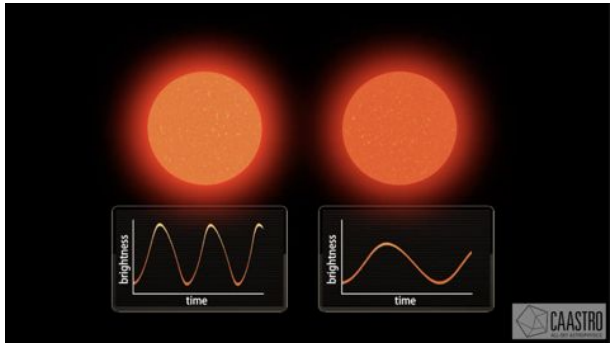
Fink, listening to the transient sky in the LSST era

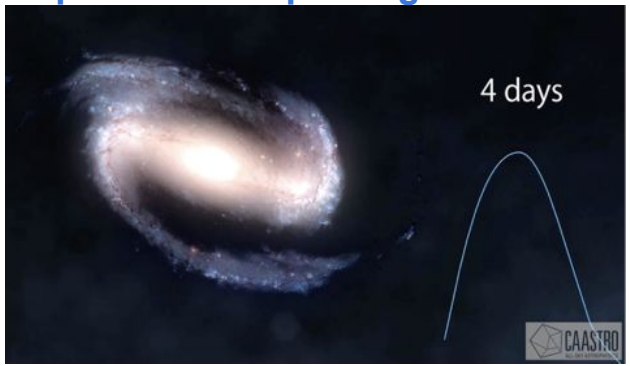
Julien Peloton
IT Department, IJCLab

The transient sky

Variable stars

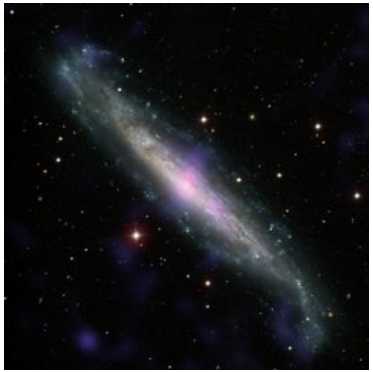


Supernovae: exploding stars



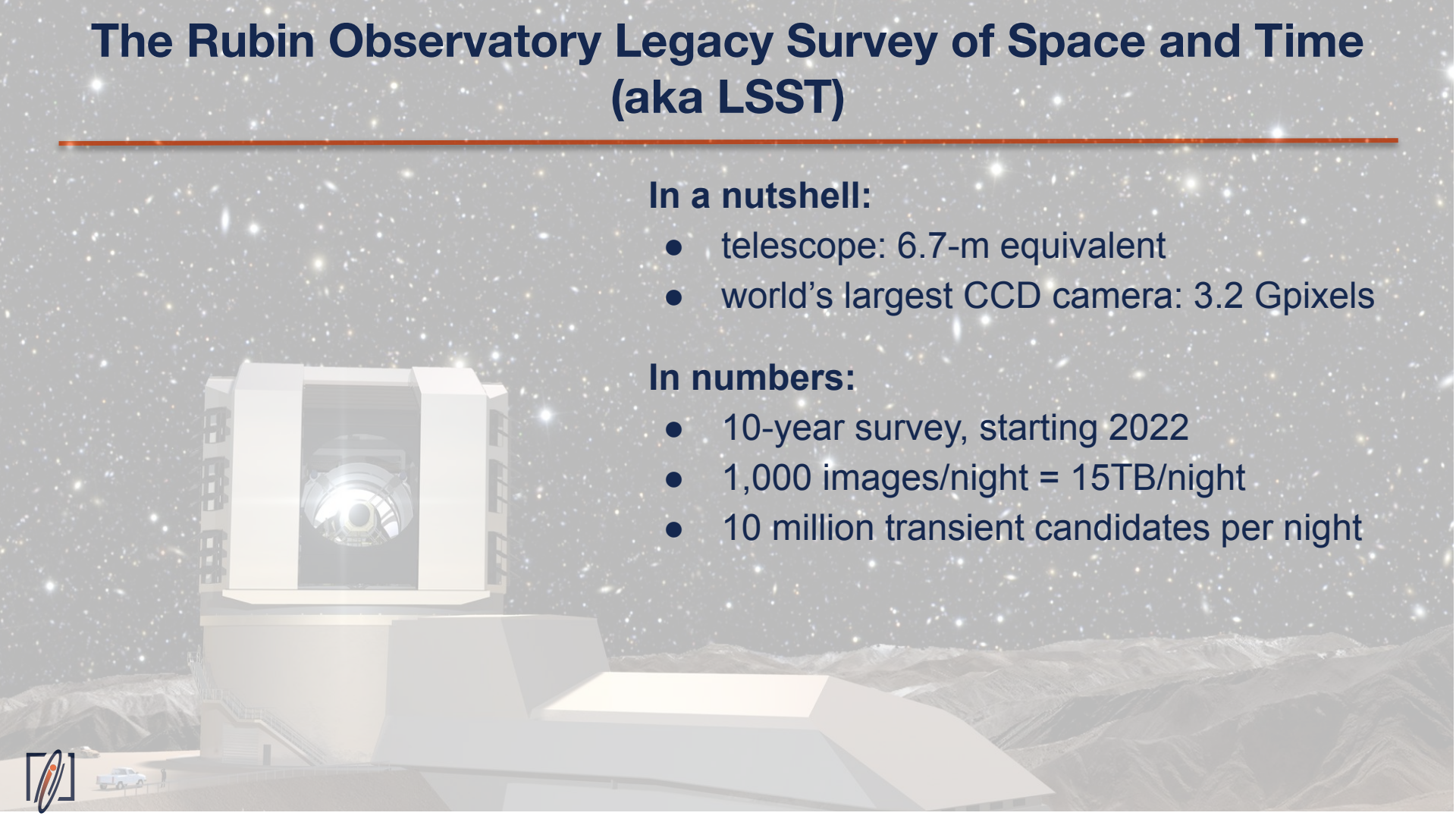
Neutron star mergers: kilonovae

Active Galactic Nuclei



+ RR Lyrae, novae, cataclysmic transients, tidal disruption events, asteroids, fast transients, calcium-rich transients, microlensing events, exoplanets transits… (the list is very long!)

