Day 1: Fundamental data and statistical literacy

Data science, statistical reasoning, and policy-making

Simon Munzert
Hertie School

Table of contents

Welcome!

Introductions

Me

I'm Simon Munzert [si'mən munsɜrt], or just Simon [saɪmən].

munzert@hertie-school.org

Professor of Data Science and Public Policy | Director of the Data Science Lab

You

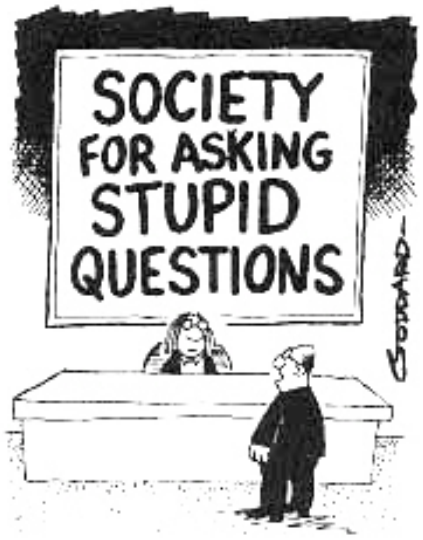
What's your name and position?

What has been your exposure to data science and statistics?

Do you have experience with quantitative evidence in your position?

Workshop etiquette

- We have a lot of ground to cover. I brought plenty of material, but ultimately you have to **signal where your interests and needs are**. I am happy to go deeper, digress, or shift attention to other topics and examples (as long as they're in my comfort zone).
- Obviously, I am all but an expert in Georgian politics or policy issues that are currently salient. For sound evidence-based reasoning about policies, domain knowledge is key, and is explicitly part of many of the data-based tools and methods we will discuss. Please **bring your own knowledge and experience to the table**.
- Please take the opportunity to **ask questions anytime**. Some of the topics might lead you out of your comfort zone. But there are no stupid questions to ask, so please bring them on.



"Excuse me, but is this The Society for Asking Stupid Questions?"

What is data science?

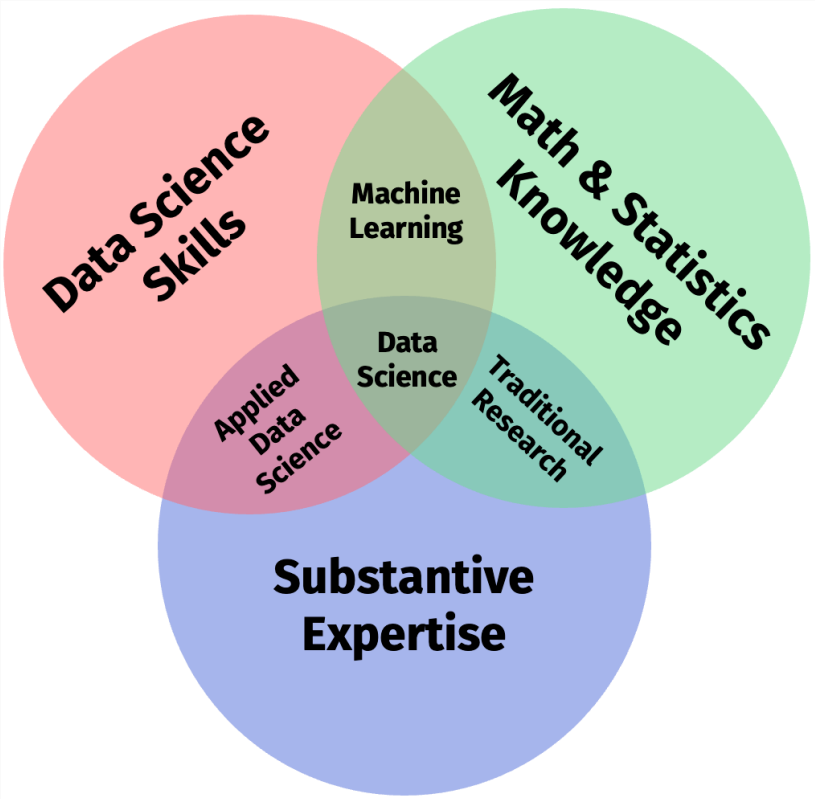
What is data science?

What is data science?

> "Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data." - Wikipedia
>
> "Data science is a concept to unify statistics, data analysis, informatics, and their related methods to understand and analyze actual phenomena with data." - Chikio Hayashi

Overall, there's **no consensus** - it is a buzzword after all. We're going to carry on with Conway's working definition.

A working definition



`Source` Drew Conway, 2010 (adapted)

Types of data-driven research and their role for policy

Hertie School

1. Description

- What is the state of the world?
- What are the trends over time?
- What are the differences between groups?

2. Explanation

- What is the effect of a policy?
- Does the effect vary across groups?
- What are the mechanisms behind the effect?

3. Prediction

- What is the path of an indicator?
- (When) will future events happen?
- What class does this observation most likely belong to?

The value for policy-making

- At the center of **monitoring**
- "How many people consume misinformation online?"
- "How many people are unemployed in a certain district?"
- "How does the distribution of income vary across educational segments of the population?"

The value for policy-making

- At the center of **evaluation**
- "Did the minimum wage increase lead to a decrease in employment?"
- "Did the campaign affect the exposure to misinformation differently across groups?"
- "Why did the intervention not lead to the expected results?"

The value for policy-making

- At the center of **forecasting** but also **targeting** and **measurement**
- "Will there be conflict?"
- "How many people will be unemployed in a certain district next year?"
- "Which individuals are most likely to be affected by a policy?"

The data science pipeline

The data science pipeline

Preparatory work

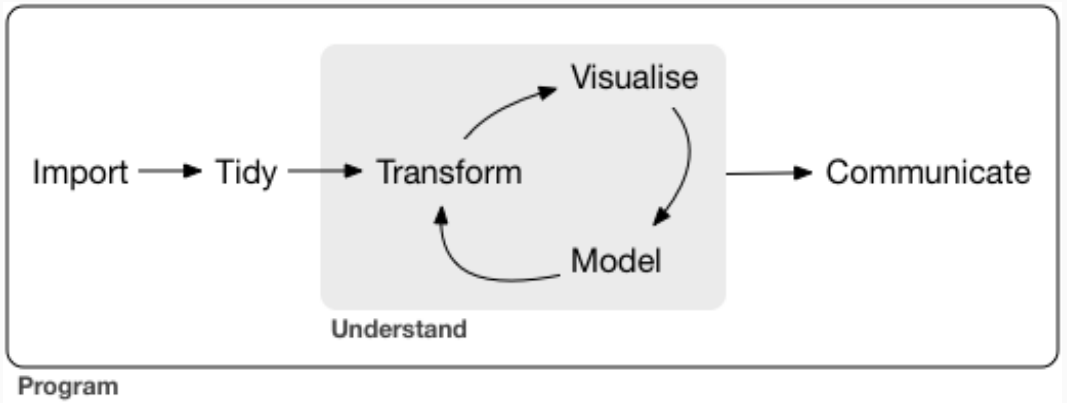
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation



`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

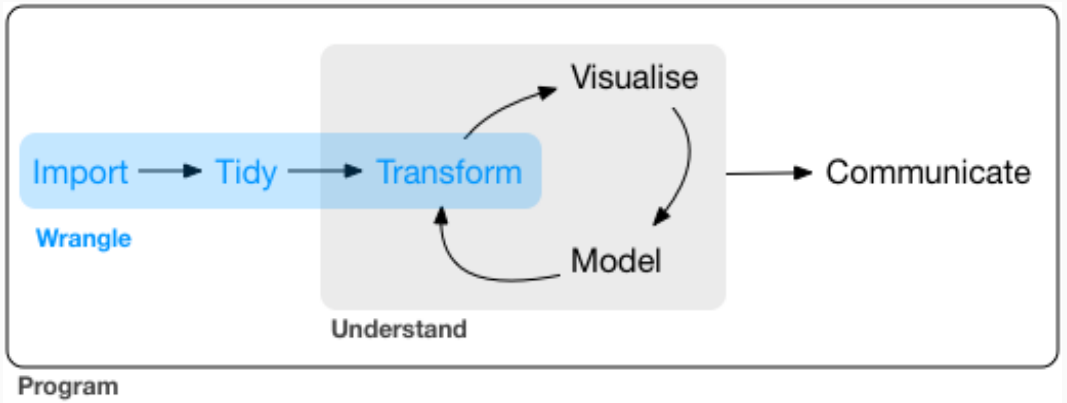
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate



`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

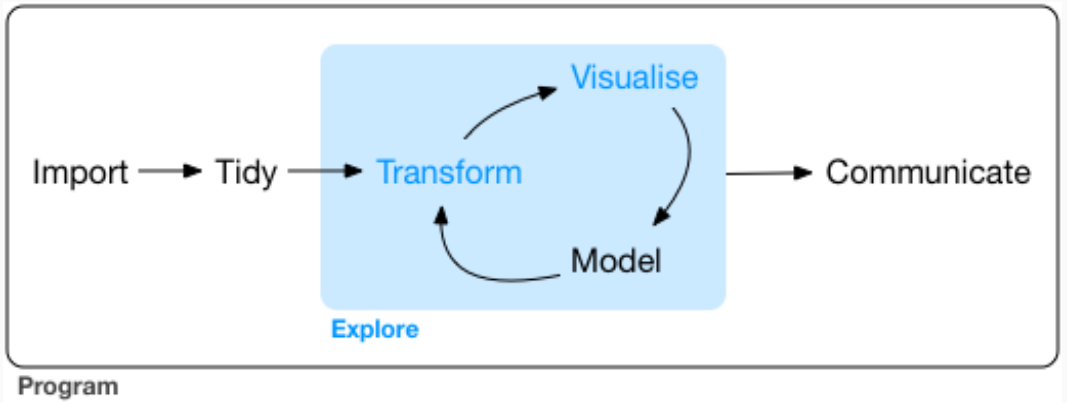
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover



`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

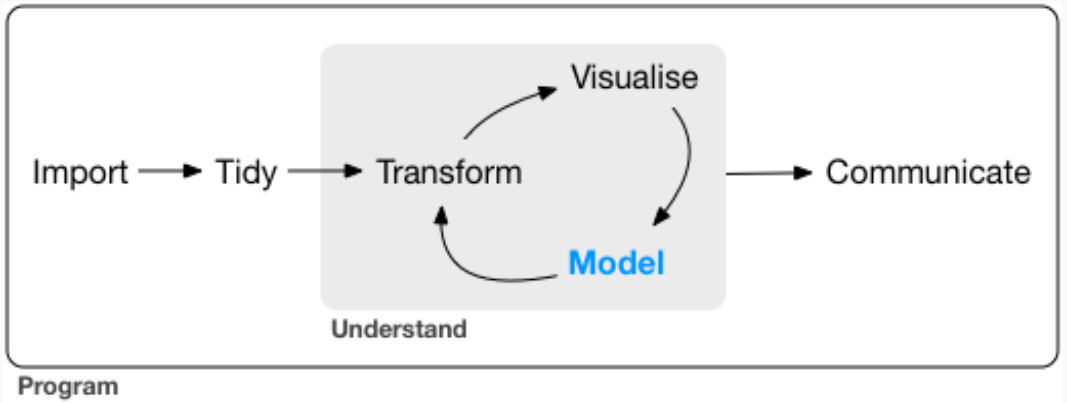
The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict



`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

The data science pipeline

Preparatory work

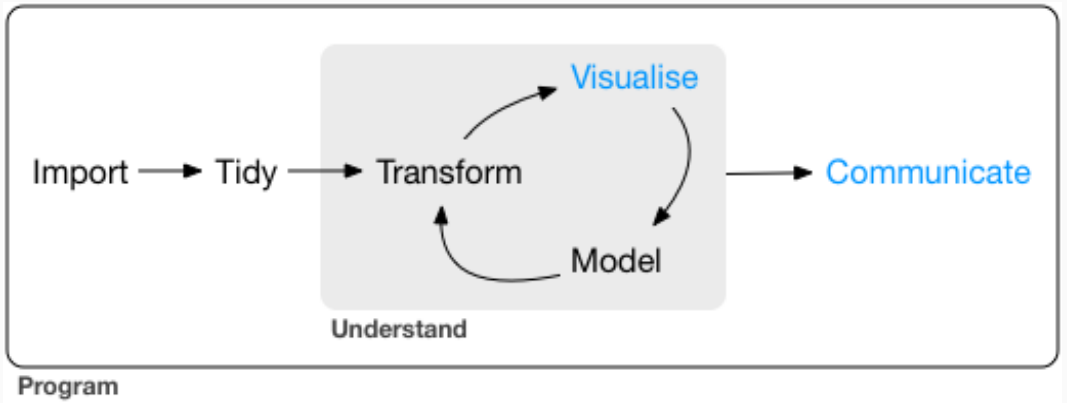
- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable



`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

The data science pipeline

Preparatory work

- **Problem definition** predict, infer, describe
- **Design** conceptualize, build data collection device
- **Data collection** recruit, collect, monitor

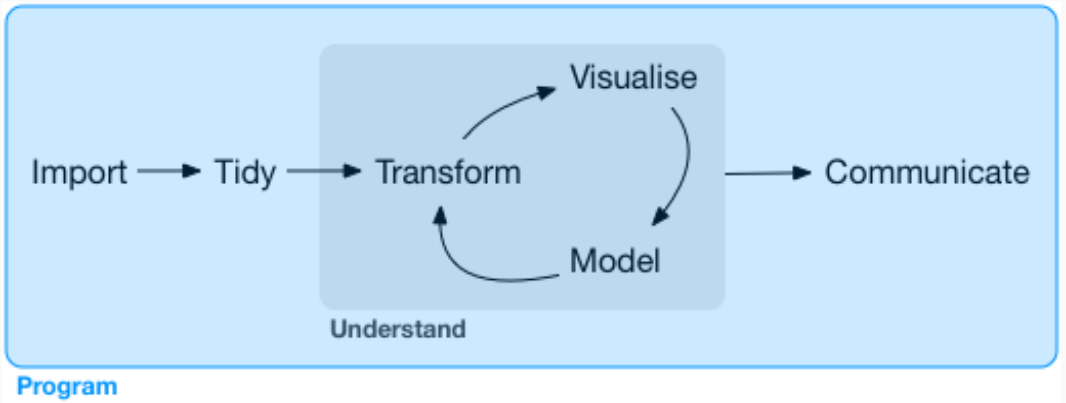
Data operation

- **Wrangle**: import, tidy, manipulate
- **Explore**: visualize, describe, discover
- **Model**: build, test, infer, predict

Dissemination

- **Communicate**: to the public, media, policymakers
- **Publish**: journals/proceedings, blogs, software
- **Productize**: make usable, robust, scalable

