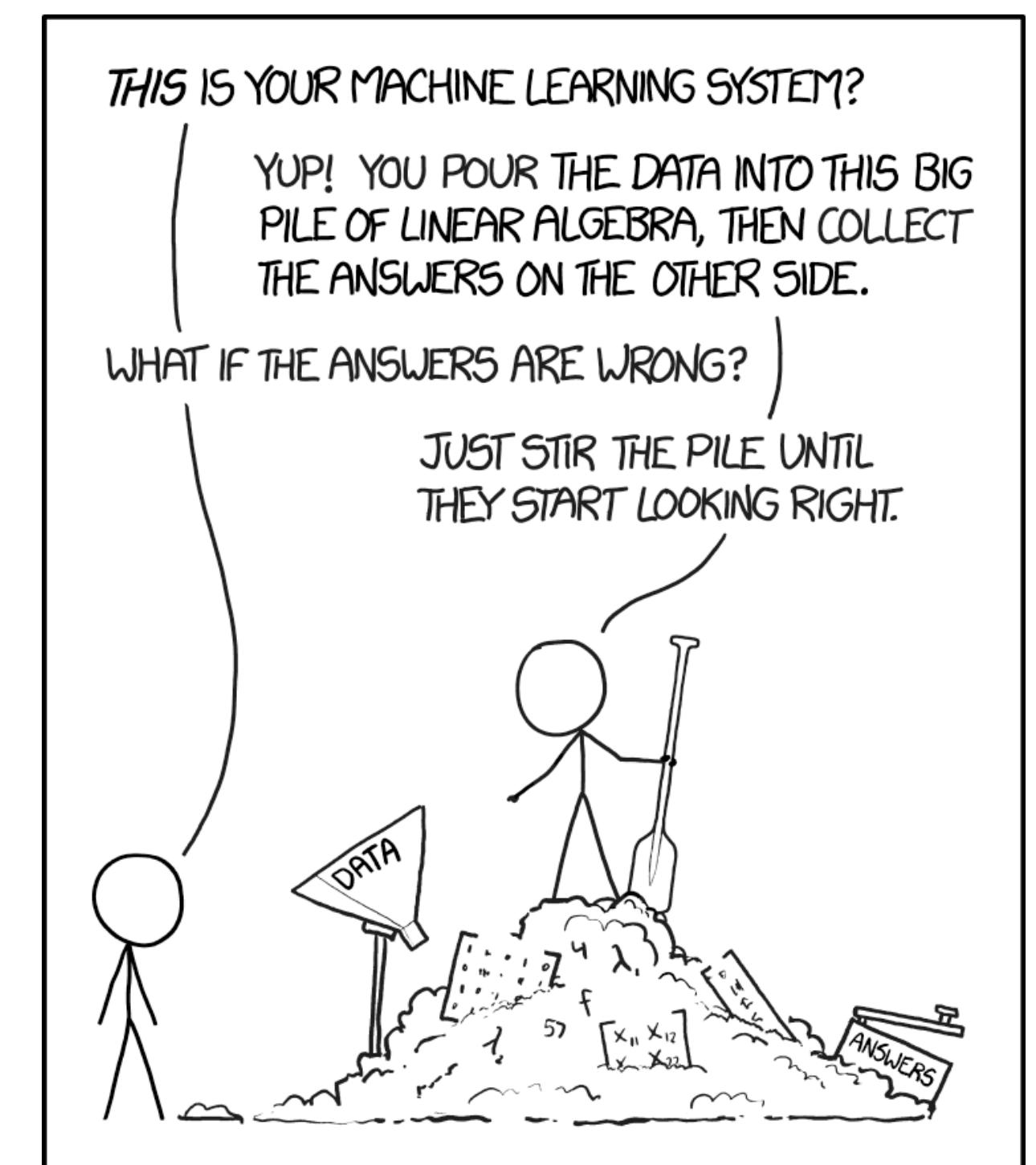


Introduction to Data Science

CS 5360 / Math 4100

Alexander Lex
alex@sci.utah.edu

Braxton Osting
osting@math.utah.edu



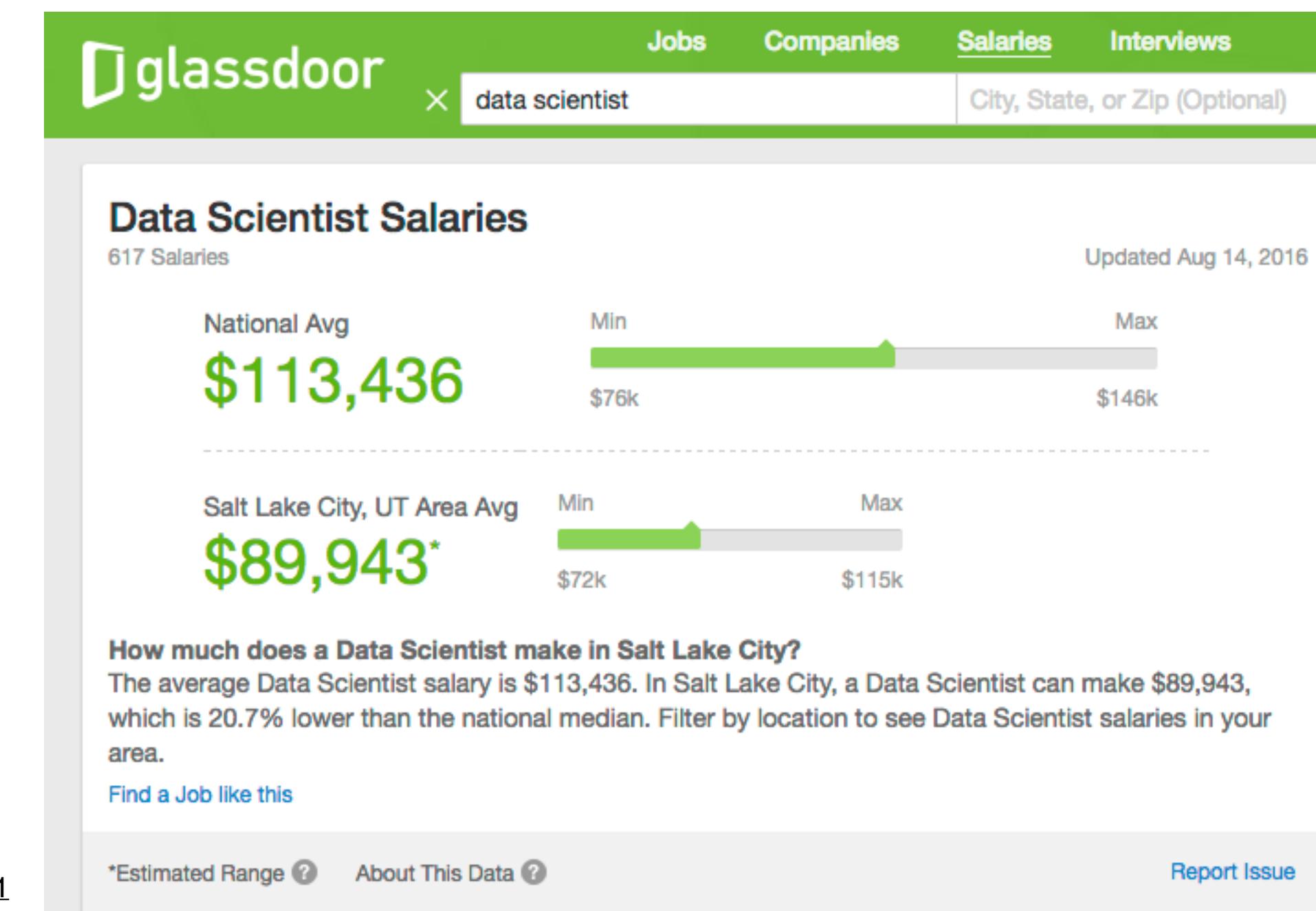
What is Data Science?

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

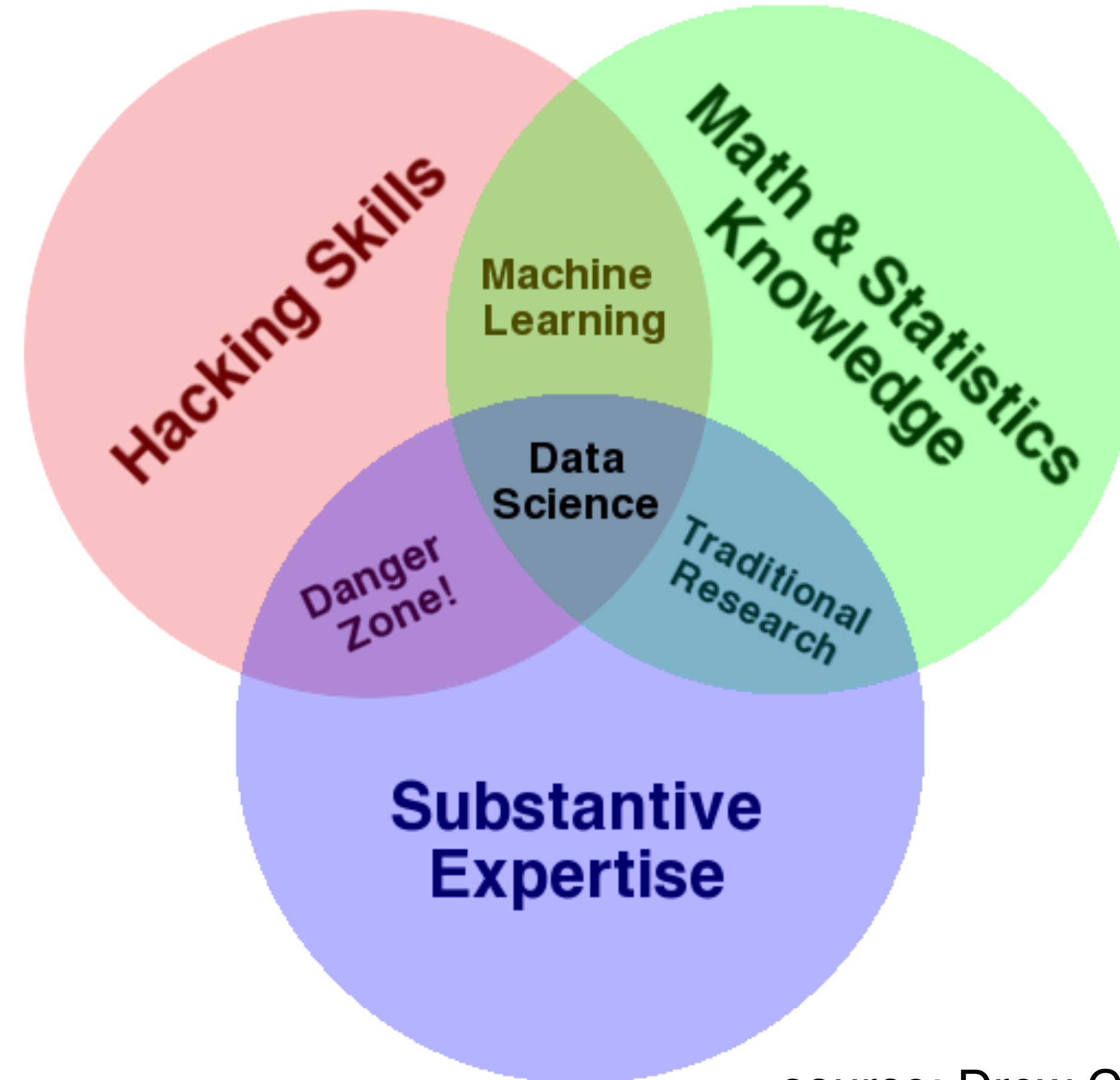
A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.



What is Data Science?



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.

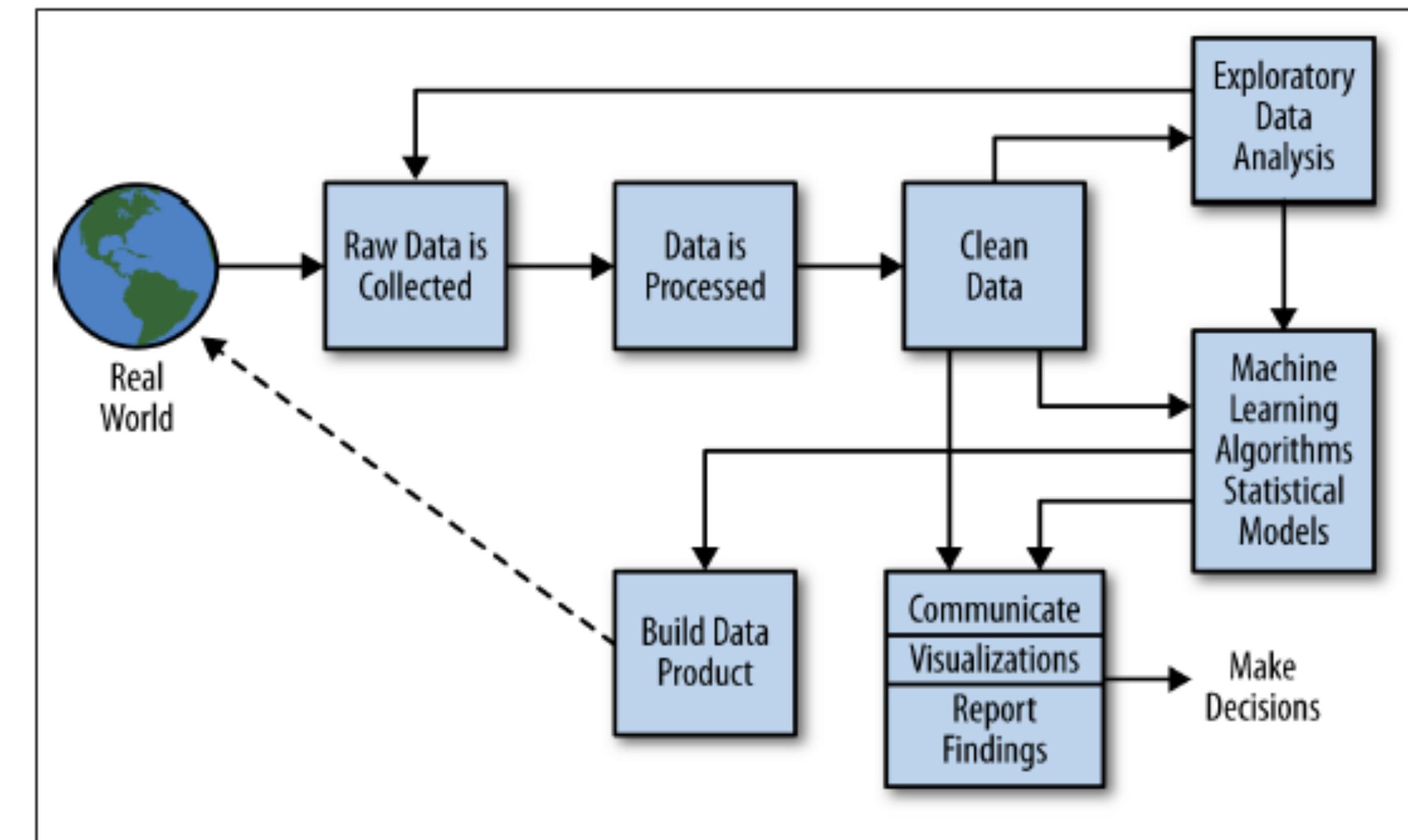


Figure 2-2. The data science process

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

Why do we care? Data is everywhere!

