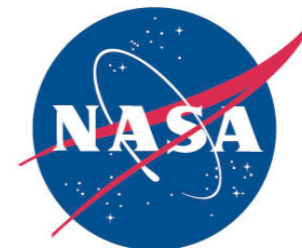


# Computational Training & Data Literacy for Domain Scientists

*Joshua Bloom*  
*UC Berkeley, Astronomy*



@profjsb

*laboratory techniques*

*Physics* **domain  
training**

*machine learning*

**statistics**

*Bayesian* *MCMC*

*GUI* *visualization*  
**advanced**

**computing**

*database* *parallel*  
*MapReduce*

**What is the toolbox  
of the modern  
(data-driven)  
scientist?**

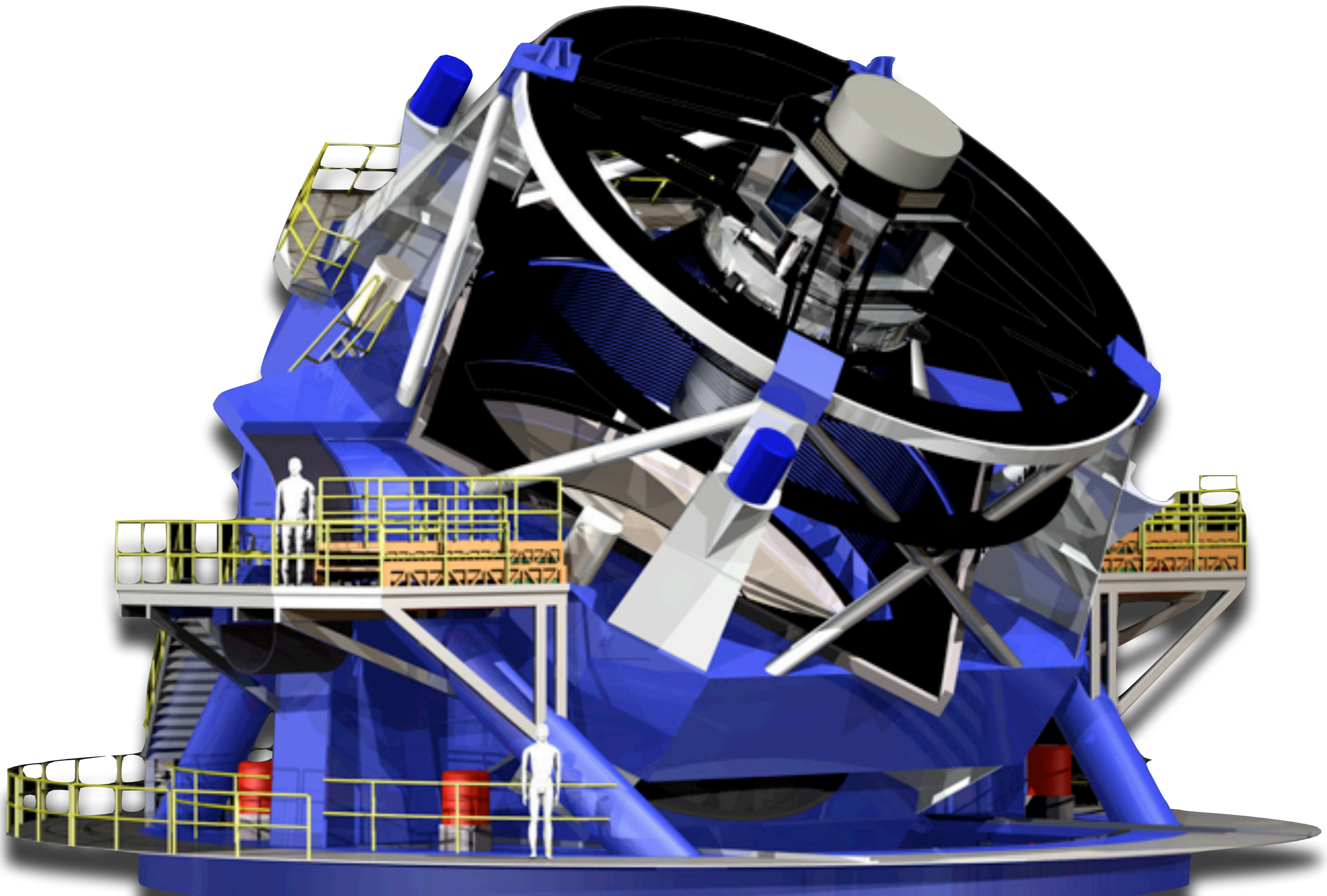


**What is the toolbox  
of the modern  
(data-driven)  
scientist?**

And...How do we teach  
this with what little time  
the students have?

