Machine Learning in the Multi-Messenger Era: Inference as a service and optimal light curve augmentation

Michael W. Coughlin
University of Minnesota

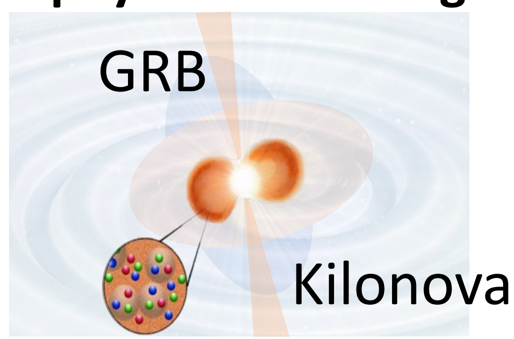
On behalf of many MMA folks
(A3D3, ZTF Machine Learning group)

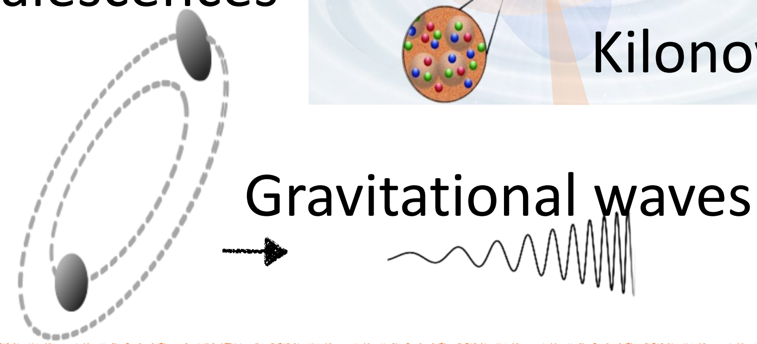
November 30, 2021

A Technical Ecosystem

Global astrophysical modeling

Compact Binary Coalescences





Dedicated instrumentation

Inference as a service

MMA Framework

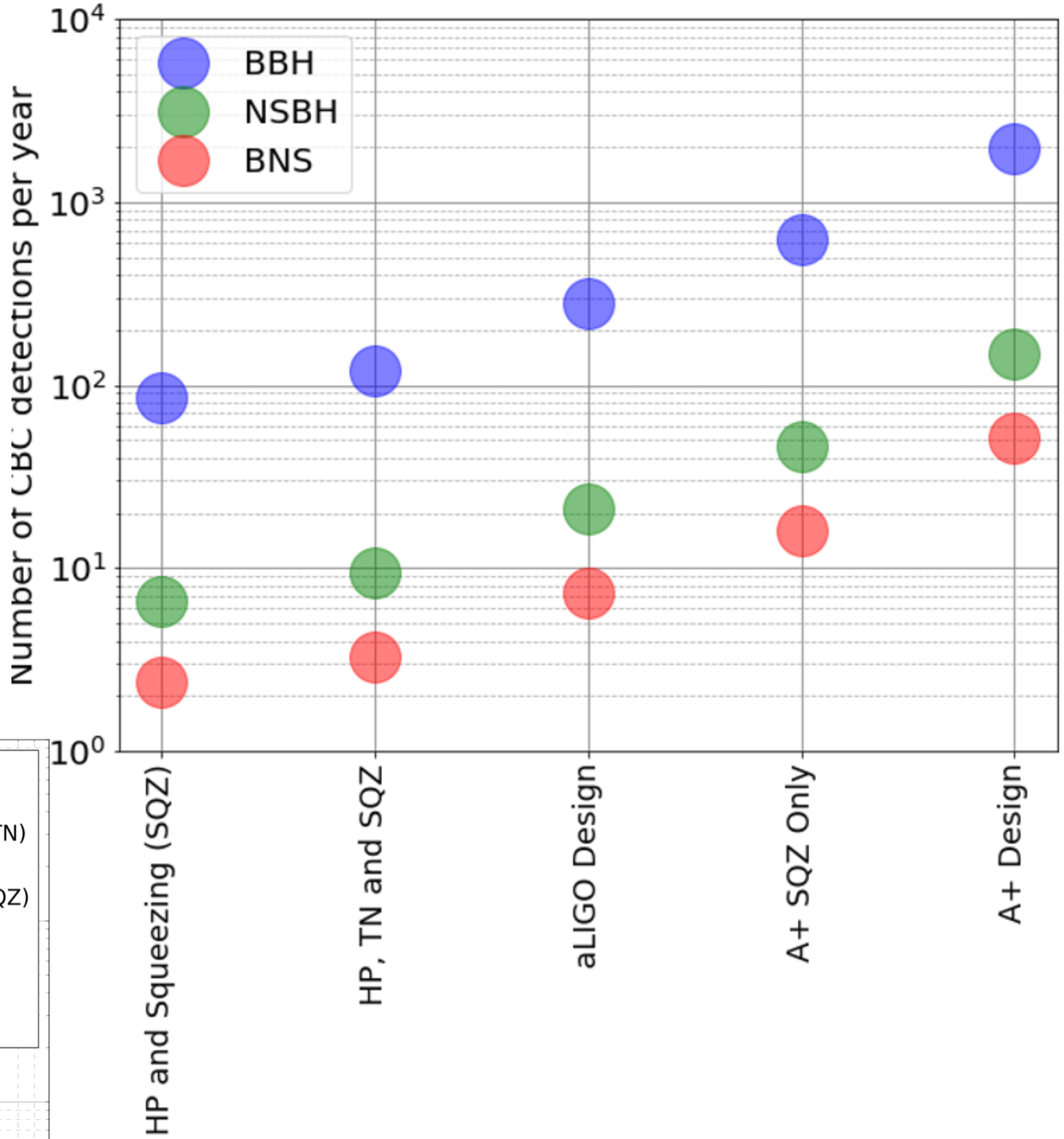
Machine Learning

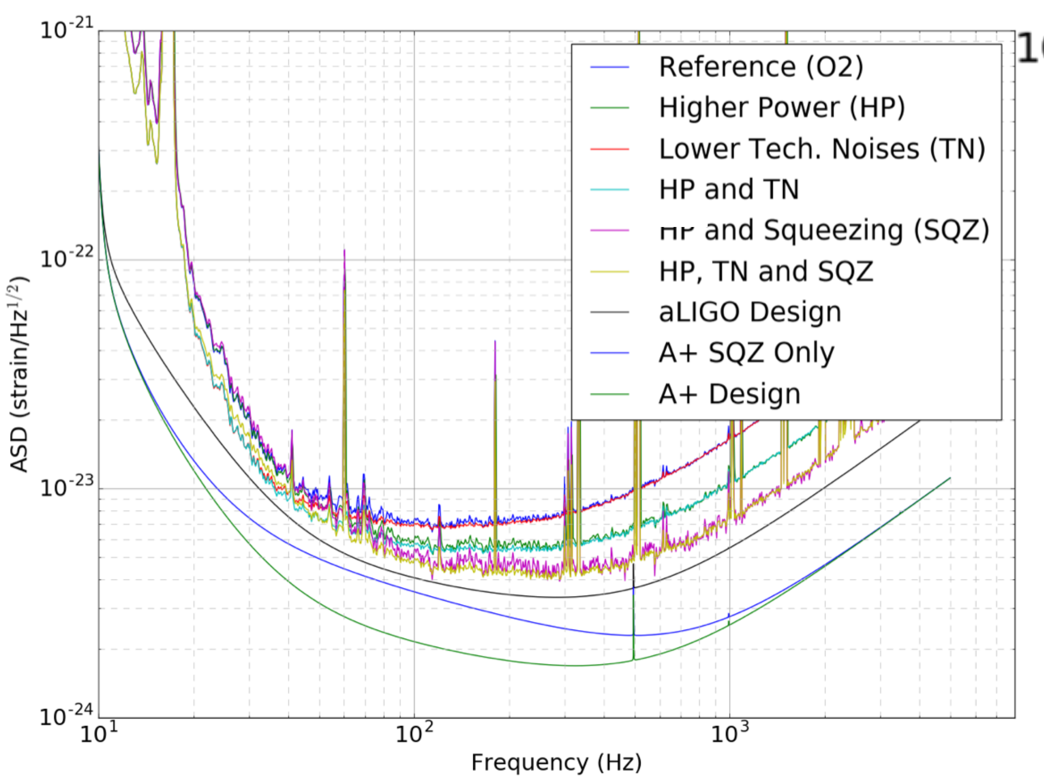
Science

Multi-Messenger science includes representation from many of the most interesting experiments today.

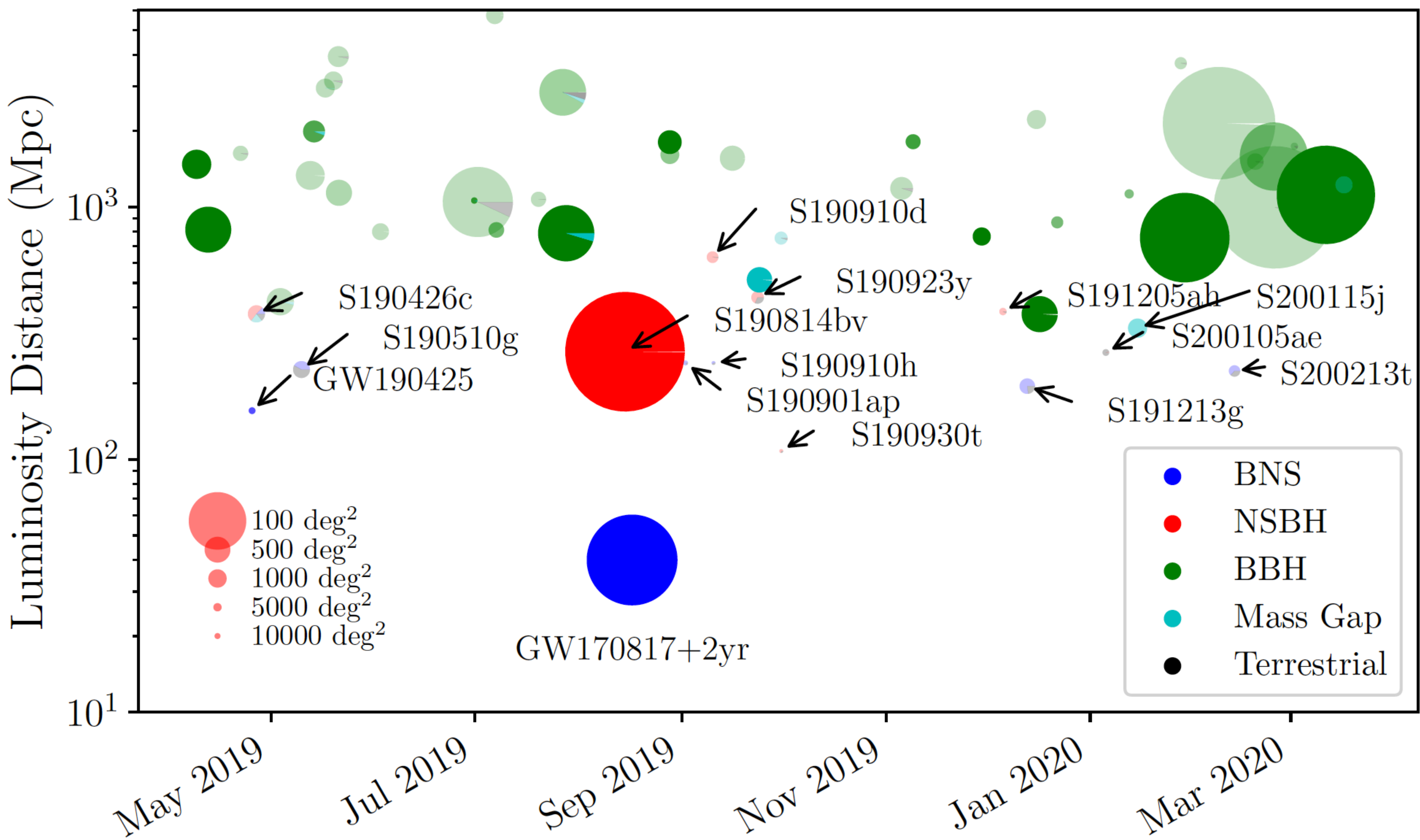
A New Era

Many detections are coming





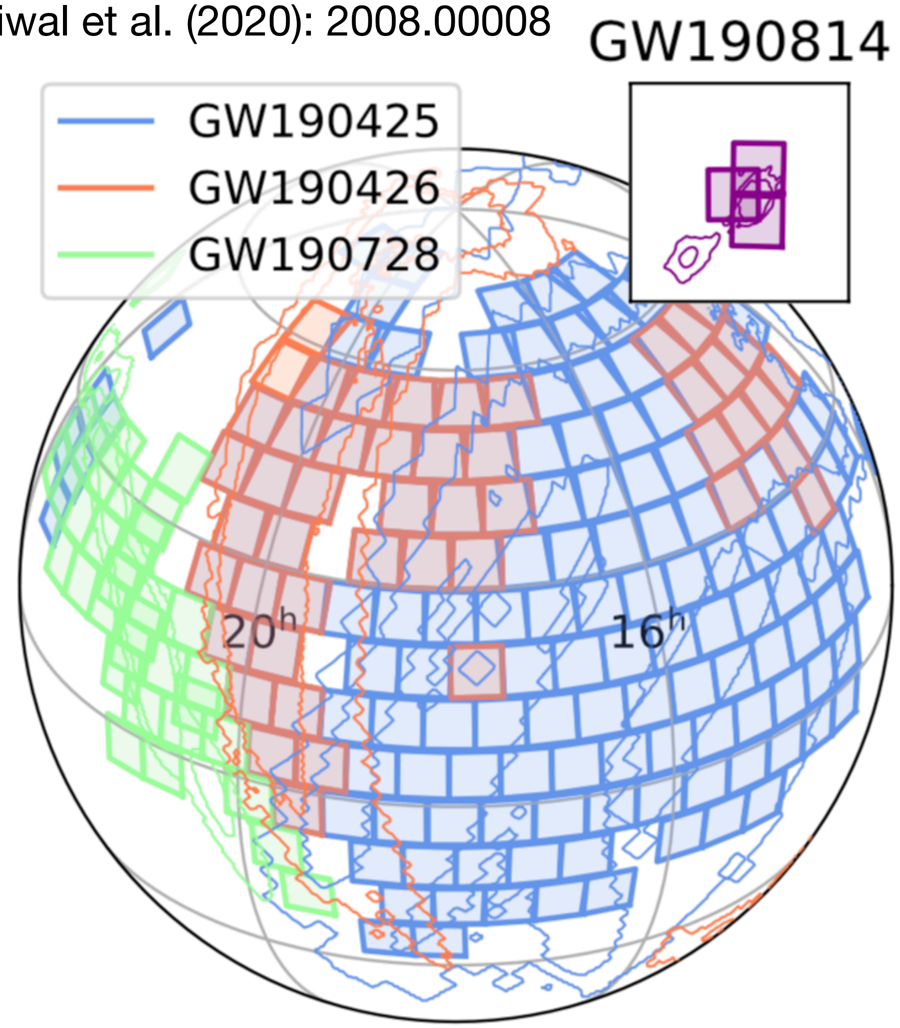
O3 Context

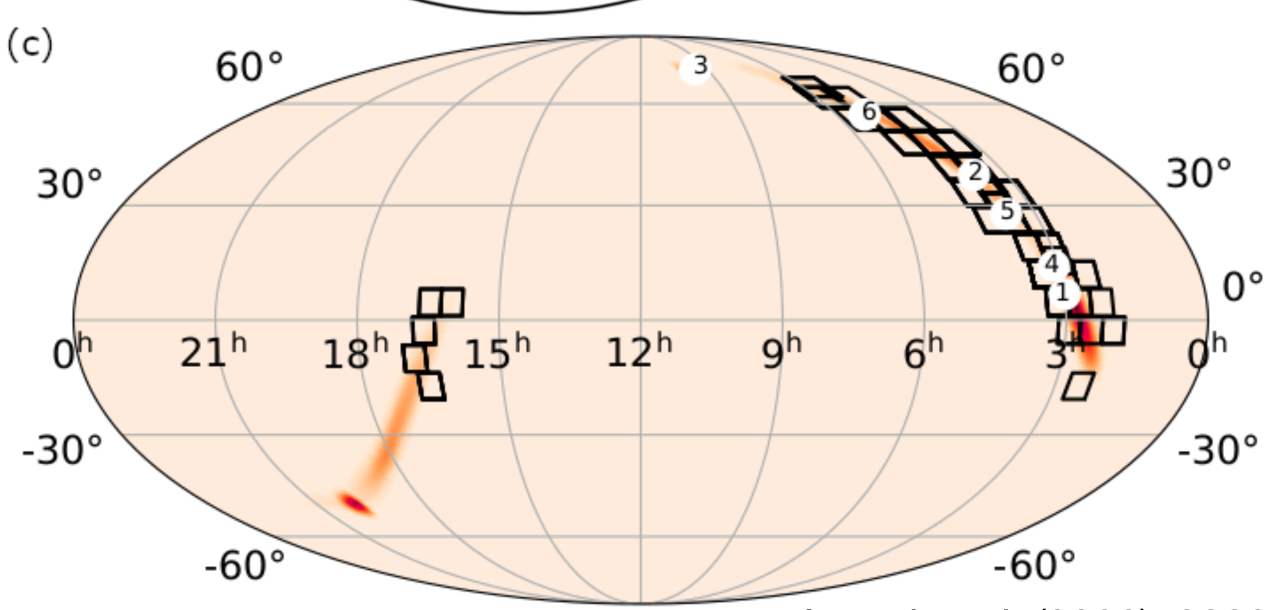


NOTE: Bigger circles (small sky areas) and very opaque (low false alarm rate) are easier to follow-up

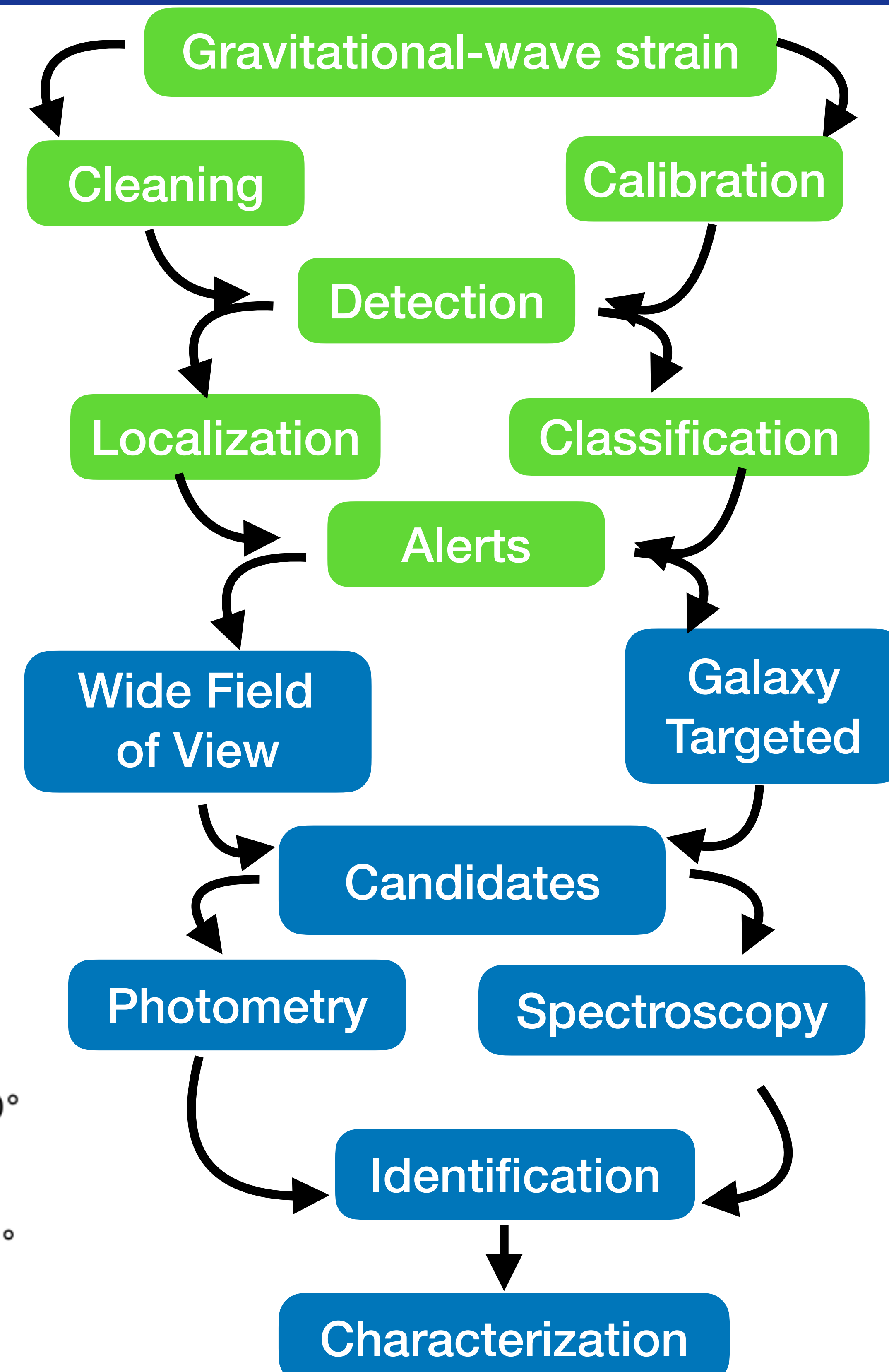
So What's the Problem?
Long Road from data to science

Kasliwal et al. (2020): 2008.00008





Anand et al. (2020): 2009.07210



Luckily, a lot of effort at this conference on problems of relevance!

