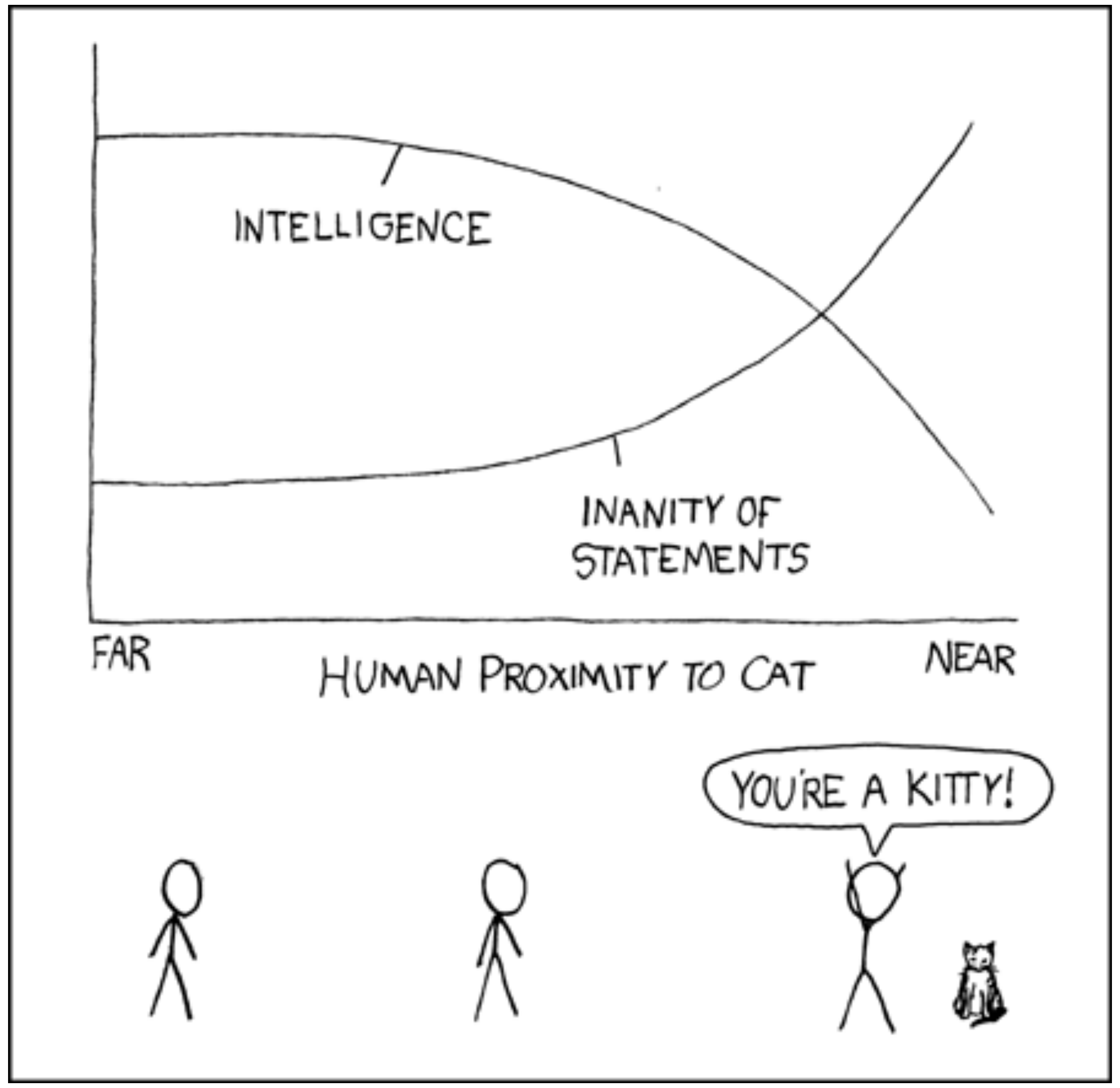
Introduction to Data Science
CS 5963 / Math 3900

Alexander Lex
alex@sci.utah.edu

Braxton Osting
osting@math.utah.edu

THE UNIVERSITY OF UTAH



[xkcd]

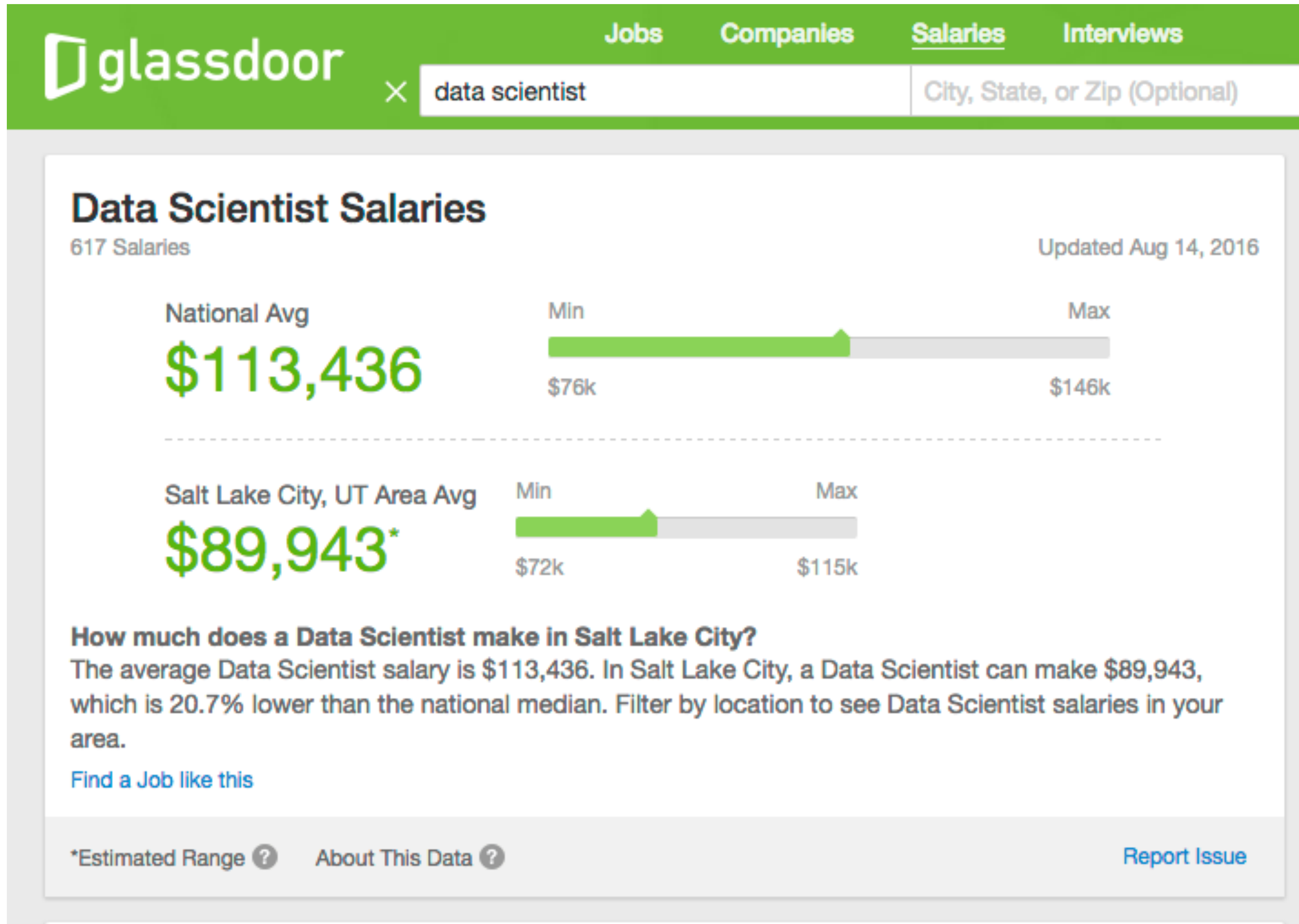
What is Data Science?

The sexiest job of the century —Harvard Buisness Review

A data scientist is a statistician who lives in San Fransisco

Data Science is statistics on a Mac

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.



Data Scientist Salaries

617 Salaries

Updated Aug 14, 2016

National Avg

$113,436

Salt Lake City, UT Area Avg

$89,943*

How much does a Data Scientist make in Salt Lake City?

The average Data Scientist salary is $113,436. In Salt Lake City, a Data Scientist can make $89,943, which is 20.7% lower than the national median. Filter by location to see Data Scientist salaries in your area.

Find a Job like this

*Estimated Range

About This Data

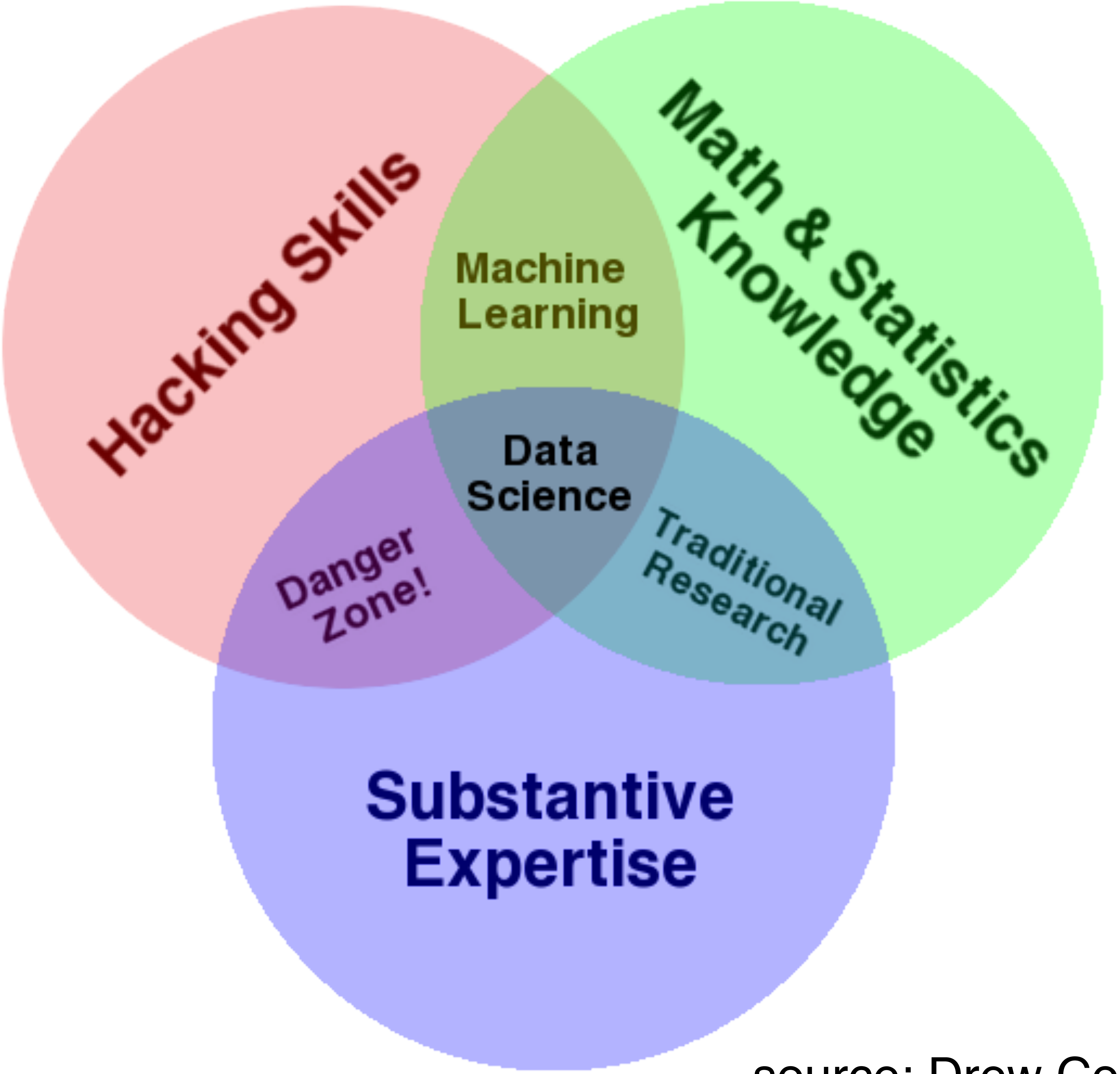
Report Issue

https://twitter.com/jeremyjarvis/status/428848527226437632/photo/1

What is Data Science?

Source: datascience.berkeley.edu

What is Data Science?



source: Drew Conway blog

What is Data Science?

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms. (Wikipedia)

Data Science closes the circle from collecting real-world data, to processing and analyzing it, to influence the real world again.

