data science in astronomy

git and GitHub

UT Austin Astronomy
grad student and postdoc seminar

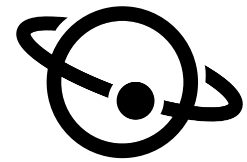
michael gully-santiago

Graduate student at UTexas
Astronomy
Aug 25, 2008 – May 2015
(projected)

Advisor: Dan Jaffe

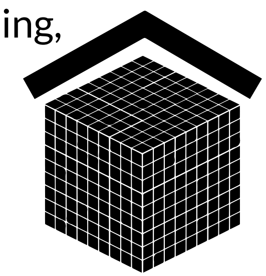
I make diffraction gratings from single crystal silicon.

I work on brown dwarfs, and have broad interests in star and planet formation.



Lately,

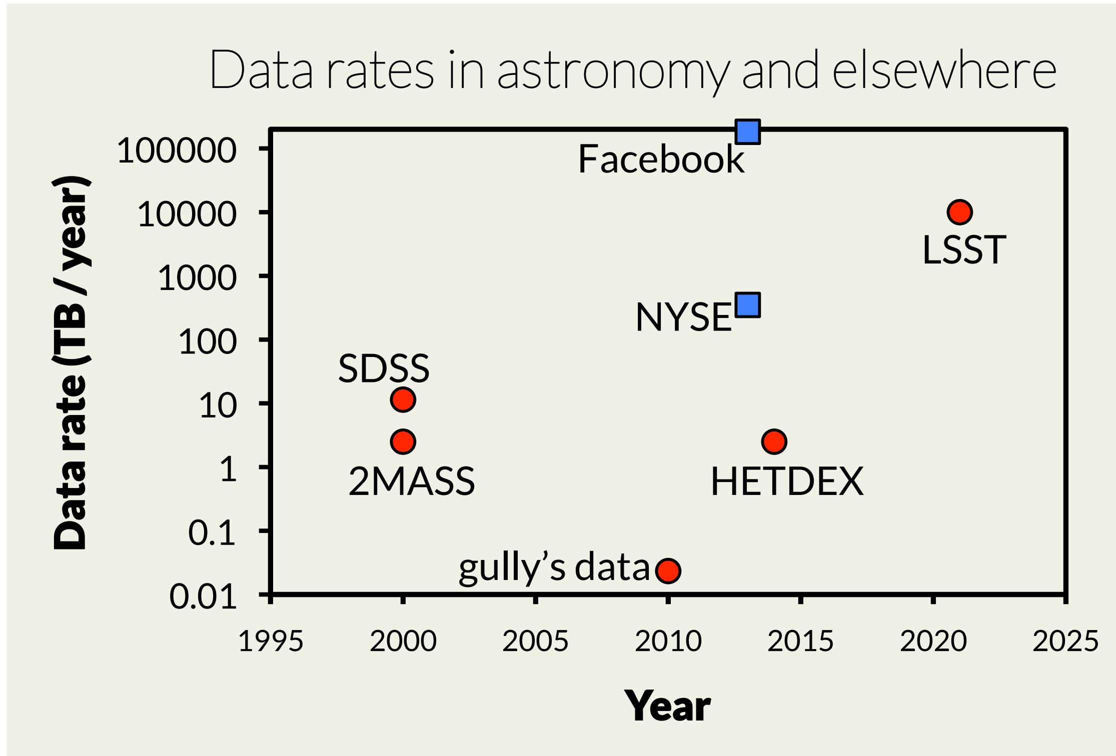
I've been building my skills in statistics, data mining, machine learning, and modern computing.



why?

The volume of data in astronomy is growing.

volume of data



sources:

SDSS Bill Howe (UW)
2MASS http://spider.ipac.caltech.edu/staff/roc/2mass/archive/data.profile.v3.html
My data set MGS
HETDEX http://hetdex.org/pdfs/research/Hill1.pdf
LSST Bill Howe (UW)
NYSE http://marciaconner.com/blog/data-on-big-data/
Facebook http://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/

The variety of data in astronomy is growing.

Here is a 94 second segment from a Coursera video.

It's from 0:30 to 2:14 of 'eScience' in Bill Howe's Introduction to **Data Science**

https://class.coursera.org/datasci-001/lecture/19

Key idea.
The skills that will be useful for astronomy already are useful for data science.

databases

Python

SQL

NoSQL

MapReduce
/Hadoop

Key idea.

The skills that will be useful for astronomy already are useful for data science.

Automated
analysis

git &
GitHub

Cloud
Computing

Machine
Learning

R

Visualizations

Key problem.
The astronomy job market is sorta tough.

Key insight.

Let’s build data science skills, because it will make our astronomy better, and better prepare us for NAPs*.

It’s a win-win.

*NAPs
Non Academic Professions (C. Lindner talk from GSPS Jan. 17, 2014)

It’s a win-win.


