

# AI from tabular data to healthcare and society

---

Gaël Varoquaux

*Inria*



This presentation

“AI” breakthroughs  
are on text and images

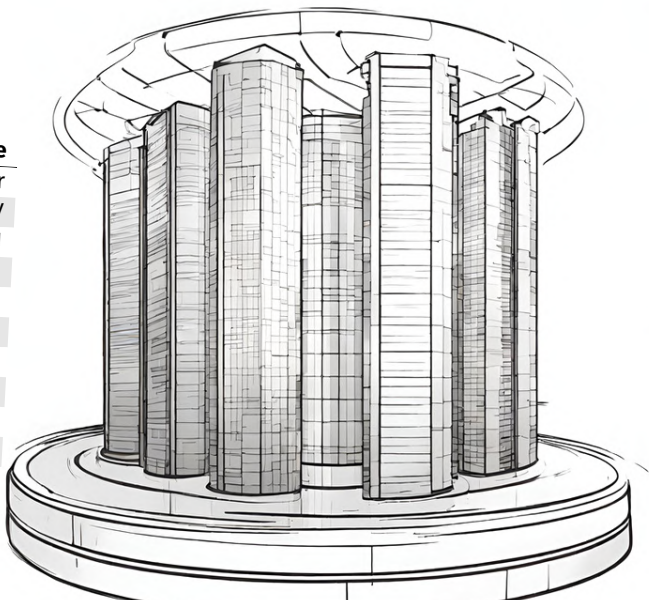
but the most precious data  
is in tables

and application-specific bias  
matter (eg in healthcare)



# 1 Tables: from data wrangling to AI

Gender	Experience	Age	Employee Position Title
M	10 yrs	42	Master Police Officer
F	23 yrs	NA	Social Worker IV
M	3 yrs	28	Police Officer III
F	16 yrs	45	Police Aide
M	13 yrs	48	Electrician I
M	6 yrs	36	Bus Operator
M	NA	62	Bus Operator
F	9 yrs	35	Social Worker III
F	NA	39	Library Assistant II
M	8 yrs	NA	Library Assistant I



# In data science most data is tabular

## Data preparation

Count, normalize, encode

Transform everything to numbers

It's the nature of statistics

We must feed the models

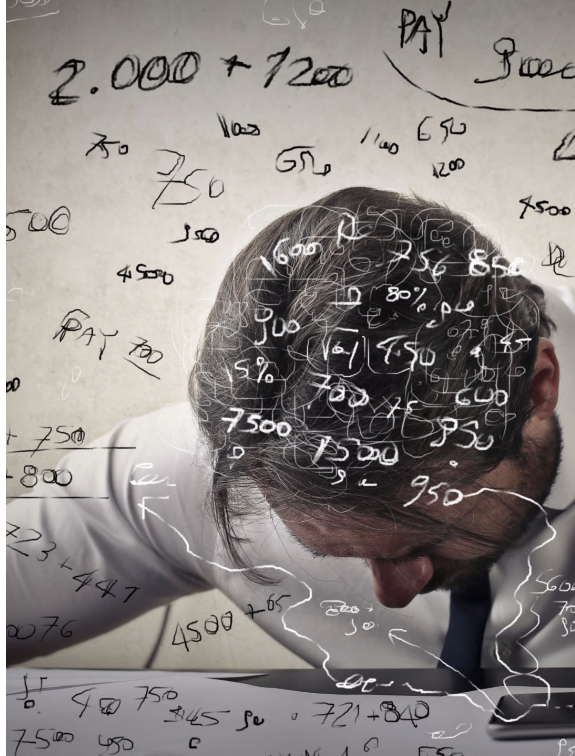


# Data preparation is exhausting

## My agenda

## Better learners to minimize preparation

# Will deep learning save us?



# Deep learning underperforms on data tables

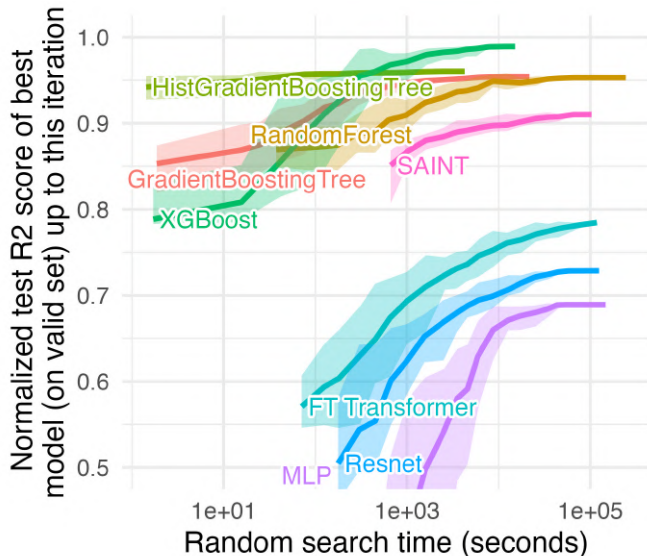
[Grinsztajn... 2022]

Tree-based methods  
out-perform tailored  
deep architectures



sklearn

HistGradientBoosting...



# Deep learning underperforms on data tables

[Grinsztajn... 2022]

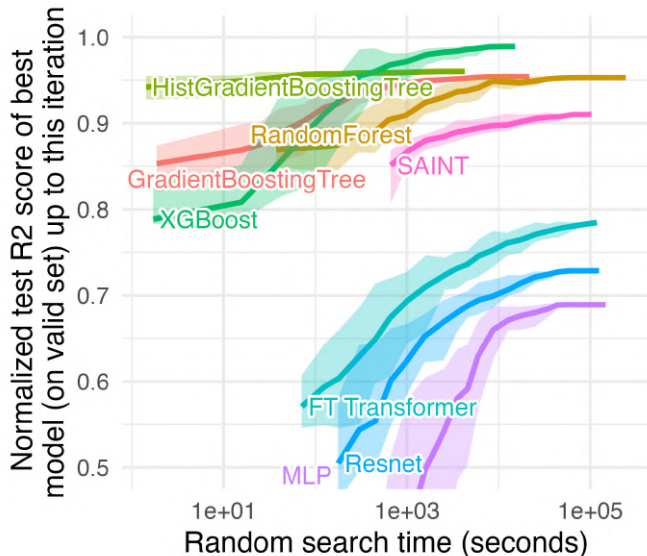
Tree-based methods  
out-perform tailored  
deep architectures

## 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

I'll come back to neural networks

First, some challenges of tables





# Missing Data

Frequent in  
health & social sciences

■  $\mathbb{R}^p \cup \{\text{NA}\}$  not a vector space



## Let us not fixate on imputation

[Le Morvan... 2021]

Impute = fill in the blanks with likely values

**Not statistically sound when missingness is informative**

*eg* fraudsters purposely not reporting information

# Let us not fixate on imputation

[Le Morvan... 2021]

Impute = fill in the blanks with likely values

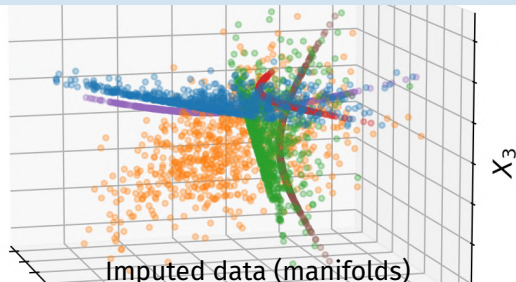
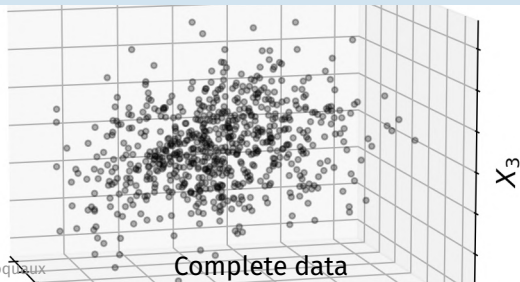
**Not statistically sound when missingness is informative**

eg fraudsters purposely not reporting information

**Imputing well is not needed to predict**

*Theorem (informal):*

a flexible learner gives asymptotically optimal prediction for all missing data mechanisms and almost all imputation.



