TÉCNICO+
FORMAÇÃO AVANÇADA

Model explainability

Local and global explanations for descriptive and predictive models

DASH: Data Science e Análise Não Supervisionada
Rui Henriques, rmch@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa

Why eXplainable AI?

- Can we trust AI agents?
  - let us view *AI agents* as agents that use predictors to act and descriptors to organize knowledge
- Critical systems… and not only…

The Big Read Artificial intelligence + Add to myFT

Insurance: Robots learn the business of covering risk

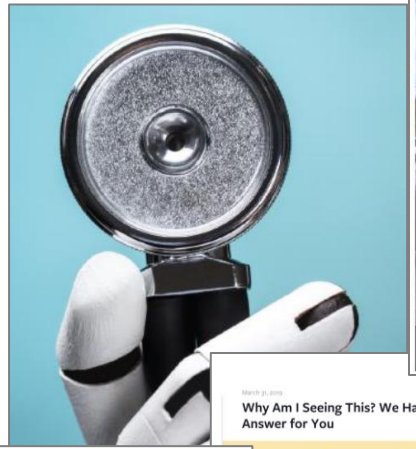
Artificial intelligence could revolutionise the industry but may also allow clients to calculate if they need protection

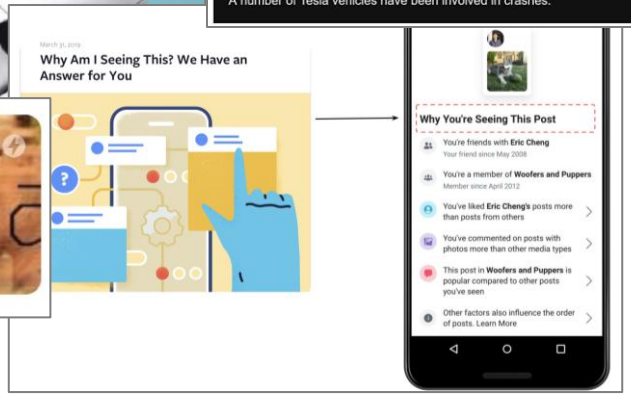
The Sun

ROBO-DOC World's first AI hospital unveiled in China with robot doctors who 'can treat 3,000 patients A DAY & will save millions'

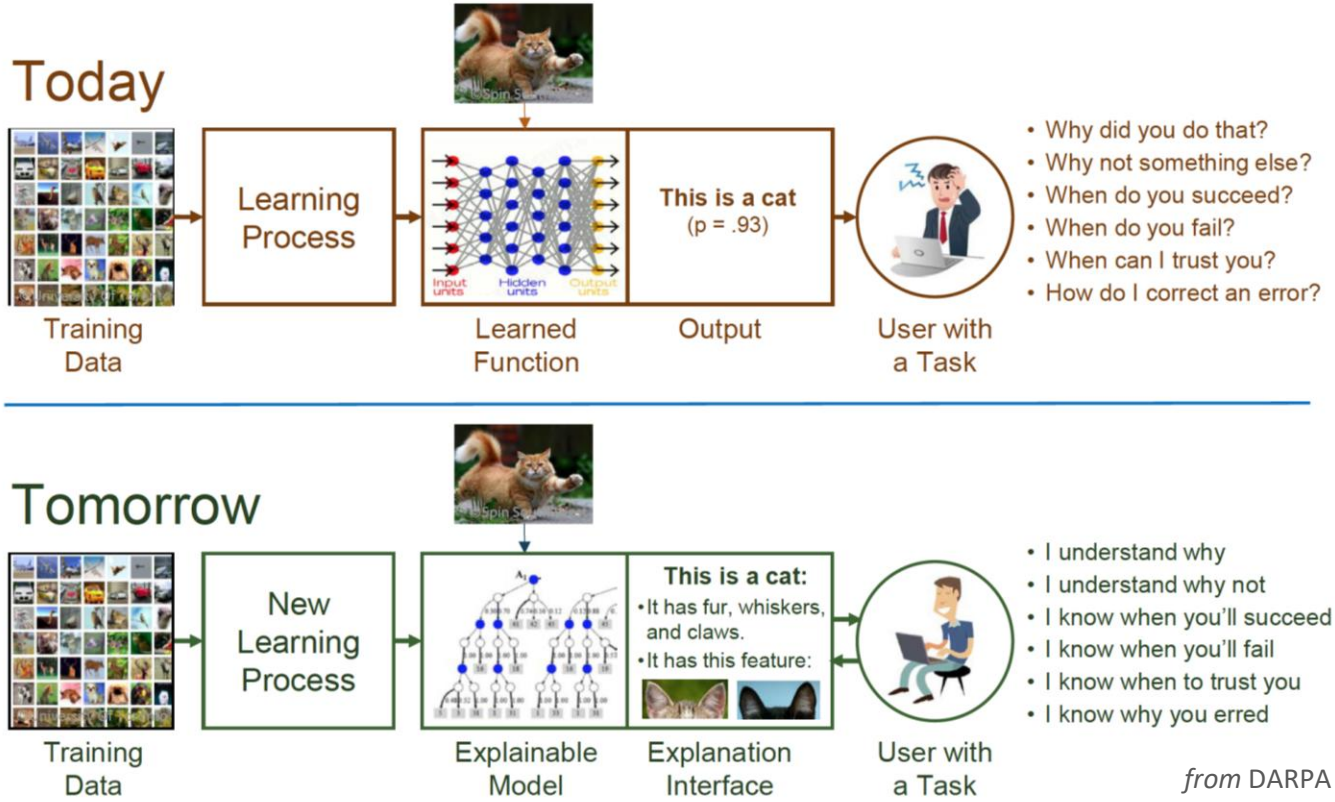
Why eXplainable AI?

- Autonomous…
  - crashes
  - misdiagnoses
  - misinfluence
  - hate





Why eXplainable AI?



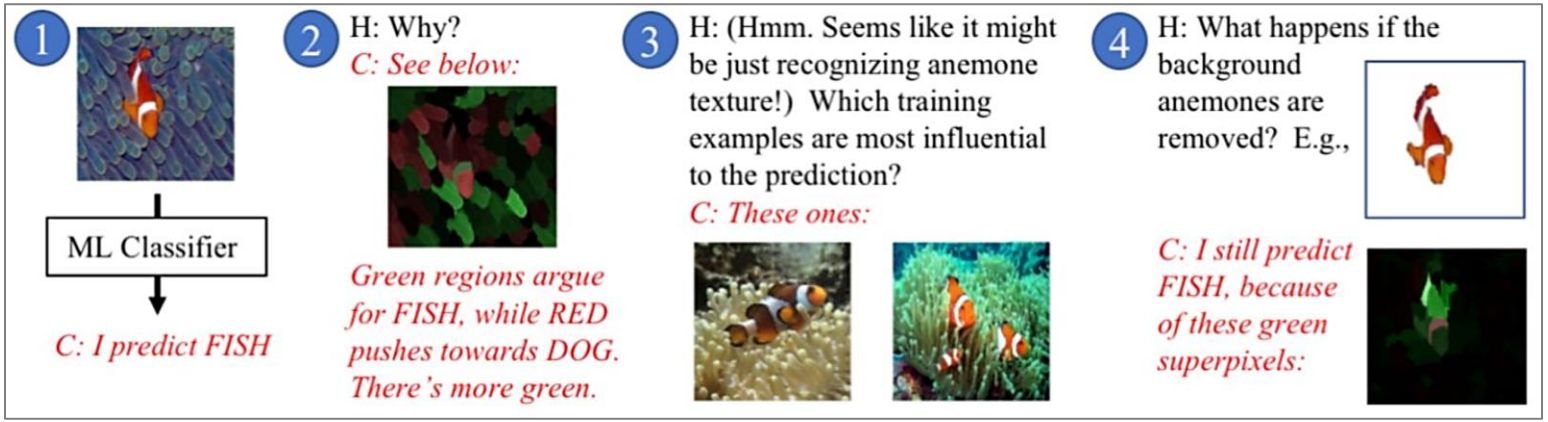
from DARPA

Outline

- **Responsible AI**
- Explainability
  - *descriptive models*
  - *predictive models*
- Post-hoc explanations
- Advances
- Evaluating XAI
- Going beyond

Interpretability *vs* explainability

- **Explainability**: ability to *<u>understand</u>* AI behavior (by humans)
  - *socially situated* understanding (assume technical and domain knowledge gaps)



Weld & Bansal, *The challenge of crafting intelligible intelligence,* 2018

- **Interpretability**: ability to understand AI behavior via *<u>simple model inspection</u>*

Trust *versus* explainability

- **Responsible AI: trustable**
    - sound/effective
    - **explainable** and **interpretable – machine/human**
    - privacy-preserving
    - uncertainty-aware
    - bias-aware
    - traceable

- Yet… if we can *understand* a model, we can decide whether to trust it or not

Explanation properties

- **Faithfulness**: do explanations represent the reasoning behind the agent behavior?
- **Usability**: can end users easily assimilate the explanations?
- **Stability**: does similar instances have similar interpretations?
- **Plausibility**: is the explanation correct or believable given current knowledge?
- **Others**
  - *transparency*
  - *traceability*
  - *ability to question*
  - …

Why explainability?

- **Efficacy criteria** (generalization capacity, actionability) **no longer enough**!
  - right to explanation…

  - safety
  - non-discrimination
  - accountability: identify *agency* and *causes* to avoid future problem
- Yet… auxiliary criteria are often **hard to quantify**
  - *e.g.* impossible to enumerate all scenarios violating safety of an autonomous car
- No clear answers in *psychology* to:
  - what constitutes an explanation? what makes some better than others?

Why explainability?

- **Performance**
  - understanding why/how the agent behaves: aids fine-tuning and optimization
- **Enhanced control**
  - ability to rapidly identify and correct mistakes
    - understanding decision process reveals vulnerabilities and flaws

- **Compliance**:
  - ensure compliance with *company policies, industry standards, government regulations*
    - critical for data scientists, auditors, and decision-makers
  - article 14 of European data protection laws (GDPR): AI must provide meaningful information on the: logic involved, significance, envisaged consequences

Why explainability?

- **Model validation**
  - separate evidence from *biases*
    - agents may rely on misrepresented demographics to make decisions
    - explanations can aid in detecting such biases, preventing ethical and legal violation

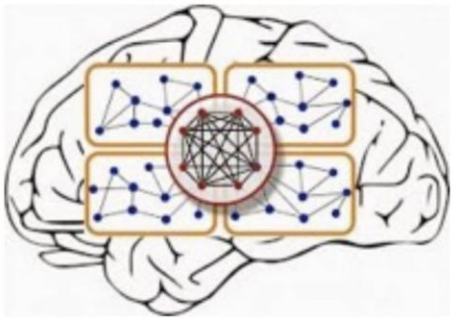
- **Model debugging**

(a) Husky classified as wolf (b) Explanation

by Guestrin et al. 2016

Why explainability?

- **Knowledge discovery**
  - allow humans to obtain new insights
  - domain experts and end users can provide realistic feedback



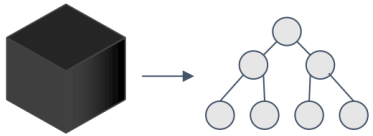
MICCAI'18

"It's not a human move. I've never seen a human play this move," says European GO champion Fan Hui. "So beautiful."

Outline

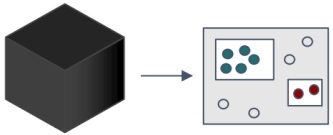
- Responsible AI
- **Explainability**
  - *descriptive models*
  - *predictive models*
- Post-hoc explanations
- Advances
- Evaluating XAI
- Going beyond

Scope of Explainability



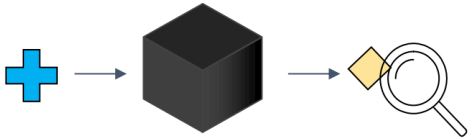
Global

How the model globally works



Subgroup

How the model behaves in data subgroups

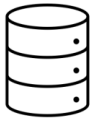


Individual/local

Explaining the reasons behind individual predictions

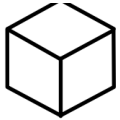
by Eliana Pastor

Explainability stages



Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering



Explainable modeling

Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

After model development
- Explaining predictions and behavior of trained models

by Eliana Pastor

Pre-modeling explainability (Kim et al.)

