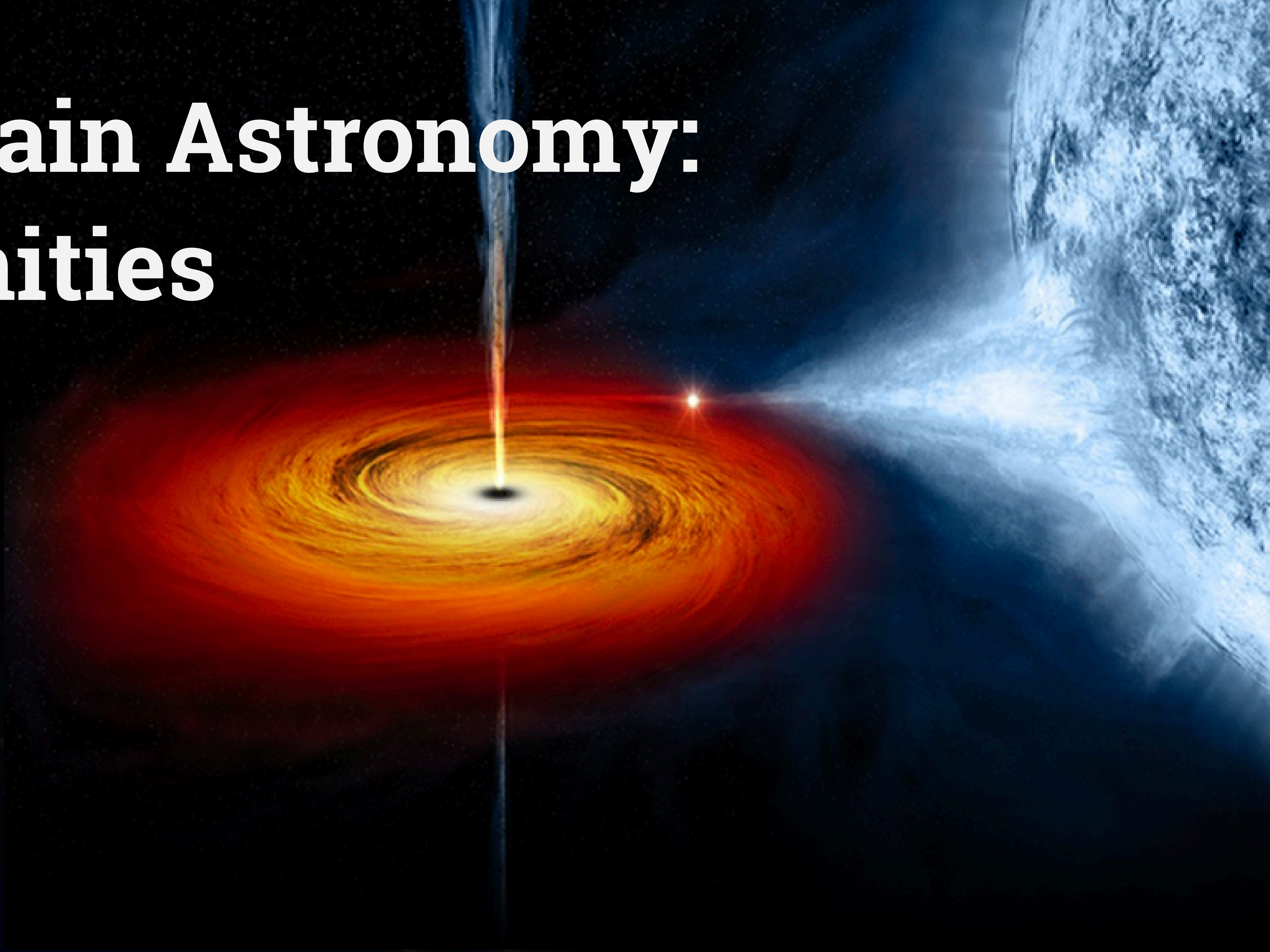
Data Science Challenges in Time Domain Astronomy: Building Methods, Tools and Communities

Daniela Huppenkothen

DIRAC, University of Washington
eScience Institute, University of Washington

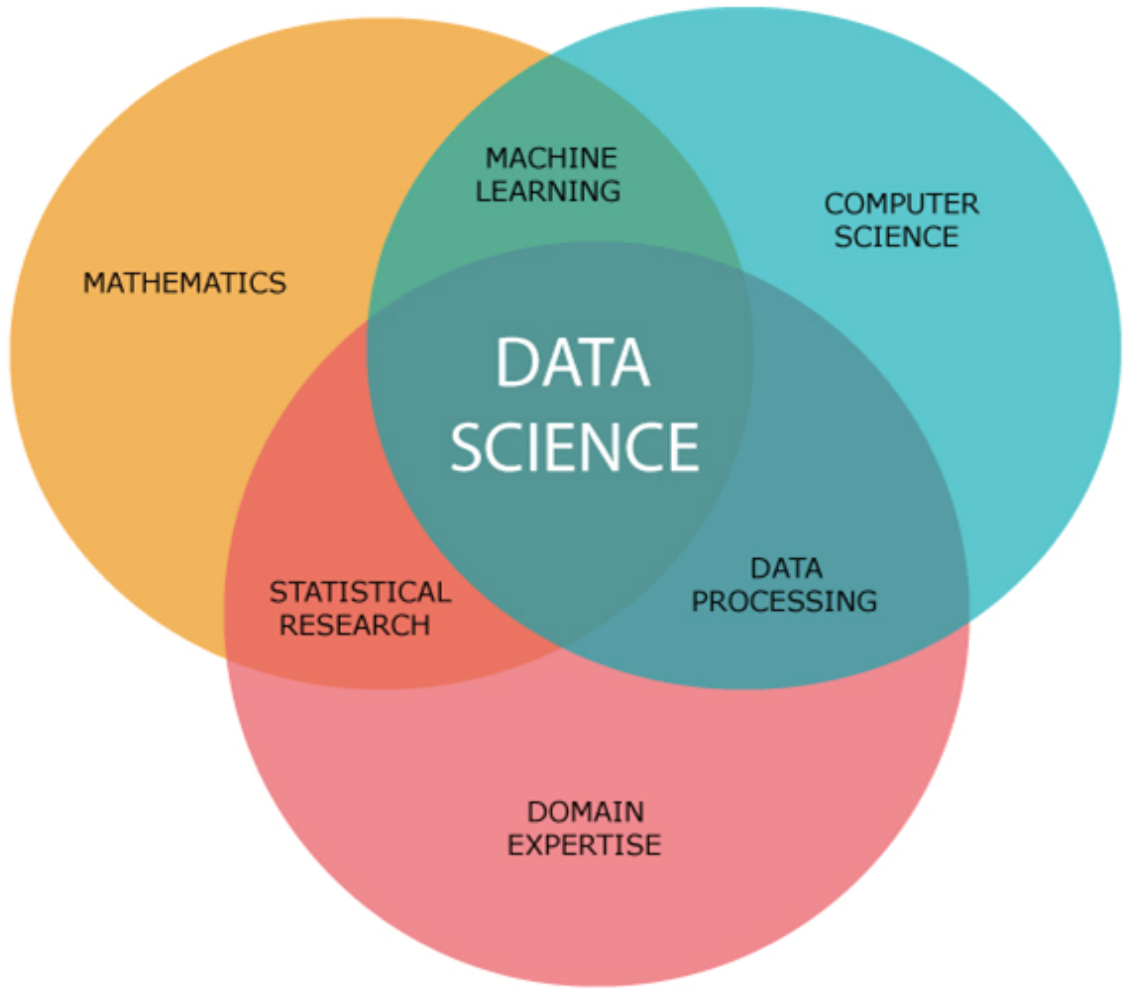


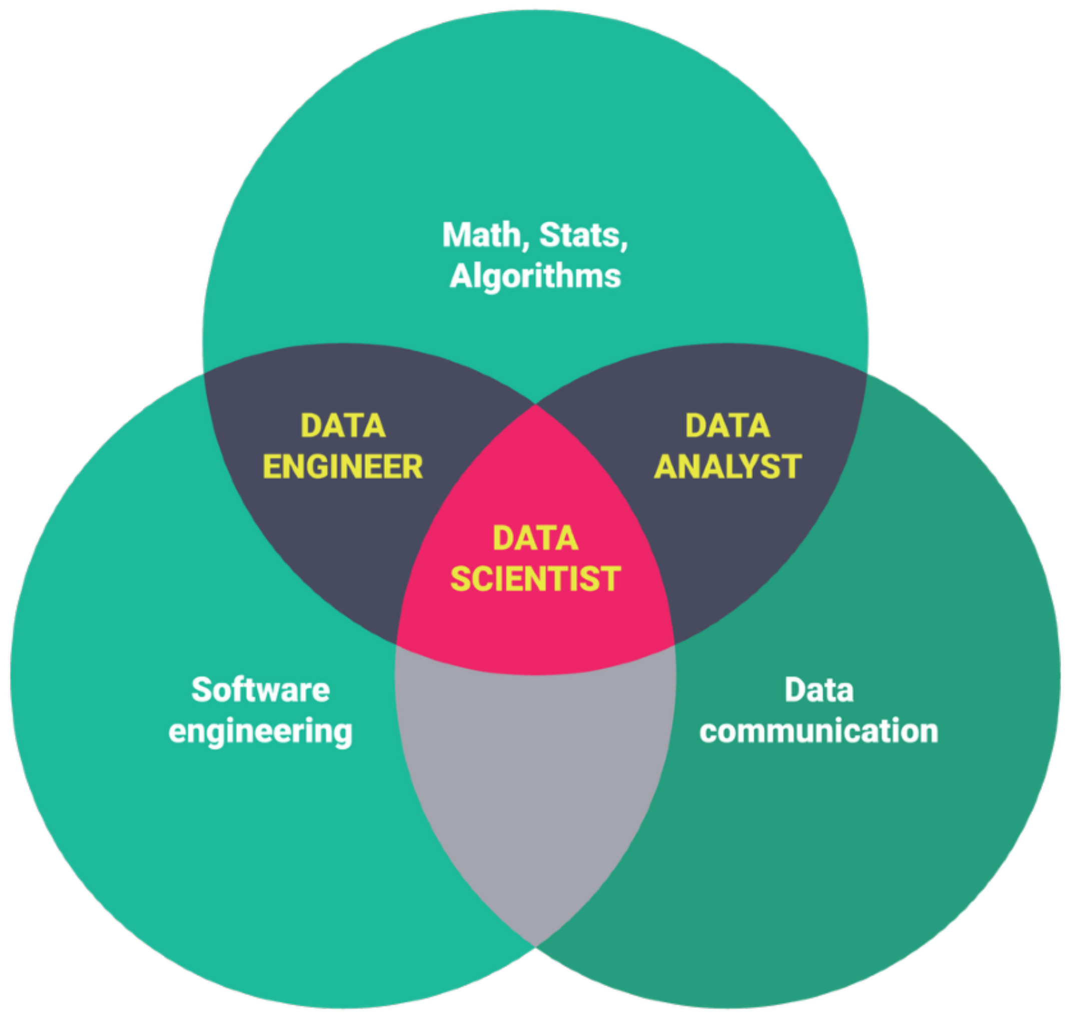
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

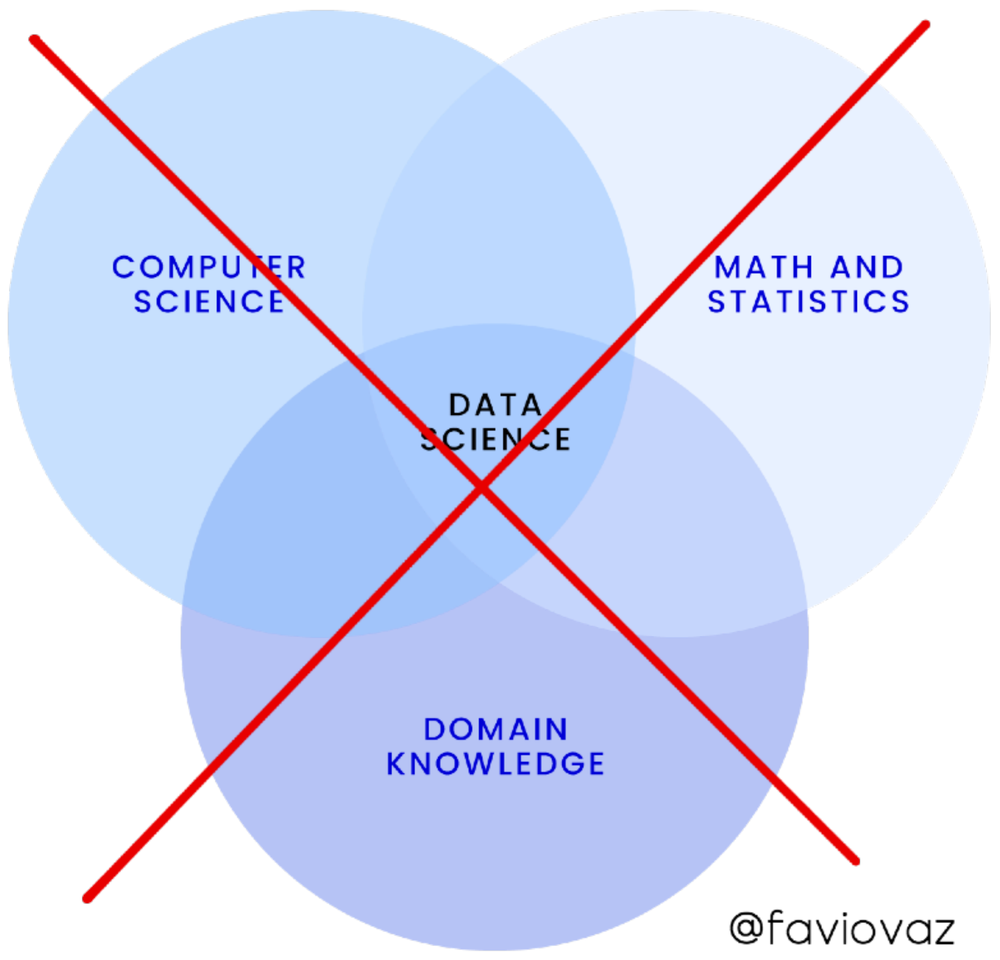
DiRAC

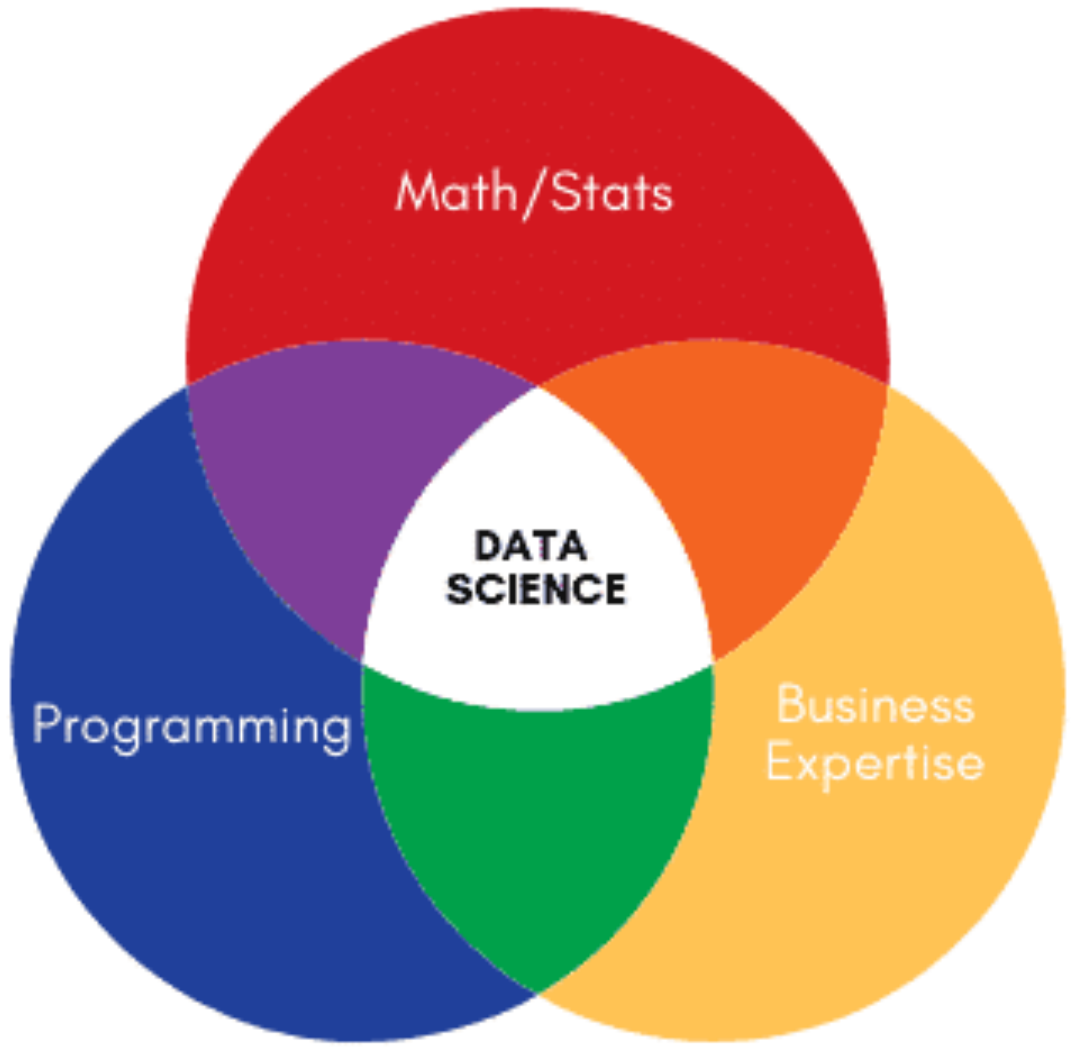
Tiana_Athriel
dhuppenkothen
✉ dhuppenk@uw.edu

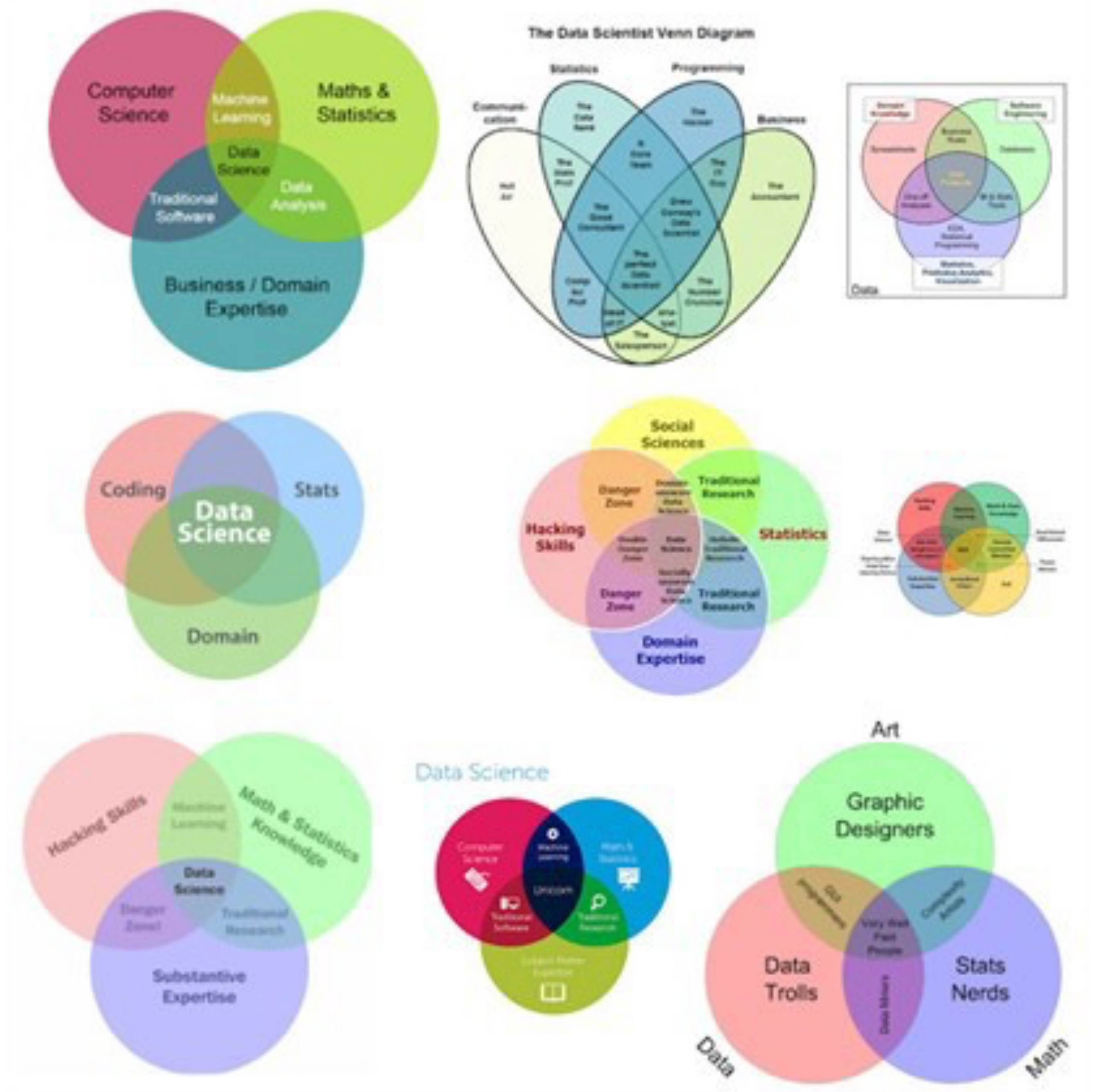
Why data science?



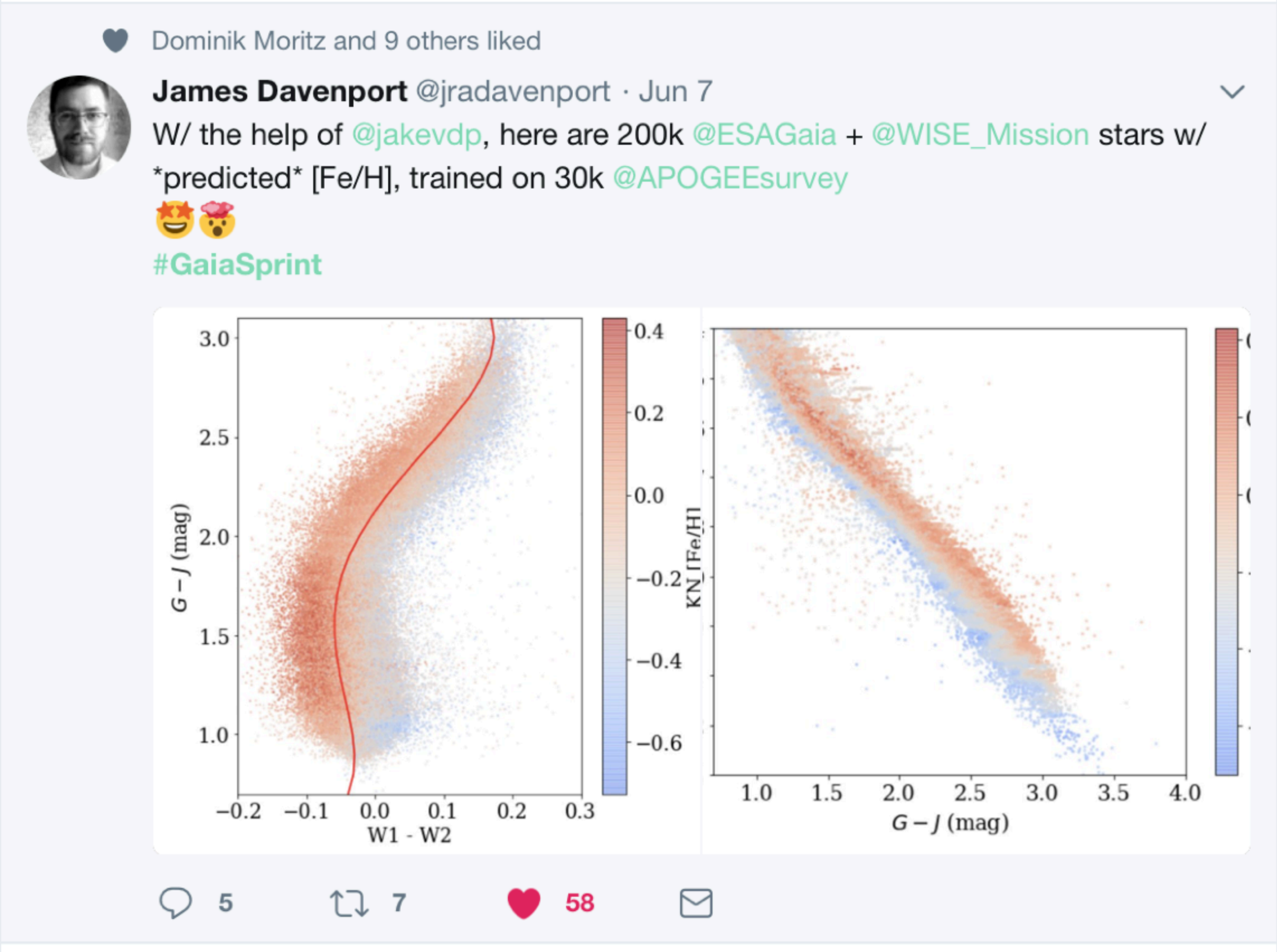








Data science is the art of drawing Venn diagrams?



