Pitt-Google Cloud-Based Broker

Michael Wood-Vasey (Pitt)
on behalf of
Christine Mazzola, Daniel Perrefort, Troy Raen (Pitt)
Ross Thomson (Google)

https://github.com/mwvgroup/Pitt-Google-Broker

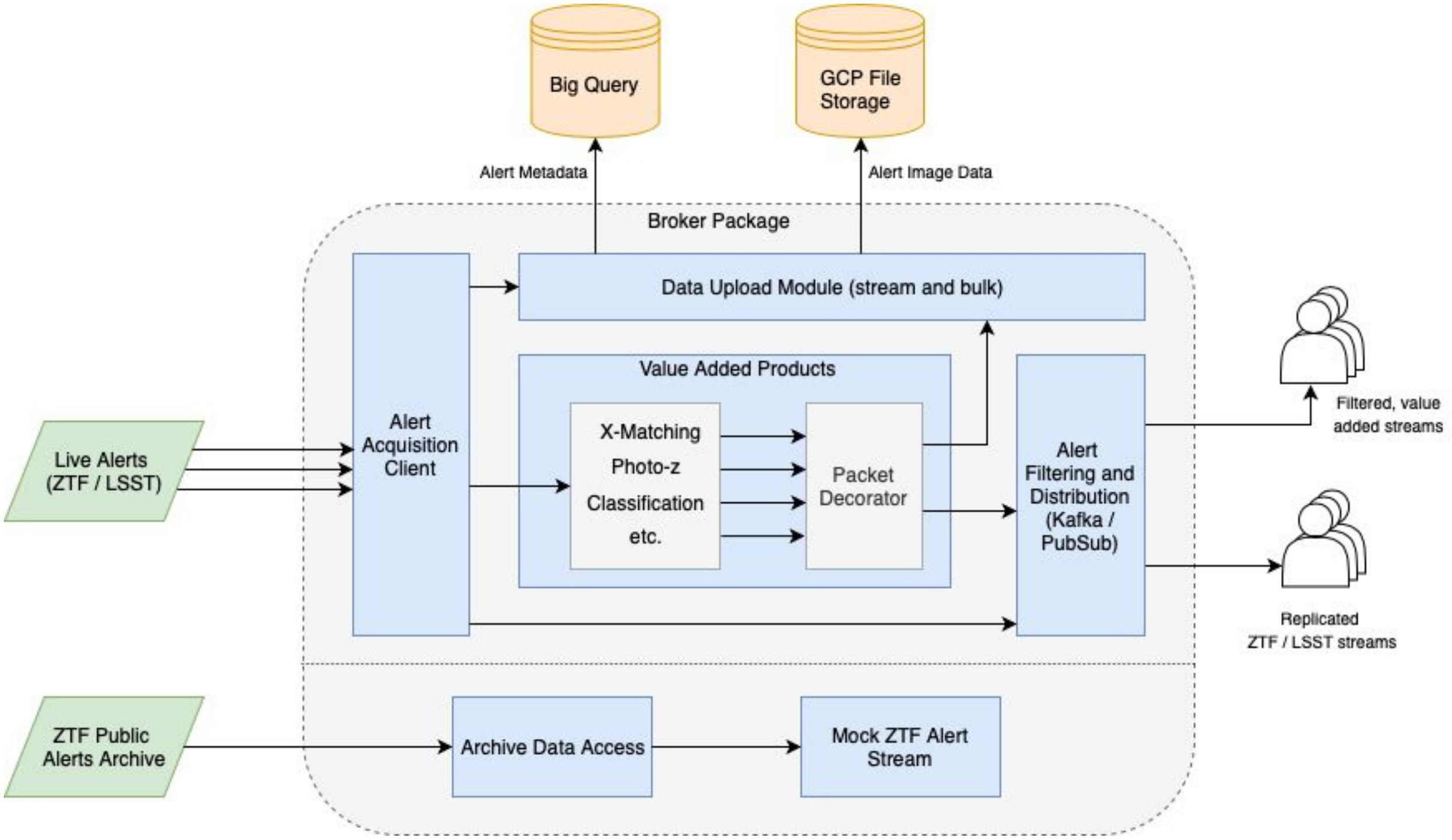
https://pitt-broker.readthedocs.io/en/latest/index.html

