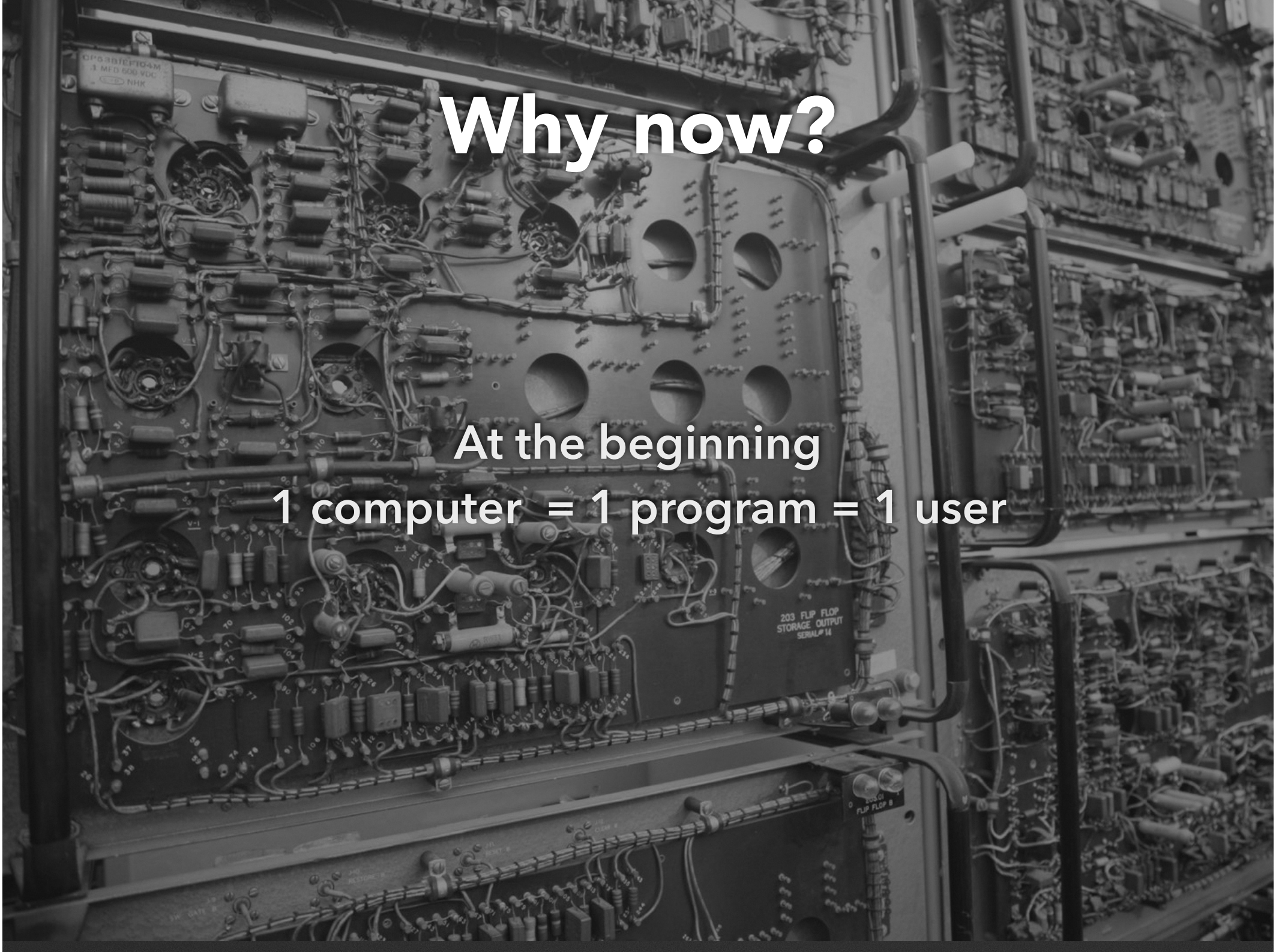
What is Data Science?

(a personal view)

Jordi Vitrià, PhD
Universitat de Barcelona

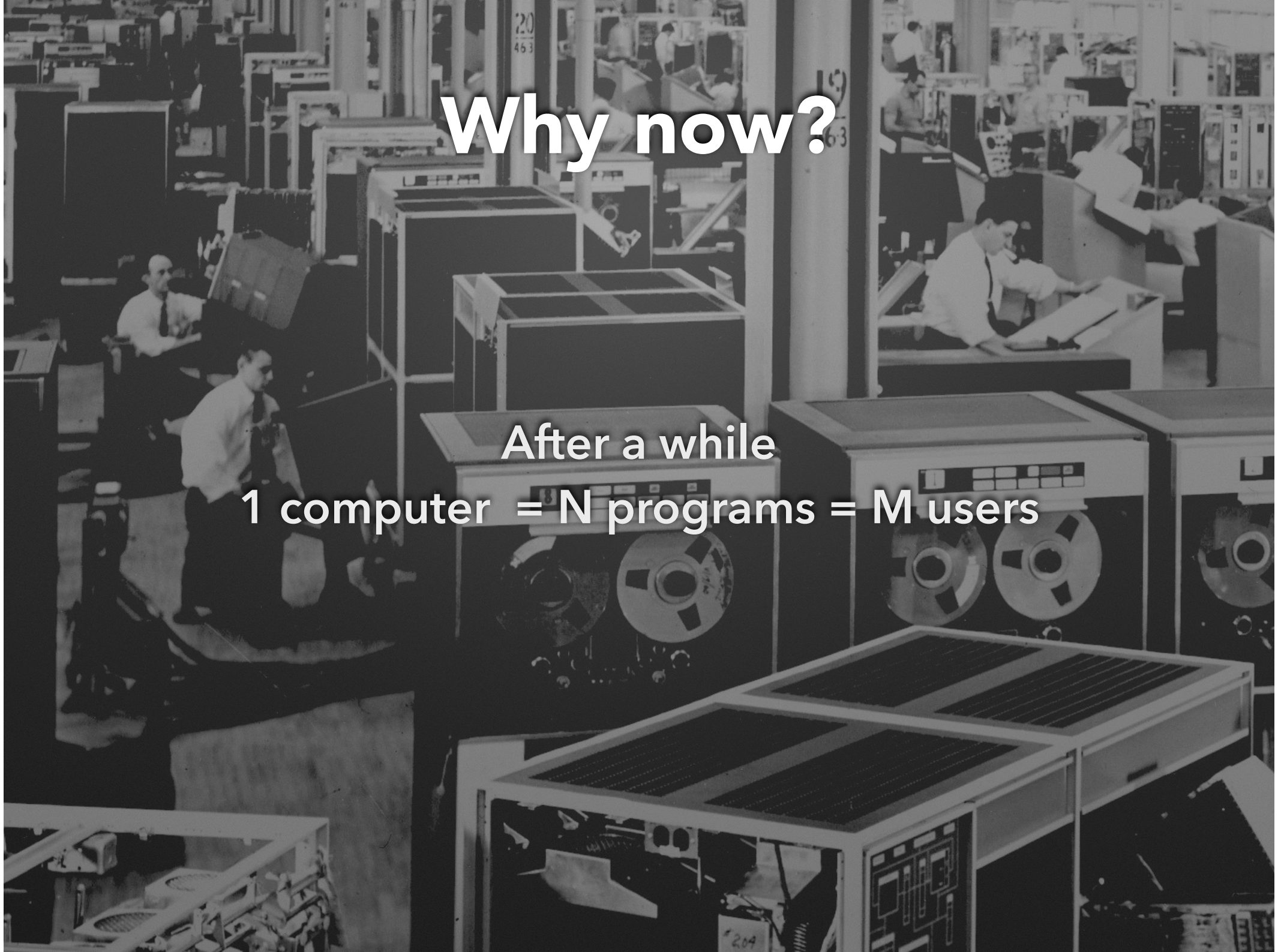

Data Science
Big Data

Taking (big)data-based decisions is not new but now it is easier.



Why now?

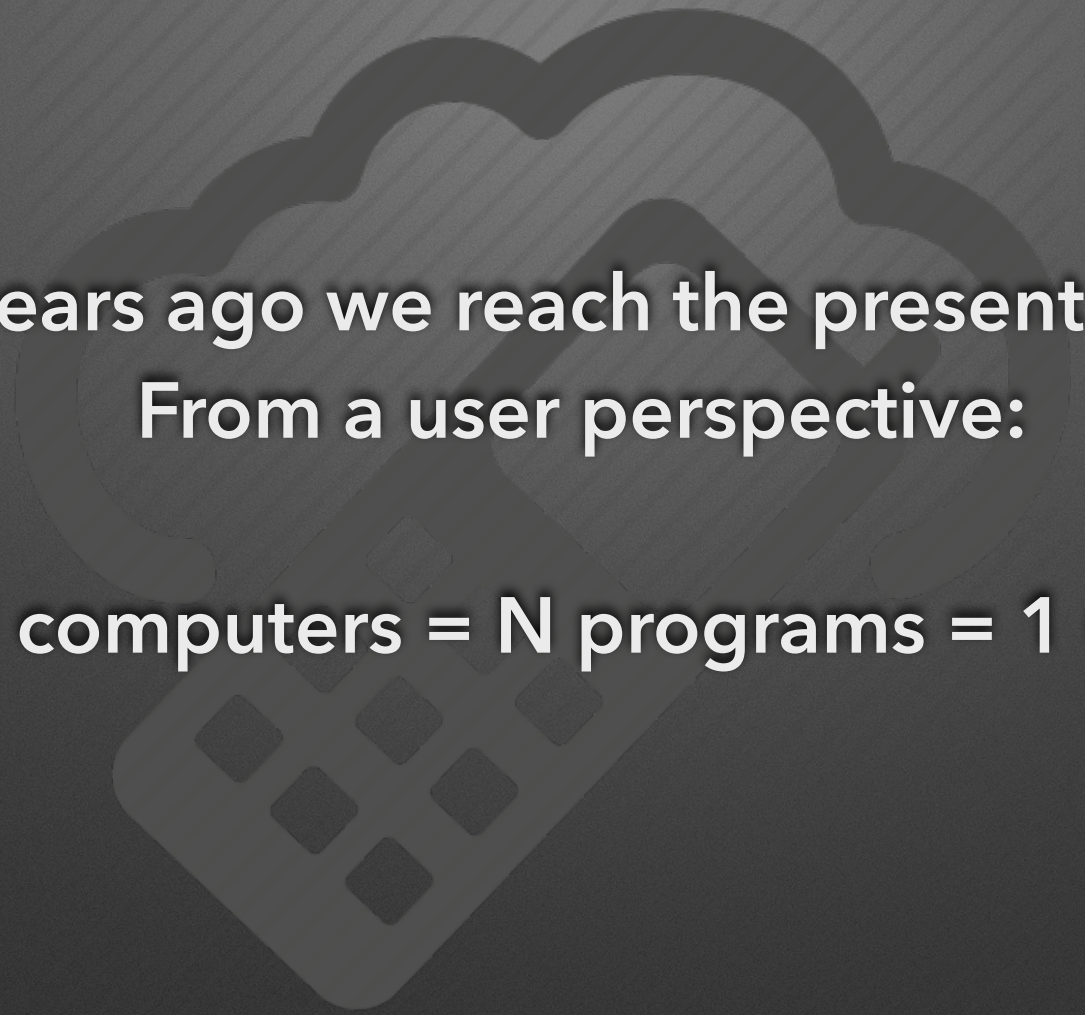
At the beginning
1 computer = 1 program = 1 user


Why now?
After a while
1 computer = N programs = M users


Why now?
Then
1 computer = N programs = 1 user

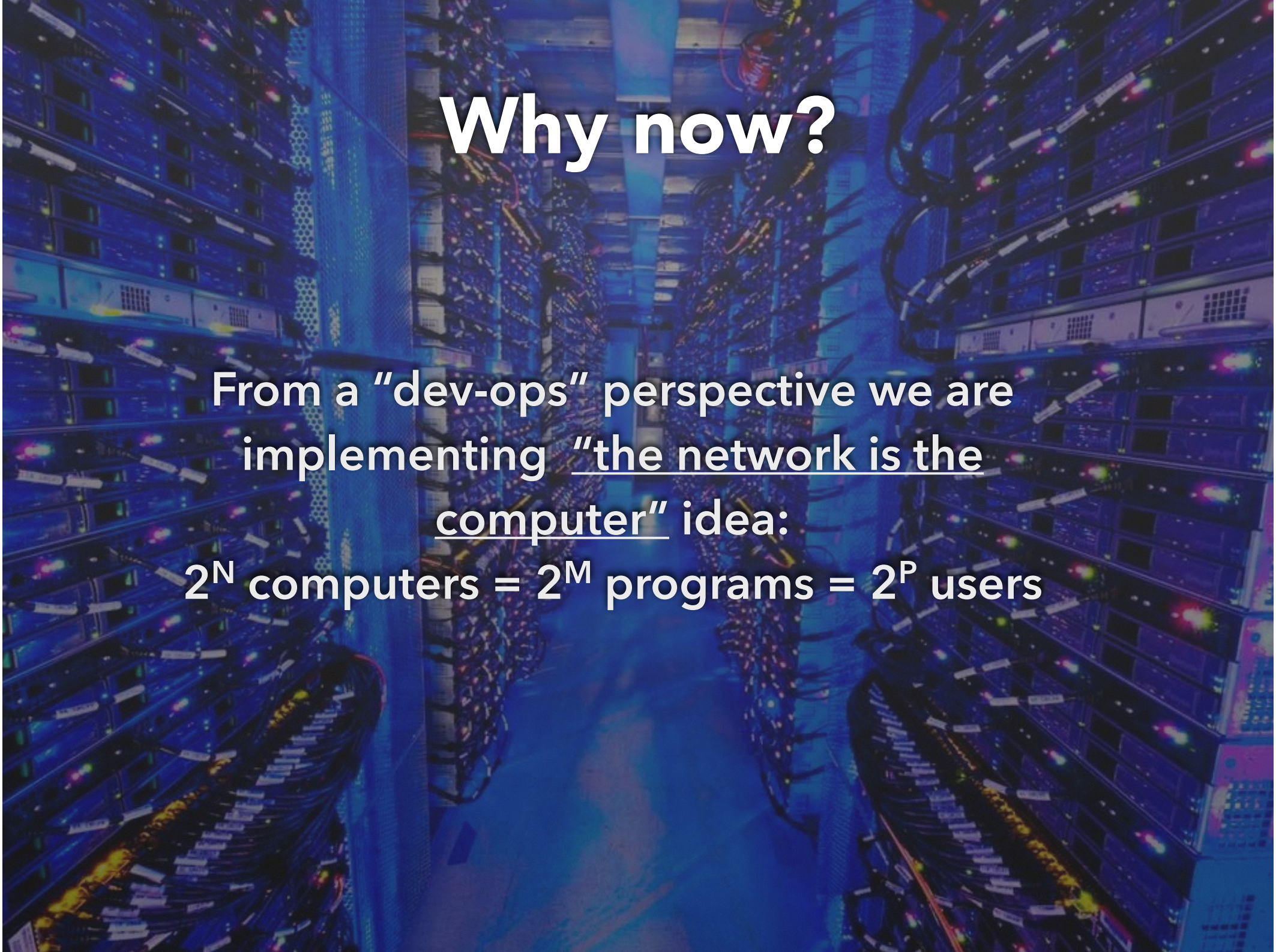
Why now?

Why now?



A few years ago we reach the present situation.
From a user perspective:

M computers = N programs = 1 user



Why now?

