# Interpretable Machine Learning

# Introduction, Motivation and History

**Learning goals**

- Why do we need interpretability?
- What have been the developments until now?

# WHY INTERPRETABILITY?

- Machine Learning (ML) has a huge potential to aid the decision-making process in various scientific and business applications due to its predictive power.
- ML models usually are intransparent black boxes, e.g., XGBoost, RBF SVM or DNNs.
  - $\rightsquigarrow$ too complex to be understood by humans
- The lack of explanation
  1. hurts trust
  2. creates barriers

$\rightsquigarrow$ Harder to adapt for critical areas with decisions affecting human life (e.g., medicine or credits).

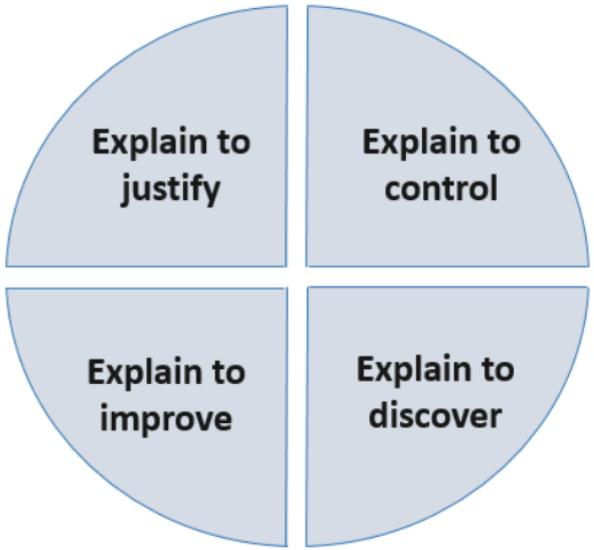
$\rightsquigarrow$ Many disciplines with required trust rely on traditional models, e.g., linear models, with less predictive performance.

# BRIEF HISTORY OF INTERPRETABILITY

- 18th and 19th century: linear regression models (Gauss, Legendre, Quetelet)
- 1940s: sensitivity analysis (SA), still used today
- Middle of 20th century: Rule-based ML, incl. decision rules and decision trees
- 2001: built-in feature importance measure of random forests
- >2010: explainable AI (XAI) for deep learning
- >2015: IML as an independent field of research
- 2018: GDPR requires explainability for some applications

# WHEN DO WE NEED INTERPRETABILITY?

- To **Justify** (and increase trust in models): investigate if and why biased, unexpected or discriminatory predictions were made.
- To **Control**: debug models, identify and correct vulnerabilities and flaws.
- To **Improve**: understanding why a prediction was made makes it easier to improve the model.
- To **Discover**: learn new facts, gather information and gain insights.



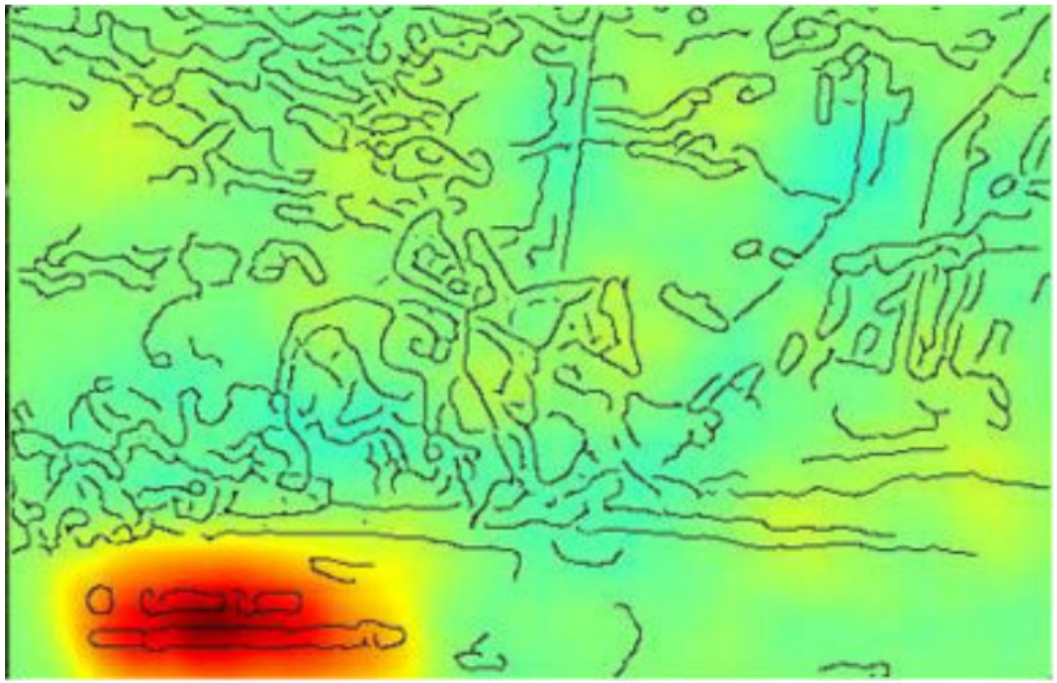
# WHY IS INTERPRETABILITY IMPORTANT?

