Explainable & Interpretable Data Science

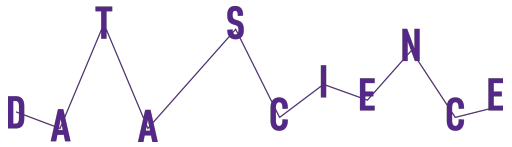
Jordi Vitrià

Feb, 2019

UNIVERSITAT DE BARCELONA

jordi.vitria@ub.edu

Departament de Matemàtiques i Informàtica

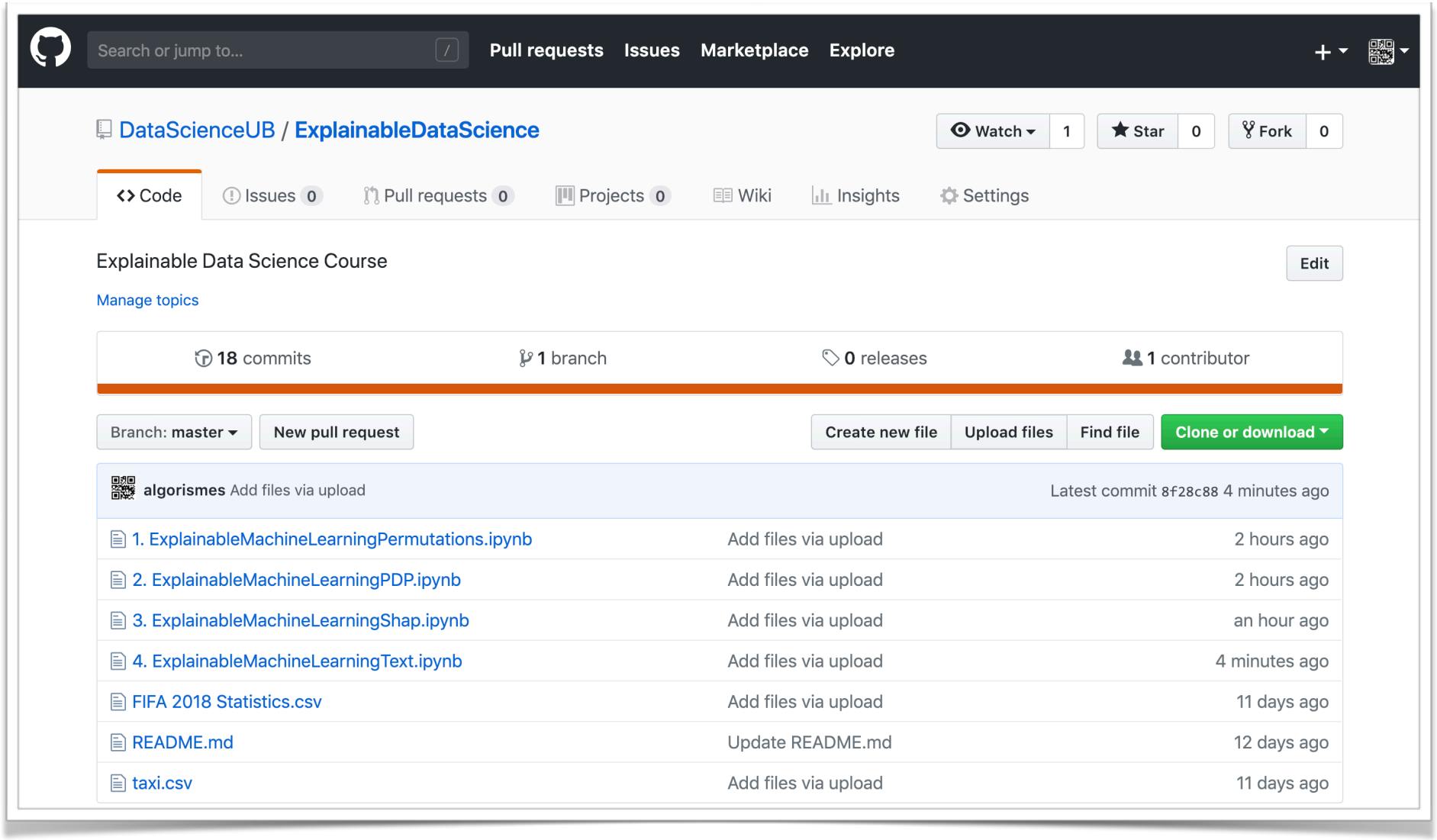


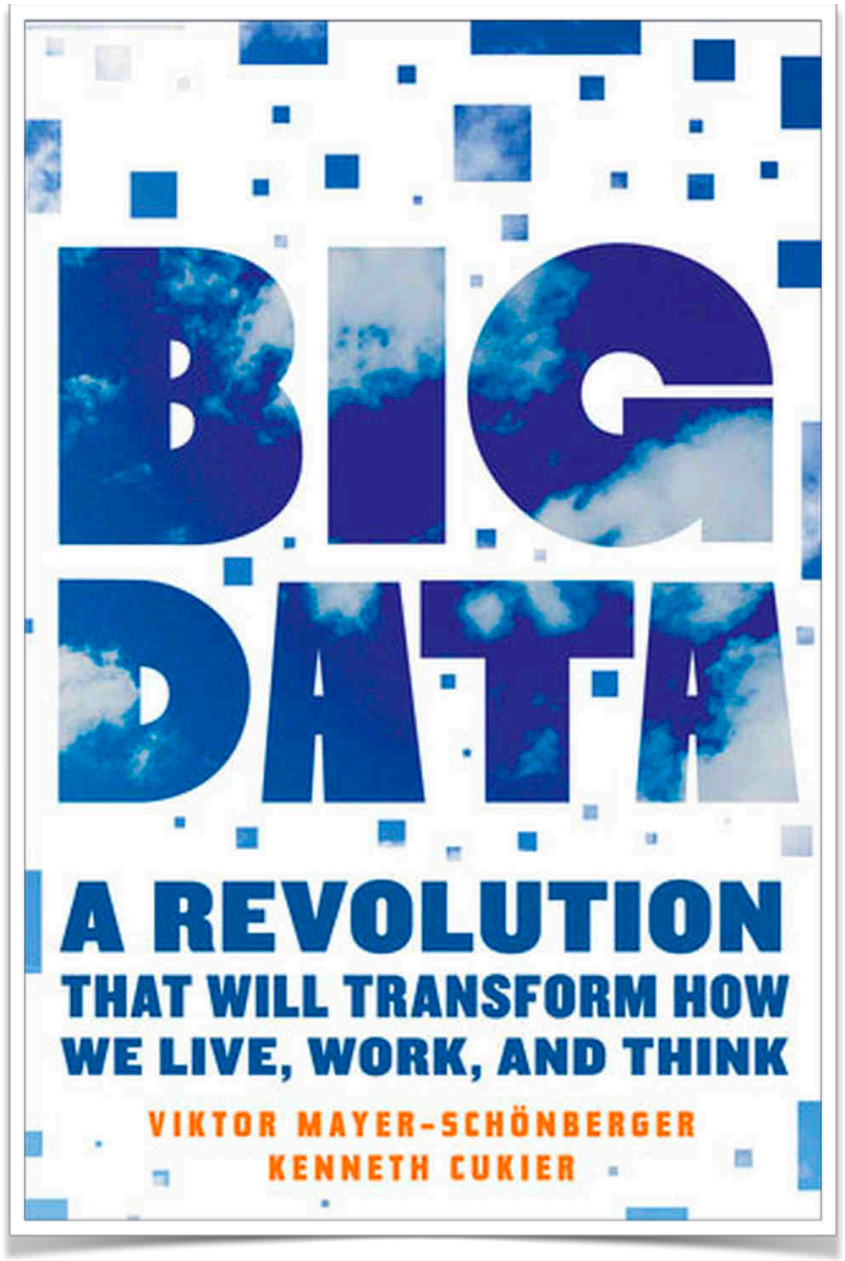
Since 2007, I am a Full Professor at the Mathematics & Computer Science Department, **Universitat de Barcelona**. Before that I spent 20 years on the faculty of the CS Department at the **Universitat Autonoma de de Barcelona**. I am the Director of the **Master in Fundamental Principles of Data Science** at UB. I am the leader of the **DataScience@UB** group, whose objective is to promote research & technology transfer in the areas of data analytics, machine learning and AI.

