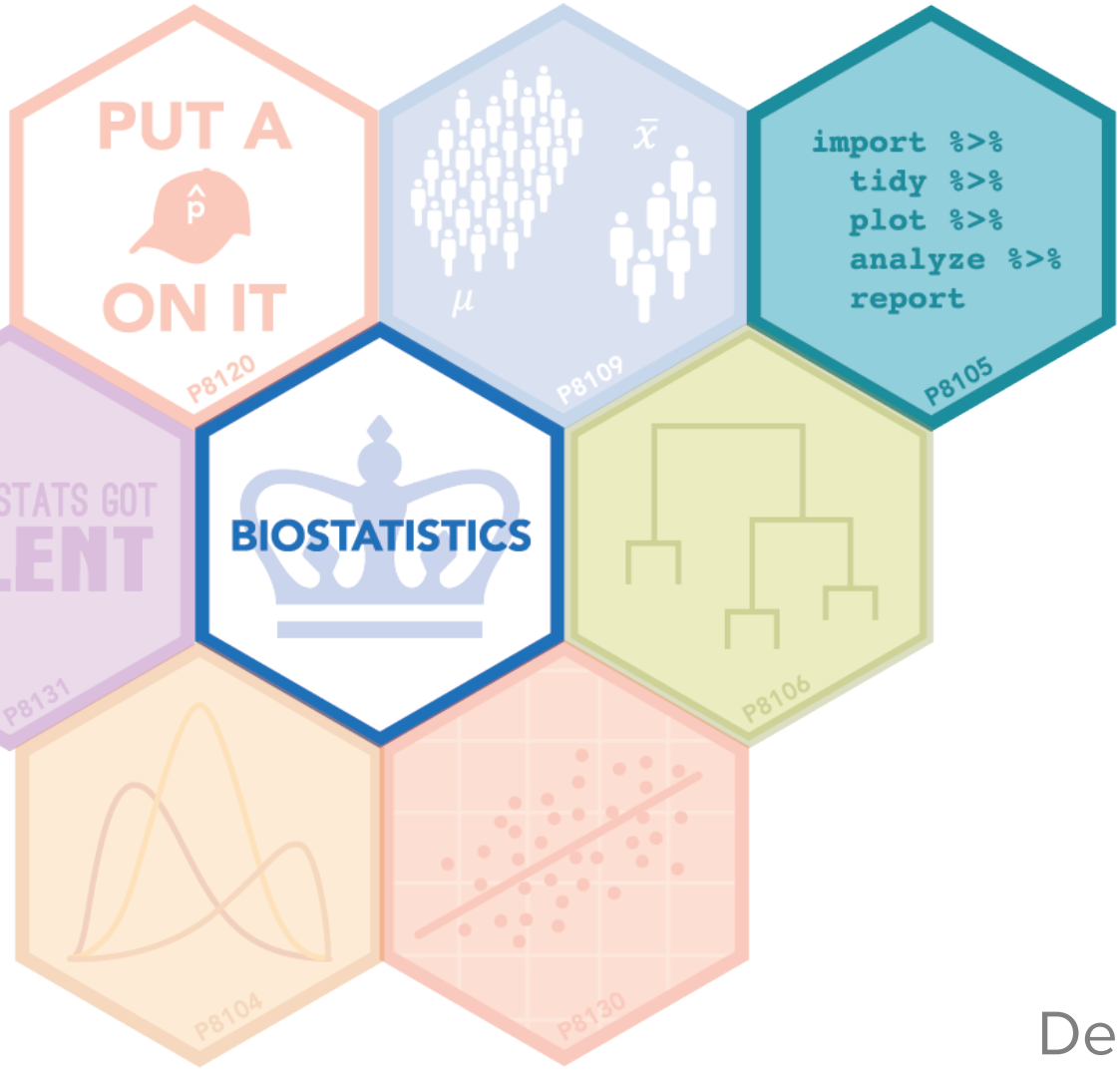


WHAT IS DATA SCIENCE?

Jeff Goldsmith, PhD
Department of Biostatistics

Data science is pretty new

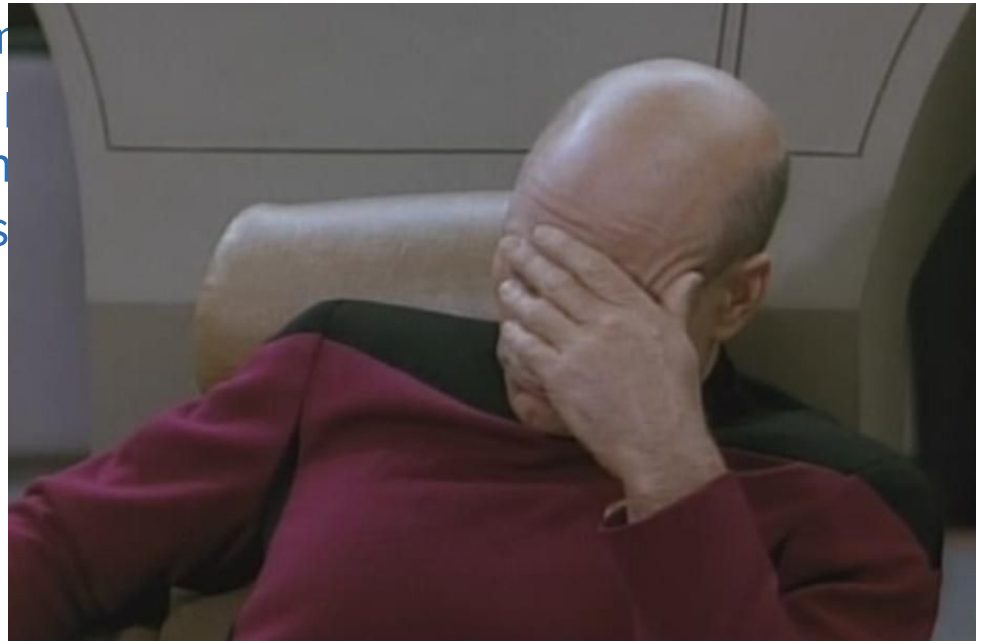
Source: Google Trends

Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- "A data scientist is just a sexier word for statistician." –Nate Silver
- "A data scientist is a better computer scientist than a statistician and is a better statistician than a computer scientist."
- "A data scientist is a statistician who is useful" – Hadley Wickham
- A data scientist is a good statistical analyst
- A data scientist is a statistician who codes in python

Some not great definitions

- Data science = statistics
- Data science = computer science
- Data science = machine learning
- Data science = statistics + computer science + machine learning
- Data scientists are big data wranglers
- "A data scientist is just a sexier word for statistician." –Nate Silver
- "A data scientist is a better com
  a better statistician than a com
- "A data scientist is a statistician
- A data scientist is a good statis
- A data scientist is a statistician



Maybe pictures will help?

