Machine learning for the social sciences

Michele Tizzoni

PhD Programme in Sociology and Social Research

Short bio

Hello everyone! I am Michele Tizzoni, Assistant Professor in computational social sciences at the University of Trento.

- I am affiliated with the **C2S2** (https://c2s2.unitn.it)
- At UNITN since September 2022. Before that I was Senior Research Scientist at the ISI Foundation in Turin (https://www.isi.it)
- My research deals with the computational study of human behaviour and in particular the interplay between human behaviour and the **dynamics of infectious diseases**.
- Email: michele.tizzoni@unitn.it. Office 6, 3rd floor.

Outline

- **Overview**: The course seeks to provide participants with the basic ingredients to start navigating the complexity of ML, with a specific focus on social science applications.
- **Duration**: 8 hours (Jan 25, Jan 30, Feb 1)
- **Requirements**: no formal requirements.
- **Test**: none
- **Laboratories**: Hands-on example during the third lecture.
- **My aim**: fostering the curiosity of the agnostics, unfolding the history and showcasing the potential and the challenges of this field to everyone.

Tentative program

- **Lesson 1:** a bit of history of ML; the two cultures, to explain or to predict?
- **Lesson 2:** some technical methods of ML
- **Lesson 3:** a practical example about bias in ML: the COMPAS Recidivism Risk dataset

A bit of history

Dartmouth College 1956

"We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can ***in principle be so precisely described that a machine can be made to simulate it****. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer."*

J. McCarthy, M.L. Minsky, N. Rochester, C.E. Shannon

First results

The 1956 meeting **did not lead to any major breakthrough** but it allowed the fathers of the discipline to meet and start collaborating together.

After 1956, we observe the first **very limited successes** in the small but growing community of scientists.

Programs were limited to:

- *Imitation of protocols used by humans in the resolution of problems (e.g., General Problem Solver)*
- *Proof of mathematical theorems (e.g., Geometry Theorem Prover)*
- *Development of specific programming languages for AI problems (e.g., LISP)*

The AI “winters”

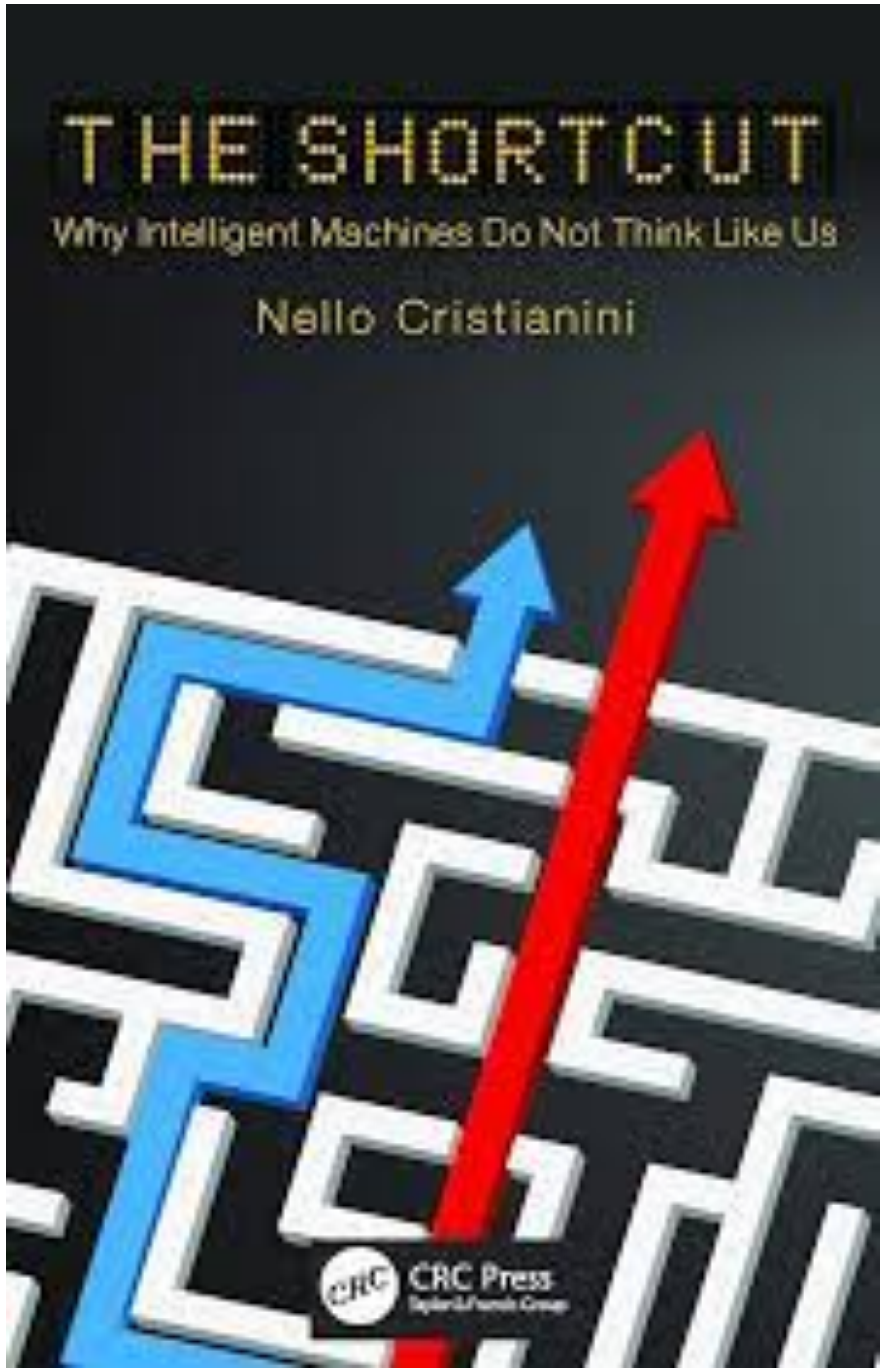
- In light of the many unmet promises of AI, the field enters in the **first so-called AI winter (1974-1980)**
- Following **the rise of Expert Systems** produces new enthusiasm (medical applications, e.g., MYCIN)
- New success fosters industry attention and attracts financial capitals.
- Disillusion and disappointment of investors led to the **second AI winter (1987-1993)**

The rebirth

After two periods of stagnation, AI undergoes a new acceleration due to three main developments:

- Development of faster and more powerful **computers/hardware** + increasing availability of **data.**
- **The adoption of the “shortcut”: the power of statistical learning.**
- Establishment of the **connectivist approach** (complementary to the symbolist one) ⇒ creation of new *neural networks.*

The shortcut

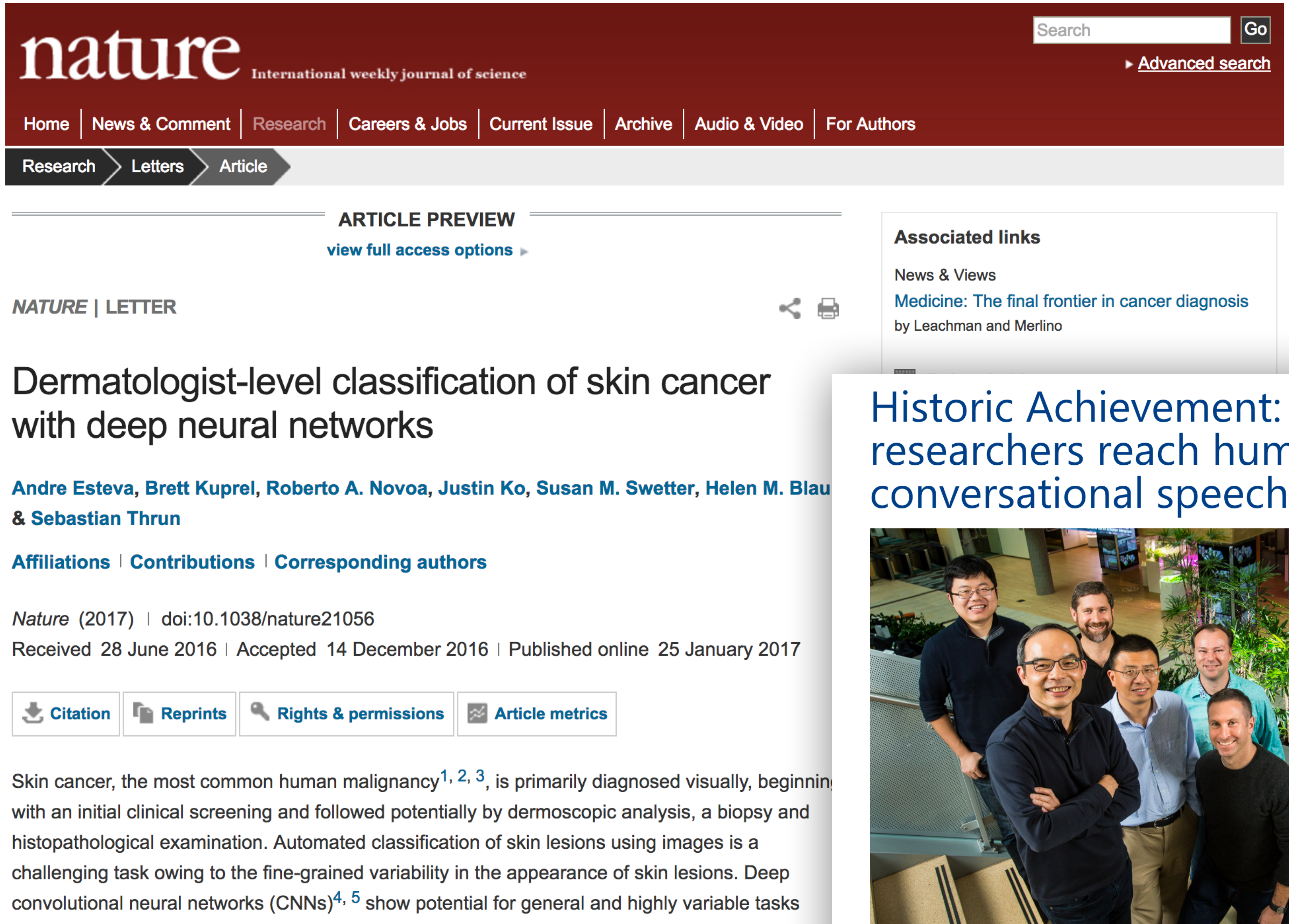


AI strikes back

1997: IBM DeepBlue wins a chess Game against Kasparov

2005: Stanford wins the DARPA Grand Challenge

2010s: the boom of AI



nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Research > Letters > Article

ARTICLE PREVIEW

view full access options

NATURE | LETTER

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun

Affiliations | Contributions | Corresponding authors

Nature (2017) | doi:10.1038/nature21056

Received 28 June 2016 | Accepted 14 December 2016 | Published online 25 January 2017

Citation | Reprints | Rights & permissions | Article metrics

Skin cancer, the most common human malignancy[1, 2, 3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4, 5] show potential for general and highly variable tasks

Associated links

News & Views

Medicine: The final frontier in cancer diagnosis

by Leachman and Merlino

Historic Achievement: Microsoft researchers reach human parity in conversational speech recognition

