Emerging Needs & Opportunities in Data-Intensive Domains: Astronomy

Joshua Bloom (@profjsb*)*
Gordon & Betty Moore Foundation Data-Driven Investigator
UC Berkeley, Astronomy

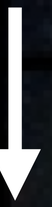
Presented to the National Academy of Sciences Roundtable on the "Intersection of Domain Expertise and Data Science"

Irvine, CA Mar 20, 2017

Democratizing Trends in Physical Sciences

Data	Decreasing cost to obtain, move, store	→	open data, more freely shared
Compute	Decreasing cost, increasing specialty	→	more accessible
Technology/ Methodology	Algorithmic innovation, software tooling	→	open source

Democratizing Trends in Physical Sciences



Competition for superior inferential capabilities

- Statistical / machine learning capabilities
- Computational prowess
- Ability to innovate methodologies

Outline

Notable data-driven success with domain + methodologies

particle physics & gravitational wave astronomy

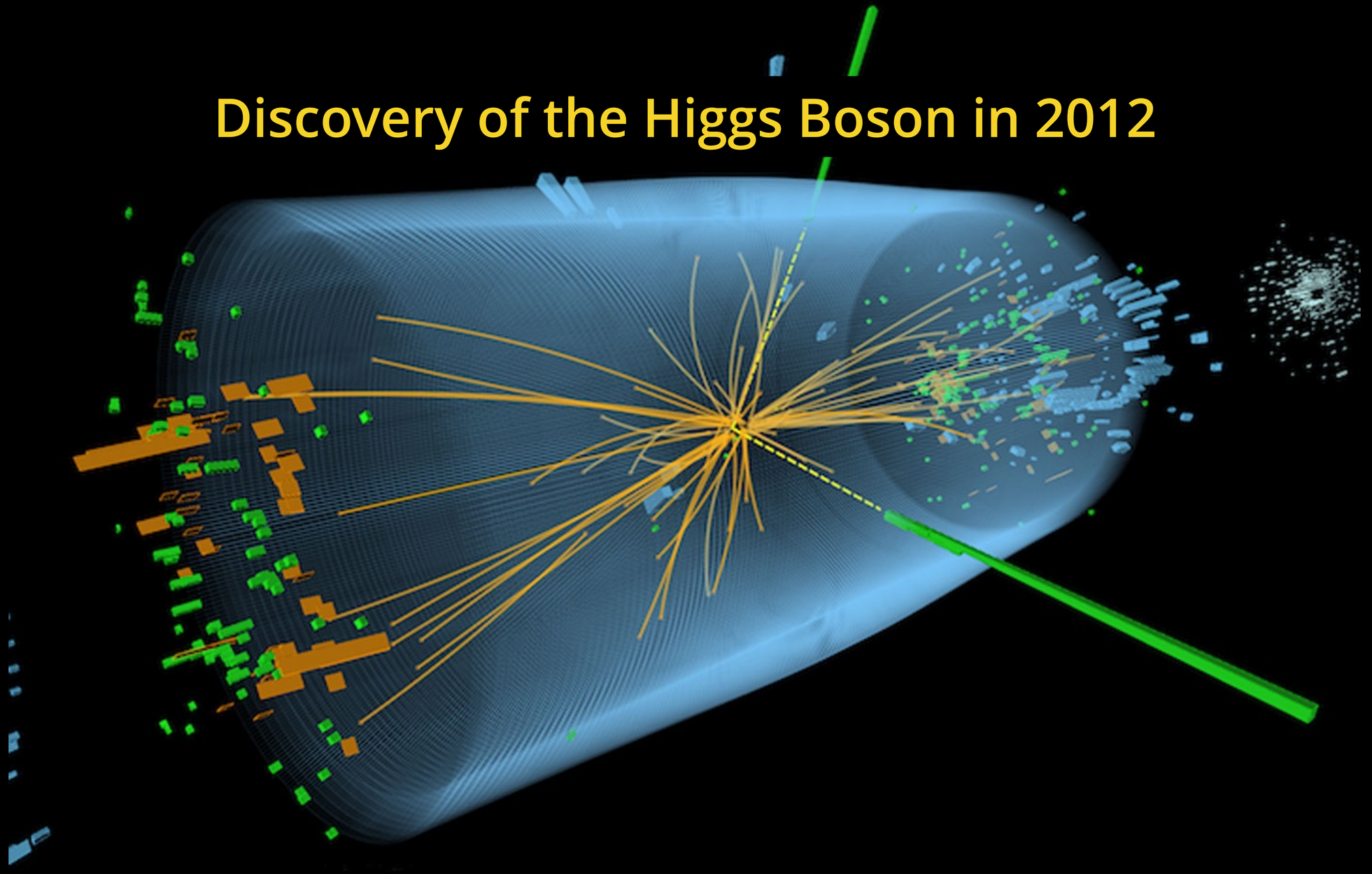
Astronomy's Data deluge

- who wins?
- examples of data science impact in astro

Educational challenges given these trends

Symbiotic relationship between data domains & methodological sciences

Discovery of the Higgs Boson in 2012



Data & Computation

- 600M collisions/s; only ~100/s are interesting
- 600 GB/s raw → 25 GB/s processed
- 3.5 MW data centers, Pb/day processing

