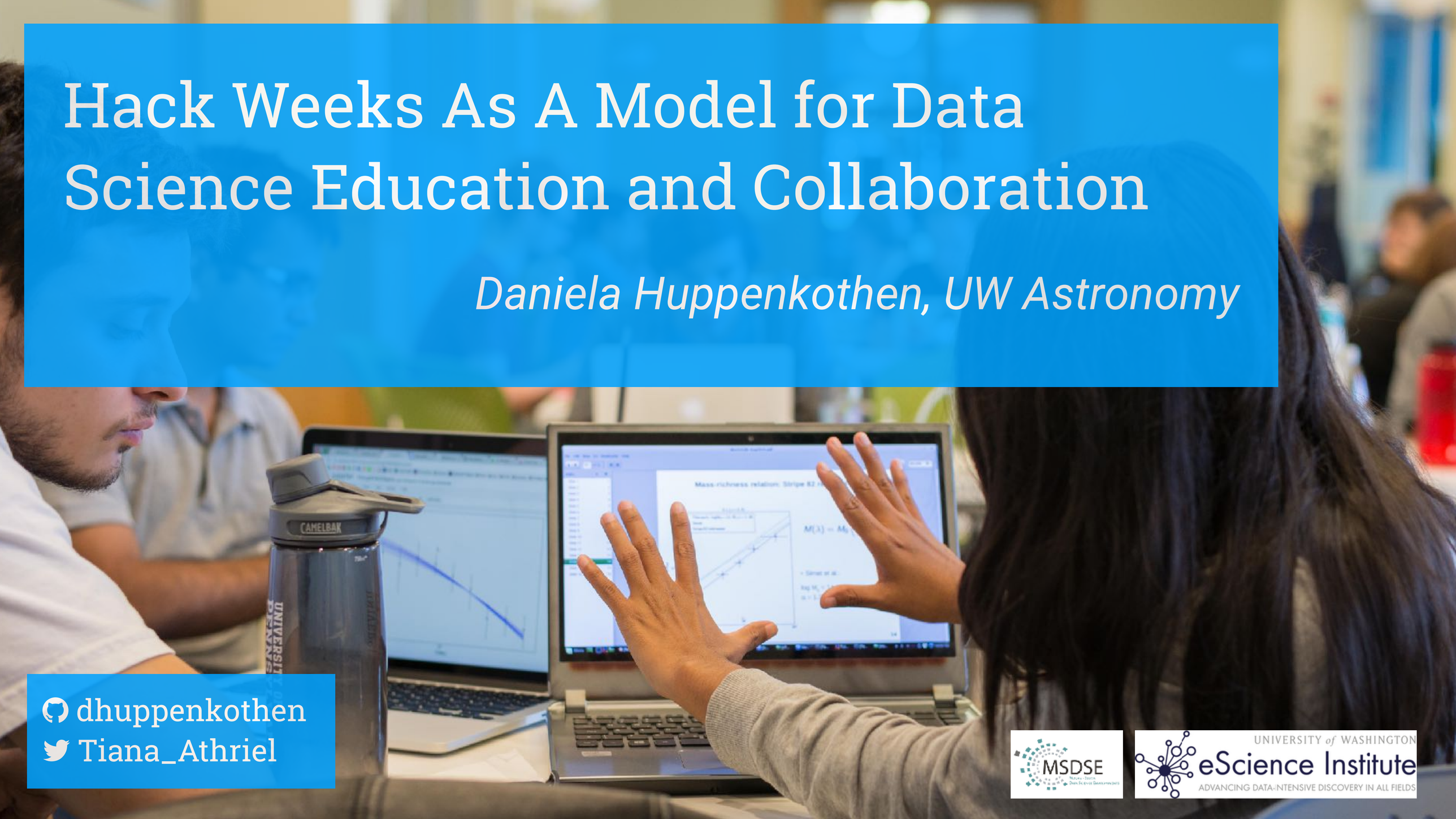
Hack Weeks As A Model for Data Science Education and Collaboration

Daniela Huppenkothen, UW Astronomy



dhuppenkothen
Tiana_Athriel

MSDSE

UNIVERSITY of WASHINGTON eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

American Astronomical Society (2018)

credit: AAS/CorporateEventImages/Phil McCarten

“The best thing about this meeting is the coffee breaks!”

“The best thing about this meeting is the coffee breaks!”

- exchange ideas
- collaboration
- networking


Can we organize a workshop
that is all coffee breaks?
Lorentz
center

- How do we improve the exchange of knowledge?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?

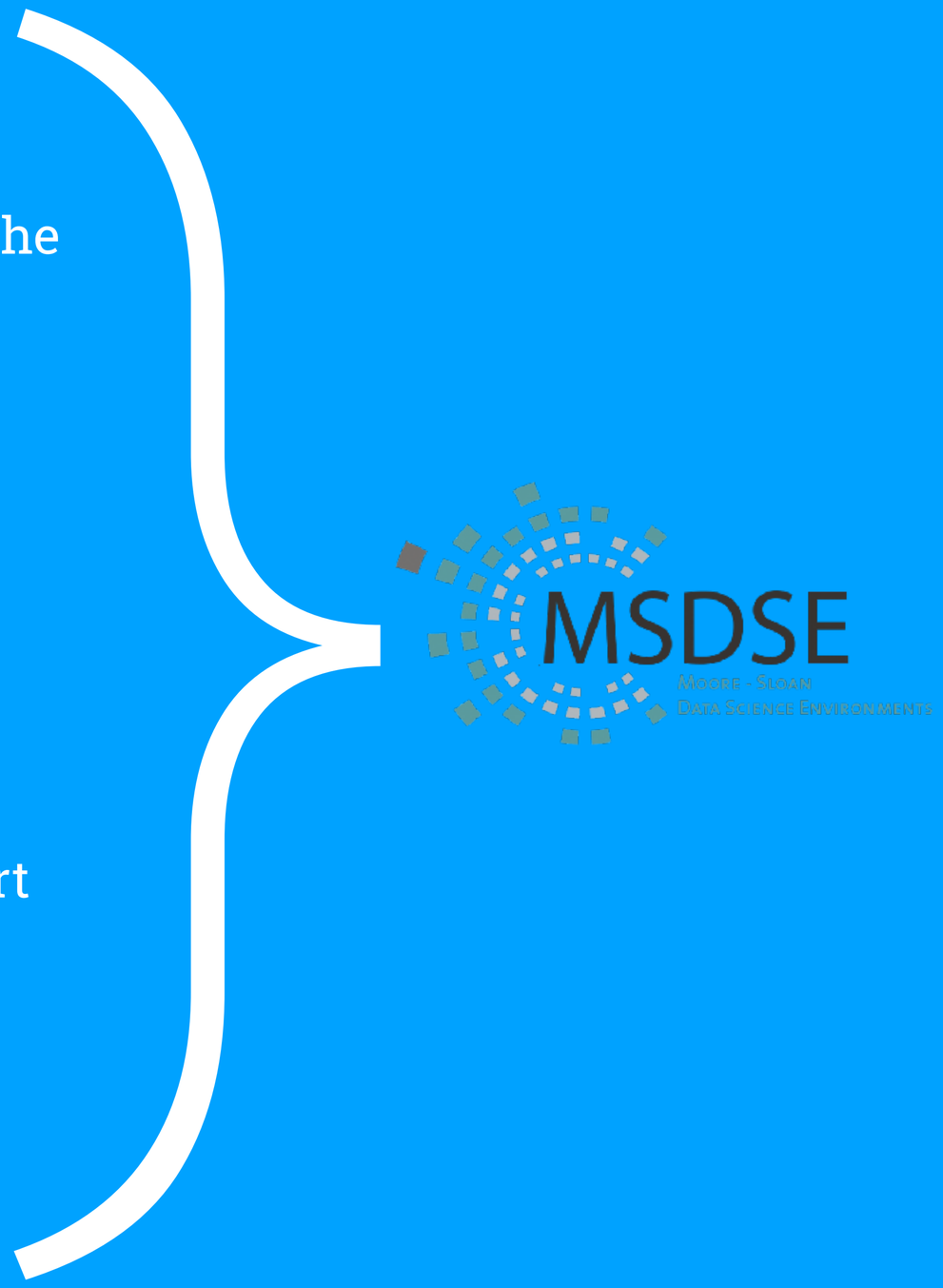
- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?
- How do we facilitate collaborations within and across fields?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?
- How do we facilitate collaborations within and across fields?
- How do we facilitate exchange of ideas, knowledge and people between academia and industry?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?
- How do we facilitate collaborations within and across fields?
- How do we facilitate exchange of ideas, knowledge and people between academia and industry?
- How do we enable researchers to brainstorm and prototype practical solutions to their problems on a short timescale?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?
- How do we facilitate collaborations within and across fields?
- How do we facilitate exchange of ideas, knowledge and people between academia and industry?
- How do we enable researchers to brainstorm and prototype practical solutions to their problems on a short timescale?
- How do we enable participation of a diverse range of researchers and make data science welcoming and inclusive?