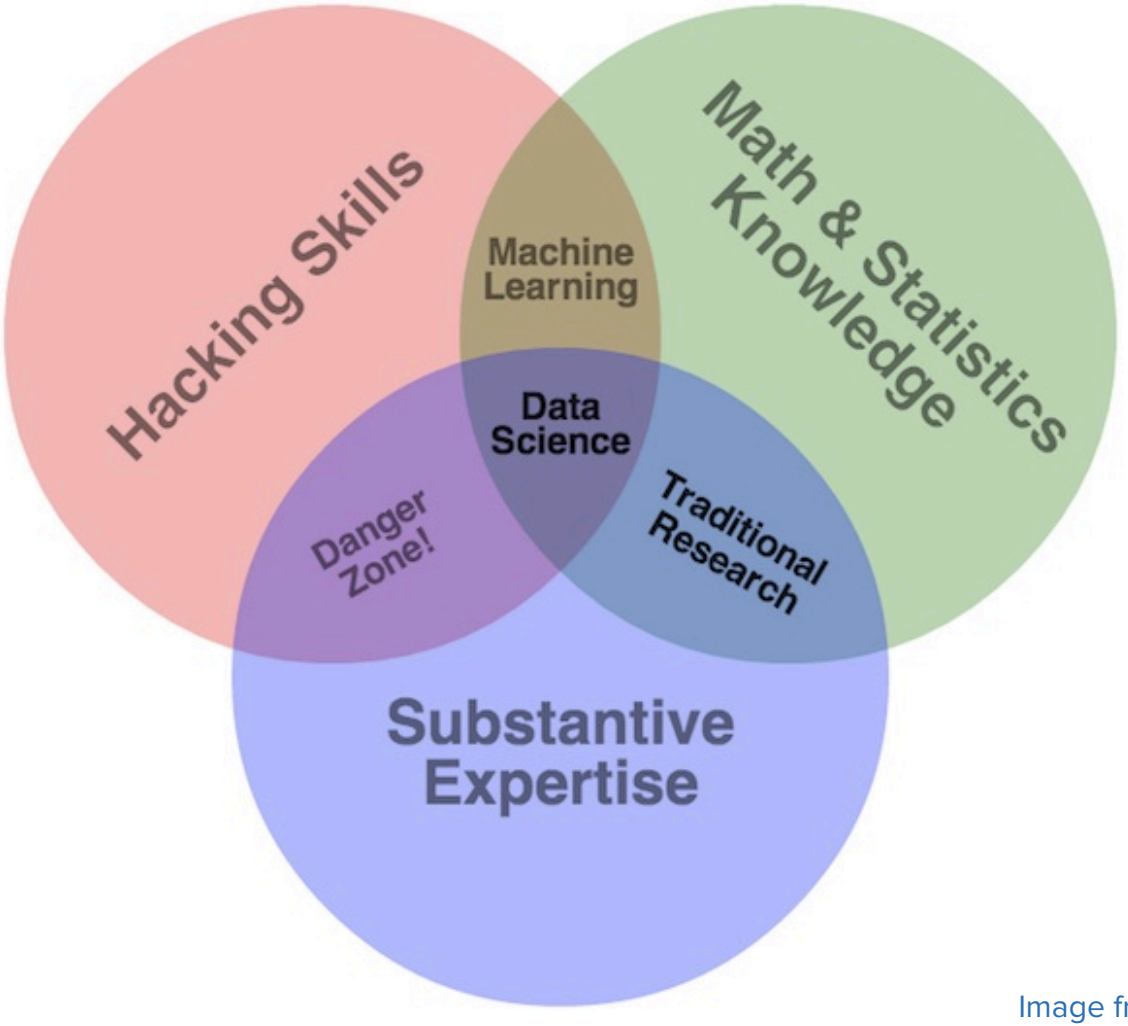
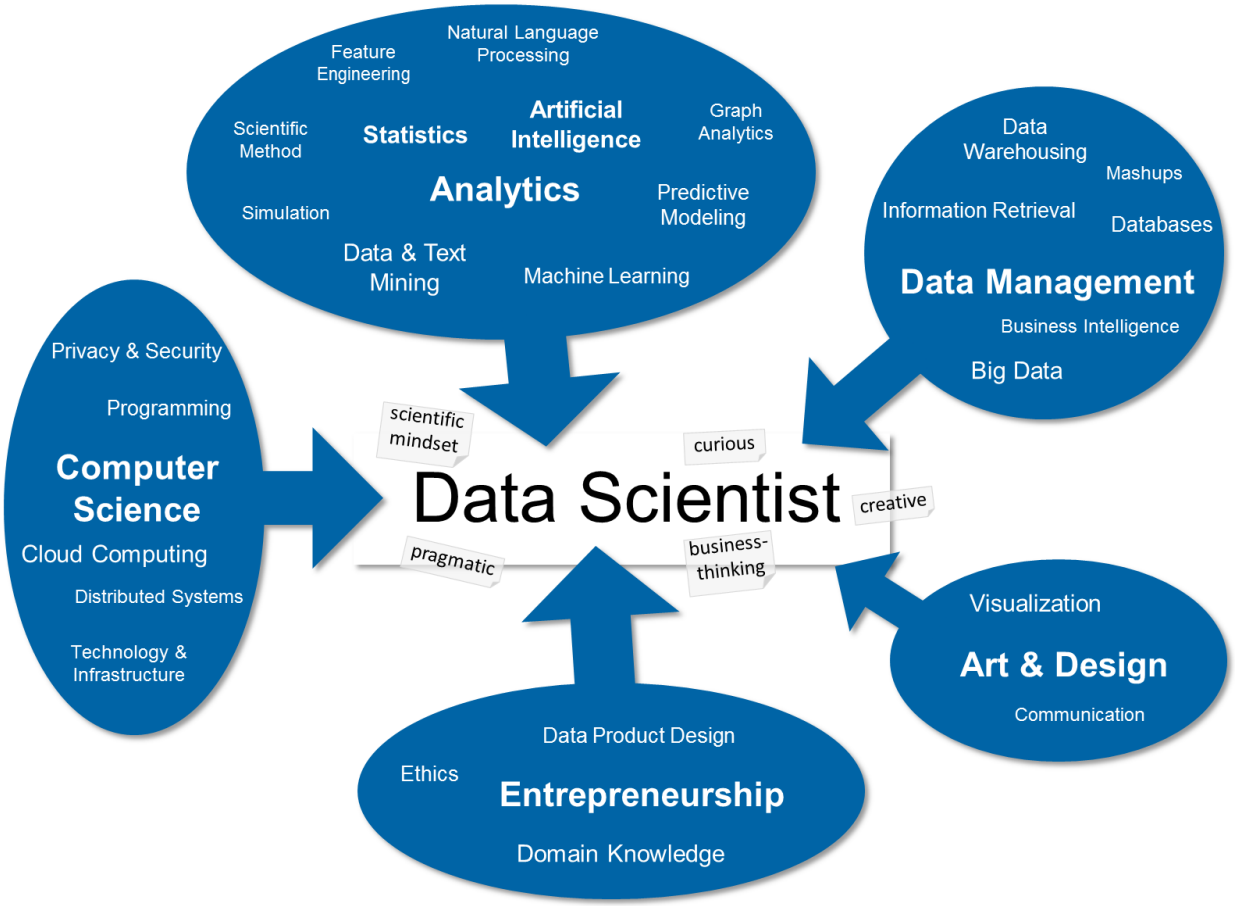


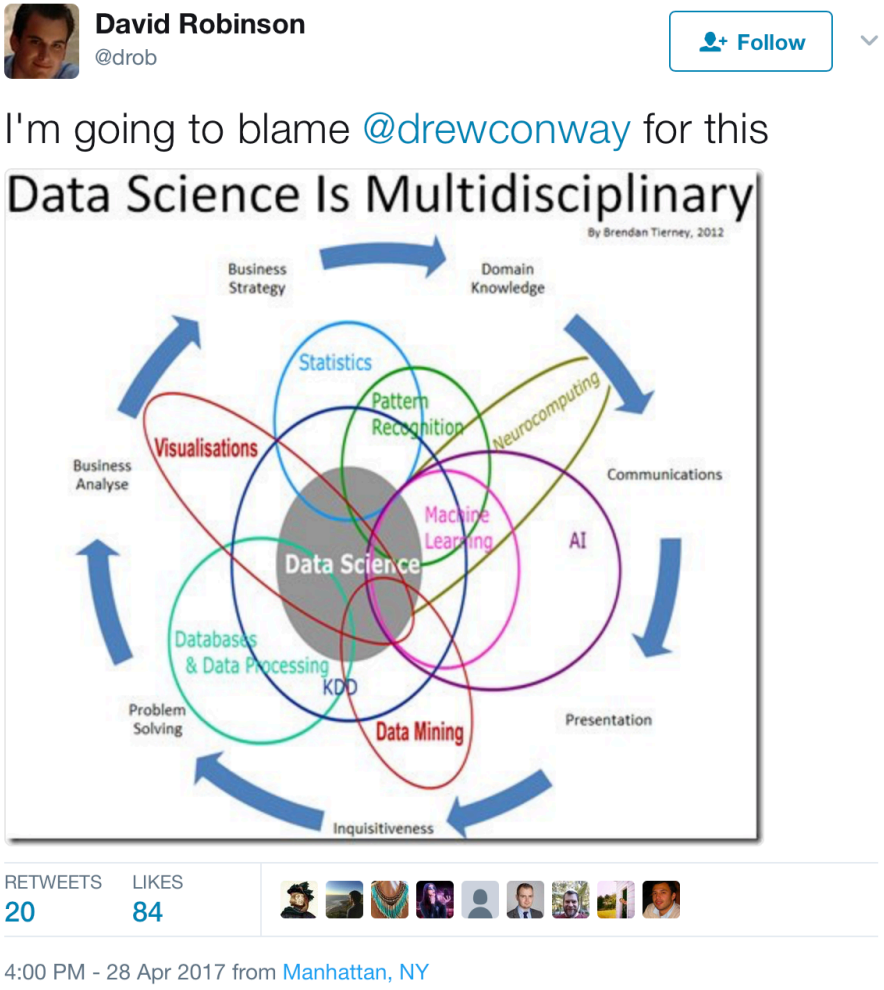
Image from Drew Conway

Maybe pictures will help?



