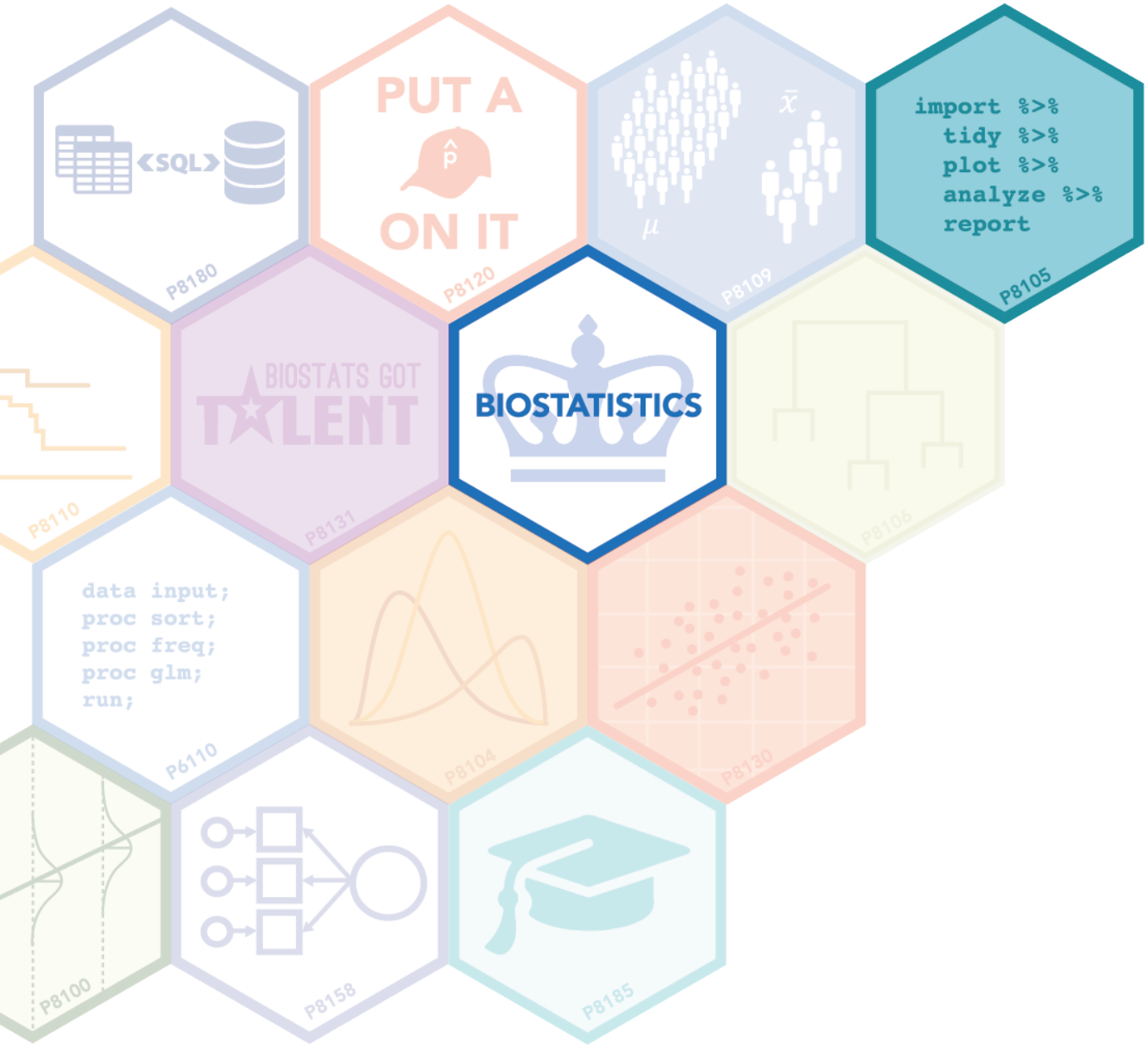


"WHAT IS DATA SCIENCE?" RE-REVISITED

Jeff Goldsmith, PhD

Department of Biostatistics

Maybe pictures will help?

