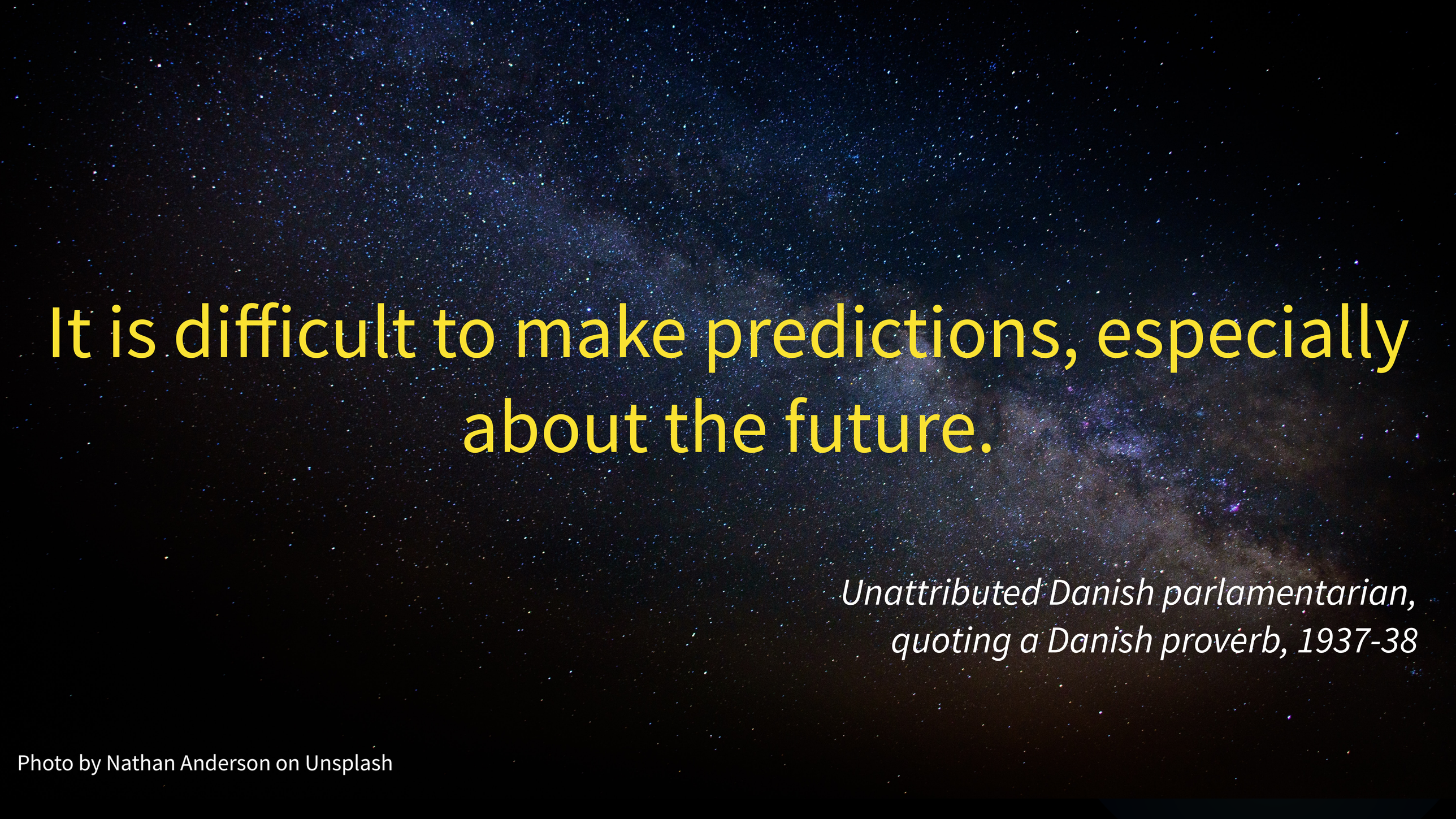
Data Science Education in 2022

Carl Howe, Director of Education
carl@rstudio.com
carlhowe.com


It is difficult to make predictions, especially about the future.
Unattributed Danish parlamentarian, quoting a Danish proverb, 1937-38
Photo by Nathan Anderson on Unsplash

RStudio's Mission:

Equip everyone, regardless of means, to participate in a global economy that rewards data literacy

RStudio: A Public Benefit Corporation

Open Source Software for Data Science

9:05 AM-10:00 AM

Keynote

Room 2

J.J. Allaire
Founder and CEO
RStudio

Open-source software is fundamentally necessary to ensure that the tools of data science are broadly accessible, and to provide a reliable and trustworthy foundation for reproducible research. This talk will delve into why open source software is so important and discuss the role of corporations as stewards of open source software. I'll also talk about how RStudio is structured and organized to pursue its mission of creating open source software for data science.

RStudio Education's mission:

Train the next million R users

#NextMillionRUsers

Nearly 100 certified instructors

RStudio Instructor Training and Certification

Our certified trainers would be happy to chat about personalized trainings or workshops that fit your needs. Or find o
become certified yourself.

