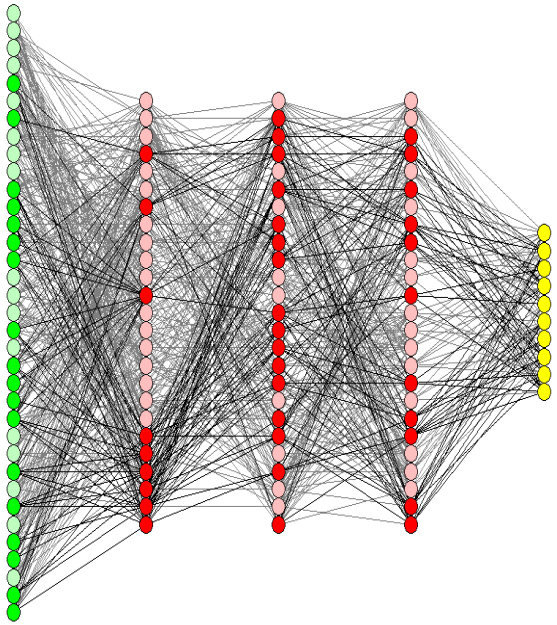
Understandable Explanations for Black-Box Models

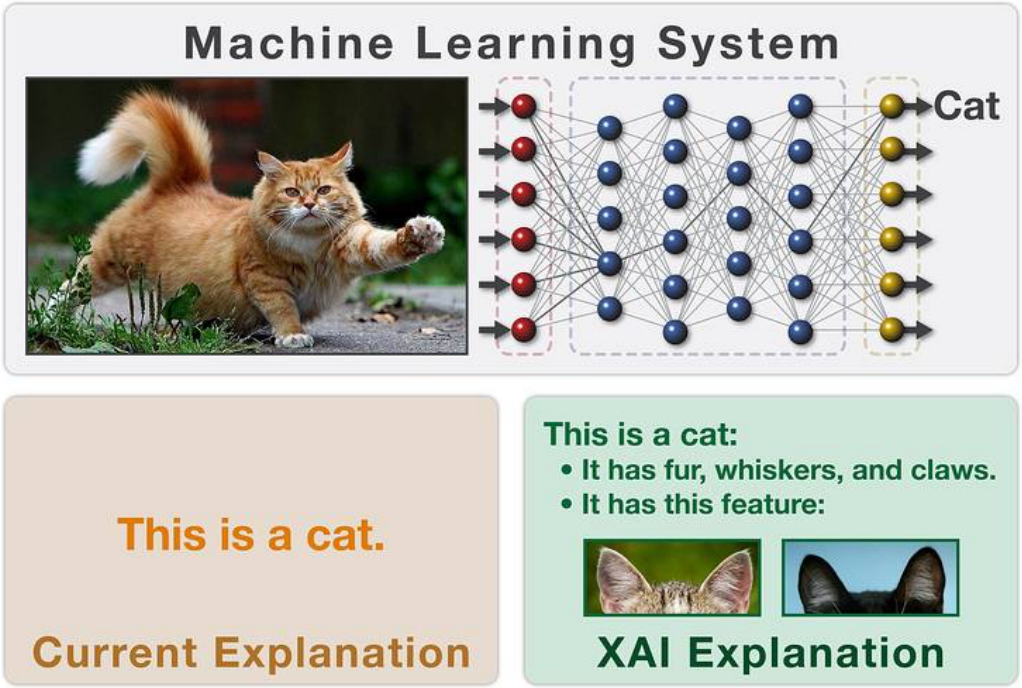
Tillman Weyde
Work with Roberto Confalonieri, Fermin Moscoso, Tarek Besold

AI in Finance

- Recent progress: **AI** for **Finance** with **Deep Learning**
  - **Prediction** – prices, performance, trends
  - **NLP** – from sentiment analysis to chatbots …
  - **Modelling customers** – credit, fraud, design, marketing
  - Possible because of **Big Data**
    - Social media, mobile devices
    - Remote sensing, image analysis …
- Problem: Deep Learning is a **Black Box**
  - for investors, banks, insurers, regulators, developers, scientists