Patrick Godwin (Pennsylvania State University)
Low-latency Noise Mitigation Techniques in Gravitational-wave Detector Data Using Auxiliary Sensor Information

Guillermo Valdes (Texas A&M University - College Station)
Virtual Talk: Acoustic noise in gravitational-wave detectors

Jenne Driggers (California Institute of Technology)
Improving the sensitivity of gravitational wave interferometers

Gabriele Vajente (California Institute of Technology)
Virtual Talk: Machine Learning and Gravitational Wave Detectors

Ik Siong Heng (University of Glasgow)
Gaussian Mixture Models for transient gravitational wave detection

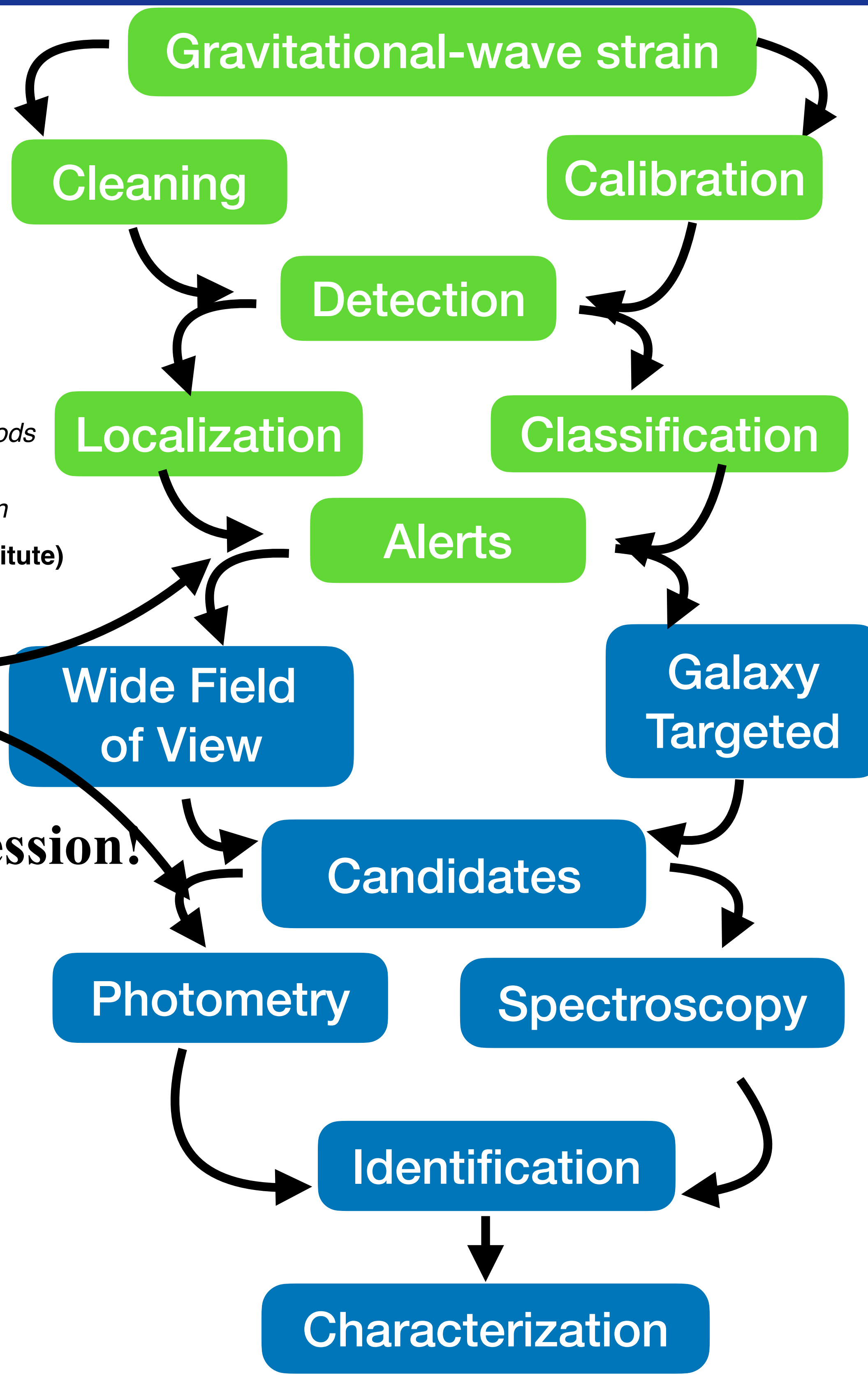
Agata Trovato (Università di Trieste)
Virtual Talk: Neural networks for gravitational-wave trigger selection in single-detector periods

John Veitch (University of Glasgow)
Virtual Talk: Computational Challenges in Gravitational Wave Parameter Estimation

Jonathan Gair (Max Planck Institute for Gravitational Physics, Albert Einstein Institute)
Virtual Talk: Rapid and robust parameter estimation for gravitational wave observations

Deep Chatterjee (University of Illinois at Urbana-Champaign)
Applications of machine learning in low-latency electromagnetic counterpart inference from gravitational waves

NOTE: Not an exhaustive copy and paste session!



Luckily, a lot of effort at this conference on problems of relevance!

Patrick Godwin (Pennsylvania State University)
Low-latency Noise Mitigation Techniques in Gravitational-wave Detector Data Using Auxiliary Sensor Information

Guillermo Valdes (Texas A&M University - College Station)
Virtual Talk: Acoustic noise in gravitational-wave detectors

Jenne Driggers (California Institute of Technology)
Improving the sensitivity of gravitational wave interferometers

Gabriele Vajente (California Institute of Technology)
Virtual Talk: Machine Learning and Gravitational Wave Detectors

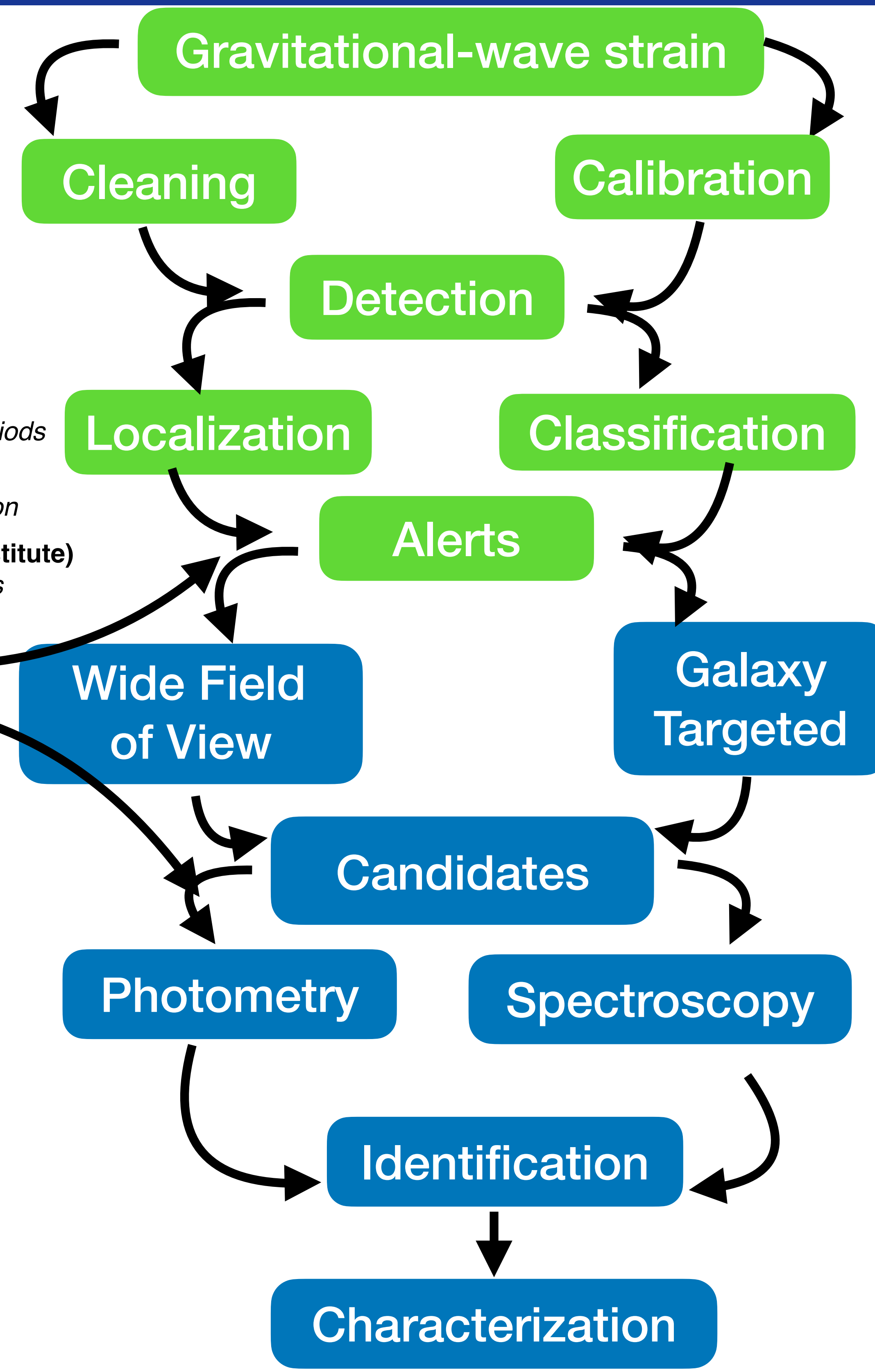
Ik Siong Heng (University of Glasgow)
Gaussian Mixture Models for transient gravitational wave detection

Agata Trovato (Università di Trieste)
Virtual Talk: Neural networks for gravitational-wave trigger selection in single-detector periods

John Veitch (University of Glasgow)
Virtual Talk: Computational Challenges in Gravitational Wave Parameter Estimation

Jonathan Gair (Max Planck Institute for Gravitational Physics, Albert Einstein Institute)
Virtual Talk: Rapid and robust parameter estimation for gravitational wave observations

Deep Chatterjee (University of Illinois at Urbana-Champaign)
Applications of machine learning in low-latency electromagnetic counterpart inference from gravitational waves



However, two questions based on this?

(i) Are we ready to apply these ML algorithms in real time?

(ii) How can we support the blue block better?

Are we ready to apply ML in real time?

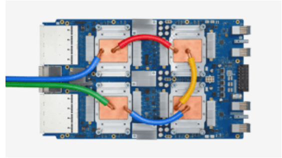
Deep Learning Programs at Inference Time

Pros

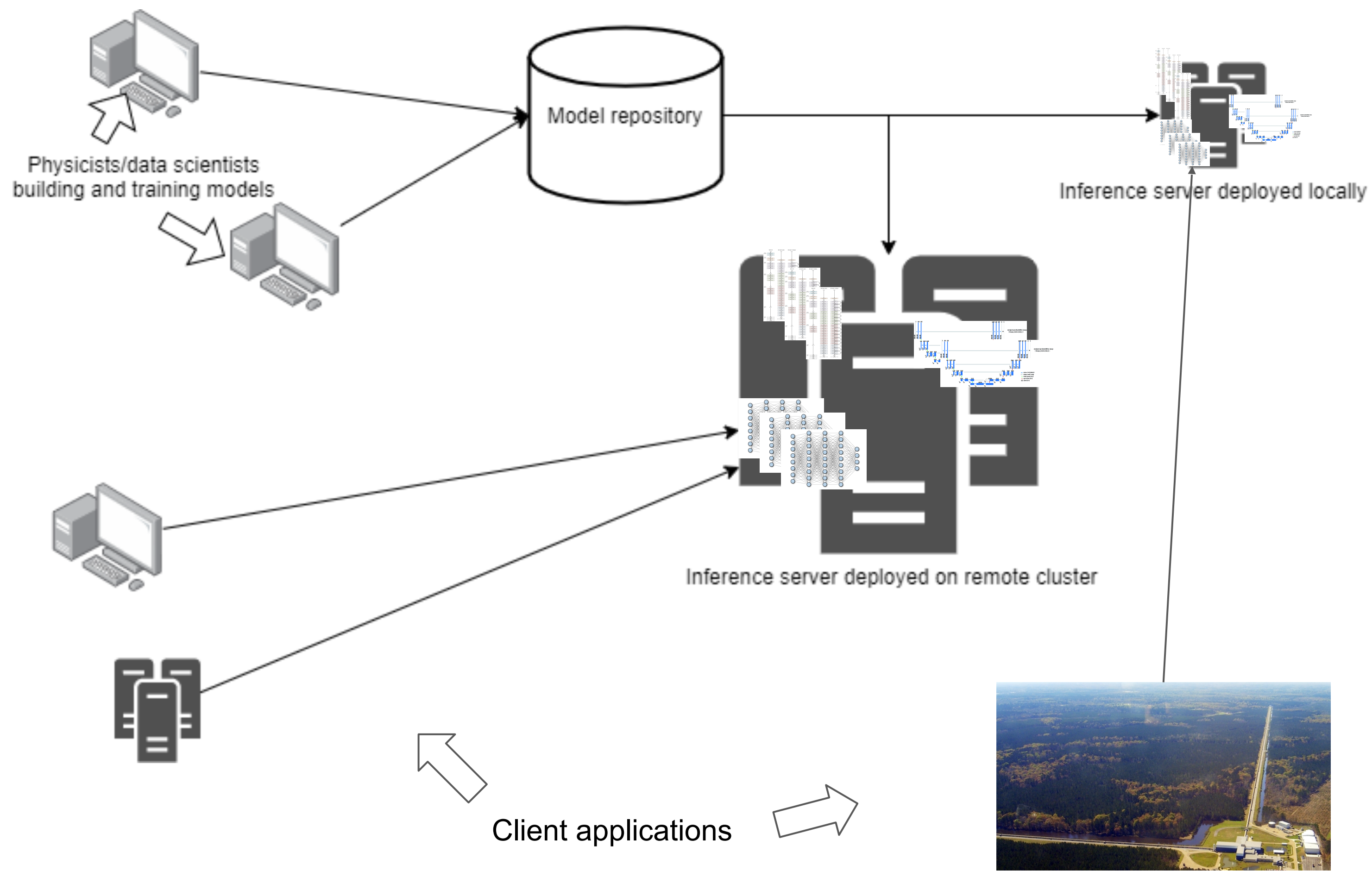
- Robust modelling capabilities
- Real-time compatible

Cons

- Resource intensive
- May require frequent updates
- Effective use requires specialized knowledge, software, and often hardware



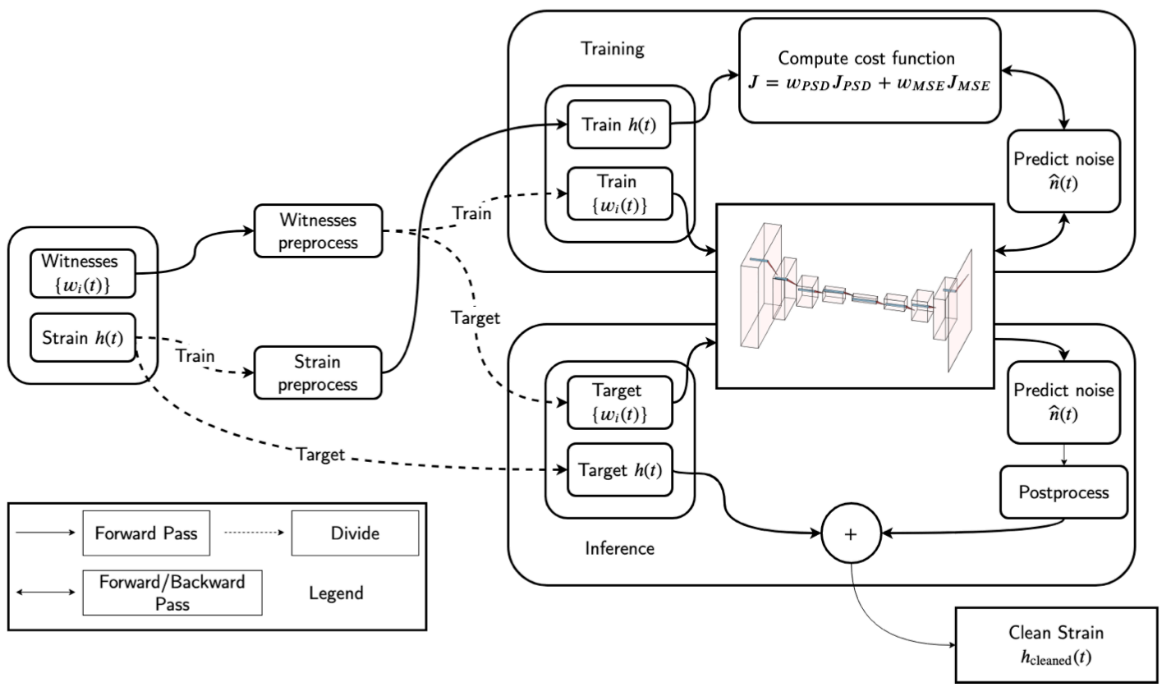
Inference as a service

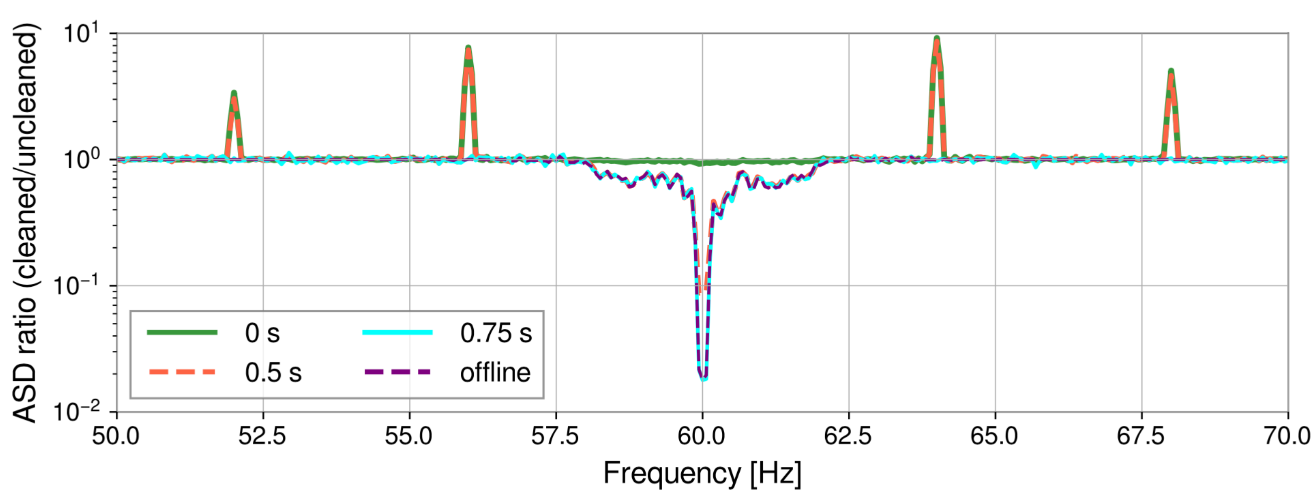


- Application for hosting trained networks and exposing them for inference via standardized client APIs
- Abstracts away details about models and their implementations
- Effectively leverages concurrent execution on heterogeneous computing resources
- Containerizations means portability and easy scale
- Centralized model repositories keep all users in sync

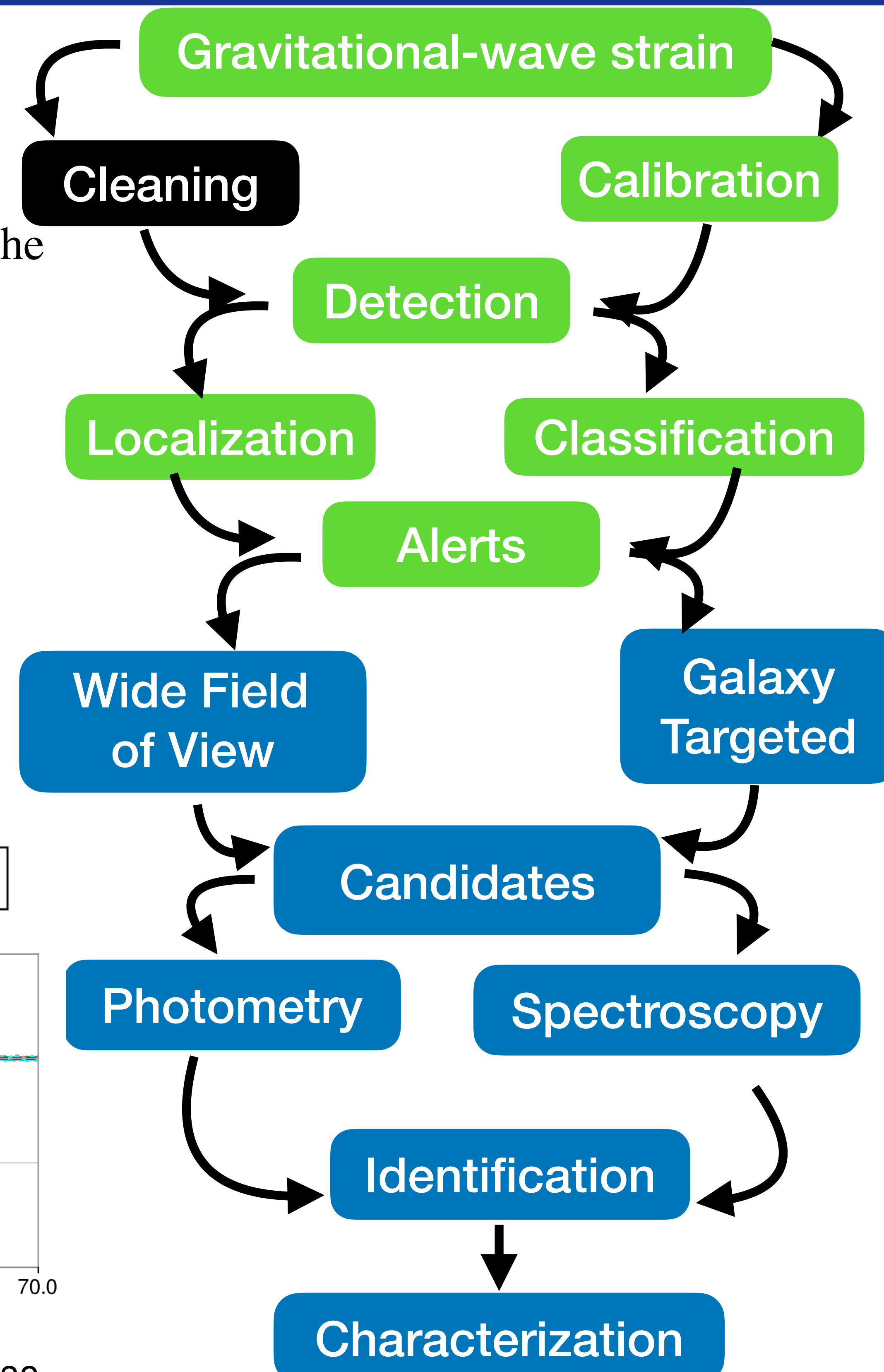
Cleaning gravitational-wave data

- Technical and environmental noises prevent GW detectors from operating at its design sensitivity
- Can use ML methods combined with on-site witness sensors to predict the detector response and remove the noise