My research statement: To understand the fundamental processes underlying the visual perception of objects to derive new computer vision algorithms that, one day, may enable machines to see.

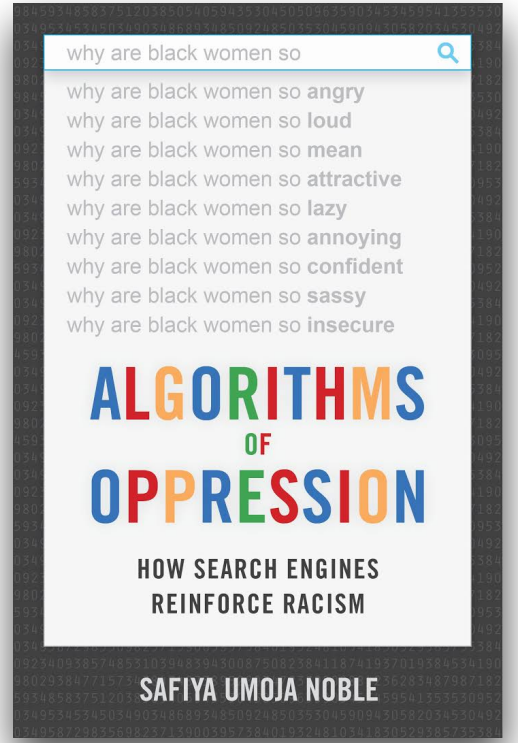
http://datascience.barcelona/
http://www.ub.edu/cvub/jordivitria/

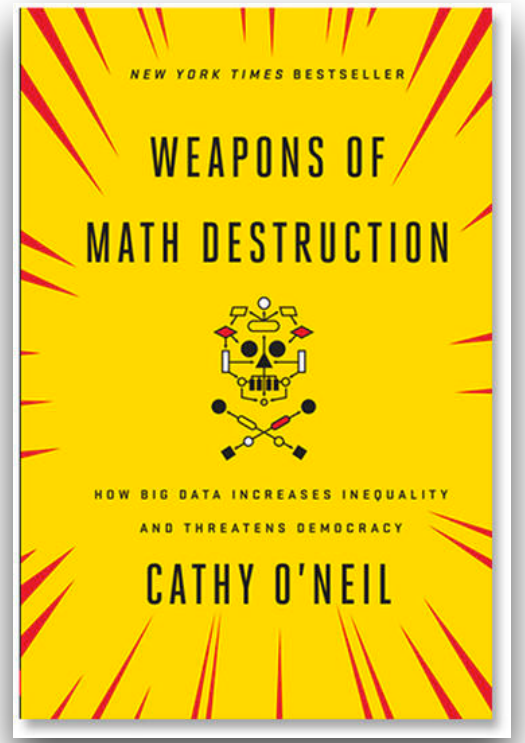
https://github.com/DataScienceUB/ExplainableDataScience

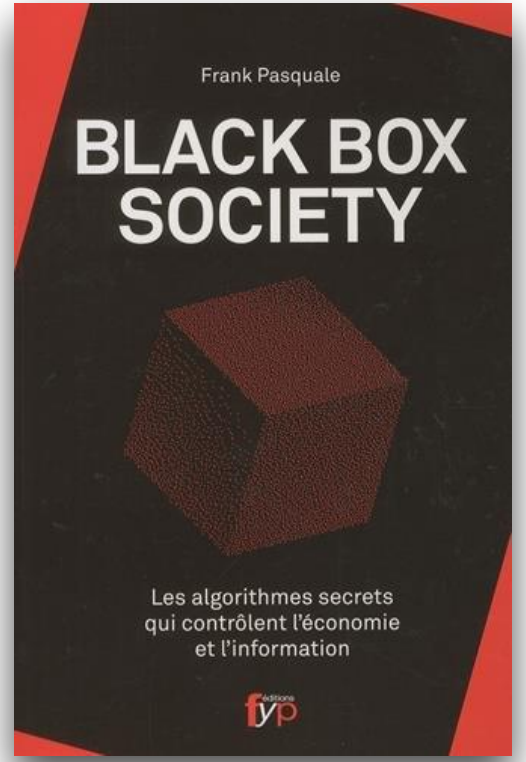



BIG
DATA
A REVOLUTION
THAT WILL TRANSFORM HOW
WE LIVE, WORK, AND THINK
VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER

March 4, 2014


why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS
OF
OPPRESSION
HOW SEARCH ENGINES
REINFORCE RACISM
SAFIYA UMOJA NOBLE


NEW YORK TIMES BESTSELLER
WEAPONS OF
MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY
AND THREATENS DEMOCRACY
CATHY O'NEIL


Frank Pasquale
BLACK BOX
SOCIETY
Les algorithmes secrets
qui contrôlent l'économie
et l'information

There was once a time when humans made important decisions about other humans.

You went to your bank manager, in person, to ask for a loan.
A human hiring committee decided which candidate would get a job.
And judges even determined what a guilty offender's sentence would be!

Now, many of those decisions are made by AIs.

There are a couple of problems with AIs making these decisions. First of all, the algorithms the machines use are often **proprietary**, meaning they aren't available to the general public or other computer scientists to examine. Second, and perhaps more troubling, AIs often interpret data in ways more **complex** than even their programmers can understand. In those cases, nobody could explain how the "black box" in an AI works, even if they wanted to.

CBC Radio · January 19

https://www.cbc.ca/radio/spark/422-1.4982026/asking-why-instead-of-how-could-better-explain-ai-decisions-1.4982038

Derrière l'algorithme de Parcoursup, un choix idéologique

La répartition des étudiants entre les universités et les filières est un problème complexe puisqu'elle s'effectue sur base d'un conflit massif entre l'offre et la demande : on dénombre plus de 880.000 candidats pour un total (à raison de 10 vœux possibles par candidat) de quelques 7.000.000 de vœux de formation *[810.000 ont in fine validé leurs voeux, NDLR]*. La résolution d'un tel conflit n'est plus sérieusement envisageable humainement. Dès lors qu'un algorithme travaille à cette mise en relation n'est pas à remettre en question. La vraie question est celle de l'objectif assigné à l'algorithme et des choix qu'il doit exécuter.