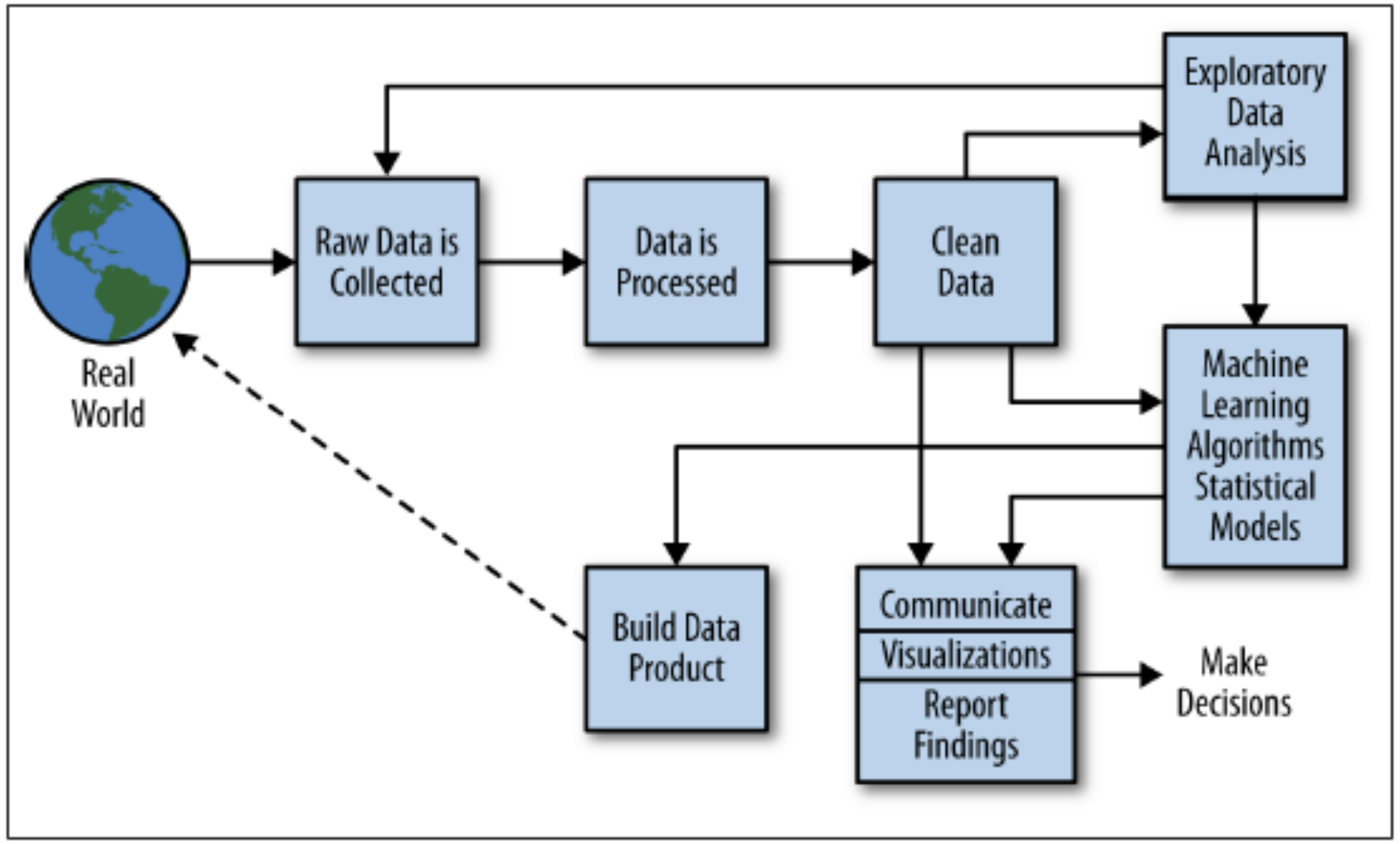


Figure 2-2. The data science process

DDS, p.41

Data Science vs. Machine Learning vs. Statistics ?!?

-> read 50 years of Data Science by David Donoho

What is Data Science?

> “The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that’s going to be a hugely important skill in the next decades, … because now we really do have **essentially free and ubiquitous data**.”

Hal Varian, Google’s Chief Economist
The McKinsey Quarterly, Jan 2009

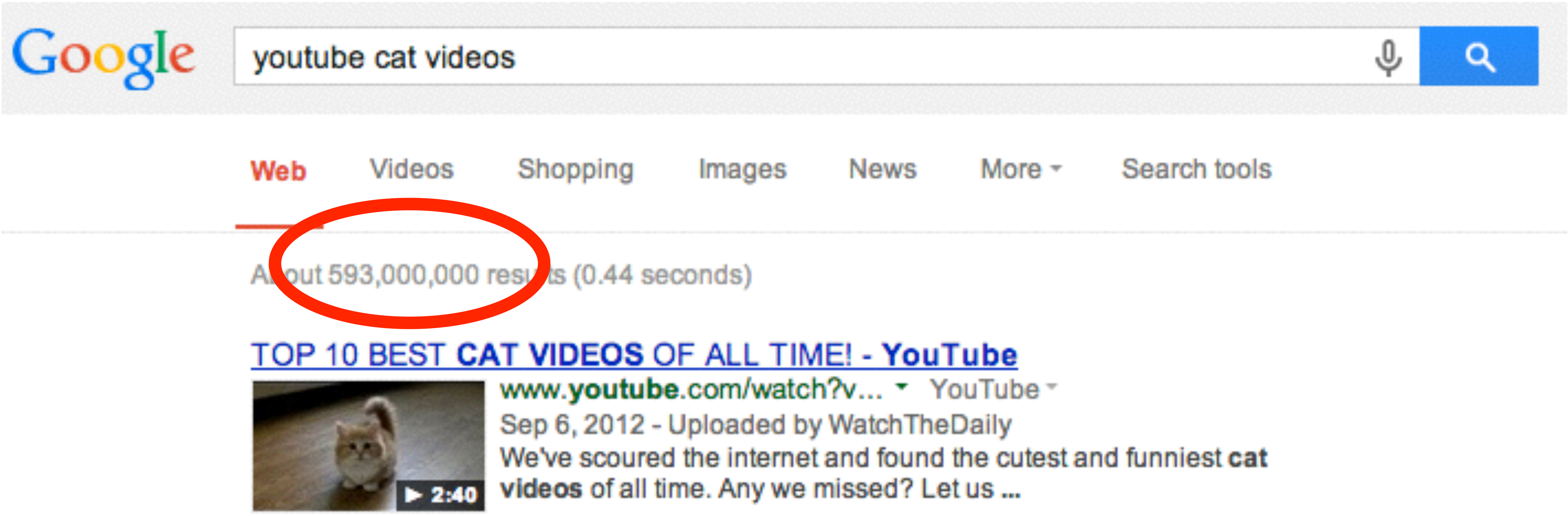
Big Data

2010: 1,200 exabytes, largely unstructured

Google stores ~10 exabytes (2013)

Hard disk industry ships ~8 exabytes/year

2.5 exabytes (2.5 billion gigabytes) generated every day in 2012

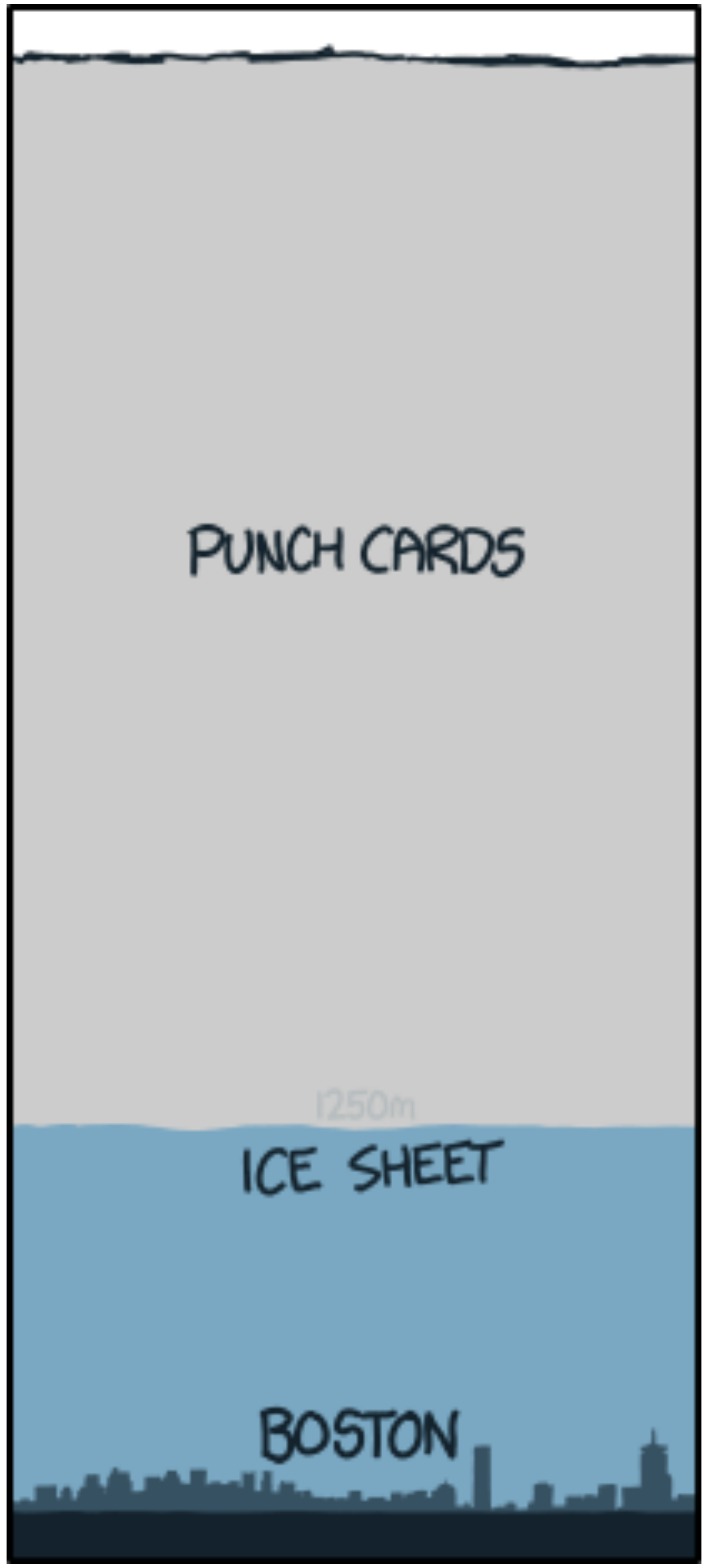


youtube cat videos

About 593,000,000 results (0.44 seconds)

TOP 10 BEST CAT VIDEOS OF ALL TIME! - YouTube
www.youtube.com/watch?v... YouTube
Sep 6, 2012 - Uploaded by WatchTheDaily
We've scoured the internet and found the cutest and funniest cat videos of all time. Any we missed? Let us ...

15 Exabytes in Punch Cards:
4.5 km over New England



In one second on the Internet there are...

http://onesecond.designly.com/

How can we leverage data?

Improve your fitness by targeted training

Improve your product

- by targeting your audience
- by considering semantics

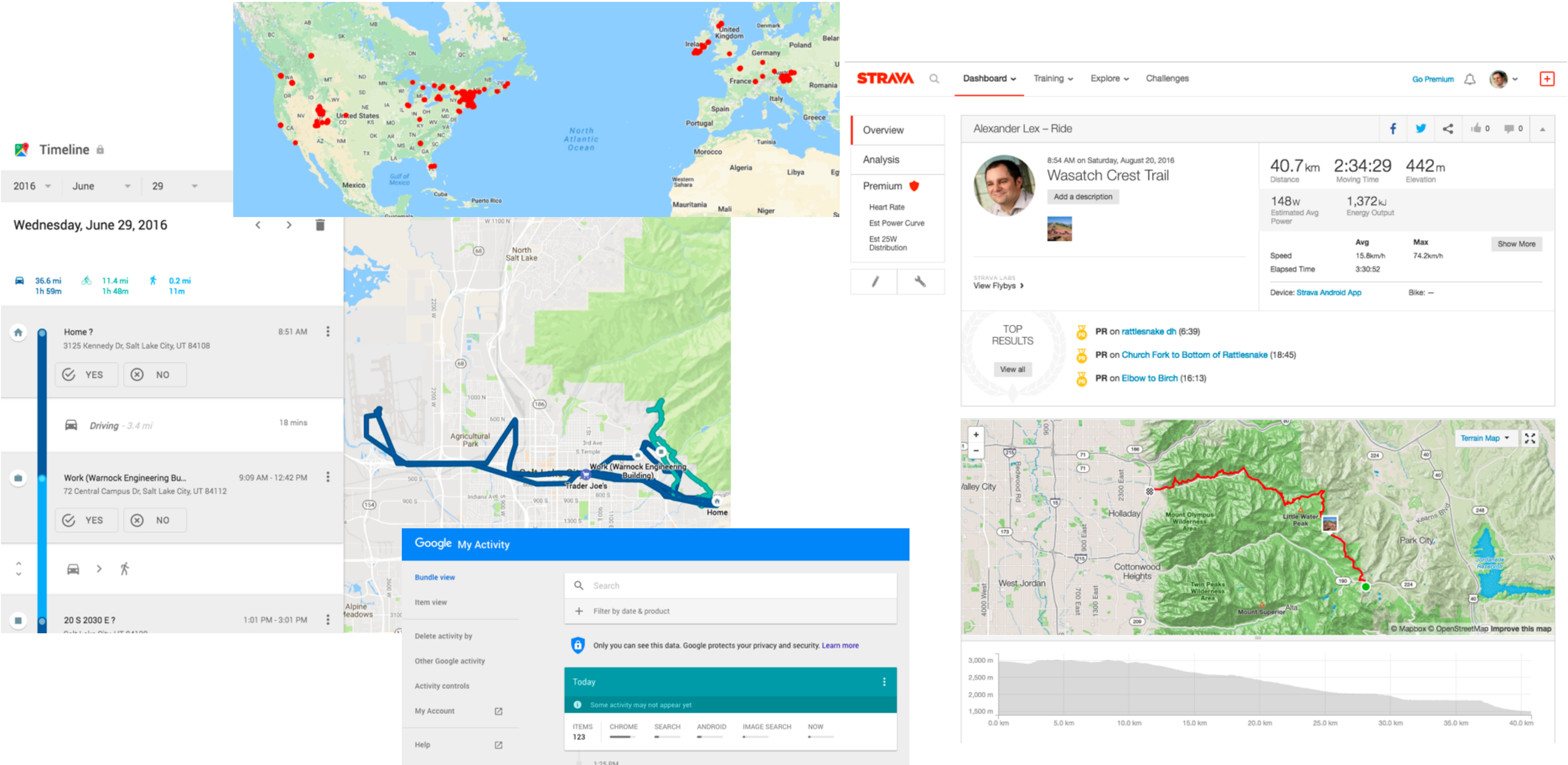
Make better decisions

- exact diagnosis, choose right medication, pick good restaurant

Predict elections, events, crowd behavior, etc.