- Machine learning is (mostly) about discovering patterns in data
- Unfortunately, it is not guaranteed that ML will identify the correct patterns
- We humans might not be able to discover patterns ML models discovered
  - That's good for science or to get new insights
  - That's bad in many practical application where unexpected behavior is not wanted
- How can you check whether the model is correct in its inference?

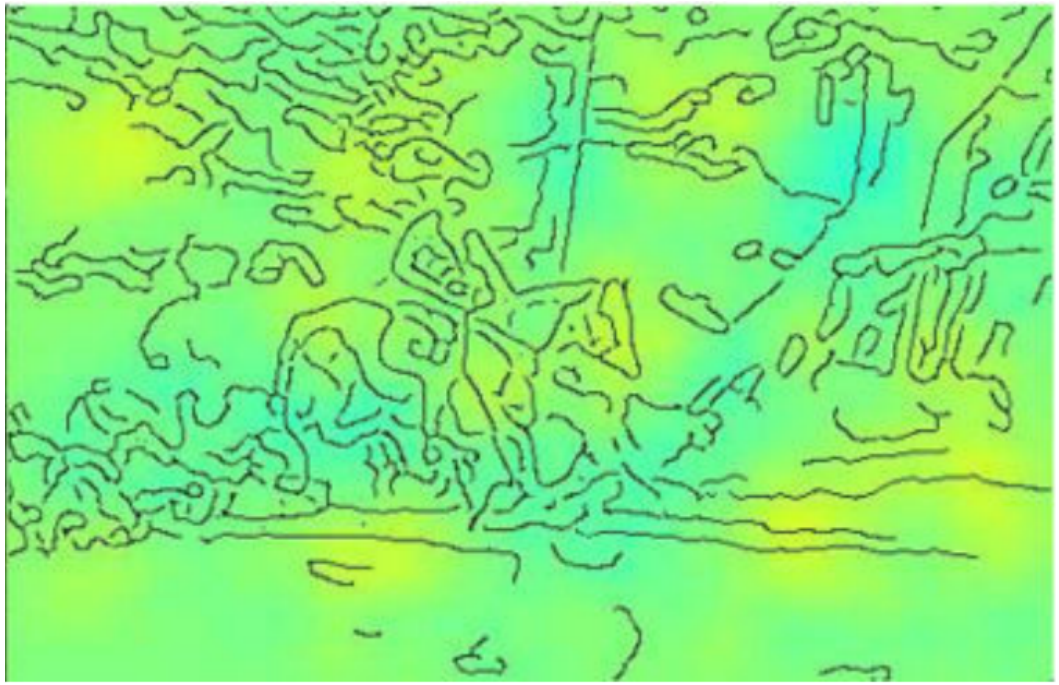
# CLEVER HANS
# HTTPS://WWW.NATURE.COM/ARTICLES/S41467-019-08987-4

# CLEVER HANS
# HTTPS://WWW.NATURE.COM/ARTICLES/S41467-019-08987-4

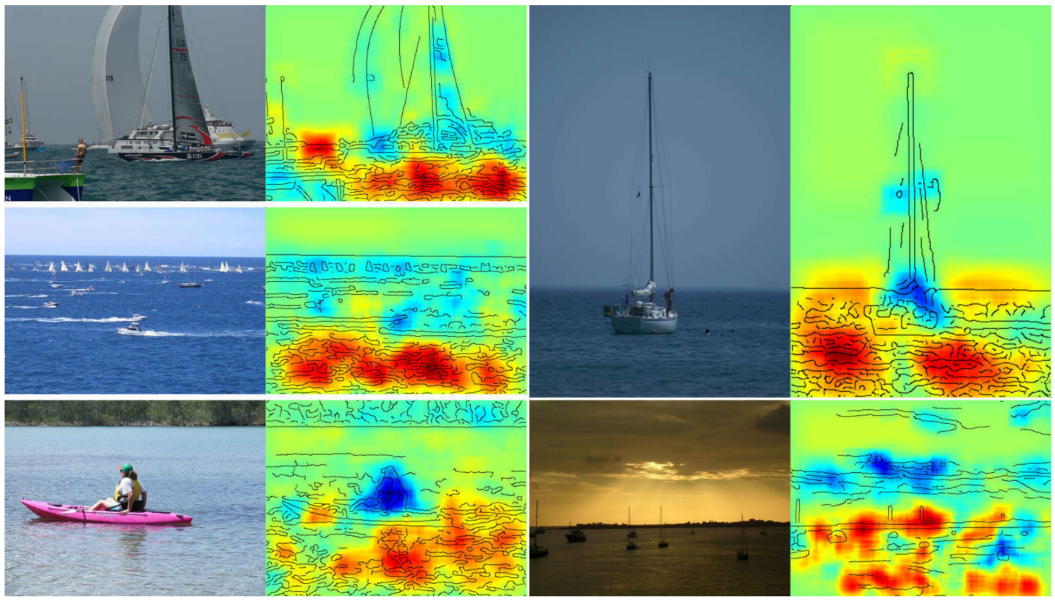
