
JSM2015
Seattle
Jennifer
Jennifer F. Bryan
Univ. of British Colombia
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

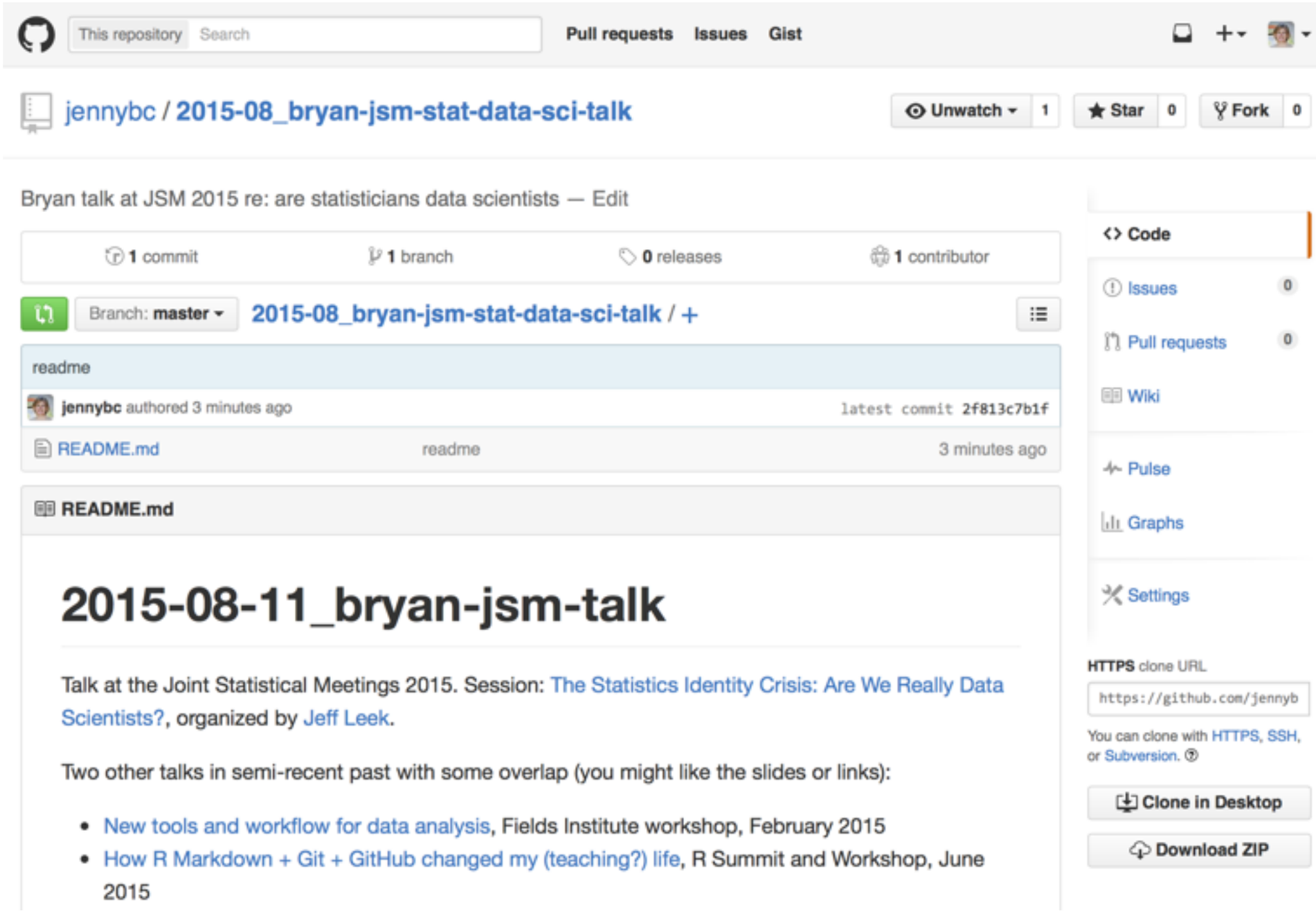
Joint Statistical Meetings 2015, Seattle, WA