- **prototypes**: representative in training
- **criticisms**: instances not well represented by the set of prototypes

Kim et al., *Examples are not Enough, Learn to Criticize! Criticism for Interpretability*. NIPS 2016

- detection using MMD-critic framework

Model-based *vs* model-agnostic

Explainability can be…

- **model-specific.** Examples:
    - integrated gradients for deep learning
    - tree visualization for associative predictors (e.g., decision trees) and descriptors

- **model-independent** (applicable to any predictor or descriptor). Examples:
    - LIME – Locally Interpretable Model Agnostic Explanations
    - SHAP – Shapley Values

Explaining AI behavior

- ***Description*: knowledge acquisition**
  - pattern discovery, modeling, clustering, anomaly analysis, summarization, representation…
  - *note*: can be supervised (e.g., discriminative patterns, annotation-aware clustering)

- ***Prediction*: decision support**
  - classification, regression, signal synthesis, prompt answering, forecasting, captioning…

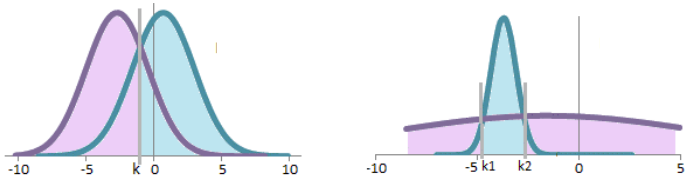
- ***Prescription***: **system optimization** (*control*)
  - data-centric simulation, multi-agent decision, reinforcement learning…

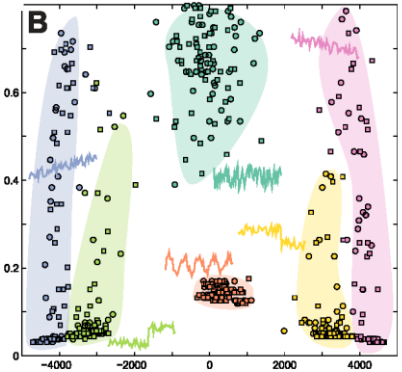
Outline

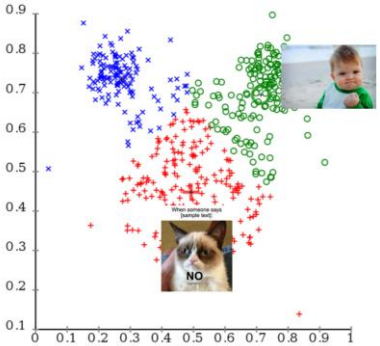
- Responsible AI
- **Explainability**
  - ***descriptive models***
  - *predictive models*
- Post-hoc explanations
- Advances
- Evaluating XAI
- Going beyond

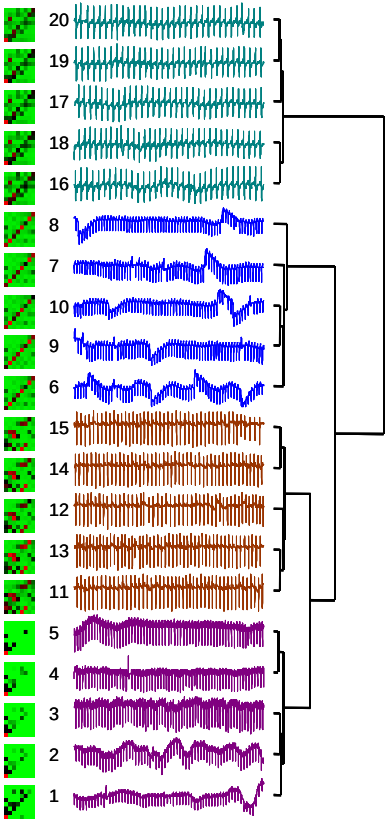
Explaining clustering

- describe **centroids**
  - **medoid** prototype (e.g., representative series/image/text) or barycenter
  - single vs. multiple anchors
  - **generative** summarization
  - **cluster-conditional distributions** of most discriminative variables





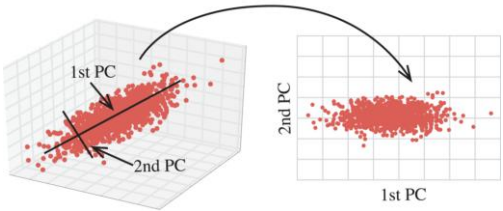


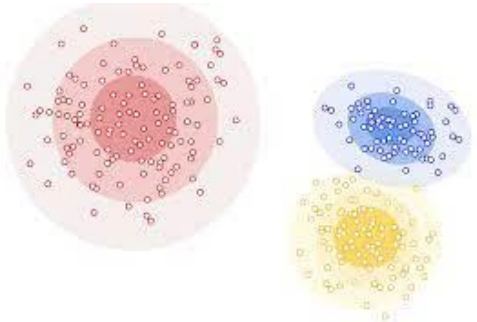


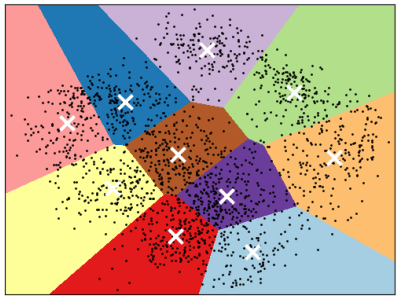
- dendrogram (hierarchical clustering)

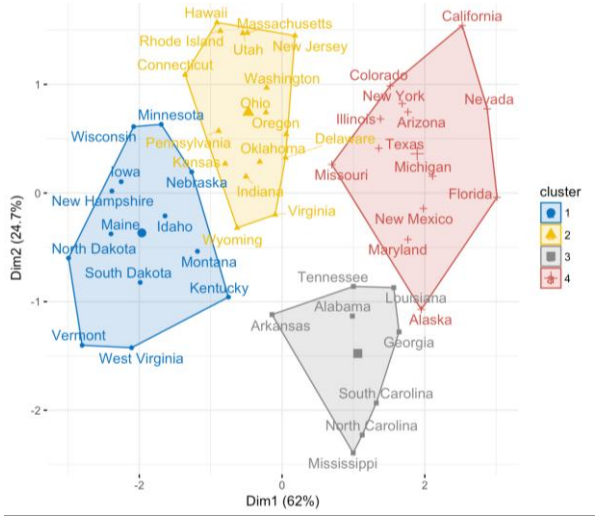
Explaining clustering

- select informative features or reduce dimensionality (uMAP, PCA, embedding)
- delineate clusters by:
  - plotting **multivariate distribution** (e.g. EM)
  - **partitioning space** (e.g. *k*-means)
  - identifying **shape** (e.g. DBSCAN)

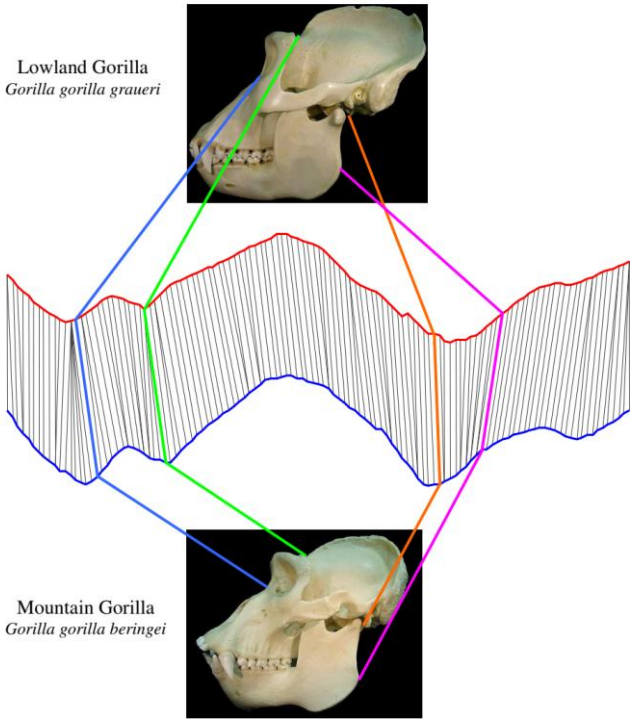


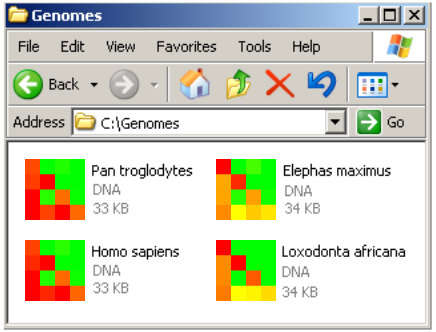




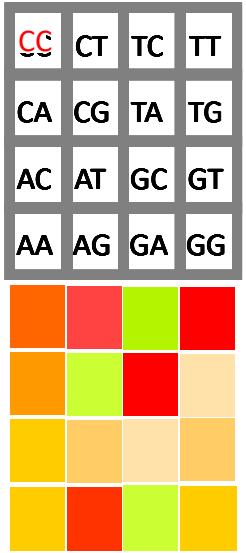


Explaining clustering (Keogh et al.)



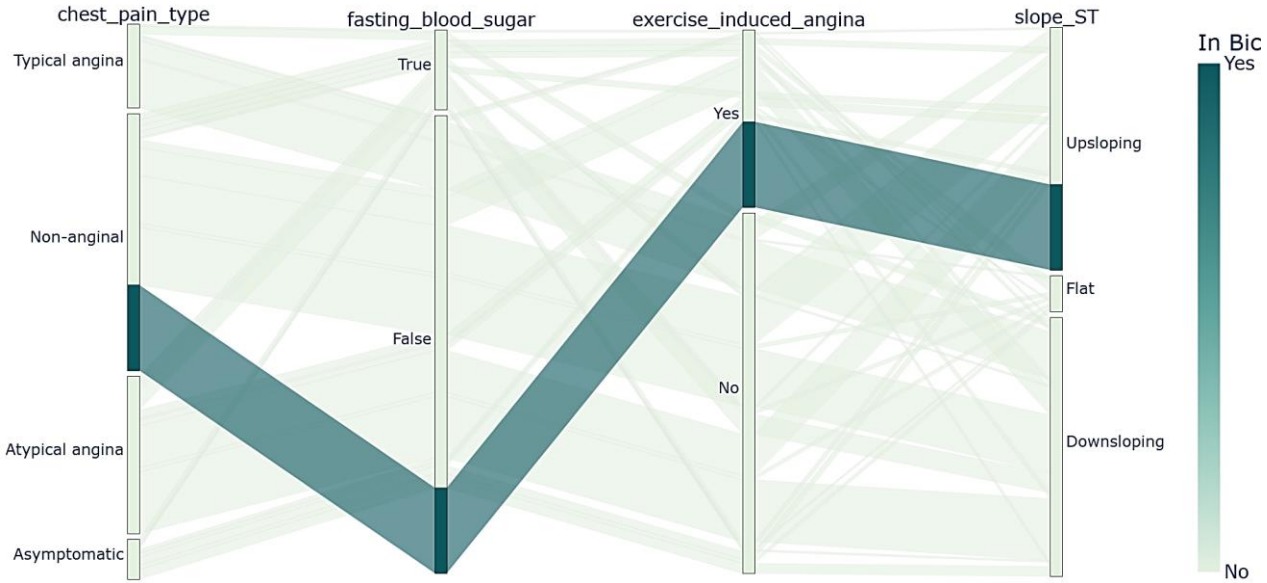


CCGTGCTAGGGCCACCTACCTTGGTCCGCCGCAAGCT
CATCTGCGCGAACCAGAACGCCACCACCTTGGGTTGA
AATTAAGGAGGCGGTTGGCAGCTTCCAGGCGCACGT
ACCTGCGAATAAATAACTGTCCGCACAAGGAGCCGA
CGATAAAGAAGAGAGTCGACCTCTCTAGTCACGACCT
ACACACAGAACCTGTGCTAGACGCCATGAGATAAGC



Explaining patterns (Gonçalves et al.)

