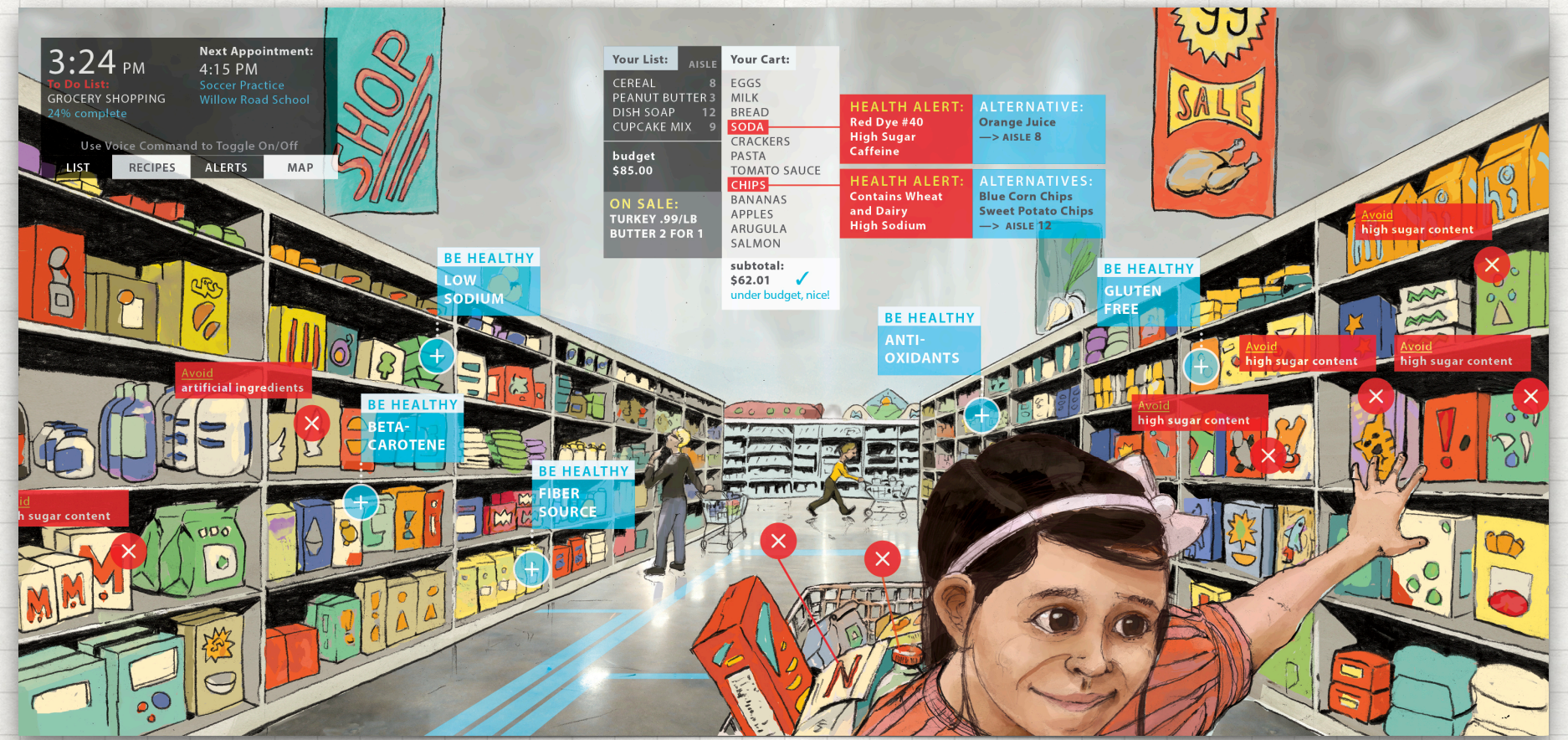
DADES I SOCIETAT

TALLER DE NOUS USOS DE LA INFORMÀTICA

UNIVERSITAT DE BARCELONA



JORDI VITRIÀ

MENTIDES: TIPUS I

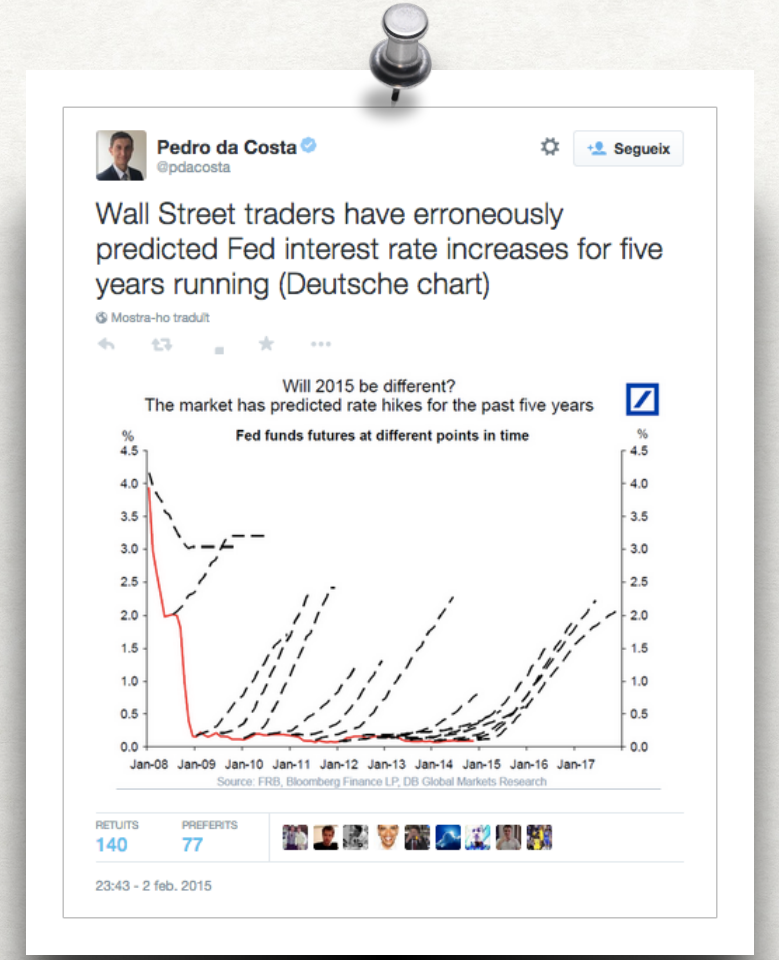
To **intentionally** develop models that don't work because it's possible to get more benefit (f.e. make more money, get more power, etc.) from a bad model than a good one.

MENTIDES: TIPUS I

MENTIDES: TIPUS I

MENTIDES: TIPUS I

MENTIDES: TIPUS II



Unintentional Lies

Something went wrong: the model, the data, the question, …

DEFINICIONS

Q: What is data science?

A: The process of finding supported answers about a special kind of reality (people, artifacts, processes, businesses, etc.).

