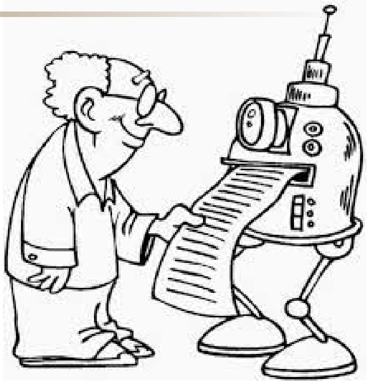
Revisiting models and uncertainty with AI

Science with AI

Gaël Varoquaux



My scientific wanderings

Physics

- Quantum physics

  Atom-interferometric tests of relativity

Brain image analysis for cognition

- Statistics, machine learning, image analysis
- Cognitive neuroscience, psychology

Machine learning for public health

  Informing policy?

From absolute quantities
to qualitative subject matters

Questions of interest

How does scientific knowledge emerge from data?

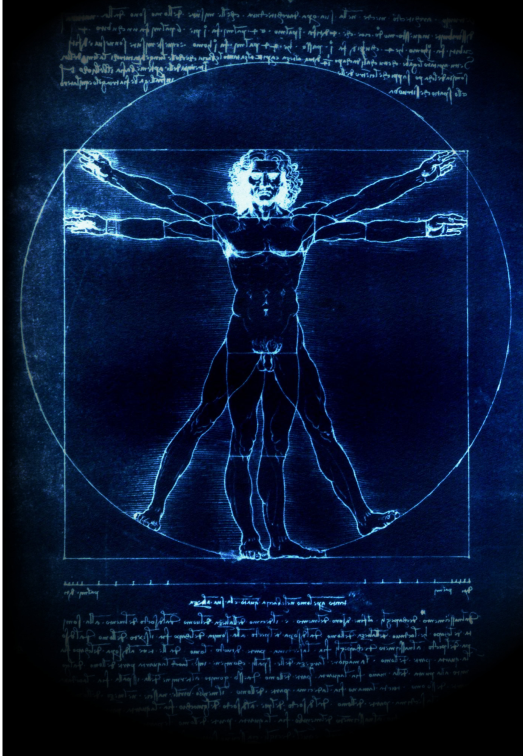
Can we have a statistical control on this process?

What role do models play?

This talk

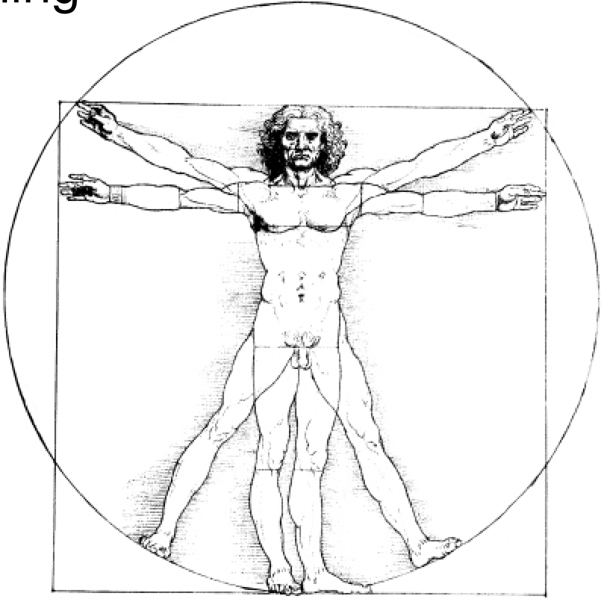
1 Rethinking modeling

2 Model uncertainty and validation



1 Rethinking modeling

AI as statistical methods
for imperfect theories



Scientific progress and statistical evidence

Dominant framework of statistical reasoning:

- Formulating a probabilistic model from mechanical hypotheses
- Integrating empirical evidence (data) by fitting this model
- Reasoning from model parameters

Rigour breaks down with wrong modeling ingredients

Science needs more reasoning from model outputs

- For statistics: robustness to mis-specification
- Generalization grounds scientific theories