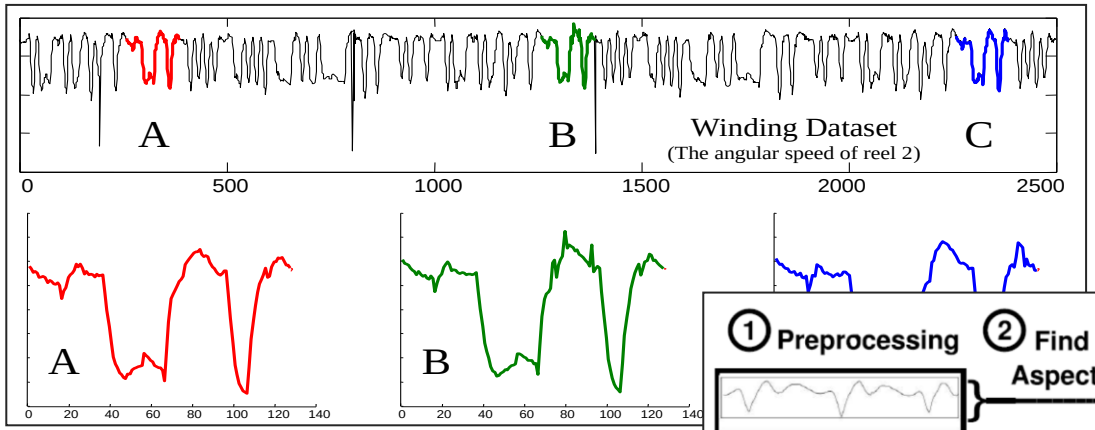
- situating **patterns against global data regularities**
- assessing *quality*, *statistical significance*, *discriminative power*…



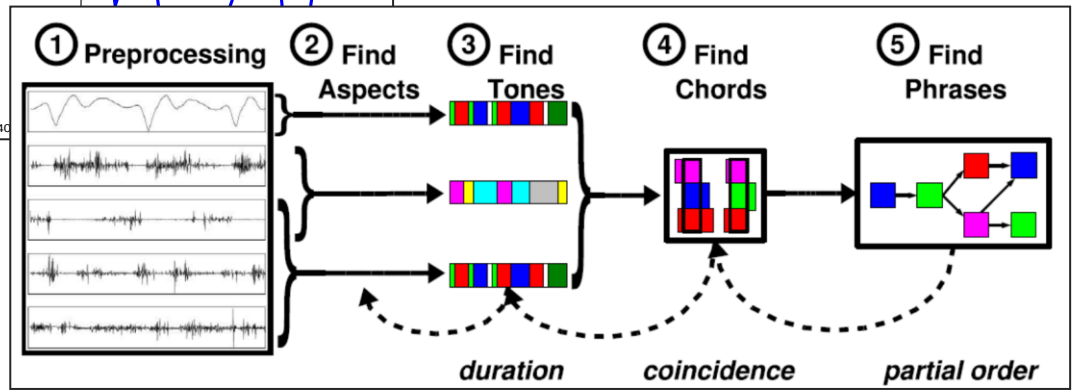
P-value = 2.568e-09
Lift: absence = 0.209
presence = 1.988

https://github.com/danielmg97/bic_parallel_coords

Explaining temporal patterns

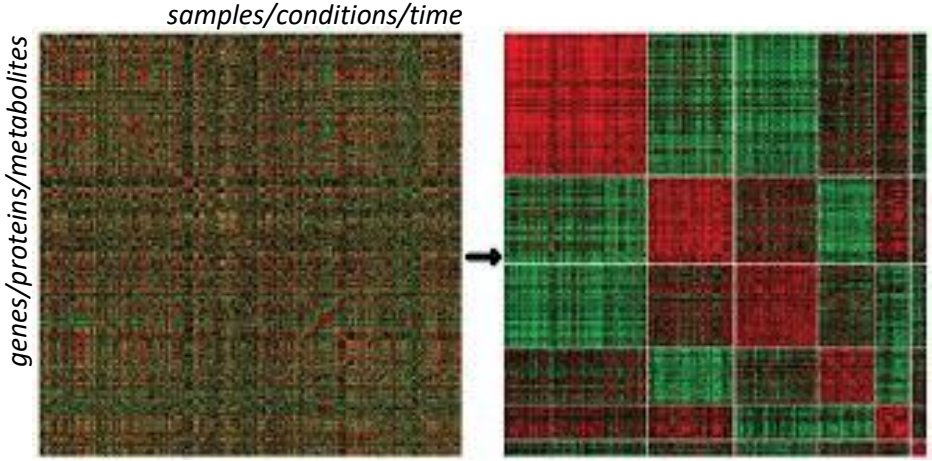


Motif discovery (Keogh et al.)

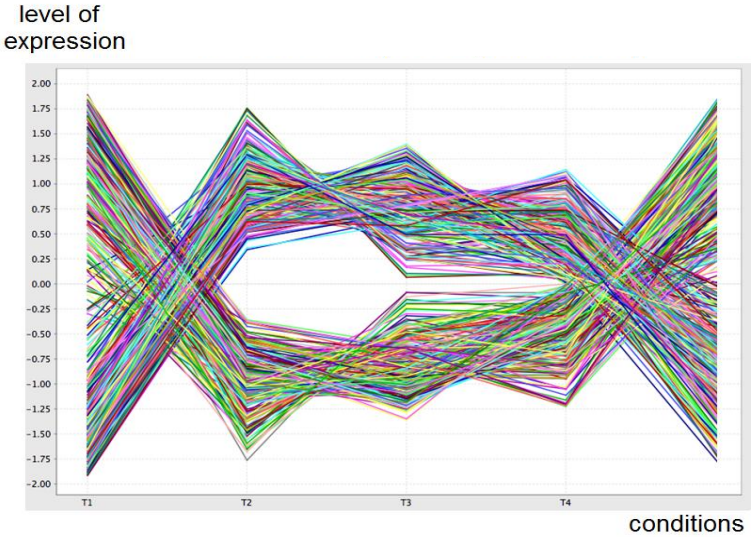


Event patterns (Mörchen et al.)

Explaining omic patterns

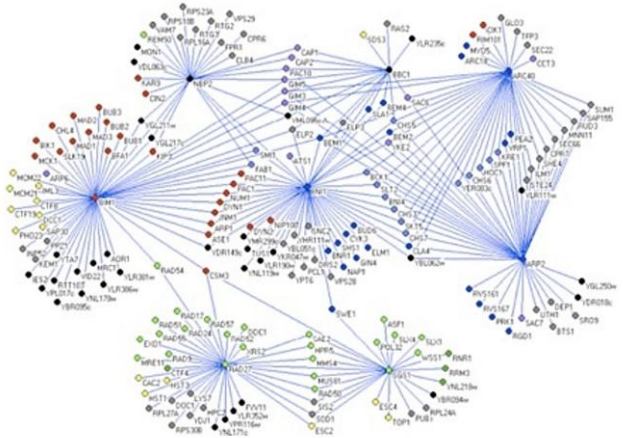


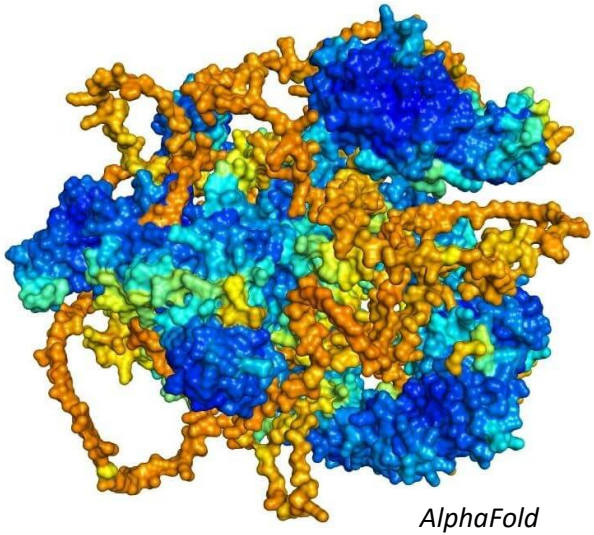
Low variance: approximate constant values



