UCSD CSS M.S. Program Capstone Proposals

Academic Year 2023-2024

Contents

1 Note from the CSS Program

These proposals are listed in order of submission, below.

- Students should go over this list, talk with their colleagues to determine which capstones are of interest to them, and form groups (according to the 'number of students' constraints).
- Students should be prepared to provide three ranked choices ("First, second and third choice"), particularly if there are more students interested in a particular capstone than slots available
- The surest way to 'guarantee' your capstone choices is by collaborating and negotiating with all of your colleagues, such that you can walk in to the last day of the bootcamp as a group with a fully settled document listing each student below their capstone choices. We have no reason to 'assign' differently if everybody's content with an arrangement.
  - If there's no group consensus, and more students want to join a capstone than there are slots available, students will be assigned randomly using something like:
  - `import random; print(random.sample(['name1', 'name2', 'name3', 'name4', 'name5', 'name6'], 4))`
  - With remaining students being reassigned according to their ranked choices
- If you have questions or concerns about a particular project which you're interested in, or would like to see (e.g.) if there's any flexibility on number of students, please contact the CSS Director

2 Evaluating California Public Utility Commission Integrated Resource Planning Analysis and Results

Website: https://centerforcommunityenergy.org/

Host: Jose Torre-Bueno jose.torrebueno@cc-energy.org Executive Director

2.1 Background:

Because utilities in California are regulated monopolies, they are guaranteed a profit on any investments they make if those investments are approved by the Public Utility Commission (CPUC). The process for doing this is called Integrated Resource Planning (IRP). For this, utilities submit proposed projects, and the CPUC runs a program called RESOLVE that is supposed to determine the electric generation and transmission resources needed that will provide reliable electricity at the lowest possible costs.

One of the inputs to RESOLVE is the assumed value of consumer solar arrays, which are referred to as behind-the-meter, or BTM. Before the most recent IRP round, the CPUC commissioned an "Avoided Cost Study" which came up with an extremely low value for BTM solar power. Using this as input, RESOLVE calculated that consumer BTM solar is not desirable, and that large scale remote solar arrays should be preferred. This then becomes the argument that increased investment by the utilities and increased rates for consumers would be needed.

Another technology that many experts think could help solve energy availability issues in California is "vehicle-to-grid" (V2G). Although V2G has been extensively researched, modeled, and tested in Europe, the CPUC did not include it at all in its RESOLVE analysis. Many back-of-the-envelope calculations, however, suggest that V2G could save a great deal of money, far more than projects the CPUC has approved or is considering.

2.2 Objectives:

The overall objective is to see if the IRP comes out differently if different assumptions are used. There are several points at which this can be tested 1. Get access to the Avoided Cost Calculator. Other organizations that have checked inputs have noted multiple incorrect assumptions. This work would require re-running with corrected assumptions and validating new inputs to RESOLVE at the earliest points in the process.

2. Regardless of how new inputs are derived, the RESOLVE model needs to be rerun with input values more in line with the values that environmental organizations recommend. We believe but currently have no way to prove to the CPUC that their policy decisions in the IRP are a consequence of invalid input assumptions. Only by running RESOLVE with our own inputs can we prove this.
3. An extension of #2 above would be to set up an environment to repeatedly run RESOLVE systematically varying the input parameters. As far as we can tell, the RESOLVE model has never been subjected to input sensitivity analysis. Since all inputs to RESOLVE are estimates of one type or another, it should not be used without understanding how sensitive the model is to input uncertainty.
4. The RESOLVE model was run without any effort to include V2G or even optimize the time-of-day EVs charge. We know that a version of RESOLVE that is supposed to incorporate V2G is under development, but that version has not been put into the public domain. This does tell us that with some Python programing could be done to add V2G to RESOLVE. This would be very important, because some simple modeling suggests V2G would have a profound impact on the resources that will need to be built in the future.

2.3 Resources and Constraints:

The Center for Community Energy (CCE) is a 501(c)3 nonprofit organization. Its mission is to promote distributed energy resources such as BTM solar. CCE is registered as an intervenor before the CPUC which means it can speak

for ratepayers in CPUC proceedings. It also means CCE is automatically notified of certain classes of filings and actions at the CPUC. All the data we will use in this project is publicly available by law. We believe that the RESOLVE model will run on a laptop. Multiple runs to explore the parameter space as described in 3 above many require access to a mainframe.