Meta skill: programming

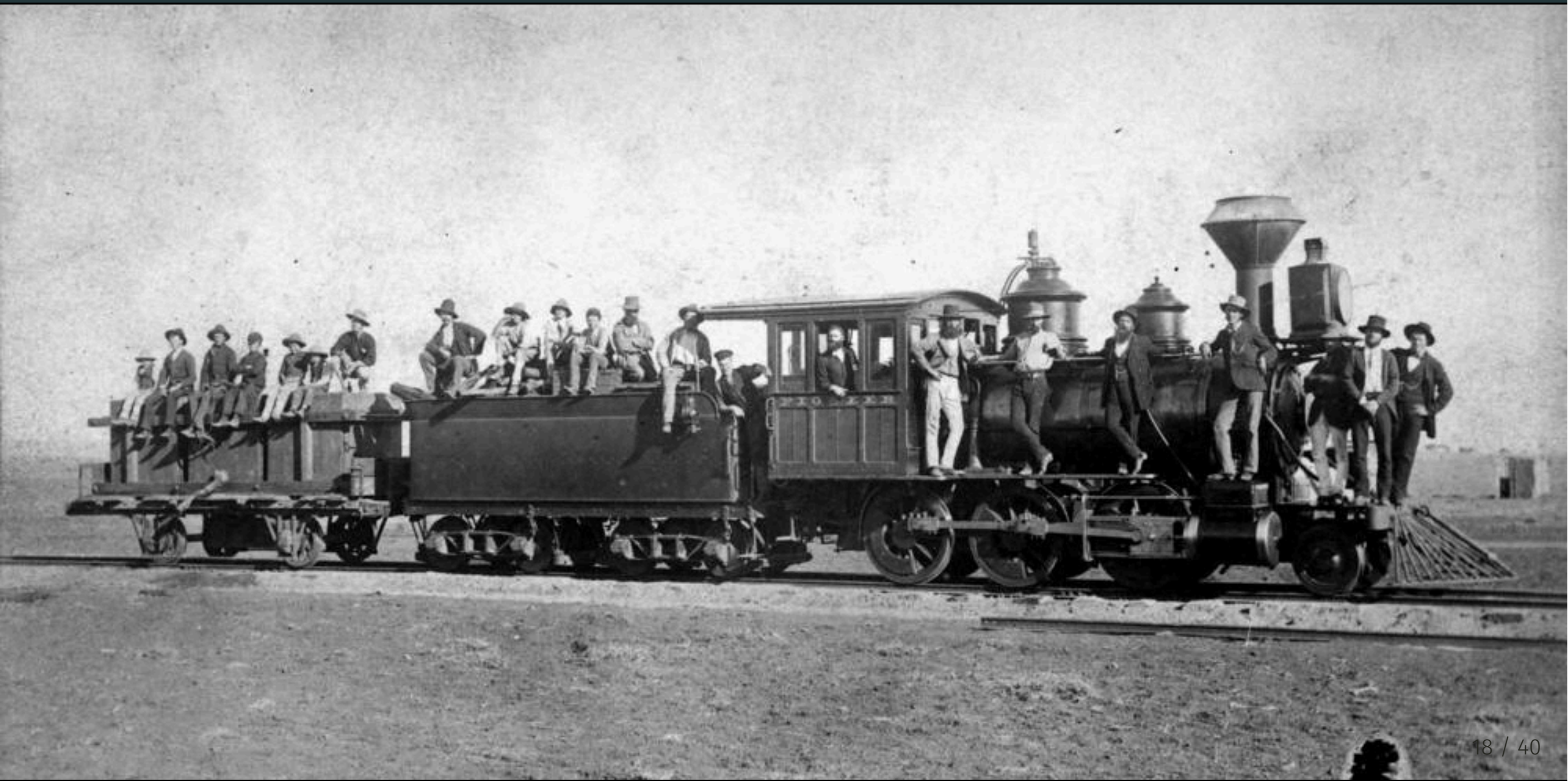


`Source` H. Wickham, M. Çetinkaya-Rundel, G. Gronemund: R for Data Science

(Data) science for public policy

Some examples of data science research informing policy

Hertie School



Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

The Billion Prices Project: Using Online Prices for Measurement and Research

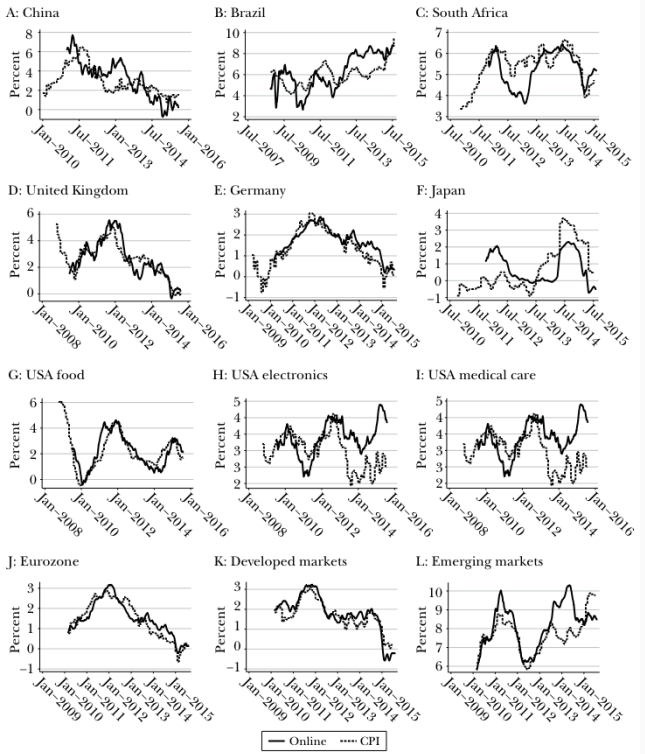
Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as "Big Data," have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate

Figure 4

Online versus Consumer Price Index (CPI) Annual Inflation Rates



Source: Authors using online price indexes computed by PriceStats and consumer price indexes sourced from the national statistical office in each country.

Notes: Figure 4 compares inflation as measured by online prices and by the offline prices in the official consumer price index for a selection of countries, sectors, and regions. Annual inflation rates for daily online price indexes are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. The series are nonseasonally adjusted. Indexes are "all-items" with the exception of China, where an online supermarket index is shown next to the official food index. Global aggregates in the last row are computed using 2010 consumption weights in each country and CPIs from official sources.

See also: https://thebillionpricesproject.com/ and https://www.pricestats.com/

Journal of Economic Perspectives—Volume 30, Number 2—Spring 2016—Pages 151–178

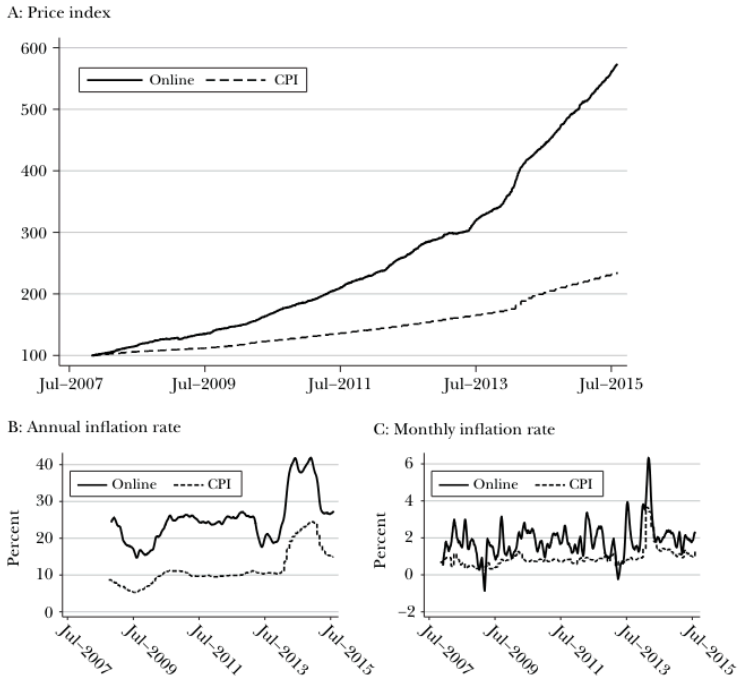
The Billion Prices Project: Using Online Prices for Measurement and Research

Alberto Cavallo and Roberto Rigobon

New data-gathering techniques, often referred to as "Big Data," have the potential to improve statistics and empirical research in economics. This paper presents one example of how this can be achieved by using the vast number of online prices displayed on the web. We describe our work with the Billion Prices Project at MIT, and emphasize key lessons that can be used for both inflation measurement and some fundamental research questions in macro and international economics. In particular, we show how online prices can be used to construct daily price indexes in multiple countries and to avoid measurement biases that distort evidence of price stickiness and international relative prices.

The basic procedure used in most countries to collect inflation data has remained roughly the same for decades. A large number of people working for national statistical offices visit hundreds of stores on a monthly or bimonthly basis to collect prices for a preselected basket of goods and services. The micro data are then processed and used to construct consumer price indexes and other related indicators. This process is expensive, complex, and often too slow for some users of the data. Infrequent sampling and slow updates to the baskets can complicate

Figure 1
Argentina



Source: Authors using online price index computed by PriceStats and the consumer price index from the national statistical office in Argentina (INDEC).
Notes: The figure compares a price index produced with online data to a comparable official consumer price index (CPI) for the case of Argentina from 2007 to 2015. It also looks at annual and monthly inflation rates using each source of data. Monthly inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average a month before. Annual inflation rates for the online index are computed as the percentage change in the average of the previous 30 days compared to the same average 365 days before. All price indexes are nonseasonally adjusted.

See also: https://thebillionpricesproject.com/ and https://www.pricestats.com/

The COMPAS algorithm to predict criminals' recidivism

Background

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a decision support tool developed by Northpointe (now Equivant) used by U.S. courts to **assess the likelihood of recidivism**
- Produced several scales (Pretrial release risk, General recidivism, Violent recidivism) based on factors such as age, criminal history, and substance abuse
- The algorithm is proprietary and its inner workings are not public

Practitioner's Guide to COMPAS Core

The Practitioner's Guide provides an overview of the COMPAS Core Module in the Northpointe Suite. The Northpointe Suite is an integrated web-based assessment and case management system for criminal justice practitioners. The Northpointe Suite has modules designed for pretrial, jail, probation, prison, parole and community corrections applications. COMPAS Core is designed for both male and female offenders recently removed from the community or currently in the community. The Practitioner's Guide to COMPAS Core covers case interpretation, validity and reliability, and treatment implications. Most of the information provided is specific to COMPAS Core. Throughout this text we use the term COMPAS Core to distinguish an element (scale, typology, decile type) specific to COMPAS Core from general elements in the Northpointe Suite, such as scales found in both COMPAS Core and COMPAS Reentry.

COMPAS is a fourth generation risk and needs assessment instrument. Criminal justice agencies across the nation use COMPAS to inform decisions regarding the placement, supervision and case management of offenders. COMPAS was developed empirically with a focus on predictors known to affect recidivism. It includes dynamic risk factors, and it provides information on a variety of well validated risk and needs factors designed to aid in correctional intervention to decrease the likelihood that offenders will reoffend.

