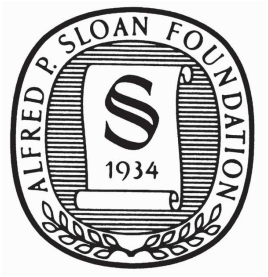
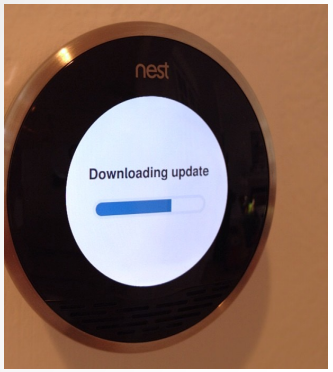


Micaela Parker
Program Coordinator

Chris Mentzel
Gordon and Betty Moore Foundation

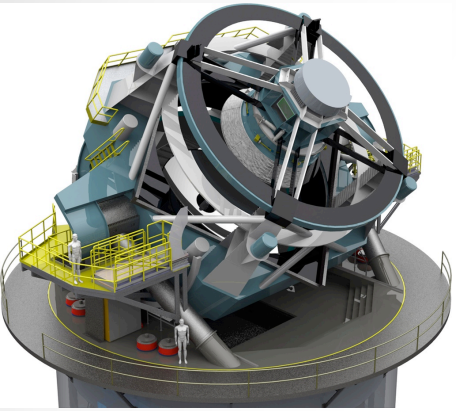
Josh Greenberg
Alfred P. Sloan Foundation

Data are being collected and used everywhere!



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

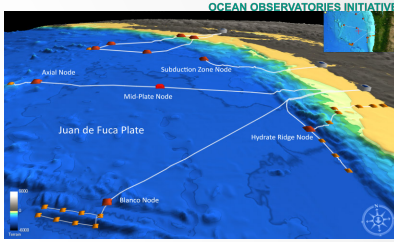
Nearly every field of discovery is transitioning from "data poor" to "data rich"



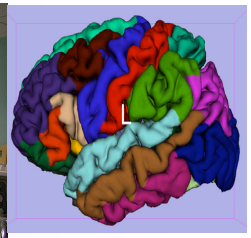
Astronomy: LSST

Physics: LHC

Digital Humanities



Oceanography: OOI



Health

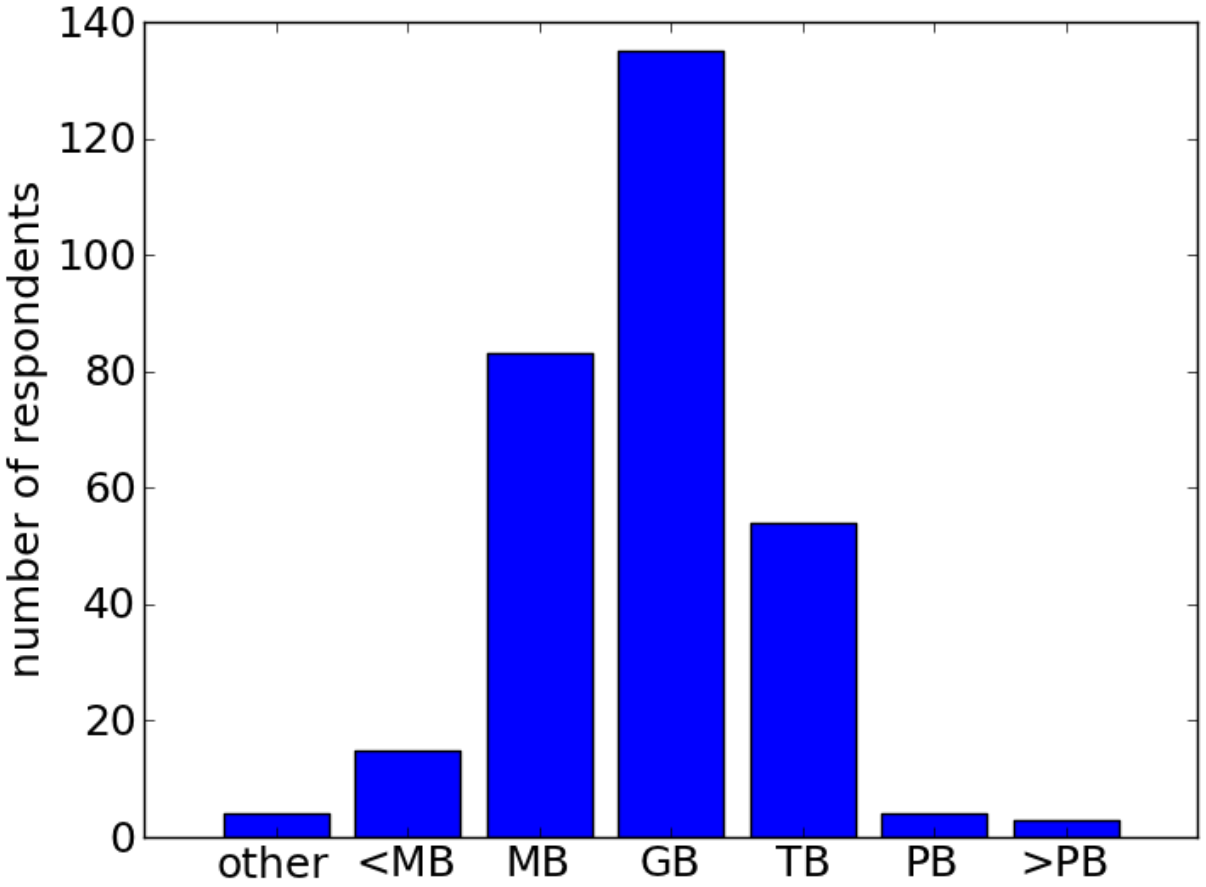
Biology: Sequencing

Economics: POS terminals

Sociology: Social Media and the Web

Data Science challenges are not just about size

How much data do you work with?