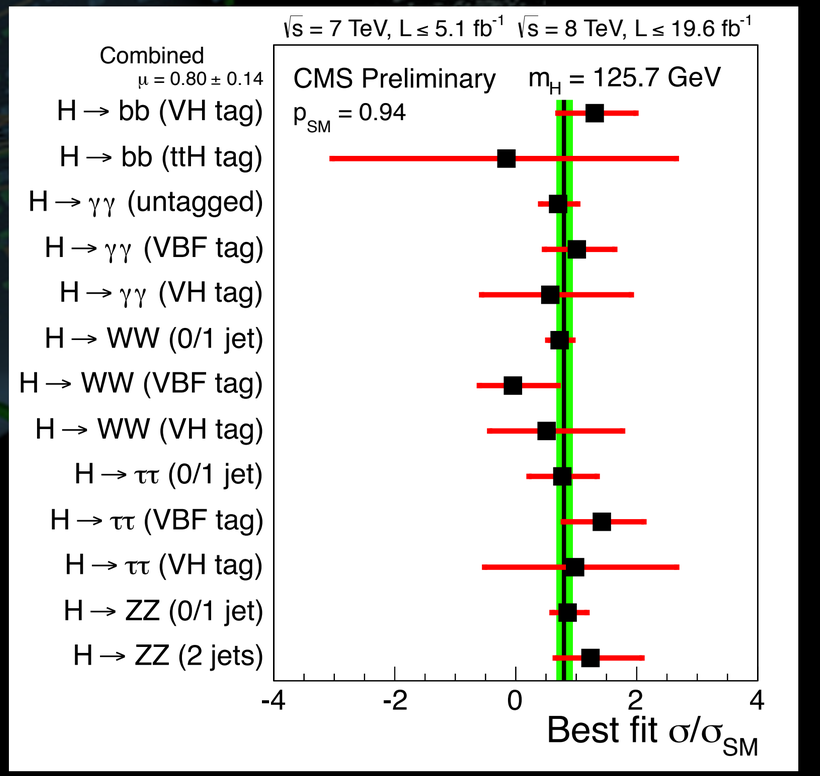
CMS PAS HIG-13-005

Inference

- "HEP phystatisticians"
- Statistics combining

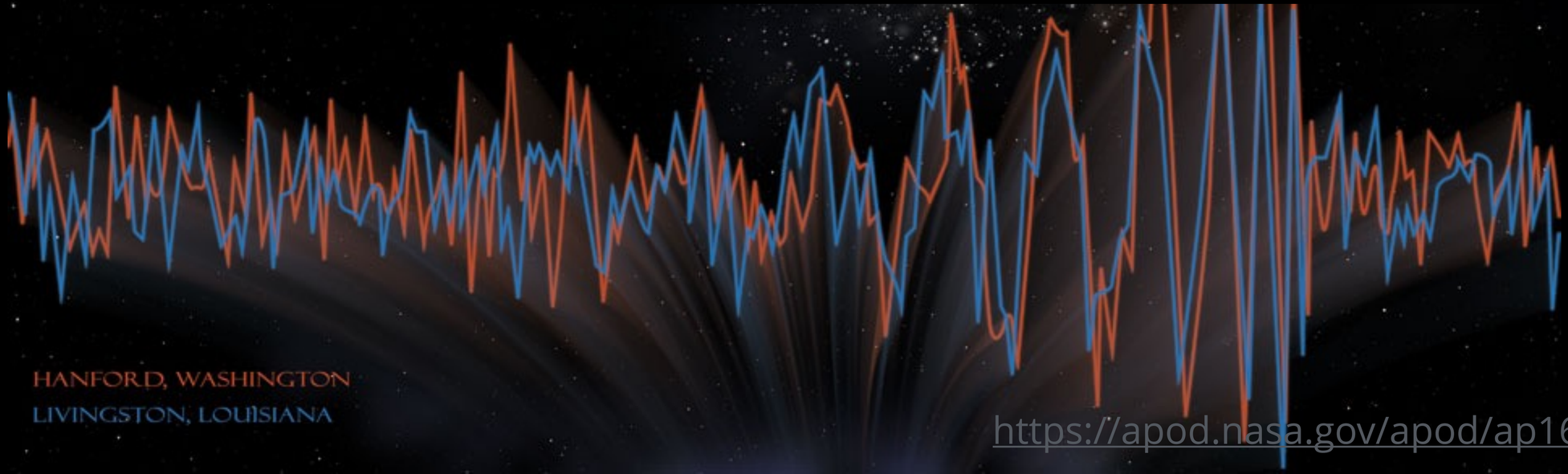
standard model (theoretical underpinning)
& observations

Horvath 1310.6839



https://home.cern/about/computing

Direct Detection of Gravitational Waves in 2016



https://apod.nasa.gov/apod/ap160211.html

LIGO, NSF, Illustration: A. Simonnet (SSU)

Data & Computation

- aLIGO Data rate: 81 MB/s, 2PB/yr
- 10^5 templates to continually match (32TB) at **5 TFLOPS**
- 20k CPUs, 12 PB cluster

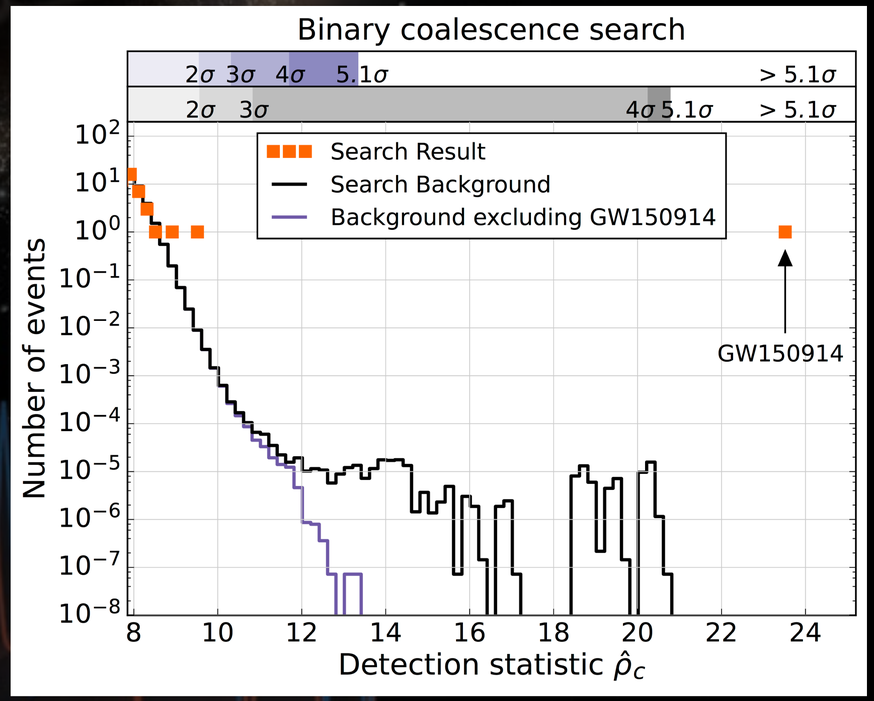
Shoemaker M060056-01.doc

INSPIRAL

Inference

- novel long-tail statistics
- Need rapid identification for multi wavelength followup

e.g. Cannon et al. 2012



HANFORD, WASHINGTON
LIVINGSTON, LOUISIANA

https://apod.nasa.gov/apod/ap160211.html

Lessons from Large-Scale Physics Experiments

- Strongly focused on limited number of high-impact discoveries
- Residual inventions (e.g., WWW, grid computing)
- **Diverse Teams**: physicists collaborating with methodological specialists
- Large collaborations producing science results **themselves**

