Machine learning for time-domain astronomy at scale

The example of Fink

Paris workshop on Bayesian Deep Learning for Cosmology and Time Domain Astrophysics 3rd ed.
20-23 May 2025

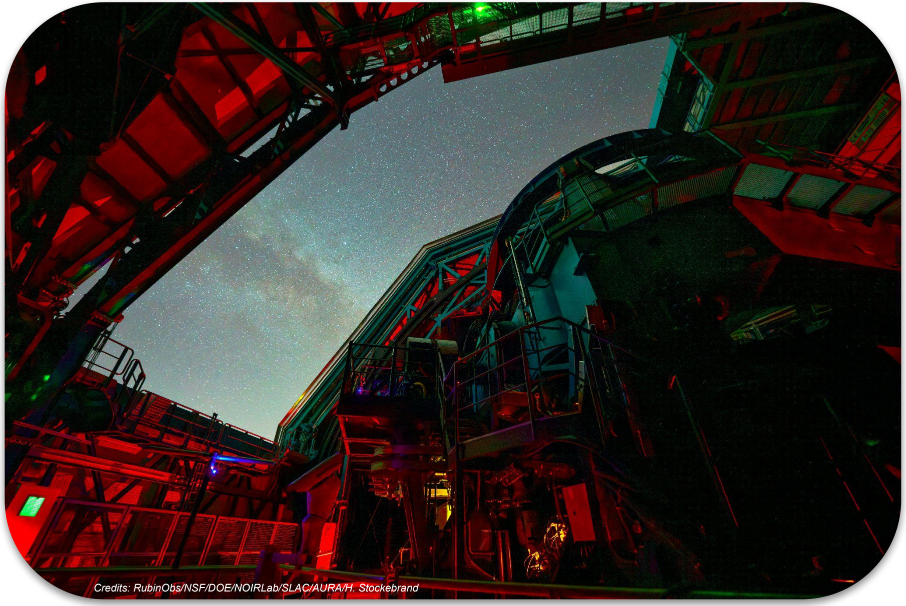
Julien Peloton
IJCLab, CNRS, Paris-Saclay

[Bigger], Better, Faster, Stronger

How did the solar system form? How do galaxies form? What is the history of the Universe? **How did we get here?**

As we build new telescopes, some questions are answered, but **more arise**

In terms of data volume and complexity, the emergence of **large-scale surveys** has been a game changer



Transient factories

Orders of magnitude

Field of View: 3.5 degrees

- 7 times the Moon!

Primary mirror diameter: 8.4 m

- 5 times average human size

Pixel camera count: 3.2 Gpixels

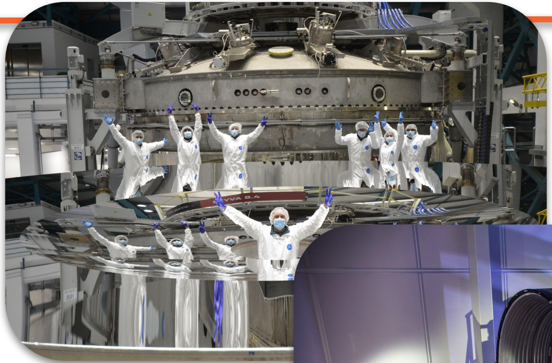
- 1000 times a smartphone

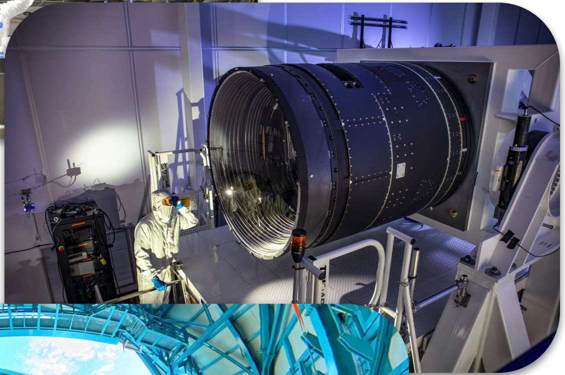
Nightly data size: 20TB/night

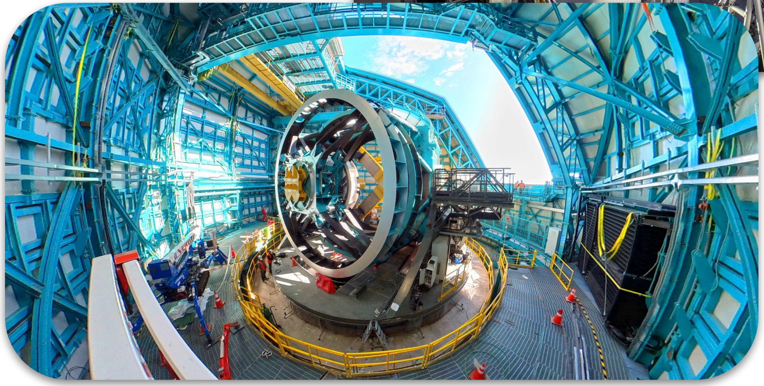
- 20 standard hard disks!

Nightly stream: 1TB/night

- 10,000,000 notifications on your smartphone







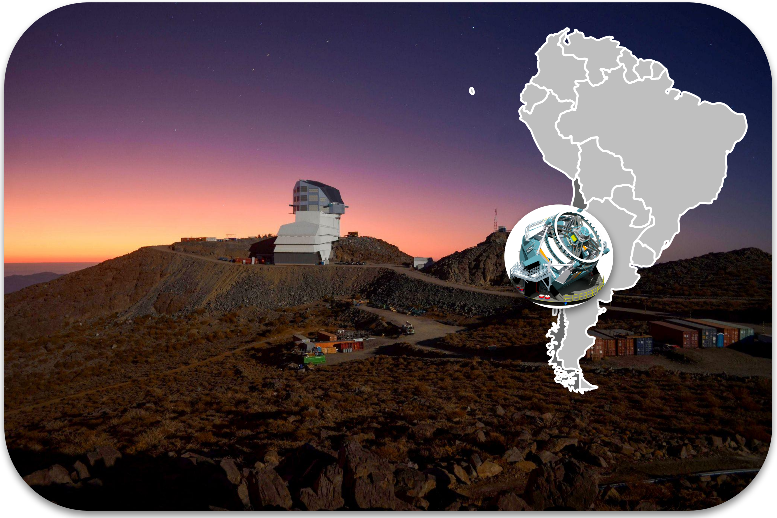
Rubin science goals

Four main science themes

- Probing dark energy and dark matter.
- Taking an inventory of the solar system.
- Exploring the transient optical sky.
- Mapping the Milky Way.

Eight science collaborations (about 1,500 scientists) - several dozens of roadmaps

Rubin/LSST data products



Rubin Observatory (2025+)

- **20TB of images / night**
- **1TB of alerts / night**: x100-x1000 above current streams
- *Everything matters a priori*

Now — **Raw Data**
Sequential 30s image, 20TB/night

60s — **Prompt Data Product**
Difference Image Analysis
Alerts: up to 10 million per night

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

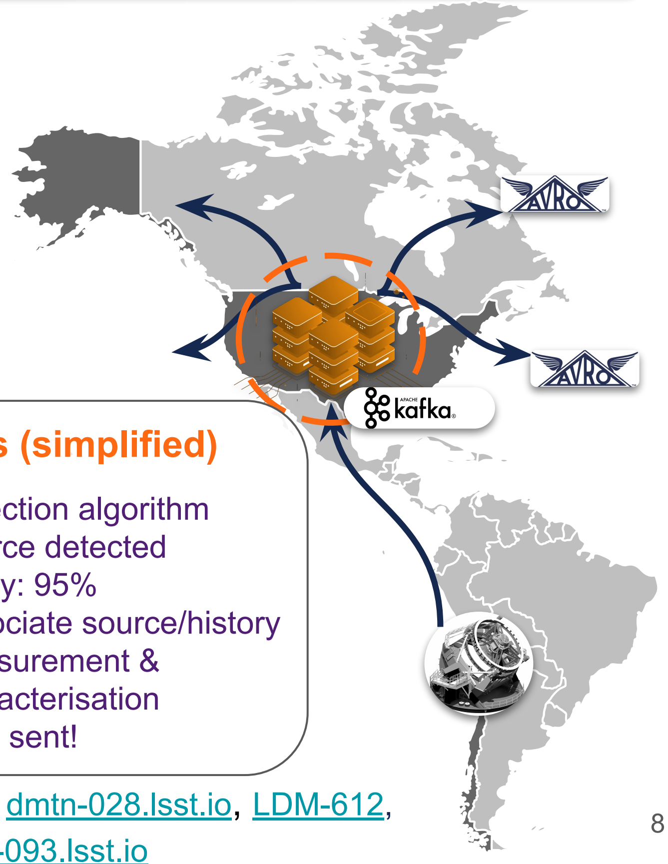
Rubin alert system

Image data sent from Chile to the USA. Alert system will identify sources that move or vary within 60 seconds.