Biology? Data-centered & computational!

Physics? Data-centered & computational!

Medicine? Data-centered & computational!

Social Sciences? Data-centered & computational!

Business? Data-centered & computational!

Why do we care? Jobs!

CS enrollments are exploding with both a growing number of majors and non-majors.

The non-majors are wise in their choices. The recent "Rebooting Jobs" report from Burning Glass and Oracle Academy shows that CS skills are the most rapidly growing skills requested in job ads, but only 18% of those job ads ask for a CS degree.

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

A screenshot of a Google search results page. The search query "youtube cat videos" is entered in the search bar. Below the search bar, there are navigation links for "Web", "Videos", "Shopping", "Images", "News", "More", and "Search tools". A red oval highlights the text "About 593,000,000 results (0.44 seconds)" which is displayed below the search bar. The first result is a link to "TOP 10 BEST CAT VIDEOS OF ALL TIME! - YouTube" with a thumbnail image of a cat.



In one second on the Internet there are...



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

The image displays four separate screenshots illustrating various types of personal data:

- Google Timeline (Top Left):** Shows activity history for June 29, 2016, including driving (36.6 mi), cycling (11.4 mi), walking (0.2 mi), and running (11m). It also shows a "Home?" location set at 3125 Kennedy Dr, Salt Lake City, UT 84108, with options to mark it as "YES" or "NO".
- Google My Activity (Bottom Left):** A summary of activity for the day, showing a driving route from "Work (Warnock Engineering Building)" at 72 Central Campus Dr, Salt Lake City, UT 84112 to "Home". The route is 3.4 miles and takes 18 mins. Other entries include "Alpine Meadows" and "20 S 2030 E?".
- Strava Dashboard (Top Right):** Displays a ride summary for "Alexander Lex – Ride" on Saturday, August 20, 2016, at 8:54 AM. The ride was 40.7 km long, took 2:34:29, and had an elevation gain of 442m. Key metrics include 148W Avg Power, 1,372kJ Energy Output, an average speed of 15.8km/h, and a maximum speed of 74.2km/h. The ride was performed on the "Wasatch Crest Trail".
- Terrain Map (Bottom Right):** A detailed map of the Wasatch Crest Trail, showing the route's elevation profile. The trail starts near Holladay and ascends through the Mount Olympus Wilderness Area, reaching peaks like Little Water Peak and Mount Superior before descending back into the valley.

Big Data in Science and Engineering

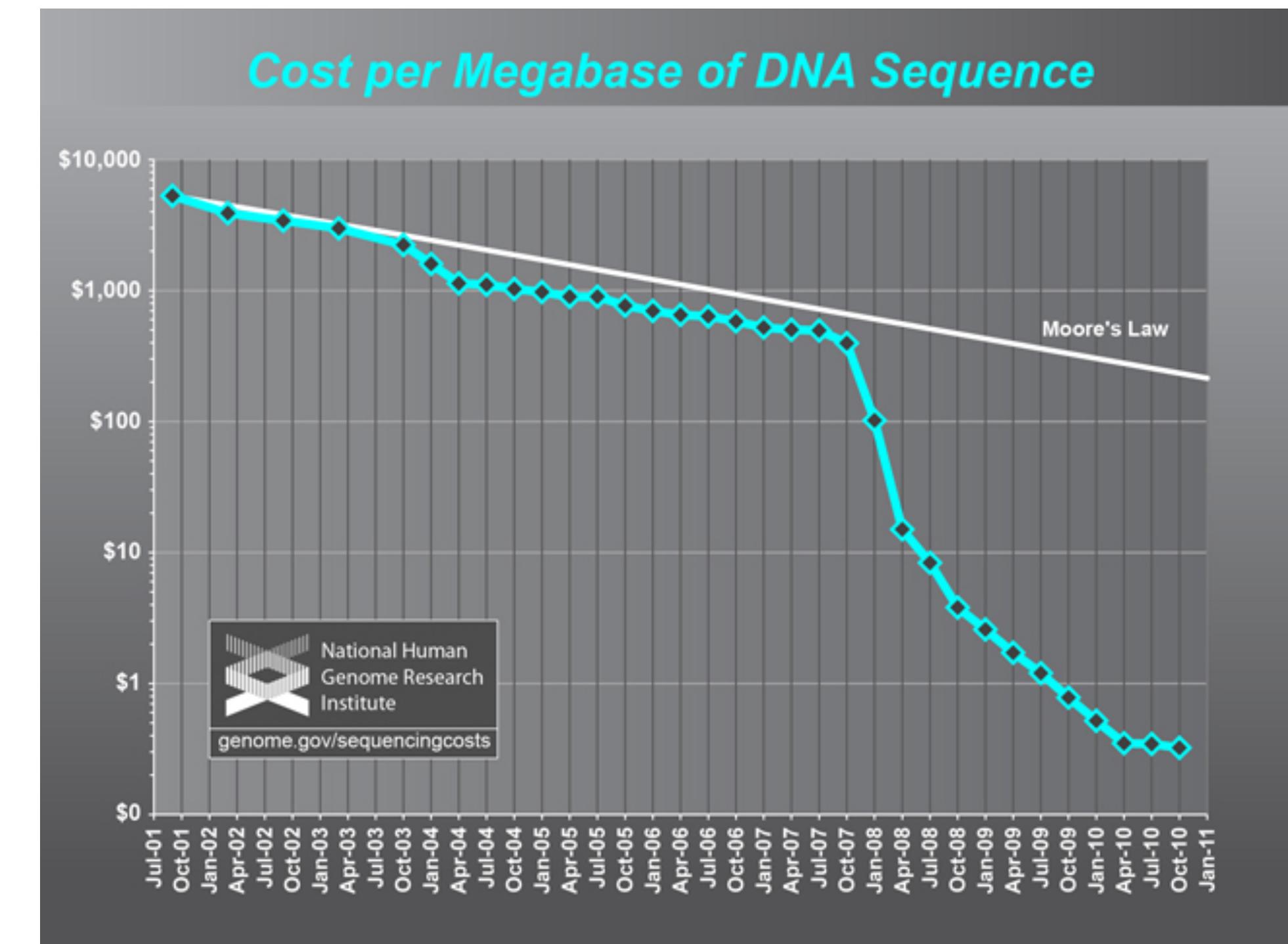
Big Data transformed science and engineering.

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?

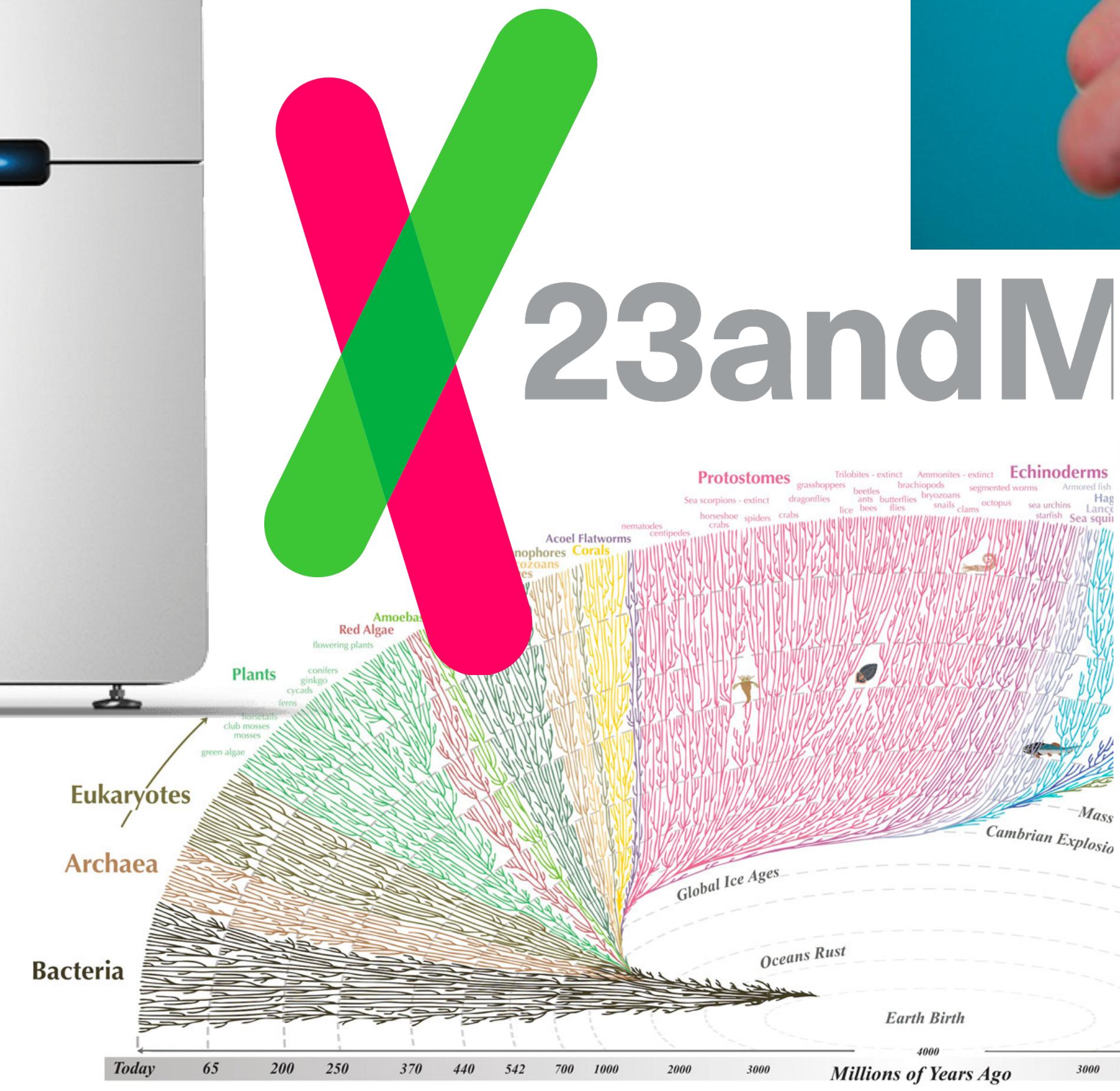
- At 15 GB of storage a piece, you'd need **20,000 Gmail accounts**. As attachments (25 MB), it would take you 12 million emails.
- A DVD-R holds 4.7 GB. You'd need **63,830 DVD-Rs, or 6,000 Blu-ray disks**.
- It takes Pandora about a day and a half to burn through a gig of mobile data. So if the CERN data was an album, you could **stream it in just over 1,230 years**.
- But its still small compared to the amount of data that the National Security Agency (NSA) 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](#))

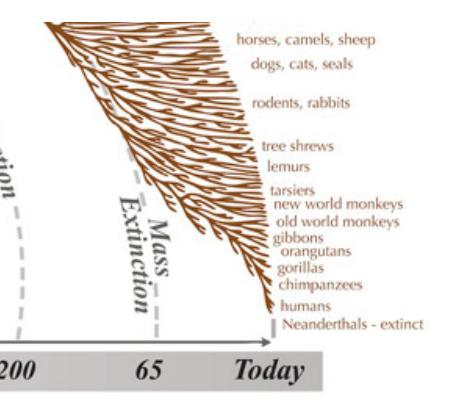
Example: Genomics



Example TCGA: 1 Petabyte



23andM



NSA Utah Data Center (Bluffdale, Utah)

Storage Capacity?

estimates vary, but NPR estimates the center will be able to handle 5 zettabytes (5 billion terabytes)



Where can you find data?

Today, a lot of data is publicly available. You probably have access to data that 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



[Home](#) [Syllabus](#) [Schedule](#) [Project](#) [Fame](#) [Resources](#)

Resources

Python

Highly Recommended Tutorials

[Learn Python the Hard Way](#)
[Code Academy](#)
[Python Cheat Sheet](#)
[Pandas Cheat Sheet](#)

Official Documentation / Resources

Data Sources

[Data.gov](#)
[Utah Data Census.gov](#)
[U.S. Bureau of Economic Analysis](#)
[Stanford Large Network Dataset C](#)
[UCI Machine Learning Repository](#)
[Dataverse Network](#)
[Infochimps](#)
[Linked Data](#)
[Guardian DataBlog](#)
[Data Market](#)
[Reddit Open Data](#)
[Climate Data Sources](#)
[Climate Station Records](#)
[CDC Data](#)
[World Bank Catalog](#)
[Free SVG Maps](#)
[UK Office for National Statistics](#)
[StateMaster](#)
[Wolfram Alpha](#)

**Who is
CS 5360 / Math 4100?**

Alexander Lex

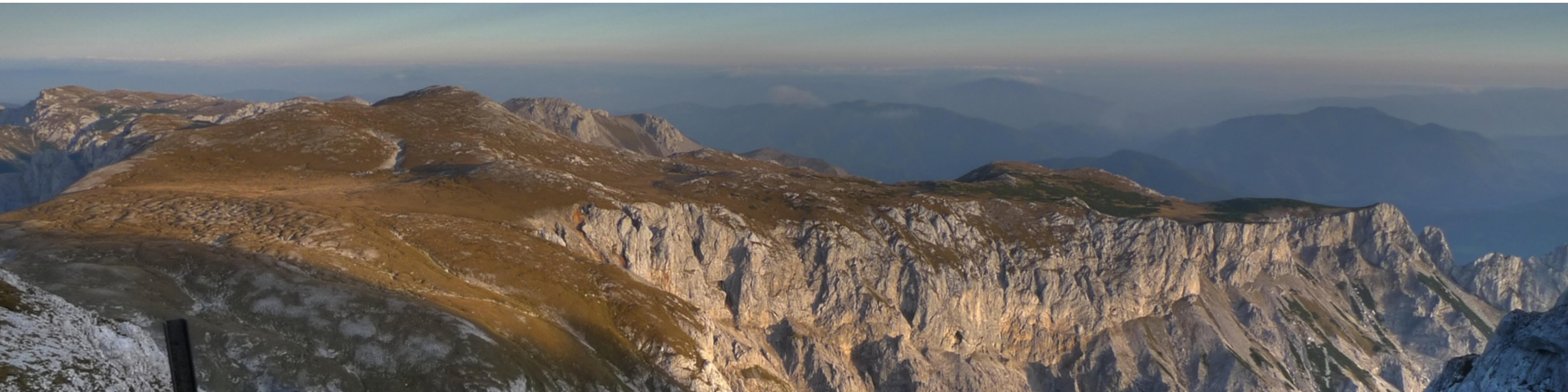
[@alexander_lex](https://twitter.com/alexander_lex)
<http://alexander-lex.net>
<http://vdl.sci.utah.edu>



Assistant Professor, Computer Science

Before that: Lecturer, Postdoctoral Fellow, Harvard

PhD in Computer Science, Graz University of Technology





Faculty

Miriah Meyer

..... Visualization Design Models, Design Studies, User-centered Design

Alexander Lex

..... Biology Visualization, Multivariate Graphs, Visualization Tools, Exploratory Visualization for Scientists

Post-Doctoral Fellows

Pascal Goffin

..... User-Centered Visualization Design, Text Visualization, Word-Scale Visualization, Visualization Tools

PhD Students

Alex Bigelow

..... Visualization Design Toolkits, Graph Data, User-centered Design

Ethan Kerzner

..... Visualization for Scientists and Engineers

Nina McCurdy

..... Visualization of multivariate, spatial + abstract data, Visualization in the Digital Humanities, Computational Creativity.

