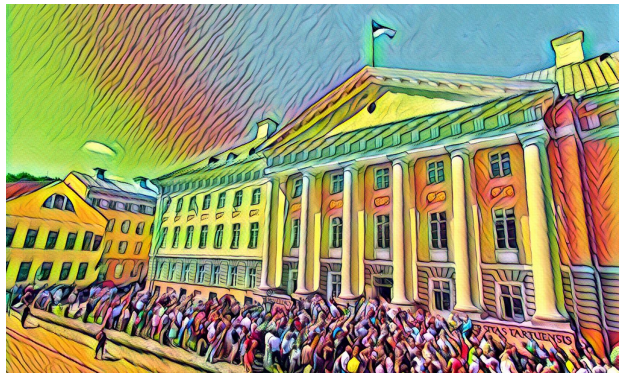
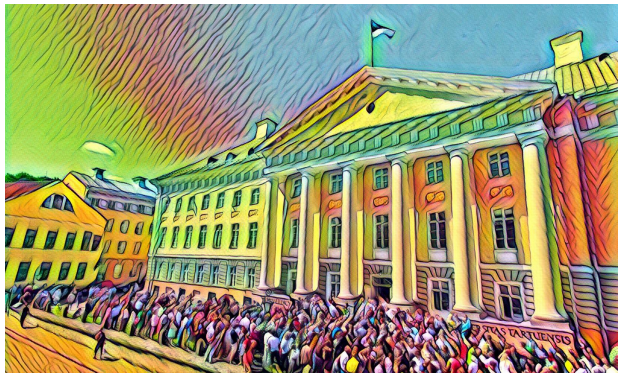




Understanding the decisions of an AI system

Meelis Kull

Institute of Computer Science

University of Tartu, Estonia

Feb 8, 2022 @ Data Science Seminar, Tartu #unitartucs

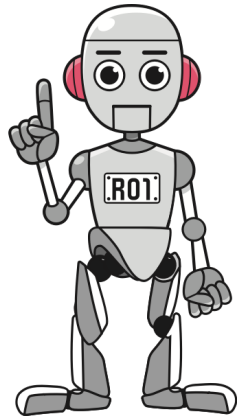
Artificial intelligence

Tool

Evolution of artificial intelligence

Tool

Co-worker



Why understand the AI?

- Why do we need to understand the decisions of an AI system?
- Why do we need to understand our co-workers?
- Errors!
  - To avoid errors
  - To mitigate the effect of errors
  - To analyze the causes of errors retrospectively
- Surprisingly good performance is also worth understanding!
- Example: a bank decides whether to give out a loan
  - AI makes preliminary decisions
  - Human overrides some of those decisions
  - Or human decides when to seek further information

Who needs to understand the AI?

- People who are affected by the decisions
  - Laws and regulations, e.g. GDPR
- Auditors who check that the AI system follows the laws
  - E.g. fairness and no discrimination
- Decision-makers (business leaders, system operators, etc.)
  - Need to understand when and how much to trust the AI system
- AI developers
  - Need to understand what kind of errors the AI makes and why
- Many other co-workers of AI
  - Highlight shortcomings of AI, learn from AI, …
- …

Application domains

- Transportation:
  - E.g. autonomous vehicles
- Healthcare:
  - E.g., diagnostics, radiology
- Finance:
  - E.g., loan decisions, fraud detections, money laundering
- Military:
  - E.g., imagery intelligence
- Retail:
  - E.g., pricing decisions, recommendations
- Social media:
  - E.g., detection of fake news, recommendations
- …

Outline

- Why do we need to understand AI?
- **Main concepts in XAI**
- Some general methods of XAI:
  - Counterfactuals
  - Surrogate models
  - Shapley values
- Explaining deep neural networks

