WARWICK
THE UNIVERSITY OF WARWICK

Introduction
CS1D6: Introduction to data and statistics

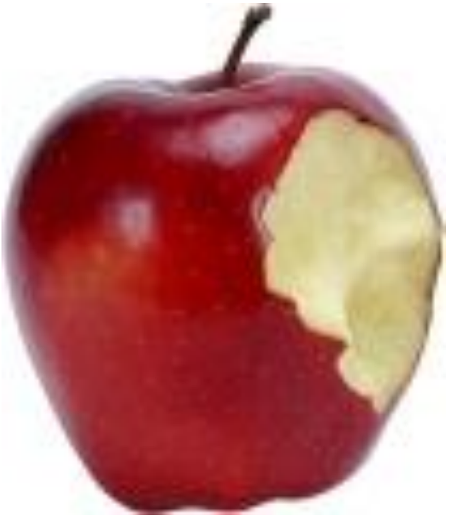
Dr. Fayyaz Minhas

Department of Computer Science
University of Warwick

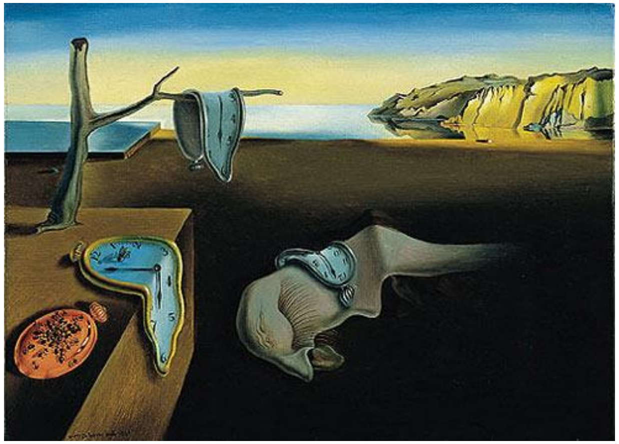
Apples

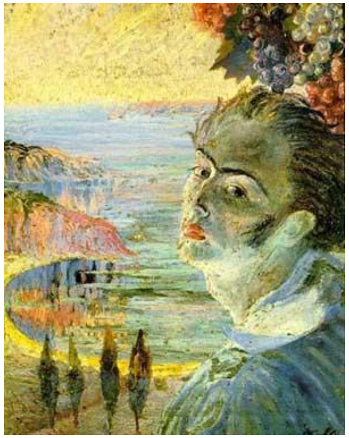
Oranges

What is this?