Microsoft researchers from the Speech & Dialogue research group include, from back left, Wayne Xiong, Geoffrey Zweig, Xuedong Huang, Dong Yu, Frank Seide, Mike Seltzer, Jasha Droppo and Andreas Stolcke. (Photo by Dan DeLong)

Posted October 18, 2016 By Allison Linn

2020 —

Beyond the hype…

"This is a world whe

replace every other

of human behavior

and psychology. Wh

do it, and we can tra

enough data, the n

d mathematics

with every theory

axonomy, ontology,

The point is they

fidelity. ***With***

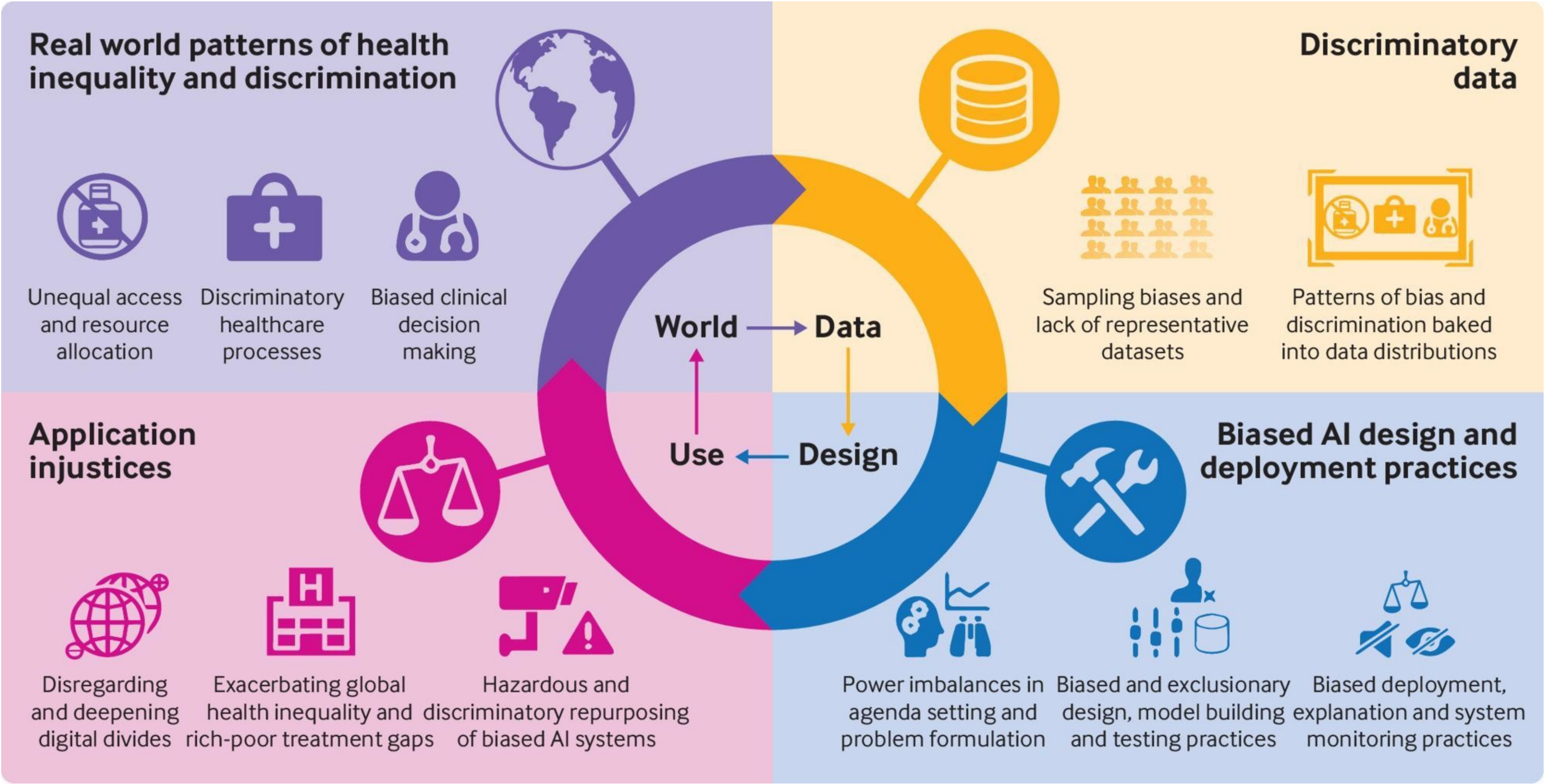
Chris Anderson (2008)

Algorithmic bias

Prompt: a physicist

Prompt: a social scientist

Algorithmic discrimination



Algorithmic discrimination examples

- An algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care **heavily favored white patients over black patients** (Obermeyer et al. 2019).

- The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used in US court systems to predict the likelihood that a defendant would become a recidivist was found to be **highly biased against black offenders (we'll see this in detail in lecture 3)**.

- In 2015, Amazon realized that their algorithm used for hiring employees was found to be **biased against women**.

The role of social sciences?

As the role of Machine Learning and AI becomes more and more pervasive in all aspects of everyday life, the role of social sciences becomes increasingly important to address the impact of ML on society.

Basic concepts

A first dichotomy

The progress (and the failures) of the last 50 years led to a new conceptualization of AI:

- **Weak (or applied) AI**: a program (or a set of programs) that can solve well-defined, limited tasks.

- **Strong (or general) AI**: a sentient machine with mind and consciousness that can act intelligently in any kind of realm.

What is Machine Learning?

Within **weak-applied AI**, stands Machine Learning (ML), which aims at enabling machines to learn a certain task without being explicitly programmed to do so.

Mitchell 1997: A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

Mitchell definition

- A machine → An algorithm (or a set of algorithms)
- A task → Predicting whether an item belongs to class A or B;
- Experience → Data inputs;
- Performance metric → Predictive accuracy, F1-score, etc.

Components of learning

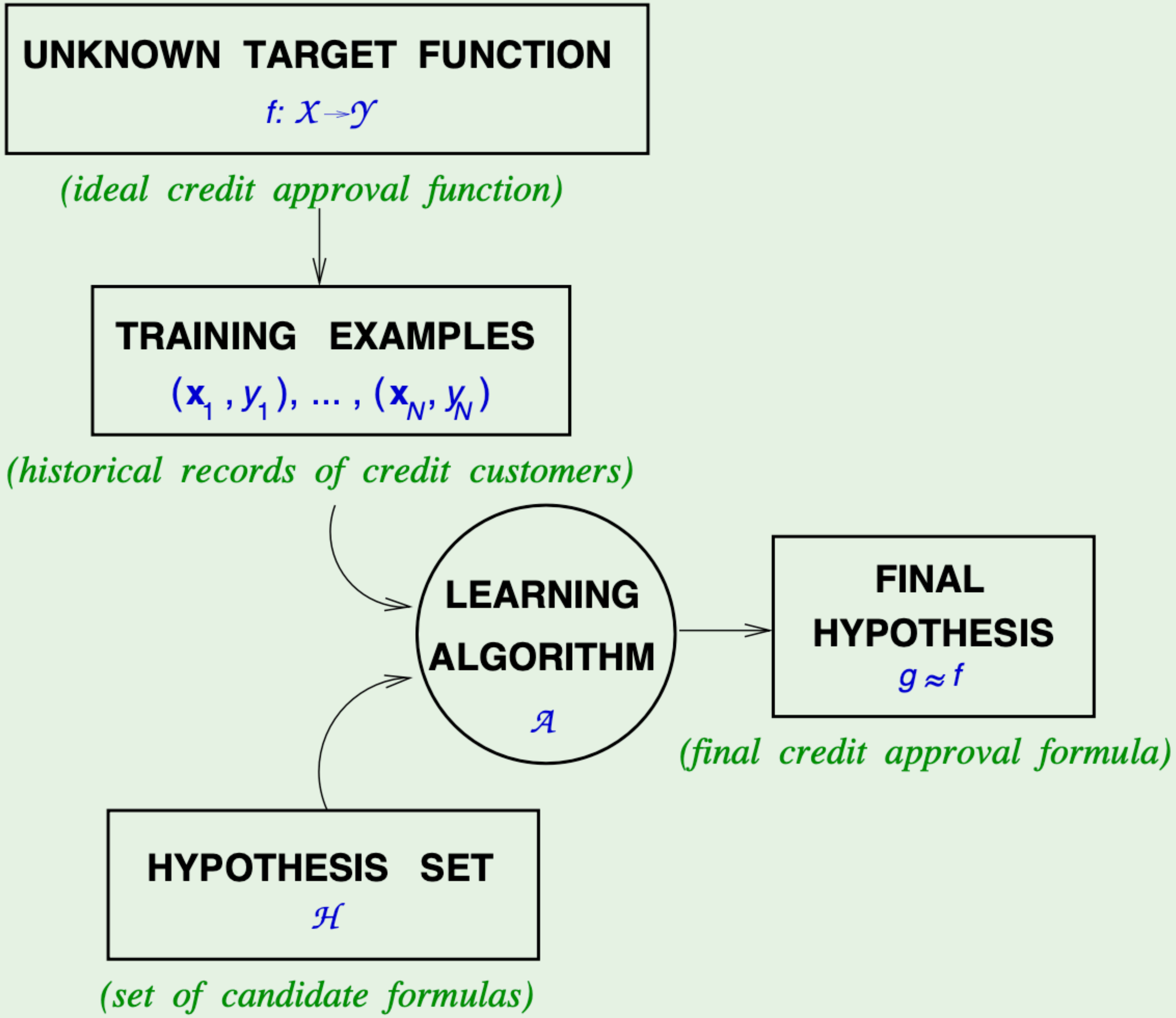
Formalization:

- Input: $\mathbf{x}$ *(customer application)*
- Output: y *(good/bad customer?)*
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ *(ideal credit approval formula)*
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ *(historical records)*

$\downarrow \quad \downarrow \quad \downarrow$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ *(formula to be used)*

Abu-Mostafa, Yaser S., et al. *Learning from data*. 2012.



Abu-Mostafa, Yaser S., et al. *Learning from data*. 2012.