Black-box phenomenological data models are good for science

Statistical evidence in science and data science

1. Model the data

Based on the knowledge and constructs of the field

& the understanding of data collection

$$m\frac{d^2}{dt^2}\vec{x} = \vec{F}$$

$$\vec{F} = q\,(\vec{E} + \frac{d}{dt}\vec{x} \times \vec{B})$$

Intelligence

Fluid intelligence

Crystallized intelligence

Statistical evidence in science and data science

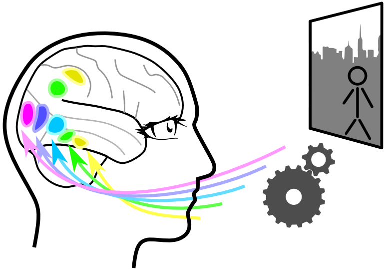
1. Model the data

   Based on the knowledge and constructs of the field
   & the understanding of data collection

2. Statistical inference
   - Fit model to data (typically maximizing likelihood)
   - Reason from the model and its parameters

Relies on statistical modeling [Cox 2006]

Example: studying brain brain activity

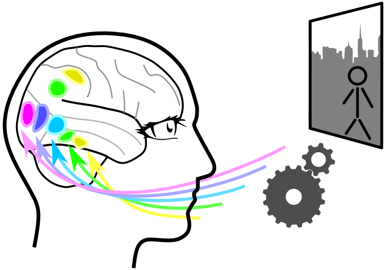


Neural support of mental process

Model of task and mental processes

$\Rightarrow$ brain maps

Example: studying brain brain activity

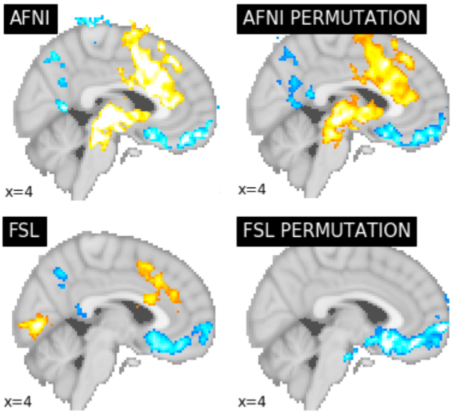


Neural support of mental process

Model of task and mental processes
$\Rightarrow$ brain maps

Uncontrolled variability

- In modeling across teams
  [Botvinik-Nezer... 2019]
- Across software for same model
  [Bowring... 2019]



Even experts cannot chose the “right” model

Teachings from history of science

Current view of physics, chemistry...

Building models from the right ingredients – “first principles”

The past

Refining relevant constructs from wrong models

The birth of mechanics

Early scientists (*eg* ancient Greece)

"natural motion of objects", no notion of force, or acceleration.

Observation of planetary motion (*eg* Kepler)

Search for regularities in planets – "harmonies"

The period squared is proportional to the cube of the major diameter of the orbit

Modern laws of dynamics (Newton)

Differential calculus ⇒ laws with force and acceleration

Unite observations of celestial and earthly motions

The birth of mechanics

Early scientists (*eg* ancient Greece)

"natural motion of objects", no notion of force, or acceleration.

Lacking key ingredients

Observation of planetary motion (*eg* Kepler)

Search for regularities in planets – "harmonies"

The period squared is proportional to the cube of the major diameter of the orbit

Phenomenological model[1] **crucial**

Modern laws of dynamics (Newton)

Differential calculus ⇒ laws with force and acceleration

Unite observations of celestial and earthly motions

Validity established by strong generalizability

Modern physics does not need phenomenological models?

Vulcan: false discovery of a planet (19th century)

Anomaly in Mercury's orbit not explained by Newtonian physics

$\Rightarrow$ invent and "observe" an additional planet, Vulcan

Theory laden observations

Modern physics does not need phenomenological models?

