

Academic
Data Science
Alliance

Academic Data Science, From Individuals to Institutions

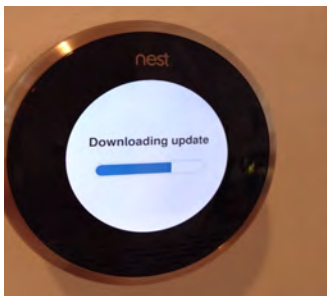
Micaela Parker, Executive Director
Academic Data Science Alliance

April 2020



Data ONE webinar

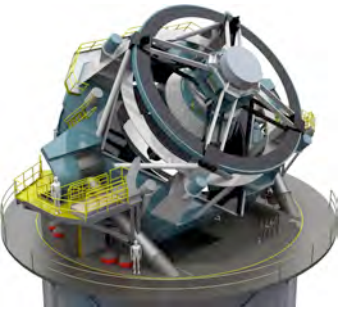
Data are being collected and used everywhere!



- 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”



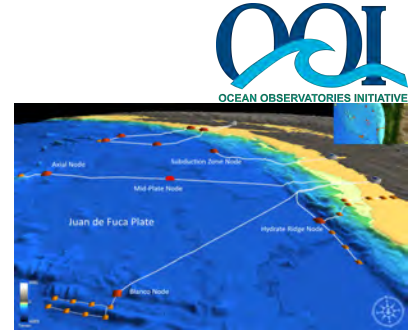
Astronomy: LSST



Physics: LHC



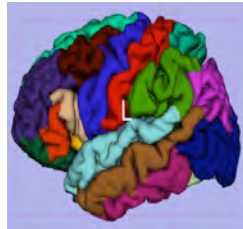
Digital Humanities



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



A bit of my personal journey

(*or*: How I knew the system was broken)



Life before data science



(circa 1997)



New mom (2002 & 2004)

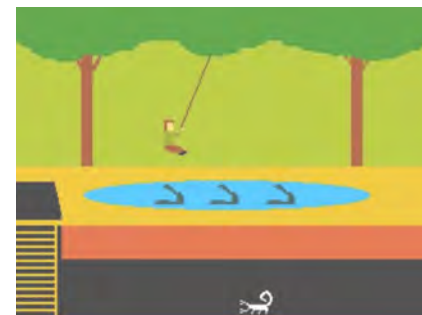


Research staff in a well-funded lab (2004-2014)



Internationally recognized researcher (2013)

Where do I go from here??

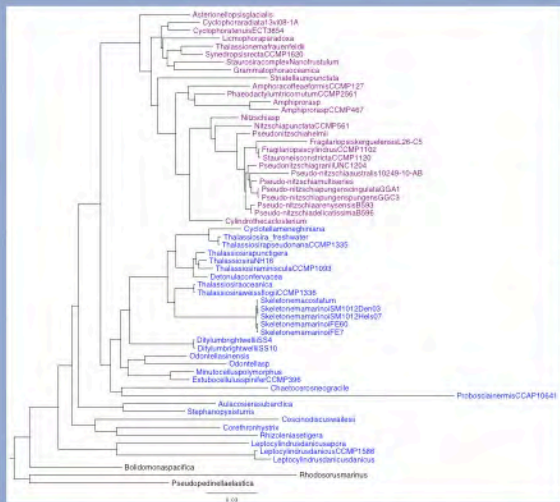


The pitfalls of a staff researcher job



These DATA are beyond me

Genomes and Transcriptomes of Lab Cultures



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 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

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.



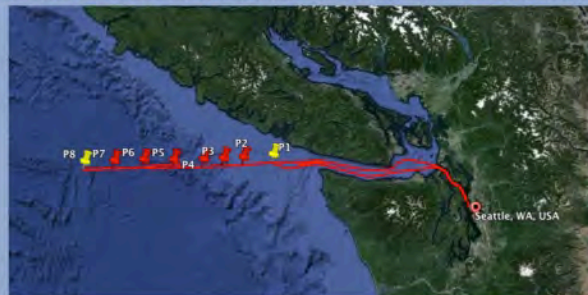
Micaela S. Parker, Ryan Groussman, E. Virginia Armbrust and the GeoMICS Consortium