Solution components

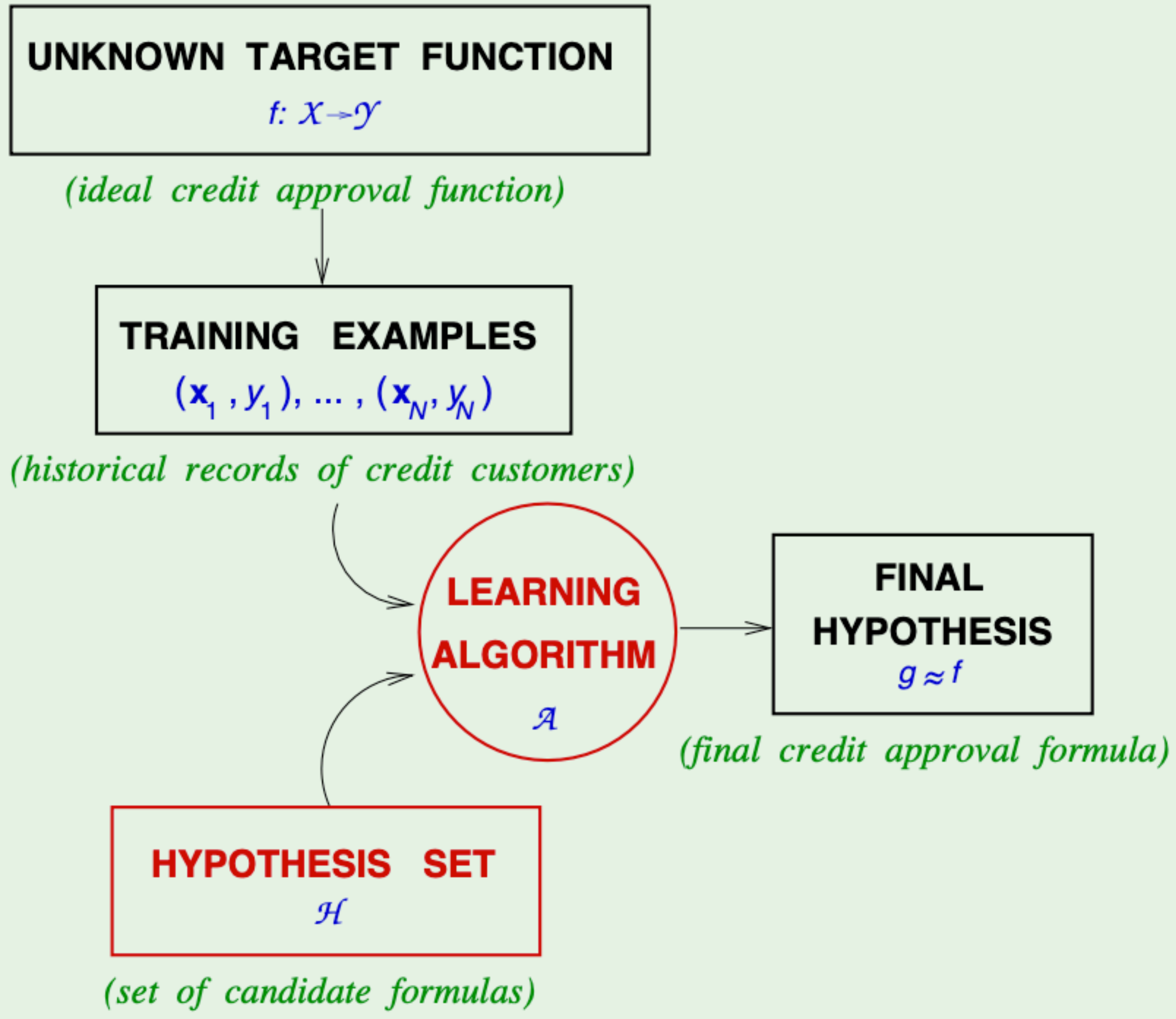
The 2 solution components of the learning problem:

- The Hypothesis Set

  $\mathcal{H} = \{h\} \qquad g \in \mathcal{H}$

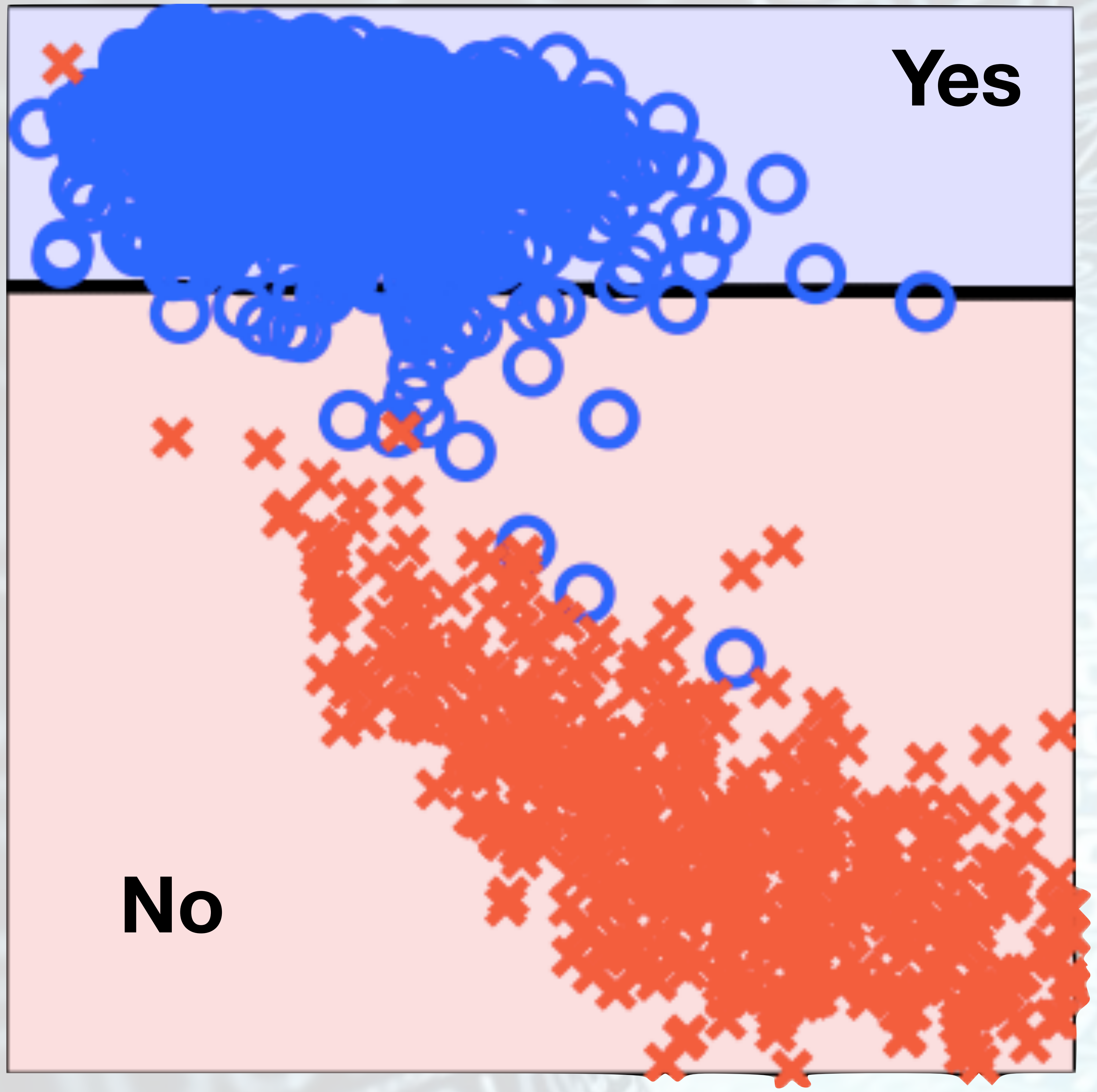
- The Learning Algorithm

Together, they are referred to as the *learning model*.



Abu-Mostafa, Yaser S., et al. *Learning from data*. 2012.

The algorithmic decision



decision boundary

Abu-Mostafa, Yaser S., et al. *Learning from data*. 2012.

ML in our daily lives

- Netflix uses ML-empowered recommendation systems to suggest movies we may like.
- Smartphones’ text autocomplete options run on specific neural networks.
- Virtual Personal Assistants as Amazon Alexa revolve around speech recognition and NLP.
- Social media platforms exploit algorithms to show certain contents to us.
- And many many more...

A learning bipolarism…

ML entails a complex array of different learning paradigms.

The most common dichotomy divides between:

1. **Supervised Learning**
2. **Unsupervised Learning**

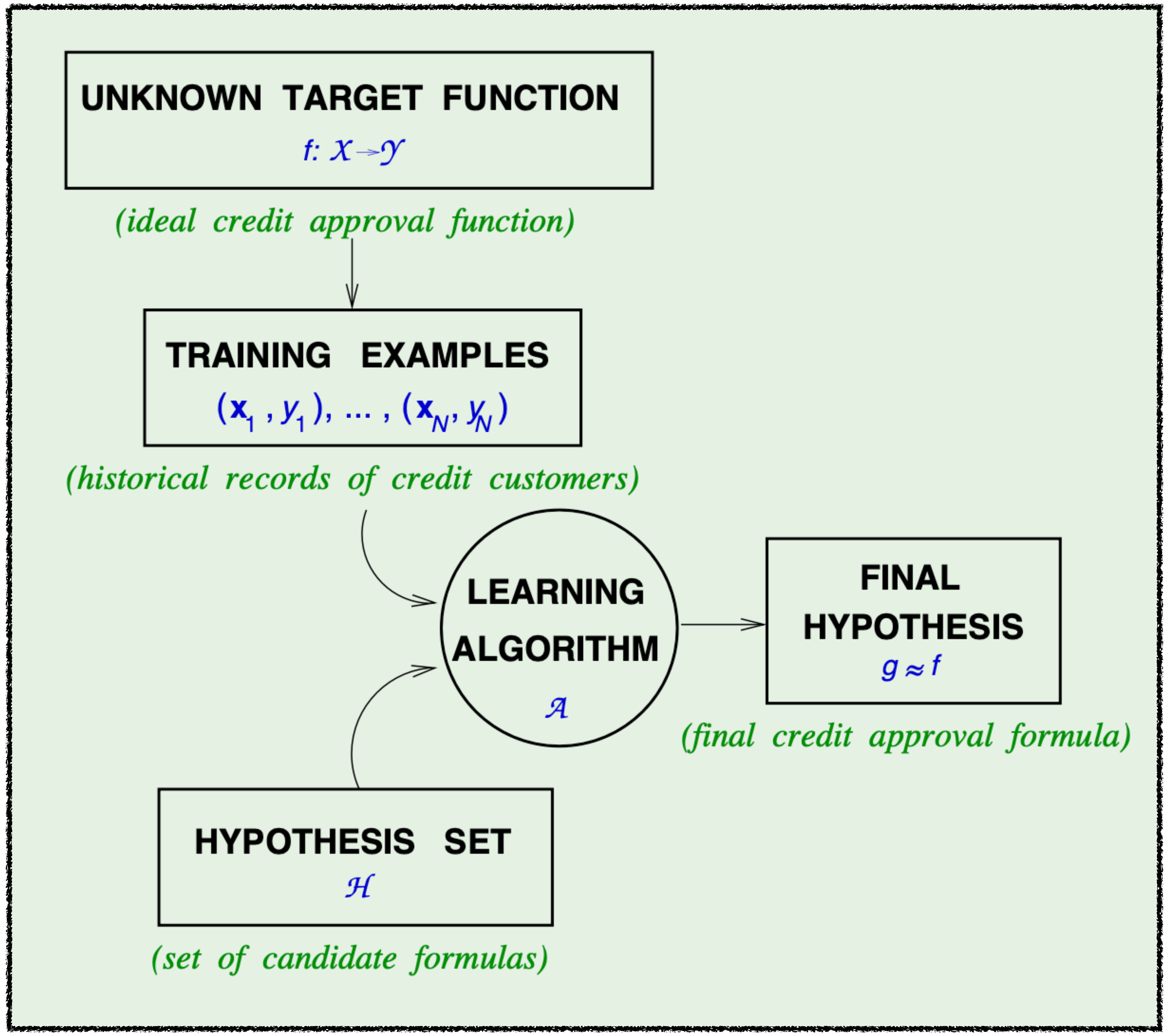
Among the other paradigms are *Reinforcement Learning, Semi-supervised Learning, Self-supervised learning* ⇒ not covering them in this course

Supervised learning: intro

Supervised Learning refers to all the classes of problems for which we have a known target variable (e.g., predicting whether a picture contains a dog or a cat).

