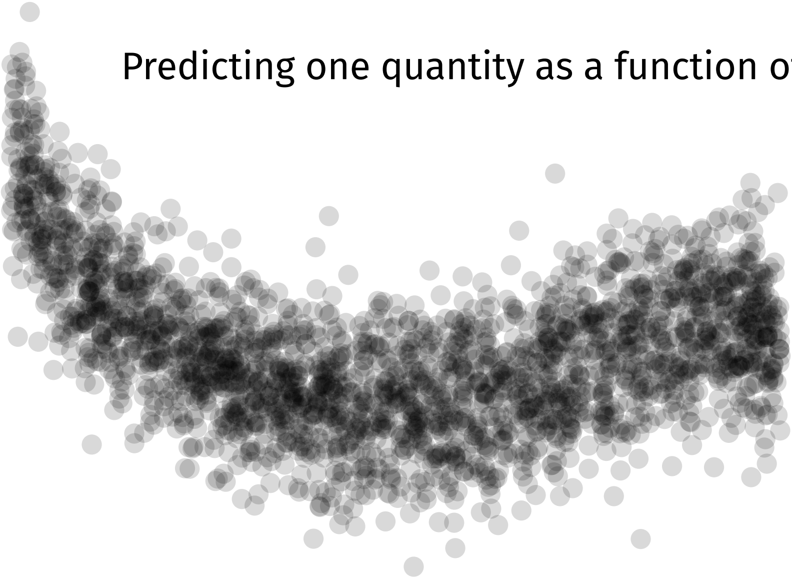
Machine learning for health:

promises and methodological challenges

Gaël Varoquaux

Inria :probabl.