2.4 Anticipated Challenges:

Much of the information the CPUC maintains is not posted in a way that makes it indexable by search engines. Locating needed information on the CPUC's servers is a challenge. However, CCE can provide guidance and training on accessing this data. Given the lack of indexing, part of the projct will necessarily require a hunt for the most current and relevant data. To be credible, it will be important for us to find the dataset(s) that were actually used by the CPUC before issuing final rulings. There will also be many data sets used in test runs and hypothetical cases. It is possible that some critical information is missing or redacted, in which case it may be necessary for CCE to use its intervenor status to issue appeal(s).

2.4.1 Are there Citizenship Requirements?

No.

2.4.2 Are there confidentiality requirements?

No.

2.4.3 Number of Students:

single student., group of two students., group of three students., group of four students., group of five students (we're trying to keep it to four, but may make exceptions for exceptional student desire)

2.4.4 Prequisite Skills beyond CSS Basics:

No."

3 MRDeep: Multilevel Regression and Post-Stratification using Deep Learning

Website: https://css.ucsd.edu

Host: Umberto Mignozzetti umbertomig@ucsd.edu Assistant Teaching Professor

3.1 Background:

Political Scientists have been relying on multilevel regression and post-stratification to extrapolate national survey results and estimate features of interest (e.g., ideology, party affiliation, opinions regarding climate change mitigation) at the local level. In this project, we will create a method to run multilevel regression and post-stratification using deep learning. The main innovation is that by customizing the architecture to fit the problem, we will be able to get estimates not only for cross-sectional data, but also for panel data.

3.2 Objectives:

I will teach the students how deep learning works at a mathematical level, and we will coordinate how to implement the architectures we need for the different problems that the users may face. In the end, we will have a wrapper for Python and a wrapper for Julia for the algorithm of our choice. A successful project will:

1. Having the students acquire a solid understanding of deep learning concepts and techniques
2. A Python wrapper using TensorFlow to have users easily estimate their models.
3. A Julia wrapper using TensorFlow to have users easily estimate their models.
4. A coauthored Journal of Statistical Software paper will be their capstone project's basis.

3.3 Resources and Constraints:

I have the datasets and the prior code all written. I also have the main paper and the lit review. The main constraints for execution of the project is that students will have to familiarize with Julia. I plan to teach them an intro to Julia during the bootcamp, and the students working with me will be using it frequently to sharpen their knowledge.

3.4 Anticipated Challenges:

1. Fall 2023: Students will be meeting with me on a weekly basis and we will study deep learning. This will require time and dedication, in a time that they will already be overwhelmed with classes.
2. Winter 2024: I will be working with students on the Python wrapper. Each student will have a weekly coding task that has to be accomplished. By the end of the quarter, we should have a beta version of the Python package up and running.
3. Spring 2024: I will be working with students on the Julia wrapper, debugging issues on the Python wrapper, and writing up together the JSS paper. By this point, they should have their capstone done.

3.4.1 Are there Citizenship Requirements?

No.

3.4.2 Are there confidentiality requirements?

No.

3.4.3 Number of Students:

group of two students.

3.4.4 Prequisite Skills beyond CSS Basics:

There are softwares that are not conventionally used in the CSS MA track."

4 Auditability, Transparency, and Accountability in Centralized Platforms

Website: https://sites.google.com/iset.ge/aramgrigoryan

Host: Aram Grigoryan a2grigoryan@ucsd.edu Assistant Teaching Professor

4.1 Background:

Many economic interactions take place in centralized platforms, where a central authority controls the data, functions, and algorithms that constitute it. For example, school districts with open enrollment increasingly rely on

centralized admissions through state of the art matching algorithms. With the rise of the internet, goods and services are ubiquitously sold by auction mechanisms in online platform. Centralized clearinghouses are especially common for allocating goods and services of high public significance, such as education, labor markets, organ transplantation, public housing, auctions for spectrum licences, carbon emissions, and electricity.

Centralized platforms operate as follows. First, a central authority (the designer) announces the regulations and rules (the mechanism) of allocation. Second, the participant share their data inputs (reports) to the designer. Lastly, the designer implements an allocation by the promised mechanism. For example, in public school admissions, the central authority are the representatives of the school district, the participants are the families with school age children in the district, and the data inputs are their school ranking lists, documentations of place of residence, etc. An important feature in all these centralized allocation platforms is that participants do not observe the reports of other participants. Hence, if during the implementation stage the designer does not use the promised mechanism to implement an allocation, the participants may have no way to observe and detect this deviation. This raises an important question of which mechanisms are transparent or easily auditable, and which are not.