Vulcan: false discovery of a planet (19th century)

Anomaly in Mercury's orbit not explained by Newtonian physics

$\Rightarrow$ invent and "observe" an additional planet, Vulcan

Theory laden observations

Particle physics builds evidence with machine learning (today)

Fundamental laws of the universe = most precise theory ever

Particle detection by discriminating physics model
with non-parametric background

"Pure" models insufficient for "dirty" reality

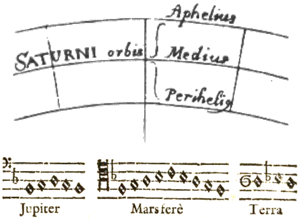
Phenomenological data fits have been crucial to science

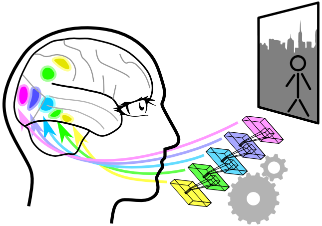
- Science uses false models as means for truer theory

  [Wimsatt 2007]

- The reductionist aesthetics of “pure” simple mathematical theories is not adapted to the messy world beyond pure physics

- Generalization or prediction failures make or break scientific theories





Statistics and scientific evidence

- Validity
- Reasonning

  = more than formal problems

Validity of scientific findings – much more than statistical validity

External validity

[Cook and Campbell 1979]

External validity asserts that findings apply beyond the study

Generalizability

Validity of scientific findings – much more than statistical validity

External validity [Cook and Campbell 1979]

External validity asserts that findings apply beyond the study

Generalizability

Constructs and their validity [Cronbach and Meehl 1955]

- Construct = abstract ingredients such as "intelligence"
- Construct validity: measures and manipulations actually capture the theoretical construct

Validity of scientific findings – much more than statistical validity

External validity [Cook and Campbell 1979]

External validity asserts that findings apply beyond the study

Generalizability

Constructs and their validity [Cronbach and Meehl 1955]

- Construct = abstract ingredients such as "intelligence"
- Construct validity: measures and manipulations actually capture the theoretical construct

Implicit realistic stances in theories

Realism = objective and mind-independent unobservable entities

Is intelligence a valid construct? How about a center of gravity?

Places implicit preferences on models beyond empirical evidence

Reasoning with statistical tools

Model reasoning [Cox 2006]

- Carefully craft a probabilistic model of the data
- Estimated model parameters are interpreted within its logic
  "data descriptions that are potentially causal" [Cox 2001]

Warranted reasoning [Baiocchi and Rodu 2021]

- Relies on warrants in the experiment (*eg* randomization)

Output reasoning [Breiman 2001, Baiocchi and Rodu 2021]

- Relies on capacity to approximate relations

Benefits of reasoning on outputs rather than models

Science needs black-box output reasoning

For statistical validity

Even expert modeling choices explore meaningful variability

- Model reasoning is conditional to the model
  parameters have a meaning in a model
- Imperfect science: 70 different teams of brain-imaging experts
  qualitatively different neuroscience findings [Botvinik-Nezer... 2020]

Analytical variability breaks statistical control

Output reasoning: milder conditions for statistical control

- Theoretical results in mispecified settings [Hsu... 2014]
- Multi-colinearity no longer an issue
- Higher-dimensional settings

$\Rightarrow$ Forces less reductionist choices

For understanding?

“Nobody understands quantum mechanics” Richard Feynman

Narrative truth versus operational truth

Humans need stories, for teaching, for intuitions, for “selling”

these simplifications are not “truth”

For ~~understanding~~ counterfactual reasoning

"Nobody understands quantum mechanics" Richard Feynman

Narrative truth versus operational truth

Humans need stories, for teaching, for intuitions, for "selling"

these simplifications are not "truth"

Counterfactual reasoning & causal inference

- We want to reason on new situations
- Causal, not correlational knowledge

  Bad health is *associated with* hospitals, but seldom *caused by*.
- Predictive models enable counterfactual reasoning if
  - they extrapolate enough
  - they build on the right variables (confounds, not colliders)

[Rose and Rizopoulos 2020, Doutreligne and Varoquaux 2023]

For broader scientific validity of findings

The only strong evidence is strong generalization

Model reasoning favors internal validity

Model reasoning often need “pure” models with little generalization

Fields without a unifying formal theory
tackle empirical evidence with overly reductionist lenses

Machine learning/AI can model the full problem space
and give testable generalization

For broader scientific validity of findings

The only strong evidence is strong generalization

Model reasoning favors internal validity

Model reasoning often need “pure” models with little generalization

Fields without a unifying formal theory
tackle empirical evidence with overly reductionist lenses