The Importance of Diatoms and Iron

Metatranscriptomes from an Ocean Transect

GeoMICS: Global scale Microbial Interactions across Chemical Surveys



← Less Iron → More Iron

- 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



Mn

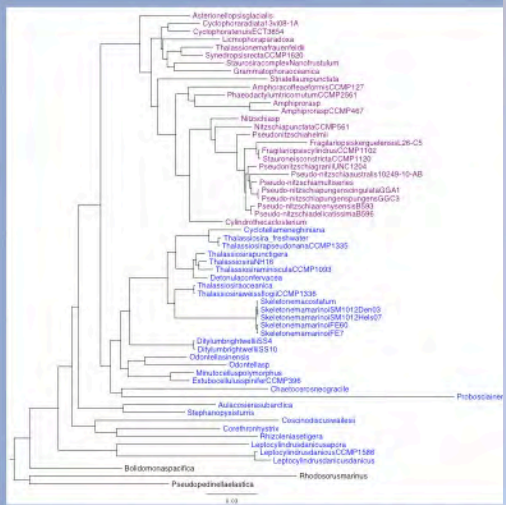
iron-responsive Clade II

Abstract: (2007) ...

Abstract: (2008) ...

Abstract: (2010) ...

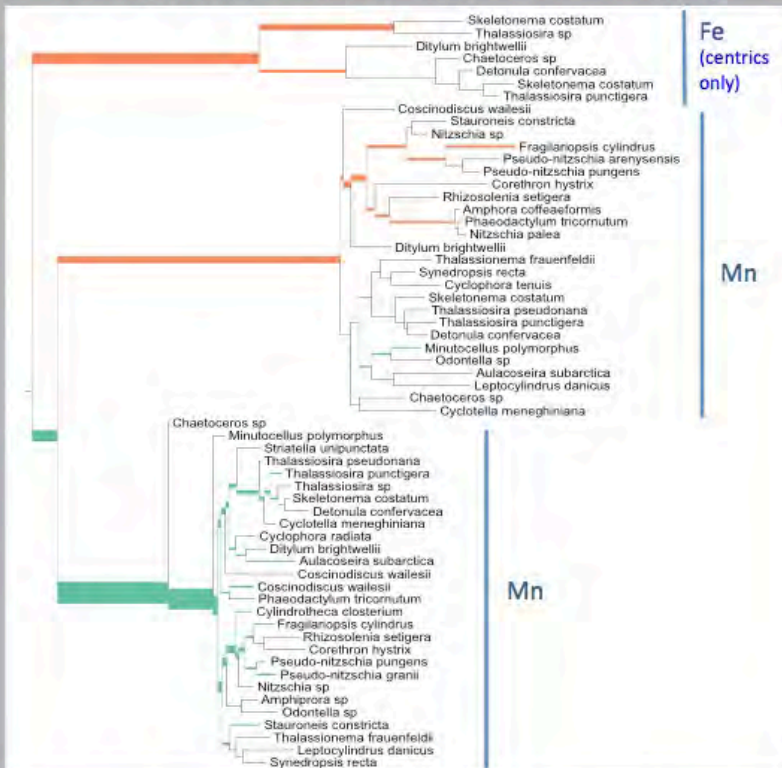
These DATA are Genomes and Transcriptomes



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.



Metalloenzymes Switch Types in Response to Metal Availability



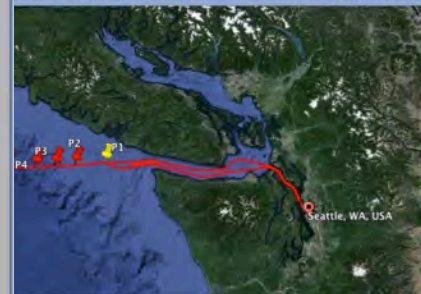
Fe/Mn-type superoxide dismutase gene tree showing one Fe clade and two Mn clades. Thus far, Fe-types have only been detected in centric diatoms. The two Mn clades recruit reads differently from the two stations.



Importance of Diatoms and Iron

Genomes from an Ocean Transect

Microbial Interactions across Chemical Surveys



More Iron

with a 1 week cruise on the R/V Thompson along a subset of Line P

lected from stations P1, P4, P6 P8; P1 and P8 (in yellow) have been

ocean chemistry across a persistent oceanographic "hot spot" in the NE

parameters were collected. An iron gradient was observed with "an order interval between P1 and P8

Iron-responsive Clade II

Abstract: (2007) *Iron-responsive superoxide dismutase* (Fe-SOD) is a key enzyme in the Fe cycle of marine diatoms. We have identified a novel Fe-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Fe-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context. We have also identified a novel Mn-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Mn-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context. We have also identified a novel Fe-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Fe-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context. We have also identified a novel Mn-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Mn-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context.

Methods: We used a combination of transcriptome and genome data to identify and characterize these genes. We used a combination of transcriptome and genome data to identify and characterize these genes. We used a combination of transcriptome and genome data to identify and characterize these genes.