… and many more applications

Example: Personal Data



Timeline

2016 June 29

Wednesday, June 29, 2016

36.6 mi 1h 59m · 11.4 mi 1h 48m · 0.2 mi 11m

Home ? — 8:51 AM
3125 Kennedy Dr, Salt Lake City, UT 84108
YES · NO

Driving - 3.4 mi — 18 mins

Work (Warnock Engineering Bu... — 9:09 AM - 12:42 PM
72 Central Campus Dr, Salt Lake City, UT 84112
YES · NO

20 S 2030 E ? — 1:01 PM - 3:01 PM

Google My Activity

Bundle view
Item view
Delete activity by
Other Google activity
Activity controls
My Account
Help

Search
Filter by date & product

Only you can see this data. Google protects your privacy and security. Learn more

Today
Some activity may not appear yet

ITEMS 123 · CHROME · SEARCH · ANDROID · IMAGE SEARCH · NOW

1:25 PM

STRAVA — Dashboard · Training · Explore · Challenges · Go Premium

Overview
Analysis
Premium
Heart Rate
Est Power Curve
Est 25W Distribution

Alexander Lex – Ride

8:54 AM on Saturday, August 20, 2016
Wasatch Crest Trail
Add a description

STRAVA LABS
View Flybys

40.7 km Distance · 2:34:29 Moving Time · 442 m Elevation
148 W Estimated Avg Power · 1,372 kJ Energy Output

	Avg	Max
Speed	15.8km/h	74.2km/h
Elapsed Time	3:30:52	

Show More

Device: Strava Android App · Bike: —

TOP RESULTS
View all

PR on rattlesnake dh (6:39)
PR on Church Fork to Bottom of Rattlesnake (18:45)
PR on Elbow to Birch (16:13)

Big Data in Science and Engineering

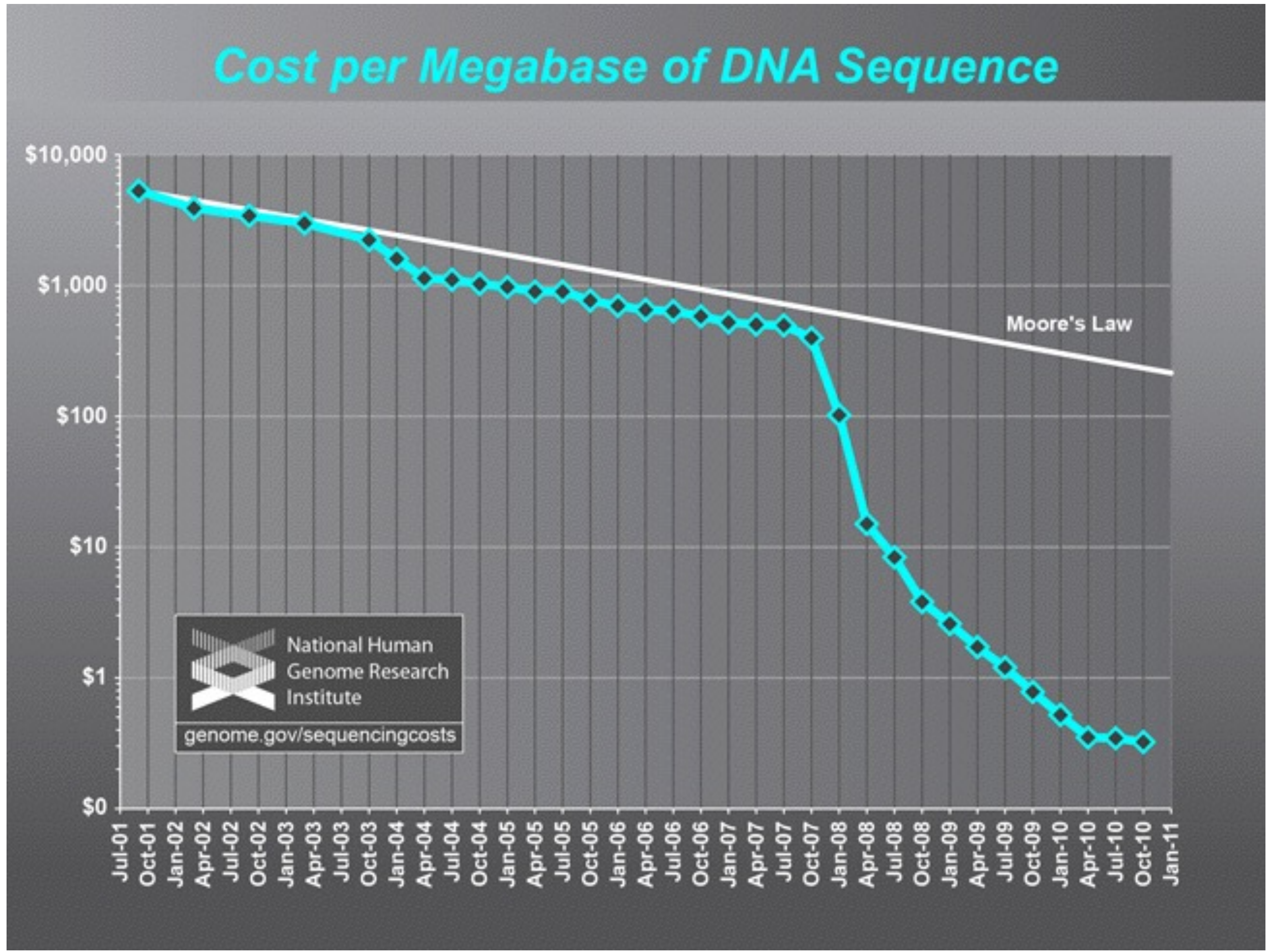
“Big Data” hasn’t just transformed industry!

It’s also transformed science and engineering. Cheap sensors (e.g. imaging) have changed the way science and engineering are done.

Examples:

- Large physics experiments and observations
- Cheaper and automated genome sequencing
- Smart buildings / cities (blyncsy)
- Geophysical imaging

Controversy: Hypothesis or data driven methods



Example: CERN Large Hadron Collider Data

CERN has publicly released over 300TB of data: CERN Open Data Portal

How much is that?

- At 15 GB of storage a piece, you'd need 20,000 Gmail accounts to store the whole shebang. If you wanted to send that much data at the max attachment size of 25 MB, it would take you 12 million emails.
- A DVD-R holds 4.7 GB. You'd need 63,830 of them to hold 300 TB.
- Your Blu-ray collection wouldn't need to expand quite so much. 6,000 discs ought to hold it.
- It takes Pandora about a day and a half to burn through a gig of mobile data. So if the CERN data was an album, you could stream it in just over 1,230 years.
- At 350 MB per hour for 4K video streaming, so if the CERN data was a 4K movie it'd probably be about 857,142 hours, or about 98 years long.
- But it ain't no thing compared to what the National Security Agency works with. Going by 2013 figures the agency released, the NSA's various activities "touch" 300 TB of data every 15 minutes or so

(Popular Mechanics Article)

Example: Genomics

Example TCGA: 1 Petabyte

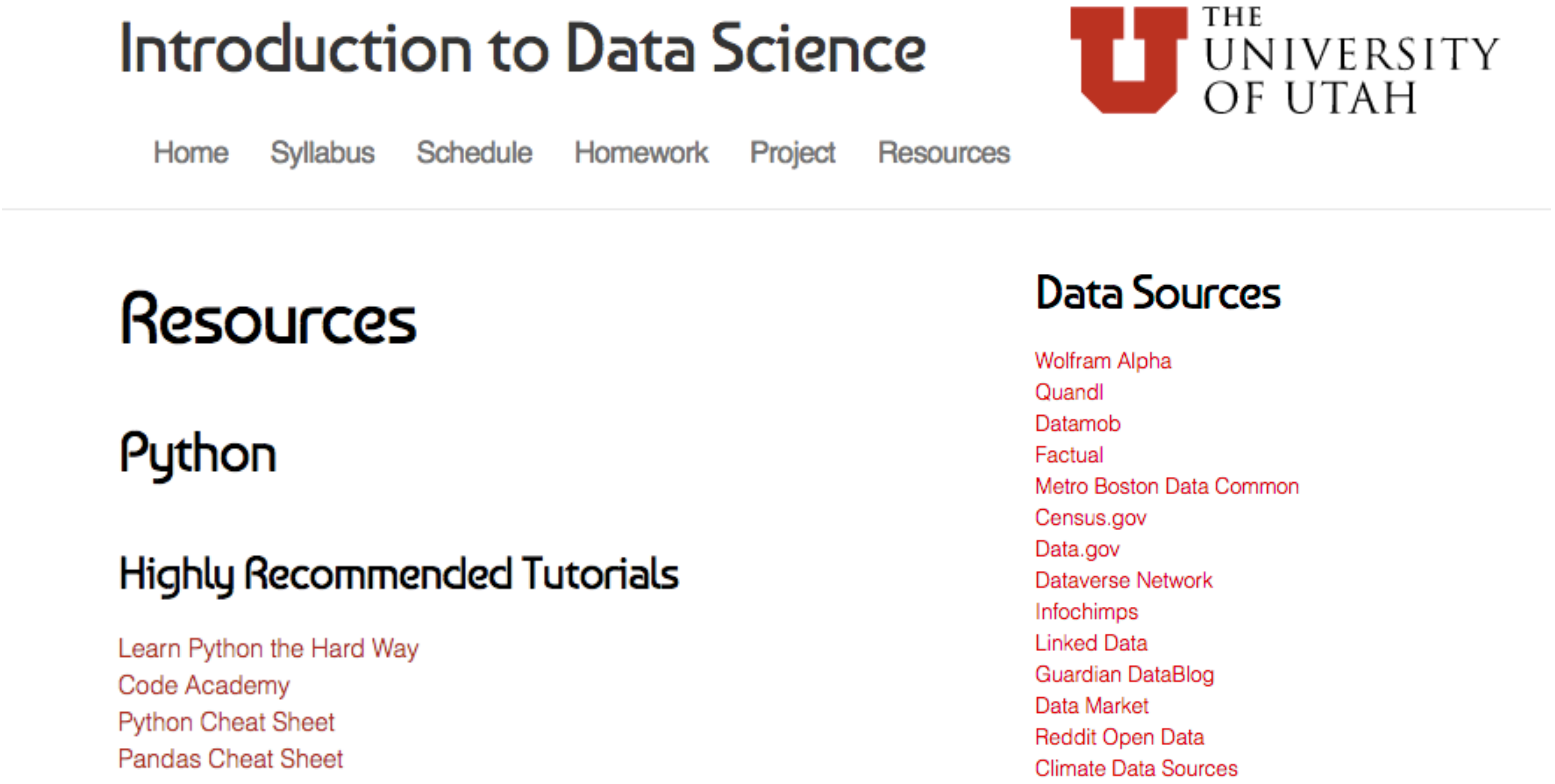
NSA Utah Data Center (Bluffdale, Utah)

Storage Capacity?

estimates vary, but Forbes magazine estimates 12 exabytes (12,000 petabytes or 12 million terabytes)

Where to find data?

Today, a lot of data is publicly available. You probably have access to data you’re interested in. If not, to get you started, we’ve provided some links to repositories on the course website.



Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Homework Project Resources

Resources

Python

Highly Recommended Tutorials

Learn Python the Hard Way
Code Academy
Python Cheat Sheet
Pandas Cheat Sheet

Data Sources

Wolfram Alpha
Quandl
Datamob
Factual
Metro Boston Data Common
Census.gov
Data.gov
Dataverse Network
Infochimps
Linked Data
Guardian DataBlog
Data Market
Reddit Open Data
Climate Data Sources

Who is CS-5963 / Math-3900?

Alexander Lex

@alexander_lex
http://alexander-lex.net
http://vdl.sci.utah.edu

Assistant Professor, Computer Science

Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology

Large, Multivariate (Biological) Networks

```
MATCH p = (s)-[x:FLIGHT]->(i)-[y:FLIGHT]->(t)  WHERE s.state in
['CA', 'OR', 'WA']  AND t.state in ['CT', 'ME', 'MA', 'RI', 'NH', 'VT',
'WA', 'TX', 'GA'] AND x.carrier = y.carrier AND x.arr_time < y.dep_time
```

Query statistics
Length 2
Paths 18389

Collapse rows: state
Collapse cols: none
Sort order: Select
Cell encoding: colormap
Path metric: path count

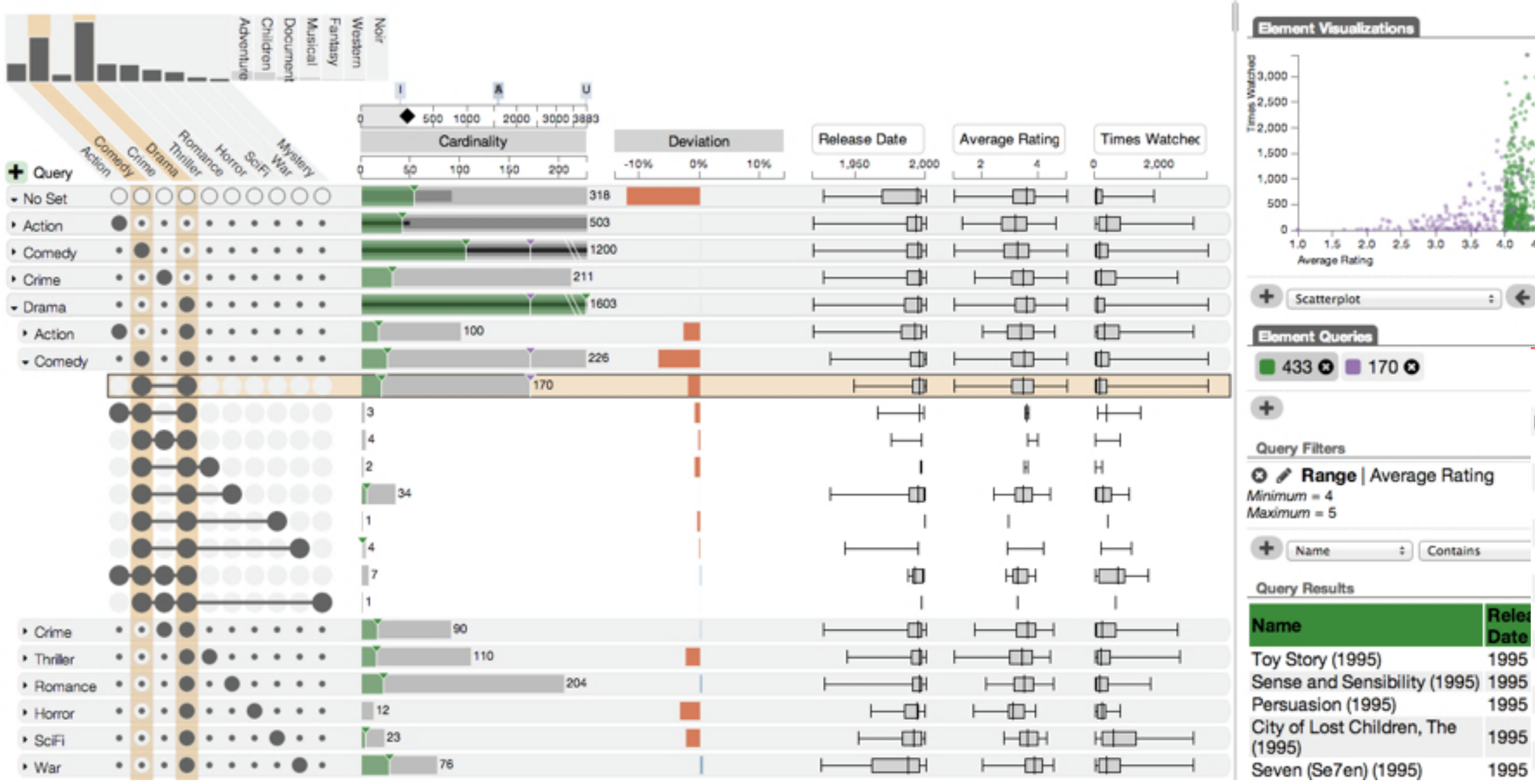
Start Hanspeter Pfis End Ben Shneiderm

Length	0	1	2	3	4
Paths	0	0	0	3	105

Pathways
Pancreatic cancer
PANCREATIC CANCER
Melanoma
Prostate cancer
ErbB signaling pathway
Selected Path
skin
lung
breast
NRAS
RALGDS
RALA
RALB
RALBP1
RAC3
RAC1
CDC42
RAC2
MAP3K1
MAP2K4
Glioma
MAPK signaling
Submit
Connectivity Matrix
degree
city
state
market
id
Intermediate Nodes
count
Node-link View
Layout
ATL
DFW
PHX
SEA
LAX
LAS
ELP
Legend
Pathfinder
Advanced Query
Path List
Path Topology
Hanspeter Pfiste
Frank van Ham
Adam Perer
Ben Shneiderman
Krzysztof Z. Gaj
Desney S. Tan
Jean-Daniel Feke
Catherine Plaisan
Jennifer Golbeck
Wendy E. Macka
Ed Huai-hsin Ch
Jeffrey Heer
CHI
TVCG
chi_publications
cited
degree
tvcg_publication