# Astronomical Data Deluge

Serious Challenge to Traditional Approaches & Toolkits



# Astronomical Data Deluge

Serious Challenge to Traditional Approaches & Toolkits

## Large Synoptic Survey Telescope (LSST) - 2020

Light curves for 800M sources every 3 days

$10^6$  supernovae/yr,  $10^5$  eclipsing binaries

3.2 gigapixel camera, 20 TB/night

## LOFAR & SKA

150 Gps (27 Tflops) → 20 Pps (~100 Pflops)

## Gaia space astrometry mission - 2014

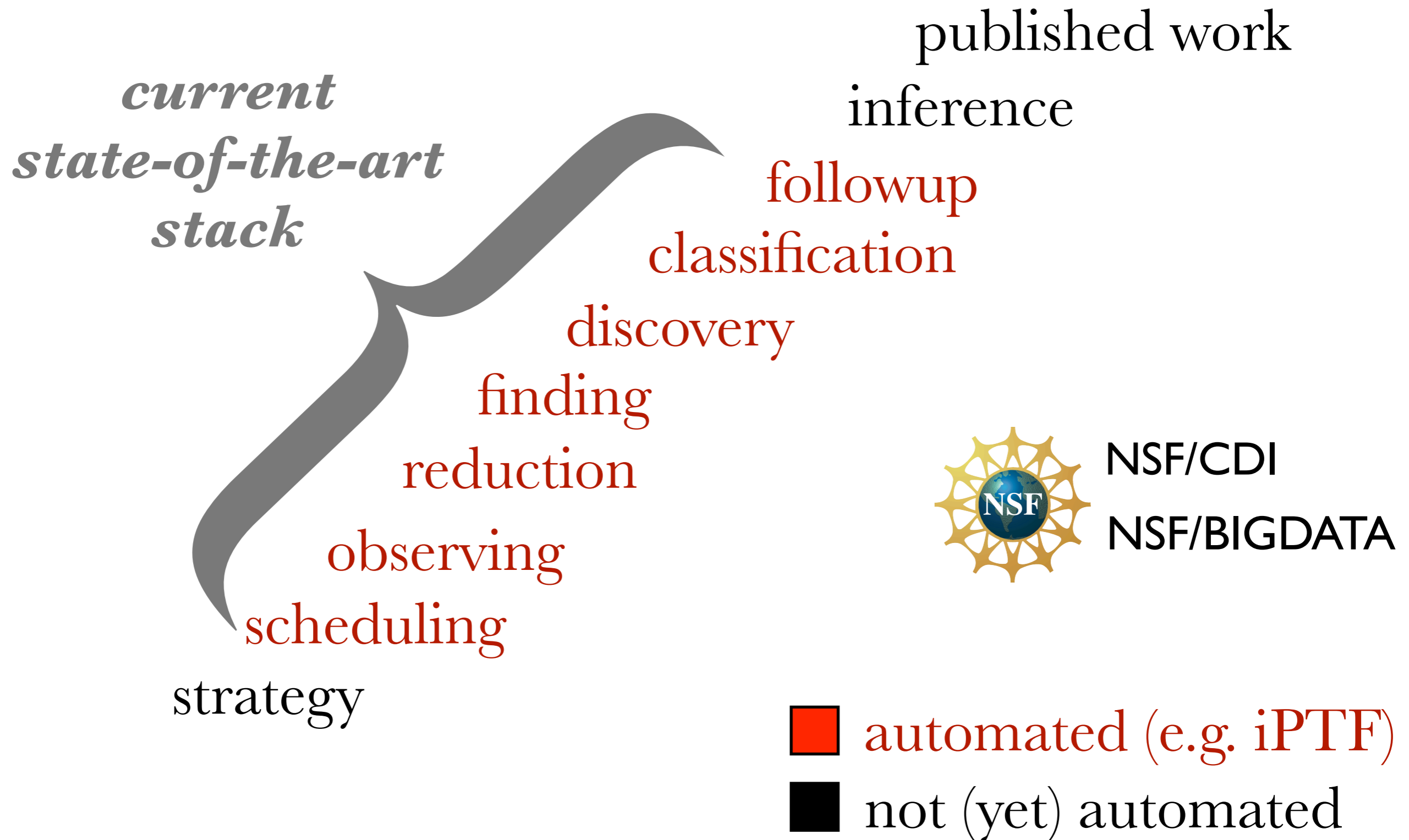
1 billion stars observed ~70 times over 5 years

Will observe 20K supernovae

Many other astronomical surveys are already producing data:

SDSS, iPTF, CRTS, Pan-STARRS, Hipparcos, OGLE, ASAS, Kepler, LINEAR, DES etc.,

# Towards a Fully Automated Scientific Stack for Transients



- ▶ Built & Deployed Real-time ML framework, discovering  $>10,000$  events in  $> 10$  TB of imaging  
→ 50+ journal articles
- ▶ Built Probabilistic Event classification catalogs with innovative active learning

<http://timedomain.org>

*Our ML framework found the  
Nearest Supernova in 3 Decades ..*



[https://www.nsf.gov/news/news\\_summ.jsp?cntn\\_id=122537](https://www.nsf.gov/news/news_summ.jsp?cntn_id=122537)

# Data-Centric Coursework, Bootcamps, Seminars, & Lecture Series



BDAS: Berkeley Data  
Analytics Stack  
[Spark, Shark, ...]



parallel  
programming  
bootcamp

## ...and entire degree programs

datascience@berkeley

### Master of Information and Data Science

The UC Berkeley School of Information invites you to learn more about the only professional data science degree delivered fully online. Answer the simple questions below to request more information.

10% Complete

About MIDS

Why Data Science?

Online Experience

### Earn a Master of Information and Data Science—Online

Now you can earn a degree in data science from anywhere in the world. The UC Berkeley School of Information offers the only professional Master of Information and Data Science

# Data-Centric Coursework, Bootcamps, Seminars, & Lecture Series

**Taught by CS/Stats**

**Aimed at Engineers &  
Programmers Heading  
Toward Industry**

datascience@berkeley

## Master of Information and Data Science

The UC Berkeley School of Information invites you to learn more about the only professional data science degree delivered fully online. Answer the simple questions below to request more information.



About MIDS

Why Data Science?