Certified Trainers

Tidyverse


Ahmad_Husain Abdullah
Jakarta, Indonesia


Garrick Aden-Buie
St. Petersburg, FL, United States


Sara Altman
San Francisco, CA, United States


Mario Annau
Vienna, Austria


D
Bristol


Malcolm Barrett
Los Angeles, CA, United States


Yanina Bellini Saibene
Santa Rosa, La Pampa, Argentina

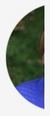

John Blischak
Akron, OH, United States


Ne
Antwe


Jenny Bryan
Vancouver, BC, Canada

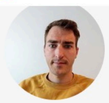

Joyce Cahoon
Raleigh, NC, United States

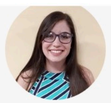

Tanya Cashorali
Boston, MA, United States


Mine Çeti
Edinbu
Ki

Studio Education

Certification Directory Become a Trainer FAQ Contact Us


Jeroen Claes
Antwerp, Belgium


Paola Corrales
Buenos Aires, Argentina


Richie Cotton
New York, NY, United States


Rhian Davies
Lancaster, United Kingdom

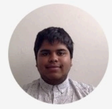

Nathan Day
Charlottesville, VA, USA


Ahmadou Dicko
Dakar, Senegal

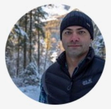

Dewey Dunnington
Halifax, NS, Canada


Tiara Dwiputri
Jakarta, Indonesia


Matthew Flickinger
Ann Arbor, MI, United States


Ian Flores Siaca
San Juan, PR, United States


Hannah Frick
London, United Kingdom


Martin Frigaard
Oakland, CA, United States

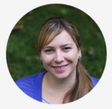

Colin Gillespie
Newcastle upon Tyne, United Kingdom


Garrett Grolemund
Clearwater, FL, United States


Ronny Hernández-Mora
San José, Costa Rica

Shiny


Malcolm Barrett
Los Angeles, CA, United States


Mine Çetinkaya-Rundel
Edinburgh, United Kingdom


Rhian Davies
Lancaster, United Kingdom

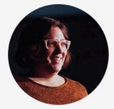

Tiara Dwiputri
Jakarta, Indonesia


Colin Gillespie
Newcastle upon Tyne, United Kingdom


Garrett Grolemund
Clearwater, FL, United States


Ted Laderas
Portland, OR, United States

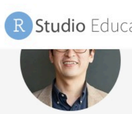

Theo Roe
Newcastle-Upon-Tyne, United Kingdom

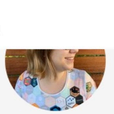

Noa Tamir
Berlin, Germany


Sebastian Voss
Essen, Germany

Studio Education

Certification Directory Become a Trainer FAQ Contact Us


Kwangchun Lee
Seoul, South Korea


Amelia McNamara
Minneapolis, MN, United States

François Michonneau
Hendaye, France

Dragoș Moldovan-Grünfeld
London, UK

Graham Parsons
Bristol, United Kingdom

Roman Popat
Edinburgh, United Kingdom

Ajeng Prastiwi
Jakarta, Indonesia

Danielle Quinn
St. John's, NL, Canada

Theo Roe
Newcastle-Upon-Tyne, United Kingdom

Zev Ross
Ithaca, NY, United States

Omayma Said
Düsseldorf, Germany

Adi Sarid
Tel Aviv, Israel

Katie Sasso Schafer
Columbus, OH, United States

Julia Silge
Salt Lake City, UT, United States

Mike Smith
Sandwich, Kent, United Kingdom

Benjamin Soltoff
icago, IL, United States

Andrea Spano
Milan, Italy

Arie Spirgel
Miami Beach, FL, United States

Meghan Sposato
Denver, CO, United States

Ryan Tanner
Denver, CO, United States

Nicole Thompson
Albuquerque, NM, United

Tiffany Timbers
Vancouver, BC, Canada

Frans van Dunné
San Carlos, Costa Rica

Greg Wilson
Toronto, ON, Canada

RStudio::conf 2020

19 Workshops

101 Teaching Staff

1,300+ students

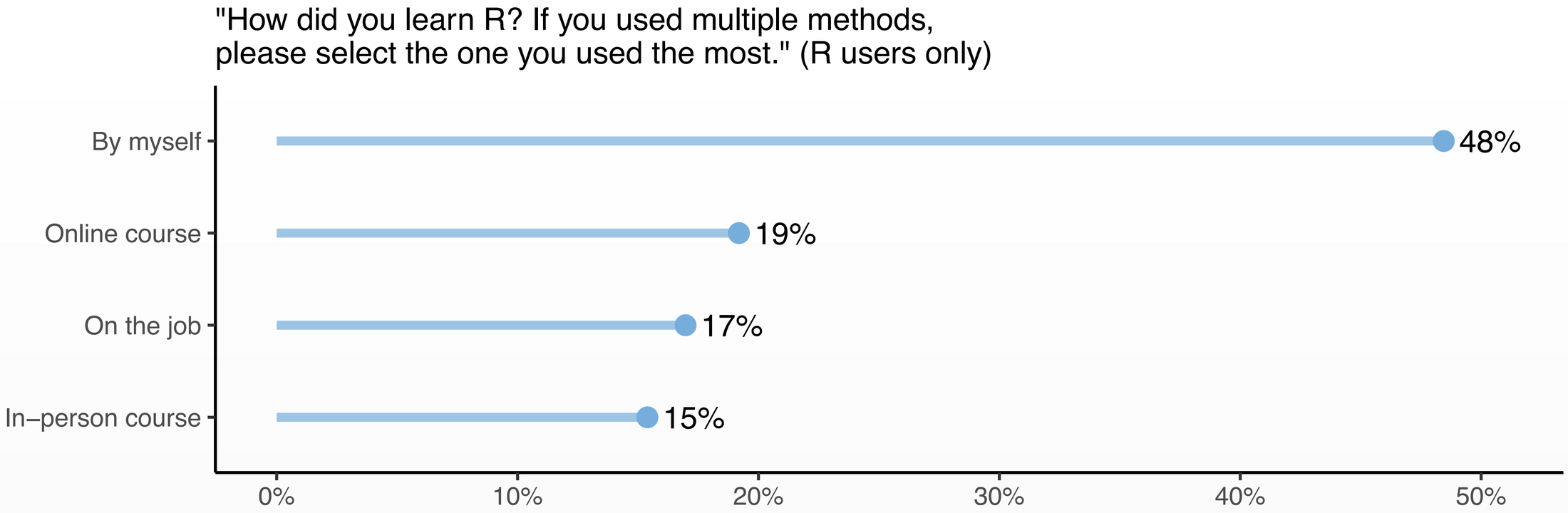
largest R education event in the world

Other RStudio Education Contributions

- Free online books
- R packages for education
- Open source workshop materials on github
- Free academic licenses for RStudio Pro products
- Annual survey of how people learn R

Photo by Nathan Anderson on Unsplash

Most R users are self-taught



https://github.com/rstudio/learning-r-survey

2019 R Community Survey, n = 1579

Data science education faces serious challenges


You're not afraid?
You will be.....

Challenges

Data explosion

1. We're drowning in data

"90% of the data in the world was created in the last 2 years"

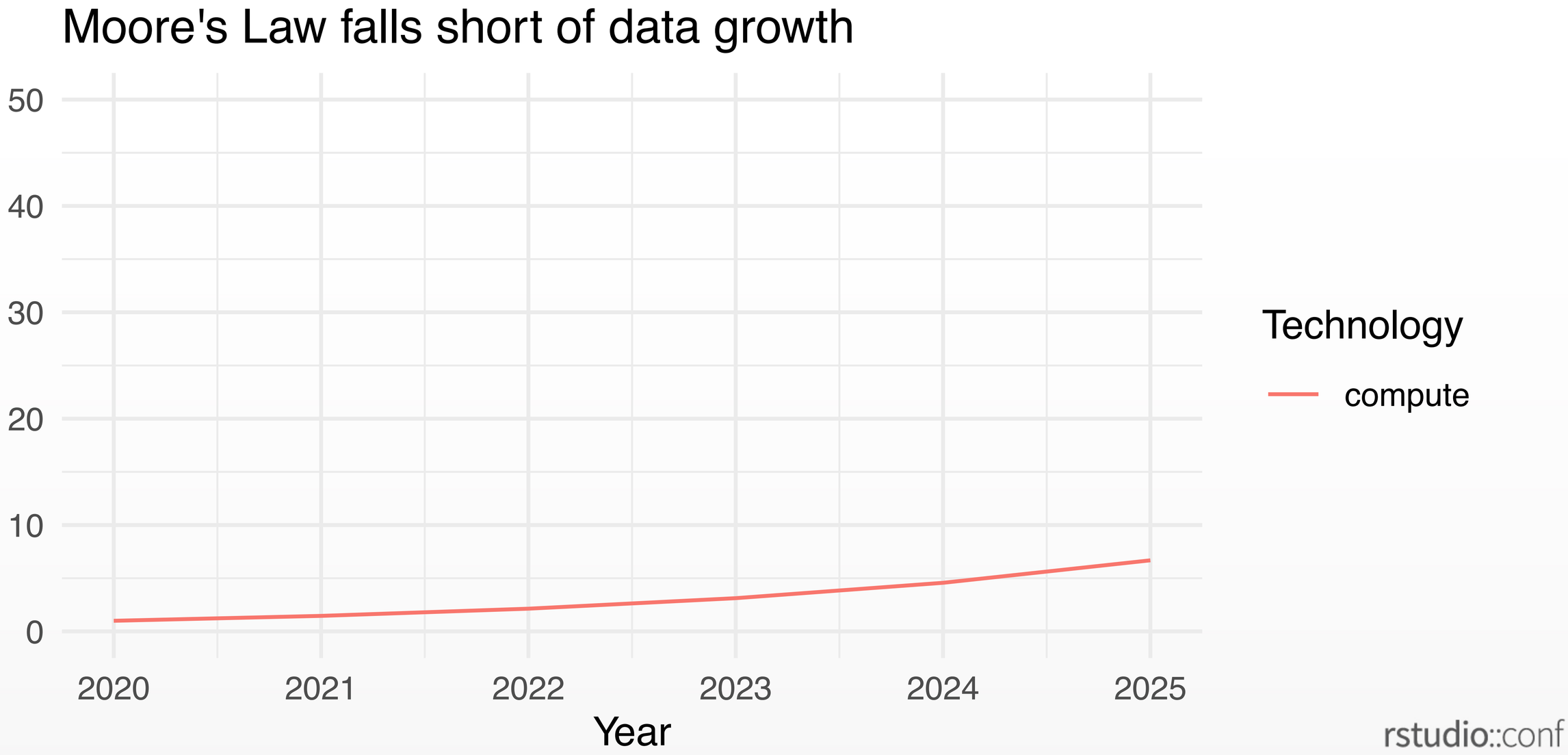
IBM Marketing Cloud, “10 Key Marketing Trends For 2017

