Academic Data Science, From Individuals to Institutions

Micaela Parker, Executive Director

*Academic Data Science Alliance*

April 2020 Data ONE webinar
INTRODUCTION

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.
A bit of my personal journey

(or: How I knew the system was broken)
Life before data science

(circa 1997)


MS and PhD in Oceanography (1999, 2004)

Research staff in a well-funded lab (2004-2014)

Internationally recognized researcher (2013)

Where do I go from here??

The pitfalls of a staff researcher job
These DATA are beyond me...
These DATA are beyond me...

Genomes and Transcriptomes of Lab Cultures

- The Marine Microbial Eukaryote Transcriptome Sequencing Project (marinemicroeukaryotes.org) aims to sequence 750 novel transcriptomes.
- 367 transcriptomes are currently available, including 47 diatom species. 31 were grown in our lab

Approximate Maximum Likelihood tree of the 18S rRNA gene from diatoms included in this study. Purple taxa are pennate diatoms, blue taxa are centric diatoms.

Metatranscriptomes from an Ocean Transect

GeoMICS: Global scale Microbial Interactions across Chemical Surveys

Reference gene trees from lab sequences are used to recruit reads from the environmental metatranscriptomes. Orange branches indicate greater abundance of reads at coastal P1 relative to off-shore P8 and Blue-green branches indicate greater abundance of reads at P8 relative to P1.

- May, 2012: GeoMICS is launched with a 1 week cruise on the R/V Thompson along a subset of Line P (stations P1 – P8, shown above).
- Metatranscriptomes have been collected from stations P1, P4, P6, P8; P1 and P8 (in yellow) have been analyzed
- Goal: microbial biogeography and ocean chemistry across a persistent oceanographic "hot spot" in the NE Pacific Ocean (Ribalet et al. 2010)
- Multiple biological and chemical parameters were collected. An iron gradient was observed with an order of magnitude difference in concentration between P1 and P8
Metalloenzymes Switch Types in Response to Metal Availability

Fe (centrics only)

Mn

Fe/Mn-type superoxide dismutase gene tree showing one Fe clade and two Mn clades. Thus far, Fe-types have only been detected in centric diatoms. The two Mn clades recruit reads differently from the two stations.

Approximate Maximum Likelihood tree of the 18S rRNA gene from diatoms included in this study. Purple taxa are pennate diatoms, blue taxa are centric diatoms.
The power of the buffet line

First All Campus Data Science Poster Session
@UW 2014
137 posters, 30+ departments
Is it time for a Career Change?
It’s ok to ask for Work/Life Balance

Job share proposal that includes:
• how it will work
• why it will benefit the organization

Sarah Stone, job share partner
met in Antarctica
It’s ok to ask for Work/Life Balance

Sarah Stone, job share partner met in Antarctica

First job-shared position in management role in UW’s history
It’s ok to ask for Work/Life Balance

Sarah Stone, job share partner met in Antarctica

First job-shared position in management role in UW’s history

CAREER PROFILES Options and Insights

SARAH A. STONE and MICAE LA S. PARKER | Program Managers, eScience Institute, University of Washington, Seattle, WA, manager@esience.washington.edu

Sarah Stone and Micaela Parker job share a program manager position for the eScience Institute at the University of Washington (UW). This position is the first management job share at UW. Because of this unique position and their shared experiences, they thought their journeys could be best described with a joint career profile for Oceanography. The profile moves back and forth between their paths, which have both similar and divergent features. Their discussion follows two themes that were pivotal in structuring both of their careers: role models and family balance. The two scientists hope their stories will inspire more women in science to push for administrative policy changes that increase job flexibility and awareness of the challenges faced by women in caregiver roles.

The Oceanography Society journal
Back to the point of this talk...

Integrating Data Science into Academia
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

Spur new methods development

University Domain Research

BUILD BRIDGES

Data Science Practice

Enable data-driven discovery
INTRODUCTION

Spur new methods development

University Domain Research

Data Science Practice

learn, use, teach

Enable data-driven discovery
Building Bridges: Our Efforts Organized into Working Groups

Scientific Theme Areas
- Biological Sciences
- Environmental Sciences
- Physical Sciences
- Social Sciences

Data Science Studies
- Career Paths and Alternative Metrics
- Education and Training
- Software Tools, Environments, and Support
- Reproducibility and Open Science
- Working Spaces and Culture
- Ethnography and Evaluation

New Data Science Methodologies Transform Discovery

Bridges
- Discovery Spurs New Data Science Methodologies

Data Science Methodologies
- Machine Learning
- Data Management
- Data Visualization / Usability
- Statistics
- Sensors
- Programming Environments
- Scalable Hardware & Software Systems
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
MSDSE’s

Software meets Education

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

Cloud-Enabled Tools for the Analysis of Subsea HD Camera 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
Brings together students and researchers with data science and domain expertise to work on focused, collaborative projects for societal benefit.
Data, Responsibly
DSSG: Impact in the Community

The Seattle Times

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
Extending Partnerships: Beyond the MSDSEs
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
Hackweeks: Growth and Evolution
Hackweeks: Growth and Evolution

**OceanHackweek 2019**
Data Science + OceanoGraphy
University of Washington
Aug. 26 - 30, 2019
(Started in 2018)

**AstroHackweek 2019**
Workshop on Water Data Science
University of Washington EsScience 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

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

- Astro Hack Week
- Geo Hack Week
- Neuro Hack Week
Hackweek Leaders and Resources

Hackweeks: Huppenkothen et al, 2018 PNAS


Toolkit: Arendt & Huppenkothen
uwescience.github.io/HackWeek-Toolkit
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.
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/
Key Takeaway

Informal intensive community-driven learning opportunities, like Hackweeks and xD workshops, quickly and effectively bring data science to campus researchers.
Challenges in the Data Science Community
“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)
Challenges

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
Challenges

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
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

Which disciplines make up academic data science in 2019?

- CS, Math, Stats, Physics, Engineering: 54%
- Social sciences, humanities: 17%
- Natural Sciences: 16%
- Business, Econ.: 8%
- MDs, RNs: 4.5%
- JDs, LLMs: 1.5%

www.obsidiansecurity.com
ADSA Activities
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
Transition MSDSE Summit to ADSA Annual Meeting

Opportunity for data savvy researchers to share and learn tools and methods outside their domain.
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:**
- Ethics
ADSA’s Career Development Network

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/
COVID-19 Data and Data Resources Page

https://www.academicdatascience.org/covid
Sign-up for our Quarterly

info@academicdatascience.org
Thank you!

micaela@academicdatascience.org

www.academicdatascience.org


@AcademicDataSci