Results: We identified a novel Fe-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Fe-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context. We have also identified a novel Mn-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Mn-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context.

Conclusions: We have identified a novel Fe-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Fe-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context. We have also identified a novel Mn-SOD gene in a pennate diatom, *Chaetoceros* sp. This gene is highly similar to the Mn-SOD gene in the centric diatom *Skeletonema costatum*, but it is located in a different genomic context.

The power of the buffet line

First All Campus Data Science Poster Session
@UW 2014
137 posters, 30+ departments



UNIVERSITY of WASHINGTON
eScience Institute



Is it time for a Career Change?



UNIVERSITY of WASHINGTON
eScience Institute



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



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

CAREER PROFILES Options and Insights

Sarah Stone, job share partner
met in Antarctica



SARAH A. STONE and MICAELA S. PARKER | Program Managers,
eScience Institute, University of Washington, Seattle, WA,
manager@escience.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.



Back to the point of this talk...

**Integrating Data Science into
Academia**



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



University
Domain
Research

BUILD BRIDGES

*Spur new methods
development*

Data
Science
Practice

*Enable data-
driven discovery*



University
Domain
Research



learn, use, teach

*Spur new methods
development*

Data
Science
Practice

*Enable data-
driven discovery*





Micaela Parker

eScience Program Manager -> eScience Executive Director -> +MSDSE Program Coordinator

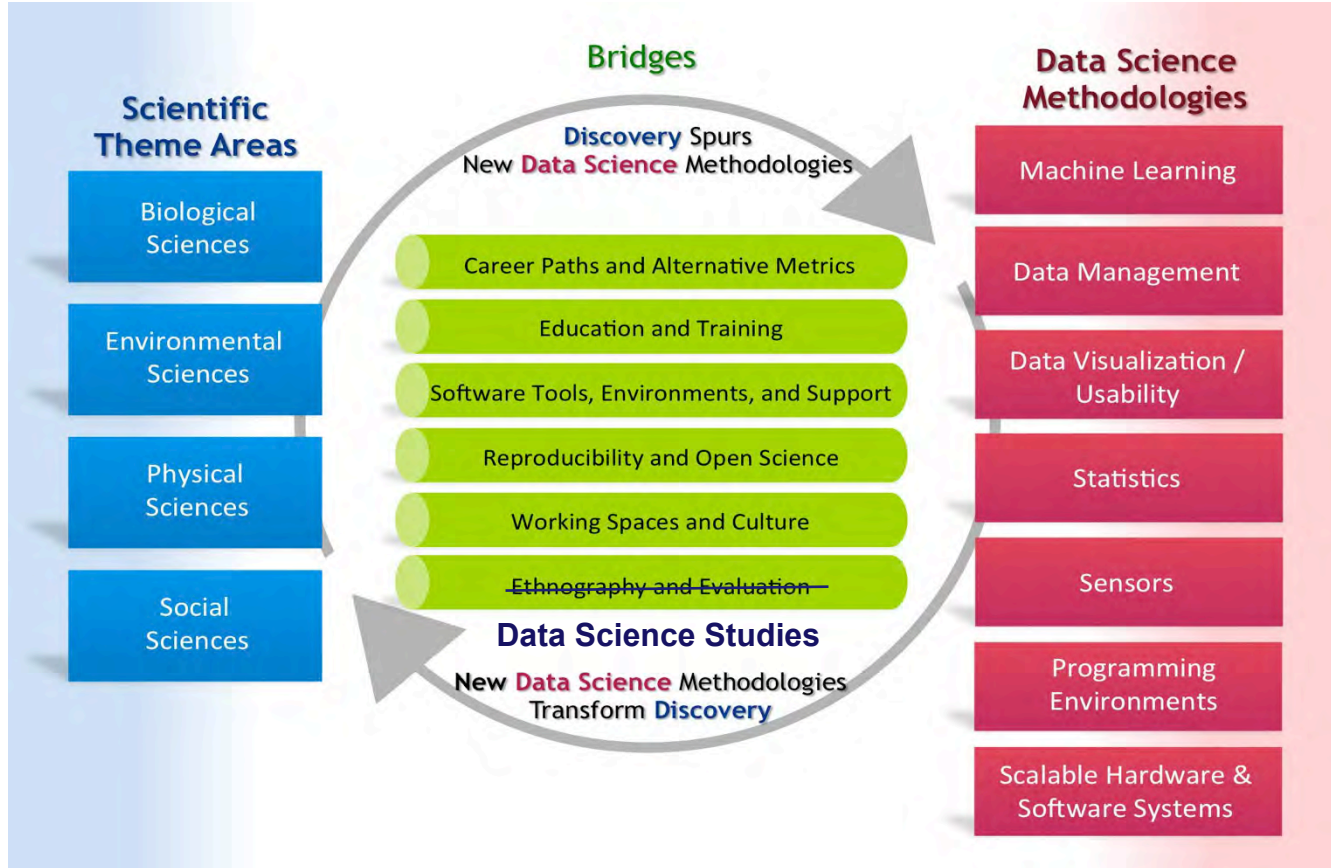
Chris Mentzel, Gordon and Betty Moore Foundation

