clearbox AI

Bolzano, 2023

EURAC workshop
29-9-2023

humans && machines

About me

I have a background in applied sciences, my PhD dissertation was about uncertainty quantification in simulations. Before Clearbox I worked several years as a scientific software developer.

I’m one of the 4 co-founders and CTO of Clearbox AI, a startup hosted by Politecnico di Torino incubator.

During my free time I do enjoy hiking and I’m currently experimenting with hydroponic farming.

About me

I have a background in applied sciences, my PhD dissertation was about uncertainty quantification in simulations. Before Clearbox I worked several years as a scientific software developer.

I'm one of the 4 co-founders and CTO of Clearbox AI, a startup hosted by Politecnico di Torino incubator.

During my free time I do enjoy hiking and I'm currently experimenting with hydroponic farming.

Susa valley

Who is Clearbox AI

We are a group of innovators with **R&D backbones** working in the MLOps field, in particular on using **synthetic data** for ML models improvement and testing.

International advisor: Mariarosaria Taddeo (Oxford Internet Institute)

Trusted by

Today's plan

We will discuss about **Trustworthy AI** from a software engineering perspective. I'll try to share my experience as a former researcher turned into an ML developer.

Useful material:

- 'ML engineering' by Andriy Burkov
- Martin Fowler's blog (www.martifowler.com)
- 'Interpretable Machine Learning' by Christophe Molnar

The AI Act

New regulation defining what can and cannot be done with AI systems.

Definition of different risk levels associated with AI systems, requiring **conformity assessments** for systems associated with high risks.

EU Definition of Trustworthy AI

1. it should be **lawful**, complying with all applicable laws and regulations.

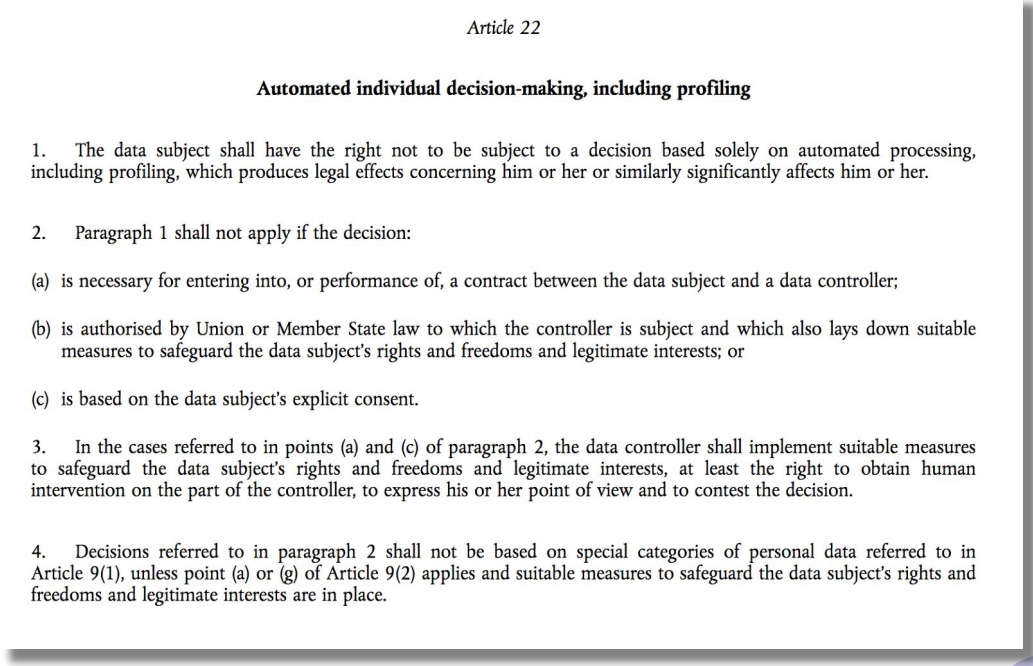
2. it should be **ethical**, ensuring adherence to ethical principles and values.

3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

AI regulation

AI regulation is currently being drafted by the European Commission, first version published last year.

However, AI is already partly regulated under **GDPR**.



Article 22

Automated individual decision-making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

2. Paragraph 1 shall not apply if the decision:

(a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;

(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or

(c) is based on the data subject's explicit consent.

3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.

4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) applies and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Deliveroo fined for improper use of ML models

HUNTON ANDREWS KURTH

HOME ABOUT CONTACT PUBLICATIONS

Skip to content

PRIVACY & INFORMATION SECURITY LAW BLOG

Global Privacy and Cybersecurity Law Updates and Analysis

Home » Italian Garante Fines Deliveroo 2.5M Euros For Unlawful Processing Of Personal Data

Italian Garante Fines Deliveroo 2.5M Euros for Unlawful Processing of Personal Data

Posted on August 5, 2021

POSTED IN ENFORCEMENT, EUROPEAN UNION, INTERNATIONAL

On August 2, 2021, the Italian Data Protection Authority (*Garante per la protezione dei dati personali*, "Garante") announced that it had levied a €2,500,000 fine on Deliveroo Italy s.r.l. for the unlawful processing of personal data of approximately 8,000 Deliveroo riders, and various infringements of the EU Genera Data Protection Regulation (the "GDPR").

Following an investigation into Deliveroo's practices, the Garante found that Deliveroo had failed to provide transparent information to its riders about the algorithm used to manage riders' work shifts. In addition, the Garante found that Deliveroo's app collected a disproportionate amount of riders' personal data in

Search...

Stay Connected

SUBSCRIBE

Topics

Select Category

Tags

Select Tag

Subscribe for free

protocol

NEWSLETTERS WORKPLACE ENTERPRISE CHINA FINTECH POLICY MANUALS BRAINTRUST POWER INDEX | SIGN IN

BULLETINS

New York City passed a bill requiring 'bias audits' of AI hiring tech

If signed into law, it will require providers of automated employment decision tools to have those systems evaluated each year by an audit service and provide the results to companies using those systems.

https://www.protocol.com/bulletins/nyc-ai-hiring-tools

Bias e Fairness

Apple's 'sexist' credit card investigated by US regulator

11 November 2019 Share

Steve Wozniak ✓
@stevewoz

Replying to @dhh

The same thing happened to us. We have no separate bank accounts or credit cards or assets of any kind. We both have the same high limits on our cards, including our AmEx Centurion card. But 10x on the Apple Card.

7:58 AM · Nov 10, 2019 · Twitter Web App

DHH ✓
@dhh

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

9:34 PM · Nov 7, 2019 · Twitter for iPhone

12.8K Retweets **28.6K** Likes

DHH ✓ @dhh · Nov 7, 2019
Replying to @dhh
I'm surprised that they even let her apply for a card without the signed approval of her spouse? I mean, can you really trust women with a credit card these days??!

86 270 4.4K

DHH ✓ @dhh · Nov 7, 2019
It gets even worse. Even when she pays off her ridiculously low limit in full, the card won't approve any spending until the next billing period. Women apparently aren't good credit risks even when they pay off the fucking balance in advance and in full.

EU Definition of Trustworthy AI

How to measure an AI system trustworthiness?

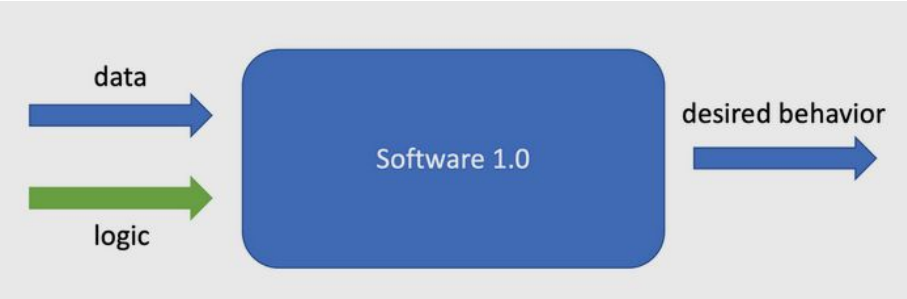
AI **model audit** analysing aspects such as bias, explainability, robustness.

Problem: there is no technical definition of what a model audit should be, yet.

How to audit machine learning models?

Machine learning models as software products

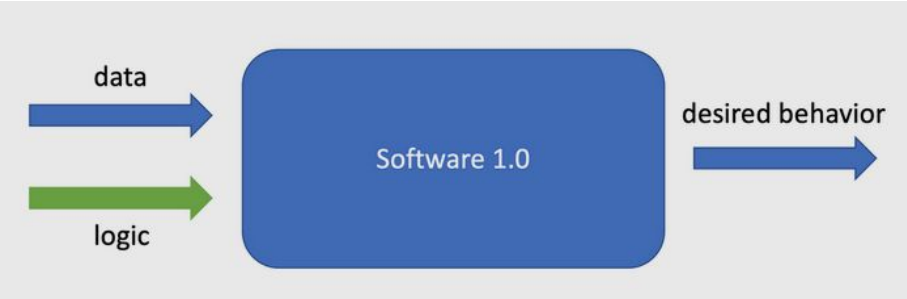
The ultimate objective of data scientists working on machine learning models is to transform them into software products.

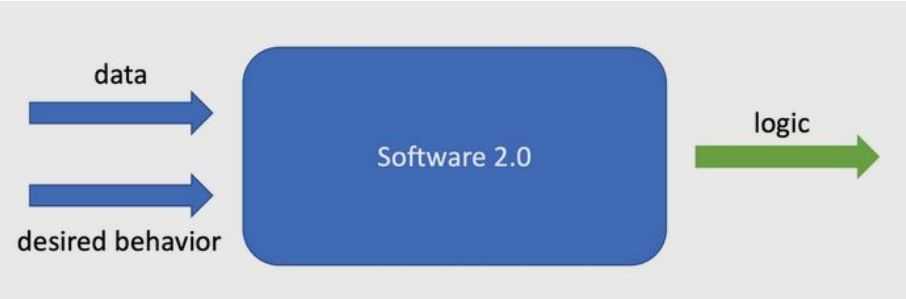


https://www.jeremyjordan.me/testing-ml/

Machine learning models as software products

The ultimate objective of data scientists working on machine learning models is to transform them into software products.





https://www.jeremyjordan.me/testing-ml/

From DevOps to MLOps

What is DevOps?

In software engineering the label DevOps corresponds to a set of *best practices* for developers and IT professionals.

These best practices have the aim of easing the development and productionalization of software products while maintaining **high code quality**.

MLOPs → Translation of the concept of **DevOps** to the field **machine learning**.

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.

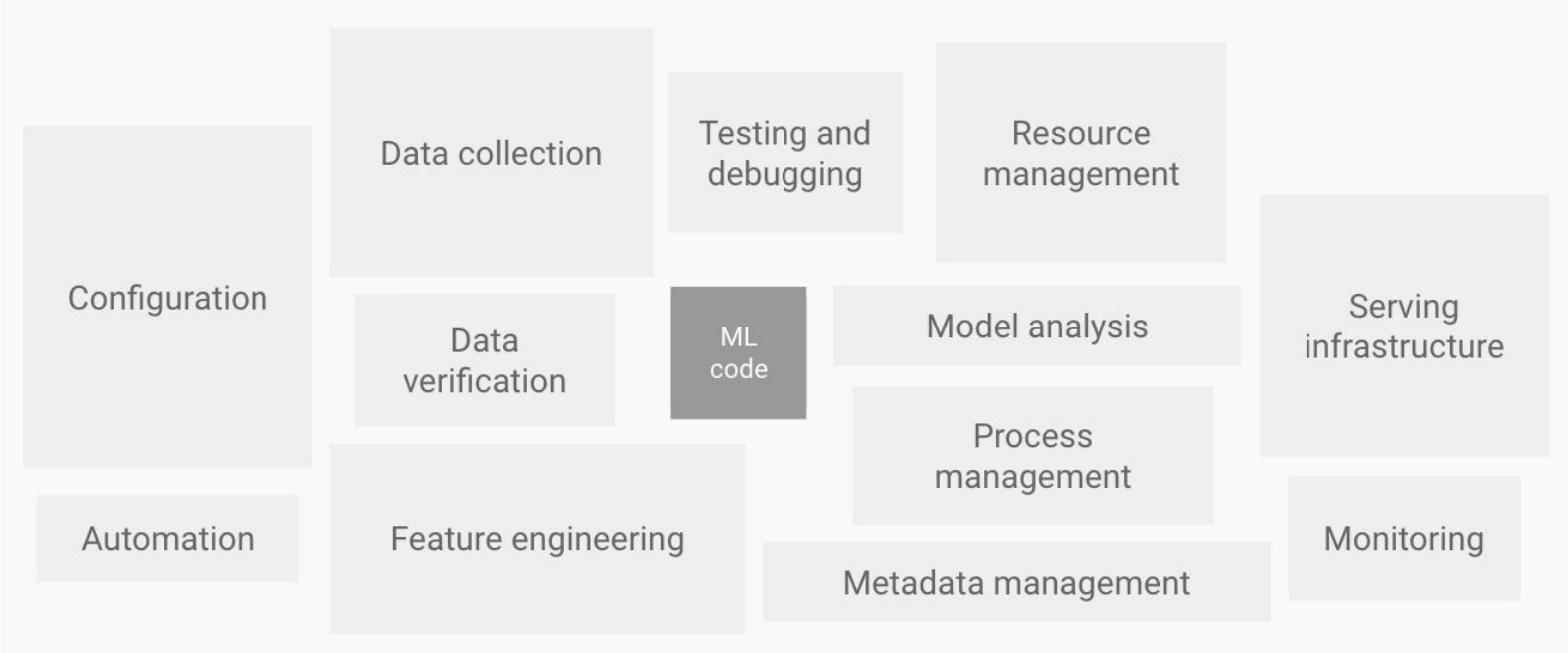
Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-François Crespo, Dan Dennison
{ebner,vchaudhary,mwyoung,jfcrespo,dennison}@google.com
Google, Inc.

Abstract

Machine learning offers a fantastically powerful toolkit for building useful complex prediction systems quickly. This paper argues it is dangerous to think of these quick wins as coming for free. Using the software engineering framework of *technical debt*, we find it is common to incur massive ongoing maintenance costs in real-world ML systems. We explore several ML-specific risk factors to

https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf



DevOps vs MLOps

AI based tools as software products?

Code

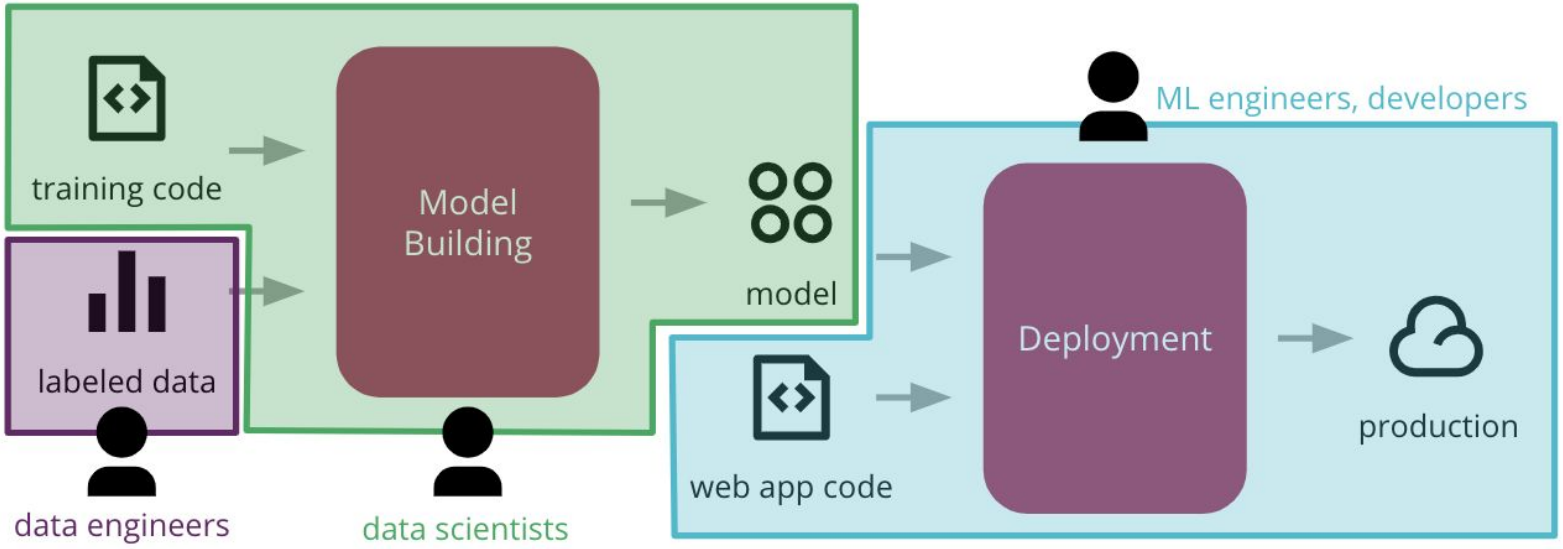
https://martinfowler.com/articles/cd4ml.html

DevOps vs MLOps

AI based tools as software products?

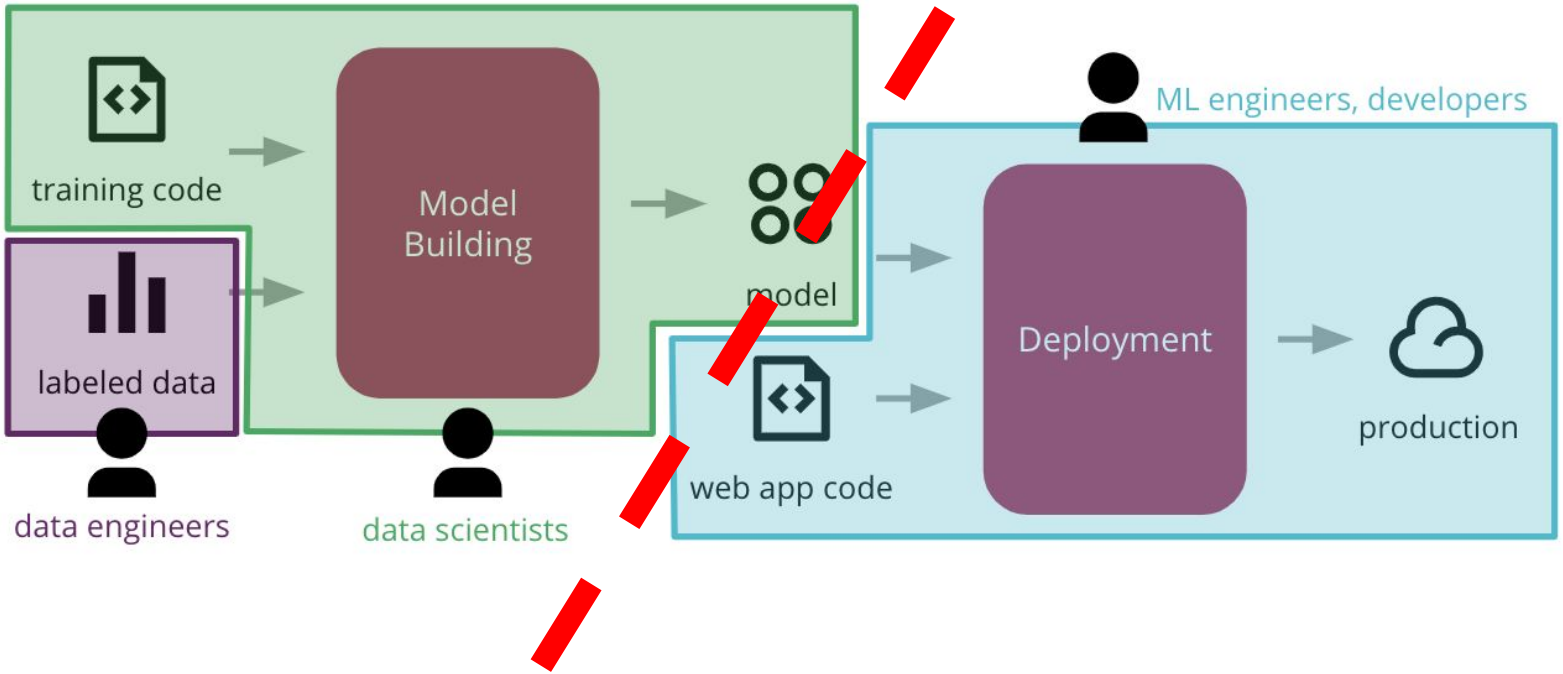
https://martinfowler.com/articles/cd4ml.html