Jimmy Moore

..... Time Series Visualization, User Interface Design, Web Programming, Human-Computer Interaction

Carolina Nobre

..... Multivariate Graphs, Genealogies, Social Networks

P. Samuel Quinan

..... Uncertainty Visualization, User-Centered Design, Cognition / Decision-Making, Color

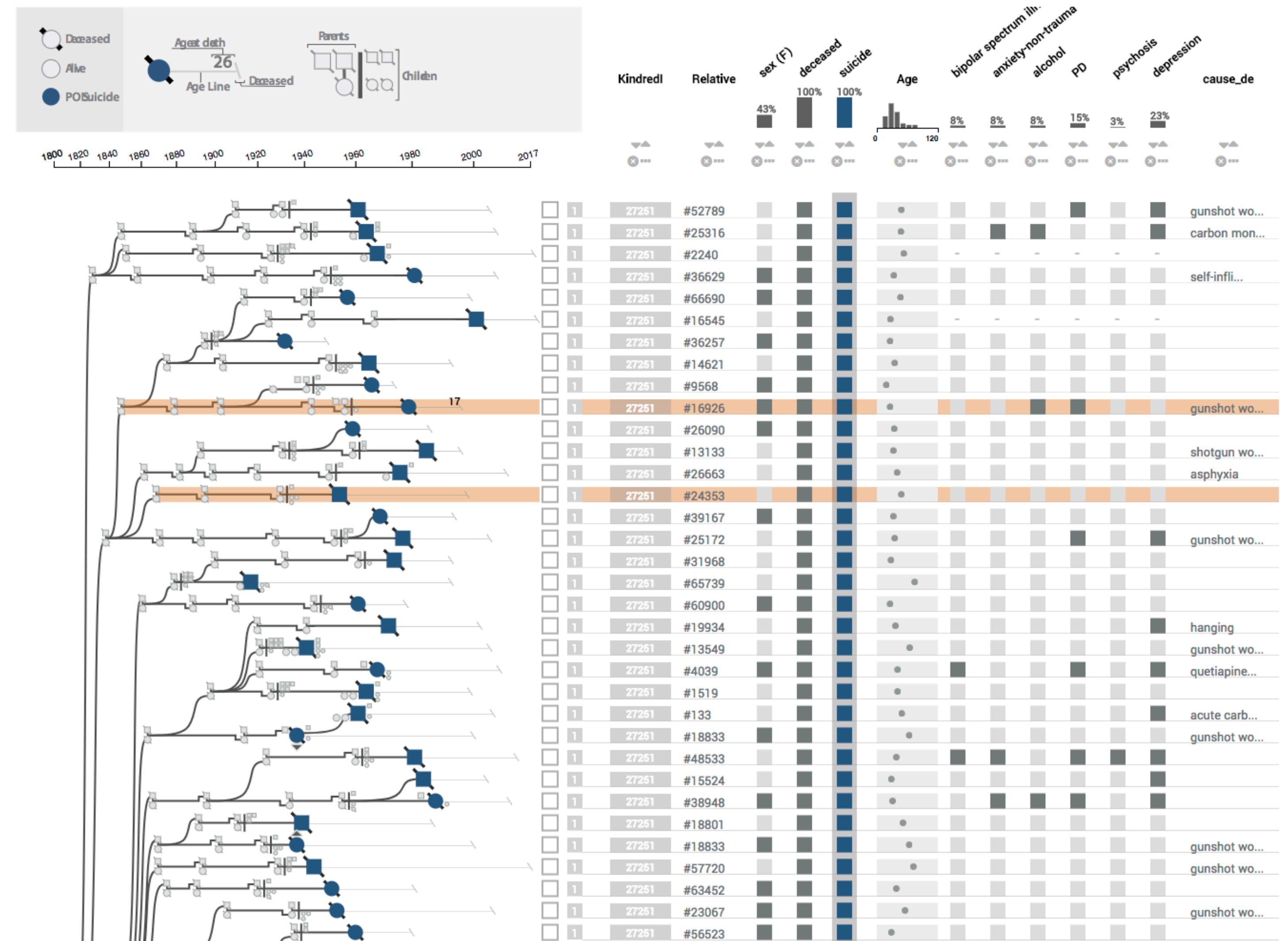
Jen Rogers

..... Interactive visualization for biological and medical data.

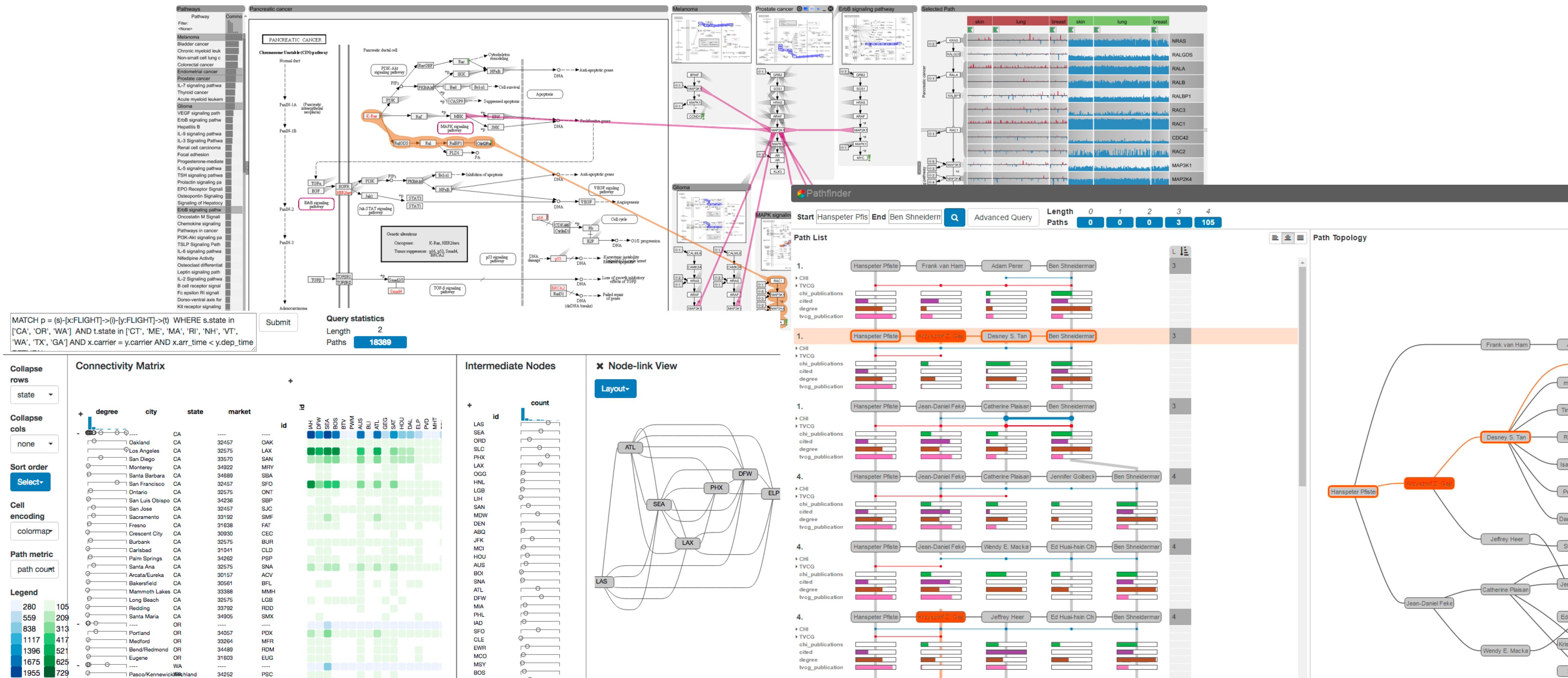
T Cameron Waller

..... Visualization of Biological Networks, Metabolism

Clinical Genealogies

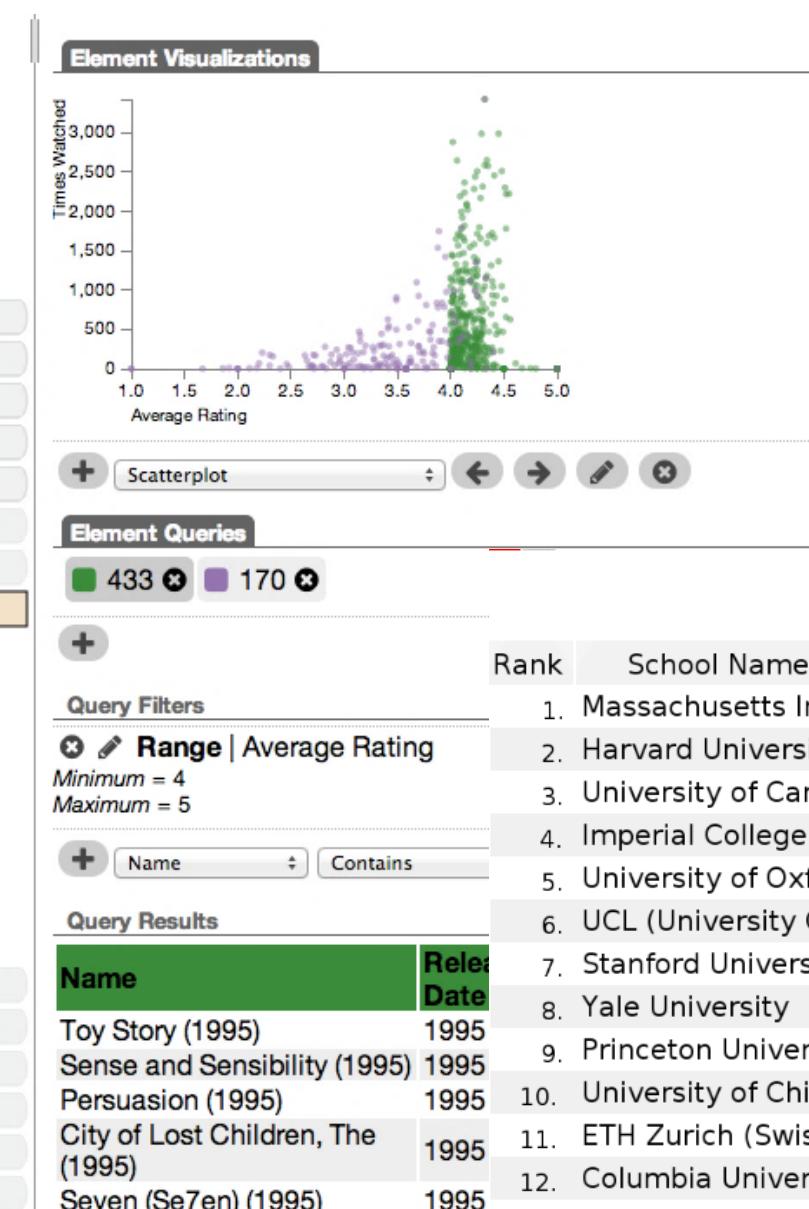
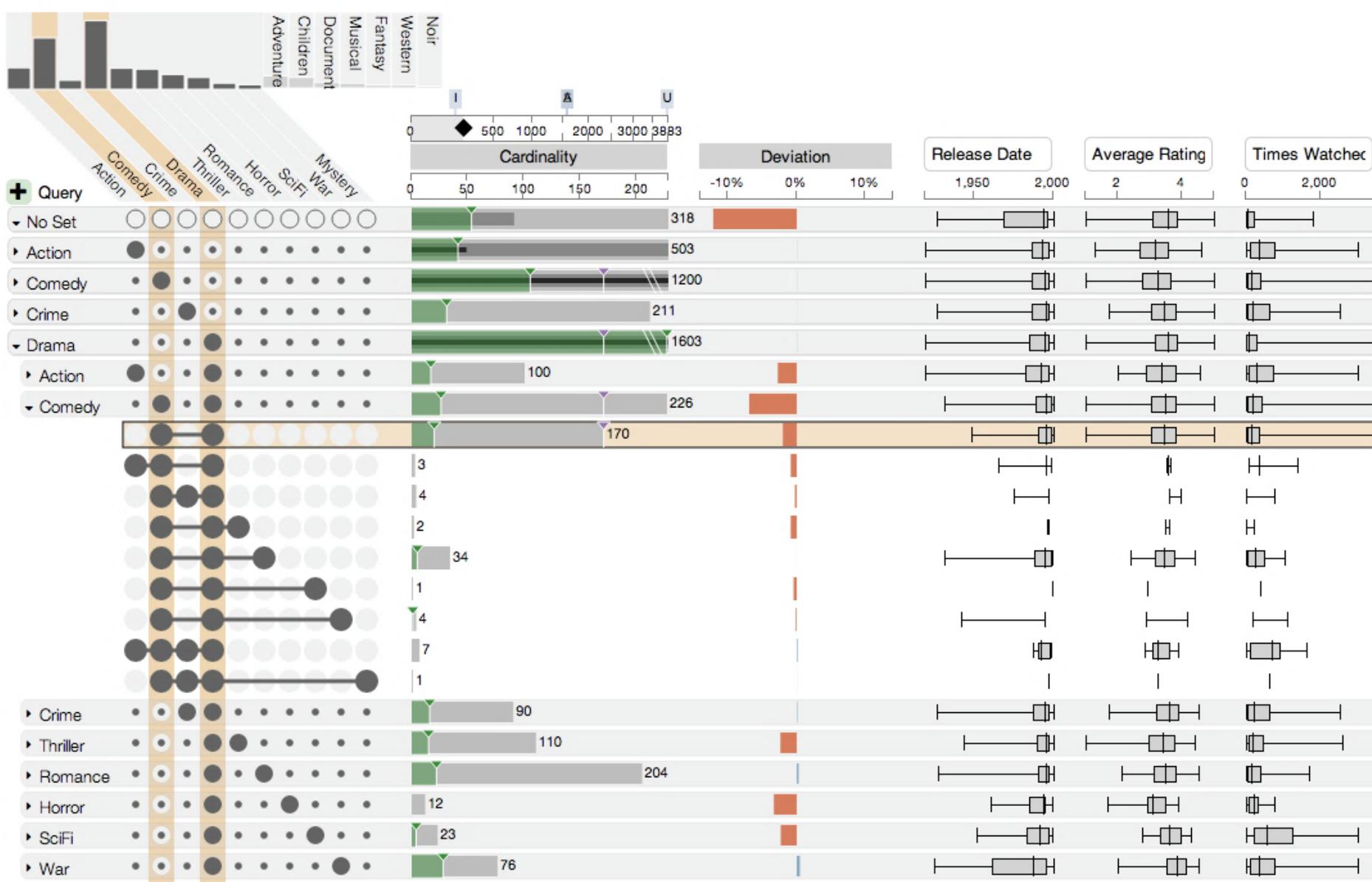


