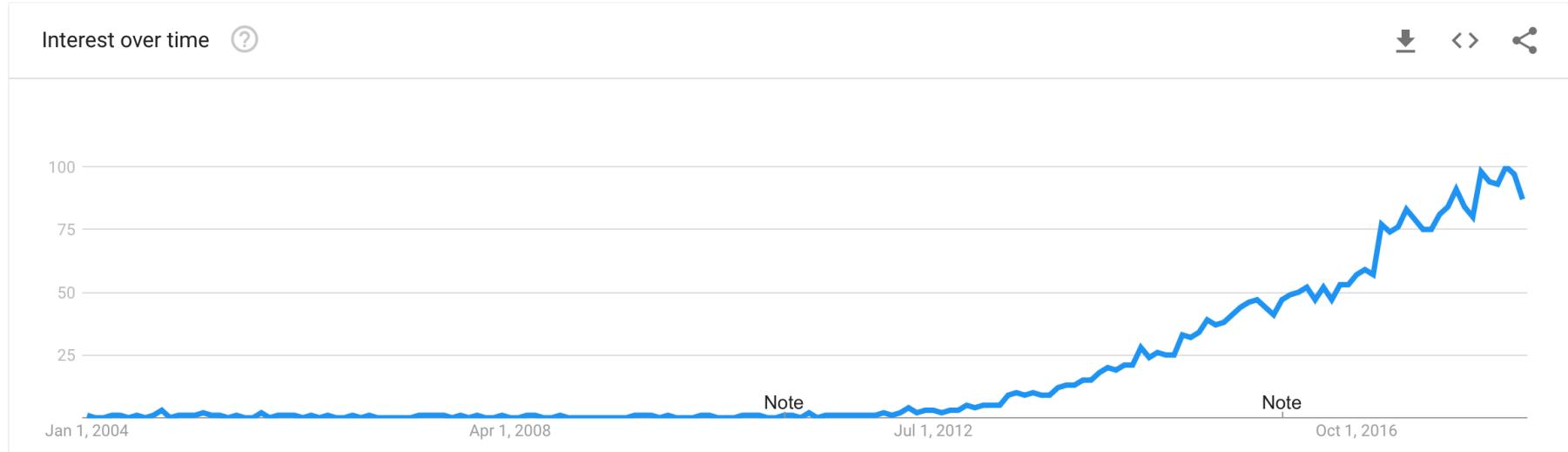


# WHAT IS DATA SCIENCE?

Jeff Goldsmith, PhD  
Department of Biostatistics

# Data science is pretty new



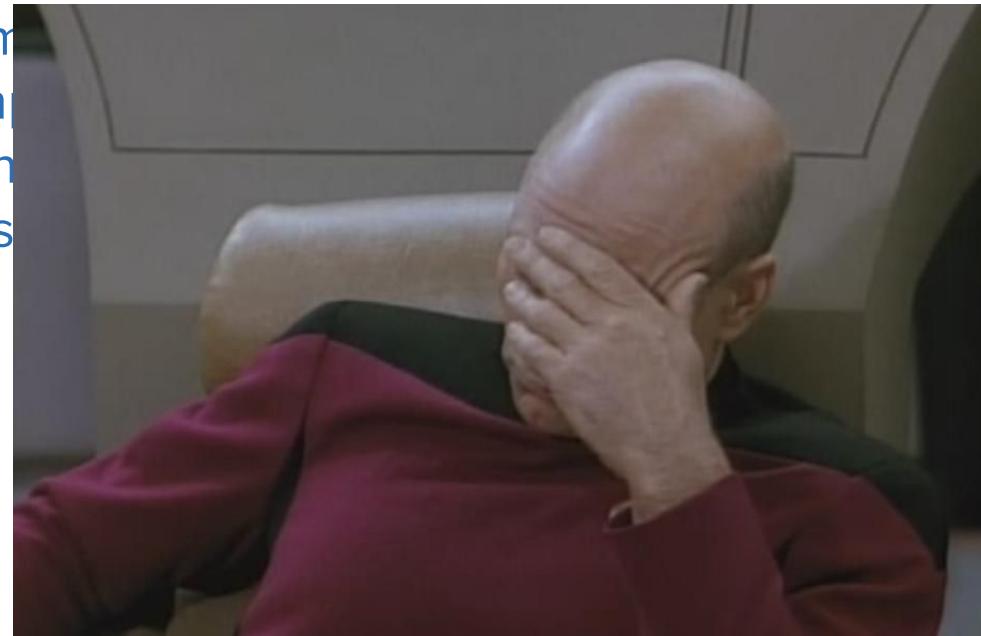
Source: Google Trends

# 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 computer scientist than a statistician”
- “A data scientist is a statistician who knows how to program”
- A data scientist is a good statistician
- A data scientist is a statistician who knows how to program