XAI – explainable artificial intelligence

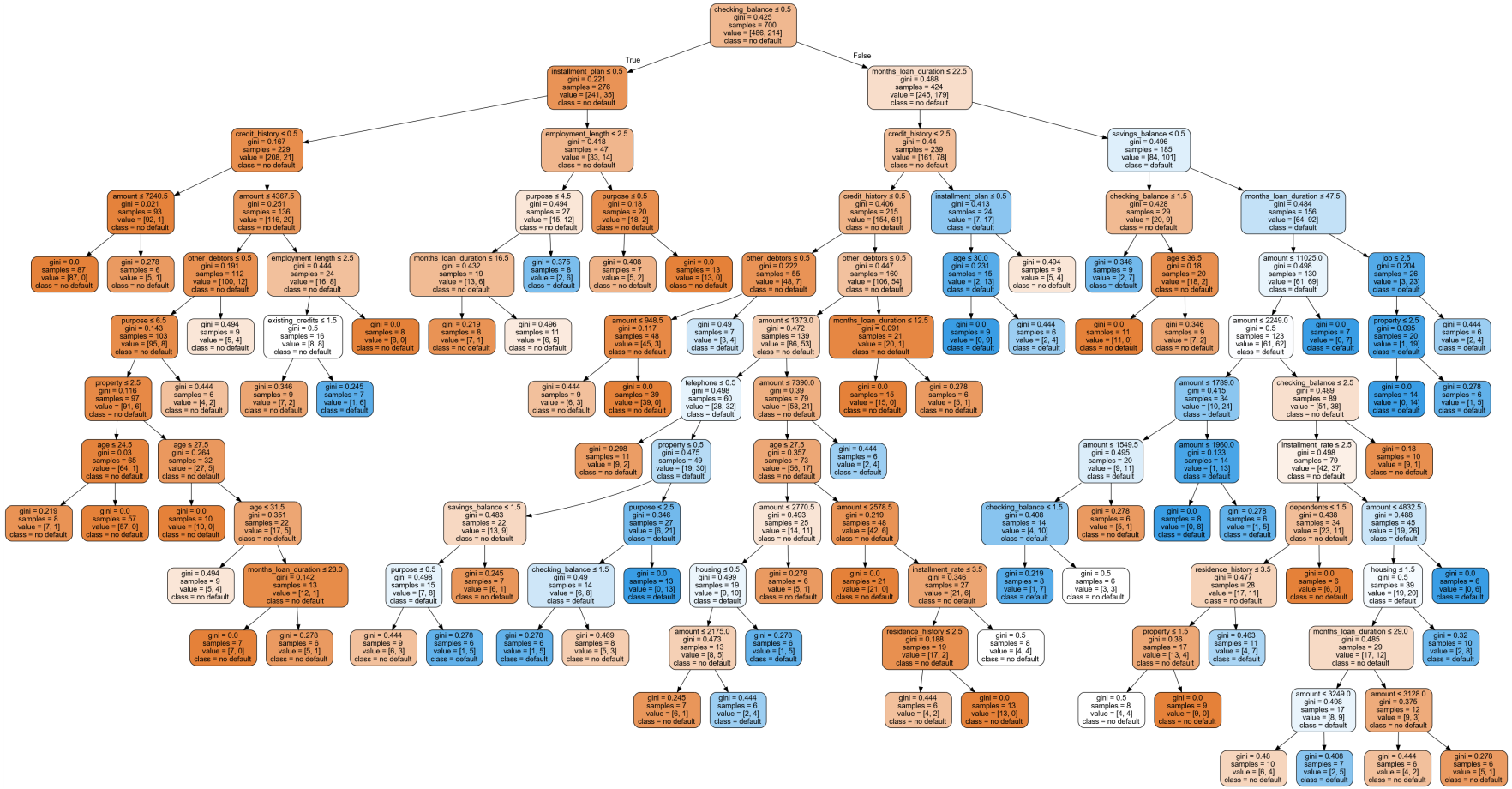
- Two common terms:
  - Interpretable AI (or interpretable machine learning)
  - Explainable AI (= XAI)
- Interpretable AI:
  - The AI system is such that we can understand (or interpret) the way it works and makes decisions
- Explainable AI:
  - There is an interface implemented between humans and the AI system which supports the process of explaining the AI system to the humans

Why is explaining and understanding hard?