Wright 2013

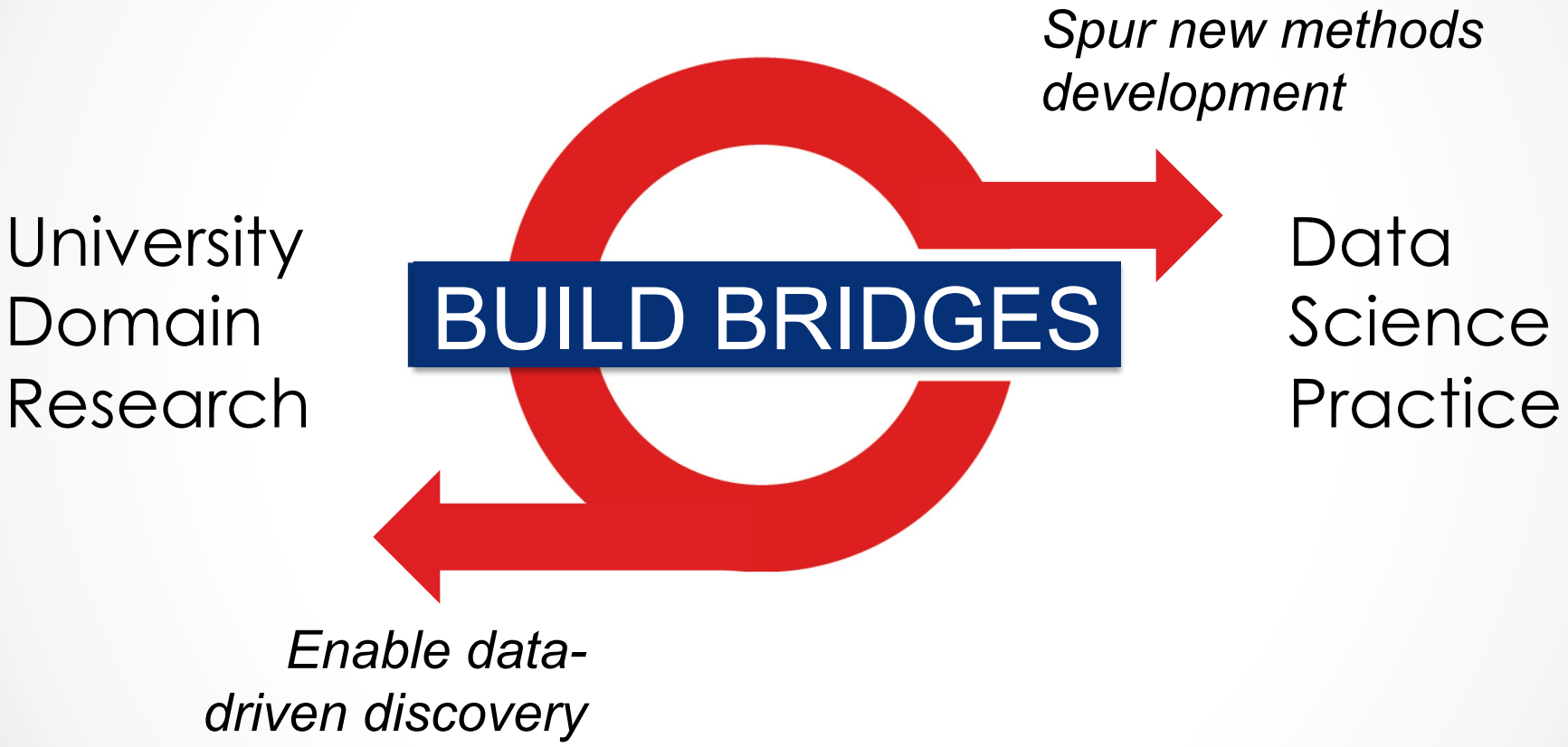
The Challenge

University
Domain
Research

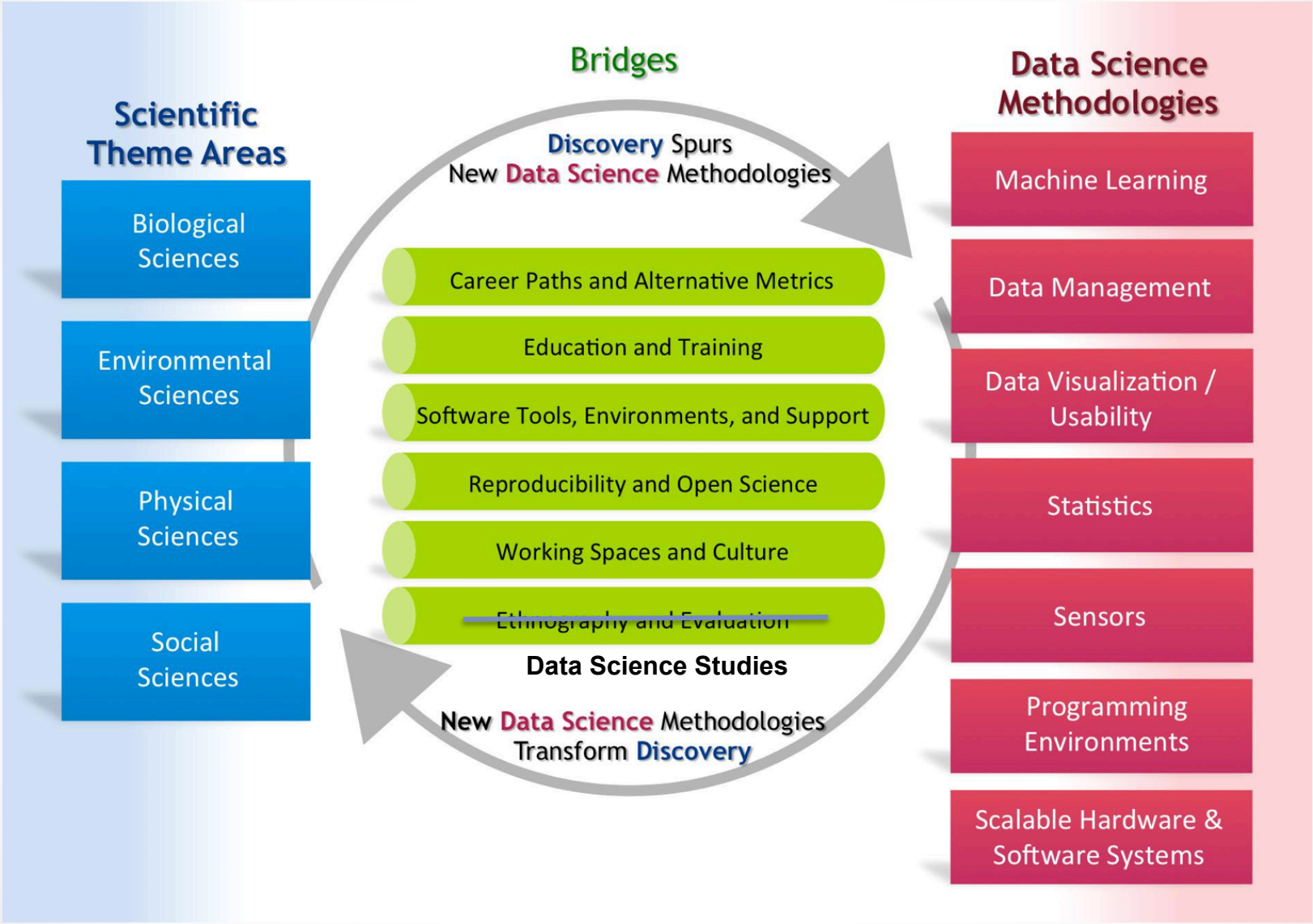
Data
Science
Practice

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

The Grand Experiment



Building Bridges: Our Efforts Organized into Working Groups



The Virtuous Cycle

Outline – a talk in two parts

- Highlights from 5 years of MSDSE Collaboration
  - cross-university, collaborative efforts
  - individual university achievements

- A few Key Takeaways and Institutional Challenges
  - the MSDSE's
  - a Landscape Survey of 20 DS Centers Nation-wide (Abt Associates)
  - a Final Evaluation of the MSDSE's by Abt Associates
  - the inaugural Data Science Leadership Summit (March 2018)

But first, a nod to some unsung heroes...

Ethnography and Evaluation → 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