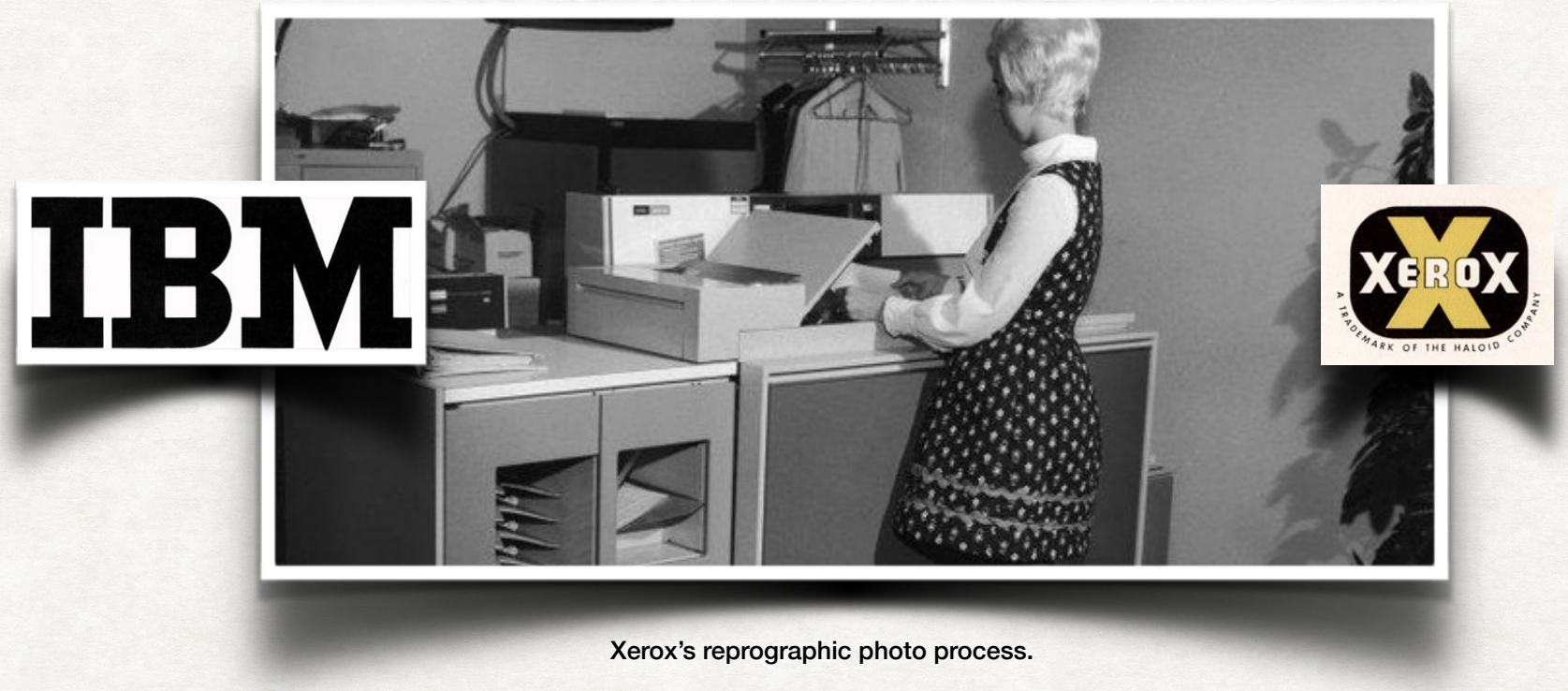
We can think of data science as a variant of the scientific method.

DEFINICIONS

Good Data Scientist =
Good Toolset + Good Dataset + **Good Skillset + Good Mindset**.

ERRORS

“Bad Questions, Bad Decisions” or “Are you solving the right problem?”



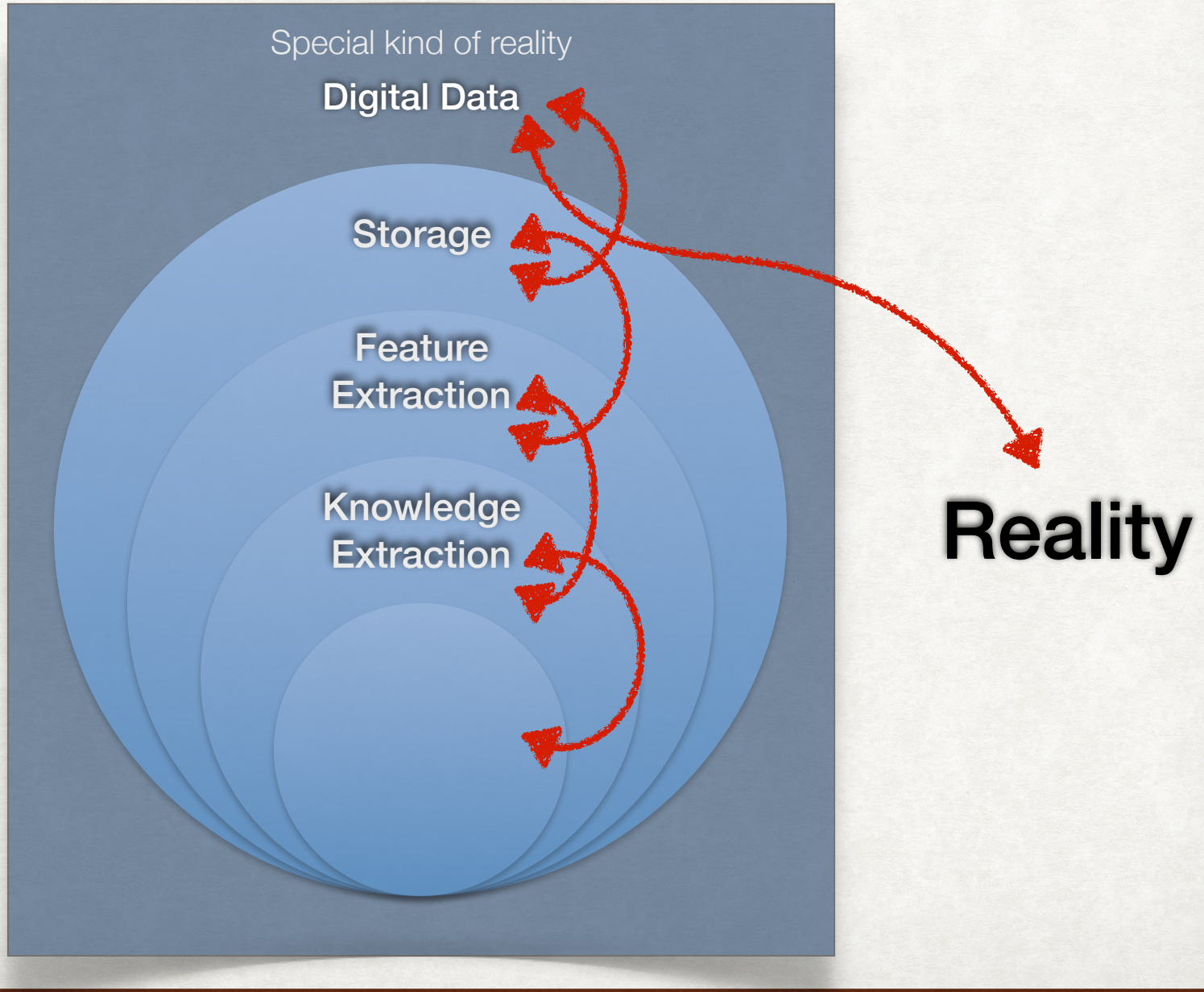
Xerox’s reprographic photo process.

Q: If a more reliable, cheaper and faster process for photocopying were available, how many more copies would people make in a given year?

The problem was framed (by both companies) as “copies from original”, ignoring a larger segment of the market: “copies of copies of copies”…

ERRORS

From questions to reality: the data science path



ERRORS

SKEPTICISM is a necessary (and heavy) element in our knapsack during the journey along the data science path.

ERRORS

The true **meaning** of the word **skepticism** has nothing to do with doubt, disbelief, or negativity.

