Master MVA

Responsible Machine Learning

Nicolas Vayatis

Lecture#3 - On Explainability

Questions raised

- Link betwen Machine Learning and human knowledge?
- What does it mean to explain?
- Academic (data scientistic) view on XAI: "whitening" the black box
  - A priori explainable models: Linear models, Decision Trees
  - Research in CS relating decision trees to symbolic AI (CNF/DNF)
  - Posthoc local/global explainability of nonlinear models
  - Post hoc counterfactual explanations
- Pragmatic view on explainability likely to rethink elements of the ML pipeline
- Engineering view on complex systems: hybrid models (simulation-based and data-driven)

Reminder: Basic components of ML algorithms (training)

- Training data
- Search space as a space of decision/prediction rules with constraints
- Optimization criterion (cost function)
- Numerical algorithm to derive an approximate solution

General view on explainability in ML

- Main question: How to translate an algorithm into natural language?
- Keep in mind that in Machine Learning, there are many algorithms:
  1. Prediction rule (a mathematical function)
  2. Training algorithm (a data-dependent procedure to build a prediction rule)
  3. A full ML pipeline including software, hardware and human-in-the-loop: labeling data, preprocessing, monitoring, recalibration, an interface with the environment and the operator
  4. Metalearning or learning-to-learn: from single system to many systems

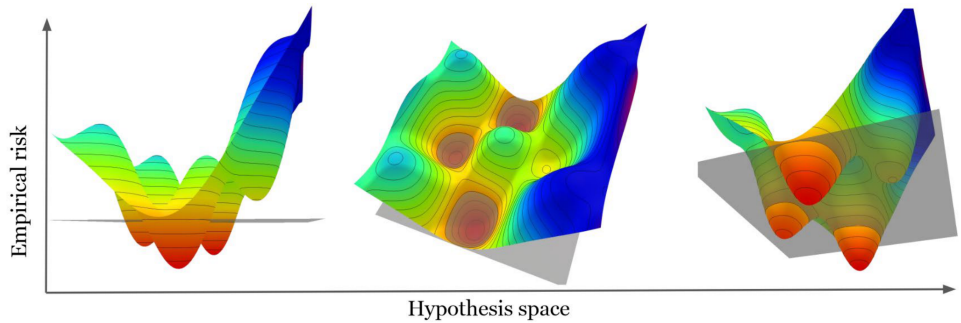
General Principles of Interpretable ML (Rudin, 2019)

- Principle 1 - An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain.
- Principle 2 - Despite common rhetoric, interpretable models do not necessarily create or enable trust – they could also enable distrust. They simply allow users to decide whether to trust them. In other words, they permit a decision of trust, rather than trust itself.
- ...

General Principles of Interpretable ML (Rudin, 2019)

- ...
- Principle 3 - It is important not to assume that one needs to make a sacrifice in accuracy in order to gain interpretability. In fact, interpretability often begets accuracy, and not the reverse. Interpretability versus accuracy is, in general, a false dichotomy in machine learning.
- Principle 4 - As part of the full data science process, one should expect both the performance metric and interpretability metric to be iteratively refined.
- Principle 5 - For high stakes decisions, interpretable models should be used if possible, rather than "explained" black box models.

The Rashomon effect (Breiman, 2001 ; Semenova, Rudin, Parr, 2022)



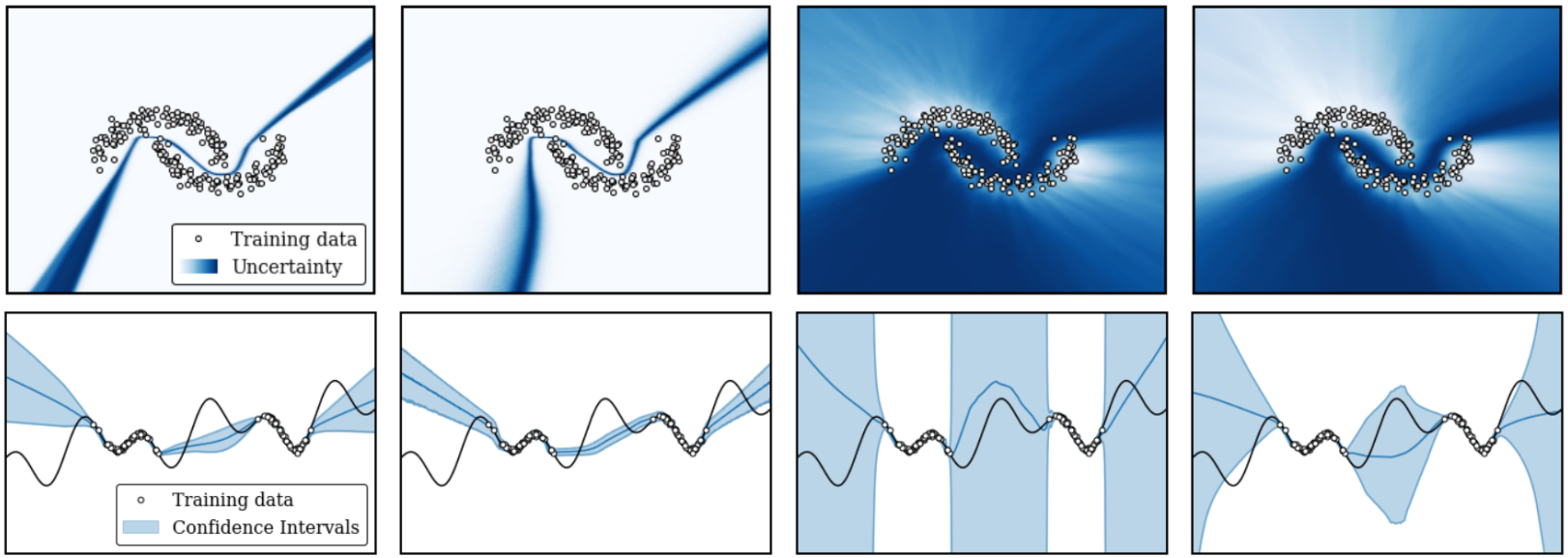
(a) From the side (b) From above (c) From underneath

