DEMYSTIFYING EXPLAINABLE AI (XAI)

Niklas Kasenburg

ALEXANDRA INSTITUTE

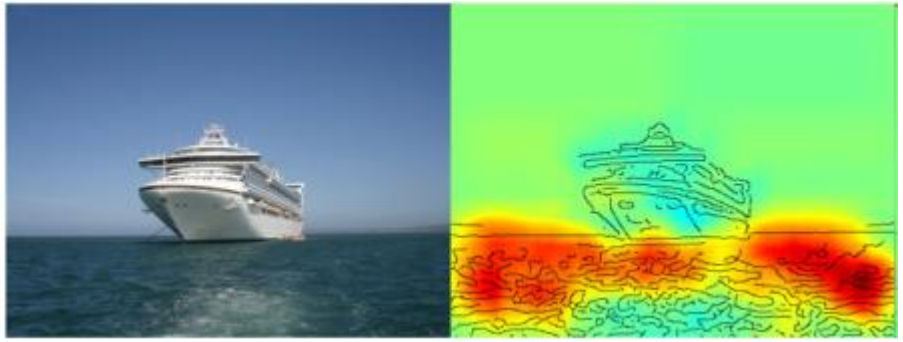
Right to the explanation in the GDPR?

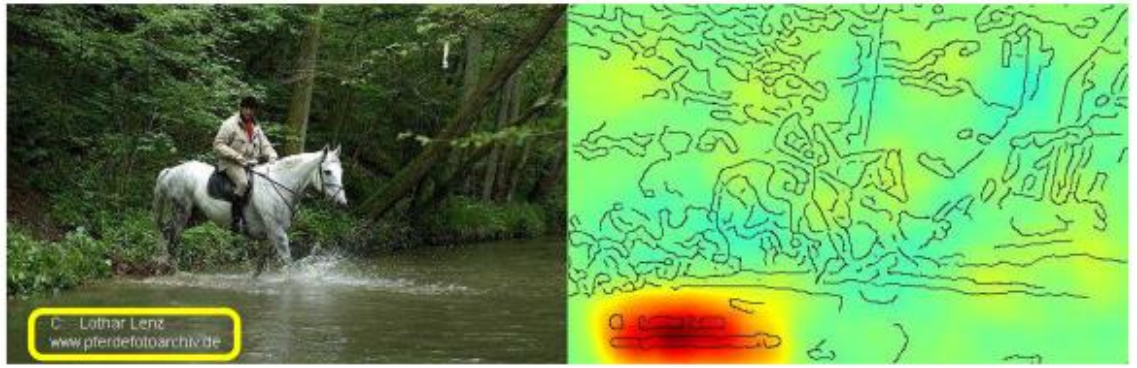
"*the existence of automated decision-making, including profiling, [...] and, at least in those cases,* ***meaningful information*** *about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."* GDPR Art. 15 1. (h)

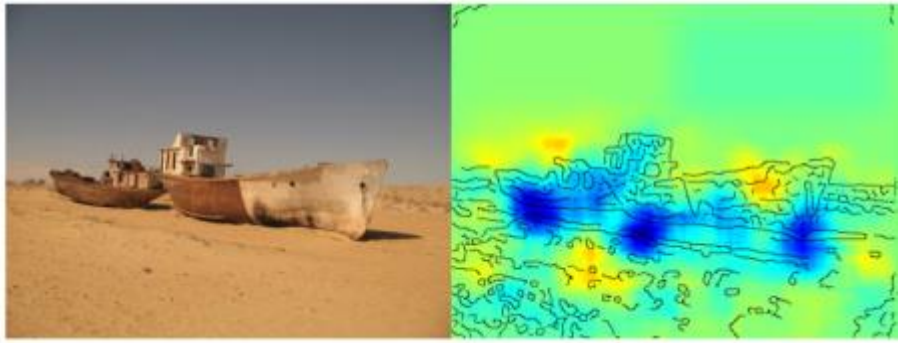
"*the data controller shall implement* ***suitable measures to safeguard*** *the data subject's rights and freedoms and legitimate interests [...]"* GDPR Art. 22 3.

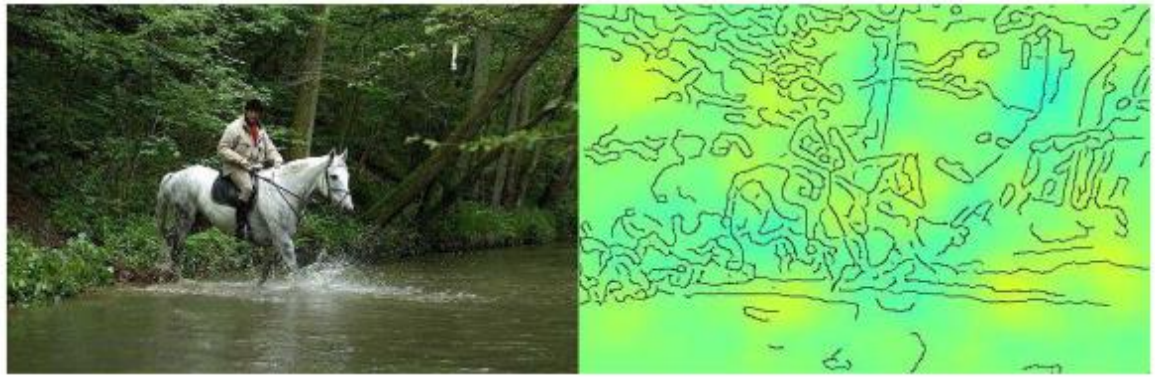
#ForanDigitalt

Clever Hans









Source: S. Lapuschkin, *Unmasking Clever Hans predictors and assessing what machines really learn*, Nature Communications 10 (1096), 2019
License: http://creativecommons.org/licenses/by/4.0/