# Maybe pictures will help?

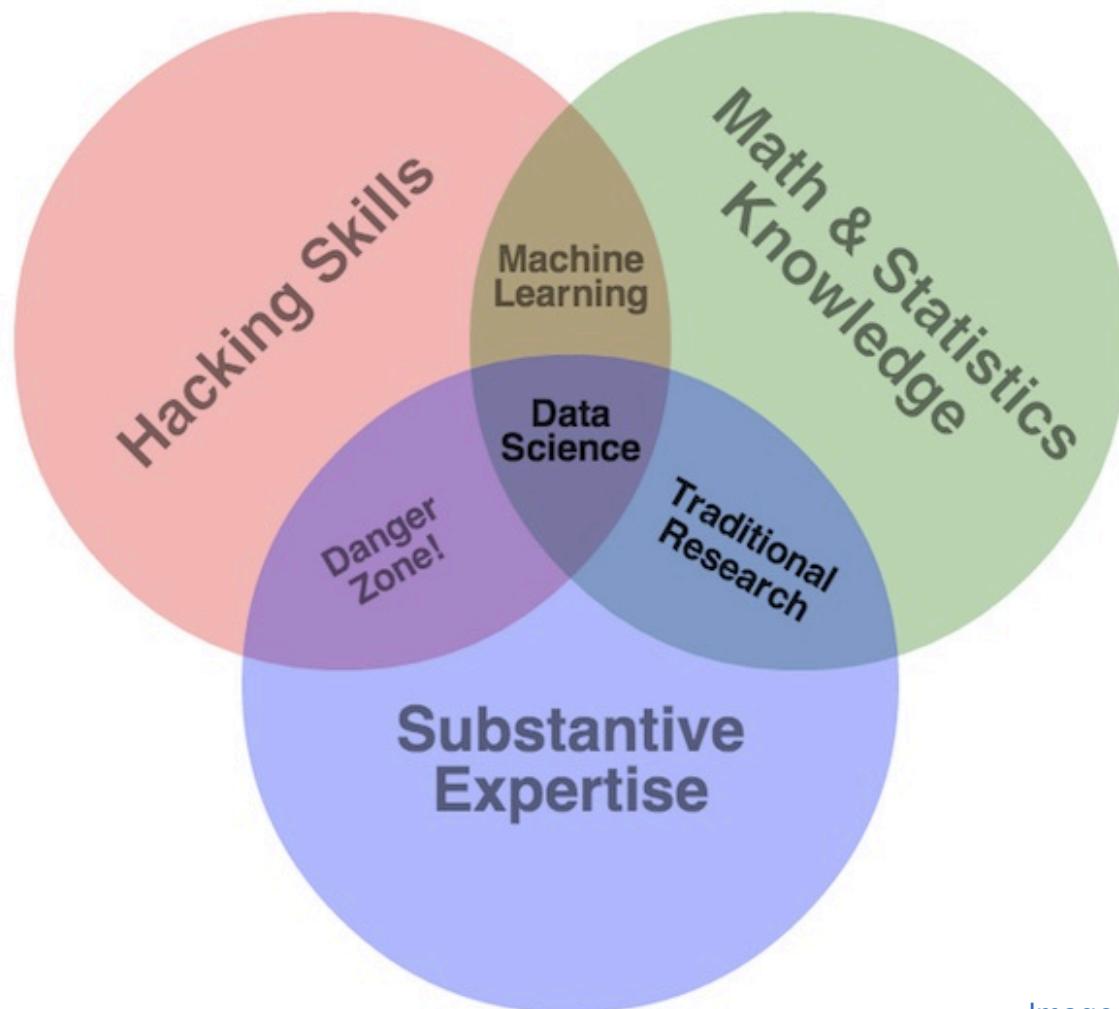
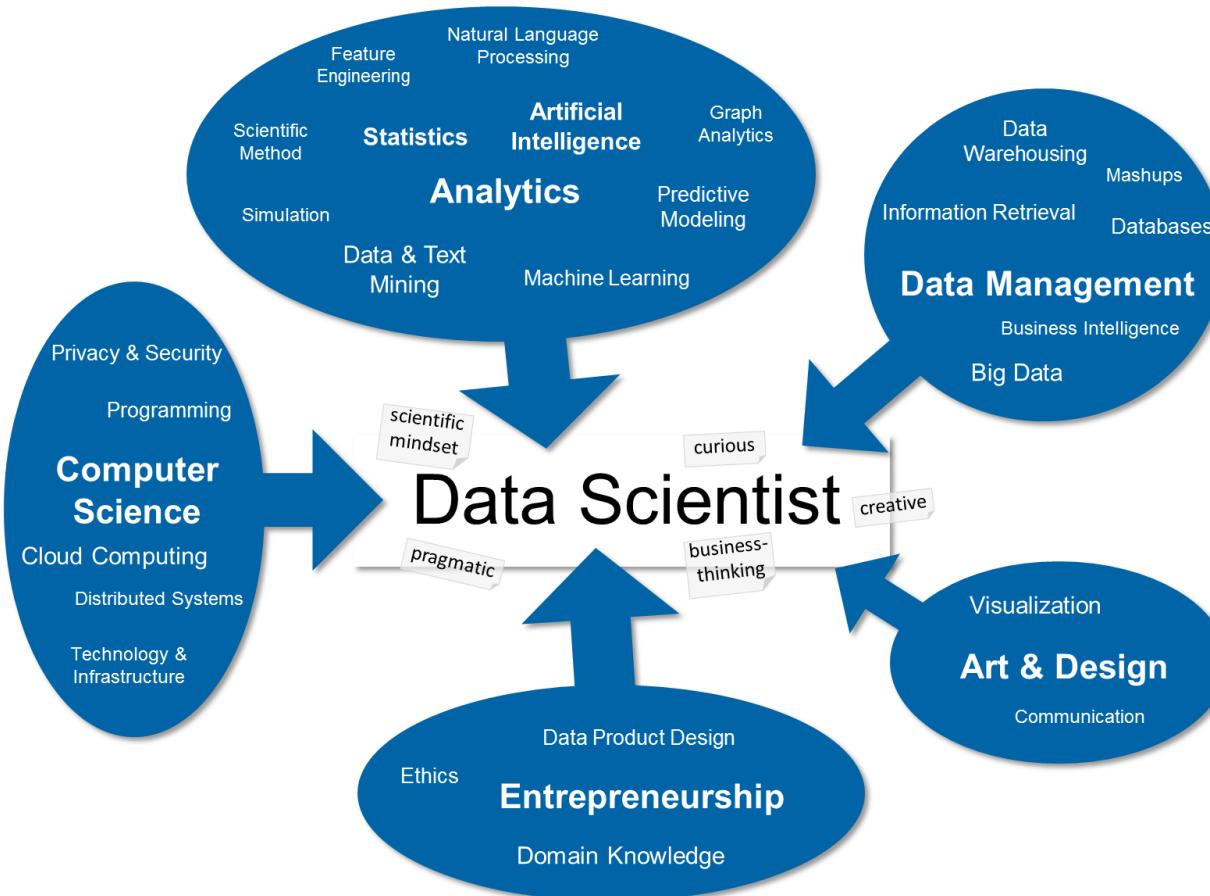


