le cnam

Statistical Challenges in Business and Industry in the Era of Machine Learning and Artificial Intelligence

Gilbert Saporta
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris
gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Applied Stochastic Models in Business and Industry

WILEY

RESEARCH ARTICLE OPEN ACCESS

Is There a Future for Stochastic Modeling in Business and Industry in the Era of Machine Learning and Artificial Intelligence?

Fabrizio Ruggeri[1] | David Banks[2] | William S. Cleveland[3] | Nicholas I. Fisher[4] | Marcos Escobar-Anel[5] | Paolo Giudici[6] | Emanuela Raffinetti[6] | Roger W. Hoerl[7] | Dennis K. J. Lin[3] | Ron S. Kenett[8] | Wai Keung Li[9] | Philip L. H. Yu[9] | Jean-Michel Poggi[10] | Marco S. Reis[11] | Gilbert Saporta[12] | Piercesare Secchi[13] | Rituparna Sen[14] | Ansgar Steland[15] | Zhanpan Zhang[16]

Correspondence: Fabrizio Ruggeri (fabrizio@mi.imati.cnr.it)

Received: 15 November 2024 | **Accepted:** 13 February 2025

Funding: This work was supported by Fundação para a Ciência e a Tecnologia and European Commission.

Keywords: artificial intelligence | machine learning | stochastic models

Outline

1. The data revolution
2. The triumph of black boxes in the 2000s and a few setbacks
3. Explainable AI
4. New requirements: interpretable and causal models
5. Machine Learning and Statistics: a mutual reinforcement
6. Concluding remarks

1. The data revolution

DEFINING THE DATA REVOLUTION

'The data revolution is: an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the 'Internet of Things,' and from other sources, such as qualitative data, citizen-generated data and perceptions data; A growing demand for data from all parts of society.'

UN Secretary-General's Independent Expert Advisory Group on a Data Revolution (A World That Counts report, page 6)

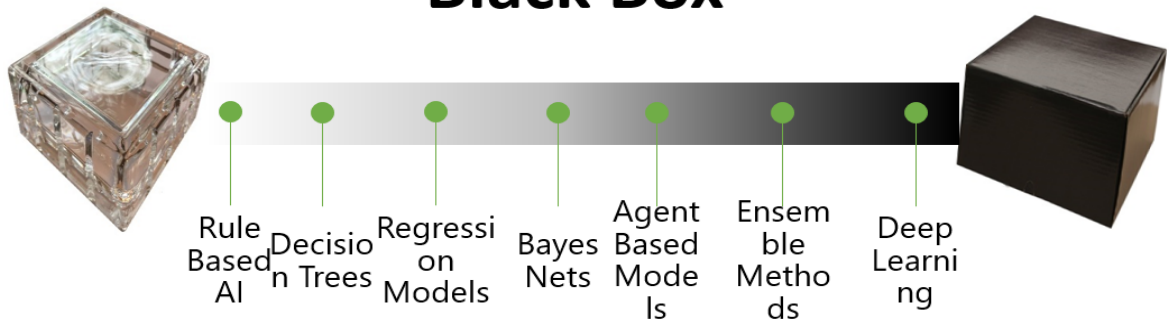
What steam was to the 19th century, and oil has been to the 20th, data is to the 21st. It's the driver of prosperity, the revolutionary resource that is transforming the nature of economic activity, the capability that differentiates successful from unsuccessful societies.

The Data Manifesto, Royal Statistical Society, 2014

2. The triumph of black boxes in the 2000s and a few setbacks

- Successful in many fields: consumer analytics, fraud detection, recommendation systems, speech recognition etc.
- Model factories
  - Systems that automatically generate predictive models with little or no human intervention. e.g.: simultaneous sales forecasts of thousands of items

From Glass Box to Black Box



- More complex models are supposed to have better accuracy
- Statistical Learning Theory proves the existence of an optimal complexity (Vapnik, 1971)

Predicting without understanding!

- Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data (Breiman, 2001).

- Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (Vapnik, 1982)

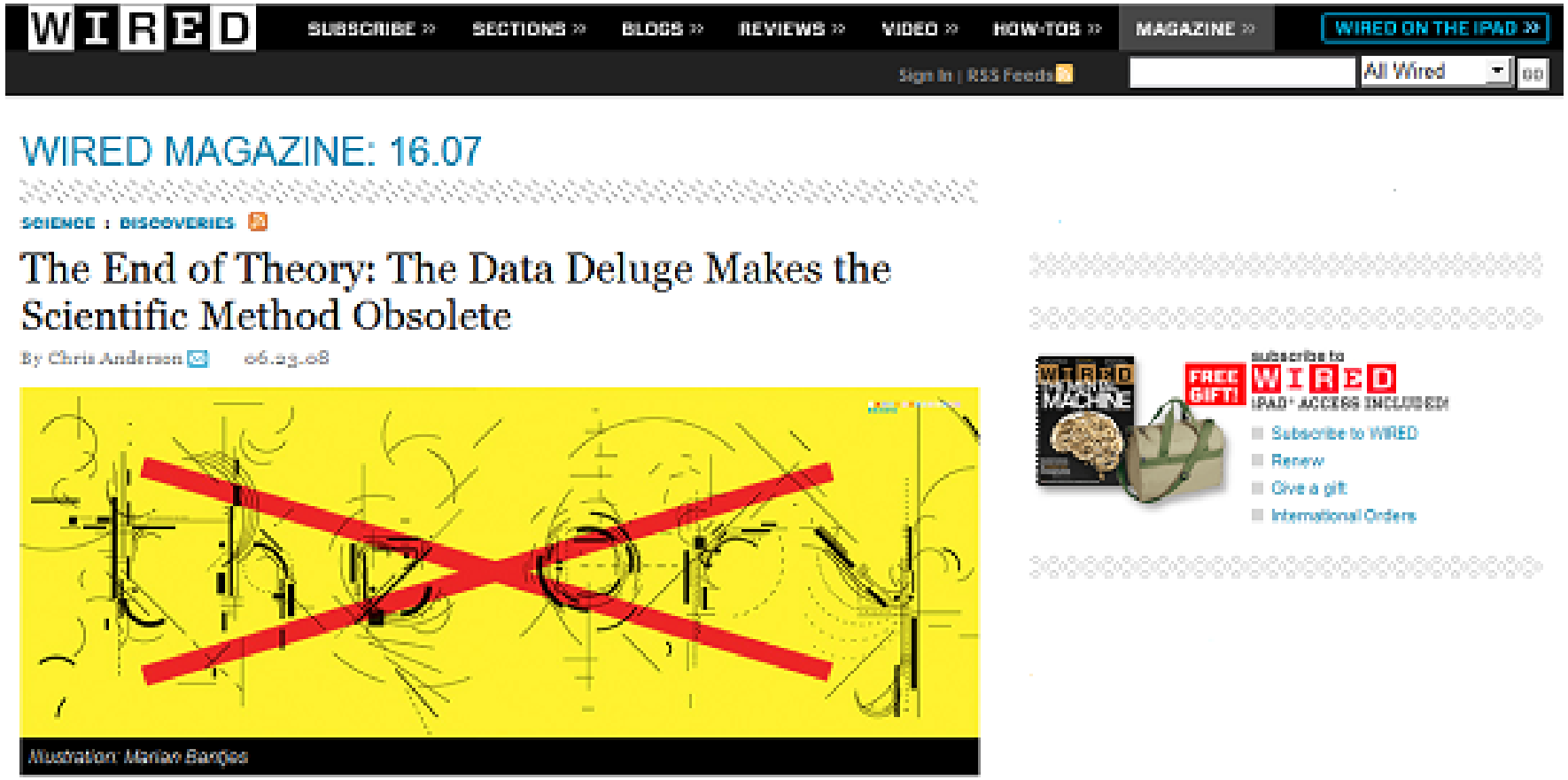
“Model” : an ambiguous term

- Generative model
  - Old and restrictive definition: a model describes the mechanism that produced the data. Usually requires a theory (physics, biology, etc.).
  - Often unworkable in business and economics
  - Recently: a machine learning model designed to create new data that is similar to its training data.
- Algorithmic model
  - Fits to data
  - Allows accurate forecasts
  - Explicit or implicit
- To understand or to predict? see Saporta (2008), Shmueli (2010)