Online Experience

## Earn a Master of Information and Data Science—Online

Now you can earn a degree in data science from anywhere in the world. The UC Berkeley School of Information offers the only professional Master of Information and Data Science

# Python Bootcamps at Berkeley



**2010:** 85 campers



**2012a:** 135 campers



# **a modern superglue computing language for science**

- ▶ high-level scripting language
- ▶ open source, huge & growing community in academia & industry
- ▶ Just in time compilation but also fast numerical computation
- ▶ Extensive interfaces to 3rd party frameworks

*A reasonable lingua franca for scientists...*

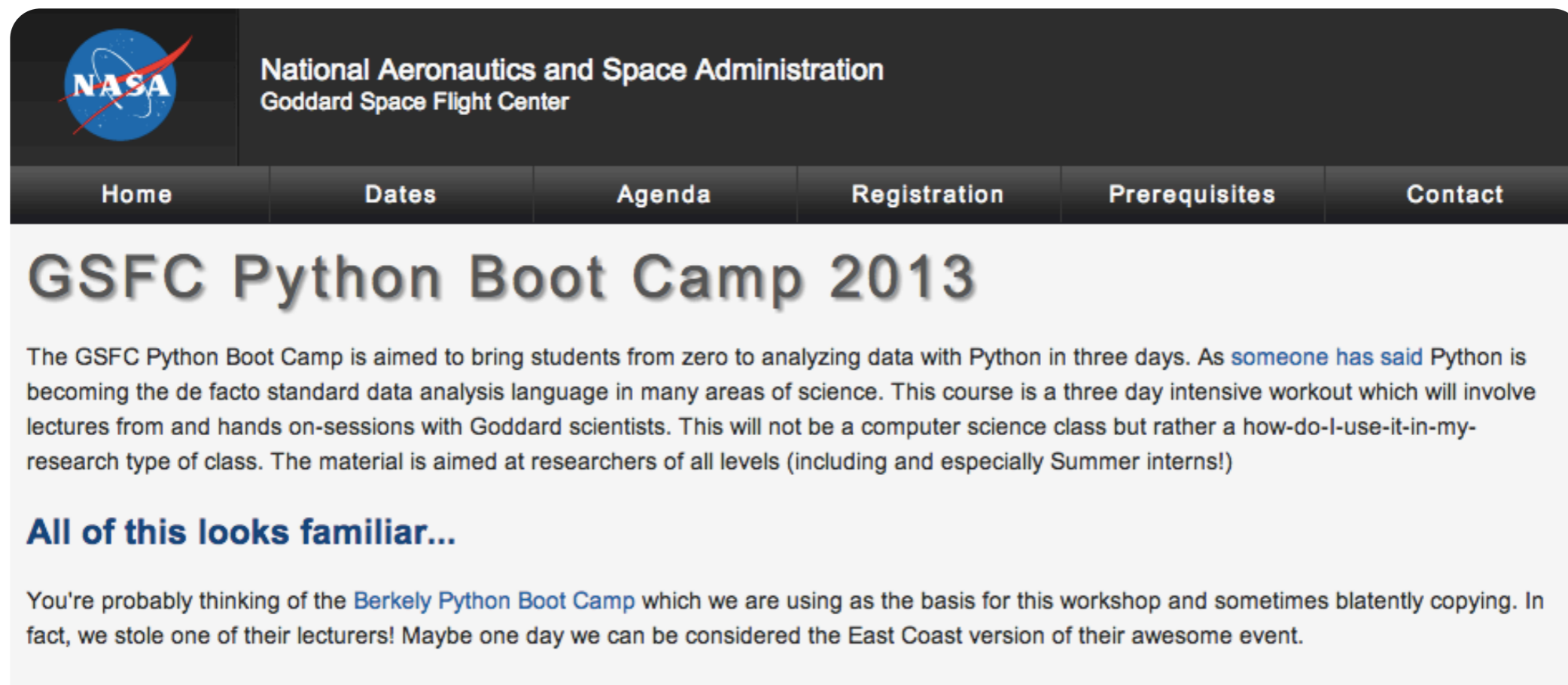
# Python Bootcamps at Berkeley

**2012b:** 210 campers

**2013a:** 253 campers

- ▶ 3 days of live/archive streamed lectures
- ▶ all open material in GitHub
- ▶ widely disseminated (e.g., @NASA)
- ▶ funded (~\$18k) by the Vice Chancellor for Research & NSF (BIGDATA)

<http://pythonbootcamp.info>



The screenshot shows the website for the GSFC Python Boot Camp 2013. The header features the NASA logo and the text "National Aeronautics and Space Administration" and "Goddard Space Flight Center". Below the header is a navigation bar with links: Home, Dates, Agenda, Registration, Prerequisites, and Contact. The main content area has the title "GSFC Python Boot Camp 2013" and a paragraph describing the camp's purpose: "The GSFC Python Boot Camp is aimed to bring students from zero to analyzing data with Python in three days. As someone has said Python is becoming the de facto standard data analysis language in many areas of science. This course is a three day intensive workout which will involve lectures from and hands on-sessions with Goddard scientists. This will not be a computer science class but rather a how-do-I-use-it-in-my-research type of class. The material is aimed at researchers of all levels (including and especially Summer interns!)"

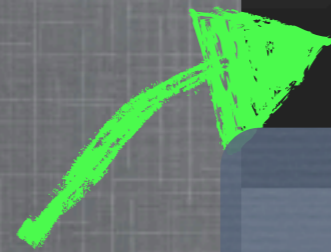
**All of this looks familiar...**

You're probably thinking of the Berkely Python Boot Camp which we are using as the basis for this workshop and sometimes blatantly copying. In fact, we stole one of their lecturers! Maybe one day we can be considered the East Coast version of their awesome event.



