

**JSM2015**



*Jennifer*

Jennifer F. Bryan  
Univ. of British Columbia  
Vancouver, BC  
Canada

**stringsAsFactors=  
HELLNO**

QUEEN OF  
SPREADSHEETS

I TWEET

**#rstats**

#rcatladies

Data Scientist



Relax,  
I am a Data Scientist™

# Teach data science and they will come

Joint Statistical Meetings 2015, Seattle, WA

Jennifer (Jenny) Bryan

Dept. of Statistics & Michael Smith Laboratories, UBC

[jenny@stat.ubc.ca](mailto:jenny@stat.ubc.ca)

<http://stat545-ubc.github.io>

<http://www.stat.ubc.ca/~jenny/>

 [@JennyBryan](https://twitter.com/JennyBryan)

 [@STAT545](https://twitter.com/STAT545)

 [@jennybc](https://github.com/jennybc)

links, files, etc. available here

The screenshot shows a GitHub repository page. At the top, there's a navigation bar with the GitHub logo, a search bar, and links for 'Pull requests', 'Issues', and 'Gist'. Below this, the repository name 'jennybc / 2015-08\_bryan-jsm-stat-data-sci-talk' is displayed, along with 'Unwatch', 'Star' (0), and 'Fork' (0) buttons. The main content area shows the repository's metadata: '1 commit', '1 branch', '0 releases', and '1 contributor'. A green 'Branch: master' dropdown is visible. Below this, a 'readme' section shows a commit by 'jennybc' from 3 minutes ago. The 'README.md' file is expanded, showing the title '2015-08-11\_bryan-jsm-talk'. The text in the README describes a talk at the Joint Statistical Meetings 2015, organized by Jeff Leek, and lists two other talks from 2015. On the right side, there's a sidebar with links for 'Code', 'Issues' (0), 'Pull requests' (0), 'Wiki', 'Pulse', 'Graphs', and 'Settings'. At the bottom of the sidebar, there's a section for 'HTTPS clone URL' with the URL 'https://github.com/jennybc/2015-08\_bryan-jsm-stat-data-sci-talk' and buttons for 'Clone in Desktop' and 'Download ZIP'.

Bryan talk at JSM 2015 re: are statisticians data scientists — Edit

1 commit 1 branch 0 releases 1 contributor

Branch: master 2015-08\_bryan-jsm-stat-data-sci-talk / +

readme

jennybc authored 3 minutes ago latest commit 2f813c7b1f

README.md readme 3 minutes ago

## 2015-08-11\_bryan-jsm-talk

Talk at the Joint Statistical Meetings 2015. Session: [The Statistics Identity Crisis: Are We Really Data Scientists?](#), organized by [Jeff Leek](#).

Two other talks in semi-recent past with some overlap (you might like the slides or links):

- [New tools and workflow for data analysis](#), Fields Institute workshop, February 2015
- [How R Markdown + Git + GitHub changed my \(teaching?\) life](#), R Summit and Workshop, June 2015

Code

Issues 0

Pull requests 0

Wiki

Pulse

Graphs

Settings

HTTPS clone URL

[https://github.com/jennybc/2015-08\\_bryan-jsm-stat-data-sci-talk](https://github.com/jennybc/2015-08_bryan-jsm-stat-data-sci-talk)

You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

Clone in Desktop

Download ZIP

[https://github.com/jennybc/2015-08\\_bryan-jsm-stat-data-sci-talk](https://github.com/jennybc/2015-08_bryan-jsm-stat-data-sci-talk)

in a wide array of academic fields,  
the **ability to effectively process data**  
is superseding other more classical modes of research



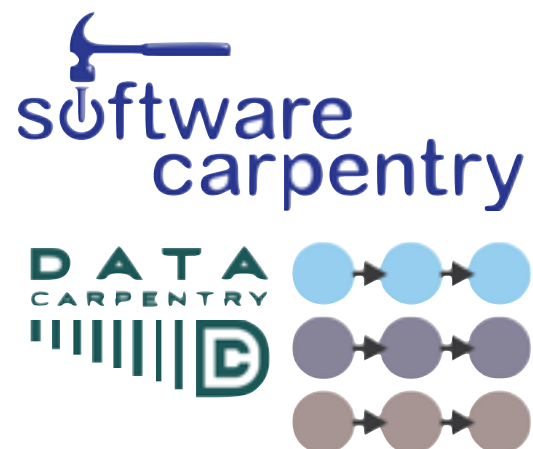


Exploratory Data Analysis  
grad course at UBC  
since 2008 (at least)