Image from Drew Conway

# Maybe pictures will help?



# Maybe pictures will help?



David Robinson

@drob

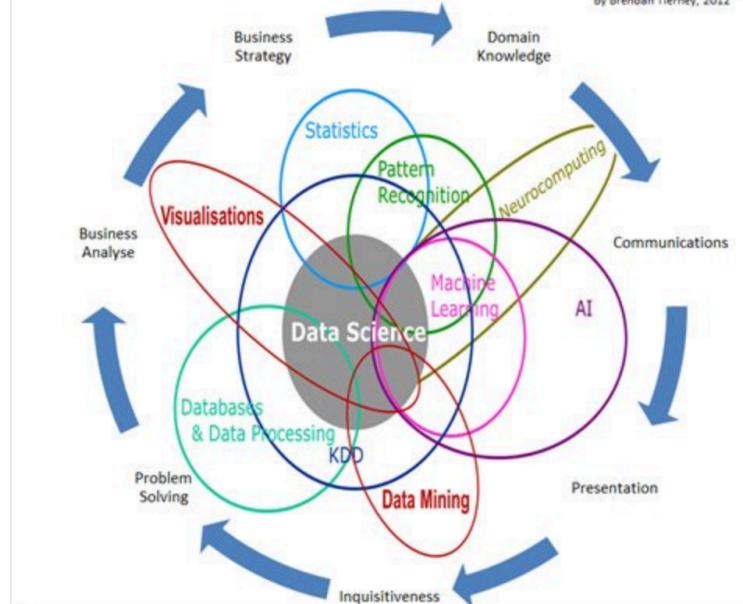
Follow



I'm going to blame [@drewconway](#) for this

## Data Science Is Multidisciplinary

By Brendan Tierney, 2012



RETWEETS    LIKES  
20            84



4:00 PM - 28 Apr 2017 from [Manhattan, NY](#)

From twitter

# Maybe pictures will help?

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

**MATH & STATISTICS**

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

**DOMAIN KNOWLEDGE & SOFT SKILLS**

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative



**PROGRAMMING & DATABASE**

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g. R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

**COMMUNICATION & VISUALIZATION**

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

[MarketingDistillery.com](http://MarketingDistillery.com) is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

**Marketing**  
DISTILLERY  
(c) Krzysztof Zawadzki

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

**MATH & STATISTICS**

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

**DOMAIN KNOWLEDGE & SOFT SKILLS**

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative



**PROGRAMMING & DATABASE**

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

**COMMUNICATION & VISUALIZATION**

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

[MarketingDistillery.com](http://MarketingDistillery.com) is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

