Explainable Artificial Intelligence

Student: Nedeljko Radulović
Supervisors: Mr. Albert Bifet and Mr. Fabian Suchanek

Introduction

Research avenues

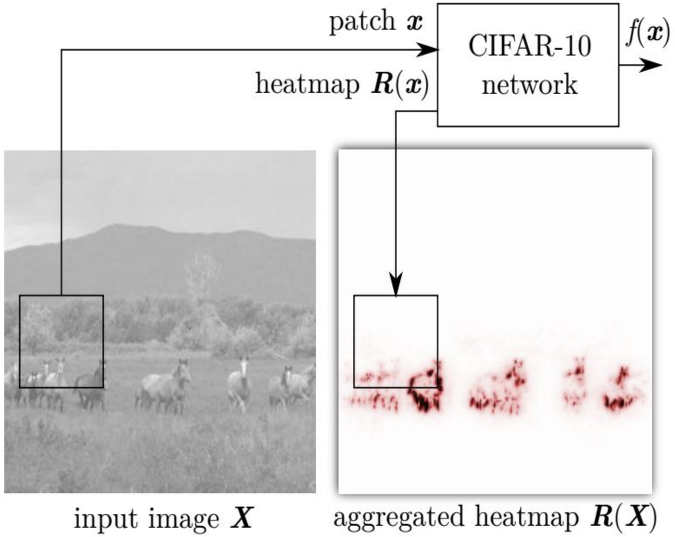
- Explainability
- Integration of first-order logic and Deep Learning
- Detecting vandalism in Knowledge Bases based on correction history

Context

- Machine Learning and Deep Learning models sometimes exceed the human performance in decision making
- Major drawback is lack of transparency and interpretability
- Bringing transparency to the ML models is a crucial step towards the Explainable Artificial Intelligence and its use in very sensitive fields

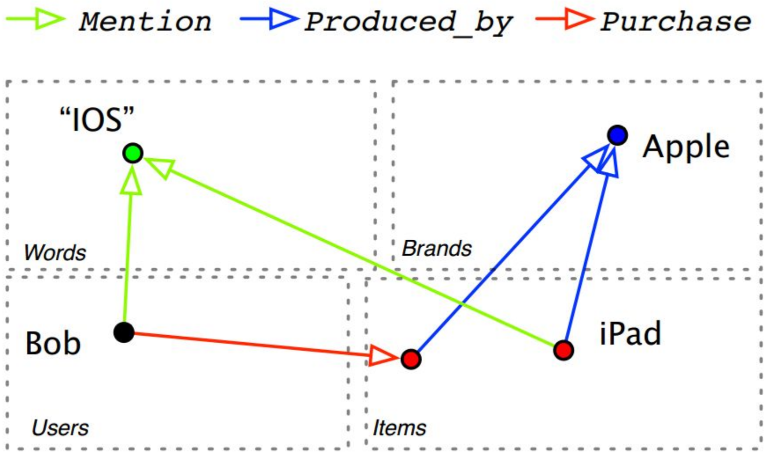
State of the art

- Exlplainable Artificial Intelligence is the topic of great interest in research in recent years
- Interpretability:
    - Using visualization techniques (mostly used in image and text classification)
- Explainability:
    - Computing influence from inputs to outputs
    - Approximating complex model with a simpler model locally (LIME)



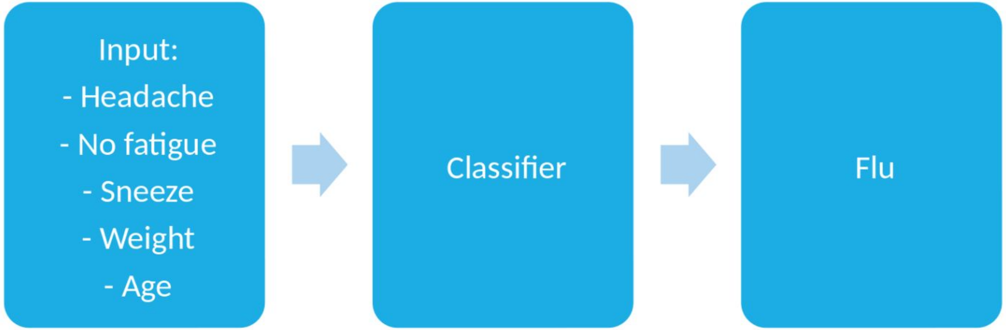
State of the art

- Attempts to combine Machine Learning and knowledge from Knowledge Bases
  - Reasoning over knowledge base embeddings to provide explainable recommendations