Dominik Moritz and 9 others liked

James Davenport @jradavenport · Jun 7

W/ the help of @jakevdp, here are 200k @ESAGaia + @WISE_Mission stars w/ *predicted* [Fe/H], trained on 30k @APOGEEsurvey

🤩🤯

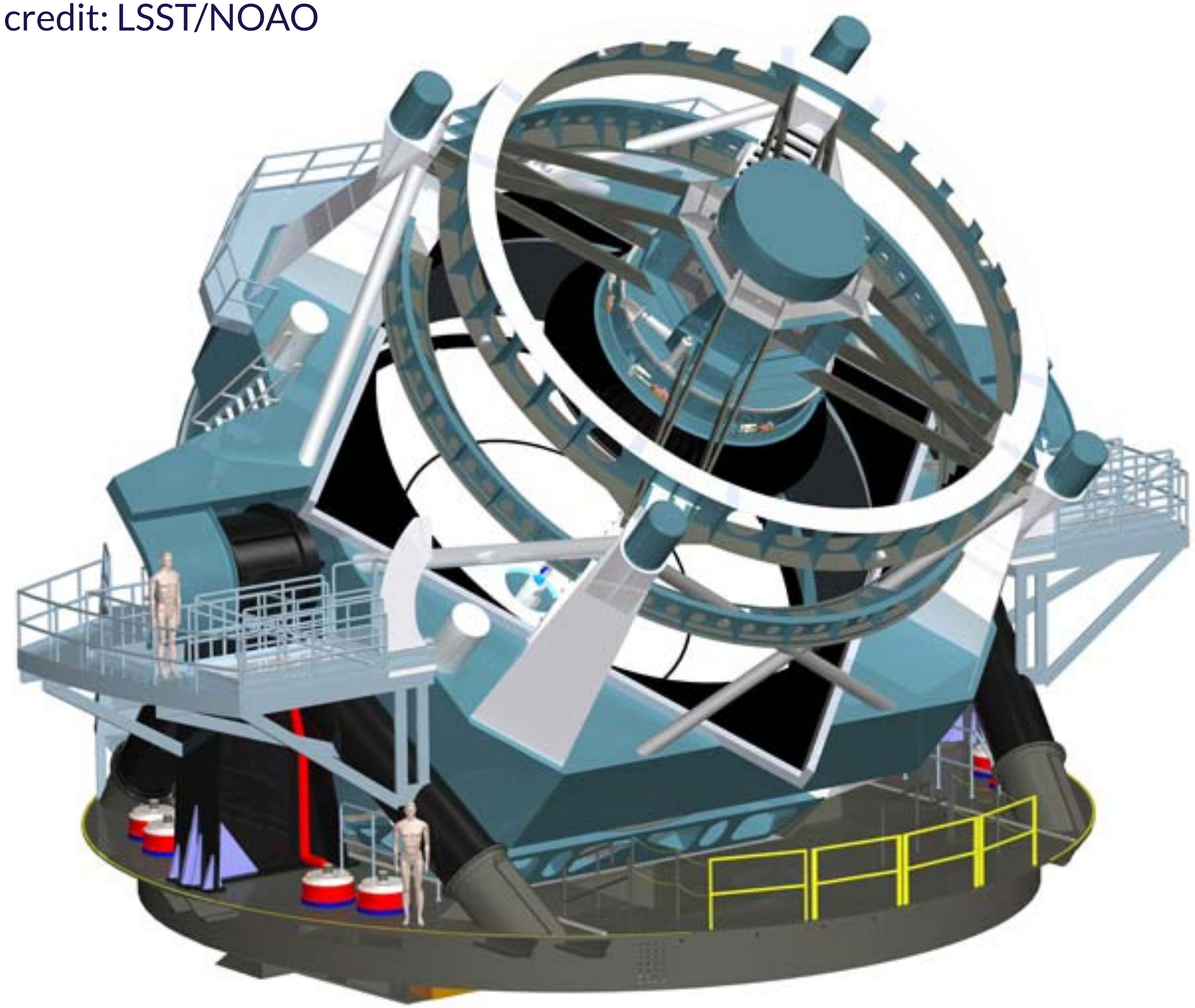
#GaiaSprint

💬 5 🔁 7 ❤ 58 ✉

Gaia

1.7 billion stars

credit: LSST/NOAO

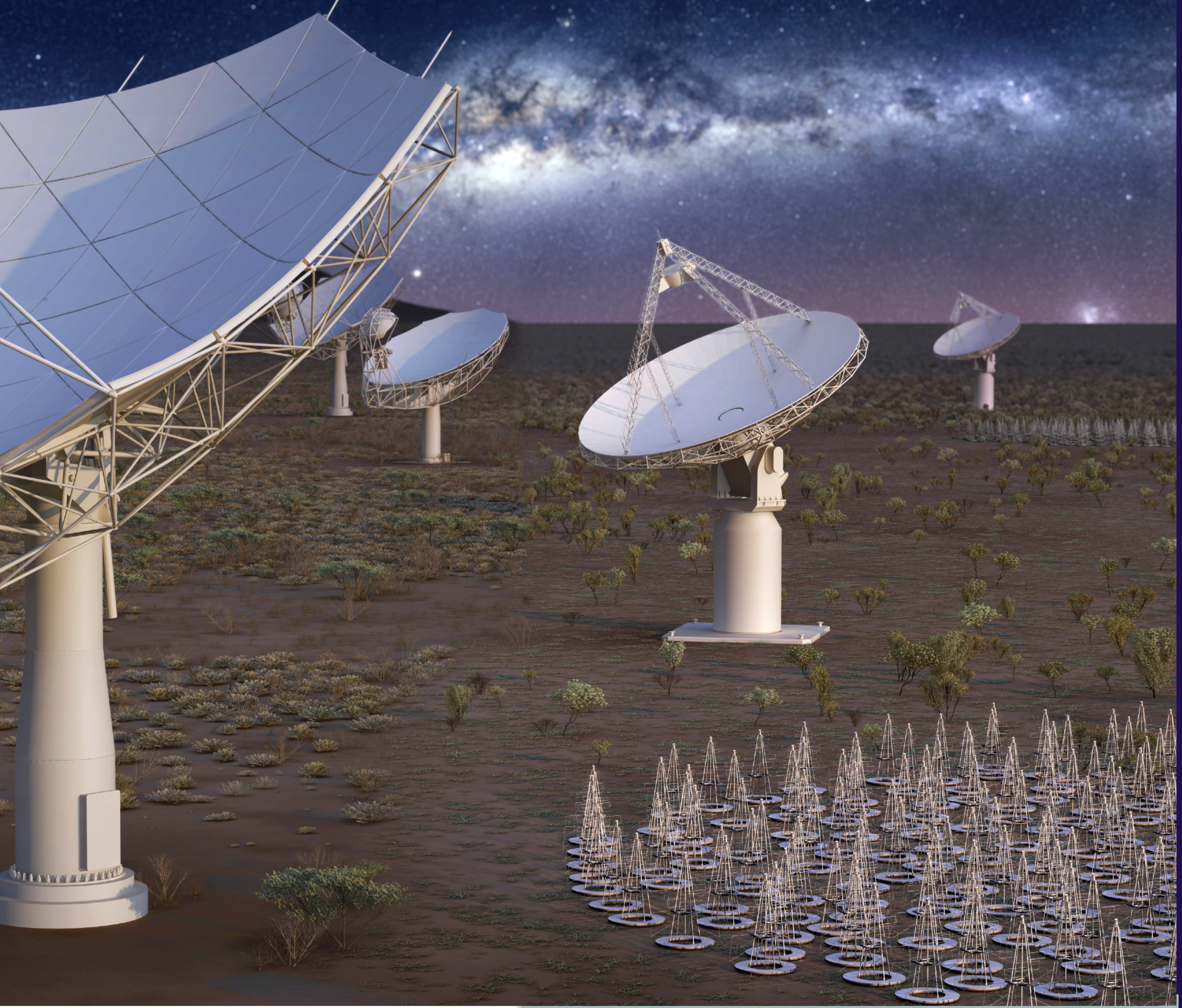


LSST

Large Synoptic Survey Telescope

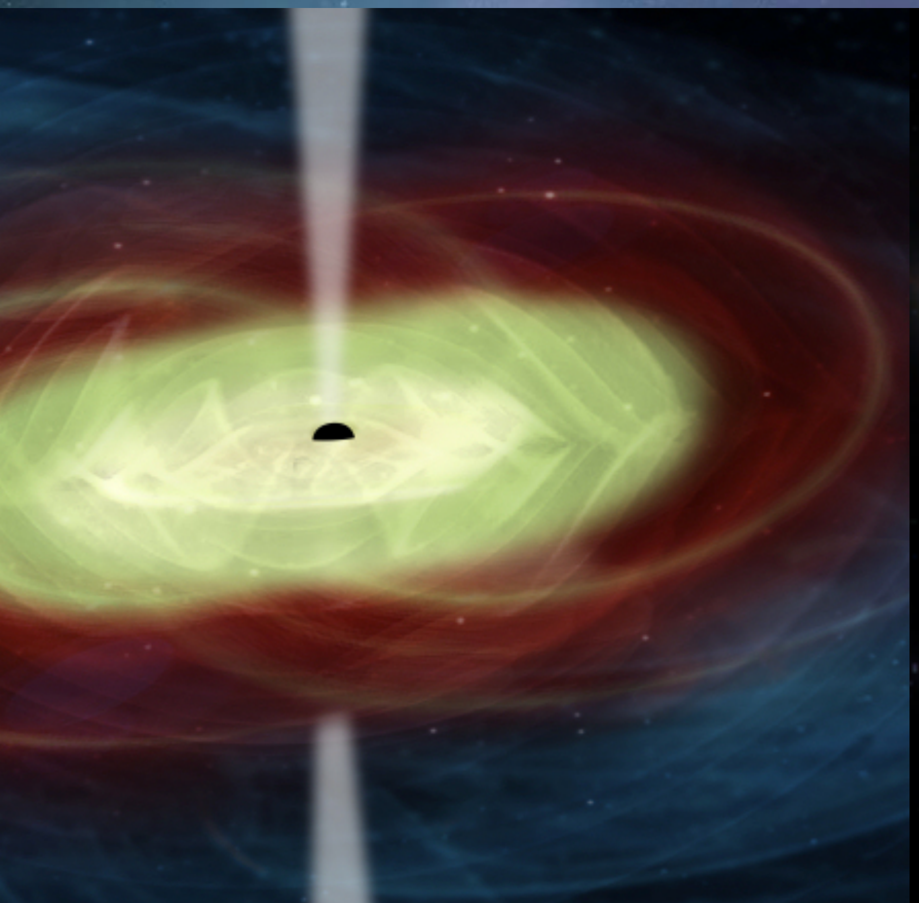
LSST

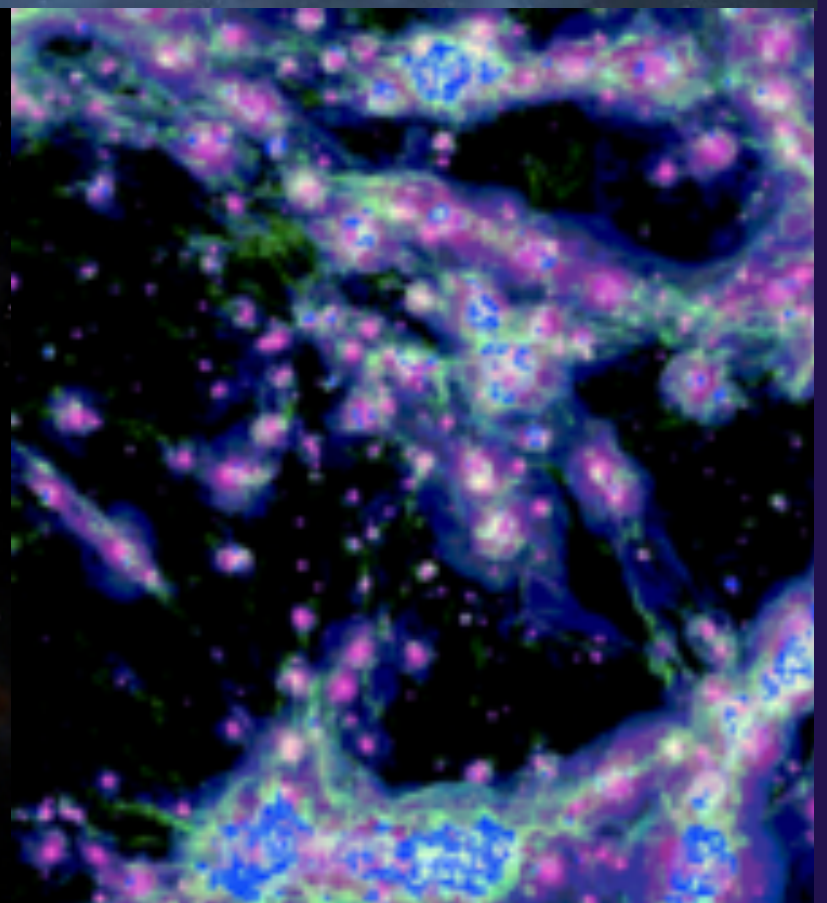
40 billion sources
10 million alerts/night



SKA

160TB raw data/second





ATHENA

100x effective area of Chandra

7.5 arcmin field of view

... but it’s not all big data!

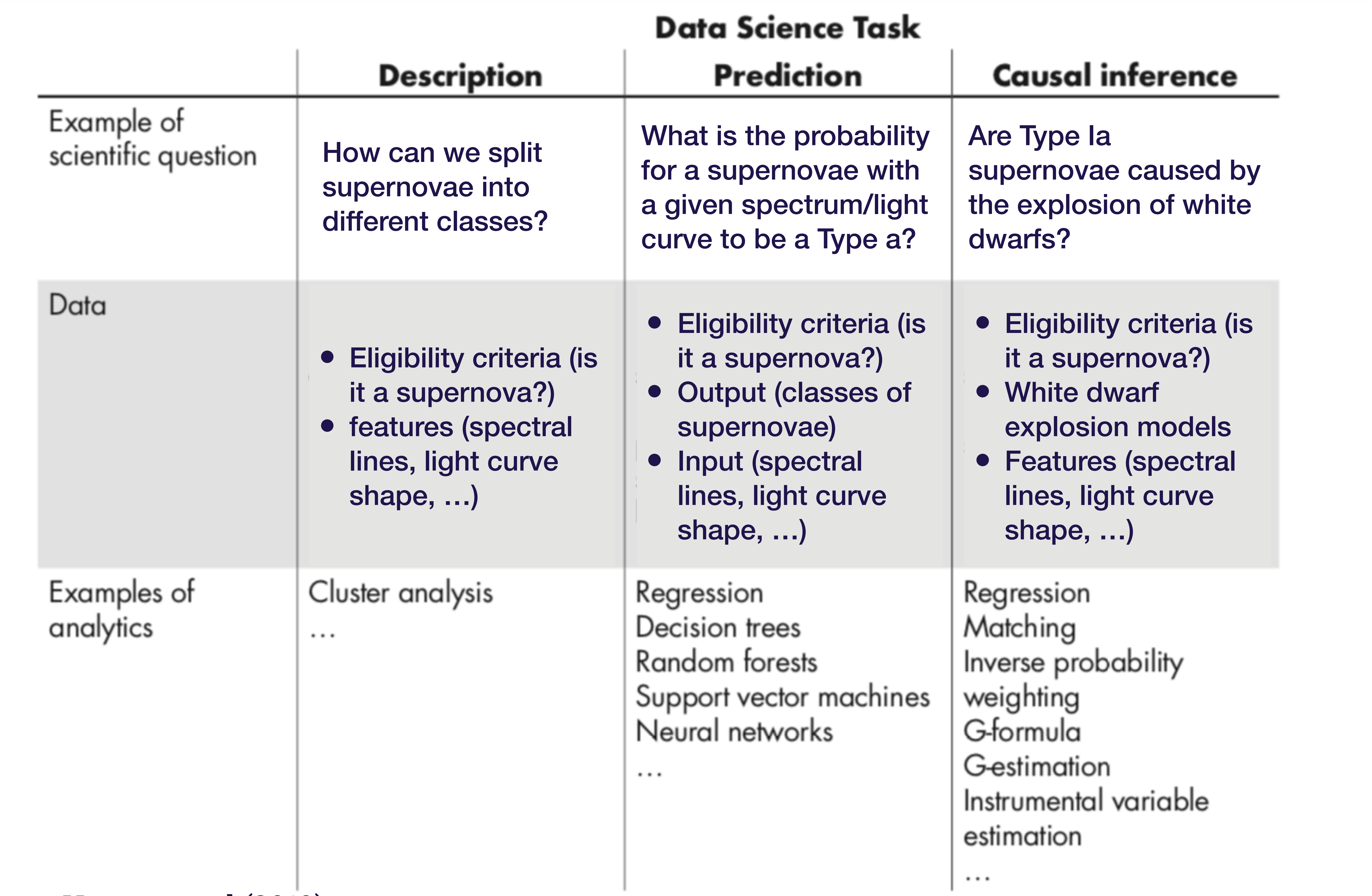
“Annoyingly **not small** (and **complex**) data”

— Genevieve Graves

How do we make sense of all this data?

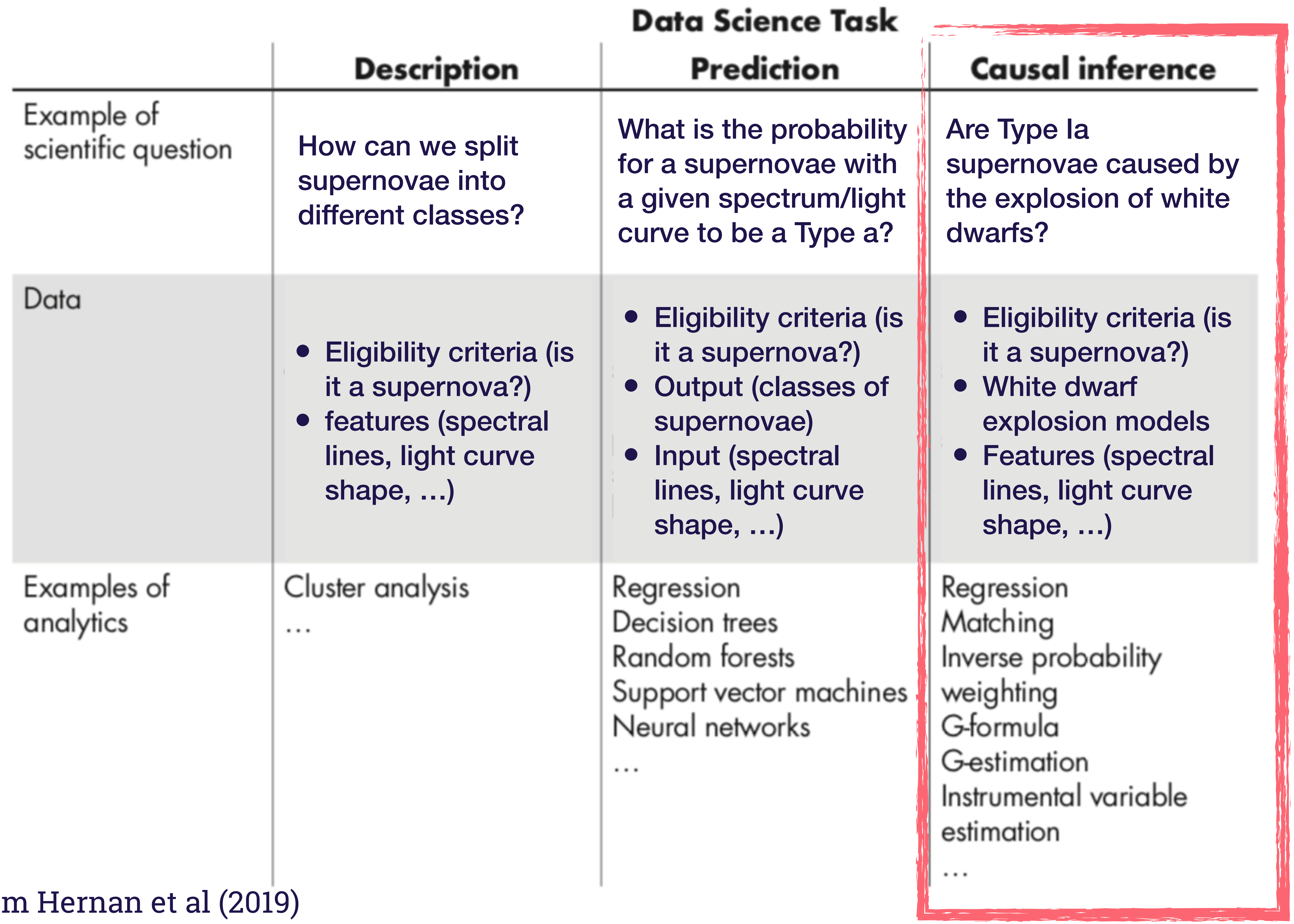
How do we make sense of all this data?

What do we want to learn?