Statistics for High Dimensional Biology  
grad course at UBC  
since 2001

w/ R. Gottardo, P. Pavlidis, G. Cohen-Freue, S. Mostafavi



Software Carpentry, Data Carpentry, Reproducible Science  
since 2012

real  
world  
data



statistical  
theory

# Data wrangling, exploration, and analysis with R

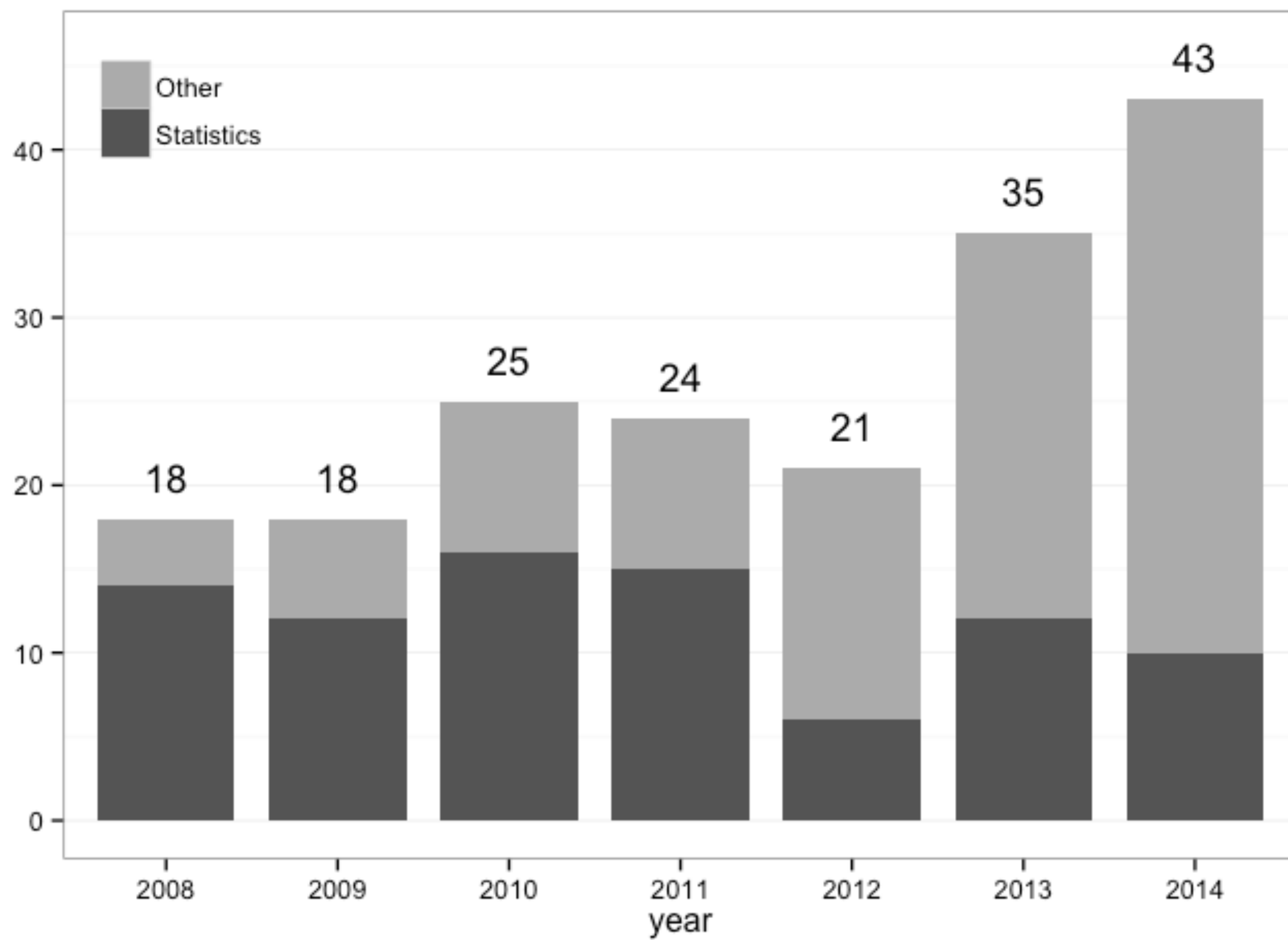
UBC STAT 545A and 547M

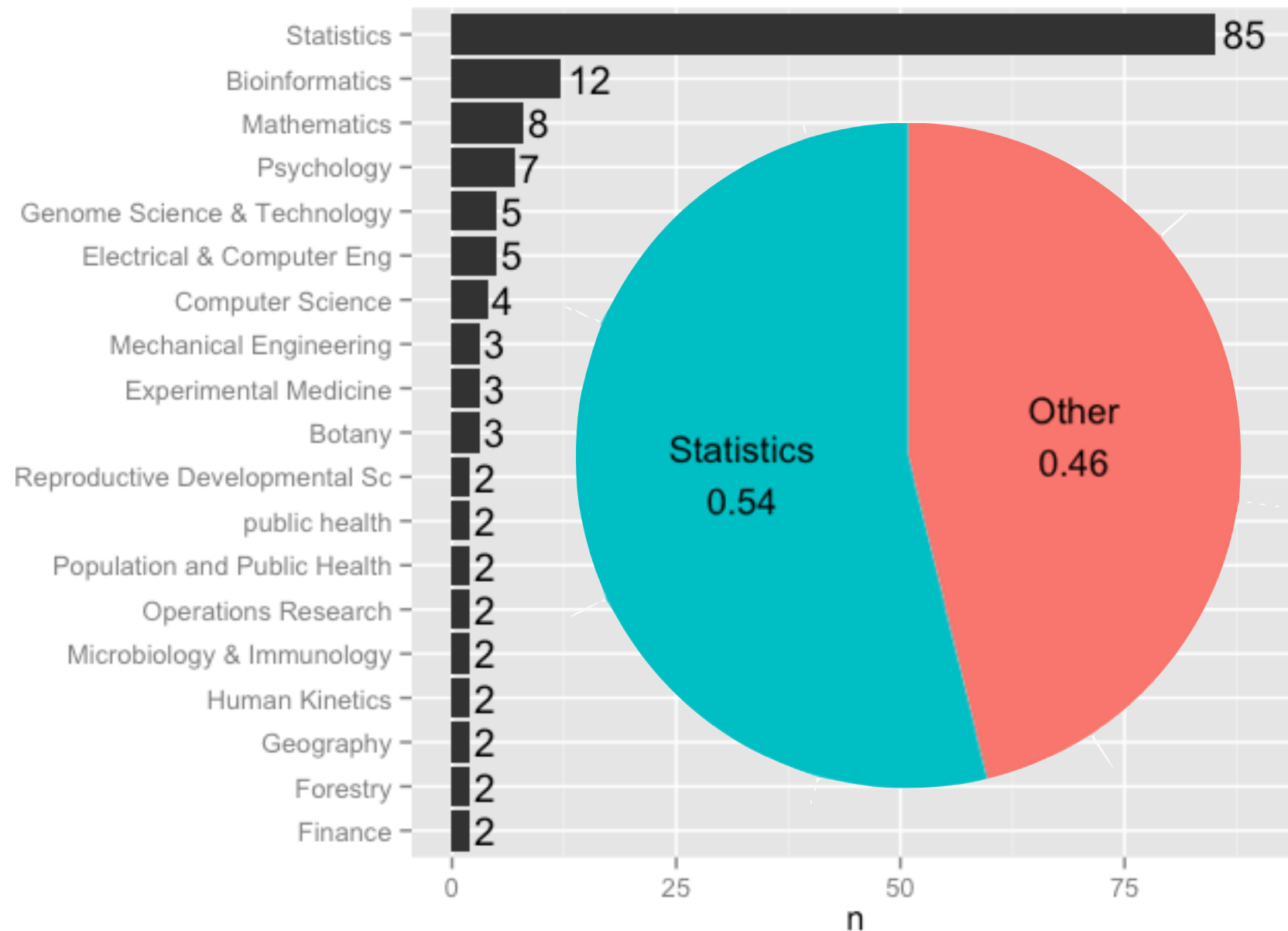
Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