- “Essentially, all models are wrong, but some are useful ”

  Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

George Box (1919-2013)



2008

Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

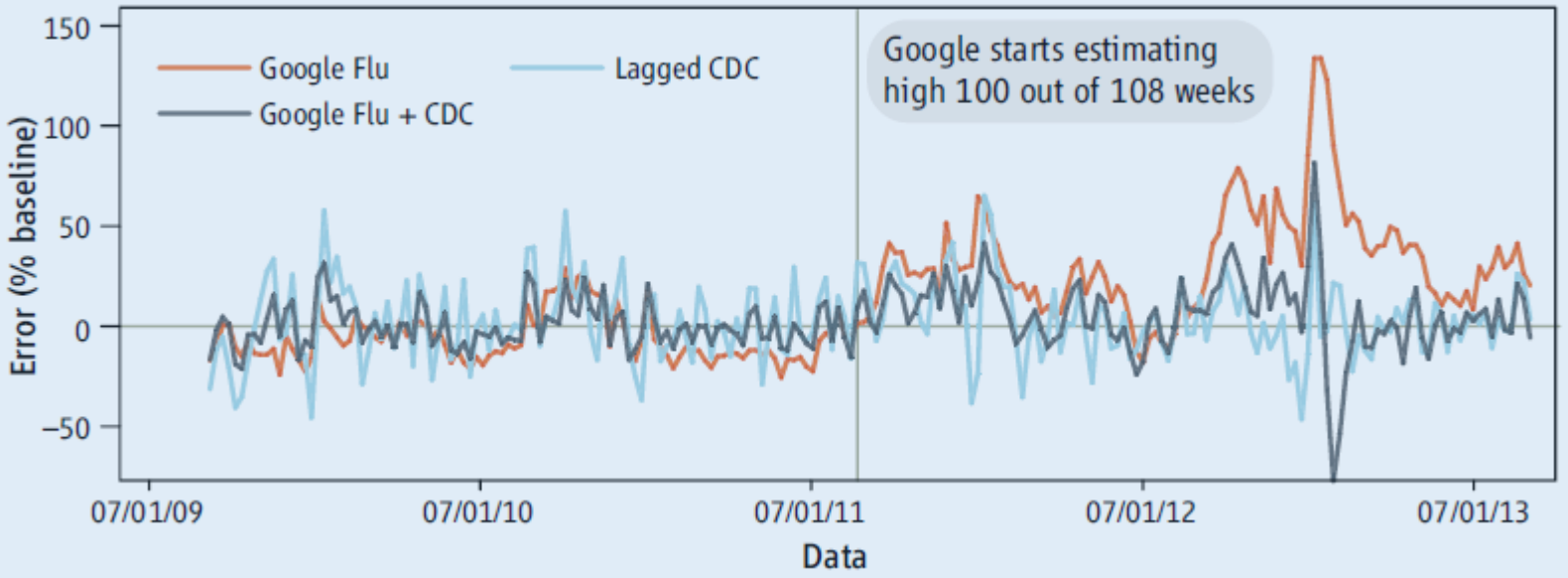
Source: DEARDAN&Friends

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014



Overestimation by 50% in 2012-2013

THE WALL STREET JOURNAL.

DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr

Updated July 1, 2015 3:41 pm ET

Black programmer Jacky Alciné said on Twitter that the new Google Photos app had tagged photos of him and a friend as gorillas.