#ForanDigitalt

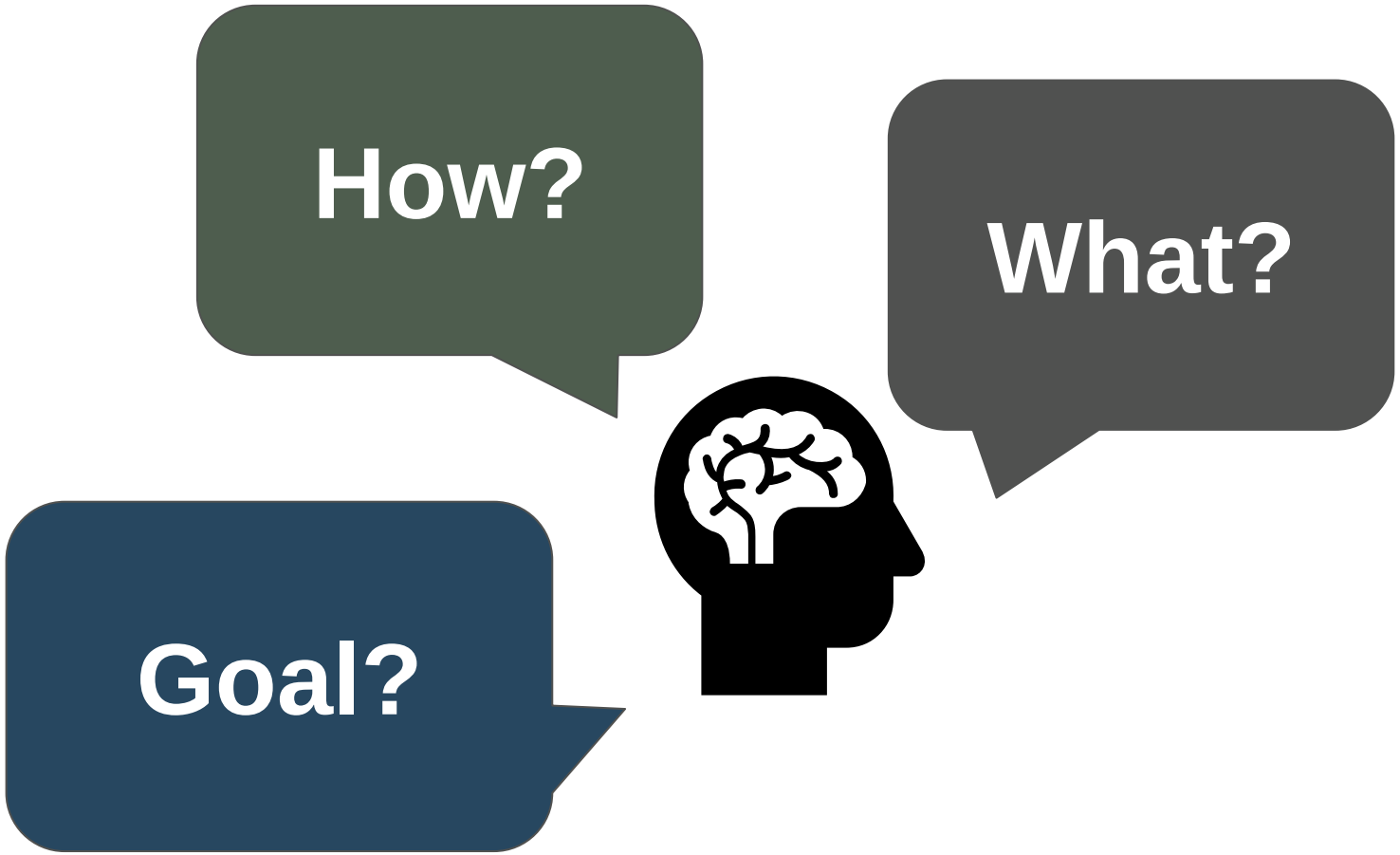
Pneumonia risk

- Model predicting risk od dying of pneumonia
- Model predicts lower risk when subject has Asthma
- Reason: Patient in training data with asthma had better medical care
- Reality: Asthma patients are a critical risk group for pneumonia

Source: R. Caruana et al., Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, In Proceedings of the 21th ACM SIGKDD, pp. 1721-1730, 2015

#ForanDigitalt

Outline



#ForanDigitalt

What is Explainable AI (XAI)?

Not only machine learning

Explainable AI: Beware of Inmates Running the Asylum

Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences

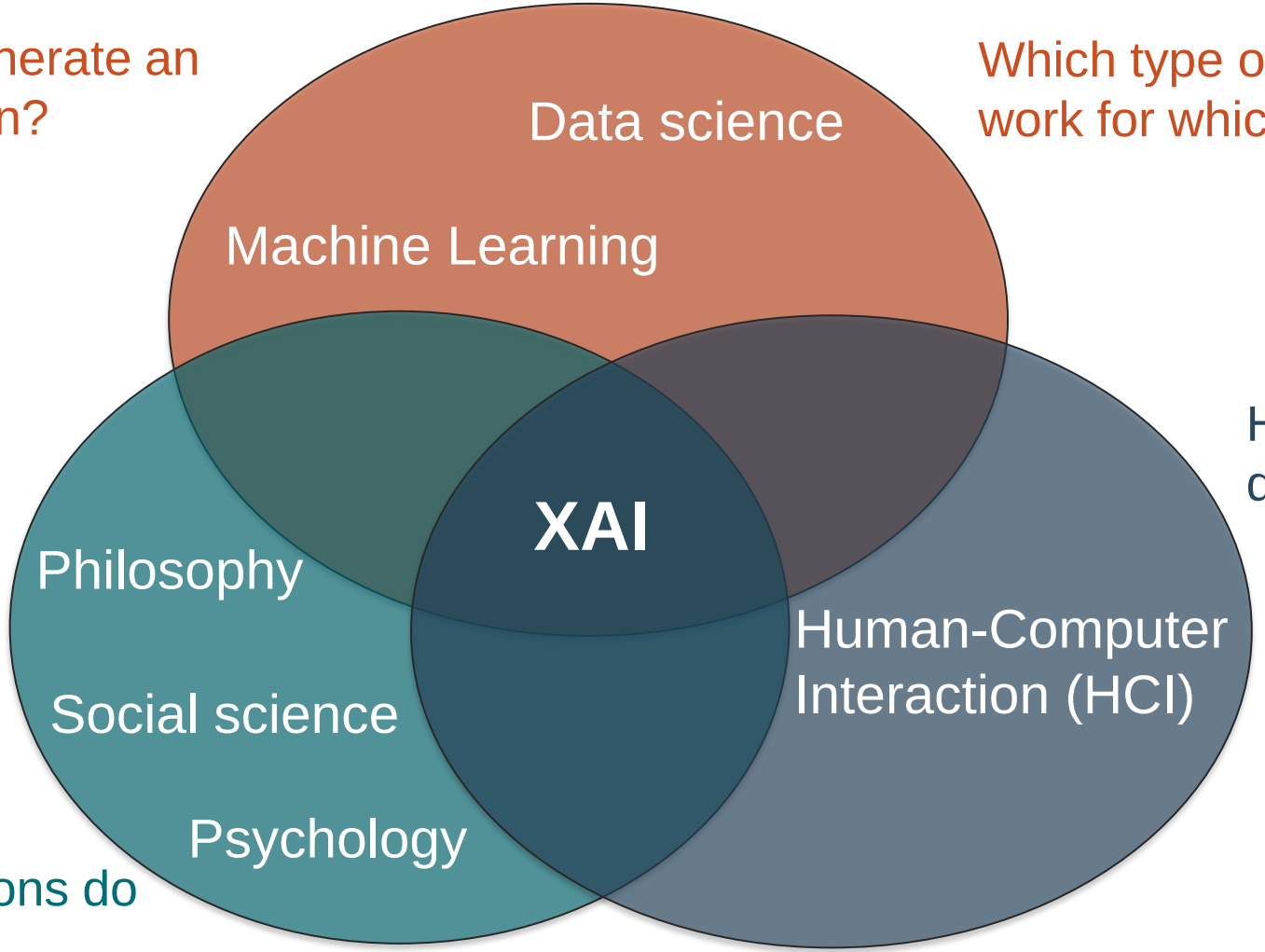
Tim Miller* and Piers Howe† and Liz Sonenberg*

#ForanDigitalt

Not only machine learning

How to generate an explanation?