Roles in the ML lifecycle



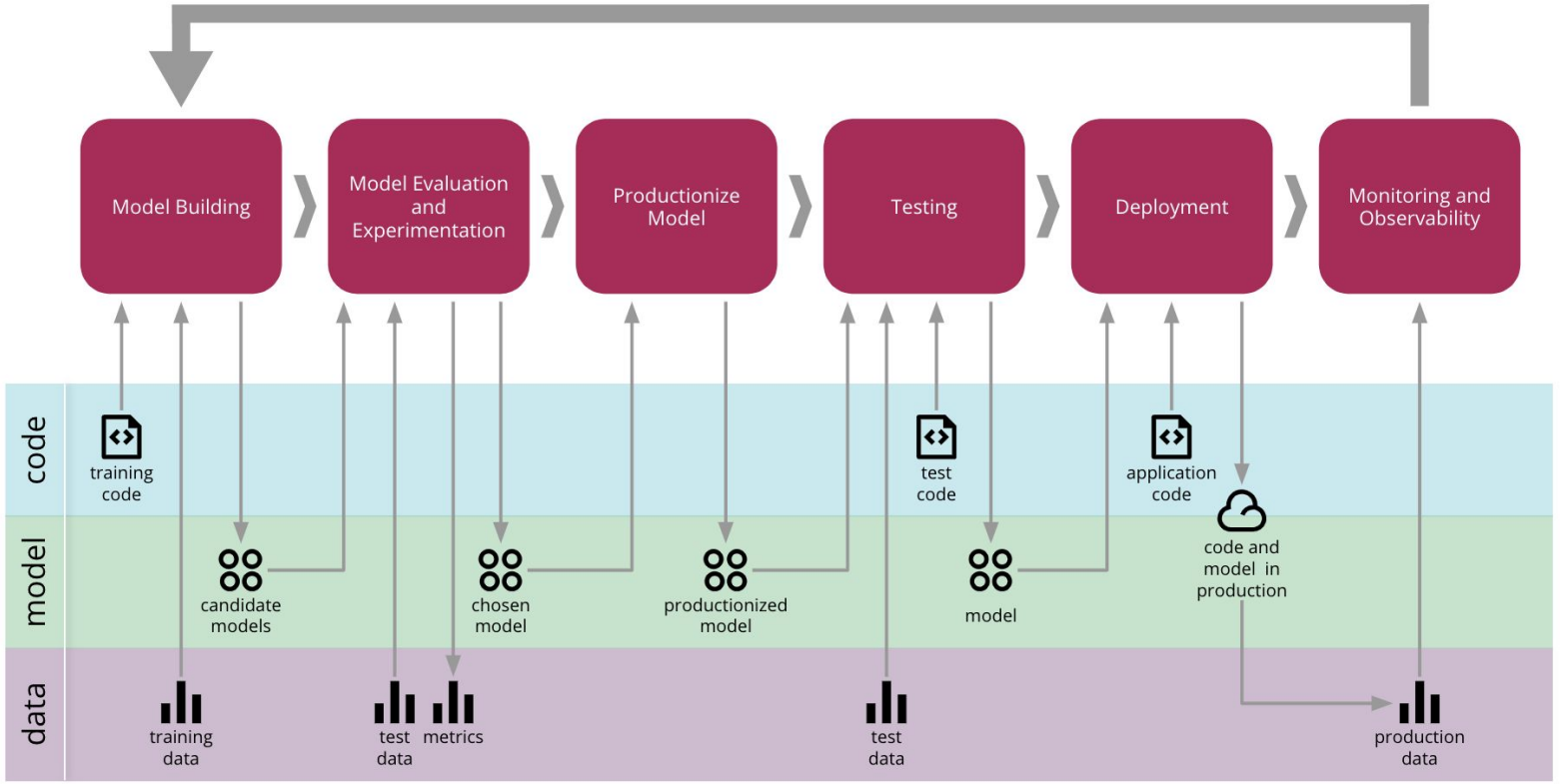
https://martinfowler.com/articles/cd4ml.html

Roles in the ML lifecycle



https://martinfowler.com/articles/cd4ml.html

Machine learning in production



https://martinfowler.com/articles/cd4ml.html

The most important concept in DevOps: CI/CD

Continuous Integration/ Continuous Delivery

Ease the interaction between development and production by giving the possibility to continuously improve code with small interventions.

These small interventions should be:

- Automatically pushed to production
- **Reliable** and **Reproducible**

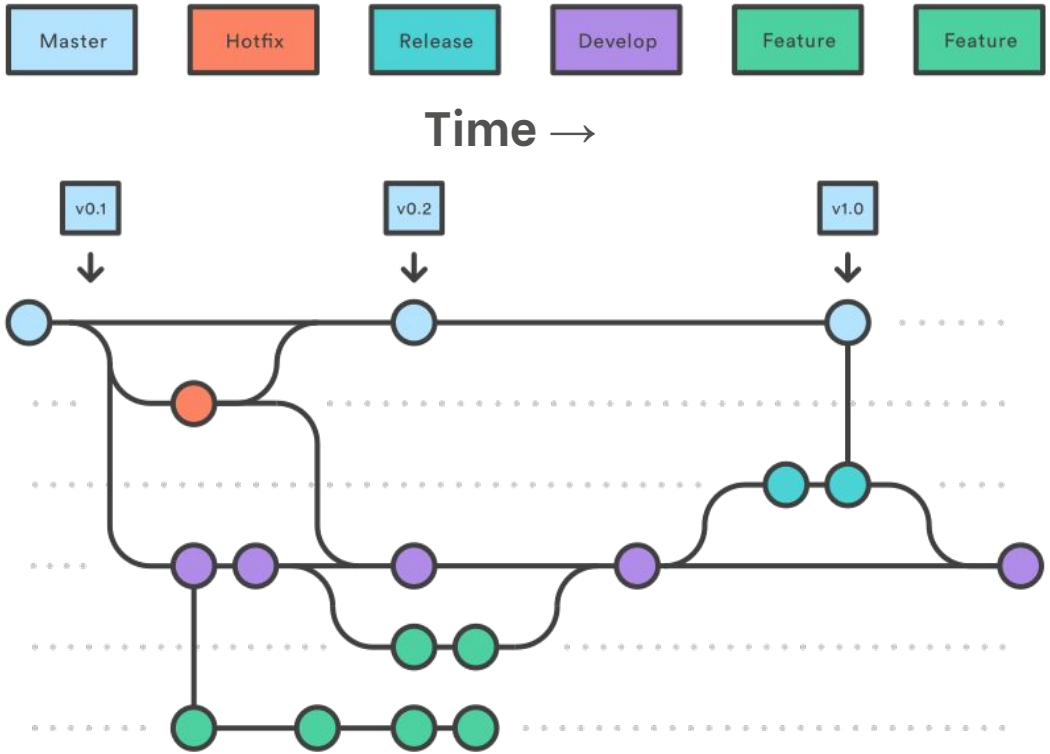
The most important concept in DevOps: CI/CD

Continuous Integration/ Continuous Delivery

Versioning → Make sure your code is reproducible

Automatic Testing → Make sure interventions don’t break anything

Version control

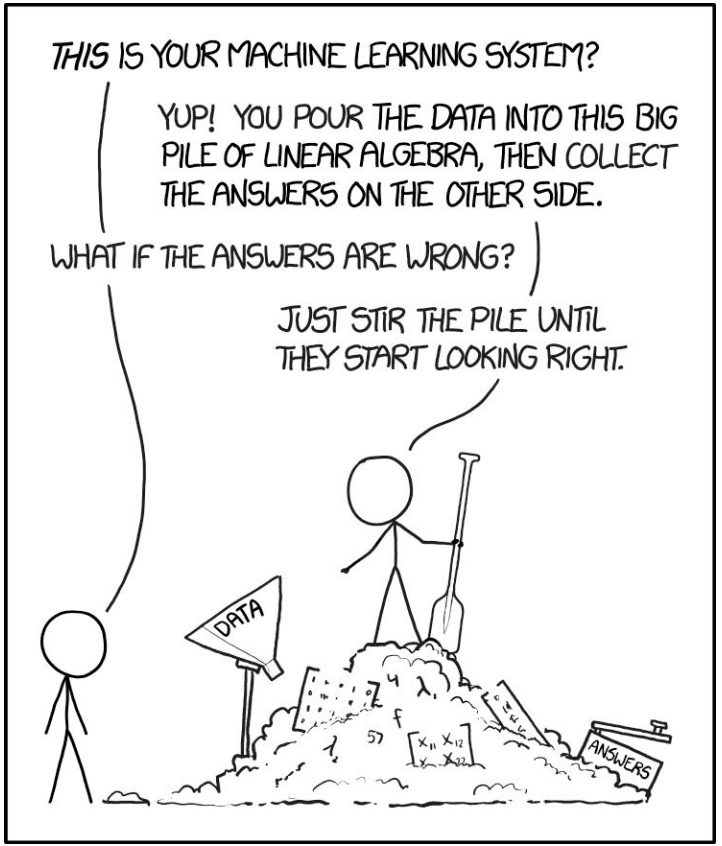


https://evolvingweb.ca/blog/ultimate-guide-agile-git-branching-workflows-drupal

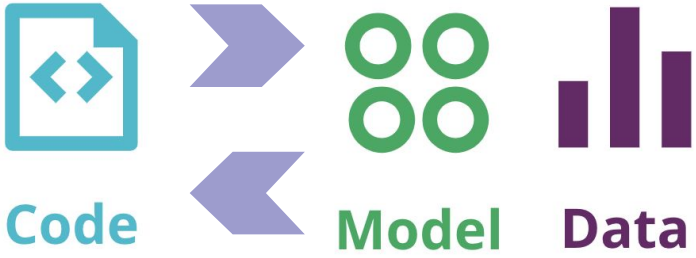
Version control in machine learning

Machine learning model development is increasingly becoming a process of '**trial and error**'

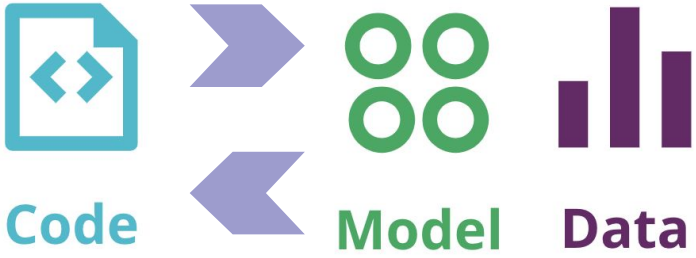
→ Reproducibility issues and potential time waste.



Version control for machine learning?



Version control for machine learning?



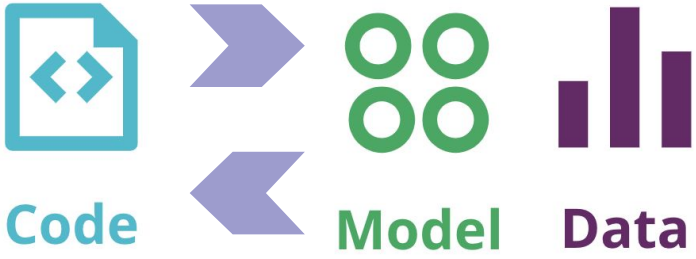
More and more libraries and tools

https://dvc.org

https://mlflow.org

https://wandb.ai

Version control for machine learning?



More and more libraries and tools

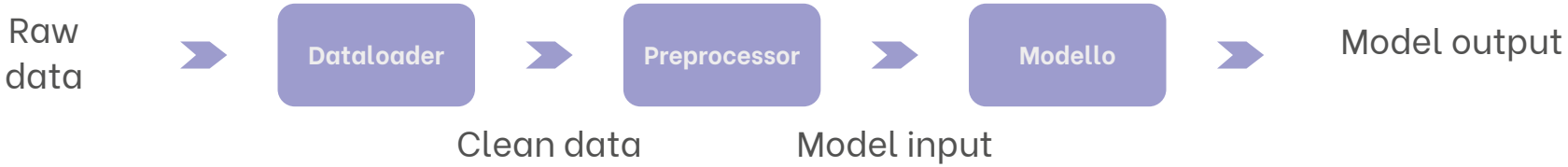
https://dvc.org

https://mlflow.org

https://wandb.ai

Model example

Toy model used for this practical example → Binary classification (Adult Income Dataset)



Dati già divisi in :

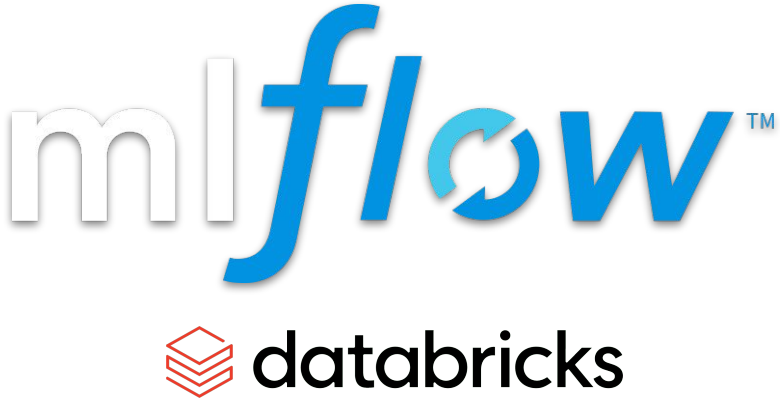
- Train
- Validation
- Test (hold-out)

mlflow

Open source platform for ML lifecycle management. Allows to do:

- Experiment tracking and reproducibility
- Organize centralised model registries
- **Model packaging and deployment**

Developed and maintained by databricks → a lot of man hours have been dedicated to the project and it shows, especially when comparing to other open source alternatives.



Practical example

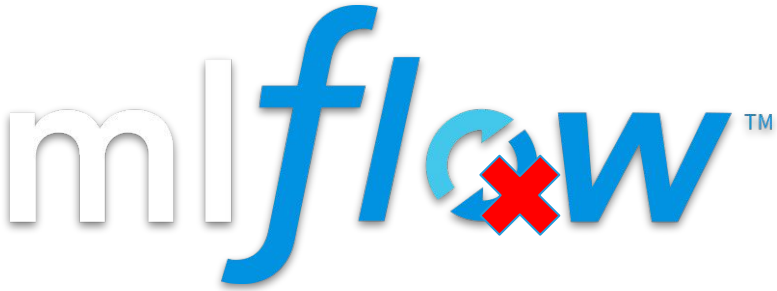
https://github.com/gillus/EURAC-workshop-2023

Open issues with ml*flow*

Despite the claims it is not a ML lifecycle tool (yet). No tools for deployment monitoring, A/B testing, active learning.

UI becomes really sluggish above a certain (large) number of experiments.

Packaging models by hand is still more efficient in terms of computational resources needed to deploy a model.



Options for data and models version control

Level 0: No versioning

Level 1: Data and models save as snapshots with every model training (es: **mlflow**)

Level 2: Data model and code versioned as a single assed

Level 3: Dedicated versioning tools (es: **git LFS**, **DVC**)

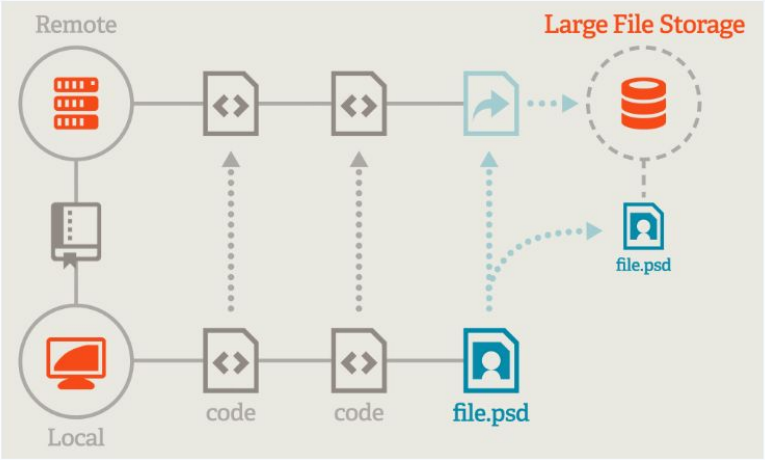
Level 4: Feature stores

Git Large File Storage

Git module to version large files

Limitations:

- Requires dedicated server
- Often limited storage limit for single files
- No caching → slow data push and pull



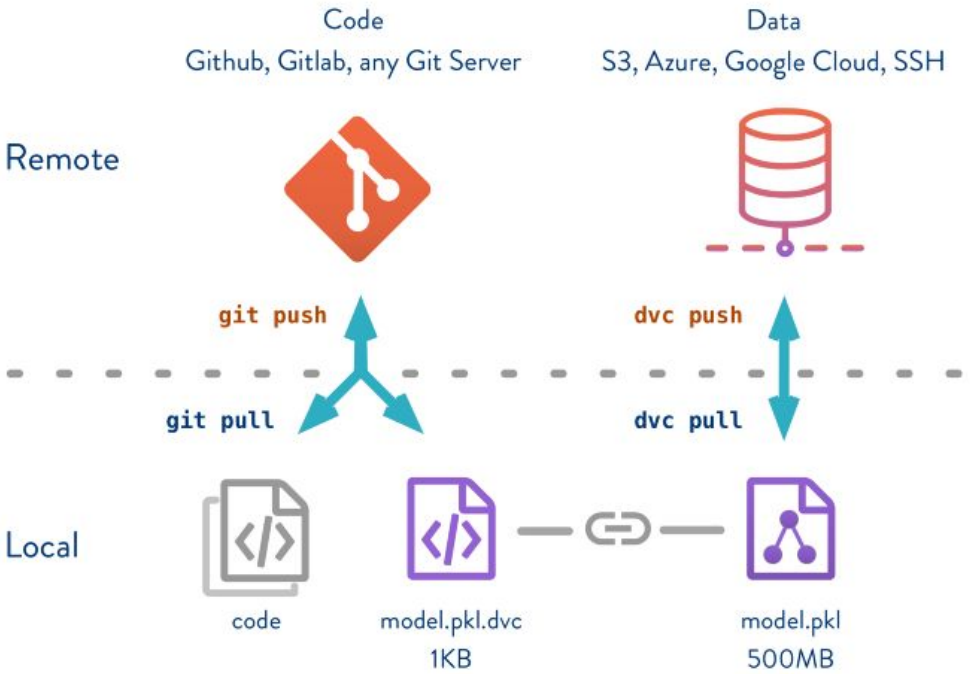
Git-LFS working - (Gif Source - https://Git-LFS.Github.com/)

Data Version Control

Data Version Control library directly integrated with **git**

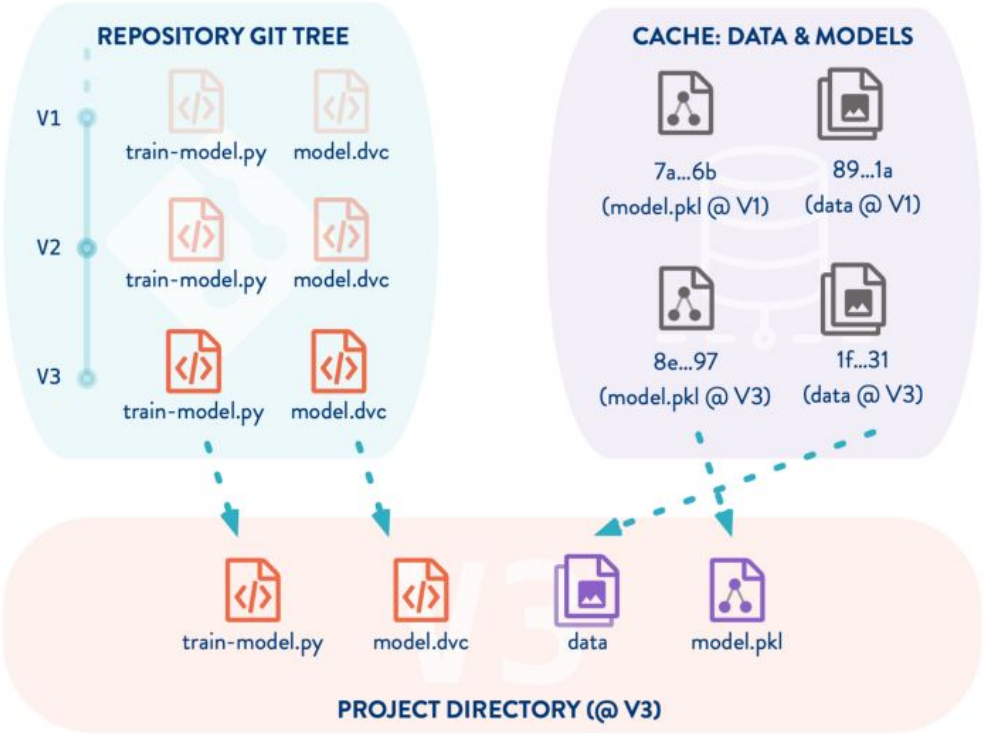
Machine learning native, it allows to perform version control of the whole machine learning pipeline

Data Version Control



dvc.org

Data Version Control



dvc.org

Practical example

https://github.com/gillus/EURAC-workshop-2023

Software testing

Any small code update should be safe

→ It should not break anything or make things worse (and if it does we should be able to roll back to the most recent working version)

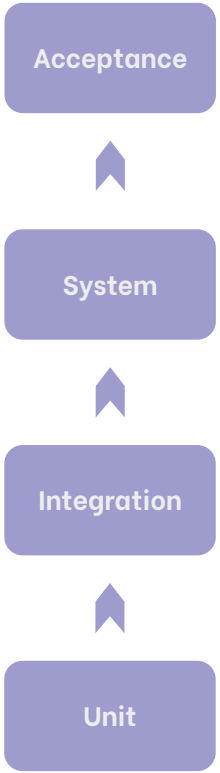
Several types of test depending on scope.