Astronomy's Data Deluge

Large Synoptic Survey Telescope (LSST) - 2020

Light curves for 800M sources every 3 days

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

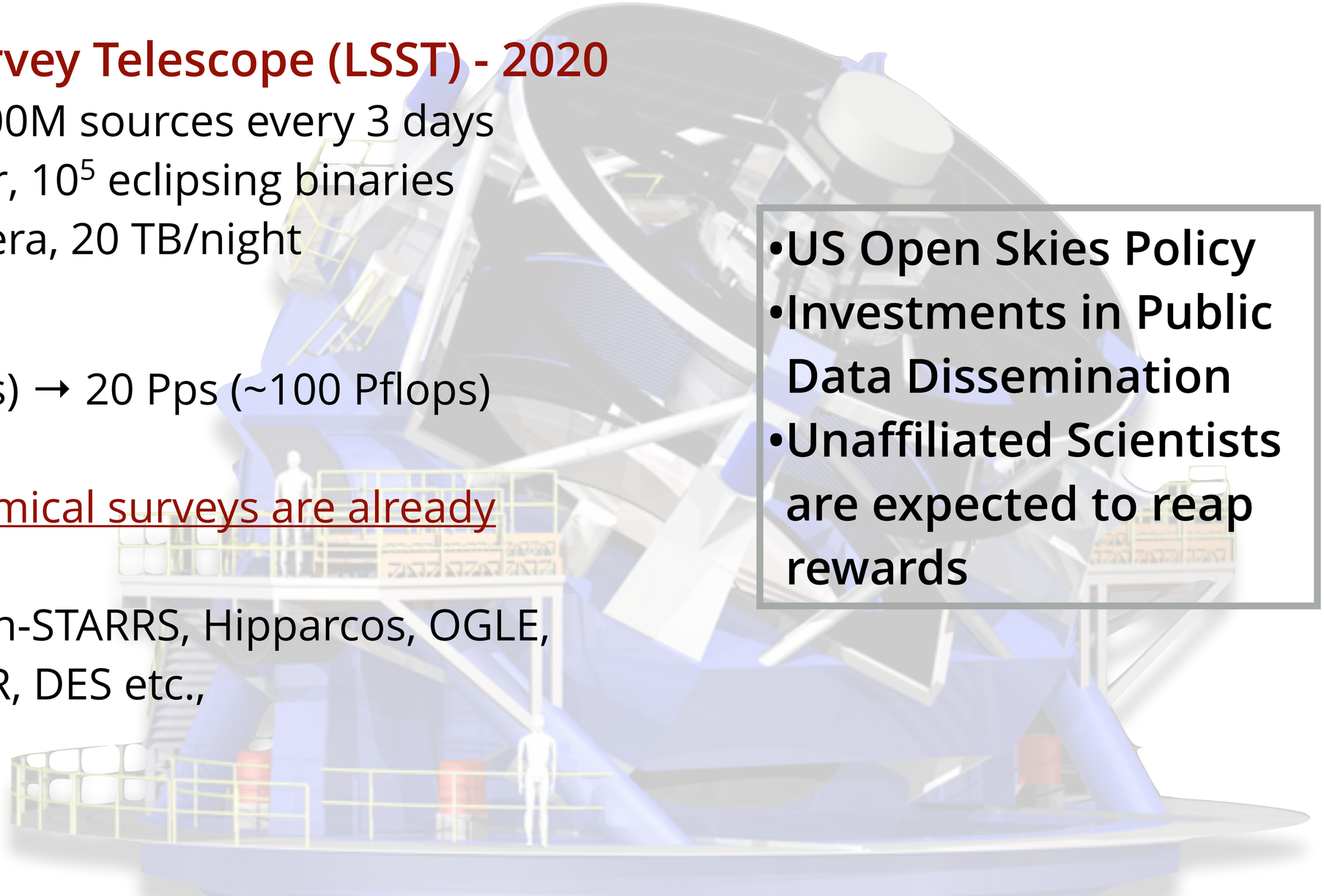
3.2 gigapixel camera, 20 TB/night

LOFAR & SKA

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

Many other astronomical surveys are already producing data:

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

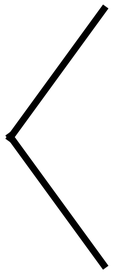


- **US Open Skies Policy**
- **Investments in Public Data Dissemination**
- **Unaffiliated Scientists are expected to reap rewards**

If anyone can get the Data, who wins?

domain expertise → **She who asks the right questions**

\+

data science 

She who answer questions better & faster than others:

- computational access
- methodological inference (e.g., machine learning)
- better story telling, dissemination of results
- reproducibility (acceptance)

Some of my work…

- Built & Deployed robust, real-time supervised machine learning framework, discovering >10,000 events in > 10 TB of imaging
  → 75+ journal articles

- Built probabilistic source classification catalogs on public archives with novel active learning approach

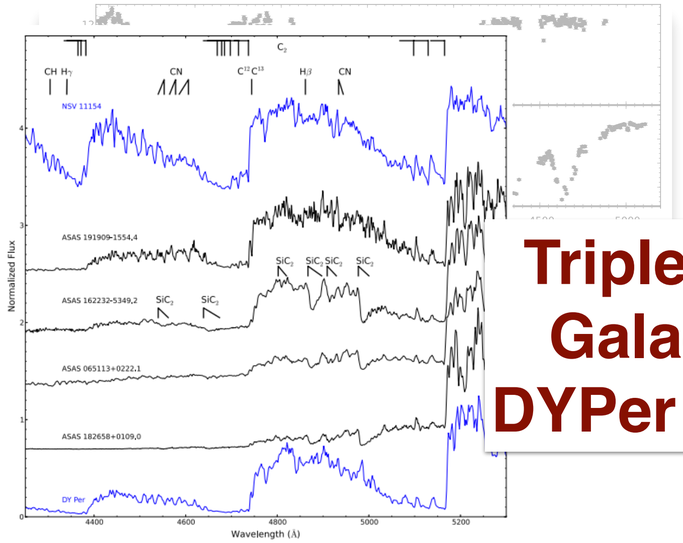
Our ML framework found the Nearest Supernova in 3 Decades ..