2020

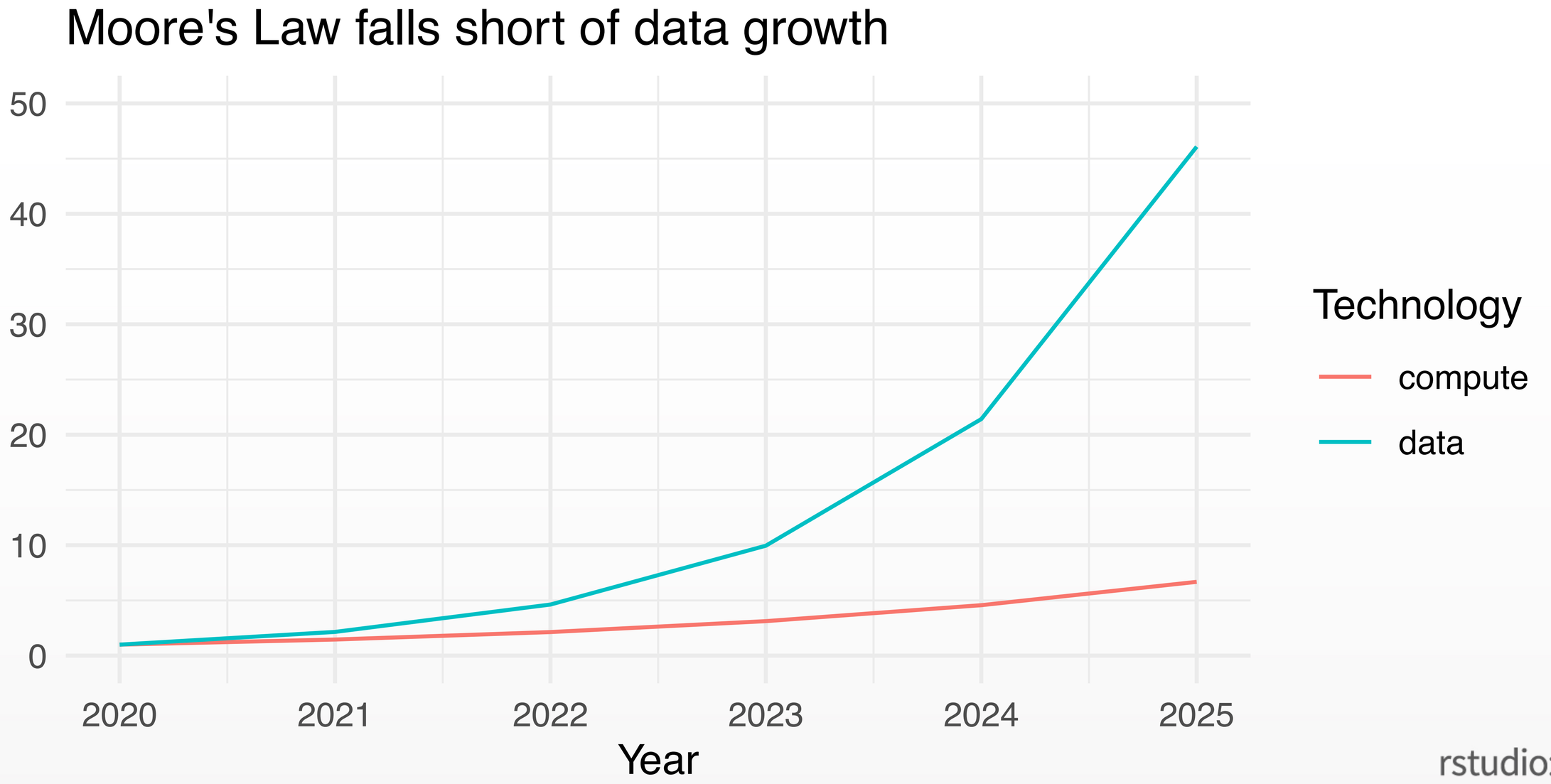
2022

2024

Moore's Law won't bail us out



Moore's Law won't bail us out



Challenges

- Data explosion
- Unreproducible results

2. We no longer trust science

PHYSICS TODAY

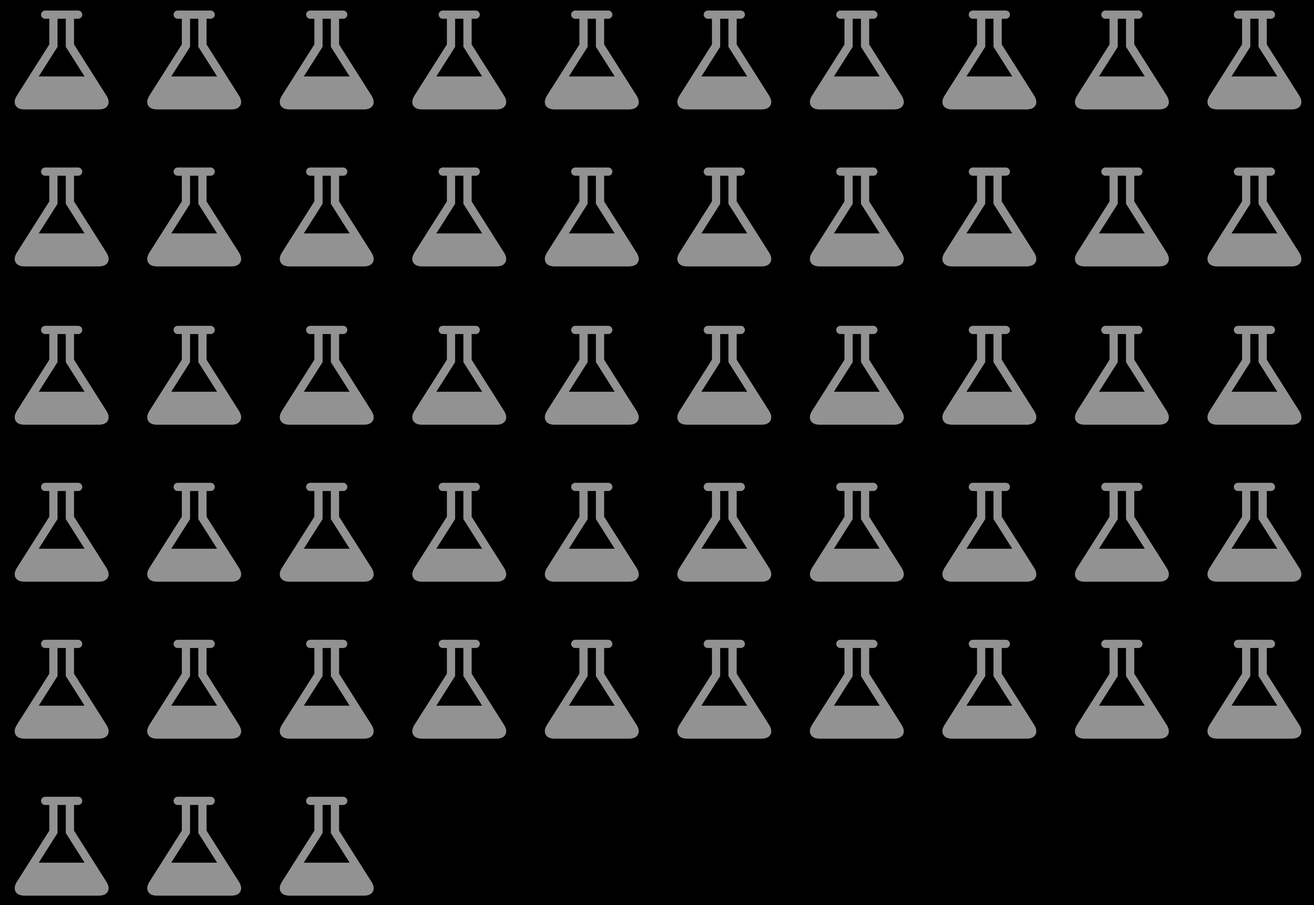
DOI:10.1063/PT.6.1.20180822a

22 Aug 2018 in Research & Technology

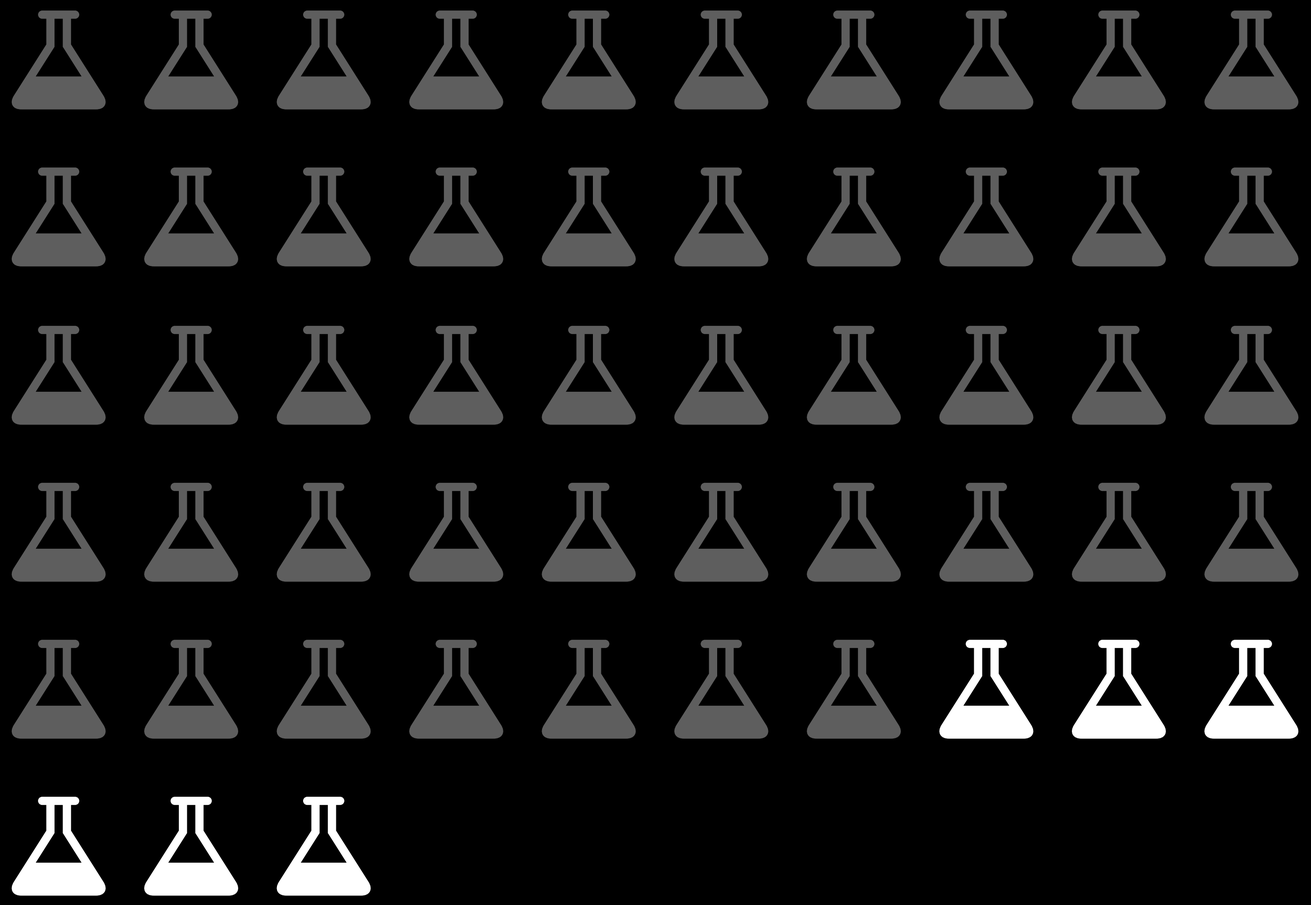
The war over supercooled water

How a hidden coding error fueled a seven-year dispute between two of condensed matter's top theorists.

