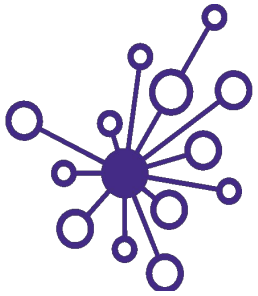


UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Data Science for Social Good

Information session for prospective project lead applicants

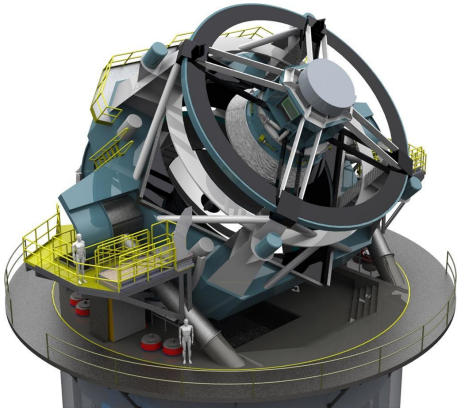
Sarah Stone
Executive Director, eScience & Program Director, DSSG
Anissa Tanweer
Research Scientist, eScience & Program Chair, DSSG

ALFRED P. SLOAN FOUNDATION

This morning

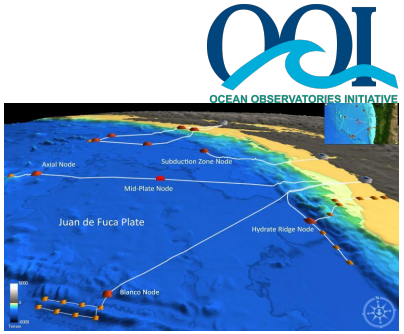
- Introduction to the eScience Institute
- Data Science for Social Good (DSSG)
  - Program overview
  - Proposal process
  - Program logistics
  - Previous projects
- Questions?

Nearly every field of discovery is transitioning from "data poor" to "data rich"

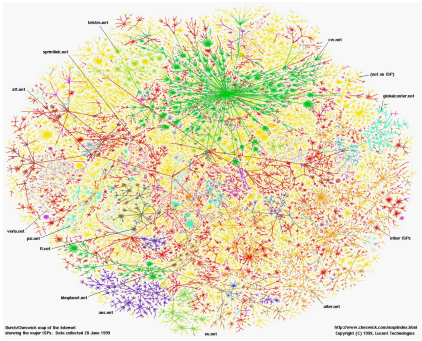


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

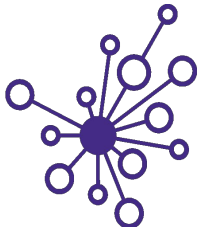
Biology: Sequencing

Economics: POS terminals

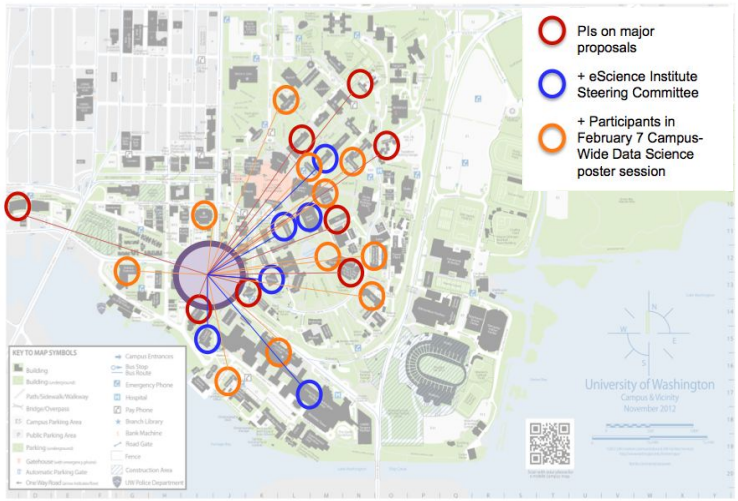
Neuroscience: EEG, fMRI

Our Mission

The eScience Institute **empowers** researchers and students in all fields to answer fundamental questions through the use of large, complex, and/or noisy data.



UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

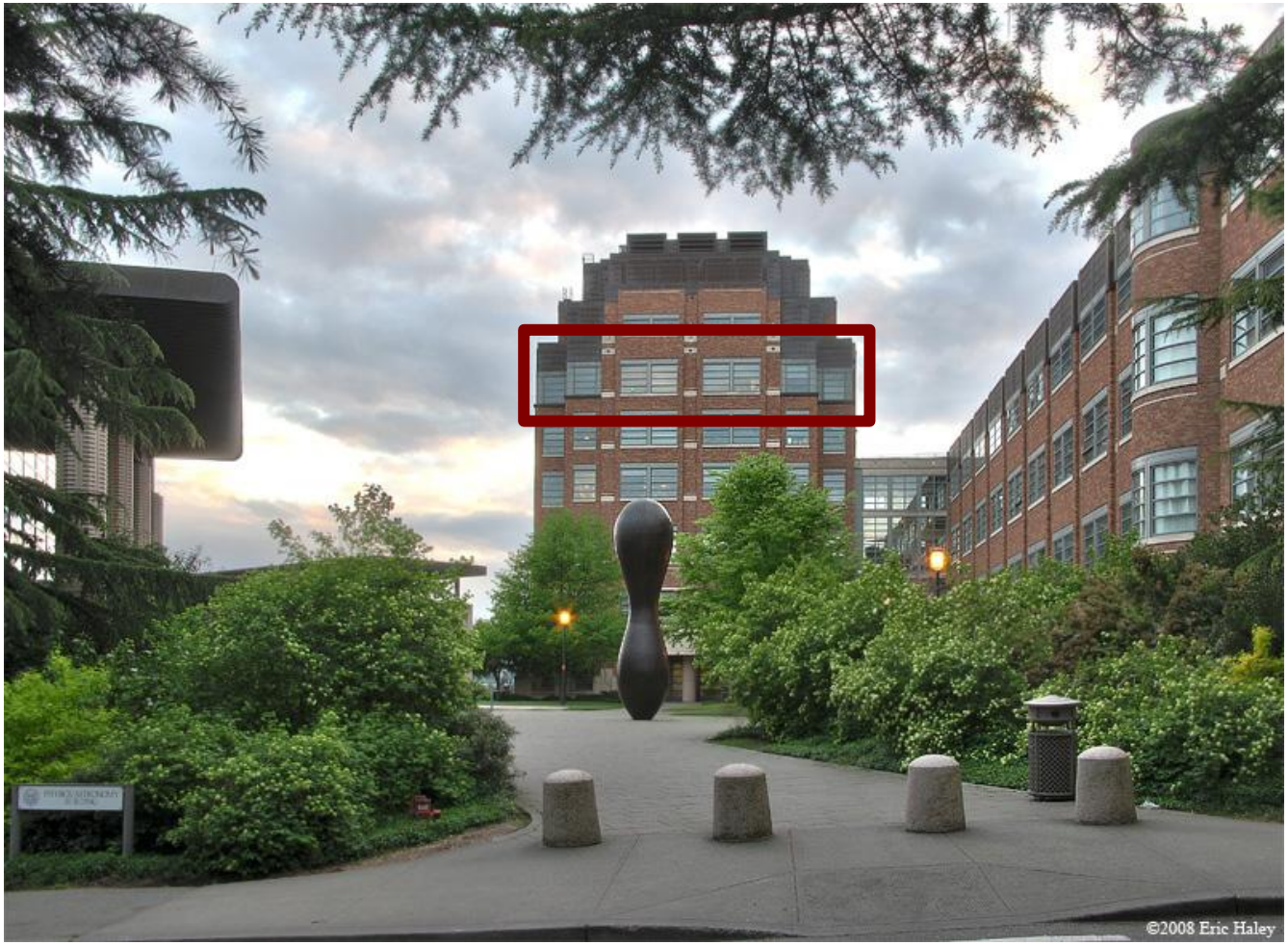


UNIVERSITY of WASHINGTON

eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

WRF Data Science Studio – a campus-wide collaboration space



UNIVERSITY of WASHINGTON

Who We Are - eScience Research Team

Director of Research

David Beck
Ph.D. Medicinal Chemistry,
Biomolecular Struct. & Design

Data Scientists

Bernease Herman
B.S. Statistics
Formerly SE at Amazon

Ariel Rokem
Ph.D. Neuroscience

Valentina Staneva
Ph.D. Applied Mathematics and Statistics

Jose Hernandez
Ph.D. Measurement & Statistics

Amanda Tan
Ph.D. Civil & Env. Engineering

Anissa Tanweer
Ph.D. Communication

Research Scientists

Anthony Arendt
Ph.D. Geophysics
APL

Bryna Hazelton
Ph.D. Astrophysics
Physics

Joe Hellerstein
Ph.D. Computer Science
IBM Research, Microsoft Research, Google (ret.)

Vaughn Iverson
Ph.D. Oceanography

Nicoleta Crisea
Ph.D. Environmental Engineering

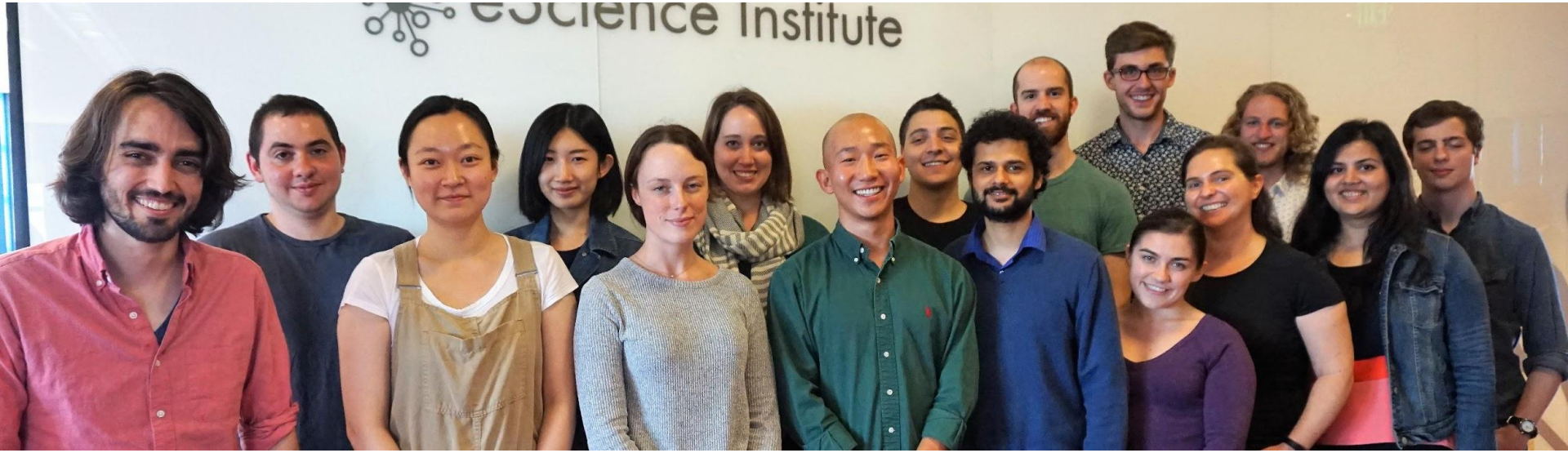
Spencer Wood
Ph.D. Zoology

Scott Henderson
Ph.D. Geological Sciences

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

UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD



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)