https://www.nsf.gov/news/news_summ.jsp?cntn_id=122537

Probabilistic Classification of 50k+ Variable Stars

15 “RCB/DYP” candidates

8 new discoveries



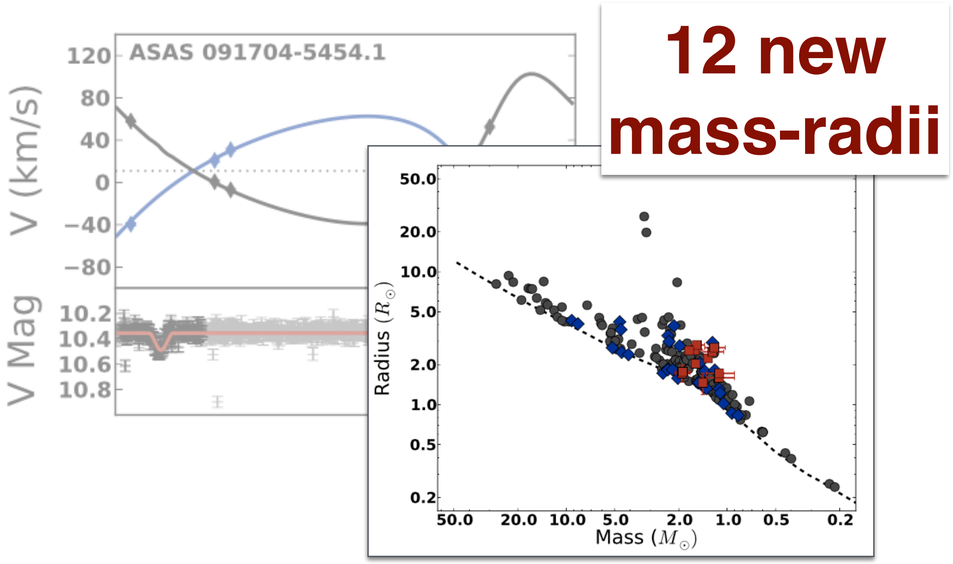
Triple # of Galactic DYPer Stars

DISCOVERY OF BRIGHT GALACTIC R CORONAE BOREALIS AND DY PERSEI VARIABLES: RARE GEMS MINED FROM ACVS

Miller, Richards, JSB,..ApJ 2012

106 “DEB” candidates

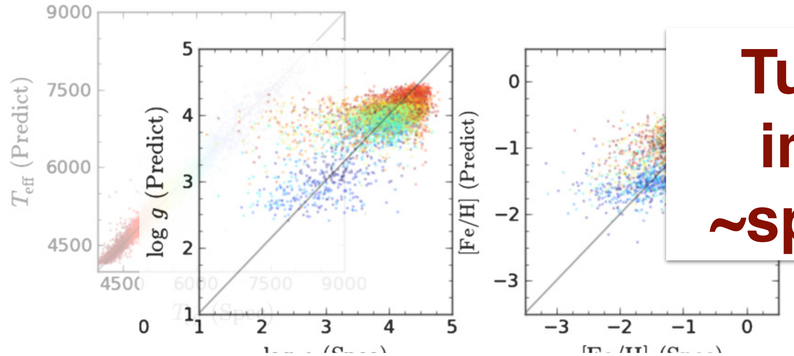
5400 Spectroscopic Targets



12 new mass-radii

The Highly-Eccentric Detached Eclipsing Binaries in ACVS and MACC

Shivvers,JSB,Richards MNRAS,2014



Turn synoptic imagers into ~spectrographs

A MACHINE-LEARNING METHOD TO INFER FUNDAMENTAL STELLAR PARAMETERS FROM PHOTOMETRIC LIGHT CURVES

Miller, JSB, Richards,..ApJ 2015

Some unsupervised work…

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

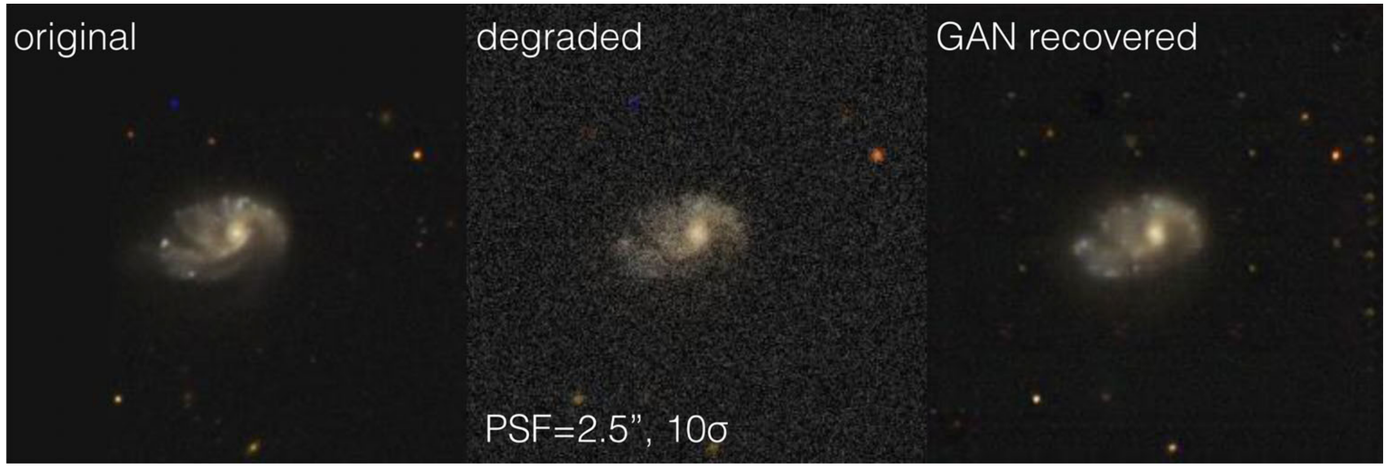
MNRAS **467,** L110–L114 (2017)
Advance Access publication 2017 January 20

doi:10.1093/mnrasl/slx008

Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit

Kevin Schawinski,[1]★ Ce Zhang,[2]★ Hantian Zhang,[2] Lucas Fowler[1] and Gokula Krishnan Santhanam[2]