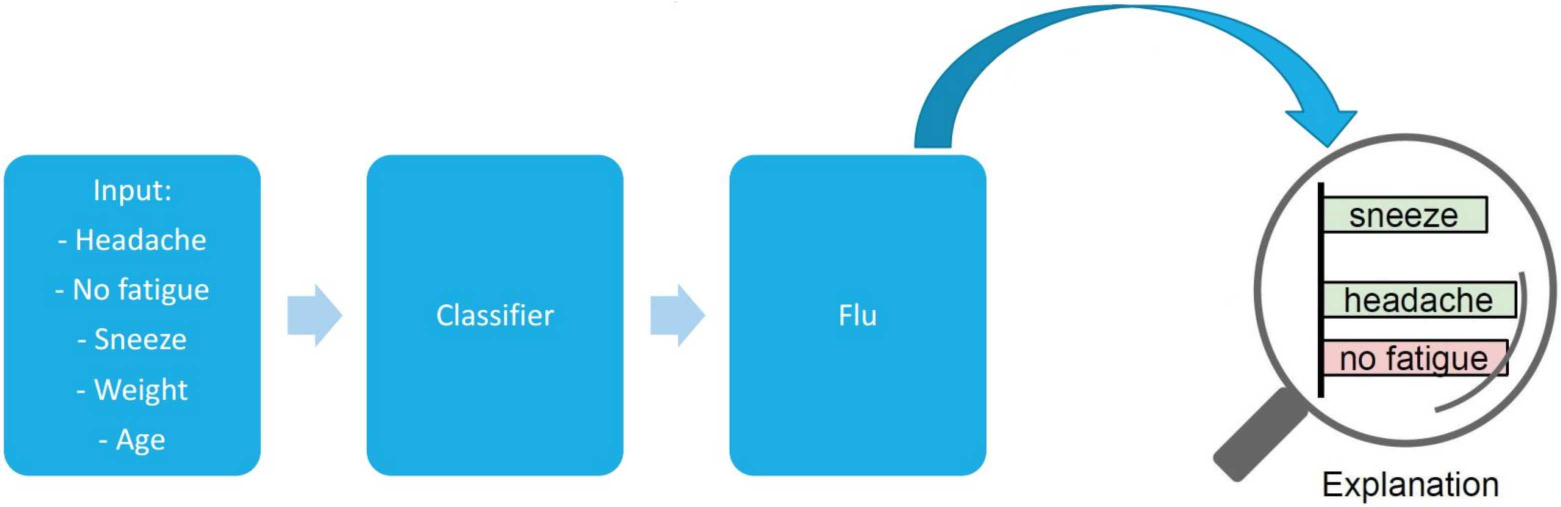


Explainability

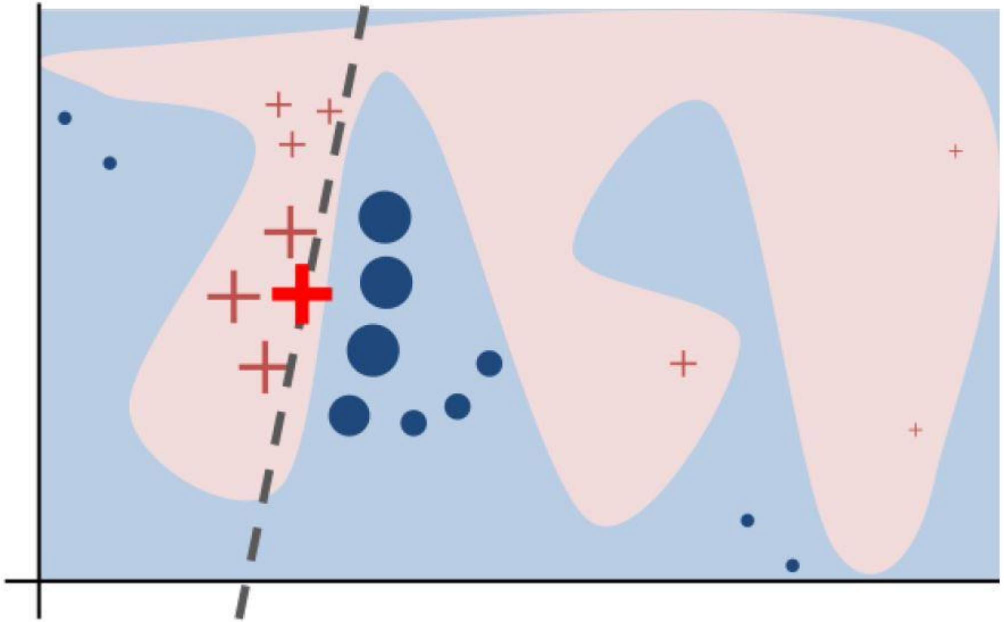
Explainability



Explainability



