Course Introduction and Motivation

About Me

data intensive statistics mostly in molecular biology

website: www.jtleek.com
twitter: @jtleek, @simplystats
blog: simplystatistics.org

About me

Course information

Instructors:

- Jeff Leek
- Elizabeth Colantuoni
- Yates Coley

TAs:

- John Muschelli

Website:

- http://www.jtleek.com/advdatasci
- https://github.com/jtleek/advdatasci

More information

- Time: MW 1:30-2:20
- Location: W2009
- Lab Time: W 12:00
- Lab Location: TBA

Requirements

- Ph.D. student (2nd year) Biostatistics
- Masters student Biostatistics
- Sorry no exceptions

What is 711?

- Historically just a methods course
- Now a combination of methods/data analysis.
- Goals
    - Teach you to think about data
    - Teach you to organize an analysis
    - Help you understand current methods
    - Get you started creating your own methods
    - Teach you practical grad school skills

Course description

Provides an intensive introduction to applied statistics and data analysis. Trains students to become data scientists capable of both applied data analysis and critical evaluation of the next generation next generation of statistical methods. Since both data analysis and methods development require substantial hands-on experience, focuses on hands-on data analysis.

Learning objectives

Upon successfully completing this course, students will be able to:

1. Obtain, clean, transform, and process raw data into usable formats
2. Formulate quantitative models to address scientific questions
3. Organize and perform a complete data analysis, from exploration, to analysis, to synthesis, to communication
4. Apply a range of statistical methods for inference and prediction

What is the point of grad school?

- Freedom
- Discover new knowledge
- Time to dive deep
- Opportunity for leadership
- Opportunity to make a name for yourself
    - R packages
    - Papers
    - Blogs
- A good presentation http://pgbovine.net/phd.htm, he also has more good resources here http://pgbovine.net/phd.htm
- Get a job

What is not the point of grad school

- Grades
- Classes
- Exams
- Proving you are smart
- Competition with other students locally

Grading philosophy

I believe the purpose of graduate education is to train you to be able to think for yourself and initiate and complete your own projects. I am super excited to talk to you about ideas, work out solutions with you, and help you to figure out statistical methods and/or data analysis. I don’t think that graduate school grades are important for this purpose. This means that I don’t care very much about graduate student grades.

TL;DR I don’t care about grades and neither will anyone else

Grading policy

That being said, I have to give you a grade, so I will use grades to help communicate your progress.

1. A - Excellent
2. B - Passing
3. C - Needs improvement

Data analysis assignments

- You will do two
- All documents should be submitted electronically
- You must submit pdfs + rmds

Grading criteria

- Did you answer the scientific question? (30%)
- Did you use appropriate statistical methods? (40%)
- Was your write-up simple, clear, and precise? (20%)
- Was your code reproducible? (10%)

Data analysis reviews

After each data analysis is turned in, they will be randomly assigned to another student for review. Your review will be due one week after it is assigned. Your comments should have the format of a typical peer review. You can find a template and instructions for these reviews herehttps://github.com/jtleek/reviews. You should include a summary of the analyses and conclusions in the project you are reviewing, any major revisions, and any minor revisions. We will also evaluate each data analysis independently to assign a grade.

About you

- Choose a row
- Start rating your comfort with these concepts 0-10:
    - 0 = whoa never heard of that
    - 5 = pretty comfortable, but would like a refresher
    - 10 = Jeff get off the stage I got this

https://docs.google.com/spreadsheets/d/1El48mUK2FVPt_i6WlsvukFoYF1uCU00MqqLeY7W46LE/edit?usp=sharing

http://bit.ly/1Js5SGx

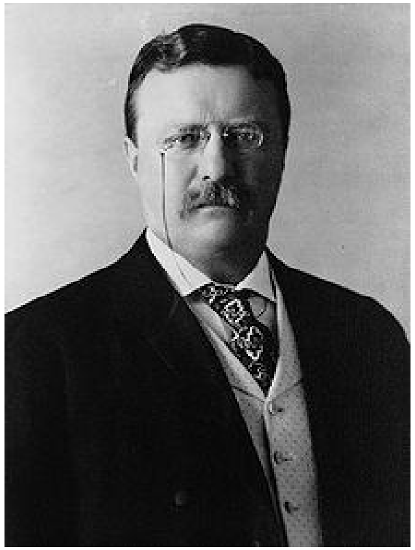
Tentative syllabus

- Version control
- Organize thyself
- EDA
- Regression and generalizations
- Smoothing
- Machine learning/prediction
- High dimensional data
- Simulations

Questions?

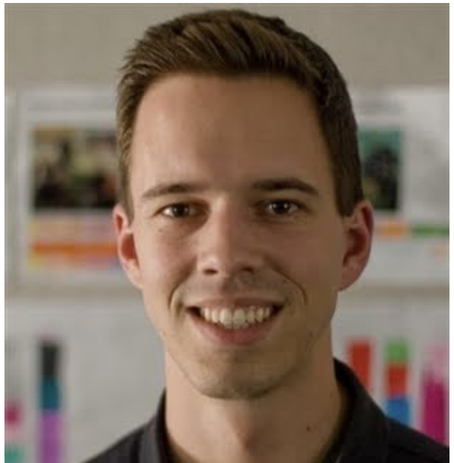
It is not the critic who counts

"It is not the critic who counts: not the man who points out how the strong man stumbles or where the doer of deeds could have done better. The credit belongs to the man who is actually in the arena, whose face is marred by dust and sweat and blood, who strives valiantly, who errs and comes up short again and again, because there is no effort without error or shortcoming, but who knows the great enthusiasms, the great devotions, who spends himself for a worthy cause; who, at the best, knows, in the end, the triumph of high achievement, and who, at the worst, if he fails, at least he fails while daring greatly, so that his place shall never be with those cold and timid souls who knew neither victory nor defeat."



Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn't have a surplus of information and have to filter it out, or you didn't have insufficient information and have to go find some?

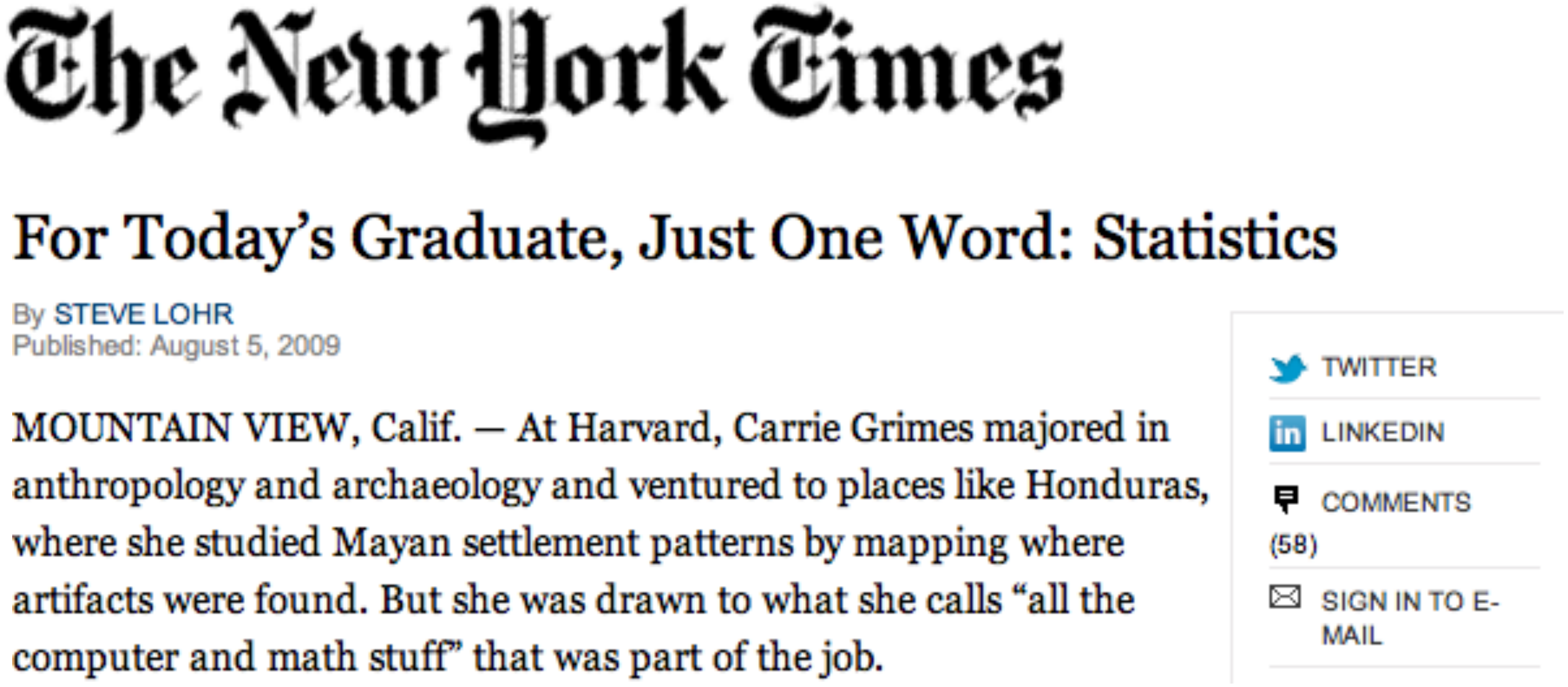
Dan Myer, Mathematics Educator



Why data science?

The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Why data science?