- How do we improve the exchange of knowledge?
- How do we remove barriers and stop fields reinventing the wheel?
- How do we teach data science to domain scientists?
- How do we facilitate collaborations within and across fields?
- How do we facilitate exchange of ideas, knowledge and people between academia and industry?
- How do we enable researchers to brainstorm and prototype practical solutions to their problems on a short timescale?
- How do we enable participation of a diverse range of researchers and make data science welcoming and inclusive?



http://astrohackweek.org

http://astrohackweek.org

Jake VanderPlas

What is a hack week?

#AstroHackWeek

#AstroHackWeek

- 5-day workshop

#AstroHackWeek

- 5-day workshop
- ~50 participants

#AstroHackWeek

- 5-day workshop
- ~50 participants
- tutorials and break-out sessions

#AstroHackWeek

- 5-day workshop
- ~50 participants
- tutorials and break-out sessions
- project work

#AstroHackWeek

- 5-day workshop
- ~50 participants
- tutorials and break-out sessions
- project work
- Lots of ☕ and 🍪

#AstroHackWeek

- 5-day workshop
- ~50 participants
- tutorials and break-out sessions
- project work
- Lots of ☕ and 🍪
- participant-driven

#AstroHackWeek

- 5-day workshop
- ~50 participants
- tutorials and break-out sessions
- project work
- Lots of ☕ and 🍪
- participant-driven
- experimental