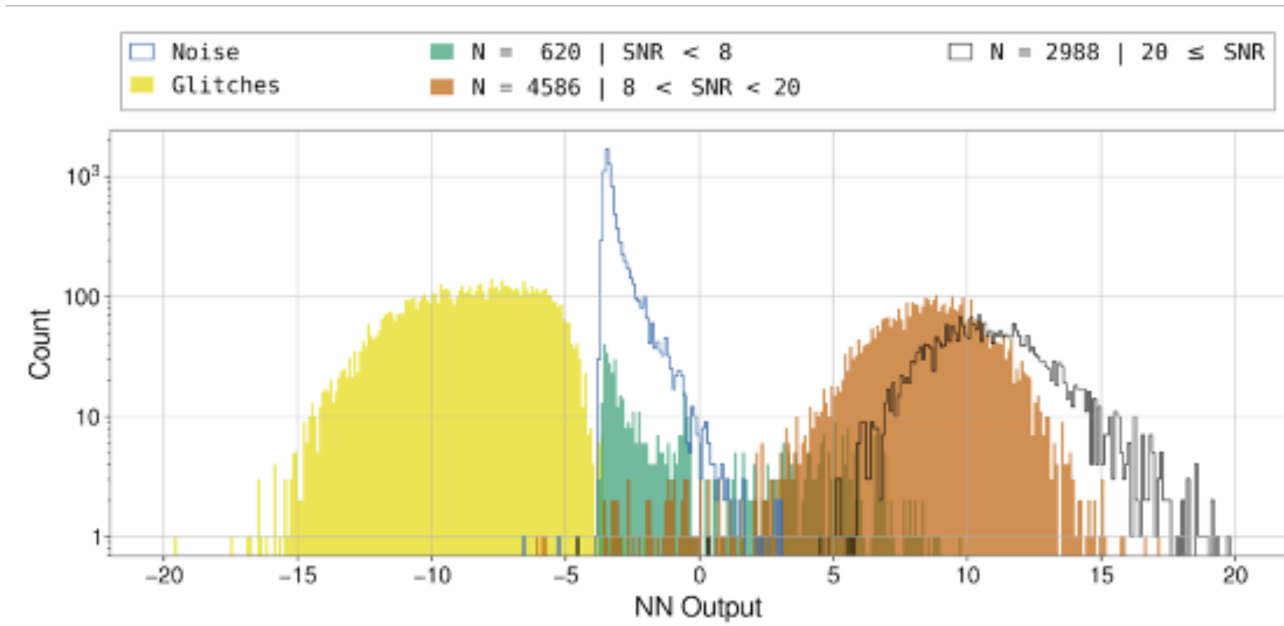


Ormiston et al. (2020): 2005.06534, Gunny et al. (2021): 2108.12430

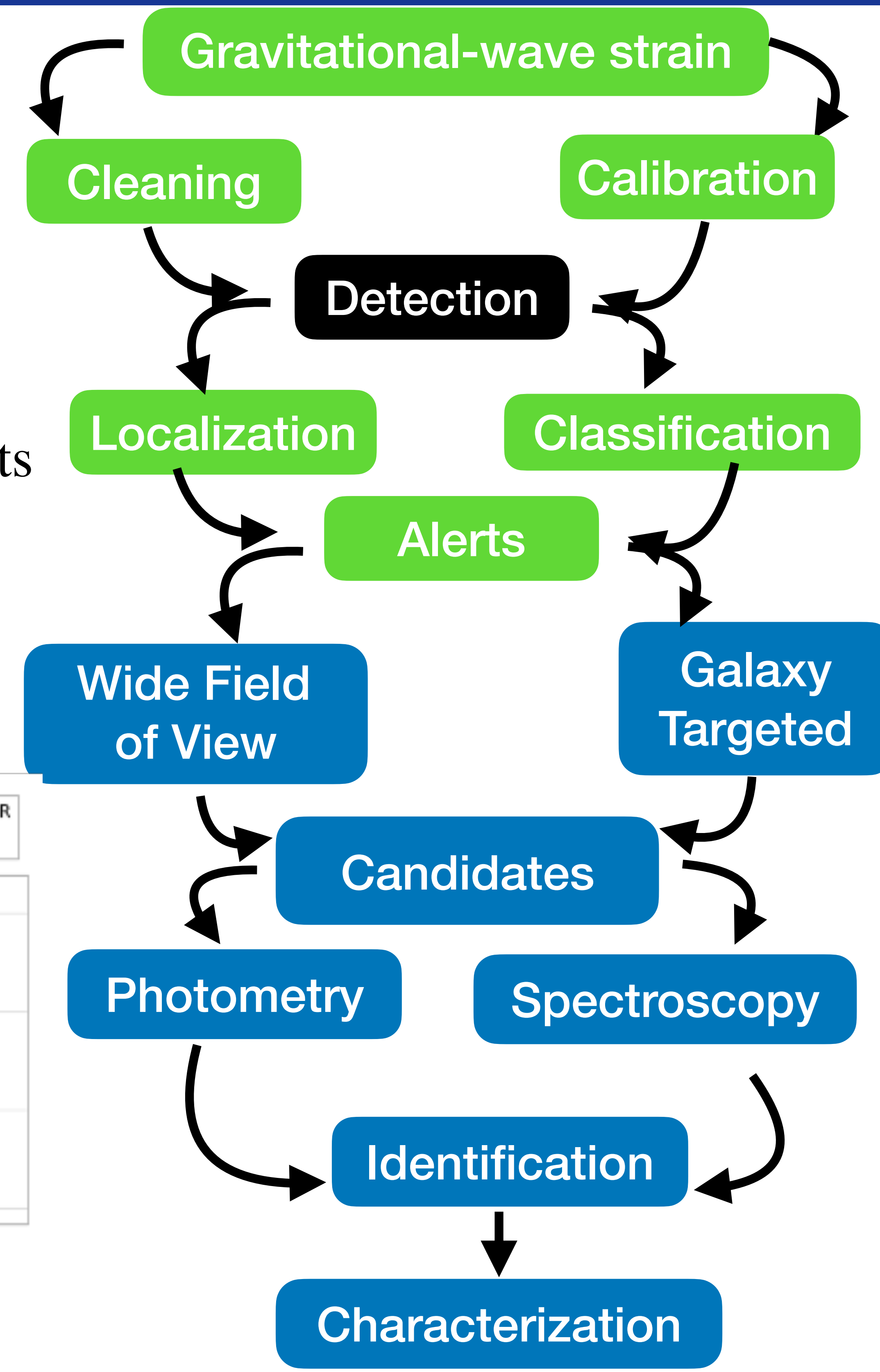


Detecting gravitational waves

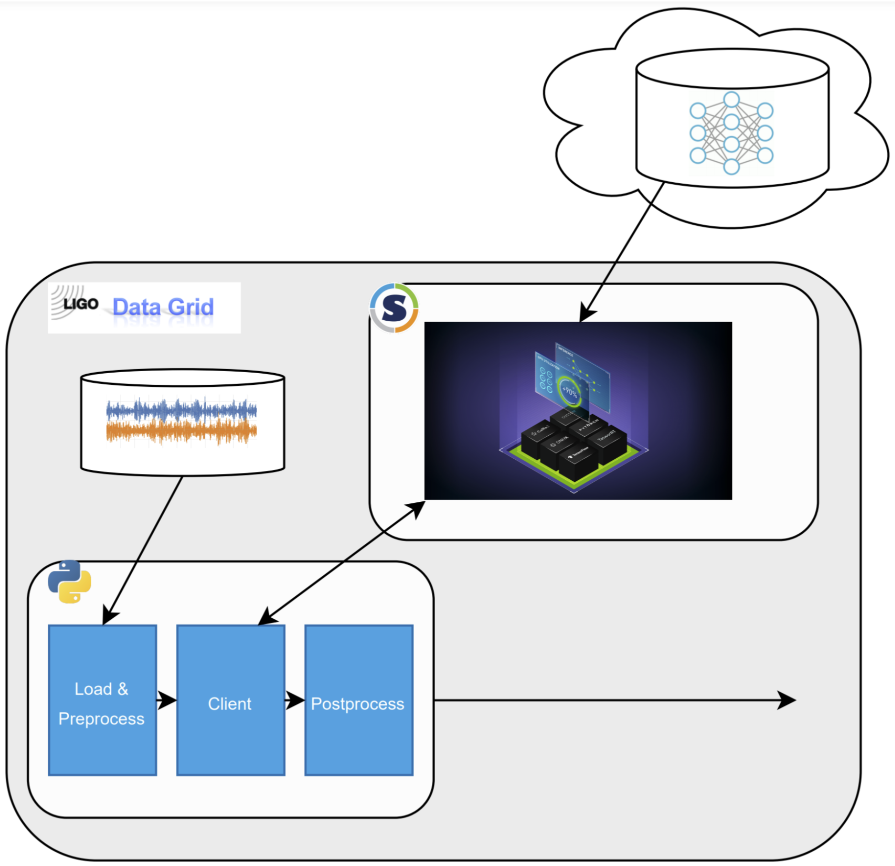
- Gravitational-wave detection is basically a solved problem for Gaussian noise, which gravitational-wave data is not
- ML methods have been shown to have the capability of meeting the speed requirements for online searches, while also being more robust to data transients
- Only BBHs (short signals) so far



Gunny et al. (2021): 2108.12430

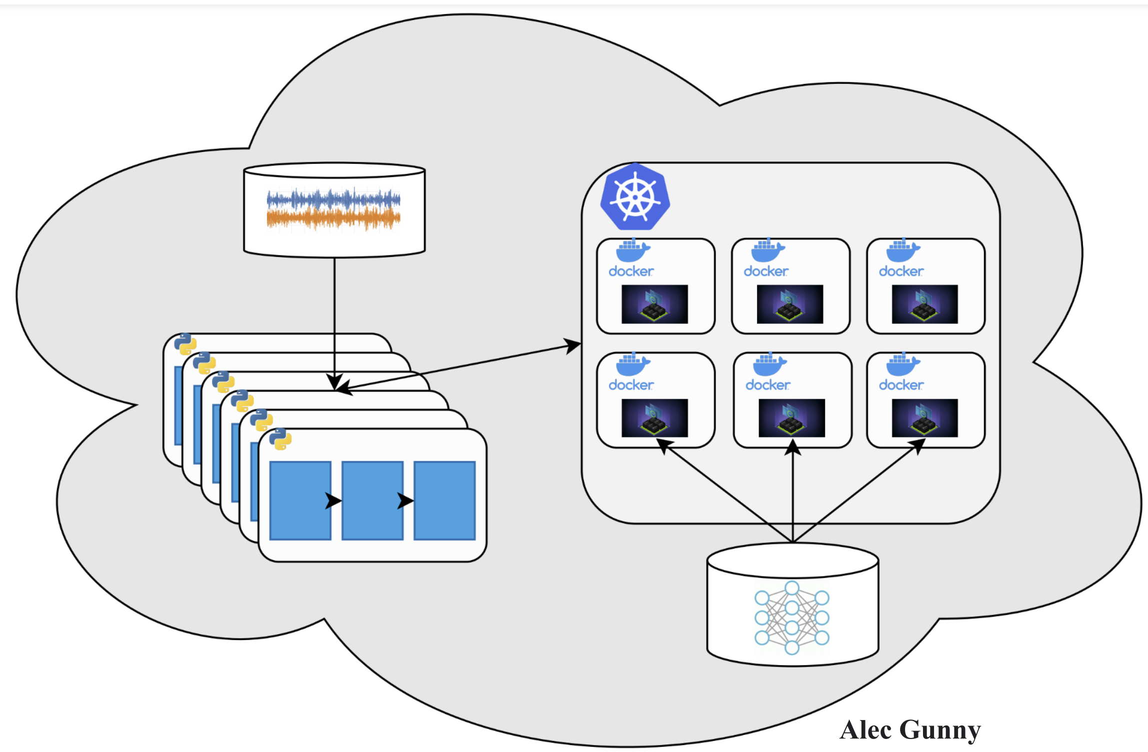


LVK Inference as a Service: Deployment Scenarios



Offline

- Maximize throughput to minimize time to completion (subject to cost constraints)
- Cloud resources leverage economies of scale
- Cleaning one month of O3 data with DeepClean
- End-to-end ensemble with DeepClean and BBHnet to estimate event likelihood over ~27 hrs. of O2 data



Online

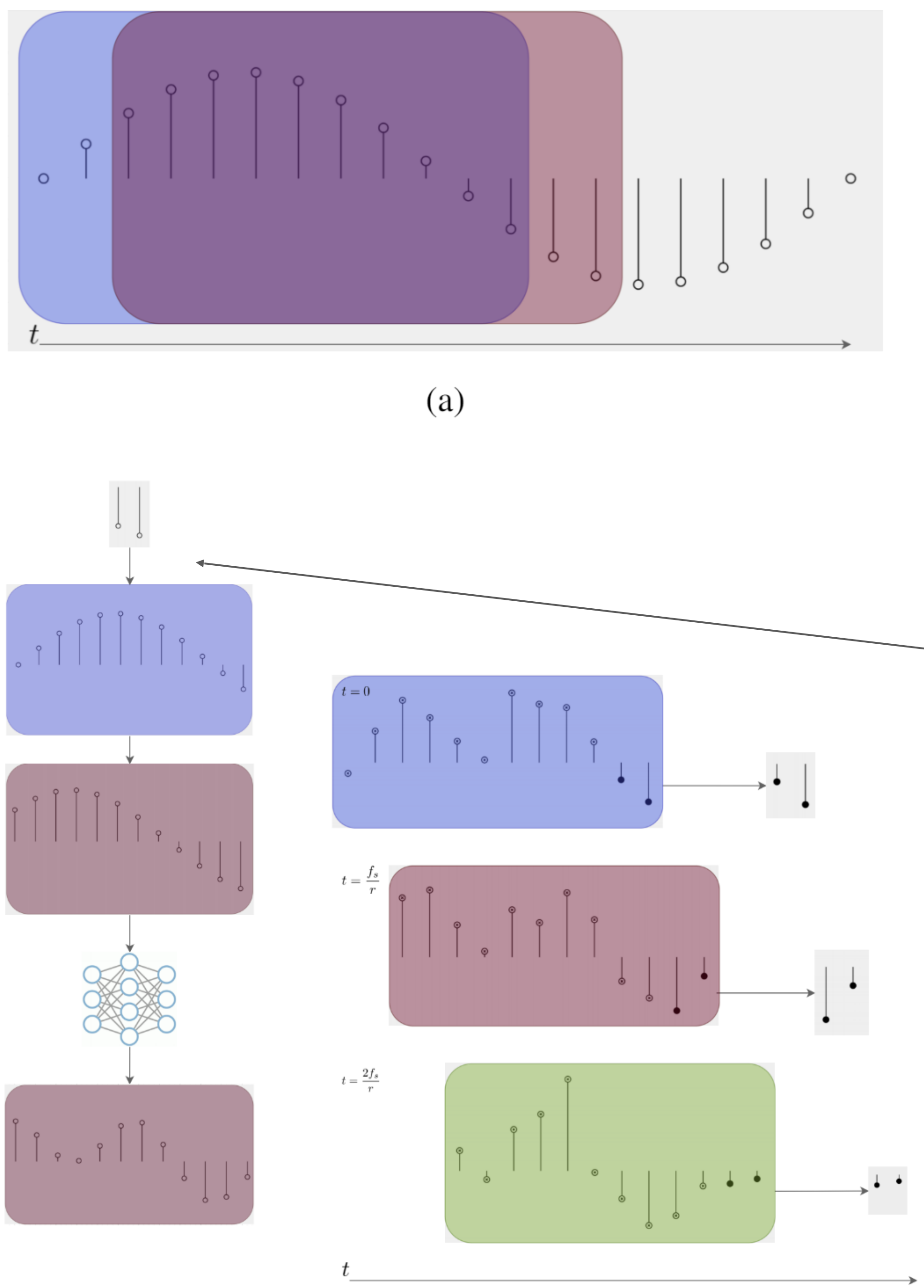
- Latency sensitive, deploy locally to minimize data transfer time
- Using DeepClean to remove noise in real-time and make cleaned strain available to downstream analyses/ searches, including BBHNet

Alec Gunny

MIT

Gunny et al. 2021: 2108.12430

Inference as a Service: for Streaming Data



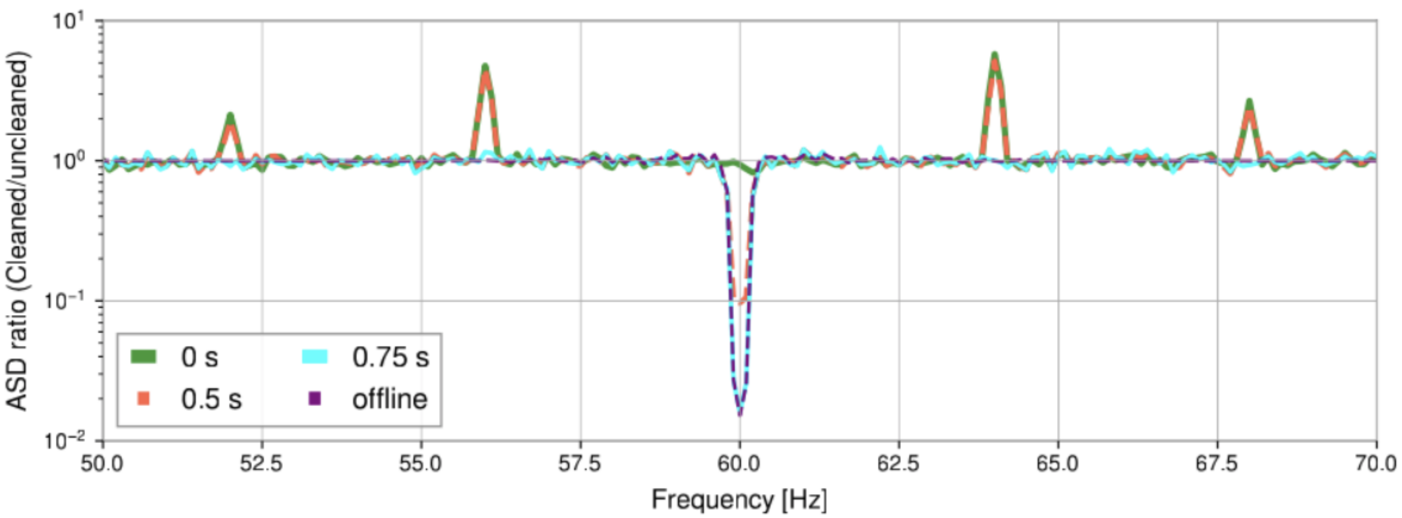
(a) (b) (c)

- DeepClean and BBHnet perform inference on fixed-length snapshots of time series
  - Rate at which snapshots are sampled fixed by an inference-time parameter *r*, the inference sampling rate
- High values of *r* compared to the length of the frame lead to substantial data overlap and redundant data transfer from client to server
- Host a "snapshotter" model on the server that maintains the current snapshot as a state
  - Only stream state updates
  - Updated snapshot gets passed to downstream models
  - Introduces a potential sequential bottleneck
- DeepClean also has overlapping output data
  - Aggregating between overlap incurs extra latency
  - Currently adopting "fully-online" solution

