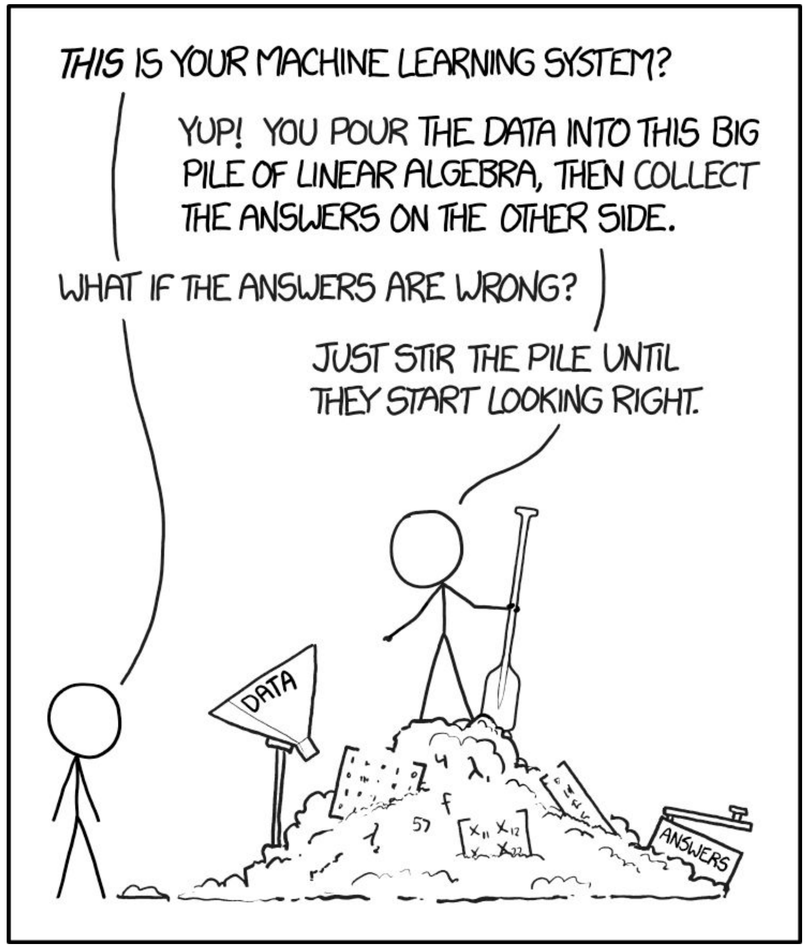
Introduction to Data Science
CS 5360 / Math 4100

Kate Isaacs
u6044649@gcloud.utah.edu

Tom Alberts
alberts@math.utah.edu



[xkcd]

Logistics

See CANVAS Module “Important Online Links” for links to the class website, YouTube channel, and office hours.

We will use PIAZZA for course discussions and Q&A, so please sign up!

All assignments will be submitted and graded in CANVAS, but see the course webpage for syllabus, schedule, and detailed course information.

(Old) Recordings

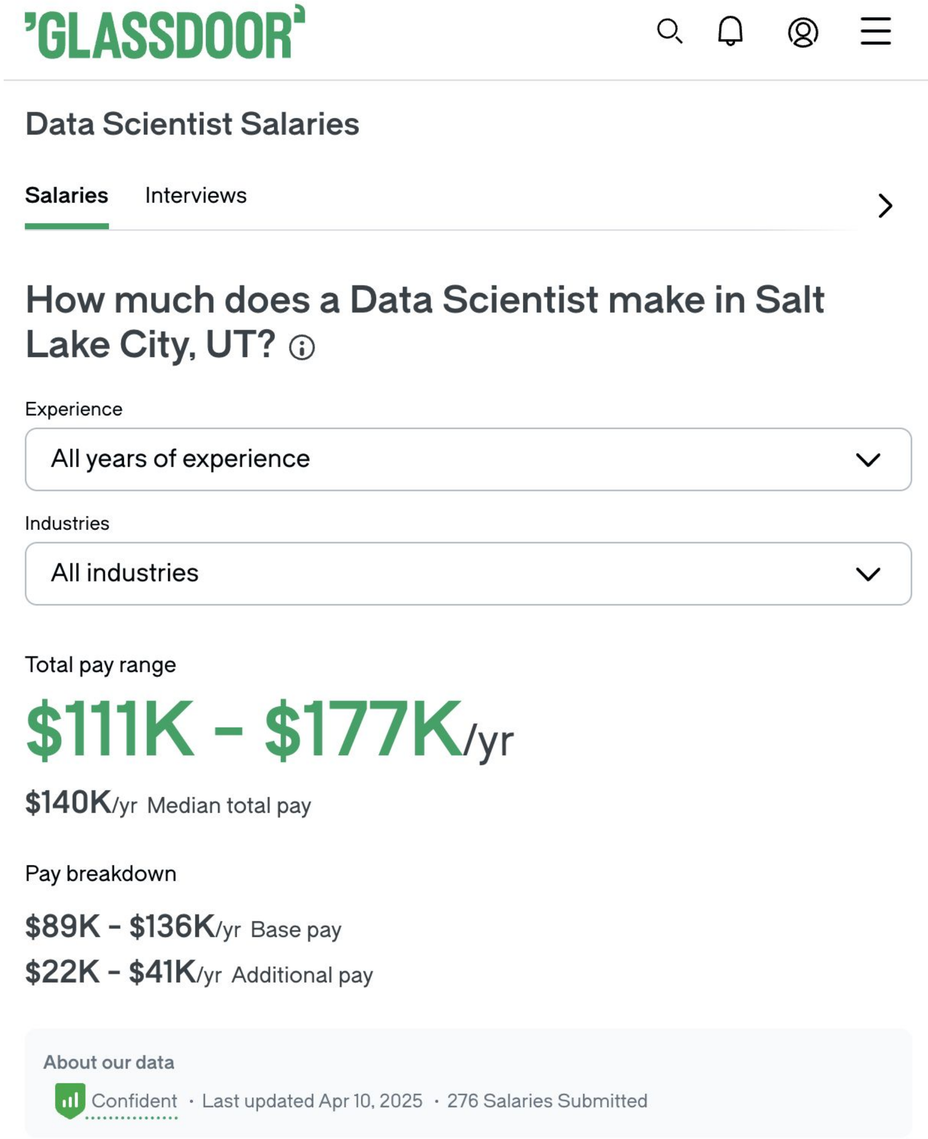
We are **not** recording this class or doing hybrid lectures per guidance from our colleges.

There are lectures from 2022 which are similar, can serve for review or backup https://datasciencecourse.net/2022/

If there is severe weather, we may do a hybrid or Zoom lecture. We will announce on Canvas. Please reach out on days you're concerned about if we haven't made a call.

What is Data Science?

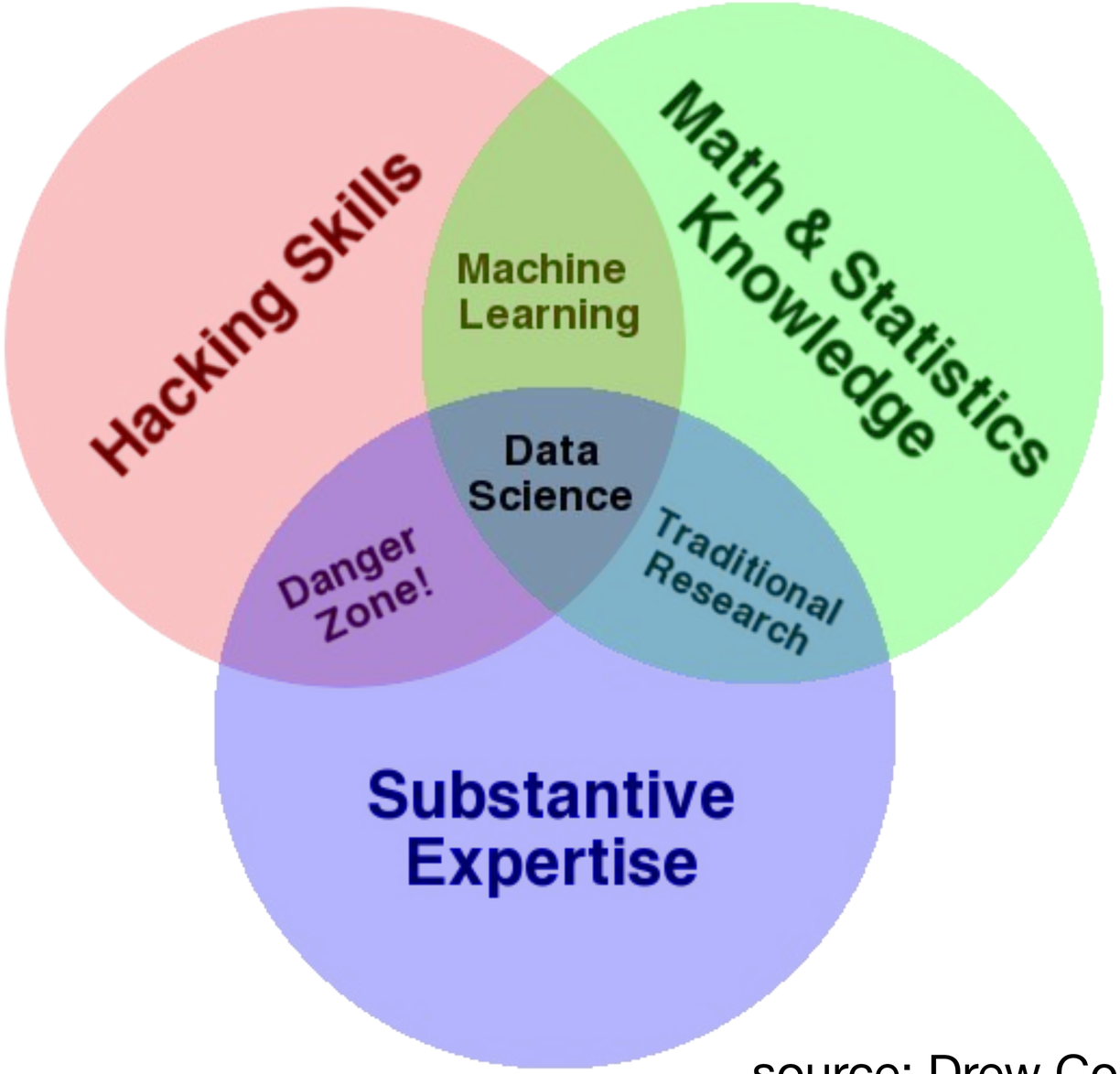
- The sexiest job of the century —Harvard Buisness Review
- A data scientist is a statistician who lives in San Fransisco
- Data Science is statistics on a Mac
- A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.



What is Data Science?

Source: datascience.berkeley.edu

What is Data Science?



source: Drew Conway blog

What is Data Science?