ILLUSTRATION: JACKY ALCINÉ AND TWITTER

Reuters

World Business Markets Sustainability Legal More

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 11, 2018 2:50 AM GMT+2 · Updated October 11, 2018

Aa

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

Algorithm "biases"

- Algorithms often simply reproduce the biases of pre-existing learning data and human decisions.
- Statistical biases
  - Non-representative sample
  - Missing data and selection bias (e.g. loan applications)
- Prejudices: societal, cultural, cognitive, ...
  - The data may be representative, but may reproduce inequalities (women's wages) and stereotypes.
- Technological biases
  - Sensors not fitted for black skins

Robustness issues

- Despite the quantity of data
- Millions (or billions) of parameters in Deep Learning models

A one pixel attack…

SHIP
CAR(99.7%)

HORSE
FROG(99.9%)

DEER
AIRPLANE(85.3%)

DEER
AIRPLANE(49.8%)

BIRD
FROG(88.8%)

SHIP
AIRPLANE(88.2%)

HORSE
DOG(70.7%)

DOG
CAT(75.5%)

BIRD
FROG(86.5%)

HORSE
DOG(88.0%)

SHIP
AIRPLANE(62.7%)

CAT
DOG(78.2%)

CAR
AIRPLANE(82.4%)

DEER
DOG(86.4%)

CAT
BIRD(66.2%)

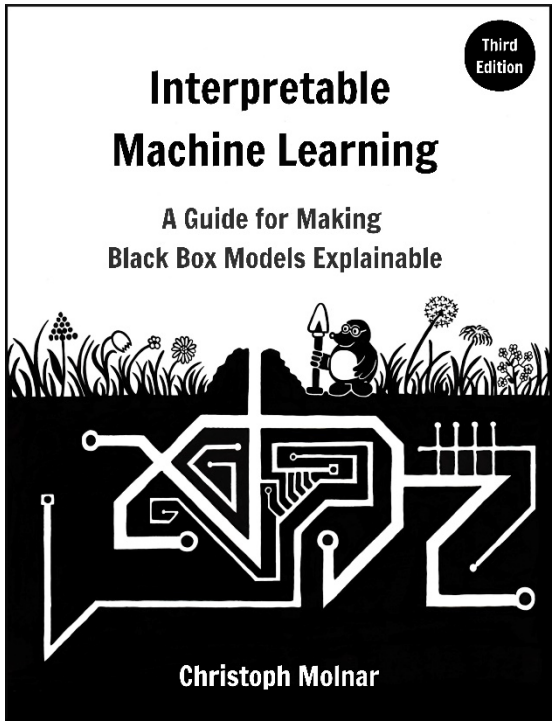
One pixel attack for fooling deep neural networks
Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi
https://arxiv.org/pdf/1710.08864.pdf, 2019

3. Explainable AI

- Is the "why" so important?
- In everyday life, we trust many processes that we do not understand: cars, television, smartphones, weather forecasts. No matter if black boxes are used.
- But when certain decisions have implications on our lives: health, employment, money, etc., the right to an explanation is necessary: UNESCO Recommendation (2021), EU AI Act (2024)

