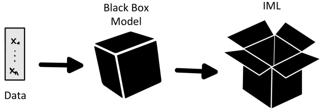
Interpretable Machine Learning

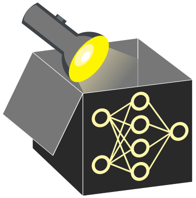
Interpretation Goals



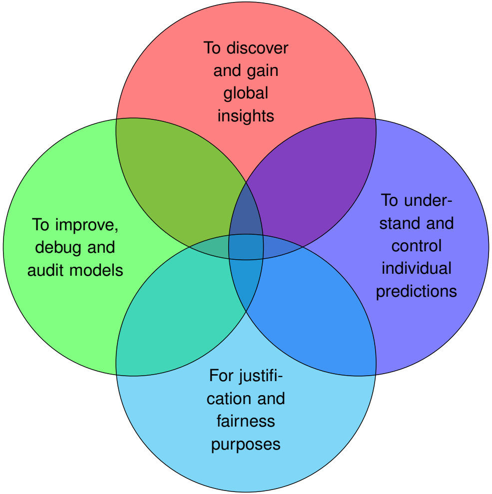
Learning goals

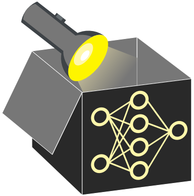
Understand Interpretation Goals:

- Global insights (discovery)
- Improve model (debug and audit)
- Understand and control individual predictions
- Justification and fairness



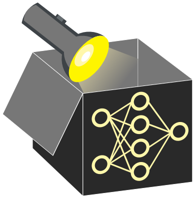
POTENTIAL INTERPRETATION GOALS





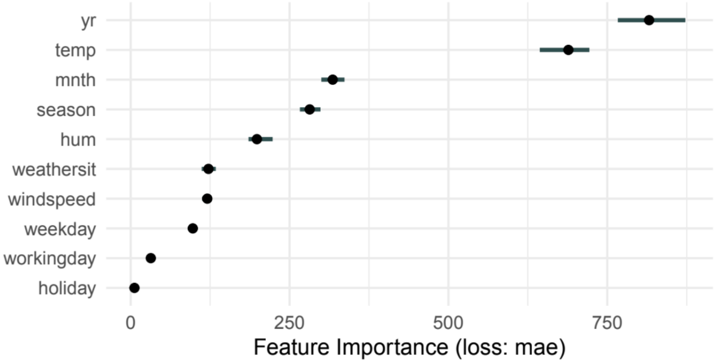
A related presentation can be found in ▸ Adadi and Berrada 2018.

DISCOVER AND GAIN GLOBAL INSIGHTS



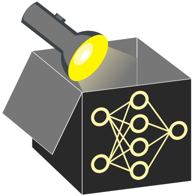
$\rightsquigarrow$ Gain insights about data, model, and underlying data-generating process

Example: Bike Sharing Dataset (predict number of bike rentals per day)
Exemplary question: Which feature influences model performance and how much?



- Year (`yr`) and Temperature (temp) most important features
- Holiday (`holiday`) less important (Can we drop it?)

IMPROVE, DEBUG AND AUDIT MODELS



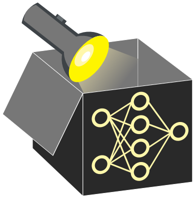
$\rightsquigarrow$ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure

IMPROVE, DEBUG AND AUDIT MODELS



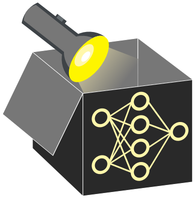
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels

IMPROVE, DEBUG AND AUDIT MODELS



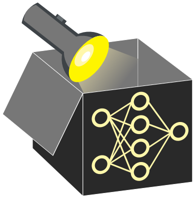
⇝ Insights help to identify flaws (in data or model), which can be corrected

Example: Neural Net Tank ▸ gwern.net

Cautionary tale (never actually happened):

- Train a neural network to detect tanks
- Good fit on training data
- Application outside training data: failure
- Reasons vary depending on input
  ⇝ NN based decision on irrelevant pixels
- E.g. model detects weather based on sky:
  ⇝ All photos with tanks show cloudy sky
  ⇝ Photos without tanks show sunny sky

IMPROVE, DEBUG AND AUDIT MODELS



⇝ Insights help to identify flaws (in data or model), which can be corrected

Comment on tank example:

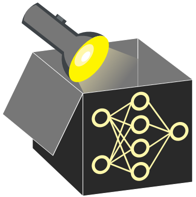
"We made exactly the same mistake in one of my projects on insect recognition. We photographed 54 classes of insects. Specimens had been collected, identified, and placed in vials. Vials were placed in boxes sorted by class. I hired student workers to photograph the specimens.

Naturally they ***did this one box at a time; hence, one class at a time.*** *Photos were taken in alcohol.* ***Bubbles would form in the alcohol. Different bubbles on different days.*** *The learned classifier was surprisingly good. But a* ***saliency map revealed that it was reading the bubble patterns*** *and ignoring the specimens.*

I was so embarrassed that I had made the oldest mistake in the book (even if it was apocryphal). Unbelievable. Lesson: always randomize even if you don't know what you are controlling for!"

