SkyPortal

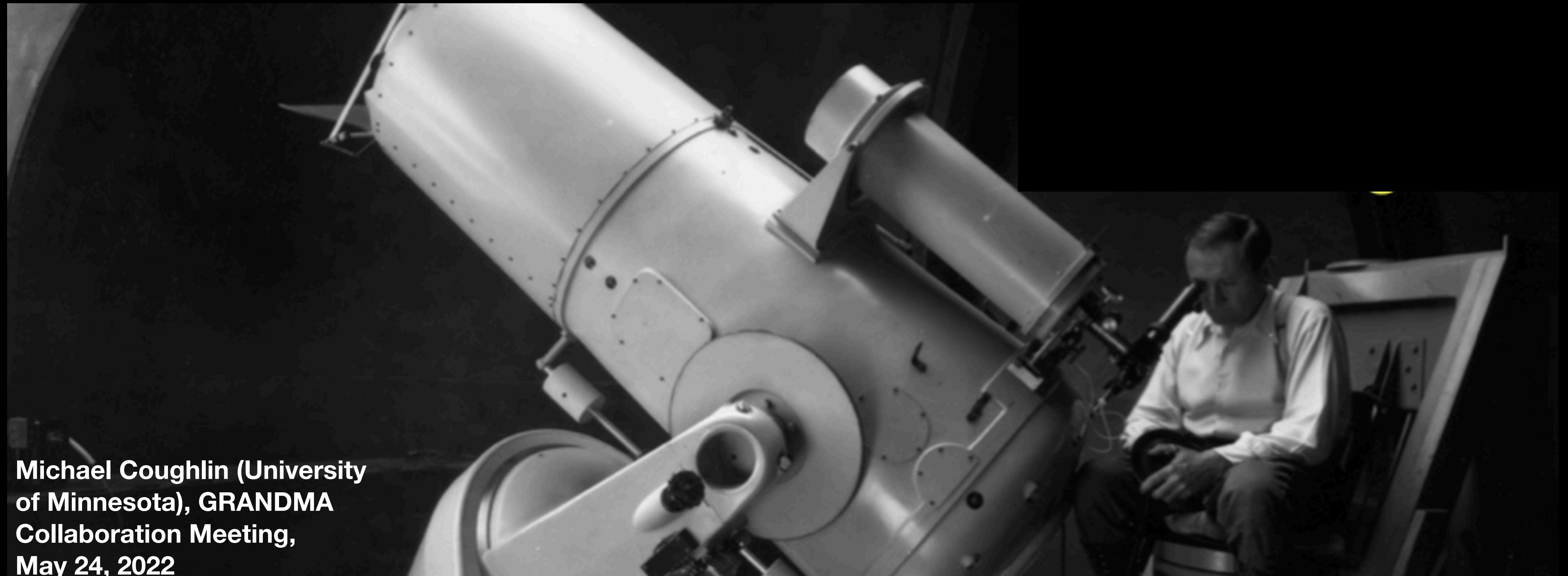## Towards a multi-messenger data science platform for the MMA era

Michael Coughlin (University of Minnesota), GRANDMA Collaboration Meeting, May 24, 2022

The History

O3 showed us how hard follow-up can be, and O4 is not going to get easier, as our interferometers remain at quite different sensitivities (plus the unfortunate situation with KAGRA's sensitivity).

However, because O3 was challenging, we built/adapted many, many quality tools both during the run and afterwards to analyze the results.

iCARE, gwemopt, simsurvey, POSSIS, gwemlightcurves/nmma, NIMBUS, various annotation pipelines, the GROWTH ToO-Marshal, the GROWTH Marshal, etc.

Table 2. Summary Statistics for O3, O4, and O5.

Run	BNS	NSBH	BBH
Median 90% credible area (deg^2) [a]			
O3	1672^{+94}_{-110}	1970^{+110}_{-110}	1069^{+43}_{-41}
O4	1820^{+190}_{-170}	1840^{+150}_{-150}	335^{+28}_{-17}
O5	1250^{+120}_{-120}	1076^{+65}_{-75}	$230.3^{+7.8}_{-6.4}$
Median 90% credible co-moving volume (10^6 Mpc^3) [a]			
O3	$6.62^{+0.97}_{-0.97}$	$44.1^{+7.4}_{-5.2}$	217^{+23}_{-16}
O4	$44.8^{+6.4}_{-6.5}$	191^{+20}_{-27}	216^{+16}_{-20}
O5	125^{+21}_{-12}	448^{+61}_{-44}	538^{+23}_{-24}
Median luminosity distance (Mpc) [a]			
O3	$176.1^{+6.2}_{-5.7}$	$337.6^{+10.9}_{-9.6}$	871^{+31}_{-28}
O4	$352.8^{+10.3}_{-9.8}$	621^{+16}_{-14}	1493^{+25}_{-33}
O5	620^{+16}_{-17}	1132^{+19}_{-23}	2748^{+30}_{-34}
Sensitive volume (10^6 Mpc^3) [ab]			
O3	$17.5^{+1.4}_{-1.3}$	$101.1^{+6.4}_{-6.1}$	1047^{+50}_{-49}
O4	$109.0^{+6.7}_{-6.5}$	558^{+26}_{-26}	4450^{+130}_{-130}
O5	590^{+29}_{-28}	2787^{+89}_{-87}	19950^{+310}_{-310}
Annual number of detections [cd]			
O3	5^{+14}_{-5}	13^{+15}_{-9}	24^{+18}_{-12}
O4	34^{+78}_{-25}	72^{+75}_{-38}	106^{+65}_{-42}
O5	190^{+410}_{-130}	360^{+360}_{-180}	480^{+280}_{-180}