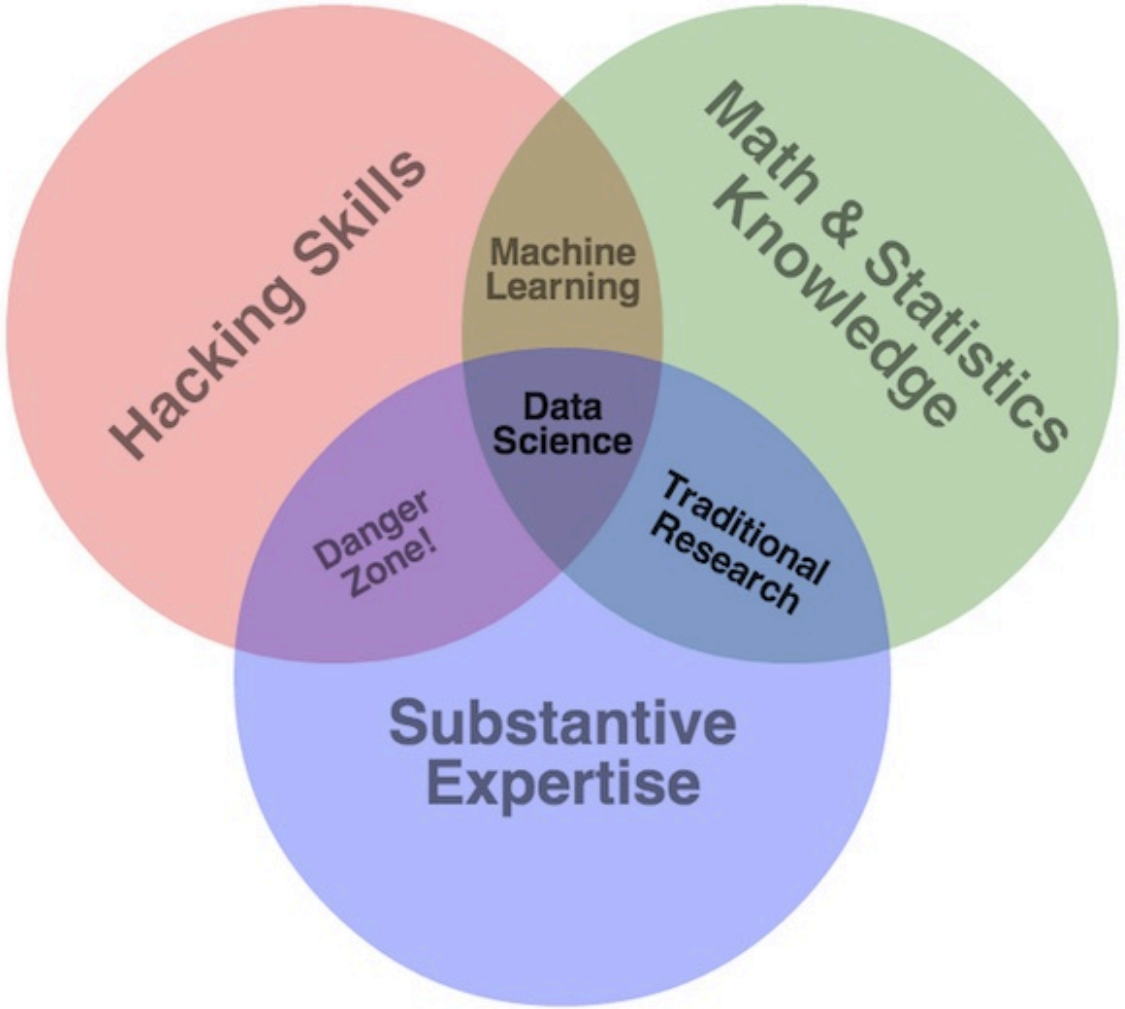
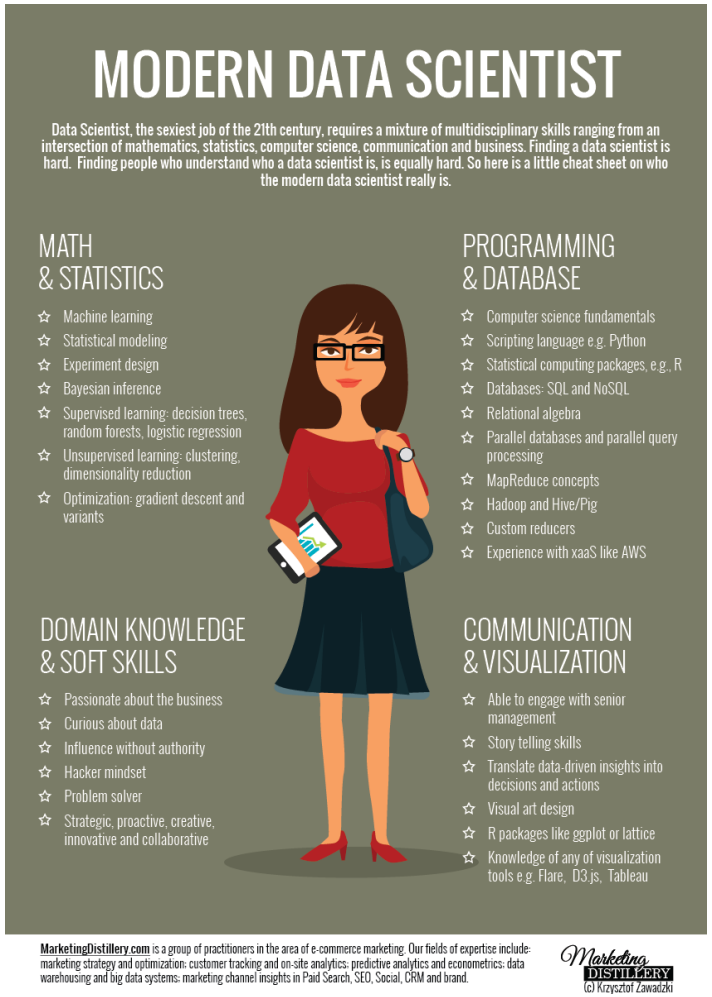
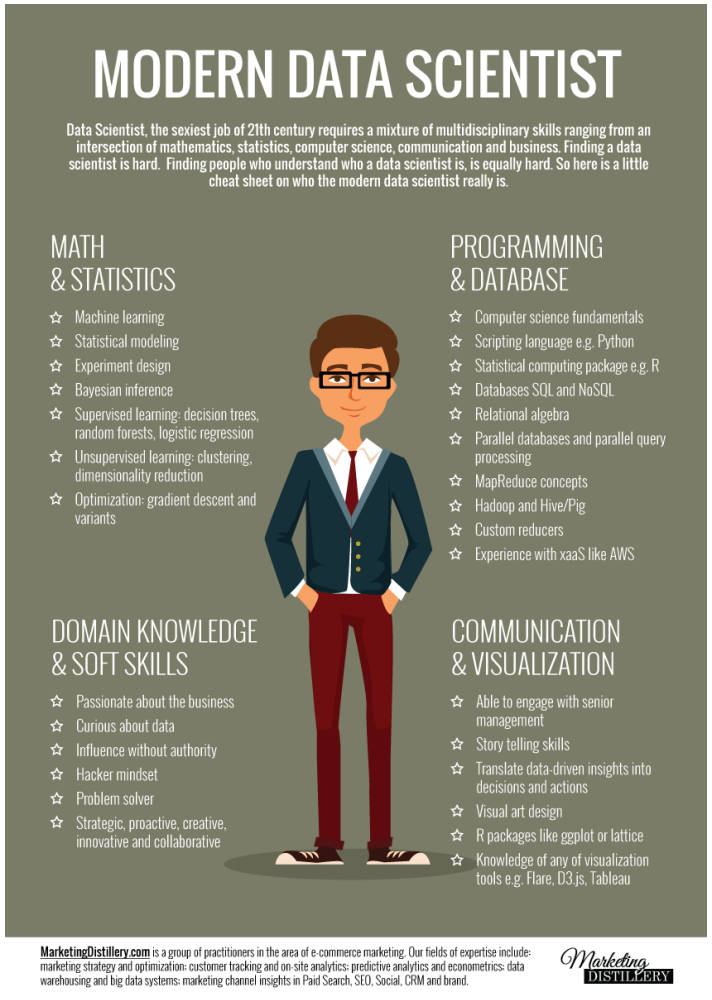


Image from Drew Conway

Maybe pictures will help?