# Let us not fixate on imputation

[Le Morvan... 2021]

Impute = fill in the blanks with likely values

**Not statistically sound when missingness is informative**

*eg* fraudsters purposely not reporting information

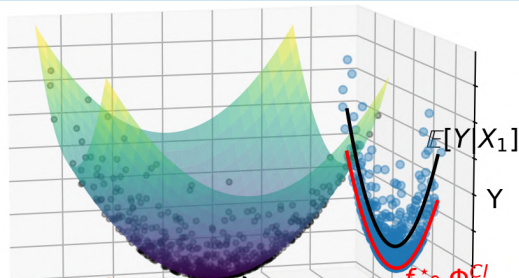
**Imputing well is not needed to predict**

*Theorem (informal):*

a flexible learner gives asymptotically optimal prediction for all missing data mechanisms and almost all imputation.

**Imputing with most likely value  
may lead to difficult prediction**

Ignores variance

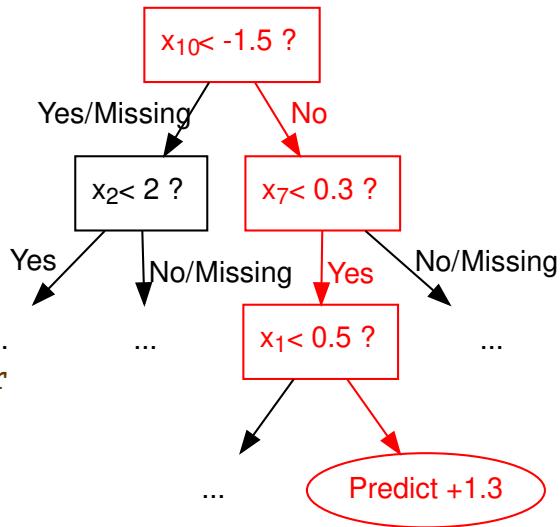


# Trees can handle missing values

[Josse... 2019]

The learner readily handles missing values

sklearn  
HistGradientBoostingClassifier



Works very well in benchmarks [Perez-Lebel... 2022]

# String entries

## Open-ended entries

- Not “categories”
- Not “entities”

### Employee Position Title

Master Police Officer

Social Worker IV

Police Officer III

Police Aide

Electrician I

Bus Operator

Bus Operator

Social Worker III



# Modeling strings, substrings

## Drug Name

---

alcohol

ethyl alcohol

isopropyl alcohol

polyvinyl alcohol

isopropyl alcohol swab

62% ethyl alcohol

alcohol 68%

alcohol denat

benzyl alcohol

dehydrated alcohol

---

## Employee Position Title

---

Police Aide

Master Police Officer

Mechanic Technician II

Police Officer III

Senior Architect

Senior Engineer Technician

Social Worker III

---

Polic...

3-gram<sub>1</sub> 3-gram<sub>2</sub> 3-gram<sub>3</sub>

# GapEncoder: String embeddings capturing latent categories

## Categories

Legislative Analyst II

Legislative Attorney

Equipment Operator I

Transit Coordinator

Bus Operator

Senior Architect

Senior Engineer Technician

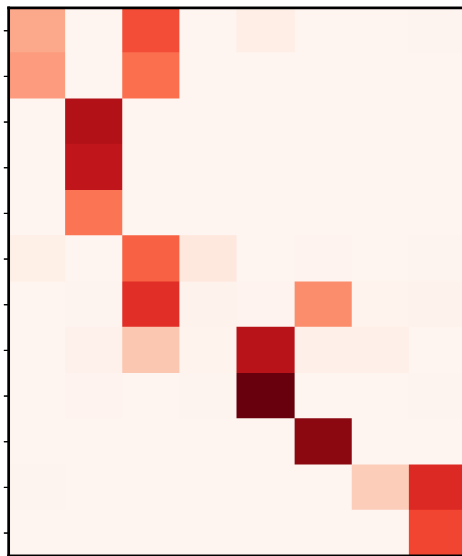
Financial Programs Manager

Capital Projects Manager

Mechanic Technician II

Master Police Officer

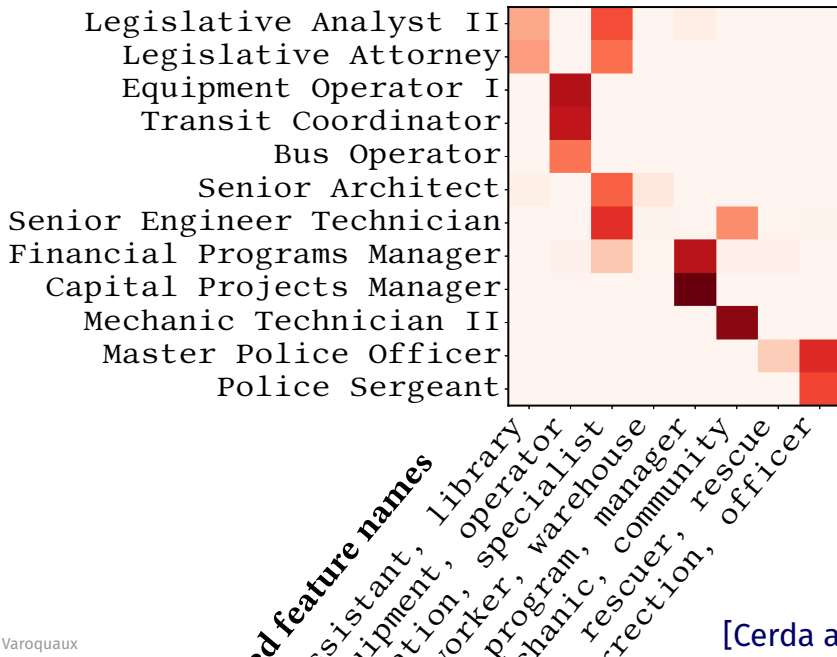
Police Sergeant





# GapEncoder: String embeddings capturing latent categories

## Plausible feature names



# Vectorizing tables: the TableVectorizer

Software: skrub

[skrub-data.org](https://skrub-data.org)

Prepping tables for machine learning



TableVectorizer

```
X = tab_vec.fit_transform(df)
```

Heuristics for different columns

■ strings with  $\geq 30$  categories  $\Rightarrow$  GapEncoder

■ date/time  $\Rightarrow$  DateTimeEncoder

■ non-string discrete  $\Rightarrow$  TargetEncoder

...

Very strong baseline

## Data tables

- Heterogeneous columns
  - Missing values
  - Open-ended strings

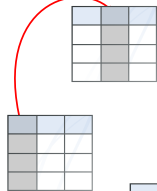
- Heterogeneous columns
  - Missing values
  - Open-ended strings



- Tree-based models
  - `sklearn HistGradientBoosting`
- Column encoding
  - `skrubg TableVectorizer`

# Learning across multiple tables

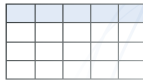
same entities



Aggregating



Analysis



# Example data-science analysis

## Real-estate market

Expected price of a property?

Predict the price from  
relevant information available

- age
- surface area
- # of rooms
- floor
- location
- ...