The Vision

We have also seen our team evolve (both in terms of people coming and going, as well as career stages). Even during these 1+ year runs, these aspects change as well.
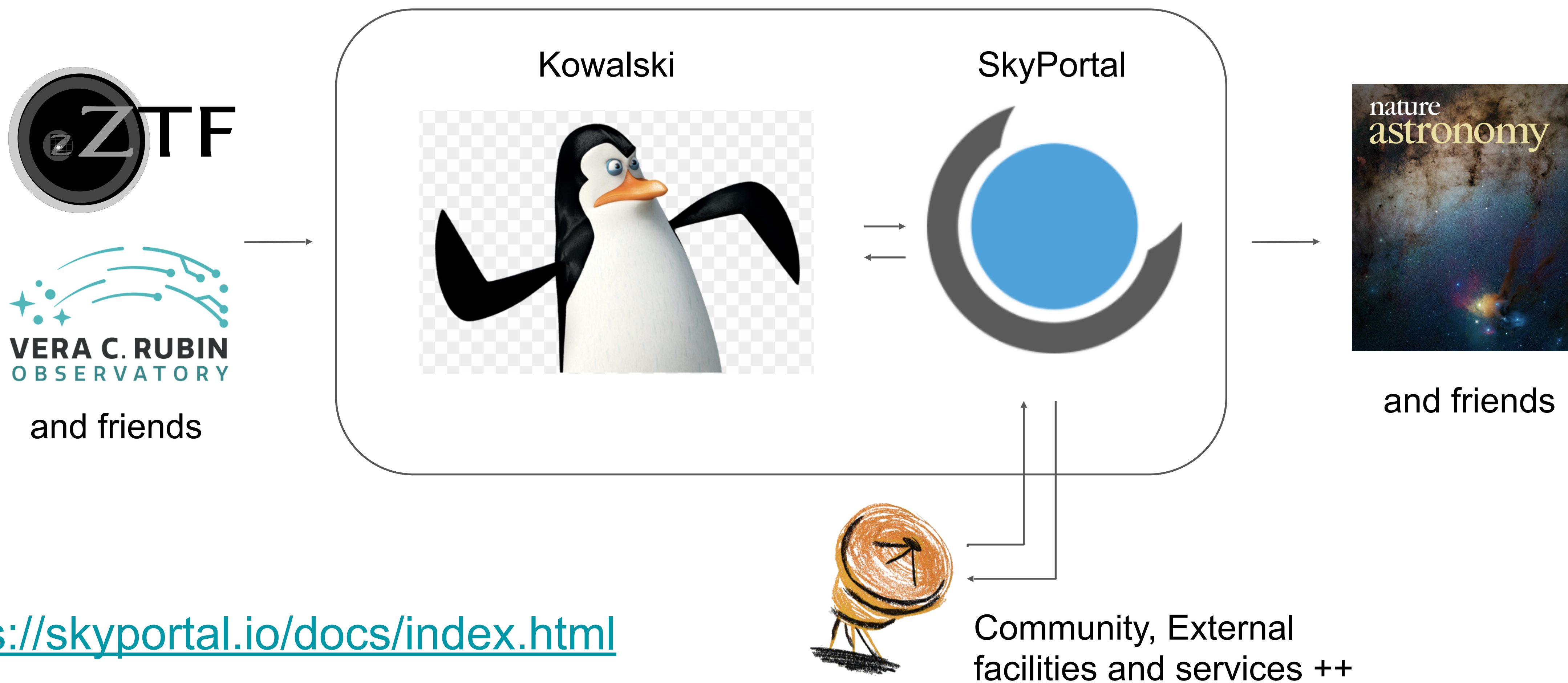
For this reason, we have set out to use SkyPortal as our one-stop-shop where many of these analyses should be available. This enables us to both save ourselves both in time and energy; the run is grueling (but obviously potentially incredibly rewarding) and limiting single points of success* is important. Even if in principle more than one person knows how to do something, it can be often the case that basically one person knows how to do it, and that's a tough spot to be in.
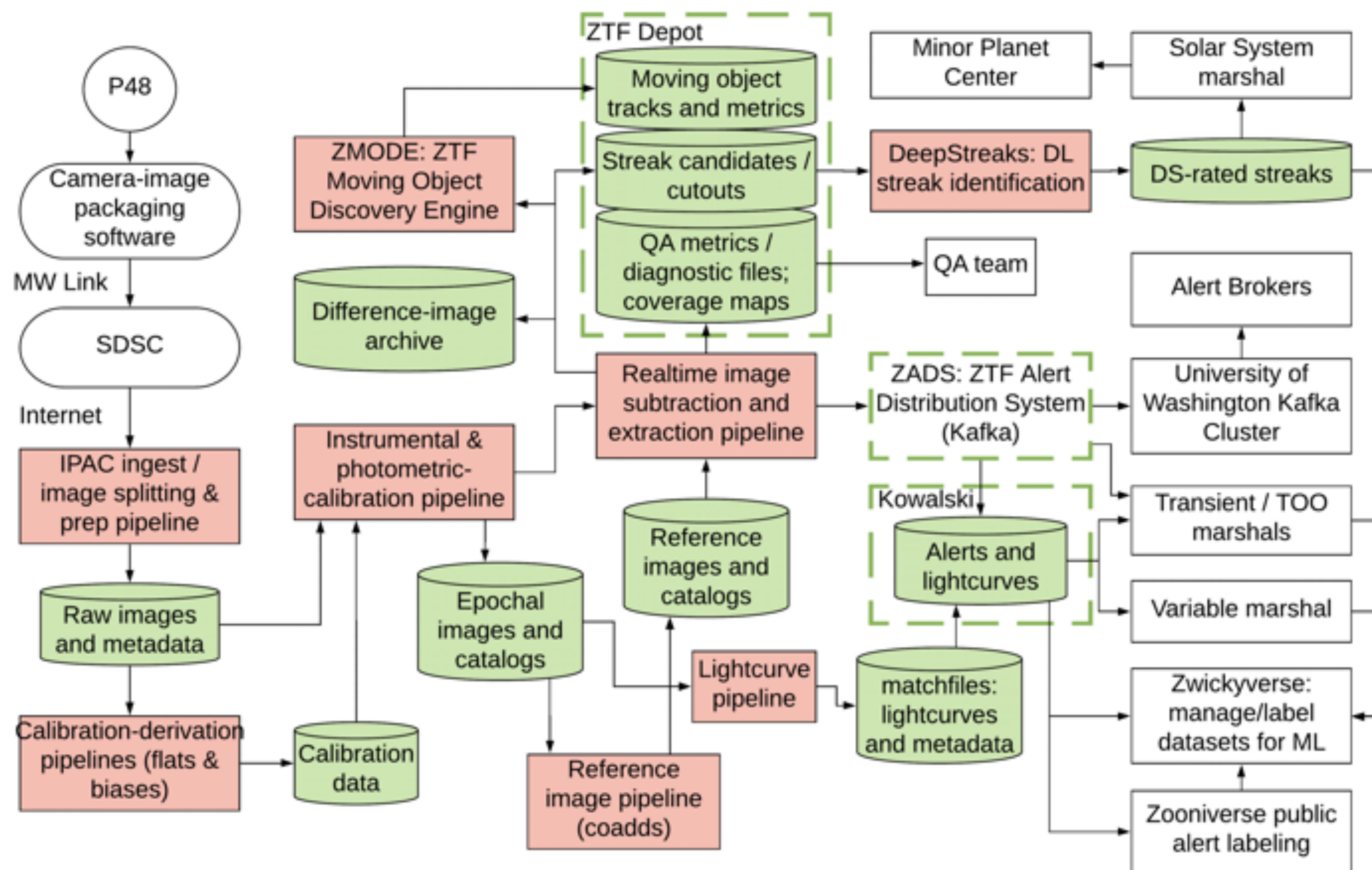
* A Patrick Brady special

SkyPortal: Schematic overview

Observe → Mine/Discover → Study and Characterize → Profit!