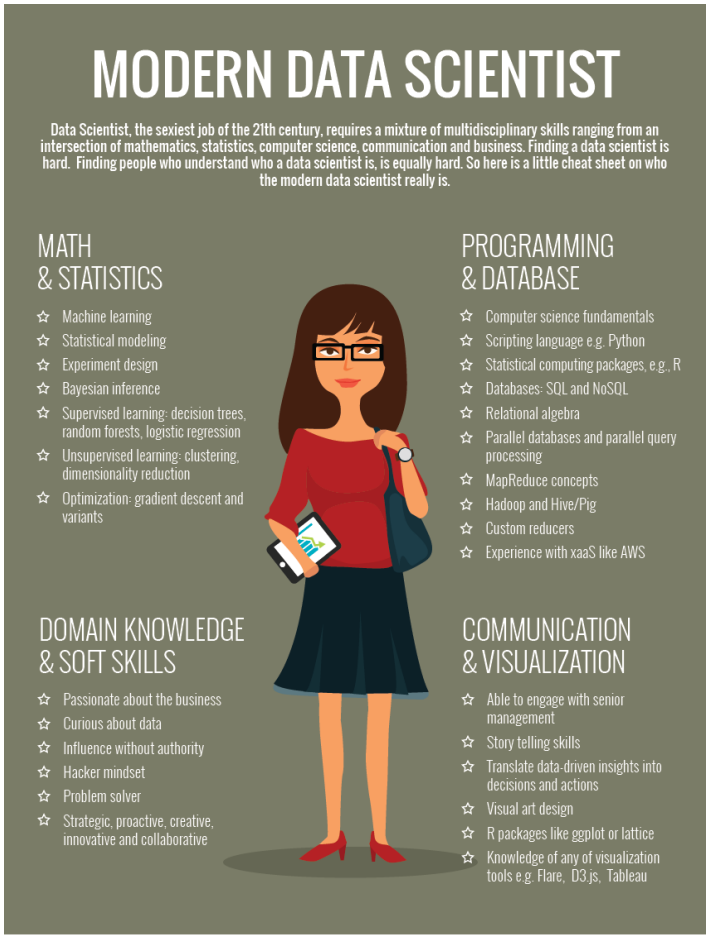
https://blog.zhaw.ch/datascience/the-data-science-skill-set/

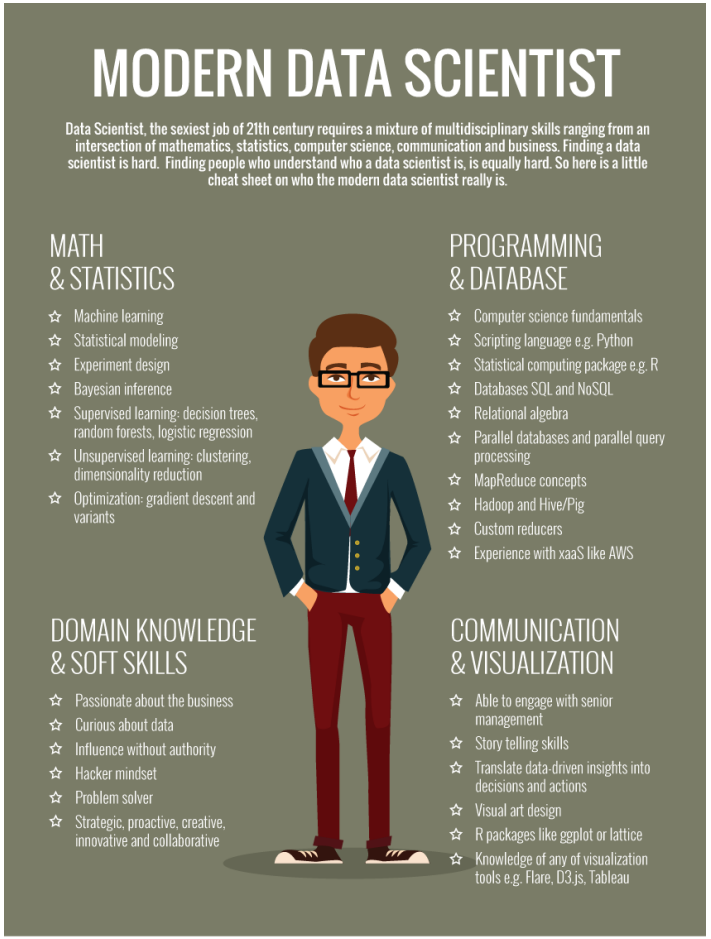
Maybe pictures will help?



From twitter

Maybe pictures will help?





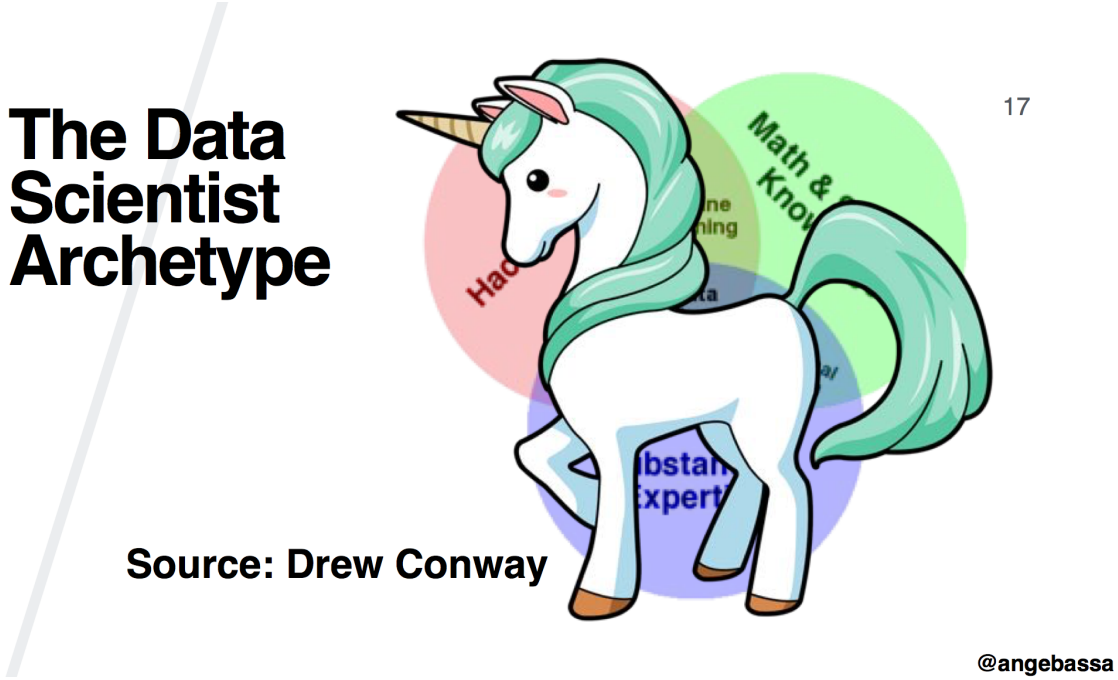
Why these definitions are bad

- "Data science is just …" definitions miss the point
  - If data science is just statistics (or machine learning, or computer science, or engineering) we wouldn't need a new term, let alone a new discipline
  - The popularity of "data science" suggests that there's a newly recognized need

- "A data scientist is a good " whatever definitions aren't helpful
  - They're almost deliberately judgmental
  - A good definition doesn't depend on opinions
  - There are "data scientists" in each discipline, but some very good statisticians / computer scientists / etc aren't "data scientists"

Why these definitions are bad

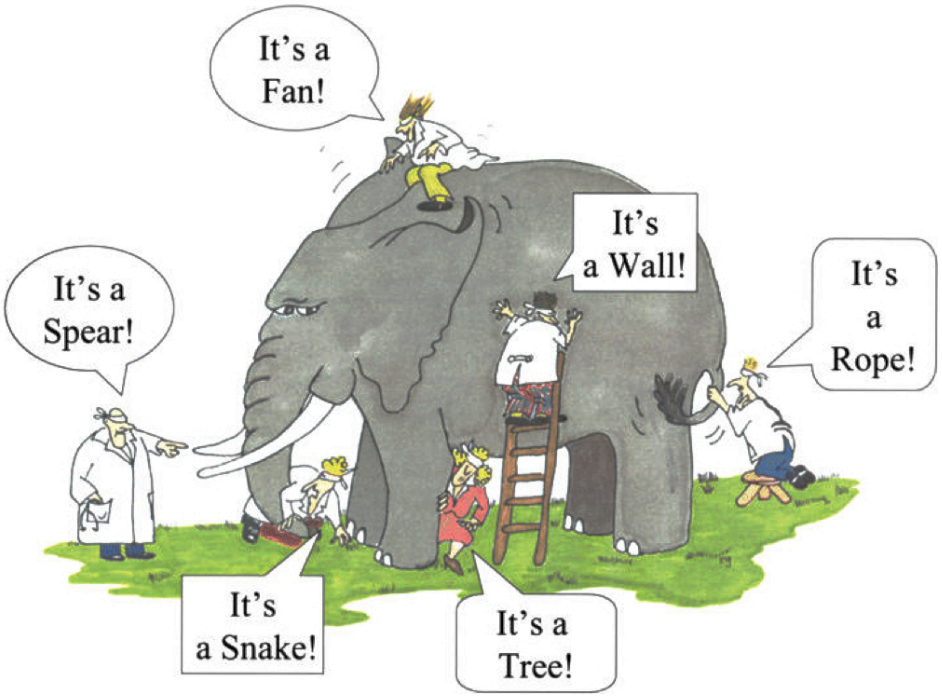
- "Data science is the combination of these 40 skills …" are unrealistic