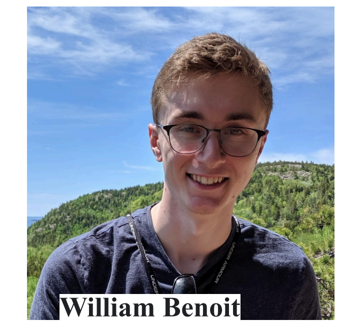
Gunny et al. 2021: 2108.12430

Performance vs aggregation latency

Performance

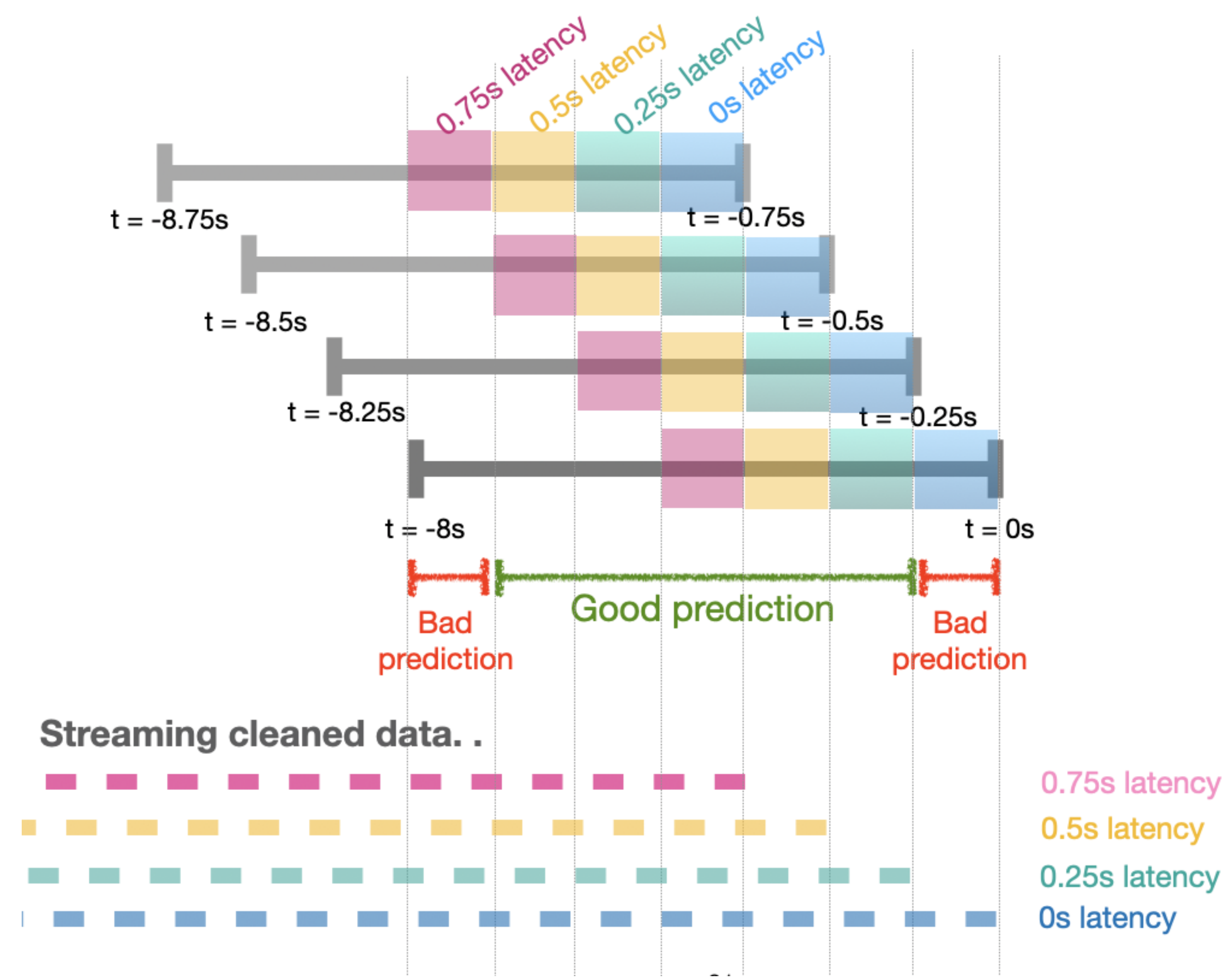


UMN



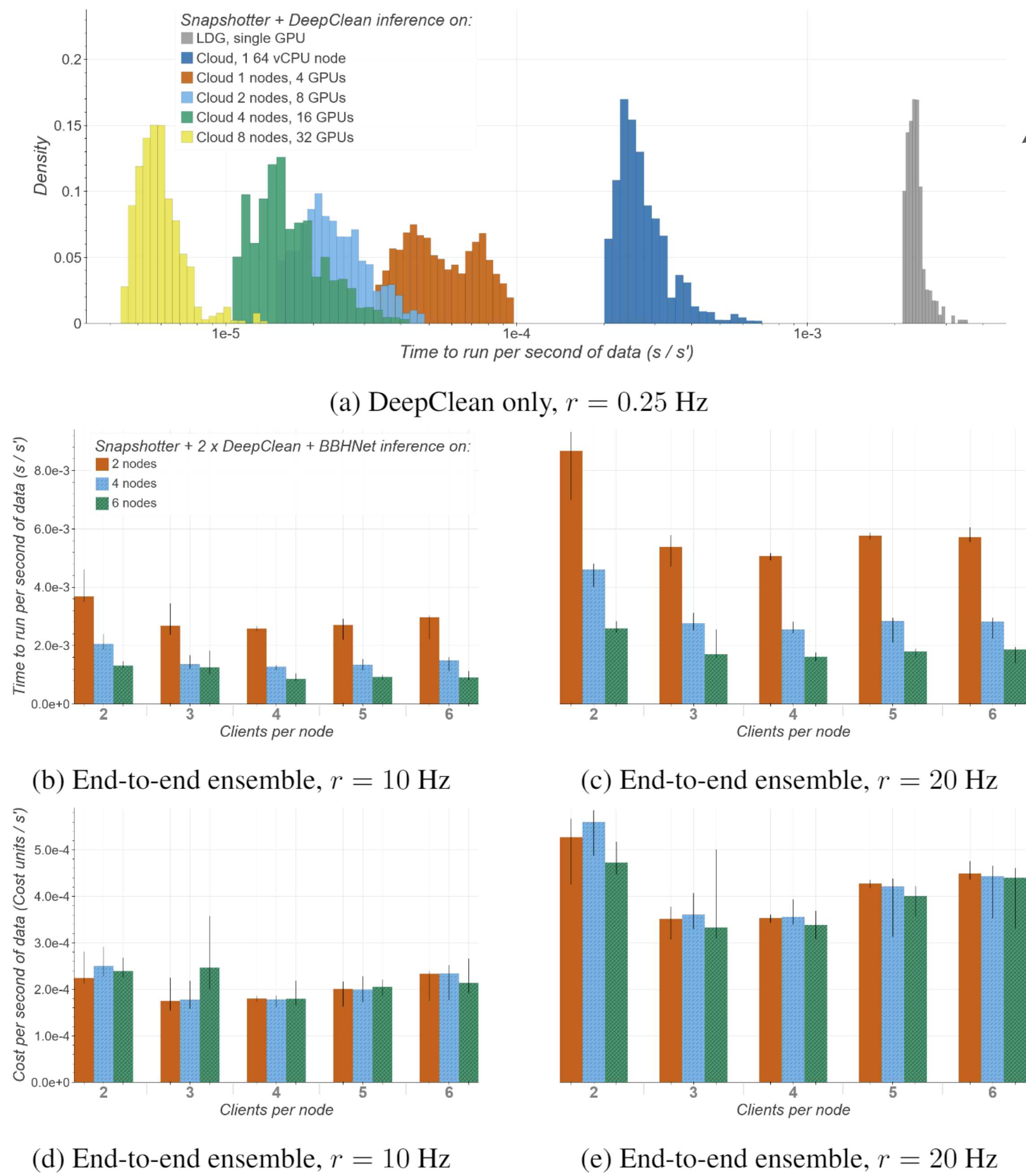
UMN

aggregation latency - the amount of data (seconds) to be excluded from the end of the segment due to quality degradation.



Gunny et al. 2021: 2108.12430

Offline Use Cases



(a) DeepClean only, $r = 0.25$ Hz

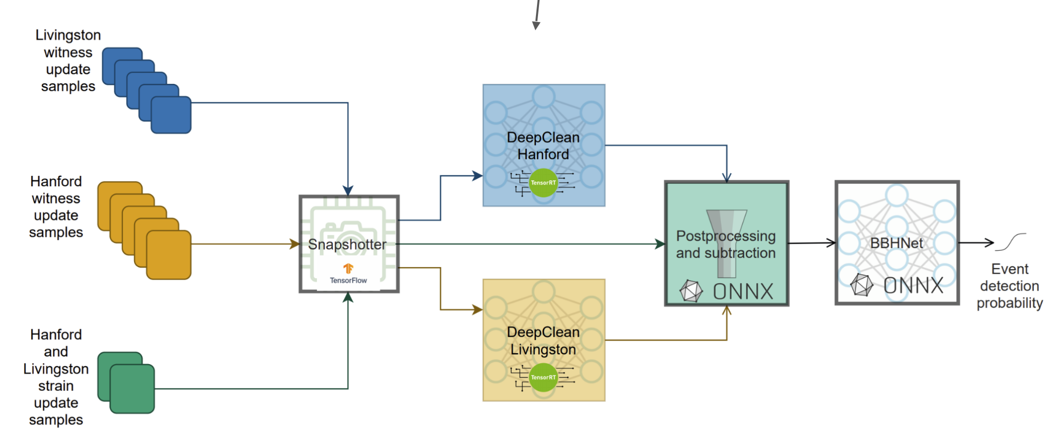
(b) End-to-end ensemble, $r = 10$ Hz

(c) End-to-end ensemble, $r = 20$ Hz

(d) End-to-end ensemble, $r = 10$ Hz

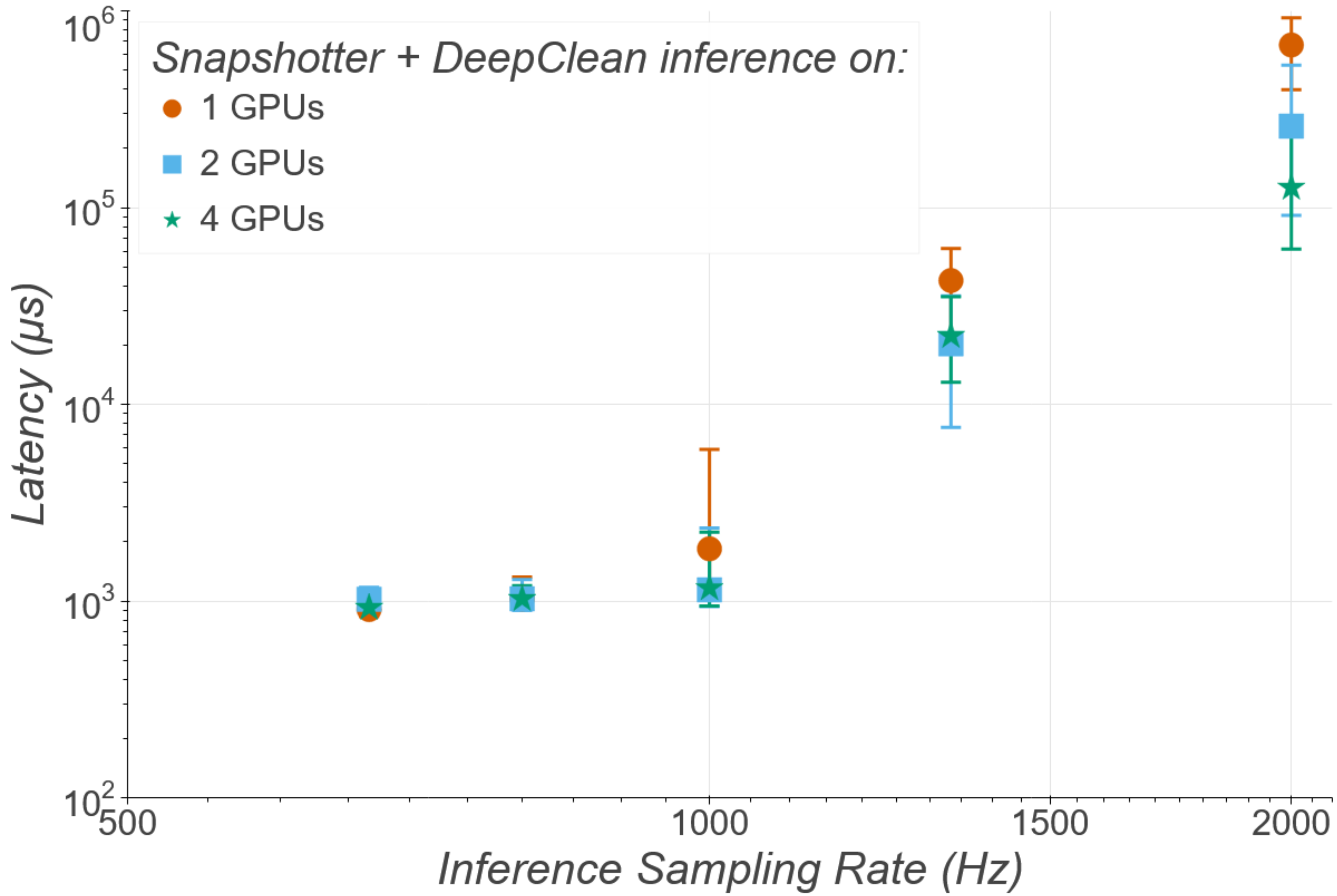
(e) End-to-end ensemble, $r = 20$ Hz

- Offline DeepClean shows the advantages of switching to IaaS model: CPU-only node with ~10x reduction in processing time, adding GPUs gives another ~5x
- Ensemble model shows economies of scale of IaaS paradigm
- Processing time decreases linearly with # of nodes, cost stays constant
  - Optimal point at “infinite” scale
  - Scale achieved with minimal additional engineering overhead
- Ensemble leverages multiple framework backends, all invisible to client users



Gunny et al. 2021: 2108.12430

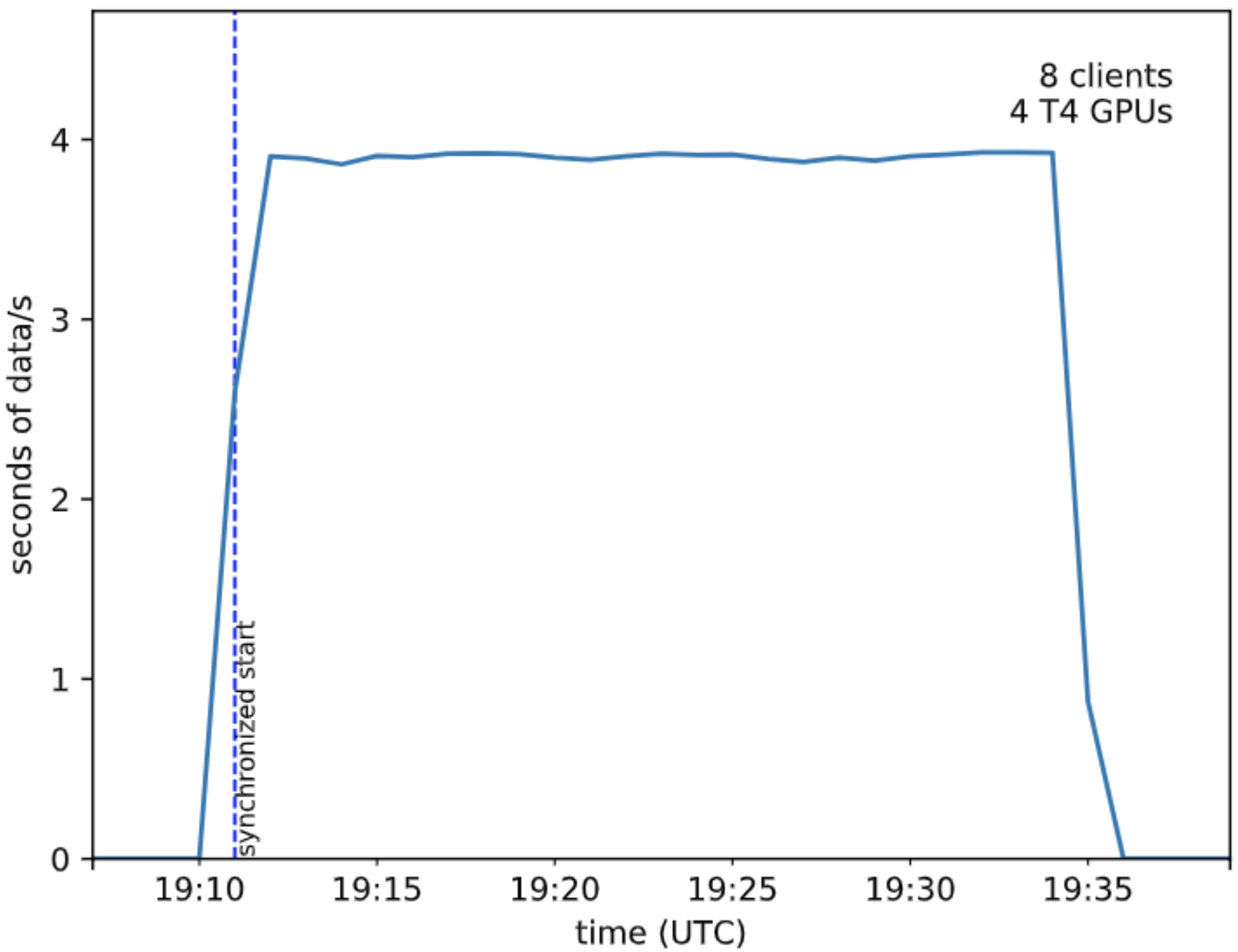
Online Use Cases



- Concurrent execution of IaaS model important at lower inference rates, keeps demand for scale low
- Scale becomes more important at higher frequency inference rates
- Bottleneck is currently sequential update to the input “state”
- Optimizing this step via HPC unlocks more advantage from additional GPUs

Tests at Scale

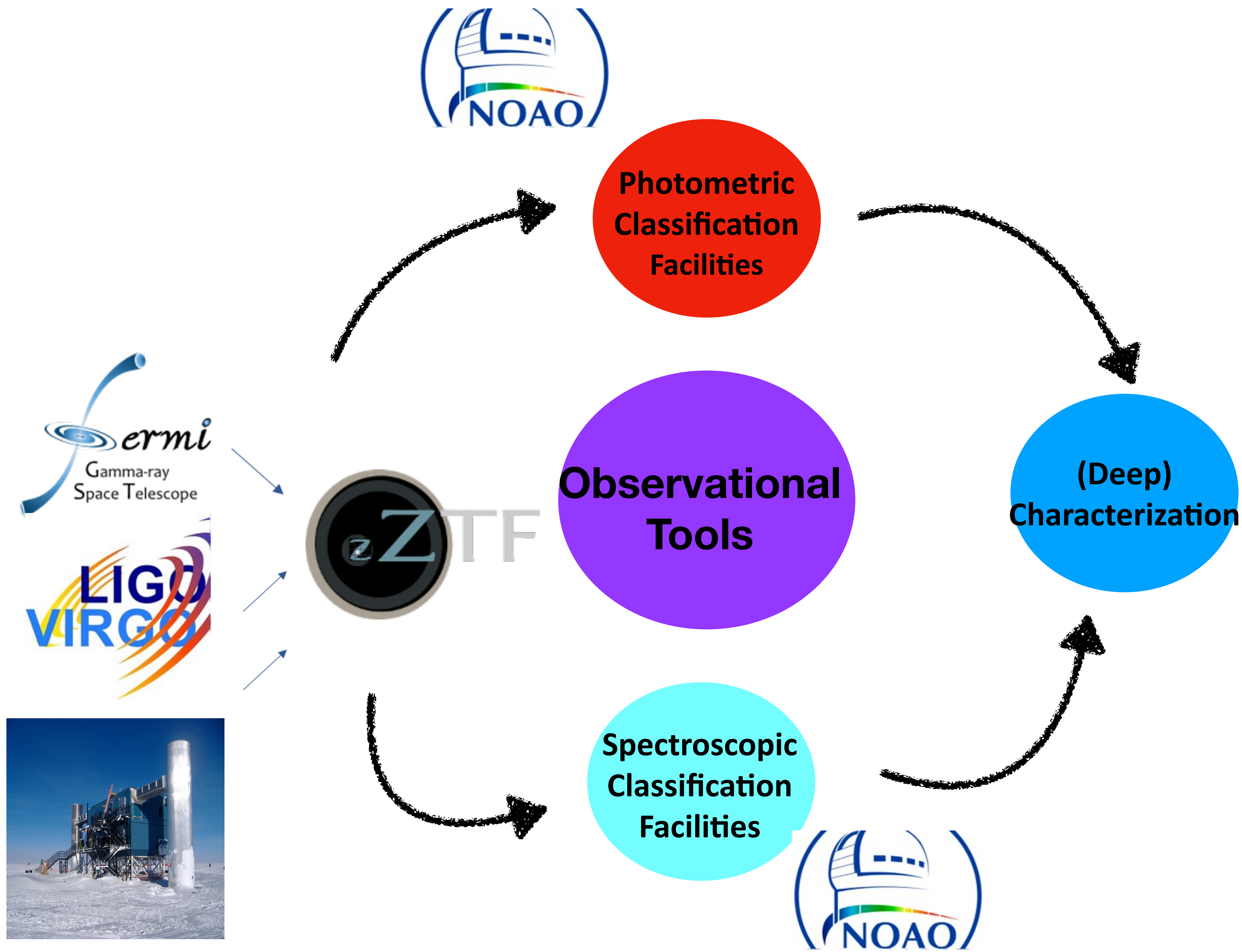
- Use HEPCloud framework to run larger tests with multiple clients/servers
  - Tests use cloud resources (GCP), submitted through HTCondor
- Jobs synchronized to start all at once, mimic realistic environment
- Able to sustain processing for full length of job
- Provides a means to manage/run large amount of jobs



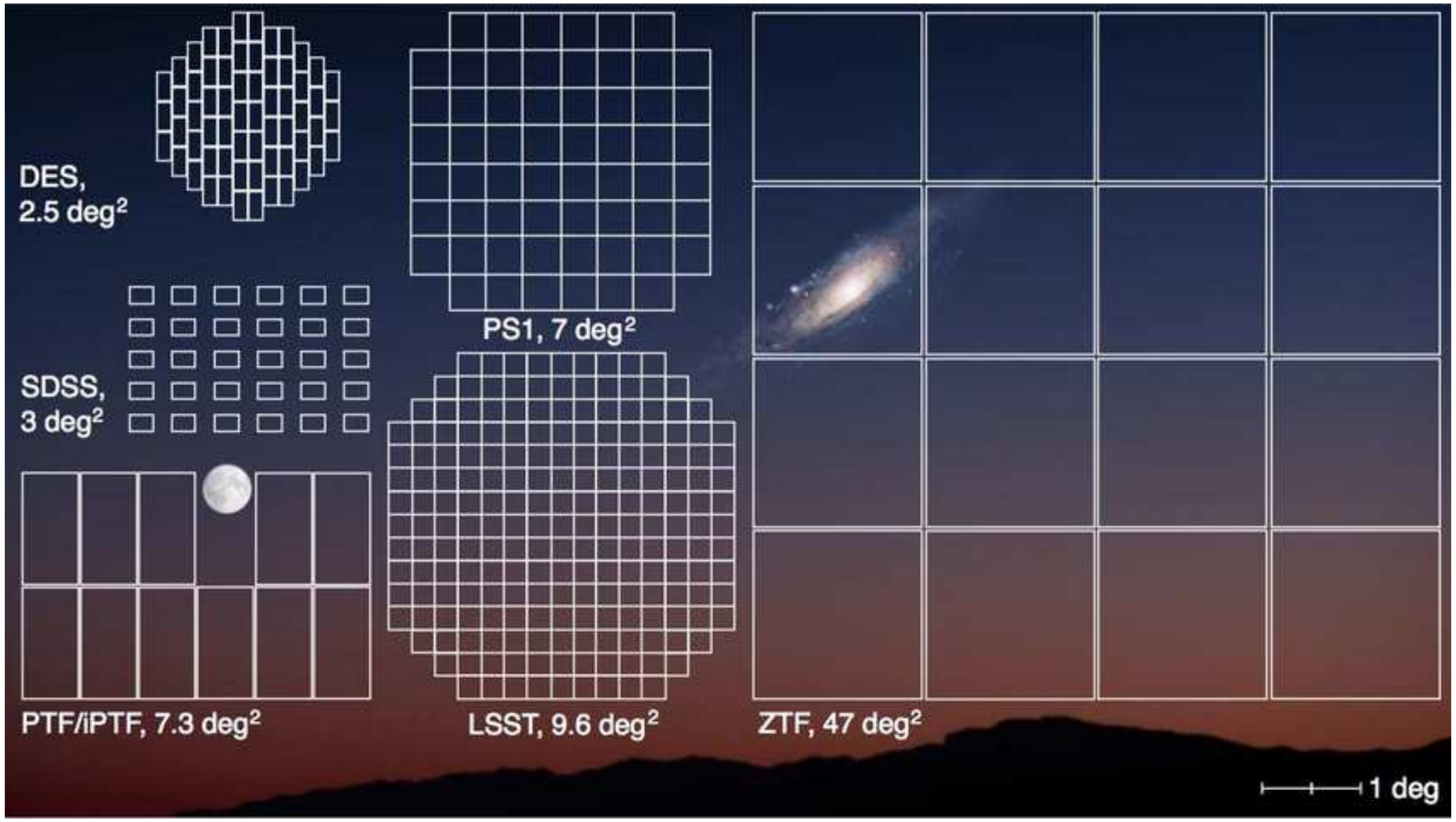
Gunny et al. 2021: 2108.12430

How can we support and perform better observations?