Josh Greenberg, Alfred P. Sloan Foundation



DataONE Webinar - April 2020

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)

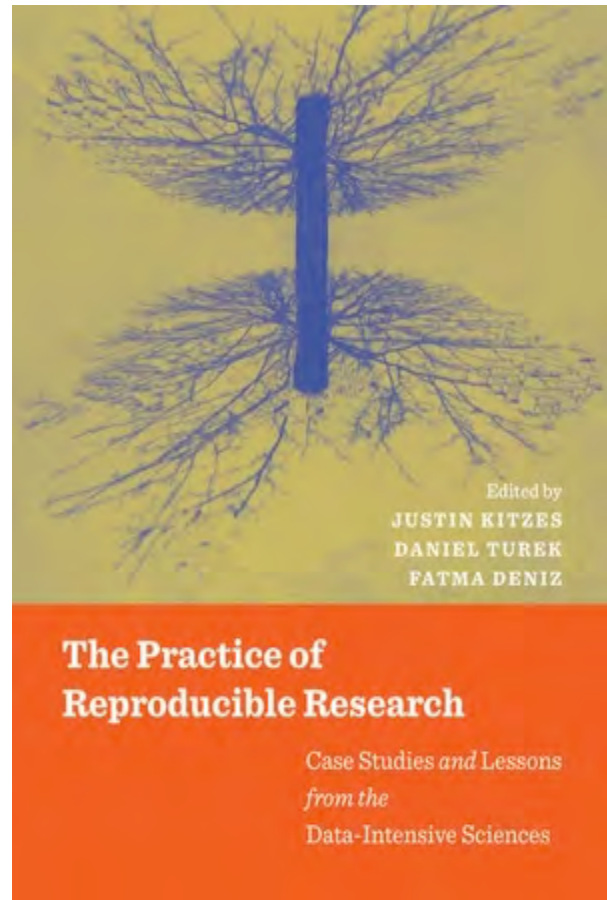


- **ReproZip**: pack your research along with all data files, libraries, environment variables and options. Anyone can reproduce the research on a different machine



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



Software meets Education



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

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

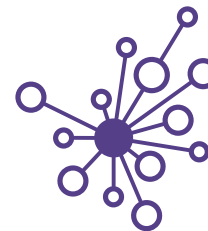
JupyterHub:

- 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)



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





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



Example Projects from the Winter Incubator

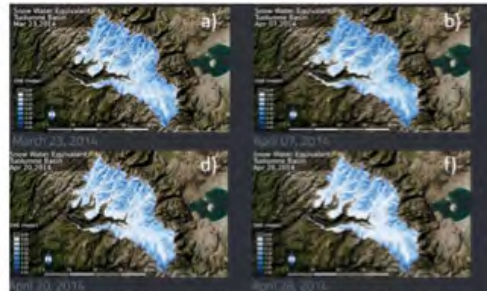


Cloud-Enabled Tools for the Analysis of Subsea HD Camera Data

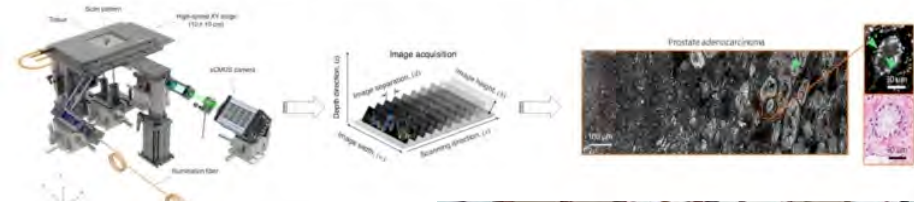
Simulating Competition in the U.S. Airline Industry



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

Education | Education Lab | Local News | Transportation

UW student project taps ORCA cards, unlocks data trove

GeekWire NEWS ▾ JOBS EVENTS ▾ RESOURCES ▾ DEALS ABOUT ▾ f t r

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 | f Share 68 | t Tweet | Share 43 | Reddit | Email

GeekWire Gala early-bird tix on sale now!