Machine learning/AI can model the full problem space
and give testable generalization

Relating to more general constructs

Theories & models are written in terms of constructs (*eg* attention)

To help generalizing across vastly different situations

Must ground these directly on observations

2 Model uncertainty and validation

Scientific criticism and reasonning on model output

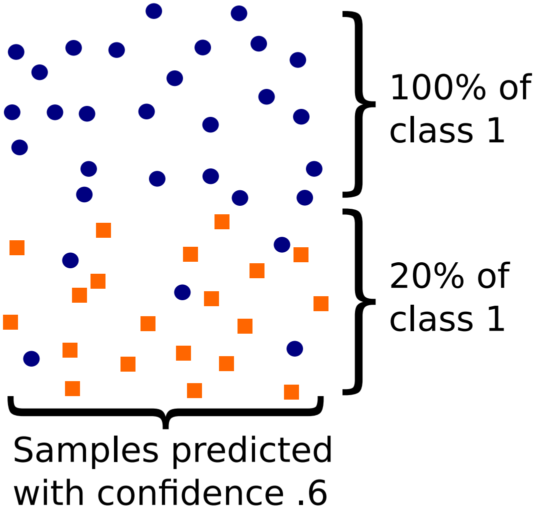
Controlling uncertainty on predictions

Applications need full
probability of error

In medicine:
harm-benefit trade-offs
[Vickers... 2016]

Calibration controls: Average error rate for all samples with score s is s

A calibrated classifier can assign **a score of .6** to individuals, but be **100% accurate on a subgroup**, and **20%** on another.



⚠ Calibration does not control individual probabilities

Metrics controlling individual probabilities

Does the classifier approach $\mathbb{P}(y|X)$?

$\mathbb{P}$ is never observed, only discrete events

Proper scoring rules

Observed (binary) label

$$\text{Brier score} = \sum_i (\hat{s}_i - y_i)^2$$

Confidence score

Minimal for $\hat{s} = P(y|X)$ (also log-loss)

Drawback: what is "good enough"?

- cannot be interpreted as an error rate
- no scale

Decomposing scoring rules into error rates

[Perez-Lebel... 2023]

Scoring rules (*eg* Brier) compound multiple aspects of error

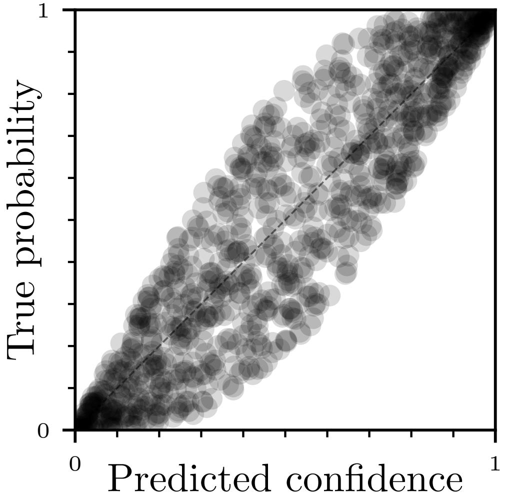
- Classifier output: $S = f(X)$
- Label probabilities: $Q = \mathbb{P}[Y|X]$
- Calibrated score[1]: $C = \mathbb{E}\Big[\mathbb{P}[Y|X]\Big|\, S\Big]$

1 Knowing the classifier output, what's the label probabilities

Scoring rule decomposition

$$\mathbb{E}\left[d(\underset{\text{Classifier output}}{S}, \underset{\text{Label distribution}}{Y})\right] = \underbrace{\underbrace{\mathbb{E}\left[d(S, \overset{\text{Calibrated score}}{C})\right]}_{\text{Calibration error}} + \underbrace{\mathbb{E}\left[d(C, \overset{\text{Expected label}}{Q})\right]}_{\text{Grouping error}}}_{\text{Epistemic error = distance to best achievable prediction}} + \underbrace{\mathbb{E}\left[d(Q, Y)\right]}_{\text{Irreducible error}}$$

An oracle calibration plot



No calibration error

$$\frac{\text{On average}}{\text{predicted confidence}} = \text{true probability}$$