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

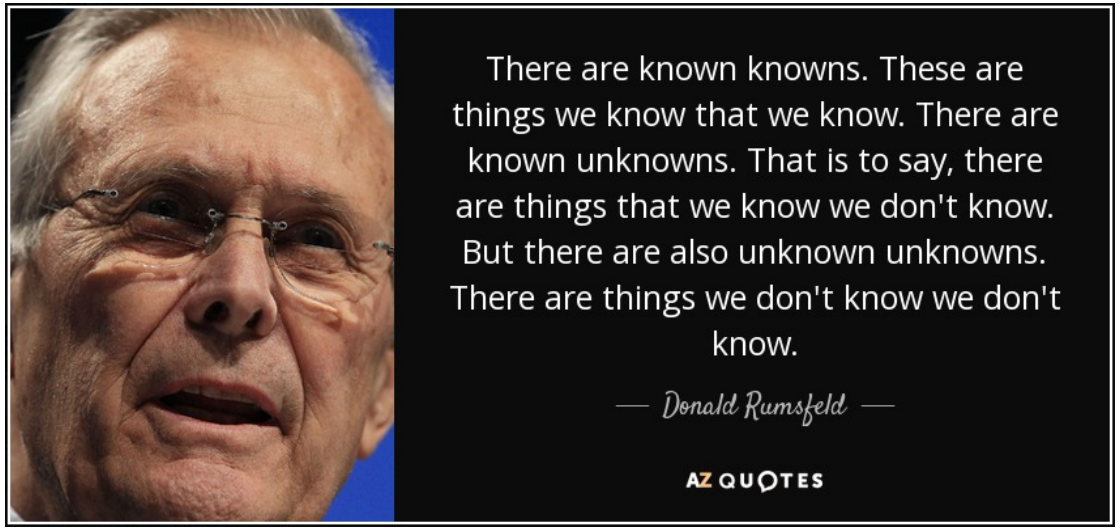
How to learn data science

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

- Pretty much the same as learning anything
  don't like to show their code

How to learn data science

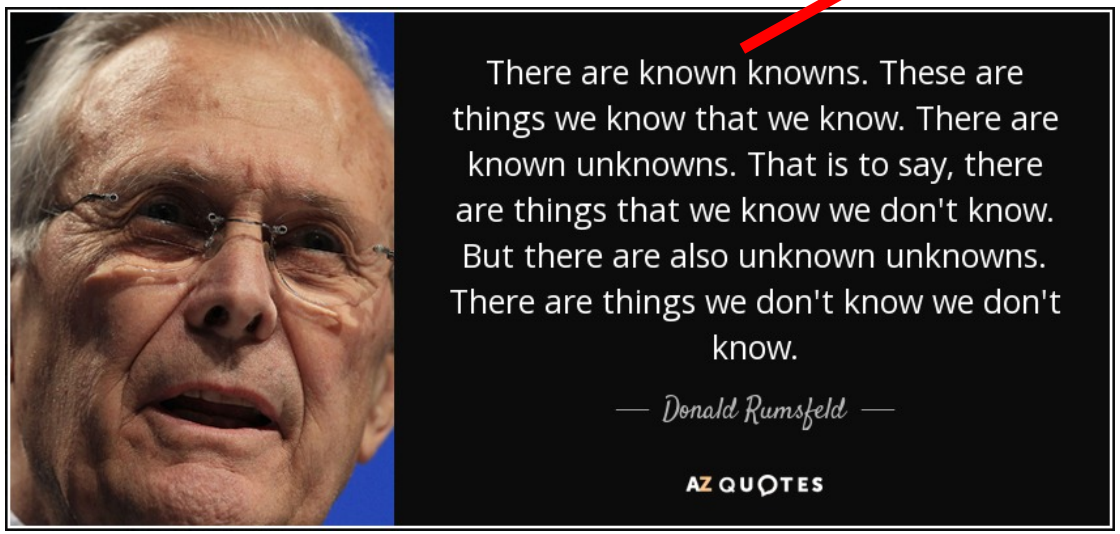
- Be on the lookout for cool stuff!



