Academic
Data Science
Alliance

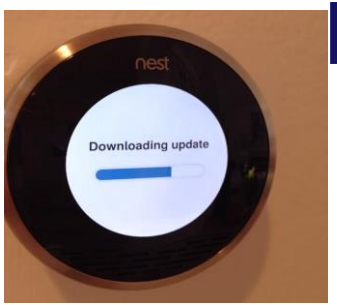
Academic Data Science: From individuals to institutions

Micaela Parker, Executive Director

Academic Data Science Alliance

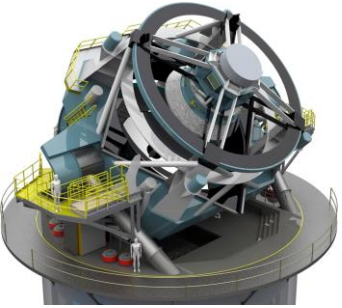
February 2020 eResearch New Zealand

Data are being collected and used h



- Smart homes
- Smart cars
- Smart health
- Smart interaction (virtual reality)
- Smart cities
- Smart discovery **

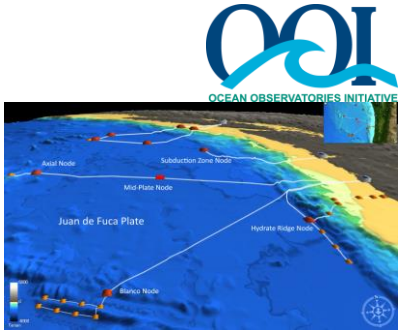
Nearly every field of discovery is transitioning from “data poor” to “data rich”



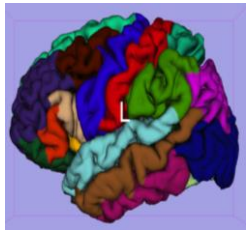
Astronomy: LSST

Physics: LHC

Digital Humanities



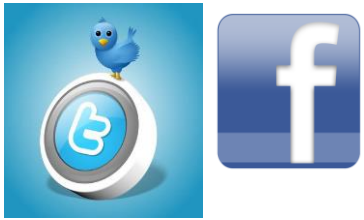
Oceanography: OOI



Health

Biology: Sequencing

Economics: POS terminals



Sociology: Social Media and the Web

University
Domain
Research

Data
Science
Practice

as **data increases in all forms and in all fields**, even some of the very best researchers struggle to generate knowledge and insight from these data

Example from my own wor

doi:10.1038/nature07539 nature

LETTERS

2008

Ferritin is used for iron storage in bloom-forming marine pennate diatoms

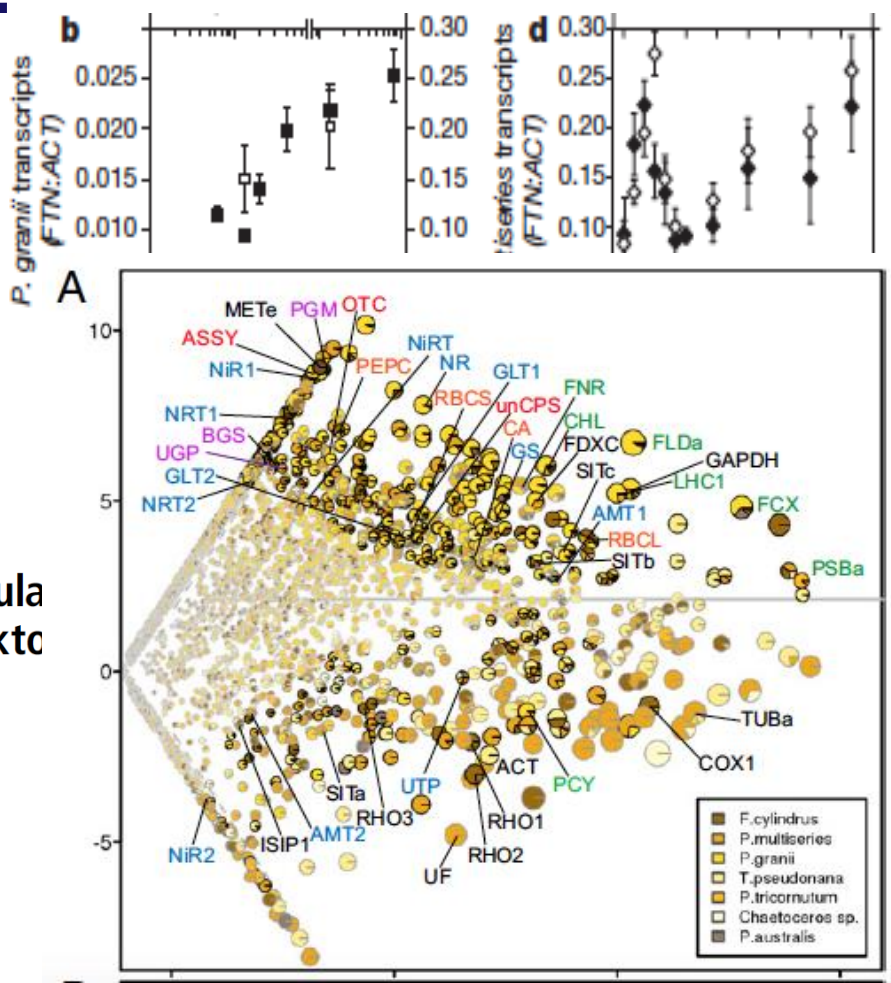
Adrian Marchetti[1]*, Micaela S. Parker[1]*, Lauren P. Moccia[2], Ellen O. Lin[1], Angele L. Arrieta[3], Francois Ribalet[1], Michael E. P. Murphy[3], Maria T. Maldonado[2] & E. Virginia Armbrust[1]