- It is hard for humans to understand other humans!
- Even harder to understand someone with a different cultural background and different skillset
- Tradeoff:
  - Better understandable explanation

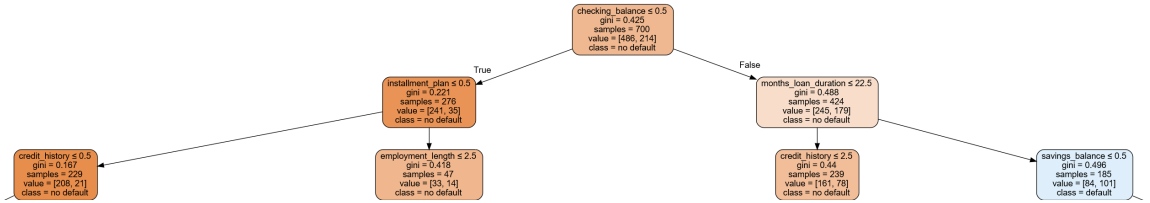
    vs
  - More accurate explanation

A model to predict personal credit risk



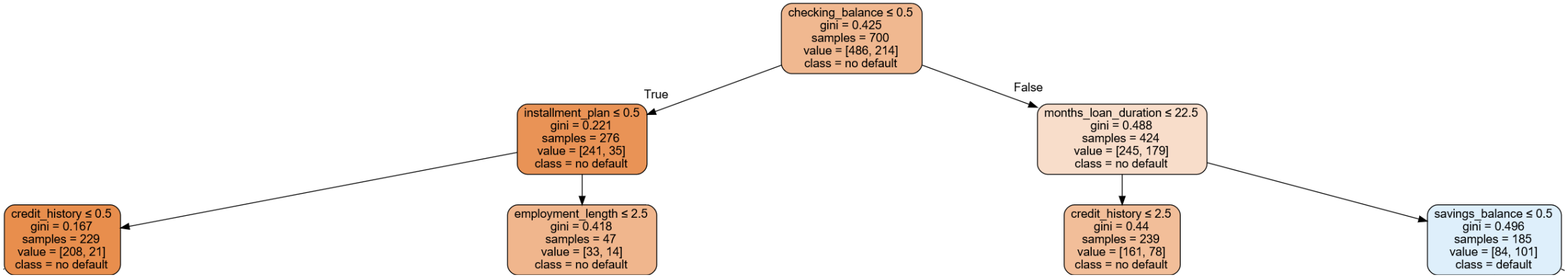
https://www.fatalerrors.org/a/using-decision-tree-to-evaluate-personal-credit-risk.html