**Marketing**  
DISTILLERY

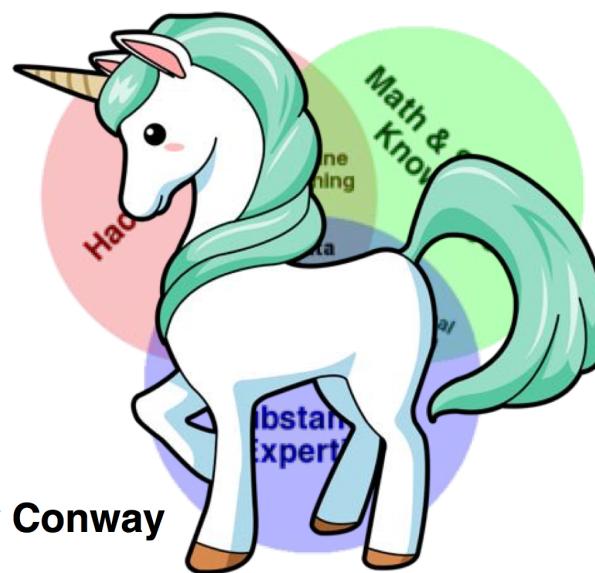
# Why these definitions are bad

- “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

## The Data Scientist Archetype



17

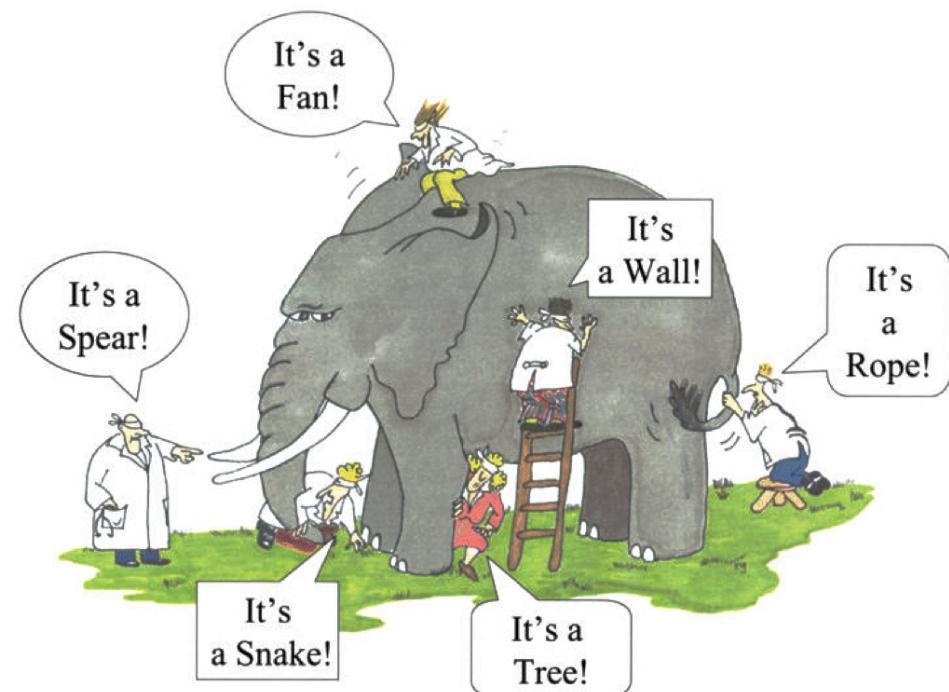
Source: Drew Conway

@angebassa

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

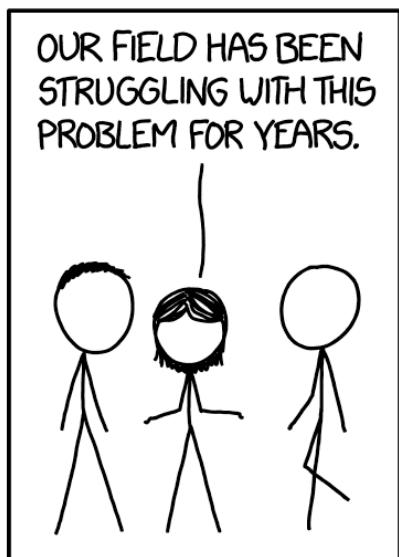
# Why these definitions are good

- 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



# Is “data science” a buzzword?

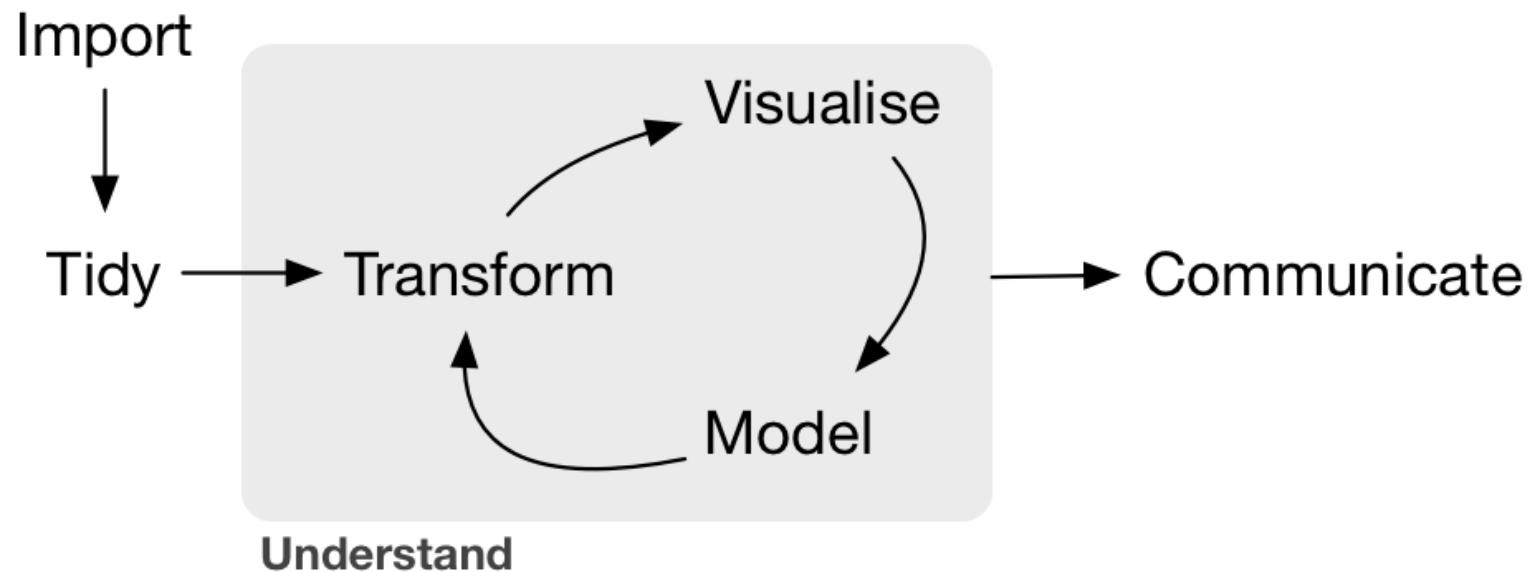
- It is used to describe a frustratingly wide collection of ideas
- It is also used in different (sometimes contradictory) ways by different people
  
- Nonetheless, the popularity of “data science” reflects an increasingly data-centric reality
- Dismissing “data science” ignores this reality, and blinds 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 anything hard because people don't like to show 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



**David Robinson** @d rob · May 19

Most coders won't answer a question without testing it. So if you don't give a reproducible example, you're asking them to make one for you

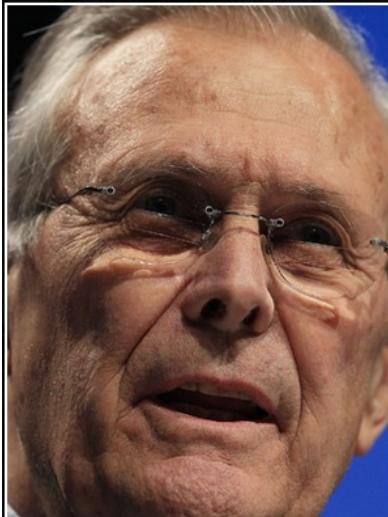
2

10

66

# How to learn data science

- Be on the lookout for cool stuff!