OBAMA
CLINTON
DEM 70
REP 41
1:51
BUSH JR.
PALIN
NBA JAM
SPORTS

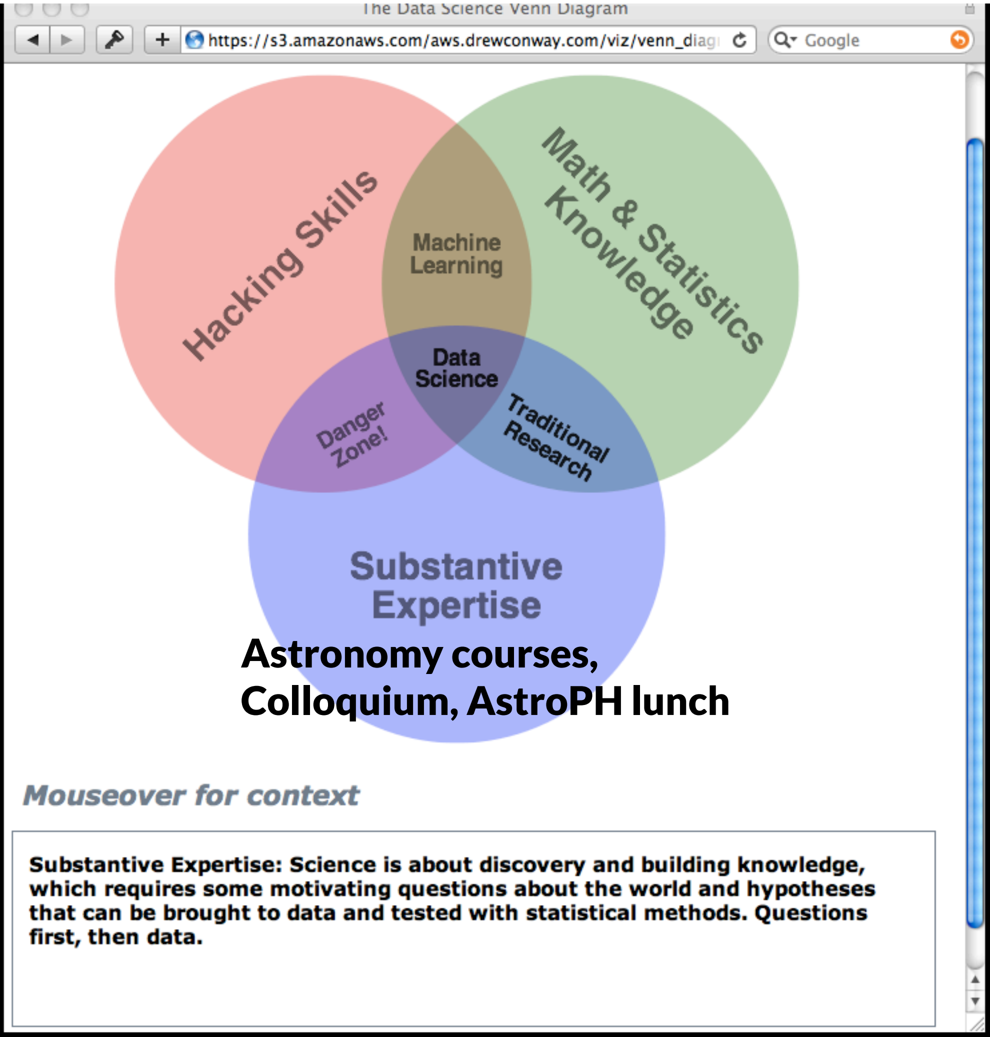
It’s a win-win.