Ashley G. Smart

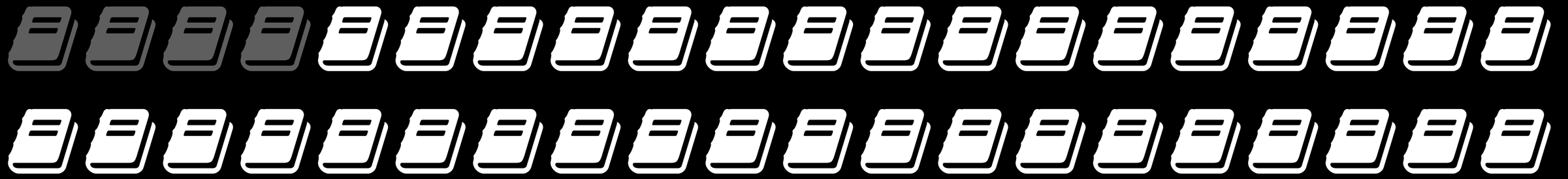


Amgen 2012



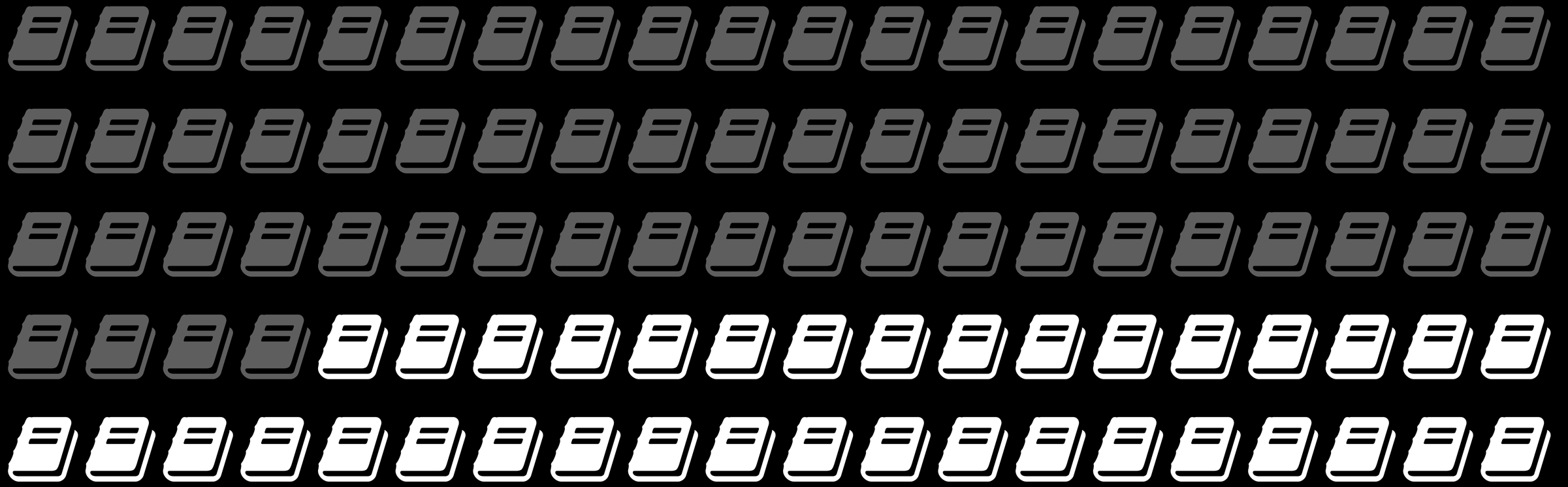
Amgen 2012: could reproduce only **6** of **53** landmark results

Psychology



36 of 100

Psychology

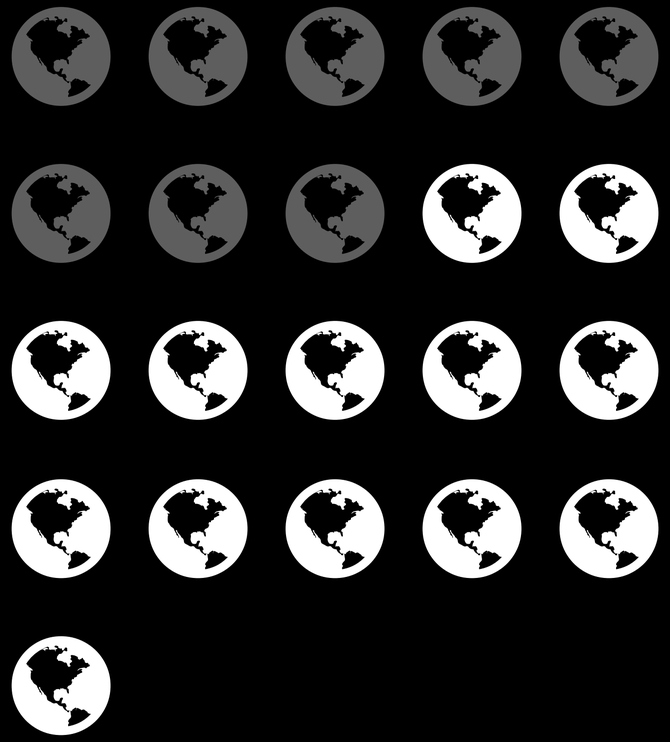


36 of 100

Economics

29 of **59**

Nature & Science



13 of **21**

Nosek et al. (2015). Estimating the reproducibility of psychological science. *Science*, 349, 6251.