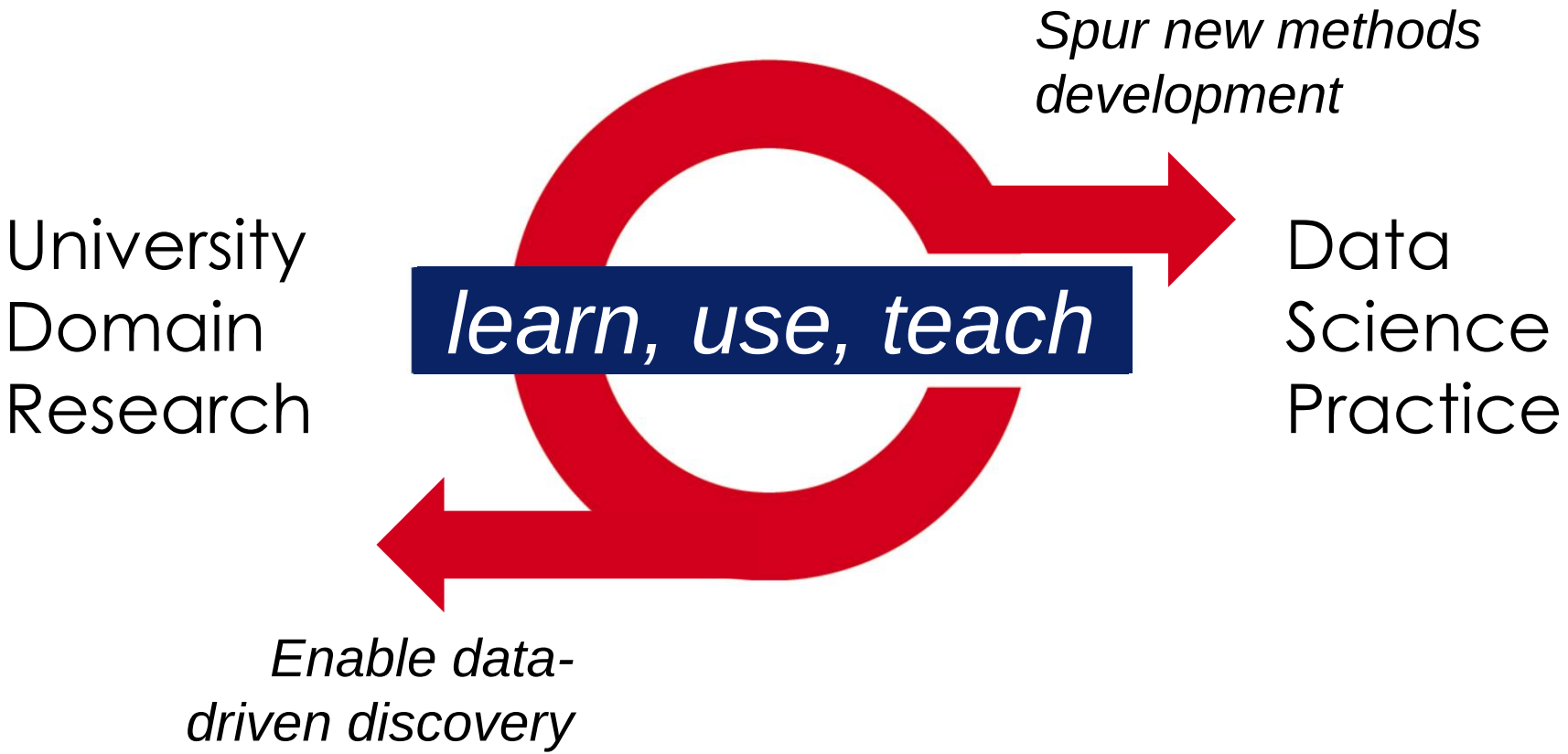
2012

Comparative metatranscriptomics identifies molecula bases for the physiological responses of phytoplankto to varying iron availability

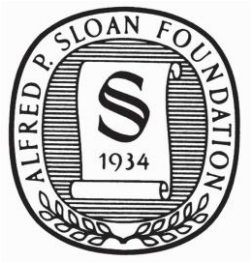
Adrian Marchetti[a,1,2,3], David M. Schruth[a,1], Colleen A. Durkin[a], Micaela S. Parker[a], Robin B. Kodner[a], Chris T. Berthiaume[a], Rhonda Morales[a], Andrew E. Allen[b], and E. Virginia Armbrust[a,2]

[a]School of Oceanography, University of Washington, Seattle, WA 98105; and [b]J. Craig Venter Institute, San Diego, CA 92121




Spur new methods development
University Domain Research
learn, use, teach
Data Science Practice
Enable data-driven discovery

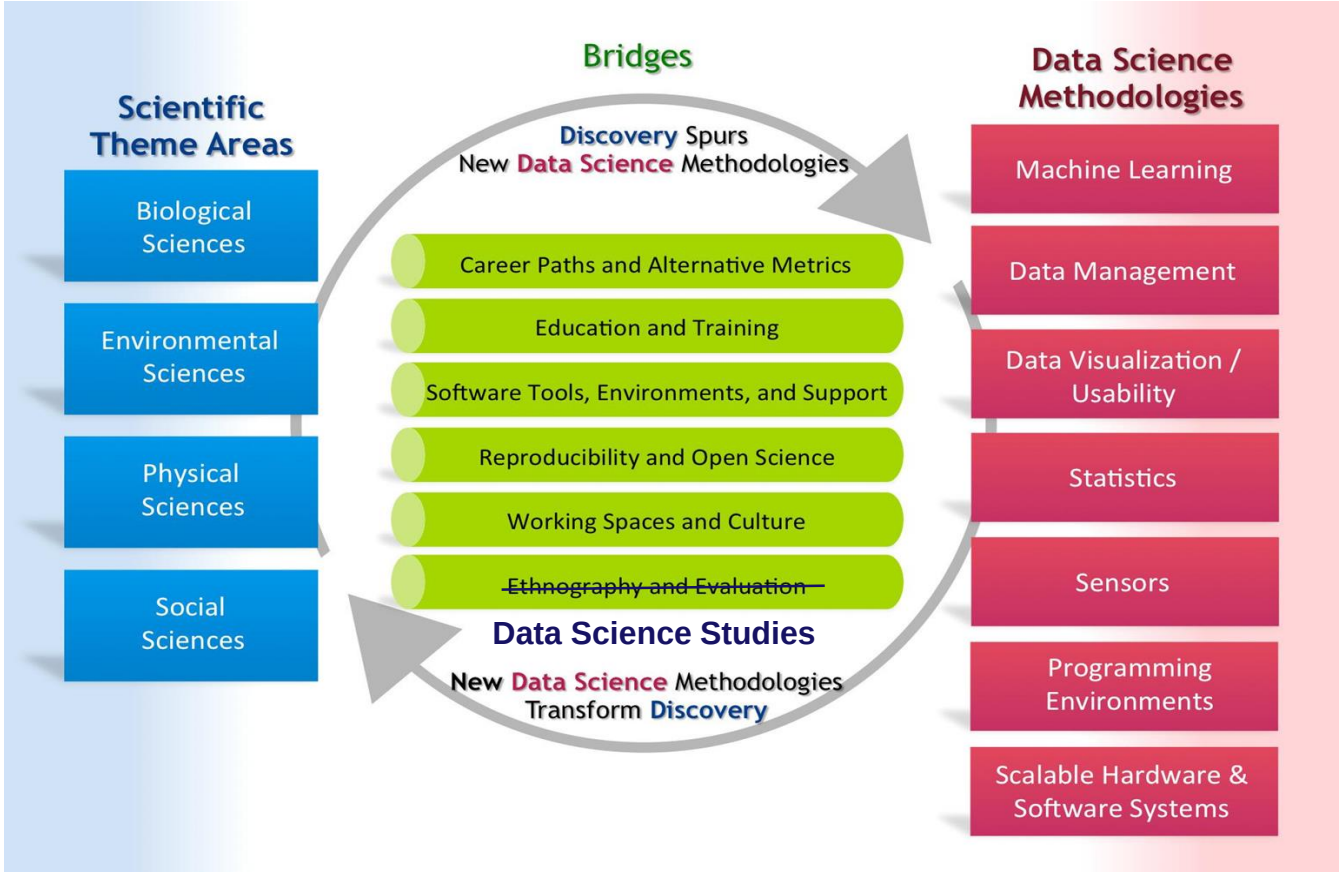
Moore-Sloan Data Science Environments



Micaela Parker
eScience Executive Director -> MSDSE Program Coordinator

Chris Mentzel, *Gordon and Betty Moore Foundation*
Josh Greenberg, *Alfred P. Sloan Foundation*

Building Bridges: Our Efforts Organized into Working Groups



Data Science Studies

to understand the complex landscape within which data science is situated, and identify and evaluate best practices...the data science of data science

- Reflective and reflexive self-evaluation

Provide immediate feedback of programs and activities = responsiveness and adaptable nature of the MSDSE's.

Raise awareness of ethical issues and surface best practices to the larger community.

