We are all data scientists

Rob Gould
rgould@stat.ucla.edu

Or: data literacy in the age of Big Data

Rob Gould
rgould@stat.ucla.edu

Outline

- Why do we need a framework like “data literacy?”
- Can’t we be satisfied with statistical literacy and statistical thinking?
- What is data literacy?
- That sounds impossible.
- Here are some projects that are trying…
- What are the next steps?

Why do we need data literacy?

Three Parables

- k-anonymity
- Pizza Girl
- Amazon Prime

Data Literacy Quiz

- What percent of the U.S. population can be uniquely identified if I know
  - zipcode
  - gender
  - birthday?

https://dataprivacylab.org/

From Latanya Sweeney Data Privacy Lab

87%

(in 2010)

https://aboutmyinfo.org/index.html

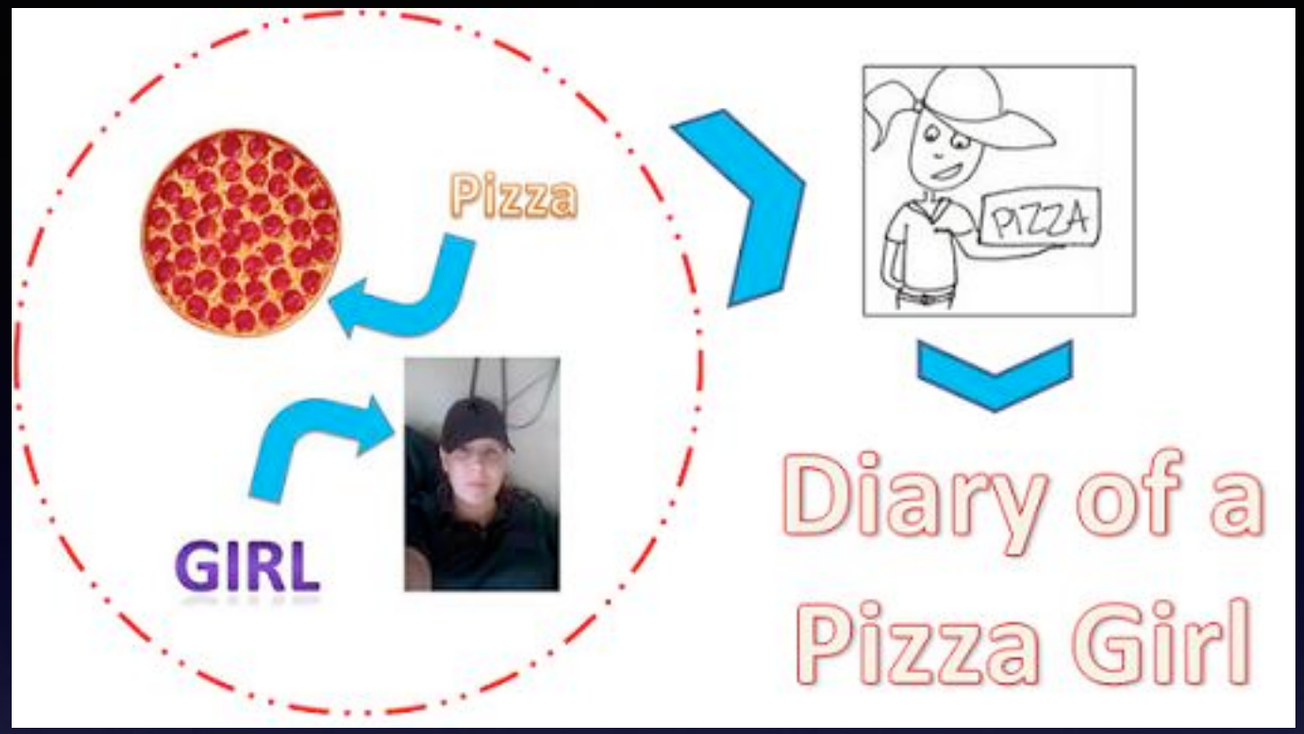
How Unique are You?

90034 (pop. 57964)
Male

Birthdate 3/7/1965 **Easily identifiable by birthdate (about 1)**

Birth Year 1965 **Lots with your birth year (about 351)**

Range 1965 to 1969 **Wow! There are lots of people in your age range (about 1759)**



Pizza girl delivers pizzas in Austin, TX.

She keeps a blog.

And she collects data

5:28: arrive
5:31: clock in
5:33–5:45: fold boxes (12 minutes)
5:45–5:48: work ovens (3 minutes)
5:48: routed
5:50: leave on delivery (11 minutes travel time)
6:01: arrive at delivery ($3 tip)
6:01: leave house (11 minutes travel time)
6:12: Arrive at store (24 minutes run time for 1 delivery)
6:12–6:15: work ovens (3 minutes)
6:15: routed
6:18: leave on delivery (15 minutes travel time)
6:33: arrive at house ($3 tip)
6:35: leave house (3 minutes travel time)
6:38: arrive at second delivery ($2 tip)
6:38: leave house (14 minutes travel time)
6:52: arrive at store (37 minutes run time on two deliveries)
6:52–7:02: work ovens (10 minutes)

How is my time most profitably spent?
Does method of payment affect my tip?

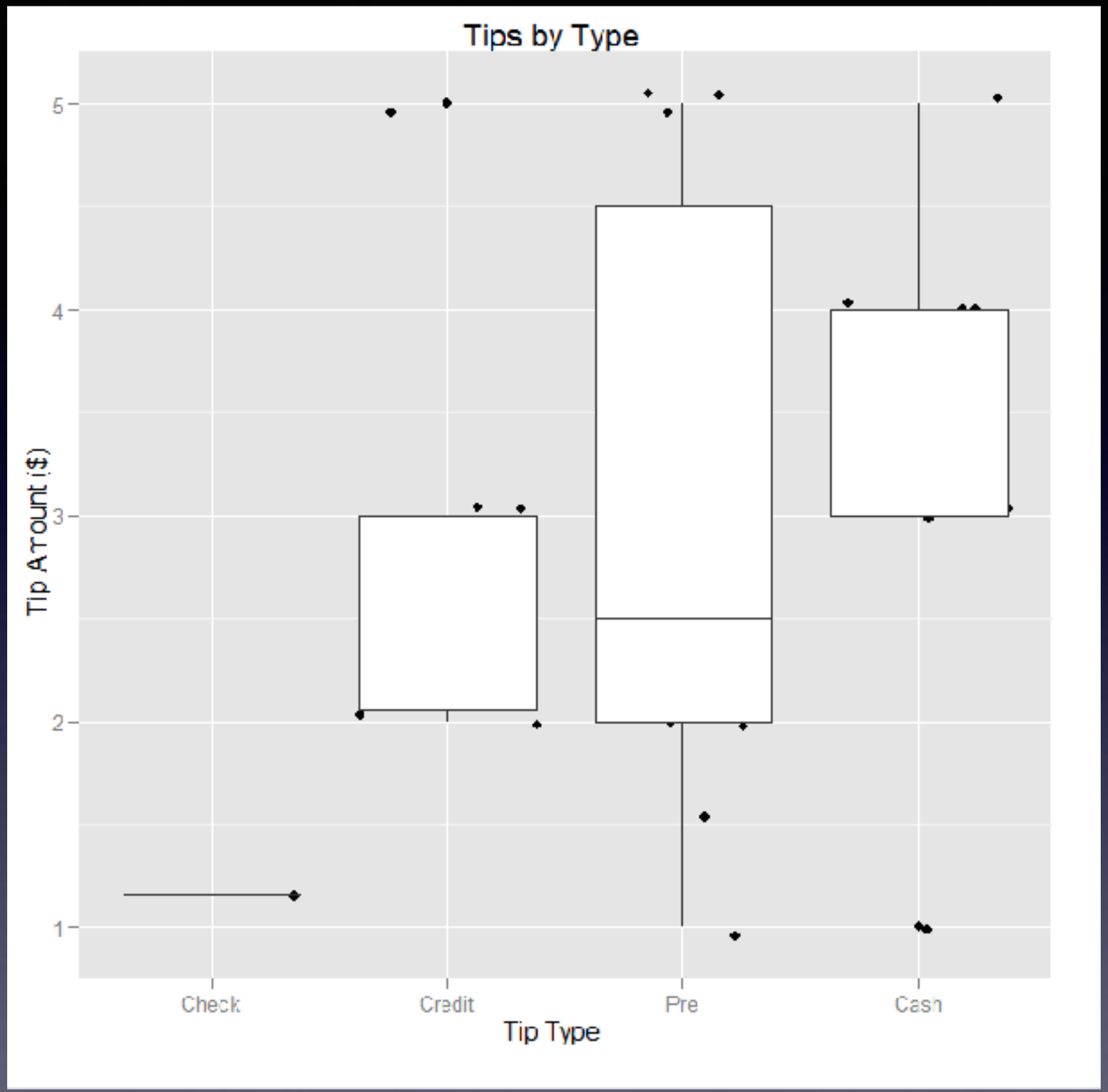
"Pizza Girl: A Statistical Analysis of a Delivery Shift: Part 1", http://slice.seriouseats.com