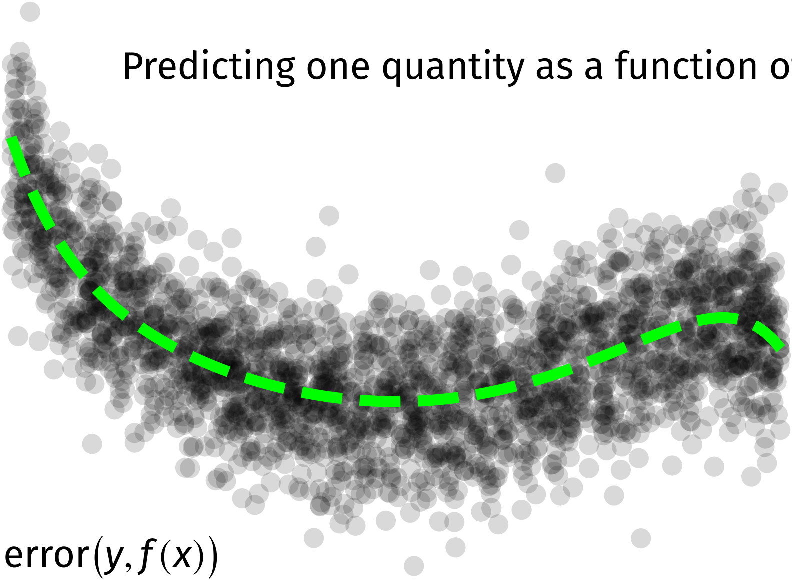
Statistical learning



Predicting one quantity as a function of others

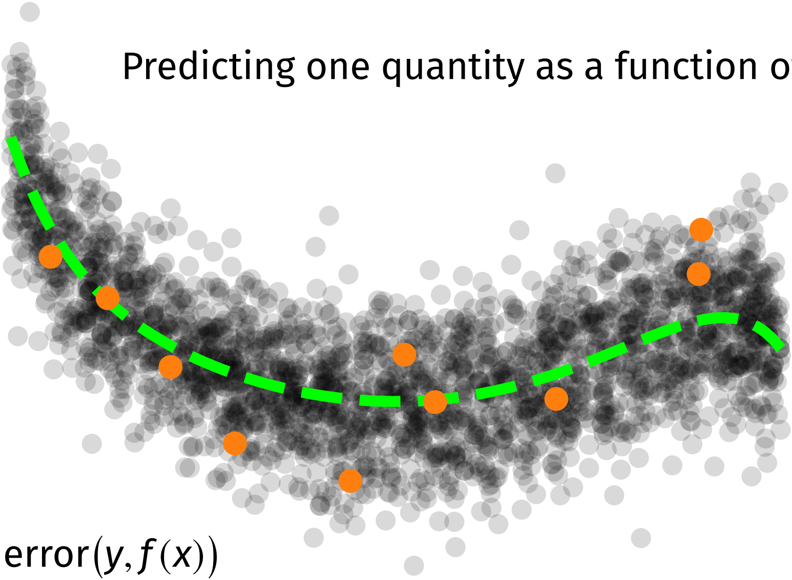
Statistical learning

Predicting one quantity as a function of others



$$\min_f \sum \mathrm{error}\big(y, f(x)\big)$$

Statistical learning

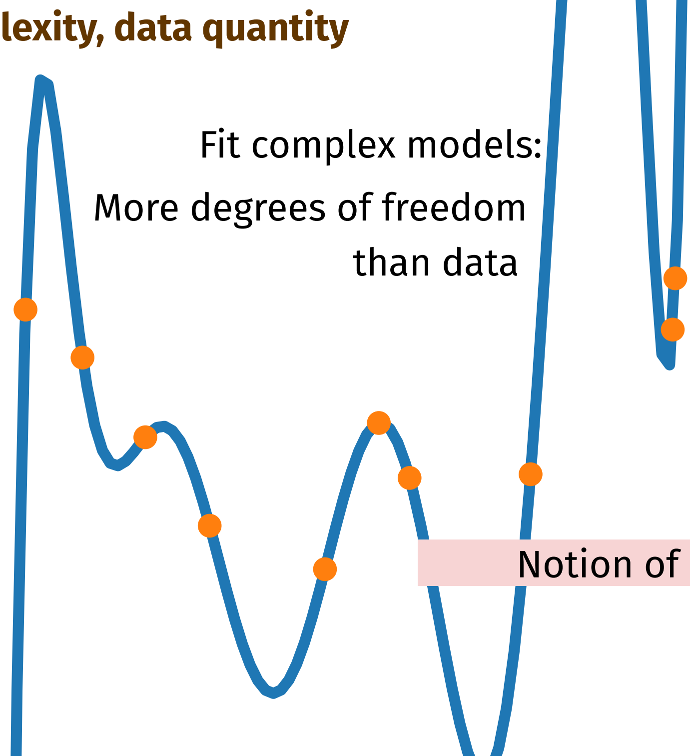


Predicting one quantity as a function of others

$$\min_f \sum \text{error}(y, f(x))$$

$\sum \neq \mathbb{E}$ Small data $\Rightarrow$ sampling noise

Model complexity, data quantity

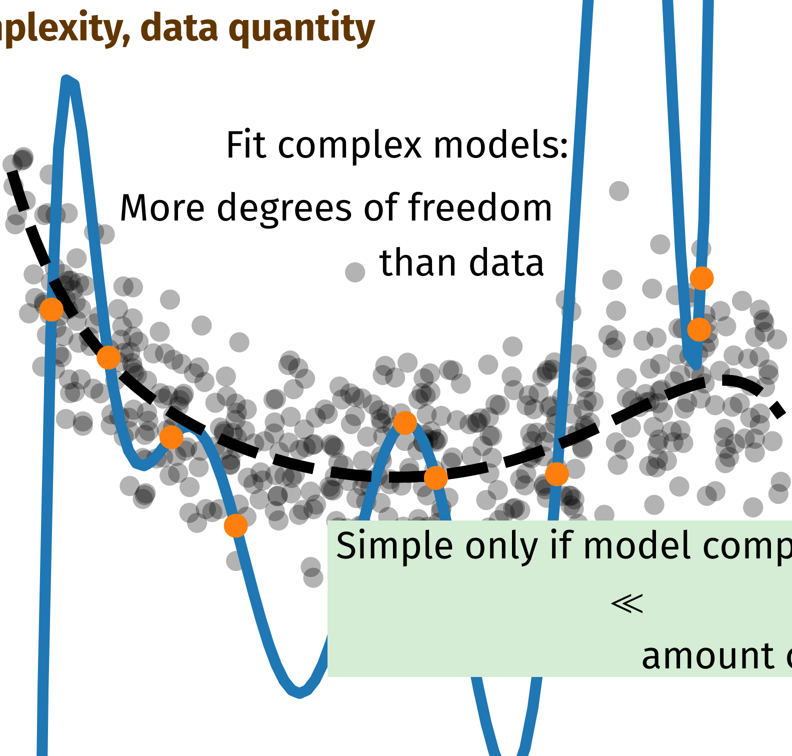


Fit complex models:

More degrees of freedom than data

Notion of overfit

Model complexity, data quantity



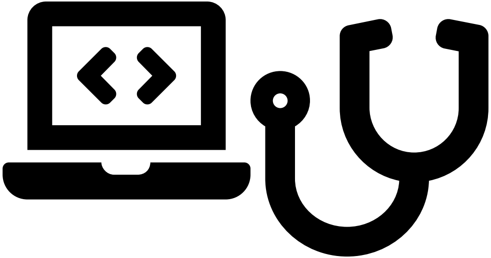
Fit complex models:

More degrees of freedom than data

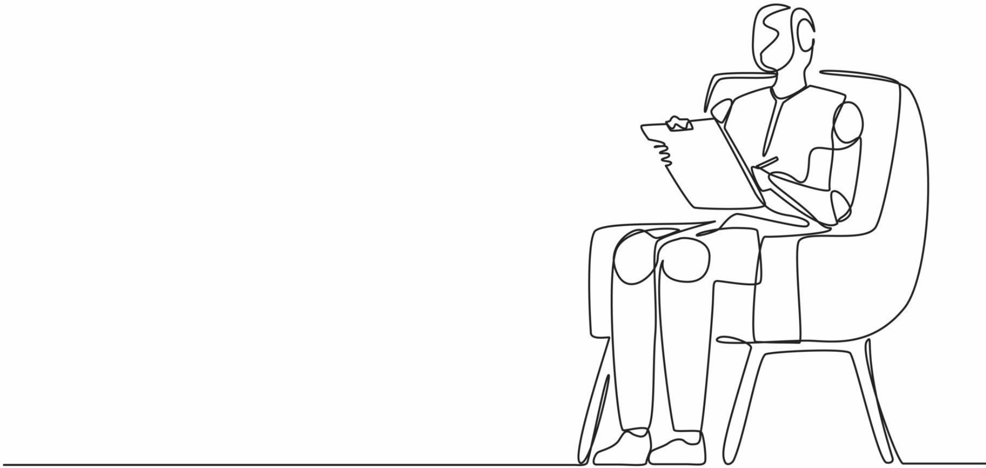
Simple only if model complexity ≪ amount of data

1. Machine learning on health data

2. Bridging the data to the application



1 Machine learning on health data



Classic machine learning tasks in medicine

Diagnostic models

From complex / incomplete data, describe patient's status

Prognostic models

Predict future evolution

Medical imaging

- Very complex data
  - High dimensional
  - Structured individual variability

- Typically, diagnostic tasks
  - “the automated radiologist”
  - seldom long-term outcomes

Medical imaging

- Very complex data
  - High dimensional
  - Structured individual variability

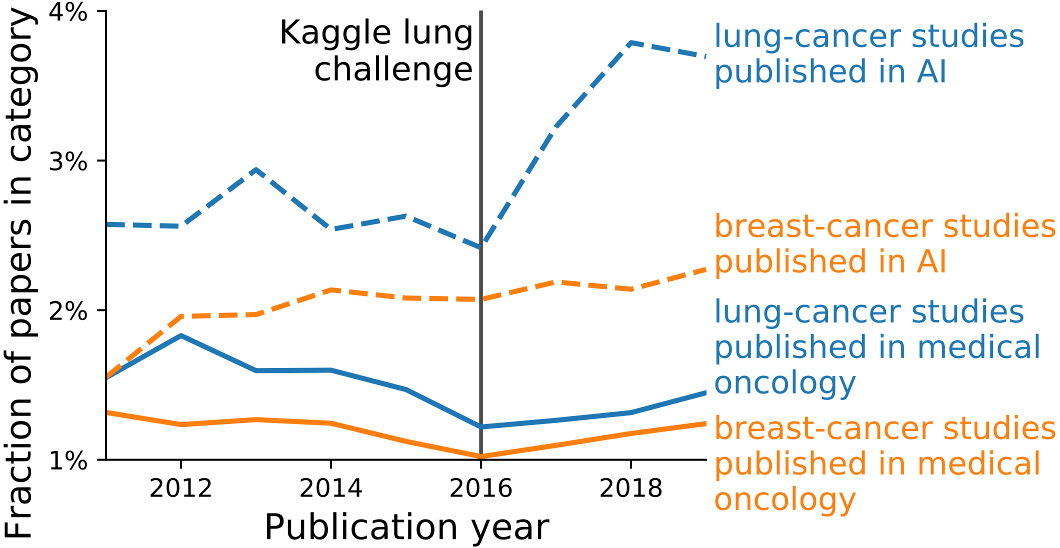
- Typically, diagnostic tasks
  - “the automated radiologist”
  - seldom long-term outcomes

Too little too late

- Data very infrequent n~1000
- Only the broken ones

Very expensive data

Driven by data availability, more than clinical relevance



A data challenge changes the field's focus

[Varoquaux and Cheplygina 2022]

Imaging is a fraction of patients' information

- An image is used within a *context*

- Cheaper data is predictive

  Questionnaires predict better mental health than brain images [Dadi... 2021]

Electronic Health records

Routine care
and administrative data

- Biological exams, doctors notes...
- Accounting, claims
- *Everything* in the hospital

Data “free”,
with a very good coverage

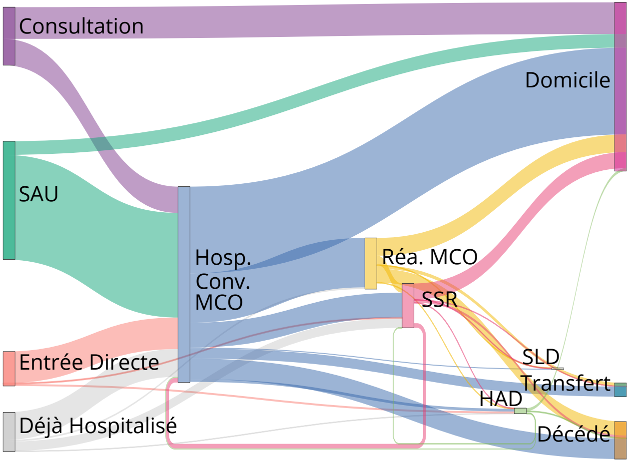
AP-HP (Paris hospitals)