<http://stat545-ubc.github.io>





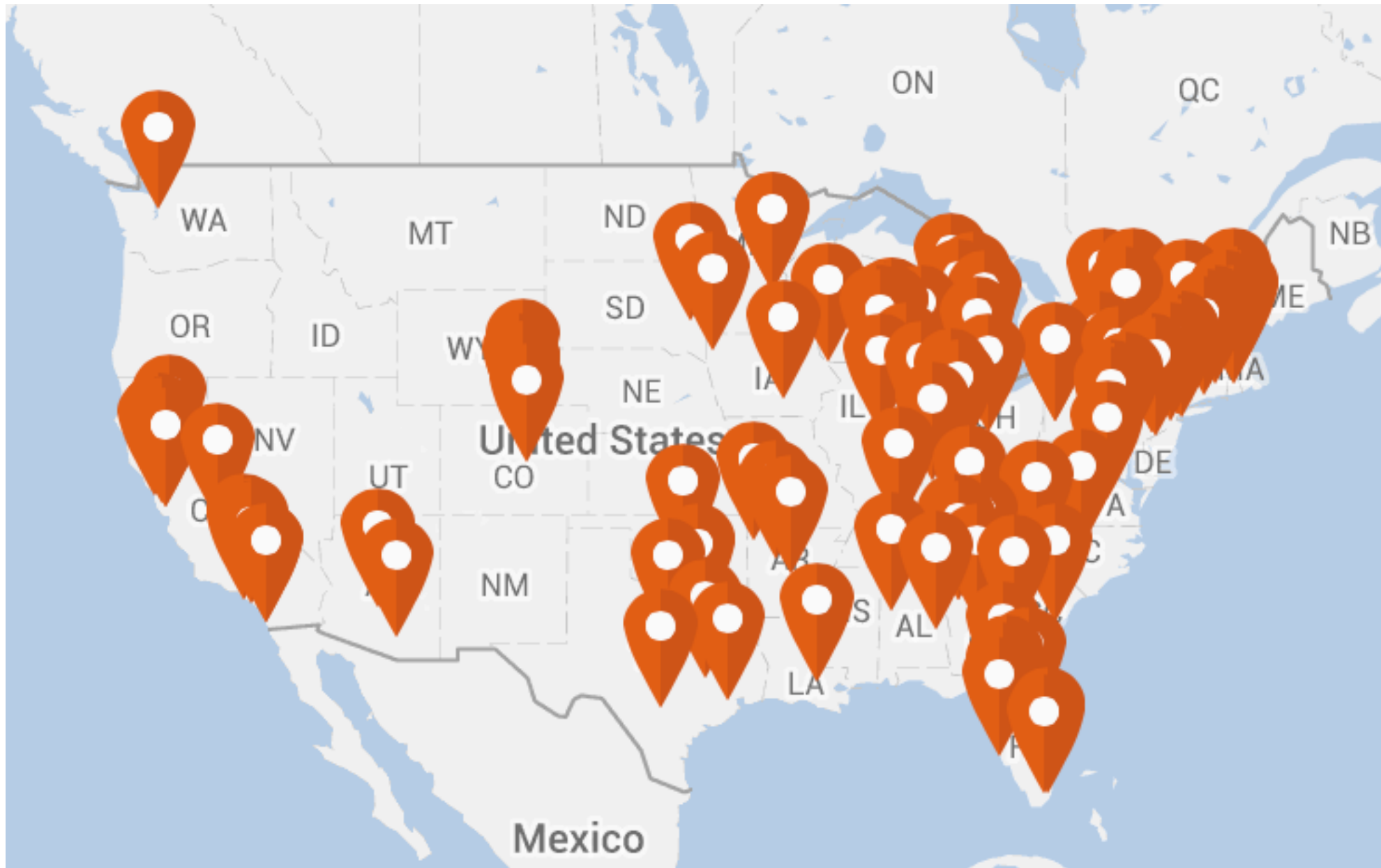


184 = cumulative enrollment 2008 - 2014

48 = # distinct programs sending students

19 = # programs with 2+ students

>300 data science degree programs  
>180 in the US alone