The Rubin Observatory Legacy Survey of Space and Time (aka LSST)



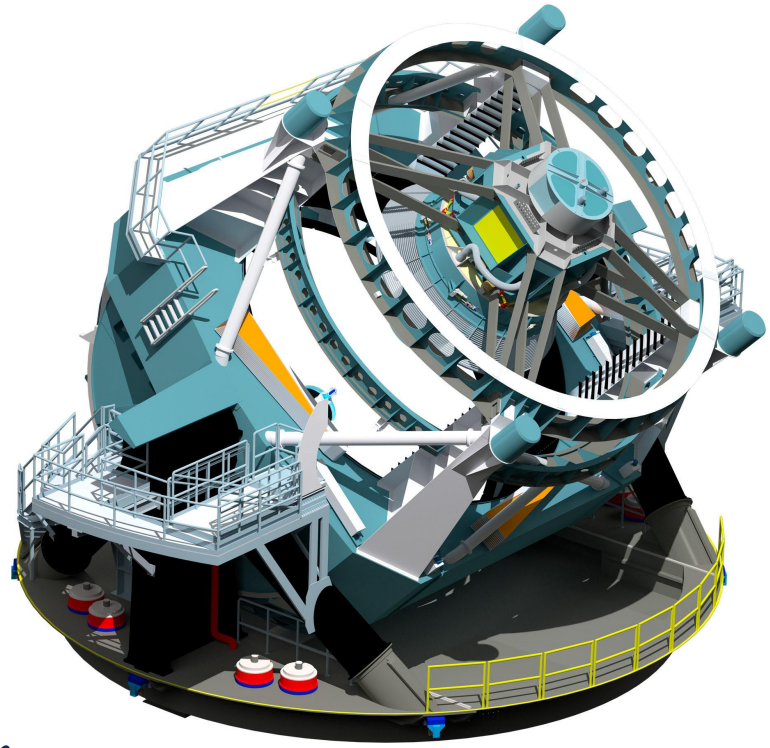
In a nutshell:

- telescope: 6.7-m equivalent
- world's largest CCD camera: 3.2 Gpixels

In numbers:

- 10-year survey, starting 2022
- 1,000 images/night = 15TB/night
- 10 million transient candidates per night

LSST data products



Now

Raw Data
Sequential 30s image, 20TB/night

60s

Prompt Data Product
Difference Image Analysis
Alerts: up to 10 million per night

Public data!

24h

Prompt Products DataBase
Images, Object and Source catalogs from DIA
Orbit catalog for ~6 million Solar System bodies

Year

Annual Data Release
Accessible via the LSST Science Platform &
LSST Data Access Centers.

End

Final 10yr Data Release
Images: 5.5 million x 3.2 Gpx
Catalog: 15PB, 37 billion objects

Alert data challenge

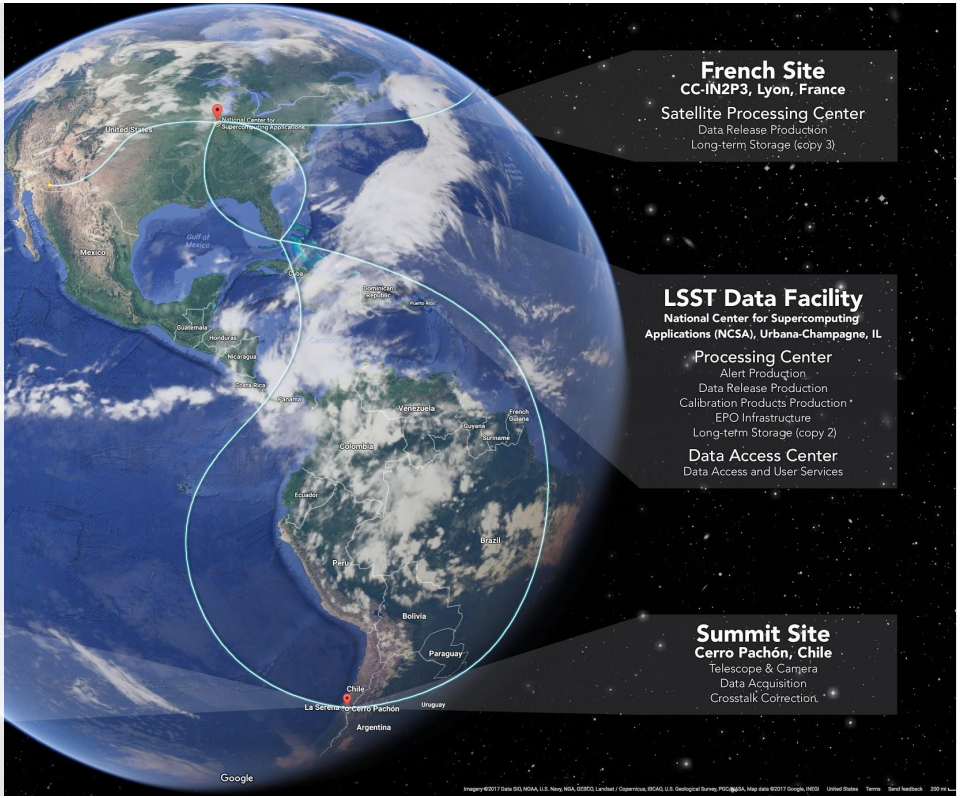
Forecasted: 10 million alerts per night…

- Current serialisation implies ~82KB/alert, 800 GB/night, 3PB in 2030.