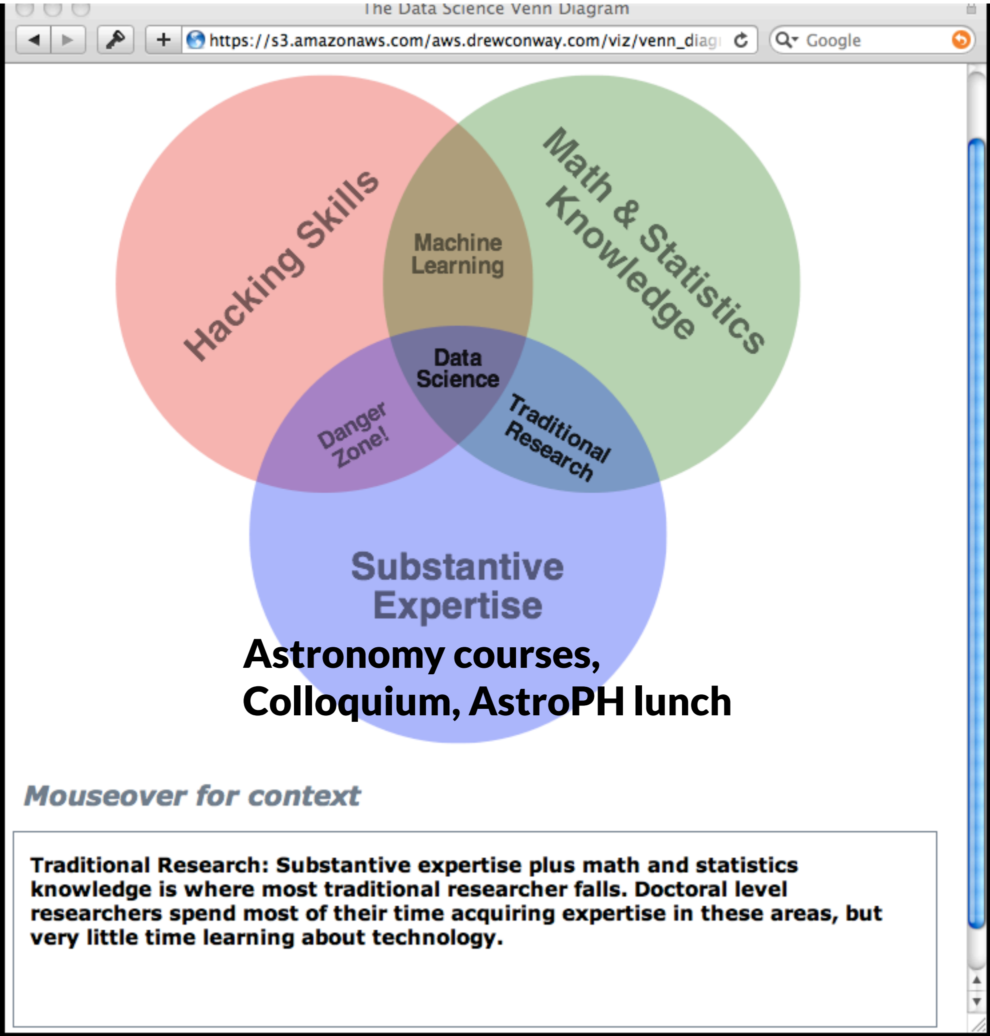

ON ACCOMPLISHED

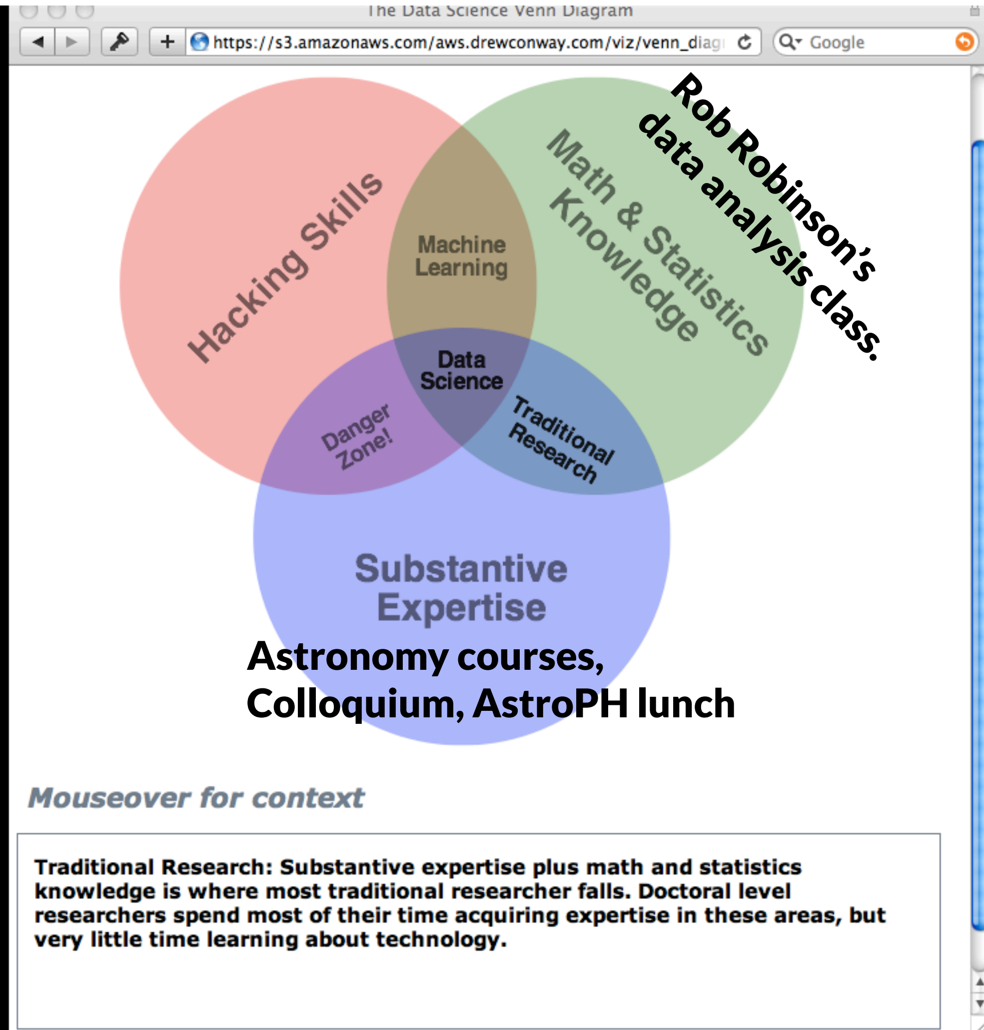
Key question.

So how do we build these skills?