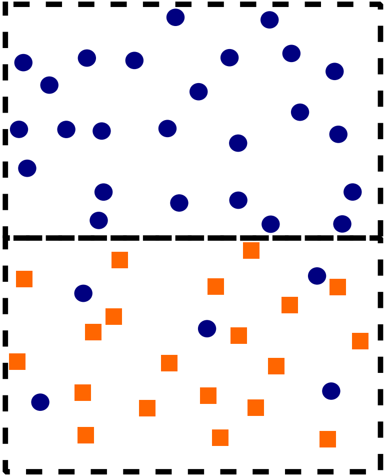
Grouping error

Classifier over-confident on some samples, under-confident on others

Measures the *dispersion* of scores

Requires access to true probabilities

[Perez-Lebel... 2023]



P(Y=1|X)
= 1

P(Y=1|X)
= .2

Estimating true probabilities on well-chosen bins

(and controlling errors due to binning)

Unlike Brier: the ideal classifier has zero grouping loss

removes the irreducible error

Controlling uncertainty on predictions

Application need full probability of error

Controling the individual probability is possible [Perez-Lebel... 2023]

Controlling more than the binary decision

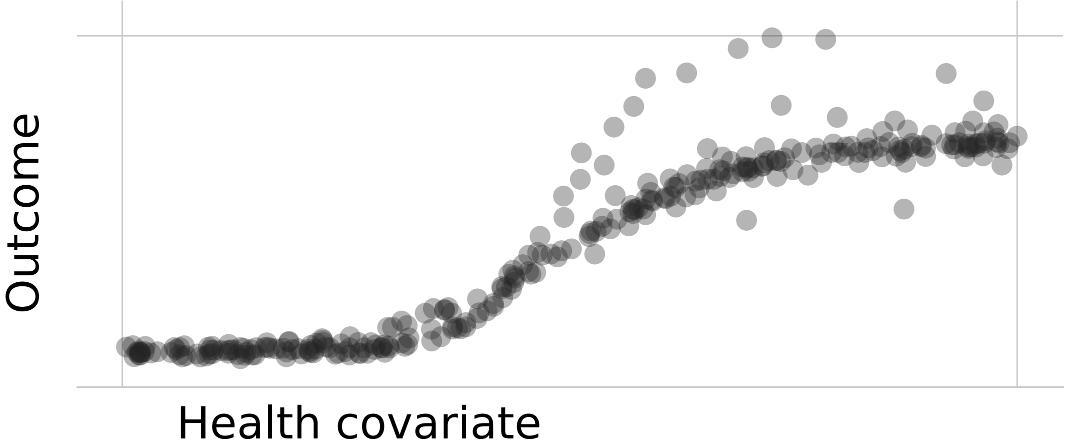
Machine-learning validation is a proxy of the error of interest

