

“WHAT IS DATA SCIENCE?” RE-REVISITED

Jeff Goldsmith, PhD

Department of Biostatistics

Maybe pictures will help?

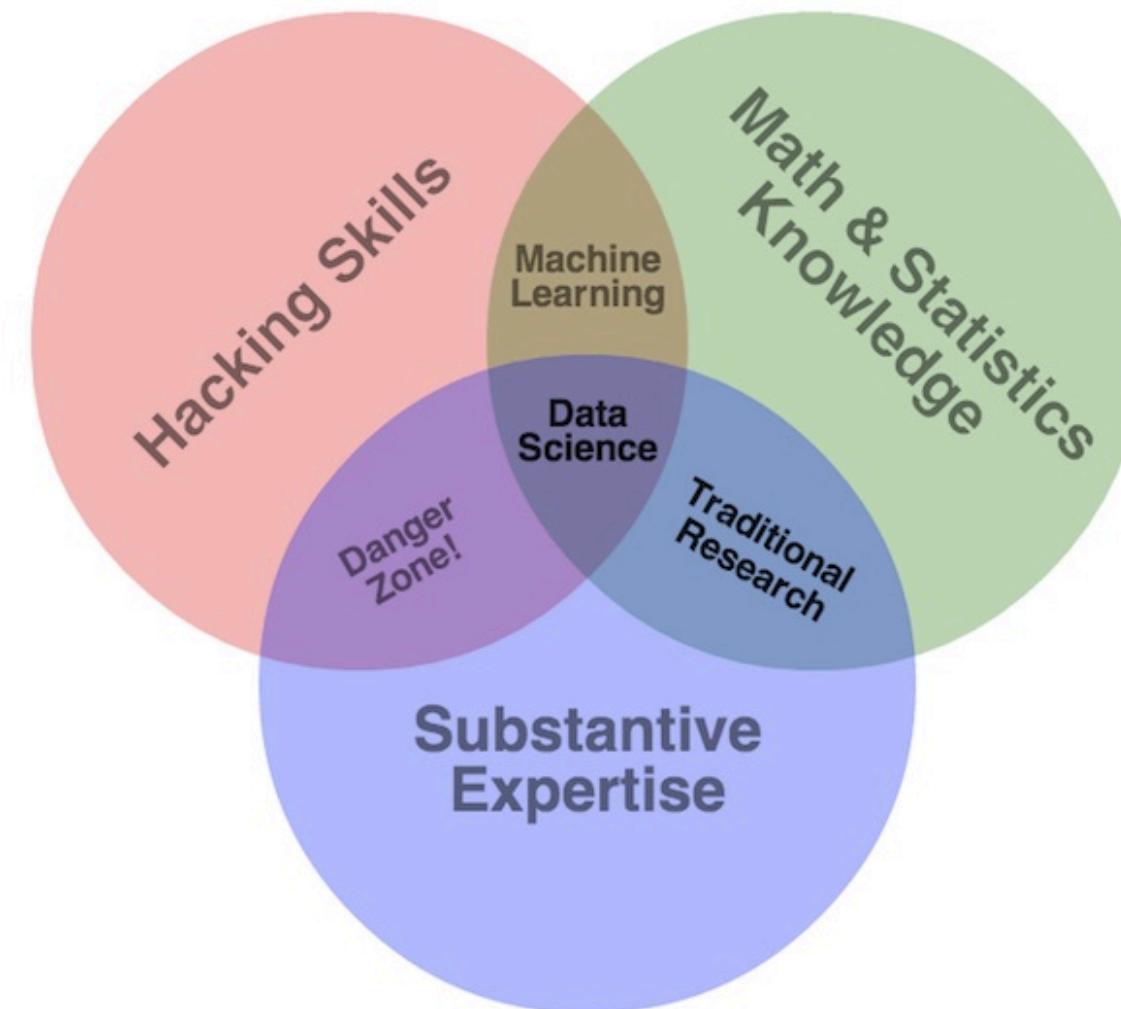


Image from Drew Conway

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

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



DOMAIN KNOWLEDGE & SOFT SKILLS

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

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 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

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



DOMAIN KNOWLEDGE & SOFT SKILLS

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

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 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

Recurring themes

- You need “data skills”
 - Data wrangling
 - Reproducibility
 - Communication
 - Analytics and modeling
- You also need a mindset
 - Intellectual curiosity
 - Ability to solve problems
 - Interest in domain, even empathy with collaborators

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

Problem solving

“I’ve interviewed a lot of people over the years.... Recently, when people have an interview, I ask a single question that I think tries to get at the point of problem solving. The question I ask is along the lines of ‘[Imagine you had access to a database of 100 million mobile devices.] What questions would you ask? What types of things do you think you could learn, and how would you go about doing it?’”