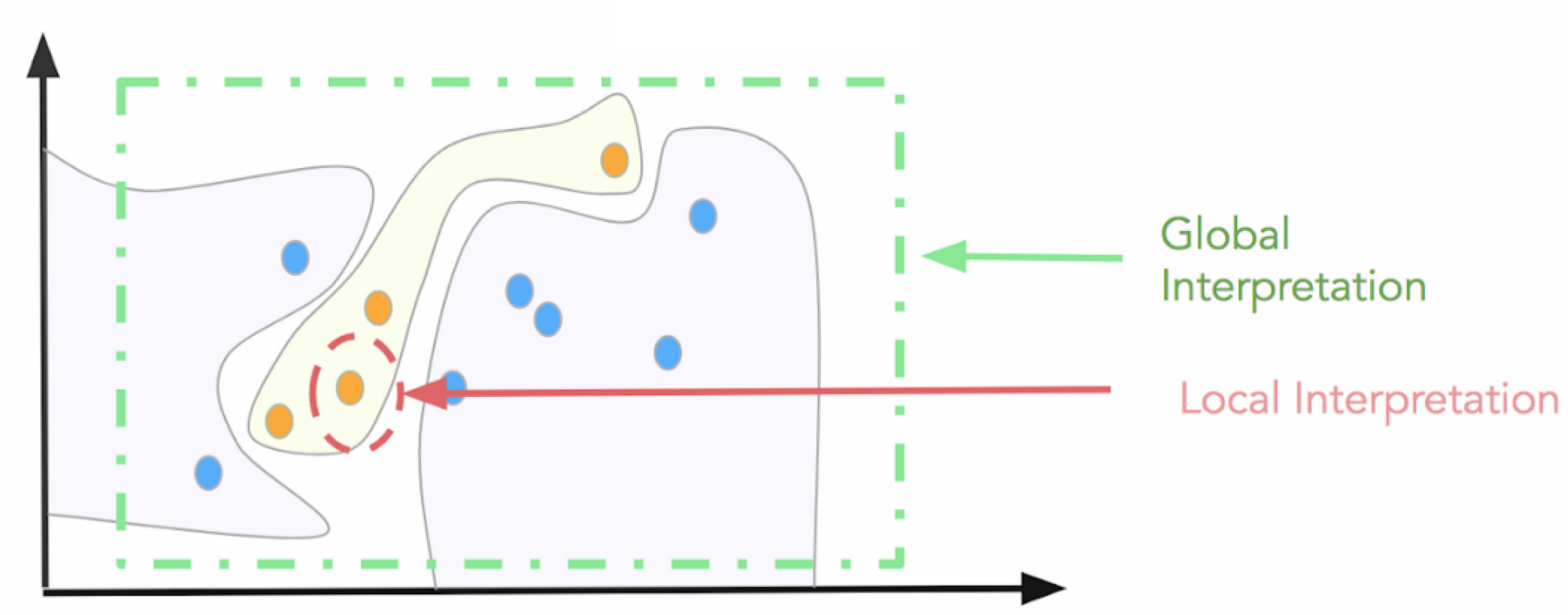
Explaining Black Box Models



- Need **transparency** and **scrutiny** for
  - **performance, robustness, efficiency, ethics, fairness**
- **DARPA** XAI Program (2016)
- **GDPR** Customer right for explanation (2018)

Explaining Black Box Models

- **Built-in:** combine performance and interpretability (rare)
- **Post-hoc:** Train model and explain afterwards (surrogate)
- **Global**: represent the whole model
- **Local:** focus on a single item