Jared P. Lander

wrote his MS thesis on

"New York Pizza: How to Find the Best" (Dec, 2008)

data from pizza blogs and menu.blog



http://www.jaredlander.com/

amazon story

- When Amazon rolled out same-day service to Amazon Prime members, they used an “unbiased” algorithm that included such things as the percent of Prime members in a neighborhood, location to distribution centers, and other neutral factors.

- But Bloomberg (2016) reported that in six major cities the neighborhoods that excluded same-day delivery were predominantly African-American. Later observers noted that the excluded areas very closely followed historic “redlining” boundaries.

- Amazon corrected this “unbiased” algorithm by extending delivery to all neighborhoods.

From NAS webinar on data science and ethics: brittany fiore-gartland, U of Washington. (http://sites.nationalacademies.org/cs/groups/cstbsi
documents/webpage/cstb_182281.pdf

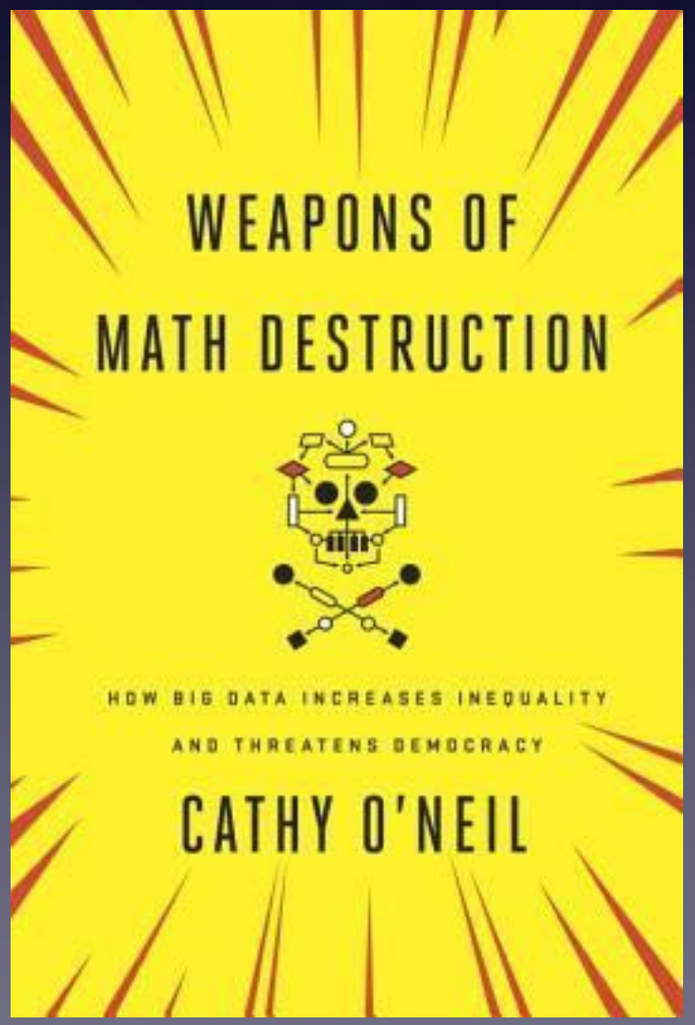
https://www.bloomberg.com/graphics/2016-amazon-same-day/

"The Amazon algorithm operates off of an inherited cartography of previous redlining efforts, which created pockets of discrimination, the consequence being that the discrimination continues to be reproduced," said Jovan Scott Lewis, a professo at the University of California, Berkeley's Haas Institute for a Fair and Inclusive Society.—

https://www.usatoday.com/story/tech/news/2016/04/22/amazon-same-day-delivery-less-likely-black-areas-report-says/83345684/

The morals

- k-anonymity reminds us that Big Data can undermine privacy.
- The Pizza Girl story reminds us that data can be empowering if we know when to collect it, how to share it, how to interpret it.
- The Amazon story reminds us that algorithms can propagate patterns of inequity



Can’t we be satisfied with statistical literacy and statistical thinking?