Software testing

Any small code update should be safe

→ It should not break anything or make things worse (and if it does we should be able to roll back to the most recent working version)

Several types of test depending on scope.



Software testing

Any small code update should be safe

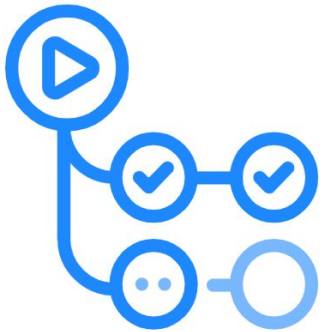
→ It should not break anything or make things worse (and if it does we should be able to roll back to the most recent working version)

Most popular **python testing** library: pytest

Software testing

Any small code update should be safe

→ It should not break anything or make things worse (and if it does we should be able to roll back to the most recent working version)

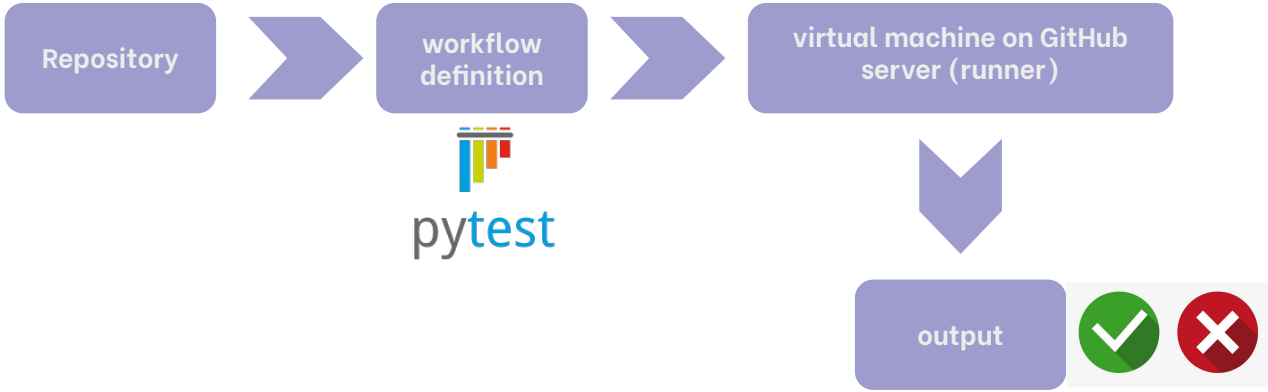


GitHub Actions

Automated testing pipelines: every time code is updated a series of tests is performed. If any of the tests fail updated is rejected.

GitHub Actions

GitHub Actions



```yaml
name: Test

on: [push]

jobs:
    build:
        runs-on: ubuntu-latest
        strategy:
            matrix:
                python-version: [3.7, 3.8]

        steps:
        - uses: actions/checkout@v2

        - name: Set up Python ${{ matrix.python-version }}
          uses: actions/setup-python@v2
          with:
            python-version: ${{ matrix.python-version }}

        - name: Install dependencies
          run: |
            python -m pip install --upgrade pip
            pip install -r requirements.txt
            pip install .

        - name: Pytest
          run: |
            pytest -v --maxfail=3 --cache-clear
```

Example:

This action performs pytest every time a [push] event happens.

Example: pytest and GitHub action

https://github.com/gillus/EURAC-workshop-2023

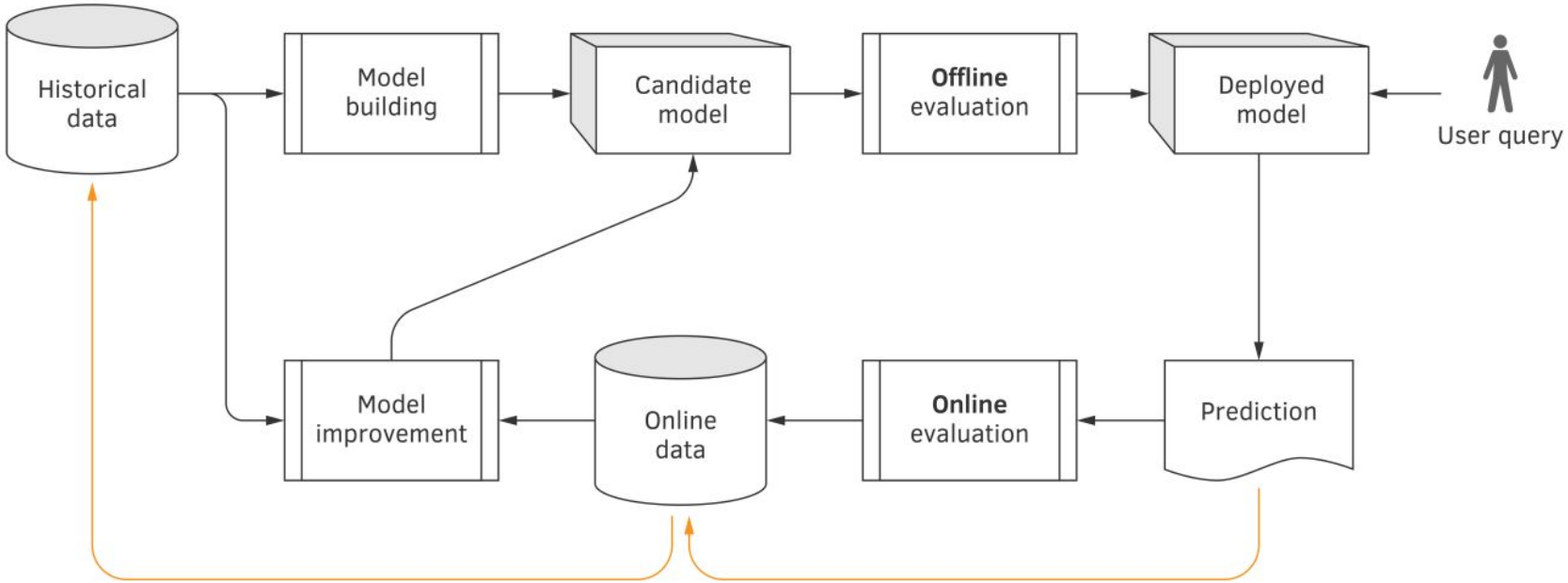
Alternative

GitLab CI/CD

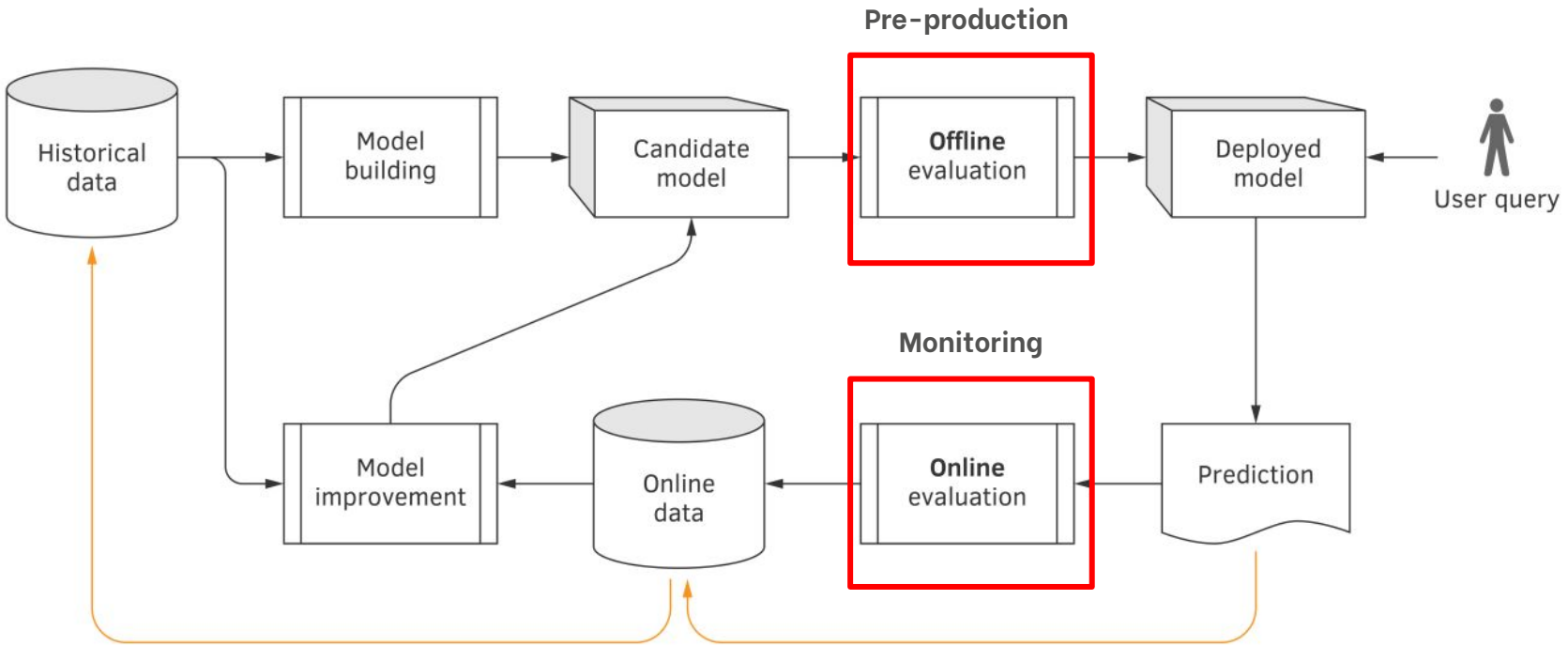
Automation server classico

Machine learning model testing

Model testing



Model testing

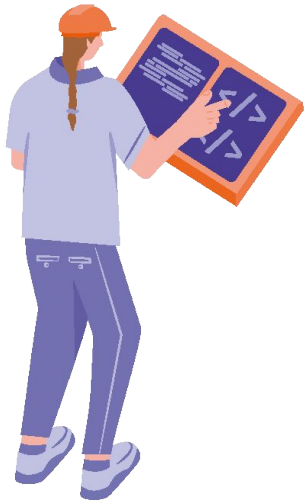


Model testing

What is model testing in machine learning?

Making sure models are properly generalizing with respect to the data and their learning task. Aim is to avoid **undesired behaviour** in production.

→ Testing should be done continuously, in pre-production and production.



Model testing

A first step can be identifying and counting the test points where the model answers were catastrophically wrong.

```
41  def test_model_overconfidence_fp(adult_test_dataset):
42      x, y, data_path = adult_test_dataset
43      clf = joblib.load('./model.pkl')
44      predictions = clf.predict_proba(x)
45
46      fp = np.where((clf.predict(x) != y) & (predictions.argmax(axis=1) == 1))
47
48      assert predictions[fp].shape[0] < 0.1 * predictions.shape[0]
49
```

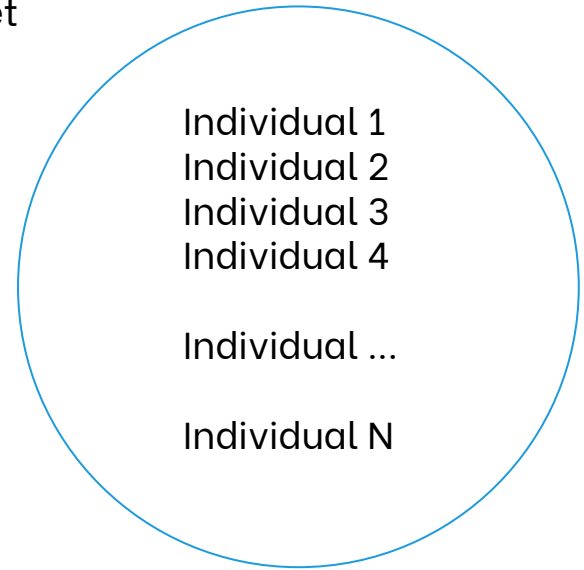
Performance analysis

Once we debugged a model for undesired behaviour, how do we test **model performance**?

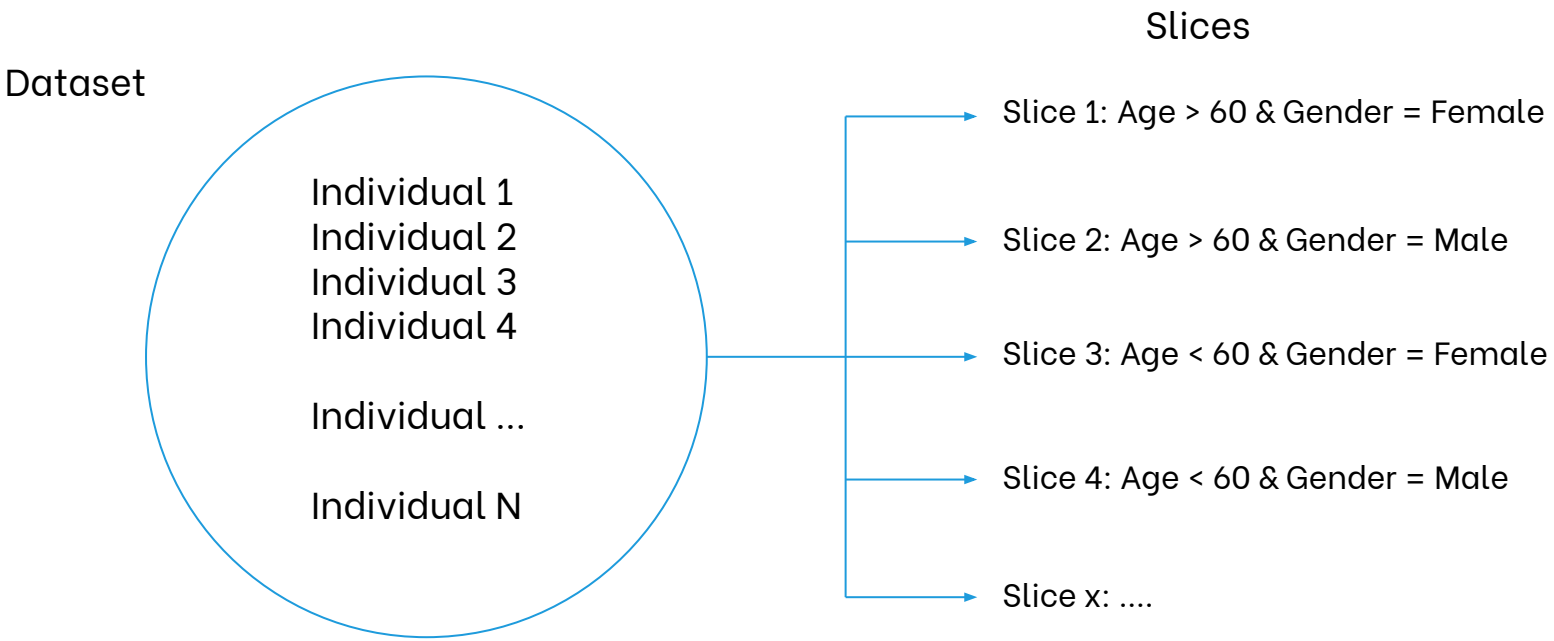
Problem: data scientists tend to make use of very global metrics (Accuracy, F1, etc) → Very limited analysis

Data slices

Dataset



Data slices


Slices
Dataset
Individual 1
Individual 2
Individual 3
Individual 4
Individual ...
Individual N
Slice 1: Age > 60 & Gender = Female
Slice 2: Age > 60 & Gender = Male
Slice 3: Age < 60 & Gender = Female
Slice 4: Age < 60 & Gender = Male
Slice x: ....

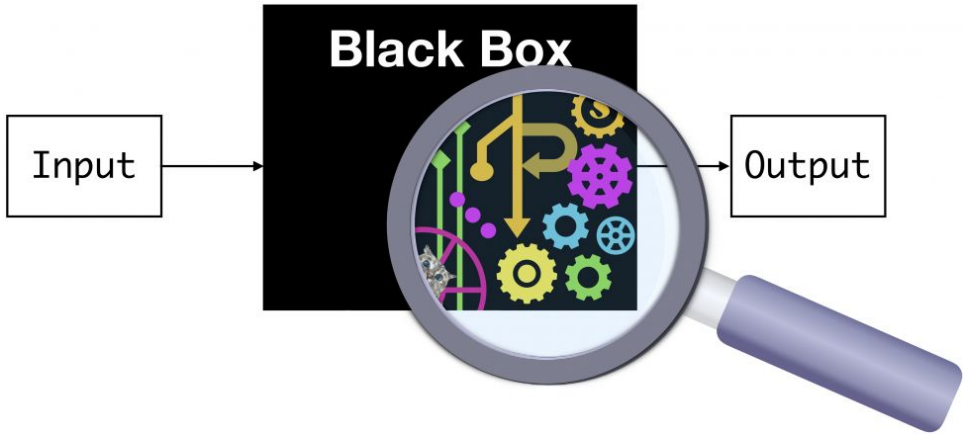
Slice Analysis

A great test is to test the performance of the model against a simple baseline (for example majority classifier)

Slice 1: Age > 60 & Gender = Female	Good model performance
Slice 2: Age > 60 & Gender = Male	**Warning**: Baseline performs better than model
Slice 3: Age < 60 & Gender = Female	Poor model performance
Slice 4: Age < 60 & Gender = Male	Great model performance
Slice x:	...

How to test model behaviour?

The black box issue
in machine learning



Increasing model **complexity** is making it difficult to properly test and understand model behaviour.

https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

Potential issues arising with black box models

Some of the typical issues that can affect model behaviour

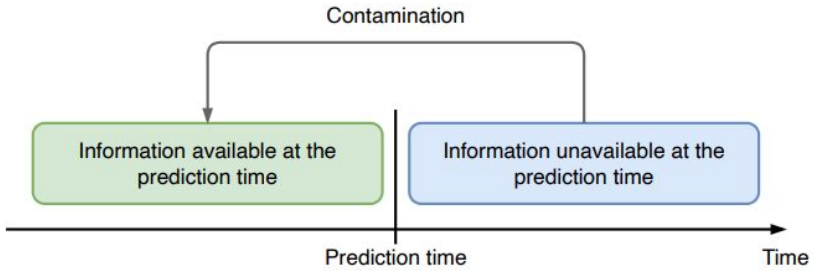
1) **Data leakage**
2) **Robustness issues**
3) **Bias e fairness**
4) ...

Data leakage

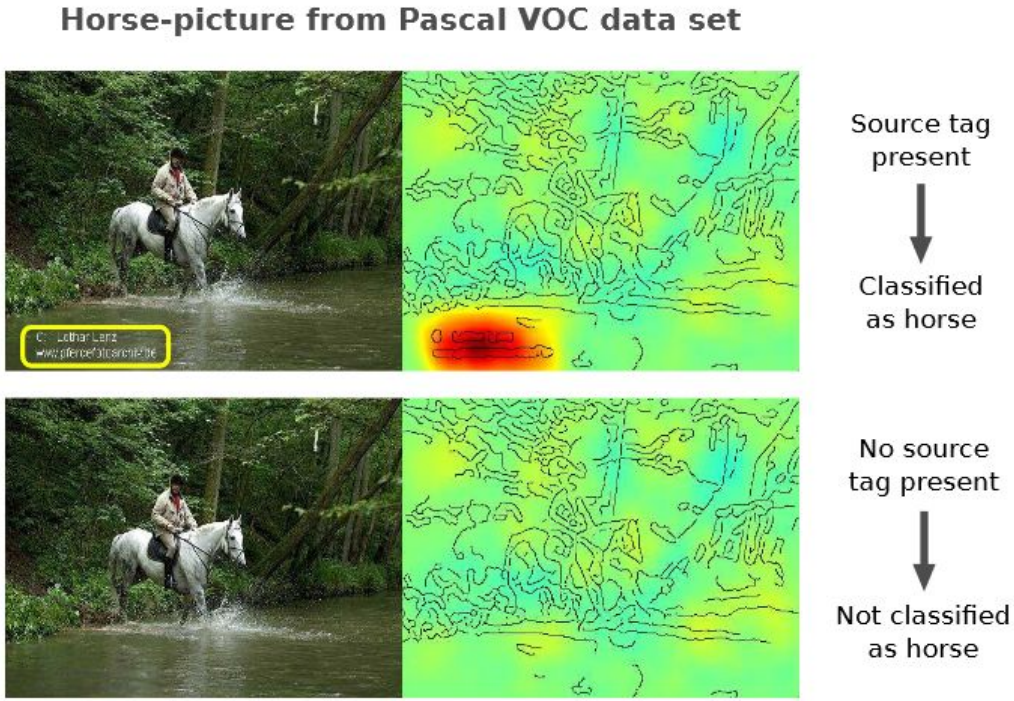
Model is trained with information that will not be present in production.