**Data Science Degrees – Analyzed and Visualized**

<http://www.kdnuggets.com/2015/07/data-science-degrees-analyzed.html>

# Data Science Bootcamp Programs

<http://yet-another-data-blog.blogspot.ca/2014/04/data-science-bootcamp-landscape-full.html>

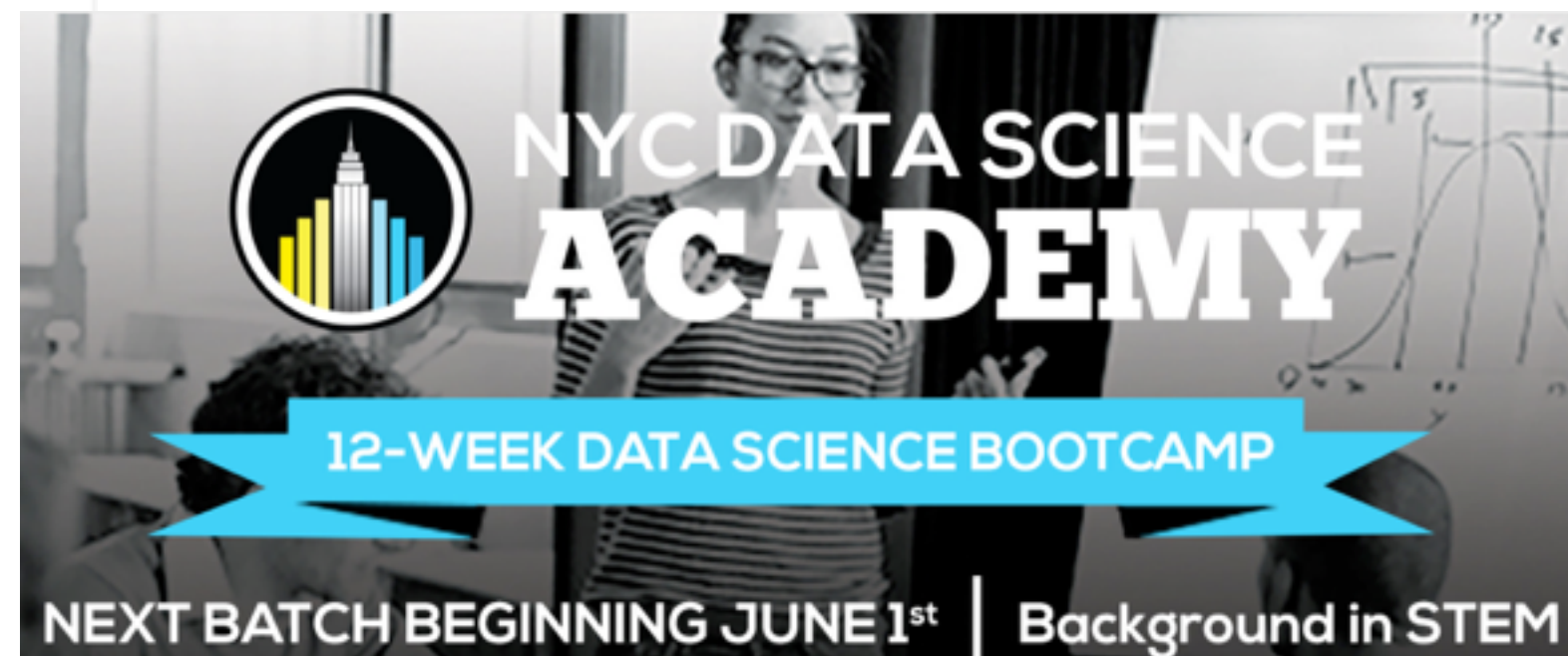
- > 14 full-time
- > 9 part-time
- > 11 online