Jennifer (Jenny) Bryan
Dept. of Statistics & Michael Smith Laboratories, UBC

jenny@stat.ubc.ca
http://stat545-ubc.github.io
http://www.stat.ubc.ca/~jenny/

@JennyBryan
@STAT545
@jennybc

links, files, etc. available here



https://github.com/jennybc/2015-08_bryan-jsm-stat-data-sci-talk

in a wide array of academic fields,

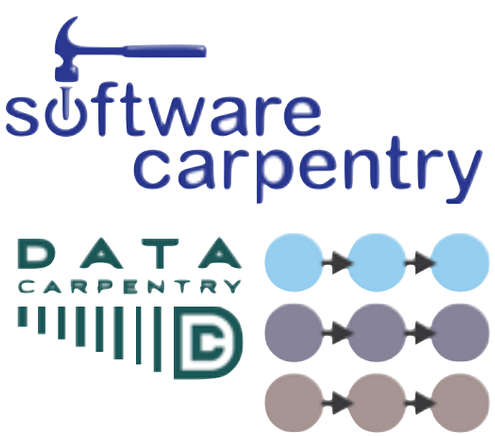
the **ability to effectively process data**

is superseding other more classical modes of research

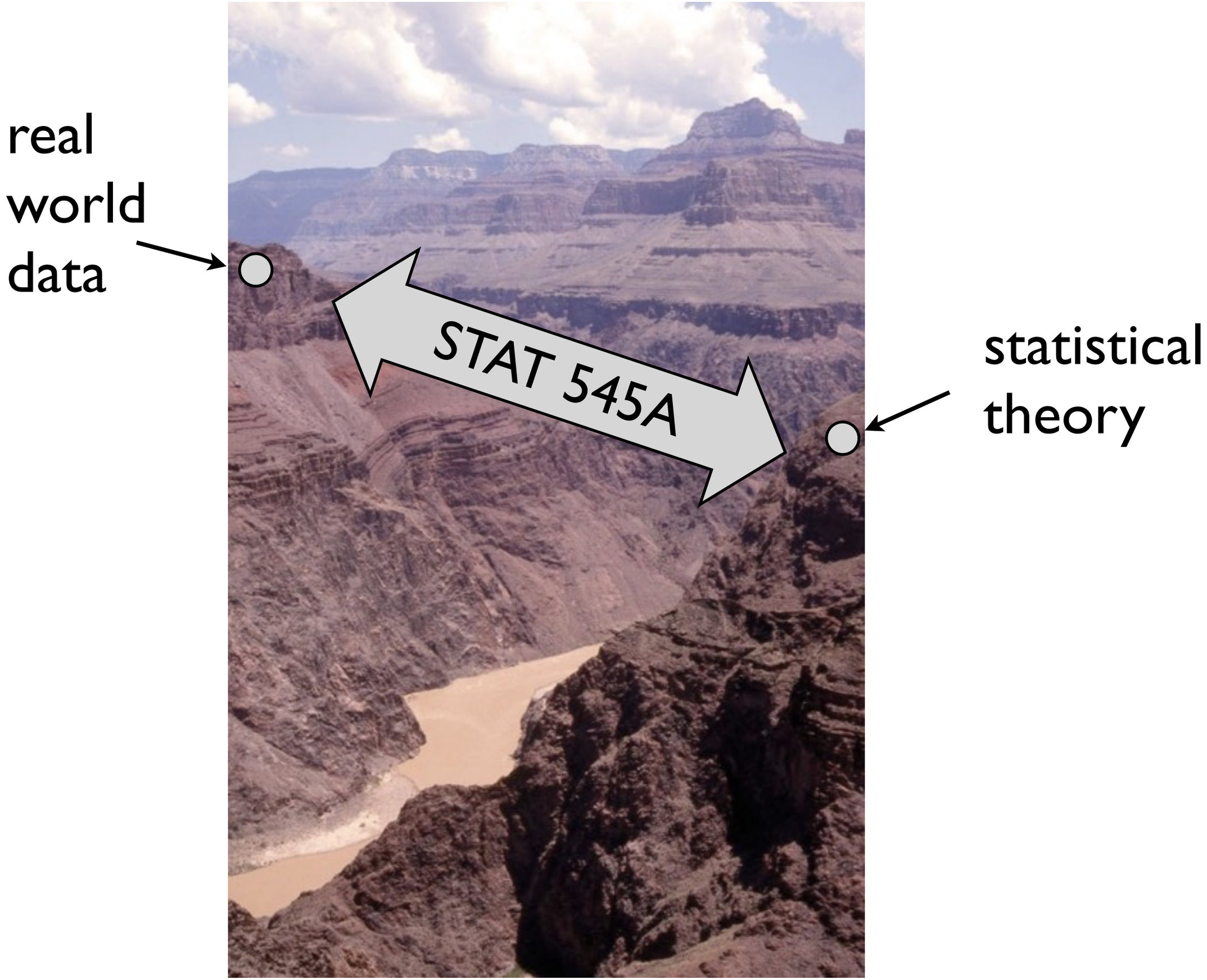
The Big Data Brain Drain: Why Science is in Trouble
http://jakevdp.github.io/blog/2013/10/26/big-data-brain-drain/