What PL's get

Intensive work on project

Exposure to new methods and approaches

Interdisciplinary teamwork

Networking opportunities

Publicity

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

Scoping meetings in preparation

Co-presence 16 hrs/wk on average

* Probably more during first 2 weeks

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)

Current student, grad and advanced undergrad

Baseline programming and stats knowledge

Eligible to work in US (can't support visas)

Strong personal statement

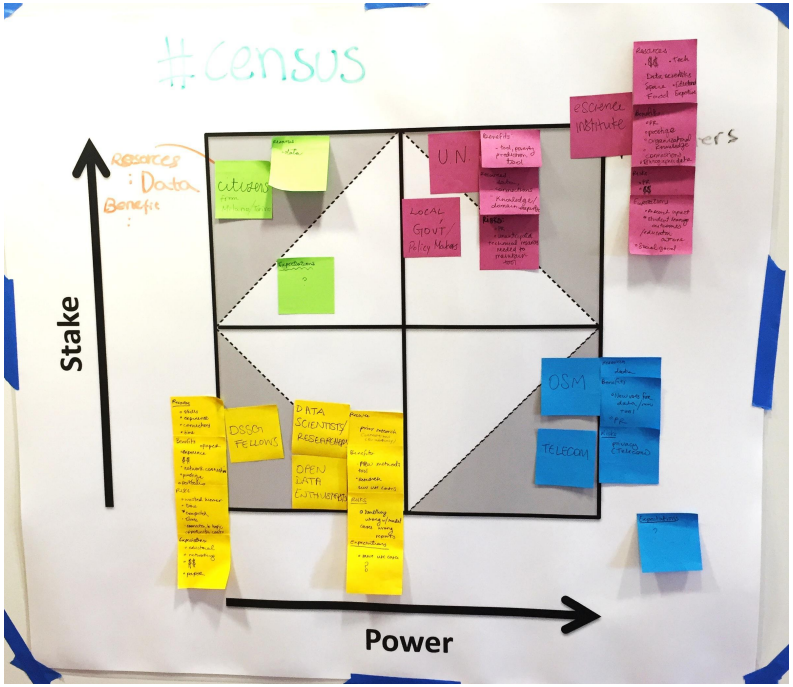
Team player

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
- Help with team management



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 Proposals is NOW OPEN!

https://escience.washington.edu/dssg-proposal

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

We encourage you to reach out and
meet with us before submitting a proposal

Office Hours:
https://escience.washington.edu/office-hours/#eScienceDataScientists

What we're looking for

- argument in support of how project will lead to positive social impact
- strong research, strong methods
- availability, commitment
- clarity and shovel-readiness
- capacity for measurable outcomes
- sustained engagement

What we *DON'T* do:

- build web portals
- app development as primary goal
- data collection

A non-exhaustive list of topical interests

- Poverty, equity, income
- Housing
- Public Education
- City planning
- Transportation
- Hazards/Resilience
- Utilities
- Economics
- Environmental issues

Technical Areas of eScience Expertise

- new platforms, new algorithms, new methods, new datasets
- working with large, heterogeneous, and noisy datasets
- scalable analytics and predictive models
- interactive visualization
- code review, publishing, and reproducibility
- online teaching materials, tutorials

Pre-Program

- 2-3 meetings with data scientists
- Project Lead orientation

First Two Weeks

- Mandatory team development workshops (may require more than 16 hrs total)
- Front-loaded tutorials

Rest of Summer

- Weekly “project spotlight” meetings
- Regularly scheduled team check-ins
- Bi-weekly check-ins with all PL’s, DS’s and administrators
- Occasional tutorials (can be on-demand)
- Visits and calls with stakeholders

End of Summer

Final presentations and reception

Important Dates

Now - Call for Proposals open

Jan. 6 - Student applications opened

Jan. 17 - Info Session

Feb. 24 - Project proposals due ***

Mar. 2 - Project short-list notifications ***

Apr. 8 - Student selection completed

Mar. - Jun. - Meetings with DS & PL

Jun. 15 - First day DSSG ***

Aug. 21 - Last day DSSG

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

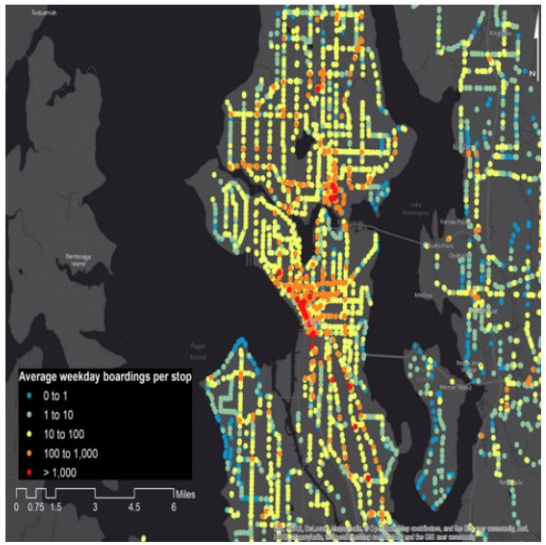
We have a broad view of what counts as data science