How to learn data science

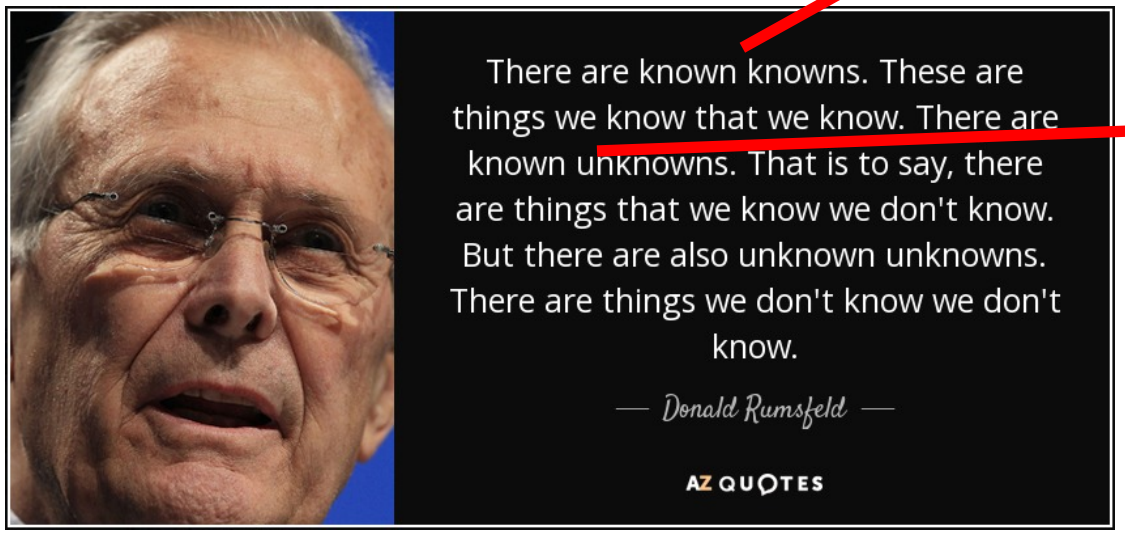
- Be on the lookout for cool stuff!

Knowledge base! :-D



How to learn data science

- Be on the lookout for cool stuff!

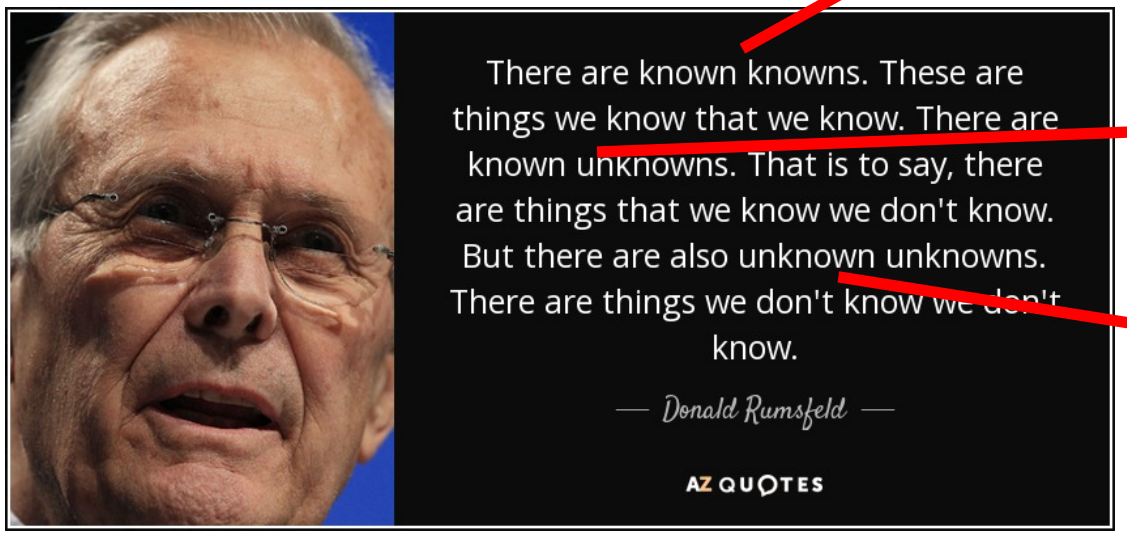


Knowledge base! :-D

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 :-(

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

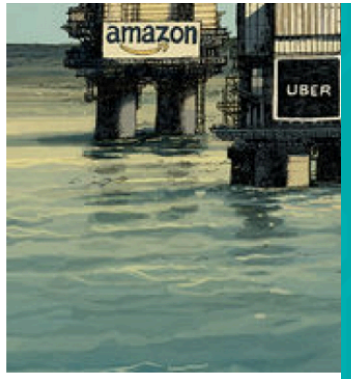
The world's most valuable resource
is no longe

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?

Google flu trends

Google

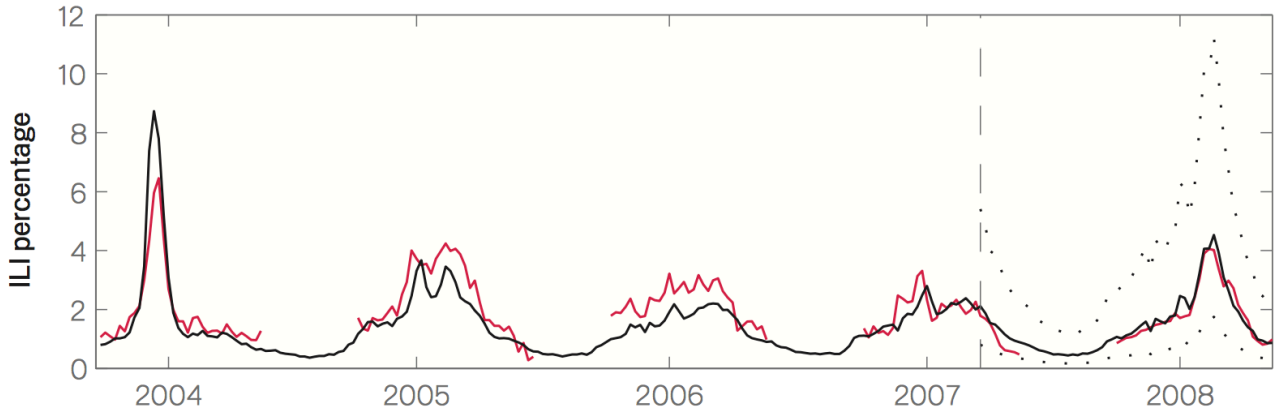
Detecting influenza epidemics using search engine query data