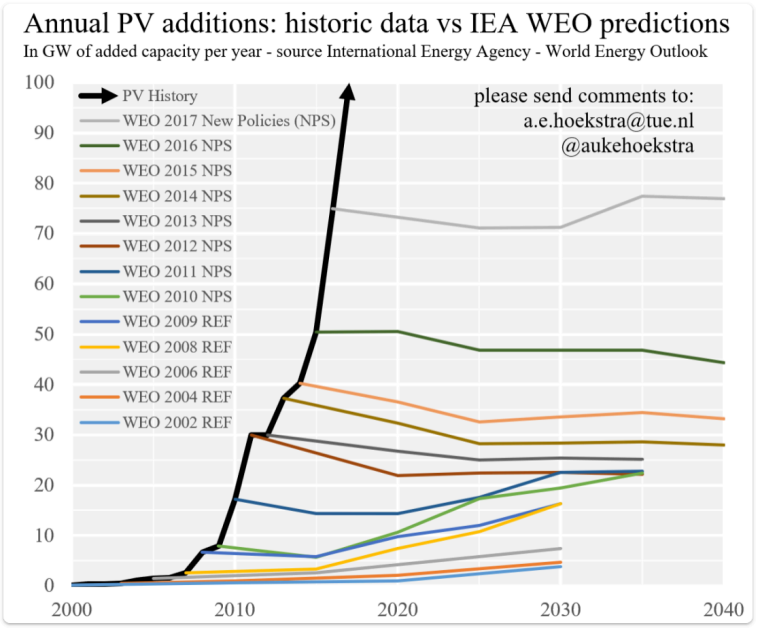
Cette décision politique et idéologique se lit dans la formule algorithmique même de Parcoursup. Cet algorithme, dont l'objectif est de mettre en relation deux objets, d'un côté des établissements, de l'autre des étudiants, est en effet inspiré par le célèbre algorithme de Gale et **Shapley**, repris par Alvin Roth, prix Nobel d'économie en 2012. Il relève au fond d'un vieux problème économique que l'on appelle l'appariement stable.

Photovoltaic growth: reality versus projections of the International Energy Agency – the 2017 update (by Auke Hoekstra)

This entry was posted on June 12, 2017 by Auke Hoekstra, in Uncategorized. Bookmark the permalink. 16 Comments

Update for 2017: the IEA is once again predicting the solar industry will stop growing. As you can see in the updated graph, yearly additions are still increasing rapidly but again the prediction of the IEA is flat. Fortunately many sources are noticing this or using "my" method for showing how far the IEA is off the mark.
Examples here, here, here, here, here, here, here, here, here, here and here. I hope the criticism will grow exponentially until the IEA learns.

This is a guest blog by Auke Hoekstra, senior adviser electric mobility at the Eindhoven University of Technology and developer of agent-based models for electric vehicles and renewable energy adoption. You can contact him at a.e.hoekstra@tue.nl or @aukehoekstra.



United Nations Univers
Centre for Policy Research

Dr Joanna Bryson

AI & Global Governance: No One Should Trust AI

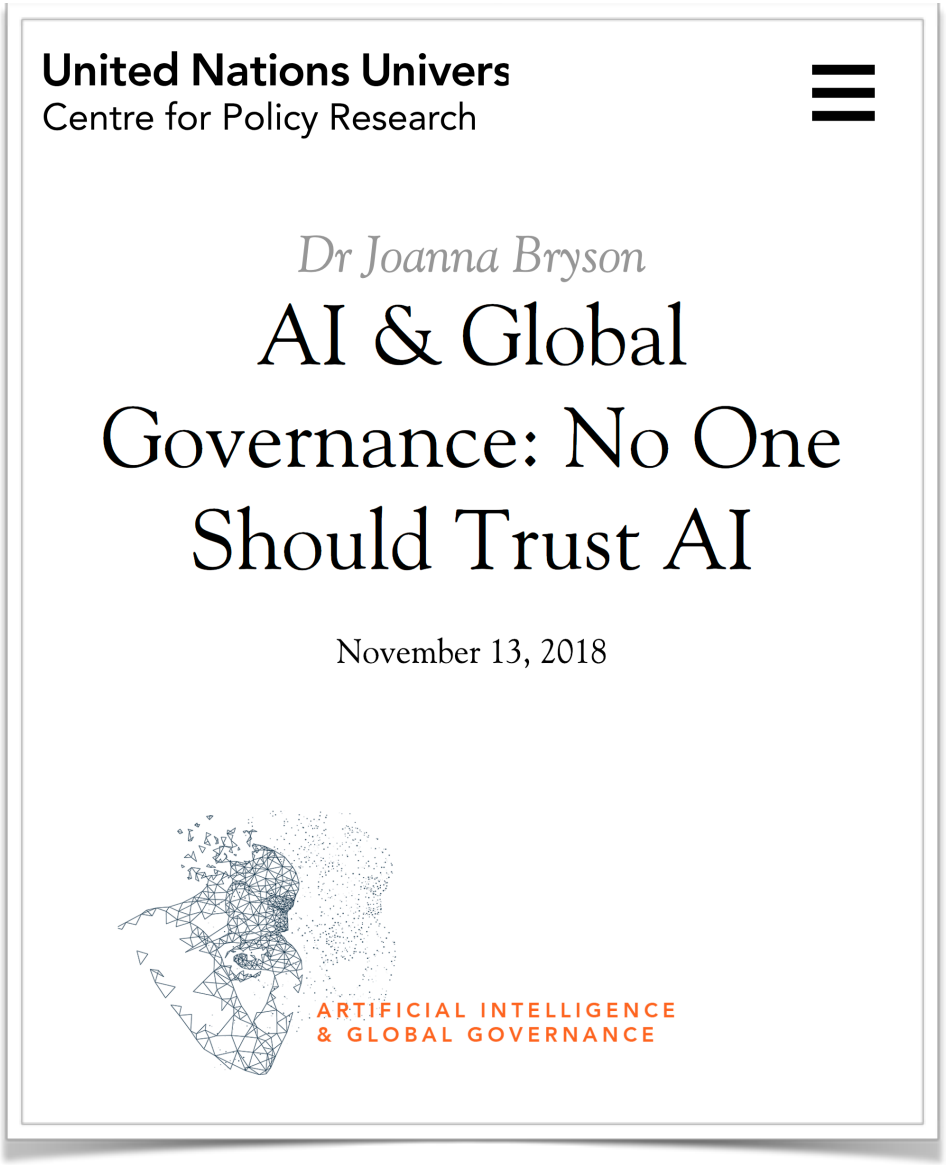
November 13, 2018

ARTIFICIAL INTELLIGENCE
& GLOBAL GOVERNANCE

Trust is a relationship between **peers** in which the trusting party, while not knowing for certain what the trusted party will do, believes any promises being made.

AI is a set of system development techniques that allow machines to compute actions or knowledge from a set of data. Only other software development techniques can be peers with AI, and since these do not "trust", no one actually *can* trust AI.

More importantly, no human should *need* to trust an AI system, because it is both possible and desirable to engineer AI for accountability. We do not need to trust an AI system, we can know how likely it is to perform the task assigned, and only that task. When a system using AI causes damage, we need to know we can hold the human beings behind that system to account.



Organizations that develop software need to be able to demonstrate due diligence in the creation of that software, including using appropriate standards for logging: what code is written (as well as when, why and by whom), which software and data libraries are used, and what hardware is used during system development.

Organizations need to undertake and document appropriate testing before the software's release and perform monitoring and maintenance while the code is in use. Like other sectors, they should be held liable unless they can prove such due diligence. These procedures lead to a safer and more stable software systems, intelligent or not.