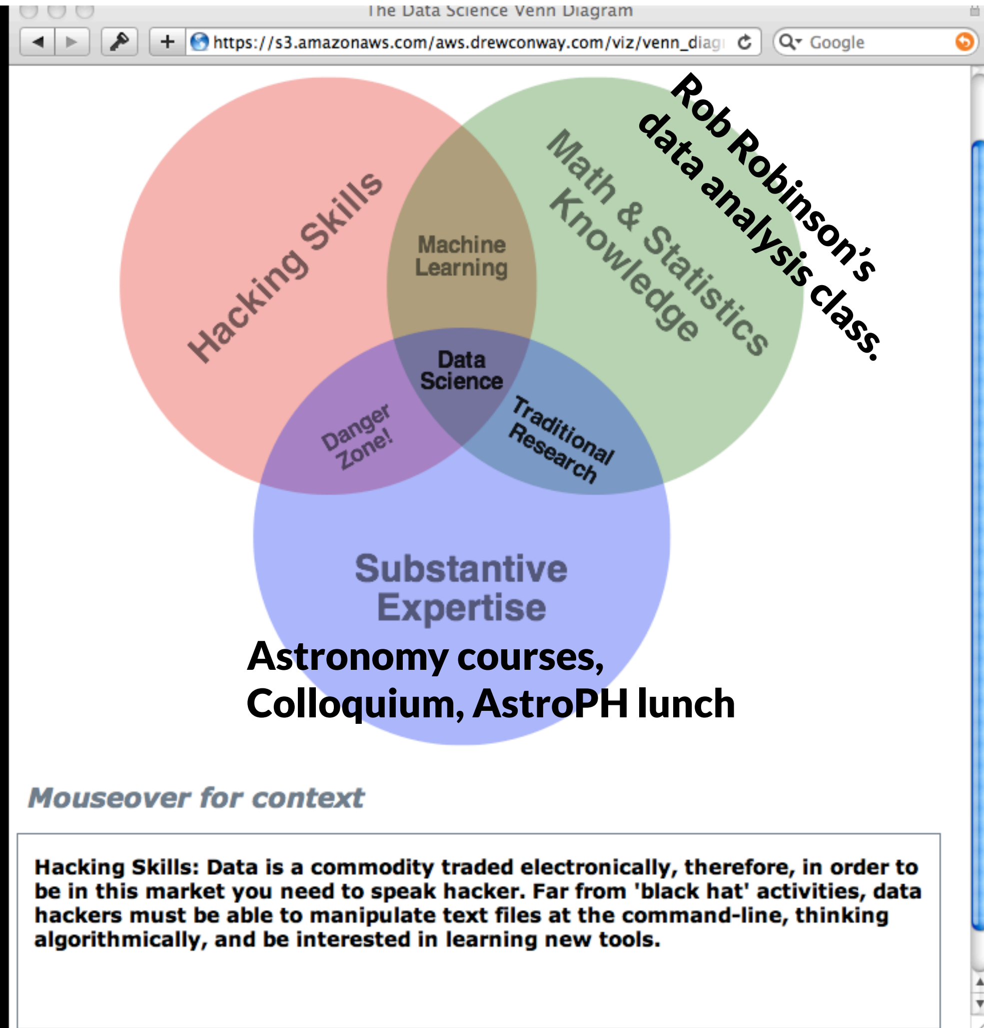

The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Mouseover for context
Substantive Expertise: Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods. Questions first, then data.