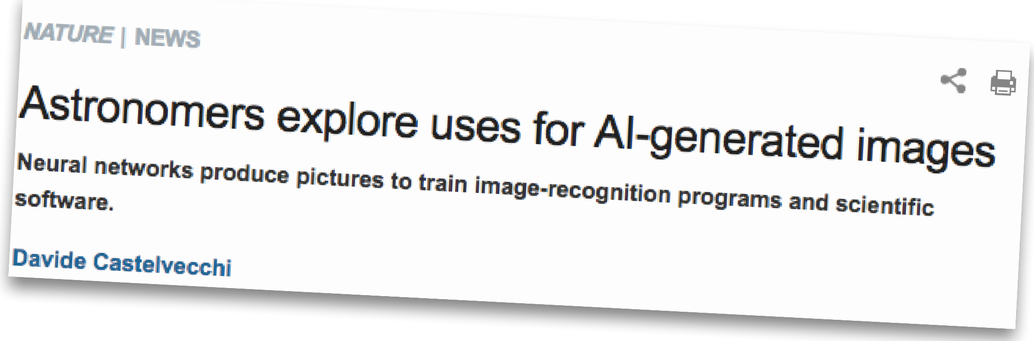
[1] *Institute for Astronomy, Department of Physics, ETH Zurich, Wolfgang-Pauli-Strasse 27, CH-8093 Zürich, Switzerland*
[2] *Systems Group, Department of Computer Science, ETH Zurich, Universitätstrasse 6, CH-8006 Zürich, Switzerland*



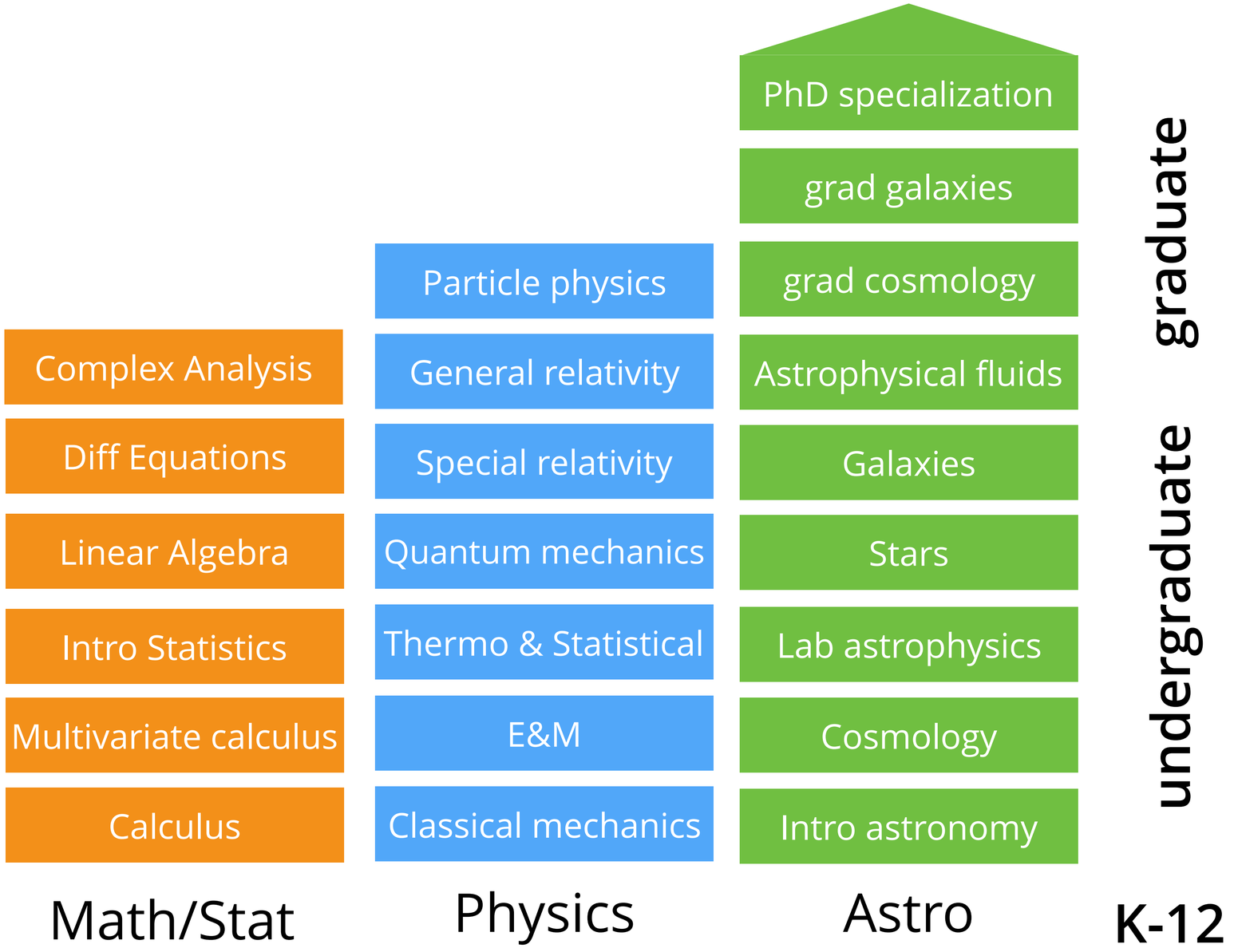
Machine Learning Is Bringing the Cosmos Into Focus