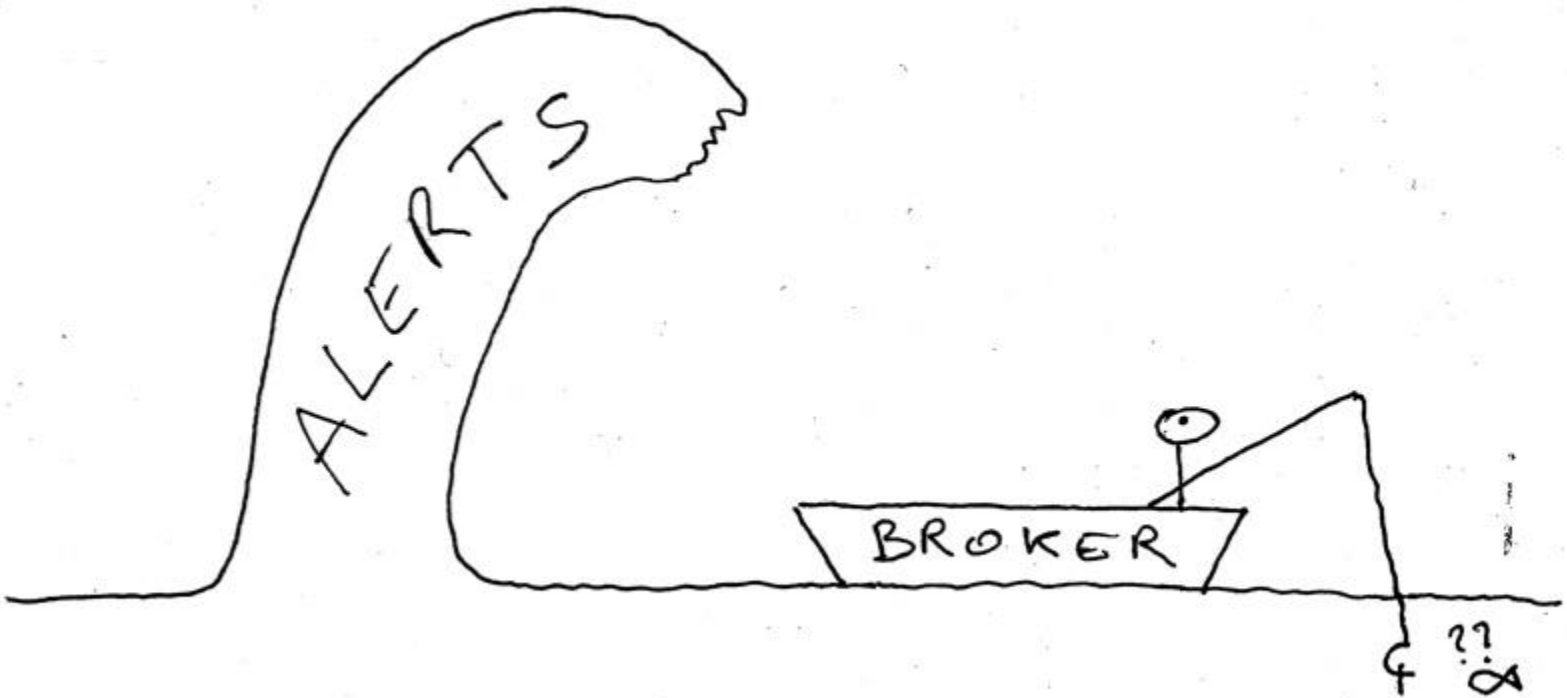
98% of alerts must be transmitted with 60 seconds of readout...

- … and processed before the next night!

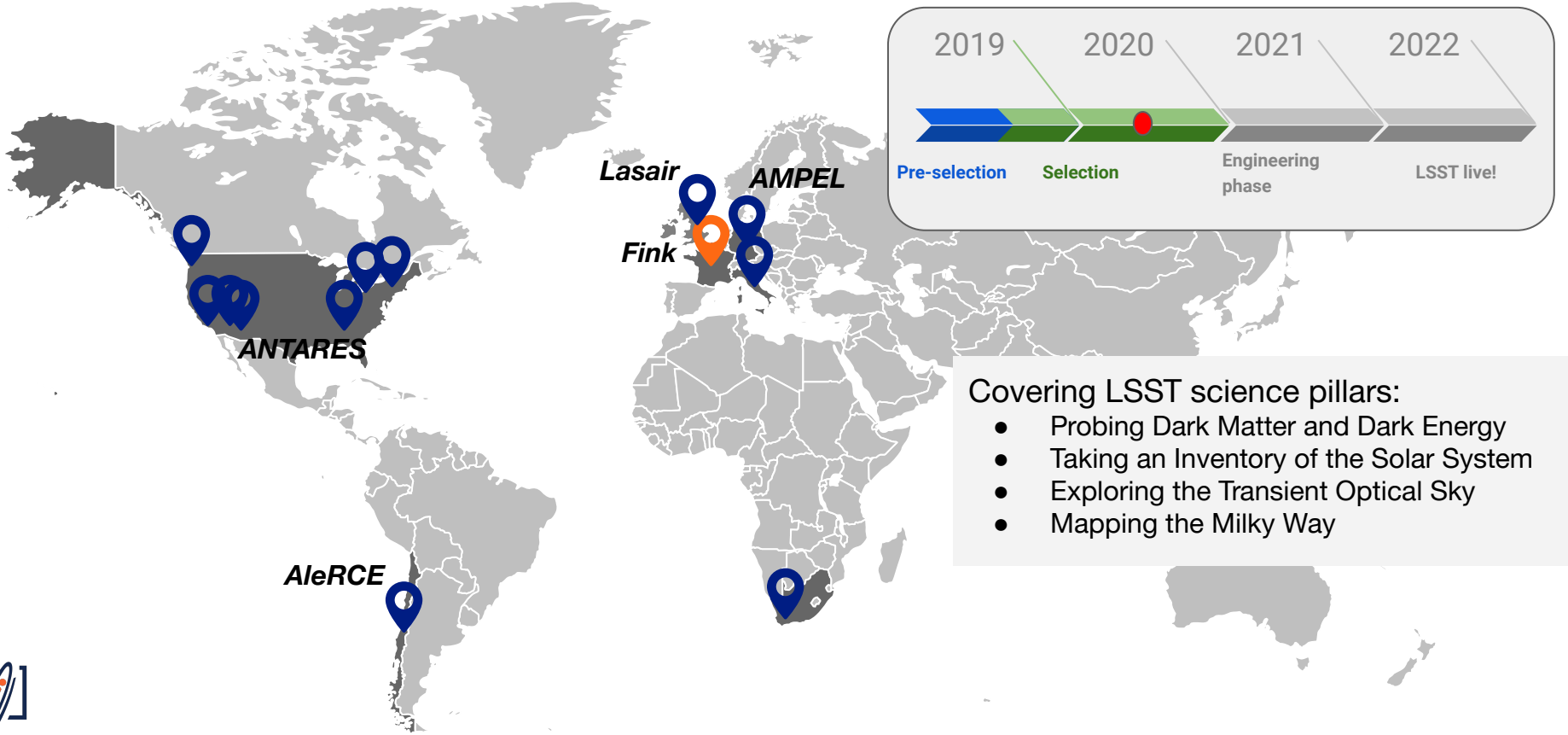
Wires to send alerts worldwide are not infinitely big...