Broader question: estimating application risks

Prediction
to support decision

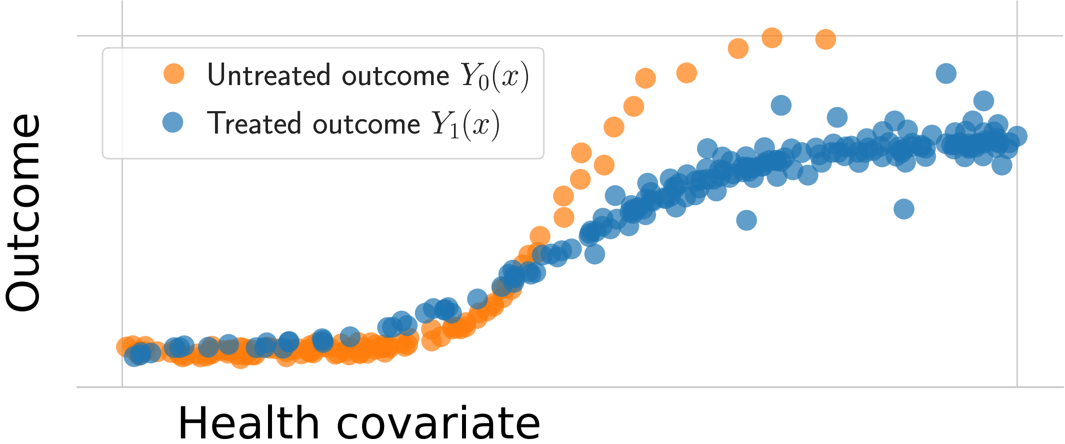
when predictors should be causal

Prognostic model: predicting a health outcome



Predictors and causal effects

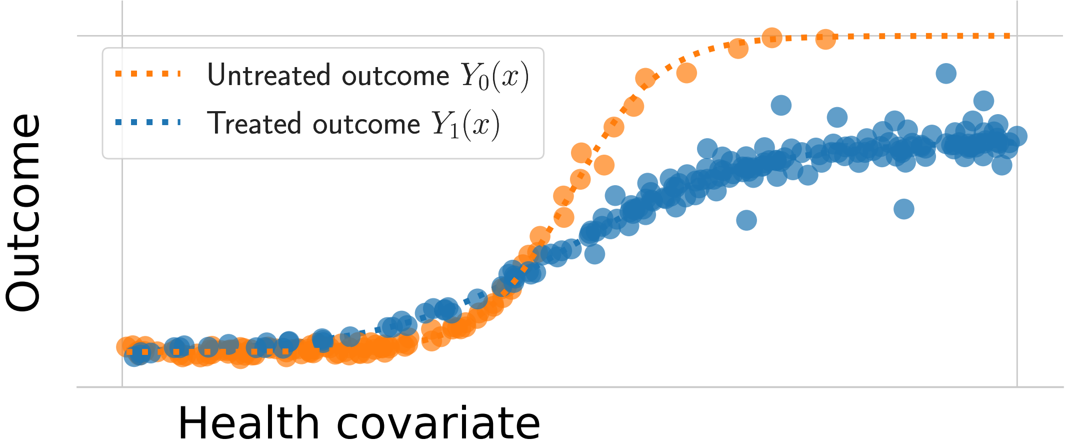
Prognostic model: predicting a health outcome



- Prediction function of intervention (treated $Y_0(x)$ vs untreated $Y_1(x)$)

Predictors and causal effects

Prognostic model: predicting a health outcome

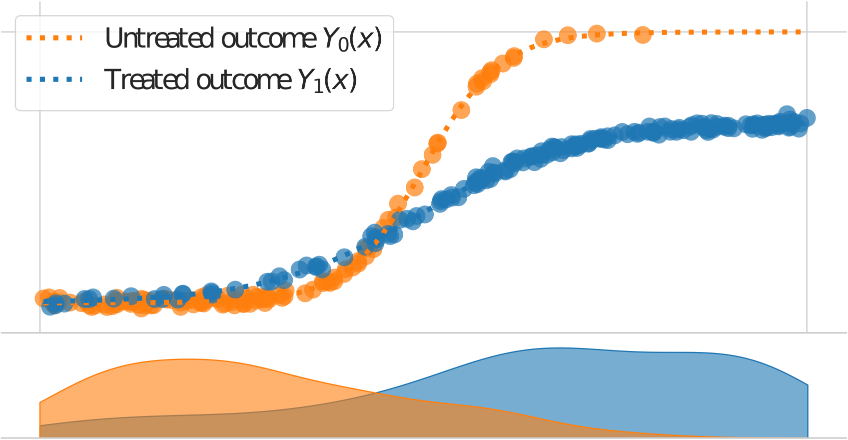


- Prediction function of intervention (treated $Y_0(x)$ vs untreated $Y_1(x)$)

- For decisions: Individual treatment effect:
  comparing predicted outcomes for the same individuals