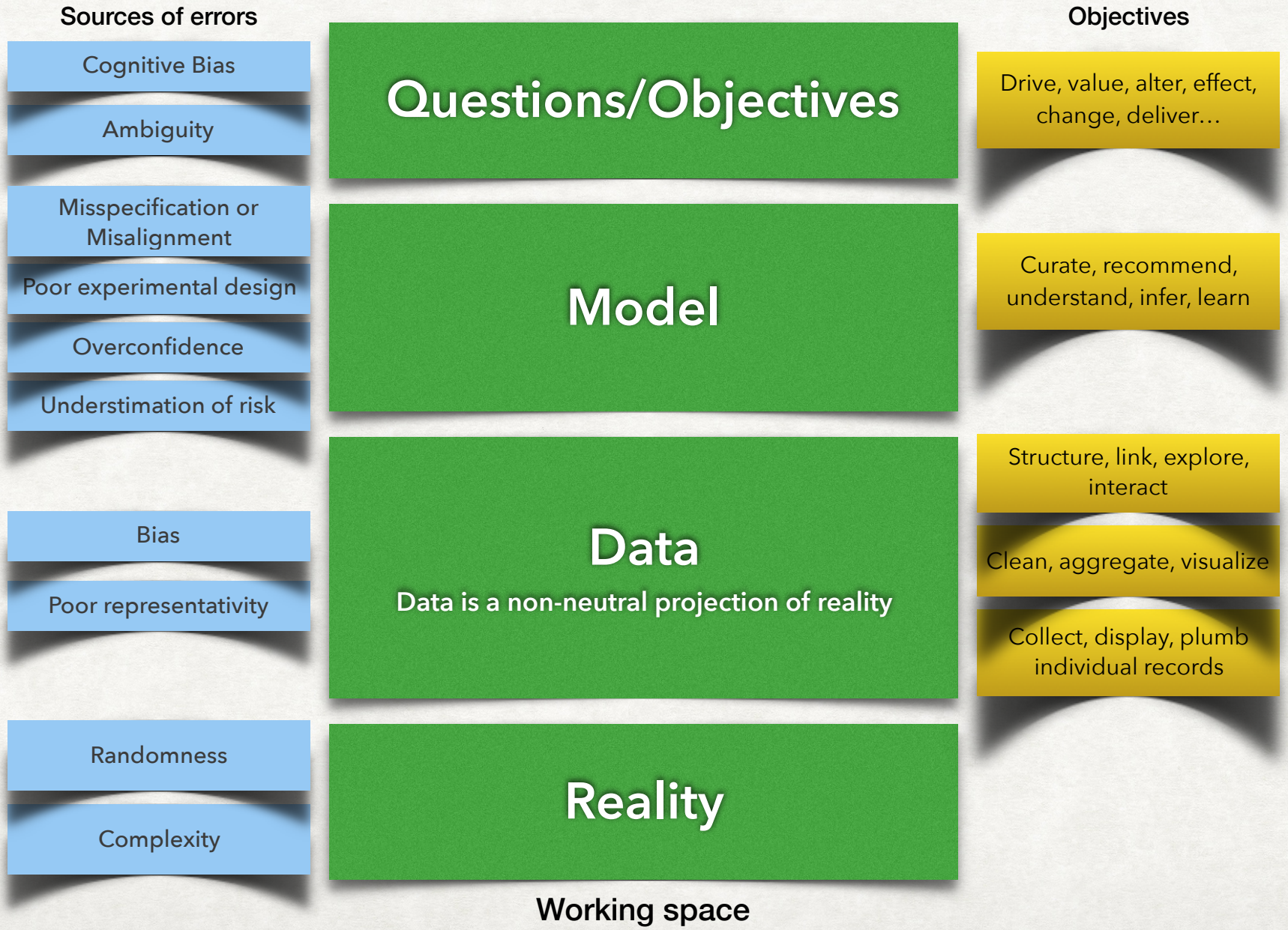
Skepticism is **the process of applying reason and critical thinking to determine validity**.

Brian Dunning

It’s easy to confuse being a skeptic with being a cynic.

ERRORS

Data science must be skeptical to be free of errors



ÈTICA

(Big) data science must be skeptical to be ethical

Data **ethics** is a set of related principles that should govern data flows in our information society, and inform the establishment of big data norms.

The big issue of big data is NOT SIZE, is GRANULARITY

Hanna Wallach

Ethical issues

Data **provenance** documents the inputs, entities, systems, and processes that influence data of interest, in effect providing a historical record of the data and its origins.

Provenance

Ensuring **privacy** of data is a matter of defining and enforcing information rules – not just rules about data collection, but about data use and retention. People should have the ability to manage the flow of their private information across massive, third-party analytical systems.

Privacy

Bias

Fairness

Inclusion means that the benefits of data analysis are accessible to nearly everyone with the right tools and connection, and can benefit everyone if enough data's available.

Inclusion

Data governance is the formal execution and enforcement of authority over the management of data and data related assets.

Governance

Solution

Putting the right question.
Answering the question by following an skeptical methodology.

The main source of Big Data is information about individuals people and their activities

Data Science is not fair or just in any meaningful way

Algorithms can be unfair

Use of convenience data

It is easy to pick up coarse-grained signals from noise, but what about fine-grained ones?

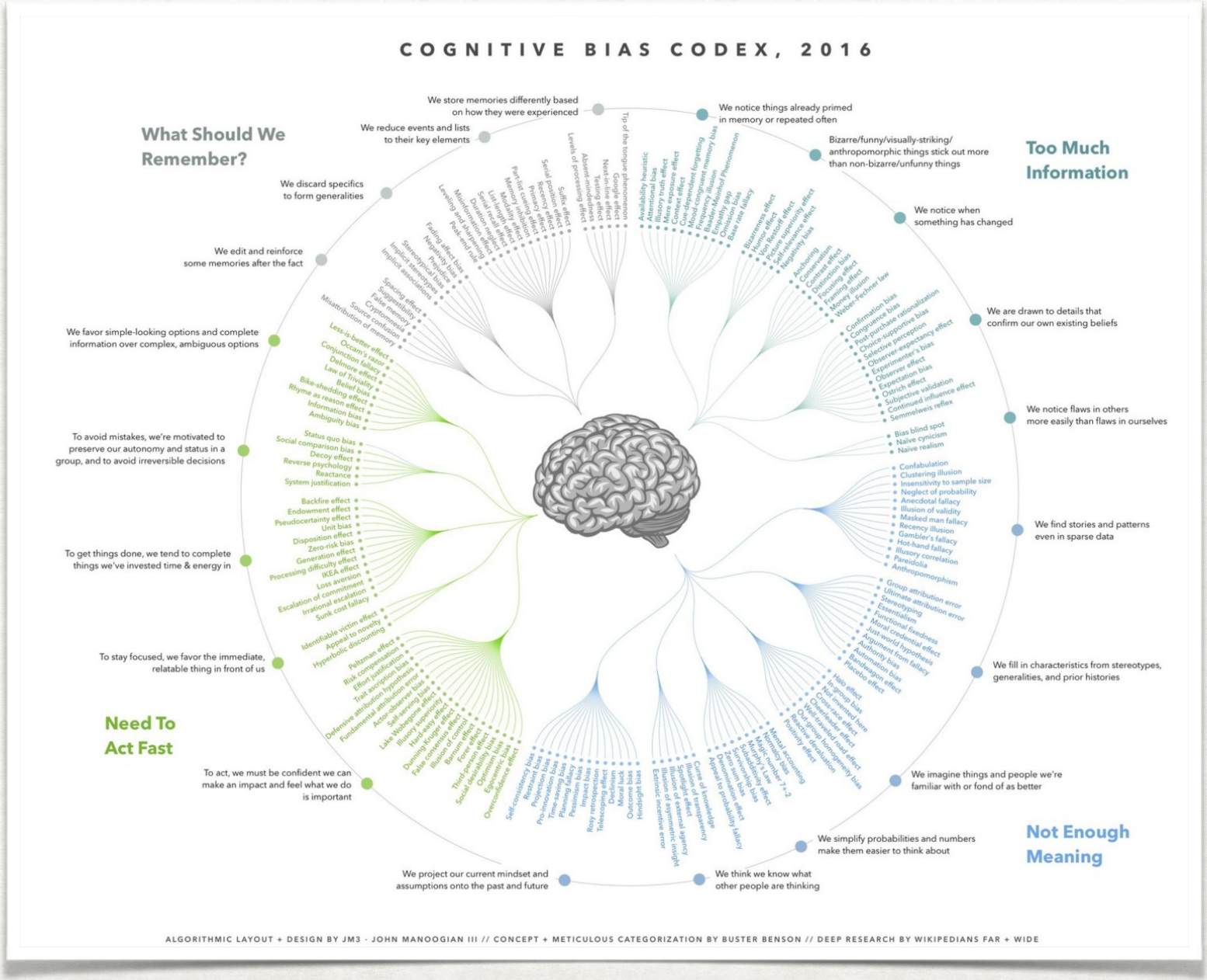
What is an accurate model?

We must give priority to question-driven Data Science.