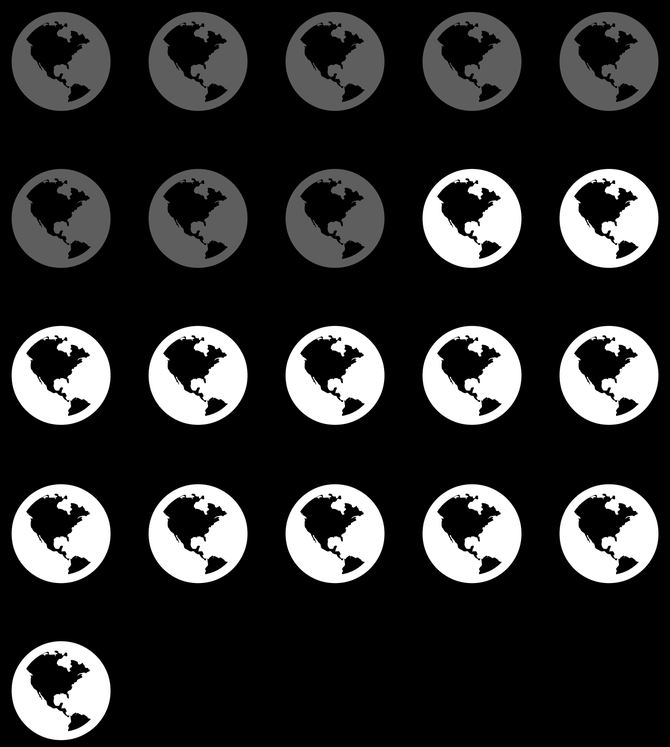
Chang AC, Li P (2015) Is Economics Research Replicable?, Finance and Economics Discussion Series 2015-083. (Board of Governors of the Federal Reserve System, Washington, DC).

Camerer, et al. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 2, 637–644.

Economics

29 of **59**

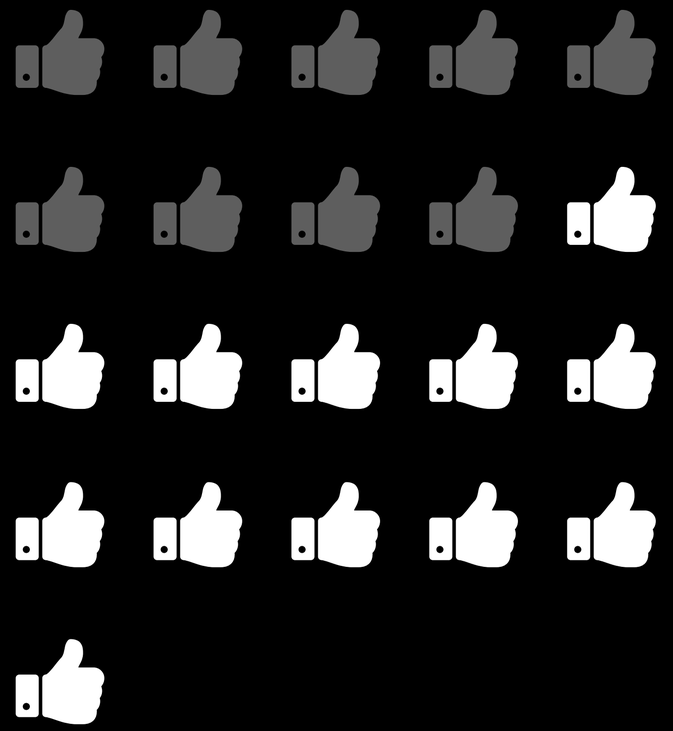
Nature & Science



13 of **21**

Coin Tosses

that were heads



12 of **21**

Why Do So Many Studies Fail to Replicate?

- The New York Times, May 2016

Psychology's Replication Crisis is Running Out of Excuses

- The Atlantic, November 2018

The Breakdown in Biomedical Research

- The Wall Street Journal, April 2017

$28 billion

£ 22B €~~24B~~

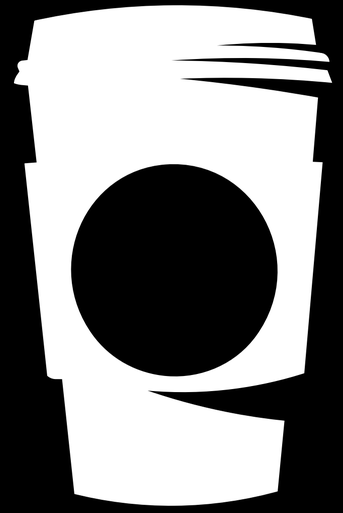
Costs to biomedical industry per year, US

Freedman LP, Cockburn IM, Simcoe TS (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

$28 billion

£ 22B € 24B

Costs to biomedical industry per year, US



Freedman LP, Cockburn IM, Simcoe TS (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

$28 billion

£ 22B € 24B

Costs to biomedical industry per year, US

Freedman LP, Cockburn IM, Simcoe TS (2015) The Economics of Reproducibility in Preclinical Research. PLoS Biol 13(6): e1002165. https://doi.org/10.1371/journal.pbio.1002165

Challenges

- Data explosion
- Unreproducible results
- Fake data

3. Data has become another way to lie

FEATURES

£350 million for the NHS: How the Brexit bus pledge is coming true

Vote Leave's success has turned the Conservatives into big spenders

James Forsyth and Fraser Nelson

Environment

Hundreds of climate sceptics to mount international campaign to stop net-zero targets being made law

Exclusive: The signatories are part of a network pushing for environmental deregulation after Brexit – and some have links with Boris Johnson's cabinet

Health & Science

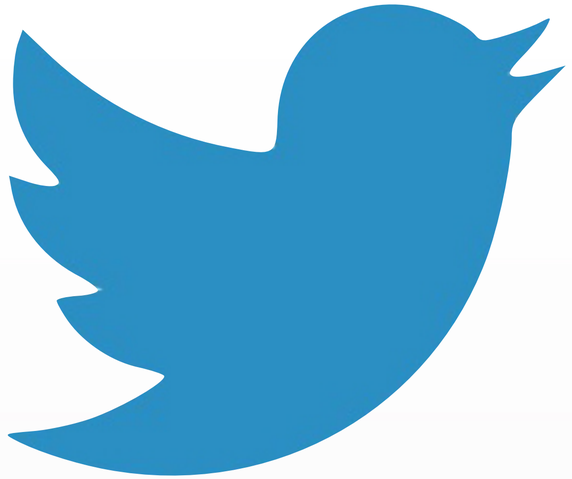
Meet the New York couple donating millions to the anti-vax movement

"They should be allowed to have the measles if they want the measles," Del Bigtree told reporters outside

When people are overwhelmed or stressed, they stop listening to traditional media.

Instead they turn to their trusted sources....





THE ROYAL SOCIETY PUBLISHING | All Journals

Home Content Information for About us Sign up Submit

ROYAL SOCIETY OPEN SCIENCE

Perspective

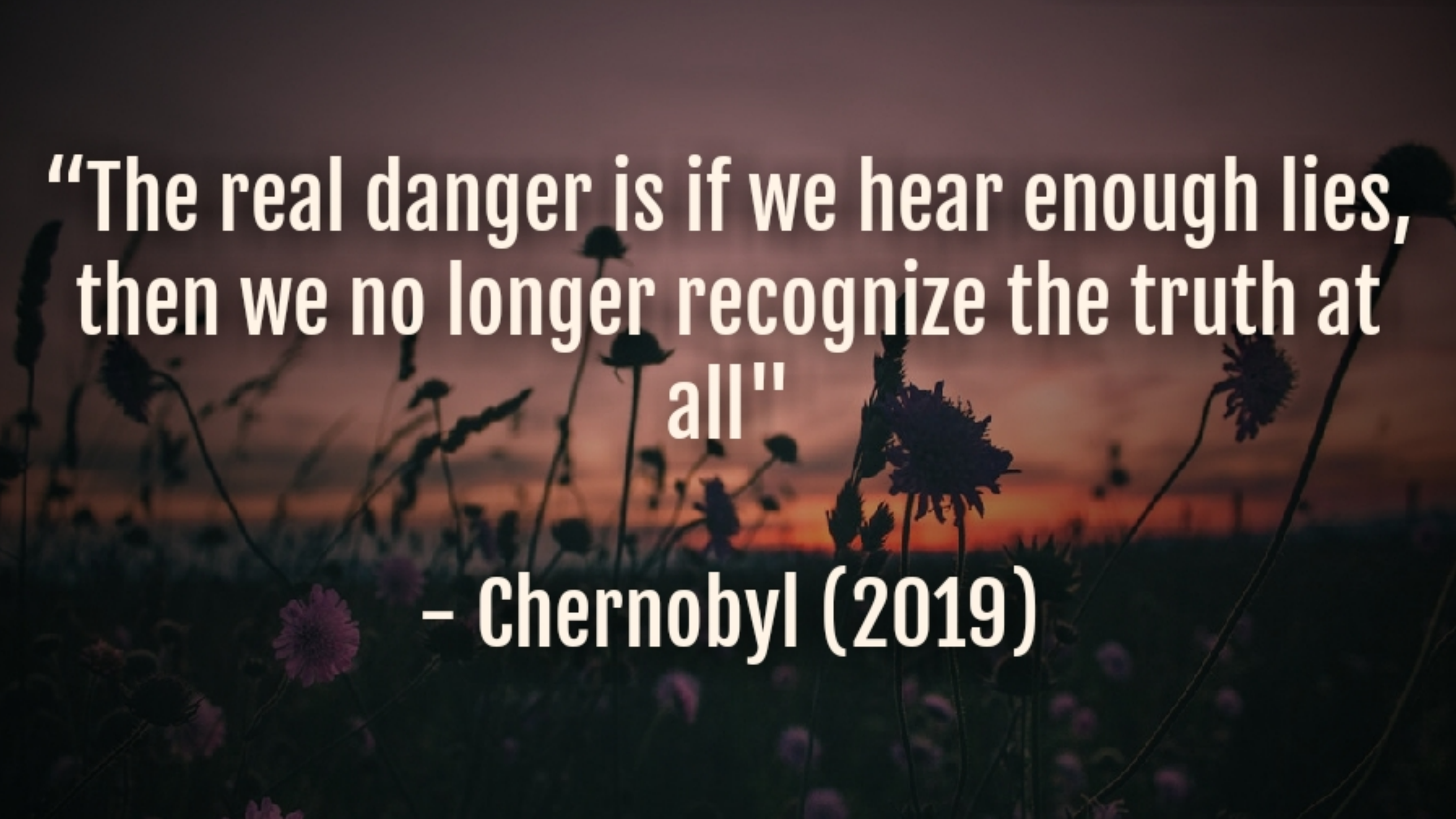
Fake science and the knowledge crisis: ignorance can be fatal