Which type of explanations work for which type of model?



Which errors do humans make when they decide?

How to support a decision?

Which type of explanations do humans use?

How to present an explanation?

Which type of explanations do humans understand?

#ForanDigitalt

Definition

explain verb

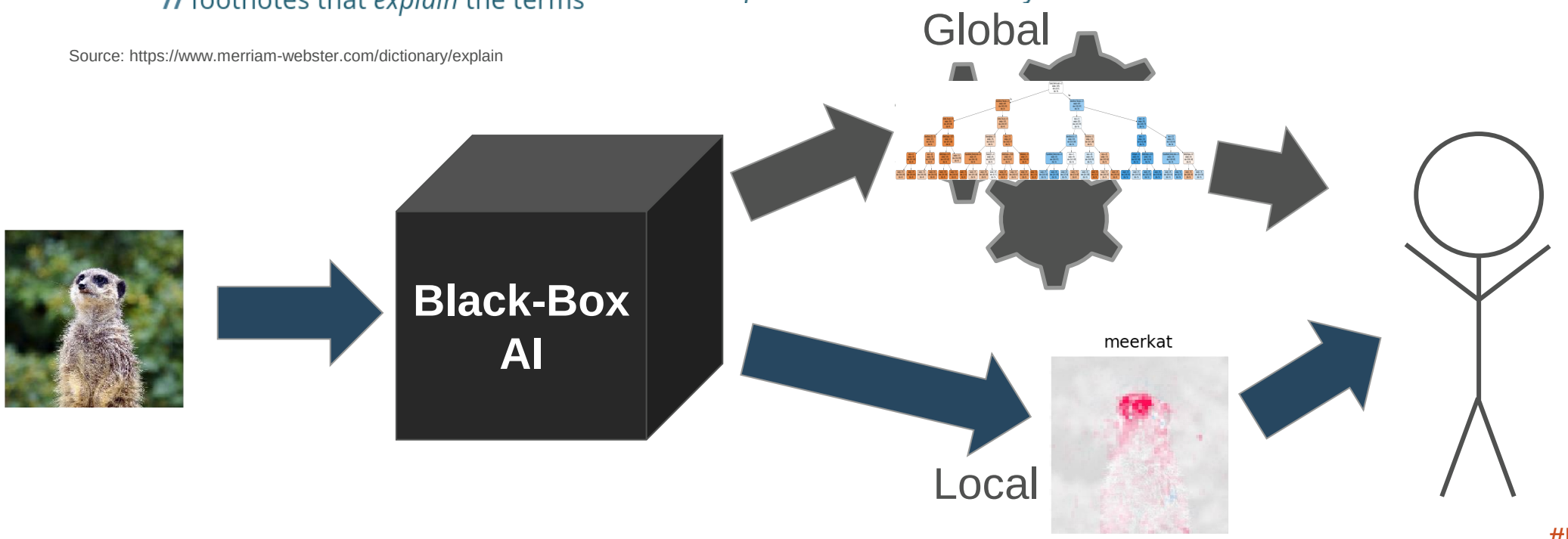
1 **a** : to make known
// explain the secret of your success

b : to make plain or understandable
// footnotes that explain the terms

2 : to give the reason for or cause of
// unable to explain his strange conduct

3 : to show the logical development or relationships of
// explained the new theory

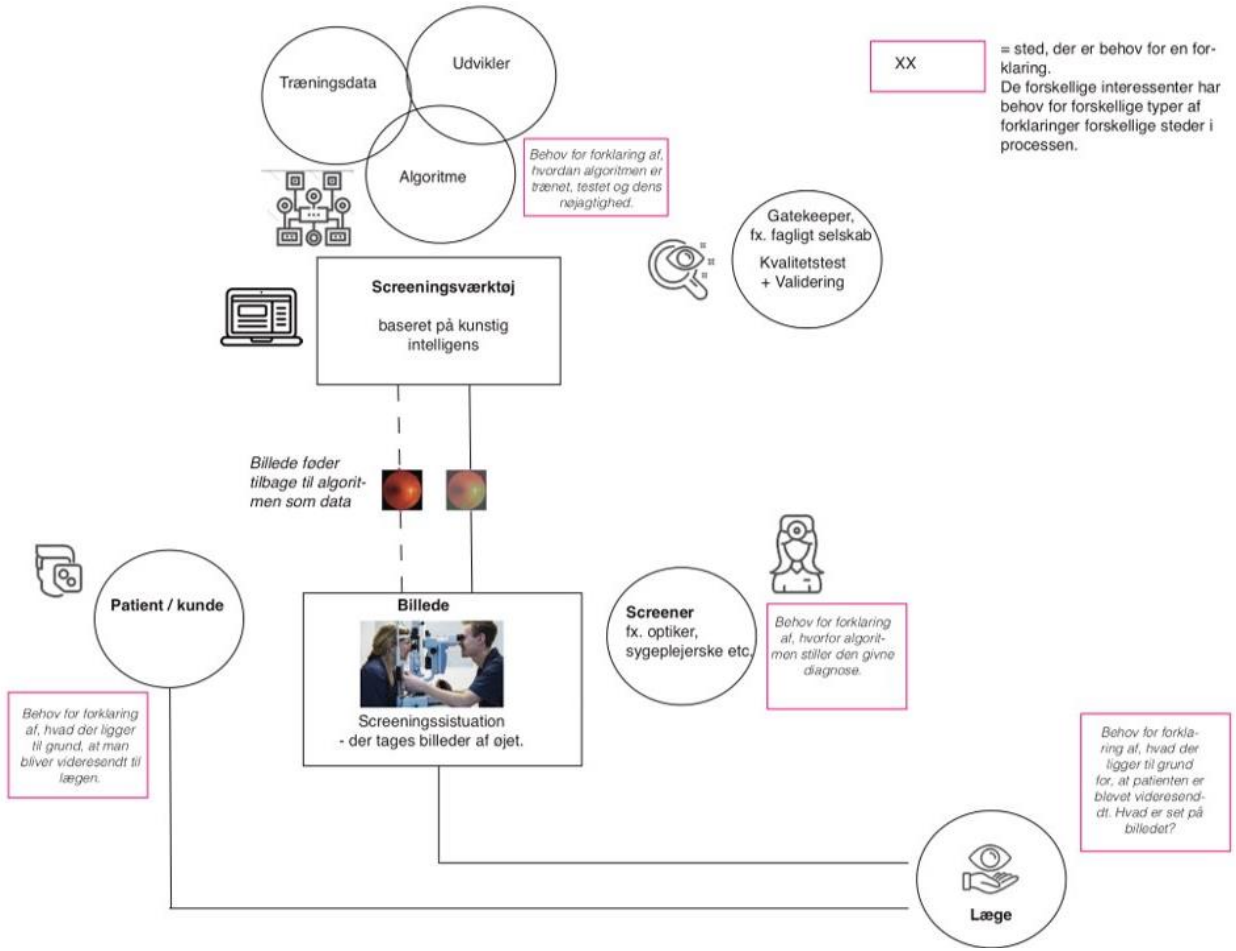
Source: https://www.merriam-webster.com/dictionary/explain



Source: https://github.com/slundberg/shap

#ForanDigitalt

Not only machine learning



#ForanDigitalt

What is the goal of using XAI?

