Introduction to Explainable Artificial Intelligence (XAI)

Szymon Bobek

Jagiellonian University
2024

https://geist.re

Data became new oil

- Soon it appears, that “unrefined” it cannot really be used
- Deep Neural Networks dominance (black box models)
- Adoption of AI in sensitive and high risk areas

The world’s most valuable resource is no longer oil, but data

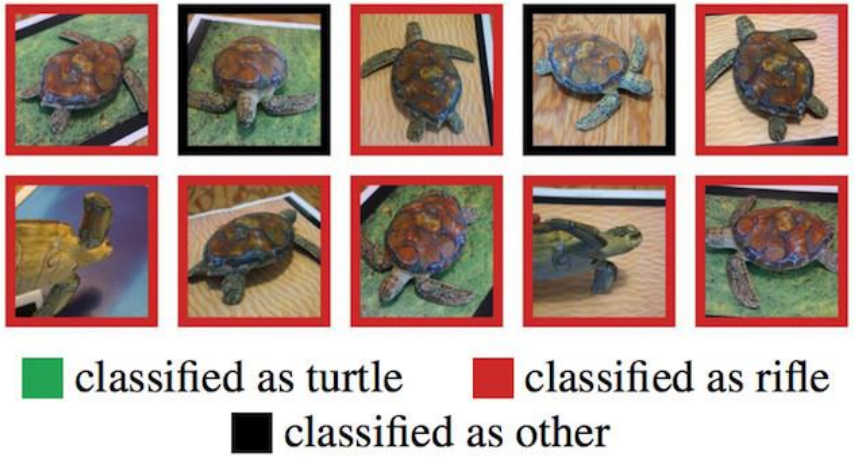
The data economy demands a new approach to antitrust rules

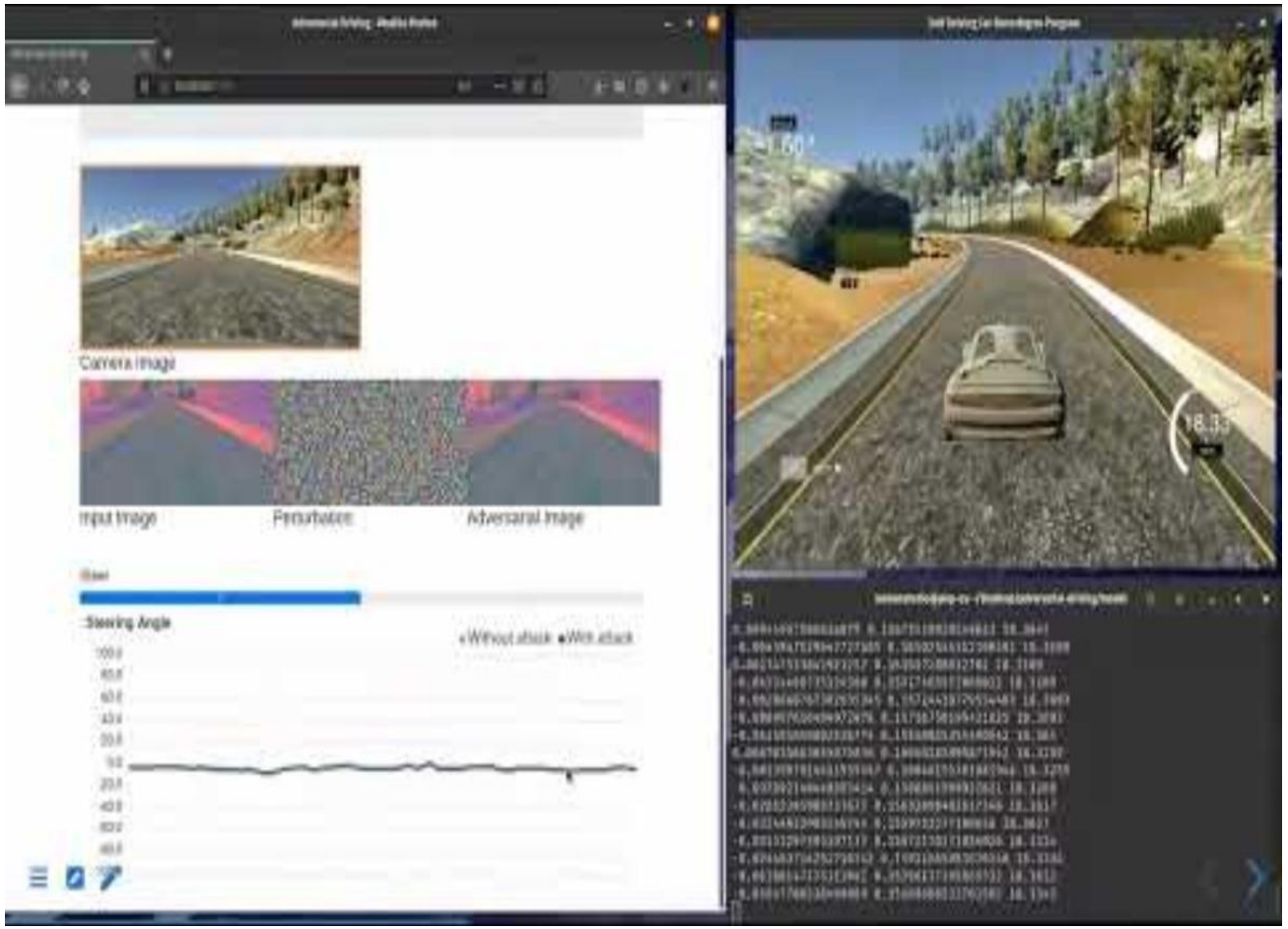
May 6th 2017

Share

Source: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

What can go wrong?





H. Wu, S. Yunas, S. Rowlands, W. Ruan and J. Wahlström, "Adversarial Driving: Attacking End-to-End Autonomous Driving," 2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, 2023, pp. 1-7, doi: 10.1109/IV55152.2023.10186386.

What can go wrong?

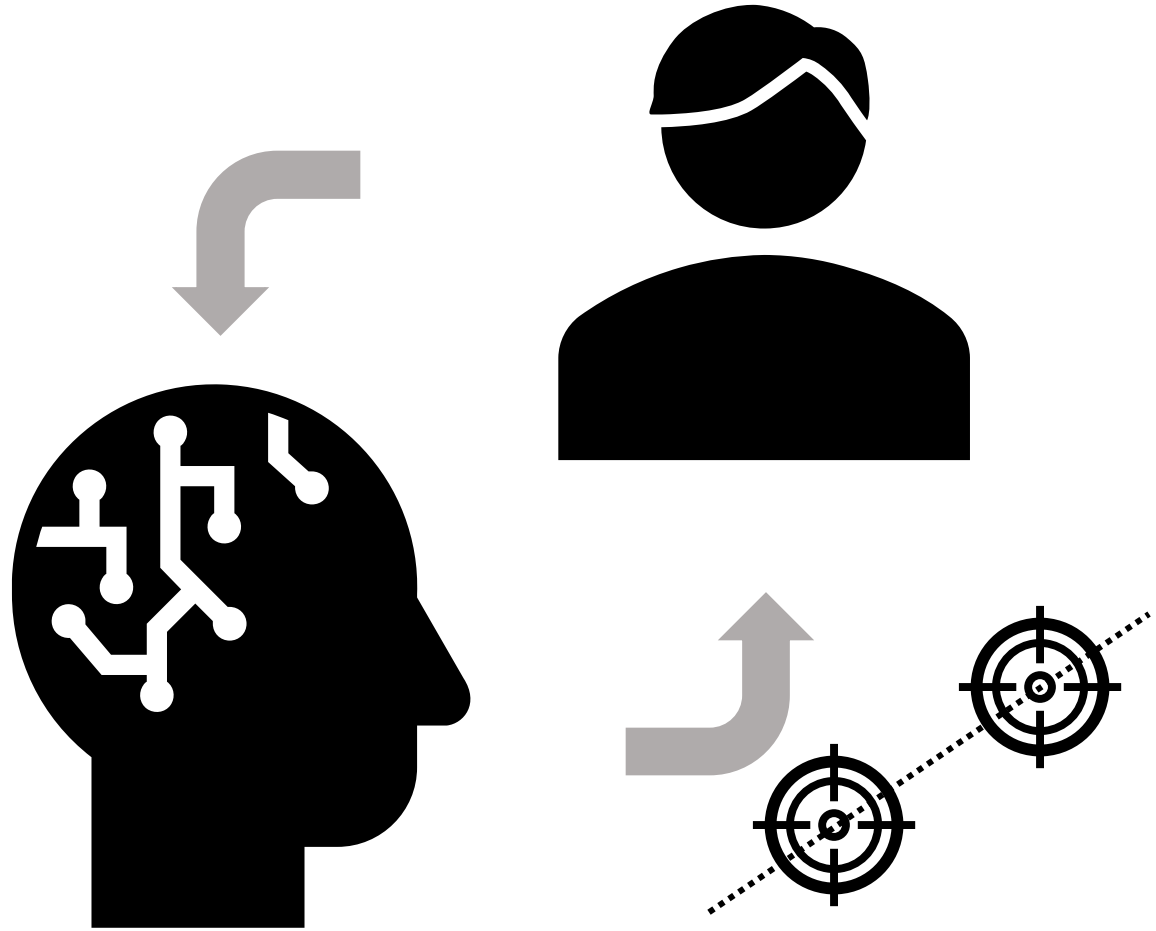
Paperclip Theory -- If you instructed a machine to optimize its paperclip production, it would eventually resort to dismantling objects such as computers, refrigerators, or any metal-based items once it depletes alternative sources of metal. This phenomenon is referred to as instrumental convergence.

https://nickbostrom.com/ethics/ai

Source: https://openai.com/research/faulty-reward-functions, 2016

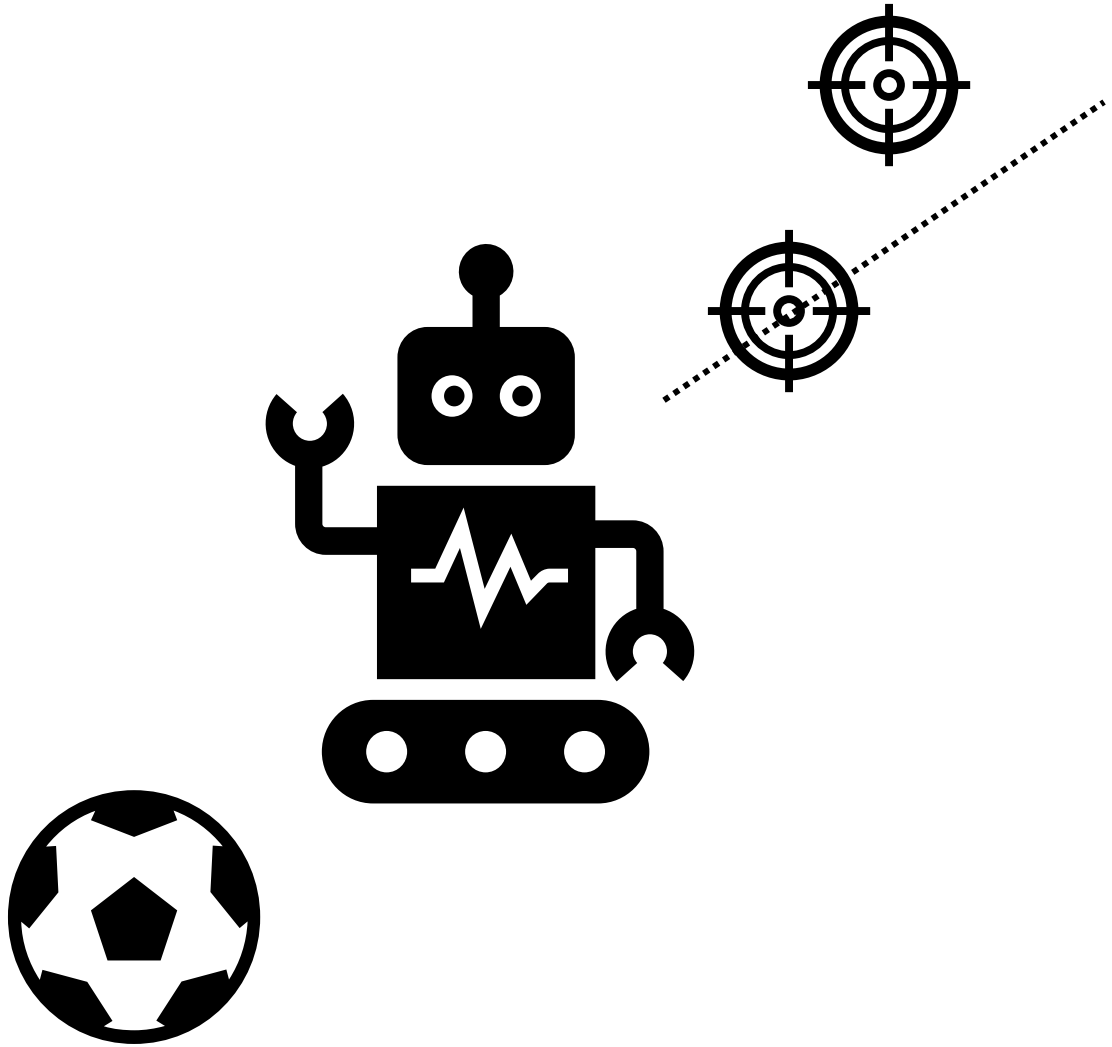
Predict the lecture quality

When we use AI, we agree on *something*



$$\max_{\theta} \ell\ell(\theta) = \max_{\theta} \sum_{i=1}^{N} \left[\mathbb{1}[y = +1] ln \frac{1}{1 + e^{-\theta x^{(i)}}} + \mathbb{1}[y = -1] ln \frac{e^{-\theta x^{(i)}}}{1 + e^{-\theta x^{(i)}}} \right]$$

When we use AI, we agree on *something*



- Large number of examples: https://vkrakovna.wordpress.com/ai-safety-resources/
- Stuart Russel: *Human Compatible: Artificial Intelligence and the Problem of Control*
- Biran Christian: The Alignment Problem: Machine Learning and Human Values
- *Max Tegmark: Life 3.0*

"But if machines are more intelligent than humans, then giving them the wrong objective would basically be setting up a kind of a chess match between humanity and a machine [...].

And **we wouldn't win** that chess match."