Henning Hopf, Alain Krief, Goverdhan Mehta and Stephen A. Matlin

Published: 01 May 2019 | https://doi.org/10.1098/rsos.190161

Abstract

Computers, the Internet and social media enable every individual to be a publisher, communicating true or false information instantly and globally. In the 'post-truth' era, deception is commonplace at all levels of contemporary life. Fakery affects science and social information and the two have become highly interactive globally, undermining


“The real danger is if we hear enough lies, then we no longer recognize the truth at all"
– Chernobyl (2019)

A true life example

The Rogoff-Reinhart Model

Two Harvard Professors publish a paper that says "When a country owes more than 90 percent of its GDP, it slides into recession."

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

Growth in a Time of Debt

By Carmen M. Reinhart and Kenneth S. Rogoff*

In this paper, we exploit a new multi-country historical dataset on public (government) debt to search for a systemic relationship between high public debt levels, growth and inflation.[1] Our main result is that whereas the link between growth and debt seems relatively weak at "normal" debt levels, median growth rates for countries with public debt over roughly 90 percent of GDP are about one percent lower than other-

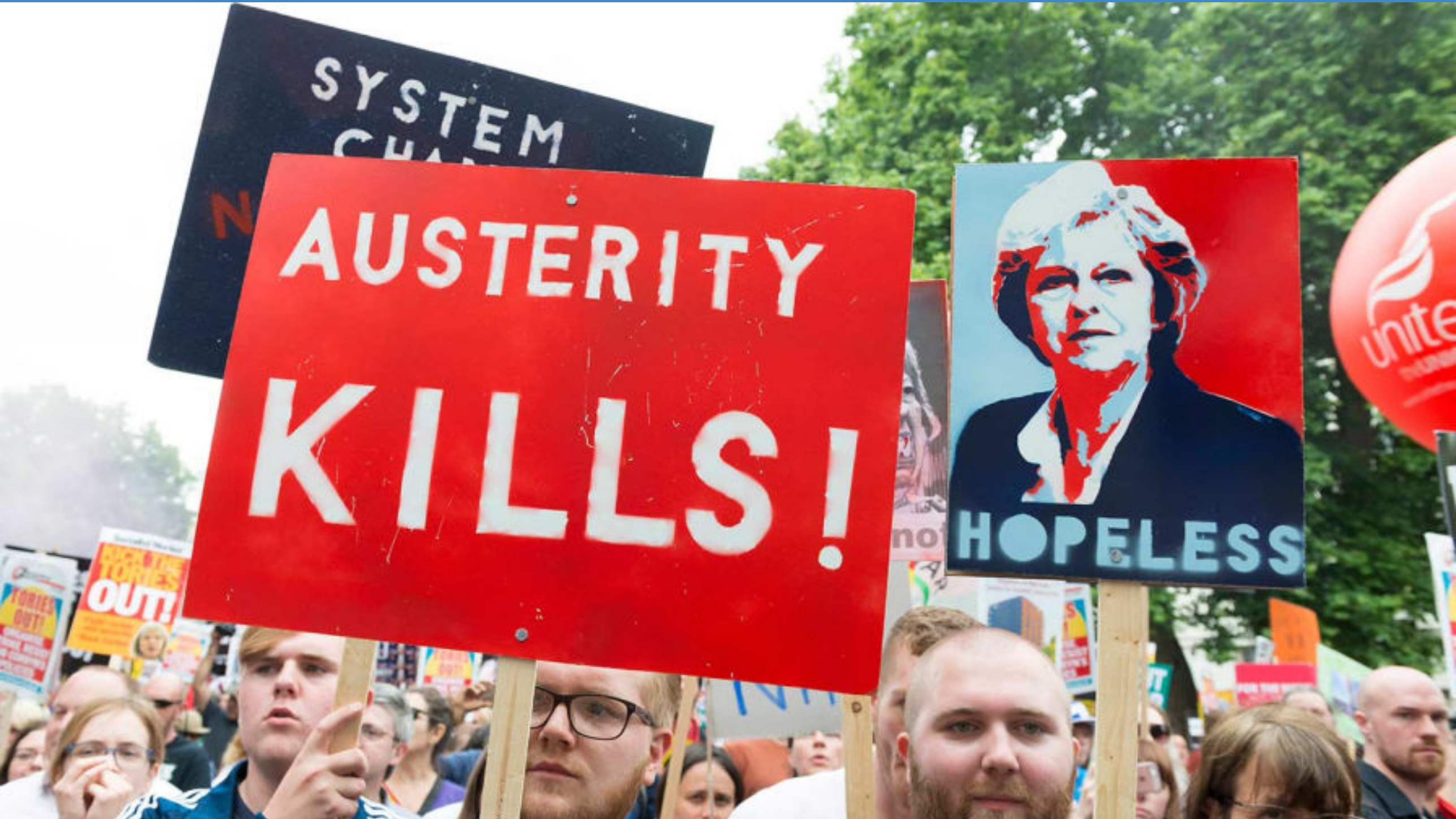
especially against the backdrop of graying populations and rising social insurance costs? Are sharply elevated public debts ultimately a manageable policy challenge?

Our approach here is decidedly empirical, taking advantage of a broad new historical dataset on public debt (in particular, central government debt) first presented in Carmen M. Reinhart and Kenneth S. Rogoff (2008, 2009b).

Thomas Herndon Discovered A Small Problem....

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	[illegible]	7.0
45	Denmark	1950-2009	3.5	1.7	2.[illegible]	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.[illegible]	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.[illegible]	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.[illegible]	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.[illegible]	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