The Observational Landscape

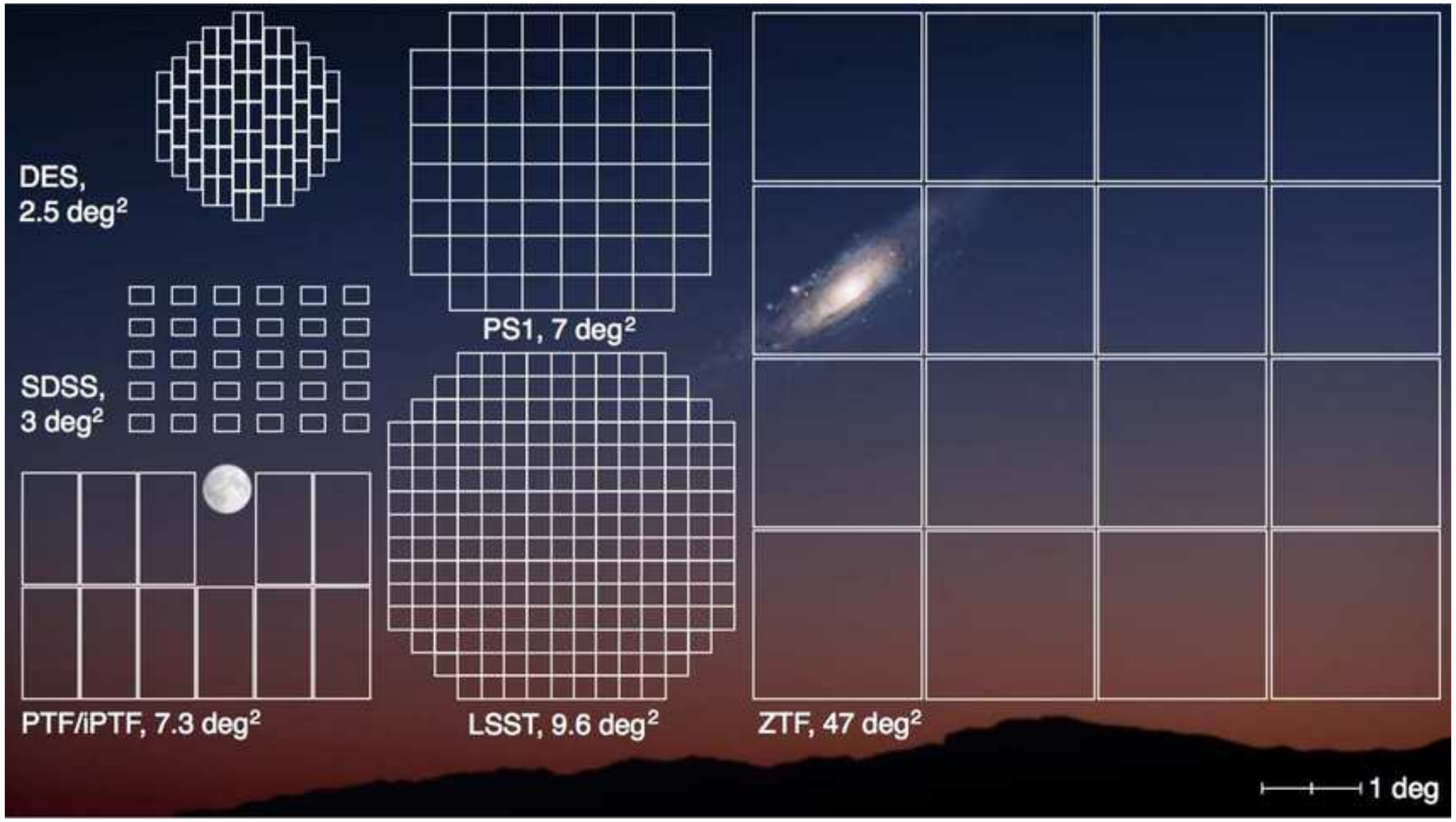


Landscape of Optical TDA



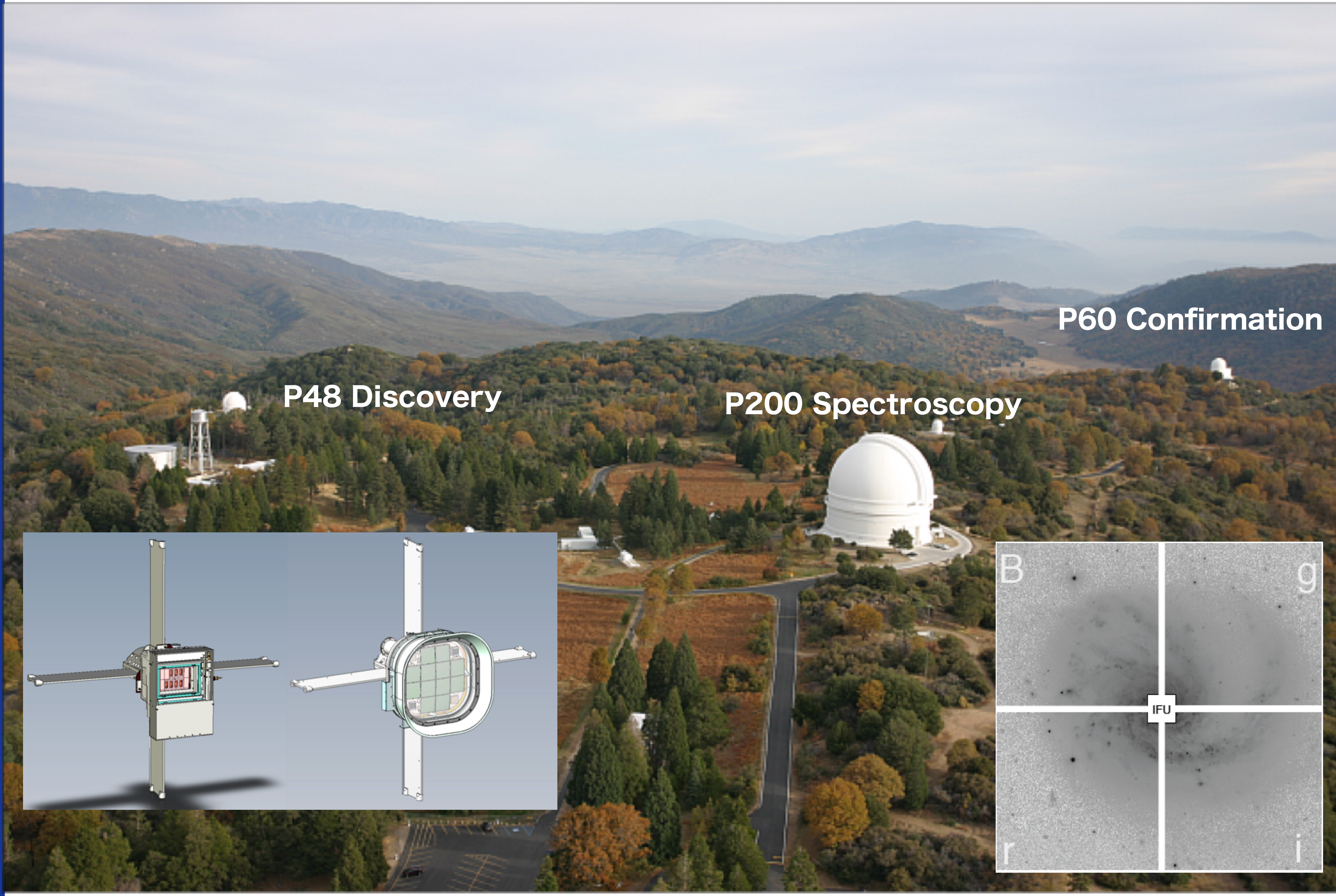
- The night sky is imaged at 17.5 mag by ASAS-SN (both hemispheres)
- The northern sky is covered by ATLAS, ZTF, and PS-1 to 19, 20.5, 21.5 over roughly two nights (ZTF issues real time, data-rich alerts)
- BlackGEM (21-22 mag; Chile) will start routine operation within this year
- Rubin is expected to become operational in 3 to 4 years

Two Different Approaches

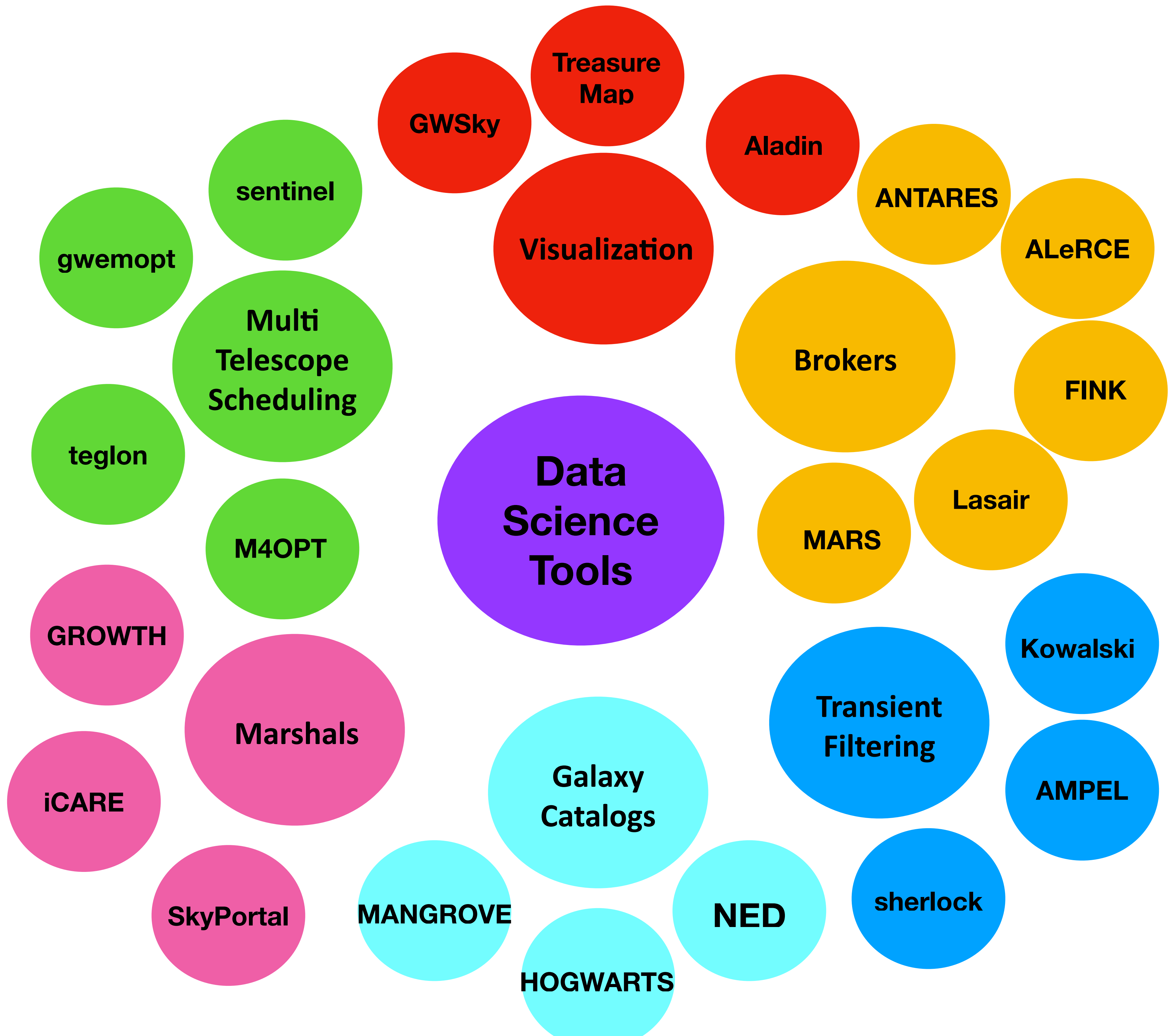


- Photometric detection followed by spectroscopic classification
– Possible for surveys which are shallow (20 mag or so)
– Spectral classification can be undertaken by existing telescopes
- Photometric detection followed by multi-band time series
– Large samples of faint objects –Much of the analysis will be statistical –Use clever techniques and filter out a small subset for further follow up