- Ethnography meets reflective and reflexive self-evaluation
- Embedded ethnographers provided immediate feedback of programs and activities –> responsiveness to issues and adaptable nature of the DSE's.
- The WG raises awareness of ethical issues and surfaces best practices to the larger community.
- In their scholarly work, they use computational, HCI, historical and ethnographic approaches to studying the practices, tools, and culture of data science





Reproducible and Open Science

NYU Center for Data Science

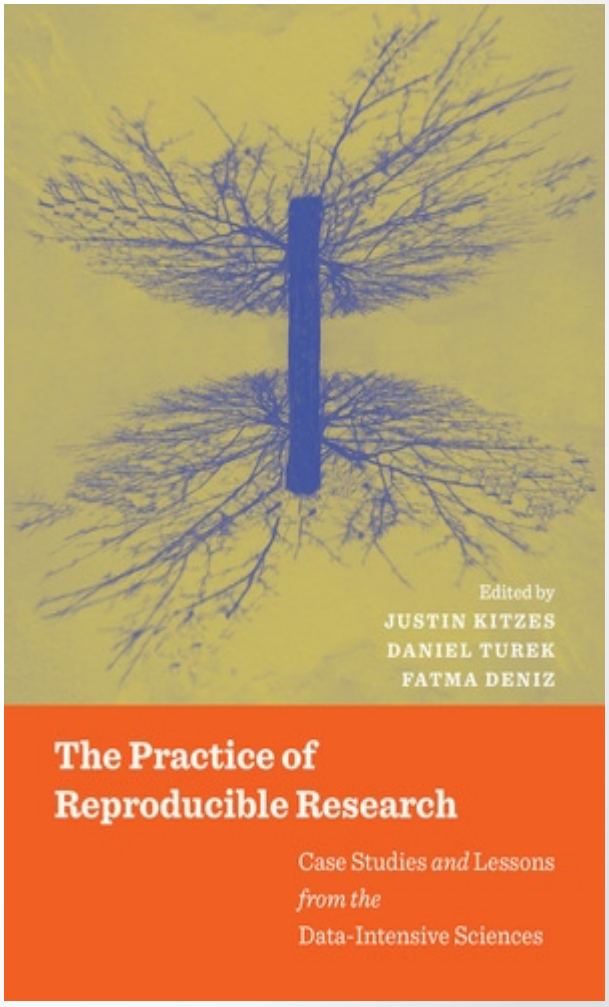
ReproZip

- Hired first reproducibility librarian in a tenure-track position!

- ReproZip: pack your research along with all necessary data files, libraries, environment variables and options. Then anybody can reproduce the research on a different machine, without tracking down and installing the dependencies, or even having to run the same operating system!