From a “dev-ops” perspective we are implementing “the network is the computer” idea:

2^N computers = 2^M programs = 2^P users

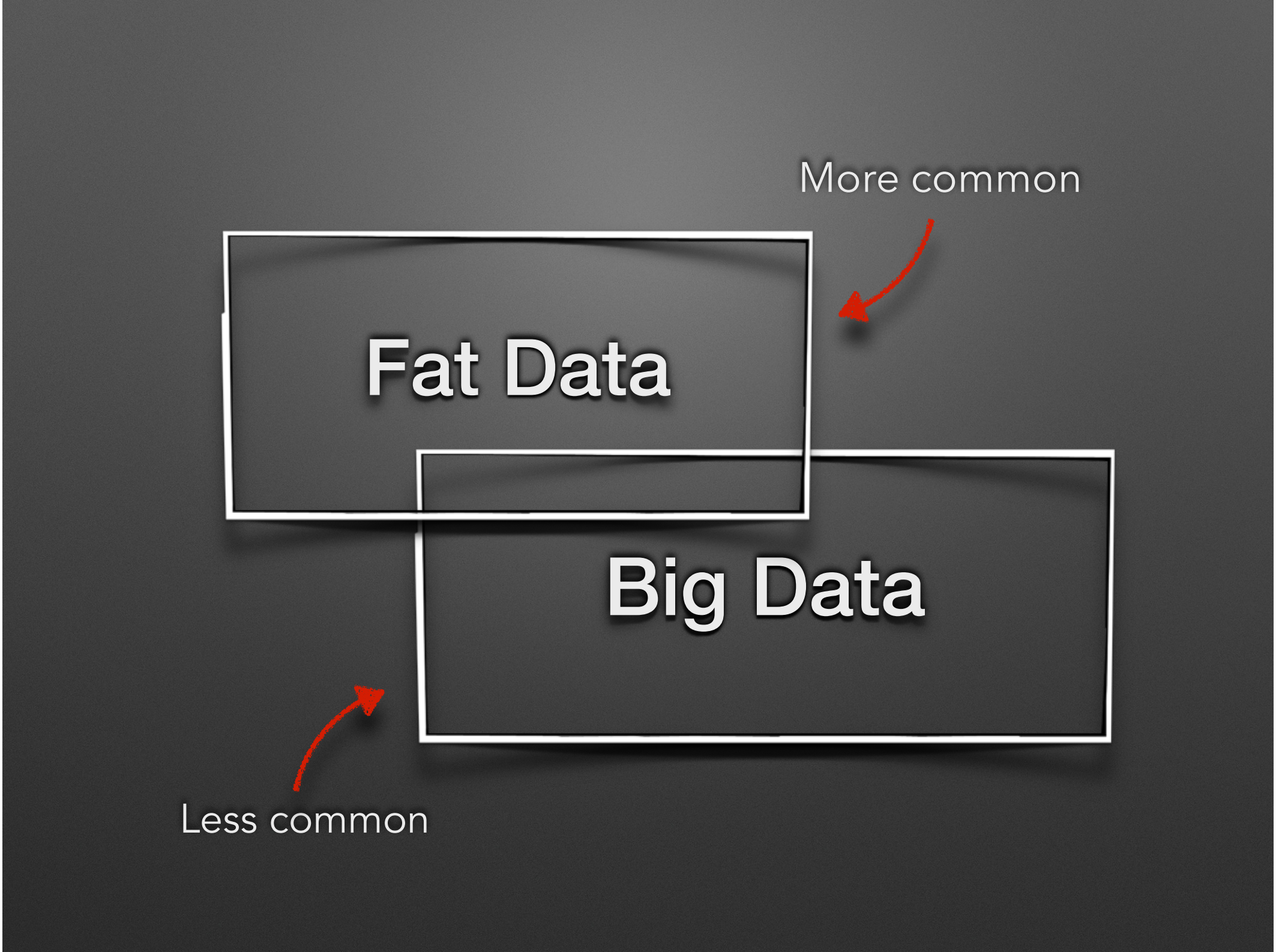
Why now?

The “cloud” is a necessary condition to process big data, but not the main cause of the Big Data fever.

Big Data

What is Big Data?

- For some people, they have big data when its size > 65536 x 256.
- In general we have big data when its size does not allow its storage and analysis in a big computer.


More common
Fat Data
Big Data
Less common

Big Data

Wal-Mart handles over one million customer transaction per hour, the information is stored on a database sized in excess of 2.5 Petabytes ($2{,}0 \times 10^{16}$ bits).

By 2016 it is likely that a typical hospital will create 665 terabytes (5.32×10^{15} bits) of data a year.

Big Data

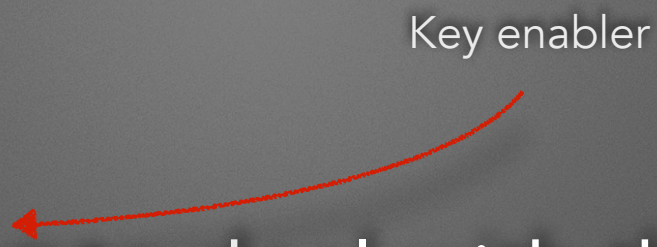
With a personal computer:

- You can find an element in a 1 MB file in less than a second.
- You can find an element in a 1 GB file in less than a minute.
- You can find an element in a 1 TB file in less than sixteen hours.
- You can find an element in a 1 PB file in less than two years.
- You can find an element in a 1 EB file in less than two thousand years.

Big Data

Big data is more than size.
It is commonly characterized with several V:

Big Data



The cloud is key to deal with the four V, but the main phenomenon behind Big Data is **datification**.

The four V are a consequence of it.

Big Data

We are rendering into data many aspects of the world that have never been quantified before:

business networks
books I'm reading
location
physical activity
consumed food
purchases
physiological signals
straight thoughts
friendship
gaze
driving behavior

Big Data

Information comes from:

- Corporate Data Bases (structured information).
- Unstructured information in documents, Wikipedia, textbooks, journals, blogs, tweets, etc.
- Images in the web, public cameras, phones, TV, YouTube, etc.
- Public APIs: smart cities, government, search engines, etc.
- Sensor Data: GPS, accelerometer, physico-chemical sensors, sociometric sensors, super-colliders, telescopes, etc.

Big Data

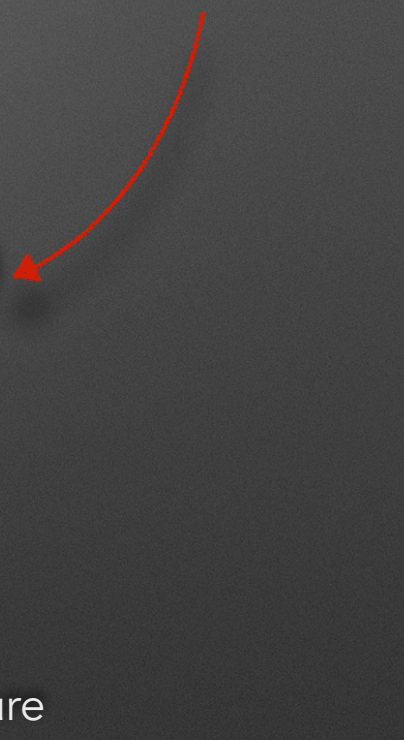
There are several Big Data flavors:

- Big multidimensional arrays (homogeneous data).
- Big tables (structured data).
- Big text.
- Big image.
- Big sound.
- Big sequential data (sensors, tweets, etc.)

Big Data

There are several problems:

- ETL (Extract, Transform, Load)
- BI/Analytics (Think you can do in SQL)
- **Advanced Analytics.**
- **Machine Learning.**
- Visualization.



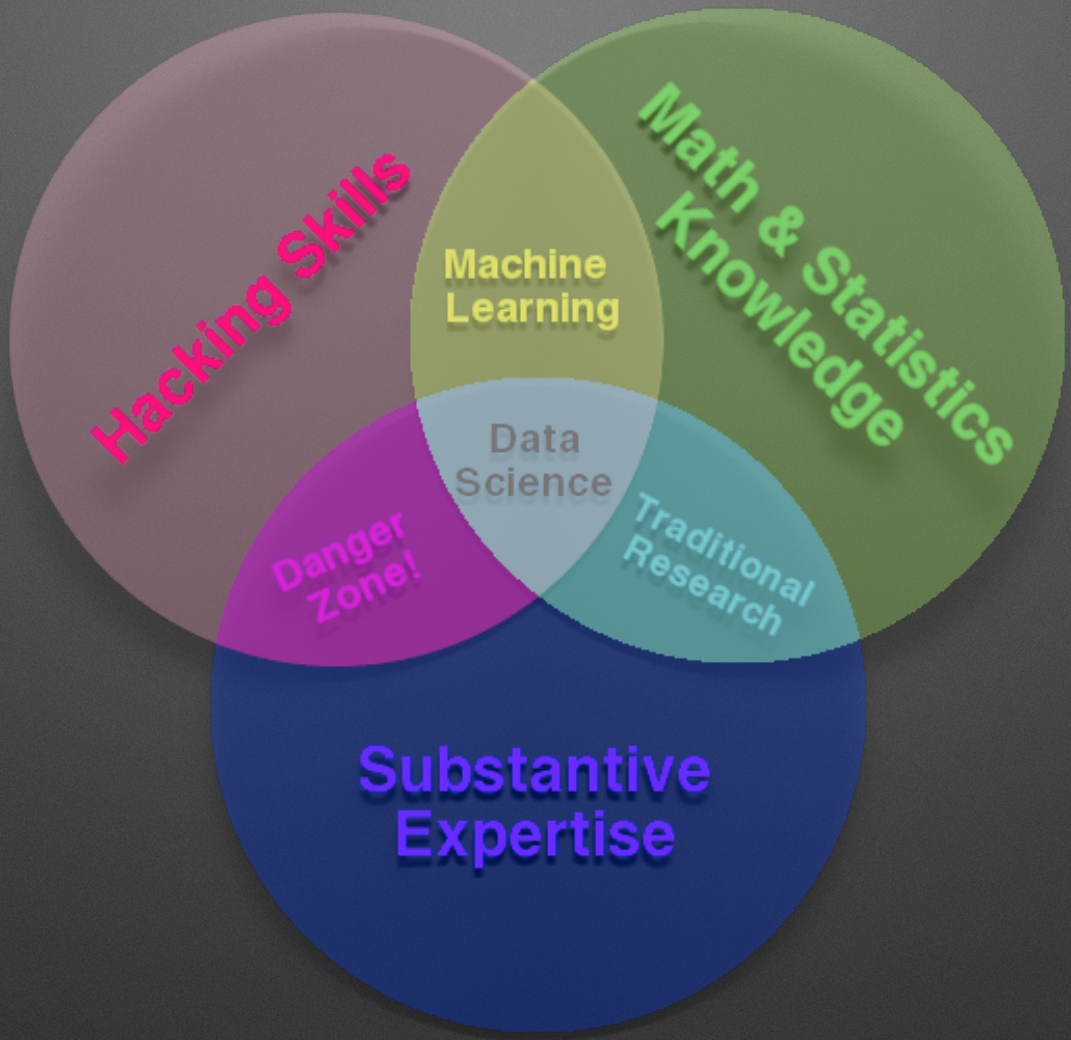
Data Science

Technology is the collection of tools, including machinery, modifications, arrangements and procedures used by humans.

Big Data is a key **technology** to process massive amounts of data (f.e. to count items).

Methodology is the systematic, theoretical analysis of the methods applied to a field of study.

Data Science is a **methodology** to define what we want to do with data, how do we evaluate our actions, what decisions can be grounded on data, how do we combine evidences from several sources, etc.



Drew Conway's Data Science Venn Diagram

Data Science

Data Science is **not** a **science** but a methodology based on multidisciplinar knowledge.

> *Currently, most company decisions are based on intuition and best practices. The alternative is to integrate data-based knowledge in the decision process.*

Data Science is a new data processing model focused on turning data into actions.

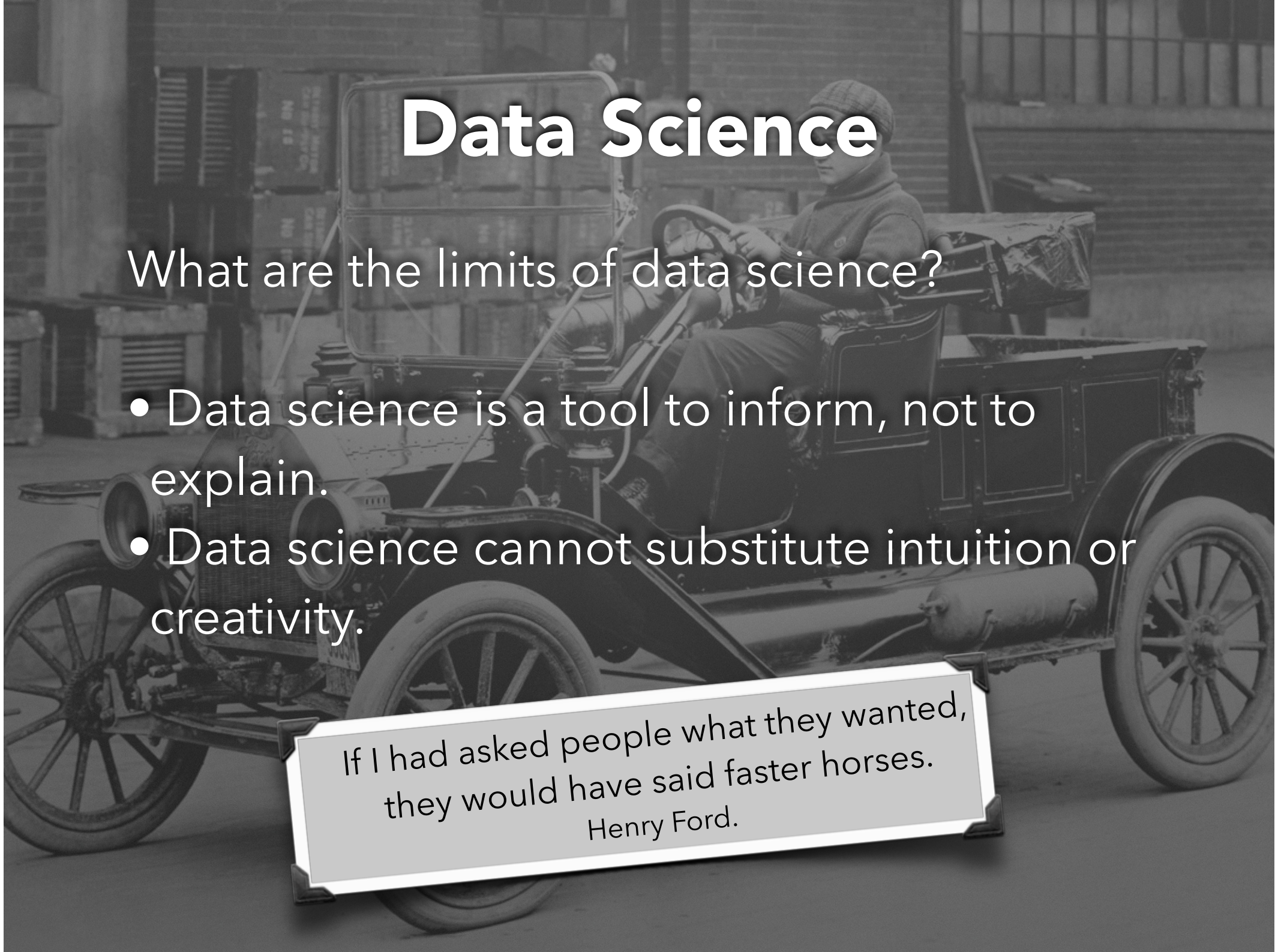
Data Science

Steps:

- Ask a question.
- Get the data. They can be heterogeneous and non structured.
- Data Processing (cleaning, munging, etc.).
- Data Analysis (computer science, linguistics, economy, sociology, etc.).
- Take a decision and act.

Data Science

Data Science is a new job!



Data Science

What are the limits of data science?

- Data science is a tool to inform, not to explain.
- Data science cannot substitute intuition or creativity.

> If I had asked people what they wanted, they would have said faster horses.
>
> Henry Ford.

Data Science

COMPANY
Mastercard

INDUSTRY
Finance

EMPLOYEES
67,000

TYPE
Behavioral Analytics

PURPOSE:

With 1.8 billion customers, MasterCard is in the unique position of being able to analyze the behavior of customers in not only their own stores, but also thousands of other retailers. The company teamed up with Mu Sigma to collect and analyze data on shoppers' behavior, and provide the insights it finds to other retailers in benchmarking reports.