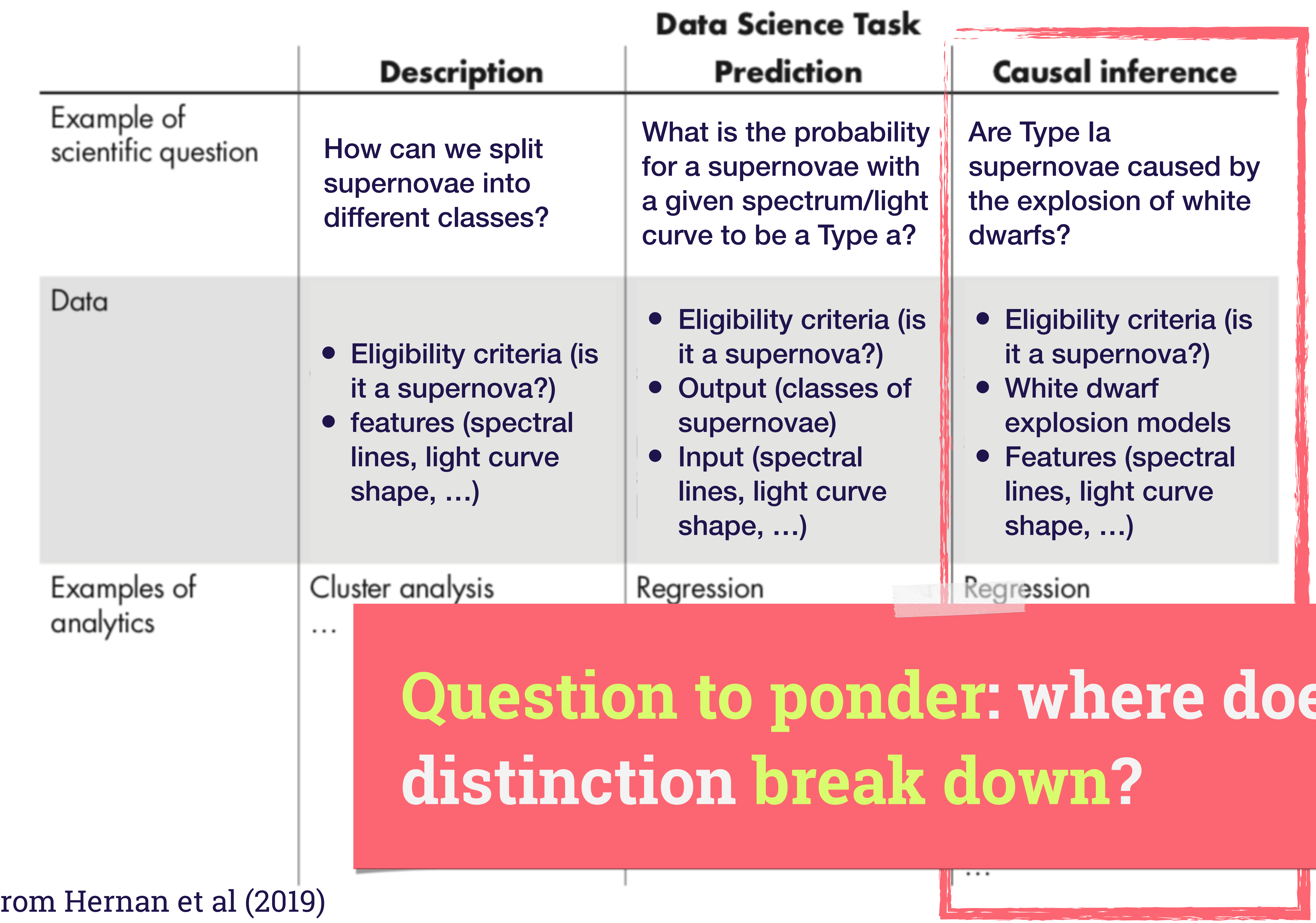
	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can we split supernovae into different classes?	What is the probability for a supernovae with a given spectrum/light curve to be a Type a?	Are Type Ia supernovae caused by the explosion of white dwarfs?
Data	• Eligibility criteria (is it a supernova?) • features (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • Output (classes of supernovae) • Input (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • White dwarf explosion models • Features (spectral lines, light curve shape, …)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

adapted from Hernan et al (2019)



	Data Science Task		
	Description	Prediction	Causal inference
Example of scientific question	How can we split supernovae into different classes?	What is the probability for a supernovae with a given spectrum/light curve to be a Type a?	Are Type Ia supernovae caused by the explosion of white dwarfs?
Data	• Eligibility criteria (is it a supernova?) • features (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • Output (classes of supernovae) • Input (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • White dwarf explosion models • Features (spectral lines, light curve shape, …)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

adapted from Hernan et al (2019)

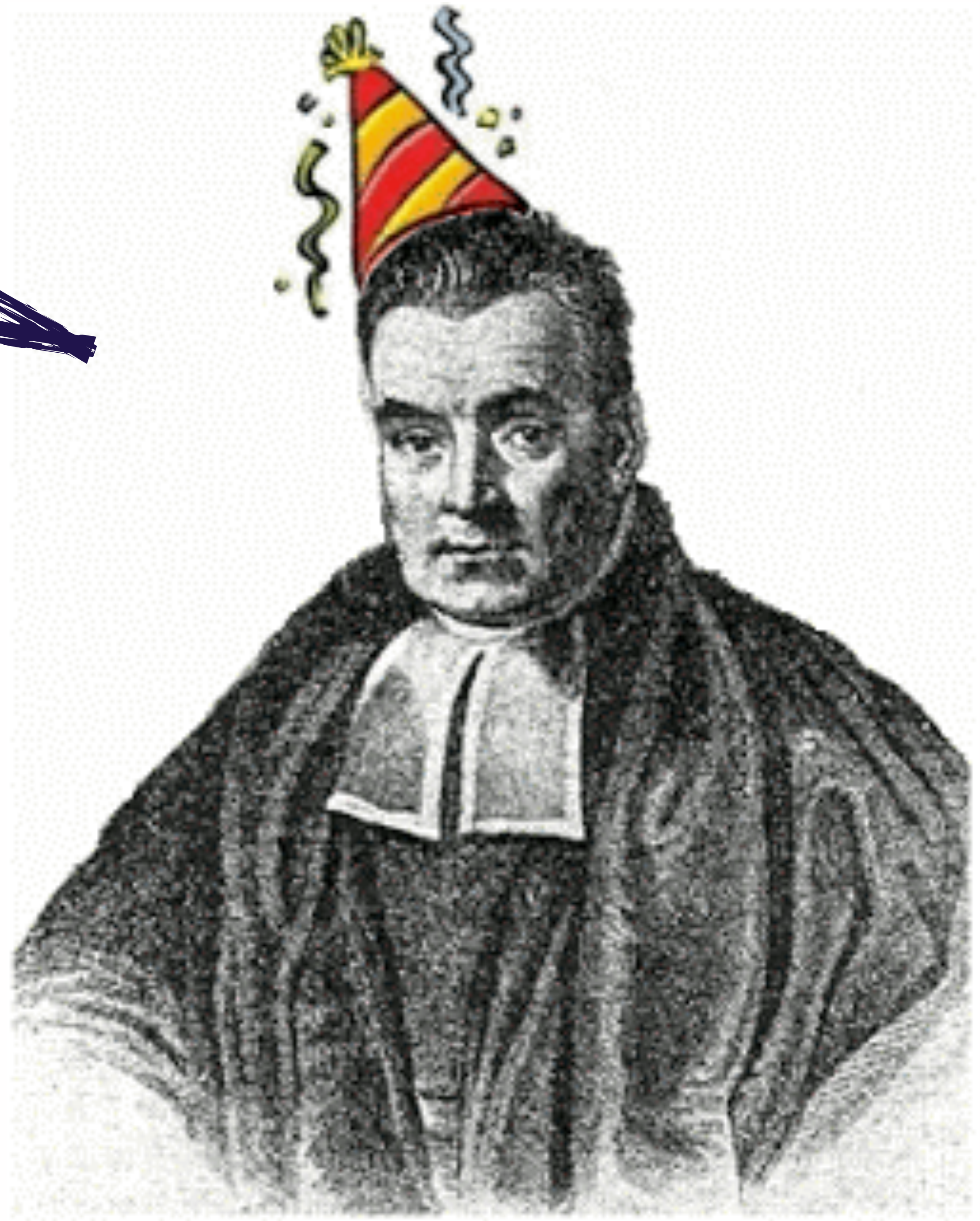


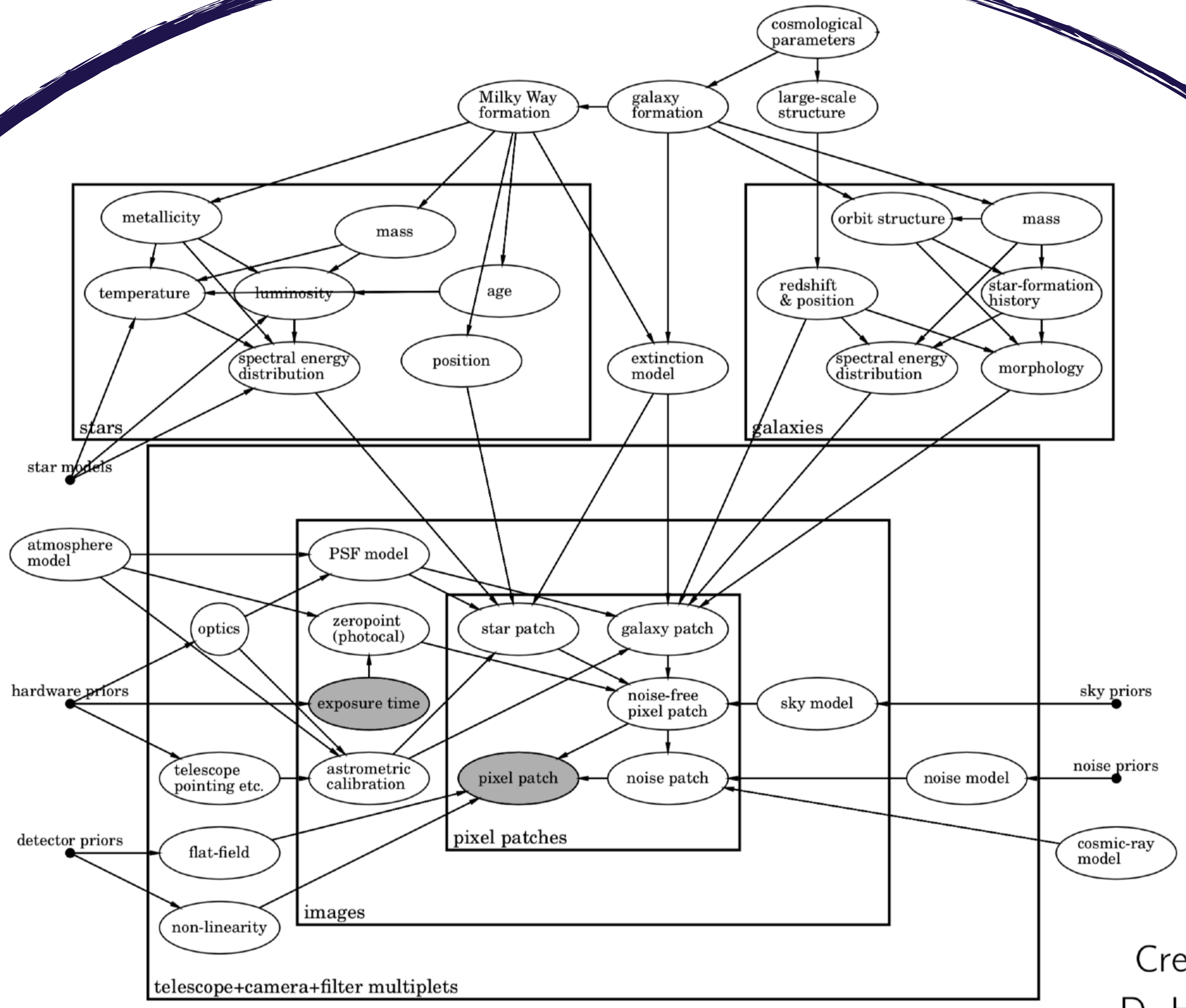
	Data Science Task		
	Description	Prediction	Causal inference
Example of scientific question	How can we split supernovae into different classes?	What is the probability for a supernovae with a given spectrum/light curve to be a Type a?	Are Type Ia supernovae caused by the explosion of white dwarfs?
Data	• Eligibility criteria (is it a supernova?) • features (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • Output (classes of supernovae) • Input (spectral lines, light curve shape, …)	• Eligibility criteria (is it a supernova?) • White dwarf explosion models • Features (spectral lines, light curve shape, …)
Examples of analytics	Cluster analysis …	Regression	Regression …

Question to ponder: where does this distinction break down?

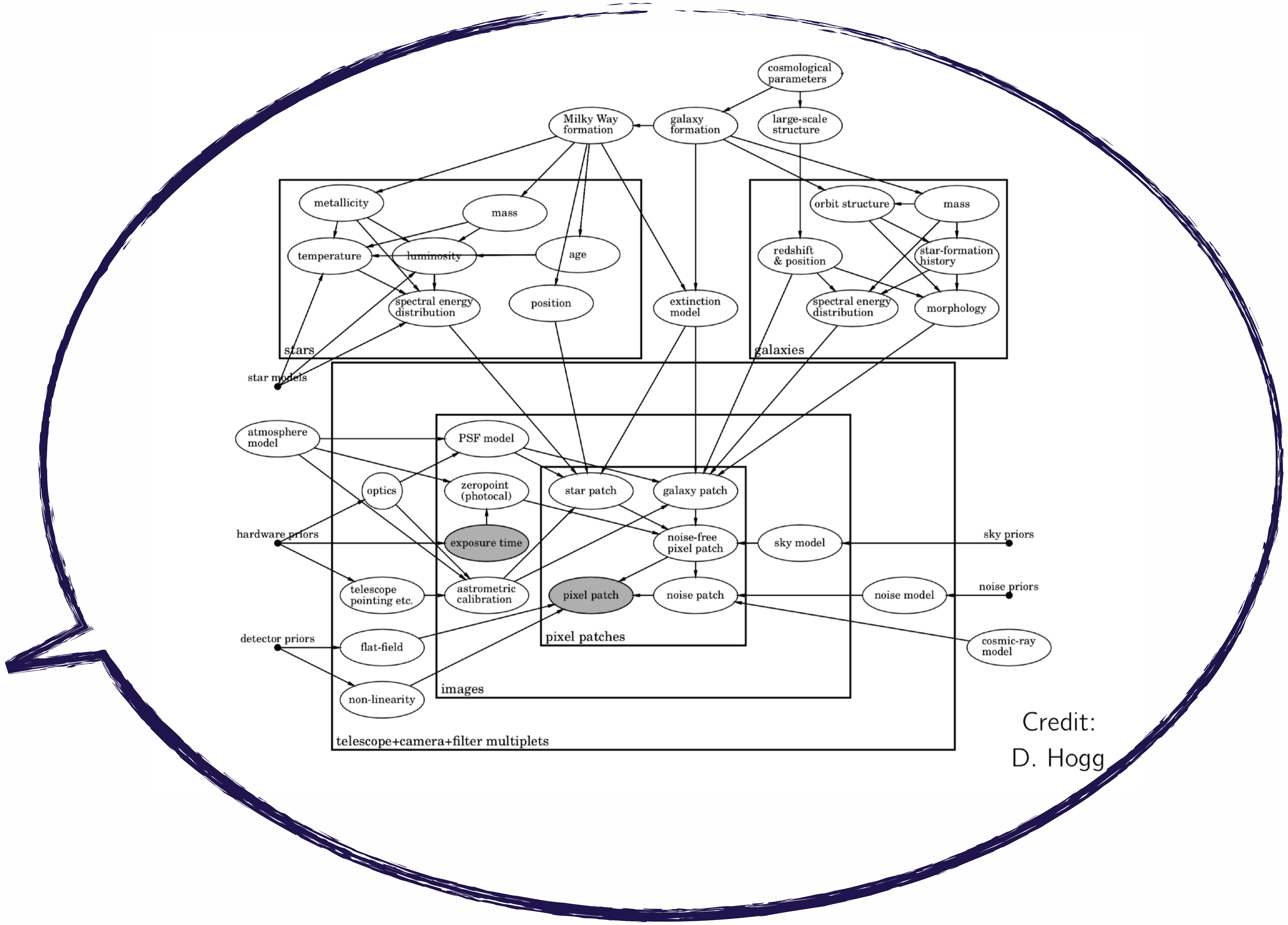
adapted from Hernan et al (2019)

You should just build a Bayesian (hierarchical) model!



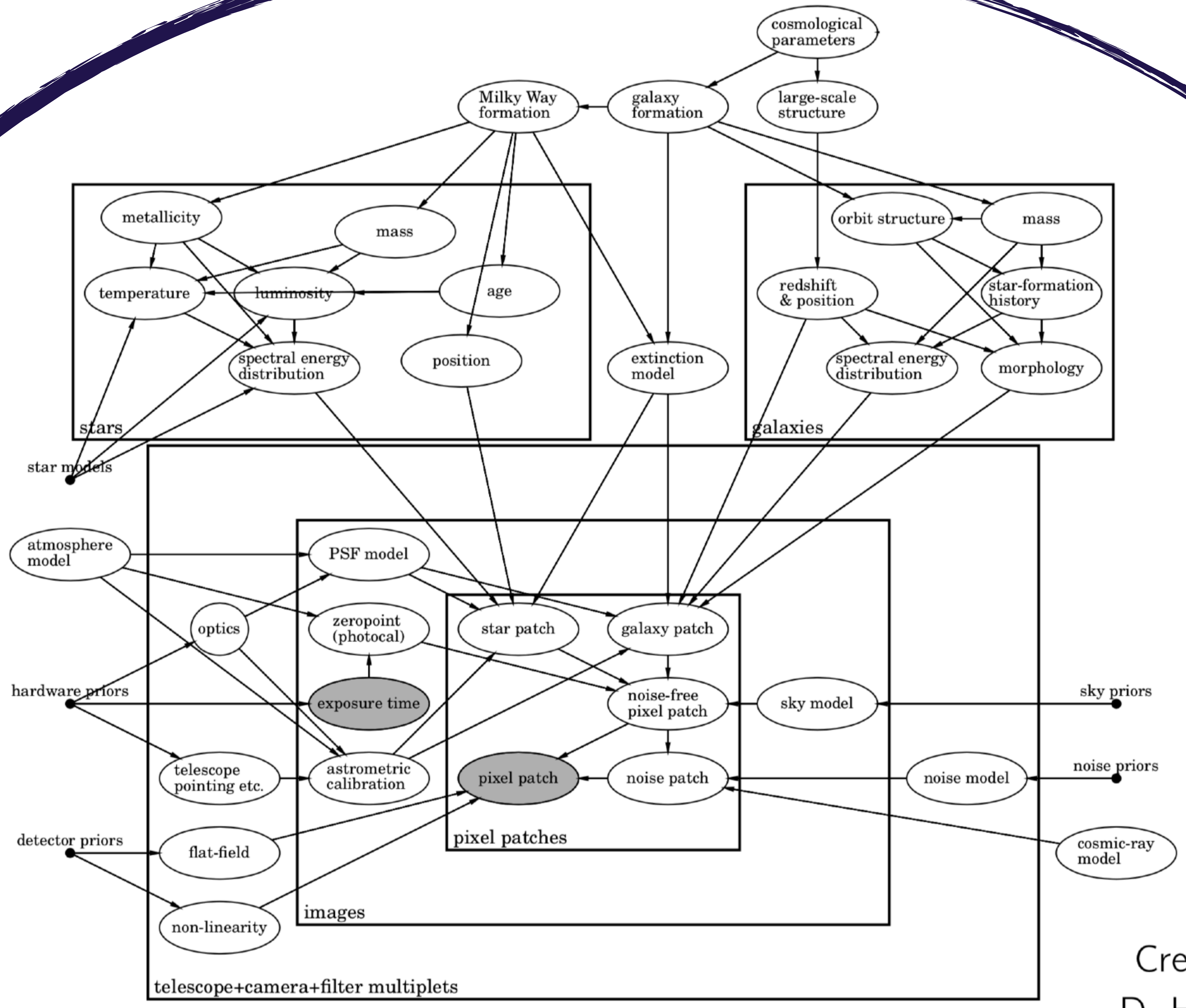


Credit:
D. Hogg

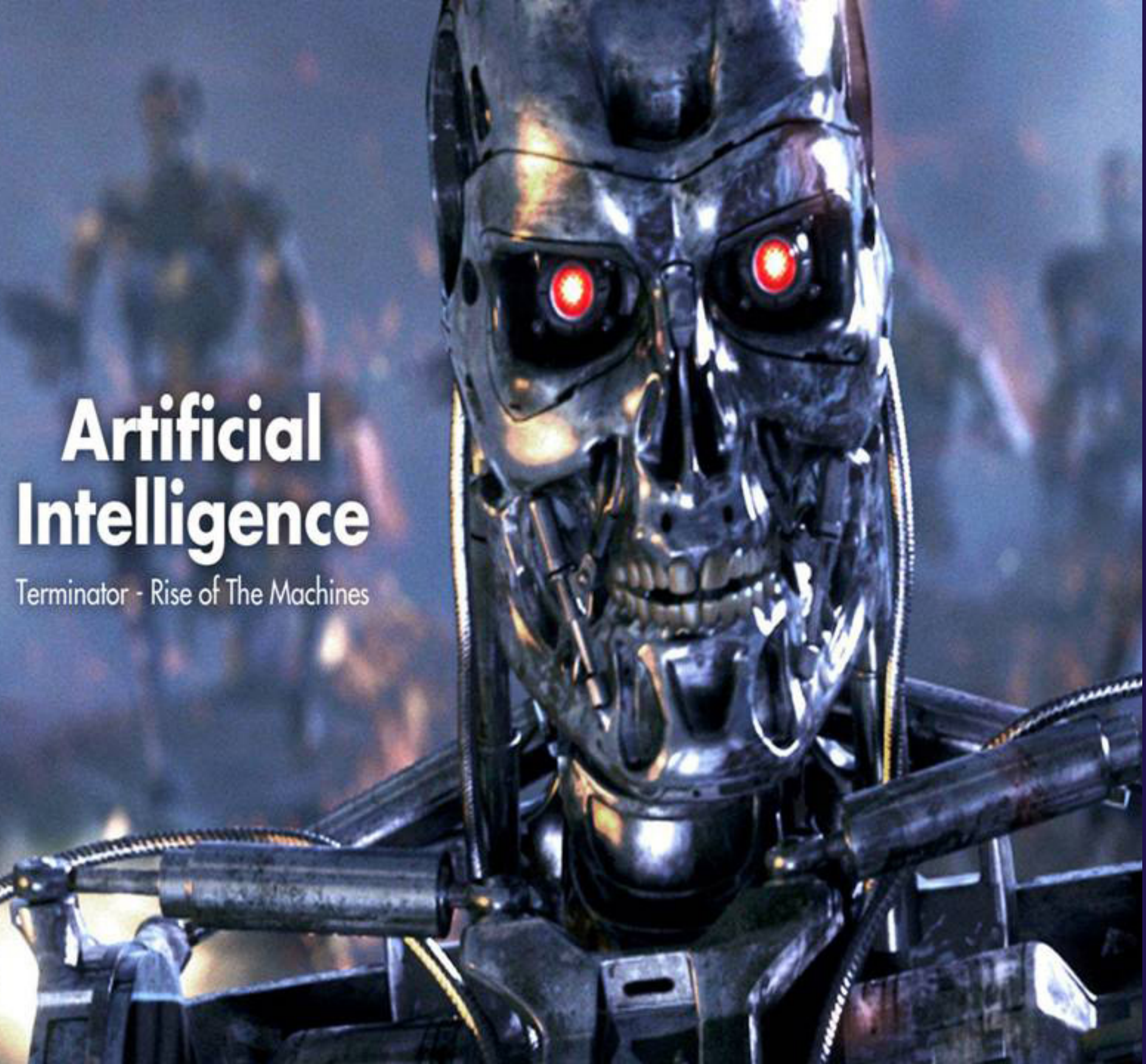


Credit:
D. Hogg

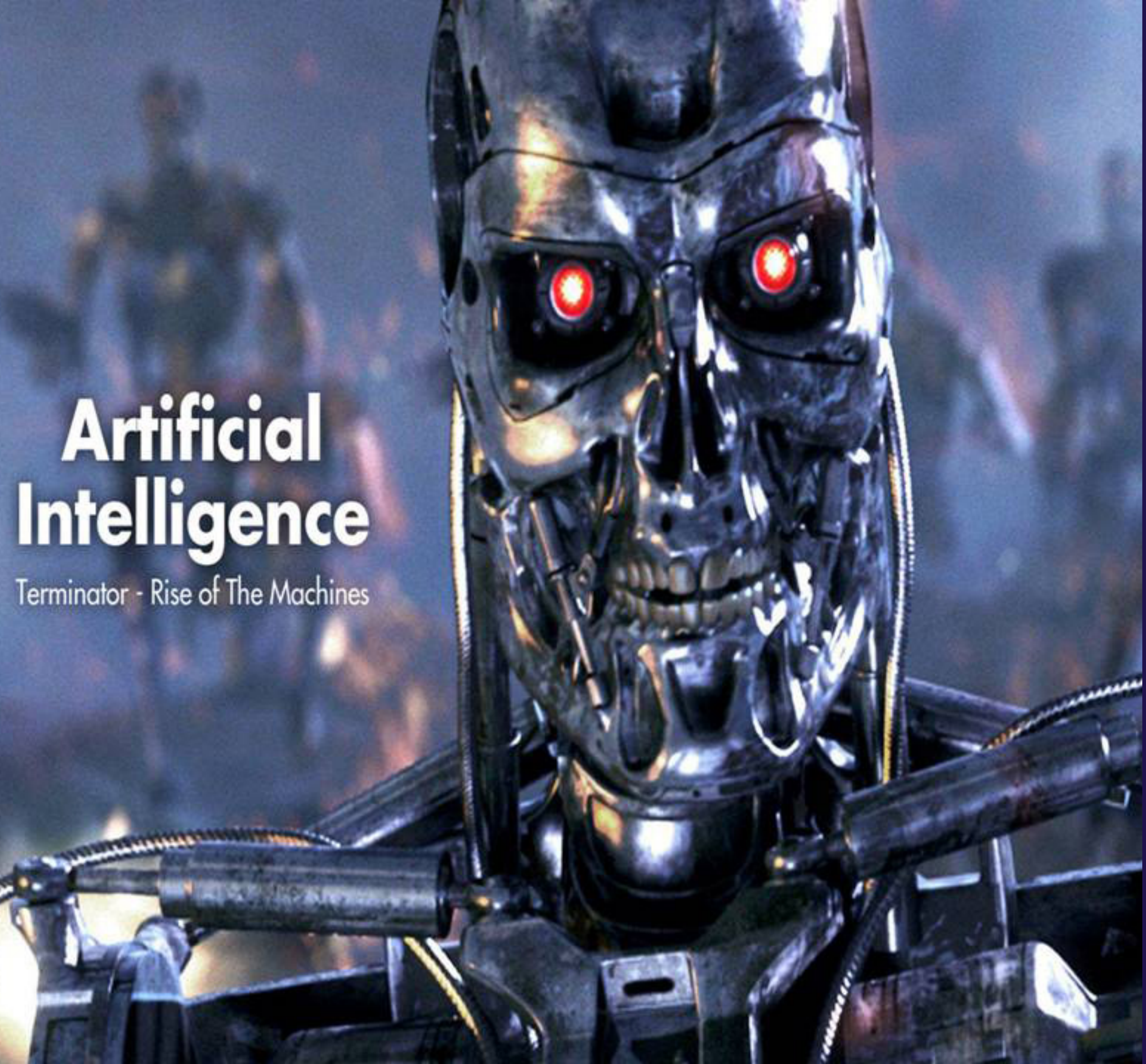
Maybe I can help?