Hackweek Mission

Collaboration

Education

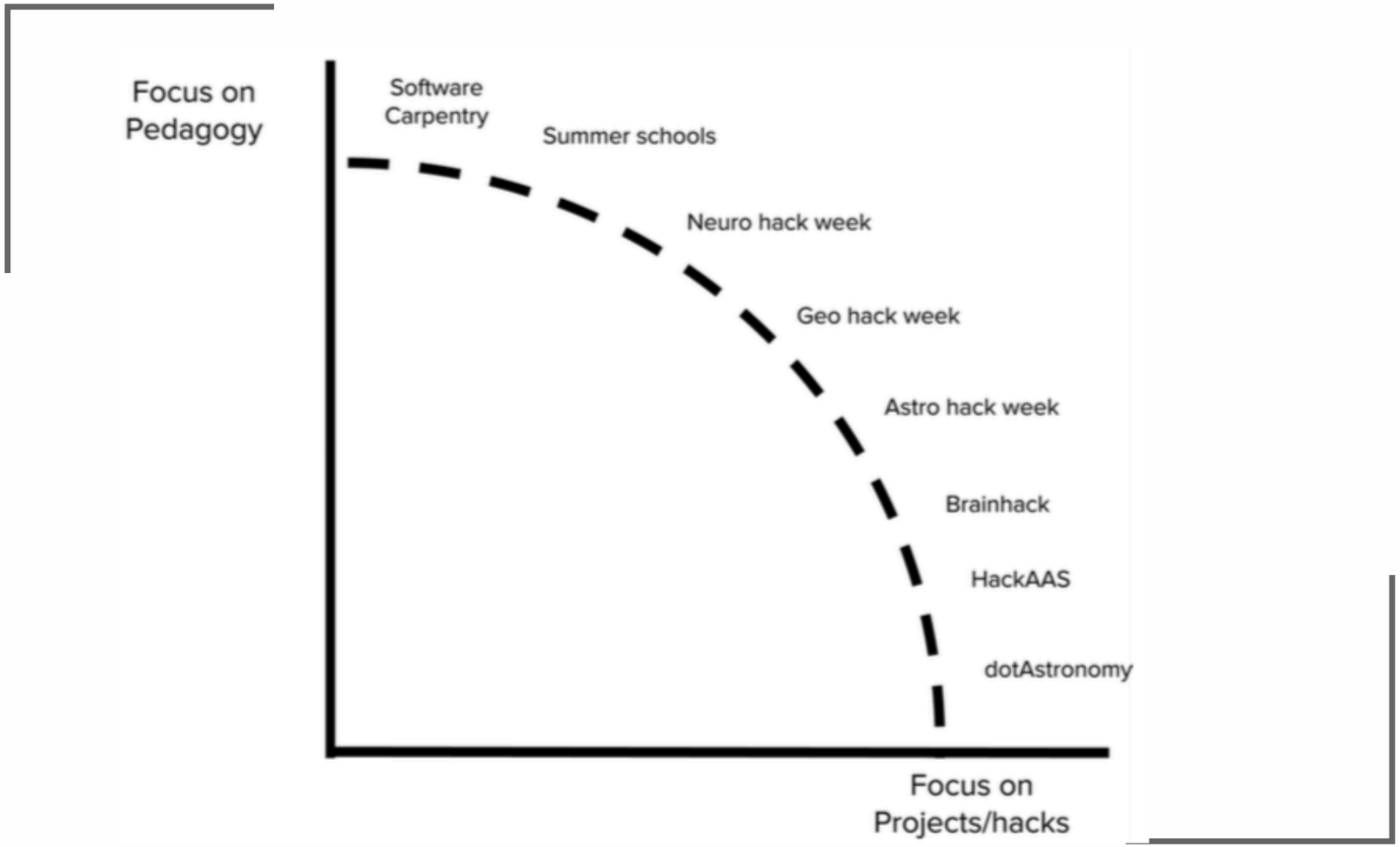
Building Software

Networking

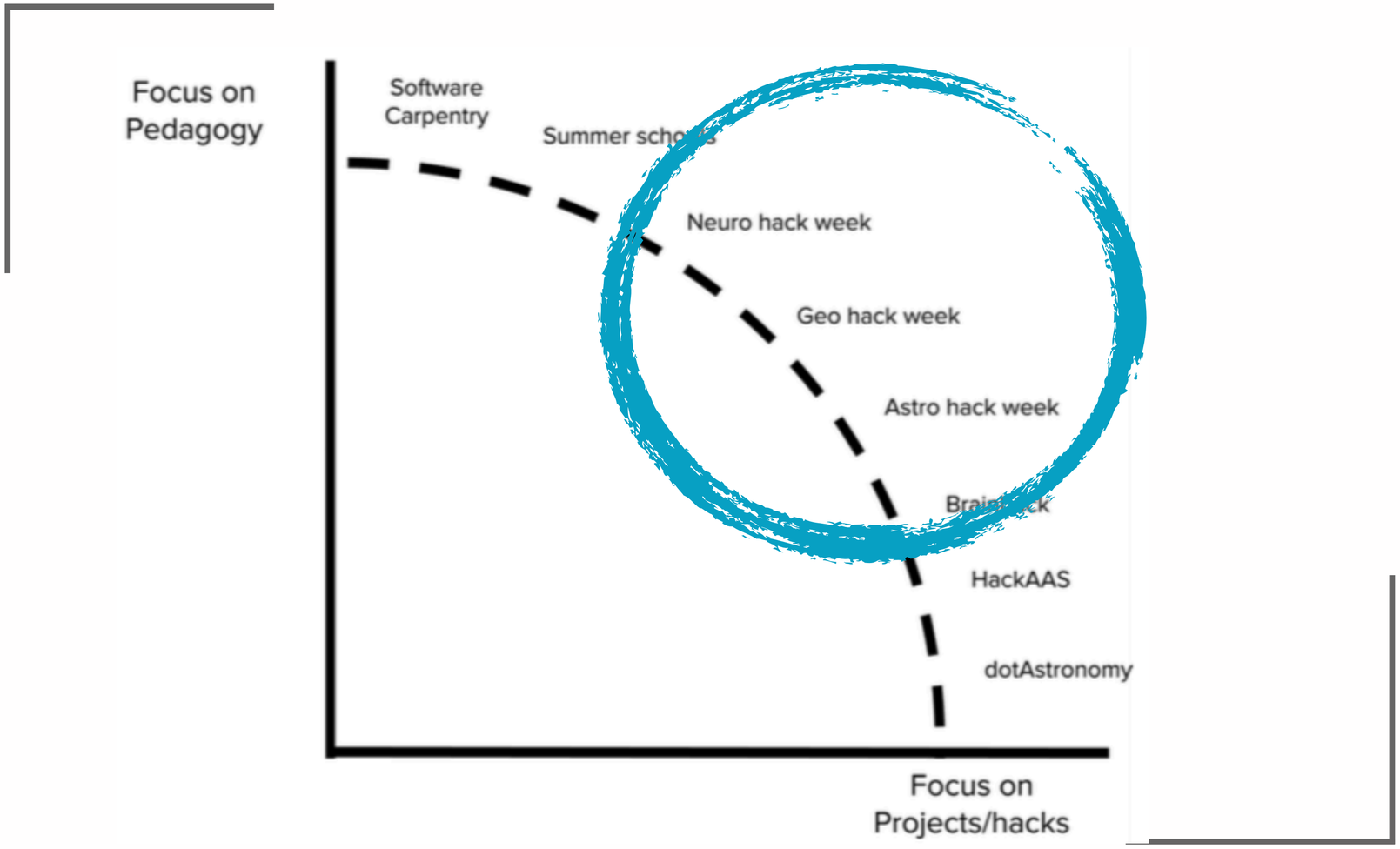
Projects

Community

credit: Anthony Arendt



http://www.pnas.org/content/early/2018/08/17/1717196115

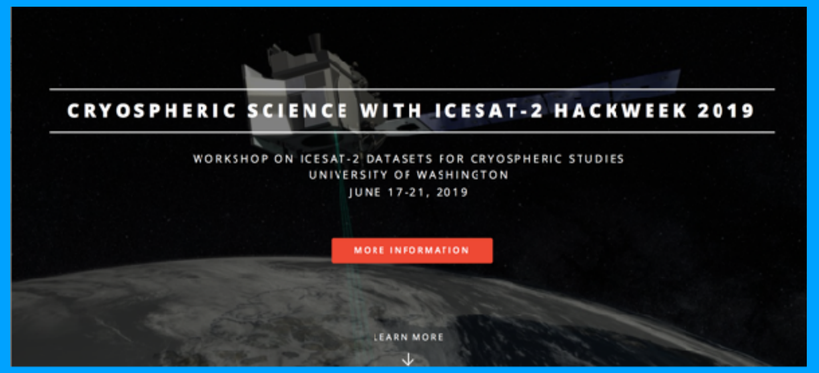


http://www.pnas.org/content/early/2018/08/17/1717196115

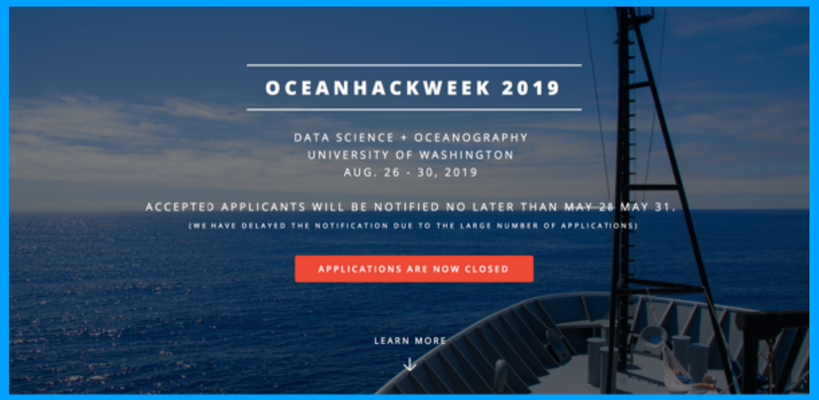
A toolbox for organizing interactive, collaborative events

https://geohackweek.github.io

https://neurohackweek.github.io



https://icesat-2hackweek.github.io



https://oceanhackweek.github.io

Hack Week

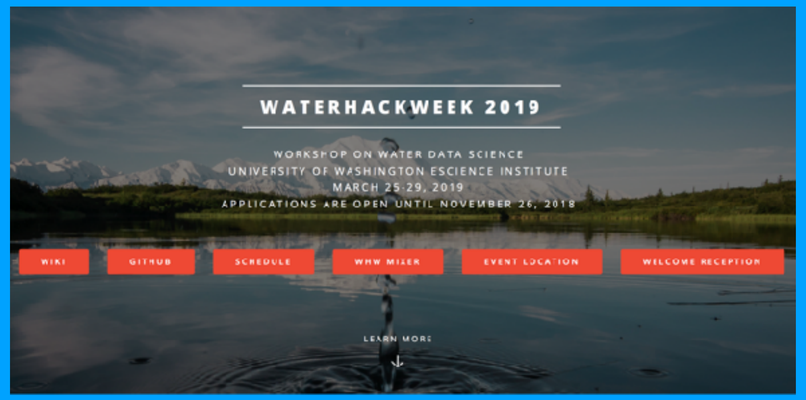
ECS Data Sciences Hack Week
May 14-19, 2018
Seattle, WA

Application Deadline: March 30, 2018

Building on the success of the first ECS Data Sciences Hack Day (October 2017), ECS is pleased to offer another opportunity at the ECS spring meeting in Seattle. In May 2018, the program will be expanded to an entire week as the next stage in ECS supporting a growing electrochemical data science and open source community. The goal of this event is to increase awareness and impact of data science tools, open source software, and shared datasets in electrochemistry and solid state science and technology, by bringing together people from different backgrounds to collaborate.

Hack Week will again be led by the very capable and engaging team from University of Washington: Dan Schwartz, David Beck, and Matt Murbach. The program will kick off on Monday, May 14 and have sessions all day Wednesday through Friday, as well as optional software training tutorials during the week. The activities will culminate with project presentations and an optional clamming expedition on Saturday, a traditional activity in the Puget Sound area.

https://www.electrochem.org/233/hack-week



https://waterhackweek.github.io

Participant-driven workshops

diversity is excellence

participant-driven != unstructured

participant-driven != unstructured

- design with the most vulnerable participants in mind
- facilitate carefully
- mitigate Impostor Phenomenon

> “At a summer school, the young learn from the old.
> At a hack week, the old learn from the young.”
>
> — David W. Hogg

"At a summer school, the young learn from the old.
At a hack week, the everyone learns from everyone else."

— Daniela Huppenkothen

Tutorials

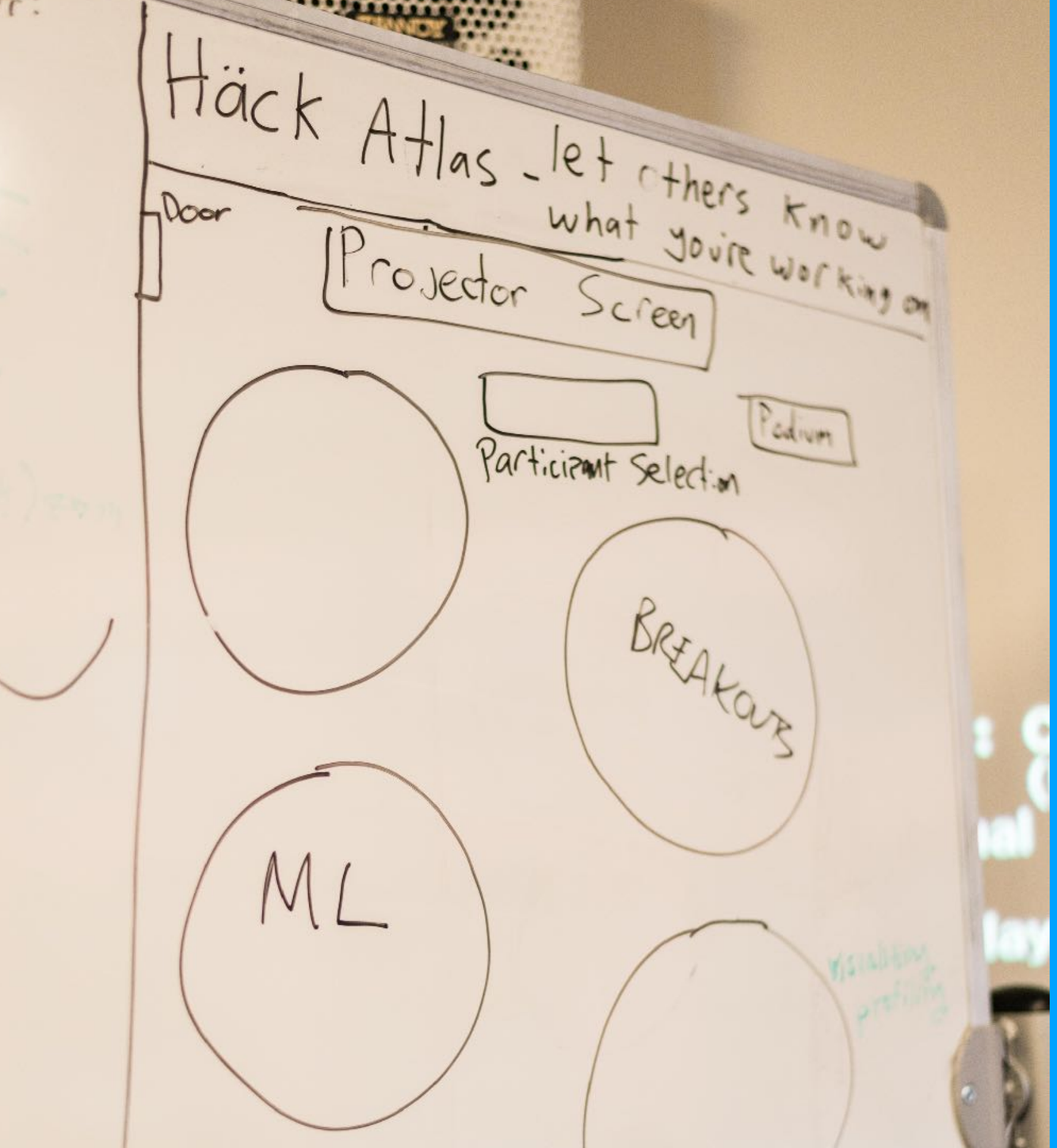
- practically oriented
- interactive
- make use of participants’ expertise

credit: Alex Alspaugh/University of Washington


MORE BREAKOUTS
(NULP)
Julia
Packaging (Python)
Bayesian model comparison
BREAKOUTS
DATABASE
TEAM DB
STOCHASTIC GRADIENT
GAIA DR1