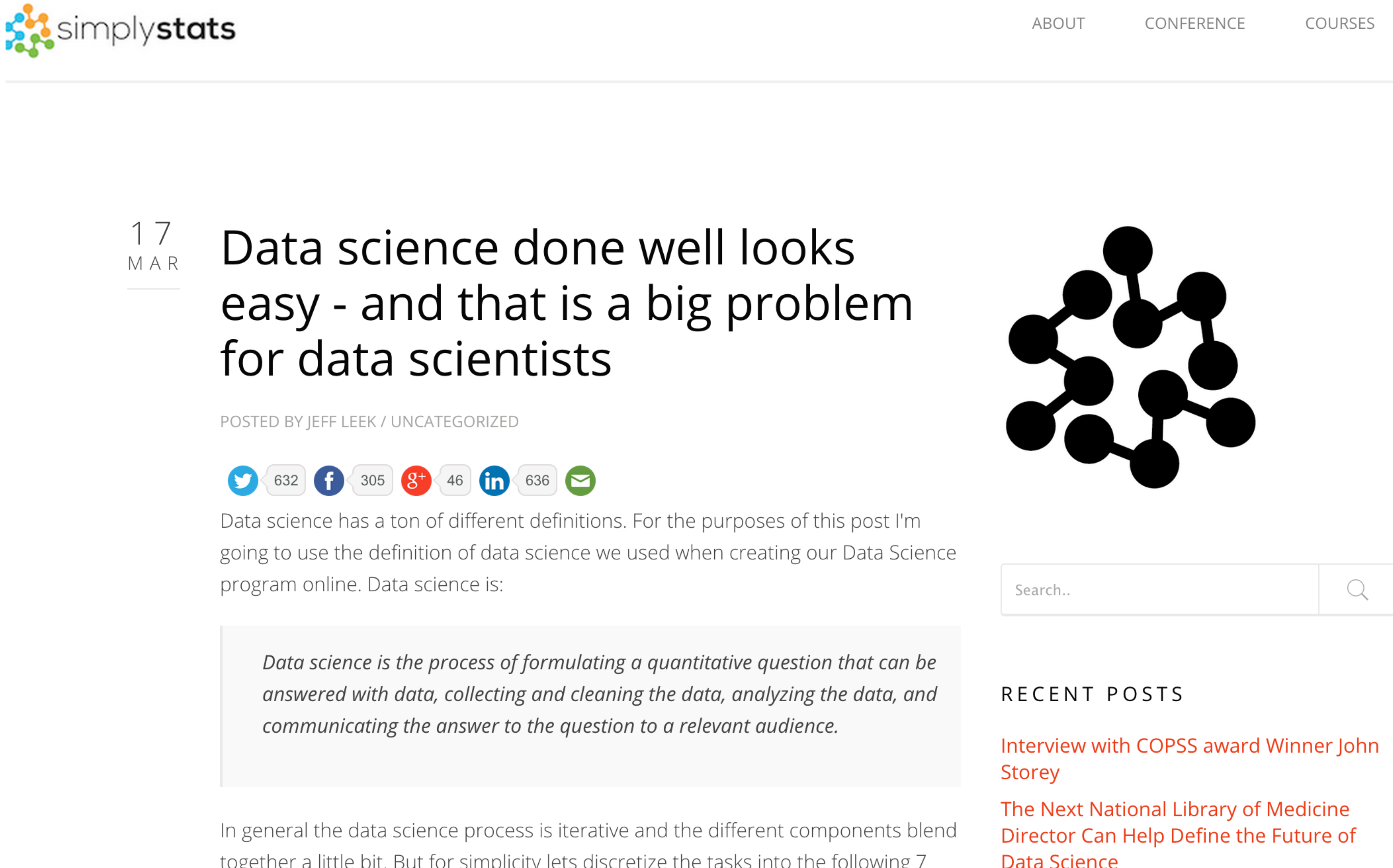
The New York Times

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

What is data science?



17 MAR

Data science done well looks easy - and that is a big problem for data scientists

POSTED BY JEFF LEEK / UNCATEGORIZED

Data science has a ton of different definitions. For the purposes of this post I'm going to use the definition of data science we used when creating our Data Science program online. Data science is:

> *Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.*

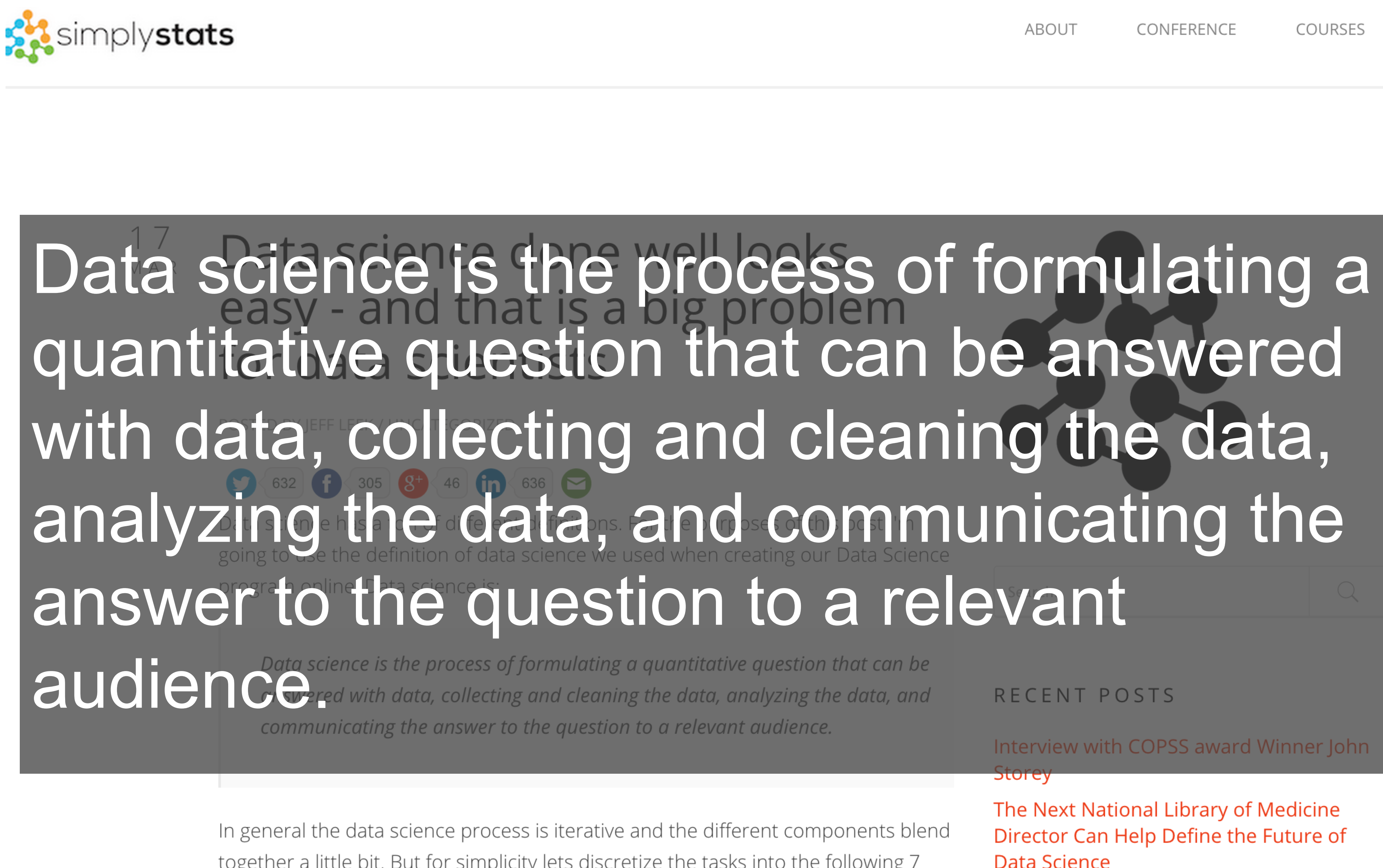
In general the data science process is iterative and the different components blend together a little bit. But for simplicity lets discretize the tasks into the following 7

RECENT POSTS

Interview with COPSS award Winner John Storey

The Next National Library of Medicine Director Can Help Define the Future of Data Science

What is data science?



Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.

Data science done well looks easy - and that is a big problem for data scientists

POSTED BY JEFF LEEK / UNCATEGORIZED

Data science has a lot of different definitions. For the purposes of this post I'm going to use the definition of data science we used when creating our Data Science program online. Data science is:

> *Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data, and communicating the answer to the question to a relevant audience.*

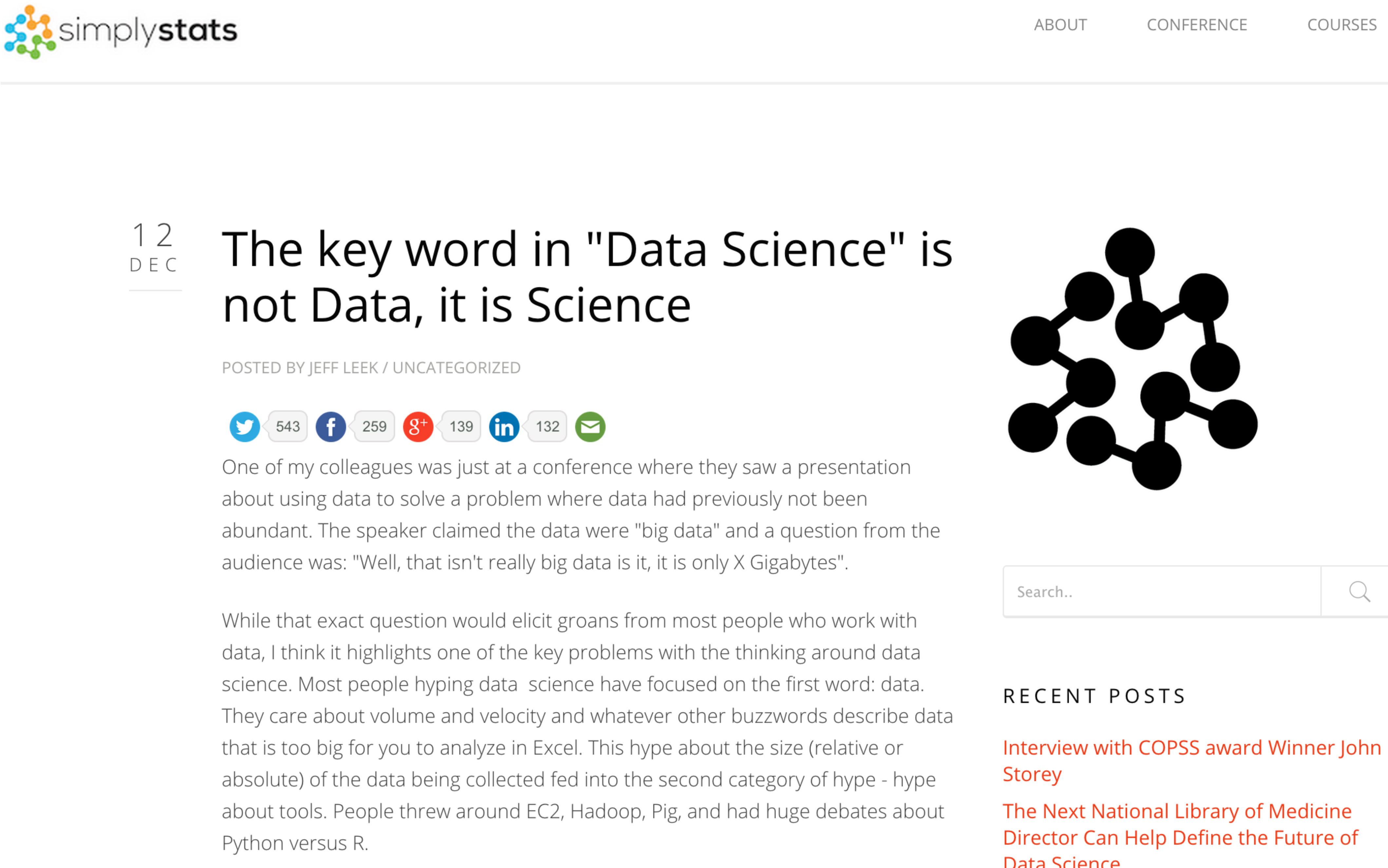
In general the data science process is iterative and the different components blend together a little bit. But for simplicity lets discretize the tasks into the following 7

RECENT POSTS

Interview with COPSS award Winner John Storey

The Next National Library of Medicine Director Can Help Define the Future of Data Science

Data science is **science**



12
DEC

The key word in "Data Science" is not Data, it is Science

POSTED BY JEFF LEEK / UNCATEGORIZED

One of my colleagues was just at a conference where they saw a presentation about using data to solve a problem where data had previously not been abundant. The speaker claimed the data were "big data" and a question from the audience was: "Well, that isn't really big data is it, it is only X Gigabytes".