https://www.youtube.com/watch?v=b9ZLXwAuUyw&app=desktop

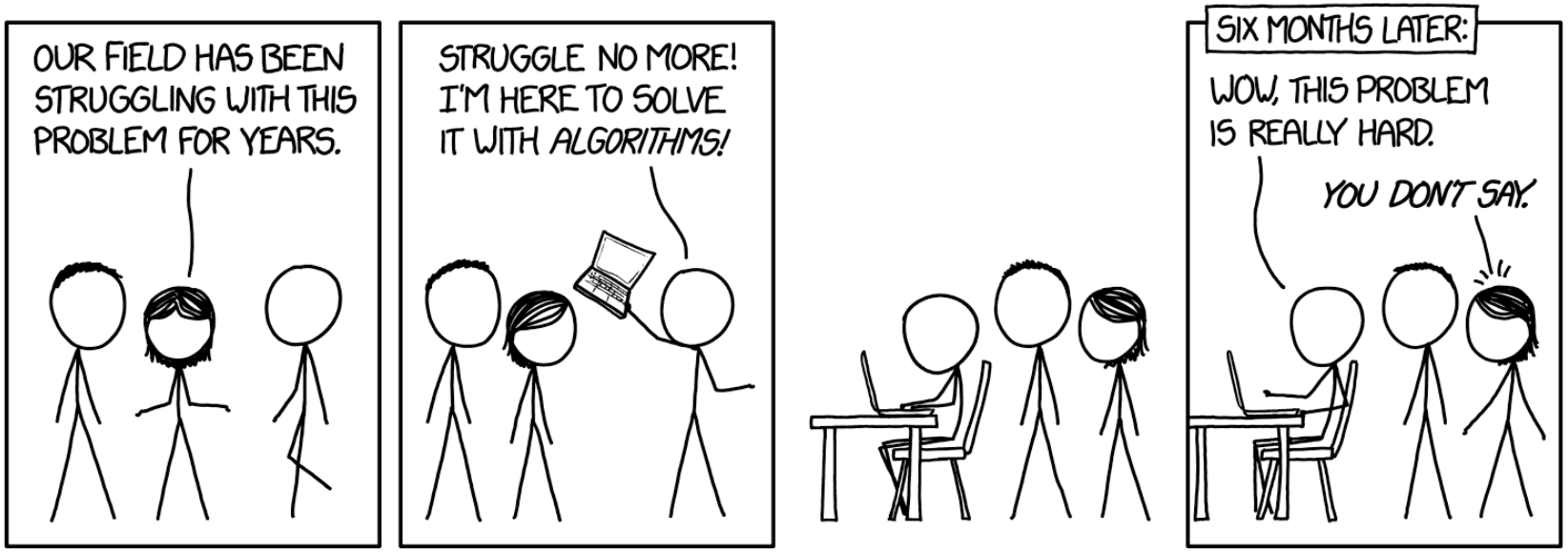
Why these definitions are good

- Kinda like the blind men and the elephant – no one perspective is completely right or completely wrong, but piling them all up isn't right either
- They give a sense of what is valued by the data science community – using data in a principled way and coding well



Why these definitions are good

- Data science is interdisciplinary
  - You do need a breadth of skills
  - You also need a particular mindset – curiosity and engagement is critical
  - You need some domain knowledge to be successful



https://www.xkcd.com/1831/

Is “data science” a buzzword?

- It is used to describe a frustratingly wide collection of ideas
- It is also used in different (sometimes contradictory) ways by different people

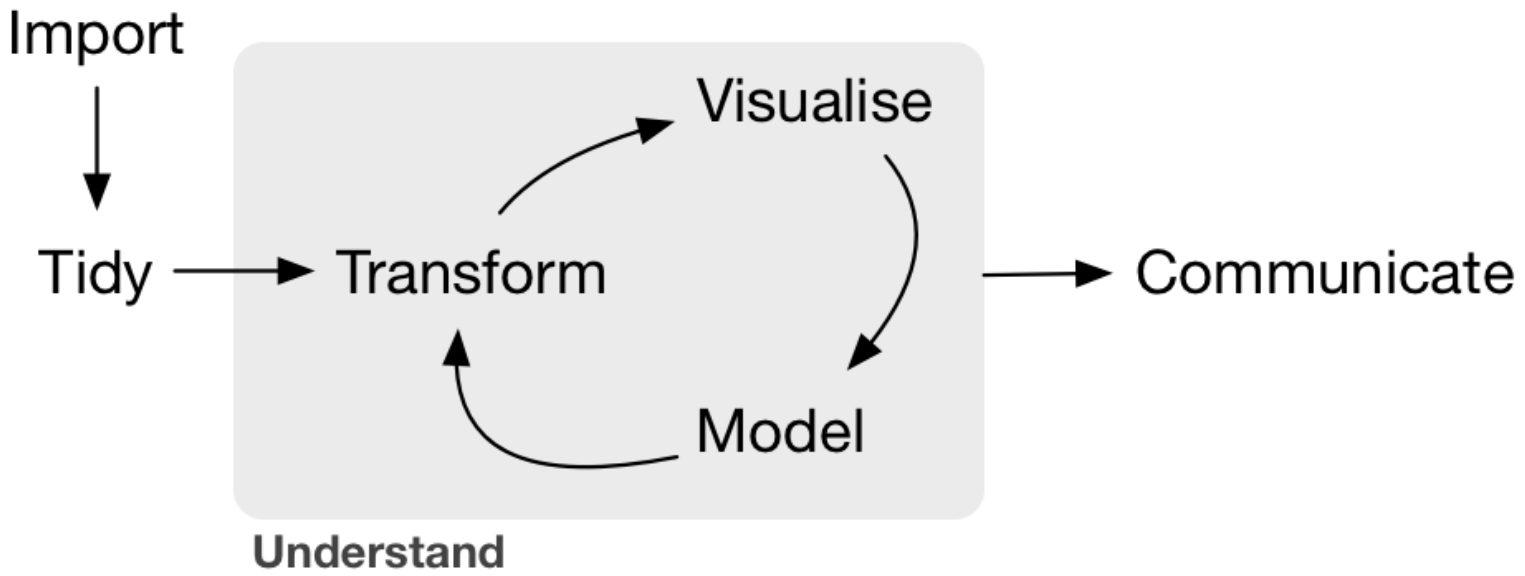
- Nonetheless, the popularity of “data science” reflects an increasingly data-centric reality
- Dismissing “data science” ignores this reality, and blinds people to the usefulness of a new perspective across disciplines

For the purpose of this class:

Data science is the use of data to formulate and answer questions in a process that emphasizes clarity, reproducibility, and collaboration, and that recognizes code as a primary means of communication.

- We'll focus mostly on process; how to answer questions through analyses are the focus of other courses

A data exploration diagram



R for Data Science

How to learn data science

- Build a broad knowledge base
- Don't be embarrassed by what you don't know
  - Corollary: don't be a jerk to people who don't know what you know
- Ask questions (well) and keep learning

- Pretty much the same as learning anything, but hard because people don't like to show their code

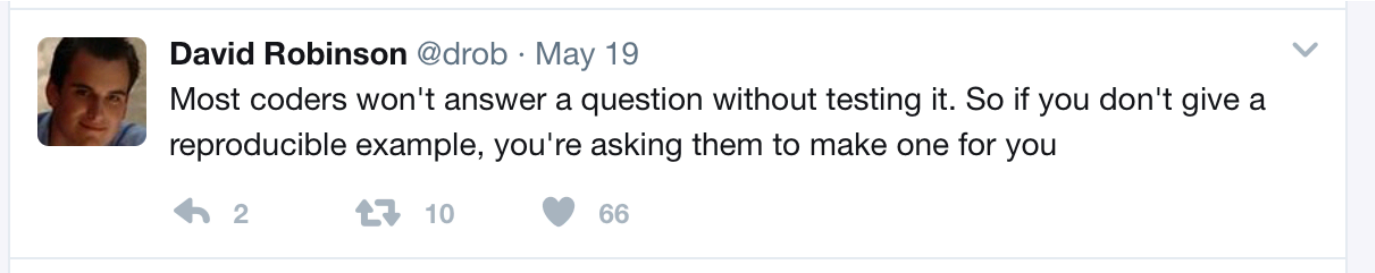
How to learn data science

- Build a broad knowledge base
- Don't be embarrassed by what you do
  - Corollary: don't be a jerk to people who do what you know
- Ask questions (well) and keep learning

- Pretty much the same as learning any hard because people don't like to sho code