Multidimensional Data

Set Visualization



Query Filters

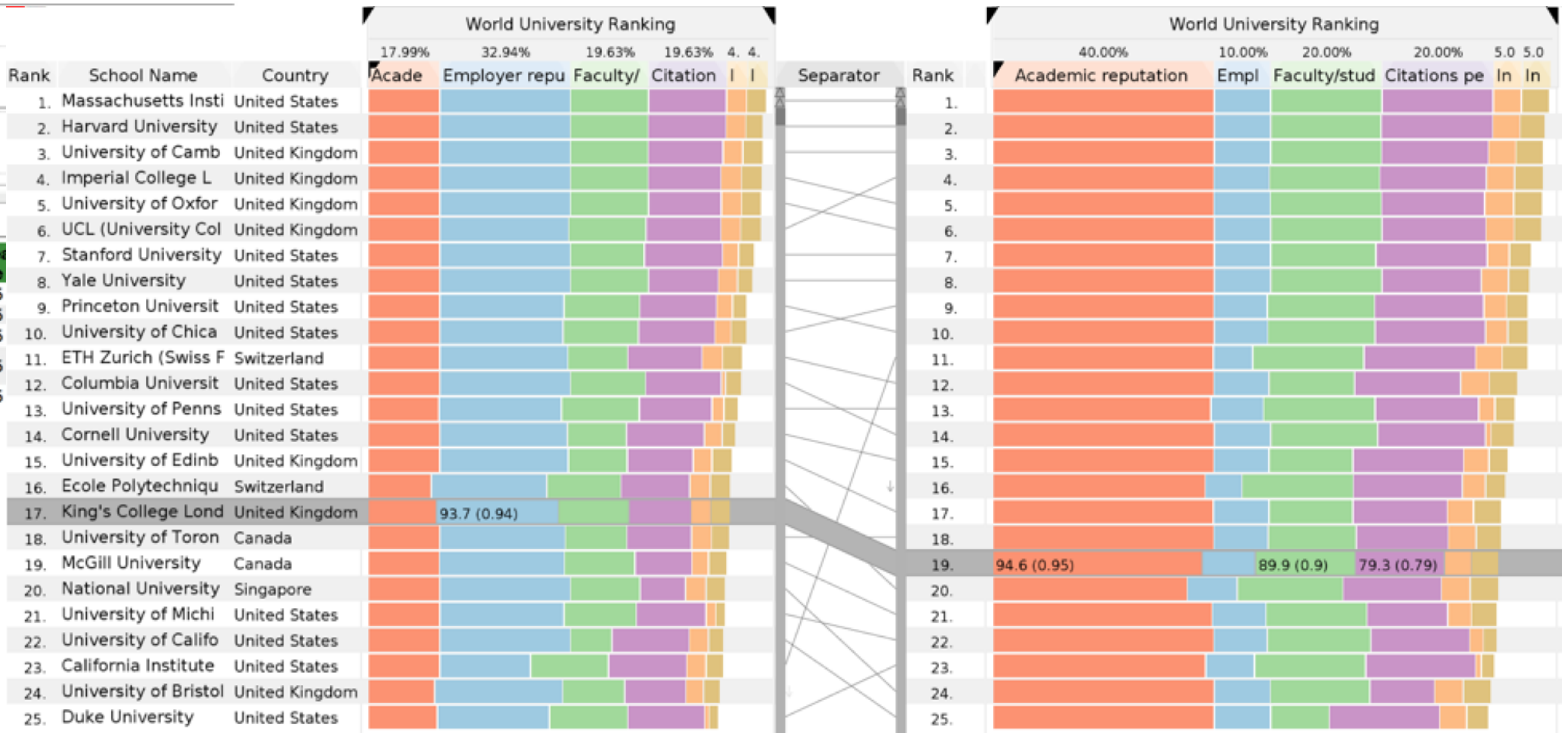
Range | Average Rating
Minimum = 4
Maximum = 5

Name ◇ Contains

Query Results

Name	Release Date
Toy Story (1995)	1995
Sense and Sensibility (1995)	1995
Persuasion (1995)	1995
City of Lost Children, The (1995)	1995
Seven (Se7en) (1995)	1995

Multivariate Rankings

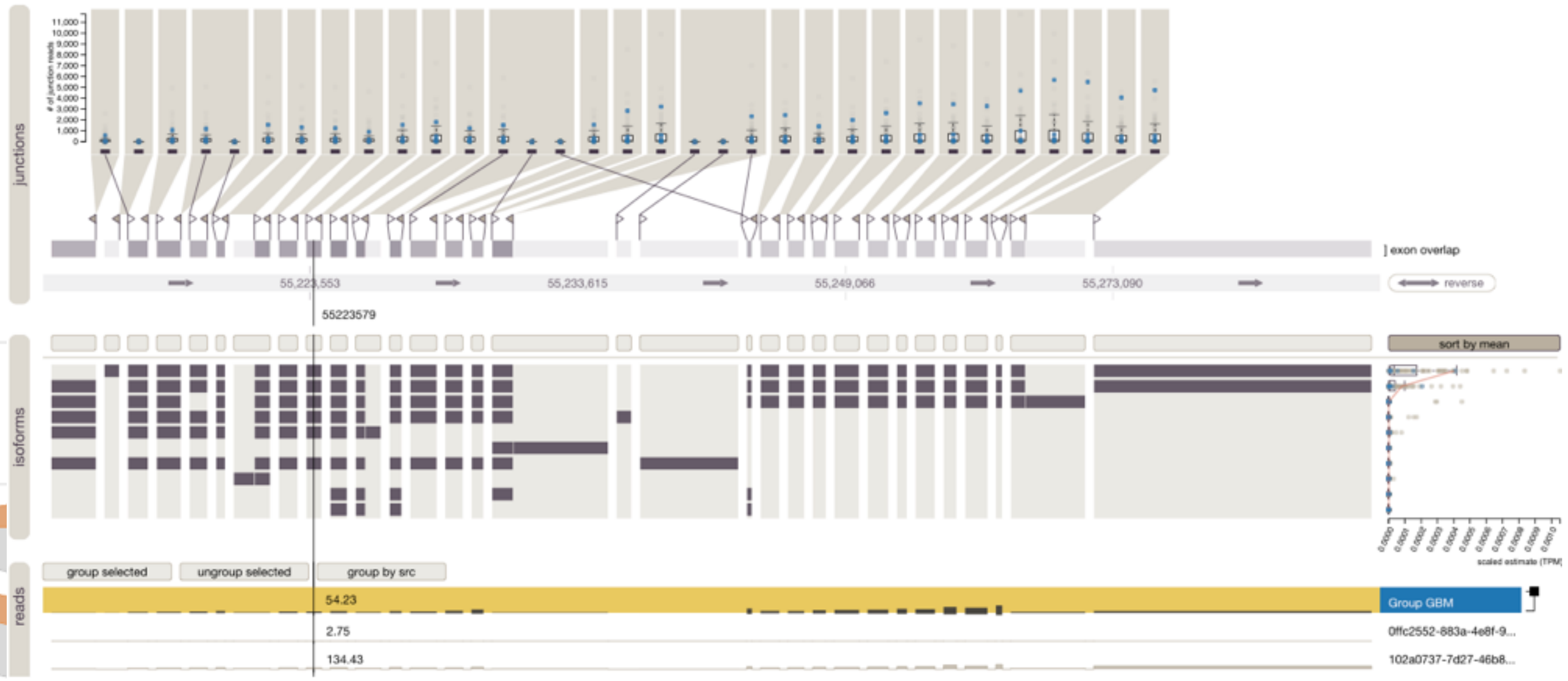


Rank	School Name	Country
1.	Massachusetts Insti	United States
2.	Harvard University	United States
3.	University of Camb	United Kingdom
4.	Imperial College L	United Kingdom
5.	University of Oxfor	United Kingdom
6.	UCL (University Col	United Kingdom
7.	Stanford University	United States
8.	Yale University	United States
9.	Princeton Universit	United States
10.	University of Chica	United States
11.	ETH Zurich (Swiss F	Switzerland
12.	Columbia Universit	United States
13.	University of Penns	United States
14.	Cornell University	United States
15.	University of Edinb	United Kingdom
16.	Ecole Polytechniqu	Switzerland
17.	King's College Lond	United Kingdom
18.	University of Toron	Canada
19.	McGill University	Canada
20.	National University	Singapore
21.	University of Michi	United States
22.	University of Califo	United States
23.	California Institute	United States
24.	University of Bristol	United Kingdom
25.	Duke University	United States

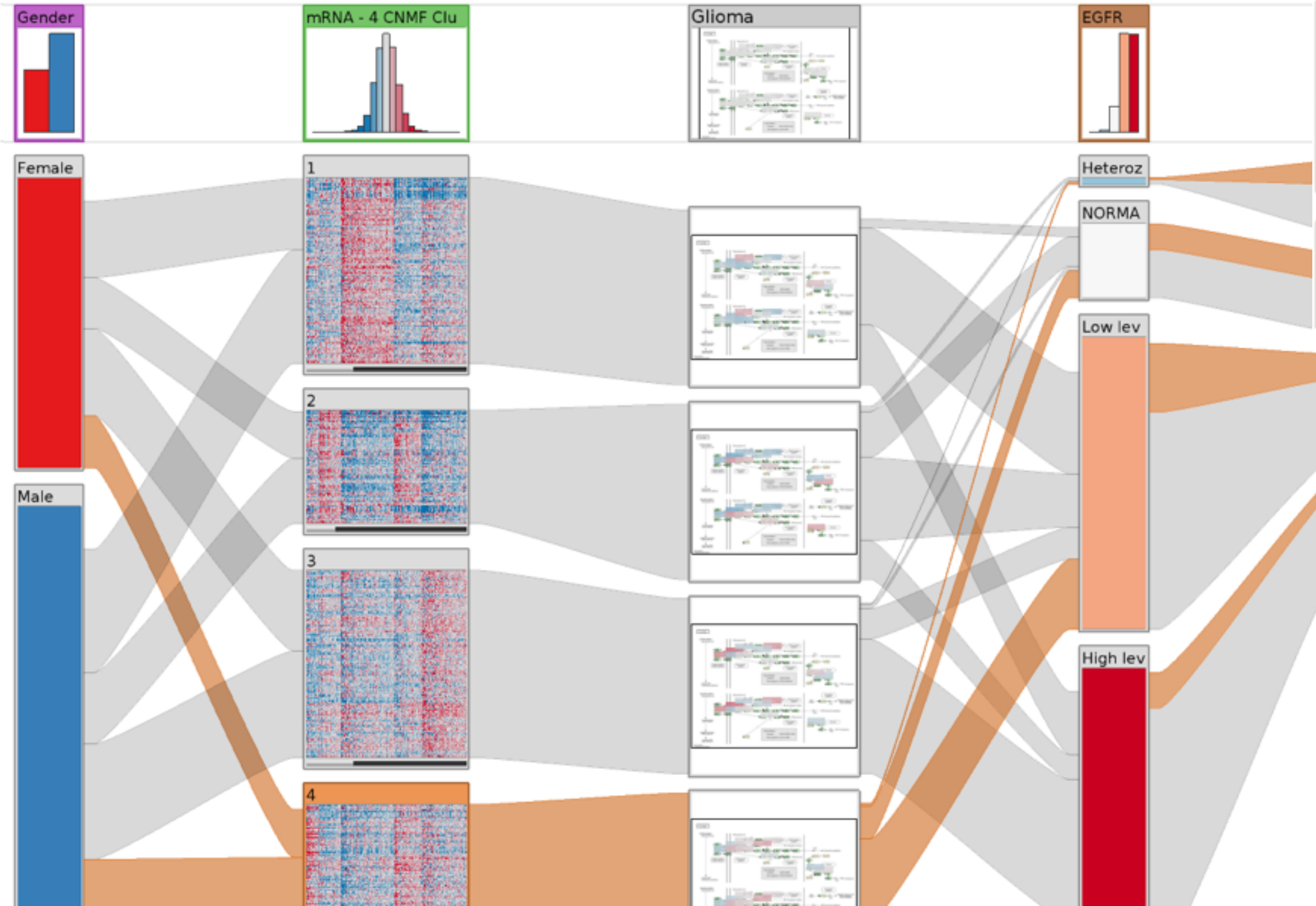
Rank
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.
21.
22.
23.
24.
25.