Credit:
D. Hogg

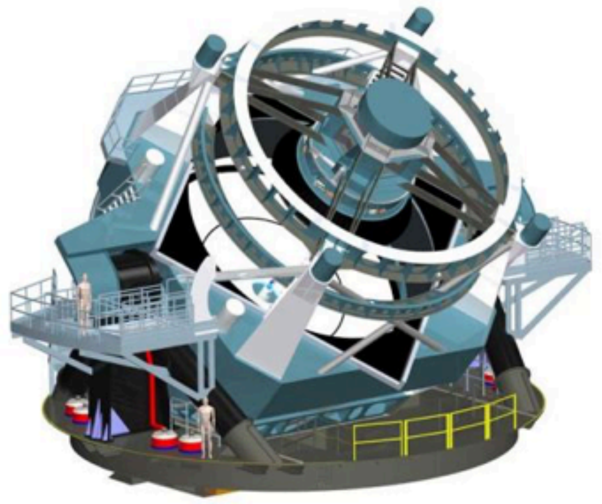


Are **astronomers** going to be replaced by a **neural network**?

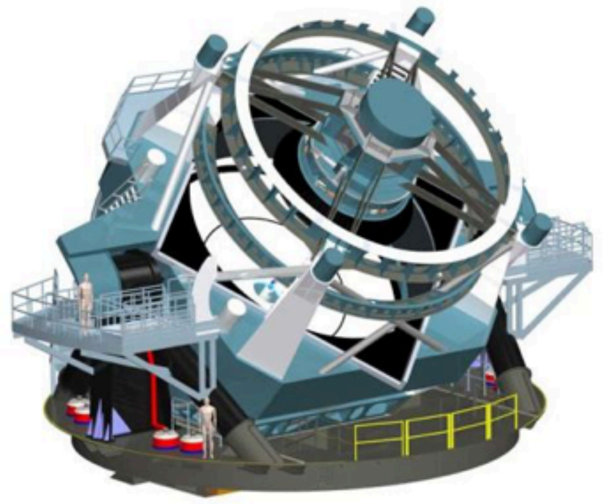


Are **astronomers** going to be replaced by a **neural network**?

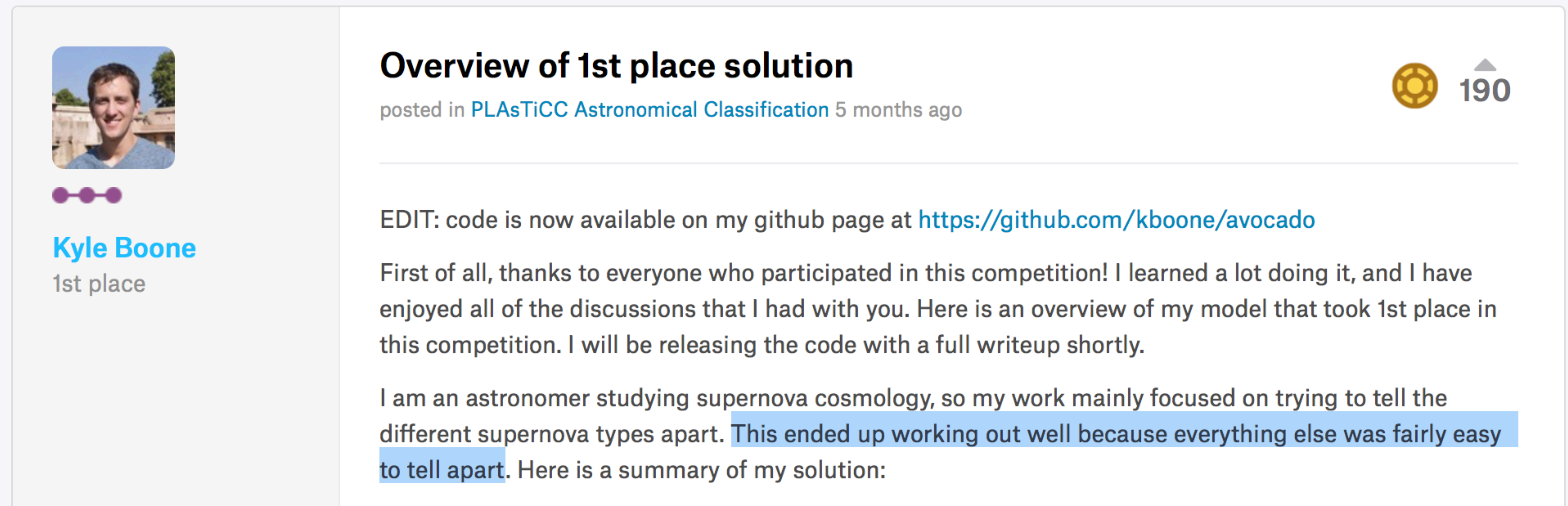
(no)



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)



Kyle Boone
1st place

Overview of 1st place solution

posted in PLAsTiCC Astronomical Classification 5 months ago

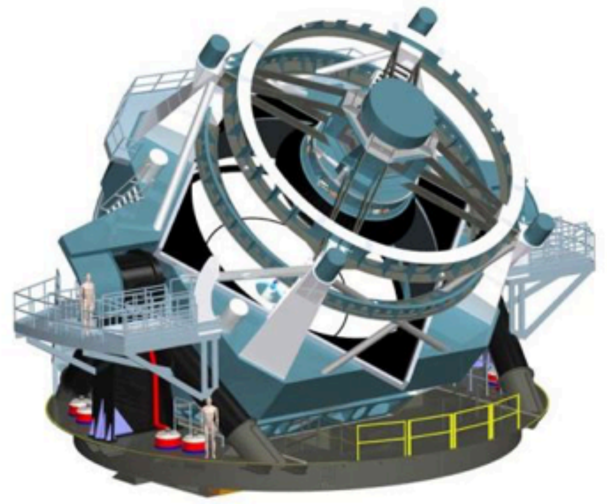
190

EDIT: code is now available on my github page at https://github.com/kboone/avocado

First of all, thanks to everyone who participated in this competition! I learned a lot doing it, and I have enjoyed all of the discussions that I had with you. Here is an overview of my model that took 1st place in this competition. I will be releasing the code with a full writeup shortly.

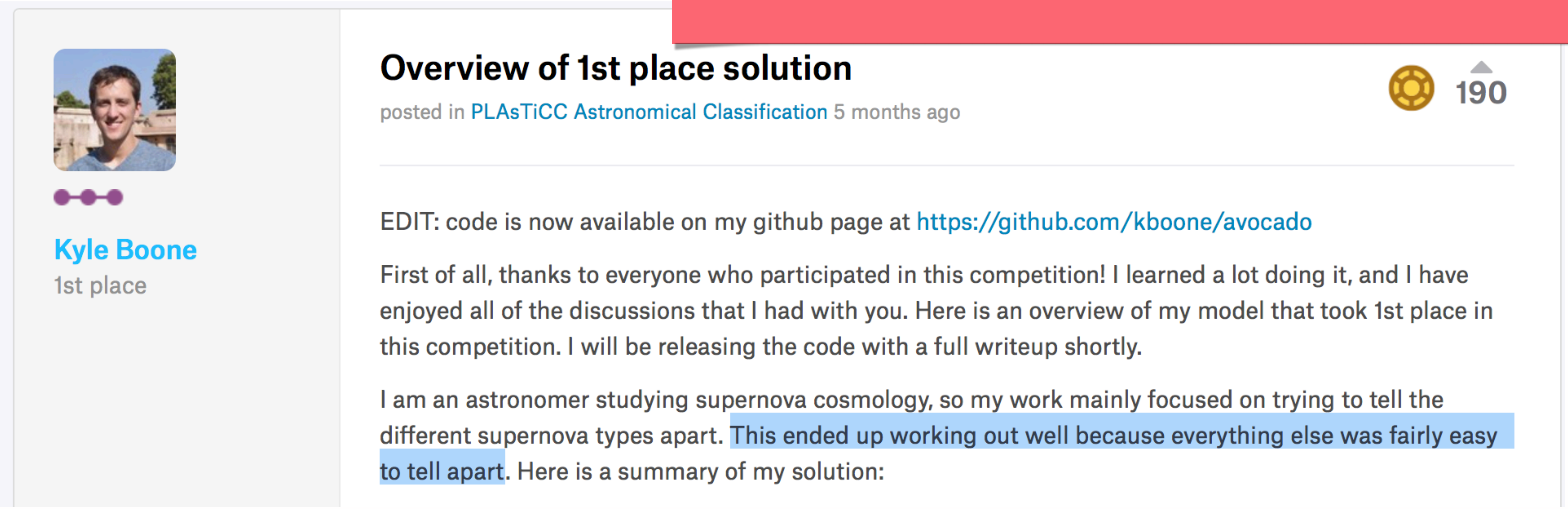
I am an astronomer studying supernova cosmology, so my work mainly focused on trying to tell the different supernova types apart. This ended up working out well because everything else was fairly easy to tell apart. Here is a summary of my solution:

Boone (2019): 1907.04690



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

> “However, **sound causal inference** requires not only **adequate data analysis techniques** but also **subject-matter expertise** about the causal structure of the problem under study”
>
> (Hernan, 2019: “Comment: Spherical Cows in a Vacuum”)



Kyle Boone

1st place

Overview of 1st place solution

posted in PLAsTiCC Astronomical Classification 5 months ago

190

EDIT: code is now available on my github page at https://github.com/kboone/avocado

First of all, thanks to everyone who participated in this competition! I learned a lot doing it, and I have enjoyed all of the discussions that I had with you. Here is an overview of my model that took 1st place in this competition. I will be releasing the code with a full writeup shortly.

I am an astronomer studying supernova cosmology, so my work mainly focused on trying to tell the different supernova types apart. This ended up working out well because everything else was fairly easy to tell apart. Here is a summary of my solution:

Boone (2019): 1907.04690

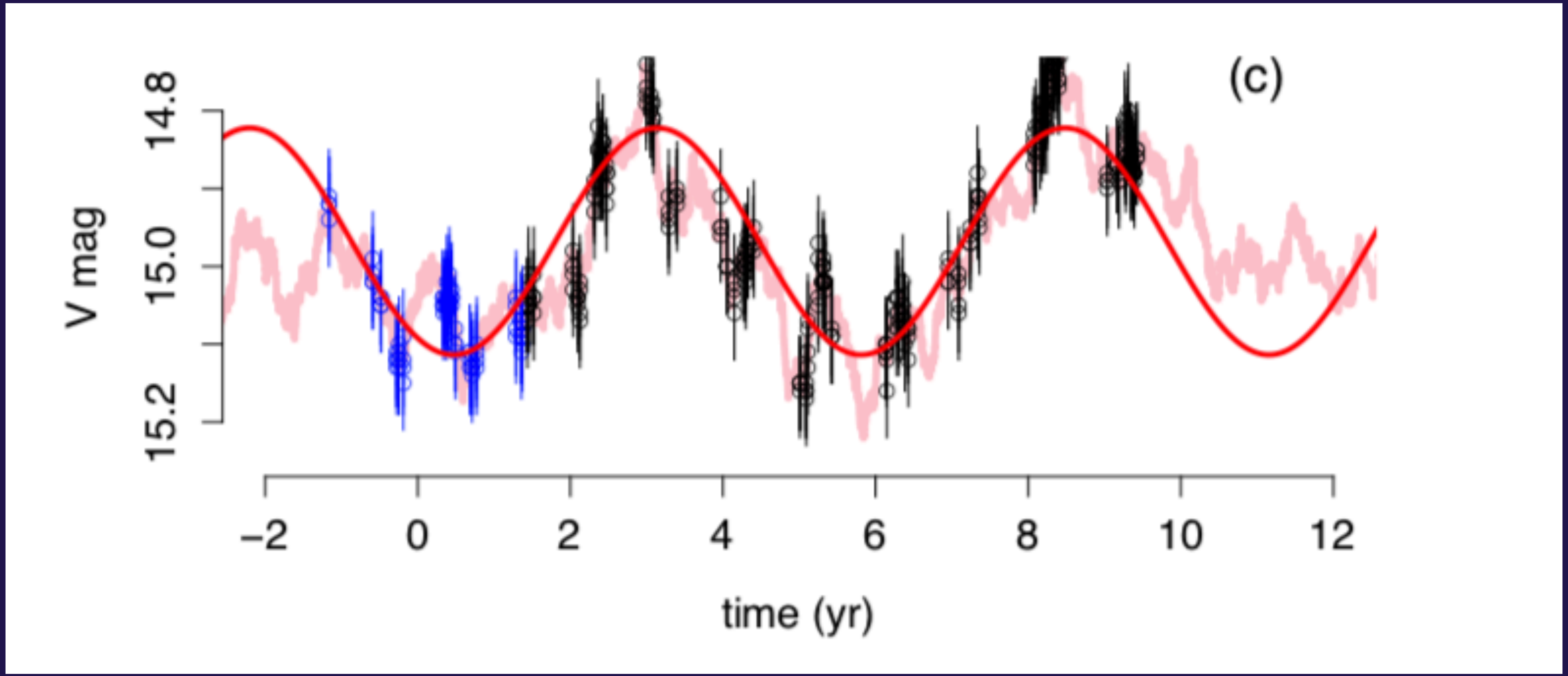
Data Science @ ADASS

- Yanxia Zhang: Photometric Redshift Estimation of Quasars by Machine Learning
- Sweta Singh: Scientific Visualisation of Extremely Large Distributed Astronomical Surveys
- Shraddha Surana: Machine Learning for Scientific Discovery
- Antonia Rowlinson: Identifying transient and variable sources in radio images
- Lightning Session 2: Automated Bayesian Inference, Supernova detection with deep learning, visualization of virual observatory data, Gaussian process modelling
- Sessions 5a and 5b: data visualization
- Lots of posters!

A Statistical View of Our Data

Our data is often …

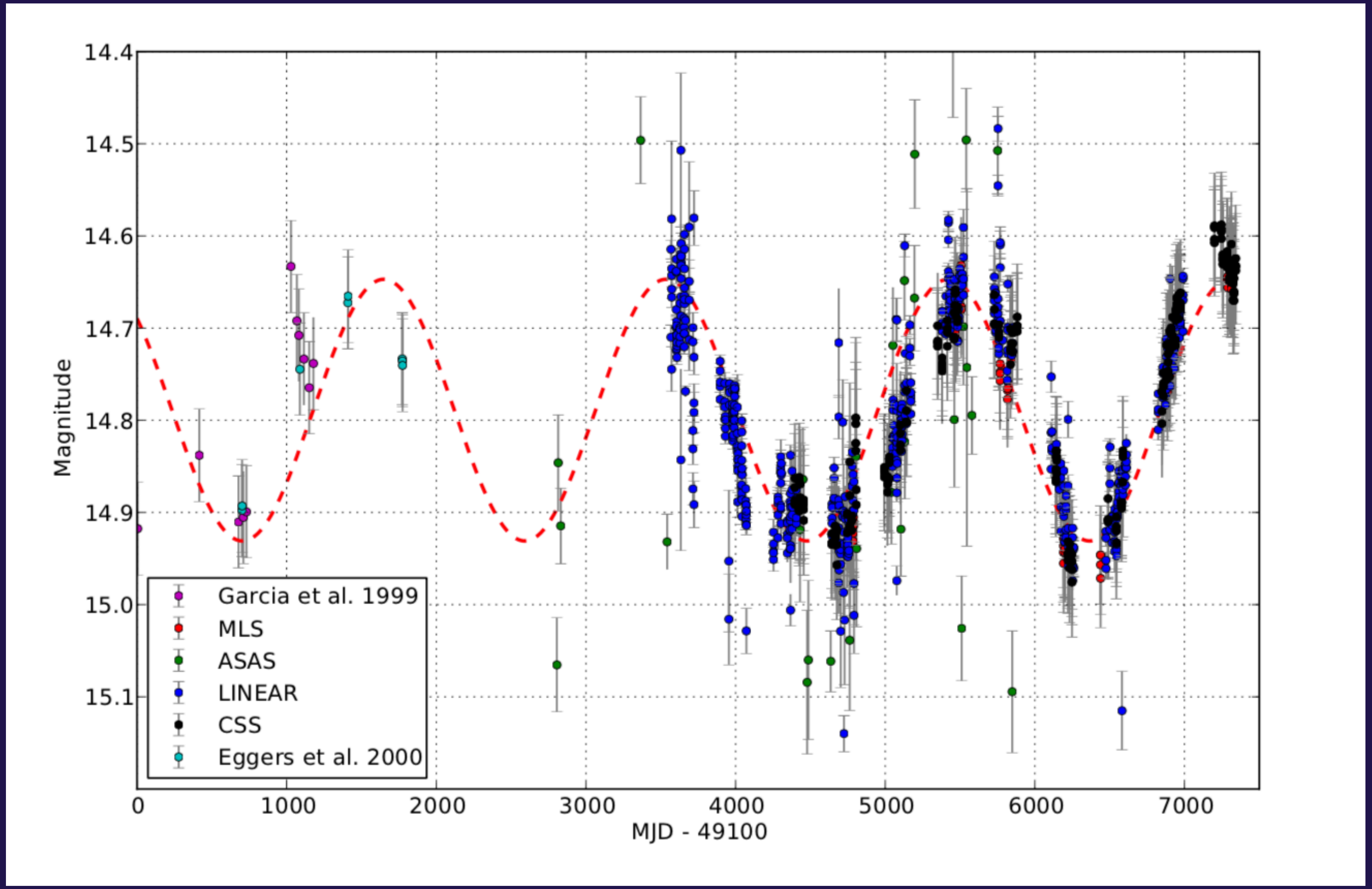
stochastic vs deterministic



Vaughan et al (2016)

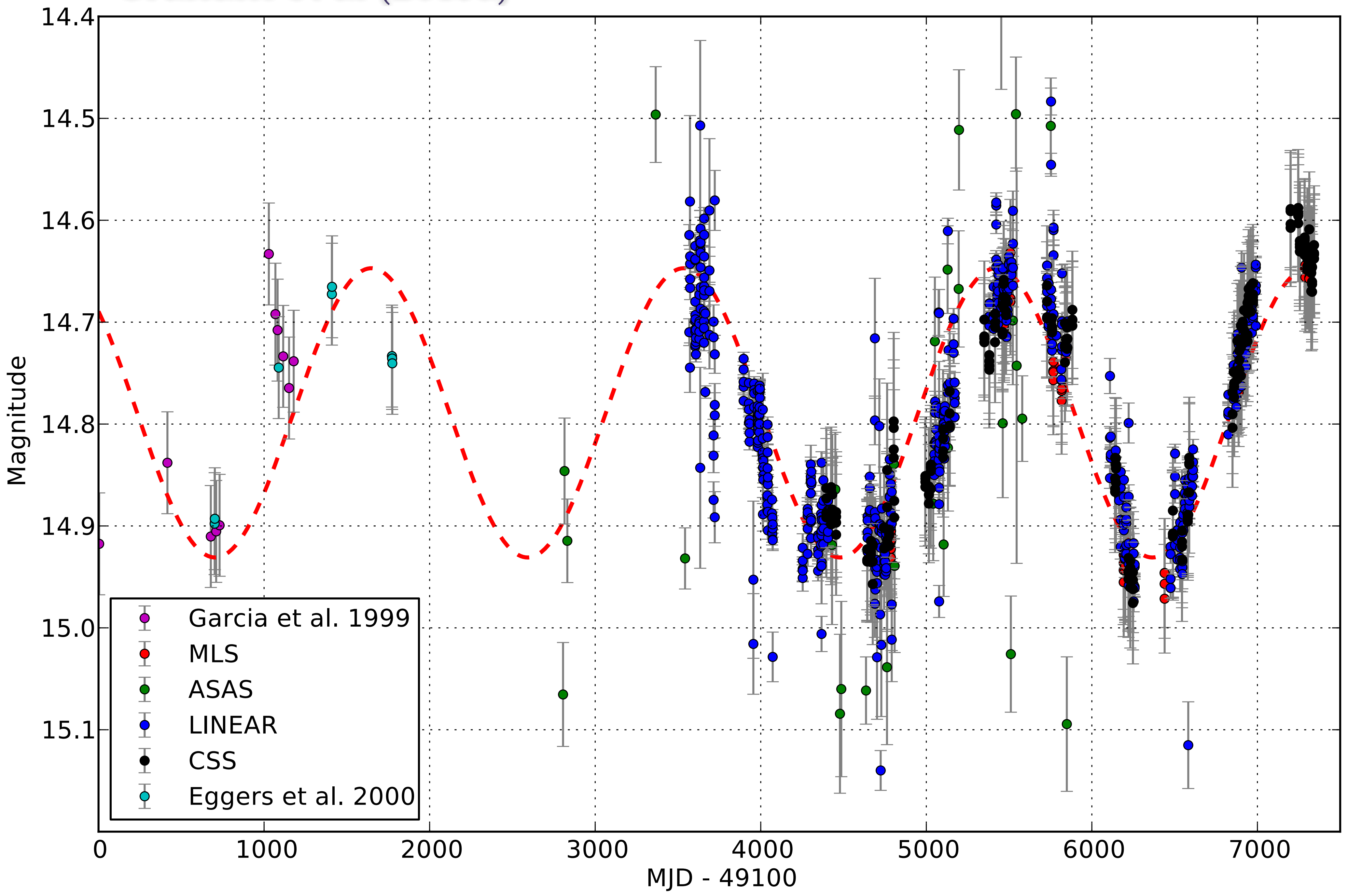
Our data is often ...