While that exact question would elicit groans from most people who work with data, I think it highlights one of the key problems with the thinking around data science. Most people hyping data science have focused on the first word: data. They care about volume and velocity and whatever other buzzwords describe data that is too big for you to analyze in Excel. This hype about the size (relative or absolute) of the data being collected fed into the second category of hype - hype about tools. People threw around EC2, Hadoop, Pig, and had huge debates about Python versus R.

RECENT POSTS

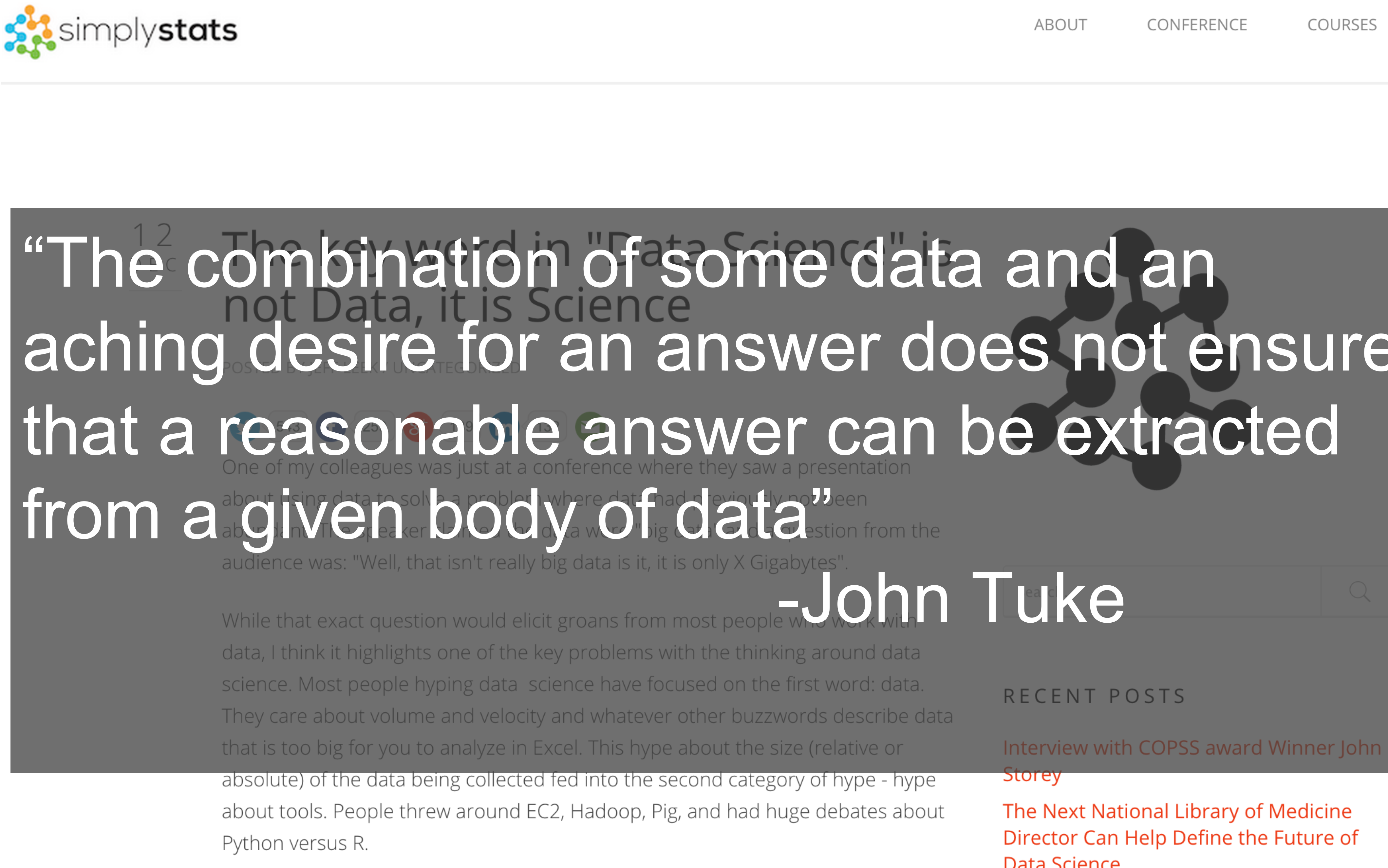
Interview with COPSS award Winner John Storey

The Next National Library of Medicine Director Can Help Define the Future of Data Science

Data science is **science**

“The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data”

-John Tuke



simplystats

ABOUT CONFERENCE COURSES

12 DEC

The key word in "Data Science" is not Data, it is Science

POSTED BY JEFF LEEK / UNCATEGORIZED

One of my colleagues was just at a conference where they saw a presentation about using data to solve a problem where data had previously not been abundant. The speaker claimed the data were "big data" and a question from the audience was: "Well, that isn't really big data is it, it is only X Gigabytes".

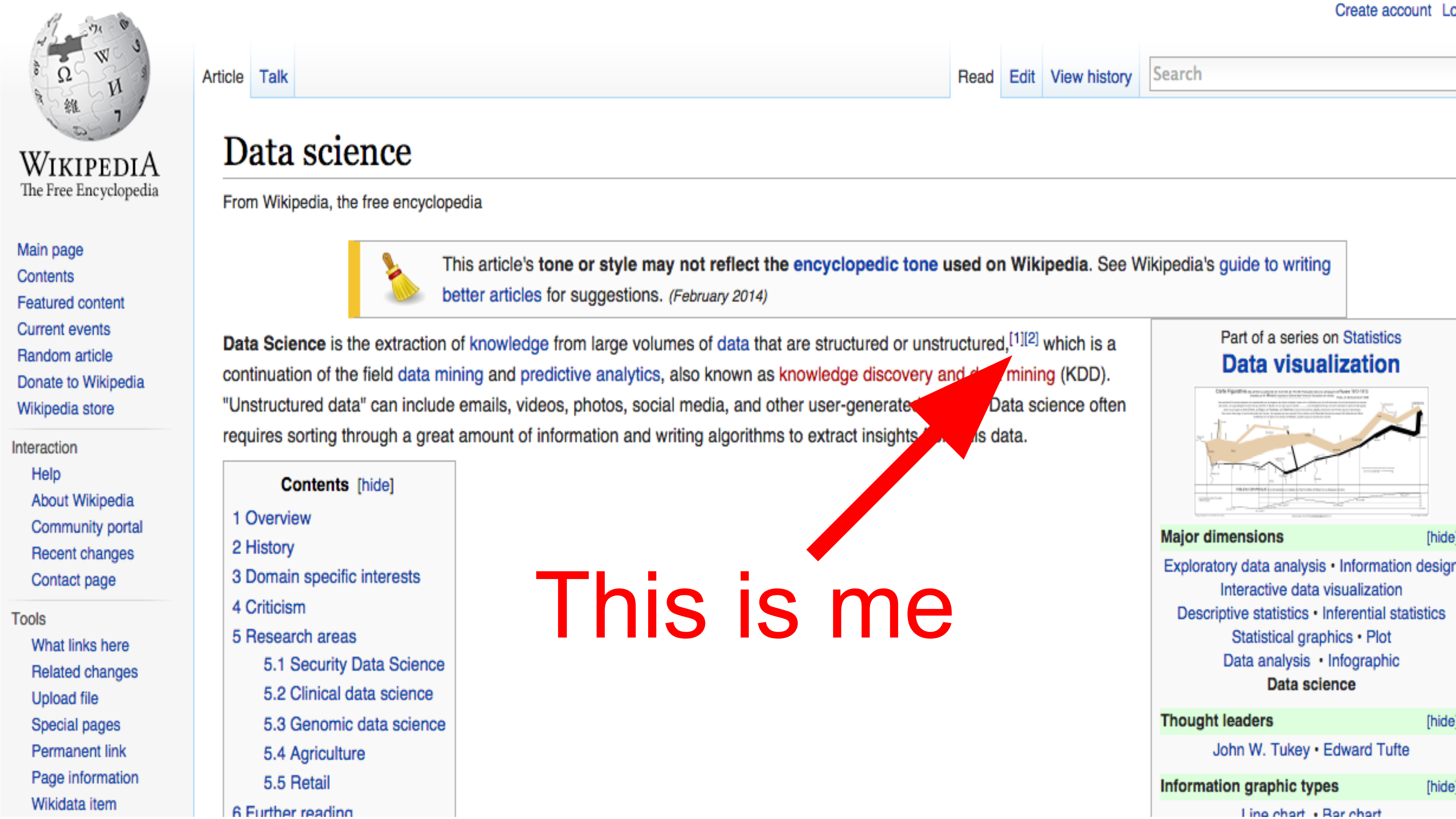
While that exact question would elicit groans from most people who work with data, I think it highlights one of the key problems with the thinking around data science. Most people hyping data science have focused on the first word: data. They care about volume and velocity and whatever other buzzwords describe data that is too big for you to analyze in Excel. This hype about the size (relative or absolute) of the data being collected fed into the second category of hype - hype about tools. People threw around EC2, Hadoop, Pig, and had huge debates about Python versus R.