- Scholarly work

Using computational, HCI, historical and ethnographic approaches to studying the practices, tools, and culture of data science

Reproducible and Open Science

- Hired first reproducibility librarian in a tenure-track position! (2018)

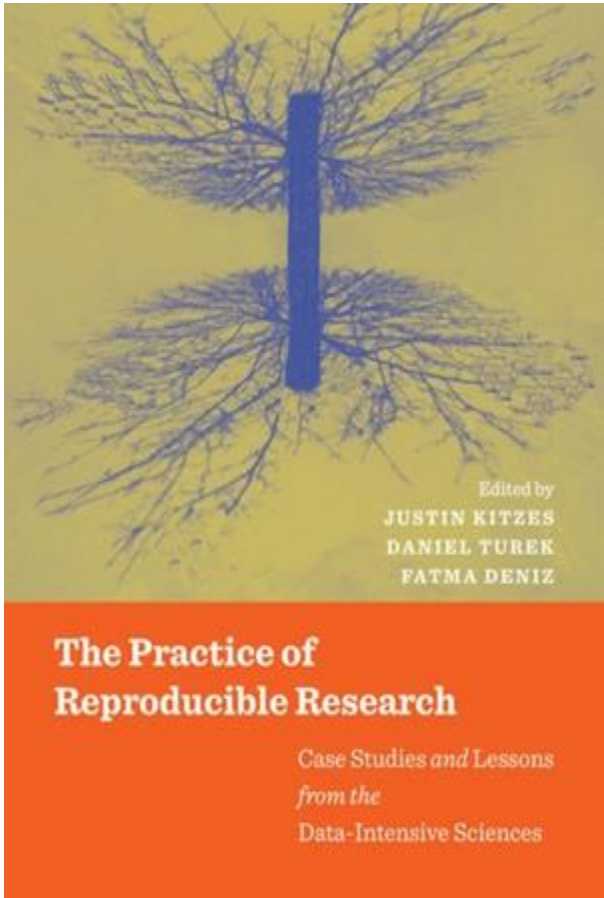
NYU Center for Data Science

- ReproZip: pack your research along with all data files, libraries, environment variables and options. Anyone can reproduce the research on a different machine

ReproZip

Case Studies Book:
a Collaborative MSDSE effort

- Collection of reproducible research workflows
- Tools, ideas, practices for real-world research projects
- Emphasis on practical aspects to make research as reproducible as possible



MSDSE's

Software meets Education

Berkeley UNIVERSITY OF CALIFORNIA

UC Berkeley Foundations of Data Science (Data 8) course:

- 1,000+ students – the fastest growing class in campus history

JupyterHub:

- Multi-user version of Jupyter Notebooks: great for classrooms!
- Jupyter Notebooks: Open-source web app for creating and sharing documents that contain live code, equations, visualizations and narrative text.

BERKELEY Institute for Data Science

jupyterhub

Campus Research Support

(The space between Office Hours and Grant Proposals)

UNIVERSITY of WASHINGTON
eScience Institute

Data Science Incubator

- Intensive data science consultation to advance research
- "Teach a person to fish" approach
- Provide a shared environment where researchers can learn from an in-house team, external mentors, and each other

UNIVERSITY of WASHINGTON
eScience Institute

Winter Incubator Program

- Quarter-long (10 weeks)
- In person engagement two days per week
  - Project Lead + Data Scientist
- Participation from faculty, grad students, staff
- 4-6 concurrent projects: Network effects among cohort beyond 1:1 interactions
  - Biology -> Political Science
  - Astronomy -> Brain Science

the "ah ha" moment!

Fruitful collaboration with potential for significant impact

MSDSE's

Example Projects from the Winter Incubator

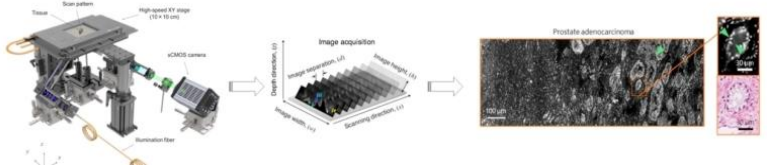
Using Social Media Data To Identify Geographic Clustering Of Anti-vaccination Sentiments

Analysis of Kenya's Routine Health Information System data

Simulating Competition in the U.S. Airline Industry

Developing a Workflow for Managing Large Hydrologic Spatial Datasets to Assist Water Resources Management and Research

3D Visualization of Prostate Cancer Using Light-Shee Microscopy

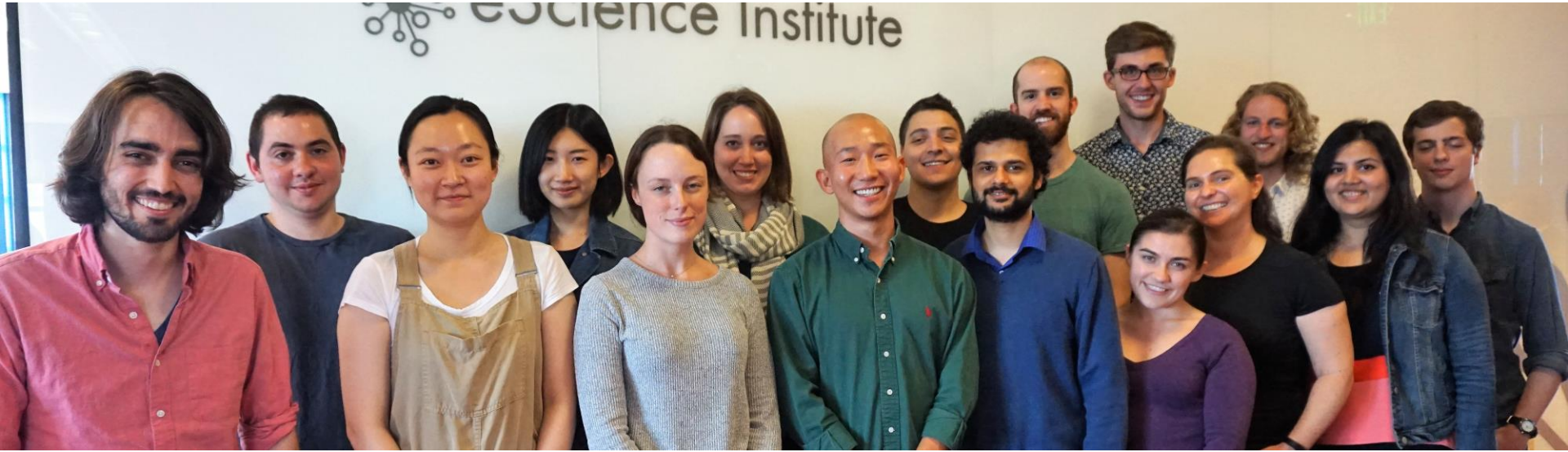


Damage Speaks: Acoustical Monitoring Framework for Structures Subjected to Earthquakes

Beyond the MSDSE's: Into the Community

Beyond the MSDSE's

UNIVERSITY of WASHINGTON
eScience Institute
DATA SCIENCE FOR SOCIAL GOOD



Brings together students and researchers with data science and domain expertise to work on focused, collaborative projects for societal benefit.

Data Science for Social Good

Project Teams

- Project Leads (1-2)
- Data Scientist Leads (1-2)
- DSSG Student Fellows (4) - highly competitive!
- Stakeholders

Example Project: Accessible Trip Routing


Cities can be incredibly complex to navigate.
For many people, technology provides the information needed to get around.

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*

Kevin’s Story

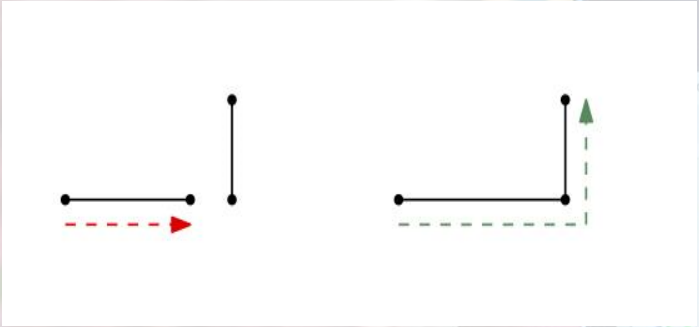
“ Using a tool like Google Maps doesn’t really help me get around. Actually sometimes this does more harm than good. I’m sent down streets I can’t cross, or up inclines that are impossible to climb. It can be deeply frustrating.”