- unevenly sampled
- heteroscedastic



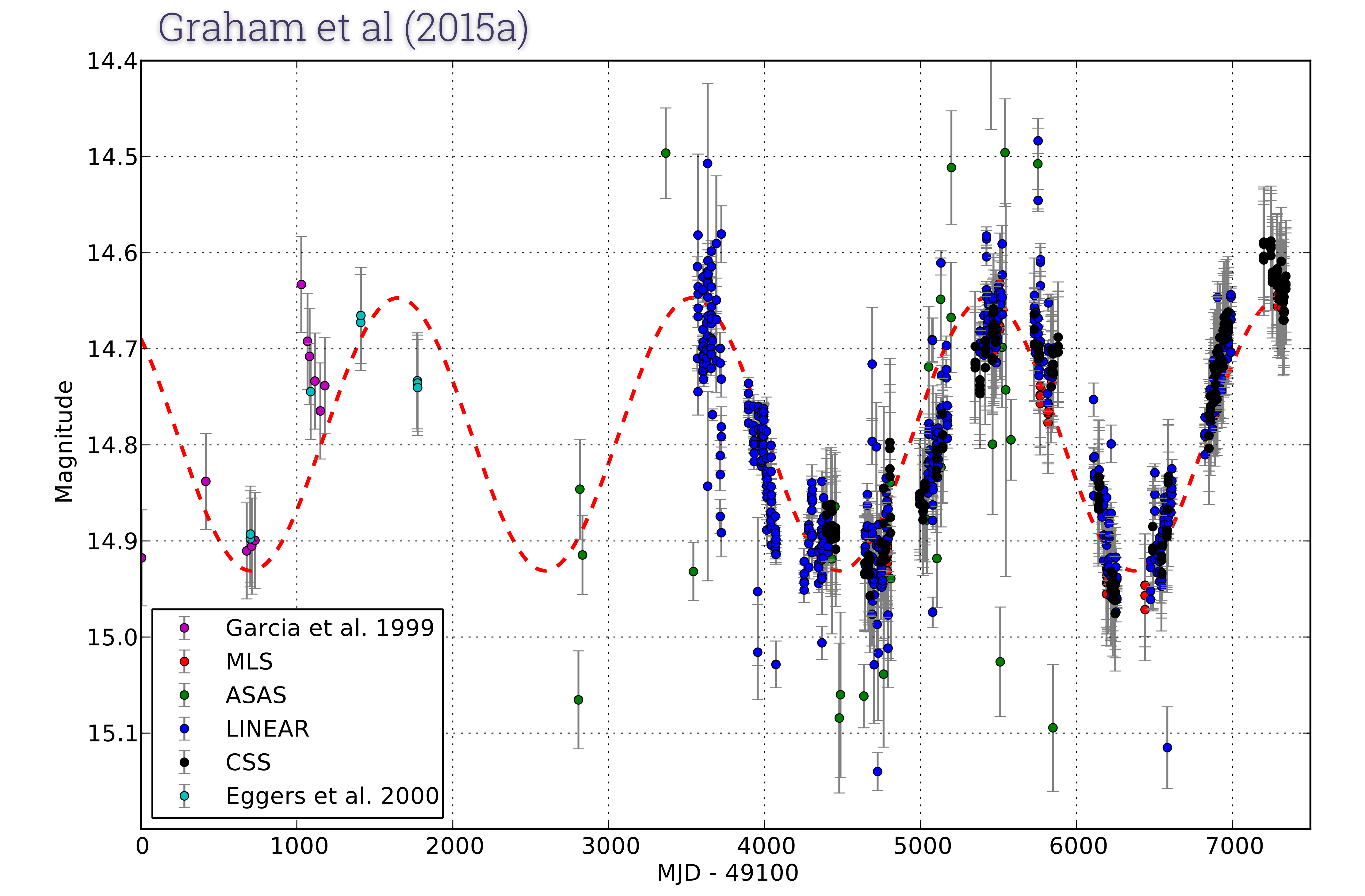
Graham et al (2015)

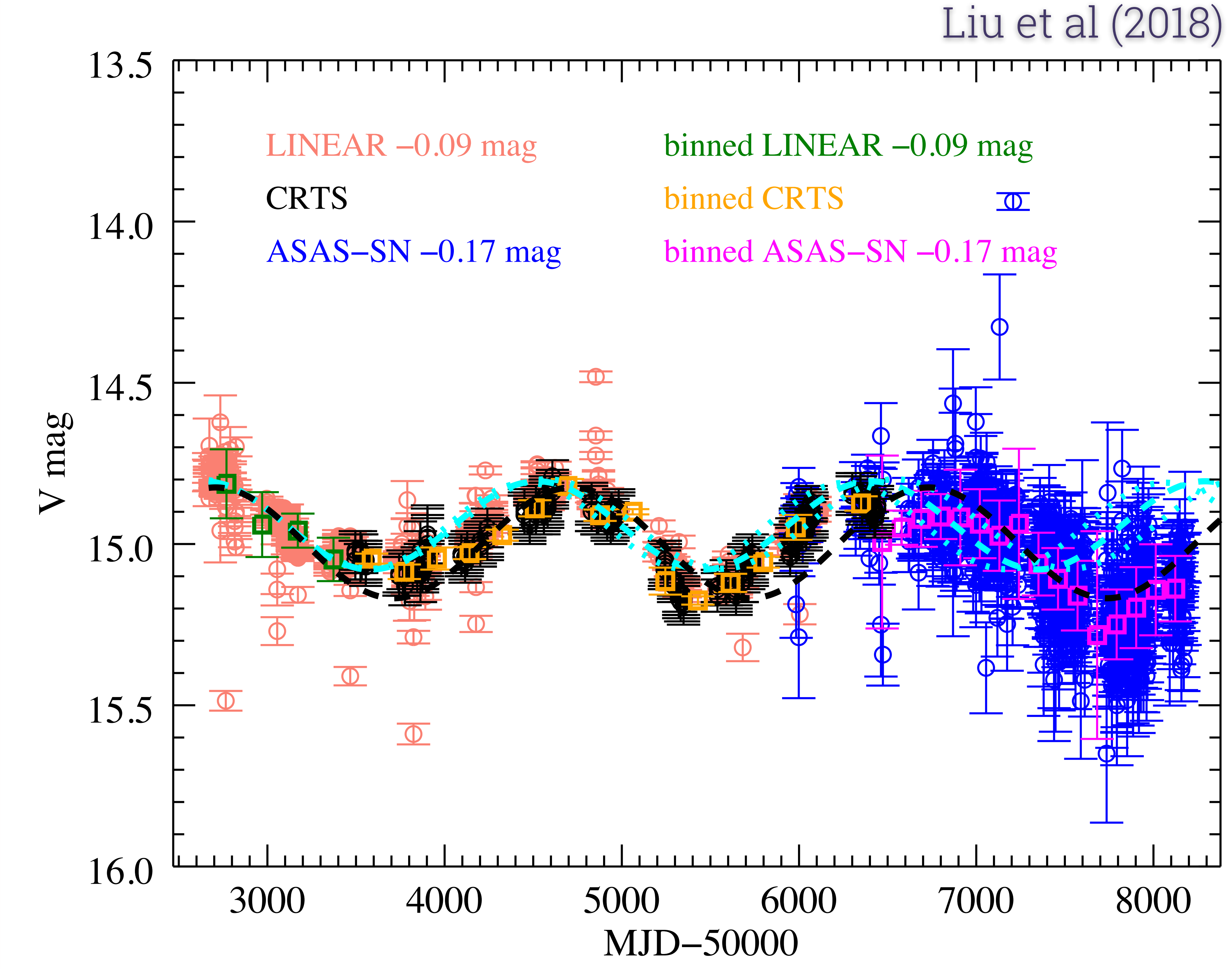
Finding Supermassive Black Hole Binaries: The curious case of PG 1302-102



Graham et al (2015a)

Finding Supermassive Black Hole Binaries: The curious case of PG 1302-102

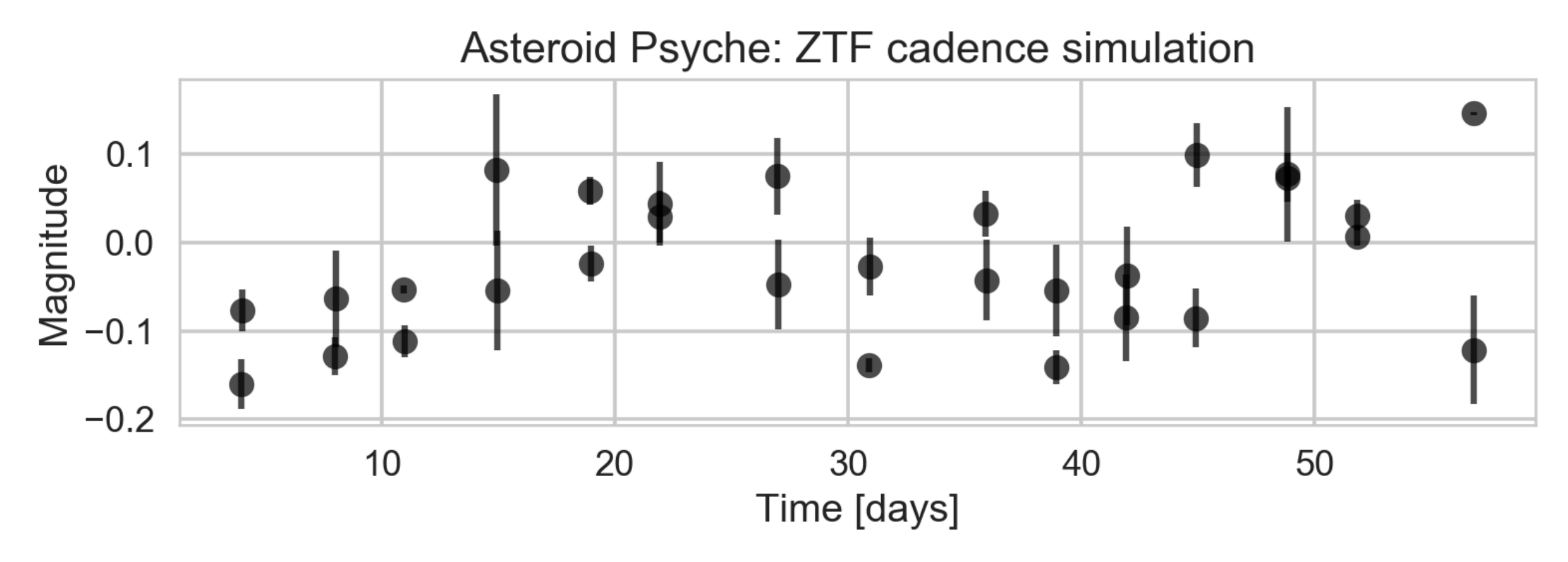

Graham et al (2015a)


Liu et al (2018)

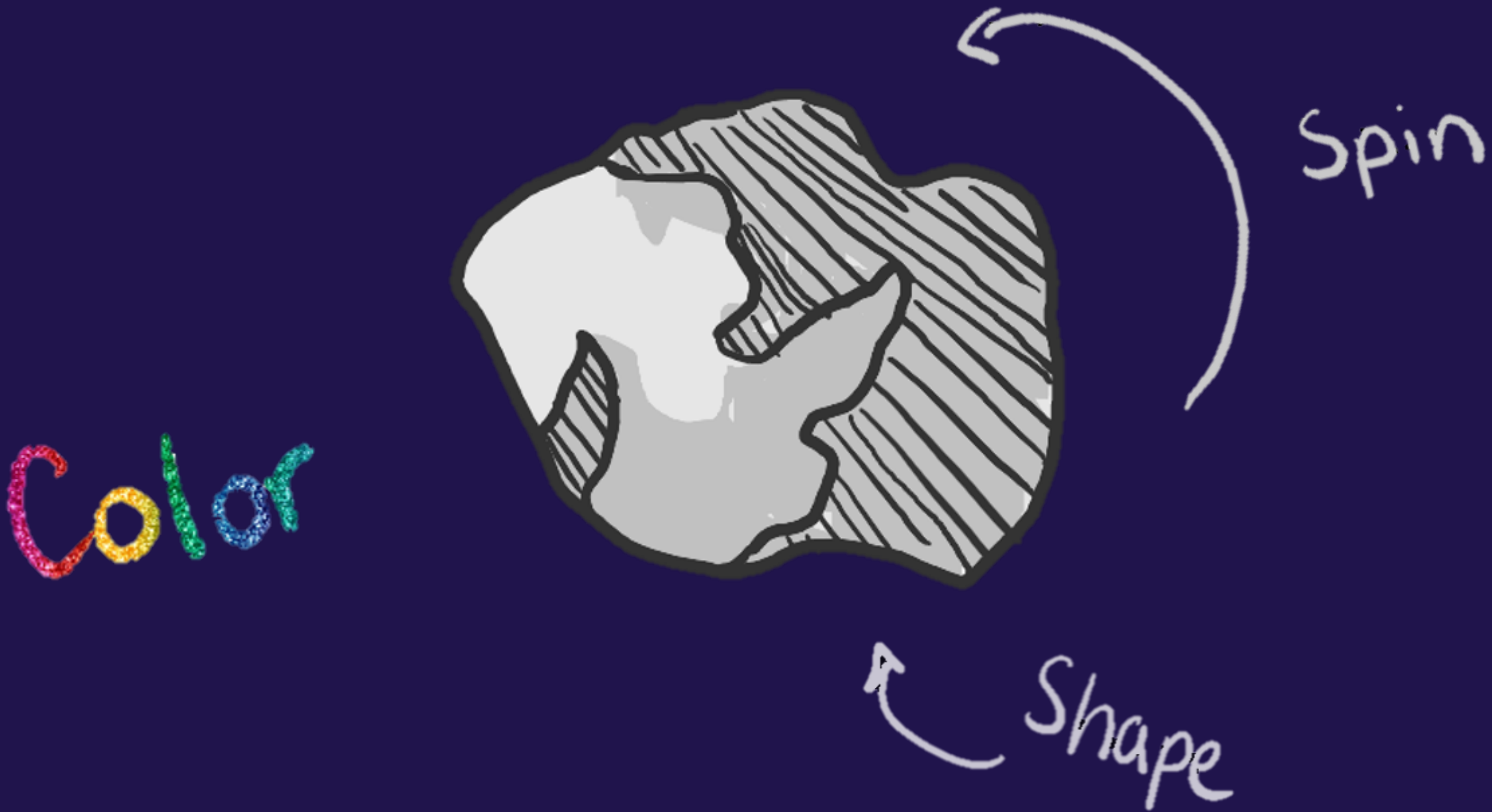
LSST will extend our baseline by a factor of ~1.5!

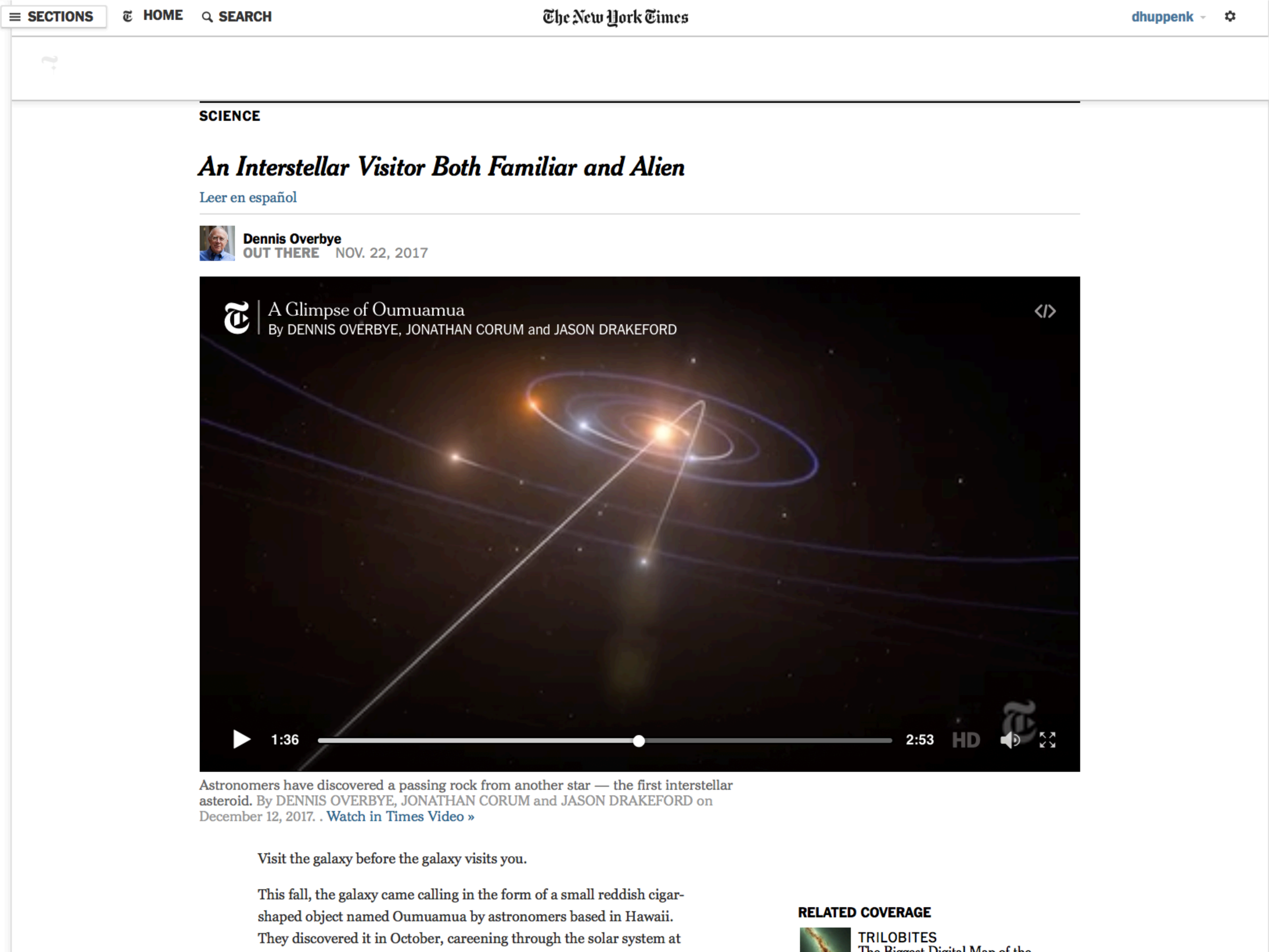
Our data is often ...

sparse



Asteroids





SCIENCE

An Interstellar Visitor Both Familiar and Alien

Leer en español

Dennis Overbye
OUT THERE NOV. 22, 2017

A Glimpse of Oumuamua
By DENNIS OVERBYE, JONATHAN CORUM and JASON DRAKEFORD

Astronomers have discovered a passing rock from another star — the first interstellar asteroid. By DENNIS OVERBYE, JONATHAN CORUM and JASON DRAKEFORD on December 12, 2017. . Watch in Times Video »

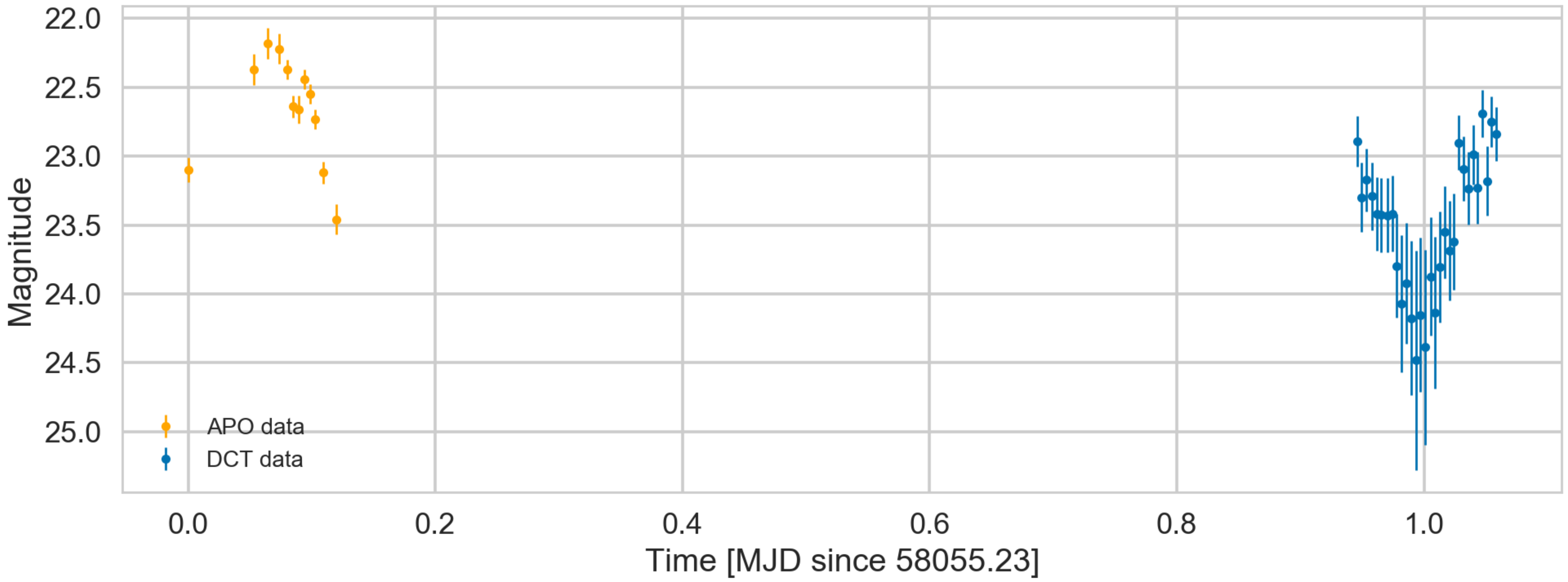
Visit the galaxy before the galaxy visits you.

This fall, the galaxy came calling in the form of a small reddish cigar-shaped object named Oumuamua by astronomers based in Hawaii. They discovered it in October, careening through the solar system at

RELATED COVERAGE

TRILOBITES
The Biggest Digital Map of the

1I'Oumuamua

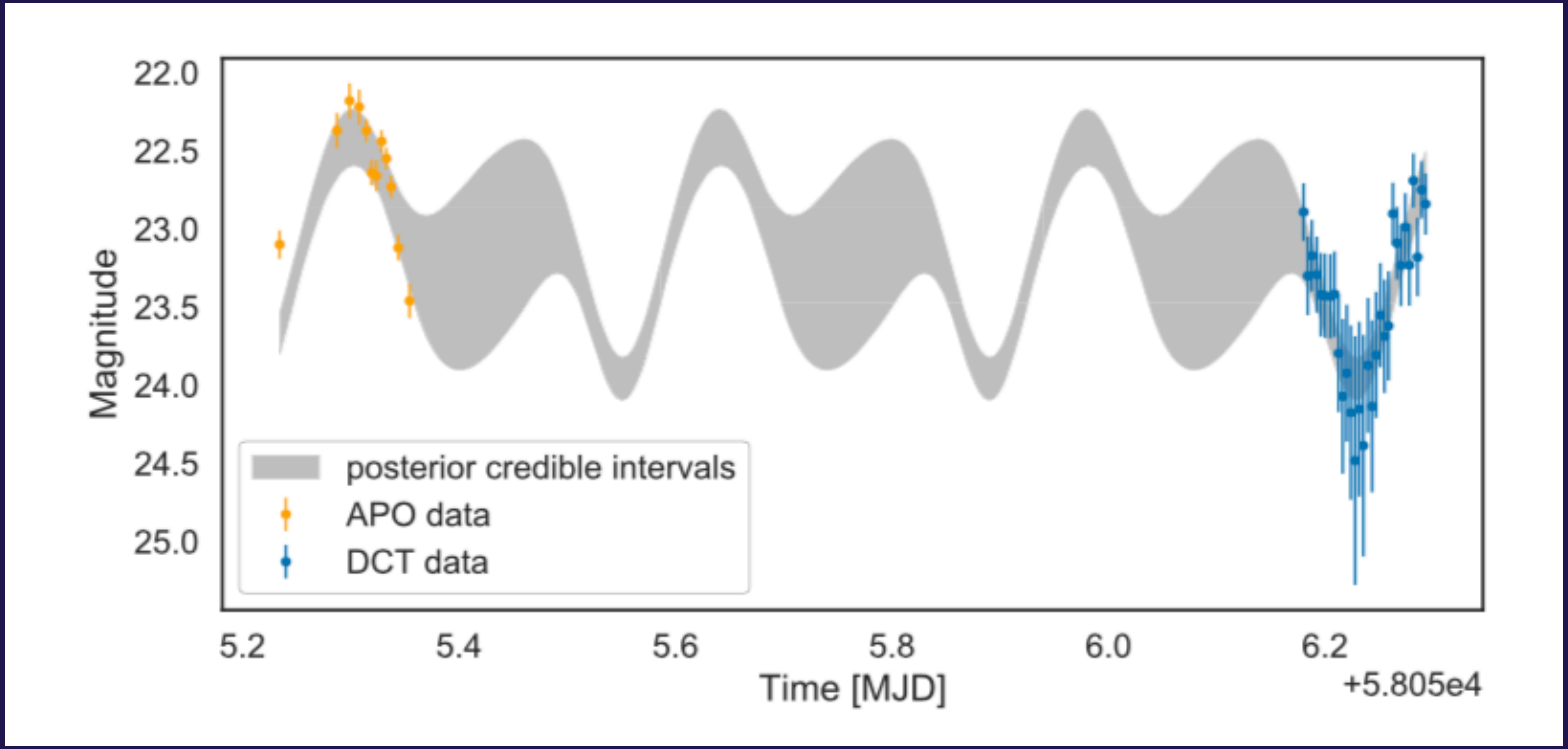


Discovery Channel Telescope +
Apache Point Observatory 3.5m

Model: Gaussian Process with periodic kernel

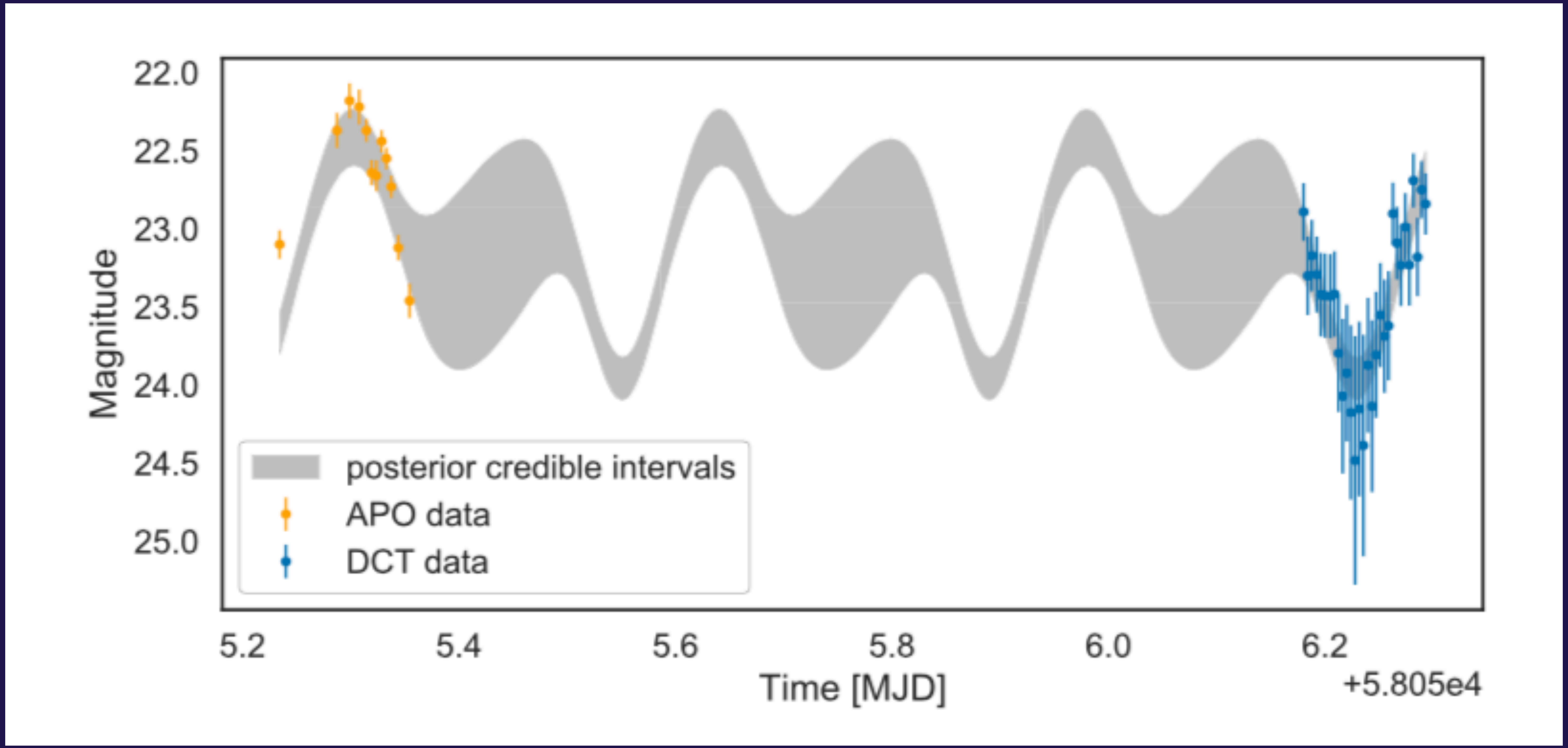
Bolin et al, incl Huppenkothen (2017)