GeekWire NEWS ▾ JOBS EVENTS ▾ RESOURCES ▾ DEALS ABOUT ▾ f t r

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 Experience 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

f t r



ADVERTISEMENT

Enter your forecasts for

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



DataONE Webinar - April 2020



UNIVERSITY of WASHINGTON
eScience Institute

 AstroData Hack Week

University of Washington

September 15-19, 2014

GEOHACKWEEK

WORKSHOP ON GEOSPATIAL DATA SCIENCE

NEUROHACKWEEK

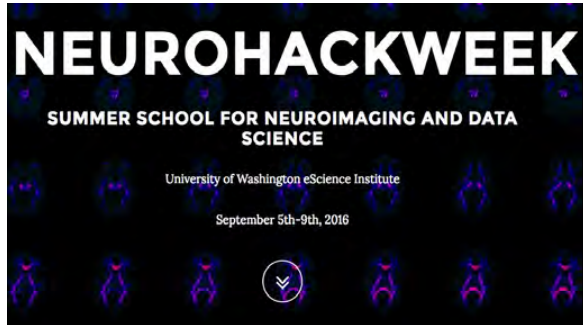
SUMMER SCHOOL FOR NEUROIMAGING AND DATA SCIENCE

University of Washington eScience Institute

September 5th-9th, 2016



Hackweeks: Growth and Evolution



Hackweeks: Growth and Evolution

OCEANHACKWEEK 2019

DATA SCIENCE + OCEANOGRAPHY
UNIVERSITY OF WASHINGTON
AUG. 26 - 30, 2019

(Started in 2018)

ASTRO HACK WEEK 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

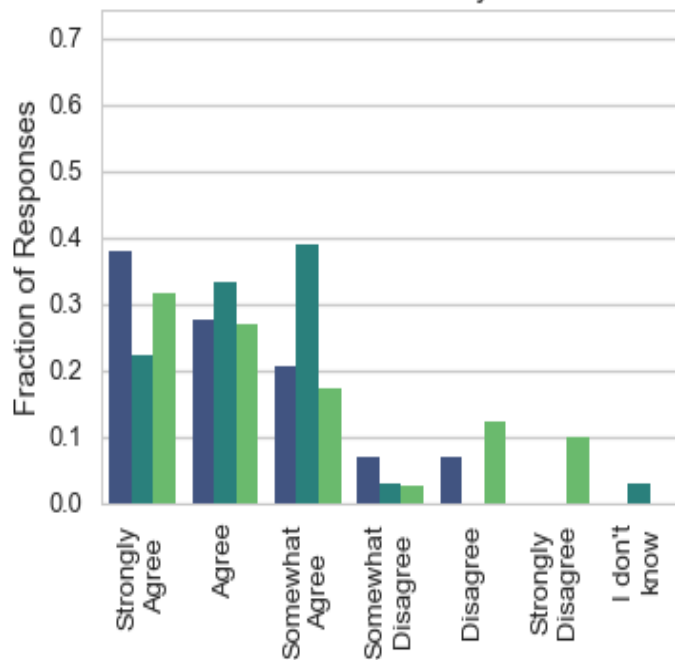
CRYOSPHERIC SCIENCE WITH ICESAT-2 HACKWEEK 2020

WORKSHOP ON ICESAT-2 DATASETS FOR CRYOSPHERIC STUDIES
UNIVERSITY OF WASHINGTON
JUNE 15-19, 2020
APPLICATION DEADLINE APRIL 3, 2020