Different users – different goals

ML engineer / developer (expert user):

- Test and improve system (debugging)
- Strengths and limits
- Behaviour in different scenarios

Source: https://xkcd.com/1838/ (https://creativecommons.org/licenses/by-nc/2.5/)

Domain	Model Purpose	Technique	Stakeholders
Finance	Loan Repayment	Feature Importance	Loan Officers
Insurance	Risk Assessment	Feature Importance	Risk Analysis
Content Moderation	Malicious Reviews	Feature Importance	Content Moderators
Finance	Cash Distribution	Feature Importance	ML Engineers
Facial Recognition	Smile Detection	Feature Importance	ML Engineers
Content Moderation	Sentiment Analysis	Feature Importance	QA ML Engineers
Healthcare	Medical Access	Counterfactuals	ML Engineers
Content Moderation	Object Detection	Adversarial Perturbation	QA ML Engineers

Source: U. Bhatt et al., *Explainable machine learning in deployment,* In Proceedings FAT* '20, 2020 [Table 1]

#ForanDigitalt

XAI as quality assurance!

Source: S. Lapuschkin *Unmasking Clever Hans predictors and assessing what machines really learn*, Nature Communications 10 (1096), 2019
License: http://creativecommons.org/licenses/by/4.0/

ALEXANDRA INSTITUTE

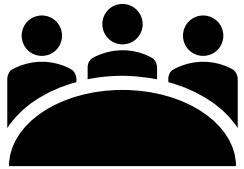
Different users – different goals

Owner and end-user:

- Trust in the system – also in unexpected scenarios
- Reasoning / decision
- Explanation of results that are opposite to expectations
- Is it responsible to use the system?

Stakeholder:

- Ethical or legal requirements
- Compliance



#ForanDigitalt

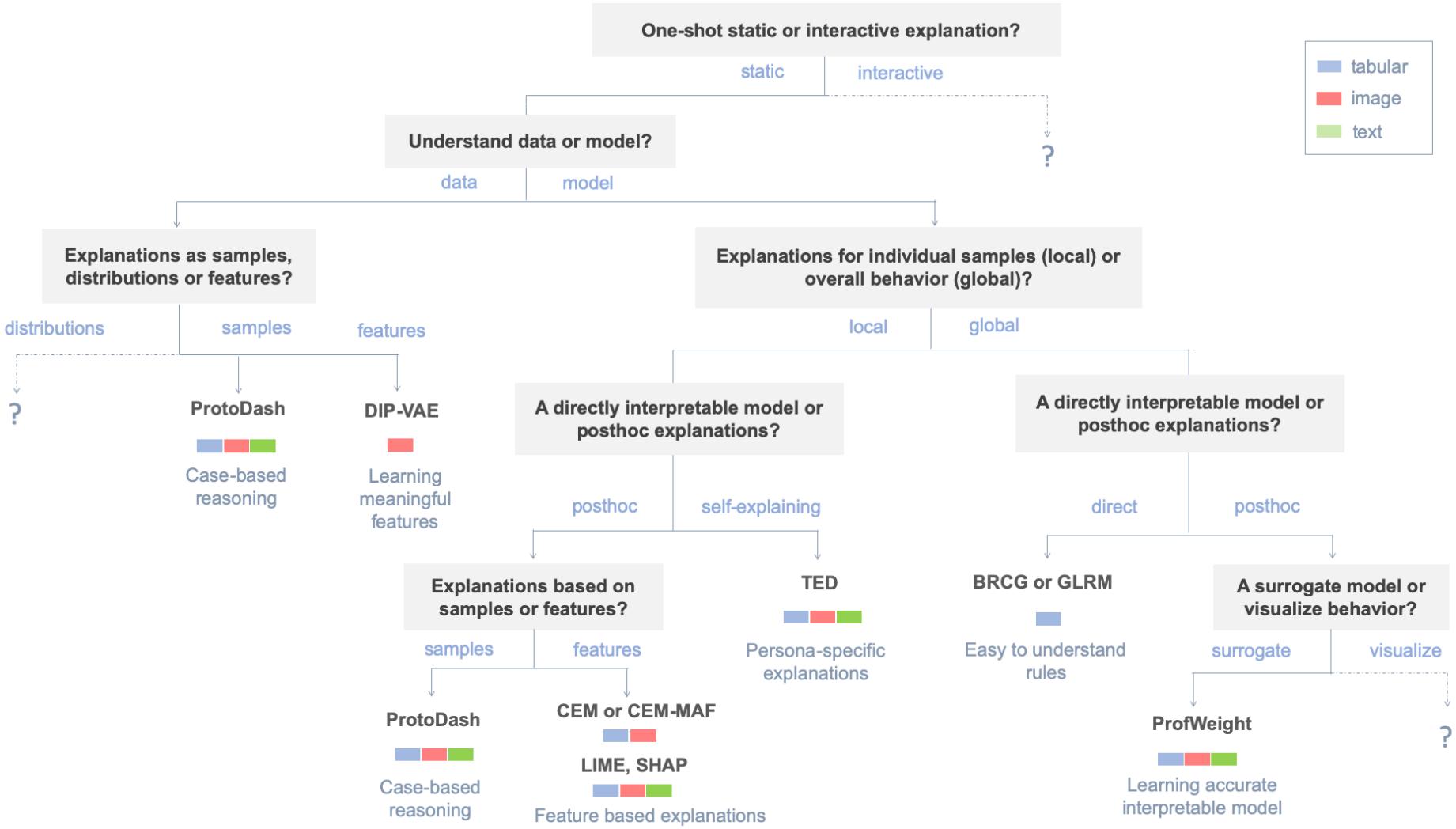
Other goals for using XAI

1. Verification: Is the result generated based on known hypotheses?
2. Additional information:
   - Result is one of many factors in a decision process
   - Explanations can "fix" wrong decisions
   - Result needs to be explained further (doctor – patient, planer – service employee, technician – business developer)
3. Simplification of the model

#ForanDigitalt

How to create explainable solutions?