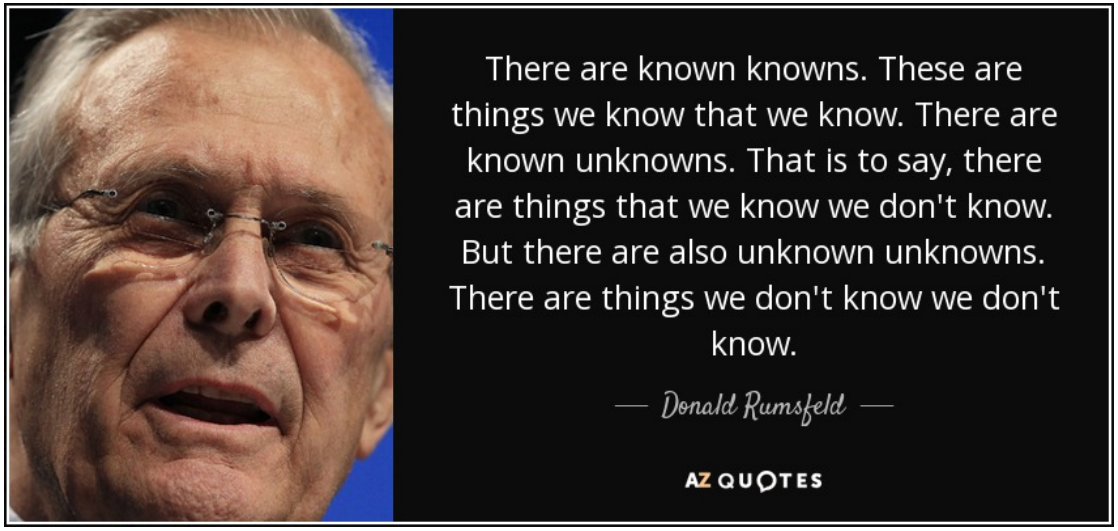
How to learn data science

- All questions are good questions, but sometimes good questions aren't asked well
- Think through what you're trying to ask
- If your code is broken, come up with a simple example that illustrates what's broken



How to learn data science

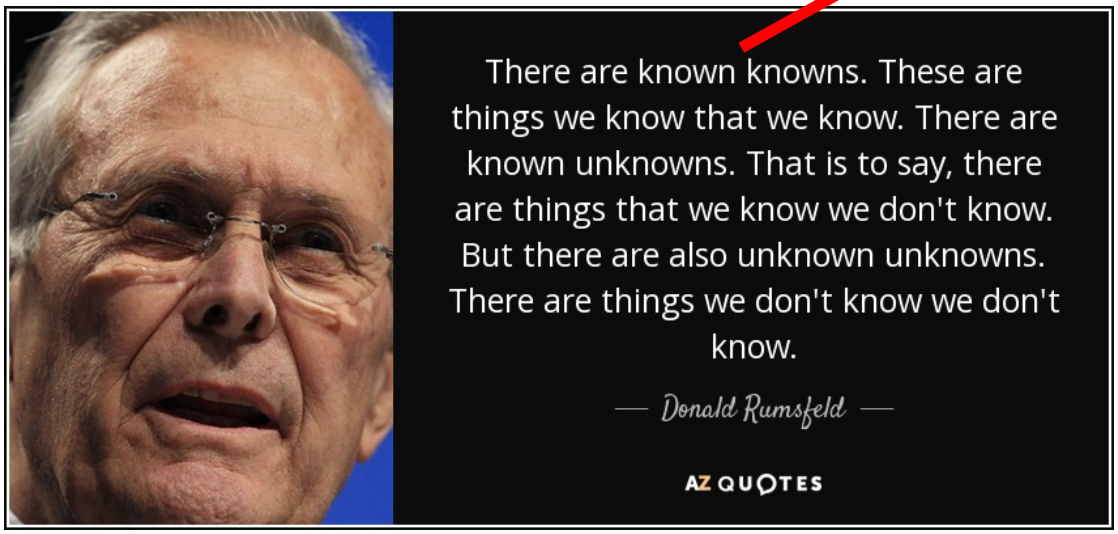
- Be on the lookout for cool stuff!



How to learn data science

- Be on the lookout for cool stuff!

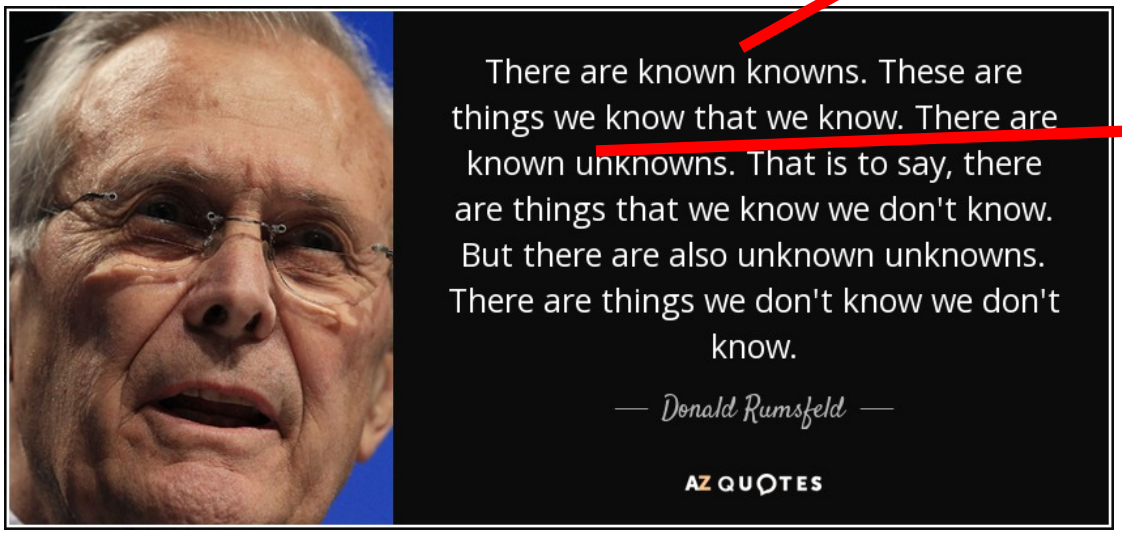
Knowledge base! :-D



How to learn data science

- Be on the lookout for cool stuff!

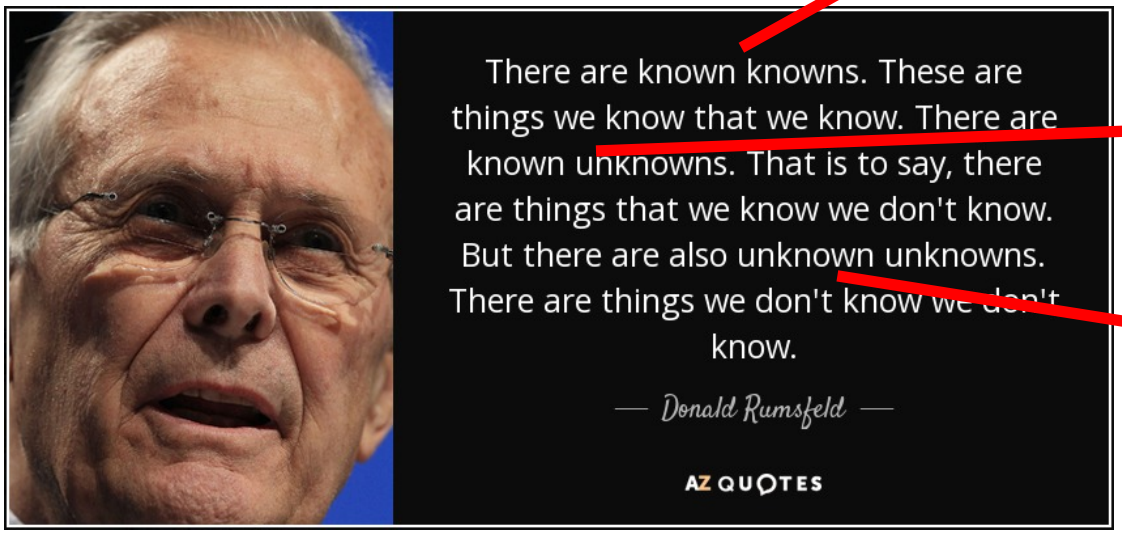
Knowledge base! :-D



Things you know exist and can learn how to do :-)

How to learn data science

- Be on the lookout for cool stuff!