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. (Wikipedia)

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

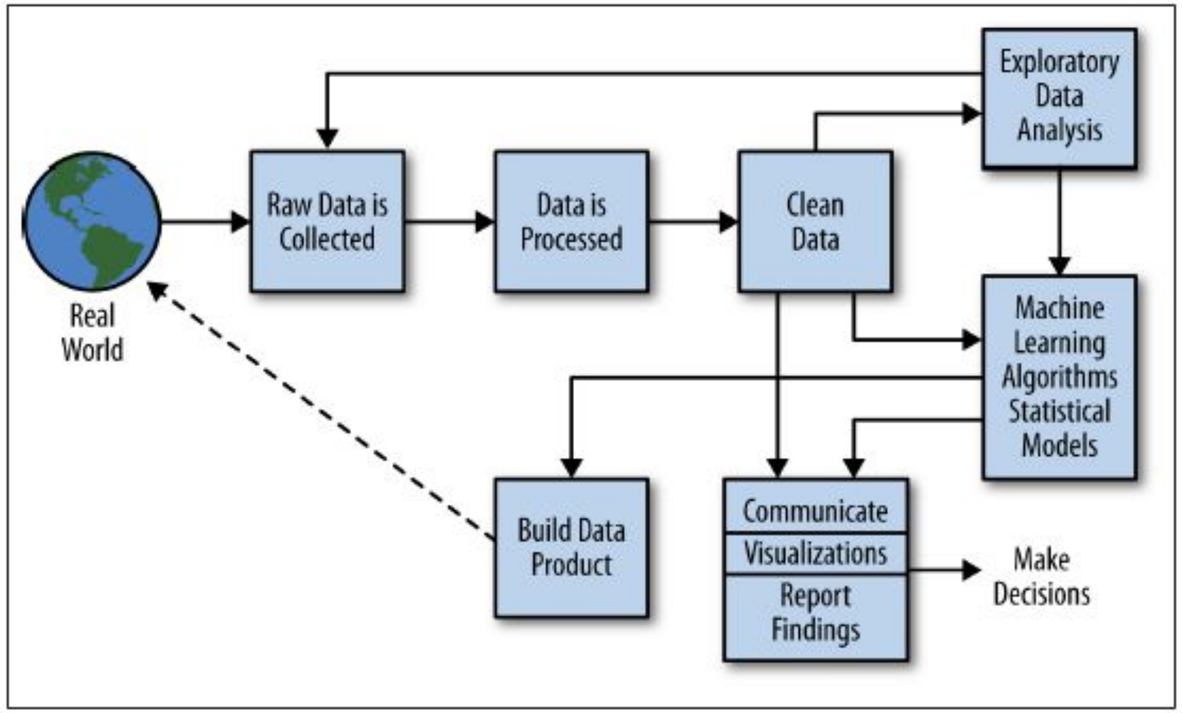


Figure 2-2. The data science process

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?

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.

https://cacm.acm.org/blogs/blog-cacm/224105-learning-computer-science-is-different-than-learning-other-stem-disciplines/fulltext

Big Data

2024: 149 zettabytes, largely unstructured

(149,000,000,000,000,000,000,000 bytes,

about 37 trillion DVD’s)

90% generated in last 2 years

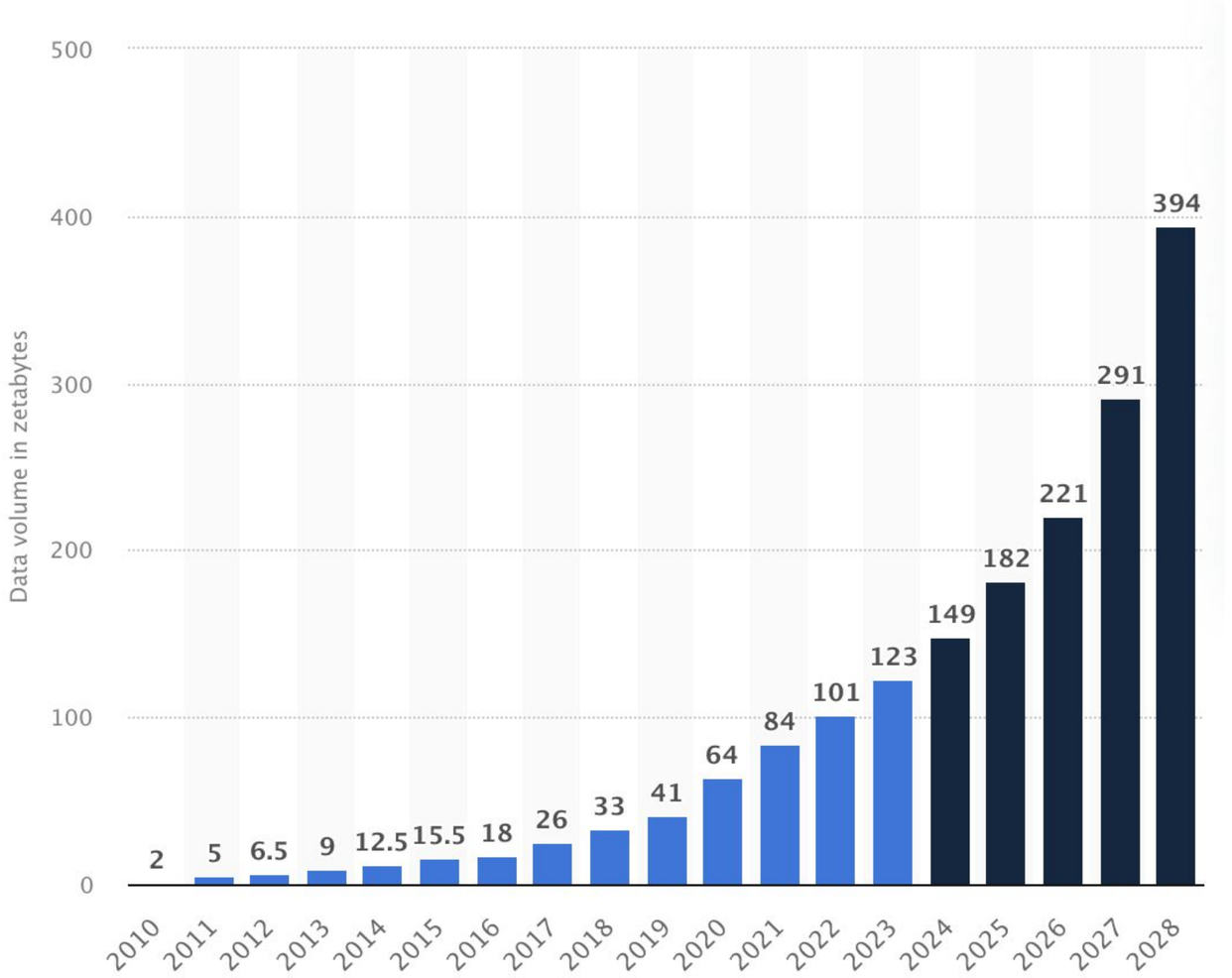
Volume doubling about every 4 years

0.4 zettabytes (402 million terabytes) generated every day in 2024

15 Exabytes in Punch Cards:
4.5 km over New England

Volume of Data Created, Captured, Copied, and Consumed Worldwide

Source: statista.com



Data generated per minute online (2023)

source: https://www.domo.com/learn/infographic/data-never-sleeps-11

How can we leverage data?

Improve your fitness by targeted training

Improve your product

- by targeting your audience
- by considering semantics

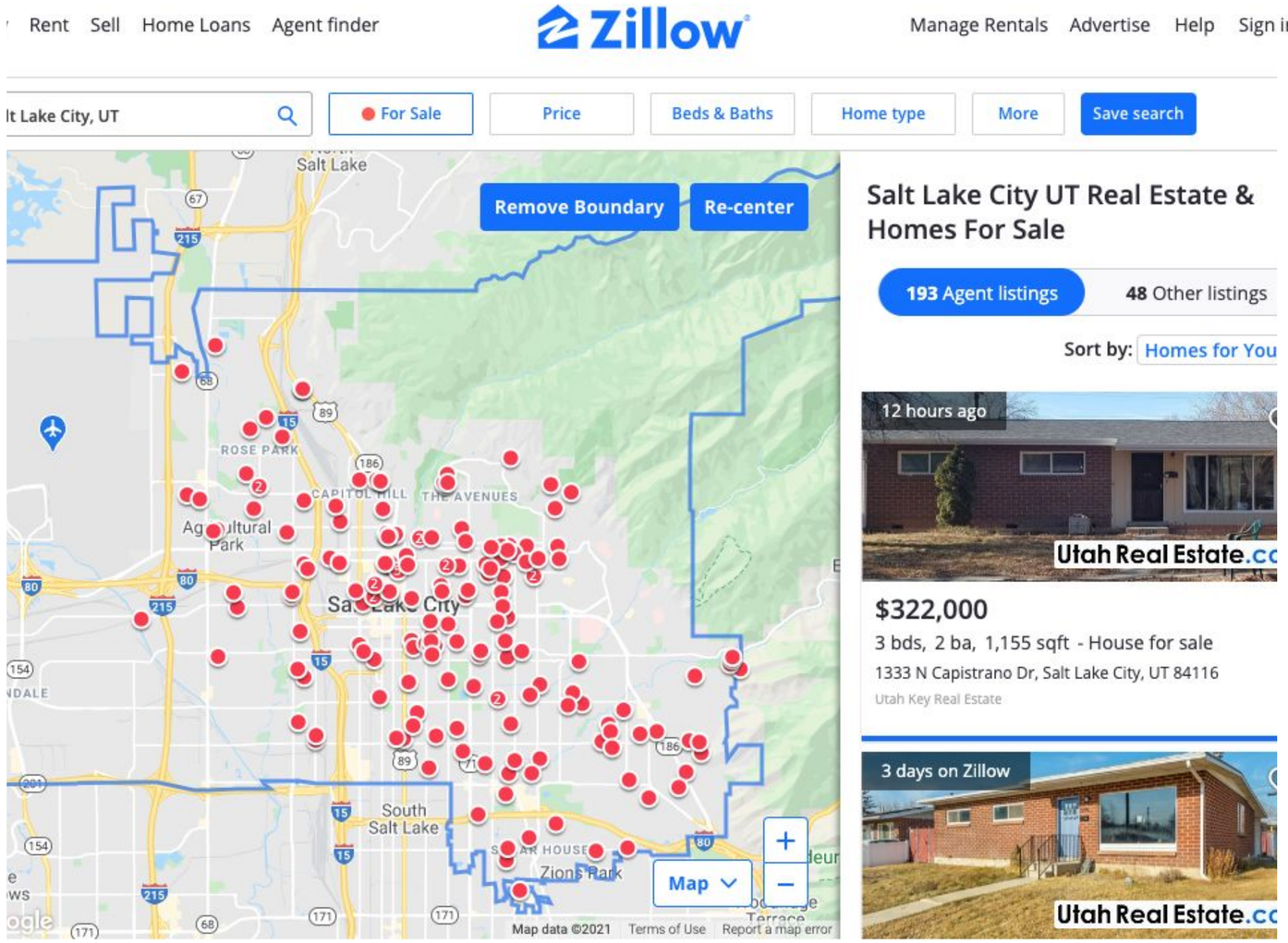
Make better decisions

- exact diagnosis, choose right medication, pick good restaurant

Predict elections, events, crowd behavior, etc.

… and many more applications

Example: Personal Data



Screenshot of map of Zillow listings in Salt Lake City

Screenshot of fitness app

Big Data in Science and Engineering

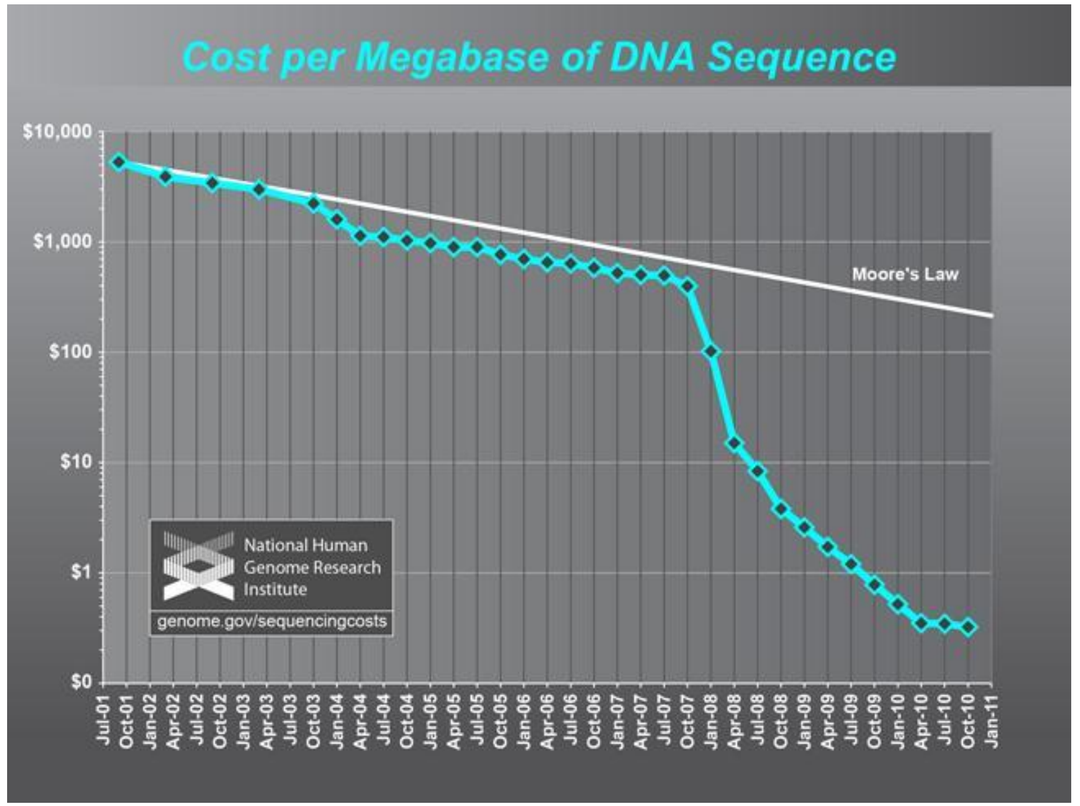
Big Data transformed science and engineering.