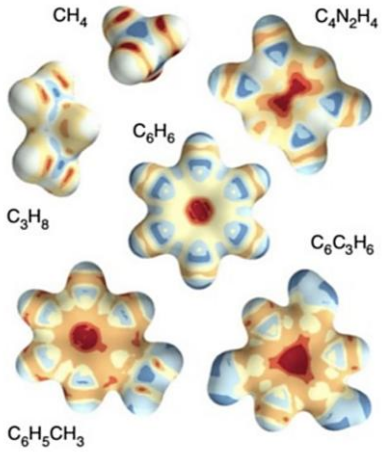
Additive and symmetric factors

Explaining structural patterns

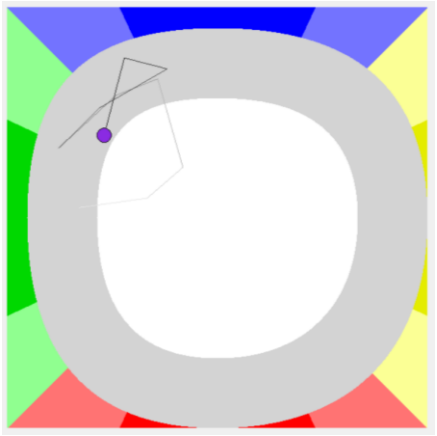


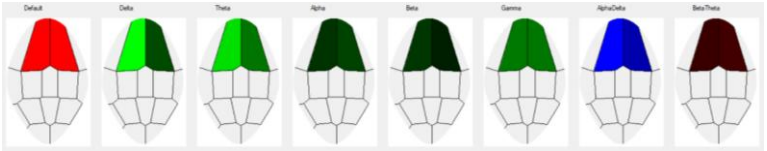


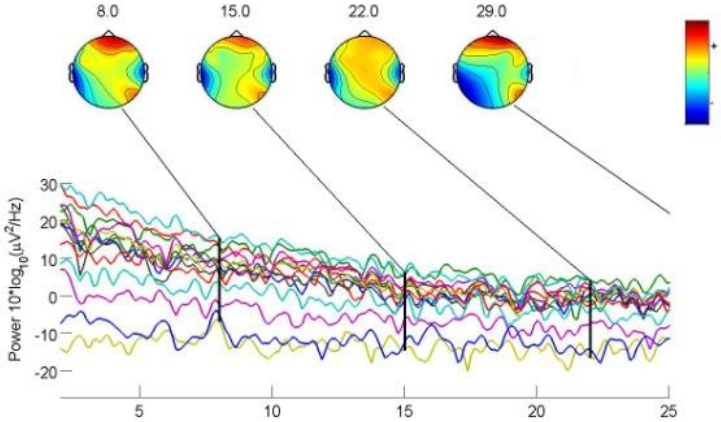
AlphaFold



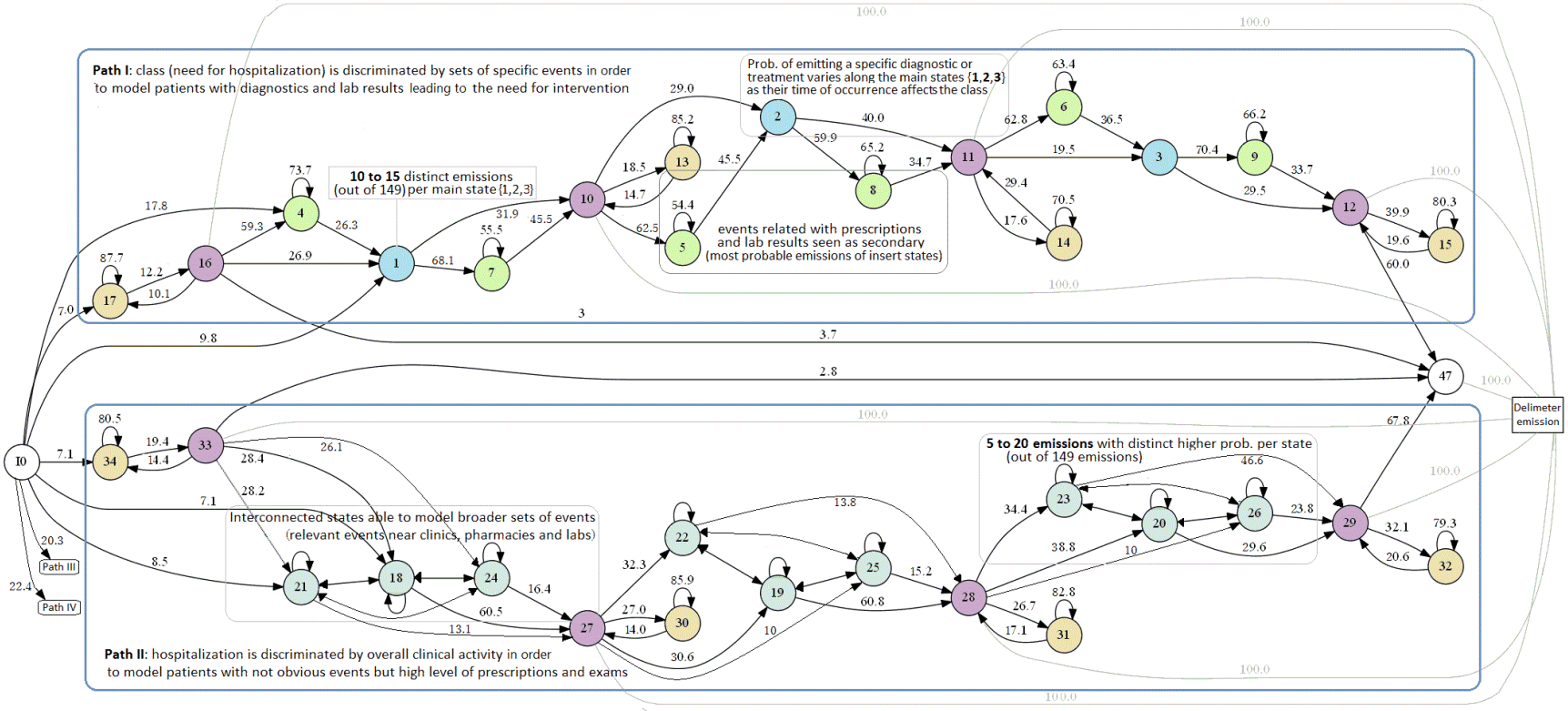
Explaining brain patterns (Henriques et al.)







Explaining events (DBNs, HMMs…)



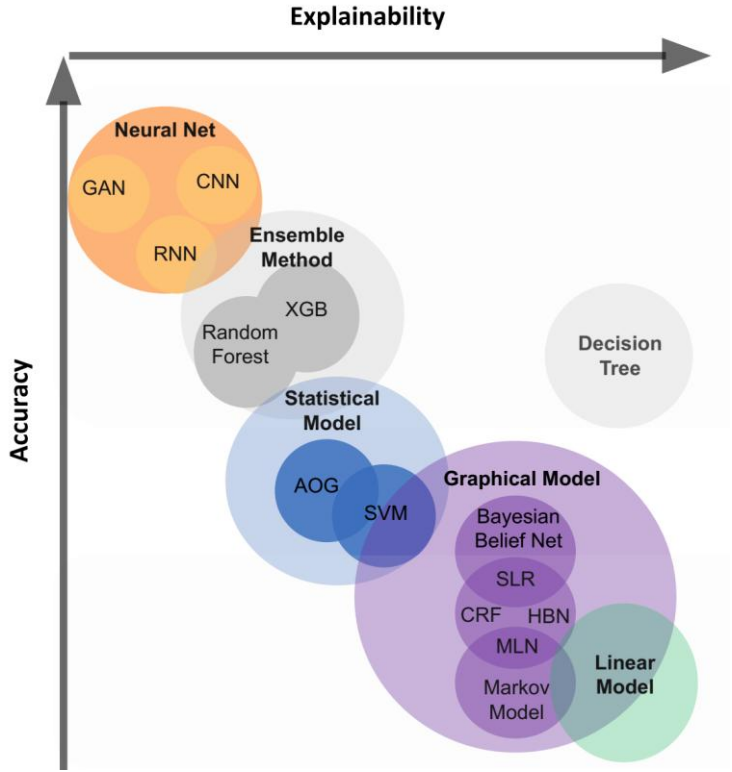
Outline

- Responsible AI
- **Explainability**
  - *descriptive models*
  - ***predictive models***
- Post-hoc explanations
- Advances
- Evaluating XAI
- Going beyond

Inherent *vs* post-hoc

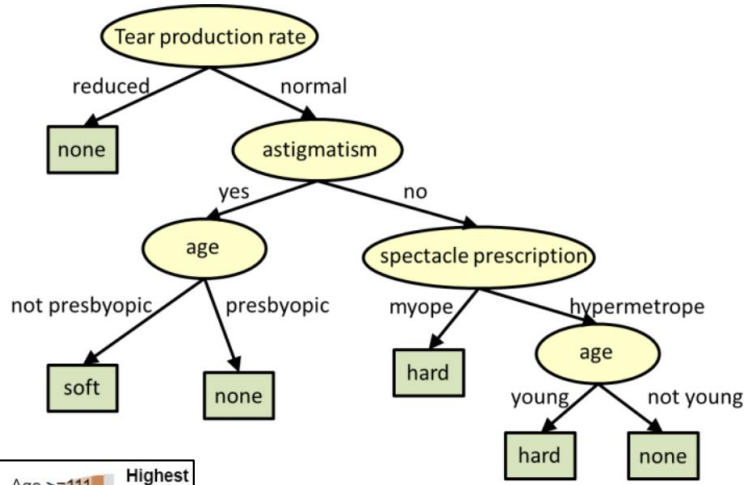
Is the explainability built into the model?

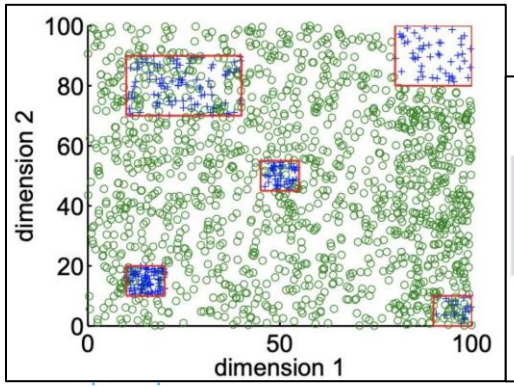
- models with ***inherent*** *interpretability*
- black-box models: we need an external method to understand them
  - ***post-hoc*** *explanations*

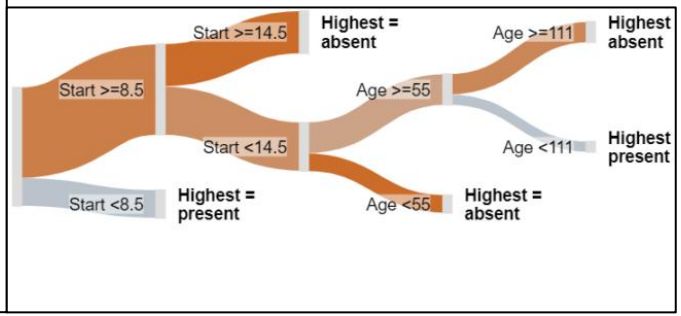


Decision trees

- Global explanations
  - tree visualizations versus space partitioning
  - coloring of nodes in accordance with
    - class (classification) or ranges (regression)
    - predictive accuracy of the paths until leaves
- Local explanations by path/point highlighting

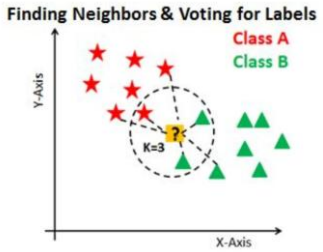


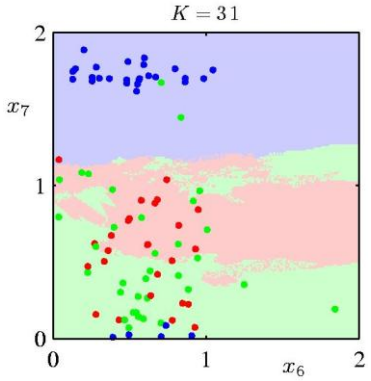


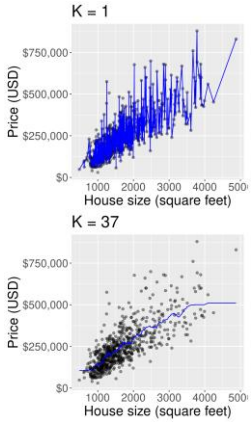


Others

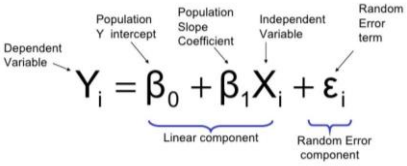
- Local learning (*on the right*)
- Linear models
- Rule-based models
- Concept-based models



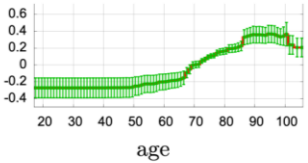




Linear models



GAMs, GA²Ms, GLMs



Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission." KDD 2015

Decision sets - Rules

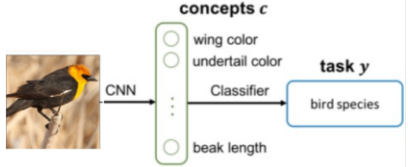
If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure≥ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

Lakkaraju et al. "Interpretable decision sets: A joint framework for description and prediction." KDD 2016

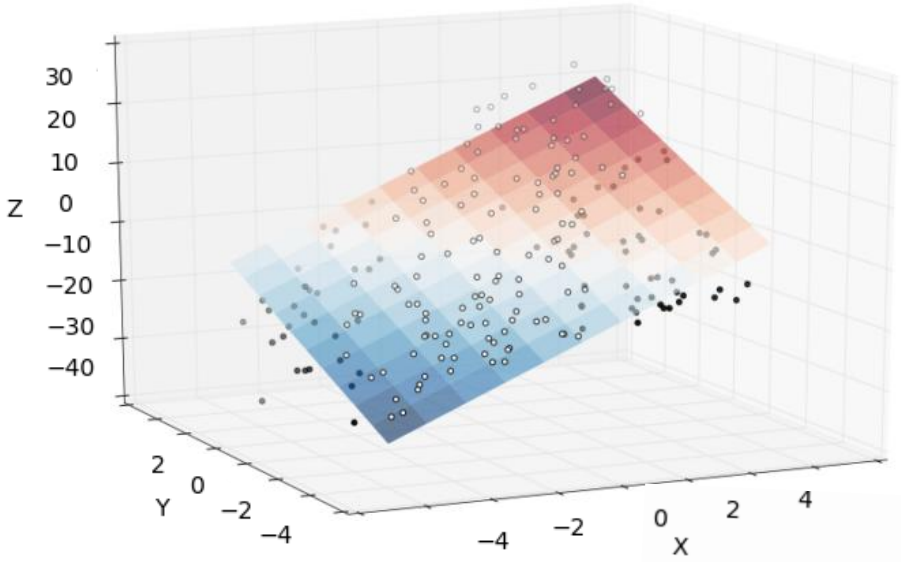
Concept-based models

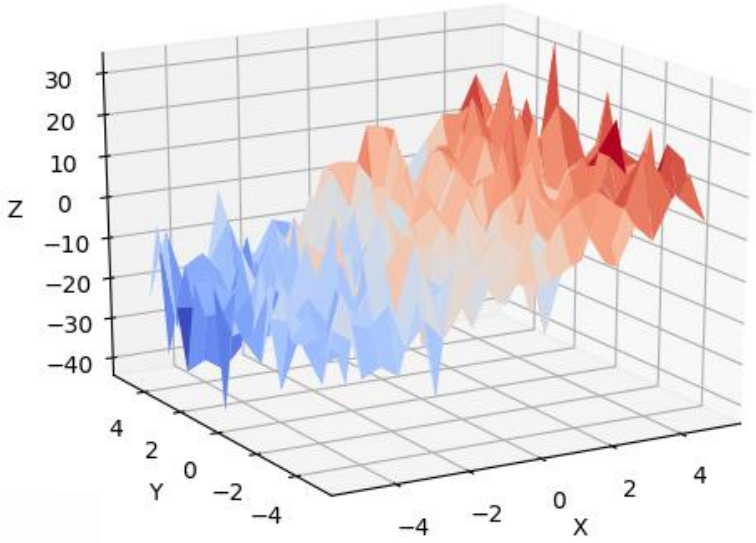


Koh, Pang Wei, et al. "Concept bottleneck models." ICML 2020.