We Will Provide

- The public alert data to **everyone**
- **Scalable replication** to N brokers
- **Semantically compressed streams** to 1,000 consumers
- Public analog of PPDB: “**AlertDB**”
- **Cross-matching** across surveys and wavelength
- **Re-playable classifications** based on publicly-available classifiers and a Bayesian belief network meta-classifier.

https://github.com/mwvgroup/Pitt-Google-Broker
https://pitt-broker.readthedocs.io/en/latest/index.html

Conceptual Design



Google Cloud Platform (GCP)

Motivating Use Cases

- To provide specific use cases, we currently plan to develop our broker around three example questions:
  - Supernova Cosmology
  - Cataclysmic Variables
  - Transient and variable populations through large-scale classification
- We welcome additional collaborators with motivating use cases to improve the reach and suitability of the broker and interface.

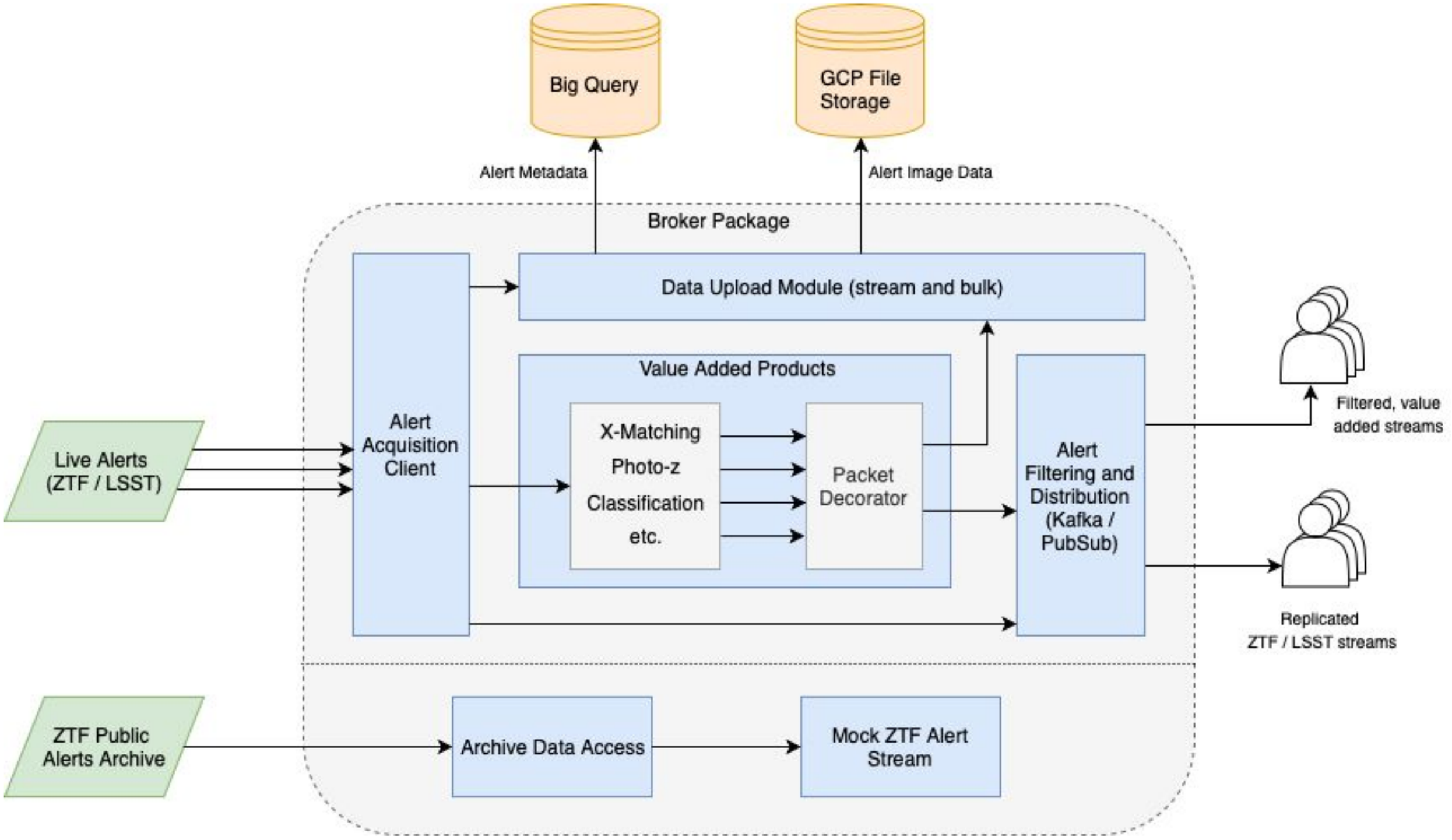
Semantically Compressed Alert Stream

- A concise version of the stream a factor of 100 times smaller can be more easily (and cheaply) distributed to a large community of up to 10,000 users.

- In a Cloud computing model, charging is often dominated by ingress/egress from data centers, regions, or processing steps.

- What are the 10(?) numbers that best summarize an event for filtering? The LSST Project DIAObject design includes space 52 (32 periodic, 20 non-periodic) numbers whose definition is TBD. What should those be?

Conceptual Design



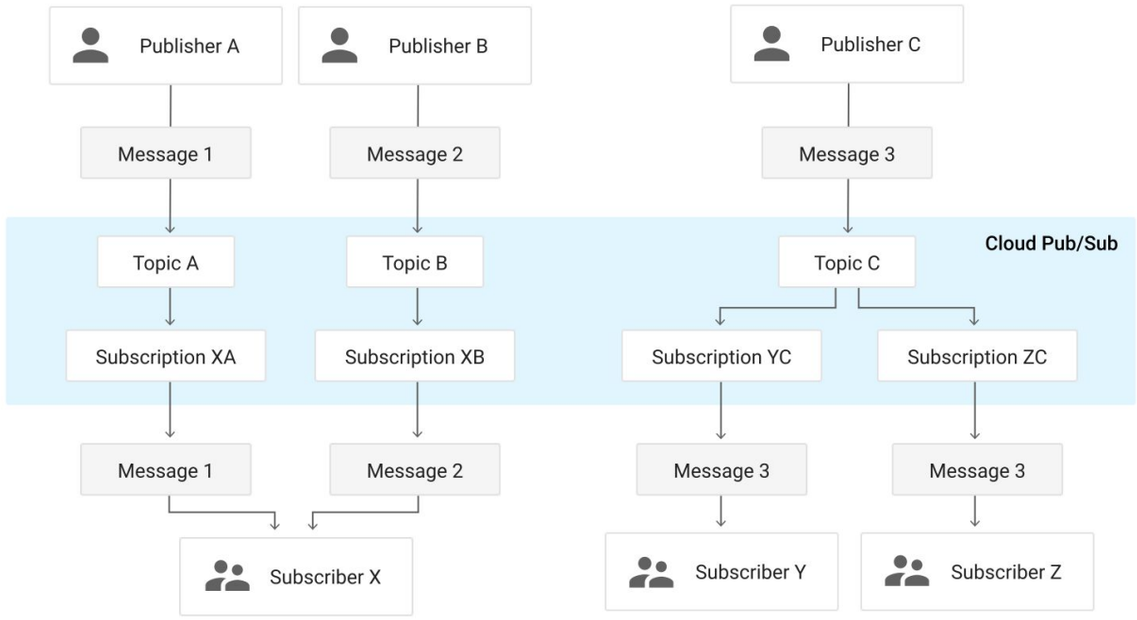
Google Cloud Platform (GCP)