LIME[1] - Explaining the predictions of any classifier



1: https://arxiv.org/abs/1602.04938

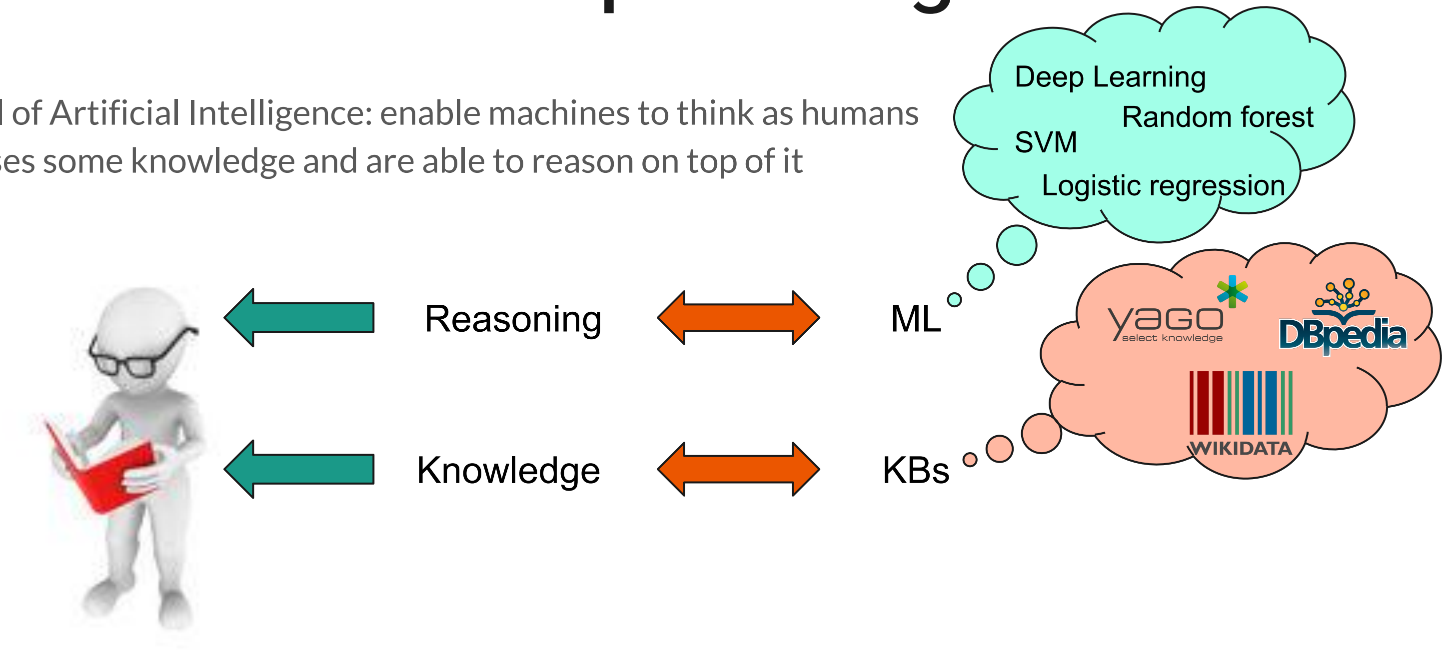
Explaining predictions in streaming setting