# Python Computing for Science

Undergraduate/Graduate Seminar Course at UC Berkeley  
(AY 250)



Part of the  
**Designated  
Emphasis in  
Computational Science &  
Engineering**  
at Berkeley

parallelism

interfacing to other languages

Bayesian inference & MCMC

visualization

hardware control

machine learning

database interaction

user interface & web frameworks

timeseries & numerical

computing

Date	Content	Instructor
Aug 27	Advanced Python Language Concepts (prepared towards Bootcamp graduates) notebook   lecture (PDF)	Josh
Sep 10	Advanced versioning, application building, debugging & testing	Isaac
Sep 17	(matplotlib) Advanced plotting and data visualization, mayavi	Stefan/Paul
Sep 24	Database interaction	Berian/Brad
Oct 1	interacting with the world (xml-rpc, urllib, sending and receiving) talking to computers (notebook)   audio IO (notebook)   lecture (part 1)   lecture (part 2)	Josh
Oct 8	timeseries & numerical computing	Stefan/Joey
Oct 15	distributed computing, large datasets (HDF5)	Josh
Oct 22	GUI (Tkinter, GTK, Traits)	Josh

“Parallel Image  
Reconstruction from  
Radio Interferometry  
Data”

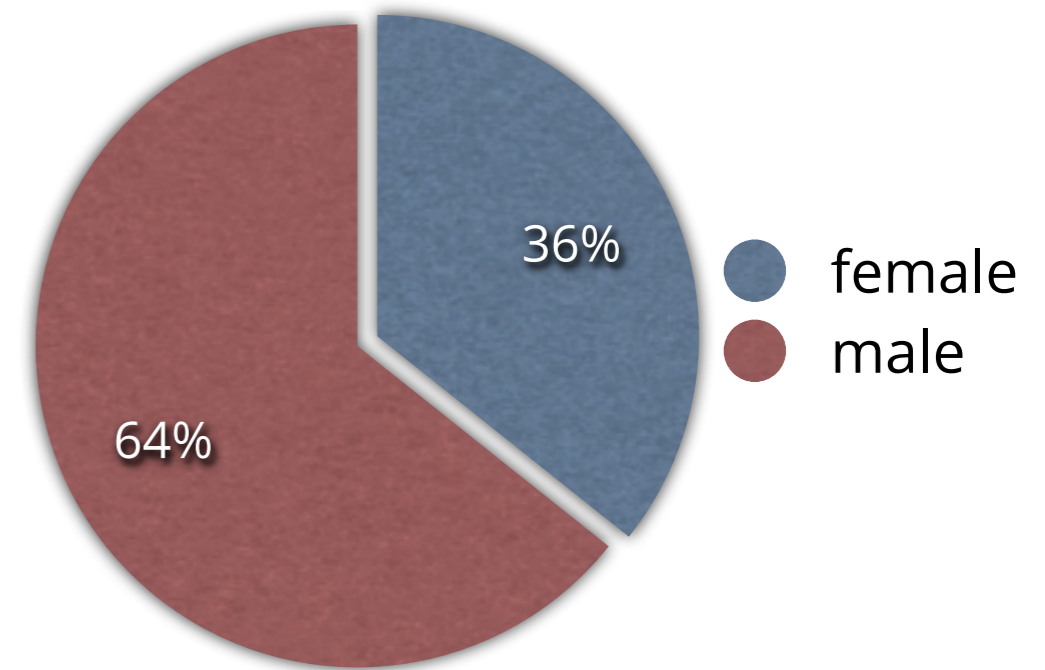
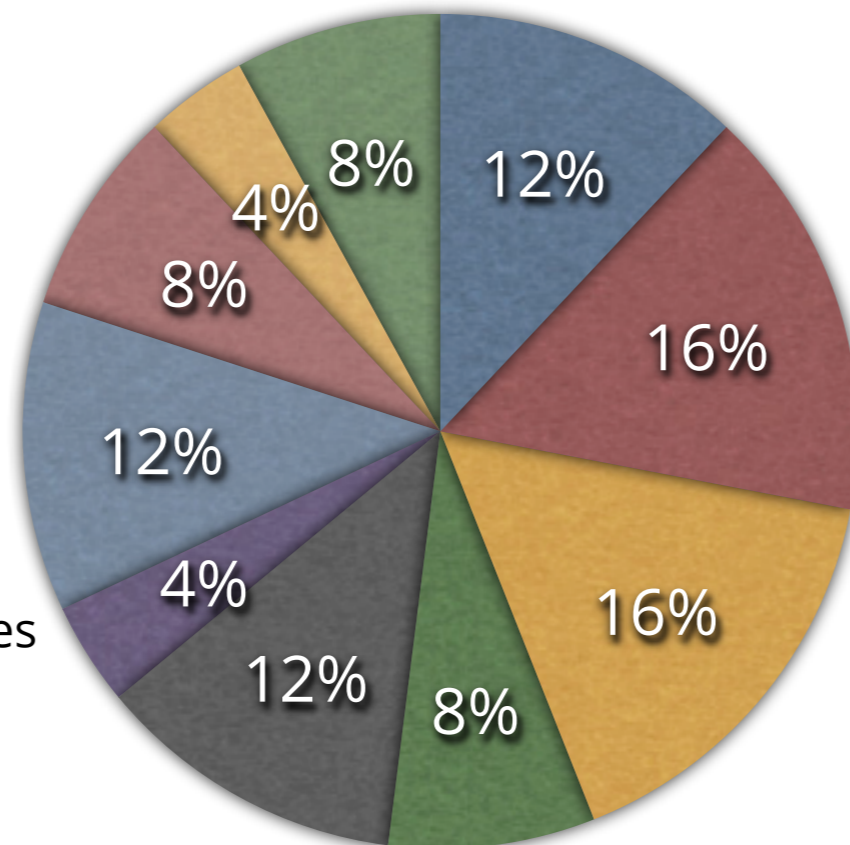
“Realtime Prediction of Activity  
Behavior from Smartphone”

“Graph Theory Analysis of  
Growing Graphs”

<http://mb3152.github.io/Graph-Growth/>

“Bus Arrival  
Time Prediction  
in Spain”