Concretely



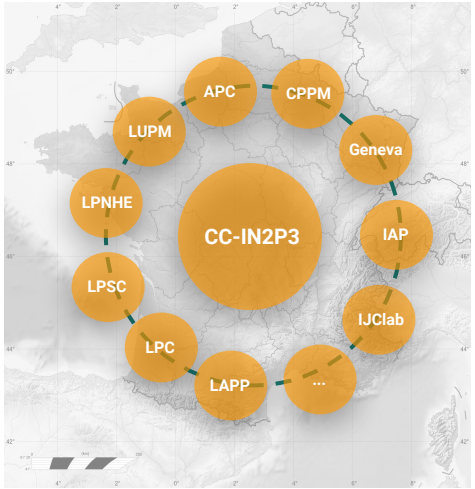
Broker landscape (2020)



Covering LSST science pillars:

- Probing Dark Matter and Dark Energy
- Taking an Inventory of the Solar System
- Exploring the Transient Optical Sky
- Mapping the Milky Way

Fink design

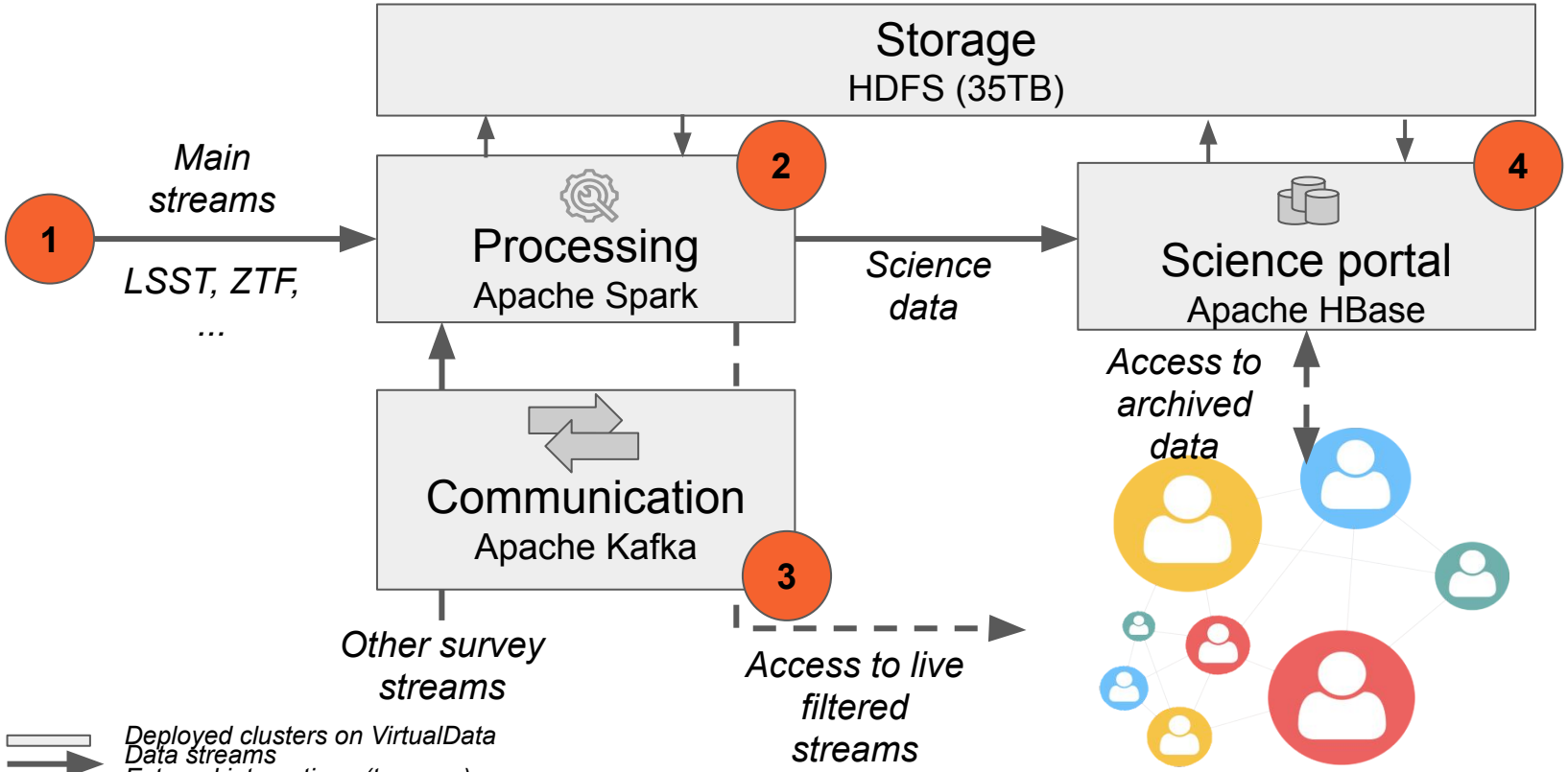


- 10 million candidates per night is about 1TB/night
- **Problem:** traditional broker tools fail at this scale
- **A solution:** distributing the load
  - Distributed computation (Apache Spark)
  - Distributed streaming (Apache Kafka)
  - Distributed database (Apache HBase)
- Fink: exploring the big data ecosystem, based on cloud infrastructures.
  - R&D started at IJCLab some years ago by your beloved engineers
  - Initial LOI signed by ~30 scientists (10 laboratories)

Working in the cloud

Cloud VirtualData (OpenStack) at University Paris-Saclay (~4,500 cores, 500TB).
Whole ecosystem of tools:

Fink in the cloud

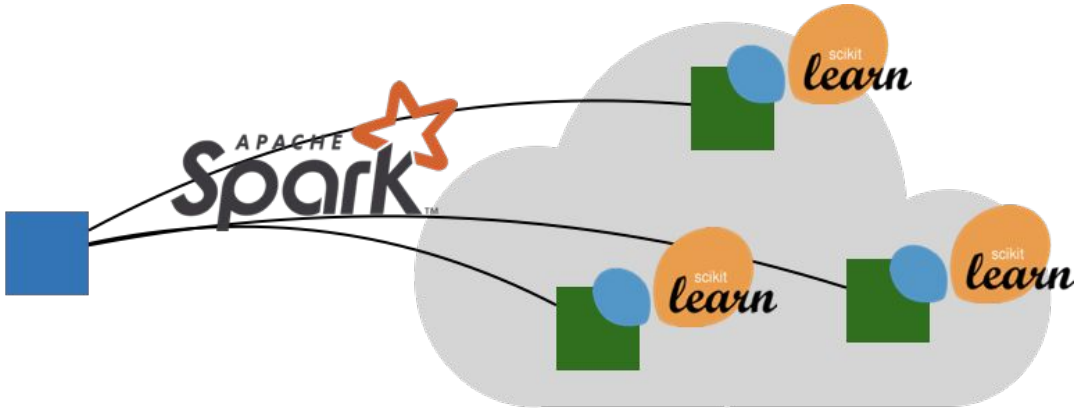


Interfacing user codes

Problem: Traditional astronomy codes are scientifically relevant, but they usually poorly scale: single machine architecture, lot of I/O, …