- Sources packaged with contextual information into world-public alert packets for distribution.

Suite of open source technologies considered for distributing alerts

- Binary serialization format: Apache Avro
- Alert distribution: Apache Kafka



Steps (simplified)

1 Detection algorithm
2 Source detected
Purity: 95%
3 Associate source/history
4 Measurement & characterisation
5 Alert sent!

Technical notes: dmtn-028.lsst.io, LDM-612, ls.st/dpdd, dmtn-093.lsst.io

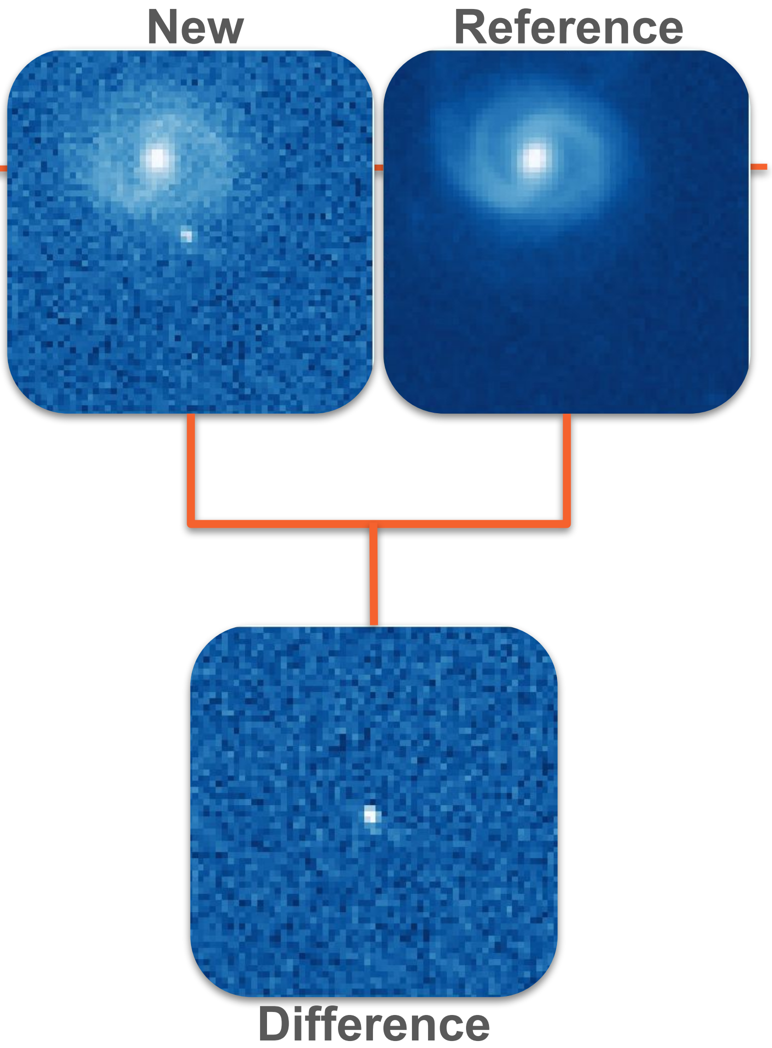
Alert content

Alerts based on **Difference Image Analysis**

Each alert contains

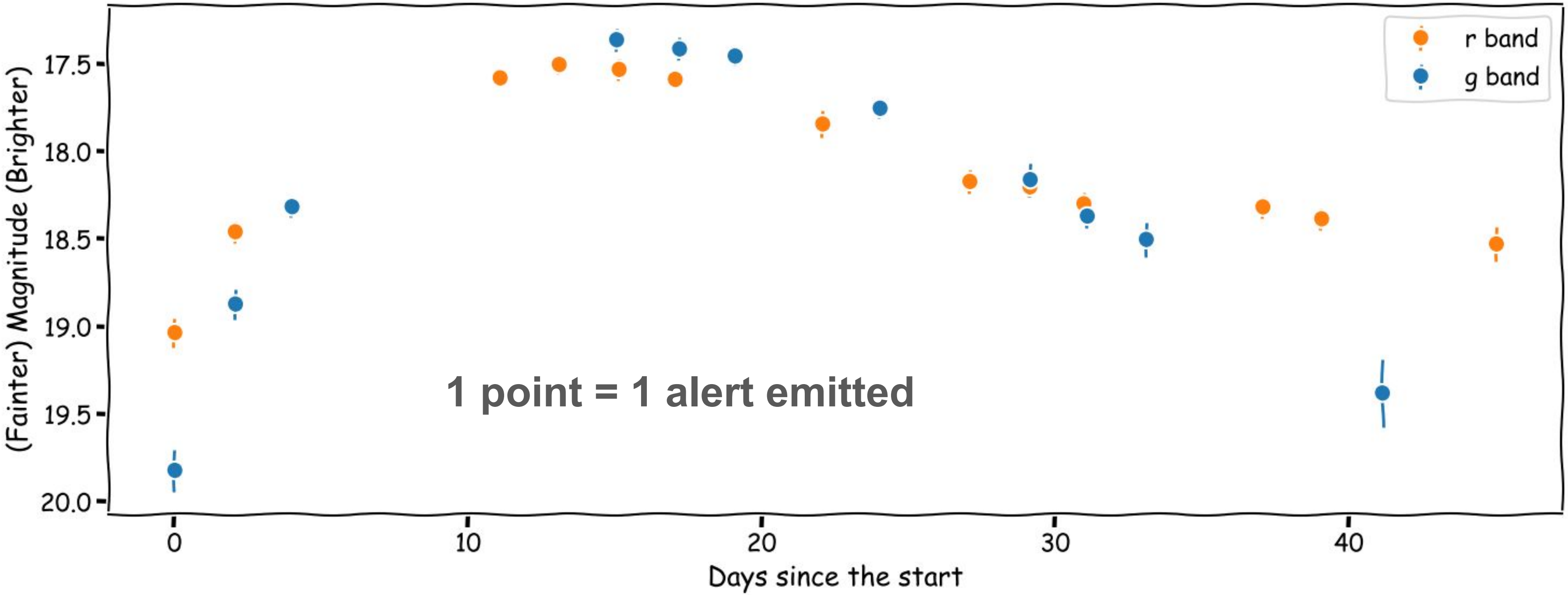
- Information about the new detection (magnitude estimate, position, ...)
- Neighbours information (Gaia, Panstarrs)
- Historical information at this position
- Small images around the detection (60x60 pixels)