Large, Multivariate (Biological) Networks



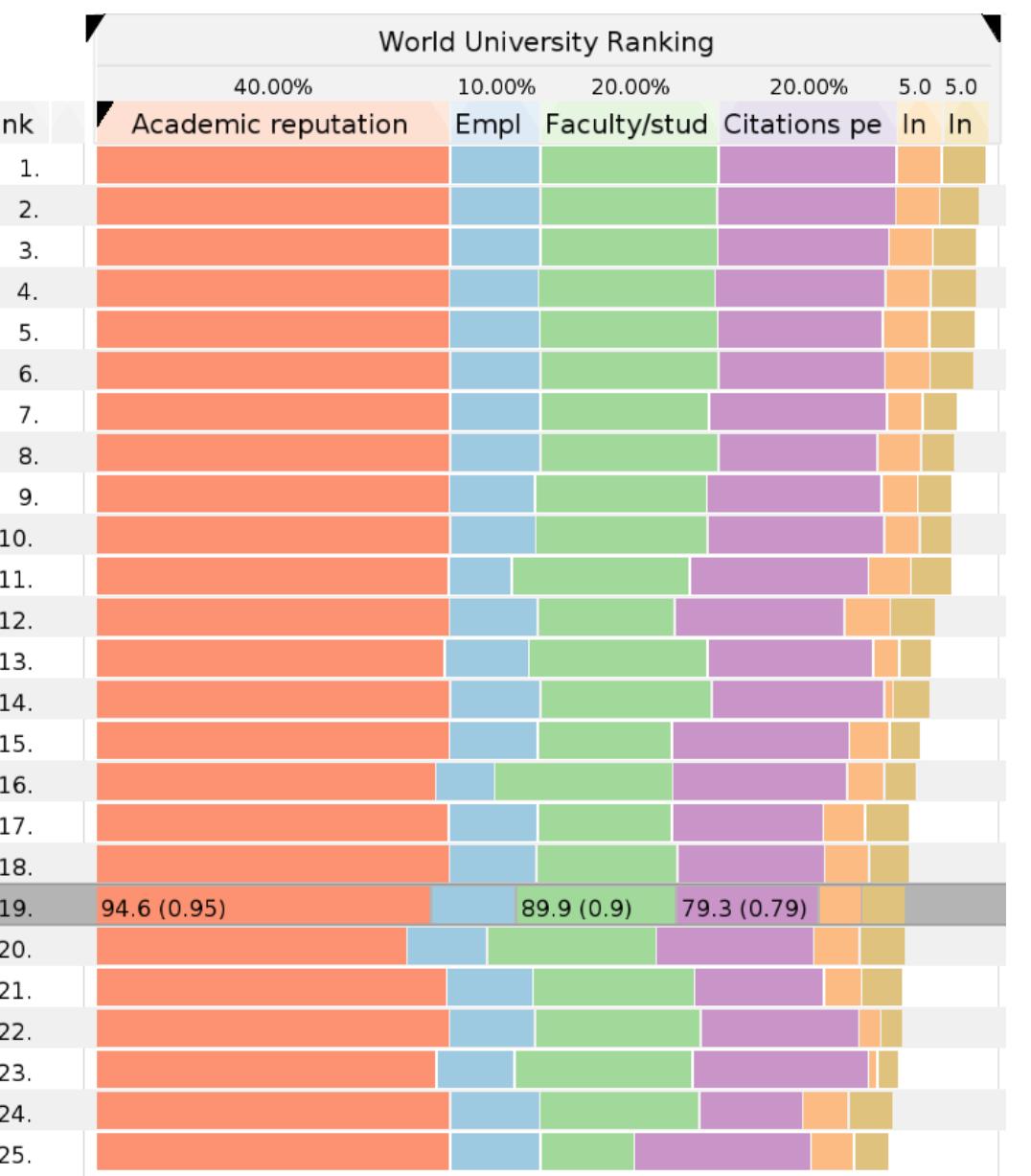
Multidimensional Data

Set Visualization



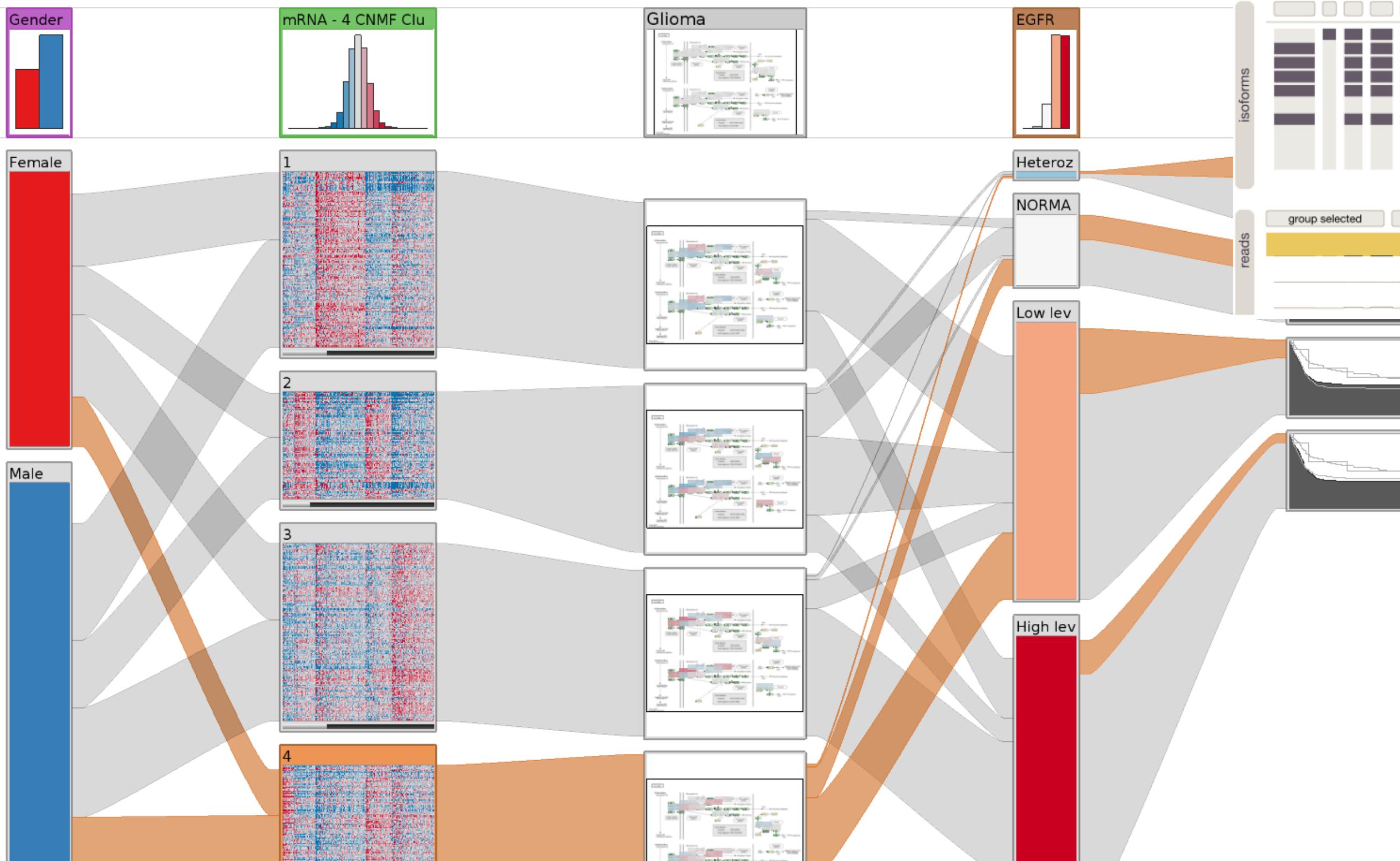
World University Ranking				
	Rank	School Name	Country	Acade
Toy Story (1995)	1995	1. Massachusetts Insti	United States	17.99%
Sense and Sensibility (1995)	1995	2. Harvard University	United States	32.94%
Persuasion (1995)	1995	3. University of Camb	United Kingdom	19.63%
City of Lost Children, The (1995)	1995	4. Imperial College L	United Kingdom	19.63%
Seven (Se7en) (1995)	1995	5. University of Oxfor	United Kingdom	4. 4.
		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	93.7 (0.94)
		18. University of Toron	Canada	
		19. McGill University	Canada	
		20. National Universit	Singapore	
		21. University of Michi	United States	
		22. University of Califfo	United States	
		23. California Institute	United States	
		24. University of Bristol	United Kingdom	
		25. Duke University	United States	