In Fink, we investigate how to best port legacy codes to scalable infrastructure

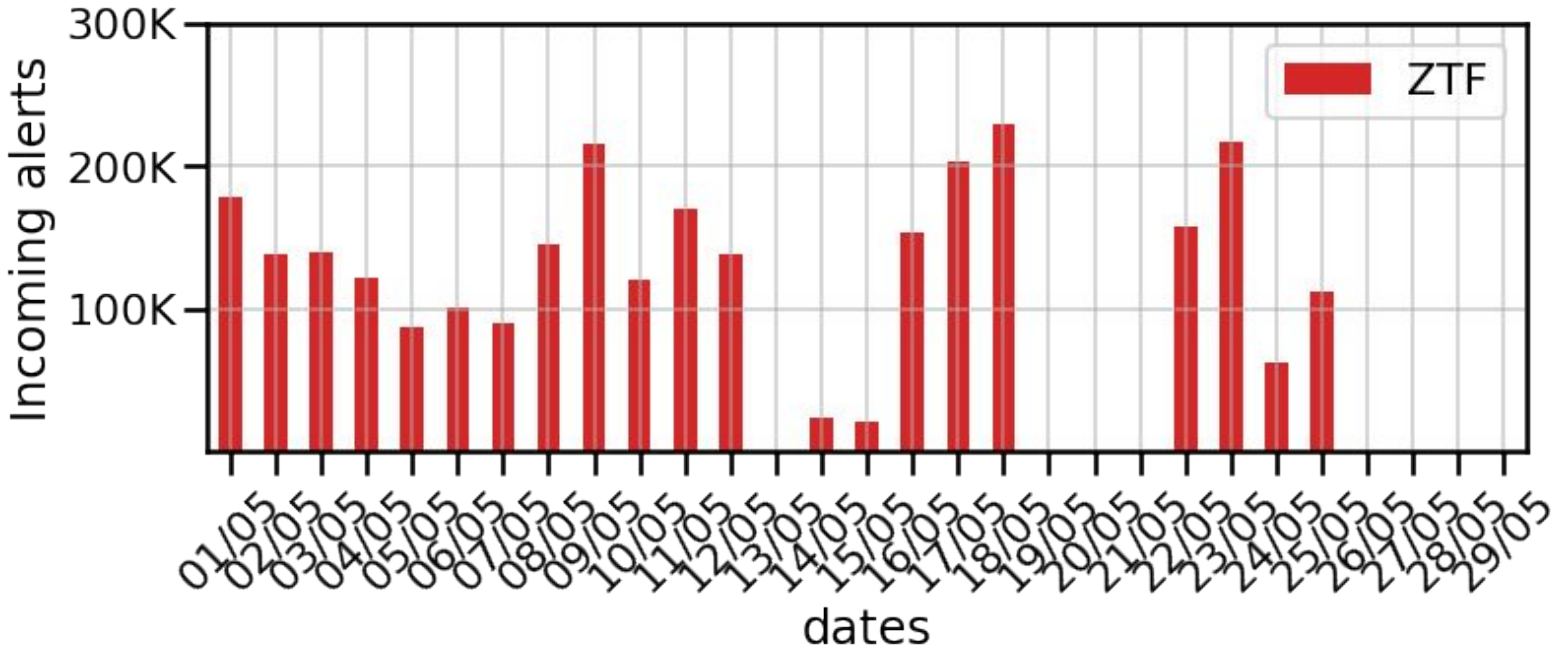
- Bridge Apache Spark with astropy/scikit learn/pytorch/…
- Or rewrite in native Spark (Scala or Python)



Processing ZTF data

We can already test Fink on real alert data

- MoU with Zwicky Transient Facility (ZTF), “pathfinder” for LSST.
- ~100,000 alerts per night (~10GB/night)

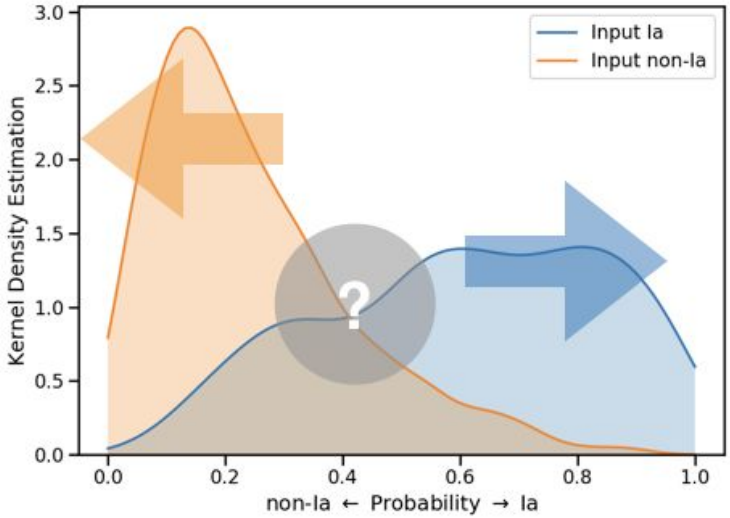


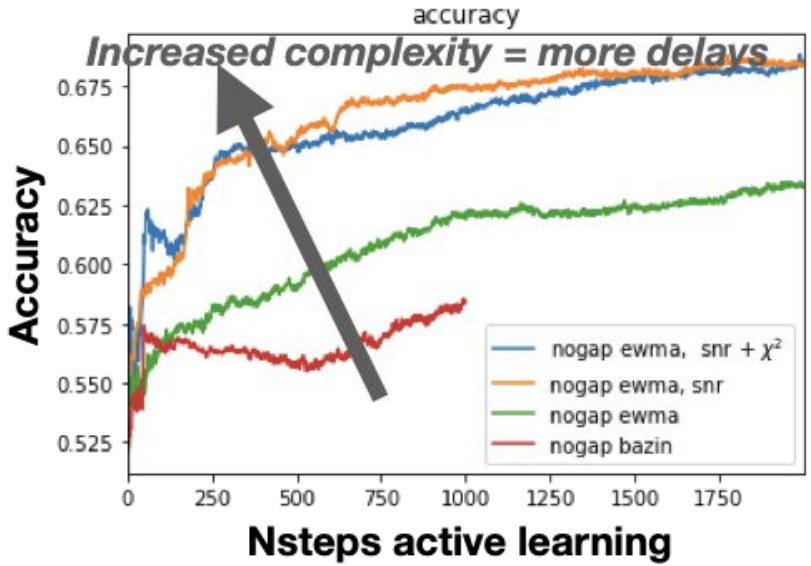
M. Leoni, E. Ishida, A. Möller

Active Learning for SN Ia

SN type Ia classification based on Random Forest classifier (Ishida et al 2019)