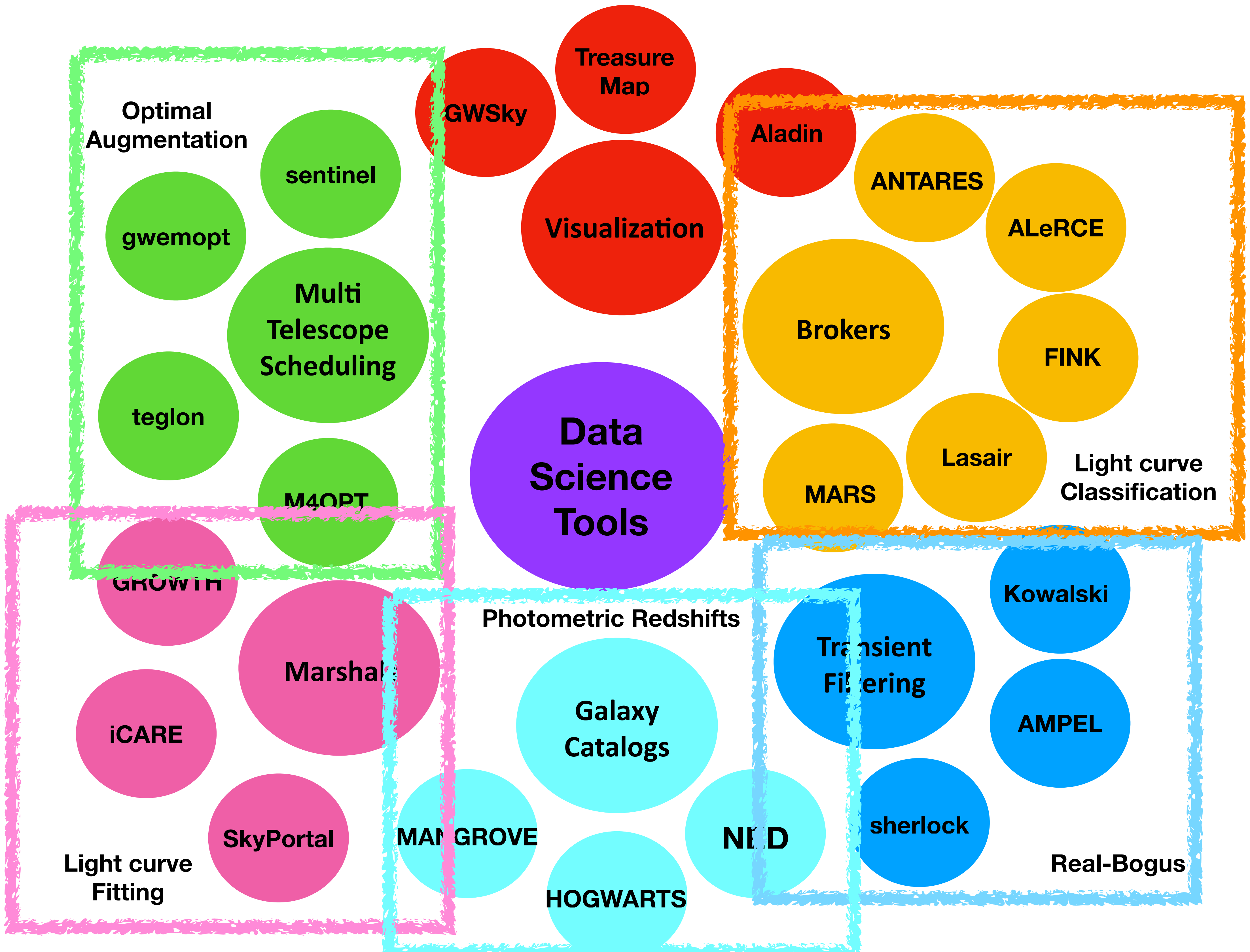
Palomar Telescopes



The Technical Landscape

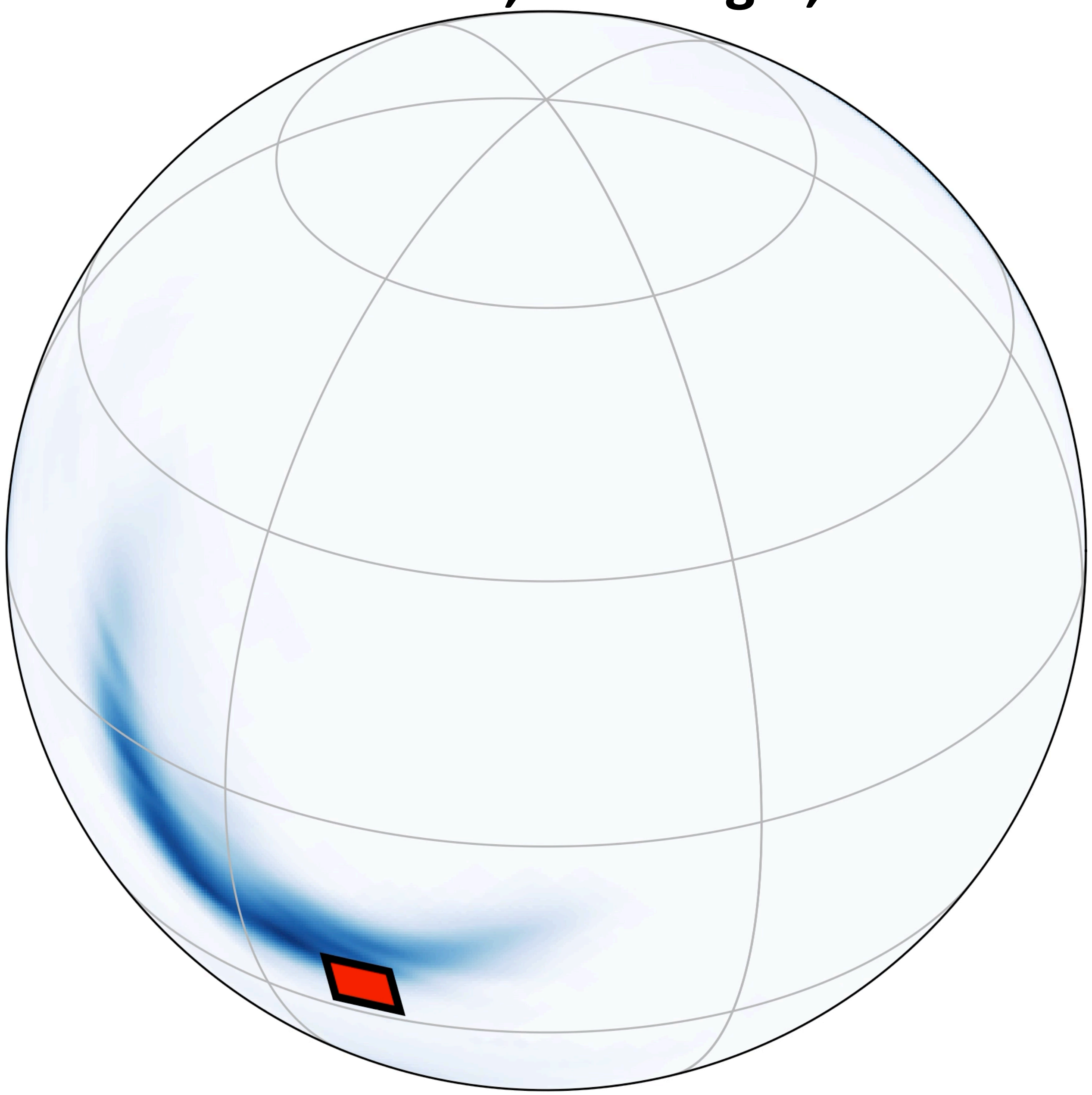


The Technical Landscape

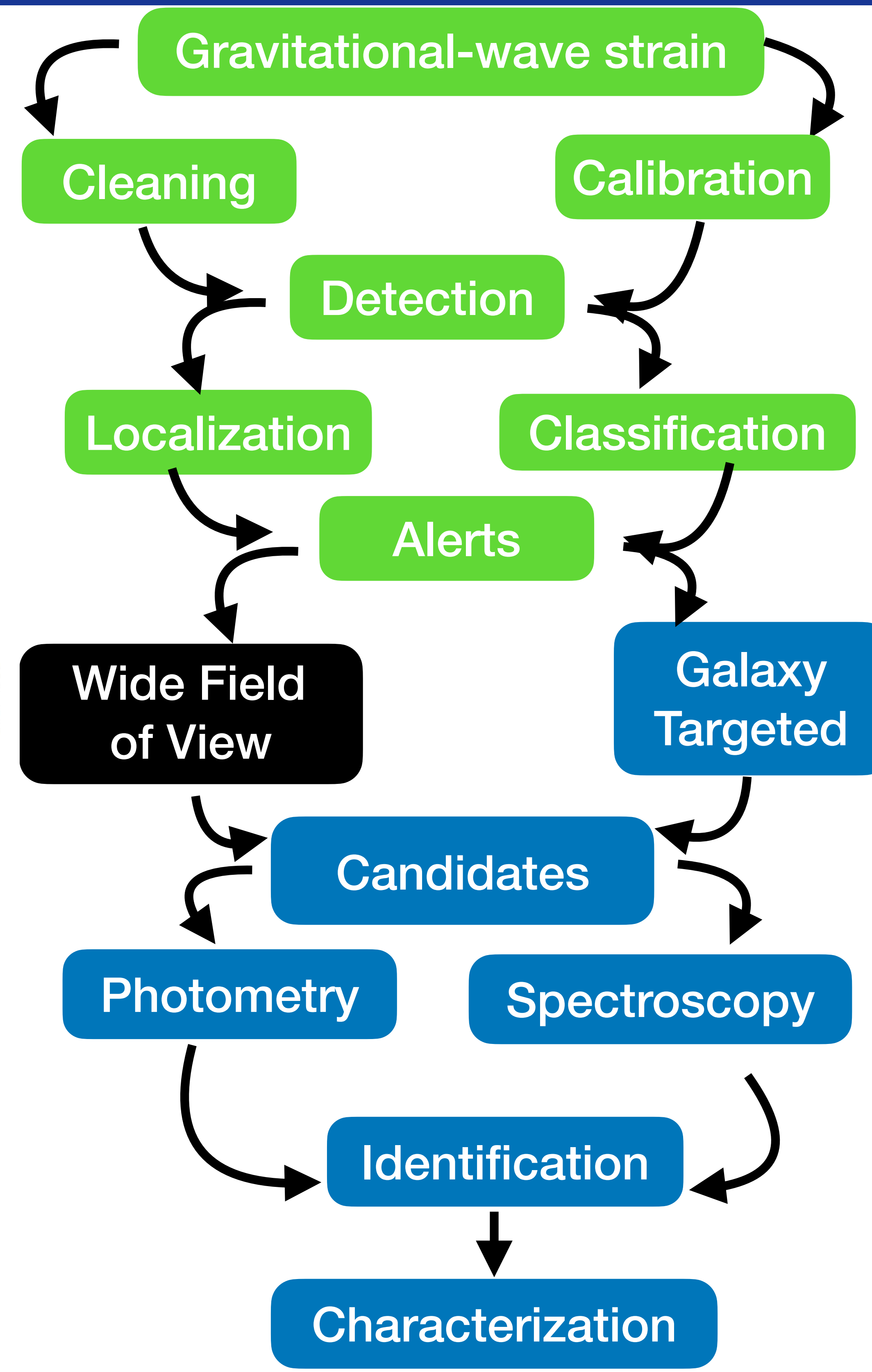


Wide Field Follow-up

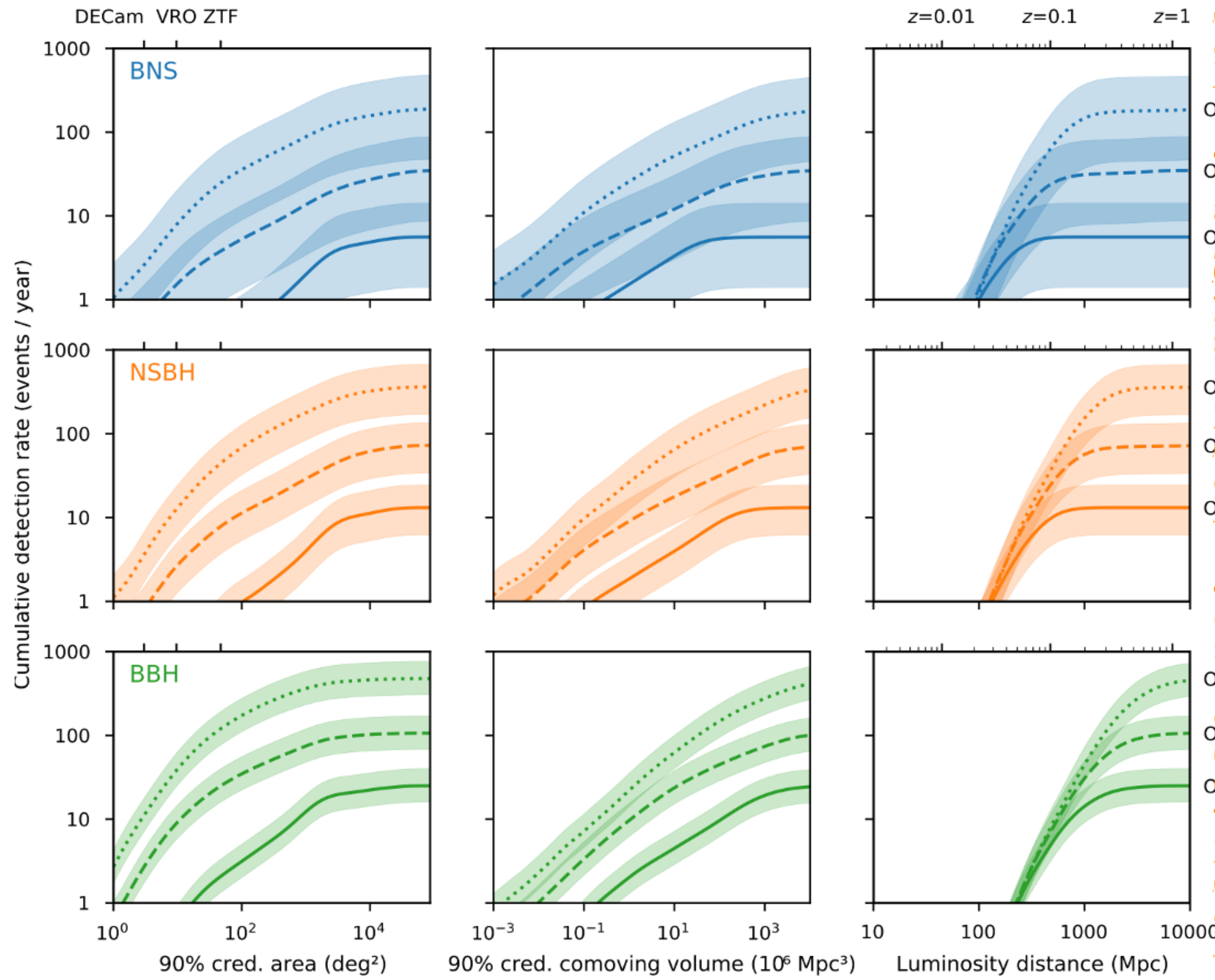
M4OPT: Mixed integer programming based scheduler, Leo Singer, PI



PC: Leo Singer, Goddard

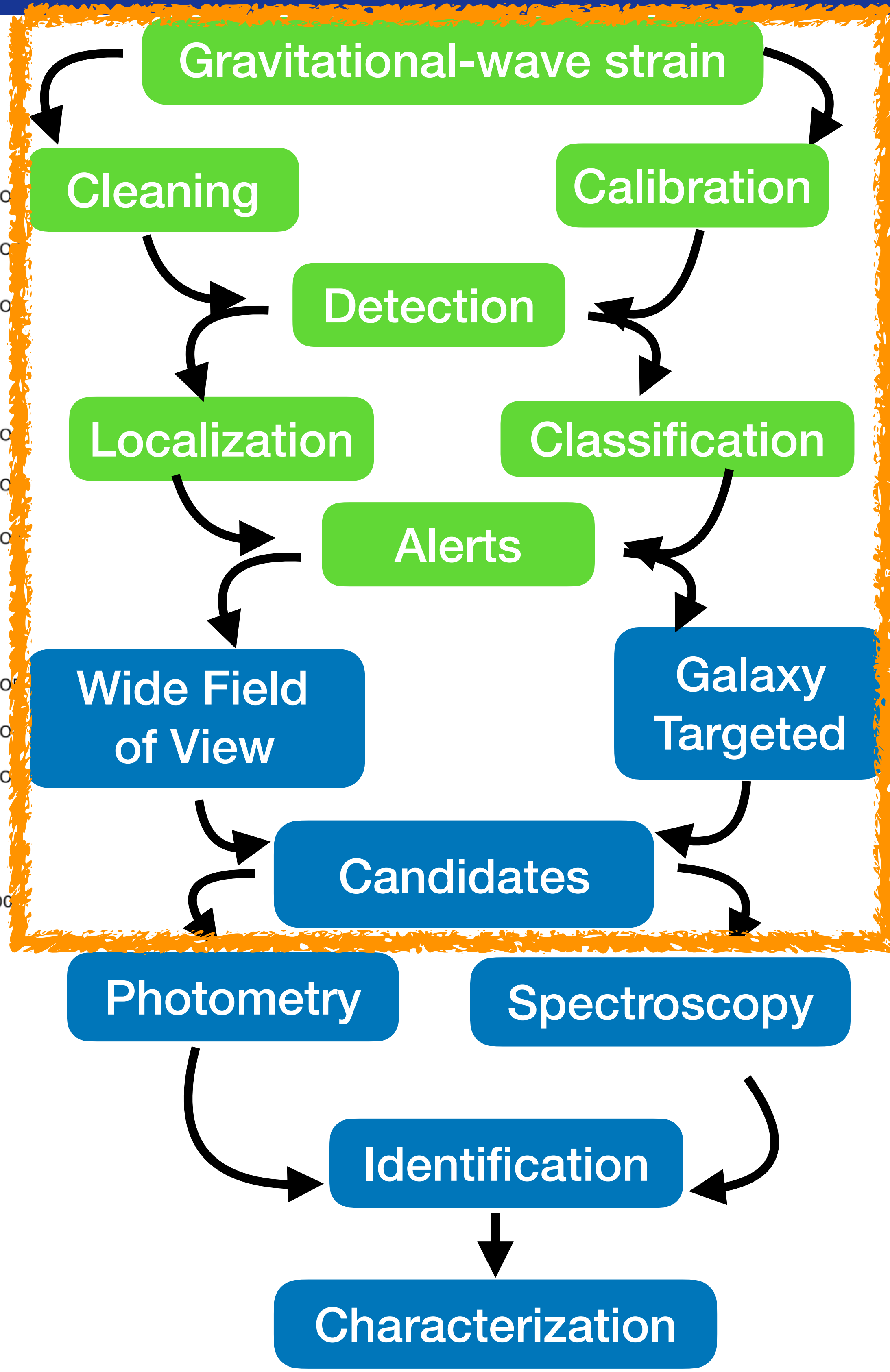


Observing Scenarios



Polina Petrov

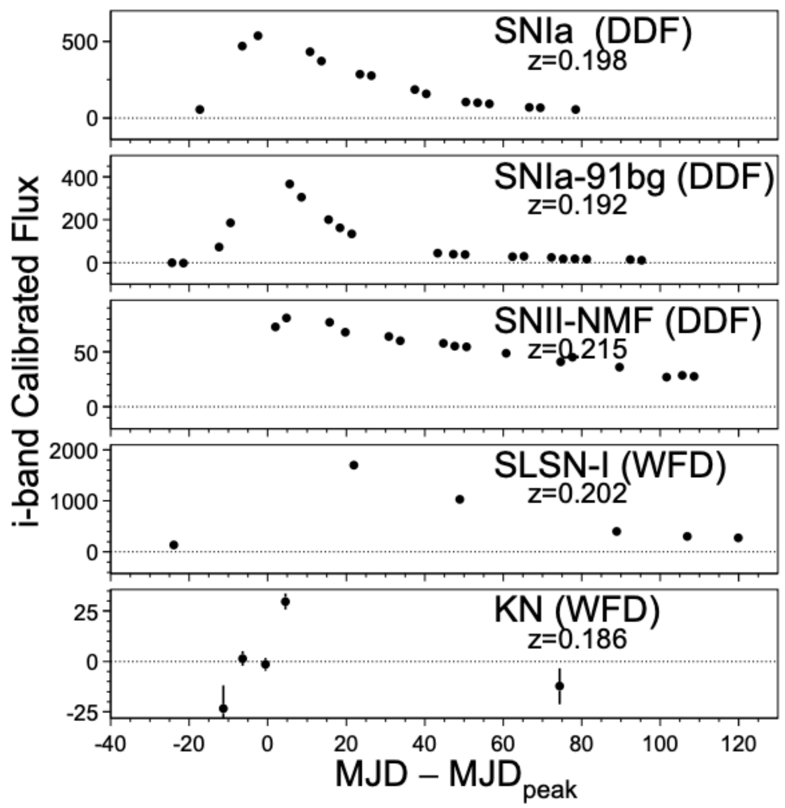
Vanderbilt



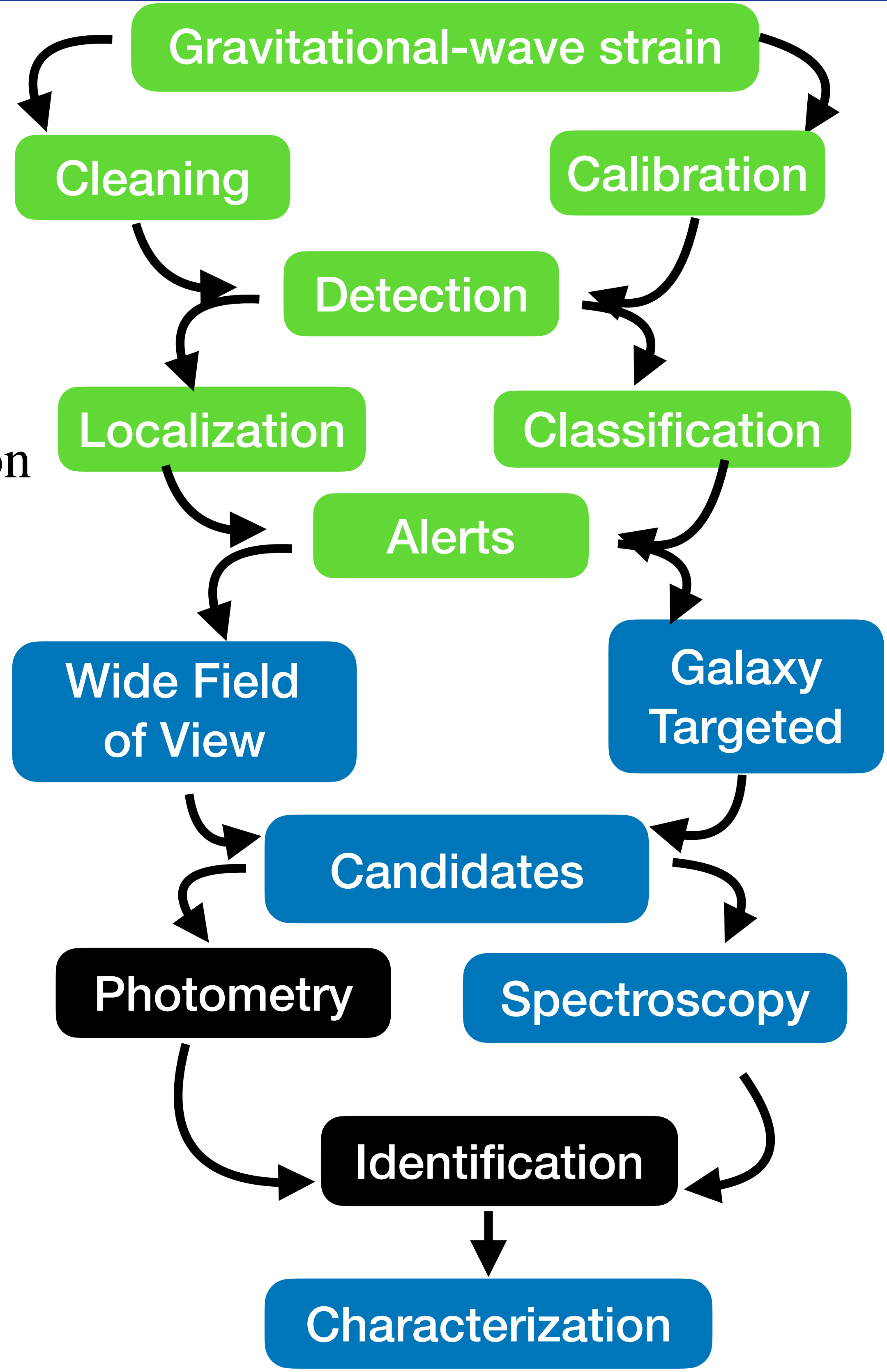
Petrov et al. 2021: 2108.07277

What data do we need?

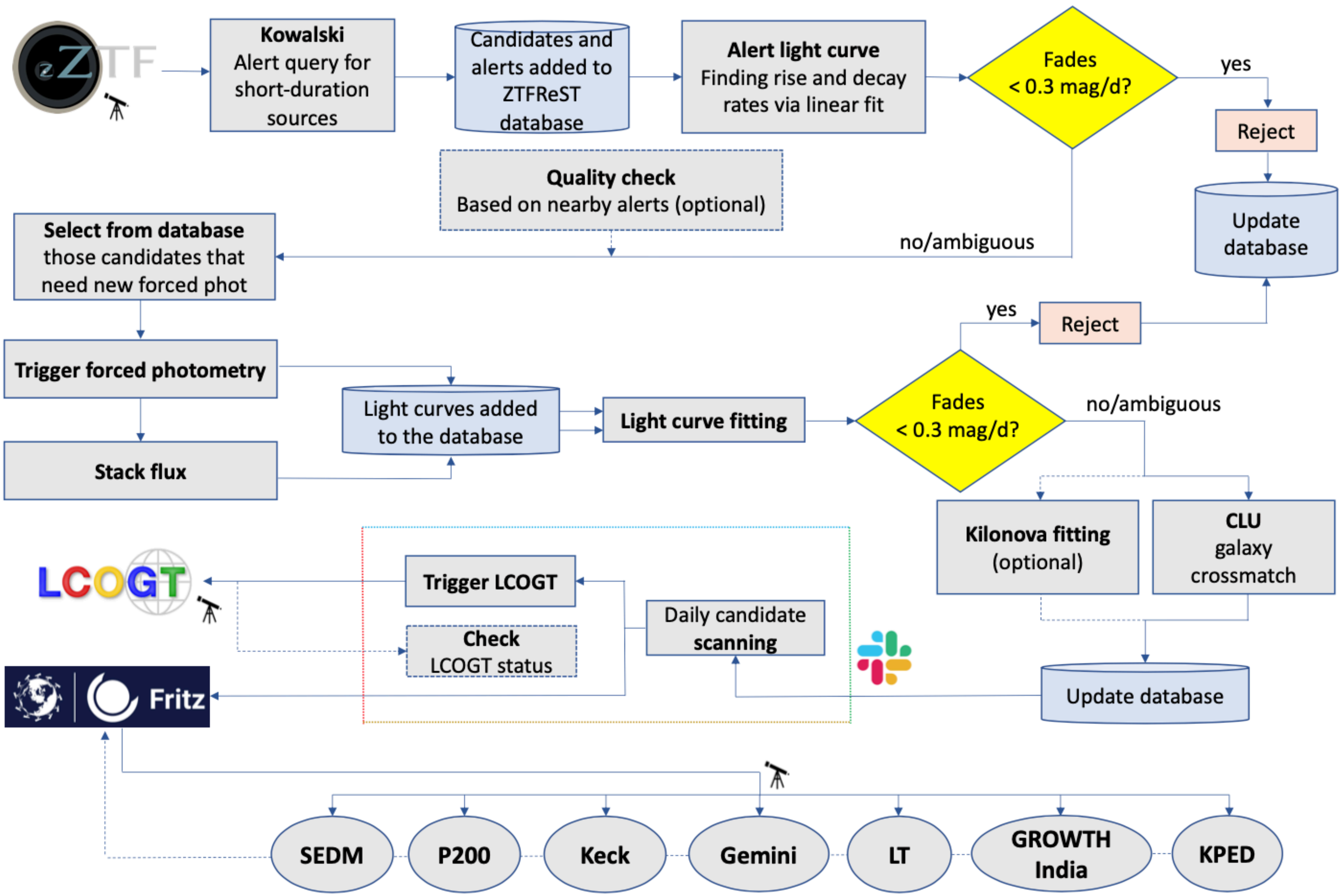
- Often, a photometric light curve is all you have available to classify it.
- Due to the many follow-up systems we have available, desire to design a system that optimizes the differentiation between models for kilonovae and other fast transients.
- Can use ML methods to speed up inference on each potential counterpart object, including when performing the GW and EM inference.



Kessler et al. (2019): 1903.11756

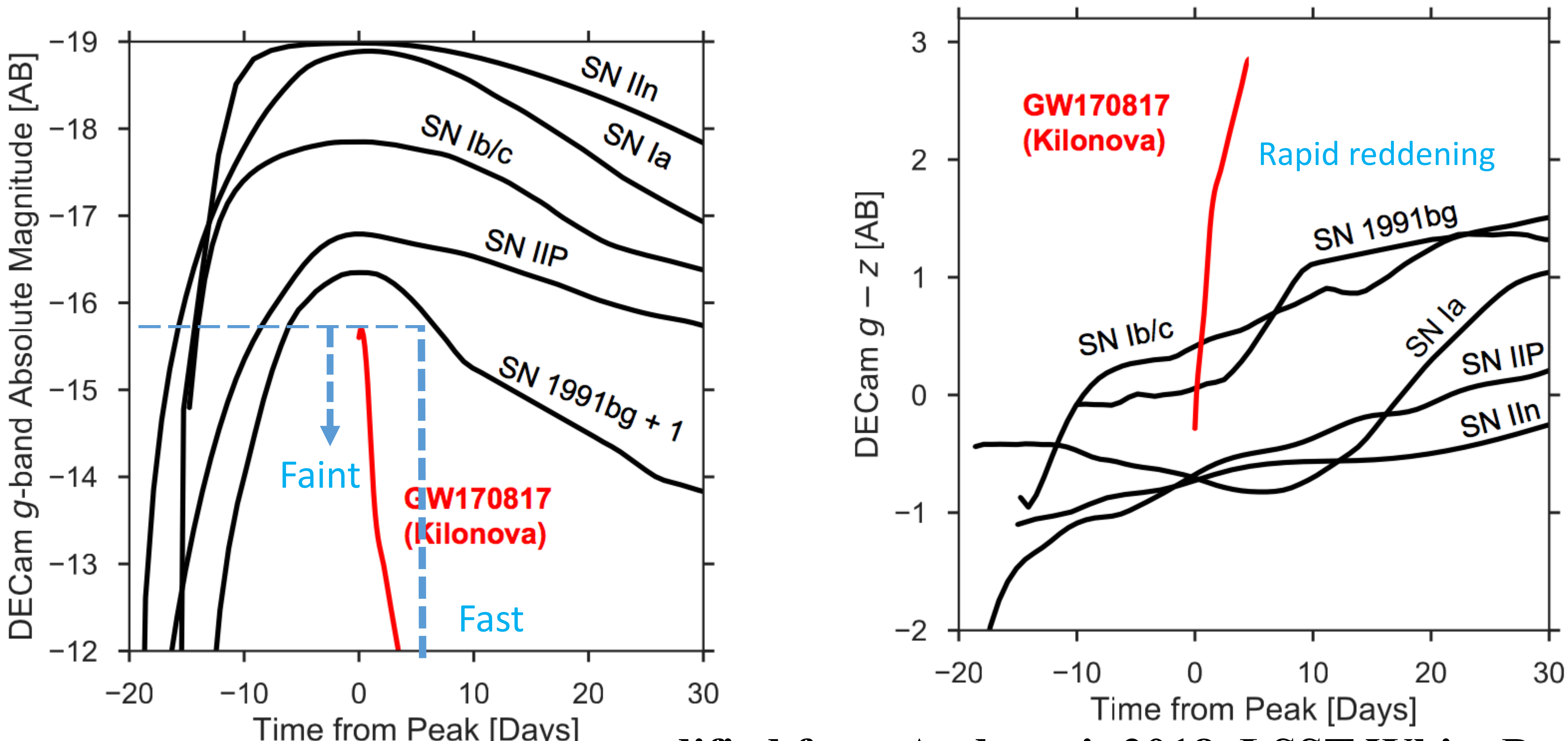


Transient Filtering



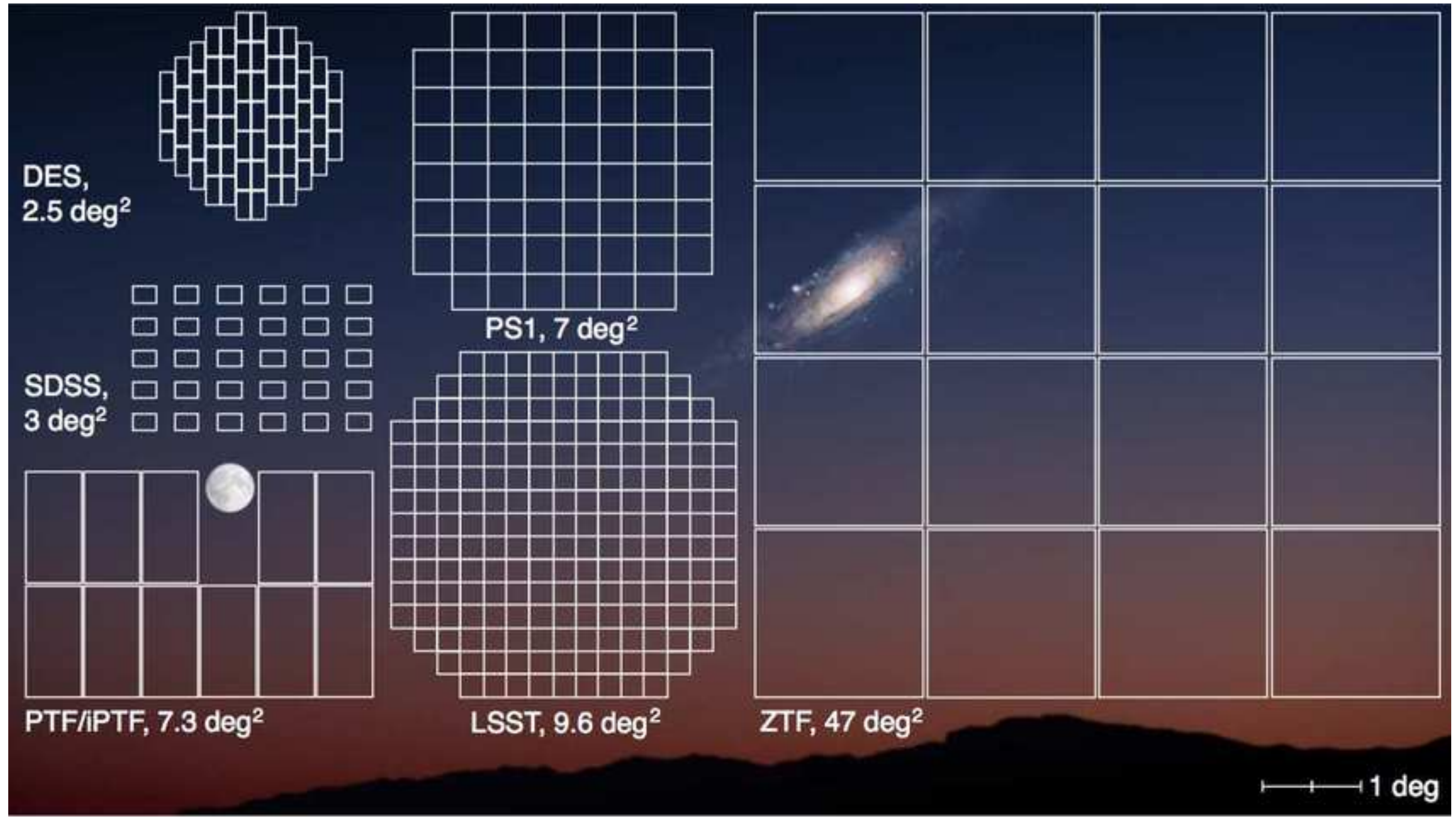
[Andreoni, Coughlin+2021, 2104.06352]