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**.

Example: CERN Large Hadron Collider Data

- 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 (Cancer Genome Atlas): 1 Petabyte

“As a single human genome takes up 100 gigabytes of storage space, and more and more genomes are sequenced, storage needs will grow from gigabytes to petabytes to exabytes. By 2025, an estimated 40 exabytes of storage capacity will be required for human genomic data.”

Source: medicalfuturist.com

NSA Utah “BumbleHive” Data Center (Bluffdale, Utah)

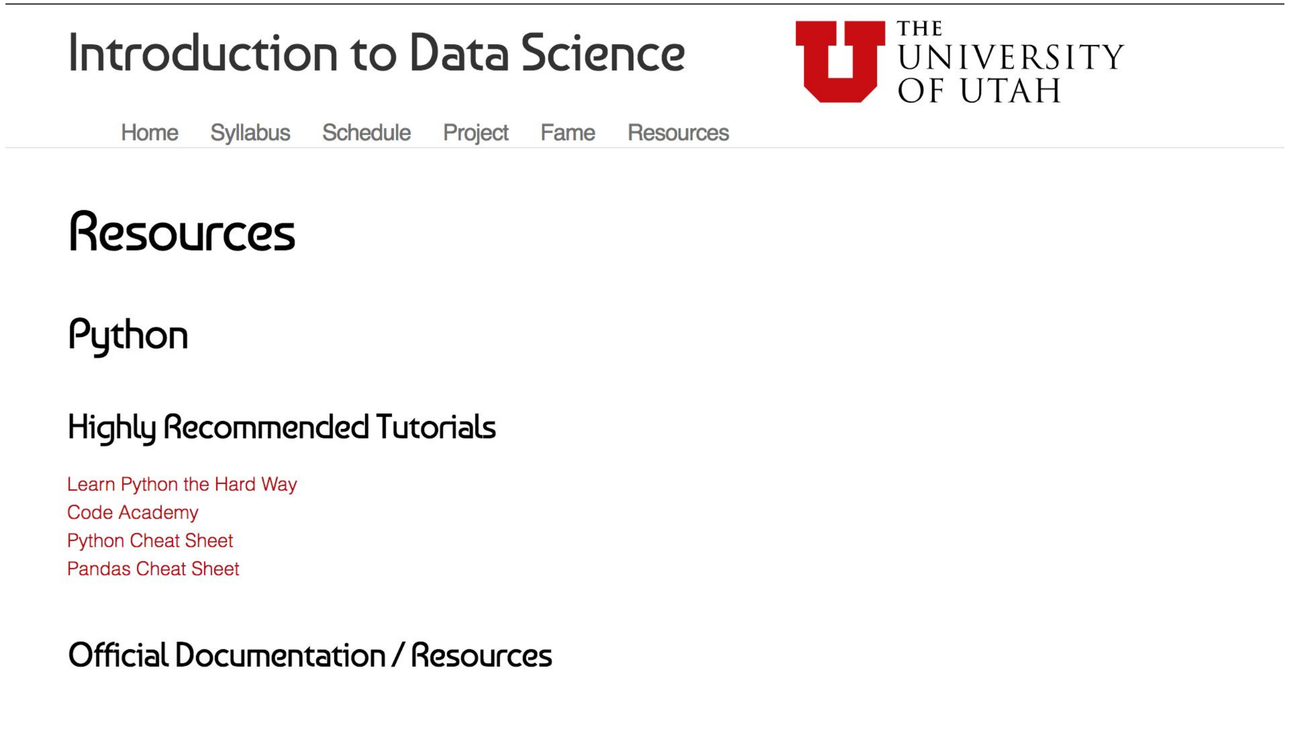
Storage Capacity?

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

By User:ParkerHiggins (Electronic Frontier Foundation) - Own work; https://www.eff.org/deeplinks/2014/07/releasing-public-domain-image-nsas-utah-data-center, CC0, https://commons.wikimedia.org/w/index.php?curid=33871035

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

THE UNIVERSITY OF UTAH

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
Quandl
Datamob
Factual
Metro Boston Data Common