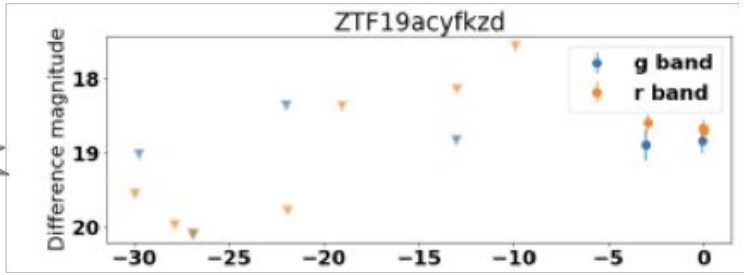
- Interfacing Apache Spark with scikit-learn
- Using Active Learning (POC)

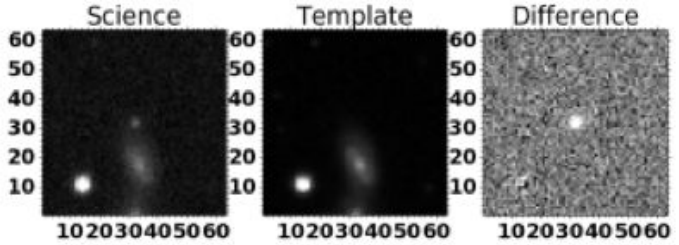




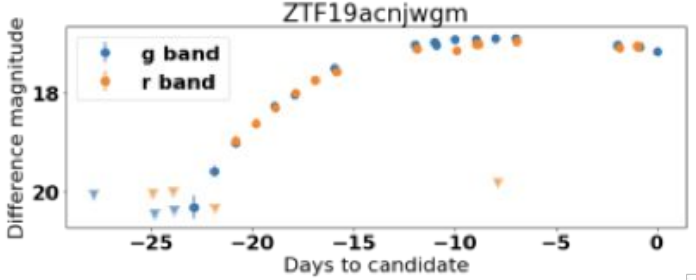
First SNe in Fink streams

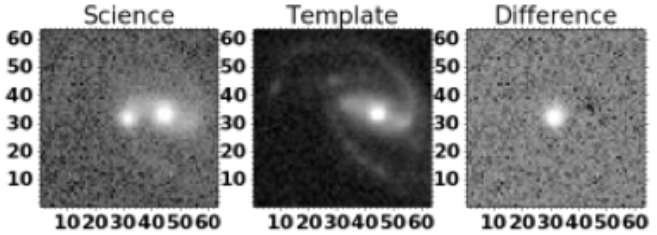
SN Ia



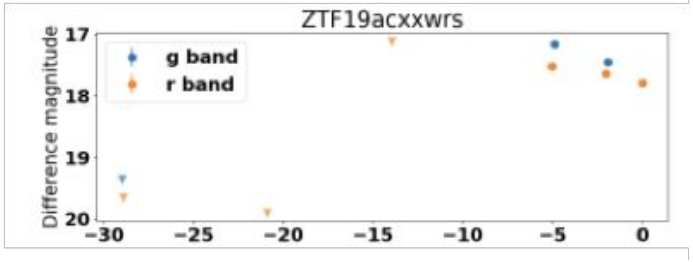


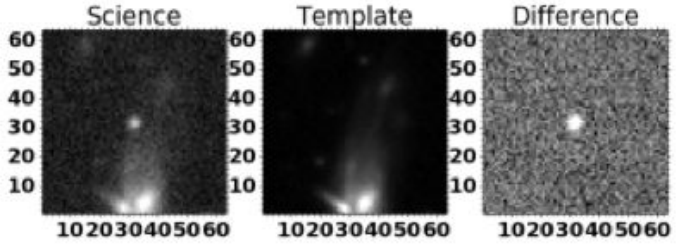
SN Ia





SN II-pec





Bayesian Neural Net for SN

Supernova classification based on Bayesian Neural Network (Möller et al 2020)

Problem: Observing time is limited and precious!

Wish: Ideally, not only classification but estimate of errors.

Our solution: Bayesian Neural Networks (Möller et al 2019).

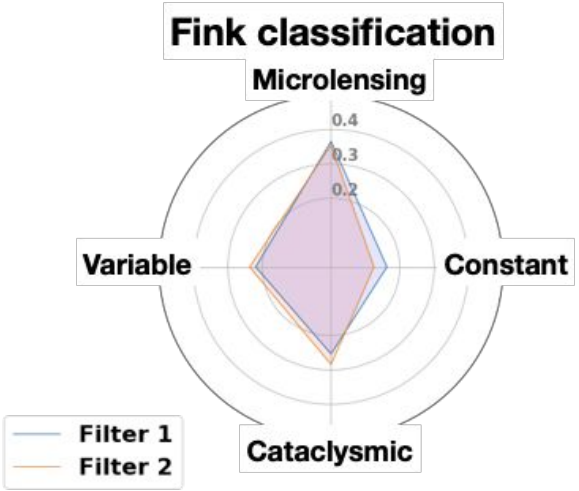
- Bayesian, convolutional, recurrent
- Tailored for big data
- Apache Spark interfaced with pytorch
- High throughput (2000 alerts/second on 100 cores)

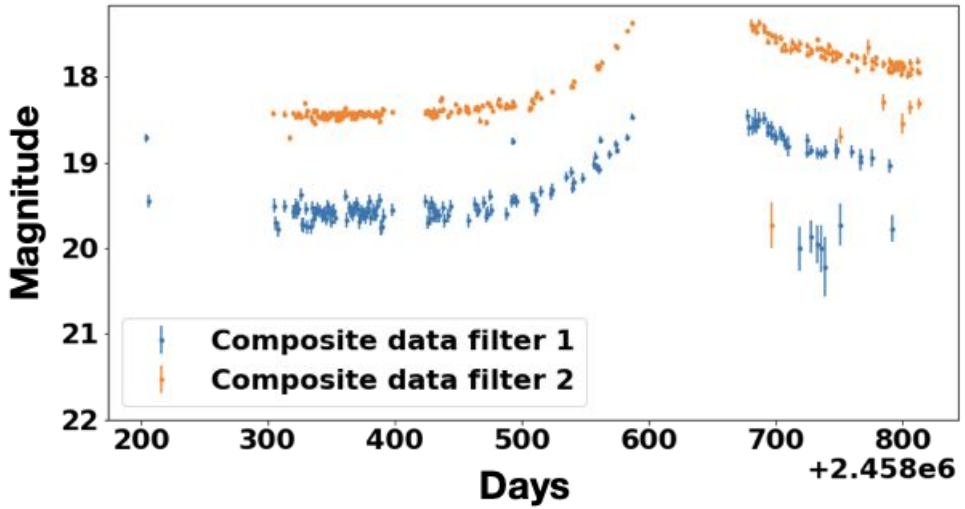
E. Bachelet, T. Blaineau, M. Moniez

Microlensing classification

Microlensing classification based on Random Forest classifier (Godines, Bachelet et al 2019): from exoplanets (hard) to black holes (still hard).