Google flu trends

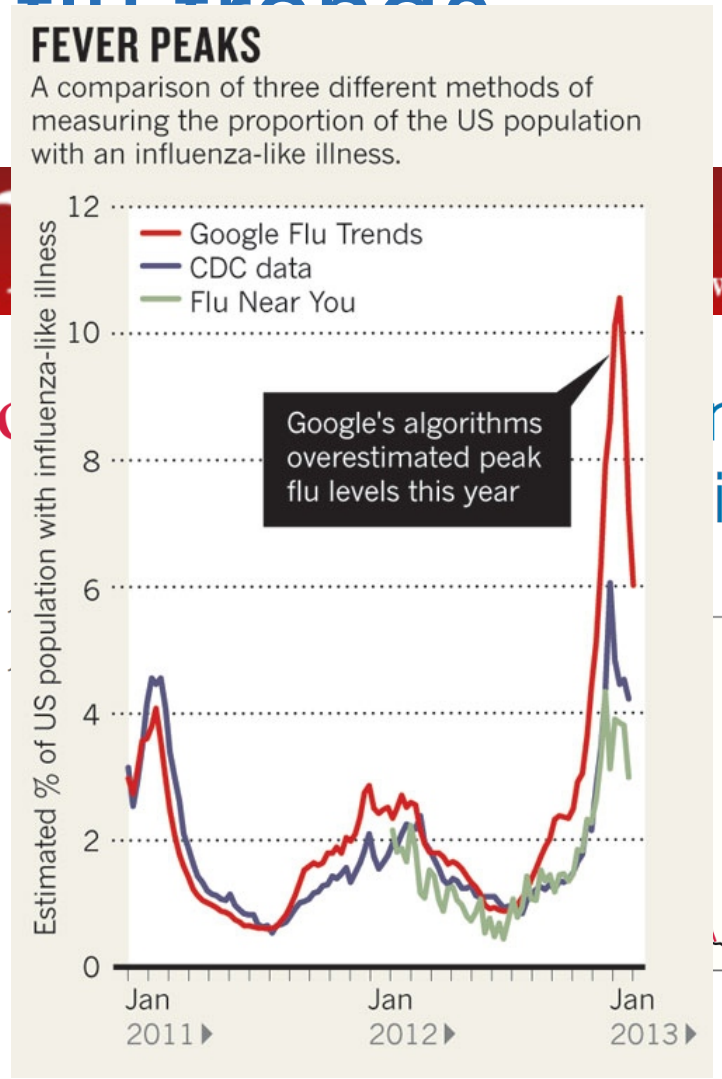
nature International weekly journal of science