From “How Industry Views Data Science Education in Statistics Departments”, Chris Volinsky’s JSM 2015 talk

Practice problem solving

- You can (and should) practice having a mindset, or a style of thinking
 - Make a habit of asking yourself what you would like to do with a data resource
 - Think about how you would accomplish it
- Be on the lookout for cool projects, and learn from them
 - Pay attention to the thought process, not just the specific tools
- Many projects need overlapping skill sets
 - You don't have to be a domain expert yourself, but you may need to work with one
 - You'll also have to communicate effectively with that person, which means at least taking an interest

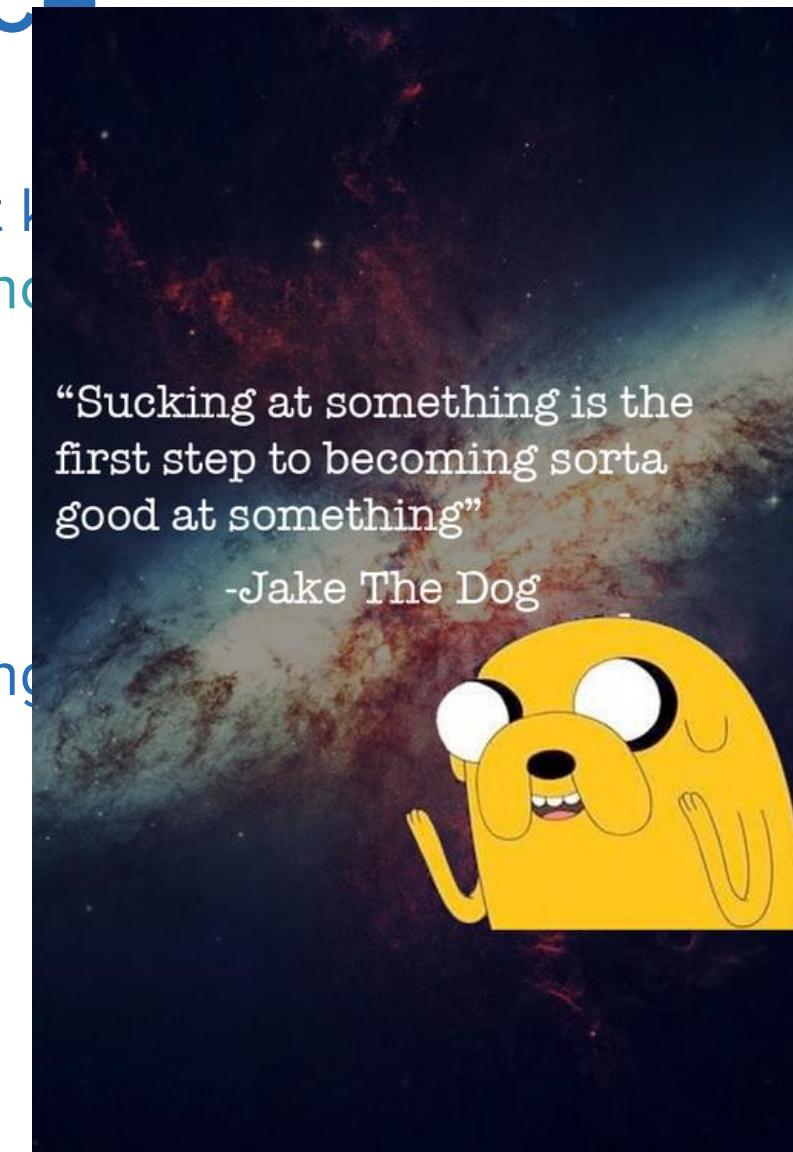
How to learn data science

- 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 don't know
 - Corollary: don't be a jerk to people who do know things
- Ask questions (well) and keep learning
- Pretty much the same as learning anything else, except that people like to show their code



How to learn data science

- Be on the lookout for cool stuff!