Kowalski

SkyPortal

and friends

and friends

Community, External
facilities and services ++

ZTF-I: data/processing flow

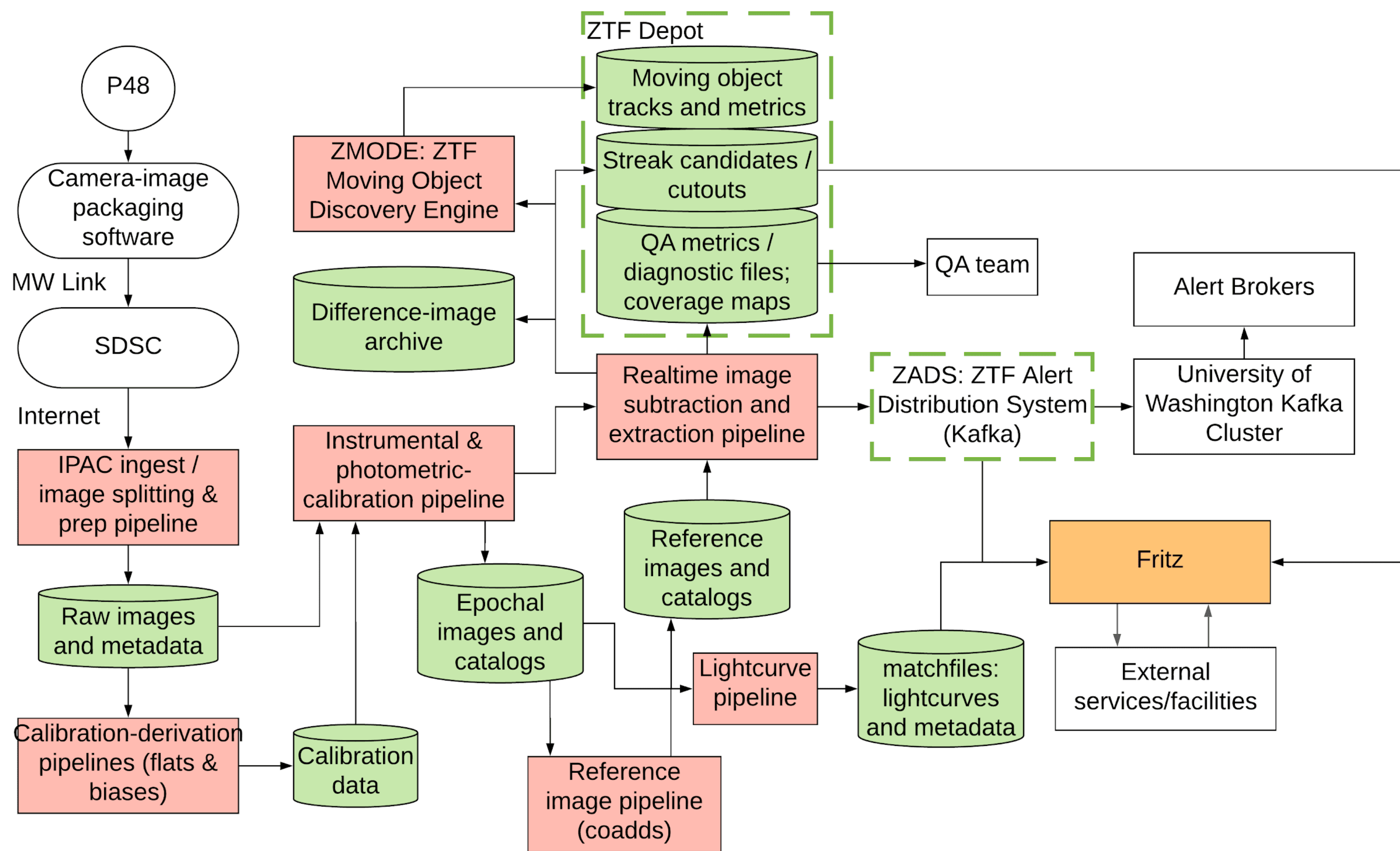ZTF acts as a discovery engine. Discoveries are followed-up using a wide range of instruments (including ZTF itself)

See Masci+ 2019

Single night	8h40m
# on-sky exposures	~700
Raw image data	~1 TB
Real-time data products	~4 TB
# unvetted 5σ alerts	~$10^5 - 10^6$
# ML-vetted alerts	~$10^3 - 10^5$
# unvetted streaks	~$10^4 - 10^6$
# ML-vetted streaks	~$10^2 - 10^3$

Nominal survey	3 x 260 n
Volume of data products	~3.2 PB
Volume of ref images	~60 TB
# CCD quad ref images	~2.8×10^5
Volume of matchfiles	~50 TB
# matchfiles	~2.8×10^5
# single-epoch PSF-fit source measurements	~800 B
# single-epoch aperture source measurements	~230 B

ZTF-II: data/processing flow

ZTF acts as a discovery engine. Discoveries are followed-up using a wide range of instruments (including ZTF itself)

Diagram labels:

- P48
- Camera-image packaging software
- MW Link
- SDSC
- Internet
- IPAC ingest / image splitting & prep pipeline
- Raw images and metadata
- Calibration-derivation pipelines (flats & biases)
- Calibration data
- Instrumental & photometric-calibration pipeline
- Epochal images and catalogs
- Reference image pipeline (coadds)
- ZMODE: ZTF Moving Object Discovery Engine
- Difference-image archive
- ZTF Depot
  - Moving object tracks and metrics
  - Streak candidates / cutouts
  - QA metrics / diagnostic files; coverage maps
- QA team
- Realtime image subtraction and extraction pipeline
- Reference images and catalogs
- Lightcurve pipeline
- matchfiles: lightcurves and metadata
- ZADS: ZTF Alert Distribution System (Kafka)
- University of Washington Kafka Cluster
- Alert Brokers
- Fritz
- External services/facilities

Single night	8h40m
# on-sky exposures	~700
Raw image data	~1 TB
Real-time data products	~4 TB
# unvetted 5σ alerts	~10^5 - 10^6
# ML-vetted alerts	~10^3 - 10^5
# unvetted streaks	~10^4 - 10^6
# ML-vetted streaks	~10^2 - 10^3

Nominal survey	3 x 260 n
Volume of data products	~3.2 PB
Volume of ref images	~60 TB
# CCD quad ref images	~2.8×10^5
Volume of matchfiles	~50 TB
# matchfiles	~2.8×10^5
# single-epoch PSF-fit source measurements	~800 B
# single-epoch aperture source measurements	~230 B