RECENT POSTS

Interview with COPSS award Winner John Storey

The Next National Library of Medicine Director Can Help Define the Future of Data Science

I defined data science



Data science

From Wikipedia, the free encyclopedia

This article's **tone or style may not reflect the encyclopedic tone used on Wikipedia**. See Wikipedia's guide to writing better articles for suggestions. *(February 2014)*

Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured,[1][2] which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and d mining (KDD). "Unstructured data" can include emails, videos, photos, social media, and other user-generate Data science often requires sorting through a great amount of information and writing algorithms to extract insights is data.

Contents [hide]

Part of a series on Statistics

Data visualization

Major dimensions [hide

Exploratory data analysis • Information desigr
Interactive data visualization
Descriptive statistics • Inferential statistics
Statistical graphics • Plot
Data analysis • Infographic
Data science

Thought leaders [hide

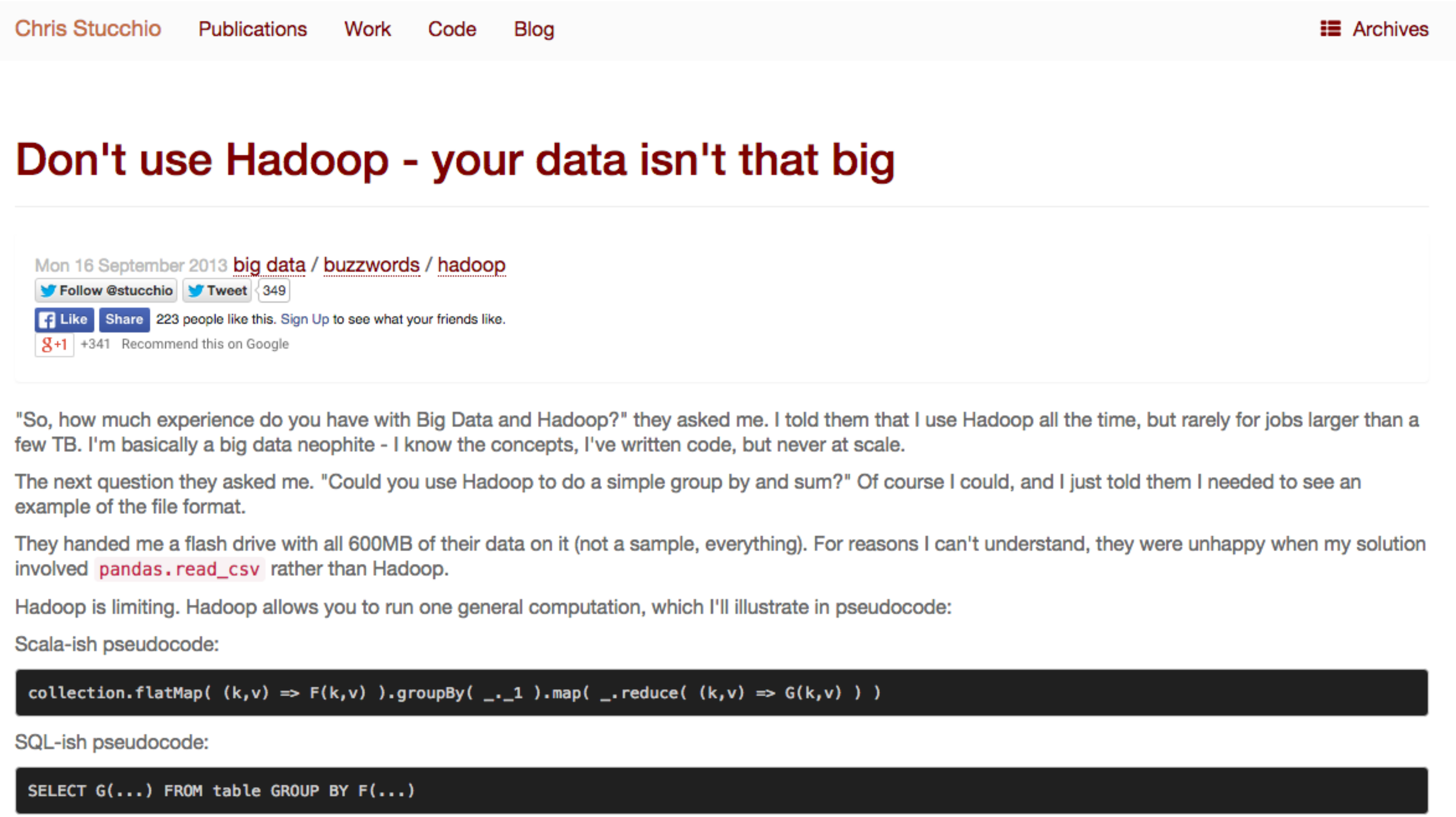
John W. Tukey • Edward Tufte

Information graphic types [hide

Line chart • Bar chart

This is me

What data science is not


Chris Stucchio
Publications
Work
Code
Blog
Archives
Follow @stucchio
Tweet
349
Like
Share
223 people like this. Sign Up to see what your friends like.
+1
+341 Recommend this on Google

Don't use Hadoop - your data isn't that big

Mon 16 September 2013 big data / buzzwords / hadoop

"So, how much experience do you have with Big Data and Hadoop?" they asked me. I told them that I use Hadoop all the time, but rarely for jobs larger than a few TB. I'm basically a big data neophite - I know the concepts, I've written code, but never at scale.

The next question they asked me. "Could you use Hadoop to do a simple group by and sum?" Of course I could, and I just told them I needed to see an example of the file format.

They handed me a flash drive with all 600MB of their data on it (not a sample, everything). For reasons I can't understand, they were unhappy when my solution involved `pandas.read_csv` rather than Hadoop.

Hadoop is limiting. Hadoop allows you to run one general computation, which I'll illustrate in pseudocode:

Scala-ish pseudocode:

```
collection.flatMap( (k,v) => F(k,v) ).groupBy( _._1 ).map( _.reduce( (k,v) => G(k,v) ) )
```

SQL-ish pseudocode:

```
SELECT G(...) FROM table GROUP BY F(...)
```

Questions and data drive data science

How do we make better beer?

Data: Measures of beer quality

Statistic:The t-statistic

What characteristics of field lead to better crops?

Data: Field characteristics, crop yields

Statistic: Analysis of variance (ANOVA)

How long do people live?

Data: Survival times of people (censored)

Statistic: Kaplan-Meier Estimator

What movies will you like?

Data: Lots of other peoples movie ratings

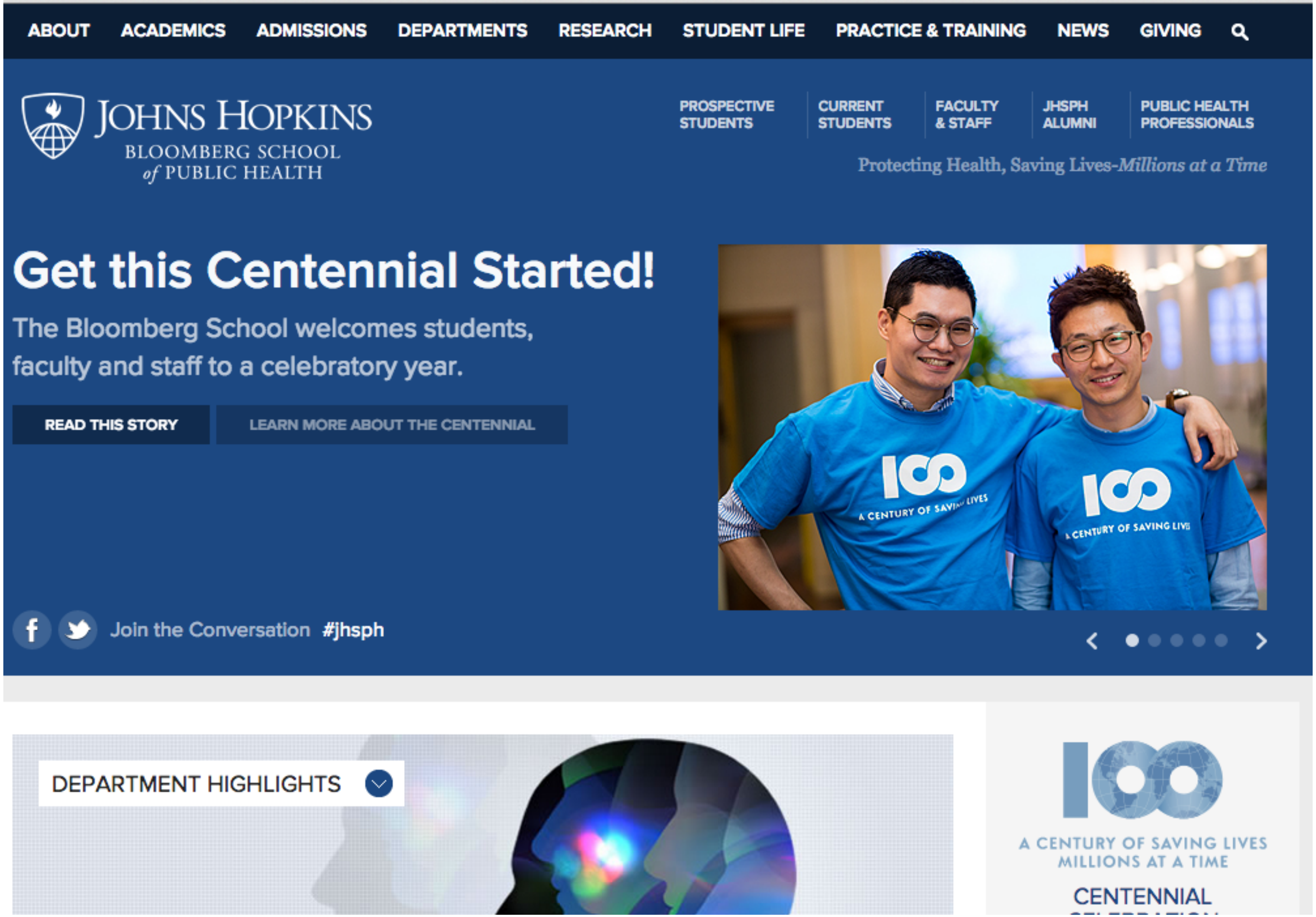
Statistic(s): Recommender systems

Sub-fields of data science

(in no particular order)

1. Biostatistics
2. Data science
3. Machine learning
4. Natural language processing
5. Signal processing
6. Business analytics
7. Econometrics
8. Text mining
9. Statistics in the social sciences
10. Statistical process control

Why are you lucky?