Multivariate Rankings

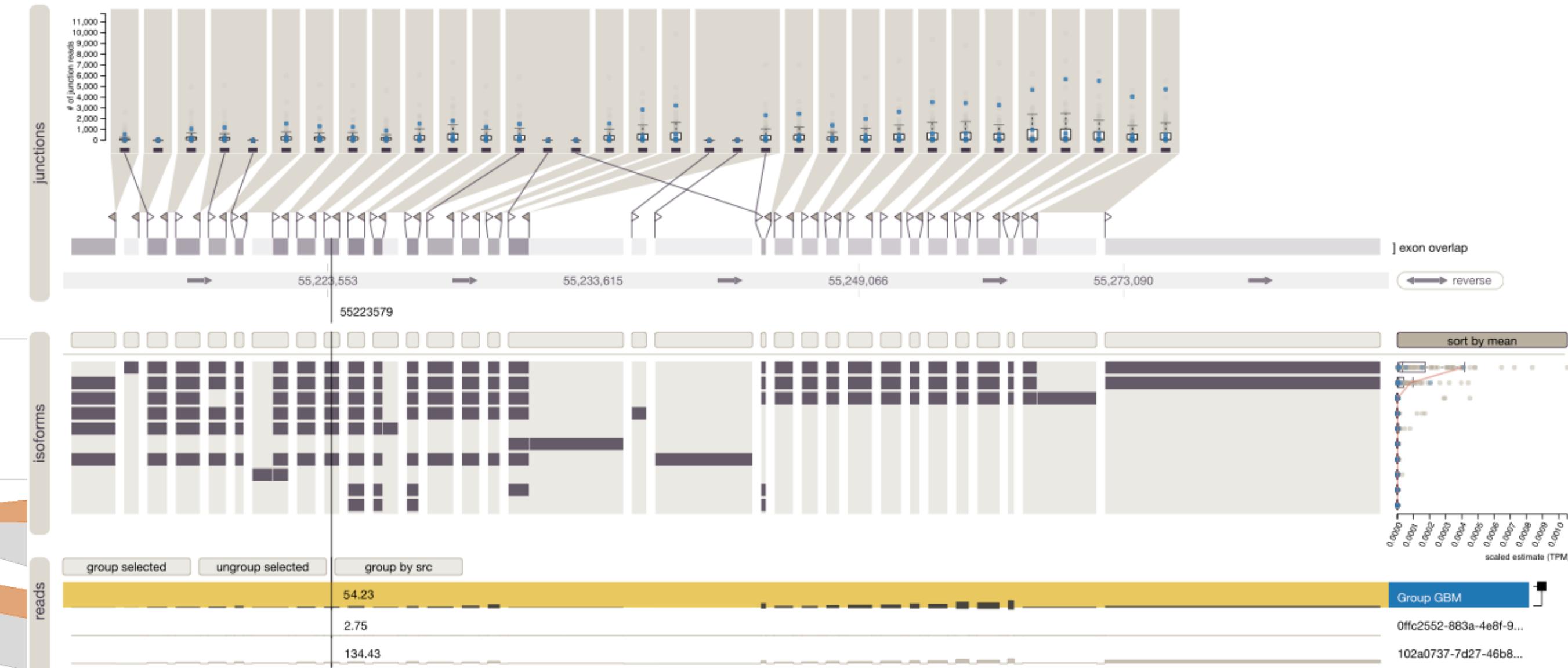


Genomic Data

Cancer Subtypes / Omics Clustering and Stratification



Alternative Splicing / mRNA-seq



Braxton Osting

Assistant Professor, Mathematics

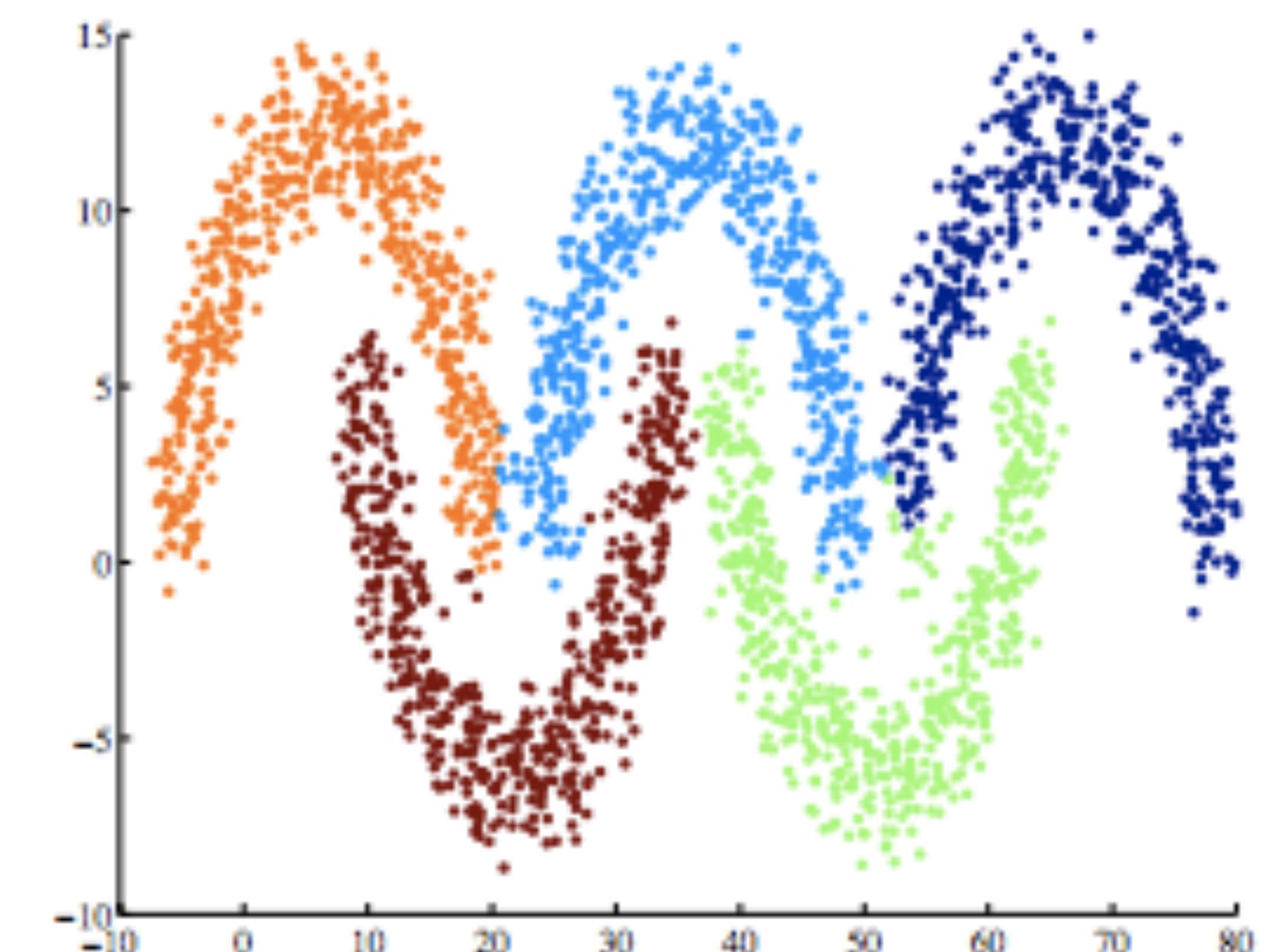
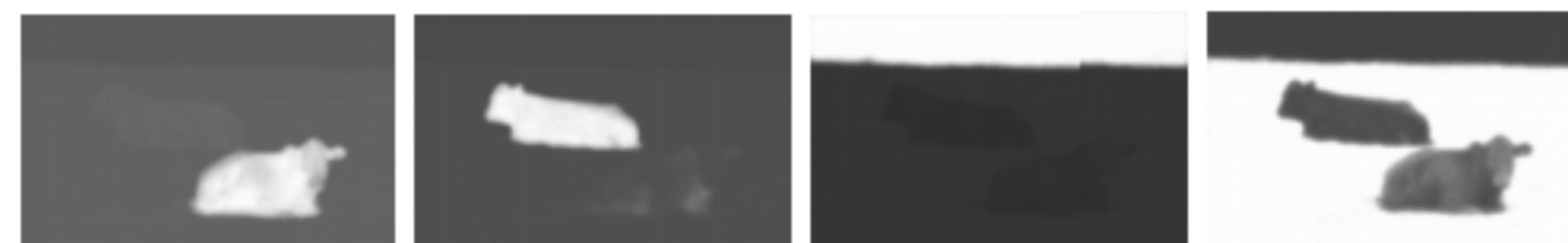
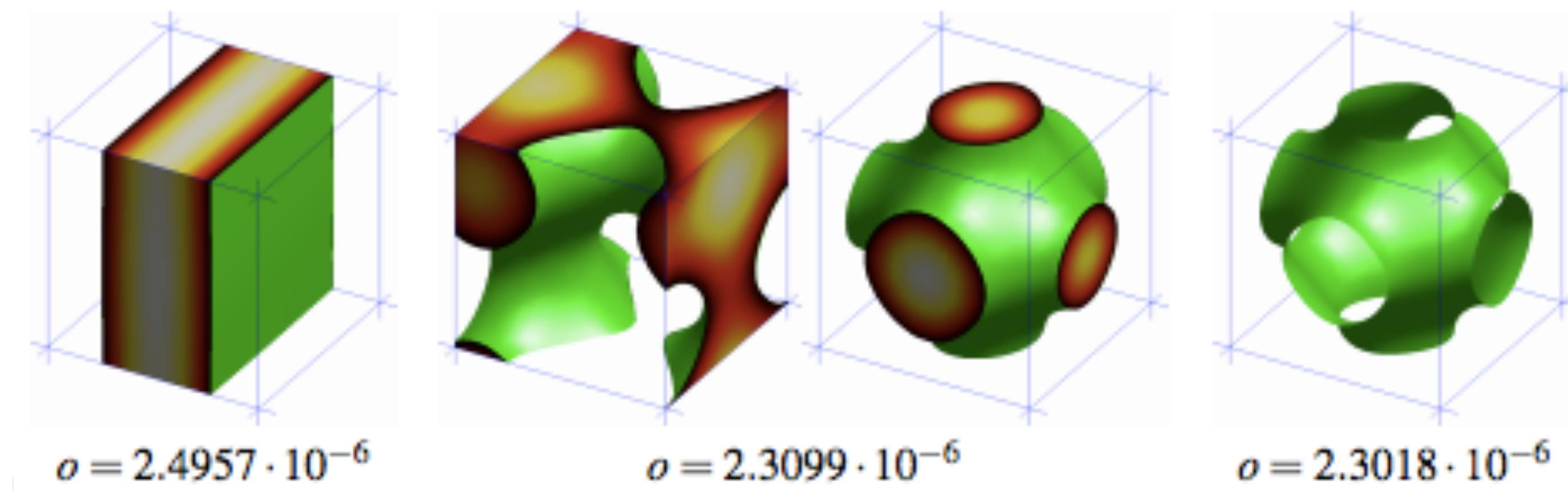
Before that: Postdoctoral Fellow, UCLA

PhD in Applied Mathematics, Columbia University



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

Partitioning, Clustering, and Image Segmentation



Statistical Ranking and Active Learning

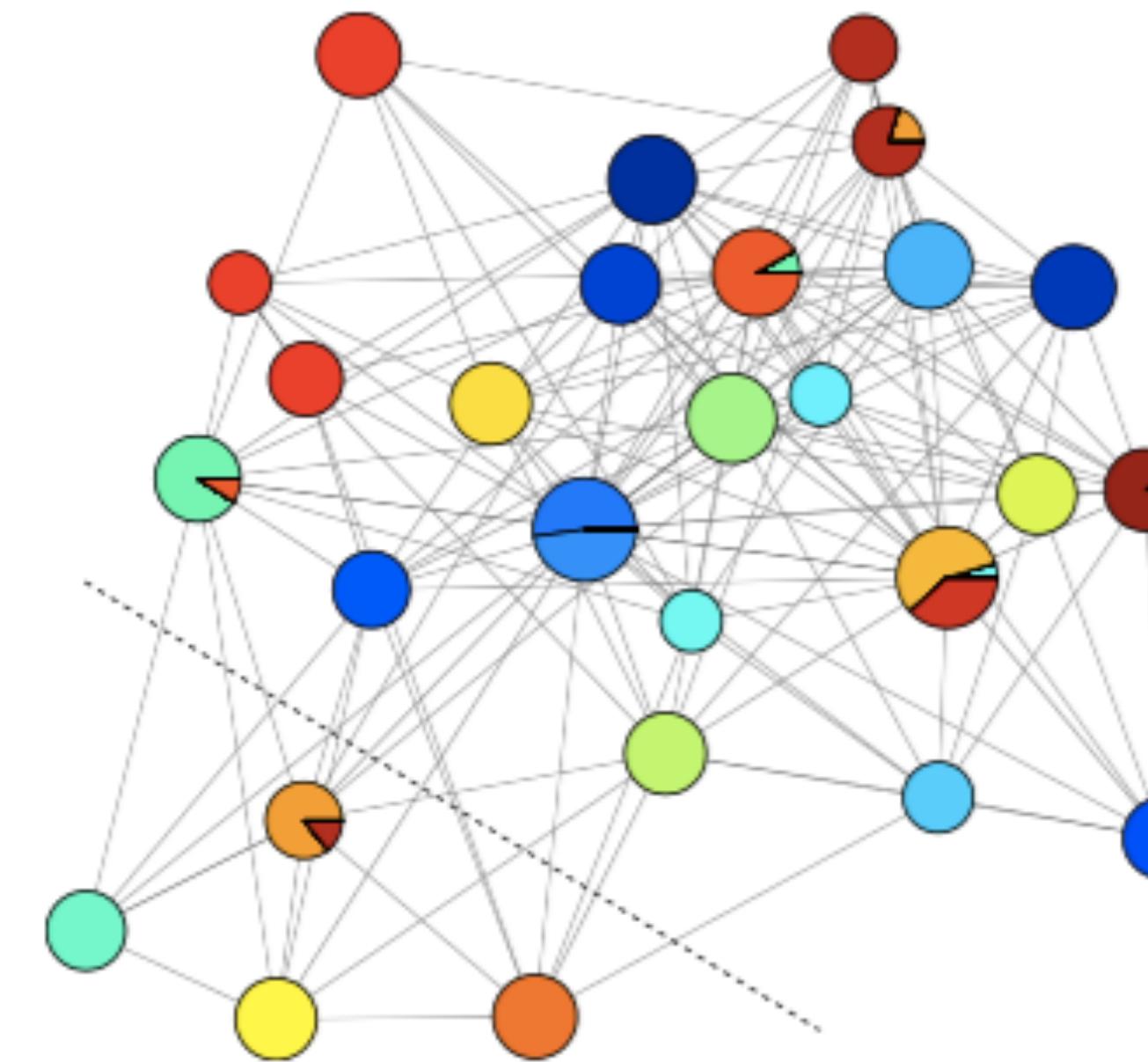
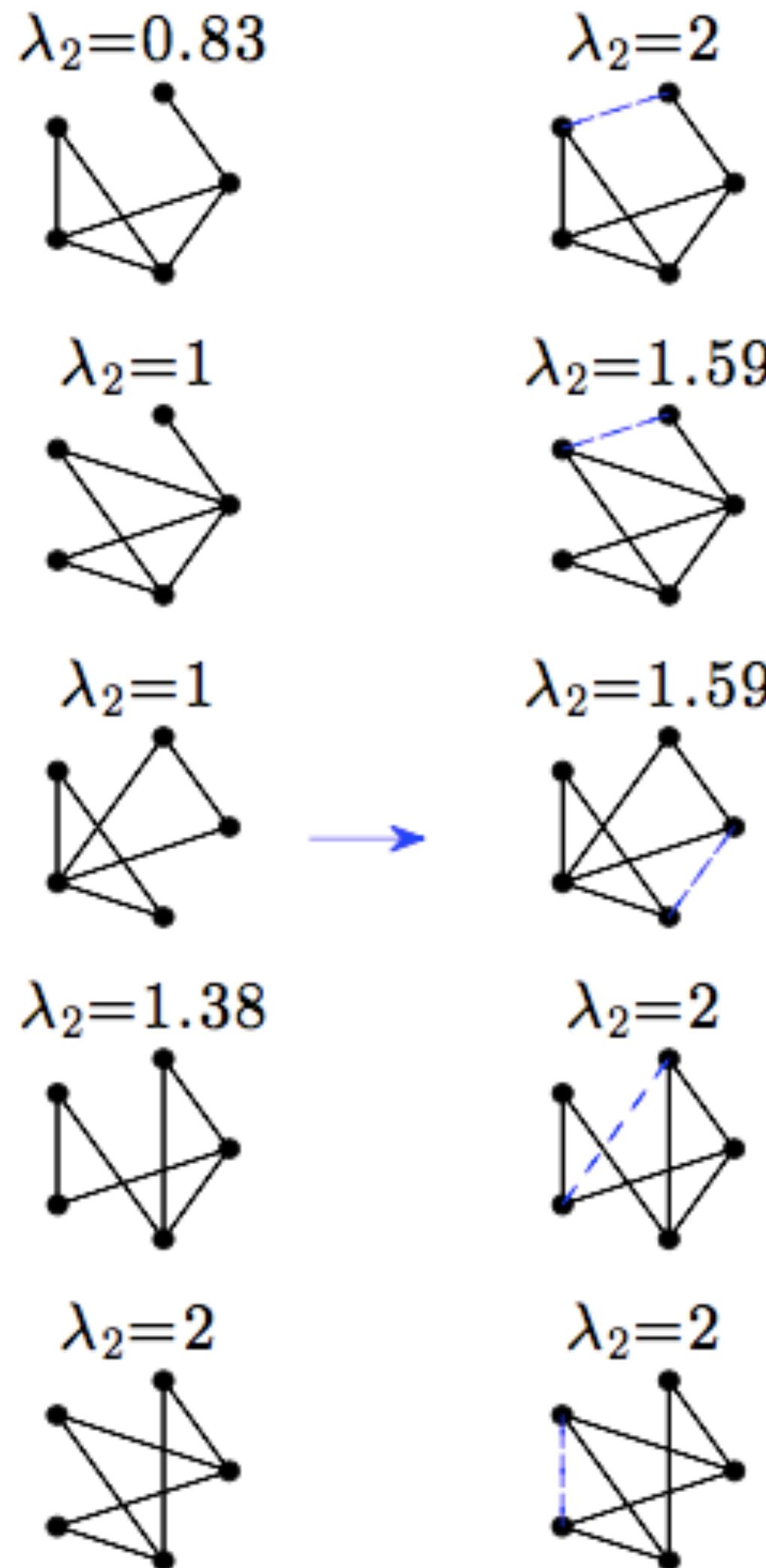
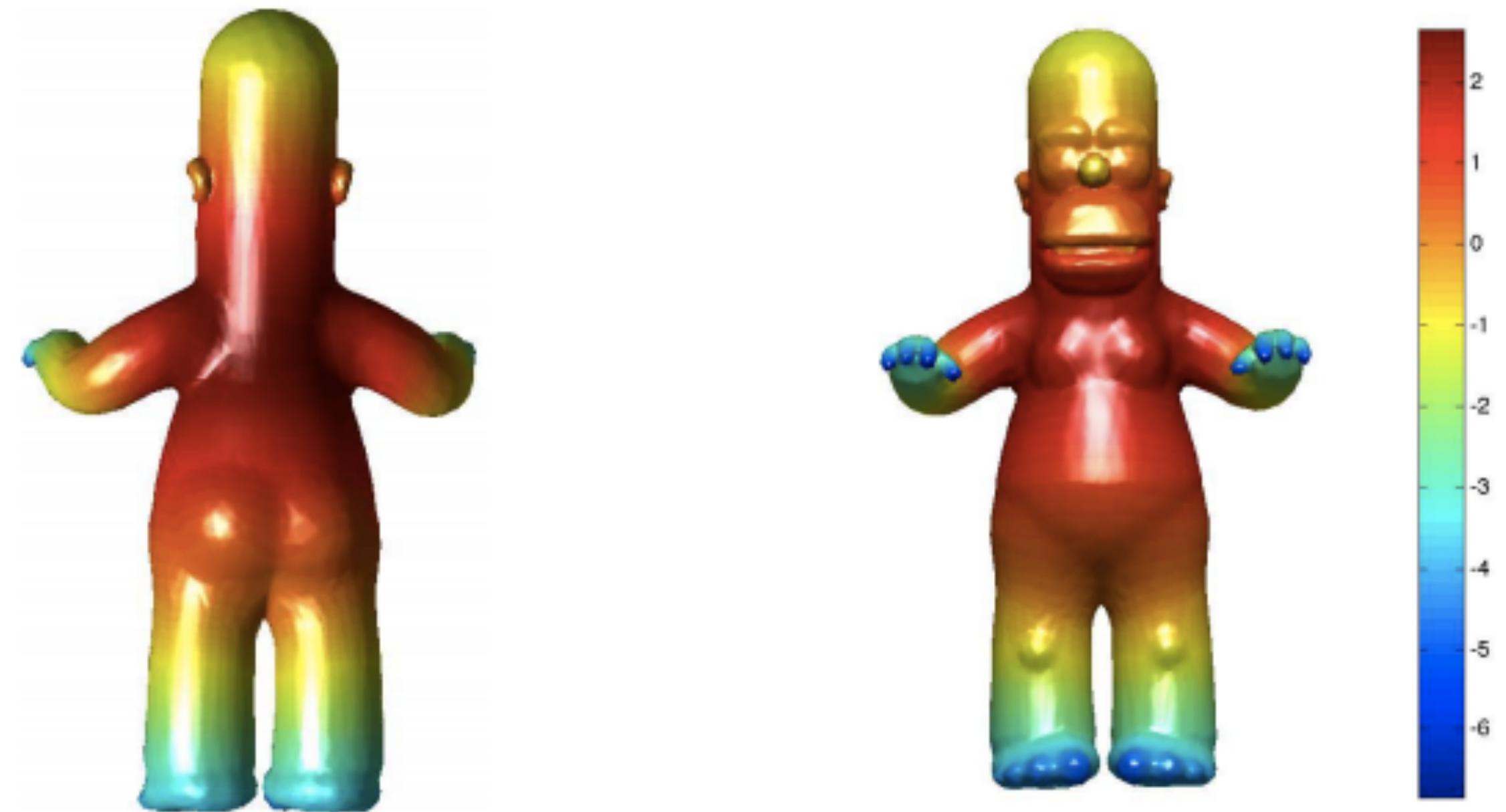
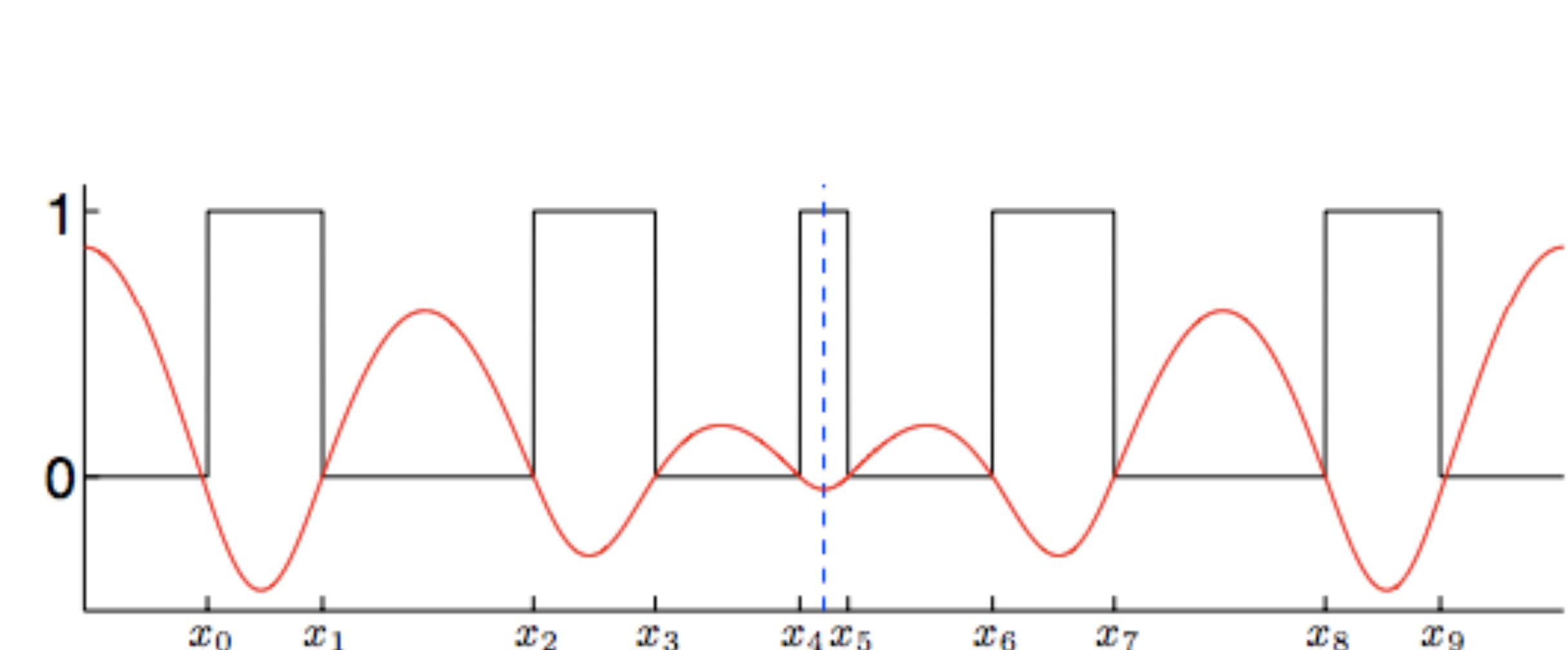
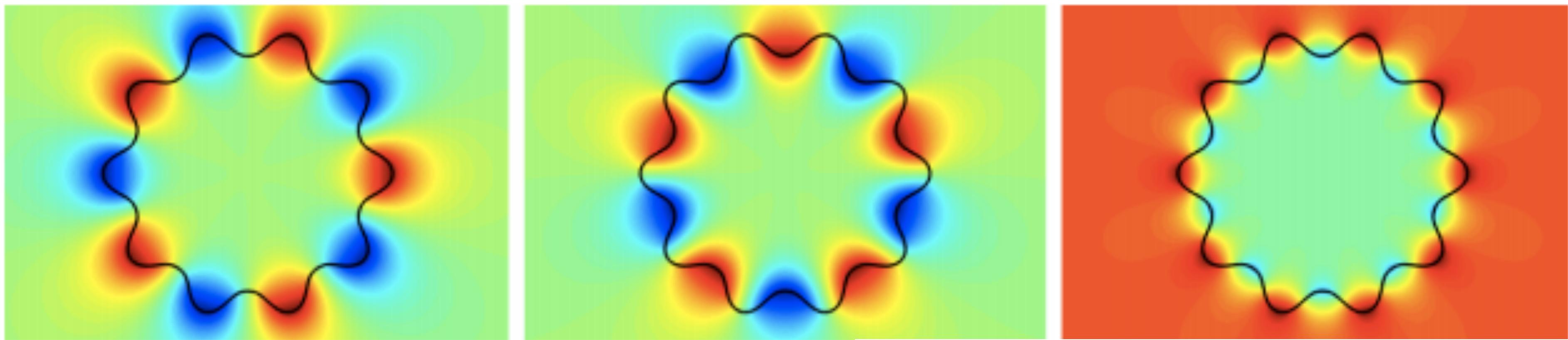