- 39 hospitals
- 8 M patients per year

Covid outbreak: Hospital management

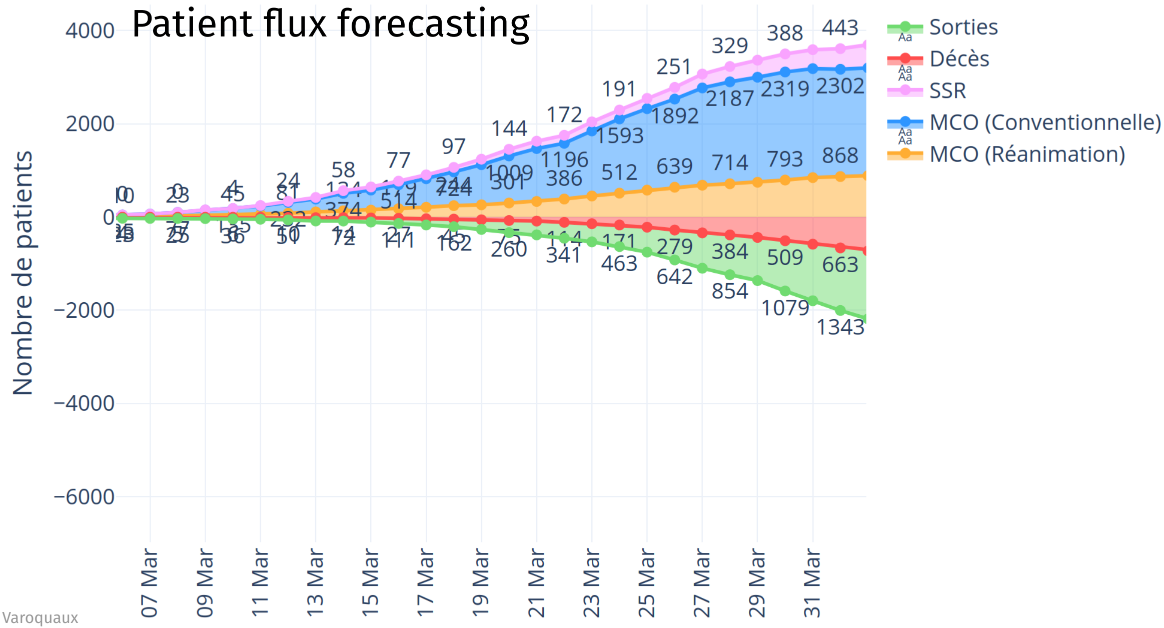
Inform
hospital-level decisions

Covid+ patient flux

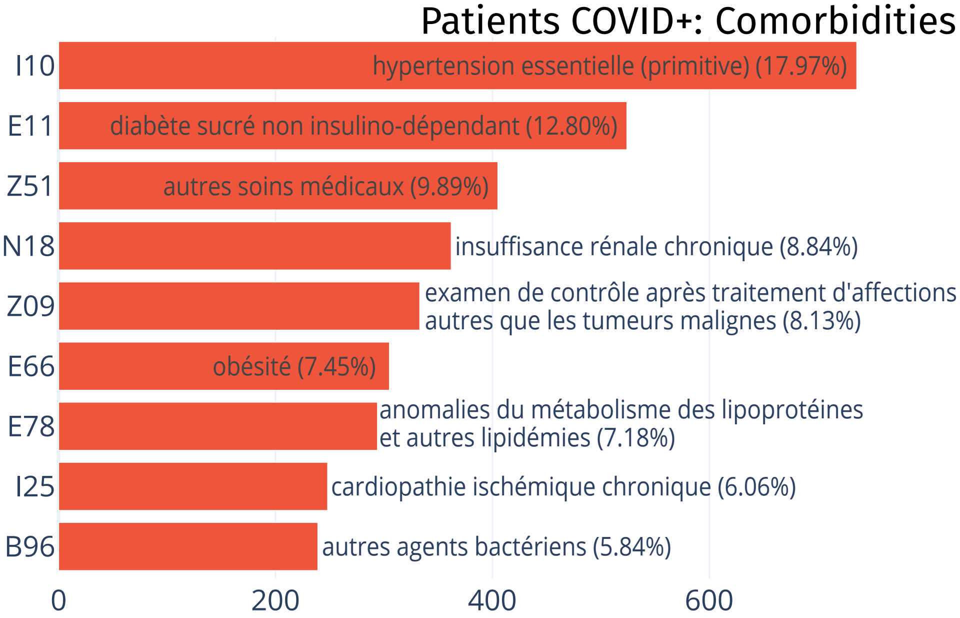


Changing reality

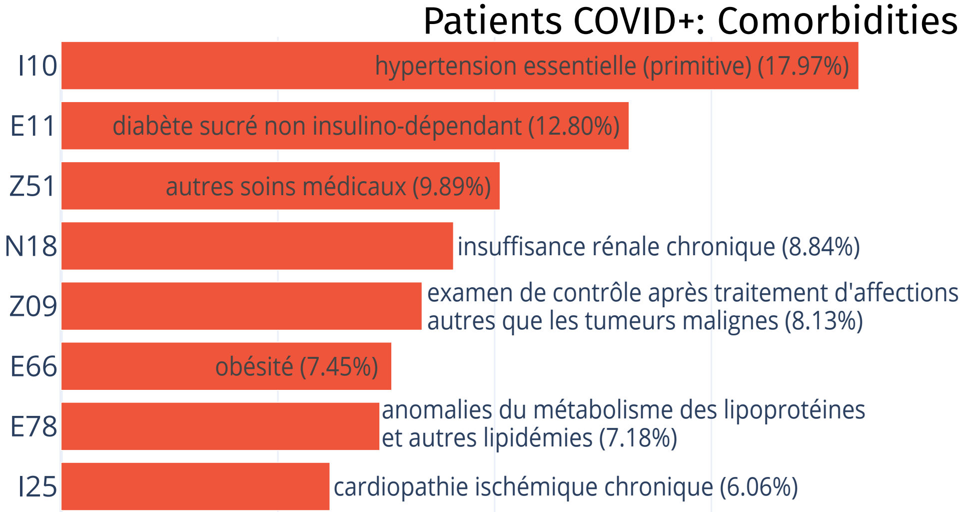
Covid outbreak: Hospital management



Covid outbreak: diagnostics



Covid outbreak: diagnostics



Machine learning to predict intensive care?

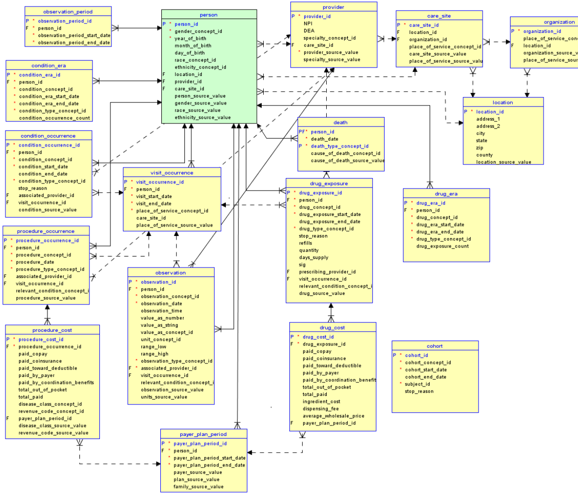
Useful for piloting, but not medical decisions

we only captured doctors' decisions, optimal or not

Pronostic modeling: A study cohort

Extracted from AP-HP's records

- 200 000 patients
- Claims: medical acts
- Biological values

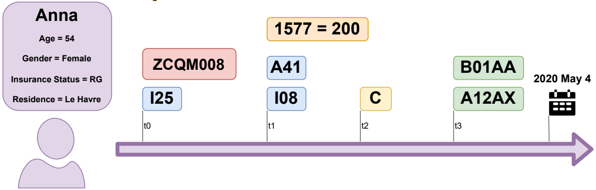


Predict future pathology?