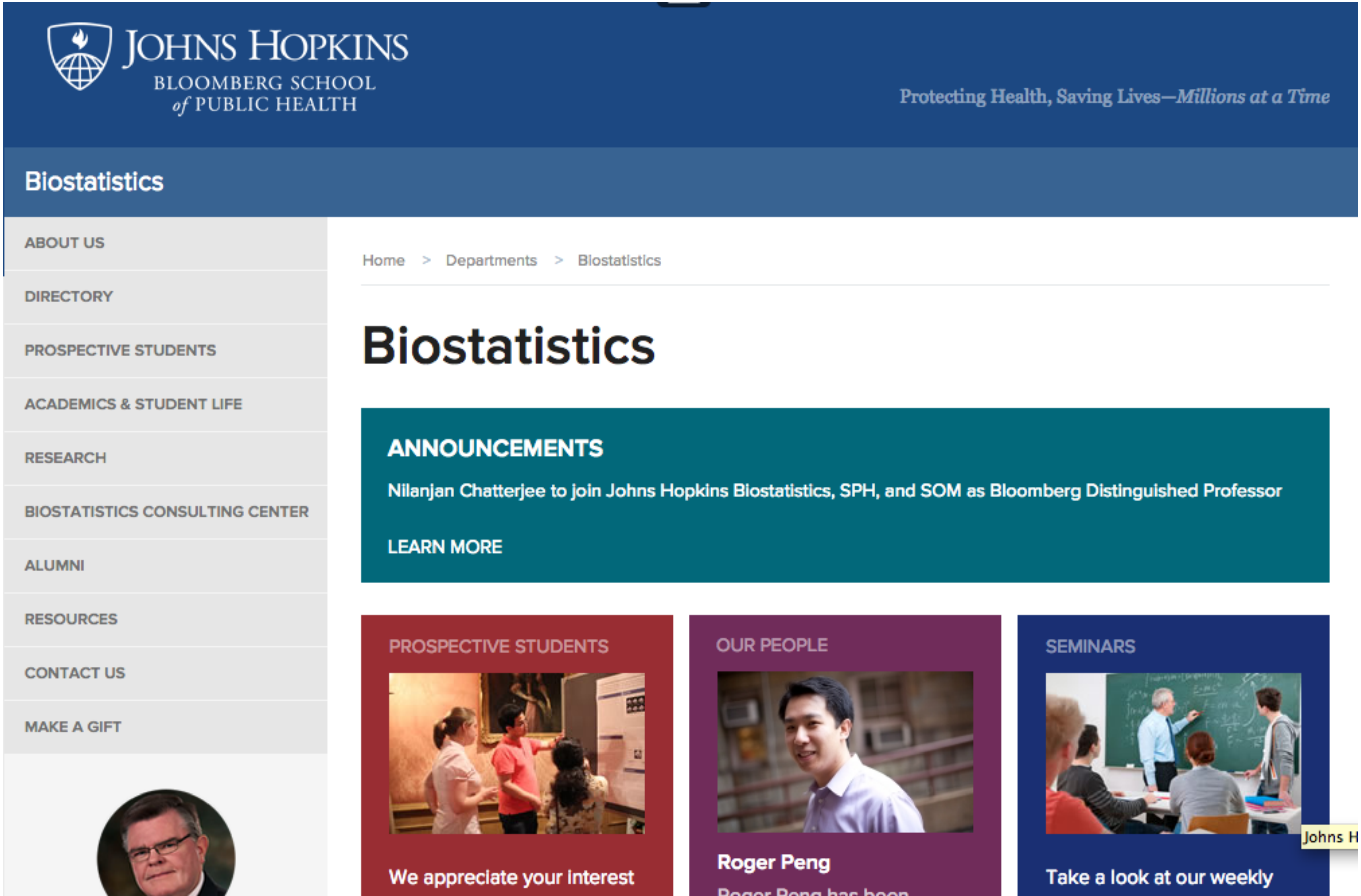
Protecting Health, Saving Lives-*Millions at a Time*

Get this Centennial Started!

The Bloomberg School welcomes students, faculty and staff to a celebratory year.

Join the Conversation #jhsph

Why are you lucky?



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Protecting Health, Saving Lives—*Millions at a Time*

Biostatistics

- ABOUT US
- DIRECTORY
- PROSPECTIVE STUDENTS
- ACADEMICS & STUDENT LIFE
- RESEARCH
- BIOSTATISTICS CONSULTING CENTER
- ALUMNI
- RESOURCES
- CONTACT US
- MAKE A GIFT

Home > Departments > Biostatistics

Biostatistics

ANNOUNCEMENTS

Nilanjan Chatterjee to join Johns Hopkins Biostatistics, SPH, and SOM as Bloomberg Distinguished Professor

LEARN MORE

PROSPECTIVE STUDENTS

We appreciate your interest

OUR PEOPLE

Roger Peng

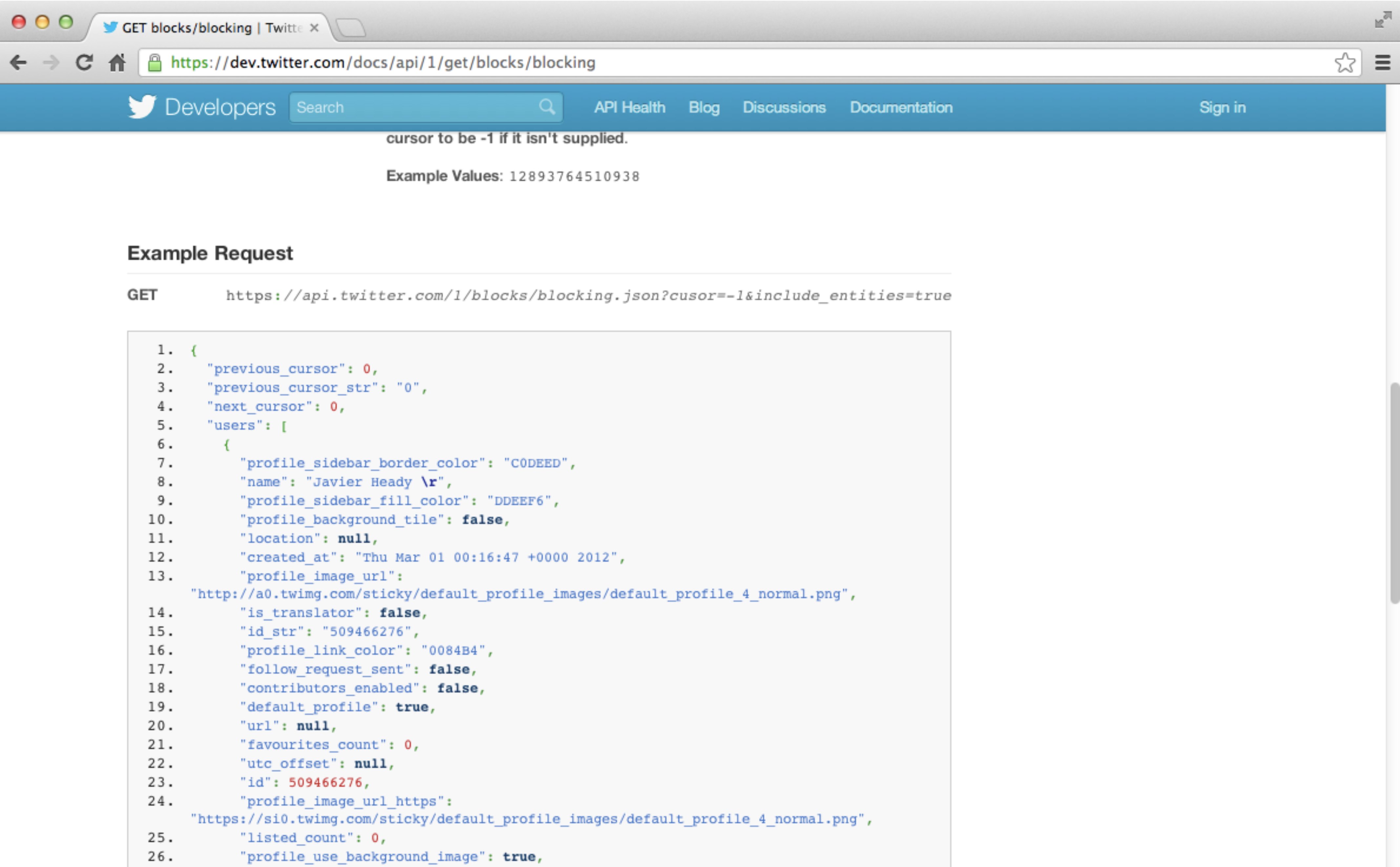
Roger Peng has been

SEMINARS

Take a look at our weekly

Johns H

Why are you lucky?


GET blocks/blocking | Twitte
https://dev.twitter.com/docs/api/1/get/blocks/blocking
Developers
Search
API Health
Blog
Discussions
Documentation
Sign in

cursor to be -1 if it isn't supplied.