Figure 3: **2011-12 NCAA Division 1 (FBS and FCS) football schedule.** Graph representation of schedule via spectral clustering by games, *top*: vertices represent teams, edges represent games, coloring indicates conference membership. *bottom*: community detection of teams (represented using pie-graphs) 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.

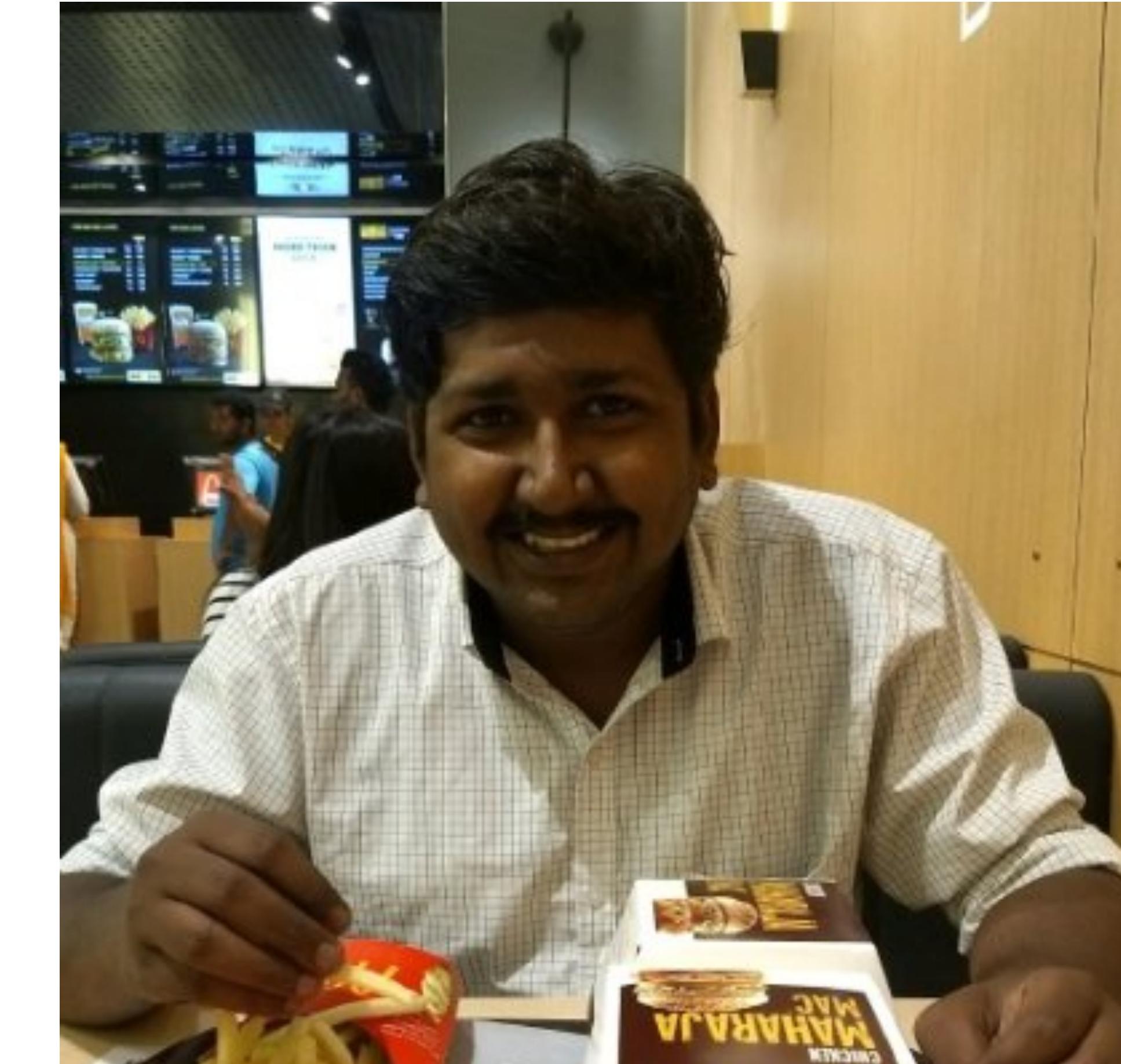
Extremal Eigenvalues



Teaching Assistants



Andy Kassen



Kiran Gadhave

Course Structure & Goals

Course Goals

Convey basic skills about each step in the data science process

data wrangling: acquire, clean, reshape, sample data

data exploration and analysis: get a feeling for the dataset, describe dataset

prediction: inferences and decisions based on data

communication

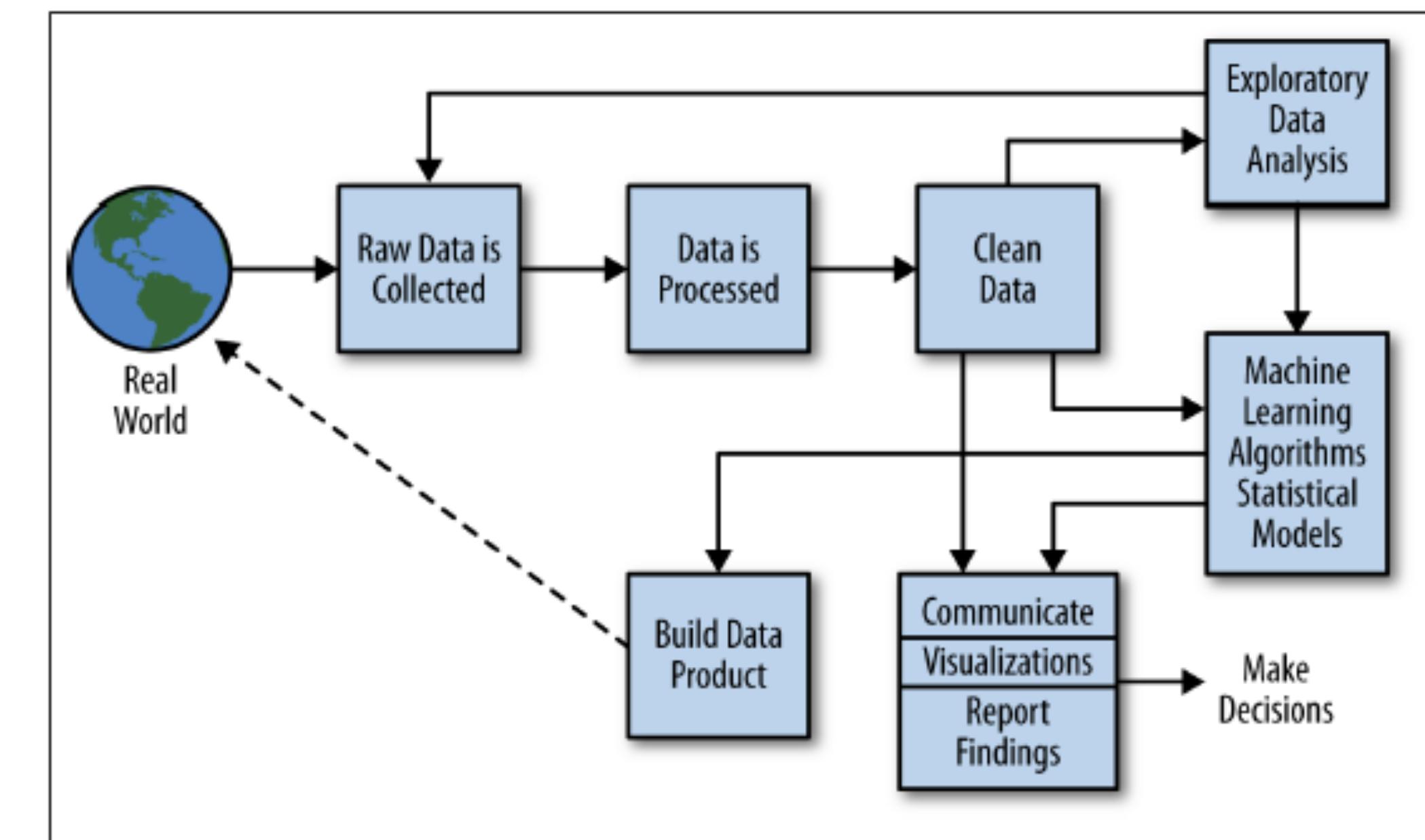


Figure 2-2. The data science process

Topics

Programming

Version Control

Data Wrangling (Pandas)