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
STAT 545A
statistical
theory

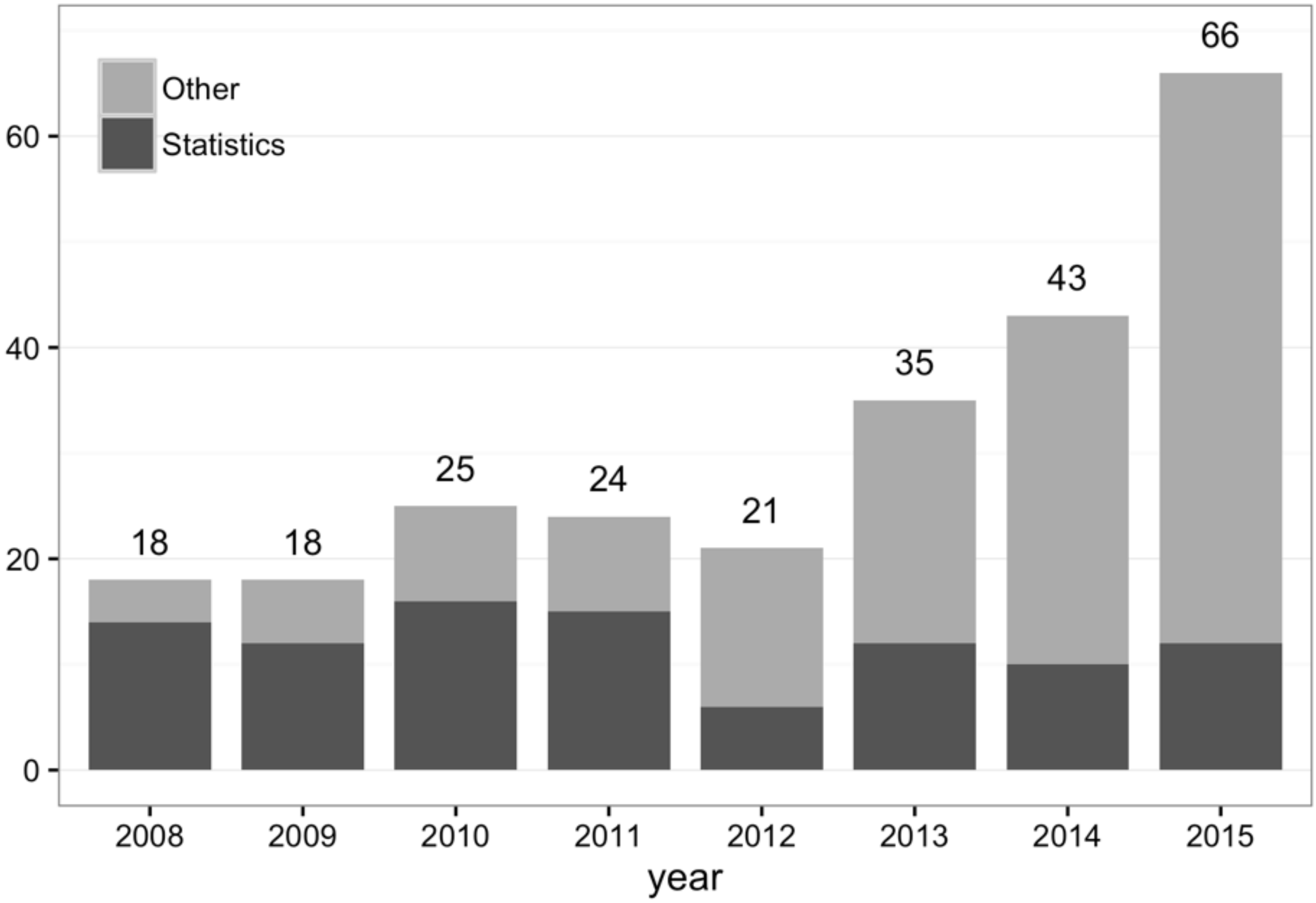
Data wrangling, exploration, and analysis with R