- Hospital re-admition
- Predict type diagnostic?

Best machine-learning approach?
- AI = deep learning
- Epidemiology = Linear model

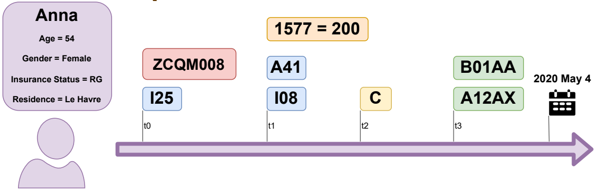
Modeling patient records: many modeling choices



1. Time-wise aggregation
   Build covariates from patient history
   - Demographics only
   - Decayed counting
   - Embeddings locally-optimized
   - National embeddings (SNDS)

Challenges
- Many different codes
- Time dimension

Modeling patient records: many modeling choices



1. Time-wise aggregation

Build covariates from patient history

- Demographics only
- Decayed counting
- Embeddings locally-optimized
- National embeddings (SNDS)

Challenges

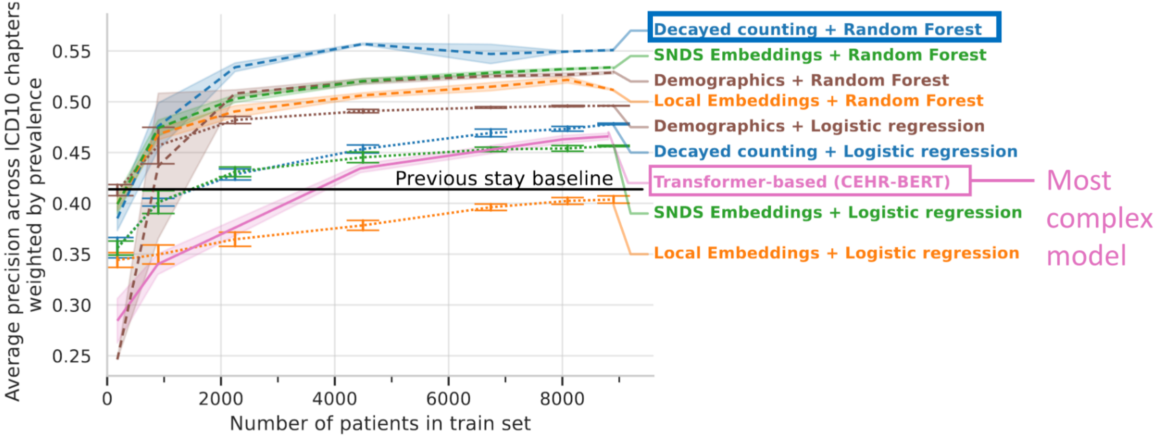
- Many different codes
- Time dimension

2. Supervised learning

- Linear model (logistic regression)
- Random forest
- Sequence model (transformer)

Benchmark a gradient of models, from simple to complex

Different models: best is not most complex



- Logistic regression = epidemiology
- Transformer = AI

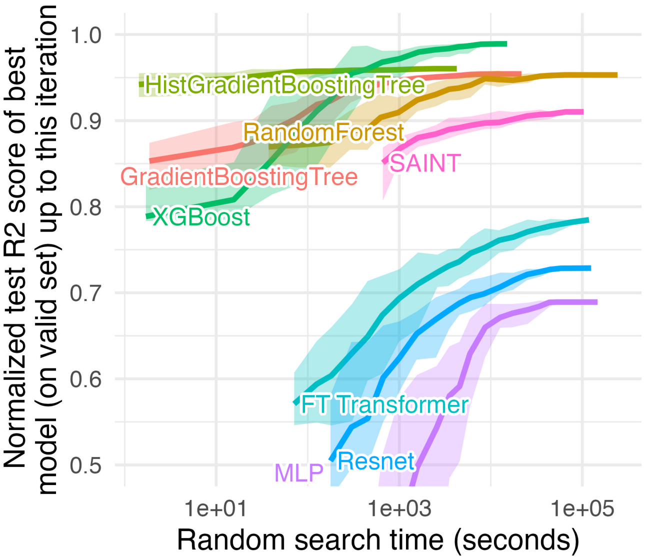
Best model = random forest

- Model from machine learning

M. Doutreligne

Why tree models > deep learning on tabular data [Grinsztajn... 2022]

Tree-based methods out-perform tailored deep architectures



Why tree models > deep learning on tabular data [Grinsztajn... 2022]

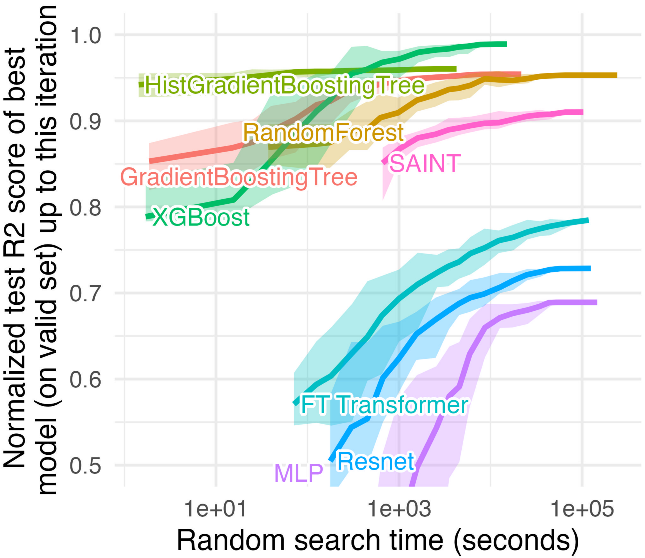
Tree-based methods out-perform tailored deep architectures

Tabular data

- Non-Gaussian marginals
- Categorical features

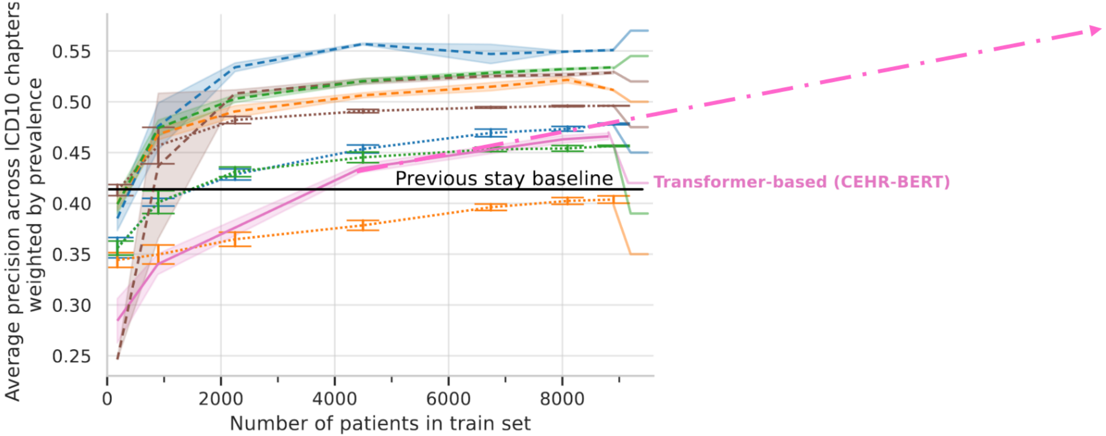
Trees' inductive bias:

- Axis-aligned
  Each column is meaningful
- Non smooth



The data's natural geometry is neither smooth nor vectorial

If we had more data

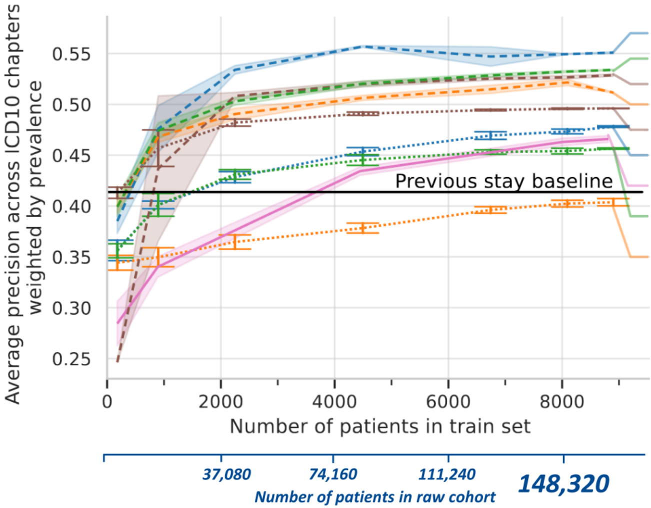


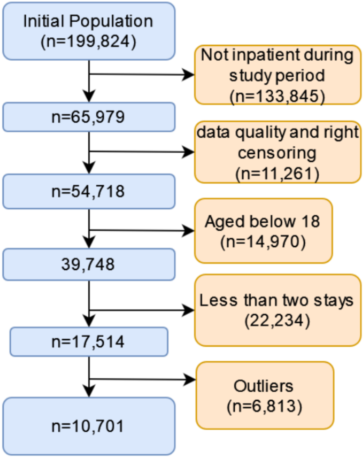
Classic machine learning trade offs:

Complex models need more data

M. Doutreligne

Why is health data small?





- Most lack data: out-patient, a single visit
- Pathologies have small prevalences

M. Doutreligne

More information: clinical notes

Clinical notes contain a huge amount of information on patients

They embed the context and the clinician’s understanding

Clinical notes are messy

```
ATCD  : BPCO post tabac, HTA , dysli
pidemie, polype colique( benin?) ttt chir
laparo med1ane, ULCERE ESTOMAC
v TRAlTEMENTS EN-COURS
Corgard
IEC. etc..
v HISTOIR3 DE LA MALADIE
Appar ition progressive de cephalée avec
cervlcalgie, sensations vert          et
poussée tensionnelle. sueurs
v EXAMEN CLINIQUE INlTIAL
 HT A sym des 2 c0té.
non calmee par loxen 20.
BDC réguliers;pas de souffle;
pas de s d'IVG ni d'IVD.
MV bilat et sym. pas de bruits surajoutés;
```

Tumeur traitée chirurgiquemen

Hyper Tension Arterielle

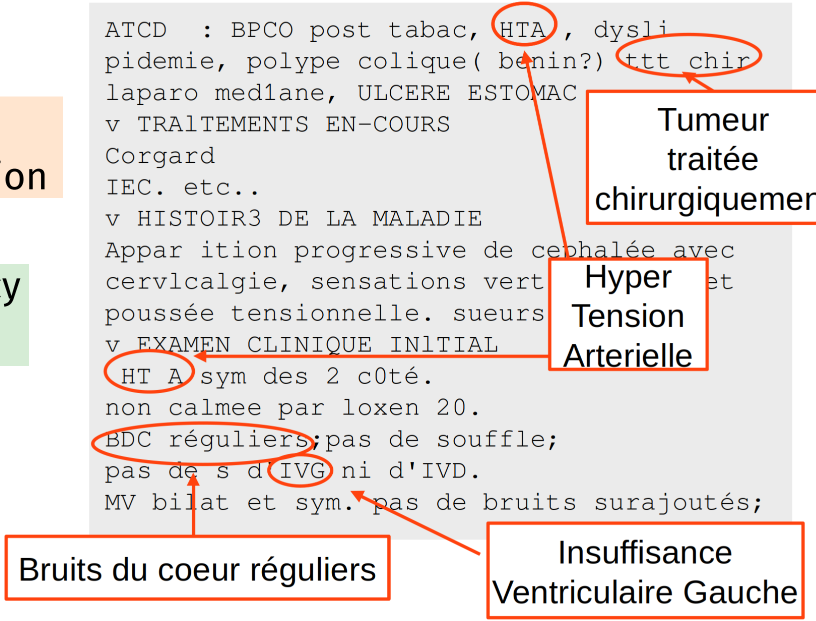
Bruits du coeur réguliers

Insuffisance Ventriculaire Gauche

Clinical notes are messy

Deep learning for information extraction

Improves accuracy from .7 to .75



Health data

- Different type of data, different type of models
  - Medical imaging: challenges of external validity
  - Text: pretrained language models and QA
  - Health records: data preparation ☺
- Always in a data-limited regime

Different goals

- Diagnostic or information extraction

  Nowcasting to help care giver
- Prognostic or future prediction
  - Help individual decision
  - Help resource management (piloting)

Predictors often fail to bring medical benefits

[Roberts... 2021] out of 62 publications on machine-learning for Covid detection on chest X-ray: none with potential for clinical use

Data often reflect an application only partly

- Information consequence of diagnostic
  - chest drain on pneumothorax X-rays [Oakden-Rayner... 2020]
  - dermatologist circling skin lesions [Winkler... 2019]

- Sampling bias (non representative of target population)

External versus internal validity

Focus on "good" prediction scores
pulls us to "beautiful" data

[Varoquaux and Cheplygina 2022]

2 Bridging the data to the application

Data may not reflect application [Varoquaux and Cheplygina 2022]

Prediction useless

- Because it builds on consequences of diagnostic
  - chest drain on pneumothorax X-rays [Oakden-Rayner... 2020]
  - dermatologist circling skin lesions [Winkler... 2019]

- Because of sampling bias

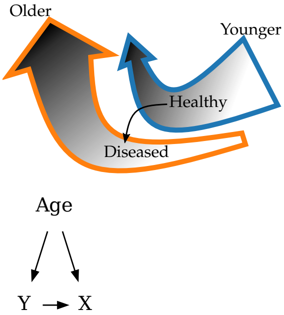
(data non representative of target population)

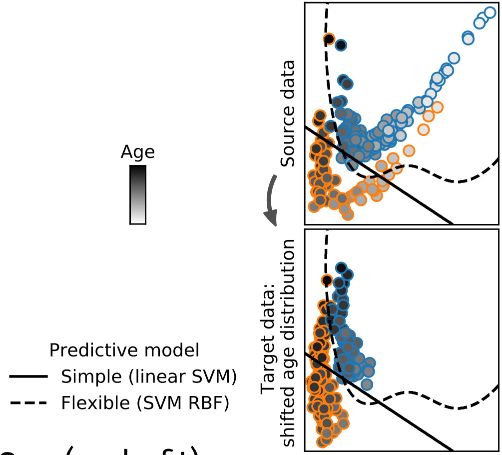
External versus internal validity

Focus on "good" prediction scores pulls us to "beautiful" data

Benin selection bias: “covariate shift” [Dockès... 2021]

