Capstone I

DSC 180A
Winter 2020

Today’s Outline

- Introduction
- The five “domains of inquiry”
- Course structure (Lec/Di/Lab)
- Syllabus + quarter I assignments

Course Resources

- Lecture Repository: https://github.com/afraenkel/DSC180A-DS-Methodology
  - Contains links to discussions, syllabus, etc.
- Computing Resources:
  - https://datahub.ucsd.edu
  - Your own computer
- Gradescope for code assignments and (some) reports.
- Course communication: Piazza for lecture and (likely) domains.

Capstone Sequence Course Goals

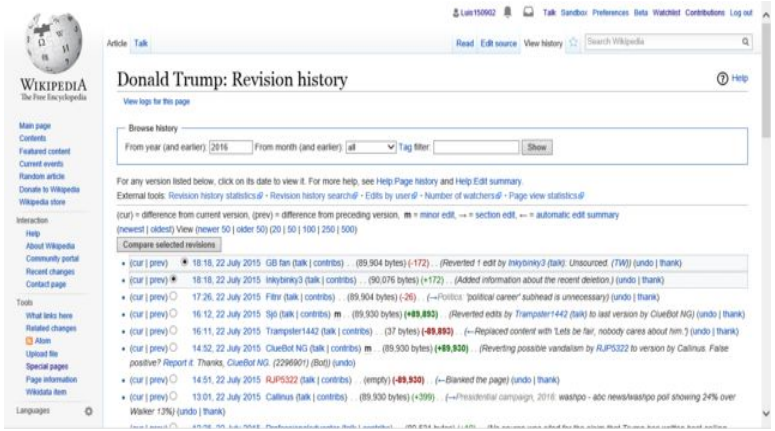
- Put together DSC skills through the lifecycle of a two-quarter project
- Learn methodological best practices for projects:
  - Reproducible and flexibly generic work
  - Effective (visual, oral) communication of work and results
- Starting an investigation with a *question* instead of a method.
- A detail-oriented pursuit of a proposal in a chosen domain.
- Produce and show off work that you are *proud of!*

Domains of Inquiry

- Choose and learn a domain of inquiry => project proposal => execute!
- Data Science capstone projects: sparse dataset in a high-dimensional space.
- Reasons to choose a domain:
  - The domain itself
  - The kind of data involved
  - The kind of methodological tooling required
- Can choose/change a domain during the first week of the quarter (space permitting).

Domain: “Wikipedia Edit Wars”

- Faculty Advisor: Prof. Molly Roberts
- Graduate Assistant: Keng-Chi Chang
- Domain Subject Keywords:
  - Social media; online conflict; information control
- Primary Data Types:
  - Unstructured text, webpage metadata
- Primary Methodological Tools:
  - ; NLP; causal inference



Domain: Quantitative Measures of Artistic Style

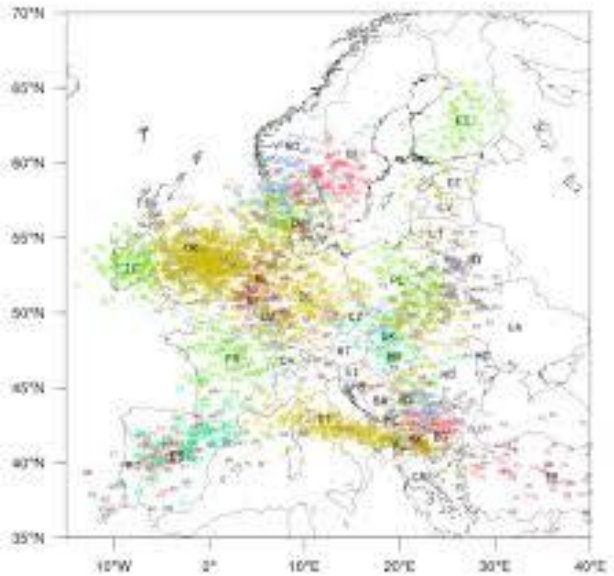
- Faculty Advisor: Prof. Robert Twomey
- Domain Subject Keywords:
  - Visual arts; artistic style
- Primary Data Type:
  - Images
- Primary Methodological Tools:
  - Image analysis; CNN