The Atlantic

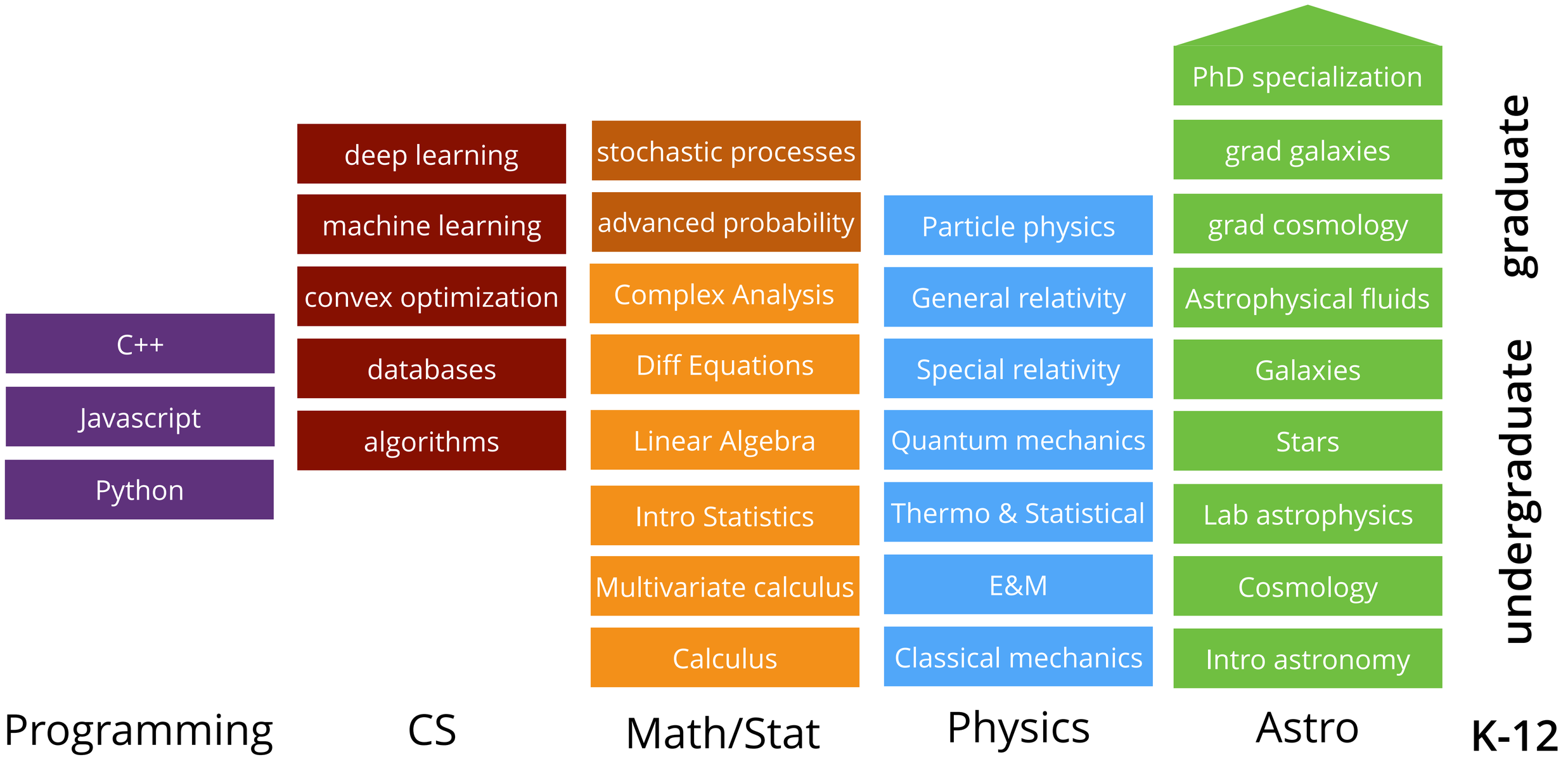
Training neural networks to identify galaxies could forever change humanity's perspective of the universe.