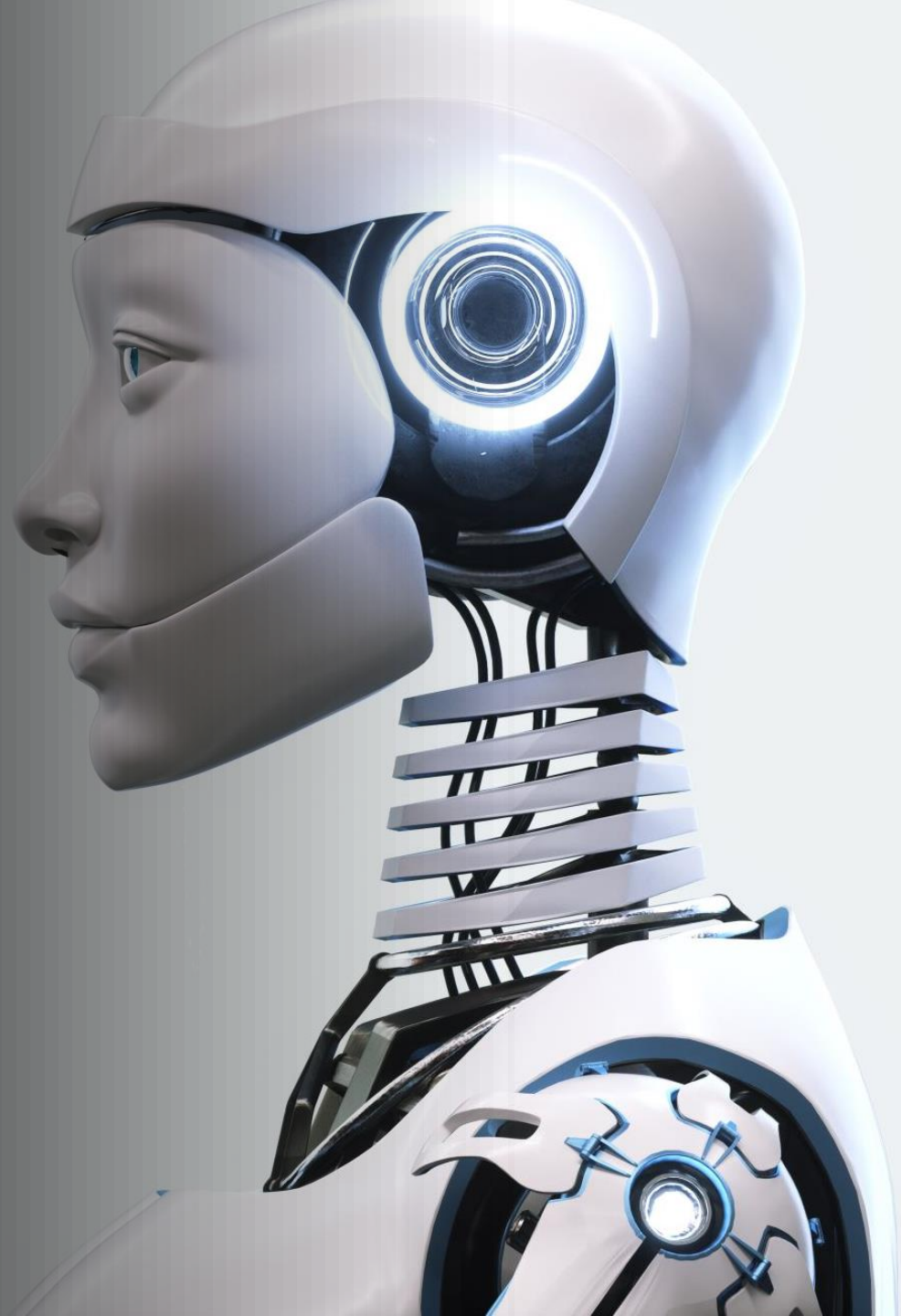
Stuart Russel

"The development of full artificial intelligence could **spell the end** of the human race"

Stephen Hawking

"If we build these devices to take care of everything for us, eventually they'll think faster than us and they'll **get rid of the slow humans** to run companies more efficiently."

Steve Wozniak



We need to know why...

- Why model made such decisions?
- What influenced the model decision mostly?
- What should be changed to change the model?
- What should be changed to change the model decision?

Otherwise, we will not be able to efficiently monitor and control their behaviour!

Where the true threats and opportunities lie?

- We should worry, but not panic
- Until 2018, it was enough to hide behind the door
- The threat is in the decision-making area, which is unrelated to the robots' motor skills or their ability to open doors.
- General AI – the last invention of humanity. And when will we invent it?
  - John McCarthy answered it very precisely in 1977 ;)
  - Stuart Russell – a few more breakthroughs in AI are needed
- Let's use AI to understand more - then there is a chance that we will make better AI in the future.

Source: https://en.wikipedia.org/wiki/STM_Kargu (CC BY 4.0)
STM Kargum, First flight 2017 – image recognition module to attack individuals

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

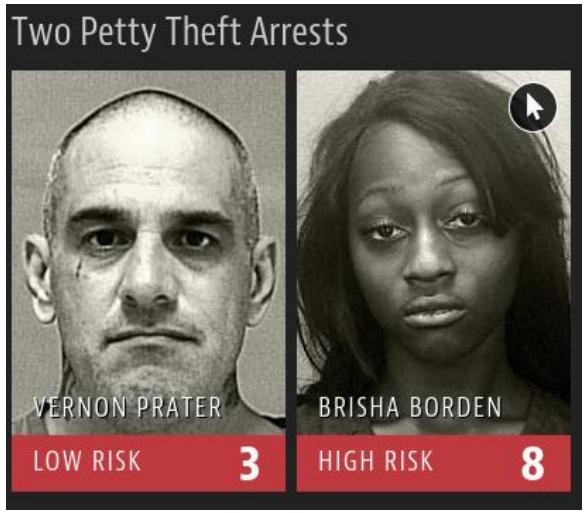
Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

High-stake decisions taken based on ML model predictions.
No dork-knob ability required to do huge harm

(2015 rok) Outdated, but still funny
https://youtu.be/g0TaYhjpOfo

XAI and ML and bias

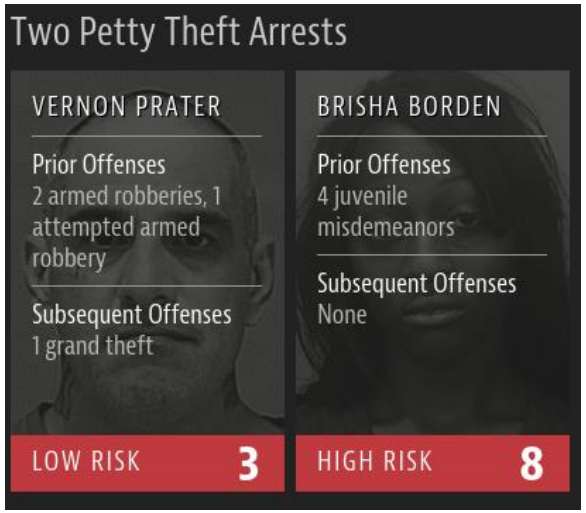
- COMPAS -- Correctional Offender Management Profiling for Alternative Sanctions system
- It was actually deployed in US and used by judges
- It was a black-box model



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

XAI and ML and bias

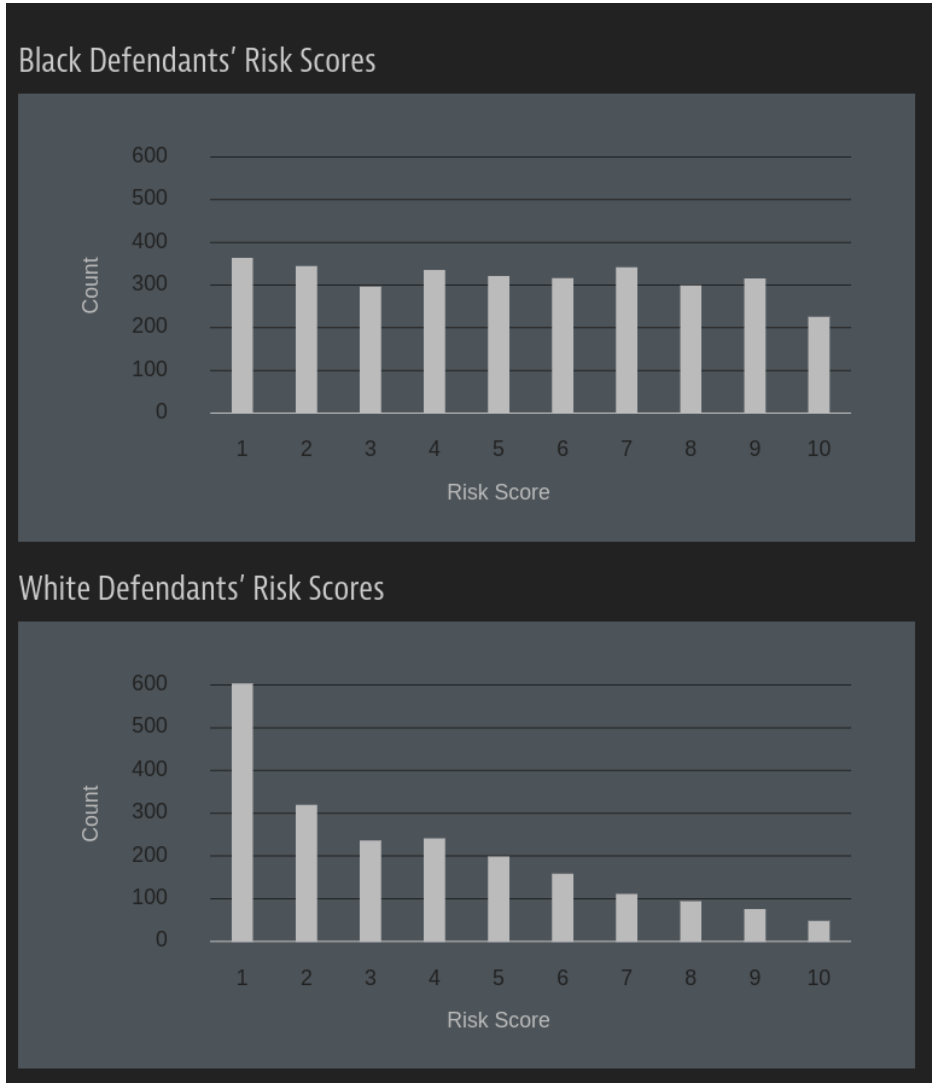
- COMPAS -- Correctional Offender Management Profiling for Alternative Sanctions system
- It was actuaslly deployed in US and used by judges
- It was a black-box model



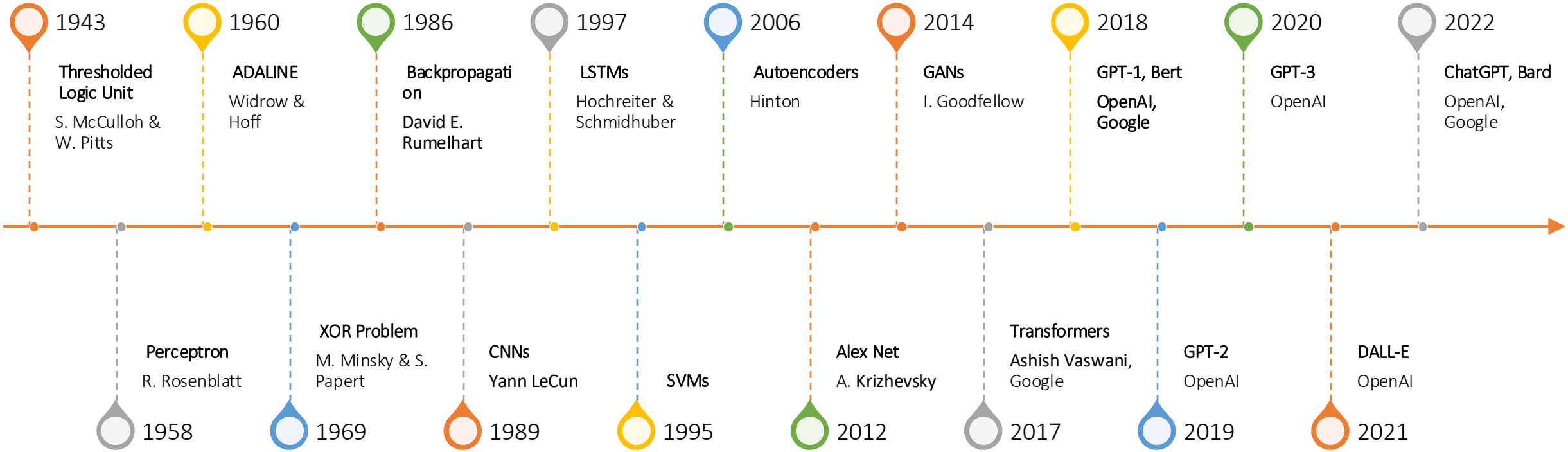
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

XAI and ML and bias

- Microsoft's racist chatbot Tay lasted 16 hours and being shut down for racist comments
- Google's still cannot find gorillas (neither MS or Apple) after issue reported by Jacky Alciné in 2015 (as of May 2023)
- Jigsaw (part of Alphabet) released dataset which purpose was prediction of online comments toxicity. There is unintended bias in the data posed by humans
- What else can go wrong?



Brief (not full) AI history



Open letter, DARPA, RODO, AI Act

OPEN-LETTER

Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter

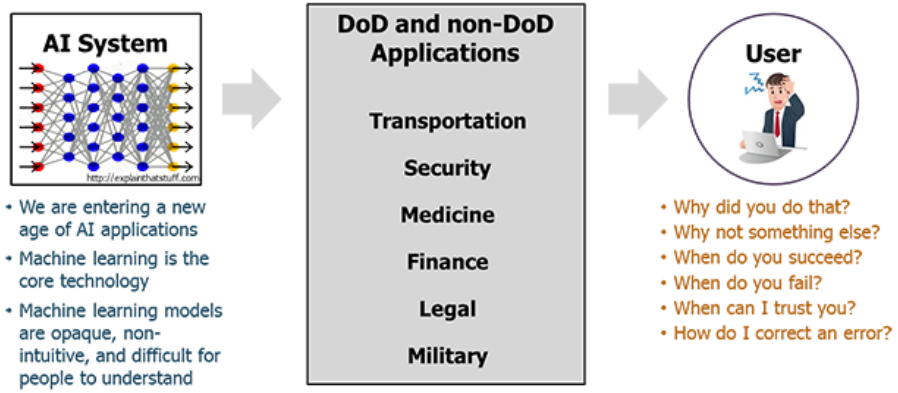
There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

Signatures
11251

Add your signature

PUBLISHED
October 28, 2015

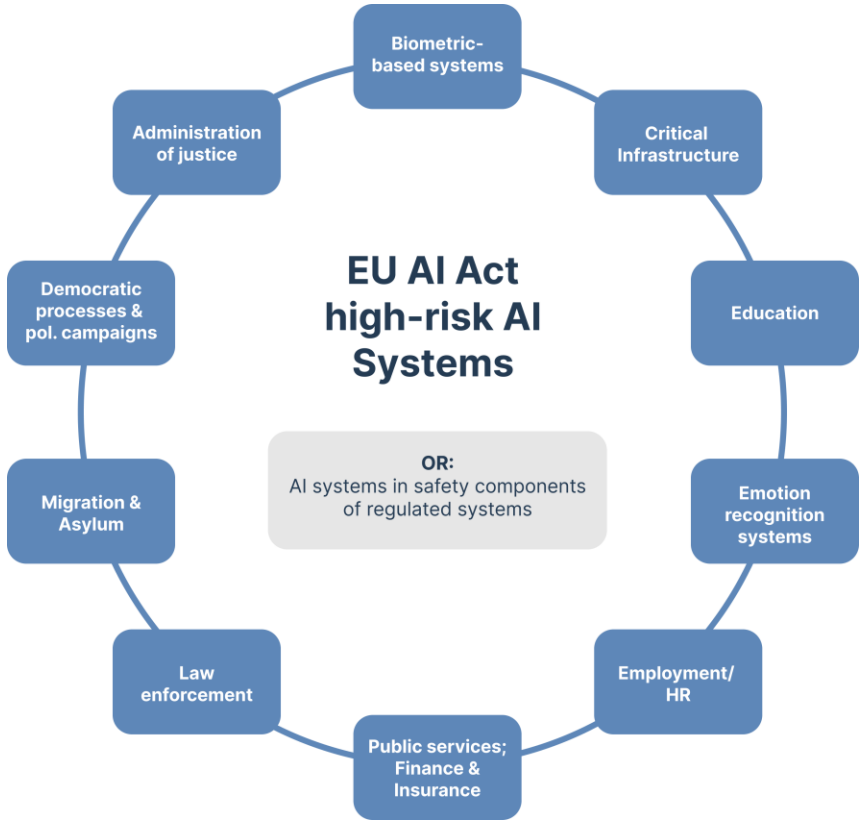
https://futureoflife.org/open-letter/ai-open-letter/