More information:
https://github.com/lsst/alert_packet

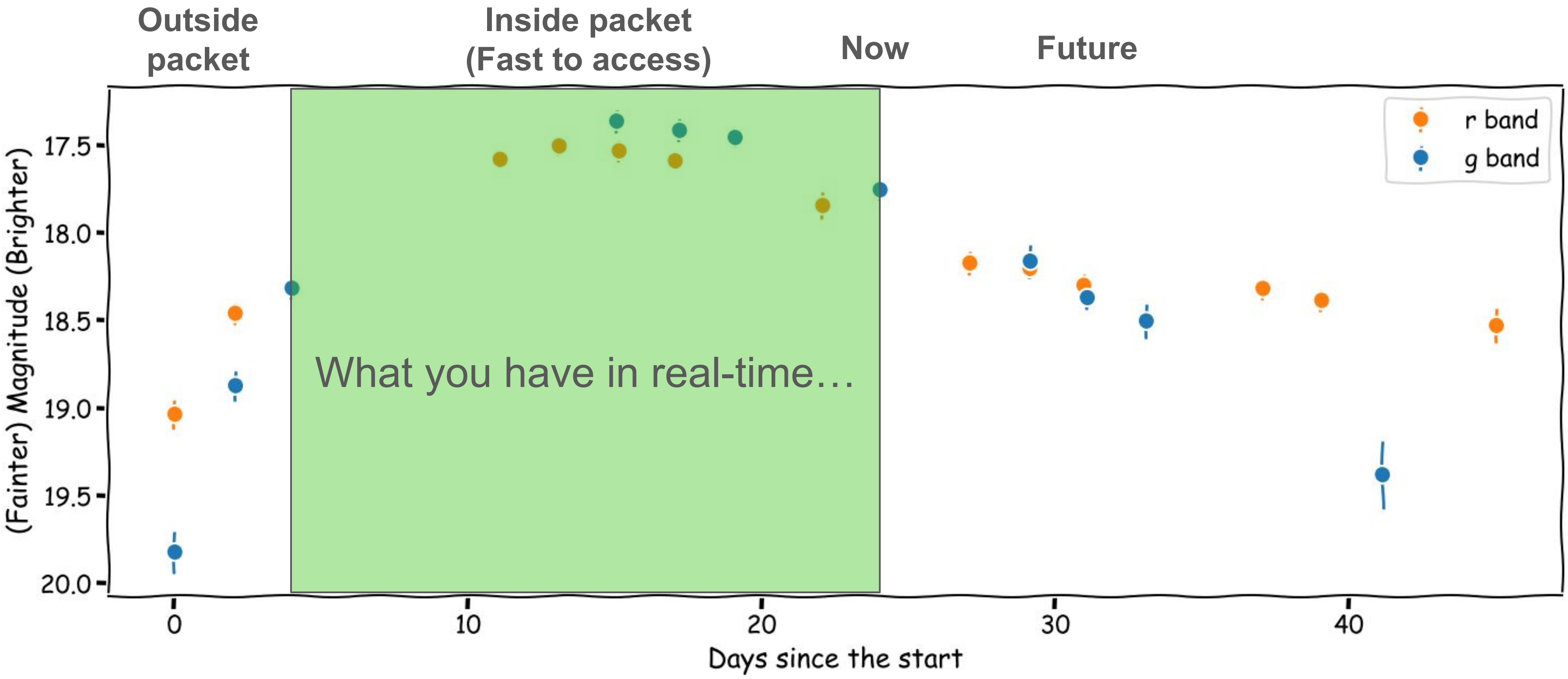


Lightcurve (Ia)

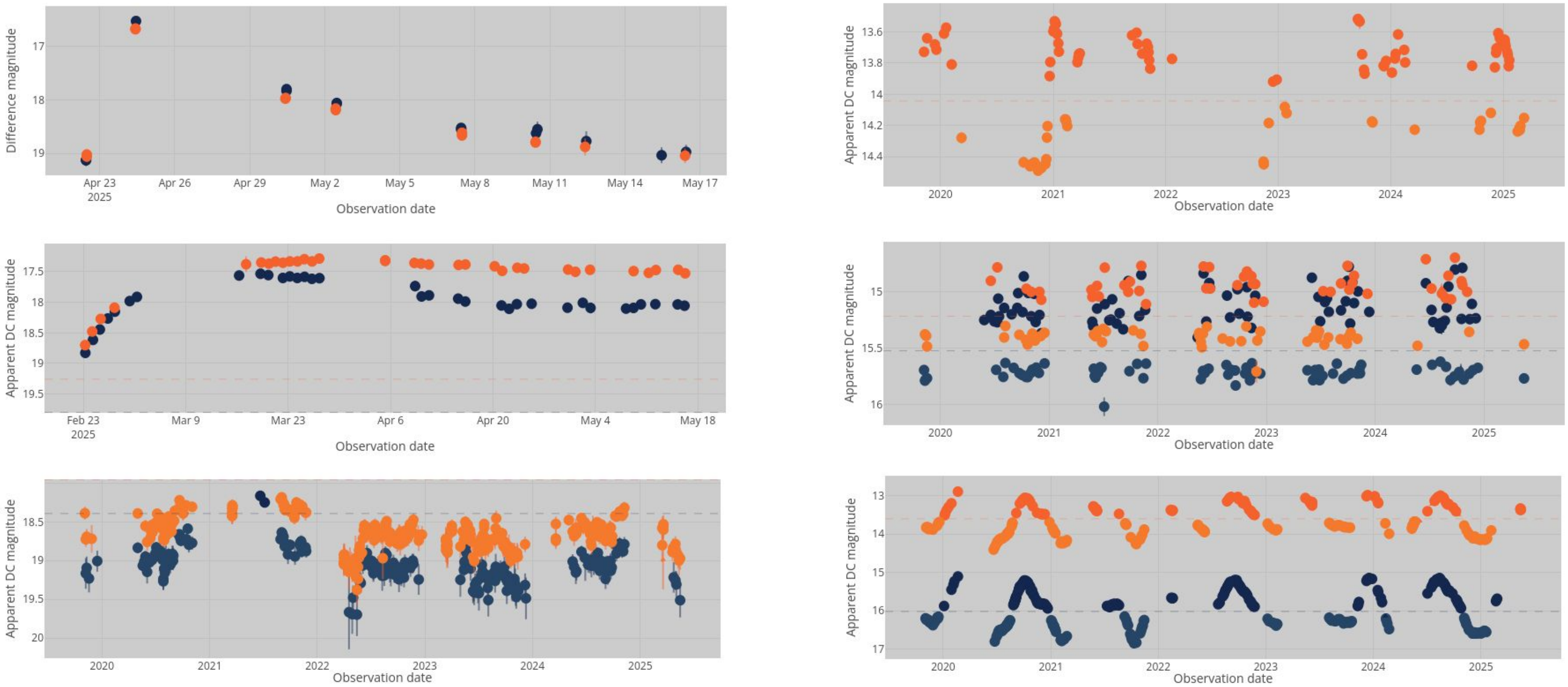
Irregularly-spaced observations. Different cadence per band.