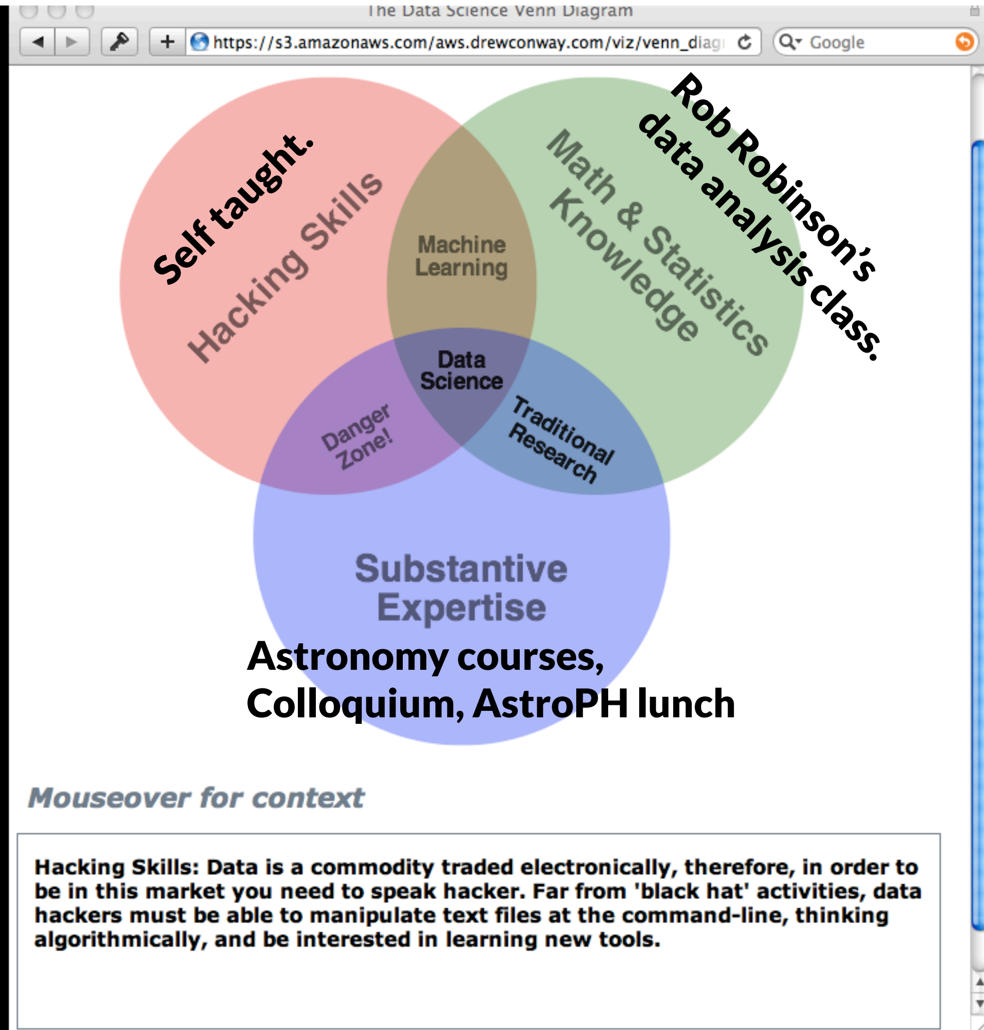

The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Astronomy courses,
Colloquium, AstroPH lunch
Mouseover for context
Substantive Expertise: Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods. Questions first, then data.


The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Astronomy courses, Colloquium, AstroPH lunch
Mouseover for context
Traditional Research: Substantive expertise plus math and statistics knowledge is where most traditional researcher falls. Doctoral level researchers spend most of their time acquiring expertise in these areas, but very little time learning about technology.

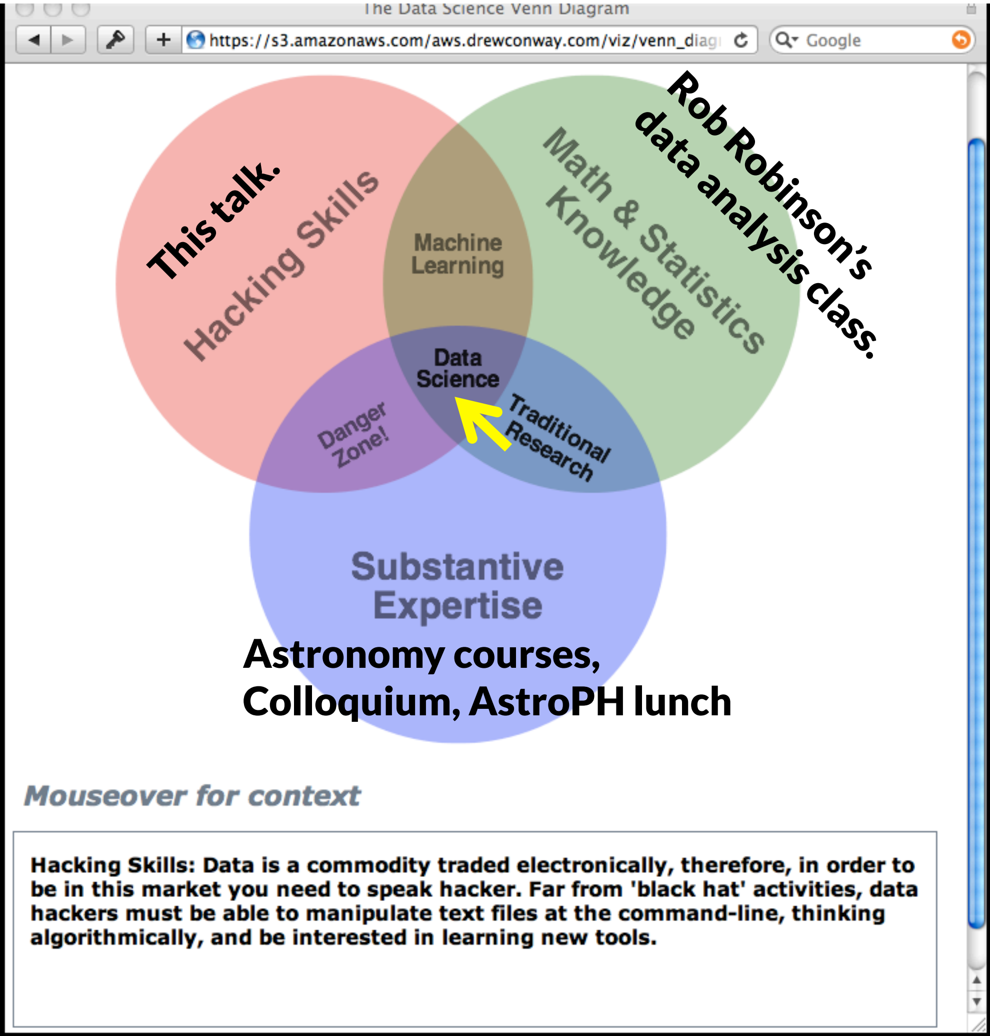


The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Rob Robinson's data analysis class.
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Astronomy courses, Colloquium, AstroPH lunch
Mouseover for context
Traditional Research: Substantive expertise plus math and statistics knowledge is where most traditional researcher falls. Doctoral level researchers spend most of their time acquiring expertise in these areas, but very little time learning about technology.