Understandable Explanations

- **Explanations** should be
  - accurate
  - actionable
  - **understandable**

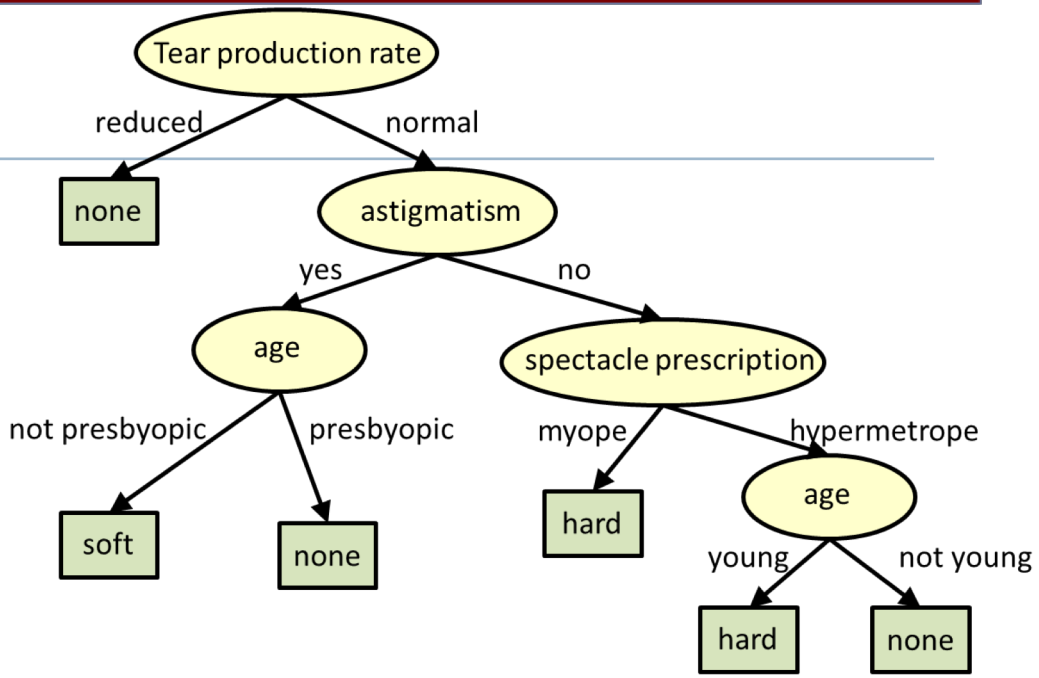
- Focus here on **understandability**
  - Goal: **match human thinking better**
  - Approach: Use **background knowledge**
  - Needed: Identify **technical concepts** for understandability
  - Test: **Experiments** with **human subjects**

Explaining with Decision Trees

- **Global post-hoc approach:**
  - Approximate trained black box model with explainable alternative (surrogate)

- **Trepan** – (Craven & Shavlik 1993)
  - Extract Decision Trees from any black box model
  - Sample output from trained model
  - Build DT on original features
  - Typically better DT than with original data alone

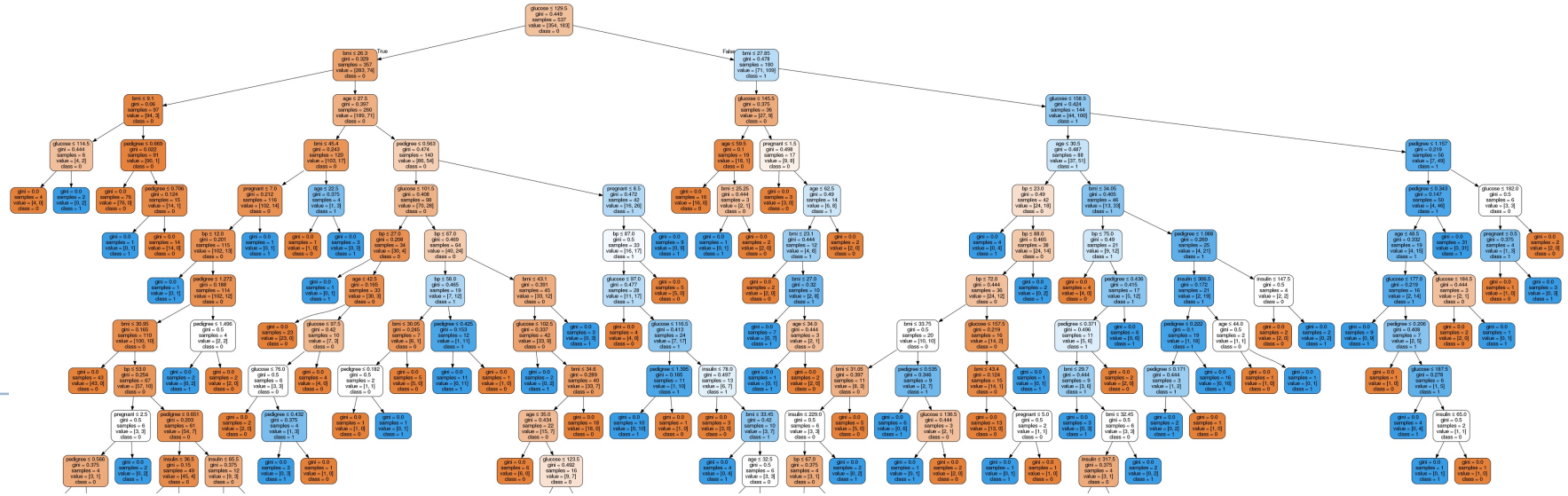
Decision Trees



- **DT**: **easy** to **understand**
  - Elements have **meaning**
  - Readable **rules**
  - **Split feature** selection by **information gain**

$$\mathrm{IG}(X_i, S) := H(S) - \sum_{j=1}^{k} \frac{|S_j|}{|S|} H(S_j)$$

- … but can become unwieldly



Bringing in Background Knowledge

- **Background knowledge** modelled in **ontologies**
  - Taxonomy as **concept hierarchy** (e.g. biology, medicine, library catalogues, product catalogues)
  - **Ontology** adds **logic** (constraints, additional information, e.g. family tree, Gene Ontology, general knowledge: DBPedia, SUMO)
- **Hypothesis**: **general concepts** are **easier to process** for humans
  - Quantified as **Information Content** IC