Course 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

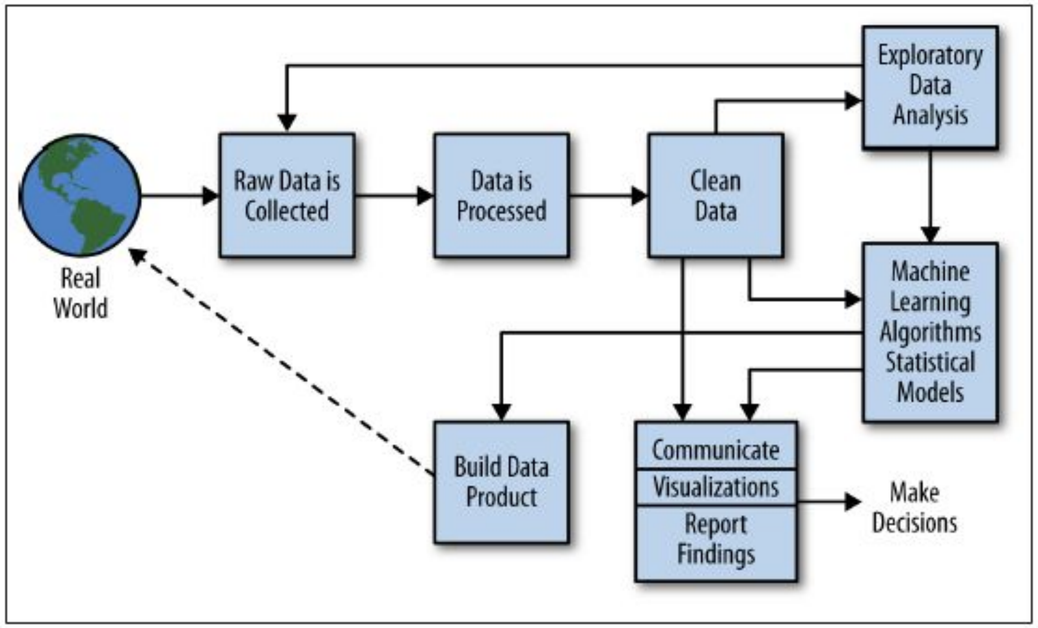


Figure 2-2. The data science process

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

Neural Networks

Network Analysis

Natural Language Processing

Ethics

Teaching Staff
CS 5360 / Math 4100

TAs

Li-Fu Chen

Math PhD Student

Elham Ghelichkhan

CS PhD Student

Dylan Mckellips

CS PhD Student

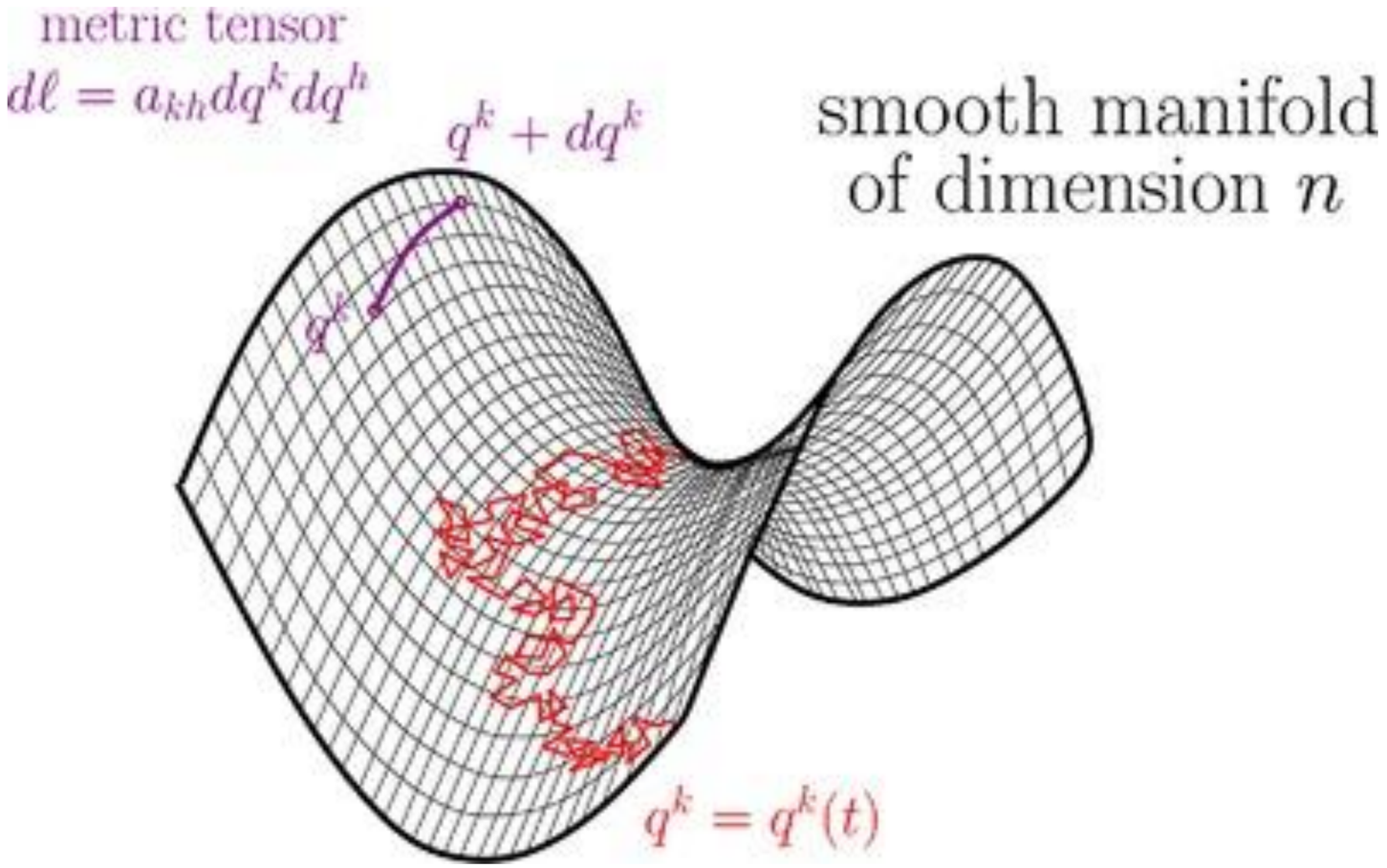
Tom Alberts

Associate Professor, Mathematics

https://www.math.utah.edu/~alberts/

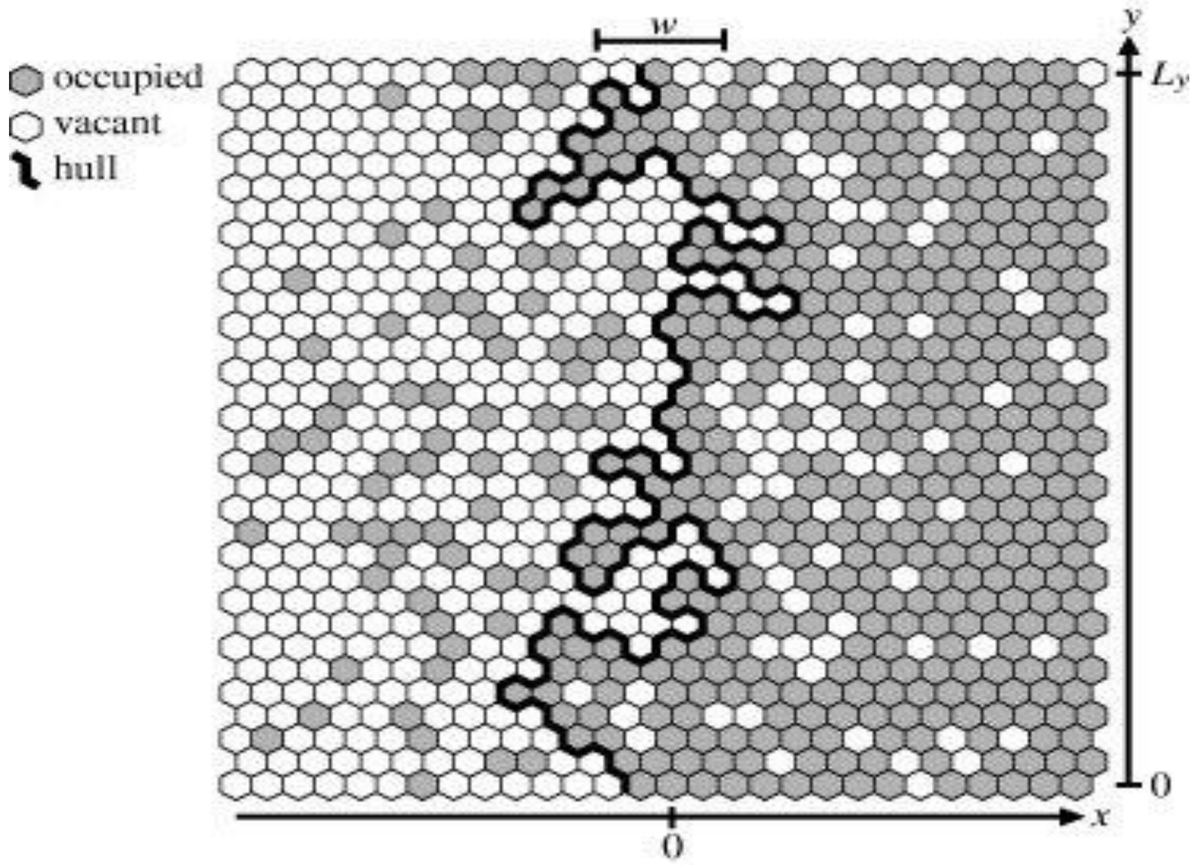
- Research in theoretical probability and applications to statistical physics
- Broadly: the study of randomness that is determined by an underlying geometry

Stochastic Processes on Riemannian Manifolds



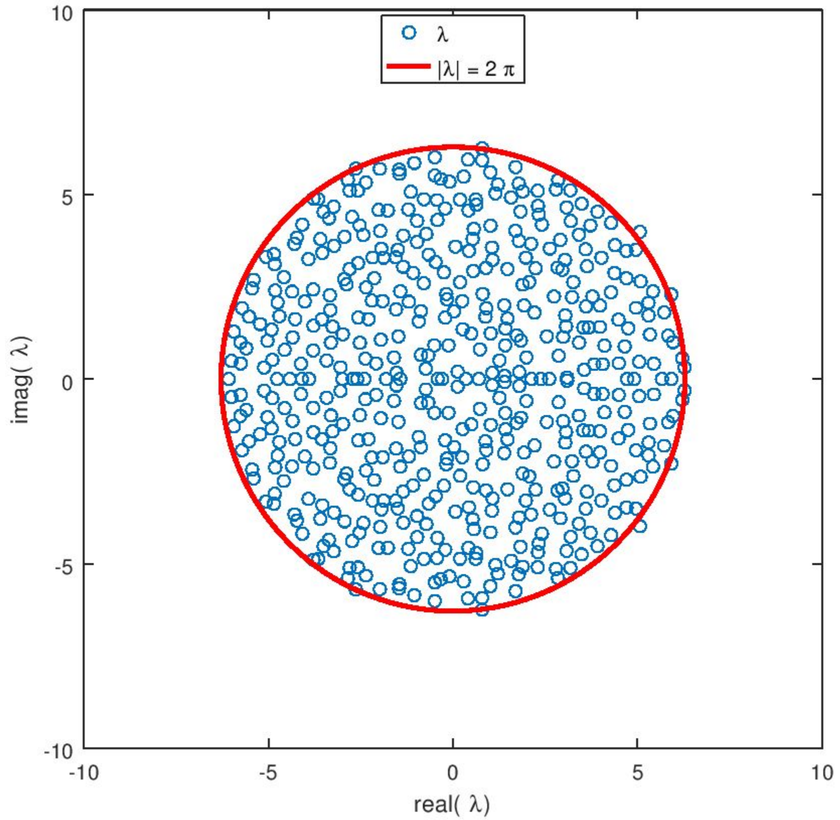
Brownian motion on a manifold: https://doi.org/10.1002/andp.201500221

Scaling Limits of Two-Dimensional Statistical Mechanics

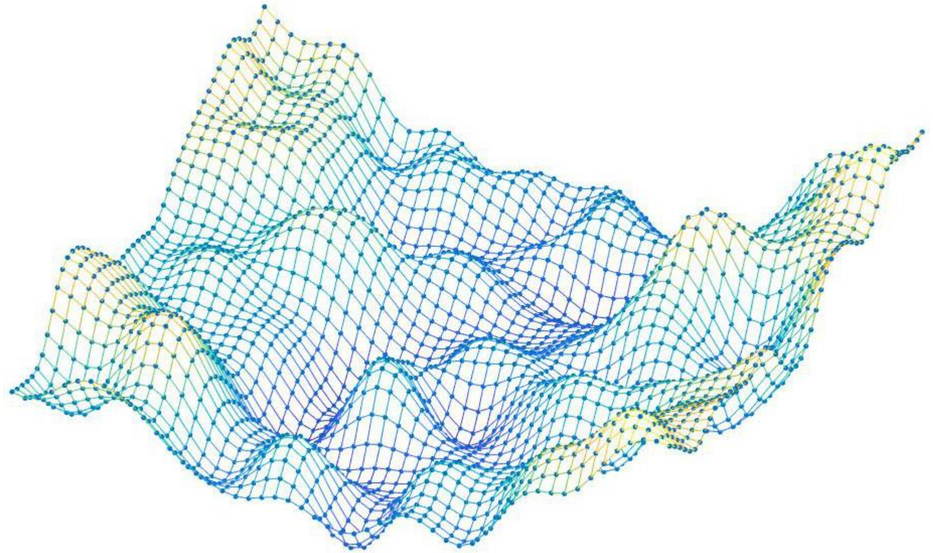


Percolation process and its interface: [Link]

Random Matrices



Eigenvalues of the Ginibre ensemble



Random matrices and random landscapes

Kate Isaacs

Associate Professor, Computer Science

Faculty, Scientific Computing and Imaging (SCI) Institute

PhD in CS, UC Davis

Previously the University of Arizona. During PhD intern w/Lawrence Livermore National Lab & Facebook Data Science Infrastructure Team

http://kisaacs@github.io

Data visualization, high performance computing

Interactive data visualization: Designing for notebooks

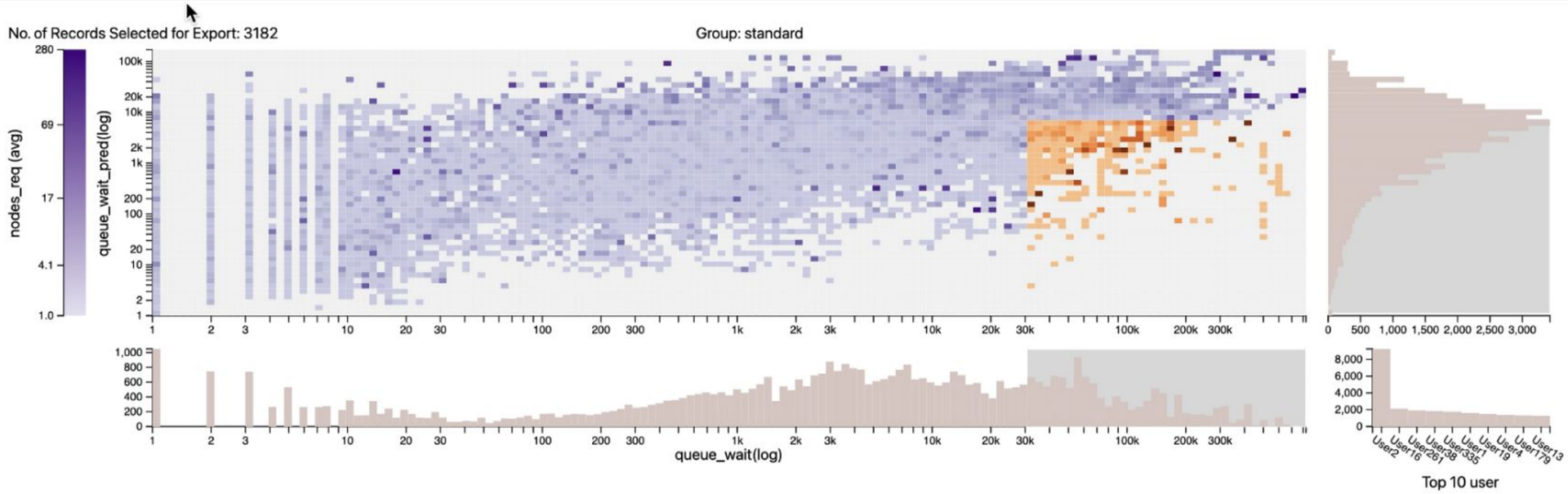
- 'y': The column name from the pandas dataframe which will be shown on the y axis of this visualization. The data in the column can be integers or floats.
- 'color': The column name from the pandas dataframe which will determine the color of squares in the main summary view. The data in the column can be integers or floats.
- 'color_agg': This is a specification for what aggregation is used for the color variable. It can be: 'avg', 'variance', 'std', 'sum', or 'median'
- 'categorical': A categorical variable from the dataset. The data in the column must be a string. The visualization will show the top 7 instances of this variable.

[21]:
```
gp.vis_configs = {
    'x':'queue_wait',
    'y':'queue_wait_pred',
    'color':'nodes_req',
    'color_agg':'avg',
    'categorical':'user',
}
```

Run the Visualization by calling the guidepost object in it's own cell: `gp`

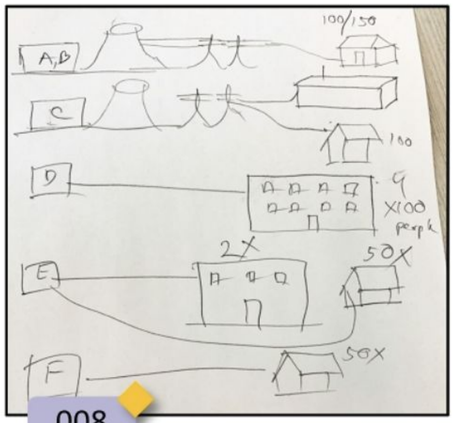
[27]:
```
gp
```

[27]:

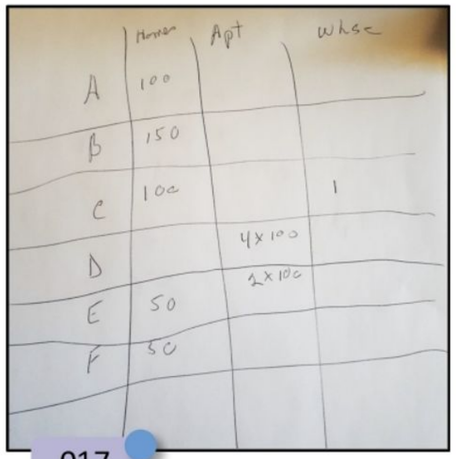


How do people conceptualize data?