# Example data-science analysis

Data may need to be  
merged across tables

City	Pop.
Paris	2.2M
Vitry	33k

City	Rent	Population
Paris	1100€	2.2M
Vitry	700€	33k
Paris	1300€	2.2M



# Example data-science analysis

Aggregations may be needed across different data granularity

City	Pop.
Paris	2.2M
Vitry	33k

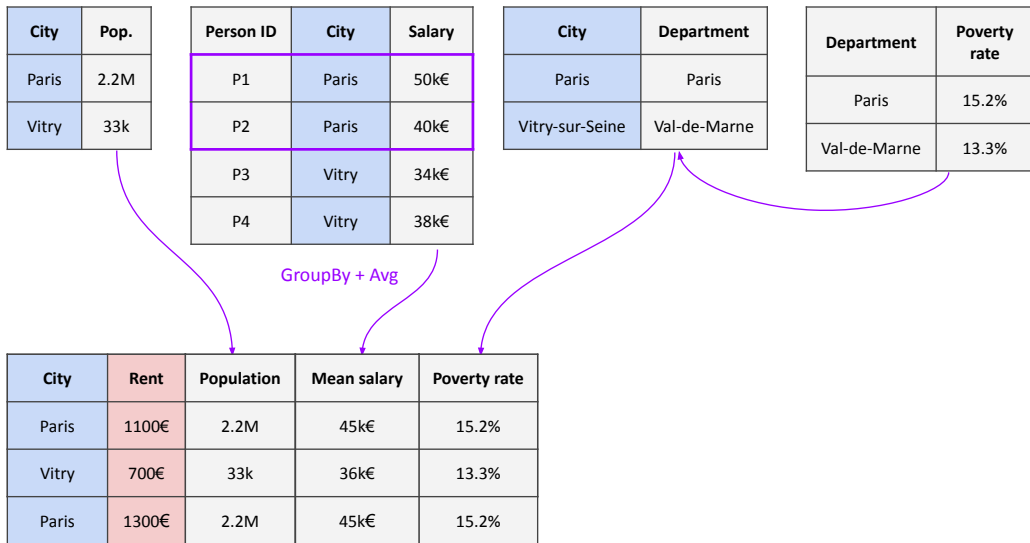
Person ID	City	Salary
P1	Paris	50k€
P2	Paris	40k€
P3	Vitry	34k€
P4	Vitry	38k€

GroupBy + Avg

City	Rent	Population	Mean salary
Paris	1100€	2.2M	45k€
Vitry	700€	33k	36k€
Paris	1300€	2.2M	45k€

# Example data-science analysis

## Multiple hops may be needed





# Example data-science analysis

## ■ Joining tables

City	Pop.
Paris	2.2M
Vitry	33k

Person ID	City	Salary
P1	Paris	50k€
P2	Paris	40k€
P3	Vitry	34k€
P4	Vitry	38k€

GroupBy + Avg

City	Rent	Population	Mean salary	Poverty rate
Paris	1100€	2.2M	45k€	15.2%
Vitry	700€	33k	36k€	13.3%
Paris	1300€	2.2M	45k€	15.2%

## ■ Aggregations

City	Department
Paris	Paris
Vitry-sur-Seine	Val-de-Marne

## ■ Multiple hops

Department	Poverty rate
Paris	15.2%
Val-de-Marne	13.3%

- Difficult for humans requires expertise on the data
- Difficult for machine learning discrete choices, combinatorial optim

# Example data-science analysis

## ■ Joining tables

City	Pop.
Paris	2.2M
Vitry	33k

Person ID	City	Salary
P1	Paris	50k€
P2	Paris	40k€
P3	Vitry	34k€
P4	Vitry	38k€

GroupBy + Avg

City	Rent	Population	Mean salary	Poverty rate
Paris	1100€	2.2M	45k€	15.2%
Vitry	700€	33k	36k€	13.3%
Paris	1300€	2.2M	45k€	15.2%

## ■ Aggregations

City	Department
Paris	Paris
Vitry-sur-Seine	Val-de-Marne

## ■ Multiple hops

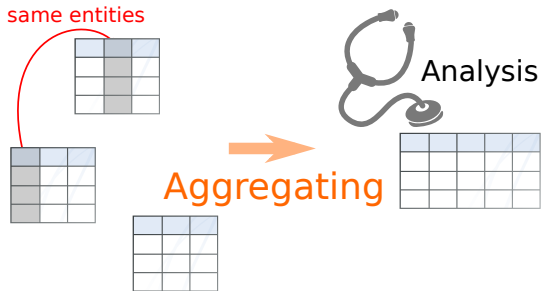
Department	Poverty rate
Paris	15.2%
Val-de-Marne	13.3%

- Difficult for humans requires expertise on the data
- Difficult for machine learning discrete choices, combinatorial optim

We need statistics and learning across tables



# Assembling data



- A “main” table
- Feature-enrichment tables



# Deep Feature Synthesis

[Kanter and Veeramachaneni 2015]

featuretools

- Greedily
  - starts from a **target table**
  - recursively joins related tables, to a given **depth**
- One-to-many relations: Computes different aggregations  
COUNT, SUM, LAST, MAX...

City	Population
Palaiseau	33k

Target table

Depth 0

City	School
Palaiseau	Lycée Camille Claudel
Palaiseau	Lycée Henri Poincaré

Depth 1

School	Students
Lycée Camille Claudel	800
Lycée Henri Poincaré	1000

Depth 2

City	Department
Palaiseau	Essonne

Department	PovertyRate
Essonne	13.3%

City	Population	COUNT( City.School)	City.Department	City.Department. PovertyRate	SUM(City. School.Students)	MAX(City. School.Students)
Palaiseau	33k	2	Essonne	13.3%	1800	800

Does not scale: # features explodes with depth and # tables

# Entity embeddings that distill information across tables

KEN: knowledge embedding with  
numbers [Cvetkov-Iliev... 2023]

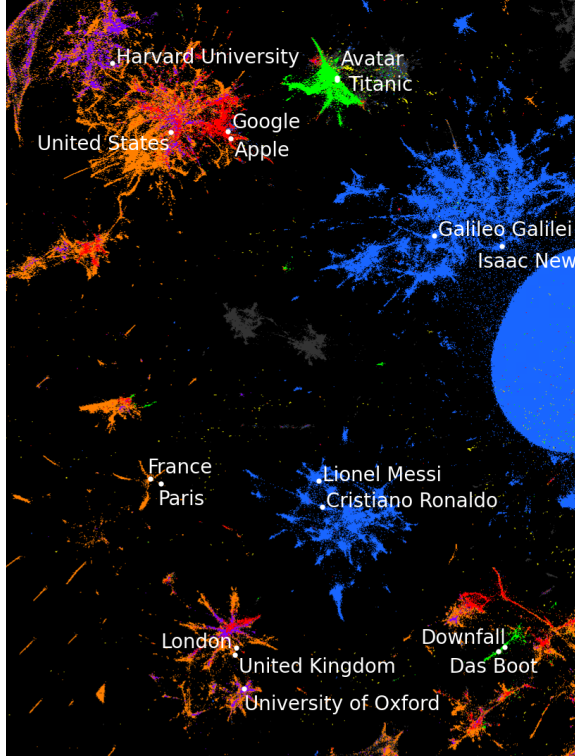
$$X \in \mathbb{R}^p$$

[soda-inria.github.io/ken\\_embeddings](https://soda-inria.github.io/ken_embeddings)

**6 million common entities**

cities, people, companies...

Example usage in skrub docs



# Needs Matching

## Morphological variant



`skrub.fuzzy_join` Hard

Should use context  
More than string similarity

Recontextualization for  
disambiguation

**State**

New York

...