Challenge of Data-Driven Domain Education Stack



Challenge of Data-Driven Domain Education Stack



Python Computing for Data Science

Graduate course at UC Berkeley

parallelism

interfacing to other languages

Bayesian inference & MCMC

visualization

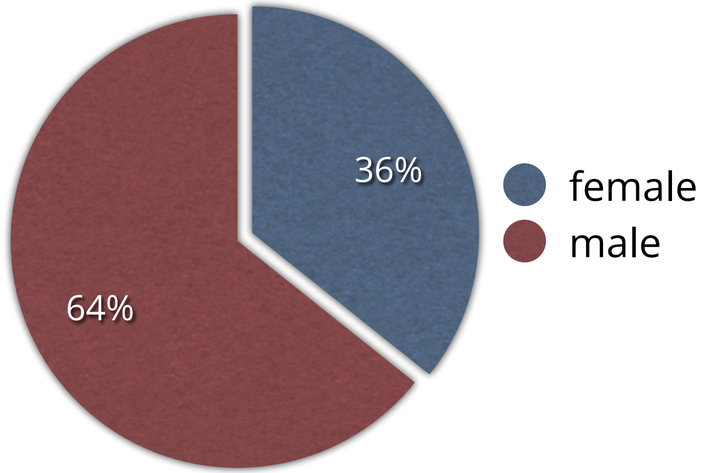
hardware control

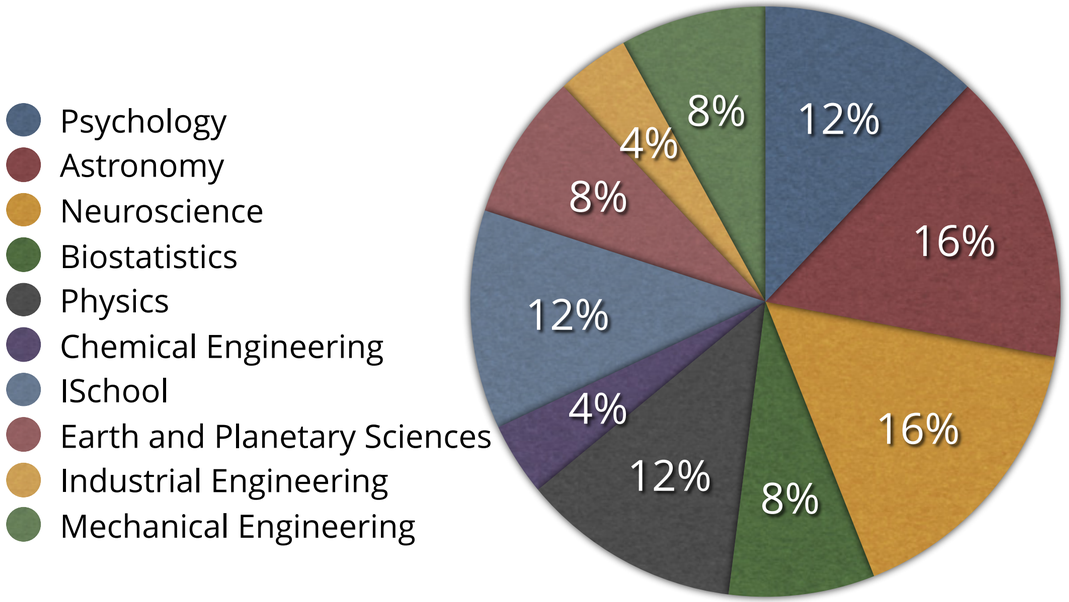
database interaction

user interface & web frameworks

timeseries & numerical computing

machine learning





http://github.com/profjsb/python-seminar

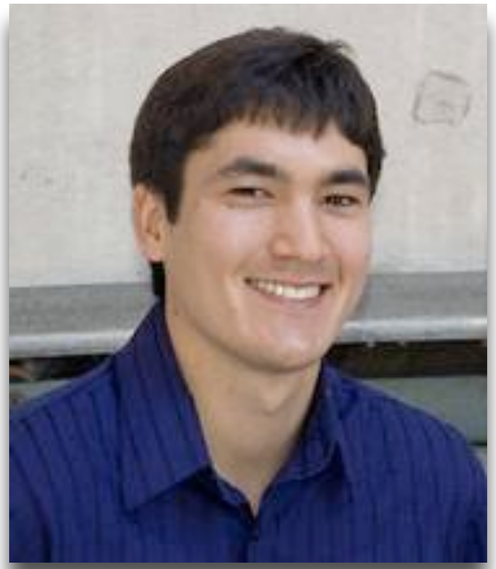
2013 statistics

Prevalence of Earth-size planets orbiting Sun-like stars

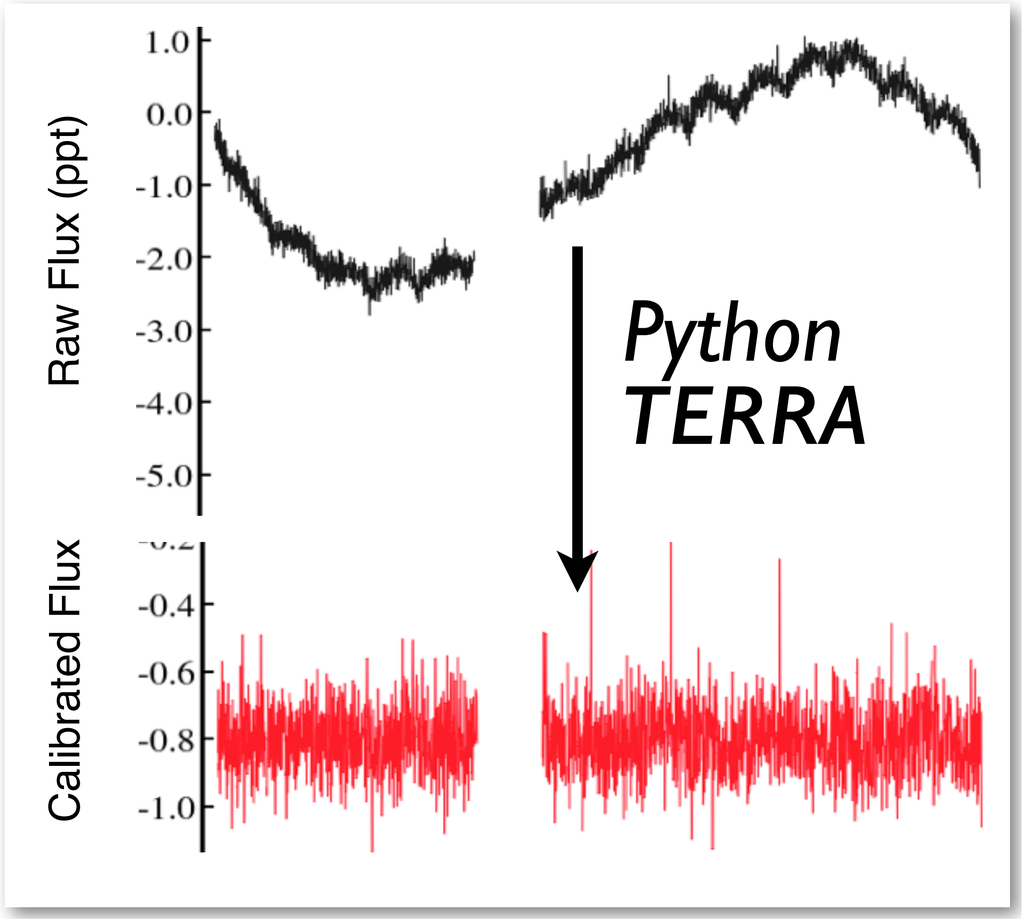
Erik A. Petigura[a,b,1], Andrew W. Howard[b], and Geoffrey W. Marcy[a]

PNAS [2014]

[a]Astronomy Department, University of California, Berkeley, CA 94720; and [b]Institute for Astronomy, University of Hawaii at Manoa, Honolulu, HI 96822



Erik Petigura
Berkeley Astro
Grad Student

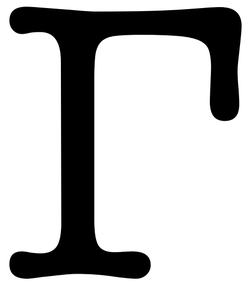


Bootcamp/Seminar Alum

DOE/NERSC computation

21st Century Eduction Mission:
Produce Gamma-shaped people

vs.



deep domain skill/knowledge/training
deep methodological knowledge/skill

deep domain or methodological skill/knowledge/training
strong methodological or domain knowledge/skill

Goal: empower teams of gamma's to excel

A Rich History of Symbiosis

Astronomers co-opt tools...

"Fleming had constructed a spyglass, by means of which visible objects, though very distant from the eye of the observer, were distinctly seen as if nearby."

- Galileo Galilei (1610)

...while astro serves as a testbed for computational exploration

“I love working with astronomers, since their data is worthless.”

- Jim Gray, Microsoft

“novelty^2 problem”