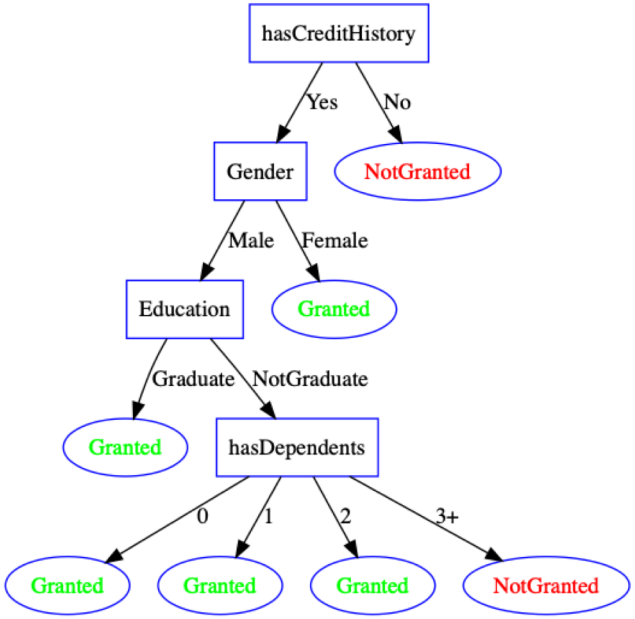
$$\mathrm{IC}(X_i) := 1 - \frac{\log\left(|\mathsf{subConcepts}(X_i)|\right)}{\log\left(|\mathsf{sub}(\mathcal{T})|\right)}$$

Trepan Reloaded: Idea

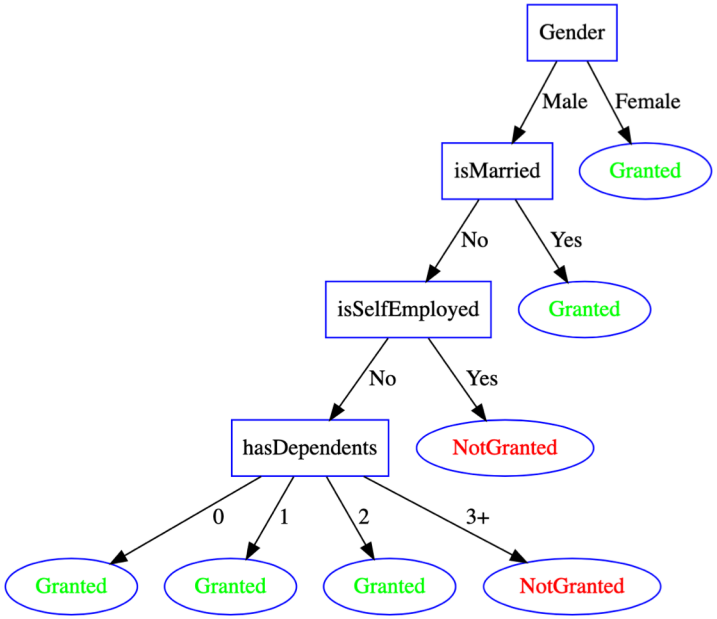
- **Prefer general features** for **split nodes**
- New **reward function** for selecting split features:

$$\mathrm{IG}'(X_i, S|\mathrm{IC}) := (1 - \mathrm{IC}(X_i))\mathrm{IG}(X_i, S)$$