**City**

New York

...



# Coming back to neural networks

## Neural networks successes in vision, text...

- Large data
- Pretrained





## Pretraining for data tables?

What prior for a bunch of numbers?

72	68	174	1
64	79	181	1
56	59	166	0
81	62	161	1

## Pretraining for data tables?

What prior for a bunch of numbers?

72	68	174	1
64	79	181	1
56	59	166	0
81	62	161	1

And now?

<i>Cardiovascular cohort</i>				
<b>Age</b>	<b>Weight</b>	<b>Height</b>	<b>Comorbidity</b>	<b>Cardiovascular event</b>
72	68	174	Diabetes	1
64	79	181	Cardiac arrhythmia	1
56	59	166	NA	0
81	62	161	Asthma	1

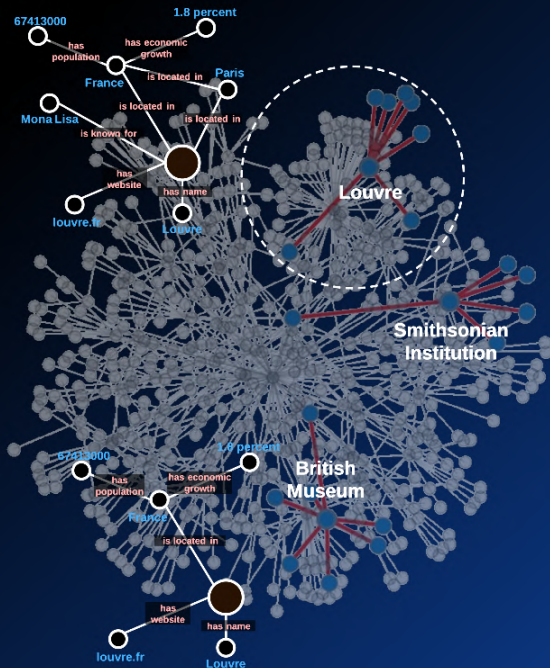
# Neural networks bring value if pretrained

**Blocker:** Integration across tables

- Entity matching  
String-level modeling
- Schema matching  
Model local relational graph

Pretrain on large data sources

Birth of tabular foundation model  
[Kim... 2024]



# CARTE: pre-trained tabular transformer

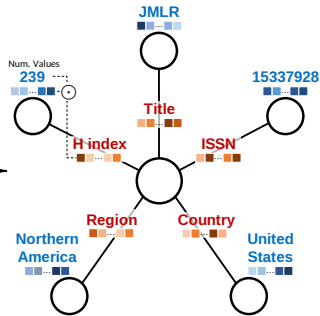
[Kim... 2024]

## Graph representation to bridge tables

- Value  
⇒ embeddings on nodes
- Column titles  
⇒ string embeddings on edges

Title	ISSN	Publisher	Country	Region	H index
				Western Europe	1331
				Northern America	239

language Model

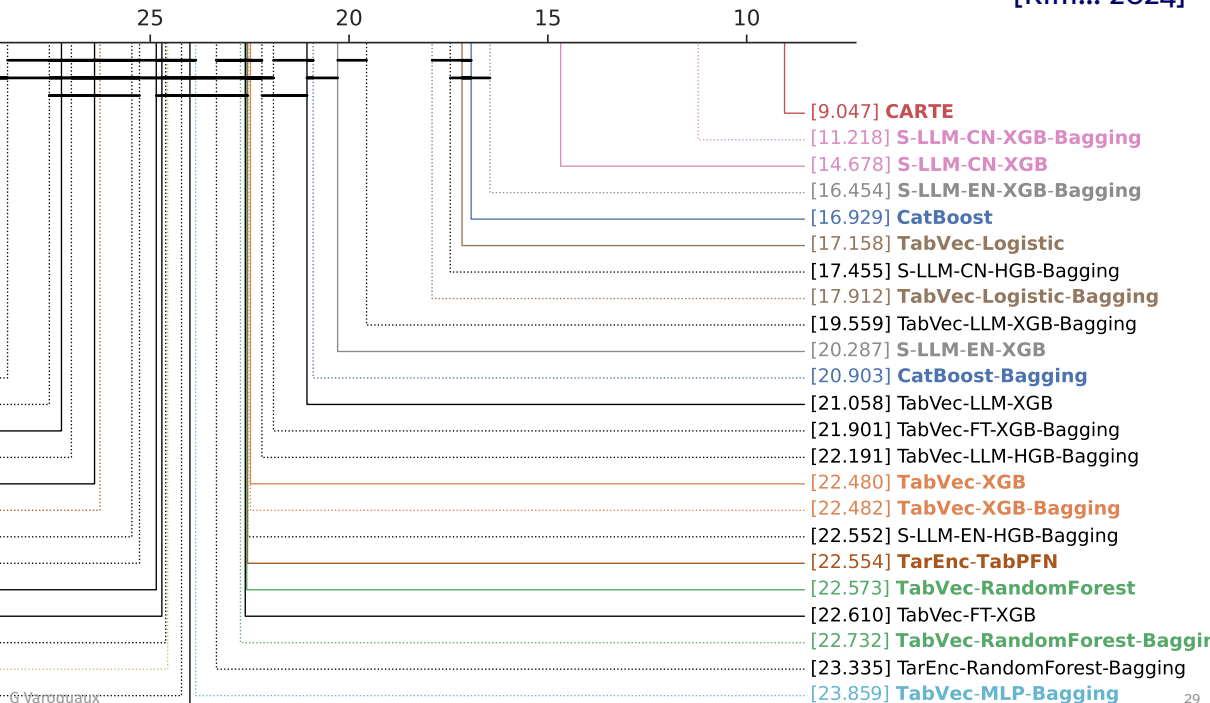


## Graph transformer architecture

- Pre-trained on large knowledge bases  
⇒ contrastive learning
- New attention accounting for relational information  
⇒ Adapted from knowledge-base embedding literature
- Brings context / enrichment to data  
⇒ particularly useful for string entries

# CARTE: facilitates small-sample learning

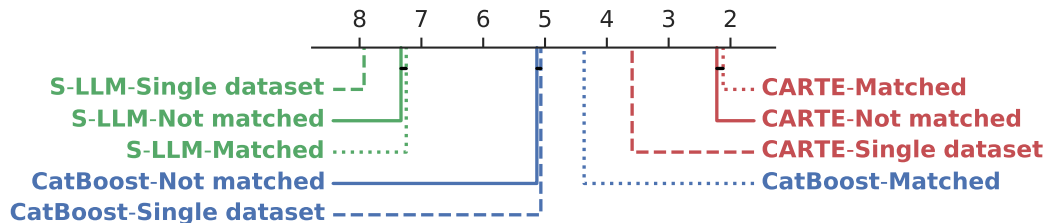
[Kim... 2024]