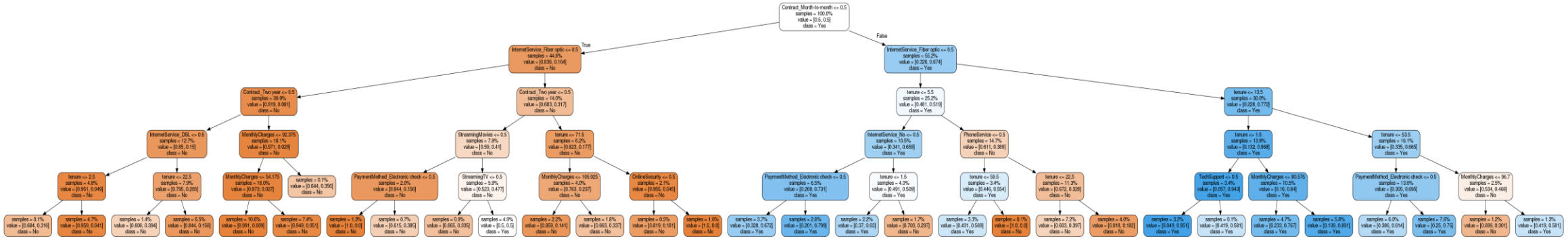
Method taxonomy



Source: https://github.com/IBM/AIX360/blob/master/aix360/algorithms/README.md

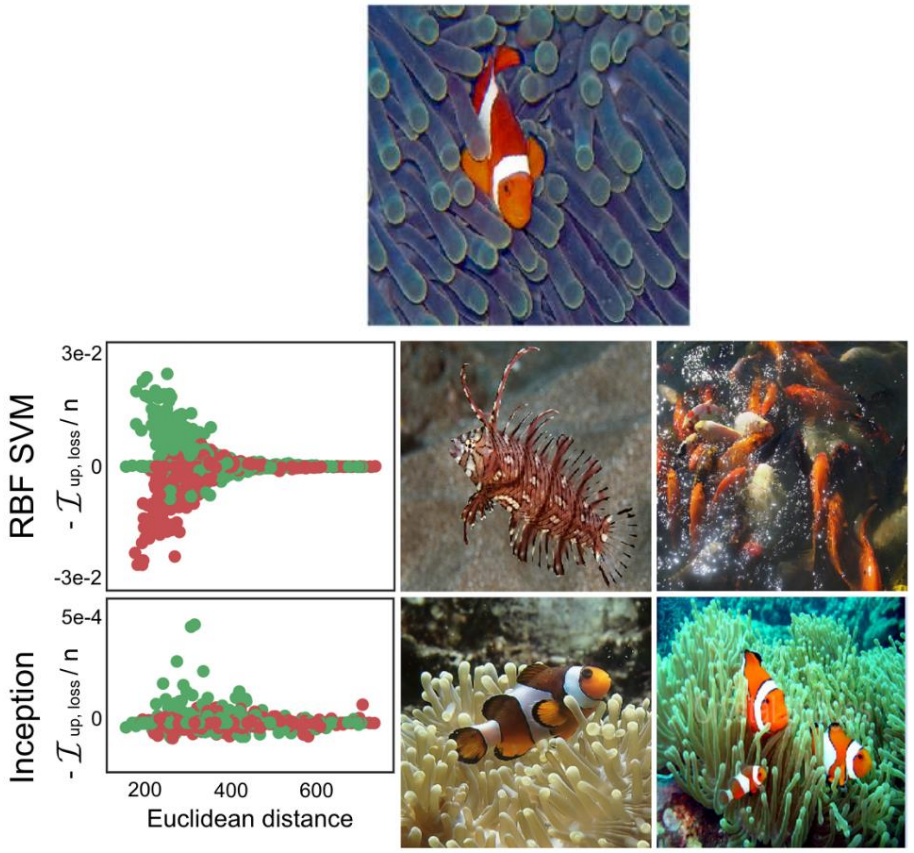
#ForanDigitalt

Rules – intrinsic local explanations



#ForanDigitalt

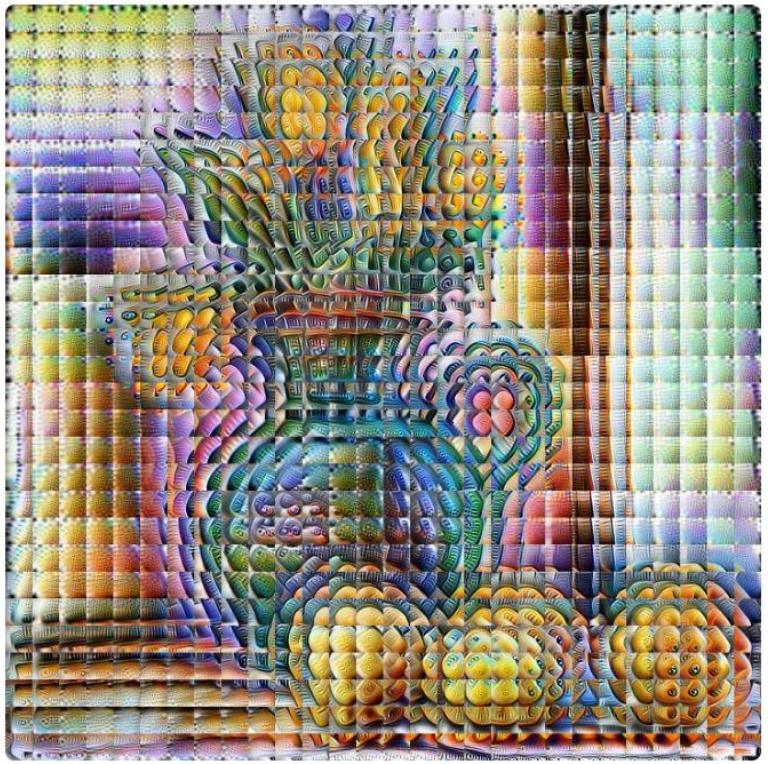
Prototypes and examples

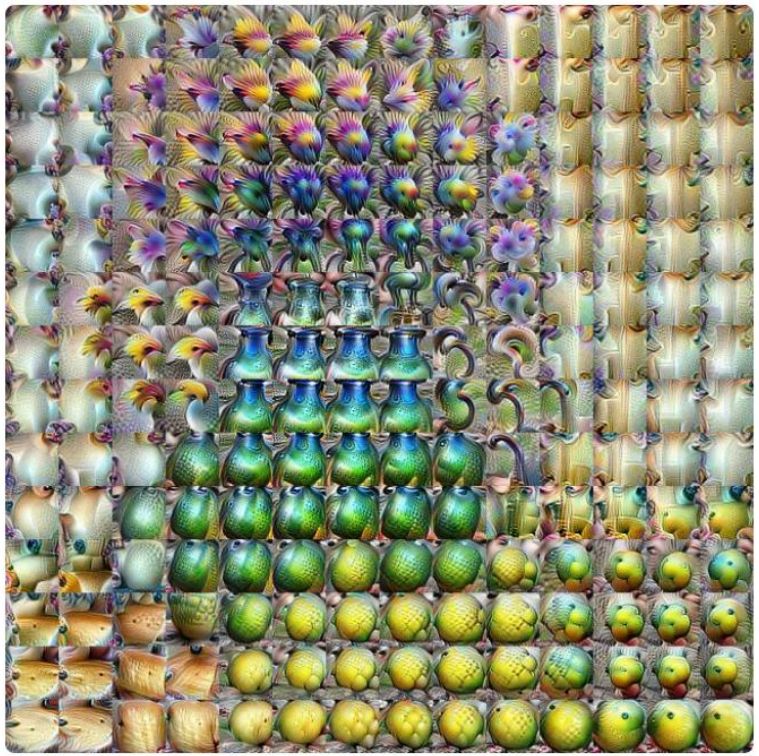


Source: P.W. Koh and P. Liang, *Understanding Black-box Predictions vis Influence Functions*, arXiv1703.04730v2 [stat.ML], July 2017