AccessMap

Seattle

TCAT The Taskar Center for Accessible Technology

Connect sidewalks

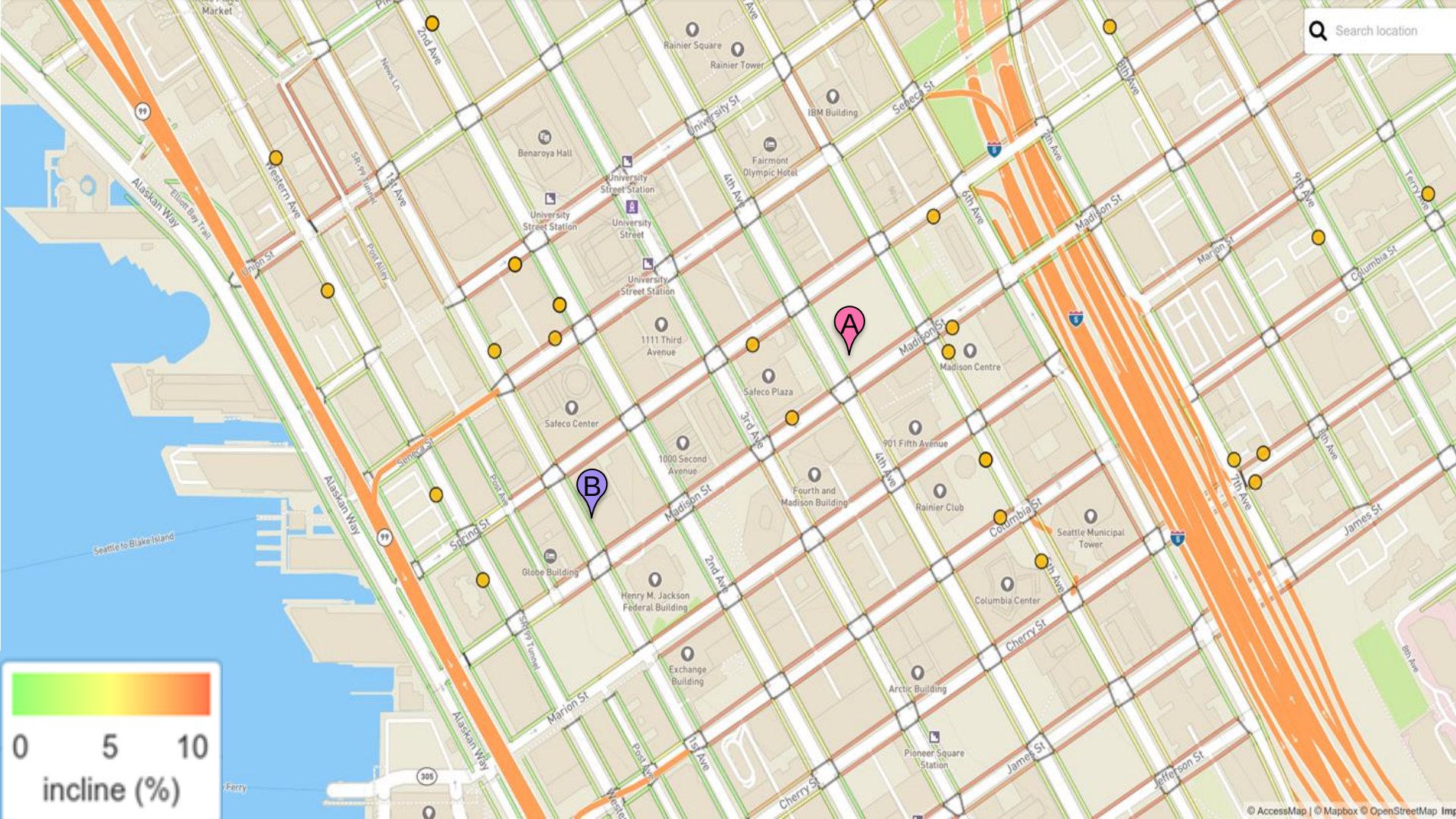


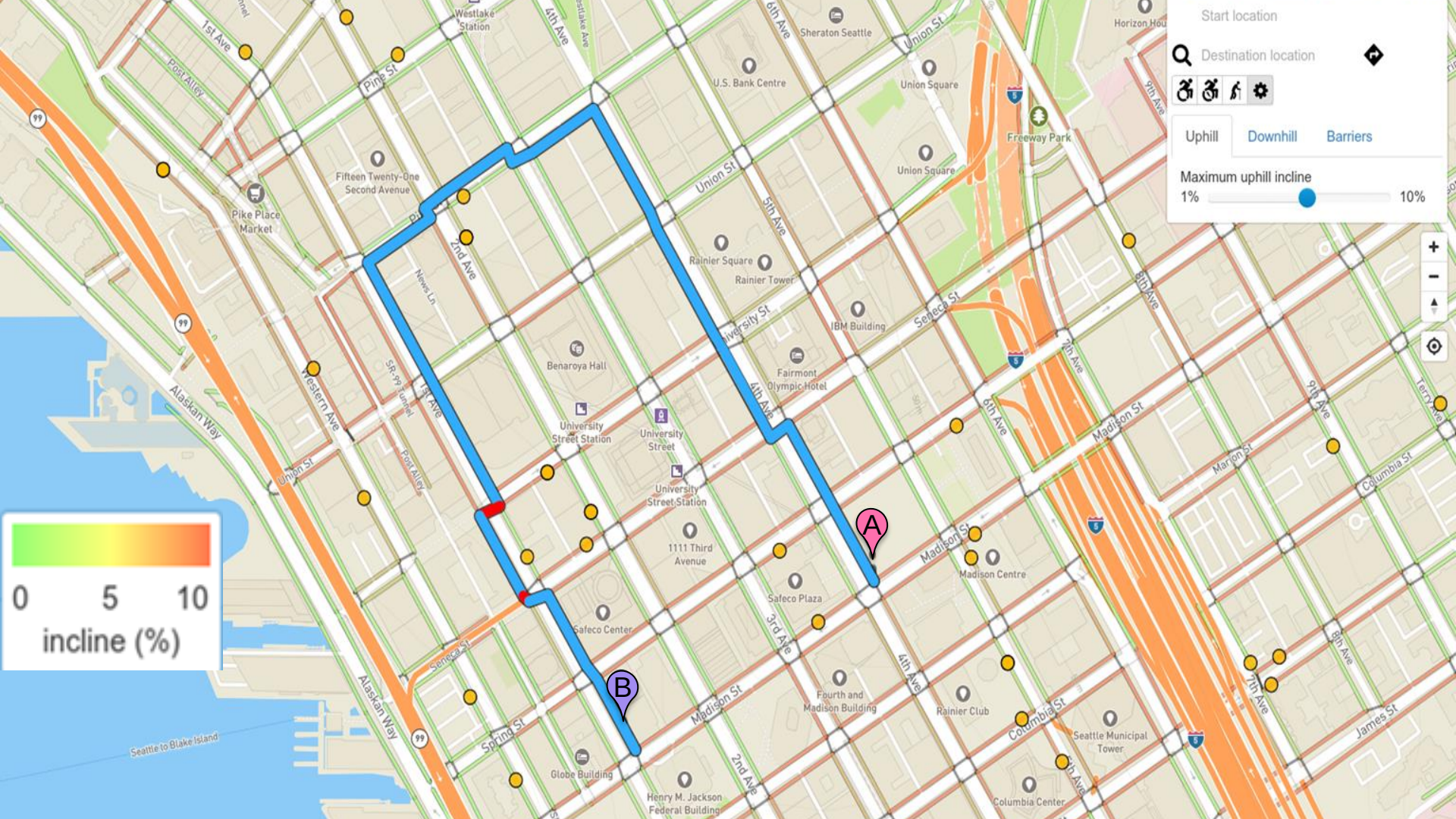
Use existing data to find the best route