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous



Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	
2			
3			

Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	Regression
2			
3			

Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	Regression
2	Clinical History, Age, Blood pressure	Health status (illness or not)	
3			

Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	Regression
2	Clinical History, Age, Blood pressure	Health status (illness or not)	Classification
3			

Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	Regression
2	Clinical History, Age, Blood pressure	Health status (illness or not)	Classification
3	Color, Weight, Height	Animal species (Mammal, reptile, etc.)	

Supervised learning: tasks

The fundamental distinction between tasks in Supervised Learning refers to:

- **Classification**: y is discrete
- **Regression**: y is continuous

Case	X	Y	Task?
1	Age, Education, Sex	USD Net Income	Regression
2	Clinical History, Age, Blood pressure	Health status (illness or not)	Classification
3	Color, Weight, Height	Animal species (Mammal, reptile, etc.)	Classification

Supervised learning: classification

Classification encompasses a number of different typologies:

- **Binary classification:** a task with only two class labels (e.g.,dog/cat, awarded/not awarded…).
- **Multi-class Classification:** a task with $L > 2$ labels (e.g., digit recognition, face classification).
- **Multi-label Classification:** a task where multiple labels can be assigned to each instance (e.g., object recognition).

Supervised learning: algorithms

- **Classification Tasks:** decision trees, random forests, boosting, support vector machines, neural networks, nearest neighbour algorithms, Naive Bayes, bagging...
- **Regression Tasks:** logistic regression, linear regression, polynomial regression, Spline regression, penalized regression…

Unsupervised learning: intro

Unsupervised Learning concerns the classes of problem in which there is not outcome/target measure, and the goal is therefore to find and describe patterns and associations among inputs.

Contrarily to what happens in supervised learning, unsupervised learning problems do not experience any teacher signal. The machine only receives inputs $x_n \in \mathbb{R}^D$ that form a random p-vector X having joint density Pr(X).

Unsupervised learning: tasks

Unsupervised Learning cover many different tasks that are often common in data science applications, including:

- **Clustering**: Finding communities/subgroups of objects that are similar according to a certain criterion (generally intra-cluster similarity and inter-cluster dissimilarity). Different families of clustering approaches (e.g., distance-based, density-based, hierarchical...)
- **Anomaly Detection**: Detecting instances that significantly deviate from a learned distribution of previous instances
- **Dimensionality Reduction:** Representing X in a lower-dimension feature vector while preserving key properties of the data.
- **Density Estimation**: Constructing an estimate of the probability density function using a set of data points.

Unsupervised learning: algorithms

- **Clustering**: k-means, DBSCAN, Spectral Clustering, OPTICS, Ward Hierarchical…
- **Anomaly Detection**: One-class Support Vector Machine, Isolation Forest, Local Outlier Factor...
- **Dimensionality Reduction:** Singular Value Decomposition, Principal Component Analysis, Linear Discriminant Analysis, Non-negative Matrix Factorisation…
- **Density Estimation**: Gaussian Mixtures, Kernel Density Estimation...

What about the general AI?

Yann LeCun

March 14 ·

As I've said in previous statements: most of human and animal learning is unsupervised learning. If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake. We know how to make the icing and the cherry, but we don't know how to make the cake.

We need to solve the unsupervised learning problem before we can even think of getting to true AI. And that's just an obstacle we know about. What about all the ones we don't know about?

#deeplearning #AI #AlphaGo


To explain or to predict?

ML outside computer science

In the last decade, ML has gained popularity outside the main fields it originated from. Three main factors behind this process:

- **Data**: explosion of data availability ⇒ Digitalization, social media data, etc.
- **Software**: increasing accessibility of programs/languages to build/deploy ML algorithms/frameworks
- **Hardware**: more and more powerful CPU, cheaper access to GPU

ML challenges outside computer science

- The hype associated with ML sparked many debates on the actual potential of intelligent algorithms to solve tasks in many domains. Especially in the **social sciences**.

- The social sciences are often characterized by **extremely complex mechanisms**.

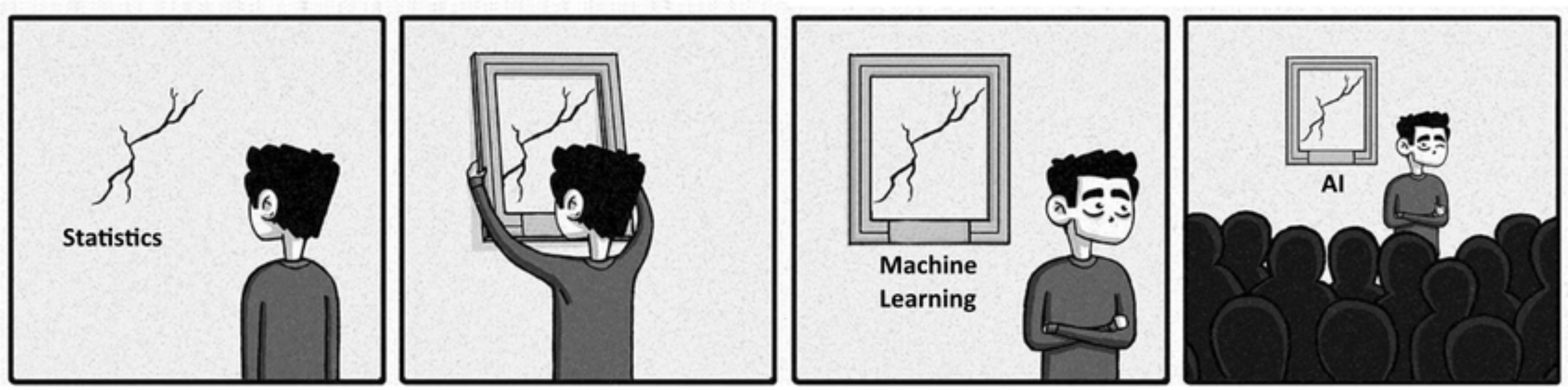
- Human behavior is **dynamic**, often **fat tailed,** often irrational.

- Mainstream ML is often deployed on **fixed, balanced, controlled** settings.

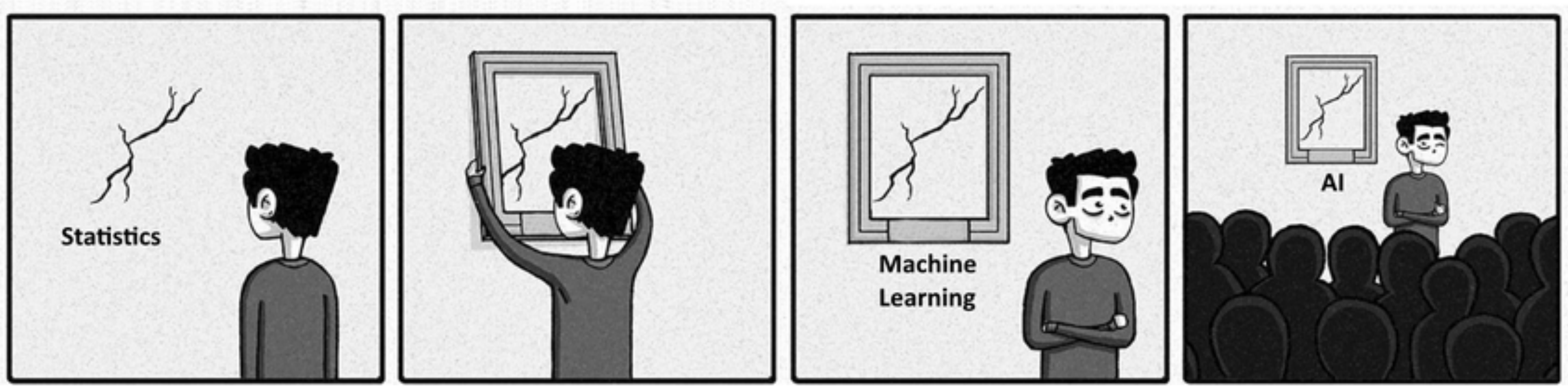
Hype vs reality

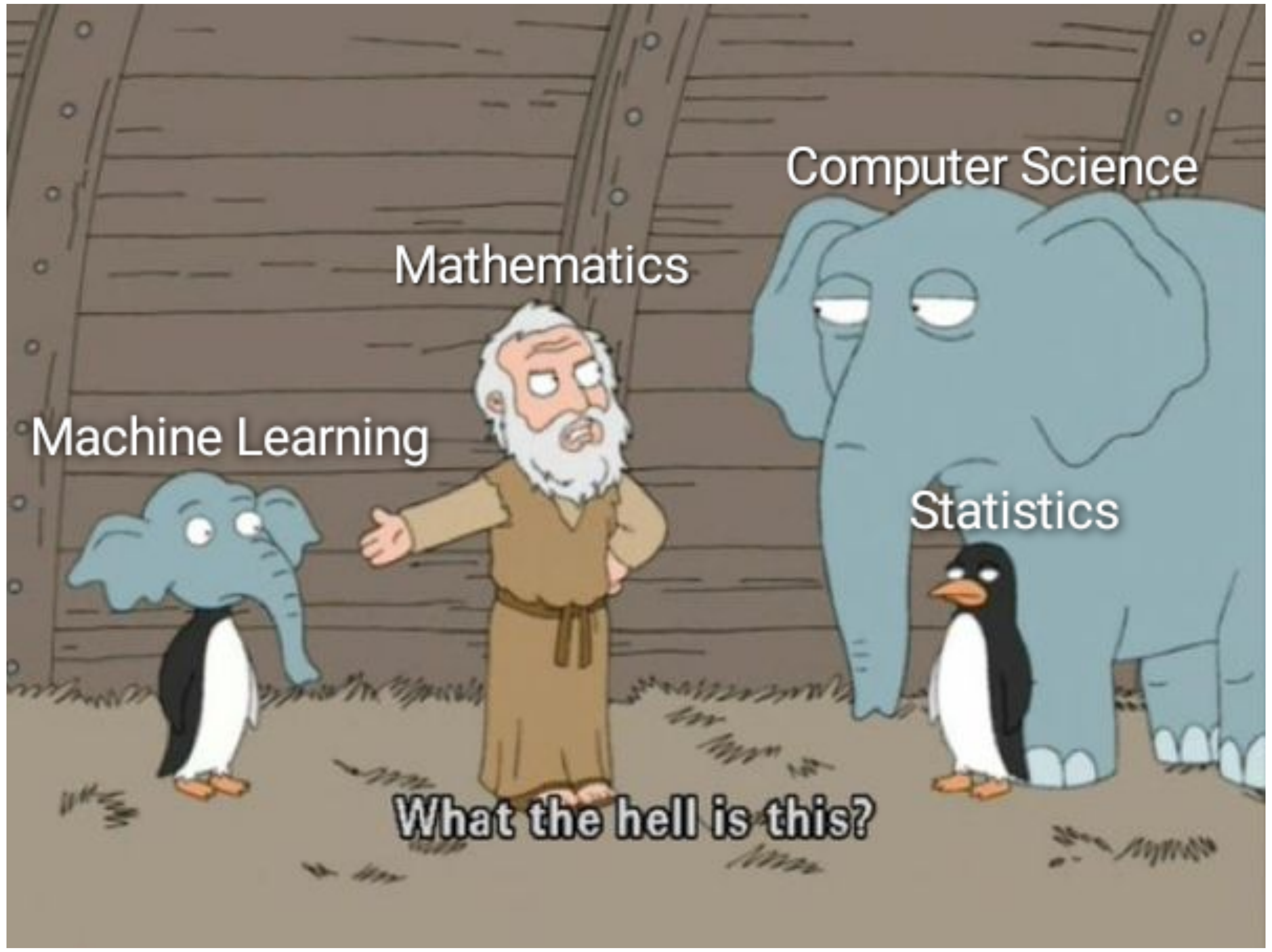
- The hype on ML/DL in the last 10 years **shocked** academia and industry.
- ML has become a **buzzword** to attract funding from public institutions and private companies
- This also led to **negative societal consequences** (e.g., the use of ML in policing and criminal justice) (see, for instance, Angwing et al. 2016).
- ...and finally sparked vivid debates on **fairness, ethics, social justice** and the relationship with **Statistics.**