- Psychology
- Astronomy
- Neuroscience
- Biostatistics
- Physics
- Chemical Engineering
- ISchool
- Earth and Planetary Sciences
- Industrial Engineering
- Mechanical Engineering



# Prevalence of Earth-size planets orbiting Sun-like stars

Erik A. Petigura<sup>a,b,1</sup>, Andrew W. Howard<sup>b</sup>, and Geoffrey W. Marcy<sup>a</sup>

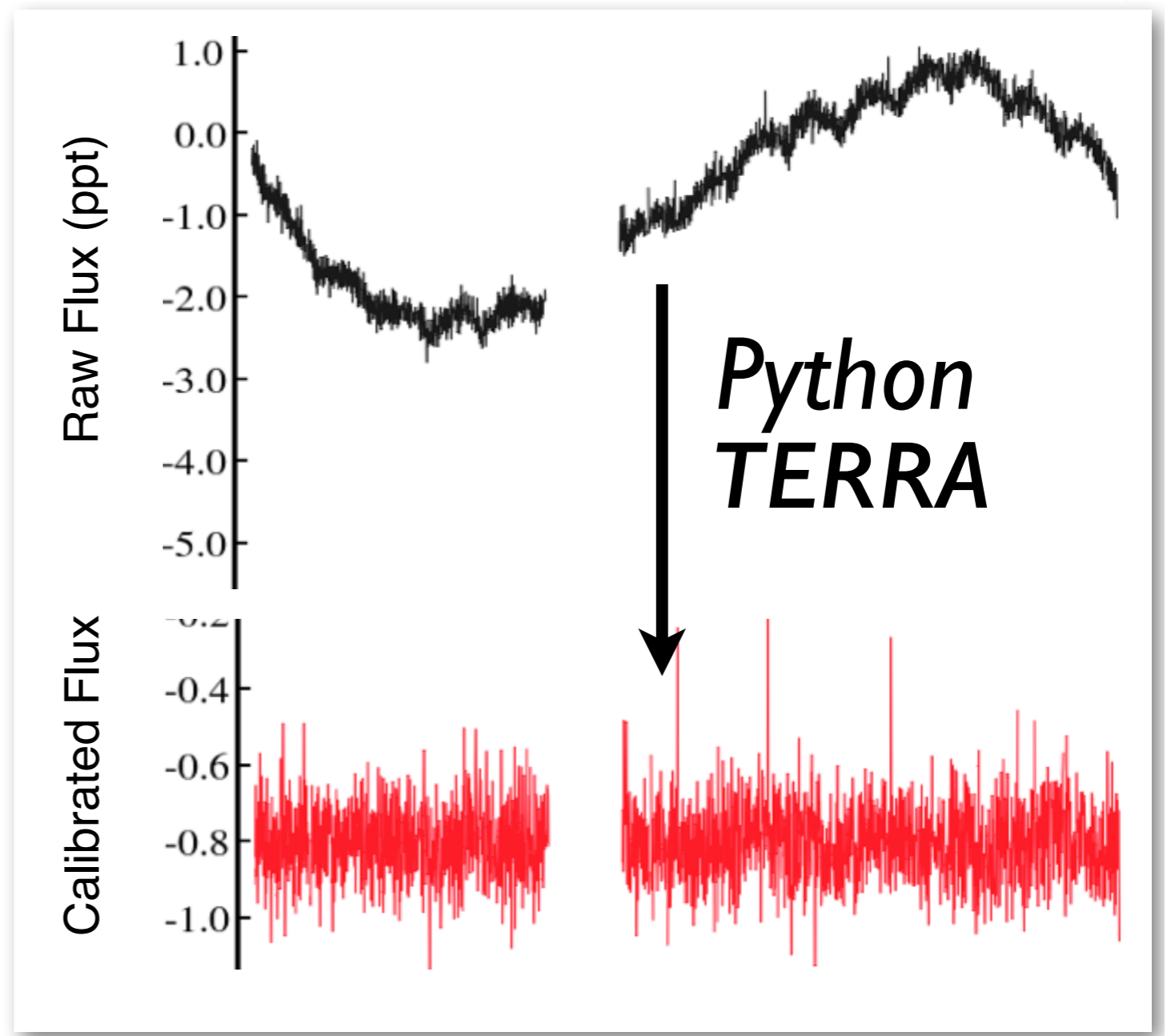
PNAS [2014]

<sup>a</sup>Astronomy Department, University of California, Berkeley, CA 94720; and <sup>b</sup>Institute for Astronomy, University of Hawaii at Manoa, Honolulu, HI 96822



Erik Petigura  
Berkeley Astro  
Grad Student

**Bootcamp/  
Seminar Alum**



DOE/NERSC computation

“Are we alone in the universe? What makes up the missing mass of the universe? ... And maybe the biggest question of all: How in the wide world can you add \$3 billion in market capitalization simply by adding .com to the end of a name?”

President William Jefferson Clinton  
Science and Technology Policy Address  
21 January 2000

“Add Data Science or Big Data to your course name to increase enrollment by tenfold.”

Joshua Bloom  
Just Now



# Python for Data Science @ Berkeley [Sept 2013]



**► Where do Bootcamps & Seminars fit into traditional domain science curricula?**

- formal coursework competes with research obligations for graduate students

**► Are they too vocational/practical for higher Ed?**

**► Who should teach them & how do we credit them?**

# Undergraduate & Graduate Training Mission

## Thinking *Data Literacy* before Thinking *Big Data Proficiency*



first this...



...then this.