Learning through hacking

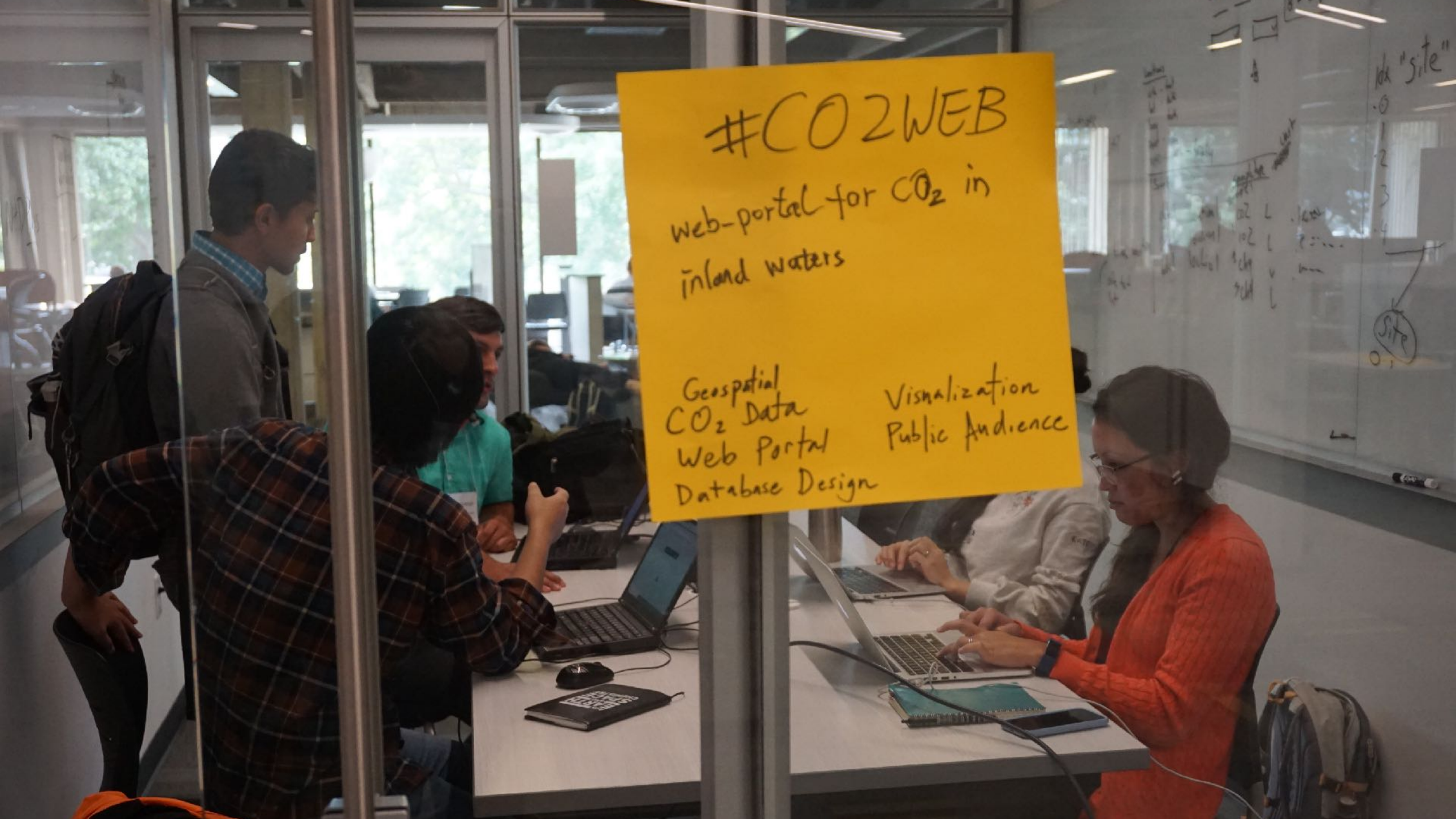
credit: Alex Alspaugh/University of Washington



Collaboration

hack (n):

A hack is a small project with a very clear goal, which should be completed by the end of the time initially allocated to it


#CO2WEB
Web-portal for CO_2 in inland waters
Geospatial
CO_2 Data
Web Portal
Database Design
Visualization
Public Audience

Does it work?

- track long-term outcomes (papers, software, …)
- evaluation via post-attendance surveys
- ethnographic work
- case studies
- team photos
- regular discussions across hack weeks

Astro Hack Week 2018

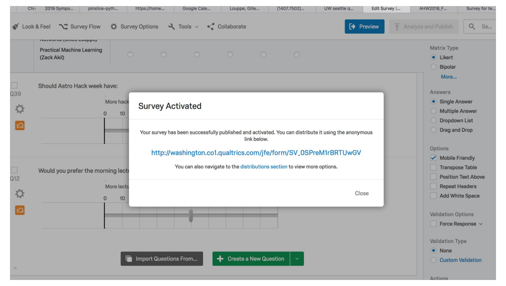
Wrap Up Slides

August 6-10, 2018

AHW 2018 Survey

(Daniela Huppenkothen + Antonia Rowlinson)

… is ready for you! (Link + password tomorrow morning!)



Tutorials for formulating problems in a Bayesian way

Leon Trapman, Mohammadjavad Vakili, Iain Murray, Andrei Igoshev, Daniel Mortlock

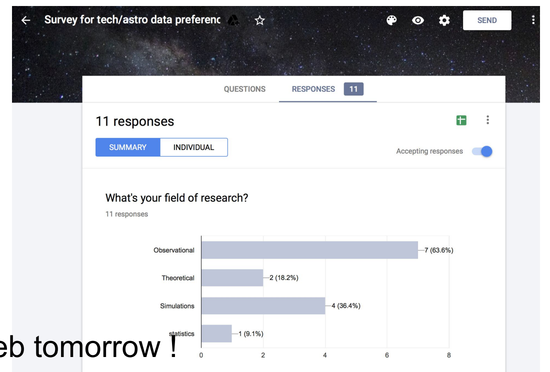
(community hack; 2018-08-09; IBM & Astro Hack Week)

1. Inferring distance to a star from a parallax measurement [A.I.; DONE]
2. Inferring cosmological parameters from power spectrum (with emulation) [M.V.]
3. Inferring luminosity of a star from parallax and flux measurements [A.I.; EXTENSION OF 1.]
4. Inferring the Solar System potential from a snapshot of planets kinematics [I.M.; PUBLISHED]
5. Inferring the mass of the Galactic halo from Magellanic clouds [I.M.; PUBLISHED]
6. Inferring the age of neutron stars from Galactic position, parallax and proper motion [A.I.]
7. Inferring dust content of a protoplanetary disk from an ALMA image [L.T.; SORT-OF-DONE]
8. Inferring whether an asteroid will hit the Earth [I.M., D.M.]
9. Inferring the properties of a merger from gravitational wave observations [A.I.]
10. Inferring which card is showing of white-white, white-black, black-black [I.M., D.M.]
11. Inferring the number density of galaxies from a survey [D.M.]

Survey for tech/astro data preference

Amruta Jaodand
Daisy Mak
Lilianne Nakazono
Norhasliza Yusof

And YOU !



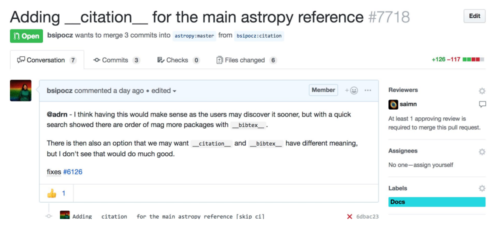
We will launch our web tomorrow !

Brigitta Sipőcz

Sentiment analysis of GitHub issue/PR comments

Fail: run into API limits after the first 437th comment, given up.

Made sure __citation__ and __bibtex__ works. It does now.