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

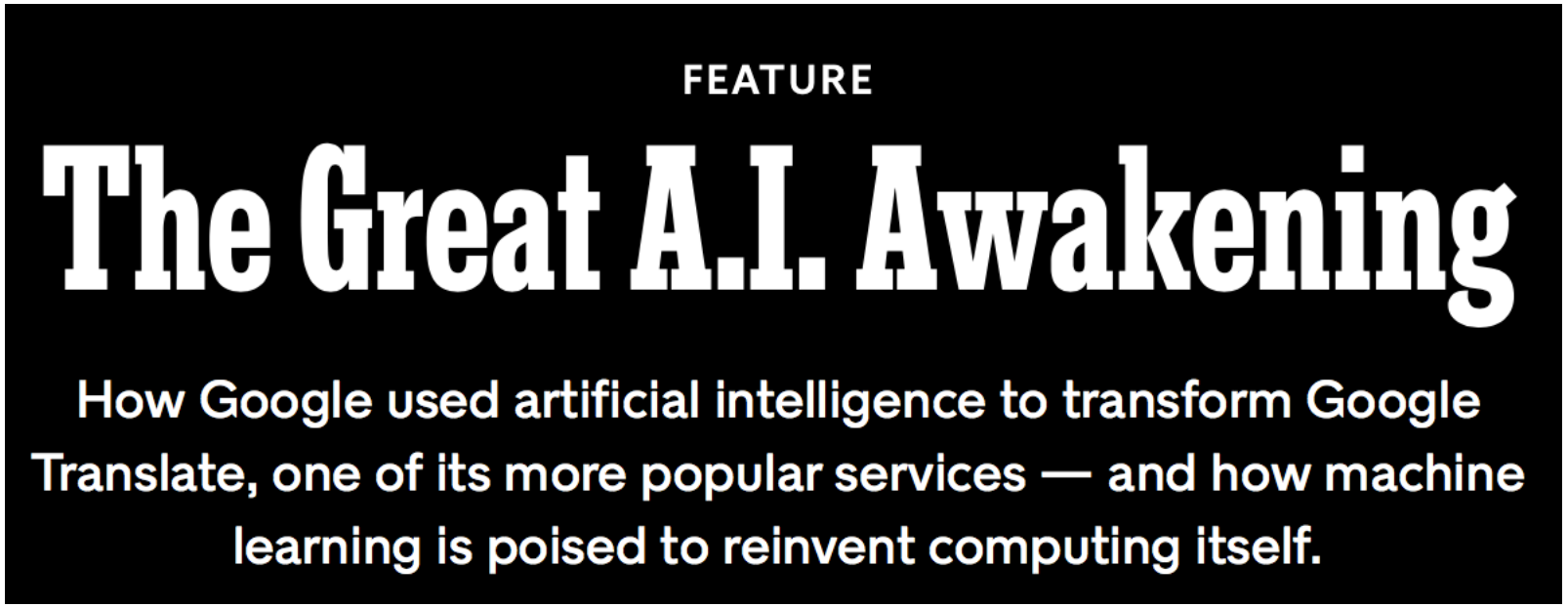
عربي

When Google got flu wrong

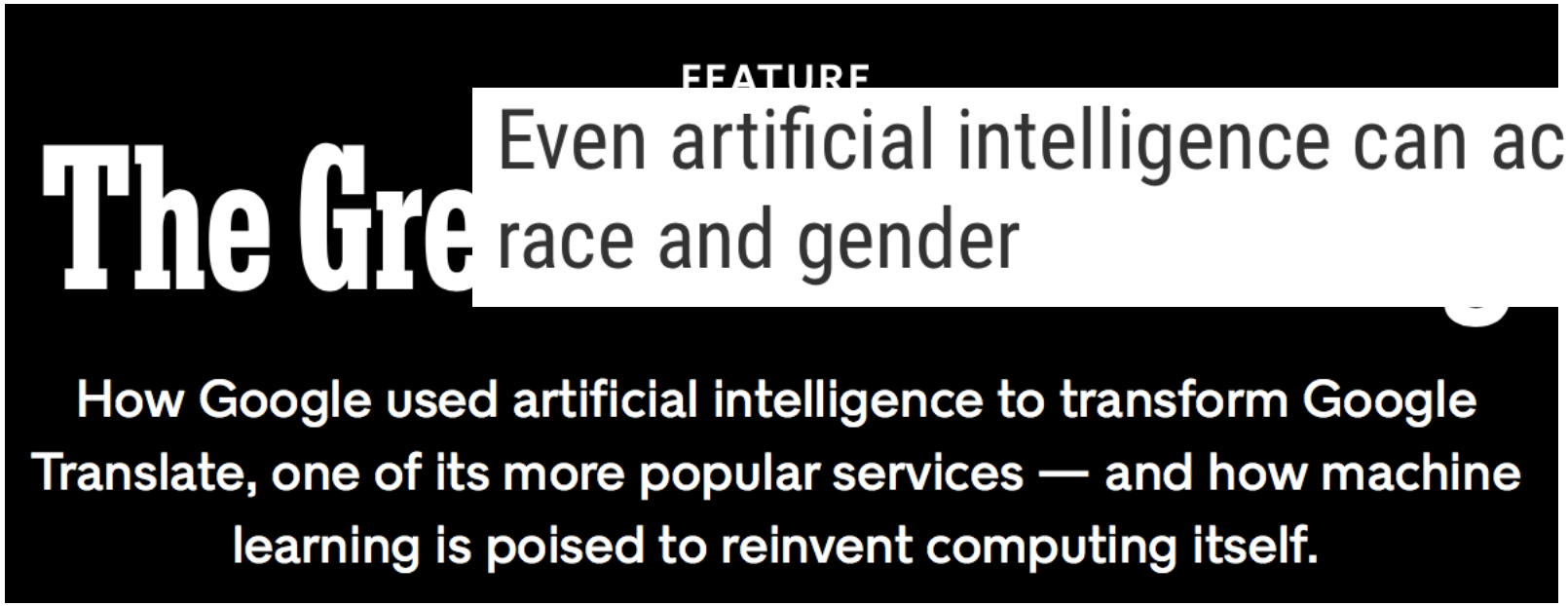
US outbreak foxes a leading web-based method for tracking seasonal flu.