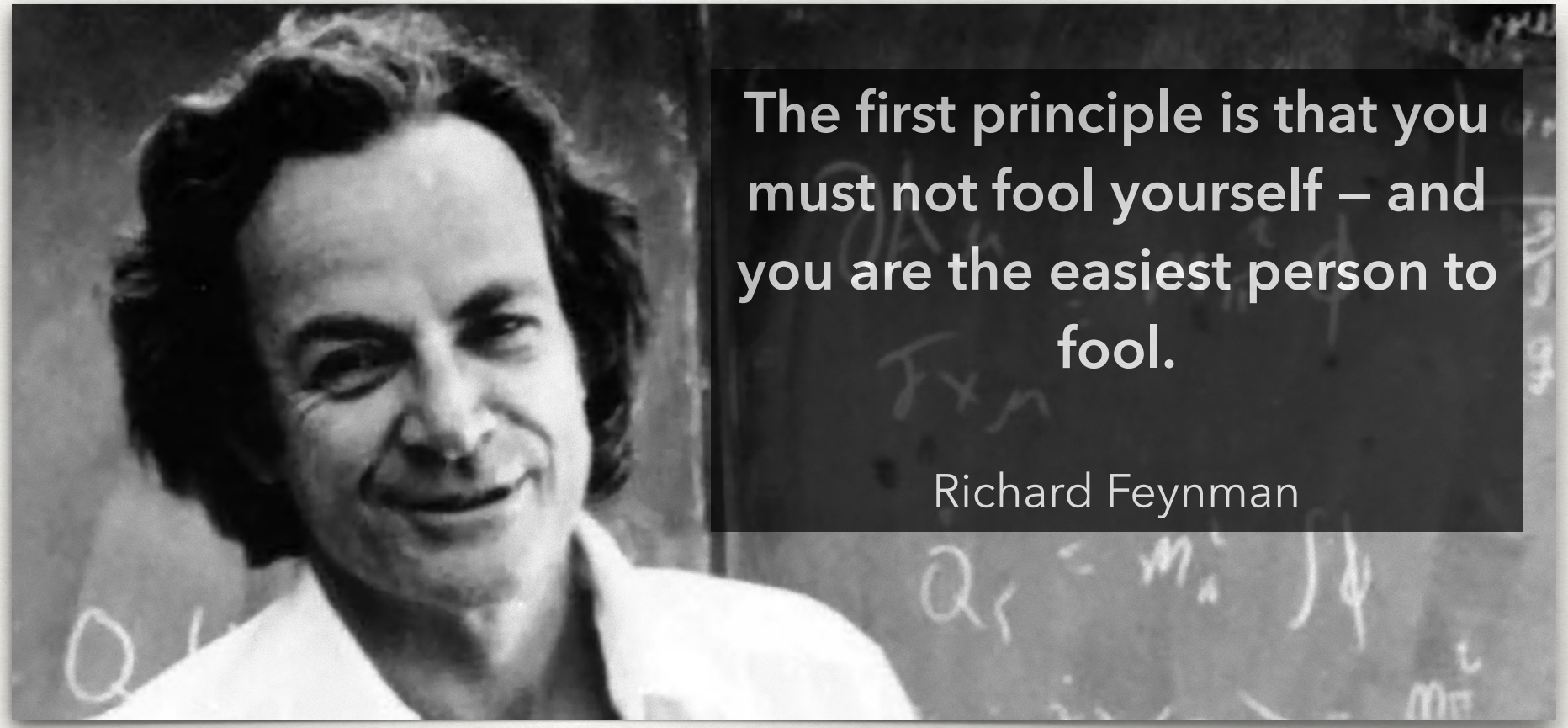
More in "Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency", by Hanna Wallach

HI HA BONES RAONS PER SER ESCÈPTIC...

BIAIXOS COGNITIUS



BIAIXOS COGNITIUS



BIAIXOS COGNITIUS

"Naïve inductivism": a belief that all scientists seeing the same data should come to the same conclusions.

In **"Of P-Values and Bayes: A Modest Proposal"**, Steven N. Goodman, 2001

By implication, anyone who draws a different conclusion must be doing so for nonscientific reasons.

It is a belief that scientific reasoning requires little more than statistical model fitting, or in our case, reporting odds ratios, P-values and the like, to arrive at the truth.

BIAIXOS COGNITIUS

When the same statistical information is conveyed in different ways, people make drastically different decisions.

The two most common modes used for communicating results are *description* and *illustration*.

Analyzed outcomes are more reliable than they actually were.

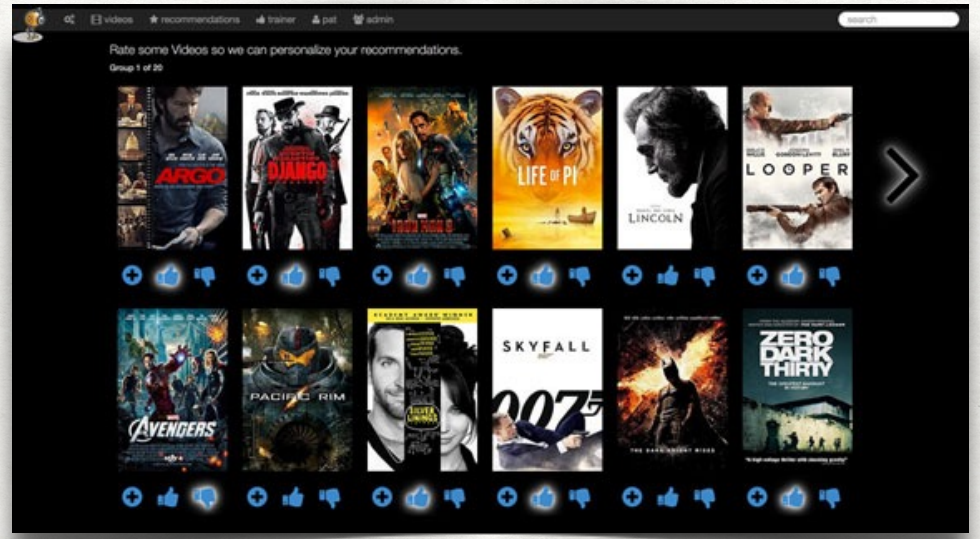
Uncertainties are more apparent but the more variables, connections, patterns are in the data, the harder it becomes to illustrate it.

Daniel Kahneman and Behavioral Economics.

MODELS ERRONIS

Algorithms are not always fair

A 95% accurate movie recommender is a great algorithm, but it can be not fair…

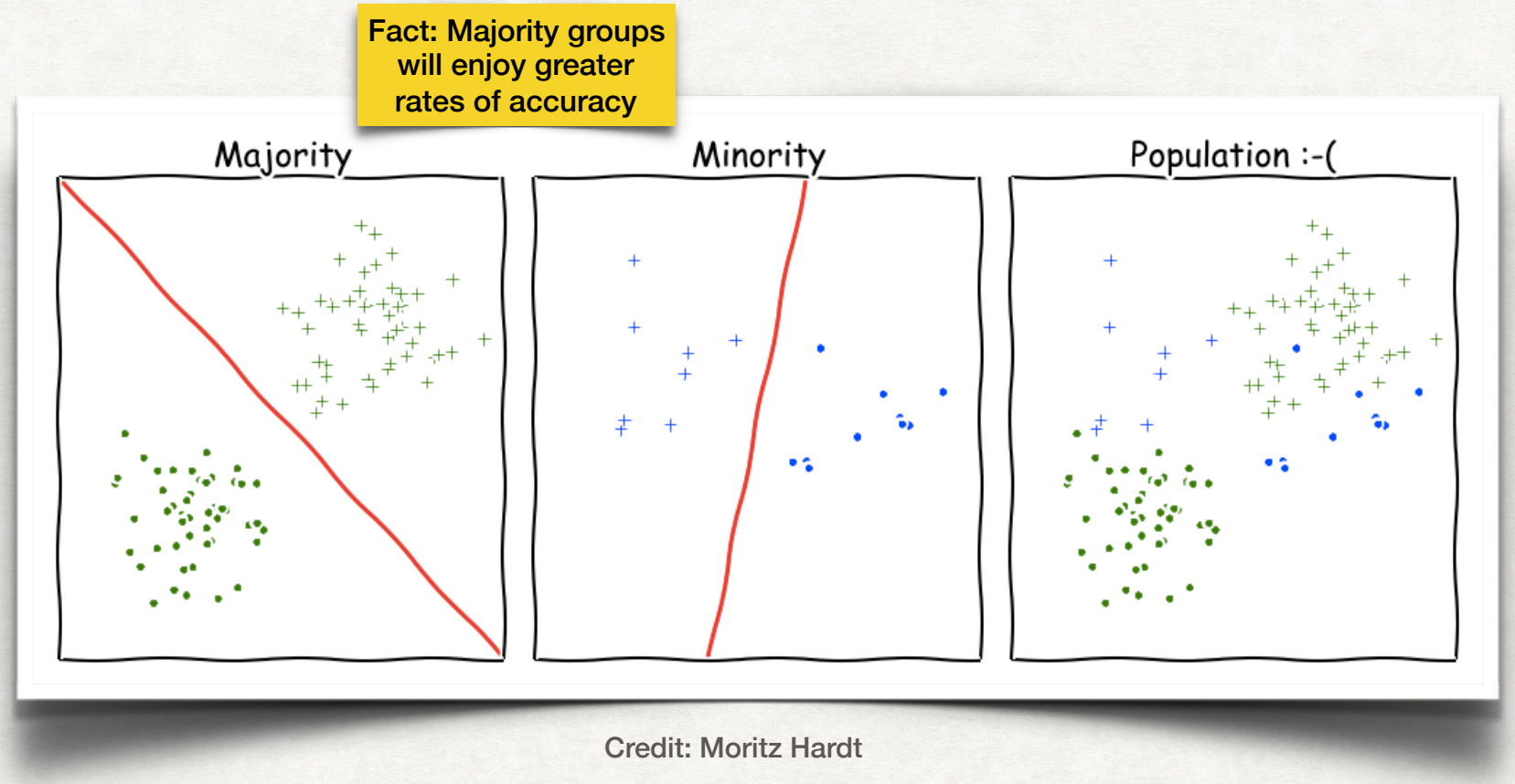


It can be 95% accurate because of noise

It can be 95% accurate because it nails recommending movies for people from major cultures but only achieves 50% for people from other cultures.