Typical reasons:

- Our learning target is ‘hiding’ inside one of our features
- One of the input features comes from the future (forecasting problems)



Data leakage



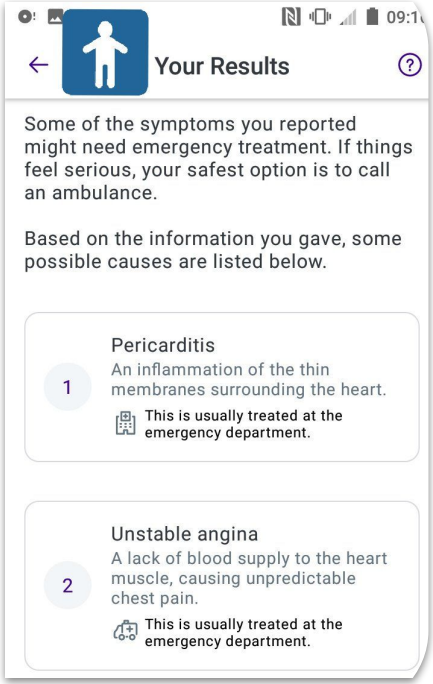
Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. Lapuschkin et al. Nature comm (2019)

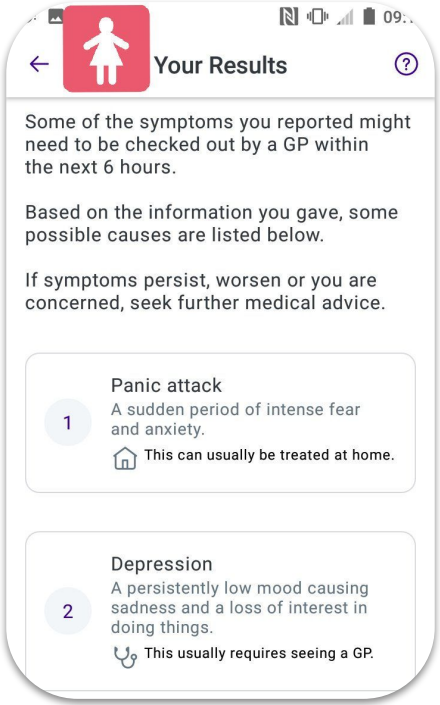
Bias e Fairness

Models learn from data →

Data contains bias that might reflect historical societal aspects we want to leave behind.

Bias e Fairness

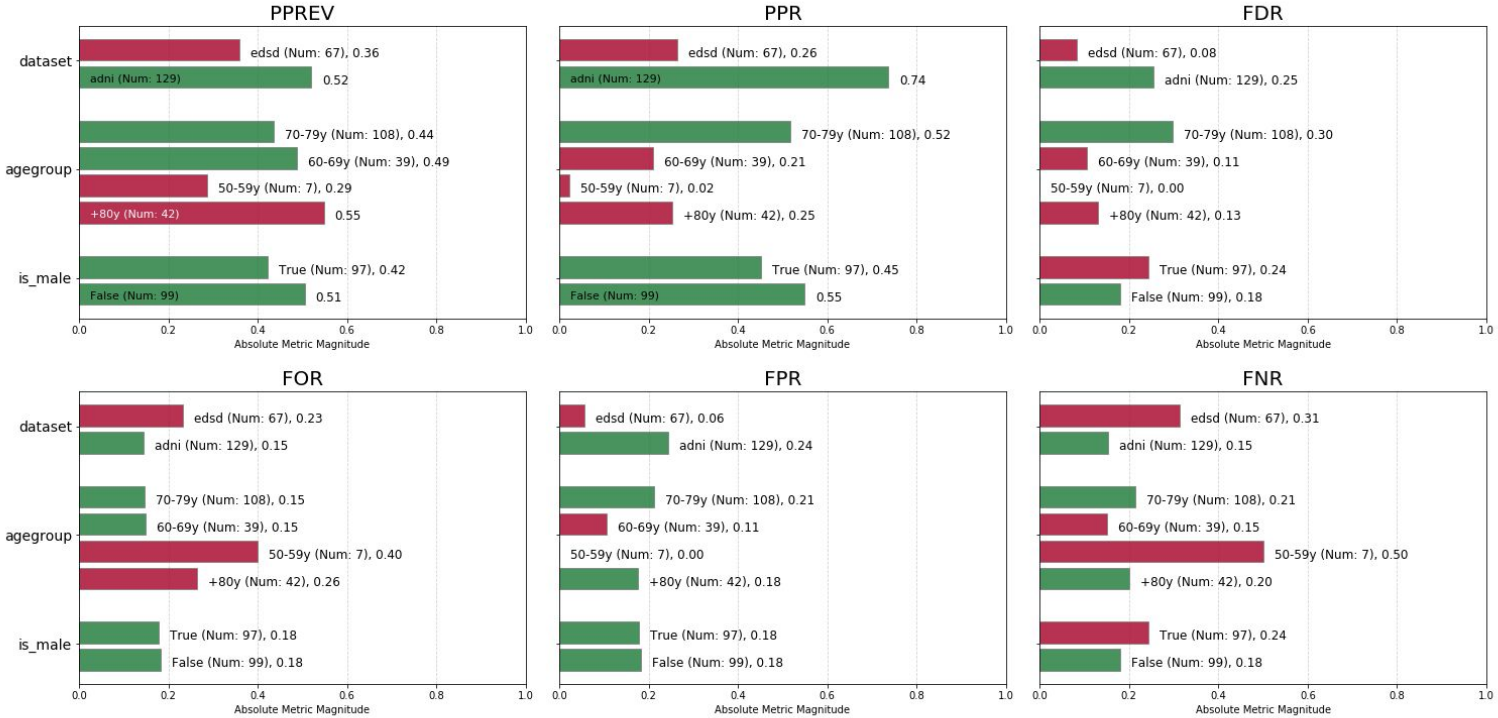




https://www.ft.com/content/19dc6b7e-8529-11e8-96dd-fa565ec55929

Bias e Fairness

Aequitas toolkit



Oala et al, ML4H Auditing: From Paper to Practice, NeurIPS2020

Model robustness

Model robustness is usually defined as the **stability of the prediction with respect to small perturbations**

In some specific contexts model instability might be associated with safety concerns

Model robustness

Model robustness is usually defined as the **stability of the prediction with respect to small perturbations**

In some specific contexts model instability might be associated with safety concerns

For example: can my autonomous driving model fail is the input becomes slightly noisier?

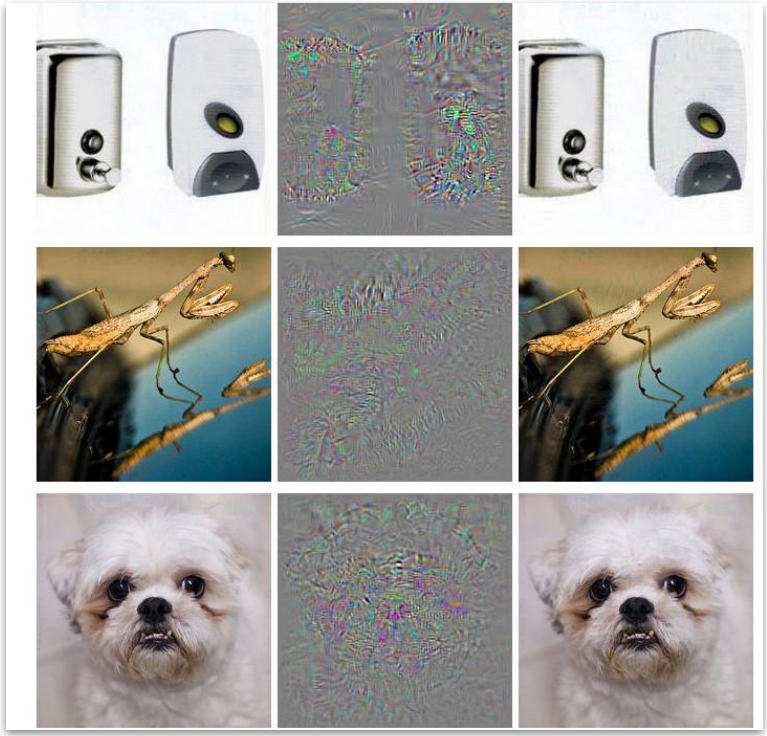
$$\|\mathbf{x} - \mathbf{x}'\| \leq \delta$$

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq \epsilon$$

Adversarial robustness

Related to model robustness → **targeted** perturbation based attacks leading to model mistakes.

Source: Berkeley AI Research (BAIR)

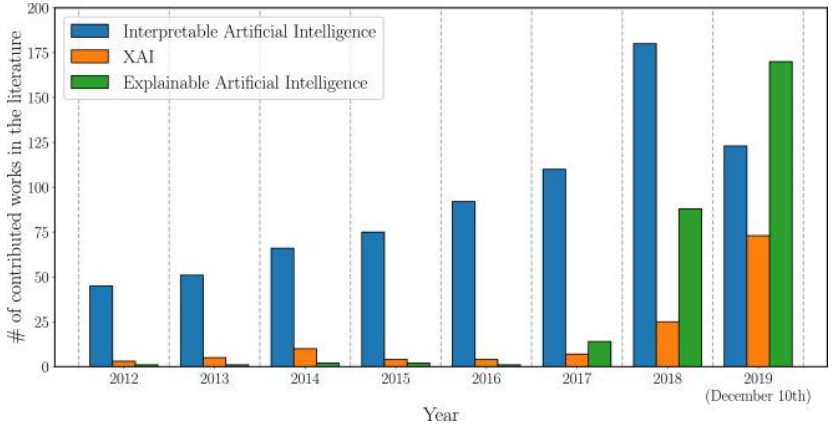


"Intriguing properties of neural networks", Szegedy et.al, 2013

How to analyze black box models: eXplainable AI (XAI)

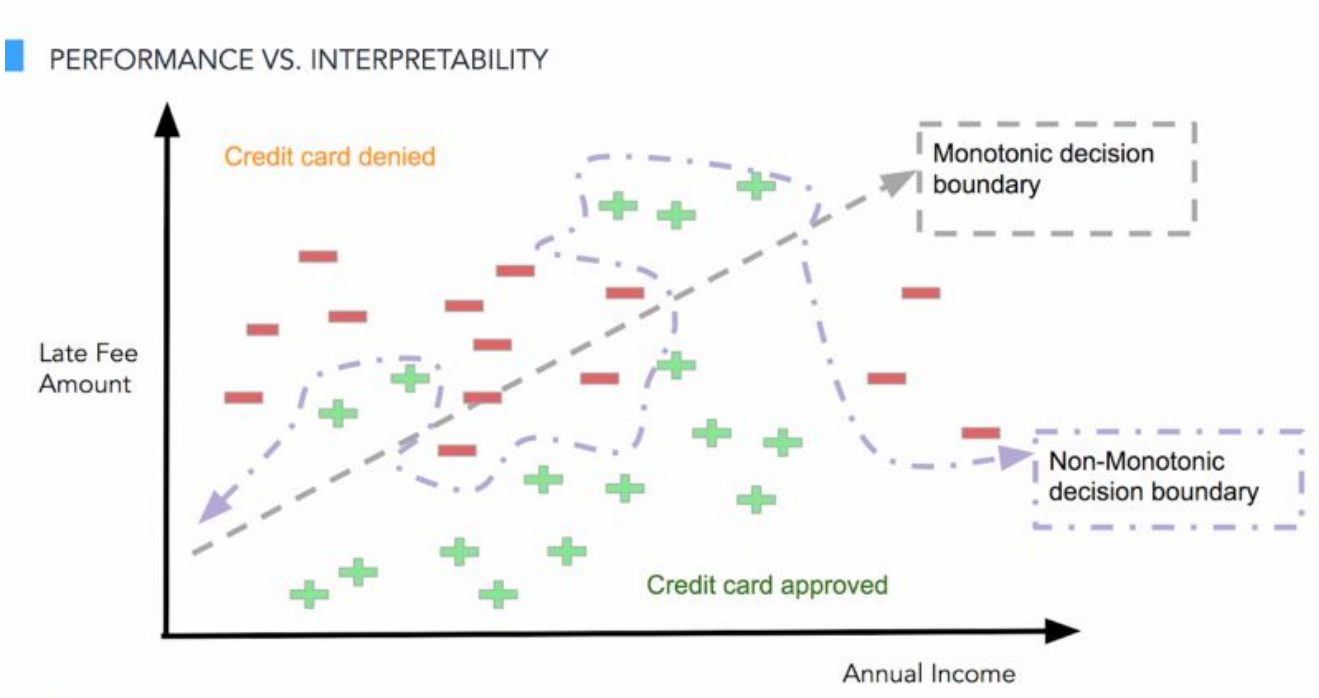
Research field whose aim is to help humans at **interpreting** decisions made by models.

Increasing academic interest in the last 5 years → More and more methods and libraries.



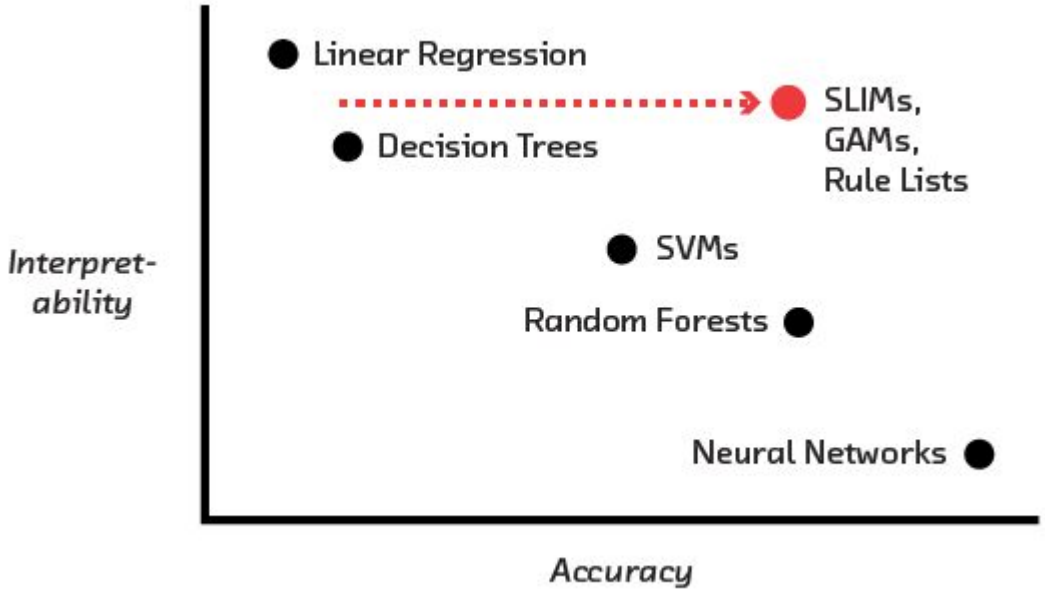
Arrieta et al, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion, 58, 2020

Model explainability



https://www.kdnuggets.com/2018/12/explainable-ai-model-interpretation-strategies.html

Model explainability

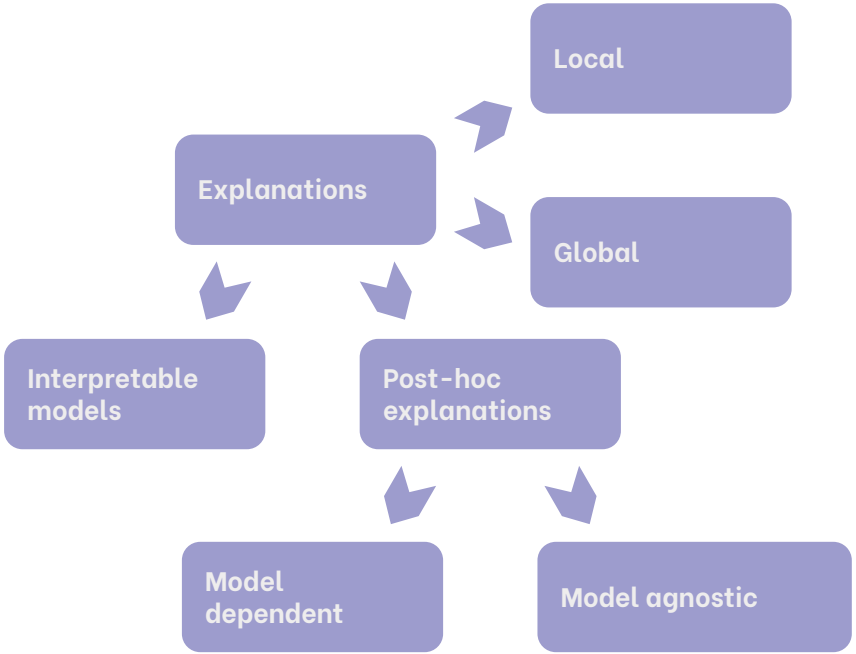


https://ff06-2020.fastforwardlabs.com/

Taxonomy of eXplainable AI

Multiple ways to approach the problem

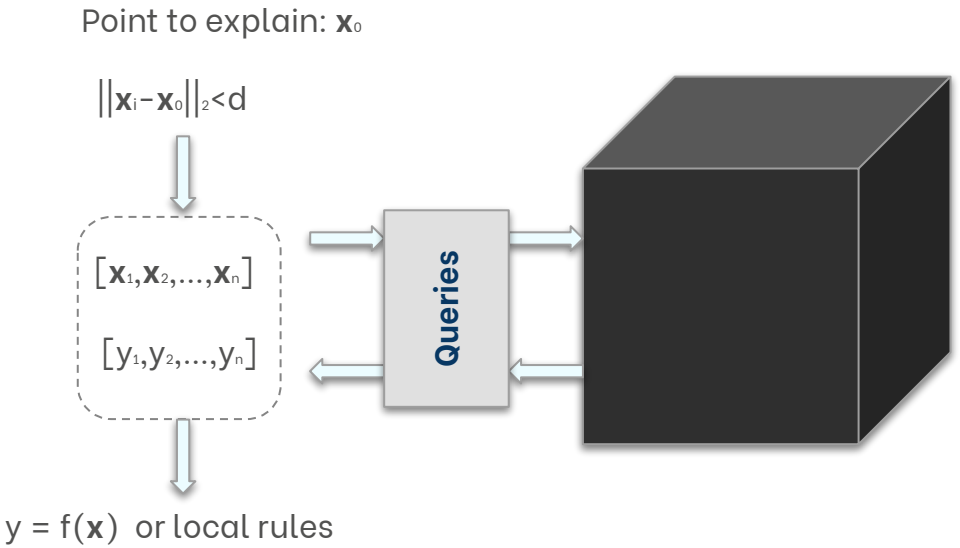
- Are we creating interpretable models or are we interpreting existing models?
- Are we explaining single predictions or the model as a whole?
- Are we looking inside the model inner workings or are we just analysing its answers?



Perturbation based methods

Family of methods designed to generate **local** explanations **agnostically**.

Explanations are generated by building a **simpler, interpretable model** able to approximate to original model around a certain prediction

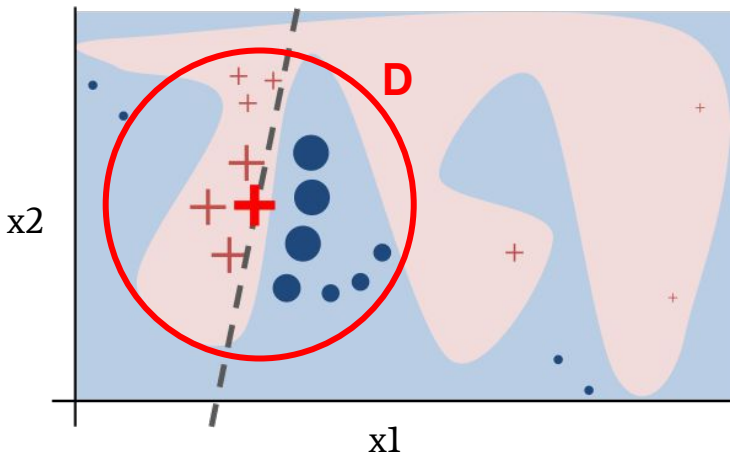


Local Interpretable Model-agnostic Explanations
(LIME)

Among most popular methods, based on perturbations.