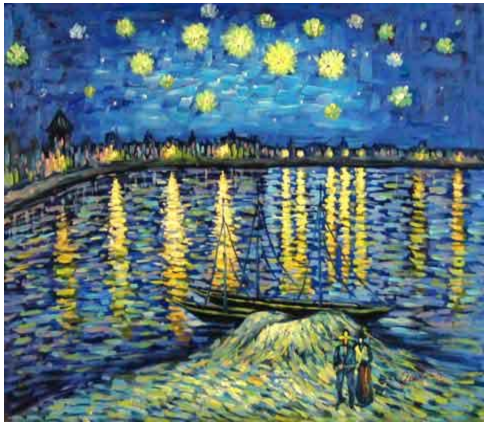


Paintings by two different painters

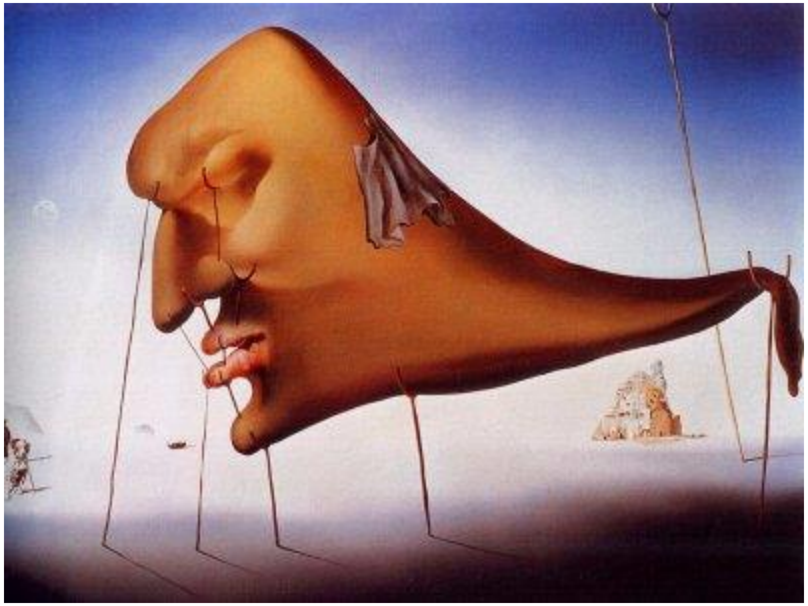




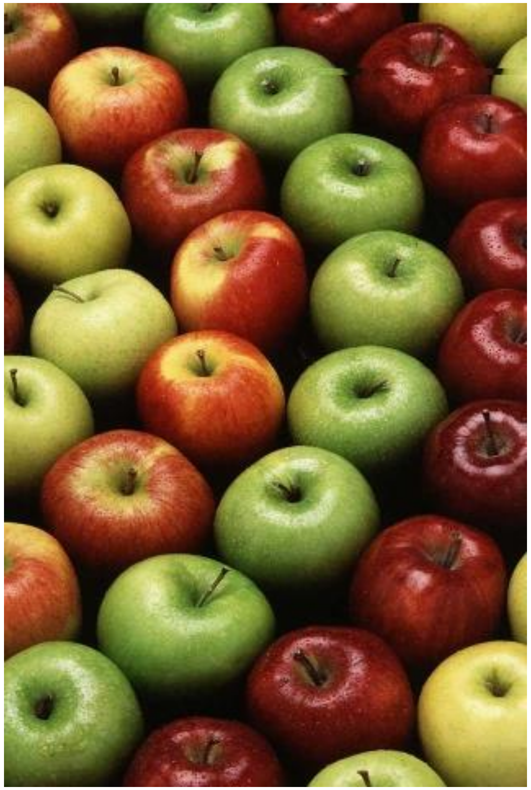
Who’s painting is this?



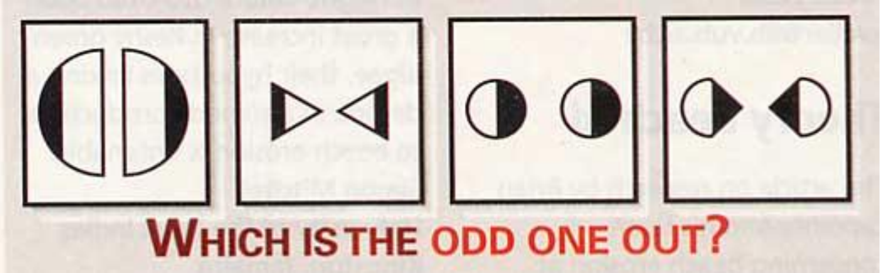
And this?



How many categories (clusters) are there?



Find the odd one out!



Predict the series

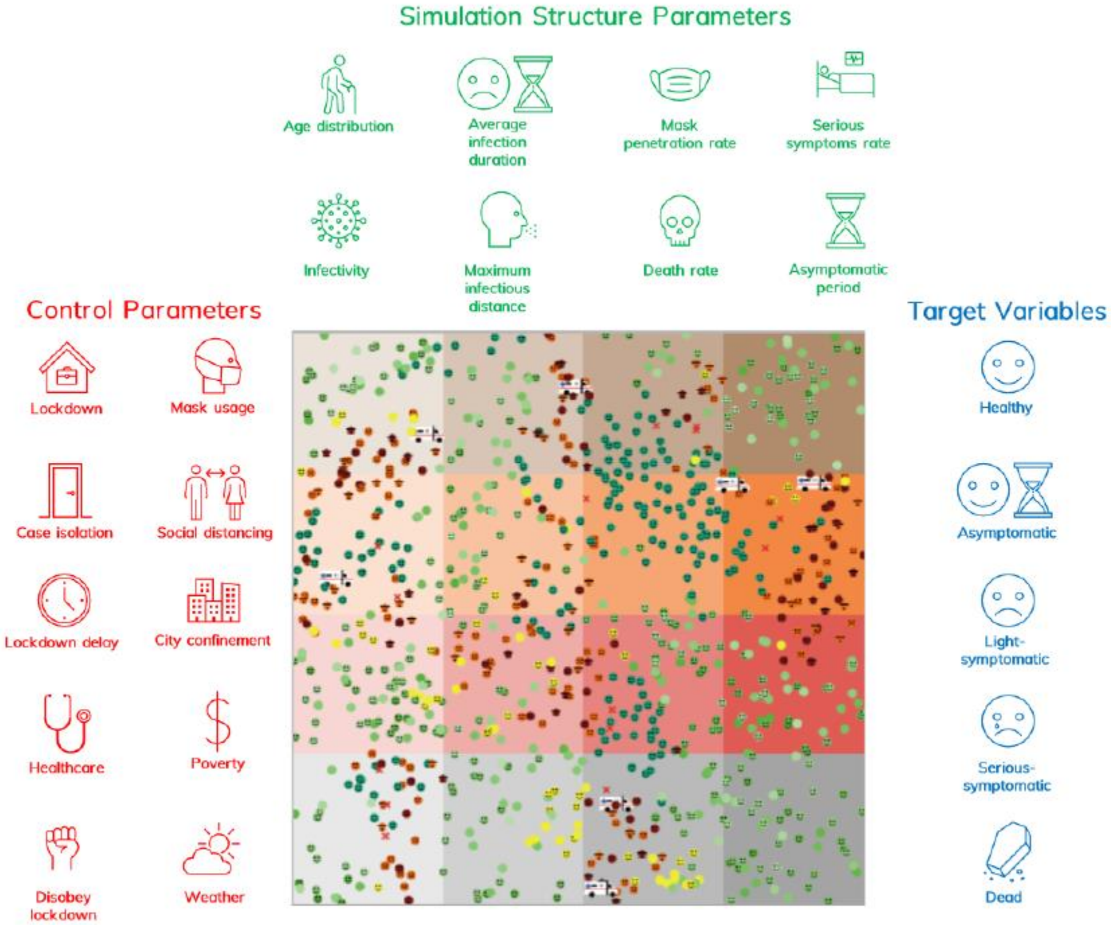
- 1,1,2,3,5,8,13,…

Question?