# Undergraduate & Graduate Training Mission

## Thinking *Data Literacy* before Thinking *Big Data Proficiency*

### Statistical Inference

#### Data analysis recipes: Fitting a model to data\*

David W. Hogg

*Center for Cosmology and Particle Physics, Department of Physics,  $\Lambda$   
Max-Planck-Institut für Astronomie, Heidelberg*

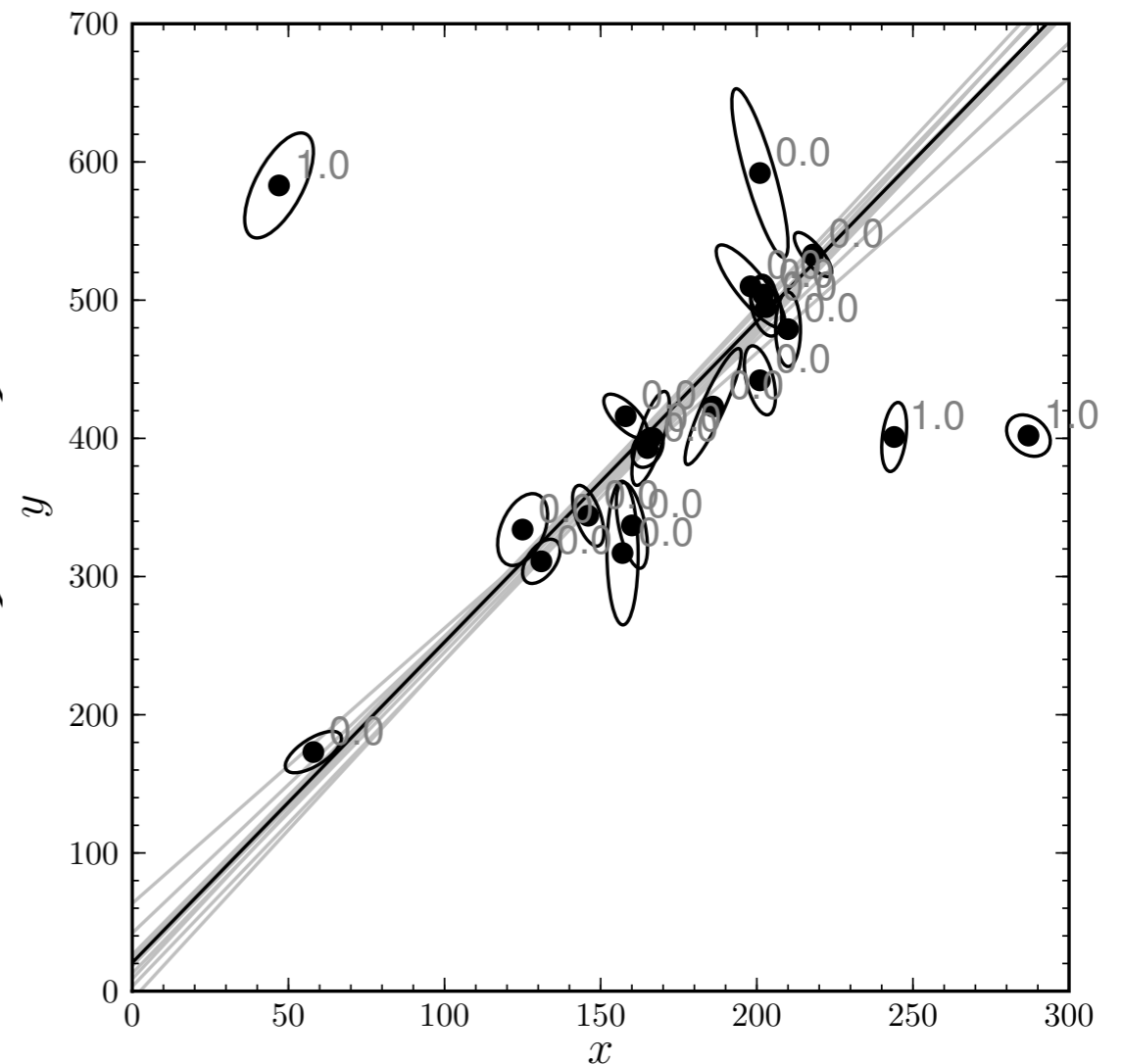
Jo Bovy

*Center for Cosmology and Particle Physics, Department of Physics,  $\Lambda$*

Dustin Lang

*Department of Computer Science, University of Toronto  
Princeton University Observatory*

arXiv:1008.4686v1



# Undergraduate & Graduate Training Mission

## Thinking *Data Literacy* before Thinking *Big Data Proficiency*

### Versioning & Reproducibility

*“Recently, the scientific community was shaken by reports that a troubling proportion of peer-reviewed preclinical studies are not reproducible.” McNutt, 2014*

<http://www.sciencemag.org/content/343/6168/229.summary>

- Git has emerged as the de facto versioning tool
- Berkeley Common Environment (BCE) Software Stack
- “Reproducible and Collaborative Statistical Data Science” (Statistics 157: P. Stark)
- Next up: Versioning (big) data?

# Exploring the Lorenz System of Differential Equations

In this Notebook we explore the Lorenz system of differential equations:

```


$$\begin{aligned} \dot{x} &= \sigma(y-x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy \end{aligned}$$


```

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters ( $\sigma$ ,  $\beta$ ,  $\rho$ ) are varied.

## Imports

First, we import the needed things from IPython, NumPy, Matplotlib and SciPy.

```
In [ ]: %matplotlib inline
```

```
In [ ]: from IPython.html.widgets import interact, interactive
        from IPython.display import clear_output, display, HTML
```

```
In [ ]: import numpy as np
```

IPython Creator



Fernando Pérez

# Exploring the Lorenz System of Differential Equations

In this Notebook we explore the Lorenz system of differential equations:

```


$$\begin{aligned} \dot{x} &= \sigma(y-x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy \end{aligned}$$


```

This is one of the classic systems in non-linear differential equations. It exhibits a range of different behaviors as the parameters ( $\sigma$ ,  $\beta$ ,  $\rho$ ) are varied.