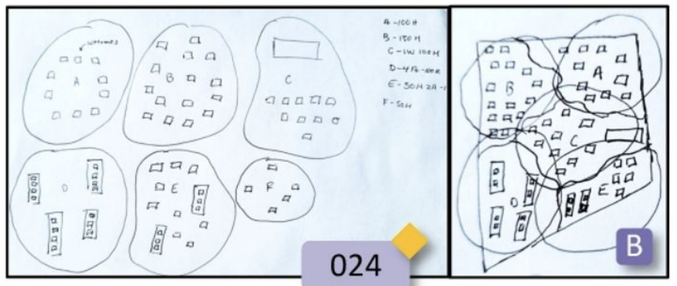
Participant ID | B Second Sketch | Computing Participant | Non-Computing Participant

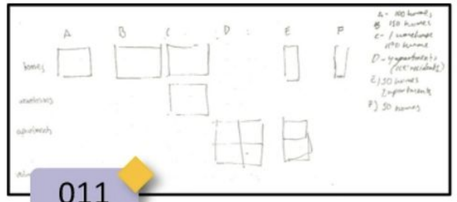


008

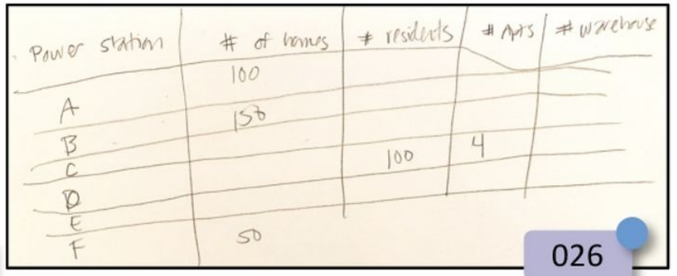


017

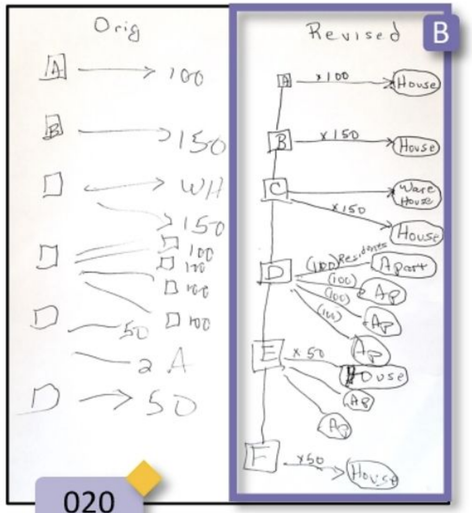




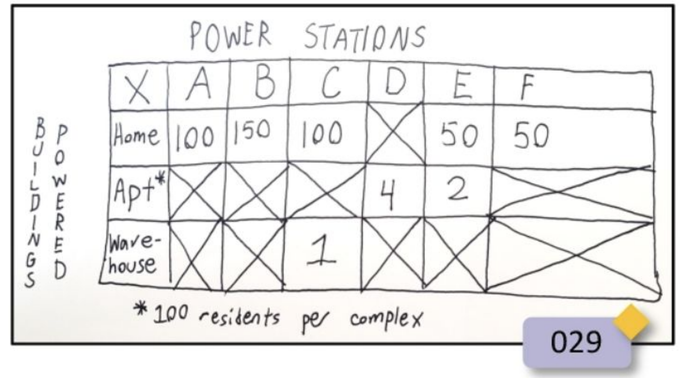
011



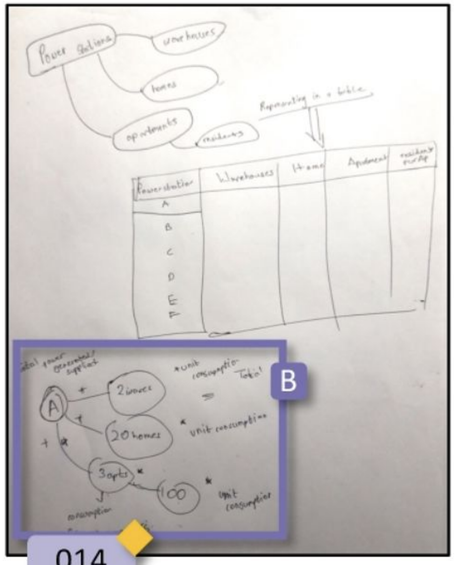
026



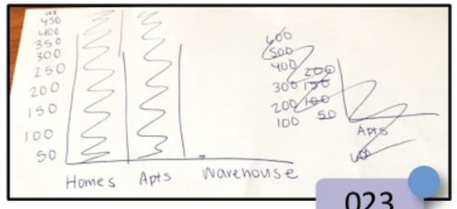
020



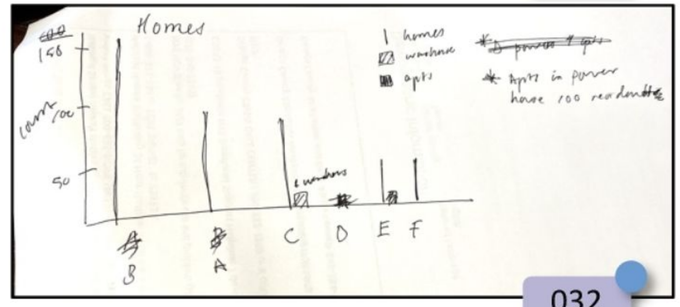
029



014

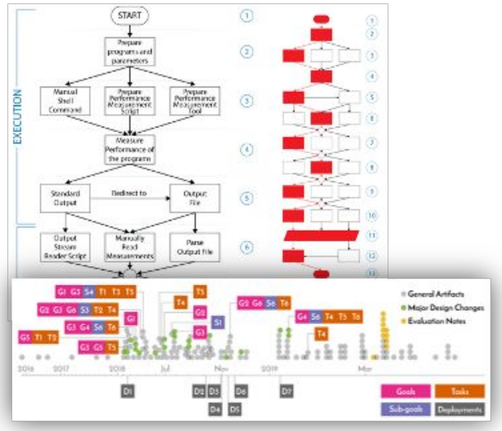


023



032

Vis Research Methodology



Vis Design Studies

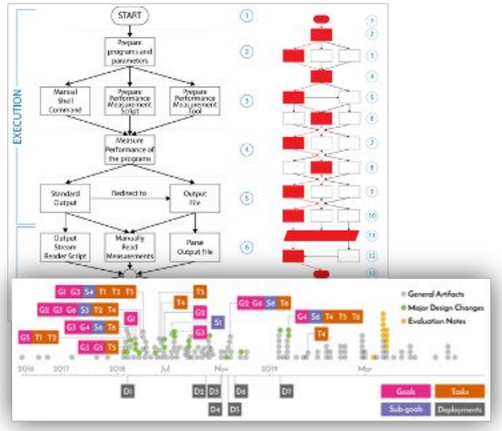
(InfoVis 2019, Ongoing)

Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature
Data structure	improvements or new uses of existing ones to support visualization
Dataset	publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation
Deployment	discussion of insights gained from real world deployment of a tool or technique
Design methodology	methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches
Evaluation methodology	methodologies that enable new ways in evaluating visualization solutions
Formalism	generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts

Contribution Types

(CG&A 2019)

Vis Research Methodology



Vis Design Studies

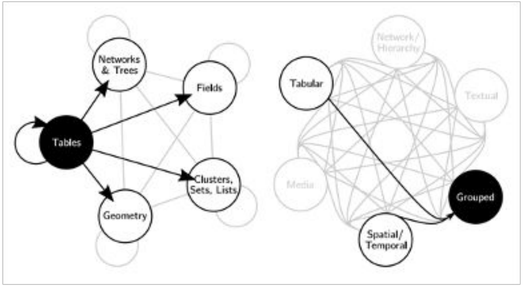
(InfoVis 2019, Ongoing)

Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature
Data structure	improvements or new uses of existing ones to support visualization
Dataset	publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation
Deployment	discussion of insights gained from real world deployment of a tool or technique
Design methodology	methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches
Evaluation methodology	methodologies that enable new ways in evaluating visualization solutions
Formalism	generalized theoretical, algorithmic, or mathematical formalisms of visualization

Contribution Types

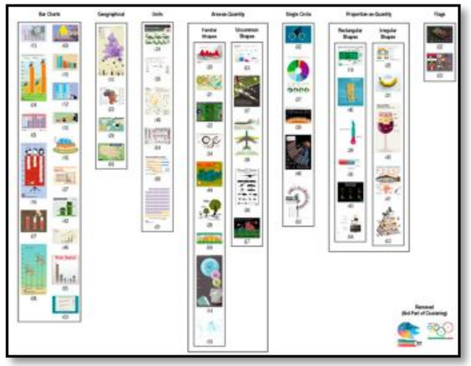
(CG&A 2019)

How People Conceptualize Data



Latent Data Abstractions

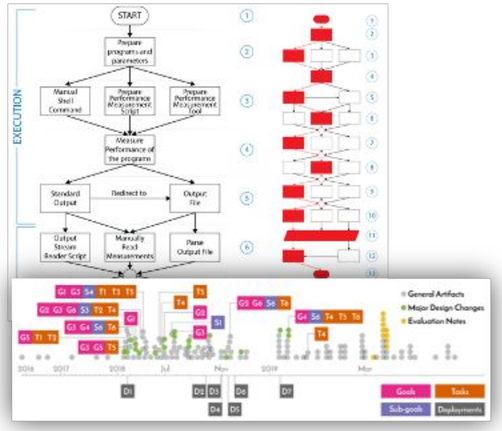
(InfoVis 2020, Ongoing)



Classification of Infographics

(Diagrams 2018)

Vis Research Methodology

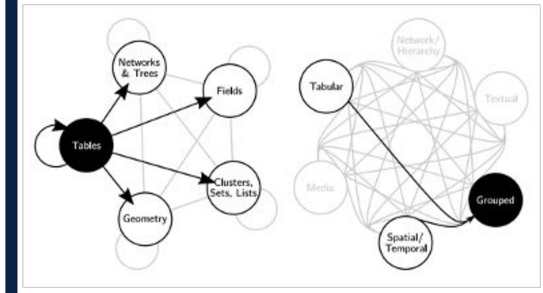


Vis Design Studies
(InfoVis 2019, Ongoing)

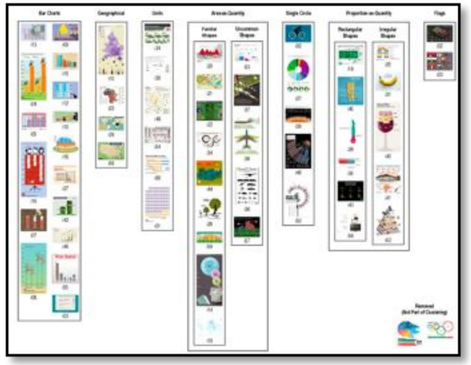
Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature
Data structure	improvements or new uses of existing ones to support visualization
Dataset	publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation
Deployment	discussion of insights gained from real world deployment of a tool or technique
Design methodology	methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches
Evaluation methodology	methodologies that enable new ways in evaluating visualization solutions
Formalism	generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts

Contribution Types
(CG&A 2019)