Incline: USGS + Sidewalk lines

Construction: data.seattle.gov Permits

Curb ramps and crosswalks: data.seattle.gov


Search location
A
B
0
5
10
incline (%)
Alaskan Way
Elliott Bay Trail
Western Ave
1st Ave
2nd Ave
3rd Ave
4th Ave
6th Ave
7th Ave
8th Ave
9th Ave
Union St
Seneca St
University St
Madison St
Spring St
Marion St
Cherry St
James St
Columbia St
Jefferson St
Post Alley
Pine St
Terry Ave
Seattle to Blake Island
Rainier Square
Rainier Tower
IBM Building
Benaroya Hall
Fairmont Olympic Hotel
University Street Station
University Street
1111 Third Avenue
Madison Centre
Safeco Plaza
Safeco Center
1000 Second Avenue
901 Fifth Avenue
Fourth and Madison Building
Rainier Club
Seattle Municipal Tower
Globe Building
Henry M. Jackson Federal Building
Columbia Center
Exchange Building
Arctic Building
Pioneer Square Station
SR 99 Tunnel
99
305
Ferry
© AccessMap | © Mapbox © OpenStreetMap Im


Start location
Destination location
Uphill
Downhill
Barriers
Maximum uphill incline
1%
10%
0
5
10
incline (%)
A
B
Pike Place Market
Fifteen Twenty-One Second Avenue
Benaroya Hall
University Street Station
University Street
1111 Third Avenue
Safeco Center
Safeco Plaza
Fourth and Madison Building
Rainier Club
Seattle Municipal Tower
Columbia Center
Henry M. Jackson Federal Building
Globe Building
Rainier Square
Rainier Tower
IBM Building
Fairmont Olympic Hotel
Madison Centre
U.S. Bank Centre
Sheraton Seattle
Union Square
Freeway Park
Westlake Station
Alaskan Way
Seattle to Blake Island
Union St
Pine St
Seneca St
Madison St
Spring St
Marion St
Columbia St
James St
1st Ave
2nd Ave
3rd Ave
4th Ave
5th Ave
6th Ave
7th Ave
8th Ave
9th Ave
Western Ave
Post Alley
SR-99 Tunnel

Beyond the MSDSE's

Impact in the Community (resonates with University Leadership)

The Seattle Times

Education | Education Lab | Local News | Transportation

UW student project taps ORCA cards, unlocks data trove

GeekWire NEWS JOBS EVENTS RESOURCES DEALS ABOUT Search

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

Post a Comment | Share 68 | Tweet | Share 43 | Reddit | Email | GeekWire Gala early-bird tix on sale now!

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

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

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

Budding UW Data Scientists Use Their Powers for Social Good

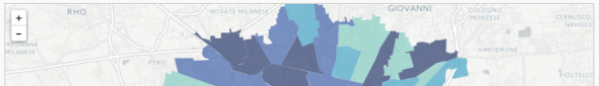
Benjamin Romano
August 24th, 2015

@bromano | @xconomy | Like Us

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

Beyond the MSDSE's

Scalable Research Impact: Community Learning Within Doma

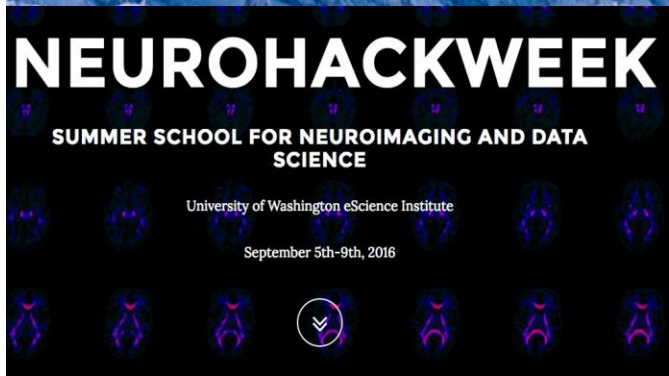
Hackweeks

shared language, shared scientific objectives

Components:

- (lots of) tutorials in introductory and state-of-the-art methodologies
- participant-driven project work in a collaborative environment
- peer-teaching and peer-learning *