Genomic Data

Alternative Splicing / mRNA-seq



Cancer Subtypes / Omics Clustering and Stratification



Braxton Osting

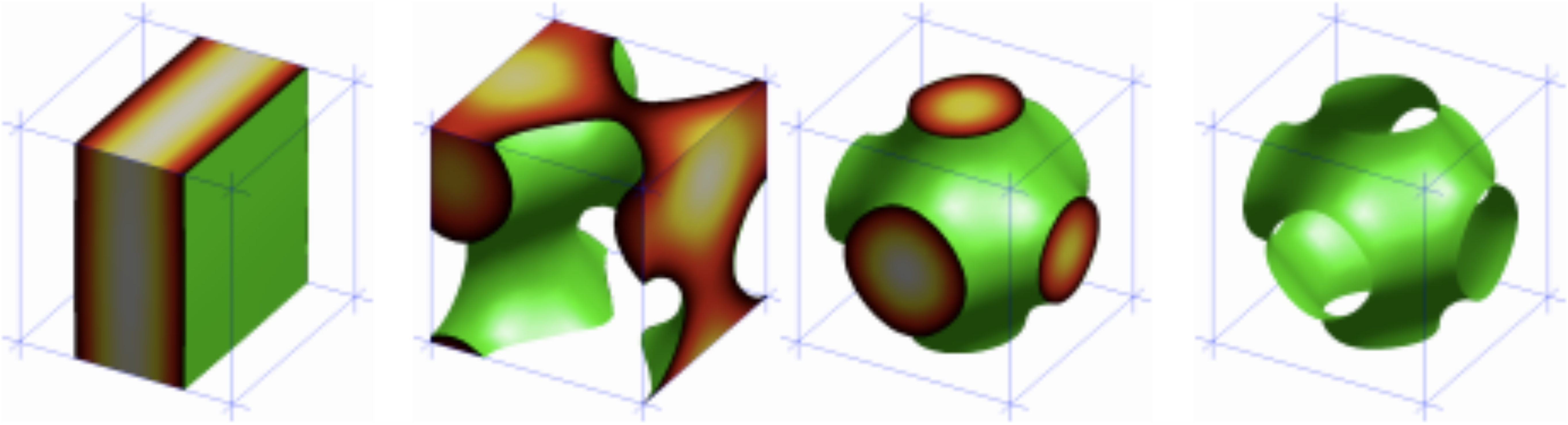
Assistant Professor, Mathematics

Before that: Lecturer, Postdoctoral Fellow, UCLA

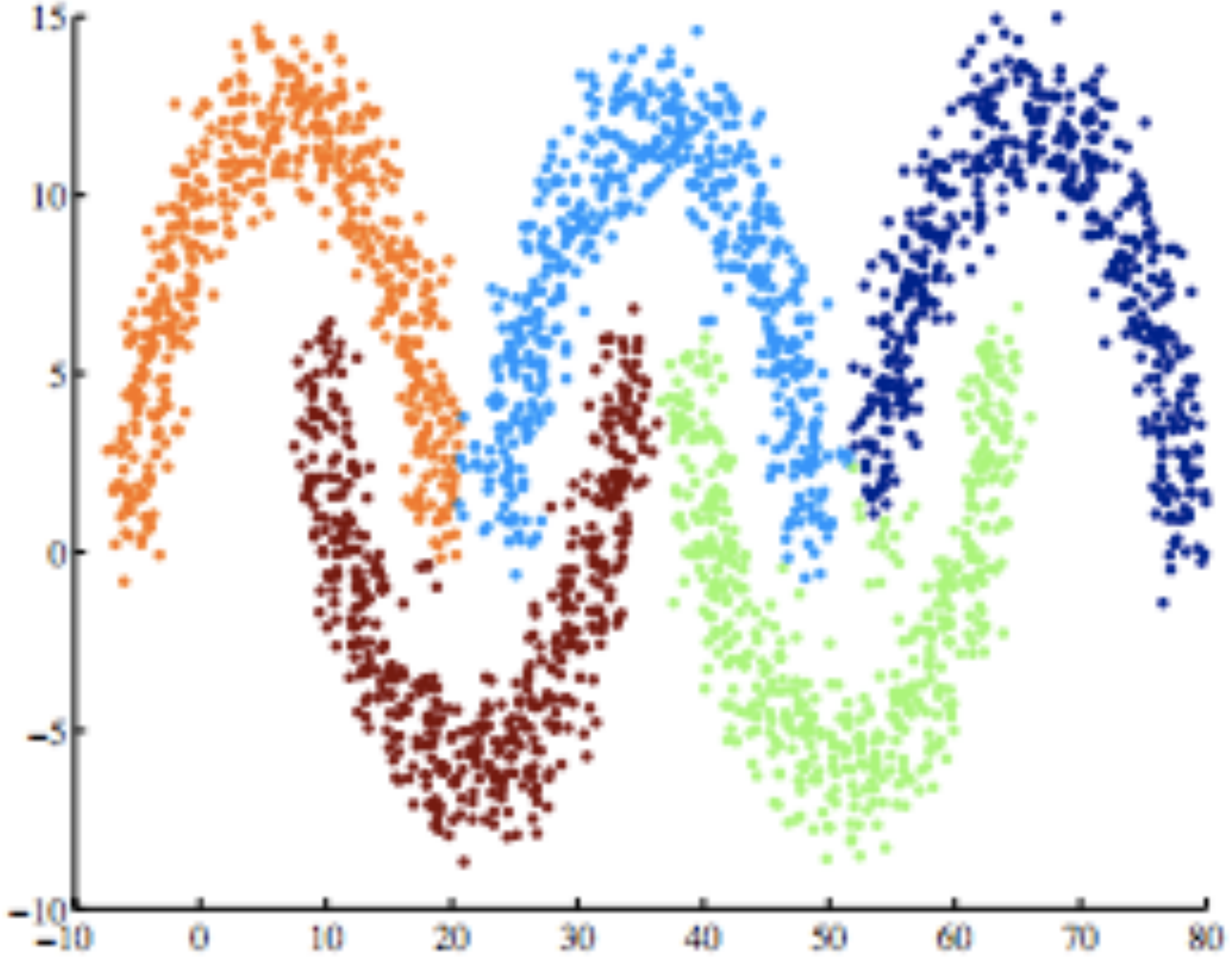
PhD in Applied Mathematics, Columbia University