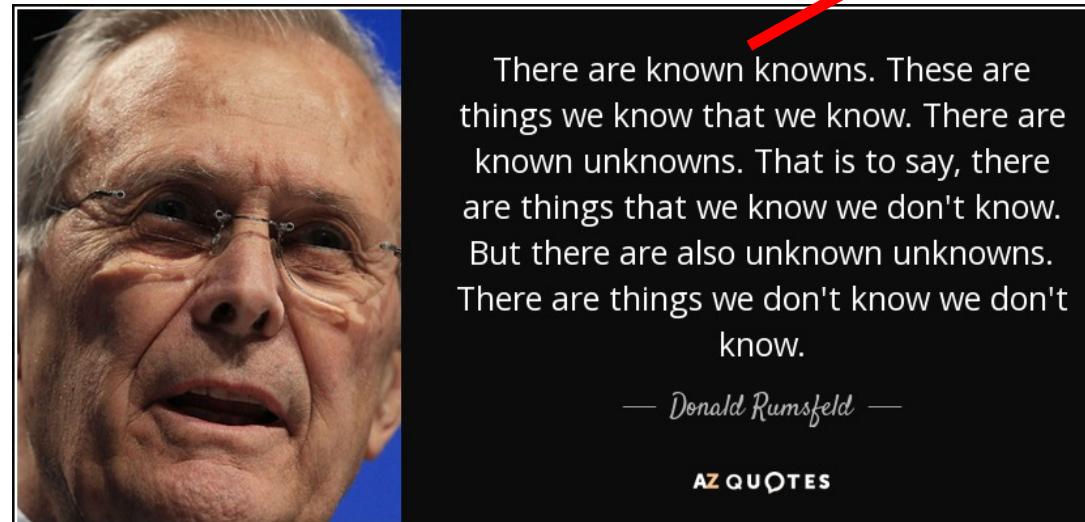
There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we know we don't know. But there are also unknown unknowns. There are things we don't know we don't know.

— Donald Rumsfeld —

AZ QUOTES

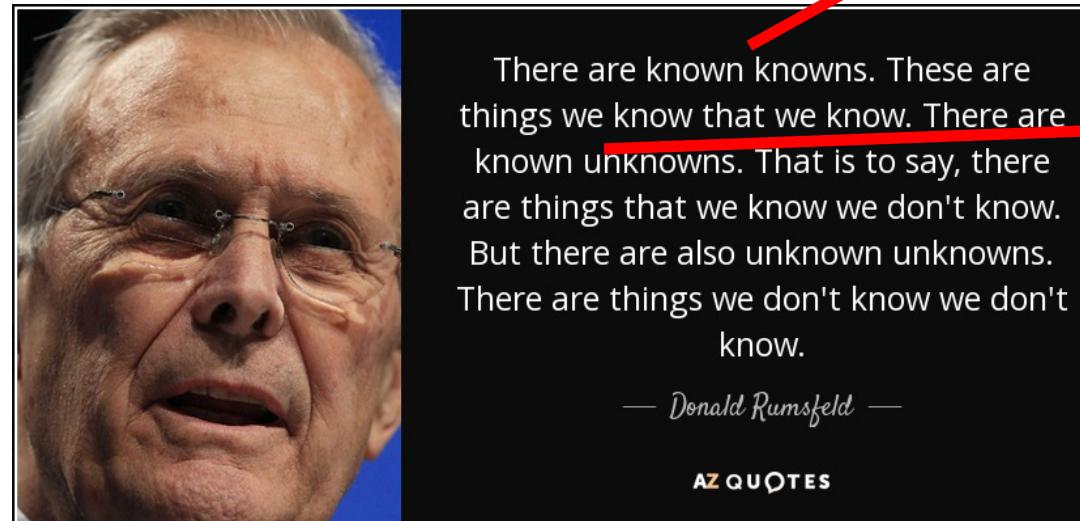
How to learn data science

- Be on the lookout for cool stuff!



How to learn data science

- Be on the lookout for cool stuff!

