WPCCS

2016

WPCCS Proceedings 2016

Friday 1st July, 2016

Contents

Full name: Dimah Al-Fraihat

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: A Comprehensive Framework for Evaluating the Effectiveness of Higher Education E-Learning Environment

Abstract: The dramatic growth and interest in providing online courses results in adopting different kinds of online learning systems in higher education institutes ranging from pure online universities that rely on e-learning as the only method of learning, to supplementary or blended ones that introduce online programs or components alongside their conventional existing programs.

With the prevalence of e-learning systems as an accepted learning strategy and delivery method, new challenges emerged to cope with this type of learning, in addition to redefining the roles of learning stakeholders and changes in learning environment.

This study aims to fully and deeply understand the factors contributing to successful evaluation of e-learning programs to help higher education institutes to design/redesign e-learning programs more efficiently, highlighting the critical factors that lead to successful and qualified e-learning programs from different e-learning stakeholders\x92 perspective (learners, instructors, faculty members).

Keywords: e-learning, evaluation, effectiveness

Abstract Word Count (Est): 141

A Comprehensive Framework for Evaluating the Effectiveness of Higher Education E-Learning Environment

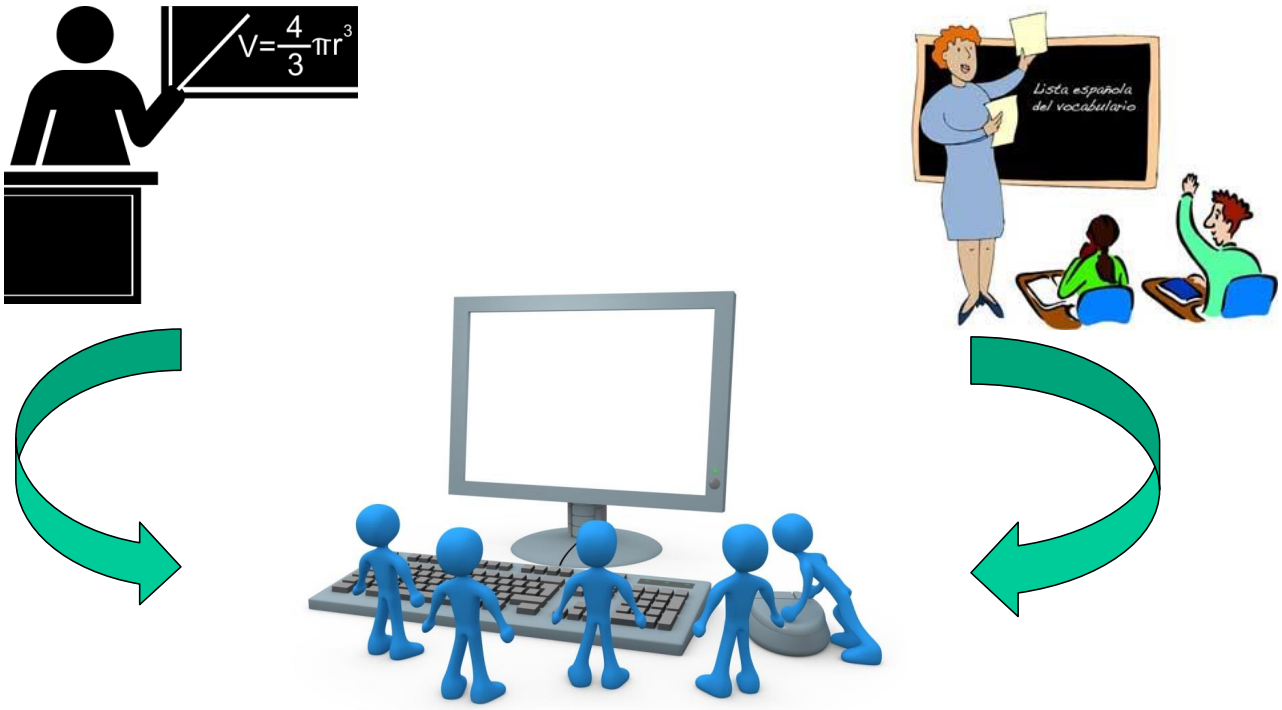
By: Dimah Fraihat
Supervisors: Mike Joy
Jane Sinclair

Contents

Overview

- The dramatic growth and increase in adopting e-learning systems.
- New challenges emerged to cope with this type of learning, in addition to redefining the roles of learning stakeholders and changes in learning environment.
- These changes debated a question by opponents who considered the quality of e-learning as subject to criticism whether this type of learning is effective or not.

- Is this type of learning capable to maintain the standards of excellence as traditional classrooms ??



The benefits and drawbacks of online learning

- No Boundaries, No Restrictions
- More Fun
- Cost Effective
- It Just Fits!

- Practical Experience
- Isolation
- Health Related Concerns

Problem Statement

- Questions related to e-learning quality standards are still unanswered in literature.
- Previous studies offer a valuable contribution for the evaluation of the quality of e-learning, but they overlook some important aspects.
- There is a need for a comprehensive framework gathering all scattered data and covering all dimensions and aspects related to e-learning quality success factors.

Why Comprehensive FW??

- The comprehensive framework is strongly needed to enable different e-learning stakeholders to examine where and how their current e-learning systems are related to each or all of the framework components.
- To provide a high-level overview of all relevant factors and dimensions that contribute to better understanding of e-learning quality success factors and utilize e-learning to its full potential.
- The factors that contribute on a better quality of e-learning systems must be framed within a big umbrella.

Significance of the Study

- This research aims to give insights for higher education institutes to strengthen the outcomes of e-learning systems, by improving planning, designing, developing, implementing, evaluating and testing of e-learning systems and components.
- Higher education institutes are expected to be the first beneficiary of this study in general and in particular e-learning developers and policy makers. They can employ the synthesized set of criteria in the proposed framework to evaluate their applications.

Research Objectives

- The ultimate goal of this research is to build an overarching framework encompassing all factors and aspects of e-learning environments that are important in one way or another to evaluating the quality of e-learning
- Rating the importance of the themes and factors to examine the relevancy of the factors by investigating the real experience from different e-learning stakeholders' point of view.

Research Questions

Research Questions

1. What are the dimensions and factors affecting the evaluation of online higher education?

Research Questions

2. What is the level of relevancy for each of these factors from the perspectives of different e-learning stakeholders?

   2.1 What are the key issues to create a successful e-learning environment?

Research Questions

3. How can the mini framework -with a range of key quality factors- be used to evaluate a selected e-learning environment?

Research Methodology

Literature Review

- Comprehensive Framework.

Data Collection

- Rank the importance of the factors in the framework.

Evaluation

- Evaluate e-course at the University of Warwick.

PhD Timeline

PhD Timeline Start Date 1-6-2015																																				
Activity	Q1			Q2			Q3			Q4			Q5			Q6			Q7			Q8			Q9			Q10			Q11			Q12		
	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
Settling	×																																			
Literature on models and Frameworks		×	×	×																																
Literature on e-learning evaluation factors			×	×	×	×																														
Top Level Framework							×																													
Research Method literature								×																												
Dimension 1,2,3,4								×	×																											
Dimension 5,6,7,8										×																										
Dimension 9,10,11,12											×																									
Writing up the detailed Framework												×																								
Writing Paper													×																							
Annual Report													×												×											
Annual Leave														×																						
Research Method Design															×																					
Initial Evaluation of the Framework																×																				
Data Collection																	×	×																		
Main Study																		×	×	×	×															
Result Analysis																																				
Writing Paper																						×														
Elicit Top Rated Factors																							×	×	×	×										
Evaluate the Framework																											×	×								
Writing Paper																													×	×						
Writing the Thesis																															×	×	×	×	×	×

Full name: Mohammed Alghamdi

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Poster

Title: Breadth First Search Level Cut

Abstract: Graph partitioning and to parallelize it is a good research topic and it has been studied for decades. The reason behind that is the inflation of graphs nowadays which cannot be processed as one graph. Therefore, it needs to be partitioned into smaller graphs. Two factors need to be considered when partitioning a graph, these includes: the balance cut and the amount of information needed to be exchanged between the subgraphs (communication cost). One way to show the balance is to use the random method of partitioning the graph. However, in many cases it cannot provide the optimal solution especially when we cannot guarantee the amount of communication cost. In our work we will use the Breadth First Search (BFS) level cut and study the effect of choosing the root on the number of partition and the communication cost.

Breadth-first search algorithm is a technique used to generally traverse in a graph. The technique carries algorithm to the solutions to many graph problems. Breadth-first search algorithm traverses the graph by breadthwards and has to use a specific queue so as help in determining the next vertex to which will be used to start the search in case there is an occurrence of a dead end in alterations.

Partitioning graph is important step when dealing with distributed graph computations. It is therefore crucial to restructure the graph layout and the BFS algorithm so as to mitigate and help avoid the problem of internode communication. In this case we use BFS algorithm has a good representation of the graph Partitioning.

Partitioning a graph usually involves the process of partitioning of its vertices so has to ensure that each vertex together with its edges and to ensure that they are owned by one processor. The vertices so that belong to the processor are the local vertices. This partitioning may involve 1-D partitioning and 2-D partitioning. In 1-D partitioning, each vertex and edges emanating from each vertex are owned by one processor. Given a vertex v, edges resulting from its edges list equals to the list of vertex indices in a given row of a matrix A. 2-D partitioning stores incident of edges on its vertices and some edges that are not.

In our work parallelizing can be done in different level. One is parallelizing using BFS, It is a principle that can be used to process all vertices which are on a single level at the same time. This means that if all the level-1 vertices are found, one can do a parallel loop that will explore from each of them to find the level-2 vertices.Another one is parallelizing the process on the sub-graph. When dealing with graph, it is often important to refer to a part of a graph, which is referred to as the sub-graph. A sub-graph has edges and vertices hence there may exist many sub-graph of a graph.This introduces a concept of graph-contraction which is used to parallelize various graph algorithms to come up with sub-graphs When this is done we end up with several sub-graphs each one being processed separately but at the same time.

Keywords: Graph partitioning, Breadth First Search

Abstract Word Count (Est): 517

Breadth First Search Level Cut

Node Degree Efficiency

Mohammed Alghamdi & Dr. Ligang He

WARWICK
THE UNIVERSITY OF WARWICK

Introduction

Graph partitioning and to parallelize it is a good research topic and it has been studied for decades. The reason behind that is the inflation of graphs nowadays which cannot be processed as one graph. Therefore, it needs to be partitioned into smaller graphs. Two factors need to be considered when partitioning a graph, these includes: the balance cut and the amount of information needed to be exchanged between the subgraphs (communication cost). One way to show the balance is to use the random method of partitioning the graph. However, in many cases it cannot provide the optimal solution especially when we cannot guarantee the amount of communication cost. In our work we will use the Breadth First Search (BFS) level cut and study the effect of choosing the root on the number of partition and the communication cost. Breadth-first search algorithm is a technique used to generally traverse in a graph. The technique carries algorithm to the solutions to many graph problems. Breadth-first search algorithm traverses the graph by breadthwards and has to use a specific queue so as help in determining the next vertex to which will be used to start the search in case there is an occurrence of a dead end in alterations. Partitioning graph is important step when dealing with distributed graph computations. It is therefore crucial to restructure the graph layout and the BFS algorithm so as to mitigate and help avoid the problem of internode communication. In this case we use BFS algorithm has a good representation of the graph Partitioning. Partitioning a graph usually involves the process of partitioning of its vertices so has to ensure that each vertex together with its edges and to ensure that they are owned by one processor. The vertices so that belong to the processor are the local vertices. This partitioning may involve 1-D partitioning and 2-D partitioning. In 1-D partitioning, each vertex and edges emanating from each vertex are owned by one processor. Given a vertex v, edges resulting from its edges list equals to the list of vertex indices in a given row of a matrix A. 2-D partitioning stores incident of edges on its vertices and some edges that are not. In our work parallelizing can be done in different level. One is parallelizing using BFS, It is a principle that can be used to process all vertices which are on a single level at the same time. This means that if all the level-1 vertices are found, one can do a parallel loop that will explore from each of them to find the level-2 vertices.Another one is parallelizing the process on the sub-graph. When dealing with graph, it is often important to refer to a part of a graph, which is referred to as the sub-graph. A sub-graph has edges and vertices hence there may exist many sub-graph of a graph.This introduces a concept of graph-contraction which is used to parallelize various graph algorithms to come up with sub-graphs When this is done we end up with several sub-graphs each one being processed separately but at the same time.

Main Objectives

1. Partition the graph in a way to speed up the process done on the graph in a cluster environment.
2. Find the relationship between the choosing root, to do the BFS, the number of portioning and the communication cost.
3. Apply two level of parallelization to increase the performance of the graph processing.
4. Reduce the communication cost between sub-graphs.
5. Speed up some current algorithm like All Pair Shortest Path (APSP) algorithm.

Materials and Methods

We will use the BFS which is an algorithm to search in different data structures. The graph is one of these data structure. As you can see in figure 1,if a BFS is applied on a graph, we will end up with tree structures which will contain a root and a number of level below it. In BFS level cut, one or more than one level will be sub-graph and the outgoing edges from a level is the information needed to be exchanged between partitions which will affect the communication cost. One benefit of BFS level cut is that every sub-graph will contact one and only one sub-graph which in many cases will be possible to aggregate the messages before they are sent and will also reduce the communication cost.

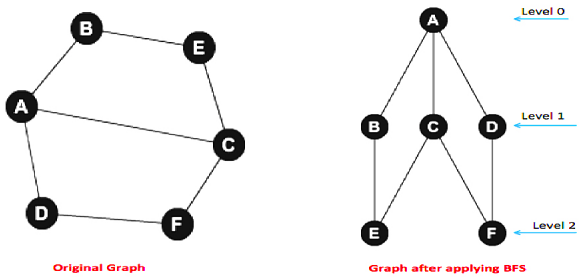


Figure 1: A graph before and after BFS

Mathematical and Results

Given an undirected graph $G = (V, E)$ were V is the vertex in the graph and E is the edge between two vertex. S is the sub-graph from G.

$$S1 + S2 + ...Si = G \tag{1}$$

We consider the number of levels in G, after applying the BFS, as L, and the number of partitions as N.

$$2 <= N <= L \tag{2}$$

There is a direct correlation between N and the communication cost, if N increases the the communication cost will increase too. just like in figure 2 if $N = 2$ - that means we partition the graph in two parts only, then we will need to exchange the information of 4 edges which are the edges between level $1and2$. where $N = L = 4$ then we need to exchange $9edges$ and that increases the communication cost. However, it could not increases the over all cost.

Another factor affect the partition process is the chosen root to start the BFS algorithm. We choose the root depending on the vertex degree to study the node degree efficiency, and we find that effaces N and L as well as the communication cost. In figure 2 we choose $Fnode$, which is a degree two node, to be the root.So, $L = 4$ and the maximum number of exchanged edges is 9.

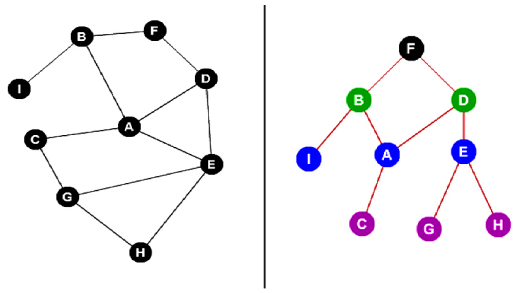


Figure 2: Left:The original graph. Right:BFS graph (degree of the root node is 2

In figure 3 we chose the highest degree node (A) and the lowest degree node (I) to be the root, and compared to average degree node F in figure 2 we found the following :- When the highest degree is the root, the number of level decreases to 3 , on the contrary, number of level is increases to 5 when the lowest degree is the root. The maximum number of exchanged edges increased to 10 in both highest and lowest degree node. Therefore, in general we can say that choosing average degree node as root will give us a good result in terms of communication cost.

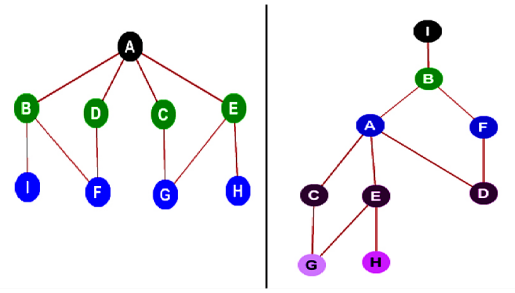


Figure 3: BFS graph. Left: highst degree node as root. Right: the lowest degree node as root

In BFS graphs we neglected the edges at the same level because it does not affect the result

Conclusions

- BFS level cut is a way to partition a graph so it can be processed in cluster environment.
- Using BFS level cut every sub-graph will contact only one sub-graph, so its possible to aggregate the messages before it is sent to help reduce the communication cost.
- The more we partition a graph the more we increase the communication cost but it does not mean that the overall cost will be increased too.
- Choosing the root depending on the degree of the node will affect the number of partition as well as the communication cost.

Forthcoming Research

We are working on applying BFS level cut on different graph and compare the results with other partitioning methods like random cut and Depth First Search. Furthermore, we are trying to create a new algorithm to increase the performance of processing a graph on different programming modules for instance the Map-Reduce and MPI and OpenMP.

References

[1] Scott Beamer, Krste Asanović, and David Patterson. Direction-optimizing breadth-first search. *Scientific Programming*, 21(3-4):137–148, 2013.

[2] Steven S Skiena. *The algorithm design manual: Text*, volume 1. Springer Science & Business Media, 1998.

[3] Hiroki Yanagisawa. A multi-source label-correcting algorithm for the all-pairs shortest paths problem. In *Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on*, pages 1–10. IEEE, 2010.

Full name: Al Alharbi

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Minimizing cost of MapReduce Jobs and Resource Allocation of Virtual Machines in the Cloud

Abstract: Cloud Computing is a demanding technology contain large data centers to provide computational infrastructures and resources over the internet .Cloud users consume the requested resources from the cloud provider with pay-as-you-go pricing plans without knowledge or specified information on the underlying infrastructure . The emphasis of processing large datasets in the future by public cloud services is estimated to be a promising future where a report estimates in 2020 almost 40% of big data analyses will be processed through public cloud services [21].MapReduce is becoming the most popular framework and used technology to analysis and store very large datasets .MapReduce is a programming model that contains of two main tasks :map task and reduce task .Input blocks are read by map tasks and then generate intermediate key and value pairs ,this intermediate key and value pairs consider as an input of reduce task. The reduce tasks merge with intermediate key and value pairs and then provide the final result of MapReduce application. Apache Hadoop framework is one of the most open source and well known platform for providing an effective execution of MapReduce programs [1].MapReduce has gained popularity in the field of academic and industry as proven to provide scalability, fault-tolerance and data integrity for handling very large datasets applications.

Keywords: MapReduce , Job Allocation ,VM Allocation ,Cloud Computing

Abstract Word Count (Est): 210

Minimizing cost of MapReduce Jobs and Resource Allocation of Virtual Machines in the Cloud

Alwaleed Alharbi
1565057

Performance Computing and Visualisation Group

WARWICK

Abstract

Cloud Computing is a demanding technology contain large data centers to provide computational infrastructures and resources over the internet .Cloud users consume the requested resources from the cloud provider with pay-as-you-go pricing plans without knowledge or specified information on the underlying infrastructure . The emphasis of processing large datasets in the future by public cloud services is estimated to be a promising future where a report estimates in 2020 almost 40% of big data analyses will be processed through public cloud services [21].MapReduce is becoming the most popular framework and used technology to analysis and store very large datasets .MapReduce is a programming model that contains of two main tasks :map task and reduce task .Input blocks are read by map tasks and then generate intermediate key and value pairs ,this intermediate key and value pairs consider as an input of reduce task. The reduce tasks merge with intermediate key and value pairs and then provide the final result of MapReduce application. Apache Hadoop framework is one of the most open source and well known platform for providing an effective execution of MapReduce programs [1].MapReduce has gained popularity in the field of academic and industry as proven to provide scalability, fault-tolerance and data integrity for handling very large datasets applications.

WARWICK

Background and issues

- Hadoop has several configuration to set up the cluster manually and therefore these setting are static .
- Some researchers develop the dynamic configuration by changing the default Hadoop files and its API scheduling.
- (Workers) are the main MapReduce tasks for processing the job where each node can responsible for number of processing of Map tasks and reduce tasks.
- Some settings can effect to the amount of memory can be given to each task .

Issues

- Performance in MapReuce jobs effected by:
- Number of nodes in the system ,Number of Map and Reduce slots ,types of nodes and allocation of mappers and reducers more efficiently.
- Scheduling (shuffling phase) Consume Time and Bandwidth between node in the cloud and money raise .

Hadoop consider the the system as Homogeneous

- Hadoop implement new schedulers such as FIFO ,Fair and capacity schedulers .therefore still the heterogeneous issue is not solved.
- Default Map slot for each node is 6
- Default reduce slot for each node is 5
- 64 mb is the default distribution size of input file.
- Each map responsible to process the amount of data by:
- Map slot = Size of input /64Mb

WARWICK

- The number of nodes can be obtained by MapReduce job is =
- Number of nodes per one MapReduce job
  = Jm/Vsm + Jr/Vsr

Jm and Jr are required to be processed for a Job.
Vsm and Vsr number of map slot and reduce slot hosted in a node.

- WARWICK

System Model

- Users submit MapReduce jobs to the cloud provider with deadline as SLA.
- Cloud provider estimate the job completion time for each type can offer (Vms Type).
- Cloud provider offer servers as heterogeneous cloud and Hadoop is installed as a service.
- Types of VMs have different price based on the performance and number of mappers and reducers.
- The cloud provider allocate VMs to the system.
- User pay the amount of money they used .

Contributions

- Minimize the cost of User's pay in the cloud when MapReduce jobs submitted by the number of mappers , reducers and number of virtual machines.
- Consider the deadline of Jobs completion time as SLA.

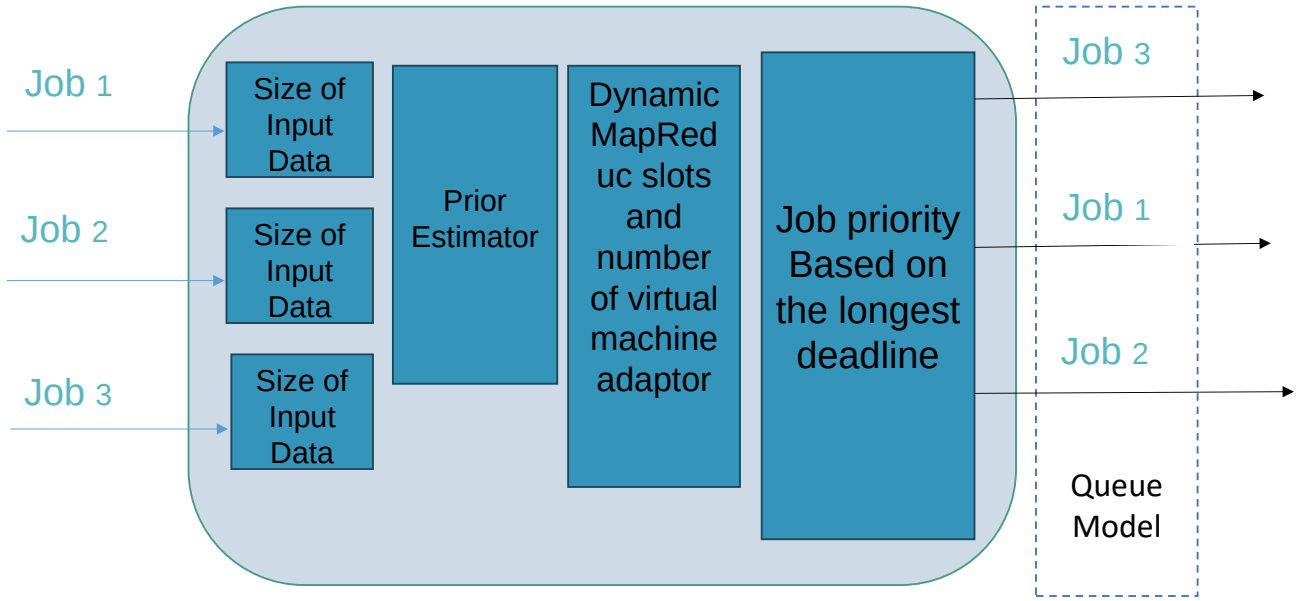
$$\text{Minimize} \quad \sum_{vn \in V} \text{Completion Time} * \text{Price of VM}$$

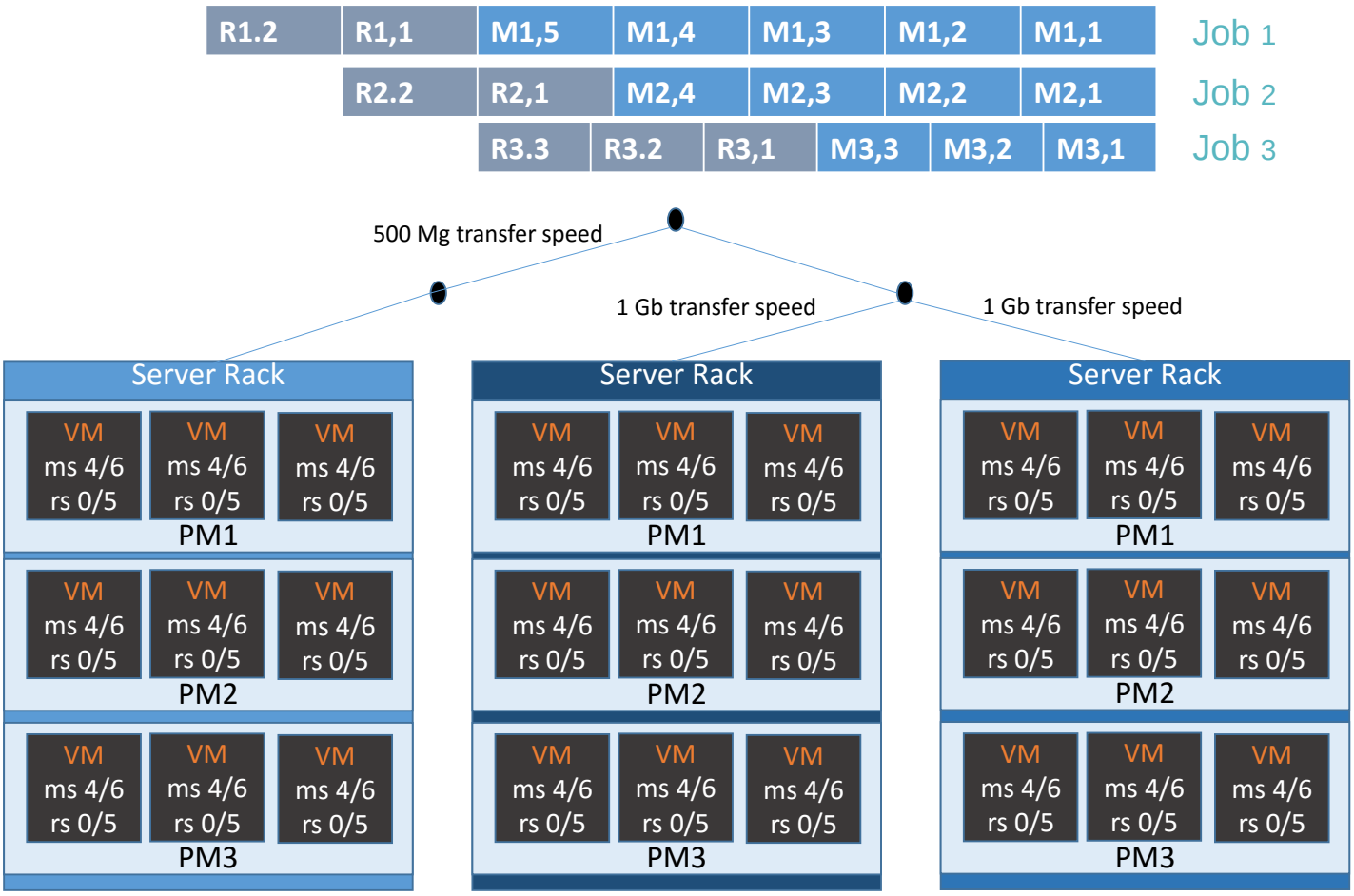
Subject to

$$\text{Completion Time} <- \text{Deadline}$$

VM type1	Price	Performance MIPS	Map Slots,Reduce Slots
Type1	$0.266 per Hour	97,125 MIPS	6,5
Type2	$0.958 per Hour	113,093 MIPS	8,6
Type3	$2.394 per Hour	238,310 MIPS	10,7

Overview Model





Experiments and considerations

- How to deal with the rejected jobs?
- Consider the running jobs during the experimental evaluation .
- Minimize the total PMs by the allocation algorithm.

Full name: Latifah Almuqren

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Twitter Analysis to Predict the Satisfaction of Telecom Company Customers

Abstract: This research is aimed at mining Arabic tweets to measure customer satisfaction toward Telecom companies in Saudi Arabia, and to predict the ratio of customer churn. This report starts with a review of previous research in using Twitter to measure user satisfaction and subjectivity analysis for Arabic. Then, it provides our approach and future plan.

Keywords: Semantic Sentiment Analysis (SSA), Arabic, Twitter, Sentiment, Customer Churn, Customer Satisfaction.

Abstract Word Count (Est): 55

Twitter Analysis to Predict the Satisfaction of Telecom Company Customers

Latifah Almuqren

Supervisor: Dr. Alexandra Cristea

Introduction

- ***Satisfied customers*** with ***company services*** make a company ***more profitable***, due to attracting new clients cost is ***five times*** more than retaining customers (Bollen et al., 2011).

- ***Social media*** is an easier tool for analysing customer satisfaction, it is more accurate tool than traditional measurement tools (Mostafa, 2013).

This presentation offers a review of the previous research on using social media to measure customer satisfaction.

Twitter

Figure1 : http://www.go-gulf.com/blog/social-media-saudi-arabia/

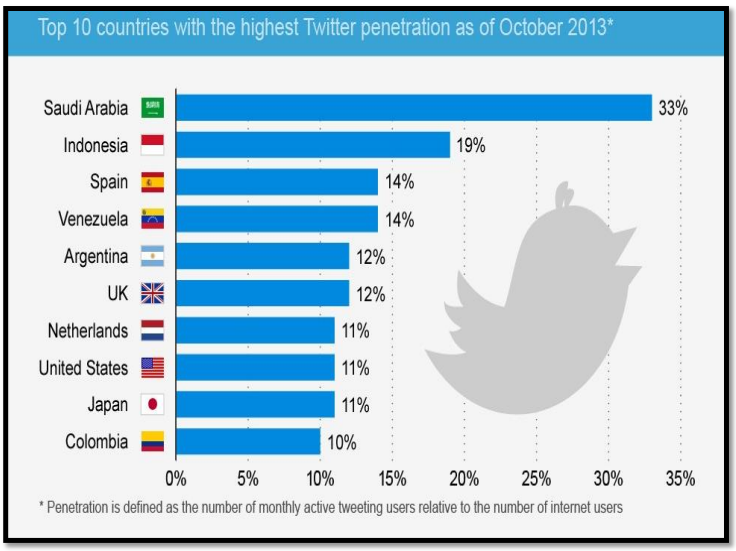


Figure2 : The United States Only Ranks 8th in Twitter Penetration (Richter, 2013)

Providing Customer Satisfaction with Social Media

- Figure 3 shows the ***customer satisfaction model*** containing all variables that link Twitter features, customer satisfaction and customer churn.

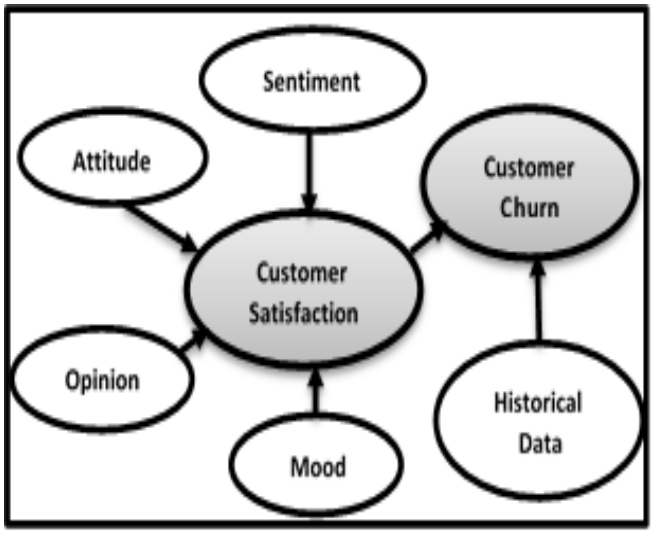


Figure 3: Our customer satisfaction model

Providing Customer Satisfaction with Social Media

- The ***most popular means*** to gather and measure ***customer satisfaction*** is through surveys.

- However, ***Data mining applications*** provide useful approaches ***to measuring customer satisfaction*** without surveys.

Literature Reviews

- Customer churn and data mining techniques:

Brandusoiu and Toderean (2013)	• Build three predictive model based in classification method to predict customer churn in telecommunication companies. • They applied three algorithms which are K-nearset neighbour, logistic regression, and Bayesian networks. • From a practical point of view, all three models have a very good performance (around 80%) in predicting churners in a mobile telecommunications company.

Literature Reviews

❖Customer satisfaction and social media:

Collines et al., (2013)

- Measured public transport rider satisfaction towards transit system services using the riders' tweets.
- This helped to improve their service quality and safety monitoring.
- The findings showed that sentiment analysis can successfully detect rider sentiments in real time toward a transit organization.

Salampasis et al., (2013)

- Analysed consumer behaviour related to food products using their microblogging messages, i.e., Tweets.
- The results showed that the success of branding can be observed by sentiments monitoring for a long time period.

Literature Reviews

❖Customer satisfaction and social media:

Sassy and Miranda (2014)

- The researchers used sentiment analysis to propose a tool for evaluating customer satisfaction in the job search company.
- The results showed that over 42% of the company's clients have positive impressions about the provided services, and 34% did not express any feelings in their remarks.

Mostafa (2013)

- Mining random tweets in Twitter to find the consumer's sentiments toward some brands by sentiment analysis .
- The findings proved that there is a positive consumer sentiment toward famous brands

Literature Reviews

❖Prediction based on Twitter sentiment:

Tsakalidis et al., (2014)

- They sought to predict the 2014 European Union based on Twitter use and opinion polls by using a lexicon-based classifier for sentiment analysis.
- The results achieved better than several baselines, including polls, prediction websites, and replication of old works.

Kampakis and Adamides (2014)

- They predict the score of the football match from twitter mining and historical data by creating three models: Performance Model based on Tweeter features, a model based on historical data, and the combined of two models.
- The results demonstrated a high accuracy if features from both datasets are combined with the final accuracy going up to 75%.

Arabic Sentiment Analysis

➢There are some attempts to build a wide polarity Arabic lexicon of Modern Standard Arabic (MSA) (Ahmed et al., 2013).

➢In contrast, there are few attempts to build an Arabic dialect lexicon, especially a Saudi dialect lexicon (Azmi and Alzanin, 2012).

➢Unfortunately, current subjectivity and sentiment analysis tools are designed mainly for the English language, and there is a dearth of tools for the Arabic context.

Research Problem

Twitter is one of the most influential sites among Saudi people (Al-Saggaf and Simmons, 2015).

Enhancing customer satisfaction is a popular topic in marketing.

Traditionally, customer satisfaction has been measured through customer interviews and questionnaires, but these cannot measure the phenomena in real time (Duwairi, and Qarbaz, 2014).

Therefore, new research is required to measure customer satisfaction based on real-time methods.

This study seeks to mine microblogging sites for the purpose of capturing user satisfaction toward telecom companies and determine how we can use such insight to offer recommendations to these companies, as well as to predict the ratio of customer churn.

This study intends to contribute to Arabic Sentiment Analysis (ASA) by building an Arabic dialect lexicon.

Research Objectives

1. To define and investigate measurable criteria for customer satisfaction towards Saudi telecommunication companies.
2. To identify, based on the above criteria, and Twitter mining Saudi telecom companies' customers' satisfaction toward company services.
3. To propose recommendations to improve the services of Saudi telecom companies.
4. To predict the potential ratio of customer churn.
5. To develop a software to help decision makers in the Saudi telecom companies to capture user satisfaction toward telecom companies.

Methodology

Approach and Methods

- The customer satisfaction model is based solely on customer sentiment, measured from Twitter, via Semantic Sentiment Analysis (SSA).

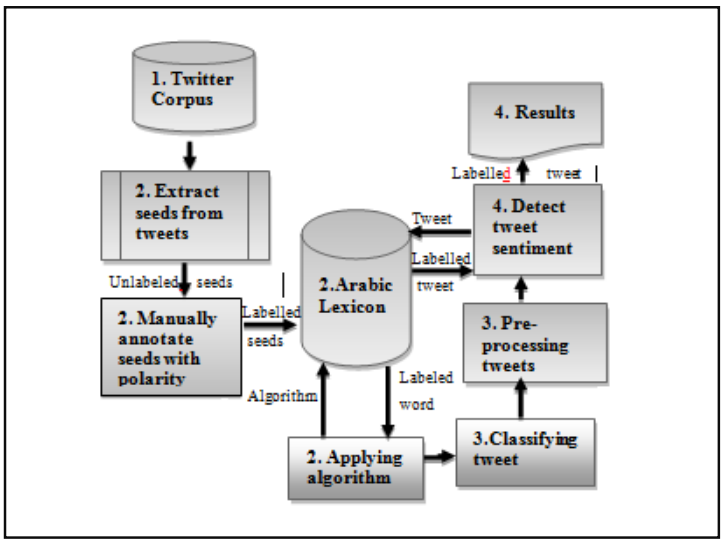
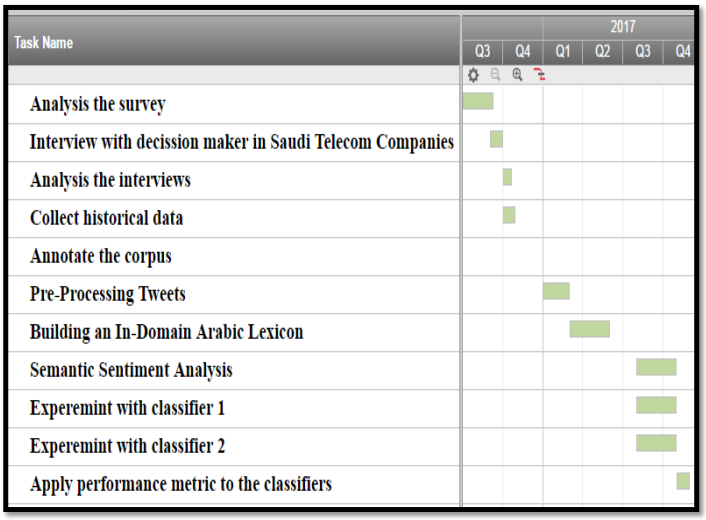


Figure 5: The Using Framework

Future Plan and Timeline



References

[1] Ahmed, S., Pasquier, M. and Qadah, G., 2013, March. Key issues in conducting sentiment analysis on Arabic social media text. In *Innovations in Information Technology (IIT), 2013 9th International Conference on* (pp.72-77). IEEE.

[2] Anon. Social Media Saudi Arabia- Statistic and Trends. [online] Available from: http://www.go-gulf.com/blog/social-media-saudi-arabia/ (Accessed 4 June 2016).

[3] Al-Saggaf, Y. and Simmons, P., 2015. Social media in Saudi Arabia: Exploring its use during two natural disasters. *Technological Forecasting and Social Change*, *95*, pp.3-15.

[4] Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), pp.1-8.

[5] Collins, C., Hasan, S. and Ukkusuri, S.V., 2013. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*, *16*(2), p.2.

[6] Duwairi, R.M. and Qarqaz, I., 2014, August. Arabic Sentiment Analysis using Supervised Classification. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on* (pp. 579-583). IEEE.

[7] Kampakis, S. and Adamides, A., 2014. Using Twitter to predict football outcomes. *arXiv preprint arXiv:1411.1243*.

[8] Mostafa, M.M., 2013. More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications*, *40*(10), pp.4241-4251.

[9] Salampasis, M., Paltoglou, G. and Giachanou, A., 2013. Using social media for continuous monitoring and mining of consumer behaviour. *International Journal of Electronic Business*, *11*(1), pp.85-96.

[10]Tsakalidis, A., Papadopoulos, S., Cristea, A.I. and Kompatsiaris, Y., 2015. Predicting elections for multiple countries using Twitter and polls. *Intelligent Systems, IEEE*, *30*(2), pp.10-17.

[11] Richter, F. The United States Only Ranks 8th in Twitter Penetration. [online] Available from https://www.statista.com/chart/1629/twitter-penetration/ (Accessed 4 June 2016).

Thank you

Full name: Najah Alsubaie

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Stain Deconvolution using a Combined Colour and Texture Framework

Abstract: With growing interest in digital slide scanners, histology image analysis is rapidly emerging as an active area of research. Several histology image analysis algorithms, such as those for mitotic cell detection, nuclei segmentation, and hormone receptor scoring, use the estimation of a stain colour to extract the required components from the image. For example, in most of the nuclei segmentation algorithms, only areas stained with Haematoxylin stain are relevant for further processing. These stain-colour based algorithms require reliable estimation of a specific stain channel out of the histology image. In this paper, we propose a novel method to estimate the stain matrix for stain deconvolution by employing both colour and textural information using the wavelet decomposition of the histology image. Stain matrix is estimated by statistically analysing the multi-resolution subcomponents of the image and incorporating that into a blind colour deconvolution framework. We conducted extensive experiments to compare the proposed method to recent state-of-the-art algorithms. Robustness of the proposed method is evaluated by performing the experiments using three different datasets of scanned slides, prepared in different labs and by different scanners.

Keywords: Histology, cancer diagnosis, stain deconvolution,wavelet decomposition, ICA

Abstract Word Count (Est): 181

Stain Deconvolution using a Combined Colour and Texture Blind Source Separation Framework

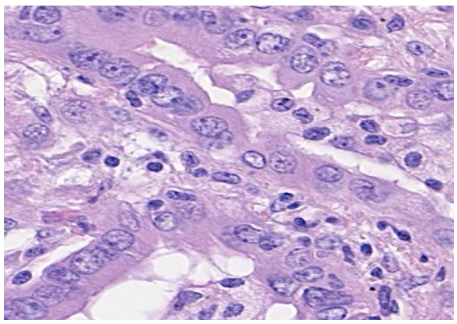
Najah Alsubaie

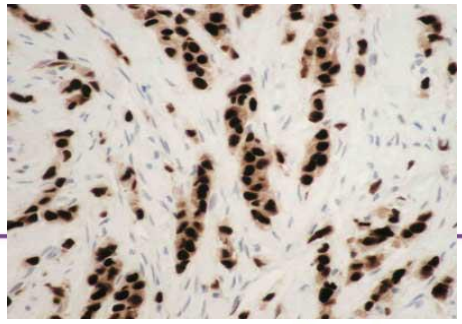
Bialab

1/7/2016

Introduction

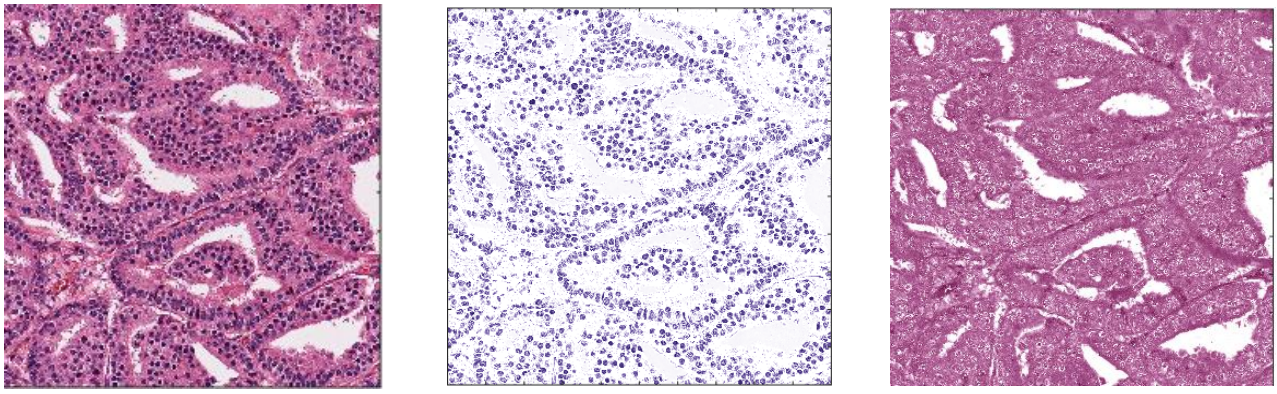
- Histology image analysis is rapidly emerging as an active area of research
  - Mitotic cell detection
  - Nuclei segmentation
  - Hormone receptor scoring





Motivation

- Use the estimation of stain colour to extract the required components from the image.



Methods

- Signal mixture $x = x_1, x_2, x_3, \ldots x_i, i >= 1$ is represented as linear combination of source signals:

$$s = s_1, \ldots, s_j, j = 1, 2, \ldots r >= 2$$

mixed by mixing parameters,

$$\begin{bmatrix} x_1 \\ x_2 \\ .. \\ x_i \end{bmatrix} = \begin{bmatrix} m_{1,1} & m_{1,2} & \ldots & m_{1,j} \\ m_{2,1} & m_{2,2} & \ldots & m_{2,j} \\ & & & \ldots. \\ m_{i,1} & m_{i,2} & \ldots & m_{i,j} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ .. \\ s_j \end{bmatrix}$$

simply,

$$\mathbf{x} = \mathbf{M}\mathbf{s}$$

Independent Component Analysis

- Tow main assumptions:
  - Sources are statistically independent.
  - At least one has non-Gaussian distribution.
- Independency among sources is a strong assumption that might not always be satisfied[1,2].

Theory of the Proposed Approach

- Raw source signals might be dependent
  - Their sub-signals can be independent of each other.
  - Representing the source signal as set of independent and dependent sub-components.

$$s = s_{j,1}, s_{j,2}, s_{j,3}, \ldots s_{j,k}, j = 1,2,\ldots r, r >= 2$$

- Applying filter *f* to allow independent sources to pass through. $s_f = fs$
- By applying f to the observed signal,

$$x_f = fx$$

Theory of the Proposed Approach

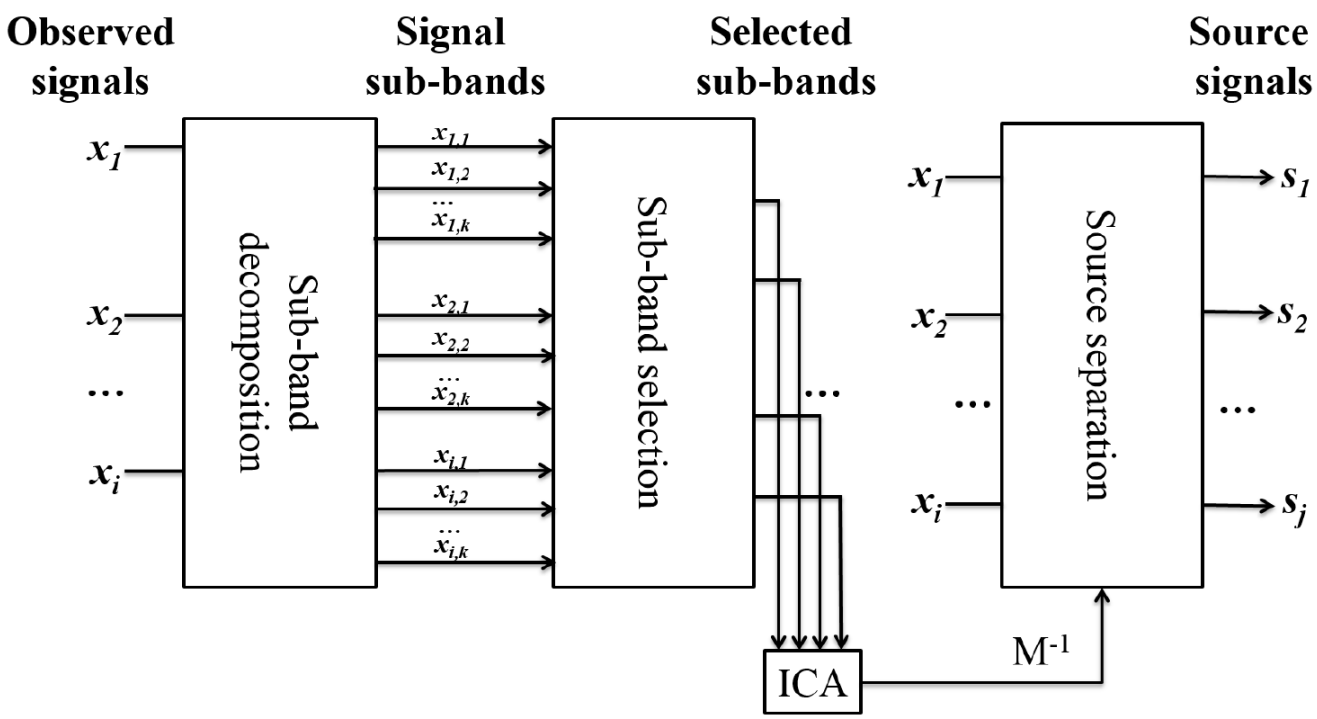
- Recall,

$$X = MS$$

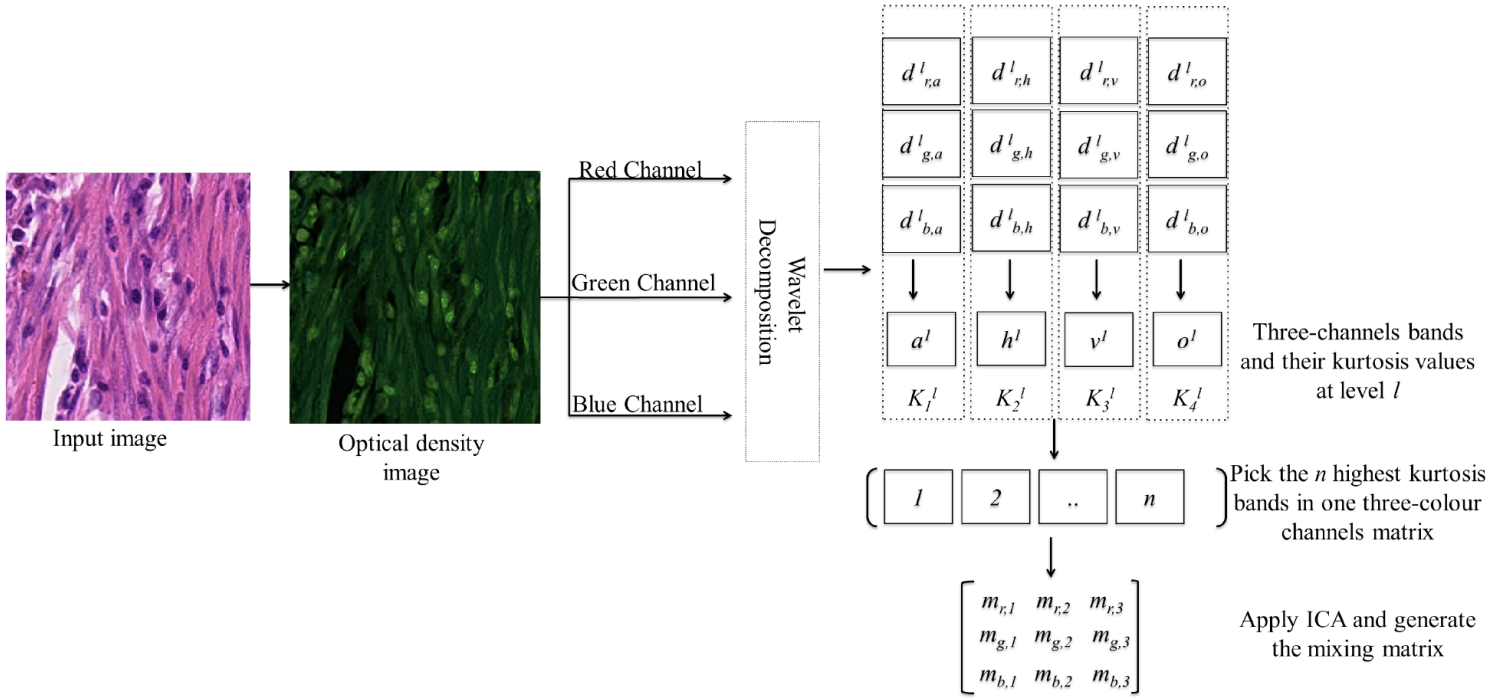
- Then, independent subcomponents of source signals could be used to generate the mixing matrix *M*

$$x_f = fx = fMs = Mfs = Ms_f$$

Theory of the Proposed Approach



Proposed Approach



Experiments and Results

- Data set preparation
- Evaluation experiments

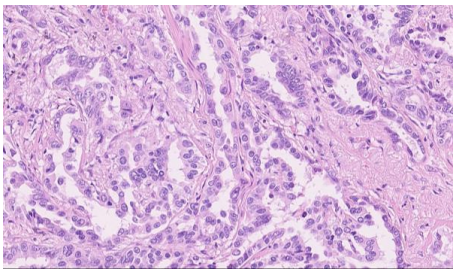
Data sets

- Stain chemicals bind differently to different tissue types
- Evaluate a stain deconvolution algorithm using different tissue types

Data Sets

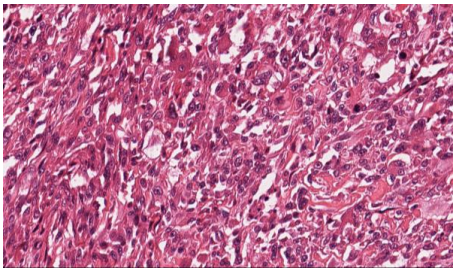
Lung Cancer Histology Images

- 40x magnification
- Omnyx VL120 scanner at UHCW.
- Two H&E whole-slide images
- Two non-overlapping 2,000x2,000 images



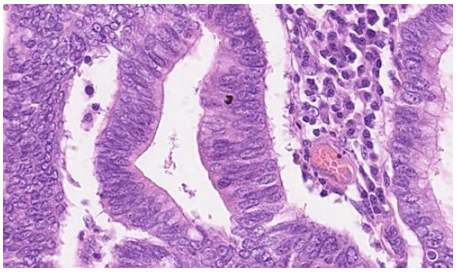
Breast Cancer Histology Images

- 40x magnification
- Aperio ScanScope XT scanners at Pathology Department of the University Medical Center Utrecht, The Netherlands
- Three H&E whole-slide images
- Two non-overlapping 2,000x2,000 images.



Colon Cancer Histology Images

- 20x magnification
- Omnyx VL120 scanners UHCW
- Seven H&E whole slide images
- Two non-overlapping 500x500 images



Experiments

- Evaluating the Estimated Stain Matrix
- Evaluating Density Map Estimation
- Assessment of Tumour Nuclei Detection using the Estimated H channel

Evaluating the Estimated Stain Matrix

- Mixing matrix ground truth (M)
  - Select pixels for each stain colour based on the biological structure they contribute to.

$$\begin{bmatrix} m_{1,1} & m_{1,2} & \dots & m_{1,j} \\ m_{2,1} & m_{2,2} & \dots & m_{2,j} \\ & & & \dots. \\ m_{i,1} & m_{i,2} & \dots & m_{i,j} \end{bmatrix}$$

- Compute the Euclidean distance between the estimated stain vector and the ground truth stain vector.

Evaluating the Estimated Stain Matrix

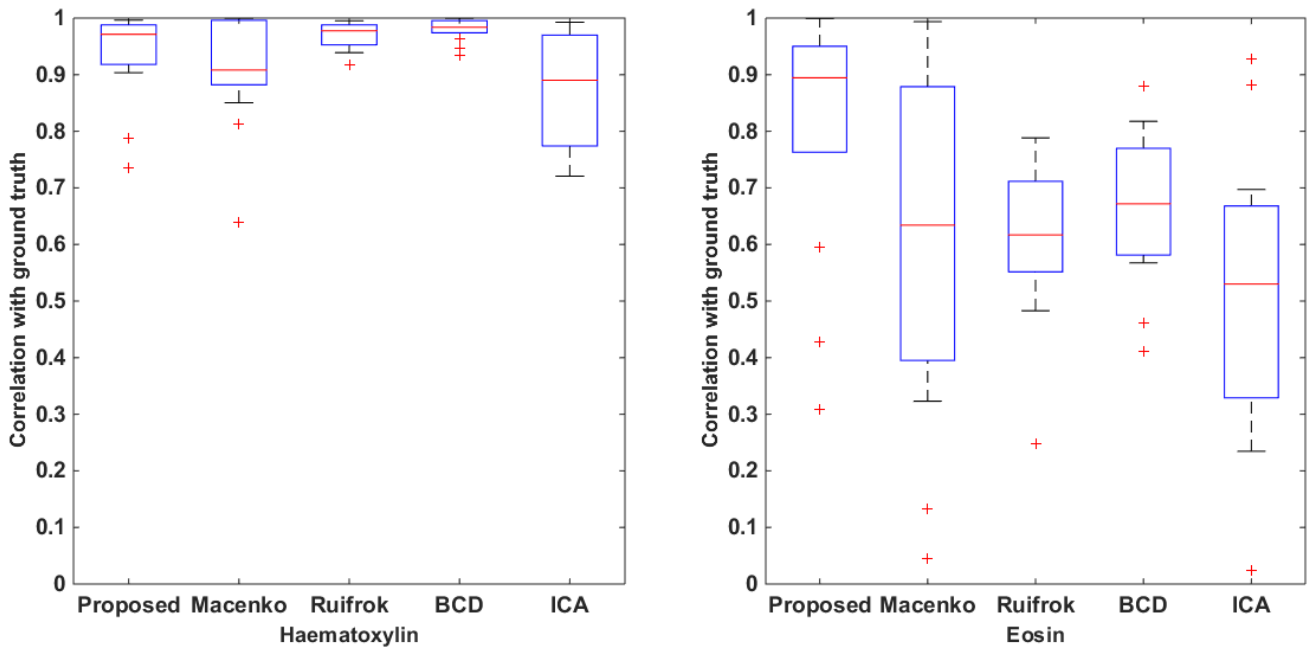
	Colon Dataset		Breast Dataset		Lung Dataset		Median	
	H	E	H	E	H	E	H	E
Proposed	**0.0623**	**0.0871**	**0.0537**	0.0585	0.1377	**0.0779**	**0.0592**	**0.085**
Macenko *et al.* [8]	0.0774	0.2002	0.1202	**0.0488**	**0.0936**	0.9949	0.0971	0.4146
Ruifrok Ruifrok and Johnston [7]	0.1980	0.2017	0.2201	0.3150	0.1826	0.1917	0.2002	0.2361
BCD [10]	*0.3789*	0.3282	0.4027	0.3606	0.3897	0.3655	0.3905	0.3514
ICA [9]	0.2795	*0.3349*	*0.5219*	*0.5607*	*0.4081*	*1.2451*	*0.4032*	*0.7136*
Median	0.198	0.2017	0.2201	0.315	0.1826	0.3655	—-	—-

Evaluating Density Map Estimation

- Distribution of a particular stain across the section.

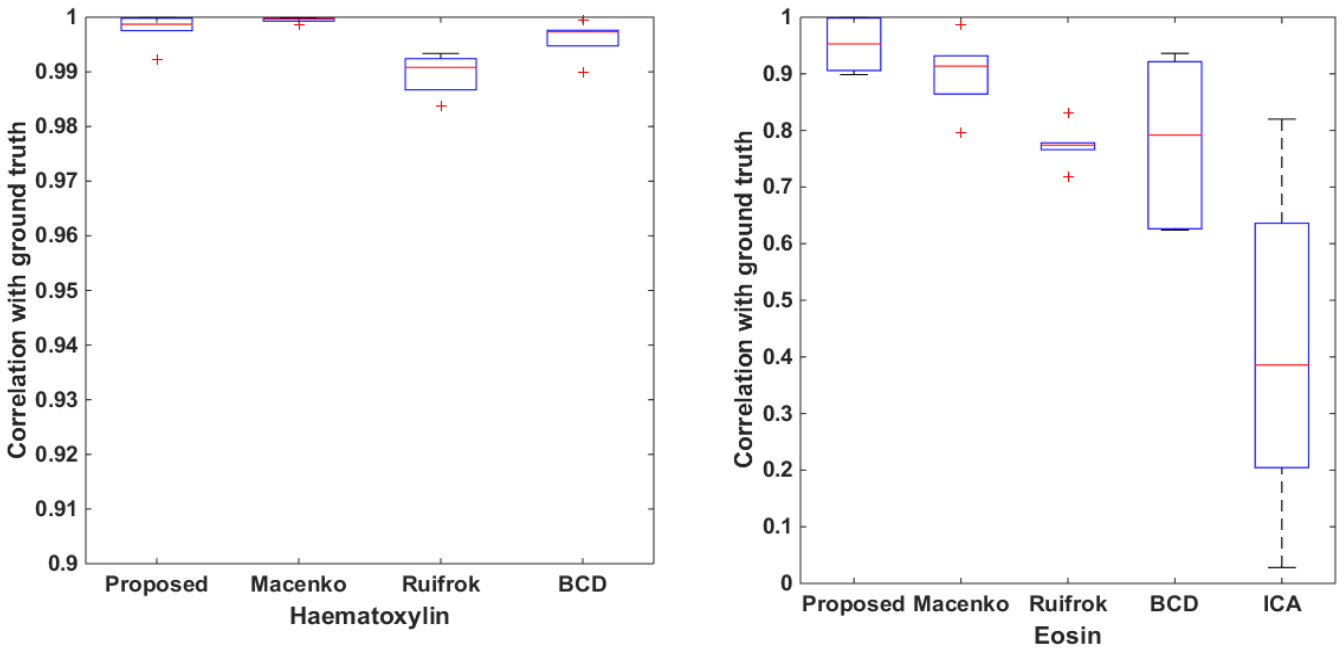
$$X = MS$$

Colon Dataset



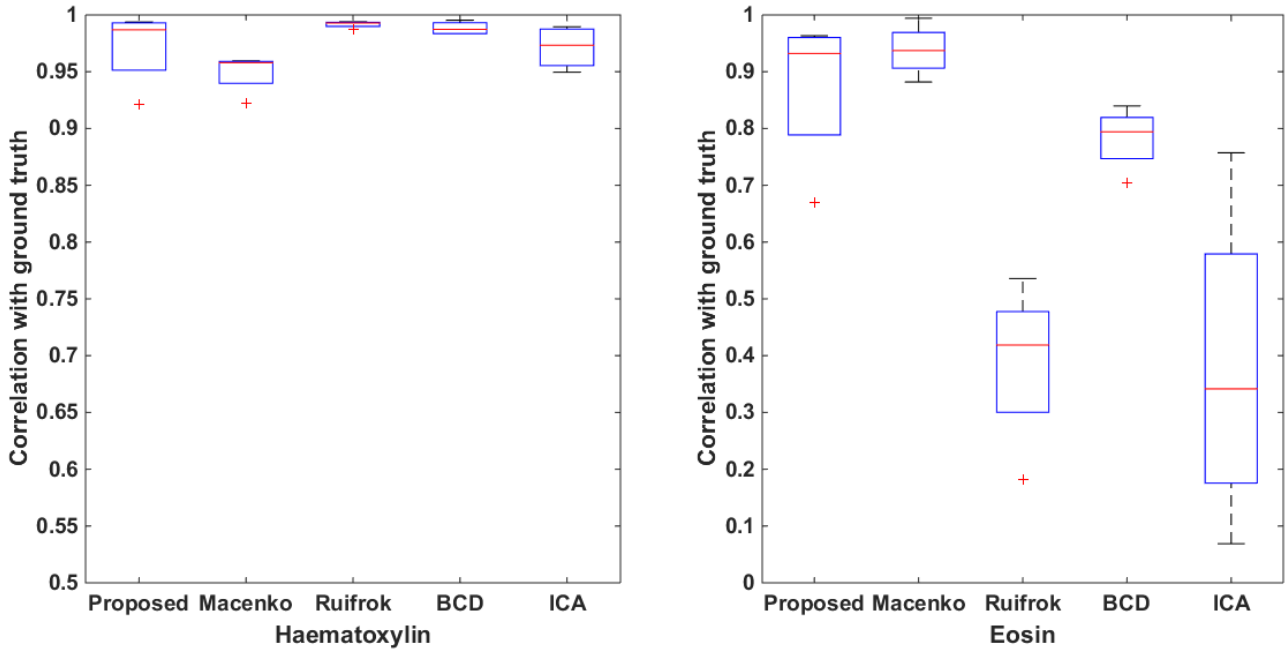
Correlation Between the Density Maps and the Ground Truth.

Breast Dataset



Correlation Between the Density Maps and the Ground Truth.

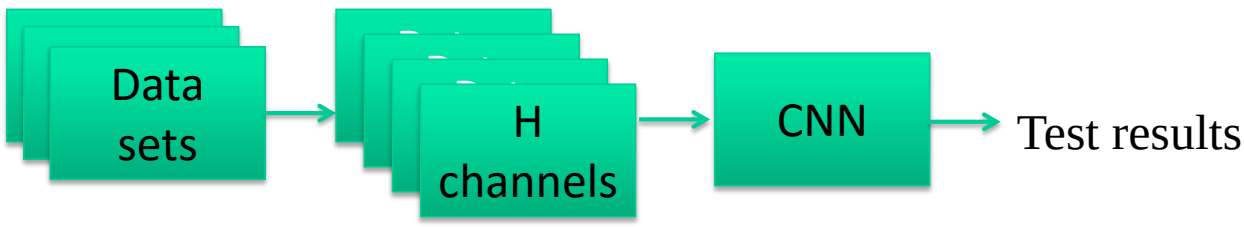
Lung Dataset



Correlation Between the Density Maps and the Ground Truth.

Assessment of Tumour Nuclei Detection

- Evaluate the generated stain channel .
- Spatially Constrained Convolutional Neural Network (SCCNN) to detect the centers of nuclei in colon histopathology images.
  - Uses H channel



Stain deconvolution Train network

Assessment of Tumour Nuclei Detection

	Precision	Recall	F1 score
Proposed	**0.838**	0.369	**0.512**
BCD [10]	0.617	0.269	0.375
Macenko *et al.* [8]	0.323	**0.488**	0.389
ICA [9]	0.217	0.383	0.277

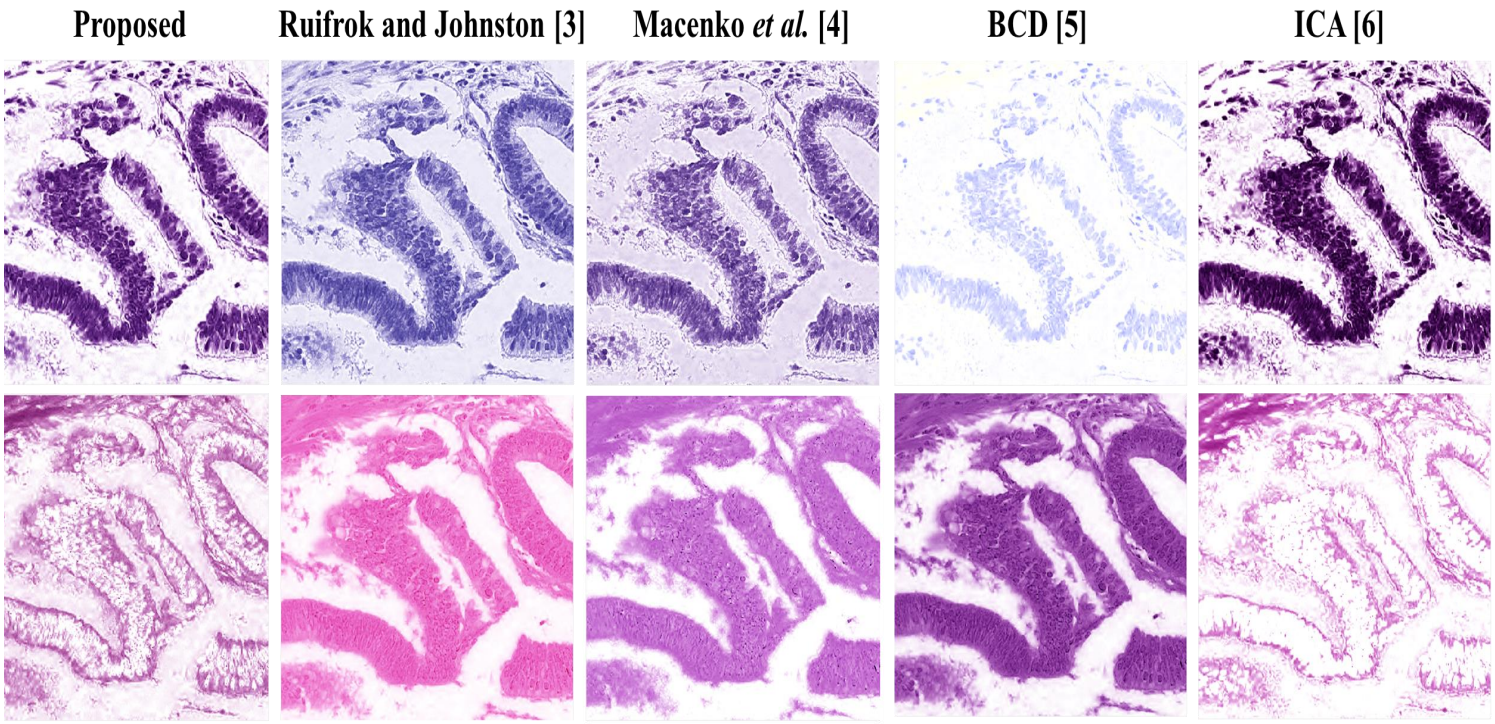
Number of Bands

- Using the data sets, we performed this experiment using
  - 5, 10,15, and 20 number of subbands.

Number of Bands

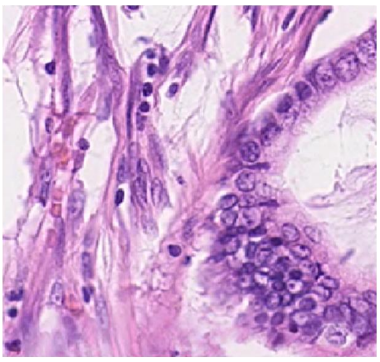
	Number of sub-bands							
	5		10		15		20	
	H	E	H	E	H	E	H	E
Colon Dataset	**0.0467**	0.0938	0.1113	0.0877	0.1316	0.1051	0.0623	**0.0871**
Breast Dataset	0.1490	0.1519	0.1077	0.1458	0.1304	0.1415	**0.0537**	**0.0585**
Lung Dataset	0.0589	0.0980	0.0457	0.0850	**0.0372**	0.0925	0.1377	**0.0779**

Qualitative Results

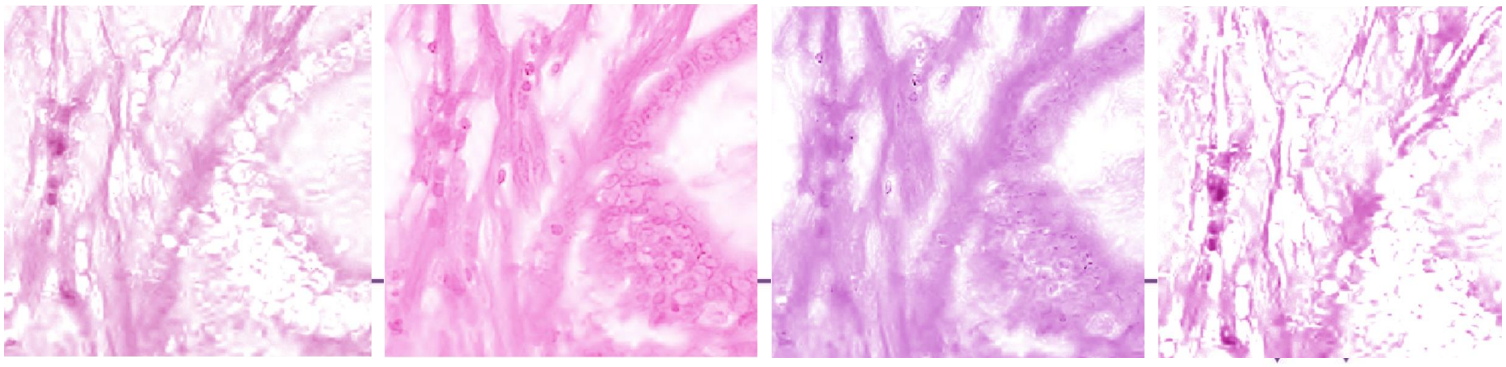


Qualitative Results

Original image



Proposed | Ruifrok and Johnston [3] | Macenko *et al.* [4] | BCD [5]



Conclusions

- Stain deconvolution is a significant pre-processing step in most of the histology image processing.
- We proposed a novel algorithm to perform stain deconvlution using the underling texture and combine that into a colour mixing model.
- Proposed algorithm could be improved by automatically adjusting the number of bands required in the deconvolution step.

Referencces

1. R. Celis, D. Romo, and E. Romero, “Blind colour separation of H&E stained histological images by linearly transforming the colour space.”Journal of Microscopy, 2015.
2. M. Gavrilovic, J. C. Azar, J. Lindblad, C. Wahlby, E. Bengtsson, C. Busch, and I. B. Carlbom, “Blind color decomposition of histological images.” IEEE transactions on medical imaging, 2013.

Acknowledgments

- Bilab group:
  - Nasir Rajpoot
  - Shan-e-Ahmed Raza
  - Korsuk Sirinukunwattana
  - Nicholas Trahearn
  - Mike Tzu-Hsi Song
- University hospitals Coventry and warwickshir
  - David Snead

Full name: James Archbold

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Indirect Concept Spread Maximisation & Diverse Spreading Dynamics

Abstract: The propagation of concepts in a population of agents is a form of influence spread, which can be modelled as a cascade from an initial set of individuals. In real-world environments there may be many concepts spreading simultaneously, through a variety of dynamics. Previous work typically does not consider the interactions between concepts with different spreading dynamics or attempting to indirectly increase the spread of a target concept through interactions with a concept that we can control. We present a generalised model of influence spread, through which we define an environment that supports diverse spreading dynamics. We discuss the problem of indirect concept maximisation and present the preliminary results from ongoing research. Initial results indicate maximising the boosting concept's spread also has the best result for the target concept, as opposed to using heuristics targeting points in the network that can be easily reached by the target concept.

Keywords: agents, influence, multi-agent systems, influence models, indirect influence

Abstract Word Count (Est): 148

Indirect Concept Spread Maximisation & Diverse Spreading Dynamics

James Archbold (archbold@dcs.warwick.ac.uk)
Nathan Griffiths (nathan@dcs.warwick.ac.uk)

Department of Computer Science, University of Warwick,
Coventry, CV4 7AL, UK

May 25, 2016

Introduction: Influence Spread

- Concepts can spread through social networks as individuals are influenced by their neighbours.
- Independent Cascade (IC): Fixed chance to be influenced by each neighbour.
- Linear Threshold (LT): Influenced based on proportion of neighbours influenced.
- Applications; the spread of epidemics or diseases, viral marketing, modelling convention emergence.

figures/infectionExample.png

Introduction: Multiple Cascades

- Influence Maximisation: If we have a budget of k nodes, how to select the k nodes that will maximise the potential spread of our chosen concept?
- Most research on influence spread has focused on single concept cascades.
- Some work on multiple cascades in single networks, but typically assume *blocking cascades* — nodes can have single concept active.
- Little work on multiple activated concepts — need to consider how concepts interact.
  - Sanz *et al.* define a model for 2 diseases interacting, affecting each other's spread.
  - Considers the diseases the *receiver* and the *infector* have active when an *infector* attempts to spread either disease to the *receiver*.

Introduction: Indirect Concept Spread Maximisation

- Assumptions:
  - An individual/node could hold more than one concept.
  - The concepts/opinions held by a node affect the concepts they adopt or spread to others.
- Can we make use of a *boosting concept* to improve the *target concept*'s chance to spread?
- What if these concepts utilise different *spreading dynamics*?