Data Science

COMPANY
Starbucks Coffee

INDUSTRY
Food & Beverage

EMPLOYEES
160,000

TYPE
Behavioral Analytics

PURPOSE:

Starbucks collects data on its customers' purchasing habits in order to send personalized ads and coupon offers to the consumers' mobile phones. The company also identifies trends indicating whether customers are losing interest in their product and directs offers specifically to those customers in order to regenerate interest.

Data Science

COMPANY
Spotify

INDUSTRY
Entertainment

EMPLOYEES
5,000

TYPE
Customer Segmentation & Behavioral Analytics

PURPOSE:

Spotify uses data from user profiles and users' playlists, and historical data on music played to provide recommendations for each user. By combining data from millions of users, Spotify is able to make recommendations even if a particular user doesn't have an extensive history with the site.

Data Science

COMPANY
Union Pacific Railroad

INDUSTRY
Transportation

EMPLOYEES
44,000

TYPE
Predictive Support

PURPOSE:

With predictive analytics and tools such as visual sensors and thermometers, Union Pacific can detect imminent problems with railway tracks in order to predict potential derailments days before they would likely occur. So far the sensors have reduced derailments by 75 percent.

Data Science

COMPANY
Coca-Cola Co.

INDUSTRY
Food

EMPLOYEES
146,200

TYPE
Market Basket Analysis

PURPOSE:

Coca-Cola uses an algorithm to ensure that its orange juice has a consistent taste throughout the year. The algorithm incorporates satellite imagery, crop yields, consumer preferences and details about the flavours that make up a particular fruit in order to determine how the juice should be blended.

Conclusions

- Big Data will be soon a **commodity** that will be used mainly for data munging and counting at scale.
- The most difficult part of **Big Data** is **Data.**
- Data Science is a new job with a bright future.

"The Incredulity of Saint Thomas" by Caravaggio

Skeptical Data Science

(to go beyond "Type I Lies")

jordi vitrià
universitat de barcelona

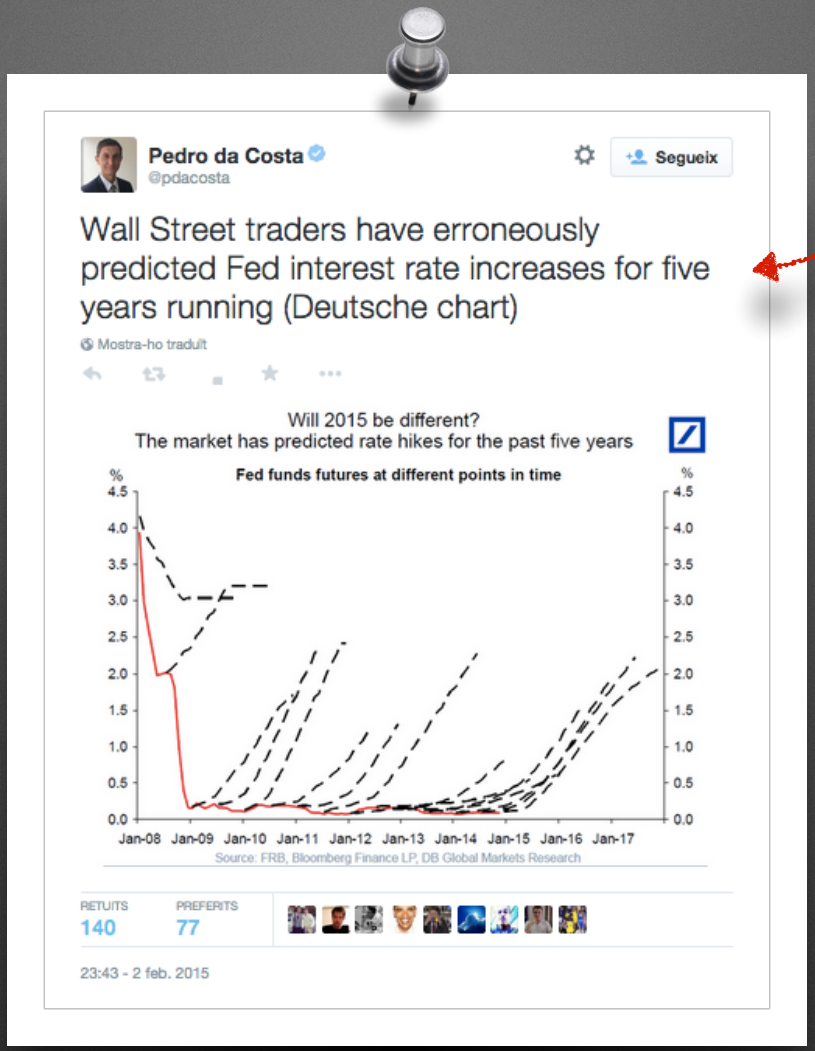
Lying with data: Type I Lies

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.

Lying with data: Type I Lies

Not today

Lying with data: Type II Lies

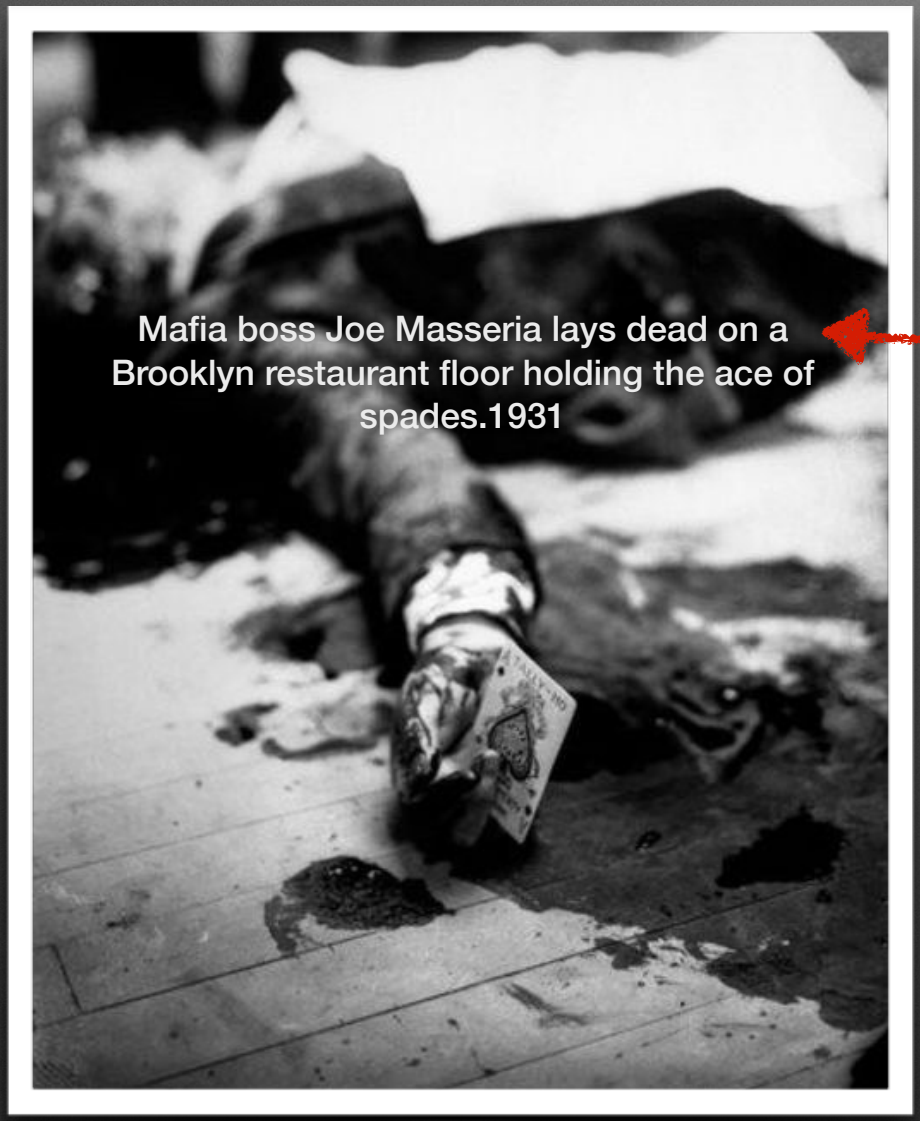


Today

Unintentional Lies

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

The backslash against big data is in full swing. *Mike Loukides, O'Reilly (2014)*



Big Data lays dead holding our algorithmic future. 2015

substitute

The backslash against big data is in full swing. *Mike Loukides, O'Reilly (2014)*

From http://www.cs.nyu.edu/faculty/davise/papers/BigDataBib.html

General Critiques

Kate Crawford, The Hidden Biases in Big Data, Harvard Business Review Blog, April 1, 2013.
Tim Harford, Big data: Are We Making a Big Mistake? Financial Times, March 28, 2014.
John Horgan, So Far, Big Data is Small Potatoes , Scientific American blog, June 9, 2014.
Gary Langer Growing Doubts about Big Data, ABC News, blog. April 8, 2014.
Gary Marcus, Steamrolled by Big Data The New Yorker (online), April 3, 2013.
Gary Marcus and Ernest Davis, Eight (No, Nine!) Problems with Big Data Op-Ed, New York Times, April 7, 2014.
Megan Scudellari, Scientists Question the Big Price Tags of Big Data, Newsweek, July 24, 2014.
(…)

Social and legal critiques

David Auerbach, You are what you click: On microtargeting, The Nation March 4, 2013.
Yian Q. Mui, Little-known firms tracking data used in credit scores Washington Post, July 16, 2011.
Frank Pasquale, The Dark Market for Personal Data New York Times, October 17, 2014.
Room for Debate, Is Big Data Spreading Inequality,? NY Times, August 6, 2014.
(…)

Education

Carol Burris, Principal uncovers flawed data in her state's official education reports Washington Post, Nov. 22, 2014
Cathy O'Neil, Value-added model doesn't find bad teachers, causes administrators to cheat "mathbabe" blog, March 31, 2013.
(…)

The Facebook Mood Manipulation Experiment

This has generated an immense literature of responses in a very short time. A very extensive bibliography is here:
James Grimmelman, The Facebook Emotional Manipulation Study: Sources The Laboratorium.

(…)

Definitions

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.

Definitions

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

+ Example

Location Data Science

Data comes from tracking customer behaviors in their physical spaces by leveraging connected mobile devices such as smartphones, existing in-venue Wi-Fi networks, low cost Bluetooth-enabled beacons, surveillance cameras, etc.

Venue owners — from retail to airports to education to amusement parks — are applying **insights** gathered from location analytics to **act** in all aspects of their business.

+ Example

- **Design**. After analyzing traffic flows in their stores (computer vision), a big box retailer realized that less than 10% of customers visiting their shoe department engaged with the self-service wall display where merchandise was stacked. By relocating the benches to increase accessibility, sales in the department increased by double digits.

- **Marketing**. A restaurant chain wanted to understand the whether or not sponsoring a local music festival had a measurable impact on customer visits. By capturing data on 15,000 visitors passing through the festival entrances (wireless beacons) and comparing it to customers who visited their restaurants two months prior to the festival and two weeks after, they concluded the festival resulted in 1,300 net new customer visits.

+ Example

- **Operations**. A grocery store chain used location analytics to understand customer wait times in various departments and check-out registers. This data not only enabled the company to hold managers accountable for wait times, but it gave additional insight into (and justification for) staffing needs for each department throughout the day and optimal times to perform disruptive tasks such as restocking shelves or resetting displays.

- **Strategy**. A regional clothing chain was concerned that opening an outlet store would cannibalize customers from its main stores. After analyzing the customer base visiting each store, they discovered that less than 2% of their main store customers visited their outlet. The upside: the outlet gave them access to an entirely new customer base with minimal impact to existing store sales.

- Example

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

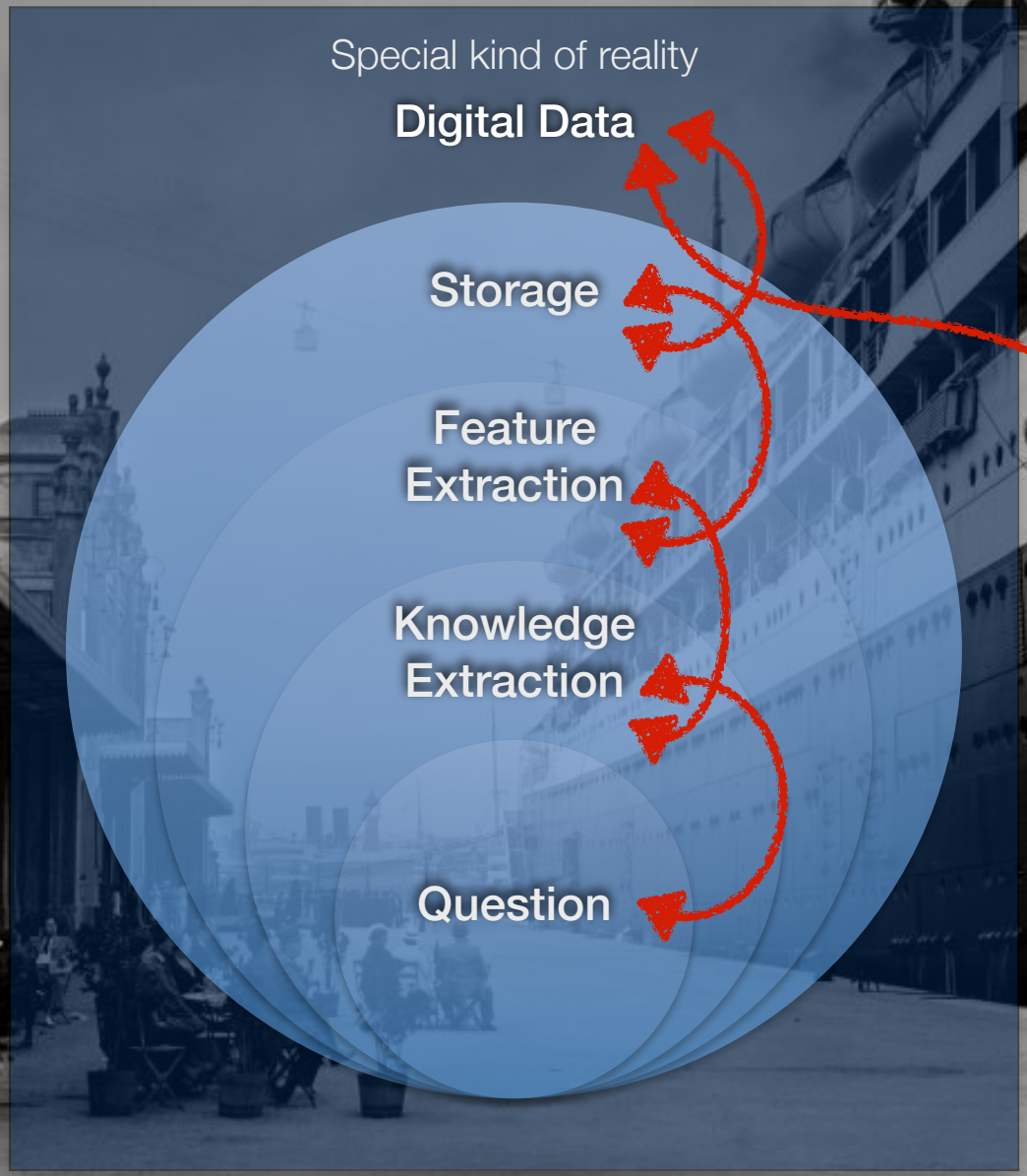
Xerox’s reprographic photo process.