In a recent series of paper, my coauthor and I investigate questions of transparency and auditability in allocation porblems (Grigoryan, 2022; Grigoryan and Moller, 2023). These questions are very timely and relevant given the salience of transparency concerns in various centralized platforms, and the general importance of algorithmic transparency and accountability for matching. More precisely, we quantify a mechanism's auditability by a measure called an auditability index, which is the size of smallest group of participants that detects a deviation. For example, if a mechanism has an auditability index of 3, it means if the designer happens to use a wrong mechanism to find an allocation, there would be some 3 individuals, whose information alone would be sufficient to detect this deviation. Hence, a smaller auditability index means that a mechanism is easier (requires less information) to be audited by the participants. We apply this theory for a broad range of centralized allocation environments.

4.2 Objectives:

There are many open questions and problems. One of them is measuring auditability index for mechanisms and problems with a computational method (simulations), as oftentimes this cannot be done analytically. In Grigoryan and Moller (2023), we have analytical conditions characterizing when a mechanism has a low or high auditability index. To understand when and how often these conditions hold, one may need to check them with a computer (e.g., brute force or smarter algorithms).

This would be a computational project solely based on simulated data, and computer generated proofs. Students will need to have working knowledge of coding (ideally, in python). Basic matching theory knowledge can be acquired during the course of the capstone project.

4.3 Resources and Constraints:

Data will be simulated. Hypotheses and open questions are known and well-defined, so it will be easier to formulate the concrete research plan.

4.4 Anticipated Challenges:

The main challenge would be to come up with efficient algorithms for computing the auditability index of various mechanisms. Basic knowledge of matching theory may be needed to understand how different mechanisms operate.

4.4.1 Are there Citizenship Requirements?

No requirements.

4.4.2 Are there confidentiality requirements?

the data will be simulated (or we may also use public data), so there are no IRB restrictions

4.4.3 Number of Students:

single student., group of two students., group of three students., group of four students.

4.4.4 Prequisite Skills beyond CSS Basics:

knowledge of basic coding in python should be sufficient to get started with the project. knowledge of algorithms, and economics is a plus, but not required"

5 Price Philanthropies Foundation Mobility Study

Website: protolab.ucsd.edu

Host: Steven Dow spdow@ucsd.edu Associate Professor of Cognitive Science

5.1 Background:

Price Philanthropies Foundation (PPF) have built four affordable housing developments (and several others in development) for the City Heights neighborhood in San Diego. The limited parking spaces have strained mobility options for the low income residents. PPF believes the current public transit options (e.g., bus/trolley) do not serve this population well. Towards the goal of creating a new rideshare pilot program, PPF has asked UCSD's Design Lab to design and analyze a household mobility survey. The survey instrument is a first step to guide how we might structure and execute a rideshare pilot for the City Heights residents.

5.2 Objectives:

The capstone team will assist with designing and analyzing a survey that will be completed by hundreds of residents in the housing developments supported by Price Philanthropy Foundations. The team will also assist in creating and analyzing a dataset to understand existing mobility conditions (i.e., an inventory of physical assets, transit lines, covered stops, parking spots). A successful project will include an extensive mixed-method analysis of datasets, a 7-8 page report on insights that could be used to structure a rideshare pilot program, and an in-person presentation to stakeholders.

5.3 Resources and Constraints:

Professors Steven Dow and Mai Nguyen will be available to advise the CSS Capstone team; the PPF partners will co-design the survey and will lead the effort to interview residents and to collect results for the survey instrument.

5.4 Anticipated Challenges:

The project execution will be dependent on PPF to collect the survey data, and so there will be a contingency plan to analyze complementary data in case there are any delays.

5.4.1 Are there Citizenship Requirements?

Not that I'm aware of

5.4.2 Are there confidentiality requirements?

Not that I'm aware of, but we will need to verify this with Price Philanthropies Foundations if the project is accepted.

5.4.3 Number of Students:

group of two students., group of three students., group of four students.