Mining Online Data for Early Identification of Unsafe Food Products

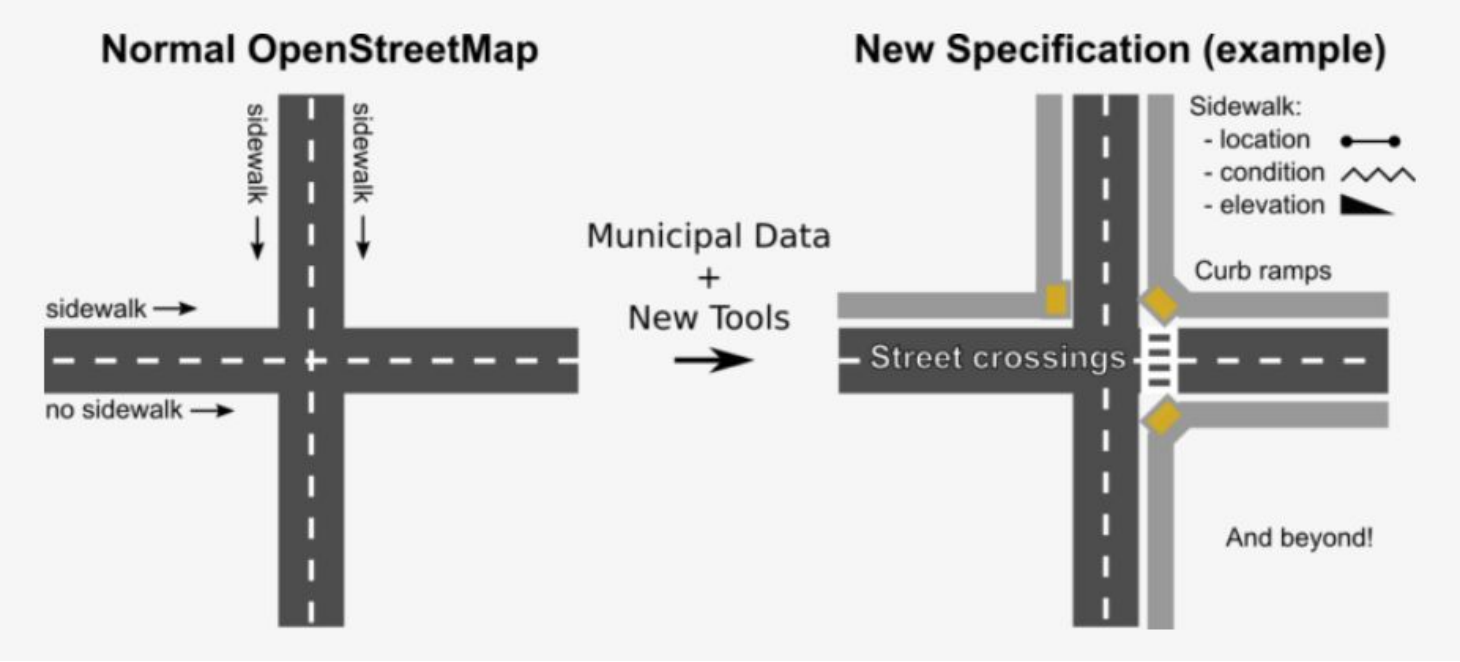
Use of ORCA data for improved transit system planning and operation

ORCA Transactions Data

Automatic Vehicle Location Data



Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks



MDG

Questions?

Contact Anissa Tanweer

tanweer@gmail.com

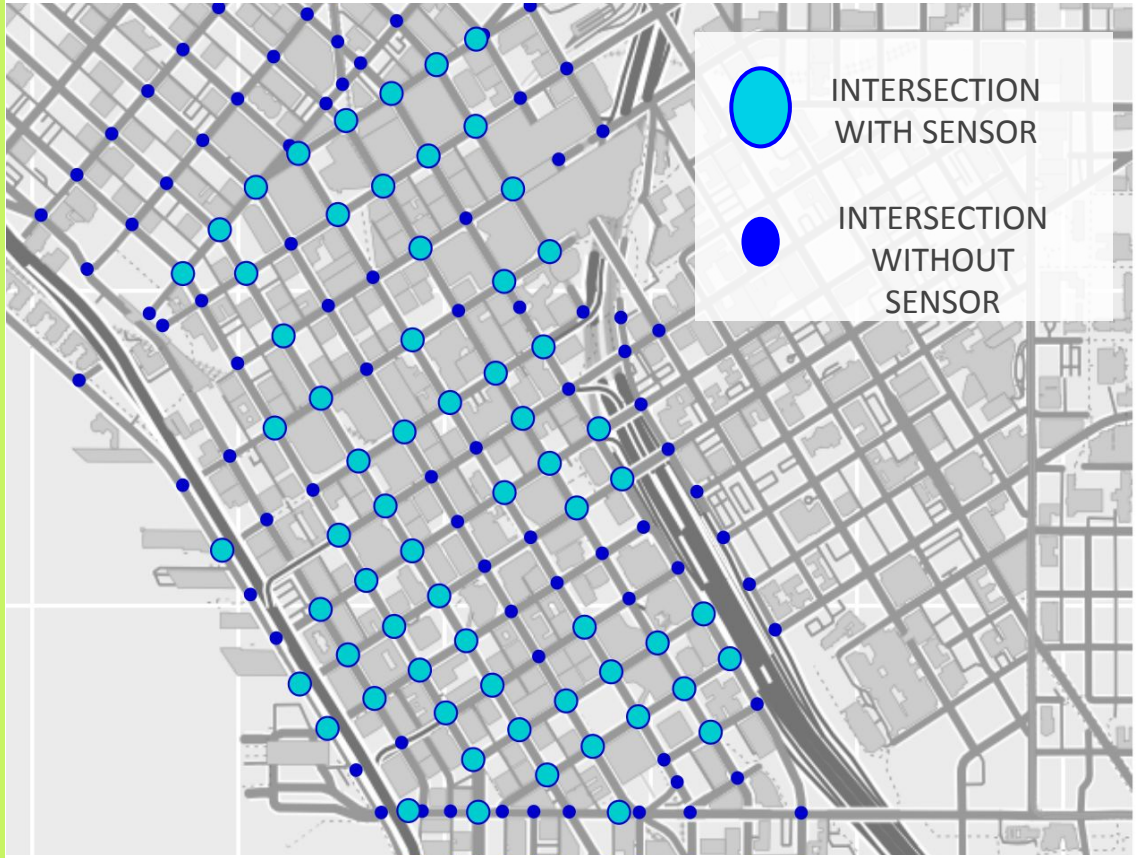
https://escience.washington.edu/dssg-proposal