Scenario

Data science tasks:

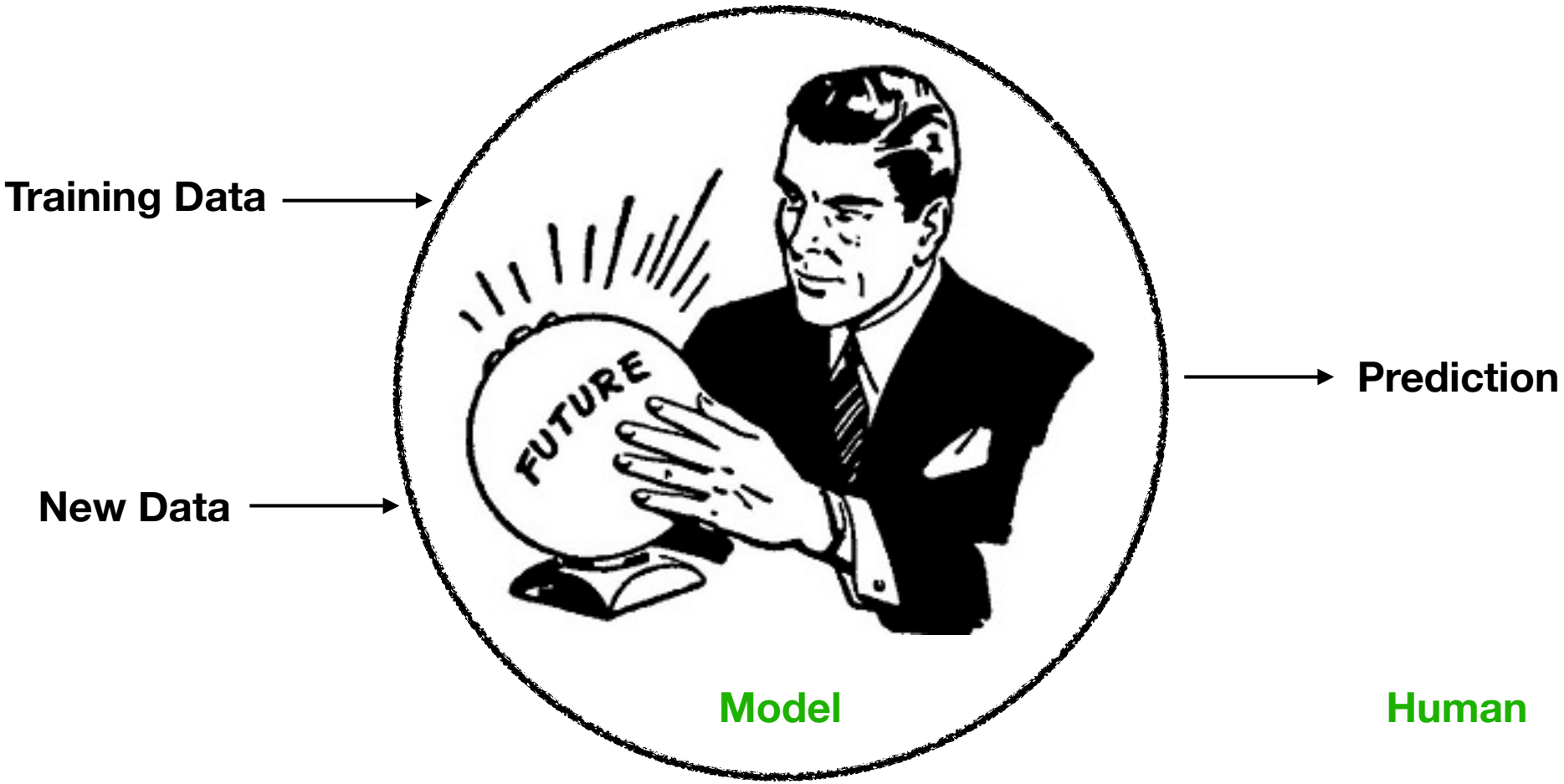
- **Description** is using data to provide a quantitative summary of certain features of the world. → What is the mean value of X?

- **Prediction** (or **association**) is using data to map some features of the world (the inputs) to other features of the world (the outputs). → How would seeing X change my belief in Y?

- **Causation**: Measuring the causal influence of a variable X in another variable Y, while excluding any influences on Y not actually due to the causal effect of X, and being able to guess what the effect will be if one performs an action. → How would my expected lifespan change if I become a vegetarian?

- **Counterfactuals**: Being able to reason about hypothetical situations, things that *could* happen. → Would my grandfather still be alive if he did not smoke?

Scenario



Why?

Explainability

Understanding reasoning behind each decision. Assuring that our knowledge is reflected in the model.

Transparency

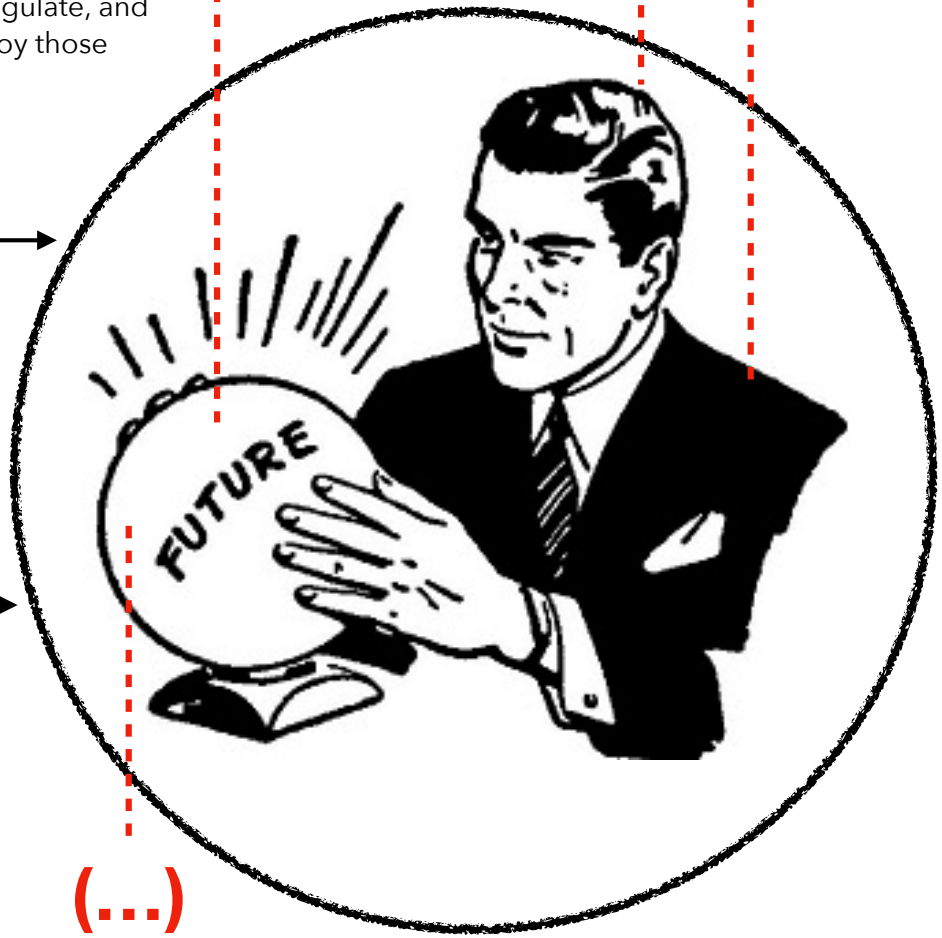
The factors/values that influence the decisions made by algorithms should be visible, or transparent, to the people who use, regulate, and are impacted by systems that employ those algorithms

Uncertainty

What is the certainty behind decisions?