5.4.4 Prequisite Skills beyond CSS Basics:

There should be members of the team who have some experience with qualitative analysis (e.g., extracting insights from quotes) to complement the quant analysis. "

6 Social Network Analysis in Education

Website: www.socionomy.net

Host: Christoforos Mamas cmamas@ucsd.edu Assistant Professor

6.1 Background:

My team and I are working on developing a social network analysis toolkit for teachers to use in K-12 settings. I would appreciate data scientists to help us with data visualizations on online environments (as the toolkit is web-based), algorithms to crunch data from students, etc, programming skills to contribute towards developing the toolkit further and other relevant support.

6.2 Objectives:

To help with debugging the toolkit To assist with data visualizations and data analysis To support with programming the toolkit and add and revise functions.

6.3 Resources and Constraints:

There are some limited resources to continue working on the project.

6.4 Anticipated Challenges:

You will be part of a team if you participate in the project. Good communication and planning would be needed to work with others. Sticking to deadlines is important as well as managing your project tasks along with your other commitments. Your involvement to the project will be around 3-5 hours weekly.

6.4.1 Are there Citizenship Requirements?

N/A

6.4.2 Are there confidentiality requirements?

Some restrictions would apply when analyzing data but this would not prevent students from completing the necessary work for their project.

6.4.3 Number of Students:

single student., group of two students., group of three students.

6.4.4 Prequisite Skills beyond CSS Basics:

We can't teach python or any other programming languages. We expect students to have knowledge of at least data analysis and visualization techniques and some intermediate to advanced programming skills. "

7 Participatory Sports Results Platform

Website: www.7sherpas.com

Host: Chris Kittler chris@7sherpas.com Founder / Director

7.1 Background:

Every weekend, there are thousands of running, triathlons, cycling races, swim meets, and other sporting events happening worldwide. All these sports generate a massive amount of data, that is usually published in the organizer's websites, storing relevant metrics such as all participants names with times, splits, pace, placing, age group, type of event, etc. However, there are at least three shortcomings in storing these metrics: First, they lack organization. Second, is not being analyzed.

Finally, participants have no place to save their results for future reference. Bringing a solution to this problem is important because sports enthusiast all over the world could save their accomplishment in one unique platform, they can also track their improvement over the years and the whole industry can benefit from the analysis of the data.

7.2 Objectives:

In this project, we will solve the problem by creating a new standard to organize, save, and analyze these results. We will develop the following products: 1) Help build a standardized platform for event organizers to publish results. 2) Build a method for participants to claim their results and store them on a personal profile. We will also perform some machine learning validation to avoid wrongfully claiming other people's results. 3) Create methodologies to analyze clusters based on performance, rank community, and use AI to make suggestions, stats, and interactions with the various profiles.

7.3 Resources and Constraints:

All these results are publicly displayed on websites, and we also can have access to the original XLS / CVS format from the timing companies. However, the format and fields will vary based on the event format and type of sport.

7.4 Anticipated Challenges:

There are several challenges that are important from the perspective of a Computational Social Scientist. To mention a few: 1) Match old records/results to the right individual, across different sports and competitions. 2) Harmonize the information that refers to the same person with different identifiers (e.g., a person named Paul Roger Smith can appear in the databases as Paul R Smith, Paul Smith, Paul Roger, etc.) 3) Using Machine Learning to detect and remove records that were wrongfully claimed.

7.4.1 Are there Citizenship Requirements?

No

7.4.2 Are there confidentiality requirements?

Yes. A confidentiality term must be signed since the project involves personally identifiable data.

7.4.3 Number of Students:

single student., group of two students., group of three students., group of four students.

7.4.4 Prequisite Skills beyond CSS Basics:

No”

7.5 Note from the CSS Director:

The Host of this particular capstone is very enthusiastic, but has been very worried about confidentiality in correspondence. He's opened up substantially from the above position, and states that the results will not be confidential, but I'm still waiting for copies of any NDAs or confidentiality documents he wishes students to sign. Although he shows signs of backing off the initial 'all code produced is a trade secret' approach, if having a fully transparent capstone, with all code open source and all analysis publicly available is particularly important to you, this Capstone likely should not be your choice.

8 Sharing Media Industry Data with the Public

Website: https://democracylab.ucsd.edu/what-we-fund/media-and-consolidation-research-organization-macro-lab

Host: Shawna Kidman & Andrew deWaard skidman@ucsd.edu / adewaard@ucsd.edu Associate Professor / Assistant Professor

8.1 Background:

We study the consolidation and financialization of the media business, a process which limits storytelling, diversity, and opportunities for creative workers in film, television, music, news, gaming, and publishing. Even though these trends intensify every year, they are extremely difficult to track and sometimes hard to understand, especially for those who lack financial literacy or data expertise. Accordingly, much of the academic work, journalistic reporting, and policy decisions around media and media production lack the high quality empirical data that Wall St. investors and media executives use to make decisions behind closed doors. This information asymmetry is not in society's best interest.