SkyPortal

- Open source (free to use, modify, and distribute)
- Scalable, API-first system, with fine-grained access control
- Multi-survey data archive and alert broker
- Interactive, mobile-friendly collaborative platform for transient, variable, and Solar system science cases
- Workhorse for ML applications: classification and labeling at scale
- Follow-up observation management: robotic and classical facilities
- Well-tested, extensive docs, CI/CD

Initiated in Feb 2020
Beta up in Sep 2020
MVP live in Nov 2020
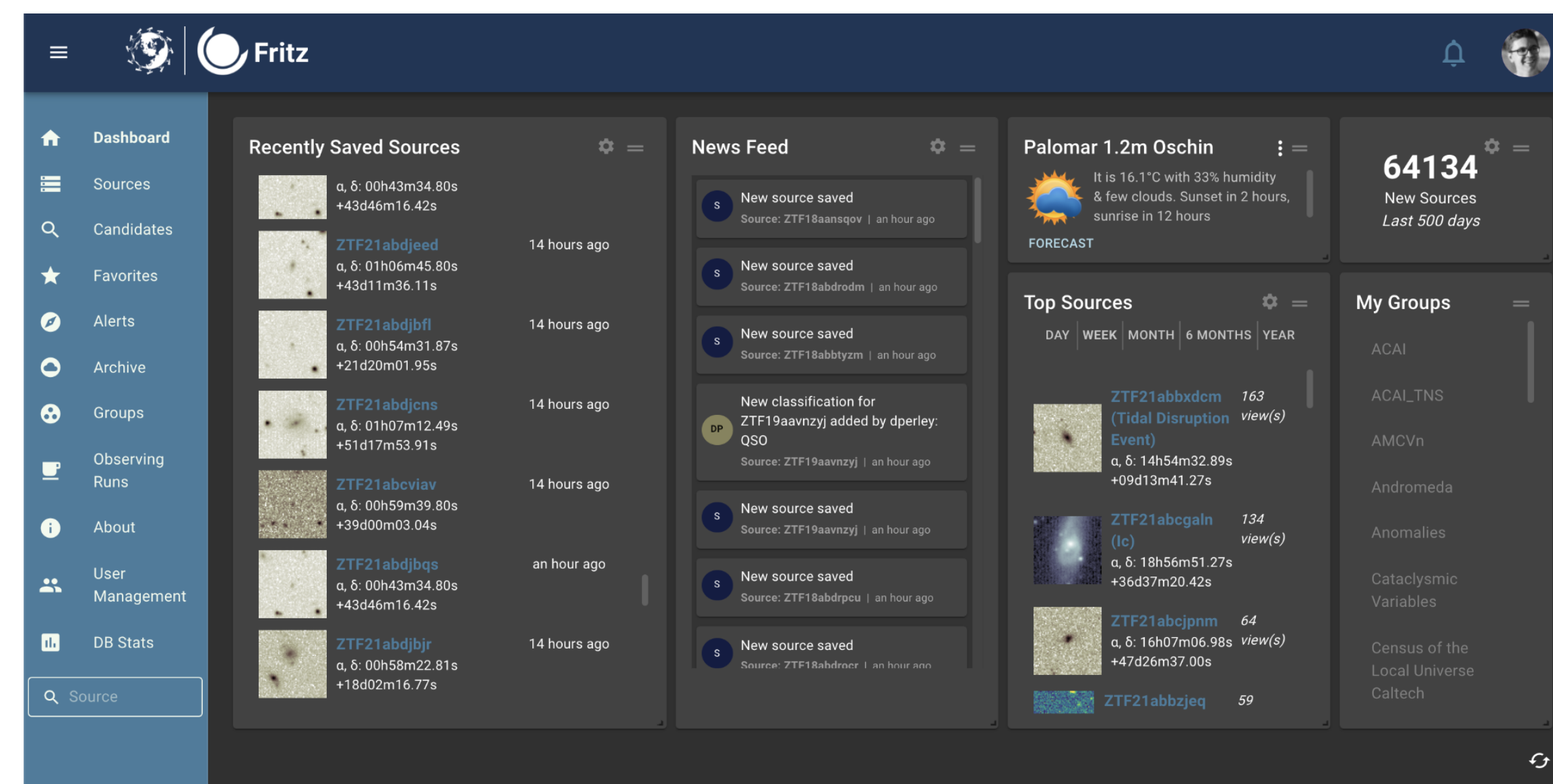
Roadmap
(In brief)

Stability/Scaling

MMA (04)

3rd Party Analysis

Stability / Scaling
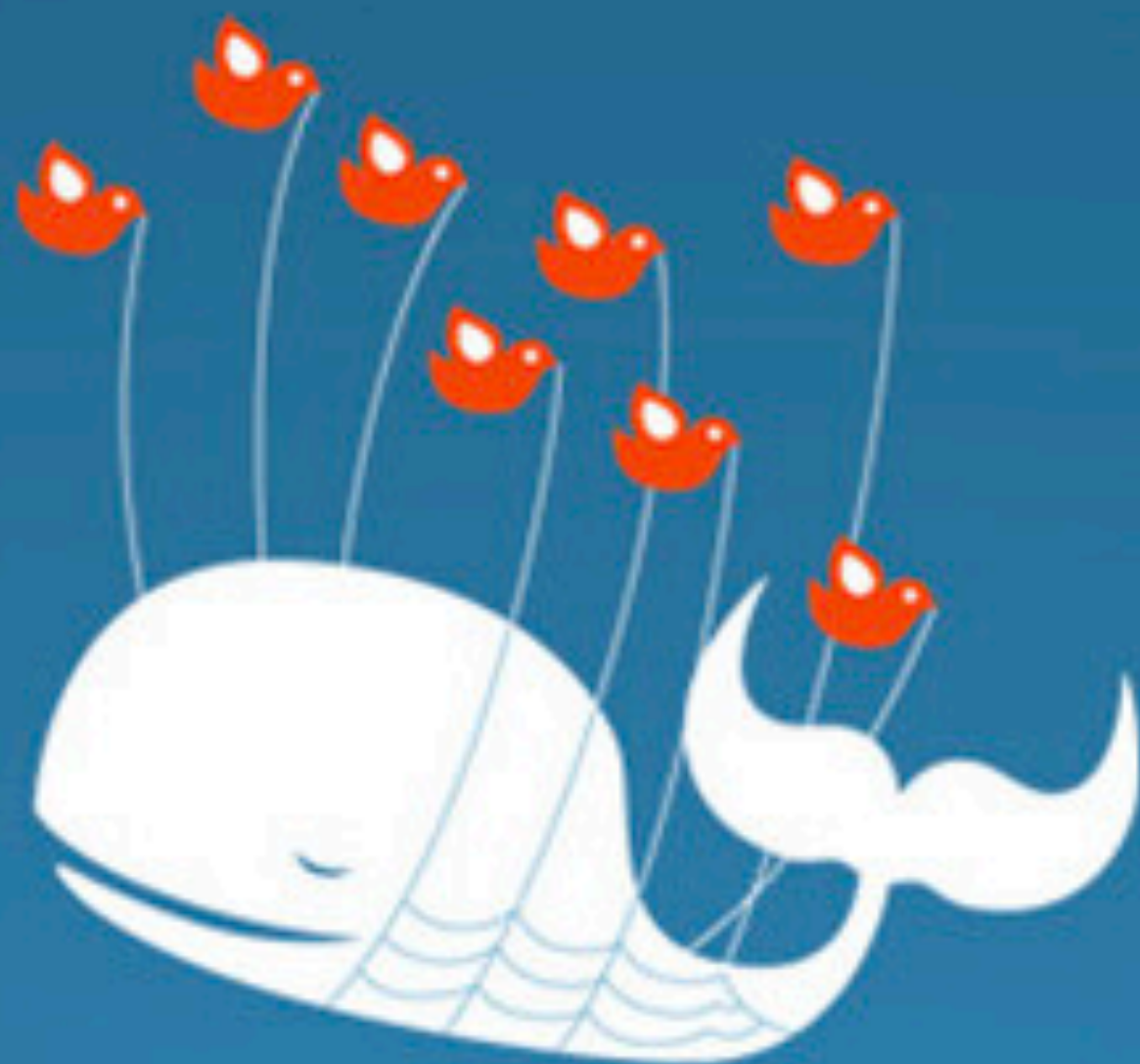
Twitter is over capacity.