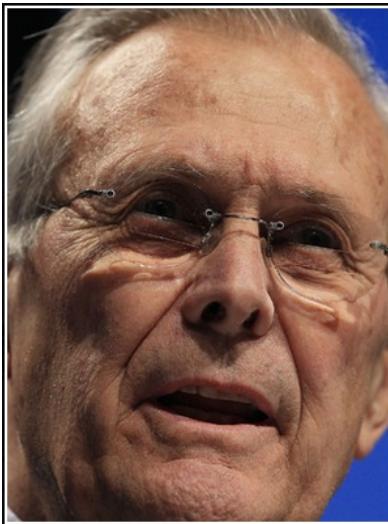
There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

— Donald Rumsfeld —

AZ QUOTES

# How to learn data science

- Be on the lookout for cool stuff!



There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

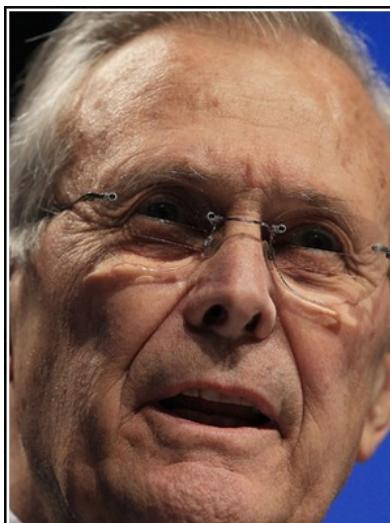
— Donald Rumsfeld —

AZ QUOTES

Knowledge base! :-D

# How to learn data science

- Be on the lookout for cool stuff!



There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

— Donald Rumsfeld —

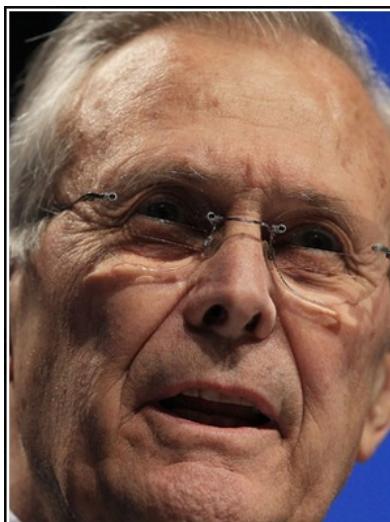
AZ QUOTES

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!



There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

— Donald Rumsfeld —

AZ QUOTES

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*



David Parkins

# Data as a resource

The world's most valuable resource  
is no longer oil.  
Illustration by [David Berman](#)

Sections ≡

The Washington Post

 BrandStudio  Content from IBM Power Systems

*The data economy does not care about oil.*



Why big data is  
"the new natural  
resource"



# Data as a resource

The world's most valuable resource  
is no longer oil.

Sections ≡

The Washington Post

wp BrandStudio



Content from IBM Power Systems

*The data economy d*



## 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™

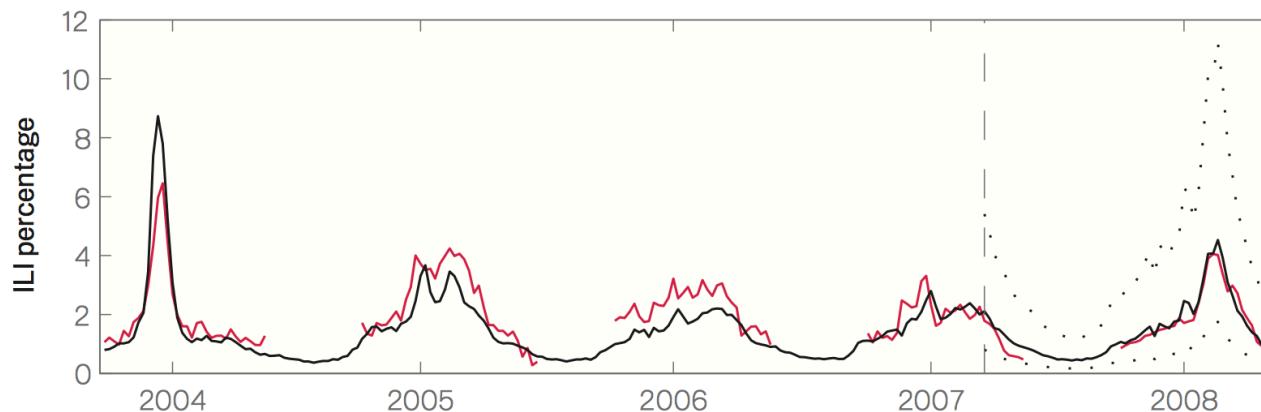
Detecting influenza epidemics using  
search engine query data