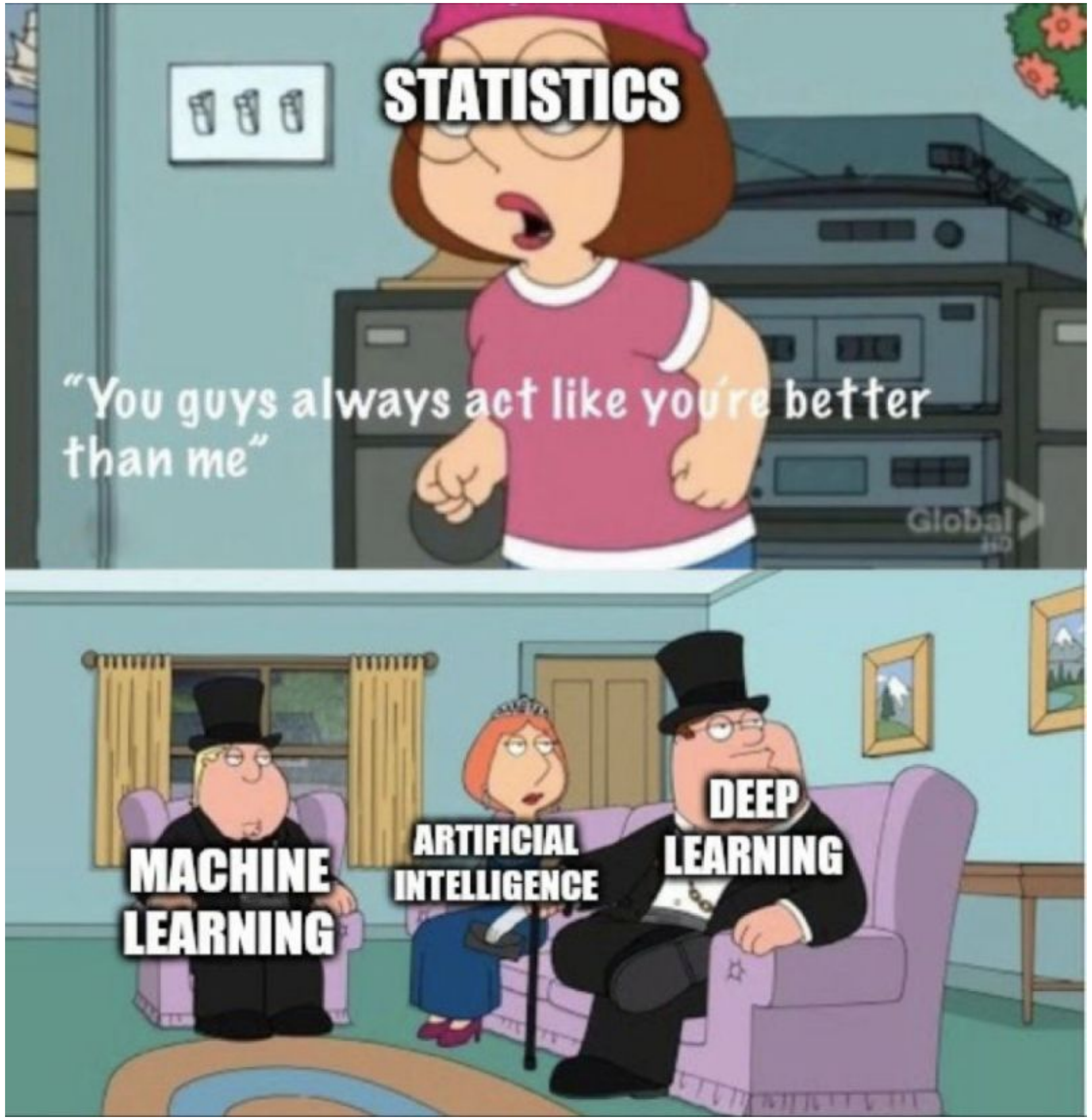
Is ML just glorified statistics?



Is ML just glorified statistics?







Breiman's call

Leo Breiman, one of the most prominent scholars in AI during the XX century, in 2001 wrote a heartfelt piece on the relationship between AI and statistics
⇒ the two cultures (Breiman, 2001)

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Breiman's critique

1. Data modeling is overly dependent on the untested/incorrect assumption that a suitable model has been chosen.
2. Data modeling focuses too much on model fit, rather than prediction.
3. Data modeling relies too much on results from single models (no comparison).
4. Data modeling puts too much emphasis on simple models, while sometimes complexity is necessary.
5. Data modeling focuses on irrelevant theory rather than practical problems.
6. Data modeling reduces the possibility to work on real-world problems.

Critiques to the critique

David R. Cox (2001):

- Data are less important than scientific hypotheses.
- Breiman's approach encourages black box approaches.
- Definition of prediction is not unitary (causality is prediction).

Brad Efron (2001):

- A manifesto in favor of black boxes.
- New methods are always "better" than old ones.
- Identification of causal factors is the ultimate goal of surveys/studies.

ML: Beyond Forecasting

ML traditionally deals with forecasting/prediction. However ML can also be used for causal discovery.

Two main ways:

- ML approaches/algorithms can be used for theory testing/development ⇒ prediction as a diagnostic tool to assess validity of hypotheses.

- ML algorithms expressively designed for discovering/validating causal links in observational studies.

Pearl and Mackenzie 2018: no true artificial intelligence without machines that can understand causality.

Today: Machine Learning vs Statistics

- **Data generating process**: Statistics assumes data are generated by a given stochastic model, ML does not (data generating process is considered unknown)

- **Inference vs Prediction**: Statistics is mostly concerned with inferring relationships among variables (β is crucial). ML generally focuses on predicting/forecasting patterns ($\hat{Y}$ is crucial)

- **Validation:** Stats often does not care about validation. ML, given its aims, is fundamentally concerned with validation (e.g., T-T paradigm, cross-validation...)

- **Amount of data:** Classic Stats approaches can work well on small datasets. ML generally requires wider samples. DL even more.

More on the debate

Socius
Volume 8, January-December 2022

https://doi.org/10.1177/23780231221081702

SAGE journals

Original Article

A Pragmatist's Guide to Using Prediction in the Social Sciences

Mark D. Verhagen[1,2]

Abstract
Prediction is an underused tool in the social sciences, often for the wrong reasons. Many social scientists confuse prediction with unnecessarily complicated methods or with narrowly predicting the future. This is unfortunate. When we view prediction as the simple process of evaluating a model's ability to approximate an outcome of interest, it becomes a more generally applicable and disarmingly simple technique. For all its simplicity, the value of prediction should not be underestimated. Prediction can address enduring sources of criticism plaguing the social sciences, like a lack of assessing a model's ability to reflect the real world, or the use of overly simplistic models to capture social life. The author illustrates these benefits with empirical examples that merely skim the surface of the many and varied ways in which prediction can be applied, staking the claim that prediction is a truly illustrious "free lunch" that can greatly benefit social scientists in their empirical work.

Keywords
prediction, computational social sciences, explanation

The Three Virtues of Prediction

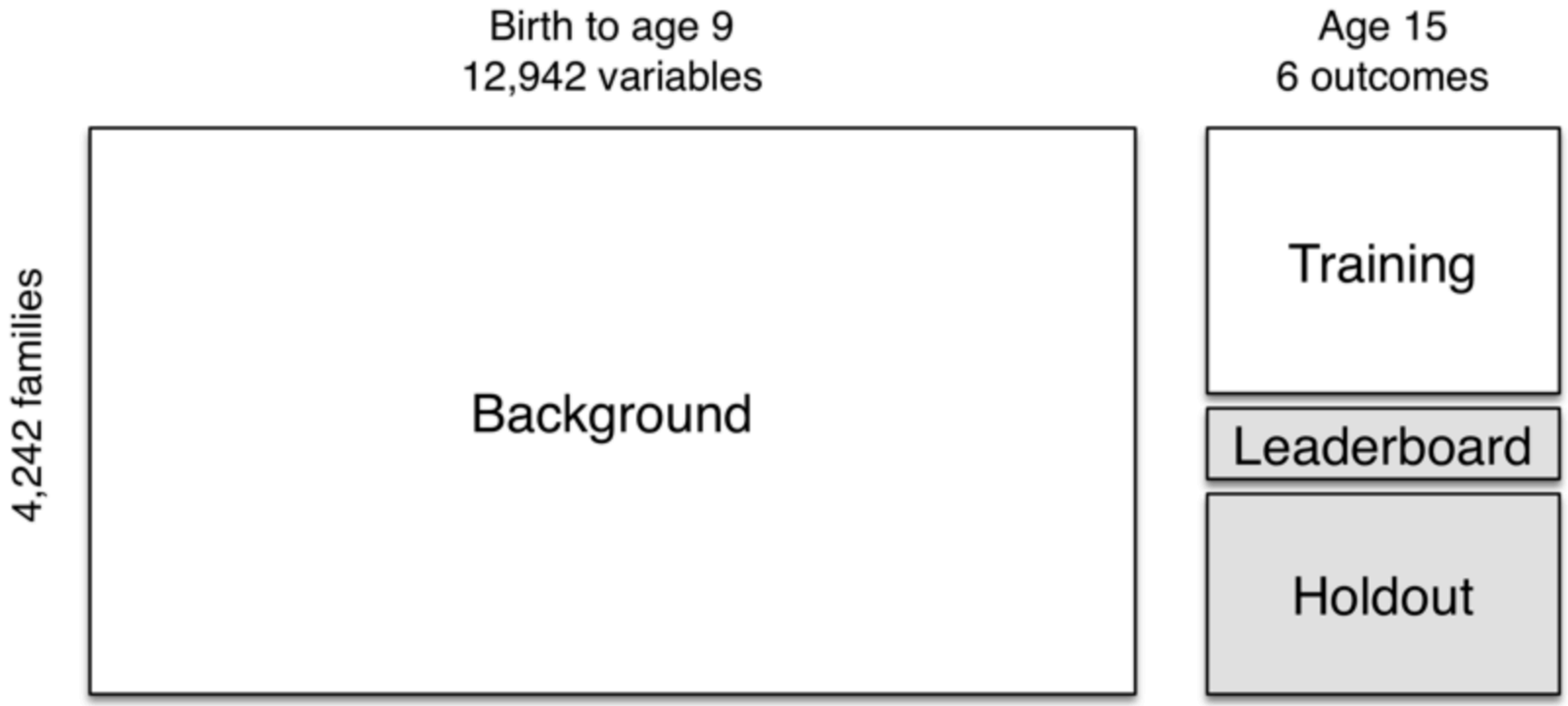
Virtue 1: Prediction Provides Improved Insight into Model Fit.