Trepan



Trepan Reloaded



Evaluation

- **User experiment**
  - Human **performance** and **subjective ratings**
    - Online, 63 subjects, age 33 (±12.23), 46f/17m
- Datasets
  - **Finance**: Kaggle Loan Dataset (selected by 34 subjects)
  - **Medical**: Cleveland Heart Disease Data (selected 29 subjects)
- **Technical factors**
  - Tree ***syntactic complexity*** for *n* leaves and *b* branches

$$\alpha\frac{n}{k} + (1-\alpha)\frac{b}{k^2}$$

Experiments I

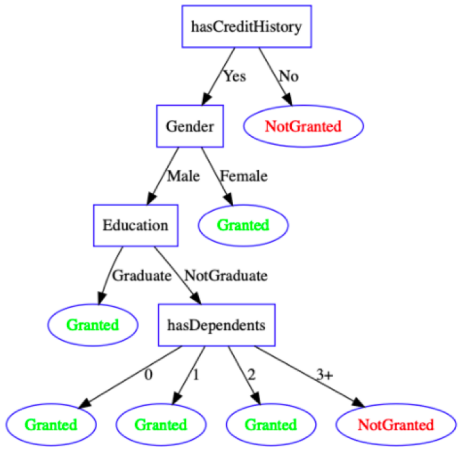
- **Classification** task
- **Determine class** of given case with DT
- 6 samples (2 each small/medium/large)
- Also rating **confidence** and **understandability**

EXAMPLE OF CLASSIFICATION TASK

In the classification task you will be asked to classify an example using a classification tree that will be shown to you.
Please have a look at this page and familiarize yourself with the task and the questions. The following pages will follow a similar pattern.

Classify the example at the *top* using the classification tree at the *bottom*.

Attribute	Value
Gender	Female
isMarried	No
hasDependents	0
Education	Graduate
isSelfEmployed	No
ApplicantIncome	3510
CoApplicantIncome	0
hasLoanAmount	76
hasLoanAmountTerm	360
hasCreditHistory	No
PropertyArea	Urban



1. The example is classified as / belongs to the class:

- Granted
- NotGranted

Experiments 2

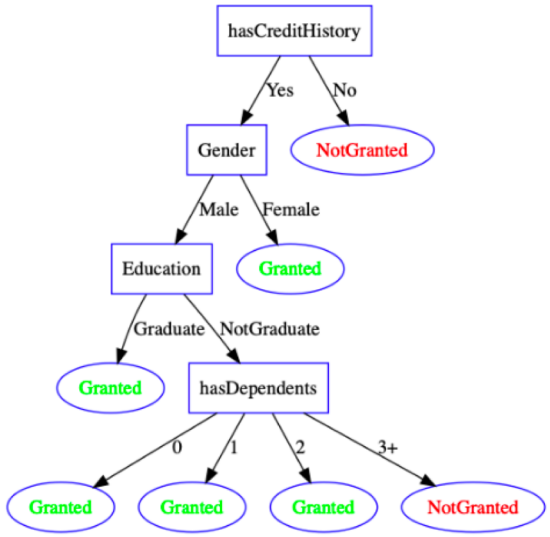
- **Inspection** task
- **Determine** if a given **statement** is **true** in a DT
- 6 samples (2 each small/medium/large)
- Also rating **confidence** and **understandability**

EXAMPLE OF INSPECTION TASK

In the inspection task you will be asked to tell whether a sentence describing (part of) a classification tree is true or
Please have a look at this page and familiarize yourself with the task and the questions. The following pages will fo

Is the **following statement true or false** with respect to the **classification tree** shown below?

You are a female; your level of education can affect the decision outcome.