<u>Case Studies Book – a Collaborative MSDSE effort</u>

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



Software meets Education

UC Berkeley Foundations of Data Science (Data 8) course with 1,000+ students – the fastest growing class in campus history

Berkeley
UNIVERSITY OF CALIFORNIA

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 in the classroom

jupyter

Files | Running | Clusters

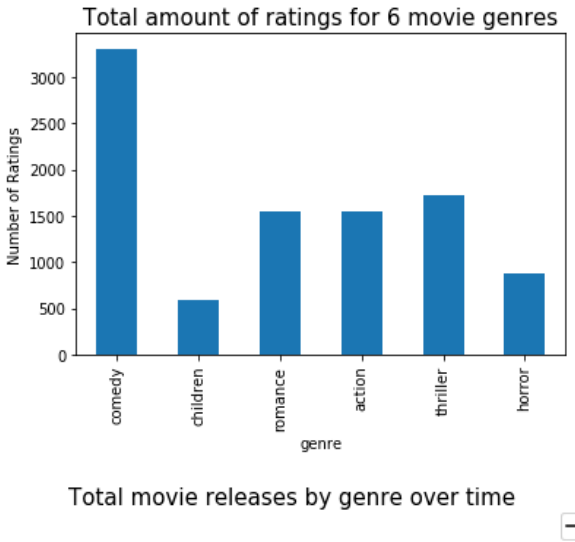
Select items to perform actions on tl

0

- adelepa
- adf
- agbriggs
- aguiao
- allykli
- aly27
- antilla
- asimon21
- bcraft
- bcraft-40uw-2eedu
- bgruenke
- bhaktib
- blbaron
- cy28
- donoh95
- dscheid
- edgaro

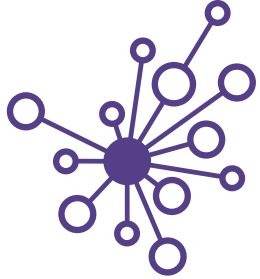
```
# Uses pandas and matplotlib to plot the THRILLER time series plot which is the bottom part of the
# second visualization (small multiples) to show the amount of releases the genre has had from the
# year 1920 to the year 2020.
totaltags4 = ['year', 'Thriller']
df4 = pd.read_table('thrillerovertime.csv', sep= ',', names= totaltags4)
df4.plot(legend = True, x = 'year')
axes = plt.gca()
axes.set_ylim([0,120])
axes.spines['bottom'].set_color('#000000')     # Similar nine lines of code as the previous section.
axes.spines['top'].set_color('#ffffff')
axes.spines['right'].set_color('#ffffff')
axes.spines['left'].set_color('#ffffff')
axes.tick_params(axis='y', colors='white')
plt.gca().get_lines()[0].set_color("red")
plt.gca().get_legend().legendHandles[0].set_color('red')
my_xticks = axes.get_xticks()
plt.xticks([my_xticks[0], my_xticks[-1]], visible=True, rotation="horizontal")
```

✓ In 2017 JupyterHub implemented at U. Washington



Data and photos from UW course: HCDE 411, courtesy of Brock Craft

Research Support



UNIVERSITY of WASHINGTON
eScience Institute

Data Science Incubator Programs

(The space between Office Hours and Grant Proposals)

- 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*

Winter Incubator Program

- Quarter-long (~10 weeks)
- In-person engagement two days per week
  - Project Lead + Data Scientist

the "ah ha" moment!

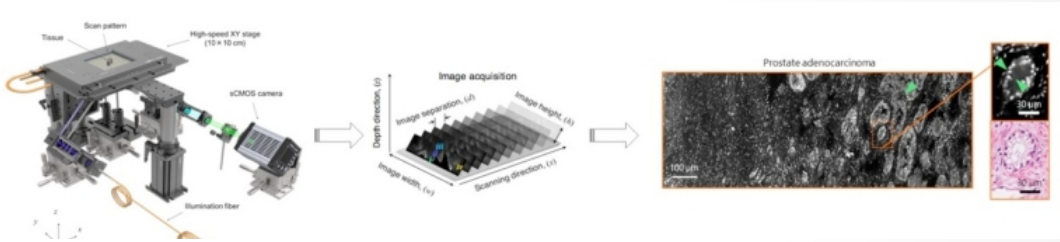
- Participation from faculty, grad students, staff
- 4-6 concurrent projects chosen by light-weight proposal process
- Network effects among cohort beyond 1:1 interactions
  - Biology -> Political Science
  - Astronomy -> Brain Science

Examples from 4 years of Incubator

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

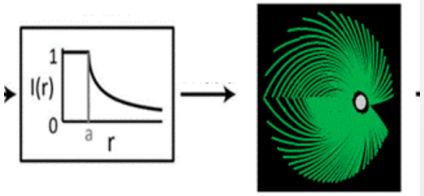
Using social media data to identify geographic clustering of anti-vaccination sentiments