(2016)
https://www.darpa.mil/program/explainable-artificial-intelligence

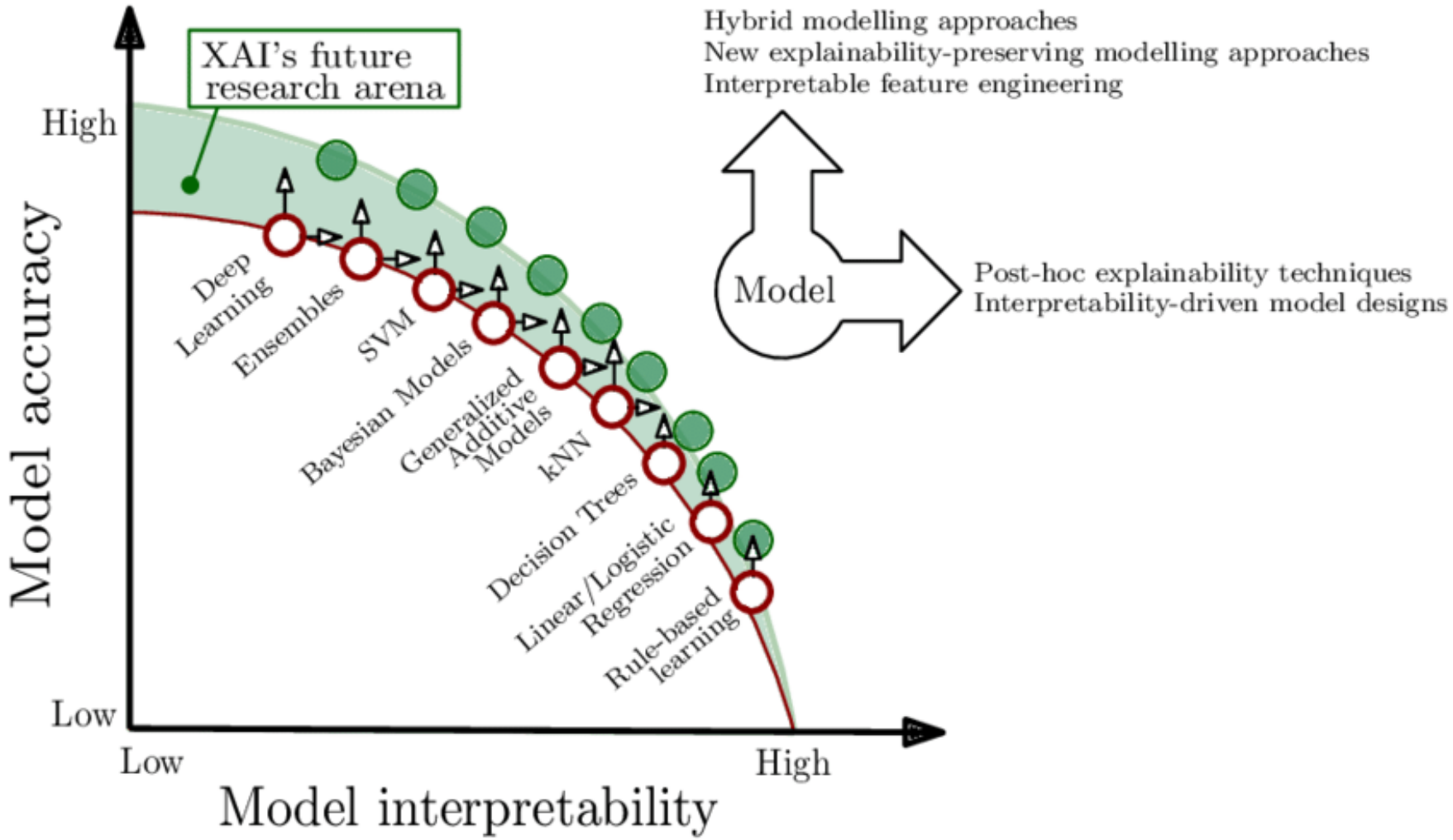
AI Act

- The AI Act classifies AI according to its risk:
  - Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI, **real-time** biometric identification systemsn, assesing emotional state of a person, predictive pricing). **Law enforcements may be allowed to use thsese** under high risk classification ;)
  - High-risk AI systems – the main focus of AI Act. Most of the AI systems used in law, or critical areas.
  - Limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
  - Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters – at least in 2021; this is changing with generative AI).



*"High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently **transparent** to enable deployers to **interpret a system's output** and use it appropriately."*

XAI and ML


XAI's future research arena
High
Model accuracy
Low
Low
High
Model interpretability
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs

XAI and ML and bias

Historical bias

- Historical bias occurs when the state of the world in which the data was generated is flawed.

Representation bias

- Representation bias occurs when building datasets for training a model, if those datasets poorly represent the people that the model will serve.

Measurement bias

- Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

Aggregation bias

- Aggregation bias occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group. (This is often not an issue, but most commonly arises in medical applications.)

Evaluation bias

- Evaluation bias occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.

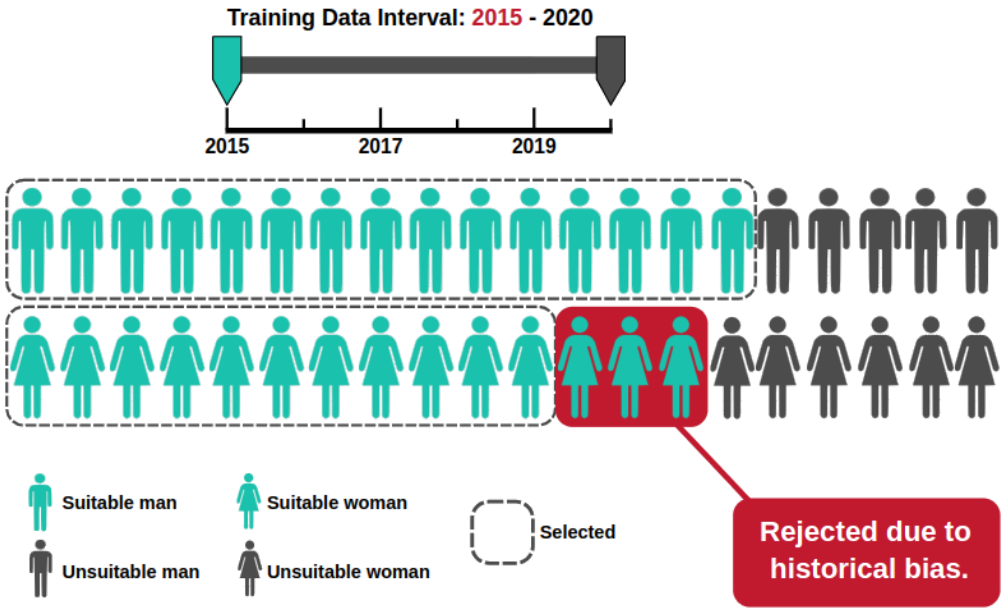
Deployment bias

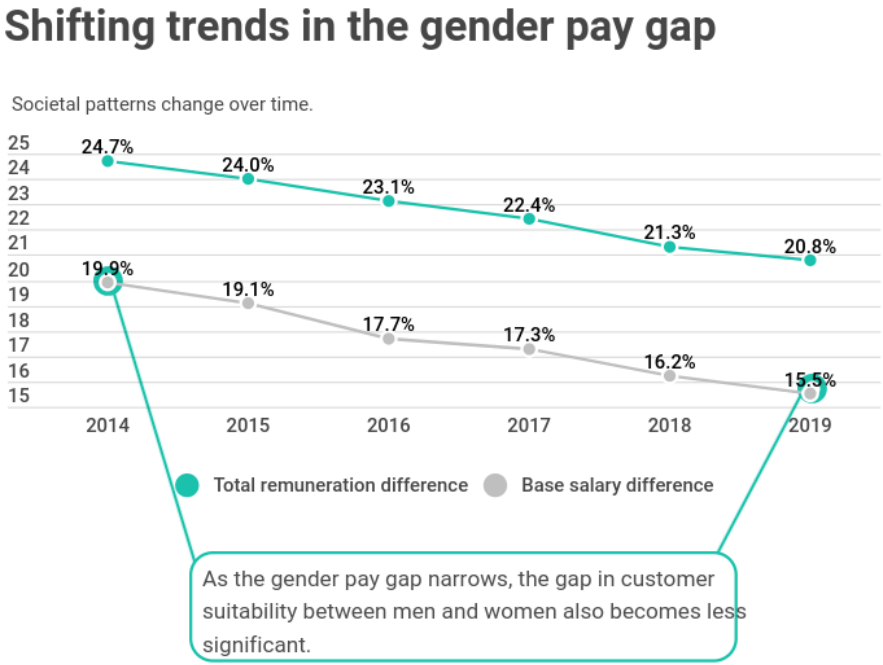
- Deployment bias occurs when the problem the model is intended to solve is different from the way it is actually used. If the end users don't use the model in the way it is intended, there is no guarantee that the model will perform well.

Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9. https://doi.org/10.1145/3465416.3483305

Historical bias

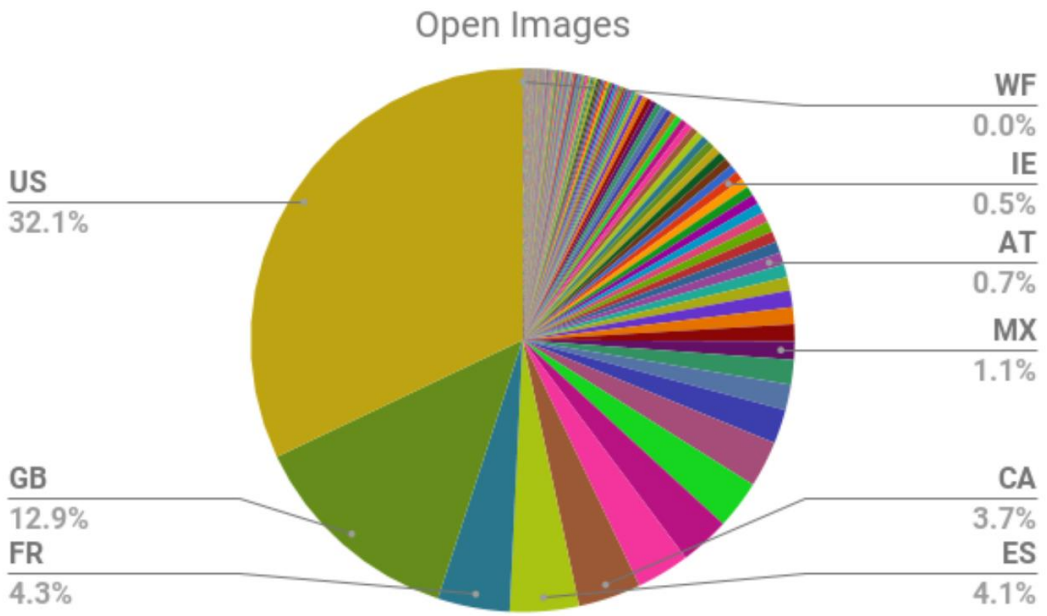
- Historical bias occurs when the state of the world in which the data was generated is flawed.

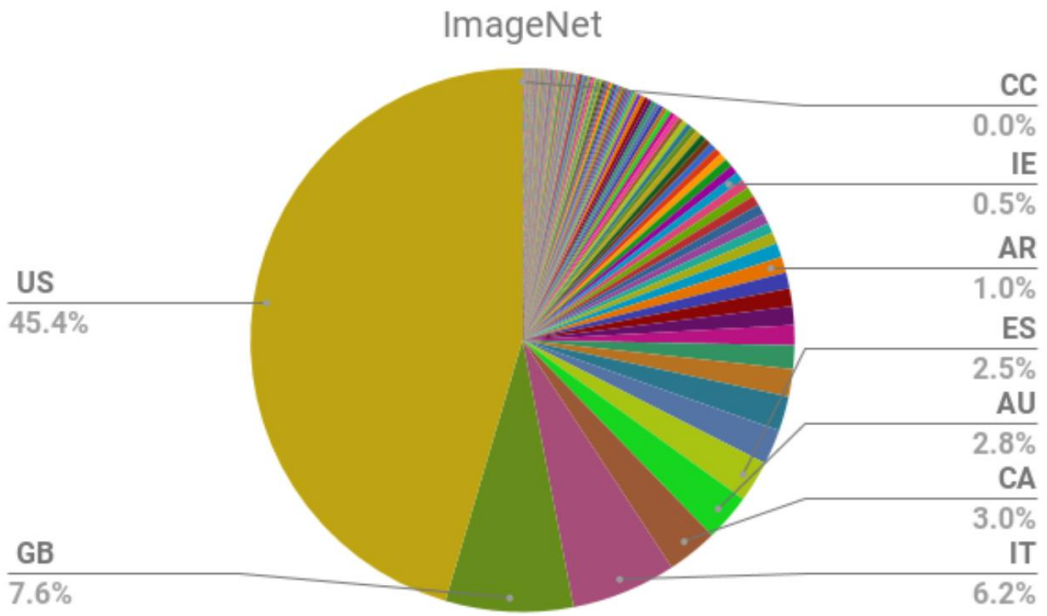




Representation bias

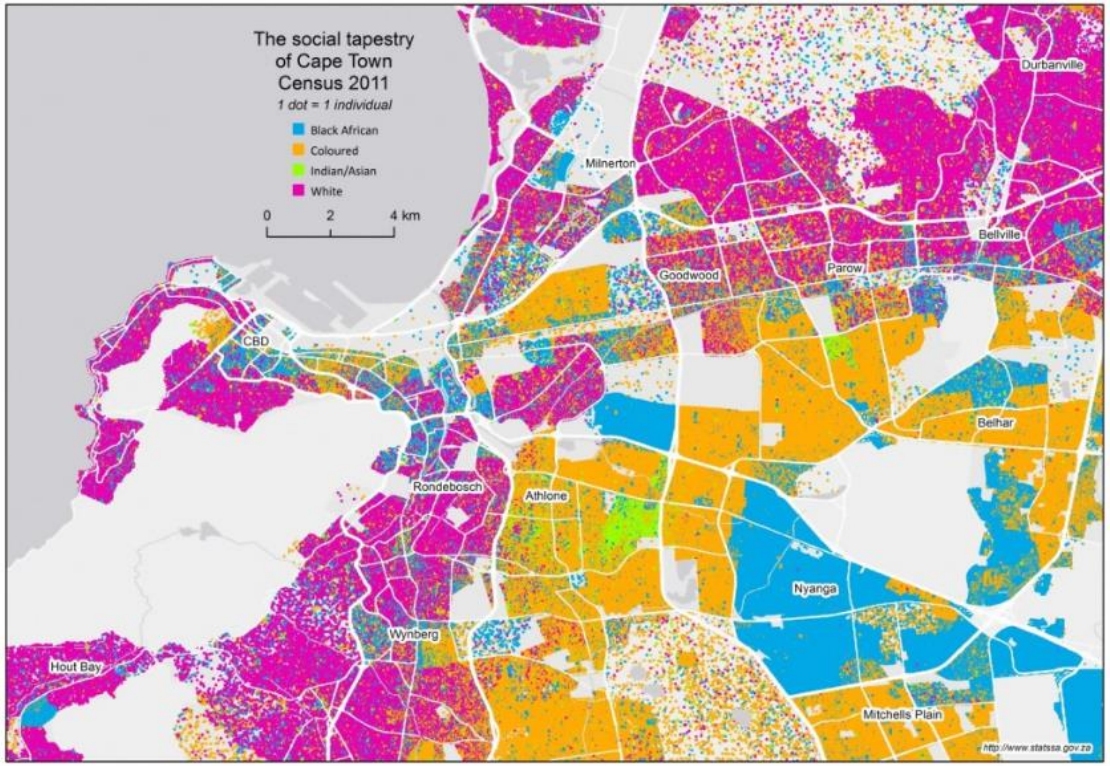
- Representation bias occurs when building datasets for training a model, if those datasets poorly represent the instances that the model will serve.