Generalised Model: Need for Model

- Many different spreading dynamic models for how concepts may spread through a network
- Different models use different attributes and interaction rules
- Results in need for a unique solution to problems in each model
- Difficult to compare models and how they may interact

Generalised Model: Requirements

- Our model must describe nodes, their attributes, concepts and how concepts interact
- Must be robust to enable the description of a wide variety of spreading models in a common language
- General model would allow for a uniform method of concept interaction between different spreading dynamics

Generalised Model: Node Description

- Each node, v, in the network has a list of attributes A_v
- Incoming neighbours, N_v^i, and outgoing neighbours, N_v^o
- Each node can have multiple concepts active, set C_v^t, describes concepts active on node v at time step t
- *Adopting Context* - $Context_{adopt}^{v,t}(c)$, describes how concepts active on v at time t affect chance to adopt concept c
- *Spreading Context* - $Context_{spread}^{v,t}(c)$, describes how concepts active on v at time t affect chance to spread concept c

Generalised Model: Concept Relationships

- Concepts can interact and affect each other's ability to spread
- *Concept relationship functions:*
  - $CR_{adopt}(c, c')$, describes the effect of c' on the chance that concept c is adopted by a *receiver* with c' active
  - $CR_{spread}(c, c')$, describes the effect of c' on the chance that c is successfully activated on a receiver when the *infector* has c and c' active.
- Each concept has set of concepts it blocks. If a concept is active on a node, no concept it blocks can simultaneously be active on that node.
- May be used by node context functions, but concept relationship functions are node independent

Generalised Model: Concept Spreading

- Relationships between nodes defined using the *influence strength*, $I^c_{m,n}(t)$, that node m can exert on node n at time t for concept c.
- The *total influence* exerted on a node at time t relating to concept c is represented by $TI^c_v(t)$
- In each time step, we update the set of active concepts for each node. For each concept, we evaluate the condition required for that concept to become active.
- For node v at time step t these updates are described as $C^t_v = C^{t-1}_v \cup c : X$, where X is the condition that must be true for c to become active on v

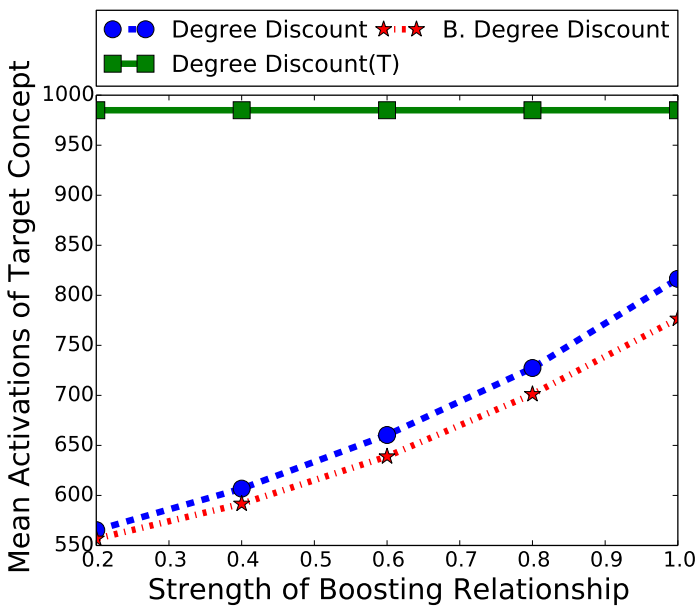
Indirect Concept Spread Maximisation: Definition

- Typically, the *spread maximisation* problem focuses on selecting seeds for a target concept
- May desire to spread a concept that we cannot directly control
- Want to see the effectiveness of placing a secondary concept to maximise the spread of a target concept
- Different selection criteria for seed nodes, we do not care about how far spread *boosting concept* is, only how it can improve spread of *target concept*

Indirect Concept Spread Maximisation: Current Work

- Currently evaluating how established heuristics perform in new problem, compared to their performance in the traditional problem.
- Using concepts that spread using the Independent Cascade spreading dynamic
- Also adapting established heuristics to value seeds based on indirect concept spread maximisation instead of normal spread maximisation
- Goals include considering reachability of a node as opposed to network reach of that node

Indirect Concept Spread Maximisation: Current Results



- Indirect concept spread maximisation seems to not provide the same level of target concept spread as normal concept spread maximisation
- If indirect concept spread maximisation is the only option, established heuristics perform better than adapted heuristics
- Suggests maximising the spread of the boosting concept provides better results than targeting nodes to be reached more easily by the target concept

Indirect Concept Spread Maximisation: Next Step

- Need to perform more robust tests to increase confidence in the result
- Then extend experiments to include concept spread limitation
- Utilise the generalised model to allow for diverse spreading dynamics, so that the relationship between different spreading dynamics can be better analysed
- Hope to facilitate the development of a general heuristic that can be applied to many spreading dynamics

Summary

- The area of influence spread has a wide range of influence spreading models, making it difficult to compare and analyse their interactions
- Through a generalised model, we can facilitate these interactions and better analyse an avenue to apply uniform environmental factors to these spreading dynamics
- Indirect concept spread methods do not currently seem to provide better performance then traditional concept spread maximisation methods.
- Effective indirect concept spread maximisation can help in situations that our target concept, such as a political opinion or product loyalty, cannot be directly accessed or controlled.

Future Work

- Non-disjoint multi-networks, attempting to spread a concept through a network that cannot be directly accessed.
- Gathering real world data to use for networks to further test out these models on real world data.

Full name: Denys Alberto Flores Armas

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Proactive Database Forensics in the BYOD Era - An Overview

Abstract: Bring-Your-Own-Device (BYOD) is an organisational trend which allows employees to use their own mobile devices in order to access corporate information assets. Although this practice increases employee productivity at work, it also poses security concerns that may expose corporate information to unauthorised disclosure and contamination. This problem is more evident when information stored in Databases becomes the attack target, urging organisations to investigate related incidents, and if possible, sanction the culprits by using relevant digital evidence about suspicious database activity. However, BYOD challenges these investigations due to the ever increasing number of uncontrolled evidence sources, and the lack of proactive database forensic approaches that can be applied to overcome the limitations of traditional digital forensic techniques in distributed environments. Hence, our current work examines the importance of developing a proactive approach to database forensics from the BYOD perspective, considering the current digital forensic challenges in the field, and introducing a STRIDE-based Threat Model for analysing BYOD threats that may hinder the investigation of disclosure and contamination of corporate information. We expect our current work can justify future research related to chain of custody provenance requirements and evidence causality when investigating databases in a BYOD environment.

Keywords: BYOD, database, information security, digital forensics, database forensics, proactive, STRIDE, threat model

Abstract Word Count (Est): 195

Proactive Database Forensics in the BYOD Era - An Overview

Presenter: Denys A. Flores
Supervisor: Dr. Arshad Jhumka
July 1st, 2016

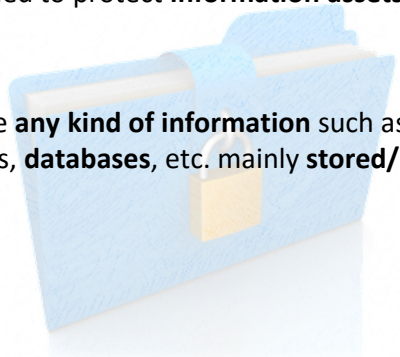
Topics

- A Brief Introduction to Information Security & Digital Forensics
- The Importance of Digital Forensics in Distributed Systems
- Challenges for Investigating Databases (DB Forensics)
- DB Forensics: Reactive vs. Proactive Approaches
- The BYOD Threat to Proactive DB Forensics
- Introducing a STRIDE-based BYOD Threat Model
- Related Work
- Conclusions

Information Security & Digital Forensics

- **Information Security** manages the **impact of risks and incidents** when **threats exploit a vulnerability** within an **information system (asset)**.
- **Controls** must be applied to protect **information assets** and **information security**.

(BS ISO/IEC, 2016)



- **Information Assets** are **any kind of information** such as documents, transcripts, audio/visual recordings, **databases**, etc. mainly **stored/processed/transmitted using electronic media**.

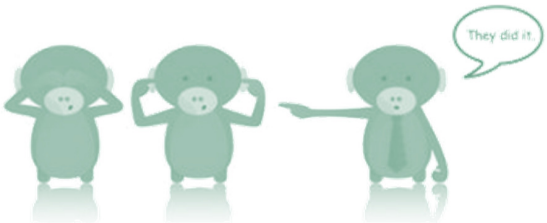
(Escrivá et al, 2013)

It is important to secure information at rest, being processed and in transit.

Information Security & Digital Forensics

- **Controls** may be policies, procedures, processes, organizational structures, hardware and software in order to **prevent materialisation of [internal/external] threats.**
- When **controls fail** an **information security incident** is introduced as **confidentiality, integrity and availability of information assets are compromised.** (BS ISO/IEC, 2016)
- **Accountability** is a security characteristic for **forensics and auditing purposes .**

(Stallings, 2011)



When there is an incident there is someone accountable for it!

Importance of Digital Forensics in Distributed Systems

- **Digital Forensics** is a forensic science used to identify, collect, preserve, analyse and present **digital evidence stored in electronic media** (information assets) during **legal proceedings** (Palmer, 2001).
- **Auditing** logs and monitors actions to **ensure compliance** (Andress, 2014)
- **In organisations, digital Forensics relies on auditing** in order to associate a security event with the perpetrator **using trustworthy digital evidence**.
- Since **digital evidence is distributed** in different locations and **collected in various electronic media** (information systems), incident response time is affected (Attoe, 2016)**.**

Digital Forensics is crucial to investigate and respond to security incidents, for example in the Cloud (Nurul, 2015)

Challenges for Database investigations (DB Forensics)

- During **data breaches**, **databases are the primary targets** and also **a great challenge for forensic investigations.**
- **DB Forensics** applies digital forensic approaches to gather suitable **digital evidence related to database activity** for presentation in a court of law (Fowler, 2007) – evidence may be **stored in different sources**.
- DB Forensics has received **very little research attention** (Hauger, 2015).
- **Generating trustworthy digital evidence** by ensuring **Chain of Custody** (evidence possession, non-repudiation, authenticity and **provenance**) during the investigation life cycle.

Traditional Digital Forensic techniques may not be suitable for DB Forensics (Khanuja, 2014).

DB Forensics: Reactive vs. Proactive

- **Reactive DB Forensics:**
  - Rely on the original database **reconstruction and recovery** (reactive controls) (Fasan, 2012).
  - **Data model and data dictionary are not considered** (Hauger, 2015).
  - **Traditional imaging and file carving** is used with ad-hoc practices for MSSQL (Fowler, 2008) and Oracle (Litchfield, 2007-2011).
  - **Its admissibility** (validity/trust) **is challenged** (Hauger, 2015).
  - **Time consuming** when dealing with short incident response periods.

Fire Fighting is not a good approach to incident response!

DB Forensics: Reactive vs. Proactive

- **Proactive DB Forensics:**
  - Rely on **analysing logged activity** (multiple evidence sources) during auditing stage when close monitoring of specific suspicious events is required (proactive controls).
  - Adopts an **e-Discovery approach** (Attoe, 2016) where digital evidence is recovered from different trusted distributed sources.
  - An **architecture with pre-defined functional requirements** is configured in a **forensically secure environment** (Digital Forensics Readiness).
  - Allow **shorter incident response time**.

Formalising Proactive DB Forensics is our challenge – can we be ready to respond efficiently?

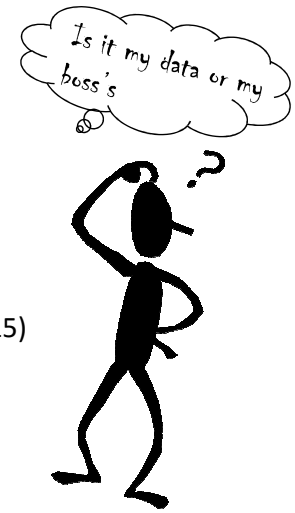
The BYOD Threat to Proactive DB Forensics

- **Bring-Your-Own-Device** (BYOD), or Dual-Use Devices is a **growing trend** encouraged by some organisations.
- **Increases employee productivity and accessibility** to corporate information assets (including databases) from anywhere at anytime.
- **Security concerns** regarding **monitoring and controlling** employee mobile device **access to information assets** (Sobers, 2015)

BYOD – Bring Your Own Disclosure??

The BYOD Threat to Proactive DB Forensics

- BYOD is a source of a vast amount of digital evidence *(Francis & Larson, 2015)*. When **corporate-owned relatively easy to handle**.
- **Insider activity has been overlooked** as only outsider attacks are seen as relevant (Pavlou et al, 2012).
- **Malicious/naive insider actions must be controlled** (Densham, 2015) to avoid corporate **information disclosure** (Pohlmann et al, 2015) and **contamination** (Downer & Bhattacharya, 2016).



In BYOD evidence ownership is problematic

A STRIDE-based BYOD Threat Model

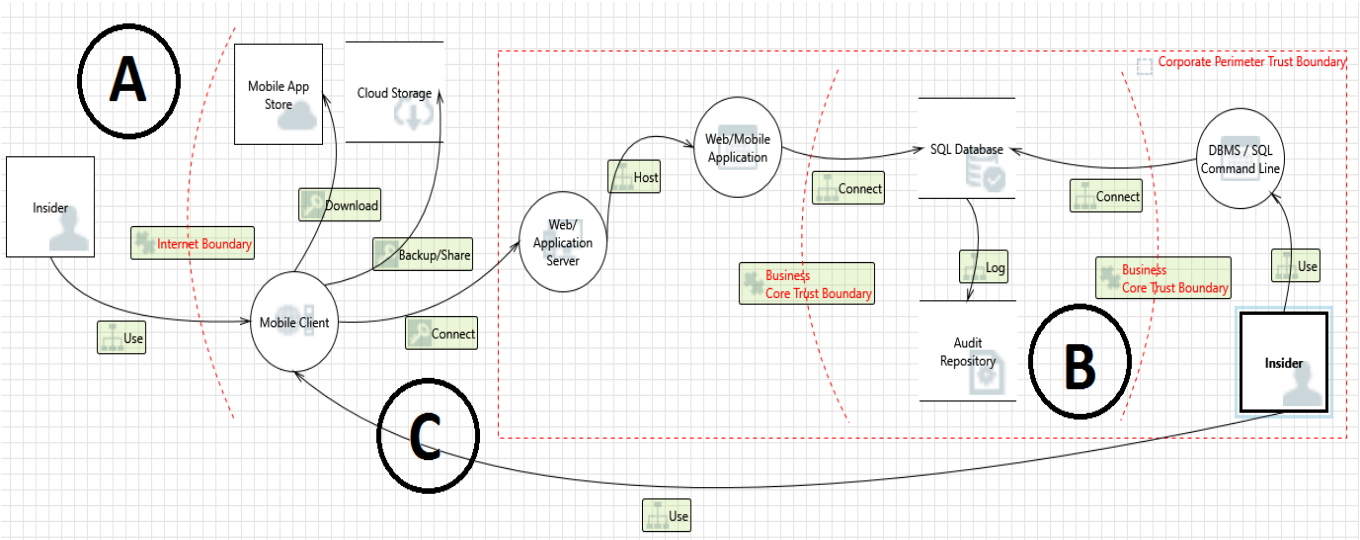
- BYOD exposes information assets (including databases) to **External and Internal Threat Contexts**

	External	Internal
Carrier	Cybercriminals (outsiders)	Trusted employees (insiders)
Threats	• Malware • Phishing • Social Engineering • Malicious Mobile Apps. • Insecure Wireless Networks • Fake Certificate Authorities • DoS	• Uncontrolled Devices • Device Misconfiguration • Unauthorised Information Sharing in Personal Clouds • Mixture of Personal and Corporate Information • Lost/Stolen/Unlinked Devices • Device Ownership

- Knowing the Threat Contexts, a **Threat Model is proposed using the STRIDE approach**

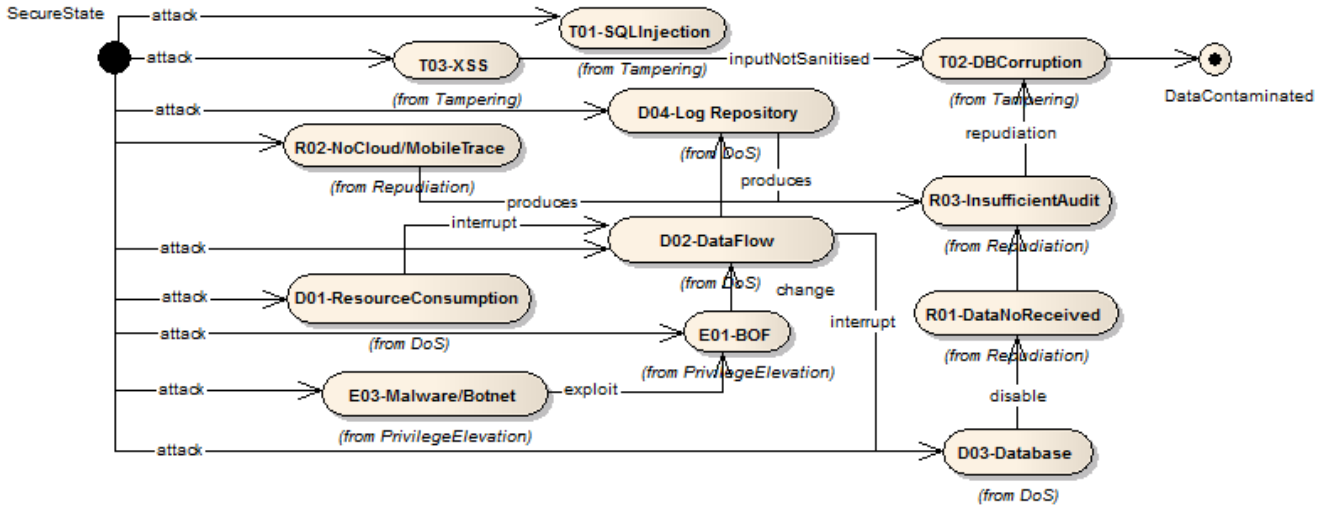
*STRIDE => **S**poofing, **T**ampering, **R**epudiation, **I**nformation Disclosure, **D**enial of Service, **E**levation of Privilege*

A STRIDE-based BYOD Threat Model



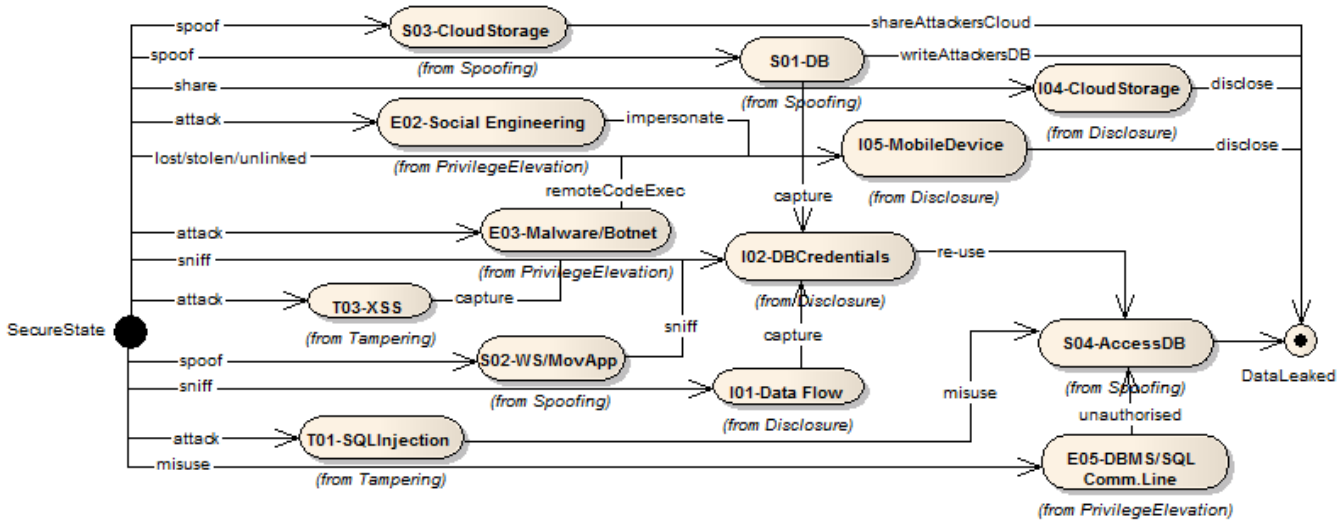
Trust Boundaries	Represents	Interacts With
A. Internet Trust Boundary (ITB)	Lower-Trust Insider Activity	Personal Cloud Mobile App Stores CPTB
B. Business Core Trust Boundary (BCTB)	Higher-Trust Insider Activity	Relational Database Audit Repository
C. Corporate Perimeter Trust Boundary (CPTB)	Internal/External Insider Interaction	ITB-located Mobile Client CPTB

A STRIDE-based BYOD Threat Model



BYOD Threats causing Information Contamination

A STRIDE-based BYOD Threat Model



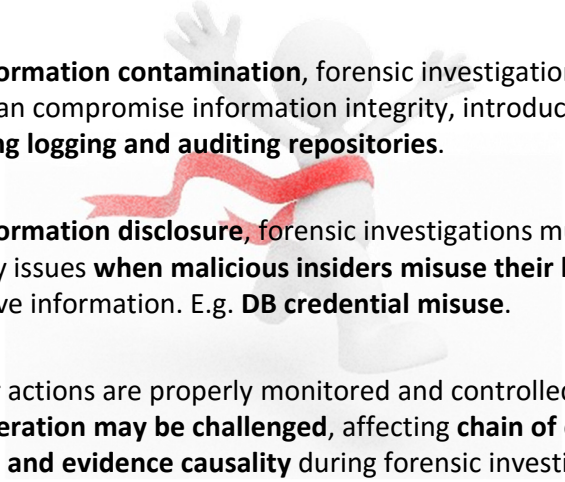
BYOD Threats causing Information Disclosure

Related Work

- Previous work on **mitigating BYOD Threats** has **not considered security and digital forensics issues**.
- **Mobile Device Management (MDM) solutions** (Sobers, 2015) provide mobile device access control, but **do not prevent/monitor information access and misuse** (i.e. disclosure and contamination).
- **STRIDE-based Threat Models** have already been applied for supporting **digital forensic readiness initiatives** (Lourida et al, 2013) without analysing threat interaction.
- **Our research provides a baseline** for understanding the environment in which proactive digital forensics initiatives may be deployed (Henry et al, 2013), considering internal and external threat interactions in the BYOD context.

Conclusions

- **Future work** towards protecting corporate information from unauthorised disclosure and contamination, **using proactive approaches.**
- Regarding **information contamination**, forensic investigations must look at insider actions that can compromise information integrity, introducing **repudiation issues when disabling logging and auditing repositories**.
- Regarding **information disclosure**, forensic investigations must be aware of information confidentiality issues **when malicious insiders misuse their high-privilege credentials** to access sensitive information. E.g. **DB credential misuse**.
- Unless insider actions are properly monitored and controlled, **proactive digital evidence generation may be challenged**, affecting **chain of custody provenance requirements and evidence causality** during forensic investigations.



WARWICK
THE UNIVERSITY OF WARWICK

Thank You!

References

- Andress, J. (2014). Chapter 1 - What is Information Security?, In The Basics of Information Security (Second Edition), Syngress, Boston, Pp. 1-22
- Attoe, R. (2016). Chapter 6 - Digital forensics in an eDiscovery world, In Digital Forensics, edited by John Sammons, Syngress, Boston, Pp. 85-98
- BS ISO/IEC (2016). 27000:2016 Information Technology-Security Techniques-Information Security Management Systems-Overview and Vocabulary, [Online], Available at: https://bsol.bsigroup.com/
- Densham, B. (2015). Three cyber-security strategies to mitigate the impact of a data breach. In Network Security, vol. 2015, 2015, pp. 5–8. [Online] Available at: http://bit.ly/1Wv9CQJ
- Downer, K. and Bhattacharya, M. (2016). BYOD security: A new business challenge. [Online] Available at: http://bit.ly/1O08xJY
- Escrivá, G. et al. (2013). Information Security. Macmillan Iberia S.A. Spain
- Fasan, O., Olivier, M. (2012). On Dimensions of Reconstruction in Database Forensics, In Seventh International Workshop on Digital Forensics & Incident Analysis.
- Francis, K. and Larson, M. (2015). Digital Forensics in the Mobile, BYOD, and Cloud Era. [Online]. Available at: http://bit.ly/1T9TxdY
- Fowler, K. (2007). SQL Server Database Forensics. [Online] Available at: http://ubm.io/1WuG9Il
- Fowler, K. (2008). SQL Server Forensic Analysis. Addison-Wesley Professional, Boston
- Hauger, W., Olivier, M. (2015). The state of Database Forensic research. In IEEE Information Security for South Africa (ISSA)
- Henry, P. et al (2013). The SANS Survey of Digital Forensics and Incident Response. [Online]. Available at: http://bit.ly/1SDdomv
- Khanuja, H. (2014). Role of metadata in forensic analysis of database attacks, In IEEE International Advance Computing Conference (IACC).
- Litchfield, D. (2007-2011). Papers on Oracle Forensics. Available at: http://www.davidlitchfield.com/security.htm
- Lourida, K., et al. (2013). Assessing database and network threats in traditional and cloud computing. [Online]. Available at: http://wrap.warwick.ac.uk/65197/
- Nurul, A., Kim-Kwang, R. (2015). Chapter 17 - Integrating digital forensic practices in cloud incident handling: A conceptual Cloud Incident Handling Model, In The Cloud Security Ecosystem, Syngress, Boston, Pp. 383-400
- Palmer, G., 2001. A Road Map for Digital Forensic Research. [Online] Available at: http://bit.ly/1bNYpaj
- Pavlou, K. et al. (2012). Achieving Database Information Accountability in the Cloud. [Online] Available at: http://bit.ly/1WuGzyh
- Sobers, A. (2015). BYOD and the Mobile Enterprise – Organisational challenges and solutions to adopt BYOD. [Online]. Available at: http://bit.ly/1Z8ZkG2
- Pohlmann, N. et al. (2015). Bring your own device for authentication (BYOD4A)–the Xign–System. In Information Security Solutions Europe (ISSE) 2015 Conference. Springer, 2015, pp. 240–250.
- Stallings, W. (2011). Network Security Essentials, 4th edition, New York, US; Prentice-Hall

Full name: Jamie Bayne

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Real-time Computer Vision for Autonomous Vehicles

Abstract: Over the past few years, the computer vision requirements of emerging autonomous vehicle have been the subject of intense academic and industrial research. Perhaps the most fundamental of these is stereo disparity, a technique for observing depth from two cameras inspired by the human visual system. Recent algorithms have achieved real-time speeds with good accuracy in certain environments, but autonomous vehicles, for which errors could be fatal, place very high demands on both. This presentation will give an overview of existing high-performance stereo disparity methods, and the challenges involved in making them run in real-time. We will then cover in less detail some related vision problems faced by autonomous vehicles.

Keywords: computer vision, stereo depth, autonomous vehicles, HPC

Abstract Word Count (Est): 110

Full name: Matthew Bradbury

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Poster

Title: DynamicSPR: A Hybrid Online Protocol for Source Location Privacy

Abstract: Wireless sensor networks (WSNs) will form the building blocks of many novel applications. Several of these applications will have to guarantee that the location of the occurrence of specific events is kept private from attackers, in what is called the source location privacy (SLP) problem. The fake source technique has been shown to be effective, however previous work has required that parameters are fixed at compile time. This poster presents an heuristic that solves this problem by determining good values online, allowing it to adapt to changing conditions. Simulations of the protocols were performed and results show that DynamicSPR provides a similar level of SLP as when parameters are optimised at compile-time. An approximate energy analysis and shows that DynamicSPR has little impact on the network lifetime and that the energy overhead it imposes is bounded.

Keywords: Source Location Privacy, Wireless Sensor Networks, Fake Sources, Energy Efficient

Abstract Word Count (Est): 136

DynamicSPR: A Hybrid Online Protocol for Source Location Privacy

Matthew Bradbury

WARWICK
THE UNIVERSITY OF WARWICK

Abstract

Wireless sensor networks (WSNs) will form the building blocks of many novel applications. Several of these applications will have to guarantee that the location of the occurrence of specific events is kept private from attackers, in what is called the *source location privacy* (SLP) problem. The fake source technique has been shown to be effective, however previous work has required that parameters are fixed at compile time. This poster presents an heuristic that solves this problem by determining good values online, allowing it to adapt to changing conditions. Simulations of the protocols were performed and results show that *DynamicSPR* provides a similar level of SLP as when parameters are optimised at compile-time. An approximate energy analysis and shows that *DynamicSPR* has little impact on the network lifetime and that the energy overhead it imposes is bounded.

Introduction

The source location privacy (SLP) problem is often framed as the panda-hunter game [5]. In this scenario a WSN has been deployed across a large area to monitor pandas, when a panda is detected a message is sent throughout the network. An attacker can use a directional antenna to find the direction from which the message came and thus trace back through the network to find the source. Our aim in providing SLP is to prevent the attacker from finding the source's location.

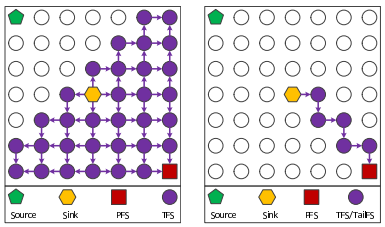
WILDSENSING [3] (which monitored badgers) and the WWF's Wildlife Crime Technology Report [1] (which tracks large animals in Namibia) are two real-world projects where providing SLP would be useful. SLP would ensure these schemes protect *context* information about the wildlife.

Previous Work

- A heuristic called *Static* was introduced [4].
- It created fake sources at a location far from the real source.
- They broadcast ⟨fake⟩ messages that are encrypted and padded to be indistinguishable to ⟨real⟩ messages.
- Parameters are fixed at compile-time.

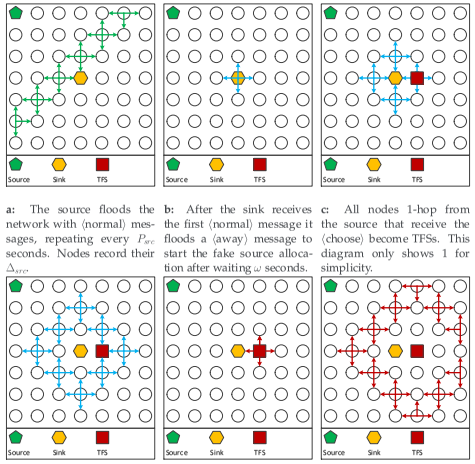
Objectives

- *Dynamic*: Determine parameters online, allowing the network to react to changing conditions [2].
- *DynamicSPR*: Reduce the energy cost of *Dynamic* by using a directed random walk away from the sink node to allocate fewer fake sources in better positions.



a: *Static* and *Dynamic* **b:** *DynamicSPR*

Figure 1: Best-case spread of fake sources for three different fake source heuristics.



a: The source floods the network with ⟨normal⟩ messages, repeating every P_{src} seconds. Nodes record their Δ_{src}.

b: After the sink receives the first ⟨normal⟩ message it floods a ⟨away⟩ message to start the fake source allocation after waiting ω seconds.

c: All nodes 1-hop from the source that receive the ⟨choose⟩ become TFSs. This diagram only shows 1 for simplicity.

d: The ⟨away⟩ flood continues, allowing nodes to record the Δ_{ss} and their Δ_{sink}.

e: The TFS starts broadcasting ⟨fake⟩ messages, for the duration of the fake source.

f: The ⟨fake⟩ message flood should lure the attacker away from the source.

Figure 2: The common actions for the *Static*, *Dynamic* and *DynamicSPR* heuristics.

Dynamic: Determining Parameters

Three major parameters were identified in *Static*: (i) the temporary fake source (TFS) duration (D_{TFS}), (ii) the temporary fake source period (P_{TFS}), and (iii) the permanent fake source (PFS) period (P_{PFS}). *Dynamic* focused on determining these parameters by reasoning about the times certain events happened.

- The D_{TFS} for node j follows the time between ⟨normal⟩ messages being received.

$$D_{TFS}(j) = \begin{cases} P_{src} - \omega & \text{if } \Delta_{sink}(j) \in \{1, \perp\} \\ P_{src} & \text{otherwise} \end{cases} \quad (1)$$

- The P_{TFS} was defined in terms of the number of ⟨fake⟩ messages ($\#_{\mathcal{F}}(j)$) that needed to be sent over the TFS duration.

$$P_{TFS}(j) = D_{TFS}(j) / \#_{\mathcal{F}}(j) \quad (2)$$

- Two strategies to determine the number of ⟨fake⟩ messages to send were investigated, Equation 3 aims to pull from the attacker's estimated position and Equation 4 aims to pull from the sink's location.

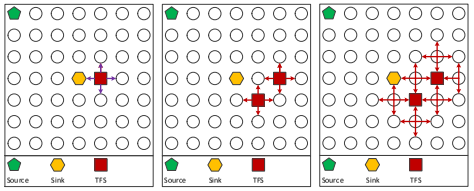
$$\#'_{\mathcal{F}}(j) = 2\Delta_{sink}(j) \quad (3)$$

$$\#'_{\mathcal{F}}(j) = \begin{cases} \Delta_{sink}(j) & \text{if } \perp \in \{\Delta_{src}(j), \Delta_{ss}\} \\ \Delta_{src}(j) - \Delta_{ss} & \text{otherwise} \end{cases} \quad (4)$$

$$\#_{\mathcal{F}}(j) = max_{\perp}(1, \#'_{\mathcal{F}}(j)) \quad (5)$$

- Finally, the PFS period was calculated so that it balanced the number of ⟨normal⟩ messages sent also taking into account the unreliability of delivering those messages ($\psi_{src}(\mathcal{F})$ is the ⟨fake⟩ message delivery ratio at the source).

$$P_{PFS}(j) = P_{src} \times \psi_{src}(\mathcal{F}) \quad (6)$$

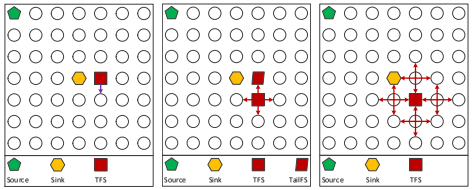


a: Choosing next fake source by broadcasting a ⟨choose⟩ message.

b: Nodes further from the source than the original TFS become fake sources.

c: The fake sources continue sending ⟨fake⟩ messages.

Figure 3: Spread of fake sources under *Static* or *Dynamic* after a TFS duration expires.



a: Choosing next fake source by unicasting a ⟨choose⟩ message, TFS becomes TailFS.

b: Original TFS becomes TailFS until it detects a further TFS.

c: TailFS detects further TFS and becomes normal again.

Figure 4: Spread of fake sources under *DynamicSPR* after a TFS duration expires.

DynamicSPR: Fake Source Allocation

The *Dynamic* heuristic focuses on improving the parameters involved with ⟨fake⟩ messages, however, it is also important to optimise the *location* of the fake sources. *Static* and *Dynamic* both allocate their fake sources in a flood from the sink node (Figure 1a) which leads to ⟨fake⟩ messages competing and colliding with one another. *DynamicSPR* uses a directed walk away from the source to allocate fewer and more targeted fake nodes (Figure 1b). This reduces collisions of messages and makes fake sources work cooperatively to pull the attacker in a single direction.

An issue is that the directed random walk is more susceptible to lost messages due to noise or collisions. To ensure reliability a new type of fake source is created called the Tail Fake Source (TailFS). A TFS becomes a TailFS after its duration expires and continues sending ⟨fake⟩ and ⟨choose⟩ messages until it receives a ⟨fake⟩ message from a further fake source (acting as an acknowledgement that a further fake source has been created).

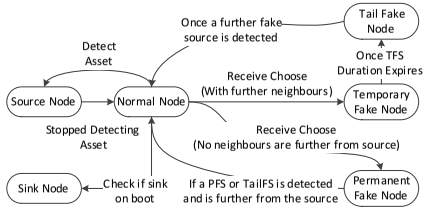
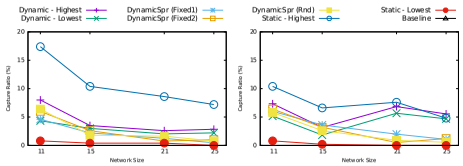


Figure 5: The conditions for nodes to transition from one type to another in *DynamicSPR*.

Results

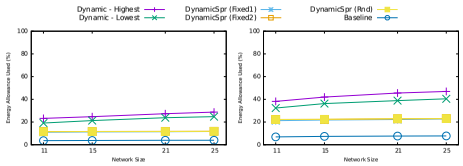
The *Static*, *Dynamic* and *DynamicSPR* heuristics were implemented in TinyOS and simulated in TOSSIM. Figure 6 shows that the capture ratio of *DynamicSPR* tends to be similar or slightly better than *Dynamic*. *Static* is included for comparison.



a: Source Period 1.0 second **b:** Source Period 0.5 seconds

Figure 6: The capture ratio of *Static*, *Dynamic* and *DynamicSPR*..

In Figure 7 the percentage of the energy budget used for each node by sending and receiving messages is shown. The total amount of energy available per day for a 2200 mAh battery at 3 V is 6.9 mAh [6]. After considering energy drain, CPU energy usage and with a 4.2% duty cycle, nodes have 1.9 mAh available. The total cost of sending (20 nAh) and receiving (8 nAh) messages can be calculated. The energy budget used is the total message cost divided by the total energy budget. Figure 7 shows that *DynamicSPR* has a small overhead compared to baseline protectionless flooding making it practical for deployments.



a: Source Period 1.0 second **b:** Source Period 0.5 seconds

Figure 7: The proportion of the energy budget used in *Dynamic* and *DynamicSPR*.

Conclusions

In this poster a new heuristic to provide SLP has been demonstrated. It has been shown to provide good levels of SLP at an affordable energy cost under realistic circumstances.

References

[1] `worldwildlife.org/projects/wildlife-crime-technology-project`. Accessed: 2016-06-03.

[2] M. Bradbury, M. Leeke, and A. Jhumka. A dynamic fake source algorithm for source location privacy in wireless sensor networks. In *14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom'15)*, pages 531–538, August 2015.

[3] Vladimir Dyo, Stephen A. Ellwood, David W. Macdonald, Andrew Markham, Niki Trigoni, Ricklef Wohlers, Cecilia Mascolo, Bence Pásztor, Salvatore Scellato, and Kharsim Yousef. Wildsensing: Design and deployment of a sustainable sensor network for wildlife monitoring. *ACM Trans. Sen. Netw.*, 8(4):29:1–29:33, September 2012.

[4] Arshad Jhumka, Matthew Bradbury, and Matthew Leeke. Fake source-based source location privacy in wireless sensor networks. *Concurrency and Computation: Practice and Experience*, 27(12):2999–3020, 2015.

[5] Pandurang Kamat, Yanyong. Zhang, W. Trappe, and C. Ozturk. Enhancing source-location privacy in sensor network routing. In *25th IEEE International Conference on Distributed Computing Systems, 2005. ICDCS 2005. Proceedings.*, pages 599–608, June 2005.

[6] Alan Mainwaring, David Culler, Joseph Polastre, Robert Szewczyk, and John Anderson. Wireless sensor networks for habitat monitoring. In *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, WSNA '02, pages 88–97, New York, NY, USA, 2002. ACM.

Full name: Ching-Chun Chang

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Reversible Data Hiding of Encrypted Signals

Abstract: The possibility of processing encrypted signals has attracted considerable research interests. In many cases, digital signals were encrypted before being processed. However, this act may prohibit many traditional signal processing approaches from being directly applied because some useful properties in natural signals no longer exist in encrypted signals. For instance, reversible data hiding exploits the spatial redundancy to embed data. Intuitively, however, it seems not possible to embed any data in an encrypted signal since the encrypted signal is typically very random and thus no redundancy can be exploited in such signal. Fortunately, in the light of successful lossless compression methods in encrypted domain, it can be expected that the task of reversible data hiding in encrypted domain is also feasible.

Keywords: Reversible Data Hiding, Encrypted Domain, Distributed Source Coding

Abstract Word Count (Est): 121

Full name: Sarunkorn Chotvijit

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Using Big Data to support improvements in Children's Services in Birmingham

Abstract: Birmingham is the most populous city outside of London, with over 1 million residents (2013 est). The city lies within the West Midlands area, which is the second largest metropolitan area in the United Kingdom. In terms of its demographics, Birmingham has a large group of children under the age of five, reflecting the high numbers of births in recent years; births are up 20% since 2001.

Birmingham City Council is the local government body responsible for the governance of the City of Birmingham. In recent years the Council has had to make large budgetary cost savings in the over of many \xa3100m. In the 2017/18 financial year the Council expects its workforce to be around 7,000 people, down from 20,000 in 2010.

Birmingham's Children's services have been rated inadequate since 2008. There were 20 investigations over child deaths from 2007 to November 2013. Consequently there are many social worker vacancies, with gaps being filled by agency workers. Responsibilities include risk assessment and decision making, victim identification and harm reduction, and co-ordinating safeguarding.

Birmingham's CareFirst system has been in operation for 10 years and contains a wealth of information on childcare cases. The total number of client records held in CareFirst, as of March 2015, is 567,119. There are a further 108,138 records which contain incomplete data.

To date there has been little analysis of this childcare data. The research involves extracting and analysing Birmingham City Council's social care data to tackle persistent failures in the service leading to poorer outcomes for children and young people and provides the technical support, tools and techniques in data analytics and data visualisation to illustrate the spatial-temporal properties of the data from the Council in order to identify the historic trends, current data patterns and predictions to support the future service provision and improved childcare in Birmingham.

Keywords: Children, Services, Safeguarding, Birmingham, City Council, City, Analysis, Big Data

Abstract Word Count (Est): 304

Full name: Henry Crosby

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: A Spatio-Temporal, Gaussian Process Regression, Real-Estate Price Predictor

Abstract: This paper introduces a novel four-stage methodology for real-estate valuation in England and Wales and then implements the output into a single real estate decision engine. The first stage of this research converts several discretized, non-uniform, spatiotemporal sales data-points into a single uniform spacetime cube, which is used to provide real-estate sales temporal singularity. Stage two utilises universal kriging, with a fifth-order polynomial kernel function, to identify the spatial dependancies related to real-estate prices. In stage three, a selection of property, economic and network features are collected and predicted. Finally, in stage four, a gaussian process method of regression is applied to the outputs of the first three stages. The resulting model is trained on a sample of 16,000 real estate transactions, and is validated against several alternative methods, including regression-kriging, an approach based on random forests, and an M5P-decision-tree model with 231,000 instances utilising r2 and RMSE. The trained model is integrated into a real estate decision engine for commercial use. Validation of the proposed model demonstrates a 96.6% goodness of fit value for residential real estate price prediction.

Keywords: Gassuian Process Regression, Real Estate

Abstract Word Count (Est): 180

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Data Driven Real Estate

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Introduction to Thesis

“Learning and Designing a Database of Key Residential Property, Land Use and Commercial Property Predictions Required to Standardize and Improve the Effectiveness of the UK's Real Estate Sector with Industrial Applications.”

Thesis Outline

C1: Introduction

C2: Background Research

C3: Proof of Concept

C4: Contribution 1 - Residential House Price Predictor

C5: Contribution 2 - Land Use Classifier

C6: Contribution 3 - Commercial Property Price Predictor

C7: Further Work

C8: Conclusions

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Thesis Outline

C1: Introduction

C2: Background Research

C3: Proof of Concept

C4: Contribution 1 - Residential House Price Predictor

C5: Contribution 2 - Land Use Classifier

C6: Contribution 3 - Commercial Property Price Predictor

C7: Further Work

C8: Conclusions

Thesis Outline

C4: Contribution 1 - Residential House Price Predictor

Sub-contirbution 1

Exploring New Data Sources to Improve UK Land Parcel Valuation

Sub-Contribution 2

A Spatially-Intensive Decision Tree Approach to Predicting the Traffic Flow of the entire UK Road Network.

Sub-Contribtuion 3

A Spatio-Temporal, Gaussian Process Regression, Real-Estate Price Predictor

Introduction

Who does it help?

Why is it so Important?

Varying motives

unprecedented leverage

market potential - 4.5% increase the UK in 2015

Why is it so Challenging?

Varying motives

data sparsity

Macro-environmental factors

Literature

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Literature and Progressions

In the UK:

1. Office for National Statistics (ONS),
2. Land registry
3. The Halifax and Nationwide Building Societies
4. The LSL Academy (formed by Pink and Aviva),

Most popular valuation models:

1. Median sales prices
2. Repeat sales
3. Hedonic regressions.

Academics Approaches:

1. RIPPER
2. C4.5
3. Adaboost
4. Case-Siller

Single House Prices

Combine 3 Techniques:

a. Space-Time Interpolation
b. Spatial Statistics
c. Machine Learning

5 Key Variables (SPENT):

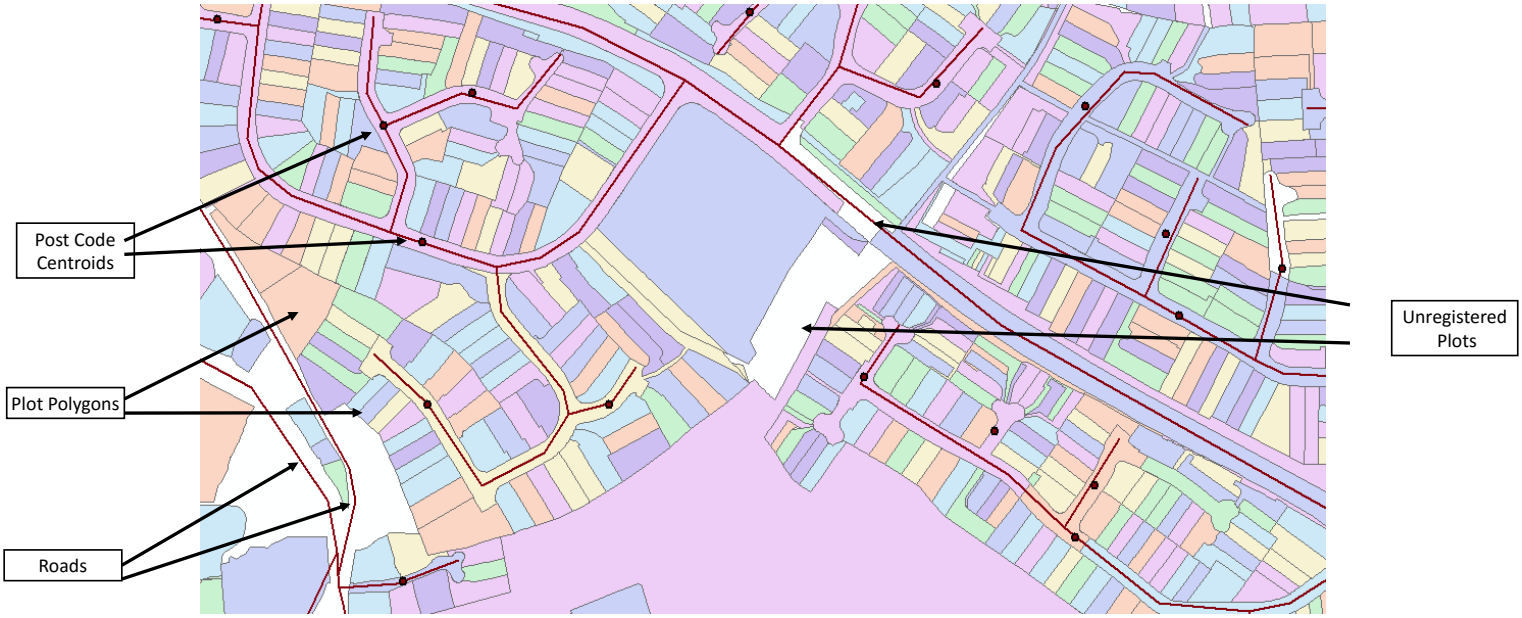
a. Space
b. Property
c. Economic
d. Network
e. Time

Method

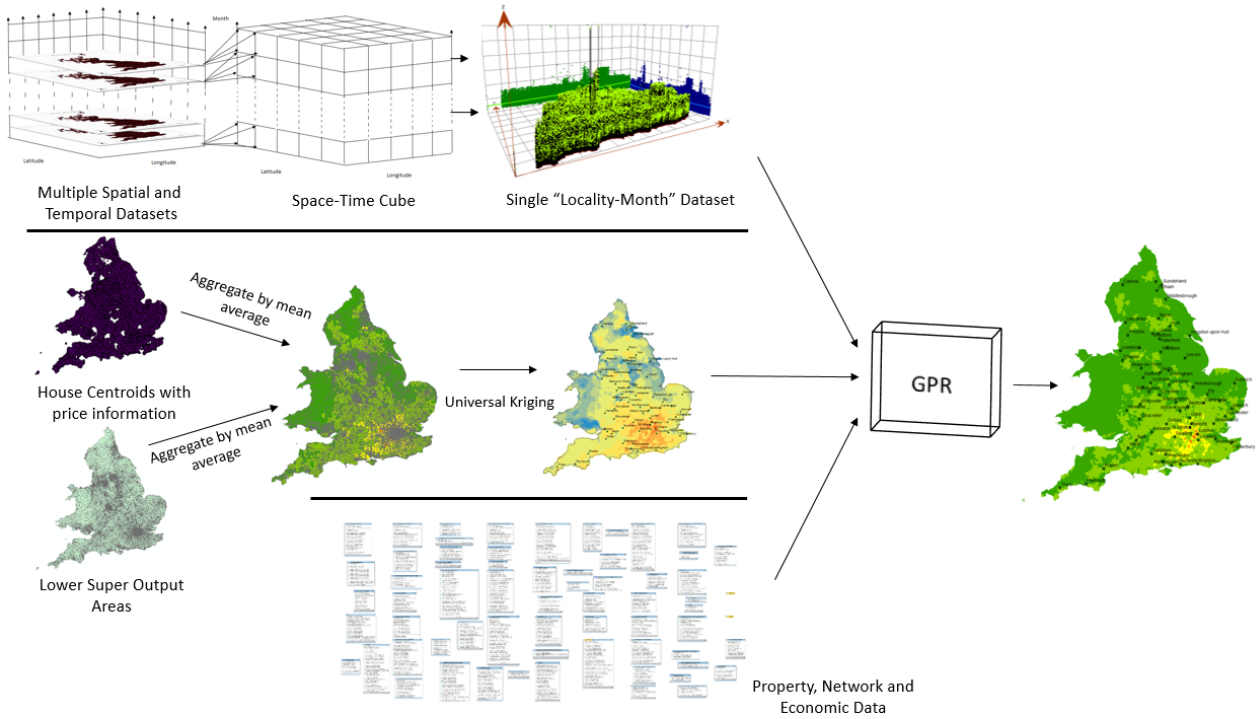
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Data Description



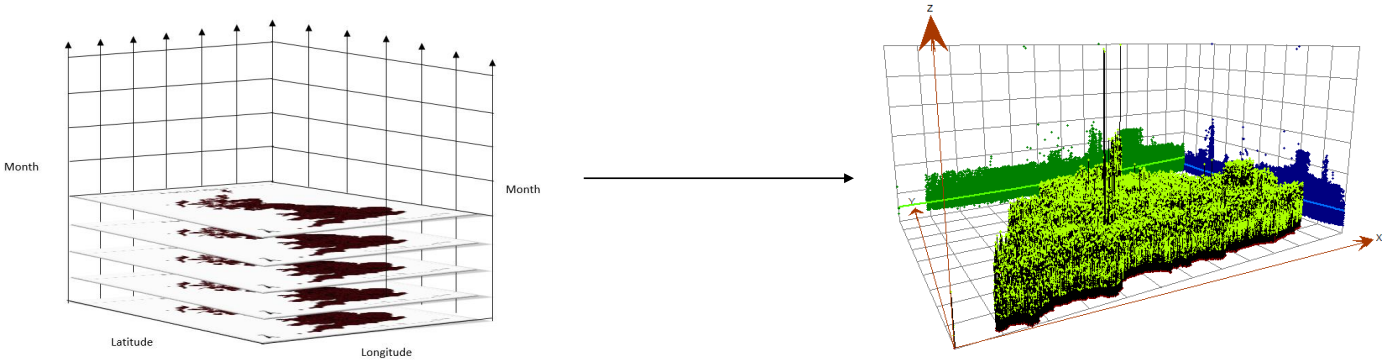
Method: Four Stages



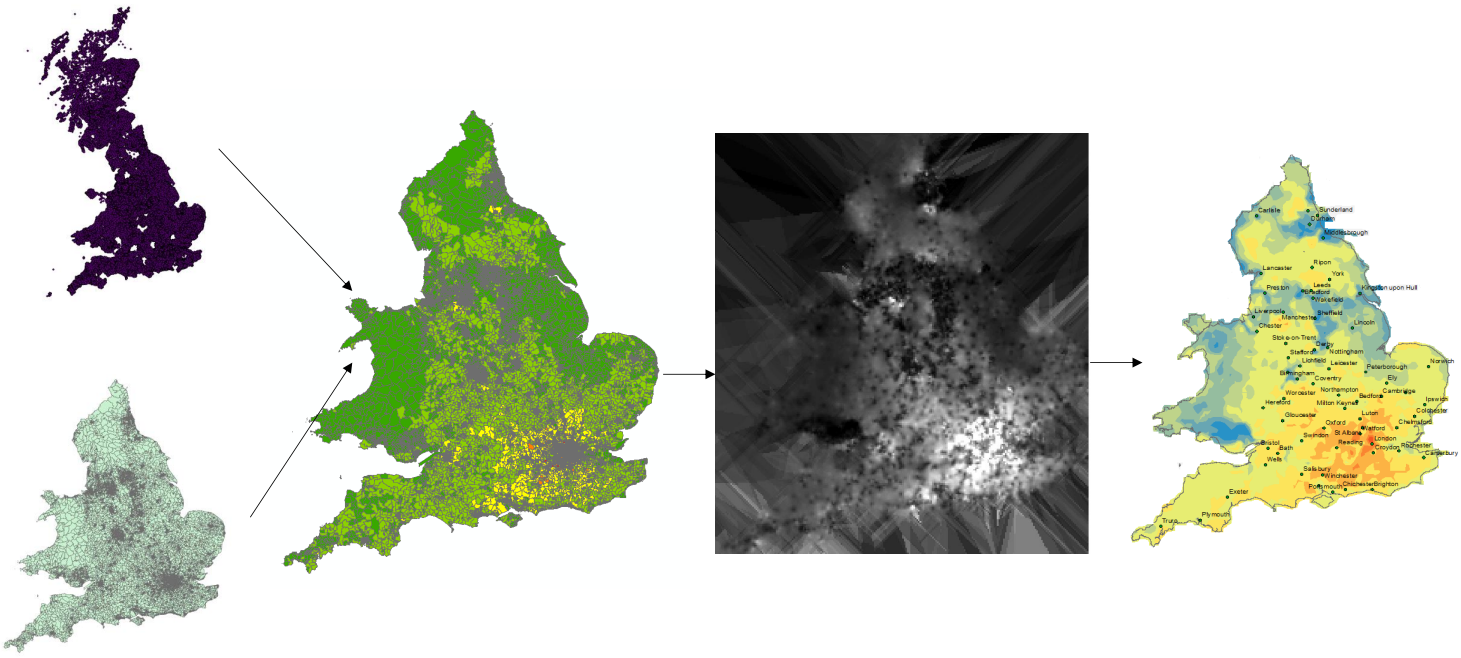
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Stage 1: Time



Stage 2: Space



WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

THE UNIVERSITY OF WARWICK

Stage 3: Economic, Property and Neighbourhood

Stage 4: Gaussian Process Regression

Gaussian Process:

$$f(x) \sim GP(m(x), k(x, x'))$$

Gaussian Process Regression:

$$f_* = f(x_t) + \epsilon_t$$

Dataset Input:

$$\mathbf{X}^s = [x_1^s, ..., x_N^s] \epsilon \, \mathbb{R}^{DXN}$$
$$\mathbf{Y}^s = [y_1^s, ..., y_N^s] \epsilon \, \mathbb{R}^{N}$$

Kernel Selection:

$$k(x, x') = \exp[\frac{-(x - x')^2}{\sigma_f^2}]$$

Hyper parameter Selection:

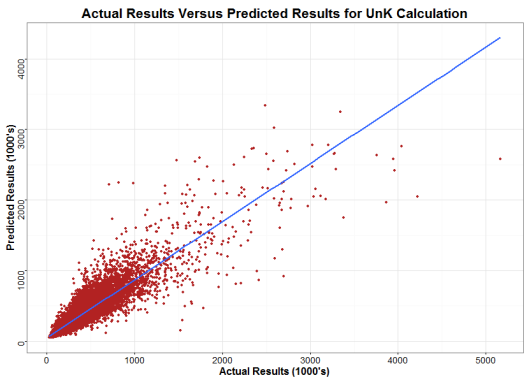
$$logp(y|X) = -\frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} y - \frac{1}{2} log|K + \sigma_n^2 I| - \frac{n}{2} log 2\pi.$$

Results

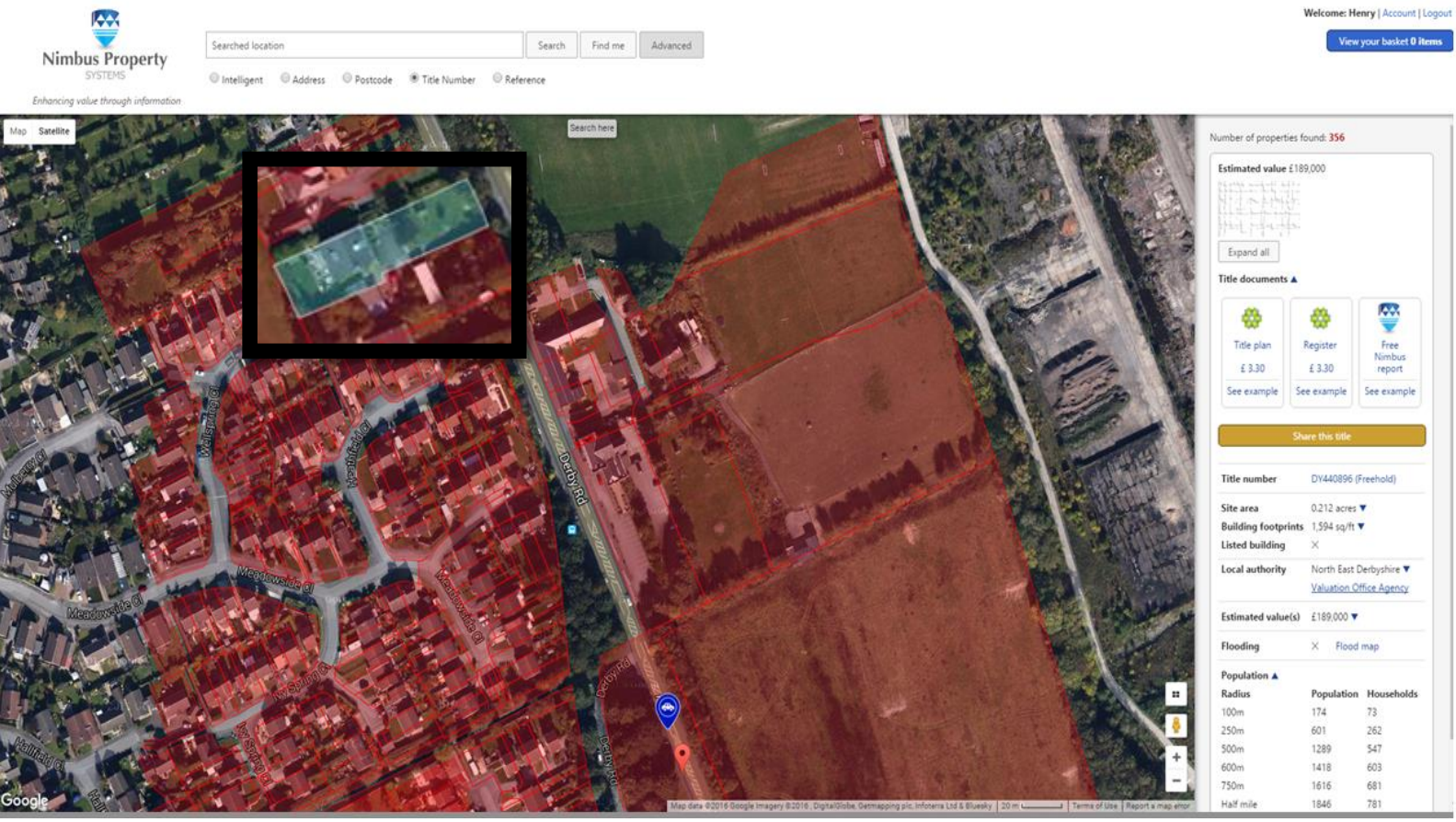
Results

	STAGE 1
Result	**Interpolation**
Sample Size	2.1m
R^2	0.710
$RMSE$	179325.8

Results: Actual Versus Predicted



Actual Versus Predicted with Space and Time



Conclusions

Conclusions

What I have shown you:

(i) A single uniform space-time cube
(ii) A spatially aware UnK calculation
(iii) A GPR on dataset of property, economic and network features.
(iv) A combined output for a new GPR.
(v) Unprecedented 96.6% accuracy
(vi) Implementation into an user-defined decision engine

Future Work:

(1) Extend to properties with no sales data
(2) Season Time Series Analysis
(3) Design of scalable GPR

Full name: Jacques Dark

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Small Summaries of Large Datasets

Abstract: Effectively utilising or analysing data often involves evaluating complicated functions of the data. However, for extremely large datasets, calculating even relatively simple statistics or properties can become slow and costly. Instead, we can build and maintain small summaries of the full data, which can be used to estimate the functions of interest.

Many summaries are known for functions over vectors (such as point queries and various norms) and there has also been a lot of work towards fundamental graph and matrix problems (such as identifying connected components, and performing linear regressions). However, real data problems continue to offer varied challenges, requiring new kinds of summaries.

Keywords: Streaming, algorithms, matrices, data

Abstract Word Count (Est): 105

Full name: Peter Davies

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Faster Deterministic Communication In Radio Networks

Abstract: We present our work on deterministic communication in radio networks. We consider an unknown radio network, in which all nodes have no prior knowledge about network topology, and provide new deterministic protocols for the fundamental communication primitives of broadcasting and wake-up. These algorithms improve upon the previous fastest known for the tasks, and can also be used as building blocks for more complex tasks.

This talk is based on joint work with Artur Czumaj.

Keywords: Radio Networks, Broadcasting, Wake-up, Distributed Computing

Abstract Word Count (Est): 74

Full name: James Dickson

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Analysing Parallel I/O in Scientific Codes through Proxy Applications

Abstract: Data movement has become a increasingly demanding aspect of large scale simulations, through mismatched advances in computational and data storage capabilities. To combat I/O performance shortfalls, a complex software stack has evolved for performing application I/O that makes use of parallel operations and increasing levels of abstraction. It is the case however that obtaining the best performance for I/O at the application level requires tuning operations at all levels in the stack to best leverage complex parallel file systems. Investigating the performance of I/O activity with production applications can be extremely time consuming and lacks the flexibility required to explore newly emerging paradigms and software. To facilitate this investigation, a workflow has been developed to profile the I/O patterns of real HPC applications and deploy a proxy application in an attempt to replicate common data storage patterns for the investigation of different elements in the parallel I/O stack.

Keywords: HPC, Parallel I/O, Profiling, Proxy Applications

Abstract Word Count (Est): 148

Analysing Parallel I/O in Scientific Codes through Proxy Applications

James Dickson

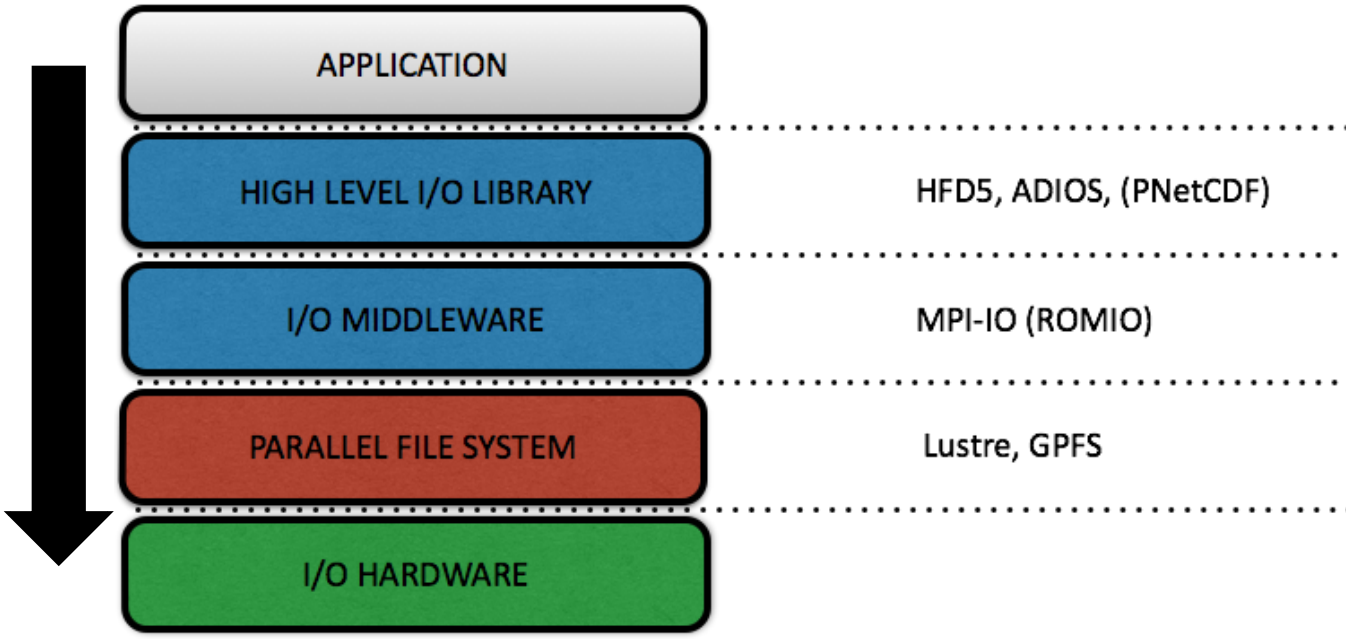
Outline

- Motivation
- Parallel I/O Stack
- Benchmarks/ Proxy Applications
- I/O Profiling and Patterns
- Future work

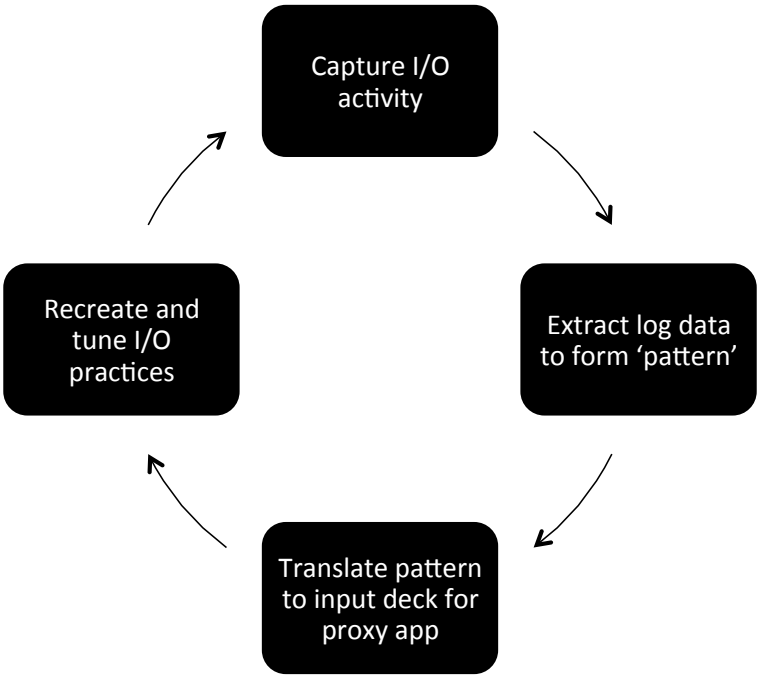
Motivation

- Understanding I/O performance is not straightforward
  - Important factors change depending on where in the I/O stack we are looking
  - File system tuning options are important but not always fully explored
  - Changes to file systems could potentially influence the strategy we want to adopt e.g. burst buffers

I/O Stack



Lifecycle



Benchmarks

- Often outdated or no longer maintained
- Specific to certain applications
  - Flash I/O
  - S3D I/O
  - Chombo I/O
- File system focused
  - IOR
  - IOZone
  - fs_test

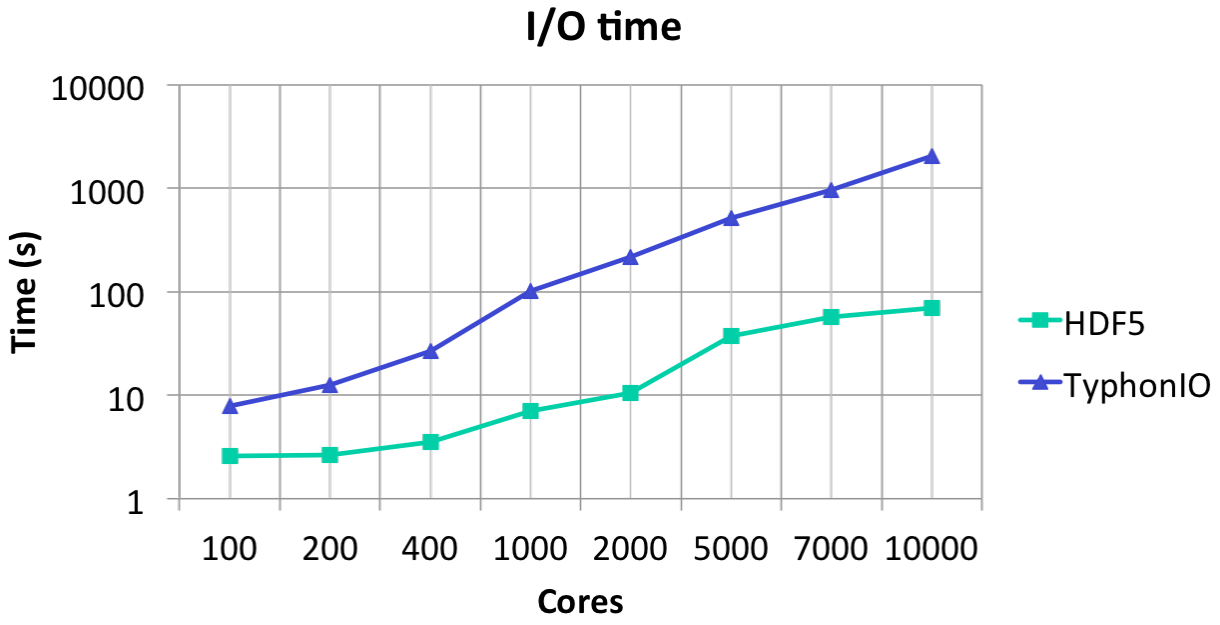
Profiling

- Capture the ‘pattern’ of I/O behavior in the most succinct and lightweight fashion – it’s possible to log every H5Dwrite(…) in its entirety but there will be a performance hit
- Darshan is widely adopted for this purpose - it can be deployed system-wide and enabled by default

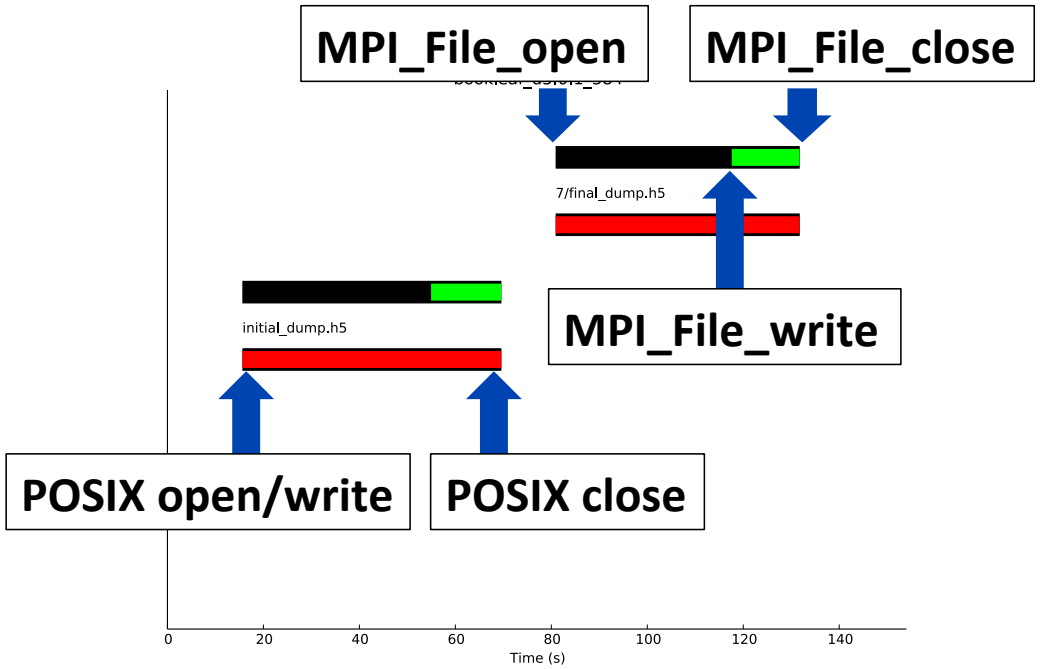
Profiling Example

<modules>	MPI-IO
<rank>	-1
<record id>	-11698698072705351164 0
<counter>	MPIIO_BYTES_WRITTEN
<value>	3063555700
<file name>	32_typhonio_000.h5
<mount pt>	/fs2
<fs type>	lustre

Profiling I/O Libraries

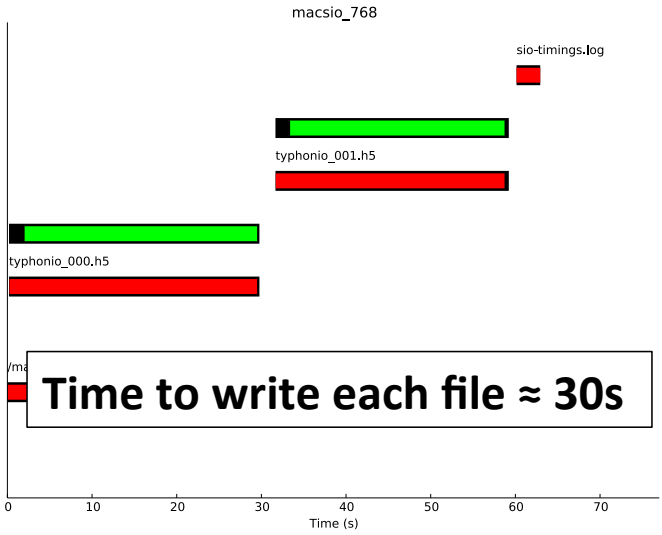


Visualising I/O Behaviour



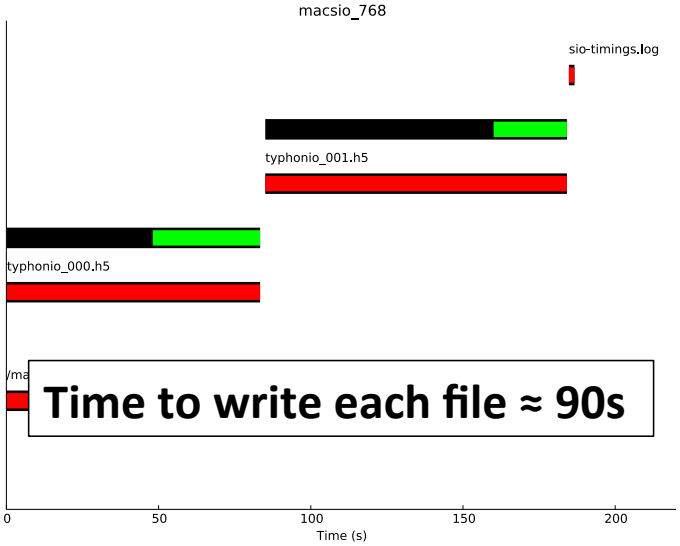
Performance Issues

Collective



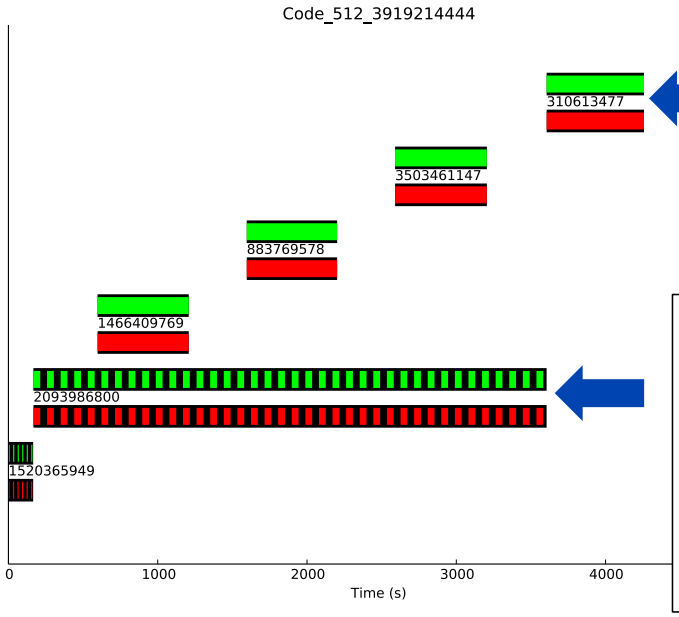
Time to write each file ≈ 30s

Independent



Time to write each file ≈ 90s

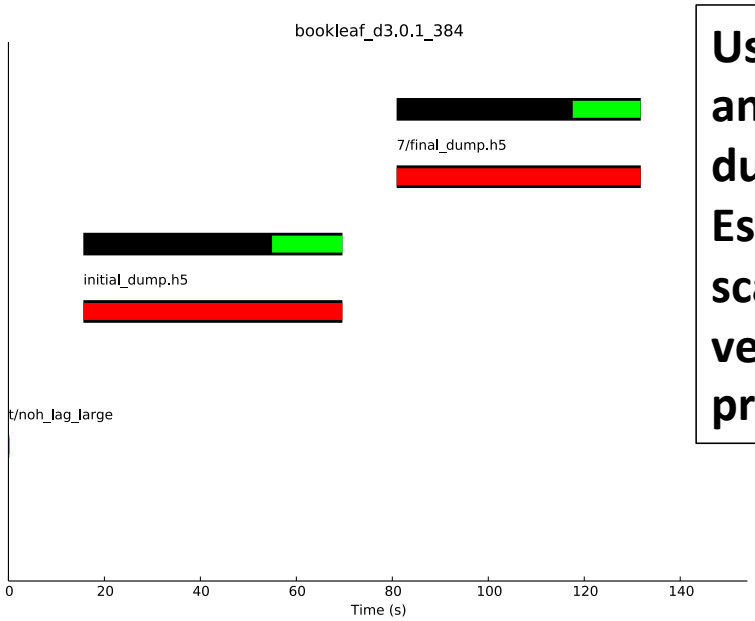
Profiling a Production Code



Checkpoint dump ≈ 147 GiB written

Vis dump (dashed line indicates r+w) ≈ 650 MiB written 11 MiB read

Profiling a MiniApp



Uses an initial and final output dump – Essentially a scaled down version of the production code

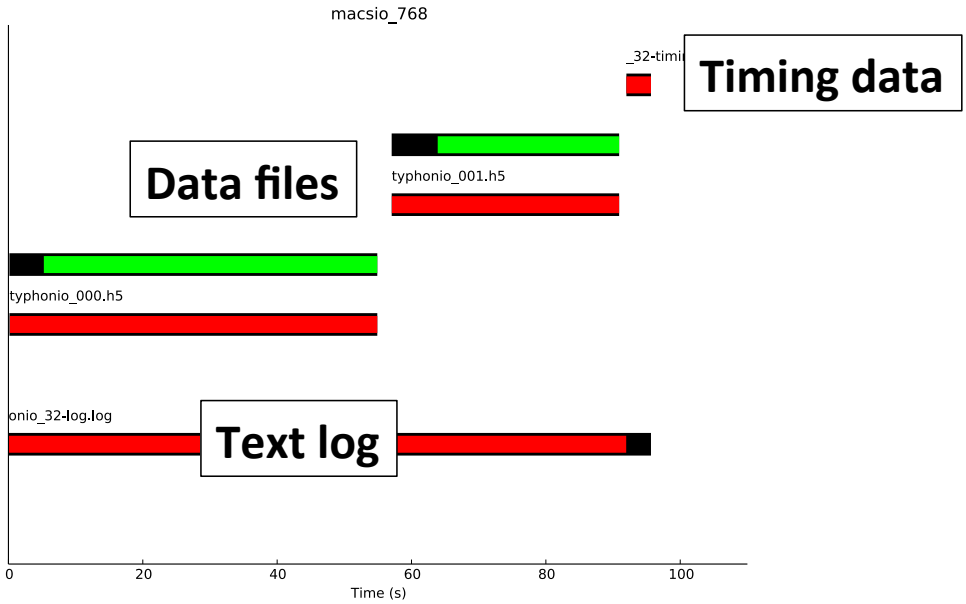
MACSio

- Developed by Mark Miller at LLNL
- Contains plugins for HDF5, SILO and Exodus
- Identical purpose to MINIO with more accessible design
- A TyphonIO plugin has been integrated to allow representation of AWE applications

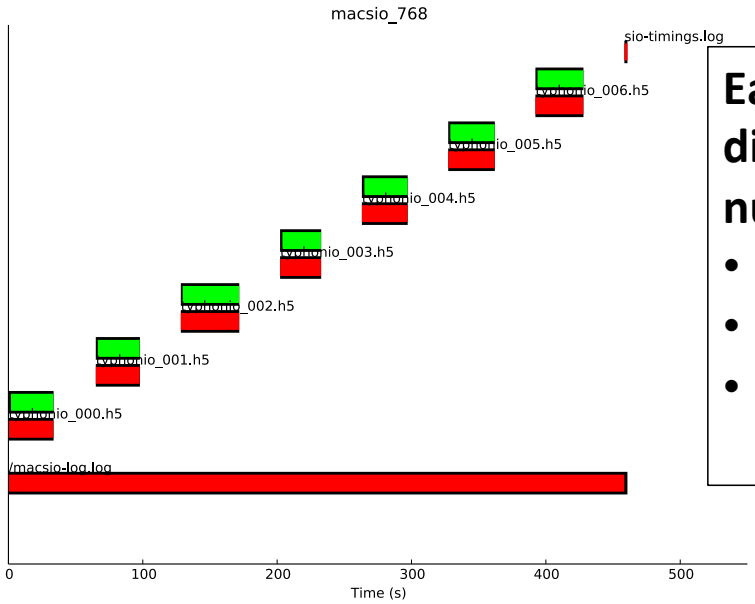
TyphonIO in MACSio

- A few issues faced:
  - Structured mesh must be created and chunk information set collectively – but mesh coordinates are written from rank 0
  - Mapping of representation required between MACSio data structures and TyphonIO

MACSio



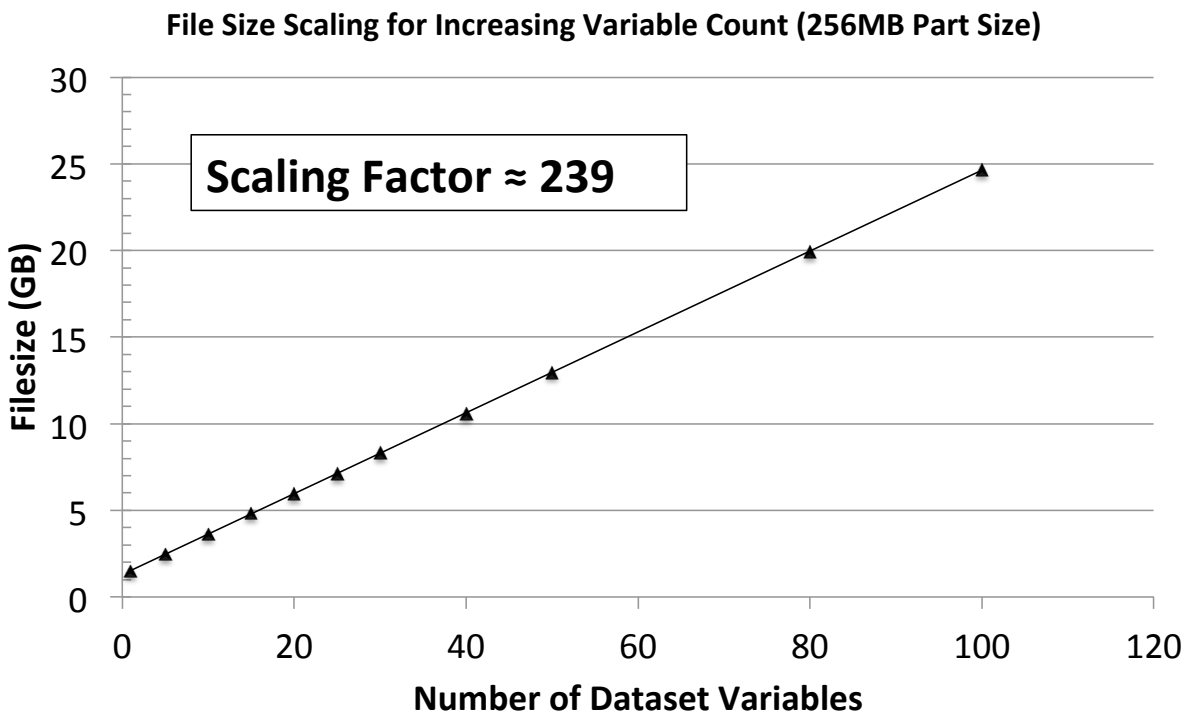
MACSio again



Easy to set different number of:

- **Output Files**
- **Variables**
- **Time buffer between I/O**

I/O Pattern Modeling



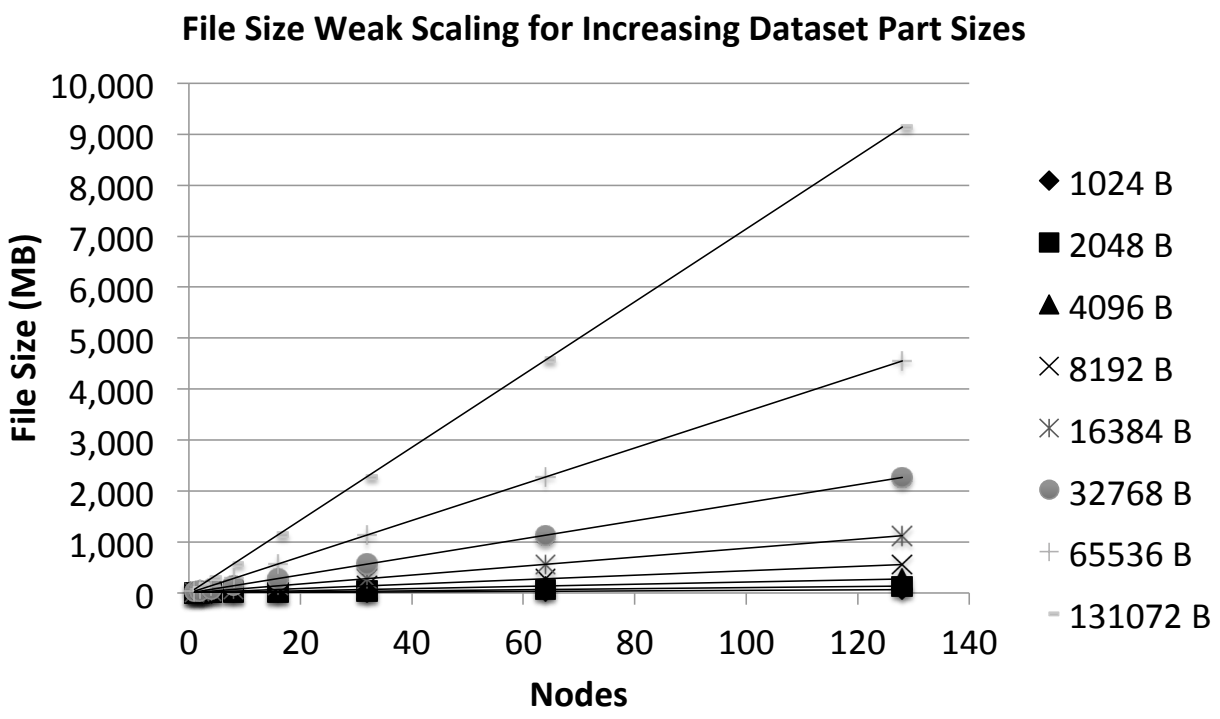
I/O Pattern Modeling

File size = Base * Part Size +
(Scaling Factor * Variable Multiplier)

File size = 5.99 * 256M + (239 * V)

Variable Count	Actual File Size	Calculated File Size
1	1534.271671	1534.271671
10	3708.030773	3690.019524
100	20461.3582	20247.49805

I/O Pattern Modeling



I/O Pattern Modeling

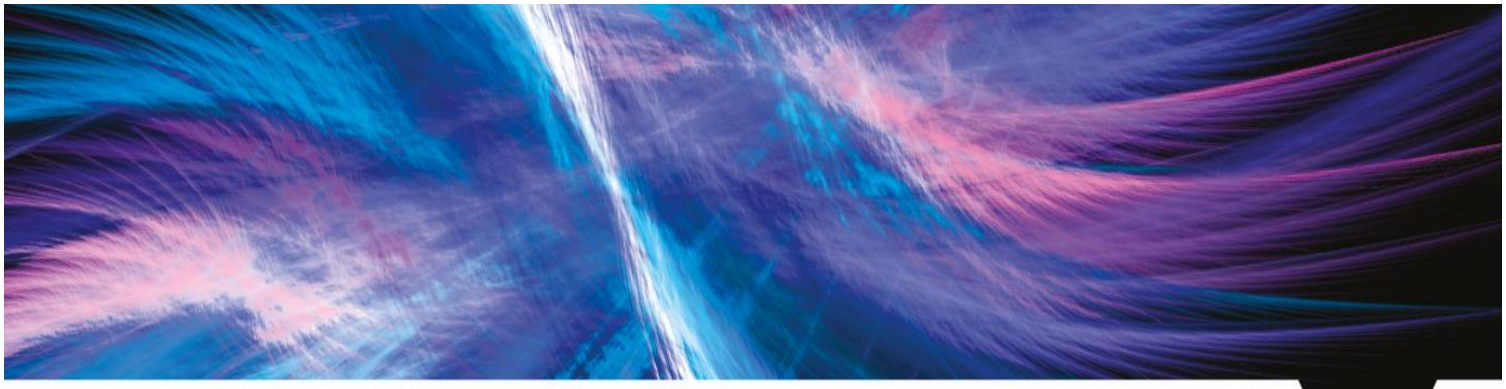
Part Size (Bytes)	Single Node File Size (Mbytes)	Scaling Factor (1 to 128 Nodes)	% Difference
1024	0.547428131	0.512237549	6.43%
2048	1.091617584	1.056427002	3.22%
4096	2.17999649	2.144805908	1.61%
8192	4.393375397	4.358184814	0.80%
16384	8.820133209	8.784942627	0.40%
32768	17.74689102	17.71170044	0.20%
65536	35.60040665	35.56521606	0.10%
131072	71.45392227	71.41873169	0.05%

Future Work

- Complete Proxy App model
- Automate the process of profiling and translation
- Investigate how changing the pattern of I/O in an application will impact performance as a whole

Future Work

- Introduce a MACSio plugin for ADIOS and compare it's performance for workload patterns to raw HDF5 and TyphonIO
- Further the investigation of how changing file system and file system tuning parameters makes or breaks I/O performance



WARWICK
THE UNIVERSITY OF WARWICK

Thank you

Questions?