Training Data →

New Data →



→ **Decision**

(...)

Why

There are different reasons that drive the demand for interpretability and explanations:

- **Human curiosity and learning**: Humans have a mental model of their environment that is updated when something unexpected happens. This update is performed by finding an explanation for the unexpected event.

- The goal of science is to **gain knowledge**, but many problems are solved with big datasets and black box machine learning models. The model itself becomes the source of knowledge instead of the data. Interpretability makes it possible to extract this additional knowledge captured by the model.

https://christophm.github.io/interpretable-ml-book/interpretability-importance.html

Why

- Machine learning models take on real-world tasks that require **safety measures** and testing. Imagine a self-driving car automatically detects cyclists based on a deep learning system. You want to be 100% sure that the abstraction the system has learned is error-free, because running over cyclists is quite bad.

- By default, machine learning models pick up biases from the training data. This can turn your machine learning models into racists that discriminate against protected groups. Interpretability is a useful debugging tool for **detecting bias** in machine learning models. It might happen that the machine learning model you have trained for automatic approval or rejection of credit applications discriminates against a minority.

- The process of integrating machines and algorithms into our daily lives requires interpretability to increase **social acceptance**. People attribute beliefs, desires, intentions and so on to objects.

- Machine learning models can only be **debugged and audited** when they can be interpreted.

https://christophm.github.io/interpretable-ml-book/interpretability-importance.html

Why

- If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily:
  - **Fairness**: Ensuring that predictions are unbiased and do not implicitly or explicitly discriminate against protected groups. An interpretable model can tell you why it has decided that a certain person should not get a loan, and it becomes easier for a human to judge whether the decision is based on a learned demographic (e.g. racial) bias.
  - **Privacy**: Ensuring that sensitive information in the data is protected.
  - **Reliability** or Robustness: Ensuring that small changes in the input do not lead to large changes in the prediction.
  - **Causality**: Check that only causal relationships are picked up.
  - **Trust**: It is easier for humans to trust a system that explains its decisions compared to a black box.

https://christophm.github.io/interpretable-ml-book/interpretability-importance.html

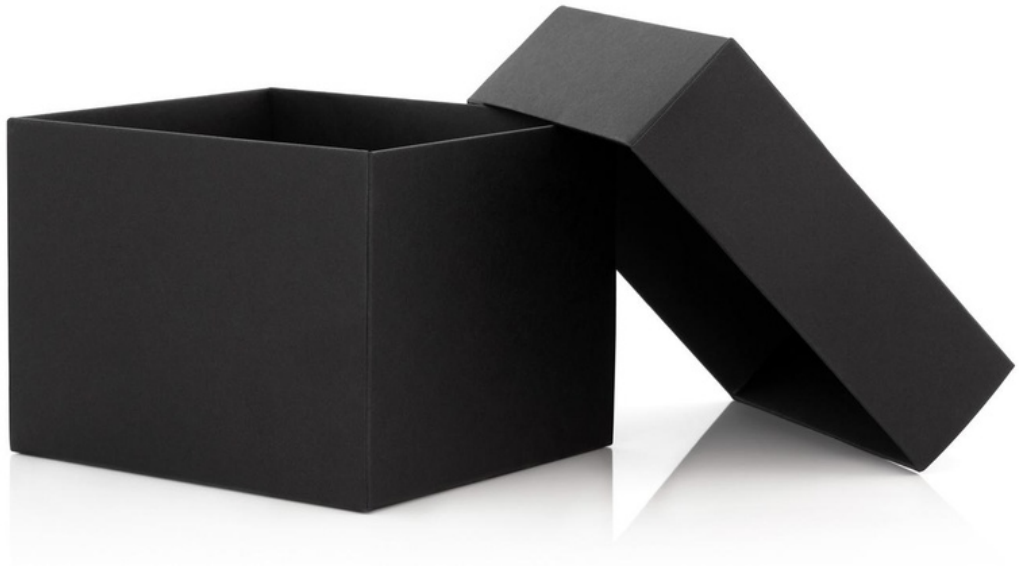
Why

We don’t need interpretability if the model has no significant impact, the problem is well studied (f.e. OCR) or if this enable people to manipulate or to game a critical system.

https://christophm.github.io/interpretable-ml-book/interpretability-importance.html

Our definitions

Explainable DS – using (or considering our models as) a black box and explaining it afterwards.



Our definitions

By black-box, we denote a predictive model which:

- has been pre-trained,
- exposes a function (typically called `predict`, to estimate the output value given a set of input values) and
- whose internal details, such as algorithm of choice, or dataset used for training, are not accessible or not even known.

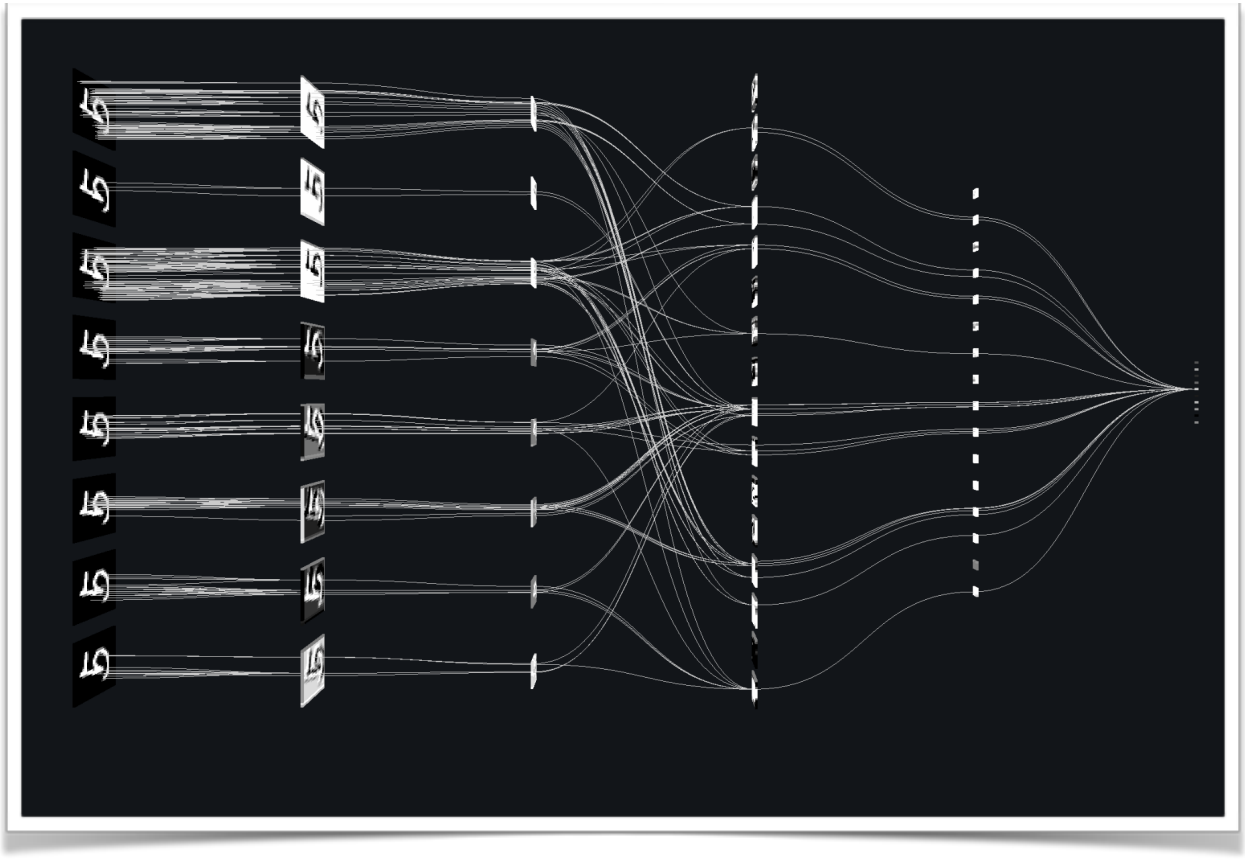
Our definitions

Sometimes, back boxes are unavoidable:

- predictive models that are available “off-the-shelf” as part of code libraries,
- machine learning servers,
- cloud-based services or
- inside domain-specific black-box software.