# Google flu trends



Google™

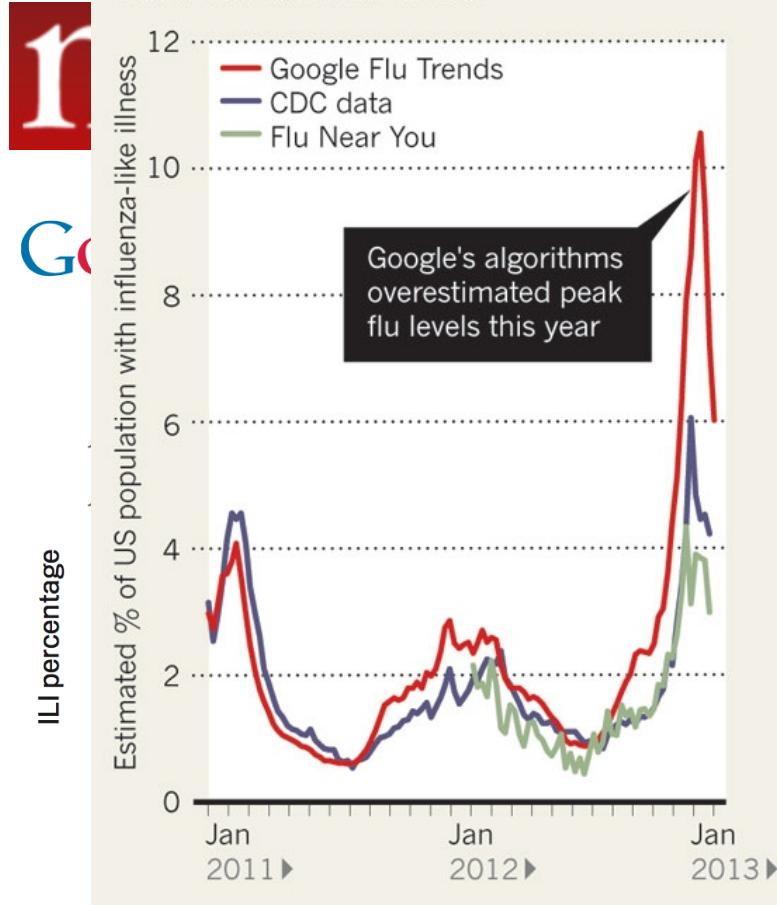
Detecting influenza epidemics using  
search engine query data



# Google flu trends

## FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



fluenza epidemics using  
in nature International weekly journal of science

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

Archive > Volume 494 > Issue 7436 > News > Article

NATURE | NEWS

عربي

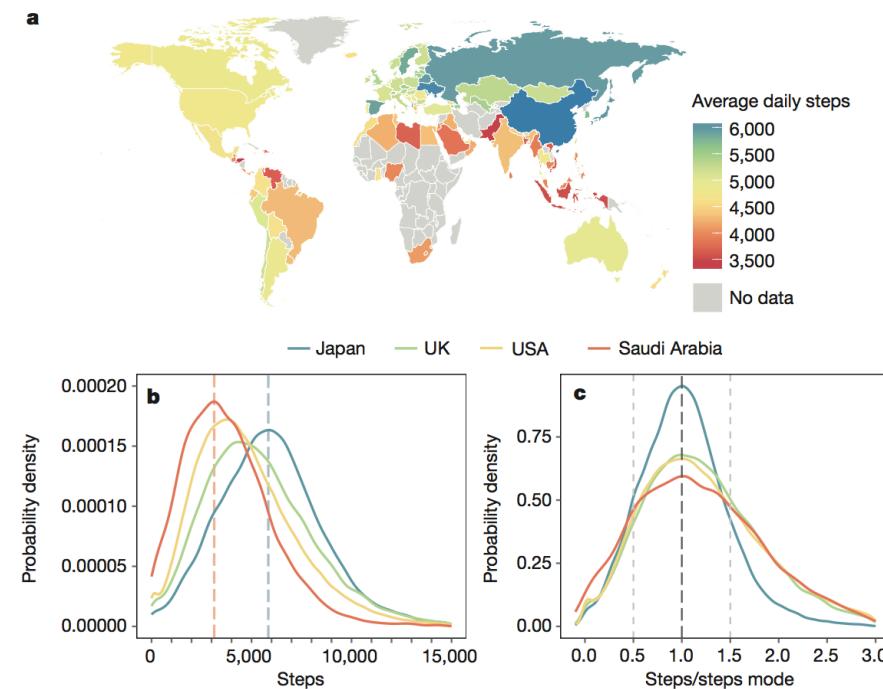
## When Google got flu wrong

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

# Activity via smartphones



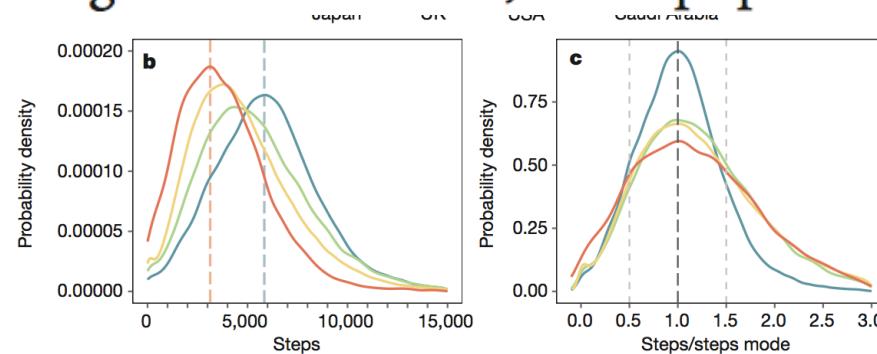
Large-scale physical activity data reveal worldwide activity inequality