Regression surfaces

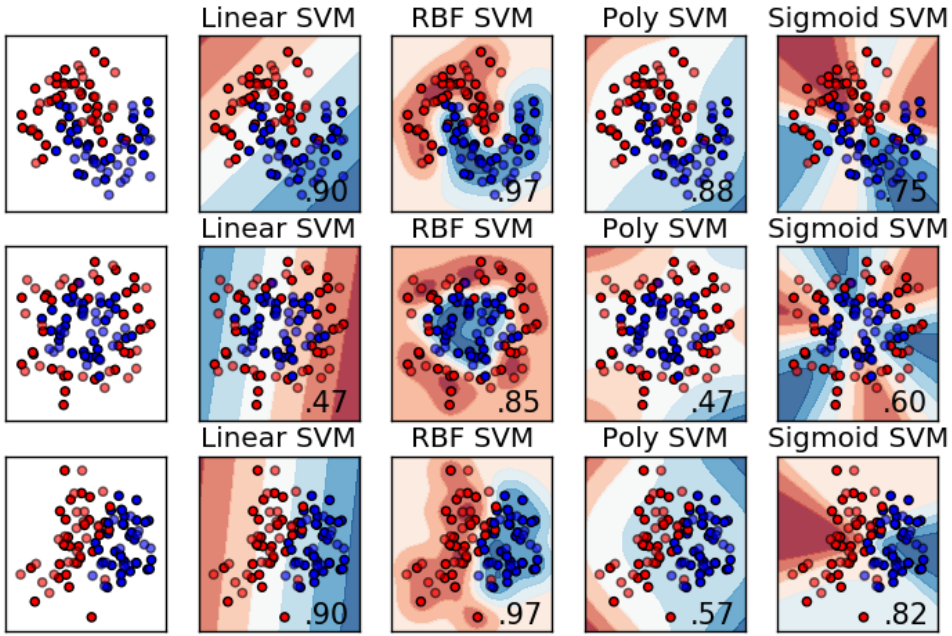
Selection/extraction of variables (transform) and regressor visualization whether linear (hyperplane) or non-linear (complex surfaces)





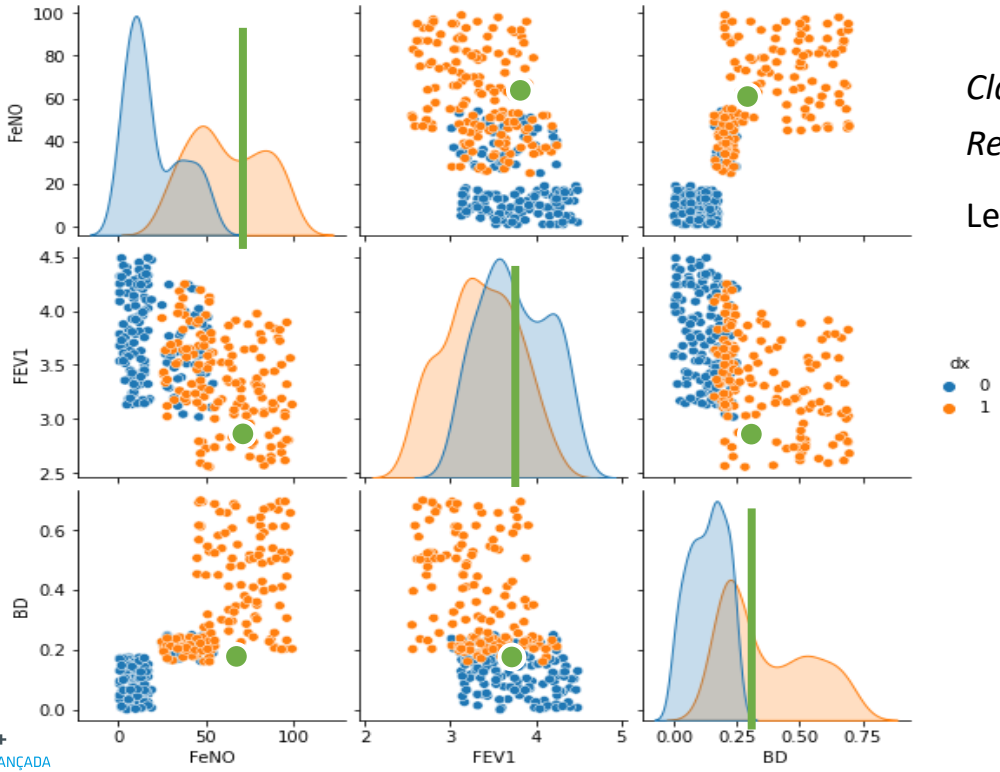
Classification boundaries

Reduce dimensionality and approximate boundaries based on posteriors



sklearn tutorials

Visualizing decisions: input view



Classification: as many colors as classes

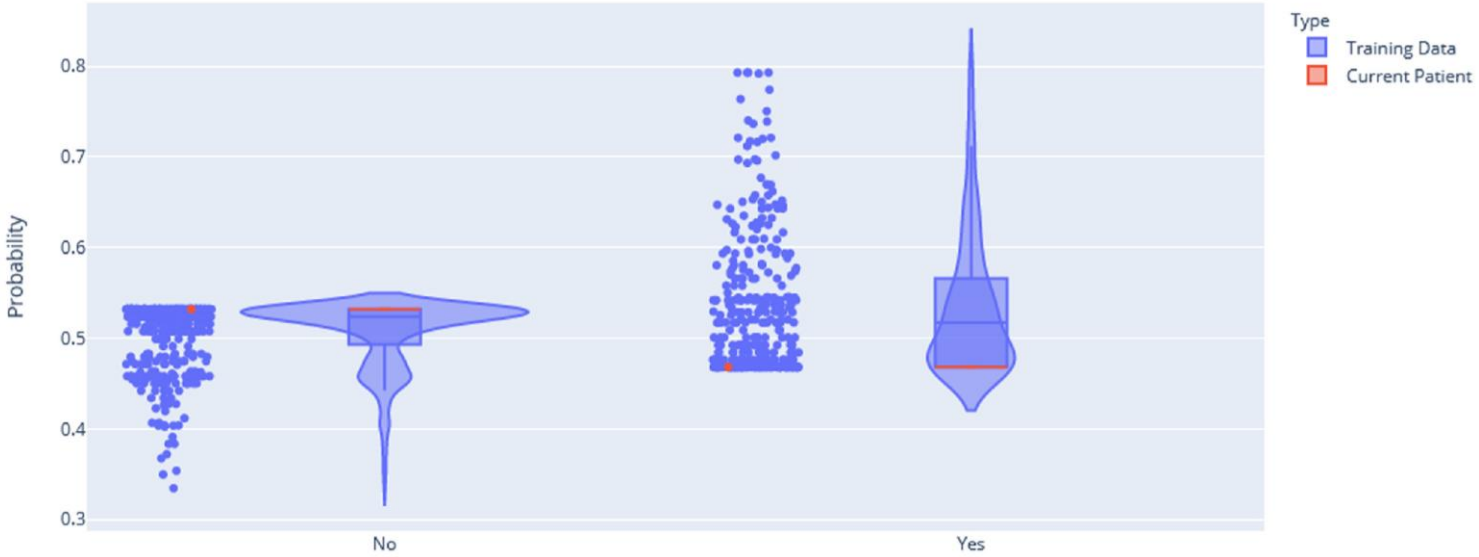
Regression: coloring scale by output value

Left example: classification, green decision

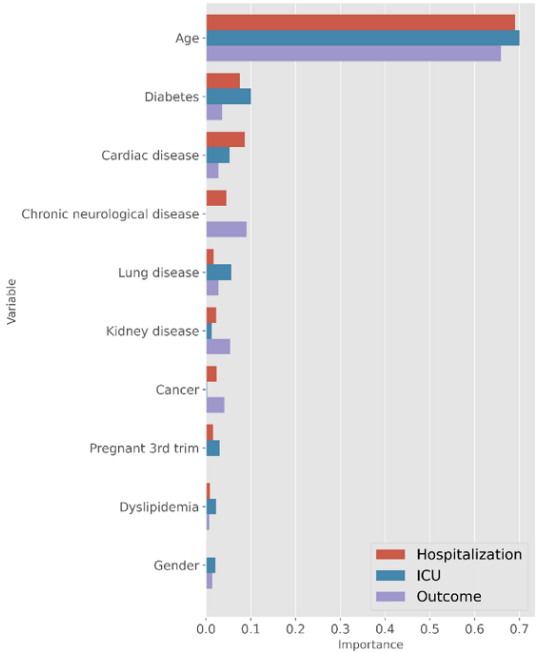
Visualizing decisions: output view

Classification: plot the class-conditional observations by $P(\mathbf{x}|c)$ and highlight the target observation

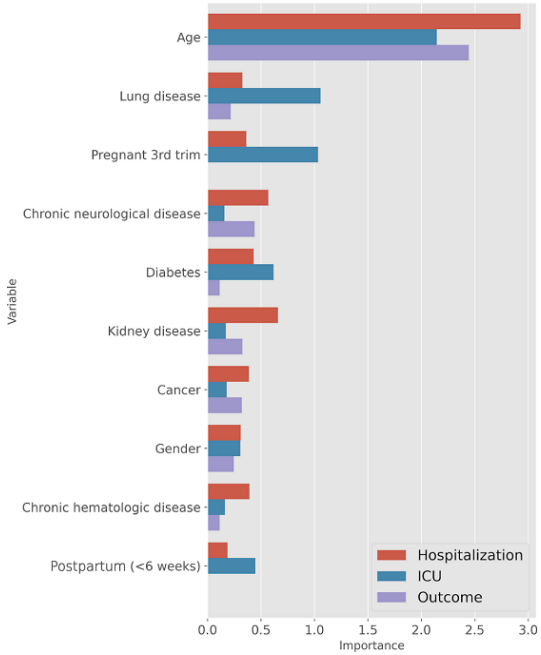
Regression: plot the distribution of observations by output value and highlight the target observation



Feature importance



(A) *XGBoost importance*



(B) *LR importance (Wald statistic)*

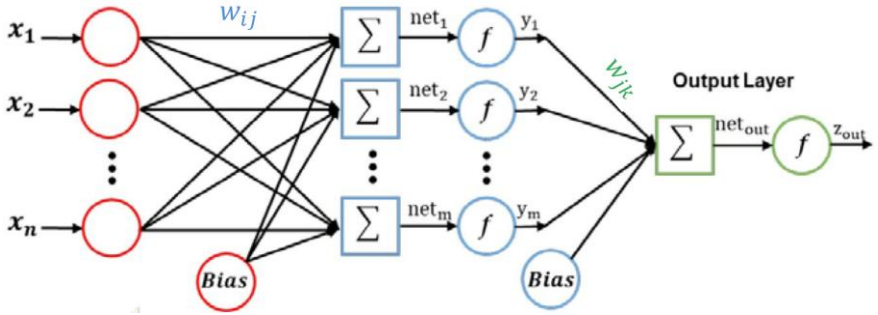
***Example** (left)*: predictability of healthcare needs for individuals testing SARS-CoV-2+