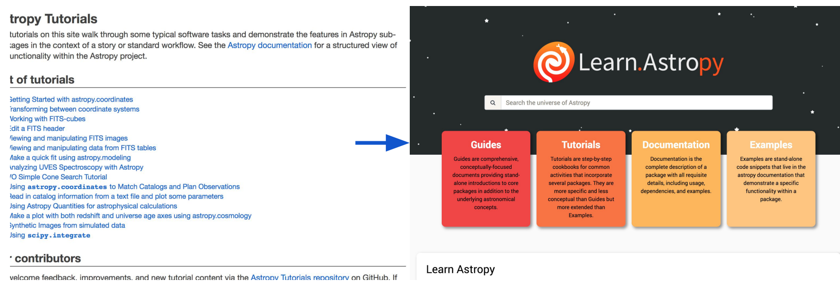


TODO: make sure **makecite** uses __citation__/__bibtex__ when available

Adrian Price-Whelan

Succeeded in getting Brigitta to attempt sentiment analysis on GitHub issue and pull request comments (but see previous slide)

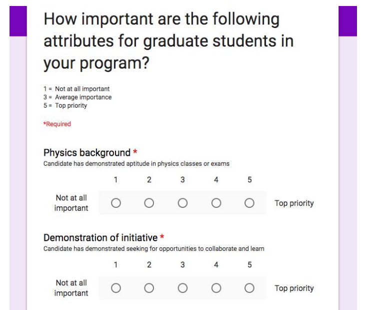
Worked on infrastructure and in progress overhaul of Astropy tutorials site



Astro Grad Admissions Optimization: questionnaire and output

Camila, Malavika, Pearce, Riccardo, Robin, Sean, Statia, Tarun

Evaluation Criteria	Super Application Stage	Interview Stage	Offer Stage
Physics Preparation	35	35	35
Computational Skills	35	35	35
Character Values	30	30	30



Based on your priorities, the following assessment tools are recommended for admissions to your program:

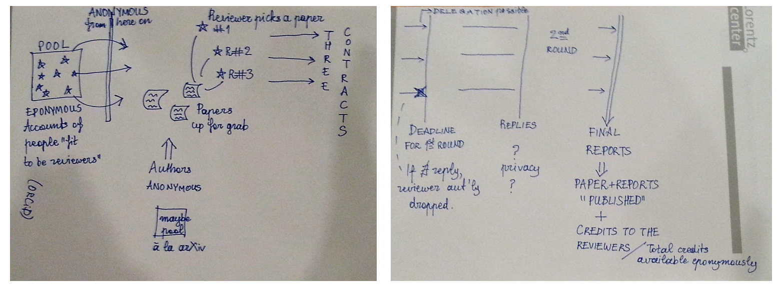
…

Questions to include in letters of reference:

"Briefly (in 5-6 sentences) describe a time that the candidate demonstrated initiative. This could include reaching out to potential mentors or collaborators, learning independently, or taking on tasks on their own."

Eleni & Alexandar, Daniel, Yusra — Peer review and the blockchain

A possible implementation of the peer-review system (as it is today) without journals, with blockchain. [More will be written in the doc...]



https://docs.google.com/document/d/1fwMtRsYj2A-NHY3pgZJ38DHZvwhtkelIXZq2UiMOoYQ

Sentiment analysis via Google NLP API - Jeroen

Steps:
- Use Github API to pull in some comments
- Created a Google Cloud Project, enabled NLP API
- Created credentials
- Use Google NLP API to parse the text

Some results of PR: https://github.com/astropy/astropy/pull/7712

I believe the grouping should work for `Time` mixins, too now that sorting is working?
Score: 0.6 Magnitude: 1.2

so if this should fail, could you add another example where shorting is failing for these columns?
Score: -0.7 Magnitude: 0.7

note that at this point `keys` had to be an `ndarray`, so all the code below dealing with a pre-made index was never being run.
Score: -0.1 Magnitude: 0.3

Score: -1 negative, 1.0 positive. Magnitude: How strong a reaction is

AHW 2019 and 2020 Venues — Lauren and Ellie

Finding venues for unique conferences is challenging, conference venues are expensive

Keep AHW affordable, look for venues that include some budget participants, small/non-existent conference costs

Venues are already booked for 2019 and looking applications for 2020

Flatiron Institute, Banff International Research Station for Mathematical Innovation and Discovery, Casa Matematica Oaxaca, Ringberg

Other ideas/suggestions ?? Add them to this document please !!

ScienceTheatre hack week

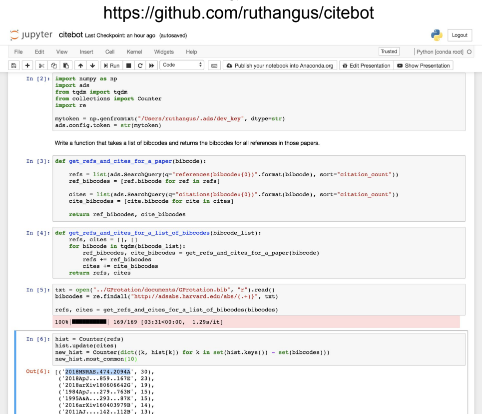
Ruth, Daniella, Pearse, Marie

- Discussed motivation, goals and objectives
- Structure
- Venue
- Funding
- Program
- Expectations and outcomes

Document here:
https://docs.google.com/document/d/1An1SW8h6SRlwmbiltnMsSG0MgBMMFGzwn_s46-oUElo/edit

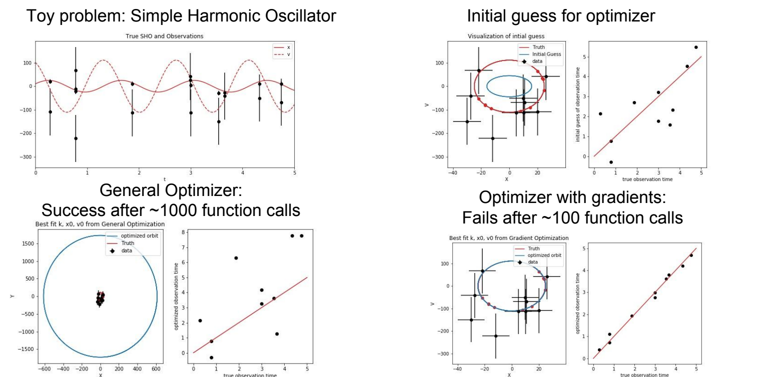
citebot

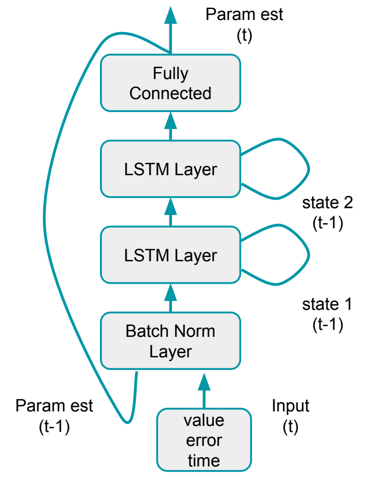
RA & DFM
https://github.com/ruthangus/citebot



Lauren Anderson, Adrian Price-Whelan, Dan Foreman-Mackey, Iain Murray

Gradients of likelihood model to use HMC samplers, or various optimization stuff

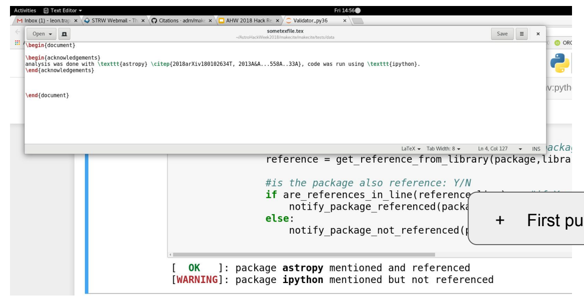




makecite —> check_cite

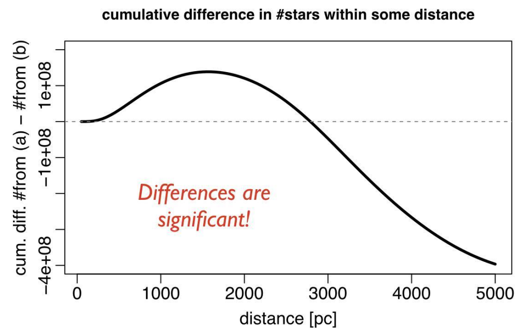
Leon Trapman

+ Adrian Price-Whelan/ Alexander Mechev / Julia Melo Rodrigues de Aguiar / Brigitta Sipőcz



Coryn Bailer-Jones, MPIA Heidelberg — *"Inverse parallax isn't distance"*

I compared no. of stars in Gaia-DR2 within some distance as computed by
(a) inverse parallax
(b) mode of posterior (from our distance catalogue: 2018AJ....156…58B)



Riccardo Buscicchio, University of Birmingham, UK

Challenge: Try not be scared by numerical integration, i.e. evaluate

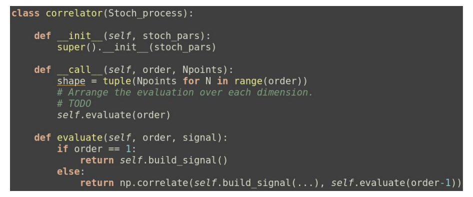
$$H_n(\tau_1,\ldots,\tau_n)=\int dt_1\ldots dt_n s(t_1-\tau_1)s(t_2-\tau_2)\ldots s(t_n-\tau_n)$$

What did we achieve:

recursive, almost-brute-force approach (thanks, Brigitta!)

$$H_n(\tau_1,\ldots,\tau_n)=\int dt_n H_{n-1}(\tau_1,\ldots,\tau_{n-1})s(t_n-\tau_n) = \mathrm{Corr}\,(H_{n-1}, s(t_n-\tau_n))$$
$$H_1(\tau)=s(t_1-\tau)$$

(soon-to-be)
Concrete outcome:



Thoughts: Any clever implementation is welcome. Btw, non-gaussianities are fun!