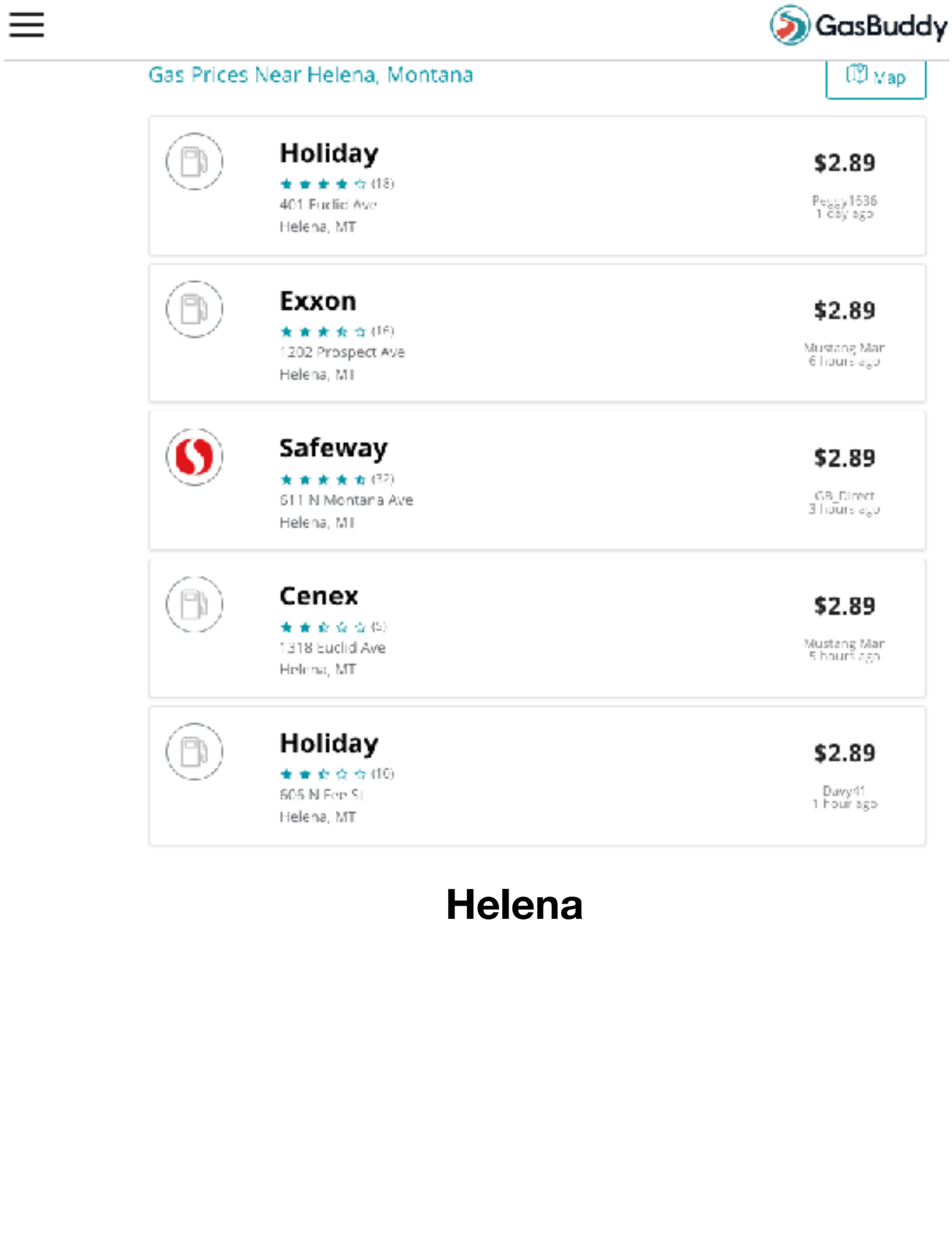
History of intro stats education

- Dark Ages
- Statistical Literacy
    - What do you need to know to read the newspaper?
- Statistical Thinking
    - How do I ask questions, consider data, analyze, interpret? (The “statistical investigation process”)
    - The Real Data movement
- Which brings us to…

The Age of Ubiquitous Data

- Ask a question, somewhere on the internet, you will find data that can (potentially) answer it.

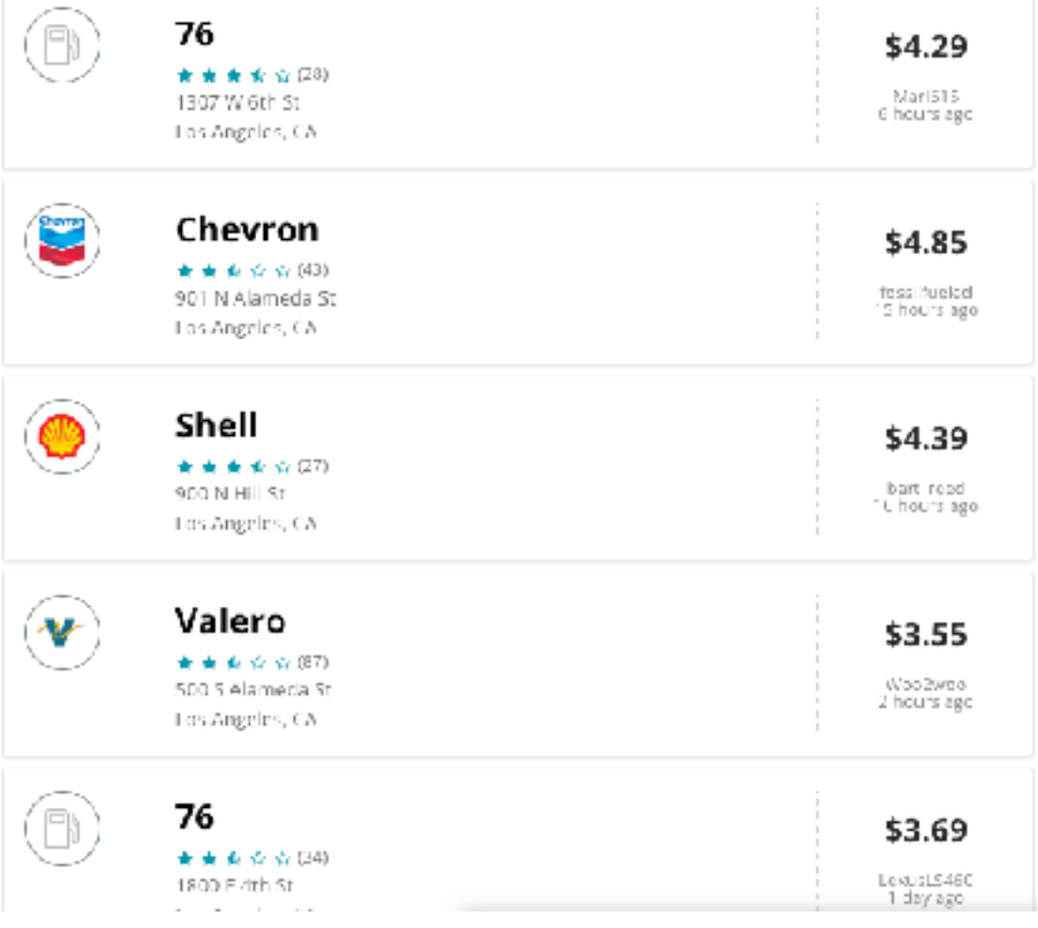
- How do gas prices in Helena compare with those in Los Angeles?



Helena

You expect me to do statistical thinking with this? How?

Los Angeles



Does standing for a few minutes increase pulse rates?

Table 24: Pulse Data

	Pulse	Group	Category
1	62	1	sit
2	60	1	sit
3	72	1	sit
4	56	1	sit
5	80	1	sit
6	58	1	sit
7	60	1	sit
8	54	1	sit
9	58	2	stand
10	61	2	stand
11	60	2	stand
12	73	2	stand
13	62	2	stand
14	72	2	stand
15	82	2	stand

estimate the difference in mean pulse between those

Table 25: Pulse Data in Matched Pairs

Pulse data: matched pairs

=	MPSit	MPStand	Difference
1	68	74	6
2	56	55	-1
3	60	72	12
4	62	64	2
5	56	64	8
6	60	59	-1
7	58	68	10

information that could affect results. It may be l
ter to *block* on a variable related to pulse. Since pec
have different resting pulse rates, the students in
experiment were blocked by resting pulse rate by p
ing the two students with the lowest resting pulse ra
then the two next lowest, and so on. One persor
each pair was randomly assigned to sit and the othe
stand. The matched pairs data are in Table 25. As in
completely randomized design, the mean difference

“The data, arranged by treatment, are in Table 24”.

“The matched pairs data are in Table 25”

Data Assignment

- Use one of the university library's data repositories to download a dataset that you find interesting. You must demonstrate that you can upload the dataset into StatCrunch or Fathom.