How? impurity-based score in XGBoost and Wald statistic in Logistic Regression

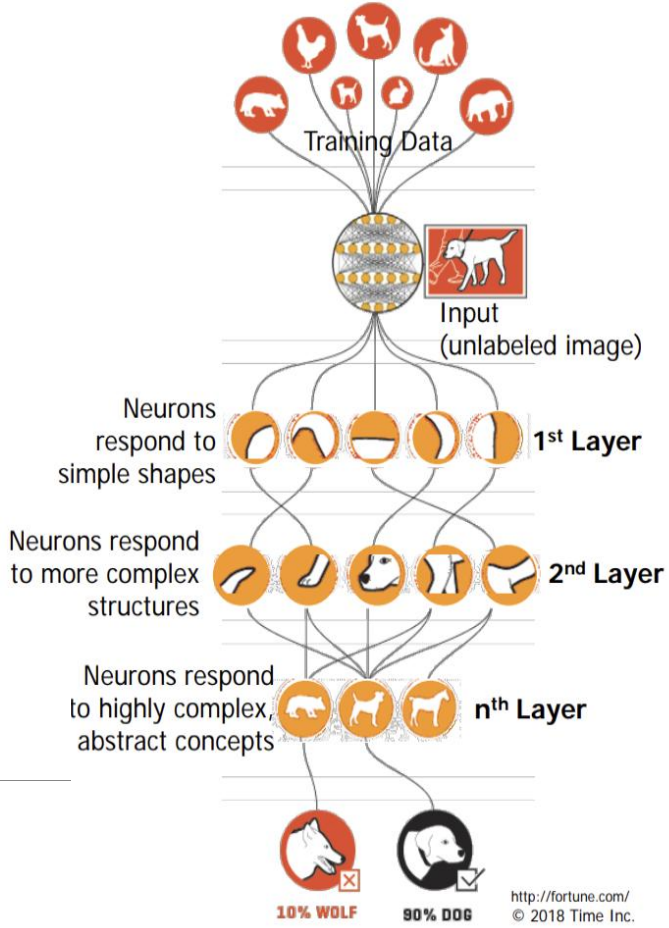
Other data structures?

I am really happy

Neural networks



$$z_{out} = f\left(\sum w_{jk}\, f\left(\sum w_{ij}\, x_i\right)\right)$$



Saliency methods for neural networks

- Heatmap based visualization

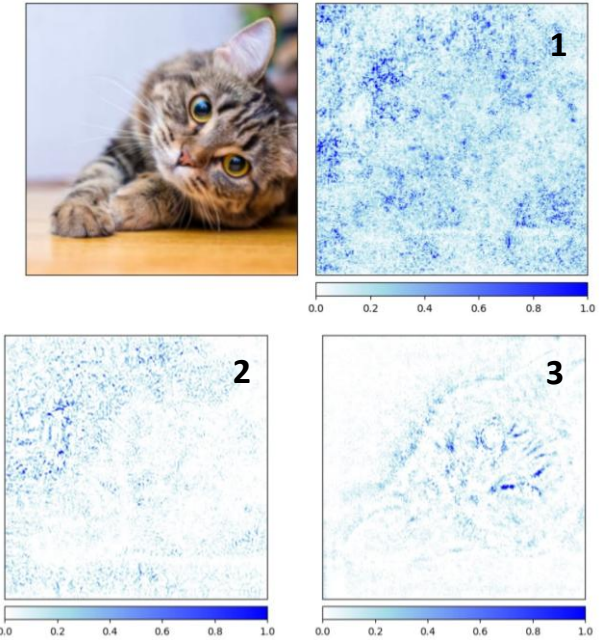
  - often needs *differentiable model* and gradient computation
  - applicable to multiple data structures
    - *multivariate* data ⇒ feature relevance
    - *image* data ⇒ pixel relevance
    - *text* data ⇒ word relevance (gradient with respect to embedding of the word)

Gradient-based explainability

- How changes in the input features affect the output directly
  - instead of computing gradients w.r.t. parameters (learn weights)…

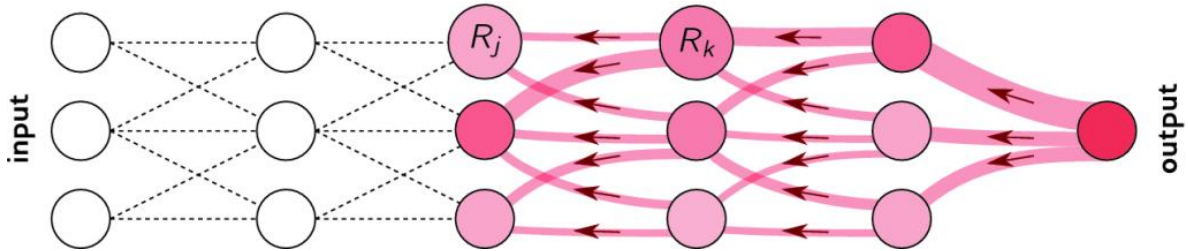
    … compute gradients w.r.t. input features
- Features with **larger gradients**:
  small changes have **higher impact** in the output

  1. VanillaGrad often produces noisy saliency maps
  2. adjust gradients in accordance with the feature value $x_i \times \frac{\partial y}{\partial x_i}$
  3. SmothGrad smooths out fluctuations (averaging gradients)



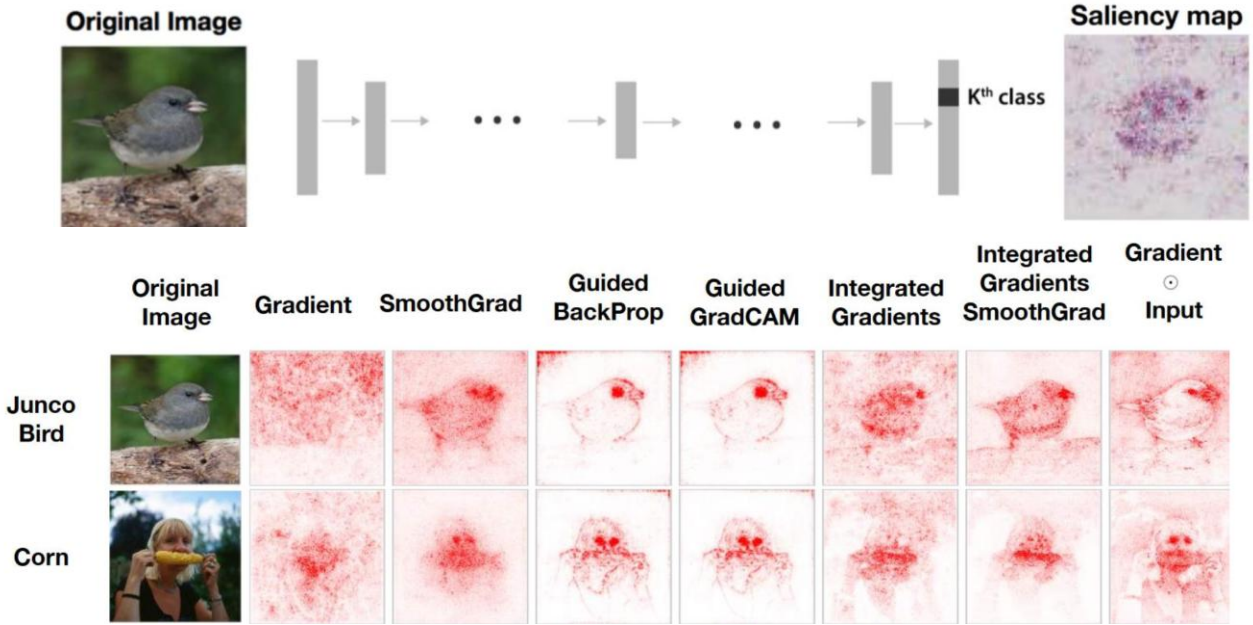
Gradient-based explainability

- **Intitution:** redistribute the contributions to the prediction score through the network neurons
  - easy case for linear neuron, yet tricky for some non-linear functions (e.g., ReLU, Sigmoid)
  - **DeepLift**
    - uses backpropagation to compare neuron activation against 'reference activation'
    - separates positive and negative contributions
  - **Layer-wise Relevance Propagation (LRP)**
    - similar to DeepLift, yet suggested for convolutional neural networks (CNNs) and recurrent neural networks (RNNs) such as LSTMs



Saliency Maps (Adebayo et al., 2018)

Gradient-based maps are inherently applicable for multivariate, text, signal data…



Outline

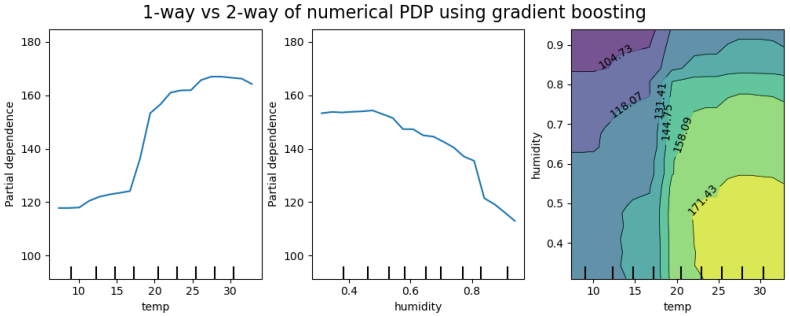
- Responsible AI
- Explainability
  - *descriptive models*
  - *predictive models*
- **Post-hoc explanations**
- Advances
- Evaluating XAI
- Going beyond

Post-hoc explanations

- **Local**
  - individual prediction explanations
    - *impact of input features*
    - *most influential examples*
    - *concepts*
    - *local decision rules*
- **Global**
  - overall model behavior
    - *partial dependence plots*
    - *global feature importance*
    - *global decision rules*

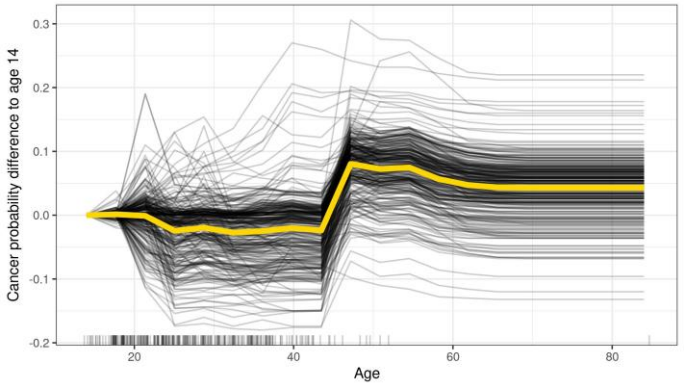
Model-agnostic plotting…

- **Partial Dependence Plots (PDP)**
  - *global visuals* of how few features influence the outcome, when others are constant



sklearn tutorials

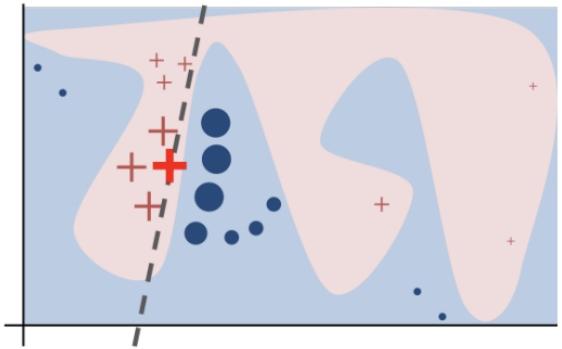
- **Individual Condition Expectations plots (ICE)**
  - *local visuals* of feature effects
    - unlike PDP, ICE separates predictions of feature effect, with one line per observation



by Molnar, Interpretable ML

LIME: Local Interpretable Model-Agnostic Explanations

- Hard to explain a complex model in its entirety?
  - How about explaining smaller/local regions?