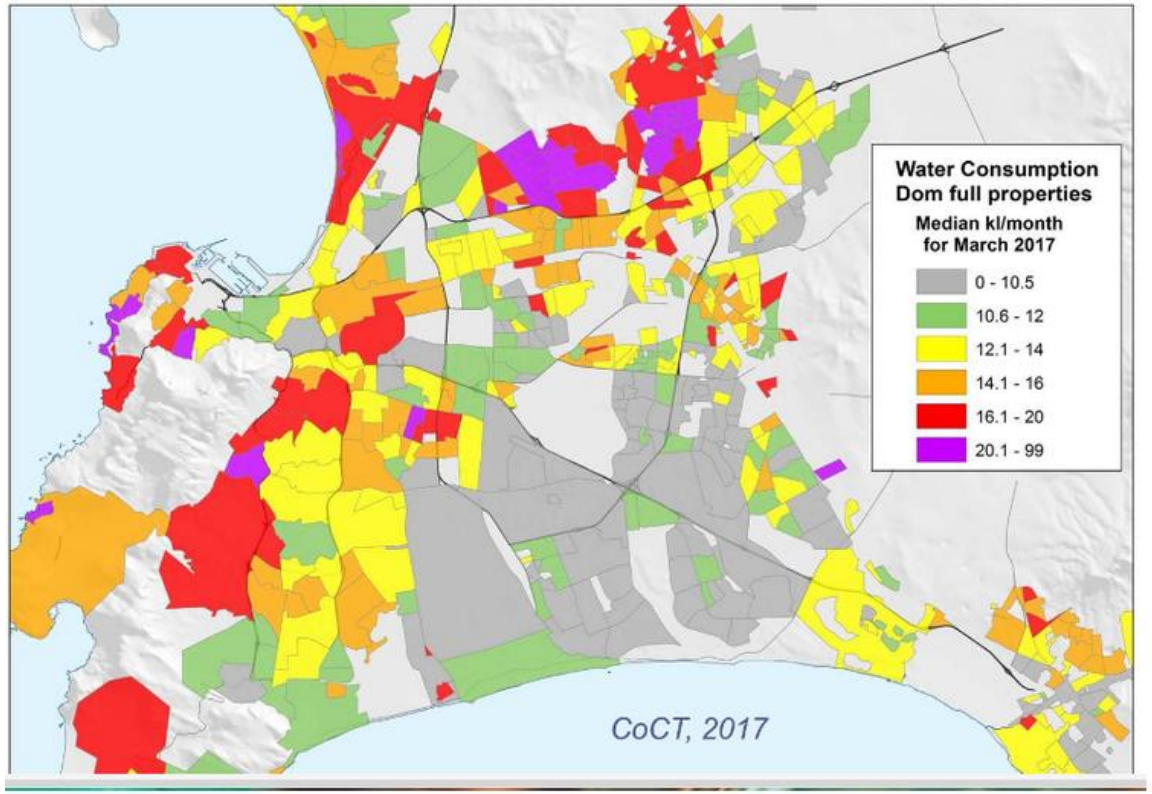




Measurement bias

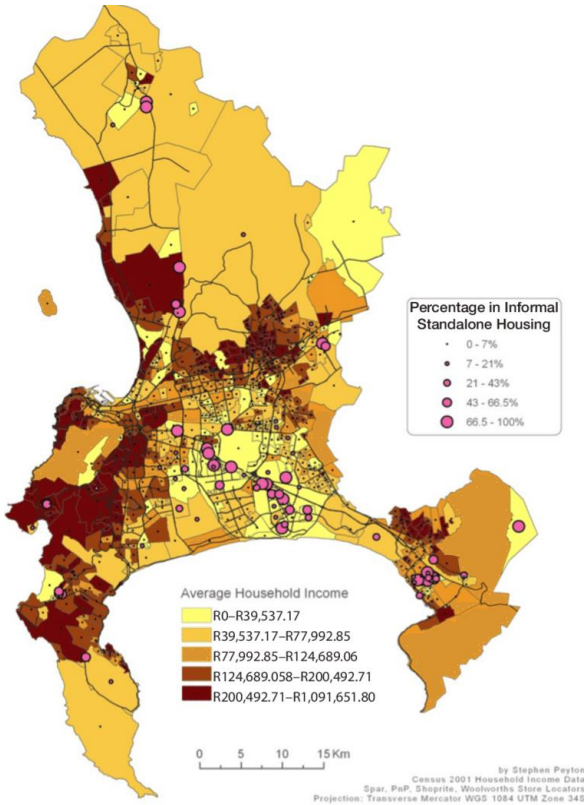
- Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

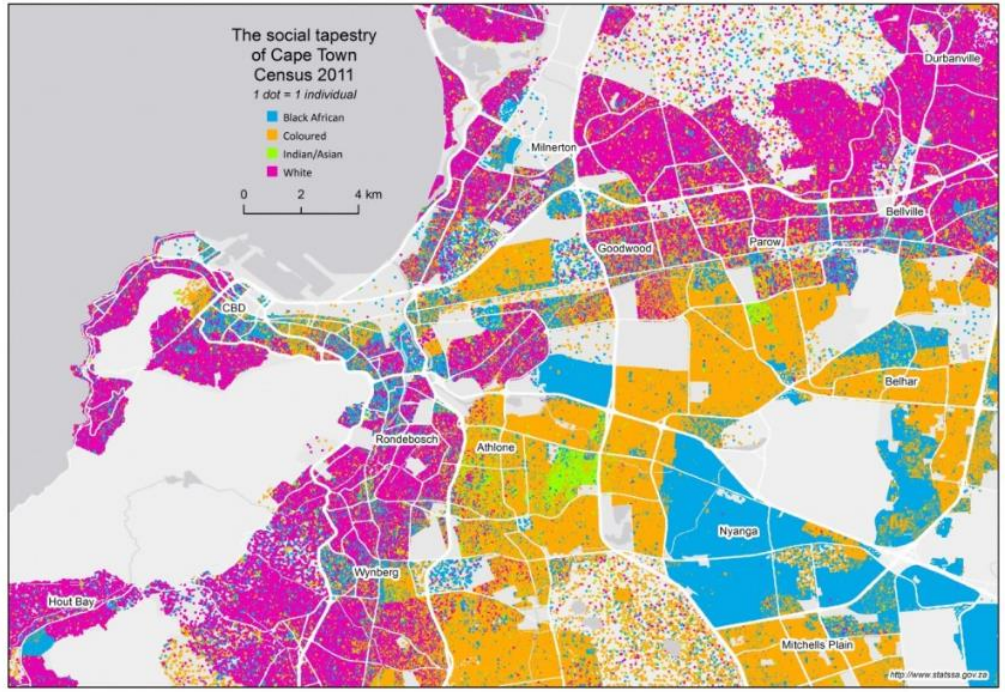


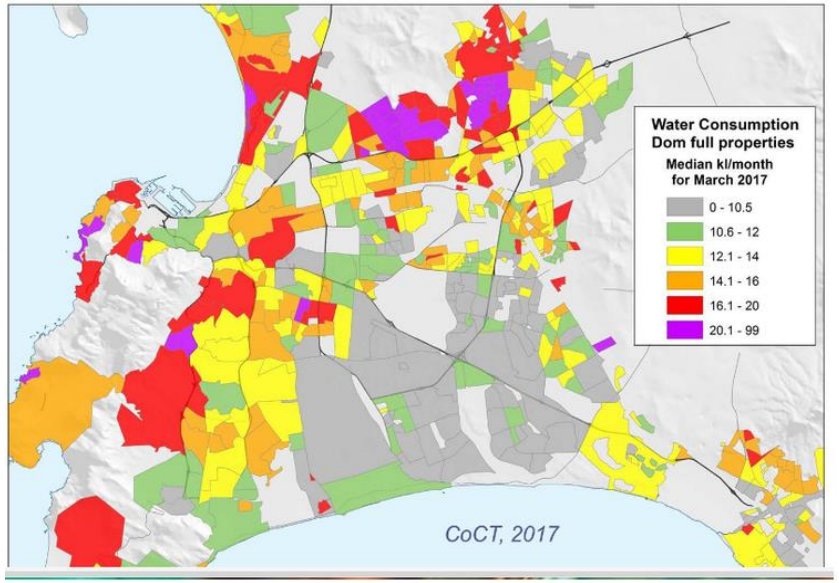


Measurement bias

- Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.





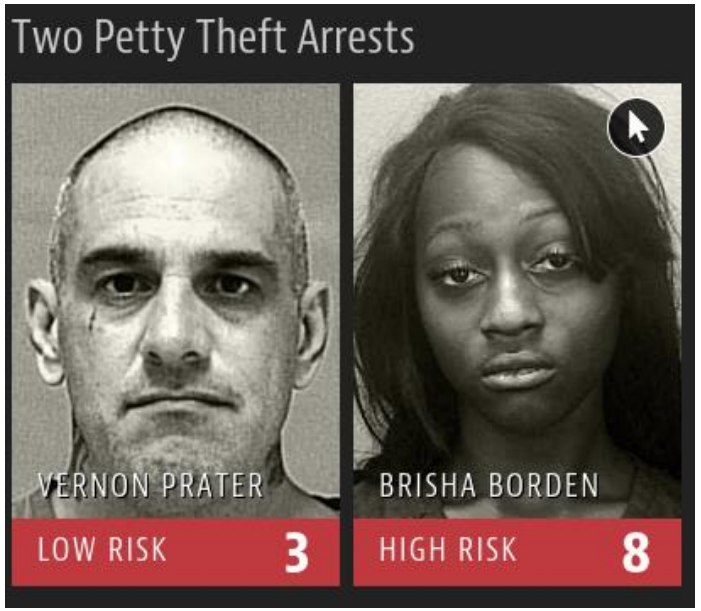


Measurement bias

- Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

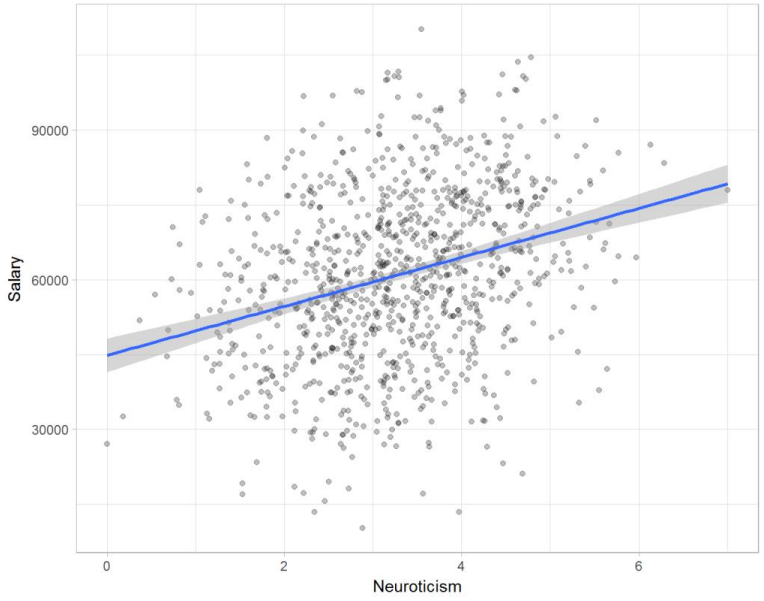
"Give me the man and I will give you the case against him"

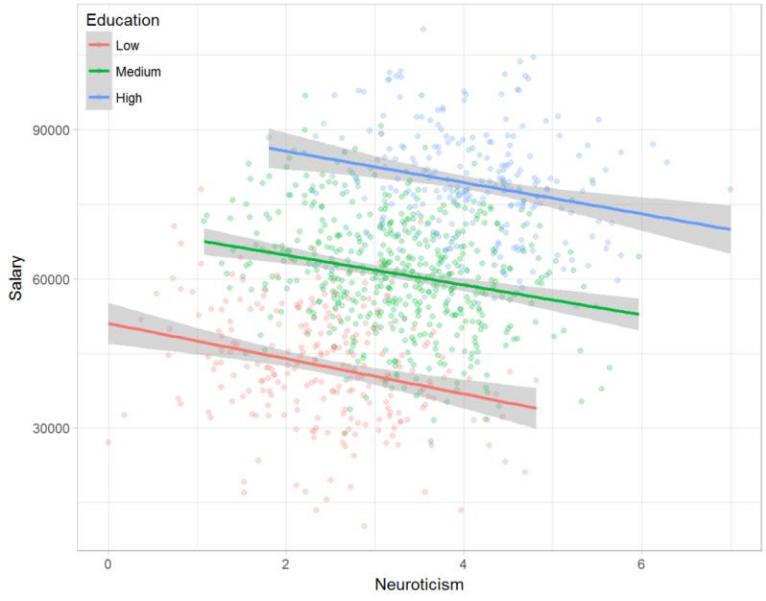
Some areas inhabitted by people of specific skin color, are more often patroled by the police

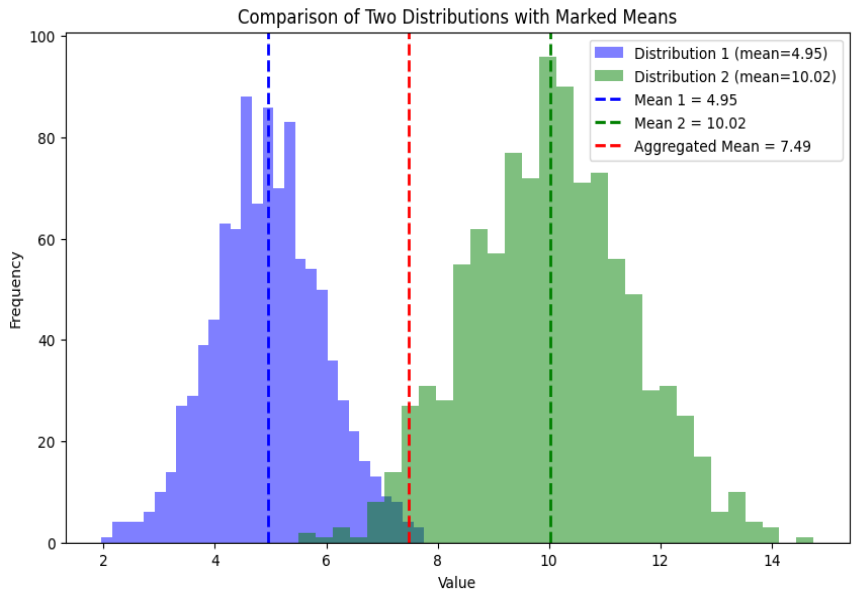


Aggregation bias

- Aggregation bias occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group. (This is often not an issue, but most commonly arises in medical applications.)