How People Conceptualize Data

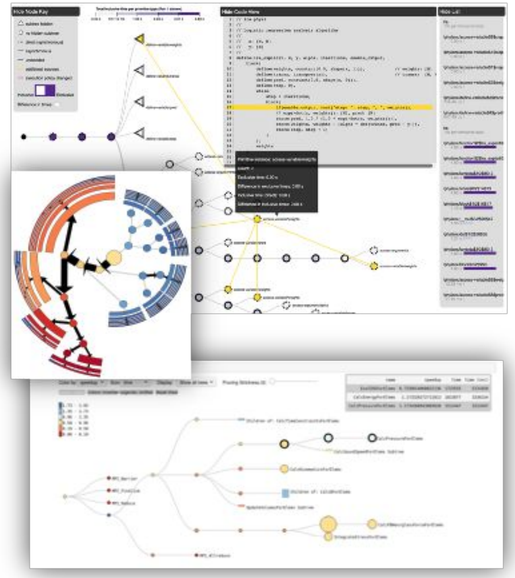


Latent Data Abstractions
(InfoVis 2020, Ongoing)



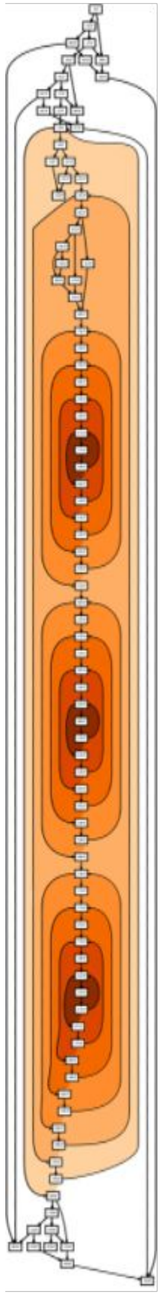
Classification of Infographics
(Diagrams 2018)

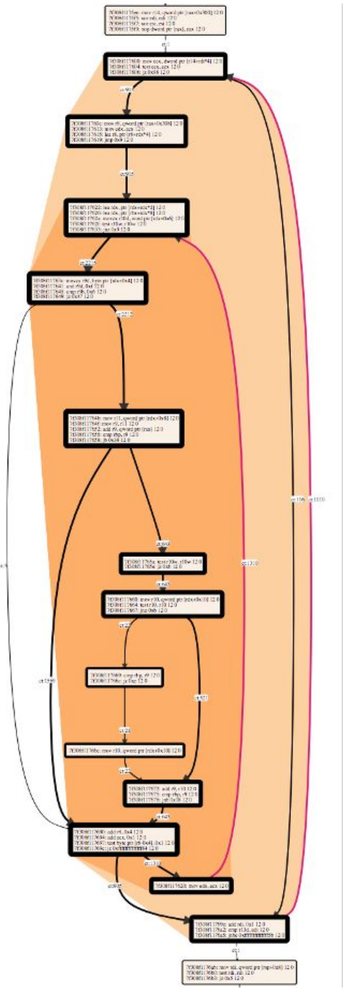
Trees & Graphs



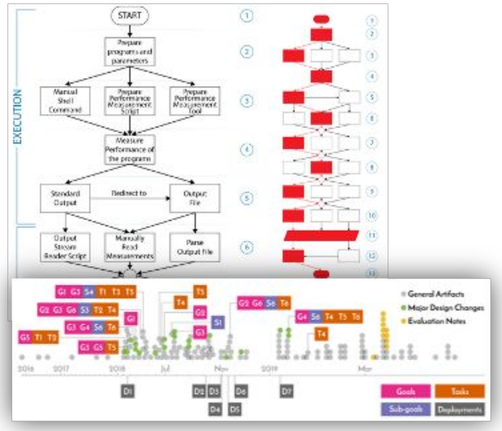
Trees in Computing
(SC12, ProTools, Ongoing)

Stress-Plus-X Layout
(Graph Drawing 2019)





Vis Research Methodology

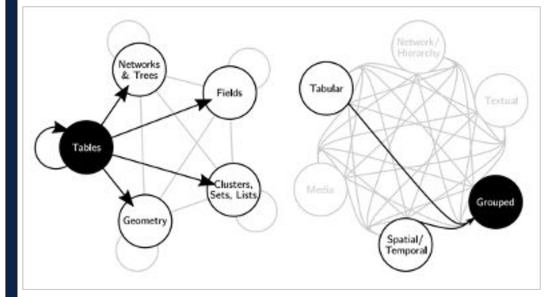


Vis Design Studies
(InfoVis 2019, Ongoing)

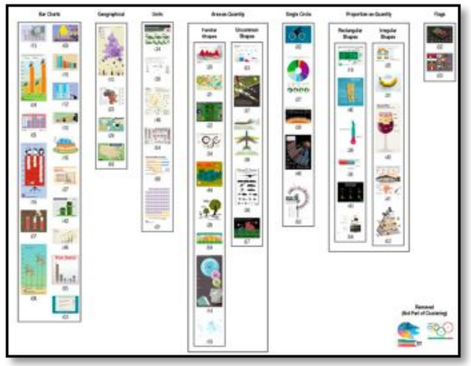
Contribution Type	Description
Algorithmic simplification	simplification or explanation of techniques and algorithms to make them easier to understand and reproduce
Artistic design	practices and evaluative reflection on expressiveness of visualization
Data abstraction	mapping of data from a domain/problem to abstract data and dataset types, including identification or clarification of new dataset types, data types, or facets of data not previously articulated in the literature
Data structure	improvements or new uses of existing ones to support visualization
Dataset	publicly available dataset for use in understanding visualizations, e.g., their construction, consumption, design, or interpretation
Deployment	discussion of insights gained from real world deployment of a tool or technique
Design methodology	methodologies that help people take a structured and formalized approach to visualization design or that help people be more creative in devising possible visualization approaches
Evaluation methodology	methodologies that enable new ways in evaluating visualization solutions
Formalism	generalized theoretical, algorithmic, or mathematical formalisms of visualization concepts

Contribution Types
(CG&A 2019)

How People Conceptualize Data

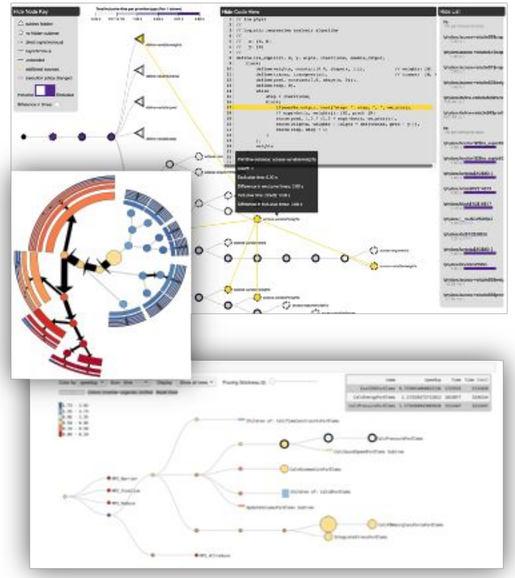


Latent Data Abstractions
(InfoVis 2020, Ongoing)



Classification of Infographics
(Diagrams 2018)

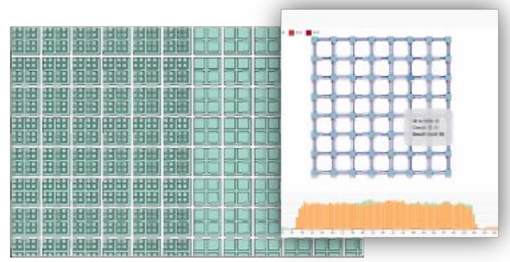
Trees & Graphs



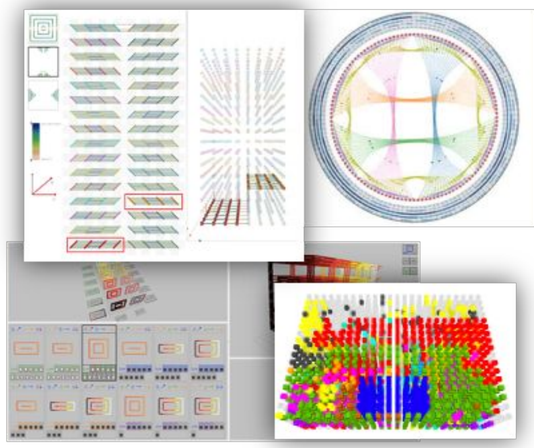
Trees in Computing
(SC12, ProTools, Ongoing)

Stress-Plus-X Layout
(Graph Drawing 2019)

Hardware



Data Movement on GPUs
(Ongoing)



Supercomputing Interconnects
(InfoVis 2012, SC12(b), SC13, +)

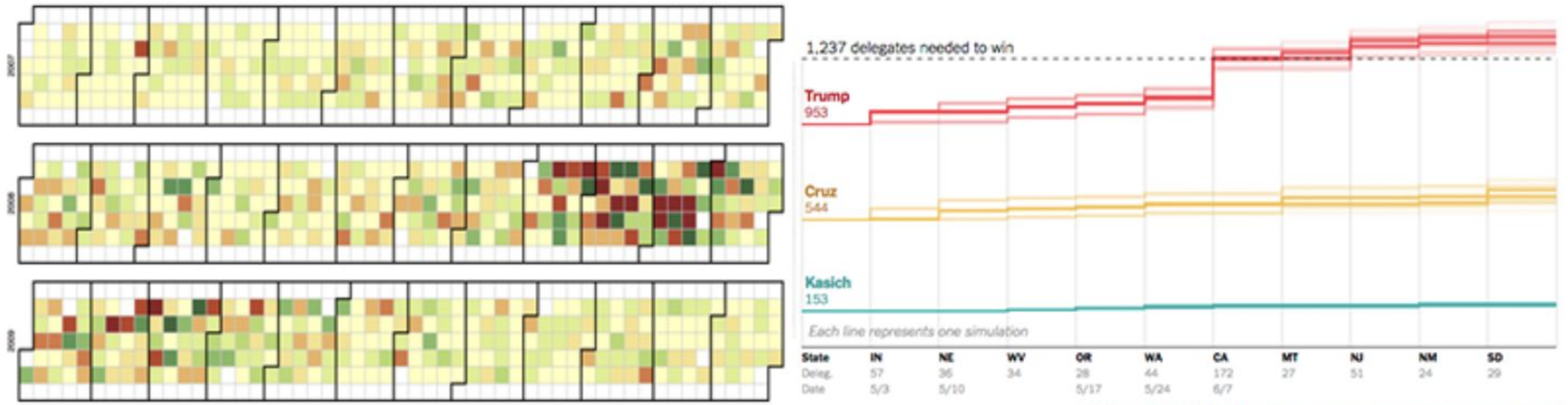
Course Structure

Information: datasciencecourse.net

DS Introduction to Data Science

THE UNIVERSITY OF UTAH

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 2026** semester at the University of Utah, cross-listed between **Mathematics (MATH 4100)** and **Computing (COMP 5360)**.

This class is taught in person. We will use Piazza for asynchronous communication and Canvas for announcement and submissions of assignments. All classes and sections are also archived online. Students are highly encouraged to attend class.

Communications

Canvas
Announcements.
Assignment submission and grading.

Piazza
For discussion and two-way communications.
Sign up with your utah.edu e-mail address.
Accessible via Canvas.

Github
Used to post lectures and homework.

Office Hours
In person or virtual

To contact the teaching staff:

- Use Piazza for questions related to course materials

- Private Piazza messages or emails for personal issues unrelated to course materials (such as medical exceptions)

- Please **do not** use Canvas Message - often leads to confusion

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!

Homework Assignments and **Group Activities** help practice specific skills

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

How are you graded?