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

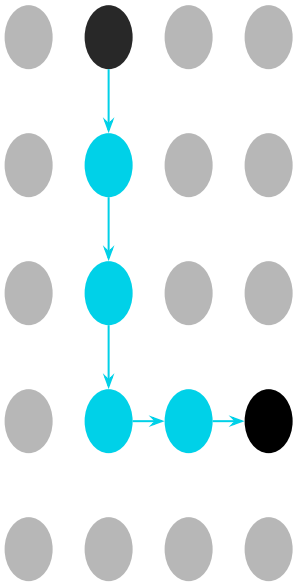


HASHED MAC	TIME	SENSOR	STRENGTH
KD98SDK8AH	8:32:01	276105	-52
8DJSKDLSX0	8:32:01	276102	-55
439WOA09A	8:32:01	265402	-75
777AJDKAL8	8:32:05	293010	-50
QKSJ239A99	8:32:07	251040	-45
DQWPPOA09	8:32:10	265402	-49
KD98SDK8AH	8:32:11	265302	-54

DATA: SENSOR GRID

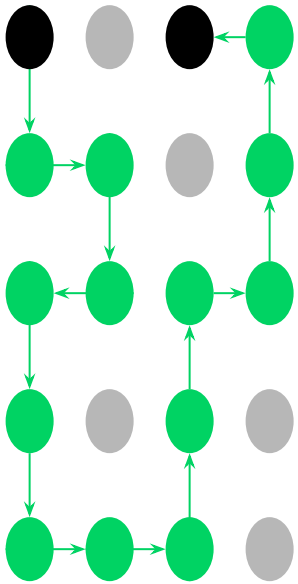
DISTANCE RATIO :: SINUOCITY

"Labeling" for Classification



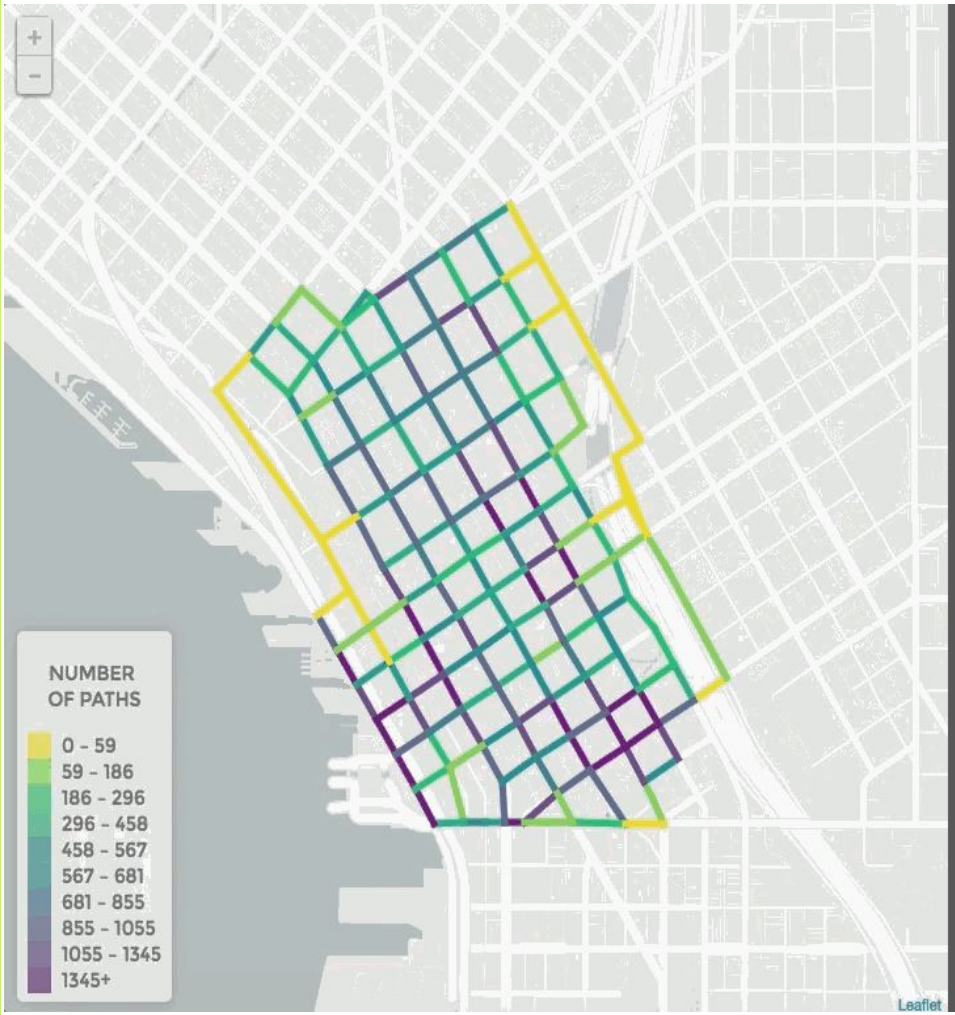
5 / 5 = 1.0

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

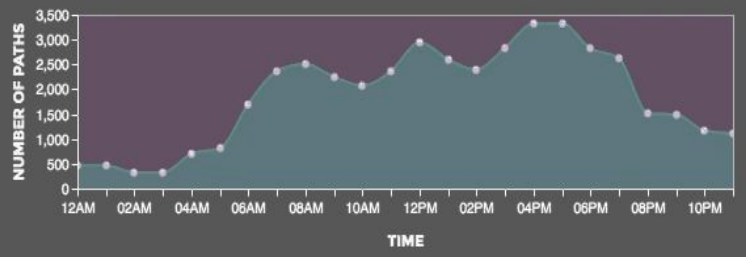
PARKING | PASSING VEHICLES

SELECT DAY

MON | TUE | WED | THR | FRI | SAT | SUN