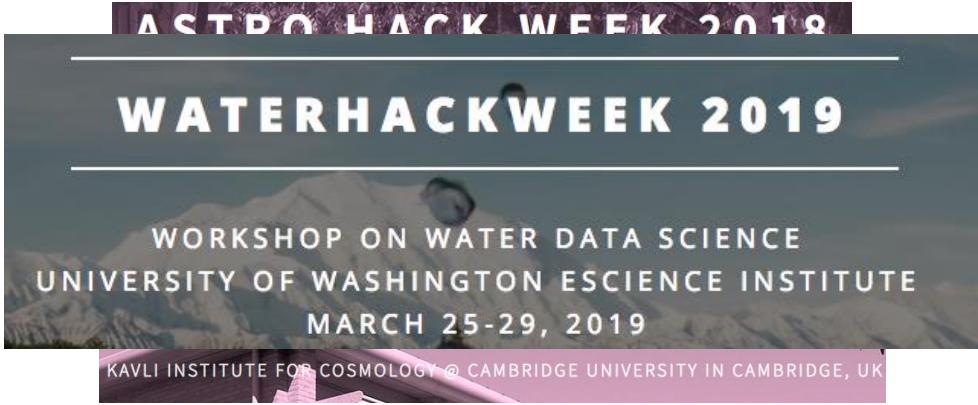
-> catalyze community



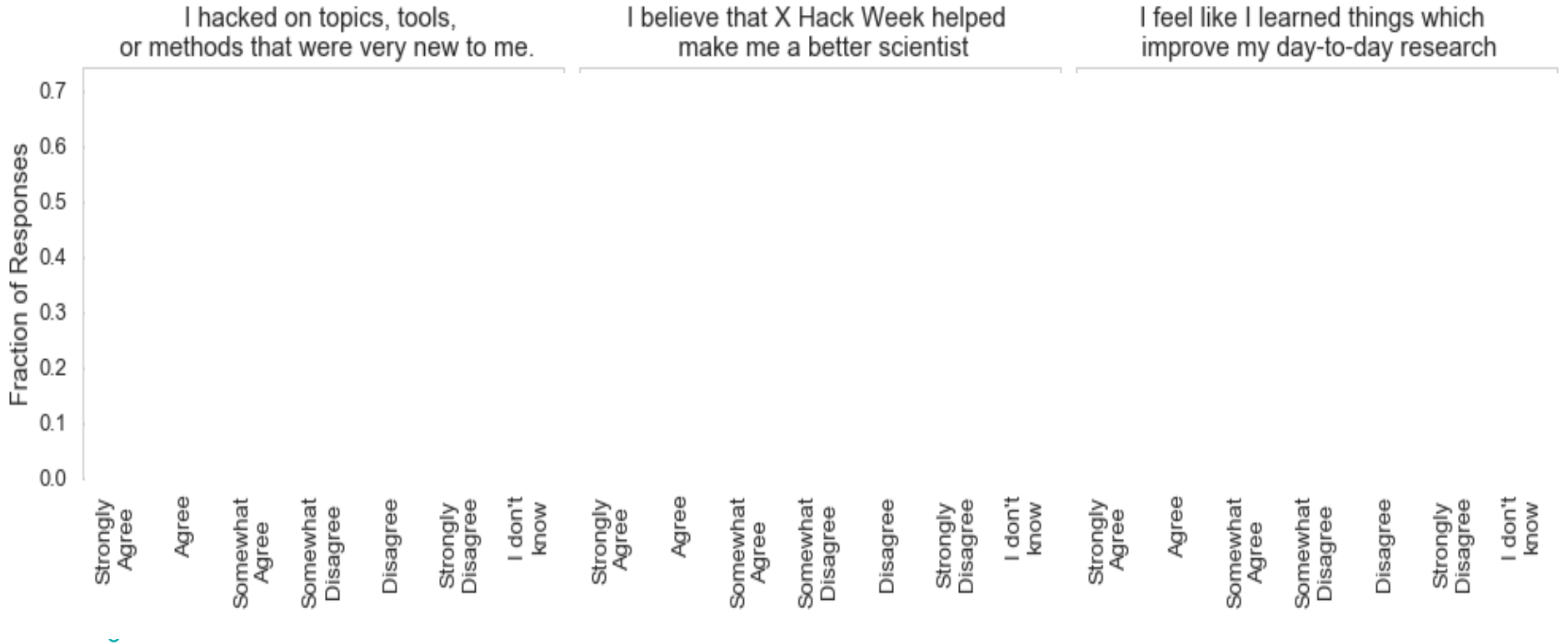
Beyond the MSDSE's

Hackweeks: Growth and Evolution

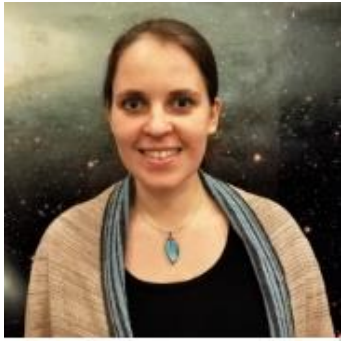
(Started in 2018)



Exit Survey Responses: Research Methods



Hackweek Leaders and Resources



Daniela Huppenkothen
Associate Director, DIRAC

David Hogg
Professor, NYU

Ariel Rokem
Senior Data Scientist, UW

Nicoleta Cristea
Research Scientist, Freshwater Initiative

Anthony Arendt
Senior Research Scientist, Polar Science Center, UW

Karthik Ram
Senior Data Scientist, UCB

Jake VanderPlas
Senior Data Science Fellow, UW

Christina Bandaragoda
Research Scientist, Civil & Evironmental Engineering

Hackweeks:
Huppenkothen et al, 2018 PNAS

Entrofy:
Huppenkothen et al, 2019
arXiv:1905.03314

Toolkit:
Arendt & Huppenkothen

uwescience.github.io/HackWeek-Toolkit

Scalable Research Impact: Community Learning Across Domains

XD Working Groups & Workshops

- XD's are methods-focused communities
  - host seminars, blogs
  - workshops: 2-3 days, include tutorials, talks by experts, and make sessions

- Inaugural ImageXD (2016):
  - 50 researchers, 14 institutions
  - computer vision, microscopy, materials imaging, photography, earth science, neuroscience, astronomy, software development, and more.

BERKELEY Institute for Data Science

XD's Growth and Evolution

- ImageXD had its 4th iteration
- Spawned:
  - TextXD (in 2017)
  - GraphXD (in 2018)

Example outcomes:

- workflows for open source image processing
- training sets for ML applications
- analysis projects

TextXD
Text Analysis Across Domains
https://www.textxd.org/

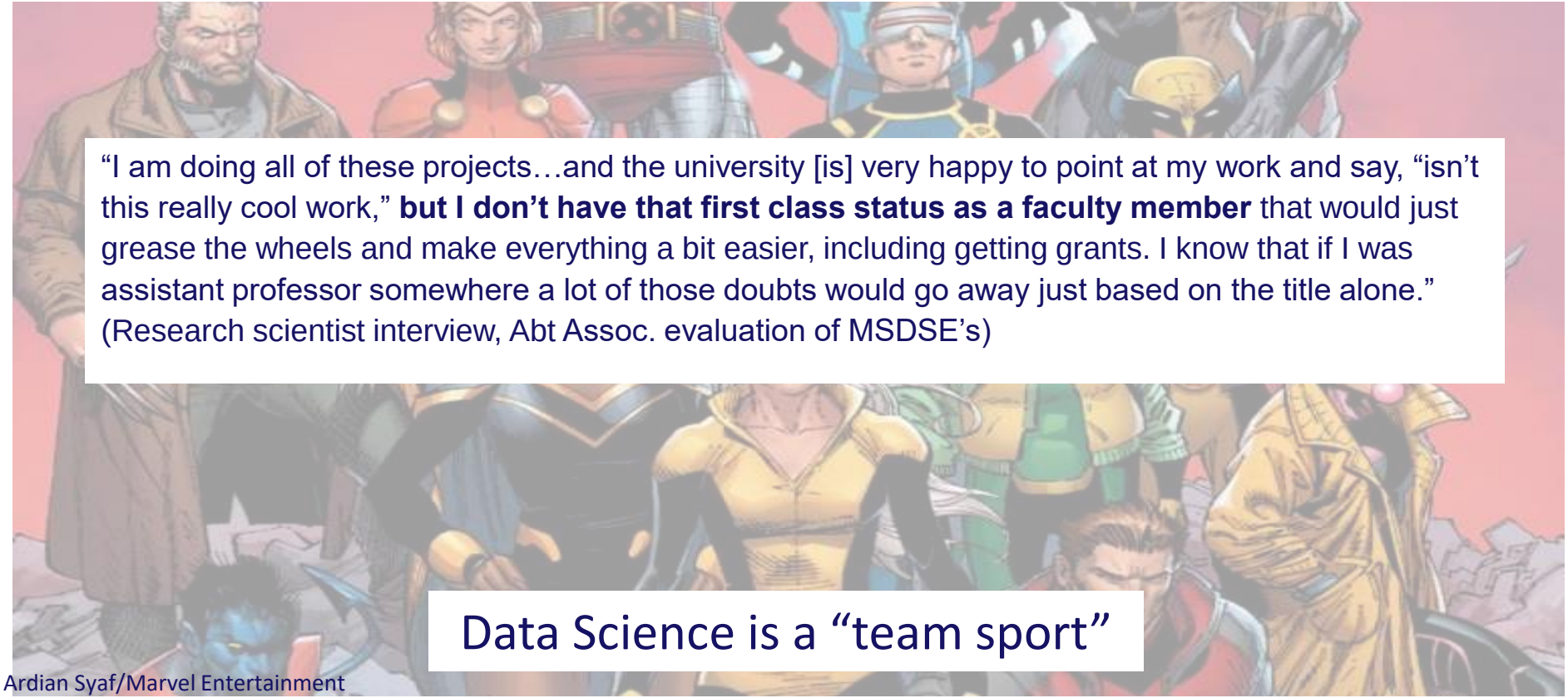
GraphXD
Graphs Across Domains

Key Takeaway

Informal intensive community-driven learning opportunities, like Hackweeks and *x*D workshops, quickly and effectively bring data science to campus researchers.

Remaining Challenges

Non-Faculty Career Paths in Academia



"I am doing all of these projects…and the university [is] very happy to point at my work and say, "isn't this really cool work," **but I don't have that first class status as a faculty member** that would just grease the wheels and make everything a bit easier, including getting grants. I know that if I was assistant professor somewhere a lot of those doubts would go away just based on the title alone." (Research scientist interview, Abt Assoc. evaluation of MSDSE's)

Data Science is a "team sport"

Ardian Syaf/Marvel Entertainment

Challenge: Viable Career Paths

Common themes from the Landscape Survey of 20 Data Science Centers (Abt Assoc.)

