Data Science for Social Good
Information session for prospective student applicants

Anissa Tanweer
Research Scientist, eScience
Program Chair, DSSG
• Brief introduction to the eScience Institute
• Data Science for Social Good (DSSG)
  – Program overview
  – Application process
  – Previous projects
• Questions?
The eScience Institute empowers researchers and students in all fields to answer fundamental questions through the use of large, complex, and/or noisy data.
WRF Data Science Studio – a campus-wide collaboration space
We Disseminate Data Science Expertise & Best Practices

- Open Office Hours
- UW Data Science Seminar & Community Seminar
- Tutorials, bootcamps, workshops, and hack weeks
  - Astrohack, neurohack, geohack
  - Software carpentry (> 400 participants since we started counting in 2015)
- Winter Incubator
- Summer DSSG
Modeled after similar programs with elements from our own Data Science Incubator.

Through the Cascadia Urban Analytics Cooperative (CUAC) we worked with the University of British Columbia to set up their pilot DSSG program in 2017.
Goals

Figure out what it means to do “good” with data science

● Train students in data science methods
● Increase data science capacity across fields and organizations
● Positively impact society
Team composition

- DSSG Student Fellows (4-5)
- eScience Data Scientist Leads (1-2)
- Project Leads (1-2)
Examples of Project Lead Affiliations

University of Washington (academia)
- Washington State Transportation Center
- Disaster Data Science Lab
- Architecture Department

Seattle Department of Transportation (gov)
Bill & Melinda Gates Foundation (philanthropy)
Conservation International (nonprofit)
Bell Labs (industry lab)
What we expect from PL’s

• Co-presence 16 hrs/wk on average
• Domain expertise
• Stakeholder engagement
• Ability to discuss and promote work
• Open & reproducible when possible (Github)
• Description of project on our website
• Acknowledgment in publications
What we expect from students

• 40 hours/week ($7,000 stipend)
• Adherence to attendance policies
• Current student, grad and advanced undergrad
• Baseline programming and stats knowledge
• International students must be eligible to work in US (can’t support visas)
• Strong personal statement
• Team player
New developments aimed at increasing access to opportunity

• Micron Opportunity Scholarship
  – Optional application process
  – Supplemental award
  – For students facing financial barriers
  – https://escience.washington.edu/micron-opportunity-scholarship

• Trying to shift to credit-based course model with financial award (TBD) instead of stipend-based internship model
What you can expect from us

• Data scientists highly experienced in cross-disciplinary collaboration
• Expertise in (non-exhaustive):
  - Machine learning  Statistical inference
  - Databases   GIS
  - Modeling   Optimization
  - Visualization   Cloud computing
• Best practices in version control, reproducibility and human-centered design
• Data science curriculum
Tutorials

- Intro to Git & GitHub
- Git and Git Workflow
- Team Management Processes
- Pandas, Geopandas, and SQL
- Python Coding Standards and Documentation
- Unit Tests
- Project Organization, Virtualization, Continuous Integration
- Pair Programming

- Object Oriented Python
- Software Design
- Machine Learning
- Web Design and Web Apps
- Cython/Dask/High Performance Python
- Vega/Altair
- Data Visualization with Tableau
- An Introduction to Visual Communication
Workshops

- Introduction to Data Science for Social Good
- Team Development Part I
- Team Development Part II
- Preparing for Stakeholder Engagement
- Stakeholder Analysis & Speculative Ethics
- Ethical Agency in Data-Rich Organizations
- Best Practices in Public Speaking
- Final Presentation Practice Talks
Other Activities

- Docathons
- Project Spotlights
- Career Conversation Luncheons
- Field research
Ethnography & Human-Centered Data Science

- study the culture & practice of data science
- provide programmatic insight
- stakeholder collaboration
- data science ethics
- human-centered design
Call for Applications is NOW OPEN!

https://escience.washington.edu/get-involved/incubator-programs/data-science-for-social-good/

https://escience.washington.edu/dssg-student-faq/

We encourage you to reach out with questions!