Full name: Roberto Leyva Fernandez

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Fast Binary-based Video Descriptors

Abstract: Action recognition is one of the top challenges in computer vision. In this talk, we present two binary-based video descriptor with outstanding characteristics in terms of recognition rate, computational times and memory requirements. The descriptors are called Binary Wavelet Differences (BWD) and Binary Dense Trajectories (BDT). Our proposed descriptors are based on the local binary patterns and produce binary vectors with a very low dimensionality. Specifically, we propose to analyze the spatio-temporal support regions of a video sequence to generate binary strings via wavelets patterns. We also propose to encode the motion information obtained from optical flow into a compact binary representation. Our evaluations on the KTH and UCF50 datasets demonstrate that our proposed descriptors achieve very competitive recognition accuracy. Moreover, they are able to attain shorter computational times and smaller memory requirements. Specifically, our proposed descriptors can be calculated up to 20X faster than orientation-based descriptors and require up to 225X less memory.

Keywords: Video Analysis, Action Recognition, Binary Data

Abstract Word Count (Est): 154

Full name: Adam Gelencser

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Poster

Title: Concept change prediction in machine learning

Abstract: The volume of data generated and captured within automotive industry during all life cycle stages is exponentially

growing. There are multiple initiatives how the available data can be utilised to increase reliability, durability,

efficiency and driver comfort. This is the case within Jaguar Land Rover as well, where data mining and machine

learning tools are more widely used than ever before. This research considers concept drift detection in the context

of vehicle data. Concept drift occurs within dynamically changing environments where the underlying distribution

of inputs or desired outputs can change significantly and to ensure accurate predictions, the models learnt must be

updated accordingly to the rate of concept drift.

Keywords: machine learning, concept drift, concept prediction, data stream

Abstract Word Count (Est): 110

Concept change prediction in machine learning

Adam Gelencser
University of Warwick
a.gelencser@warwick.ac.uk

Abstract

The volume of data generated and captured within automotive industry during all life cycle stages is exponentially growing. There are multiple initiatives how the available data can be utilised to increase reliability, durability, efficiency and driver comfort. This is the case within Jaguar Land Rover as well, where data mining and machine learning tools are more widely used than ever before. This research considers concept drift detection in the context of vehicle data. Concept drift occurs within dynamically changing environments where the underlying distribution of inputs or desired outputs can change significantly and to ensure accurate predictions, the models learnt must be updated accordingly to the rate of concept drift.

Introduction

Normally, machine learning models are trained on historical data, but data from the past may not be relevant in the future after a while and as a consequence the models will become less and less accurate over time. There are existing solution to deal with the above problem (e.g. periodic retraining, online learning etc.), but current challenge is how to keep the existing models as much up-to-date as possible and at the same time minimise the processing and the amount of data to be stored for the periodic retraining [1].

In much of the literature, e.g. in [3], the applications where concept change occurs and require some form of adaptation, are categorised as follows:

1. Monitoring and control, which includes detection of abnormal behaviours or adversary activities on the web, computer networks, telecommunications, financial transactions and time series forecasting tasks.
2. Personal assistance and information applications, which include recommender systems, categorization and organization of textual information, customer profiling for marketing, spam filtering etc.
3. Decision support, which includes predictive analytics e.g. evaluation of credit worthiness, demand prediction, like electricity, food in supermarkets, crime in a region etc.
4. In artificial intelligence applications concept change is usually called dynamic environment as they usually include a wide spectrum of moving and stationary systems, which interact with changing environment, for instance robots, mobile vehicles, smart household appliances, etc.

As can be seen from the above, concept change is recognised in a wide variety of domains and affects many machine learning applications.

Main Objectives

This research aims to examine concept change and its effects in automotive industry and offer a practical approach to deal with change in data:

1. to investigate the extent and likely consequences of concept change in vehicle induced data;
2. to ascertain methodologies currently applied in other domains to manage changes in concepts in machine learning;
3. to determine concept change handling techniques which are appropriate within the limitations of vehicle environments as well.

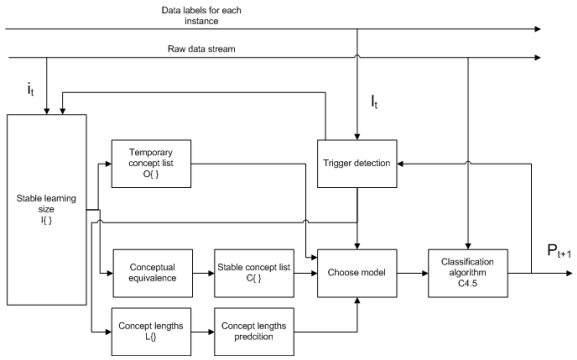


Figure 1: The architecture of the extended RePro algorithm

Method

A special and particularly interesting case in concept changes is the one of recurring concepts. This may happen for instance when the environment is subject to seasonal variations which repeats over time. This problem introduces a new challenge in that apparently obsolete data or learned concepts may well be relevant again in the future. It would thus be profitable to exploit this past knowledge as soon as it is appropriate rather than learn a new concept from the incoming data stream. This requires, however, that previous concepts are stored and that their relevance to the current situation be quickly recognized. In recent years, several proposals have been put forward to meet this challenge, some within the ensemble-based approach. Most algorithms work in a reactive manner, meaning they monitor the data or the accuracy of the model for potential concept change signals and when one present they either start learning a new model or choose the best available from the ensemble.

The first approach which stepped beyond this point, was an algorithm called RePro which collects previously seen concepts and also, how they follow each other. This is the basis for establishing the transition probabilities for each concept. The assumption in this is that concepts follow each other in a predefined order in which case the RePro algorithm is capable of predicting the next likely concept based on historical changes.

The work carried out so far builds on the above by adding further prediction capability. It not only predicts which concepts will come next, but also when it will change. This is done by recording concept lengths in the form of data samples for each concept and using linear regression predicting the next length, at which point the concept is automatically swapped to the predicted one.

Results

The following results show that this approach yields to a higher accuracy and during concept changes the accuracy measure is not dropping to much lower levels. The overall accuracy measure with the length predication is 96.25% while the original RePro algorithm has an accuracy of 92.17%.

In the below graph it is visible that in case of accurate length prediction the accuracy of the model remains high.

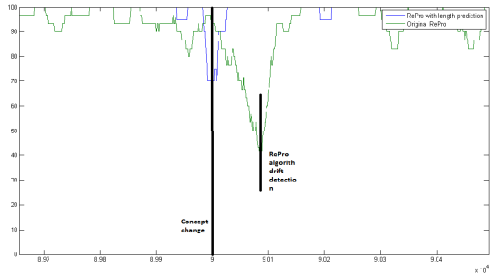


Figure 2: Accuracy of the original RePro and extended with length prediction algorithm

On the accuracy graph below the prediction of the start time of the new concept is too early and hence the accuracy falls for a short period but the algorithm switches back to the original model until the new model becomes more accurate.

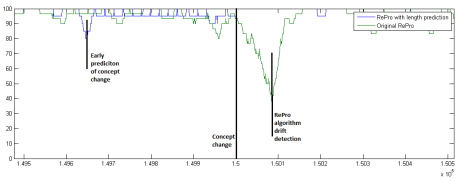


Figure 3: Accuracy of the original RePro and extended with length prediction algorithm

Conclusions

The work improves on the existing strategies for prediction of concept changes under data streams with recurring concept drifts. Algorithms are developed and experimentally validated, and compared to the selected existing adaptive learning algorithms and passive training strategies (no adaptivity).

The work so far focused on using the length of concepts as a prediction factor for concept changes. As the above results illustrate this approach can a be a very good predictor of concept changes and thus improving the accuracy of models. However due to its strong requirement on static length for concepts it is only providing its full value in certain cases.

Forthcoming Research

Future work will look at removing the strong assumptions on the data and the concept length and also how previously not seen concepts can be formed based on historical concept changes.

References

[1] Pedro Domingos Geoff Hulten, Laurie Spencer. Mining Time-Changing Data Streams. *ACM KDD*, 2001.

[2] Mykola Pechenizkiy I. Zliobaite, Jorn Bakker. OMFP: An Approach for Online Mass Flow Prediction in CFB Boilers. *Discovery Science*, 5808, 2009.

[3] Albert Bifet Mykola Pechenizkiy Abdelhamid Bouchachia Joao Gama, Indre Zliobaite. A survey on concept drift adaptation. *ACM Computing surveys*, 1(1), 2013.

[4] L. I. Kuncheva. Classifier ensembles for detecting concept change in streaming data: Overview and perspectives. *2nd Workshop SUEMA*, 2008.

[5] R. S. Michalski M. A. Maloof. Incremental rule learning with partial instance memory. *Artificial Intelligence*, 154, 2004.

[6] Rudolf Kruse Mirko Bottcher, Martin Spott. Predicting Future decision trees from evolving data. 2008.

[7] Kyosuke Nishida. Learning and Detecting Concept Drift. 2008.

[8] I. Renz R. Klinkenberg. Adaptive information filtering: Learning in the presence of concept drifts. *ICML*, 1998.

[9] Xingquan Zhu Ying Yang, Xindong Wu. Mining in anticipation for concept change: Proactive-reactive prediction in data streams. 2005.

[10] I. Zliobaite. Learning under concept drift: an overview. Technical report, 2009.

Full name: Zhuoer Gu

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: An Efficient Method for CPU Host Load Motif Discovery

Abstract: Mining repeated patterns (often called motifs) in CPU utilization of computers (also called CPU Host Load) is of fundamental importance. Many recently emerging applications running on high performance computing systems rely on motif discovery for various purposes, including efficient task scheduling, energy saving, etc.. In this paper, we propose an efficient motif discovery framework for CPU host load. The framework is elaborately designed to take into account the important properties in host load data. The framework benefits from its ability of on-line discovery and the adaptivity to work with massive data. The experiments are conducted in this paper and the experimental results show that the proposed method is effective and efficient.

Keywords: Pattern Discovery , Motif, Host Load, Time Series

Abstract Word Count (Est): 111

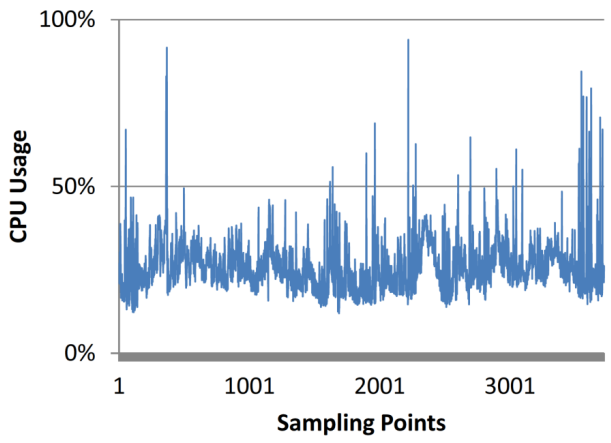
An Efficient Method for CPU Host Load Motif Discovery

Zhuoer Gu

THE UNIVERSITY OF WARWICK

What is CPU Host Load?

- CPU host load is the percentage of CPU time consumed in a series of time slots.
- CPU host load, often sampled and collected with time, is time series.



What is a Motif?

- A *match* in a time series data, is two subsequences of the data whose distance is lower than a given threshold.
- A motif in a time series dataset is a set of subsequence. Each subsequence in the set matches to every other subsequence.

Why Mining Motifs?

It is very valuable to discover motifs in CPU host load in a Cloud system. Many applications that utilize motifs in time series data can also be applied to work with CPU host load data, including

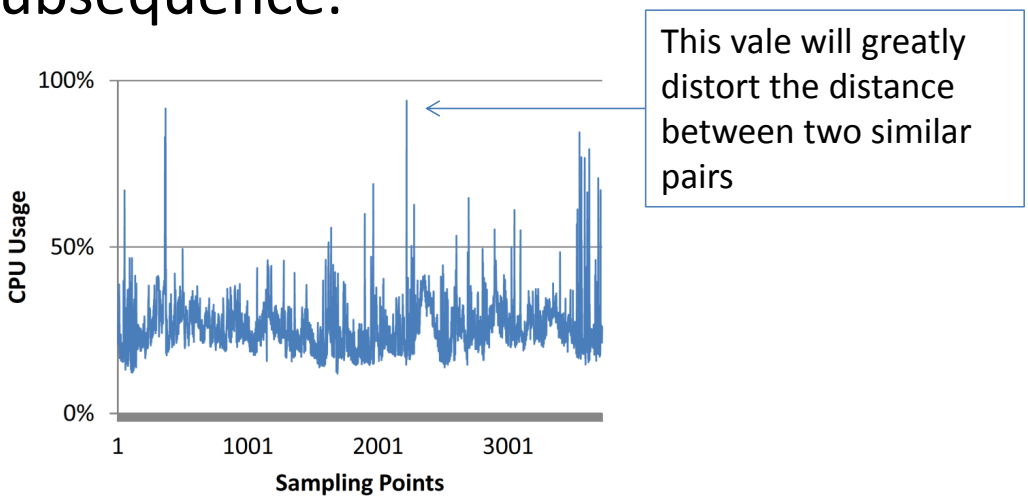
- The algorithms for mining association rules in time series data based on pattern discovery
- Classification algorithms that are based on building typical prototypes of each class
- Anomaly detection
- Finding periodic patterns

Why it is Difficult?

- To find motifs in a single series of data, we need to consider every possible subsequence in the data. For all the susequences, we need to calculate every pair of them to find match.
- The time complexity of this could be $O(n^5)$ to $O(n^6)$!

Why it is Difficult?

- The fluctuation nature of CPU host load makes it a problem to locate a pair of match subsequence.



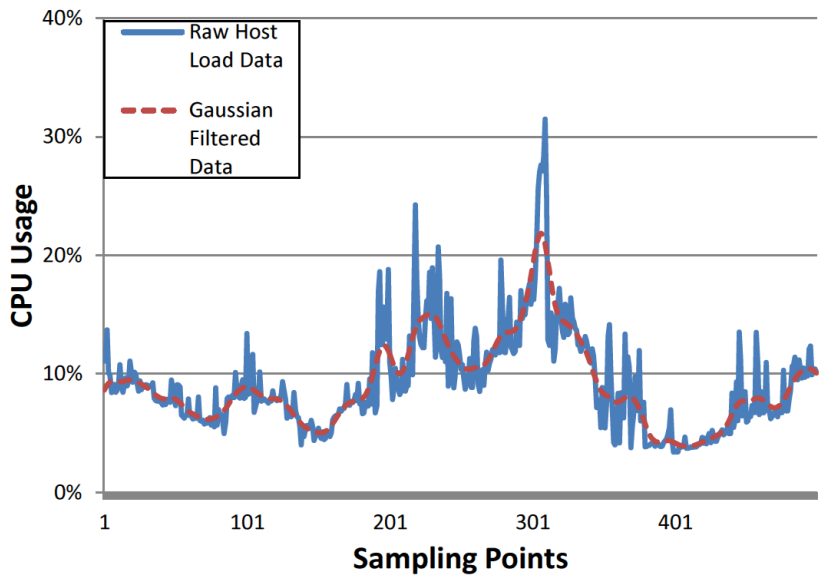
Our Tasks

- Reduce noise
- Reduce time complexity

Noise Reduction

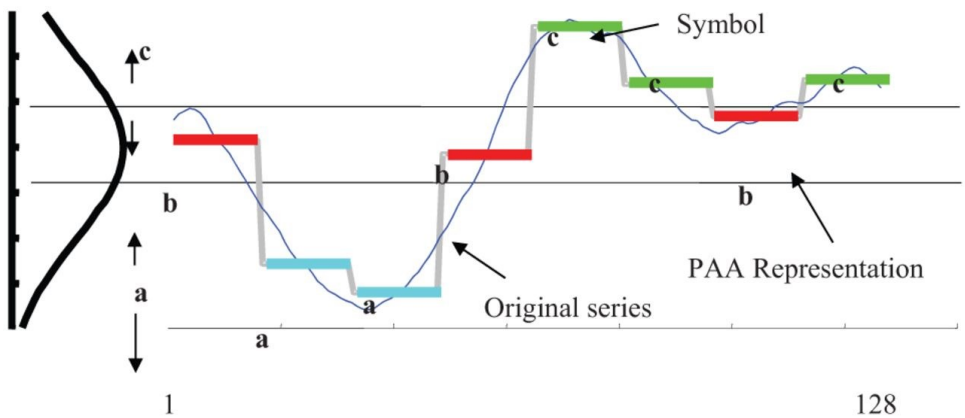
- High frequency noise in host load data can be smoothed with low-pass filters
- We use the Gaussian filter to remove noise
- Gaussian filter: $G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$
- Applying Gaussian filter is a convolution of Gaussian function and signal
- The filter takes nearby values with weights to form a new current value

Noise Reduction



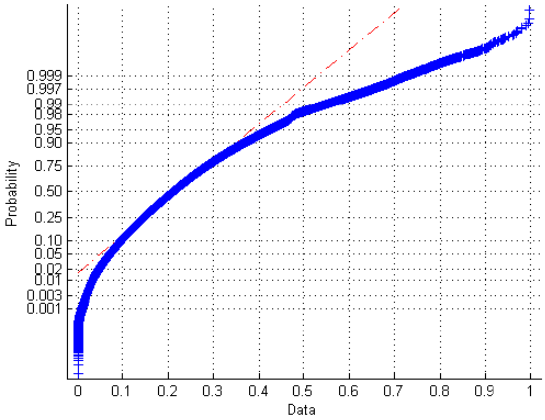
Dimension and Numerosity Reduction

- PAA representation
- Modified SAX representation
- They provide lower-bounding!



Modified SAX

- SAX preassumes data to be Gaussian distribution
- However our data distribution is like this:



Normal probability plot of host load data

- Applying SAX to this data may harm the effectivity of the method greatly!

Modified SAX

- Take a random subset of the whole dataset as a sample dataset
- Measure the distribution of the subset to infer the real distribution of the whole dataset
- Use this distribution as the evidence of SAX

Motif Discovery

- A brute force algorithm as mentioned:

```
Input:CPU host load: P, Q;
      Maximum allowed distance: R;
Output: Set of motifs
Brute_force_motif_search(P,Q,R)
{
 foreach (Host_load subsequence_P in P)
 {
    foreach (Host_load subsequence_Q in Q)
    {
      if (Distance(subsequence_P,
                   subsequence_Q) < R)
      {
       motif_set.add(subsequence_Q);
      }
    }
 }
}
```

Which I can not have any result with a time series data with 3000 data points before I get impatient

Motif Discovery

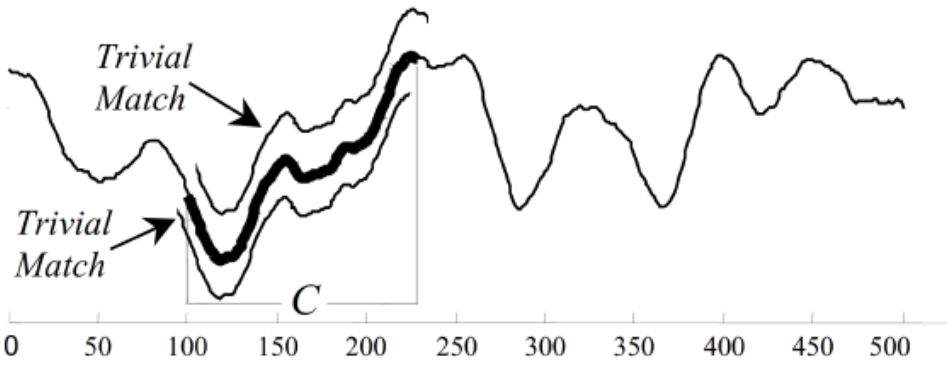
- This time we modified the brute force algorithm to find only longest motifs, witch reduces the search space greatly.

```
Input:CPU host load: P, Q;
      Maximum allowed distance: R;
Output: Set of motifs
LongestPossible(P,Q,R)
{
  for (i=0; i<P.length; i++)
  {
    for (j=0; j<Q.length; j++)
    {
      if (|P[i]-Q[j]|<R)
      {
        disntance=|P[i]-Q[j]|;
        count_p=i;
        count_q=j;
        while(sqrt(distance)<R)
        {
          distance+=power(P[count_p]-Q[count_q],2);
          count_p++;
          count_q++;
        }
        motif_set.add(Q[j to count_j-1)
      }
    }
  }
}
```

- However the algorithm gives many repeated motif

Motif Discovery

- Trivial match is useless to our task:



- Another place we can save time!

Motif Discovery

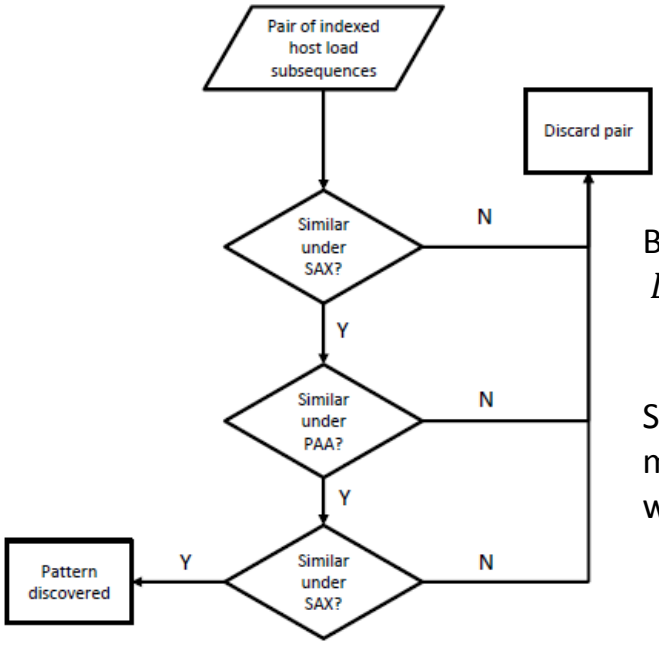
- This time trivial matches are skipped!

```
Input:CPU host load: P, Q;
      Maximum allowed distance: R;
Output: Set of motifs
TrivialMatchSkip(P,Q,R)
{
  for (i=0; i<P.length; i++)
  {
    for (j=0; j<Q.length; j++)
    {
      if (|P[i]-Q[j]|<R)
      {
        disntance=|P[i]-Q[j]|;
        count_p=i;
        count_q=j;
        while(Sqrt(distance)<R)
        {
          distance+=power(P[count_p]-Q[count_q],2);
          count_q++;
;       }
        i = i + (count_p - i) / 2;
        j = j + (count_q - j) / 2;
        motif_set.add(Q[count_j-1 to i]);
      }
    }
  }
}
```

Cascade Discovery

- Search in indexed space improves efficiency
- Meanwhile reduces accuracy
- Lower bounded indexing methods allow cascade discovery

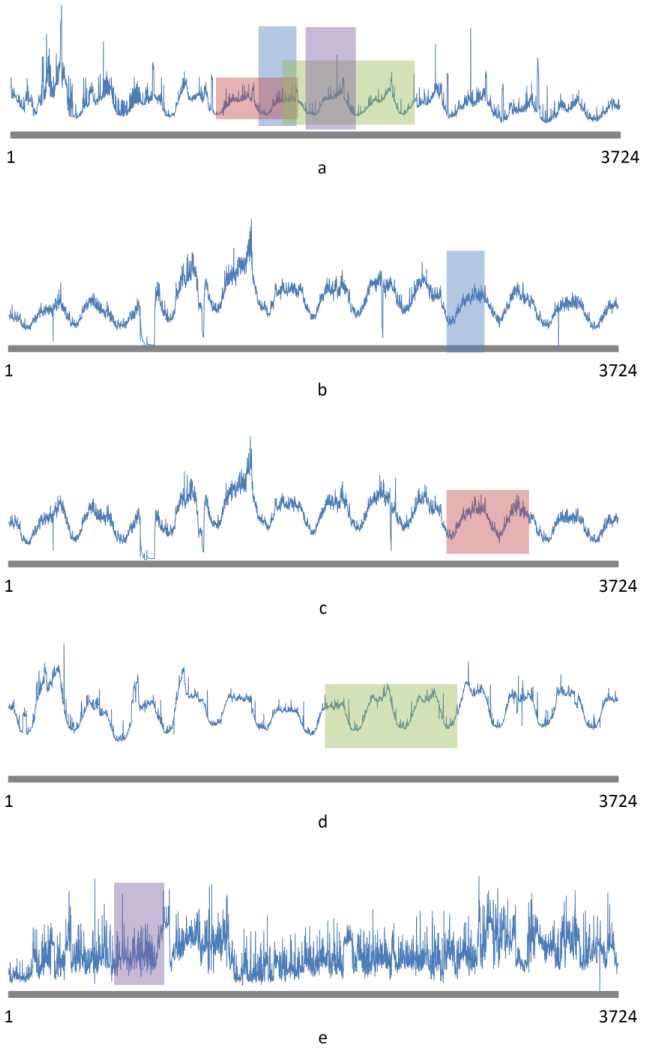
Cascade Discovery



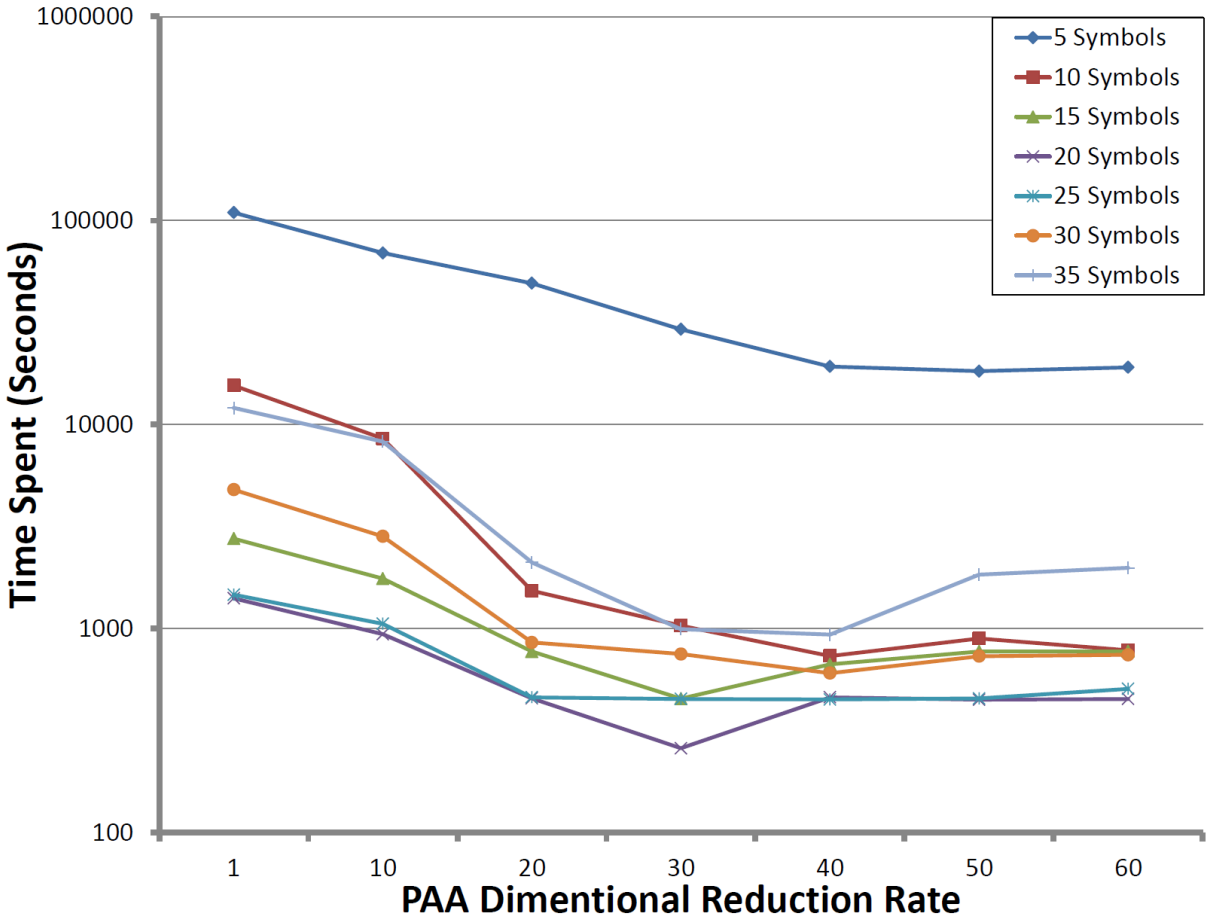
Because
$D_{SAX}(T_1, T_2) \geq D_{PAA}(T_1, T_2) \geq D(T_1, T_2)$

So any pair is a pattern under higher index must be a pattern under lower stage as well!

- Similar patterns are in same colour



WARWICK



Thank You!

THE UNIVERSITY OF
WARWICK

Full name: Chen Gu

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Poster

Title: phantom walkabouts in wireless sensor networks

Abstract: As wireless sensor networks (WSNs) have been applied across a spectrum of application domains, the problem of source location privacy (SLP) has emerged as a significant issue, particularly in security-critical situations. In the seminal work on SLP, phantom routing was proposed as a viable approach to address the issue. However, recent work has shown the limitations of phantom routing to handle issues such as multiple sources. We address some of those limitations and develop a novel version of phantom routing, namely phantom walkabouts, which are phantom routes of variable lengths. We show that phantom walkabouts provide state-of-the-art levels of source location privacy with often better message complexity (hence, energy usage).

Keywords: wireless sensor network, phantom routing, phantom walkabouts

Abstract Word Count (Est): 110

Phantom Walkabouts in Wireless Sensor Networks

Chen Gu

WARWICK
THE UNIVERSITY OF WARWICK

Abstract

As wireless sensor networks (WSNs) have been applied across a spectrum of application domains, the problem of source location privacy (SLP) has emerged as a significant issue, particularly in security-critical situations. In the seminal work on SLP, phantom routing was proposed as a viable approach to address the issue. However, recent work has shown the limitations of phantom routing to handle issues such as multiple sources. We address some of those limitations and develop a novel version of phantom routing, namely phantom walkabouts, which are phantom routes of variable lengths. We show that phantom walkabouts provide state-of-the-art levels of source location privacy with often better message complexity (hence, energy usage).

Introduction

A wireless sensor network (WSN) consists of a number of tiny devices, known as sensor nodes or motes, that can sense different attributes of the environment and use radio signals to communicate among themselves[1].

Threats to privacy in monitoring applications can be considered along two dimensions: (i) content-based threats and (ii) context-based threats. Context-based privacy threats focus on the context in which messages are broadcast and how information can be observed or inferred by attackers. Context is a multi-attribute concept that encompasses situational aspects of broadcast messages, including environmental and temporal information. On the other hand, content-based privacy threats relate to use of the content of the messages broadcast by sensor nodes. There has been much research addressing the issue of providing content privacy.

It is often desirable for the source of sensed information to be kept private in a WSN. Algorithms or techniques that protect this contextual information are said to provide source location privacy (SLP). A WSN set up to forward the information collected about an asset would typically consist of the following: a dedicated node for data collection called a *sink node*, many nodes that are involved in sending information from these assets called *source nodes*, and many other nodes in the network used to route/relay messages over multiple hops from the sources to the sink.

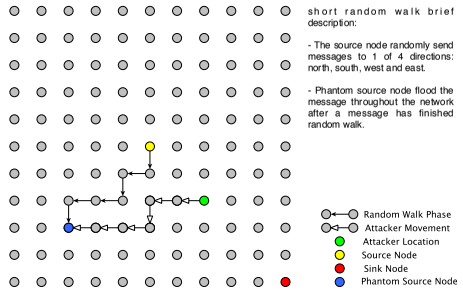
Main Contributions

1. The phantom walkabouts is a novel and more general technique than phantom routing, that help achieve a better trade-off between SLP and energy.
2. Via extensive simulations, we show the viability of phantom walkabouts. For example, under certain parametrization, phantom walkabouts achieve extremely high SLP with acceptable message overhead.

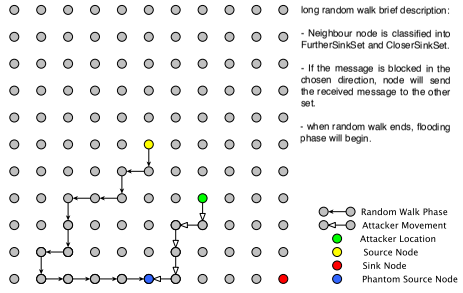
Theory

Phantom walkabouts are basically phantom routing with variable random walk lengths. Differing from phantom routing where the length of the random walk is small, phantom walkabouts perform sequences of short and long random routing.

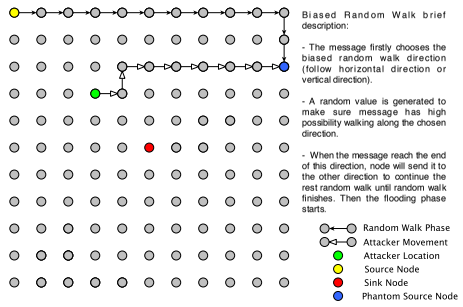
0.1 Short Random Walk Routing (phantom routing)



0.2 Long Random Walk Routing



0.3 Biased Random Walk Routing



A random value $\theta \in [0, 1]$ is a value to make sure message has high possibility walking along the direction δ. The node decide to send this message to the neighbour node by following equation:

$$f(\delta, \theta, \alpha) = \begin{cases} \delta & \text{if } \theta \in [0, \alpha] \\ \{H, V\} \setminus \{\delta\} & \text{otherwise} \end{cases} \quad (1)$$

0.4 Phantom Walkabouts

The phantom walkabouts technique extends the phantom routing protocol by adopting variable lengths of phantom routing. When a source node routes messages using phantom walkabouts, a message m_i is selected to either go on a short random walk of length s or long random walk of length l, according to the following:

$$PA(m, n) = \begin{cases} s & \text{if } m_i \bmod (m+n) \in (0, m], m, n \in \mathbb{N}^+ \\ l & \text{otherwise} \end{cases} \quad (2)$$

The sequencing of messages looks like as follow:

$$\underbrace{M_s, \cdots, M_s}_{m}, \underbrace{M_l, \cdots, M_l}_{n}, \underbrace{M_s, \cdots, M_s}_{m}, \underbrace{M_l, \cdots, M_l}_{n} \cdots$$

Results

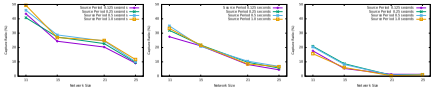


Figure 1: $PA(1, 0)$: Using **short** random walks

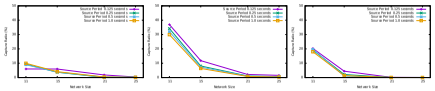


Figure 2: $PA(1, 1)$: Using **alternating short and long** random walks

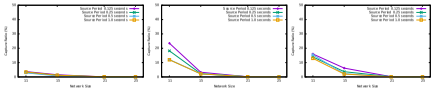


Figure 3: $PA(1, 2)$: Using **alternating short and two long** random walks

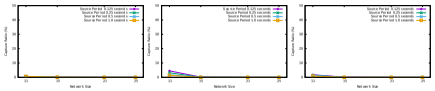


Figure 4: $PA(0, 1)$: Using **long** random walks

Figure 5: Source Location Privacy: Source Corner configuration with 1, 2 and 3 sources respectively

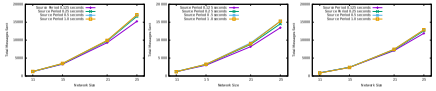


Figure 6: $PA(1, 0)$: Using **short** random walks

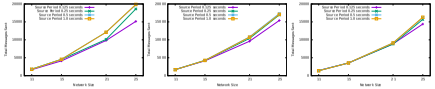


Figure 7: $PA(1, 1)$: Using **alternating short and long** random walks

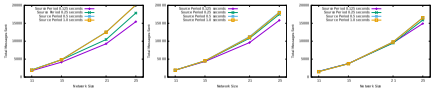


Figure 8: $PA(1, 2)$: Using **alternating short and two long** random walks

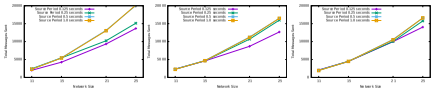


Figure 9: $PA(0, 1)$: Using **long** random walks

Figure 10: Messages Sent Source Corner with 1, 2 and 3 sources respectively

- The level of source location privacy increases (i.e., capture ration decreases) with increasing message rate. However, the higher number of messages lead to a much lower safety period, meaning that it is difficult for an attacker to capture the source within the safety period.
- The level of source location privacy increases with increasing number of sources.
- the alternating between a short and a long random walk in phantom walkabouts yields, in general, a higher level of SLP than when using short random walks during phantom routing, especially for the larger-sized networks.

Conclusions

- Phantom walkabouts propose to interleave sequences of short random walks and long random walks to attempt to make the attacker move in the wrong direction, as opposed to phantom routing (with small random walks) where an attacker moves towards the source, while reducing message overhead.
- Phantom walkabouts provide much better levels of SLP at certain parameterisation, albeit at only a small message overhead over phantom routing.

Forthcoming Research

- investigate phantom walkabouts with variable short and long random walks.
- with different network topologies.

References

[1] A. Mainwaring, J. Polastre, R. Szewczyk, and D. Culler. Wireless sensor networks for habitat monitoring. In *Proceedings of the First ACM International Workshop on Wireless Sensor Networks and Applications*, 2002.

Full name: Neha Gupta

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Poster

Title: SMA for City

Abstract: Social media data is widespread and can tell much about citizen's opinions to government provided services and campaigns. Raw social media data is biased towards different demographics and social patterns. De-biasing the data with cultural and technological awareness is important in the city context, to create valuable insights into citizen's perspectives of public services. In this research, we examine how different data sets can be combined to provide greater reliability to social media analytics. We also study how messages can spread, or be spread, effectively by targeting specific descriptions or social media audiences. Social media intelligence has applications in policing, social feedback, population monitoring and politics; our focus will be the citizen engagement of urban-centric government services.

Keywords: Social media Analytics, Citizen Engagement, Natural Language Processing

Abstract Word Count (Est): 117

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES

WARWICK

De-biasing Social Media Data to understand Citizen Engagement with Government Services

Social media data is widespread and can tell much about citizen's opinions to government provided services and campaigns. Raw social media data is biased towards different demographics and social patterns. De-biasing the data with cultural and technological awareness is important in the city context, to create valuable insights into citizen's perspectives of public services. In this research, we examine how different data sets can be combined to provide greater reliability to social media analytics. We also study how messages can spread, or be spread, effectively by targeting specific descriptions or social media audiences. Social media intelligence has applications in policing, social feedback, population monitoring and politics; our focus will be the citizen engagement of urban-centric government services.

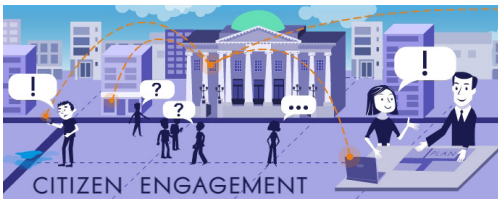
Related Work

Social media monitoring for government business goes beyond brand monitoring. The intention of social media monitoring by government is to understand citizen feedback, improve community services, emergency management in real time, create situational awareness for crime prevention, stop the manifestation of rumours and broadcast a message about a policy yet to unveil for constructive feedback.

Linders highlights how the presence of social media technology is empowering citizens to play an active role in the function of their government [1]. He suggests three implementations of government engagement with the citizen - citizen sourcing, government as a platform and 'do it yourself' government.

Fernandez et al unearth interesting findings regarding what kind of audience share their opinion about policy making on Twitter in Germany [2]. They concluded that high volume conversations around policy topics do not typically come from citizens, but are instead from news agencies and other organizations. Therefore, the underlying question remains, how much is a government actually engaging with its citizens and, is this feedback biased by a view shared by a small group of organizations.

Recent research on 31 data-enabled cities to understand citizen engagement with government, highlights the strong differences between the respective popularity of social media services in each city. The most-used social media platforms are Twitter, YouTube and Facebook [3]. Therefore, a more detailed content analysis as well as a more differentiated analysis of users must be conducted, taking into account the demographics of the city and linking social media data with other sources of data to achieve a better insight into government services.



In an increasingly complex world, citizens' input is a critical resource for policy-making (OECD, 2009)

Other research identifies how to engage the right individuals at the right time on social media, to effectively propagate information within a given time frame [4]. They develop a feature-based model that learns people's exhibited social behaviour, including the content of their tweets and social interactions, to characterize their willingness and readiness to propagate information to other users. They also develop a model based on a user's previous retweeting wait-times, to predict their next retweet time.

Aim of this Research

Social media data alone cannot answer how engaged people are with government. There is also a growing tendency towards citizen co-production, whereby citizens take on the role of partner in the definition and delivery of public services. To improve citizen representation and responsiveness to a government service or initiative, social media feedback about a government service should be combined with other datasets. To create this affinity with the citizen, it is also important to understand the parameters which accelerate information diffusion in social media spaces.

"You have to understand your target audience in three ways. How they discover information, their preferences for consumption and what motivates them to take action."

Lee Odden,
CEO of TopRank Online Marketing

Therefore, this research investigates:

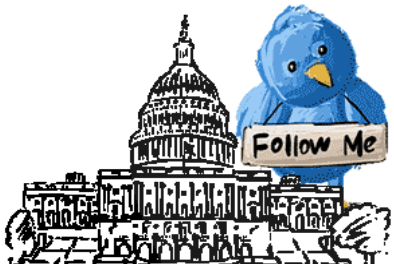
1. How we can combine social media data with other datasets for better representation of citizen feedback on government services and initiatives;

2. How we can increase the reach of government messaging by understanding viral trends on social media;

3. How the emergence of new 'social contracts' empower the public in playing a more active role in the functioning of government.

Industry Links and Research Impact

Until recently, social media use by government bodies has been based on the information "push" strategy. However, latest research promises that if government resources are committed to engage through a "pull" or "networking" strategy, the inclusion of diverse public opinion will help government organizations increase the quality and effectiveness of their processes [5, 6].



New York hired a chief digital officer in January 2011 and has since had a fully-fledged social media strategy

Social media analytics has been shown to help businesses gather competitive intelligence and understand better their business environments, suppliers, and competitors. Based on a report published by Techcrunch [7] the most successful brands monitor and measure their social media index using various tools. One of the most successful companies to implement social media intelligence is Dell, which has more than 21 million social connections on Facebook and Twitter. Through their site, Dell has collected over 24,000 ideas from crowdsourcing social feedback, of which 550 have been implemented in their product line since 2007.

Microstrategy, IBM Cognos, Oracle and SAS also incorporate social media intelligence into their existing business platforms. Their application blends social insights with more-traditional data, including private conversation data from call center transcripts, transactional, behavioural, referral and web analytics.

Through our relationship with CUSP and the City Analytics Unit in New York, we are fortunate to have access to large city-related social media archives.

References

[1] D. Linders, "From e-government to we-government: Defining a typology for citizen coproduction in the age of social media," Gov. Inf. Q., vol. 29, no. 4, pp. 446–454, 2012.

[2] M. L. T. Cossio, et al, "Using Social media to Inform Policy Making," Uma ética para quantos?, vol. XXXIII, no. 2, pp. 81–87, 2012.

[3] A. Mainka, S. Hartmann, W. G. Stock, and I. Peters, "Government and Social Media: A Case Study of 31 Informational World Cities," 2014 47th Hawaii Int. Conf. Syst. Sci., pp. 1715–1724, 2014.

[4] J. Lee, Kyumin and Mahmud, Jalal and Chen, Jilin and Zhou, Michelle and Nichols, "Who Will Retweet This ? Automatically Identifying and Engaging Strangers on Twitter to Spread Information," Proc. 19th Int. Conf. Intell. User Interfaces, 2014.

[5] I. Mergel, "A framework for interpreting social media interactions in the public sector," Gov. Inf. Q., vol. 30, no. 4, pp. 327–334, 2013.

[6] I. Mergel, "Social media adoption and resulting tactics in the U.S. federal government," Gov. Inf. Q., vol. 30, no. 2, pp. 123–130, 2013.

[7] "techcrunch." [Online]. Available: http://techcrunch.com/2011/09/13/dachis-group-debuts-social-business-index-think-of-it-as-klout-for-companies/.

Neha Gupta
PhD 2014 - 2018

Full name: Suncica Hadzidedic

Year of Study: PhD Year 3

Department: Computer Science

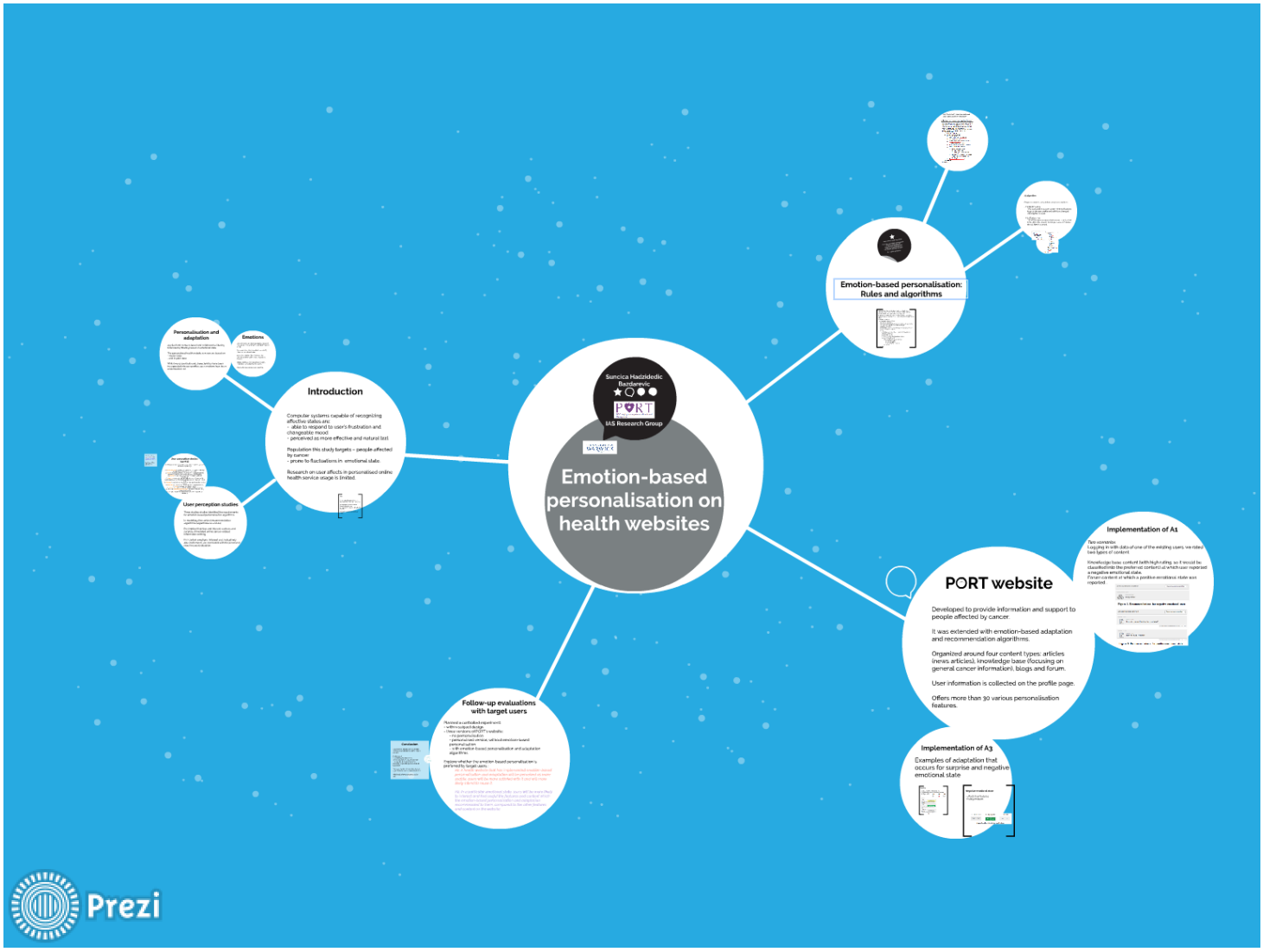
Submission Type: Presentation

Title: Emotion-based personalisation on health websites

Abstract: Affective computing has received substantial attention in the recent time. However, its application to personalised online health services is understudied. In this paper, we show, based on intensive user studies, how emotions affect user interaction with health websites. With regard to this, we propose adaptive techniques applicable to cancer-oriented health websites, by modeling algorithms for content recommendation, as well as website layout and navigation adaptation, that generate results tailored to users' emotional state. The emotional state of a user is measured as the intensity of 9 basic emotions a user experiences at the start of website use. Furthermore, we present the implementation of the algorithms to a real-life health website for people affected by cancer. Finally, initial evaluations are performed, and large-scale evaluations are designed.

Keywords: personalisation; adaptation; emotions; health websites; recommendation algorithms; filtering techniques

Abstract Word Count (Est): 125


Suncica Hadzidedic Bazdarevic
PORT
IAS Research Group
THE UNIVERSITY OF WARWICK
Emotion-based personalisation on health websites
Introduction
Computer systems capable of recognizing affective status are:
- able to respond to user's frustration and changeable mood
- perceived as more effective and natural (a)
Population this study targets – people affected by cancer
- prone to fluctuations in emotional state
Research on user affects in personalised online health service usage is limited.
Personalisation and adaptation
Emotions
User perception studies
Emotion-based personalisation: Rules and algorithms
PORT website
Developed to provide information and support to people affected by cancer
It was extended with emotion-based adaptation and recommendation algorithms.
Organized around four content types: articles (news articles), knowledge base (focusing on general cancer information), blogs and forum.
User information is collected on the profile page.
Offers more than 30 various personalisation features.
Implementation of A1
Implementation of A3
Examples of adaptation that occurs for surprise and negative emotional state
Follow-up evaluations with target users
Conclusion
Prezi


Suncica Hadzidedic Bazdarevic
PORT
NGO helping cancer patients in Bosnia and Herzegovina
IAS Research Group
THE UNIVERSITY OF WARWICK
Emotion-based personalisation on health websites
Prezi

Personalisation and adaptation

Applied both content-based and collaborative filtering
Extended by filtering based on emotional state

The personalised health website services are based on:
- explicit data
- and implicit data

Whilst many user traits and characteristics have been incorporated into user profiles, user emotions have been understudied [32]

Emotions

Basic emotions, are more appropriate to use in research on interaction with a computer system [8, 43].

We observe individual emotions,and classify them into emotional states

9 emotions studied: Fear, Sadness, Joy, Surprise, Shame, Guilt, Anger, Disgust, and Interest.

Various methods for measuring emotions - objective and subjective measures.

This study relied on user self-reporting.

Introduction

Computer systems capable of recognizing affective states are:
- able to respond to user's frustration and changeable mood
- perceived as more effective and natural [22].

Population this study targets – people affected by cancer
- prone to fluctuations in emotional state.

Research on user affects in personalised online health service usage is limited.

Goal

Explore emotion recognition in personalised online health systems.

Create algorithms for emotion-based adaptation and personalisation on an existing health website.

Evaluate the website with target users.

User perception studies (cont'd)

R2 findings were used to match the adaptation to the website layout and navigation (algorithm A3).

C3. Embarrassment is negatively correlated with the feature (F1) **filtering recommended content** on the user's profile

C2. Fear is negatively correlated with 3 features - (F2) **search outcomes presentation**, (F3) **filtering search outcomes** and (F4) **blogging**.

C3. Sadness is negatively correlated with (F2) **search outcomes presentation**.

C4. Surprise is negatively correlated with 2 features - (F5) **user profile customisation** and (F6) **external links** about each data field in profile editing.

C5. Boredom is negatively correlated with (F7) **commenting** on the website's content.

C6. Boredom correlates negatively with (F8) and (F9) **adapting text size**.

C7. Guilt is in a negative correlation with (F6)

C8. Interest is in a positive correlation with (F9) **adapting text colour**.

C9. Pride usage negative correlation with (F10) negatively correlated with 3 features: F6, F8 and F9

R3. Being in a **negative emotional state** users prefer **adaptivity** over adaptability

User perception studies

Three studies studies identified the requirements for emotion-based personalisation algorithms.

In modelling the content recommendation algorithms (algorithms A1 and A2).

R1.1 implied that fear and interest, sadness and surprise, stimulated online cancer-related information seeking.

R1.2 Certain emotions (interest and, indicatively, also excitement), are correlated with the perceived need for personalisation.

Prezi

Introduction

Computer systems capable of recognizing affective states are:
- able to respond to user's frustration and changeable mood
- perceived as more effective and natural [22].

Population this study targets – people affected by cancer
- prone to fluctuations in emotional state.

Research on user affects in personalised online health service usage is limited.

Goal

Explore emotion recognition in personalised online health systems.

Create algorithms for emotion-based adaptation and personalisation on an existing health website.

Evaluate the website with target users.

Prezi

Goal

Explore emotion recognition in personalised online health systems.

Create algorithms for emotion-based adaptation and personalisation on an existing health website.

Evaluate the website with target users.

Personalisation and adaptation

Applied both content-based and collaborative filtering
Extended by filtering based on emotional state

The personalised health website services are based on:
- explicit data
- and implicit data

Whilst many user traits and characteristics have been incorporated into user profiles, user emotions have been understudied [32]

E

Basic emotio
research on i
[6, 43].

We observe i
them into en

9 emotions s
Surprise; Sha
Interest.

Various meth
- objective ar

This study rel

ing

d on:

en
been

Emotions

Basic emotions, are more appropriate to use in research on interaction with a computer system [6, 43].

We observe individual emotions,and classify them into emotional states

9 emotions studied: Fear; Sadness; Joy; Surprise; Shame; Guilt; Anger; Disgust; and Interest.

Various methods for measuring emotions - objective and subjective measures.

This study relied on user self-reporting.

C7. Guilt is in a negative correlation with **F1**;
C8. Interest is in a positive correlation with F9: **adapting text colour**;
C9. Pre-usage negative emotional state is negatively correlated with 3 features: **F1, F5** and **F8**.
C10. Being in a **negative emotional state** users prefer **adaptivity** over adaptability

User perception studies

Three studies studies identified the requirements for emotion-based personalisation algorithms.

In modelling the content recommendation algorithms (algorithms A1 and A2).

R1.1 implied that fear and interest, sadness and surprise, stimulated online cancer-related information seeking.

R1.2 Certain emotions (interest and, indicatively, also excitement), are correlated with the perceived need for personalisation.

into:

oy and

adness,
nd guilt

oredom

User perception studies (cont'd)

R2 findings were used to match the adaptation to the website layout and navigation (algorithm A3).

C1. Embarrassment is negatively correlated with the feature (F1) **filtering recommended content** on the user's profile.
C2. Fear is negatively correlated with 3 features – (F2) **search outcomes presentation, (F3) filtering search outcomes** and (F4) **blogging**.
C3. Sadness is negatively correlated with (F2) **search outcomes presentation**.
C4. Surprise is negatively correlated with 2 features – (F5) **user profile customisation** and (F6) **information bubbles** about each data field in profile editing.
C5. Boredom is negatively correlated with (F7) **commenting** on the website's content.
C6. Boredom correlates negatively with **F5** and F8: **adapting text size**;
C7. Guilt is in a negative correlation with **F1**;
C8. Interest is in a positive correlation with F9: **adapting text colour**;
C9. Pre-usage negative emotional state is negatively correlated with 3 features: **F1, F5** and **F8**.
C10. Being in a **negative emotional state** users prefer **adaptivity** over adaptability

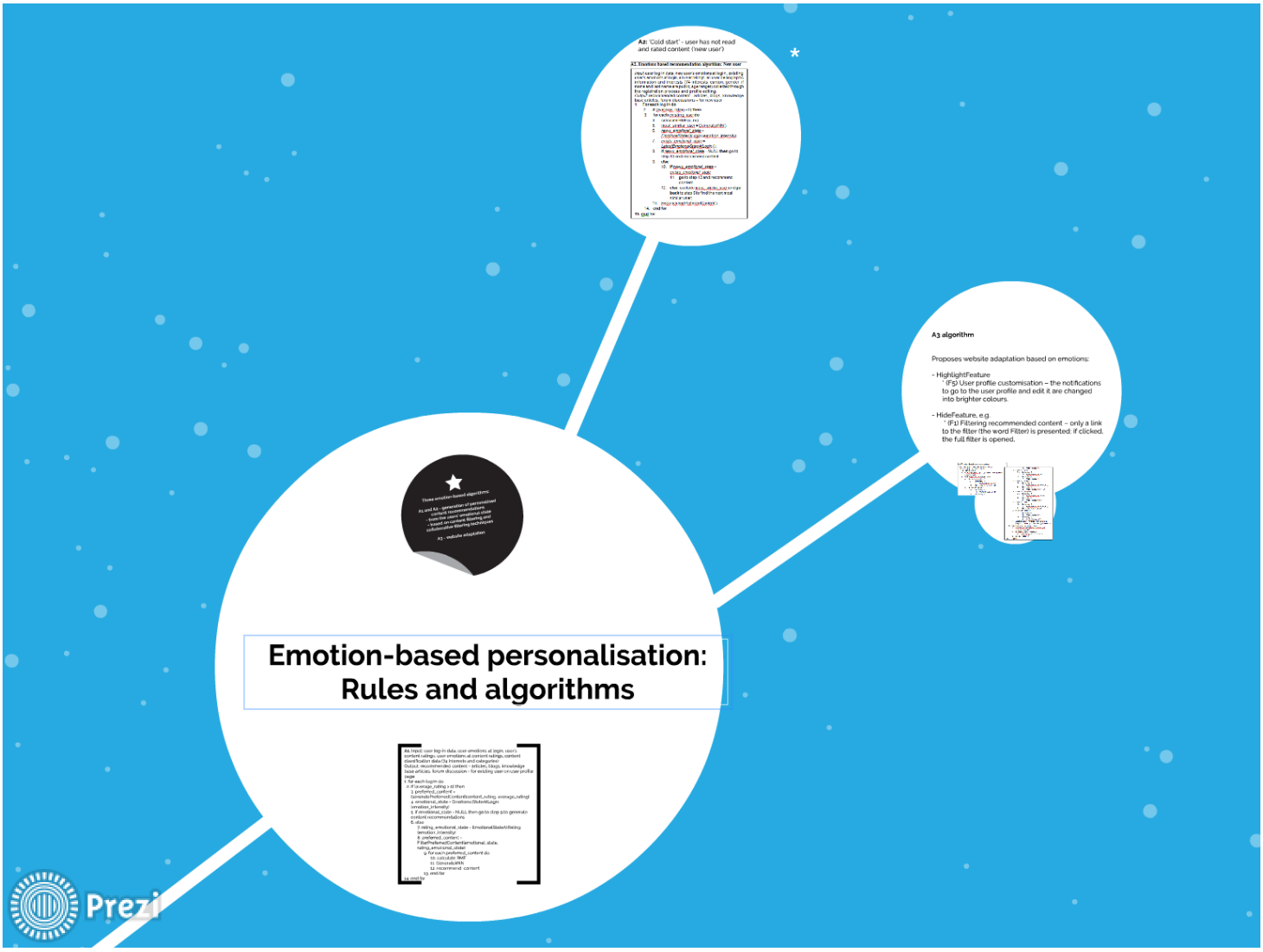
Prezi

9 emotions classified into:

- 3 **positive**: interest, joy and surprise
- 6 **negative**: anger, sadness, disgust, fear, shame and guilt

Embarrassment and *boredom* were mapped to
- *shame* and *disgust*, respectively.

Prezi



Three emotion-based algorithms:

A1 and A2 - generation of personalised content recommendations
- from the users' emotional state
- based on content filtering and collaborative filtering techniques

A3 - website adaptation

A1. Input: user log-in data, user emotions at login, user's content ratings, user emotions at content ratings, content classification data (74 interests and categories)
Output: recommended content - articles, blogs, knowledge base articles, forum discussion - for existing user on user profile page

```
1. for each log in do
  2. if (average_rating > 0) then
     3. preferred_content =
     GeneratePreferredContent(content_rating, average_rating)
     4. emotional_state = EmotionalStateAtLogin
     (emotion_intensity)
     5. if emotional_state = NULL then go to step 9 to generate
     content recommendations
     6. else
        7. rating_emotional_state = EmotionalStateAtRating
        (emotion_intensity)
        8. preferred_content =
        FilterPreferredContent(emotional_state,
        rating_emotional_state)
           9. for each preferred_content do
              10. calculate RMF
              11. GeneratekNN
              12. recommend  content
           13. end for
14. end for
```

A2: 'Cold start' - user has not read and rated content ('new user')

A2. Emotions based recommendation algorithm: New user

Input: user log in data, new user's emotions at login, existing user's emotions at login, all user ratings, all user demographic information and interests (74 interests, canton, gender, if name and last name are public, age ranges) collected through the registration process and profile editing.
Output: recommended content - articles, blogs, knowledge base articles, forum discussions - for new user

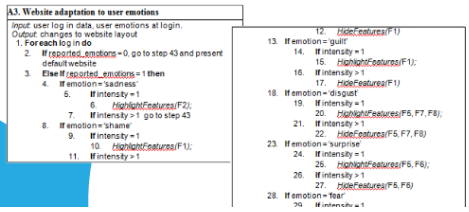
```
1.  For each log in do
    2.  if (average_rating = 0) then
    3.  for each existing_user do
        4.  calculate RMF(u, nu)
        5.  most_similar_user = GeneratekNN();
        6.  newu_emotional_state =
            EmotionalStateAtLogin (emotion_intensity)
        7.  existu_emotional_state =
            LatestEmotionalStateAtLogin ();
        8.  if newu_emotional_state = NULL then go to
            step 13 and recommend content
        9.  else
            10. if newu_emotional_state =
                existu_emotional_state
                11. go to step 13 and recommend
                    content
            12. else, exclude most_similar_user and go
                back to step 5 to find the next most
                similar user;
        13. RecommendPreferredContent();
    14. end for
15. end for
```

A3 algorithm

Proposes website adaptation based on emotions:

- HighlightFeature
  * (F5) User profile customisation – the notifications to go to the user profile and edit it are changed into brighter colours.

- HideFeature, e.g.
  * (F1) Filtering recommended content – only a link to the filter (the word Filter) is presented; if clicked, the full filter is opened.



3. Website adaptation to user emotions

Input: user log in data, user emotions at login.
Output: changes to website layout

```
1. For each log in do
   2.  If reported_emotions = 0, go to step 43 and present
       default website
   3.  Else If reported_emotions = 1 then
       4.  If emotion = 'sadness'
           5.  If intensity = 1
               6.  HighlightFeatures(F2);
           7.  If intensity > 1  go to step 43
       8.  If emotion = 'shame'
           9.  If intensity = 1
               10. HighlightFeatures(F1);
           11. If intensity > 1
               12. HideFeatures(F1)
       13. If emotion = 'guilt'
           14. If intensity = 1
               15. HighlightFeatures(F1);
           16. If intensity > 1
               17. HideFeatures(F1)
       18. If emotion = 'disgust'
           19. If intensity = 1
               20. HighlightFeatures(F5, F7, F8);
           21. If intensity > 1
               22. HideFeatures(F5, F7, F8)
       23. If emotion = 'surprise'
           24. If intensity = 1
               25. HighlightFeatures(F5, F6);
           26. If intensity > 1
               27. HideFeatures(F5, F6)
       28. If emotion = 'fear'
           29. If intensity = 1
               30. HighlightFeatures(F2, F3, F4);
           31. If intensity > 1 go to step 43
       32. If emotion = 'interest'
           33. If intensity = 1
               34. HideFeatures(F9);
           35. If intensity > 1
               36. HighlightFeatures(F9)
       37. Else If emotion ≠ ('interest', 'fear', 'surprise',
           'disgust', 'shame', 'guilt', 'sadness') go to step 43
   38. Else if reported_emotions > 1 then
       39. emotional_state = EmotionalStateAtLogin
           (emotion_intensity)
       40. if emotional_state = 'negative'
               41. HideFeatures(F1, F5, F9, F10);
       42. else go to step 43
   43. end for
```

PORT website

Developed to provide information and support to people affected by cancer.

It was extended with emotion-based adaptation and recommendation algorithms.

Organized around four content types: articles (news articles), knowledge base (focusing on general cancer information), blogs and forum.

User information is collected on the profile page.

Offers more than 30 various personalisation features.

Logging in with data c
two types of content

Knowledge base cont
classified into the pre
a negative emotional
Forum content at whi
reported .

Figure 2. R

Figure 3. I

Emotion tool

One of the essential features is the instrument for measuring emotions.

- a dialog box
- a yellow icon with emoticon
- clicking on ET expands it
- automatically pops-out at login and rating

e

support to

daptation

rticles

Implementation of A1

Two scenarios
Logging in with data of one of the existing users, we rated two types of content

Knowledge base content (with high rating, so it would be classified into the preferred content) at which user reported a negative emotional state.
Forum content at which a positive emotional state was reported .

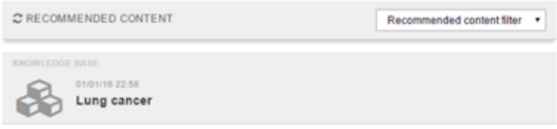


Figure 2. Recommendations for negative emotional state

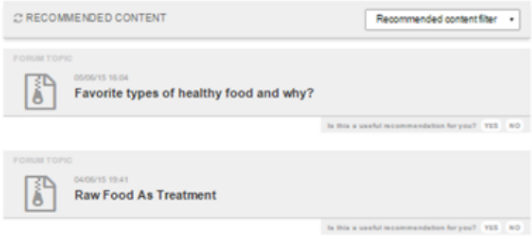
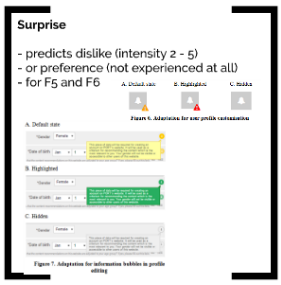


Figure 3. Recommendations for positive emotional state

Implementation of A3

Examples of adaptation that occurs for surprise and negative emotional state



Negative emotional state

- affects four features
- hiding features

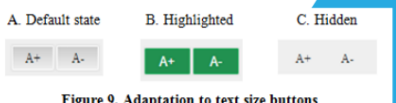


Figure 9. Adaptation to text size buttons

Surprise

- predicts dislike (intensity 2 - 5)
- or preference (not experienced at all)
- for F5 and F6

Figure 6. Adaptation for user profile customisation

Figure 7. Adaptation for information bubbles in profile editing

Negative emotional state

- affects four features
- hiding features

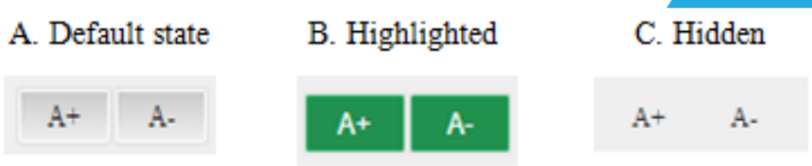


Figure 9. Adaptation to text size buttons

Follow-up evaluations with target users

Planned a controlled experiment:
- within-subject design
- three versions of PORT's website:
  - no personalisation
  - personalised version, without emotion-based personalisation
  - with emotion-based personalisation and adaptation algorithms.