SELECT TIME

ALL DAY | MORNING | MIDDAY | EVENING



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

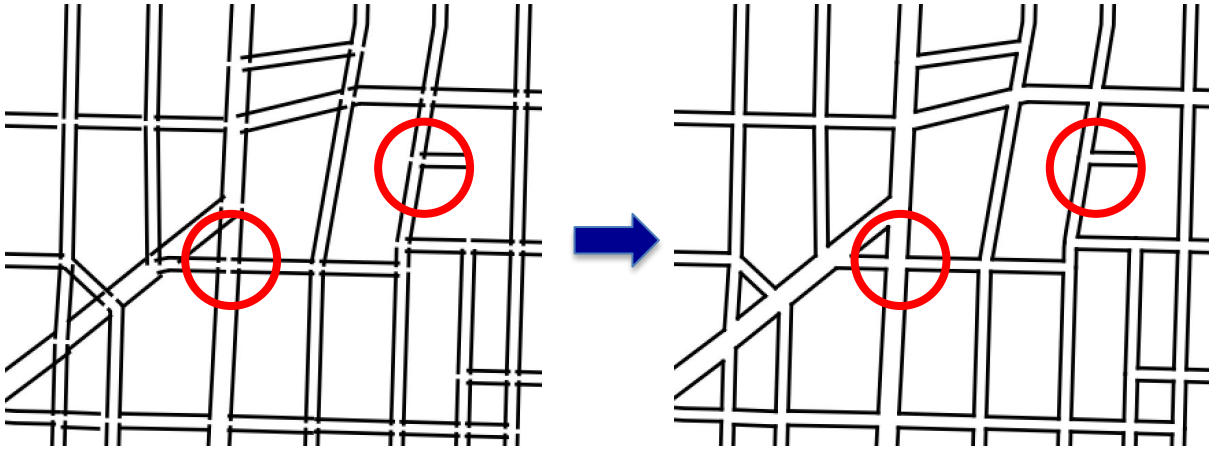
- Mining Online Data for Early Identification of Unsafe Food Products
- Use of ORCA data for improved transit system planning and operation
- Global Open Sidewalks: Creating a shared open data layer and an OpenStreetMap data standard for sidewalks
- CrowdSensing Census: A heterogenous-based tool for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

54.5 million

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

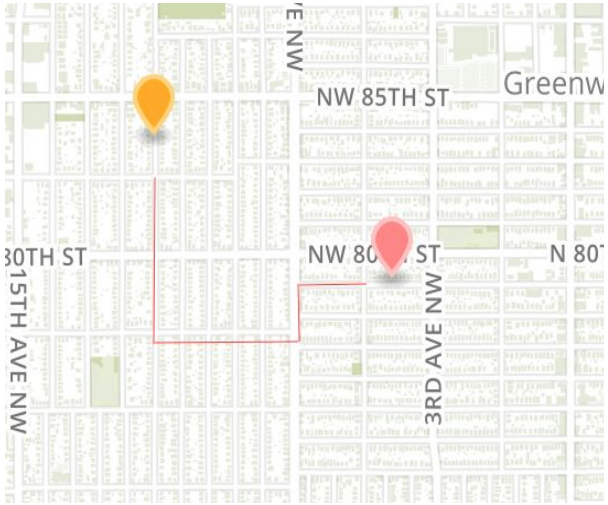
U.S. Census Bureau, *Americans With Disabilities: 2010, issued July 2012*