A model to predict personal credit risk



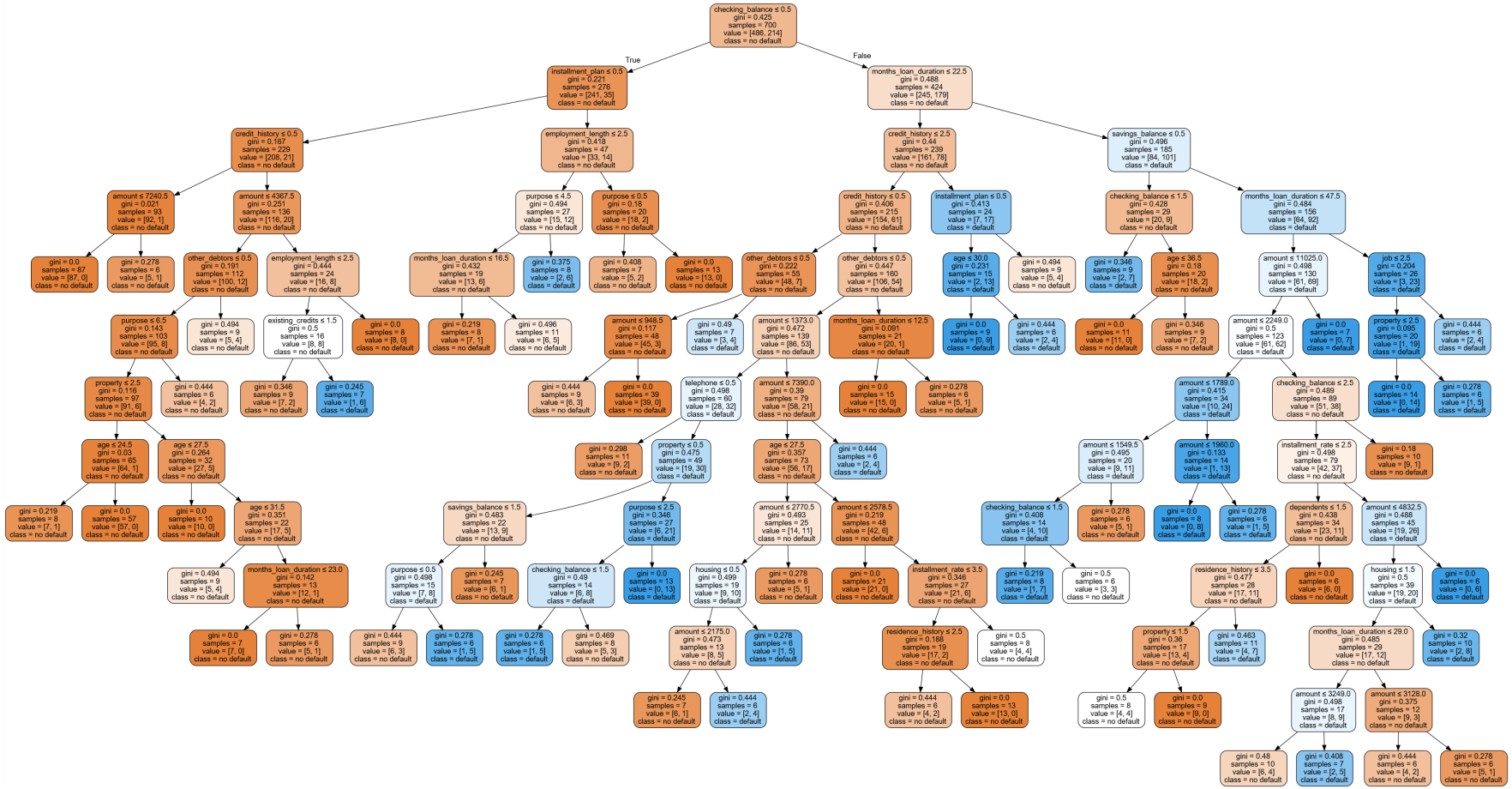
https://www.fatalerrors.org/a/using-decision-tree-to-evaluate-personal-credit-risk.html

A model to predict personal credit risk



https://www.fatalerrors.org/a/using-decision-tree-to-evaluate-personal-credit-risk.html

A model to predict personal credit risk



https://www.fatalerrors.org/a/using-decision-tree-to-evaluate-personal-credit-risk.html

Challenges of explaining AI

- Decision trees can grow big
- Random forests and gradient boosting can contain hundreds of trees
- Decision processes of neural networks can include thousands / millions / billions of calculation operations

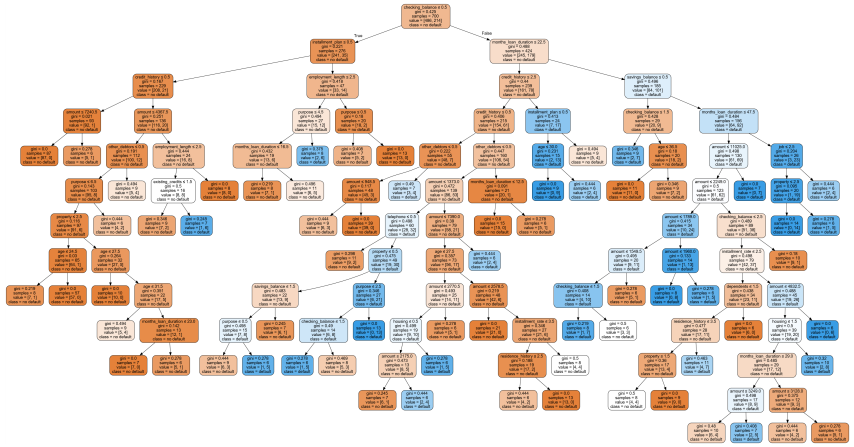
Four principles of XAI

- Explanation:
  - AI system must supply evidence, support; or reasoning for each decision made by the system.
- Meaningful:
  - the explanation must be understandable by, and meaningful to, its users
- Accuracy:
  - the explanation must reflect accurately the system's processes.
- Knowledge limits:
  - AI systems must identify cases that they were not designed to operate in and, therefore, their answers may not be reliable.