# Activity via smartphones

**nature** International weekly journal of science

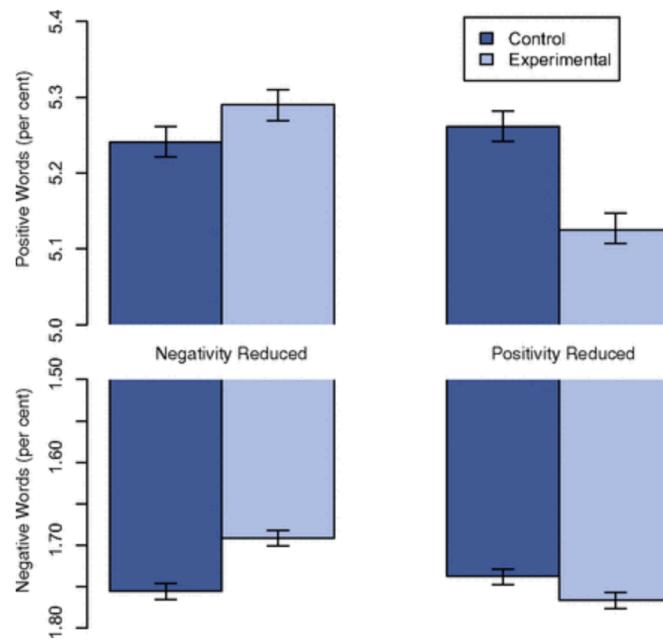
We study 68 million days of minute-by-minute step recordings from 717,527 anonymized users of the Argus smartphone application developed by Azumio. The dataset includes recordings of physical activity for free-living individuals from 111 countries (Fig. 1a). We focus on the 46 countries with at least 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



Experimental evidence of massive-scale emotional contagion through social networks

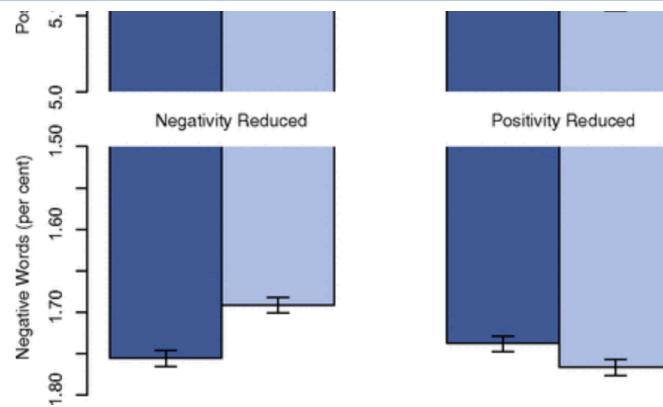


# Facebook experiments

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.

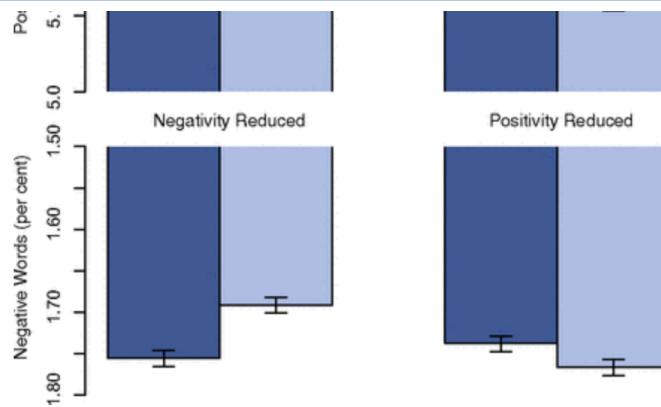


# BBC | Facebook experiments

## NEWS

### 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 | Facebook experiments

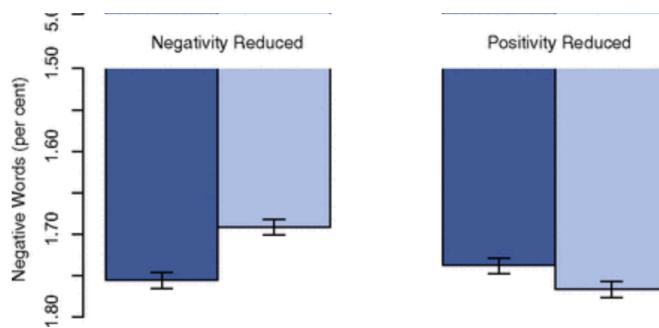
NEWS

## Facebook admits failings over emotion manipulation study

the guardian

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 | oook experiments

NEWS

**Facebook admits failings over emotion manipulation study**

theguardian

cebook, that emotional states can be pple 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 | Facebook experiments

