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 everywhere!

- 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”
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
INTRODUCTION

Example from my own work

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

Adrian Marchetti², Micaela S. Parker², Lauren P. Moccia², Ellen O. Lin¹, Angele L. Arrieta¹, Francois Ribalet¹, Michael E. P. Murphy², Maria T. Maldonado² and E. Virginia Armbrust¹

2008

Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability

Adrian Marchetti²,³, David M. Schruth², Colleen A. Durkin², Micaela S. Parker², Robin B. Kodner², Chris T. Berthiaume⁶, Rhonda Morales⁶, Andrew E. Allen⁶, and E. Virginia Armbrust²,³

²School of Oceangraphy, University of Washington, Seattle, WA 98105; and ³Craig Venter Institute, San Diego, CA 92121

2012
INTRODUCTION

Spur new methods development

University Domain Research

Data Science Practice

learn, use, teach

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

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
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.
Campus Research Support
(The space between Office Hours and Grant Proposals)

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
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

Fruitful collaboration with potential for significant impact
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-Sheet Microscopy

Damage Speaks: Acoustical Monitoring Framework for Structures Subjected to Earthquakes
Beyond the MSDSE’s: Into the Community
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.

“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.”
Connect sidewalks

Use existing data to find the best route

Incline:

+ Sidewalk lines

Construction:

Permits

Curb ramps and crosswalks:
Beyond the MSDSE's

Impact in the Community (resonates with University Leadership)

UW student project taps ORCA cards, unlocks data trove

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

Budding UW Data Scientists Use Their Powers for Social Good

Student projects leapfrog governments and industry in ‘Data Science for Social Good’ program
Beyond the MSDSE’s

Scalable Research Impact: Community Learning Within Domains

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

OceanHackWeek 2019
Data Science + Oceanography
University of Washington
Aug. 26 - 30, 2019
(Started in 2018)

WaterHackWeek 2019
Workshop on Water Data Science
University of Washington EScience Institute
March 25-29, 2019
Kavli Institute for Cosmology @ Cambridge University in Cambridge, UK

Cryospheric Science with ICESAT-2 Hackweek 2020
Workshop on ICESAT-2 datasets for Cryospheric Studies
University of Washington
June 15-19, 2020
Application Deadline April 3, 2020
Exit Survey Responses: Research Methods

- I hacked on topics, tools, or methods that were very new to me.
- I believe that X Hack Week helped make me a better scientist.
- I feel like I learned things which improve my day-to-day research.
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 & Environmental Engineering

Hackweeks:
Huppenkothen et al, 2018 PNAS

Entrofy:
Huppenkothen et al, 2019
arXiv:1905.03314

Toolkit:
uwescience.github.io/HackWeek-Toolkit
Beyond the MSDSE’s

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.
Beyond the MSDSE’s

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

https://www.textxd.org/
Beyond the MSDSE’s

Key Takeaway

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

Graphic: Laura Norén ©2019 Obsidian Security
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
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
“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

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/
Thank you!

micaela@academicdatascience.org

www.academicdatascience.org
Early Career Support: The Data Science Co-Op
Beyond the MSDS's

Data, Responsibly
Beyond the MSDSE’s

Exit Survey Responses: Open Science

I am embarrassed to put my code & data online

I am afraid that if I put my code & data public, I will be scooped
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)