Data Acquisition

Web Scraping

Web APIs

Databases

Basic Stats

Hypothesis Testing

Visualization

Regression

Classification

Logistic Regression, K-Nearest
Neighbors, SVM, Decision Trees,
Neural Networks

Clustering

Dimensionality Reduction

Network Analysis

Natural Language Processing

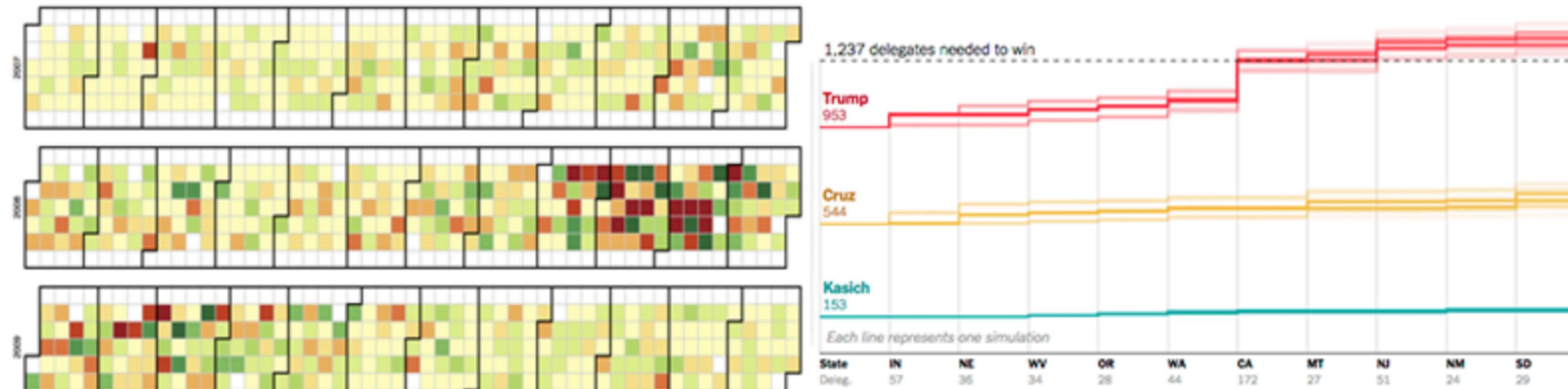
Ethics

Information datasciencecourse.net

Introduction to Data Science



Home Syllabus Schedule Project Fame Resources



D3 Calendar Chart | How the delegate race could unfold

Introduction to Data Science is a **three-credit course**, offered in the **Spring 2019** semester at the University of Utah, cross-listed between **Mathematics (MATH 4100)** and **Computing (COMP 5360)**.

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?

Communicate

Slack Team

<https://datasciencecourse2019.slack.com/>

Used for announcements and discussions. Sign up with your utah.edu e-mail address.

Canvas

<https://utah.instructure.com/courses/537442>

Used only for grading

Github

<https://github.com/datascience-course/2019-datascience-lectures>

<https://github.com/datascience-course/2019-datascience-homeworks>

Used to post lectures and homework

Office Hours

Braxton: Tuesdays, 5:00-6:00,
LCB 118

Alex: Thursdays, 5:00 - 6:00,
WEB 3887

Andy: Wednesdays, 10:00 -
11:30, LCB 326

Kiran: Friday, 11:00 - 12:30,
MEB 3167

E-Mail

alex@sci.utah.edu

osting@math.utah.edu

Course Components

Lectures introduce theory and coding

includes both short, hands-on coding exercises and longer, in-depth coding examples

Based on a published Jupyter notebook on GitHub

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/difficulty

Start early!

Due on Fridays, late days: -10% per day, up to two days.

Final Project: 40%

Teams, two milestones

Schedule

Lectures:

Tue / Th 3:40-5:00

WEB L101

Calendar

Link

Lectures frequently involve computer activities.

Bring your own computer!

Have Python, etc installed
(see HW0)

Introduction to Data Science



[Home](#) [Syllabus](#) [Schedule](#) [Project](#) [Fame](#) [Resources](#)

Schedule

MATH 4100 / COMP 5360

Today January 2019

2023-2024 School Year

 Print Week Month Agenda 

Sun	Mon	Tue	Wed	Thu	Fri	Sat
30	31	Jan 1	2	3	4	5
6	7	8	9	10	11	12
		3:40pm Intro Data S 5pm Braxton Osting	10am Andy Kassen c	HW0 Due 3:40pm Intro Data S 5pm Alex Lex Office	11am Kiran Office Ho	
13	14	15	16	17	18	19
		3:40pm Intro Data S 5pm Braxton Osting	10am Andy Kassen c	3:40pm Intro Data S 5pm Alex Lex Office	Add/Drop Deadline Homework 1 Due 11am Kiran Office Ho	
20	21	22	23	24	25	26
		3:40pm Intro Data S 5pm Braxton Osting	10am Andy Kassen c	3:40pm Intro Data S 5pm Alex Lex Office	Homework 2 due 11am Kiran Office Ho	
27	28	29	30	31	Feb 1	2
		3:40pm Intro Data S 5pm Braxton Osting	10am Andy Kassen c	3:40pm Intro Data S 5pm Alex Lex Office	Homework 3 due 11am Kiran Office Ho	

+ Google Calendar

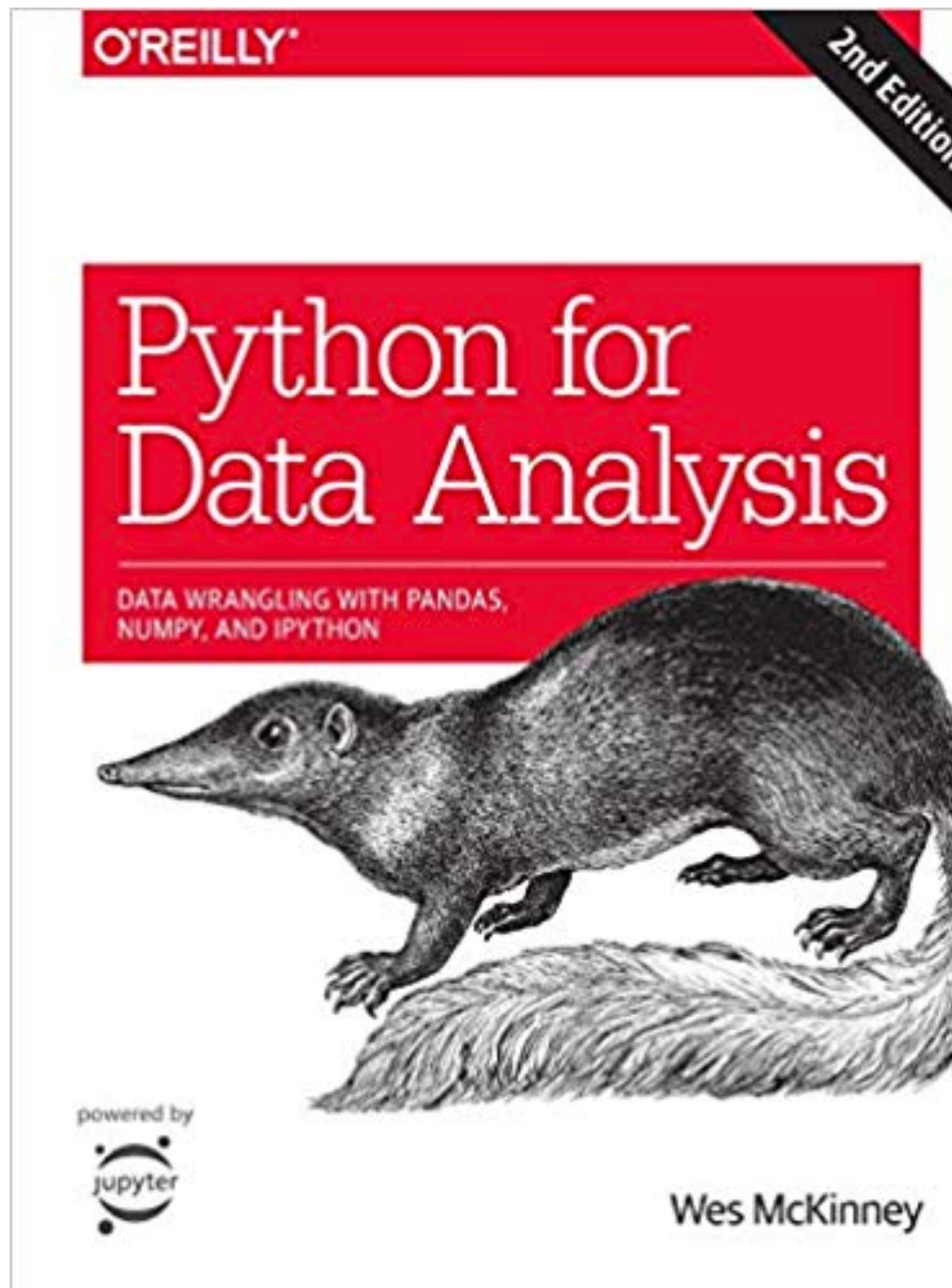
Subject to change.

Week 1

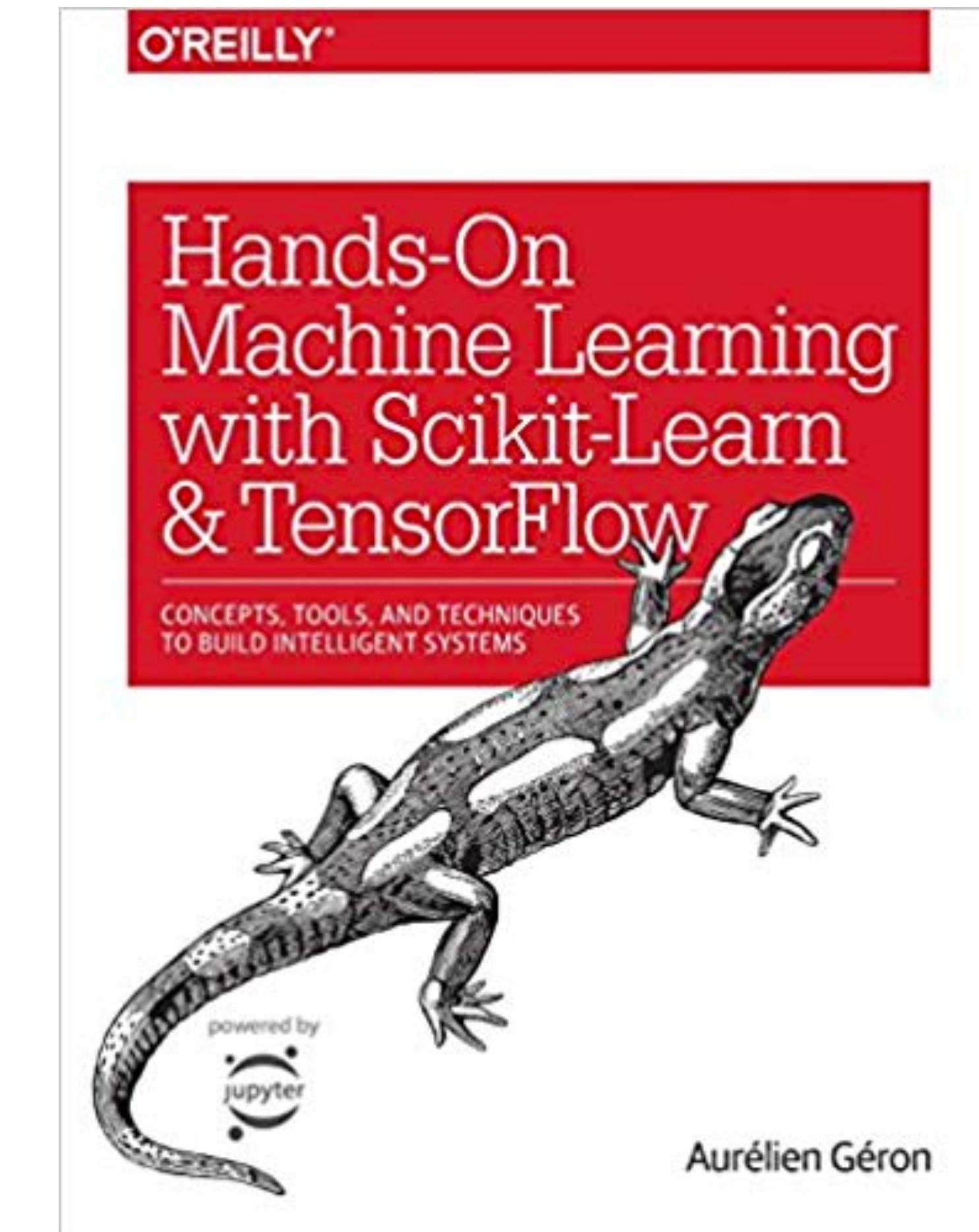
Lecture 1: Introduction

Tuesday, Jan. 8

Books



Primary Text for Readings
Available for free on Campus: [link](#)



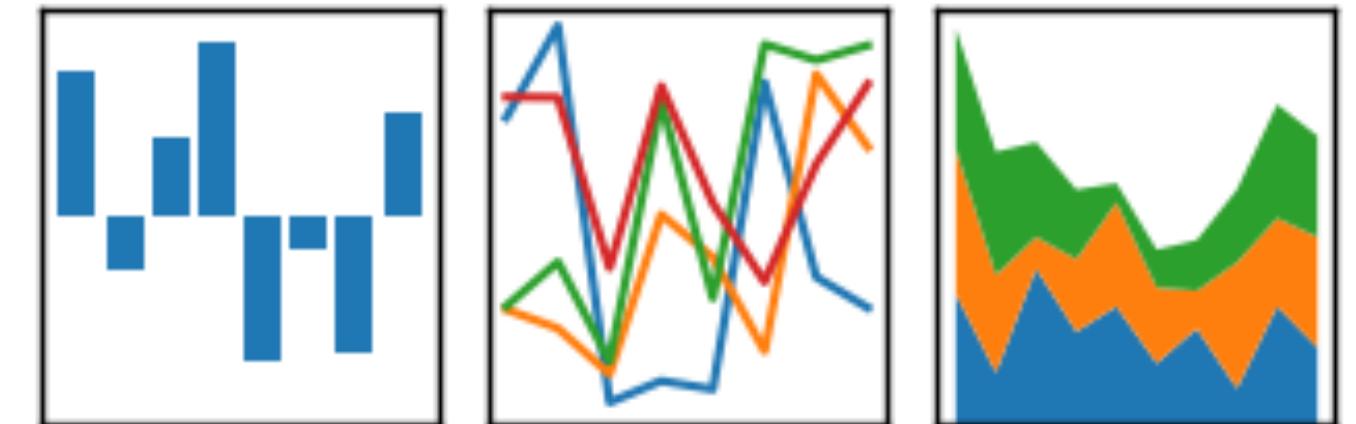
Supplementary Text
Available for free on Campus: [link](#)

Programming



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.