- (at least 2000 observations, 5 variables, at least 2 numerical variables)

Students' Questions

- I uploaded a date, but a strange "random" number was uploaded instead. (data formats: dates, strings, characters, floating)

- Are observations people, or records? (hierarchical structures)

- I got several files zipped together. Which one is the data? (file extensions, managing files)

- Which document do I download? The documentation says it provides only SAS/SPSS/Stata/ASCII, but not 'csv". (file extensions, managing files)

- What to do with fixed format? (data storage)

- I click "download all files" but nothing opens up. (file management)

- I can't upload a .tsv file. (file extensions)

- I need step-by-step instructions. (?)

What More Do We Need?

- So the notion of statistical literacy needs to be expanded to allow students to access, manage, handle the data that they will be statistically thinking about and

- account for living in a menacing world of Big Data and

- adjust for an age in which we have a notion of “our data”, not just data that belongs to scientists and professionals.

What else?

- Ethical considerations. “When you consent to use an app, and give away your rights to your data, you can’t possibly forsee all of the future uses of that data.” (Fiore-gartland, paraphased from her webinar for the NAS)

- The role of algorithms and black-boxes. Companies that target vulnerability, as read from Facebook and Google. (Weapons of Math Destruction)

- Measuring predictive success. What does it mean to make a prediction? What is a good predictive success rate? How are these measured? With what uncertainties?

~~Real Data Movement~~

Data First Movement

aka data literacy

What is data literacy?

I don’t know.

Data Literacy

- In 2011 at USCOTS I made a stab at a definition, based on the need to prepare citizens to live in a world in which there was both plentiful, complex, and rich data as well as powerful and inexpensive analysis tools.

2011

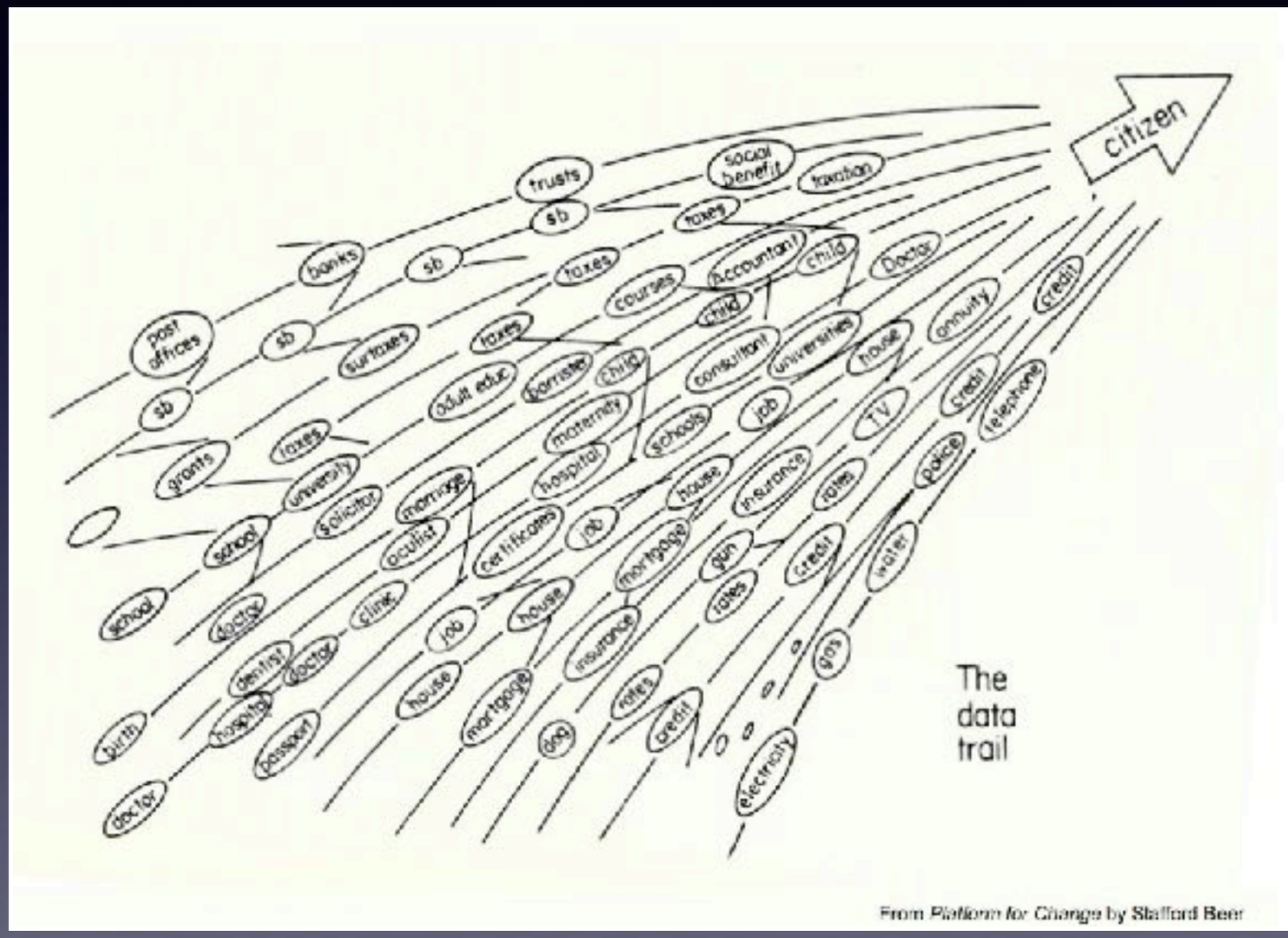
Citizen statisticians

- Participate in formal and informal data gathering
- e.g. OpenMaps, DidYouFeelIt, CENS, twitter

www.openmaps.eu

2011 Citizen Statisticians

- Aware of their own "data trail"
- Aware of the harm that could be done to themselves or others through data aggregation



(From *Platform for Change*, Beer, S., 1975)

2011

Citizen Statisticians

- Can identify opportunities to improve selves through sharing and analyzing data.

