Our Goal in this AK: design, develop & test a System Causability Scale

**TU. Important Definition:**
- Causability := a property of a person (Human)
- Explainability := a property of a system (Computer)

**TU. First Homework for you:** make yourself familiar ...
- [https://www.tensorflow.org/tutorials](https://www.tensorflow.org/tutorials)

**TU. Our Playground**

**TU. System Usability Scale**

**TU. The challenge**

Why did the algorithm do that?  
Can I trust these results?  
How can I correct an error?  

A possible solution  

- The domain expert can understand why ...
- The domain expert can learn and correct errors ...
- The domain expert can re-enact on demand ...

**TU. Usability**

Usage Indicators  

Means  

Knowledge  

**TU. Science is to test crazy ideas – Engineering is to put these ideas into Business**

LV 706.046 Summer Term 2019  
Monday, March, 18  
From measuring Usability to measuring Causability: Methods of Explainable AI  
Assoc.Prof. Dr. Andreas Holdinger  
Holdinger-Group, HCI-KDD, Institute for Medical Informatics/Statistics  
Medical University Graz, Austria  
Institute of Intensive Systems & Data Science  
Graz University of Technology, Austria  

a.holdinger@tugraz.at  
https://hci-kdd.org/intelligent-user-interfaces-2019

**References:**
Methods of Explainable AI

1. Gradients
2. Sensitivity Analysis
4. Optimization (Local-IME - model agnostic, BETA transparent approximation, ...)
5. Deconvolution and Guided Backpropagation
6. Model Understanding
   - Feature visualization, Inverting CNN
   - Qualitative Testing with Concept Activation Vectors TCAV
   - Network Dissection

TU Methods of ex-AI

TU Important Definition: Ground truth

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
- Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
- Empirical inference = drawing conclusions from empirical data (observations, measurements)
- Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
- Causal inference is an example of causal reasoning.

TU Empirical Inference Example

\[ y = \sum a_i \cdot f_i(x) + b \]
\[ y = a \cdot x \]

Gottfried W. Leibniz (1646-1716)
Hermann Wey (1885-1955)
Vladimir Vapnik (1936)
Alexey Cheremnychev (1938-2014)
Gregory Chaitin (1947)

TU Remember: Reasoning = "Sensemaking"

- Deductive Reasoning = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct, or else the conclusion cannot be determined.
  - Based on the truth of premises: A \(\land\) B, B \(\land\) C, therefore A \(\lor\) C
- Inductive reasoning = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true or false
  - generates hypotheses and uses DR for answering specific questions
- Abductive reasoning = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

TU Gradients

TU Gradients

Remember: hard inference problems

- High dimensional [curse of dim., many factors contribute]
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - * = Def.: a sequence or collection of random variables is independent if and only if each random variable is independent of all others.

}

https://repository.tu-berlin.de/handle/1813/19383
**TU Gradients**

![Image of gradient diagrams]

**TU LRP Layer-Wise Relevance Propagation**

![Image of LRP diagrams]

**TU A NN-classifier during prediction time**

![Diagram showing a neural network classifier]

**TU**

**Definition 1.** A heatmap $R(x)$ is conservative if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

$$\forall x : f(x) = \sum R_x(x).$$

**Definition 2.** A heatmap $R(x)$ is positive if all values forming the heatmap are greater or equal to zero, that is:

$$\forall x, p : R_p(x) \geq 0.$$

**Definition 3.** A heatmap $R(x)$ is consistent if it is conservative and positive. That is, it is consistent if it complies with Definitions 1 and 2.

![Diagram showing heatmap properties]

---

Future Work

- Explainable AI with Deep Tensor and Knowledge Graph


Explanations in Artificial Intelligence will be necessary

"Does your car have any idea why my car pulled it over?"

https://www.newyorker.com/cartoon/a9697

IBM is doing it now: teaching meaningful explanations

Teaching Meaningful Explanations

- What is a good explanation?
- (obviously if the other did understand it)
- Experiments needed!
- What is explainable/understandable/intelligible?
- When is it enough (Sättigungsgrad – you don’t need more explanations – enough is enough)
- But how much is it ...

Seemingly trivial questions ...?