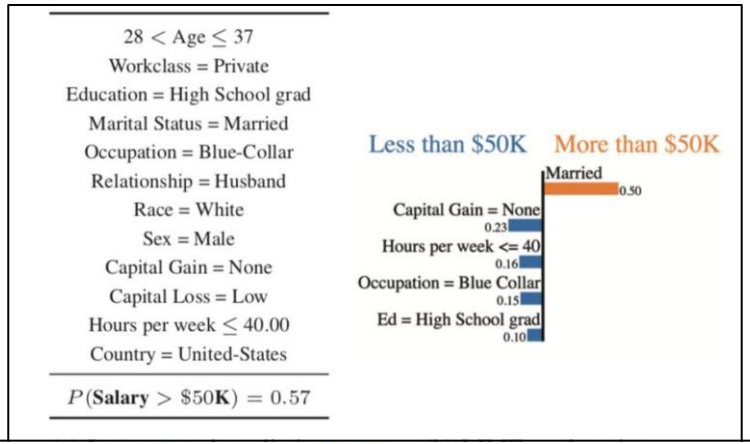


- LIME aims to see what happens inside an algorithm by capturing feature interactions
  - **local**: performs multi-feature perturbations around a prediction and measures results
  - **global**: SP-LIME is an extension for selecting representative and non-redundant predictions to further provide a global view of the model

LIME: local explanations

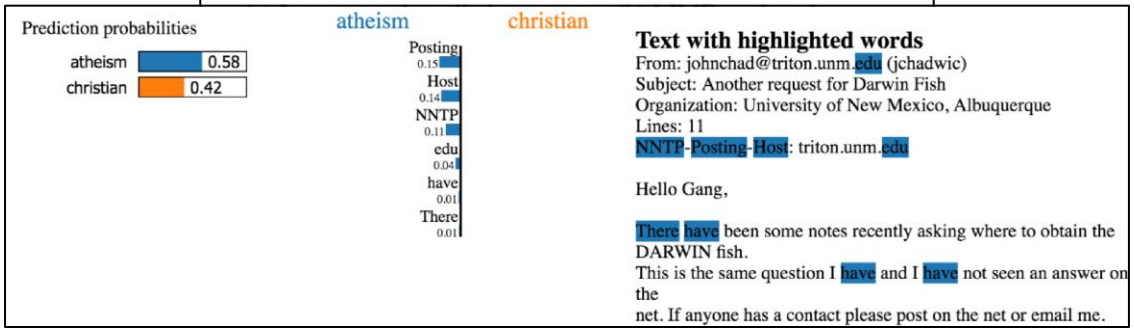
… applicable to

- multivariate data



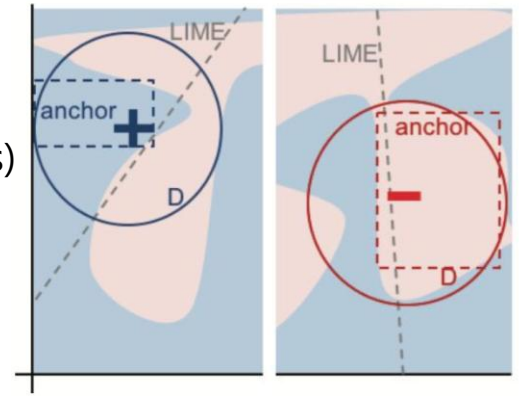
Ribeiro et al. 2018 (AAAI)

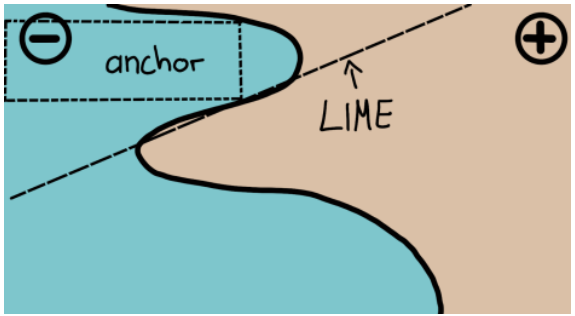
- text data



Anchors

- Proposed by the same creators as **LIME** (Ribeiro et al., 2018)
  - explains individual predictions using IF-THEN rules (anchors) that support (anchor) the predictions well enough
  - *anchors* make LIME local linear boundaries clearer

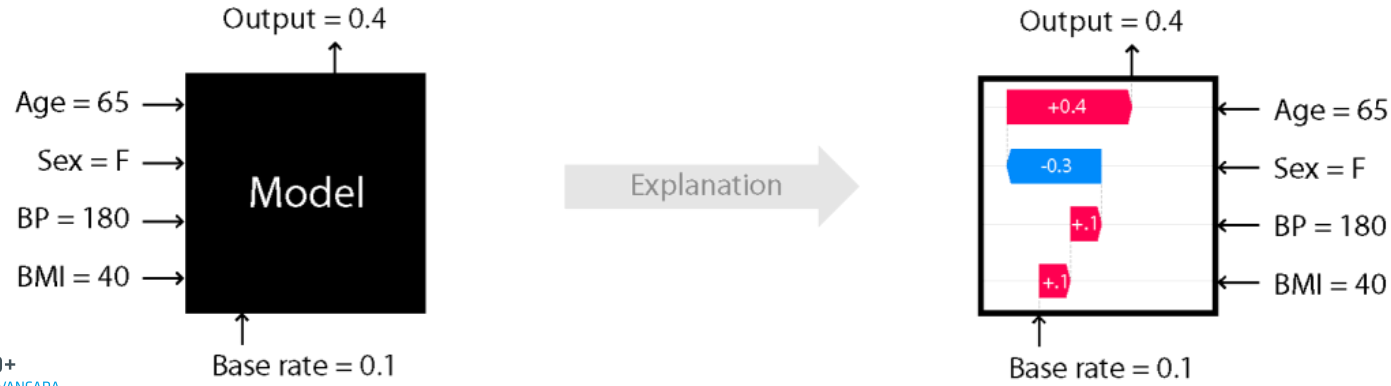




IF Country = United-States **AND** Capital Loss = Low
AND Race = White **AND** Relationship = Husband
AND Married **AND** 28 < Age ≤ 37
AND Sex = Male **AND** High School grad
AND Occupation = Blue-Collar
THEN PREDICT Salary > $50K

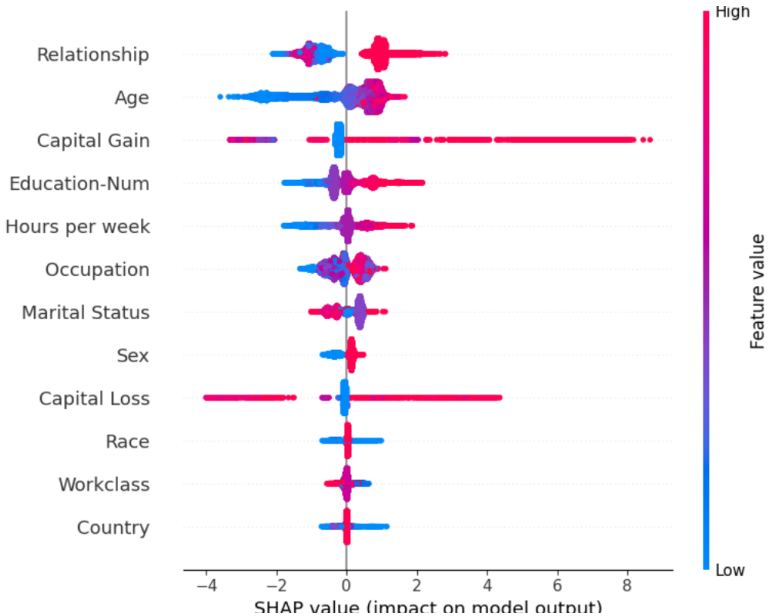
SHAP: SHapley Additive exPlanation

- Guiding univariate principle: removing a feature to quantify its *local* importance (average over different values of the feature if models does not support missings)
- SHAP: considers the **contribution of multiple features**
  - uses the game theory to assign feature importances: coalition game where **players** are the *features* in the input, **gain** is the *prediction*, feature **attributions** are the *shapley values*
  - end result produces additive explanations/attributions



SHAP: SHapley Additive exPlanation

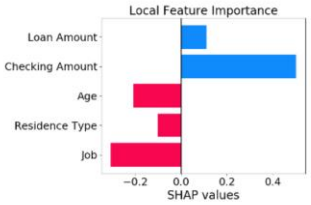
- Local feature importance: average marginal contribution over all possible coalitions
- *Extensions* for **global feature importance**
  - density scatter plots over shapley values across observations
- *Exact SHAP*: exponential in the number of features ⇒ need for heuristic approaches
- Extensions to better approximate specific models
  - **DeepSHAP** to handle the impact of specific neural processing operations (e.g., softmax)
  - **ProfWeight** to transfer knowledge from pre-trained models (details in "*Improving Simple Models with Confidence Profiles*")



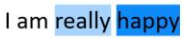
Feature importance

- from ***local***...

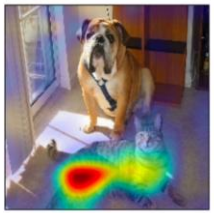
Structured data



Text



Images



(c) Grad-CAM 'Cat'

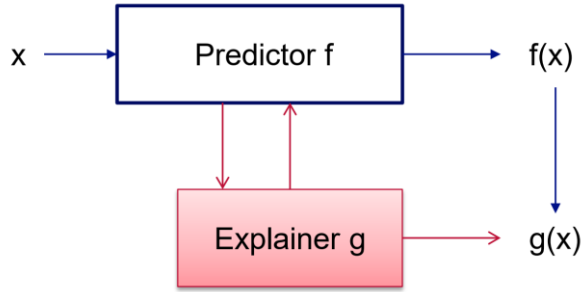
- ... to ***global***
  - **in-model setting**: wide diversity of principles (recall impurity-based, Wald statistics)
  - **post-hoc setting**: *permutations*
    - shows the decrease in the score (accuracy, F1, R^2) of a model when a single feature is randomly shuffled, revealing the importance of a given feature
    - relationship between a feature and target, useful for non-linear/opaque models
    - global LIME and SHAP

Outline

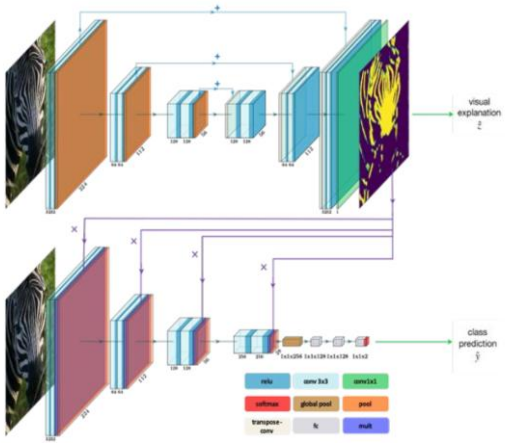
- Responsible AI
- Explainability
  - *descriptive models*
  - *predictive models*
- Post-hoc explanations
- **Advances**
- Evaluating XAI
- Going beyond

FLINT: learning interpretable networks

(by Parekh et al. 2020)



- **Jointly learn f predictor and g explainer**
  - mutual benefits: g can even inform/reshape f

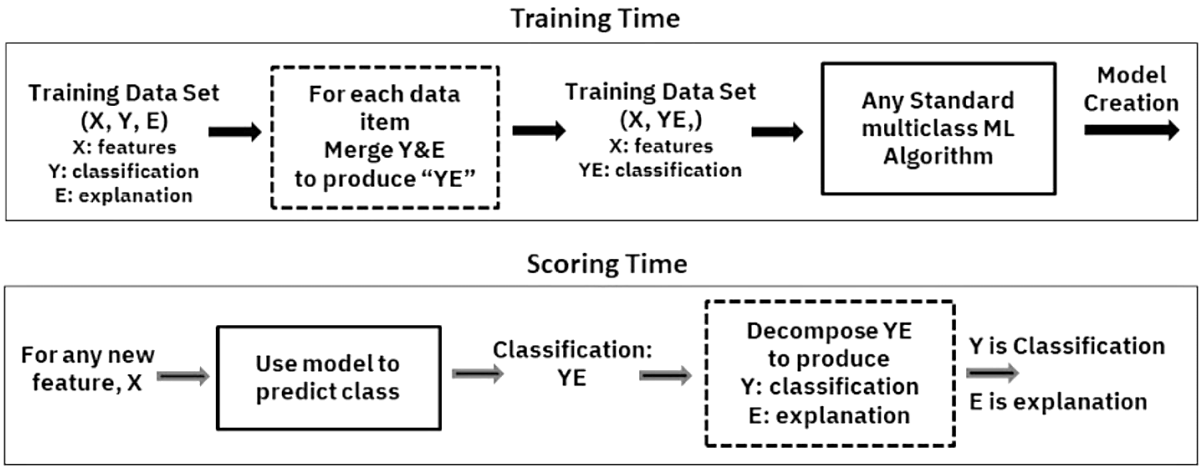


Variants (including *Rio-Torto et al. 2020*)

- How to ensure fidelity (not separate reasoning)?
  - joint learning in a 3-phased training process
  - custom loss functions

Explanations-in-the-loop (Hind et al.)