Most non-faculty positions in academia:

- are temporary appointments (1-2 year) on “soft” money
- have non-competitive salaries
- lack an obvious promotion path

“I think there is a degree of structural change going on in the academy, but I think that it's happening very slowly…Do these kind of positions of leadership that are not tenure-track faculty get created? If not, I'll ***probably end up going to work for some other non-profit, open source type of place****.”* (Staff data scientist)

“Mentoring for the data scientists and research scientists to help them figure out what to do strategically for themselves, their careers, it isn’t something that is really addressed now, and it is hard because these are new jobs in academic research which means ***we need more mentoring not less****.”* (Staff data scientist)

Challenge: Viable Career Paths

What can universities do to compete?

- PI status!
- “Competitive” salaries and titles (”Professor of Practice”?)
- Highlight the advantages of a university: intellectual environment and opportunities to mentor and teach
- Give them the ability to mentor students and postdocs
- Elevate software and workflow contributions to “publication count” in hiring and tenure reviews
- And early career mentorship

Institutional Challenges

- Greatest challenge: navigating the university's political landscape and persuading the faculty that they would benefit from a data science center.
  - **engage the university community in the design process**.

- Defining the "place" for a Data Science Center. Is it its own School? Is it a core element of the university? Part of the Libraries or Research IT, or both? Or wholly independent?
  - **dedicated space and a strong emphasis on collaboration**, **interdisciplinarity, and community building**. (Virtually all entities in the landscape survey are administratively based outside of any one department or school)

- Faculty involvement: Balance the engagement expectations and departmental obligations.
  - **Provide teaching releases or access to discretionary funding to support their research while they support the work of the data science center**

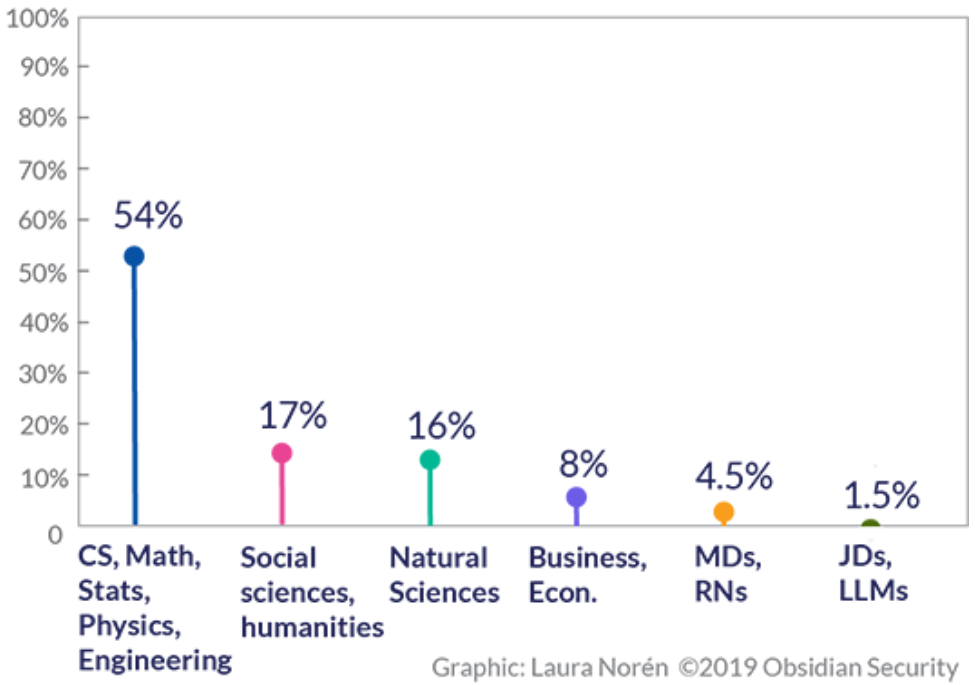
Community Challenge for Data Science: Diversity

“We have a chance to get it right from the beginning”

Who's Building Your AI? A Research Brief

by Laura Noren, Gina Helfrich, and Steph Yeo

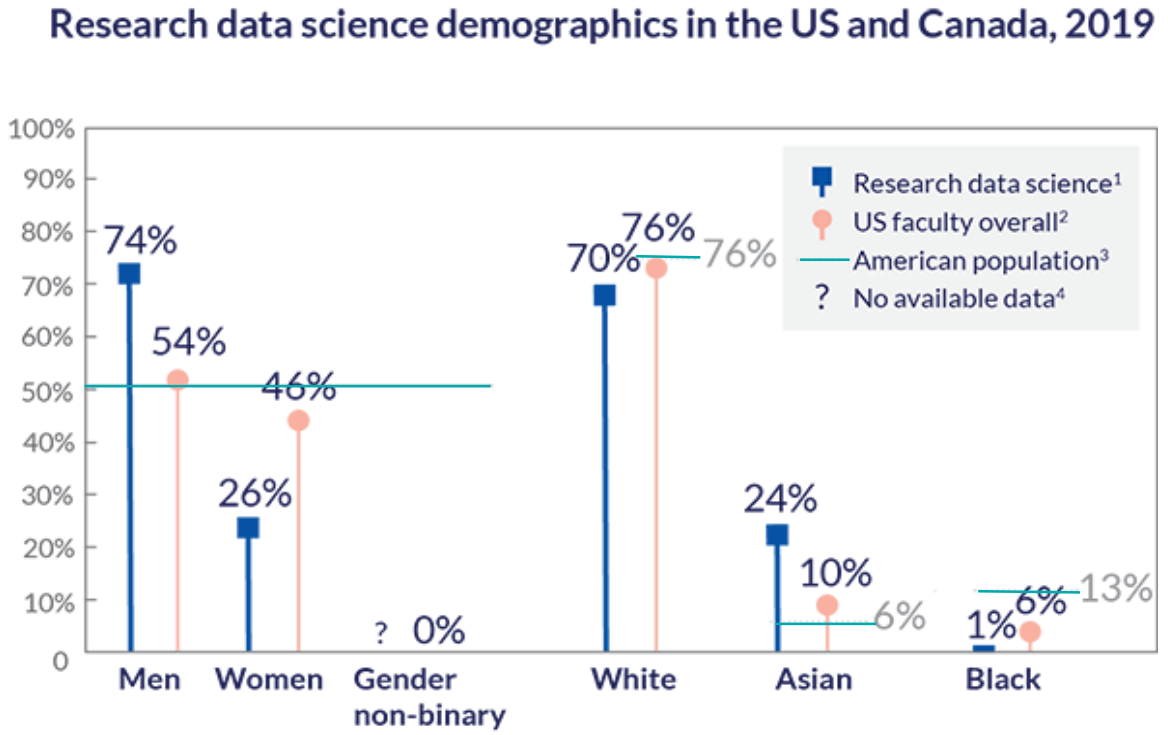
- ~3300 individuals, 41 data science and/or AI research centers, US and Canada
- gathered the data manually, mostly from institutional websites
- Each institute was given a chance to review and correct the data

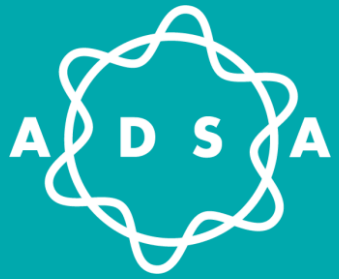


Who's Building Your AI? A Research Brief

by Laura Noren, Gina Helfrich, and Steph Yeo

- The authors recognize the problem with lumping diverse cultures into these broad race categories; versus political implications of reporting nationality.
- 4% of sample did not fit into the white, black or Asian categories

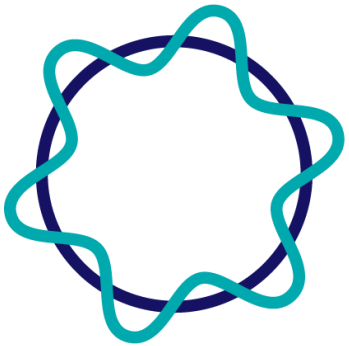




The Academic Data Science Alliance

The Academic Data Science Alliance

a community-building organization that supports university researchers in their efforts to learn, use, and teach data-intensive methodologies and responsible applications



Academic
Data Science
Alliance

Where do we begin?