Emily Keller
efkeller@uw.edu
What we’re looking for in DSSG fellows

In individuals:

• Baseline programming and research methods training
• Motivation for wanting to participate
• Strong teamwork
• Experience with research and “social good”
• Commitment to diversity

Across the cohort:

• Range of disciplinary backgrounds and expertise
• Range of technical abilities
• Range of educational experience levels
Selection Process

• Initial screening of applications
• Committee review of threshold candidates
• Interviews with top candidates (~20% of total applicant pool) - Late March
• Solicitation of info from short-listed candidates - Early April
• Notification of admission offer or waitlist - Mid April
Important Dates

Jan. 6 - Student applications opened
Jan. 23 - Student Info Session
Feb. 17 - Student Fellow Applications due
Mid March - Notification of interview invitation
Late March - Interviews held
Early April - Notification of shortlisting
Mid April - Notification of admission offer or waitlist status
June 15th - First day of program
August 21st - Last day of program
A non-exhaustive list of topical interests

• Poverty, equity, income
• Housing and homelessness
• Public education
• City planning
• Transportation
• Hazards/Resilience
• Utilities
• Economic impacts
• Environmental issues
We have a broad view of what counts as data science
Mining Online Data for Early Identification of Unsafe Food Products

This drink is the worst!
These were bad and made me sick. Not great. The worst!

My stomach hurt so bad after drinking this and I got sick! So bad...worst time of my life thus far.

This drink is a great drink!
Use of ORCA data for improved transit system planning and operation

ORCA Transactions Data

Automatic Vehicle Location Data

[Map showing average weekday boardings per stop]
Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks
Summer 2019 Projects

ADUniverse: Evaluating the Feasibility of (Affordable) Accessory Dwelling Units in Seattle

**Project leads:** Rick Mohler, Associate Professor, Department of Architecture, University of Washington; and Nick Welch, Senior Planner, City of Seattle Office of Planning and Community Development

**Data science lead:** Joseph Hellerstein

Developing an Algorithmic Equity Toolkit with Government, Advocates, and Community Partners

**Project lead:** Mike Katell, PhD Candidate, UW Information School

**Data science lead:** Bernease Herman

Understanding Congestion Pricing, Travel Behavior, and Price Sensitivity

**Project lead:** Mark Hallenbeck, Director, Washington State Transportation Center, University of Washington

**Data science lead:** Vaughn Iverson

Natural Language Processing for Peer Support in Online Mental Health Communities

**Project leads:** Tim Althoff, Assistant Professor, Computer Science & Engineering, University of Washington; and Dave Atkins, Research Professor, Psychiatry and Behavioral Sciences, University of Washington

**Data science lead:** Valentina Staneva
Questions?

Application process, eligibility, etc:
Emily Keller (efkeller@uw.edu), Program Coordinator

Program content, etc:
Anissa Tanweer (tanweer@gmail.com), Program Chair

https://escience.washington.edu/get-involved/incubator-programs/data-science-for-social-good/
Summer 2017 DSSG

- Improving transit services using ORCA data – Washington State Transportation Center
- Strengthening capacities, knowledge and data sharing platforms for sustainable development – Vital Signs
- Can traffic sensor data detect vehicle cruising? - Seattle Department of Transportation
- The 'Equity Modeler': examining just development in Seattle - Department of Urban Design and Planning and Department of Architecture
Can traffic sensor data detect vehicle cruising?
- w/ the Seattle Department of Transportation

Anamol Pundle
## DATA: SENSOR GRID

### Sensor Data Table

<table>
<thead>
<tr>
<th>HASHED MAC</th>
<th>TIME</th>
<th>SENSOR</th>
<th>STRENGTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD98SDK8AH</td>
<td>8:32:01</td>
<td>276105</td>
<td>-52</td>
</tr>
<tr>
<td>8DJSKDLSX0</td>
<td>8:32:01</td>
<td>276102</td>
<td>-55</td>
</tr>
<tr>
<td>439WOA09A</td>
<td>8:32:01</td>
<td>265402</td>
<td>-75</td>
</tr>
<tr>
<td>777AJDKAL8</td>
<td>8:32:05</td>
<td>293010</td>
<td>-50</td>
</tr>
<tr>
<td>QKSJ239A99</td>
<td>8:32:07</td>
<td>251040</td>
<td>-45</td>
</tr>
<tr>
<td>DQWPOOA09</td>
<td>8:32:10</td>
<td>265402</td>
<td>-49</td>
</tr>
<tr>
<td>KD98SDK8AH</td>
<td>8:32:11</td>
<td>265302</td>
<td>-54</td>
</tr>
</tbody>
</table>
DISTANCE RATIO :: SINUOCITY