COMPAS was first developed in 1998 and has been revised over the years as the knowledge base of criminology has grown and correctional practice has evolved. In many ways changes in the field have followed new developments in risk assessment. We continue to make improvements to COMPAS based on results from norm studies and recidivism studies conducted in jails, probation agencies, and prisons. COMPAS is periodically updated to keep pace with with emerging best practices and technological advances.

In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/needs data are critical. COMPAS was designed to optimize these practical factors. We acknowledge the trade-off between comprehensive coverage of key risk and criminogenic factors on the one hand, and brevity and practicality on the other. COMPAS deals with this trade-off in several was; it provides a comprehensive set of key risk factors that have emerged from the recent criminological literature, and it allows for customization inside the software. Therefore, ease of use, efficient and effective time management, and case management considerations that are critical to best practice in the criminal justice field can be achieved through COMPAS.

`Source` Practitioner's Guide to COMPAS Core

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

Source ProPublica 2016

Machine Bias*

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner

There's software used across the country to predict future criminals. And it's biased against blacks.

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances—which belonged to a 6-year-old boy—a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late—a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting $86.35 worth of tools from a nearby Home Depot store (Figure 6.1.1).

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden—who is black—was rated a high risk. Prater—who is white—was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Scores like this—known as risk assessments—are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts—as is the case in Fort Lauderdale—to

* Julia Angwin, Jeff Larson, Surya Mattu, Lauren Kirchner, "Machine Bias," *ProPublica* (May 23, 2016). Reprinted with permission.

DOI: 10.1201/9781003278290-37

The COMPAS algorithm to predict criminals' recidivism

The ProPublica and other investigations

- In 2016, ProPublica published an investigation showing that COMPAS was **biased against African Americans**
- **Bias:** The algorithm was more likely for African Americans to wrongly predict that defendants would re-offend.
- **Accuracy:** only 20% of people predicted to commit violent crimes actually went on to do so (in a later study estimated with 65%, still worse than a group of humans with little expertise)

SCIENCE ADVANCES | RESEARCH ARTICLE

RESEARCH METHODS

The accuracy, fairness, and limits of predicting recidivism

Julia Dressel and Hany Farid*

Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and sentencing decisions. Proponents of these systems argue that big data and advanced machine learning make these analyses more accurate and less biased than humans. We show, however, that the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise. In addition, despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear predictor with only two features.

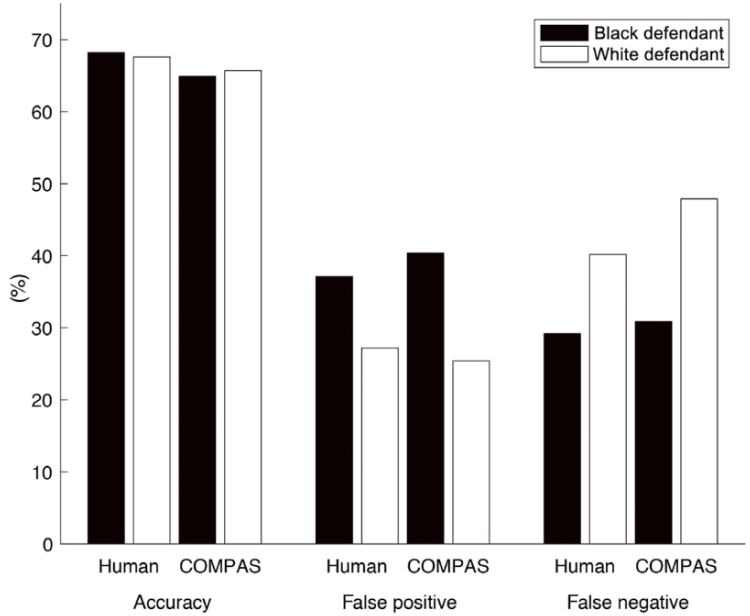


Fig. 1. Human (no-race condition) versus COMPAS algorithmic predictions (see also Table 1).

`Source` Dressel and Fair, 2018, Science Advances

Like-minded sources on Facebook are prevalent but not polarizing

https://doi.org/10.1038/s41586-023-06297-w

Received: 21 December 2022

Accepted: 7 June 2023

Published online: 27 July 2023

Open access

Check for updates

Brendan Nyhan[1,25], Jaime Settle[2,25], Emily Thorson[3,25], Magdalena Wojcieszak[4,5,25], Pablo Barberá[6,25], Annie Y. Chen[7], Hunt Allcott[8], Taylor Brown[6], Adriana Crespo-Tenorio[6], Drew Dimmery[6,24], Deen Freelon[9], Matthew Gentzkow[10], Sandra González-Bailón[9], Andrew M. Guess[11,12], Edward Kennedy[13], Young Mie Kim[14], David Lazer[15], Neil Malhotra[16], Devra Moehler[6], Jennifer Pan[17], Daniel Robert Thomas[6], Rebekah Tromble[18,19], Carlos Velasco Rivera[6], Arjun Wilkins[6], Beixian Xiong[6], Chad Kiewiet de Jonge[6,26], Annie Franco[6,26], Winter Mason[6,26], Natalie Jomini Stroud[20,21,26] & Joshua A. Tucker[22,23,26]

Many critics raise concerns about the prevalence of 'echo chambers' on social media and their potential role in increasing political polarization. However, the lack of available data and the challenges of conducting large-scale field experiments have made it difficult to assess the scope of the problem[1,2]. Here we present data from 2020 for the entire population of active adult Facebook users in the USA showing that content from 'like-minded' sources constitutes the majority of what people see on the platform, although political information and news represent only a small fraction of these exposures. To evaluate a potential response to concerns about the effects of echo chambers, we conducted a multi-wave field experiment on Facebook among 23,377 users for whom we reduced exposure to content from like-minded sources during the 2020 US presidential election by about one-third. We found that the intervention increased their exposure to content from cross-cutting sources and decreased exposure to uncivil language, but had no measurable effects on eight preregistered attitudinal measures such as affective polarization, ideological extremity, candidate evaluations and belief in false claims. These precisely estimated

How do social media feed algorithms affect attitudes and behavior in an election campaign?

Andrew M. Guess[1]*, Neil Malhotra[2], Jennifer Pan[3], Pablo Barberá[4], Hunt Allcott[5], Taylor Brown[4], Adriana Crespo-Tenorio[4], Drew Dimmery[4,6], Deen Freelon[7], Matthew Gentzkow[8], Sandra González-Bailón[9], Edward Kennedy[10], Young Mie Kim[11], David Lazer[12], Devra Moehler[4], Brendan Nyhan[13], Carlos Velasco Rivera[4], Jaime Settle[14], Daniel Robert Thomas[4], Emily Thorson[15], Rebekah Tromble[16], Arjun Wilkins[4], Magdalena Wojcieszak[17,18], Beixian Xiong[4], Chad Kiewiet de Jonge[4], Annie Franco[4], Winter Mason[4], Natalie Jomini Stroud[19], Joshua A. Tucker[20]

We investigated the effects of Facebook's and Instagram's feed algorithms during the 2020 US election. We assigned a sample of consenting users to reverse-chronologically-ordered feeds instead of the default algorithms. Moving users out of algorithmic feeds substantially decreased the time they spent on the platforms and their activity. The chronological feed also affected exposure to content: The amount of political and untrustworthy content they saw increased on both platforms, the amount of content classified as uncivil or containing slur words they saw decreased on Facebook, and the amount of content from moderate friends and sources with ideologically mixed audiences they saw increased on Facebook. Despite these substantial changes in users' on-platform experience, the chronological feed did not significantly alter levels of issue polarization, affective polarization, political knowledge, or other key attitudes during the 3-month study period.

Reshares on social media amplify political news but do not detectably affect beliefs or opinions

Andrew M. Guess[1]*, Neil Malhotra[2], Jennifer Pan[3], Pablo Barberá[4], Hunt Allcott[5], Taylor Brown[4], Adriana Crespo-Tenorio[4], Drew Dimmery[4,6], Deen Freelon[7], Matthew Gentzkow[8], Sandra González-Bailón[9], Edward Kennedy[10], Young Mie Kim[11], David Lazer[12], Devra Moehler[4], Brendan Nyhan[13], Carlos Velasco Rivera[4], Jaime Settle[14], Daniel Robert Thomas[4], Emily Thorson[15], Rebekah Tromble[16], Arjun Wilkins[4], Magdalena Wojcieszak[17,18], Beixian Xiong[4], Chad Kiewiet de Jonge[4], Annie Franco[4], Winter Mason[4], Natalie Jomini Stroud[19], Joshua A. Tucker[20]

We studied the effects of exposure to reshared content on Facebook during the 2020 US election by assigning a random set of consenting, US-based users to feeds that did not contain any reshares over a 3-month period. We find that removing reshared content substantially decreases the amount of political news, including content from untrustworthy sources, to which users are exposed; decreases overall clicks and reactions; and reduces partisan news clicks. Further, we observe that removing reshared content produces clear decreases in news knowledge within the sample, although there is some uncertainty about how this would generalize to all users. Contrary to expectations, the treatment does not significantly affect political polarization or any measure of individual-level political attitudes.