Please wait a moment and try again. For more information, check out **Twitter Status.**

"Debugging performance issues is really hard. But it's not hard due to a lack of data; in fact, the difficulty arises because there is too much data…It's easy to confuse causes with symptoms"

Building the plane while we fly it…
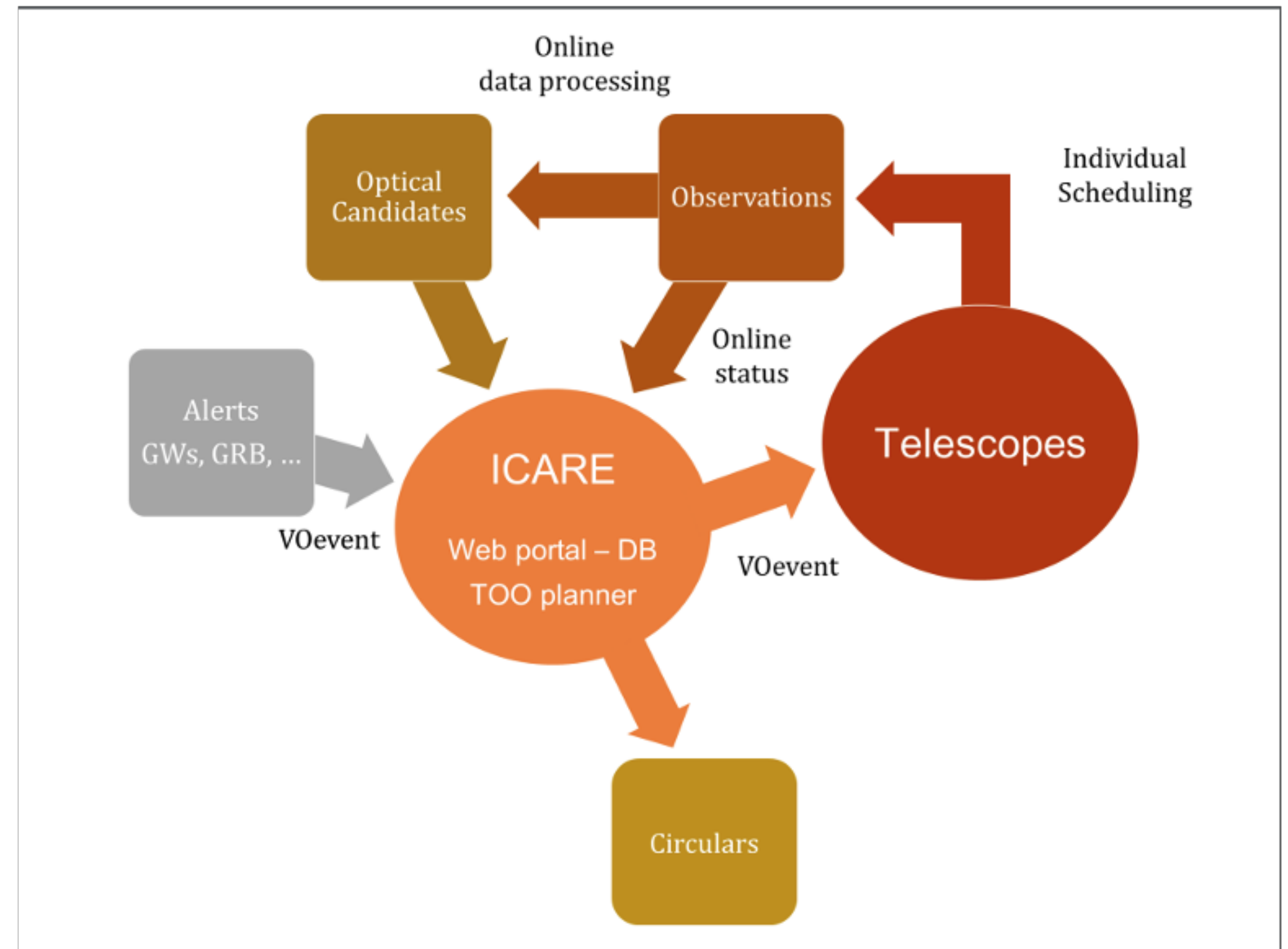
…and (trying to) squash bugs big and small.