- Idea behind LIME is to use simple models to explain predictions
- Use already interpretable models - Decision trees
- Build Decision tree in the neighbourhood of the example
- Use the paths to leaves to generate explanations
- Use Hoeffding Adaptive Tree in streaming setting and explain how predictions evolve based on changes in the tree

Integration of First-order logic and Deep Learning

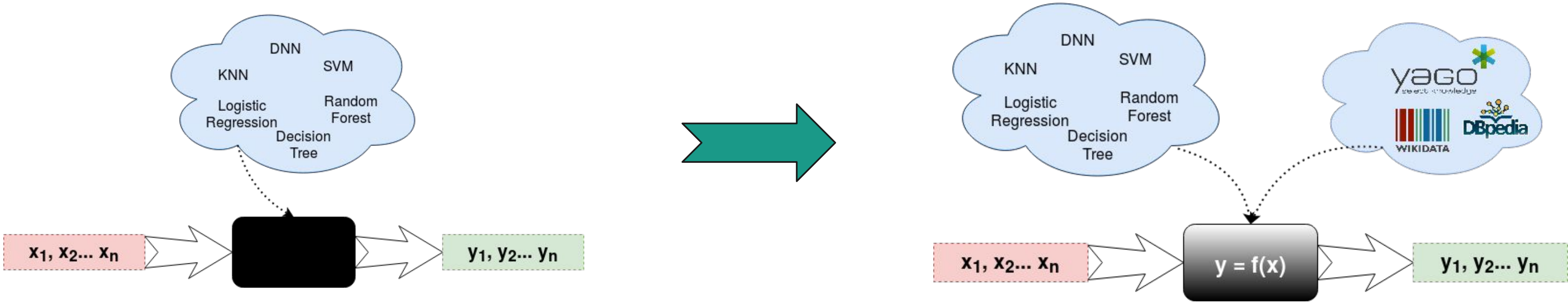
Integration of FOL and Deep Learning

- Ultimate goal of Artificial Intelligence: enable machines to think as humans
- Humans posses some knowledge and are able to reason on top of it



Integration of FOL and Deep Learning

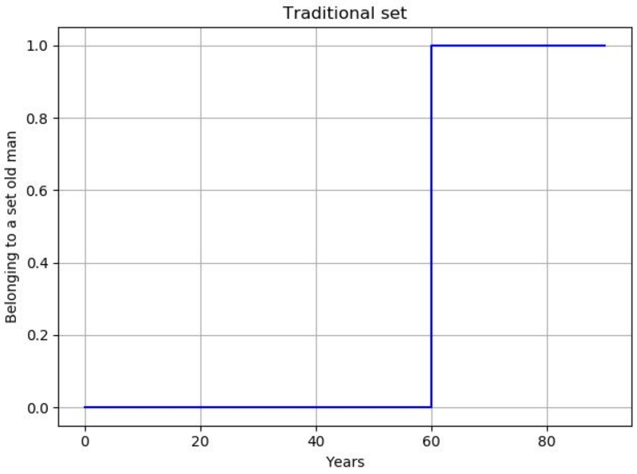
- There are several questions that we want to answer through this research:
  - How can KBs be used to inject meaning into complex and uninterpretable models, especially deep neural networks?
  - How can KBs be used more effectively as (additional) input for deep learning models?
  - How we can adjust all these improvements for streaming setting?

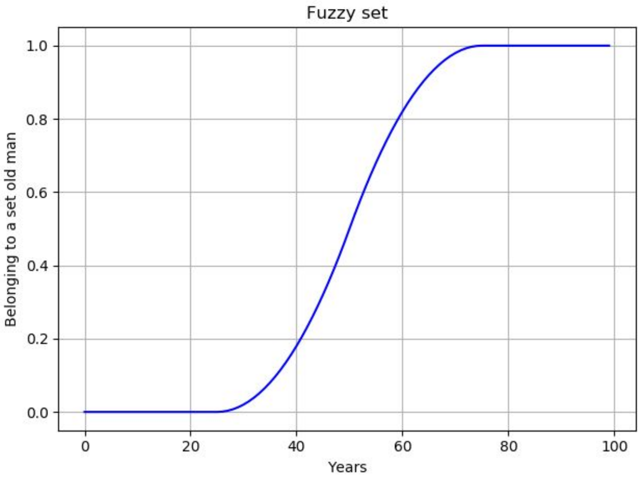


Main Idea

- Explore symbiosis of crisp knowledge in Knowledge Bases and sub-symbolic knowledge in Deep Neural Networks
- Approaches that combined crisp logic and soft reasoning:
  - Fuzzy logic
  - Markov logic
  - Probabilistic soft logic