Cloud Pub/Sub

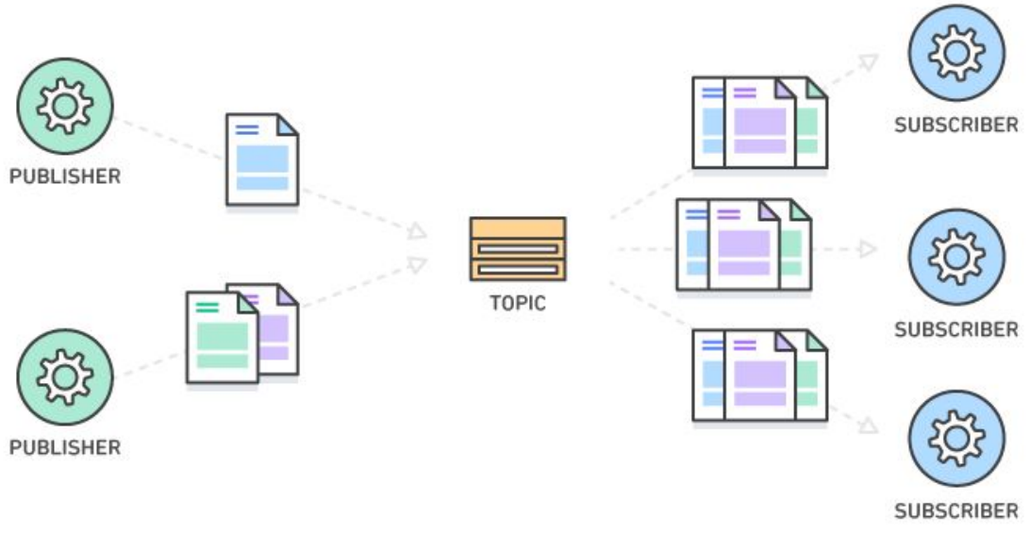
- Publishers and Subscribers organized around Topics.

- Topics can be flexibly defined.
  All
  SNela at z>0.2
  Transients in LMC
  Microlensing toward Galactic bulge, …

- Infrastructure supported under GCP and AWS.



https://cloud.google.com/pubsub/docs/overview



https://aws.amazon.com/pub-sub-messaging/

AlertDB for the World

- An approximate analog to the LSST Prompt Processing Products Database (PPDB) that provides database functionality to the key elements of the LSST alert stream.

- Allow people with time interests on the scale of hours to have access to the information on the alert stream without having to wait for the 24-hour update of the planned Prompt Processing Products Database.

- Provide PPDB-like access to anyone in the world.

- Maintain information sufficient reproduce state at any given date to re-play classifications for filtering and simulations.

Providing both Streams and AlertDB through Google

- The alert stream will be ingested into a BigQuery database, which will be on the scale of billions of entries.

- BigQuery supports streaming ingest and large numbers of simultaneous users.

- Streams supported over either Kafka or Pub/Sub.
  - If you want the full stream, Kafka is fine
  - Pub/Sub can do the full stream, and is likely particularly better suited to Topic for subsets.

Planned APIs

- Direct connections to the Kafka stream or Pub/Sub from within Google Cloud or the outside world ($$$).

- SQL access to the accumulated AlertDB through BigQuery.

- A higher-level Python API to support more abstract operations

- A higher-level Python API to support integration with additional external databases and resources

- A service to which Target and Observations Monitor (TOM) system can connect.