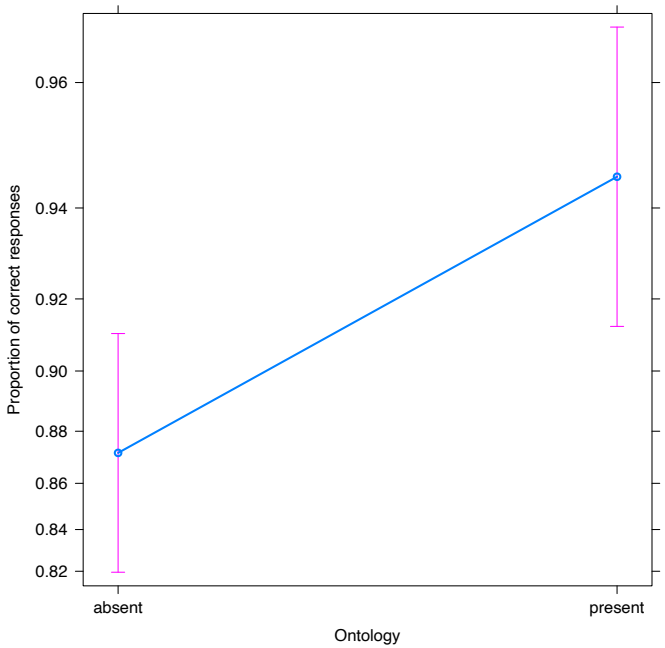


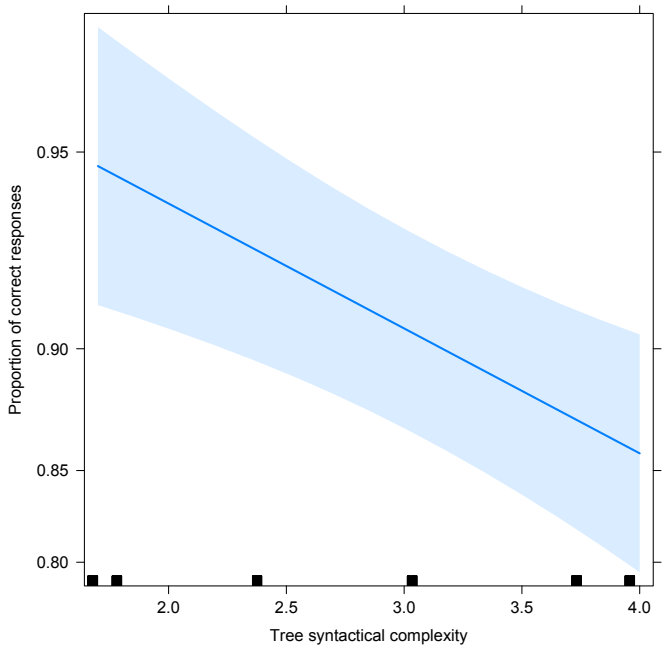
1. The above statement is:

- ○ True
- ◉ False

Results Experiments 1 & 2

- **Correctness** of responses: mixed-effect logistic regression (task, ontology, syntactical complexity)

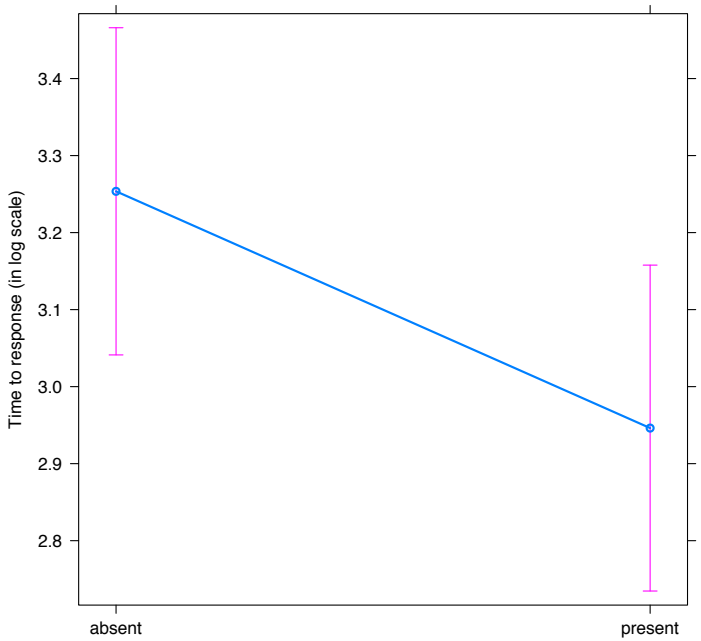


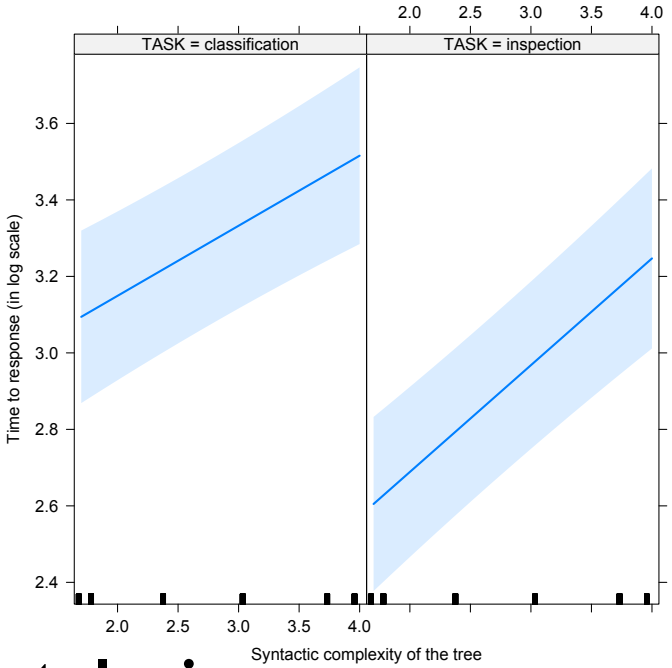


- Responses **with ontology** are **more often correct**
- Higher **syntactic complexity** makes task **harder**
- Subjective **understandability** very **similar**
- All effects **significant** at $p < 0.01$