ASTRO HACK WEEK LOCAL/REGIONAL EDITION

Lilianne, Stephanie, David

Goals:
- To further extend reach to people who want to learn about astrohack (tools, etc) and its topics but couldn't afford to come to international venues, have fewer resources or were not accepted to the workshop.
- To lessen language barriers. For example, not everyone could speak English so if in a regional/local setting, if everyone speaks Portuguese then easier to teach or implement the workshop.
- To encourage people to learn new topics beyond their choice of study and engage them to use these topics for their perusal, expand skills and learning.
- To accomplish good activities for the astronomy society in the local country and in the general public as a whole. MORE IDEAS? SUGGESTIONS?

Link:
https://docs.google.com/document/d/1xRiE6CGYTSHQ6K2iEnprLmj3u9fMVUUnclxX-pxdAPU/edit?usp=sharing

Motivator - Chrome Extension

Boris L
Nicolas A

Including great historical quotes from Ru Paul, Beyonce, your grandma, etc.

Deep Time Series

Alexandar, Brigitta, Ellianna, Gilles, Nicolas, Pearse, Rodrigo, Rohan, Ruth, Tarun

Goal

There is a lot of information about the mass, age and rotation period of a star in its light curve but our physical models and the tools we use to extract this information are flawed. We postulate that we can do better with RNNs.

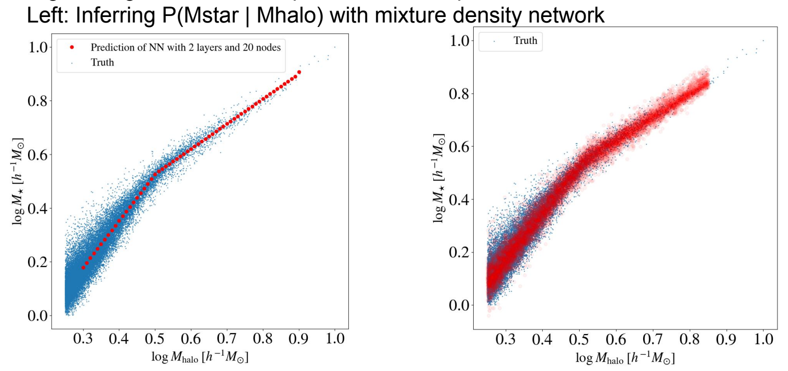
Learnings
- Data pre-processing is hard.
- RNNs are cool.
- RNNs are expensive - try other approaches first!

Next Steps
1. Run this architecture on 16K Kepler Red Giants star data
2. Apply a Generative Adversarial Network?
3. First few key features to investigate from Kepler data: Mass, age, rotation period

Link to learnings document: https://tinyurl.com/yauuw98z

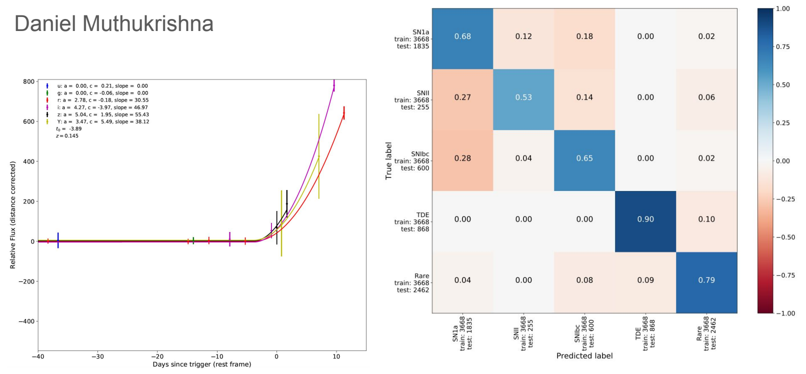
Mohammadjavad Vakili with lots of helpful practical advice from Cole Clifford

Inferring the central galaxy stellar mass-halo mass relation with Neural nets:
Right: Regression with simple tensorflow implementation of FC NN
Left: Inferring P(Mstar | Mhalo) with mixture density network



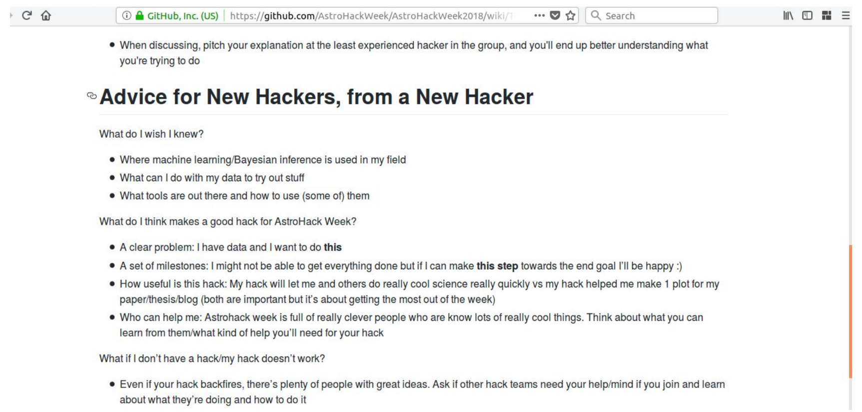
Recurrent Neural Net (GRU) and 1D CNN for early transient light curve classification

Daniel Muthukrishna



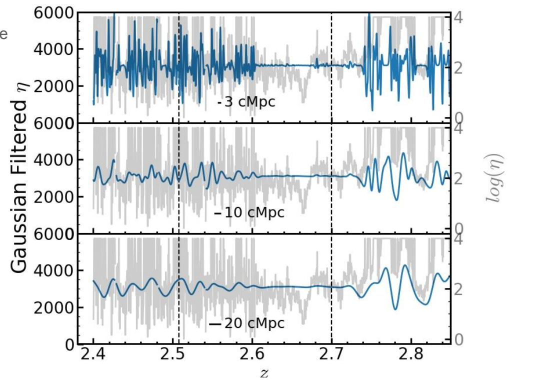
How to make the most out of AstroHack Week

Pearse Murphy



Using Gaussian Bandpass to Filter Data

Sean Morrison
Laboratoire d'Astrophysique de Marseille



Pearse Murphy, Trinity College Dublin, Ireland

Challenge: Compute 512 FFTs on ~2 million points without killing my computer

What did we achieve: Found a python wrapper for the FFTW library and implemented it. Unfortunately there was no significant speed up.

Concrete outcome: A "lazy" solution is to do 512/N FFTs on N computers at the same time and collect data at the end

Thoughts: I might join a different hack - feature recognition with machine learning type of thing.

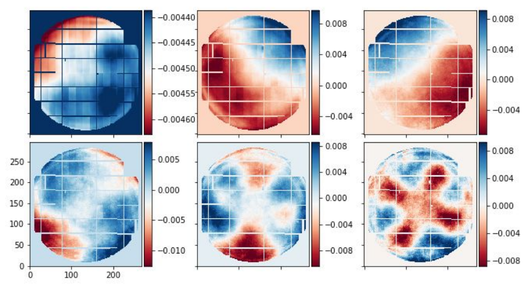
Yusra AlSayyad (Princeton University/LSST)

Yusra AlSayyad (Princeton University/LSST)

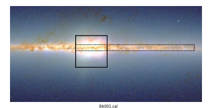
Goal: Explore HSC backgrounds

Produced some interesting eigen-backgrounds for the Y-band.

Thanks to Rodrigo, Matthew, Nicolas for brainstorming with me

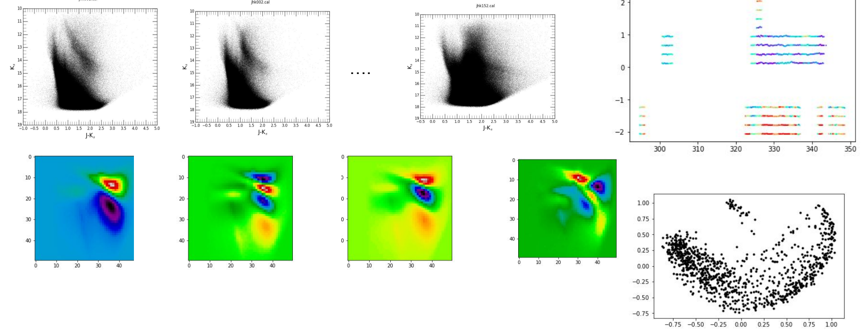


Efşan Sökmen - Iain Murray



Thanks to : Eleni Petrakou and Andrei Igoshev

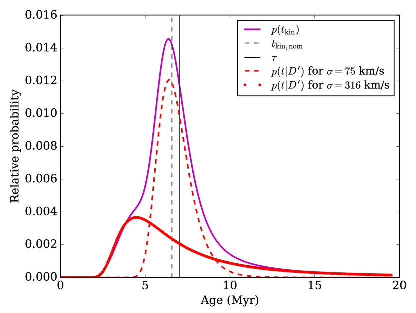
PCA on Stellar Populations in the Southern Plane - VVV survey



Andrei Igoshev, Technion, Israel

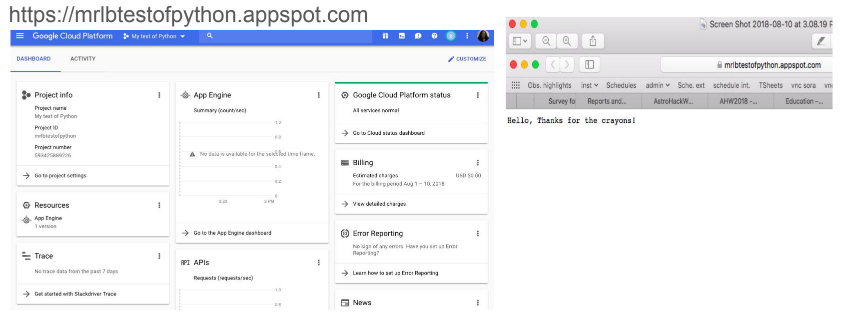
Deriving the posterior for a variable which depends on many measured values but not measured itself.