Our definition

Interpretable DS – using a model that is not black box.



Scope

We can look for answers to different questions:

- How does the trained model make predictions?
- How do parts of the model affect predictions?
- Why did the model make a certain prediction for an instance?
- Why did the model make specific predictions for a group of instances?

Evaluation

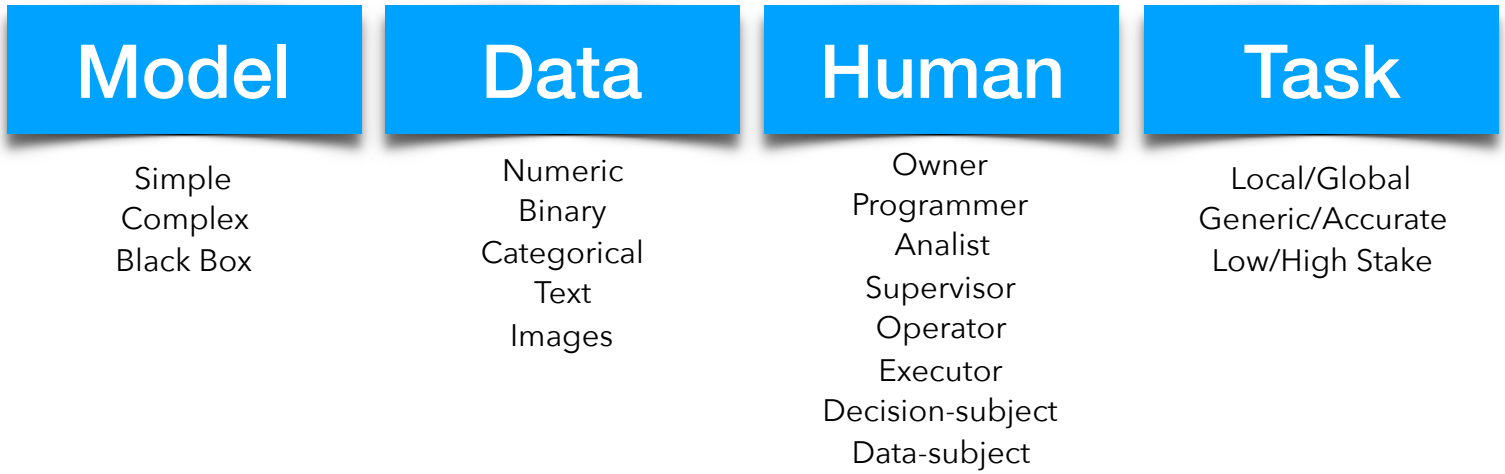
Doshi-Velez and Kim (2017) propose three main levels for the evaluation of interpretability:

- **Application level evaluation (real task)**: Put the explanation into the product and have it tested by the end user. A good baseline for this is always how good a human would be at explaining the same decision.
- **Human level evaluation (simple task)** is a simplified application level evaluation. The difference is that these experiments are not carried out with the domain experts, but with laypersons. An example would be to show a user different explanations and the user would choose the best one.
- **Function level evaluation (proxy task)** does not require humans. For example, it might be known that the end users understand decision trees. In this case, a proxy for explanation quality may be the depth of the tree. Shorter trees would get a better explainability score.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards A Rigorous Science of Interpretable Machine Learning," no. MI: 1–13. http://arxiv.org/abs/1702.08608.

Ingredients

The explanation depends on several factors.
There is **no best explanation**.

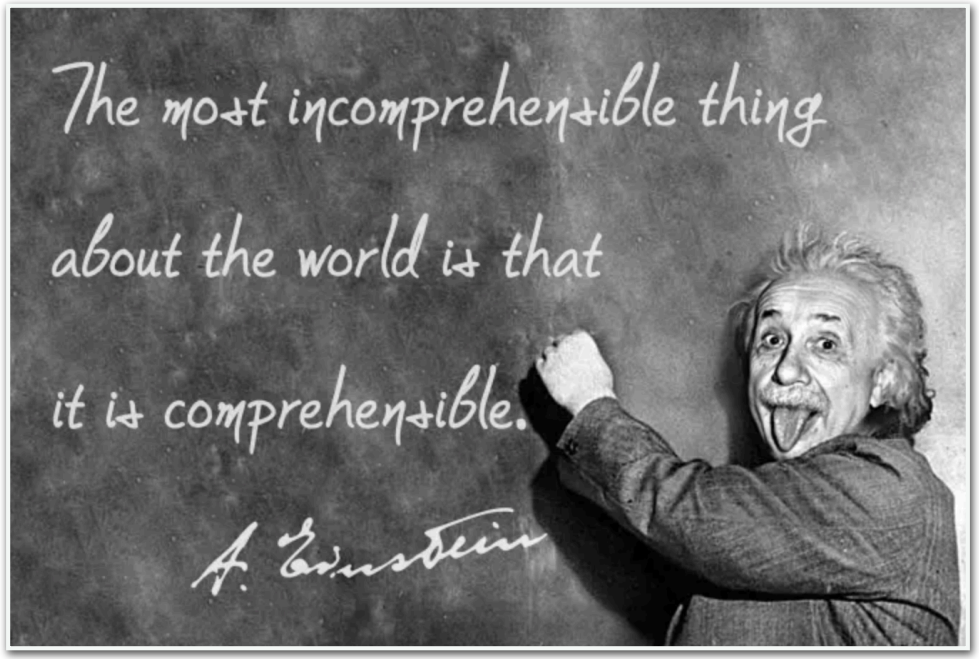


What is an explanation?

R.Feynman

https://www.youtube.com/watch?v=Q1IL-hXO27Q

What is an explanation?



An explanation is the **answer to a why-question** (Miller 2017).

- Why did not the treatment work on the patient?
- Why was my loan rejected?

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269.

What is an explanation?

A good explanation is:

- **Contrastive**. Humans usually do not ask why a certain prediction was made, but **why this prediction was made *instead of another prediction***. The solution for the automated creation of contrastive explanations might also involve finding prototypes or archetypes in the data.

- **Selected**. People do not expect explanations that cover the actual and complete list of causes of an event. We are used to selecting **one or two causes** from a variety of possible causes as **THE** explanation.

- **Social**. The social context determines the content and nature of the explanations. Getting the social part of the machine learning model right depends entirely on your specific application.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269.

What is an explanation?

A good explanation is:

- **Focused on the abnormal**. People focus more on causes that had a small probability but nevertheless happened.
- **Thruthful**. The explanation should predict the event as truthfully as possible, which in machine learning is sometimes called **fidelity**.
- **Consistent** with prior beliefs of the explainee. This is difficult to integrate into machine learning!
- **General and probable**. A cause that can explain many events is very general and could be considered a good explanation. Generality can easily be measured by the feature's support, which is the number of instances to which the explanation applies divided by the total number of instances.