Asymmetric ideological segregation in exposure to political news on Facebook

Sandra González-Bailón[1]*, David Lazer[2], Pablo Barberá[3], Meiqing Zhang[3], Hunt Allcott[4], Taylor Brown[3], Adriana Crespo-Tenorio[3], Deen Freelon[1], Matthew Gentzkow[5], Andrew M. Guess[6], Shanto Iyengar[7], Young Mie Kim[8], Neil Malhotra[9], Devra Moehler[3], Brendan Nyhan[10], Jennifer Pan[11], Carlos Velasco Rivera[3], Jaime Settle[12], Emily Thorson[13], Rebekah Tromble[14], Arjun Wilkins[3], Magdalena Wojcieszak[15,16], Chad Kiewiet de Jonge[3], Annie Franco[3], Winter Mason[3], Natalie Jomini Stroud[17,18], Joshua A. Tucker[19,20]

Does Facebook enable ideological segregation in political news consumption? We analyzed exposure to news during the US 2020 election using aggregated data for 208 million US Facebook users. We compared the inventory of all political news that users could have seen in their feeds with the information that they saw (after algorithmic curation) and the information with which they engaged. We show that (i) ideological segregation is high and increases as we shift from potential exposure to actual exposure to engagement; (ii) there is an asymmetry between conservative and liberal audiences, with a substantial corner of the news ecosystem consumed exclusively by conservatives; and (iii) most misinformation, as identified by Meta's Third-Party Fact-Checking Program, exists within this homogeneously conservative corner, which has no equivalent on the liberal side. Sources favored by conservative audiences were more prevalent on Facebook's news ecosystem than those favored by liberals.

The Meta US 2020 Election study

Hertie School

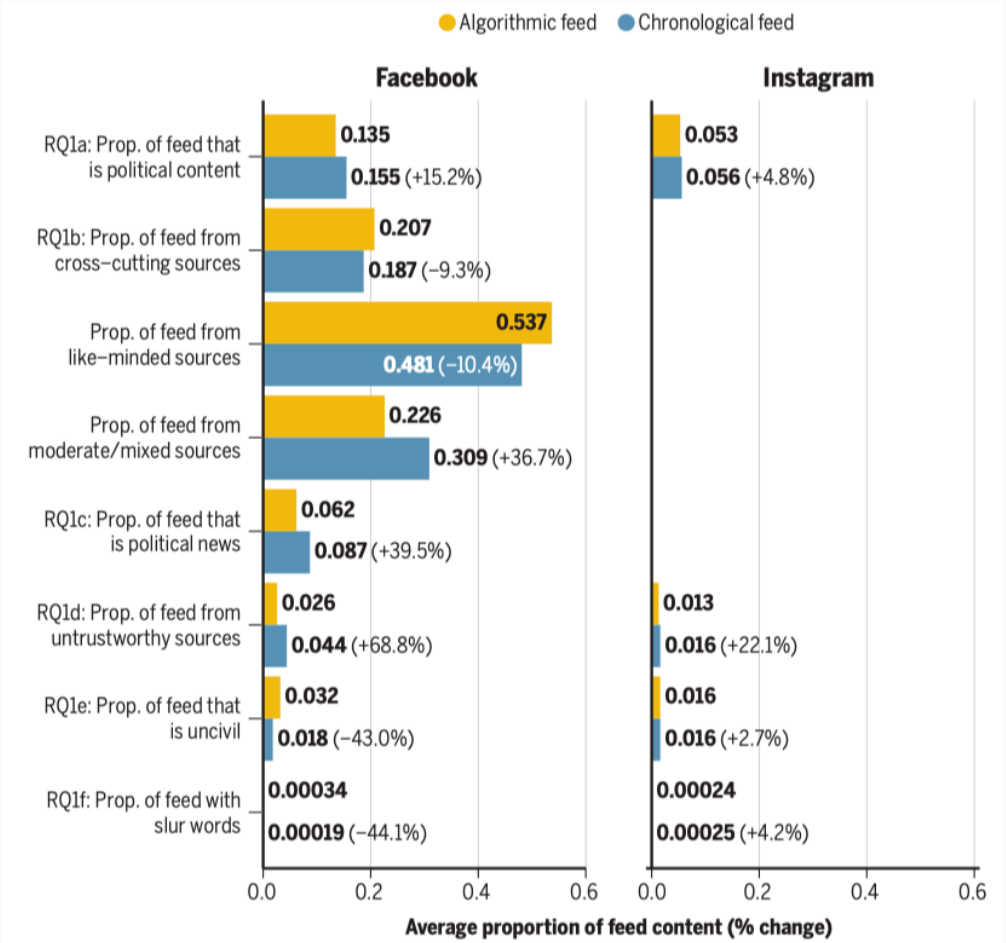


Fig. 2. Estimated changes in prevalence of feed content on both Facebook and Instagram. (**Left**) Facebook. (**Right**) Instagram. Values are average unweighted proportions within each group, with percent changes relative to the Algorithmic Feed control group in parentheses. All differences are significant at the $p < 0.005$ level, except RQ1f for Instagram ($p < 0.05$); confidence intervals are thus not shown. RQ1b and RQ1c were not tested for Instagram because political and ideology classifications are not available on that platform. Fully specified regression models with survey weights are reported in the SM, section S2.2.

`Source` Guess et al. 2023, Science

The Meta US 2020 Election study

Hertie School

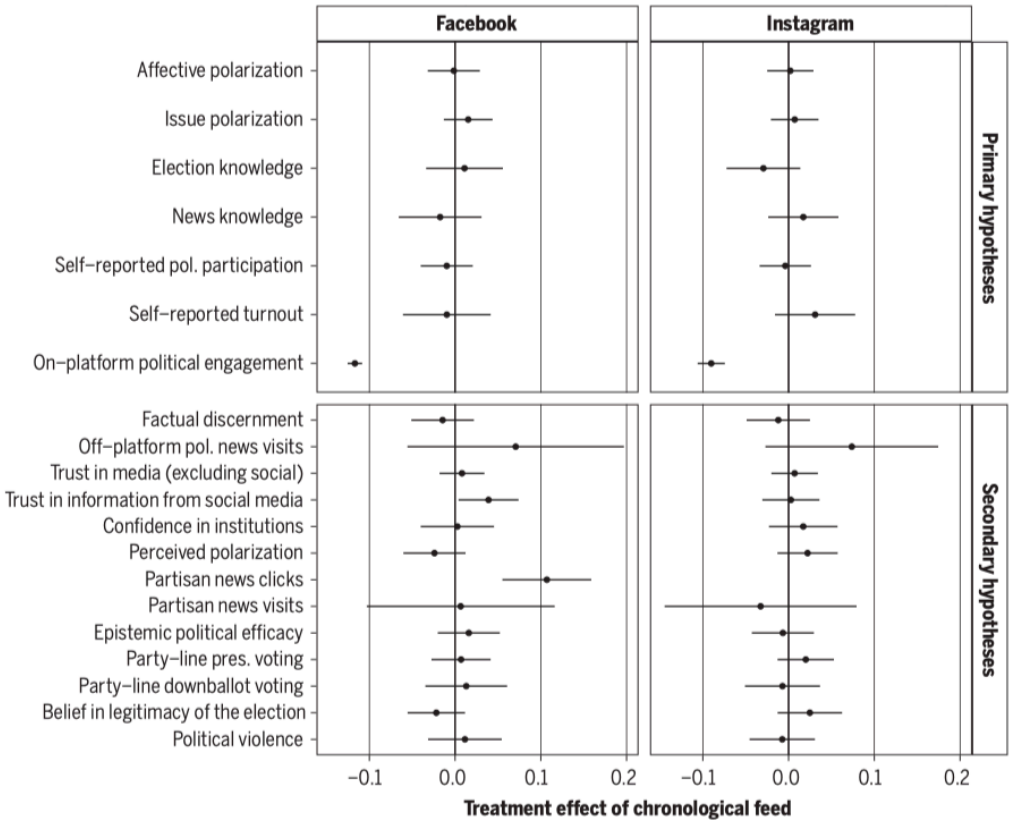


Fig. 3. Population average treatment effects of the Chronological Feed, relative to the Algorithmic Feed control group, on both Facebook and Instagram. (**Left**) Facebook. (**Right**) Instagram. Estimates are presented in standard deviations with 95% confidence intervals (not adjusted for multiple comparisons). Partisan news clicks are estimated only for Facebook because source-level estimates of political ideology are not available for Instagram. pol., political; pres., presidential.