#ForanDigitalt

Learned representation

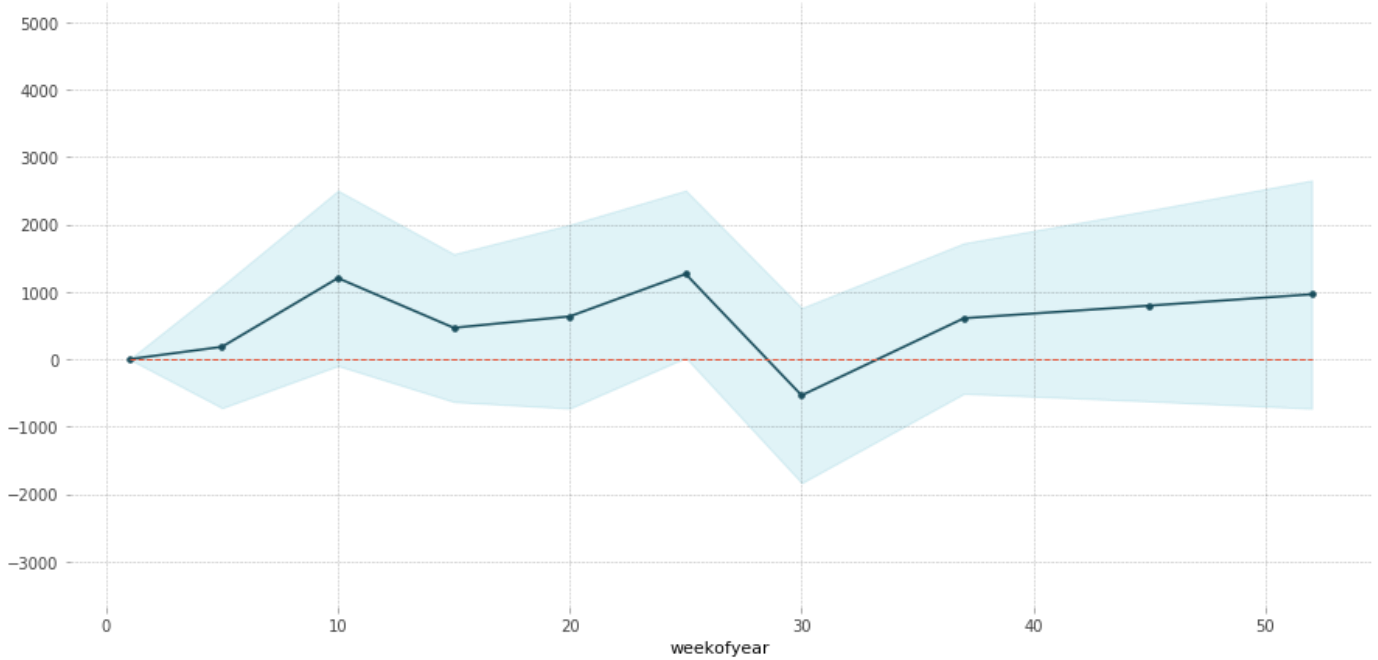




Source: Olah, et al., *The Building Blocks of Interpretability*, Distill, 2018.

#ForanDigitalt

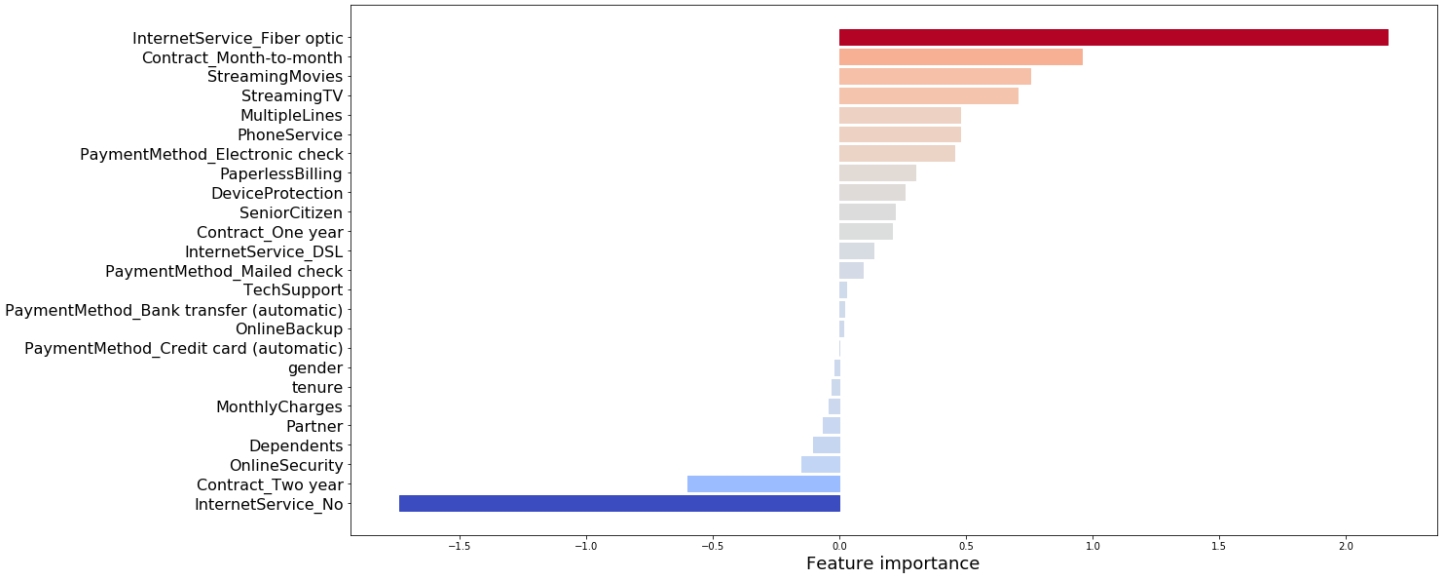
Partial dependence plot – global post-hoc explanations