The covariate X change, but the link $X \rightarrow y$ is preserved

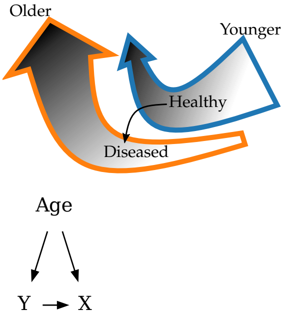


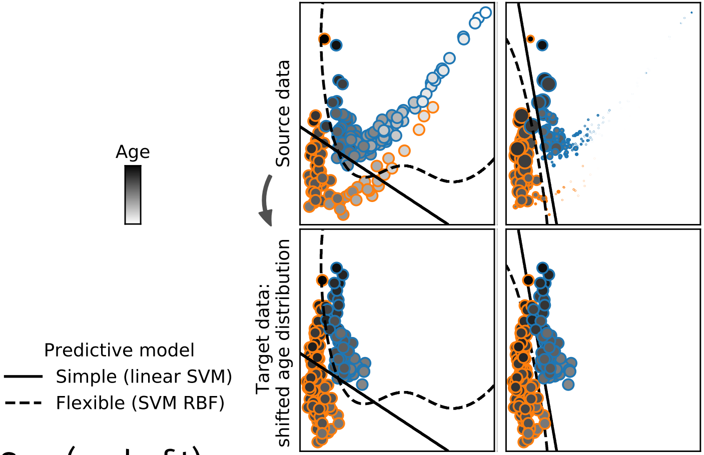


- A “simple” model fails (underfit)
- A flexible model succeeds, with enough data

Benin selection bias: "covariate shift" [Dockès... 2021]

The covariate X change, but the link $X \rightarrow y$ is preserved



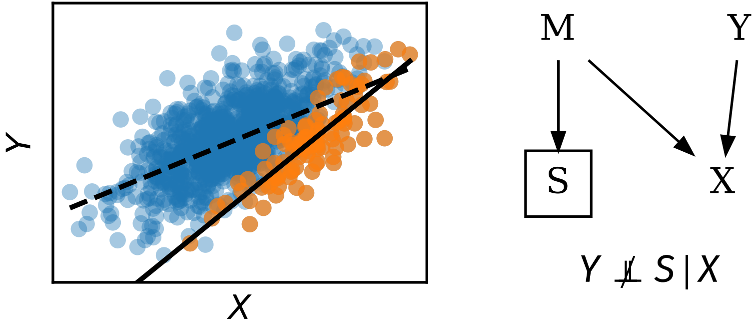


- A "simple" model fails (underfit)
- A flexible model succeeds, with enough data
- Reweighting helps for simple models or limited data

When selection bias breaks association

[Dockès... 2021]

An example: Selection based on M



$Y \not\perp S \,|\, X$

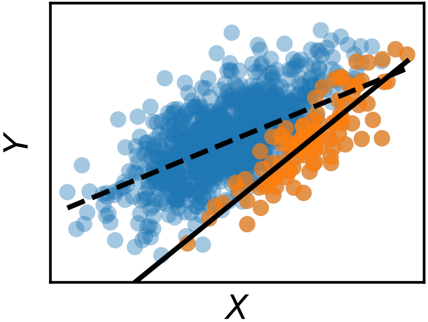
A common cause to selection S and the data (X, Y)

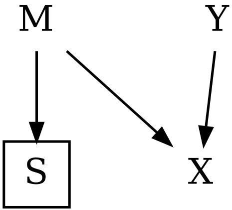
distorts the association between X and Y

When selection bias breaks association

[Dockès... 2021]

An example: Selection based on M





$Y \not\perp S \mid X$

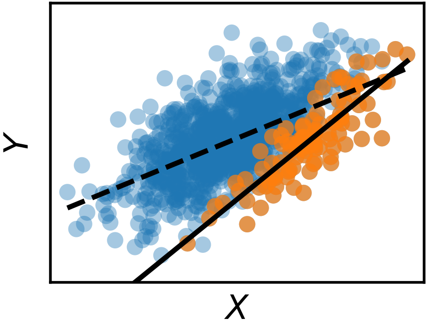
A common cause to selection S and the data (X, Y)

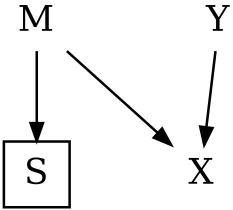
distorts the association between X and Y

More data, bigger models won't solve the problem

When selection bias breaks association

An example: Selection based on M





$Y \not\perp S \mid X$

A common cause to selection S and the data (X, Y) distorts the association between X and Y

More data, bigger models won't solve the problem

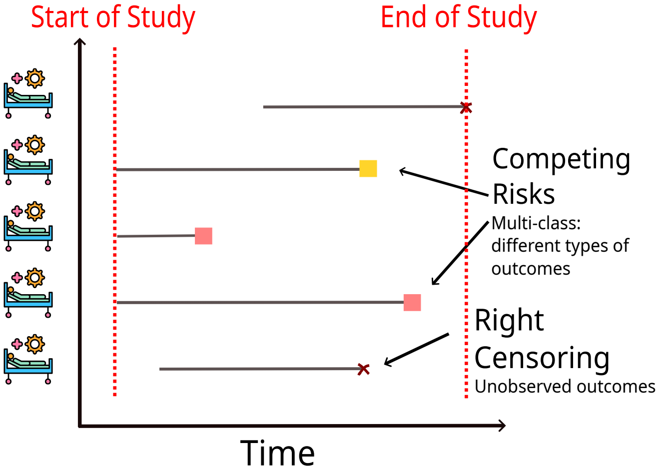
Next, I'll expand a couple common cases

Censored data

Outcomes not yet observed
Survival analysis

Survival analysis

Individuals not observed long enough to know their outcomes



Naive approach biased: *eg* even for a long-lasting disease, in a week-old outbreak the mean illness duration < 1 week

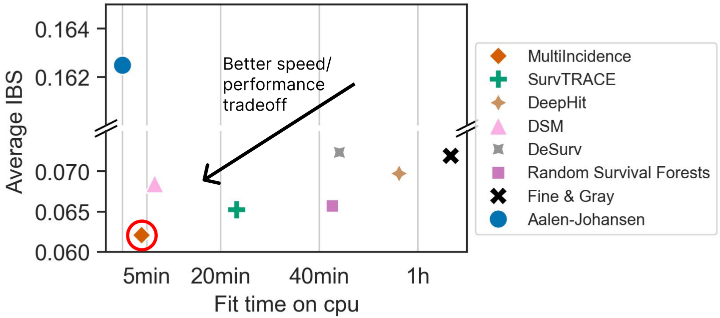
A marked case of selection bias

[Alberge... 2024]

Survival analysis: compensation terms in the loss [Alberge... 2024]

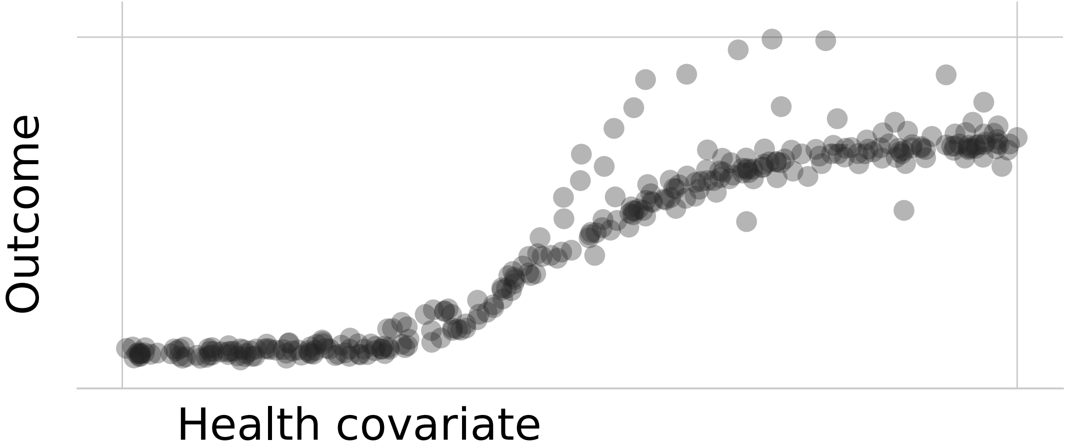
- Compute probability of censoring (increases with time)
- Weight samples by inverse probability
- Recovers true outcome probabilities
- Can be used with stochastic solvers