MMA Infrastructure for O4

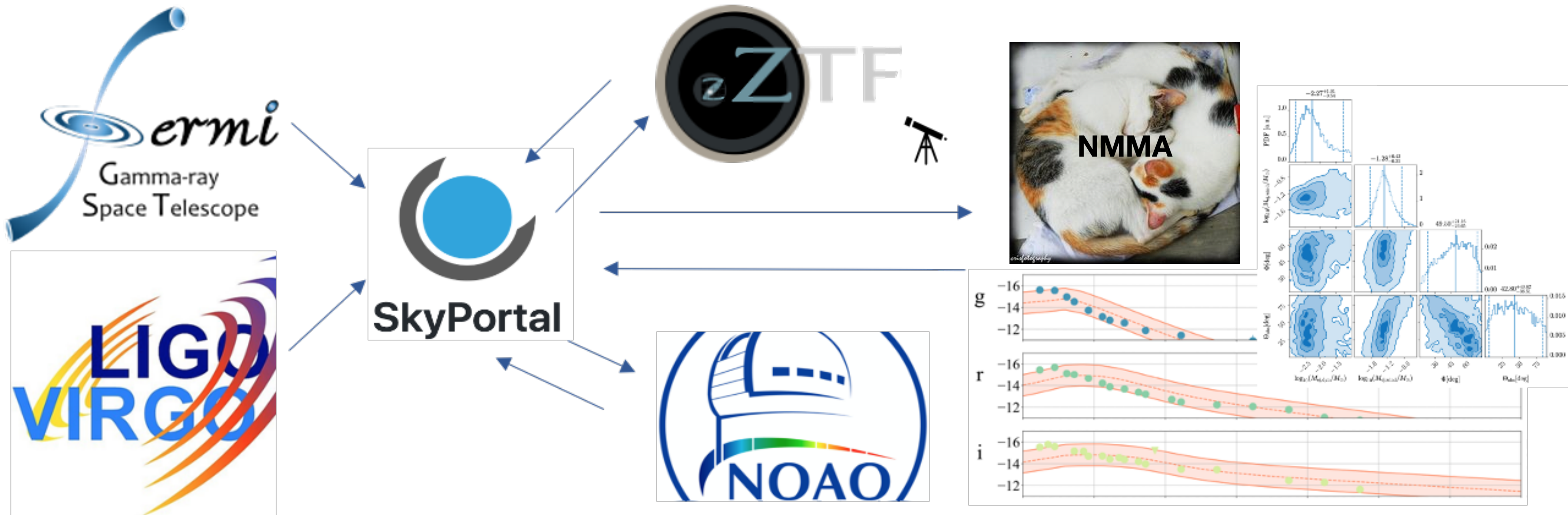# Inherits from two very successful projects during O3 and beyond: GRANDMA's iCARE and GROWTH's ToO Marshal

Ahumada et al. 2105.05067, Anand and Coughlin et al. 2009.07210, Andreoni and Goldstein et al. 1910.13409, Antier et al: 1910.11261, 2004.04277, Coughlin et al.: 1907.12645, etc.

A vision for the O4 workflow

Goal: Inform follow-up decisions

Goal: Remove difference between code we use to vet candidates in low latency and code we use to do science

What do I hope is available by O4?

- Light curve fitting: fit to supernova, kilonova, GRB, etc. models with the click of a button (led by Berkeley, support from UMN / OCA for NMMA)
- Follow-up prioritization suggestion (led by Caltech/Ari with ORACLE code)
- Creation of space-time MOCs for ease of internal and external observation distribution (led by UWM)
- Improved scheduling with M4OPT (led by Goddard)
- Upperlimits/efficiency assessment: triggering of simsurvey/NIMBUS so we can quote sensible efficiencies in GCNs/papers (led by UWM, Stockholm, UMN and others). Are we tooled up for things like using our own survey limits as priors? GW170817 as a prior? etc?
- Shifts: fine-grained tooling to receive notifications of events / observations in time-zone appropriate ways (led by OCA/IJCLAB)
- Rapid/large-scale time-series interaction, i.e. data sets like from KPED or Chimera (led by Berkeley)

What do I need from you?

- **All:** Missing features in the to-do list that you remember / expect you will find essential for O4
- **Those with existing observing programs (especially on those with APIs accessible through SkyPortal):** Try to trigger your programs through the website and work with us to fix any bugs / missing features. I know it will probably fail the first time or two, but this will allow us to be set by the time we need all hands on deck.
- **Those with existing code you use to interact in cool ways with SkyPortal but are otherwise poorly/tersely documented:** Take a few minutes to add some text and code to the SkyPortal Markdown docs (https://github.com/skyportal/skyportal/tree/master/doc)
- **Those with existing code you use that should probably be enabled within SkyPortal:** Post an issue / talk to me about what it does / check the big board
- **Those with some hiring power:** Send us your undergrads / summer students willing to learn some JS / React.js, we are sorely in need of some front-end help.

What do I need from you?

- **All:** Missing features in the to-do list that you remember / expect you will find essential for O4
- **Those with existing observing programs (especially on those with APIs access** ... e website and wo... y fail the first time or ... on deck.
- **Those ...Portal but are otherw...** ...e text and code to ...ortal/tree/ master/...

- **Those with existing code you use that should probably be enabled within SkyPortal:** Post an issue / talk to me about what it does / check the big board
- **Those with some hiring power:** Send us your undergrads / summer students willing to learn some JS / React.js, we are sorely in need of some front-end help.
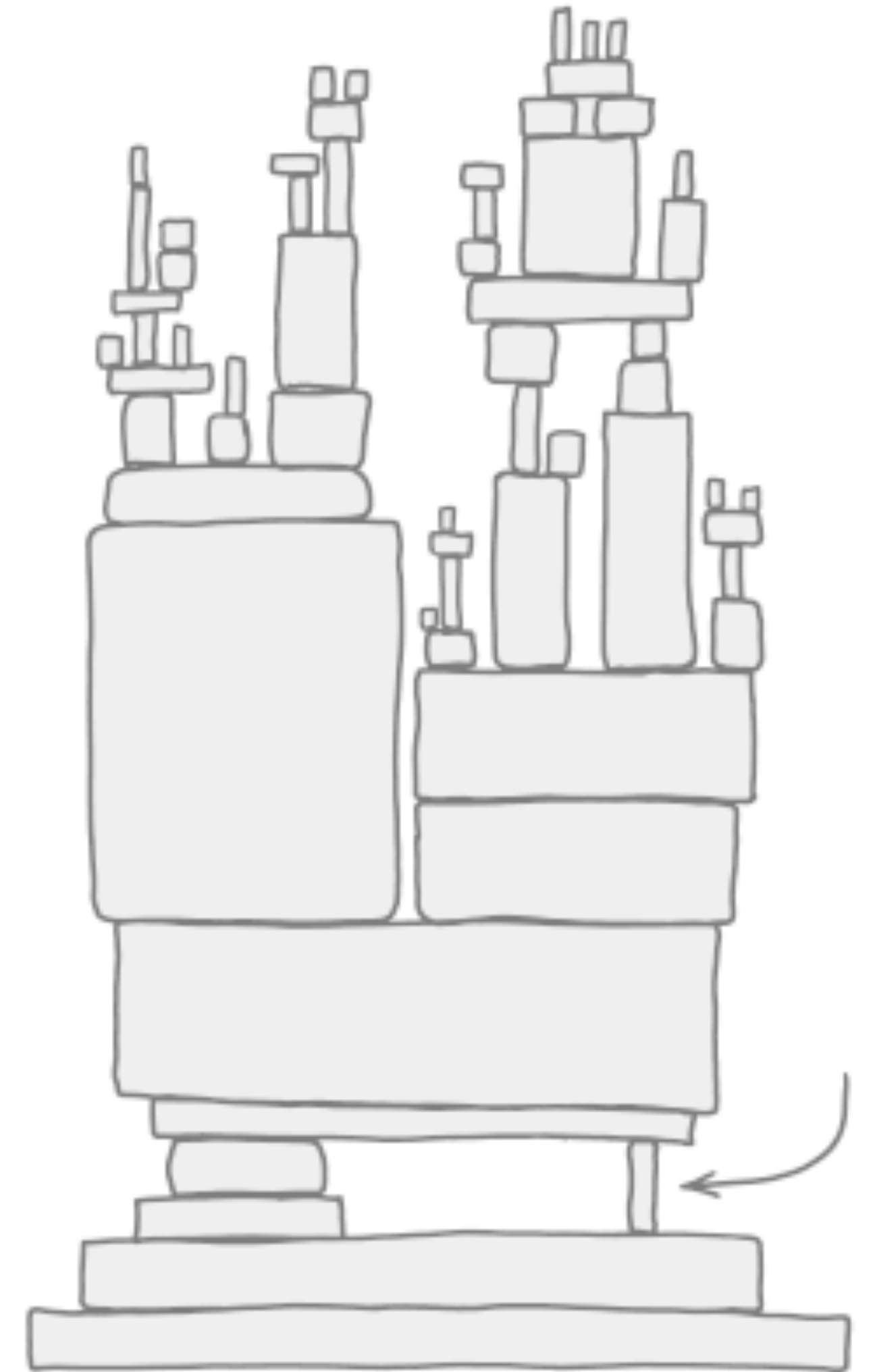
Experience so far…