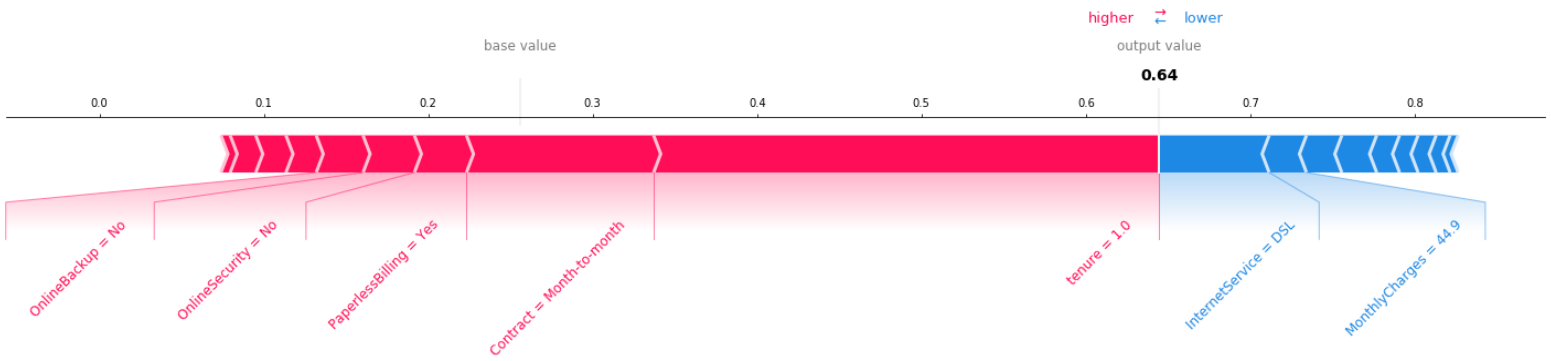


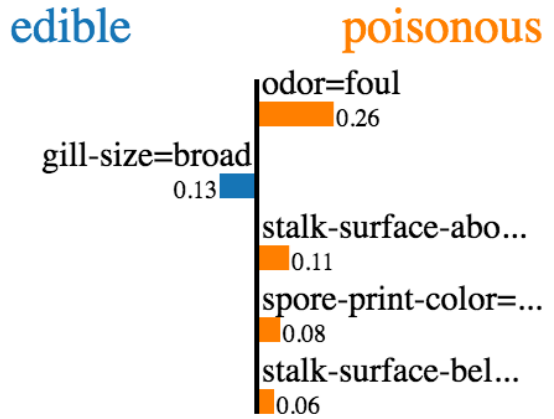
Source: https://github.com/SauceCat/PDPbox/blob/master/tutorials/pdpbox_regression.ipynb

#ForanDigitalt

Feature attribution (importance)



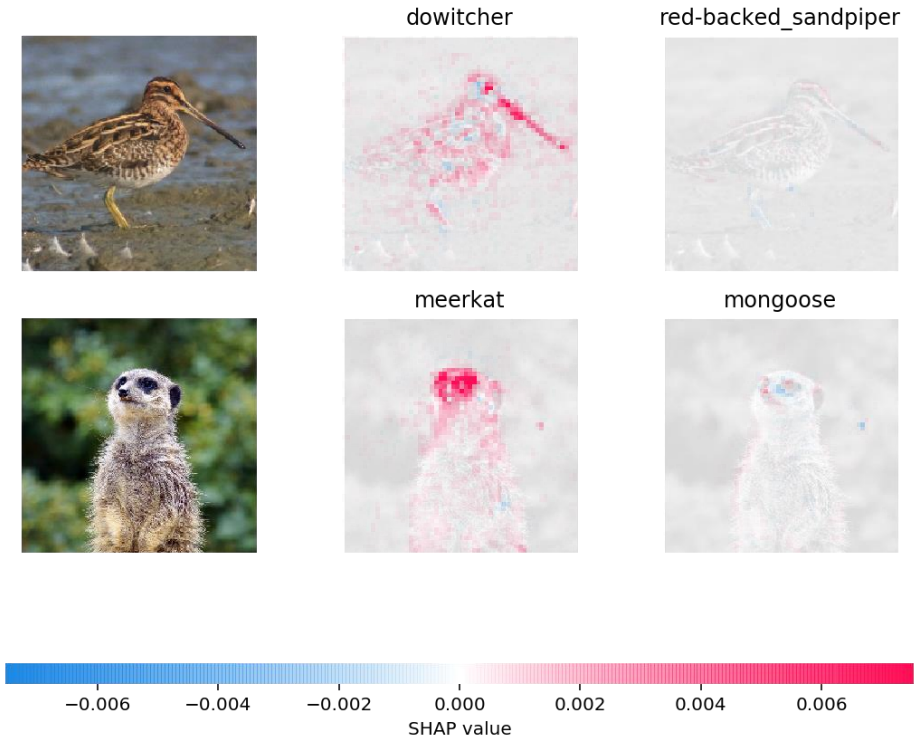




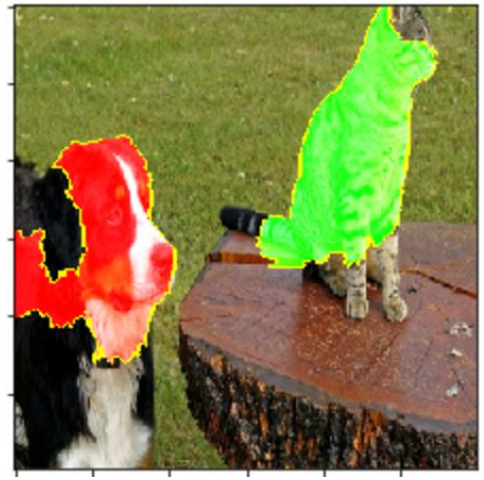
Source: https://github.com/marcotcr/lime

#ForanDigitalt

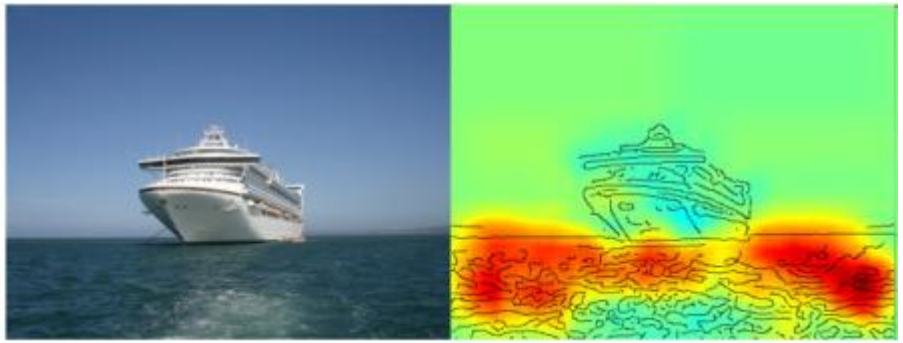
Feature attribution – images



Source: https://github.com/slundberg/shap



Source: https://github.com/marcotcr/lime



Source: S. Lapuschkin, *Unmasking Clever Hans predictors and assessing what machines really learn*, Nature Communications 10 (1096), 2019