Explore whether the emotion-based personalisation is preferred by target users :

H1. A health website that has implemented emotion-based personalisation and adaptation will be perceived as more usable, users will be more satisfied with it and will more likely intend to reuse it.

H2. In a particular emotional state, users will be more likely to interact and find useful the features and content which the emotion-based personalisation and adaptation recommended to them, compared to the other features and content on the website.

Prezi

Conclusion

Explored the introduction of emotion-based personalisation to online health domain

Challenge of:
- combining emotions with personalisation on a health website
- being able to connect specific emotions to specific types of desired adaptation

Proposed design of 3 emotion-based algorithms for cancer-related websites

Initial results favour personalisation, but...

The initial, requirement-based evaluations with real users

show a good match for the extracted correlations between sentiments and adaptation features.

promising a more appropriate way of designing adaptation for this especially sensitive group of users, who are cancer patients and their caregivers.

Prezi

Suncica Hadzidedic Bazdarevic
PORT
IAS Research Group

Emotion-based personalisation on health websites

Introduction

Computer systems capable of recognizing affective status are:
- able to respond to user's frustration and changeable mood
- perceived as more effective and natural (a)

Population this study targets – people affected by cancer:
- prone to fluctuations in emotional state

Research on user affects in personalised online health service usage is limited.

Personalisation and adaptation

Emotions

User perception studies

Emotion-based personalisation: Rules and algorithms

PORT website

Developed to provide information and support to people affected by cancer.

It was extended with emotion-based adaptation and recommendation algorithms.

Organized around four content types: articles (news articles), knowledge base (focusing on general cancer information), blogs and forum.

User information is collected on the profile page.

Offers more than 30 various personalisation features.

Implementation of A1

Two scenarios:
Logging in with data of one of the existing users, we noted two types of content

Knowledge base content (with high rating, so it would be classified into the preferred content) at which user reported a negative emotional state.
Forum content at which a positive emotional state was reported.

Implementation of A3

Examples of adaptation that occurs for surprise and negative emotional state

Follow-up evaluations with target users

Planned as a controlled experiment
- within-subject design
- three versions of PORT website:
  - non-personalised version
  - personalised version, ratings based
  - emotion-based personalisation and adaptation algorithm

Purpose: whether the emotion-based personalisation is preferred by target users

Hypothesis: Users will prefer the version with emotion-based personalisation and adaptation and be prevented or more quickly alerted when they are stressed with. It and will more deep identify issues.

H0: In a stressful or emotional state users will be more likely to choose the non-personalised version of the website. Or the emotion-based personalisation and adaptation recommended by them, compared to the other versions, will lead to a decrease in the website

Conclusion

Full name: Vikki Houlden

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Greenspace Characteristics and Mental Wellbeing in the Urban Environment

Abstract: With over half the global population now living in cities, urban planners are faced with the challenge of creating environments to accommodate increasing numbers of residents and an changing way of life; they also have a social responsibility to ensure that neighbourhoods can be positive, happy places for their inhabitants. However, research shows that mental health and wellbeing are generally poorer in urban than rural areas. While mental wellbeing has been linked to increased longevity, productivity and societal prosperity, it can be strongly influenced by our surroundings. Evolutionary theories suggest that humans are happiest in natural environments, which are typically designed into urban areas at the expense of buildings, although it is not yet known which types of urban greenspace are most beneficial, and to who.

This presentation gives an overview of the current state of research regarding mental wellbeing and urban greenspace. Primary studies have involved national-level regressions of proportional greenspace across England, and will advance using GIS-based analyses to categorise different greenspace characteristics alongside modelling area-level accessibility, to investigate how different individuals interact with their local greenspace, and uncover the driving factors in this relationship.

Social-level interventions, such as improving urban greenspace, can be a cost-effective and far-reaching method of improving mental health and wellbeing, and there may be implications for wider issues such as the environment and economy.

Keywords: greenspace, mental wellbeing, nature, health, GIS

Abstract Word Count (Est): 221

Full name: Peng Jiang

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Optimization on Virtualization Platform for Parallel Jobs

Abstract: Virtualization technology, such as Xen, allows multiple virtual machines (VM) to co-exist in a single multi-core physical machine. Multiple VMs may host different processes of the same parallel program. Ideally, the multiple VMs should be scheduled to run simultaneously by the VM manager. However, the existing schedulers in Xen do not support this type of co-scheduling. Therefore, a scheduler with supporting parallel jobs is urgently required by industry as parallel working is tending to become main stream in design software.

Some solutions have been proposed recent years, most of them are trying to solve this problem in a traditional way, namely, scheduling. Some novel scheduling algorithms haven been proposed. However, traditional scheduling only approach can not solve this problem properly. Here I propose a new thought of solution: Re-queuing mechanism. This mechanism is an ideal solution to our identified problem, which manipulates running queue without interfering the original features of Credit Scheduler. It keeps the fairness brought by Credit scheduler, and also improves the performance of parallel jobs on Xen.

Keywords: Xen Parallel

Abstract Word Count (Est): 170

Full name: Jack Kirton

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Source Location Privacy-Aware Data Aggregation Scheduling for Wireless Sensor Networks

Abstract: Source Location Privacy is a specific problem to a set of WSNs charged with the task of asset monitoring. In a WSN such as this, one has a sink (that collects data from the entire network) and a source (the currently broadcasting node). A node is said to broadcast when it becomes aware of the presence of a particular asset. This information is then routed back to the sink. The problem arises when we wish to keep the location of the asset private, as one could simply find the source of data and they would have captured the asset. SLP is the art of preventing this case.

Many different methods of providing SLP have been devised, usually involving some combination of phantom routing and fake sources. These methods both provide protection at the routing layer of the network. Solutions at the routing layer often impose a high message overhead on the SLP-aware protocol.

The purpose of this presentation is to show a solution for SLP-awareness at the MAC layer of a TDMA network. In this solution, a simple Data Aggregation Scheduling (DAS) algorithm will be devised for the basic TDMA network. This will then be built upon by discussing methods of creating diversionary routes through the network finally deploying such methods into an SLP-aware DAS algorithm with the goal of providing some level of SLP and having reduced message overhead compared to routing-based approaches.

Keywords: Source Location Privacy; Wireless Sensor Networks; TDMA Schedule; DAS

Abstract Word Count (Est): 234

Full name: Elena Kochkina

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Opinion Mining from Textual and Visual Data in Social Media

Abstract: Internet users spend the majority of their time online on social media sites exchanging information in the form of text, image or video. Researchers have been increasingly using data from social media platforms to address questions about individual and group behaviour, sentiment and opinion .

Since 'a picture is worth a thousand words' information is often disseminated in terms of images as well as text, especially when one intends to make an impression on their audience.Thus it is crucial to perform analysis of both text and images to obtain a better understanding of any information conveyed, its veracity, or the mindset of individuals. My project focuses on multimodal analysis of social media, it seeks to combine information extracted from images and text to gain insights on people's stance or sentiment towards various issues and to understand phenomena such as rumour spread and opinion formation.

I have approached this task by exploring existing Distributional Semantic Models (DSMs), which create representations of words as high-dimensional real-valued vectors, that can be trained through unsupervised algorithms and used as features instead of hand-engineered ones. Initially, I looked into application and properties of text-based DSMs such as word2vec and then multimodal ones. Currently I am applying context-based word embeddings in a tweet stance classification task using LSTM, a type of recurrent neural network, which is well suited to processing time series data and can have memory properties. My present work will feed into my broad aim of multimodal analysis of social media.

Keywords: social media, opinion mining, neural networks

Abstract Word Count (Est): 247

Full name: Tejas Kulkarni

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Beyond the Geometric Mechanism.

Abstract: Getting access to information is one of the key challenges in any data science project. And many important datasets (e.g medical records) are not opened to the general public due to privacy concern. And that's where the field of statistical privacy comes into picture. However, any method for anonymizing sensitive datasets must resolve a trade-off between utility and privacy. My general area of research is differential privacy which is a theoretical framework for quantifying dataset privacy. Our overall goal is to subject machine learning/data mining algorithms to differential privacy and provide strong utility guarantees.

We are studying the conflict between privacy and database utility as an optimization problem and we have proposed a formal framework to encode our desired privacy requirements as linear programming constraints. We are interested in finding a generic taxonomy of privacy preserving mechanisms based on privacy-utility promises they provide.

Main result: Inspired from the geometric mechanism (a standard procedure for perturbing data, guaranteeing differential privacy), we have proposed novel structural properties for probability mass function for these mechanisms and studied impact of enforcing these properties over utility of a mechanism. As objective function, we have used standard l0 norm (number of error a mechanism makes). l1 norm (computes mean of a mechanism,) and l2 norm (variance of a mechanism). This project is a good mixture of theoretical and applied skills.

Keywords: algorithms, database privacy, mechanisms,statistical database

Abstract Word Count (Est): 225

Full name: Tim Law

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Optimisation of a Molecular Dynamics Simulation of Chromosome Condensation

Abstract: We present optimisations applied to a bespoke biophysical molecular dynamics
simulation designed to investigate chromosome condensation. Our primary focus is on domain-specific algorithmic improvements to determining short-range interaction forces between particles, as certain qualities of the simulation render traditional methods less effective. We implement tuned versions of the code for both traditional CPU architectures and the modern many-core architecture found in the Intel Xeon Phi coprocessor and compare their effectiveness. We achieve speed-ups starting at a factor of 10 over the original code, facilitating more detailed and larger-scale experiments.

Keywords: hpc,optimisation,molecular dynamics,many-core

Abstract Word Count (Est): 89

Full name: Zhenyu Li

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Convergence of HPC and Big Data Analytics

Abstract: High Performance Computing (HPC) and Big Data Analytics (DA) are two communities that grew out of two different needs, process speed and the ability to process large volume of data. The characteristics of HPC applications are small input and large output, whereas big DA is the opposite. As such, HPC is traditionally used in scientific programming, and big data analytics in commerce. Recent trends of development, that HPC applications start to make use of large volume of scientific data and data analytic algorithms becoming more complex, create a convergence of need for the HPC and DA communities. However, the software and hardware architectures of HPC and DA systems are not compatible and have been diverging for the past decade. For instance, HPC and DA have entirely different software stacks, programming models and tools. As such, a system that is capable of running both types of jobs is hugely desired. In this presentation, we will be exploring the challenges and opportunities for bringing the HPC and DA architectures together. We will also be talking about how we propose to solve this problem and an executive plan of accomplishing it.

Keywords: high performance computing, big data analytics

Abstract Word Count (Est): 188

Convergence of HPC and Big Data Analytics

Presenter: Zhenyu Li (1st yr. Mphil/PhD)

Supervisor: Prof. Stephen Jarvis

Table of Content

Table of Content

HPC & Big Data Analytics

HPC:

1. High-Performance (Scientific) Computing
2. HPC grew out of speed to solve algorithmically complex problems
3. Small Input, Large Output

Big Data Analytics:

1. Data analytics/Map-Reduce grew out of the need to process large volume of data
2. Focus on resilience & programmability
3. Large Input, Small Output

HPC & Data Analytics – Traditional Applications

HPC:

1. Finite Element Analysis: LS-DYNA, CTH
2. Computational Fluid Dynamics: OpenFOAM, Overflow
3. Chemistry & Material: VASP, TBMD
4. Atmospheric Modeling & Climate Prediction: 3DVAR, 4DVAR, WRF

Big Data Analytics:

1. E-Commerce:
    1. Web-browsing histories
    2. Social media sentiments
    3. Purchase Histories and Trends

Table of Content

WARWICK

HPC and Data Analytics - Convergence

1. Convergent Needs:
   1. For HPC:
      1. More data processing power, as scale of physics simulation grows
      2. Higher-level programming languages and models
   2. For Data Analytics:
      1. More speed, as data grow larger and algorithms more complex

HPC & Data Analytics
- Case Study : Atmospheric Modelling

Data Ingestion & Preprocessing →

Alternating Physics Simulation & Data Analytics →

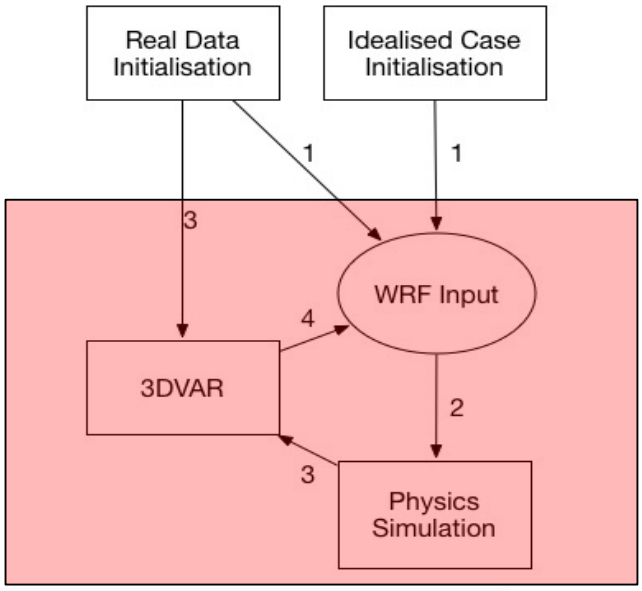


Figure 1: Schematics of WRF

HPC & Data Analytics
- Case Study : Atmospheric Modelling

Data Assimilation - minimization of the cost function:

$$J(x) = J_b + J_o$$
$$= \frac{1}{2}(x - x^b)B^{-1}(x - x^b)$$
$$+ \frac{1}{2}(y - y^o)(E + F)^{-1}(y - y^o)$$

Table of Content

WARWICK

HPC & Data Analytics - Divergence : Software Stack

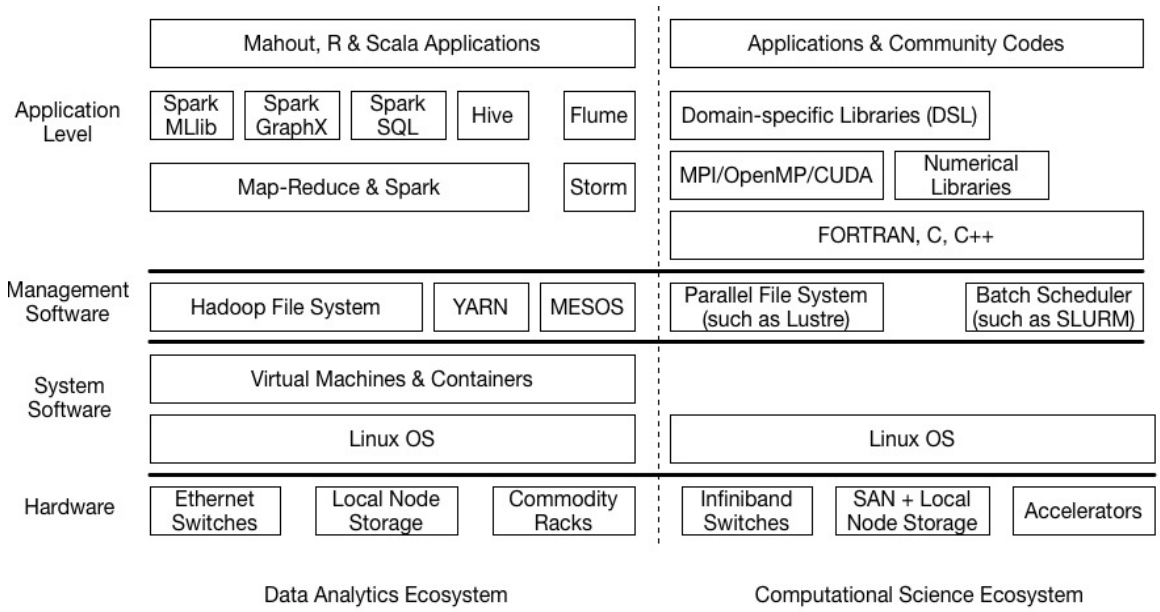


Figure 2: HPC & Data Analytics Ecosystem

HPC & Data Analytics - Divergence : Hardware

Table 1:List of Commercial HPC & DA Systems in 2016

Name	CPU	Accelerator	Interconnect	Storage	Application
Cray XC	Intel Xeon E5	Intel Xeon Phi	Aries	Network Attached	HPC
Cray CS	Intel Xeon E5	Intel Xeon Phi, NVidia Tesla	FDR, Infiniband	Network Attached	HPC
Cray Urika-GX	Intel Xeon E5	None	Aries	HDD, SSD	Data Analytics
IBM Power S822	Power8	None		HDD, SSD	Data Analytics
IBM Power S822LC	Power8	FPGA, NVidia GPU		HDD,SSD	High Performance Data Analytics
IBM Power S824L	Power8	NVidia GPU		HDD,SSD	High Performance Data Analytics

Table of Content

WARWICK

HPC & Data Analytics - Unification

Problem Statement:

How to build a system that is capable of running both HPC and Data Analytic jobs?

HPC & Data Analytics - Unification

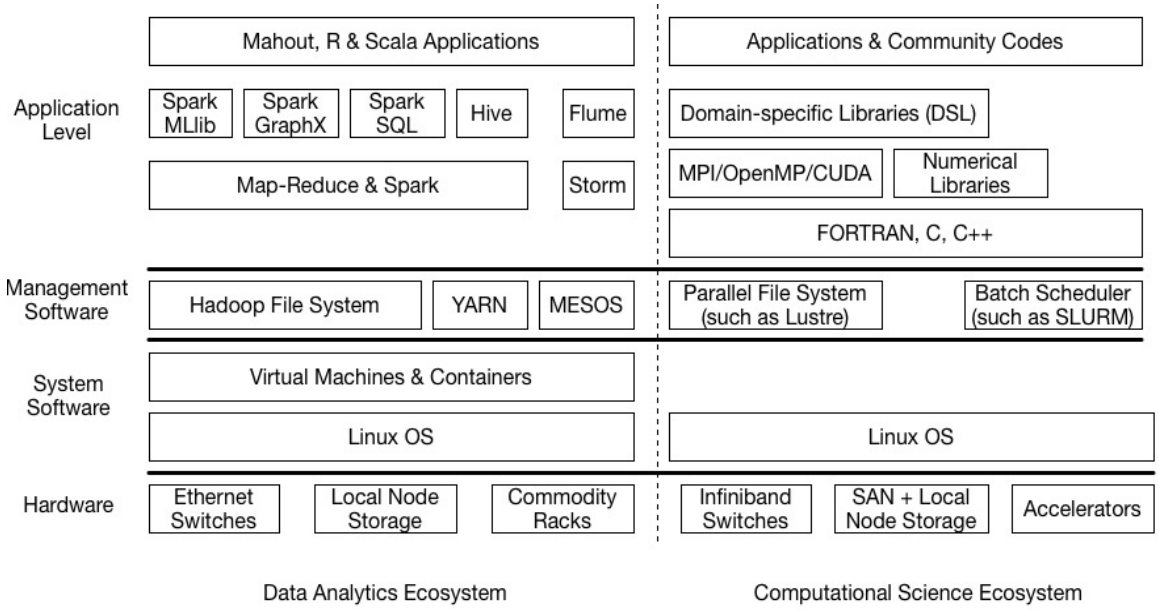


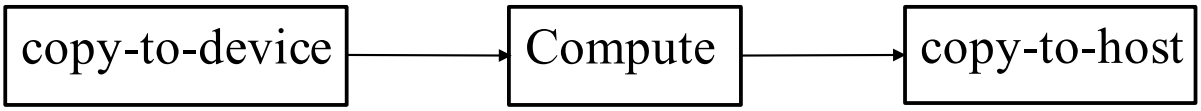
Figure 2: HPC & Data Analytics Ecosystem

Table of Content

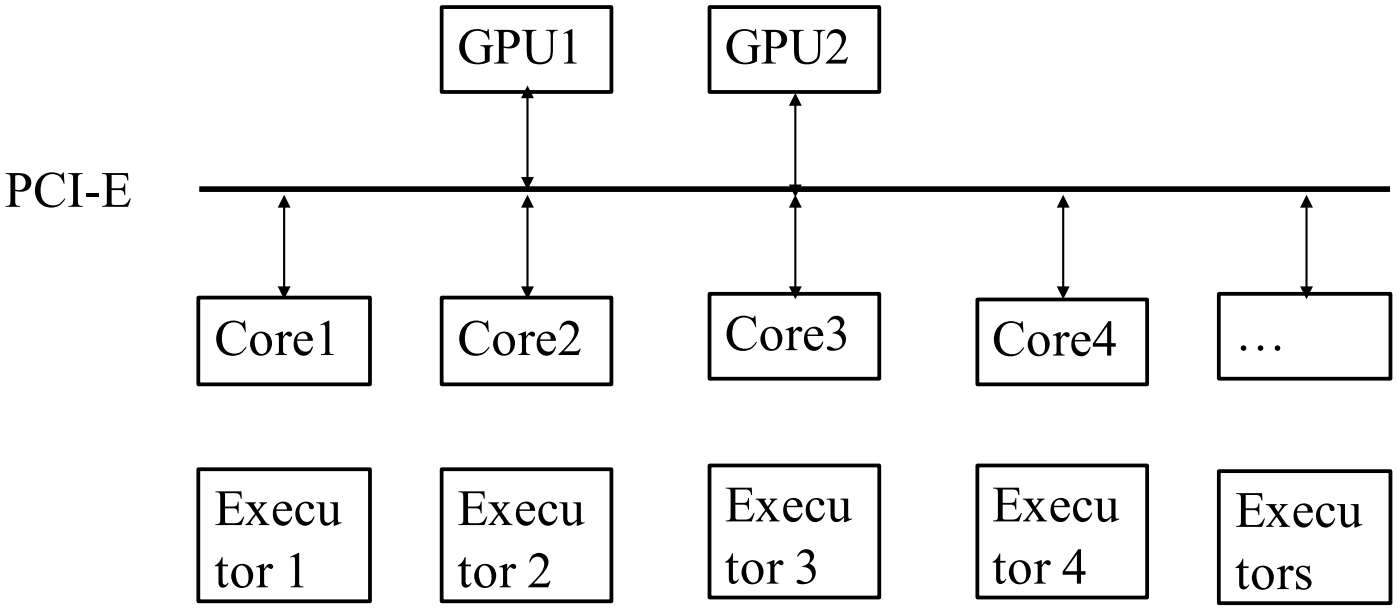
WARWICK

Progress –
Stage 1: Acceleration of Map-Reduce

Common Pattern for Accelerator Computation



Progress – Stage 1: Acceleration of Map-Reduce



Progress –
Stage 1: Acceleration of Map-Reduce

Challenges:

1. Interface gap between Java/Scala/R/Python code and accelerator code
2. Efficient Concurrent Execution of Map-Reduce Tasks
   i. Non-exclusive access to device
   ii. Handling out-of-memory conditions
   iii. Handling persistent data
   iv. Transparent Swap-in Swap-out
   v. Co-scheduling GPU tasks

Progress –
Stage 1: Acceleration of Map-Reduce

Middleware Solution:

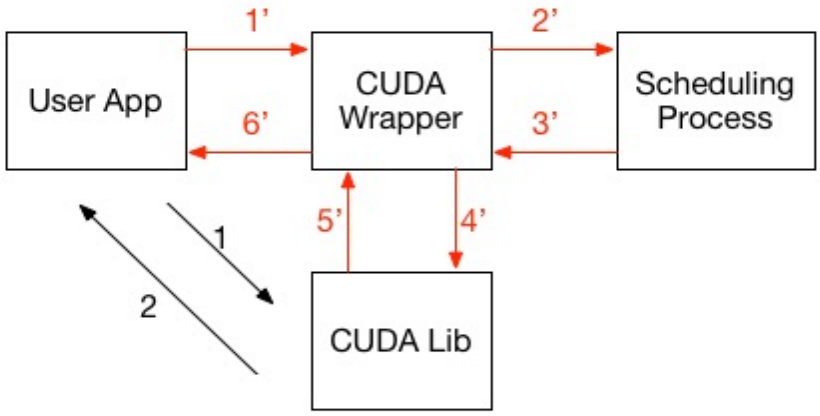


Figure 3: CUDA Middleware Design

Progress –

Stage 1: Acceleration of Map-Reduce

Creating Proxy Apps (GPU + Spark):

1. Machine Learning Algorithms:
   1. Linear Regression (Stochastic Gradient Descent)
   2. K-Means
   3. Random Forest
   4. Matrix Factorization
2. HPC Apps on Spark:
   1. Distributed Matrix Arithmetic (Matrix-Matrix, Matrix-Vector, Vector-Scalar Multiplications)
   2. Optimisation Algorithms: Gradient Descent, Conjugate Gradient, etc.

Table of Content

WARWICK

Next Stage – Further Unification

1. Workload Manager:
    1. HPC: SLURM, PBS
    2. Data Analytics: YARN, MESOS
    3. Enable non-exclusive access to accelerator resources
2. File System:
    1. HPC: Lustre (Network Connected)
    2. Data Analytics: HDFS (Local Nodes)

Q & A

Bibliography

1. Daniel A Reed J. Exascale Computing and Big Data [Internet]. Cacm.acm.org. 2016 [cited 20 June 2016]. Available from: http://cacm.acm.org/magazines/2015/7/188732-exascale-computing-and-big-data
2. Data Assimilation | ECMWF [Internet]. Ecmwf.int. 2016 [cited 20 June 2016]. Available from: http://www.ecmwf.int/en/research/data-assimilation
3. Michalakes J, Dudhia J, Gill D, Henderson T, Klemp J, Skamarock W, Wang W. The weather research and forecast model: software architecture and performance. InProceedings of the 11th ECMWF workshop on the use of high performance computing in meteorology 2005 Sep 20 (pp. 156-168).
4. Barker D, Huang XY, Liu Z, Auligné T, Zhang X, Rugg S, Ajjaji R, Bourgeois A, Bray J, Chen Y, Demirtas M. The weather research and forecasting model's community variational/ensemble data assimilation system: WRFDA. Bulletin of the American Meteorological Society. 2012 Jun;93(6):831-43.
5. Barker DM, Huang W, Guo YR, Bourgeois AJ, Xiao QN. A three-dimensional variational data assimilation system for MM5: Implementation and initial results. Monthly Weather Review. 2004 Apr;132(4):897-914.
6. Huang XY, Xiao Q, Barker DM, Zhang X, Michalakes J, Huang W, Henderson T, Bray J, Chen Y, Ma Z, Dudhia J. Four-dimensional variational data assimilation for WRF: formulation and preliminary results. Monthly Weather Review. 2009 Jan;137(1):299-314.

Full name: Shan Lin

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Person Re-identification

Abstract: People tracking has to face many issues in video surveillance scenarios. One of the most challenging aspect is to re\x96identify people across different cameras. This presentation is about general approaches to solve this problem and discuss some challenges and difficulties. I will discuss some interesting approaches in Features Representations and Metric Learning. In feature representation, one interested method I am focusing on is finding correspondence by salient regions. By finding some small salient regions of people such as vivid colour bag, hat or clothing, person identities can be recognised across the surveillance cameras with different views. In learning method, I tried to cover some commonly used distance learning metric such as KISSME and RankSVM, Last but not least, I will address many good public datasets for evaluating the performance of people re-id models.

Keywords: Person Re-identification, Features Representations, Metric Learning

Abstract Word Count (Est): 133

Person Re-identification

Literature Review

Shan Lin

Department of Computer Science
University of Warwick

WARWICK
THE UNIVERSITY OF WARWICK

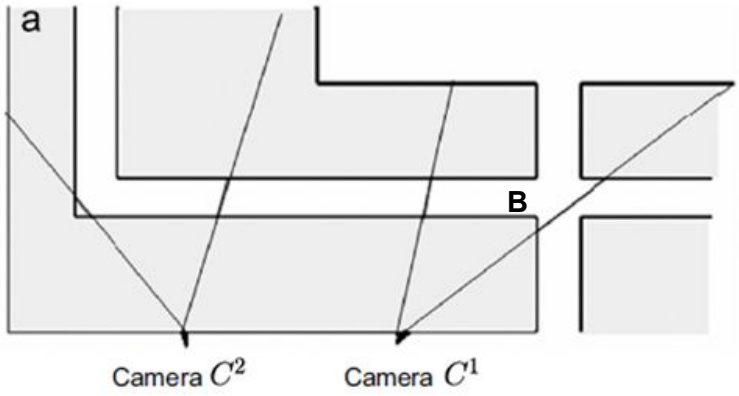
Outline

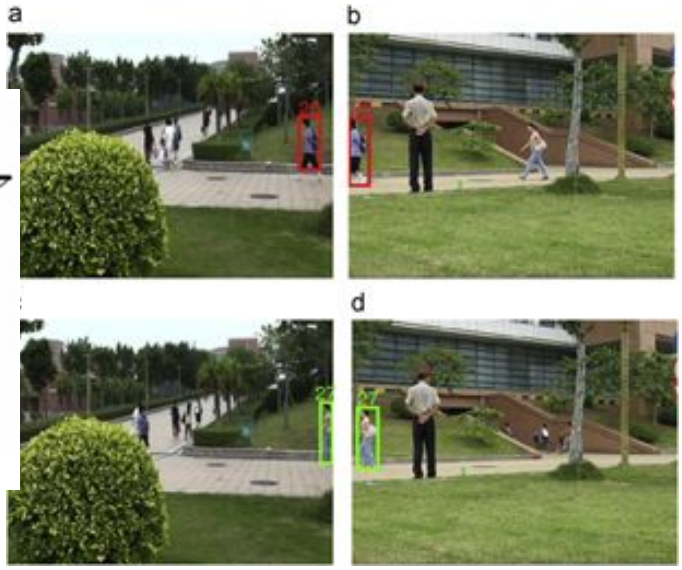
- Introduction
- Challenges
- Feature Representation
- Learning Metrics
- Datasets

Introduction

Person Re-identification Problem:

- In **non-overlapping** camera networks,
- matching the same individuals across **multiple cameras**.





Person Re-ID Challenges

View/Pose Changes

Camera **A** at Time **T**

Camera **B** at Time **T+Δt**

Person Re-ID Challenges

Illumination

Need illumination-invariant features or light-amending process

Person Re-ID Challenges

Occlusions

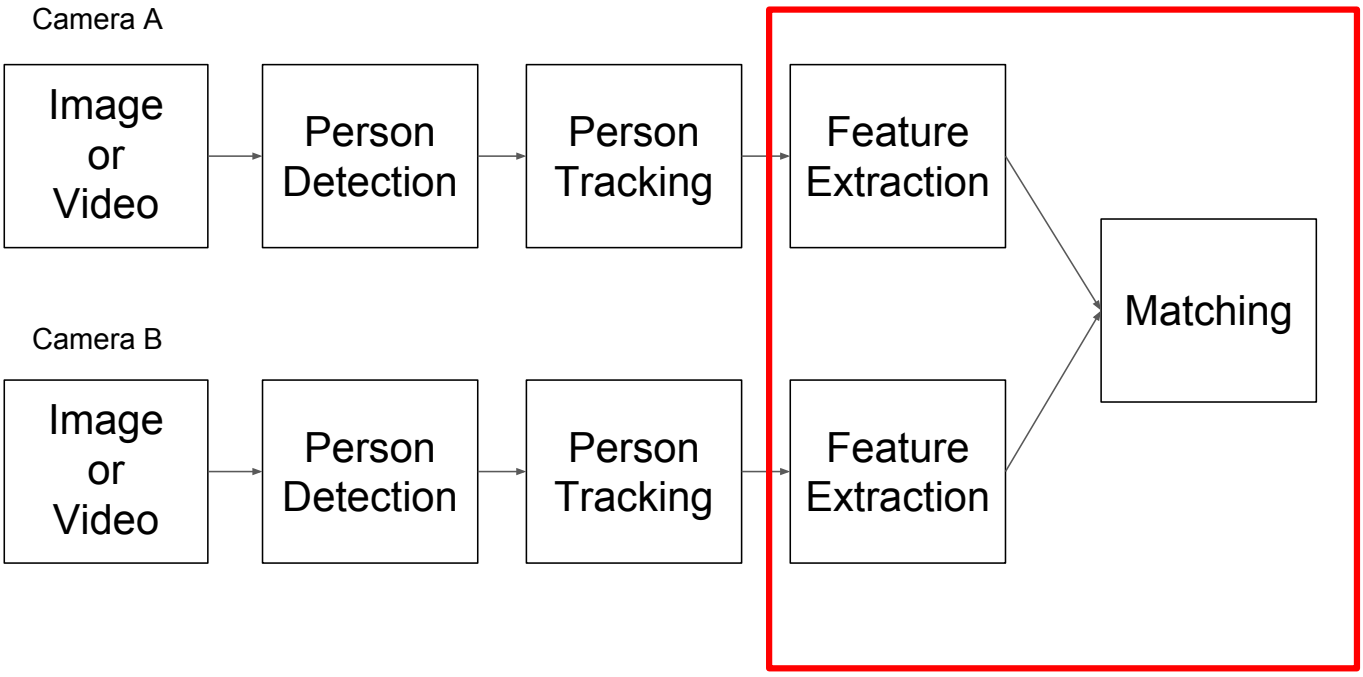
Carried objects occlude the person appearance

People can occlude by external objects

Person Re-ID Challenges

Intra-class Variant and Limited Sample size

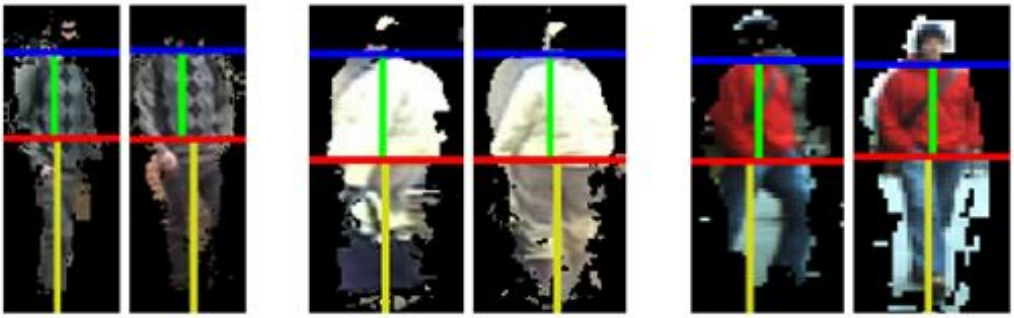
Re-Identification System



Feature Representation by Segmentation

Framework:

1. Segmenting the image into small sections.
    a. Eg. separates regions with strongly different appearance
    b. Separates Head,Body,Legs(Left to Right)
2. Extracting features from each sections.
3. Re-id by part-to-part comparison.



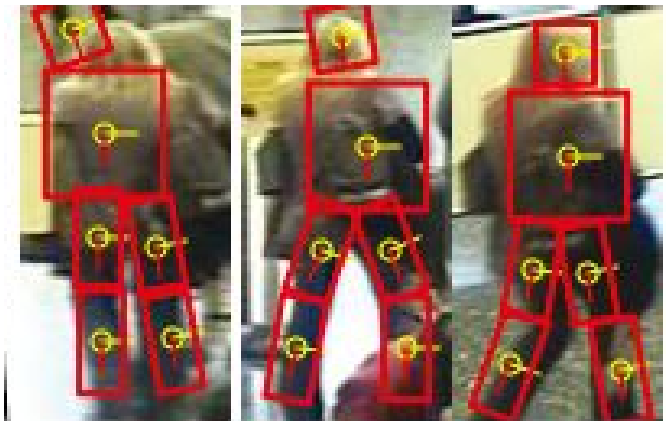
Feature Representation by Detection

Framework:

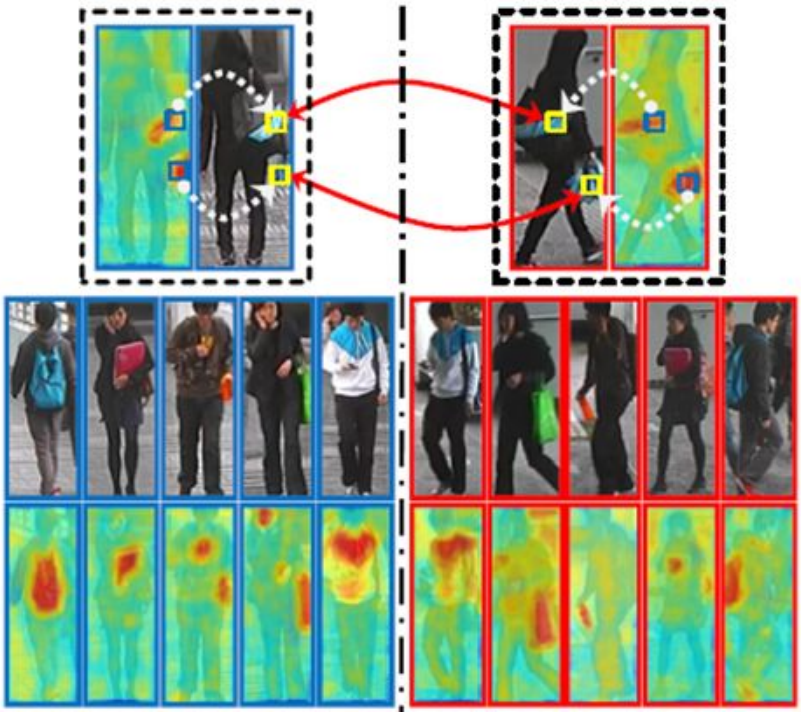
1. Constructing pictorial structures model for every image by detecting different body parts

2. Extracting features from each body part.

3. Re-id by part-to-part comparison.



Feature Representation by Salience



Intuition:

Recognizing person identities based on some small salient region

Salient Region:

Discriminative region which making a person standing out from their companions.

Feature Representation by Salience

Dense Correspondence: widely in face and scene alignment

1. Each human image is densely segmented into a grid of local patches.
2. **Dense Color Histogram:** LAB
3. **Dense SIFT :** complementary feature to color histograms
4. Dense color histograms and dense SIFT features are concatenated as the final multi-dimensional descriptor vector for each patch.

Feature Representation by Salience

Adjacency Constrained Search:
Restricting the search range for each image patch.

$$\hat{\mathcal{S}}(x^{A,p}_{m,n}, \mathbf{x}^{B,q}) = \{\mathcal{T}^{B,q}(b) | b \in \mathcal{N}(m)\}, \quad \forall x^{A,p}_{m,n} \in \mathcal{T}^{A,p}(m),$$

$$\mathcal{N}(m) = \{m-l, ..., m, ...m+l\}$$

Similarity Score:
Euclidean distance between patch feature x and y.

Feature Representation by Salience

salient patches are those possess uniqueness property. They distributed far away from other patches.

$$score_{knn}(x^{A,p}_{m,n}) = D_k(X_{nn}(x^{A,p}_{m,n})),$$ **e:**

Dk(): the distance between the patch and its k-th nearest neighbor.

The salient patches can only find limited number of visually similar neighbors.

Feature Representation by Salience

Bi-directional weighting for patch matching:

Patches in red boxes are matched in dense correspondence with the guidance of corresponding salience scores in dark blue boxes.

Distance Metric Learning

Distance metric learning:

- shift the focus from feature selection
- to improve re-id accuracy.
- Learning appropriate distance metrics
  - Keep feature from the same class closer
  - Different class farther apart
  - Ultimately, maximizes the interclass distance Or minimizes the intraclass distance

Distance Metric Learning

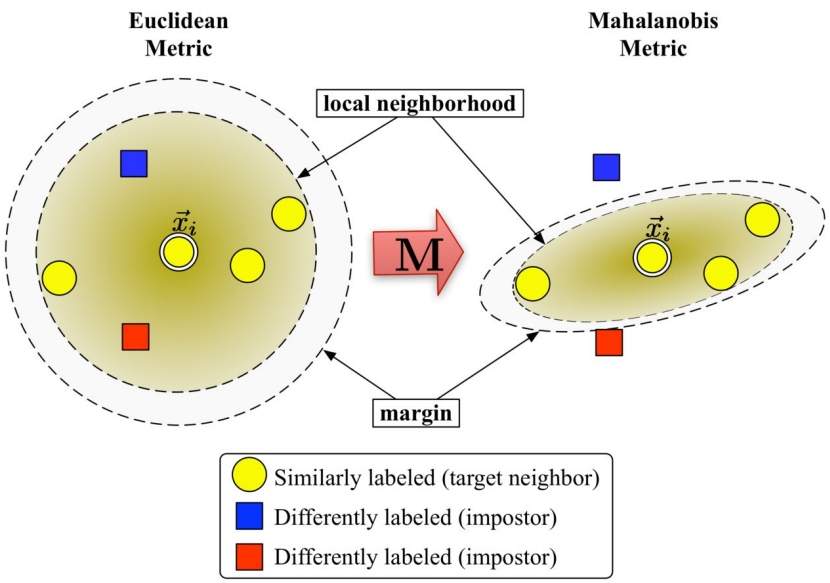
RankSVM:

- Treated as a relative ranking problem
- Learn a relative ranking to the probe image
- Set of weak SVM based ranker of different features
- Combine to build a stronger ranker by using ensemble learning
- No False Match Rejection

Distance Metric Learning

LMNN:

- Large margin nearest neighbor
- False Match Rejection



Datasets

VIPeR

- 632 person, 2 images per person.
- Occlusions.
- Pose/View changes, shot by two cameras.
- Heavy illumination changes.

Datasets

GRID

- 250 Probe Images
- 250 True Match.
- Additional 775 picture.

- Occlusions.
- Pose/View changes
- illumination changes.
- **Scale Variations**

Datasets

CUHK Dataset

CUHK01 Dataset

- 3884 images
- 972 pedestrians

- Occlusions.
- Pose/View changes, shot by two cameras.
- illumination changes.

Datasets

CUHK Dataset

CUHK03 Dataset

- 13,164 images
- 1,360 pedestrians

- Occlusions.
- Pose/View changes, shot by two cameras.
- illumination changes.

Datasets

CUHK Dataset

CUHK02 Dataset

- Extended version of CUHK01 by adding 3 different camera view setting

- Occlusions.
- Pose/View changes, shot by five cameras.
- illumination changes.

Used for evaluating re-id algorithms under *different camera view transforms.*

Thank You

Q and A

Full name: James Marchant

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: On Dynamic Networks and Convention Emergence

Abstract: Within multiagent systems (MAS) coordination and cooperation amongst agents is of paramount importance. Due to the nature of these systems it is often difficult or impractical to define a cohesive and complete set of rules on expected agent behaviour in any given situation. The idea of convention emergence, that of letting the population of agents learn, amongst themselves, the best and consistent course of action to take in a certain situation, alleviates many of these concerns. Encouraging timely emergence of these conventions and perhaps influencing the direction that they should take whilst still maintaining minimal top-down control is an active area of research. This work explores the nature of this convention emergence, how to influence it and how to destabilise existing conventions within dynamic, ever-changing networks.

Keywords: Conventions, Multi-Agent Systems, Dynamic Networks

Abstract Word Count (Est): 126

On Dynamic Networks and Convention Emergence

James Marchant (james@dcs.warwick.ac.uk)

Department of COMPUTER SCIENCE

WARWICK

Introduction

- Coordinated actions reduce costs and increase efficiency.
- *Conventions* allow coordinated actions to emerge.
- Agents are often situated within a *topology* and interactions limited to neighbours.
- *Fixed Strategy* (FS) agents are often used to encourage and direct convention emergence.
- We explore the effect of *dynamic topologies* on convention emergence and utilise a new FS placement heuristic, LIFE-DEGREE, to examine particular network features.

WARWICK

Background: Conventions

- Conventions are socially-accepted rules or standards, exhibited as strategy choice.
- *"An equilibrium everyone expects in interactions that have more than one equilibrium".*
- Able to facilitate *coordination* with few assumptions about agent architecture.
- Differ from *norms*; no punishment.
- Online *emergence* requires little computational overhead and deals with issues faced in offline generation (lack of knowledge, complexity).

WARWICK

Background: Fixed Strategy Agents

- Agents who utilise same strategy regardless of others' actions.
- FS agents shown to influence much larger populations.
  - Sen *et al.* show 5 agents influencing 3,000.
- Placing at influential locations increases effect.
  - Placement by metrics offers improvements over random placement.

WARWICK

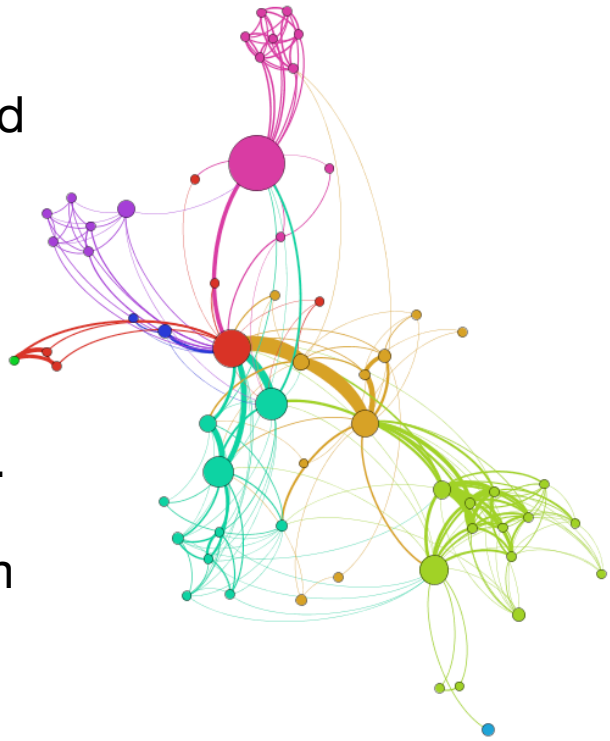
Convention Emergence Model

- We adopt the general approach from the literature.
- Agents are situated in a topology, interactions limited to neighbours.
- Receive payoff from *n-action coordination game*.
  - Previous work often suffers from small convention spaces.
- Utility-based strategy selection; *simplified Q-Learning* based on payoff used to inform agent decisions.
- Agents are also able to *explore* the action space randomly.

WARWICK

Dynamic Topologies

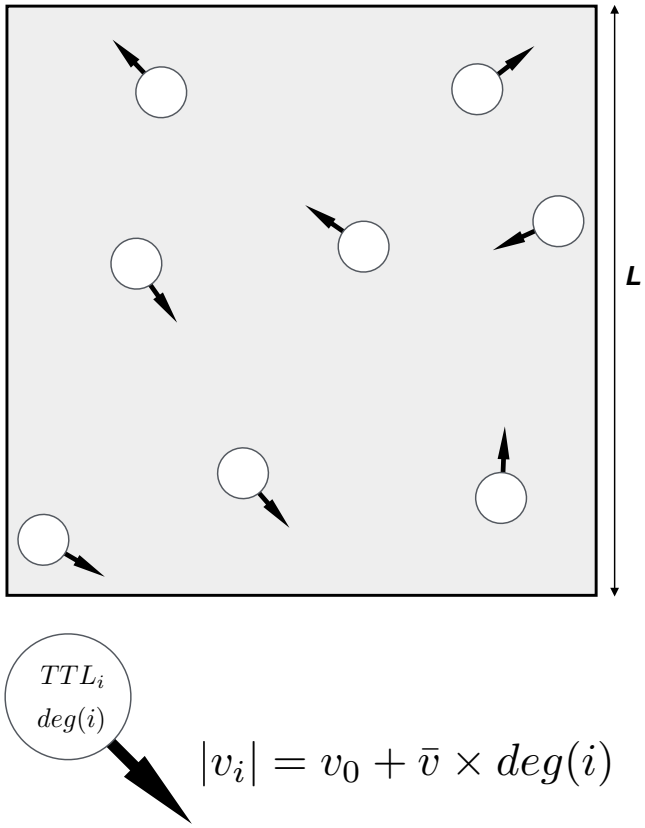
- Most previous work has focussed on static topologies.
- We explore convention emergence within a dynamic setting using a particle-based model created by González et al.
- Savarimuthu et al. explored norm emergence on this model.



WARWICK

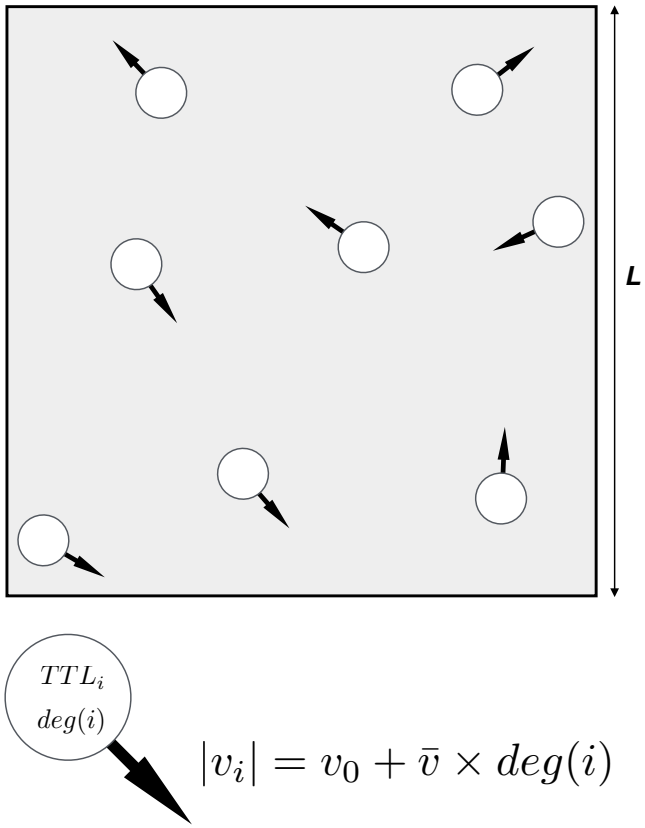
González Model

- Builds a dynamic topology using N particles located in a 2D box of sides L to represent the agents.
- Particles have a speed proportional to their degree.
- Collisions create an edge between the agents.
- Exhibits *preferential attachment*.



González Model

- Particles also have a Time-To-Live (TTL).
- Chosen uniformly at random from [0, $\mathcal{T}_l$] upon creation.
- When TTL = 0, agent and all its edges are removed.
- A new agent/particle is created at the exact same location.
  - Population remains static, edges don't.



WARWICK

Placement Heuristic

- FS agents are placed using a number of heuristics comparable to those in static topologies.
- Degree-based placement differs; changing degrees introduces two approaches:
  - Static Degree: FS agents only chosen once/upon expiration. Ignores changed to agent degree.
  - Updating Degree: FS agents are chosen every timestep. Keeps track of changing degrees but more expensive/FS agents may move around too much.

Placement Heuristic

- We also introduce a new heuristic, LIFE-DEGREE.
- Uses the concept of *agent age* (expected remaining TTL).
  - Unique concept, not in static topologies.
- Allows exploration of the effect of selecting older agents against younger ones.
- Combines age and normalised degree in weighted sum.
- Also performed as Static and Updating.

$$\text{LIFE-DEGREE}(n) = \omega \times deg(n) + (1 - \omega) \times E_{rTTL}(n)$$

Experimental Setup

- Experimental population set to 1,000 agents.
- Q-Learning learning and exploration rates set to 0.25.
- The payoff is calculated using the 10-action coordination game.
  - Agents have 10 possible choices, positive reward if they choose the same, negative otherwise.
- We use 100 runs per simulation and the aggregate results are presented here.

WARWICK

Initial Intervention

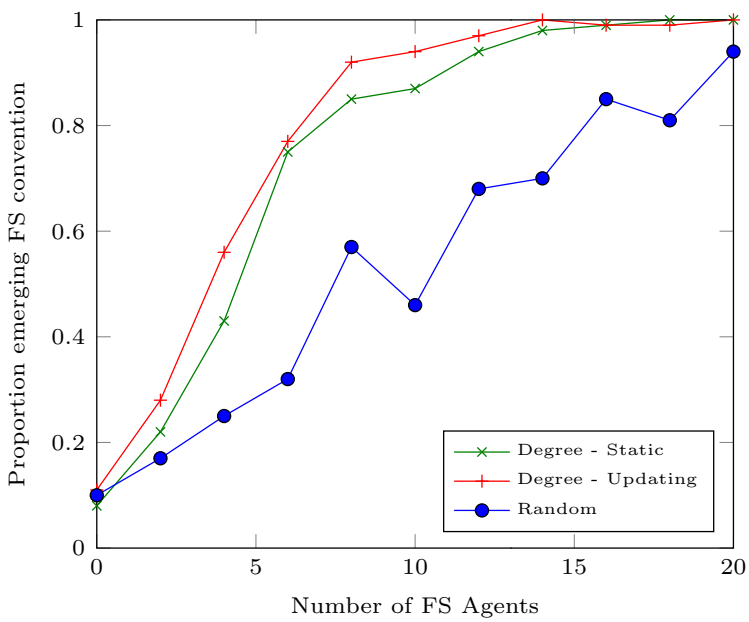


Fig 2: Traditional Heuristics

- With no FS agents expected proportion given action space (10 actions).
- Attempted to influence with traditional heuristics.
  - Number of FS agents needed to guarantee change higher than in static networks.
- Little difference between Static and Updating.

WARWICK

Initial Intervention: Static

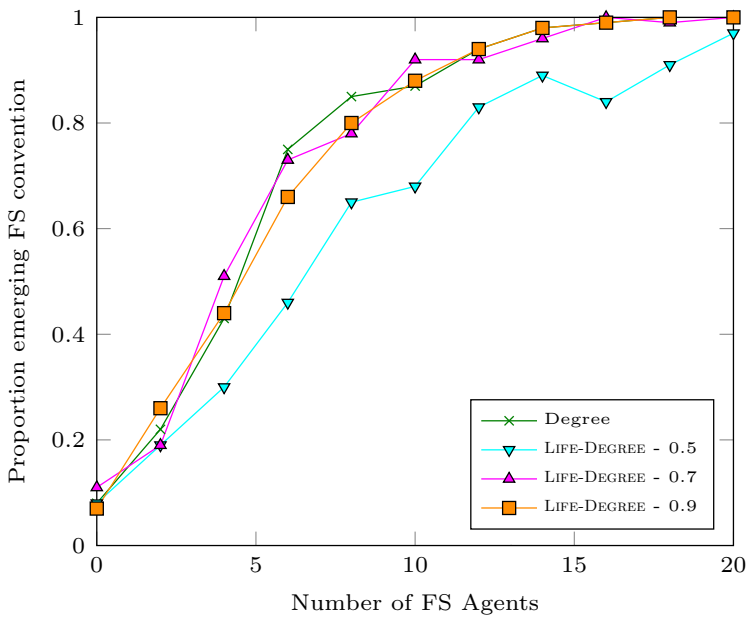


Fig 3: Static Life-Degree

- When using Static Life-Degree, equal weighting between degree and life results in substantially worse performance.
- Regardless of weighting, doesn't outperform pure degree.
- Indicates that efficacy much more dependent on an agent's ability to influence others, regardless of lifetime.

WARWICK

Initial Intervention: Updating

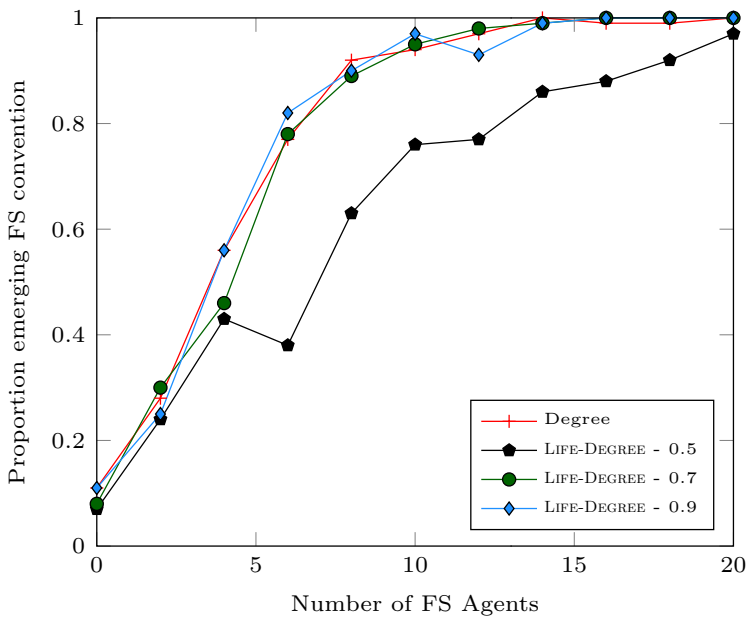


Fig 4: Updating Life-Degree

- Updating Life-Degree exhibits similar performance to Static.
- Life-Degree never meaningfully exceeds the pure degree heuristic, regardless of weighting.
- There is little difference between Static and Updating performance indicating that initial intervention doesn't rely on up-to-date degree information.

WARWICK

Late Intervention: Static

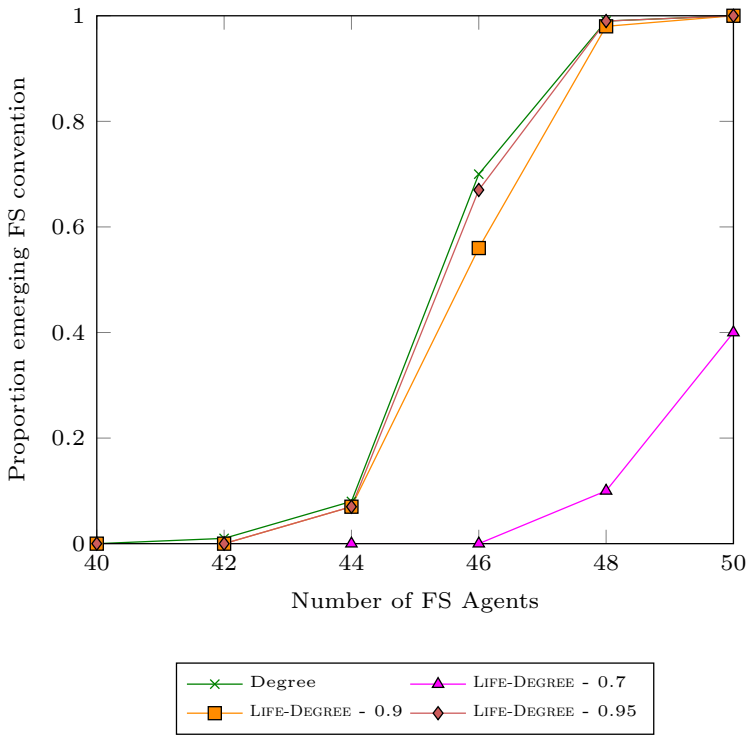


Fig 5: Static Life-Degree, Late Intervention

- FS agents can also be used to *destabilise* existing conventions.
- We allow a convention to emerge naturally (by waiting 5000 timesteps) then introduce FS agents using a different action from the convention.
- Random performance v. poor so not included.

Late Intervention: Static

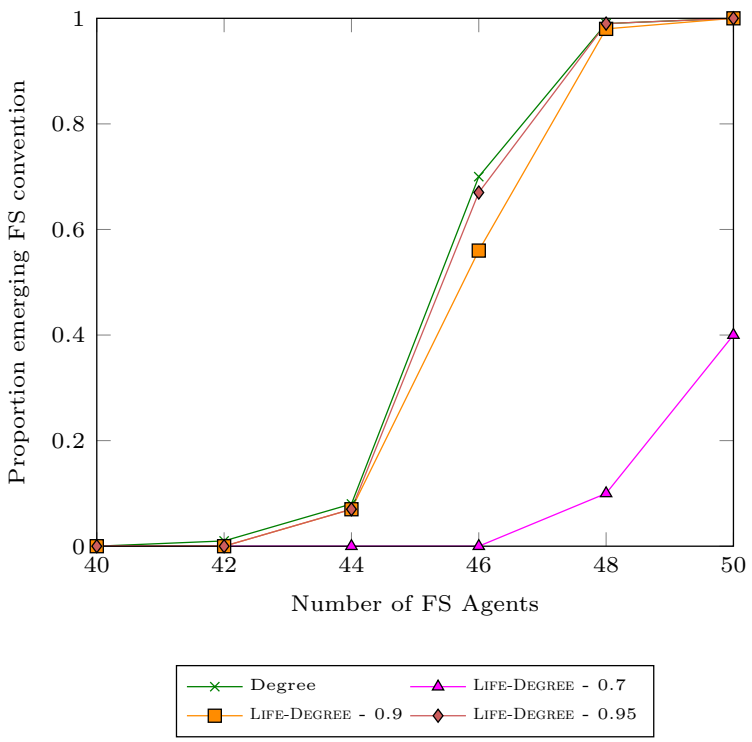


Fig 5: Static Life-Degree, Late Intervention

- Much larger number of FS agents needed to have effect
  - Mirrors findings on static topologies.
- Transition from “no-effect” to “guaranteed change” occurs over small range; indicates a critical point.
- Similar to initial intervention, pure degree performs best.
  - Higher weighting needed to reach similar behaviour; late intervention more sensitive to diluting degree heuristic.

Late Intervention: Updating

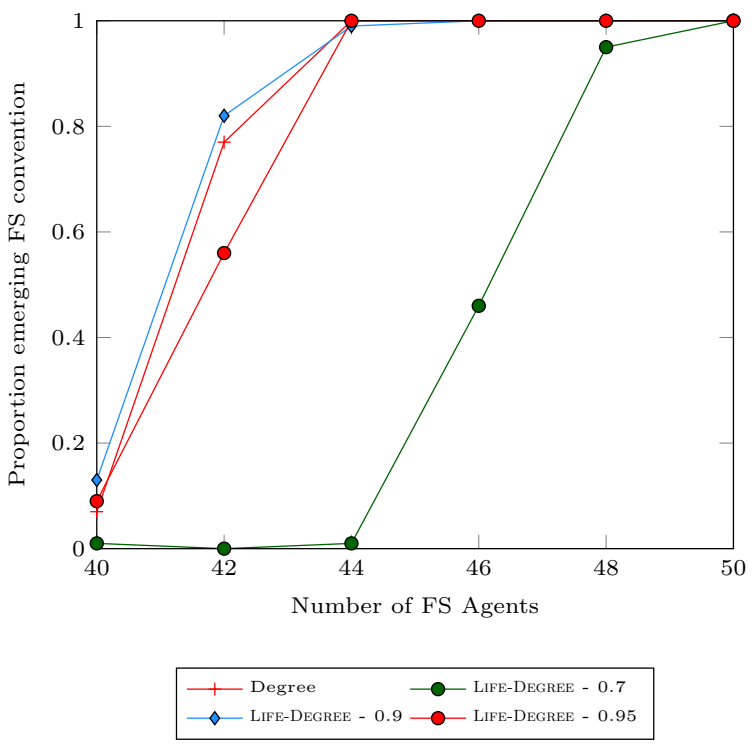


Fig 6: Updating LIFE-DEGREE, Late Intervention

- Updating heuristics perform substantially better than their static equivalents.
  - Effect not seen in initial intervention.
  - Destabilisation benefits from up-to-date degree placement to maximise the effect of the agents.
- LIFE-DEGREE performance still worse; agent age can safely be ignored with no detrimental effects in all scenarios.

WARWICK

Additional Dynamic Topologies: RASH

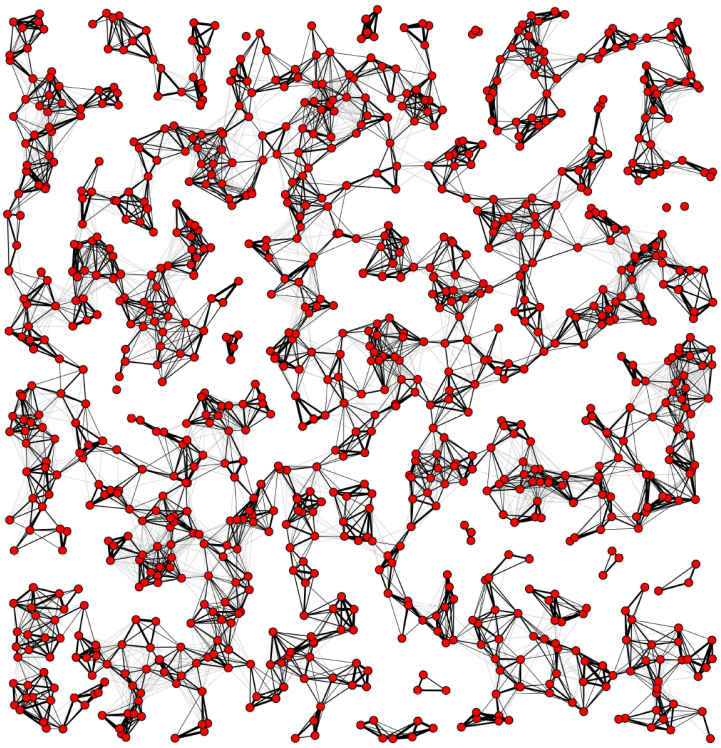
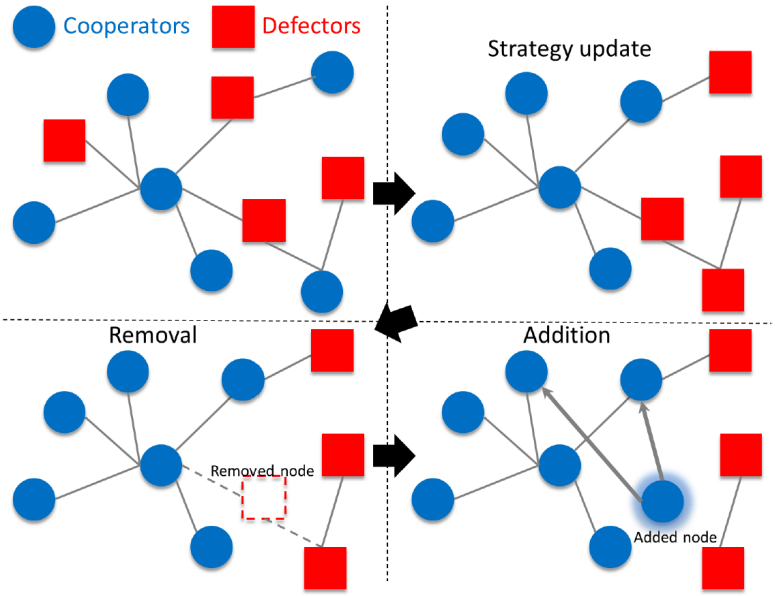


Fig 7: RASH Network

- RASH by zu Erbach-Schoenberg et al.
- Mimics gatherings with acquaintances
- Shared histories
- Localisation of nodes in 2D grid

WARWICK

Additional Dynamic Topologies: Ichinose



- Dynamic topology by Ichinose et al.
- Creates a Scalefree graph using Barasi-Albert generation.
- Rewires the graph using four different techniques: RR, RP, TR, TP

Fig 7: Ichinose Network

WARWICK

Summary

- We explored the nature of convention emergence in dynamic topologies, showing that it occurs naturally as in static topologies.
- We showed there is a minimum level of connectedness needed for convention emergence to occur.
- We found that consideration of agent longevity offers no improvement to FS agent placement; agent degree (a measure of their influence) is the most important factor.
- We found that late intervention/destabilisation benefits from having up-to-date information on agent degrees, something that initial interventions don't exhibit.

WARWICK

Questions?

WARWICK

On Dynamic Networks and Convention Emergence

James Marchant (james@dcs.warwick.ac.uk)

Department of COMPUTER SCIENCE

WARWICK

Full name: Helen McKay

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Personalisation Through Transfer Learning with Concept Drift

Abstract: Bespoke applications provide personalised functionalities to user but can be expensive to produce. To reduce costs, off-the-shelf applications utilise manual input to tailor functionalities to user\x92s personal preferences. Manual inputs must be frequently updated and maintained to ensure they effectively represent the user\x92s desires, for example, smart heating systems often require users to provide a heating schedule defining their desired heating times and temperatures with respect to behavioural patterns. Relying on users to frequently modify manual inputs can become a burden on users as inputs can become redundant over time. To overcome this dependency, machine learning techniques can be utilised to predict personalised preferences from sensing the surrounding environment. Online learning methodologies can be used within real-world environments to make predictions in real-time using data streams, preventing the need to store large amounts of historical data. User preferences and real world environments are dynamic in nature, causing concept drifts that require learning algorithms to update or change their prediction models to remain effective. To build effective models, learning algorithms require initial training data, preventing useful predictions from initially being made for new users, this is known as the cold start problem. Transfer learning can be utilised to overcome the cold start problem by transferring knowledge from existing users to new users, reducing the time taken to achieve an effective model. Transfer learning can also be used to provide recommendations of beneficial personalisations users were unaware of.

Keywords: Online learning, concept drift, transfer learning

Abstract Word Count (Est): 236

Personalisation through Transfer Learning with Concept Drift

Helen Mckay

Project Overview

- Personalise application functionalities to suit user behavioural patterns
- Three components:
  - Online machine learning
  - Concept drifting data streams
  - Transfer learning
- Simulation environment: Smart Home Heating System

Smart Home Heating

- Aim: Regulate household temperature.
- Solution 1: Bespoke system
  - Not reproducible or cost effective
  - Have to foresee the future
- Solution 2: Heating schedules
  - Dependant on user inputs

Smart Home Heating

- Using machine learning
  - Sense surrounding environment
  - Model behavioural patterns
  - Predict desired personalised functionalities
- Online learning, processing data streams
- Behavioural patterns are subject to change over time

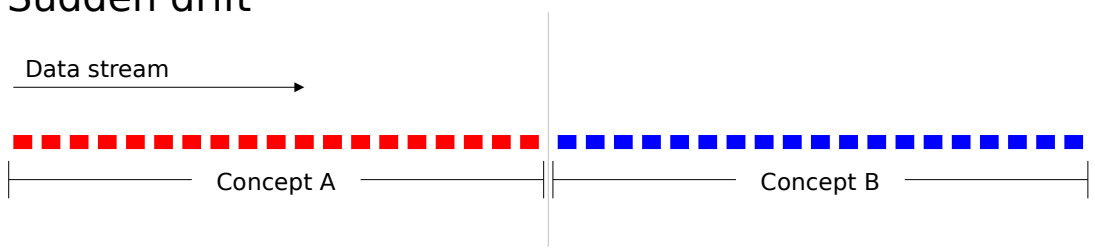
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
EPSRC Engineering and Physical Sciences Research Council
JAGUAR LAND ROVER

Online learning and Concept Drift

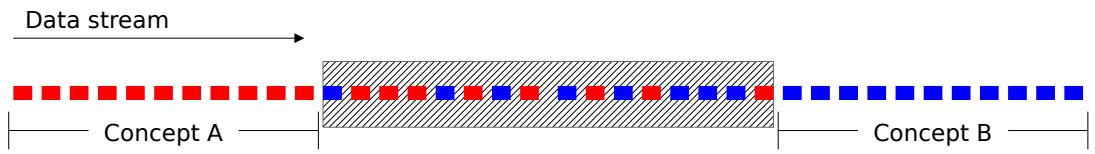
- Modelling in real-world environments
  - Retain a short history of instances and cannot foresee future instances
- Real-world environments and human desires are dynamic
  - Changes in the target concept, induced by uncertainty about the future is known as concept drift
- Concept Drift
  - Real drift – change in preferences (target concept)
  - Virtual drift – change in distribution of instances

Speed of drift

- Sudden drift



- Gradual drift



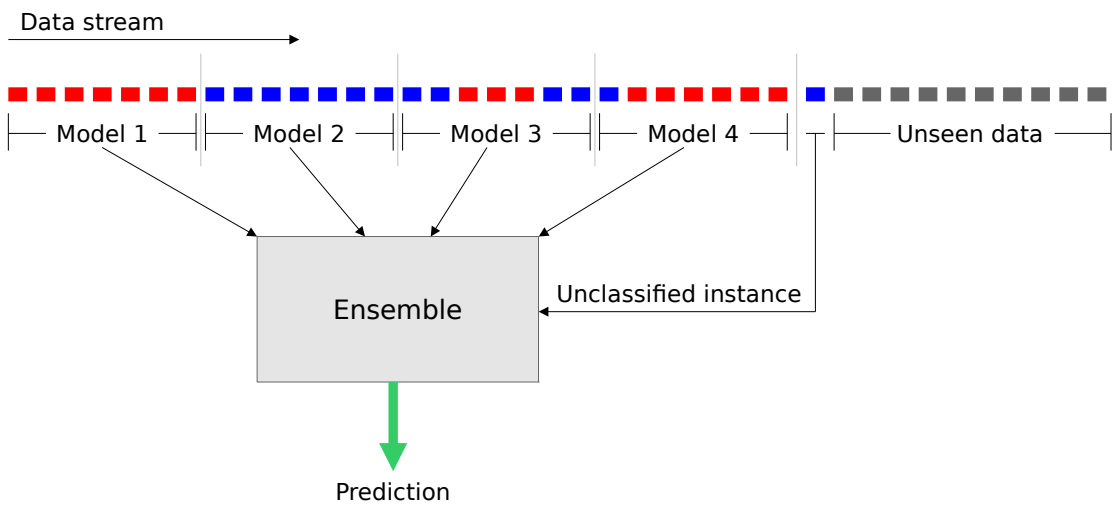
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
EPSRC Engineering and Physical Sciences Research Council
JAGUAR LAND ROVER

Drift Detection Strategies

- Continually update model with newly seen instances
- Ensembles
  - Weight models depending on their age
  - Partition dataset to train each model in the ensemble
- Sliding window and trigger points
  - Identifies the precise location a concept changes
  - Use a window to capture a small subset of the most recent instances
  - A drift is detected if the accuracy of the window drops below a threshold
  - Learn a new model based on the instances within the window

Drift Detection Strategies

- Ensembles

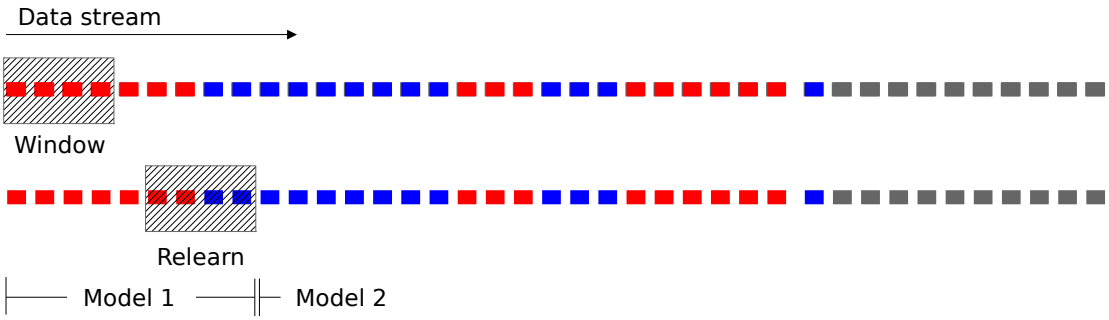


Drift Detection Strategies

- Continually update model with newly seen instances
- Ensembles
  - Weight models depending on their age
  - Partition dataset to train each model in the ensemble
- Sliding window and trigger points
  - Identifies the precise location a concept changes
  - Use a window to capture a small subset of the most recent instances
  - A drift is detected if the accuracy of the window drops below a threshold
  - Learn a new model based on the instances within the window

Drift Detection Strategies

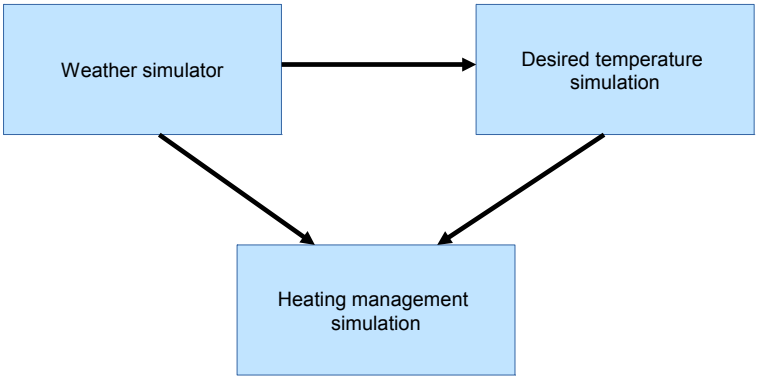
- Sliding window



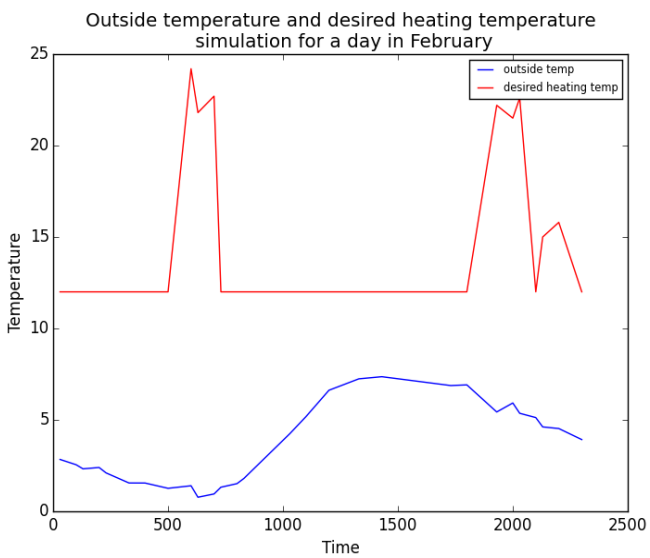
  - Window size
  - Error threshold

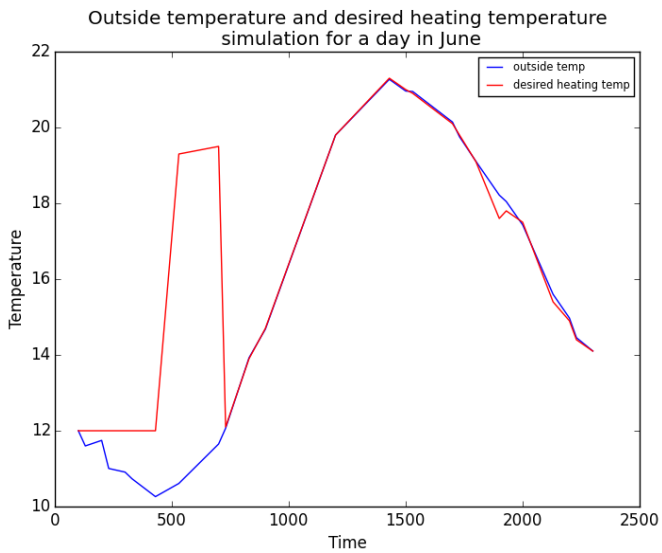
Smart Home Heating Simulation

- Components:

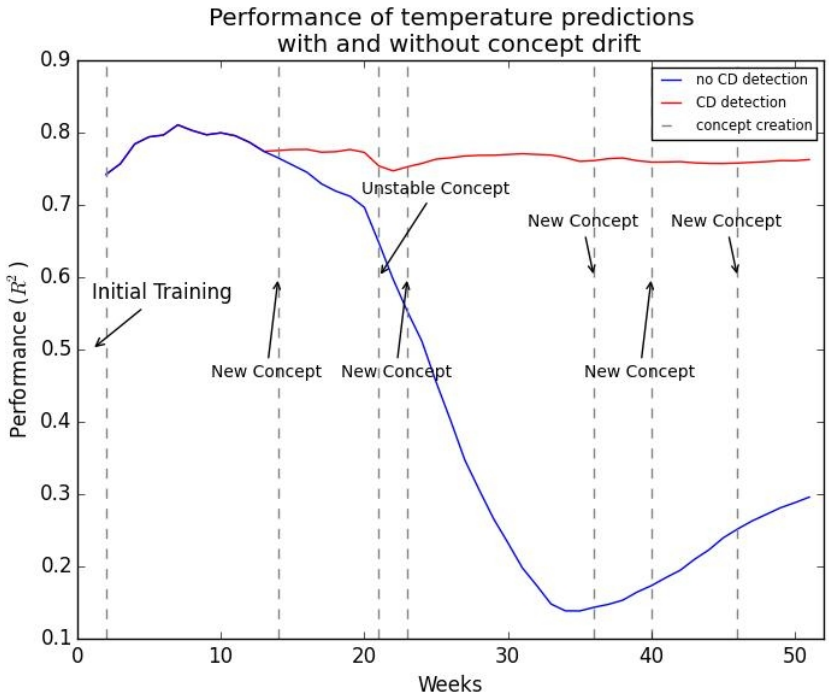


Simulations and Concept Drift





WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
EPSRC Engineering and Physical Sciences Research Council
JAGUAR LAND ROVER



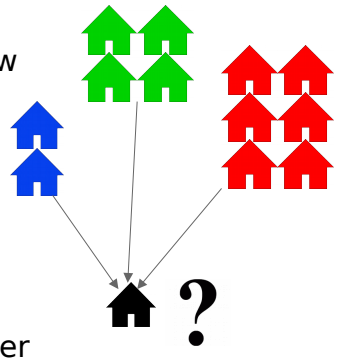
The cold start problem

- Machine learning algorithms need training data to create effective models
- The cold start problem occurs when insufficient information is available to build an effective model
- Typically this occurs for the duration of 2 weeks, whereby the application cannot make effective decisions or predictions

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
EPSRC Engineering and Physical Sciences Research Council
JAGUAR LAND ROVER

Transfer learning

- Transfer learning can be used to address the cold start problem
  - Use knowledge obtained from existing users and transfer to a new user
- Need to be able to identify similar users
- Other uses of TL:
  - Proactive concept drift detection
  - Provide recommendations of personalisable functionalities the user was unaware of



Questions?

WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
EPSRC Engineering and Physical Sciences Research Council
JAGUAR LAND ROVER

Full name: Olfat Mirza

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Style Analysis for Source Code Plagiarism Detection

Abstract: Plagiarism has become an increasing problem in higher education in recent years. A number of research papers have discussed the problem of plagiarism in terms of text and source code and the techniques to detect it in various contexts. There is a variety of easy ways of copying others\x92 work because the source code can be obtained from online source code banks and textbooks, which makes plagiarism easy for students. Source code plagiarism has a very specific definition, and Parker and Hamblen define plagia-rism on software as \x93A program that has been produced from another program with a small number of routine transformations\x94. The transformations can range from very simple changes to very difficult ones, which can be one of the six levels of program modifications that are given by Faidhi and Robinson. Cod-ing style is a way to detect source code plagiarism because it relates to programmer personality without affecting the logic of a program, and can be used to differentiate between different code authors.

The immediate objective of this research is to identify a suitable data set (BlackBox) where it can be applicable to apply appropriate algorithm to analyse the Coding Style in that dataset.

Keywords: Source Code Plagiarism Detection, Style Analysis, Coding Style.

Abstract Word Count (Est): 197

Full name: Andrew Owenson

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Proxy-app led evaluation of optimisations of an unstructured CFD code

Abstract: Optimising a large production code to utilise modern processor capabilities can take a considerable amount of time, and the efficacy of a particular optimisation may not be known until evaluation which could lead to much time wasted. One such code is Hydra, an unstructured mesh CFD simulation engine that operates on meshes of edges and vertices, consisting of 50K lines of code. To speed up the evaluation of optimisations, a proxy-app of Hydra has been developed with only 1500 lines of code. This has very similar data-access patterns as that of a core algorithm in Hydra, which allows for the rapid evaluation of potential optimisations to that algorithm.

In my recent work I have used this proxy-app to evaluate changes to the layout of the mesh data with respect to vectorisation - transposition of arrays of multi-dimensional data, aligning memory allocations to cache boundaries, and reordering the placement of points and edges in memory. I also evaluate cache blocking, and parallelisation of the proxy-app within a many-core processor with shared memory, evaluating OpenMP scheduling strategies and different graph colouring schemes to allocate workload to threads.

Keywords: Optimisation, performance, parallellisation, vectorisation, OpenMP, CFD, unstructured

Abstract Word Count (Est): 185

Full name: Vangelis Pitidis

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: The geological dimension of urban resilience: Impact of geo-hazards in resilient urban design and policy

Abstract: The notion of urban resilience has recently emerged as an evolved amalgamation of previous urban development concepts such as sustainability and Disaster Risk Reduction. Building on its application in various scientific disciplines, resilience has been portrayed as a transdisciplinary approach to urban development, based on the idea of continuous adaptation to abrupt change through acknowledging the uncertainty of contemporary cities. It also constitutes as a step forward to current scientific discourse since it attempts to prepare the city for \x91expecting the unexpected\x92 by creating a favourable milieu for local authorities and stakeholders to help the city endure a variety of (un)expected perturbations by remaining functional, evolve through inevitable transformations and finally develop as a consistent unit to a more resilient and sustainable habitat.

Even though the content of urban resilience derives from and applies to a wide spectrum of disciplinary fields, it is widely accepted that geology has been largely neglected from that discourse. Geological conditions rarely prohibit development entirely but instead they can induce regulations for the suitability of land, the urban design and the cost of development. Current urban planning approaches tend to take the ground into account from a hazard mitigation point of view, while the real question should focus on its capacity to accommodate future development, depending on its attributes. Greater proactive engagement with the geological characteristics of urban settings as well as appropriate use of money, technology and expertise is capable of leading to a more successful and integrated response to luring external disturbances.

Keywords: urban resilience, hazard, geology, equilibrium

Abstract Word Count (Est): 249

The geological dimension of urban resilience: Impact of geo-hazards in resilient urban design and policy

Vangelis Pitidis

Supervisors: Prof. Jon Coaffee
Prof. Stephen Jarvis

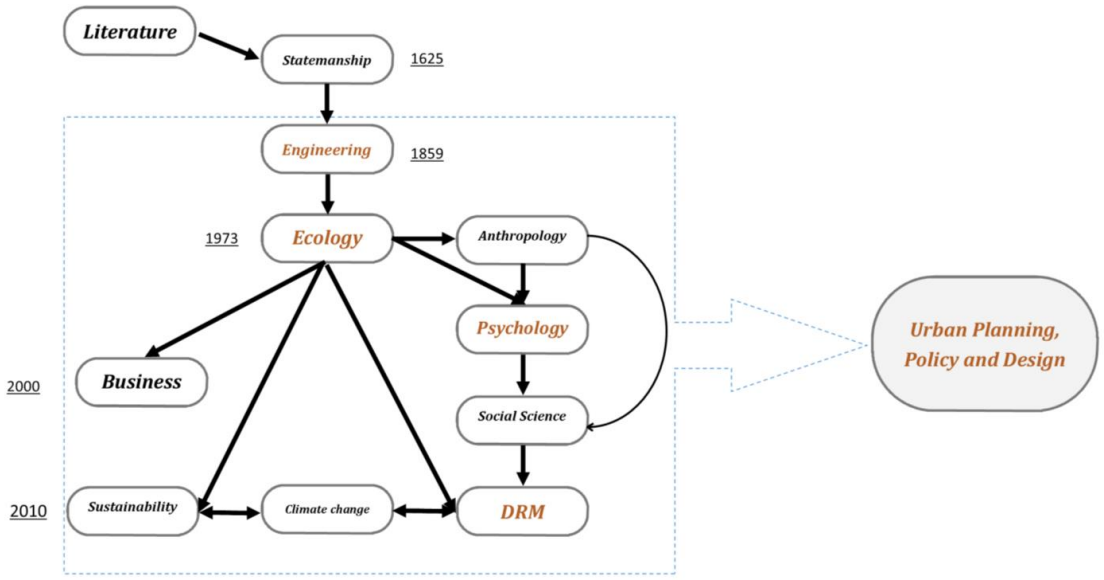
WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Overview

- Definition of resilience
- Journey of the term across time and different disciplines
- Emergence of urban resilience
- Resilient urban planning and design
- Negligence of geology from the urban resilience discourse
- Presentation of a case study
- Conclusions
- Further steps

WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Evolution of resilience as a term across disciplines



Basic contrasting definitions

Engineering	Ecological
Static	Dynamic
Equilibrium	Dis-equilibrium
Stability	Disruption
Expected	Unexpected
Efficiency	Existence
End state	Transition state

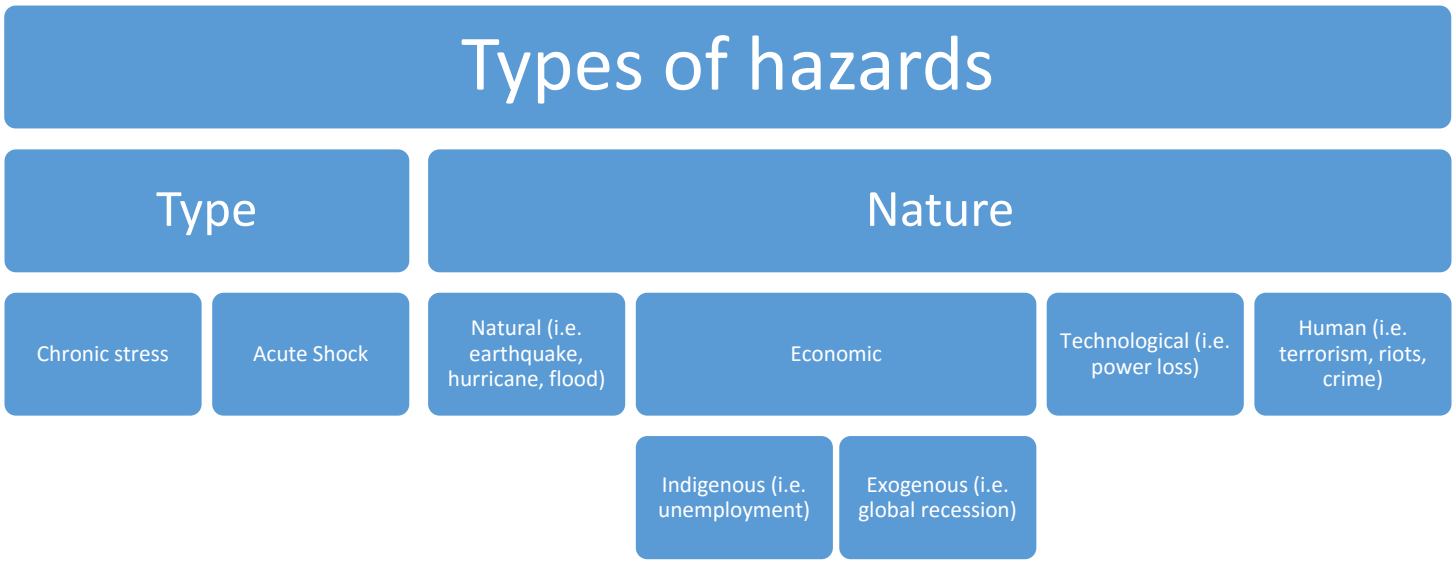
WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

What is resilience?

'Resilience is a measure of the persistence of systems and their ability to absorb change and disturbance and still maintain the same relationships between populations and state variables'.

Holling, 1973

Hazards



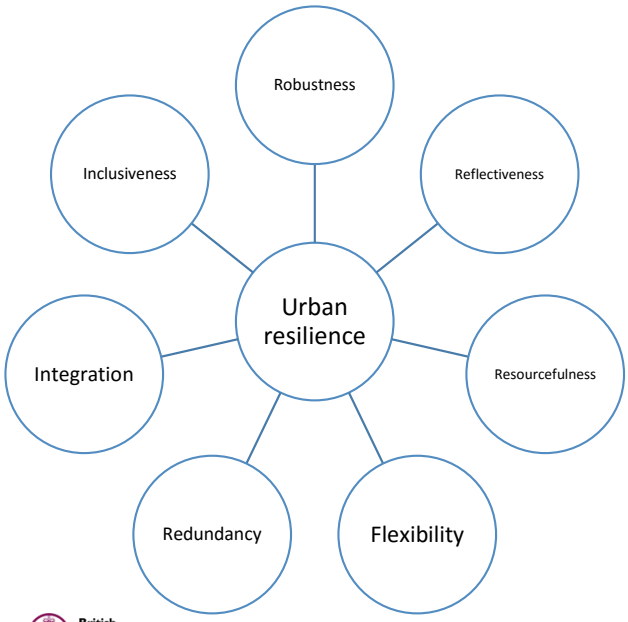
WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

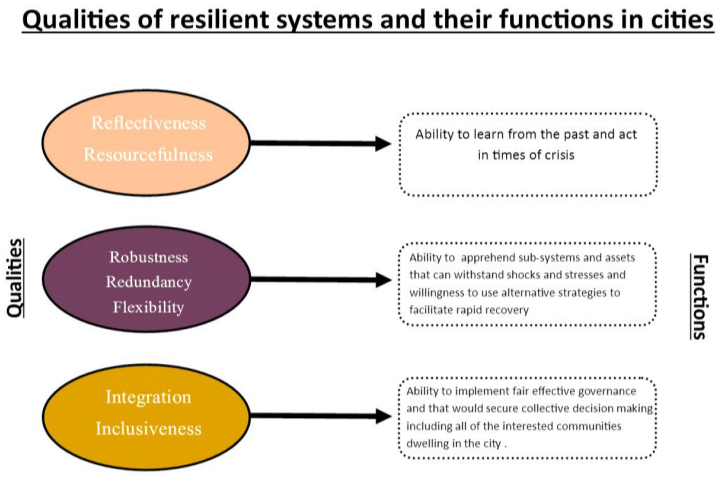
What is urban resilience?

Urban resilience is the capacity of the city's components and subsystems to prepare for all kinds of shocks, absorb their impact by adapting to the new reality and transform to a more efficient entity.

Rodin, 2014

Qualities of urban resilience





WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Resilient urban planning and design

Multifunctionality of urban space

- Mixing uses and combining functions.

Redundancy and modularisation

- Spread the risks across time, geographical areas and multiple subsystems.

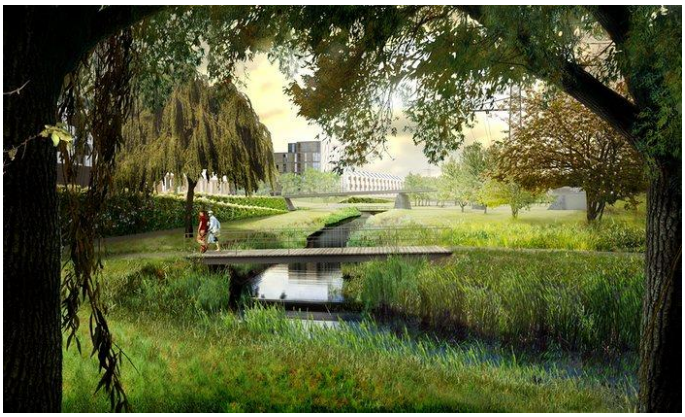
Adaptive planning and design

- Safe to fail experimental projects implementation and monitoring of their outcome.

Multi-scale connectivity

- Short feedback loops and quicker response in time of hazard. Smooth flow of materials people, information. Aesthetic contribution as well.

Resilient urban planning and design



Flood resilient design in Barking and Dagenham

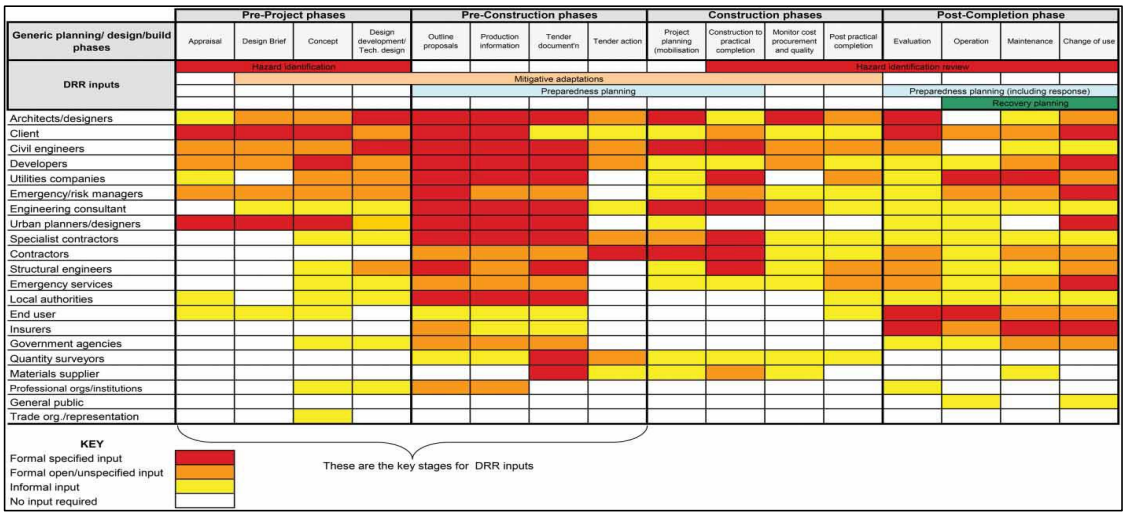
Bicycle routes in the Netherlands

WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Negligence of geology from the urban resilience discourse

- Geological dimensions of resilience have been widely neglected in resilience policies.
- Many models have attempted to connect resilience and the built environment.
- Difference between resilient infrastructure and infrastructure that builds resilience.
- Many examples of resilience policies failing due to lack of geological insight.

Negligence of geology from the urban resilience discourse



(Bosher, 2014)

Steps towards integrating geology in the resilience discourse

- Illustrations of ground suitability for different uses.
- 3D and 4D urban planning tools connecting the **ground properties** with the **land use.**
- Connection of geo-environmental models with social and economic models.
- Monitoring of individual development in urban areas.
- Implementation of resilient planning and design.

(Bricker et al, 2015)

Assessment of a resilience policy implementation

- City of Thessaloniki
- Part of the 100 Resilient Cities Network
- Two phases of implementation.
  - Preliminary Resilience Assessment, PRA (3 months)
  - Resilience Strategy Development (9 months)
- Five axis for the implementation of the resilience policy
  a) Data Driven Thessaloniki
  b) Ecosystem that promotes new working modules
  c) Thermaikos Bay: introducing water in the everyday life of the people
  d) Mobility as a driver for business continuity and environmental mitigation
  e) Co-ownership of public space

Conclusions

- Urban resilience is a very 'hot' topic in contemporary urban studies.
- It reflects the long journey of resilience across different disciplines, as explored above
- Holistic approaches provide a good chance for re-theorising urban risk.
- Resilient urban design and planning should reflect urban resilience qualities and be self-monitored by them
- Geology needs to be included in this multi-disciplinary approach through targeted actions.

Further steps

- Categorisation of different types of geo-hazards.
- Formulation of a new dataset through the combination of geological and socio-economic already existing datasets.
- Generation and testing of a model for the quantification of urban resilience.
- Provision of policy and design directives.
- Incorporation of 3D modeling in the generated model.
- Application in a UK case study.

References

- Bricker S., Reeves H., Campbell S.D.G., Price S. (2015), 'The ground beneath cities: where should future development occur?',Thought piece by the British Geological Survey
- Desouza K. and Flanery T. (2014), 'Designing, planning and managing resilient cities: A conceptual framework', *Cities*, **35**: pp. 89-99
- Holling, C. S. (1973), 'Resilience and stability of ecological systems', *Annual Review of Ecology and Systematics* **4** (1): pp 1-23
- Rodin J. (2014), 'The resilience Dividend: Managing disruption, avoiding disaster, and growing stronger in an unpredictable world', New York: Profile Books
- 100 Resilient Cities Network (2016), available at: URL<http://www.100resilientcities.org/> , (access: 26/4/2016)

WARWICK
WARWICK INSTITUTE FOR THE SCIENCE OF CITIES
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Questions?

Full name: Caroline Player

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Trust in a Multi Agent System

Abstract: This project so far has been a literature review and initial experiments in to the extension of trust algorithms to include stereotypes and tag-based work with the aim of aiding cooperation in an agent based system. Trust between two agents denotes the expectation of an interaction with that agent, and requires some prior experience with that agent to make this assessment. It is not always the case that this data exists, and with the use of stereotypes and tags we hope to improve initial assessments, leading to a more realistic and robust trust evaluation mechanism.

Keywords: Agents, Trust, Tags, Stereotypes

Abstract Word Count (Est): 95

Full name: Lee Prangnell

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Poster

Title: Adaptive Quantization Matrices for HD and UHD Display Resolutions in Scalable HEVC

Abstract: HEVC contains an option to enable custom quantization matrices (QMs), which are designed based on the Human Visual System (HVS) and a 2D Contrast Sensitivity Function (CSF). Visual Display Units (VDUs), capable of displaying video data at High Definition (HD) and Ultra HD (UHD) display resolutions, are frequently utilized on a global scale. Video compression artifacts that are present due to high levels of quantization, which are typically inconspicuous in low display resolution environments, are clearly visible on HD and UHD video data and VDUs. The default HVS-CSF QM technique in HEVC does not take into account the video data resolution, nor does it take into consideration the associated VDU\x92s display resolution to determine the appropriate levels of quantization required to reduce unwanted video compression artifacts. Based on this fact, we propose a novel, adaptive quantization matrix technique for the HEVC standard, including Scalable HEVC (SHVC). Our technique, which is based on a refinement of the current HVS-CSF QM approach in HEVC, takes into consideration the display resolution of the target VDU for the purpose of minimizing video compression artifacts. In SHVC SHM 9.0, and compared with anchors, the proposed technique yields important quality and coding improvements for the Random Access configuration, with a maximum of 56.5% luma BD-Rate reductions in the enhancement layer. Furthermore, compared with the default QMs and the Sony QMs, our method yields encoding time reductions of 0.75% and 1.19%, respectively.

Keywords: HEVC; Quantization Matrix; Human Visual System; Contrast Sensitivity Function.

Abstract Word Count (Est): 235

ADAPTIVE QUANTIZATION MATRICES FOR HD AND UHD RESOLUTIONS IN SCALABLE HEVC

Lee Prangnell and Victor Sanchez

Department of Computer Science, University of Warwick, Coventry, England, UK

Introduction

- The proposed Adaptive Quantization Matrix (AQM) technique is a novel refinement to the default Human Visual System and Contrast Sensitivity Function (HVS-CSF) based Quantization Matrix (QM) technique in HEVC.
- The proposed AQM technique employs a display resolution parameter and a Euclidean distance parameter for the purpose of minimizing video compression artifacts in high resolution VDUs.
- These two parameters represent the resolution of the target VDU and the importance of transform coefficients within a Transform Block (TB) with respect to the resolution of the VDU.

HVS-CSF Default QM Technique in HEVC

- The HVS-CSF based 8×8 QM technique in [1] is employed as the default intra QM in HEVC [2].
- HEVC supports up to 32×32 TBs.
- Default 16×16 and 32×32 QMs, for the corresponding TB sizes, are obtained by upsampling and replicating the 8×8 default intra and inter QMs (see Fig. 1).

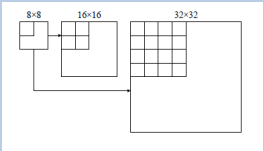


Fig. 1. The replication of 8×8 QMs to construct 16×16 and 32×32 QMs for corresponding TB sizes.

The HVS-CSF approach in [1], including the associated Modulation Transfer Function (MTF), is employed to produce a 2D Frequency Weighting Matrix (FWM), $H(f)$, comprising floating point values, from which the default intra QM in HEVC is derived. $H(f)$ is computed in (1):

$$H(f) = \begin{cases} a(b + cf) \times e^{-c(f)^d}) & \text{if } f' > f_{max.} \\ 1.0 & \text{otherwise,} \end{cases} \quad (1)$$

where f is the radial frequency in cycles per degree of the visual angle subtended represented in two dimensions such that $f = f(u,v)$, where f_{max} denotes the frequency of 8 cycles per degree (i.e., the exponential peak). The MTF is computed with the constant values a=2.2, b=0.192, c=0.114 and d=1.1. The normalized radial spatial frequency, $f'(u,v)$, is defined using angular dependent function $S(\theta(u,v))$. Both $f'(u,v)$ and $S(\theta(u,v))$ are quantified in (2)-(5).

$$f'(u,v) = \frac{f(u,v)}{S(\theta(u,v))} \quad (2) \qquad f(u,v) = \frac{\pi}{180 \sin^{-1}(1/\sqrt{1+dis^2})} \times \sqrt{f(u)^2 + f(v)^2} \quad (3)$$

$$S(\theta(u,v)) = \frac{1-s}{2}\cos(4\theta(u,v)) + \frac{1+s}{2} \quad (4) \qquad \theta(u,v) = \arctan\left(\frac{f(u)}{f(v)}\right) \quad (5)$$

where dis represents the perceptual viewing distance of 512mm and s is the symmetry parameter with a value of 0.7. Parameter s ensures that the floating point values in $H(f)$ are symmetric. As s decreases, $S(\theta(u,v))$ decreases at approximately 45 degrees; this, in turn, increases $f'(u,v)$ and decreases $H(f)$. The discrete horizontal and vertical frequencies are computed in (6):

$$f(u) = \frac{u-1}{\Delta \times 2N}, \text{ for } u = 1, 2..., N; \qquad f(v) = \frac{v-1}{\Delta \times 2N}, \text{ for } v = 1, 2..., N; \quad (6)$$

where Δ denotes the dot pitch value of 0.25mm (approximately 100 DPI) and N is the number of horizontal and vertical radial spatial frequencies. A static dot pitch value of 0.25mm is utilized to compute FWM $H(f)$.

- The default QM technique in HEVC does not take into account the specific display resolution of a target VDU, nor does it take into account the importance of transform coefficients in a TB with respect to the resolution of the target VDU.
- The dot pitch can be the same value for a multitude of VDU display resolutions depending on the pixel density of the VDU. The default QM technique does not take this into account.

Related Work

- The method in [3] involves adjustments to the parameter selection of the HVS-CSF QM technique in [1]. This refinement produces a modified FWM, from which the intra and inter QMs are derived. This technique does not take into account the target VDU's display resolution.
- In [4], the authors propose a novel intra QM method that modifies the weighting values in the QM by employing a normalized exponent variable. Similar to the method in [3], this technique does not take into account the target VDU's display resolution.

Proposed AQM Technique

- We focus on integrating the proposed intra and inter AQMs into SHVC to produce a two-layered bit-stream for SNR scalability. Lower levels of quantization are applied to the EL for the purpose of reducing video compression artifacts in high resolution VDUs (see Fig. 2).

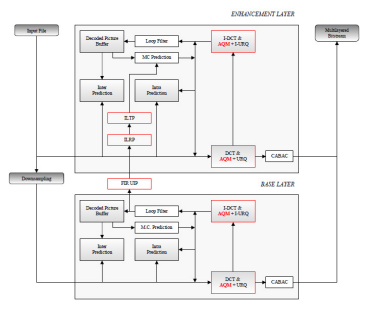


Fig. 2. A block diagram of the AQM technique in SHVC. The AQM technique is highlighted in red. This diagram shows how our method operates on an inter-layer basis in SHVC.

Our technique is based on parameter $A_{i,j}$, which is applied to each element of $H(f)$ located at position (i,j), denoted as $H_{i,j}$. $H'_{i,j}$, is computed in (7).

$$H'_{i,j} = H_{i,j}^{A_{i,j}} \quad (7)$$

Parameter $A_{i,j}$ is computed as a function of two parameters in (8):

$$A_{i,j}(d_{i,j}, w) = e^{-\left(\frac{d_{i,j}}{w}\right)} \in [0,1] \quad (8)$$

where $d_{i,j}$ is the normalized Euclidean distance between the DC transform coefficient and the current coefficient located at position (i,j) in a TB, and w is the display resolution parameter. Euclidean distance $d_{i,j}$ is computed in (9):

$$d_{i,j} = \sqrt{\frac{(i_1 - i_2)^2 + (j_1 - j_2)^2}{(i_1 - i_{max})^2 + (j_1 - j_{max})^2}} \in [0,1] \quad (9)$$

where (i_1, j_1), (i_2, j_2), (i_{max}, j_{max}) represent the position of the floating point values in $H(f)$ associated with the DC coefficient, the current coefficient and the farthest AC coefficient, respectively. Each $A_{i,j}$ value decreases as the display resolution parameter w decreases. The w parameter is quantified in (10) and the VDU's normalized hypotenuse value p is computed in (11); parameter w rapidly decreases as p increases (see Fig. 3).

$$w = h_t^{-p} \in (0,1] \quad (10) \qquad p = \frac{h_a}{h_t} \in (0,1] \quad (11)$$

where h_t is the VDU's maximum hypotenuse value, in pixels, and where h_a is the VDU's actual hypotenuse value in the pixel domain; h_t and h_a are computed in (12) and 13), respectively:

$$h_a = \sqrt{x^2 + y^2} \quad (12) \qquad h_t = \sqrt{x_{max}^2 + y_{max}^2} \quad (13)$$

where (x, y) represent the horizontal and vertical dimensions of the target VDU, respectively, and (x_{max}, y_{max}) represent, respectively, the maximum possible horizontal and vertical dimensions of the target VDU.

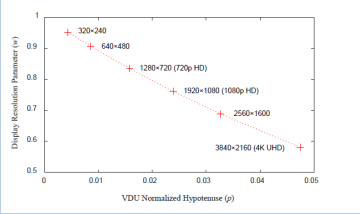


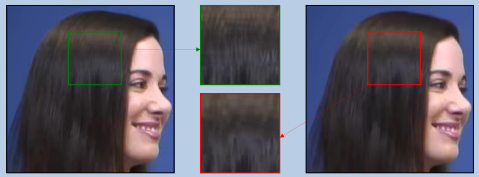
Fig. 3. Display resolution parameter w rapidly decreases as the VDU's normalized hypotenuse value p, and its display resolution, increases.

Experimental Setup

- **Reference Software:** SHVC SHM 9.0 (latest version).
- **Video Quality Metric:** BD-Rate.
- **Anchors:** Default QMs in HEVC in [1, 2] and the Sony QM technique in [3].
- **QPs:** 22, 27, 32, 37.
- **Encoding Configurations:** All Intra (Main), Low Delay (Main) and Random Access (Main).
- **Scalable Bit-Stream:** Two-layered bit-stream. The BL is aimed at HD 720p VDUs (1280×720) and the EL is aimed at 4K VDUs (3840×2160).

Results & Discussion

- In Table 1, we tabulate the average BD-Rate (for EL and BL). In Fig. 4, we show an example of the reconstruction improvements attained by our technique.



(a) (b)

Fig 4. A single frame of the HD KristenAndSara video sequence, coded using a QP = 30. Fig 4 (a) shows the improvement of the reconstruction quality of the frame using the AQMs designed for a 4K VDU with QP = 30 versus the default QMs in HEVC, as shown shown in Fig 4 (b).

Table 1. BL and EL average BD-Rate results of the proposed AQM technique compared with anchors. The results in green indicate performance improvements, the results in black indicate no improvements and the results in red indicate negative results.

Proposed AQM Technique versus Default QMs																								
Class	All Intra			Low Delay B			Random Access			Class	All Intra			Low Delay B			Random Access		
	Y %	U %	V %	Y %	U %	V %	Y %	U %	V %		Y %	U %	V %	Y %	U %	V %	Y %	U %	V %
A (BL)	-0.6	-0.5	-0.3	-2.3	-2.6	-2.6	-3.0	-5.0	-5.4	A (BL)	-2.2	0.0	-0.2	-3.2	-3.9	-4.7	-4.8	-6.9	-7.2
B (BL)	-0.4	-0.1	0.0	-2.1	-2.2	-2.5	-2.9	-4.5	-5.0	B (BL)	-1.1	-1.1	-1.7	-2.3	-4.7	-6.5	-3.0	-6.6	-8.2
C (BL)	-0.4	0.2	0.2	-3.1	-3.3	-2.9	-3.6	-5.4	-5.0	C (BL)	-2.3	-3.9	-4.8	-1.9	-5.5	-5.3	-4.8	-11.2	-10.8
D (BL)	-0.4	0.2	0.2	-2.1	-2.3	-2.6	-2.3	-3.9	-2.9	D (BL)	-2.3	-0.1	-3.2	-1.8	-2.1	-3.0	-2.7	-5.4	-5.1
E (BL)	-0.2	0.2	0.3	-2.9	-3.6	-4.4	-3.3	-2.8	-2.8	E (BL)	-0.8	0.2	0.5	-3.5	-4.7	-5.7	-4.3	-3.4	-3.1
Average	**-0.4**	**0.0**	**0.1**	**-2.5**	**-2.8**	**-3.0**	**-3.0**	**-4.3**	**-4.2**	**Average**	**-1.7**	**-1.0**	**-1.9**	**-2.5**	**-4.2**	**-5.0**	**-3.9**	**-6.7**	**-6.9**
Class	**All Intra**			**Low Delay B**			**Random Access**			**Class**	**All Intra**			**Low Delay B**			**Random Access**		
	Y %	**U %**	**V %**	**Y %**	**U %**	**V %**	**Y %**	**U %**	**V %**		**Y %**	**U %**	**V %**	**Y %**	**U %**	**V %**	**Y %**	**U %**	**V %**
A (EL)	-0.8	-0.5	-0.1	-19.7	-19.1	-19.0	-37.4	-39.7	-40.1	A (EL)	-2.3	0.0	-0.1	-52.7	-56.0	-56.6	-56.5	-58.7	-59.2
B (EL)	-0.4	-0.1	0.0	-12.3	-14.2	-14.4	-40.4	-43.7	-44.5	B (EL)	-1.3	-1.1	-1.6	-36.2	-42.9	-43.0	-50.6	-55.5	-56.9
C (EL)	19.0	21.1	21.5	-2.9	-2.6	-2.2	-30.0	-32.0	-31.9	C (EL)	-45.0	-47.8	-48.2	-19.9	-24.1	-23.7	-44.4	-49.1	-49.1
D (EL)	-3.2	-3.7	-3.6	-6.3	-6.9	-7.0	-32.5	-35.2	-34.5	D (EL)	-52.6	-56.9	-57.2	-28.5	-31.6	-31.8	-39.1	-42.5	-42.1
E (EL)	-1.6	-2.0	-2.6	-6.0	-6.7	-6.9	-29.3	-31.3	-31.1	E (EL)	-33.5	-35.4	-35.0	-6.6	-6.9	-7.1	-33.5	-35.4	-35.0
Average	**2.60**	**2.96**	**3.16**	**-9.44**	**-9.90**	**-9.90**	**-33.92**	**-36.38**	**-36.42**	**Average**	**-26.94**	**-28.24**	**-28.42**	**-28.78**	**-32.30**	**-32.44**	**-44.82**	**-48.24**	**-48.50**

Most Significant BD-Rate Improvement Results Compared with Anchors (EL versus EL)

- **Sequence:** Class B HD (Default QMs in HEVC) & Class A UHD 4K Sequence (Sony QMs).
- **Encoding Configuration:** Random Access (Main).
- **BD-Rate versus Default QMs in HEVC:** -40.4% (Y), -43.7% (Cb) and -44.5% (Cr).
- **BD-Rate versus Sony QMs:** -56.5% (Y), -58.7% (Cb) and -59.2% (Cr).

The very high EL versus EL BD-Rate improvements are mainly due to the increased accuracy of inter-layer prediction for the EL. That is, our technique yields improved reconstruction of the BL, which allows for a more accurate prediction for the EL.

Encoding & Decoding Times Improvement Results Compared with Anchors

- **Encoding & Decoding Times versus Default QMs in HEVC:** -0.75% and -4.67%, respectively.
- **Encoding & Decoding Times versus Sony QMs :** -1.19% and -2.82%, respectively.

A more accurate prediction of the EL from the BL decreases the workload of the entropy coding process, which leads to improved encoding and decoding times in the proposed AQM technique.

Conclusions

A novel AQM technique for Scalable HEVC is proposed to improve video reconstruction quality, thereby reducing the visibility of compression artifacts on high resolution VDUs. Compared with anchors, the proposed method yields important coding efficiency and visual quality improvements, with a maximum luma BD-Rate improvement of 56.5% in the EL. In addition, the proposed technique attains modest encoding and decoding time improvements.

[1] C. Wang, S. Lee and L. Chang, "Designing JPEG quantization tables based on human visual system," *IEEE International Conference on Image Processing*, Kobe, Japan, 1999, vol. 2, pp. 376-380.

[2] V. Sze, M. Budagavi and G. J. Sullivan, "Quantization Matrix," *in High Efficiency Video Coding (HEVC): Algorithms and Architecture*, Springer International Publishing, 2014, pp. 158-159.

[3] M. Haque, A. Tabatabai and Y. Morigami, "HVS model based Default Quantization Matrices," *Document JCTVC-G880*, 2011, pp. 1-13.

[4] S. Jeong and B. Jeon, "Newer Quantization Matrices for HEVC," *Document JCTVC-I0126*, 2012, pp. 1-8

Full name: Ayman Qahmash

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Qualitative study to investigate the relationship between mathematics and programming through students' perceptions

Abstract: Generally, mathematical abilities have a certain degree of importance in computer science education and a correlation between mathematics and programming. In this qualitative study, the relationship between programming and mathematics had been investigated through computer science students' perception. I interviewed nine computer science students asking them a questions to capture their perception on how mathematical abilities can be implemented in programming and what is the relationship between mathematics and programming. The correlation between different mathematics concepts such as calculus and discrete mathematics had been investigated. The results show that the students find that discrete mathematics could be more applicable to introductory programming module more than calculus whereas data structures and functional programming module could relate positively to student performance in calculus more than discrete mathematics.

Keywords: Computer Science Education

Abstract Word Count (Est): 125

Full name: Farrukh Qazi

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Cybercrimes Investigations & Enigmatic Challenges in Modern Computing Era

Abstract: Where the convergence of information and communication systems has transformed the approach of global data processing, concurrently, it has also opened myriad windows of opportunities for cyber criminals, whether it\x92s a cyber-enabled or cyber-dependent crime.

Freely accessible anti-forensics tools/ techniques and online collaborations facilitate to circumvent the security of digital assets, which ultimately presents multi-layered challenges for cyber forensics investigators within the corporate\x92s distributed digital estate. Cyber attacks or compromises (both internal / external) do require an immediate investigation followed by well defined detect, analyse and protect policy. Identifying and preserving fragile digital artifiacts is the prime objective of cyber forensics team in timely and methodical fashion. Court admissible evidence collection, correctness, completeness, authenticity and integrity are very strategical when big data is scattered through complexed virtualized, cloud and mobile based outsourced computing environments. In this presentation, my research curiosity shows some of those vital fragments from the cyber forensics threat landscape e.g. tamper proofing, secure provenance, evidence preservation and non-repudiation so to work out closely with practitioners for some achievable process based framework to tackle these challenges in future.

Keywords: Cyber forensics; Non-Repudiation; Tamper Proofing; Secure Provenance; Evidence Prservation

Abstract Word Count (Est): 180

WARWICK
THE UNIVERSITY OF WARWICK

WPCCS
2016

Cybercrimes Investigations & Enigmatic Challenges in Modern Computing Era

Farrukh Qazi – *(Ist Year PhD Student)*
Supervisor: Dr. Arshad Jhumka
Department of Computer Science
The University of Warwick

Abstract

Where the convergence of information and communication systems has transformed the approach of global data processing, concurrently, it has also opened myriad windows of opportunities for cyber criminals, whether it's a cyber-enabled or cyber-dependent crime.

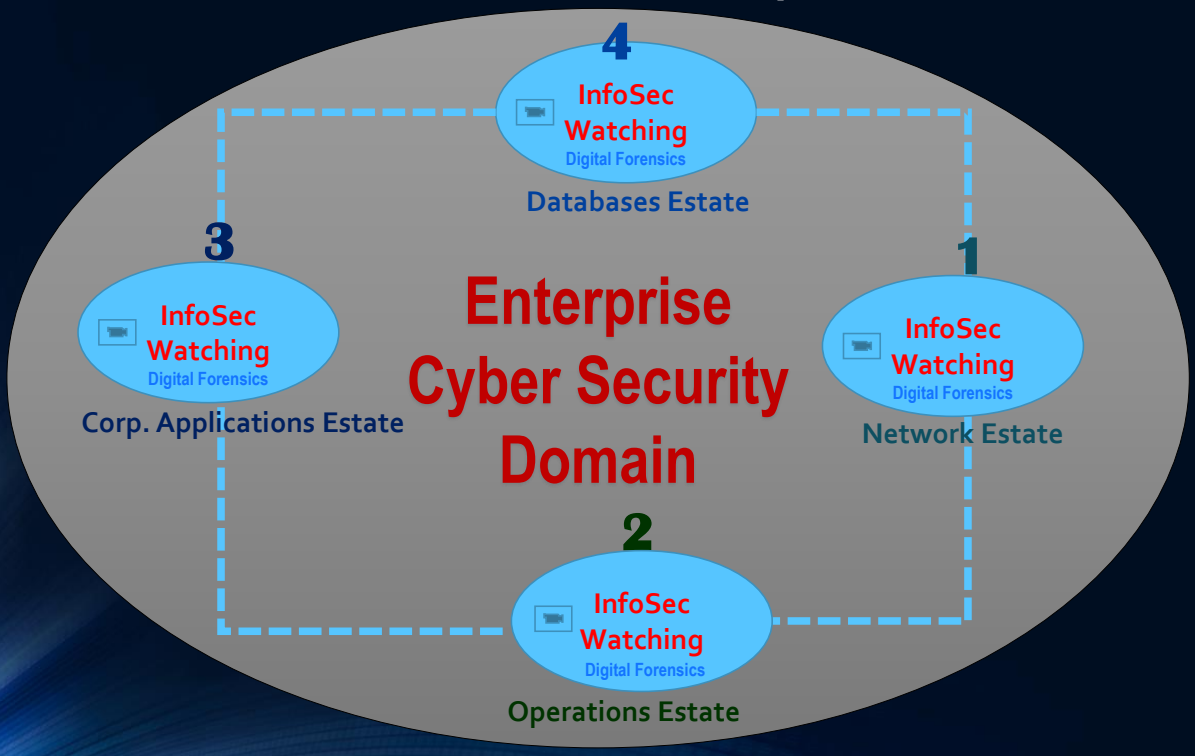
Freely accessible anti-forensics tools/ techniques and online collaborations facilitate to circumvent the security of digital assets, which ultimately presents multi-layered challenges for cyber forensics investigators within the corporate's distributed digital estate. Cyber attacks or compromises (both internal / external) do require an immediate investigation followed by well defined detect, analyse and protect policy. Identifying and preserving fragile digital artifiacts is the prime objective of cyber forensics team in timely and methodical fashion. Court admissible evidence collection, correctness, completeness, authenticity and integrity are very strategical when big data is scattered through complexed virtualized, cloud and mobile based outsourced computing environments. In this presentation, my research curiosity shows some of those vital fragments from the cyber forensics threat landscape e.g. tamper proofing, secure provenance, evidence preservation and non-repudiation so to work out closely with practitioners for some achievable process based framework to tackle these challenges in future.

AGENDA

- 01 Information Security Paradigm
- 02 Cyber Forensics – 4 x Ws & Scope
- 03 Cyber Forensics - Challenges
- 04 Cyber Forensics – My Research Interests

01 Information Security Paradigm

Cyber Security & Digital Forensics
The Relationship



Global Cyber Security Threats – A Scary Count!!!..

Global Internet Users

2000: 400m

2015: 3.2 b

Source : (ICT Facts & Figures 2015)

Constant Stream of Cyber Warfare

Global Cyber Attacks Live Show…
http://map.norsecorp.com/#/

Cyber Threats

New Malware Variants (Added in Each Year)

2014	2015
317M	431M
–	+36%

Mobile Threats

New Mobile Vulnerabilities

2013	2014	2015
127	168	528
–	+32%	+214%

System's Vulnerabilities

New Vulnerabilities

2013	2014	2015
6,787	6,549	5,585
–	-4%	-15%

Web Threats

Websites Found with Malware

2013	2014	2015
1 in 566	1 in 1,126	1 in 3,172

[Source: Symantec Internet Security Threat Report 2016]

Theoretical Motivation

1

Locard's Exchange Principle

Whenever two objects have an interaction, they leave some sort of artefacts behind, it's just matter of discovering them.

2

Rumsfeld Metrix

1. Known knowns 	2. Known unknowns 
3. Unknown knowns 	4. Unknown Unknowns 

Crime Scene Investigation - The Protocol

Cyber Security/ Forensics Investigation – The Protocol

Digital Forensics Best Practices

- ACPO
- NIST
- ENISA
- SWGDE
- US-CERT
- Other Frameworks for Digital Forensics

02 Cyber Forensics – 4 x Ws & Scope

Digital Forensics – 4 X Ws

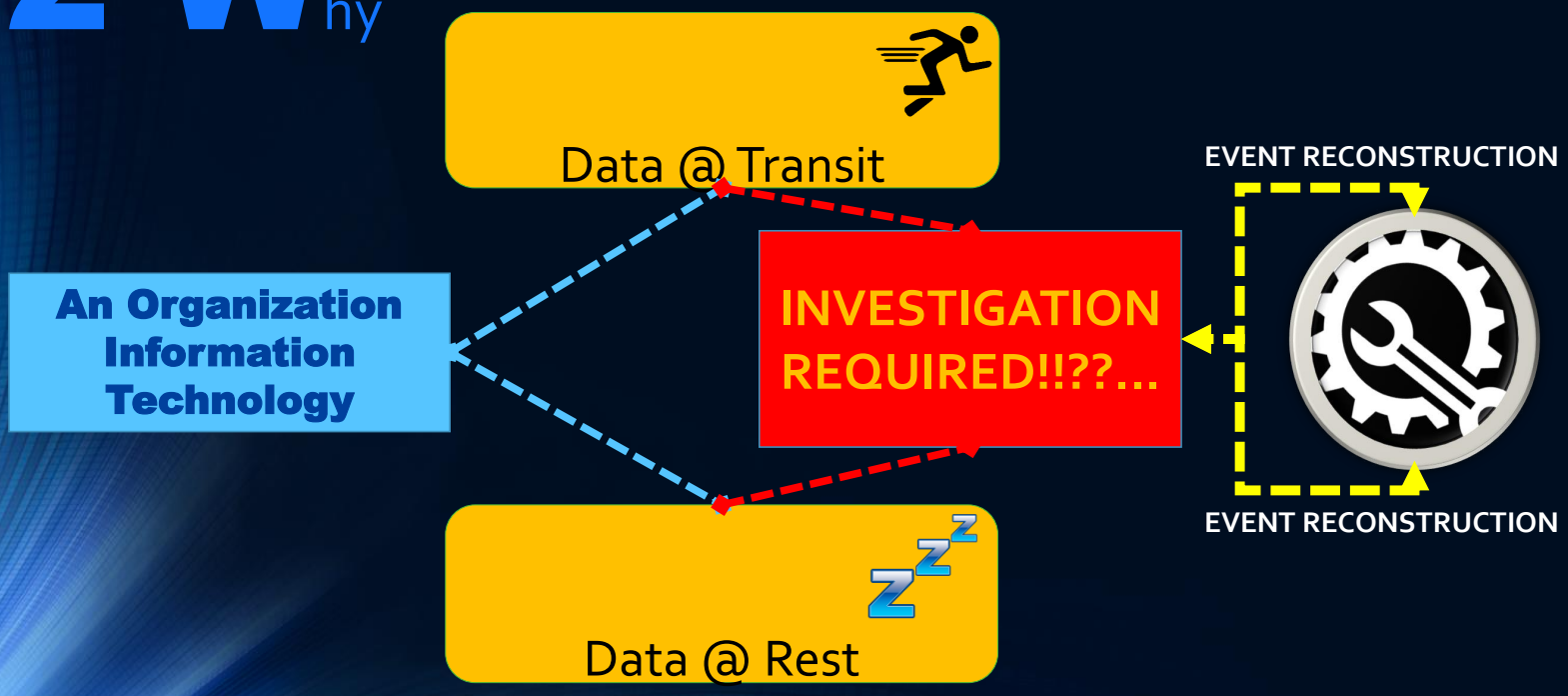
1-What

"It is considered the application of science to the **identification**, **collection**, **examination**, and **analysis** of **data** while **preserving** the **integrity of the information** and maintaining a strict **chain of custody** for the data."

[Ref; NIST-SP800-86 Integrated DF Techniques into IR]

Digital Forensics – 4 X Ws

2-Why



3-Where

Digital Forensics – 4 X Ws

Where is the Evidence

- Physical
- Local
- Volatile
- Owned

Vs.

- Virtual
- Remote
- Non-Volatile
- Outsourced

4-When

Digital Forensics – 4 X Ws

Generally 2 x Approaches

Pro-active

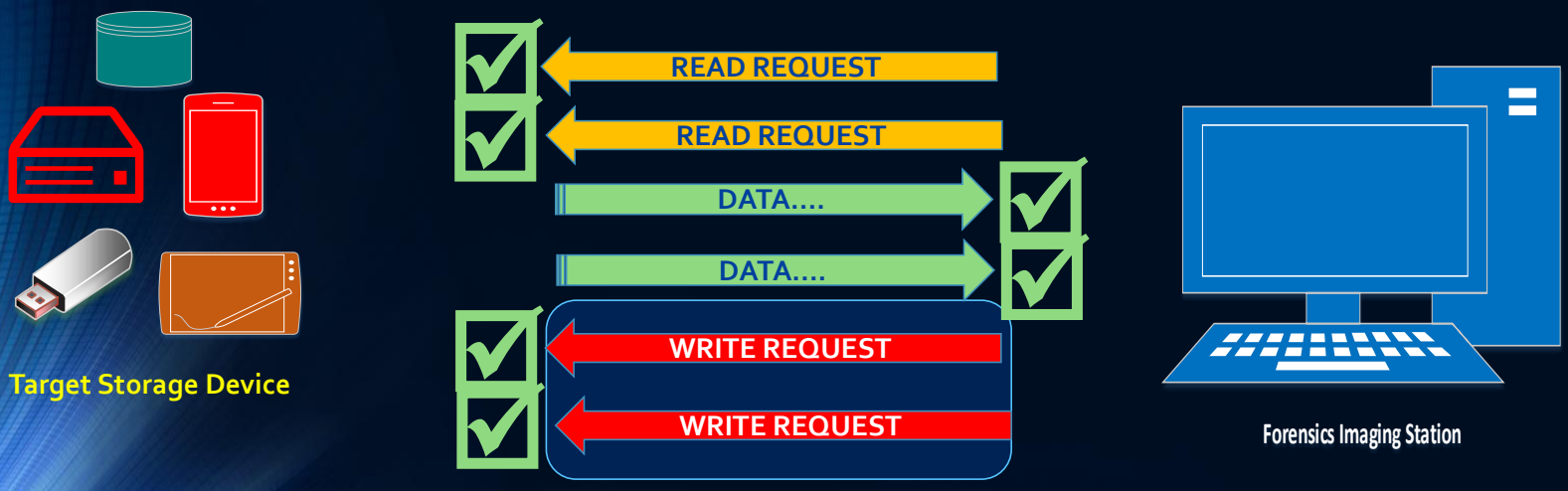
- Pre Cyber Incident/ Breach
- Continuous Monitoring
- Detect>Response>Comply

Re-active

- Post Cyber Incident/ Breach
- Swift Response
- Detect>Response>Comply

03 Cyber Forensics - Challenges

Digital Forensics – Data Read/ Write Operation

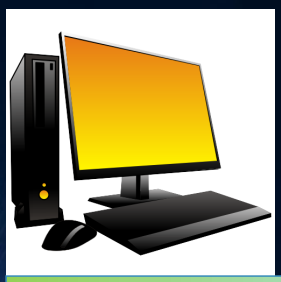


Digital Forensics – Evidence Imaging

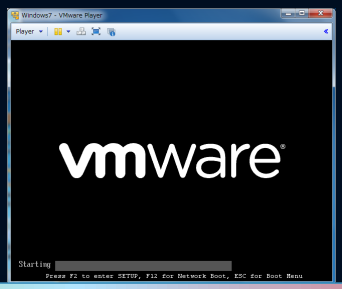
Digital Forensics – Challenges

1. Operational
2. Technical
3. Legal

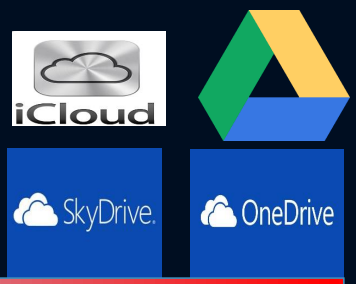
Physical Machines



Virtualization based



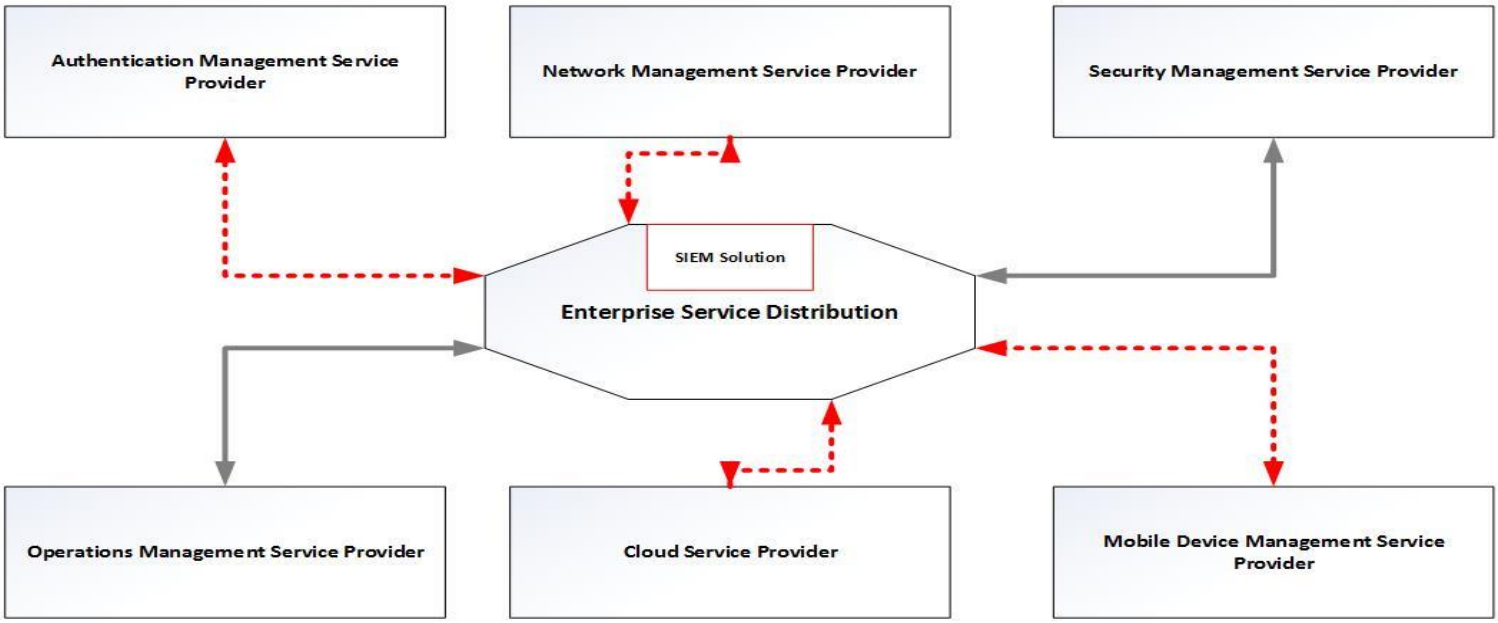
Cloud based

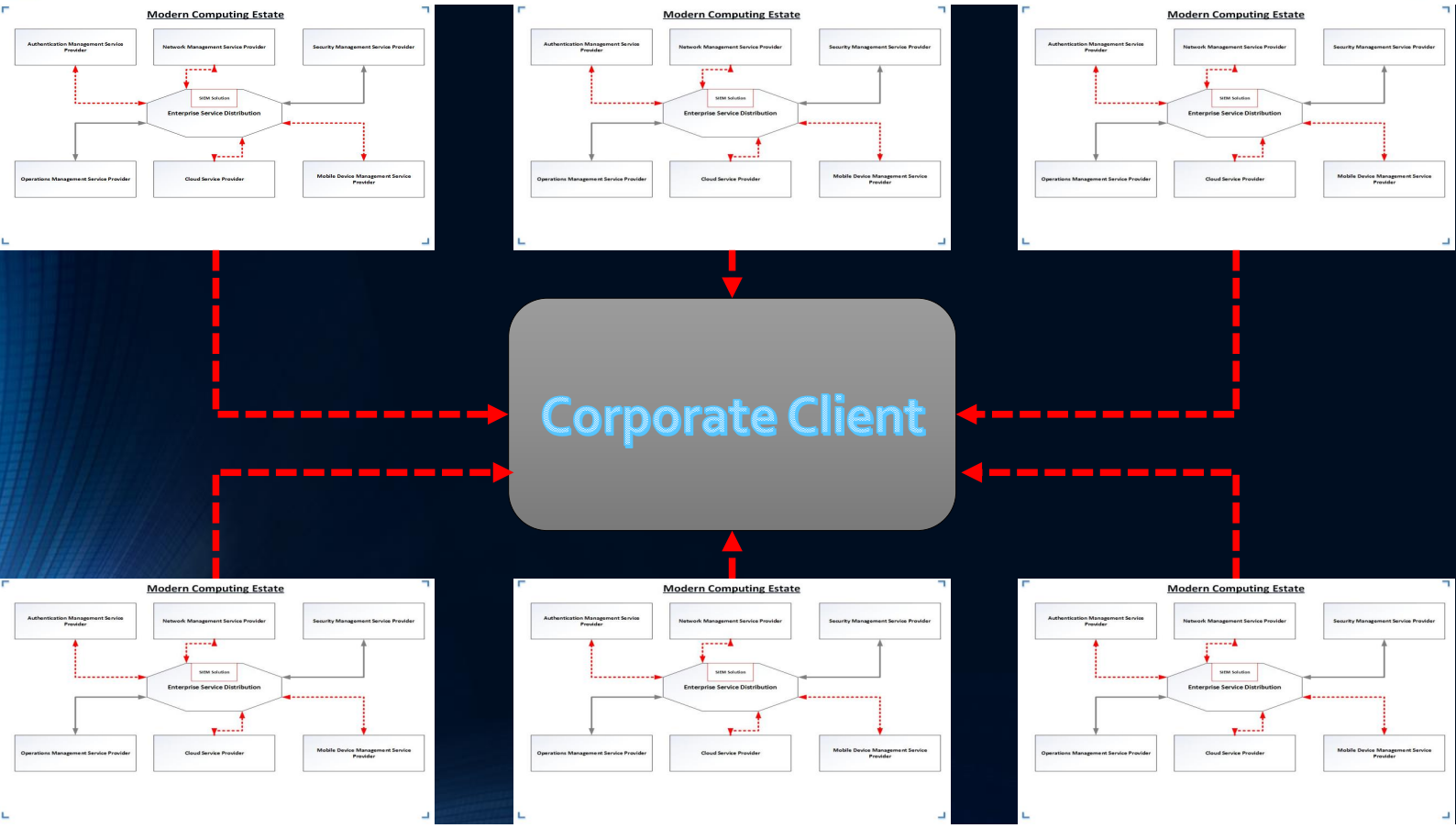


Digital Investigation Analysis Complexity

Investigation Progression & Complexity Level....

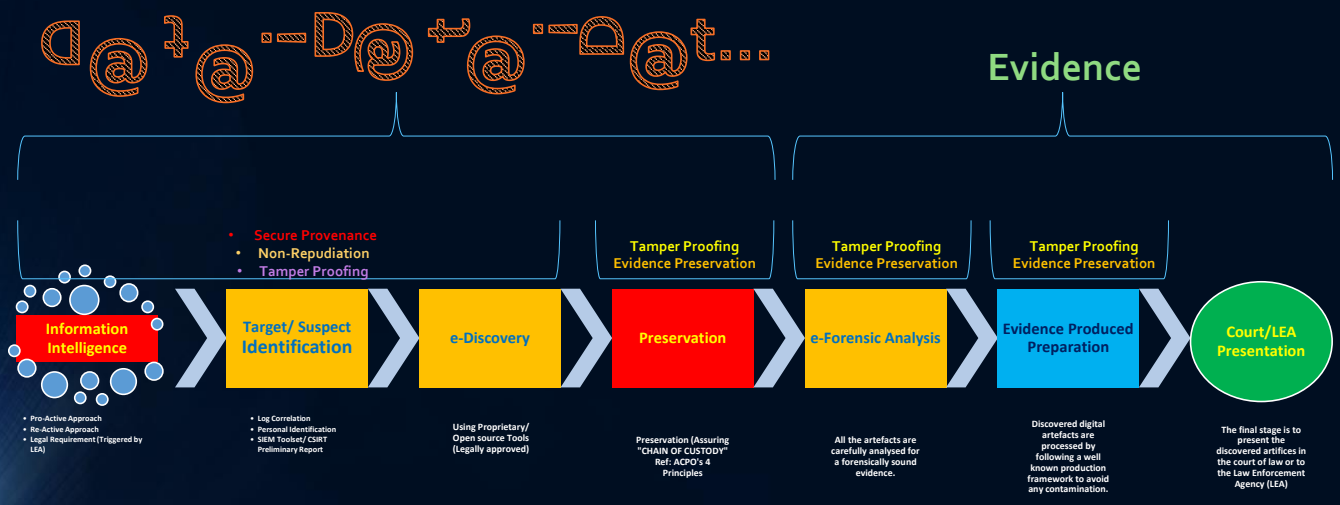
Modern Computing Estate



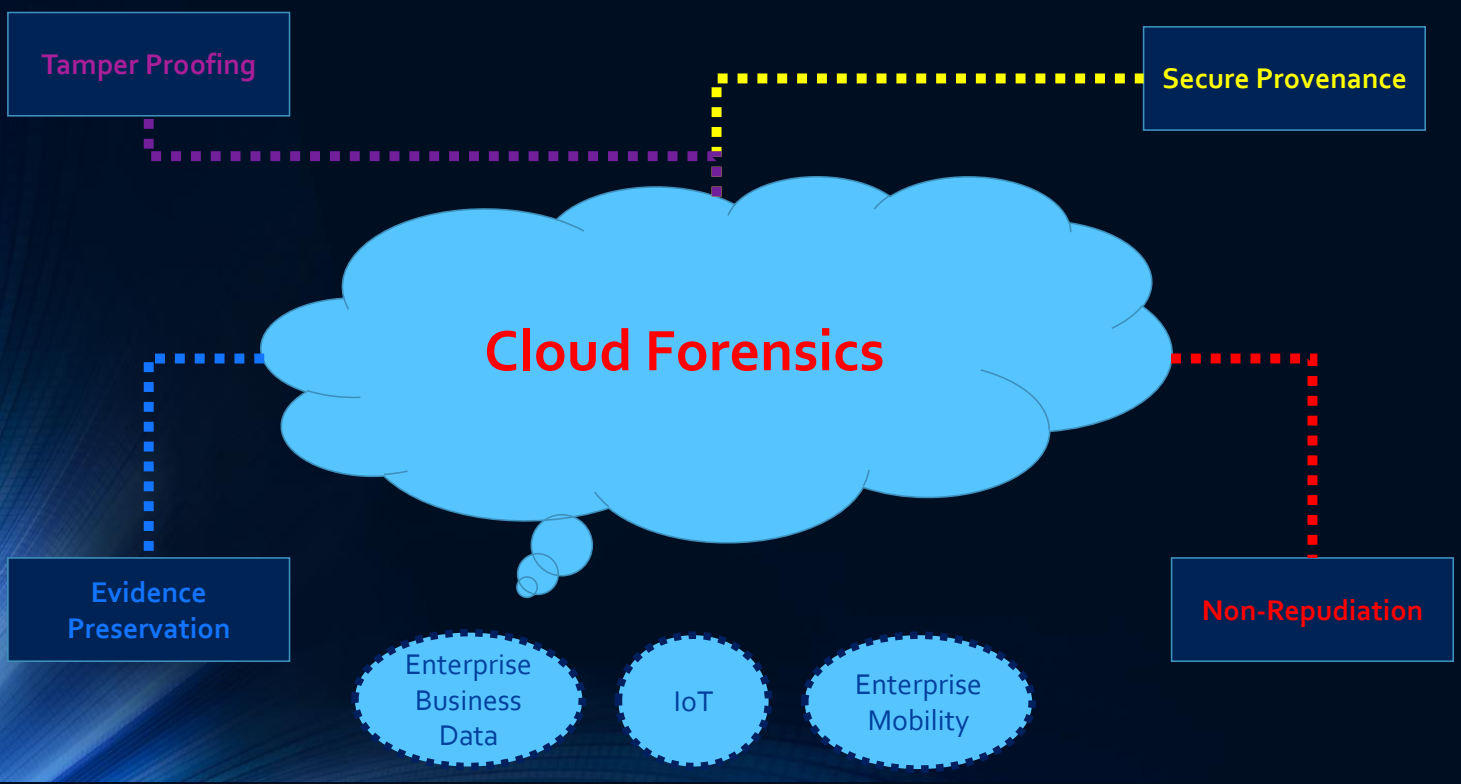


04 Cyber Forensics – My Research Interests

Digital Forensics
The Process



Digital Forensics – My Research Area!!!…



Digital Forensics – My Research Area!!!…

Secure Provenance

"The fact of coming from some particular source; origin, a record of the ultimate derivation and passage of an item through its various owners."
[*Oxford English Dictionary*]

Simmhan et al. The key information which assists to locate the original history of a data and starting from its original sources.

- Provenance can be used to trace the audit trail of data
- Determine resource usage

Digital Forensics – My Research Area!!!…

Non- Repudiation

"Strong and substantial evidence of the identity of the signer of a message and of message integrity, sufficient to prevent a party from successfully denying the origin, submission or delivery of the message and the integrity of its contents."

[Ref: American Bar Association: american1996digital]

Kuhn et al. that Non-Repudiation acknowledges proofs and identity of both the sender and the recipient and transmission receipts.

[Ref: NIST : kuhn2001introduction]

Digital Forensics – My Research Area!!!…

Tamper Proofing

- Anti-Anti-Forensics Techniques
- Data Integrity
- Tampering Techniques
  - Anti-Forensics
  - Malware
  - Rootkits
  - Reverse Engineering
- Physical Attacks
  - Critical systems/ security/ application logs deletion
  - Critical systems/ security/ application logs Modification
- Cybercrime as a Service
- Cloud Storage - Especially Multi-Tenants Data – who's the owner of this data?

Digital Forensics – My Research Area!!!…

- Data is Volatile
- Data is Fragile
- Data is Big
- Data is All Over
- Data is Outsourced
- Data is stored remotely

"Forensically Soundness / Court Admissibility"

Evidence Preservation

My Research lifecycle

Questions?

Farrukh Qazi
f.a.qazi@warwick.ac.uk

Thanks!!!…

References

Alqahtany, S., et al. (2015). Cloud Forensics: A Review of Challenges, Solutions and Open Problems. Cloud Computing (ICCC), 2015 International Conference on.

Association, A. B., et al. (1996). Digital Signature Guidelines: Legal Infrastructure for Certification Authorities and Secure Electronic Commerce, Section of Science and Technology, American Bar Association.

Kent, K., et al. (2006). "Sp 800-86. guide to integrating forensic techniques into incident response."

Kuhn, D. R., et al. (2001). Introduction to public key technology and the federal PKI infrastructure, DTIC Document.

Locard, E. (2008). Locard's Exchange Principle.

Logan, D. C. (2009). "Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry." Journal of Experimental Botany **60(3): 712-714.**

QPM, D. J. W. (2012). Good Practice Guide for Computer-Based Electronic Evidence.

Schatz, B. and M. Cohen (2010). Refining evidence containers for Provenance and accurate data Representation. Advances in Digital Forensics VI, Springer**: 227-242.**

Simmhan, Y. L., et al. (2005). "A survey of data provenance in e-science." ACM Sigmod Record **34(3): 31-36.**

Zawoad, S. and R. Hasan (2013). Digital forensics in the cloud, DTIC Document.

Full name: John Rahilly

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Poster

Title: The Predictable City?

Abstract: The entwined fates of the city and its people are inexorably and interminably bound to the consequences of various forms of public policy intervention. From determining access to education for a child to the provision of state pensions in later years, policy-makers influence life at every stage of development and have the capacity to irreversibly trans- form our futures. However, the policy-making and evaluation process has consistently been analysed as inadequate. Despite recent moves towards an evidence based policy approach, the primary means of impact evaluation remain retrospective analysis following implementation.

Various modelling methods have been utilised in regards to particular public policy interventions, but it can be contended do not adequately inform the decision making process. Consequently, this re- search will seek to examine current methods of policy evaluation and the use of city modelling and simulation as an informational tool, before experientially analysing their effectiveness in predicting the out- comes associated with three defined policy interventions within the broad spheres of judicial, social and planning policy.

Based upon this analysis it is intended to identify the essential requirements for an optimised suite of complex policy modelling, simulation and impact assessment models, which will be developed as a practical, quantitative analytical testing ground for policy and legislative interventions.

Keywords: City Modelling, Policy Making, Policy Evaluation, Complex Adaptive Systems

Abstract Word Count (Est): 209

The Predictable City?

Improving Policy-Making Through the Application of City Modelling and Simulation

John Rahilly

Centre for Doctoral Training in Urban Science

j.rahilly@warwick.ac.uk

WARWICK

THE UNIVERSITY OF WARWICK

Abstract

The entwined fates of the city and its people are inexorably and interminably bound to the consequences of various forms of public policy intervention. From determining access to education for a child to the provision of state pensions in later years, policy-makers influence life at every stage of development and have the capacity to irreversibly transform our futures. However, the policy-making and evaluation process has consistently been analysed as inadequate. Despite recent moves towards an evidence based policy approach, the primary means of impact evaluation remain retrospective analysis following implementation.

Various modelling methods have been utilised in regards to particular public policy interventions, but it can be contended do not adequately inform the decision making process. Consequently, this research will seek to examine current methods of policy evaluation and the use of city modelling and simulation as an informational tool, before experientially analysing their effectiveness in predicting the outcomes associated with three defined policy interventions within the broad spheres of judicial, social and planning policy.

Based upon this analysis it is intended to identify the essential requirements for an optimised suite of complex policy modelling, simulation and impact assessment models, which will be developed as a practical, quantitative analytical testing ground for policy and legislative interventions.

"Experience is the only propehcy of wise men."

de Lamartine., 1847

Introduction

The entwined fates of the city and its people are inexorably and interminably bound to the caprices of public policy intervention. From determining access to education to the provision of state pensions, policy-makers are the unseen actor with influence over life at every stage of development and have the capacity to irreversibly transform our futures (*Birkland., 2014*). However, the policy making process generally remains a recondite system (*Hallsworth and Rutter., 2011*), reportedly characterised by persistent failures to realise preconceived outcomes (*Hallsworth et al., 2011*) and identified as inducing significant unintended consequences (*Bovens et al., 2002*).

The consequences of public policy interventions which are perceived as defective can be critical to both society and governments (*Howlett., 2012*). Ineffective policy can potentially perpetuate a problem paradigm or even create new issues within a society, allied to which significant governmental costs can be incurred in relation to subsequent mitigation or termination (*McConnell., 2010*) and ultimately erode trust in the system of governance (*Hallsworth and Rutter., 2011*).

Increasing governmental attention has been focused upon improving the policymaking process (*Hallsworth et al., 2011*) therefore, it seems a prescient moment in which to examine the current paradigm and consider novel approaches.

Margaret Thatcher's resignation in 1990 can in part be attributed to the estimated 20 billion (equivalent to 2015) costs associated with the failure of the *Poll Tax*.

Policy Making

Most conventional models related to policy-making have historically reflected a simple, broadly cyclical process (*Anderson., 1975; May and Wildavsky., 1978; Brewer and de Leon., 1983; Hallsworth et al., 2011*) founded upon principles of systems thinking (*Birkland., 2014*). Such models generally include, but are not restricted to distinct stages of Agenda setting, Policy formulation and Decision making, Implementation, Evaluation and Maintenance, Succession or Termination (*Jann and Wegrich., 2007*) [refer to **figure 1**].

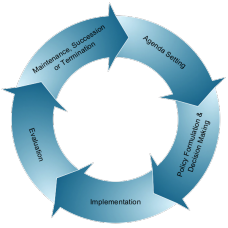


Figure 1: Graphic of Public Policy Cycle as defined by *Jann and Wegrich., (2007)*

Policy Evaluation

Reflective of the burgeoning prevalence of 'outcome' oriented policy (*Sanderson., 2002*), the evaluation stage has come to be dominated by a focus upon the efficacy of interventions in relation to intended aims (*Vedung., 2009*).

However, it is acknowledged there is a significant issue relating to the identification of causality between an observed effect and implementation of a policy intervention (*Pearce and Raman., 2014*). In order to address this problem governments have increasingly begun to adopt a *Randomised Controlled Trial* approach based upon piloted interventions (*Haynes et al., 2012*).

It is suggested that the use of an RCT approach enables the most robust analysis (*Haynes et al., 2012*), but crucially can be subject to ethical objection (*Oakley et al., 2006*) and time constraint (*Pearce and Raman., 2014*).

Aim and Objectives

It is the primary aim of this research to develop a comprehensive understanding of the requirements for an holistic, exploratory modelling system, in which to evaluate the impacts associated with different public policy options.

In order to achieve this aim the following objectives have been identified:

1. Theoretically examine the extent to which 'big data' addresses issues related to causal inference within policy evaluation;
2. Evaluate the impacts of defined policy case study examples;
3. Investigate extant policy impact assessment simulation tools in relation to defined case study examples;
4. Evaluate and identify aspects of relevant policy simulation models;
5. Propose a novel model for the simulation of policy impacts.

City Modelling and Policy Simulation

City Modelling

"This digital world which parallels the material, now gives us unprecedented power to understand and explore cities."

Batty., 2009

Initially the modelling of cities tended to focus upon the physical environment (*Hamilton and Curwell., 1998*), but over the course of the last decade there have been advanced efforts to develop holistic systems, which incorporate elements of Land-Use Transportation, Urban Dynamics, Agent-Based Models and Micro-simulation (*Batty., 2009*).

Extant city models have broadly been based upon two categories of data (*Hamilton et al., 2005*);

- Spatial data, which relates to the physical form of a city [refer to **figure 2**]

and

- Thematic data, which relates to the functions of a city

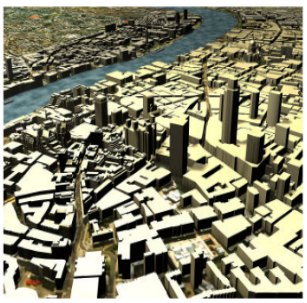


Figure 2: Visualisation of London in Miller-Hare model is used to support planning decision-making (*Batty et al., 2000*)

Policy Simulation

Simulation models are increasingly recognised as integral tools to analyse the impacts of policy, based upon both cost and viability (*Majstorovic et al., 2015*).

Different forms of policy simulation models have been utilised within the economic and planning spheres for a prolonged period (*Brenner and Werker., 2009*). However, similar developments in regards to social policy represent a less established field (*Heidelberg and Desai., 2015*).

Exploratory models (*Majstorovic et al., 2015*) have been developed by governments around the world, with one of the most comprehensive simulations undertaken by the *Statistique Canada* Social Policy Simulation Database and Model.

Consideration will be given to the value offered by various types of simulation tools, including system dynamics, micro-simulation, Agent-Based Models and algorithmic models.

It is hypothesised that combined data derived from city models and principles from existing policy simulation may offer new insights into predictive policy-making.

Research Impact

Arguably the need for the policy process to become more efficient and effective is greater than ever, with governmental budgets having been decimated by the austerity agenda (*Lowndes and Pratchett., 2012*).

Therefore, the development of a suite of robust policy simulation models holds the potential to revolutionise policy-making at both local and national levels.

References

Anderson, J.E. (1975). Public Policymaking. New York: Praeger. Batty, M. (2009). Urban modeling. International Encyclopedia of Human Geography. Oxford, UK: Elsevier.

Birkland, T. A. (2014). An introduction to the policy process: Theories, concepts and models of public policy making. Routledge.

Bovens, M. A., T'Hart, P., and Peters, B. G. (Eds.). (2002). Success and failure in public governance: A comparative analysis. Edward Elgar Publishing.

Brenner, T., and Werker, C. (2009). Policy advice derived from simulation models. Journal of Artificial Societies and Social Simulation, 12(4), 2.