## Imports

First, we import the needed things from IPython, NumPy, Matplotlib and SciPy.

```
In [ ]: %matplotlib inline
```

```
In [ ]: from IPython.html.widgets import interact, interactive
from IPython.display import clear_output, display, HTML
```

```
In [ ]: import numpy as np
```

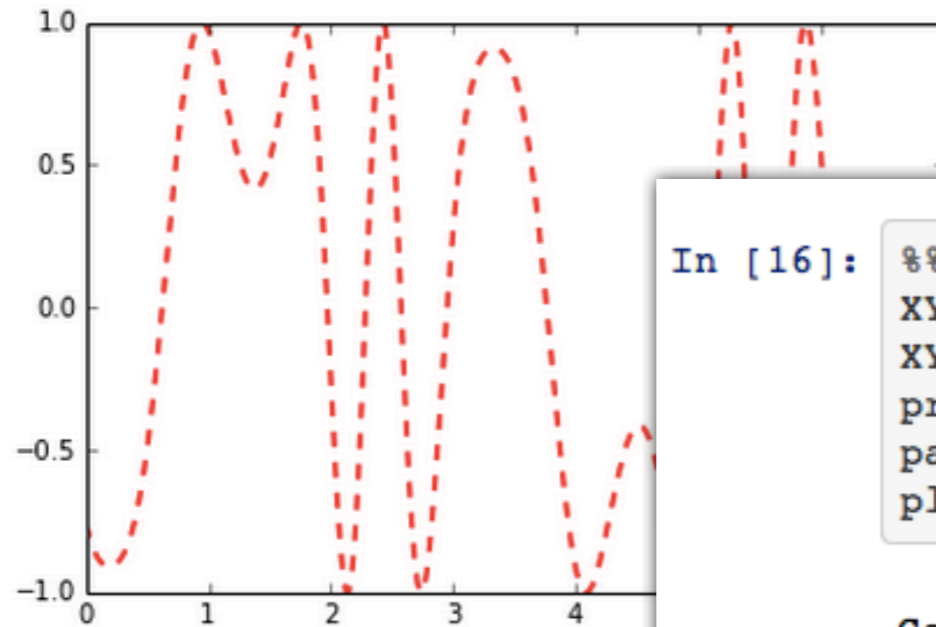
IPython Creator



Fernando Pérez

```
In [6]: %%julia
# Note how we mix numpy and julia:
x = linspace(0,2*pi,1000); # use the julia linspace
y = sin(3*x + 4*np.cos(2*x)); # use the numpy cosine and julia sine
plt.plot(x, y, color="red", linewidth=2.0, linestyle="--")
```

```
Out[6]: [<matplotlib.lines.Line2D at 0x3a80150>]
```



Julia array operations work and return nati

# Julia

IPython notebook  
is ~agnostic to the  
backend

```
In [16]: %%R -i X,Y -o XYcoef
XYlm = lm(Y~X)
XYcoef = coef(XYlm)
print(summary(XYlm))
par(mfrow=c(2,2))
plot(XYlm)
```

Call:

```
lm(formula = Y ~ X)
```

Residuals:

1	2	3	4	5
-0.2	0.9	-1.0	0.1	0.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.2000	0.6164	5.191	0.0139 *
X	0.9000	0.2517	3.576	0.0374 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7958 on 3 degrees of freedom

Multiple R-squared: 0.81, Adjusted R-squared: 0.7467

F-statistic: 12.79 on 1 and 3 DF, p-value: 0.03739

# R

**“novelty<sup>2</sup> problem”**

**Extra Burden for Forefront Scientists**

Established CS/Stats/Math *in Service*  
of novelty in domain science

*vs.*

Novelty in domain science driving &  
informing novelty in CS/Stats/Math

<https://medium.com/tech-talk/dd88857f662>

# Berkeley Institute for Data Sciences (BIDS)

- ▶ Physical Space & New Entity dedicated to the Moore/Sloan Data Science principles
- ▶ Goal: rich resource and ecosystem for domain scientists to connect & collaborate with methodologists