Definition 1 (Rashomon set)**.** *Given $\theta \geq 0$, a data set S, a hypothesis space $\mathcal{F}$, and a loss function ϕ, the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$ is the subspace of the hypothesis space defined as follows:*

$$\hat{R}_{set}(\mathcal{F}, \theta) := \{f \in \mathcal{F} : \hat{L}(f) \leq \hat{L}(\hat{f}) + \theta\},$$

where $\hat{f}$ is an empirical risk minimizer for the training data S with respect to loss function ϕ: $\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \hat{L}(f)$.

Out-of-distribution uncertainty: regimes of accuracy may induce regimes of explainability



(a) Deep Ensemble (b) MC-Dropout (c) MaxWEnt (d) MaxWEnt + Clip

From (De Mathelin, Deheeger, Mougeot, Vayatis, 2023)

Notions of ML explainability

• **Causal explanation**: logical rules or key drivers or combination of factors which explain the prediction
• **Statistical explanation**: statistics of the training set, e.g. ≪ you are diagnosed X, you have 40% of chances to recover ≫
• **Contextual explanation**: could be traced by clustering the data set by date for instance
• **Explanation of the impact of the prediction**: relies on post processing the prediction outcome within the decision process
• **Counterfactual explanations:** variables for which small perturbation induce opposite decision
. . . and tools: XAI by DARPA, WHAT-IF by Google, and more (check Boza-Evgeniou, INSEAD Working paper, 2021)

How much complexity can be handled by Humans?

• **Sparsity - The law of the seven (+/-2) objects**
George A. Miller. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", Psychological Review, 1956.
• **Expository power - Humans are seriously limited in estimating the association between three or more variables.**
D Jennings, TM Amabile, and L Ross. Informal covariation assessment: Data-based vs. theory-based judgments. In Judgment under uncertainty: Heuristics and biases, pages 211–230, 1982
• **Monotonicity – ≪ We understand what we already know ≫**
Stefan Rüping. Learning interpretable models. PhD thesis, Universität Dortmund, 2006.

Global explainability of linear models

Result table for a simple linear model

Summary of Fit

RSquare	0.025173
RSquare Adj	0.021194
Root Mean Square Error	1.119951
Mean of Response	0.799874
Observations (or Sum Wgts)	247

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	7.9354	7.9354	6.3266
Error	245	307.30094	1.25429	**Prob > F**
C. Total	246	315.23634		0.0125