Johns Hopkins University

## Data Science

A Specialization on Coursera: Your Pathway to Expertise

Final Capstone Project created with:





# Key Aspects of Program

- Curriculum designed completely from scratch
- 9 courses (free or \$49 signature track)
- 1 capstone project course w/ industry partnership
- Total signature track cost (modular): \$490
- Each course is four weeks
- Every course runs every month
- Quizzes, in video quizzes, programming assignments and peer assessment projects
- All content open source with permissive license on GitHub

# Johns Hopkins DSS via Roger Peng

- Total Time Running: **13 months**
- Avg. Monthly Enrollment: **182,507**
- Avg. Monthly SigTrack: **12,771** (7%)
- Overall Course Completion Rate: **6%**
- Signature Track Course Completion Rate: **67%**
- Capstone Enrollment: **663** (10/2014), **1041** (3/2015)

# Johns Hopkins DSS via Roger Peng

# 1158

Data Science  
Specialization  
completers  
(first 13 months)

Statistics Master's Degrees		2011	2012	2013	2003-2013
1	Columbia University in the City of New York	242	288	294	1943
2	Rutgers University-New Brunswick	47	62	79	576
3	Ohio State University-Main Campus	45	43	25	486
4	Stanford University	39	30	54	414
5	University of Michigan-Ann Arbor	44	47	55	407
6	University of Illinois at Urbana-Champaign	46	36	61	373
7	California State University-East Bay	49	43	55	354
8	Cornell University	35	51	54	346
9	Michigan State University	44	36	25	341
10	North Carolina State University at Raleigh	29	38	28	329

## Data Science: The End of Statistics? by Larry Wasserman

<https://normaldeviate.wordpress.com/2013/04/13/data-science-the-end-of-statistics/>

## Data science: how is it different to statistics? by Hadley Wickham

<http://bulletin.imstat.org/2014/09/data-science-how-is-it-different-to-statistics%E2%80%89/>