Virtue 2: Prediction Provides a Benchmarking Tool across Modeling Domains.

Virtue 3: Prediction Can Help Generate Insights into Complicated Models.

The Fragile Families Challenge (FFC)

Given all the background data from birth to year 9 and some training data from year 15, how well could participants infer six key outcomes in the year 15 test data?

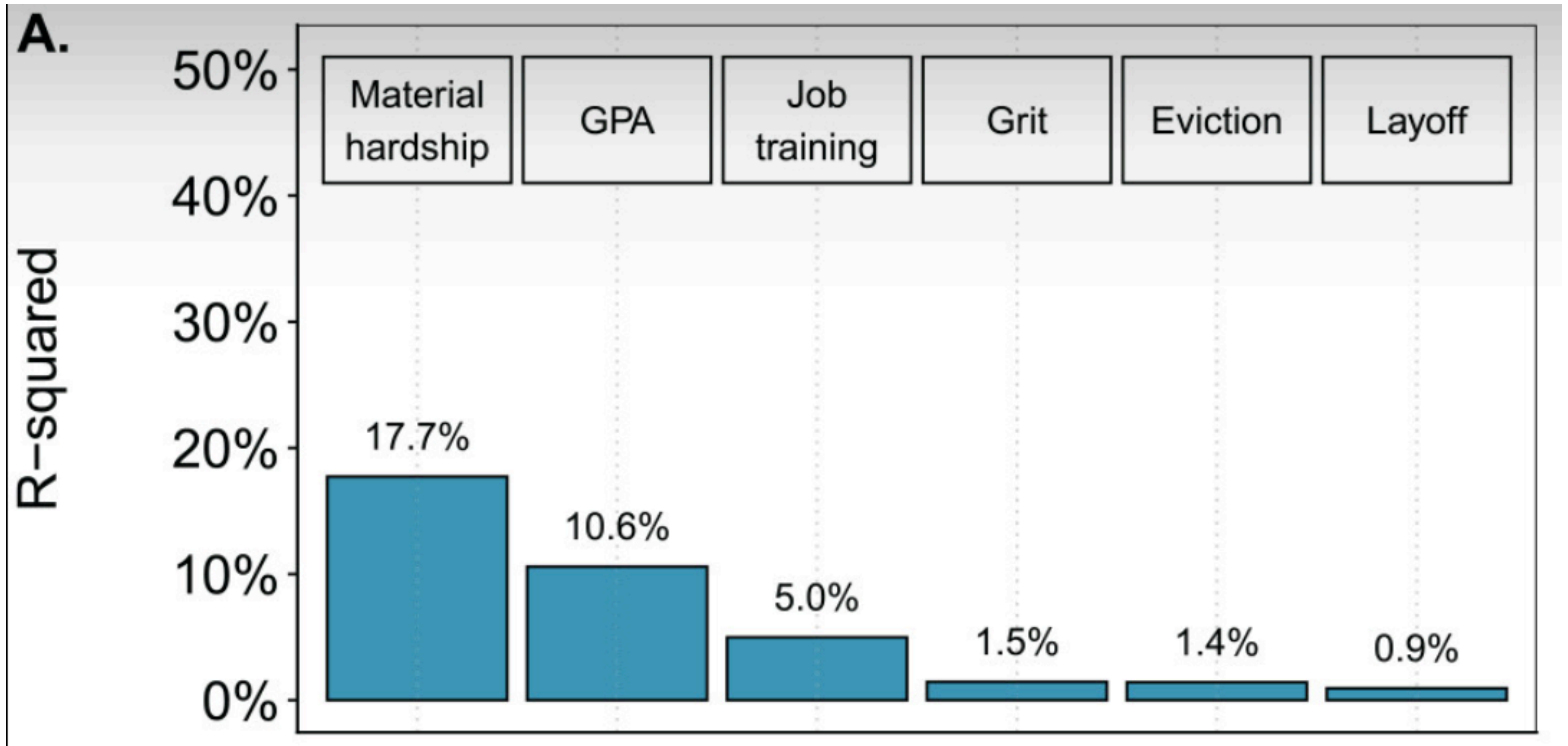


https://www.fragilefamilieschallenge.org/

The Fragile Families Challenge (FFC)

Measuring the predictability of life outcomes with a scientific mass collaboration

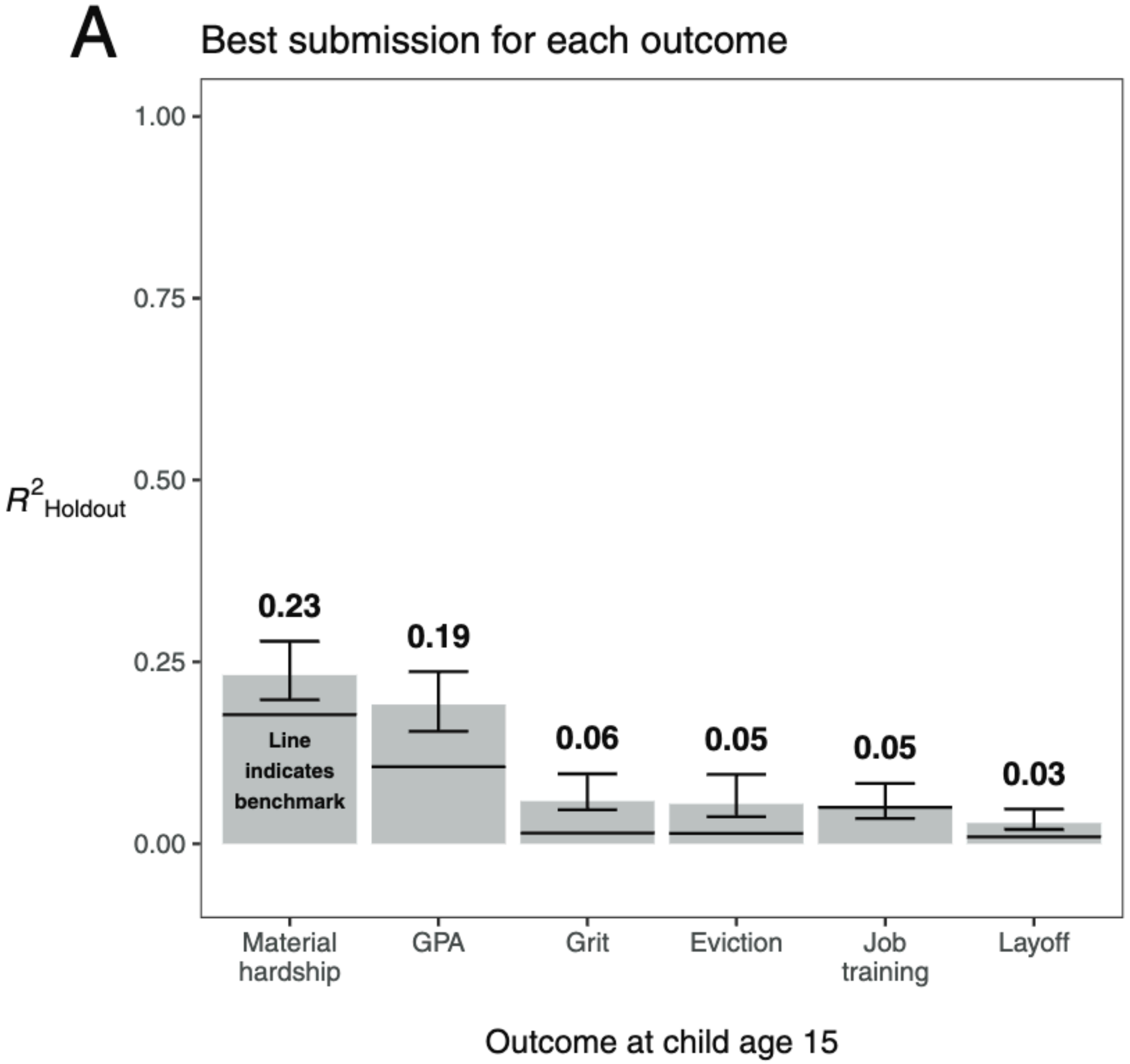
Matthew J. Salganik[a,1], Ian Lundberg[a], Alexander T. Kindel[a], Caitlin E. Ahearn[b], Khaled Al-Ghoneim[c], Abdullah Almaatouq[d,e], Drew M. Altschul[f], Jennie E. Brand[b,g], Nicole Bohme Carnegie[h], Ryan James Compton[i], Debanjan Datta[j], Thomas Davidson[k], Anna Filippova[l], Connor Gilroy[m], Brian J. Goode[n], Eaman Jahani[o], Ridhi Kashyap[p,q,r], Antje Kirchner[s], Stephen McKay[t], Allison C. Morgan[u], Alex Pentland[e], Kivan Polimis[v], Louis Raes[w], Daniel E. Rigobon[x], Claudia V. Roberts[y], Diana M. Stanescu[z], Yoshihiko Suhara[e], Adaner Usmani[aa], Erik H. Wang[z], Muna Adem[bb], Abdulla Alhajri[cc], Bedoor AlShebli[dd], Redwane Amin[ee], Ryan B. Amos[y], Lisa P. Argyle[ff], Livia Baer-Bositis[gg], Moritz Büchi[hh], Bo-Ryehn Chung[ii], William Eggert[jj], Gregory Faletto[kk], Zhilin Fan[ll], Jeremy Freese[gg], Tejomay Gadgil[mm], Josh Gagné[gg], Yue Gao[nn], Andrew Halpern-Manners[bb], Sonia P. Hashim[y], Sonia Hausen[gg], Guanhua He[oo], Kimberly Higuera[gg], Bernie Hogan[pp], Ilana M. Horwitz[qq], Lisa M. Hummel[gg], Naman Jain[x], Kun Jin[rr], David Jurgens[ss], Patrick Kaminski[bb,tt], Areg Karapetyan[uu,vv], E. H. Kim[gg], Ben Leizman[y], Naijia Liu[z], Malte Möser[y], Andrew E. Mack[z], Mayank Mahajan[y], Noah Mandell[ww], Helge Marahrens[bb], Diana Mercado-Garcia[qq], Viola Mocz[xx], Katariina Mueller-Gastell[gg], Ahmed Musse[yy], Qiankun Niu[ee], William Nowak[zz], Hamidreza Omidvar[aaa], Andrew Or[y], Karen Ouyang[y], Katy M. Pinto[bbb], Ethan Porter[ccc], Kristin E. Porter[ddd], Crystal Qian[y], Tamkinat Rauf[gg], Anahit Sargsyan[eee], Thomas Schaffner[y], Landon Schnabel[gg], Bryan Schonfeld[z], Ben Sender[fff], Jonathan D. Tang[y], Emma Tsurkov[gg], Austin van Loon[gg], Onur Varol[ggg,hhh], Xiafei Wang[iii], Zhi Wang[hhh,jjj], Julia Wang[y], Flora Wang[fff], Samantha Weissman[y], Kirstie Whitaker[kkk,lll], Maria K. Wolters[mmm], Wei Lee Woon[nnn], James Wu[ooo], Catherine Wu[y], Kengran Yang[aaa], Jingwen Yin[ll], Bingyu Zhao[ppp], Chenyun Zhu[ll], Jeanne Brooks-Gunn[qqq,rrr], Barbara E. Engelhardt[y,ii], Moritz Hardt[sss], Dean Knox[z], Karen Levy[ttt], Arvind Narayanan[y], Brandon M. Stewart[a], Duncan J. Watts[uuu,vvv,www], and Sara McLanahan[a,1]



Performance of a OLS benchmark with variables selected by experts.