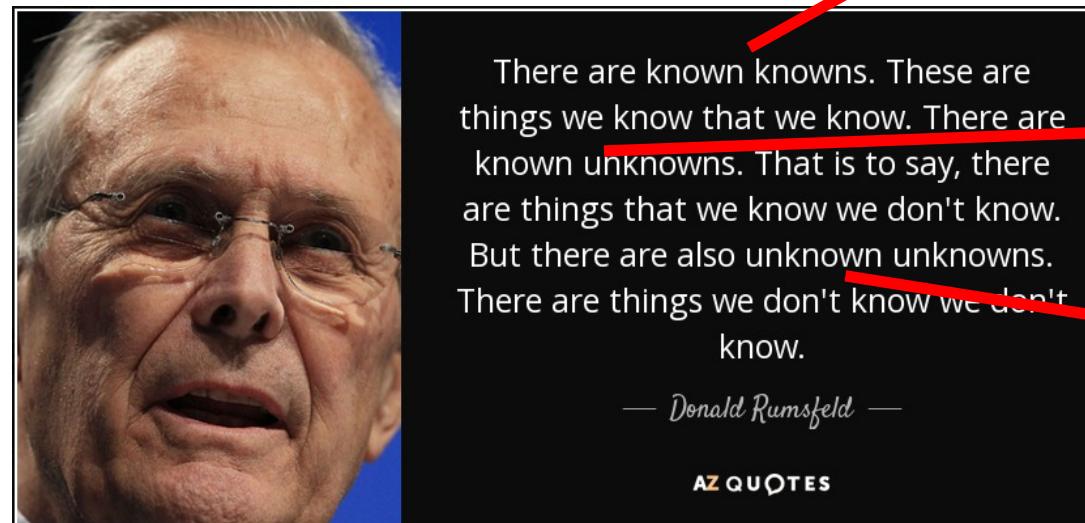


Knowledge base! :-D

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

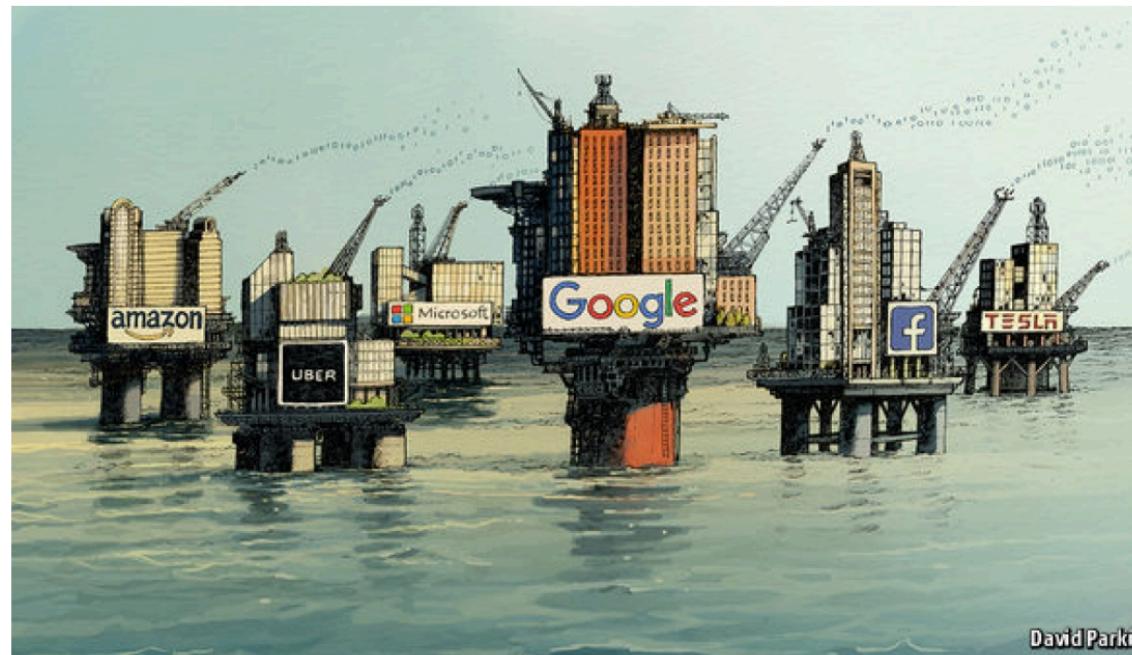
Things you know
exist and can
learn how to
do :-)

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

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 longer oil, but data

Sections ≡

The Washington Post

wp BrandStudio i Content from IBM Power Systems

The data economy does not



Why big data is
"the new natural
resource"



Data as a resource

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

[Sections](#)  BrandStudio  Content from **IBM Power Systems**

The data economy does not have to be a zero-sum game.