## NEWS

Fa  
ma

th

"Fa  
to

Fo

Faceb  
Science

ut



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

For

# Limitations of data

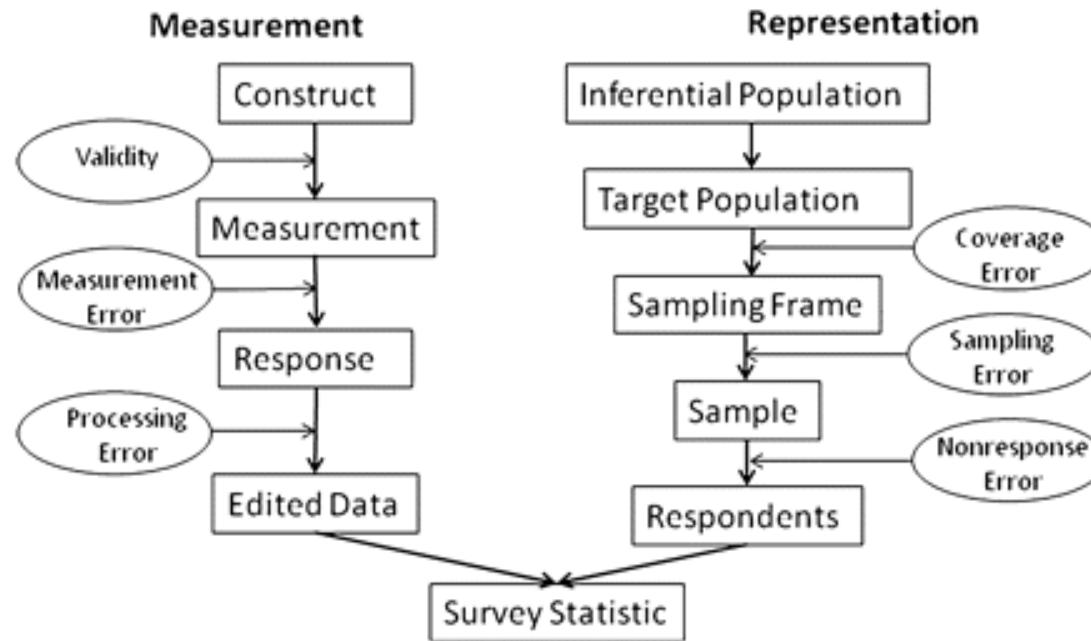
- Not trying to gang up on Google, Facebook and smartphone
  - In each case, these are smart people doing interesting thing with cool data
- These cases point to challenges to be overcome, and are important opportunities
  - How can public health practitioners engage with non-traditional partners in 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?

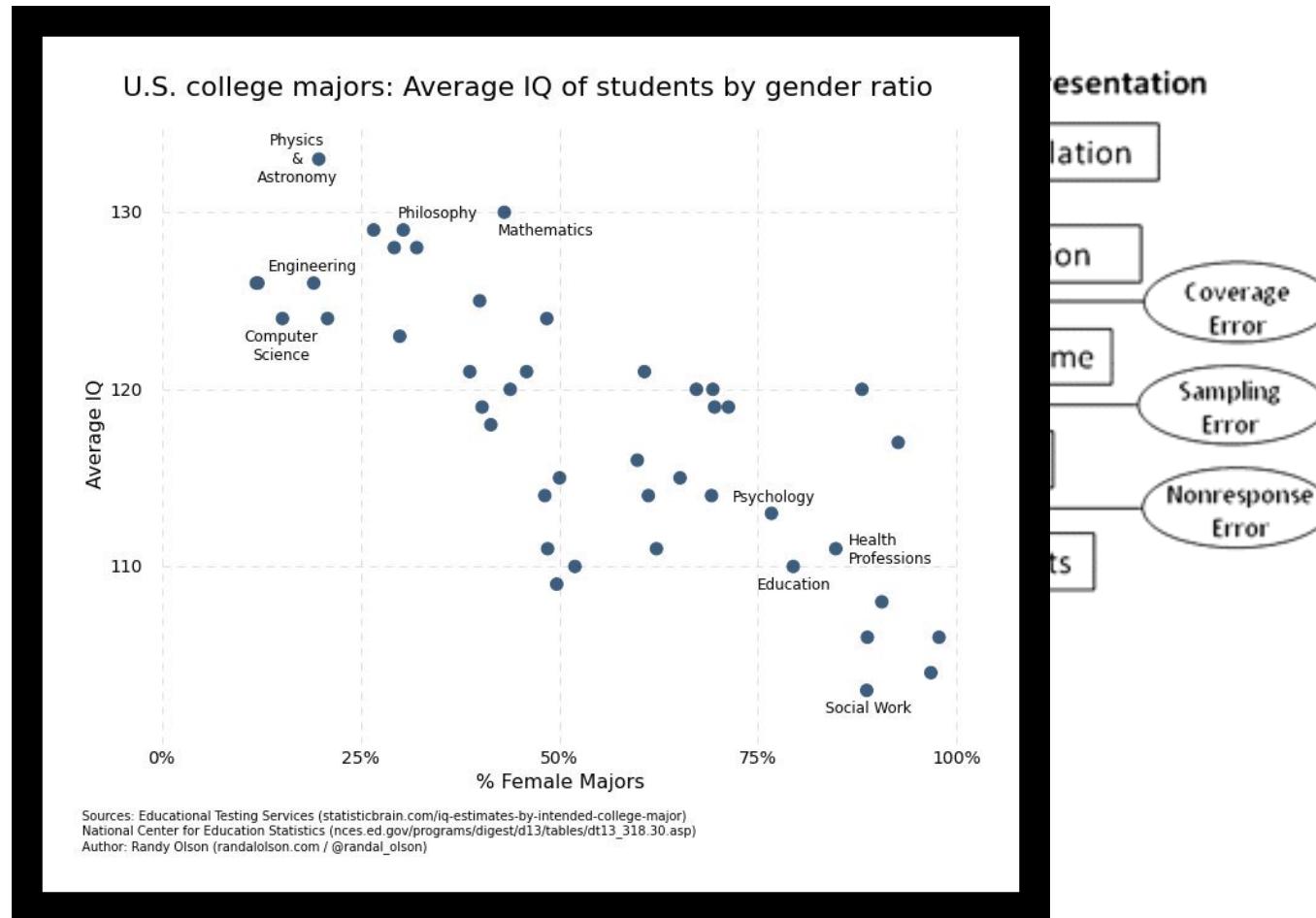
- Who is represented?
- What do measurements mean? Can they be trusted?
- Will I uncover associations? Causes?
- Can I develop new hypotheses or confirm existing ones?
- Can I design an intervention, or evaluate the success of one?

# 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.

Sources: Ed  
National Ce  
Author: Ran

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.