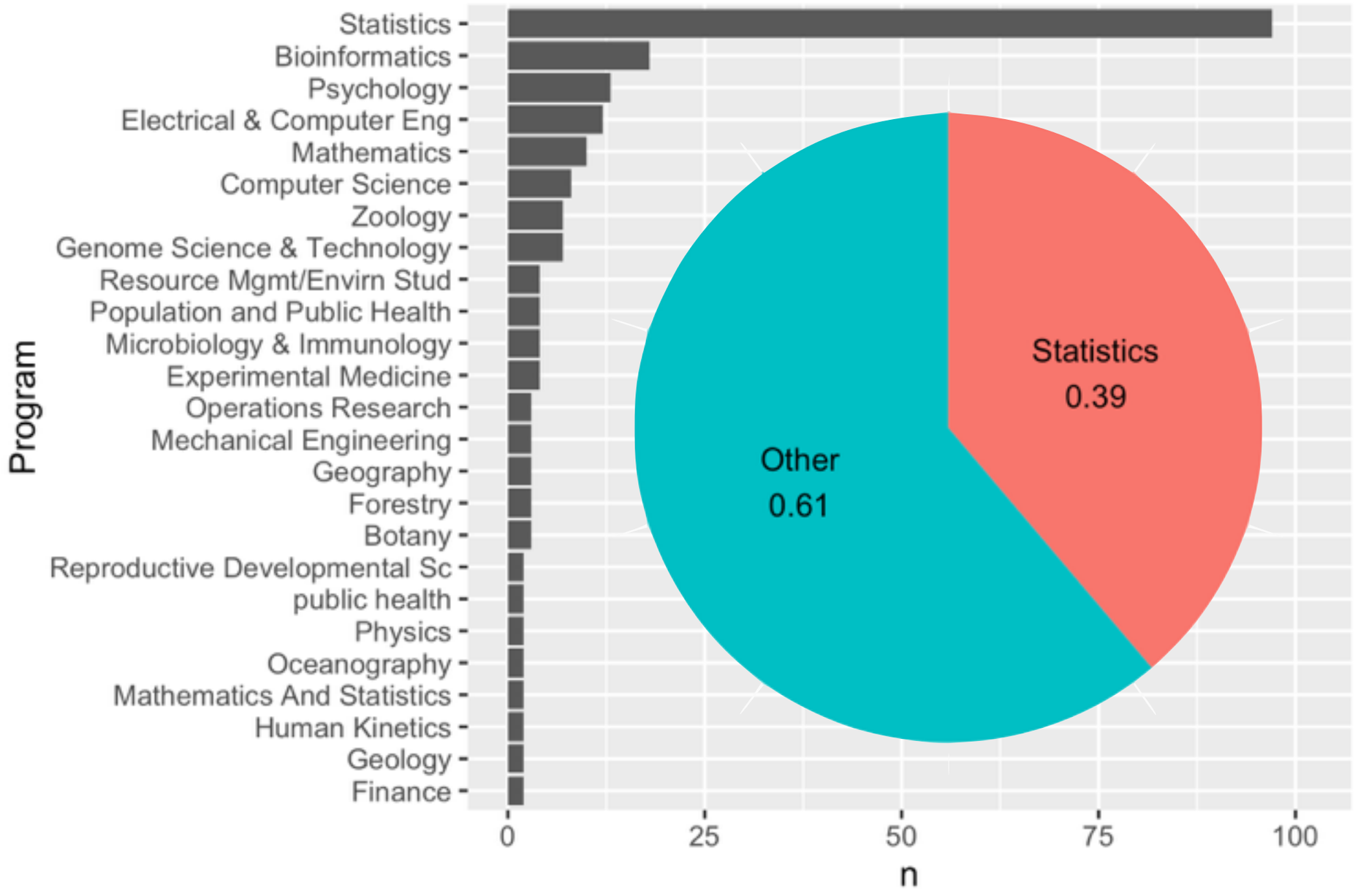
UBC STAT 545A and 547M

Learn how to

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

http://stat545-ubc.github.io


Other
Statistics
60
40
20
0
18
18
25
24
21
35
43
66
2008
2009
2010
2011
2012
2013
2014
2015
year