Faster, better, than more complex schemes



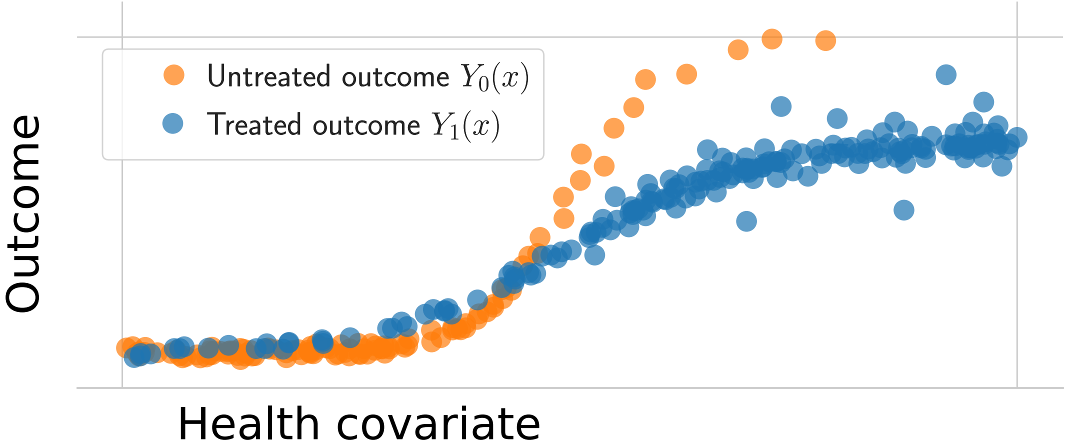
Prediction
to support decision

Prediction for decision making: causal effect



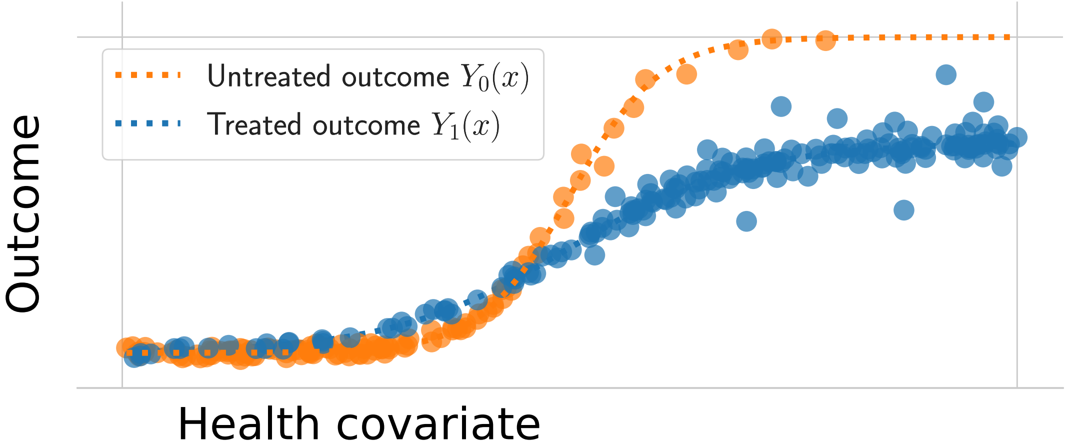
- Can a predictive model orient intervention choices?

Prediction for decision making: causal effect



- Can a predictive model orient intervention choices?
- We need the outcome as function of an intervention of interest

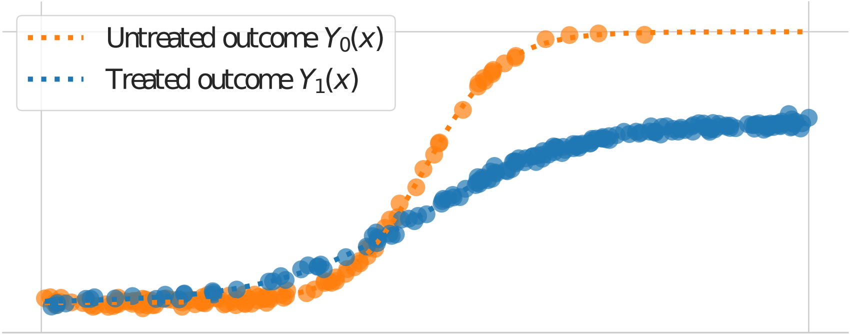
Prediction for decision making: causal effect



- Can a predictive model orient intervention choices?
- We need the outcome as function of an intervention of interest
- The proper quantity is the Individual treatment effect:
  comparing predicted outcomes for the same individuals

Causal inference and selection bias

[Doutreligne and Varoquaux 2023]

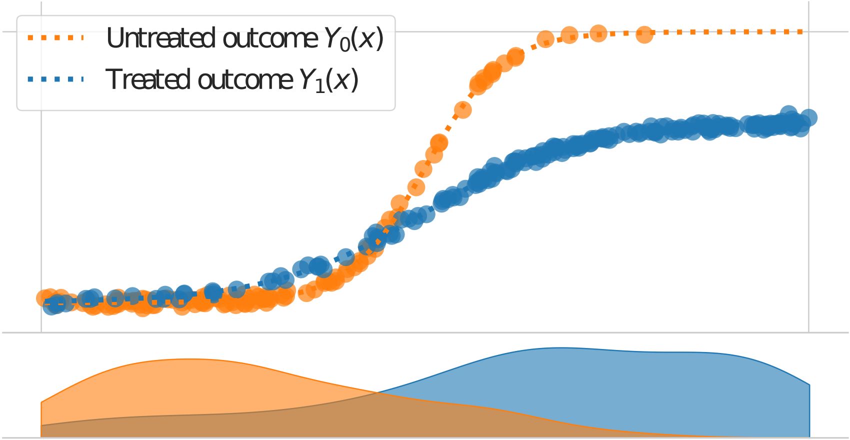


- Only one potential outcome observed per individual

  Machine learning to extrapolate across individuals

Causal inference and selection bias

[Doutreligne and Varoquaux 2023]

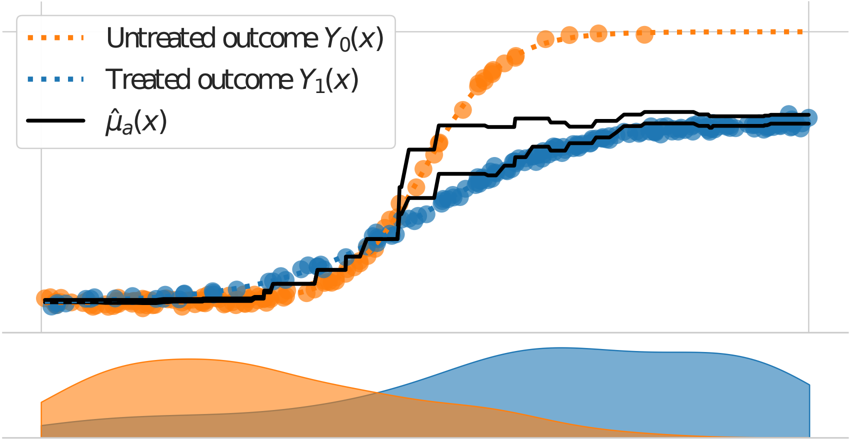


- Healthy individuals did not receive the treatment

  (selection bias compared to balanced intervention distribution)