Brewer, G., and deLeon, P. (1983). The Foundations of Policy Analysis. Monterey, Cal.: Brooks, Cole.

Hallsworth, M. and Rutter, J. (2011). Making Policy Better: Improving Whitehalls Core Business. Institute for Government [pdf]. [Available at: http://www.instituteforgovernment.org.uk/sites/default/files/publications/Making

Hallsworth, M., Parker, S. and Rutter, J. (2011) Policy Making in the Real World: Evidence and Analysis. Institute for Government. [pdf]. [Available at: http://www.instituteforgovernment.org.uk/sites/default/files/publications/Making

Hamilton, A., Curwell, S., and Davies, T. (1998). A simulation of the urban environment in Salford. In Proceeding of the CIB World Building Congress, Gvle, Sweden, CIB.

Hamilton, A., Wang, H., Tanyer, A. M., Arayici, Y., Zhang, X., and Song, Y. H. (2005). Urban information model for city planning. Journal of Information Technology in Construction (ITCon), 10, 55-67.

Haynes, L., Goldacre, B., and Torgerson, D. (2012). Test, learn, adapt: developing public policy with randomised controlled trials. Cabinet Office-Behavioural Insights Team.

Heidelberg, R. L., and Desai, A. (2015). Simulation Rules: The Role of Simulation in Policy Inquiry. Administrative Theory and Praxis, 37(1), 1-17.

Howlett, M. (2009). Policy analytical capacity and evidencebased policymaking: Lessons from Canada. Canadian public administration, 52(2), 153-175.

Jann, W., and Wegrich, K. (2006). 4 Theories of the Policy Cycle. Handbook of public policy analysis, 43.

Lowndes, V., and Pratchett, L. (2012). Local governance under the coalition government: Austerity, localism and the Big Society. Local government studies, 38(1), 21-40.

Majstorovic, D., Wimmer, M. A., Lay-Yee, R., Davis, P., and Ahrweiler, P. (2015). Features and added value of simulation models using different modelling approaches supporting policy-making: a comparative analysis. In Policy Practice and Digital Science (pp. 95-123). Springer International Publishing.

May, J.P., and Wildavsky, A. (ed.) (1978). The Policy Cycle. Beverly Hills, CA: Sage.

McConnell, A. (2010). Policy success, policy failure and grey areas in-between. Journal of Public Policy, 30(03), 345-362.

Oakley, A., Strange, V., Bonell, C., Allen, E., and Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. BMJ (Clinical research ed.), 332(7538), 413-416.

Pearce, W., and Raman, S. (2014). The new randomised controlled trials (RCT) movement in public policy: challenges of epistemic governance. Policy sciences, 47(4), 387-402.

Sanderson, I. (2002). Evaluation, policy learning and evidencebased policy making. Public administration, 80(1), 1-22.

Vedung, E. (2009). Public policy and program evaluation. Transaction Publishers.

Full name: Joao Marcal Ramos

Year of Study: MSc (Research)

Department: Computer Science

Submission Type: Presentation

Title: Minimizing Decomposable Submodular Functions

Abstract: Submodular function minimization is a fundamental optimization problem that arises in several applications in machine learning and computer vision. The problem is known to be solvable in polynomial time, but general purpose algorithms have high running times and are unsuitable for large-scale problems. Recent work have used convex optimization techniques to obtain very practical algorithms for minimizing functions that are sums of 'simple' functions and, in particular, decomposable submodular functions. A new analysis of the convergence rate of the current algorithms for this problem is presented from the point of view of image segmentation (ie, graph cut) as well as a new 'hybrid' algorithm which takes advantage of the intuition gained in this analysis.

Keywords: convex optimization, machine learning, algorithms, discrete mathematics

Abstract Word Count (Est): 114

Full name: Mohammed Rehman

Year of Study: PhD Year 4 or later

Department: Computer Science

Submission Type: Presentation

Title: Phenomenography, culture and mobile learning

Abstract: Phenomenography is a qualitiative research methodology that originated in the 1970's, in an attempt to investigate learning through the eyes of the learner. In this presentation I will discuss how phenomenography was used to collect data from participants in 3 countries (Japan, UK and China) in order to understand how students engage with mobile learning and the extent to which their cultural context influences their attitudes towards, and perceptions of, learning that involves the use of mobile devices.

Keywords: mobile learning, methodology, phenomenography, culture

Abstract Word Count (Est): 78

Full name: Shenyuan Ren

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Developing Power-aware Scheduling Mechanisms for Virtualized Environments

Abstract: in such a system is a critical metric to measure
the sustainability and eco-friendliness of the system. This paper develops three power-aware scheduling
strategies in virtualized systems managed by Xen, which is a popular virtualization technique. These three
strategies are the Least performance Loss Scheduling (LLS) strategy, the No performance Loss Scheduling
(NLS) strategy and Best Frequency Match (BFM) scheduling strategy. These power-aware strategies are
developed by identifying the limitation of Xen in scaling the CPU frequency and aim to reduce the energy
waste without sacrificing the jobs running performance in the computing systems virtualized by Xen. LLS
works by re-arranging the execution order of the VMs. NLS works by setting a proper initial CPU frequency
for running the VMs. BFM reduces energy waste and performance loss by allowing the VMs to jump the
queue so that the VM that is put into execution best matches the current CPU frequency. Scheduling for both
single core and multi-core processors are considered in this paper. The evaluation experiments have been
conducted and the results show that compared with the original scheduling strategy in Xen, the developed
power-aware scheduling algorithm is able to reduce energy consumption without reducing the performance
for the jobs running in Xen.

Keywords: Scheduler, Power-aware,Cloud Computing

Abstract Word Count (Est): 203

Developing Power-aware Scheduling Mechanisms for Virtualized Environments

Shenyuan Ren

July 1st,2016

Contributions:

– This paper develops three power-aware scheduling strategies in virtualized systems managed by Xen:

1) The ***Least performance Loss*** Scheduling (LLS) strategy
2) The ***No performance Loss*** Scheduling (NLS) strategy
3) The ***Best Frequency Match*** (BFM) scheduling strategy

–These power-aware strategies are developed by identifying the limitation of Xen in ***scaling the CPU frequency*** and aim to reduce the energy waste without sacrificing the jobs running performance in Xen:

1) LLS works by ***re-arranging the execution order*** of the VMs
2) NLS sets ***a proper initial CPU frequency*** for running the VMs
3) BFM allows the VMs ***to jump the queue***

– Scheduling for both ***single-core*** and ***multi-core processors*** are considered in this paper.

A Motivating Example

– Under the OnDemand governor, Dynamic Voltage Frequency Scaling ***(DVFS)*** is used to adjust the CPU frequency on demand.

– The execution time slice of a VM is ***30ms*** by default. The power governor checks and changes, if necessary, the CPU frequency every ***10ms*** (called the ***frequency scaling slice***) by ***at most one level***.

– Scheduler :SEDF (Simple Earliest Deadline First). In SEDF, the CPU requirement for each VM can be translated to ***the deadlines*** by which a VM has to start running. In each scheduling round, SEDF puts the VM with ***the earliest deadline*** into execution.

A Motivating Example

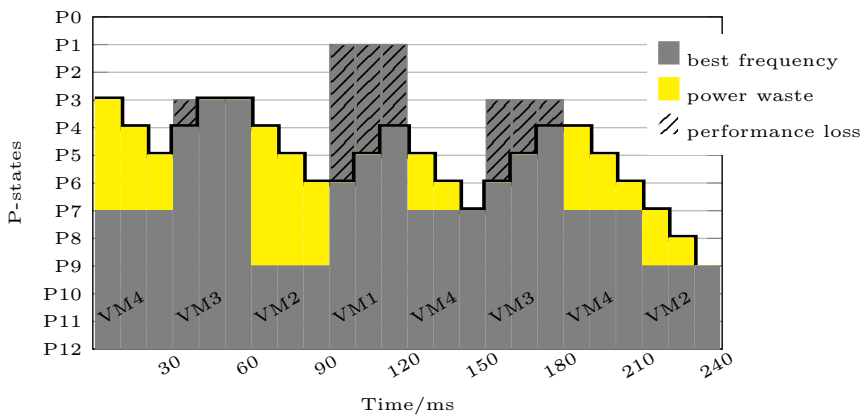


Figure: Energy waste and performance loss under DVFS in Xen

Performance Model

Assume a set of independent tasks T = $\{T_1, T_2, ..., T_n\}$, task T_i runs in VM_i.

f_i: the best frequency of VM_i

t_i: the execution time of task T_i runs at the frequency of f_i

P_s: the scheduling time slice, which is 30ms by default

P_f: the frequency scaling slice, which is 10ms

$F_t(f_i')$: the function of execution time over CPU frequency

Equivalent execution rate $c(f_i')$:

$$c(f_i') = \begin{cases} \frac{t_i}{F_t(f_i')} & if f_i' < f_i, \\ 1 & if f_i' \geq f_i. \end{cases} \tag{1}$$

Performance Model

$f_i'(j)$: the frequency which task T_i runs at in the jth time interval

$c(f_i'(j))$: the execution rate of T_i in the jth time interval

Determine the number of intervals that VM_i uses to complete task T_i:

$$\sum_{j=0}^{m} c(f_i'(j))P_f \ \geq t_i \tag{2}$$

The ***total execution time*** of task T_i:

$$t_i' = \sum_{j=0}^{m} P_f \tag{3}$$

Energy Consumption Model

C: the capacitance being switched per clock cycle

V: the voltage

A: the activity factor

The power consumption ***rate*** of CPU for running T_i (denoted by r_i):

$$r_i(f_i') = A \times C \times V^2 \times f_i' \tag{4}$$

The ***total energy consumption*** of task T_i, denoted by e_i:

$$e_i = \sum_{j=0}^{m} r_i(f_i'(j))P_f \tag{5}$$

The Scheduling Strategy with Least Performance Loss

Theorem 1 gives the "Least Performance Loss" Scheduling strategy (LLS), namely ***the execution order*** of the tasks that leads to the least performance loss.

Theorem 1

Given a set of tasks, $T = \{T_1, T_2, \ldots, T_n\}$, the best CPU frequency of T_i is f_i. Assume $f_1 \leq f_2 \leq \ldots \leq f_n$. If the current CPU frequency is f, then given the current CPU frequency, the LLS strategy is to run the tasks in the following order, where T_r's frequency f_r is the highest frequency that is less than the current frequency f.

Namely, the execution order is to start from T_r, go up to T_n in the increasing order of frequency and then come down to T_1 in the decreasing order of frequency, and that the upward and downward execution pattern in terms of frequency repeat until all tasks have been completed.

The Scheduling Strategy with Least Performance Loss

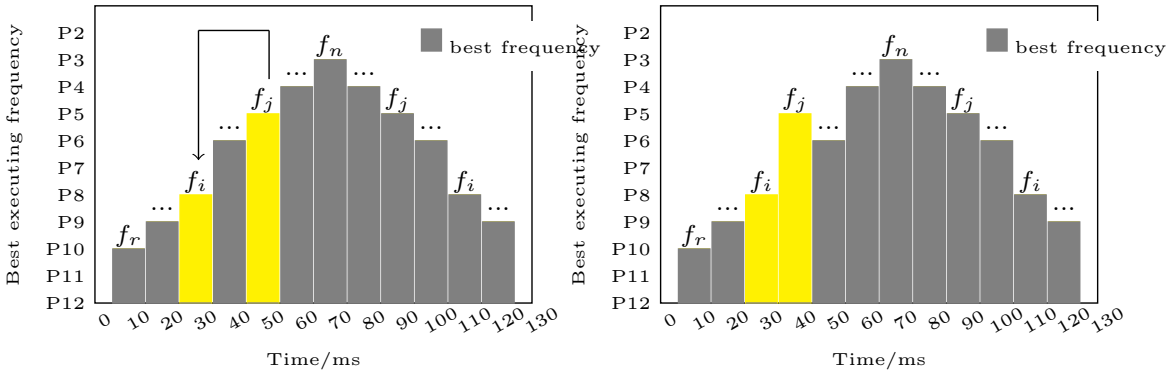


Figure: Least Performance Loss Scheduling Strategy & Illustration of changing the execution position of a randomly selected task

The Scheduling Strategy with No Performance Loss

NLS calculates ***the initial CPU frequency*** that the CPU needs to be set with in order for all VMs to run ***without performance loss***.

NLS scheduling strategy uses below equation to determine the execution frequency at every frequency scaling slice:

$$f_{k'}(i-1) = max(f_k(i) - \Delta f, f_{k'}) \tag{6}$$

where,

$$k' = \begin{cases} k & \text{if } (i-1)\text{-th frequency scaling slice is in } T_k\text{'s scheduling slice} \\ k-1 & \text{if } (i-1)\text{-th frequency scaling slice is in } T_{k-1}\text{'s scheduling slice} \end{cases} \tag{7}$$

The Scheduling Strategy with No Performance Loss

Note that although NLS guarantees no performance loss, it may cause ***energy waste***.

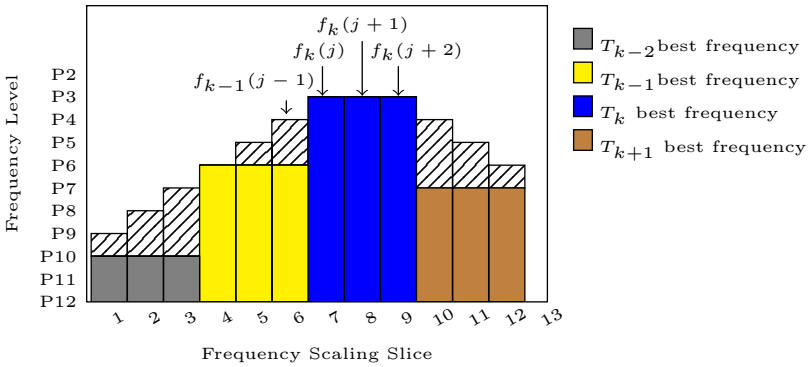


Figure: "No Performance Loss" Scheduling Strategy

The Scheduling Strategy with No Performance Loss

Input: Tasks T_1,T_2,...,T_n in the run-queue, whose best running frequencies are f_1,f_2,...,f_n; T_K is the task with the highest best frequency f_K; j is the index of the first frequency scaling slice in T_K's scheduling slice

Output: $f_1(1)$;

```
k = K, f_k(j) = f_K;
for (i = j; i ≥ 2; i − −) do
    if (j − 1)th frequency scaling slice is in T_k's scheduling slice then
        k' = k
    else
        k' = k − 1;
    end if
    f_k'(i − 1) = max(f_k(i) − Δf, f_k');
end for
```

Algorithm 1: No performance loss scheduling strategy

Power-aware SEDF Scheudling (Single-core)

Assume that a set of VMs T_i $(1 \leq i \leq n)$ are in the run queue of a single core. If the first VM in the run-queue has ***the smallest gap*** between its the best frequency and the current executing frequency, the VM will be scheduled.

However, if there are other VMs in the queue which have ***smaller frequency gaps*** than the first VM, BFM will first identify this VM, for instance T_j. Before allowing task T_j to jump the queue, BFM needs to make sure that all the VMs queueing before T_j satisfy Inequality 8.

$$\forall L_i < L_j \quad d_i - [t_c + (L_i + 1) * P_s] \geq 0 \tag{8}$$

Power-aware SEDF Scheudling (Single-core)

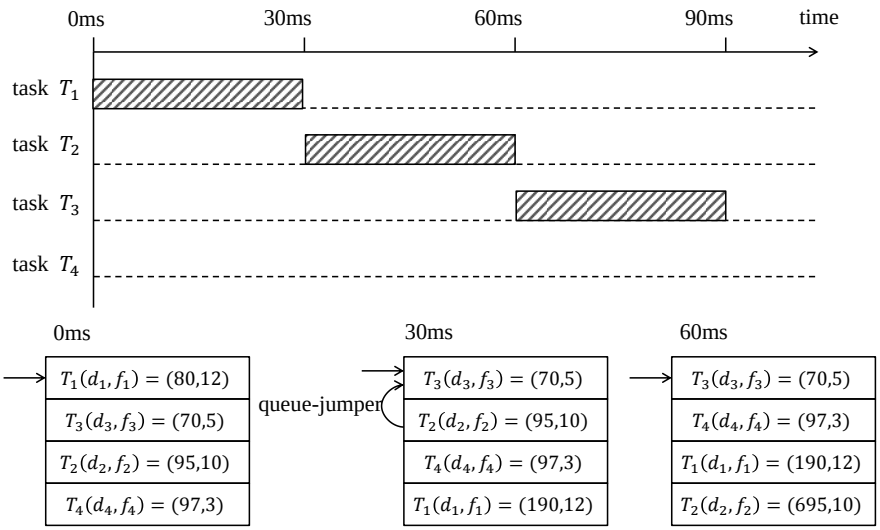


Figure: An example of the BFM scheduling strategy

Power-aware SEDF Scheudling (Single-core)

Input: A set of of VMs, T_i $(1 \leq i \leq n)$, with their CPU requirements (p_i, s_i, x_i); the best frequency of T_i, f_i; scheduling slice P_s; frequency scaling slice P_f; current time t_c; T_i's deadline d_i; T_i's location in the queue, L_i;

```
for The end of each scheduling slice do
    Calculate the deadlines of all tasks, sort the tasks in the ascending order of
    deadline in queue Q
    Obtain the VM with the earliest deadline, denoted by T_e
    for all Tasks in the run queue do
        Calculate the frequency gap g_i between the best frequency of task T_i and the
        current frequency f
    end for
    Sort the tasks' frequency gaps in the ascending order
    Get the first frequency gap, g_k, in the frequency-gap sorting queue and denote
    the corresponding task by T_k
    if T_k is T_e, i.e., g_e is the minimal gap
     then
        Schedule T_e to run
    else
        Set flag = 0
        while g_k is no more than g_e do
            if Each task T_i before T_k in queue Q satisfies d_i − [t_c + (L_i + 1) ∗ P_s] ≥ 0
            then
                Schedule T_k to run next
                Set flag = 1
                Exit while
            else
                Get the next frequency gap, g_k, in the frequency-gap sorting queue and
                denote the corresponding task by T_k
            end if
        end while
        if flag = 0 then
            Schedule T_e to run next
```


Power-aware SEDF Scheudling (Multi-core)

1) When allocating a set of VMs to a set of cores, BFMM tries to allocate the VMs with the ***closest best frequencies*** to the same core as long as the CPU requirement of the VMs on the same core can be satisfied.

2) When BFMM allocates T_i and finds that T_i's CPU requirement cannot be met in the core, it moves to ***the next VM*** and check if the next VM can be allocated the core. The process continues until all VMs are examined for the current core.

3) BFMM then moves to the next core and tries to allocate VMs to the core. This process repeats until all VMs are allocated.

Power-aware SEDF Scheudling (Multi-core)

Input: A set of VMs $\{T_i\}(1 \leq i \leq n)$; the frequency of T_i, f_i; the CPU requirement of T_i, $\frac{s_i}{p_i}$); a set of cores $C_j, (1 \leq j \leq m)$;

```
for all VMs do
    Sort the VMs in the ascending order of frequency and obtain the sorted list of
    {T_1, T_2, ..., T_n}, i.e., f_1 ≤ f_2 ≤ ... ≤ f_n
end for
The current core C_c is initialized to be C_1, i.e., c = 1
for T_1 → T_n do
    if VM T_i has not been allocated then
        Calculate the total CPU requirement of VMs allocated to C_c, denoted by Σ_c
        if s_i/p_i + Σ_c ≤ 100%
         then
            allocate T_i to core C_c
        else
            for j = i + 1; j ≤ n; j++; do
                if VM T_j satisfies s_j/p_j + Σ_c ≤ 100% then
                    Allocate T_j to core C_c
                end if
            end for
            c++
        end if
    end if
end for
```

Algorithm 3: The VM allocation method in BFMM

Evaluation

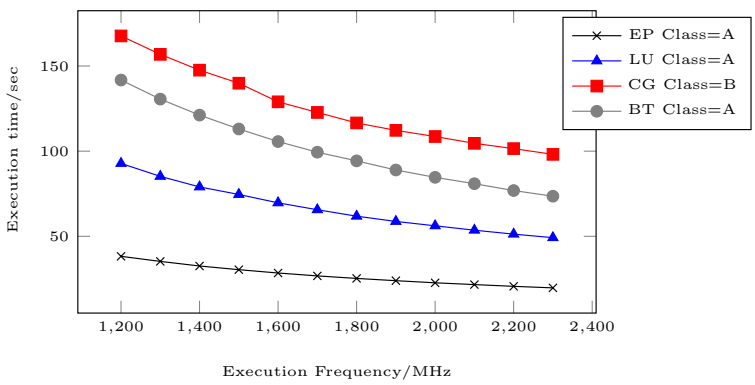
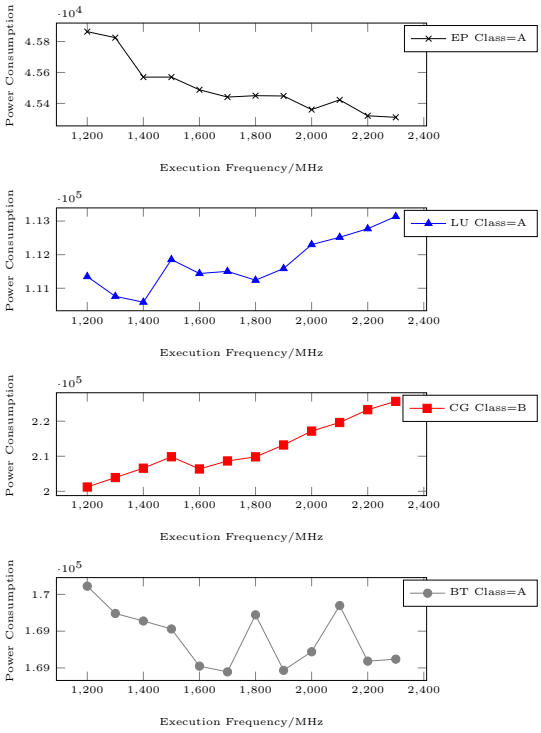


Figure: The execution times of different benchmark tasks running with different frequencies

Evaluation



Experiments on Single-core processors

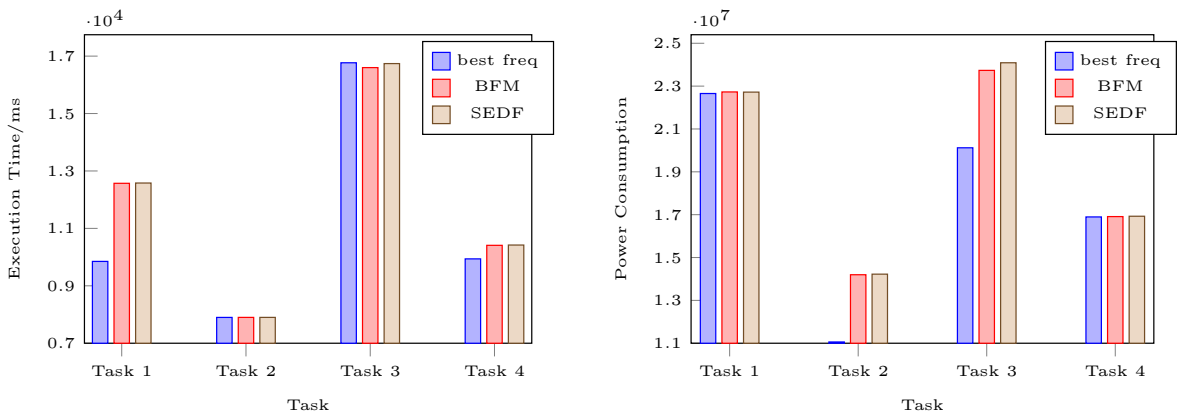


Figure: Execution times and Power consumption of tasks on a single core under SEDF and BFM

Experiments on Multi-core processors

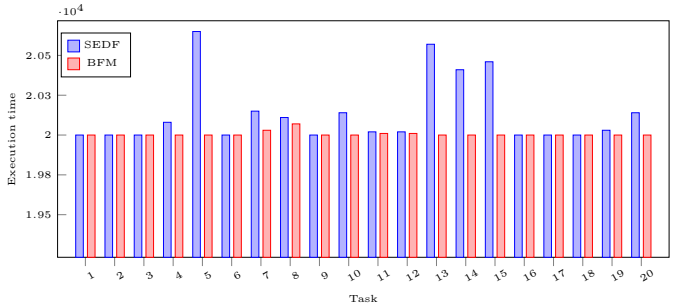


Figure: Execution times of tasks on a multi-core processor under SEDF and BFMM

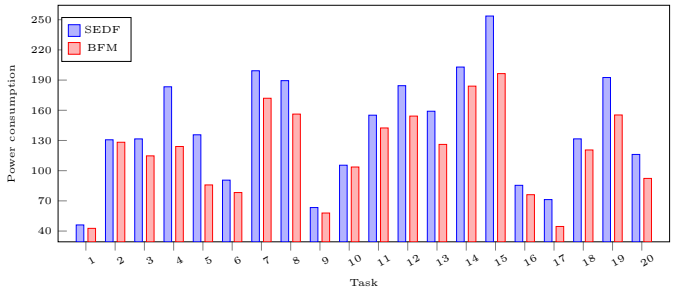


Figure: Power consumption of tasks on a multi-core processor under SEDF and BFMM

Thanks for listening!

Full name: Ade Shonola

Year of Study: PhD Year 4 or later

Department: Computer Science

Submission Type: Poster

Title: Security Issues in M-learning

Abstract: This study aims to examine the security threats or concerns that users may face when using mobile devices for educational purposes. While the challenges of adopting mobile learning in universities are enormous, this study identifies the critical security challenges. It also seek to discover the components of m-learning environment that are vulnerable to risks, threats and attacks.

Keywords: Mobile security, m-learning, mobile device

Abstract Word Count (Est): 57

Security Issues in M-learning

Ade Shonola, Mike Joy
Department of Computer Science, University of Warwick

WARWICK
THE UNIVERSITY OF WARWICK

Introduction

Mobile learning is becoming popular among learners and educators due to rapid growth in academic technologies. The use of mobile and portable devices in education has potential to motivate new approaches to learning and the prospects of implementing m-learning are numerous ranging from flexibility to portability and adaptability. While educators are using mobile devices as teaching aids, students are using them as learning tools. Academic researchers are also using their portable gadgets for collaboration. Handheld devices, however can potentially become vulnerable and pose serious threats to confidentiality, integrity and privacy of users if adequate consideration is not given to the security aspects of mobile learning.

Motivation

This study aims to examine the security threats or concerns that users may face when using mobile devices for educational purposes. While the challenges of adopting mobile learning in universities are enormous, this study identifies the critical security challenges. It also seek to discover the components of m-learning environment that are vulnerable to risks, threats and attacks.

Main Research Questions

1. What are the threats including security that might affect the implementation of mobile learning in higher education institutions?

2. In what ways are the security of m-learning devices breached?

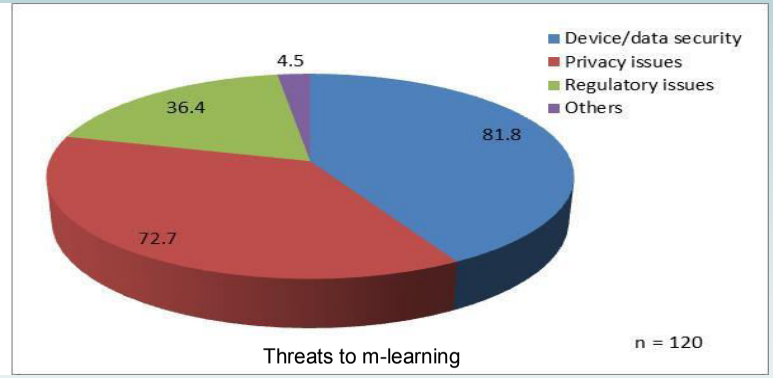
3. How can these security threats or issues be assessed and reduced?

4. What are the responsibilities of the stakeholders in ensuring risk free m-learning?

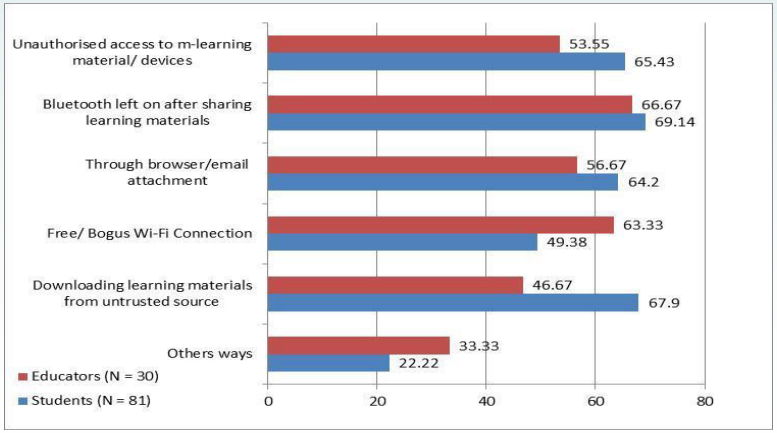
Methods

- ❑ Reviewing academic literature, research based articles and journal publications relevant to the case study (Universities in Nigeria)
- ❑ Collecting quantitative data through random sampling of interested participants using BSREC approved questionnaire.
- ❑ Exploring qualitative research methodology by interviewing experts and academics in m-learning and computer security field.

Result Analytics

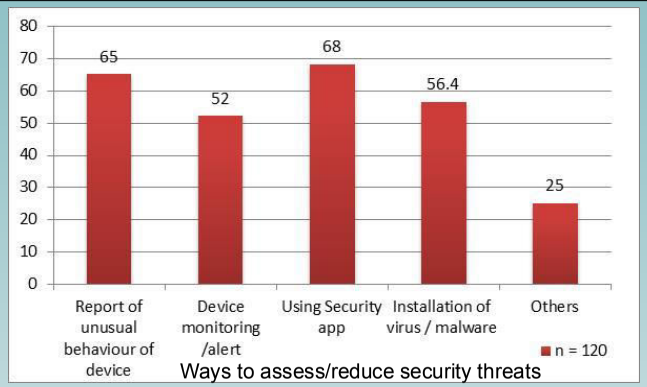


Threats to m-learning



How security is breached

Result Analytics Contd.



Ways to assess/reduce security threats

Conclusion

M-learning promoters (tutors and education institutions) should focus not only on developing m-learning content but also take responsibilities in ensuring threat free m-learning environment by creating awareness among users as well as ensuring up-to-date mobile devices are used for learning.

Future research

In order to increase security in m-learning, future research work is to develop security awareness and enhancement app for android mobile devices. The purposes of the app include:

- ❑ Security Awareness rising
- ❑ Vulnerabilities Scanning
- ❑ Threats Reporting

References

Shonola, S. A and Joy, M.S (2015) 'Security of m-learning System: A Collective Responsibility'. International Journal of Interactive Mobile Technologies (iJIM) pp.64 – 70 http://dx.doi.org/10.3991/ijim.v9i3.4475

M-learning Security Enhancement App

Ade Shonola, Mike Joy
Department of Computer Science, University of Warwick

WARWICK
THE UNIVERSITY OF WARWICK

Introduction

Mobile devices have been playing vital roles in modern day education as students can access or download learning materials on their smartphones and tablets, they can also install educational apps on their devices and study at anytime, anywhere. However, many mobile apps are not secure and some users are unaware of the growing mobile threats. There have been many security incidents with mobile gadgets ever since these devices are becoming popular and widely used, most especially in open operating systems (La Polla, et al., 2013). The need to provide adequate security for portable devices being used for learning cannot be underestimated.

Objective

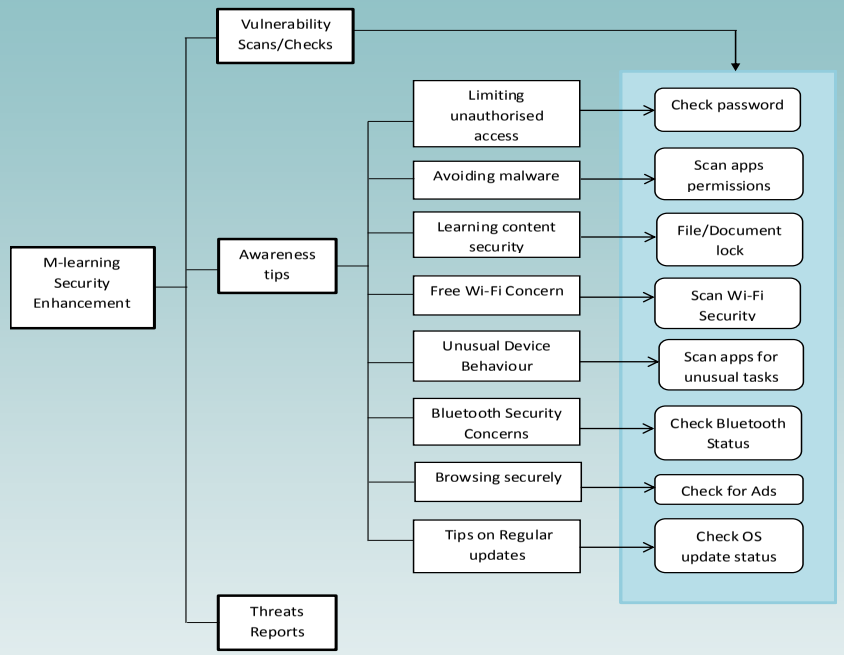
M-learning Security Enhancement app is designed and developed for android smart mobile devices to enhance the security of m-learning systems for end-users, most especially the students. The app promotes security awareness among learners by informing them about security issues and providing tips on how to avoid them . The app also identifies some significant security weaknesses in learning devices by scanning for vulnerabilities and presenting reports on the security threats that are identified.

Security Enhancement Activities

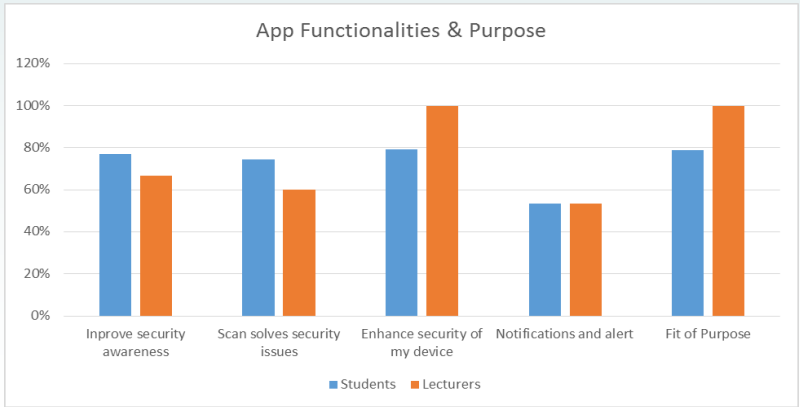
Once the app is installed on a device, user can:

- ❑ Start the app and read security awareness tips in each section. This raises their knowledge on security issues.
- ❑ Scan/check for any vulnerability in their m-learning device at appropriate section.
- ❑ Get regular notifications/alerts from the app regarding any security issues and suspicious app by starting the scanner service.
- ❑ Get security reports on weaknesses in their device as well as appropriate recommendations.
- ❑ Avoid further security breaches by following the various recommendations given in the m-learning security enhancement reports from time to time.

Enhancement Model



Evaluation Analytics



Feedback & Conclusion

Some of the feedback on the app functionalities are:

- Many students have gained security knowledge through the use of the app,
- Students get timely notifications from the app regarding any security issues and suspicious app.
- The app gives students opportunity to practice simple security tasks.
- One participant said, 'through the app, I enjoy testing the Wi-Fi connection anywhere I go by checking the Wi-Fi security'.
- Another feedback said 'my knowledge on adware and spyware has improved. I have also installed Advert blocker as recommended, thus prevent advert pop ups on my screen unnecessarily'.
- The app monitors other apps which may be malware or spyware, through its scanner services.
- The app is fit for purpose because it helps to solve some of the security issues that our students have encountered in the past.
- Above all, the app does what it says because it provides extra security checks in addition to normal device security. Thus, the app enhances the in-built security features of m-learning devices.

The development of m-learning security enhancement app was necessary in order to raise students' awareness on security, augment existing security in m-learning devices and provide information on reducing threats in their devices.

Future research

To carry out experimental study on securing copyrighted m-learning materials in form of Digital Right Management (DRM) and if possible, incorporate it in the Security Enhancement app.

References

- La Polla, M., Martinelli, F., & Sgandurra, D. (2013). A survey on security for mobile devices. Communications Surveys & Tutorials, IEEE, 15(1), 446-471.

Full name: Eliot Shore

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Visual SLAM for Automatic Parking Spot Detection

Abstract: Advanced Driver Assistance Systems (ADAS) are technologies designed to improve vehicle safety or the driving experience. They include systems for automating work of the driver, for example adaptive cruise control, and sensory systems such as driver monitoring systems and pedestrian detection systems.

This project is to design a system for a vehicle with front facing cameras to map a car park as it drives around for the automatic detection of empty parking spots. The technique used is visual SLAM (Simultaneous Localisation and Mapping), a research field historically developed for autonomous robot navigation.

SLAM systems work by making sensory observations of the environment to estimate a map (in the case of visual SLAM these will be features detected in each frame of the video). Next, the robot/vehicle position is calculated within that map, and based on an estimate of movement, the new position within its' environment is predicted before updating its' map again. This is a chicken-and-egg problem since position cannot be known without a map, but an accurate map cannot be generated without a position estimate. Furthermore, in a dynamic environment where landmarks move, classical SLAM methods fail. Therefore, performing SLAM in environments where spaces switch between occupied and empty is an active research area.

Current work on the project is to improve feature matching accuracy between images taken with wide-angle lenses.

Keywords: Simultaneous Localisation and Mapping, Wide-angle lens, SIFT, Feature matching

Abstract Word Count (Est): 222

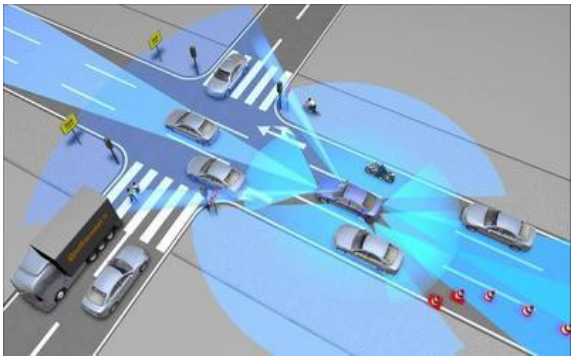
Visual SLAM for Automatic Parking Spot Detection

WARWICK

WARWICK

Smart Technology Environment

- Advanced Driver Assistance Systems (ADAS)
- Autonomous Vehicles



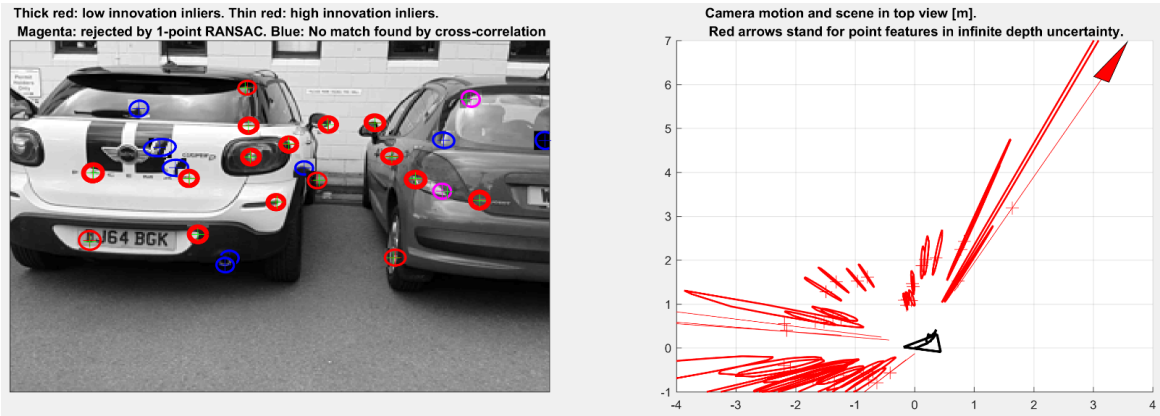
Advanced Driver Assistance Systems

- Adaptive cruise control
- Emergency braking
- Sign recognition
- Pedestrian detection
- Pothole detection
- Car-to-car communication

Jaguar Land Rover, *Pothole Alert*

WARWICK

Parking Spot Detection and SLAM



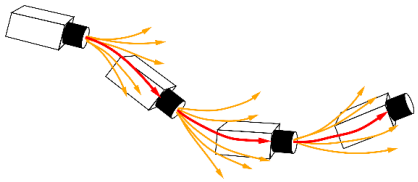
What is SLAM?

- Simultaneous Localisation and Mapping
  - Chicken-and-egg problem
  - Computing a camera's pose whilst mapping its' environment at the same time
  - Historically developed for robot navigation

SLAM Problem

- Given controls:
  - $u_{1:T} = \{u_1, u_2, \ldots u_T\}$
- And observations:
  - $z_{1:T} = \{z_1, z_2, \ldots z_T\}$
- Wanted:- map:
  - m
- And path of travel:
  - $x_{1:T} = \{x_1, x_2, \ldots x_T\}$

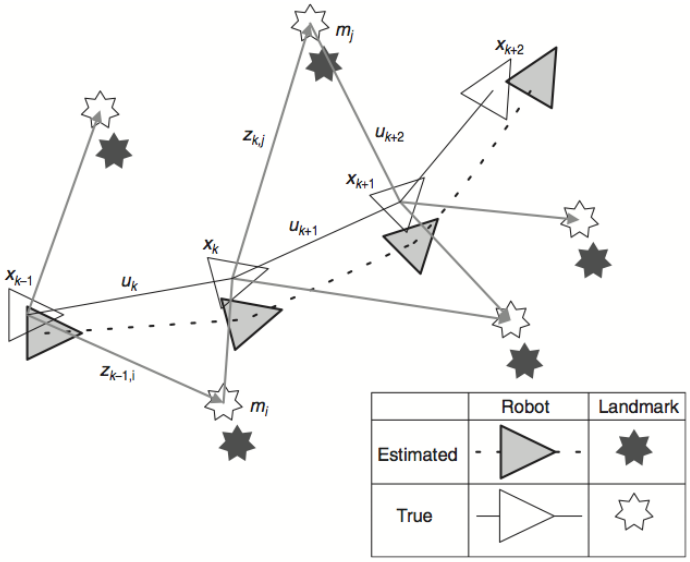
- The map and path estimate are probabilistic:
- $p(x_{0:T}, m \mid z_{1:T}, u_{1:T})$



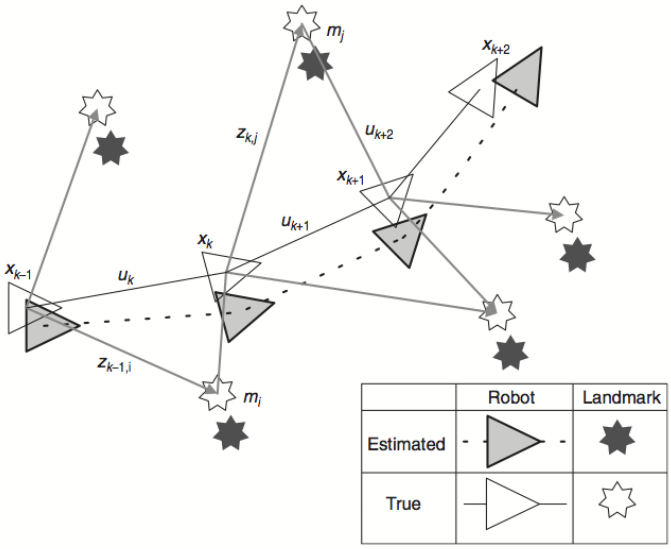
Davison (2004), *Real-time 3D SLAM with wide-angle vision*

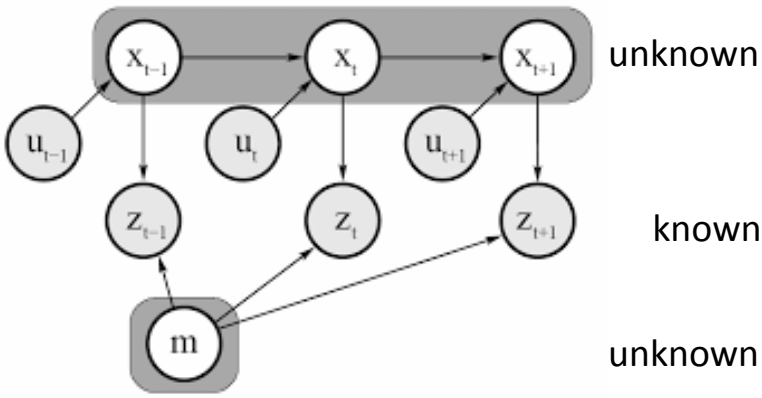
SLAM Process

1. The sensor observes its environment and measures the location of landmarks relative to itself.
2. The sensor is moved and updates its new estimated pose based on some form of motion prediction data.
3. The sensor observes and measures the position of landmarks relative to itself.
   -The sensor is not where it was expected to be.
4. Update its estimate of where it is based on its new observations.
5. The updated pose estimate should be more accurate than its original prediction.



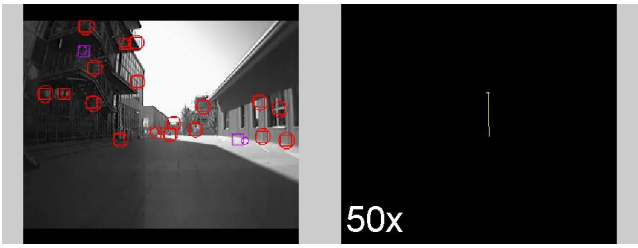
Graphical Model





Arrows represent dependencies between variables

What is SLAM?



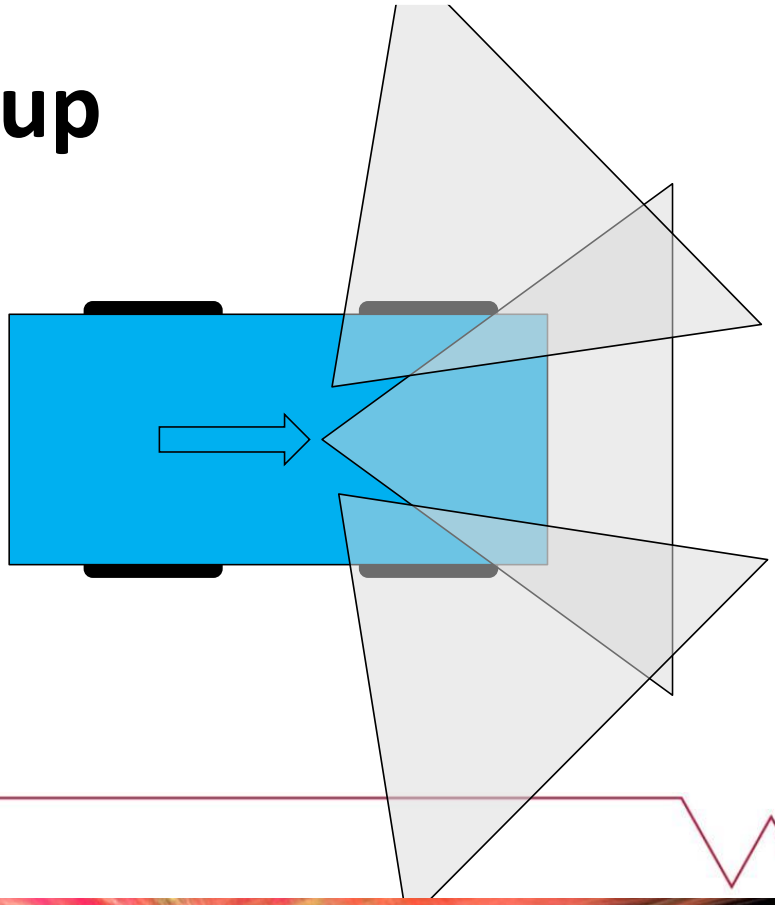
Civera (2010), *1-Point RANSAC for EKF Filtering- Application to Real-Time Structure from Motion and Visual Odometry*

Why is SLAM Hard?

- Uncertainty increases with motion
- Data association is not trivial
  - Depends on what sensor data is available
  - Visual SLAM uses a camera
  - Dependent on matching image features

Camera Setup

- Cameras mounted on car
- Partially overlapping fields of view

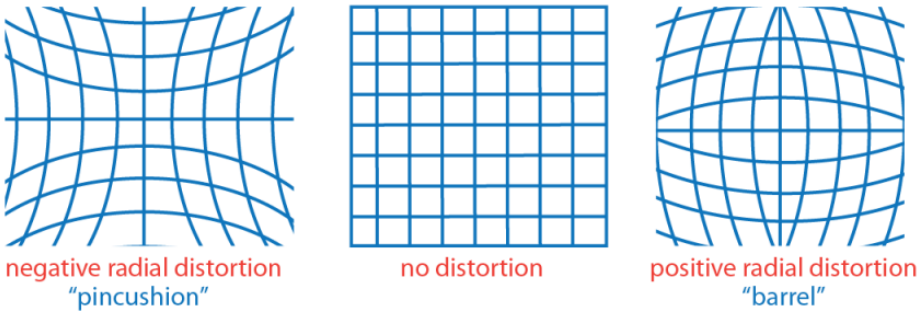


Current Work – Wide Angle Vision

Lenses and Distortion

Distortion Correction

$$u_d = f_x(u, v)$$
$$v_d = f_y(u, v)$$



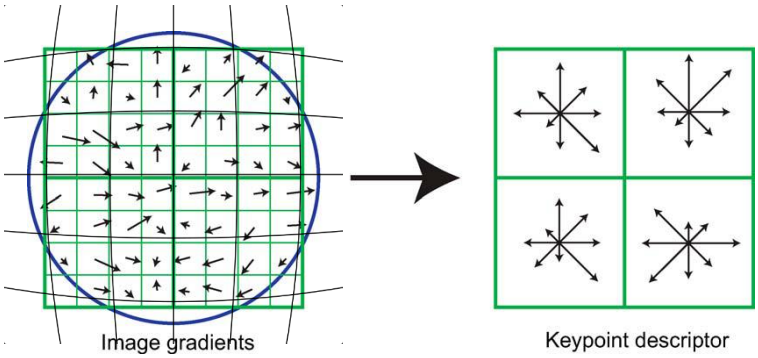
Mathworks, *Camera Calibration*

$$u = f_x^{-1}(u_d, v_d)$$

$$v = f_y^{-1}(u_d, v_d)$$

SIFT Local Image Gradients

$$\mathsf{J_f} = \begin{bmatrix} \frac{\partial u_d}{\partial u} & \frac{\partial u_d}{\partial v} \\ \frac{\partial v_d}{\partial u} & \frac{\partial v_d}{\partial v} \end{bmatrix}$$



Adapted from Lowe (2004) *Distinctive Image Features from Scale-Invariant Key points*

Questions?

Full name: Hasliza Sofian

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Random Slot Selection for TDMA-based MAC Protocol in Distributed WSN

Abstract: All sensors in WSN are expected to manage their own network. Most of the time, sensors have no data to send and are switched into sleep mode to conserve energy. Decent MAC protocols are responsible\xa0to manage the data transmission over the network. Typical multi-hop setups employ distributed or centralized network to disseminate data. Focusing on adaptive TDMA-based MAC protocol slot-allocation for distributed topology, TRAMA and TDMA-W adapt random slot selection for initial slot before the nodes contending for specific slot with other sharing-slot nodes. The loser nodes will simply pick out other slots and the procedure continues until all transmitters have their own unique slots. During high traffic load, more nodes have more data to send in one frame. Choosing a slot randomly causes time and energy inefficient due to a higher number for contention occurred and greater number of control packet have to be transmitted. FLAMA experiments proved higher number of control packet transmission leads to time delay and increases the energy consumption. We study about developing a novel scheme to choose the slot randomly, but within defined slot boundaries. This scheme identifies maximum number of slots that a node may assign to itself, thus, fulfils single or multi-slot demand. With this scheme, the probability of a node within its' 2-hops neighbours to initially choose the same slot decreases, less slot schedule update packets is required, and multi-slot allocation is equally divided among nodes.

Keywords: TDMA, slot-assignment.

Abstract Word Count (Est): 235

Full name: Noor Hasimah Ibrahim Teo

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: ONTOLOGIES FOR AUTOMATIC QUESTION GENERATION

Abstract: Advances in Semantic Web technologies have created interest among researchers in developing ontology-based applications in numerous research areas. One such research area is the ontology-based question generation that aims to benefit instructors by providing support and intelligent assistance to the automatic generation of questions, and hence increasing the amount of staff time available for developing new skills and constructing worthwhile feedback for students. However, current technologies for automatic generation of questions do not work very well and are not very sophisticated. The main objective of this work is to improve the way (in which) question are generated automatically, by incorporating a semantic understanding of the question domain. Useful questions at different question taxonomy are constructed through semantic strategies that use an ontology and template based approach. Initial experimental results have shown the effectiveness of these approaches in generating questions.

Keywords: Semantic, Ontology, Question Generation

Abstract Word Count (Est): 139

Full name: Malkiat Thiarai

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Using data to improve services to vulnerable children and adults

Abstract: Using historical data from Birmingham City Council\x92s social care system, we have looked the range of services provided to vulnerable children, young people and older adults over a 6 year period from 2010-2015

The aim is to understand the data held, its quality and how it can be used to support service delivery in future years

Keywords: Personal data, Children's Services, Local Authority

Abstract Word Count (Est): 56

Warwick Postgraduate Colloquium
in Computer Science
Malkiat Thiarai
July 2016

Using data to improve services to vulnerable children and adults

- Using historical data from Birmingham City Council's social care system, we have looked the range of services provided to vulnerable children, young people and older adults over a 6 year period from 2010-2015

- The aim is to understand the data held, its quality and how it can be used to support service delivery in future years

Financial Challenges

- The research is being used to inform the Council's future service delivery and planning against the backdrop of the financial challenges its faces
- BCC estimates that between 2010 and 2019 it will have saved £815m from its controllable budget

The data

- Derived from personal data and includes 18 attributes, including date of birth, age, ethnicity, religion, full postcode and cost
- On the name of the individual was removed for research purposes
- Details of the service received by the individual were provided

Element types

358 different elements can be provided to a service user depending on an assessment of need

CHEFREFE	CHILDRENS EXT. FOSTERING - RESERVATION FEE
CHEFRESD	CHILDRENS EXT. FOSTERING SIBLING AND RES
CHEFRESO	CHILDRENS EXT. FOSTERING SOLO AND RESERVATION
CHEFRESV	CHILDRENS EXT. FOSTERING RESERVATION FEE
CHEFSDIS	CHILDRENS EXT. FOSTERING - SIBLING DISCOUNT
CHEFSIBL	CHILDRENS EXT. FOSTERING SIBLING DISCOUNT
CHEFSLDI	CHILDRENS EXT. FOSTERING - SIBLING L / T DISCOUNT
CHEFSOFE	CHILDRENS EXT. FOSTERING - SOLO DISCOUNT
CHEFSOLO	CHILDRENS EXT. FOSTERING SOLO PLACEMENT
CHEFSTFE	CHILDRENS EXT. FOSTERING - STANDARD FEE
CHEFSTND	CHILDRENS EXT. FOSTERING STANDARD FEE
CHEFVARC	CHILDRENS EXT. FOSTERING VARIATION CONTINUING

CHEFVARO	CHILDRENS EXT. FOSTERING VARIATION ONE-OFF
CHEFVART	CHILDRENS EXT. FOSTERING VARIATION TRANSFER
CHERDET	CHILDRENS EXTERNAL RES FOR DISABLED WITH EDUCATION
CHERDIT	CHILDRENS EXTERNAL RESIDENTIAL DISABLED HOME
CHERDRT	CHILDRENS EXTERNAL RESIDENTIAL DISABLED RESPITE
CHEREAS	CHILDRENS EXTERNAL RESIDENTIAL ASSESSMENT
CHEREDT	CHILDRENS EXTERNAL RESIDENTIAL WITH EDUCATION
CHEREFAS	CHILDRENS EXTERNAL RES FAMILY ASSESSMENT
CHEREST	CHILDRENS EXTERNAL RESIDENTIAL HOME
CHEREXT	CHILD EXTERNAL RESIDENTIAL EXTRA NEEDS

Element types

DEENLTBL	Nursing Client L/T Dementia External Block
DEENSTBL	Nursing Client S/T Dementia External Block
DEERLTBL	Residential Client L/T Dementia External Block
DEERSTBL	Residential Client S/T Dementia External Block
DEESTNTM	Nursing Client Dementia Ext S/T Bed Based
DEESTRRM	Residential Carer Dementia Ext S/T Bed Based
DEESTRTM	Residential Client Dementia Ext S/T Bed Based
DEETERTM	Residential Client Dementia Ext T/P Bed Based
DEILTRTM	Residential Long Term Client Internal Dementia

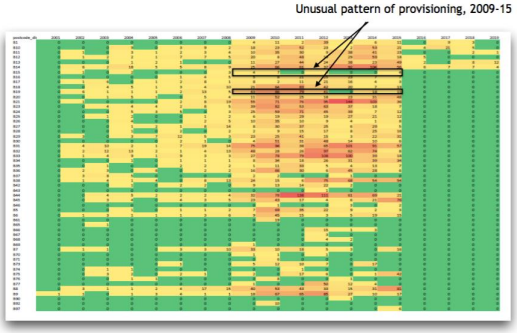
HSSU1865	Home Support 18-64 Ext Community Based
HSSU65PL	Home Support 65 Plus Ext Community Based
HSSUAUTS	Home Support Autism Spectrum Ext Community Based
HSSUBINJ	Home Support Brain Injury Ext Community Based
HSSUDEME	Home Support Dementia Ext Community Based
HSSULDIS	Home Support Learning Disab Ext Community Based
HSSUMACT	Home Support Mental Health Act Ext Community Based

LDEHSECH	Learning Disab Ext. - Extra Care Sheltered Housing
LDEHSGCO	Learning Dis. - Ext. - General Contracted
LDEHSGNC	Learning Dis. - Ext. - General Non-Contracted
LDEHSQDS	Learning Dis External - Quick Discharge Service
LDEHSSCO	Learning Dis. - Ext. - Specialist Contracted
LDEHSSNC	Learning Dis. - Ext. - Specialist Non-Contracted
LDEHSUPO	Learning Dis. - Ext. - Supported Living Outreach
LDEHSUPP	Learning Dis. - Ext. - Supported Living
LDELRET	Learning Dis. - Ext - L/T Residential - Recharge

MHDODAYC	Mental Health - External Day Centre Attendance
MHDOINTC	Mental Health - Internal Day Centre Attendance
MHEBLACT	Mental Health - External - Block Activity
MHEHSGCO	Mental Health External - General Contracted
MHEHSGNC	Mental Health External - General Non-Contracted
MHEHSQDS	Mental Health External - Quick Discharge Service
MHEHSSCO	Mental Health External - Specialist Contracted
MHEHSSNC	Mental Health External - Specialist Non-Contracted
MHEHSUPP	Mental Health Ext Supported Living
MHELRET	Mental Health - Ext - L/T Residential - Recharge
MHELTBL	Mental Health - External Residential - BLOCK

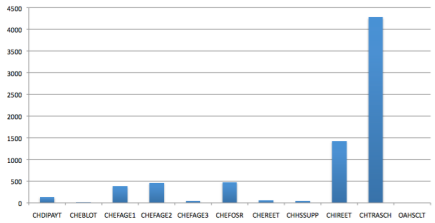
Initial research

Analysis to date - heat map postcode/year

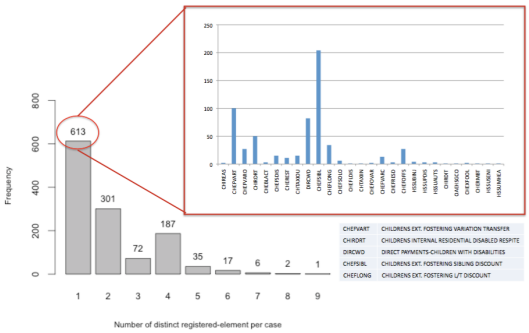


Analysis to date - services registered 2009-14

CHEFAGE1	CHILDRENS EXT. FOSTERING – AGE BAND 1
CHEFAGE2	CHILDRENS EXT. FOSTERING – AGE BAND 2
CHEFOSR	CHILDRENS EXTERNAL FOSTERING
CHIREET	CHILDRENS INTERNAL RESIDENTIAL SERVICE
CHITRASCH	CHILDRENS TRANSPORT TO SCHOOL

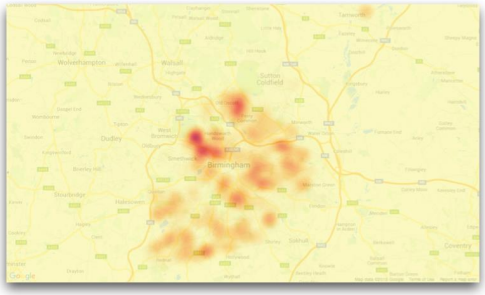


Analysis to date - number of distinct registered elements per case (age 0 - 11)



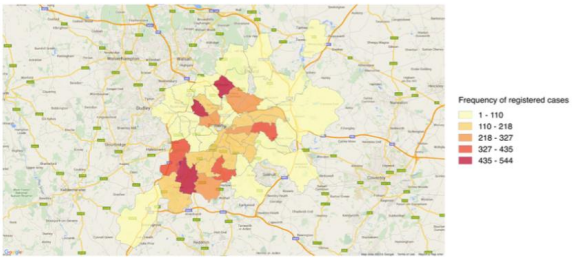
Initial research

Density heat map (age 0 - 11)

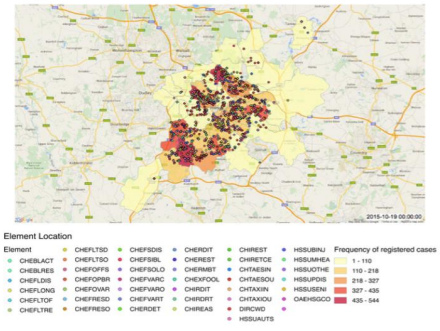


Density heat map (age 0 - 11)

Postcodes of particular interest: B21; B31; B44

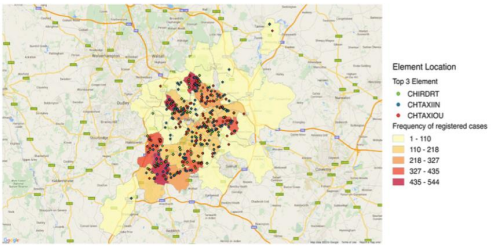


Distribution of elements (age 0 - 11)



Initial research

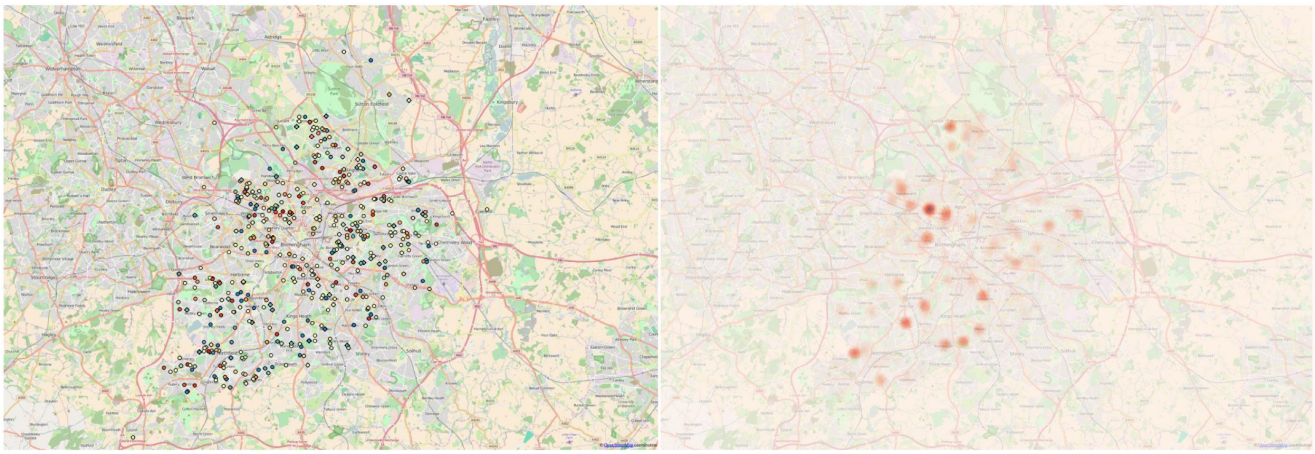
Distribution of elements (age 0 - 11)



Distribution of elements, B44 0, (age 0 - 11)

Distribution of elements, B21 8, (age 0 - 11)

Location of top 20 elements : Age 0-11 (2010-2012)

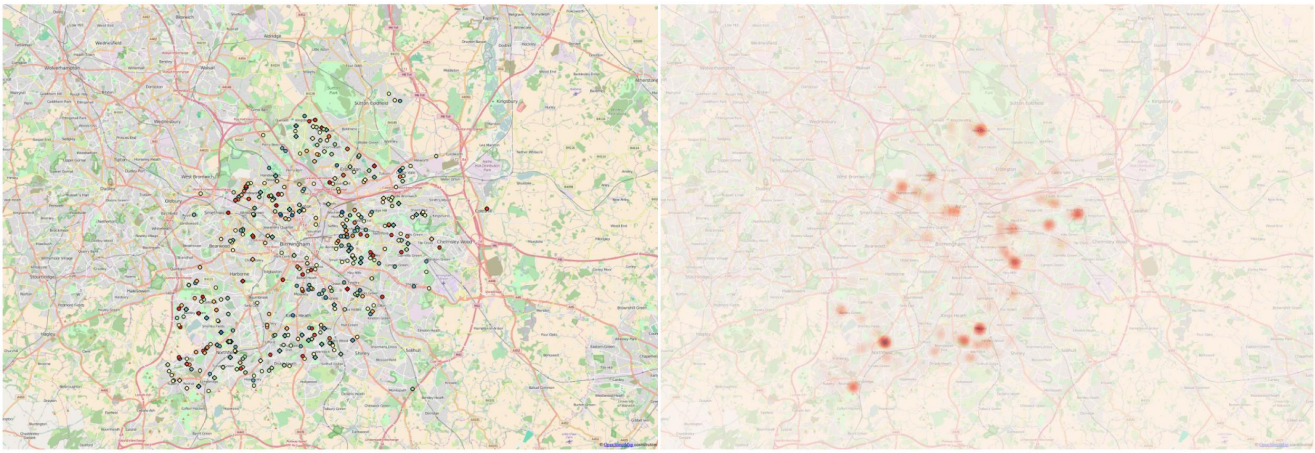


Element Provision

Top20 Element Name 2010-2012 Age 0-11

- CHEBLACT
- CHEFLONG
- CHEFLTSD
- CHEFLTSO
- CHEFOAGE
- CHEFOFFS
- CHEFRESD
- CHEFRESV
- CHEFSIBL
- CHEFSOLO
- CHEFSTND
- CHEFVARC
- CHEFVARO
- CHEFVART
- CHEREST
- CHEXFOOL
- CHIRDIT
- CHIRDRT
- CHIRETCE
- DIRCWD

Location of top 20 elements : Age 0-11 (2013-2015)

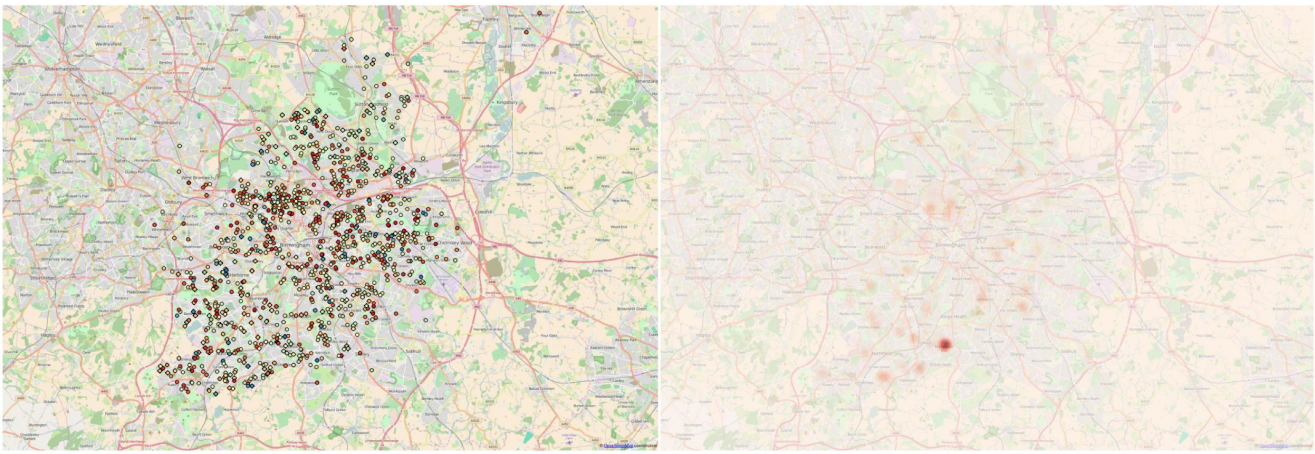


Element Provision

Top20 Element Name 2013-2015 Age 0-11

- CHEBLACT
- CHEFLONG
- CHEFLTRE
- CHEFLTSD
- CHEFLTSO
- CHEFOAGE
- CHEFOFFS
- CHEFRESD
- CHEFRESV
- CHEFSIBL
- CHEFSOLO
- CHEFSTND
- CHEFVARC
- CHEFVARO
- CHEFVART
- CHEREST
- CHIRDRT
- DIRCWD
- HSSUAUTS
- HSSUBINJ

Location of top 20 elements : Age 11-25 (2010-2012)

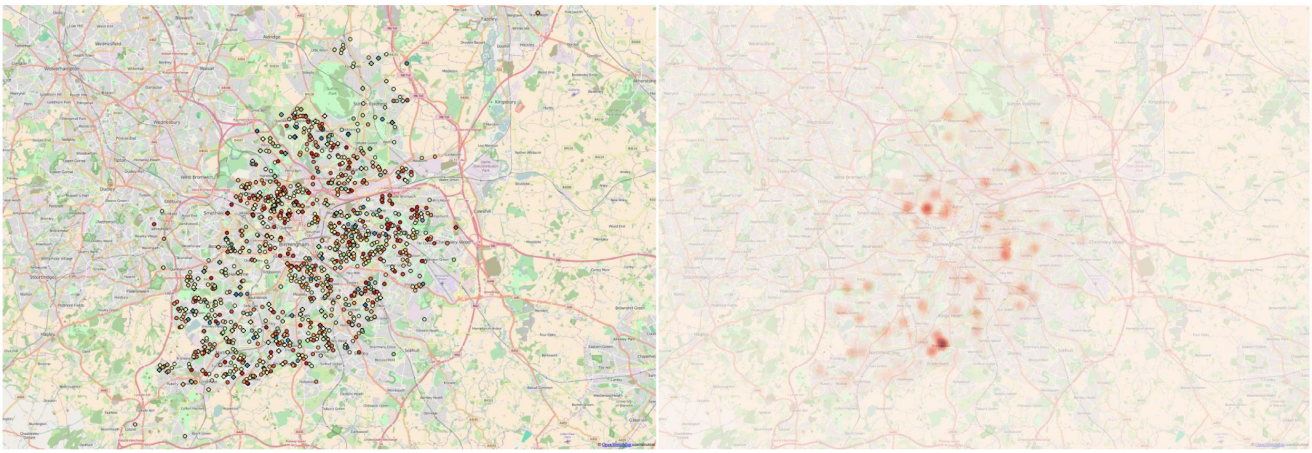


Element Provision

Top20 Element Name 2010-2012 Age 11-25

- CHEBLACT
- CHEFLONG
- CHEFSTND
- CHEFVART
- CHEREST
- CHERSET
- CHIRDRT
- CHIREAS
- CHIREST
- CHIRETCE
- CUASCACV
- DIRCWD
- DIRPAY
- LDDODAYC
- LDESTRR
- LDESTRT
- LDESVRT
- LDETVRT
- LDISTRR
- LDISTRT

Location of top 20 elements : Age 11-25 (2013-2015)

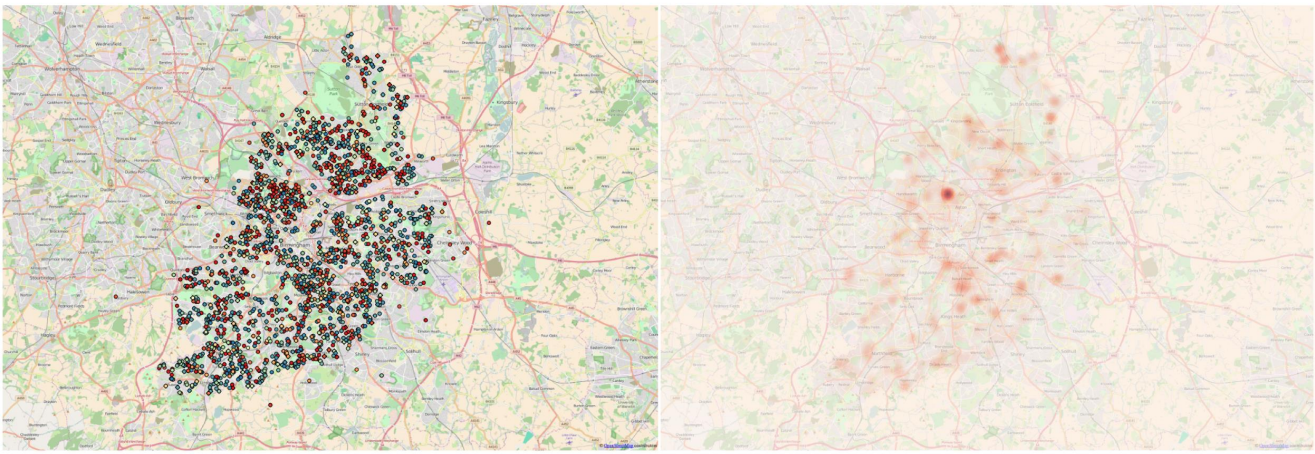


Element Provision

Top20 Element Name 2013-2015 Age 11-25

- CHEBLACT
- CHEFLONG
- CHEFOAGE
- CHEFSIBL
- CHEFSTND
- CHEFVARC
- CHEFVART
- CHEREST
- CHERPRYT
- CHERSET
- CHIRDRT
- DIRCWD
- DIRPAY
- HSSUAUTS
- HSSULDIS
- HSSUPDIS
- LDESTRT
- LDESVRT
- LDISTRT
- PDESVRT

Location of top 20 elements : Age 25-65 (2010-2012)

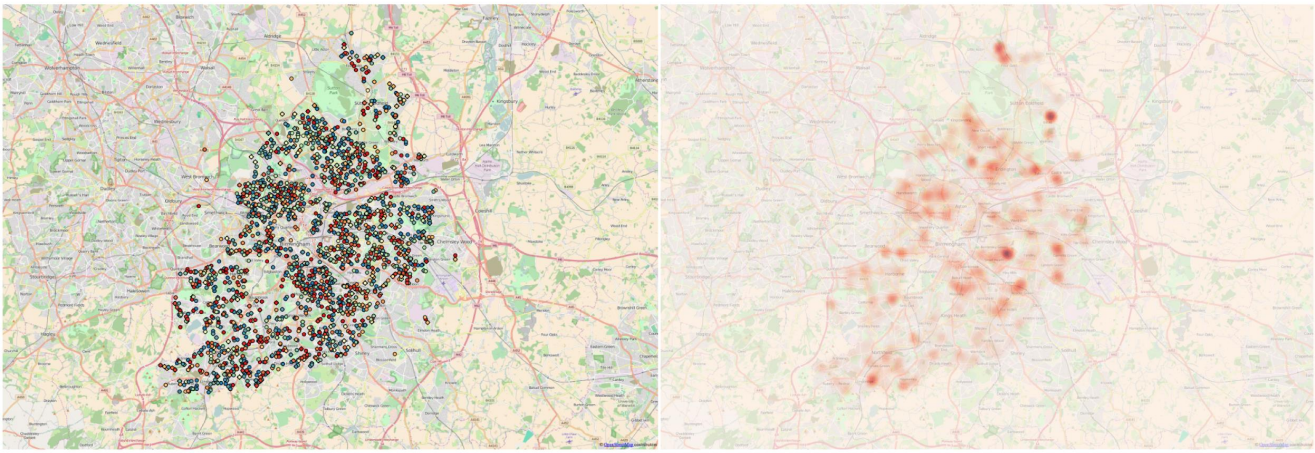


Element Provision

Top20 Element Name 2010-2012 Age 25-65

- DIRPAY
- LDDODAYC
- LDDOINTC
- LDEHSGCO
- LDELVRT
- LDEPPCNO
- LDEPPCOH
- LDEPPFLE
- LDESTRR
- LDESTRT
- LDESVRR
- LDESVRT
- LDISTRR
- LDISTRT
- MHEHSSCO
- OAEHSGCO
- PDDODAYC
- PDEHSGCO
- PDIHSGEN
- SMETVRT

Location of top 20 elements : Age 25-65 (2013-2015)