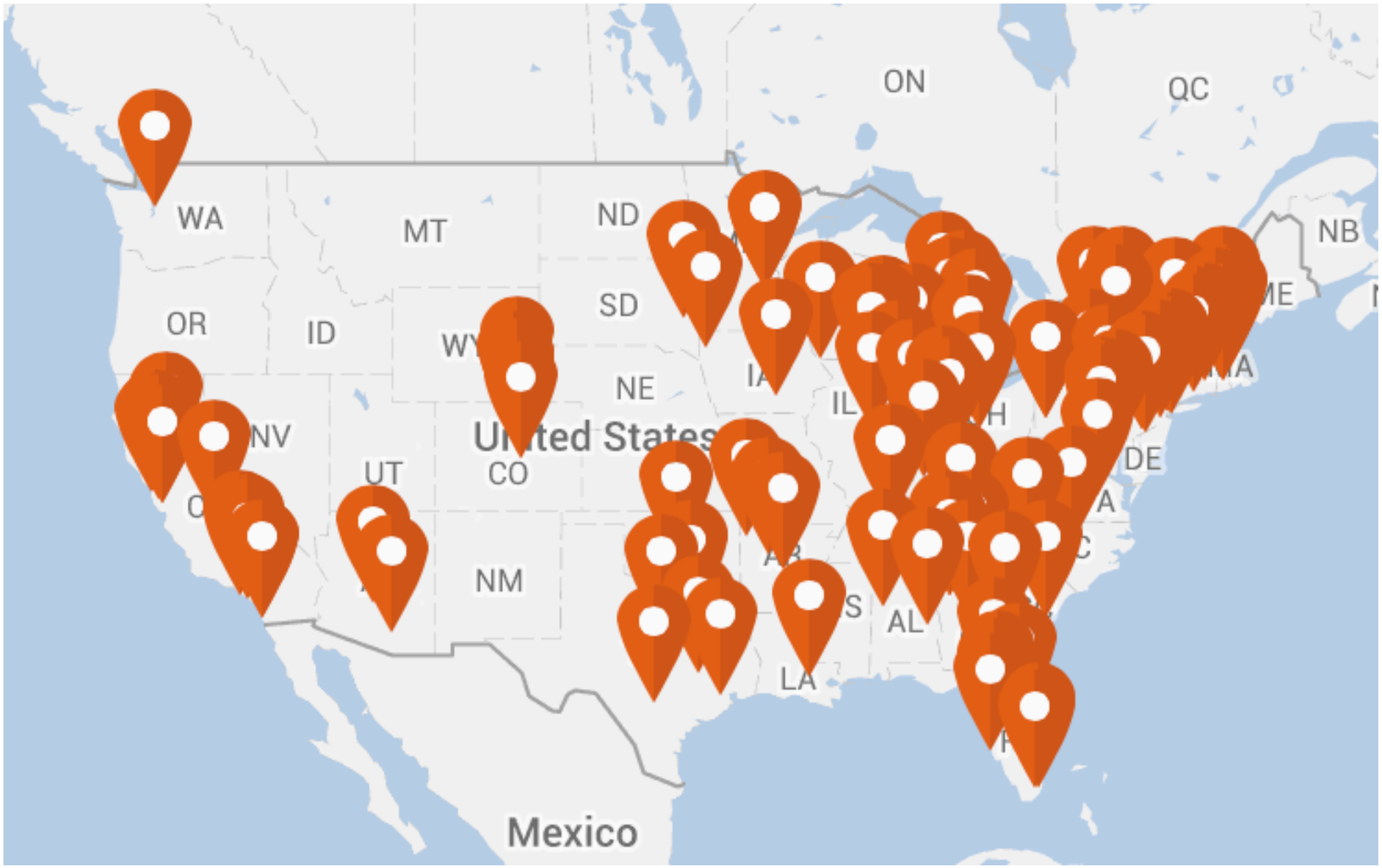


250 = cumulative enrollment 2008 - 2015

54 = # distinct programs sending students

25 = # 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:

SwiftKey.

Data Science for Social Good | People | Projects | Events | Get Involved | Blog | Media

The Eric & Wendy Schmidt

Data Science for Social Good

Summer Fellowship

THE UNIVERSITY OF CHICAGO

The Data Science for Social Good Summer Fellowship

Johns Hopkins DSS via Roger Peng

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

http://community.amstat.org/blogs/steve-pierson/2014/02/09/largest-graduate-programs-in-statistics

50 years of Data Science
by David Donoho
https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

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

Tukey, 1962

… as I have watched mathematical statistics evolve,
I have had cause to wonder and to doubt....