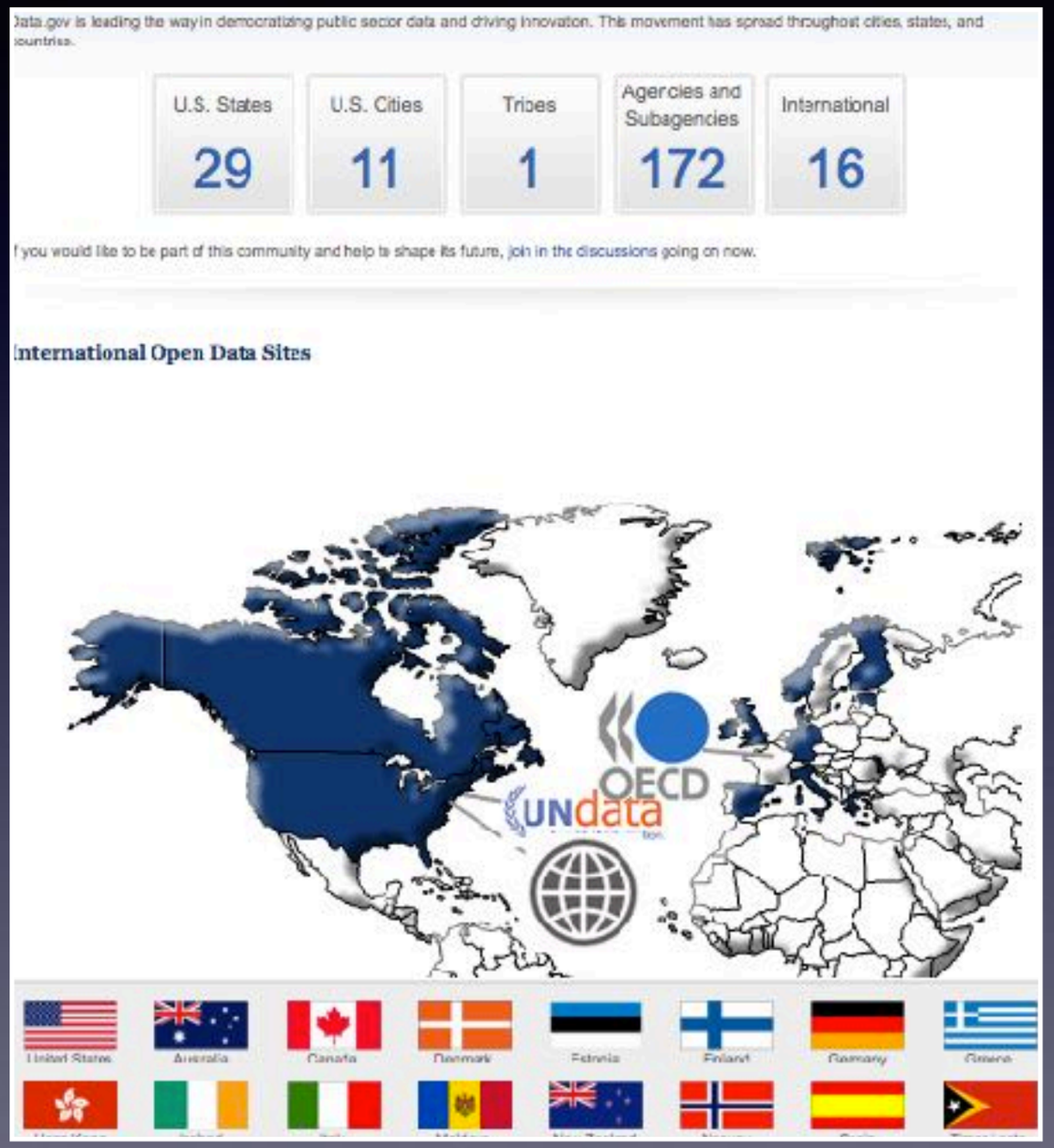
- have the skills and desire to share data with a wide range of people.

Boxes
703 routed
705 out
710 711 $2
721 723 $3
734 in
Cleainin
741 routed
744 in
7:58 $3
812 813 $3
816 817 $3
832 in
Mop
837 route
839 out
845 847 $5
853 in
Phone
Makeline
903 route
905 out
912 915 $2
920 in
922 925 out
932 933 stiffed
942 zero
950 in 953 ccash out

2011

Citizen Statistians

- Know how to find data
- Can think critically about data sources
- Know how to move data from the source to an analysis tool



Data Literacy is…

"Data literacy is the ability to understand, find, collect, interpret, visualize, and support arguments using quantitative and qualitative data" (Deahl, p 41, 2007) MIT Civic Data Design Lab

…the capacity to understand, create and manage [their own] data in meaningful, efficacious and ethical ways. (Bowler, et. al 2017) (Information Sciences)

Oceans of Data Institute

The data literate individual understands, explains and documents the utility and limitations of data by becoming a critical consumer of data, controlling his/her personal data trail, finding meaning and taking action based on data. S/he can identify, collect, evaluate, analyze, interpret, present and protect data. —2015

http://oceansofdata.org/topic/data-literacy

That does not sound like an introductory statistics course

- The good news is that a growing number of people are realizing that we need a K-16 and beyond curriculum in

Some Projects that are Trying

“[tools and resources to] contribute to young people’s ability to understand quantitative evidence about key social phenomena that permeate civic life.” http://www.procivicstat.org/

ProCivicStat

Promoting civic engagement via explorations of evidence

HOME | TEACHING MATERIALS | **OTHER RESOURCES** | PUBLICATIONS | PEOPLE AND EVENTS

Data Sets

We have assembled a collection of resources and associated metadata that can support teaching and learning. These enable tutors to create materials relevant to local circumstances, perhaps using lesson structures, data visualisations and ideas set out in our teaching materials.

Eurostat
Link to news, data and publications from Eurostat
http://ec.europa.eu/eurostat

OECD Data
Find, compare and share the latest OECD data: charts, maps, tables and related publications.
http://data.oecd.org

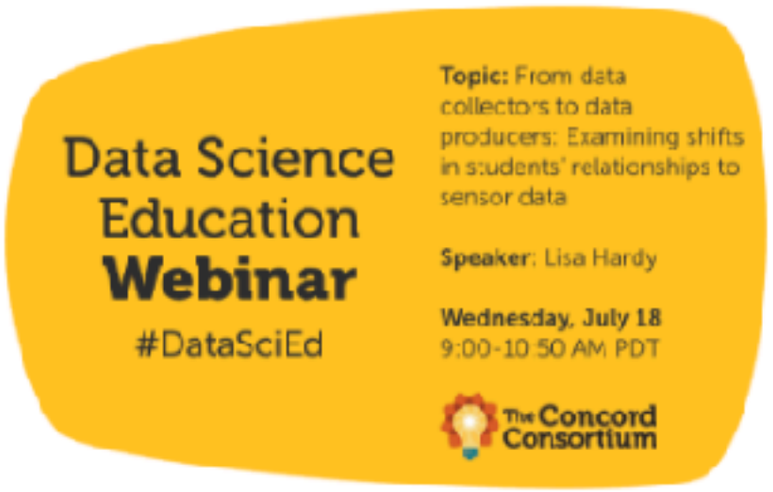
Other Key Sources for International Comparisons
United Nations
World Bank

Data in Gapminder World
A list of all the indicators displayed in Gapminder World

Data Science Education Meetups

Over 100 thought leaders from organizations around the U.S. and four continents generated innovations in technology, and teaching and learning at the Concord Consortium's first Data Science Education Technology conference in February 2017. We continue to gather thought leaders in data science education through a series of webinars and informal meetups at conferences.

Data Science Education Meetups



Speaker: Lisa Hardy
July 18, 9-10:50AM PDT
RSVP Now

Speaker: Andee Rubin
September 19, 9-10:50AM PDT
RSVP Now

Speaker: Jennifer Kahn
October 25, 9-10:50AM PDT
RSVP Now

https://concord.org/meetup/

Secure | https://codap.concord.org

Apps | Outlook email App | HCA stuff | Data sets | List of countries by... | Mobilize | MZ RStudio Sign In | StatCrunchThis | home - Mobilize Wiki | Other Bookmarks

CODAP

About Forums Contact Help Try CODAP

Common Online Data Analysis Platform (CODAP)

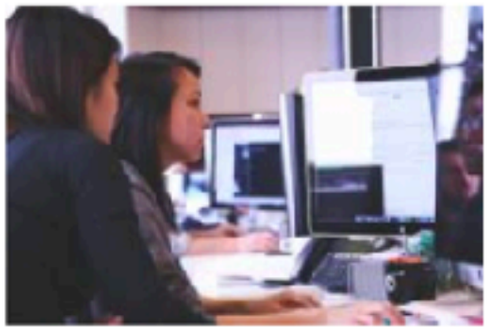
Open-source software for dynamic data exploration

For Educators For Developers

CODAP is a free web-based data tool designed as a platform for developers and as an application for students in grades 6–14.