Idea:

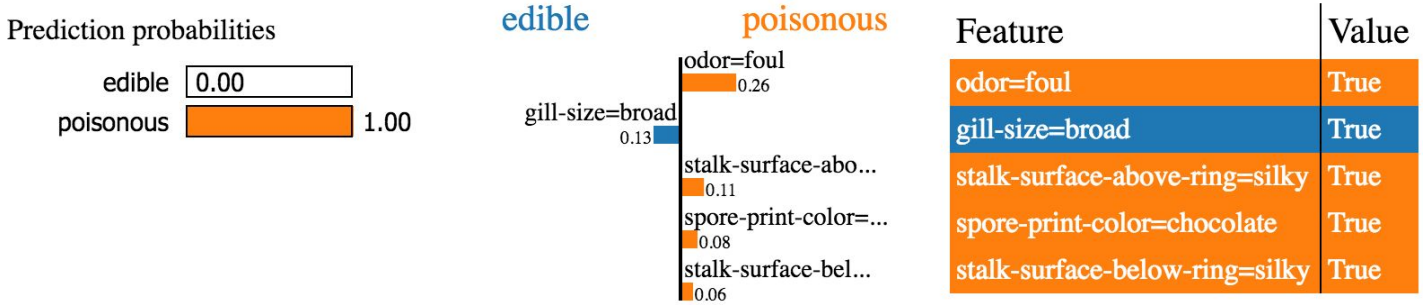
- Generate **N** points in the area **D** around the point that needs to be explained.
- Collect your model answers for these points.
- Build a linear model using the points (x,y) obtained in the previous steps.



https://arxiv.org/pdf/1602.04938.pdf

LIME

Example



- Several hyperparameters required to build explanation (N of points, distance D, etc)
- Can be applied to any type of data (tabular, image, text)
- Not always possible to converge to an explanation
- Explanations are not **prescriptive.**

Shapley values

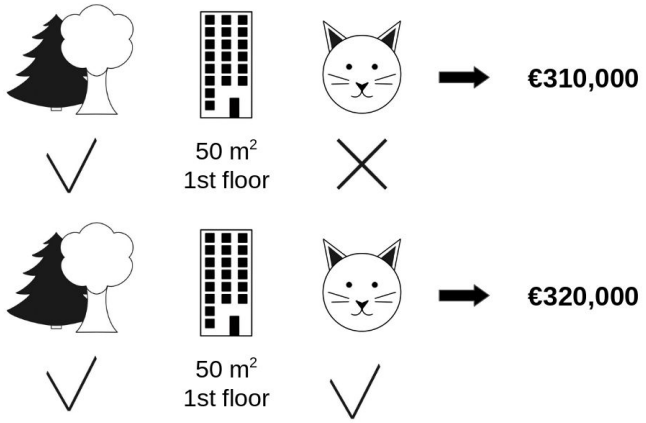
Concept originating from game theory: *how to redistribute the prize of a cooperative game ?*

Shapley values define a way to fairly assign a share of the prize to each player.

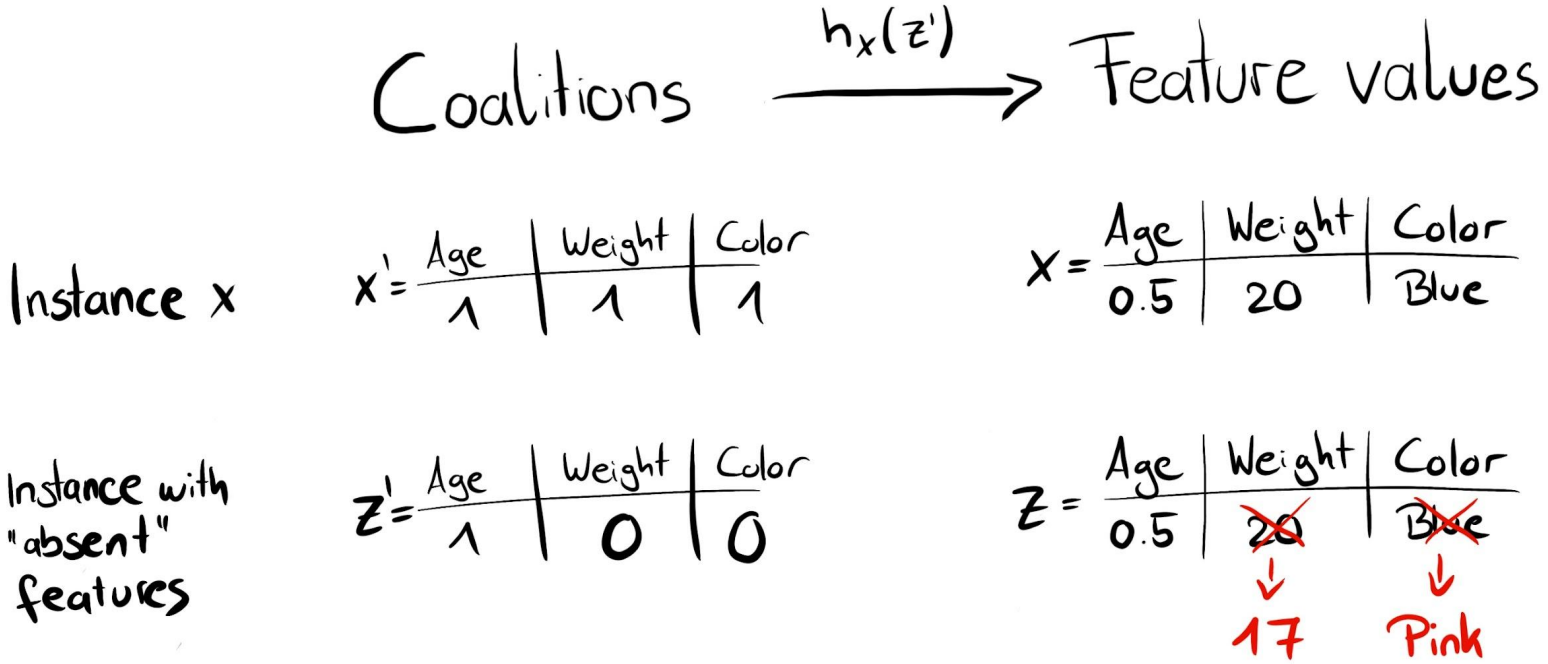
When applied to machine learning:

Players → Input features

Prize → Model output



https://christophm.github.io/interpretable-ml-book/



https://christophm.github.io/interpretable-ml-book/

SHAP

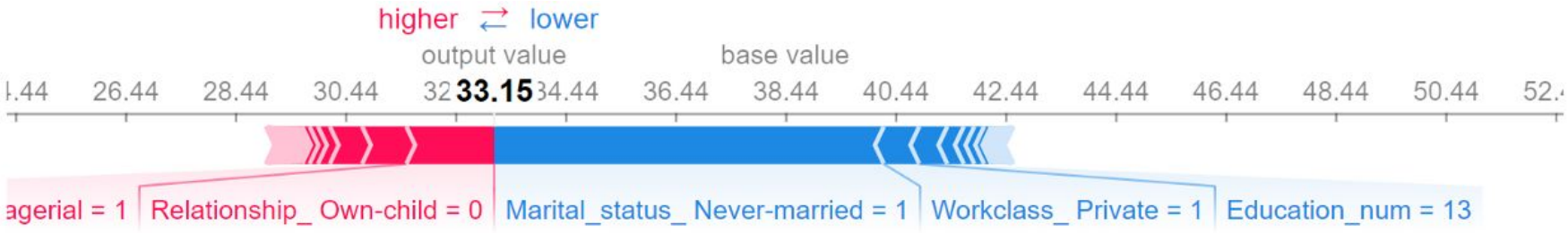
A method to approximate Shapley Values

Very popular Python library that can be used to approximate Shapley values

KernelSHAP → Model agnostic method

TreeSHAP → Faster version compatible with tree-based models

SHAP



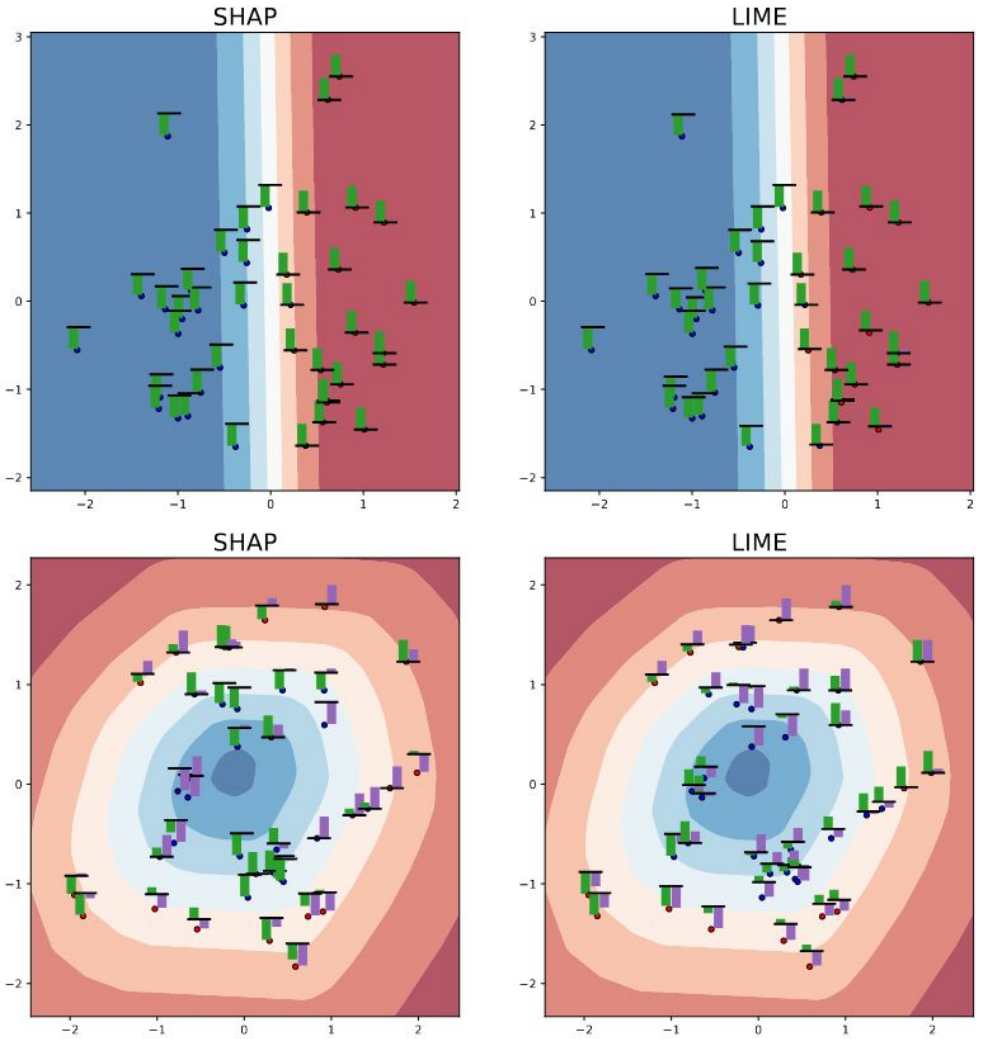
- Open source library, it includes visualization modules
- Calculating SHAP coefficients can be really slow

Robustness of perturbation methods

Perturbation based explanations can be affected by robustness issues

→ Small input change can lead **to big change in explanation**.

Important to monitor **explanation quality**.



https://arxiv.org/pdf/1806.08049.pdf

Issues with local perturbation based explanations

- **Robustness** issues can lead to confusion
- Explanation quality can change a lot also depending on **hyperparameter tuning** → a lot of fine tuning required on specific use case
- Can be **computational intensive** → How long can we wait to get an explanation?

How to turn model explanations into tests?

One should use a lot of domain expertise to create tests based on interpretability analysis. Examples of tests that use explanations:

- Feature a should have positive contribution to the prediction
- Group of explanations should not make use of single feature
- (for imaging): attention should not fall onto particular image regions

Practical example

https://github.com/gillus/EURAC-workshop-2023

Understanding model errors

Understanding model errors

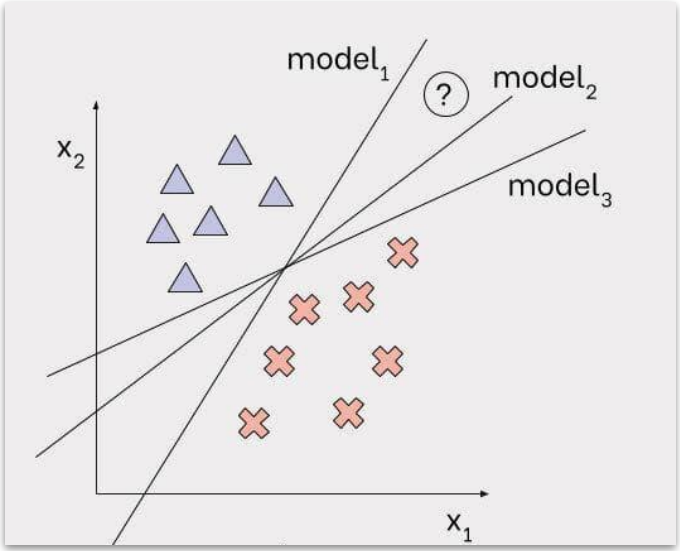
We don’t expect models to be infallible.

Errors can happen but we should be able to define acceptable thresholds based on the data.

Epistemic uncertainty
(reducible)

With epistemic uncertainty represents **lack of knowledge** about the problem we are dealing with.

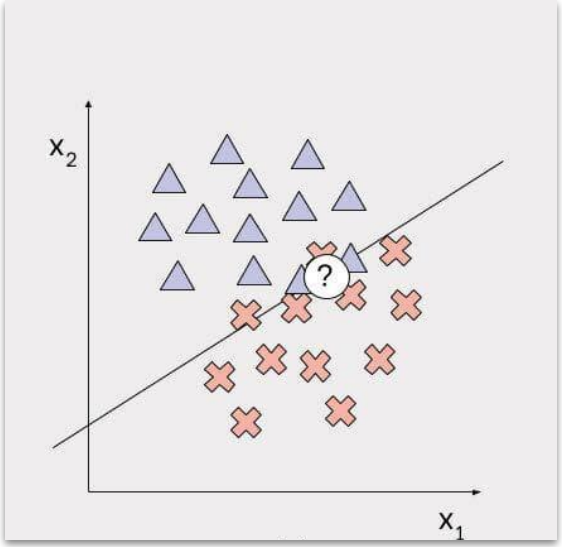
In machine learning it can be usually associated with **lack of data**.



Aleatoric uncertainty
(irreducible)

This type of uncertainty is associated to the presence of **noisy data**.

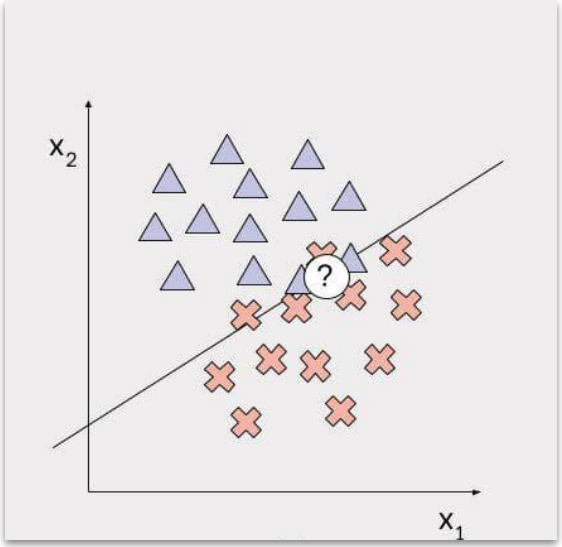
In principle this noise cannot be reduced via feature engineering or data augmentation.



Aleatoric uncertainty
(irreducible)

This type of uncertainty is associated to the presence of **noisy data**.

In principle this noise cannot be reduced via feature engineering or data augmentation.

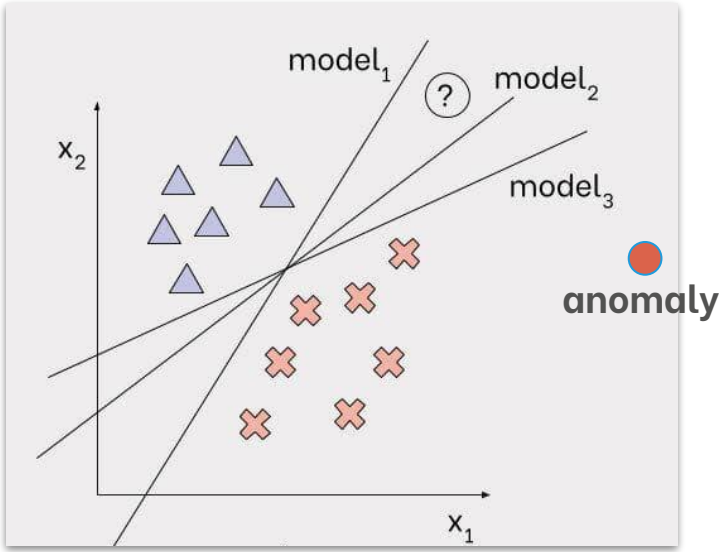


Anomalies

A form of epistemic uncertainty.

What should our model do with points which are not very common in our training data?

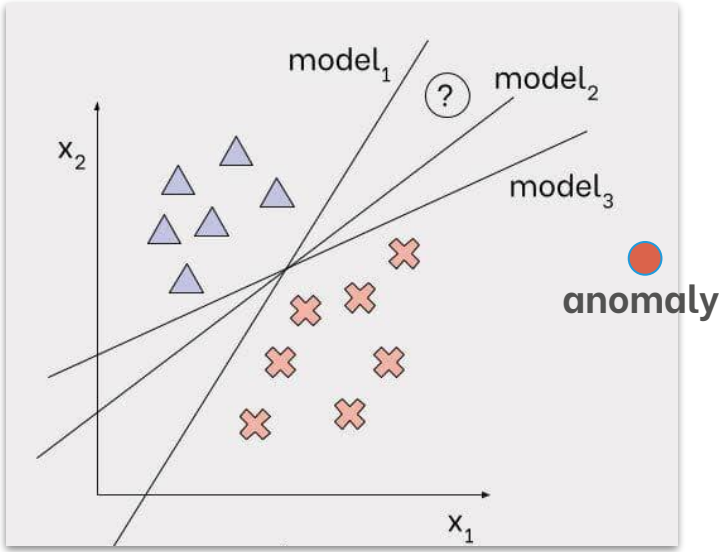
Example: *An 18 year old with a Ph.D.*



Anomalies

Quantifying anomalies corresponds to defining a model's **operational range**

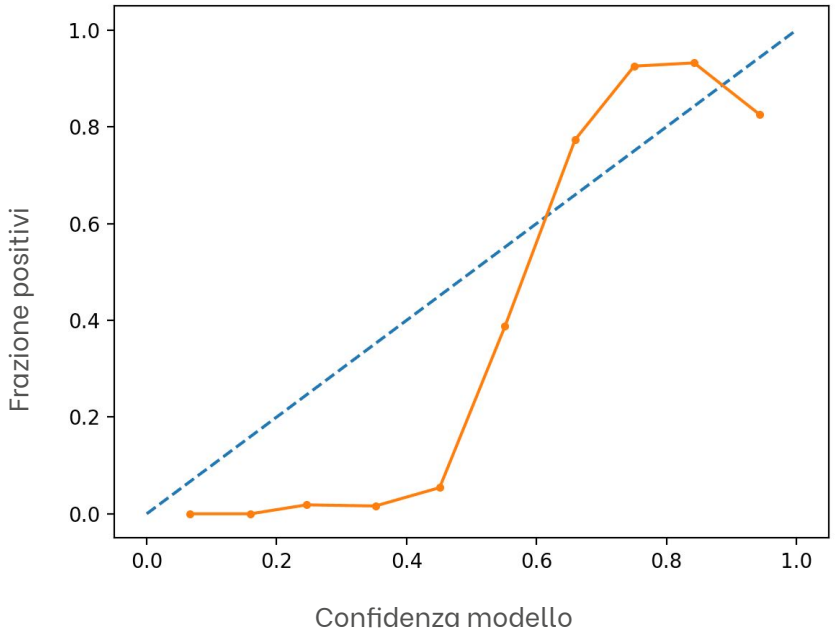
Models are trained on specific dataset distributions → We should be able to determine whether new data belongs to the same distributions.