I have come to feel that my central interest is in
data analysis…

Chambers, 1993

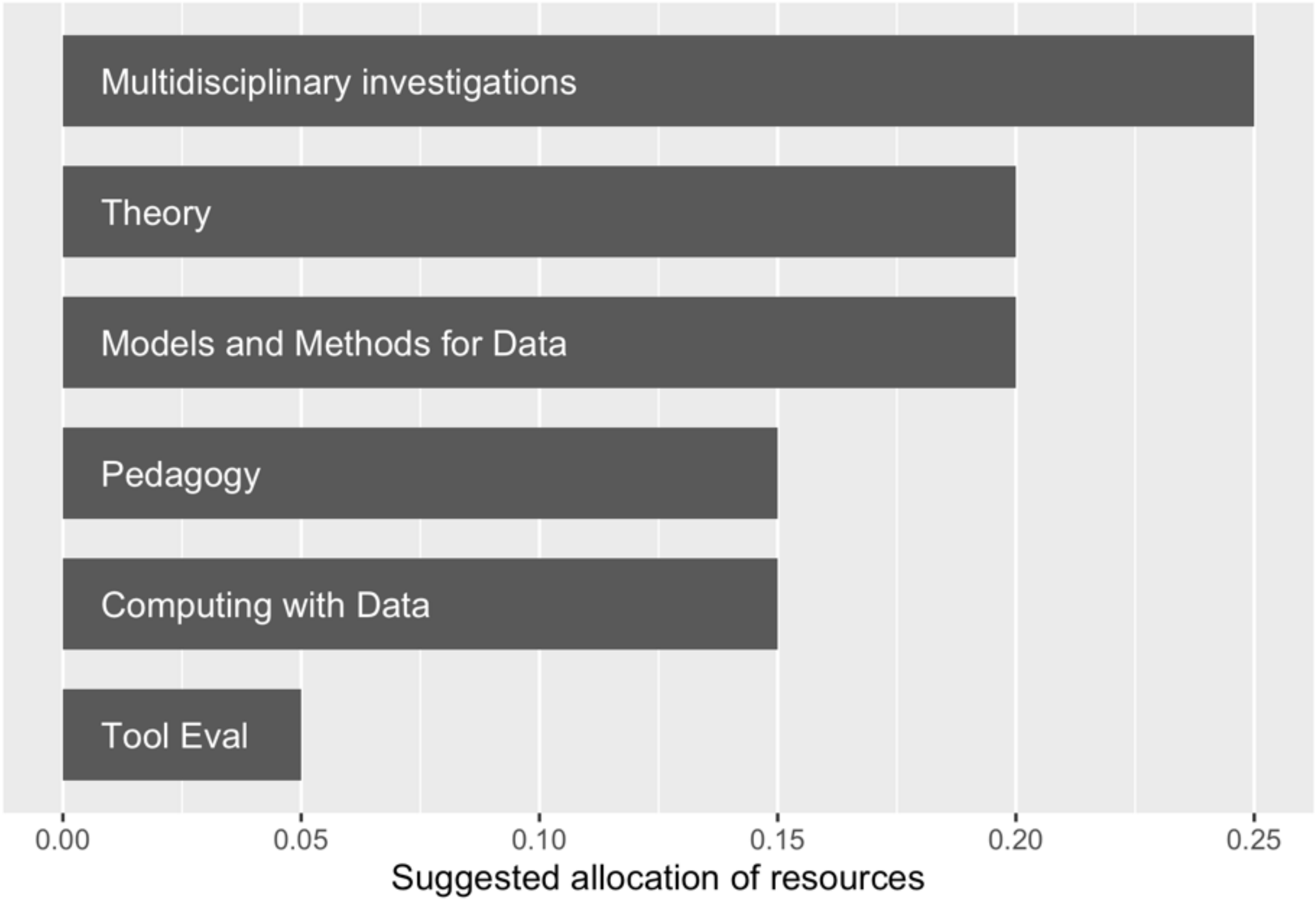
The statistics profession faces a choice:

- traditional topics – data analysis supported by mathematical statistics
- a broader viewpoint – based on an inclusive concept of learning from data

The latter course presents severe challenges as well as exciting opportunities.

The former risks seeing statistics become increasingly marginal.