8.2 Objectives:

We aim to create a research community around a polished, publicly accessible website that analyzes, visualizes, simplifies, and explains the media industry data we've been gathering with the help of student researchers for the past two years. We want anyone to be able to visit this website, learn about media consolidation, find data of particular interest to them, and use dashboards to create graphs or charts that they can use in their own work (dissertations, articles, petitions, reports, etc.). To achieve this goal, we need to organize and analyze what we have, see what kind of new insights we can generate, and produce the visuals to easily communicate these insights, in effect turning hard numbers into narrative.

8.3 Resources and Constraints:

This project is housed under the MACRO Lab (https://macrolab.vercel.app/), which has received small grants from the Democracy Lab in the Department of Communication and from the Division of Social Sciences. We've used these funds to develop a website we aim to maintain for many years into the future. We've also used funds to pay for data as well as for researchers to clean the data and to collect additional data. In order to publish, we need to transform some of what we've purchased through analysis; and we need to find more and better ways to turn the raw data we've gathered into meaningful visuals. Our data includes financial info on individual companies (such as revenue and market capitalization), records of mergers and acquisitions, data about decades of film and television titles, and information related to individual sectors/topics, e.g. streaming subscriber numbers, movie theater & concert tickets, DVD & CD sales, venture capital investments in media, etc.

8.4 Anticipated Challenges:

Learning more about how the media and tech industries work. Making financial data meaningful, clear, and interesting to a public that lacks financial literacy. Combining data from multiple sources that is not always consistent. Figuring out how to use data about film, television, and other creative work to produce meaningful analysis about culture. Using innovative visualizations and analysis to bridge the inherent divide between an audience that is creatively and qualitatively oriented, and data that is empirical and needs cleaning/formatting.

8.4.1 Are there Citizenship Requirements?

No

8.4.2 Are there confidentiality requirements?

No

8.4.3 Number of Students:

group of four students.

8.4.4 Prequisite Skills beyond CSS Basics:

No"

9 Feasibility Analysis of Income Restricted Affordable Housing Projects for The Homelessness Crisis in San Diego

Website: https://homelessnesshub.ucsd.edu/

Host: Feiyang Sun f1sun@ucsd.edu Assistant Teaching Professor

9.1 Background:

According to the San Diego Regional Task Force on Homelessness, 15,327 people in San Diego experienced homelessness for the first time between 2021 and 2022. Numerous factors contribute to the current homelessness crisis and each individual's circumstances are unique. However, it is widely agreed that the lack of affordable housing in the San Diego region contributes to high rates of housing precarity and homelessness. Every eight years California jurisdictions must update the Housing Element in their general plan to accommodate their Regional Housing Needs Assessment (RHNA) allocation of housing units which includes market rate and income-restricted

affordable units. However, the jurisdictions are not required to build these units. As long as a local jurisdiction has demonstrated to the State of California's Office of Housing and Community Development (HDC) that its zoning map and regulations can feasibly accommodate its allocation, it is considered to be in compliance with the RHNA program's requirements.

9.2 Objectives:

Urban Land Institute (ULI) San Diego-Tijuana proposes to collaborate with the Homelessness Hub at UC San Diego to evaluate the feasibility of developing income-restricted affordable housing on these sites, particularly in terms of securing the necessary financing and development approvals. Currently, the project aims to answer the following questions: 1) Where will the affordable units be built?; 2) Who will develop the units?; 3) Is there adequate funding to build these units?; and 4) What are the barriers to securing the development approvals to build these units? The findings will be shared in a white paper and a local symposium.

9.3 Resources and Constraints:

The Homelessness Hub is a non-partisan hub for research, education, policy, and action on homelessness. It conducts unbiased and data-driven research in order to inform local policy and action aimed at reducing homelessness. With expertise in GIS and spatial analysis, Homelessness Hub has created a database of maps related to affordable housing and homelessness in the San Diego region. It recently created a detailed data layer that geospatially identifies all of the proposed sites for future affordable housing development as identified in the recently updated Housing Elements for all of the jurisdictions in San Diego County based on the most recent RHNA allocations.