PROJECTS

DATA ANALYTICS TECHNICIAN ADVANCEMENT (DATA) PROGRAM



Oceans of Data Institute will partner with Columbus State Community College (OH) in a new ATE-NSF funded project: Data Analytics Technician Advancement (DATA) Program. The project will establish a DATA Pathway in the central Ohio region to increase the supply of qualified data analytics technicians. Multiple entry and exit points will offer maximum flexibility to address the shifting needs of industry employers and accommodate the diverse interests and experiences of students of different ages and life stages. ODI will draw upon previous NSF-funded work, including ODI's *Creating Pathways for Big Data Careers,* to facilitate the design of work-based learning activities and supportive materials for students enrolled in DATA. Specifically, ODI will work with Columbus State and its industry...

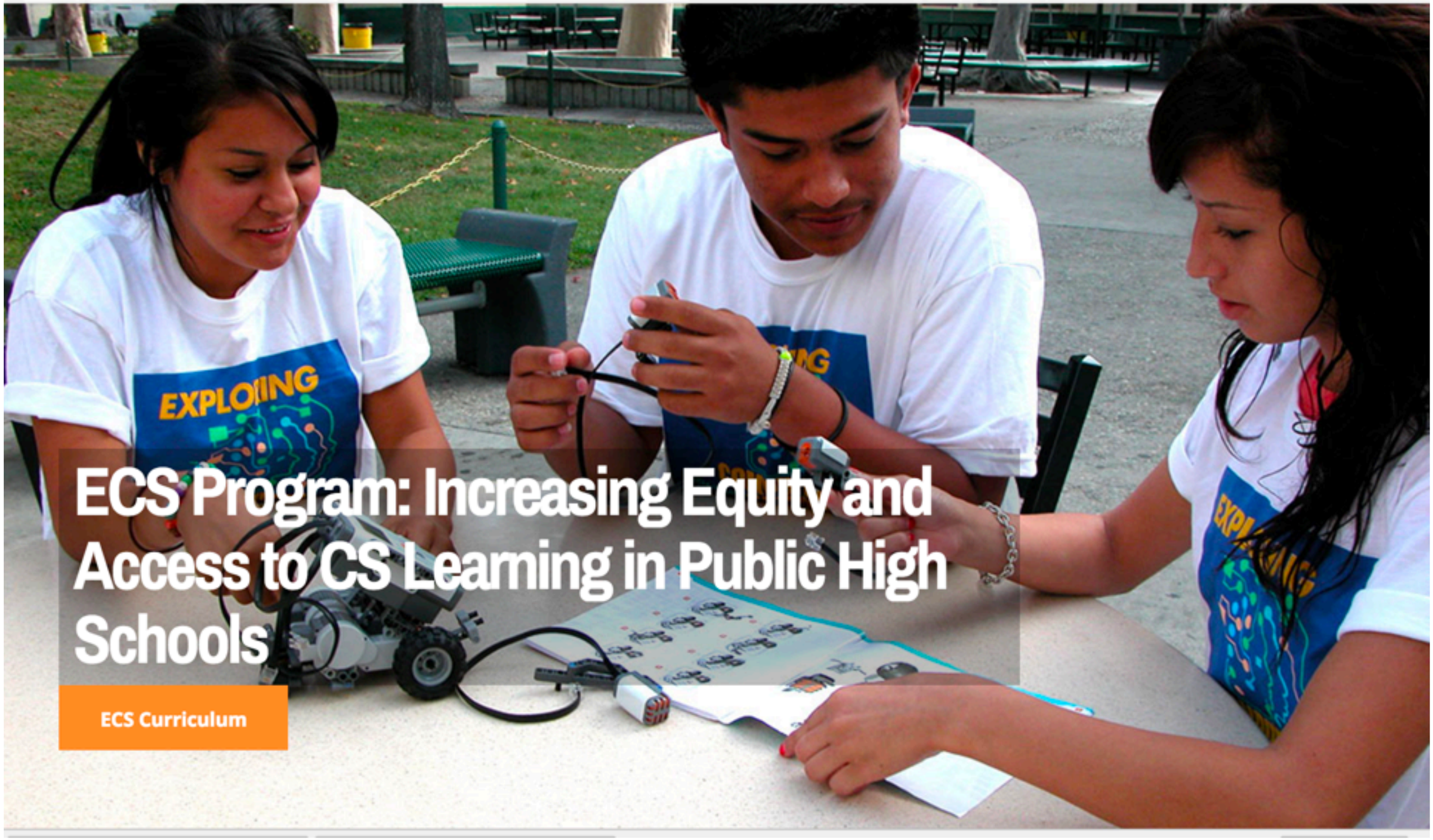
Read more

Two-Year College Data Science Summit

May 10-11, 2018 (awaiting final confirmation), Washington, DC

With funding from the National Science Foundation, this workshop will bring together a diverse group of participants to make recommendations for two-year college data science programs, keeping in mind the needs of each of three student populations:

1. Those seeking employment following an associate's degree
2. Those seeking transfer to four-year programs
3. Those seeking certificate programs and college-level courses in data science for professional development



http://www.exploringcs.org/

Others

- Park City Math Institute Guidelines (2016). DeVeaux et al, https://www.amstat.org/asa/files/pdfs/EDU-DataScienceGuidelines.pdf

- http://www.bootstrapworld.org/: curricular modules K-12

- Chris Wild's Data Science MOOCS for K-12 https://www.mooc-list.com/instructor/chris-wild

- "Data Science Goes to School" Initiative in Germany, funded by Deutsche Telekom Stiftung and led by U. of Paderborn (Rolf Biehler)

- ASA and Royal Statistical Society curriculum (Nicholas Fisher).

- AMATYC sub-committee on Data Science (Brian Kotz)

- Handbook on Research in Statistics Education (2017), Ben-Zvi, et al, Springer-Verlag

IDS

Introduction to Data Science

- *Introduction to Data Science* is a year-long statistics & computing course designed for secondary students who have taken basic Algebra. Typically, IDS is taken by students in their third year of high school.

- California's two public university systems require four years of math for admission and accept IDS as one of those years.

- Currently taught at 20 schools within the Los Angeles Unified School District (LAUSD) and in another 10 schools at 6 local school districts.

http://www.mobilizingcs.org/introduction-to-data-science

Components

- In-class activities with guidelines for teachers to develop conceptual understandings, terms, methods.

- Computer labs where students learn the R language and practice data analysis (using PS data, open data, data scraped for purposes of this class)

- Participatory Sensing “campaigns” where students collect and analyze data. Multivariate data-visualization “dashboard” to generate hypotheses and questions.