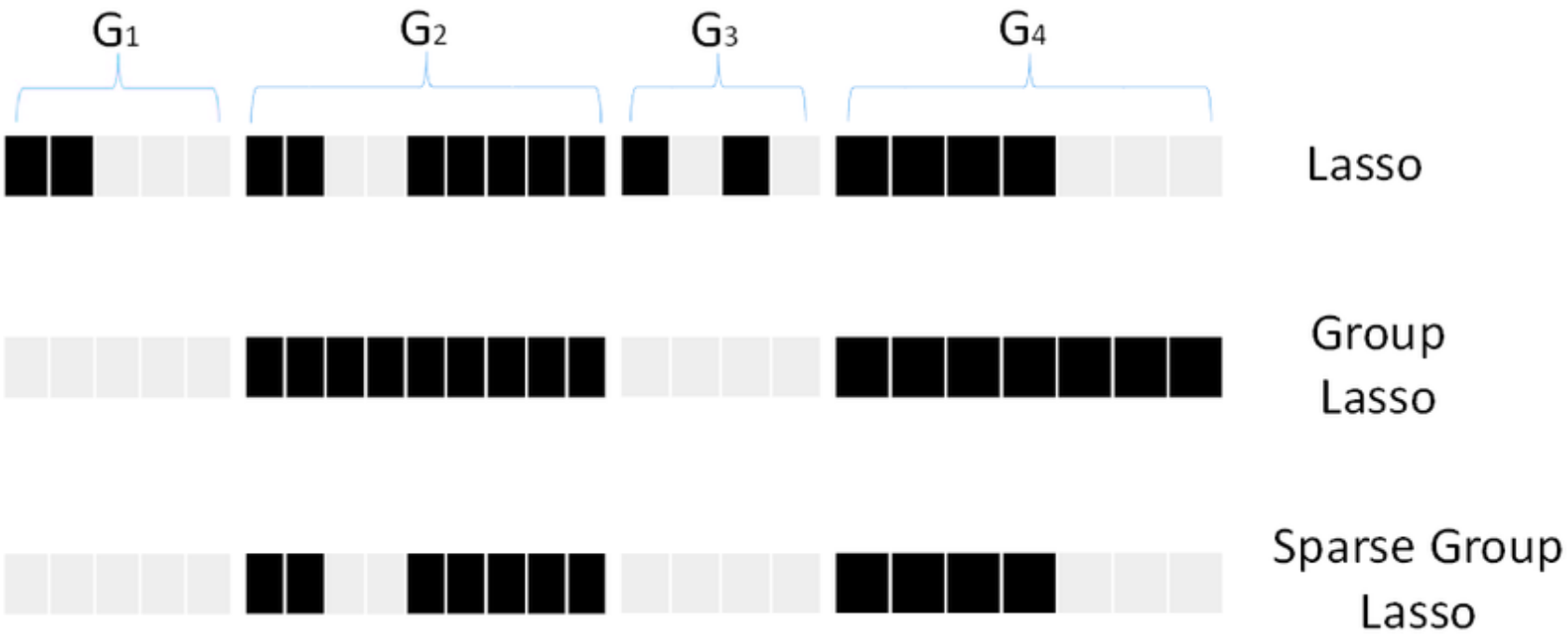
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.2885955	0.596146	3.84	0.0002
Initial TL (mm)	-0.015478	0.006153	-2.52	0.0125

Interpretability in linear models

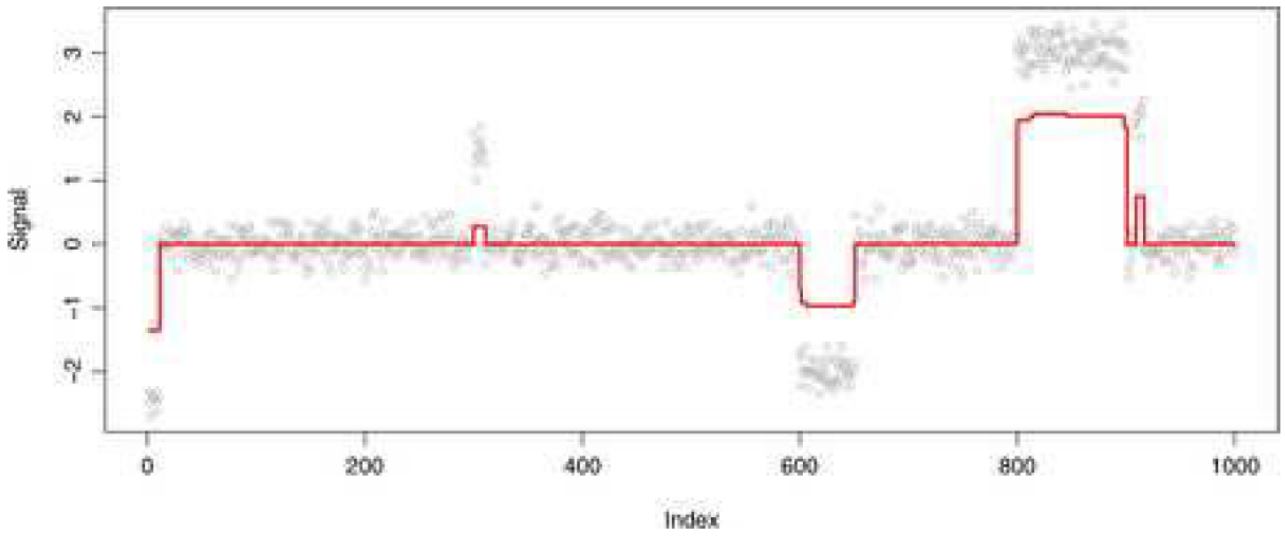
- General interpretation in small dimensions
  - Sensitivity analysis: effect of increment of one variable on the outcome (all other variables being fixed
  - Variable importance measured with the t-statistic (coefficient divided by it standard error)
  - Individual Conditional Expectation plot (ICE): plots the variations of the outcome with respect to the variations of a single variable (Goldstein et al. , 2015)
  - Partial Dependence Plots (PD): same but with respect to a subgroup of variables
- In high dimensions:
  - Sparsity and structured sparsity
  - Integer linear models and interpretable constraints (Ustun and Rudin, 2016)

Sparsity and Structured sparsity



- The idea is to reduce the dimension of the linear model while preserving the underlying structure
- Similar with matrix estimation (see Savalle, Richard, Vayatis, 2012)

Other example: Fused LASSO



- Enforcing temporal coherence leads to adding a penalty term:

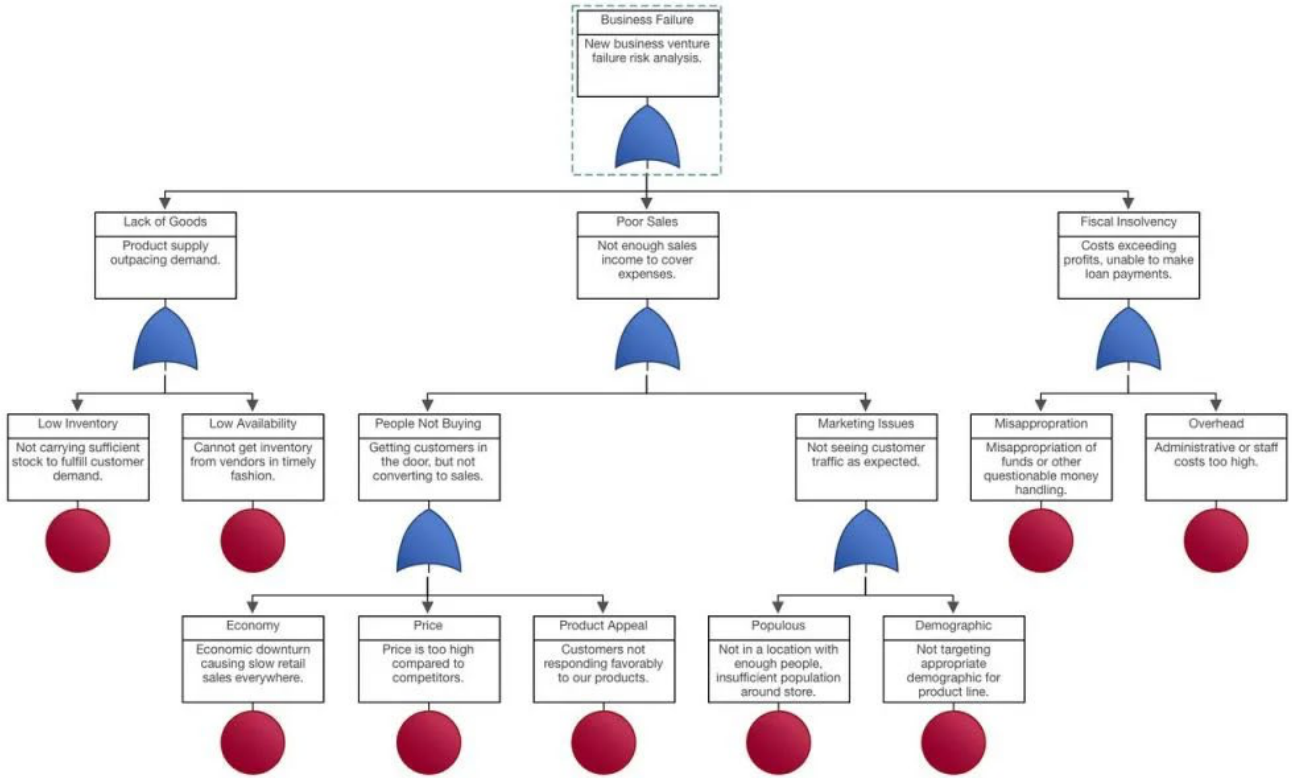
$$\widehat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|_1 + \mu \sum_{j=2}^{d} |\beta_j - \beta_{j-1}| \right\}$$

Global explainability of nonlinear models

Interpretability in nonlinear models

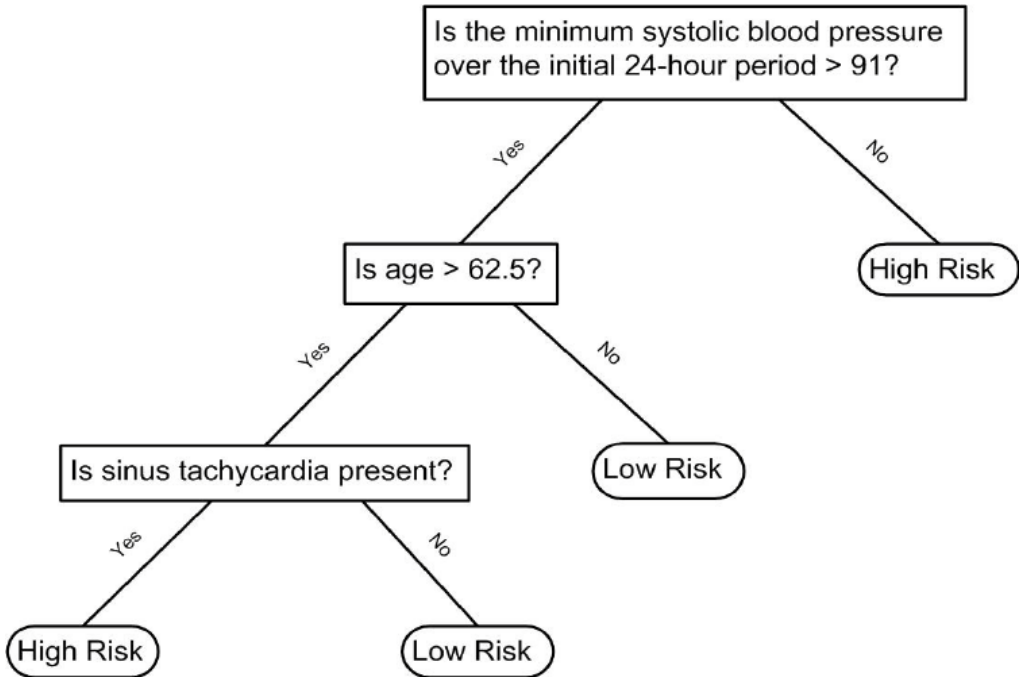
- Ideal interpretable systems are expert systems
- Decision trees
  - do provide by construction inference data-driven explanations
- Random Forests
  - Variable importance
- Issue with tree size and redundancy (Izza, Ignatiev, Marques-Silva, 2022)
- Modern view : causal inference (Prosperi, M., Guo, Y., Sperrin, M. et al. , 2020)