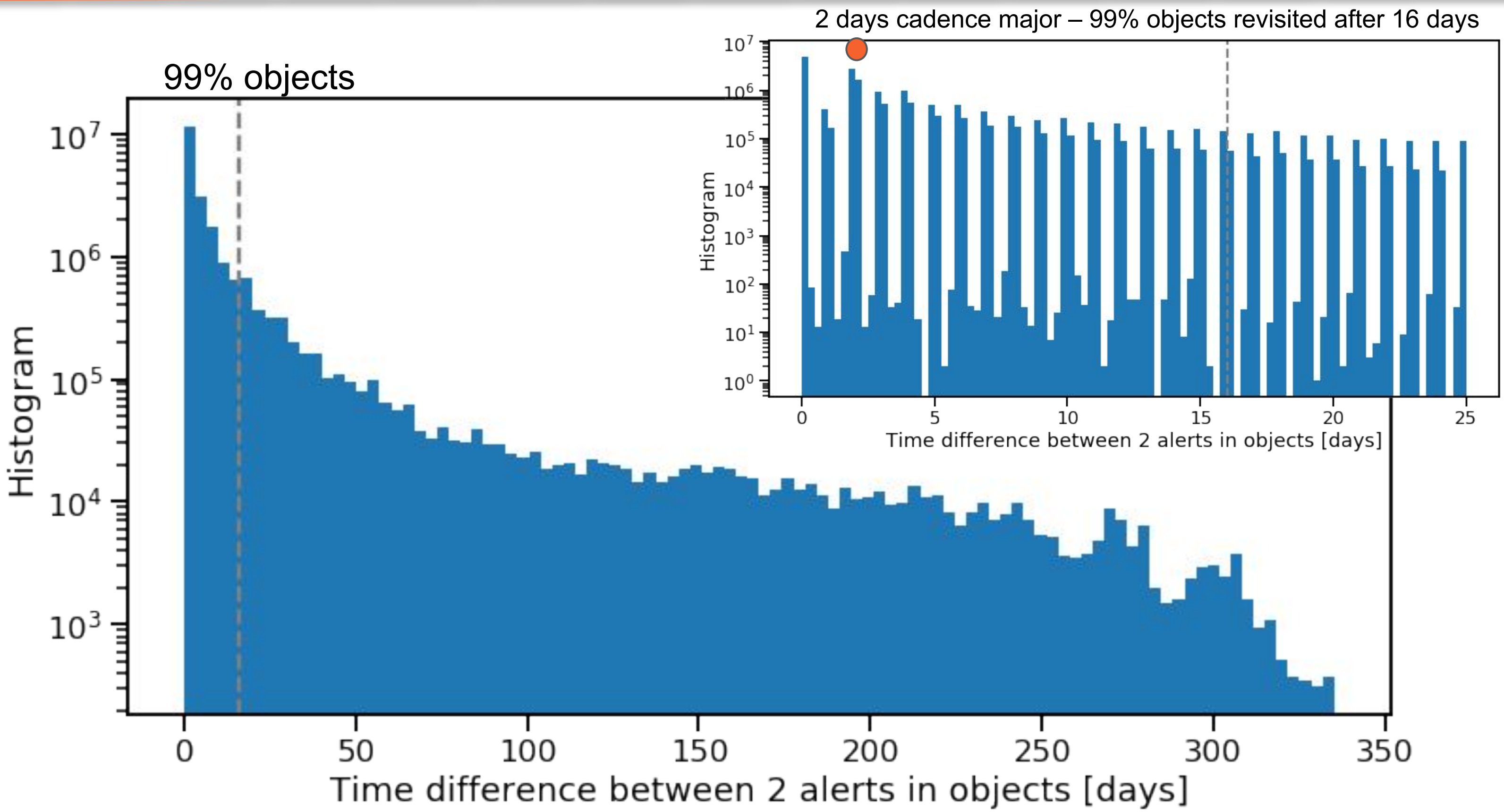
Lightcurve (Ia)



Lightcurves…



Cadence for ZTF (2021)



The rise of brokers

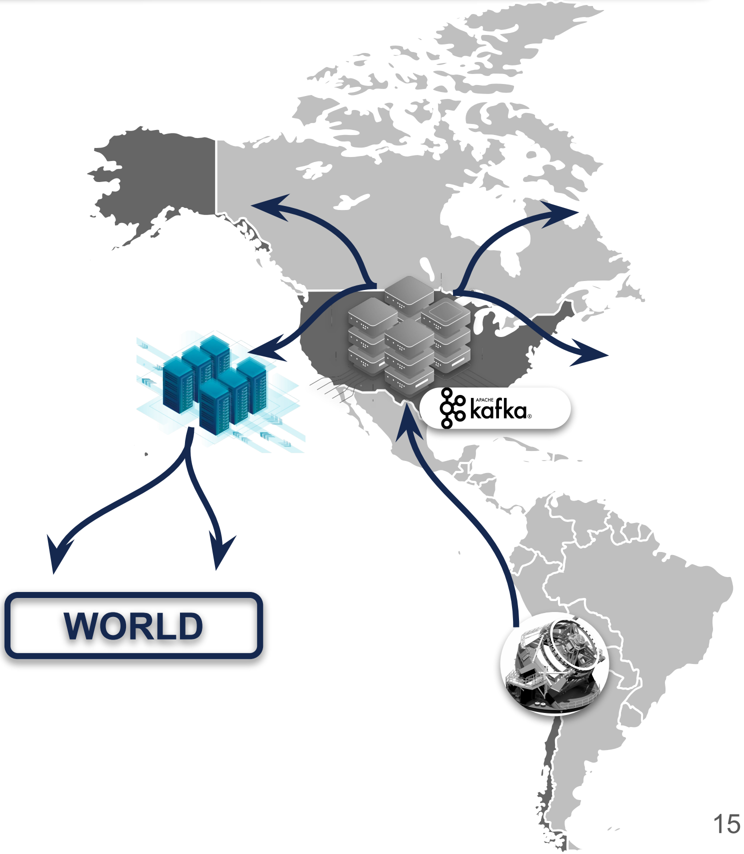
Rubin brokers

Broker: system ingesting, **classifying**, filtering, and redistributing alerts. Service to the community.

Rubin will send the full alert stream to seven brokers; others and individuals will operate downstream.

- ALERCE, AMPEL, ANTARES, Babamul, [Fink](), Lasair, Pitt-Google

All already operating on ZTF (300k alerts/night), and testing deployment of the Rubin Alert Distribution system.



Fink in a nutshell

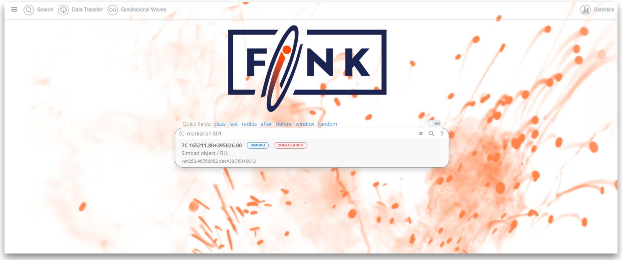
https://fink-broker.org

Fink is a broker serving the scientific community by ingesting, classifying, filtering, and redistributing alerts from telescopes and surveys.

As of 2025: 70+ collaborators, 15 countries

Services deployed on large OpenStack clouds (UPSaclay & CC-IN2P3)

- Scalable to millions of alerts per night



Operating 24/7 since 2019, serving 100+ unique users per day (scientists, follow-up facilities & amateurs)

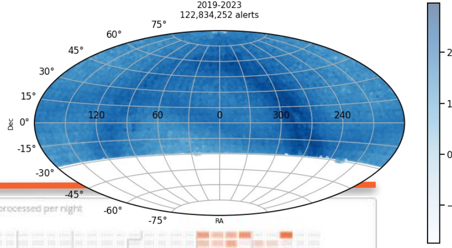
What do we do?

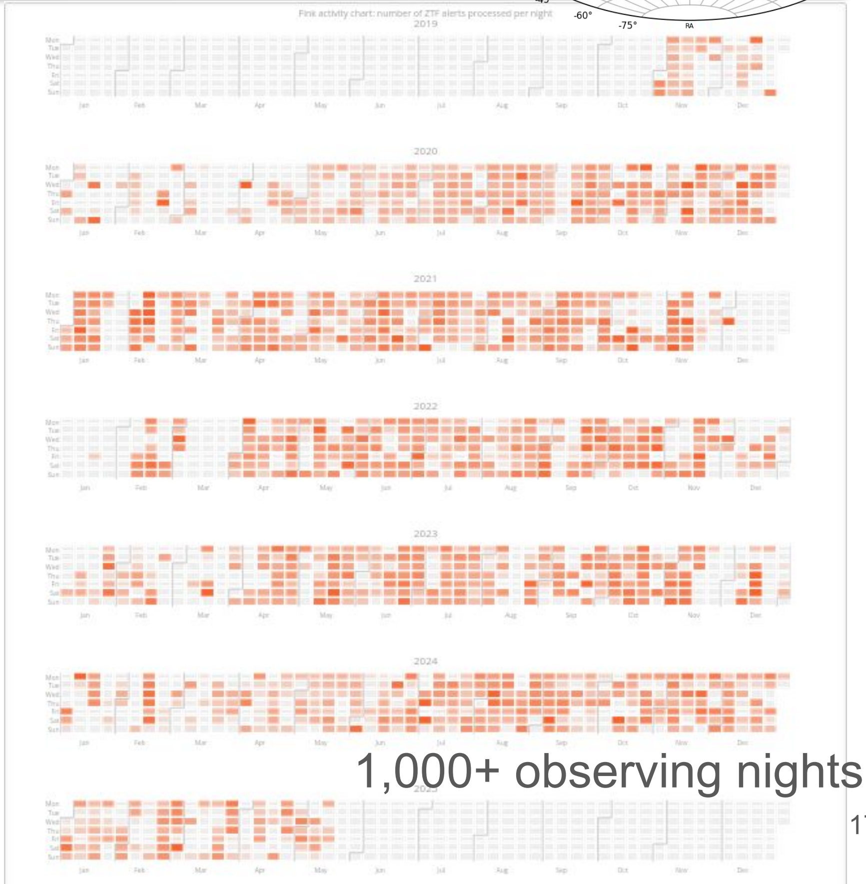
Scientific roadmap is completely open

- Satellites & debris detection
- Solar System science
- Galactic science: microlensing events, cataclysmic variables, YSO…
- Extra-galactic science: supernovae, gamma-ray bursts, blazars, kilonovae, tidal disruptive events, …
- Anomaly detection, hostless transients, …