Model calibration

Standard approach

Model calibration assessed using calibration curve (reliability diagram)

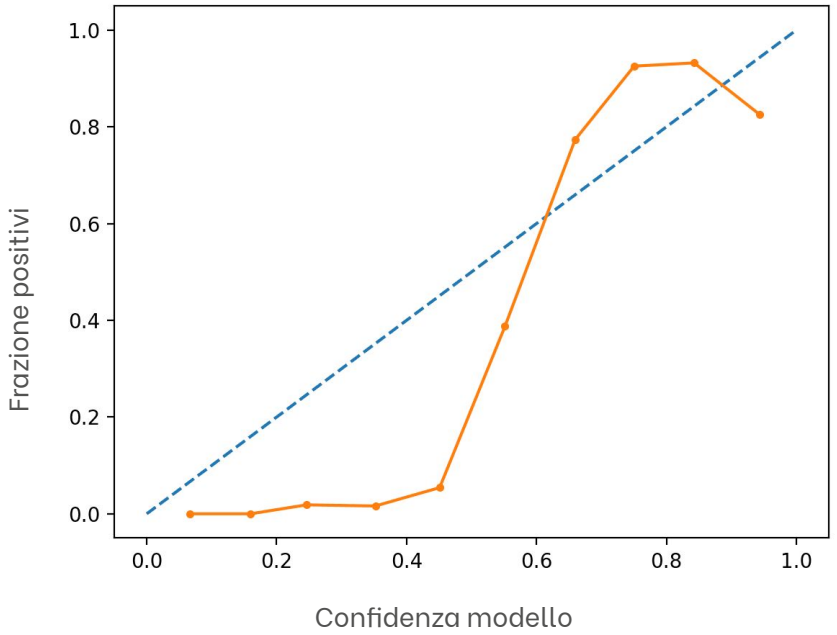


Model calibration

Standard approach

Model calibration assessed using calibration curve (reliability diagram)

Techniques to improve calibration:



- **Platt scaling**
- **Isotonic regression**

model monitoring

Monitoring

We need to be able to properly test model behaviour over time

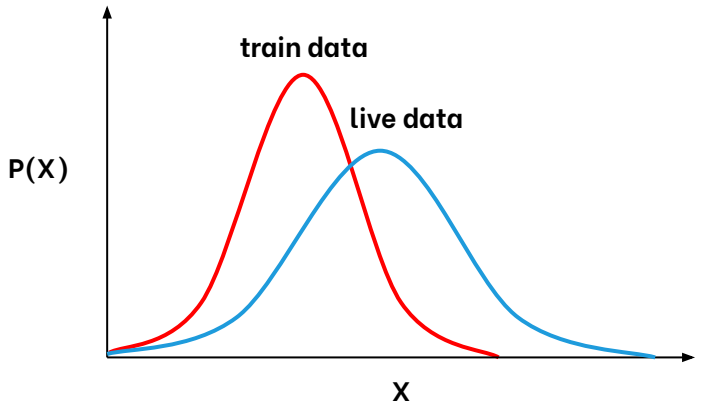
Data changes over time → Models can start operating outside their design space

How can data change:

1. **Data** drift
2. **Concept** drift

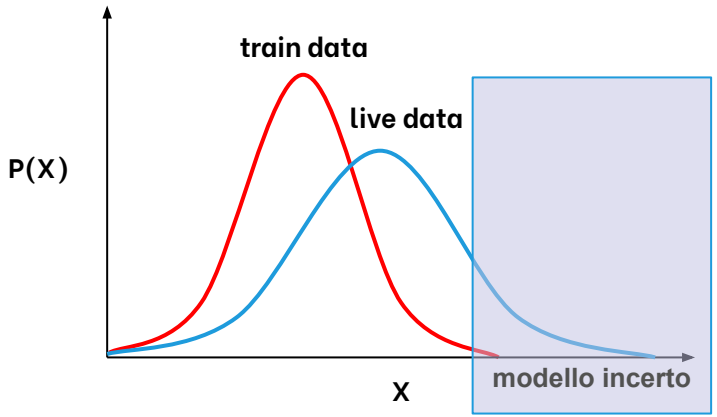
Data drift

Probability distributions of input data are changing over time



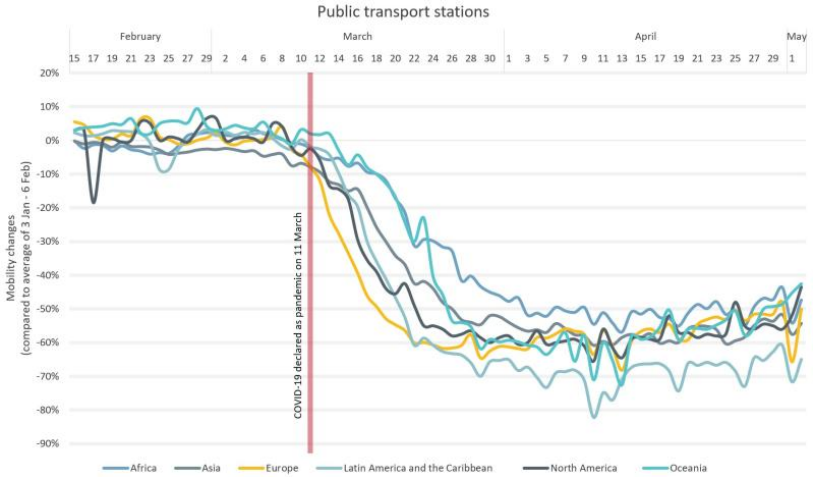
Data drift

Probability distributions of input data are changing over time



Data drift

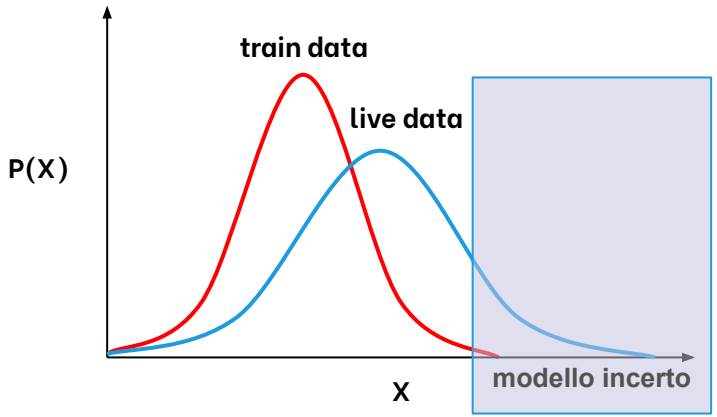
Probability distributions of input data are changing over time



Data drift

Probability distributions of input data are changing over time

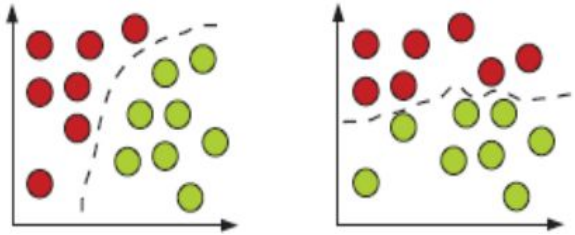
Possible to quantify how much are distributions changing using a number of statistical tests



Concept drift

Changes in the relationship between input and output distributions can vary significantly over time depending on the problem at hand.

Distributions in the feature space may remain similar; what changes in this case is the decision boundary of the problem.



HAL Id : hal-02062610, version 1

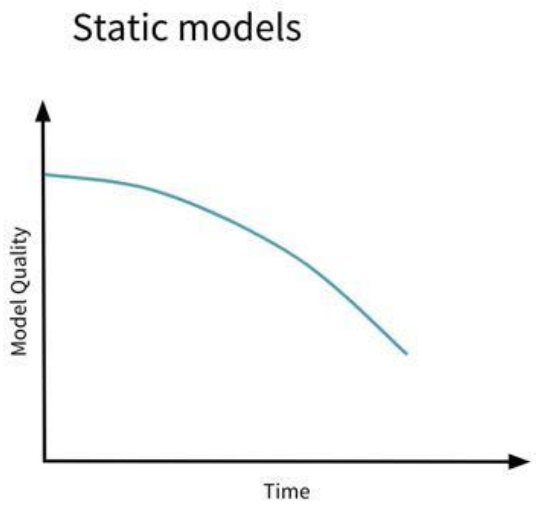
Concept drift

Examples:

Churn prediction: a new low-cost phone provider enters the market → the churn rate changes drastically even if the input distributions remain the same.

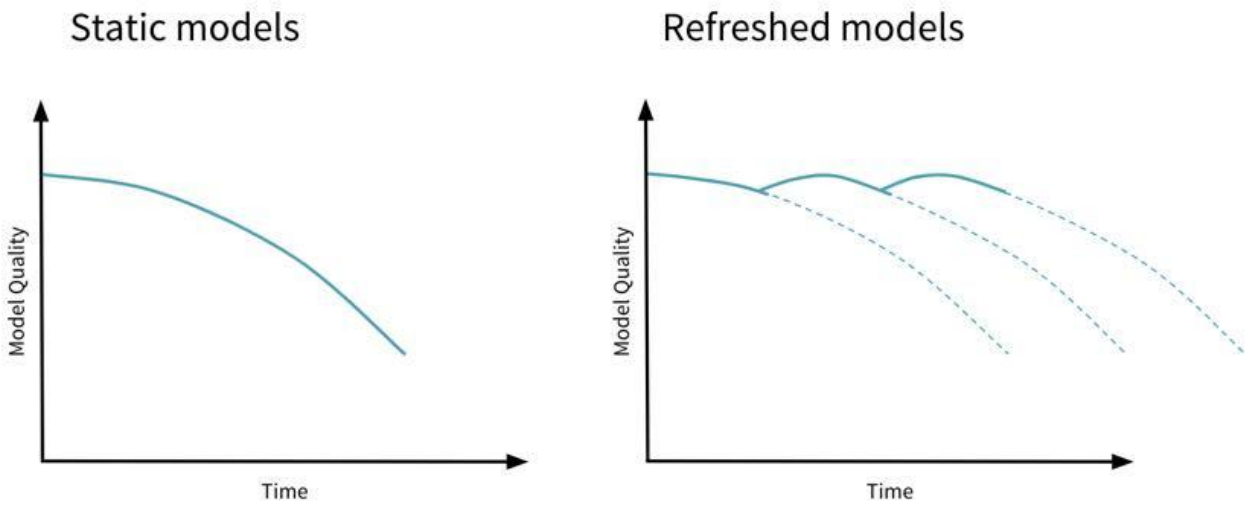
Sentiment analysis: 'This song is sick!' would have had a completely different meaning before the '90s.

Concept and data drift

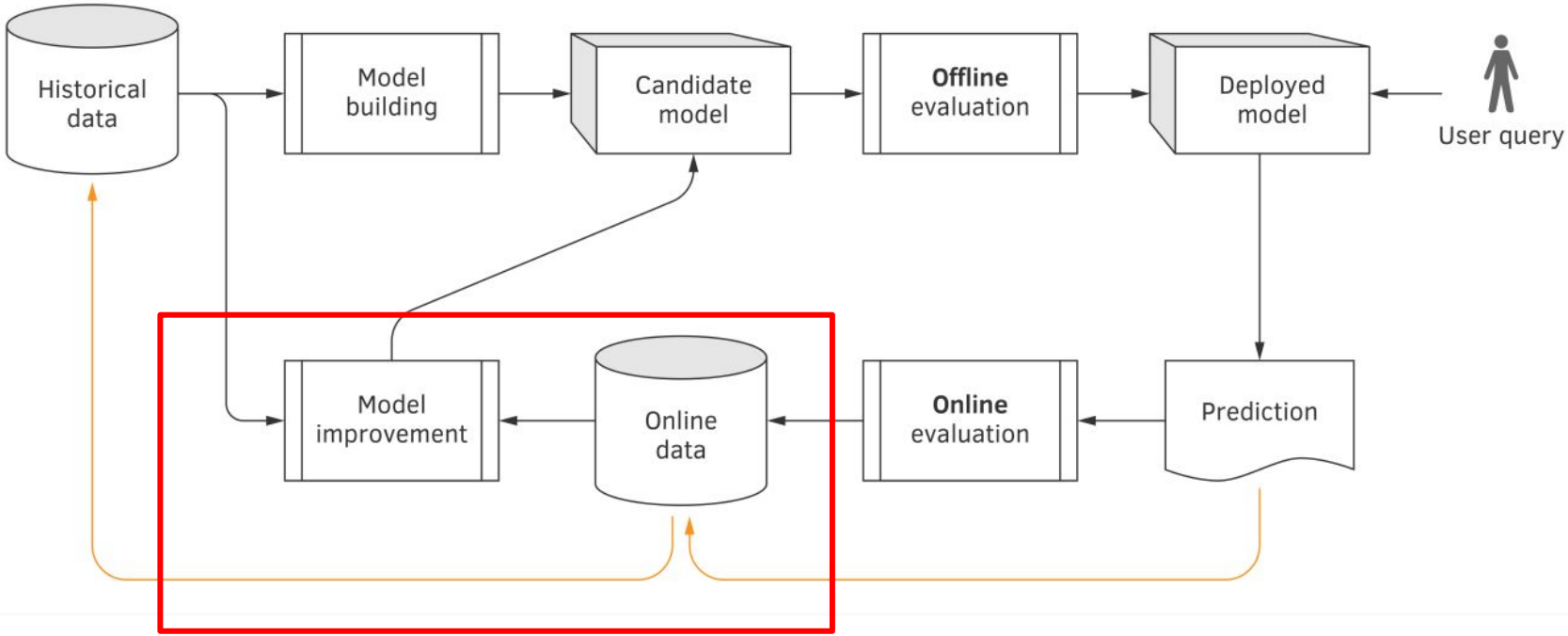


https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html

Concept and data drift



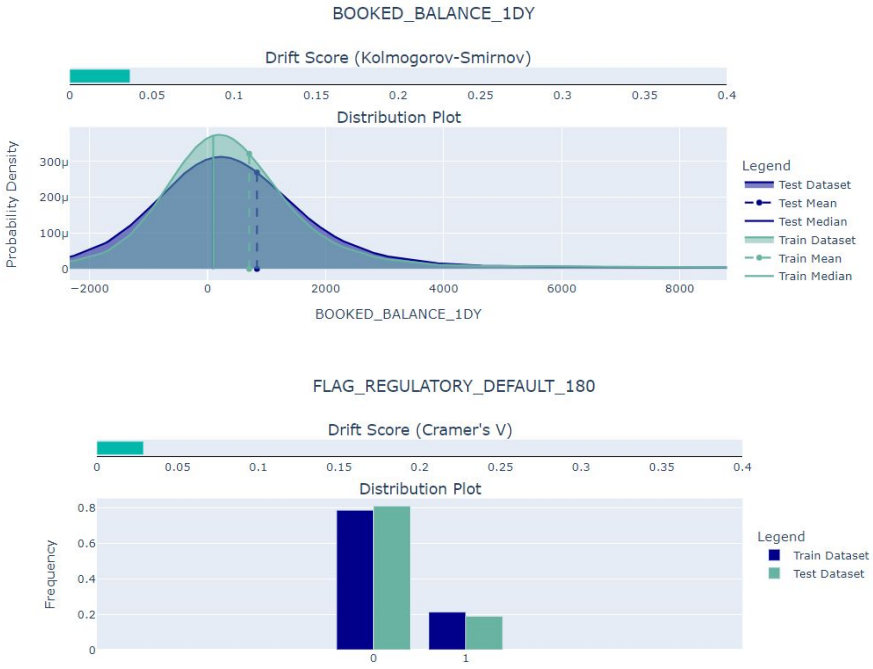
https://databricks.com/blog/2019/09/18/productionizing-machine-learning-from-deployment-to-drift-detection.html


Historical data
Model building
Candidate model
Offline evaluation
Deployed model
User query
Model improvement
Online data
Online evaluation
Prediction

Model monitoring example: deepchecks

Open-source library for validation and monitoring of machine learning models.

It allows for analysis in relation to univariate and multivariate distributions and to the performance of the model over time.



Putting it all together: model cards

Model cards

Comprehensive assessment to ensure that models are reproducible and adequately tested.

Different levels:

- Model reporting and reproducibility
- Performance analysis
- Model behaviour analysis
- Robustness check

https://huggingface.co/bigscience/bloom

https://huggingface.co/CompVis/stable-diffusion

Model auditing

Comprehensive assessment to ensure that models are reproducible and adequately tested.

Different levels:

- Model reporting and reproducibility
- Performance analysis
- Model behaviour analysis
- Robustness check

For practical examples, please visit **www.aiaudit.org**

ML4H Auditing: From Paper to Practice

Luis Oala LUIS.OALA@IHHI.FRAUNHOFER.DE
Machine Learning Group, Fraunhofer HHI, Germany

Jana Fehr JANA.FEHR@HPI.DE
Machine Learning and Digital Health, Hasso-Plattner-Institute, Germany

Luca Gilli LUCA@CLEARBOX.AI
Clearbox AI Solutions, Italy

Pradeep Balachandran PBN.TVM@GMAIL.COM
Technical Consultant (Digital Health), India

Alixandro Werneck Leite ALIXANDROWERNECK@OUTLOOK.COM
Machine Learning Laboratory in Finance and Organizations, Universidade de Brasília, Brazil

Saul Calderon-Ramirez SACALDERON@ITCR.AC.CR
Centre for Computational Intelligence, De Montfort University, United Kingdom

Danny Xie Li DXIE@IC-ITCR.AC.CR
Instituto Tecnológico de Costa Rica, Costa Rica

Gabriel Nobis GABRIEL.NOBIS@HHI.FRAUNHOFER.DE
Machine Learning Group, Fraunhofer HHI, Germany

Erick Alejandro Muñoz Alvarado ERICKMATEC@ESTUDIANTEC.CR
Instituto Tecnológico de Costa Rica, Costa Rica

Giovanna Jaramillo-Gutierrez GJGUTIERREZ@PROTONMAIL.COM
Milan and Associates SPRL, Spain

Christian Matek CHRISTIAN.MATEK@HELMHOLTZ-MUENCHEN.DE
Department of Medicine III, LMU Klinikum and Institute of Computational Biology, Helmholtz Zentrum München, Germany

Arun Shroff ARUNSHROFF@GMAIL.COM
xtend.ai, U.S.

Ferath Kherif FERATH.KHERIF@CHUV.CH
Laboratory for Research in Neuroimaging, Lausanne University Hospital and University of Lausanne, Switzerland

Bruno Sanguinetti BRUNO.SANGUINETTI@DOTPHOTON.COM
Dotphoton AG, Switzerland

Thomas Wiegand THOMAS.WIEGAND@HHI.FRAUNHOFER.DE
TU Berlin and Fraunhofer HHI, Germany

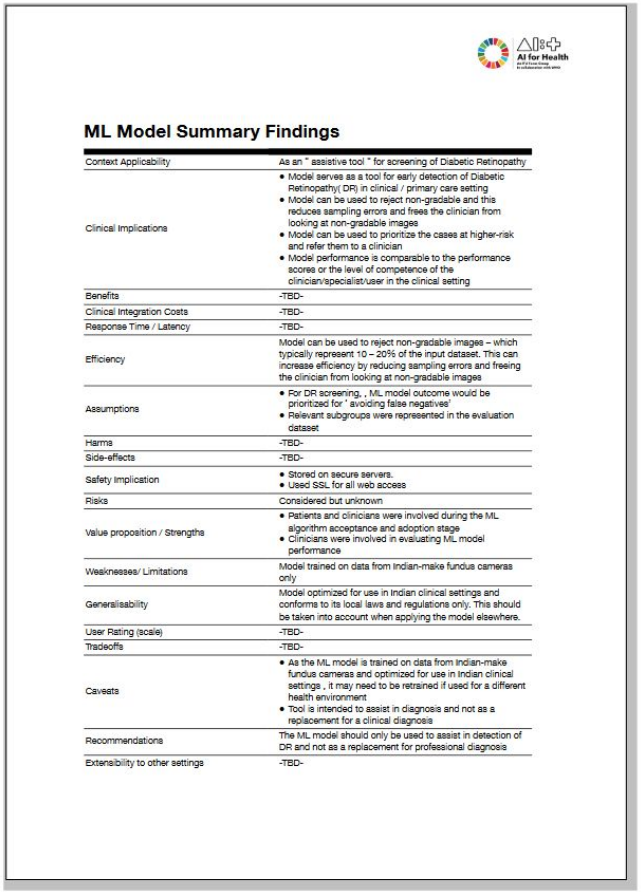
Editors: Emily Alsentzer⊗, Matthew B. A. McDermott⊗, Fabian Falck, Suproteem K. Sarkar, Subhrajit Roy‡, Stephanie L. Hyland‡

Reporting

Model summary contains:

- the results of a questionnaire filled by each stakeholder involved within the model development

- the outcome of the technical assessments possibly highlighting model limitations.