Expert systems provide causal explanations



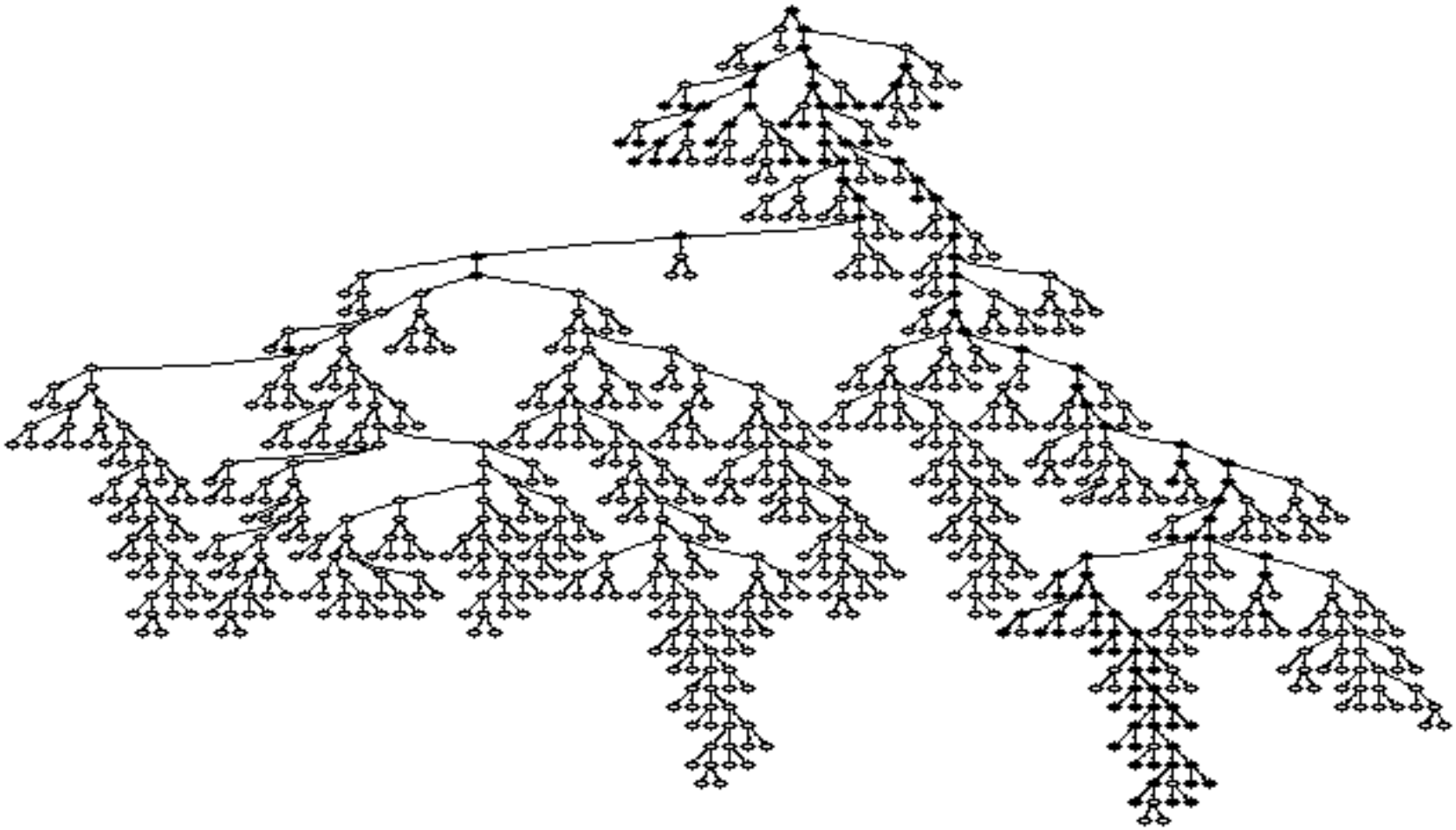
Example of a fault tree with logical gates combining unitary events to explain the top event