- Practicums: projects to tie units together

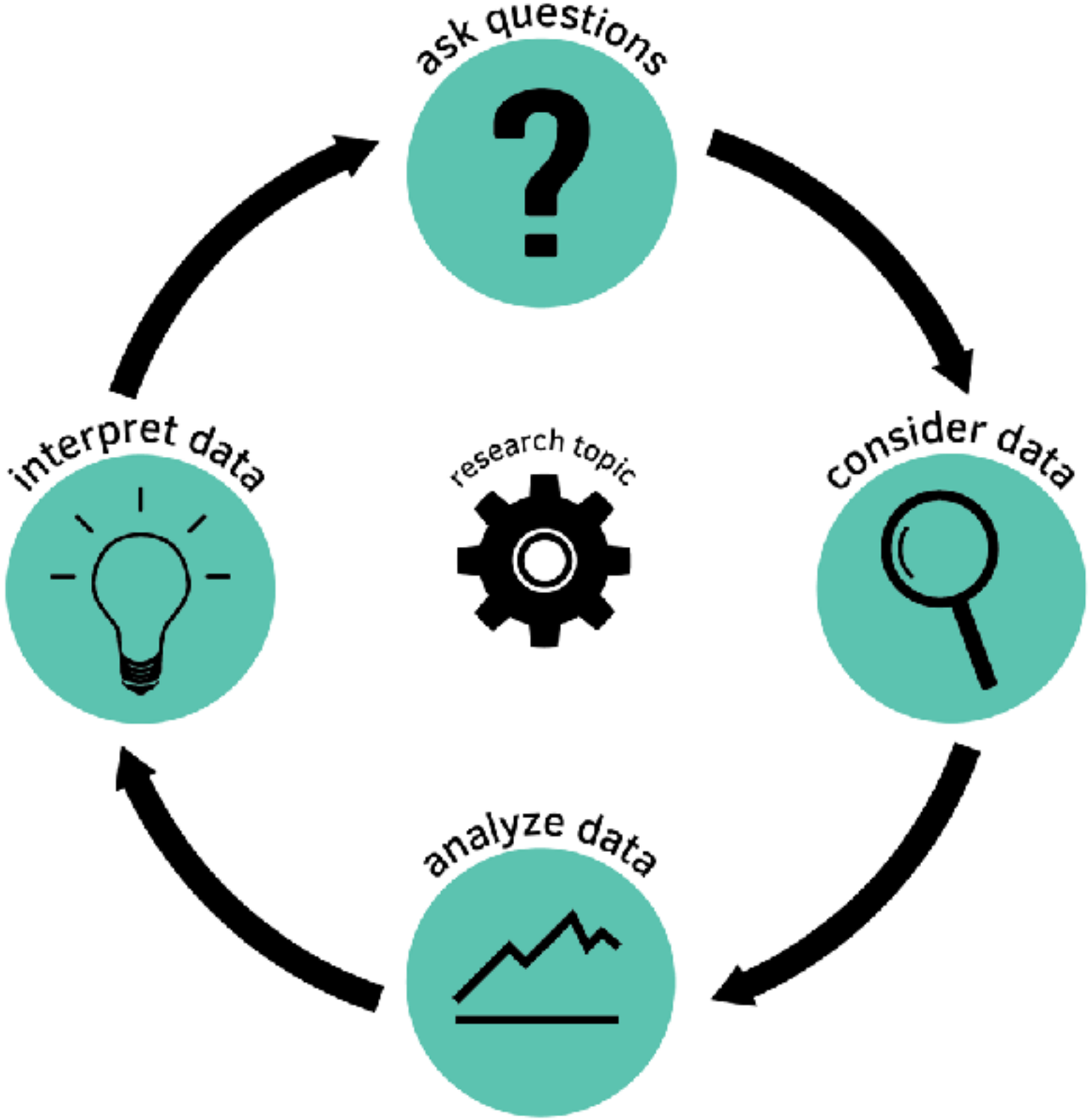
- Work with data that contain a variety of types (dates, GPS, images); collected under modern paradigms (sensors, human sensors); large (but not Big).

Data collection campaigns using mobile devices

- Food habits (Unit 1)
- Time Use (Unit 1)
- Personality Color
- Stress/chill (Unit 3)
- Class designed (emotions, eating after 7pm, appearance)
- Water use (Unit 4)

The Data Cycle

(modeled after the Statistical Investigation Cycle in GAISE K-12)



4 Units, each about 9 weeks

- Unit 1: Focus on data
  “ This unit will introduce the idea of ‘data,’ fundamental to the rest of the course”

- Unit 2: Informal inference using randomization paradigm
  “ This unit deepens the informal reasoning skills developed in Unit 1 by enriching students' technical vocabulary and developing more precise analytical tools. Most importantly, this unit introduces the formal concept of probability as a tool for understanding that sometimes patterns observed in data are not ‘real.’"

- Unit 3: Data Collection
  “ focuses on data collection methods, including traditional methods of designed experiments and observational studies and surveys. It introduces students to sampling error and bias, which cause problems in analysis made from survey data”

- Unit 4: Predictions, Multivariate
  “ This unit will develop modeling skills, beginning with learning to fit and interpret least squares regression lines and learning to use regression to make predictions. Students will learn to evaluate the success of these predictions and so compare models for their predictive accuracy.”

★ Introduction to data, cultural issues, distributions

- Visualization of distributions
- Exploratory data analysis/summary statistics

- Basic probabilities through simulation

★ Informal inference with randomization based testing

- The Normal distribution

- Controlled experiments/random assignment
- Observational studies, confounding factors

★ Survey Sampling/writing questions

★ Humans as sensors to collect data

★ Scraping data from html tables on the internet

- One-variable regression, prediction emphasis

★ Multiple-regression, prediction emphasis

★ Model Eliciting Activity as summary activity

★ Classification and Regression Trees, Clustering with K-means

Some Topics in Labs

- Rstudio commands
- Visualizations
- Importing data
- Subsetting data
- Cleaning names, categories, strings
- loops, generating random numbers, shuffling data
- Merging data
- Long-to-wide data
- Finding probabilities with distributions
- organizing data with XML
- Creating/using Testing and training data

What are the next steps?

Trajectories

- Tim Erickson has identified what he calls “data moves”: skills and operations on data that help students prepare and analyze data.

    - What is the complete set of data moves? How can they be ordered? What concepts and understandings are required to be in place before the data moves can be taught?

- Bill Finzer has identified “Data habits of mind”. Same questions. How do they relate to the data moves?

- Kim et. al (upcoming publication in TISE) introduce the notion of ‘tame’ data. They claim that converting data from wide to narrow is an advanced skill and should be postponed. What evidence do we have to evaluate claims such as these?

- Can students learn statistics with R? What’s gained? What’s lost?

What do our students think?

- Bowler et al (2017*): What do students think the term “data” means? 2 camps: Data as a product of scholarly enterprise; Data in terms of the networked, digital world

- Konold et al (2017, SERJ): Students naturally organize multivariate data as hierarchical; table data is harder

- Haldar et al (2017, to appear in TISE): beginning students can recognize hierarchical data in more formal representations given the appropriate tools

*** Proceedings from 80th annual meeting of association for information science & technology, Washington, dC, Oct 27-Nov 1, 2017**

Professional Development

- Where will teachers learn these skills?