- **Train** AI systems to jointly provide a prediction and its **explanation** (deails on *Teaching AI to Explain its Decisions*, AIES 2019)
  - training tuples <X, Y, E> where E are **rationales** (human annotations to explain labels)



Explanations-in-the-loop (Hind et al.)

- **Pros**
  - explainability directly in the training process
  - teach the model what is important for us (as humans)
  - alignment to human reasoning and values
  - explanations can be tailored for the target audience

- **Cons**
  - require annotations/rationales (in fact, paper tests the approach on synthetic rationales)
  - rationale limitations (rules)
  - fidelity: explanations may not necessarily reflect inner workings but human expectations
    - faithfulness versus plausibility trade off

Outline

- Responsible AI
- Explainability
  - *descriptive models*
  - *predictive models*
- Post-hoc explanations
- Advances
- **Evaluating XAI**
- Going beyond

Do (post-model) explanations make sense?

- often do not (in a *human-understandable* manner)...

- C. Rudin, "*Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*", Nature ML 2019

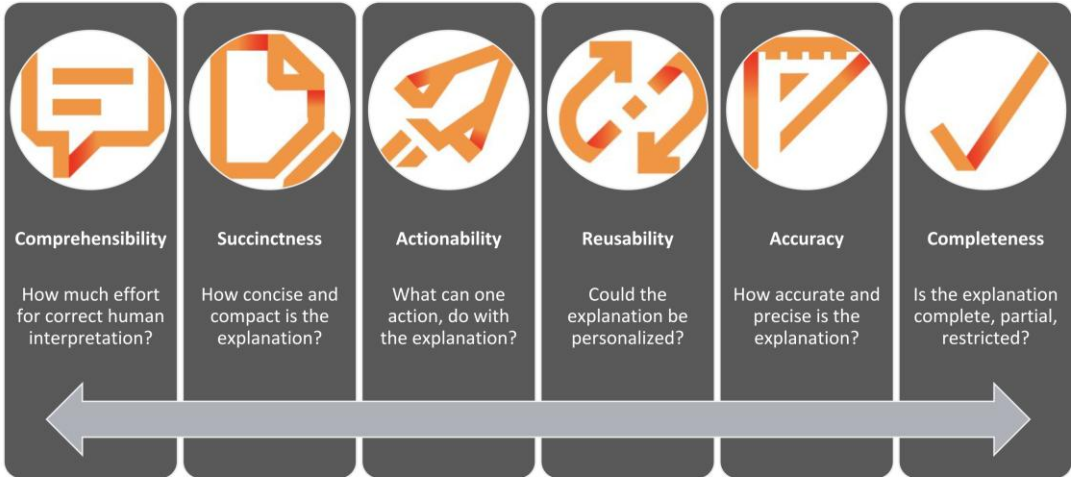
Explanations can mislead

- **We cannot assume that the explanations are by default faithful!**
  - no guarantee that a plausible explanation reflects the inner reasoning of the model
  - transparency may be at odds with broader objectives of AI, such as:
    - prefer *interpretable models* over *accurate* ones to convince decision makers (e.g. short-term goal of building medical trust clashing with long-term safety)

- **Do not blindly embrace explanations**!
  - post-hoc explanations can seem plausible but be misleading
    - most do not claim to open up the black-box
    - dependent on available data (couple *explainability* with *generalization ability*)
    - many only provide *plausible* explanations for *local behavior*

Evaluation

- Establish **guiding criteria**



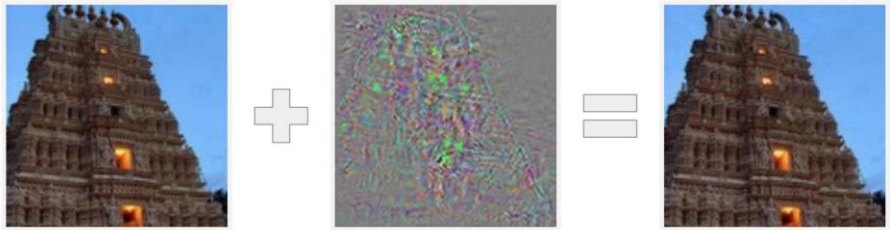
Source: Accenture Point of View. Understanding Machines: Explainable AI. Freddy Lecue, Dadong Wan

- **Assess** explanations
  - test mathematical properties
  - *monitor end users* behavior and interaction with explanations
  - conduct *(semi-)structured interviews* with (non-)experts

Evaluation

Nauta et al. *From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI*, *ACM CSUR*, 2023

- **Consistency**: identical inputs should have id. explanations ⇒ e.g. invariance to different f initializations
- **Continuity**: similar inputs should have similar explanation ⇒ stability for slight variations



- **Constrastivity**: whether explanations convey discriminative content against alternative conditions ⇒ *target sensitivity*: how features vary between decisions (the higher the difference, the better)
- **Covariate complexity** ⇒ assess how disentangled are the covariates of an explanation

Evaluation

Nauta et al. *From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI*, *ACM CSUR*, 2023

- **Presentation**
  - **compactness** ⇒ size, redundancy
  - **composition** (organization and structure of the explanation) ⇒ user studies (usability)
  - **confidence** ⇒ assess uncertainty (e.g. TED framework)

- **Utility**
  - **context** and **actionability** (relevance to stakeholders' needs)
    ⇒ user studies, quantitative assessment against domain knowledge
  - **plausibility** (reasonableness) ⇒ agreement among explainers or experts
  - **controllability** (how much a user can correct or interact with an explanation) ⇒ user studies

Evaluation

Nauta et al. *From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI*, *ACM CSUR*, 2023

- **Faithfulness:** alignment with inner working

  ⇒ use synthetic data or state-of-the-art explainers to produce reference explanations and evaluate how closely the explanation reflects the true one (agreement)

  - **fidelity** ⇒ verify if the outcomes of *f* and the explanations match

  - **correctness** (*comprehensiveness*, whether explanation captures all relevant elements) and **completeness** (*sufficiency*, whether the highlighted elements are sufficient to explain an output)
    ⇒ incremental perturbations to assess comprehensiveness/sufficiency

Outline

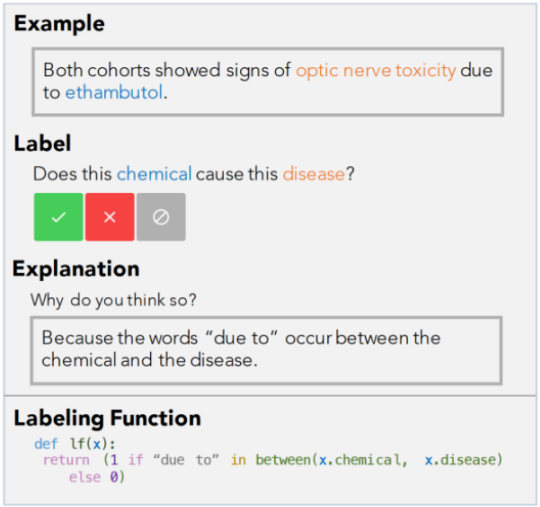
- Responsible AI
- Explainability
  - *descriptive models*
  - *predictive models*
- Post-hoc explanations
- Advances
- Evaluating XAI
- **Going beyond**

Going beyond classic explanations

- **Human-friendly explanations**
  - enhancing explanations with findings from *social sciences* and *human behavioral studies*
- **Conversational explanations**: consider the social context (to who?)
  - often preferred format is verbal explanation if it is explaining to lay-users

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

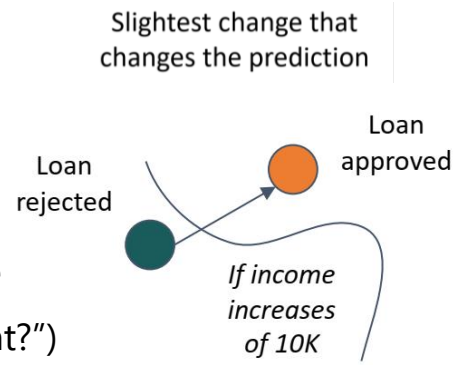
Rajani et al. 2019



Hancock et al. 2018

Going beyond classic explanations

- **Conterfactual explanations** (also called *differential* or *contrastive*)
  - explains why this prediction was made instead of another ("Why Q rather than R?")
    - single *versus* multiple instances ("Why didn't I get a mortgage when my friends did?")
  - often given by the smallest changes in features that alter the outcome
  - enhances trust and actionability ("what if I lower the requested amount?")



- Diverse Counterfactual Explanations **(DiCE)** by Watcher et al.
  - ***closeness*** to the predefined output and inputs
  - ***diversity*:** multiple counterfactuals
  - ***feasibility***: satisfy user constraints (e.g., age, amount)

Going beyond classic explanations

- **Selective explanation** (*sparse explanation*):
  - a *minimal set of features* that help justify the prediction is preferred
  - users often do not expect a complete cause for a decision

- **Credible explanations**: explanation consistent with *prior knowledge*

Outline

- **Responsible AI**
- **Explainability**
  - *descriptive models*
  - *predictive models*
- **Post-hoc explanations**
- **Advances**
- **Evaluating XAI**
- **Going beyond**

References

- Ribeiro, Singh, Guestrin, ***Why Should I Trust You?*** *Explaining the Predictions of Any Classifier*, KDD 2016
- Beaudouin et al., ***Flexible and Context-Specific AI Explainability****: A Multidisciplinary Approach,* arXiv 2020
- Wexler et al. ***The What-If Tool****: Interactive Probing of Machine Learning Models*, IEEE TVCG 2020
- Christoph Molnar's online book on ***Interpretable Machine Learning***, 2020
- Adebayo et al., ***Sanity Checks for Saliency Maps***, NeurIPS 2018
- Alvarez-Melis and Jaakola, ***Self-Explaining Neural networks***, NeurIPS 2018
- Hendricks et. al, ***Generating Visual Explanations***, ECCV 2016
- Plumb et al. ***Regularizing Black-box Models*** *for Improved Interpretability*, arXiv 2019

Further readings…

- Silva and Gonçalves "*Explainable AI:* ***unveiling what machines are learning***"
- Carvalho et al. "***Machine Learning Interpretability****: A Survey on Methods and Metrics*"
- Sequeira et al. "*An exploratory study of interpretability for face presentation attack detection*"
- Chen et al. "***Concept whitening*** *for interpretable image recognition*"
- Silva et al. "***Interpretability-Guided Content-Based Medical Image Retrieval***"
- Sundararajan et al. "***Axiomatic Attribution for Deep Networks***"
- Bach et al. "*On Pixel-Wise Explanations for Non-Linear Classifier Decisions by LWR Propagation*"

Thank you!

Rui Henriques

rmch@tecnico.ulisboa.pt