Decision trees also provide causal explanations and they can be learned from data

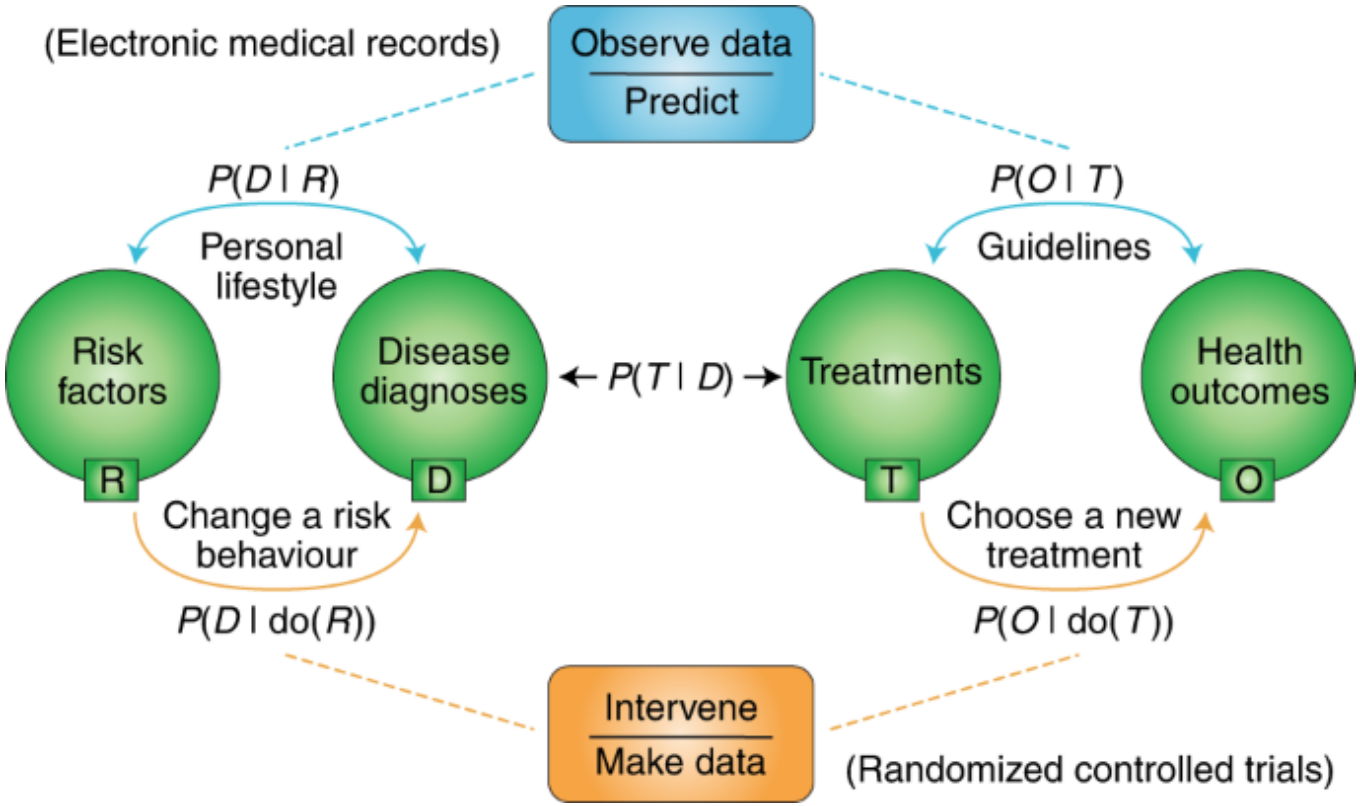


Reference: CART Book
Breiman, Friedman, Olshen, and Stone (1984).

But what if the optimal tree looks like this?