- Different timescales (days/months/year)
- Using PCA + Random Forest (Apache Spark + scikit-learn)



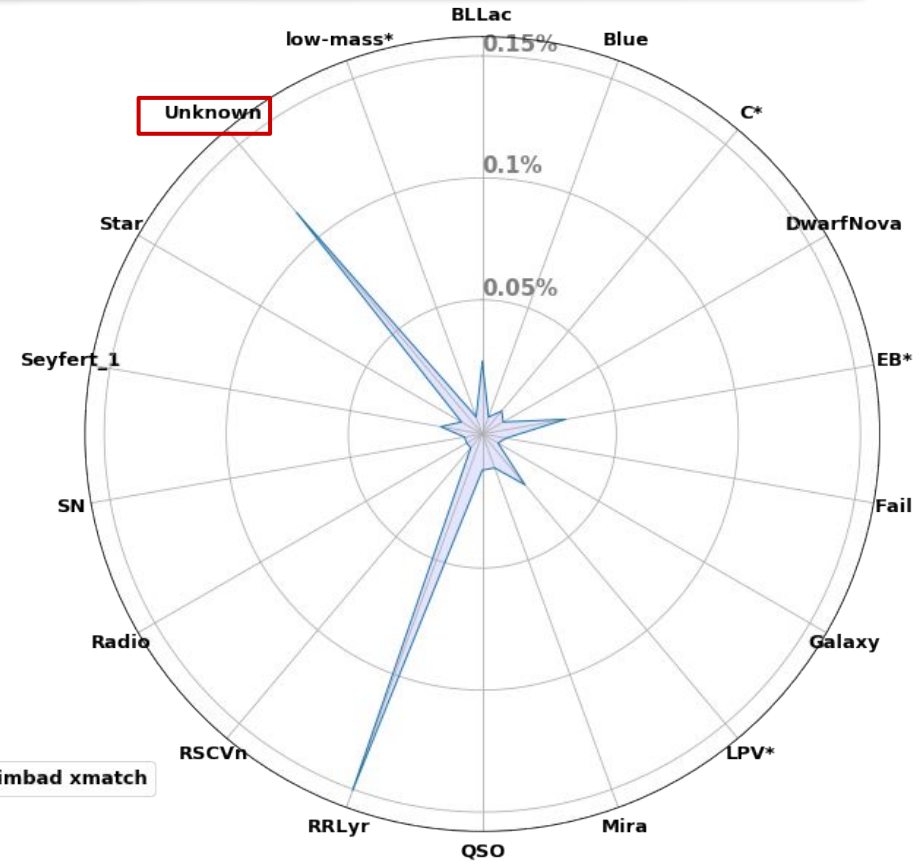


Large cross-match

Cross-match service

- **(Naive) question:** Are there already known objects in the stream?
- **Problem:** Standard cross-match is very slow! (we have 10,000 alerts / 30 seconds)
- **Our solution:** Distributed cross-match using xmatch service @ CDS Strasbourg.

Cross-match of thousands of objects per second.



Coordination

Identifying interesting LSST alerts is only part of the story: we need coordination with other facilities, follow-up resources and existing networks.

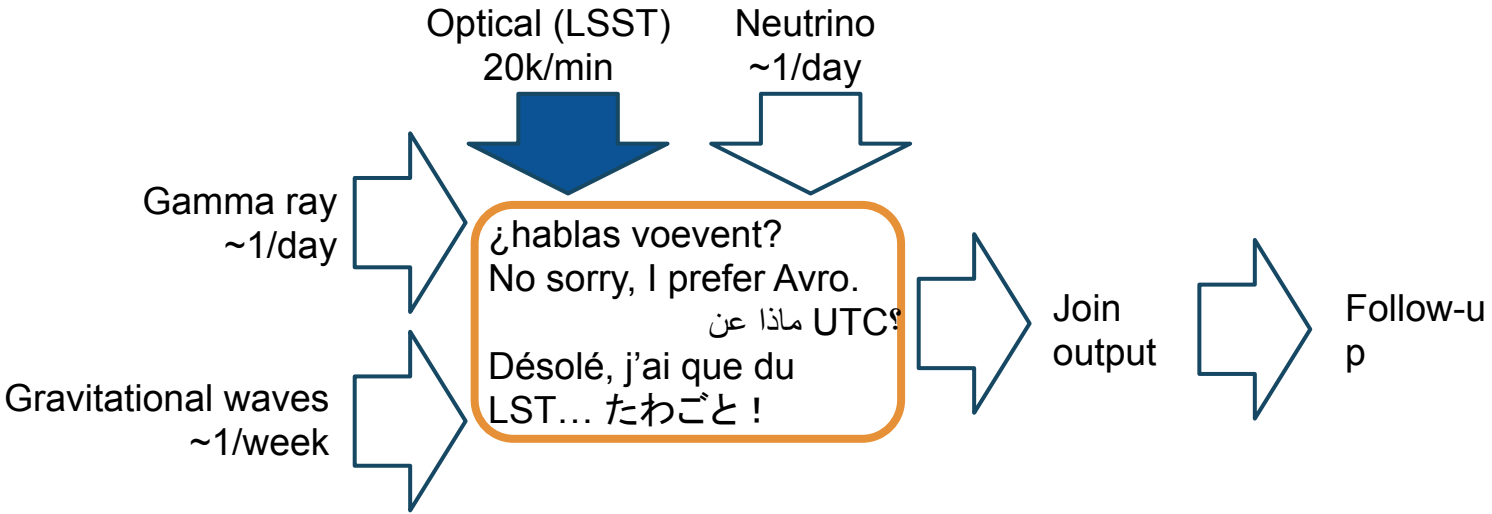
- Your expertise is important to us!
- Discussions and work with teams from: CTA, Integral, KM3NET, SVOM, ...

We will regularly publicize a prioritized list of targets for each science case that should be followed in order to improve future estimates.

- How to integrate this in the current landscape given the scale?
- How to coordinate with existing follow-up resources (ToO, TOM or TNS) and surveys?