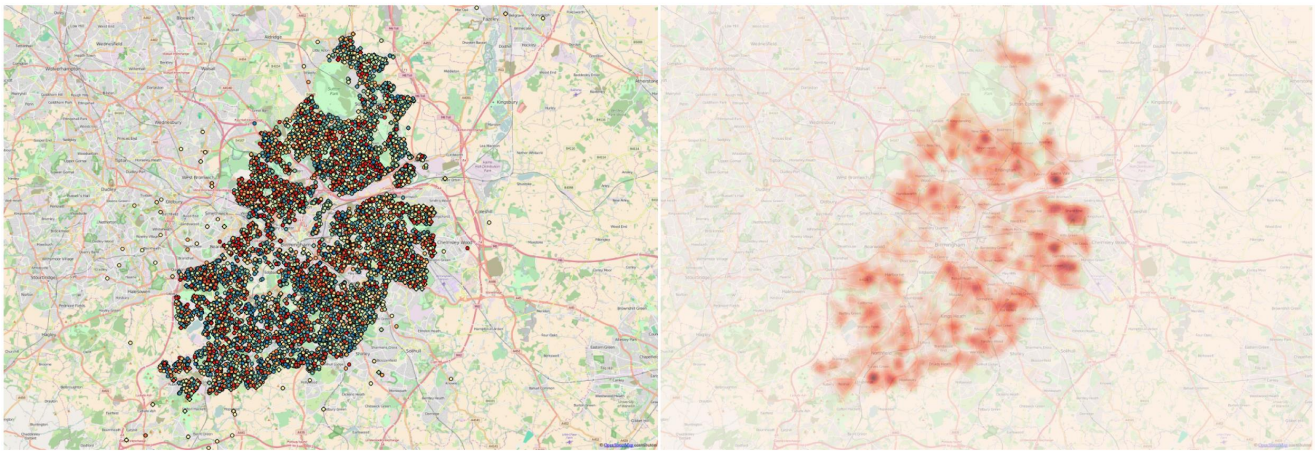


Element Provision

Top20 Element Name 2013-2015 Age 25-65

- DIRPAY
- HSSU1865
- HSSU65PL
- HSSULDIS
- HSSUMHEA
- HSSUPDIS
- LDEHSUPP
- LDESTRR
- LDESTRT
- LDESVRT
- LDISTRR
- LDISTRT
- LDITERT
- OAEHSQDS
- OAIHSENB
- OAIHSENH
- PDEHSGCO
- PDEHSQDS
- PDIHSENH
- PDIHSGEN

Location of top 20 elements : Age 65-90 (2010-2012)

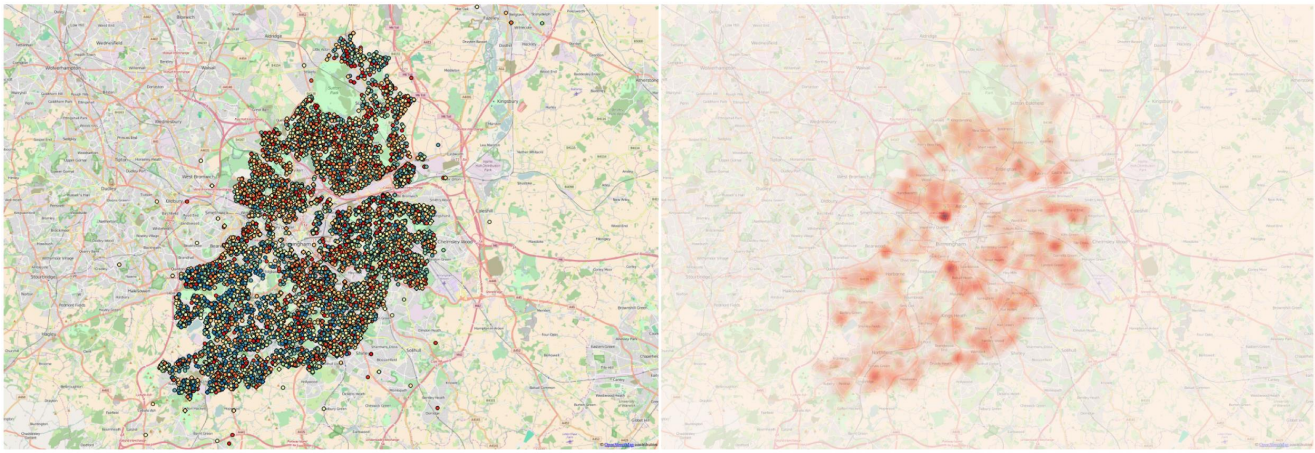


Element Provision

Top20 Element Name 2010-2012 Age 65-90

- DIRPAY
- HSSU65PL
- OADODAYC
- OADOINTC
- OAEHSECS
- OAEHSGCO
- OAEHSQDS
- OAEHSSCO
- OAELTNT
- OAELTRT
- OAELVRT
- OAESTNT
- OAESTRT
- OAICEXT
- OAIHSECS
- OAIHSENB
- OAIHSENH
- OAIHSGEN
- OAISTRT
- PDEHSGCO

Location of top 20 elements : Age 65-90 (2013-2015)



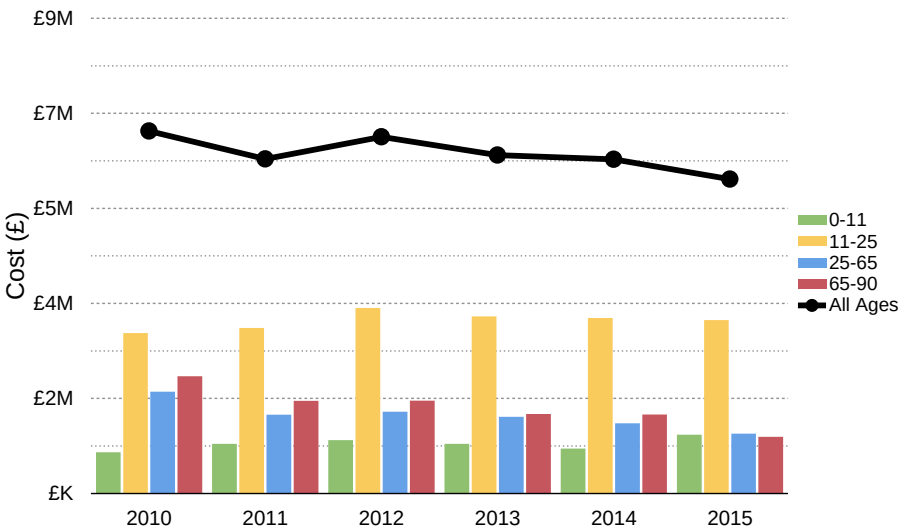
Element Provision

Top20 Element Name 2013-2015 Age 65-90

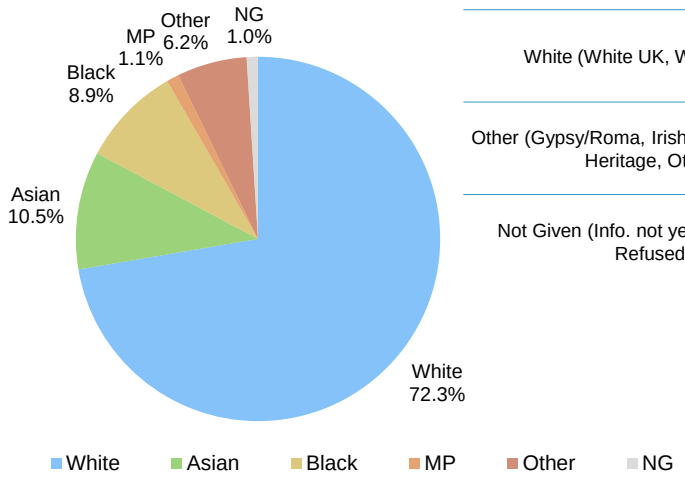
- DIRPAY
- HSSU65PL
- OADODAYC
- OADOINFC
- OADOINTC
- OAEHSECH
- OAEHSGCO
- OAEHSQDS
- OAELTNT
- OAELTNTM
- OAELTRT
- OAELTRTM
- OAENBACT
- OAESTRT
- OAESTRTM
- OAICEXT
- OAIHSENB
- OAIHSENH
- OAIHSGEN
- OAISTRT

Annual cost of top 20 elements

	Cost (£)
2010	6,671,860
2011	6,159,413
2012	6,566,631
2013	6,228,763
2014	6,151,460
2015	5,787,769
Total	37,565,896



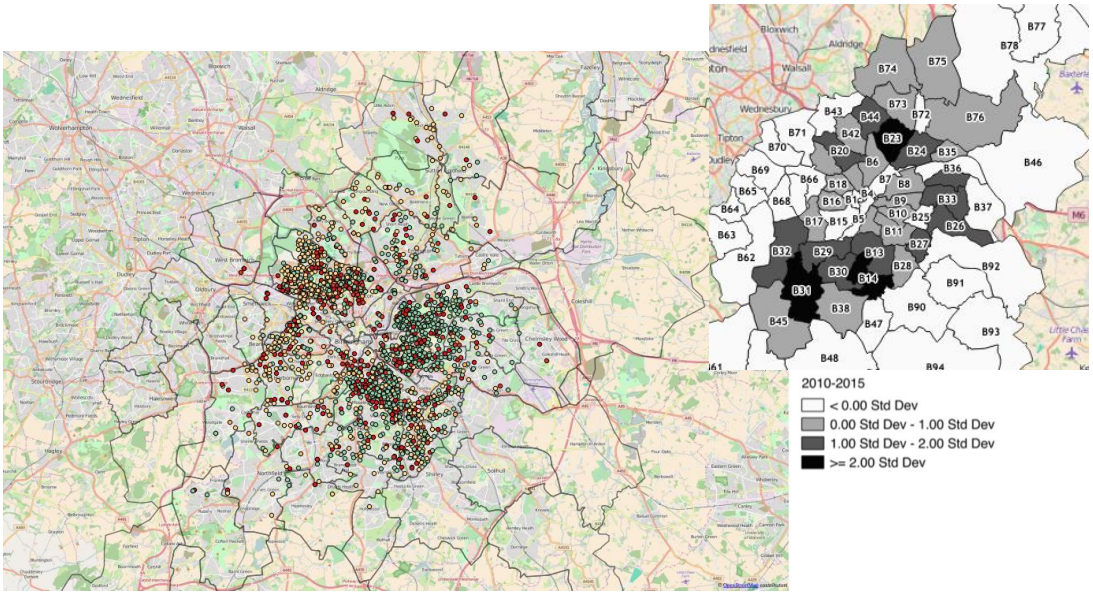
Ethnicity Profile



	Number of Unique People	Per Cent
Asian (Bangladeshi, Chinese, Indian, Pakistani, Asian Other)	2890	10.46
Black (Black African, Black African Caribbean, Black - Other)	2460	8.9
Mixed Parentage (White+Asian, White+Black African, White+Black Caribbean, Other Mixed Background)	316	1.1
White (White UK, White Other)	19953	72.25
Other (Gypsy/Roma, Irish, Traveller of Irish Heritage, Other)	1710	6.2
Not Given (Info. not yet obtained and Refused)	289	1

Example of ethnicity profile

All Asian from 2010-2015 (10.5%)

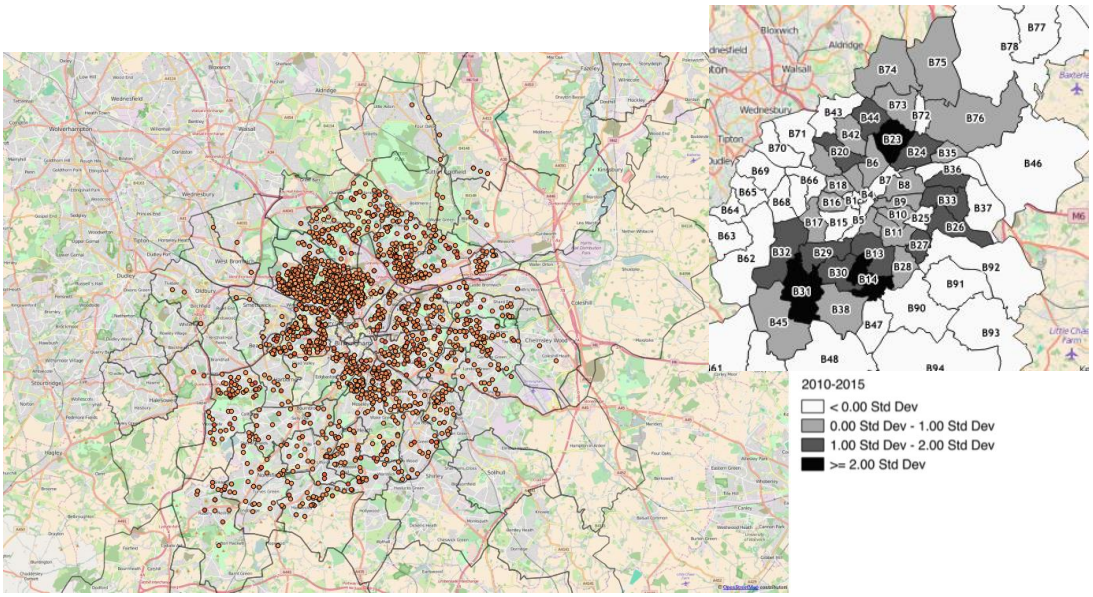


Ethnic Groups

- Asian Other
- Bangladeshi
- Black - Other
- Black African
- Black-African Caribbean
- Chinese
- Gypsy/Roma
- Indian
- Information not yet obtained
- Irish
- Mixed Parentage - Other Mixed Background
- Mixed Parentage - White and Asian
- Mixed Parentage - White and Black African
- Mixed Parentage - White and Black Caribbean
- Other
- Pakistani
- Refused
- Traveller of Irish Heritage
- White - Other
- White - UK

Example of ethnicity profile

All Black from 2010-2015 (8.9%)



Ethnic Groups

- Asian Other
- Bangladeshi
- Black - Other
- Black African
- Black-African Caribbean
- Chinese
- Gypsy/Roma
- Indian
- Information not yet obtained
- Irish
- Mixed Parentage - Other Mixed Background
- Mixed Parentage - White and Asian
- Mixed Parentage - White and Black African
- Mixed Parentage - White and Black Caribbean
- Other
- Pakistani
- Refused
- Traveller of Irish Heritage
- White - Other
- White - UK

How the research is being used

- From the data extracted a mapping tool has been created and shared with the Council to use to carry out more detailed analysis
- The level of detail about the individual and the location of the service allows for much more granularity of the analysis
- By retaining the data ‘in-house’ we are able to use far more personally identifiable data to better understand what services are provided, where, to who and at what cost over any period of time in the 6 years

How the research is being used

- Using historic data we can begin to identify past trends and understand how demand in the future may emerge
- Demonstrate the value of data held by public authorities, in particular, personal data
- Look at data held about assessments over the same period – of which there are about 2.7m records to identify how demand is met and where it is not

Emerging challenges and issues

- Further use of personal data for research and analytical purposes needs to address issues of privacy, governance and ethics
- Risks include accepting the data at face value without further validation or scrutiny
- Awareness and understanding by citizens of the use of their personal data (and others) in making decisions about meeting their needs

Thank you.

Questions.

Full name: Alasdair Thomason

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Context Trees: Augmenting Geospatial Trajectories with Context

Abstract: Exposing latent knowledge in geospatial trajectories has the potential to provide a better understanding of the movements of individuals and groups. Motivated by such a desire, this work explores augmenting trajectories with land usage information to summarise the context behind user actions into a single data structure, called a context tree. We propose a method for context tree construction, and provide concrete implementations for each stage. Through evaluation of the construction method and analysis of the properties of generated context trees, we demonstrate the foundation for understanding and modelling behaviour afforded. Summarising user contexts into a single data structure gives easy access to information that would otherwise remain latent, providing the basis for better understanding and predicting the actions and behaviours of individuals and groups. Finally, we evaluate the utility of context trees through a sample application, that of predicting the future location and contexts of individuals.

Keywords: Clustering, Context, Land Usage, Spatiotemporal Data, Trajectories

Abstract Word Count (Est): 147

Full name: Nataliya Tkachenko

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Poster

Title: Predicting the impact of urban flooding using open data

Abstract: This study aims to explore whether there is a relationship between search patterns for flood risk information on the web and how badly localities have been affected by flood events. We hypothesize that localities where people stay more actively informed about potential flooding experience less negative impact than localities where people make less effort to be informed. Being informed, of course, does not hold the waters back; however, it may stimulate (or serve as an indicator of) such resilient behaviours as timely use of sandbags, relocation of possessions from basements to upper floors and/or temporary evacuation from flooded homes to alternative accommodation. We make use of open data to test this relationship empirically. Our results demonstrate that although aggregated web search reflects average rainfall patterns, its eigenvectors predominantly consist of locations with similar flood impacts during 2014\x962015. These results are also consistent with statistically significant correlations of web search eigenvectors with flood warning and incident reporting datasets.

Keywords: Flooding, information-seeking, Google Analytics

Abstract Word Count (Est): 157

Predicting the impact of urban flooding using open data

Nataliya Tkachenko*[1], Rob Procter[1,2], Stephen Jarvis[1,2]

*@FloodSmartCity; [1] Warwick Institute for the Science of Cities; [2] Department of Computer Science, University of Warwick

ABSTRACT

This study aims to explore whether there is a relationship between search patterns for flood risk information on the web and how badly localities have been affected by flood events. We hypothesize that localities where people stay more actively informed about potential flooding experience less negative impact than localities where people make less effort to be informed. Being informed, of course, does not hold the waters back; however, it may stimulate (or serve as an indicator of) such resilient behaviours as timely use of sandbags, relocation of possessions from basements to upper floors and/or temporary evacuation from flooded homes to alternative accommodation. We make use of open data to test this relationship empirically. Our results demonstrate that although aggregated web search reflects average rainfall patterns, its eigenvectors predominantly consist of locations with similar flood impacts during 2014–2015. These results are also consistent with statistically significant correlations of web search eigenvectors with flood warning and incident reporting datasets.

BACKGROUND

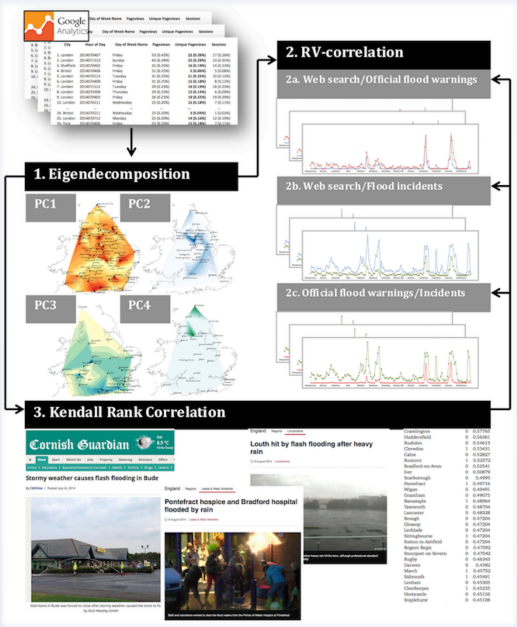
The phenomenon of flooding is extremely complex and subject to change. Incidents are no longer restricted to obvious areas where a river or stream exists; many urban floods are simply caused by huge amounts of rain falling very quickly (flash floods) in an area where the drainage system is unable to cope or due to unexpected underground basin recharge and rise of the groundwater levels [1]. As a consequence, there is an emerging motivation to understand how accurate our knowledge can be about flood risk—its location, timing and duration—and how data collection and analysis can assist us.

To optimize the design of its web-based services, the UK Environment Agency has installed Google Analytics on its live flood warning pages. Interest has therefore emerged in analyzing records of flood warning information seeking, which, coupled with geolocation records, could be potentially useful not only for web designers, but also for flood risk modelers. In this study, we analyse whether web-based information seeking about flood risk can help us understand how badly those locations have been, or may be, affected.

METHODOLOGY

We have developed a three-step analysis (figure below) aimed at testing whether earlier or later engagement with flood risk information on the web is correlated with actual hazard outcomes in cities with similar web search behaviour patterns.

First, we analysed spatio-temporal patterns in the annual Google Analytics dataset, in order to connect information needs to the geography of human activity on the web. The procedure of eigendecomposition, which is often referred to as principal component analysis or nonlinear dimensionality reduction applied to time-series data, is a well-recognized and understood pattern recognition technique. Decomposed in this way, the data allow us to subsequently determine the basic dimensions of the relationships between, for example, information-seeking activity in a number of cities and other factors, such as rainfall patterns, official flood warnings and rates of flood-related incidents in those locations.



We then visualized relationships between official flood risk information issued to the public, reported incidents due to surface water flooding to properties and flood-related information seeking on the web. For this purpose, we constructed three data matrices and employed paired RV (matrix correlation) analysis.

Finally, we used Kendall's τ non-parametric measure of correlation between four pairs of two ranked variables: each of PC1–4 and corresponding binary values, which encode whether that locality had been recorded as affected or severely affected by flooding during the study period April 2014–March 2015.

RESULTS

We found that early engagement with flood risk information correlates with less severe hazard outcomes in the set of locations with similar search behaviour patterns, which are defined here as *eigencities*. Although total search volumes globally correspond to national average rainfall patterns, it can be argued that different principal components are responses, not necessarily linear, to rainfall patterns at different spatial scales: from national to regional and to the most local. In such instances, principal components, or *eigencities*, represent a surrogate measure of the relationship between actual natural phenomena (e.g. precipitation levels) and local knowledge/habits/ preferences of how to deal with the prospective or ongoing flood event.

	N l α≥0.5	Warnings	Reporting	W:R	τ
PC1	96	0.372*	0.533**	0.211 NS	0.283***
PC2	37	0.187 NS	0.317*	0.401*	0.054 NS
PC3	41	0.643**	0.796**	0.550**	(-0.266)**
PC4	30	0.408*	0.191 NS	0.195 NS	(-0.044) NS

* p<0.1, ** p<0.05, *** p<0.001

REFERENCES

[1] Handmer J and Proudley B (2007) Communicating uncertainty via probabilities: the case of weather forecasts. *Environmental Hazards* 7: 79-87

ACKNOWLEDGEMENTS

The lead author gratefully acknowledges funding by the UK Engineering and Physical Sciences Research Council (grant no. EP/L016400/1), the EPSRC Centre for Doctoral Training in Urban Science. This project is conducted in collaboration with the British Geological Survey (BGS) and the authors are grateful for their support in this research.

Full name: Adam Tsakalidis

Year of Study: PhD Year 2

Department: Computer Science

Submission Type: Presentation

Title: Mood Prediction Using Social Media and Mobile Phone Data

Abstract: Recent developments in Natural Language Processing have provided deeper insights in understanding, modelling and predicting human behaviour as revealed online. In this context, mood prediction \x97 predicting someone's mood based on the content that she shares online \x97has attracted a lot of interest lately, in an attempt to create online sensors that can capture the mood of a population in real time. However, in most cases this task is studied on the large scale (e.g., a timeline of all tweets within a country). Such macro-level approaches benefit from the availability of large-scale data. However, they have two main drawbacks: (a) they are user-agnostic and (b) they are assuming that a document (e.g., a tweet) classified as 'happy' implies that the author of the document is feeling 'happy'. In the current work, a different (micro-level) approach is used to tackle these issues. During a total period of 12 months, we have been monitoring the social media accounts and SMS messages of 28 subjects, who have been completing two well-established psychometric tests on a daily basis. We show that the mood of a user as defined by the two tests can be predicted using strictly the text that she has shared (R\x882=.62) and we present a way of incorporating other mobile phone-related data in our analysis to further boost our performance. We aim at applying our approach to the city-level, moving from a well-defined micro-level to the macro-level analysis.

Keywords: mood prediction, text mining, social media

Abstract Word Count (Est): 237

Full name: Ian Tu

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: Occupant State Monitoring

Abstract: Determining vehicle occupant state has become a very active field as of late due to the arrival of autonomous cars and intelligent transport systems. As a result, the vehicle industry is striving to create a driving environment where vehicle systems are smartly fitted and controlled to allow for a safer and more pleasant driving experience. To enable this to happen, vehicle occupant state monitoring systems should be developed to help analyse, model and optimise the driving experience. This talk aims to review the current state-of-the-art vehicle occupant monitoring systems, their general framework, and also present any recent advances in the field of vehicle occupant state monitoring.

Keywords: Occupant State, Vehicle Occupant, Computer Vision, Machine Learning

Abstract Word Count (Est): 106

[Occupant State Monitoring]

Ian Tu

Introduction

What is “Occupant State Monitoring”?

- A system that collects observable information about vehicle occupants
- Uses data to assess occupant’s current behavioural state in order to adjust to personal preferences and achieve optimal vehicle action
- Applications:
  - Active safety – Alert and adjust to activity that could influence the occupant’s safety
  - Improve in-car experience – Link current mood/activity to a preferable journey
- Why bother?
  - Now feasible
  - Technological advances
  - Advent of autonomous cars

Behavioural State Analysis
Categories

- **Distraction**
  - Visual
  - Cognitive
  - Auditory
  - Biomechanical
- **Fatigue**
  - Local Physical
  - General Physical
  - Central Nervous
  - Mental

Behavioural State Analysis
Effects

- **Distraction**
  - Narrow outward inspection, more time looking in single direction
  - Pupil dilation, increased heart rate, increased skin temperature
  - Irregular vehicle control, erratic speed, drifting

- **Fatigue**
  - Less frequent head motions, increased face touching/scratching, yawning
  - Increased blink rate, increase percentage time eye closed
  - Decreased steering performance, more lane deviation

Detecting Occupant State

Measures of occupant state

- **Subjective report** - Karolinska Sleepiness Scale
- **Biological** – Heart Rate, EEG, ECG, EOG
- **Physical** – PERCLOS, head pose, eye gaze, posture
- **Vehicle performance** – Seat pressure, steering wheel position, accelerator position, vehicle telemetry
- **Hybrid –** Mixture of the above: Eye gaze + Vehicle data, Head Pose + Vehicle Data, Eye Gaze + Head Pose + Pupil Dilation + Heart rate

Machine learning

Non-linear models

Human state not linear so non-linear techniques are required:

- Neural Networks
- Support vector machines
- Boosting
- Bayesian Networks
- Hidden Markov Models

Methodology

General System Framework

- **Sensor** – Cameras, vehicle sensors, GPS, heart rate monitors, …
- **Sensor Output Post Processing** – Background removal, noise removal, …
- **Feature Extraction** – Face, Eyes, Mouth, Arms, …
- **Classifier** – SVM, NN, etc.
- **Occupant State Result**
- **Vehicle Action**

Challenges

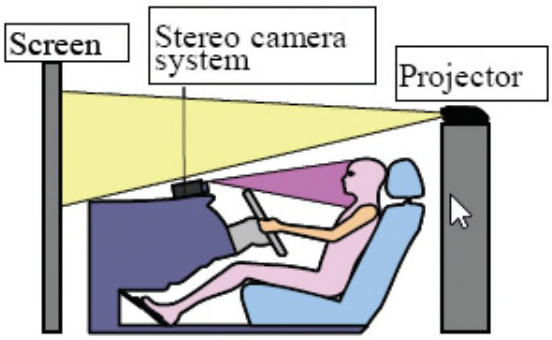
Why is it still an open problem?

- **Occupant differences** – facial features, body posture, etc.
- **In-car conditions** – passengers, layout & setup, noise, temperature
- **Outdoor conditions** – night, day, overcast, wet, humid, hot
- **Camera systems** – where to put cameras, what type of cameras, camera cost
- **Methodology** – no set procedure, some work well under constrained conditions but are not robust under real driving conditions

State-of-the-art

Miyaji et al. – Stereo Camera System

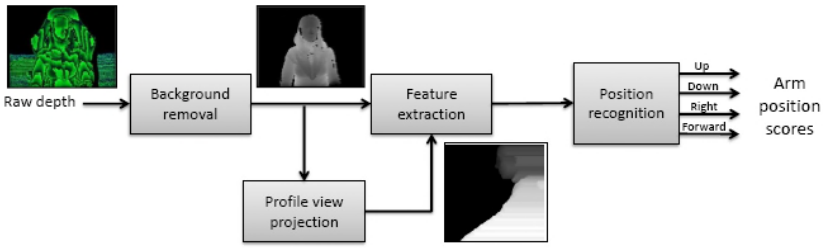
- Cognitive loads were conversation (describing a route they took) and arithmetic (subtracting 7 continually starting at 1000)
- Used the standard deviations of the amount of eye movement and the amount of head movement
- Average value of the pupil diameter, pupil dilation under cognitive load
- Average value of the heart rate R-waves on ECG, decreases under load
- AdaBoost/SVM



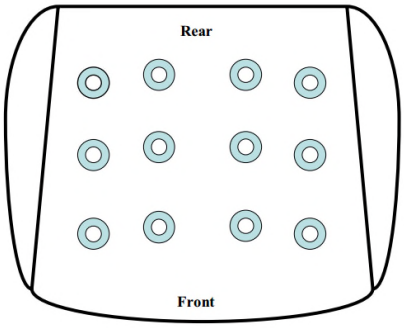
State-of-the-art

Craye et al. – RGB Depth Sensor

- Kinect Sensor
- Features:
  - Eye State
  - Head Pose
  - Facial Expression
  - Arm Position
- AdaBoost/HMM Classifier
- Driving Simulator



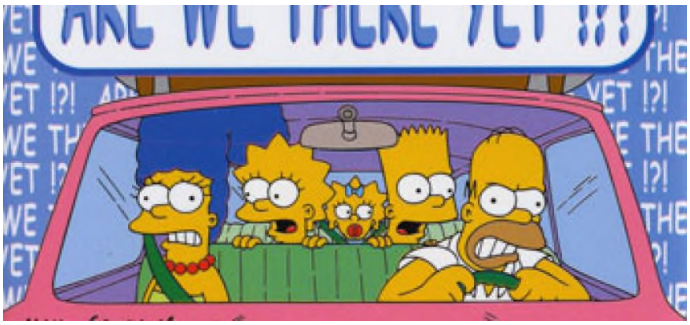
State-of-the-art
Occupant Classification System



- Kennedy et al:
  - Classify into 3 categories adults, children under 6 and empty seat
  - Seat pressure detectors
  - Neural Networks
- Peréz-Jiménez et al:
  - Face and seatbelt detection
  - Cascade classifiers and k-nearest neighbour classifier

Occupant State Monitoring
Initial Plan

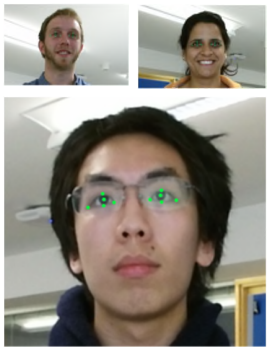
- Occupant Monitoring System
  - **Sensor** – Cameras, vehicle sensors, GPS, heart rate monitors, …
  - **Sensor Output Post Processing** – Background removal, noise removal, …
  - **Feature Extraction** – Mouth, Eyes, Heart Rate, Arms, …
  - **Machine Learning Classifier**
  - **Occupant State Result**
  - **Vehicle Action**



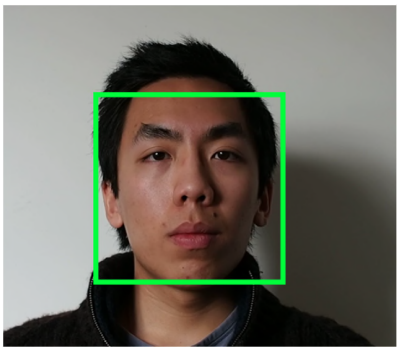
- Initial Plan
  - **Sensor choice –** Camera vision
  - **Feature choice –** Eyes
    - Closely linked with state, widely used
    - Robust and real-time applicable
  - Build a key component of an **Occupant Monitoring System** - Eye Detector

Occupant State Monitoring

Overview - Pipeline

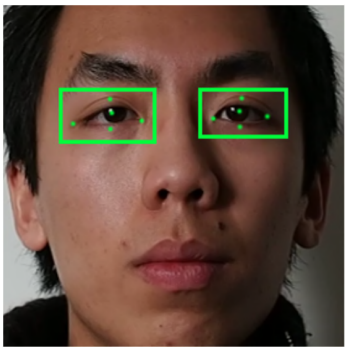


Train model with labelled images (internal & external datasets)



Input: Image
Output: Face region

Input: Face region
Model prediction.
Output: Eye Region



Final Result

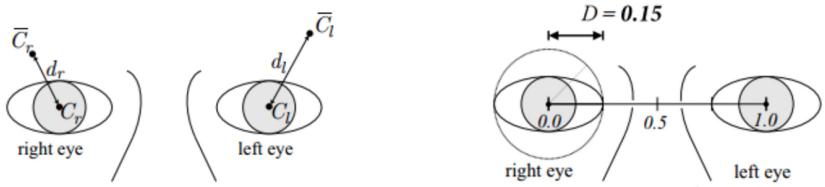
Occupant State Monitoring

Results

Testing

BioID dataset has 1521 images of 23 subjects

Method	D < 0.05	D < 0.10	D < 0.15
Jesorsky et al. (2001)	40%	79%	93%
Valenti et al. (2008)	**84%**	91%	**99%**
Timm et al. (2011)	82.5%	**93%**	95%
Ren et al. (2014)	77%	92%	96%
Proposed Method	**79%**	**96%**	**99%**

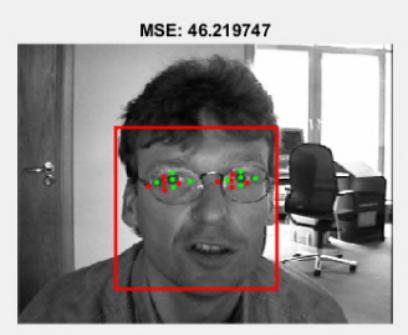


$$D = \frac{\max(d_l, d_r)}{\|C_l - C_r\|}$$

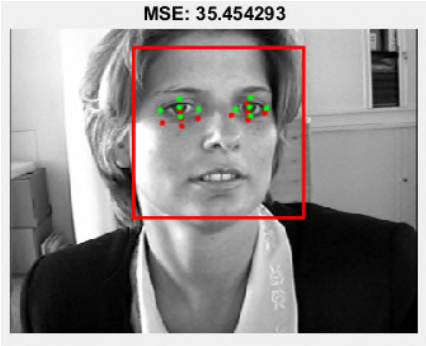
$d_r =$ distance between estimated eye point and true eye point

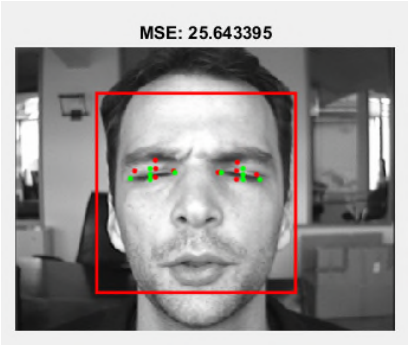
C_r is true point of eye

Occupant State Monitoring
Results

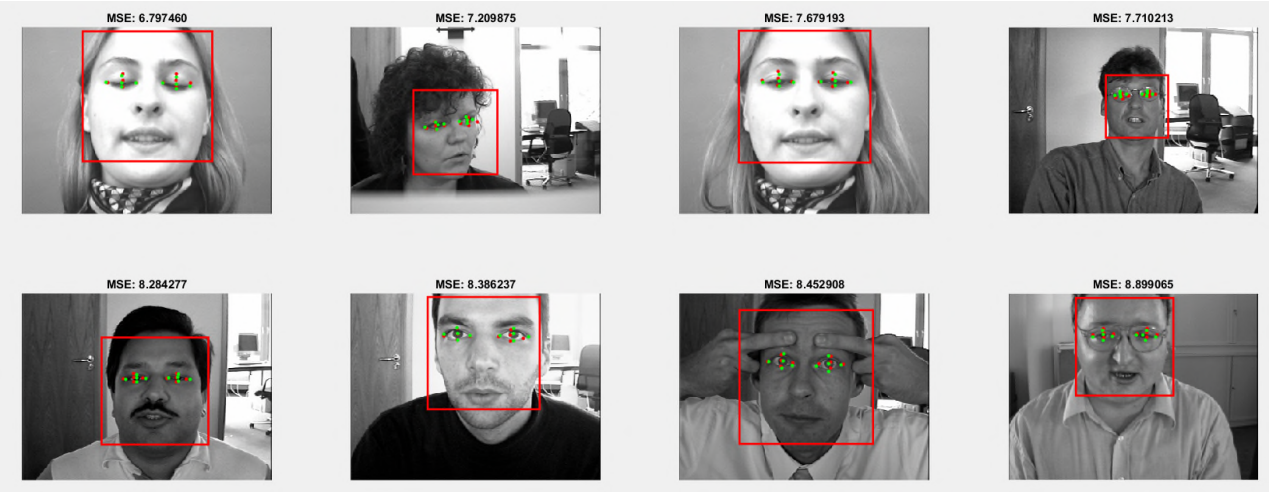


Eye Detector
- Mean Squared Error (MSE)



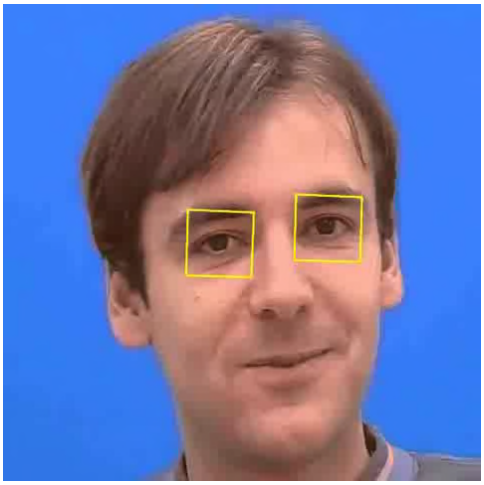


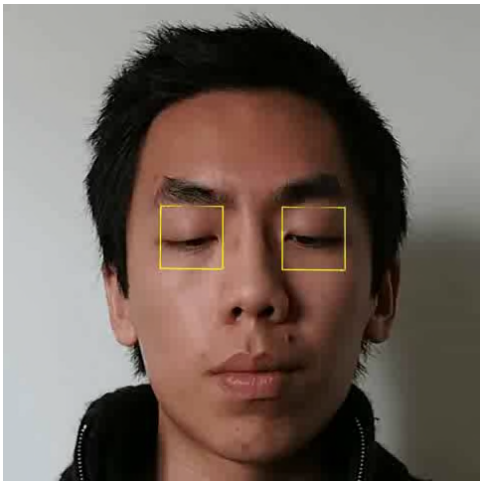
Occupant State Monitoring
Results



Occupant State Monitoring
Showcase

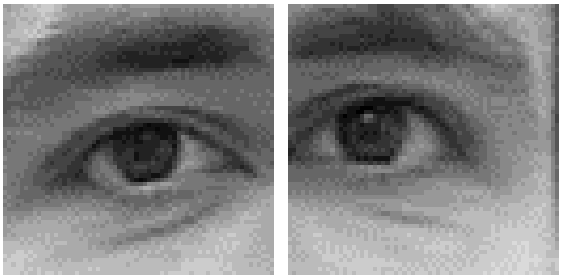
Component Demo – Eye Detector

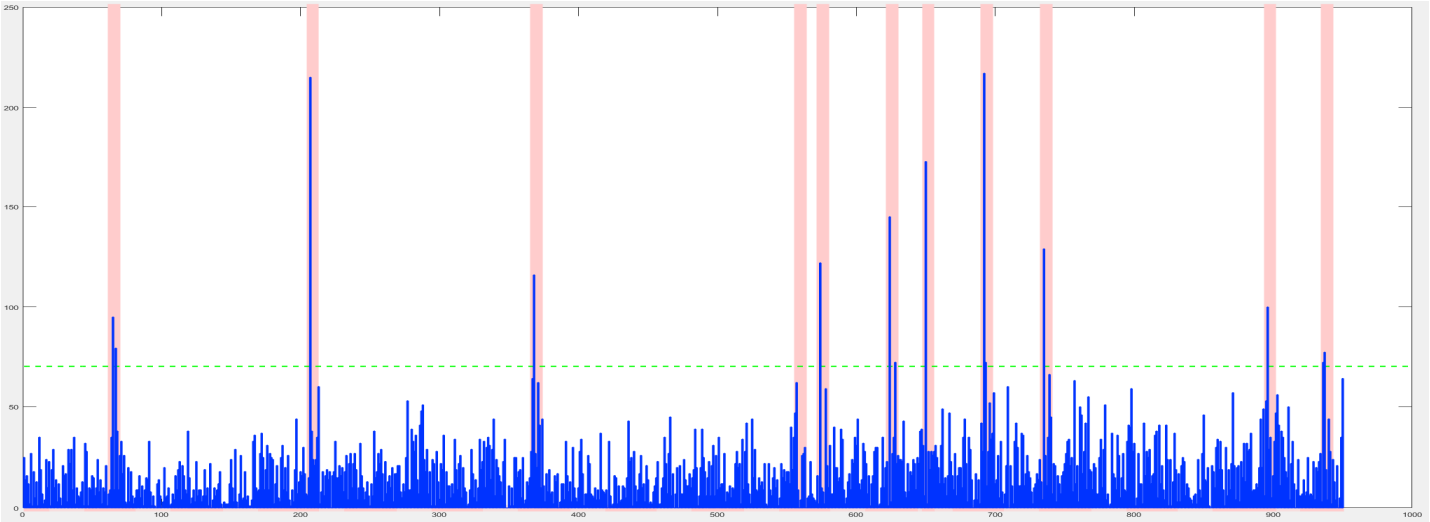




Occupant State Monitoring
Showcase

Component Demo – Blink Detector





The Outlook

Open Questions & Improvements

- Include more meaningful features to improve model
- Investigate relationship between occupant behaviour and optimal vehicle parameters
- Testing under various traffic and environmental scenarios
- Identifying typical nuances and behaviours between specific occupant groups (male vs female, young vs old, aggressive vs conservative people)

Full name: Monisa Abdul Wahab

Year of Study: PhD Year 3

Department: Computer Science

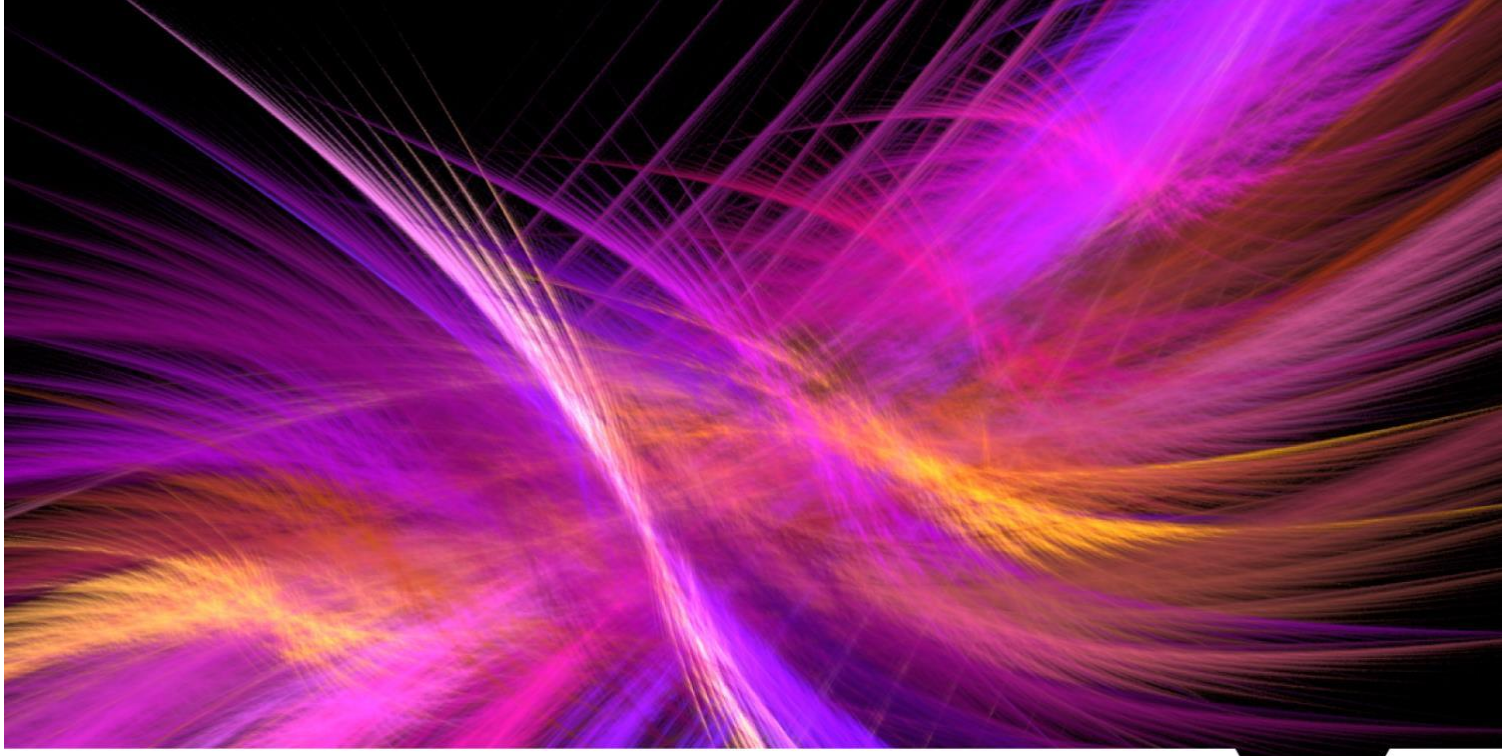
Submission Type: Presentation

Title: A Gamification Approach to the Design and Development of MARA ePortfolio (MeP) in Kolej Profesional MARA (KPM), Malaysia

Abstract: Games are part of day to day life that could model behaviours. Gamification, on the other hand, is the process of using game thinking and game mechanics to solve problems and to engage users which are different from serious games application. Quite a number of world\x92s larest company had already deployed at least one gamified application because they understand the potential of game mechanics as powerful motivators of behavior. Learner motivation is very important to engage users to use e-learning application. The e-portfolio is an emerging technology solution for assessing student achievement and showcasing learning evidence that gives significant benefits to the students and educators. Students do not value the portfolio as a lifelong learning strategy while the pure intention of it is to initiate lifelong learning strategy. A major challenge with electronic portfolio today is to maintain the learner\x92s intrinsic motivation to willingly engage in the portfolio process. However, very limited research has been done to explore the possibility of embedding game mechanics in e-learning applications specifically the e-portfolio applications to raise engagement. This research uses a web blog and game-like techniques like points, badges, and leaderboard that would engage learners and trigger their motivation to further use the e-portfolio application. This presentation will specifically focus on the design and development of the gamified e-portfolio application while the results are still underway and will not be presented. The MARA ePortfolio (MeP) is an effort to include the gamification elements inside an e-learning application to make it interesting and enjoyable experience for the learners to uncover the potential of gamification in education.

Keywords: e-learning, e-portfolio, gamification, gamified, game mechanics, user engagement

Abstract Word Count (Est): 262



WPCCS 2016

By: Monisa Abdul Wahab

A Gamification Approach to MARA ePortfolio (MeP)

Design, Development and Implementation

Table of content

Table of content

Introduction

- What is an e-portfolio?
  - It is an electronic collection of evidence that shows your learning journey over time (Barrett 2010).
  - Wide applied to many higher education institutions in many countries.
  - Provide students a user-centered learning facility to manage information
  - New in Malaysia, specifically in Majlis Amanah Rakyat (MARA) higher education institutions

Introduction

- Despite the significant benefits of e-portfolio systems in education, one of the most disturbing issue is user engagement.

Introduction

- What is gamification?
  - is the use of game play elements for non-game applications (Muntean 2011).
  - By applying game mechanics and dynamics to tasks and e-learning processes we can increase user engagement with an e-learning application and its specific tasks (Muntean 2011).

Introduction

  - Successfully used in many web based businesses to increase user engagement but a very limited research has been done to explore the idea of integrating games elements in e-portfolio systems as a tool to increase student motivation and engagement.

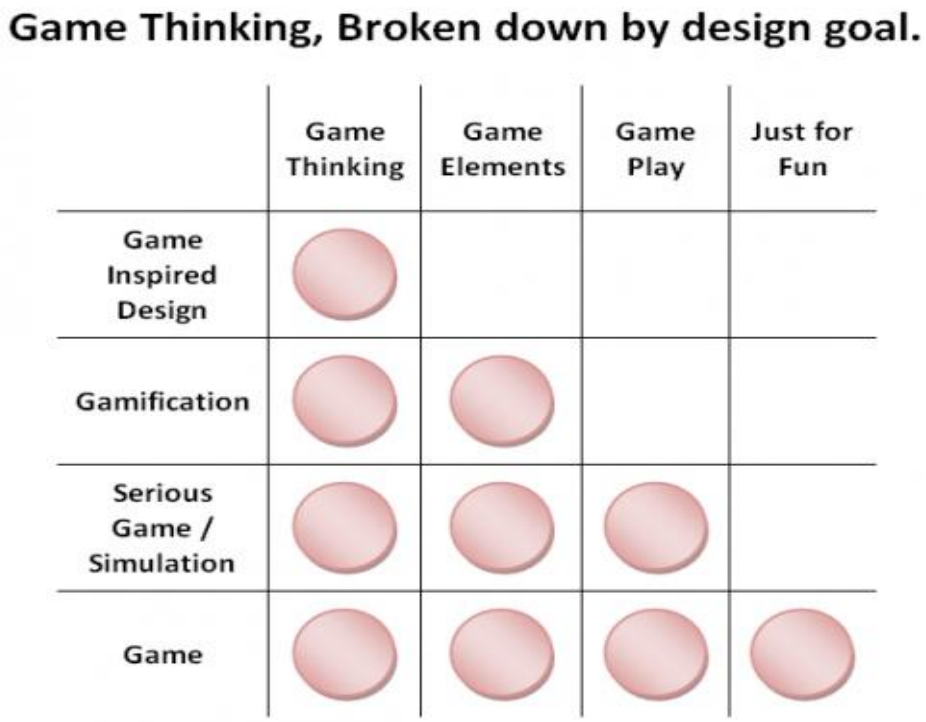
Introduction

- Game mechanics and game dynamics (adapted from Bunchball (2010)).

Table 1: Game elements

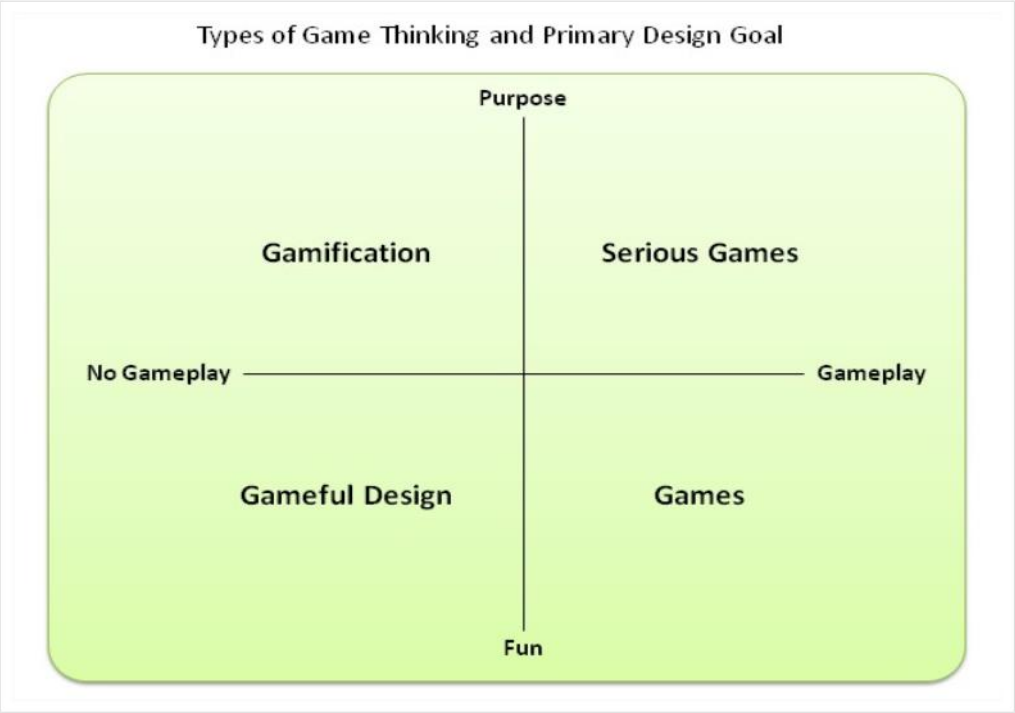
Game mechanics	Game dynamics
Points	Reward
Levels	Status
Trophies, badges, achievements	Achievement
Virtual goods	Self expression
Leaderboards	Competition
Virtual gifts	Altruism

Game thinking



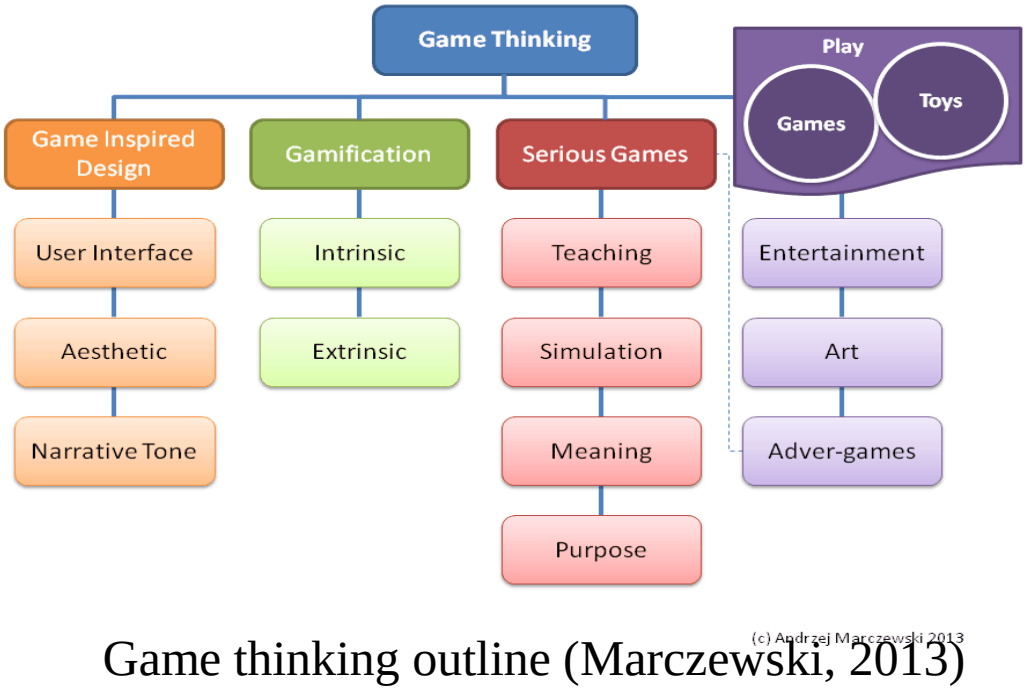
Game thinking broken down by design goal (Marczewski, 2013)

Game thinking



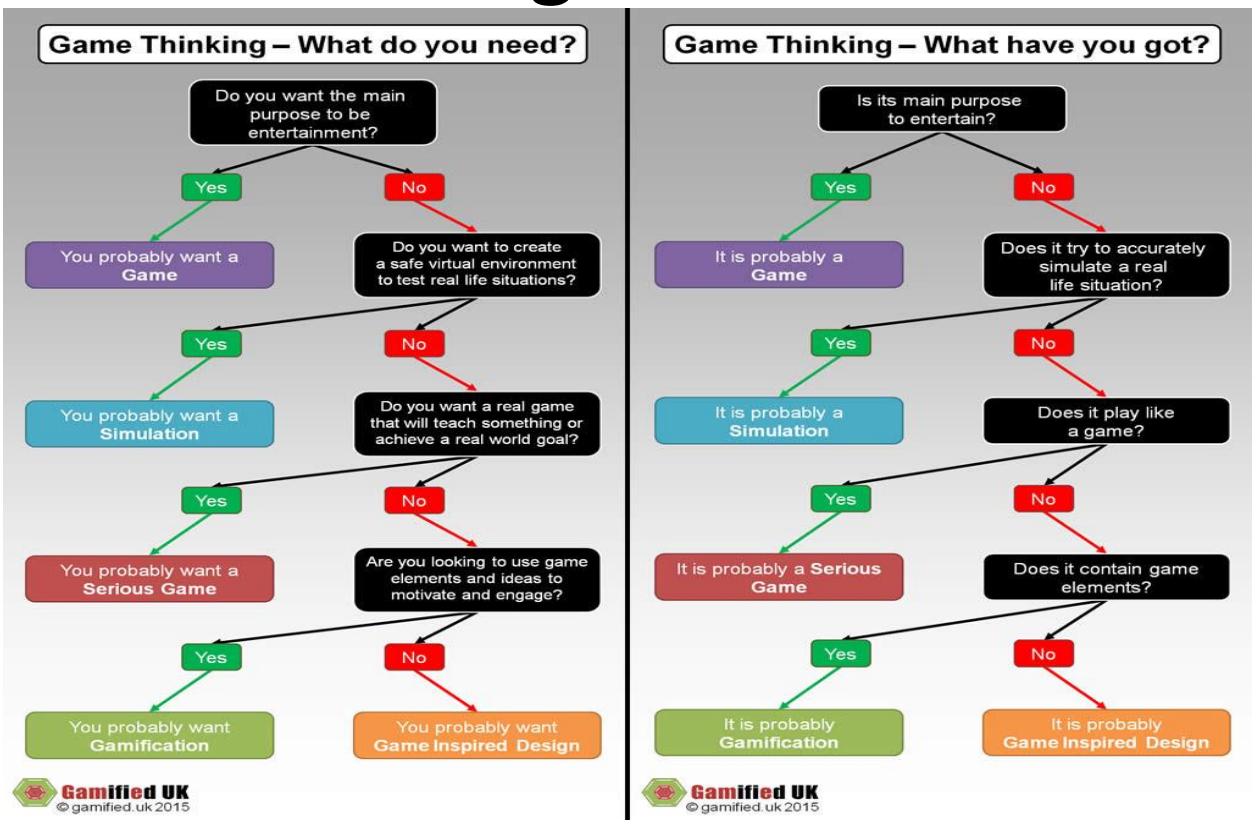
Type of Game Thinking and Primary Design Goal (Marczewski, 2013)

Game thinking



Game thinking outline (Marczewski, 2013)

Game thinking decision trees



Game thinking decision trees (Marczewski, 2013)

Introduction

- Who is Kolej Profesional MARA (KPM)?
  - A MARA higher education institutions that offers a range of courses from preparatory to higher national diploma level.
  - Have 6 campuses throughout Malaysia.
  - Currently populated with students ages between 17-26 years (generation Z or digital natives) who attained average results in Malaysian Certificate of Education (SPM).
  - Interested in implementing e-portfolio but worries about user engagement issues.

Introduction

- Why KPM, Malaysia?
  - higher education institution, smaller context that mimic the broader context, accessible.

Problem statement

- There are significant benefits of using the e-portfolio system in education but there are also known user engagement issues.
- A major challenge with e-portfolio today is to maintain learner intrinsic motivation to willingly engage in the portfolio process (Barrett, 2005).

Problem statement

- How to solve user engagement issues in e-portfolio implementation?
  - By applying game mechanics and dynamics to tasks and e-learning processes we can increase user engagement with an e-learning application and its specific tasks (Muntean 2011).

Summary of Literature Review

- Meeus et al. (2006) stated that portfolio in education is:
  - Student centered;
  - Competence oriented;
  - Cyclical with regard to action and reflection;
  - Multimedia oriented
- The use of this game-like techniques can increase the feel of ownership and purpose when engaging tasks (Muntean, 2011).

Summary of Literature Review

- The main goal of gamification is to rise the engagement of users by using game-like techniques such as scoreboards and personalized fast feedback (Flatla, Gutwin & Nacke, 2011).
- Schools have been using game-like elements in classroom activities like points and badges as a reward for desired behaviors to motivate student (Lee and Hammer, 2011).

Summary of Literature Review

- Barrett (2005) suggested the use of multimedia tools and the use of web log or "blogs" and "wikis" to engage learner.
- The greater learner control over e-portfolio content, purpose, and process will lead to more intrinsic motivation (see figure 1)

Summary of Literature Review

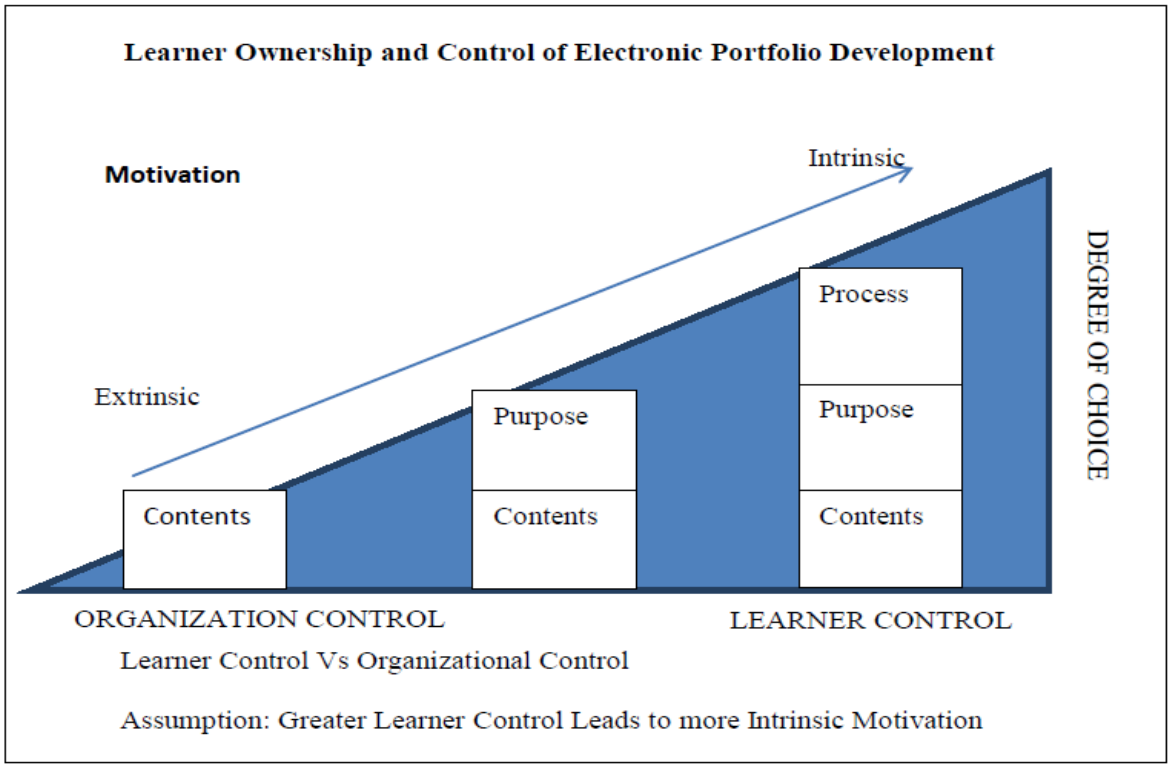


Figure 1 (Barrett 2005)

Main research question

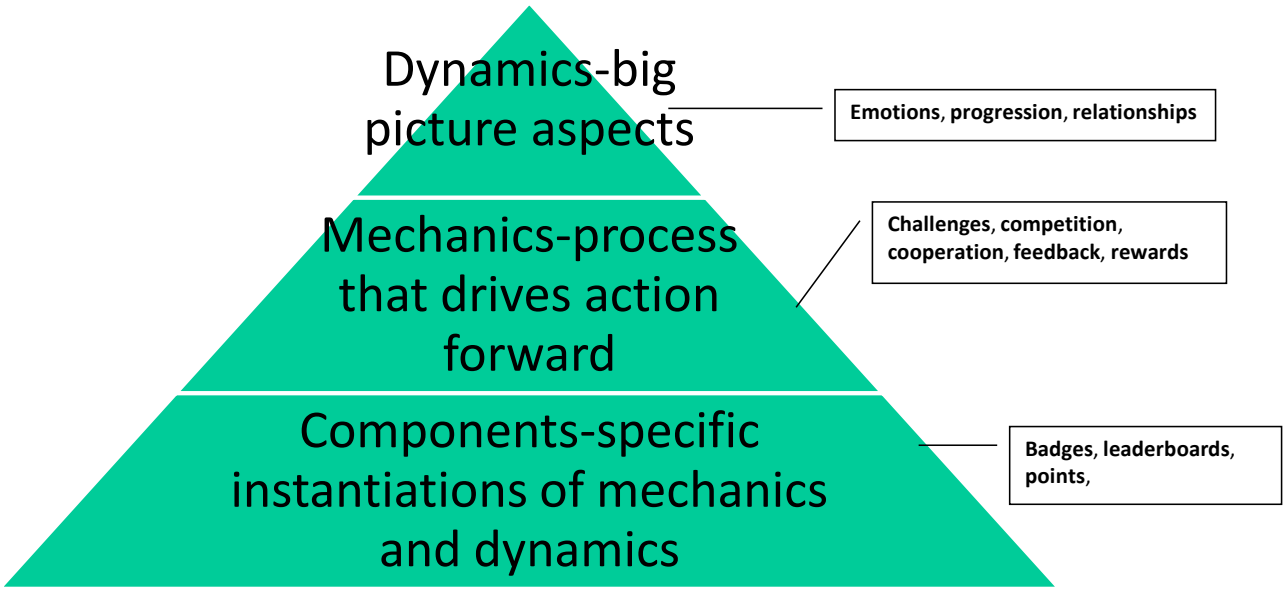
- Can we improve user motivation and user engagement in an e-portfolio system by applying game elements?

Principles of gamification

- Principles of gamification (Flatla et al., 2011)
  - Goal orientation
  - Achievement
  - Reinforcement
  - Competition
  - Fun orientation

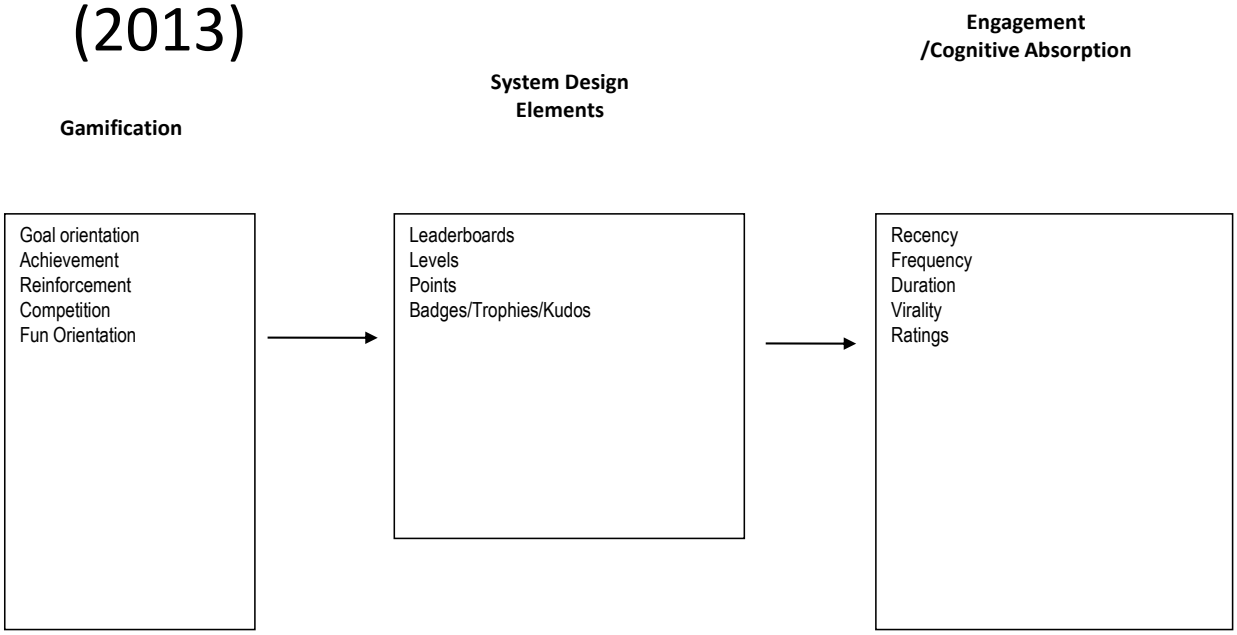
Pyramid of gamification elements

- Pyramid of gamification elements (Werbach and Hunter, 2012)



Adapted gamification framework

- Gamification framework by Nah et al. (2013)



MeP Design

- By using points, badges and leaderboard as extrinsic motivators we hope to trigger required intrinsic values like belonging, autonomy, power, mastery, meaning, learning, self knowledge, and fun.
- Points : keep score, determine completed tasks, connect to rewards, provide feedback, display progress, data for the MeP admin, can be used to represent anything.

MeP Design

- Badges : representation of achievements, flexibility, signaling of importance, collections, social display.
- Leaderboard/ranks : ranking (feedback on competition), pride
- MeP main features:
  - Register new user
  - Give access to authorise user and denied access to aunauthorised user

MeP Design

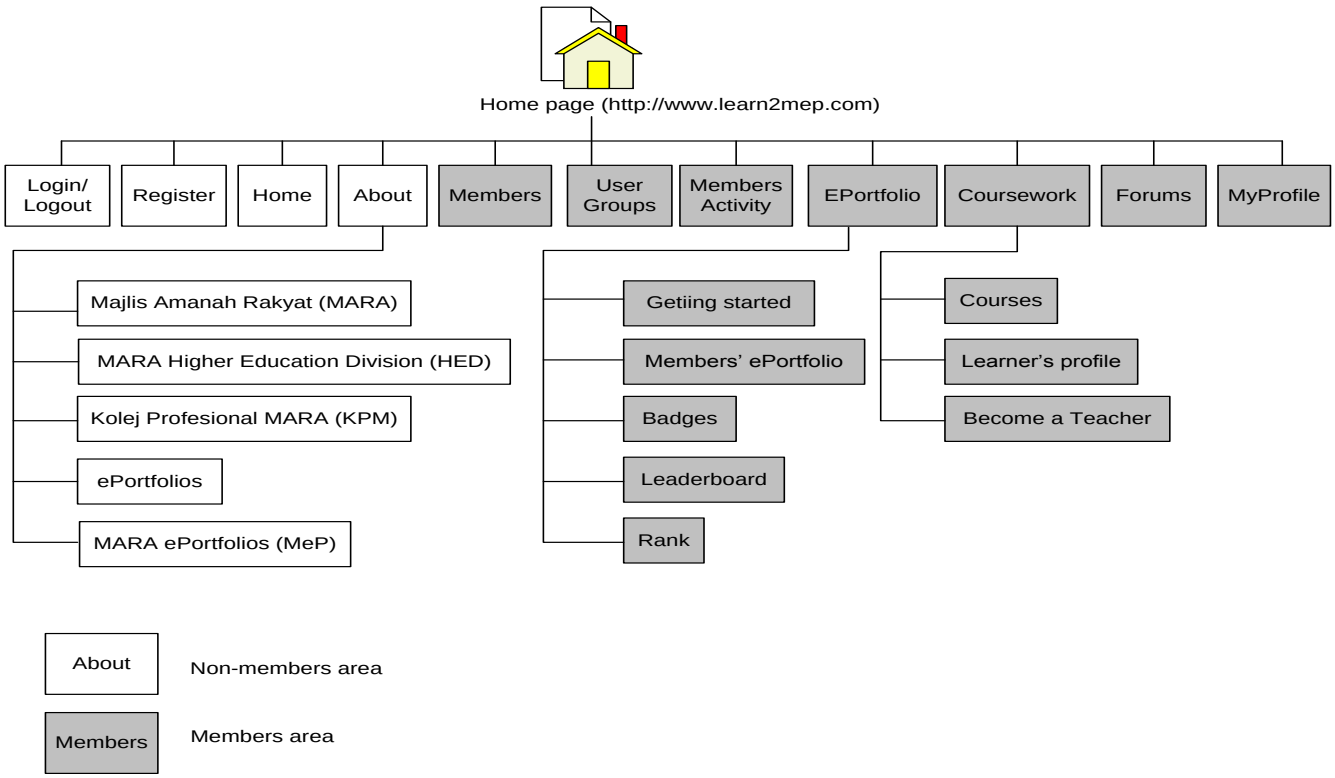
- Create, update and manage their profile (public portfolio).
- Create, update and manage their personal blog (personal portfolio)
- Allow registered users to give comments and feedback to member's uploaded media and content
- Allow registered users to mark any member's updates/content as favourite
- Award points and badges based on user activities

MeP Design

- Show leaderboard of active users
- Rank users based on accumulated points

➢ The design process of MeP started from March 2015 until May 2015.

MeP Site Map



MeP Development

- The development process of MeP started from June 2015 until January 2016
- This application has been developed using Wordpress, PHP, MySQL, HTML, CSS, JS

MeP Implementation

- The implementation phase took place from February to March 2016 and the evaluation of MeP has been done afterwards.
- Participants has been asked to complete a post survey and participate in the interview sessions on a voluntary basis.

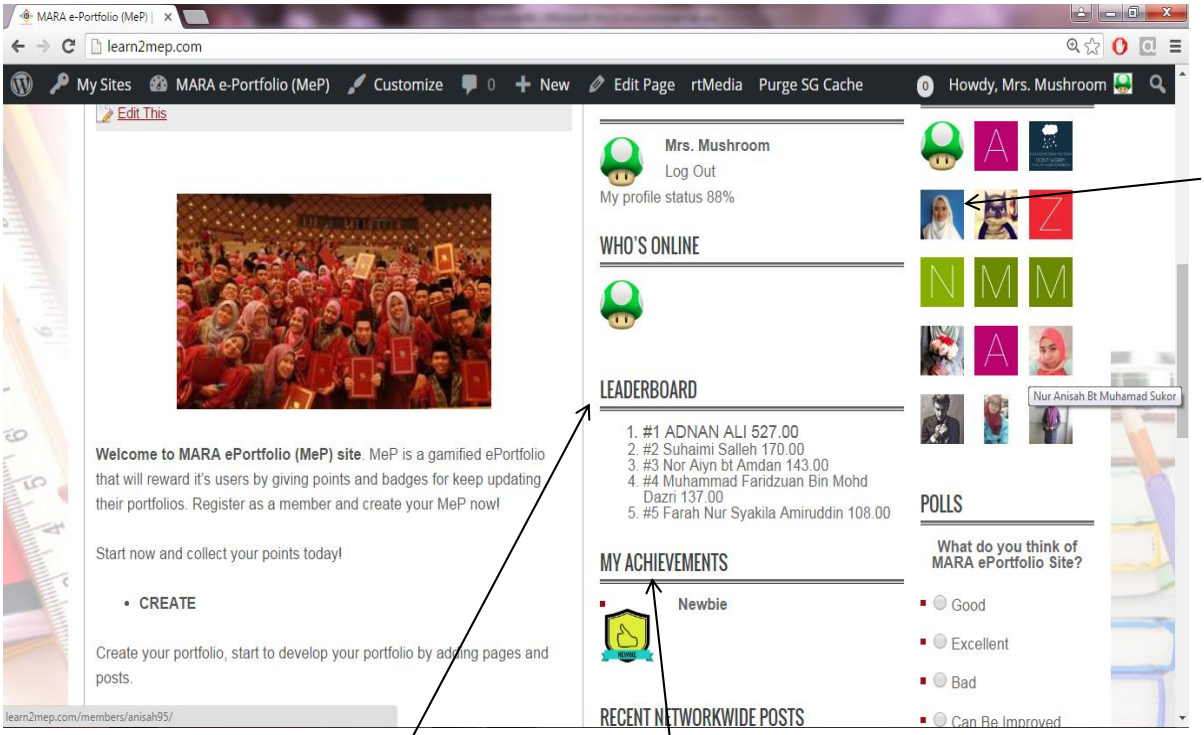
MeP Evaluation

- The evaluation of MeP has been done on April 2016 in the form of post survey (online questionnaire), interview and e-portfolio samples.
- Currently, we are working on data analysis of the collected data and the results has not yet ready.

Conclusion

- In this research, we developed a gamified e-portfolio for KPM students to use based on the literature, synthesize the findings from the review, and demonstrate it through MARA ePortfolio (MeP) application.
- The design and implementation of the MeP was successful and data collection activities has been done. The data analysis of the post survey and interview is still underway thus, the results of the implementation is currently unavailable.