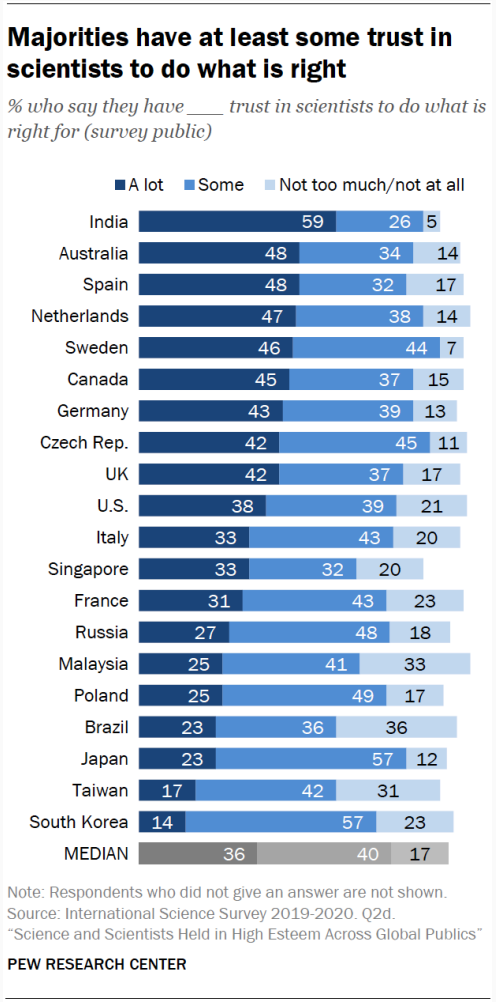
`Source` Guess et al. 2023, Science

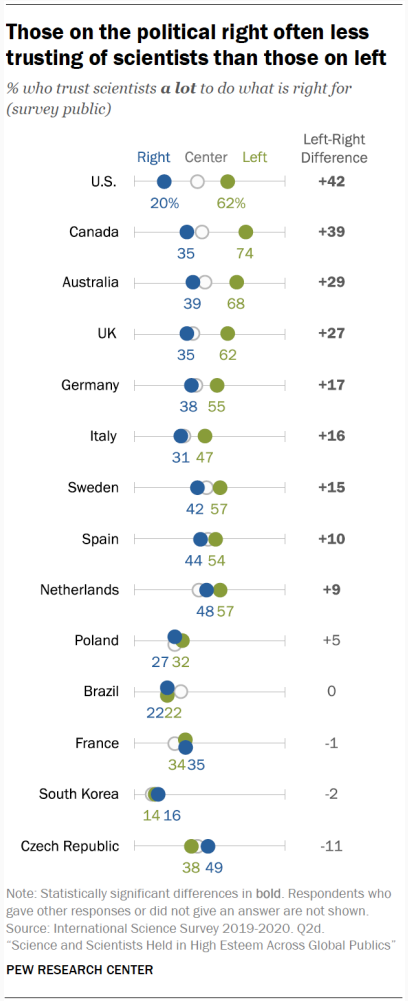
Fostering good data science for good public policy

Hertie School

Trust in science

Hertie School





Source Pew 2020

Trust in science

Hertie School

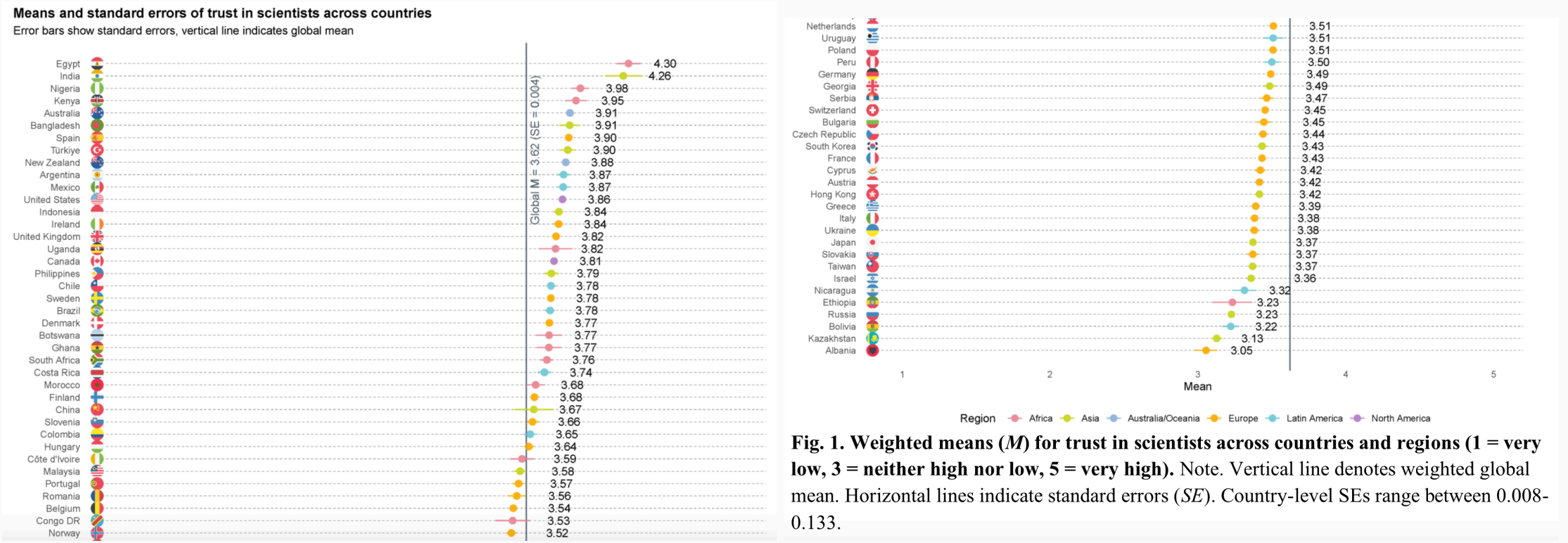


Fig. 1. Weighted means (*M*) for trust in scientists across countries and regions (1 = very low, 3 = neither high nor low, 5 = very high). Note. Vertical line denotes weighted global mean. Horizontal lines indicate standard errors (*SE*). Country-level SEs range between 0.008-0.133.