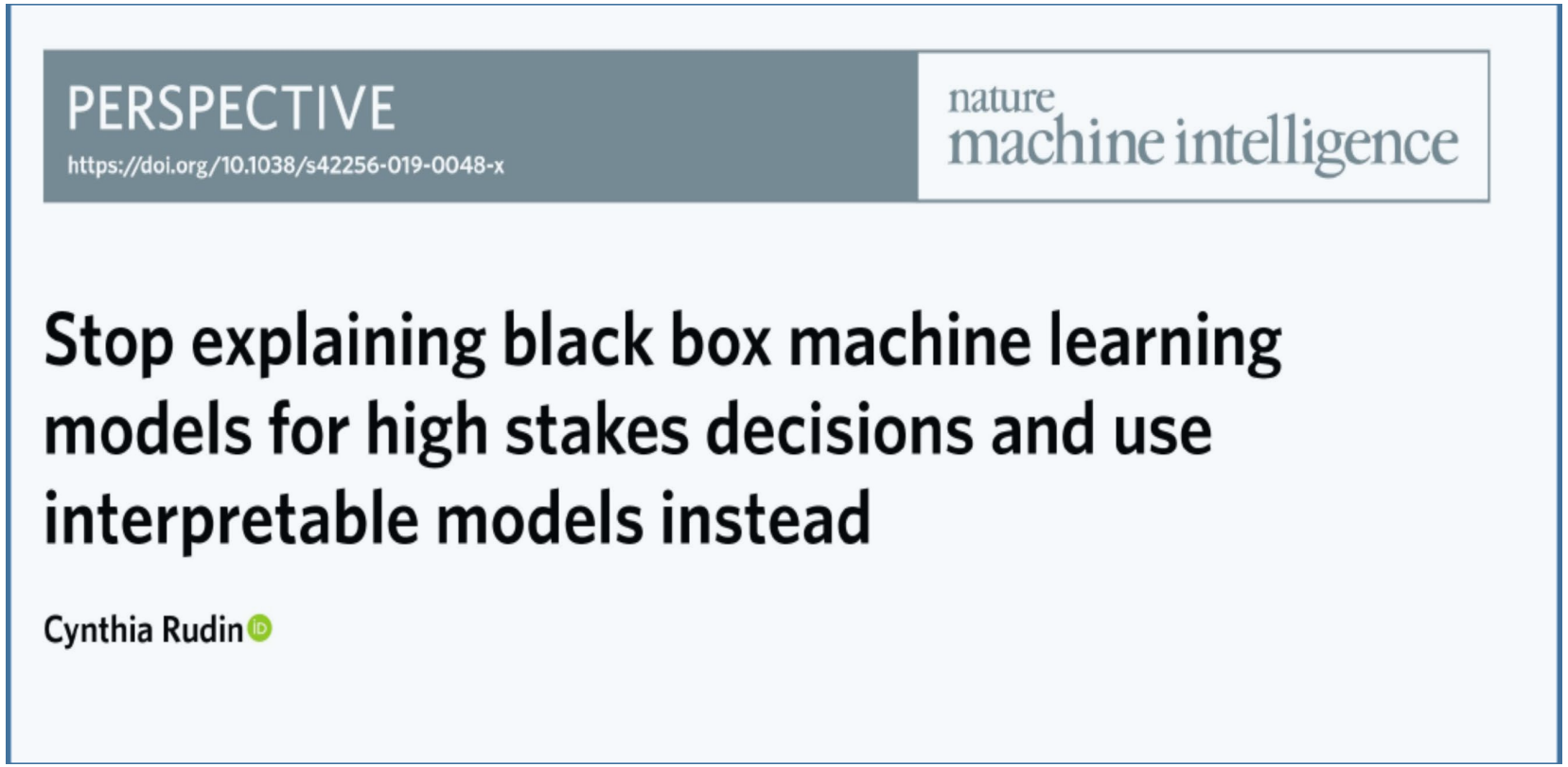
Interpretability vs explainability

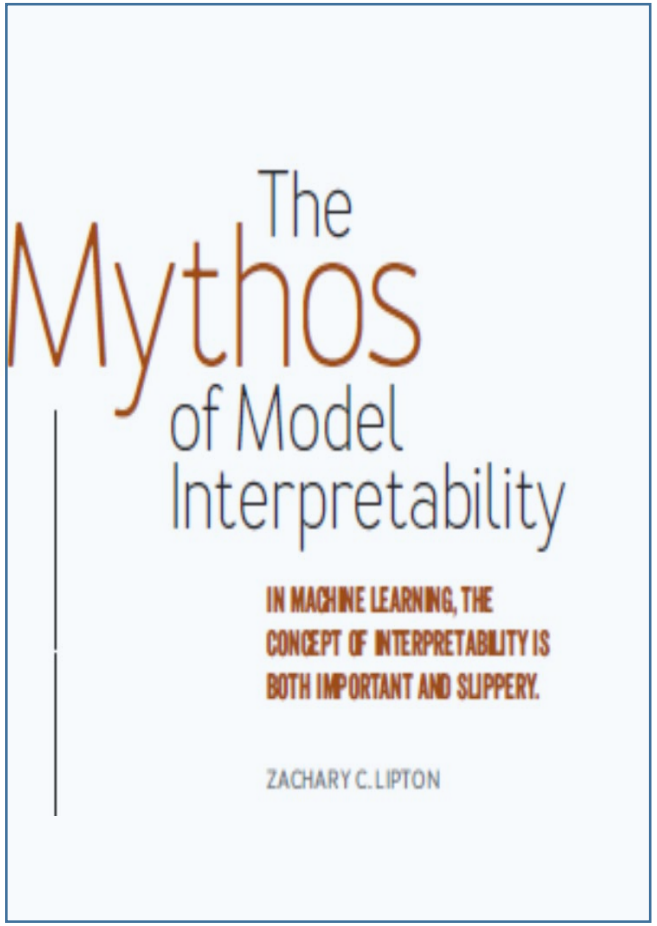
- Terms often used interchangeably
- Interpretability
  - Refers to **simple** and transparent algorithms: logical models (trees, ...), linear (sparse, ...), knn.
- Explainability
  - the ability to explain or present in terms understandable to a human being
  - Generally **post-hoc** (open the black box)
    - Local or global
    - Specific or **agnostic**