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*”…

Definitions

From questions to reality: the data science path



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

What

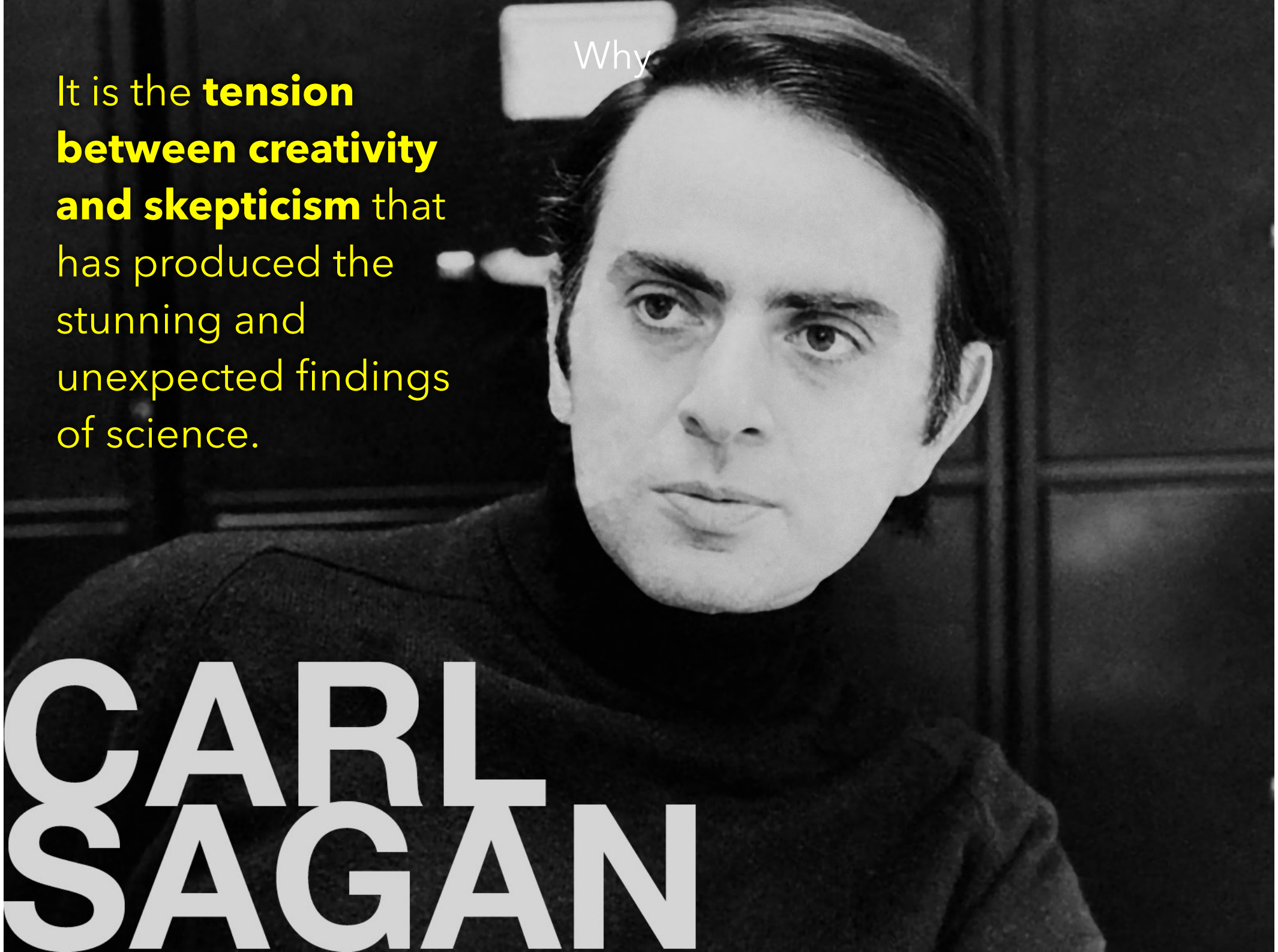
The true **meaning** of the word **skepticism** has nothing to do with doubt, disbelief, or negativity.

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

Brian Dunning

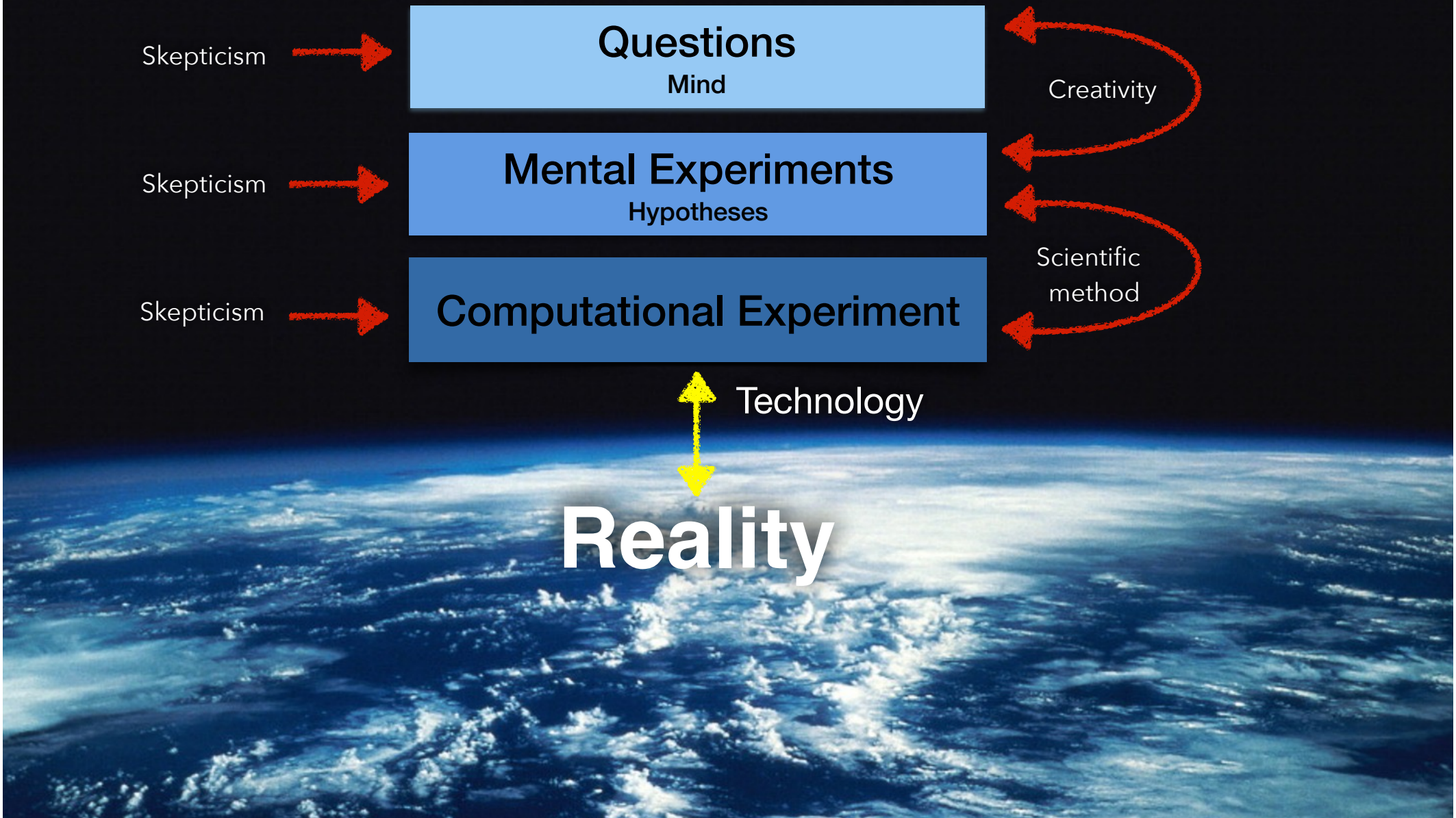


It’s easy to confuse being a skeptic with being a cynic.


Why
It is the **tension between creativity and skepticism** that has produced the stunning and unexpected findings of science.
CARL SAGAN

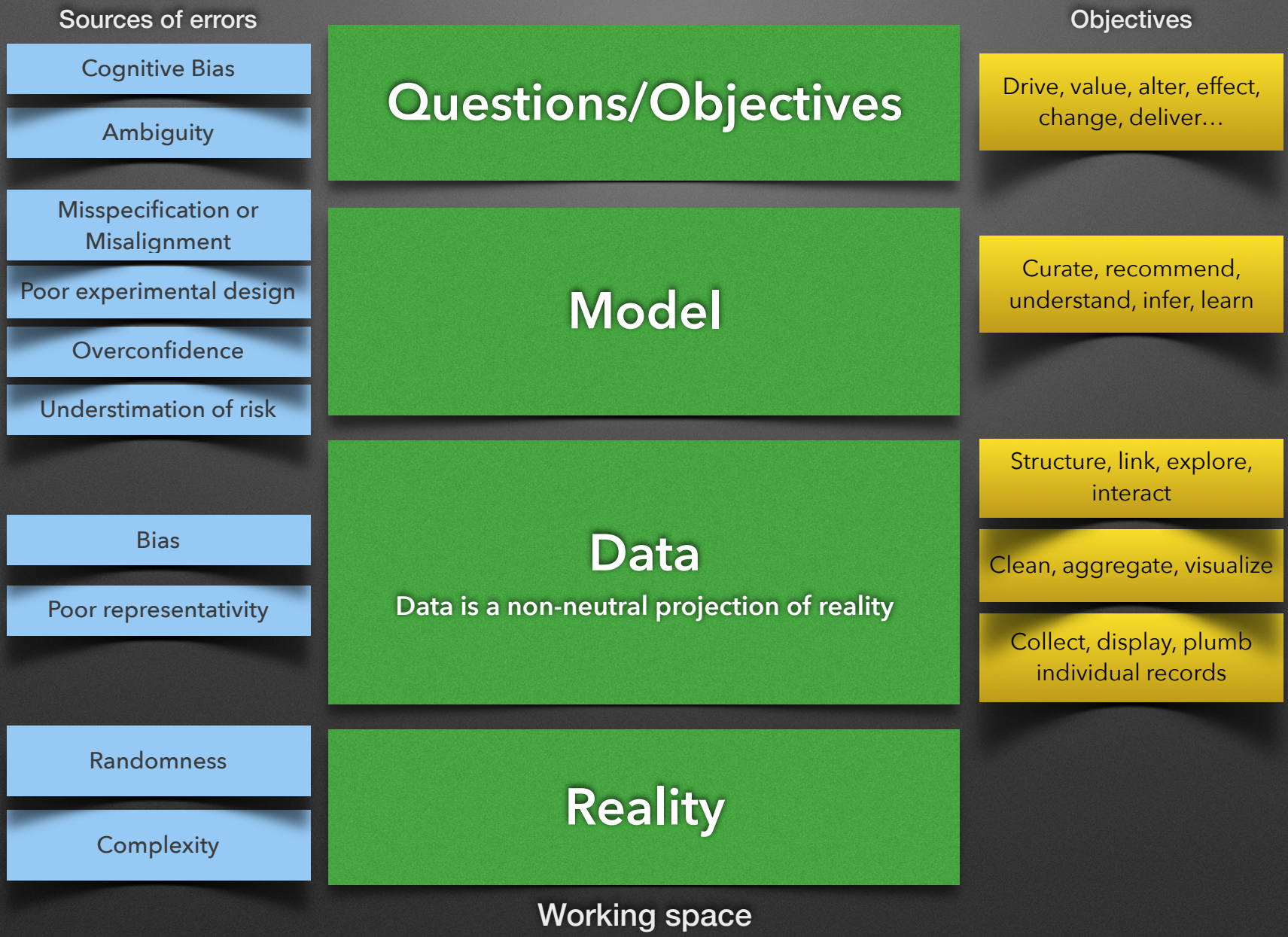
When

From question to reality: data science process



Why

Data science must be skeptical to be free of errors



Why

(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.

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.

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.

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

- Provenance
- Privacy
- Bias
- Fairness
- Inclusion
- Governance

Solution

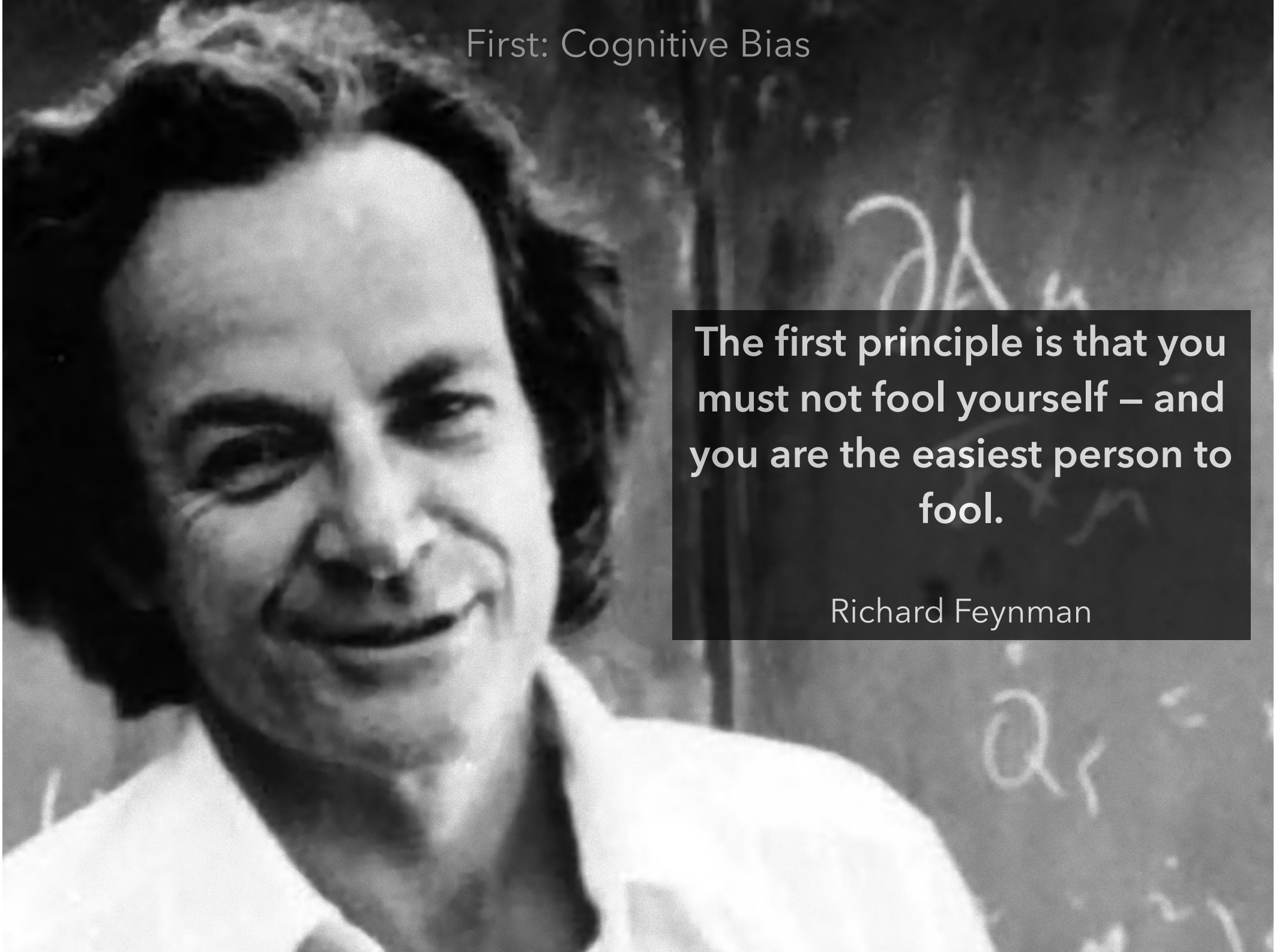
Putting the right question.
Answering the question by following an skeptical methodology.