MODELS ERRONIS

Algorithms are not always fair



Credit: Moritz Hardt

Even if two groups of the population admit simple classifiers, the whole population may not.

MODELS ERRONIS

Facebook

Facebook apologises over 'cruel' Year in Review clips

The social network apologised after it showed one user a photograph of his recently-deceased daughter in its 'Year in Review' feature; many others had similar complaints

Alex Hern

Monday 29 December 2014 15.11 GMT

A Facebook Year in Review posted to Twitter by Julieanne Smolinski. Photograph: Twitter

MODELS ERRONIS

theguardian
Winner of the Pulitzer prize 2014

home › tech | UK | world | politics | sport | football | opinion | culture | business | all

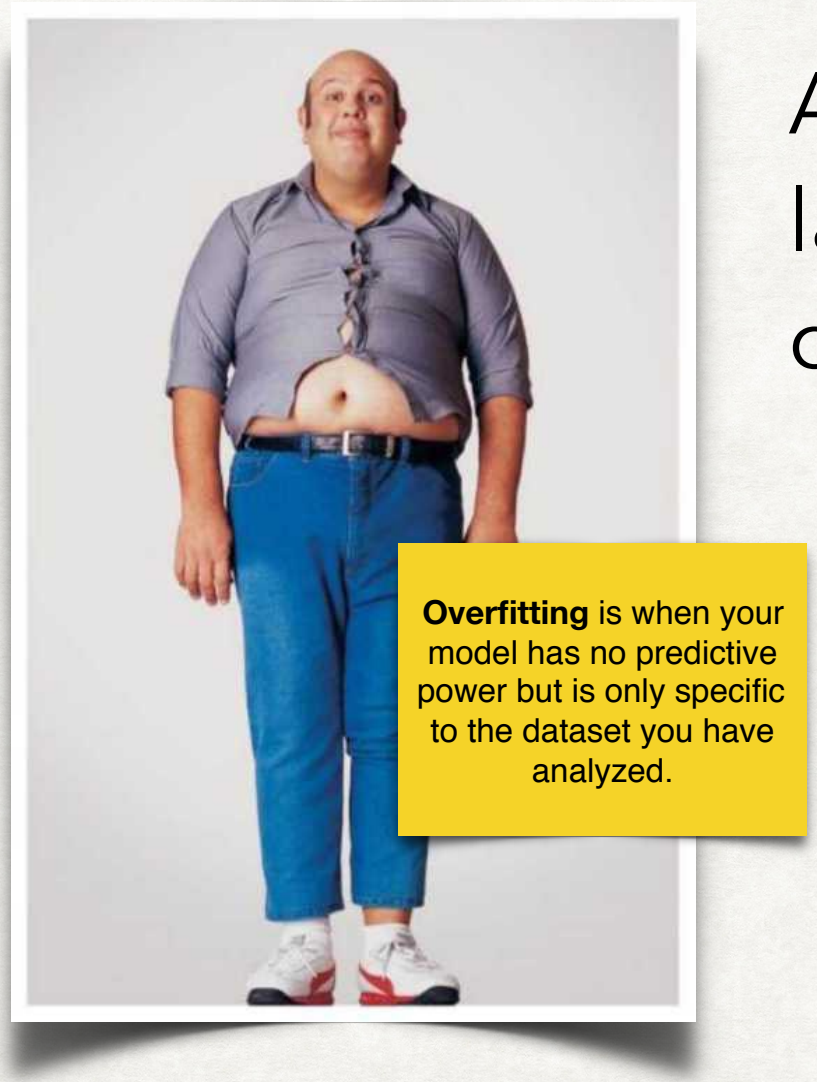
Flickr

Flickr faces complaints over 'offensive' auto-tagging for photos

Auto-tagging system slaps 'animal' and 'ape' labels on images of black people, and tags concentration camps with 'jungle gym' and 'sport'

The famous train tracks leading into Auschwitz, which were labelled "sport" by Flickr's algorithm. Photograph: Christopher Furlong/Getty Images

MODELS ERRONIS



Overfitting is when your model has no predictive power but is only specific to the dataset you have analyzed.

A perfect model with lack of generalization capabilities.

MALES DADES

Sentence length and word sophistication have been found to correlate well with the scores of human graders, but they cannot be the base of a program for grading student essays…..

MALES DADES

Data can be subject to unexpected feedback effects.

SECTIONS HOME SEARCH The New York Times

Bits

Search Bits SEARCH

Disruptions: Data Without Context Tells a Misleading Story

By NICK BILTON FEBRUARY 24, 2013 11:00 AM 4 Comments

Google's Flu Predictor overestimated how many people had the flu this flu season. Erik S. Lesser/European Pressphoto Agency

PREVIOUS POST
Samsung's New 8-Inch Tablet Takes on the iPad Mini

NEXT POST
Dell's Intentions Get a Hard Look

THE BITS DAILY UPDATE

Every weekday, **get the latest technology news,** analysis and buzz from around the web — delivered to your inbox.

SIGN UP FOR OUR NEWSLETTER See a Sample »

SCUTTLEBOT *News from the Web, annotated by our staff*

Uber Strengthens Driver Background Checks in India
UBER BLOG | The ride-hailing company has partnered with First Advantage, a global screening agency, to do better background checks in the region, weeks after a woman was allegedly raped by her Uber driver. - *Mike Isaac*

Why I Am Not a Maker
THE ATLANTIC | When tech culture only celebrates creation, it risks ignoring those who teach, criticize, and take care of others. - *Jenna Wortham*

Facebook's Privacy Policy Reviewed by Hamburg Data Regulator
BLOOMBERG | Says Johannes Caspar, the data protection commissioner of Hamburg: "I think it's problematic that Facebook wants to exchange user data between all of its

MALES DADES

"**The Pepsi Challenge**" data illustrated that customers preferred the taste of Pepsi over that of coke.

Data is the product of experiments. Designing good experiments is an art.

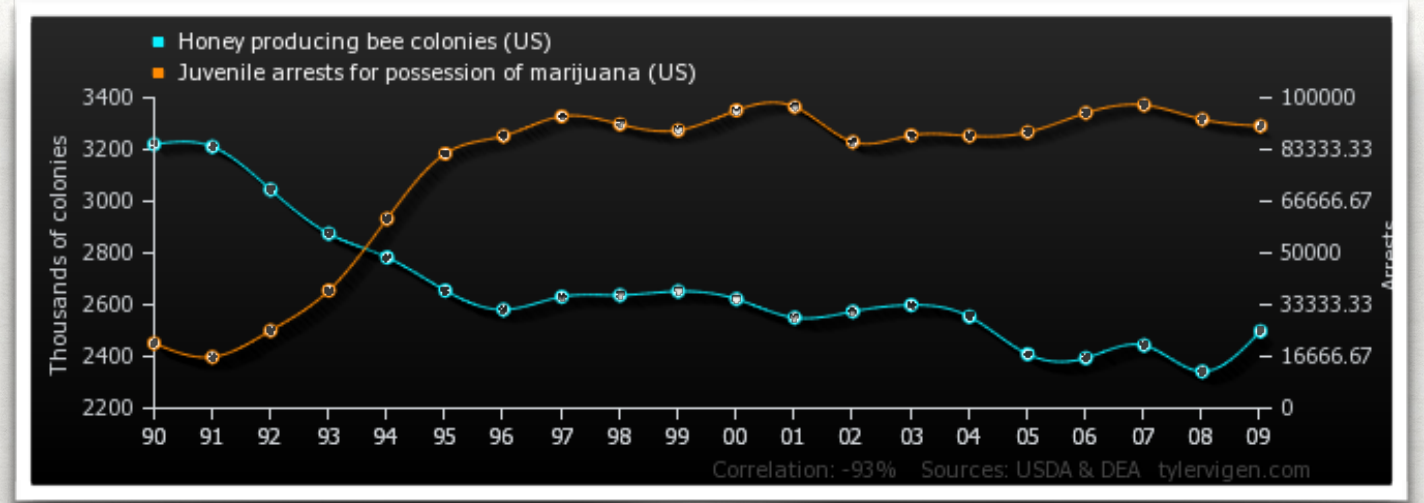
There were two "fallacies" in the study. **First**, the study used "sips" from small paper cups; sweeter taste is preferred in small sips, but not in larger consumption. **Second**, a possibly more "valuable" data point was not shared; consumers who "discovered" through the survey that they preferred the taste of Pepsi, still preferred to buy Coke: "We're a Coke household."