Academic Misconduct

You are welcome to **discuss** the course's ideas, material, and homework with others in order to better understand it, but **the work you turn in must be your own** (or for the project, your's and your teammate's). For example, you must **write your own code**, design your own visualizations, and critically evaluate the results in your own words. You may not submit the same or similar work to this course that you have submitted or will submit to another. Nor may you provide or make available solutions to homework to individuals who take or may take this course in the future.

We will **automatically check for plagiarism** in all your submissions.

SoC Cheating Policies

See also the SoC Academic Misconduct Policy:

http://www.cs.utah.edu/wp-content/uploads/2014/12/cheating_policy.pdf

You will fail the class if you cheat.

A “strike” will be recorded.

CoE and SoC Policies

[https://www.coe.utah.edu/wp-content/uploads/pdf/faculty/
semester_guidelines.pdf](https://www.coe.utah.edu/wp-content/uploads/pdf/faculty/semester_guidelines.pdf)

<http://www.cs.utah.edu/socguidelines/>

Course Policies

Review Syllabus for:

Collaboration, Cheating and Plagiarism

Missed Activities and Assignment Deadline

Late Policies

Regrading Policies

Respect for Diversity

American with Disabilities Act

Sexual Misconduct

Student Name and Personal Pronoun

This Week

HW0, including course survey

Make sure to complete this before class on Thursday. Use office hours!

Introduction to programming in python

Readings:

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

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

HW 0

<https://github.com/datascience-course/2019-datascience-homeworks/tree/master/HW0>

README.md 

Homework 0

Introduction to Data Science - MATH 4100 / COMP 5360.

This homework is due before class on Thursday, January 10th.

Welcome to MATH 4100 and Computing 5360 - Introduction to Data Science. In this class, we will be using a variety of tools that will require some initial configuration. To ensure everything goes smoothly moving forward, we will set up the majority of those tools in this homework. This homework will not be graded, but **it is essential that you complete it before the second lecture** as it sets up the tools that we will be using in class for exercises.

1. Survey

This is a class about data, so we also want to have some data about you! Please complete the course survey [located here](#). It should only take a few moments of your time.

2. Introduction

Once you are signed up for the class and have access to [Slack](#), introduce yourself to your classmates and course staff by introducing yourself in the #general channel. Include your name/nickname, your affiliation, why you are taking this course, and tell us something interesting about yourself (e.g., an unusual hobby, past travels, or a cool project you did, etc.). Also tell us whether you have experience with data science.

Github

Github is a web-based hosting service for version control using git.

We'll discuss git and github extensively in a later lecture.

The basics are described in the README.md file located [here](#).

The screenshot shows a GitHub repository page for a README.md file. The title 'Introduction to Data Science - Homeworks' is displayed prominently. Below the title, there is a course website link: <http://datasciencecourse.net>. A note explains that the repository contains homework directories and recommends using git to clone and update it. It also suggests using GitHub Desktop or specific commands for updates. Two sections are detailed: 'Initial Step: Cloning' with a command to clone the repository, and 'Updating' with a command to pull updates. Both sections include explanatory text and a note about creating a local folder on the computer.

README.md

Introduction to Data Science - Homeworks

Course website: <http://datasciencecourse.net>

This repository will contain directories with all homeworks. You can manually download the files for each homework, but we recommend that you use git to clone and update this repository.

You can use [GitHub Desktop](#) to update this repository as new homeworks are published, or you can use the following commands:

Initial Step: Cloning

When you clone a repository you set up a copy on your computer. Run:

```
git clone https://github.com/datascience-course/2019-datascience-homeworks
```

This will create a folder `2019-datascience-homeworks` on your computer, with the individual homeworks in subdirectories.

Updating

As we release new homeworks, or if we discover mistakes and update an already released homework description, you'll have to update your repository. You can do this by changing into the `2019-datascience-homeworks` directory and executing:

```
git pull
```

That's it - you'll have the latest version of the homeworks.

Next Week

Data Structures and Pandas

Introduction to Descriptive Statistics

HW1 due

About You

Current Enrollment

Math 4100: 59

COMP 5360: 54

Trouble enrolling? send an email

Class Survey

First year for anyone?

Who is an undergraduate?

Who is a MS student?

Who is a PhD student?

Math? Biology? Other Sciences? Engineering? Humanities? Business? Other?

Who knows Python?

R?

Matlab?

C / C++?

Java?

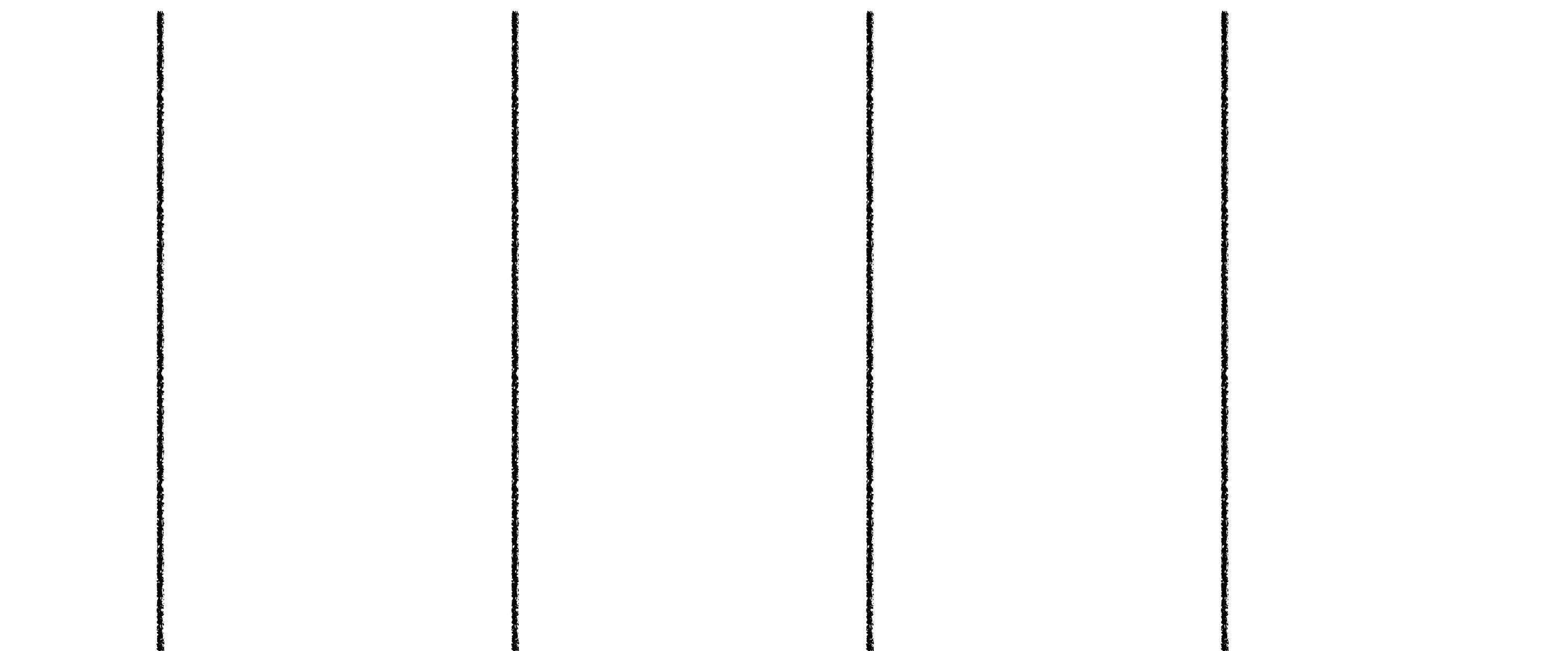
Other languages?

Who has programmed for 1 year? 5 years? More?

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

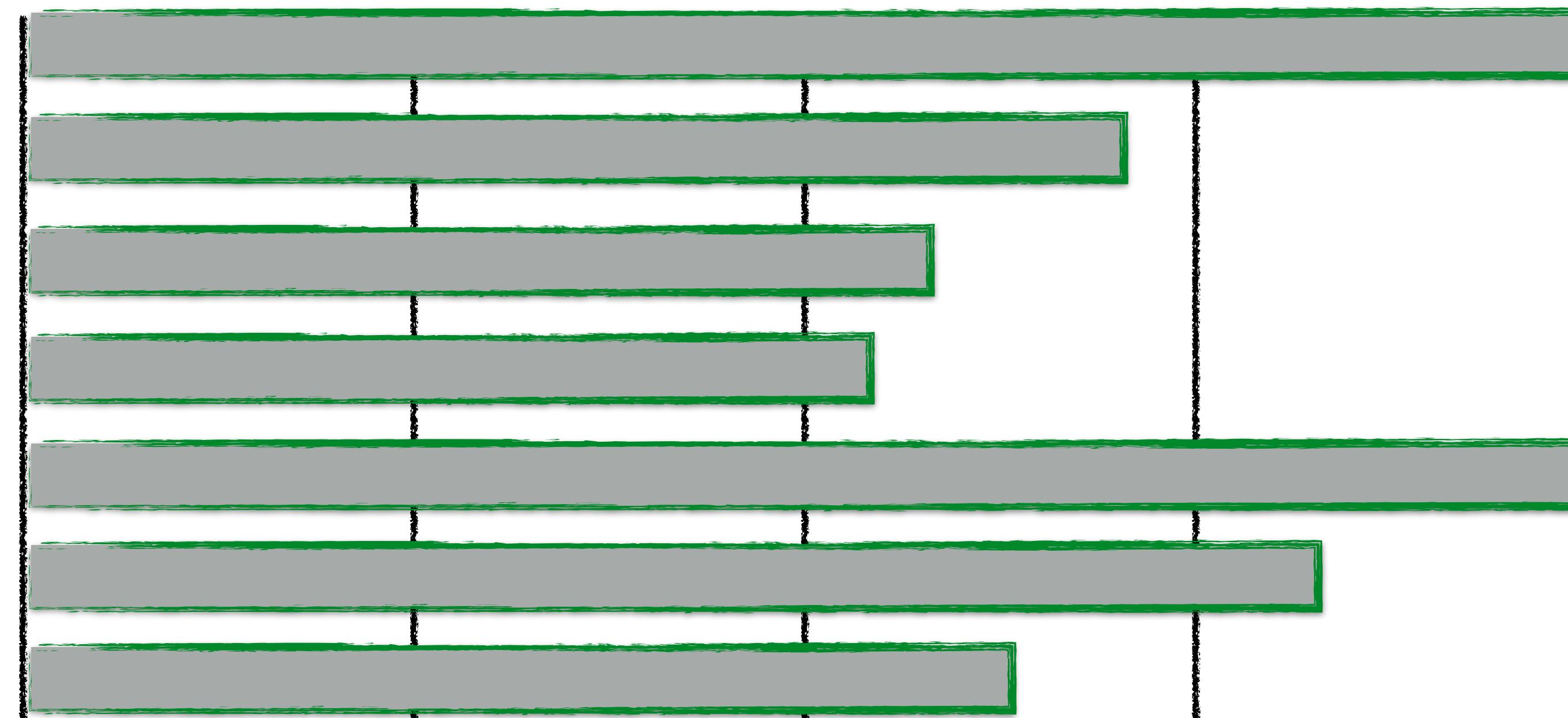


In addition, in the comments section, please write any particular subjects you'd like to see covered in class.

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



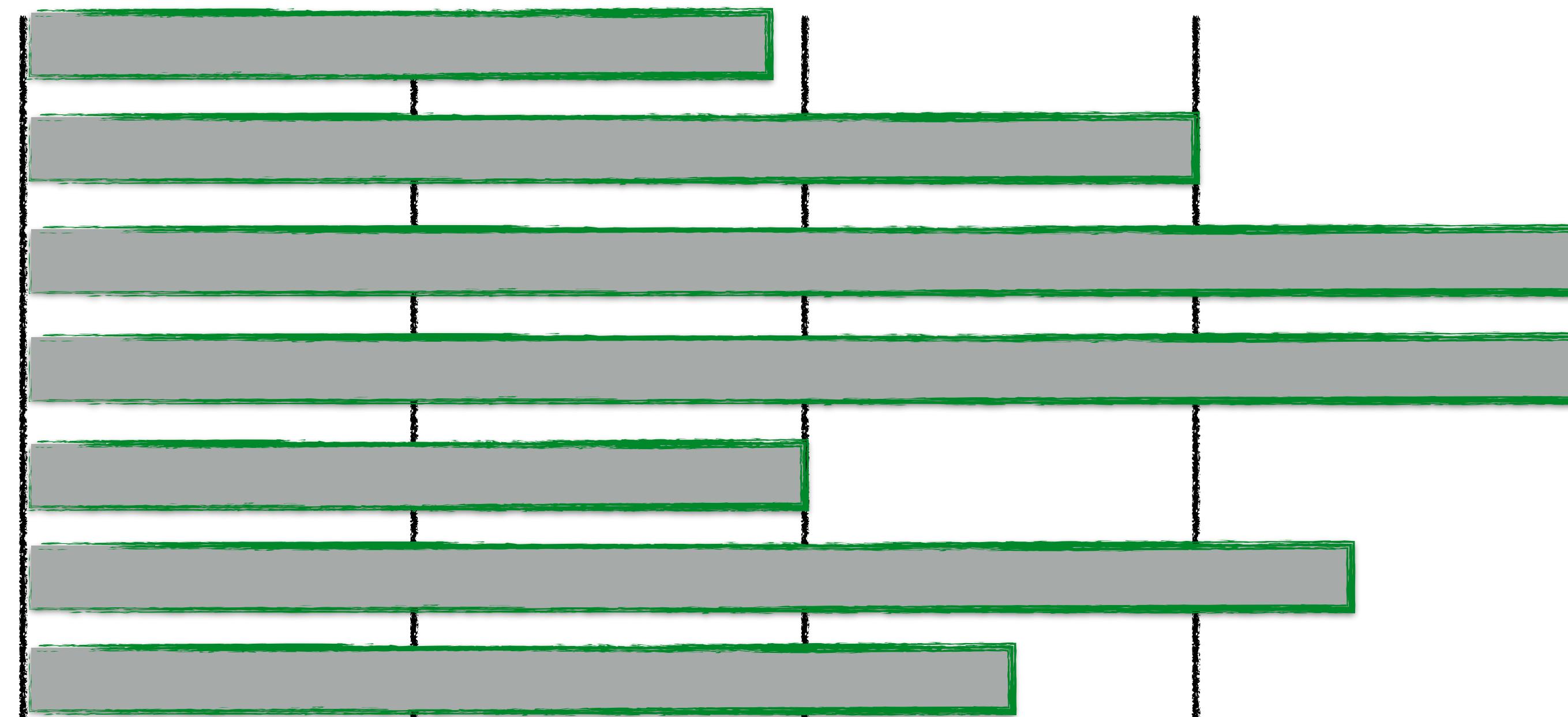
1 - little knowledge

5 - Expert

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



1 - little knowledge

5 - Expert