`Source` Cologna et al. 2024

Trust in science

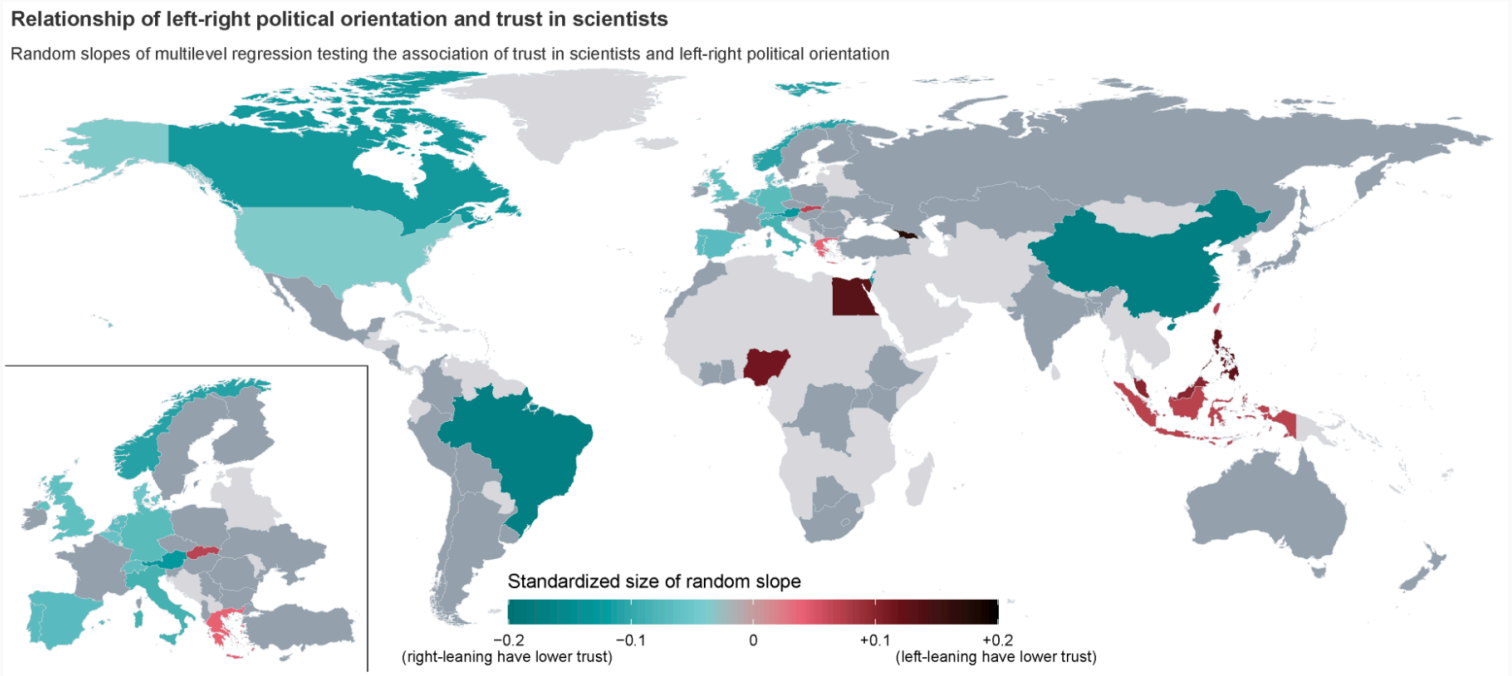


Fig. 3. Relationship of left-right political orientation and trust in scientists. Figure visualises standardised random slopes for political orientation (1 = left – 5 = right), which were extracted from a weighted linear multilevel regression model that explained trust in scientists (1 = very low, 3 = neither high nor low, 5 = very high) across countries and contained random intercepts and slopes of political orientation across countries. Countries with significant effects ($p < .05$) are displayed in colours: Countries coloured in shades of blue show a positive association of left-leaning orientation and trust in scientists (i.e., right-leaning have lower trust). Countries coloured in shades of red show a positive association of right-leaning orientation and trust in scientists (i.e., left-leaning have lower trust) Countries with non-significant effects are shaded in dark grey. Countries with no available data are shaded in light grey.

Source Cologna et al. 2024

Trust in data science

Hertie School

Data scientists have the potential to help save the world

By Leo Borrett May 17, 2017

With an untold number of crises emerging every year, big data is becoming increasingly important for helping aid organisations respond quickly to chaotic and evolving situations.

HOW DATA SCIENCE IS SAVING LIVES

AVINASH N Sep 29 · 2 min read

For all the people first priority is about their life. Life is one of the most precious thing in the world. Can Data Science techniques save life, is it possible? Yes, using Data Science techniques to analyze large data sets today has a huge impact on saving lives.

Health

Artificial intelligence and covid-19: Can the machines save us?

Analytics And Data Science

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

How AI Will Save Thousands of Lives

Sepsis is the problem; data are the cure

Drew Smith, PhD Jan 10, 2020 · 5 min read ★

STUDENTS

Data Science: Why It Matters and How It Can Make You Rich

Trust in data science?

The Cambridge Analytica case: What's a data scientist to do?

The Cambridge Analytica controversy has highlighted data ethics issues especially dear to early career stage data scientists

Researchers just released profile data on 70,000 OkCupid users without permission

By Brian Resnick | @B_resnick | brian@vox.com | May 12, 2016, 6:00pm EDT

An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor

Its creators said they could use facial analysis to determine if someone would become a criminal. Critics said the work recalled debunked "race science."

Data Failed the Election, But There's Still Hope for Business

Everyone is blaming data for failing to predict Trump's win. But it's the data handlers who need the real reexamination.

The replication crisis

What the crisis is about

- The finding that many scientific studies are difficult or impossible to reproduce.
- Reproducibility is a cornerstone of science as an enterprise of knowledge generation → bad.

Factors fueling the replication crisis

- Solo, silo-ed investigators limited to small sample sizes
- Wrong incentives in science
- No pre-registration of hypotheses being tested
- Post-hoc cherry picking of hypotheses with best P values
- Only requiring P < .05
- No replication
- No data sharing

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

Published research findings are sometimes refuted by subsequent evidence, with ensuing confusion and disappointment. Refutation and controversy is seen across the range of research designs, from clinical trials and traditional epidemiological studies [1–3] to the most modern molecular research [4,5]. There is increasing concern that in modern research, false findings may be the majority or even the vast majority of published research claims [6–8]. However, this should not be surprising. It can be proven that most claimed research findings are false. Here I will examine the key factors that influence this problem and some corollaries thereof.

The Essay section contains opinion pieces on topics of broad interest to a general medical audience.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful. "Negative" is actually a misnomer, and the misinterpretation is widespread. However, here we will target relationships that investigators claim exist, rather than null findings.

As has been shown previously, the probability that a research finding is indeed true depends on the prior probability of it being true (before doing the study), the statistical power of the study, and the level of statistical significance [10,11]. Consider a 2×2 table in which research findings are compared against the gold standard of true relationships in a scientific field. In a research field both true and false hypotheses can be made about the presence of relationships. Let R be the ratio of the number of "true relationships" to "no relationships" among those tested in the field. R is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2×2 table, one gets PPV = $(1 - \beta)R/(R - \beta R + \alpha)$. A research finding is thus

Citation: Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2(8): e124.

Abbreviation: PPV, positive predictive value

John P. A. Ioannidis is in the Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece, and Institute for Clinical Research and Health Policy Studies, Department of Medicine, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, United States of America. E-mail: jioannid@cc.uoi.gr

Competing Interests: The author has declared that no competing interests exist.

DOI: 10.1371/journal.pmed.0020124

`Source` Ioannidis 2005/PLOS Medicine

Goals of this workshop

1. Thinking hard about causality

- Develop expectations that imply testable statements about effects.
- Prioritize designs that help isolate causal effects.
- Care about **internal validity**.

2. Measure the policy options and outcomes you want to learn about

- Find good empirical representations of your concepts of interest.
- Observe and/or manipulate wisely.
- Care about **measurement validity**.

3. Make conclusions about the real world

- Generalize with care.
- Don't oversell or misinterpret your findings.
- Account for uncertainty.
- Care about **external validity and statistical conclusion validity**.

Learning goals for this workshop

Hertie School

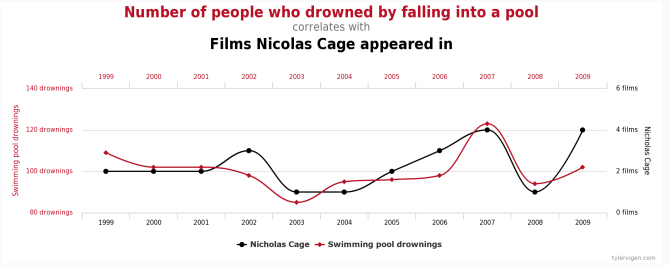
	Data science literacy					
Day	Statistical literacy	Causal reasoning	Data literacy	AI literacy	Evidence consumption	Ethical reasoning
1 - Fundamental data and statistical literacy	✅	✅	✅		✅	
2 - Policy evaluation and impact assessment	✅	✅			✅	
3 - AI and big data for policy-making			✅	✅		✅
4 - Informed consumption of evidence	✅				✅	
5 - Data visualization and communication			✅		✅	
6 - Data management and ethics			✅	✅		✅

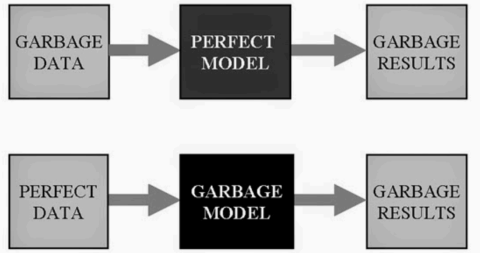
Calling out bad evidence when you see it

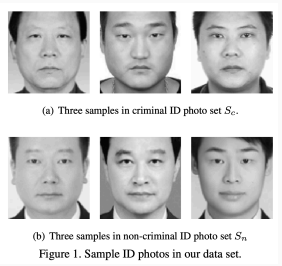
Hertie School

1. Learn not to be fooled by

- big data
- garbage data
- garbage models
- weird samples
- claims of generality
- statistical significance
- implausibly large effect sizes
- highly precise forecasts
- overfitted models

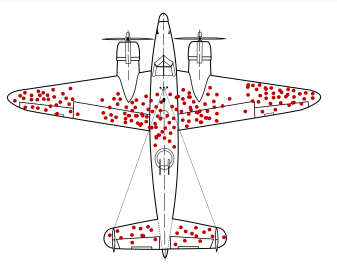


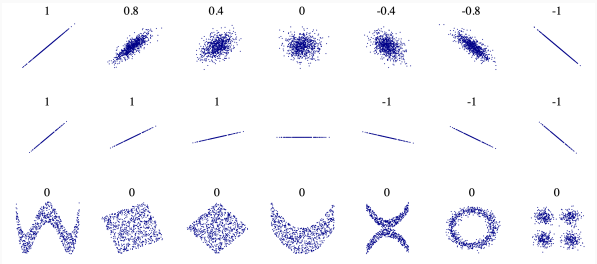



(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.