# CLEVER HANS
# HTTPS://WWW.NATURE.COM/ARTICLES/S41467-019-08987-4



# CLEVER HANS
# HTTPS://WWW.NATURE.COM/ARTICLES/S41467-019-08987-4



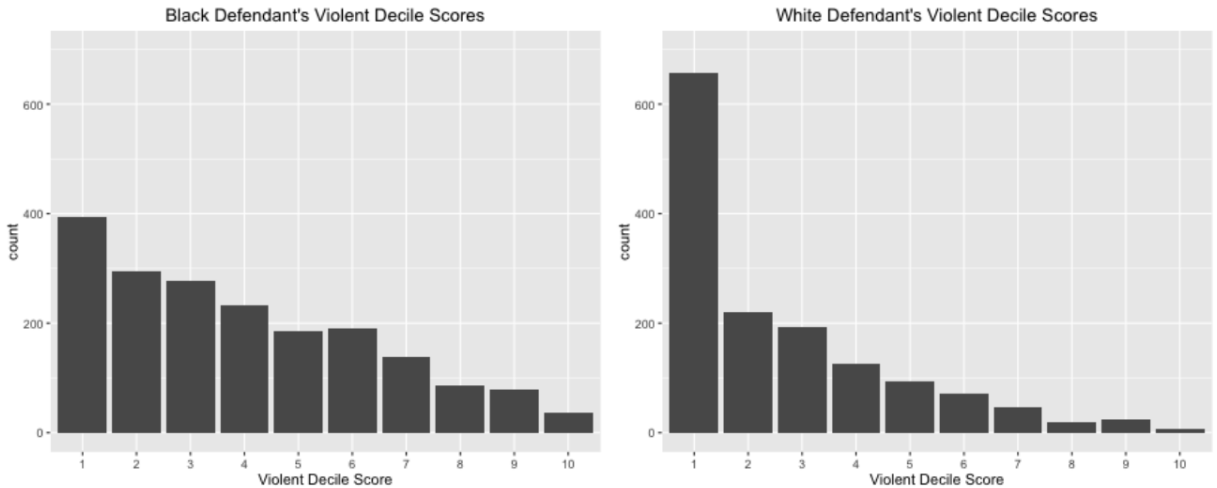
# CLEVER HANS
# HTTPS://WWW.NATURE.COM/ARTICLES/S41467-019-08987-4



# COMPASS

- Correctional Offender Management Profiling for Alternative Sanctions
- predict recidivism risk
  - i.e., criminal re-offense after previous crime, resulting in jail booking
  - different risk levels: high risk, medium risk or low risk
- evaluation based on a questionnaire the defendant has to answer