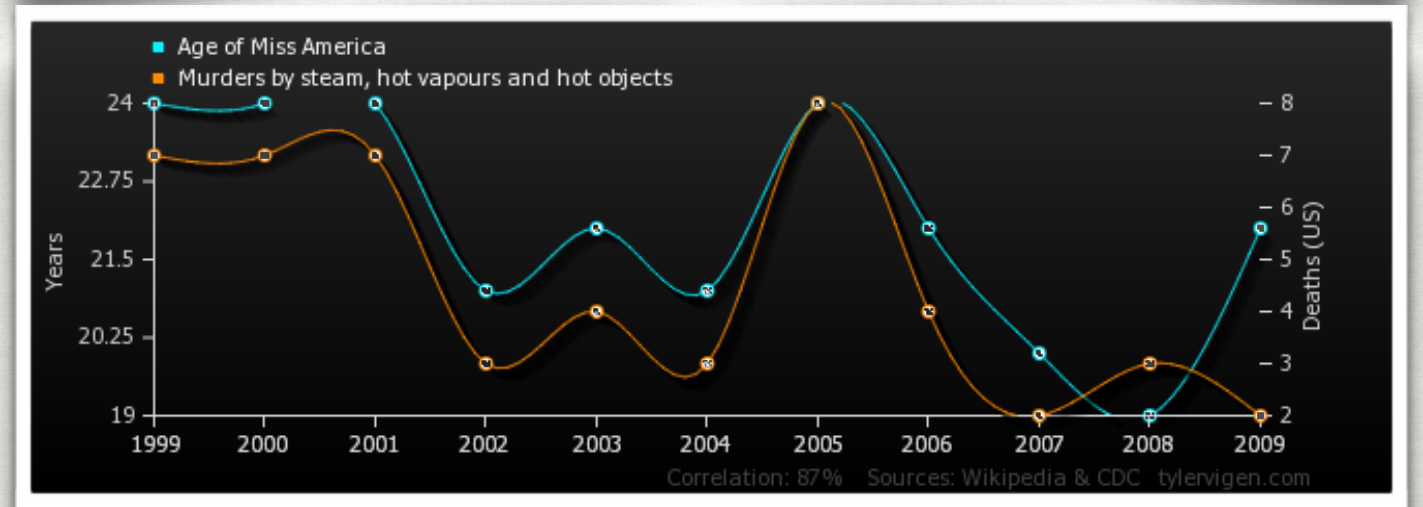
CASUALITAT

If you have enough data correlation is as good as causation. False!

The more data you have the more spurious correlations will show up. True!

CASUALITAT





Source: Spurious Correlations
http://www.tylervigen.com/

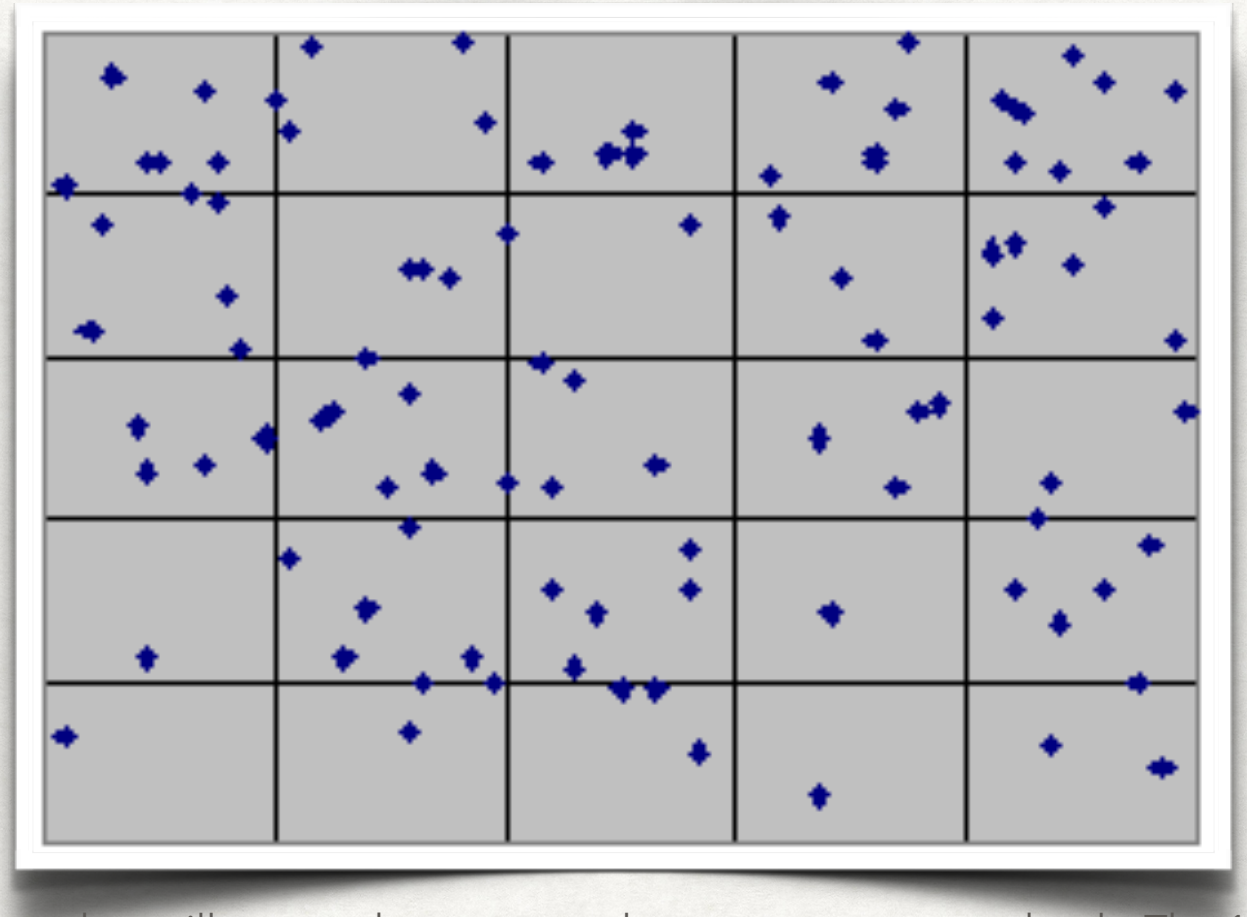
CASUALITAT

Randomness does not mean absence of structure or partial order.

Random means unpredictable.

CASUALITAT

Some occurrences of increased (or decreased) rates of cancer are due to random variation. This is particularly true where small numbers of people are involved.