Model: Gaussian Process with periodic kernel



Bolin et al, incl Huppenkothen (2017)

Model: Gaussian Process with periodic kernel

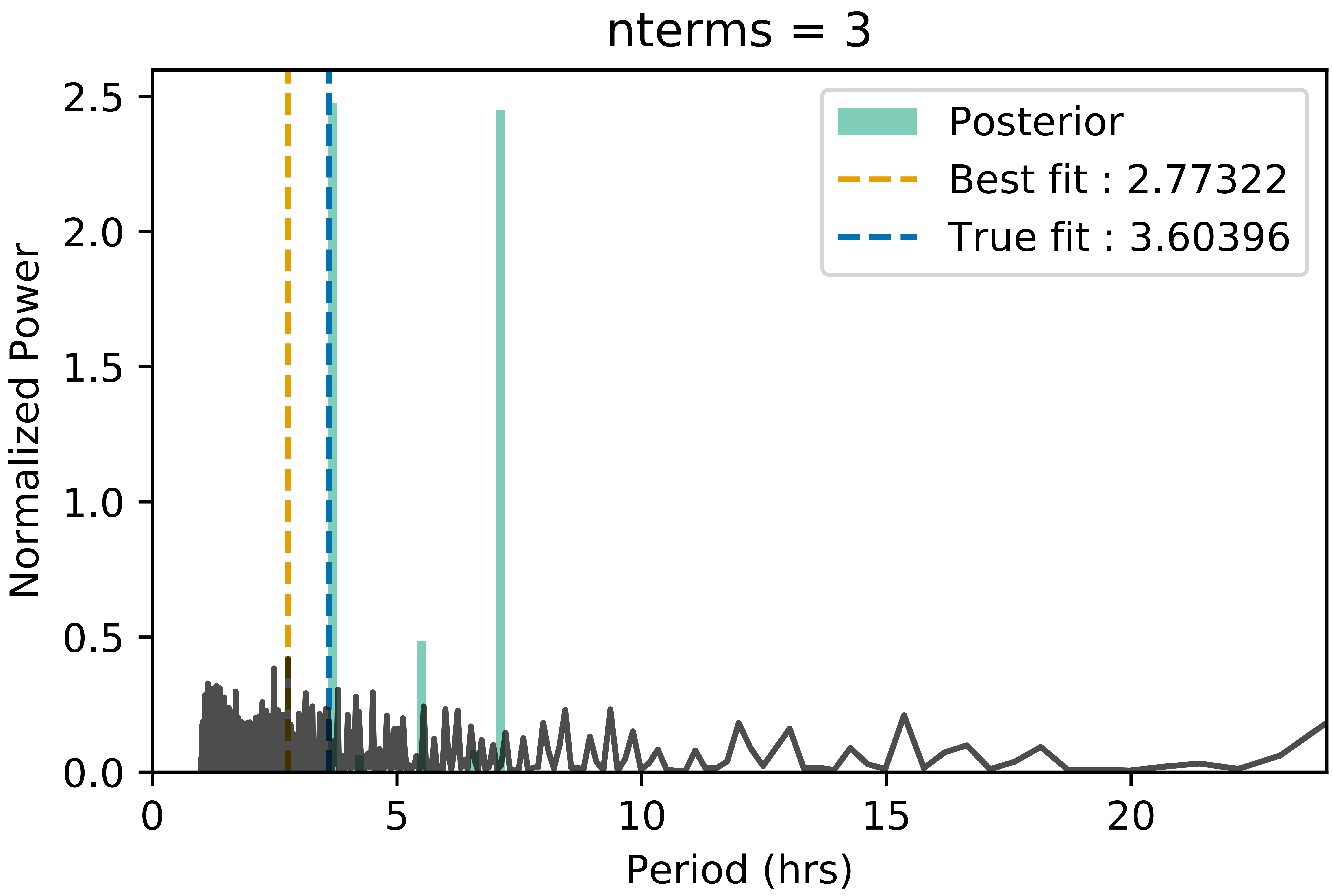


better period estimate with less data

Bolin et al, incl Huppenkothen (2017)

result: 1I’Oumuamua is **extremely elongated**, and **probably tumbling**

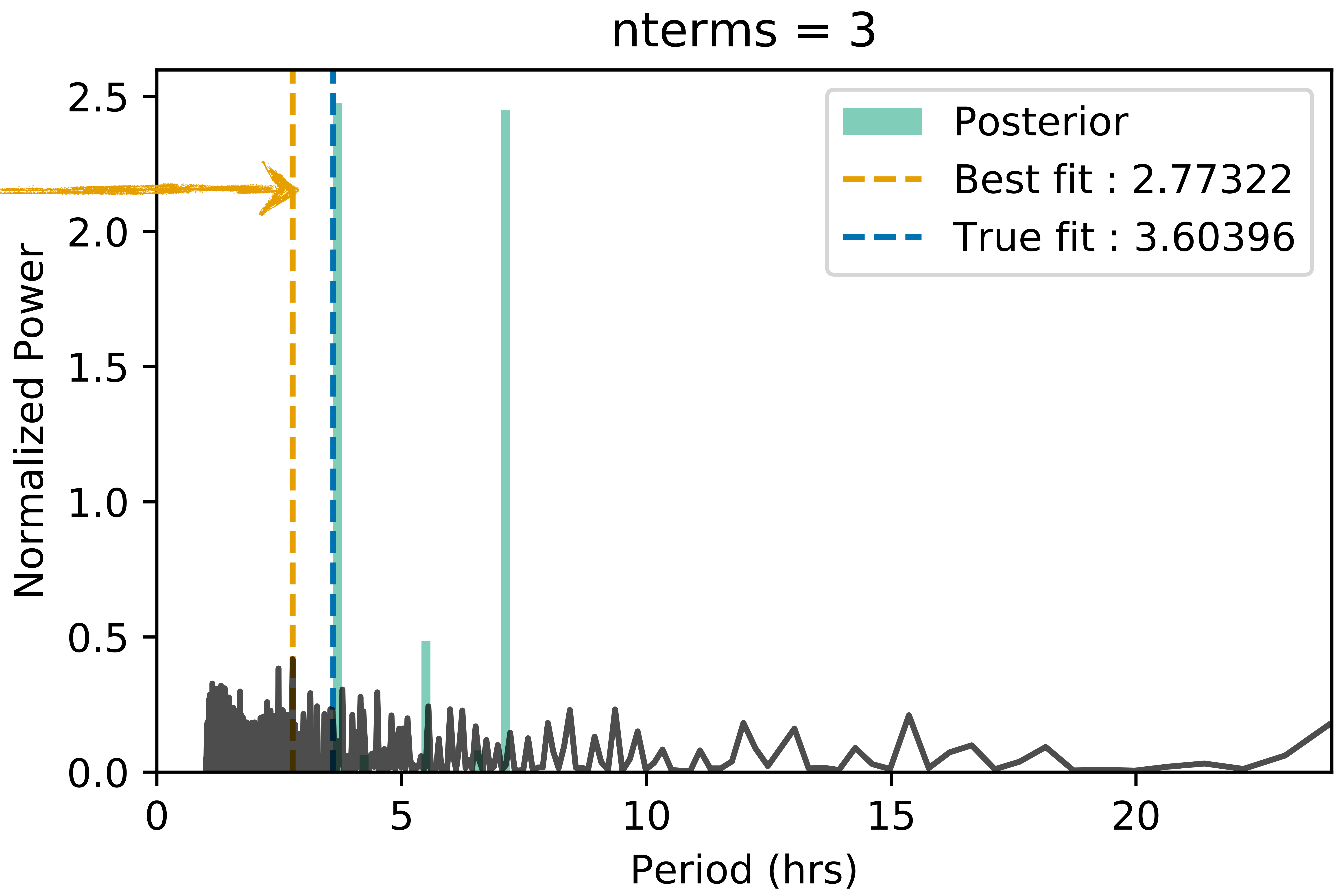
Gaussian Processes for Sparse Asteroid Light Curves



Lindberg, Huppenkothen et al (in prep)

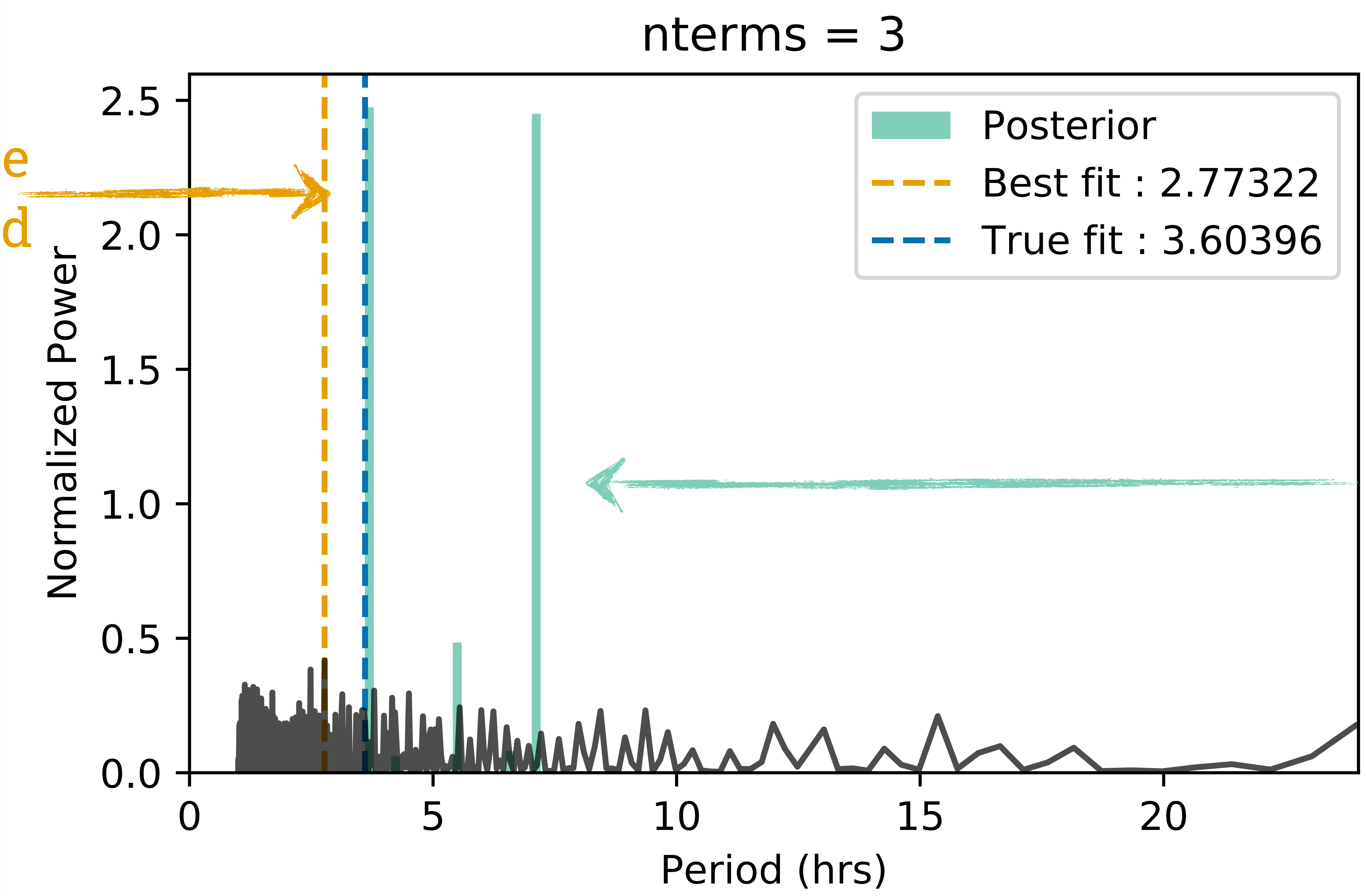
Gaussian Processes for Sparse Asteroid Light Curves

Lomb-Scargle best-fit period



Lindberg, Huppenkothen et al (in prep)

Gaussian Processes for Sparse Asteroid Light Curves

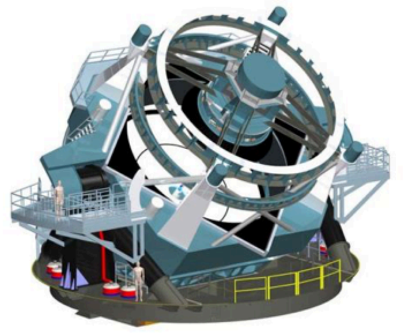


Lomb-Scargle best-fit period

posterior probability

Lindberg, Huppenkothen et al (in prep)

Training data sets are biased



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

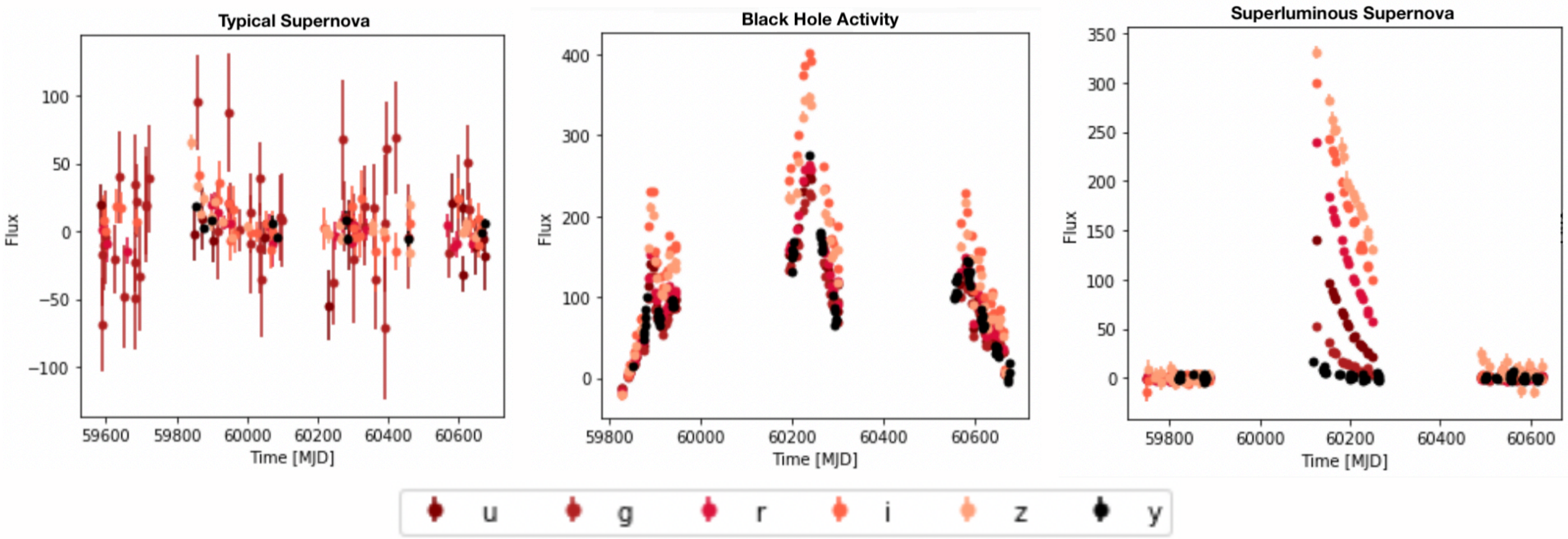
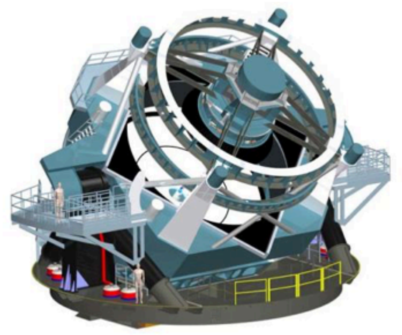


Figure Credit: Leah Fulmer

see also: The PLAsTiCC Team (2018)

Training data sets are biased

- 8000 objects in training data set: bright, low-redshift objects



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

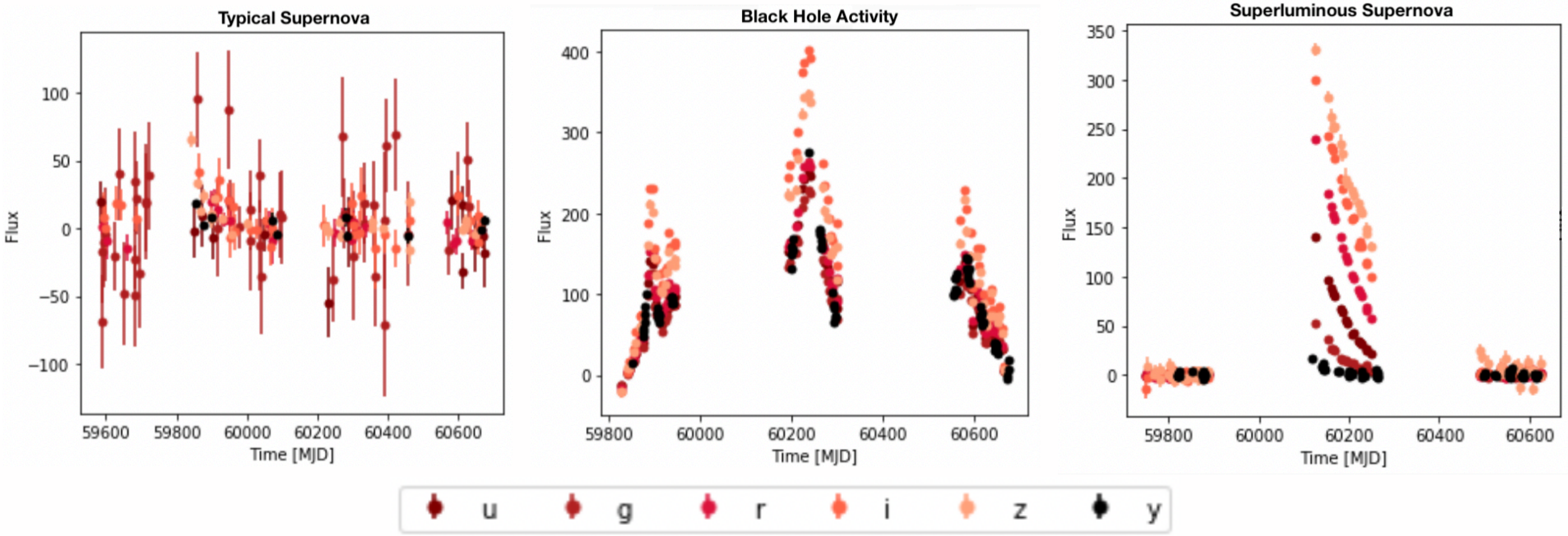
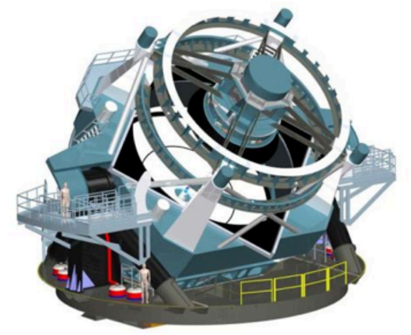


Figure Credit: Leah Fulmer

see also: The PLAsTiCC Team (2018)

Training data sets are biased

- 8000 objects in training data set: bright, low-redshift objects
- 3.5 million objects in test set: fainter, more distant objects



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

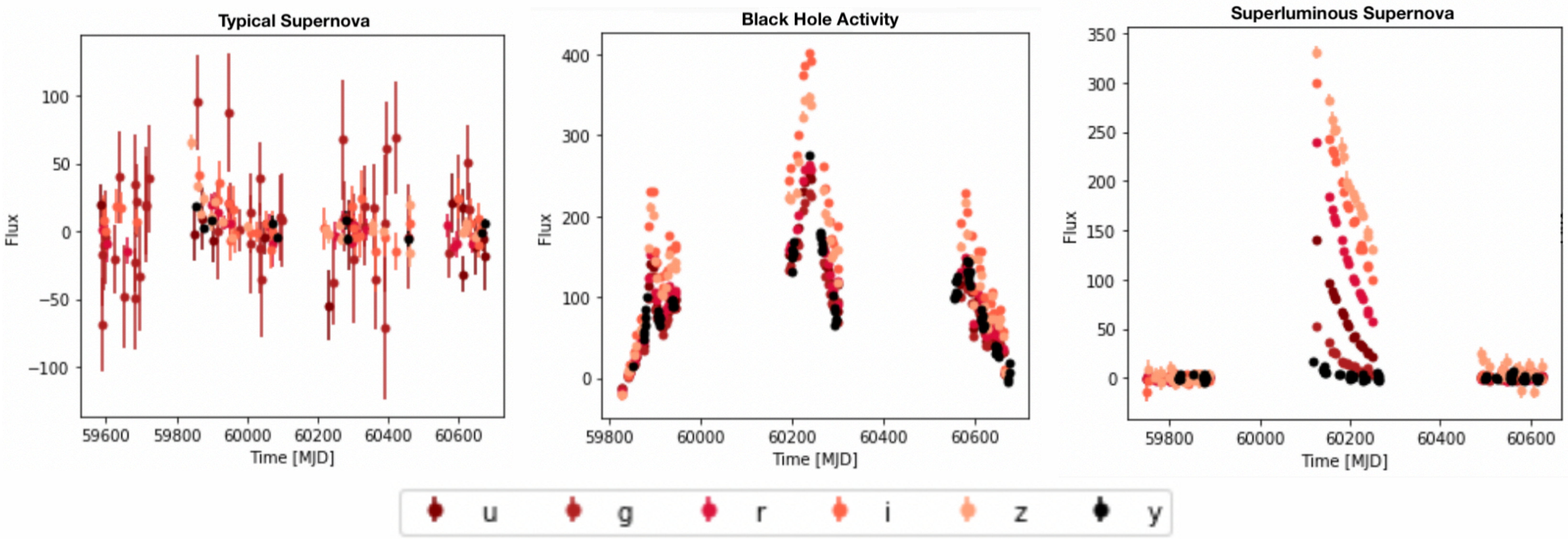
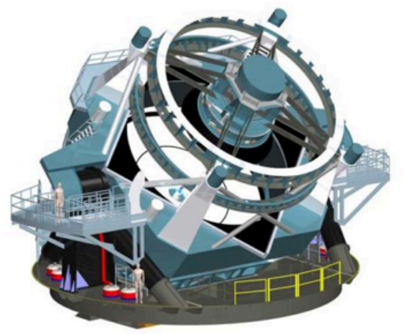