Phillips, P. J., Hahn, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2020) Four principles of explainable artificial intelligence. NIST, Gaithersburg, Maryland. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=933399

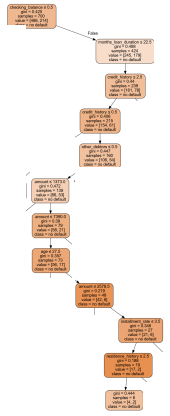
Global vs local explanations

- Global explanation:
  - How does the AI system make decisions?
  - What information does it use or not use?
  - What bits of information contribute more to the decisions?



Global vs local explanations

- **Local** explanation:
  - How did the AI system make the decision on **this specific instance**?
  - What information **did** it use or not use?
  - What bits of information **turned out to contribute more to the decision**?



Properties of XAI

- Transparent vs opaque model
- Model-agnostic vs model-specific XAI method
- Global vs local explanations

Outline

- Why do we need to understand AI?
- Main concepts in XAI
- **Some general methods of XAI:**
  - Counterfactuals
  - Surrogate models
  - Shapley values
- Explaining deep neural networks

Local model-agnostic approaches of XAI

- Counterfactuals
  - What feature value should be changed for the decision to change?
  - E.g. If the person would earn 10% more per year, the loan would be given
- Feature importances
  - Which features contributed to the decision?
  - How important were these contributions?
- Surrogate models
  - A new interpretable model fitted locally near this instance
  - Can provide feature importances, but potentially more as well

Local Interpretable Model-agnostic Explanations (LIME)

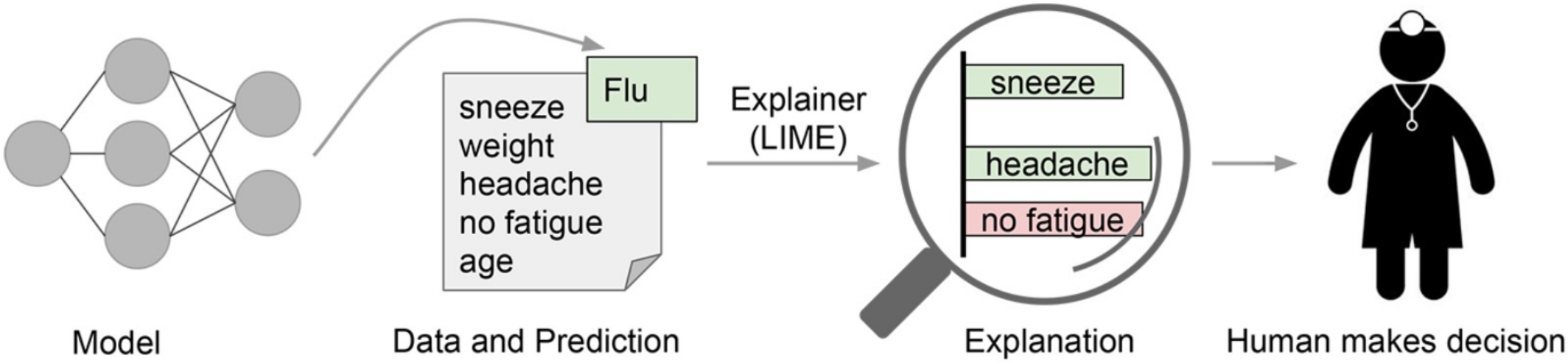


Figure 1. Explaining individual predictions to a human decision-maker. Source: Marco Tulio Ribeiro.