The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Rob Robinson's data analysis class.
Astronomy courses, Colloquium, AstroPH lunch
Mouseover for context
Hacking Skills: Data is a commodity traded electronically, therefore, in order to be in this market you need to speak hacker. Far from 'black hat' activities, data hackers must be able to manipulate text files at the command-line, thinking algorithmically, and be interested in learning new tools.



The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
Self taught.
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Rob Robinson's data analysis class.
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Astronomy courses, Colloquium, AstroPH lunch
Mouseover for context
Hacking Skills: Data is a commodity traded electronically, therefore, in order to be in this market you need to speak hacker. Far from 'black hat' activities, data hackers must be able to manipulate text files at the command-line, thinking algorithmically, and be interested in learning new tools.


The Data Science Venn Diagram
https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diag
Google
This talk.
Hacking Skills
Machine Learning
Math & Statistics Knowledge
Rob Robinson's data analysis class.
Data Science
Danger Zone!
Traditional Research
Substantive Expertise
Astronomy courses, Colloquium, AstroPH lunch
Mouseover for context
Hacking Skills: Data is a commodity traded electronically, therefore, in order to be in this market you need to speak hacker. Far from 'black hat' activities, data hackers must be able to manipulate text files at the command-line, thinking algorithmically, and be interested in learning new tools.

Our strategy.

Let’s follow Brian Mulligan’s advice, and focus on just a few things.

databases

Python

SQL

NoSQL

MapReduce /Hadoop

Our strategy.

Let's follow Brian Mulligan's advice, and focus on just a few things.

Automated analysis

git & GitHub

Cloud Computing

Machine Learning

R

Visualizations

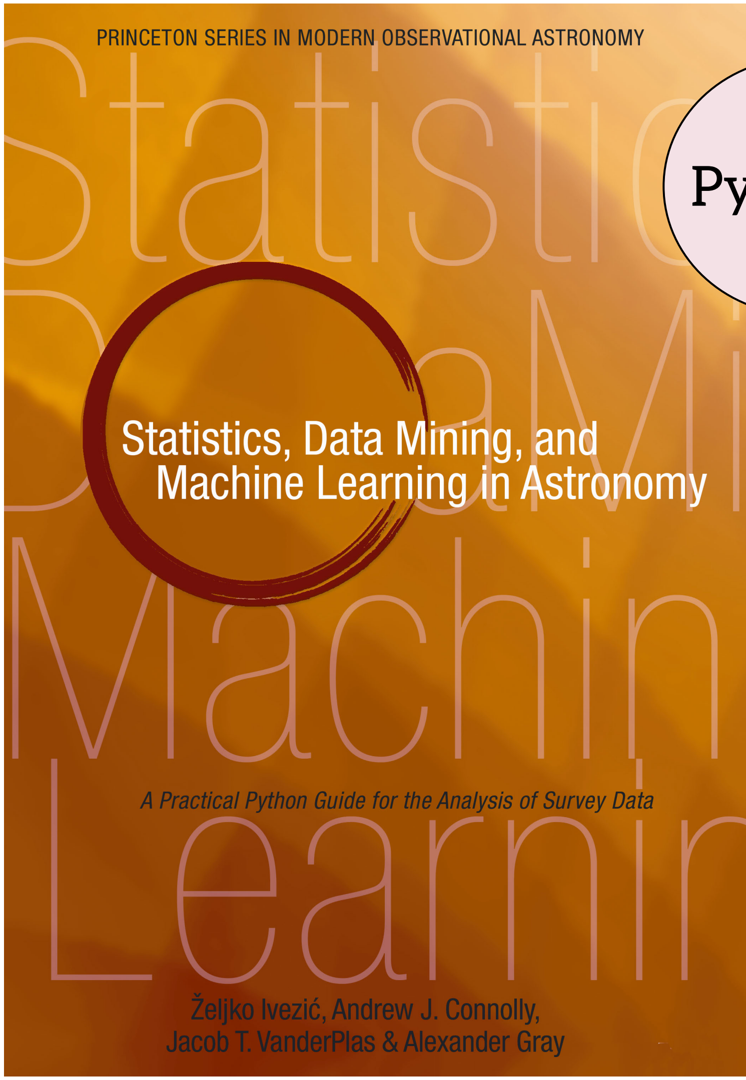
focus on just

a few things.

Python

Machine Learning

git & GitHub



Python

These are the main topics of our data science in astronomy meetup.

gigayear.weebly.com/data-science.html

Machine Learning

mailing list
http://eepurl.com/LdArH

git &
GitHub

Here is an attempt at a live github demo.

complicated selection function is discussed in §4.9, and censored data are discussed in the context of regression in §8.1.