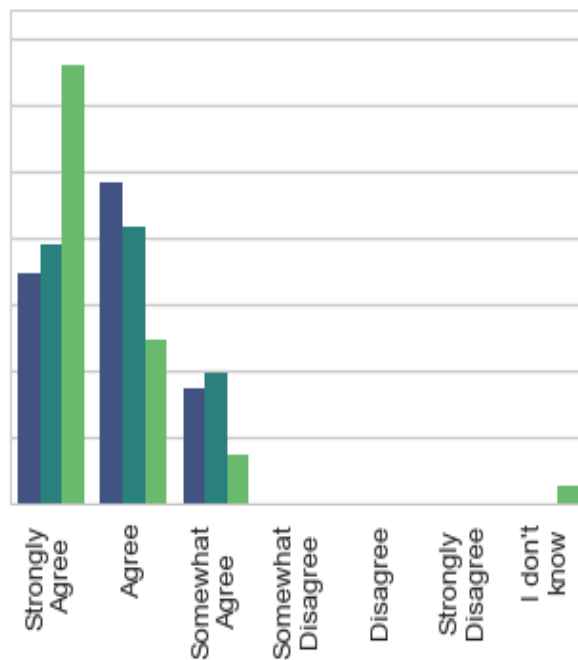


Exit Survey Responses: Research Methods

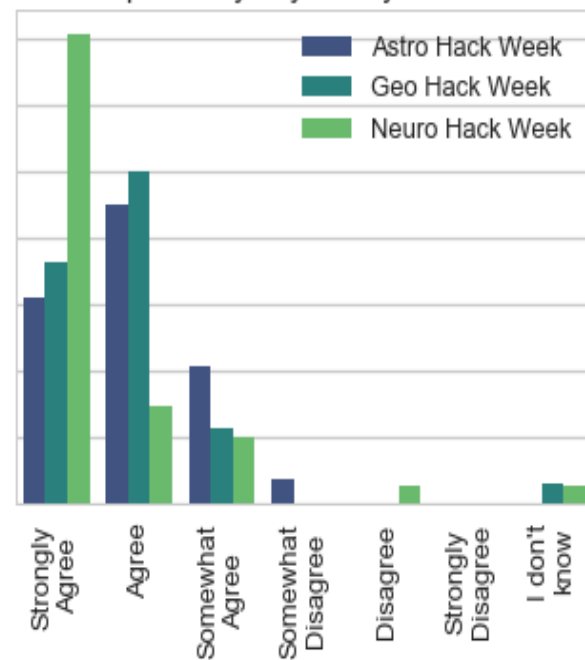
I hacked on topics, tools, or methods that were very new to me.



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

Hackweeks:

Huppenkothen et al, 2018 PNAS

Entropy:

Huppenkothen et al, 2019 arXiv: 1905.03314

Toolkit:

Arendt & Huppenkothen

uwescience.github.io/HackWeek-Toolkit



Anthony Arendt
Senior Research Scientist,
Polar Science Center, UW



Karthik Ram
Senior Data Scientist, UCB



Jake VanderPlas
Senior Data Science Fellow, UW

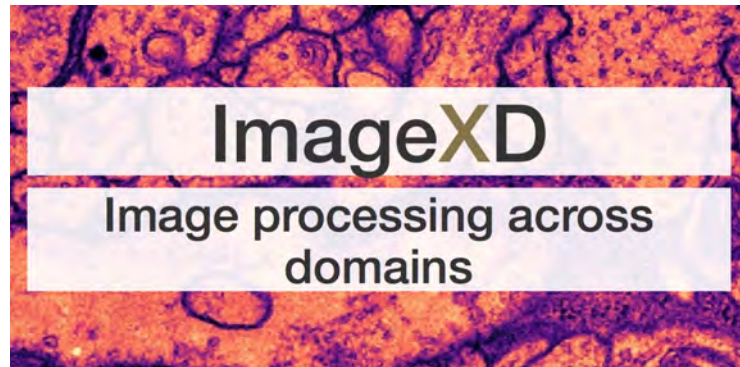


Christina Bandaragoda
Research Scientist, Civil & Environmental Engineering

Community Learning Across Domains

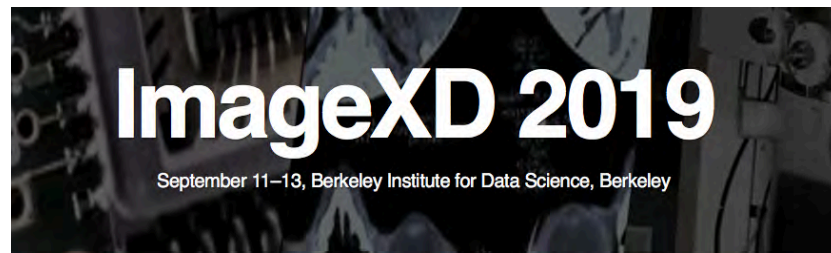
XD Working Groups & Workshops

- 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/>



GraphXD
Graphs Across Domains



Key Takeaway

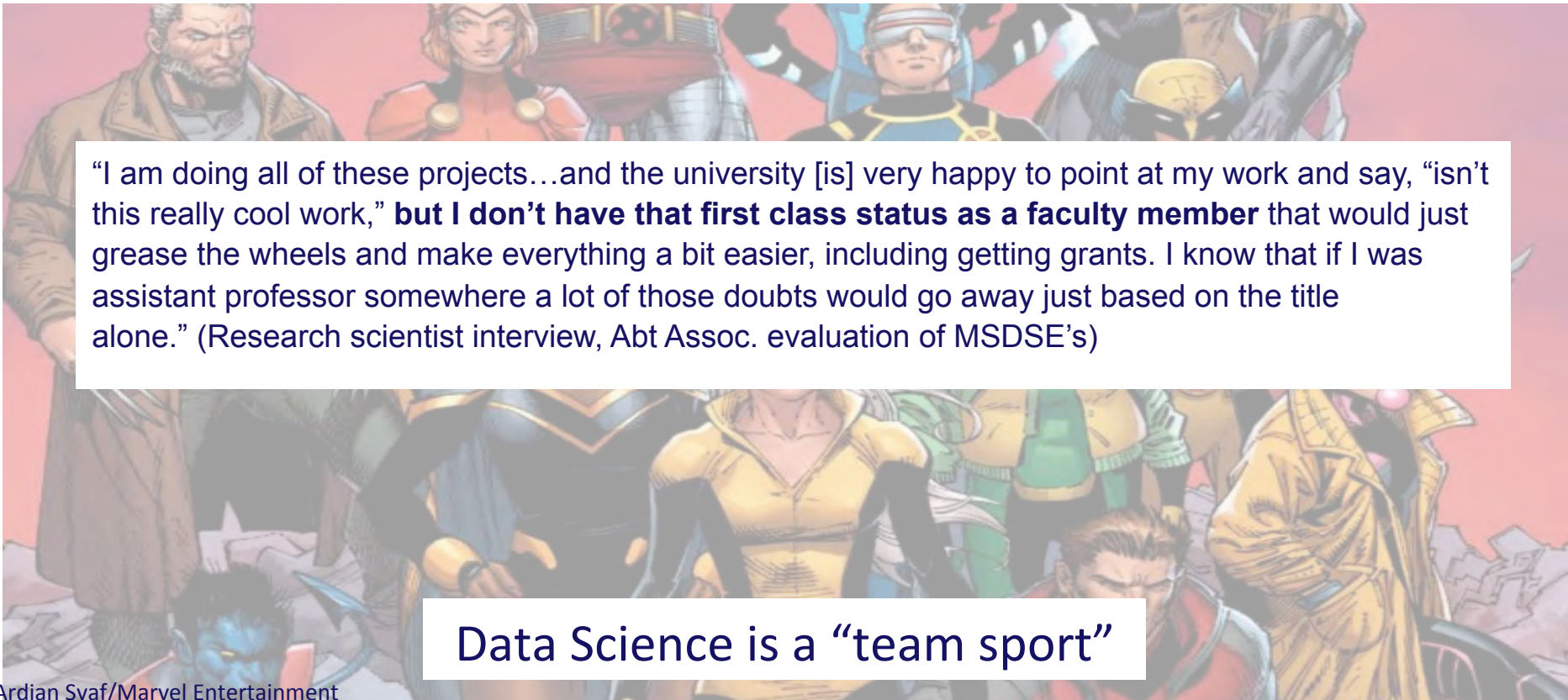
Informal intensive community-driven learning opportunities, like Hackweeks and xD workshops, quickly and effectively bring data science to campus researchers.



Challenges in the Data Science Community



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



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”**



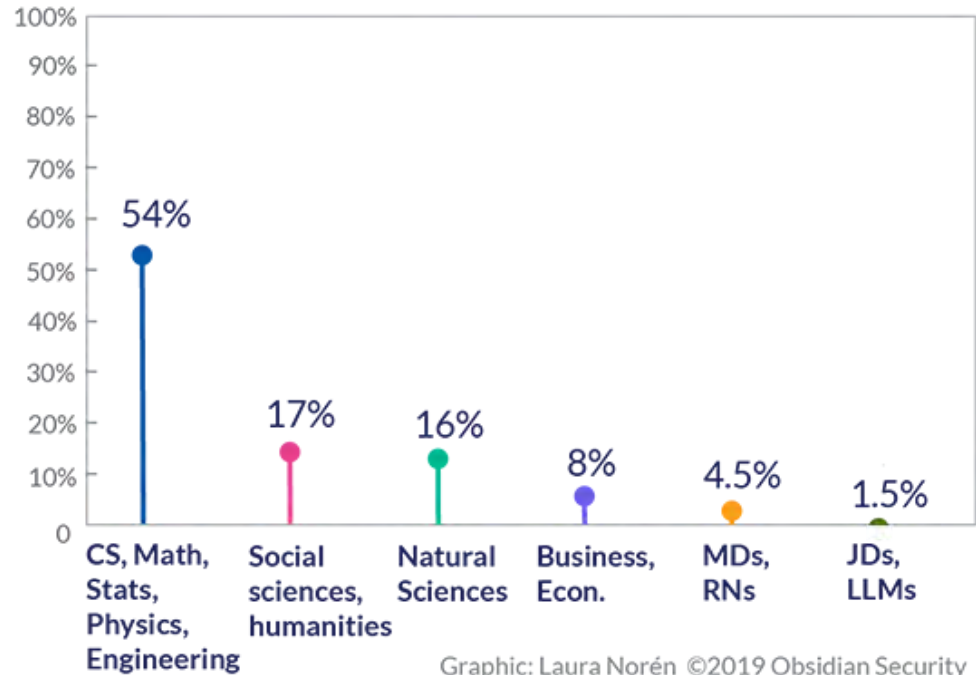
Who's Building Your AI? A Research Brief