- 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

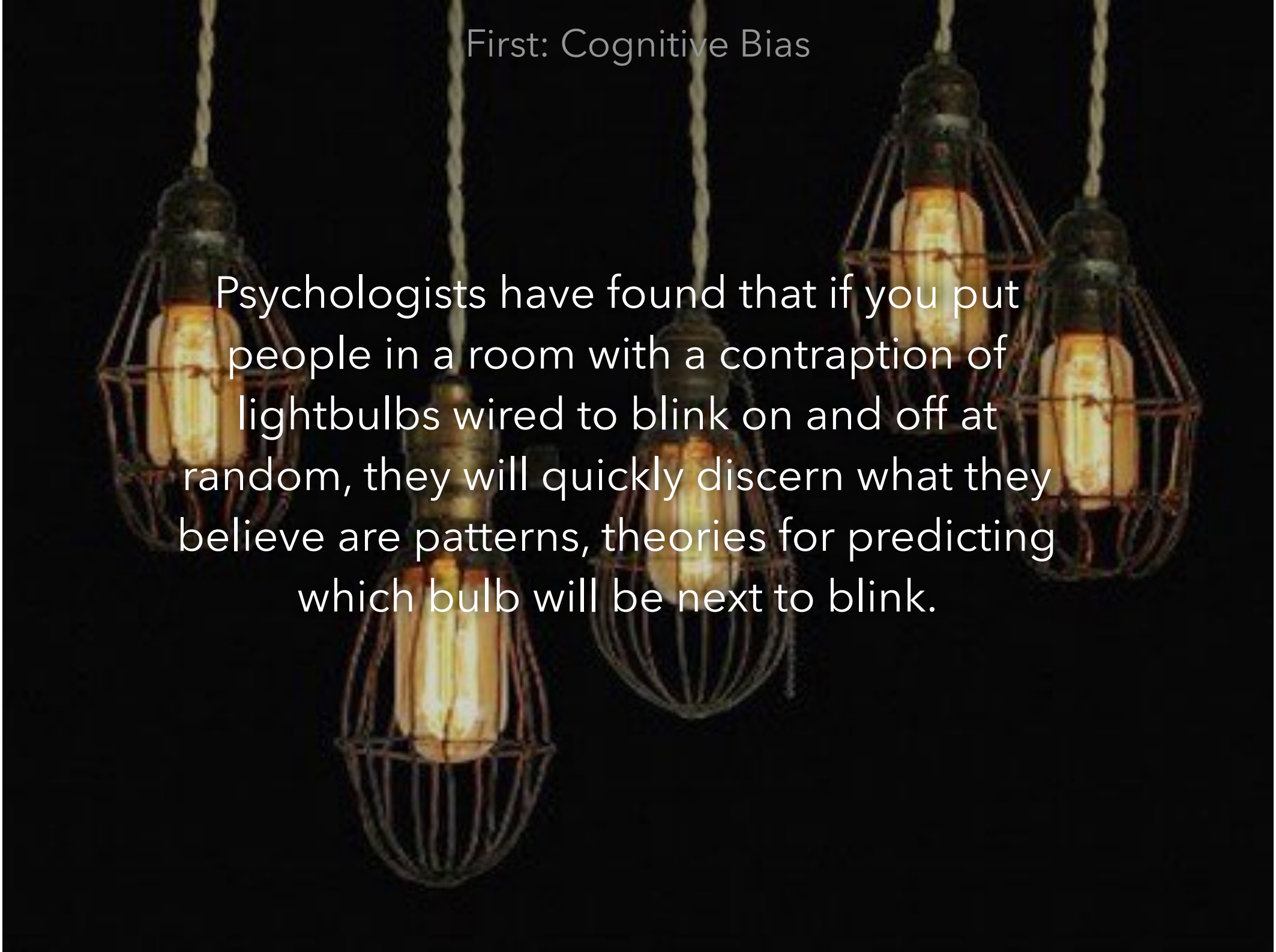

5 good reasons
to be skeptical

First: Cognitive Bias

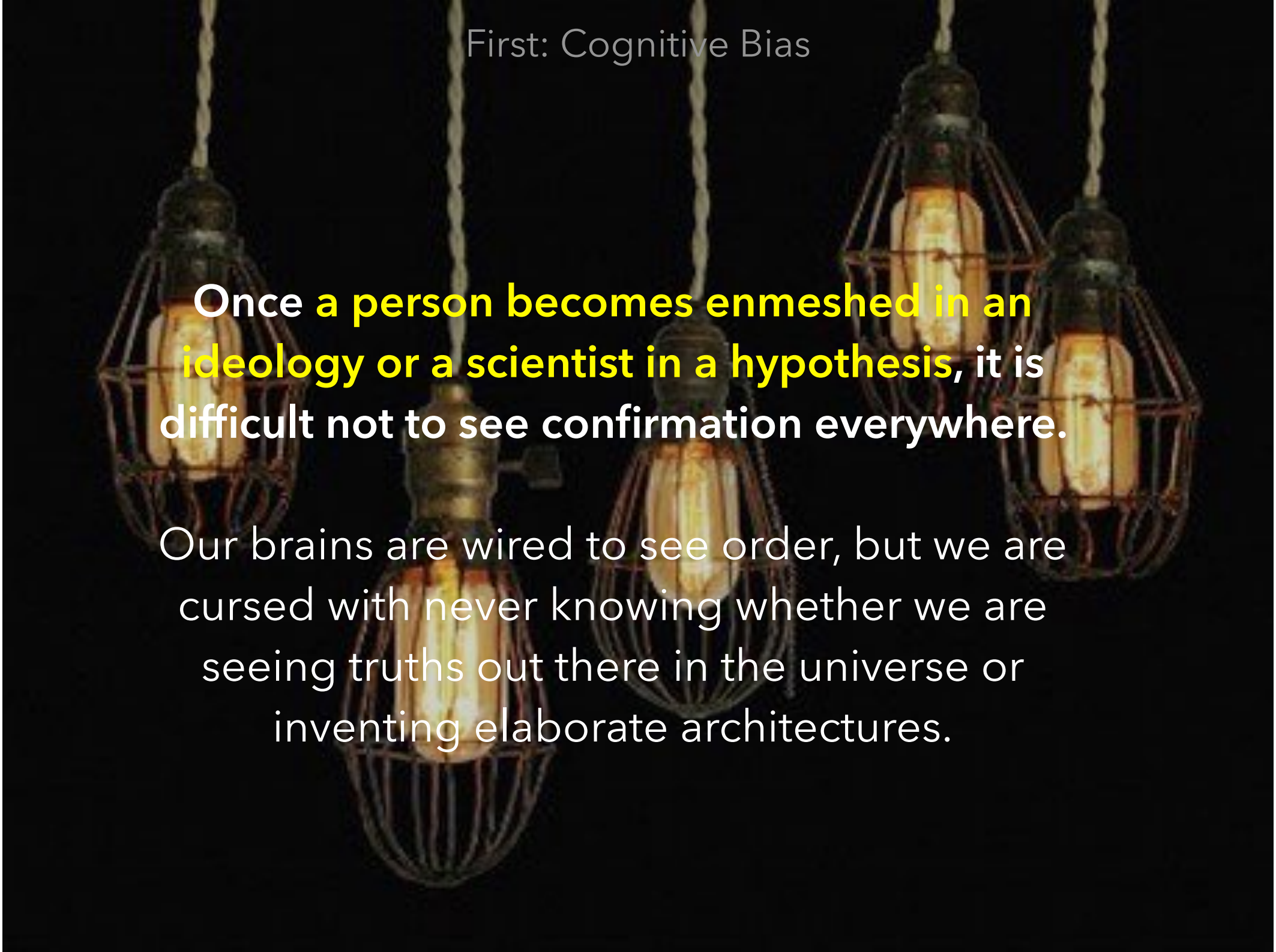


The first principle is that you must not fool yourself — and you are the easiest person to fool.

Richard Feynman



Psychologists have found that if you put people in a room with a contraption of lightbulbs wired to blink on and off at random, they will quickly discern what they believe are patterns, theories for predicting which bulb will be next to blink.



First: Cognitive Bias

Once a person becomes enmeshed in an ideology or a scientist in a hypothesis, it is difficult not to see confirmation everywhere.

Our brains are wired to see order, but we are cursed with never knowing whether we are seeing truths out there in the universe or inventing elaborate architectures.

Following your intuition can be misleading:

Monty Hall problem and Bayesian Reasoning.

First: Cognitive Bias

"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.

First: Cognitive Bias

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

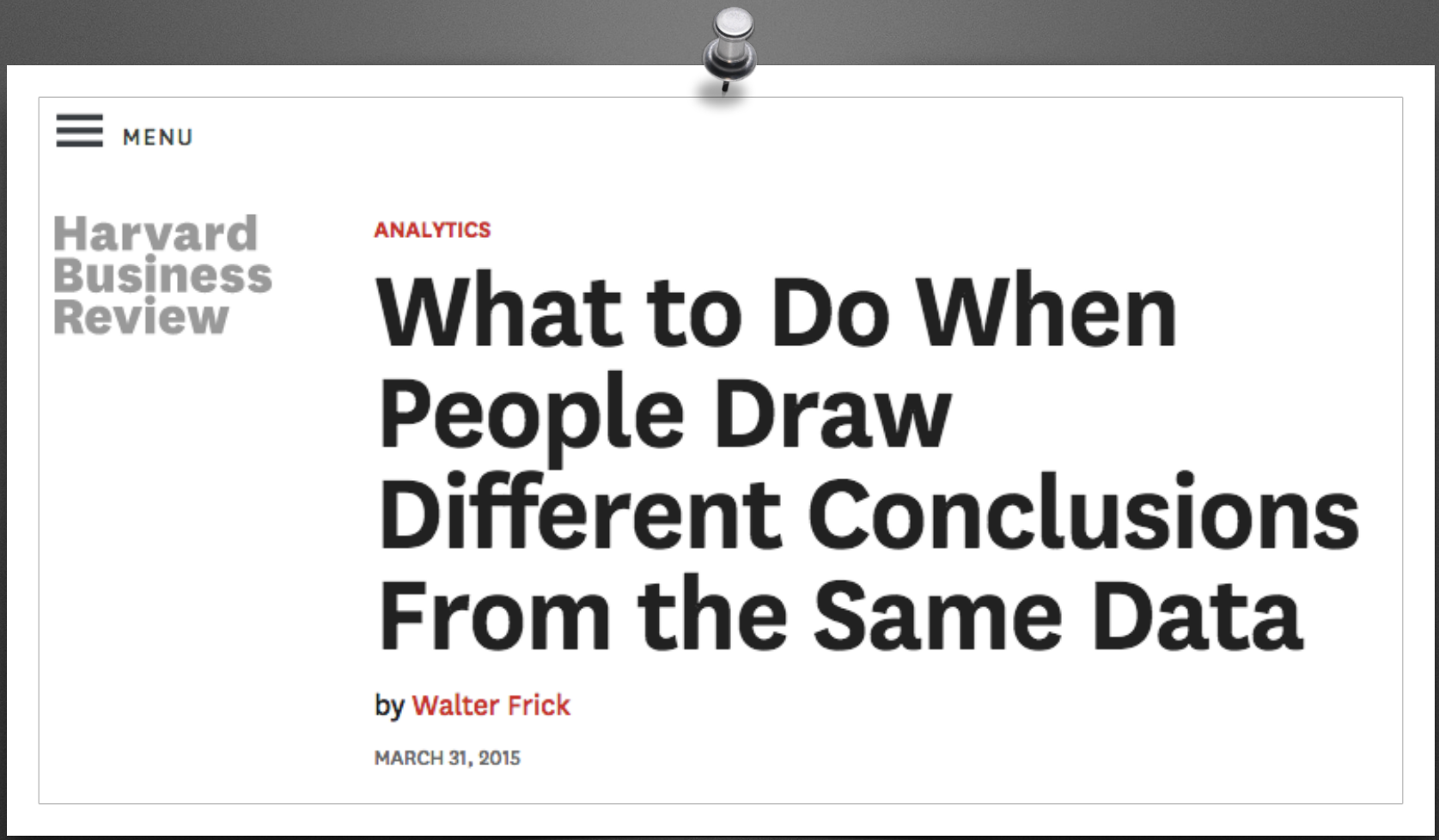
The two most common modes used for communicating results are *description* and *illustration*.

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.

First: Cognitive Bias



MENU

Harvard Business Review

ANALYTICS

What to Do When People Draw Different Conclusions From the Same Data

by Walter Frick

MARCH 31, 2015

The researchers recruited **61 analysts** (mostly academics) and asked them to assess **whether soccer referees were more likely to give red cards to players with darker skin tones**. The analysts split up into **29 teams**, and were given a dataset that included numerous variables about both players and referees.

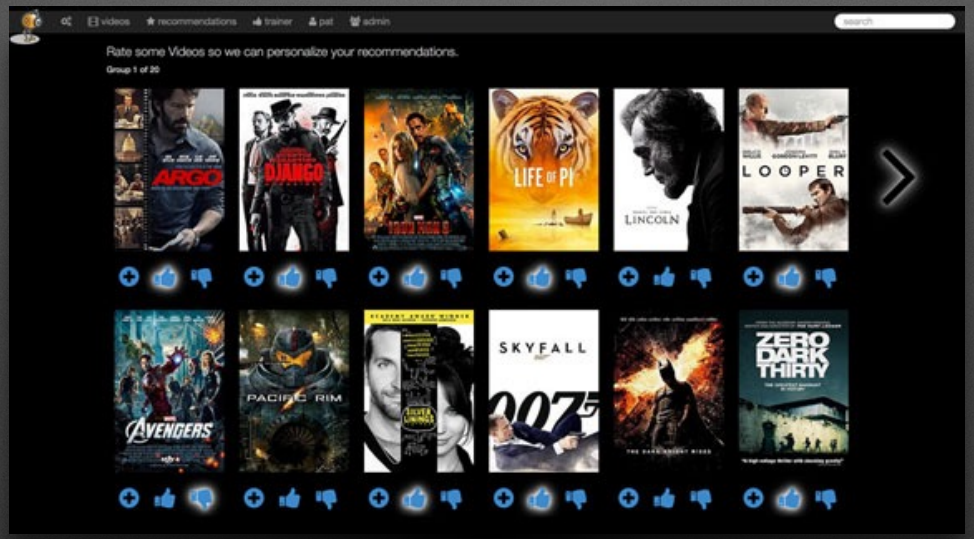
Each team devised their **method for answering the question, and then shared that approach** – but not any results – with the group. The result was a heated debate over which methods were defensible, and which were not. If you're looking for a correlation between skin tone and red cards received, does it make sense to control for the position the player plays? What about the country their team is located in, or how many yellow cards they've received?

From 29 teams came 21 different sets of variables. Different teams also used different statistical models.

Not surprisingly, then, they came to different conclusions. 20 of the teams found a statistically significant relationship between a player's skin color and the likelihood of receiving a red card. Nine teams found no significant relationship.