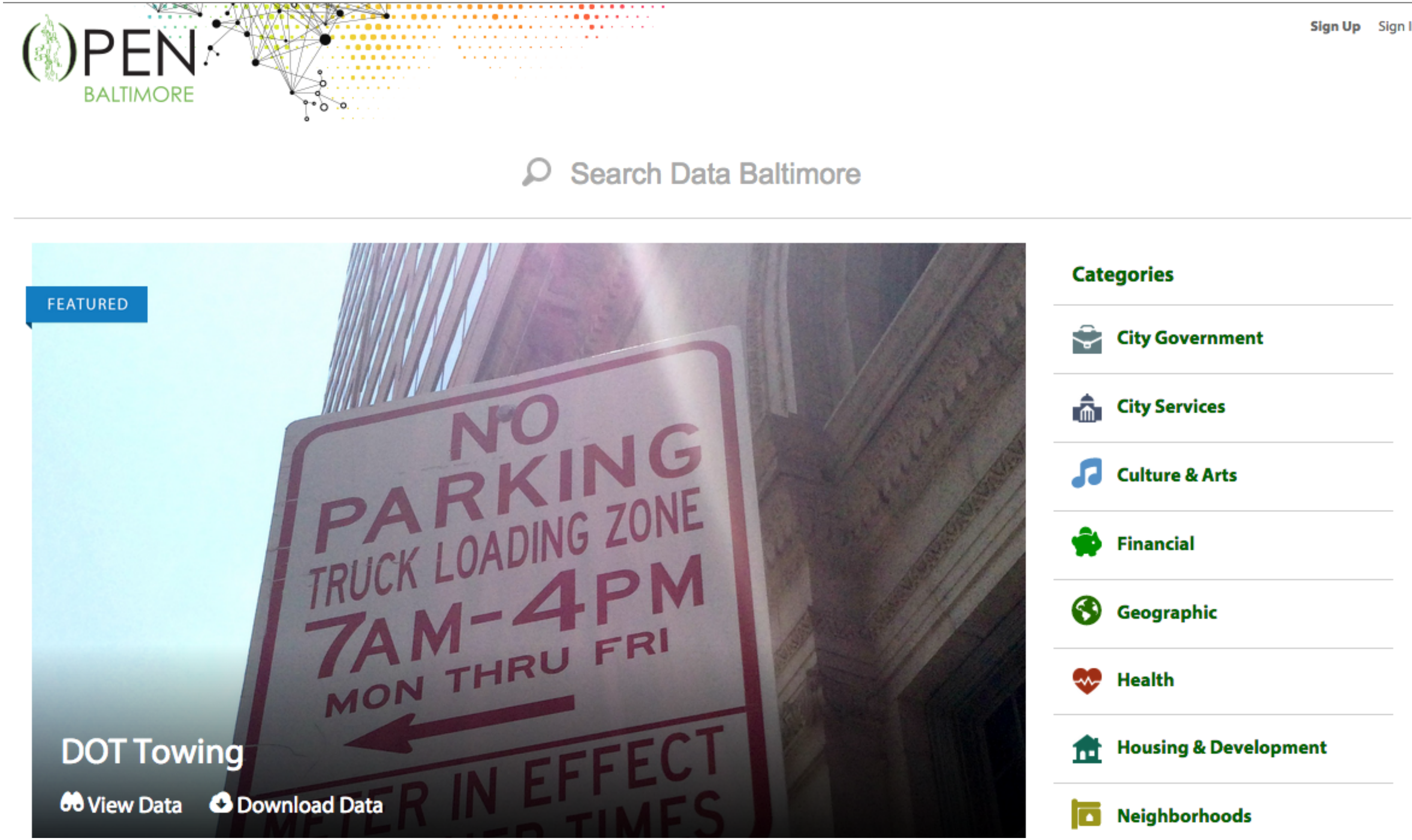
Example Values: `12893764510938`

Example Request

GET `https://api.twitter.com/1/blocks/blocking.json?cusor=-1&include_entities=true`

```
{
  "previous_cursor": 0,
  "previous_cursor_str": "0",
  "next_cursor": 0,
  "users": [
    {
      "profile_sidebar_border_color": "C0DEED",
      "name": "Javier Heady \r",
      "profile_sidebar_fill_color": "DDEEF6",
      "profile_background_tile": false,
      "location": null,
      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
      "profile_image_url":
"http://a0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "is_translator": false,
      "id_str": "509466276",
      "profile_link_color": "0084B4",
      "follow_request_sent": false,
      "contributors_enabled": false,
      "default_profile": true,
      "url": null,
      "favourites_count": 0,
      "utc_offset": null,
      "id": 509466276,
      "profile_image_url_https":
"https://si0.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
      "listed_count": 0,
      "profile_use_background_image": true,
```

Why are you lucky?



OPEN BALTIMORE

Sign Up Sign I

Search Data Baltimore

FEATURED

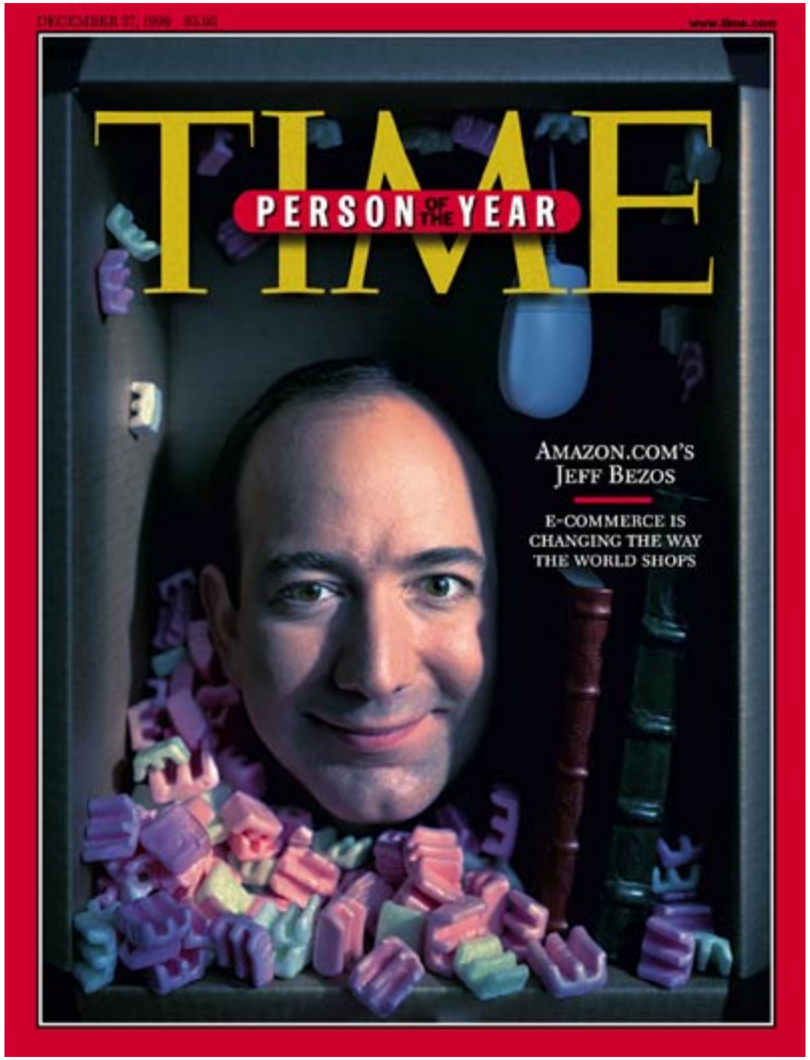
DOT Towing

View Data Download Data

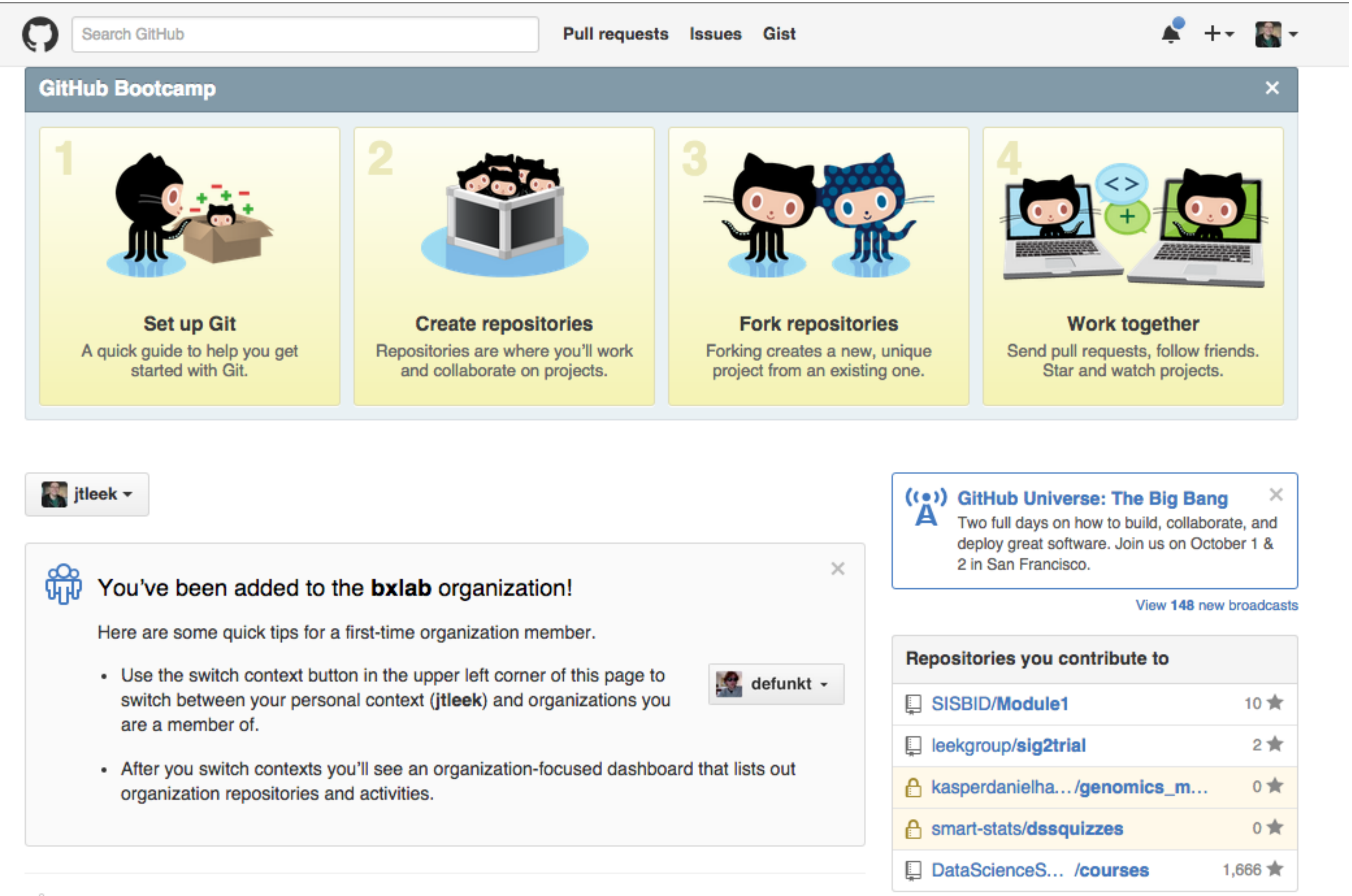
Categories

- City Government
- City Services
- Culture & Arts
- Financial
- Geographic
- Health
- Housing & Development
- Neighborhoods

Why are you lucky?