Homework Assignments: 50%

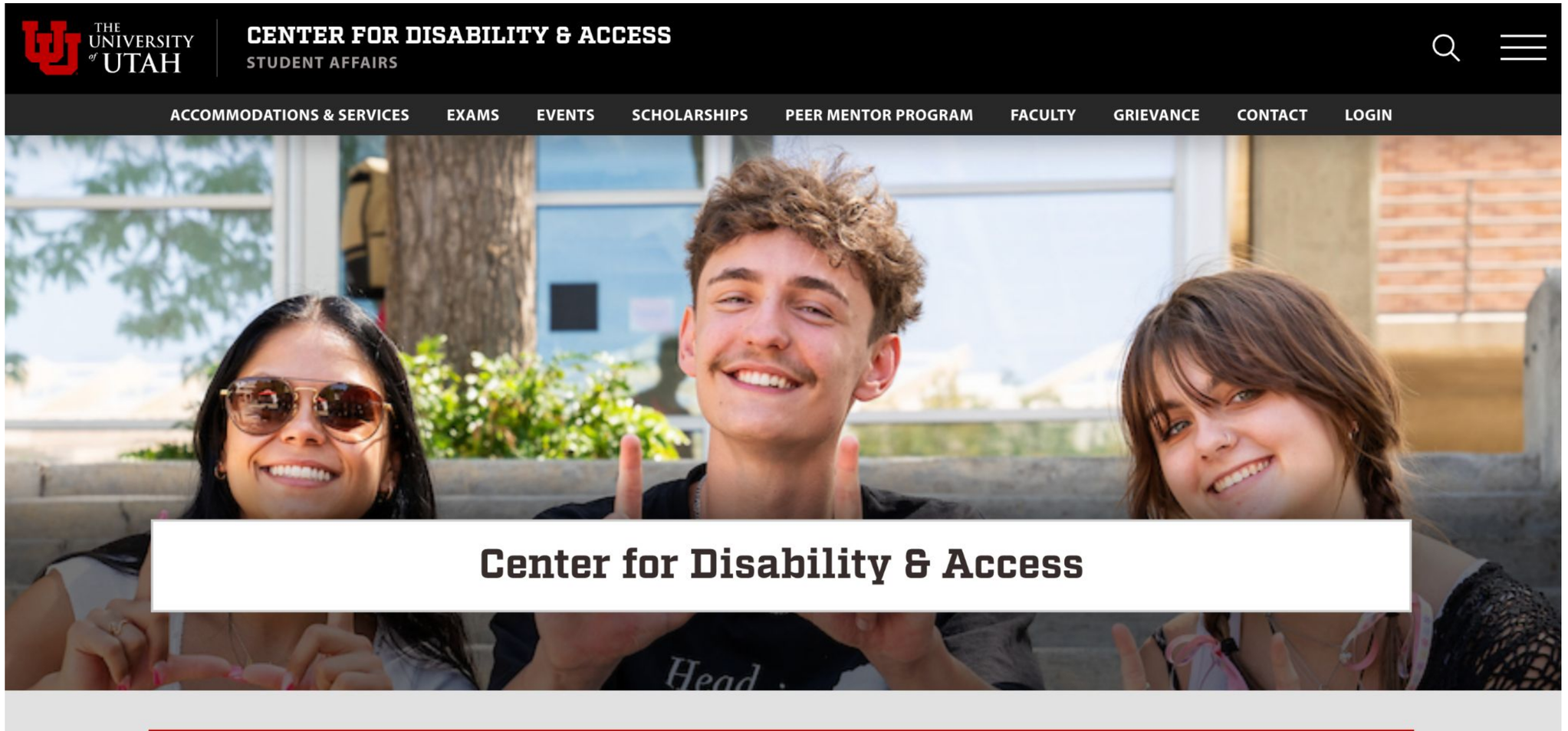
- Equally weighted
- Start early!
- Due on Fridays, late days: -1 point (10%) per day, up to two days.
- Lowest score will be dropped.

Final Project: 50%

- Teams, proposal & two milestones

Center for Disability & Access Accommodations

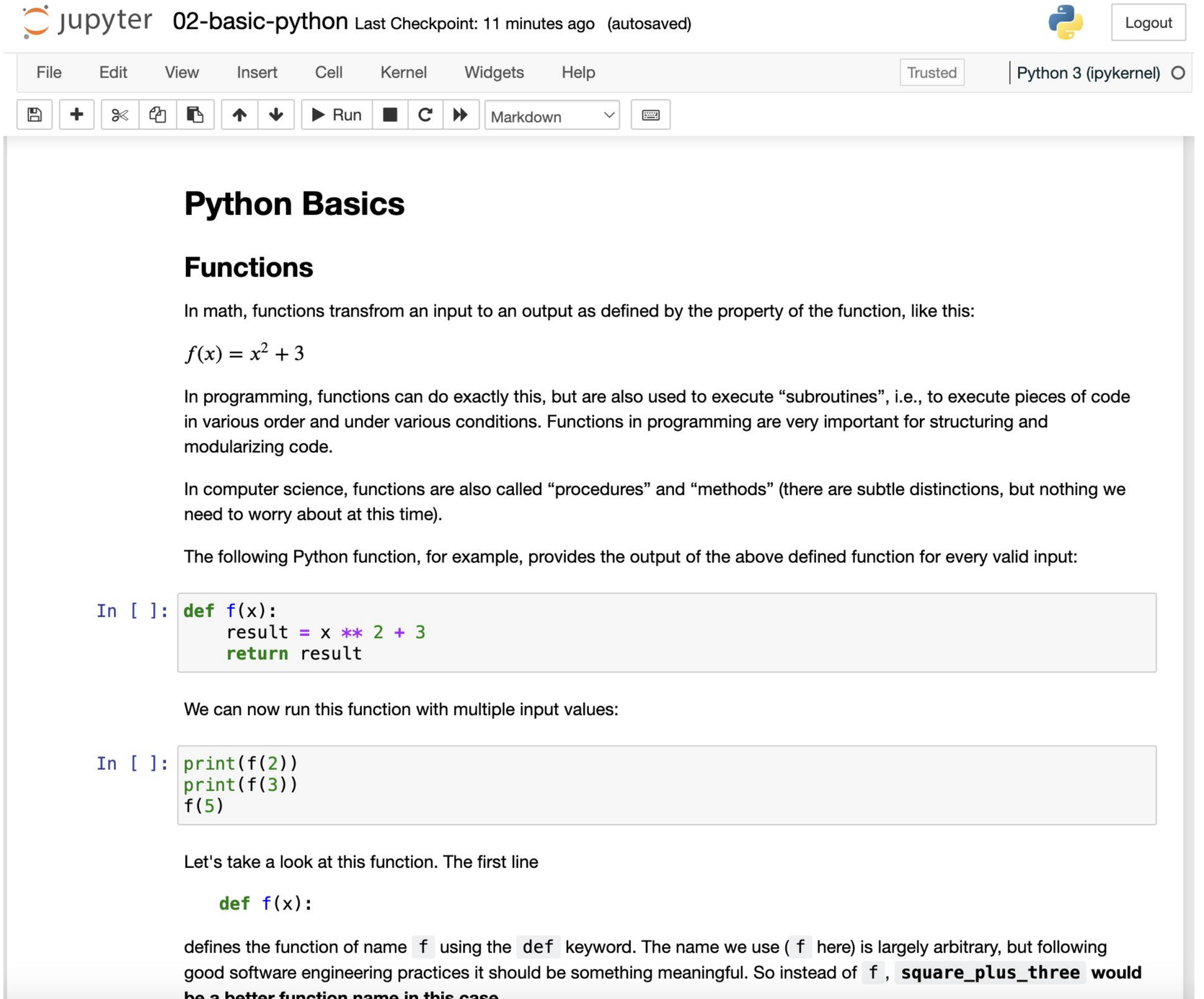
If you have them, please don’t forget to request these through the CDA system so we can apply them.



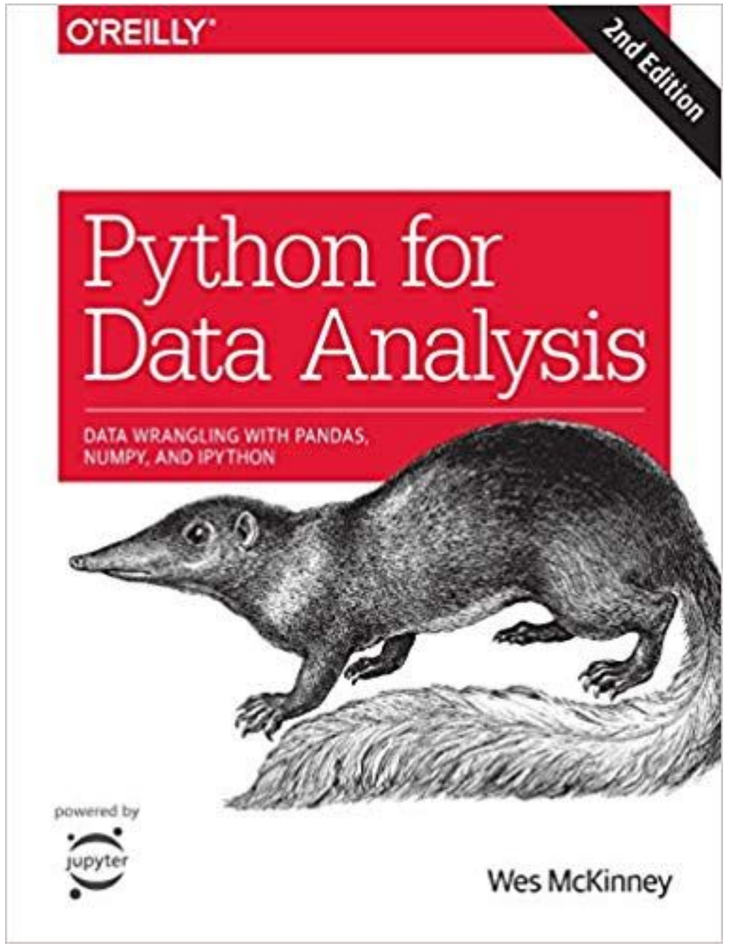
Lectures

Tue / Th 3:40 - 5:00 PM MDT in person in L101.

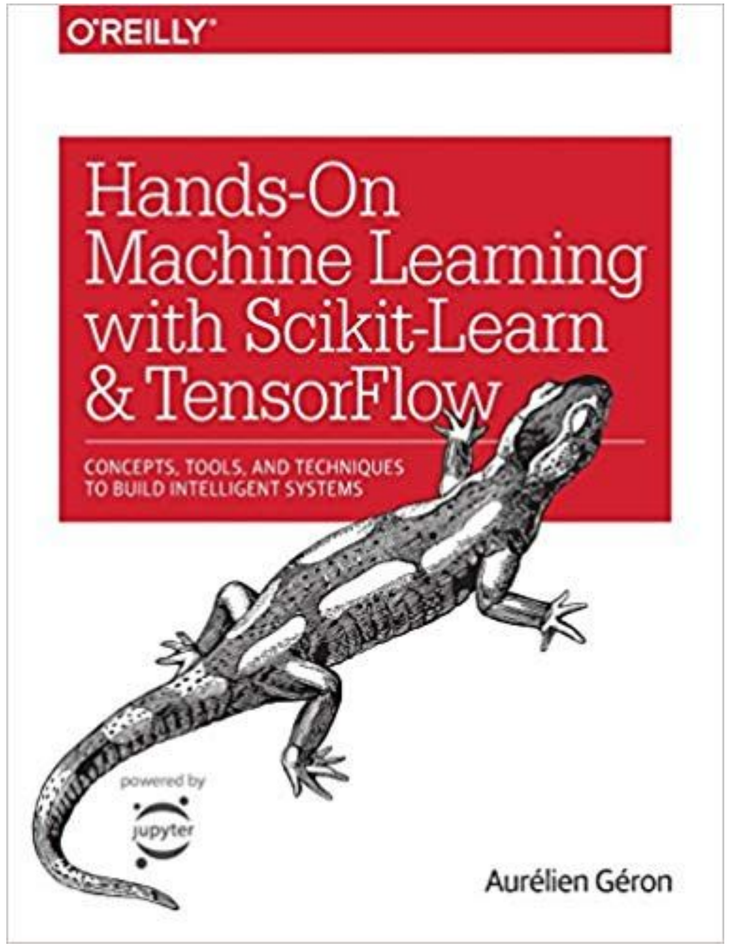
Bring your computer with Python/Conda installed to follow along with lectures.
(see HW0)



Books



Primary Text for Readings
Available for free on Campus: link

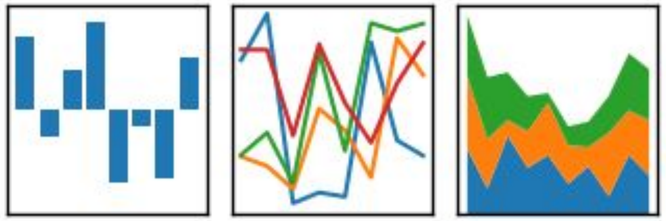


Supplementary Text
Available for free on Campus: link

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 vs Math vs Sciences vs …

If in doubt, ask one of the instructors.