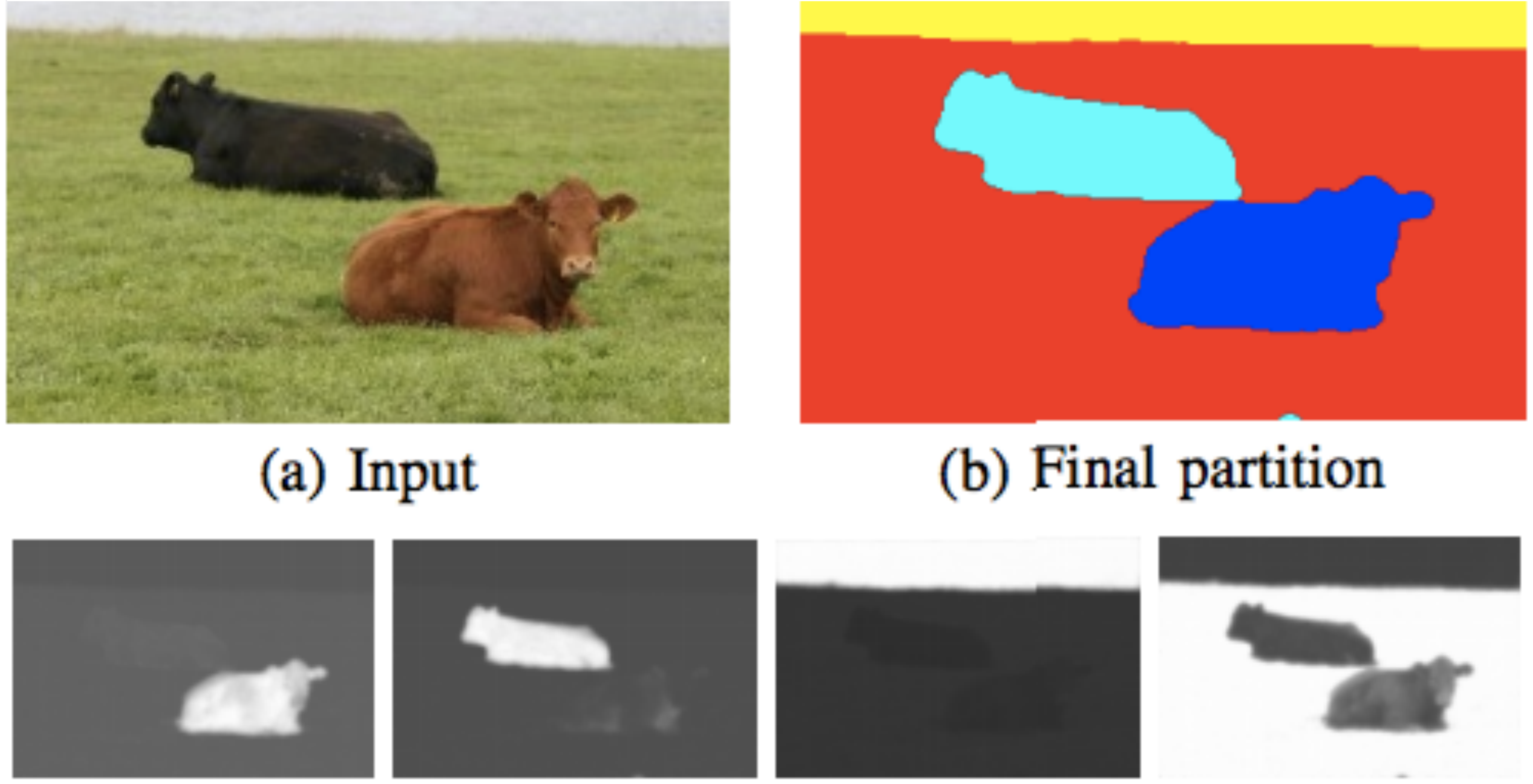
http://math.utah.edu/~osting

Partitioning, Clustering, and Image Segmentation



$o = 2.4957 \cdot 10^{-6}$ $o = 2.3099 \cdot 10^{-6}$ $o = 2.3018 \cdot 10^{-6}$



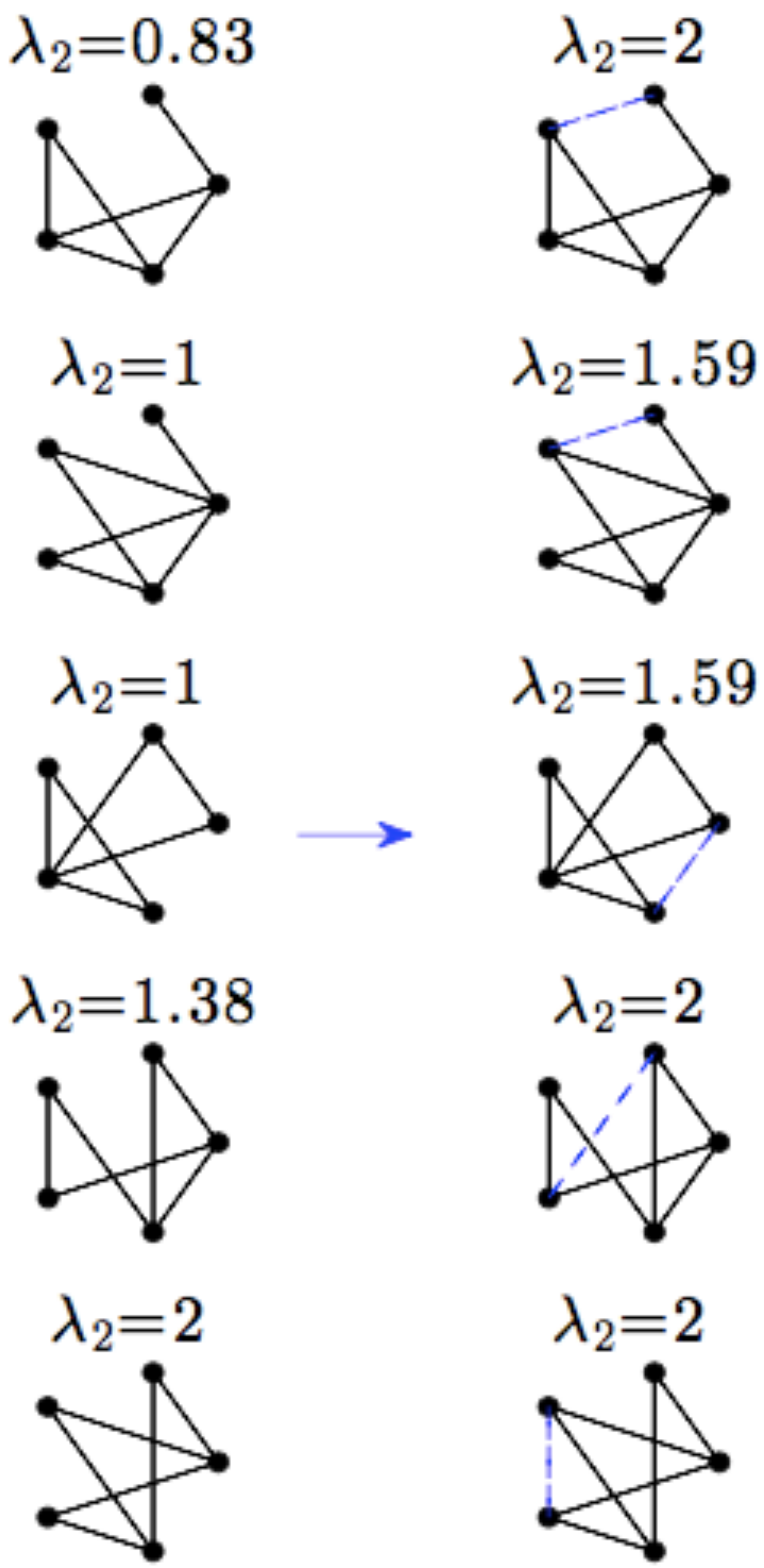


(a) Input

(b) Final partition

(c) Ground states u_ℓ

Statistical Ranking and Active Learning



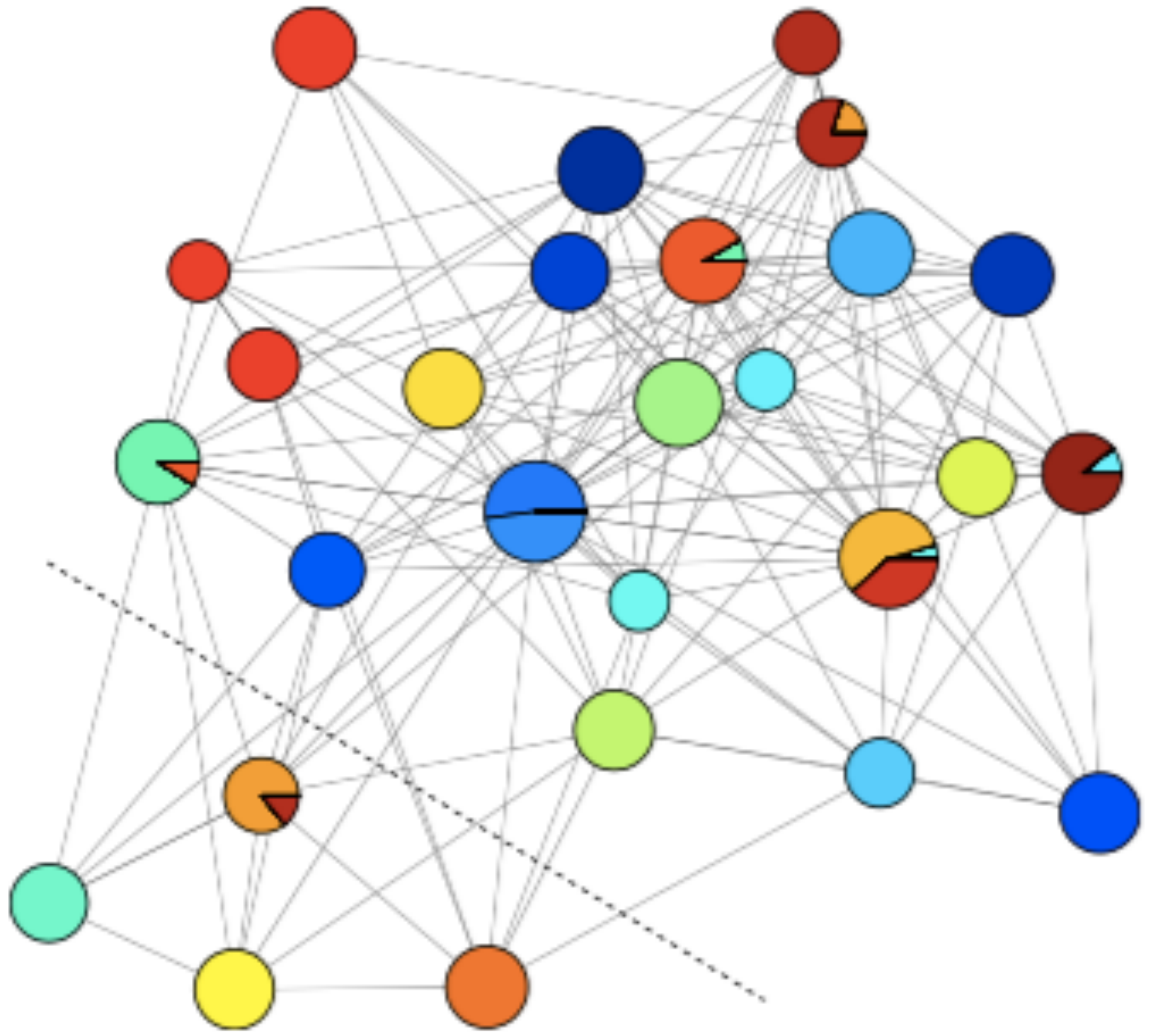
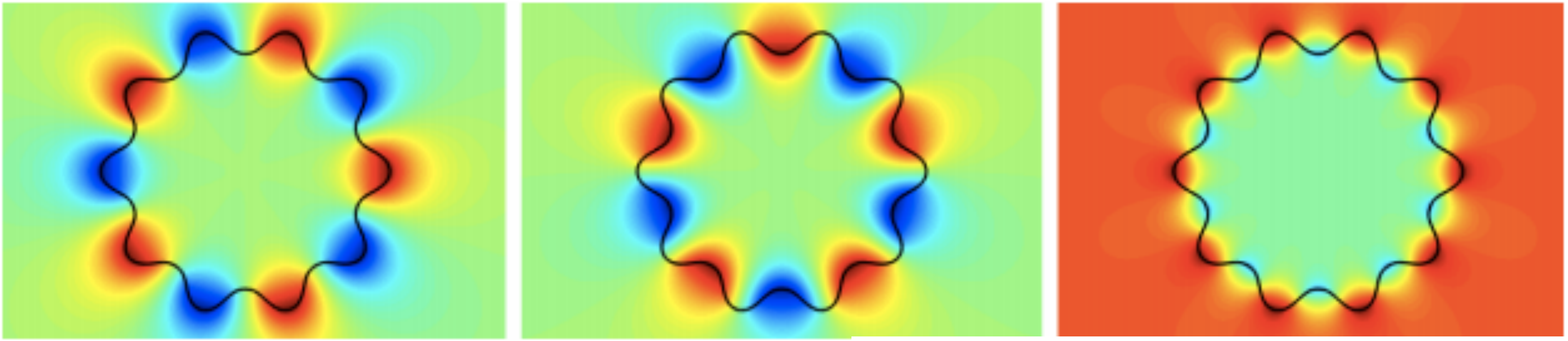
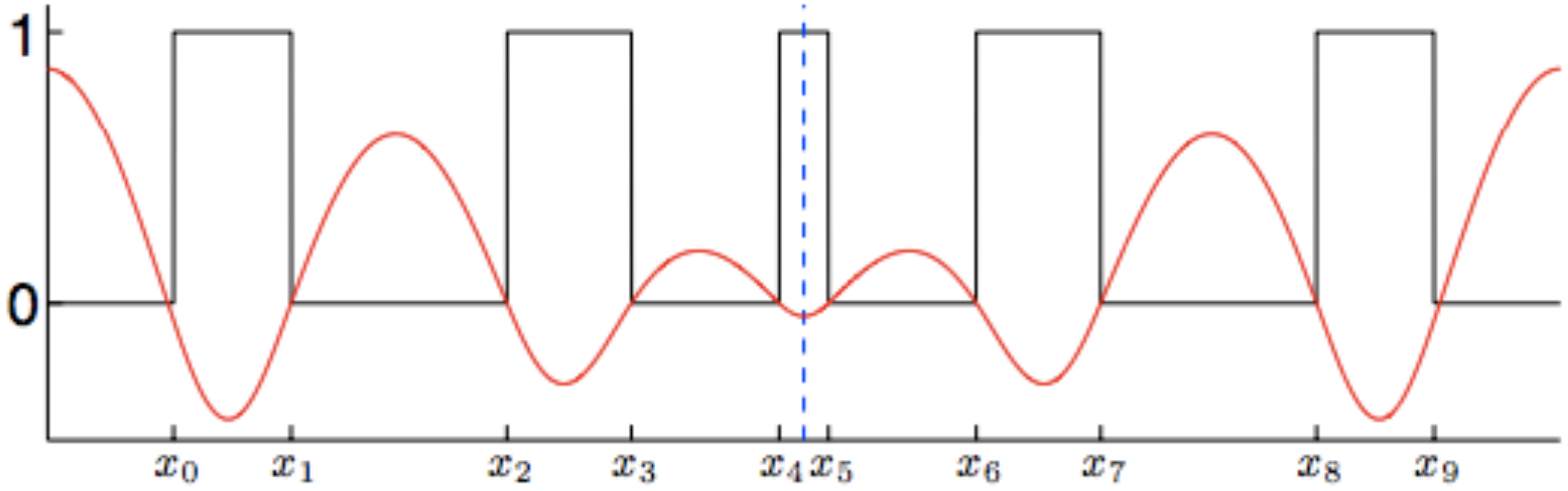
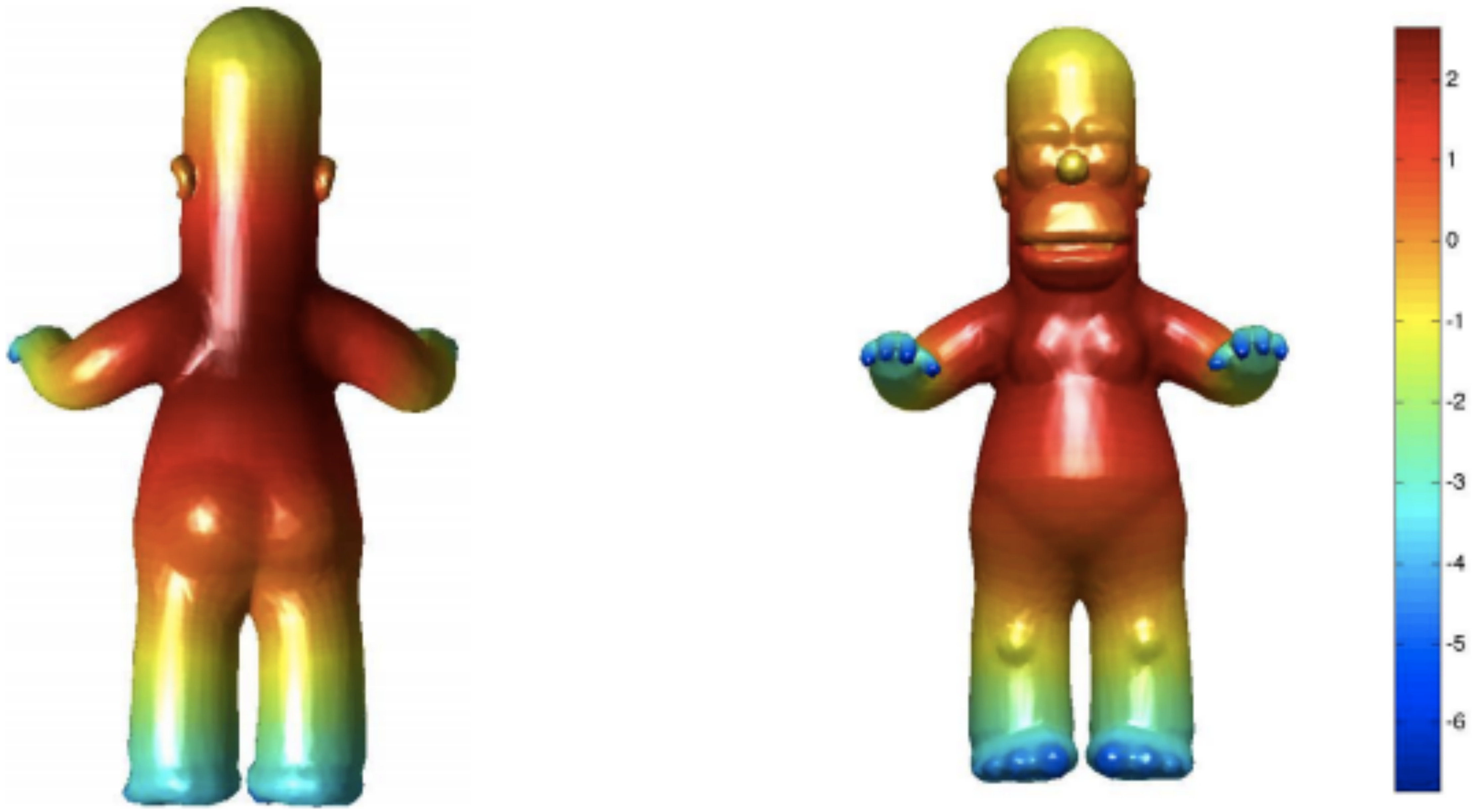


Figure 3: **2011-12 NCAA Division 1 (FBS and FCS) football schedule.** Graph representation of schedule via spectral clustering by games, *top:* vertices represent teams, edges represent games, coloring indicates conference membership. *bottom:* community detection of teams (represented using pie-graphs) reveals that teams primarily play within their own conference. The dashed lines indicate an edge cut which is discussed in the text. See §5.3.

Extremal Eigenvalues







Teaching Assistants

Olivia Dennis

Magdalena Schwarzl

Structure & Goals

Course Goals

Convey basic skills about each step in the data science process

data wrangling: acquire, clean, reshape, sample data
data exploration: get a feeling for the dataset
prediction: inferences and decisions based on data
communication