3D Visualization of Prostate Cancer Using Light-Sheet Microscopy



Scalable Manifold Learning for Large Astronomical Survey Data

Improved Stimulation Protocols for Sight Restoration Technologies



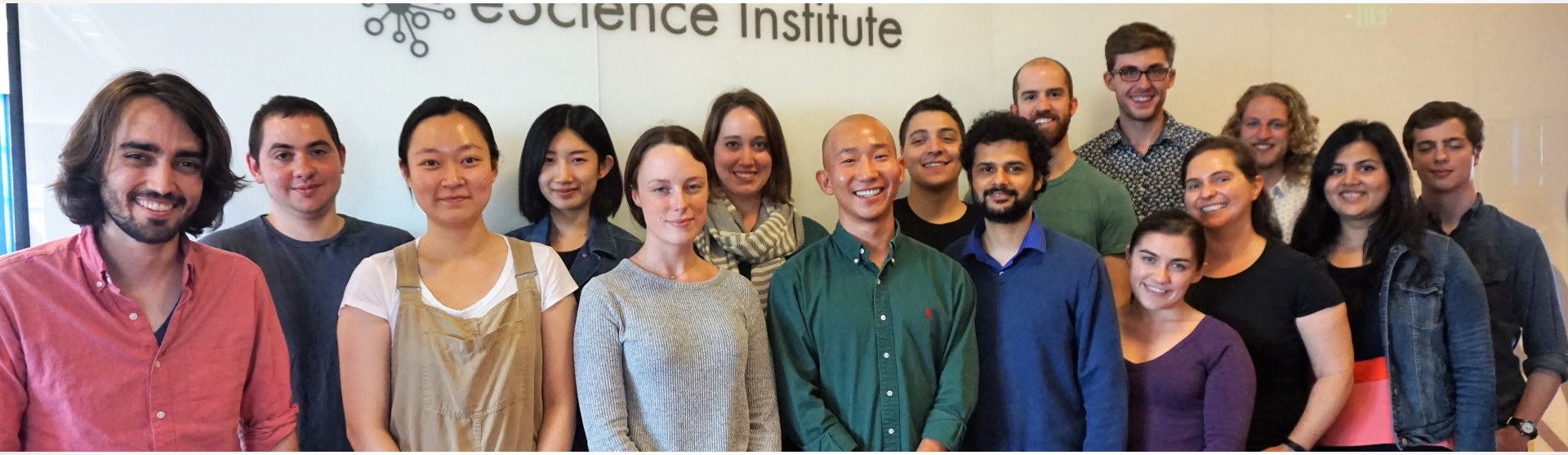
Simulating Competition in the U.S. Airline Industry

Analysis of Kenya's Routine Health Information System data

Damage Speaks: Acoustical Monitoring Framework for Structures Subjected to Earthquakes

Summer Incubator Program

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.

Examples from 3 years of Data Science for Social Good

Open Sidewalk Graph for Accessible Trip Planning
The Taskar Center for Accessible Technology

Predictors of Permanent Housing for Homeless Families
Bill and Melinda Gates Foundation

Mining Online Data for Early Identification of Unsafe Food Products
Institute for Health Metrics and Evaluation, Department of Global Health

Use of ORCA data for improved transit system planning and operation
Washington State Transportation Center

Strengthening capacities, knowledge and data sharing platforms for sustainable development
Conservation International, 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

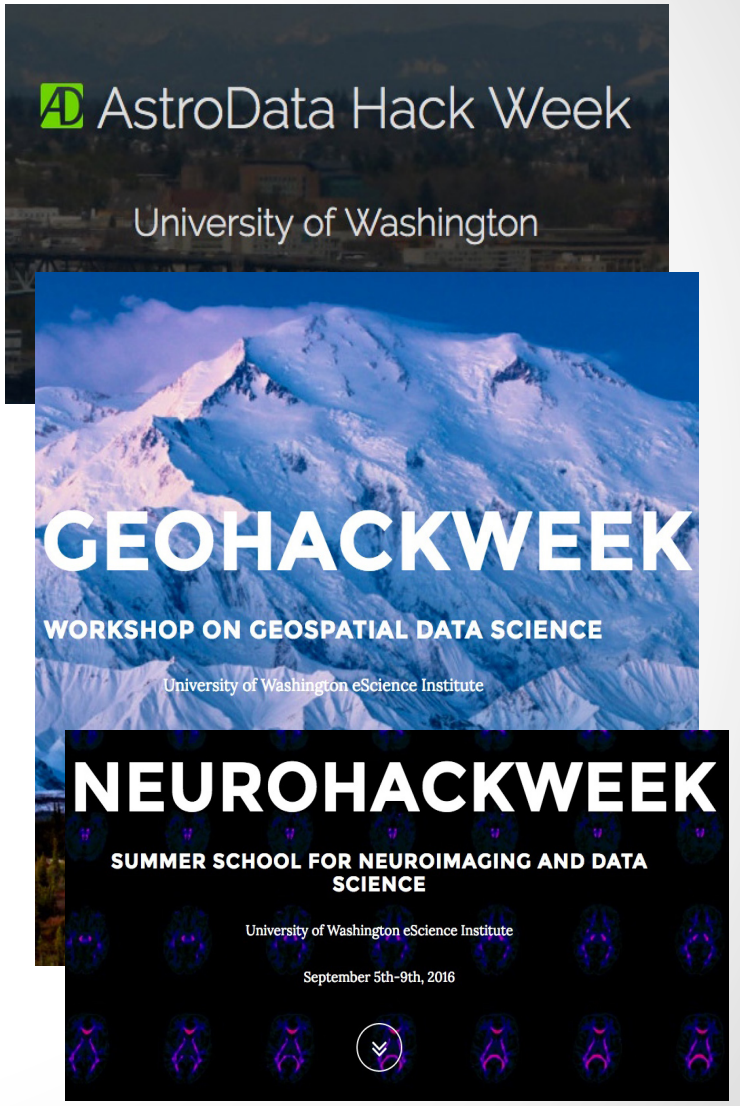
Scalable Research Impact: Community Learning Within Domains

Hackweeks

- building a culture of practice and developing resources within an existing domain-specific community

- week-long, 3 main components:

    - tutorials in state-of-the-art methodology
    - project work in a collaborative environment
    - peer-teaching and -learning



MSDSE
Moore - Sloan
Data Science Environments

Scalable Research Impact: Community Learning Within Domains

Hackweeks

- domain-focused communities

- week-long, three components:
  - tutorials
  - project work
  - peer-learning

Signs of Success

- AstroHackweek's 5th iteration goes international to Leiden, Netherlands in 2018
- NeuroHackweek → NeuroHackademy, 2 week summer program
- New Hackweeks this year: WaterHack, OceanHack, SocioHack(?)
- Outcomes include papers, software, and results (e.g. renewable energy sourcing in the state of WA)

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.