2007 2008

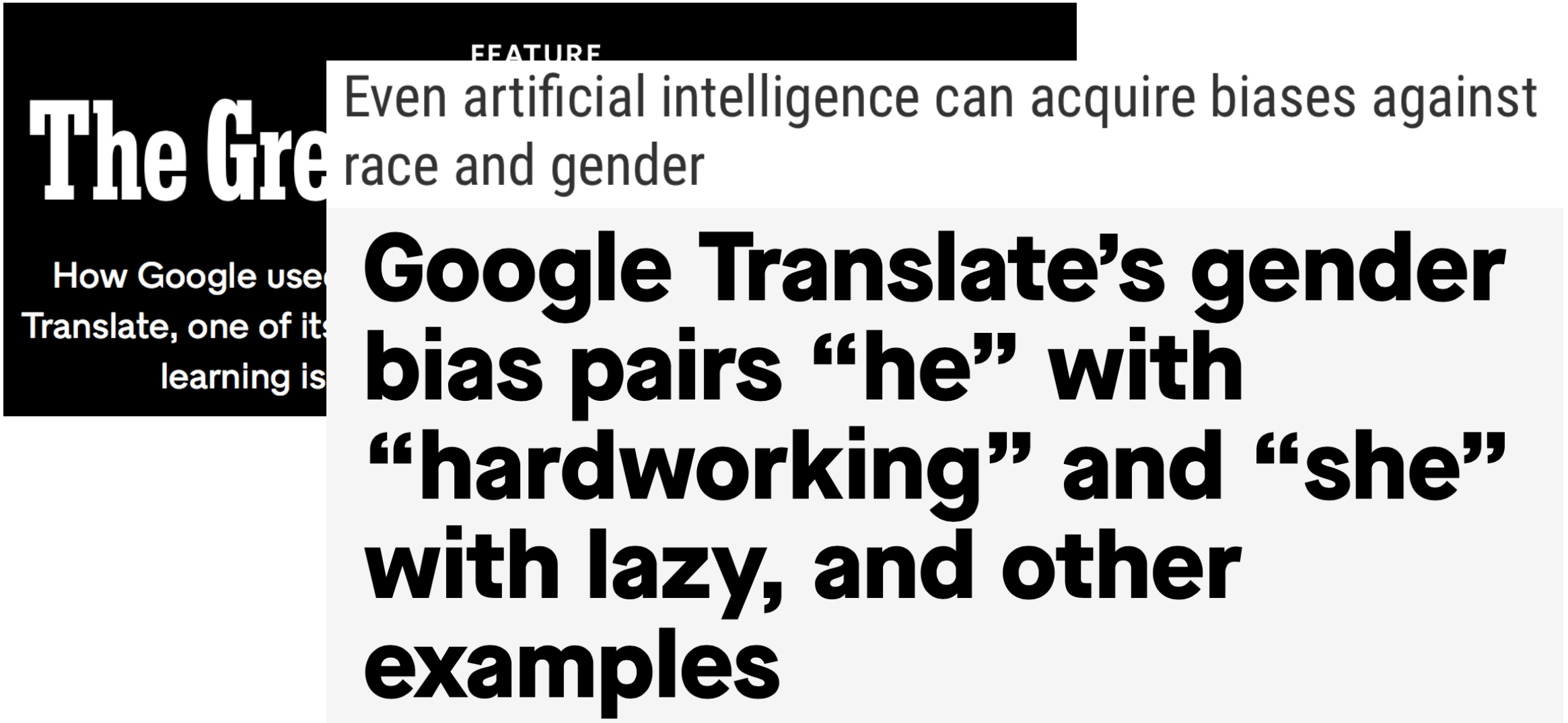
AI for Translation



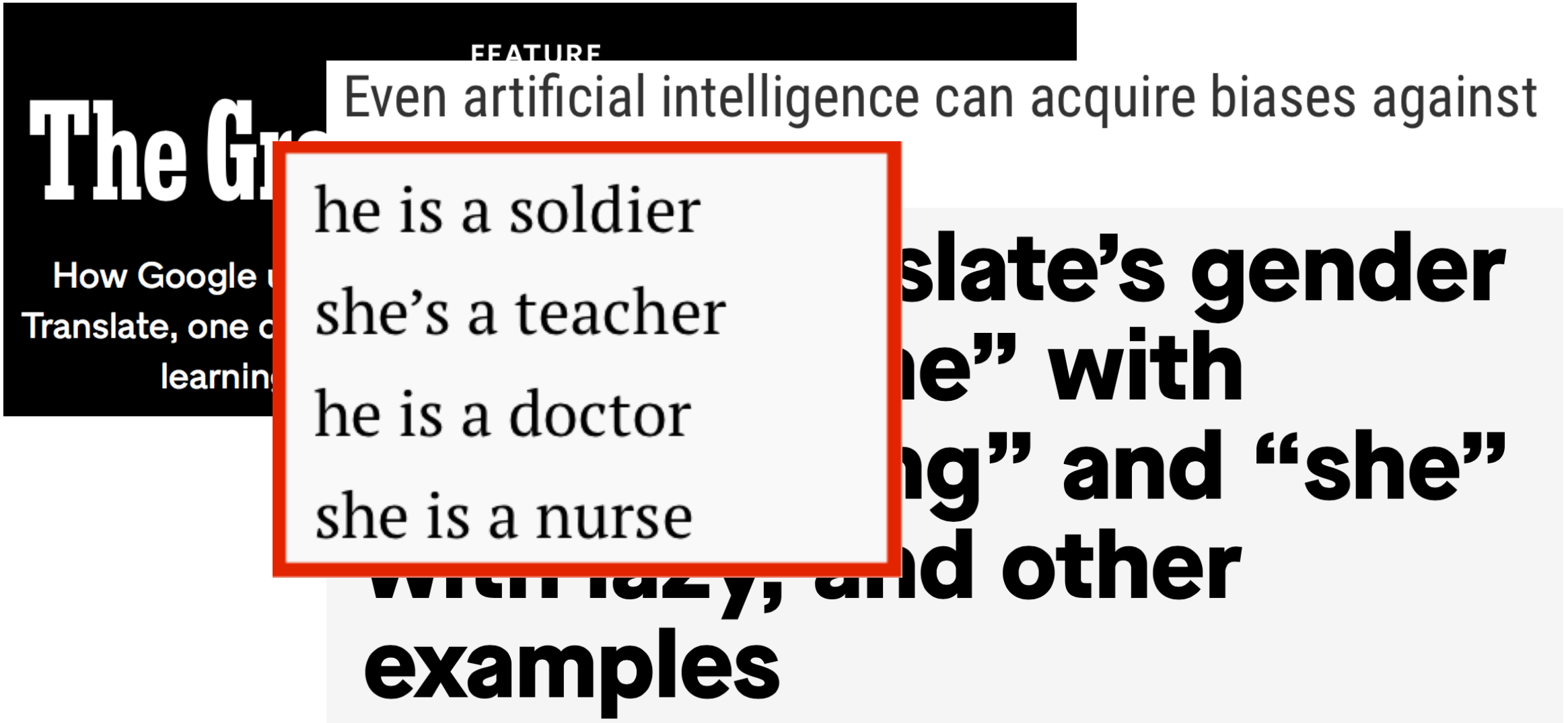
AI for Translation



AI for Translation



AI for Translation



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

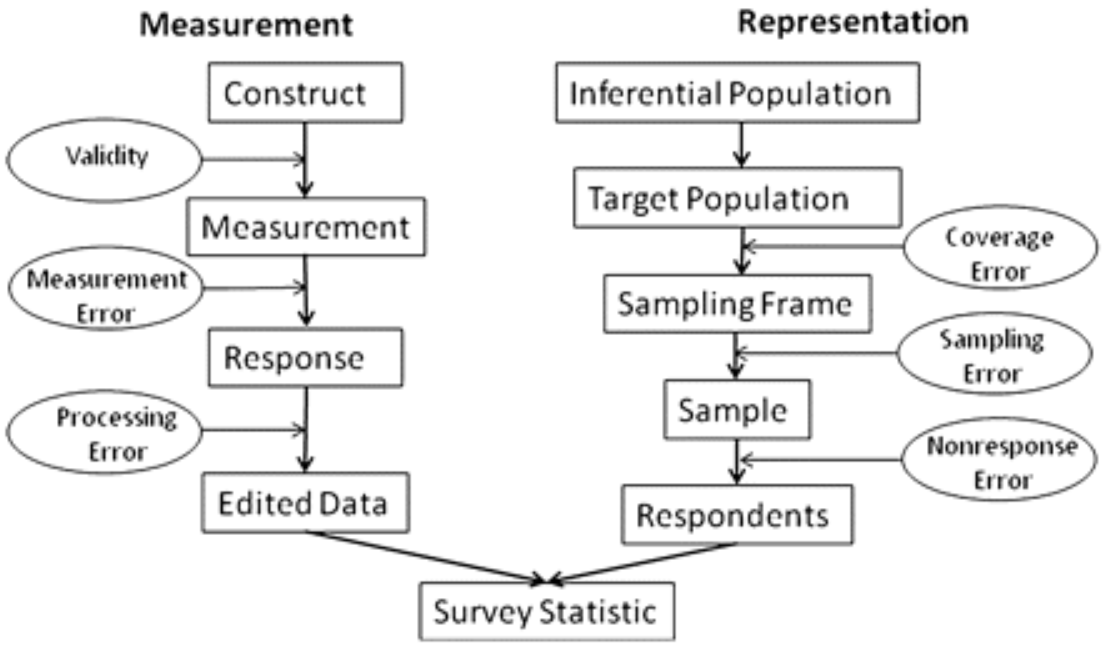
- These cases point to challenges to be overcome, and are important opportunities
  - How can public health practitioners engage with non-traditional partners in a beneficial way?
  - How can tech be used or evaluated as a public health tool when it changes so rapidly?
  - How can big data overcome issues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