Modern view: causal inference

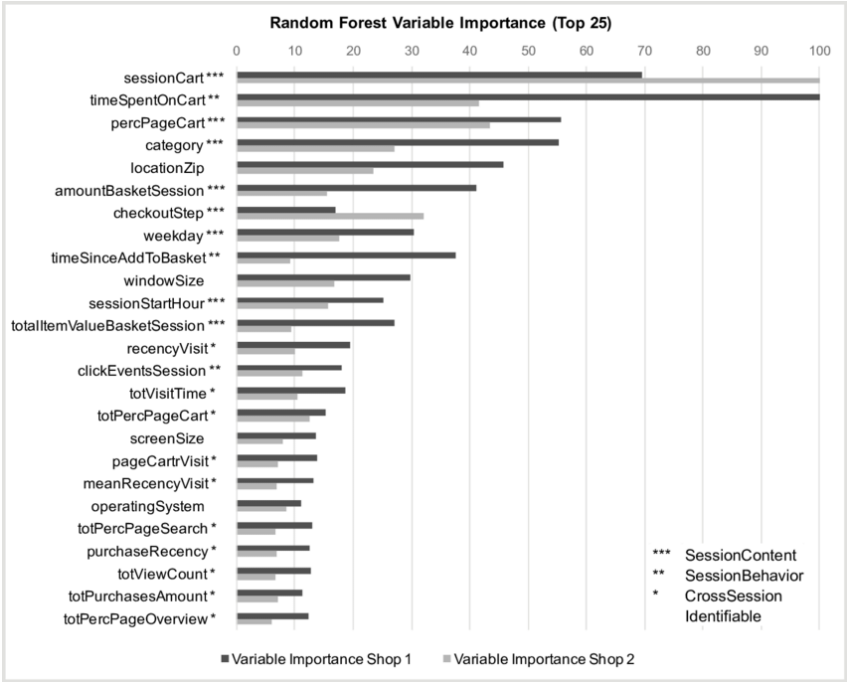


From Prosperi, M., Guo, Y., Sperrin, M. *et al.* Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell* **2,** 369–375 (2020). https://doi.org/10.1038/s42256-020-0197-y

Variable importance in ensemble of trees

Variable importance is based on counting how many times a variable is used in a split along the decision trees with a coefficient which is proportional to the decrease of impurity at that very node.

Example from [Baumann, Annika; Haupt, Johannes; Lessmann, Stefan; and Gebert, Fabian (2019) "The Price of Privacy - An Evaluation of the Economic Value of Collecting Clickstream Data," *Business & Information Systems Engineering*: Vol. 61: Iss. 4, 413-431.]



Local explainability: LIME/SHAP/counterfactual

Posthoc analysis: local approximation

LIME: Local Interpretable Model-Agnostic Explanations
[Ribeiro, Singh, Guestrin, (2016)]

Consider a black-box model f and an evaluation point x, then:

1. Generate a collection $\pi(x)$ of points similar to x (perturbations of x or instances in the vicinity of x
2. Select an interpretable set of functions (linear, decision trees, ensemble of shallow trees)
3. Estimate a local approximation $g \in \mathcal{G}$ of the black-box f based on the following optimization problem:

$$\text{loc-explanation}(x) \in \underset{g \in \mathcal{G}}{\arg\min} \{L(f, g, \pi(x)) + \Omega(g)\}$$

where L discrepancy measure between f and g using a collection $\pi(x)$ of points similar to the evaluation point x and $\Omega(\cdot)$ regularizer which enforces interpretability/sparsity

Illustration of LIME

from [Ribeiro, Singh, Guestrin, (2016)]

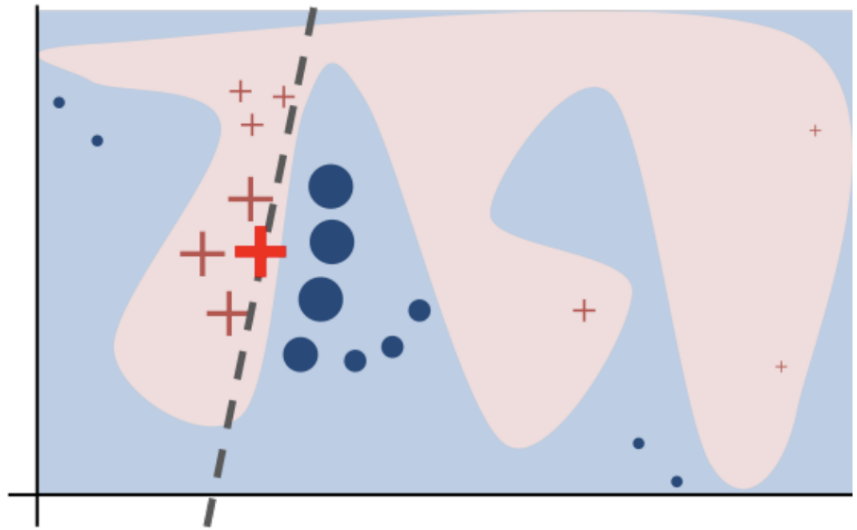


Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

Posthoc analysis: Shapley additive explanation

from [Lundberg and Lee, NeurIPS'17]

Definition 1 Additive feature attribution methods *have an explanation model that is a linear function of binary variables:*

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where $z' \in \{0,1\}^M$, M *is the number of simplified input features, and* $\phi_i \in \mathbb{R}$.

Theorem 1 *Only one possible explanation model* g *follows Definition 1 and satisfies Properties 1, 2, and 3:*

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \tag{8}$$

where $|z'|$ *is the number of non-zero entries in* z', *and* $z' \subseteq x'$ *represents all* z' *vectors where the non-zero entries are a subset of the non-zero entries in* x'.