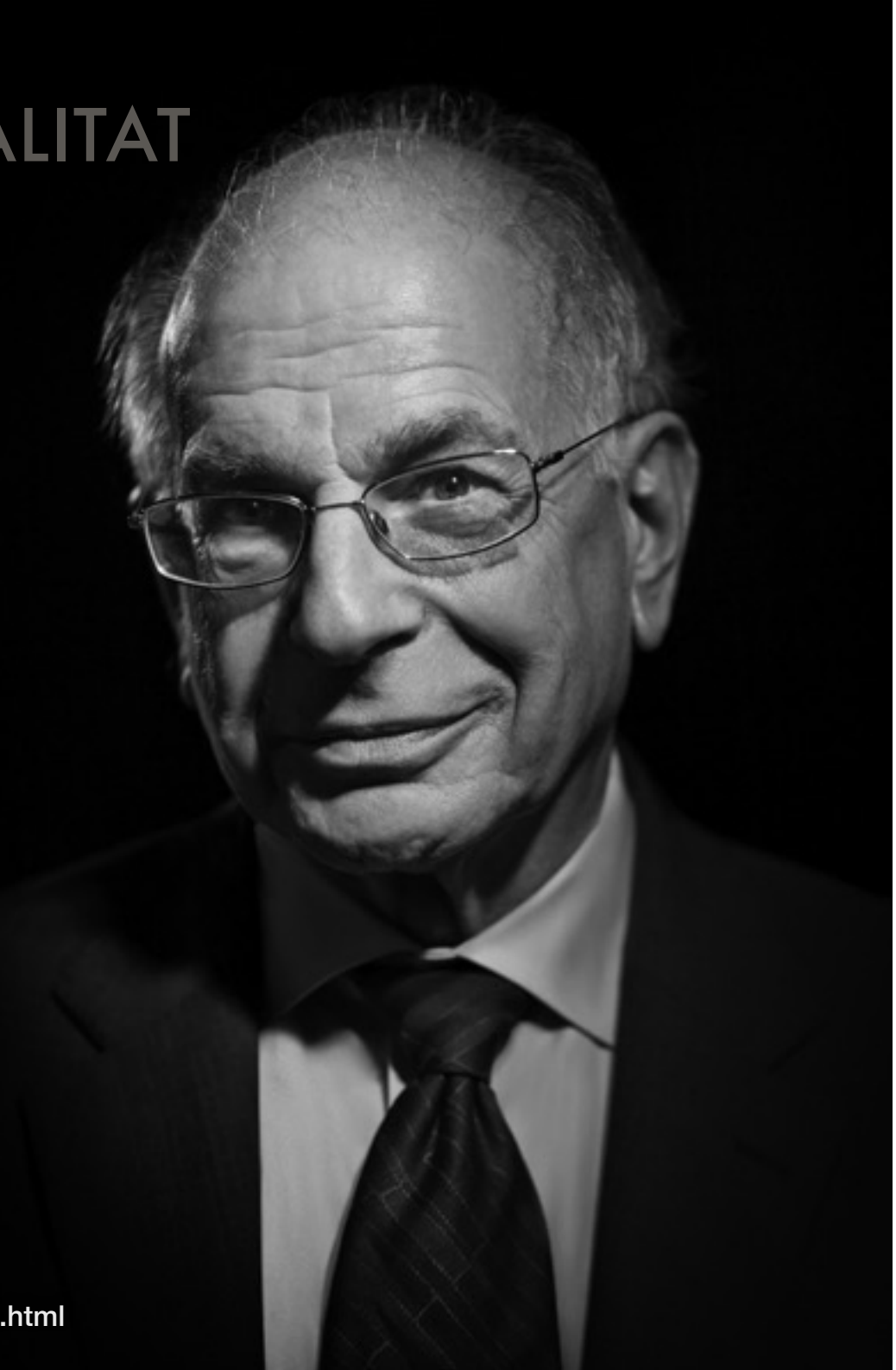


The figure above illustrates how cancer clusters can occur randomly. The 100 dots on this grid were randomly generated. In theory, there should be four dots in each of the 25 areas. But some have only one dot and others have many more than four.

CASUALITAT

In 2012 Professor Kahneman wrote:

"90% of the students who saw the CRT in normal font made at least one mistake in the test, but the proportion dropped to 35% when the font was barely legible. You read this correctly: performance was better with the bad font."

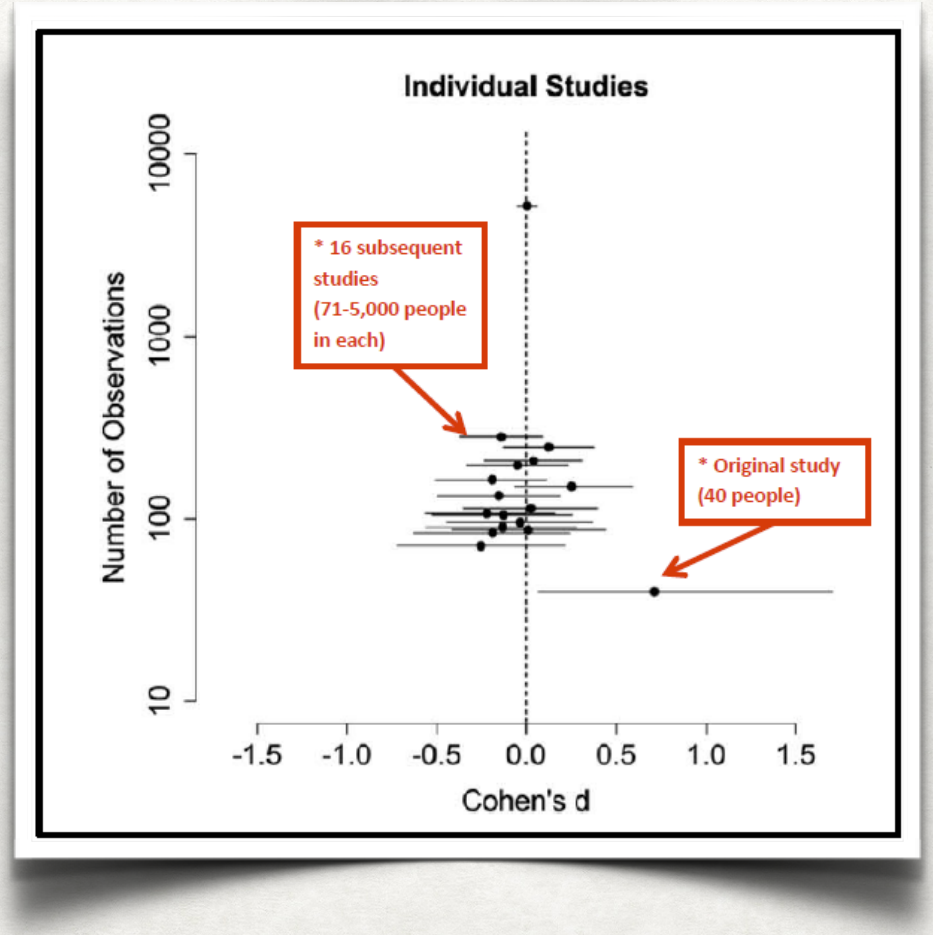


http://www.terryburnham.com/2015/04/a-trick-for-higher-sat-scores.html

CASUALITAT

The original paper reached its conclusions based on the test scores of 40 people.

If you analyze a total of over 7,000 people by looking at the original study and 16 additional studies:



http://www.terryburnham.com/2015/04/a-trick-for-higher-sat-scores.html

CASUALITAT

paper	comment	citations as of April 20, 2015	citations as of today
Alter et al. (2007). "Overcoming intuition: metacognitive difficulty activates analytic reasoning." Journal of Experimental Psychology: General 136(4): 569.	Original paper showing hard-to-read leads to higher scores	344	click for current count
Thompson et al. (2013). "The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking." Cognition 128(2): 237-251.	Paper contradicts Alter at. al by reporting no hard-to-read effect.	38	click for current count
Meyer et al. (2015). "Disfluent fonts don't help people solve math problems." Journal of Experimental Psychology: General 144(2): e16.	Our paper summarizing the original study and 16 others.	0 (this "should" increase at least as fast as citations for Alter et. al, 2007)	click for current count

http://www.terryburnham.com/2015/04/a-trick-for-higher-sat-scores.html

ÈTICA

Skepticism means to ask a simple question: What is the most evil thing that can be done with my model?

ÈTICA

Unfair use of data.

You cannot access to our “best offer” prices because you have the capacity to pay more.

ÈTICA

Use of bad features.

Credit denial because of racial identity.

ÈTICA

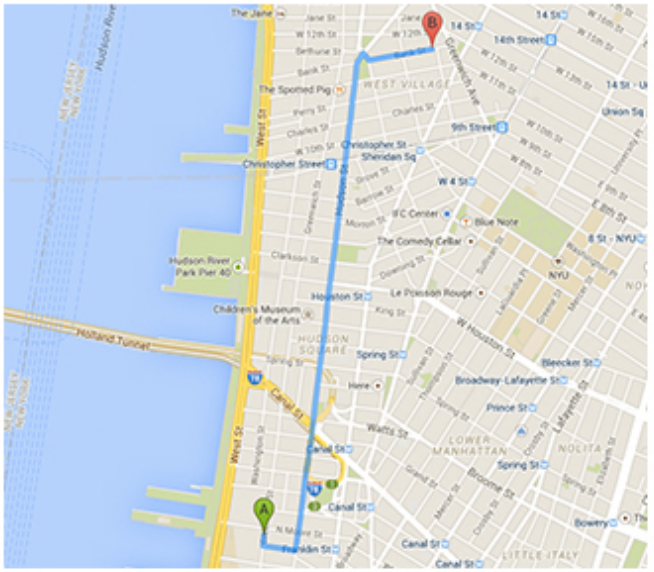
Privacy.