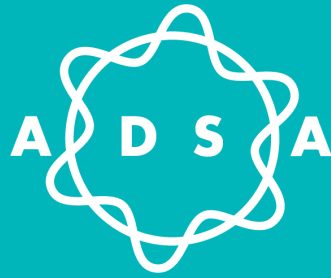
by Laura Noren, Gina Helfrich, and Steph Yeo

- ~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

www.obsidiansecurity.com

Which disciplines make up academic data science in 2019?



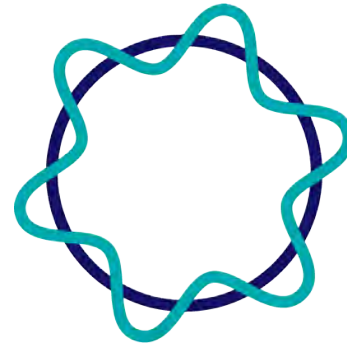


ADSA Activities



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



**Academic
Data Science
Alliance**



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

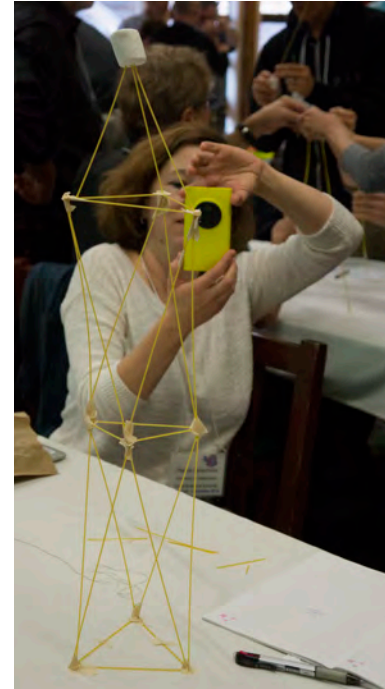
bring together thought leaders in our community
to tackle pressing challenges throughout the year

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**

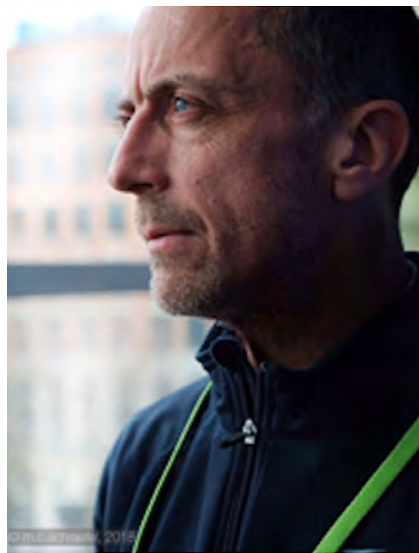


Data Science Community Newsletter

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/>



COVID-19 Data and Data Resources Page

<https://www.academicdatascience.org/covid>

Datasets

Analytic Tools

Academic Research Article Collections

Events and Conversations

Challenges



Funding Opportunities

Data Visualizations

Computing Resources

Research Tracking

Support Networks

Other Collections of Resources



Sign-up for our Quarterly

info@academicdatascience.org



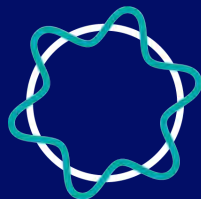
**Academic
Data Science
Alliance**

Welcome to our first ADSA Quarterly!

April 2020

Here you will find updates on the activities of the Academic Data Science Alliance, event reminders, and some guest spots for shout-outs in our community. Enjoy!





Academic
Data Science
Alliance

Thank you!



micaela@academicdatascience.org

www.academicdatascience.org



<https://adsa-slack-auto-invite.herokuapp.com/>



@AcademicDataSci