Dev/Ops Personnel

Currently: 7 code contributors are mostly
those donating their time

⚠️ 🙏 Michael, Jada (UMN),
Guy, Josh, Stefan (Berkeley),
Theo, Thomas (OCA/GRANDMA)

We appreciate bug 🐞 reports but we *really*
need code contributors 👩‍💻

How to Engage

- Document Issues on the SkyPortal GitHub

- Contribute code via pull requests - we're happy to onboard you!

- Lots of (unique) learning opportunities in development and operations

Lesson Learned: open source is 💚

- Fritz/SkyPortal/Kowalski is open source
  - A huge part of devs is volunteer labor of love
  - While the core dev team is relatively small (in terms of effective person-hours), dozens of people have [contributed](#) meaningfully
- Leveraging what GitHub/OSS has to offer
  - Issues to track bugs and feature requests
  - PRs + thorough code review
  - GitHub Actions as the CI/CD platform
  - Don't argue about style, enforce pre-commit hook (black, flake8, eslint...)
- Project management tool: ZenHub
  - Should be as close as possible to GH

Lesson Learned: team and community is 💚

- Extraordinary individuals with a broad range of expertise
  - Respectful + supportive environment (that naturally extends to mutual admiration) allows the team to go far (check out this article for the spirit)
  - Critical but fair and open-minded review of ideas and code allows to iterate fast, converging on better solutions
  - Staying in sync: Slack + Daily 15-min stand-ups + weekly 1h meetings
- Constructive community feedback is essential for success
  - Beta-testers!
  - Enormously useful for finding/fixing bugs and implementing new features
  - Need to be clear about the communication channels: a dedicated Slack channel for smaller issues + GH issue templates for larger stuff
  - Critically important: prioritization + clear big picture for the project