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?

Google flu trends



Google™

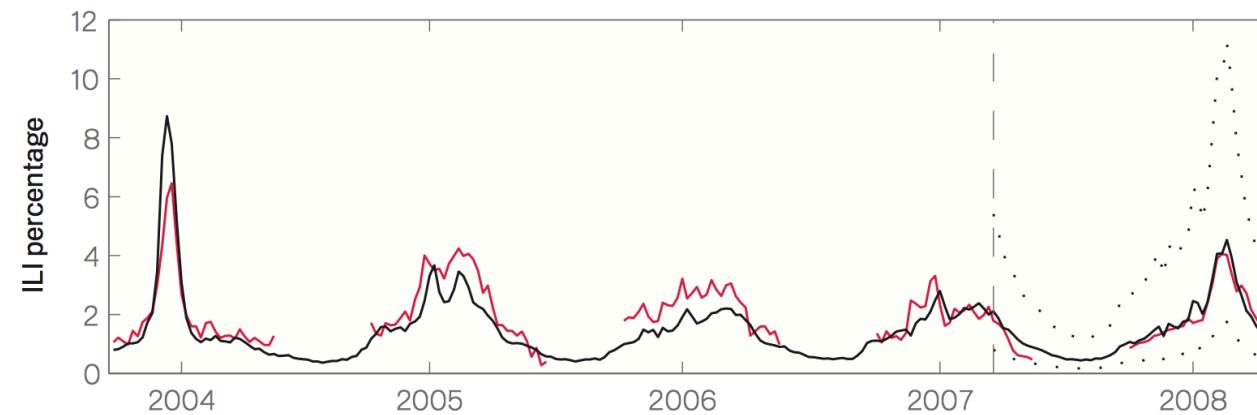
Detecting influenza epidemics using
search engine query data

Google flu trends

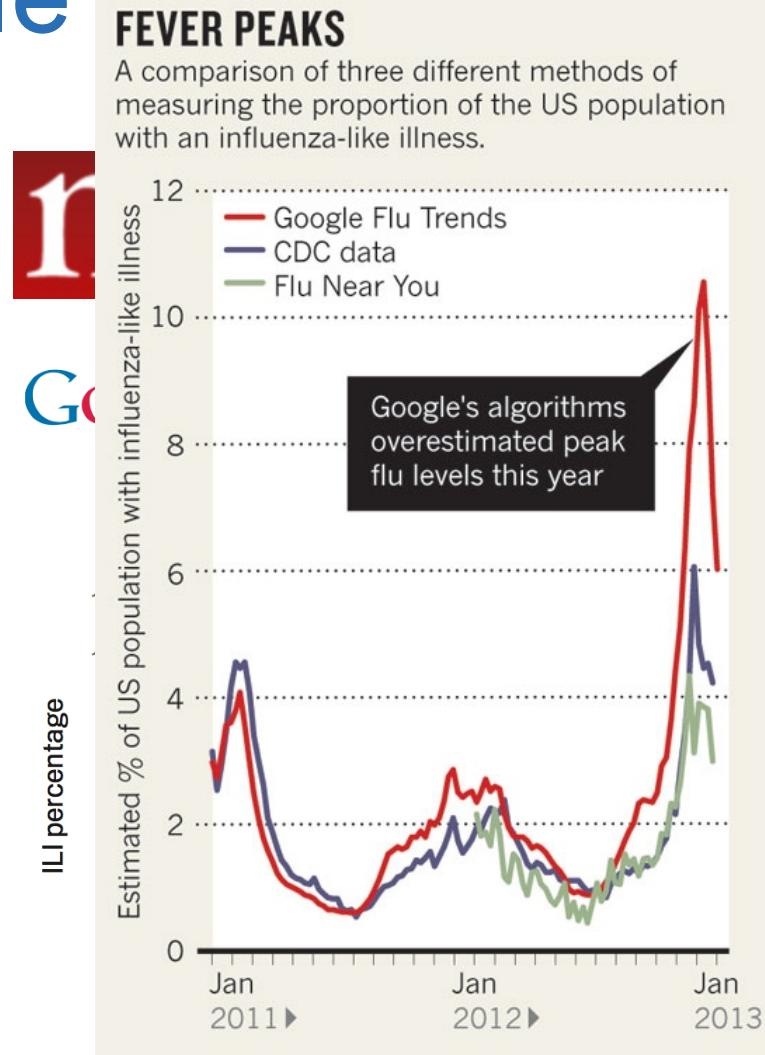


Google™

Detecting influenza epidemics using
search engine query data



Google flu trends



weekly journal of science

flu trends epidemics viruses

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

عربی

When Google got flu wrong

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

AI for Translation

FEATURE

The Great A.I. Awakening

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

AI for Translation

FEATURE

The Great Translation Experiment Even artificial intelligence can acquire biases against race and gender