A public health lens

How can we use these data to improve health?

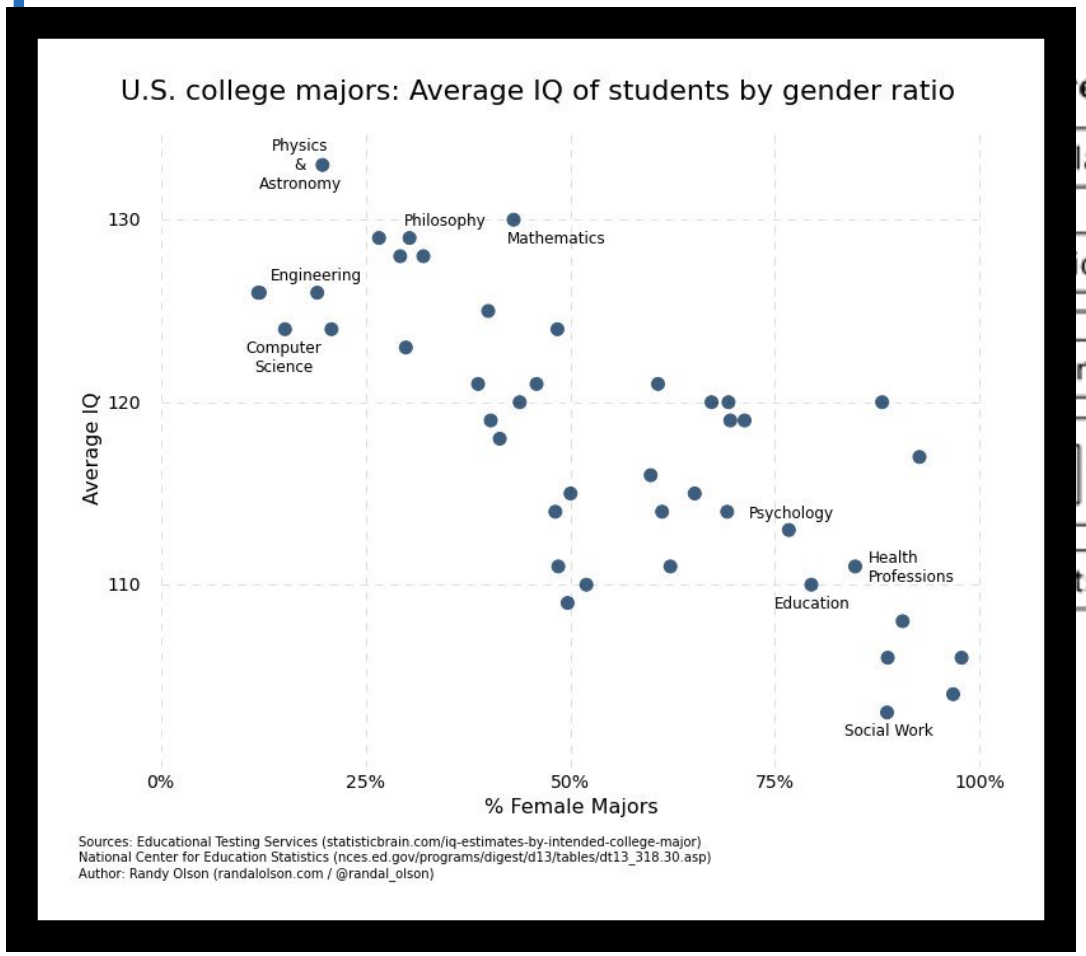
- Who is represented?
- What do measurements mean? Can they be trusted?
- Will I uncover associations? Causes?
- Can I develop new hypotheses or confirm existing ones?
- Can I design an intervention, or evaluate the success of one?

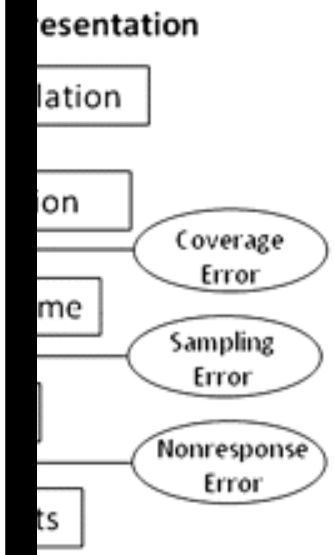
Be skeptical about data



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



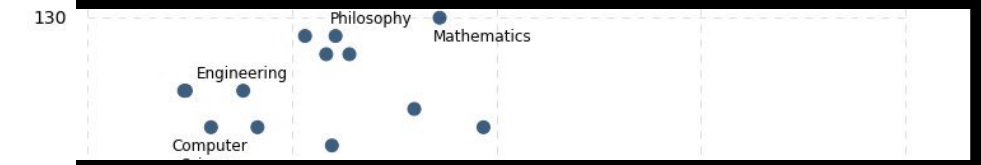


From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
via “Data Alone Isn’t Ground Truth” by Angela Bassa

Be skepti

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 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