# CARTE: facilitates small-sample learning

## Across multiple tables

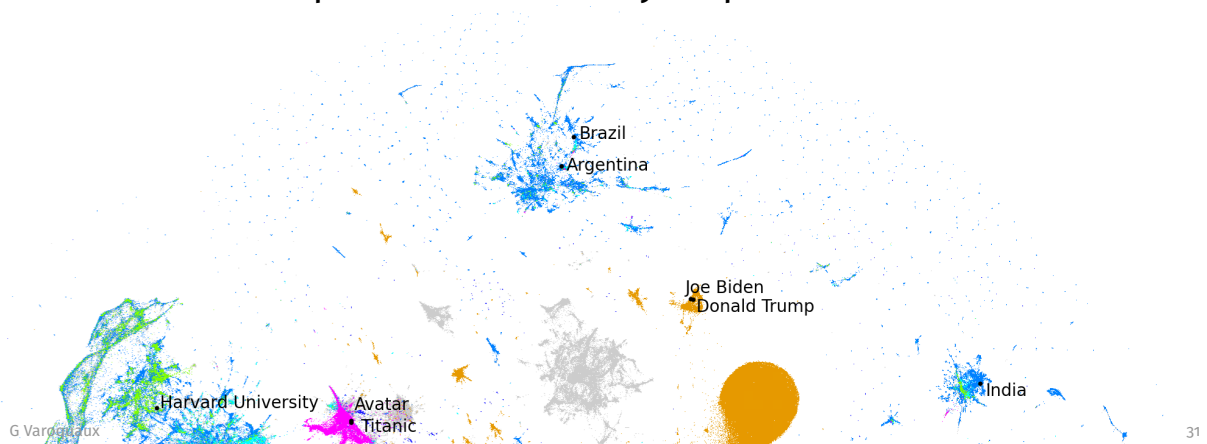
Fine tuning on multiple tables together



# Representation learning + rich machine learning

## Can partly automate data preparation

- Optimizing data transformation / representations for a task
- Continuous representations really help



# Skrub: software bridging databases to scikit-learn

## Prepping tables for machine learning

Prepping tables for machine learning

[skrub-data.org](https://skrub-data.org)



Database operations

- separating train/test time
- that can be optimized

Pipeline

- Joins & aggregations
- Mostly columnar

Optimized for learning:

- Supervising join discovery
- Vector representations, of strings, fuzzyness

API in flux



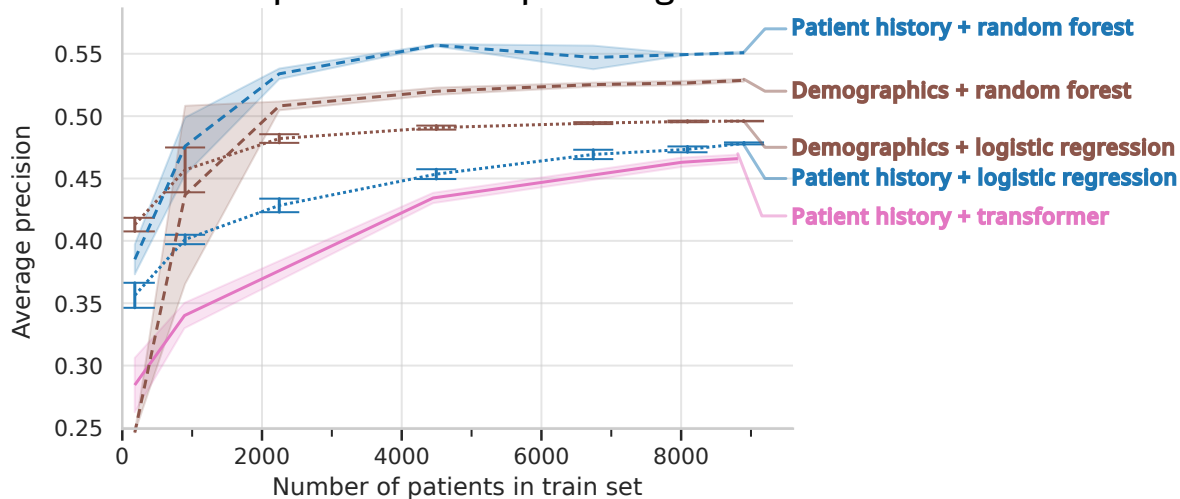
## 2 Application value

Some lessons from healthcare



# How to best use rich health database: 150 000 person history

## Prevention: predict future pathologies



**The fanciest ML model doesn't predict best**

# Needs more than big AI

- Fanciest doesn't always outperform
- Data may not reflect application



## 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

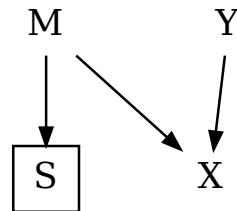
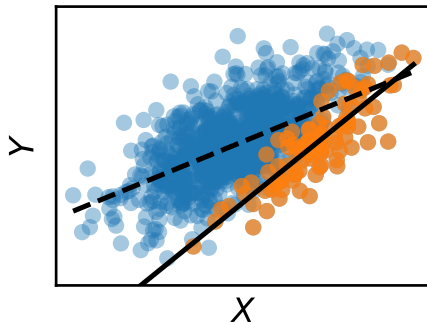
“Good” prediction scores,  
but not on useful outcomes



# When selection bias breaks association

[Dockès... 2021]

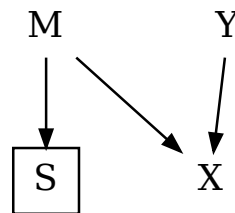
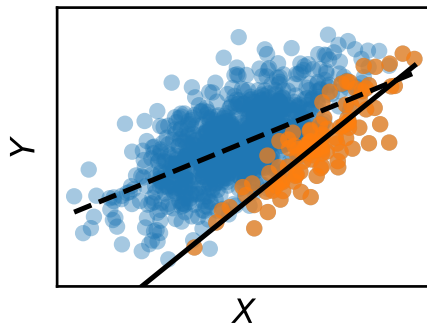
An example: Selection based on  $M$



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

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

An example: Selection based on  $M$

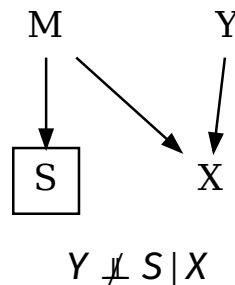
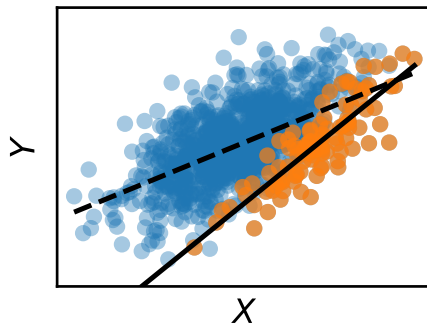


$$Y \not\perp S | X$$

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

**More data, bigger models won't solve the problem**

An example: Selection based on  $M$



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 common cases**

# **Censored data**

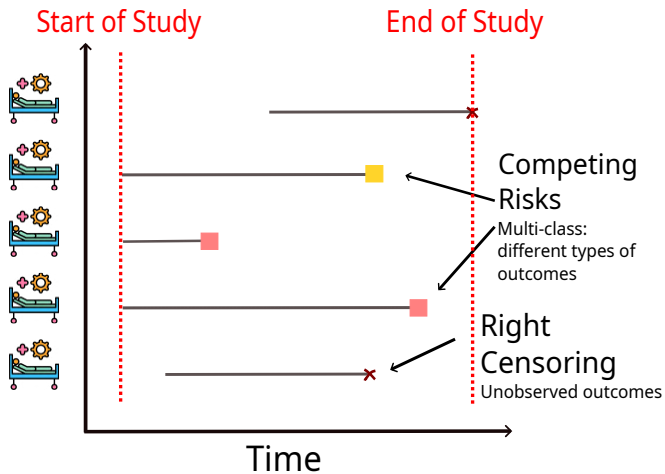
**Outcomes not yet observed  
Survival analysis**