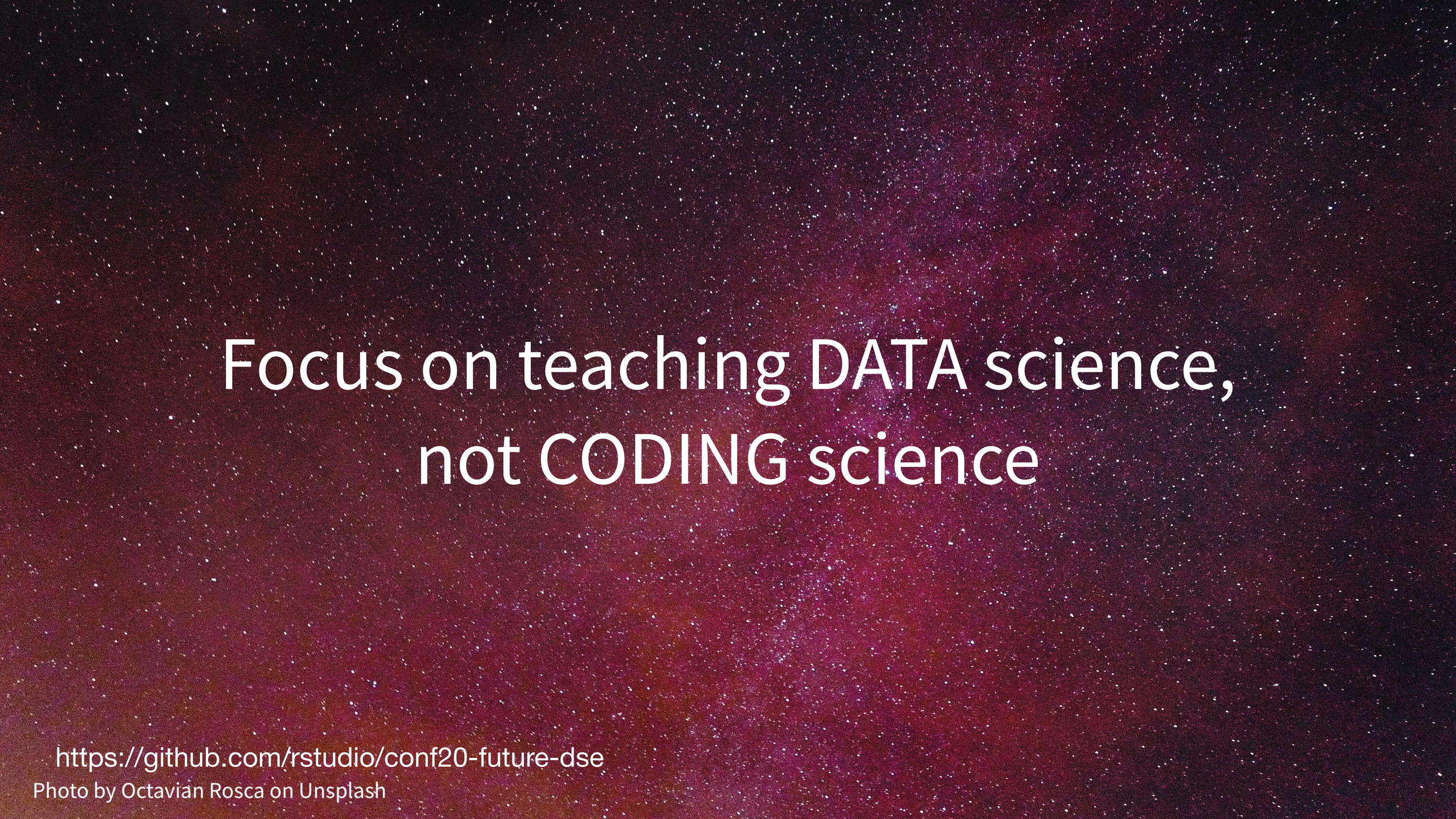
Rogoff and Reinhart's results weren't reproducible, yet they led to this....


SYSTEM
AUSTERITY
KILLS!
HOPELESS
TORIES OUT!
unite

But we see hope

An analyst is just someone who reports the obvious

before anyone else does



https://github.com/rstudio/conf20-future-dse

Photo by Octavian Rosca on Unsplash

Challenges

- Data explosion
- Unreproducible results
- Fake data

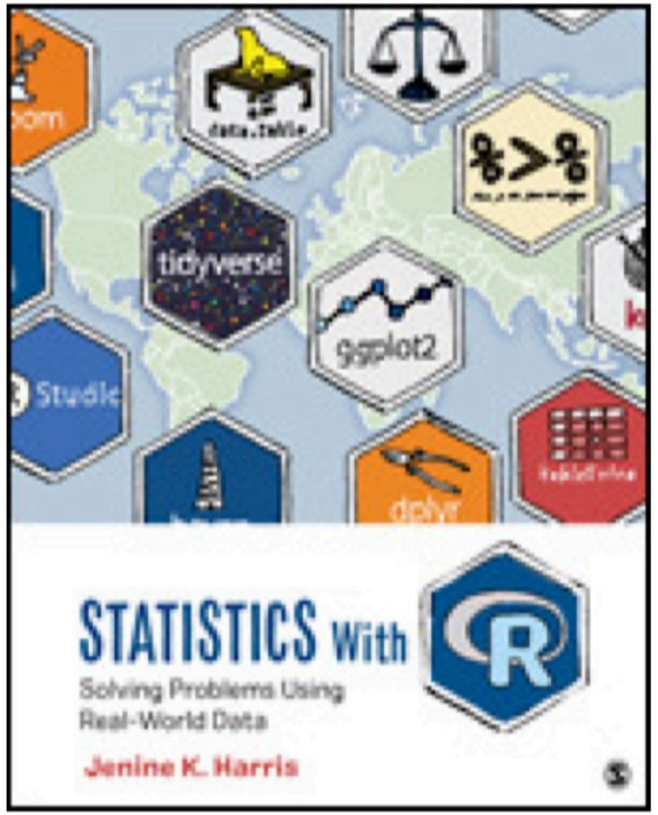
Changes

- Embrace statistics

1. Data science embraces statistics on real data

Statistics because data is too big

Real data because the real world is messy



Statistics With R

Solving Problems Using Real-World Data

Jenine K. Harris - Washington University in St.Louis, USA

January 2020 | 784 pages | SAGE Publications, Inc

Challenges

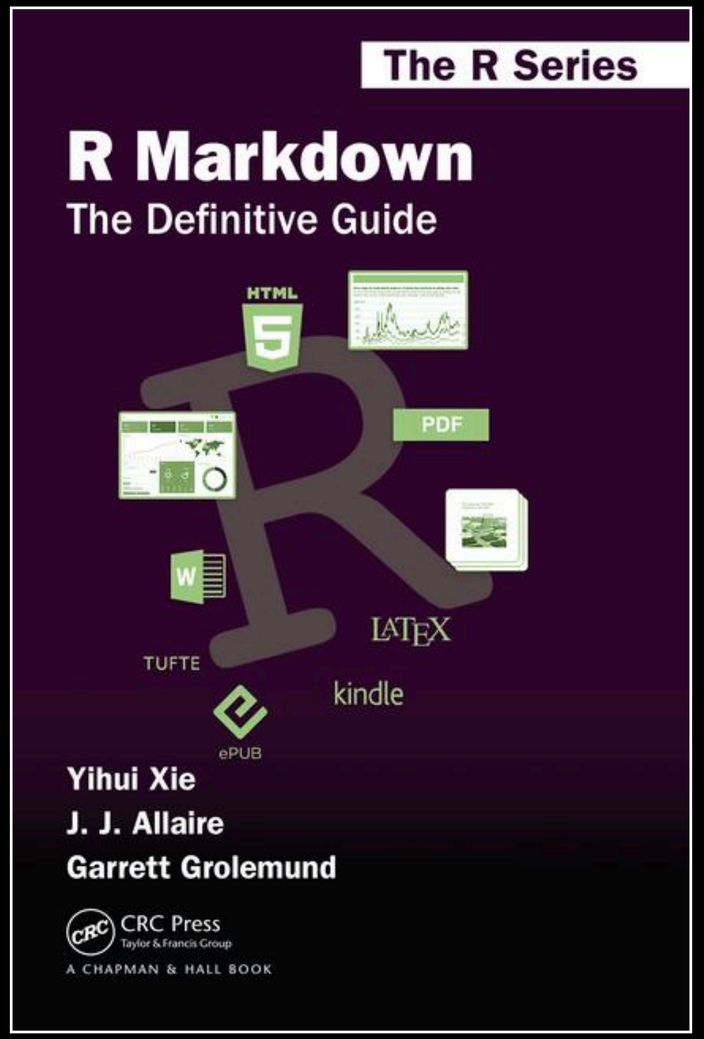
- Data explosion
- Unreproducible results
- Fake data