Miller, Tim. 2017. "Explanation in Artificial Intelligence: Insights from the Social Sciences." arXiv Preprint arXiv:1706.07269.

It is a difficult problem!

The New York Times

YOUR MONEY ADVISER

In California, Gender Can No Longer Be Considered in Setting Car Insurance Rates

By Ann Carrns

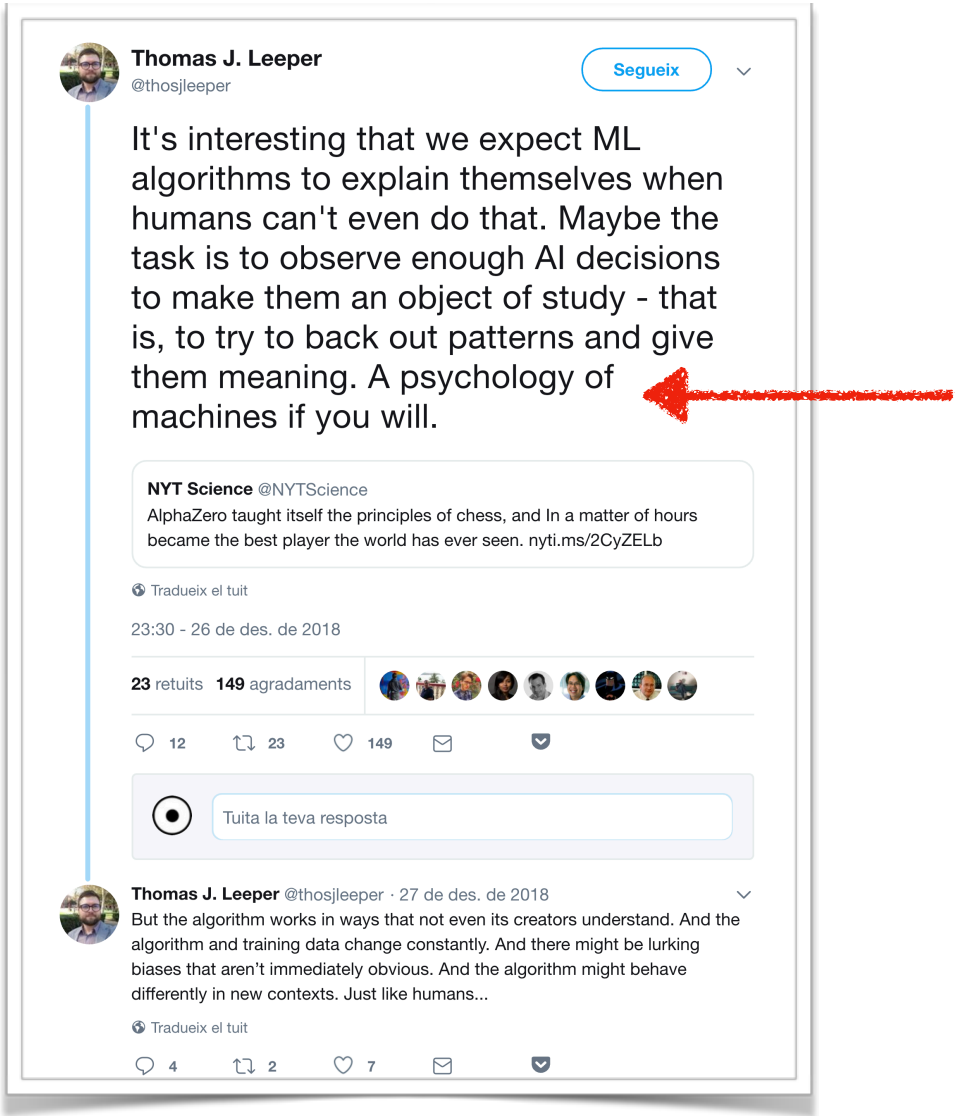
Jan. 18, 2019

Minh Uong/The New York Times

California joined about a half-dozen states this month in banning the use of a person's gender when assessing risk factors for car insurance, a change that could potentially alter rates for scores of drivers across the state.


Delip Rao @deliprao · 18 h
Not only does the gender variable has to be decorrelated with each of the other other individual factors, but also has to be decorrelated to any arbitrary function of any arbitrary subset of the other facts. For anyone wondering, this perfect decorrelation is next to impossible.
Tradueix el tuit
1
4
Delip Rao @deliprao · 18 h
Why caution must be exercised in understanding/accepting such things, or we are simply living in an illusion of progress as opposed to actual progress.
Tradueix el tuit
1
1
7
Delip Rao @deliprao · 17 h
So, if this perfect decorrelation is impossible, what is one to do? 1) Always be vigilant, and 2) Always measure disparate outcomes wrt to the sensitive variables. Remember that justice should not be blind, and that statistical fairness should be the aim than absolute fairness.
Tradueix el tuit
3
6

Skeptical Point of View



Thomas J. Leeper
@thosjleeper

Segueix

It's interesting that we expect ML algorithms to explain themselves when humans can't even do that. Maybe the task is to observe enough AI decisions to make them an object of study - that is, to try to back out patterns and give them meaning. A psychology of machines if you will.

NYT Science @NYTScience
AlphaZero taught itself the principles of chess, and In a matter of hours became the best player the world has ever seen. nyti.ms/2CyZELb