# Survival analysis

Individuals not observed long enough to know their outcomes

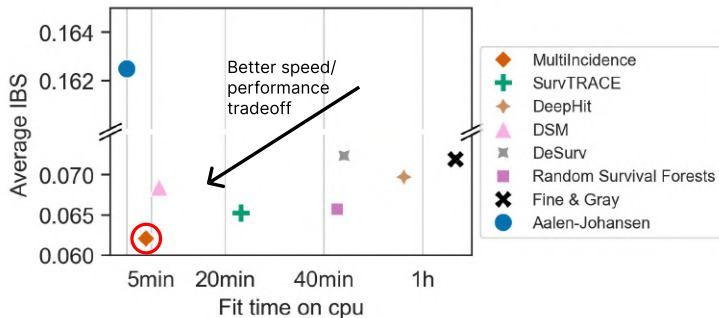


**Naïve approach biased:** *eg* even for a long-lasting disease, in a week-old outbreak the mean illness duration  $< 1$  week

A marked case of selection bias

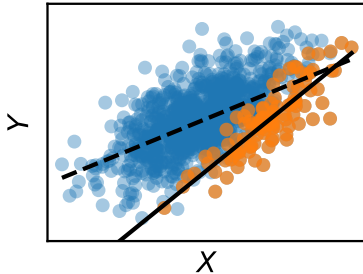
- Compute probability of censoring (increases with time)
- Weight samples by inverse probability
- Recovers true outcome probabilities
- Can be used with stochastic solvers

Faster, better, than  
more complex  
schemes



## Data may not reflect application

- Data result for a historical process
- Biases not solved by bigger data



Need (often tricky) corrections



### 3 A broader picture



# Big neural networks are the long-term solution

Towards foundation models for tables

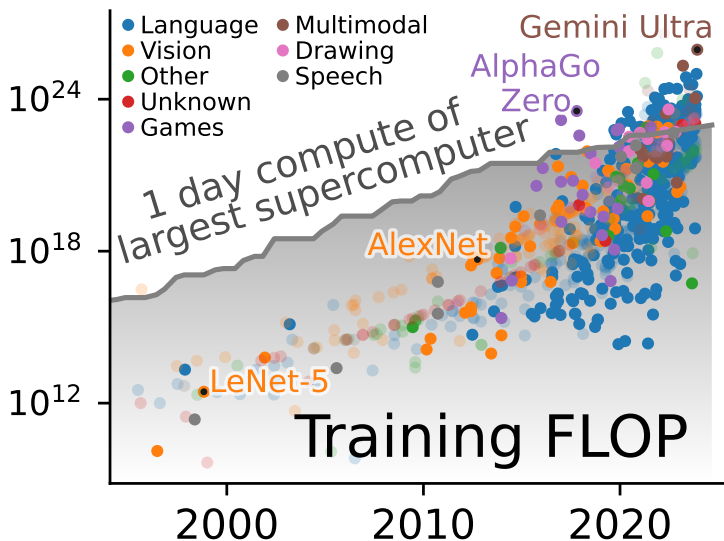
AlexNet and LLMs have shown **data + GPU solves everything**

*“our results can be improved simply by waiting for faster GPUs and bigger datasets”*  
[Krizhevsky... 2012]

*“leverag[ing] computation [is] ultimately the most effective [...] The ultimate reason for this is Moore’s law”*  
[Sutton 2019]

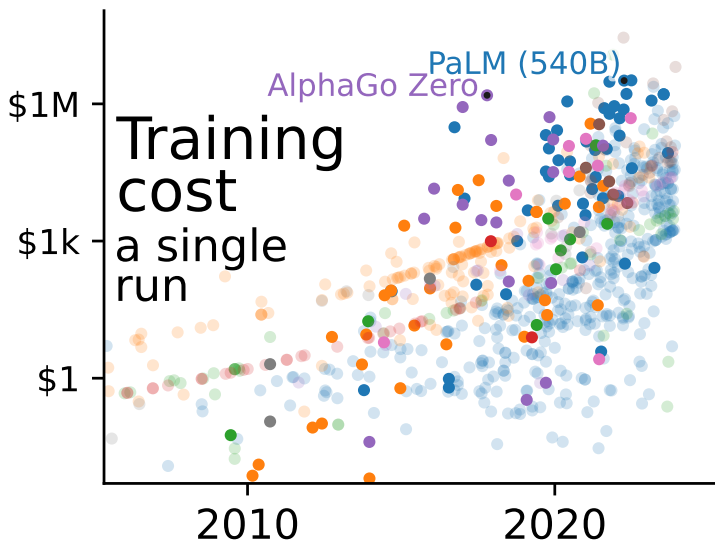
**We just keep going bigger, fancier**

# Big neural networks are the long-term solution – not



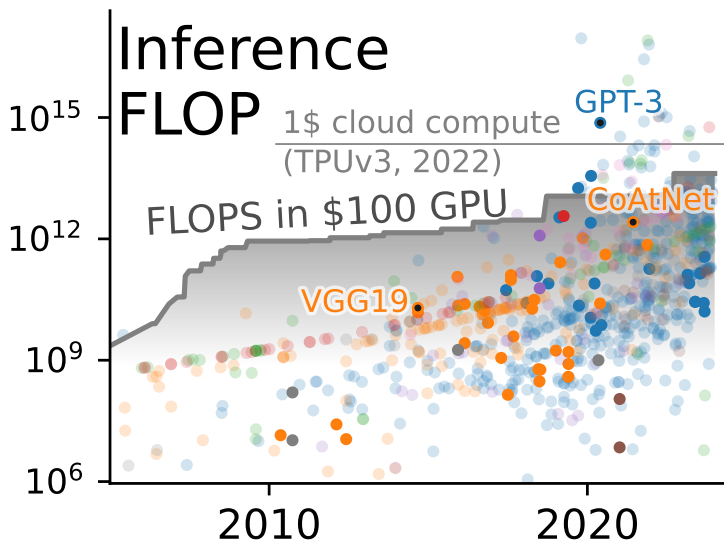
**AI compute overtook largest supercomputers**

# Big neural networks are the long-term solution – not



**Rebound effect:** demand increase beats efficiency gains

# Big neural networks are the long-term solution – not



**Unsustainable inference – Costs don't add up**



Are we the  
baddies?



# Tech's social norms

How do we choose what we work on?

Social norms of success  
**Bigger models, beating benchmarks**

(face recognition, GANs, LLMs)

Big tech

The value system of big players  
defines the cool



# Scale benefits some actors

Detailed individual data  
Social networks

Large compute  
GPU and cloud providers

**Big tech wins**



## Scale benefits some actors

High-quality scanners

Large hospitals

Latest phones

Rich, urban, and young

A widening gap between urban  
centers and forgotten rurals

Fuels unrest



# Choose what we facilitate

We can make our choices

🔧 the technology we create

👤 who we enable

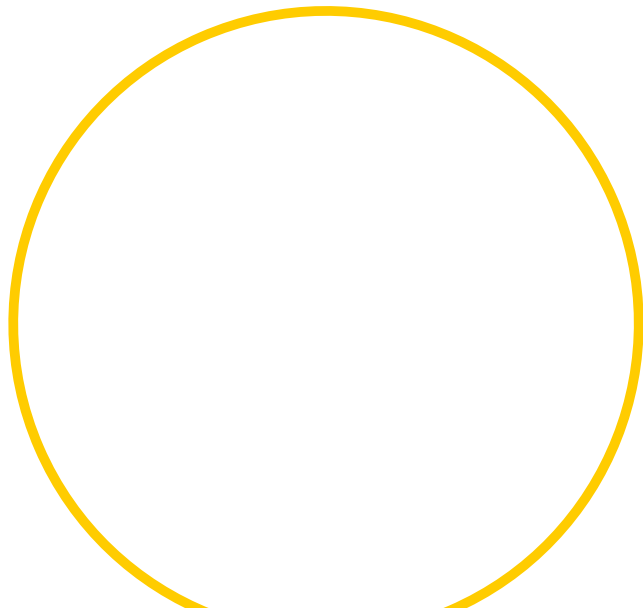
Shapes how the value of tech  
is distributed