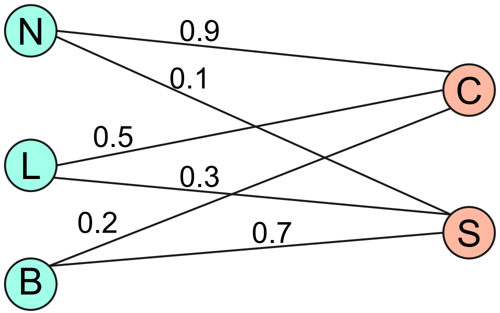
Fuzzy logic - Fuzzy set





Fuzzy logic - Fuzzy relation and Fuzzy graph

close to	Chicago	Sydney
New York	0.9	0.1
London	0.5	0.3
Beijing	0.2	0.7



Markov Logic and Probabilistic Soft Logic

- First-order logic as template language
- Example:
  - Predicates: *friend, spouse, votesFor*
  - Rules:

$$friend(Bob, Ann) \wedge votesFor(Ann, P) \rightarrow votesFor(Bob, P)$$

$$spouse(Bob, Ann) \wedge votesFor(Ann, P) \rightarrow votesFor(Bob, P)$$

Markov Logic

- Add weights to first-order logic rules:

$$friend(Bob, Ann) \wedge votesFor(Ann,P) \rightarrow votesFor(Bob, P) : [3]$$

$$spouse(Bob, Ann) \wedge votesFor(Ann,P) \rightarrow votesFor(Bob, P) : [8]$$

- **Interpretation:** Every atom (*friend(Bob, Ann), votesFor(Ann,P), votesFor(Bob, P), spouse(Bob, Ann)*) is considered as random variable which can be: *True* or *False*
- To calculate probability of an interpretation:

$$P(I) = \frac{exp(\sum_{r \in I} weight)}{\sum_{all\ I} exp(\sum_{r \in I} weight)}$$

Probabilistic Soft Logic

- Add weights to first-order logic rules:

$$friend(Bob, Ann) \wedge votesFor(Ann,P) \rightarrow votesFor(Bob, P) : [3]$$

$$spouse(Bob, Ann) \wedge votesFor(Ann,P) \rightarrow votesFor(Bob, P) : [8]$$

- **Interpretation:** Every atom (*friend(Bob, Ann), votesFor(Ann,P), votesFor(Bob, P), spouse(Bob, Ann)*) is mapped to soft truth values in range [0, 1]
- For every rule we compute distance to satisfaction:

$$d_r(I) = max\{0, I(r_{body}) - I(r_{head})\}$$

- Probability density function over *I*:

$$f(I) = \frac{1}{Z} exp[- \sum_{r \in R} weight\ (d_r(I))],\ Z = \int_I exp[- \sum_{r \in R} weight\ (d_r(I))]$$

Detecting vandalism in Knowledge bases based on correction history

Detecting vandalism in KBs based on correction history

- Collaboration with Thomas Pellissier Tanon
- Based on a paper: “Learning How to Correct a Knowledge Base from Edit History”
- Wikidata project
- Wikidata is a collaborative KB with more than 18000 active contributors
- Huge edit history: over 700 millions edits
- Method uses previous users corrections to infer possible new ones

Detecting vandalism in KBs based on correction history

- Prospective work in this project:
  - Release history querying system for external use
  - Try to use external knowledge (Wikipedia articles) to learn to fix more constraints violations
  - Use Machine Learning to suggest new updates
  - Use data stream mining techniques

Thank you!

Questions, ideas… ?

Research avenues

- Explainability
- Integration of first-order logic and Deep Learning
- Detecting vandalism in Knowledge Bases based on correction history