Results Experiments 1 & 2

- Response **time** (correct responses): mixed-effect linear regression (task, ontology, syntactical complexity)



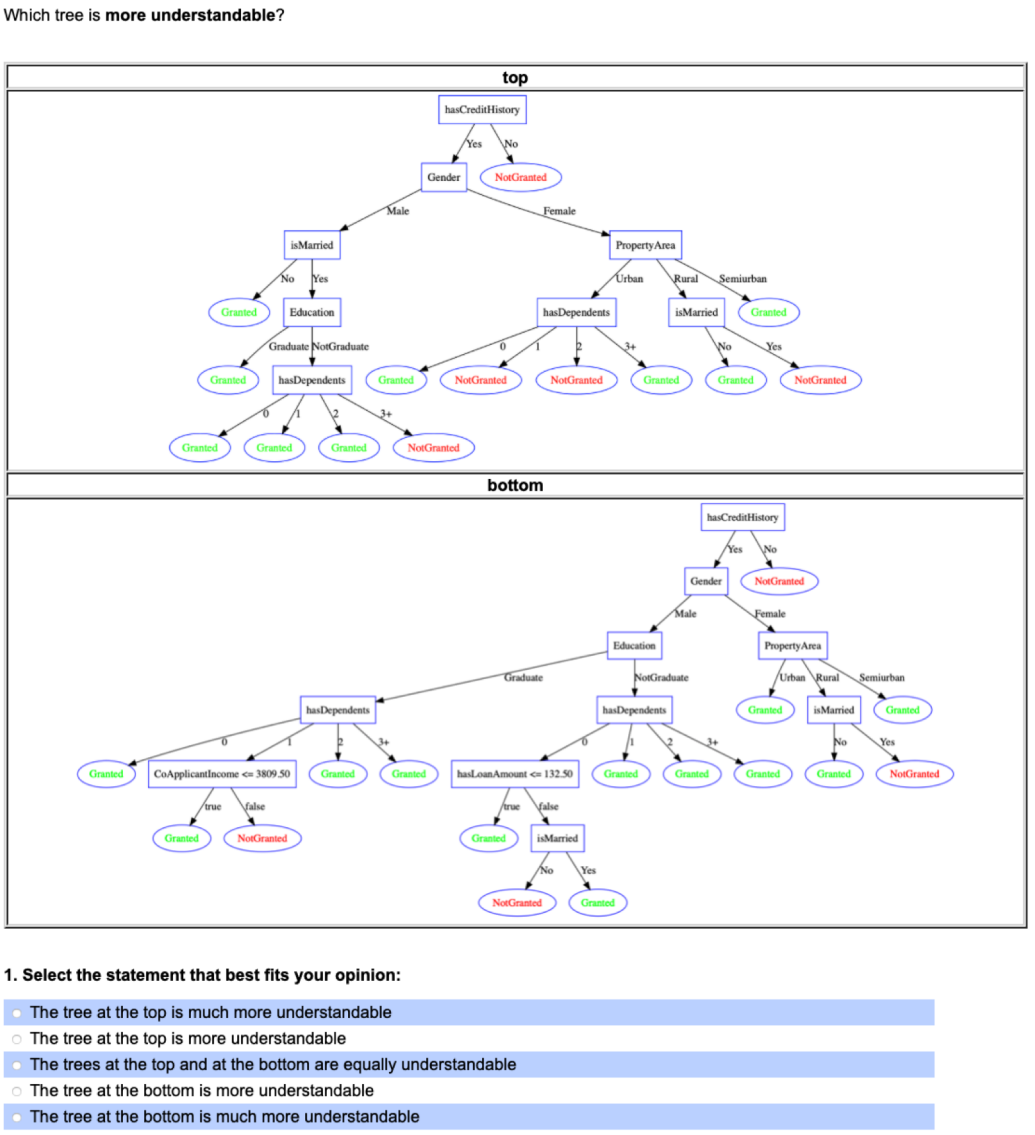


- Correct responses **quicker** with **ontologies**
- Higher **syntactic complexity** makes task **harder**
- **Confidence** results similarly significant
- All effects **significant** at $p < 0.01$