Automated cleaning of sidewalk data through computational geometry



powered by data from:
SDOT/Socrata
Google API

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



OpenStreetMap (OSM)

Simplifying the user process

Current practice

Our Proposal

The Seattle Times

Education | Education Lab | Local News | Transportation

UW student project taps ORCA cards, unlocks data trove

Originally published August 19, 2016 at 10:21 pm | *Updated August 21, 2016 at 6:37 pm*

GeekWire NEWS JOBS EVENTS RESOURCES DEALS ABOUT Search

STRATOS
MEDICAL DEVICE DEVELOPMENT IS RISKY. LET STRATOS' ENGINEERS CLEAR THE PATH FOR PRODUCT DEVELOPMENT.
learn more >>

Trending: Hewlett Packard Enterprise cuts staff as it struggles to keep pace in the cloud

Newsletter signup Space & Science

Could Amazon reviews keep you from getting sick? Researchers analyze text to predict food recalls

BY CLARE MCGRANE on August 28, 2016 at 11:16 am

Xconomy Xperience Tech + Life EXOME Biotech + Health Our Regions Tech Channels Meet the Xconomists Our Events

ADVERTISEMENT

Seattle Home Seattle Events Local Jobs Archives Xconomists VC / M&A Deals

Budding UW Data Scientists Use Their Powers for Social Good

Benjamin Romano
August 24th, 2015
@bromano @xconomy Like Us

Earn a degree in the field of data science these days and your ticket is punched: Google, Amazon, Facebook, leading-edge academic research, a well-funded startup

GeekWire NEWS JOBS EVENTS RESOURCES DEALS ABOUT

Trending: Microsoft reveals the 'Xbox Onesie' and the internet goes nuts

Could data help solve Seattle's transportation challenges?

BY CLARE MCGRANE on August 20, 2016 at 3:30 pm

Student projects leapfrog governments and industry in 'Data Science for Social Good' program

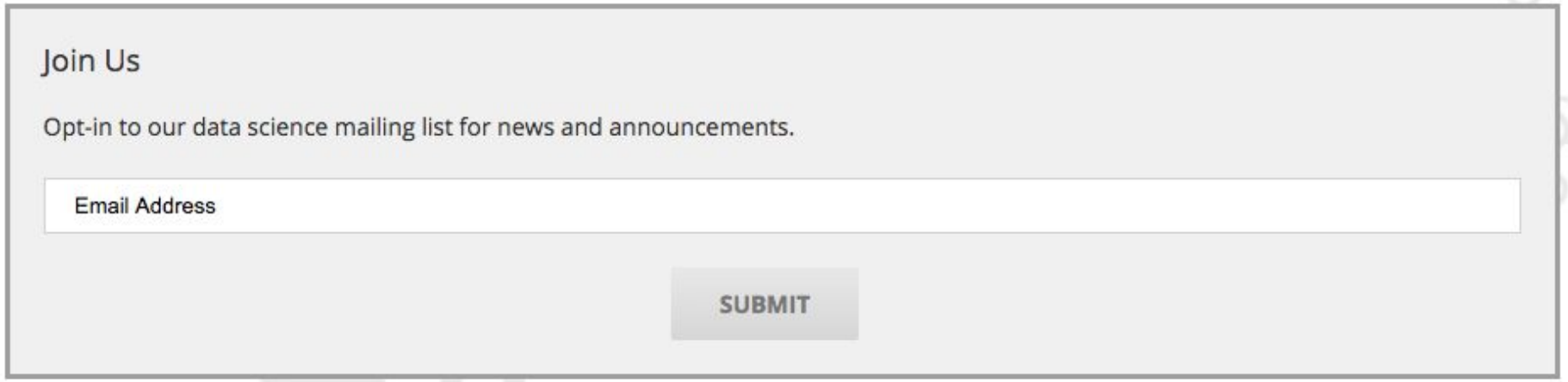
Posted Aug 26, 2016 by Devin Coldewey, Contributor

Next Story

ADVERTISEMENT
Enter your forecasts for a chance to win prizes totaling $1.2 million.

UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

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



Follow us on social media…

twitter *@uwescience*

facebook *uwescienceinstitute*

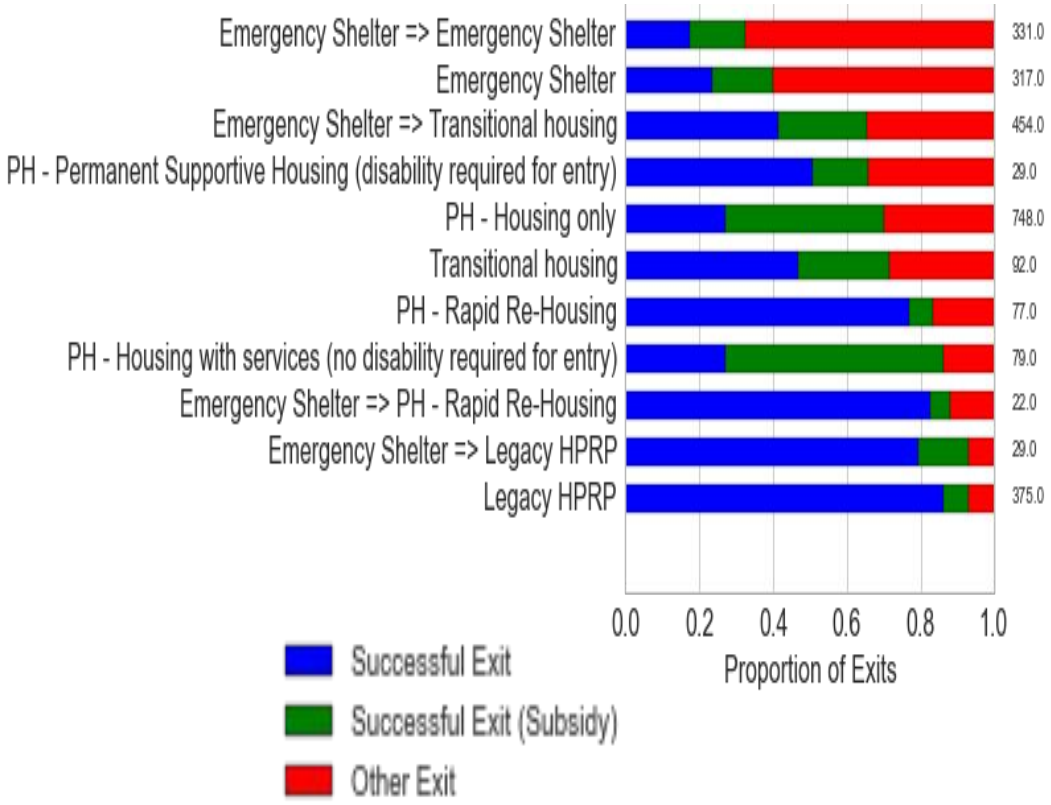
Novel Analyses of Homeless Family Trajectories through Programs

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

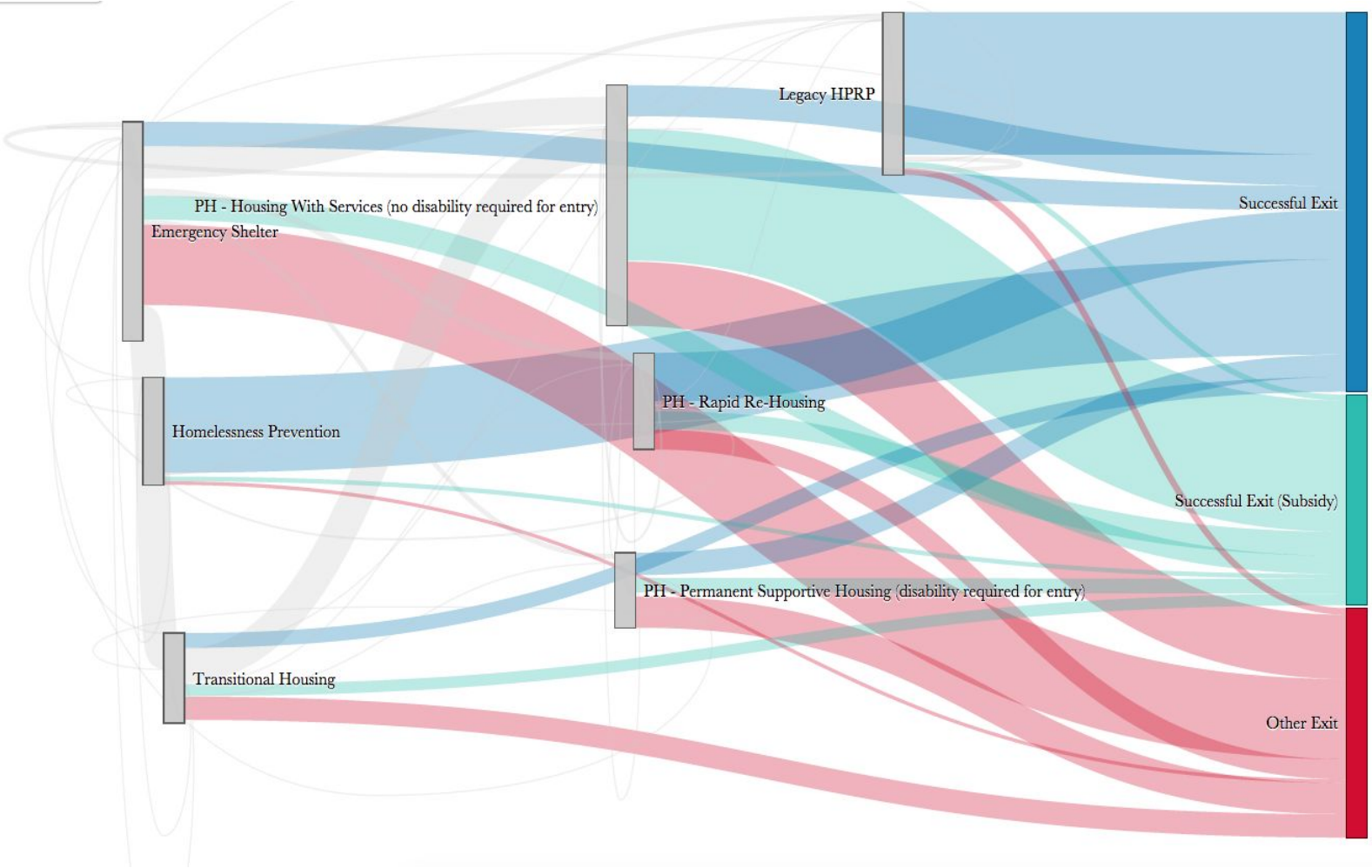
The DSSG team

- 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



Novel Analyses of Family Trajectories through Programs – Sankey Diagram



Using D3, technology developed in Jeff Heer's group

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.