# The advancement of knowledge

Courtesy of Matt Might

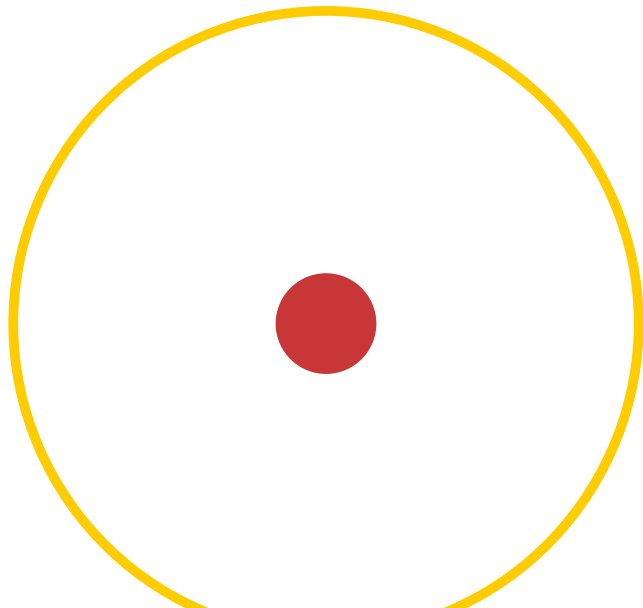
Imagine a circle that contains human knowledge



# The advancement of knowledge

Courtesy of Matt Might

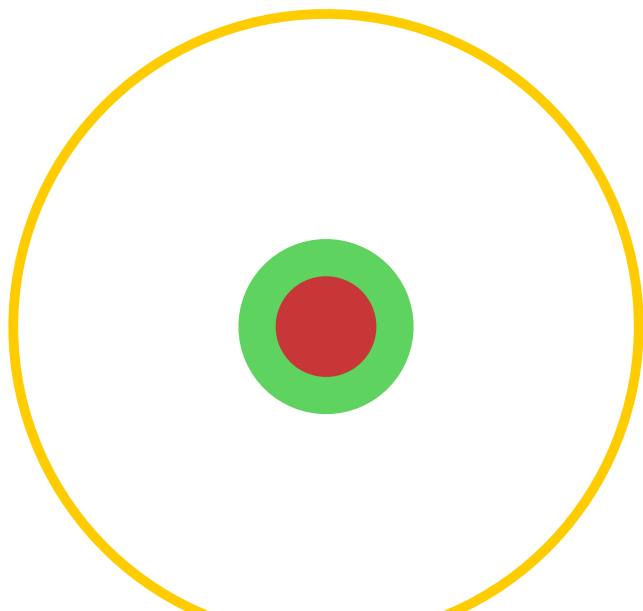
By the time you finish elementary school, you know a little



# The advancement of knowledge

Courtesy of Matt Might

High school takes you a little bit further

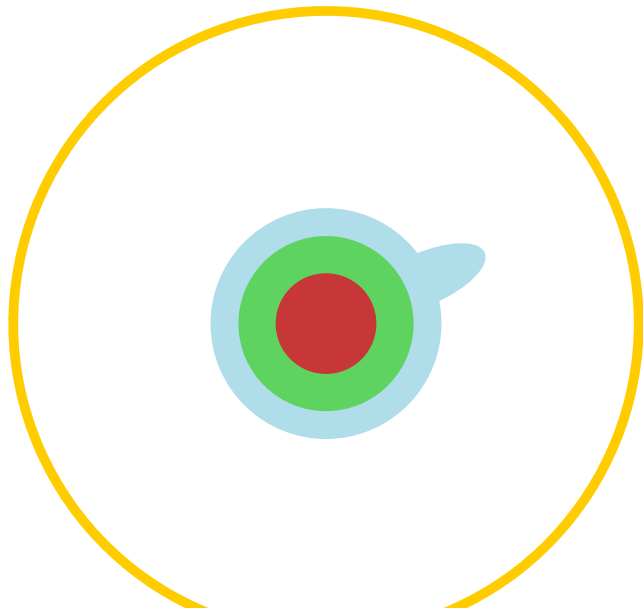




# The advancement of knowledge

Courtesy of Matt Might

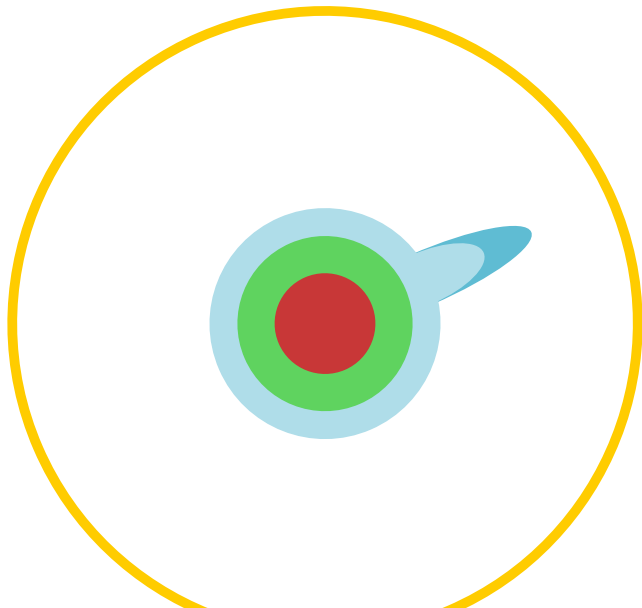
With a bachelors degree, you gain a speciality



# The advancement of knowledge

Courtesy of Matt Might

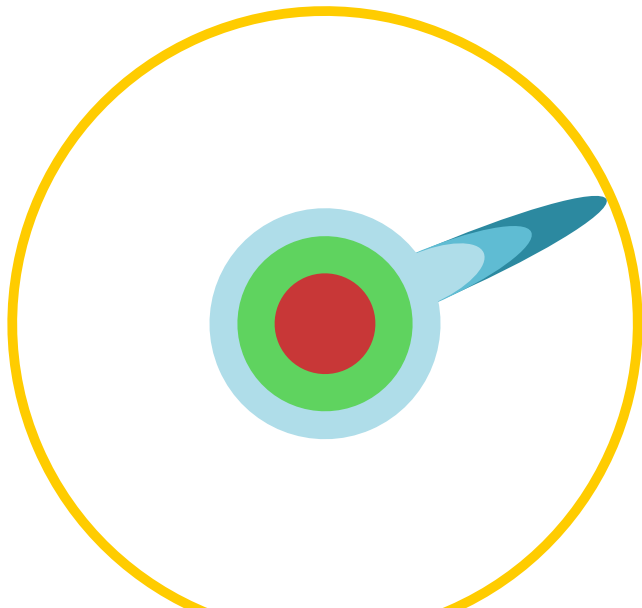
A master's degree deepens this speciality