One man's trash is another man's treasure

https://fink-broker.org/papers





1,000+ observing nights

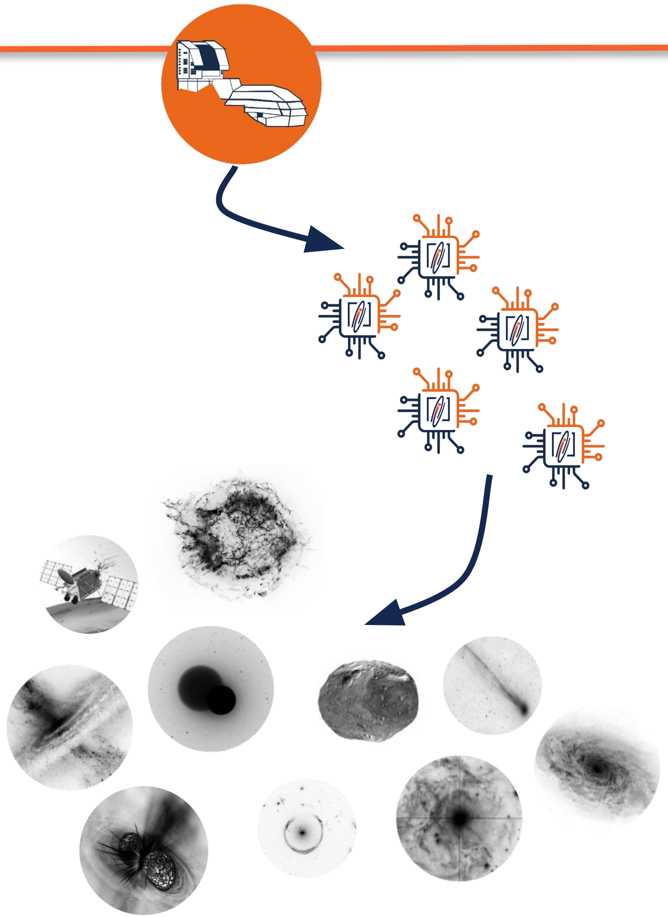
Philosophy

Time-domain astronomy is broad. Opening {code, data, expertise} is the key for success.

The large volume of data impacts possibilities. Providing services & guidance is necessary.

Fink was designed to be modular and to facilitate the inclusion of personalized analyses

- Fink core team provides infrastructure & technical assistance to access data and run analyses (cloud-based)
- Users extend Fink scientific capabilities by providing scientific codes



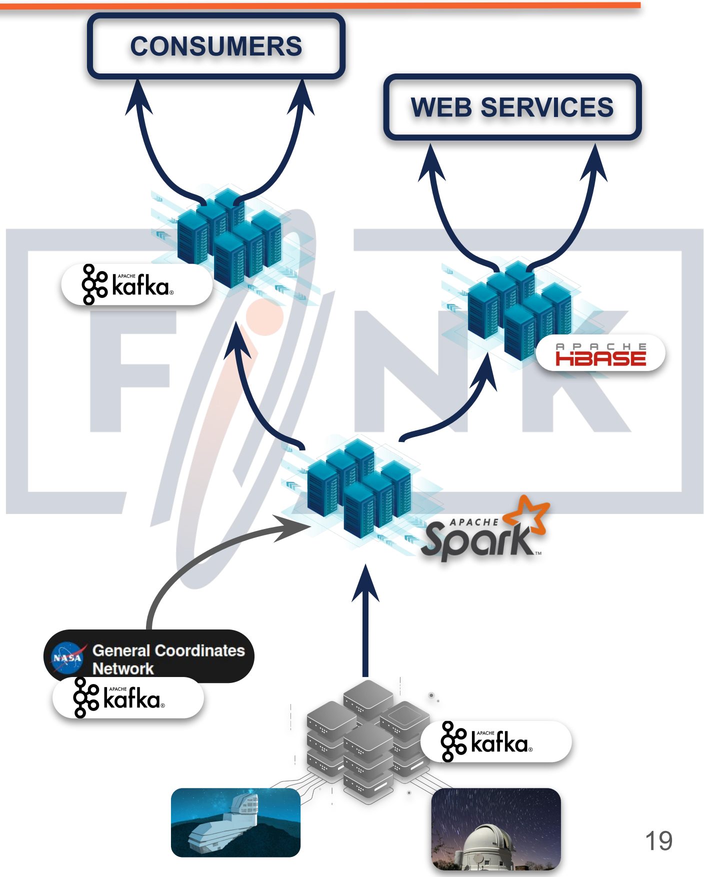
Fink platform

Fink deployed on cloud platforms (UPSaclay, CC-IN2P3). Using distributed computing:

- Computing (Spark), database (HBase), streaming (Kafka), storage (Ceph & HDFS)
- Orchestration: Mesos & Kubernetes
- Autoscaling based on the load

ML: Pytorch, Tensorflow, scikit-learn (note: CPU only!)

Versioning of **code, model, data**



ML in Fink

Some reasons to use ML

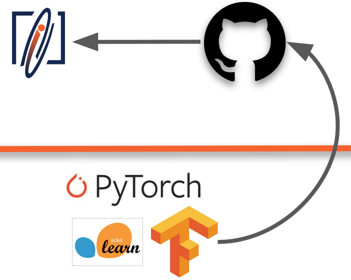
Some reasons why you would like to introduce ML

- Accelerate slow tasks
- Help exploring un-modeled parts of a parameter space (discovery)
- Help in the decision-making process (follow-up)

This presentation will **not** cover everything

- Focus on **some aspects** highlighting the role of ML as an **assistant**
- Focus on **science performance** & **usability** by the community

Challenges



- **Evolving ecosystem**
  - New release/library every 6 months!
  - Containers & plug-and-play architecture
  - Training ahead of time & regular hackathons.
- **Evolving analysis:** how to continuously update classification models?
  - As the surveys go, we learn more
  - Real data is far from available simulations (non-representative, sparse, irregular, noisy…)
  - Models need to (should?) be retrained regularly
- **ML lifecycle can be sporty!**
  - Keeping track of the training sets, dependencies, metric comparisons
  - Gitops/MLFlow
- **Scaling (beyond what we have):** Interfacing with external resources

Some real-life examples

Dark energy
Enabling the discovery of fast transients
Terra incognita
Next-gen classifiers
Deep model compression
Classification problem

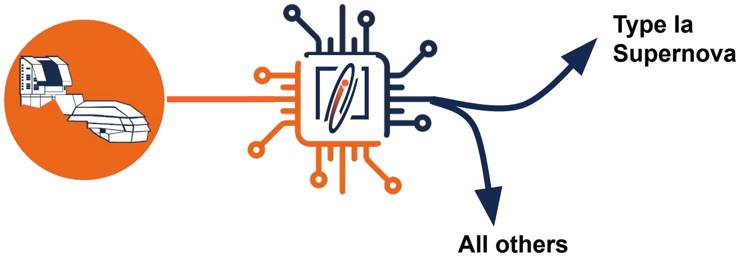
Measuring distances

Goal: Studying *dark energy* using Type Ia Supernovae properties

Difficulty: Finding *early* Type Ia Supernovae!