The key point when accounting for truncated data is that the data likelihood of a single datum must be a properly normalized pdf. The fact that data are truncated enters analysis through a renormalization constant. In the case of a Gaussian error distribution (we assume that σ is known), the likelihood for a single data point is

$$p(x_i|\mu, \sigma, x_{\rm min}, x_{\rm max}) = C(\mu, x_{\rm min}, x_{\rm max}) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma}\right), \tag{4.11}$$

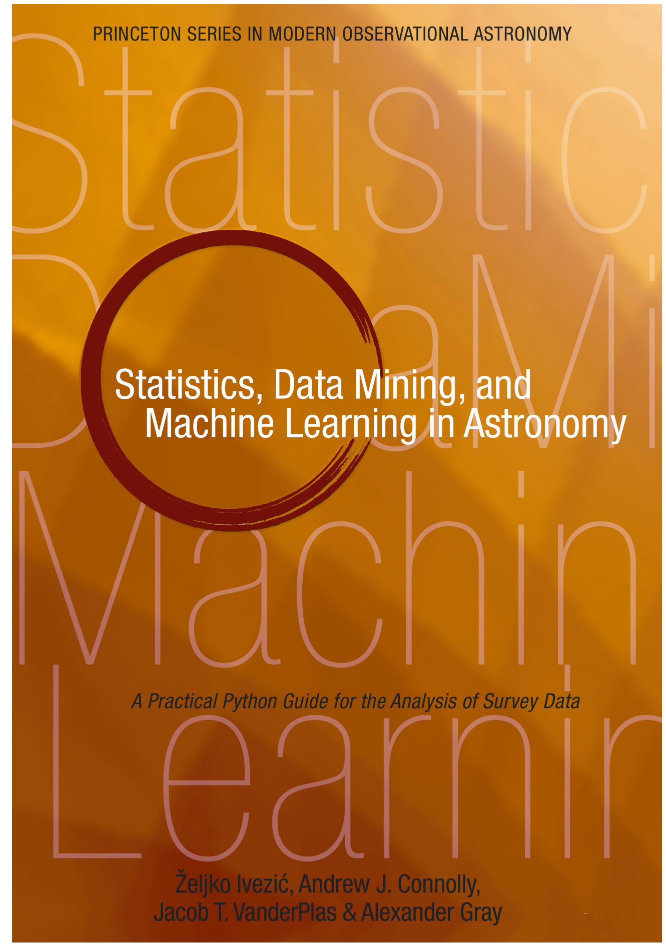
where the renormalization constant is evaluated as

$$C(\mu, \sigma, x_{\rm min}, x_{\rm max}) = (P(x_{\rm max}|\mu, \sigma) - P(x_{\rm min}|\mu, \sigma))^{-1} \tag{4.12}$$

with the cumulative distribution function for Gaussian, P, given by eq. 3.48.

The log-likelihood is

$$\ln L(\mu) = \text{constant} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + N \ln\left[C(\mu, \sigma, x_{\rm min}, x_{\rm max})\right]. \tag{4.13}$$



git and GitHub demo

pull request to astroML code base

Visit astroML github page: https://github.com/astroML

1) Update the README.md file with this new text:

Page 130: The denominator of the argument of the exponential of Eq. (4.11) should be sigma squared, not sigma, to better match Eq. (3.43) and lead to Eq. (4.13).

2) git status, git add, git commit, git push

3) Perform a pull request on GitHub

Thank you.

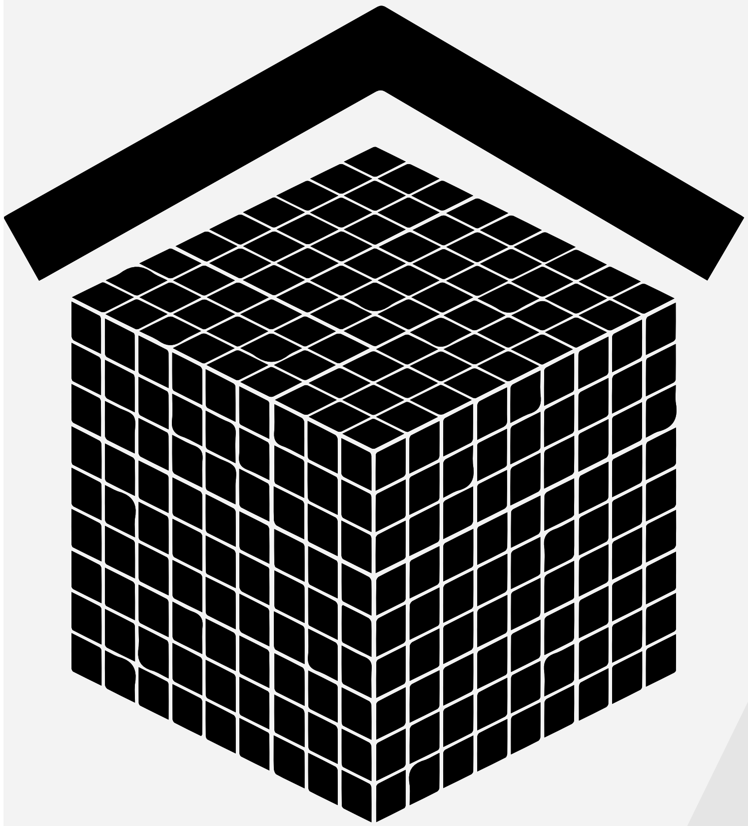
This presentation is available for download on speakerdeck

gully@astro.as.utexas.edu | **astronomer and engineer**

attribution to:

Pierre TORET, from The Noun Project

Sá Ferreira - Purple Matter, from The Noun Project



Open questions for discussion

Is this all worth it?
Will this put more papers in the ApJ?
When is the best time to invest?
Is it still useful if I'm not collaborating?