“Labeling” for Classification

5 / 5 = 1.0
PROBABLY NOT CRUISING

14 / 2 = 7.0
PROBABLY CRUISING
CRUISING IN DOWNTOWN SEATTLE

As part of The Data Science for Social Good Program at The University of Washington, the Traffic Cruising Team has produced a heatmap to identify cruising in the downtown Seattle area.

SELECT TYPE OF CRUISING

PARKING

SELECT DAY

MON TUE WED THR FRI SAT SUN

SELECT TIME

ALL DAY MORNING MIDDAY EVENING

NUMBER OF PATHS

0 - 59
59 - 186
186 - 296
296 - 458
458 - 567
567 - 681
681 - 855
855 - 1055
1055 - 1345
1345+

NUMBER OF PARKING PATHS

0 - 59
59 - 186
186 - 296
296 - 458
458 - 567
567 - 681
681 - 855
855 - 1055
1055 - 1345
1345+

TIME

0AM 12AM 01AM 02AM 03AM 04AM 05AM 06AM 07AM 08AM 09AM 10AM 11AM 12PM 01PM 02PM 03PM 04PM 05PM 06PM 07PM 08PM 09PM 10PM
Projects - Years 1 & 2

2015
• Open Sidewalk Graph for Accessible Trip Planning
• Assessing Community Well-being through Open Data and Social Media
• Predictors of Permanent Housing for Homeless Families
• Rerouting Solutions and Expensive Ride Analysis for King County Paratransit

2016
• Mining Online Data for Early Identification of Unsafe Food Products
• Use of ORCA data for improved transit system planning and operation
• Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks
• CrowdSensing Census: A heterogenous-based tool for estimating poverty
54.5 million

People in the USA need assistive devices or have trouble walking more than a quarter mile.

U.S. Census Bureau, *Americans With Disabilities: 2010, issued July 2012*
Automated cleaning of sidewalk data through computational geometry

powered by data from: SDOT/Socrata Google API

<table>
<thead>
<tr>
<th>Step</th>
<th>Running Time</th>
<th>Solved (All)</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connecting T-Gaps</td>
<td>~3.9s</td>
<td>3,837 (4,352)</td>
<td>88.2</td>
</tr>
<tr>
<td>Intersection Cleaning</td>
<td>~23.6s</td>
<td>38,844 (44,700)</td>
<td>86.9</td>
</tr>
<tr>
<td>Polygon Cleaning</td>
<td>~10min</td>
<td>7,283 (8,035)</td>
<td>90.6</td>
</tr>
<tr>
<td>Connecting Subgraphs</td>
<td>~23.2s</td>
<td>39,913 (45,265)</td>
<td>88.1</td>
</tr>
</tbody>
</table>
OpenStreetMap (OSM)

Simplifying the user process

Current practice

Our Proposal
Sign up for our mailing list at escience.washington.edu

Follow us on social media…

  twitter   @uwescience
  facebook  uwescienceinstitute
Novel Analyses of Homeless Family Trajectories through Programs

When homeless families engage in services and programs, what factors are most likely to lead to a successful exit?

The DSSG team

• developed algorithms to identify ‘families’ and to identify ‘episodes’ of homelessness including back-to-back, or overlapping enrollments in individual programs

• devised innovative ways to visualize and analyze the ways families transition between programs
The DSSG team created interactive visualizations to facilitate exploration of the data by the stakeholders. This diagram shows the proportional flow from one program to another, as well as the eventual outcome.