2025

controversies





3.1 Variable importance measures

3.1.1 Specific methods

- It is often believed that simple models, such as linear or logistic regression, are easily interpretable.
- Generally untrue!
- Except in the case of orthogonal designs, the parameter values hardly reflect the importance of the variables.

- More than 14 methods to quantify the importance of variables in linear models!(Grömping, 2015, Wallard, 2015)

`R` package `relaimpo`

b's (not normalized)
Joint contribution (not normalized)
Squared semipartial correlations
Squared raw correlations
Squared standardized b's
Sequential SS, from left to right
Sequential SS, from right to left
Pratt
CAR scores/Gibson
Green et al.
Fabbris
Genizi/Johnson
LMG
PMVD

3.1.2 Agnostic methods

Permutation variable importance (Breiman, 2001)

- Shuffling the values of each predictor provokes an increase of the prediction error
- The worse the performance, the more important this predictor was
- Easy to understand approach, applicable to any kind of model
- Must be repeated and averaged
- Importances are not additive
- Can lead to physically impossible unit pairs and outliers.

Shapley value

Inspired by game theory

Prediction task = *game*

Predictor = *player*

Predictor subset = *coalition*

Prediction = *payoff*

- Nice mathematical properties
  - Including additivity and uniqueness under certain conditions.
  - Allows to decompose an individual prediction (local values)
  - Global importance of a predictor: average of the local values on all units

3.2 Surrogate models

"A surrogate model is an interpretable model that is trained to approximate the predictions of a black box model" (Molnar, 2020)

- Can be global or local, agnostic or specific.
- Agnostic means that it can be applied to any learning model.
- A surrogate model tries to approximate the black box model, not to fit the data.
- Trees, linear models are the preferred alternative models.

- **Example: Linearizing a kernel classifier** (Liberati, Camillo, Saporta, 2017)

- A credit scoring case: 75 000 « good » and 10 000 « bad » small businesses asking an italian bank for a credit

  - The best classifier was a SVM with a Cauchy kernel

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + \mathbf{b} \quad K(\mathbf{x}_i, \mathbf{x}) = \frac{1}{1 + \|\mathbf{x}_i - \mathbf{x}\|^2}$$

  - Difficult to use. Professionals prefer an additive scoring rule

- Solution: Reconstruction of the kernel discriminant function through a linear regression where $f(\mathbf{x})$ is the target and the original variables are the predictors

Discriminant rules	Correct classification rates		
	Bad class	Good class	Overall
Cauchy	71.48	74.78	73.86
Logistic regression	54.91	89.88	79.90
FLDA	61.88	56.57	58.08
Reconstructed	73.80	74.51	74.30

4. New requirements: interpretable AND causal models

- Measuring the importance of a variable does not answer this question: what would be the answer if one or more predictors were changed intentionally or unintentionally?
  - Measuring the effect of a variable "all other things being equal" not always feasible.
  - Changing X_j may change the values of other predictors if they are causally related.