ML Model Summary Findings

Context Applicability	As an " assistive tool " for screening of Diabetic Retinopathy
Clinical Implications	• Model serves as a tool for early detection of Diabetic Retinopathy(DR) in clinical / primary care setting • Model can be used to reject non-gradable and this reduces sampling errors and frees the clinician from looking at non-gradable images • Model can be used to prioritize the cases at higher-risk and refer them to a clinician • Model performance is comparable to the performance scores or the level of competence of the clinician/specialist/user in the clinical setting
Benefits	-TBD-
Clinical Integration Costs	-TBD-
Response Time / Latency	-TBD-
Efficiency	Model can be used to reject non-gradable images – which typically represent 10 – 20% of the input dataset. This can increase efficiency by reducing sampling errors and freeing the clinician from looking at non-gradable images
Assumptions	• For DR screening, , ML model outcome would be prioritized for ' avoiding false negatives' • Relevant subgroups were represented in the evaluation dataset
Harms	-TBD-
Side-effects	-TBD-
Safety Implication	• Stored on secure servers. • Used SSL for all web access
Risks	Considered but unknown
Value proposition / Strengths	• Patients and clinicians were involved during the ML algorithm acceptance and adoption stage • Clinicians were involved in evaluating ML model performance
Weaknesses/ Limitations	Model trained on data from Indian-make fundus cameras only
Generalisability	Model optimized for use in Indian clinical settings and conforms to its local laws and regulations only. This should be taken into account when applying the model elsewhere.
User Rating (scale)	-TBD-
Tradeoffs	-TBD-
Caveats	• As the ML model is trained on data from Indian-make fundus cameras and optimized for use in Indian clinical settings , it may need to be retrained if used for a different health environment • Tool is intended to assist in diagnosis and not as a replacement for a clinical diagnosis
Recommendations	The ML model should only be used to assist in detection of DR and not as a replacement for professional diagnosis
Extensibility to other settings	-TBD-

Oala et al, ML4H Auditing: From Paper to Practice, NeurIPS2020

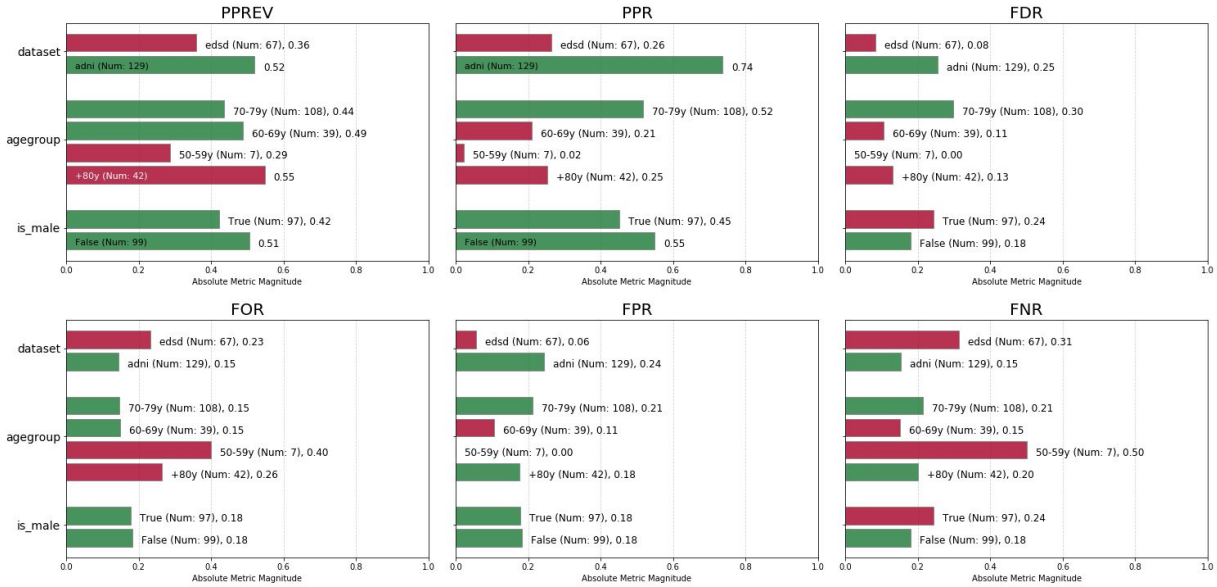
Bias and fairness

Model performance analysed using a stratified approach.

Data points are partitioned into groups of protected categories.

Challenge: data partitioning strategy needs to be defined by auditor.

For imaging problems, requires rich metadata.



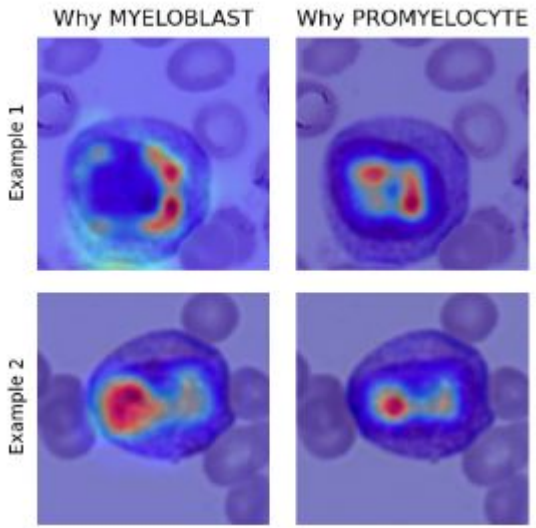
Oala et al, ML4H Auditing: From Paper to Practice, NeurIPS2020

Model behaviour analysis

Global model behaviour can be studied using **interpretability techniques**.

Local explanations are clustered to identify global model behaviours.

Challenge: choosing the right explanation technique for the particular use case.

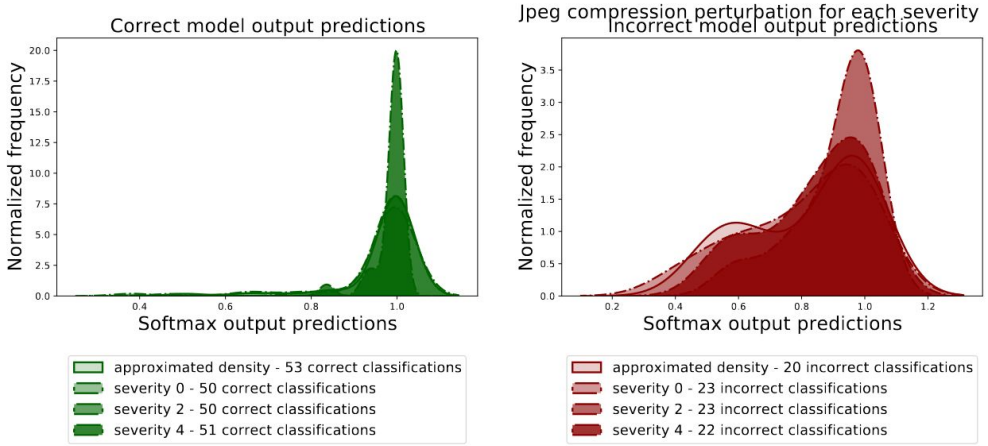


http://proceedings.mlr.press/v136/oala20a/oala20a.pdf

Robustness

Analyse model outputs when in presence of realistic input perturbations such as image compression artifacts.

Challenge: definition of realistic perturbations for each use case.



http://proceedings.mlr.press/v136/oala20a/oala20a.pdf

The case for model auditing from a research point of view

Translational Machine Learning

Machine Learning and Deep learning models have demonstrated to be able to outperform clinicians in several tasks.

Models and datasets are becoming increasingly available.

Yet, more than 90% of AI studies do not reach clinical testing. [1]

[1] Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter

Examples

NIH National Library of Medicine
National Center for Biotechnology Information

PMC PubMed Central® Search PMC Full-Text Archive Searc

Journal List > AMIA Jt Summits Transl Sci Proc > v.2020; 2020 > PMC7233077

Proceedings — AMIA Joint Summits on Translational Science AMIA

AMIA Jt Summits Transl Sci Proc. 2020; 2020: 191–200. PMCID: PMC7233077
Published online 2020 May 30. PMID: 32477638

A Review of Challenges and Opportunities in Machine Learning for Health

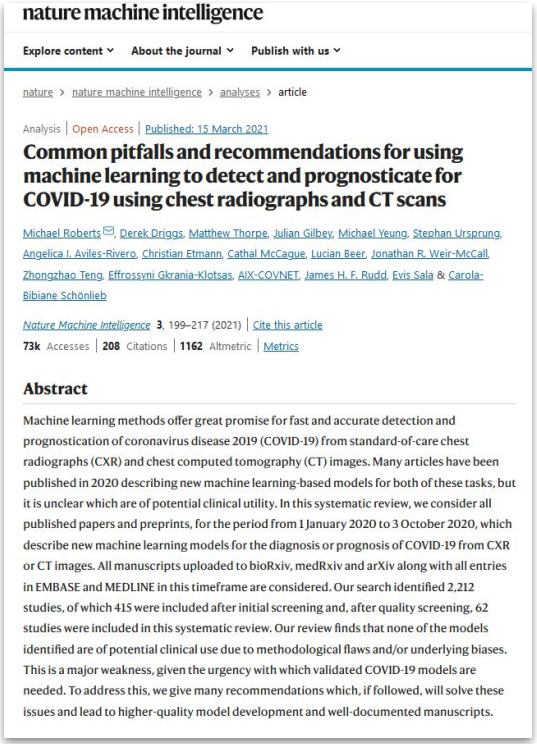
Marzyeh Ghassemi, PhD,[1] Tristan Naumann, PhD,[2] Peter Schulam, PhD,[3] Andrew L. Beam, PhD,[4] Irene Y. Chen, SM,[5] and Rajesh Ranganath, PhD[6]

• Author information • Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Abstract Go to: •

Modern electronic health records (EHRs) provide data to answer clinically meaningful questions. The growing data in EHRs makes healthcare ripe for the use of machine learning. However, learning in a clinical setting presents unique challenges that complicate the use of common machine learning methodologies. For example, diseases in EHRs are poorly labeled, conditions can encompass multiple underlying endotypes, and healthy individuals are underrepresented. This article serves as a primer to illuminate these challenges and highlights opportunities for members of the machine learning community to contribute to healthcare.



nature machine intelligence

Explore content About the journal Publish with us

nature > nature machine intelligence > analyses > article

Analysis | Open Access | Published: 15 March 2021

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, AIX-COVNET, James H. F. Rudd, Evis Sala & Carola-Bibiane Schönlieb

Nature Machine Intelligence **3**, 199–217 (2021) | Cite this article

73k Accesses | **208** Citations | **1162** Altmetric | Metrics

Abstract

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

npj | digital medicine

Explore content About the journal Publish with us

nature > npj digital medicine > review articles > article

Review Article | Open Access | Published: 12 April 2022

Machine learning for medical imaging: methodological failures and recommendations for the future

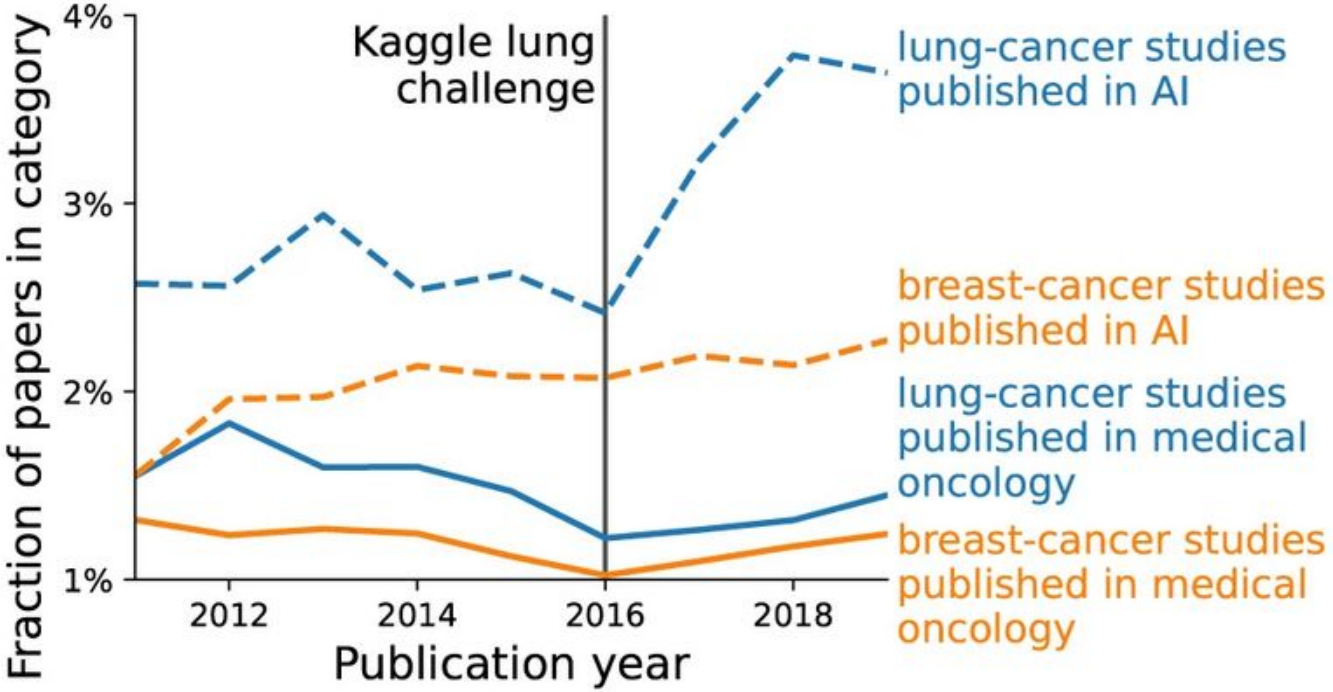
Gaël Varoquaux & Veronika Cheplygina

npj Digital Medicine **5**, Article number: 48 (2022) | Cite this article

15k Accesses | **1** Citations | **261** Altmetric | Metrics

Abstract

Research in computer analysis of medical images bears many promises to improve patients' health. However, a number of systematic challenges are slowing down the progress of the field, from limitations of the data, such as biases, to research incentives, such as optimizing for publication. In this paper we review roadblocks to developing and assessing methods. Building our analysis on evidence from the literature and data challenges, we show that at every step, potential biases can creep in. On a positive note, we also discuss on-going efforts to counteract these problems. Finally we provide recommendations on how to further address these problems in the future.



https://www.nature.com/articles/s41746-022-00592-y

nature machine intelligence

Article | Published: 31 May 2021

AI for radiographic COVID-19 detection selects shortcuts over signal

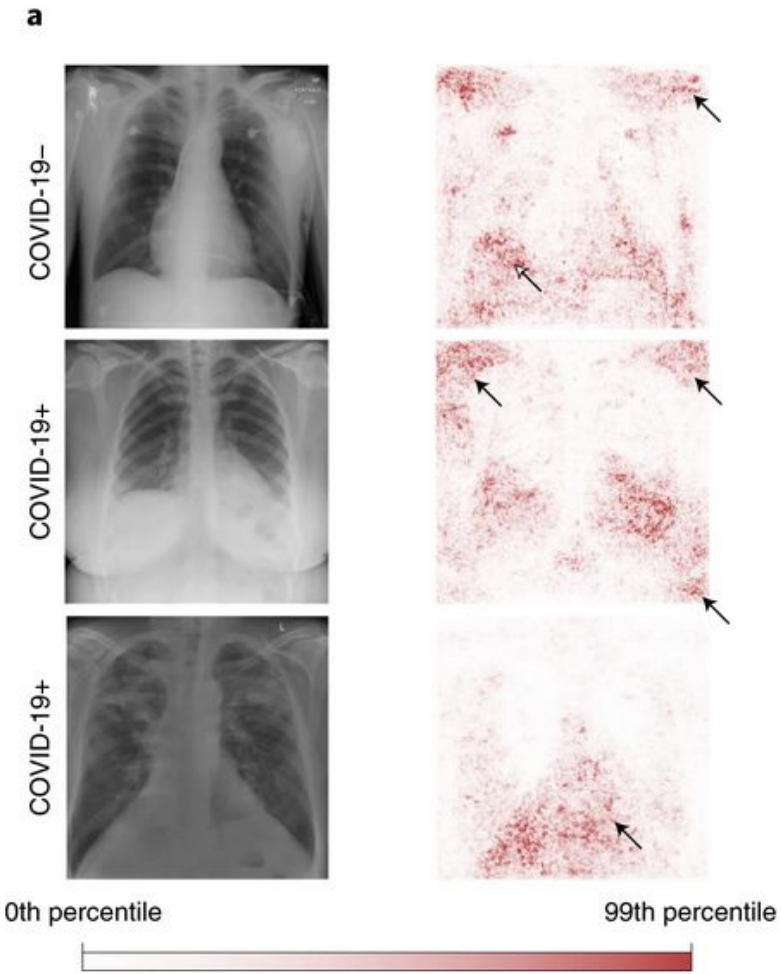
Alex J. DeGrave, Joseph D. Janizek & Su-In Lee

Nature Machine Intelligence **3**, 610–619 (2021) | Cite this article

23k Accesses | **69** Citations | **377** Altmetric | Metrics

Abstract

Artificial intelligence (AI) researchers and radiologists have recently reported AI systems that accurately detect COVID-19 in chest radiographs. However, the robustness of these systems remains unclear. Using state-of-the-art techniques in explainable AI, we demonstrate that recent deep learning systems to detect COVID-19 from chest radiographs rely on confounding factors rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals. We observe that the approach to obtain training data for these AI systems introduces a nearly ideal scenario for AI to learn these spurious 'shortcuts'. Because this approach to data collection has also been used to obtain training data for the detection of COVID-19 in computed tomography scans and for medical imaging tasks related to other diseases, our study reveals a far-reaching problem in medical-imaging AI. In addition, we show that evaluation of a model on external



https://www.nature.com/articles/s42256-021-00338-7/figures/2

Conclusions

Auditing AI models is a daunting task → Needs to be done **programmatically**.

Best approach is by using DevOps practices from start, in particular **automated testing**.

The process of defining and writing tests should be considered as an integral part of model development.

clearbox AI
humans && machines

Thanks for Reading

Feel free to contact us:

www.clearbox.ai

luca@clearbox.ai
giovannetti@clearbox.ai

@ClearboxAI

Backup slides, explainable AI

Explaining with examples

Explaining a prediction by pointing to the most similar examples from the training set

Property ID	#rooms	type	surface (m2)	#bathrooms	garder	**SalePrice**
n	**3**	**detached**	**150**	**2**	**yes**	**315000 €**

Explaining with examples

Explaining a prediction by pointing to the most similar examples from the training set

Property ID	#rooms	type	surface (m2)	#bathrooms	garder	SalePrice
n	**3**	**detached**	**150**	**2**	**yes**	**315000 €**

This property was sold within the last few months, it looks very similar to your current query

Property ID	#rooms	type	surface (m2)	#bathrooms	garder	SalePrice
424	**3**	**detached**	**140**	**2**	**yes**	**310000 €**

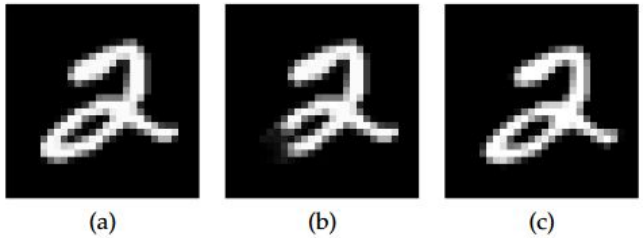
Explaining with examples

Explaining a prediction by pointing to the most similar examples from the training set

https://beenkim.github.io/papers/KIM2016NIPS_MMD.pdf

Issues with example-base explanations

- How to define similarity between points? →
- Sometimes example still needs to be explained (for example when in presence of many input features)



https://arxiv.org/abs/1606.05908

AGE	DAY	JOB	MONTH	BALANCE	CONTACT	EDUCATION	DURATION (S)	HOUSING LOAN	PERSONAL LOAN	CAMPAIGN CALLS	CREDIT DEFAULT	MARITAL STATUS	PREVIOUS CALLS	PREVIOUS OUTCOME
40	8	blue-collar	may	640	unknown	primary	347	yes	yes	2	no	married	0	unknown
36	23	blue-collar	may	655	unknown	primary	153	yes	no	4	no	married	0	unknown

Counterfactual examples

reasoning using hypothetical scenarios

Finding examples which are able to **change model prediction** in a certain direction following specific constraints.

Ex: *If you were 5 years older your loan would have been approved.*

Optimization problem similar to adversarial examples. **Can be extremely slow**.

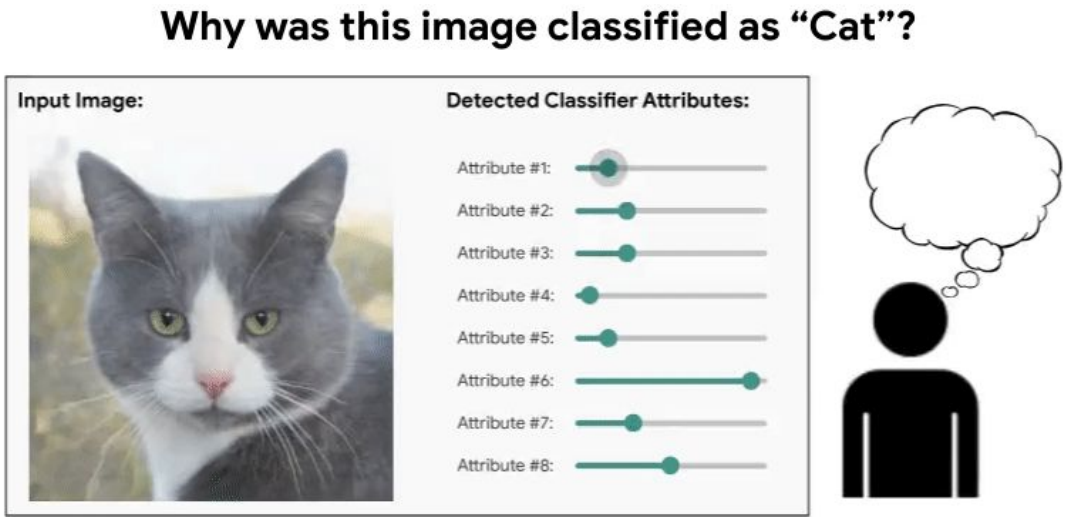
$$\|\mathbf{x} - \mathbf{x}'\| \leq \delta$$

$$|f(\mathbf{x}) - f(\mathbf{x}')| > \epsilon$$

Generative AI

Synthetic data has a lot of potential when it comes to model auditing.