Figure Credit: Leah Fulmer

see also: The PLAsTiCC Team (2018)

Training data sets are biased

- 8000 objects in training data set: bright, low-redshift objects
- 3.5 million objects in test set: fainter, more distant objects
- training data properties are non-representative of the test set



PHOTOMETRIC LSST ASTRONOMICAL TIME-SERIES CLASSIFICATION CHALLENGE (PLASTICC)

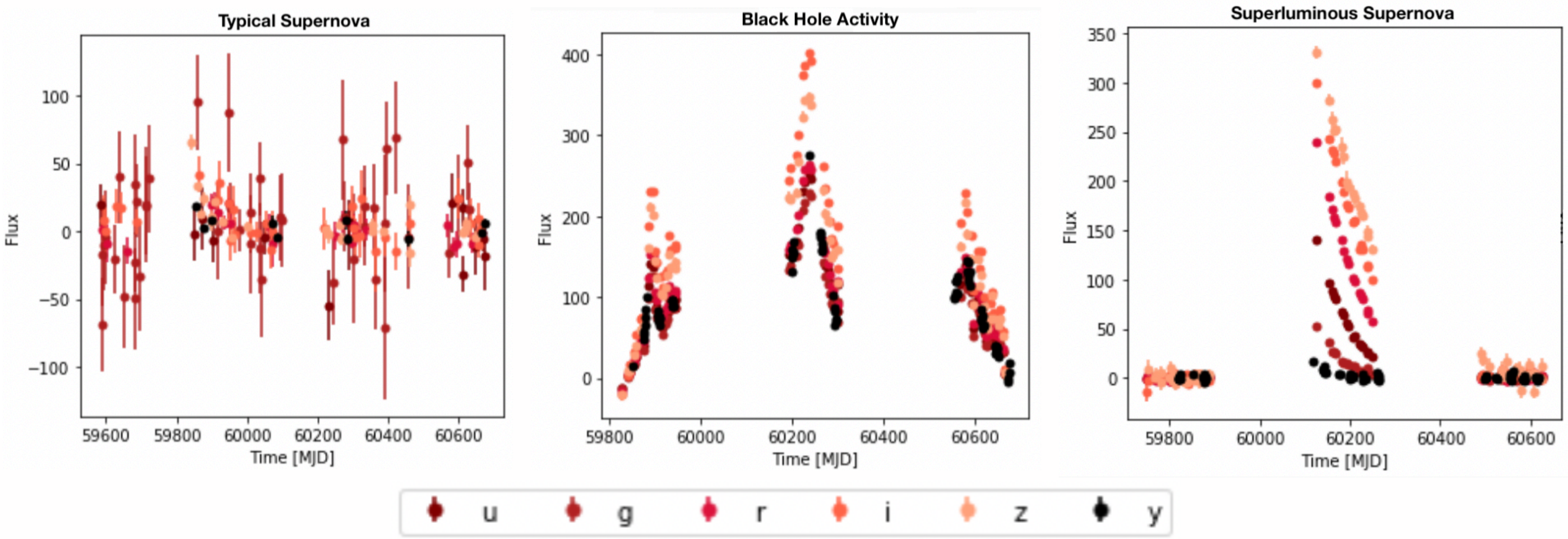


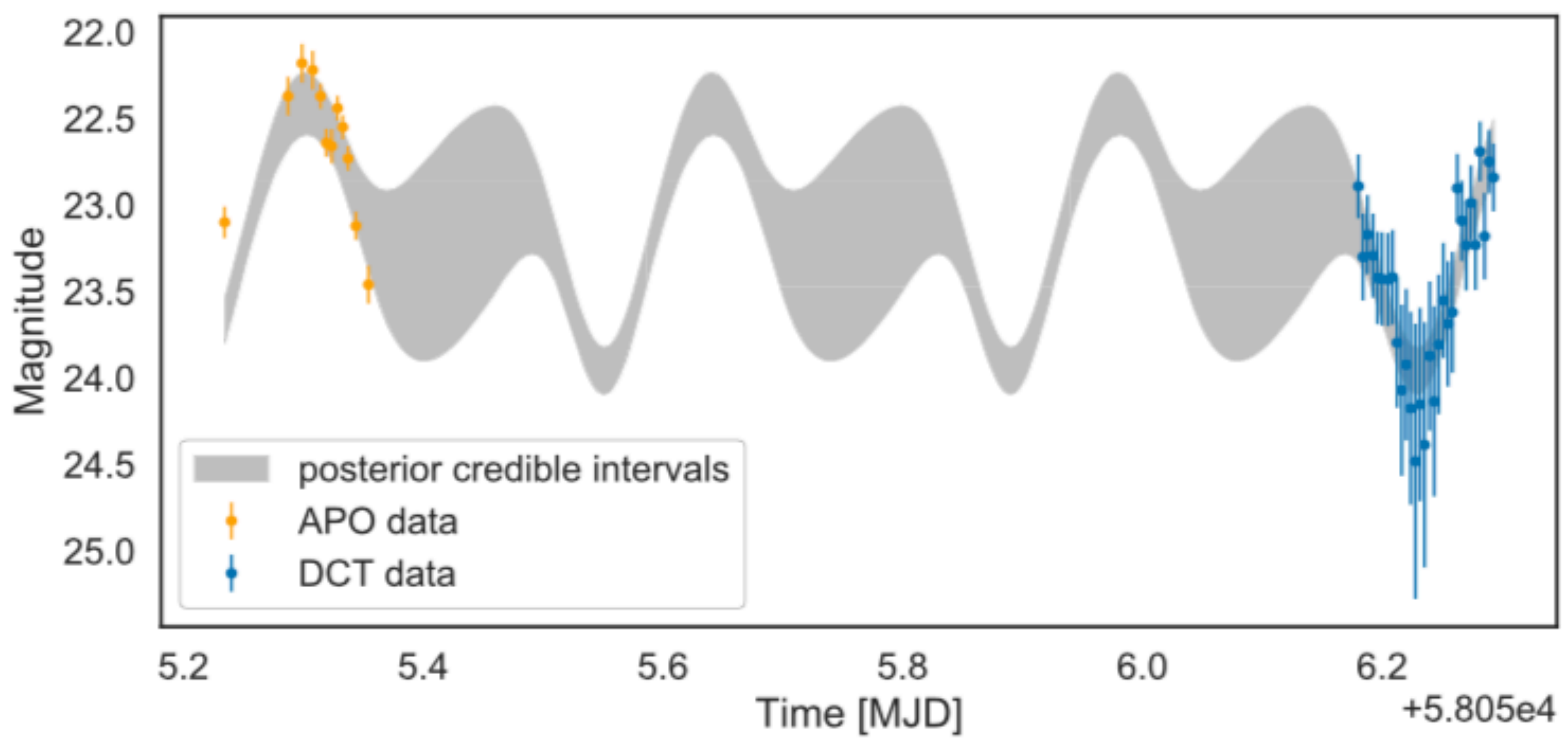
Figure Credit: Leah Fulmer

see also: The PLAsTiCC Team (2018)

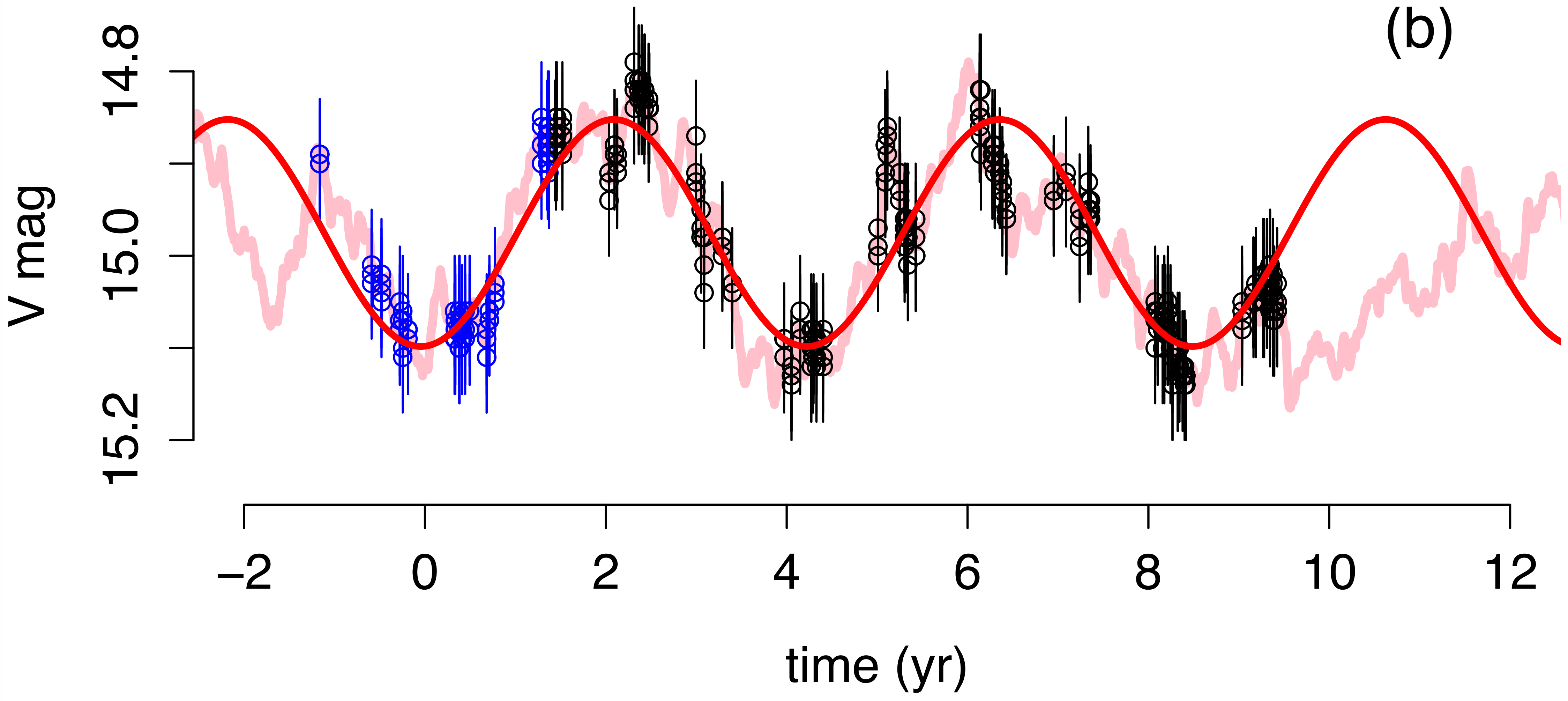
(Some) Current Major Challenges in Time Domain Astronomy

- uneven sampling
- heteroscedasticity
- non-stationarity
- multi-wavelength data sets
- modeling multiple dimensions simultaneously (time, energy, polarimetry, …)
- computational scaling
- classification of sources and time series

Asteroid light curves are a simpler analogue to finding periods in AGN



Bolin et al, incl Huppenkothen (2017)

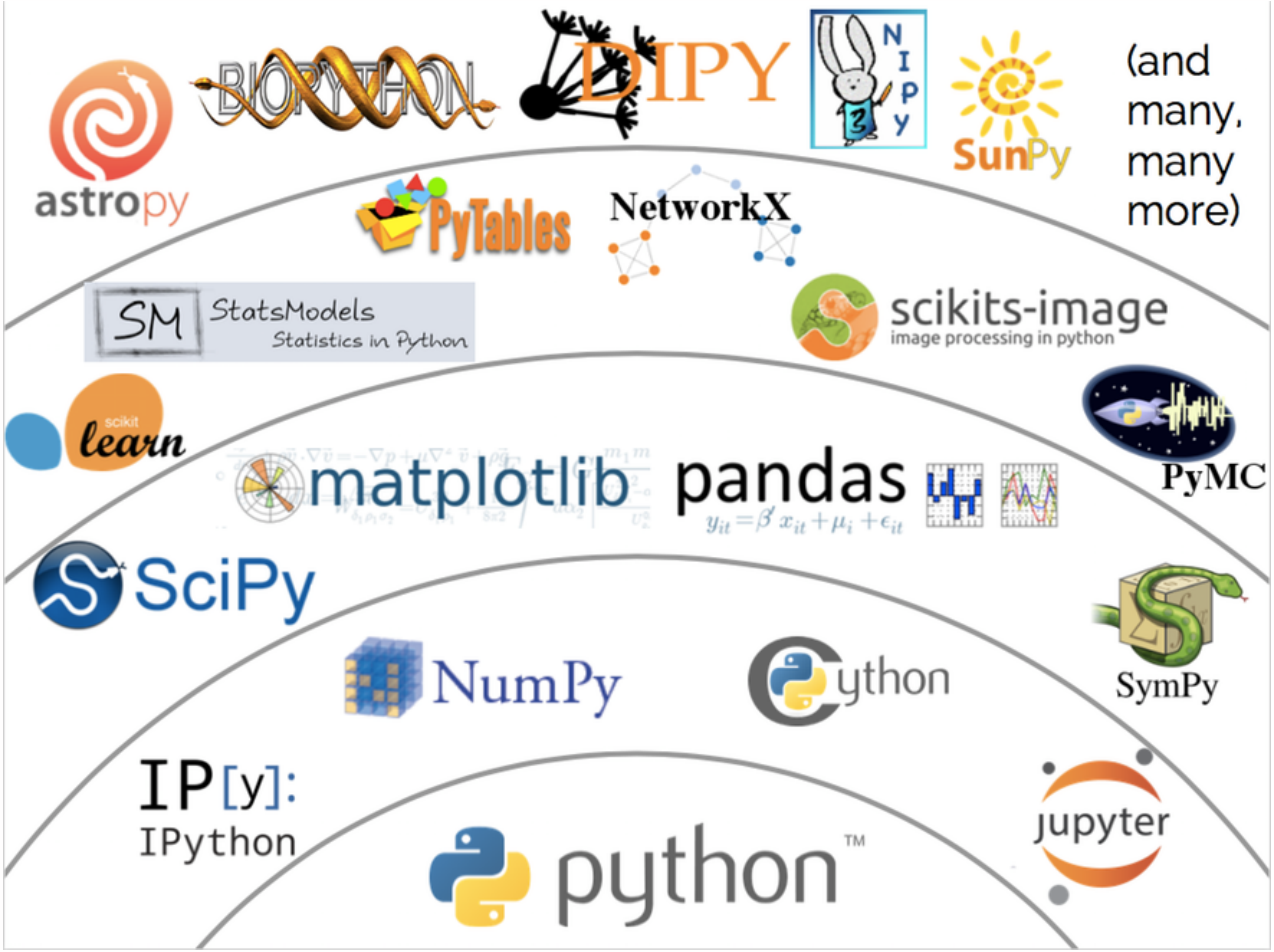


Vaughan et al , incl
Huppenkothen (2016)

The good news: many of the really hard problems have simpler equivalents in other areas of astronomy or even other scientific domains

Solving data science challenges through communities of practice

see also: http://msdse.org/reports/



Credit: Jake VanderPlas

https://speakerdeck.com/jakevdp/the-unexpected-effectiveness-of-python-in-science

Stingray

The Next Generation Spectral-Timing Software

- 3 lead developers/maintainers (Huppenkothen, Bachetti, Stevens)
- ~10 contributors
- 6 completed Google Summer of Code Projects
- astropy-affiliated project

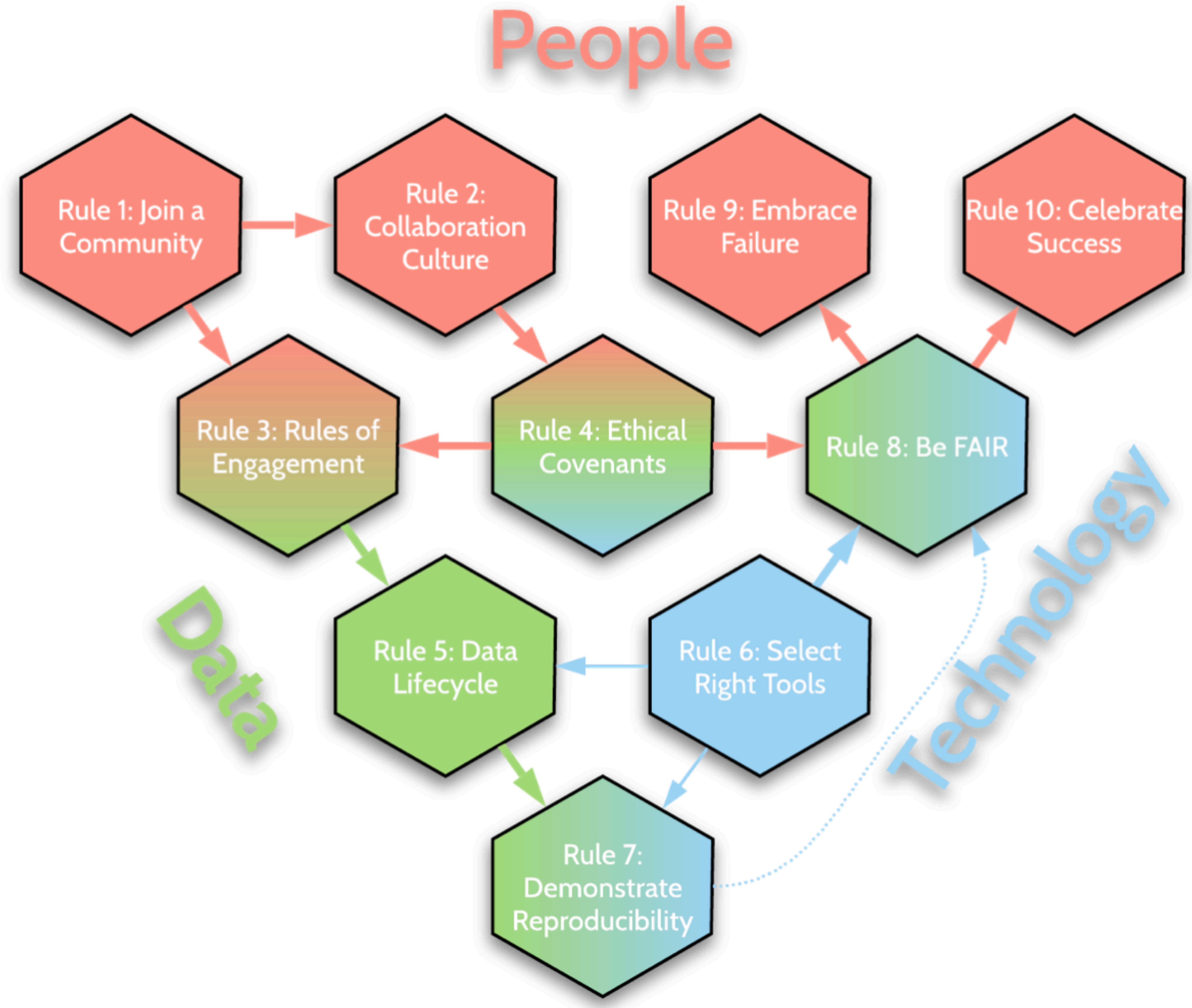
http://stingraysoftware.github.io

Huppenkothen et al (2019)

The largest data science challenges will be solved through collaboration across fields

see also: http://msdse.org/reports/

Ten Simple Rules for Researchers Engaging in Data Science and Domain Science Collaboration


People
Rule 1: Join a Community
Rule 2: Collaboration Culture
Rule 9: Embrace Failure
Rule 10: Celebrate Success
Rule 3: Rules of Engagement
Rule 4: Ethical Covenants
Rule 8: Be FAIR
Data
Rule 5: Data Lifecycle
Rule 6: Select Right Tools
Technology
Rule 7: Demonstrate Reproducibility

Nord // Lemon Labs (in prep)

#Astro Hack Week

#Astro Hack Week

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

credit: Anthony Arendt

Hackweek Mission

Collaboration

Education

Building Software

Networking

Projects

Community

GEOHACKWEEK 2018

WORKSHOP ON GEOSPATIAL DATA SCIENCE
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE
SEPT 10 - 14, 2018

https://geohackweek.github.io

NEUROHACKADEMY

SUMMER SCHOOL IN NEUROIMAGING AND DATA SCIENCE

University of Washington eScience Institute

July 30th - August 10th, 2018

https://neurohackweek.github.io

WATERHACKWEEK 2019

WORKSHOP ON WATER DATA SCIENCE
UNIVERSITY OF WASHINGTON ESCIENCE INSTITUTE
MARCH 25-29, 2019

APPLICATIONS ARE OPEN
ANNOUNCEMENTS WILL BE MADE SHORTLY.

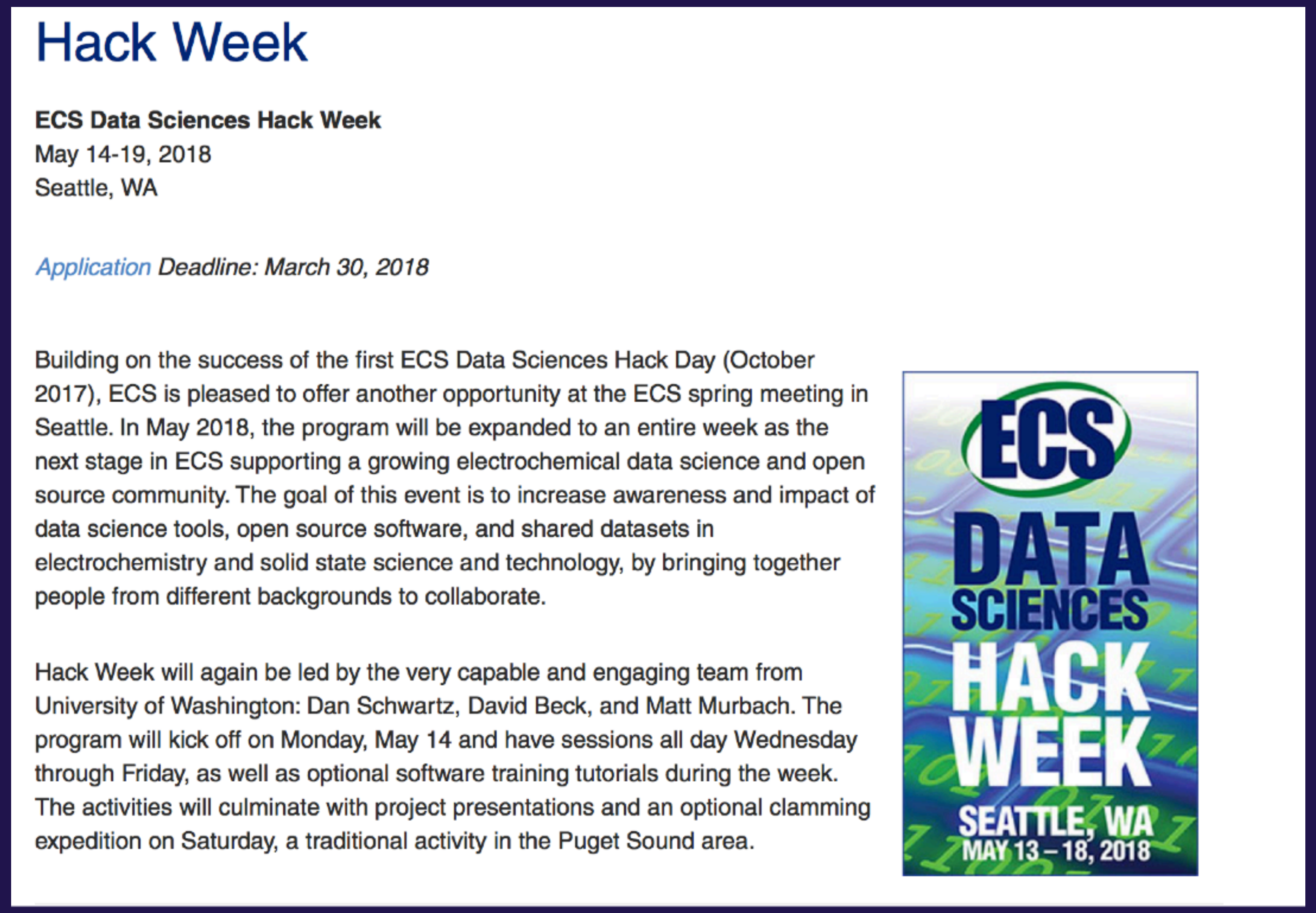
https://waterhackweek.github.io

OCEANHACKWEEK 2018

DATA SCIENCE + OCEANOGRAPHY
UNIVERSITY OF WASHINGTON
AUG. 20 - 24, 2018

APPLICATIONS HAVE BEEN CLOSED.
ANNOUNCEMENTS WILL BE MADE SHORTLY.