Why do we want fancy strategies? Kilonovae - Hard to find



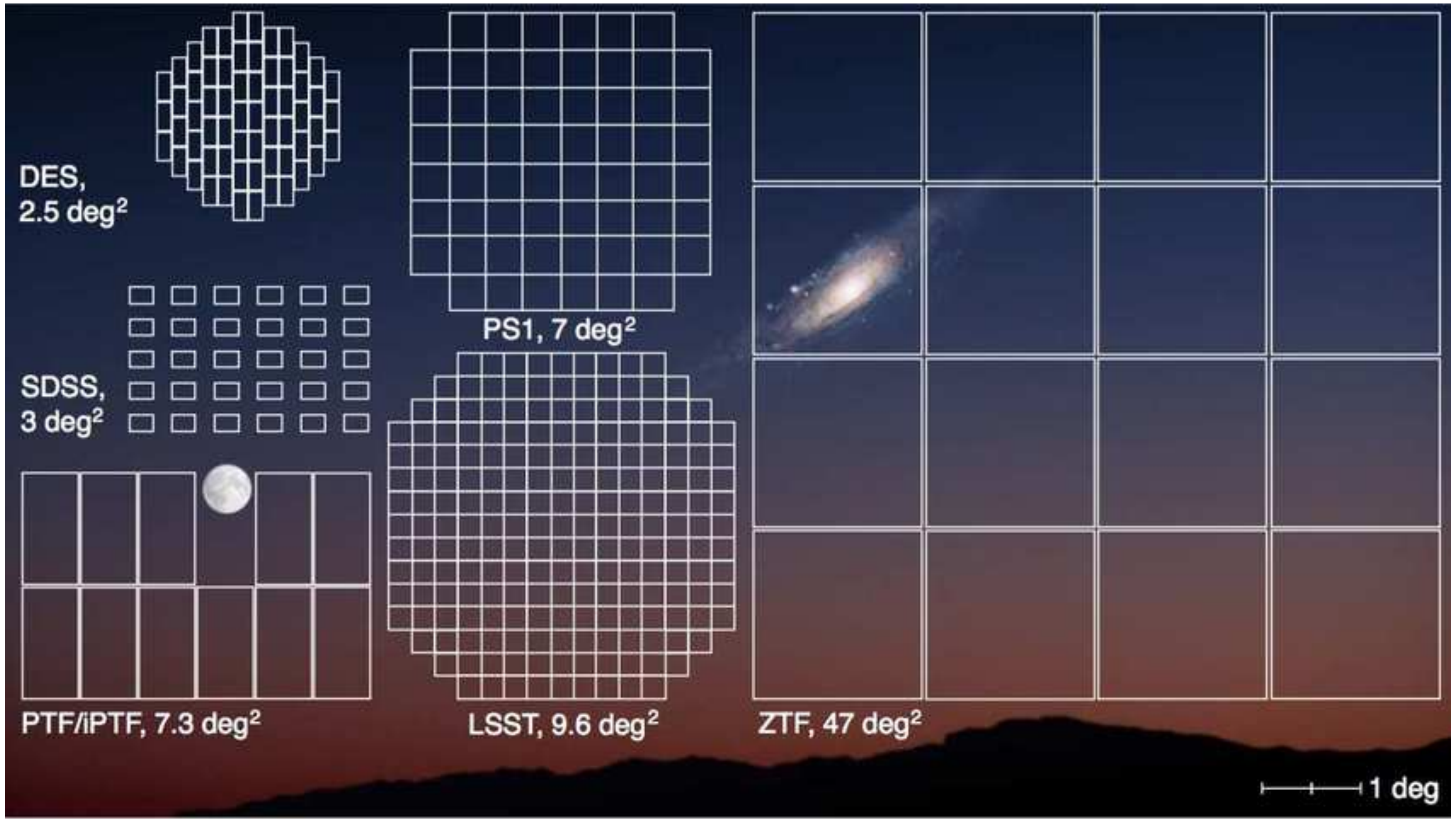
modified from Andreoni+2018, LSST White Paper

Landscape of Optical TDA



- **The night sky is imaged at 17.5 mag by ASAS-SN (both hemispheres)**
- **The northern sky is covered by ATLAS, ZTF, and PS-1 to 19, 20.5, 21.5 over roughly two nights (ZTF issues real time, data-rich alerts)**
- **BlackGEM (21-22 mag; Chile) will start routine operation within this year**
- **LSST is expected to become operational in 3 to 4 years**

Two Different Approaches



- Photometric detection followed by spectroscopic classification

– Possible for surveys which are shallow (20 mag or so)

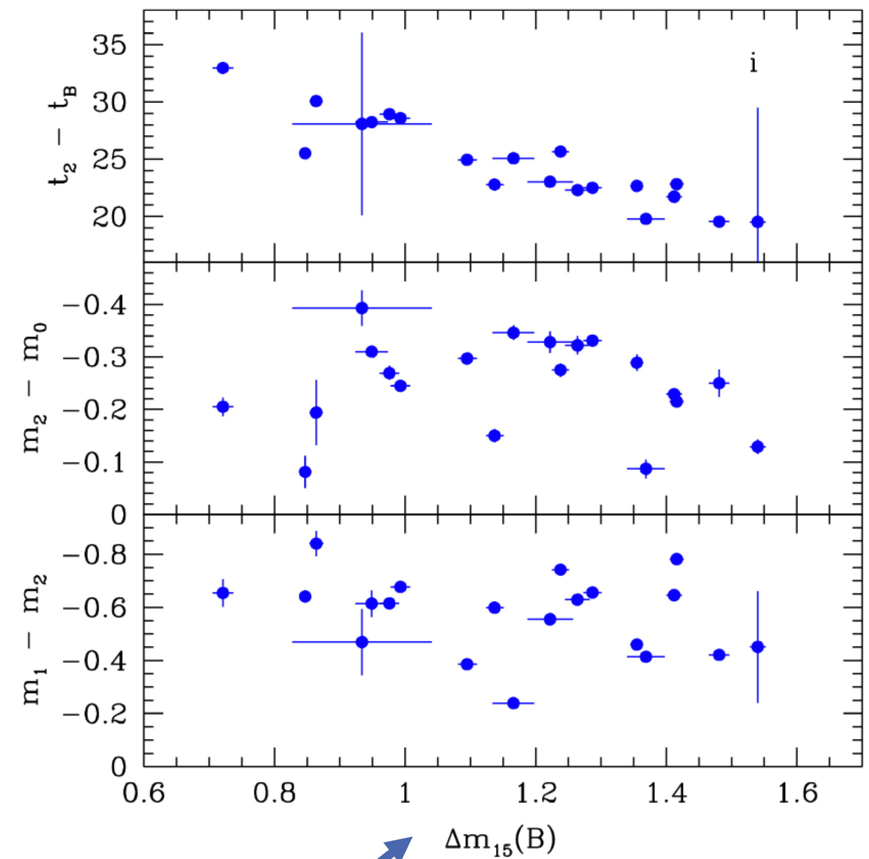
– Spectral classification can be undertaken by existing telescopes

- Photometric detection followed by multi-band time series

– Large samples of faint objects –Much of the analysis will be statistical –Use clever techniques and filter out a small subset for further follow up

Value-driven Real-time follow-up

- As is, data will be insufficient for full science inference without additional follow-up
  - e.g. extracting physics from light curves
- Need to perform value-driven follow-up
  - Volume of alerts far exceeds the ability to follow-up with limited and/or expensive follow-up resources
- Augment sparse Rubin LCs to improve constraints on SALT2 (supernova) models
- We augment photometry to (branch-normal) SN Ia LCs from ZTF-I public survey (g and r) using P48 in g,r, and i
  - i-band important for precisely estimating H0 (Burns+ 2018)
  - Second peak could help probe SN Ia explosion mechanisms
- Broadly an optimal real-time resource allocation problem and not restricted to SALT2



Peak luminosity

Folatelli+ 2010

Niharika (Ari) Sravan

Tyler Barna

UMN

Caltech

Sraven et al. 2021: submitted

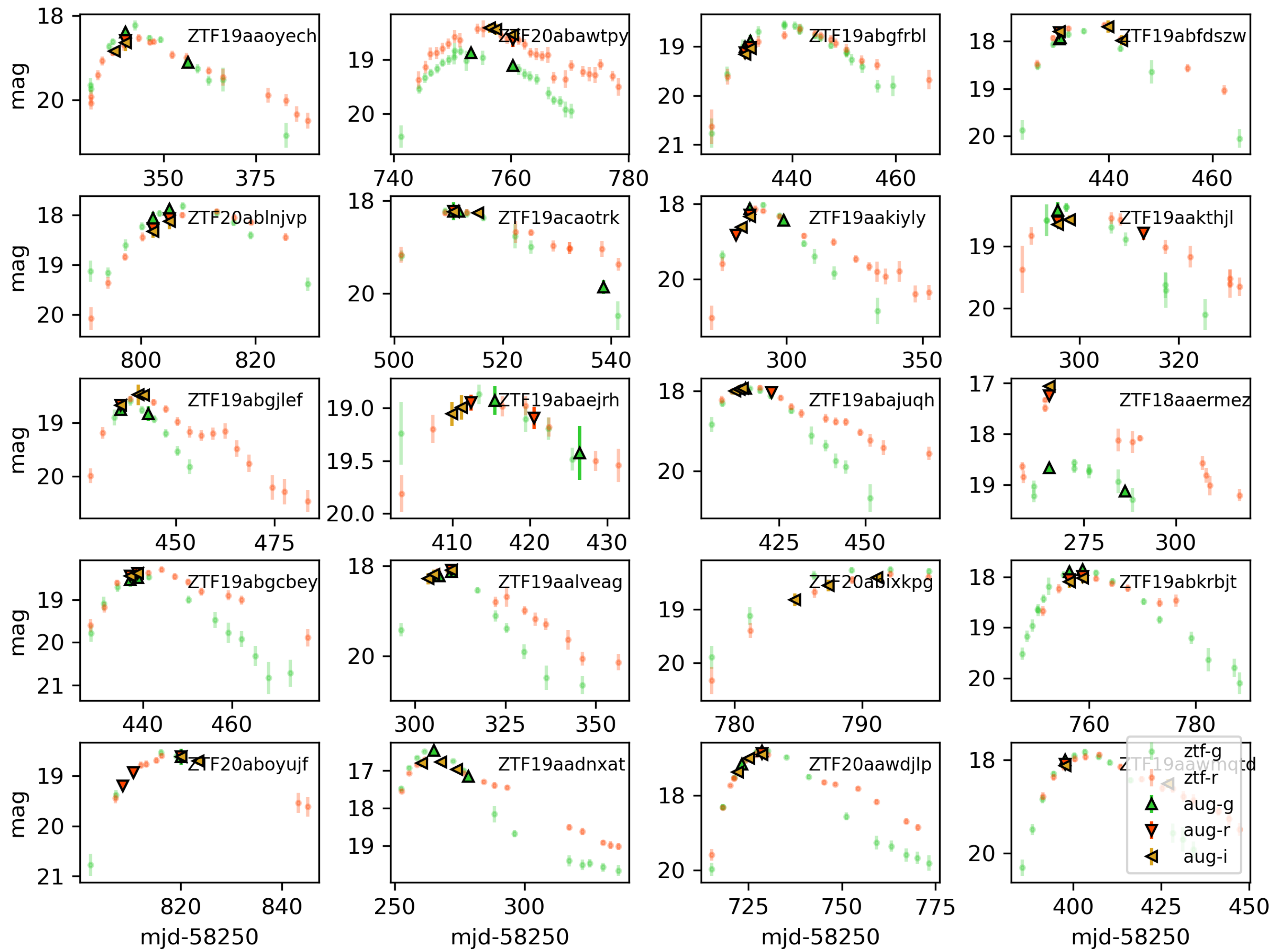
Results

10-20 % Median improvement in parameters over random allocation

Gap filling

Resolves phase with high variability/ diversity:

- Around peaks and valleys

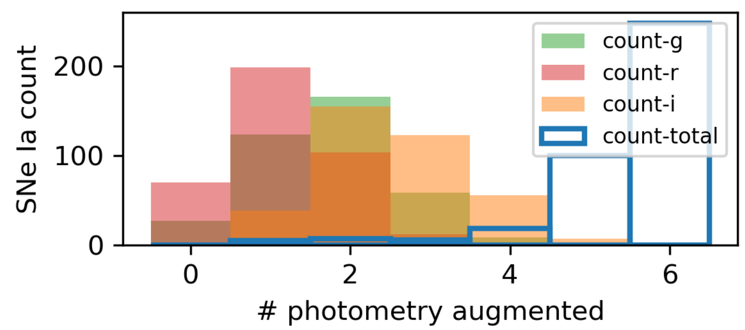


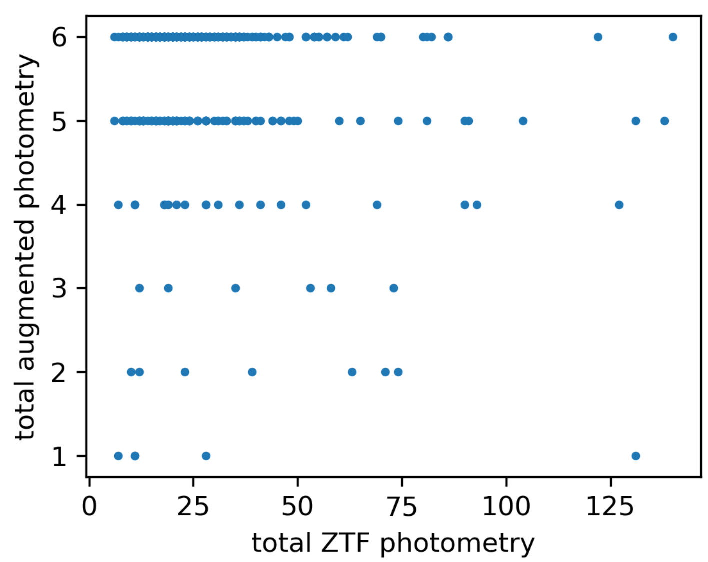
Sraven et al. 2021: submitted

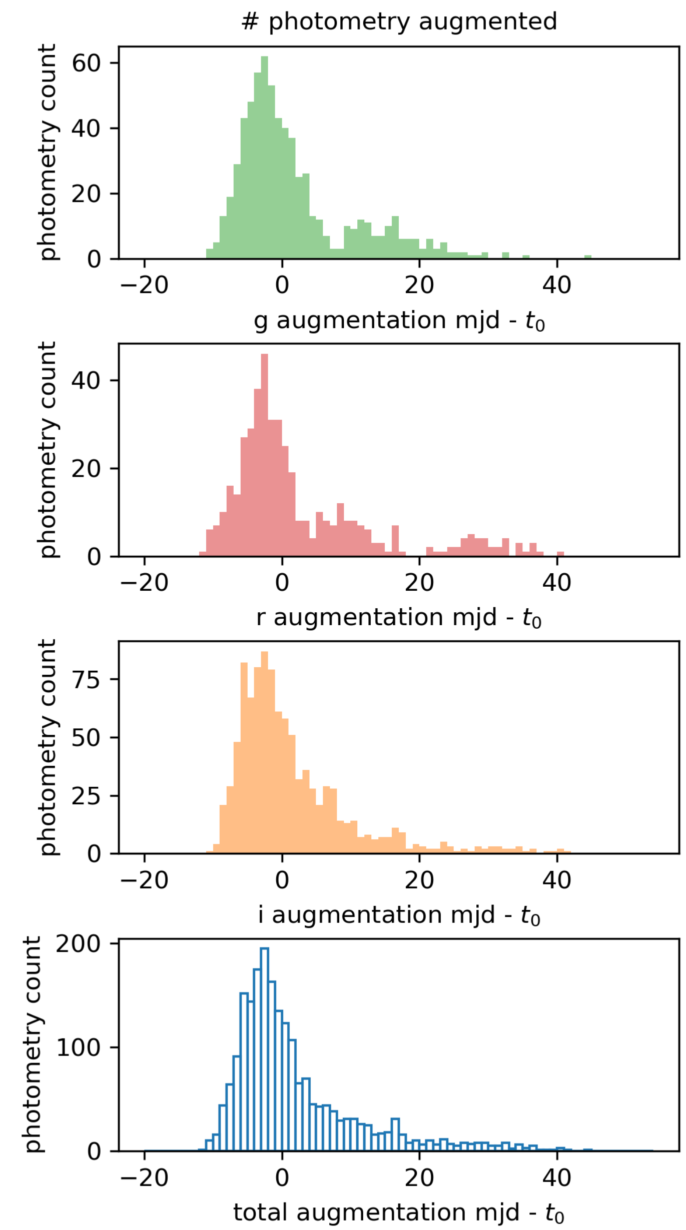
Results

Interesting notes:

- More in g due to sparser sampling
- 3-5% more improvement for SNe Ia > 18.5mag
  - Even better prospects for Rubin





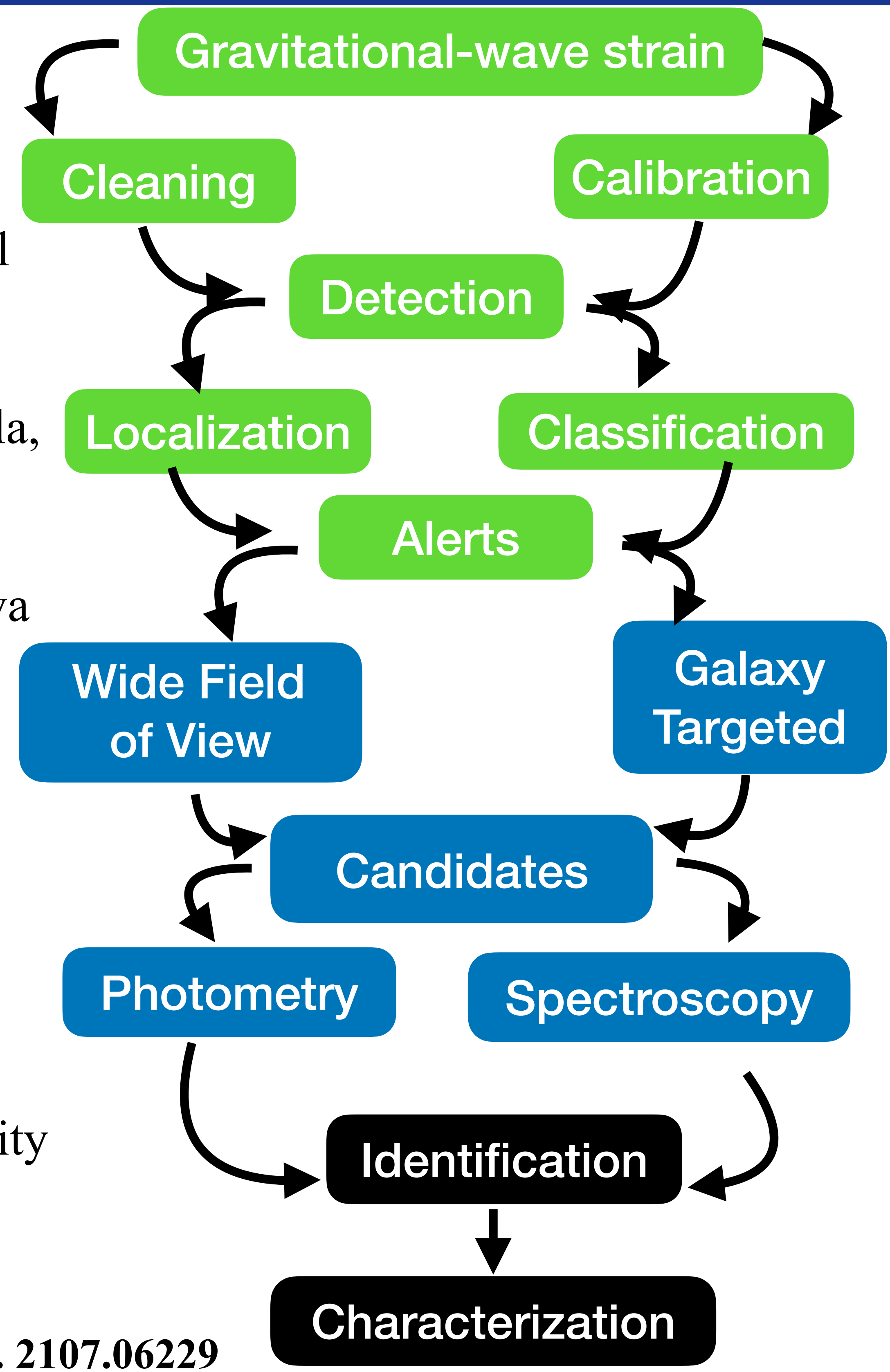


Sraven et al. 2021: submitted

So… what then?