Tradueix el tuit

23:30 - 26 de des. de 2018

23 retuits **149** agradaments

12 23 149

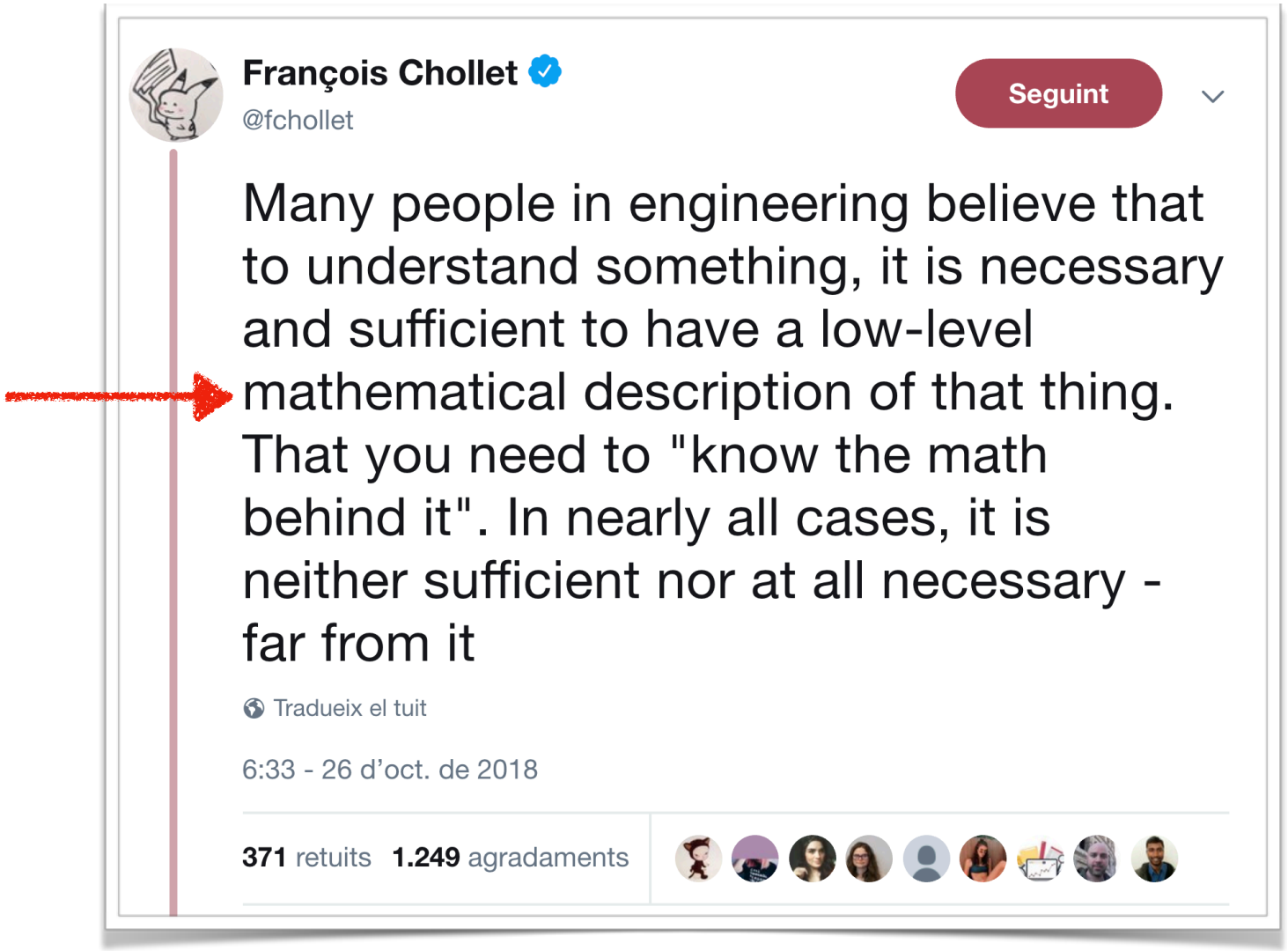
Tuita la teva resposta

Thomas J. Leeper @thosjleeper · 27 de des. de 2018
But the algorithm works in ways that not even its creators understand. And the algorithm and training data change constantly. And there might be lurking biases that aren't immediately obvious. And the algorithm might behave differently in new contexts. Just like humans...

Tradueix el tuit

4 2 7

Skeptical Point of View



François Chollet
@fchollet

Seguint

Many people in engineering believe that to understand something, it is necessary and sufficient to have a low-level mathematical description of that thing. That you need to "know the math behind it". In nearly all cases, it is neither sufficient nor at all necessary - far from it

Tradueix el tuit

6:33 - 26 d'oct. de 2018

371 retuits **1.249** agradaments

Simpson’s Paradox

Suppose you're suffering from kidney stones and go to see your doctor. The doctor tells you two treatments are available, treatment A and treatment B. You ask which treatment works better, and the doctor says:

"*Well, a study found that treatment* ***A has a higher probability of success than treatment B***."

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson's Paradox

You start to say "*I'll take treatment A, thanks!*", when your doctor interrupts:
"*But the same study also looked to see which treatment worked better, depending on whether patients had large kidney stones or small kidney stones.*"

You say "*Well, do I have large kidney stones or small kidney stones*"? As you speak the doctor interrupts again, looking sheepish, and says:
"*Actually, it doesn't matter. You see, they found that treatment B has a higher probability of success than treatment A, regardless of whether you have large or small kidney stones.*"

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson's Paradox

Example:

It sounds impossible. But it's true: an actual study was done in which treatment TB was found to work with higher probability than treatment TA, for both large and small kidney stones, despite the fact that treatment TA works with higher overall probability than Treatment TB.

Here's the numbers from the study:

C. R. Charig, D. R. Webb, S. R. Payne, O. E. Wickham (March 1986)	**Treatment TA helps**	**Treatment TB helps**
Large kidney stones	69% (55 / 80)	**73%** (192 / 263)
Small kidney stones	87% (234 / 270)	**93%** (81 / 87)
All patients	**83%** (289 / 350)	78% (273 / 350)

http://michaelnielsen.org/reinventing_explanation/index.html

Simpson’s Paradox

The practical significance of Simpson's paradox surfaces in **decision making situations** where it poses the following dilemma:

Which data should we consult in choosing an action, the aggregated or the partitioned?

Simpson’s Paradox

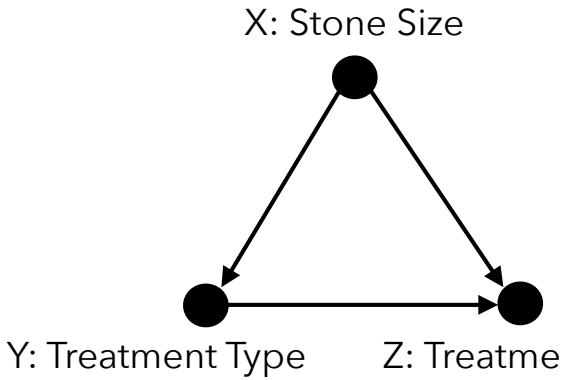
In the Kidney Stone example above, it is clear that if one is diagnosed with "Small Stones" or "Large Stones" the data for the respective subpopulation should be consulted and Treatment TB would be preferred to Treatment TA.

But what if a patient is not diagnosed, and the size of the stone is not known; would it be appropriate to consult the aggregated data and administer Treatment TA?

This would stand contrary to common sense; a treatment that is preferred both under one condition and under its negation should also be preferred when the condition is unknown.

Solving Simpson's Paradox by using Causal Reasoning

In our problem, **kidney stone size** X and **treatment type** Y are both causes of success Z. X may be a cause of Y if other doctors are assigning treatment based on kidney stone size. Clearly there are no other causal relationships between X, Y, and Z. Y comes after X so it cannot be its cause. Similarly Z comes after X and Y. **Since X is a common cause (confounding variable), it should be measured**.



Causal Diagram

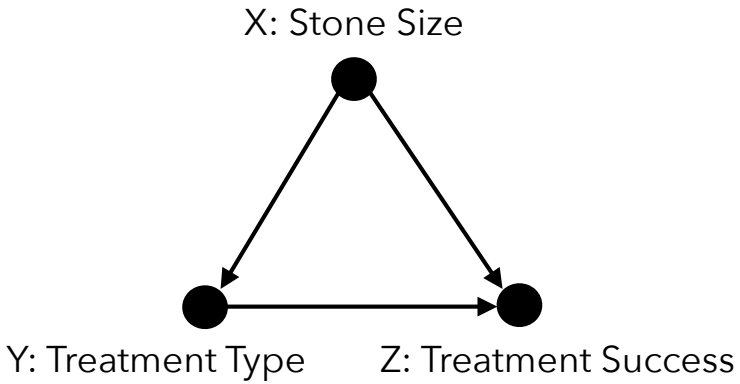
Solving Simpson's Paradox by using Causal Reasoning

X (stone size) is confounder of Y (treatment type) and Z (treatment success). For an unbiased estimate of of the effect of Y on Z we must adjust the confounder.

We can do it by looking at the data for X separately, then taking the (weighted) average:

$$\Pr(z \mid do(x)) = \sum_{y} \Pr(y) \Pr(z \mid x, y)$$

Solving Simpson's Paradox by using Causal Reasoning



	Treatment TA helps	Treatment TB helps
Large kidney stones	69% (55 / 80)	**73%** (192 / 263)
Small kidney stones	87% (234 / 270)	**93%** (81 / 87)
Average	78%	**83%**