2. Consume policy-relevant evidence effectively and efficiently

3. Apply elements of data science to policy-making

What we're not going to cover

Programming

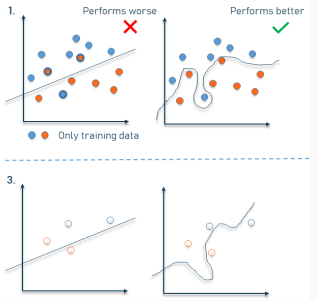
- Python, R, SQL, etc. skills are essential for data science
- The learning curve is steep and requires practice
- We're happy to provide a glimpse behind the curtain if that is of interest, and provide additional resources

Active modeling

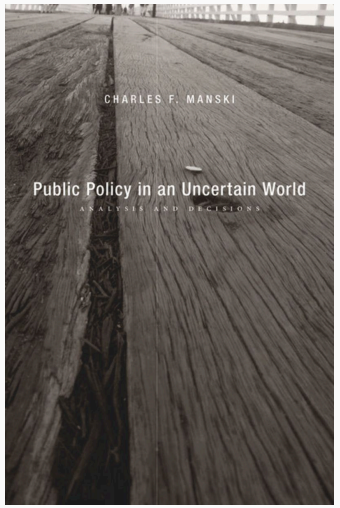
- Building designs and models - explanatory and predictive - requires more theoretical and practical knowledge than we can cover in this workshop
- Focusing on the principles of statistical and causal reasoning should be sufficient to critically assess designs and models

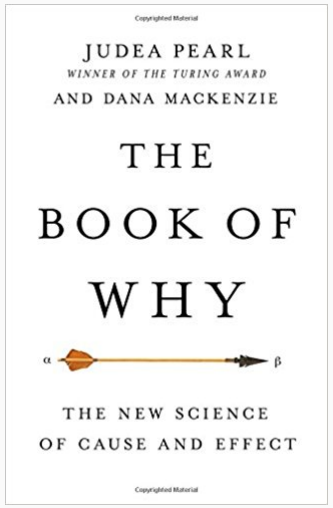
Advanced machine learning, NLP

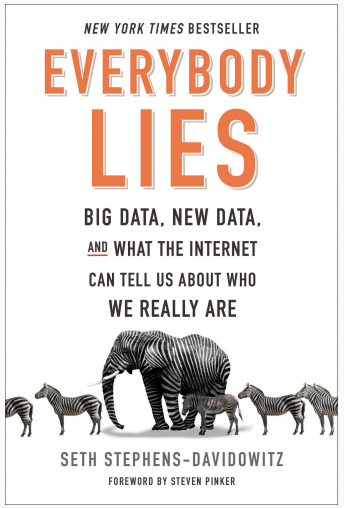
- ML, DL, NLP are technologies that drive many of the most exciting applications of data science
- Understanding what happens under the hood requires a solid foundation in math and stats
- We will focus on fundamental elements of ML-based research

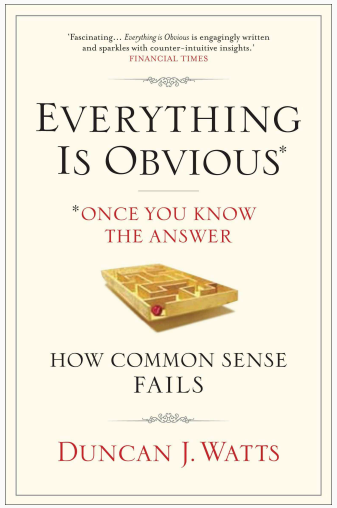


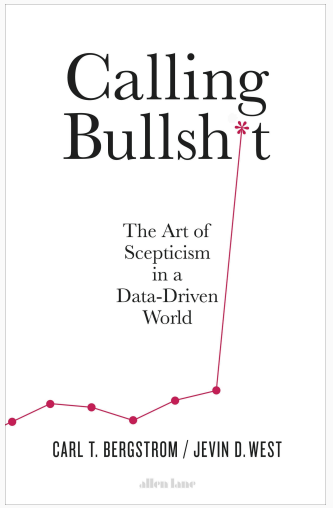
Further reading

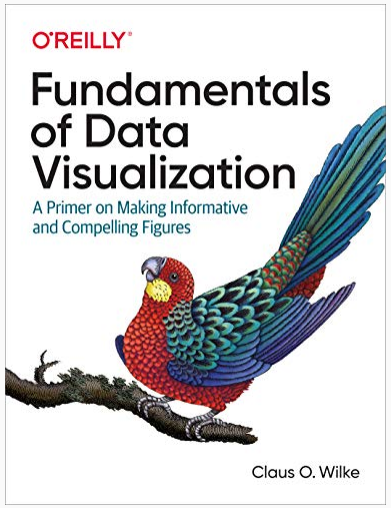


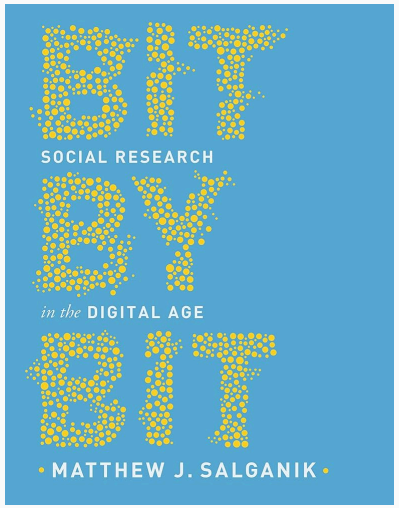


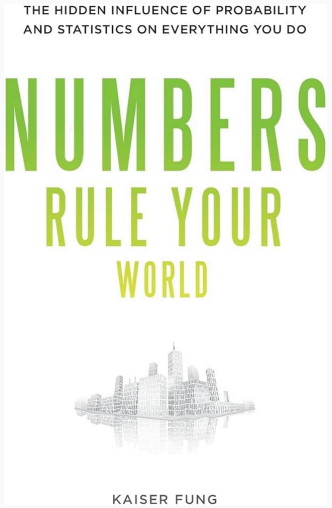


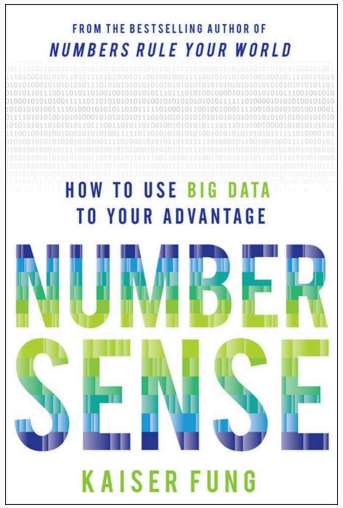


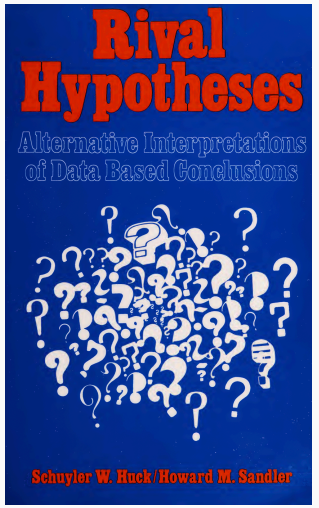


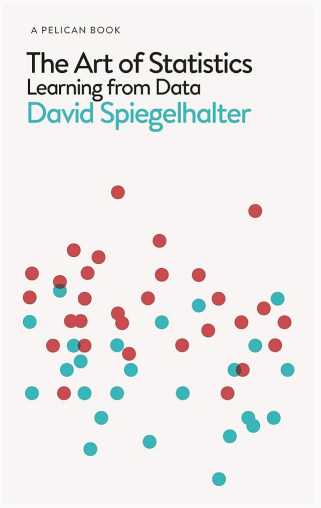


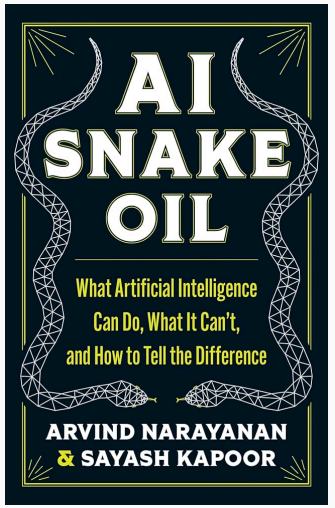












Further listening