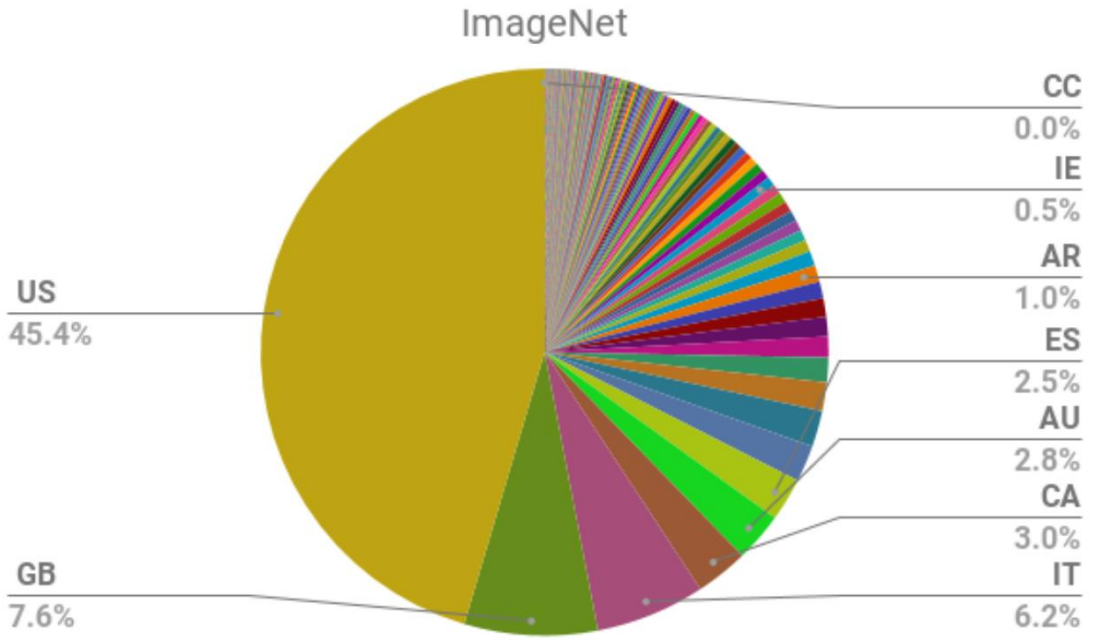






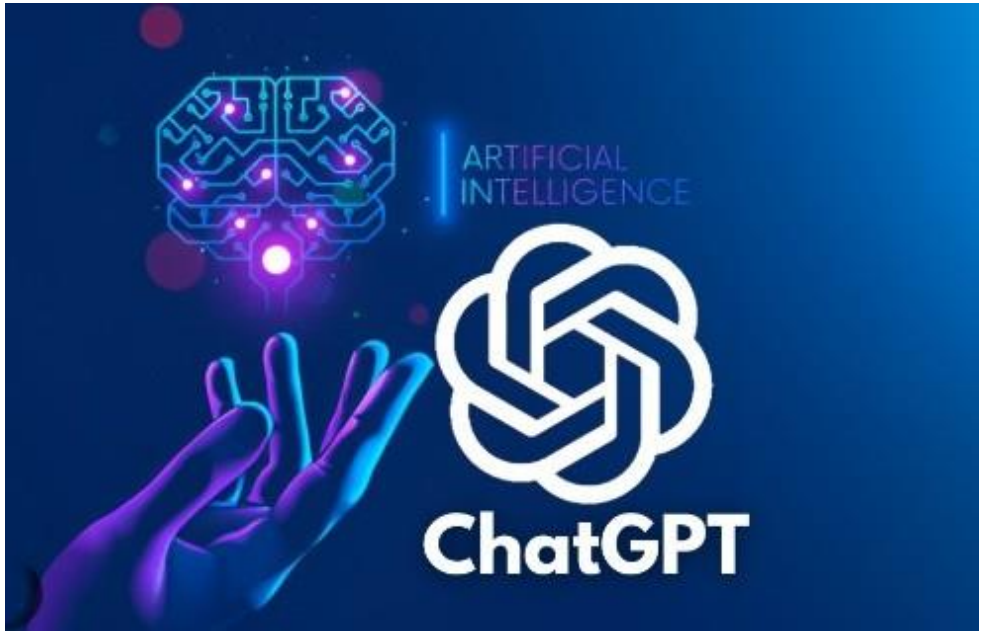
Evaluation bias

- Evaluation bias occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.

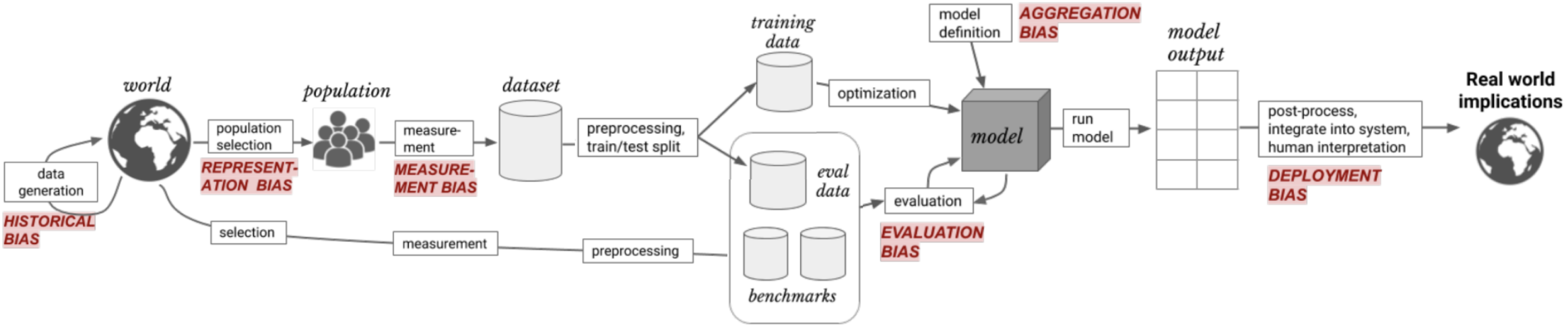


Deployment bias

- Deployment bias occurs when the problem the model is intended to solve is different from the way it is actually used. If the end users don’t use the model in the way it is intended, there is no guarantee that the model will perform well.



XAI and ML and bias



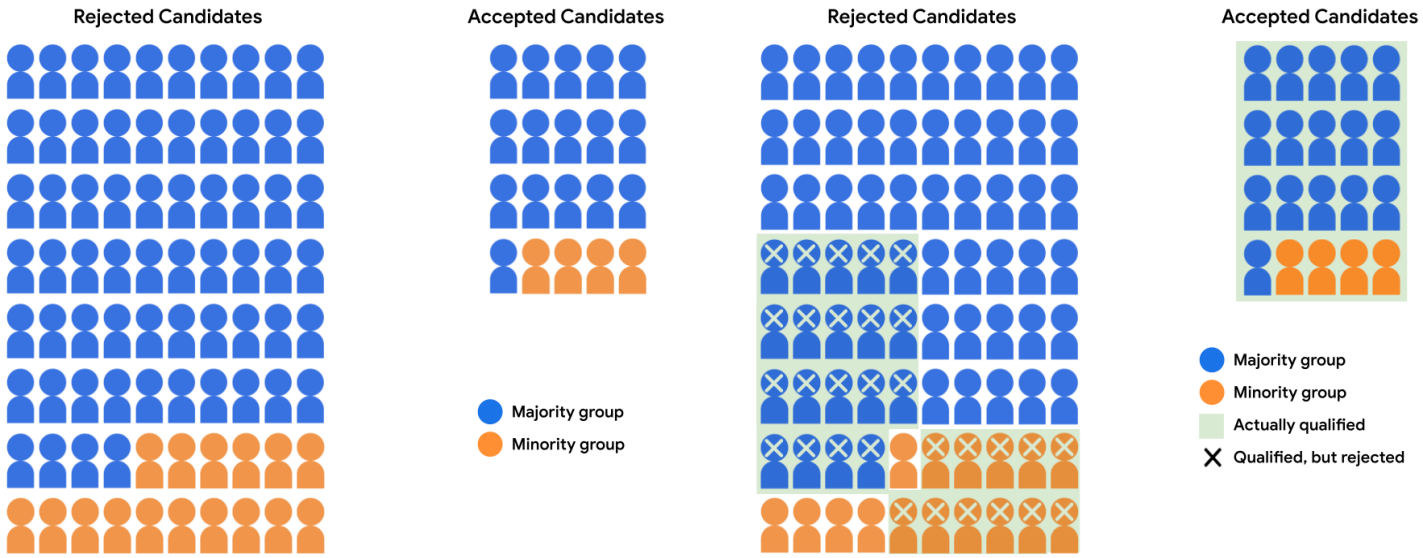
Harini Suresh and John Guttag. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21). Association for Computing Machinery, New York, NY, USA, Article 17, 1–9. https://doi.org/10.1145/3465416.3483305

Meassuring and mitigating bias

- **Equalized odds vs. Demographic parity**
- Equality of Opportunity
- Disparate Impact
- Counterfactual Fairness

$Pr(\boldsymbol{Y}|D{=}unprivileged) - Pr(\boldsymbol{Y}|D{=}privileged) = 0$

Prediction should be independent of the attribute D (for instance skin color). In equalized odds, we can add more constrints (e.g. qualitfication)

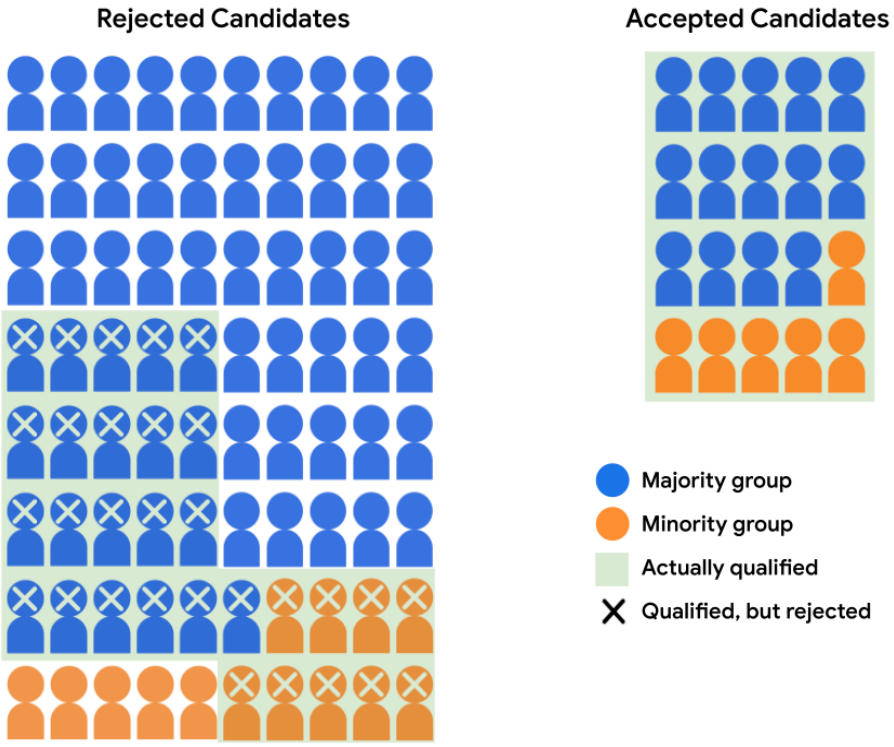


Meassuring and mitigating bias

- Equalized odds vs. Demographic parity
- **Equality of Opportunity**
- Disparate Impact
- Counterfactual Fairness

$Pr(\mathbf{Y=1}|D=unprivileged, Y=1) - Pr(\mathbf{Y=1}|D=privileged, Y=1) = 0$

Prediction should be independent of the attribute D (for instance skin color) but only for a specific class

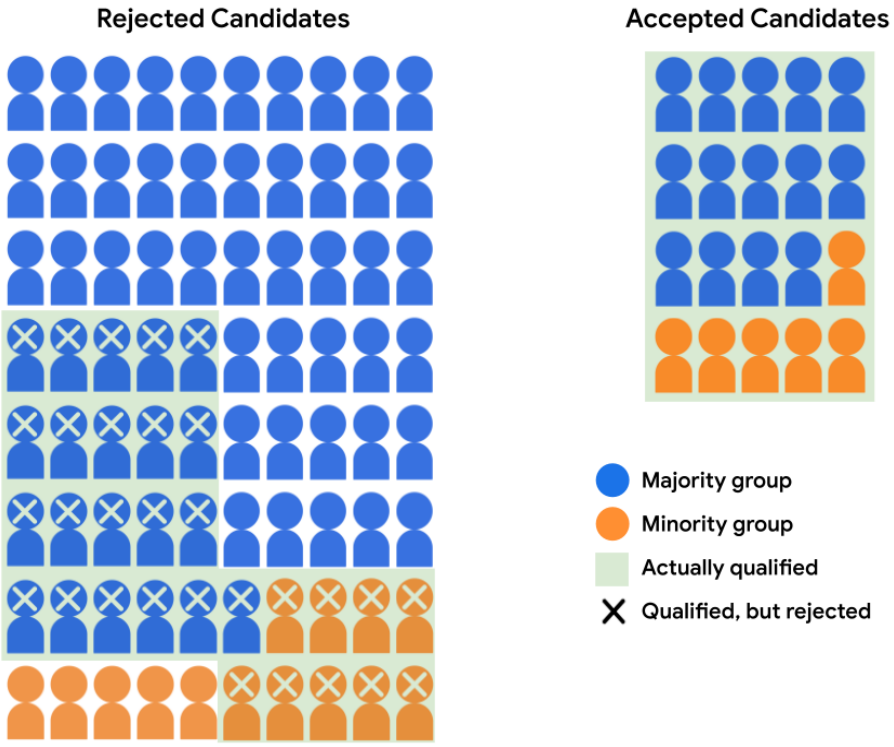


Meassuring and mitigating bias

- Equalized odds vs. Demographic parity
- Equality of Opportunity
- **Disparate Impact**
- Counterfactual Fairness

$$\frac{Pr(Y=1|D=unprivileged)}{Pr(Y=1|D=privileged)}$$

Similat to equal opportunity, but measuring it as a ratio, not difference

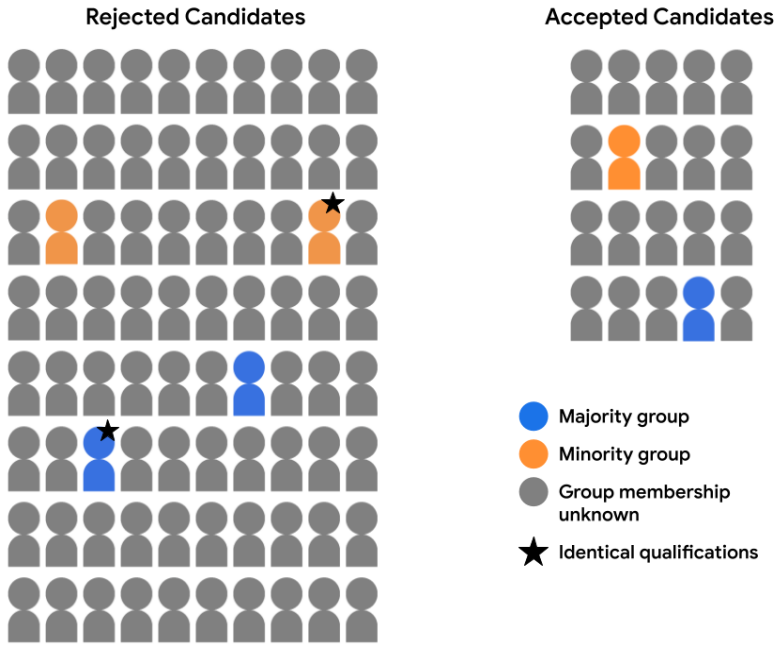


Meassuring and mitigating bias

- Equalized Equalized odds vs. Demographic parity
- Equality of Opportunity
- Disparate Impact
- **Counterfactual Fairness**

$Pr(\boldsymbol{Y}_i|D=unprivileged) - Pr(\boldsymbol{Y}_i|D=privileged) = 0$

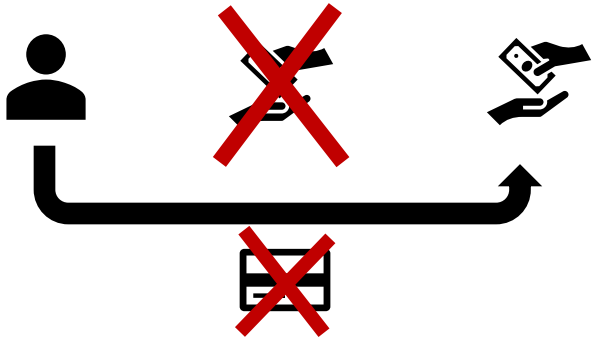
We only have informaiton about *D* for small subset of data. Wantoto check whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more sensitive attributes.