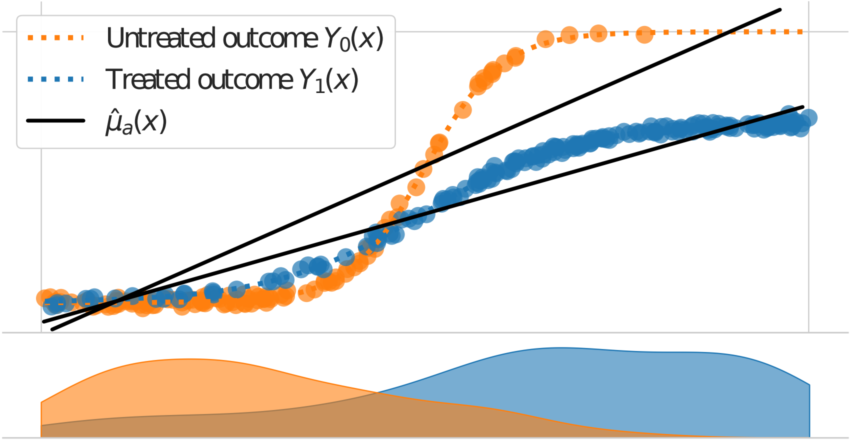
Causal inference and selection bias

[Doutreligne and Varoquaux 2023]



- Healthy individuals did not receive the treatment
  (selection bias compared to balanced intervention distribution)
- Good risk-minimizer associates treatment to negative outcomes

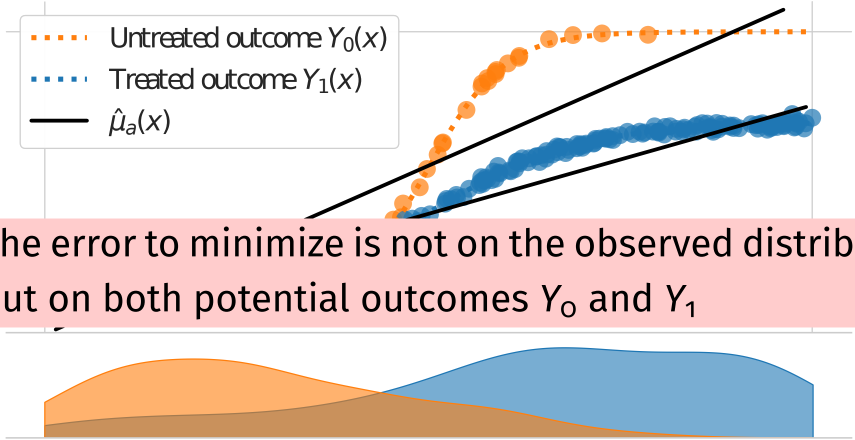
Causal inference and selection bias [Doutreligne and Varoquaux 2023]



- Healthy individuals did not receive the treatment
  (selection bias compared to balanced intervention distribution)
- Good risk-minimizer associates treatment to negative outcomes
- A worse predictor gives better causal inference

Causal inference and selection bias

[Doutreligne and Varoquaux 2023]



The error to minimize is not on the observed distribution but on both potential outcomes Y_0 and Y_1

- Healthy individuals did not receive the treatment
  (selection bias compared to balanced intervention distribution)
- Good risk-minimizer associates treatment to negative outcomes
- A worse predictor gives better causal inference

Inputs of predictors for decision making [Doutreligne... 2023]

- Using post-intervention information gives inapplicable prediction

  eg a drug lowers the blood pressure

  prediction with post-intervention measures of blood pressure

Inputs of predictors for decision making

[Doutreligne... 2023]

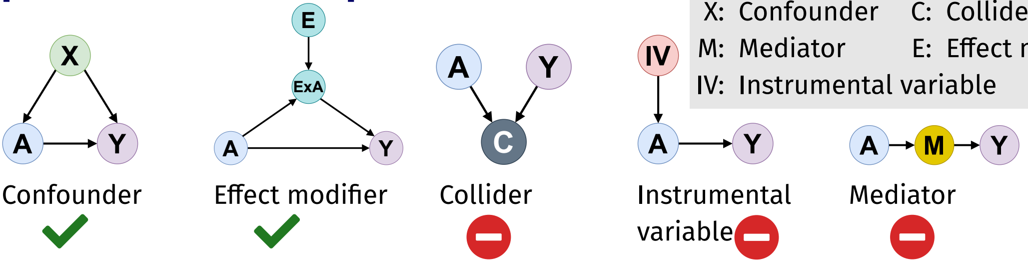
- Using post-intervention information gives inapplicable prediction

  eg a drug lowers the blood pressure

  prediction with post-intervention measures of blood pressure

- Causal criteria for variable inclusion:
  [Pearl and Mackenzie 2018]

A: Intervention Y: Outcome
X: Confounder C: Collider
M: Mediator E: Effect modifier
IV: Instrumental variable



Many caveats with temporal data (*eg* health records), see [Doutreligne... 2023]

Prediction to support decision

- Contrast predictions of potential outcomes

  best causal inference ≠ best usual predictor [Doutreligne and Varoquaux 2023]

- Don't predict from consequences of intervention

Data may not reflect application

It's not a question of best predicting y given X

- Survival: individuals not observed long enough
- Causality: observed only one potential outcome per individual

More data, bigger learner won't fix the problem

Need dedicated compensations

The soda team: Machine learning for health and social sciences

Machine learning for statistics
Causal inference, biases, missing values

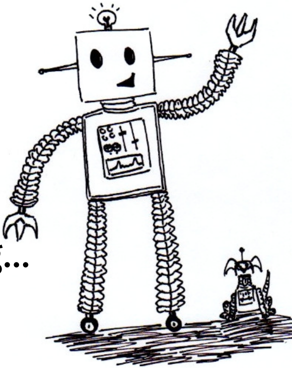
Health and social sciences
Epidemiology, education, psychology

Tabular relational learning
Relational databases, data lakes

Data-science software
scikit-learn, joblib, skrub

Better machine learning for health

Health records, routine care = close to practice



Bridge the data to the application

- The health outcome is the focus
- But we seldom observe it without bias, censoring...
  - Survival, for pronostic models
  - Causality, for decision models

The data results from prior choices, existing practice

Better evaluation

- Better metrics close to application
- Account for variance in benchmarks

Avoid the race to scale

@GaelVaroquaux

References I

J. Alberge, V. Maladière, O. Grisel, J. Abécassis, and G. Varoquaux. Teaching models to survive: Proper scoring rule and stochastic optimization with competing risks. *arXiv preprint arXiv:2406.14085*, 2024.

K. Dadi, G. Varoquaux, J. Houenou, D. Bzdok, B. Thirion, and D. Engemann. Population modeling with machine learning can enhance measures of mental health. *GigaScience*, 10(10):giab071, 2021.

J. Dockès, G. Varoquaux, and J.-B. Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 2021.

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

M. Doutreligne, T. Struja, J. Abecassis, C. Morgand, L. A. Celi, and G. Varoquaux. Causal thinking for decision making on electronic health records: why and how. *arXiv preprint arXiv:2308.01605*, 2023.

L. Grinsztajn, E. Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

References II

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

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, ... Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):1–8, 2022.

G. Varoquaux and O. Colliot. Evaluating machine learning models and their diagnostic value. `https://hal.archives-ouvertes.fr/hal-03682454/`, 2022.

References III

J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, ... Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.