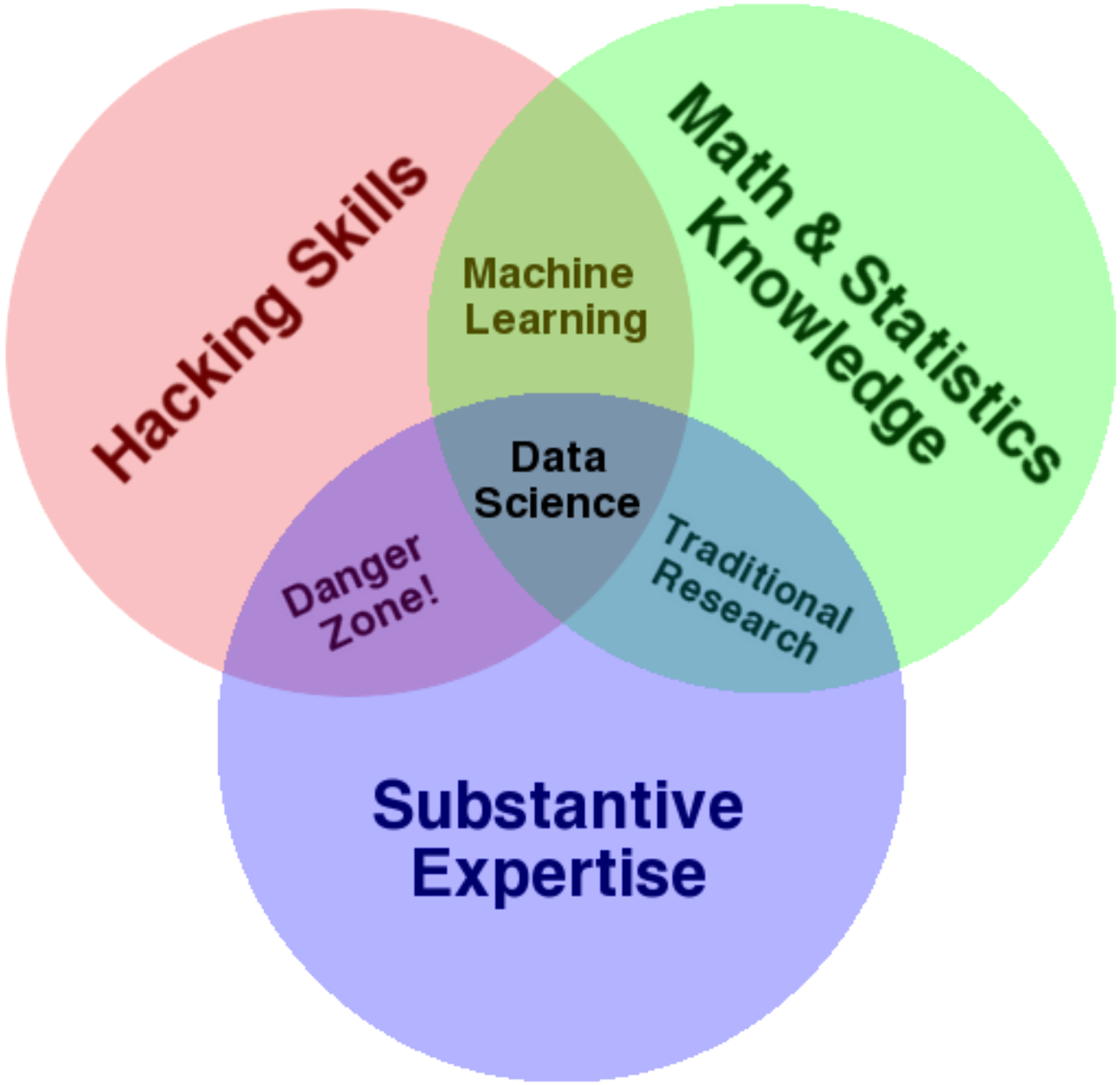
- California is about to create a computer science certificate. But CS isn't a math course, it's a science course.

- Data science is probably more of a math course. (It is in California.) So who trains these teachers?

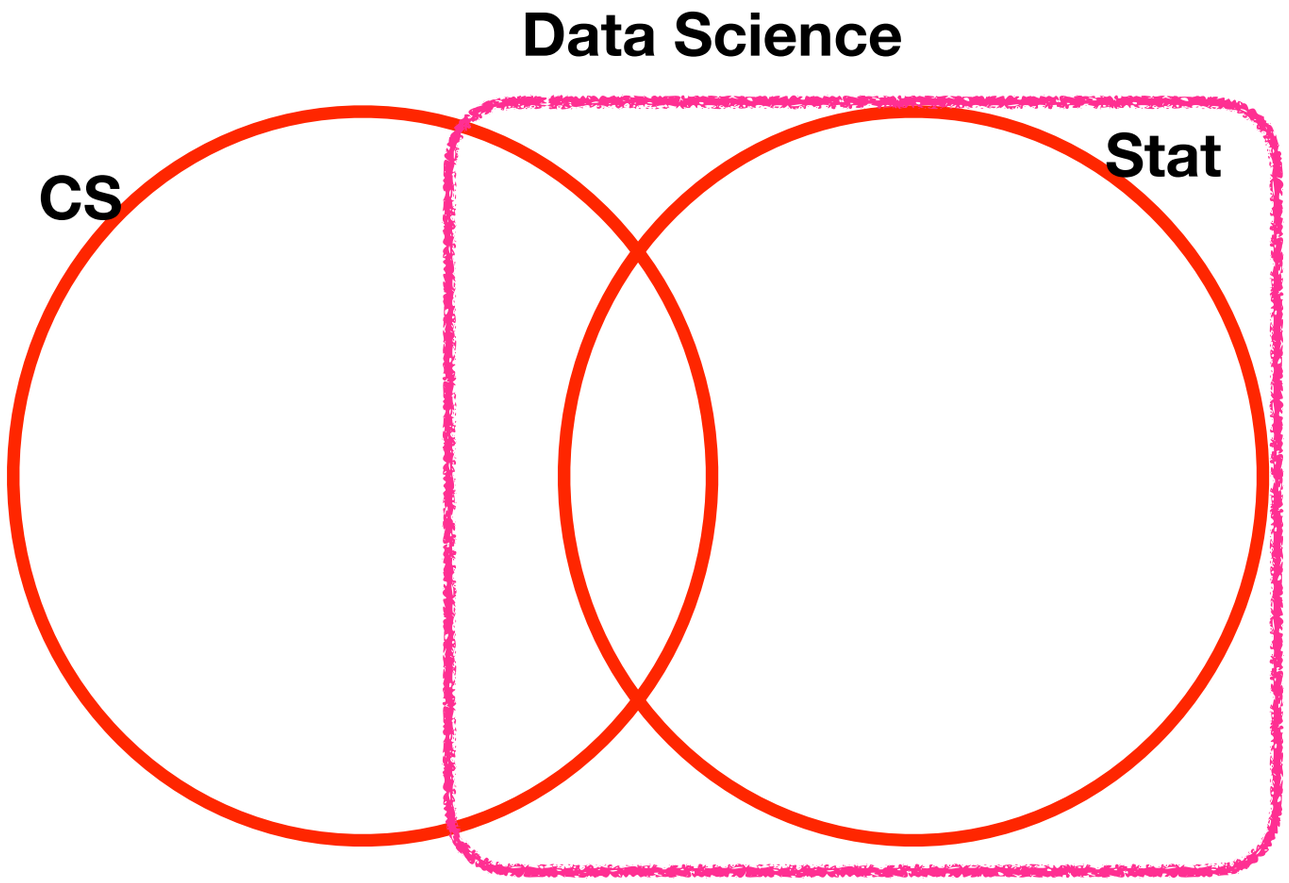
- Pre-service teaching is fairly entrenched. Do we take time from the math curriculum? The science curriculum?

Thanks!

rgould@stat.ucla.edu



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram


Data Science
CS
Stat