Posterior is derived.



FUN with GOOGLE CLOUD PLATFORM Python APP Engine

https://mrlbtestofpython.appspot.com

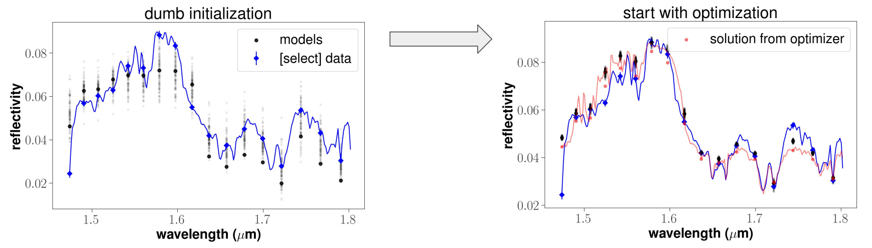


Improved parameter estimation: planet spectra and RT models

Statia Cook, Columbia Univ./AMNH

Help from: Iain Murray (!!), Lauren Anderson, Becky Steele, Brigitta Sipőcz, Daniela

- runs with emcee are slow, don't always converge well, not sure if method good for my 10-15 model parameters
- Initial idea: test out something other than emcee
- Actual "hack": try optimizing first, work with simplified data and model (for speed)



Results: works better! I learned that using optimizer first is a good idea :-)

Cardboard Universe: tinyurl.com/3dexoplanets

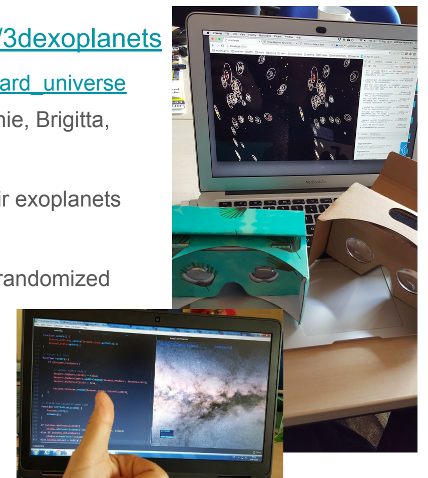
https://github.com/beckysteele/cardboard_universe

Team: Matt, Ellie, David, Efsan, Stephanie, Brigitta, Yanett, Becky

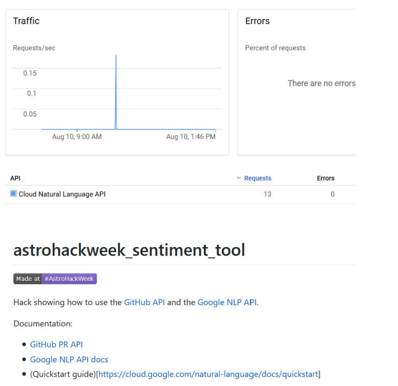
Challenge: Zoom through stars and their exoplanets using Google Cardboard + Three.js

Achieved: In-browser prototype ready (randomized systems only)

Next steps: Connect Exoplanet Archive data to 3D simulation, input a 360 deg view with a Milky Way background, and make it Google Cardboard-able



Jeroen Bédorf - Leiden University/Observatory

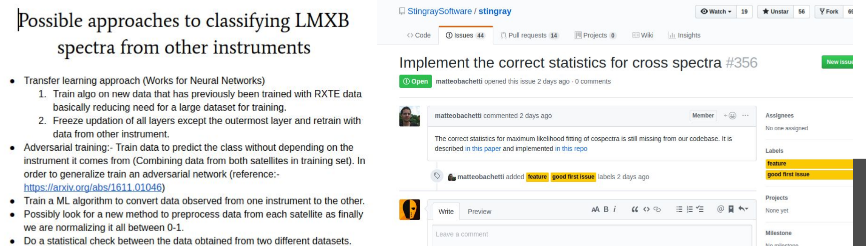


Google APIs, challenges involved: Finding creditcard details, access permissions, service user roles, including credentials in the API request, accessing the results, enabling the correct APIs, installing the correct Python packages.

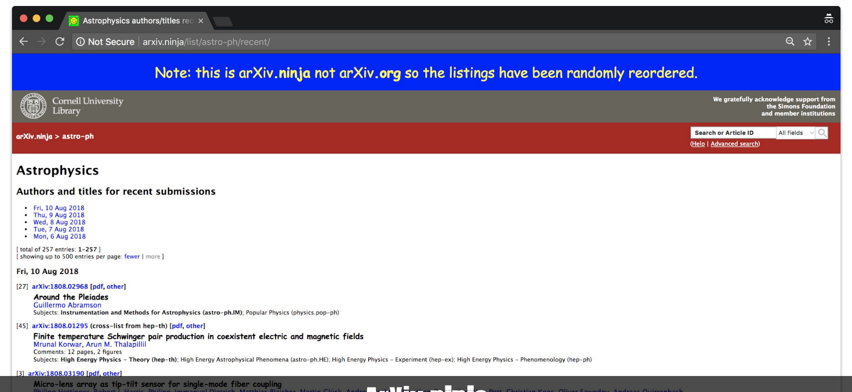
https://github.com/jbedorf/astrohackweek_sentiment_tool

Rohan Pattnaik

Personal Hack Objectives:
- Compile a list of approaches to classify spectra from other instruments
- Get started with open source development



ArXiv.ninja



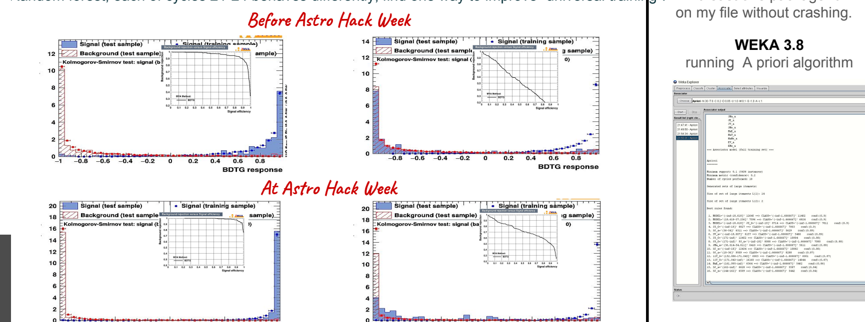
ArXiv.ninja
Dan F-M // Adrian P-W
github.com/dfm/arxiv.ninja

BIG DATA METHODS FOR EXTRACTING RELATIONS BETWEEN THE TIMING OF SOLAR FLARES AND PLANETARY POSITIONS

Eleni Petrakou

Indications exist for a relation between them.
Goals of the week: Take a solid step in classification + Kickstart associative rules mining.

ONE SOLID STEP IN CLASSIFICATION:
Random forest; each of cycles 21-24 behaves differently; find one way to improve "universal training".



KICKSTART RULES:
= At least one package running on my file without crashing.

WEKA 3.8
running A priori algorithm

AstroCapital

Goal: Create a web-page which educates and allows astronomers to communicate about tech-enabled all-things-astro.