Causal inference: distribution shift

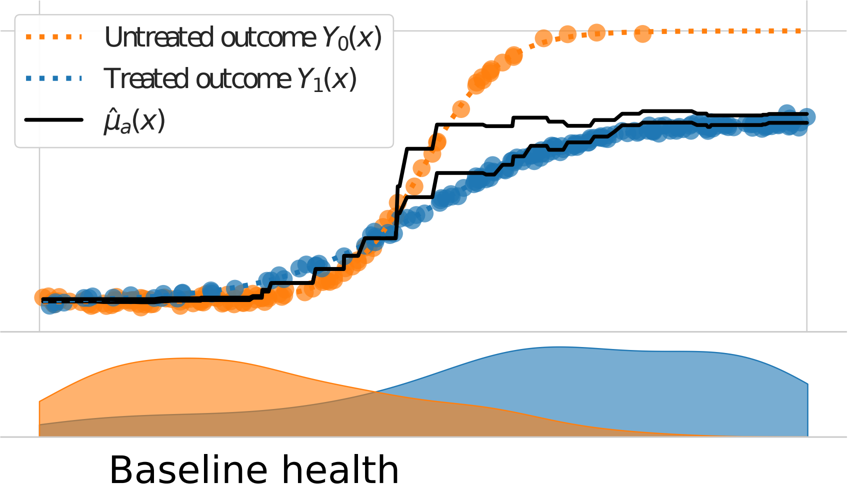
[Doutreligne and Varoquaux 2023]



- Healthy individuals did not receive the treatment

Causal inference: distribution shift

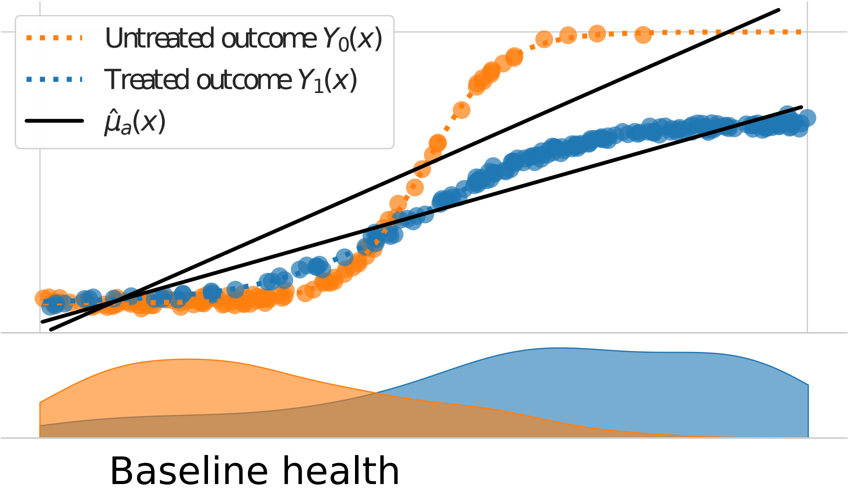
[Doutreligne and Varoquaux 2023]



- Healthy individuals did not receive the treatment
- The model associates treatment to negative outcomes

Causal inference: distribution shift

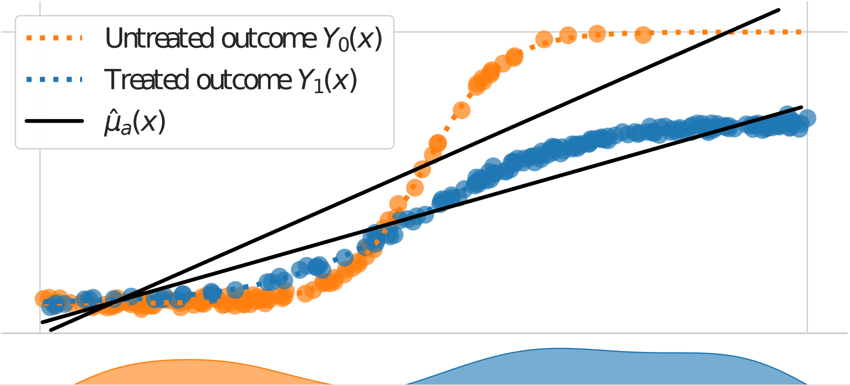
[Doutreligne and Varoquaux 2023]



- Healthy individuals did not receive the treatment
- The model associates treatment to negative outcomes
- A worse predictor gives better causal inference

Causal inference: distribution shift

[Doutreligne and Varoquaux 2023]



Standard cross-validation / predictive accuracy not good

Must weight equally errors on treated vs untreated outcome

- Healthy individuals did not receive the treatment
- The model associates treatment to negative outcomes
- A worse predictor gives better causal inference

Selecting predictors for treatment

[Doutreligne and Varoquaux 2023]