Solution: using ML to navigate in a sea of transient events each night

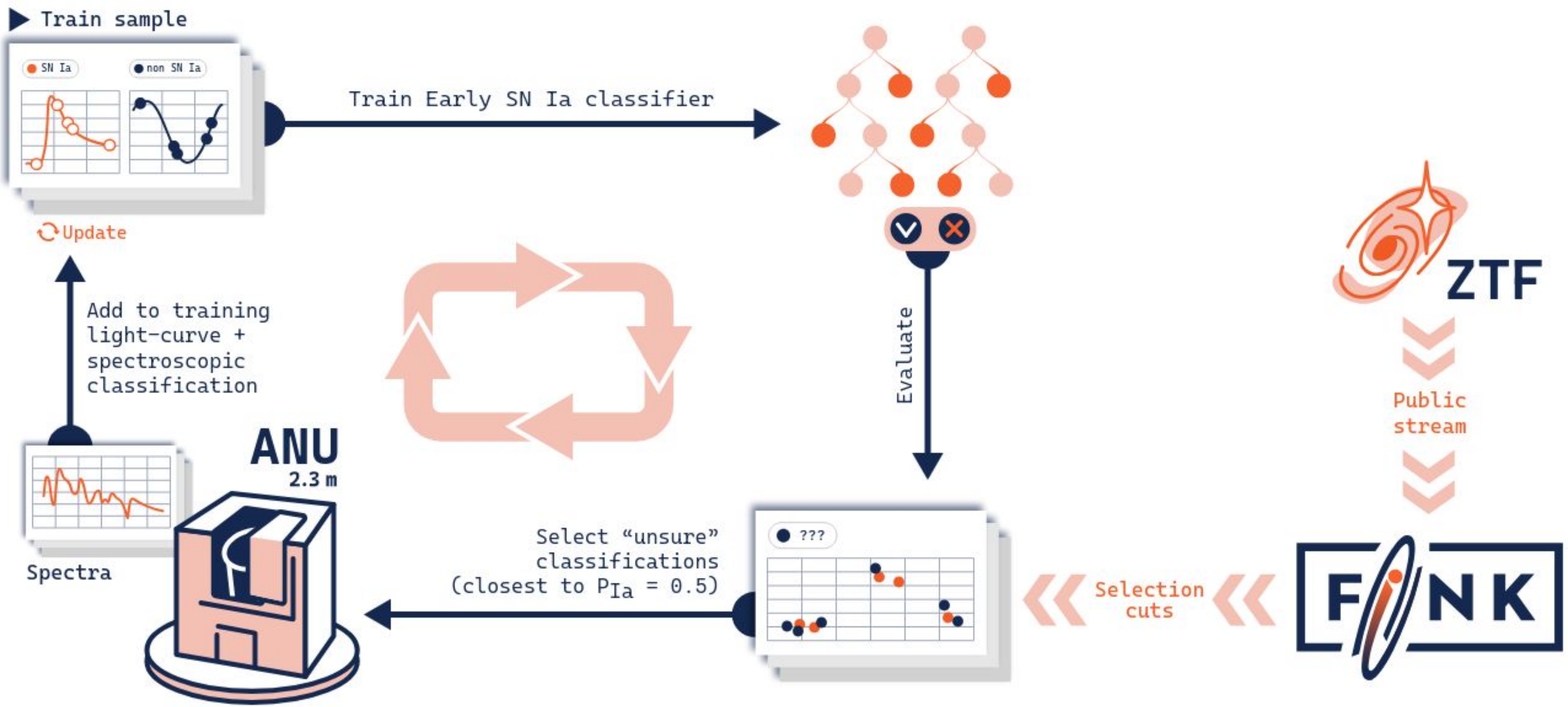
- Combination of Random Forest & DL techniques *(see Anais Möller talk)*
- Fast inference to isolate candidates
  - 0-5 candidates each night
- Data sent to TNS for follow-up
- Data Representativity? *Active Learning*




```
Number of candidates sent to TNS: 2854
Number of classified candidates: 1294
--------------
type
(TNS) SN Ia                   1060
(TNS) SN II                     75
(TNS) SN Ia-91T-like            40
(TNS) SN IIn                    25
(TNS) SN Ic                     24
(TNS) SN Ib                     16
(TNS) SN Ic-BL                  10
(TNS) SN Ia-pec                  7
(TNS) SN IIb                     6
(TNS) SLSN-I                     5
(TNS) SN Iax[02cx-like]          4
(TNS) SN Ib/c                    3
(TNS) SN Ibn                     3
(TNS) TDE                        2
```

Leoni et al 2022
Möller et al 2019

Real-time active learning



Möller et al 2025

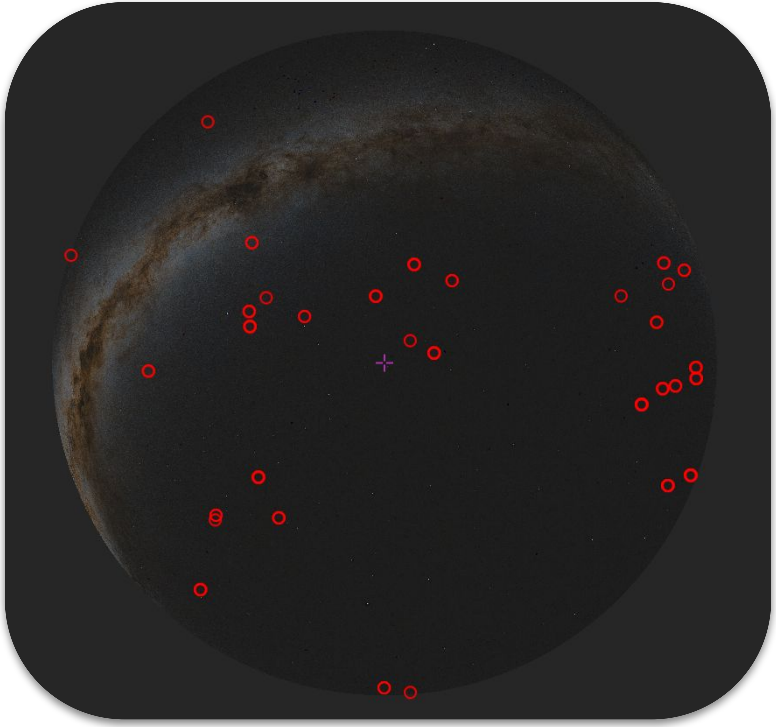
How to access data?

Getting data at this scale is not trivial (training, visualisation, …)

All data available:

- Real-time: instant message applications: Kafka, Slack, Telegram, …
- Historical: REST API & web applications (fink-portal.org)

Open data & science with 200M alerts.