# The advancement of knowledge

Courtesy of Matt Might

Research papers take you to the edge of human knowledge



# The advancement of knowledge

Courtesy of Matt Might

Once you are at the boundary, you focus, you push



# The advancement of knowledge

Courtesy of Matt Might

And one day it yields



# The advancement of knowledge

Courtesy of Matt Might

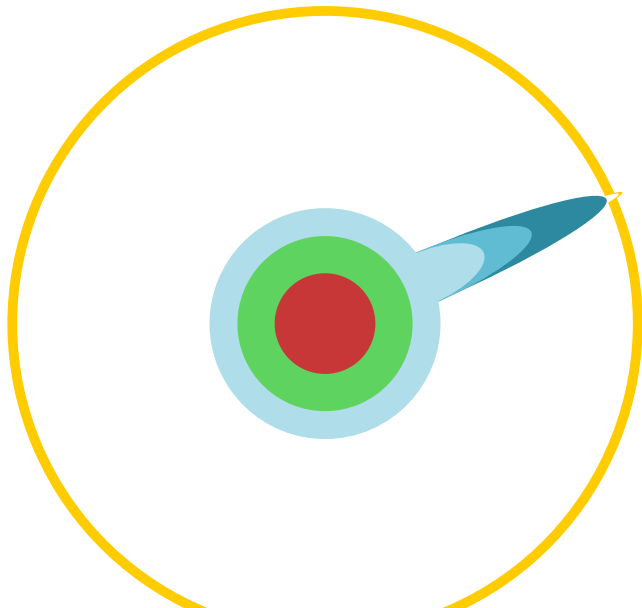
Of course, the world looks different to you now



# The advancement of knowledge

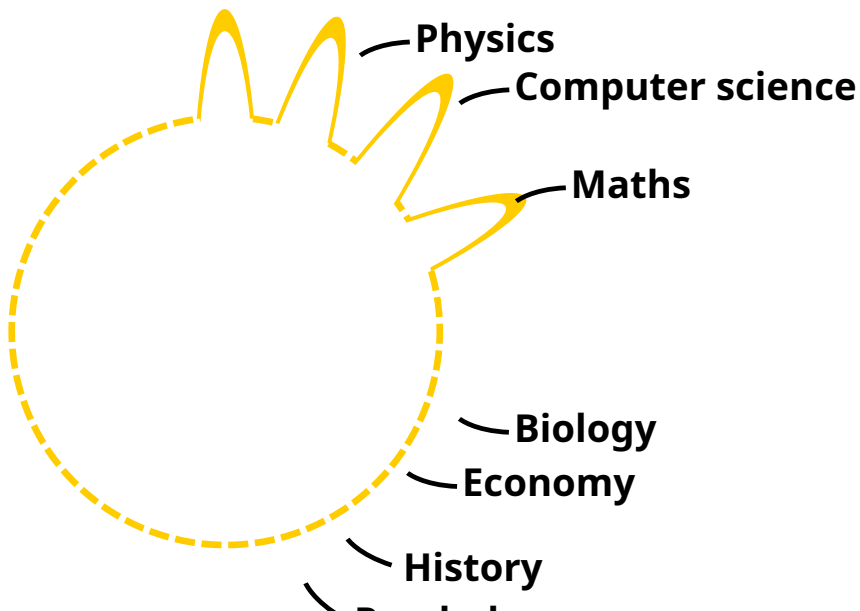
Courtesy of Matt Might

But don't forget the big picture



# The advancement of knowledge

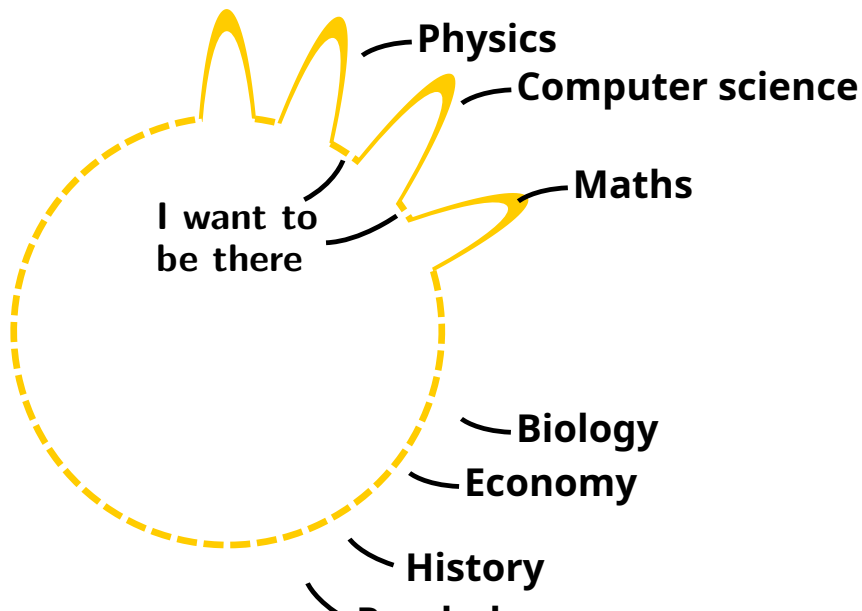
This is an optimistic view





# The advancement of knowledge

This is an optimistic view



■ Don't forget to go back to the big picture

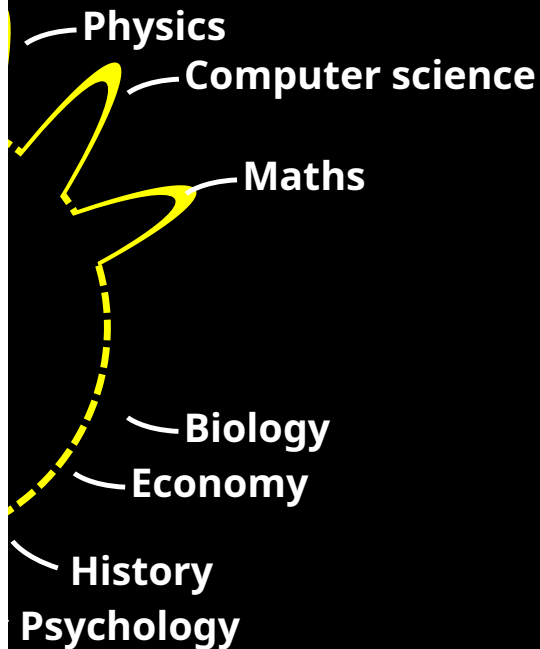
■ Research communication:

1. Why?
2. What?
3. How?

Experts (reviewers) overly focus on 3.

■ Success

- Short-term incentives = focus
- Long-term success = big picture



## Short term / long term

### Short term: Publish or perish

- Answer trendy questions
- Master math / coding
- Write well

### Long term

- Questions & skills to move forward
  - Work with the right people
- Empowering and kind

Learn your way through the system



# The soda team: Machine learning for health and social sciences

## Tabular relational learning

Relational databases, data lakes

## Health and social sciences

Epidemiology, education, psychology

## Machine learning for statistics

Causal inference, biases, missing values

## Data-science software

scikit-learn, joblib, dirty-cat



## AI from tabular data to healthcare and society

- Tables: Improving machine learning, drawing from databases
  - diminishing returns of imputation, categorical encoding..
  - the skrub software
  - CARTE foundation model: contextualising numbers
- Application value, eg as in health
  - Not the biggest model, but most appropriate
- We can focus on value, rather than scale



## 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.
- P. Cerda and G. Varoquaux. Encoding high-cardinality string categorical variables. *Transactions in Knowledge and Data Engineering*, 2020.
- A. Cvetkov-Iliev, A. Allauzen, and G. Varoquaux. Relational data embeddings for feature enrichment with background information. *Machine Learning*, pages 1–34, 2023.
- J. Dockès, G. Varoquaux, and J.-B. Poline. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*, 10(9):giab055, 2021.
- 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.
- J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

## References II

- J. M. Kanter and K. Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2015.
- J. Kim, L. Grinsztajn, and G. Varoquaux. Carte: pretraining and transfer for tabular learning. *arXiv soon*, 2024.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What’s a good imputation to predict with missing values? *NeurIPS*, 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.

## References III

- A. Perez-Lebel, G. Varoquaux, M. Le Morvan, J. Josse, and J.-B. Poline. Benchmarking missing-values approaches for predictive models on health databases. *GigaScience*, 11, 2022.
- R. Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1), 2019.
- G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022.
- 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.