## Data Science, Big Data and Statistics — can we all live together? by Terry Speed

<http://www.chalmers.se/en/areas-of-advance/ict/calendar/Pages/Terry-Speed.aspx>



**pick zero or one:**

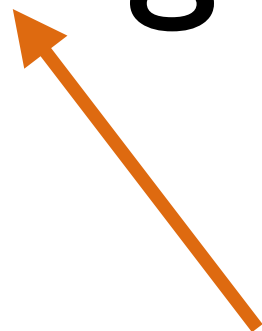
**data science is 'just' statistics**

**data wrangling is not statistics**

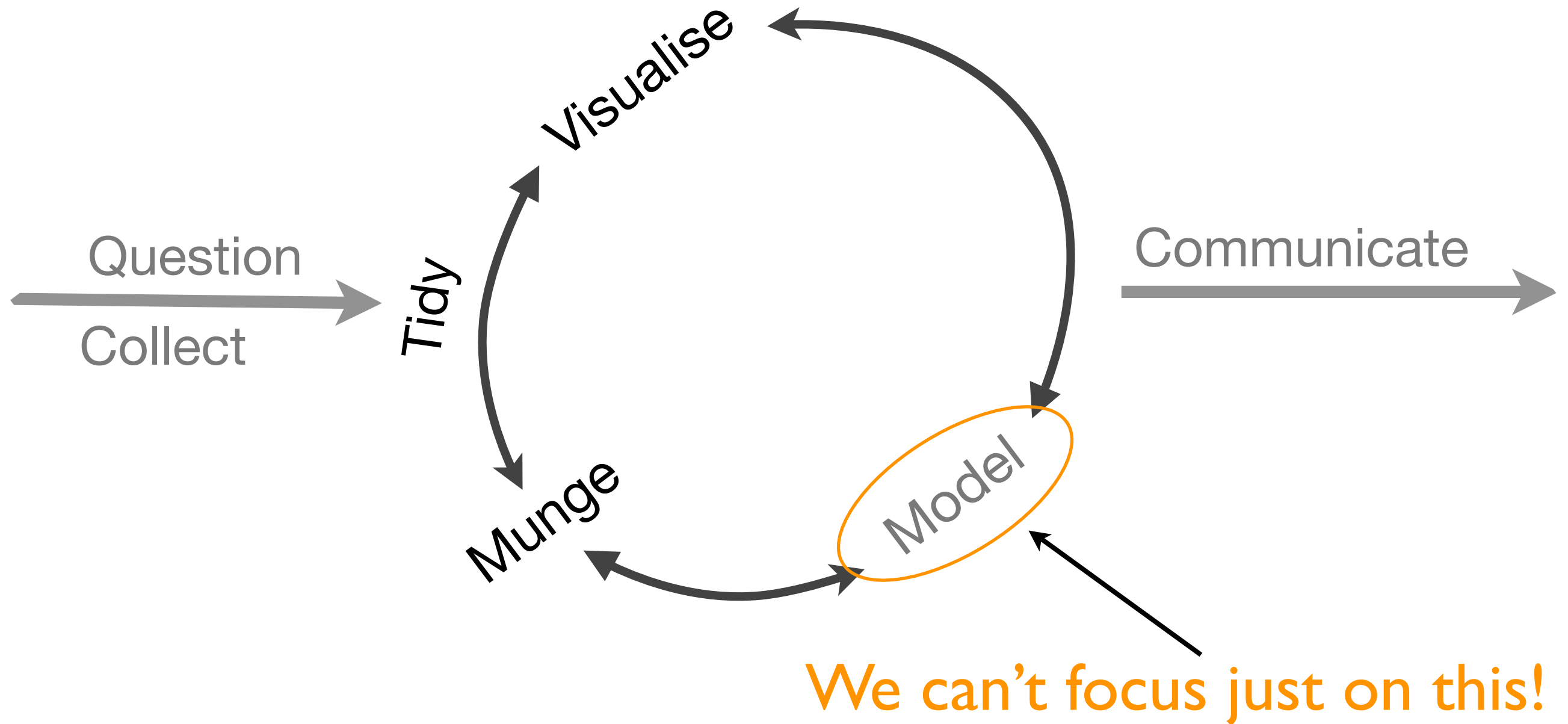
pick zero or one:

data science is 'just' statistics

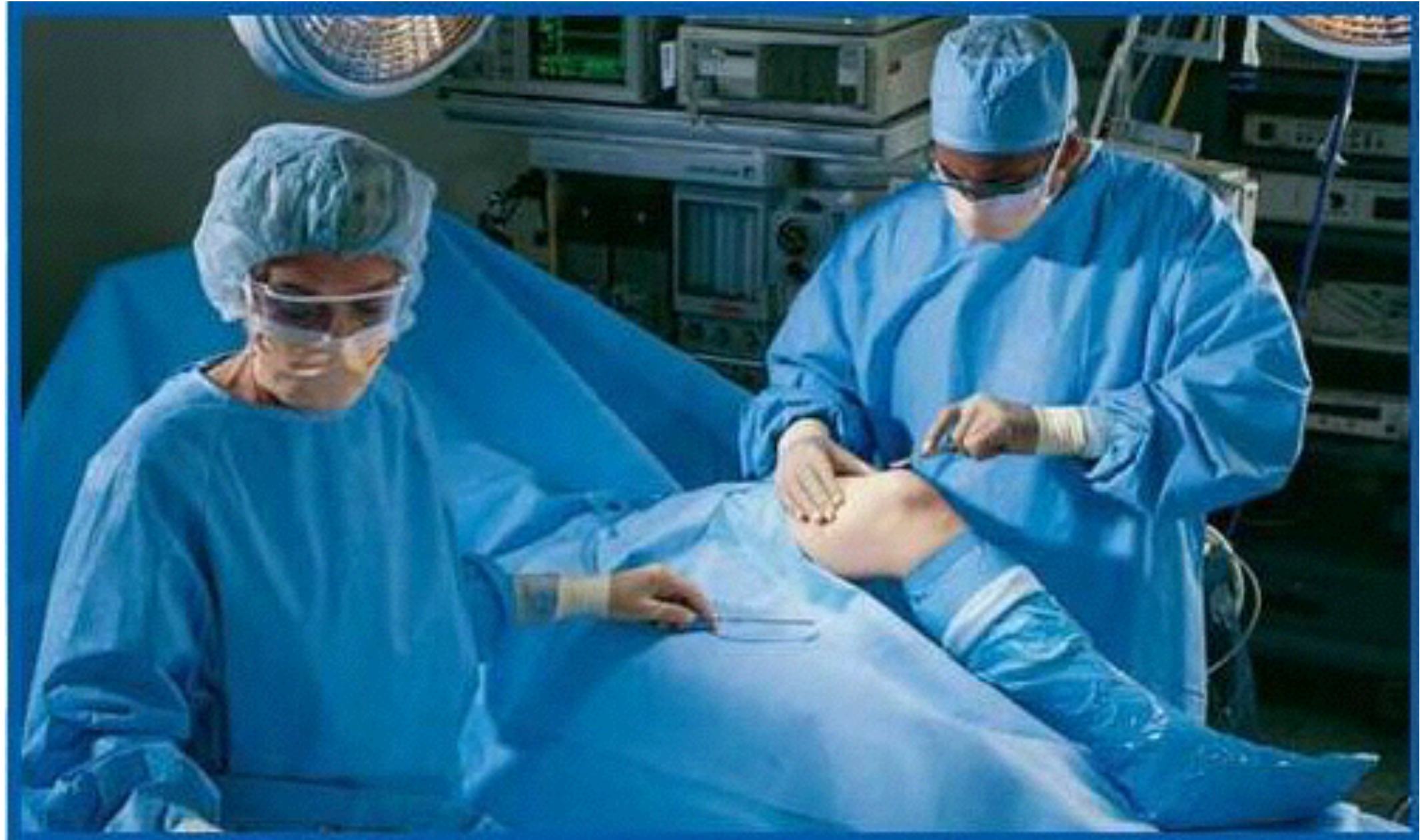
~~data wrangling~~ is not statistics



placeholder for a whole  
slew of things



No one is going to prepare your data for you.





# How STAT 545 projects go sideways: An Incomplete List

inability to

- ... scrape data off the web

- ... get data from an API

- ... parse JSON or XML

utter defeat by date times

text encoding fiascos

ineptitude with regular expressions

R scripts that consume infinite time and RAM

software installation gong shows

We cannot expect anyone to  
know anything we didn't  
teach them ourselves.

Sarah Bryce

We cannot teach anyone  
something if we don't (sort of)  
know it ourselves.

Me

Related: I love my TAs.





# STAT 545 now

~~permission~~ requirement to invest time in setting up tools and to develop proficiency

“simple” descriptive stats  
exploration through visualization

tame data from “the wild”

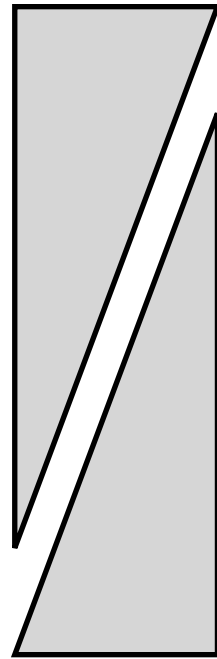
alpha to omega: raw data to a web page/app

readiness for open science and automation

STAT 545 = 1 semester, 3 contact hours/wk

R markdown

Git(Hub)



Data wrangling, cleaning, munging



Visualization

8 weeks

(R chops, in general)



Automation & pipelines



R packages

4 weeks



Shiny



Web APIs and scraping

**some conversation starters ...**

MOOCs and weekend bootcamps are great

BUT I have concerns about all this stuff living outside the regular academic envelope

Do we signal it isn't that important?

What are career implications for those who embrace?

Are we in denial about the need to make room for this in our regular programs?

To a very great degree, daily work by other people sounds easy -- certainly easier than what we have to do.

Gretchen Rubin



Don't study artifact, study nature.

Consider: Behind every wildly successful tool there's probably a very powerful abstraction.

jenny@stat.ubc.ca

http://stat545-ubc.github.io

http://www.stat.ubc.ca/~jenny/



@JennyBryan



@STAT545



@jennybc