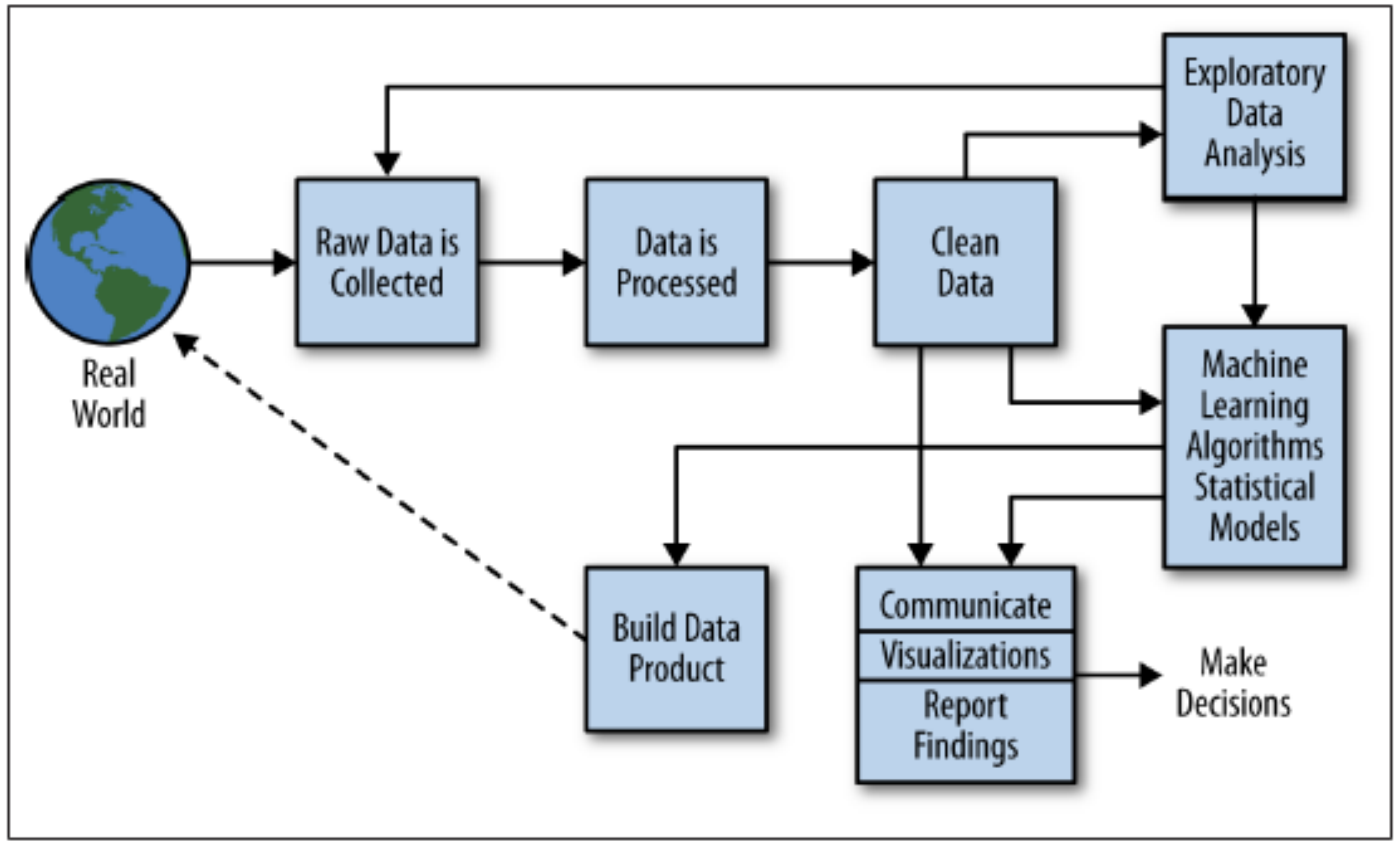
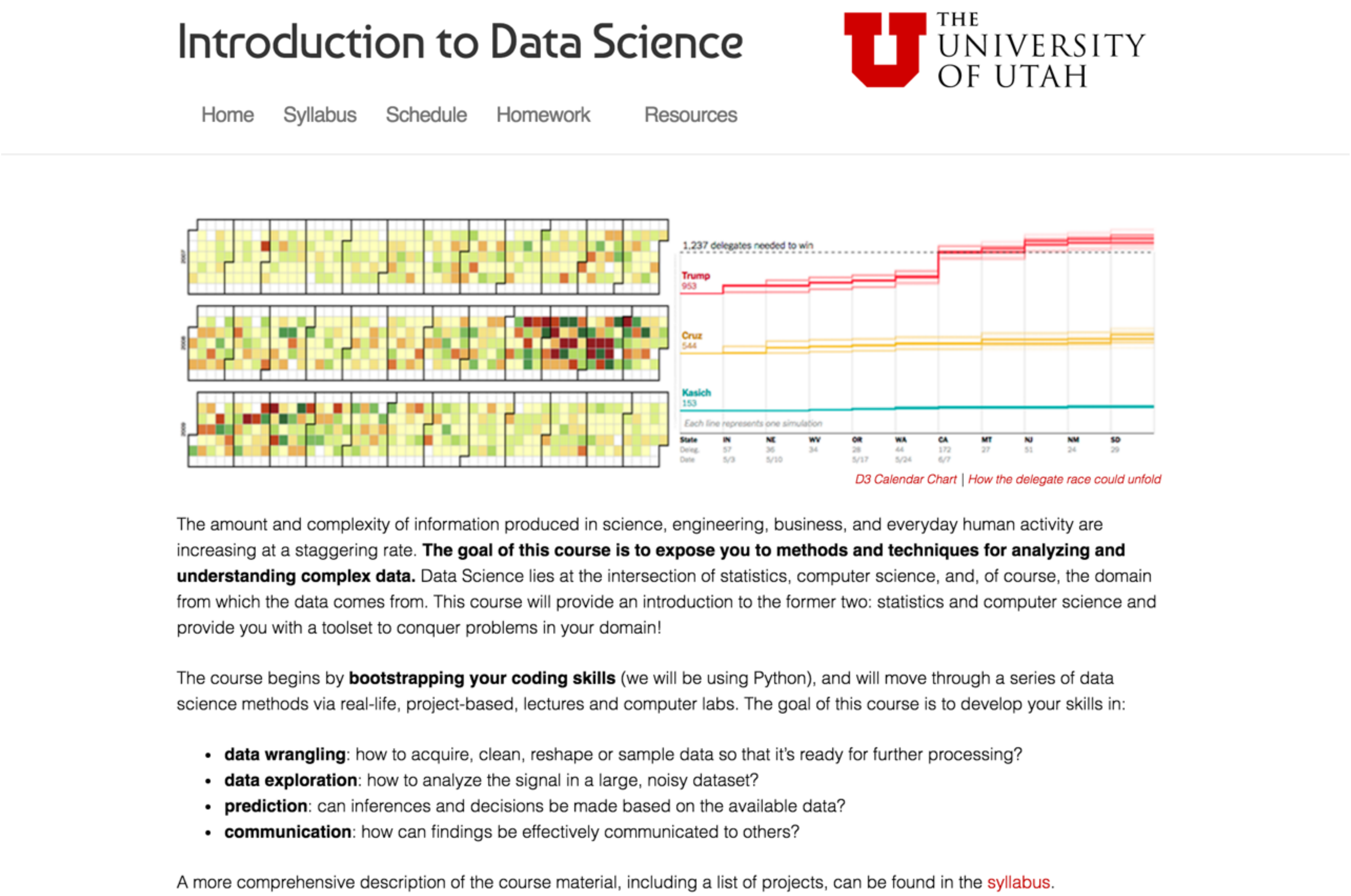

Real World
Raw Data is Collected
Data is Processed
Clean Data
Exploratory Data Analysis
Machine Learning Algorithms Statistical Models
Build Data Product
Communicate Visualizations Report Findings
Make Decisions

Figure 2-2. The data science process

Information datasciencecourse.net


1,237 delegates needed to win
Trump
953
Cruz
544
Kasich
153
Each line represents one simulation
State
IN
NE
WV
OR
WA
CA
MT
NJ
NM
SD

Introduction to Data Science

THE UNIVERSITY OF UTAH

Home Syllabus Schedule Homework Resources

D3 Calendar Chart | *How the delegate race could unfold*

The amount and complexity of information produced in science, engineering, business, and everyday human activity are increasing at a staggering rate. **The goal of this course is to expose you to methods and techniques for analyzing and understanding complex data.** Data Science lies at the intersection of statistics, computer science, and, of course, the domain from which the data comes from. This course will provide an introduction to the former two: statistics and computer science and provide you with a toolset to conquer problems in your domain!

The course begins by **bootstrapping your coding skills** (we will be using Python), and will move through a series of data science methods via real-life, project-based, lectures and computer labs. The goal of this course is to develop your skills in:

- **data wrangling**: how to acquire, clean, reshape or sample data so that it's ready for further processing?
- **data exploration**: how to analyze the signal in a large, noisy dataset?
- **prediction**: can inferences and decisions be made based on the available data?
- **communication**: how can findings be effectively communicated to others?

A more comprehensive description of the course material, including a list of projects, can be found in the syllabus.

Communicate

Canvas

- https://utah.instructure.com/courses/389967/
- Please use forum for all general questions - code, concepts, etc.
- Only use e-mail for personal inquiries

Office Hours

- Alex: Thursdays, 3:30 - 4:30, WEB 3887
- Braxton: Wednesdays, 4:00-5:00, LCB 116
- TAs: Thursdays, 3:30 - 5:30, room TBA

E-Mail

- alex@sci.utah.edu
- osting@math.utah.edu

Course Components

Lectures introduce theory, simple examples in code

Labs Short coding tutorials, longer examples

- Based on a published Jupyter notebook on website
- Strongly related to homework assignments
- Applications!

Homeworks help practice specific skills

Final Project gives you a chance to go through the complete data science process

How are you graded?

Homework Assignments: 60%

- Varying value, depending on length/difficult
- Start early!
- Due on Fridays, late days: -10% per day, up to two days.

Final Project: 40%

- Teams, two milestones

Advise: put away your devices!

No Computers, Tablets, Phones in lectures

- except when used for labs / exercises

Switch off, mute, flight mode

Why?

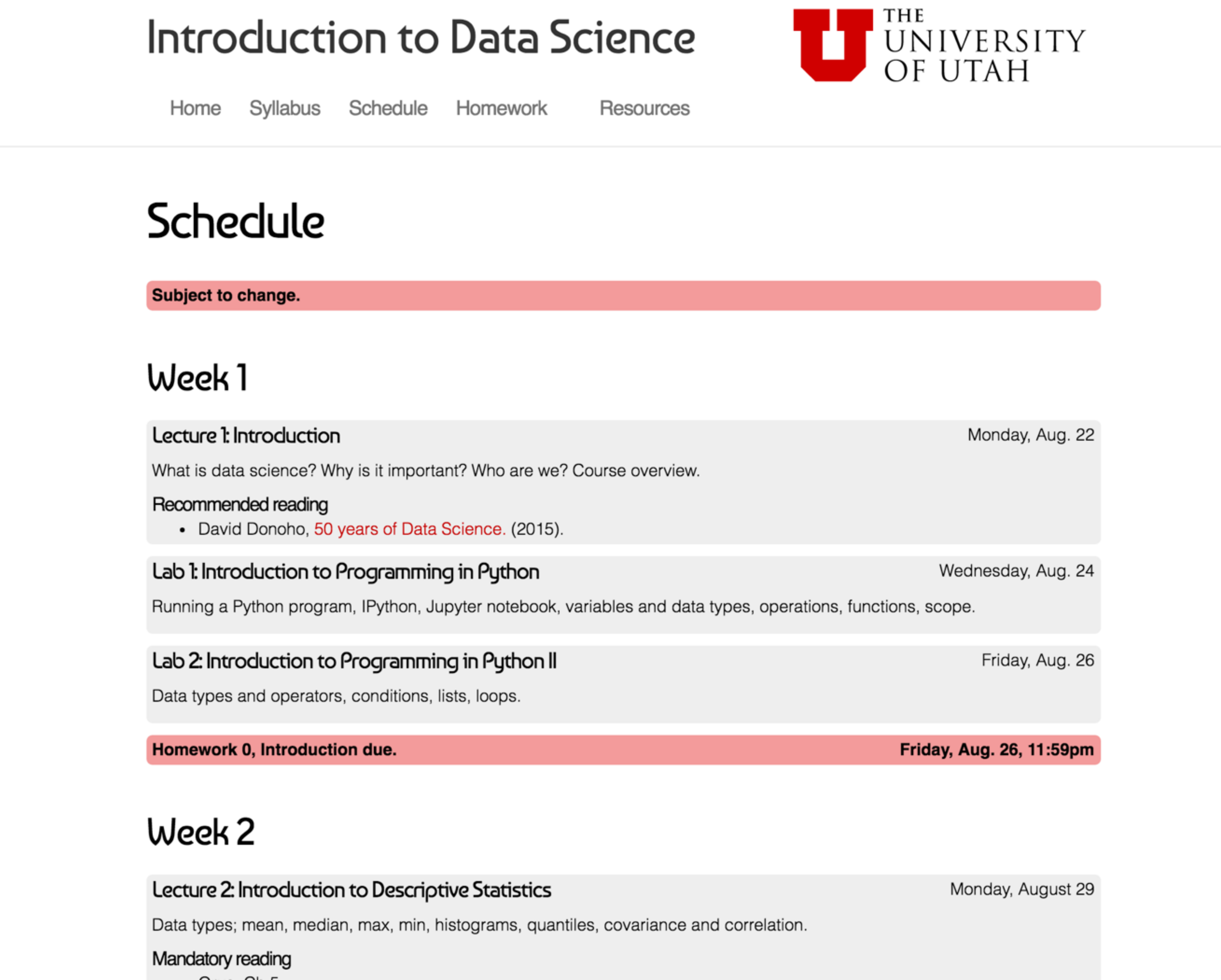
- It’s better to take note by hand
- Notifications are designed to grab your attention

Applies to Theory lectures, coding along in technical lectures encouraged

Schedule

Lectures:
MWF 3:05 - 3:55 PM
WEB L114

Labs at least once per week.
Bring your own computer!
Have Python, etc installed
(see HW0)



Schedule

Subject to change.

Week 1

Lecture 1: Introduction — Monday, Aug. 22

What is data science? Why is it important? Who are we? Course overview.

Recommended reading

- David Donoho, 50 years of Data Science. (2015).

Lab 1: Introduction to Programming in Python — Wednesday, Aug. 24

Running a Python program, IPython, Jupyter notebook, variables and data types, operations, functions, scope.

Lab 2: Introduction to Programming in Python II — Friday, Aug. 26

Data types and operators, conditions, lists, loops.

Homework 0, Introduction due. **Friday, Aug. 26, 11:59pm**

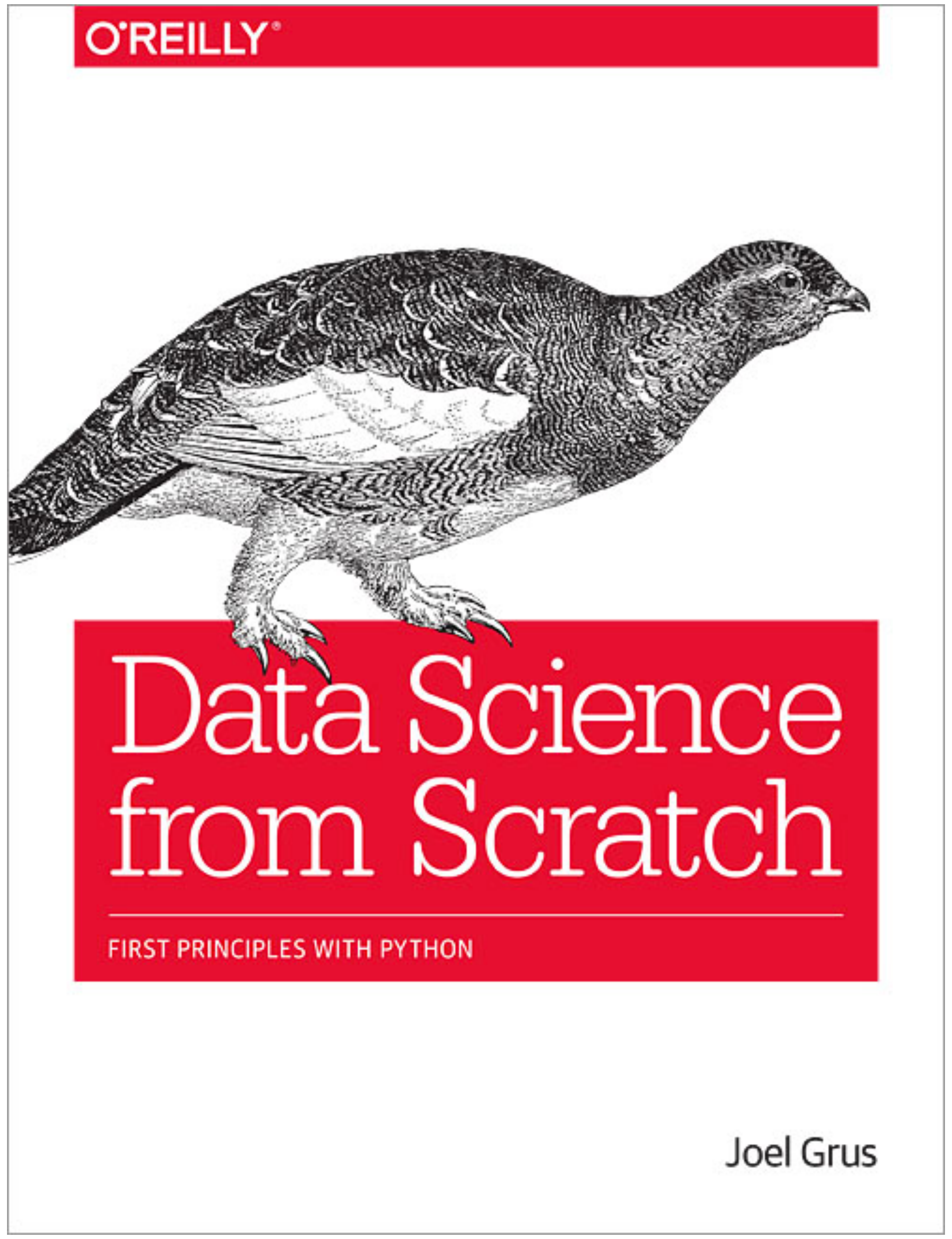
Week 2

Lecture 2: Introduction to Descriptive Statistics — Monday, August 29

Data types; mean, median, max, min, histograms, quantiles, covariance and correlation.