Scalable Research Impact: Community Learning Across Domains

XD 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.

TextXD
Text Analysis Across Domains

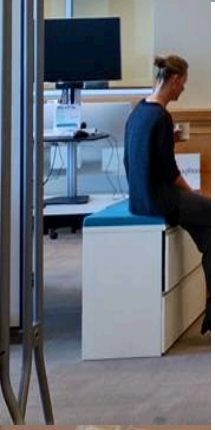
GraphXD
Graphs Across Domains

Signs of Success

- ImageXD had its 3rd iteration, spawned: TextXD (in 2017), GraphXD (in 2018)
- Example outcomes: blueprints for open source image processing, training sets for ML applications, analysis projects

Key takeaway

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



“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.”

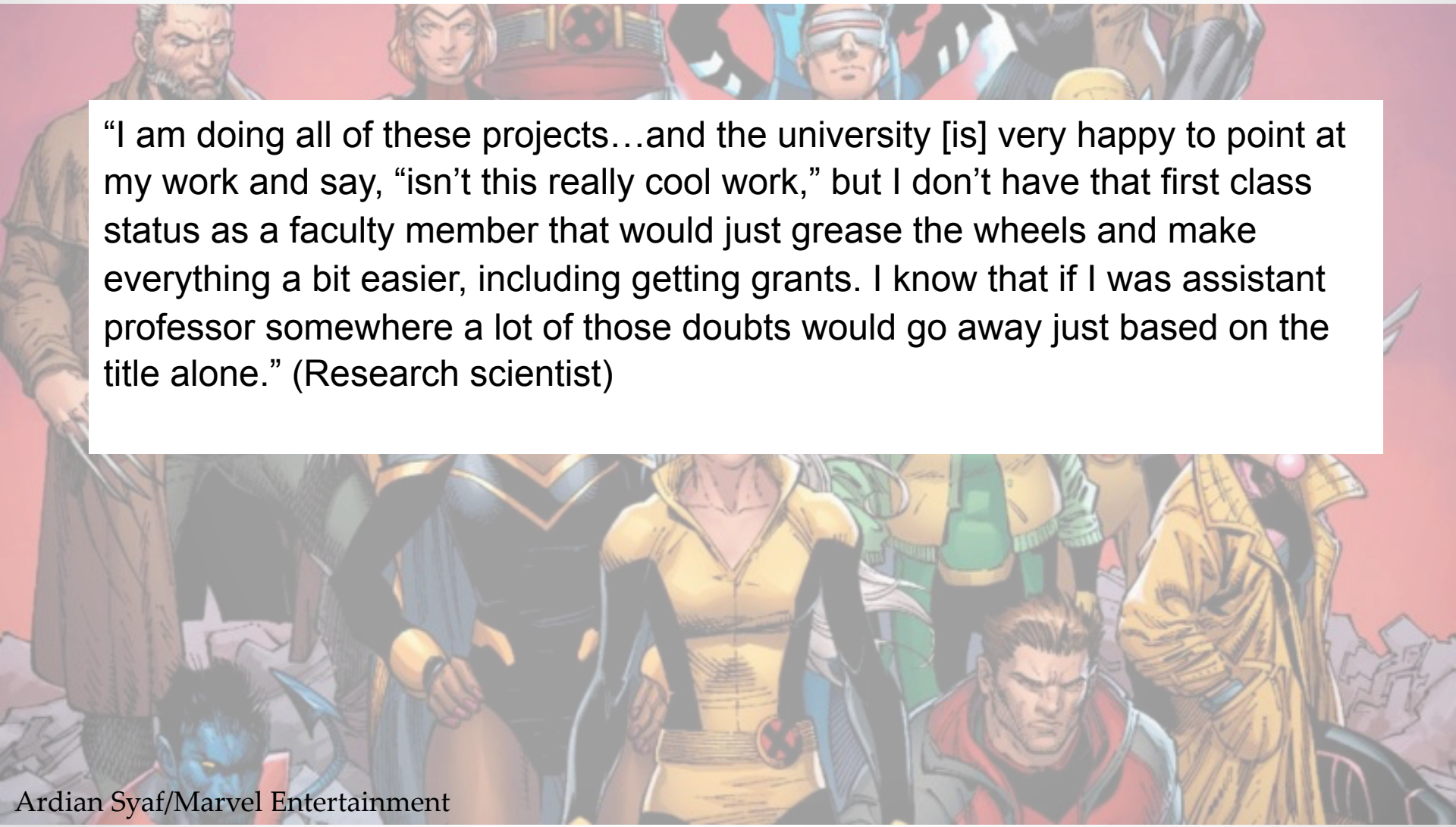
(Interview of MSDSE participant, Abt Associates Final Evaluation)

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

Career paths for academic data scientists

Data Science is a "team sport"



"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)

Ardian Syaf/Marvel Entertainment

Challenge: Viable career paths to attract and retain data science talent

Common theme from the Landscape Survey of 20 Data Science Centers
(Abt Associates)

- Academic labs struggle to obtain computational support they need
  - Salaries for data scientists on the market exceed full professor salaries
  - Most academic data scientist positions are contingent on grant funding
- Data scientist positions are difficult to create at a university
  - Many universities don't have a prestigious tier for staff to match faculty lines

How can academia make these positions more attractive?

Challenge: Viable career paths to attract and retain data science talent

- PI status
- Highlight the advantages of the university environment: a more intellectual environment and opportunities to mentor and teach
- Give them the ability to mentor students/postdocs
- "Competitive" salaries and titles ("Professor of Practice"?)
- And early career mentorship!