Multi-messenger: the SVOM case

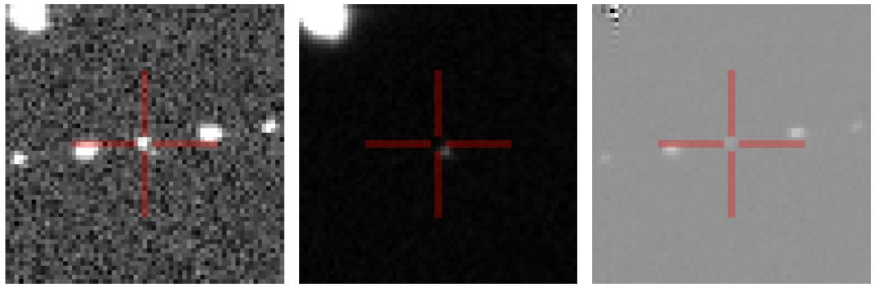
- **Goal**: Continuous cross-match of LSST with SVOM (both ways).
- **Problems:** How to quickly react without overwhelming the systems?
- **Currently**: Exploring ZTF with GRB surveys (SWIFT/Fermi) + simulations
- **Future:** Selection function using frequency extrapolation, synergies ground + space, joint scanning strategies...



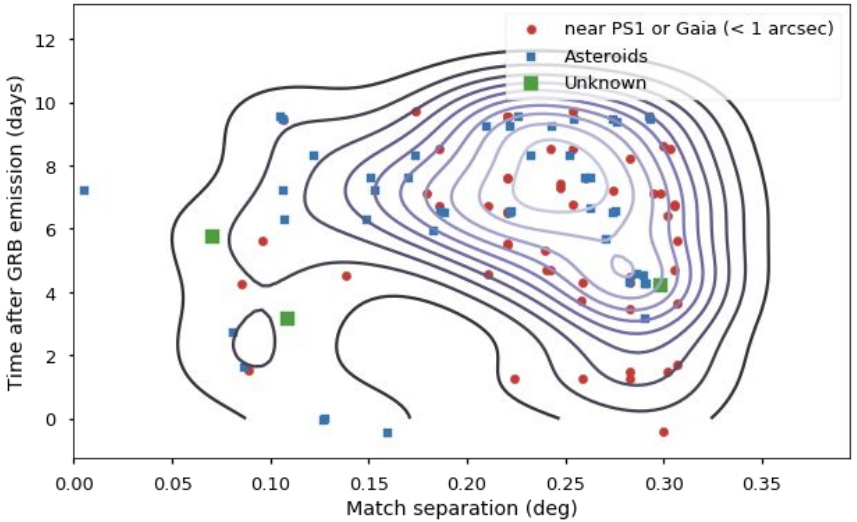
Solar system physics

A lot of solar system physics as well

- Many asteroids (or satellites…)!



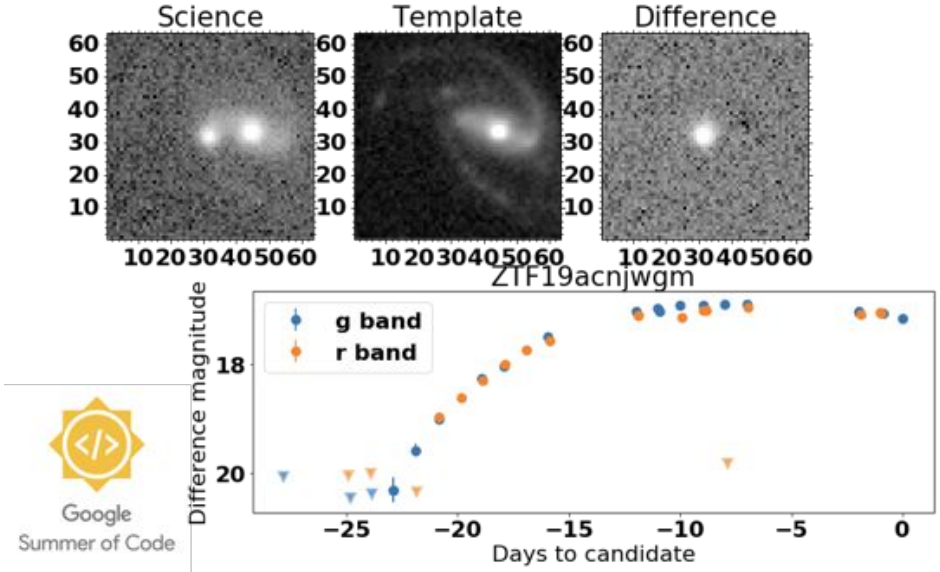
ZTF x Fermi (w/ D. Turpin)

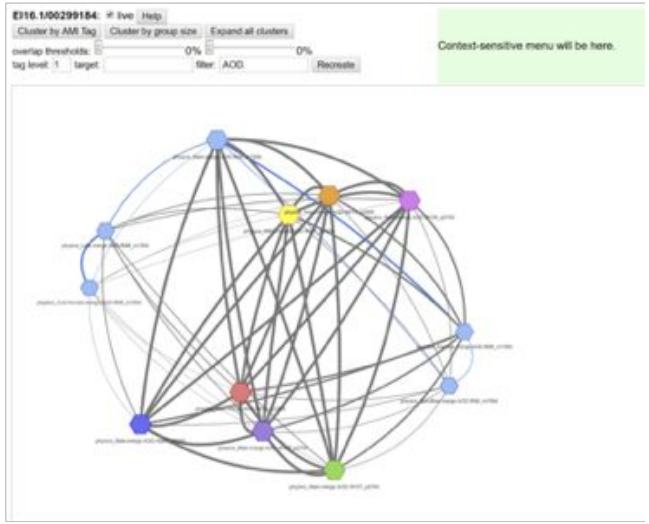


Accessing Fink data

Two entry points for users:

- Kafka streams (Apache Kafka cluster deployed)
- Science Portal (under construction): distributed database + graph solution





Google Summer of Code

Towards LSST

What will change with LSST?

- Volume of data x100
- Wider alert population
- Richer alert packets

Simulation tools for LSST deployed in the cloud (Kafka cluster)

- We can simulate up to 10x LSST rates!

Continuous R&D projects @ IJCLab, e.g.

- Improving storage layer to enforce data integrity (C. Arnault)
- Deployment using Kubernetes (S. Pateyron)
- Introducing Graph DB for visualising data at Petascale (J. Hrivnac)
- Distributed Machine Learning to classify objects faster than light (M. Leoni)

Take away

Fink is a broker designed to tackle LSST alert big data challenges

- Enabling science by applying state-of-the-art technology.

Technology Readiness Level (TRL) 6/9.

- Still under development (deadline: December 2020)
- Fink is already processing ZTF data stream (MoU 2020).
- First science modules deployed and testing capabilities beyond expectations: SNe, GRB, microlensing, ...

We need you!

- Full broker proposal (end 2020).
- More science cases & technology deployment to come


FINK
https://fink-broker.org
LSST Project/NSF/AURA