How Google used artificial intelligence to transform Google Translate, one of its more popular services — and how machine learning is poised to reinvent computing itself.

AI for Translation

The Great

How Google used
Translate, one of its
learning is

FEATURE

Even artificial intelligence can acquire biases against race and gender

Google Translate’s gender bias pairs “he” with “hardworking” and “she” with lazy, and other examples

AI for Translation

The Grid

How Google Translate, one of the world's most popular learning tools, can reinforce gender bias

FEATURE

Even artificial intelligence can acquire biases against

he is a soldier
she's a teacher
he is a doctor
she is a nurse

With lazy, and other
examples

slate's gender
“he” with
“ng” and “she”
and other

Limitations of big data

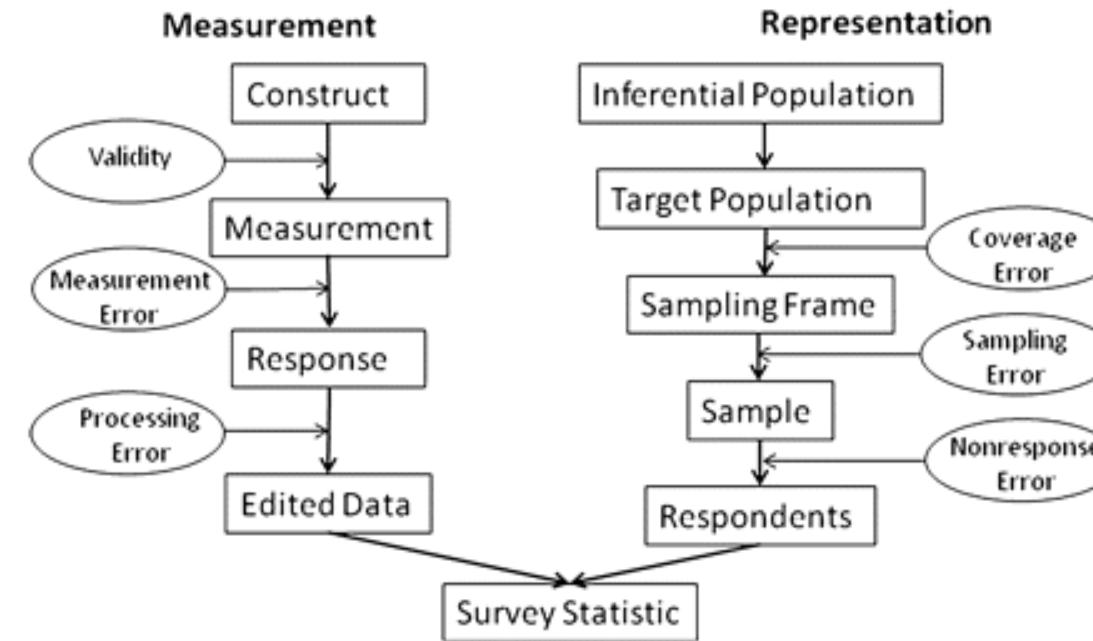
- Not trying to gang up on Google, Facebook and smartphone
 - In each case, these are smart people doing interesting thing with cool data
- These cases point to challenges to be overcome, and are important opportunities
 - How can public health practitioners engage with non-traditional partners in a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