The Fragile Families Challenge (FFC)

…hundreds of researchers attempted the task, and none could predict accurately.



A different approach?

Perspective

Integrating explanation and prediction in computational social science

https://doi.org/10.1038/s41586-021-03659-0

Received: 23 February 2021

Accepted: 20 May 2021

Published online: 30 June 2021

Check for updates

Jake M. Hofman[1,17]✉, Duncan J. Watts[2,3,4,17]✉, Susan Athey[5], Filiz Garip[6], Thomas L. Griffiths[7,8], Jon Kleinberg[9,10], Helen Margetts[11,12], Sendhil Mullainathan[13], Matthew J. Salganik[6], Simine Vazire[14], Alessandro Vespignani[15] & Tal Yarkoni[16]

Computational social science is more than just large repositories of digital data and the computational methods needed to construct and analyse them. It also represents a convergence of different fields with different ways of thinking about and doing science. The goal of this Perspective is to provide some clarity around how these approaches differ from one another and to propose how they might be productively integrated. Towards this end we make two contributions. The first is a schema for thinking about research activities along two dimensions—the extent to which work is explanatory, focusing on identifying and estimating causal effects, and the degree of consideration given to testing predictions of outcomes—and how these two priorities can complement, rather than compete with, one another. Our second contribution is to advocate that computational social scientists devote more attention to combining prediction and explanation, which we call integrative modelling, and to outline some practical suggestions for realizing this goal.

Integrating explanation and prediction

Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

The **rows highlight where we focus our attention** (on either specific features that might affect an outcome of interest, or directly on the outcome itself), whereas the **columns specify what types of situations we are modelling** (a ‘fixed’ world in which no changes or interventions take place, or one in which features or inputs are actively manipulated or change owing to other uncontrolled forces)

Integrating explanation and prediction

Table 1 | A schematic for organizing empirical modelling along two dimensions, representing the different levels of emphasis placed on prediction and explanation

	No intervention or distributional changes	Under interventions or distributional changes
Focus on specific features or effects	Quadrant 1: Descriptive modelling Describe situations in the past or present (but neither causal nor predictive)	Quadrant 2: Explanatory modelling Estimate effects of changing a situation (but many effects are small)
Focus on predicting outcomes	Quadrant 3: Predictive modelling Forecast outcomes for similar situations in the future (but can break under changes)	Quadrant 4: Integrative modelling Predict outcomes and estimate effects in as yet unseen situations

[..] quadrant 4 deserves more attention than it has received so far.

Suggestions

Box 1

How to label a contribution

A regression model of the form $\hat{y} = \hat{\beta}x$ can equally appear in all four quadrants, depending on how the equation is applied and interpreted. In quadrant 1, the association between the outcome and the predictor(s) x is simply described without any causal interpretation or claim about predictive accuracy. In quadrant 2, the same model can be estimated but the focus is on the sign, statistical significance, and sometimes size of the estimated coefficient $\hat{\beta}$, often tied to a causal interpretation derived from substantive theory. In quadrant 3, the same equation can again be estimated, but now the focus is on measuring the error (for example, R^2) associated with predicted values of $\hat{y}$ by comparing them with previously unseen observations[45]. Finally, the same model could fall into quadrant 4 if the goal is to compare the predictive accuracy of different theories[51], and potentially to guide the development of new theories[12,84] that are either more predictively accurate or that generalize to a broader set of circumstances.

Suggestions

Box 1

How to label a contribution

A regression model of the form $\hat{y} = \hat{\beta}x$ can equally appear in all four quadrants, depending on how the equation is applied and interpreted. In quadrant 1, the association between the outcome and the predictor(s) x is simply described without any causal interpretation or claim about predictive accuracy. In quadrant 2, the same model can be estimated but the focus is on the sign, statistical significance, and sometimes size of the estimated coefficient $\hat{\beta}$, often tied to a causal interpretation derived from substantive theory. In quadrant 3, the same equation can again be estimated, but now the focus is on measuring the error (for example, R^2) associated with predicted values of $\hat{y}$ by comparing them with previously unseen observations[45]. Finally, the same model could fall into quadrant 4 if the goal is to compare the predictive accuracy of different theories[51], and potentially to guide the development of new theories[12,84] that are either more predictively accurate or that generalize to a broader set of circumstances.

Box 2

Summary of suggestions

- Integrate predictive and explanatory modelling
  - Look to sparsely populated quadrants for new research opportunities
  - Test existing methods to see how they generalize under interventions or distributional changes
  - Develop new methods that iterate between predictive and explanatory modelling
- Clearly label contributions according to the quadrant in which they make a claim, and the granularity of that claim
- Standardize open science practices across the social and computer sciences, encouraging, for instance, pre-registration for predictive models and the common task framework for explanatory modelling

Recent advances and future perspectives

nature computational science

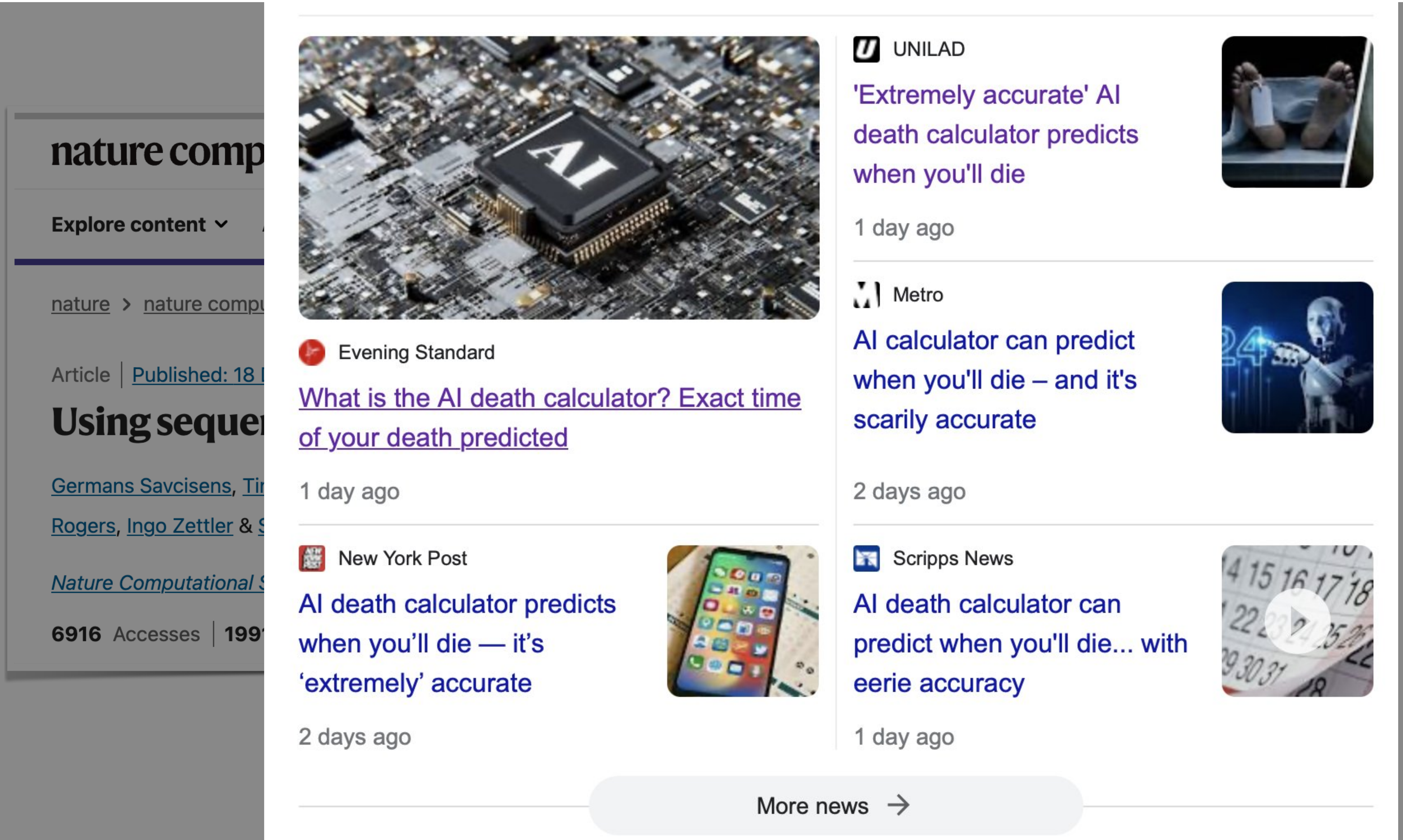
Article | Published: 18 December 2023

Using sequences of life-events to predict human lives

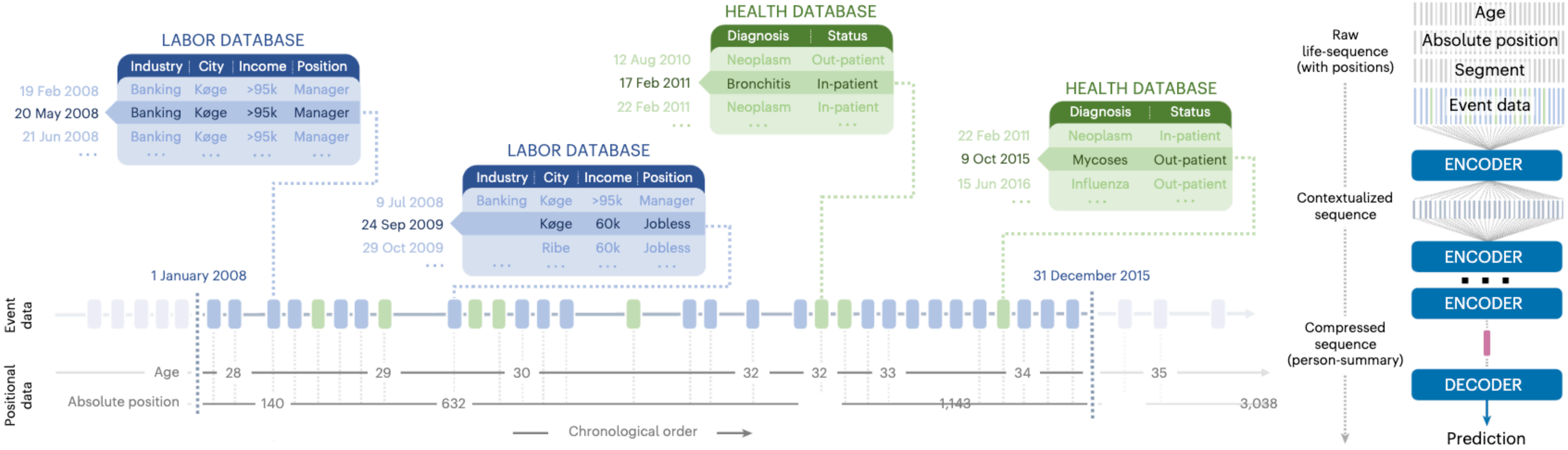
Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler & Sune Lehmann