Ribeiro, M.T., Singh, S. and Guestrin, C
"Why should I trust you?" Explaining the predictions of any classifier.
KDD 2016

Shapley values

- Originally introduced in 1951 in cooperative game theory by Lloyd Shapley, later awarded with the Nobel Prize in Economics in 2012
- Suppose K partners are contributing to an enterprise, how should they divide the profits?
- Example:
  - 2 partners
  - Partner A alone would create the profit of 1M Euros
  - Partner B alone would create the profit of 2M Euros
  - Two partners together create the profit of 4M Euros
  - How much should each partner get out of this 4M Euros?

Shapley values

- Example:
  - f({1})=1
  - f({2})=2
  - f({1,2})=4
- General formula for payoffs according to Shapley values:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

- Uniquely derived from the axioms:
  - Efficiency, Symmetry, Linearity, Null player.

Shapley values

- Example:
  - f({1})=1
  - f({2})=2
  - f({1,2})=4
- General formula for payoffs according to Shapley values:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

- Answer:
  - The profit of 4M should be shared as 1.5M and 2.5M EUR.

Shapley values for machine learning

- Features are the partners of the cooperation
- Decision is the profit
- Shapley values show how much of the decision should be attributed to each feature
- This method unifies many earlier AI interpretation methods

Lundberg & Lee. "A unified approach to interpreting model predictions." *NeurIPS 2017*

Outline

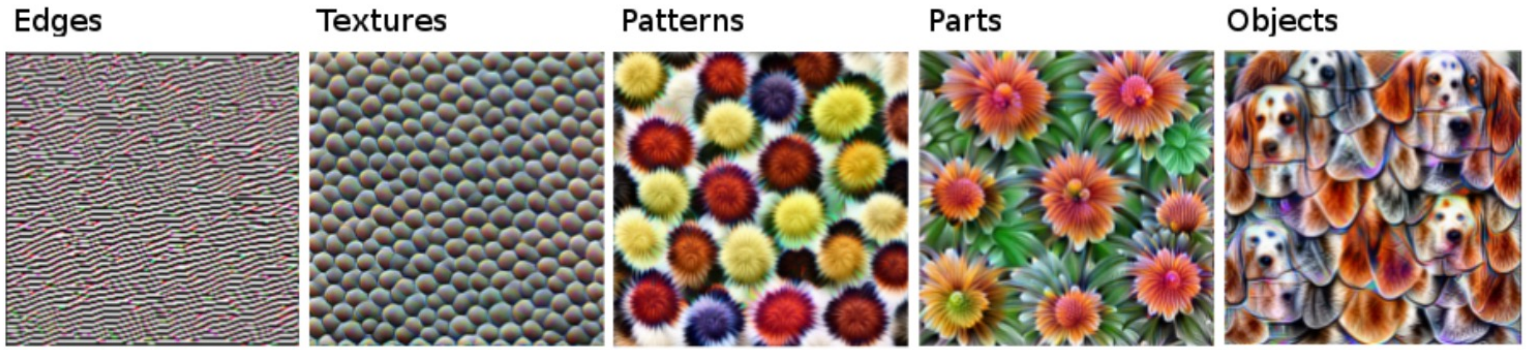
- Why do we need to understand AI?
- Main concepts in XAI
- Some general methods of XAI:
  - Counterfactuals
  - Surrogate models
  - Shapley values
- **Explaining deep neural networks**

Interpreting deep nets in computer vision

- Problems with many earlier methods :
  - All features (=pixels) contribute
  - No single feature (=pixel) is enough to change the decision

Global interpretations of neural nets

- What do the units in different layers represent within the neural net?
- Example units from 5 layers of a CNN:

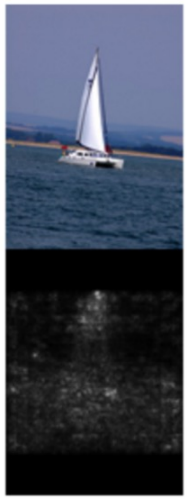
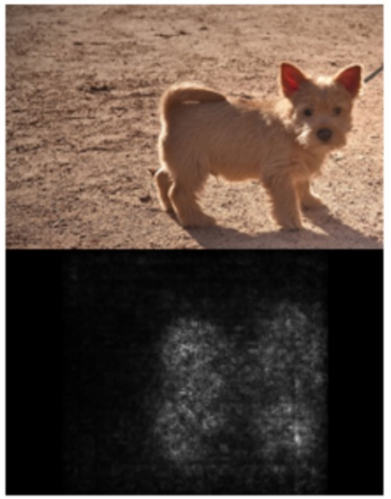