#ForanDigitalt

Feature attribution – text

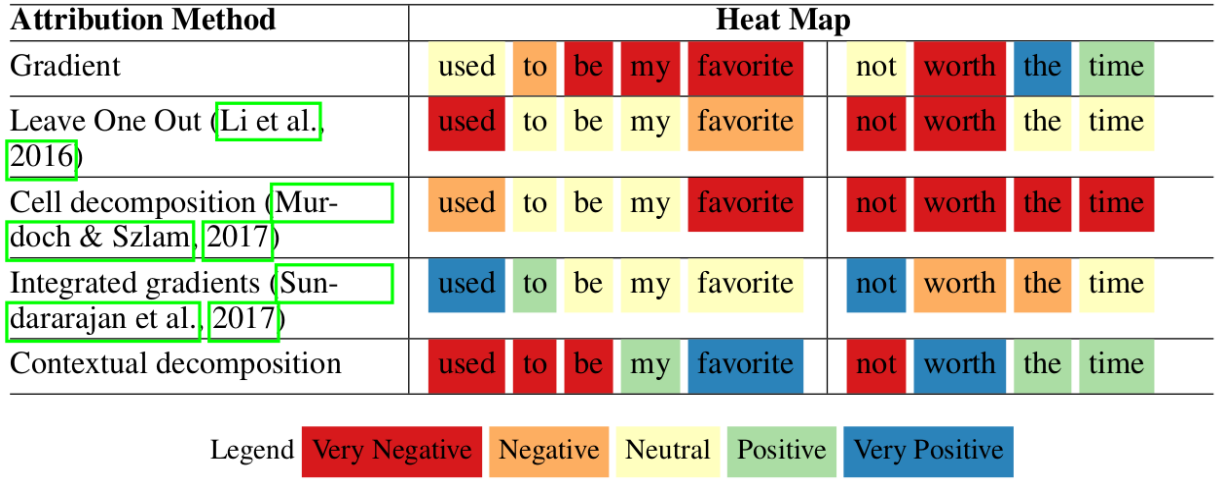
a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate ... ... m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Source: T. Lei et al., *Rationalizing Neural Predictions*, arXiv:1606.04155v2 [cs.CL], November 2016

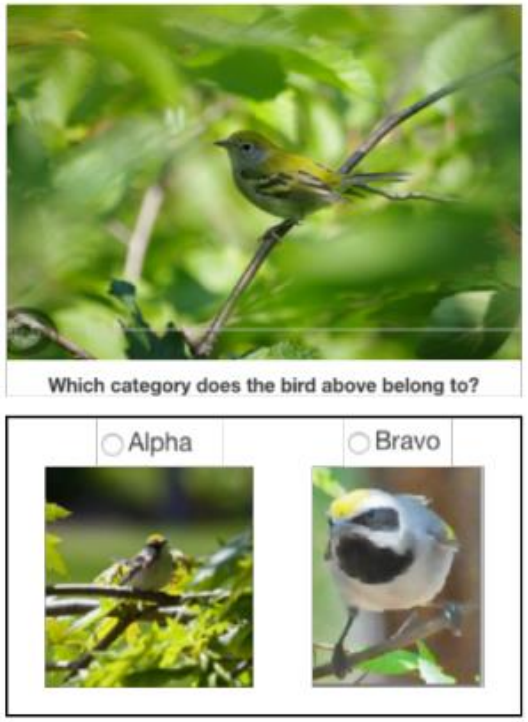


Source: W. J. Murdoch et al., *Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs*, arXiv:1801.05453v2 [cs.CL], April 2018

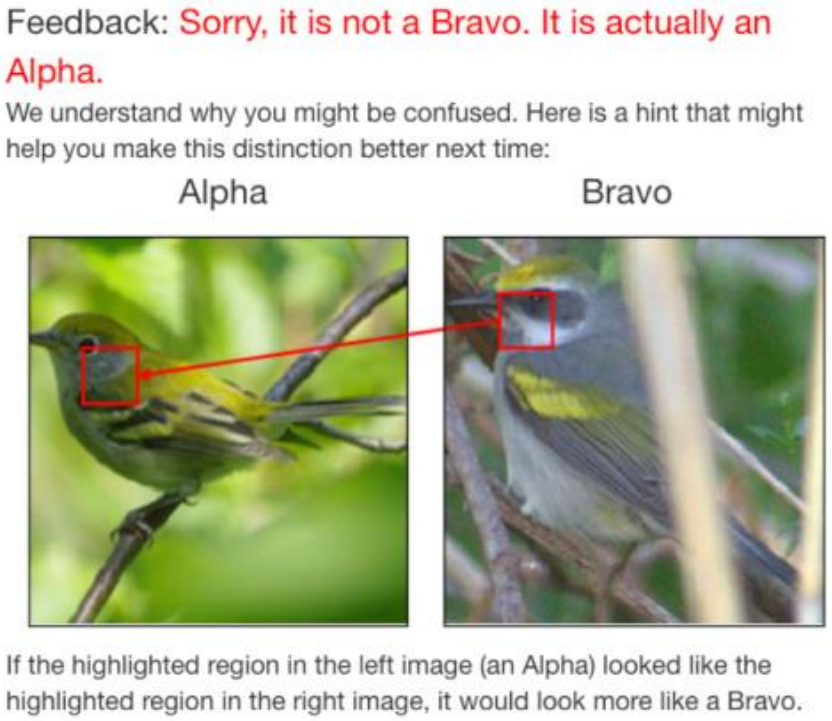
#ForanDigitalt

Application example

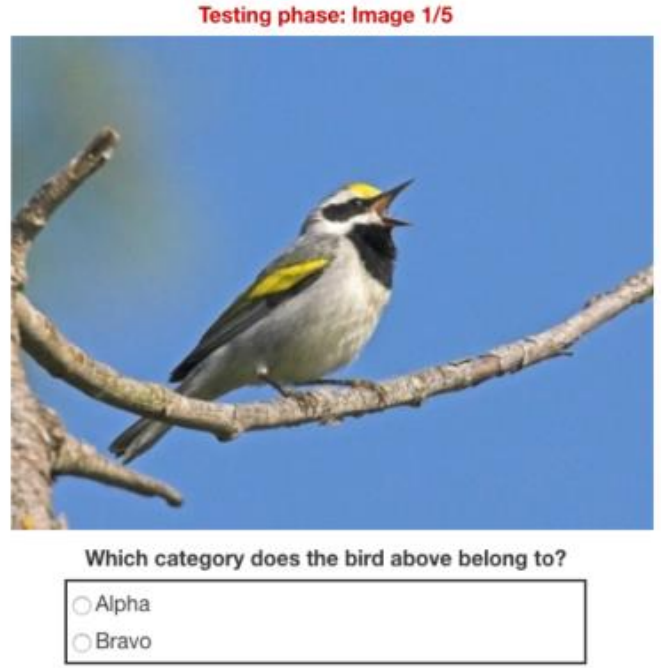
Learning tool



(a) Training Interface



(b) Feedback



(c) Testing Interface

Source: Y. Goyal et al., *Counterfactual Visual Explanations*, arXiv:1904.07451v2, June 2019

#ForanDigitalt

Is that enough?

"the existence of automated decision-making, including profiling, [...] and, at least in those cases, ***meaningful information*** *about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject."* GDPR Art. 15 1. (h)

ALEXANDRA INSTITUTE

“Transparency” in whole life-cycle

- Transparency about the:
  - Objective
  - Training process and data
  - Model and parameter choices
  - Testing and calibration
  - Monitoring
- All stakeholders (owner, data scientist, auditor) should be involved
- See also: https://medium.com/@szymielewicz/black-boxed-politics-cebc0d5a54ad

#ForanDigitalt

Thank you for your time

niklas.kasenburg@alexandra.dk

+45 93 50 85 40

https://www.linkedin.com/in/niklas-kasenburg-7428ba91/

ALEXANDRA INSTITUTE