▸ Thomas G. Dietterich

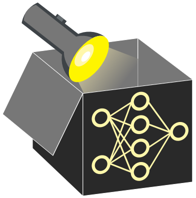
DEBUG AND AUDIT



- Nearly all computer programs have bugs
  - ⇝ Minimizing such bugs extremely relevant
- Process with multiple steps to locate, understand and solve a problem
  - ⇝ Classical debugging
- **In ML** we have a program (learner) writing another program (model)
- How to debug or audit programs which contain ML models?
- Based on a single cross-val score?

⇝ Being able to interpret your model will always be helpful – if possible!

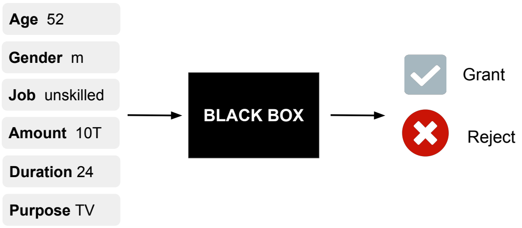
UNDERSTAND & CONTROL INDIVIDUAL PREDICTIONS



$\rightsquigarrow$ Explaining individual decisions can prevent unwanted actions based on the model

Example: Credit Risk Application

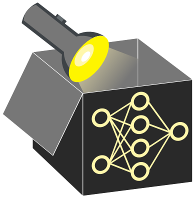
$\mathbf{x}$: customer and credit information; y: grant or reject credit



Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

UNDERSTAND & CONTROL INDIVIDUAL PREDICTIONS

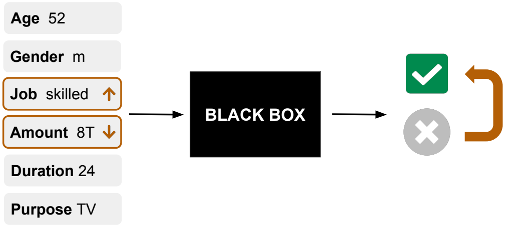


$\rightsquigarrow$ Explaining individual decisions can prevent unwanted actions based on the model

Example: Credit Risk Application

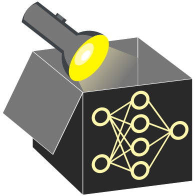
$\mathbf{x}$: customer and credit information; y: grant or reject credit

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**



“If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

JUSTIFICATION AND FAIRNESS



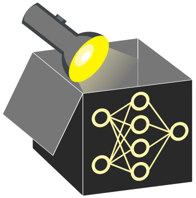
$\rightsquigarrow$ Investigate if and why biased, unexpected or discriminatory predictions were made

Example: COMPAS

- COMPAS: Correctional Offender Management Profiling for Alternative Sanctions
- Commercial tool used in courts to assess a defendant's risk of re-offending
- Predicts **recidivism risk**:
  - Likelihood of an individual with a past offense is arrested again
  - Features: race, gender, age, number of prior prison sentences, ...
  - Output: COMPAS score from 1 (low risk) to 10 (high risk) risk of recidivism
- Based on a questionnaire completed by the defendant

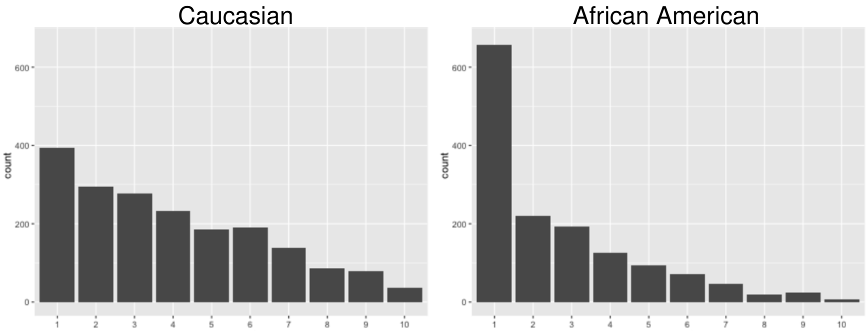
JUSTIFICATION AND FAIRNESS: COMPAS

▸ Larson et al. 2016



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

Descriptive data analysis of the target (COMPAS score) by a feature encoding race:



COMPAS score: 1 (low risk) to 10 (high risk)

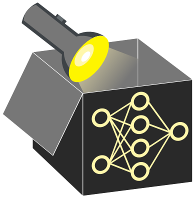
⇝ Model skewed towards low risk for Caucasians

⇝ Strong indication that the model is discriminating against African American

⇝ Use IML to investigate if and how much the model uses the defendant's race

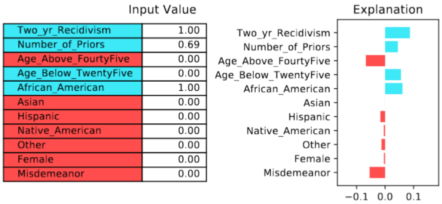
JUSTIFICATION AND FAIRNESS: COMPAS

▸ **Alvarez-Melis and Jaakkola 2018**



⇝ Investigate if and why biased, unexpected or discriminatory predictions were made

IML: Analyze how strongly a feature influences an individual prediction (e.g., LIME):



- Pick a defendant
- LIME quantifies a feature's impact on the defendant's COMPAS score
- `African_American` has a large positive weight on COMPAS score
- Occurs for many individuals, see ▸ XAI Stories ⇝ Suggests racial bias