Algorithms are not always fair

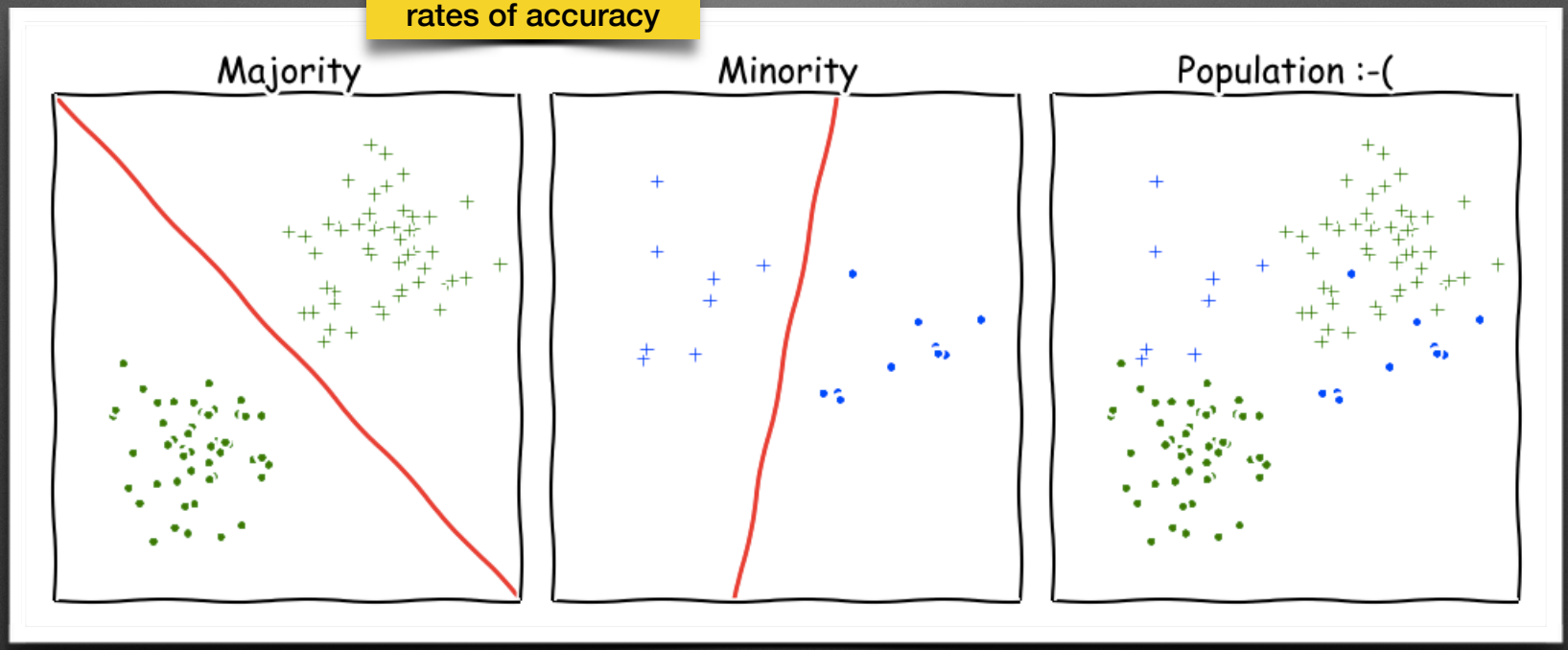
A 95% accurate movie recommender is a great algorithm, but it can be not fair…



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.

Algorithms are not always fair



Credit: Moritz Hardt

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

Second: Wrong models

Fact: Personalization is dangerous.

sign in | join us | search

jobs | dating | more | UK edition

theguardian
Winner of the Pulitzer prize

UK | world | sport | football | comment | culture | economy | lifestyle | fashion | environment | tech | money | travel | all sections

home › tech | games

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

Shares 10,565 | Comments 369

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

Advertisement
Next time you transfer money,
TransferWise.com
TransferWise
get started

Second: Wrong models

theguardian
Winner of the Pulitzer prize 2014

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

Second: Wrong models

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.

Third: Bad Data

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…..

Third: Bad Data

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

Third: Bad Data

"**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."

Third: Bad Data

Fact:

One summer, 132 cats were brought to Manhattan Animal Medical Center, they had fallen out of open windows. Of the 132 cats, 128 survived, some from 32 story falls.

Interpretation:

The New York Times wrote of the miraculous survival instincts of cats; the ability to “adjust rotation orientation” during a fall, “flying-squirrel” aerodynamics, joints and muscles acting as “shock absorbers.” The media discussion lasted for months. *"On Landing Like a Cat - It's a Fact" - New York Times, August 22, 1989.*

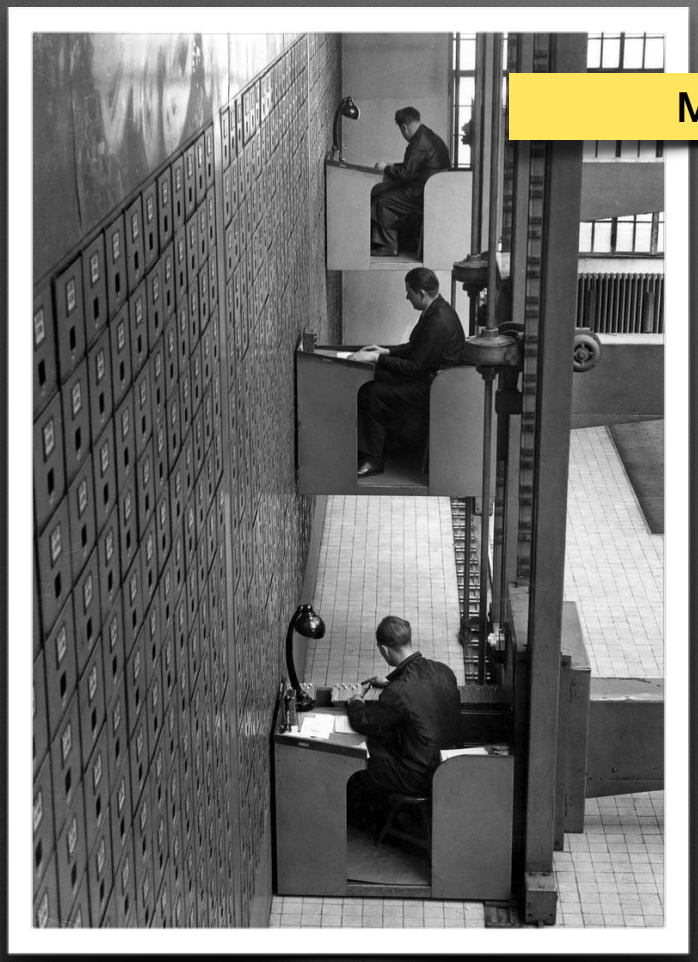
Third: Bad Data

The data point that ended the Super Cat conversation involved data sampling. One woman interviewed said, *"my poor cat must have been the exception. She fell out our 9th story window and died. Of course, I did not bring her body to the hospital, nor did I report it…"*

Use of convenience data

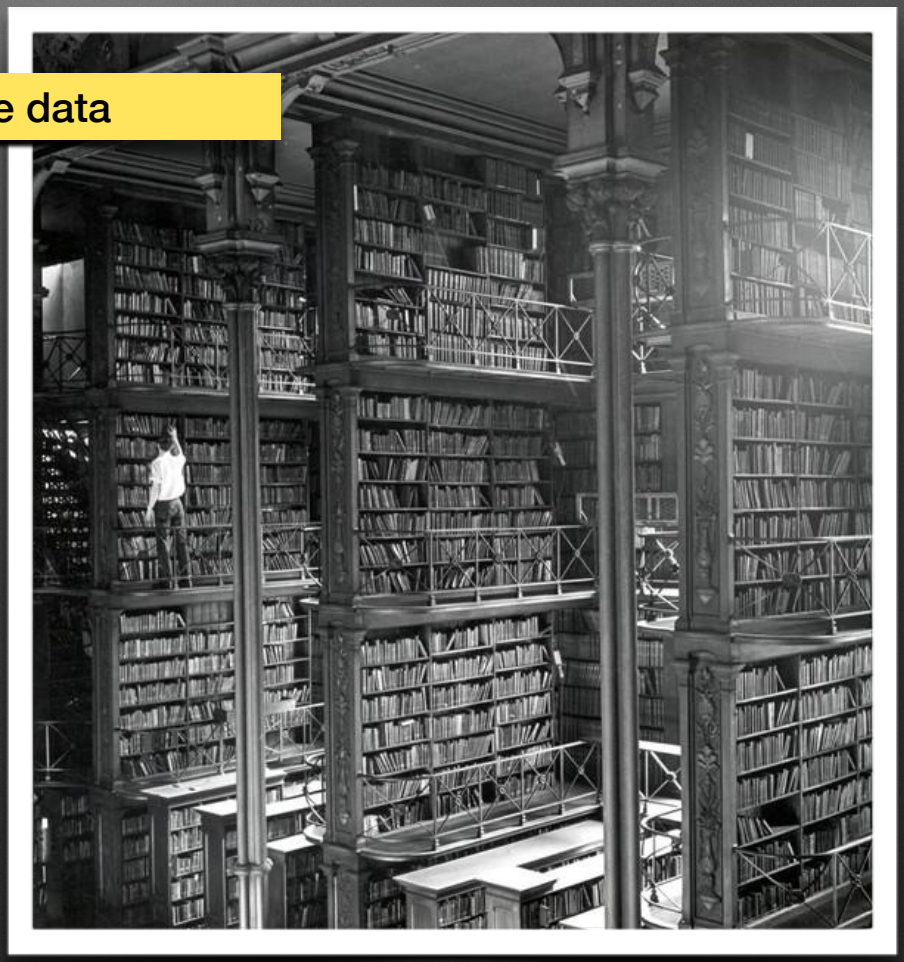
The **original survey had only included cats that had survived their falls**. When cats that died on impact were included the study it was no longer newsworthy.

Is it always better to have more samples (Big N) and more features (Big P)?



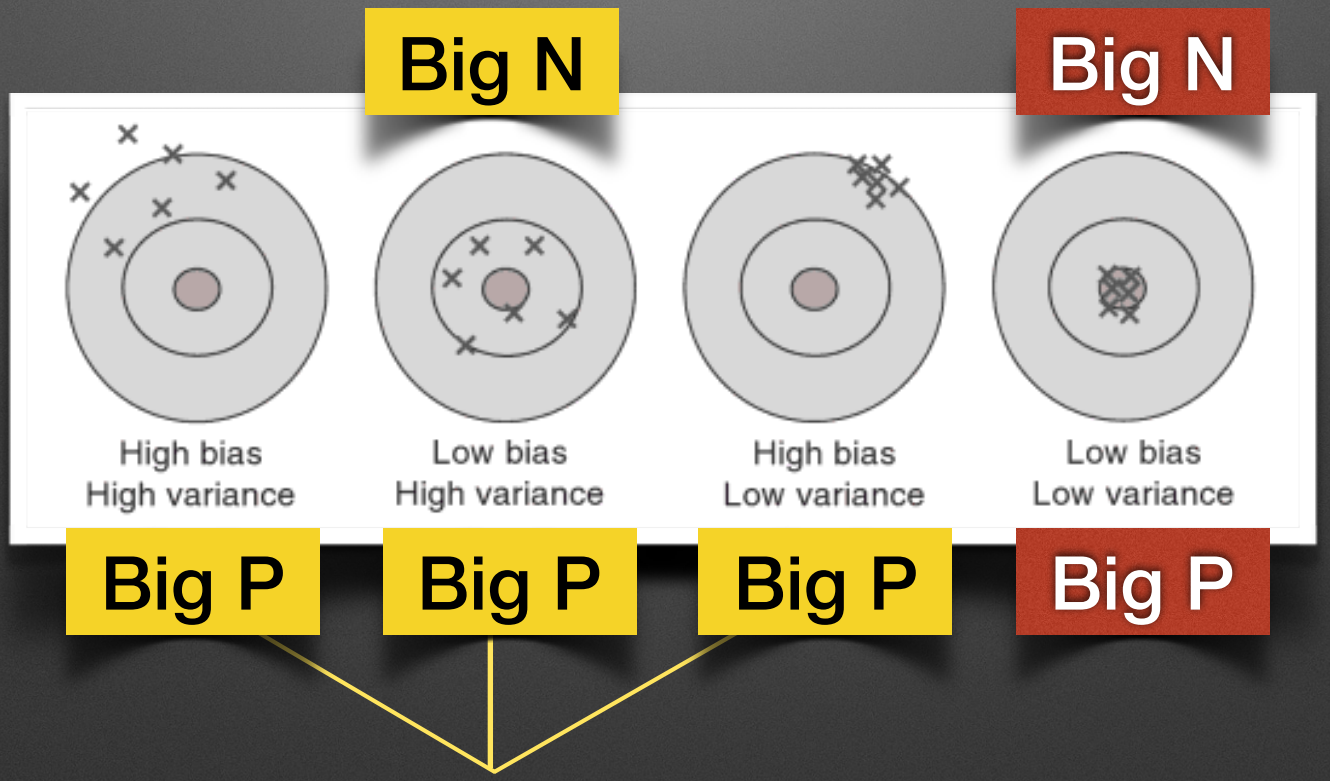
Archivists, France, 1937.

More data



The Public Library of Cincinnati, Ohio (built in 1874)

Is it always better to have more samples (Big N) and more features (Big P)?

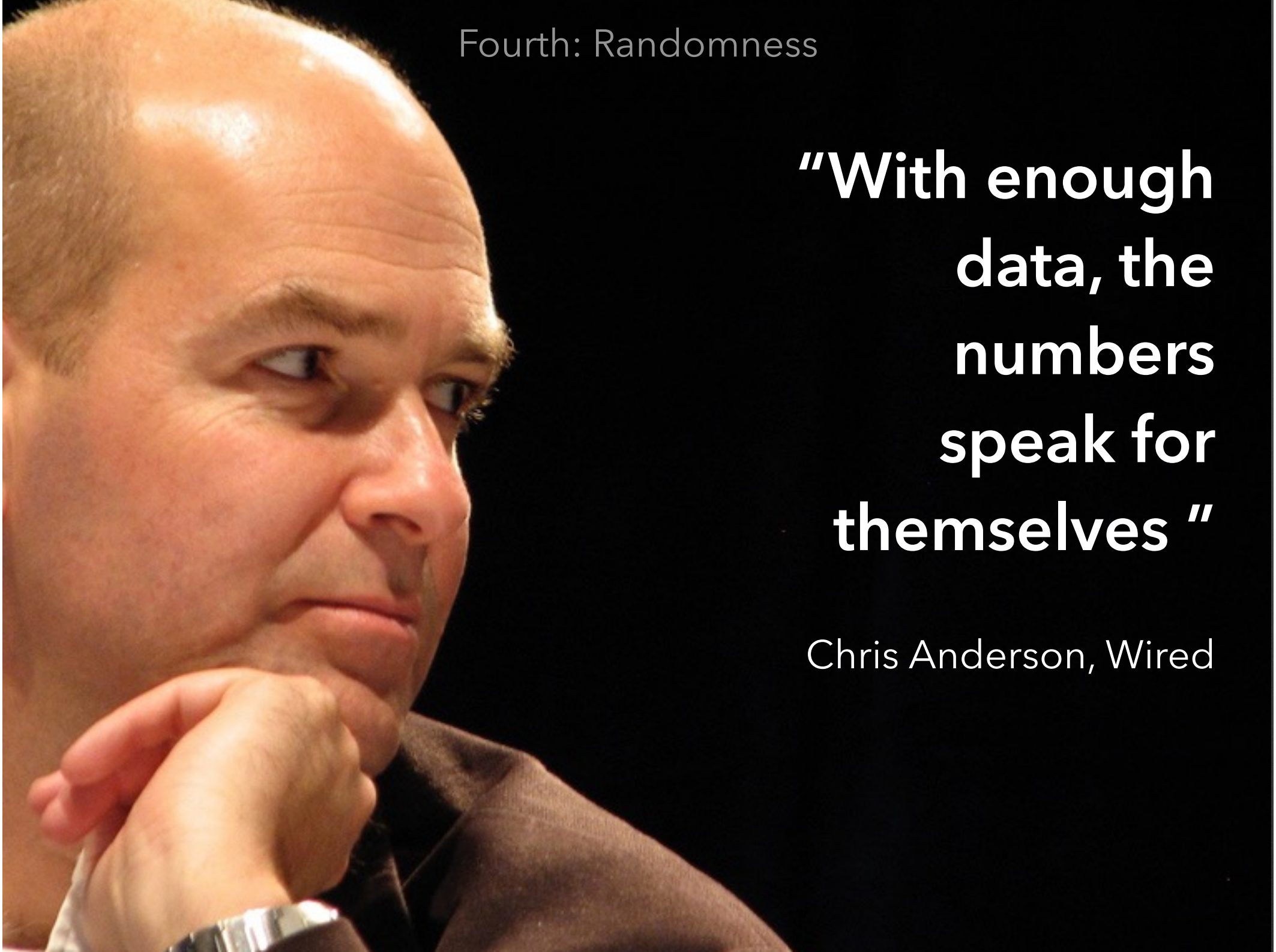


If they are meaningful!