Why are you lucky?



GitHub Bootcamp

1 **Set up Git**
A quick guide to help you get started with Git.

2 **Create repositories**
Repositories are where you'll work and collaborate on projects.

3 **Fork repositories**
Forking creates a new, unique project from an existing one.

4 **Work together**
Send pull requests, follow friends. Star and watch projects.

jtleek

You've been added to the **bxlab** organization!

Here are some quick tips for a first-time organization member.

- Use the switch context button in the upper left corner of this page to switch between your personal context (**jtleek**) and organizations you are a member of.
- After you switch contexts you'll see an organization-focused dashboard that lists out organization repositories and activities.

defunkt

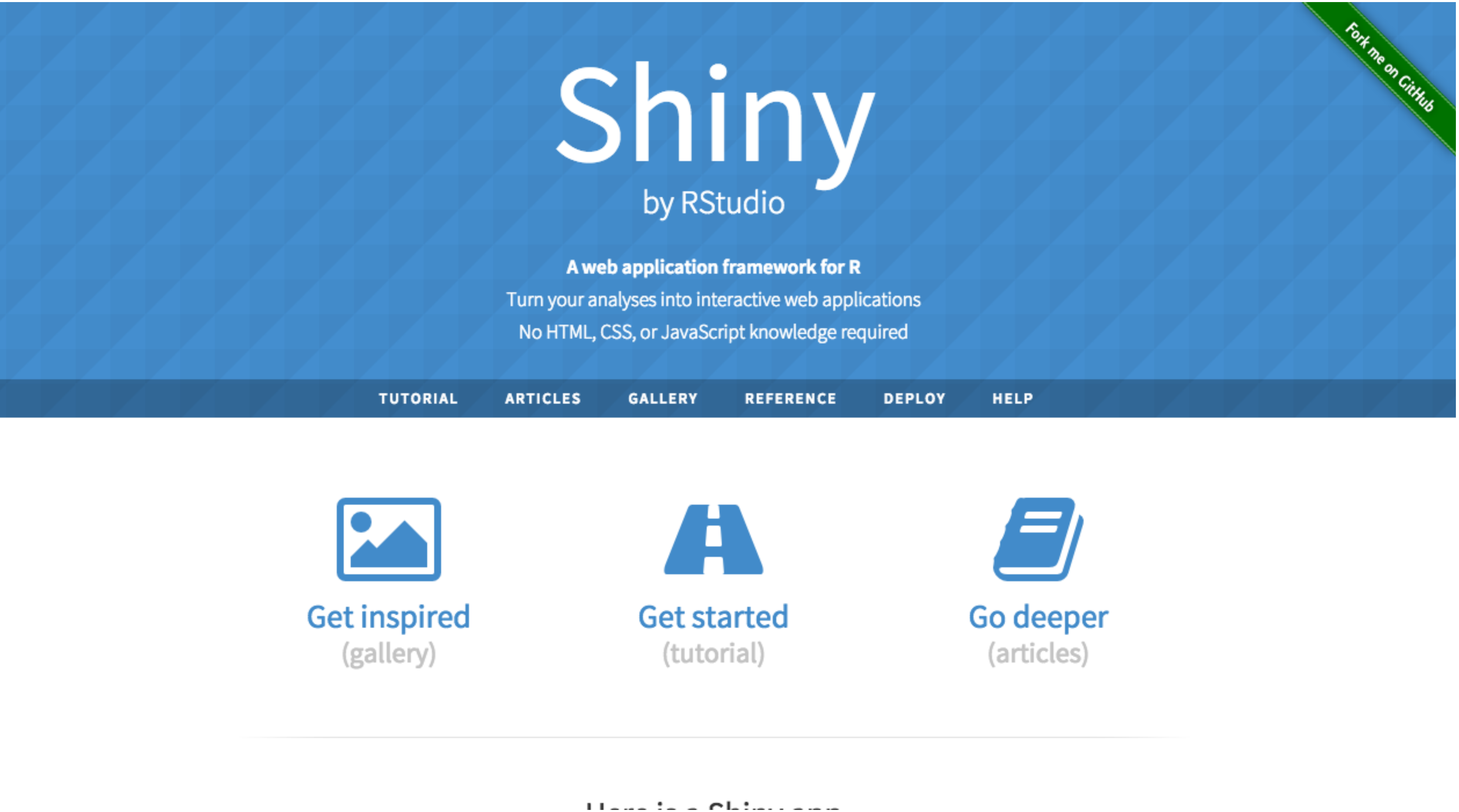
GitHub Universe: The Big Bang
Two full days on how to build, collaborate, and deploy great software. Join us on October 1 & 2 in San Francisco.

View **148** new broadcasts

Repositories you contribute to

SISBID/**Module1**	10 ★
leekgroup/**sig2trial**	2 ★
kasperdanielha…/**genomics_m…**	0 ★
smart-stats/**dssquizzes**	0 ★
DataScienceS… /**courses**	1,666 ★

Why are you lucky?



Shiny

by RStudio

A web application framework for R

Turn your analyses into interactive web applications

No HTML, CSS, or JavaScript knowledge required

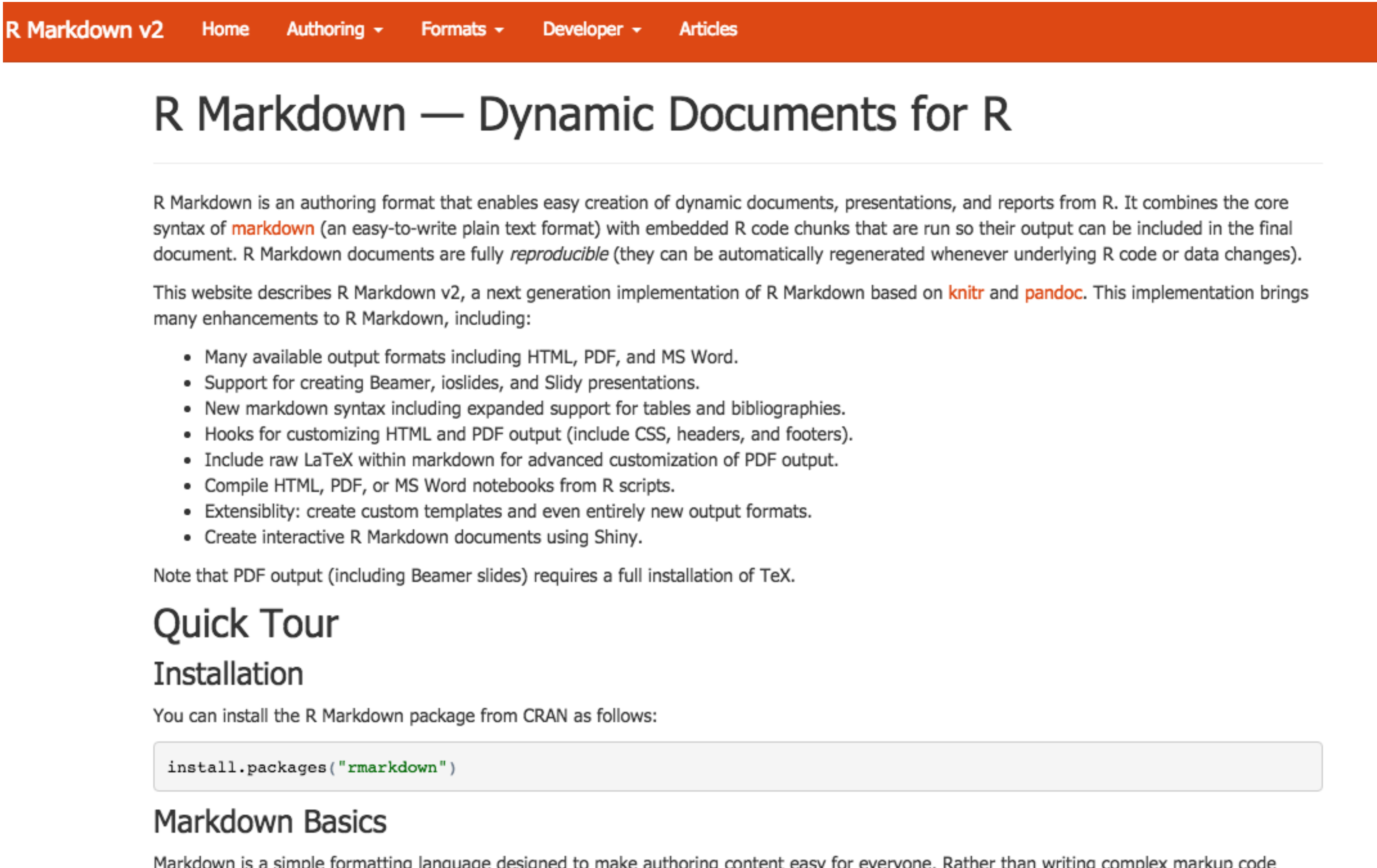
Get inspired (gallery)

Get started (tutorial)

Go deeper (articles)

Here is a Shiny app

Why are you lucky?



R Markdown — Dynamic Documents for R

R Markdown is an authoring format that enables easy creation of dynamic documents, presentations, and reports from R. It combines the core syntax of markdown (an easy-to-write plain text format) with embedded R code chunks that are run so their output can be included in the final document. R Markdown documents are fully *reproducible* (they can be automatically regenerated whenever underlying R code or data changes).

This website describes R Markdown v2, a next generation implementation of R Markdown based on knitr and pandoc. This implementation brings many enhancements to R Markdown, including:

- Many available output formats including HTML, PDF, and MS Word.
- Support for creating Beamer, ioslides, and Slidy presentations.
- New markdown syntax including expanded support for tables and bibliographies.
- Hooks for customizing HTML and PDF output (include CSS, headers, and footers).
- Include raw LaTeX within markdown for advanced customization of PDF output.
- Compile HTML, PDF, or MS Word notebooks from R scripts.
- Extensiblity: create custom templates and even entirely new output formats.
- Create interactive R Markdown documents using Shiny.

Note that PDF output (including Beamer slides) requires a full installation of TeX.