Posthoc analysis: counterfactual explanations

Given a black-box f and an instance x, find counterfactual instances which are similar but would flip the label of x through f [Wachter,Mittelstadt, Russell (2018)]

- The counterfactual instances are synthetic and can be obtained by solving an optimization problem of the form:

$$\text{counterfactual}(x) \in \arg\min_{x'} \left\{ \ell(f(x'), y') + \lambda d(x', x) \right\}$$

  where ℓ loss function which provides pointwise evaluation of discrepancy between prediction and label and d is a distance between instances, λ is a smoothing parameter

- It is possible to obtain several counterfactual instances by tuning the smoothing parameter λ

Counterfactuals
Example from credit scoring

Instance x of interest

age	sex	job	housing	savings	checking	amount	duration	purpose
58	f	unskilled	free	little	little	6143	48	car

Counterfactual explanations x':

age	sex	job	amount	duration	o_2	o_3	o_3	$\hat{f}(x')$
		skilled		-20	0.108	2	0.036	0.501
		skilled		-24	0.114	2	0.029	0.525
		skilled		-22	0.111	2	0.033	0.513
-6		skilled		-24	0.126	3	0.018	0.505
-3		skilled		-24	0.120	3	0.024	0.515
-1		skilled		-24	0.116	3	0.027	0.522
-3	m			-24	0.195	3	0.012	0.501
-6	m			-25	0.202	3	0.011	0.501
-30	m	skilled		-24	0.285	4	0.005	0.590
-4	m		-1254	-24	0.204	4	0.002	0.506

Complex phenomena: from explainability to knowledge...

Understanding of complex systems

Motivations and Challenges for Digital Twins in health, industry...

1. Understanding/Anticipating the behavior of the full system at scale

   Challenge #1: system-level (multiscale, multiphysics...) and lifecycle-level modeling

2. Benefits of simulations to reduce the number of physical experiments (if possible)

   Challenge #2: computational feasibility (real-time in some cases)

3. Assessment of interventions on the system

   Challenge #3: Blending expertise from physics and data in the grey zone

The grey zone of physical models

The limits of physics to describe complex systems

1. only fundamental mechanisms explained by physical laws, while variability in real systems may arise from unreported/unobserved/unrecorded factors
2. highly refined physical models bring heavy simulations which may not be compatible with real-life operations (need to run in real-time for instance)
3. sensitivity to high dimensional parameters may compromise the ability to monitor all possible outcomes (experimental design may suffer from the curse of dimensionality)

The grey zone of data-driven models

The limits of data-driven models (parametric statistics or machine learning)

1. Complex models (e.g. neural networks) trained on high dimensional databases with small sample size may overfit
2. Out of the domain of training data, prediction accuracy cannot be guaranteed
3. Knowledge transfer to perform cold start on a new instance of the complex system (say new factory, new design) raises fundamental issues on the use and validity of any statistical model
4. Adoption of data-driven models by human experts requires some level of interpretability/explainability/transparency which cannot be provided through functions with complex dependencies (like neural networks).

Discussion

- Explainable AI as an academic field may not be that useful to address societal/industrial needs
- Probably there might not be a universal notion of explainability, it might even be user-specific as $\ll$ we understand what we already know $\gg$...
- Interesting lead that ML may be itself a tool for scientific exploration and explainability of complex phenomena

AI for good

Sendhil Mullainathan

An algorithmic approach to reducing unexplained pain disparities in underserved populations

Emma Pierson, David M. Cutler, Jure Leskovec, Sendhil Mullainathan & Ziad Obermeyer

Nature Medicine **27**, 136–140 (2021) | Cite this article

16k Accesses | **114** Citations | **738** Altmetric | Metrics

Abstract

Underserved populations experience higher levels of pain. These disparities persist even after controlling for the objective severity of diseases like osteoarthritis, as graded by human physicians using medical images, raising the possibility that underserved patients' pain stems from factors external to the knee, such as stress. Here we use a deep learning approach to measure the severity of osteoarthritis, by using knee X-rays to predict patients' experienced pain. We show that this approach dramatically reduces unexplained racial disparities in pain. Relative to standard measures of severity graded by radiologists, which accounted for only 9% (95% confidence interval (CI), 3–16%) of racial disparities in pain, algorithmic predictions accounted for 43% of disparities, or 4.7× more (95% CI, 3.2–11.8×), with similar results for lower-income and less-educated patients. This suggests that much of underserved patients' pain stems from factors within the knee not reflected in standard radiographic measures of severity. We show that the algorithm's ability to reduce unexplained disparities is rooted in the racial and socioeconomic diversity of the training set. Because algorithmic severity measures better capture underserved patients' pain, and severity measures influence treatment decisions, algorithmic predictions could potentially redress disparities in access to treatments like arthroplasty.

"I think that neuroscience will be again a source of inspiration and will keep playing a fundamental role for the advances of AI: indeed the best definition of intelligence we have is the one defined by Turing, which is deeply filled with human intelligence concepts. Therefore, studying how our brain works will definitely support us in engineering areas of computer vision and machine learning. However, more than this, I strongly believe interdisciplinarity will be at the core of AI future development. Breakthrough discoveries can only be derived from the cooperation among computer scientists, engineers and neuroscientists: good news for institutions such as Center for Brains, Minds and Machines of MIT or Computation and Cognition Lab of Stanford and other universities, where there is a good mix of engineering competencies, neuroscience and cognitive knowledge. Such capabilities are still maybe out of the focus in several departments of tech giants, where the side of engineering solutions is privileged."

from [Interview of Prof. Tomaso Poggio (MIT), March 2019.]