"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)

More Challenges and Lessons Learned

- Establishment of a Center
  - 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.
- Foundational elements of Data Science Centers
  - 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 a data science center
- Credit for Software
  - Elevating software and workflow contributions to "publication count" in hiring and tenure reviews
- Paths to sustainability
  - Data science as a core element of the university: Part of the Libraries or Research IT, or both?

Addressing the Challenges: Community Building For Institutions

Annual Meetings

- Data Science Leadership Summit
  - opportunity for thought leaders to discuss challenges and lessons learned as academia adapts to the data science revolution

- MSDSE Data Science Summit
  - opportunity for data savvy researchers to share and learn tools and methods outside their domain

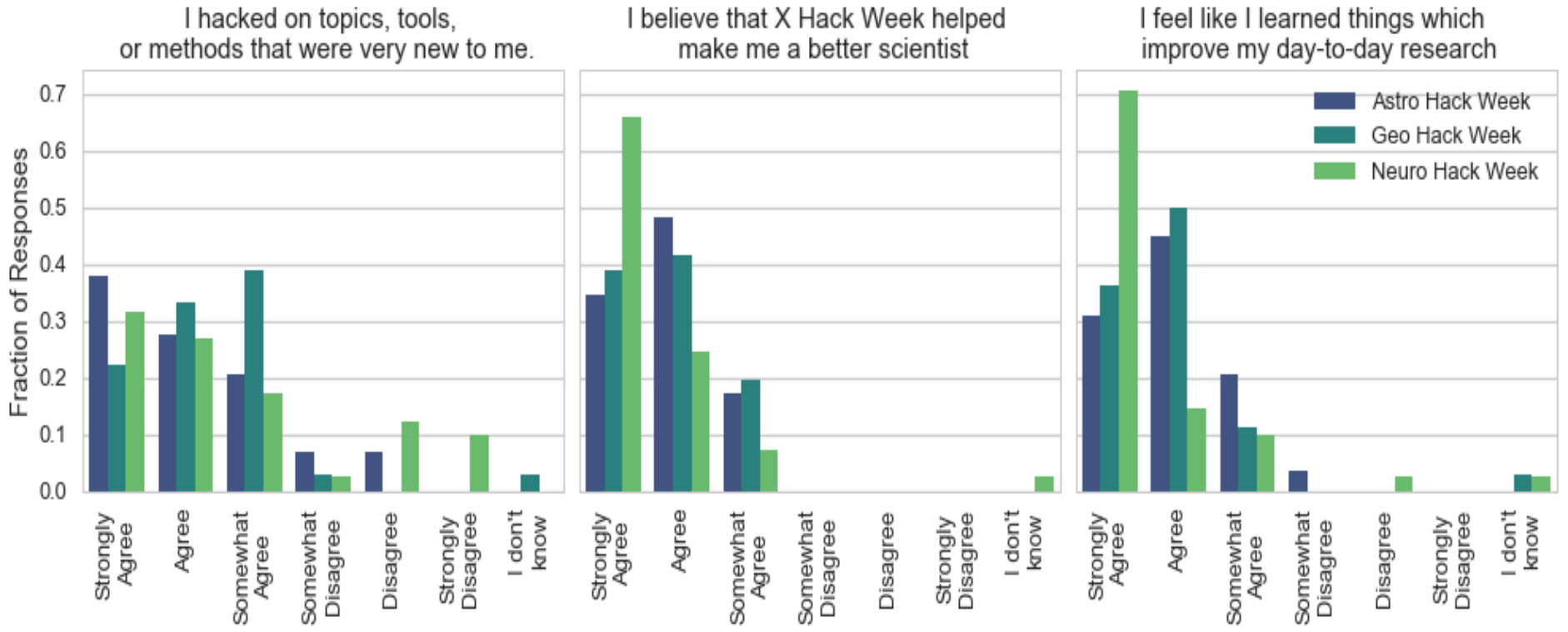
Thank you!