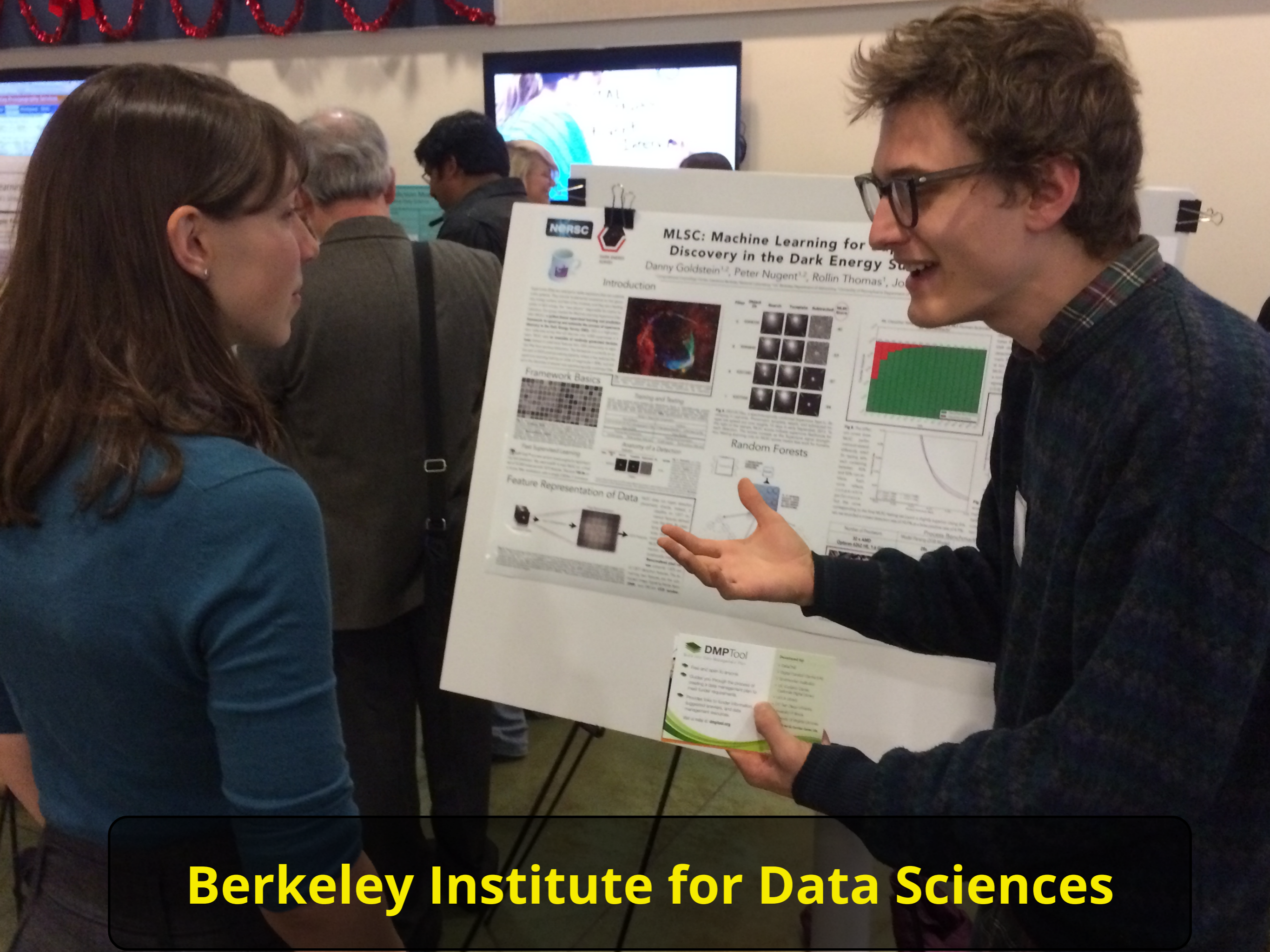


ALFRED P. SLOAN  
FOUNDATION

“Bold new partnership launches to harness potential of data scientists and big data”



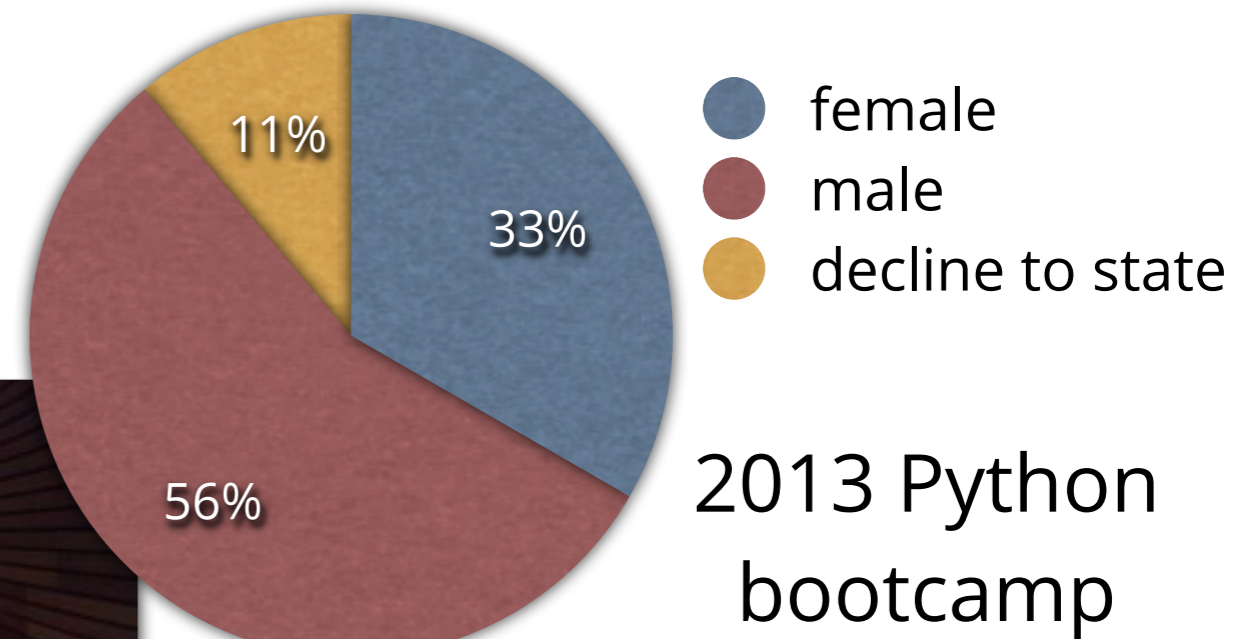
**Berkeley Institute for Data Sciences**



**Berkeley Institute for Data Sciences**

# Towards an Inclusive Ecosystem

## Expanding Participation Among Underrepresented Groups



- 2013 Python Seminar: 36% women
- 2013 AMP Camp: < 5% women at
- This Workshop: 2 women out of 22 speakers

# Summary

- ▶ **Domain Science increasingly dependent upon methodological competencies**
- ▶ **Higher-Ed Role of such training still TBD**
  - **formal courses competes for time**
- ▶ ***Data Literacy* before *Big Data Proficiency***
- ▶ **Need to create inclusive, collaborative environments bridging domains & methodologies**

Thank you.

@profjsb