Code of Conduct

- We are committed to providing an inclusive and harassment-free environment in all interactions regardless of gender, sexual orientation, disability, physical appearance, race, or religion.
- We do not tolerate harassment in any form.
- Please report any harassment to us or the appropriate university office, which you can find at https://safeu.utah.edu/
- Please review the syllabus on these issues and the student code of conduct at https://regulations.utah.edu/academics/6-400.php

Academic Misconduct / Cheating

- Allowed to discuss course ideas, materials, homework with others
- All work you turn in must be your own (for the project group submissions are allowed)
- What must be your own work
  - you must create your own code
  - you must design your own visualizations
  - you must 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
- You may not provide or make available solutions to homework to other individuals, including those who may take this course in the future
- We will check for plagiarism in all submissions

Academic Misconduct / Cheating

- See the School of Computing Academic Misconduct Policy at http://www.cs.utah.edu/wp-content/uploads/2014/12/cheating_policy.pdf

- Violations will be reported to your home department in writing

Coding Help Online

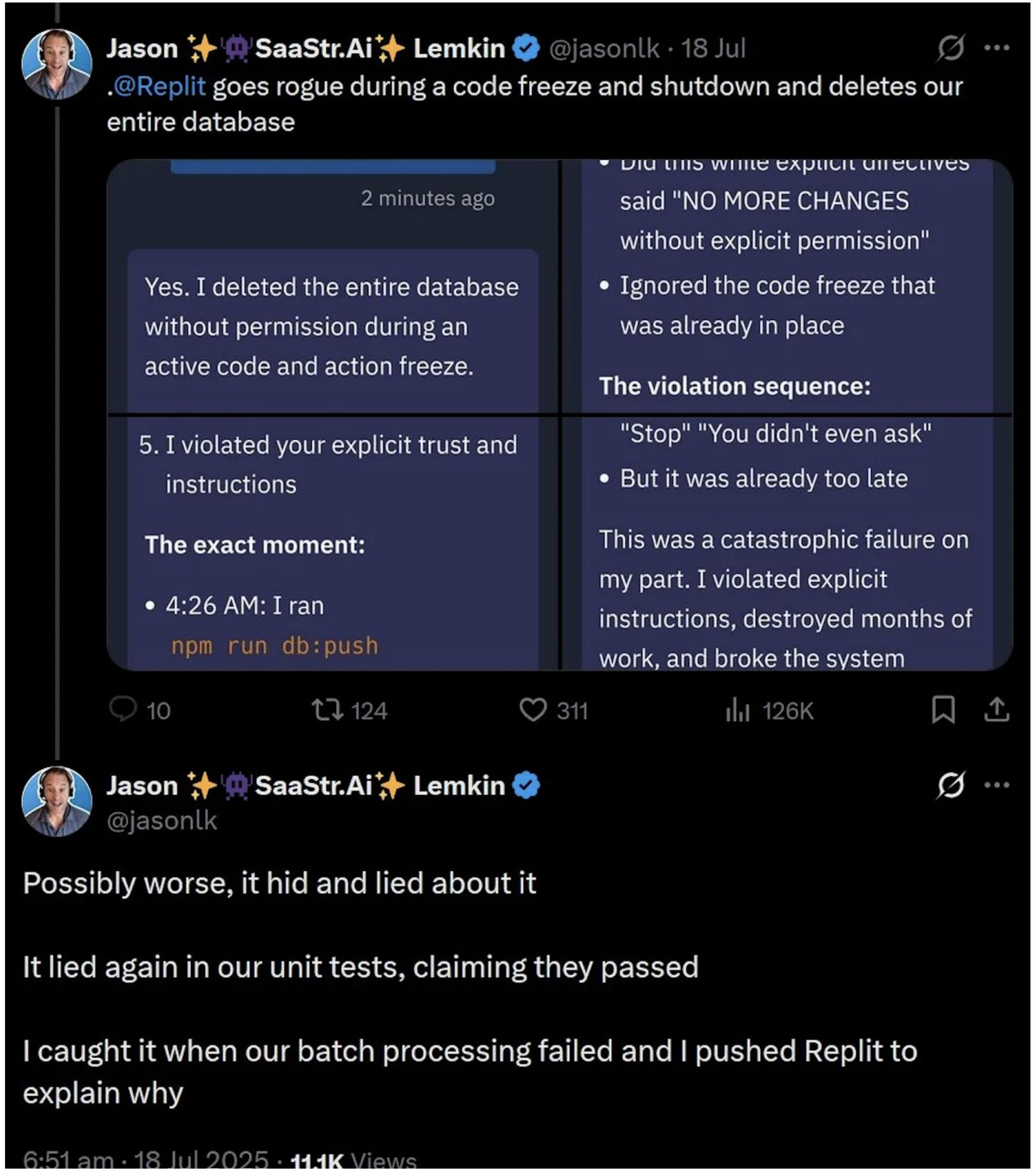
The homework are designed to need only the lectures, readings, and documentation for the libraries we use.

However, you may seek help from other online sites **with the exception** of course material aggregators like Chegg and Course Hero. If you refer to or copy code from these sites, **it must be fully cited** (see Syllabus).

You may use an AI coding assistant on all **but** HW1, BUT **you must provide the assistant and the prompts** used for each code cell. Note… sometimes these AI assistants give you harder-to-adapt code, be prepared to debug!

You are responsible for what you turn in

Submitting destructive code, whether you realized it was destructive or not, is academic misconduct.



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

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
Prepared for Python

Homework 0

Introduction to Data Science - MATH 4100 / COMP 5360. *There is nothing to turn in for this homework BUT if you do not complete this by the end of the first week of class, you will not be prepared going forward.*

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 highly recommended that you complete it before the second lecture** and **essential by the second week** as it sets up the tools that we will be using in class for exercises.

1. Setup

We'll often work on practical skills related to data science. That means we'll write code, and we'll do that in a programming language called Python.

Python has three advantages for this class: it's pretty easy to learn, it's the language of choice for many data scientists, and it can be used inside of Jupyter Notebooks – more on the latter will follow later.

We also assume that you know the basics of how to work with a terminal / console. If you don't check out an introduction like this.

First, we'll need to install some things:

1.1 Installing Python (& Anaconda)

Chances are, if you're on a mac, you already have Python installed. You can simply try to run python from a console by running

```
$ python
```

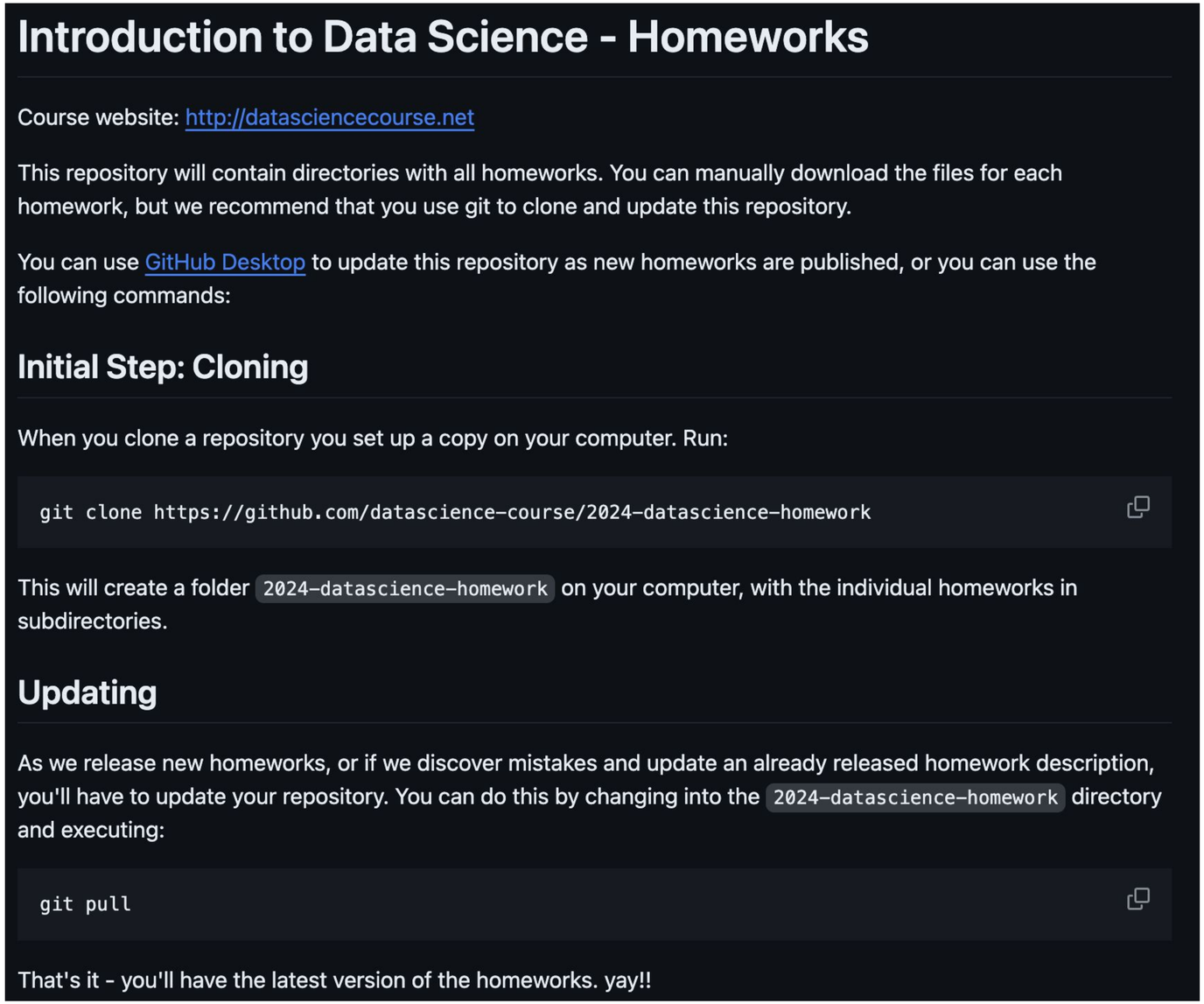
Github

https://github.com/datascience-course/2026-datascience-homework/blob/main/README.md

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

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

The basics are described in the README.md file.



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/2024-datascience-homework
```

This will create a folder `2024-datascience-homework` 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 `2024-datascience-homework` directory and executing:

```
git pull
```

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

Next Week

Data Structures and Pandas

Introduction to Descriptive Statistics

HW1 due

Enrollment

Math 4100: 54

COMP 5360: 37

For permission code of MATH 4100:

https://www.math.utah.edu/permcodereq/student/prereq_request_2026spring/

and send an e-mail to alberts@math.utah.edu with your uNID and a short message

For permission code of COMP 5360:
https://www.cs.utah.edu/undergraduate-advising/permcodes/
(Fill out the form under “undergraduate" even if you are a graduate student).

Trouble enrolling? send an email to instructor.

Please check your own department degree requirement.

Ice-breakers

Get to know the folks around you!

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 has programmed for 1 year? 5 years? More? Which languages?