Experiments 3

- **Comparison** task (subjective)
- Rating **which DT** is more **understandable**
- 3 samples

- **Result**: DT with **ontology subjectively more understandable** (p<0.01)



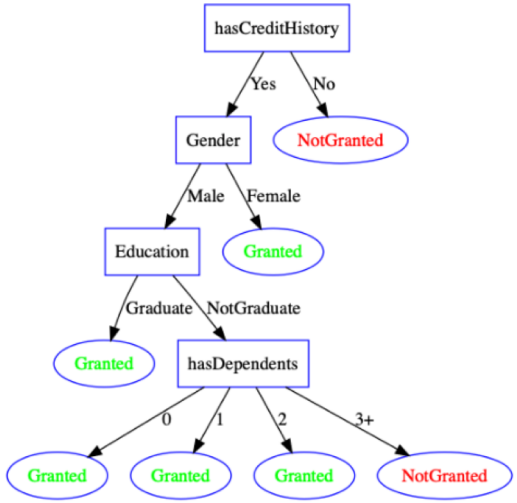
Experiments 4

- **Empowerment** (actionability)
- Determine how you could **change** the **outcome** for a given case with a given DT

EXAMPLE OF EMPOWERMENT TASK

In the empowerment task you will be asked to specify what event or what action you could take according to the decision tree to change a decision outcome.
Please have a look at this page and familiarize yourself with the task and the questions. The following pages will follow a similar pattern.

Specify what event could change the decision outcome.
In providing the answer, notice that you **can also change the premises provided**.



Please have a look at the question and answer, you do no have to provide any answer here.

1. You are a male and you have 3 children: is there an action (only 1) that you could take according to the decision tree to become eligible for the loan?
[You can also change these premises].

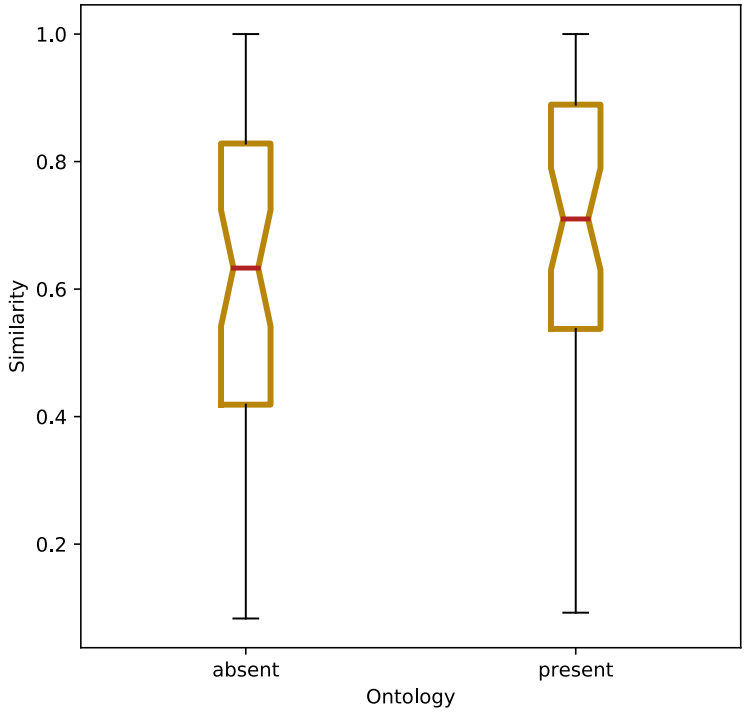
- (•) Yes
- ( ) No
- ( ) Not applicable

2. If yes, what is such an action/event?

The action/event is: become a graduate

Results Experiments 4

- Evaluation
  - **Sematic similarity** of **free text** answers
  - SpaCy with **pre-trained word embeddings**
  - Analysed for similarity with **pre-defined answers**
- **Ontology significantly** increases **correctness** and reduces **response time**



Conclusions

- **Explanation generation with Trepan Reloaded**
  - Integrate **semantic background knowledge**
  - Hypothesis: **general concepts** are **more understandable**
- **Experimental results**
  - **Human performance** and **subjective understandability improved significantly in all tasks**
  - **Hypothesis robustly confirmed**
- **Ontologies** makes **Decision Trees more effective** for human use
- **Future work**
  - Apply with **more ontologies** and different **use cases**
  - Automate **ontology selection** and **mapping**
  - **Fine tune** understandability for different domains and tree structures

Thank you!