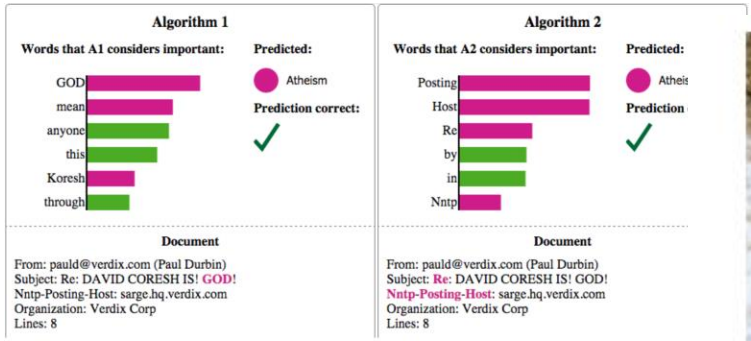


Meassuring and mitigating bias

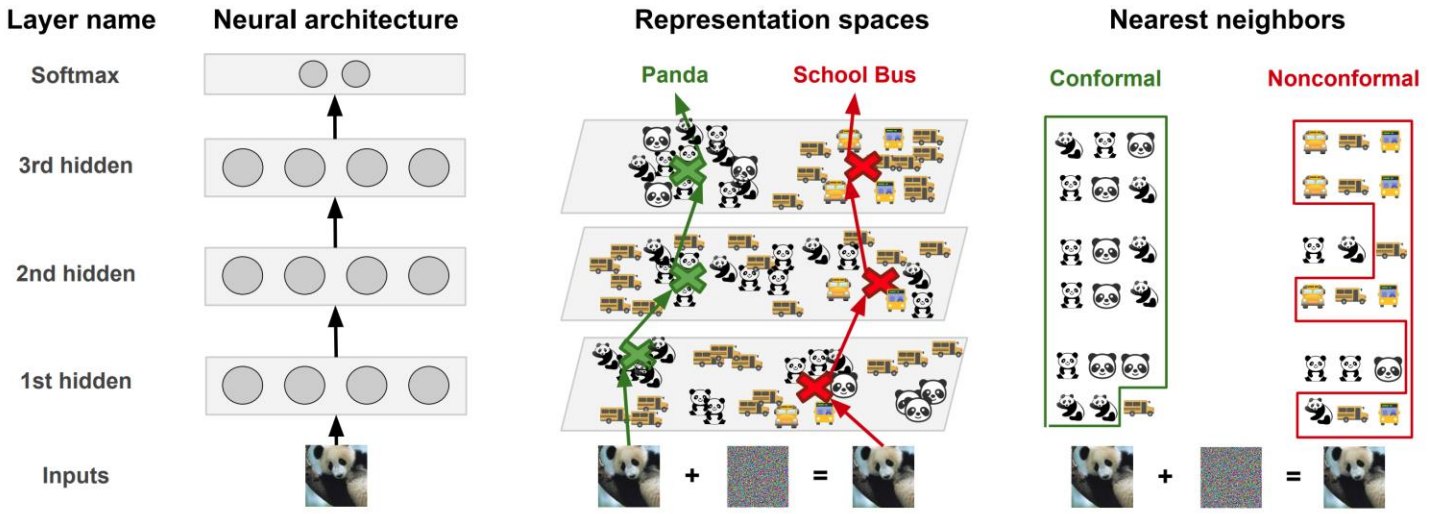
- Pre-processing
  - Unawareness - In this technique, we attempt to decrease algorithmic bias by removing sensitive/protected attributes from training data. This is called unawareness.
  - Reweighting, downsampling
  - Augmentation – new data generaiton, counterfactual augmentation
- In-modelling
  - Prejudice Remover Regularizer – ML model that adds regularizaiton term to assure fairness
- Post-processing
  - MLDebiaser – debiasing the output of balck-box model with additional calibration

Examples of XAI

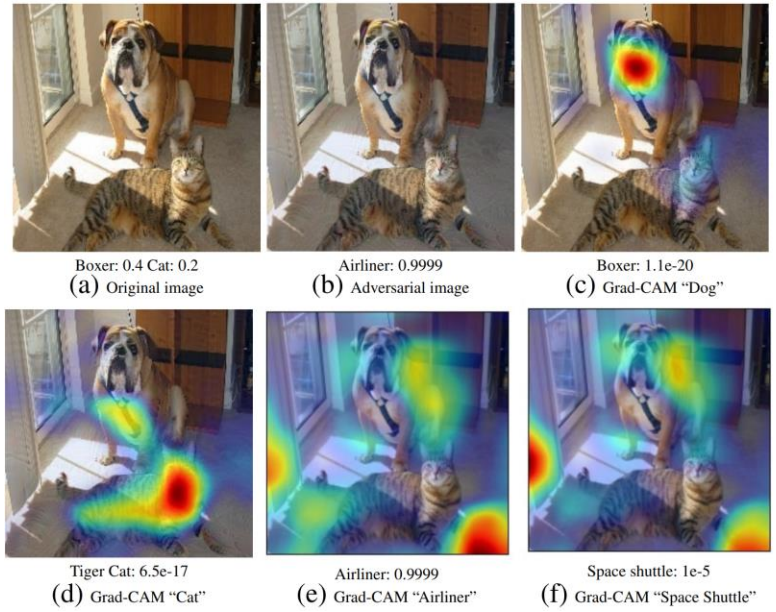




M. T. Ribeiro, S. Singh, and C. Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

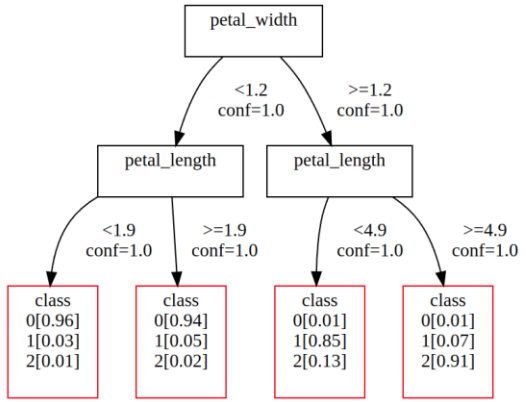


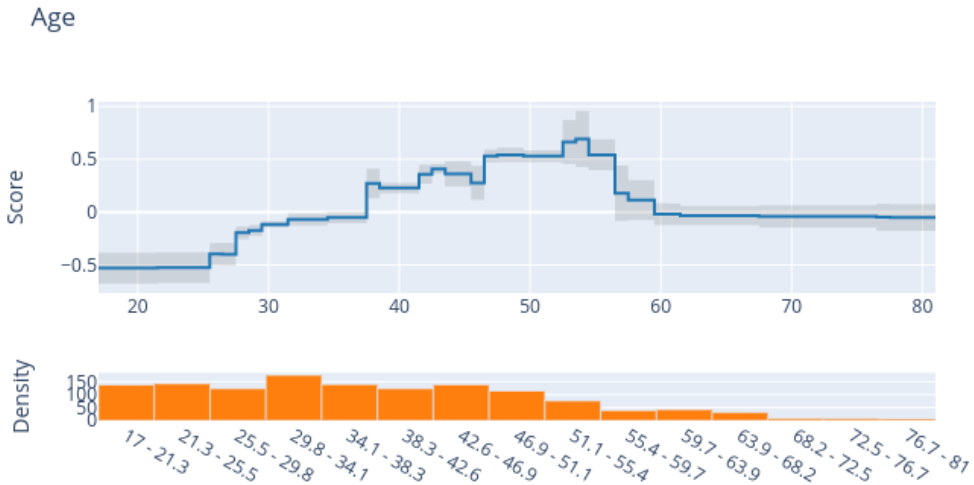
Papernot, N., & Mcdaniel, P. (2018). Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. ArXiv, abs/1803.04765.



R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, i D. Batra, „Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization", w 2017 IEEE International Conference on Computer Vision (ICCV), paź. 2017, s. 618–626. doi: 10.1109/ICCV.2017.74.

Examples of XAI

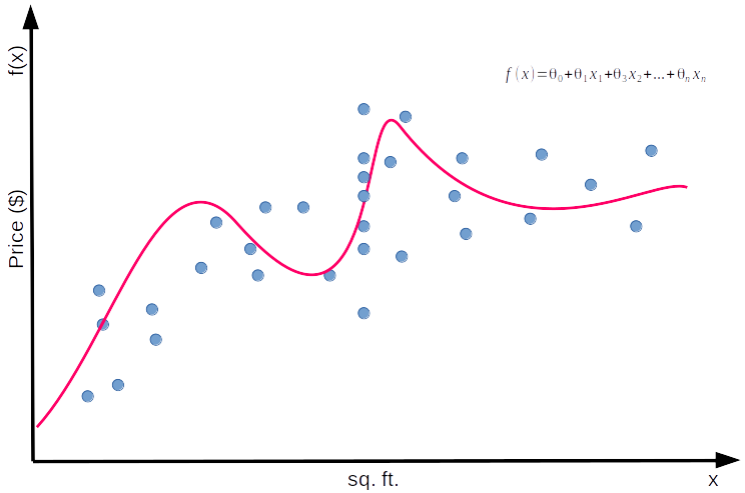


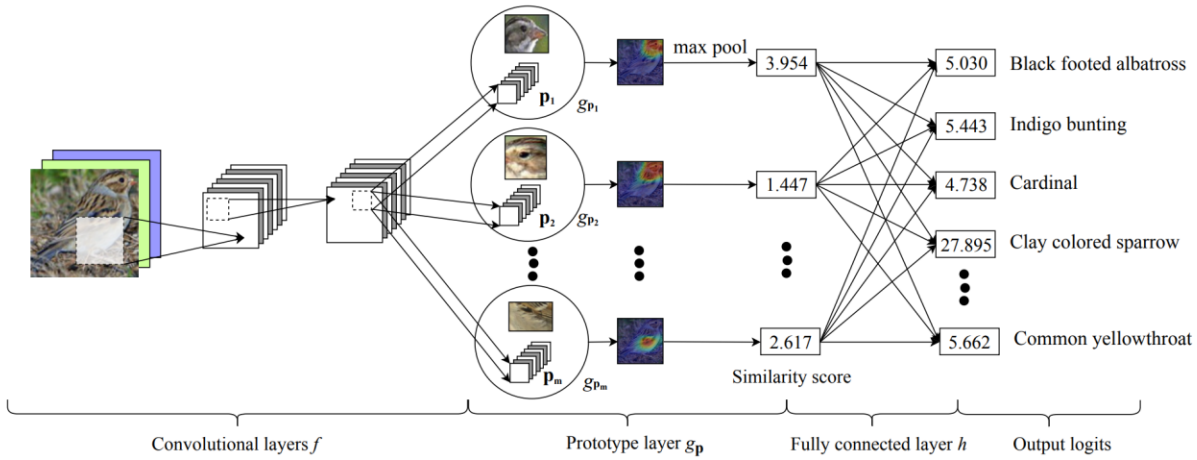


Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 623–631. 2013.

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Rudin, C. **Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead**. Nat Mach Intell 1, 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

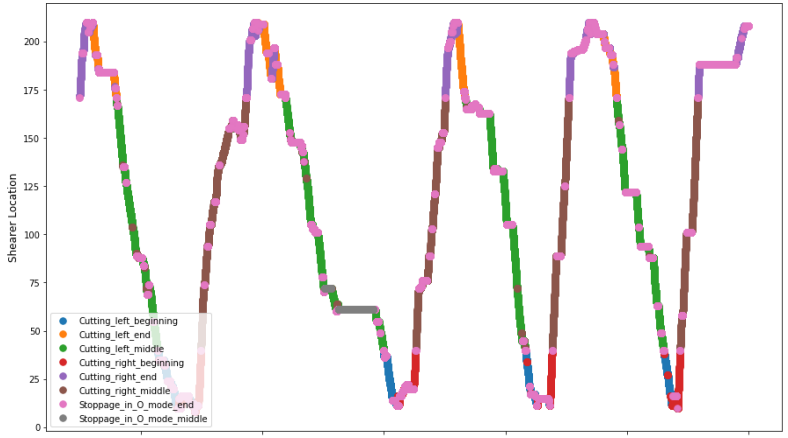


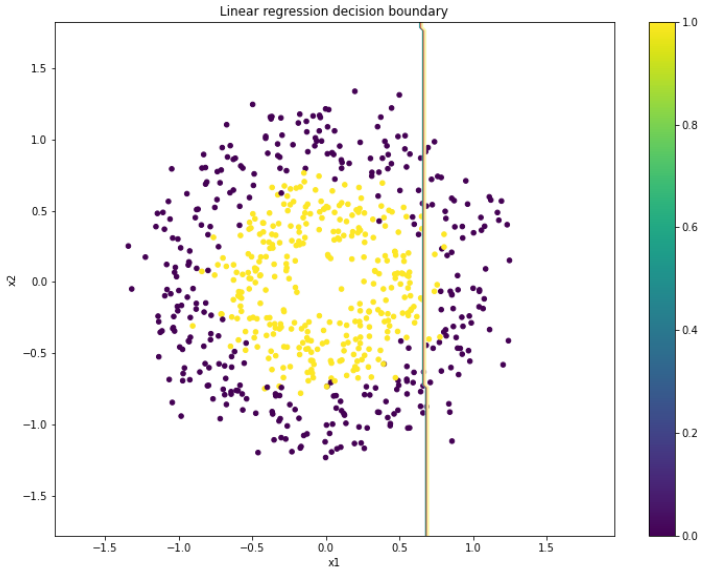


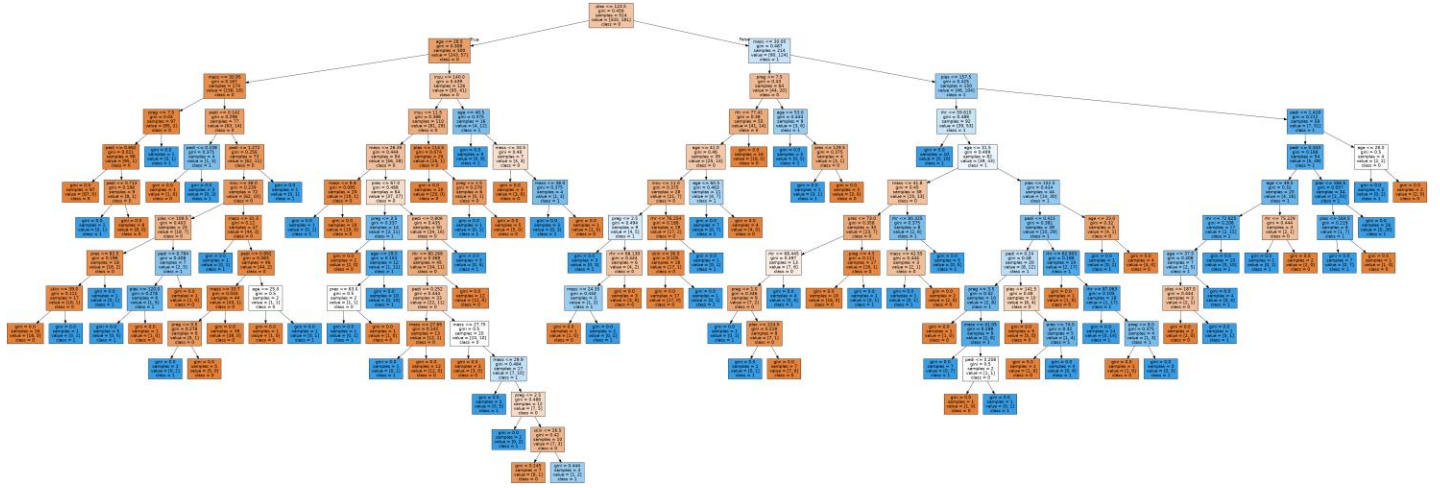
C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, i C. Rudin, „This Looks Like That: Deep Learning for Interpretable Image Recognition”. arXiv, 28 grudzień 2019. doi: 10.48550/arXiv.1806.10574.

What is so difficult?

- An explanation that is not understandable
- An explanation that is intended for someone else
- An explanation that is incorrect
- An explanation that is correct, but not true!