4.1 Confusion between correlation and causality

- Regression, ML models are not causal, but are often used as if they were, resulting in many disappointments...
- Seeing is not doing (Pearl & Mackenzie, 2018)

$$P(Y \mid X = x) \neq P(Y \mid do(X = x))$$

- Difficult to infer causality from observational data
  - Create pairs of twins, one treated, the other untreated.
  - Propensity score matching (Rosenbaum et Rubin, 1983)

4.2 Controlled experiments

- As Box et al. put it, "To find out what happens when you change something, it is necessary to change it." … the best way to answer causal questions is usually to run an experiment.
  - Eg AB Testing in marketing, web advertising (Bottou,2013)
- Experiments are easier in industry, agriculture than in economics, business or health studies because of the human factor:
  - Difficult to reproduce
  - A **counterfactual** question: How would an individual have behaved in the absence of treatment?

Causal inference and counterfactual reasoning

The basic identity (Varian, 2016):

Outcome for treated −Outcome for untreated
= [Outcome for treated −Outcome for treated if not treated]
+ [Outcome for treated if not treated – Outcome for untreated]
= Impact of treatment on treated + selection bias

Counterfactual part : « Outcome for treated if not treated" or what would have happened if they had not been treated?

- The basic identity shows the interest of **randomized trials**: selection bias has an expectation of zero, hence the possibility of estimating causal effect.

5. Machine Learning and Statistics: a mutual reinforcement

No antagonism between statistical modeling and Machine Learning. ML and Data Sciences are a 21st-century counterpart to statistics, just as the emergence of computer science in the second half of the 20th century led to the development of multivariate statistics followed by Big Data analytics.

5.1 ML enriched statistical thinking

Machine Learning has enriched the statistician's toolbox and, above all, has provided him with the fundamental concept of generalization and the need to go beyond simply fitting a model on the basis of training data alone.

The three samples procedure for selecting a a model inside a family of models

- Learning set: estimate parameters for the set of models in competition
- Test set : choice of the best model in terms of prediction error
  - NB Reestimation of the final model with all available observations
- Validation set : estimate the performance for future data. « Generalization »

NB Parameter estimation ≠ performance estimation

- **Elementary?**

  - Not that sure…
  - Have a look on publications in econometrics, epidemiology, .. prediction is rarely checked on a hold-out sample (except in time series forecasting)

5.2 Statistical culture and Machine Learning

- Statisticians can provide ML practitioners with:
  - their knowledge of biases and of robust methods,
  - their feel for data (missing, outliers, etc.), and their culture

  to avoid reinventing techniques such as categorical data encoding or principal component analysis!
- As B.Efron claimed: “Those who ignore Statistics are condemned to reinvent it”

6. Concluding remarks

- Challenges
    - High dimensional data
    - Categorical data
    - Will LLM replace statisticians?
    - Will synthetic (or *silicon*) data replace real data?

References

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.
- Bottou, L. et al. (2013). Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, 14, 3207–3260,
- Breiman, L., (2001). Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231
- Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, *26*(4), 745–766.
- Grömping, U. (2015). Variable importance in regression models. *WIREs Computational Statistics,* 7, 137-152.
- Liberati, C., Camillo, F., Saporta, G. (2017). Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Advances in Data Analysis and Classification*, *11*(1), 121-138.
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31-57.
- Molnar, C. (2025). *Interpretable machine learning , A Guide for Making Black Box Models Explainable*, https://christophm.github.io/interpretable-ml-book.

- Pearl, J., Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.
- Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215
- Rudin, C., Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, *1*(2).
- Saporta, G. (2008). Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- Schölkopf, B., von Kügelgen, J. (2022). From statistical to causal learning. In *Proceedings of the International Congress of Mathematicians* (Vol. 7, pp. 5540-5593)
- Shmueli, G. (2010). To explain or to predict?. *Statistical science*, *25*(3), 289-310.
- Srinivasan, R., Chander, A. (2021). Biases in AI Systems: A survey for practitioners. *Queue*, *19*(2), 45-64
- Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*, (translated by Samuel Kotz), Springer
- Varian, H. (2016). Causal inference in economics and marketing, *PNAS* , 113, 7310-7315
- Wallard, H. (2015). Using explained variance allocation to analyse importance of predictors. In *16th ASMDA conference proceedings* (Vol. 30), 1043-1054,