# COMPAS MODEL ANALYSIS HTTPS://WWW.PROPUBLICA.ORG/ARTICLE/HOW-WE-ANALYZED-THE-COMPAS-RECIDIVISM-ALGORITHM




$\rightsquigarrow$ Strong indication that the model is discriminating black defendants

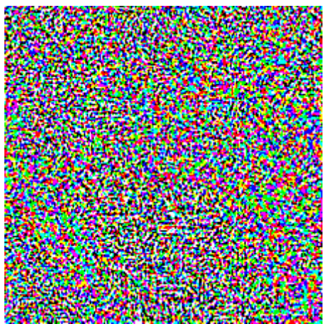
# OTHER EXAMPLES ML FAILED IN

- Credit and insurance scoring
- Medical applications
  - Identification of diseases
  - Chance of recovering
  - Recommendations of treatments
- Crime predictions
- Rating job applications
- ...

$\rightsquigarrow$ GDPR (aka DSGVO) requires that for some applications predictive models have to be explainable

# ADVERSARIAL EXAMPLES HTTPS://ARXIV.ORG/PDF/1412.6572.PDF

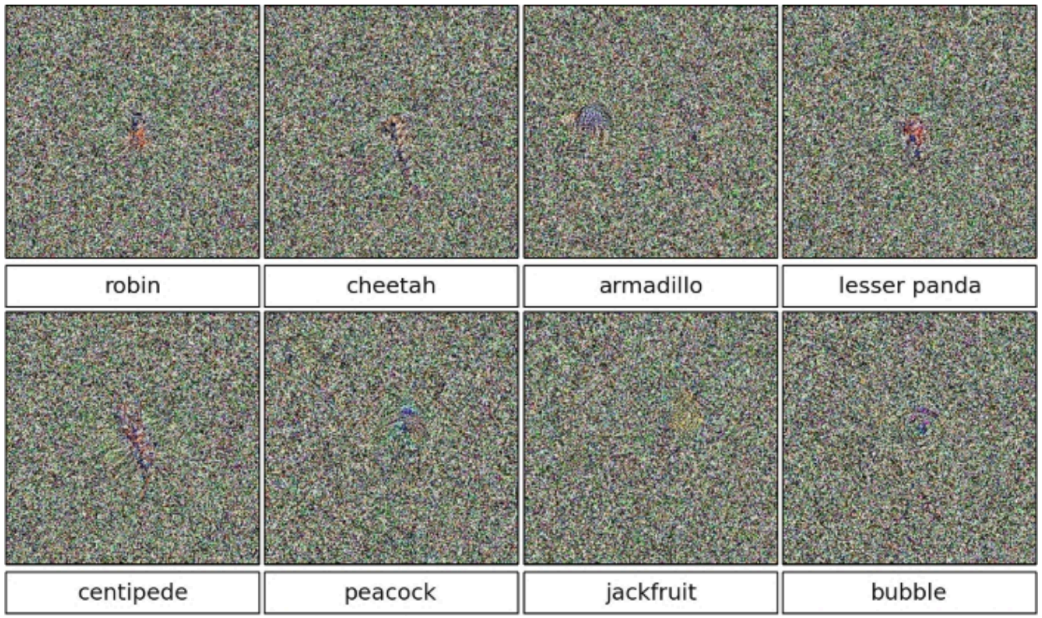
**Panda**



**Airplane**

$\rightsquigarrow$ ML Models might not capture human-like understanding.

# ADVERSARIAL NOISE HTTPS://ARXIV.ORG/PDF/1412.6572.PDF




$\rightsquigarrow$ **Adversarial Noise:** Noise that is imperceptible to **humans** but results in incorrect classification results

# ADVERSARIAL EXAMPLES HTTPS://ARXIV.ORG/PDF/1412.6572.PDF