micaela@msdse.org

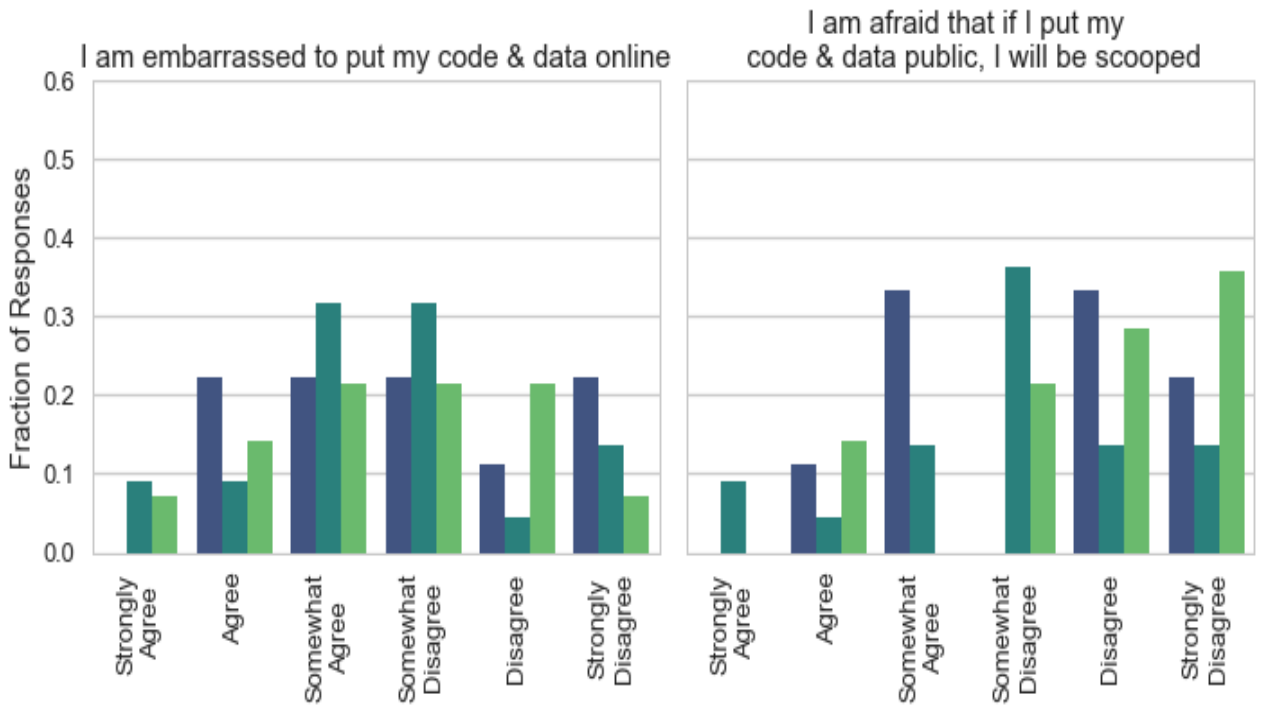
http://msdse.org

"Creating Institutional Change in Data Science"
Chronicles of Higher Ed, Mar 2018
http://msdse.org/

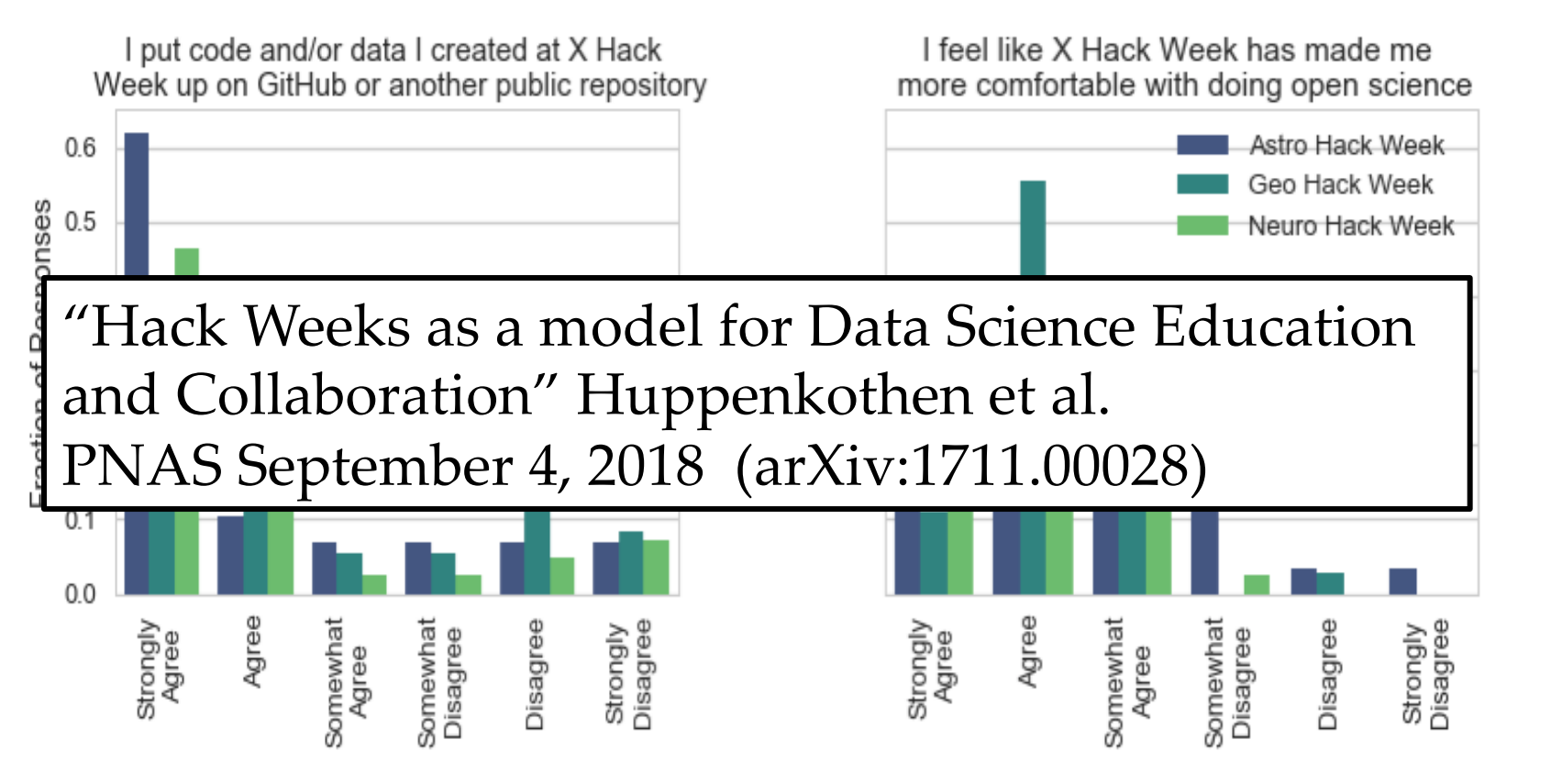
Exit Survey Responses: Research Methods



Exit Survey Responses: Open Science



Exit Survey Responses: Open Science



"Hack Weeks as a model for Data Science Education and Collaboration" Huppenkothen et al. PNAS September 4, 2018 (arXiv:1711.00028)