Fourth: Randomness

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!**

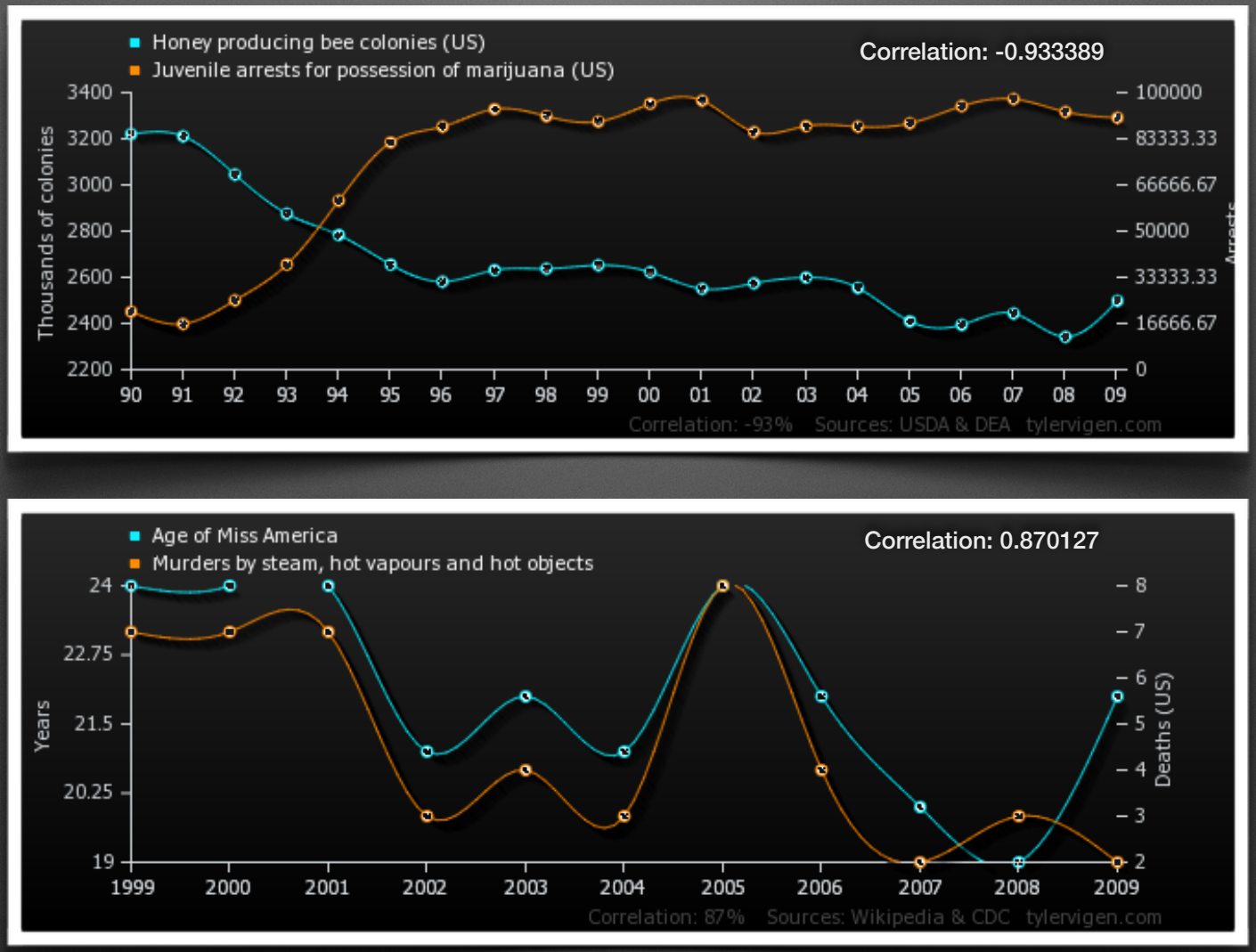


“With enough data, the numbers speak for themselves ”

Chris Anderson, Wired


“The numbers have no way of speaking for themselves”
Nate Silver

Fourth: Randomness



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

Fourth: Randomness

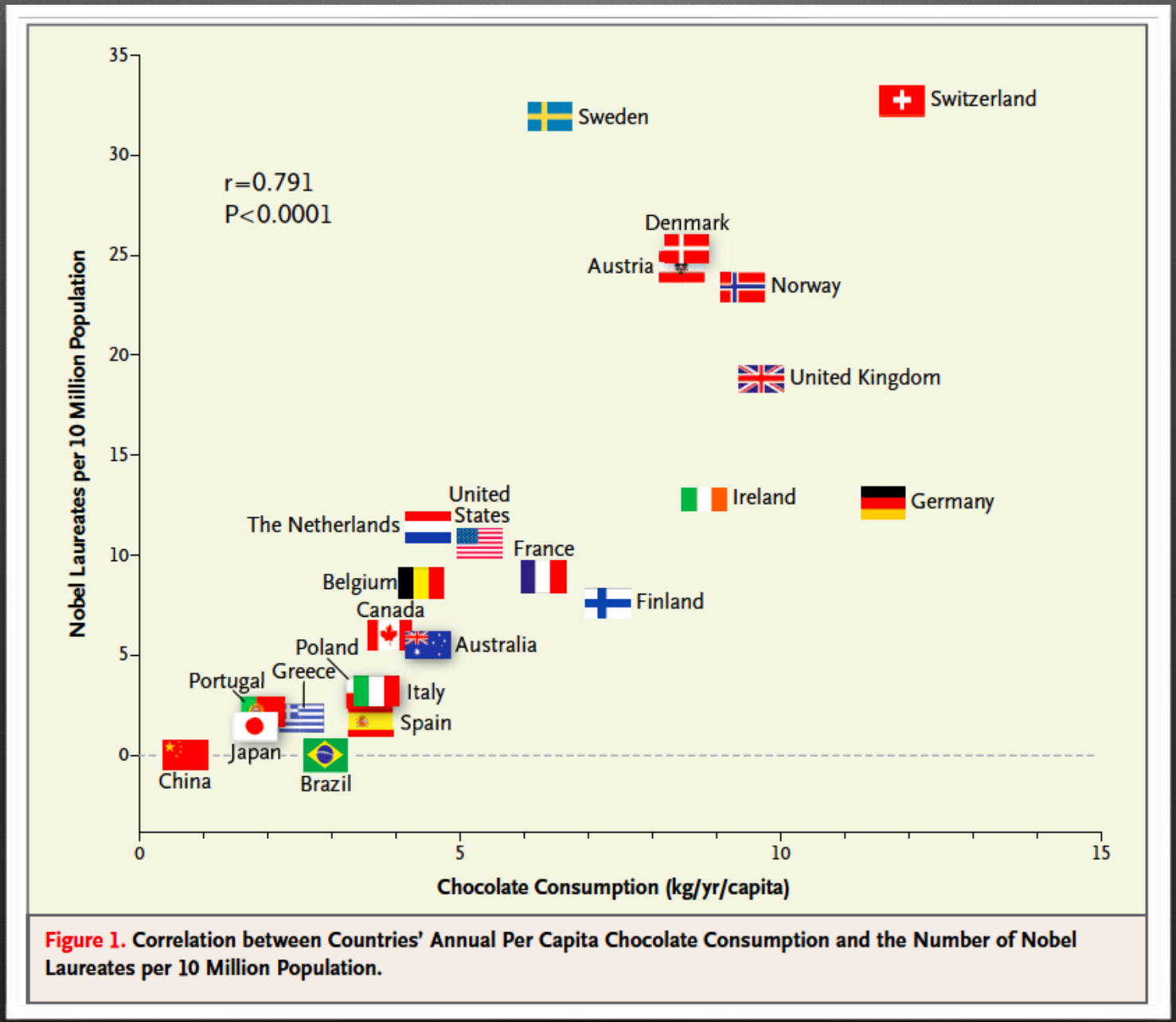


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Fourth: Randomness

INTRODUCING: NEJM Knowledge+
A new way to learn – watch the video >>

Welcome | Renew, Subscribe or Create Account | Sign In

The NEW ENGLAND JOURNAL of MEDICINE

SUBSCRIBE OR RENEW
Includes NEJM iPad Edition, 20 FREE Online CME Exams and more >>

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME | Keyword, Title, Author, or Citation | Advanced Search

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

N Engl J Med 2012; 367:1562-1564 | October 18, 2012 | DOI: 10.1056/NEJMon1211064

Share:

Article | References | Citing Articles (20)

Dietary flavonoids, abundant in plant-based foods, have been shown to improve cognitive function. Specifically, a reduction in the risk of dementia, enhanced performance on some cognitive tests, and improved cognitive function in elderly patients with mild impairment have been associated with a regular intake of flavonoids.[1,2] A subclass of flavonoids called flavanols, which are widely present in cocoa, green tea, red wine, and some fruits, seems to be effective in slowing down or even reversing the reductions in cognitive performance that occur with aging. Dietary flavanols have also been shown to improve endothelial function and to lower blood pressure by causing vasodilation in the peripheral vasculature and in the brain.[3,4] Improved cognitive performance with the administration of a cocoa polyphenolic extract has even been reported in aged Wistar–Unilever rats.[5]

Since chocolate consumption could hypothetically improve cognitive function not only in individuals but also in whole populations, I wondered whether there would be a correlation between a country's level of chocolate consumption and its population's cognitive function. To my knowledge, no data on overall national cognitive function are publicly available. Conceivably, however, the total number of Nobel laureates per capita could serve as a surrogate end point reflecting the proportion with superior cognitive function and thereby give us some measure of the overall cognitive function of a given country.

Access Provided By:
CRAI UNIVERSITAT DE BARCELONA

TOOLS
PDF | Print | Download Citation | Supplementary Material | E-Mail | Save | Article Alert | Reprints | Permissions | Share/Bookmark

TOPICS
Diet/Nutrition
Neurology/ Neurosurgery

MORE IN
Commentary
October 18, 2012

TRENDS
Most Viewed (Last Week)

IMAGES IN CLINICAL MEDICINE
Occipital Calcification and Celiac Disease
[35,960 views]
April 17, 2014 | R.G. Cury and C.H. Moreira

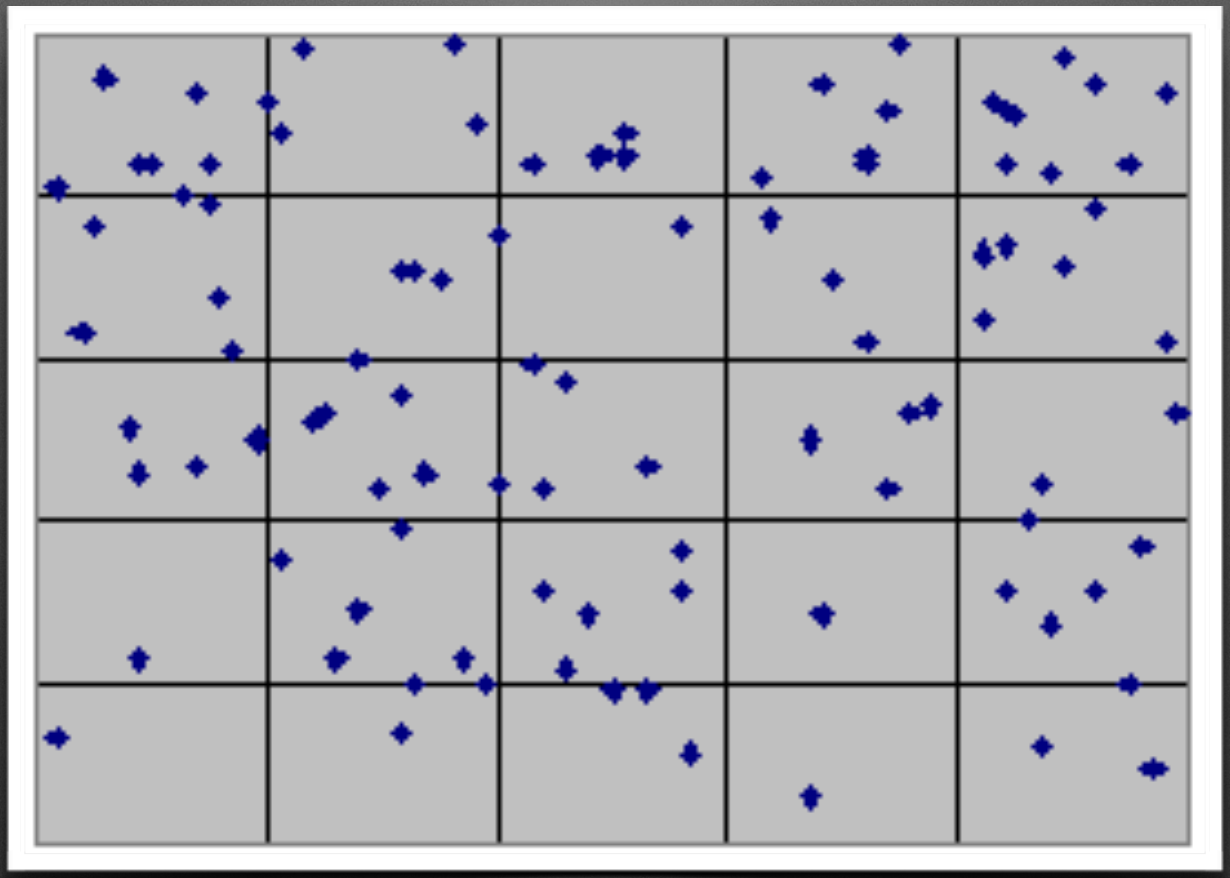
Fourth: Randomness

Randomness does not mean absence of structure or partial order.

Random means unpredictable.

Fourth: Randomness

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.

Fourth: Randomness

In 2012 Professor Kahneman wrote:

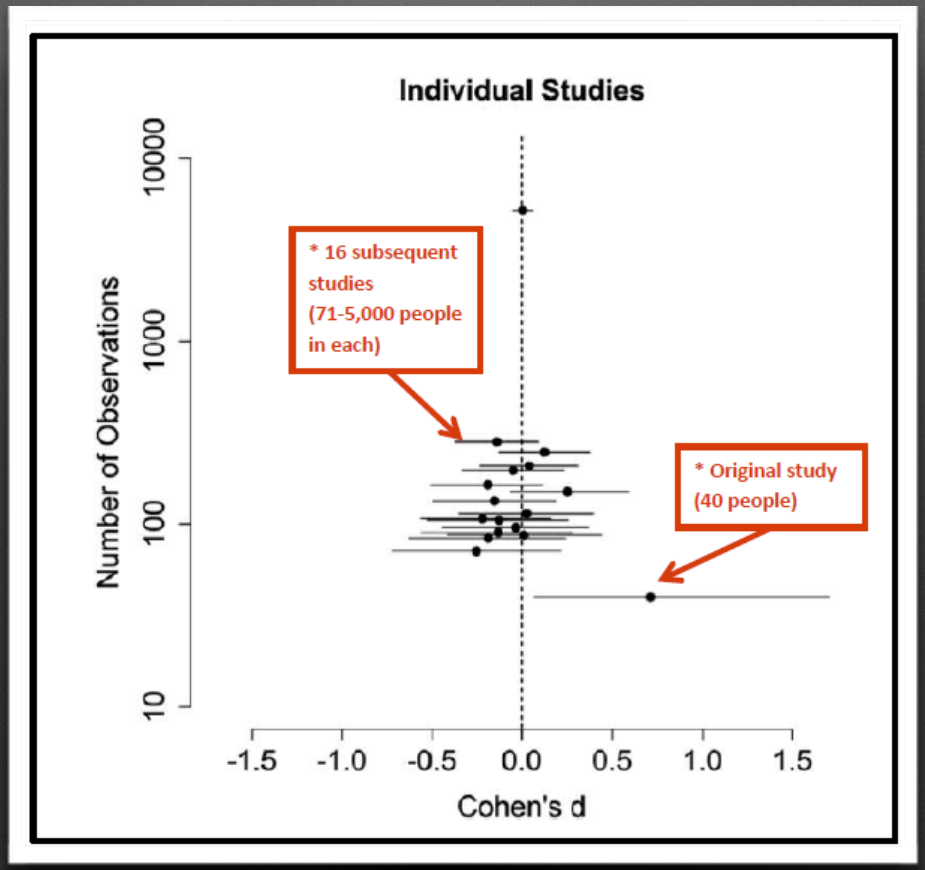
"90% of the students who saw the CRT in normal font made at least one mistake in the test, but the proportion dropped to 35% when the font was barely legible. You read this correctly: performance was better with the bad font."