Knowledge base! :-D

Things you know exist and can learn how to do :-)

Things you don't know exist and can't use :-(

DS twitter starter pack

- Follow these people to add some “knowns” to your repertoire

- @AmeliaMN
- @dataandme
- @drewconway
- @drob
- @hadleywickham
- @hmason
- @hspter
- @_inundata
- @jennybryan
- @johnmyleswhite
- @jtleek
- @juliasilge
- @kara_woo
- @kwbroman
- @rdpeng
- @seanjtaylor
- @sgrifter
- @statpumpkin
- @xieyihui
- @rOpenSci
- @rstudio
- @simplystats

Data as a resource

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Data as a resource

The world's most valuable resource is no longe

The data economy d

Sections ☰ The Washington Post

wp BrandStudio ⓘ Content from **IBM Power Systems**

Why big data is "the new natural resource"

Data as a resource

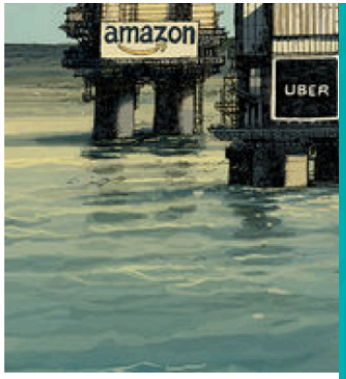
The world's most valuable resource
is no longe

The data economy d

Sections ☰ The Washington Post

wp BrandStudio ⓘ Content from **IBM Power Systems**

Is Data The New Oil? How One Startup Is Rescuing The World's Most Valuable Asset



"the new natural resource"

Data in health and medicine

Data are everywhere

- Clinical trials
- Observational studies
- Genomics
- Medical imaging
- Microbiome

Data in health and medicine

Data are everywhere

- Clinical trials
- Observational studies
- Genomics
- Medical imaging
- Microbiome

- Electronic health records?
- Mobile health technologies?
- Twitter posts?
- Search terms?
- Social networks?

A public health lens

How can we use these data to improve health?

- Improve surveillance, leading to better prevention efforts?
- Better understanding of mechanisms?
- More precise and more effective outreach?

- Doing something simple and useful is better

Google flu trends

Google

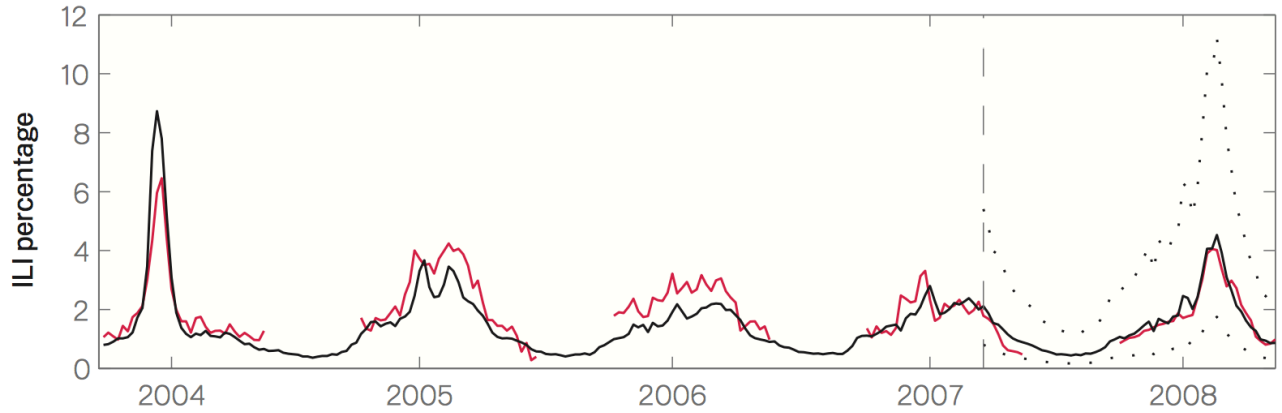
Detecting influenza epidemics using search engine query data

Google flu trends

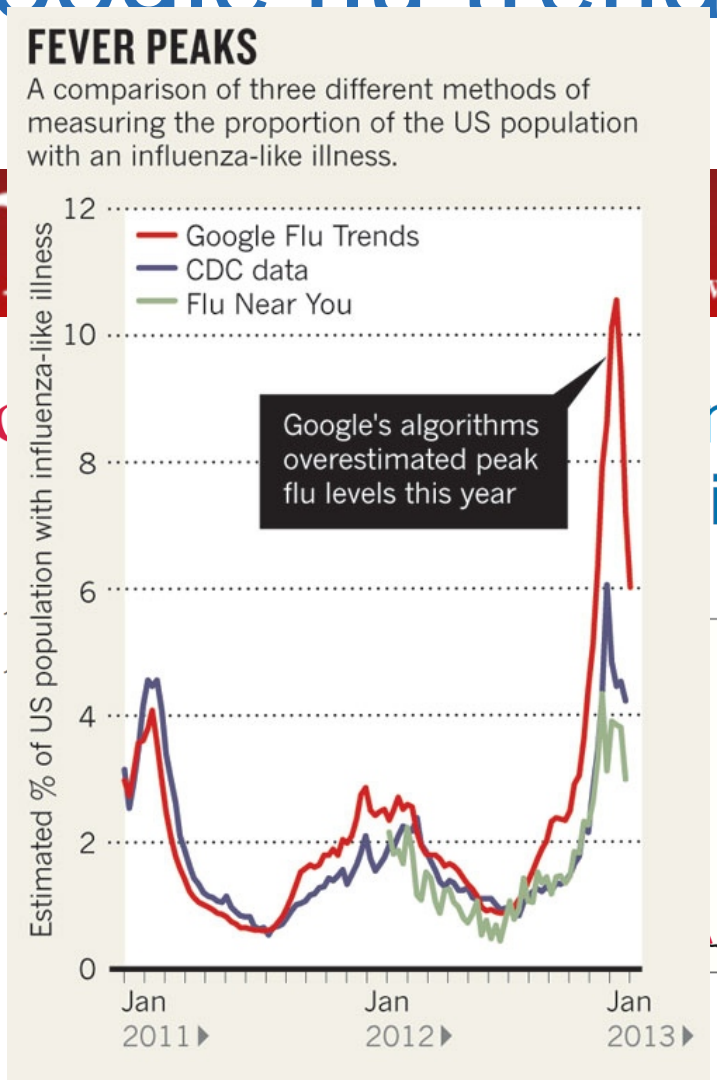
nature International weekly journal of science

Google

Detecting influenza epidemics using search engine query data



Google flu trends



weekly journal of science

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive

Archive > Volume 494 > Issue 7436 > News > Article

NATURE | NEWS

عربي

When Google got flu wrong

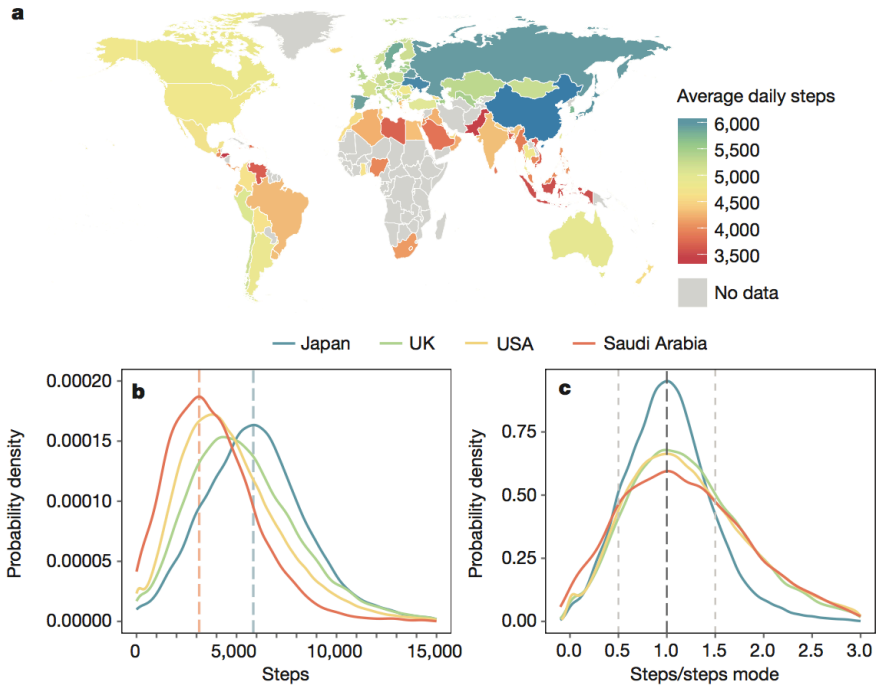
US outbreak foxes a leading web-based method for tracking seasonal flu.