Quick Tour

Installation

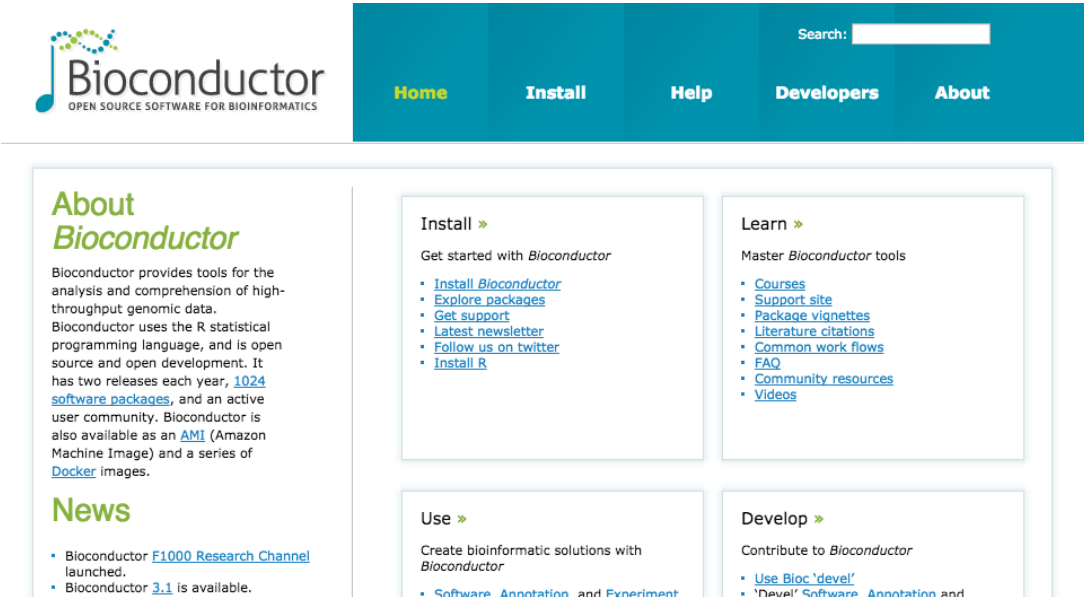
You can install the R Markdown package from CRAN as follows:

```
install.packages("rmarkdown")
```

Markdown Basics

Markdown is a simple formatting language designed to make authoring content easy for everyone. Rather than writing complex markup code

Why are you lucky?



About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1024 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.

News

- Bioconductor F1000 Research Channel launched.
- Bioconductor 3.1 is available.

Install »

Get started with *Bioconductor*

- Install *Bioconductor*
- Explore packages
- Get support
- Latest newsletter
- Follow us on twitter
- Install R

Learn »

Master *Bioconductor* tools

- Courses
- Support site
- Package vignettes
- Literature citations
- Common work flows
- FAQ
- Community resources
- Videos

Use »

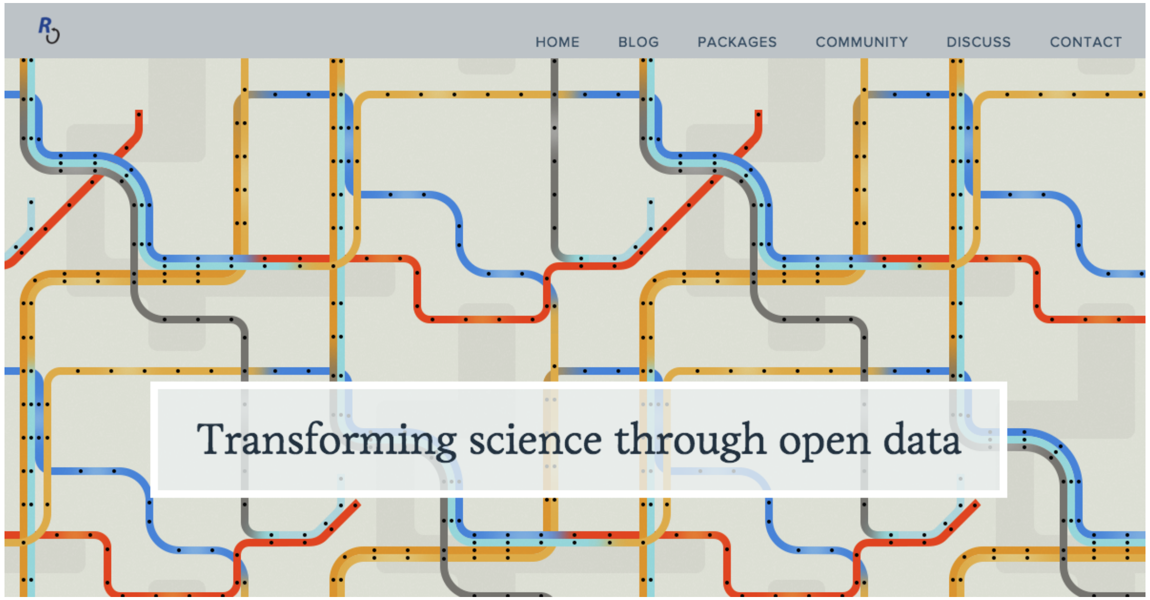
Create bioinformatic solutions with *Bioconductor*

- Software, Annotation, and Experiment

Develop »

Contribute to *Bioconductor*

- Use Bioc 'devel'
- 'Devel' Software, Annotation and



Transforming science through open data

Why you are lucky?

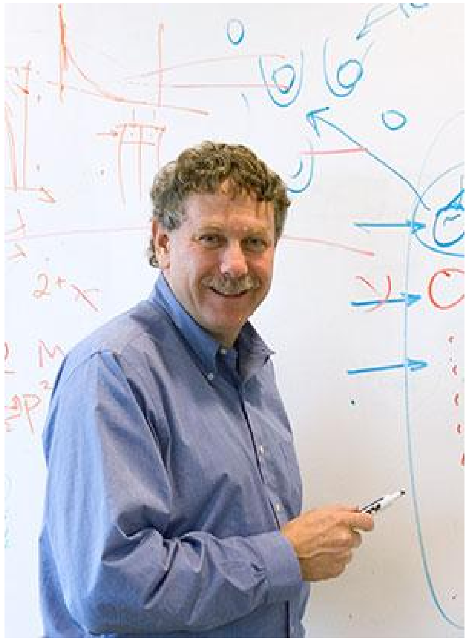
- You are at the best school of public health and medicine in the world
- You are in the oldest/best department of Biostatistics in the world
- Data online is free and abundant
- It is the "1999" of data

Who is a data scientist?

A person who can find, analyze, and visualize data to both identify patterns **and** determine if they are real.

Daryl Morey

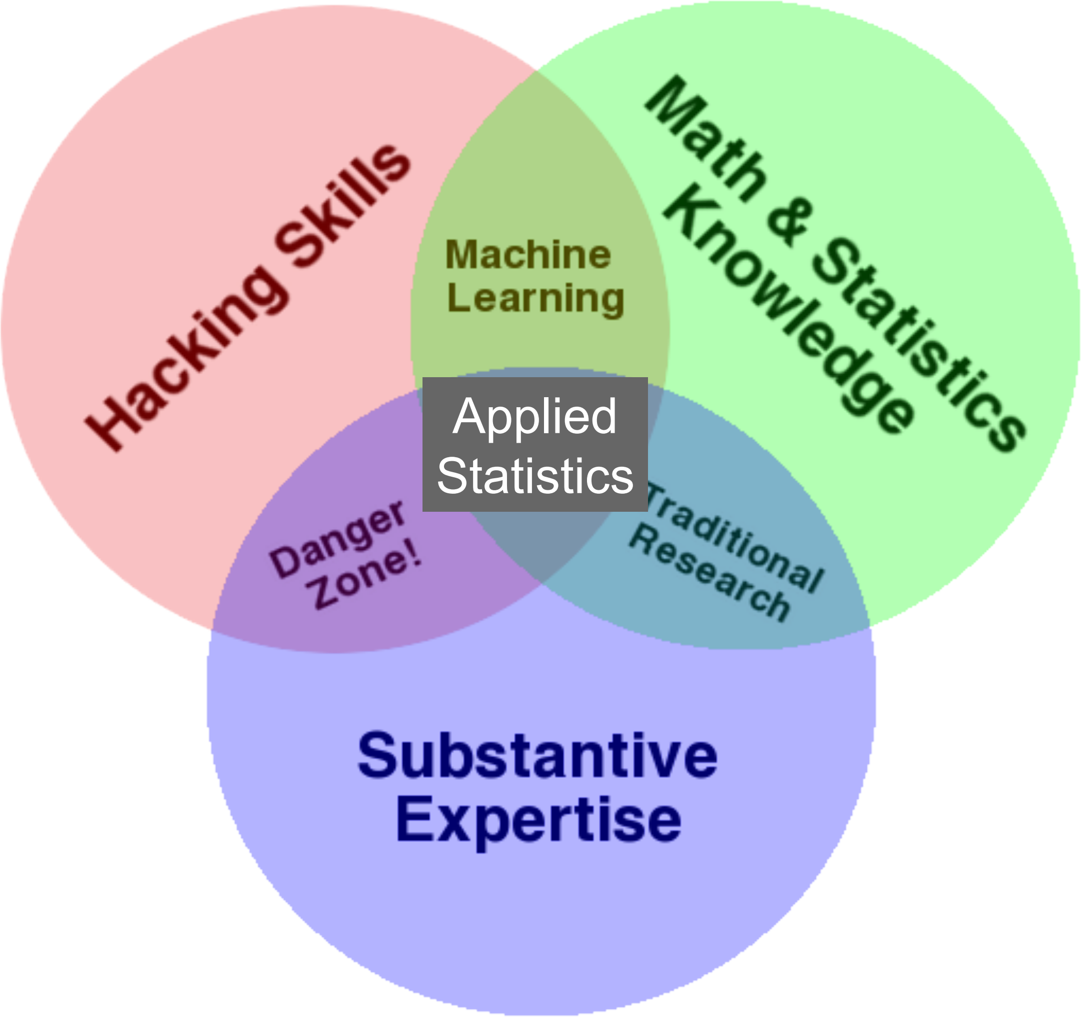
Eric Lander



Nate Silver

Hilary Mason

What is applied statistics?



-Drew Conway

What will this course cover?

1. Translating questions into data analyses
2. Obtaining, organizing, and cleaning data
3. Performing a complete data analysis:
    1. Exploration
    2. Algorithm/model definition
    3. Analysis
    4. Synthesis and communication.
4. Statistical and computational tools

The key challenge in data analysis

Ask yourselves, what problem have you solved, ever, that was worth solving, where you knew knew all of the given information in advance? Where you didn't have a surplus of information and have to filter it out, or you didn't have insufficient information and have to go find some?

-Dan Meyer