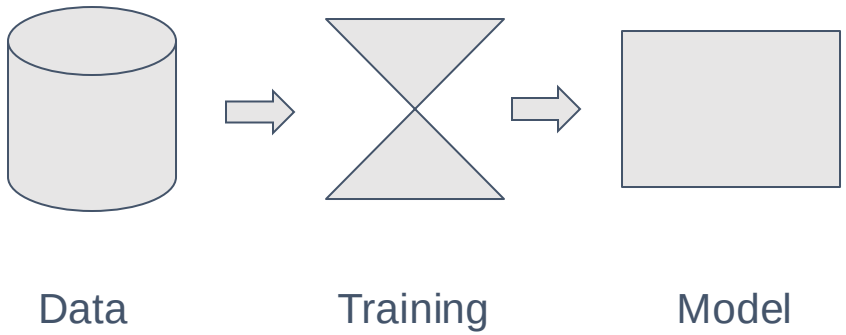






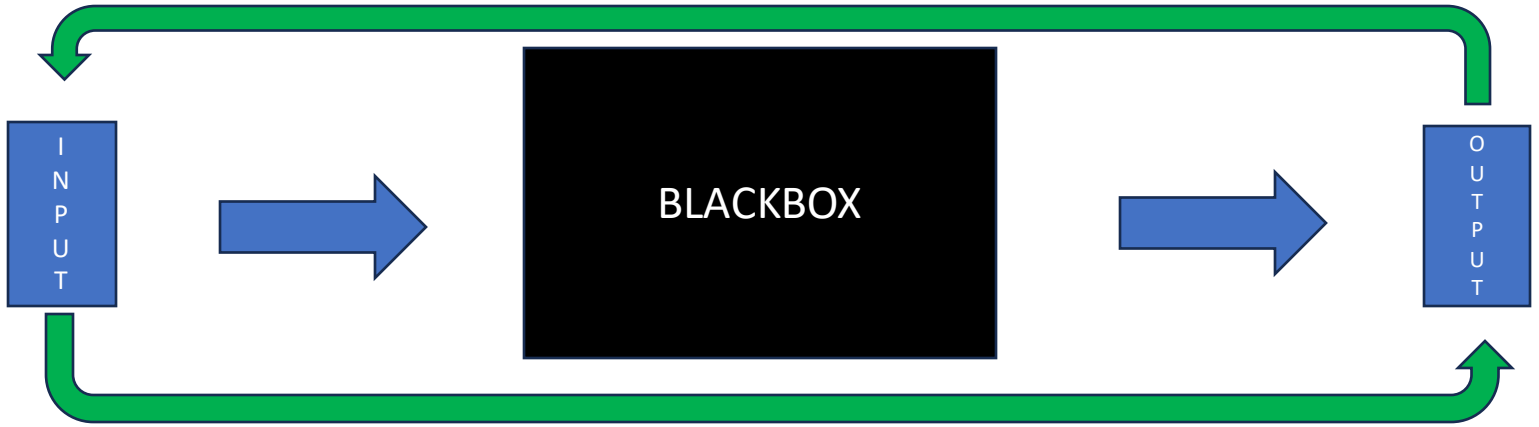
Pre-modelling, Post-modelling methods, In-modelling

- Pre-modeling methods
  - Exploratory data analysis, knowledge discovery
  - Prototypes-critics
- Post-modelling methods
  - Global: PDP, ICE, ALE, Feature importances, surrogate models, etc.
  - Local: SHAP, LIME, Lore, LUX, GradCam, etc.
  - Counterfactual explanations: DICE, Wach, CEM, etc.
  - Adversarial examples
- In-modelling methods
  - Simple models: Linear regression, decision trees, etc
  - ProtoPNet, Self Explainable Neural Networks, etc.
  - Explainable Boosting Machines

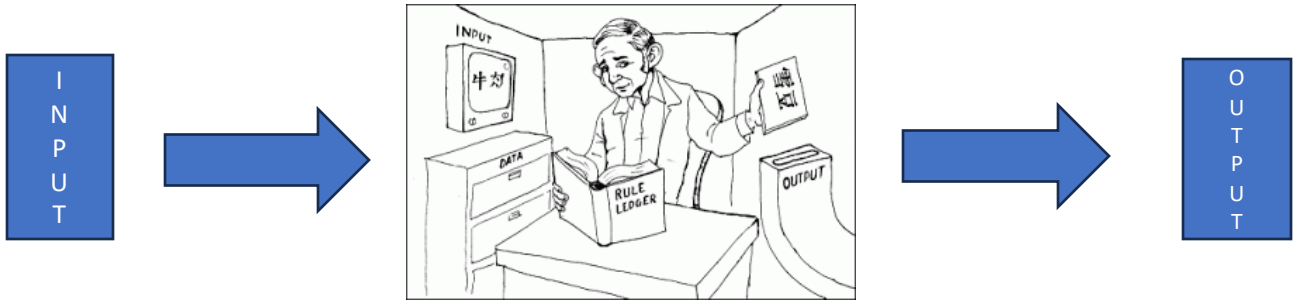


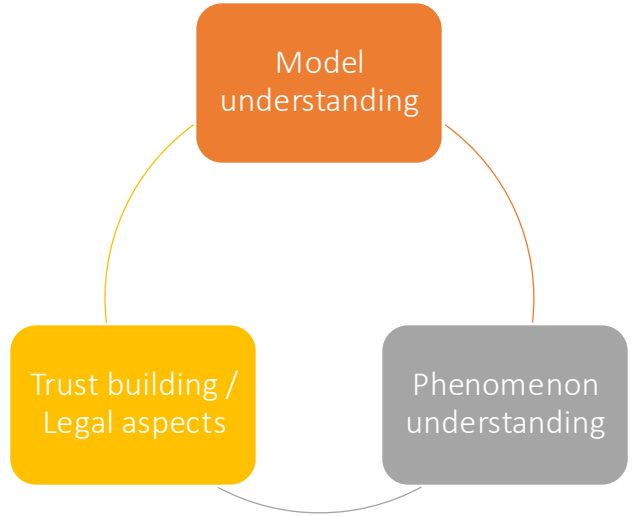
Glassbox and whitebox

XAI

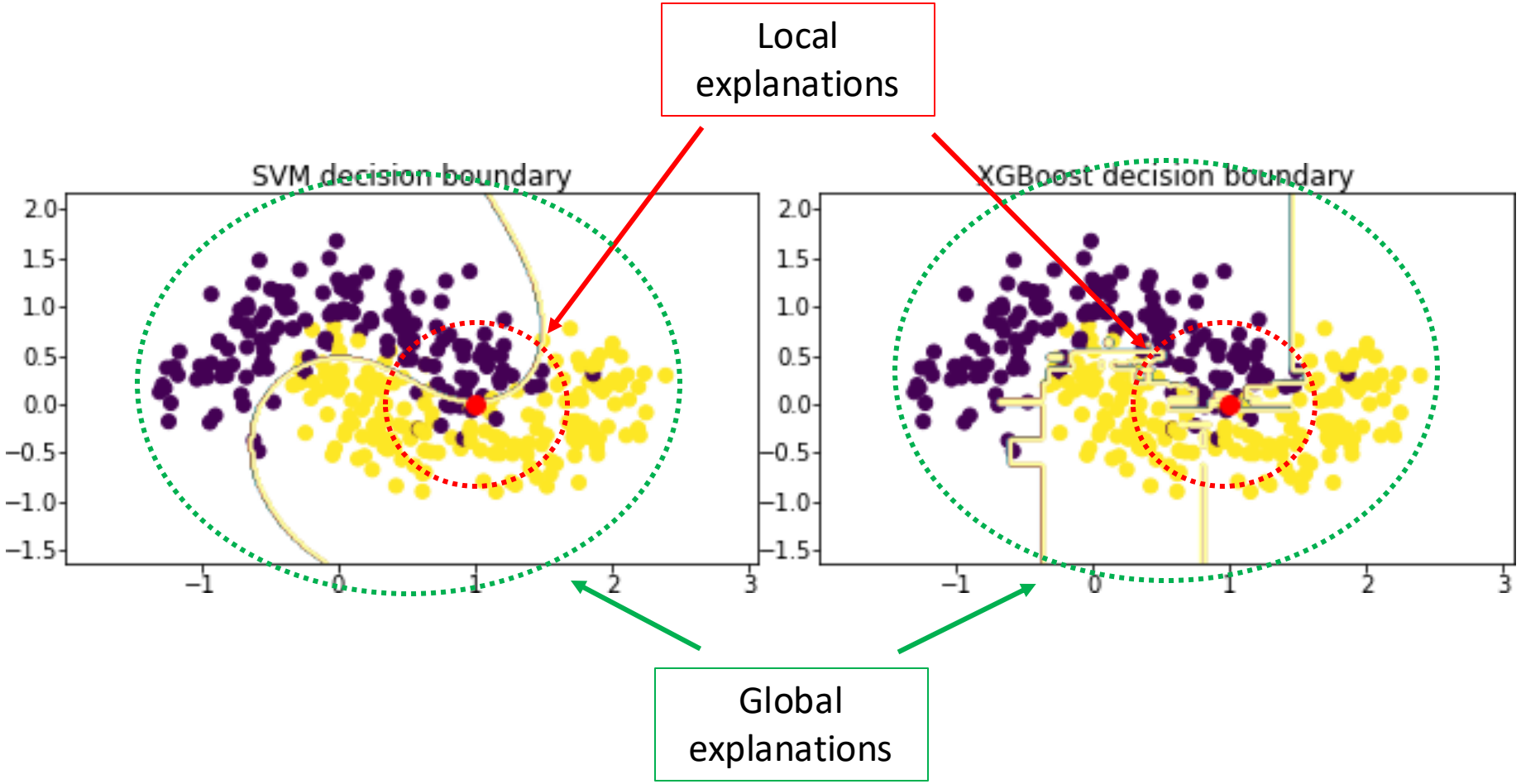


Glassbox



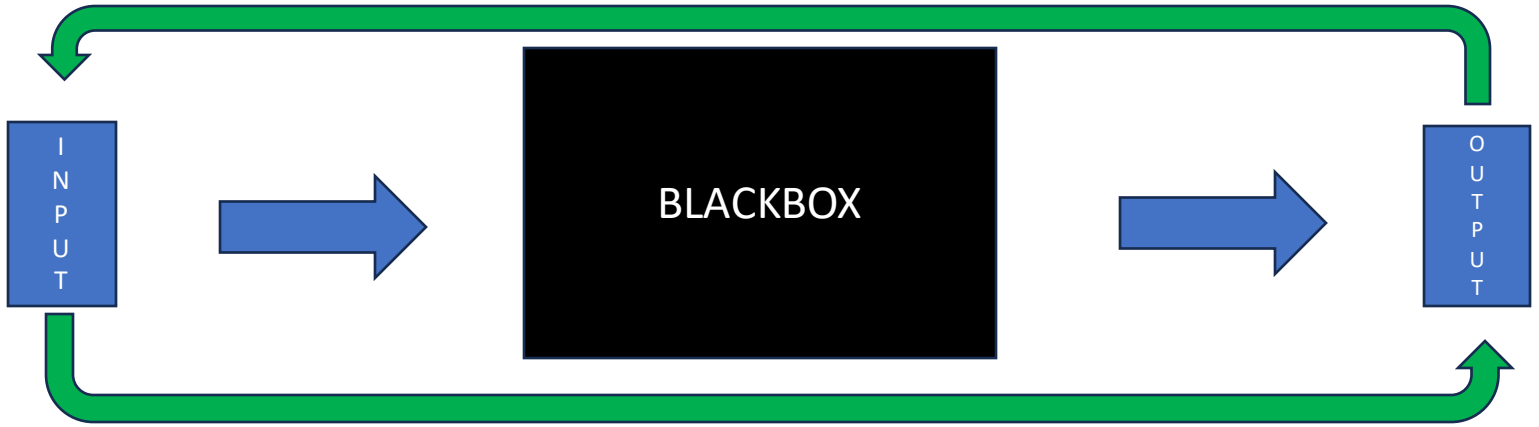


Local vs Global explanations

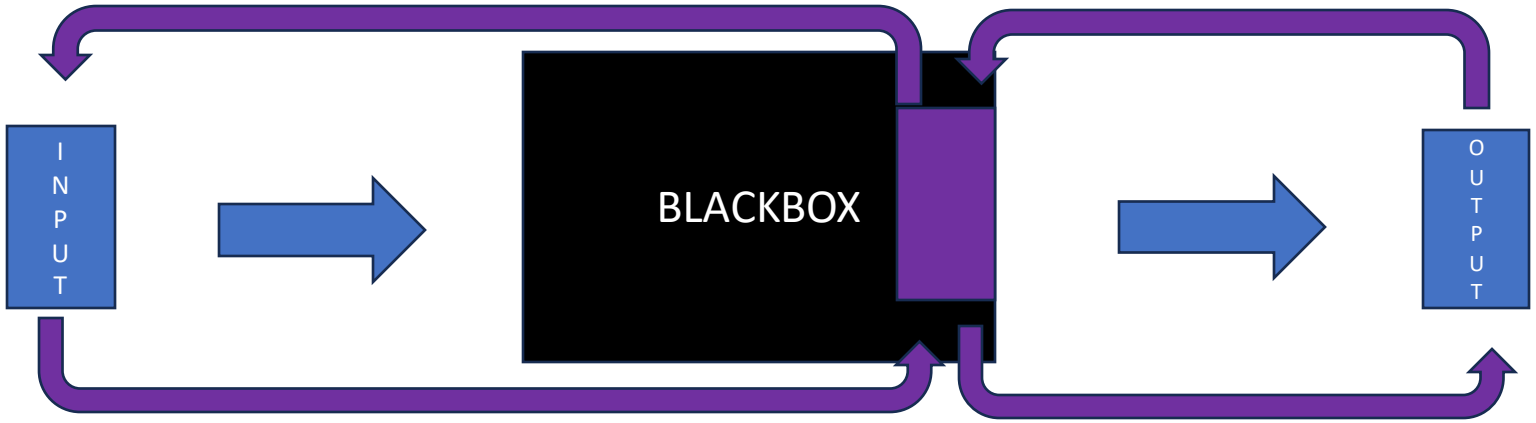


Model-agnostiv vs. Model specific

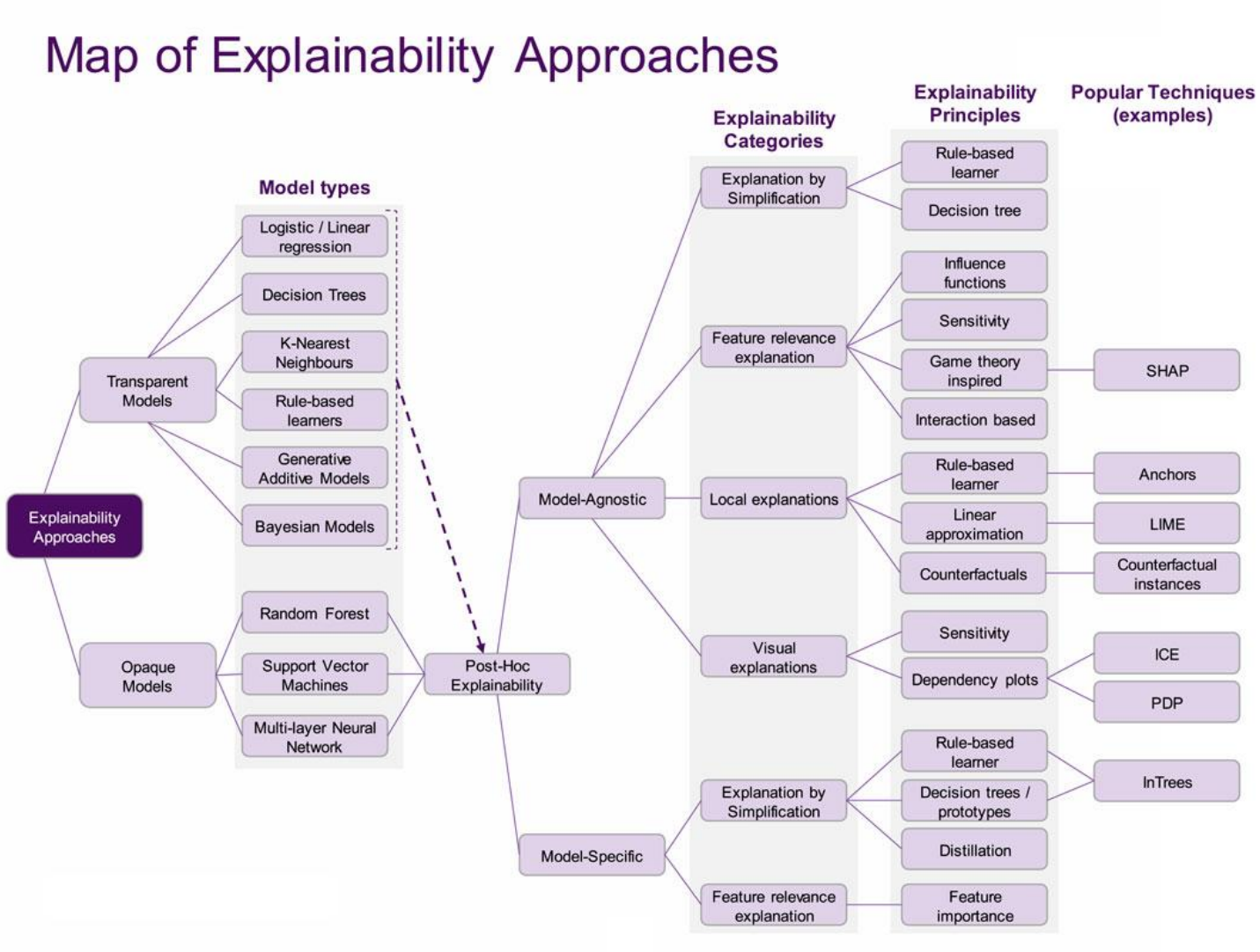
Model-agnostic



Model-specific



Classification of XAI methods



Is XAI a new oil to Artificial Intelligence?