MeP User Interface



Profile picture

Leaderboard

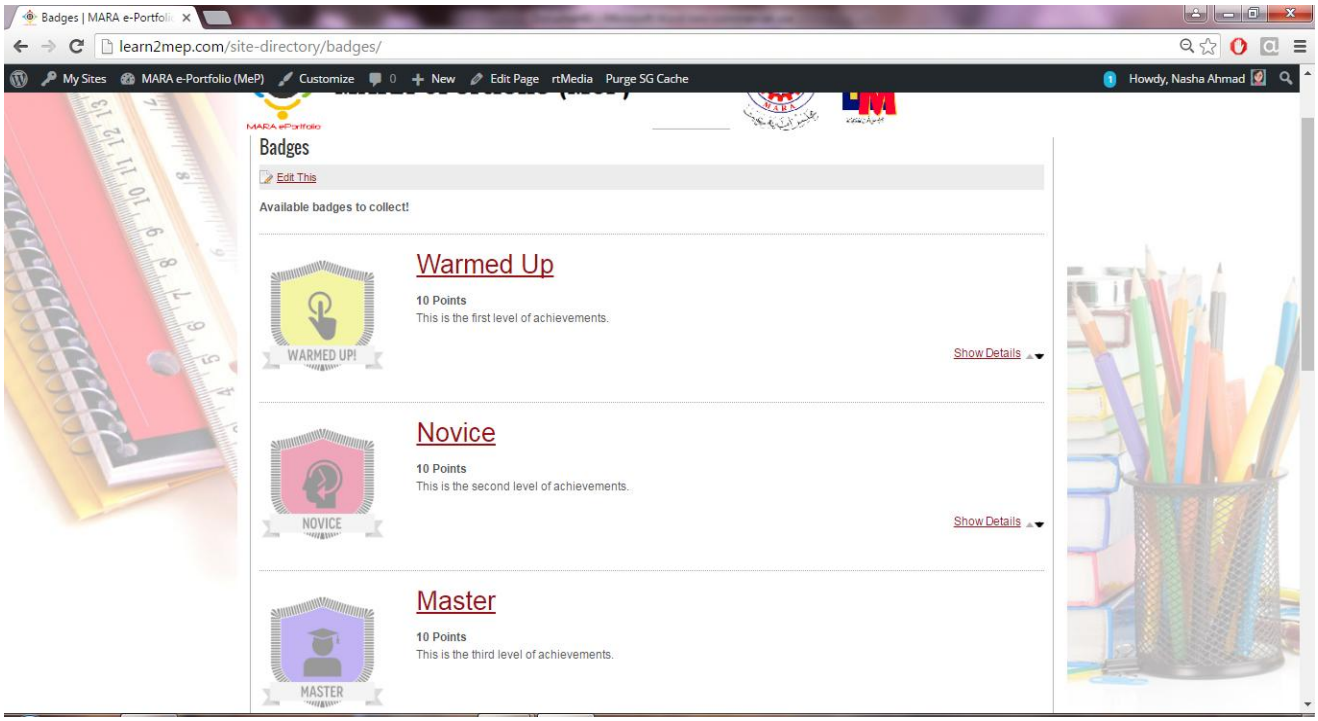
Achievements

MeP User Interface

Notifications

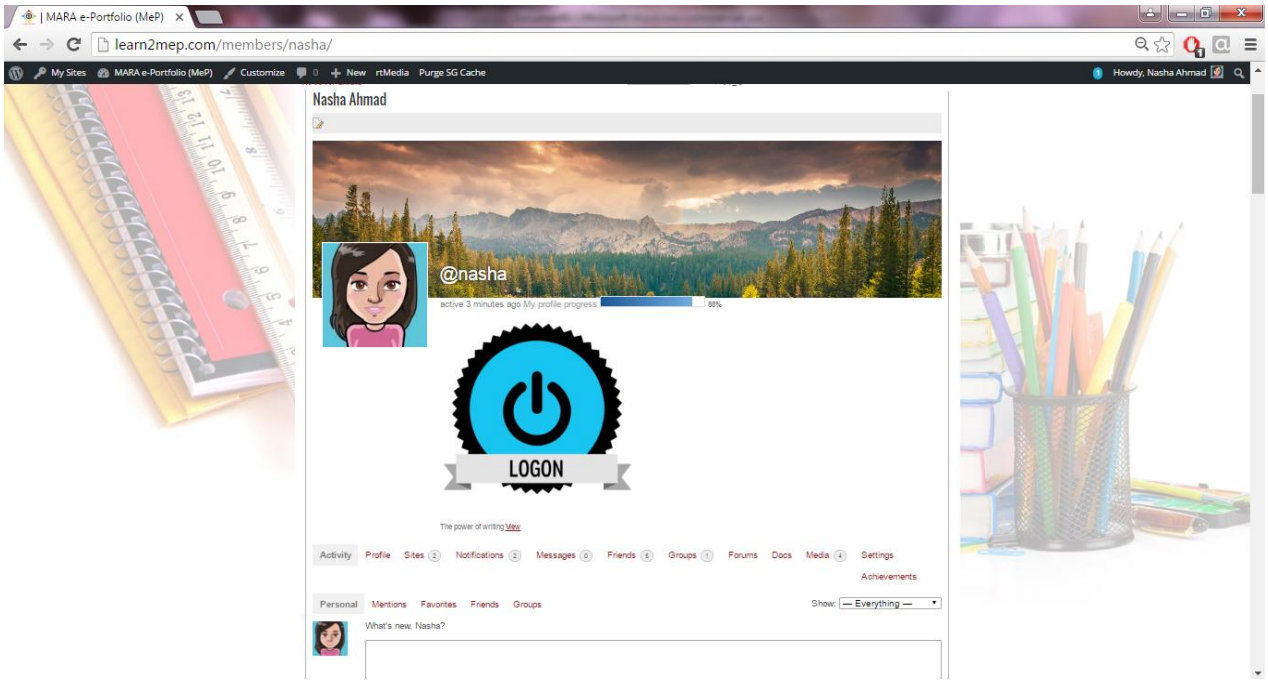
MeP User Interface

- Available badges to collect



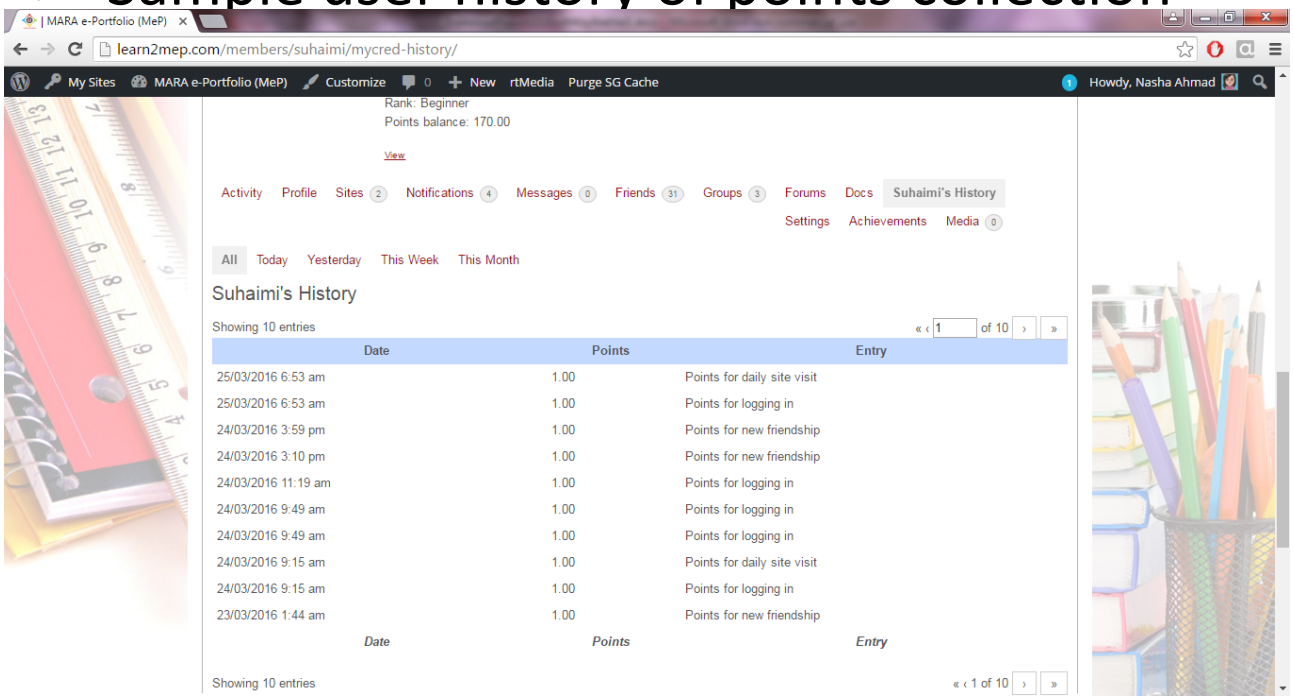
MeP User Interface

- User profile page



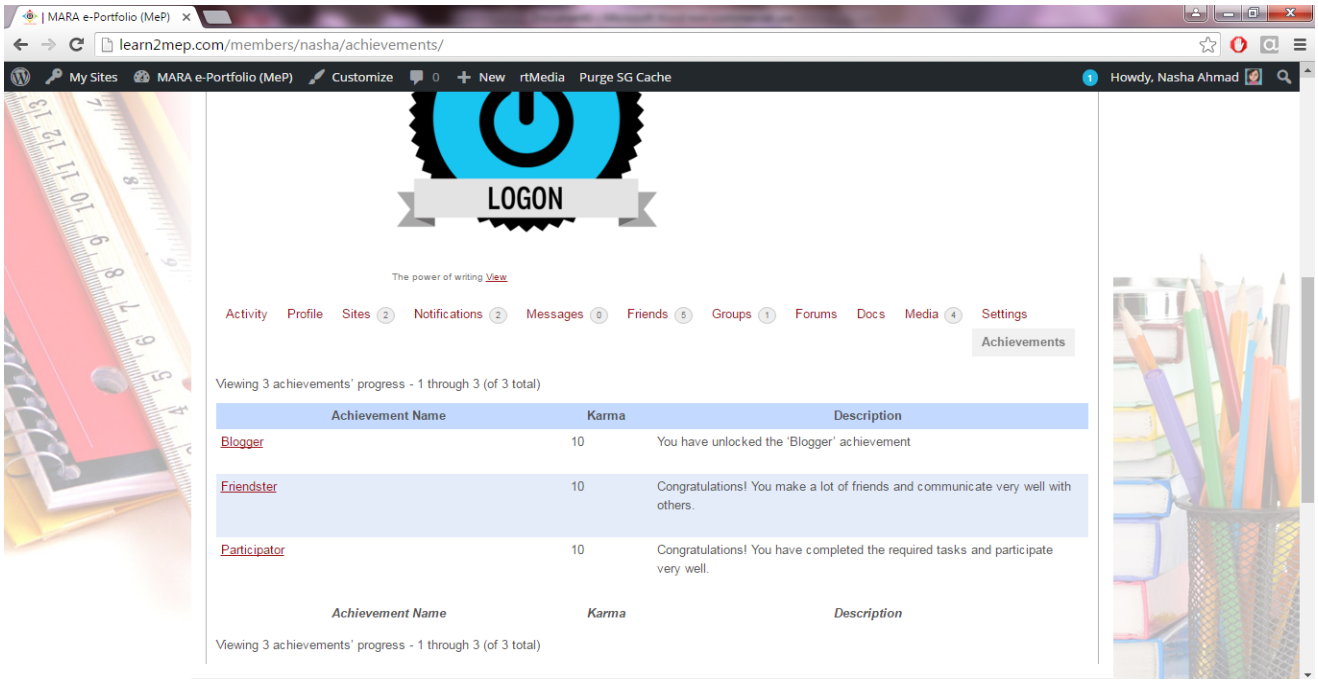
MeP User Interface

- Sample user history of points collection



MeP User Interface

- User achievement



Full name: Greg Watson

Year of Study: PhD Year 1

Department: Computer Science

Submission Type: Presentation

Title: People Re-Identification with Articulated Appearance Matching

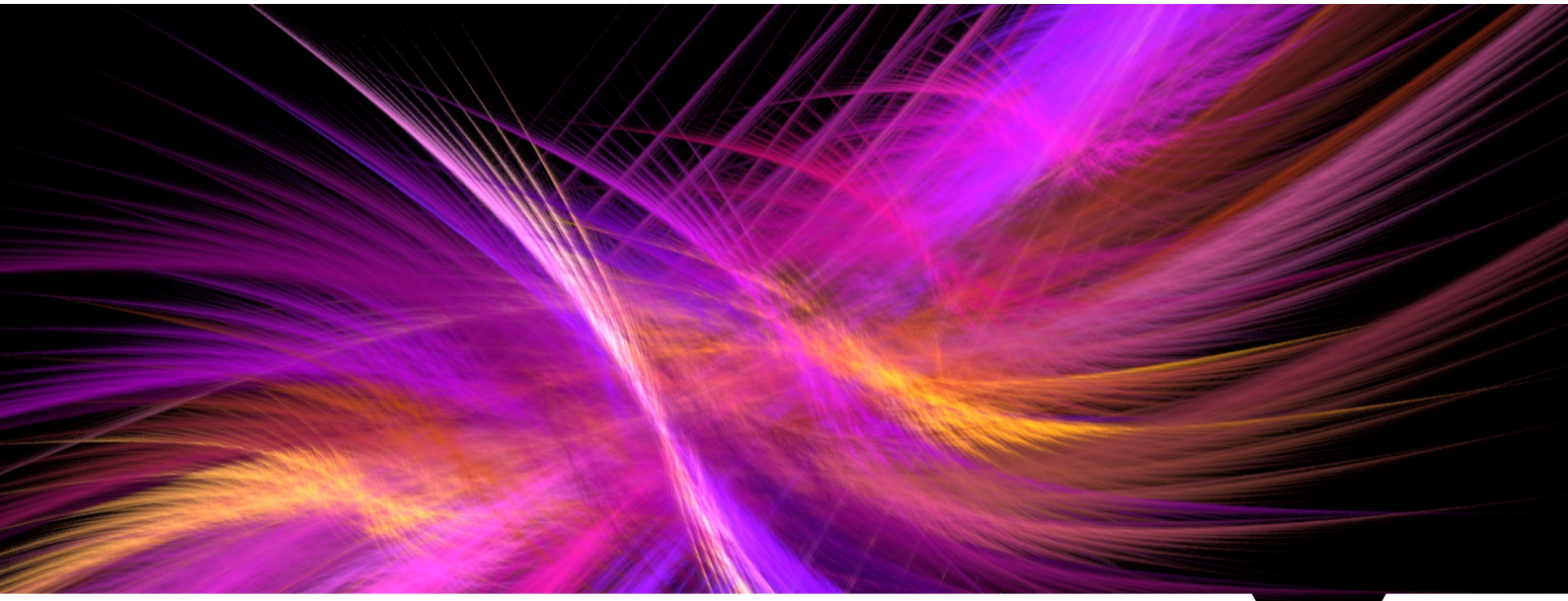
Abstract: People Re-identification is about identifying whether or not a person captured in one camera can be accurately identified with a second camera, at a different location and time.

Generally, the first step in this task is to eliminate the background from the image, in order to only extract features from the foreground. Many different methods have been proposed to do this, and I will discuss the methods which I have investigated; These methods mainly revolve around using machine learning in order to learn the locations of the foreground through interpretation of the person\x92 position in the image.

By using these techniques, it is then possible to subdivide a person into individual limbs and match them separately. Such methods can improve the matching rate further. State-of-the-art methods are currently achieving a high Rank-1 rate, which is defined as the amount of images matched on the first attempt. I will relate my work to the state-of-the-art methods and compare their performance.

Keywords: Person Re-identification, Image Analysis, Feature Matching, Data

Abstract Word Count (Est): 159

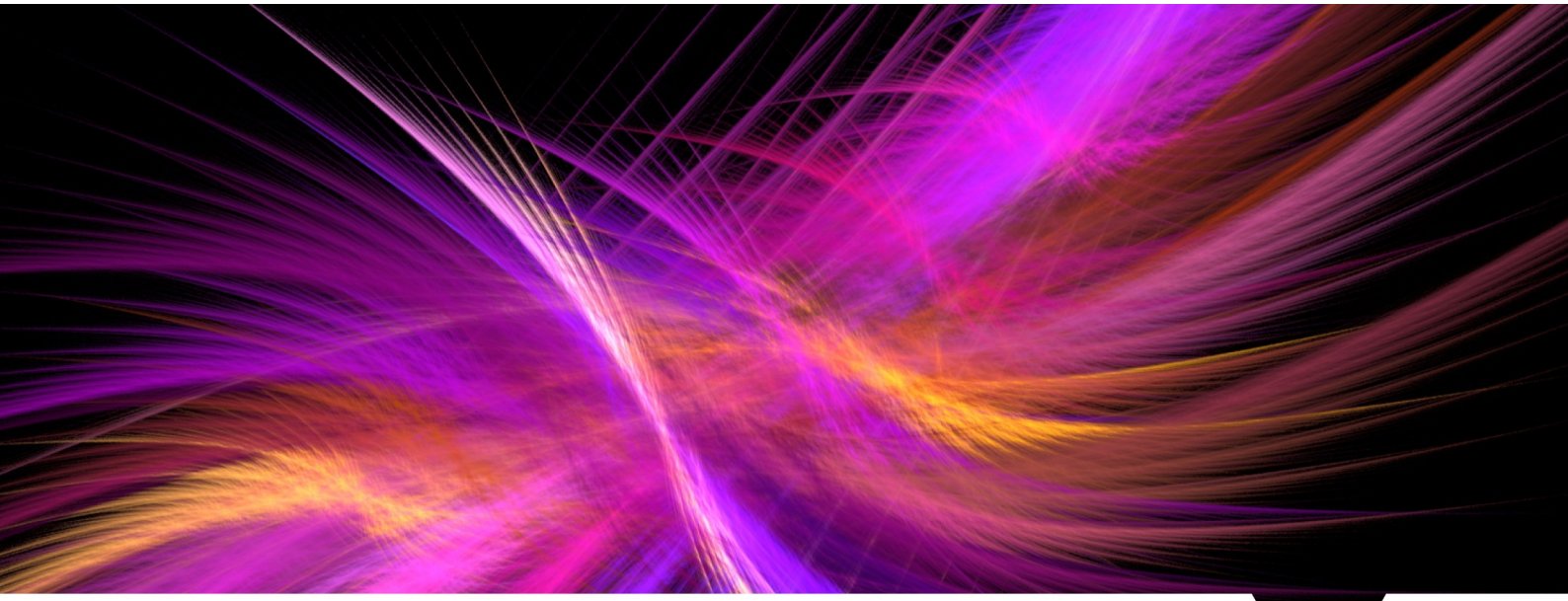


People Re-Identification with Articulated Appearance Matching

Greg Watson

First Year PhD Student, Department of Computer Science

Supervisor: Dr. Abhir Bhalerao



Part 1 – Definition of ReID

What is 'People Re-Identification'?

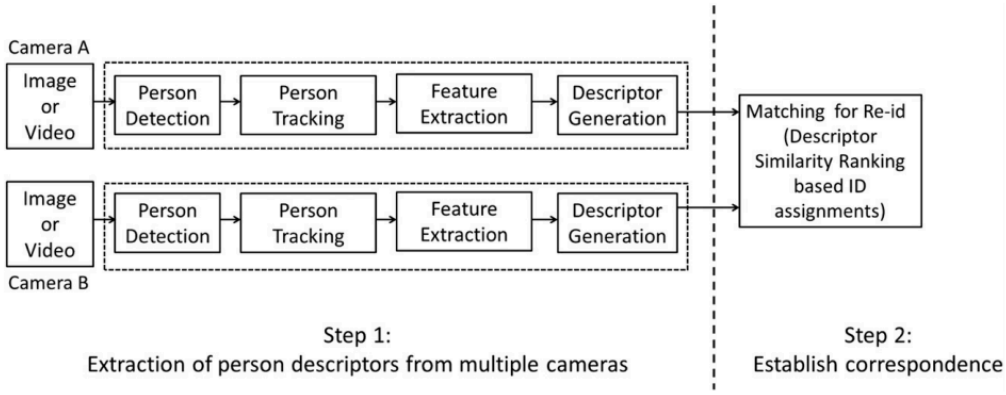


Fig. 3. Re-ID system diagram.

- Matching an individual or group of people between **different**, **non-overlapping cameras** at **different locations** and at **different times**.

Bedagkar-Gala, Apurva, and Shishir K. Shah. "A survey of approaches and trends in person re-identification." *Image and Vision Computing* 32.4 (2014): 270-286.

Challenges

- Different images of the same person may have different:
  - Views
  - Lighting
  - Resolution
- The same person will look different at different times.

Challenges - Examples

The same, and look similar.

The same, and look different.

Different, and looks similar.

Method

For each test image x in the probe set

Compare to all test images y in the gallery set

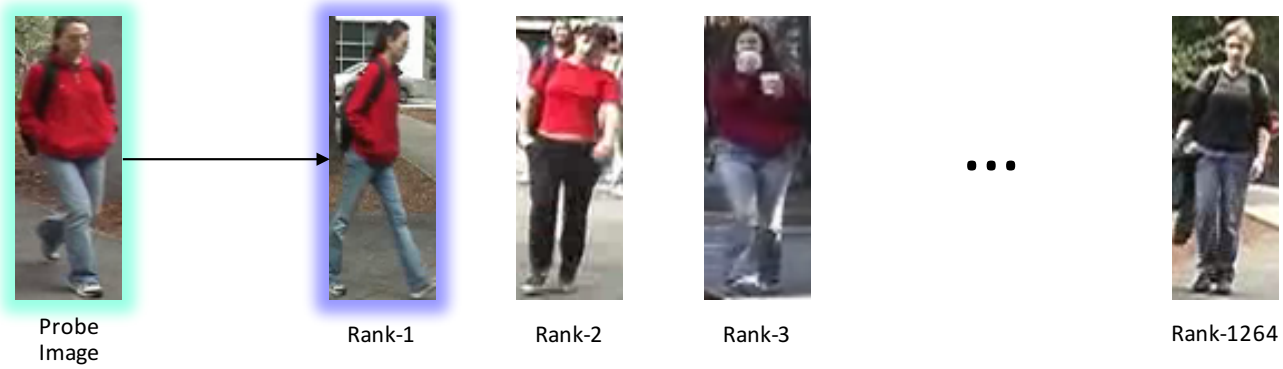
Compute the distance between the two, $D(I_x, I_y)$

End

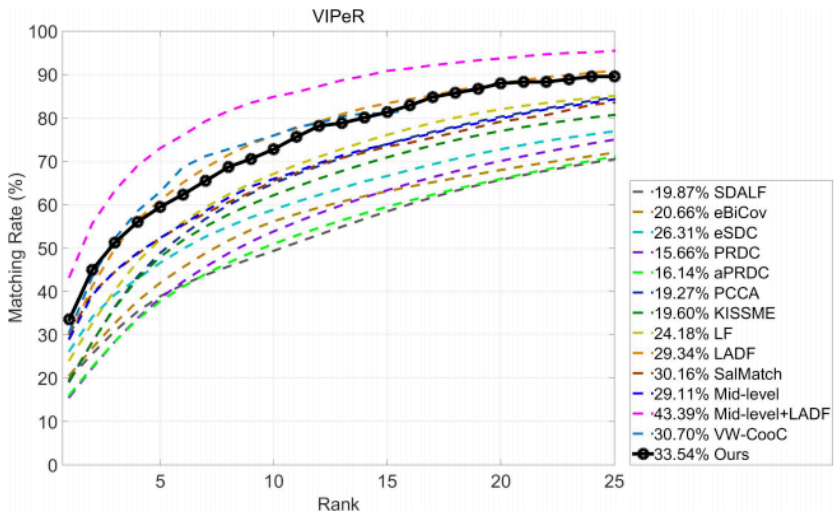
Find x and y such that $D(I_x, I_y)$ is the smallest.

Performance Measure: Rank-n

- The percentage of images correctly matched in ’n’ ‘guesses’.



State-of-the-Art Performance



Model-Based and Model-Free

- Model-Based:
  - Strong prior on image contents.
  - Typically a person model, e.g. head, torso, legs.
  - Skeleton models.
- Model-Free:
  - Hidden or weak semantic model by learning.
  - Deep Learning.

Model-Based: Skeleton Fitting

- Train up a shape fitting approach:
  - Shapes are locations of limbs, i.e. skeleton.
  - Mark shapes on a set of images.
  - Learn a regressor between appearance and shape.

Why use a skeleton?

- Separate foreground (human) from background (treated as noise).
- Localize limbs allowing limb-to-limb feature matching.
- Would not catch potentially discriminating information **outside of the foreground**.

Spatial Localisation

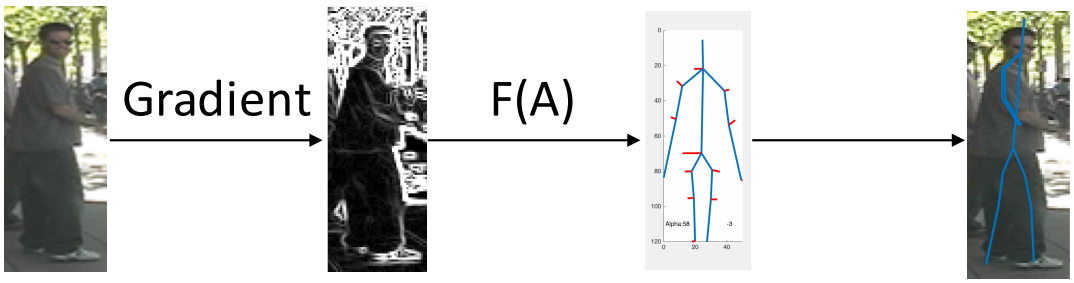
- Matching limb-by-limb would lead to better results than matching body-by-body.
- PROBLEM: Accurate limb-fitting method required.

Appearance to Shape Regression

- s = f(A)
  - ‘s’ is the target shape parameters (set of points of skeleton). s = $\{x_1, y_1, x_2, y_2, \ldots, x_n, y_n\}$.
  - A is the image appearance (or feature map). ‘A’ could be a gradient image.
  - f is a mapping function.

Appearance to Shape Regression

- This is a very high dimensional problem:
  - E.g. A is N x M dimensions, s is 15 x 2:
    - ‘s’ expands to 28 x 2 with ‘Limb Widths’.
  - We can reduce dimensionality of both A and s.

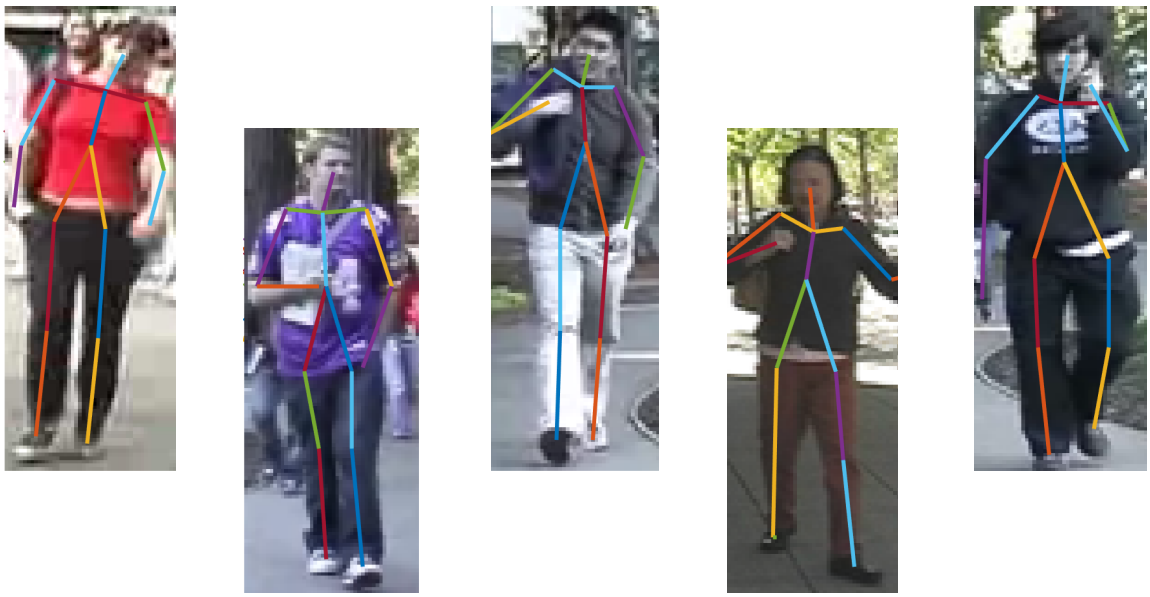


Solving for the Mapping Function

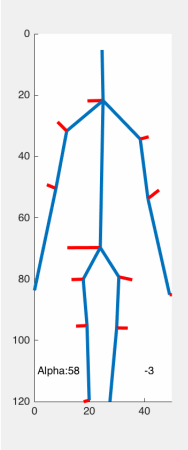
- There are two options:
  1. Solve as a linear problem, s = AI, using Linear Least Square Regression.
  2. Solve it more generally (non-linearly) using a Neural Network.

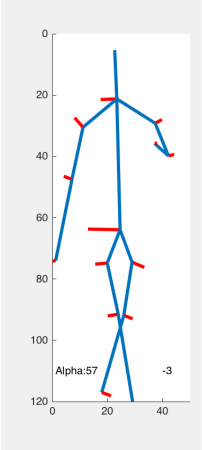
In both cases, training examples are necessary.

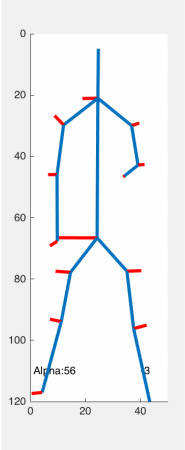
Training Data Examples



Dimensionality Reduction of s



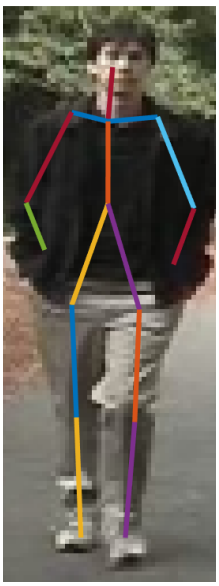




Linear Regression – Good Results

Left: Pre-Labelled Skeleton

Right: Predicted Skeleton



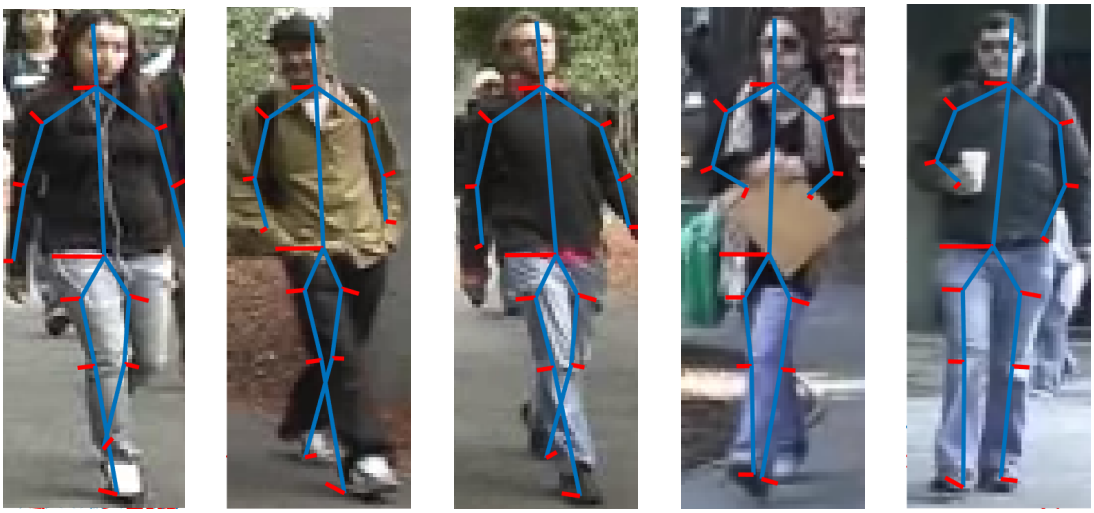
Linear Regression – Bad Results

Left: Pre-Labelled Skeleton

Right: Predicted Skeleton

Results

- Able to show distinct poses a lot better.



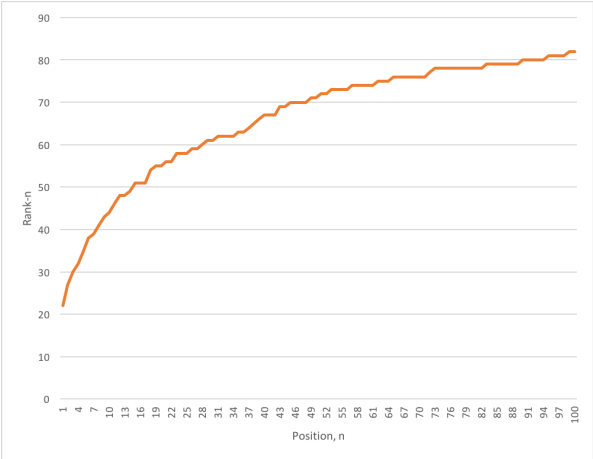
Feature Extraction with the Skeleton

- Aim:
  - $D(F(I_i \mid s_i), F(I_j \mid s_j))$

- Where F is a set of features extracted from image I given an estimate of s.

Current Ranking

- Rank-1: 22%
- Rank-2: 27%
- Rank-3: 31%
- ...
- Rank-10: 44%



WARWICK
THE UNIVERSITY OF WARWICK

Conclusions

- ReID is a generic matching problem.
- There are **Model-Based** and **Model-Free methods**.
- I will look at improving results by **Feature Weighting** and **Distance Metric Learning**.
- I will move to investigate Model-Free methods.

Questions?

Full name: Qiang Zhang

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Wavelet Appearance Pyramids for Landmark Detection and Pathology Classification: Application to Lumbar Spinal Stenosis

Abstract: Appearance representation and feature extraction of anatomy or anatomical features is a key step for segmentation and classification tasks. We focus on an advanced appearance model in which an object is decomposed into pyramidal complementary channels, and each channel is represented by a part-based model. We apply it to landmark detection and pathology classification on the problem of lumbar spinal stenosis. The performance is evaluated on 200 routine clinical data with varied pathologies. Experimental results show an improvement on both tasks in comparison with other appearance models. We achieve a robust land- mark detection performance with average point to boundary distances lower than 2 pixels, and image-level anatomical classification with accuracies around 85%.

Keywords: wavelets appearance pyramids, lumbar spinal stenosis

Abstract Word Count (Est): 113

Wavelet Appearance Pyramid

for Landmark Detection and Pathology Classification
application to Lumbar Spinal Stenosis

Qiang Zhang[1], Abhir Bhalerao[1], Caron Parsons[23], Emma Helm[2], Charles Hutchinson[23]

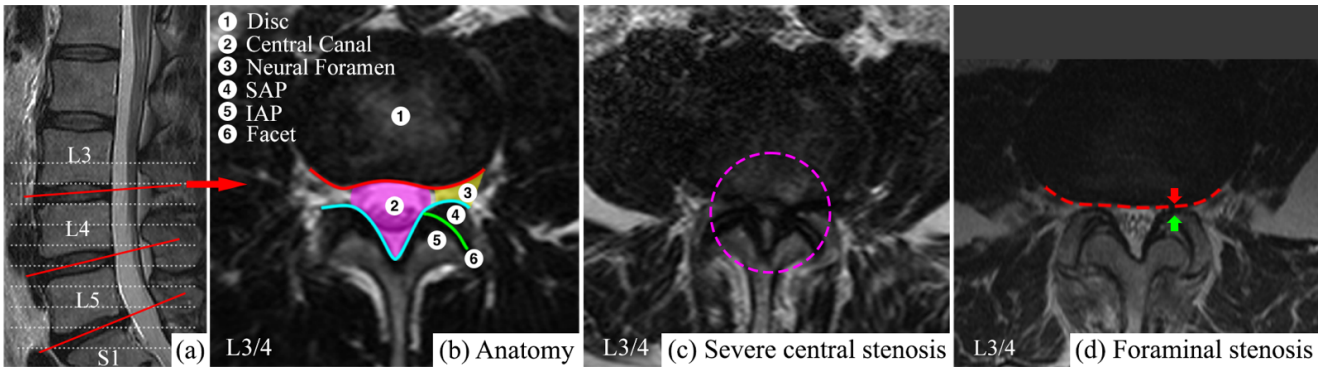
WARWICK

1. Department of Computer Science, University of Warwick, UK
2. Division of Heath Sciences, University of Warwick, UK
3. Department of Radiology, University Hospital Coventry of Warwickshire, UK

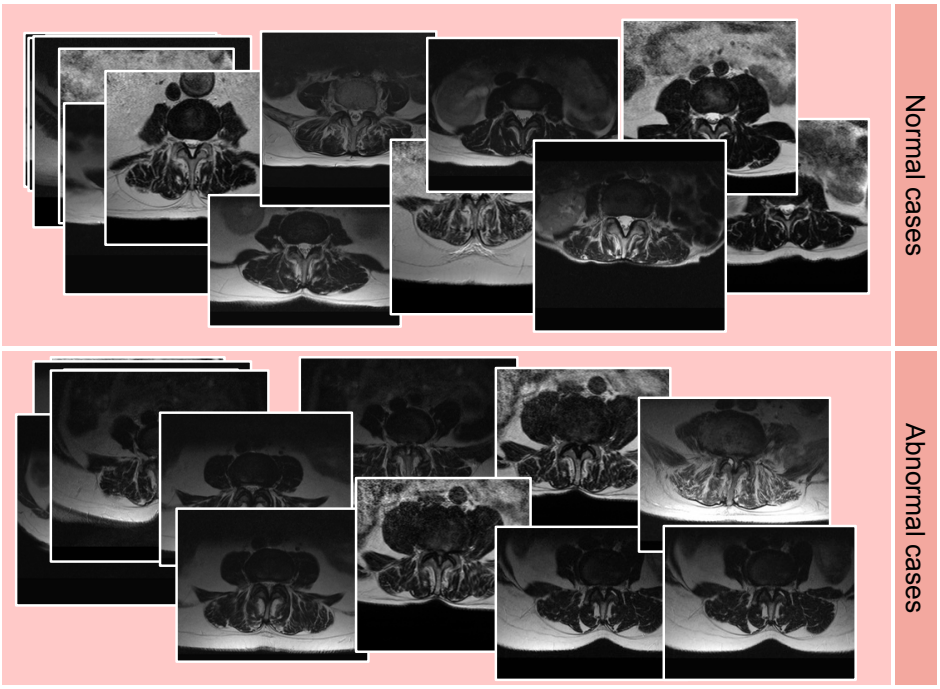
[to appear on MICCAI 2016]

Clinical background

- Lumbar spinal stenosis

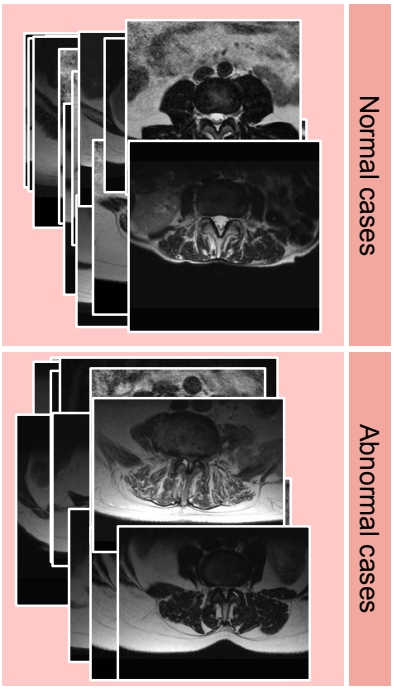


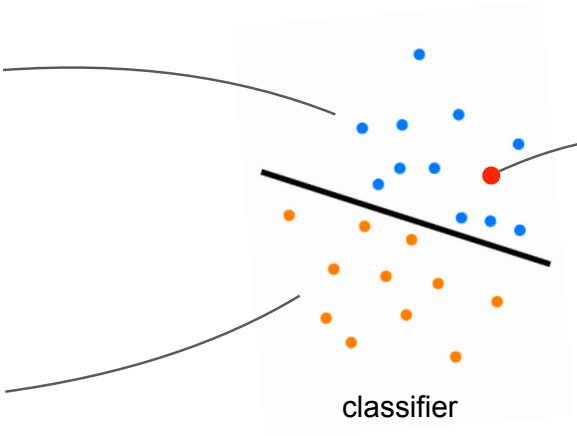
Problem to solve: Diagnosis based on radiological images



Normal or Abnormal ?

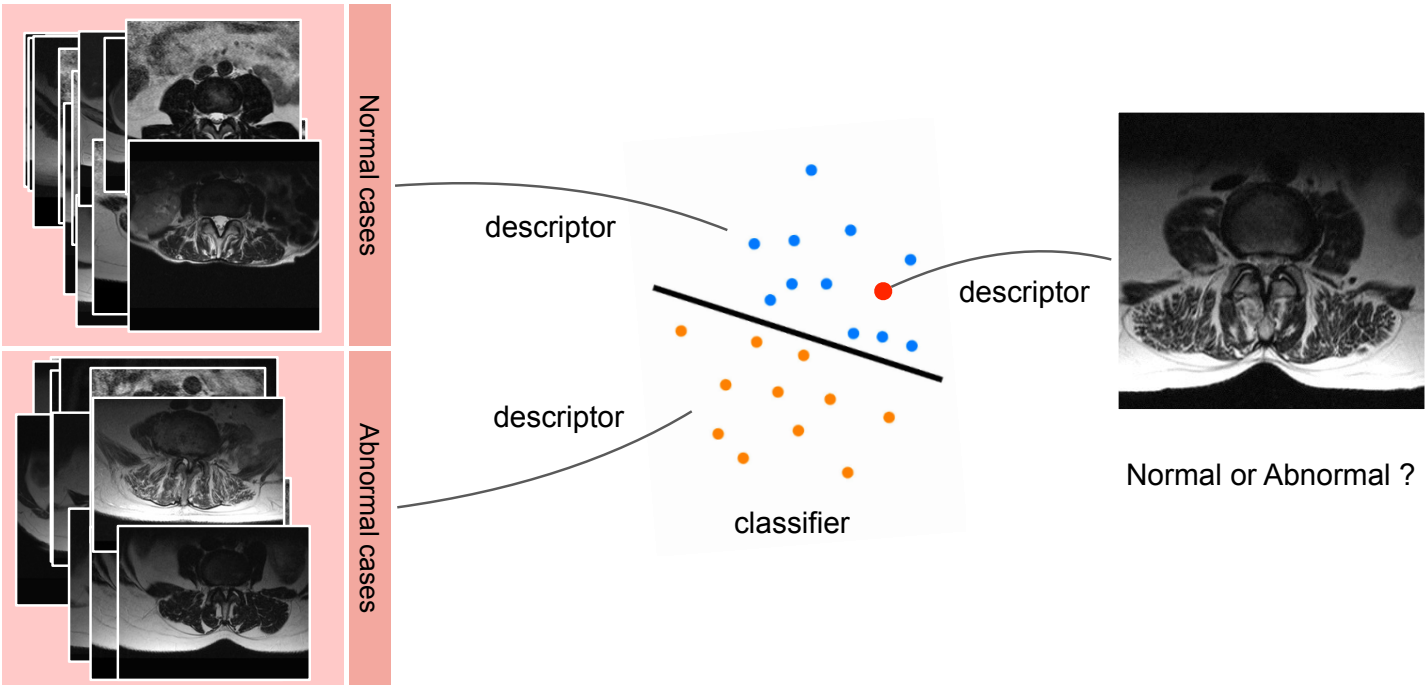
Classification





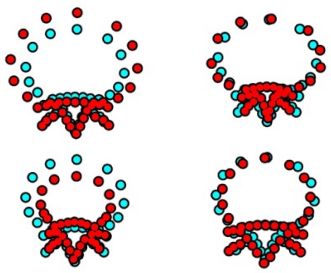
Normal or Abnormal ?

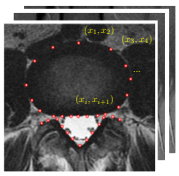
Classification

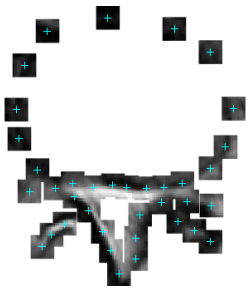


Related work

- Shape (point distribution model)







Examples are

- Active shape models
- Constrained local models
- Deformable part models

Related work

- Part-based models

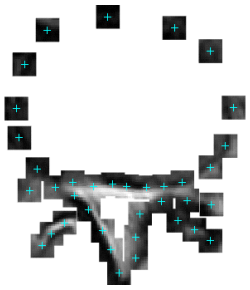
  A collection of parts with geometrical constraints on their spatial configuration

  — Part descriptor

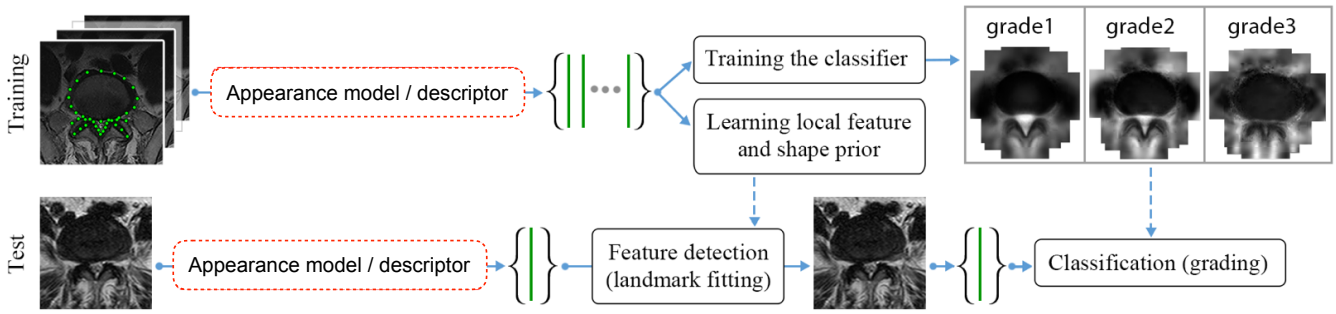
  How to represent the local appearance

  — Shape regularisation

  How to learn the shape variations in the training samples and use it to constrain the shape within plausible variation during testing

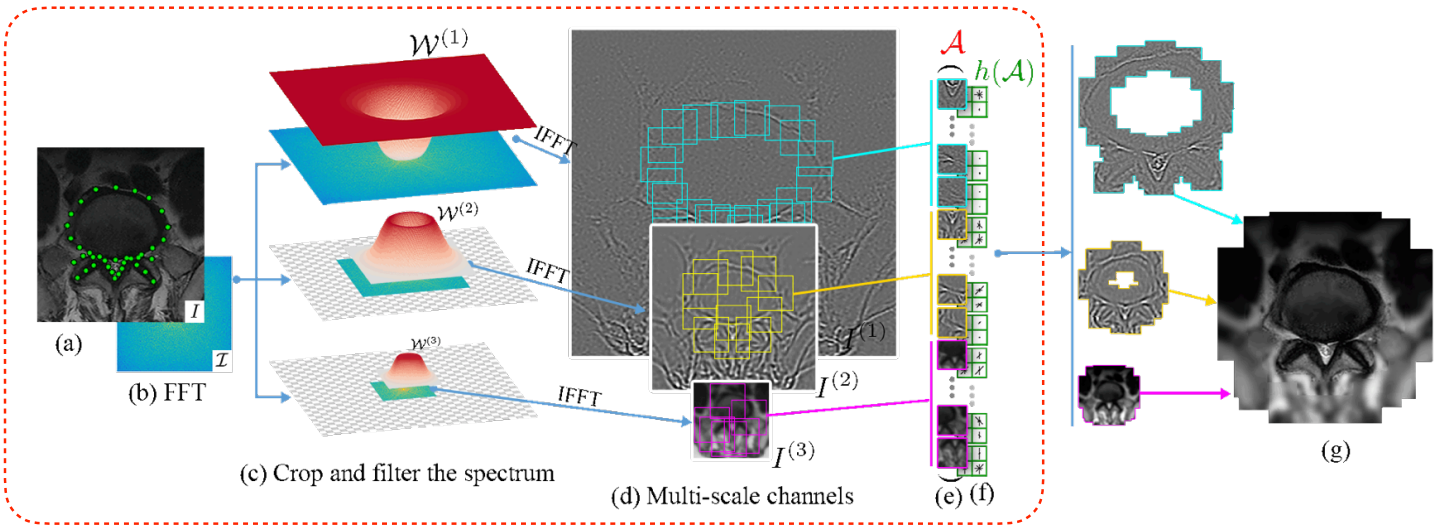


Pipeline of a standard approach



- To represent the anatomy with coherent appearance models or feature descriptors, and vectorise the representations as inputs for training a classifier.
  - Landmark detection
  - Classification

The proposed wavelet appearance pyramid



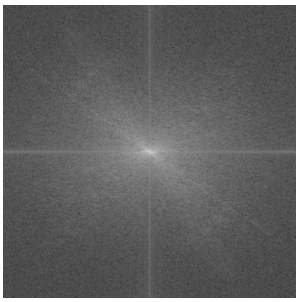
Notations

- Appearance pyramid [1]
  - An assembly of part based models built on an Gaussian image pyramid
  - A multi-scale part based active appearance model
  - Modelling the appearance variability with local rigid translations of the parts as well as linear modes of the assembly of the parts
- Wavelet appearance pyramid
  - An appearance pyramid built on an wavelet image pyramid
  - Decompose the appearance into pyramidal channels at complementary scale ranges with wavelets, and represent each channel with a part based model.

[1] Zhang, Q., Bhalerao, A., Dickenson, E., & Hutchinson, C. (2016). **Active appearance pyramids** for object parametrisation and fitting. *Medical Image Analysis*, *32*, 101-114.

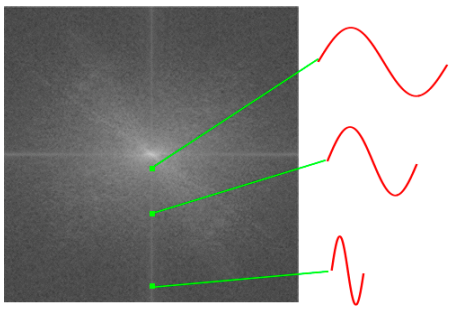
Explicit scale selection in the Fourier domain

Spectrum

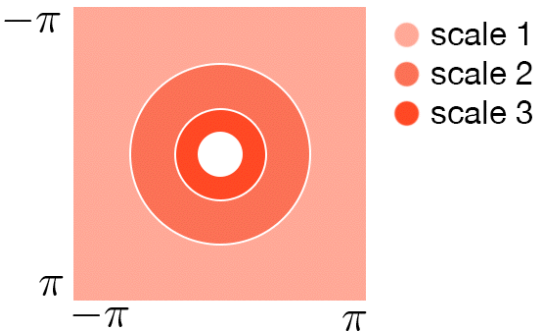


Explicit scale selection in the Fourier domain

Spectrum

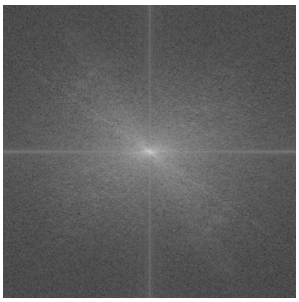


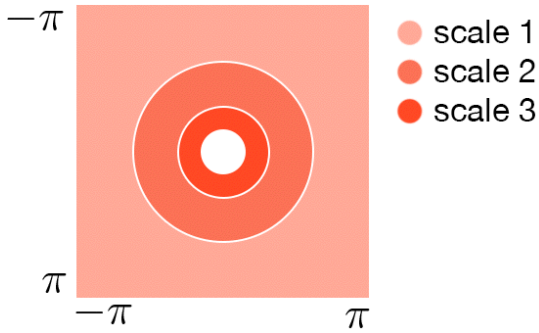
Subband



Explicit scale selection in the Fourier domain

- Looking for a basis function who is:
  - As smooth as possible: Large vanishing moments.
  - Uniform: The sum over scales is a constant.





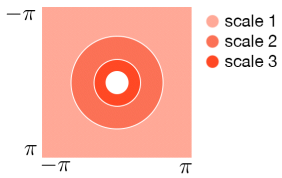
Explicit scale selection in the Fourier domain

- Looking for a basis function who is:
  - As smooth as possible: Large vanishing moments.
  - Uniform: The sum over scales is a constant.

- Loglets [1]:

$$\mathcal{W}(\boldsymbol{u}; s) = \mathrm{erf}\left(\alpha \log\left(\beta^{s+\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right) - \mathrm{erf}\left(\alpha \log\left(\beta^{s-\frac{1}{2}} \frac{\rho}{\rho_0}\right)\right)$$

$$\boldsymbol{u} = [\rho, \theta]$$

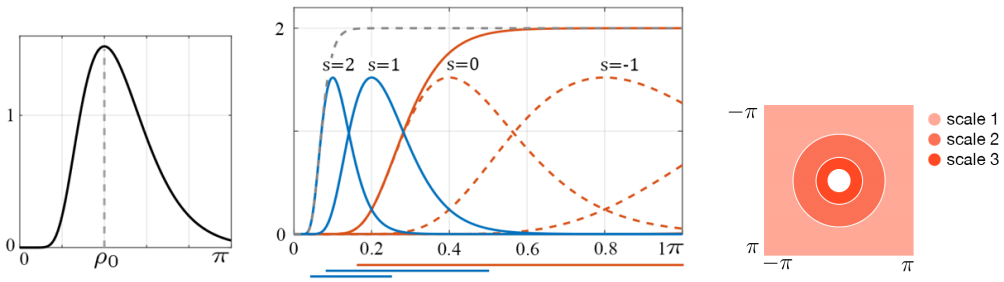


[1] Knutsson, Hans, and Mats Andersson. "Loglets: Generalized quadrature and phase for local spatio-temporal structure estimation." *Image Analysis*. Springer Berlin Heidelberg, 2003. 741-748.

Explicit scale selection in the Fourier domain

- Loglets:

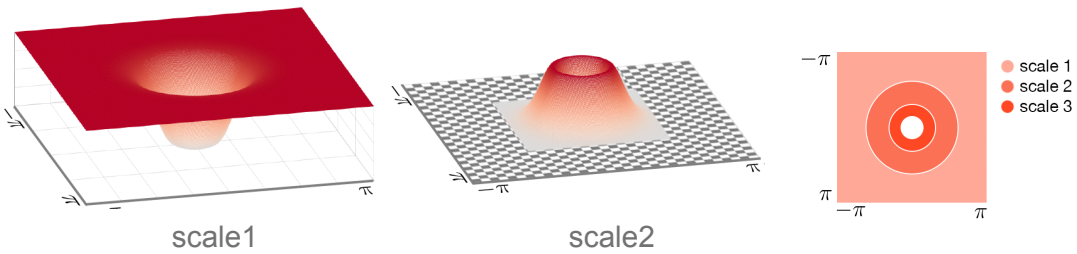
$$\mathcal{W}(\boldsymbol{u};s)=\mathrm{erf}\left(\alpha\log\left(\beta^{s+\frac{1}{2}}\frac{\rho}{\rho_0}\right)\right)-\mathrm{erf}\left(\alpha\log\left(\beta^{s-\frac{1}{2}}\frac{\rho}{\rho_0}\right)\right)$$

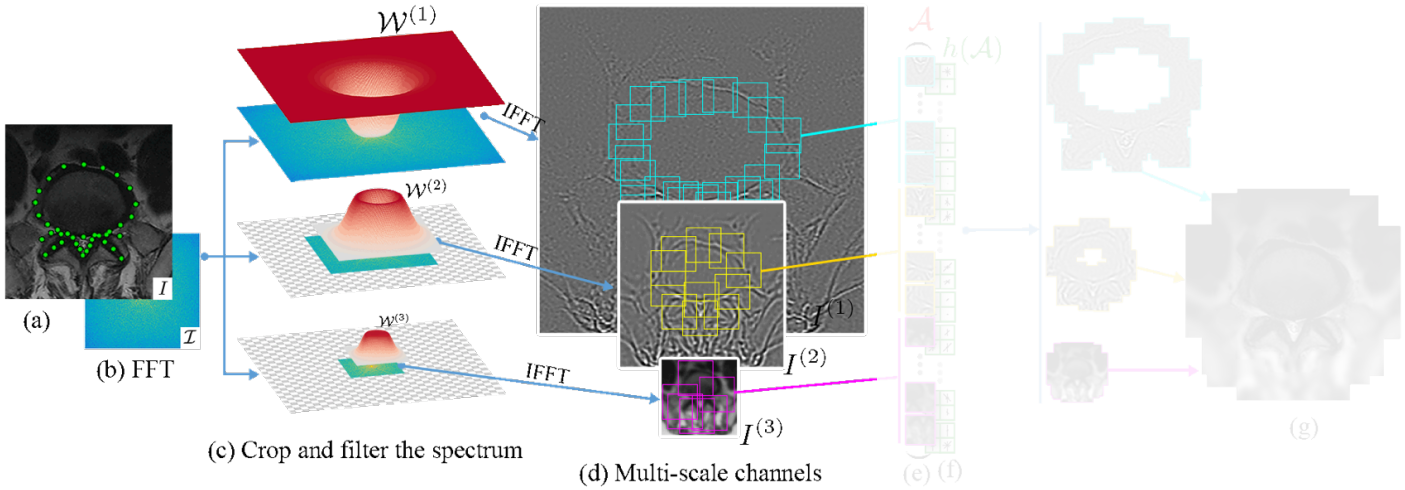


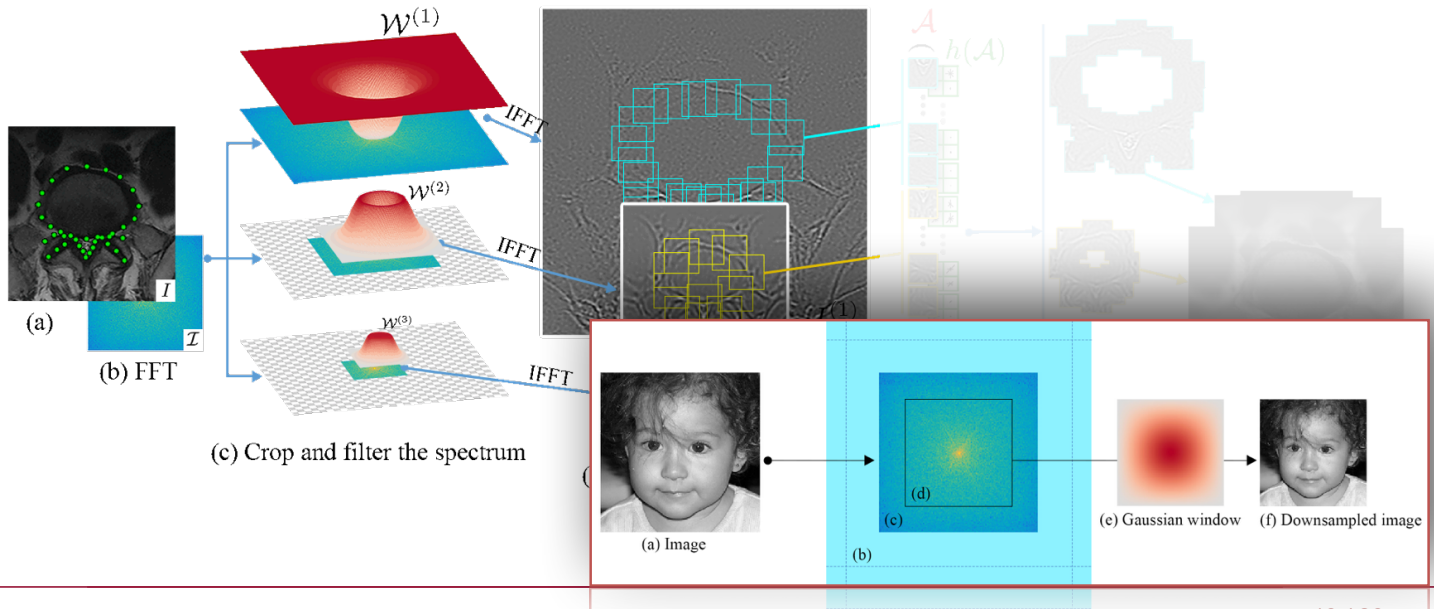
Explicit scale selection in the Fourier domain

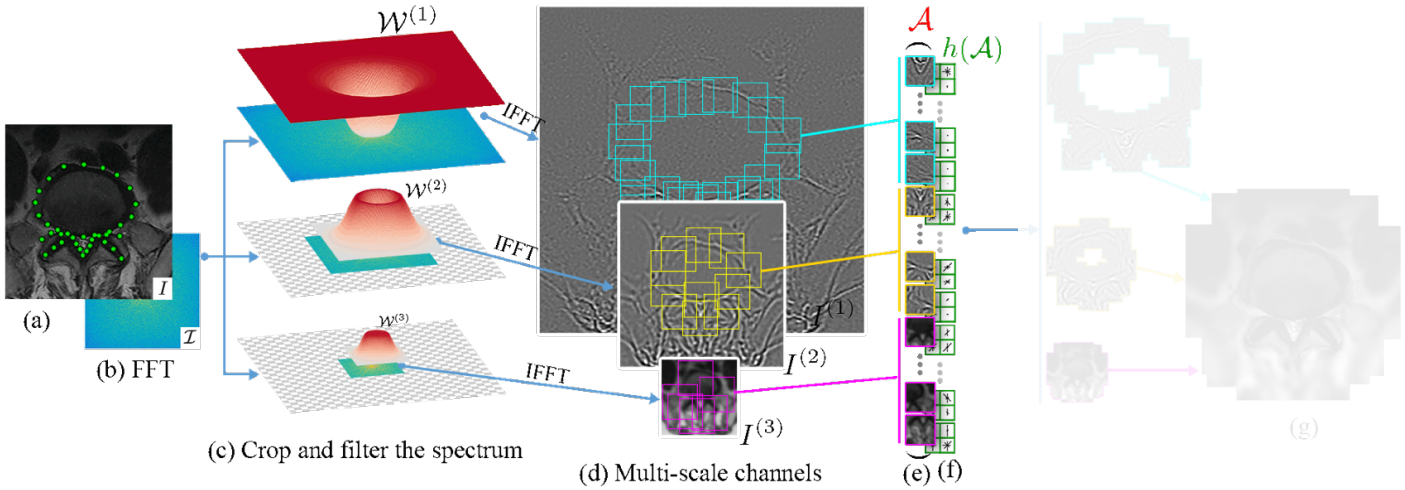
- Loglets:

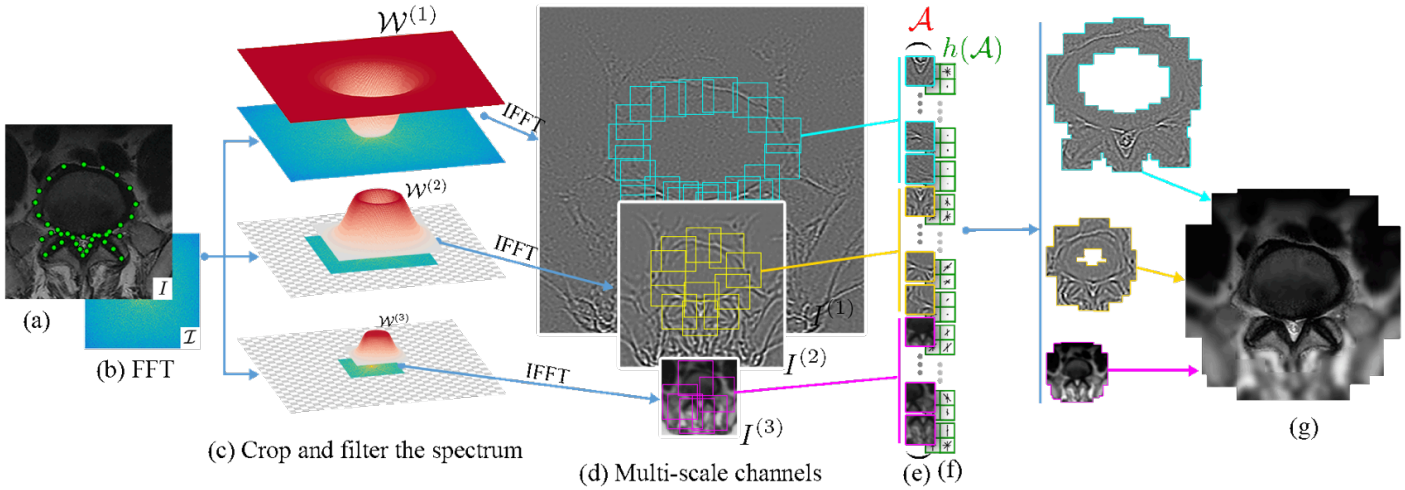
$$\mathcal{W}(\boldsymbol{u};s)=\mathrm{erf}\left(\alpha\log\left(\beta^{s+\frac{1}{2}}\frac{\rho}{\rho_0}\right)\right)-\mathrm{erf}\left(\alpha\log\left(\beta^{s-\frac{1}{2}}\frac{\rho}{\rho_0}\right)\right)$$



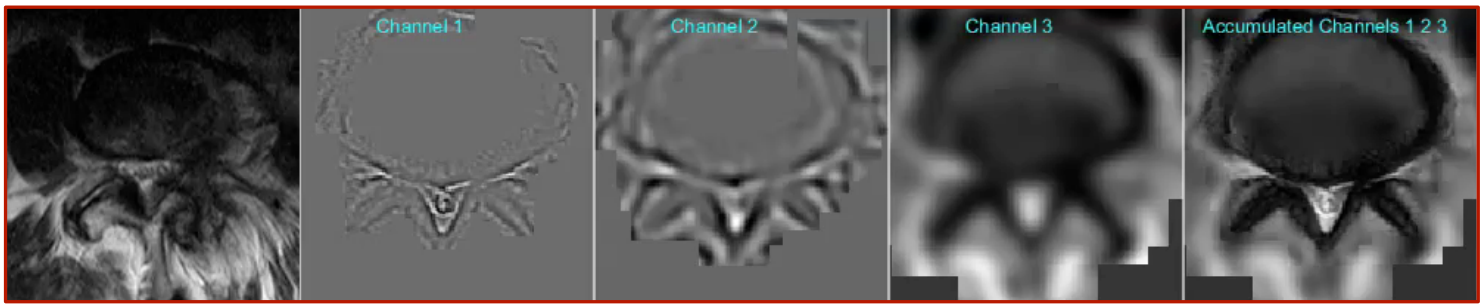

$\mathcal{W}^{(1)}$
IFFT
$\mathcal{W}^{(2)}$
IFFT
$\mathcal{W}^{(3)}$
IFFT
I
(a)
$\mathcal{I}$
(b) FFT
(c) Crop and filter the spectrum
$I^{(1)}$
$I^{(2)}$
$I^{(3)}$
(d) Multi-scale channels
$\mathcal{A}$
$h(\mathcal{A})$
(e) (f)
(g)


$\mathcal{W}^{(1)}$
IFFT
$\mathcal{W}^{(2)}$
IFFT
$\mathcal{W}^{(3)}$
IFFT
I
(a)
$\mathcal{I}$
(b) FFT
(c) Crop and filter the spectrum
$I^{(1)}$
$\mathcal{A}$
$h(\mathcal{A})$
(a) Image
(b)
(c)
(d)
(e) Gaussian window
(f) Downsampled image

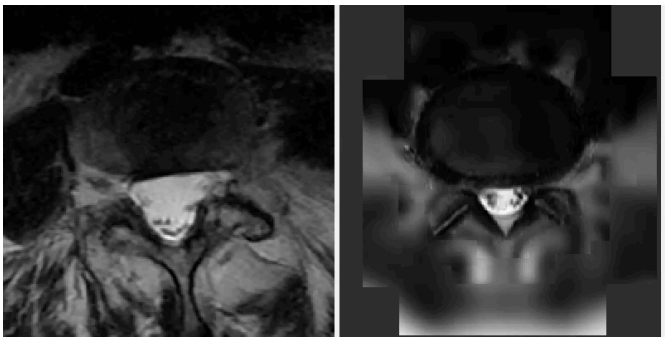

$\mathcal{W}^{(1)}$
$\mathcal{W}^{(2)}$
$\mathcal{W}^{(3)}$
IFFT
IFFT
IFFT
I
$\mathcal{I}$
(a)
(b) FFT
(c) Crop and filter the spectrum
$I^{(1)}$
$I^{(2)}$
$I^{(3)}$
(d) Multi-scale channels
$\mathcal{A}$
$h(\mathcal{A})$
(e) (f)
(g)

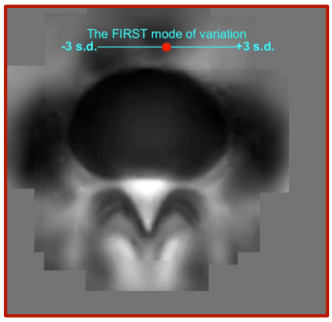

$\mathcal{W}^{(1)}$
$\mathcal{W}^{(2)}$
$\mathcal{W}^{(3)}$
IFFT
IFFT
IFFT
I
$\mathcal{I}$
(a)
(b) FFT
(c) Crop and filter the spectrum
$I^{(1)}$
$I^{(2)}$
$I^{(3)}$
(d) Multi-scale channels
$\mathcal{A}$
$h(\mathcal{A})$
(e) (f)
(g)

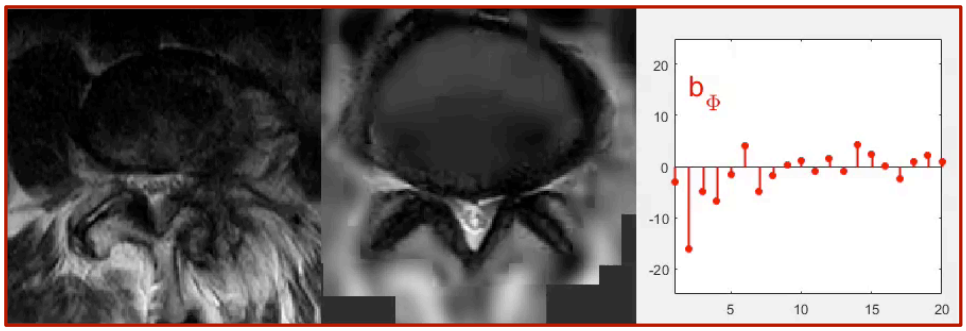
- The variability among the population can be approximated by local rigid translations of the parts in each channel, and the appearance changes by linear modes of the assembly of the parts.



- Groupwise correspondence is built.





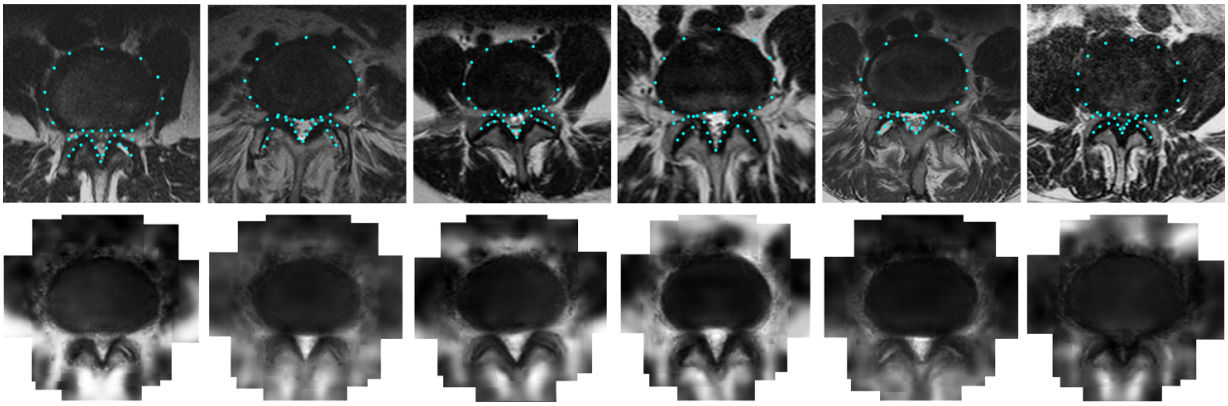


- The most significant variations in the population can be learned by PCA and the dimensionality reduced by preserving the first several significant components, which span a feature space. A WAP therefore can be represented in the feature space by a compact set of parameters b_\Phi, which can be used as inputs of a classifier.

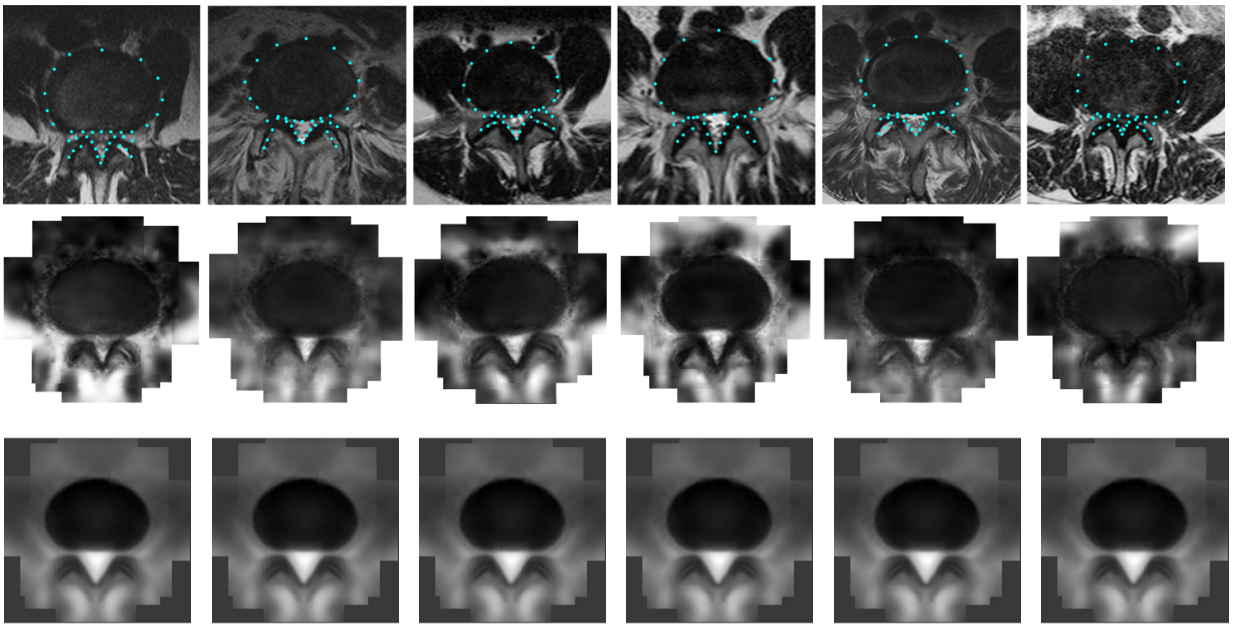
Evaluation

Table 1: Performance of landmark detection

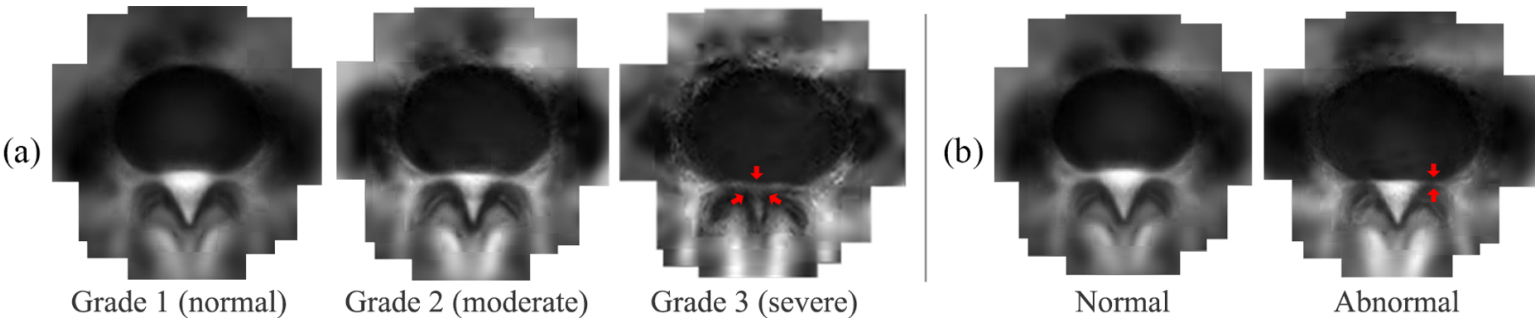
Metrics	AAM	ASM	CLM	WAP$^-$	WAP
PtoBD (in pixels)	3.10 ± 1.29	2.51 ± 1.32	2.34 ± 1.15	1.95 ± 0.92	**1.87±0.73**
DSC (%)	90.6 ± 4.9	92.1 ± 5.2	92.4 ± 5.2	93.9 ± 3.3	**94.7±2.6**



Evaluation



Evaluation



Evaluation

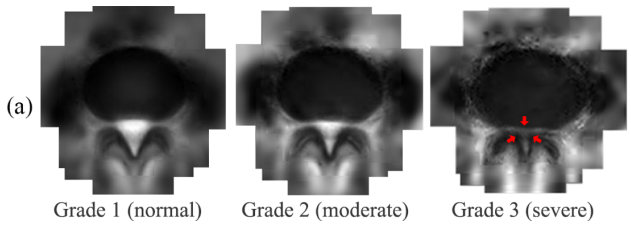


Table 2: Agreement of classification and grading of central stenosis

Method	Accuracy (%) of classification			MAE of grading			RMSE of grading		
	L3/4	L4/5	L5/S1	L3/4	L4/5	L5/S1	L3/4	L4/5	L5/S1
ASM	79.1±4.8	77.4±4.3	81.7±4.5	0.25	0.31	0.20	0.55	0.67	0.48
AAM	70.1±7.1	69.7±7.3	71.3±8.8	0.41	0.44	0.32	0.72	0.79	0.58
CLM	81.0±4.9	82.4±4.5	82.7±4.4	0.23	0.25	0.23	0.53	0.56	0.52
WAP$^-$	80.7±4.9	82.1±4.6	84.7±4.2	0.23	0.25	0.18	0.53	0.58	0.47
WAP	**84.7±4.6**	**84.5±4.3**	**85.9±4.2**	**0.19**	**0.21**	**0.16**	**0.48**	**0.54**	**0.44**

Evaluation

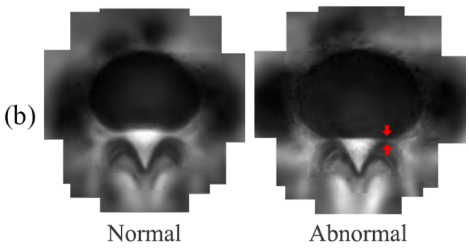


Table 3: Accuracy (%) of classification of foreminal stenosis

Anatomy	ASM	AAM	CLM	WAP$^-$	WAP
L3/4	83.3 ± 3.8	73.3 ± 5.5	83.1 ± 4.7	84.3 ± 4.1	**85.0±3.9**
L4/5	82.4 ± 4.6	76.2 ± 5.8	83.3 ± 4.3	86.9 ± 3.9	**87.8±3.5**
L5/S1	81.8 ± 4.7	74.5 ± 5.7	82.9 ± 4.5	85.2 ± 4.3	**85.7±4.3**

Wavelet Appearance Pyramid for Landmark Detection and Pathology Classification: application to Lumbar Spinal Stenosis

Thank you

WARWICK

Full name: Yiwei Zhou

Year of Study: PhD Year 3

Department: Computer Science

Submission Type: Presentation

Title: Sentiment and Similarity Analysis of User-generated Content

Abstract: The development of social media has provided the researchers new opportunities to understand people and events in real-world through the massive user-generated data.

However, the unstructured, informal, redundant characteristics of user-generated data have also brought up new challenges for traditional natural language processing approaches.

In this thesis, we proposed several natural language processing approaches, to answer the demands of different application scenarios.

In particular, we focused on the area of sentiment analysis and similarity analysis, and proposed several broadly applicable algorithms.

We extensively evaluated our approaches on various publicly available datasets, from multilingual Wikipedia to tweets, to illustrate their effectiveness.

We made several findings, from the existence of sentiment bias and representation differences on multilingual Wikipedia, to the quickness and comprehensiveness of tweets in reporting real-world events.

We built several applications, from detecting the reputation-influential sentences on Wikipedia, to employing Wikipedia contexts to improve the performance of entity-centric news retrieval applications, to generating timeline for high-impact events in real-time.

Keywords: sentiment analysis, similarity analysis, user-generated content

Abstract Word Count (Est): 159