9.4 Anticipated Challenges:

1) Work with both small and big data sets. The project will work with finance (pro forma) data of existing affordable development units. Since there're only a small number of projects in San Diego, the data will be limited in number of observations but high dimensional in terms of relevant variables. It requires some creativity and innovation to combine the rich pro forma data of existing projects with large spatial data of potential future development sites.
2) Understand real estate finance. The project requires learning of some basic concepts of real estate development and finance.

9.4.1 Are there Citizenship Requirements?

No citizenship issues.

9.4.2 Are there confidentiality requirements?

No confidentiality issues.

9.4.3 Number of Students:

group of two students.

9.4.4 Prequisite Skills beyond CSS Basics:

No other prerequisite skills required."

10 Global Design Index Research & Development: Measuring the Impact of Design

Website: https://designlab.ucsd.edu/

Host: Dr. Mai Nguyen mainguyen@ucsd.edu Director, Design Lab

10.1 Background:

The World Design Organization (WDO) in collaboration with past World Design Capitals (WDCs), Taipei University and global design thought leaders are developing a design index based on international data that embodies the implicit impact of design on a city's economy and SDGs. The data is expected to drive cities to increase design investments in their own policy making. Under the framework of this research model, through the collection and analysis of international indicators, WDO can analyze the components of different indexes and try to define the ""power of design"" dimension.

10.2 Objectives:

Collecting, assessing, and validating Regional design indicators to provide additional data sources and data sets to make the Design Index relevant and meaningful to our region (San Diego Tijuana).

10.3 Resources and Constraints:

Potential collaboration with consulting partner (i.e Deliotte), SD & TJ EDC, City of San Diego and City of Tijuana

10.4 Anticipated Challenges:

Complex system Consistency, relevance, scale, accuracy of regional data sets for comparison and usage Design definition and priority of design data sets that are regionally relevant Timezone Language barriers with Taipeii partner

10.4.1 Are there Citizenship Requirements?

No

10.4.2 Are there confidentiality requirements?

No

10.4.3 Number of Students:

group of two students., group of three students., group of four students.

10.4.4 Prequisite Skills beyond CSS Basics:

No"

11 Political Messaging and Trade Policy

Website: https://polisci.ucsd.edu/

Host: J. Lawrence Broz jlbroz@ucsd.edu Professor and Chair

11.1 Background:

How, when, why and to whom do politicians send messages, and to what extent are they persuasive? How does this vary across parties, administrations, branches of government and time? Can we find systematic evidence of "dog whistles"? With trust in democratic institutions being perilously low, it is imperative to understand the impact of political messaging, how to mitigate its negative consequences, and how to amplify its benefits. However, researchers know almost nothing about the policy content of government messaging, its impact on public attitudes, political behavior (e.g. voting, turnout), or stakeholder expectations, yet we know these messages are important. Our research agenda explores the content of political messages across all government sources available online – social media, congressional committees, press releases, laws and regulations – and across formats such as text, images and videos in order to analyze their impact.

11.2 Objectives:

Our focus in this project is on a single policy dimension – globalization. Our research seeks to shed light on the connection between tectonic shifts in the geopolitical landscape and domestic politics. Our objective is to understand the nature and content of public messages from politicians regarding globalization – tweets, press releases, congressional records, Facebook posts, etc. This project is aimed at developing an essential tool to categorize, analyze and standardize messages across various sources, and the outcome of the project is intended to be used for future academic research about political messaging.

11.3 Resources and Constraints:

To measure political messaging, we will use proprietary data from Voxgov (https://www.voxgov.com) on all government sources available online, including social media, congressional committees, press releases, laws and regulations. The Voxgov organization currently accesses about 50 million items across 10,000 government sources, though we focus on a subset of the data related to the World Trade Organization (WTO) which has about 70,000 items. The focus on the WTO reflects our current substantive interests, but the study of political messaging can be extended to any salient policy area, from education to global warming. Proposed product: We seek at least one student to help us build an algorithm housed within a graphical user interface (GUI). The tool would convert non-standardized media items (text, images, video) into a searchable database that can be used by social scientists in analytical studies. The database is ultimately intended to help contribute to cutting-edge research on politics and economics. For each item in the data, we would like to collect information including, but not limited to: (a) the main subject(s) of the document, (b) sentiment about each subject, (c) if the item references another item or event, (d) if that item is factual, (e) does the item contain news about the WTO and if so, which news event is referenced, (f) if the author is using persuasive language and if so, who they are attempting to persuade (e.g. other politicians, constituents, donors, foreign governments, foreign citizens).