Mandatory reading

Books



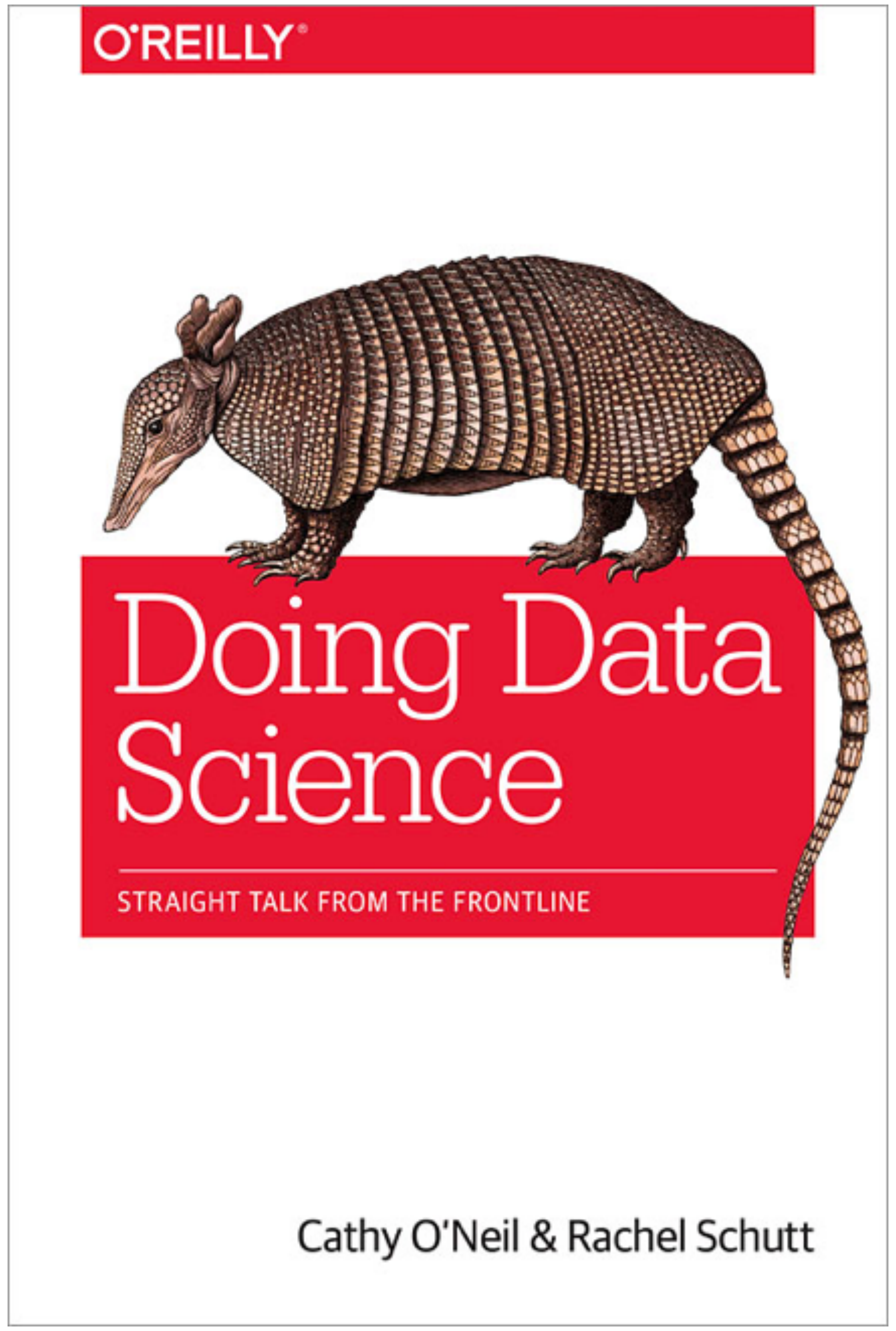
O'REILLY

Data Science from Scratch

FIRST PRINCIPLES WITH PYTHON

Joel Grus

Primary Text for Readings
Available for free on Campus:
http://proquest.safaribooksonline.com/9781491901410



O'REILLY

Doing Data Science

STRAIGHT TALK FROM THE FRONTLINE

Cathy O'Neil & Rachel Schutt

Supplementary Text

Programming

NumPy

python™

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Is this course for me ???

Prerequisites

Programming experience

- Python, C, C++, Java, etc.

Calculus 1

- UU Math 1170, 1210, 1250 1310, 1311 or equivalent

Willingness to learn new software & tools

- This can be time consuming

You will need to build skills by yourself!

- Engineering vs Computer Science

If in doubt, ask one of the instructors.

This Week

HW0, including course survey

Introduction to programming (two labs)

Readings:

Cathy O’Neil and Rachel Schutt, Doing Data Science. (2014) Chapter 1.

David Donoho, 50 years of Data Science. (2015).

Next Week

HW1 due

Introduction to Descriptive Statistics

Data Structures and Pandas

Office hours start!

About You

Enough about us! Please submit a “data science profile”

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise

1 - little knowledge 5 - Expert

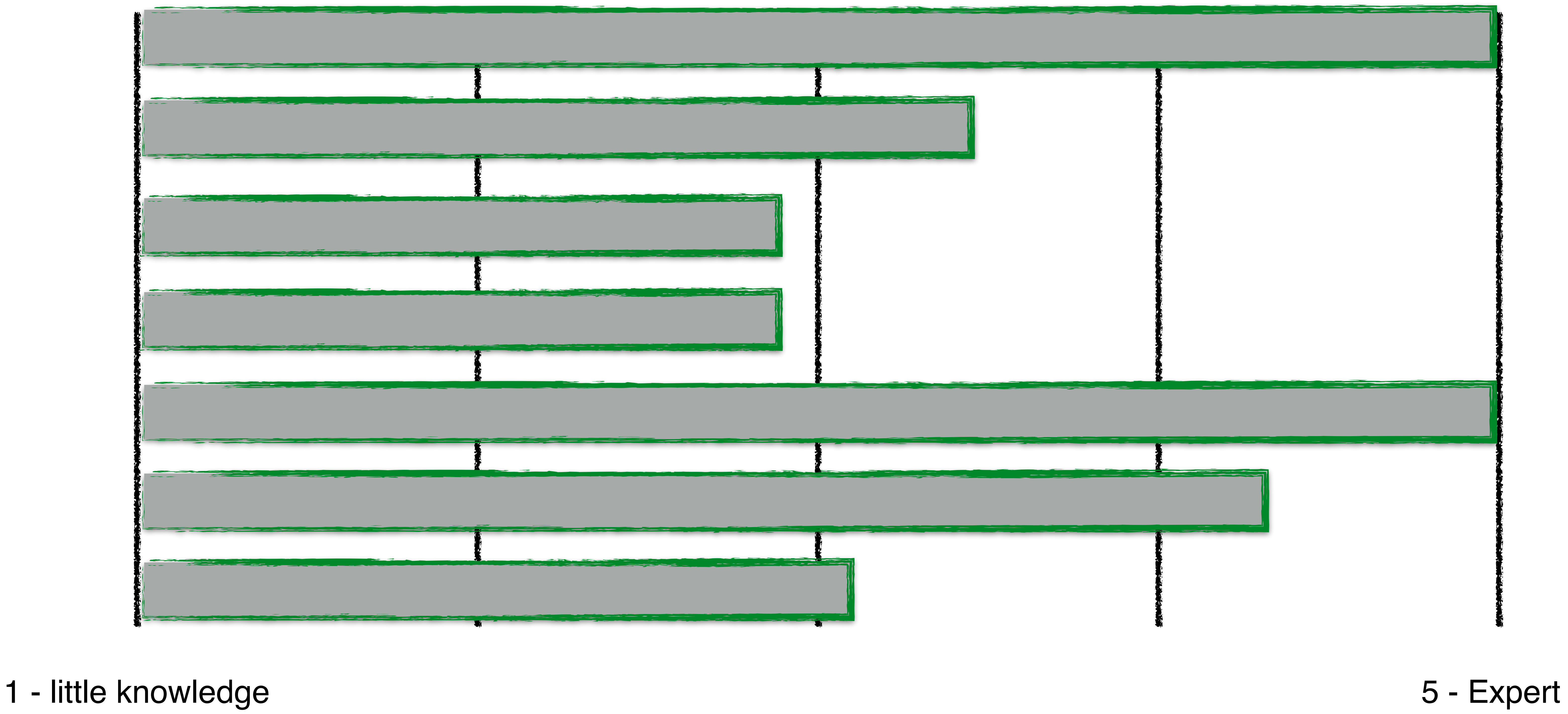
In addition, in the comments section, please write any particular subjects you'd like to see covered in class.

[O’Neil+Schutt (2013), p.10]

Alex’s Data Science Profile

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise

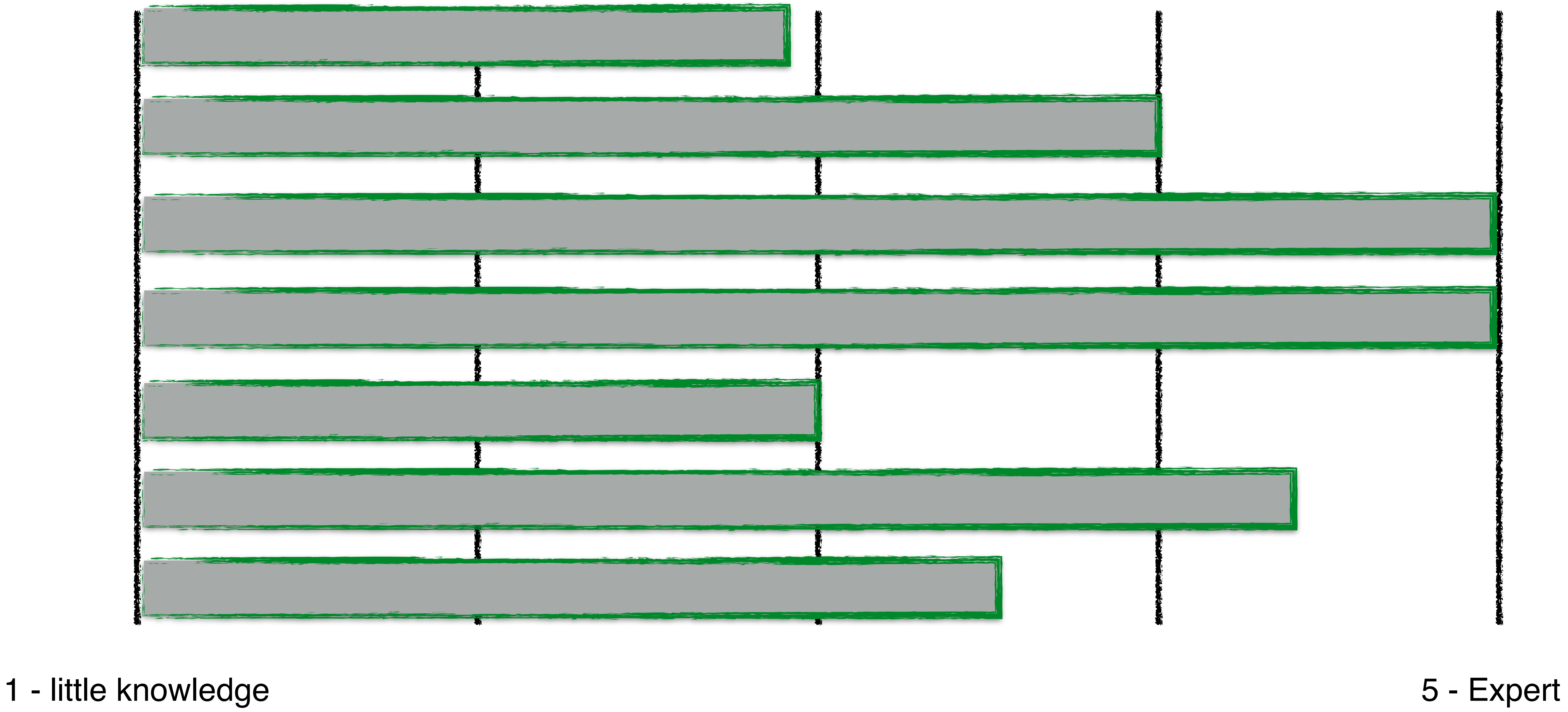


[O’Neil+Schutt (2013), p.10]

Braxton’s Data Science Profile

Please fill out this survey, rating yourself on a scale of 1-5 (5=expert) with respect to your skill level along the following seven dimensions:

1. Data Visualization
2. Machine Learning
3. Mathematics
4. Statistics
5. Computer Science
6. Communication
7. Domain Expertise



[O’Neil+Schutt (2013), p.10]