Enabling the discovery of fast transients

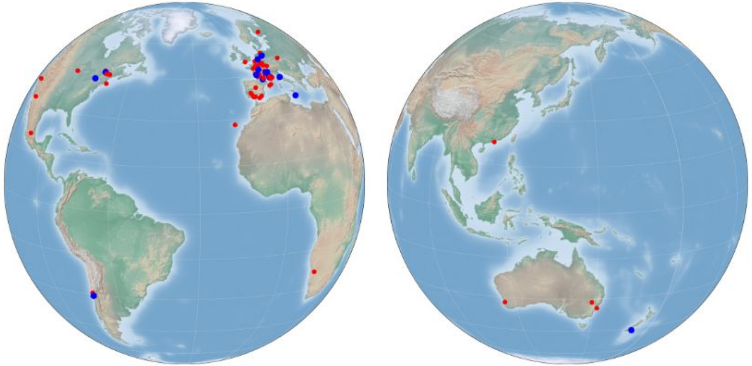
Fink classifies in real-time ZTF alerts from transient phenomena (~200k/night)

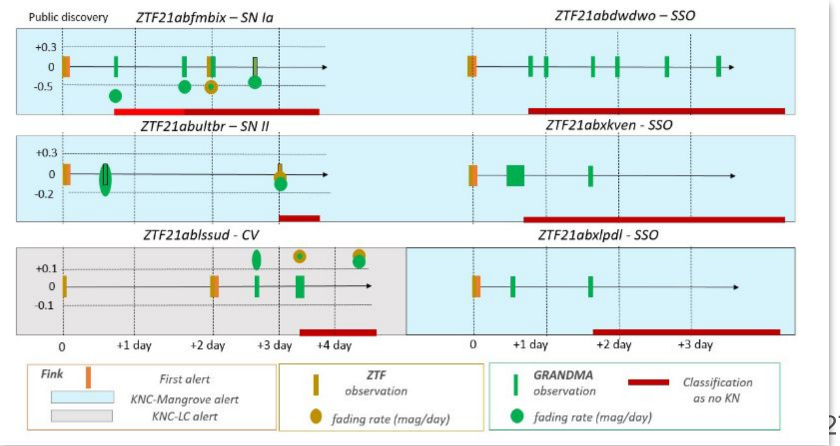
Selected fast transients (~1/night) are sent to the GRANDMA network in real-time for potential follow-up

- ML techniques (Random forest)
- Rate-based consideration
- Contextual consideration

Citizen science program in parallel

[GRANDMA Collaboration 2022 MNRAS 515 4, 6007-6022](#)
[B. Biswas et al 2023 A&A 677, A77](#)
[M. W. Coughlin et al 2023 ApJS 267 31](#)





Terra incognita

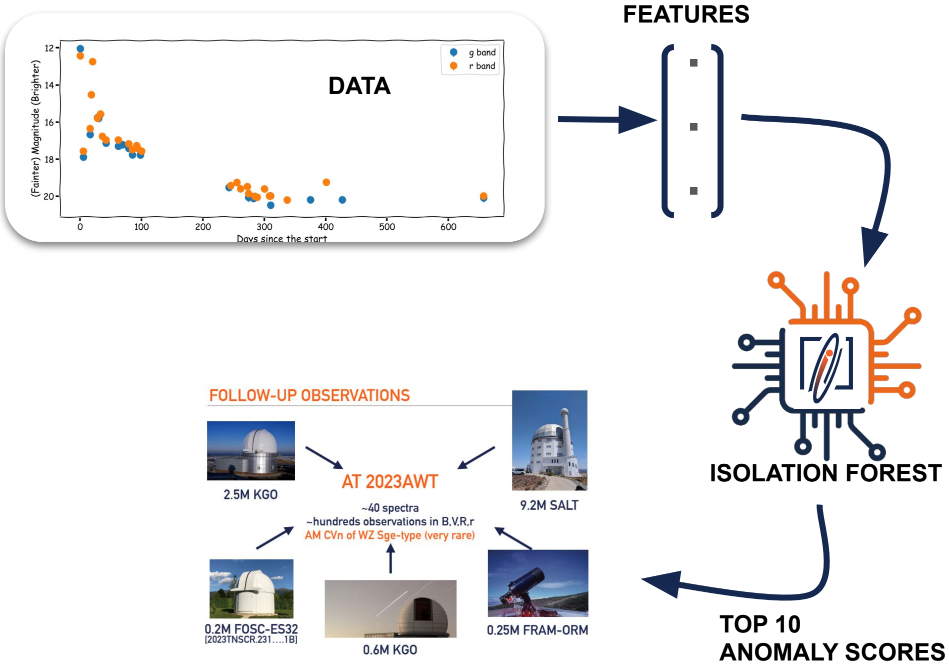
https://snad.space

Discovering unmodeled or unexpected phenomena presents a great challenge

Identifying patterns or behaviors that deviate from *a norm*

Concept of **anomaly detection**

- Anomalous *compared to* something
- Often mathematical description in terms of statistical features associated to a class of transient



AM CVn-type stars (Helium dwarf novae): Interacting Binary White Dwarfs. Close binary systems with ultra short periods (5-70 minutes).
3 systems discovered as of 2024.
Courtesy: M. Pruzhinskaya

Personalization

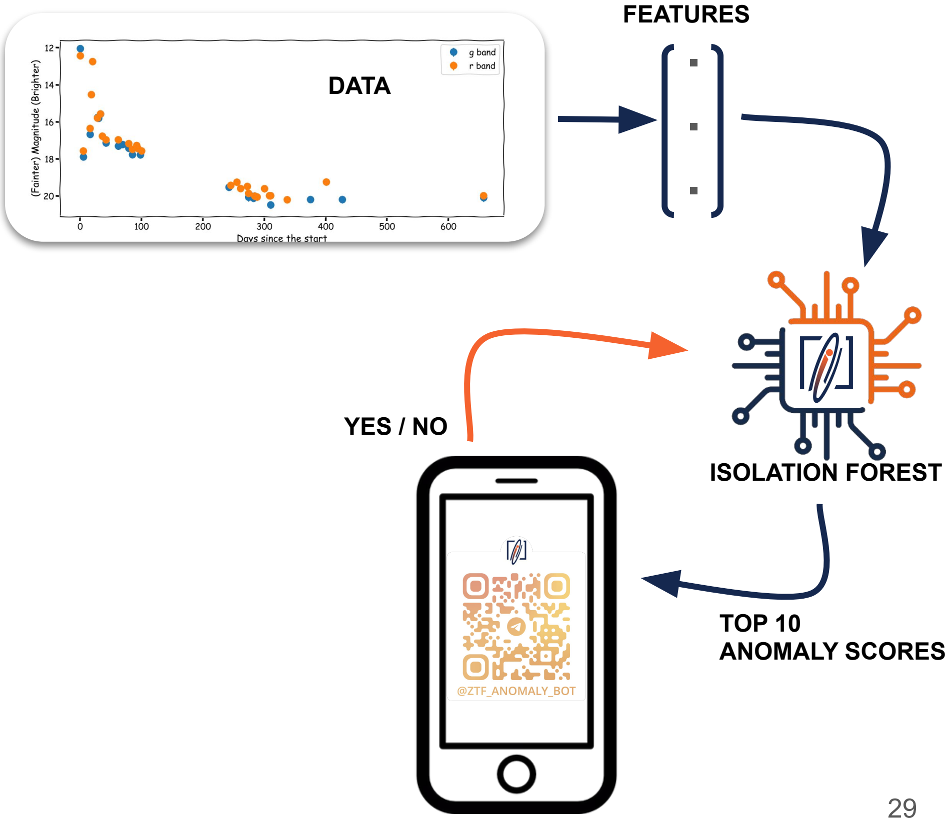
One [hu]man's trash is another [hu]man's treasure

Automatic adjustment of models based on the preferences of astronomers

For night N and person P:

- Objects are scored by a model M_p
- Top 10 scores are shown to P
- **P votes on objects**
- Weights of M_p are updated

New objects for night N+1 are scored, etc.

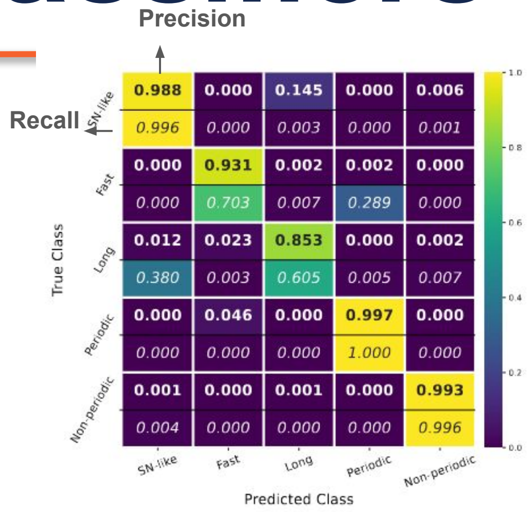


Preparing next-gen classifiers

Classifiers are not really cross-survey as each survey has its own specificity (depth, cadence, filter coverage, …).

The ELaSTICC challenge (LSST-DESC) provided some LSST-like data. Going from ZTF to Rubin was not trivial.