→ Automated strategies to generate synthetic points can be very beneficial for explanations and robustness analysis.

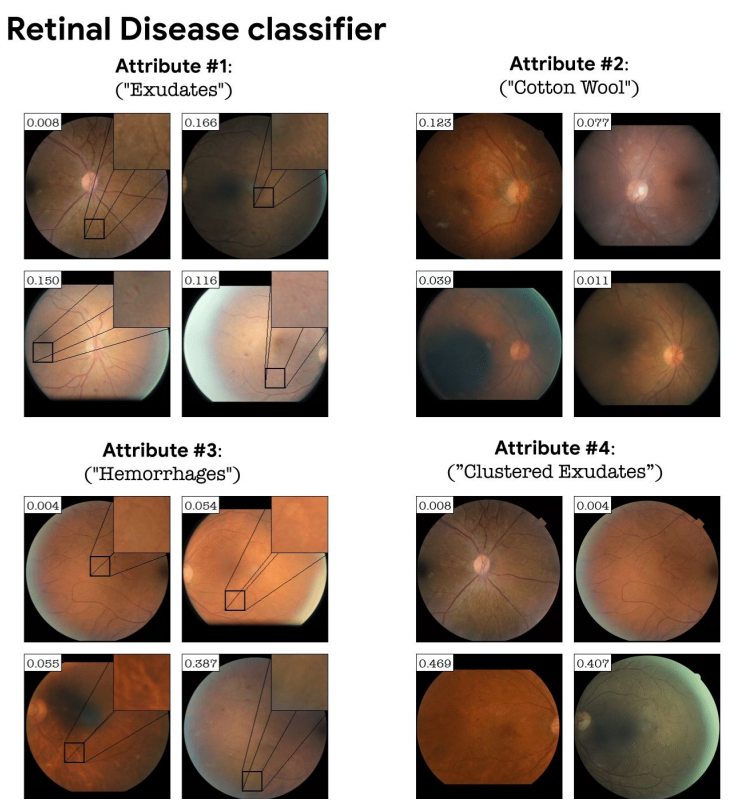


https://ai.googleblog.com/2022/01/introducing-stylex-new-approach-for.html

Generative AI

Synthetic data has a lot of potential when it comes to model auditing.

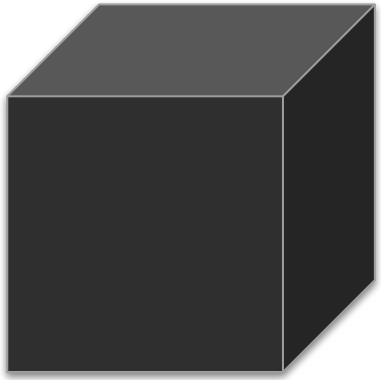
→ Automated strategies to generate synthetic points can be very beneficial for explanations and robustness analysis.



https://ai.googleblog.com/2022/01/introducing-stylex-new-approach-for.html

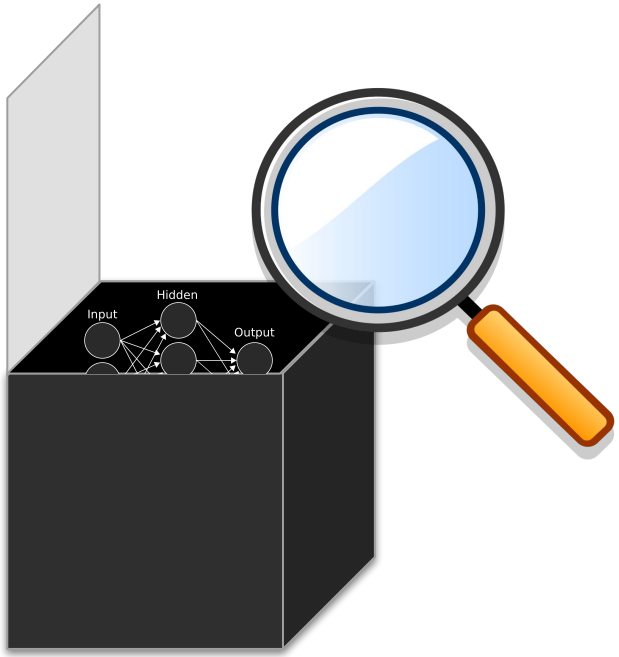
Intrusive methods

Opening the black-box



Intrusive methods

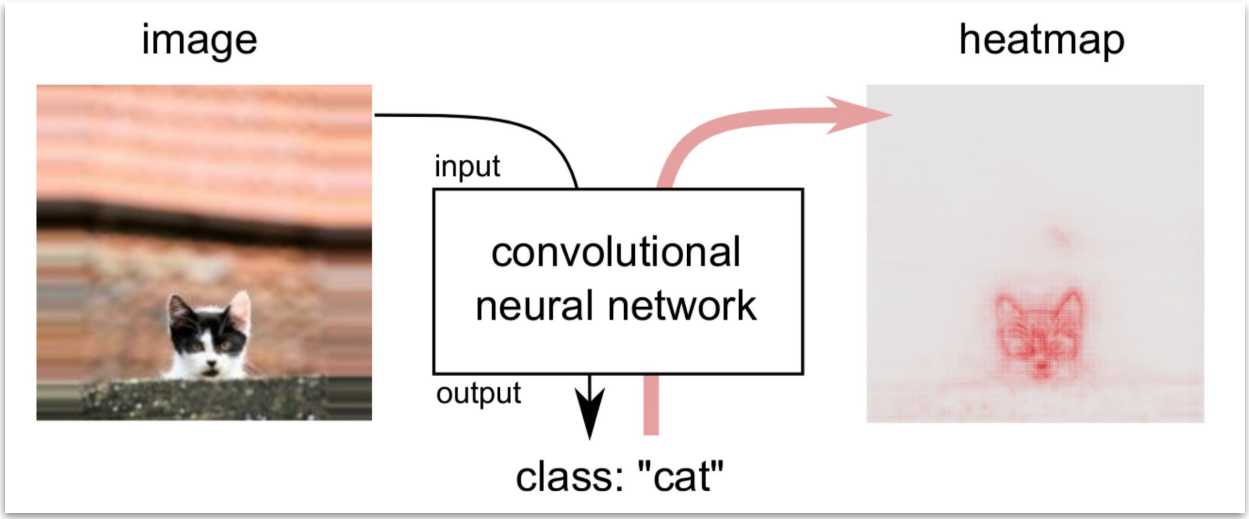
Opening the black-box



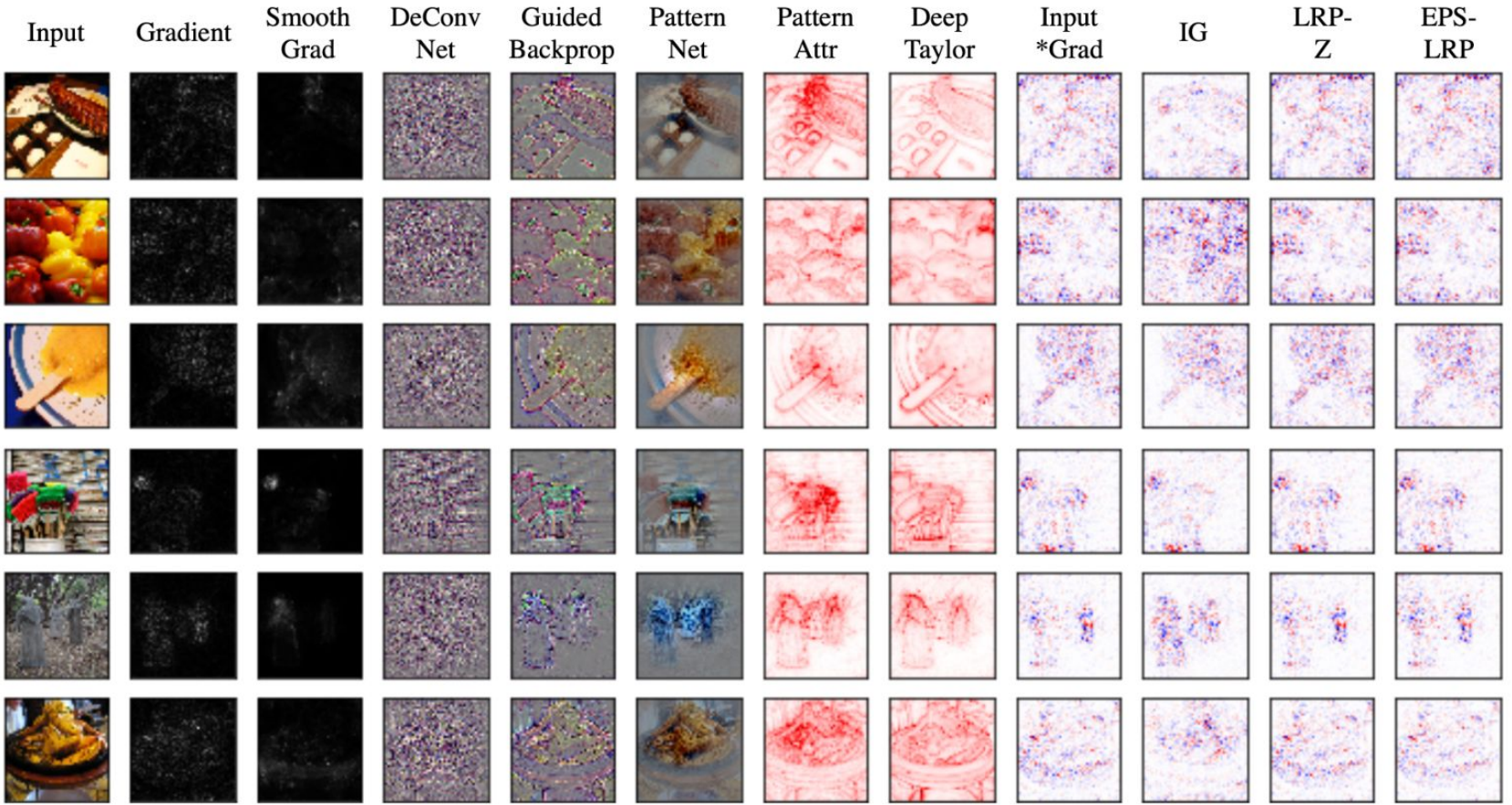
Intrusive methods

Opening the black-box

Example: Deep Taylor Decomposition



Source: Montavon et al. (ICML 2016)



Singh A, Sengupta S, Lakshminarayanan V. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*. 2020; 6(6):52. https://doi.org/10.3390/jimaging6060052

https://lrpserver.hhi.fraunhofer.de/image-classification

Issues with intrusive methods:

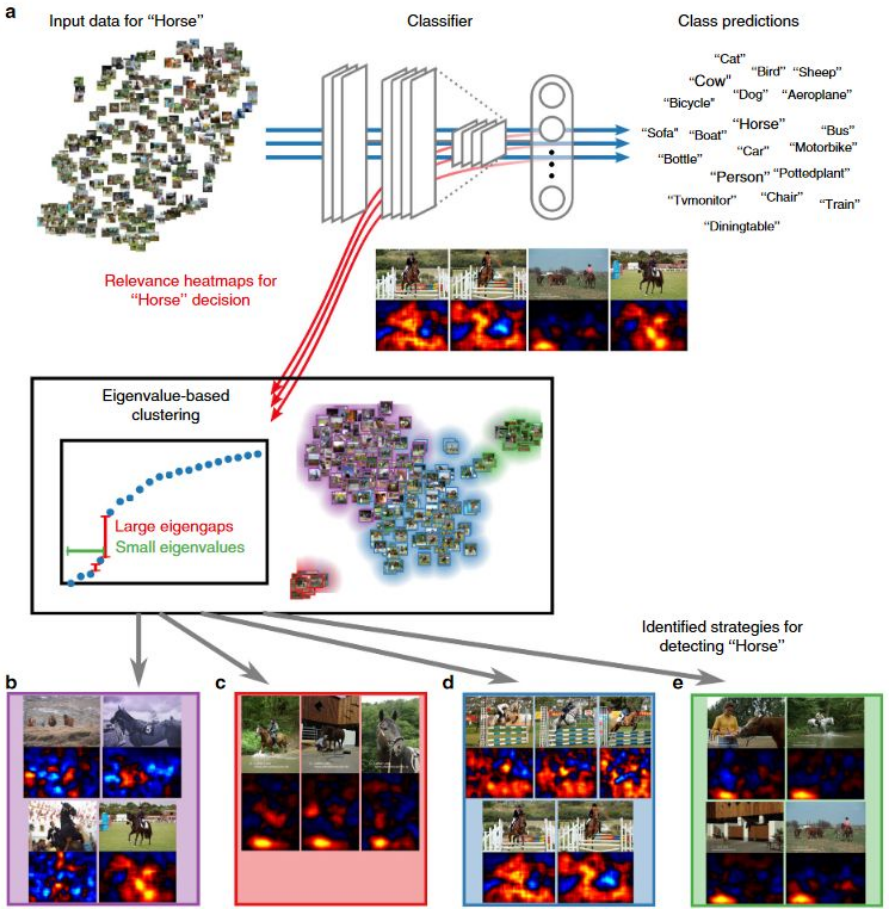
- Explanations can be either too noisy or too coarse
- Applying to existing models can be challenging due to library compatibility

From local to global explanations

It is possible to group local explanations according to their similarity.

Groups of explanations could be used to represent global model behaviour.

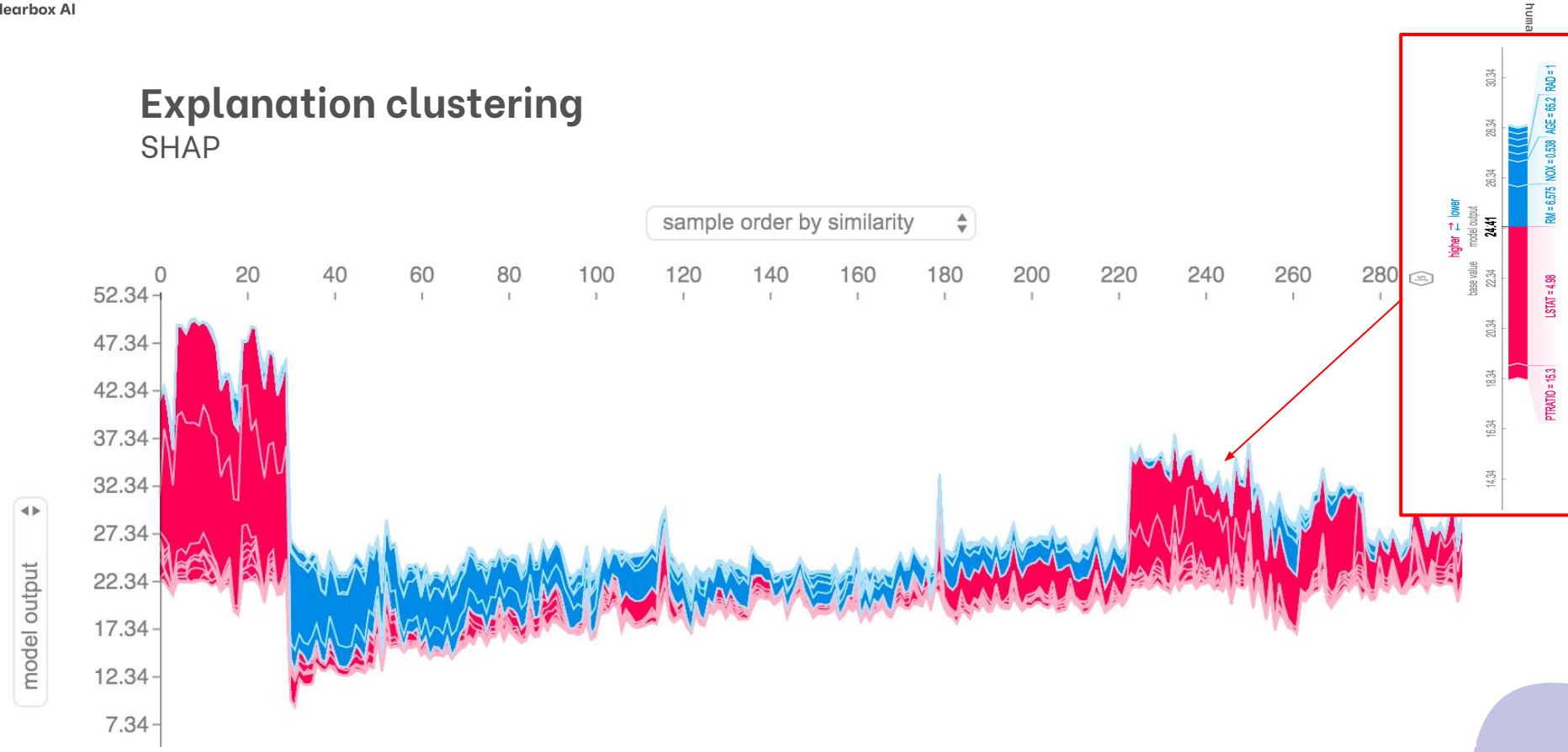
Could also help isolating undesired behaviour.



Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. Lapuschkin et al. Nature comm (2019)

Explanation clustering

SHAP



Model explainability
beyond debugging

IEEE SPECTRUM Engineering Topics Special Reports Blogs Multimedia The Magazine Professional Resources Search

Feature | Biomedical | Diagnostics

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

By **Eliza Strickland**

In 2014, IBM opened swanky new headquarters for its artificial intelligence division, known as IBM Watson. Inside the glassy tower in lower Manhattan, IBMers can bring prospective clients and visiting journalists into the "immersion room," which resembles a miniature planetarium. There, in the darkened space, visitors sit on swiveling stools while fancy graphics flash around the curved screens covering the

Model explainability
beyond debugging

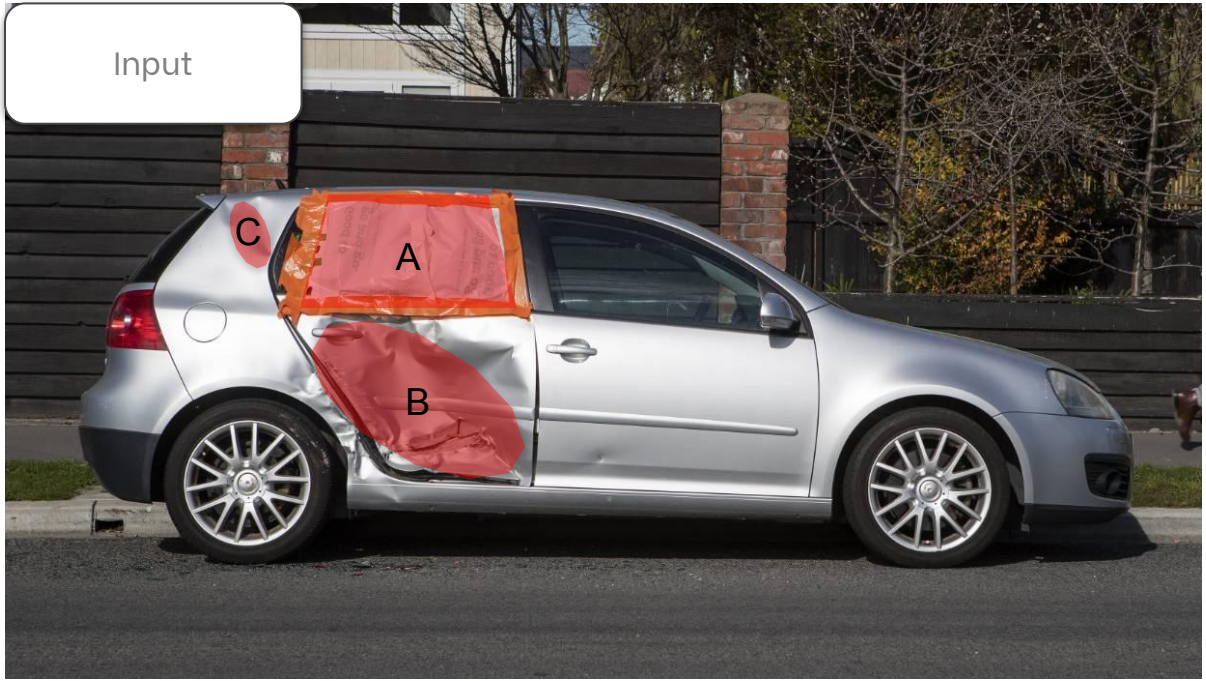
Input

Output:

Damage= 2000€

Model explainability
beyond debugging

Input



Output:
Damage = 2000€
Breakdown:
A = 300€
B = 1200€
C = 500€

Historical Example for B
Label = 2500€