Domain: Fair and Predictive Policing

- Faculty Advisor: Prof. Aaron Fraenkel
- Graduate Assistant: Rebecca Fraenkel
- Domain Subject Keywords:
  - Discrimination, Data Journalism
- Primary Data Type:
  - Administrative Data, Geo Data, Time Series.
- Primary Methodological Tools:
  - Causal Inference

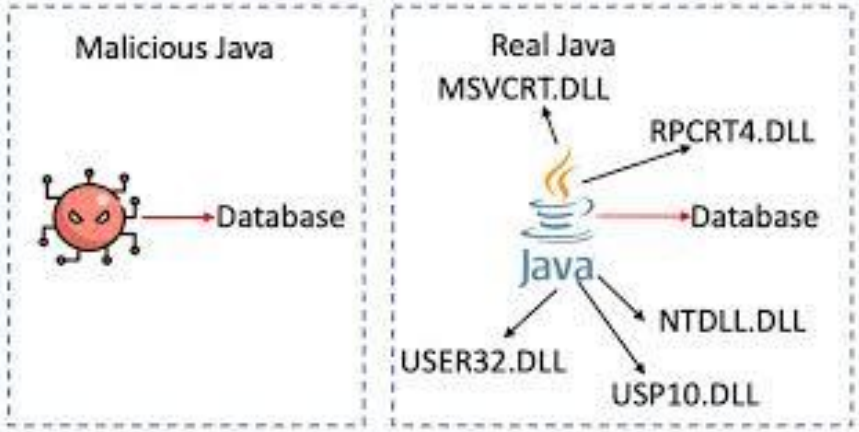
Domain: Clustering the Human Genome

- Faculty Advisor: Prof. Shannon Ellis
- Domain Subject Keywords:
  - Genetic Ancestry; Gene Sequencing
- Primary Data Type:
  - Genetic Sequence Data (POPRES)
- Primary Methodological Tools:
  - Clustering; Outlier Detection; Statistics



Domain: Malware and Heterogeneous Graphs

- Faculty Advisor: Prof. Aaron Fraenkel
- Graduate Assistant: Shivam Lakhotia
- Domain Subject Keywords:
  - Cyber-Security, Code Learning, Graphs
- Primary Data Type:
  - Computer code
- Primary Methodological Tools:
  - Graph Techniques



Course Structure: 3 components

Lecture: methodological guidance (Mon)

Discussion: learning the chosen domain in groups (Wed)

Labs: one-on-one help with work (Fri)

Each is important and *required*

Lecture Goals

- Learn and use Data Science best practices for projects
  - Apply *directly* to project in domain!
  - Projects will be graded according to using these best practices
- Topics:
  - Flexible, organized project structure and software development
  - Using multiple environments
  - Version control
  - Deployment and builds
  - (Responsible) data handling
  - Broader impacts and ethics

Discussion Section Structure

- Learn the domain and pursue a proposal *guided* by domain expert.
  - *You* are responsible for learning material and doing data analyses.
  - Come ready each Wednesday to *actively* discuss the material and results.
  - **Coming to section prepared is mandatory and necessary for the success of capstone!**
- Discussions are for engaging contextual questions, data, and conclusions.
  - Sections should *not* deal with coding problems.
  - Data Scientists must translate problems about code into the language of the domain/data.
- Discussion are for:
  - Engaging with the domain and the questions at hand.
  - A place to ask for clarification about the data generation process.
  - Brainstorm with peers about how possible proposals.

Syllabus

Component	% of Grade
Methodology HW	10%
Discussion Section Participation	10%
Domain result replication (3 reports)	30%
Domain result replication (workflow)	20%
Project proposal	30%

Assignments: Learning the Domain

1. The Data
    a. Report: Explanation of the DGP and schema; design choices.
    b. Code: Data ingestion code, using best practices.
2. Cleaning and EDA
    a. Report: Summarize results of EDA; defense of choice of cleaning code.
    b. Code: Cleaning and EDA code, using best practices.
3. The Result Replication
    a. Report: Summary of result of the ‘replication’, shortcomings, and possible improvements.
    b. Code: Replication code that produces results, using best practices.

Assignments: The Proposal

- Worked on in groups (same as your project).
- Write and submit a proposal, with background research.
- Rehearse and deliver a 2-3 minute elevator pitch (general audience)
- Create a skeleton workflow for the project (github repo with boilerplate).

Your group will work on and present the project in Quarter 2!