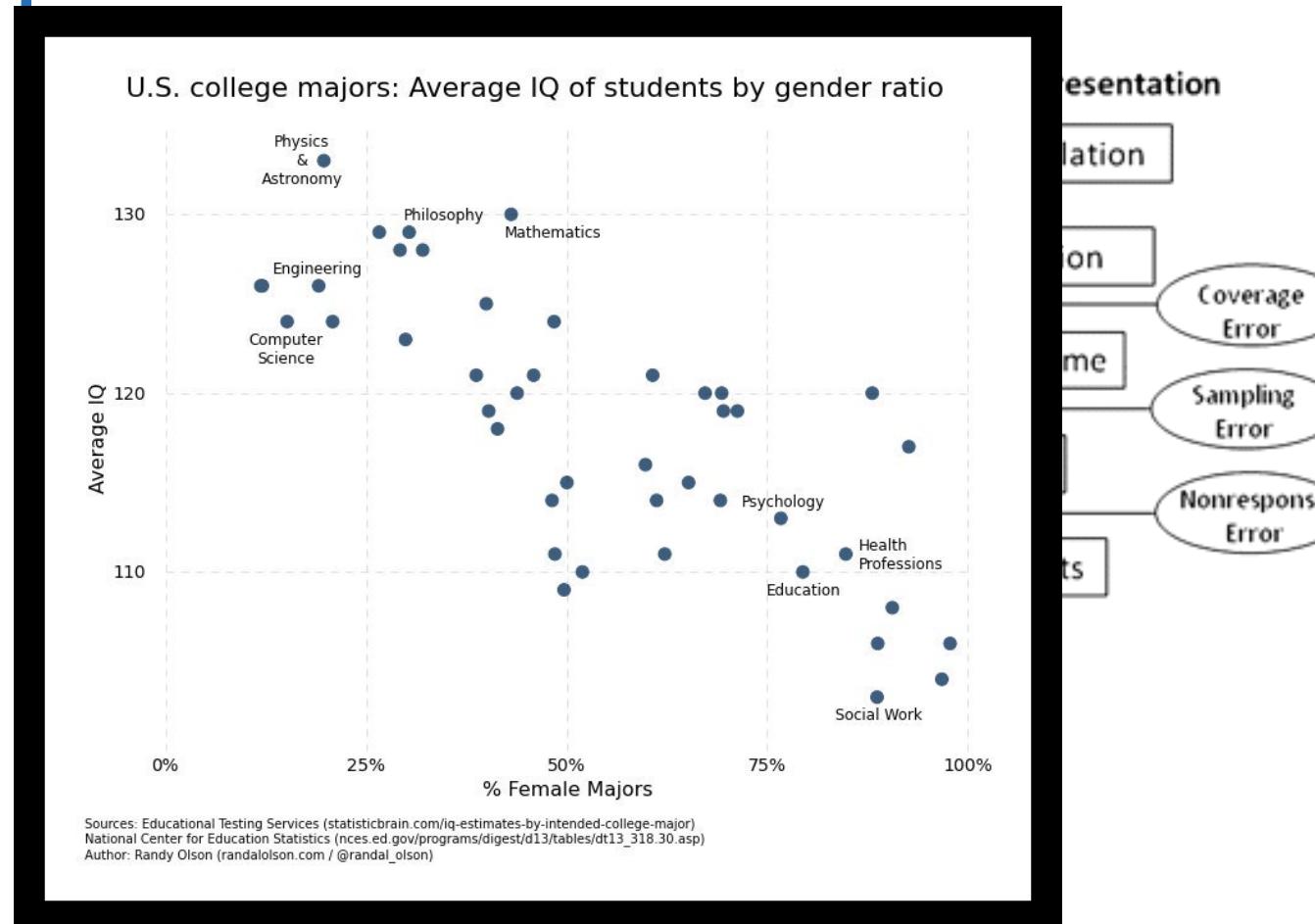
- Who is represented?
- What do measurements mean? Can they be trusted?
- Will I uncover associations? Causes?
- Can I develop new hypotheses or confirm existing ones?
- Can I design an intervention, or evaluate the success of one?

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



representation

selection

time

me

nts

Coverage
Error

Sampling
Error

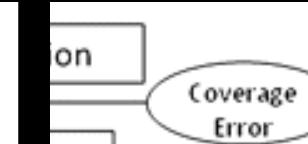
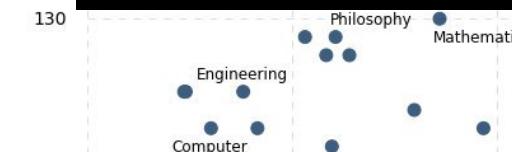
Nonresponse
Error

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

Be skeptical

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: Economic
National Center for Education Statistics
Author: Randall Kroszner

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

A caveat before you leave ...

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.

Final thoughts