Are we getting what we want from the Dept.?
How do we build synergies within the Dept.?
How to build momentum, overcome inertia

extras

Global Resources

codeschool.com is a great way to quickly learn git

try.github.io is a great way to try the basics of git

astroml.org contains Astronomy specific machine learning code

coursera.org/course/datasci has free online videos aas.org/posts/story/2014/01/astrophysics-code-sharing-ii-sequel

Making Your Work More Valuable by Giving It Away
Benjamin Weiner (University of Arizona)

NSF Policies on Software and Data Sharing
Daniel Katz (National Science Foundation)

The Astropy Project's Self-Herding Cats Development Model
Erik Tollerud (Yale University)

Costs and Benefits of Developing Out in the Open
David W. Hogg (New York University)

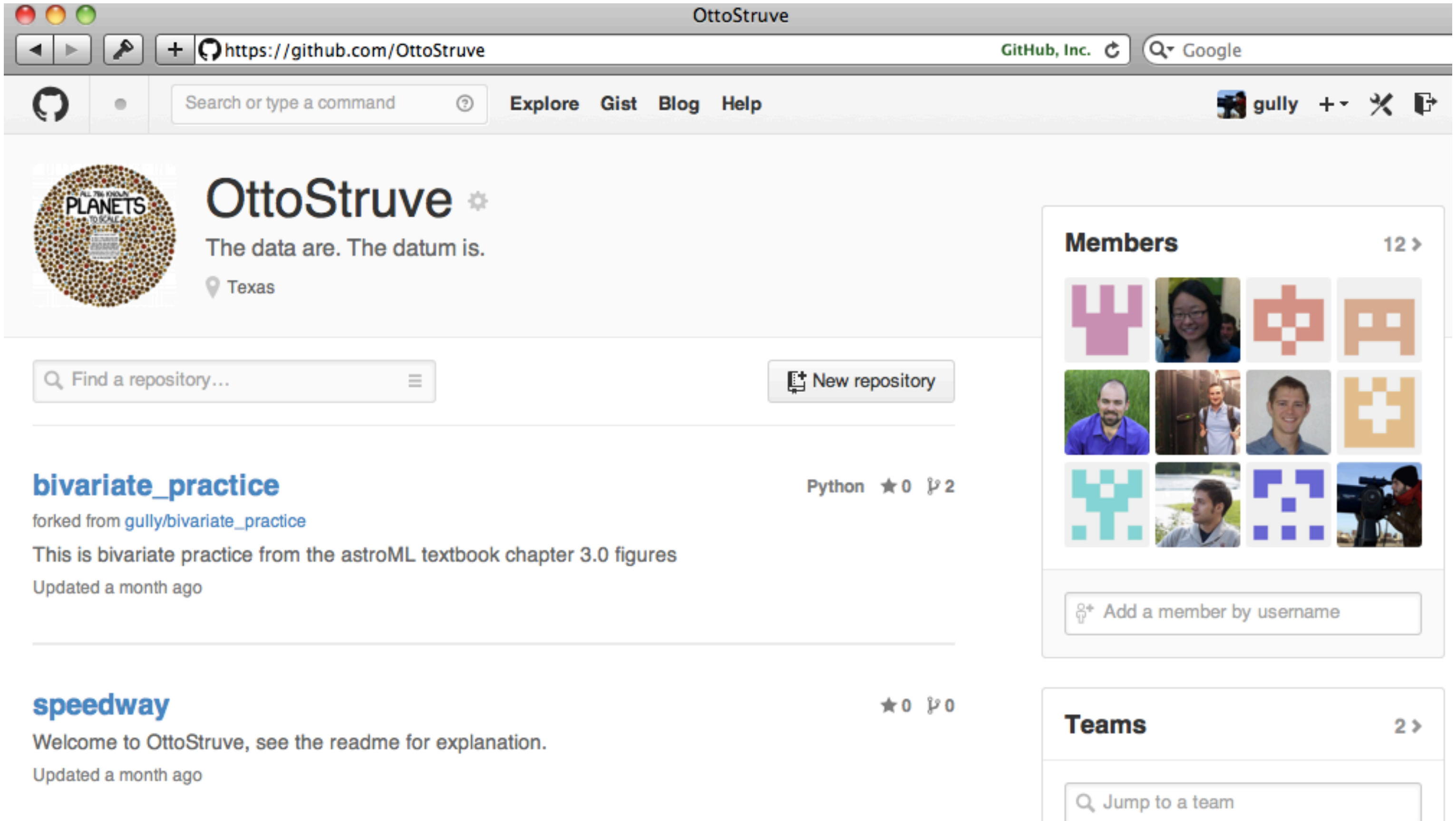
Local Resources

UT Austin data science in astronomy meetup- times vary

Next week's grad student town hall- *(& proposal to astro Faculty)*

Friday, Feb 7 at 1pm in the classroom

UT Austin Astronomy GitHub Organization: OttoStruve



UT Austin Astronomy GitHub Organization: OttoStruve