Lemma – rewriting of outcome model:

$$\text{(R-decomposition)} \qquad y(a) = m(x) + \big(a - e(x)\big)\tau(x) + \varepsilon(x; a)$$

$$\text{(Conditional mean outcome)} \qquad m(x) \stackrel{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x],$$

$$\text{(Propensity score)} \qquad e(x) \stackrel{\text{def}}{=} \mathbb{P}[A = 1|X = x].$$

Model-selection procedure

1. Compute m and e on train set (with standard ML tools)
2. On test set, use adjusted risk ("doubly robust"):

$$R\text{-risk}(f) = \mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\Big[\Big(\,(Y - m\,(X)) - (A - e\,(X))\,\tau_f\,(X)\,\Big)^2\Big]$$

[Nie and Wager 2021]

Prediction
to support decision

- A causal question
- R-risk

Raising the bar

Machine learning research

- Addressing distribution shifts
- Better model validation

Beyond technosolutionism

- Stop the overfitting
- Right focus
- Right incentives

The soda team: Machine learning for health and social sciences

Machine learning for statistics
Causal inference, biases, missing values

Health and social sciences
Epidemiology, education, psychology

Tabular relational learning
Relational databases, data lakes

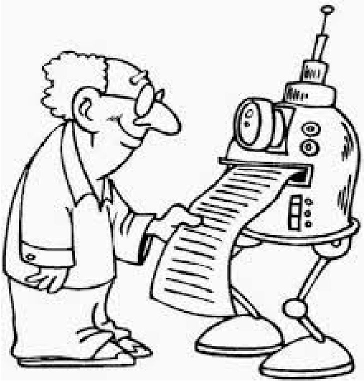
Data-science software
scikit-learn, joblib, skrub

AI gives statistical methods for imperfect theories [Varoquaux 2021]

- Model reasoning has no guarantees for imperfect models
- Scientific roadblocks are on model ingredients, not functional forms

- Gauge models more on their predictions than their ingredients
- Scientific inference from model predictions as in [Eickenberg... 2017]

counterfactual reasoning, model comparison, feature importances



Model validation from outputs

- Uncertainty beyond calibration [Perez-Lebel... 2023]
- Causal reasonning [Doutreligne and Varoquaux 2023]
- Machine-learning evaluation [Varoquaux and Colliot 2023]

@GaelVaroquaux

References I

M. Baiocchi and J. Rodu. Reasoning using data: Two old ways and one new. *Observational Studies*, 7(1):3–12, 2021.

R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, A. Adcock, ... Variability in the analysis of a single neuroimaging dataset by many teams. *bioRxiv*, 2019.

R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, ... Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810):84–88, 2020.

A. Bowring, C. Maumet, and T. E. Nichols. Exploring the impact of analysis software on task fmri results. *Human brain mapping*, 40(11):3362–3384, 2019.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

T. Cook and D. Campbell. *Quasi-experimentation: Design and analysis issues for field settings 1979 Boston*. MA Houghton Mifflin, 1979.

References II

D. R. Cox. [statistical modeling: The two cultures]: Comment. *Statistical science*, 16(3): 216–218, 2001.

D. R. Cox. *Principles of statistical inference*. Cambridge university press, 2006.

L. J. Cronbach and P. E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52:281, 1955.

M. Doutreligne and G. Varoquaux. How to select predictive models for decision making or causal inference? 2023. URL `https://hal.science/hal-03946902`.

M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

D. Hsu, S. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14, 2014.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

A. Perez-Lebel, M. L. Morvan, and G. Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. *ICLR*, 2023. URL `https://arxiv.org/abs/2210.16315`.

References III

S. Rose and D. Rizopoulos. Machine learning for causal inference in biostatistics. *Biostatistics*, 21(2):336–338, 2020.

G. Varoquaux. Ai as statistical methods for imperfect theories. In *NeurIPS 2021 AI for Science Workshop*, 2021.

G. Varoquaux and O. Colliot. Evaluating machine learning models and their diagnostic value. In *Machine learning and brain disorders*. 2023.

A. J. Vickers, B. Van Calster, and E. W. Steyerberg. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, 352, 2016.

W. C. Wimsatt. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press, 2007.