Public NYC Taxicab Database Lets You See How Celebrities Tip

J.K. Trotter
Filed to: DATA 10/23/14 1:00pm

134,190 18

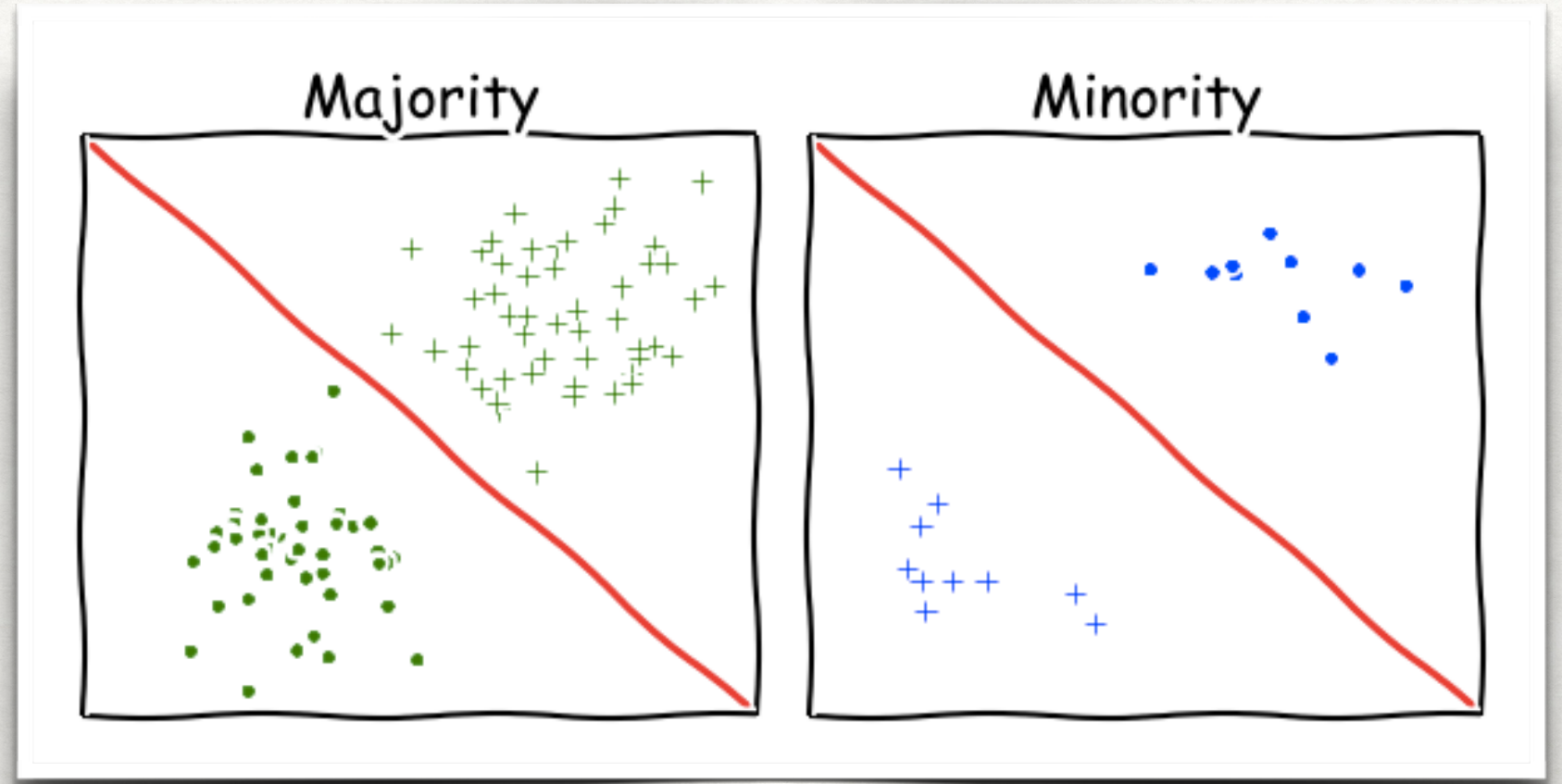
BRADLEY COOPER



JULY 8, 2013 • 7:34 PM - 7:44 PM
376 GREENWICH ST. TO 13 BANK ST.
$9.00 FARE • CASH; UNKNOWN TIP • ©SPLASH

ÈTICA

Inclusion.



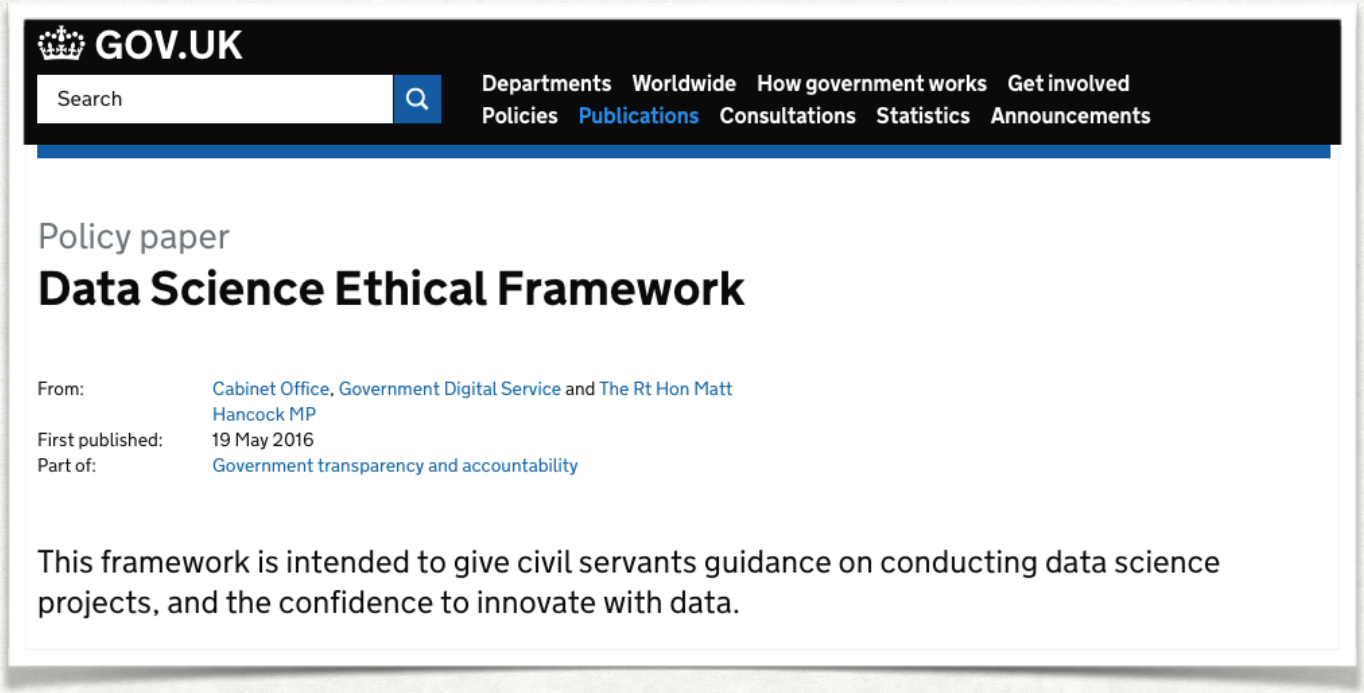
The job of predicting should be managed as medicine:

PRIMUM, NON NOCERE

A good model is a model that is useful even when it fails.

Source: N.Silver, The signal and the noise, 2012

EXEMPLE:
DATA SCIENCE ETHICAL FRAMEWORK



GOV.UK

Search

Departments Worldwide How government works Get involved
Policies Publications Consultations Statistics Announcements

Policy paper

Data Science Ethical Framework

From: Cabinet Office, Government Digital Service and The Rt Hon Matt Hancock MP
First published: 19 May 2016
Part of: Government transparency and accountability

This framework is intended to give civil servants guidance on conducting data science projects, and the confidence to innovate with data.