MSDSE 3 Annual Summits

opportunity for data savvy researchers to share and learn tools and methods outside their domain

Transition MSDSE Summit to ADSA

The ADSA Leadership Summit

leaders of academic data science initiatives, and faculty interested in creating new initiatives on their campuses

- to form an academic leaders community for data science;
- to share best practices where they face similar challenges and opportunities;
- to take collective responsibility in preparing next-generation data scientists to contribute in the best interests of society

ADSA

The Leadership Summit

"I did less reading of my email this week than at any conference in recent memory."

Special Interest and Working Groups

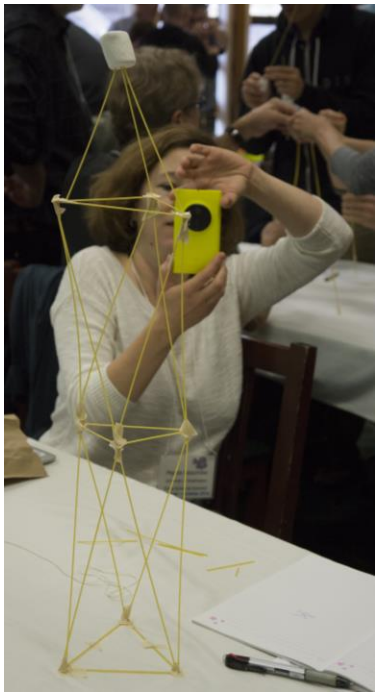
bring together thought leaders in our community to tackle pressing challenges throughout the year

Special Interest Groups:

- Education
- Diversity, Equity, Inclusion

Working Group:

- Code of Ethics



Early Career Support: The Data Science Co-Op

Mission statement

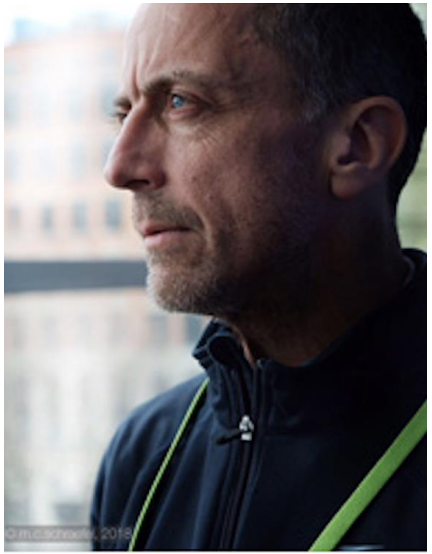
- **trusted and growing community** of (mostly academic) data scientists
- **peer-powered culture**
- collaborative infrastructure and opportunities **helping us share our expertise**
- align with academic values like **transparency, inclusion, publishing, and openness**

Data Science Community Newsletter

Sign up here.

The Data Science Community Newsletter (DCSN) is a witty, informative weekly newsletter launched in 2015 and wholly supported by the Academic Data Science Alliance. It is written by Laura Norén and curated by Brad Stenger.

https://cds.nyu.edu/newsletter/



Academic
Data Science
Alliance

Thank you!

micaela@academicdatascience.org

www.academicdatascience.org

ADSA

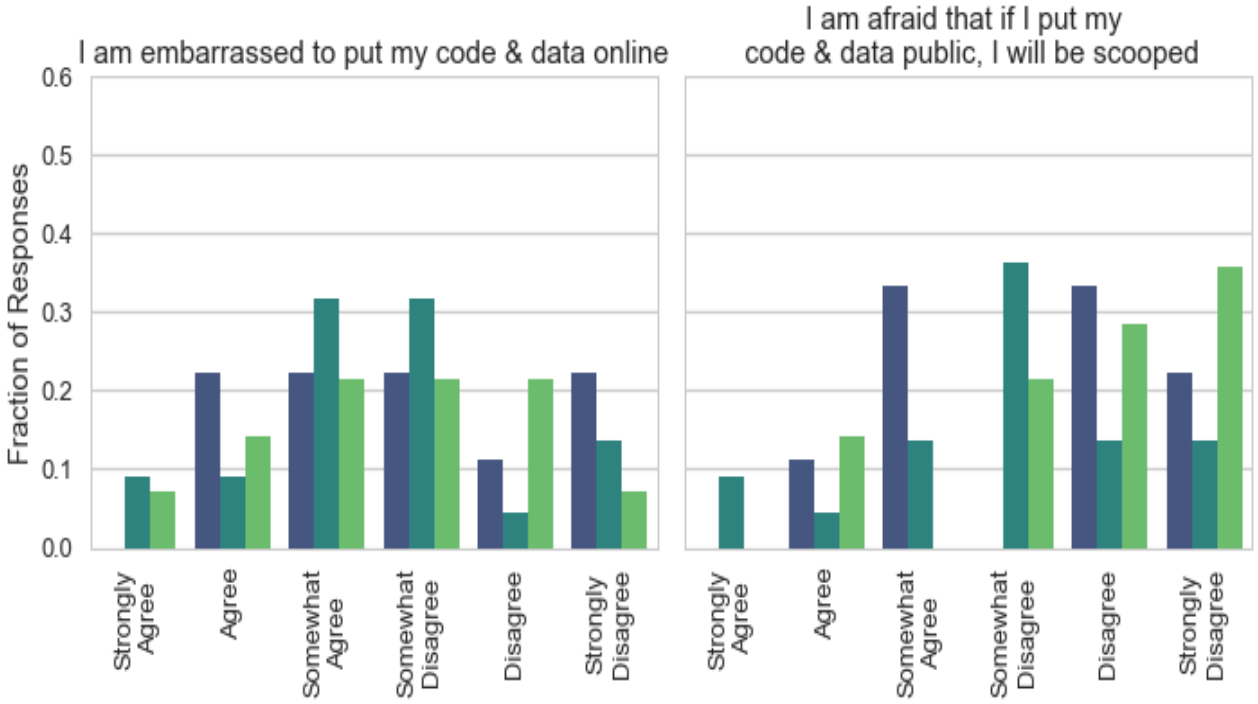
@AcademicDataSci

Early Career Support: The Data Science Co-Op

Beyond the MSDSE's

Data, Responsibly

Exit Survey Responses: Open Science



Designing Working Spaces and Culture

- Neutral space on campus for collaboration
  (Partner with campus libraries)
- Take advantage of the "water cooler effect"
- Design Considerations
  - Drop-in open workspace, small & large meeting rooms
  - Hot desks & casual seating, flexible & transformable
  - Writeable surfaces

“One thing that I think we talk a lot about and I think has been verified, is that **having a neutral space on campus is important**. We’re not viewed as part of the computer sciences department or another department in particular. There’s this sort of **Switzerland effect**, you’re outside of the departmental silos. People come here and are more likely to collaborate across disciplines than they might otherwise be if they were all going to somebody’s particular department.”

(Interview of MSDSE participant, Abt Associates Final Evaluation)