2007 2008

Activity via smartphones

nature International weekly journal of science

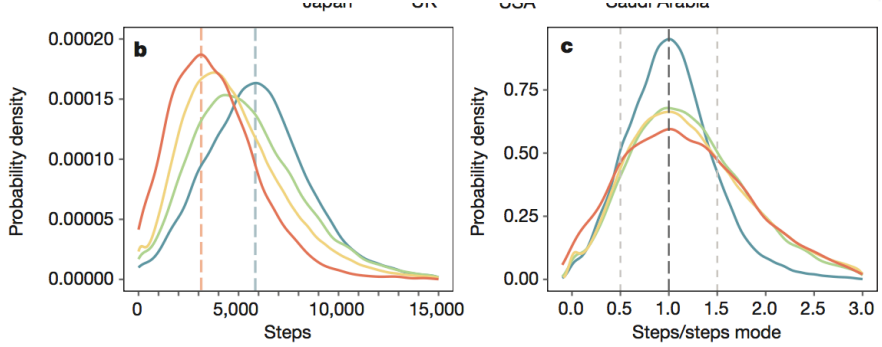
Large-scale physical activity data reveal worldwide activity inequality



Activity via smartphones

nature International weekly journal of science

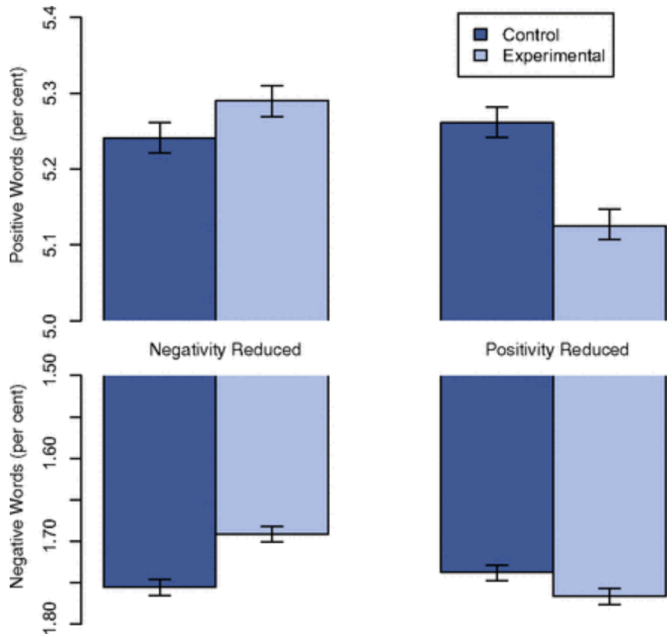
We study 68 million days of minute-by-minute step recordings from 717,527 anonymized users of the Argus smartphone application developed by Azumio. The dataset includes recordings of physical activity for free-living individuals from 111 countries (Fig. 1a). We focus on the 46 countries with at least 1,000 users (Supplementary Table 1); 90% of these users were from 32 high-income countries and 10% were from 14 middle-income countries (including five lower-middle-income countries; Methods). The average user recorded 4,961 steps per day



Facebook experiments

PNAS

Experimental evidence of massive-scale emotional contagion through social networks

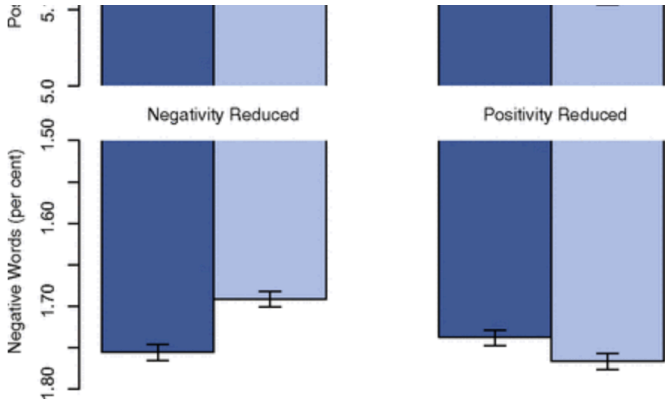


Facebook experiments

PNAS

Significance

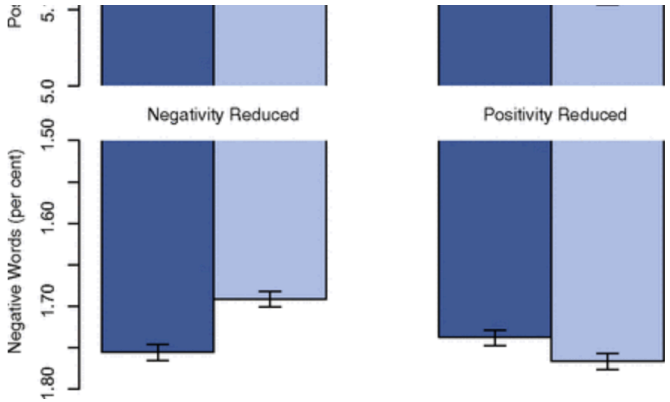
We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.



BBC NEWS Facebook experiments

Facebook admits failings over emotion manipulation study

We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.



BBC ook experiments

NEWS

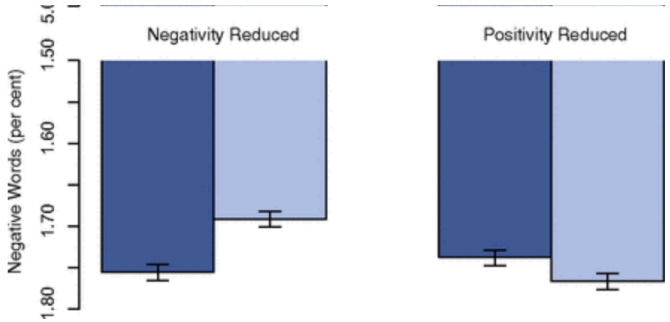
A C

Facebook admits failings over emotion manipulation study

theguardian

cebook, that emotional states can be
ople to experience the same emotions without
emotional contagion occurs without direct

Facebook reveals news feed experiment to control emotions



BBC ook experiments

NEWS

Facebook admits failings over emotion manipulation study

theguardian

cebook, that emotional states can be
ople to experience the same emotions without
emotional contagion occurs without direct

Facebook reveals news feed experiment to control emotions

Forbes / Tech

Facebook Manipulated 689,003 Users' Emotions For Science

BBC ook experiments

NEWS

Fa

ma

th

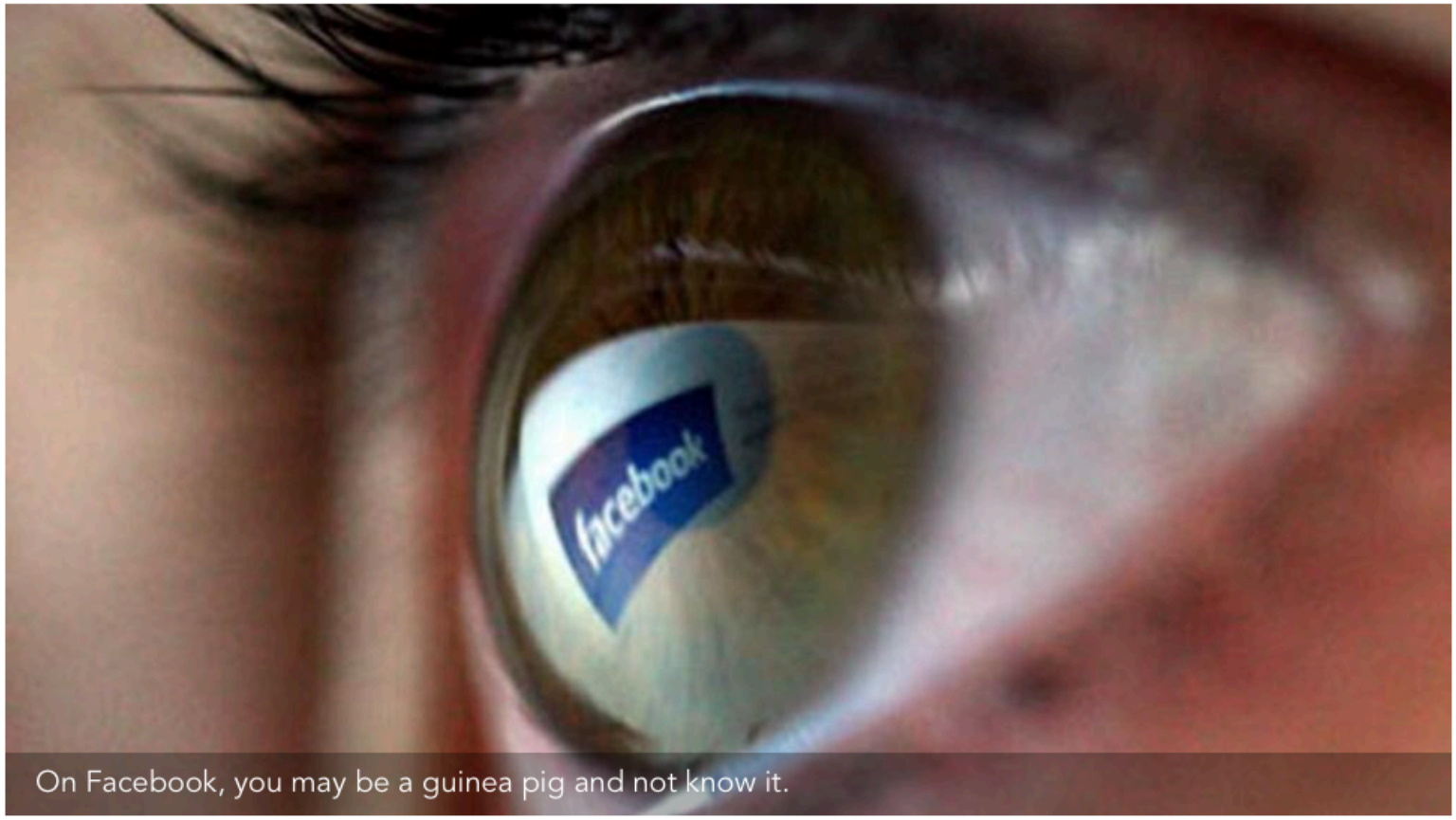
"Fa

c

to

Fo

ut



On Facebook, you may be a guinea pig and not know it.