- Consider the vectors
  - $X_1=[1\ 2\ 1\ 4]^T$
  - $X_2=[2\ 4\ 2\ 4]^T$
  - $X_3=[0\ 0\ 0\ 4]^T$
  - $X_4=[3\ 6\ 3\ 4]^T$
  - $X_5=[4\ 8\ 4\ 4]^T$
- To store each vector, how many dimensions (or variables) do we need?

Learning to drive

How can we simulate COVID-19?

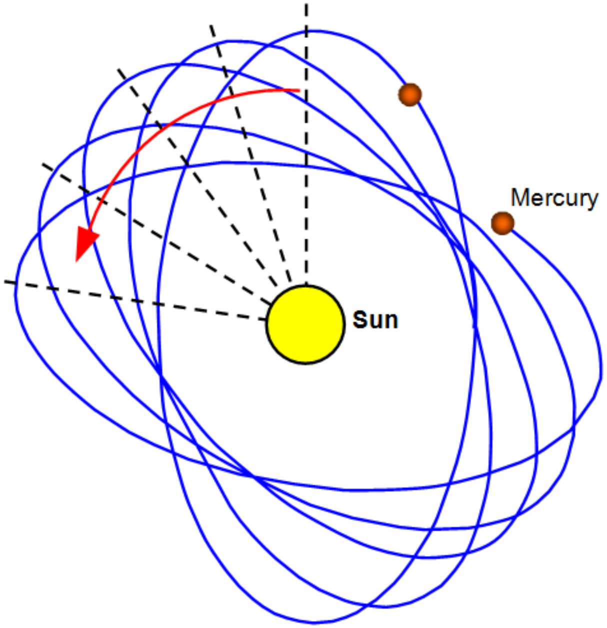


Questions

• How were you able to recognize that the object shown was indeed an apple?	Classification
• How were you able to discriminate between the paintings from two different painters?	Classification
• How were you able to find out the different types of apples in the picture?	Clustering
• How did you manage to find the next number in the series?	Regression
• How were you able to find which dimension was redundant?	Dimensionality Reduction
• How were you able to find the odd one out?	Anomaly Detection
• Learning to drive / write?	Reinforcement learning

Example: Human Learning

- Science is based on developing and testing hypothesis that “explain” our universe
- For example:
  - Newton’s Formula F = ma explains the motion of an object of mass m when a force F is applied to it
  - Scientists observed that Newtonian mechanics does not “explain” the motion of mercury properly
  - This led to the development of theory of relativity by Einstein which explains it!!
- We constantly try to develop and refine models of the world and the universe
- However <u>sometimes</u> it gets hard!



Why do we need computers?

- …ATTCGAGGATTACACCGTAAGAAATTT…
- …ATCGCCTGATTACATATATACCGTTGG…
- ….AGATTAAATCGTTCGATTCACATTGAC
- **Deduction vs. Induction Reasoning**
- **High dimensions**
- **Required Reading**
  - Halevy, Alon, Peter Norvig, and Fernando Pereira. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems*, 2009.

Statistical analysis

- Central Focus
  - Data representation
  - Identify/discover patterns
    - Discriminate
    - Regress
    - Cluster
    - Identify anomalies
  - Drawing meaningful conclusions

How to get there?

- Basic statistics, programming, visualization
- Understanding Univariate Sampling
- Multivariate Analysis
- Multivariate Visualization
- Understanding Correlation
- MV Correlation
- Linear Models
- Dimensionality Reduction: PCA
- Clustering
- Regression: OLS
- GLMs
- Applications
- Limitations

Tentative

Day	9am	10am	11am	Noon	1pm	2pm	3pm	4pm
Mon	Intro/C1 Lecture	Intro to Python	C1/C2 Lecture	Lunch	C2 Lecture	C1/2 Lab	C1/2 Lab	C1/2 Lab
Tues	C3 Lecture	C3 Lecture	C3 Lab	Lunch	C4 Lecture	C4 Lecture	C4 Lab	C4 Lab
Wed	Introduction	Sampling	C5 Lab / Catchup	Lunch	MVA	Linear Models & Visualization	Lab	Lab
Thur	PCA	LDA	Lab	Lunch	LDA	Clustering	Lab	Lab
Fri	Regression	Regression	Lab	Lunch	Applications	Applications & Limitations	Lab	Lab

End of Lecture-1

> We want to make a machine that will be proud of us.
>
> - Danny Hillis