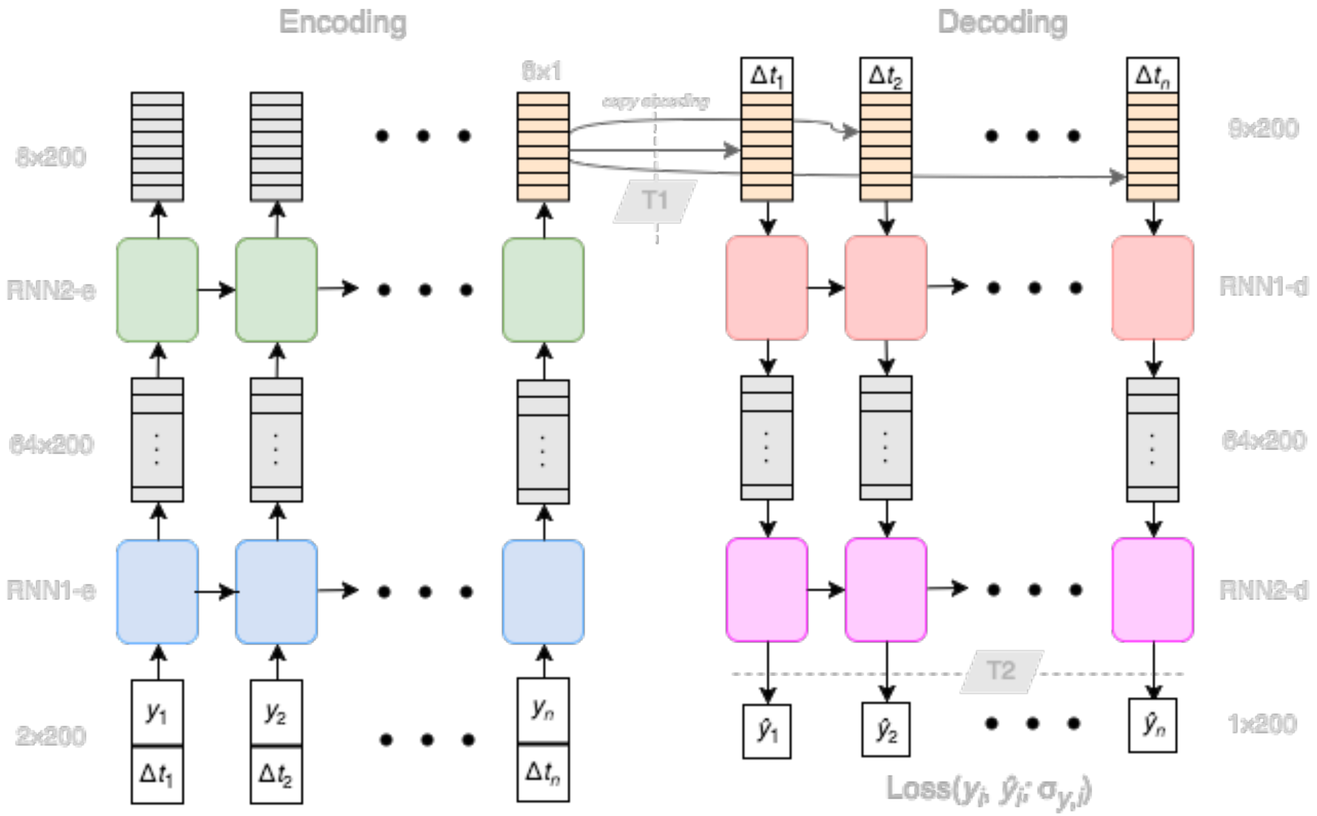
Established CS/Stats/Math *in Service*
of novelty in domain science

vs.

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

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

Deep Learning for (Astronomical) Timeseries Inference



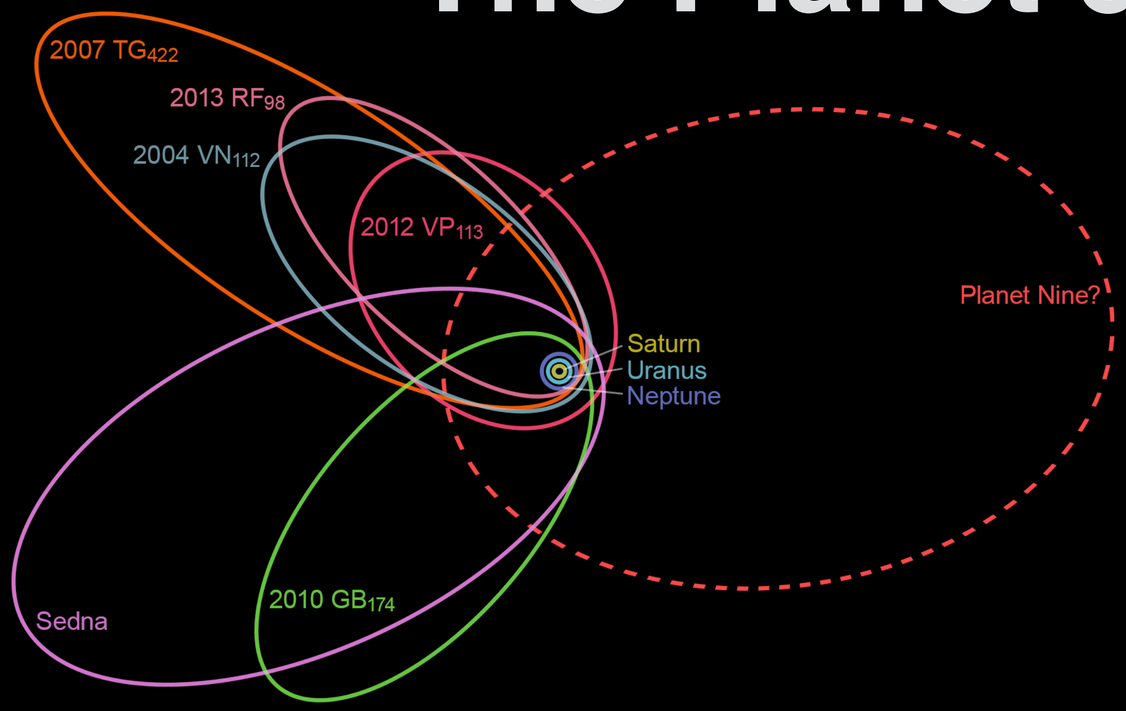
Network for RNN+GRU

"Automated feature extraction for irregularly-sampled time series using deep neural networks"

Naul, JSB+, in prep

Stats PhD from Stanford

The Planet 9 Opportunity



hypothesized to **exist** using mostly public data about comet orbits + deep domain knowledge + sophisticated computing

Batygin & Brown 1601.05438

My Prediction: the discovery data of Planet X have already been obtained & reside in existing public data archives. It will be a group of clever astronomers & statisticians with a lot of compute resources that will make the retrospective discovery of the century…

https://www.forbes.com/sites/jillianscudder/2017/01/21/astroquizzical-care-planet-nine

Summary

- Democratizing trends in the physical sciences
- Winners will be those domain experts with superior inferential capabilities, not themselves, but in teams
- Data science education for domains should be focused on learning enough to work with those methodological domains
- Symbiosis: Physical domain science data/questions can also foster novel methodological approaches

Emerging Needs & Opportunities in Data-Intensive Domains: Astronomy

Thanks!

Joshua Bloom
@profjsb