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

Fourth: Randomness

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

Fourth: Randomness

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

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

Unfair use of data.

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

Use of bad features.

Credit denial because of racial identity.

Fifth: Ethics

Privacy.

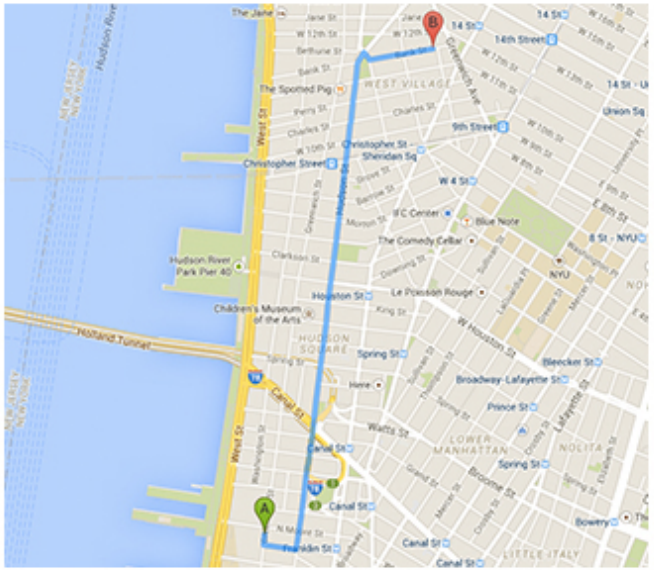
Privacy.

Public NYC Taxicab Database Lets You See How Celebrities Tip

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

134,190 18

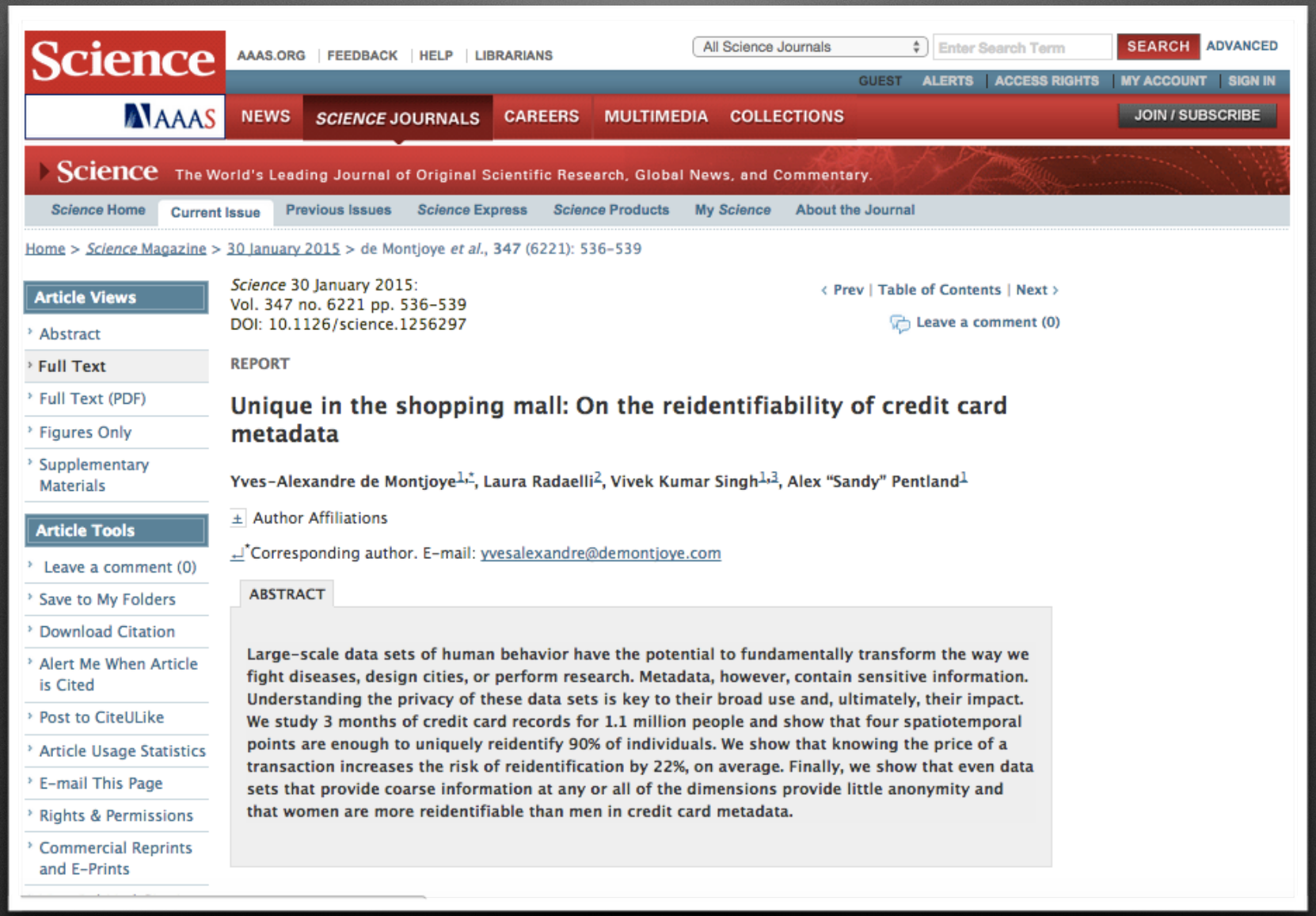
BRADLEY COOPER



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

Fifth: Ethics

Privacy.



Science AAAS.ORG | FEEDBACK | HELP | LIBRARIANS
All Science Journals | Enter Search Term | SEARCH | ADVANCED
GUEST | ALERTS | ACCESS RIGHTS | MY ACCOUNT | SIGN IN
NEWS | SCIENCE JOURNALS | CAREERS | MULTIMEDIA | COLLECTIONS | JOIN / SUBSCRIBE
Science The World's Leading Journal of Original Scientific Research, Global News, and Commentary.
Science Home | Current Issue | Previous Issues | Science Express | Science Products | My Science | About the Journal
Home > Science Magazine > 30 January 2015 > de Montjoye et al., 347 (6221): 536-539

Article Views
Abstract
Full Text
Full Text (PDF)
Figures Only
Supplementary Materials
Article Tools
Leave a comment (0)
Save to My Folders
Download Citation
Alert Me When Article is Cited
Post to CiteULike
Article Usage Statistics
E-mail This Page
Rights & Permissions
Commercial Reprints and E-Prints

Science 30 January 2015:
Vol. 347 no. 6221 pp. 536-539
DOI: 10.1126/science.1256297

‹ Prev | Table of Contents | Next ›
Leave a comment (0)

REPORT

Unique in the shopping mall: On the reidentifiability of credit card metadata

Yves-Alexandre de Montjoye[1,*], Laura Radaelli[2], Vivek Kumar Singh[1,3], Alex "Sandy" Pentland[1]

± Author Affiliations

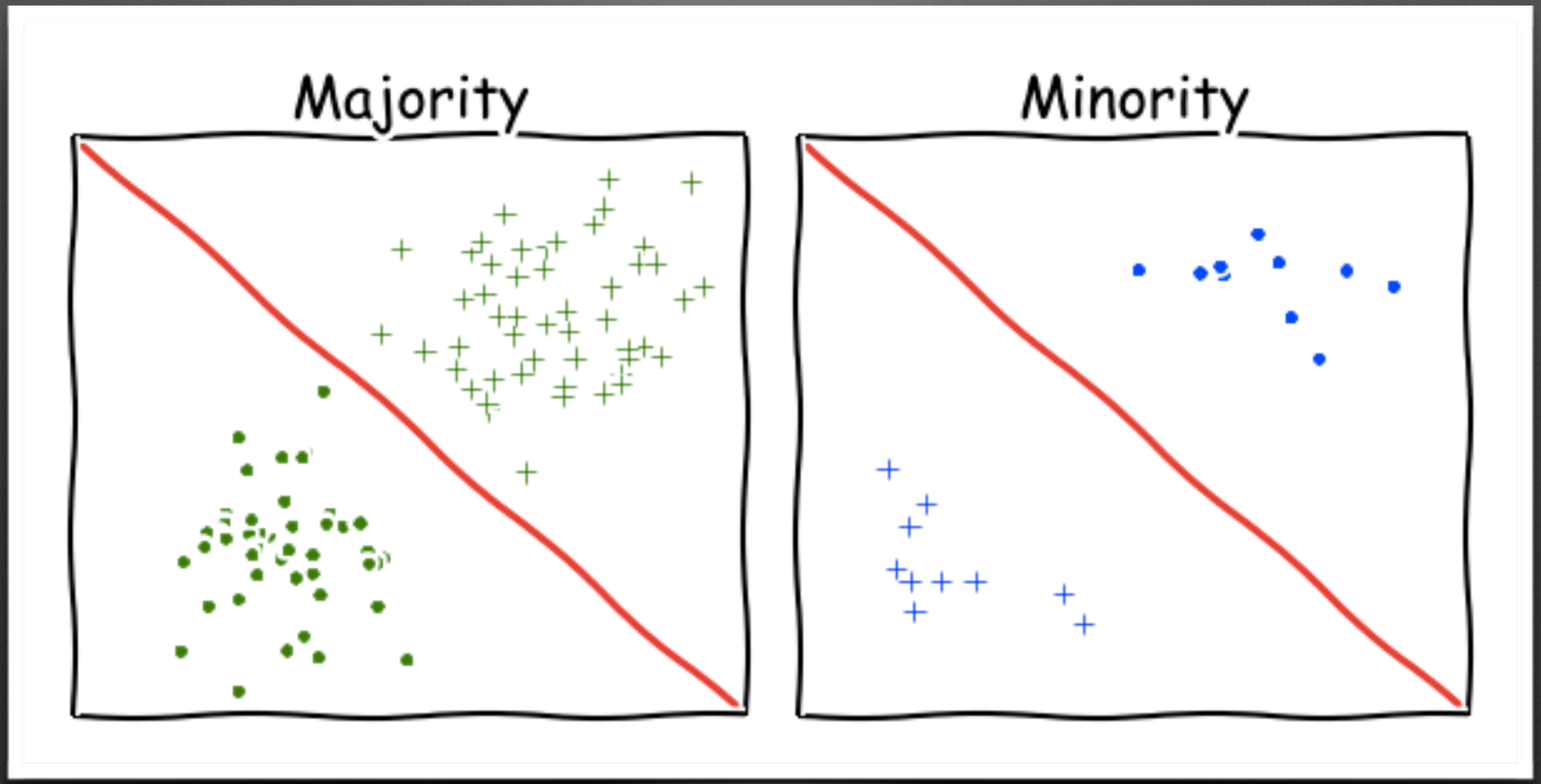
*Corresponding author. E-mail: yvesalexandre@demontjoye.com

ABSTRACT

Large-scale data sets of human behavior have the potential to fundamentally transform the way we fight diseases, design cities, or perform research. Metadata, however, contain sensitive information. Understanding the privacy of these data sets is key to their broad use and, ultimately, their impact. We study 3 months of credit card records for 1.1 million people and show that four spatiotemporal points are enough to uniquely reidentify 90% of individuals. We show that knowing the price of a transaction increases the risk of reidentification by 22%, on average. Finally, we show that even data sets that provide coarse information at any or all of the dimensions provide little anonymity and that women are more reidentifiable than men in credit card metadata.

Fifth: Ethics

Inclusion.

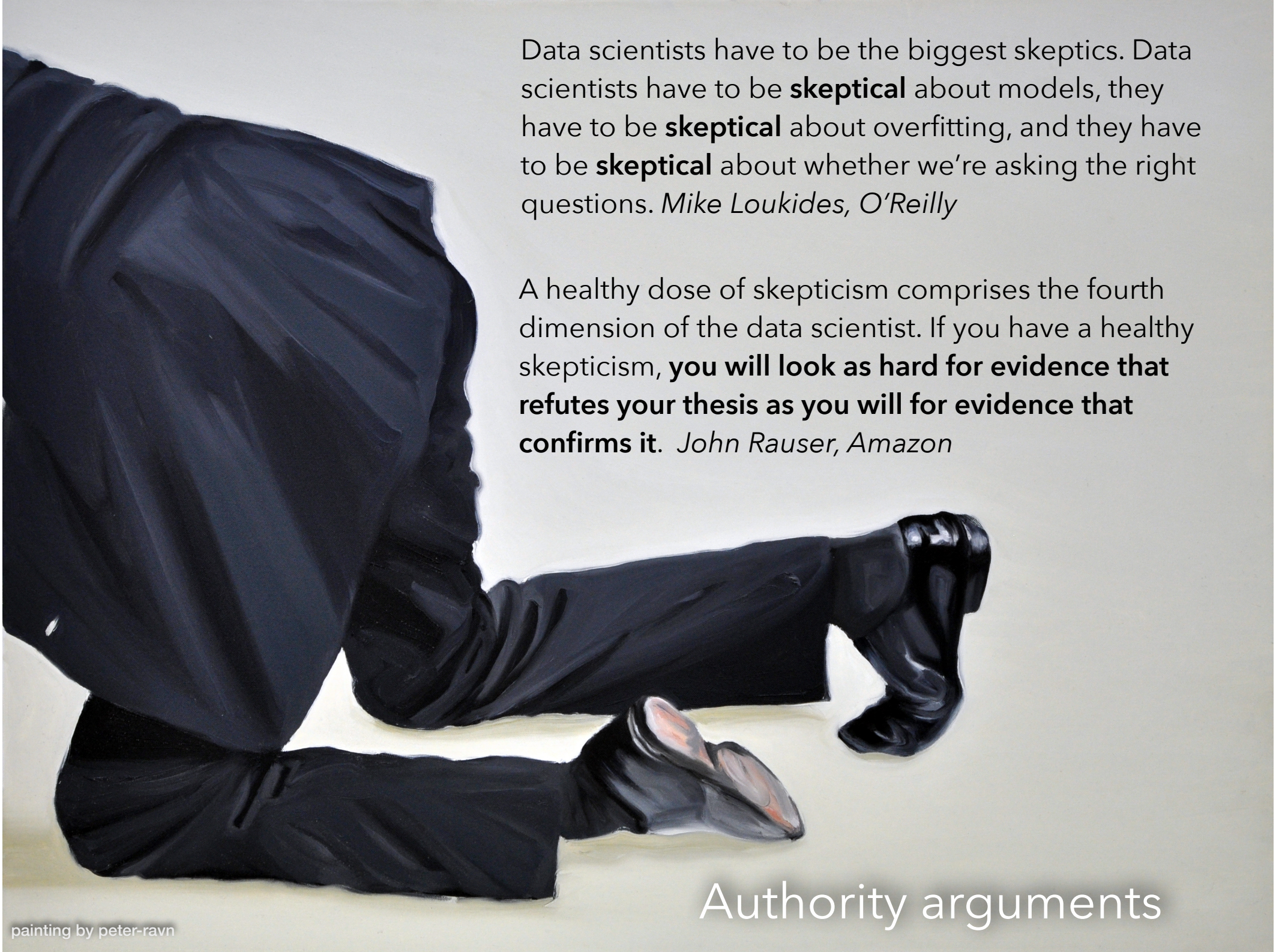


Credit: Moritz Hardt

Authority arguments

painting by peter-ravn

(…) Whereas in finance we need to worry about models manipulating the market, in data science we need to worry about models manipulating people, which is in fact scarier. Modelers, if anything, have a bigger responsibility now than ever before. *Cathy O'Neal, Mathbabe*



Data scientists have to be the biggest skeptics. Data scientists have to be **skeptical** about models, they have to be **skeptical** about overfitting, and they have to be **skeptical** about whether we're asking the right questions. *Mike Loukides, O'Reilly*

A healthy dose of skepticism comprises the fourth dimension of the data scientist. If you have a healthy skepticism, **you will look as hard for evidence that refutes your thesis as you will for evidence that confirms it**. *John Rauser, Amazon*

painting by peter-ravn

Authority arguments

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