Extra difficulty: accessing training & test data at this scale



Some classifiers updated from ZTF to Rubin, some new using recent architecture (CATS – using Multivariate LSTM Fully Convolutional Network)

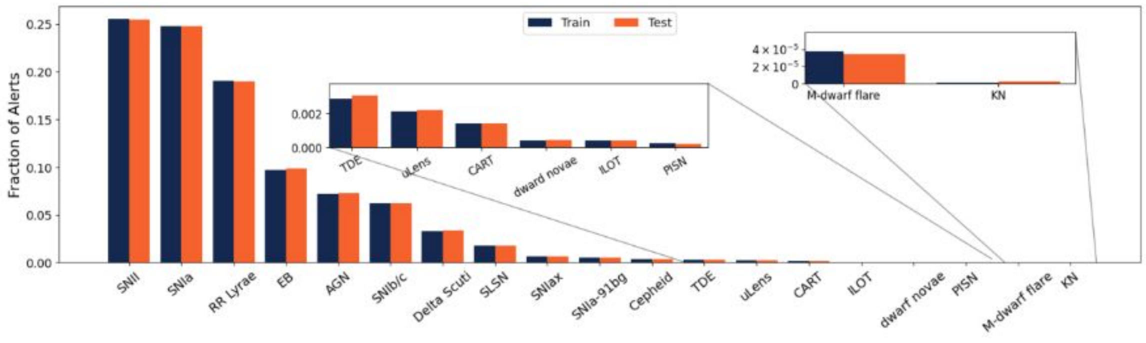


Fig. 1: ELAsTiCC class distribution for our training (dark blue) and test (orange) sets.

[Fraga et al 2024]

Deep model compression

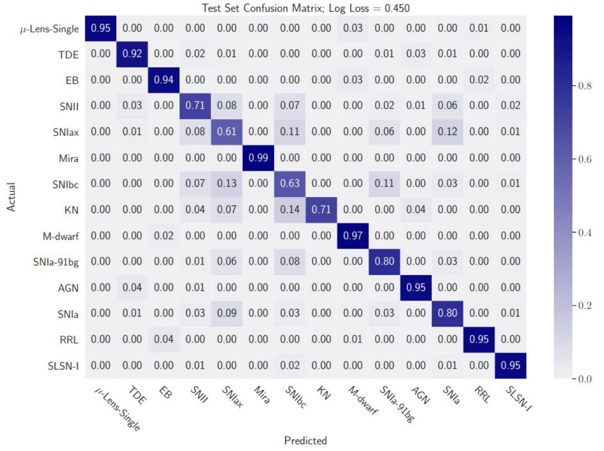
Complex models are often big, and slow. Our goal is to run deep learning on extremely resource constrained devices.

Perform **deep compression** on time-series transformer to produce lightweight models:

- **Weight-clustering:** grouping the weights of each layer.
- **Weight-quantization:** reducing the number of bits used for representing numbers (32 -> 8)
- **Weight-pruning:** removing unimportant weights (does not preserve accuracy)
- **Serialisation format:** FlatBuffers (TFLite)

Improve latency (up to x15,000), reduce model size (up to x20) and preserve accuracy.

Compression Method	Model Size (kB)	Load Latency (s^{-3})	Inference Latency (s)	Loss
Baseline	1100	6324.145	0.333	0.968
Baseline + Huffman	244	6015.565	0.224	0.968
Clustering	892	5559.868	0.227	0.836
Clustering + Pruning	688	5721.021	0.230	1.017
Clustering + Huffman	240	4991.857	0.223	0.836
Clustering + Pruning + Huffman	128	5251.288	0.228	1.017
†Clustering	92	0.426	0.046	0.836
†Clustering + Quantization	**60**	**0.271**	**0.043**	**0.834**

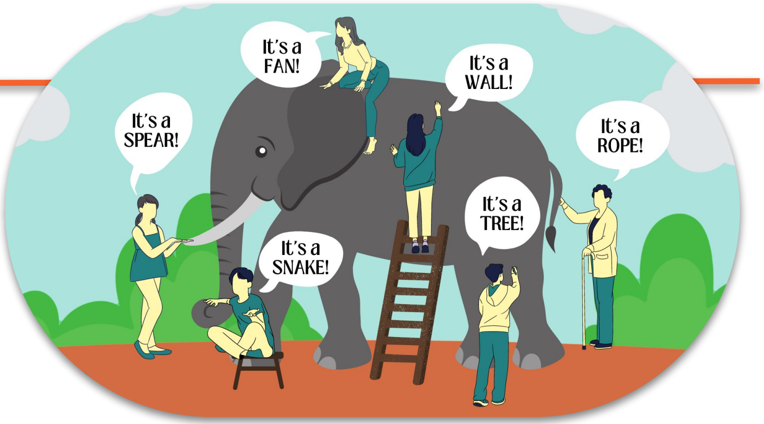


[Allam et al 2023]

Classification

Alert classification vs **object** classification

- Example ZTF25aahmbaq, ZTF21aalxxzn
- Partial view and time evolution

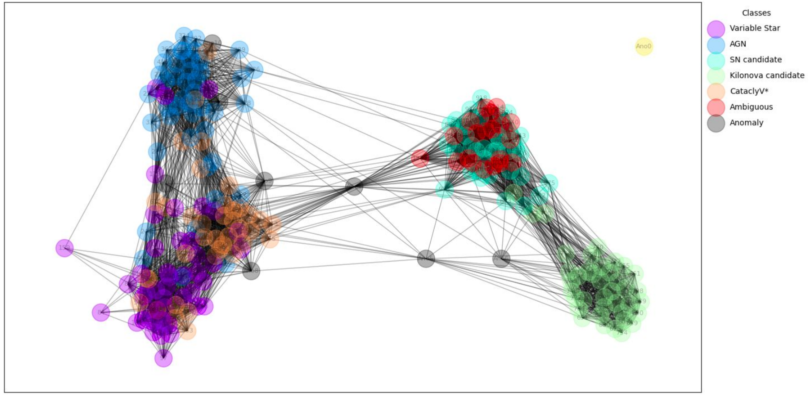


Two strategies: big fat classifier (the truth) versus personalised classifiers (the truths)

How to provide coherent classification scheme (if any)? How to combine classification scores?

Exploring graph-based solutions

- How to define similarity?
  - ML for graph construction
- Free lunch: recommendation system!



Thinking beyond

Computerization

A broker team cannot only be engineers talking to engineers

- The interfaces with the communities of users are crucial
- Companies have usually their own team dedicated to this

We are facing a LOT of sociological problems

- The use of computers is not innocent, even in 2025

Some pressing questions for all broker teams:

- How to make sure tools fit user skills and needs?
  - How to reach and teach the user base?
- How to make sure tools are used efficiently?
- How to make sure tools can be flexible enough to be adapted?

Our duty is not only to perform (best results on biggest machines), but make it usable by normal people with commodity hardware.

Broader impact

Societal Impact: broader than just astronomy, often driven by external forces. Mix of excitement and apprehension, unclear path forwards.

Change in Work Methods: AI is changing traditional approaches to data analysis and observation. Not always required, but can replace large parts of analyses. Can be daunting for those accustomed to conventional techniques.

Confrontation with Computing: More and more confronted to advanced computing tools and infrastructures. Many may lack the necessary programming and data analysis skills, creating a barrier to effectively keep doing science.

Need for Abstraction/simplification: How to effectively understand and utilize AI tools? Often not intuitive for non-experts. *Generative AI* can streamline processes, but many open questions…

Takeaway messages

New facilities and telescopes generate **unprecedented data volumes**, more and more complex. And it doesn't seem to want to stop anytime soon.

The role of ML as an **assistant** in handling astronomical data is well established:

- Accelerate slow tasks
- Help exploring un-modeled parts of a parameter space
- Help in the decision-making process

The most difficult challenges lie in the changes brought about in **work methods** and the **required skills** to continue advancing science

- Personalisation is a key. Training is another one.
- Do not develop only for yourself – think about other users!

Questions?

https://fink-broker.org