- Another unit:

Local explanations: Vanilla Gradient (Saliency Maps)

- If we do a tiny change in one pixel, how much does it affect the predicted probability for the particular class?
- Can be efficiently calculated using gradients



Simonyan, K., Vedaldi, A. and Zisserman, A., 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*. ICLR 2014 workshop

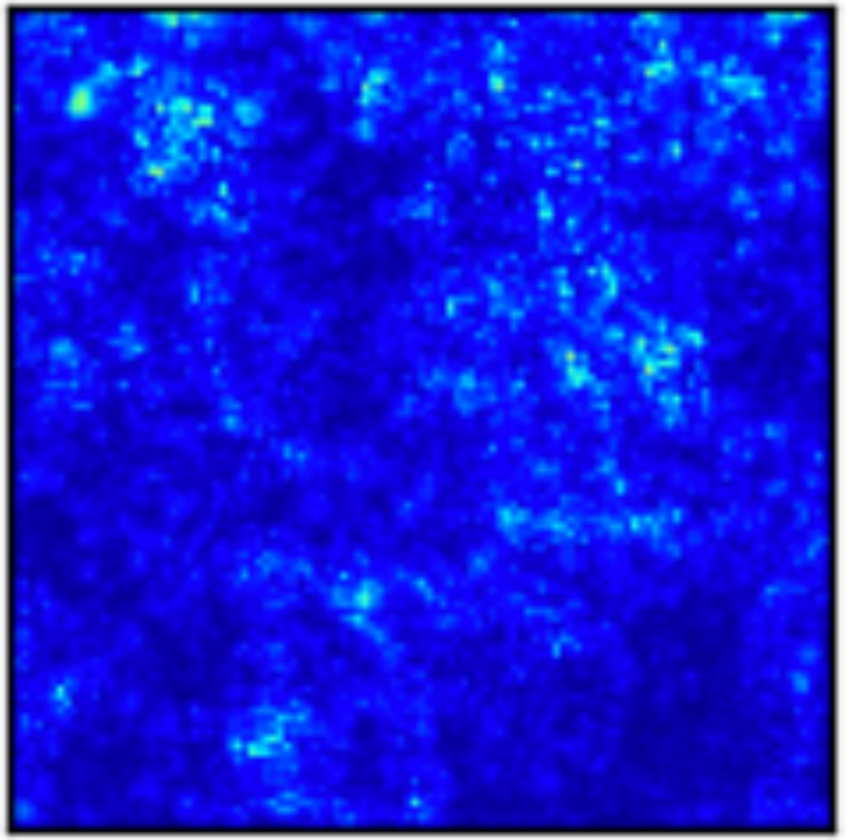
Octopus misclassified as an eel

Original image:

Octopus misclassified as an eel

Original image:

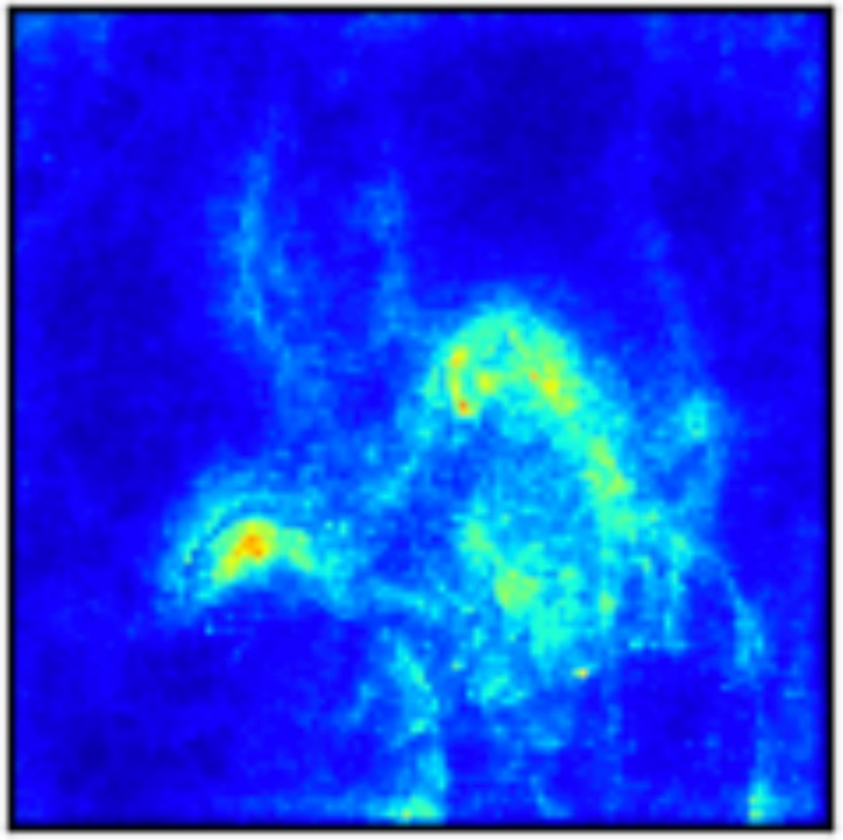
Vanilla gradient:



Octopus misclassified as an eel

Original image:

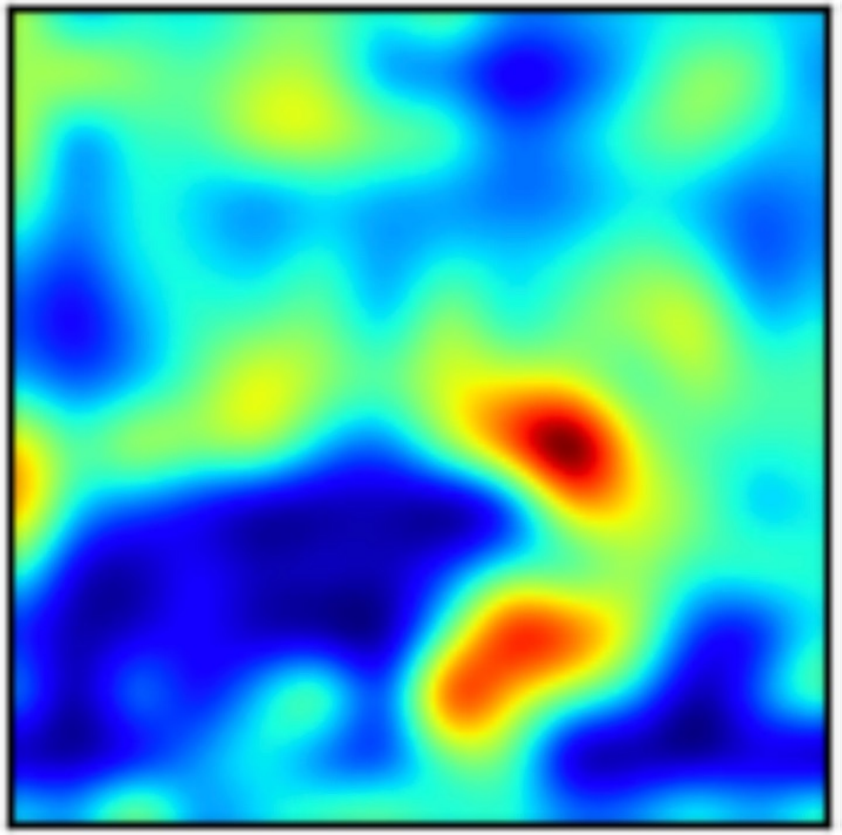
Smoothgrad:



Octopus misclassified as an eel

Original image:

Grad-CAM



Outline

- Why do we need to understand AI?
- Main concepts in XAI
- Some general methods of XAI:
  - Counterfactuals
  - Surrogate models
  - Shapley values
- Explaining deep neural networks

Questions?

Outline

- Why do we need to understand AI?
- Main concepts in XAI
- Some general methods of XAI:
  - Counterfactuals
  - Surrogate models
  - Shapley values
- Explaining deep neural networks