Why:
- Lack of such a platform
- Efficient, new ways to go from data to astronomy/science
- Open platform for everyone
- share, contribute and stay updated!
- Central astro-tech resource hub
- Connecting ex-/non-astronomers to the astro community

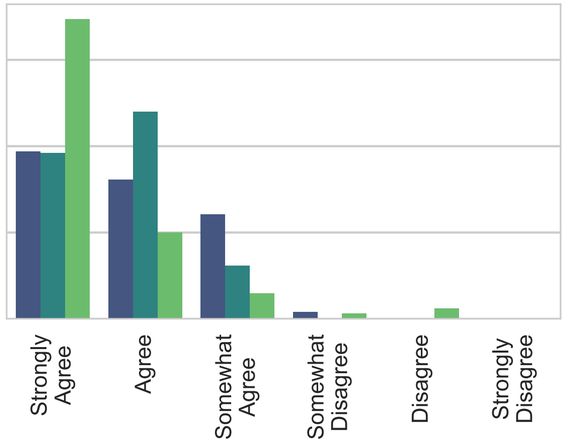
Status: Survey and Web

What next:
- Let us know if you can contribute to any sections (e.g. writing blogs)

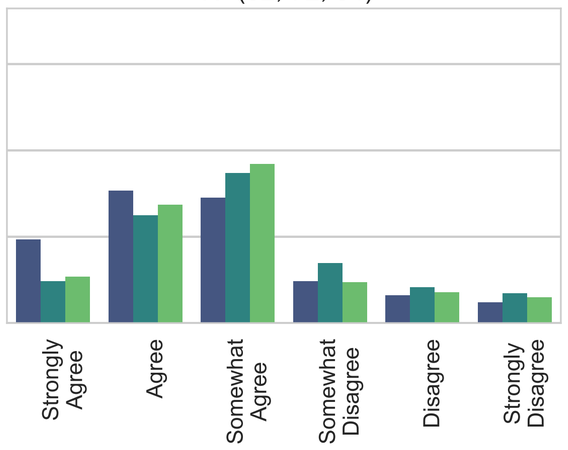
AmrutaJaodand
Daisy Mak
LilianneNakazono
Zach Akil
NorhaslizaYusof
Becky Steel
Mohammadjavad Vakili

Survey Results

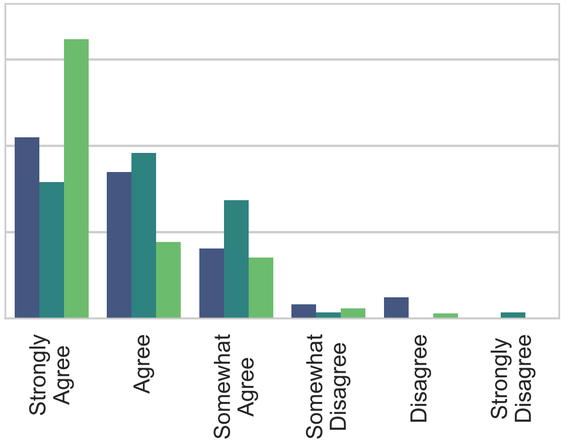
(b) I feel like I learned things which improve my day-to-day research; N=(62, 73, 85)



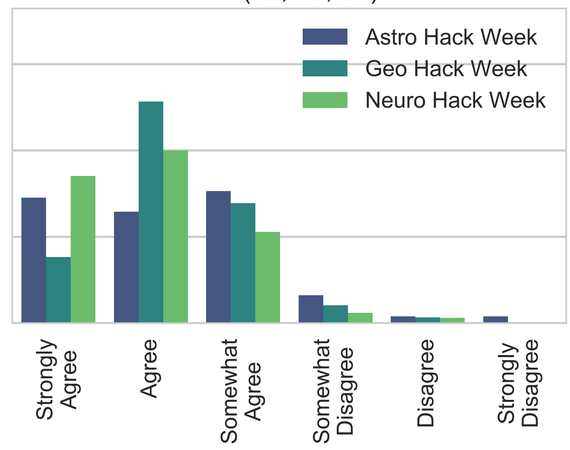
(d) I ended up teaching things to other people; N=(62, 72, 84)



(e) I felt like I built valuable connections at X Hack Week; N=(62, 73, 85)



(h) I feel like X Hack Week has made me more comfortable with doing open science; N=(62, 72, 85)



Take-Away Lessons

build a community first

credit: eScience Institute

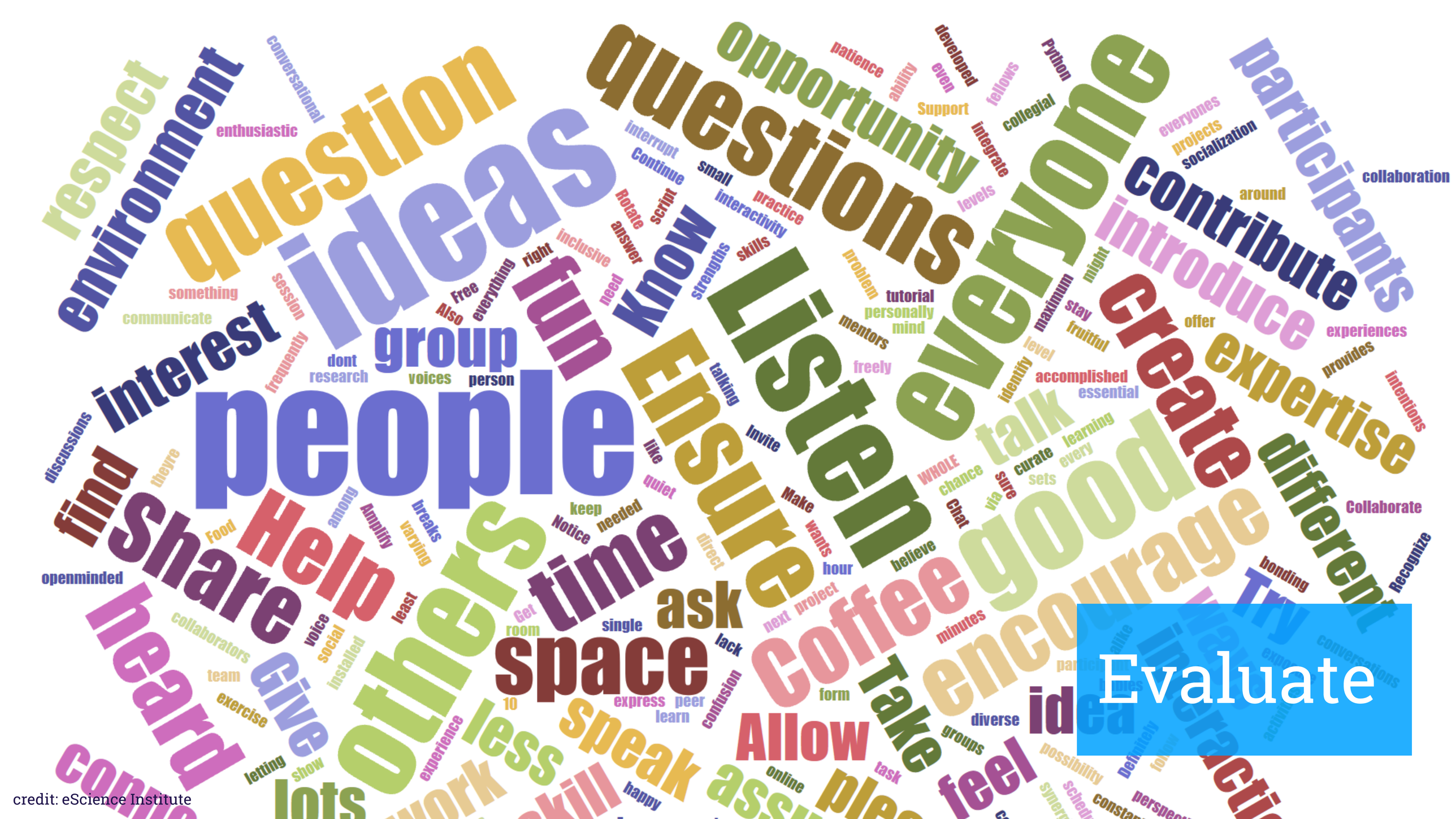
build a culture that empowers people to ask fundamental (and trivial) questions

credit: Alex Alspaugh/University of Washington

Adapt concepts and ideas to your community's needs


ASTRO
HACK
WEEK
Häck Atlas - let others know what you're working on
Projector
ML
Deep Learning
LUNCH SET-UP

Experiment



credit: eScience Institute

http://www.pnas.org/content/early/2018/08/17/1717196115

Hack Weeks as a model for Data Science Education and Collaboration

Daniela Huppenkothen[a,b,c,1], Anthony Arendt[d,e], David W. Hogg[b,a,f,g], Karthik Ram[h], Jake VanderPlas[e], and Ariel Rokem[e]

+ extensive supplementary materials

+ this talk: https://github.com/dhuppenkothen/sdss-talk

+ living checklist: https://docs.google.com/document/d/1PKcb7J3Xx3fjA-oHeKtvLdcvWmA3tCqs8T2CXAsv0U8/edit?usp=sharing

dhuppenkothen

dhuppenk@uw.edu

Tiana_Athriel

http://www.pnas.org/content/early/2018/08/17/1717196115

Hack Weeks as a model for Data Science Education and Collaboration

Daniela Huppenkothen[a,b,c,1], Anthony Arendt[d,e], David W. Hogg[b,a,f,g], Karthik Ram[h], Jake VanderPlas[e], and Ariel Rokem[e]

+ extensive supplementary materials

+ this talk: https://github.com/dhuppenkothen/sdss-talk

+ living checklist: https://docs.google.com/document/d/1PKcb7J3Xx3fjA-oHeKtvLdcvWmA3tCqs8T2CXAsv0U8/edit?usp=sharing

Come and chat with us!

dhuppenkothen

dhuppenk@uw.edu

Tiana_Athriel