Cleveland, 2001



Donoho, 2015

Greater Data Science

- Data Exploration and Preparation
- Data Representation and Transformation
- Computing with Data
- Data Modeling
- Data Visualization and Presentation
- Science about Data Science

Full recognition of the scope of GDS would require … major shifts in teaching.

pick zero or one:

data science is ‘just’ statistics

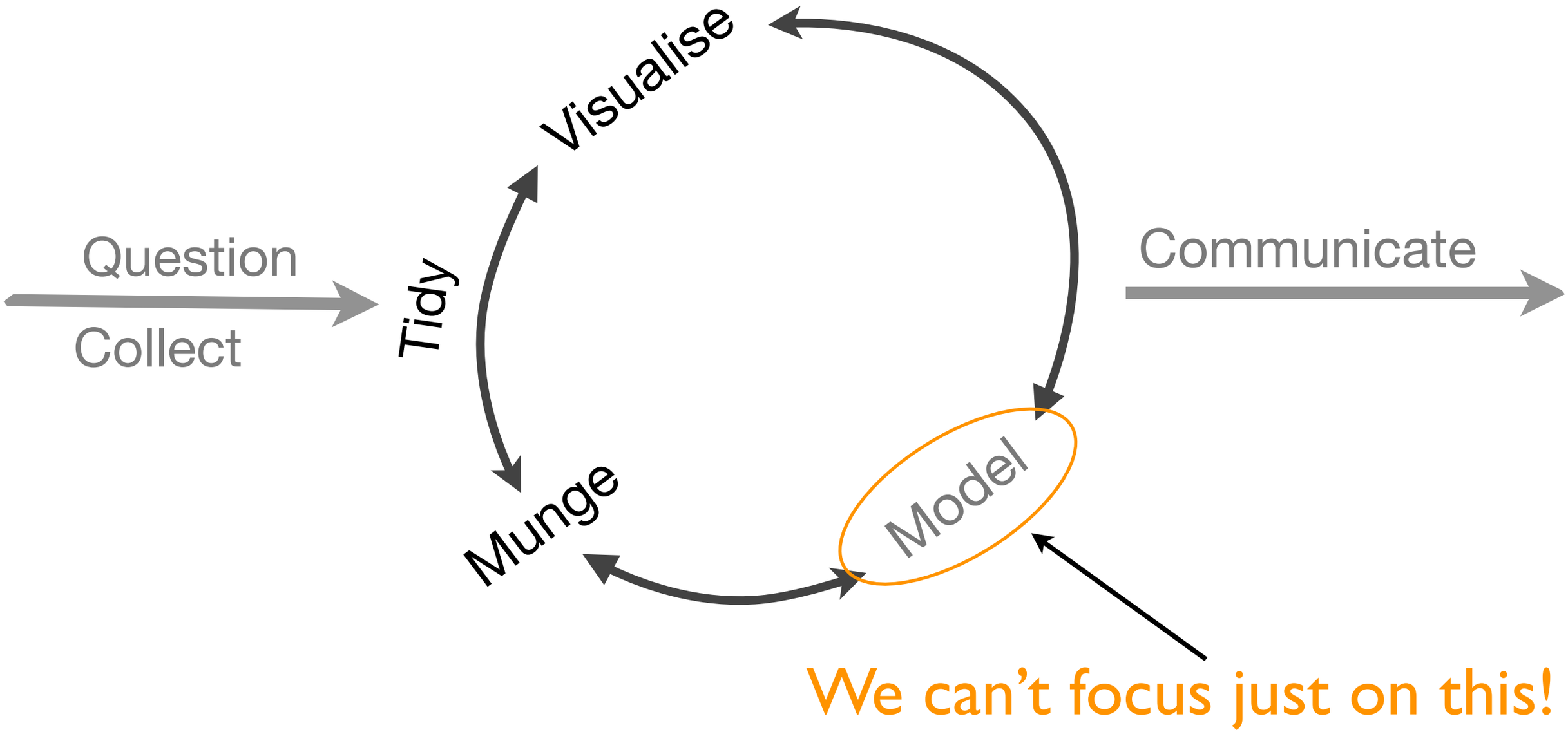
data wrangling is not statistics

pick zero or one:

data science is ‘just’ statistics

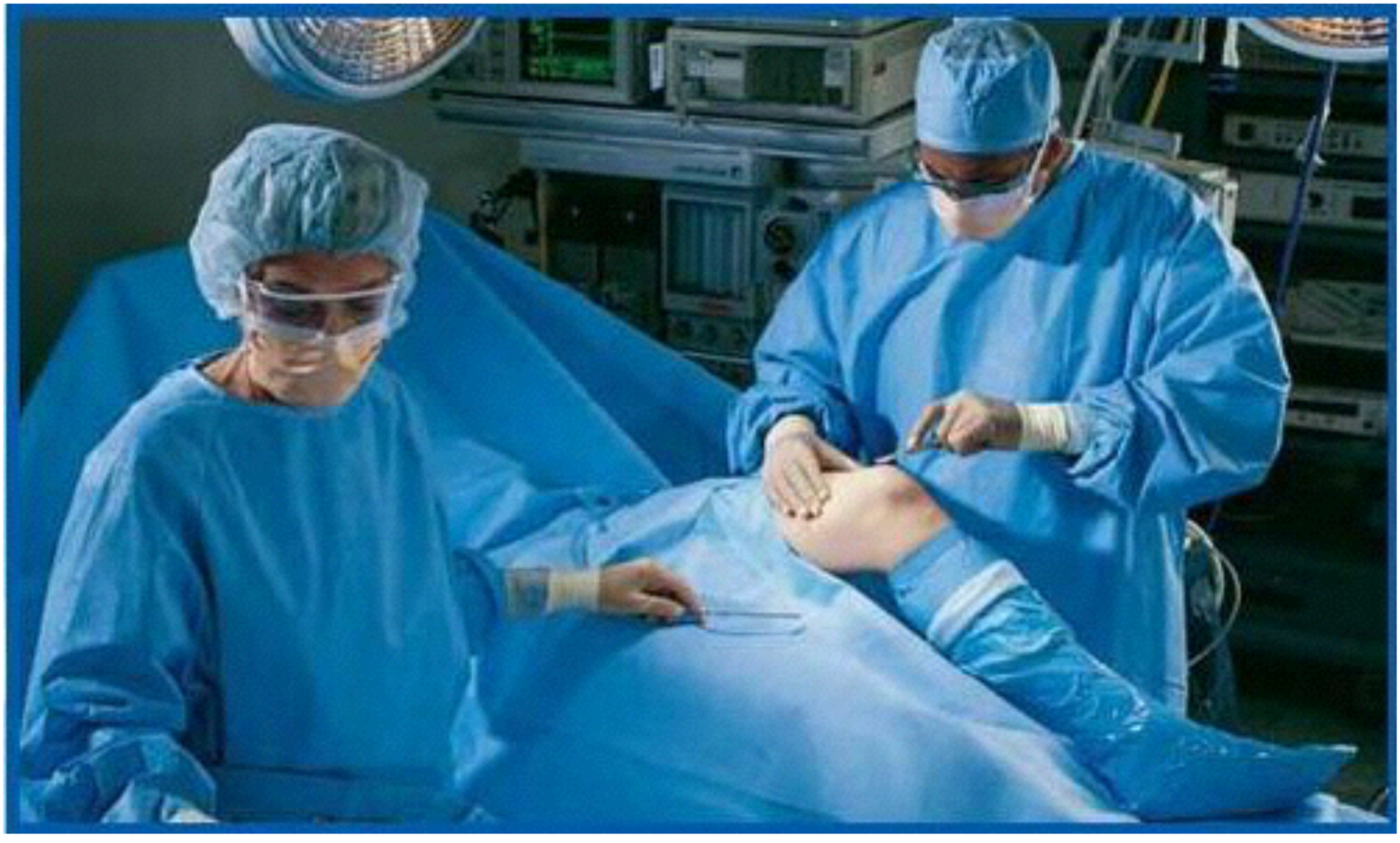
~~data wrangling~~ is not statistics

placeholder for a whole slew of things



Slides from Hadley Wickham's talk in the Simply Statistics Unconference http://t.co/D931Og8mq3

No one is going to prepare your data for you.



http://www.aryamanpharma.com/wp-content/uploads/2014/02/571.jpg

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, including the web + APIs

readiness for open science and automation

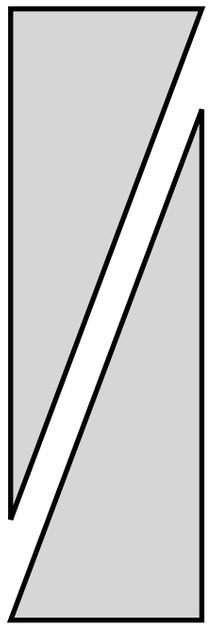
create an R package

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

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 that 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.

Don’t over-study mathematical complexity while under-solving real world complexity.

jenny@stat.ubc.ca
http://stat545-ubc.github.io
http://www.stat.ubc.ca/~jenny/

@JennyBryan
@STAT545
@jennybc