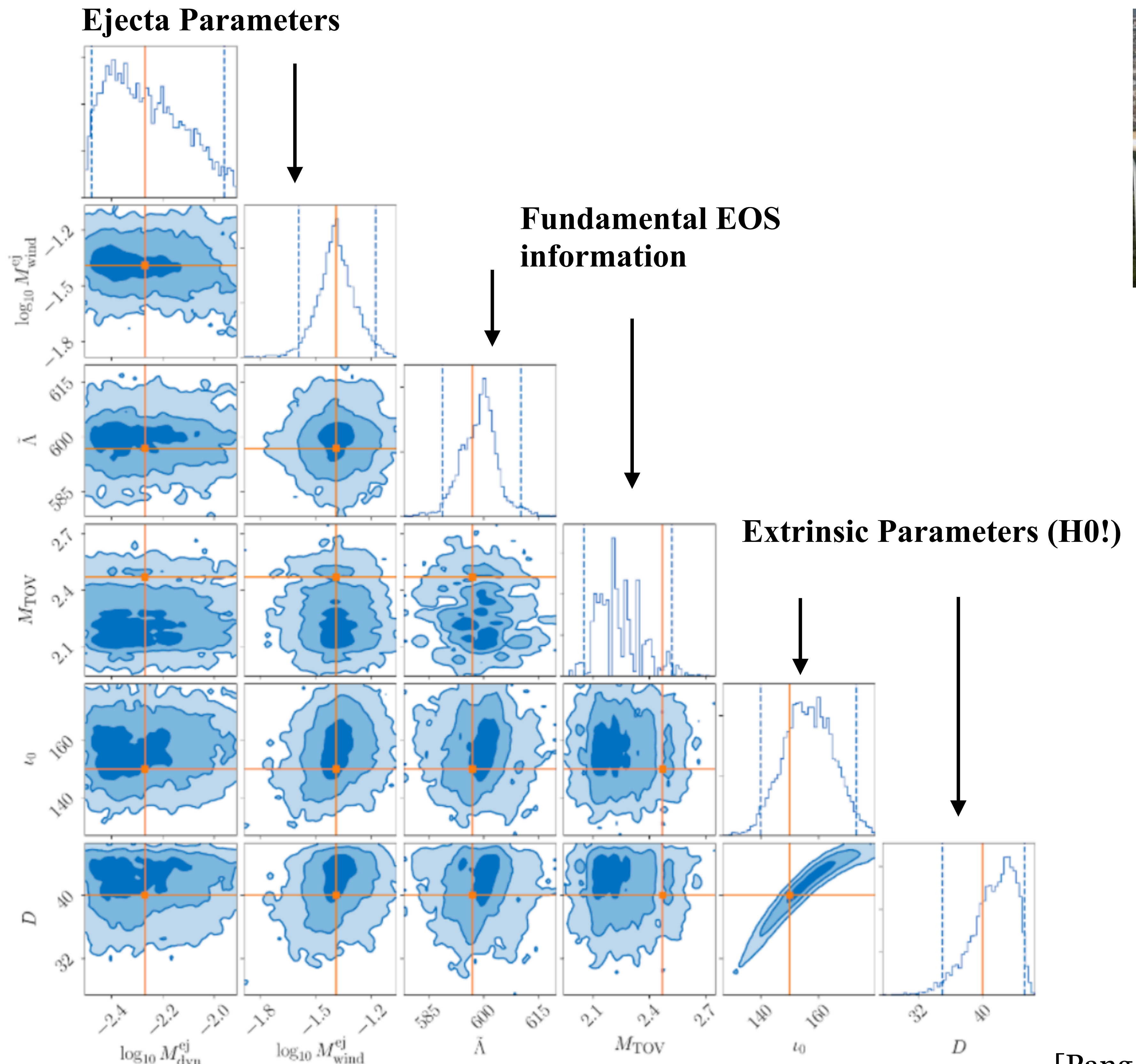
NMMA: A Fully Bayesian Joint-Inference Pipeline

- gravitational-wave data analysis using parallel bilby
- kilonova modelling with various models (Bulla, Kasen, etc.)
- gamma-ray burst afterglow fits (also supernova models from sncosmo)
- chiral effective field theory to simulate the neutron-star EOS
- neutron-star maximum mass and NICER constraints, fits to relate ejecta parameters to progenitor parameters using numerical relativity



Used for many analyses at this point:
Ahumada et al. 2105.05067, Dietrich et al. 2002.11355,
Tews et al. 2007.06057, Pang et al. 2105.08688, Huth et al. 2107.06229

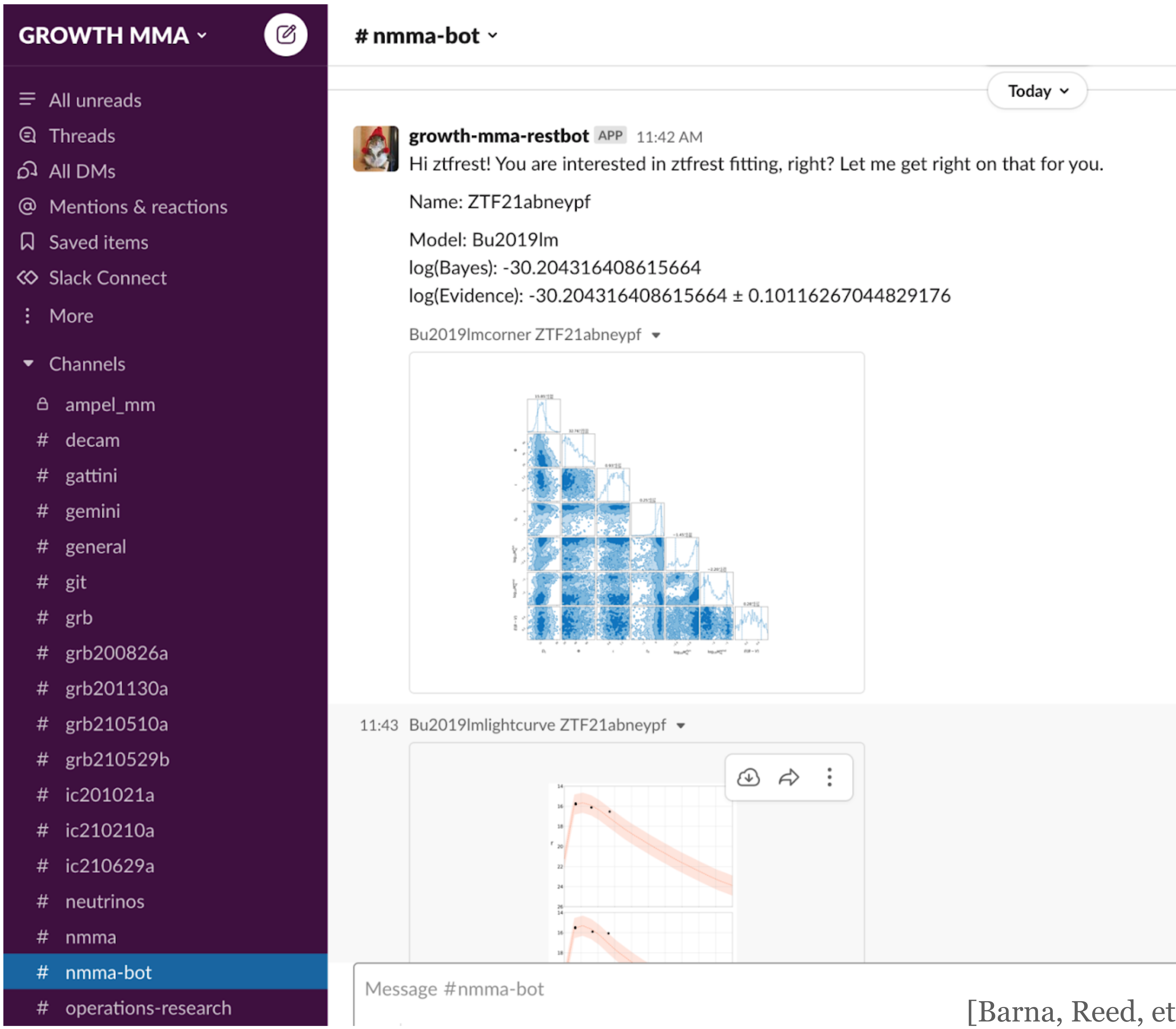
NMMA: A Fully Bayesian Joint-Inference Pipeline



- Extract science (and filtering criteria) in one place!
- For example, immediately extract information on ejecta, neutron star physics, and cosmology (in case of a host galaxy).

[Pang+, 2021, in prep]

Online Filtering



UMN

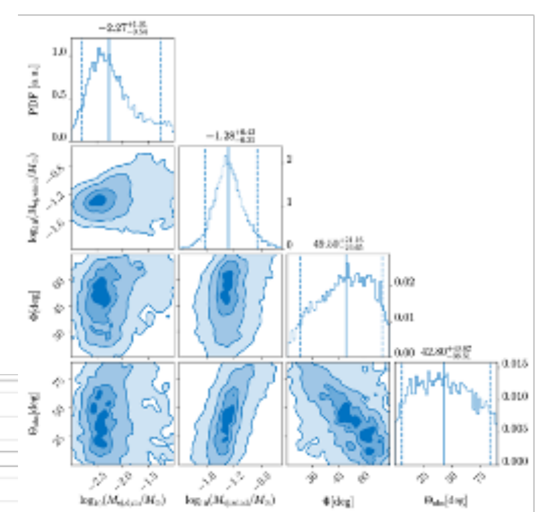
UMN

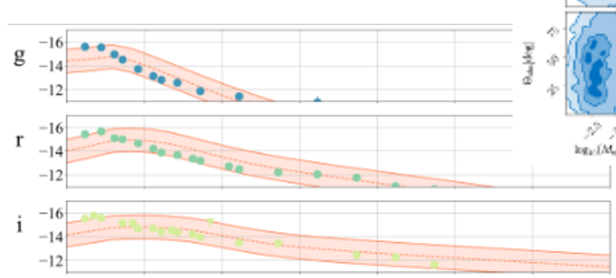
[Barna, Reed, et al., in prep]

Improving community follow-up

SkyPortal

NMMA



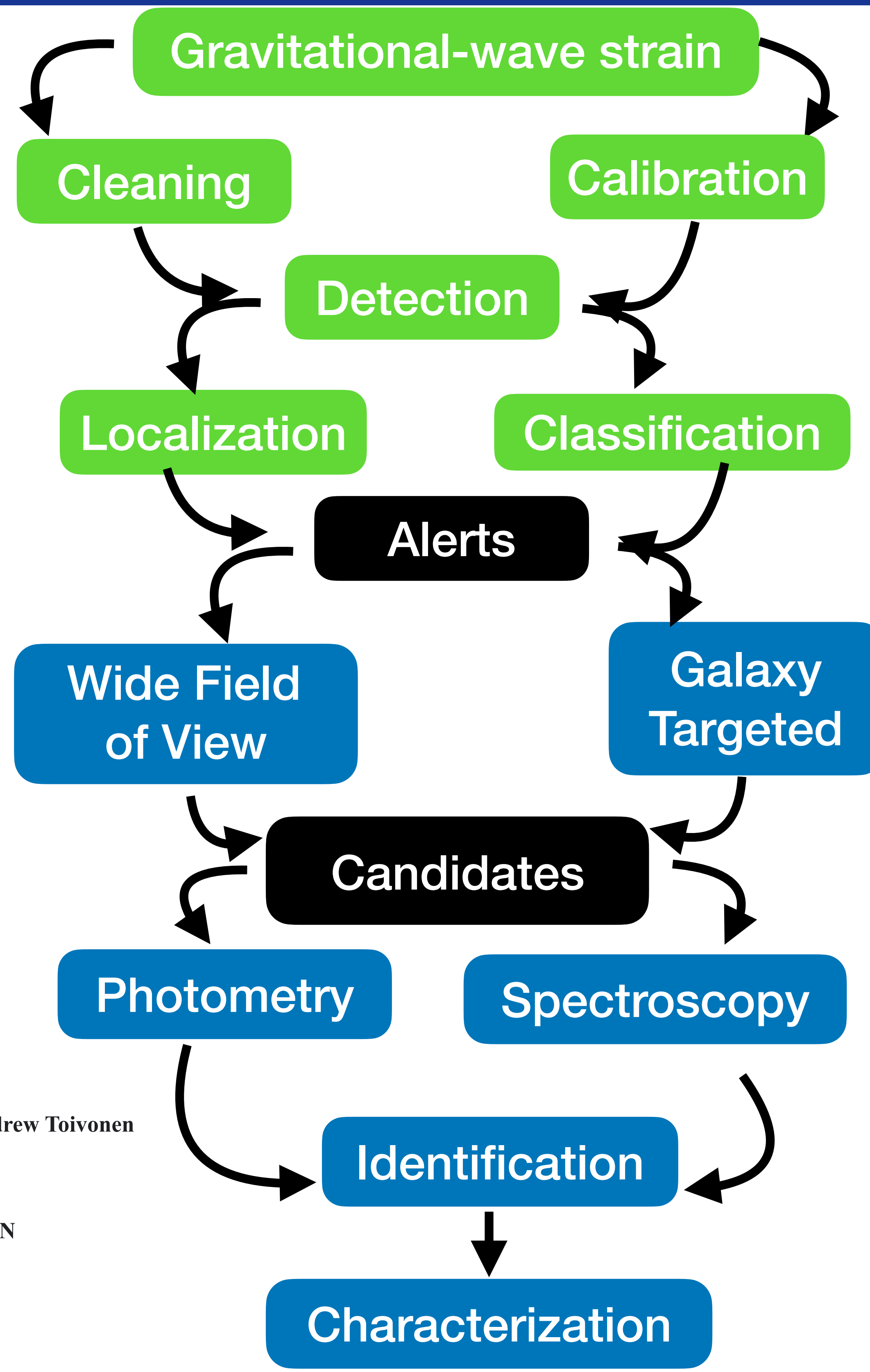


Gargi Mansingh

American

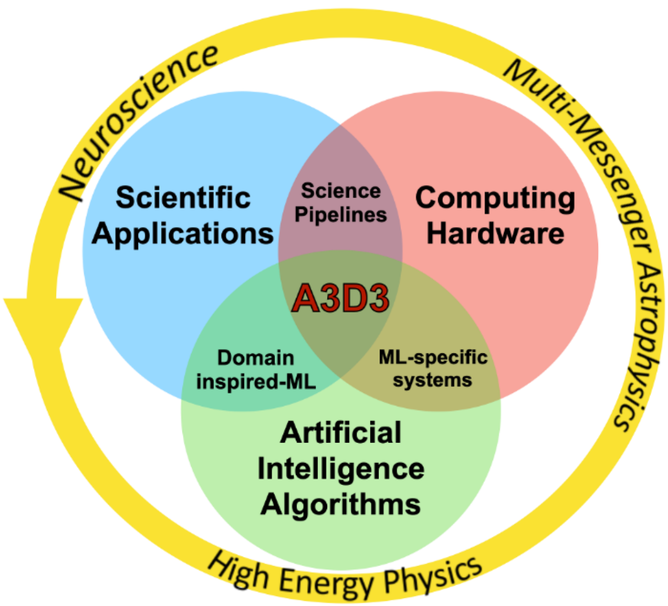
Andrew Toivonen

UMN



A3D3

Accelerated AI Algorithms for Data-Driven Discovery



Join us!

- Monday 9 am Central: Technical MMA call (anything related to gravitational-wave counterpart searches)
- Tuesday 2 pm Central: A3D3 ML Detection meeting
- Thursday 8:30 am Central: A3D3 KAGRA meeting
- Friday 2 pm Central: A3D3 Inference as a Service meeting

Summary and Perspectives

IAAS

- IaaS model represents a powerful way to bring advantages of deep learning to bear in gravitational-wave astronomy
- Deploying and optimizing IaaS pipelines requires aligning benefits of scale with constraints of problem
- Several applications currently under development to increase number and speed of event detections during O4

ML based follow-up

- Technique interpretable
  - Good start before invoking reinforcement learning (also for benchmarking)
- Data acquisition failure tolerant
  - Latency intolerant
- Main uncertainty is the reliability of simulated augmented photometry
  - Need verification with real data
- Extensions of work include:
  - Variable observing cost, observing season based budget
  - Other choices of utility including prior building, model discrimination

My own perspective on the areas of greatest need:

- Can the initial promise of ML applications in terms of detection and PE make its way from BBH signals (short) to BNS signals (long)?
- Can ML provide optimal follow-up strategies to rule in and out specific transients as sources given limited telescope time and sensitivity?
- Is ML the key to a truly MMA pipeline, with inference on GW, optical, GRB, etc. data sets?

Thank you!

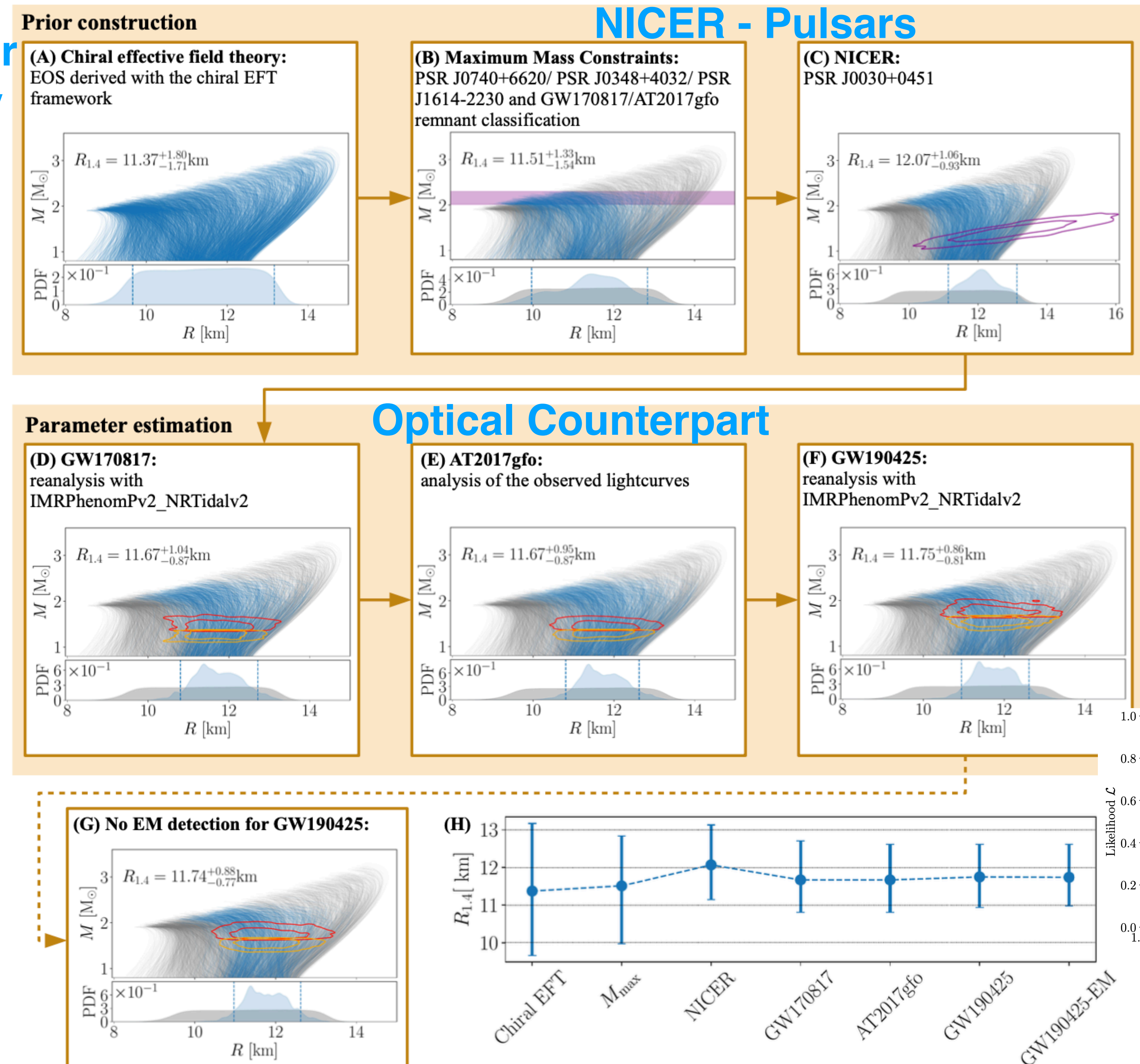
MMA Equation of State Constraints

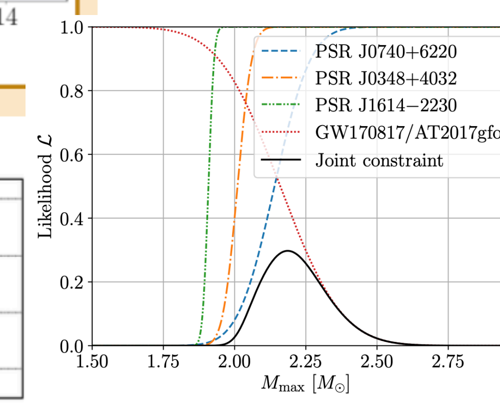
Nuclear Theory

NICER - Pulsars

Optical Counterpart

GWs





[Dietrich, Coughlin, et al., Science]

Methodology

The Goal

- Given an observing budget decide how to augment LCs in real-time (adapt to collected data) such that we maximize expected utility (EU) for the augmented LC (realized at the end of the episode)
  - Utility conveys our preference for an outcome given a decision. We choose utility as pseudo* A-optimality = minimum SALT2 parameter variances in sncosmo
  - Since we do not assume that redshift is known as the SN is taking place, we solve for it and minimize its uncertainty as well. Assume known SN sky location, MW extinction.

The Algorithm

- On each day estimate EU of action space {no action, g, r, i, gr, ri, ig, gri} given observed data and expected data (stochastic) under no action. Remaining budget allocated randomly**
  - Outcome states given actions and expected future data estimated using encoder-decoder LSTM trained on 105 simulated ZTF SNe Ia
- Take modal action with least cost having max EU*** per N simulated future outcomes
  - Augmentations from 2-D Gaussian Process fit to full LC and fed back for the next day

* because `sncosmo` solves chi^2 minimization   max likelihood
** substitutes expected **optimal** actions for expected **naïve** actions
***   -greedy and tuned to maximize median A-optimality for all validation SNe Ia

Sraven et al. 2021: submitted