Faceb For Science

Limitations of data

- Not trying to gang up on Google, Facebook and smartphone
  - In each case, these are smart people doing interesting thing with cool data

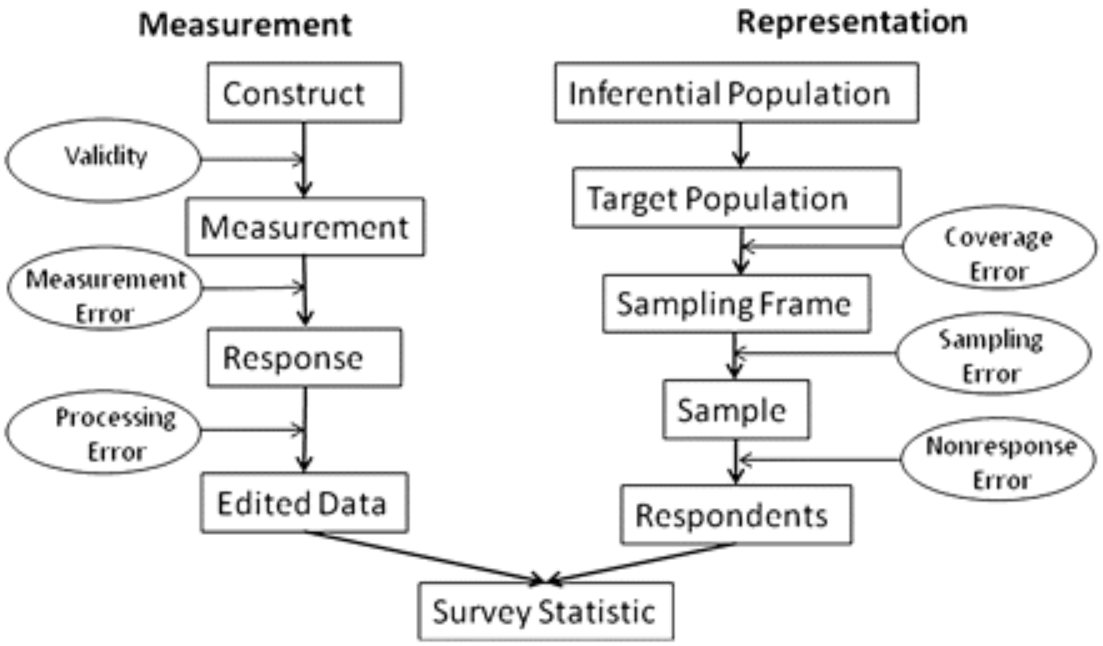
- These cases point to challenges to be overcome, and are important opportunities
  - How can public health practitioners engage with non-traditional partners in a beneficial way?
  - How can tech be used or evaluated as a public health tool when it changes so rapidly?
  - How can big data overcome issues of selection bias and access?

A public health lens

How can we use these data to improve health?

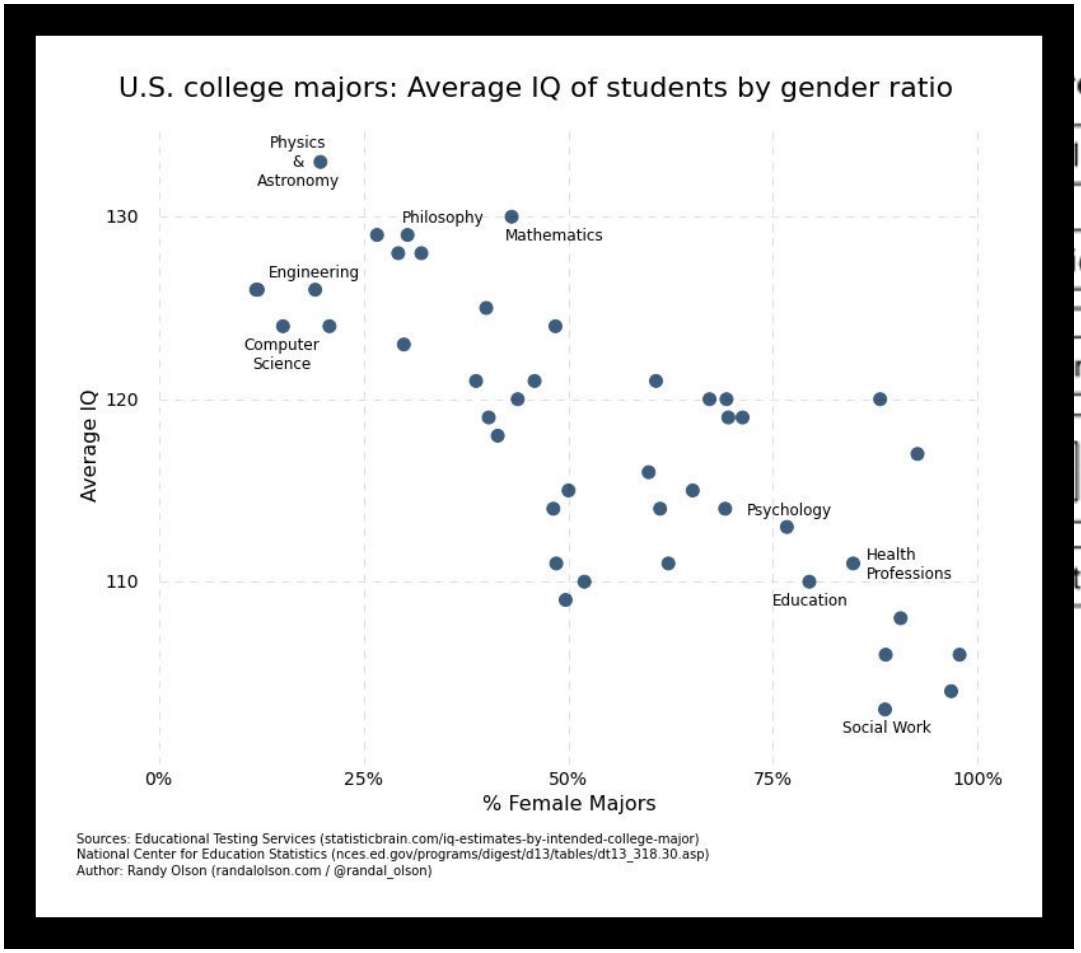
- Who is represented?
- What do measurements mean? Can they be trusted?
- Will I uncover associations? Causes?
- Can I develop new hypotheses or confirm existing ones?
- Can I design an intervention, or evaluate the success of one?

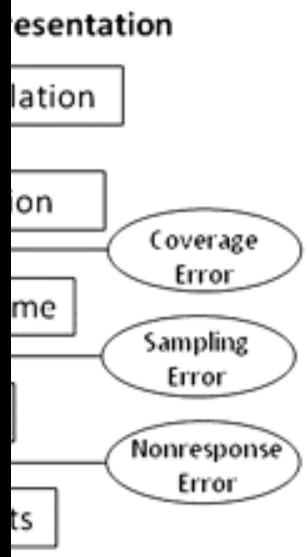
Be skeptical about data



From "Total Survey Error: Past, Present, and Future" (Groves and Lyberg)
via "Data Alone Isn't Ground Truth" by Angela Bassa

Be skeptical about data



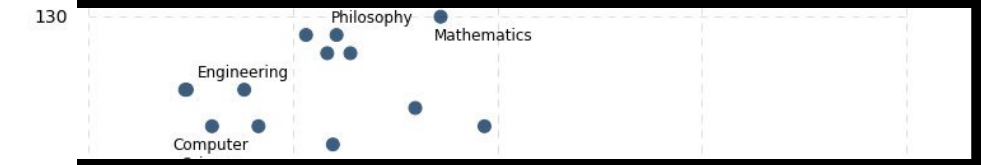


From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
via “Data Alone Isn’t Ground Truth” by Angela Bassa

Be skeptical about data

Untrustworthy Data Will Lead to Poor Conclusions

Trusting all data as if it were fact is a dangerous proposition.



So, can any data ever be trusted?

The short answer is… *it depends*. Skepticism is not a free pass to disregard data you disagree with. It's a tool to ensure that the conclusions derived from data are reliable and do, in fact, reflect reality.

You also shouldn't trust data just because it "proves" a point that you're already inclined to believe. It's probably even more important to be skeptical of extraordinary claims with which your heuristics already naturally align.

From "Total Survey Error: Past, Present, and Future" (Groves and Lyberg)
via "Data Alone Isn't Ground Truth" by Angela Bassa

A caveat before starting …

People sometimes confuse fancy methods for data science.

Don't Do That.

A simple method applied to good data and clearly communicated

is **much** better than

a fancy method that no one understands applied to bad data.