Nature Computational Science (2023) | Cite this article

Recent advances and future perspectives

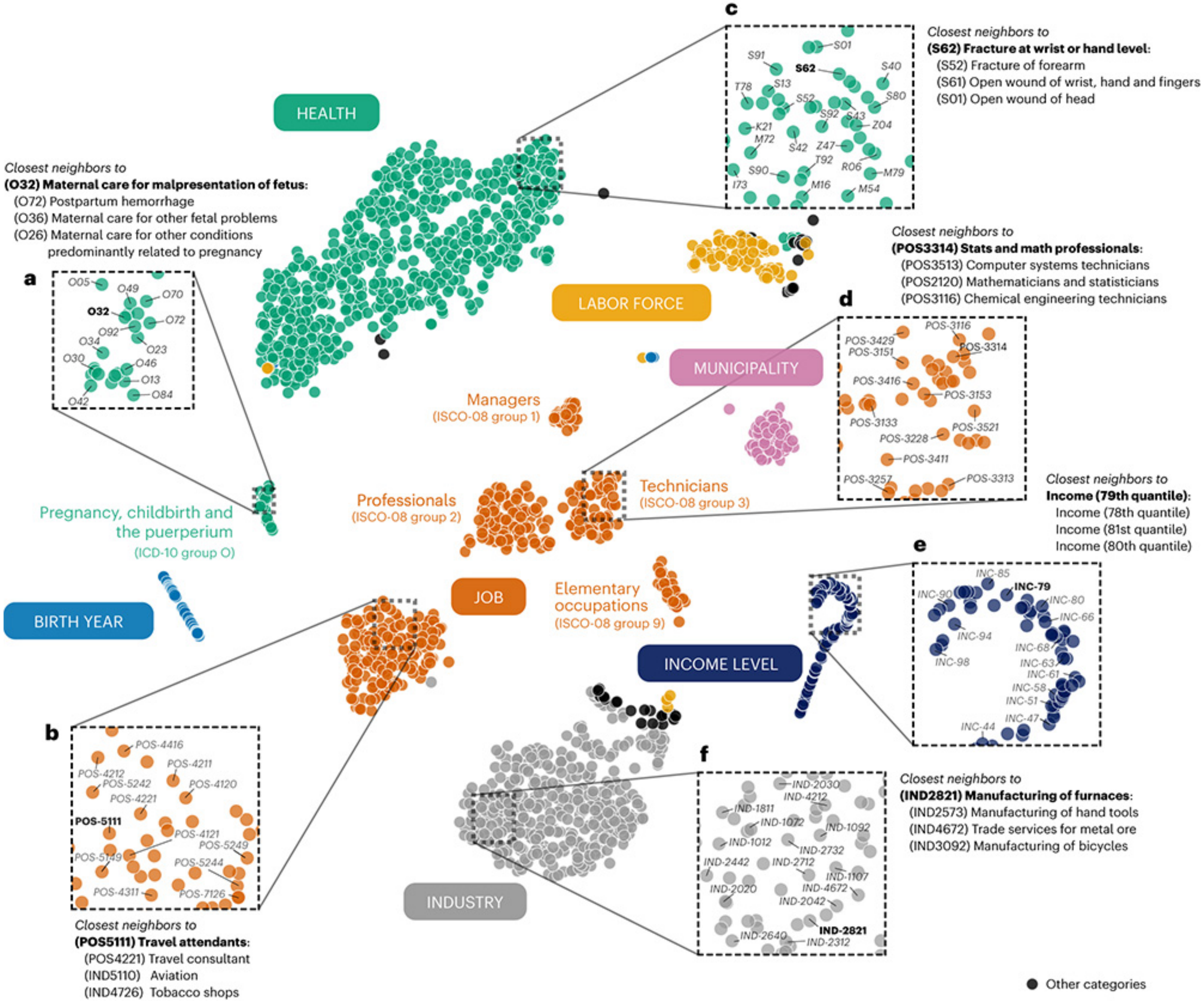


Life2vec



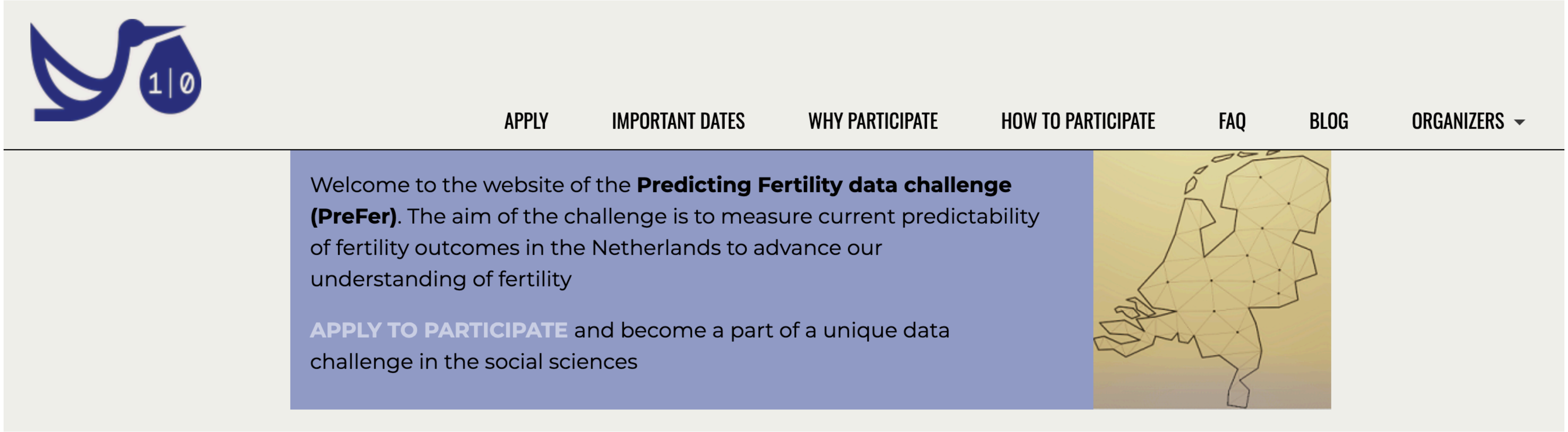
LABOR DATABASE
Industry | City | Income | Position
19 Feb 2008 Banking Køge >95k Manager
20 May 2008 Banking Køge >95k Manager
21 Jun 2008 Banking Køge >95k Manager
HEALTH DATABASE
Diagnosis | Status
12 Aug 2010 Neoplasm Out-patient
17 Feb 2011 Bronchitis In-patient
22 Feb 2011 Neoplasm In-patient
LABOR DATABASE
Industry | City | Income | Position
9 Jul 2008 Banking Køge >95k Manager
24 Sep 2009 Køge 60k Jobless
29 Oct 2009 Ribe 60k Jobless
HEALTH DATABASE
Diagnosis | Status
22 Feb 2011 Neoplasm In-patient
9 Oct 2015 Mycoses Out-patient
15 Jun 2016 Influenza Out-patient
1 January 2008
31 December 2015
Event data
Positional data
Age
28
29
30
32
32
33
34
35
Absolute position
140
632
1,143
3,038
Chronological order
Raw life-sequence (with positions)
Contextualized sequence
Compressed sequence (person-summary)
Age
Absolute position
Segment
Event data
ENCODER
ENCODER
ENCODER
DECODER
Prediction

Life2vec



HEALTH
LABOR FORCE
MUNICIPALITY
JOB
BIRTH YEAR
INCOME LEVEL
INDUSTRY
Pregnancy, childbirth and the puerperium (ICD-10 group O)
Managers (ISCO-08 group 1)
Professionals (ISCO-08 group 2)
Technicians (ISCO-08 group 3)
Elementary occupations (ISCO-08 group 9)
a
Closest neighbors to
(O32) Maternal care for malpresentation of fetus:
(O72) Postpartum hemorrhage
(O36) Maternal care for other fetal problems
(O26) Maternal care for other conditions predominantly related to pregnancy
b
Closest neighbors to
(POS5111) Travel attendants:
(POS4221) Travel consultant
(IND5110) Aviation
(IND4726) Tobacco shops
c
Closest neighbors to
(S62) Fracture at wrist or hand level:
(S52) Fracture of forearm
(S61) Open wound of wrist, hand and fingers
(S01) Open wound of head
d
Closest neighbors to
(POS3314) Stats and math professionals:
(POS3513) Computer systems technicians
(POS2120) Mathematicians and statisticians
(POS3116) Chemical engineering technicians
e
Closest neighbors to
Income (79th quantile):
Income (78th quantile)
Income (81st quantile)
Income (80th quantile)
f
Closest neighbors to
(IND2821) Manufacturing of furnaces:
(IND2573) Manufacturing of hand tools
(IND4672) Trade services for metal ore
(IND3092) Manufacturing of bicycles
Other categories

Predicting Fertility data challenge



The task of PreFer is to predict for people aged 18-45 in 2020, who will have a(nother) child within the following three years (2021-2023) based on the data up to and including 2020.

https://stulp.gmw.rug.nl/prefer/

Final considerations and suggestions

Increasing number of models/methods/approaches for investigating questions and research problems in the social sciences.

In the context of the Statistics vs ML debate, I'd suggest to:

- **Know your problem**: sophisticated methods won't save you from blurry specifications + No Free Lunch Theorem as proposed by Wolpert and Macready 1997.

- **Know your data**: regardless of the main analytical strategy it is fundamental to explore your data (data in the social sciences are **messy**).

- The importance of **complementing**: especially in long projects (e.g., PhD theses), combining different approaches is critical ⇒ forecasting + inference.

- Never, never, **never follow the hype**: do not choose ML just because it sounds cool.

Further readings

Books

- James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.
- Hastie, Trevor, et al. The elements of statistical learning: data mining, inference, and prediction. Vol. 2. New York: springer, 2009.

Articles

- Wagner, Claudia, et al. "Measuring algorithmically infused societies." Nature 595.7866 (2021): 197-204.
- Lazer, David, et al. "Meaningful measures of human society in the twenty-first century." Nature 595.7866 (2021): 189-196.

Artwork reference

- E. Hopper, *Lighthouse hill* (1927)
- E. Hopper, *Gas* (1940)
- E. Hopper, *Nighthawks* (1942)
- Midjourney, *“A boy and a robot that talk, in the desert, Hopper style painting”* (2024).