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

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

Experiment

ASTRO
HACK
WEEK
Häck Atlas - let others know what you're working on
Door
Projector Screen
Photo-Z
Participant Selection
Podium
Breakout
ML
COFFE E
Door
Deep Learning
LUNCH SET-UP

Evaluate

credit: eScience Institute

ASTRO HACK WEEK 2019

AUG 26 — AUG 30, 2019

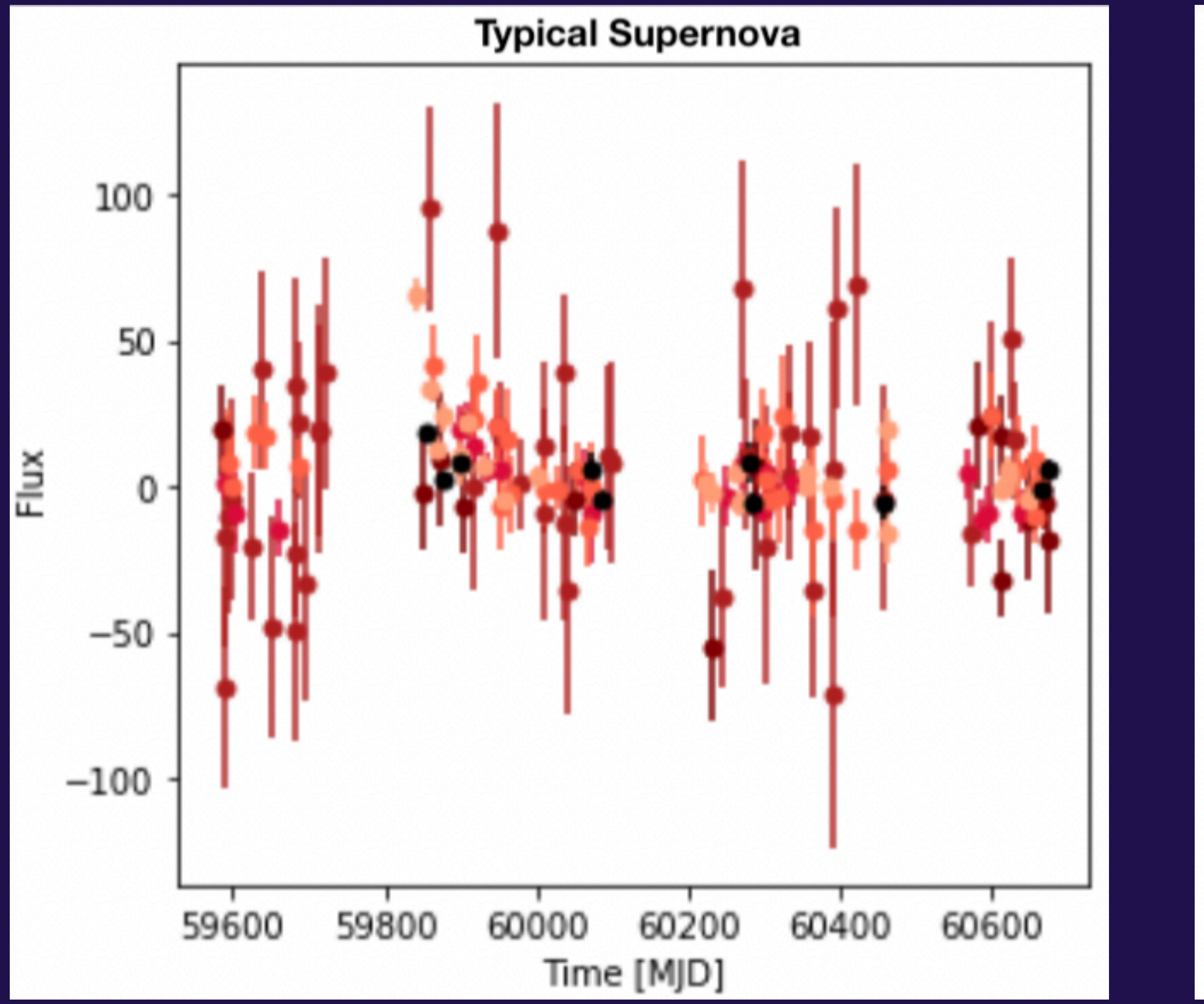
KAVLI INSTITUTE FOR COSMOLOGY @ CAMBRIDGE UNIVERSITY IN CAMBRIDGE, UK

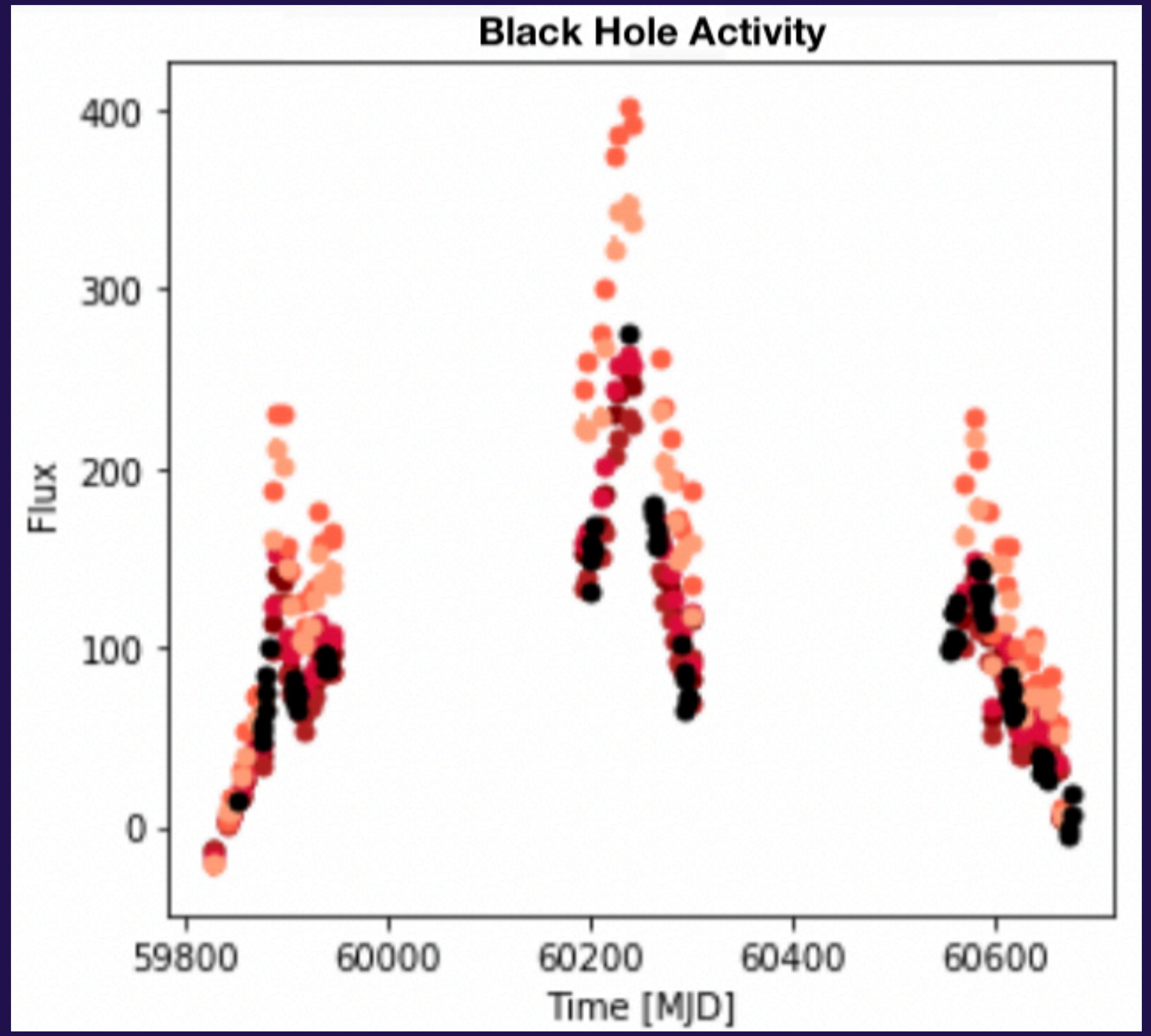
stay tuned for Astro Hack Week 2020!

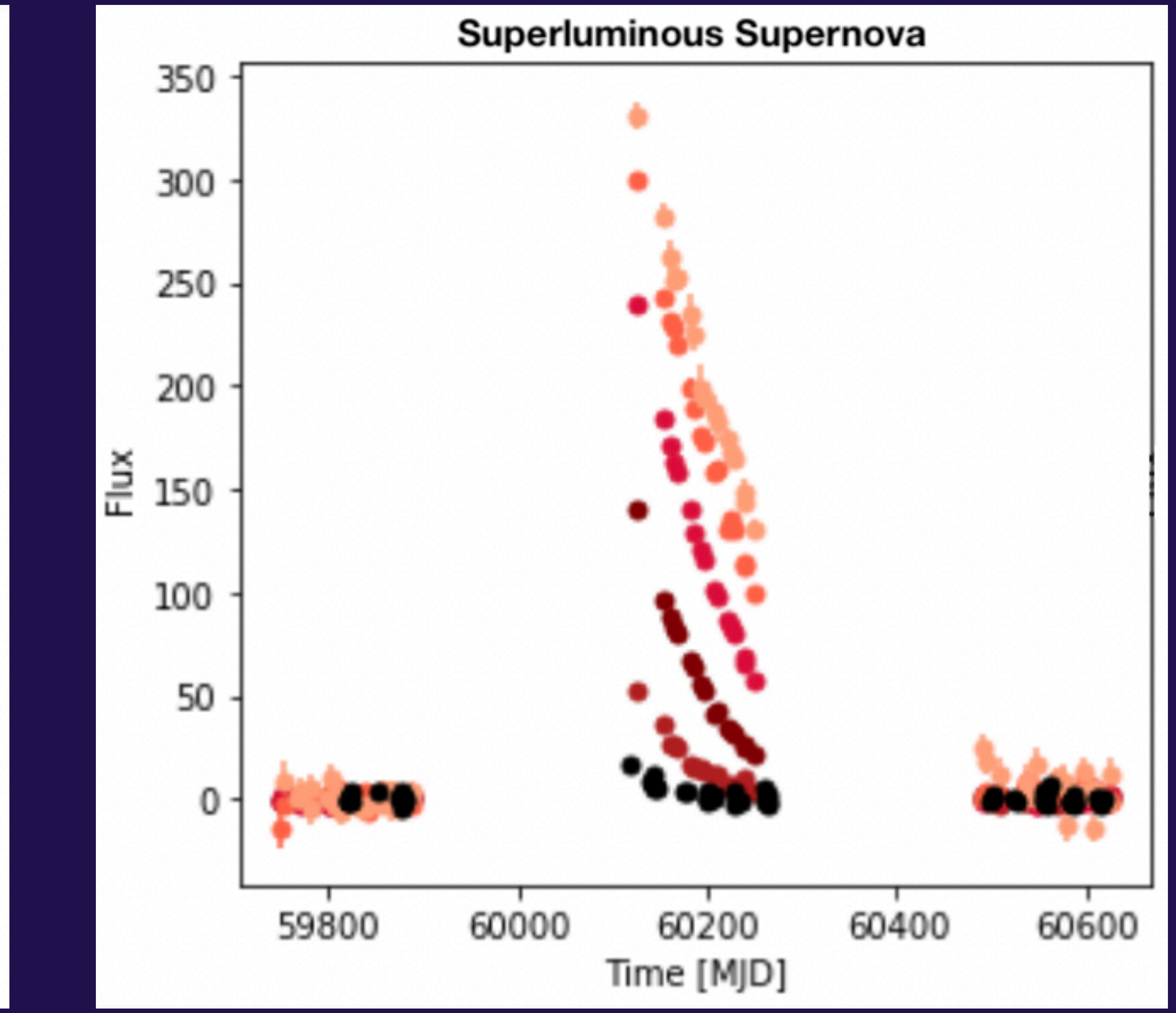
http://astrohackweek.org

Conclusions

Astronomical time domain data sets are complex, unevenly sampled, heteroscedastic, sparse and biased

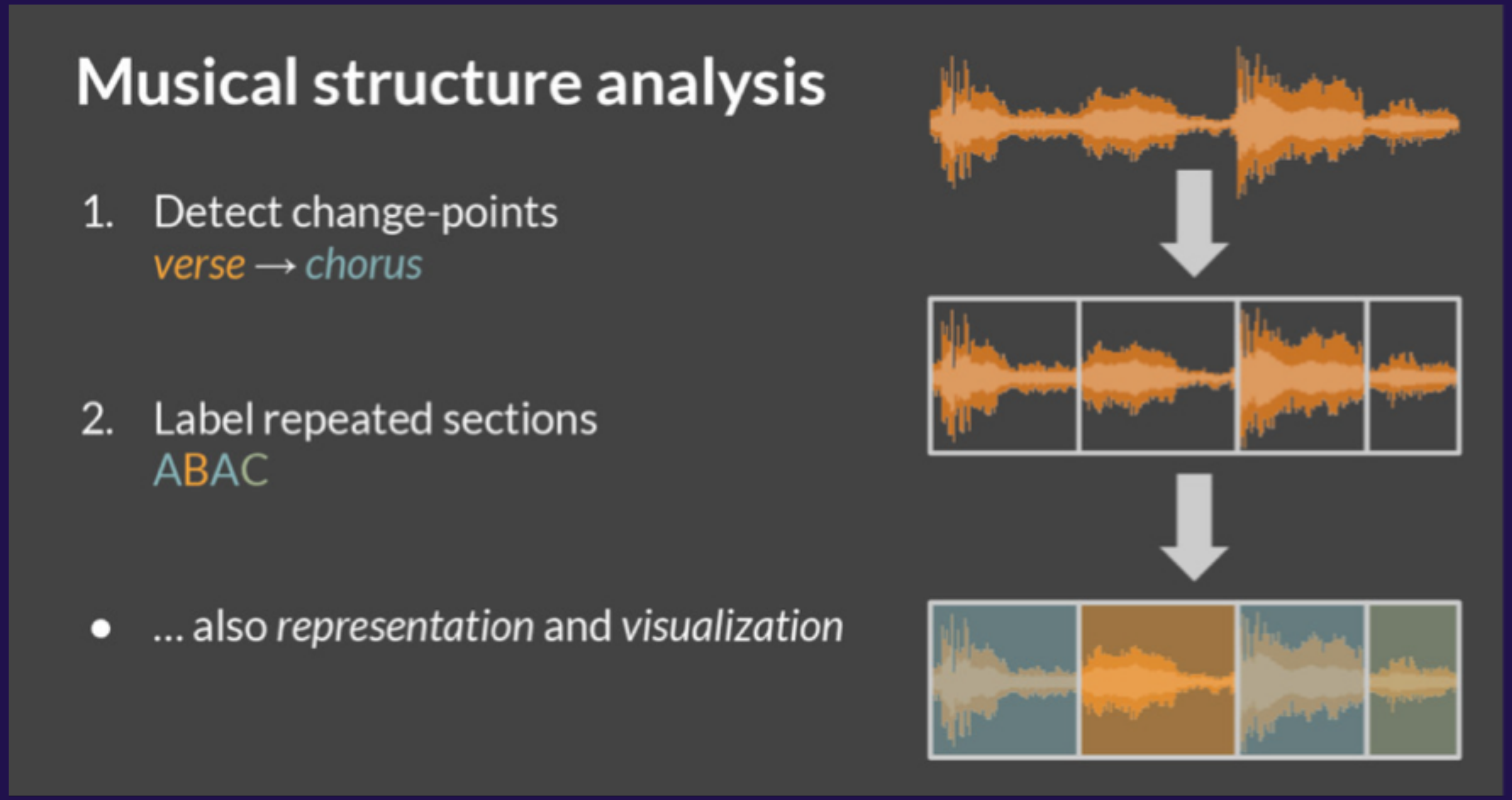

Typical Supernova


Black Hole Activity


Superluminous Supernova

This often makes the application of standard tools difficult

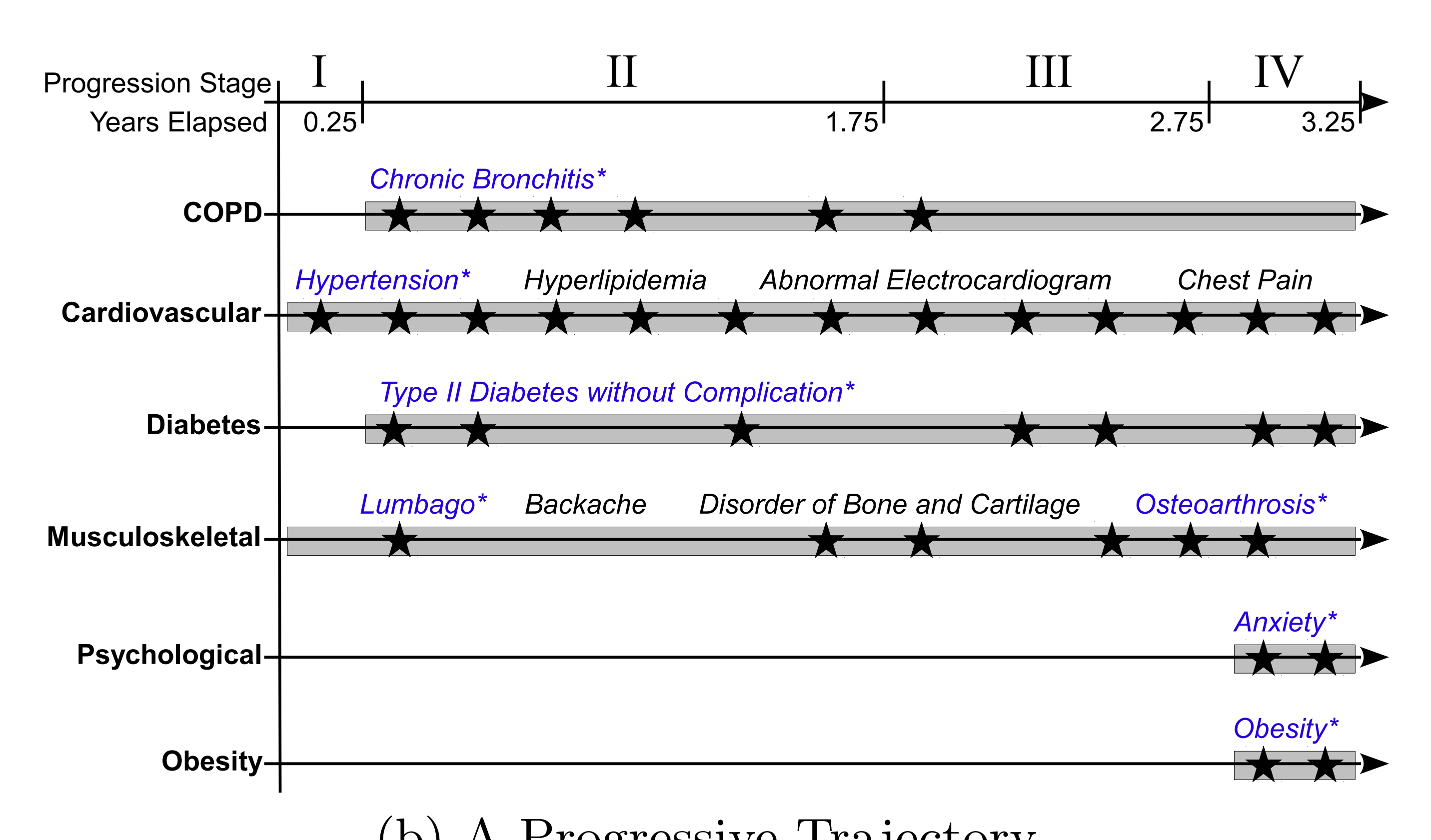
There are many data analysis challenges that are shared across scientific domains



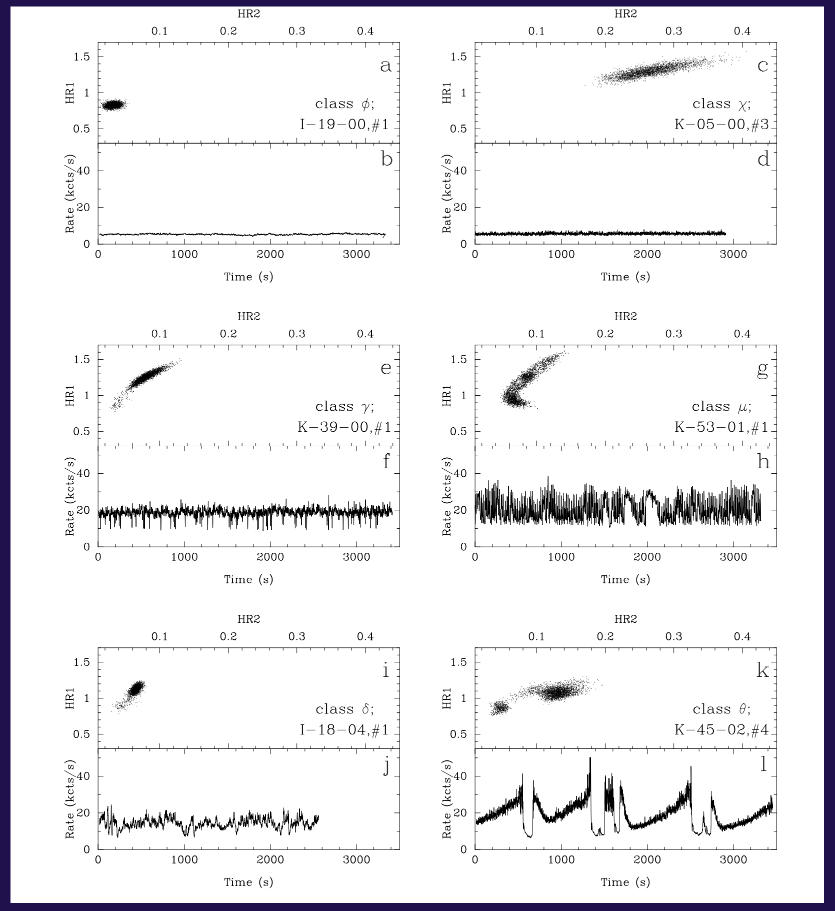
Musical structure analysis

1. Detect change-points
   verse ⟶ *chorus*
2. Label repeated sections
   ABAC

- … also *representation and visualization*



(b) A Progressive Trajectory



Data science provides shared venues and a common language to solve these problems across domains

One of the major challenges of interdisciplinary research is the language barrier between fields

We need people who can translate across disciplinary jargon

Interdisciplinary and data science research requires **community building** first and foremost

community building ≠ putting people in the same room

What can we learn from this community?
What mistakes should we avoid making?