11.4 Anticipated Challenges:

The items come in a variety of formats: raw text, PDF, video, and/or images. One challenge is to convert the items into a common format. In addition, sentiment must be minimally categorized as "positive'' or "negative" and along a continuous scale. Content(s) of each item must be verified against credible news and information sources; ideally, we would like a rating for how accurate the item is. Furthermore, these items contain specialized economics and political concepts which may be unfamiliar to people outside of these disciplines. Students will work with the principal investigators (PIs) to help guide them in subject areas in which they are not familiar in order to best structure their algorithm.

11.4.1 Are there Citizenship Requirements?

No citizenship issues.

11.4.2 Are there confidentiality requirements?

VoxGov data is proprietary and is not for public use. However, the main product of the project can be included on student portfolio websites using a subset of publicly available VoxGov data.

11.4.3 Number of Students:

group of three students.

11.4.4 Prequisite Skills beyond CSS Basics:

No."

12 CUPA: Contemporary Ukrainian Poetry Archive

Website: https://literature.ucsd.edu/people/faculty/aglaser.html

Host: Amelia Glaser amglaser@ucsd.edu Professor of Literature, Chair in Judaic Studies, Associate Director of the Institute for Arts and Humanities

12.1 Background:

As a scholar of literature, I have been fascinated by East European poets' use of social media. Over the past two years, I have worked with students to build an interactive online corpus of poems, posted by Ukrainian poets to Facebook. Together, we have been working to expand this archive and visualize the data that the archive gives us. Students working on this project should have an interest in digital humanities, computational methods for analyzing language (ie NLP), and an interest in web design. Students should also have an interest in poetry, and how readers respond to it. Fluency in a Slavic language is not required, though it is a plus.

The study of poetry has long been considered solely a qualitative pursuit. However, scholars working at the intersection of literature and the digital humanities have observed that quantitative analysis of poetic trends can help us to understand cultural, social, and formal phenomena. My current research involves following contemporary poets from Ukraine who have shared their work on social media (specifically Facebook) since the outbreak of the Donbas war in 2014. Together with my students, I have developed a database of poetry shared to Facebook, and we have tracked basic metrics around these poems, including "likes," "shares," and comments, as well as hand-curated data about formal genre and theme. Using the tools of natural language processing (NLP) with this archive of poetry, we have begun to examine what a dataset can tell us about how poetry changes, and how responses to it change, during political turmoil. We have also developed an interactive poetry archive, which is currently open to scholars by request.

I welcome applications from students with a background in programming, web-design, and natural language processing who are interested in helping to build this archive, and/or design analyses using this or a related corpus of literary/artistic work. Student collaborators need not be able to read Ukrainian or Russian, but, in addition to basic programming skills, they should have some interest in how poetry interacts with cultures of the world.

Students will be required to attend a weekly meeting to share their individual progress. Each student will develop a unique project that either helps to further develop the existing Contemporary Ukrainian Poetry Archive; uses the data collected to develop models that help to explain trends in poetry posted to social media; or develops a similar or related model for use with another community of writers (for example, a historical group of poets publishing in specific journals or another contemporary group of artists using social media).

12.2 Objectives:

We hope to further expand the database to include more poems, and to develop methods of automatically tagging poems by theme and keyword. We also hope to use multilingual BERT and other tools to probe what the corpus of recent poems can teach us about how poetry interacts with current events.

12.3 Resources and Constraints:

Students should come to the project with some knowledge of programming and data visualization. Any direct work with the archive must remain private until the group has prepared it for publication. Students developing their own related datasets may be able to publish these with proper citation and attribution.

12.4 Anticipated Challenges:

Students must be sensitive to confidentiality and privacy issues surrounding social media and the ongoing war in Russia/Ukraine.

12.4.1 Are there Citizenship Requirements?

no

12.4.2 Are there confidentiality requirements?

The data being collected is somewhat sensitive given the ongoing war in Ukraine. Therefore, we have kept the database private for the time being. Students involved in the project must exercise discretion in publishing findings related to contemporary Ukrainian and/or Russian poetry and poets.

12.4.3 Number of Students:

single student., group of two students., group of three students., group of four students.

12.4.4 Prequisite Skills beyond CSS Basics:

Students should have a solid command of English. In addition, they should be willing/able to read poetry (in any language)."