François Chollet ✔ @fchollet · Apr 12 ···
Doing everything your users request is the shortest path to ruining your product -- especially when these users come from a legacy product and just want the new thing to be like the old thing
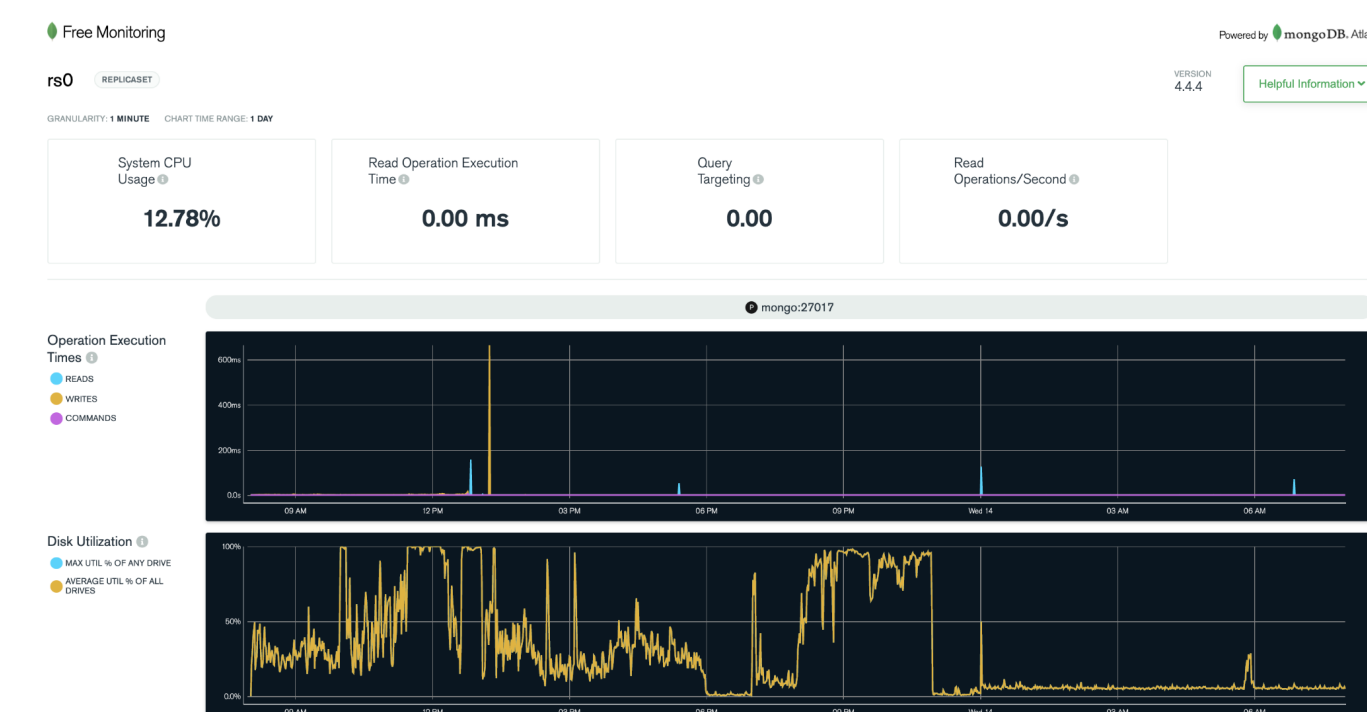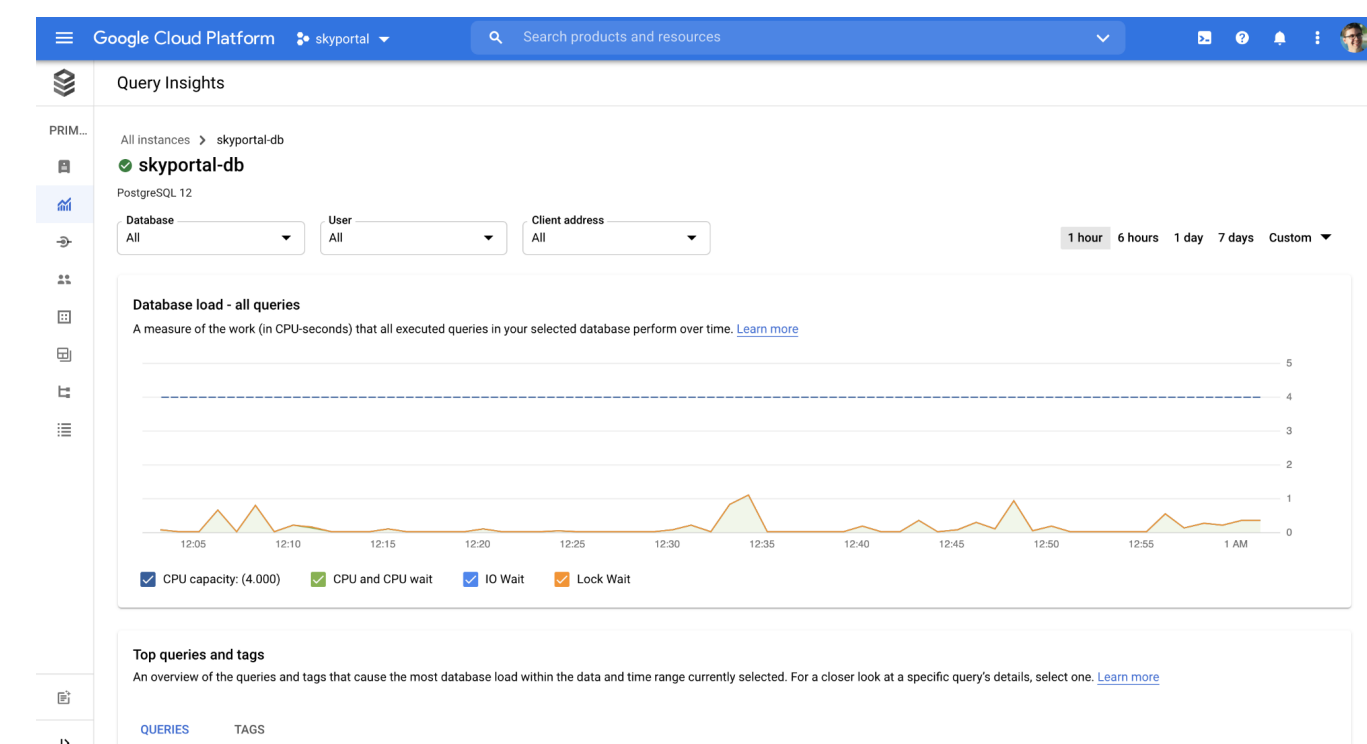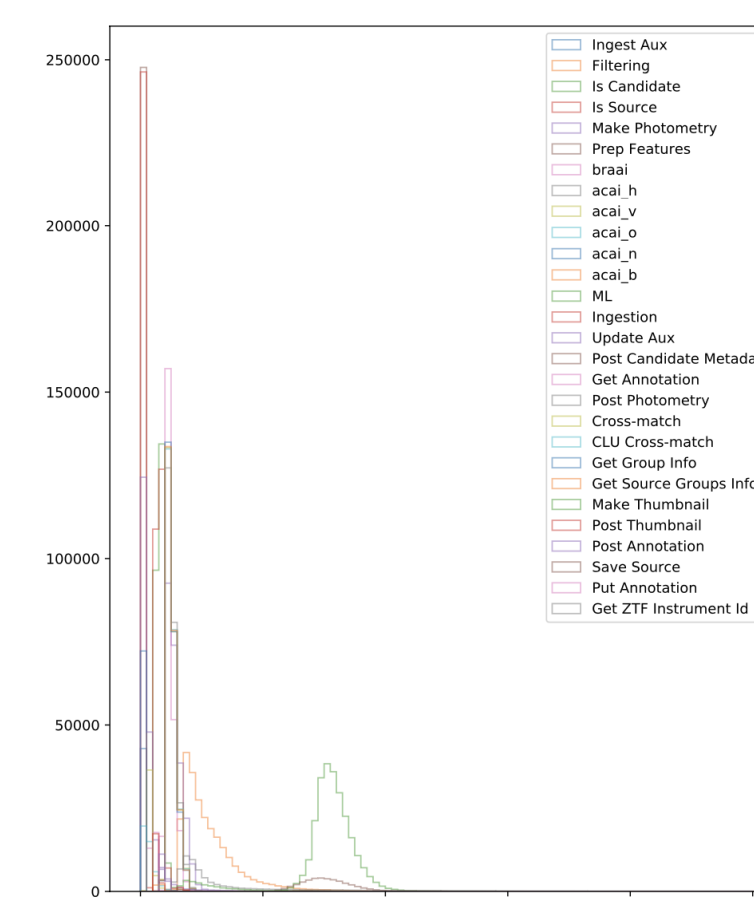
Lesson Learned: testing

- You are not testing your code enough
  - No, you're not!
  - From unit testing to integration testing through API and frontend, every bit is essential
  - Is that docker image still building from scratch? Note the word "continuous" in CI
  - Never underestimate the scale of a disaster that six innocently-looking lines of code can bring
  - Database migrations should be tested both ways - roll-backs are more common than we'd wish
  - Understand (and embrace!) flakiness
- Staging environment
  - Helps catch a lot of bugs before they have a chance to reach production, e.g. innocently-looking migrations that can take forever

Lesson Learned: production is really hard

- Even harder is to deliver updates/new features to prod
- Testing *is* your friend, but it won't catch everything that can happen
    - Running a subset of the test suite on a read-only replica of the prod db
- Weekly (at least) deployment to prod
- Resilient infrastructure for deployment
    - Monitoring the performance of the different components
    - Query Insights on the GCP
    - API endpoint response times, temporal evolution

Thank you!