- Or maybe XAI is new 42 of Artificial Intelligence?
- To make use of explanation, you need to understand what question it answers
- There are explanations which are technically correct, but not useful or not understandable to the addressee
- There are explanations that are correct but inconsistent with domain knowledge, or with ML model

Goebel, R. et al. (2018). **Explainable AI: The New 42?**. In: Holzinger, A., Kieseberg, P., Tjoa, A., Weippl, E. (eds) Machine Learning and Knowledge Extraction. CD-MAKE 2018. Lecture Notes in Computer Science(), vol 11015. Springer, Cham. https://doi.org/10.1007/978-3-319-99740-7_21

This is not a new idea

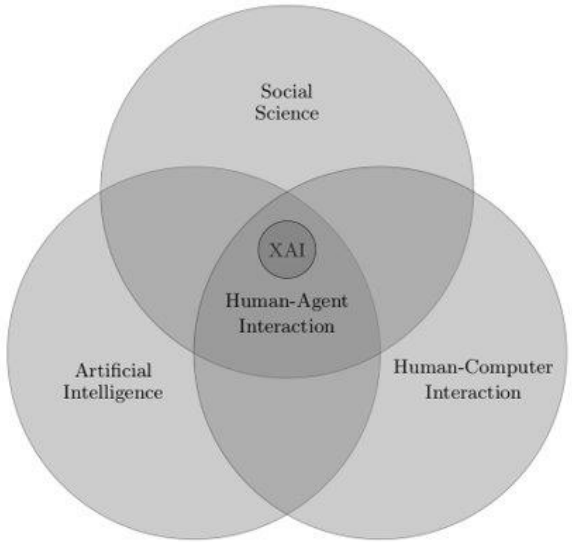
To explain an event is to provide some information about its causal history.
In an act of explaining, someone who is in possession of some information about the causal history of some event - explanatory information, I shall call it - tries to convey it to someone else. - David Lewis

Different approaches

- Intelligibility of the system
- Interpretability of models
- Explainability of ML models

Old topic

- Expert systems
- Recommender systems
- Context-aware systems
- Machine learning



Why XAI is non trivial

*In an **act** of explaining, **someone** who is in possession of **some information***

Artificial intelligence / XAI

Most often feature importance

*about the **causal history of some event** - explanatory information,*

Why input to the model generated such output

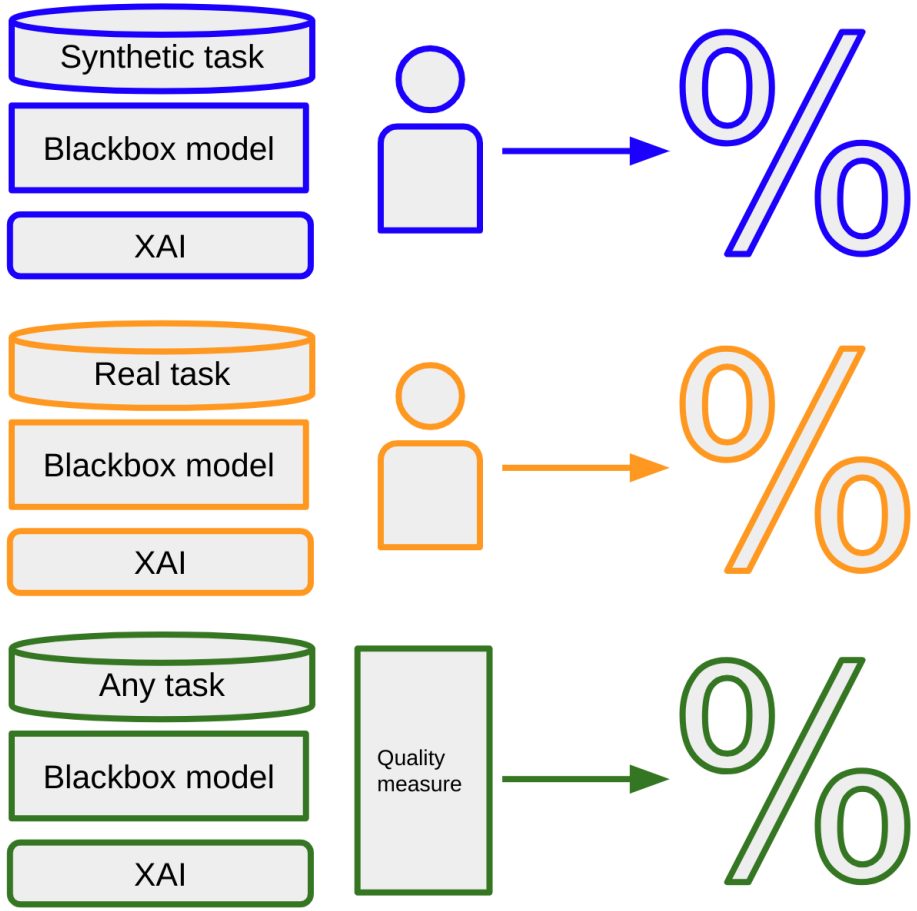
*I shall call it - tries to **convey it to someone else**.*

Human

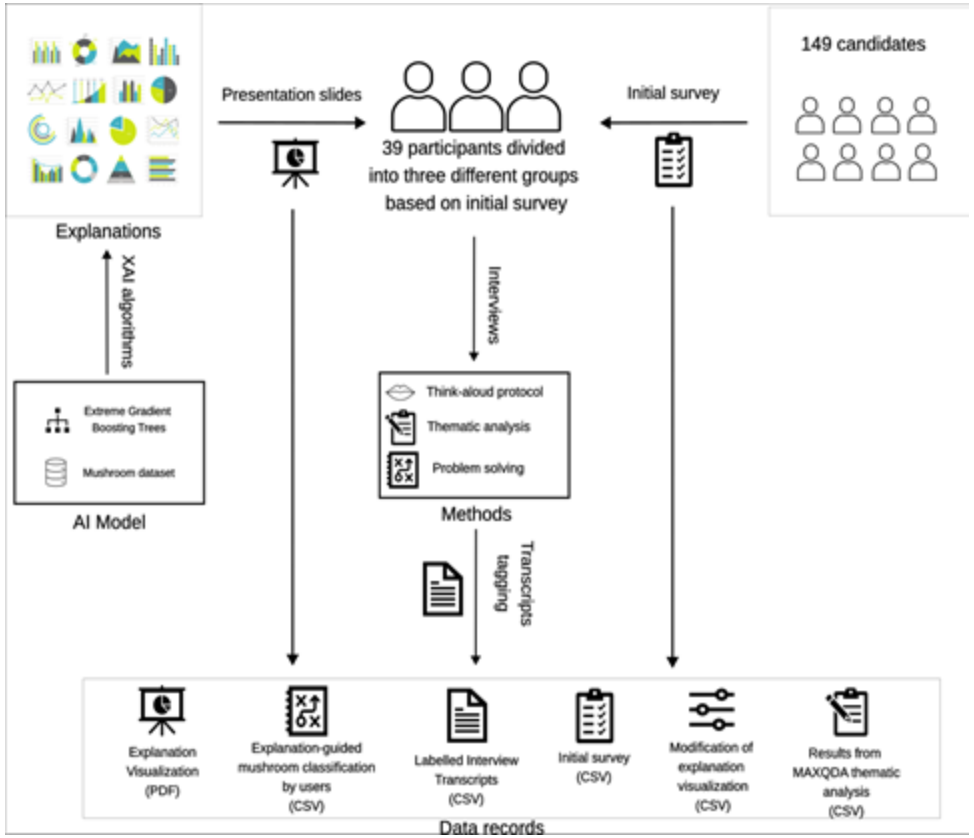
Evaluation of XAI methods

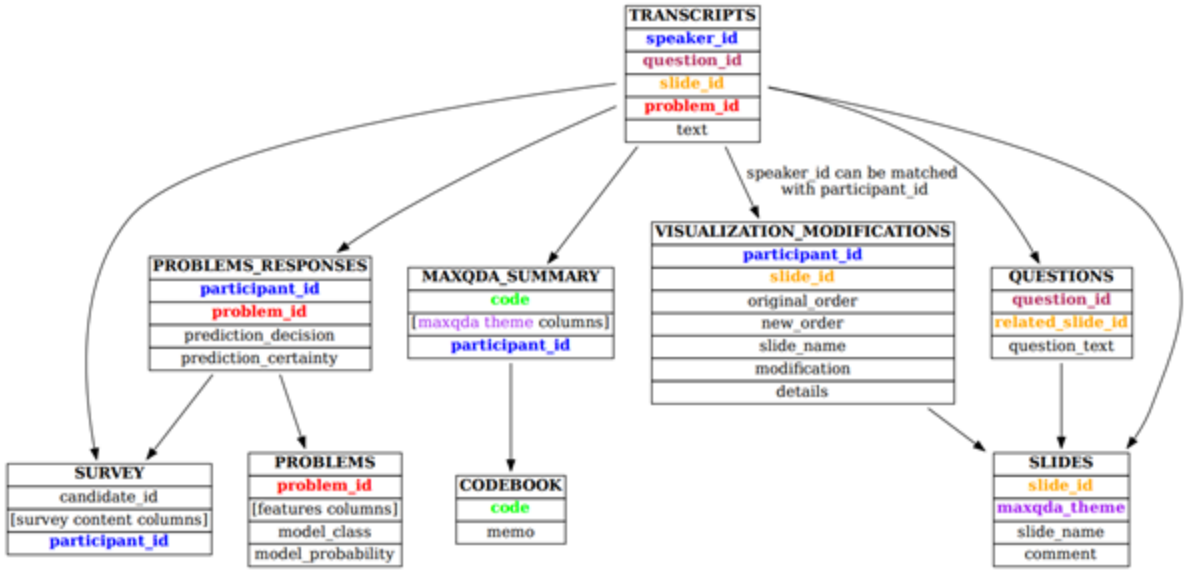
- Types of evaluation approaches
  - **Human-grounded**
  - **Application-grounded**
  - **Functional**
- Popular Quality measures
  - Fidelity (local and global)
  - Stability
  - Consistency
  - Coverage
  - Certainty
  - Representativeness
  - Simplicity/Comprehensability
- Ready to use frameworks
  - Quantus

https://github.com/understandable-machine-intelligence-lab/Quantus



User study on comprehensibility of XAI methods

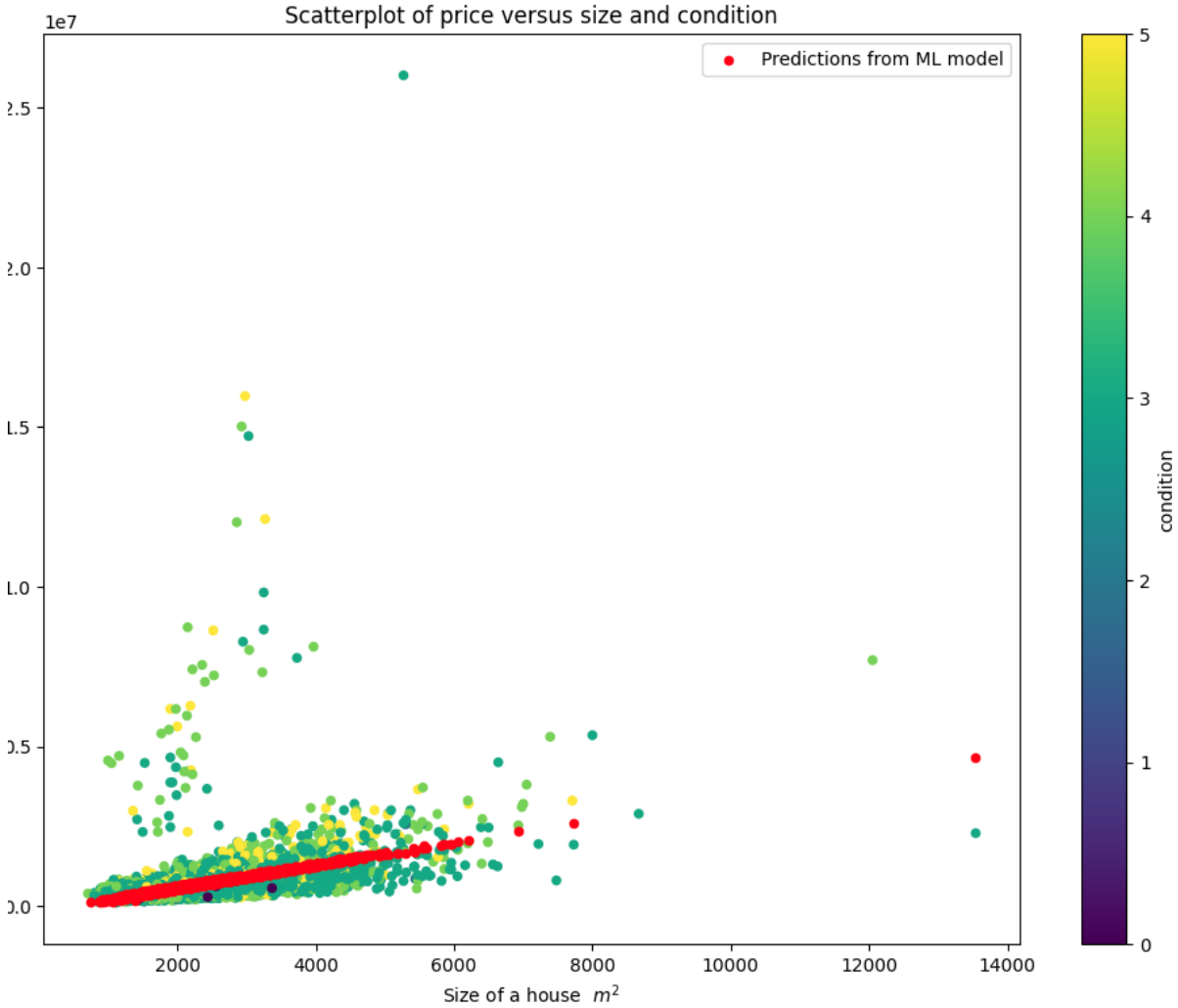




zenodo

10.5281/zenodo.11448395

All in all, everything starts from the data



Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re