Changes

- Embrace statistics
- Public processes

2. Good data science relies on R Markdown computational documents and open processes

http://**bookdown.org/yihui/rmarkdown**



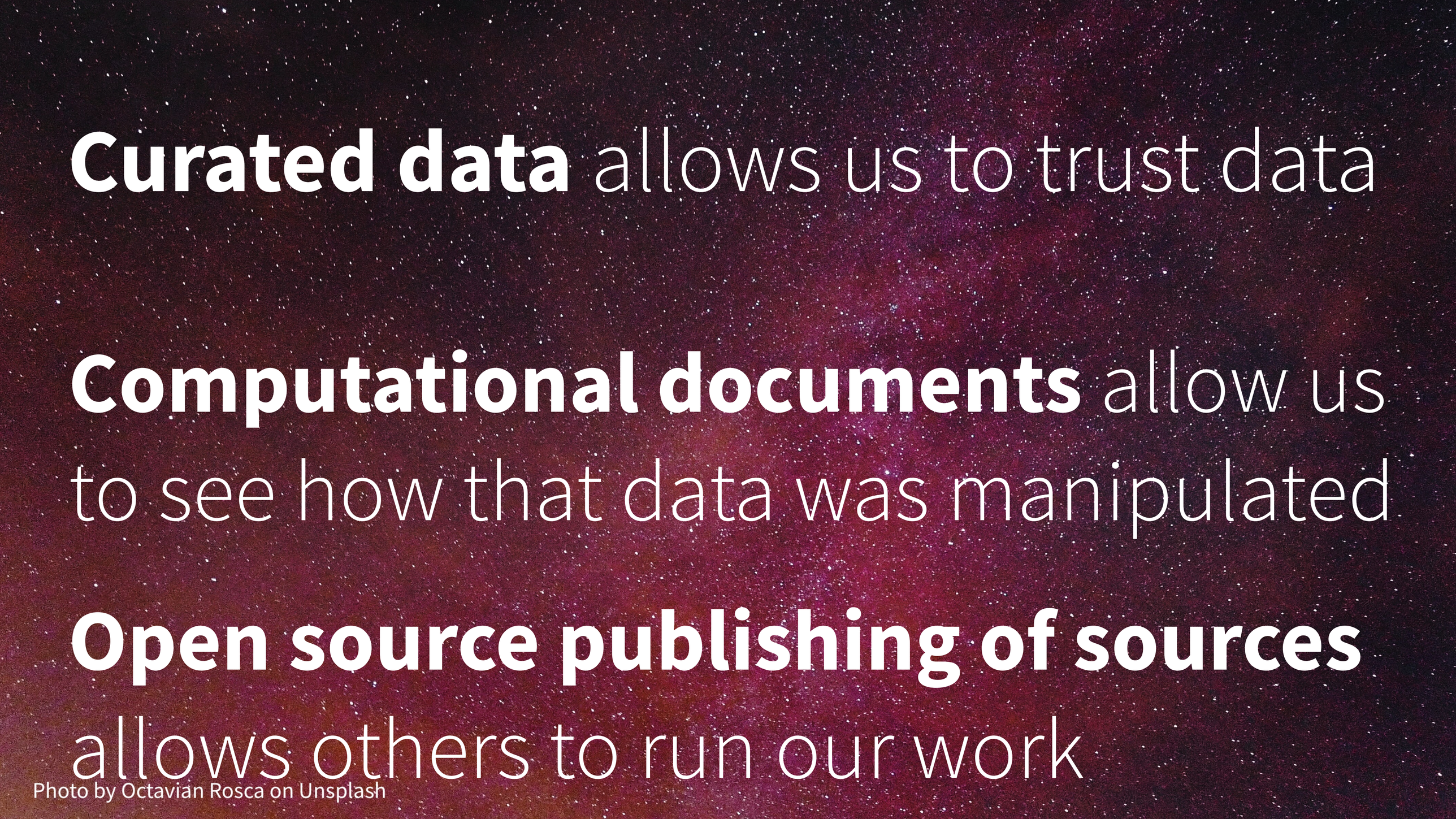


Photo by Octavian Rosca on Unsplash

Challenges

- Data explosion
- Unreproducible results
- Fake data

Changes

- Embrace statistics
- Public processes
- Authoritative storytellers

3. We must democratize data science

R Ladies

Our democratized experts must be viral storytellers

Who would you trust more to diagnose your illness?

1. A data scientist who learned some medicine?
2. A doctor who learned some data science?

Stephan Kadauke

Assistant Lab Director - CHOP

Division of Pathology Informatics

Course Director - Reproducible Clinical Data Analysis with R/RStudio

RStudio Certified Instructor

A delightful series of modules to learn statistics and R coding for students, scientists, and stats-enthusiasts.

desirée de leon

DESIRÉE DE LEON

nsf-grfp fellow

emory university

Photo by Etienne Girardet on Unsplash

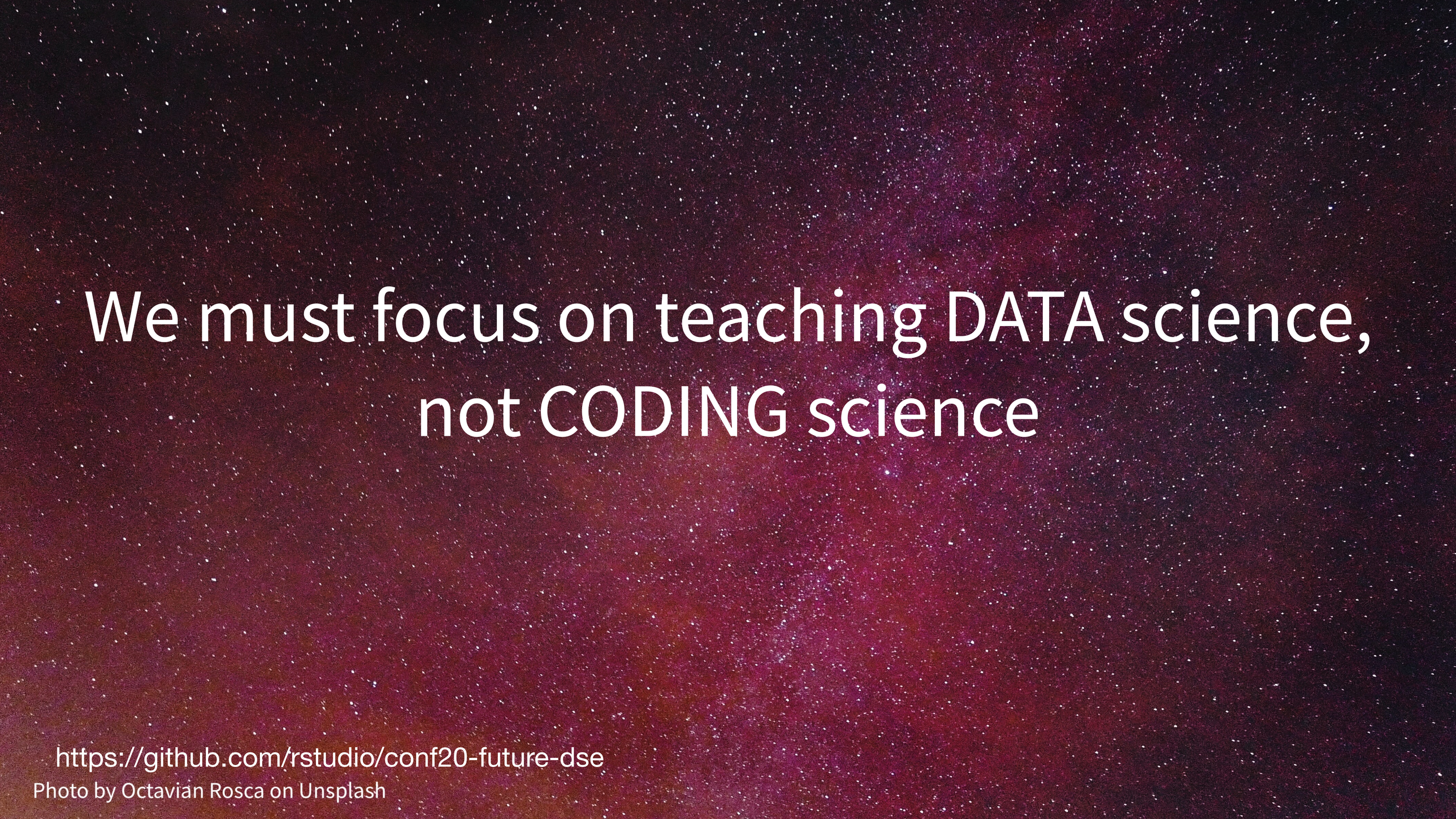
rstudio::conf

Challenges

- Data explosion
- Unreproducible results
- Fake data

Changes

- Embrace statistics
- Public processes
- Authoritative storytellers



https://github.com/rstudio/conf20-future-dse

Photo by Octavian Rosca on Unsplash


Too soon to tell, it is....


RStudio®
Each of you can help us
overcome the new
challenges in data
science education


RStudio®
Help everyone, regardless of means, participate in a global economy that rewards data literacy


RStudio®
Materials at:
https://github.com/rstudio/conf20-future-dse