Classification and Cross-Matching

- Will load publicly-available datasets (Gaia, SDSS, Pan-STARRS, …) into Google Cloud to facilitate cross-matching

- Will implement multiple publicly-available classifiers representing a range of objects (e.g., SuperNNova (Möller+19), UPSILoN (Kim+16), …)

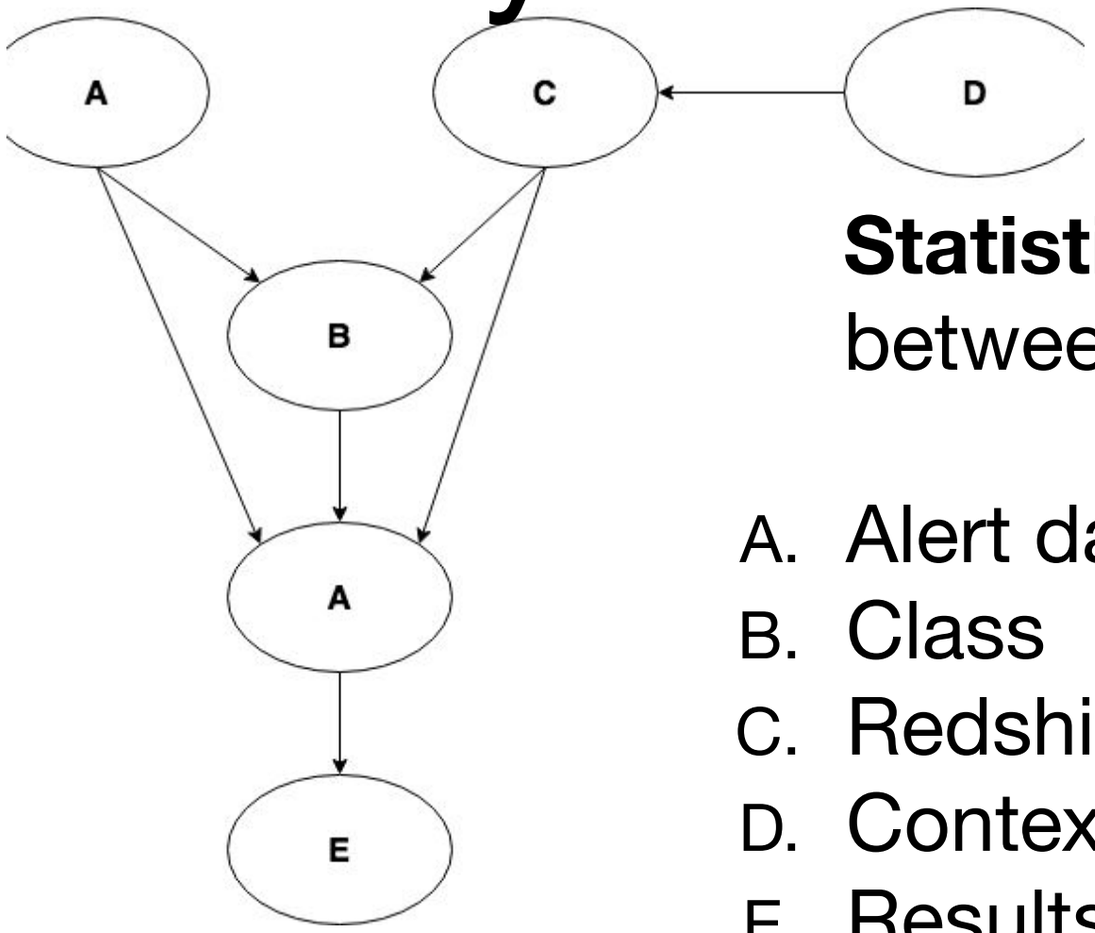
- Broker meta-classifier will be structured as a Bayesian belief network (see next slides)

- Classifications and cross-matching will be re-playable

- Codes are being included in a modular way to allow individuals to move their analyses between the 60-second, 24-hour, and yearly data with connections to LSST Science Platform for those with LSST Data Rights.

- API hooks will provide access to any stage of the pipeline (e.g., run your own classifier)

Bayesian Belief Network

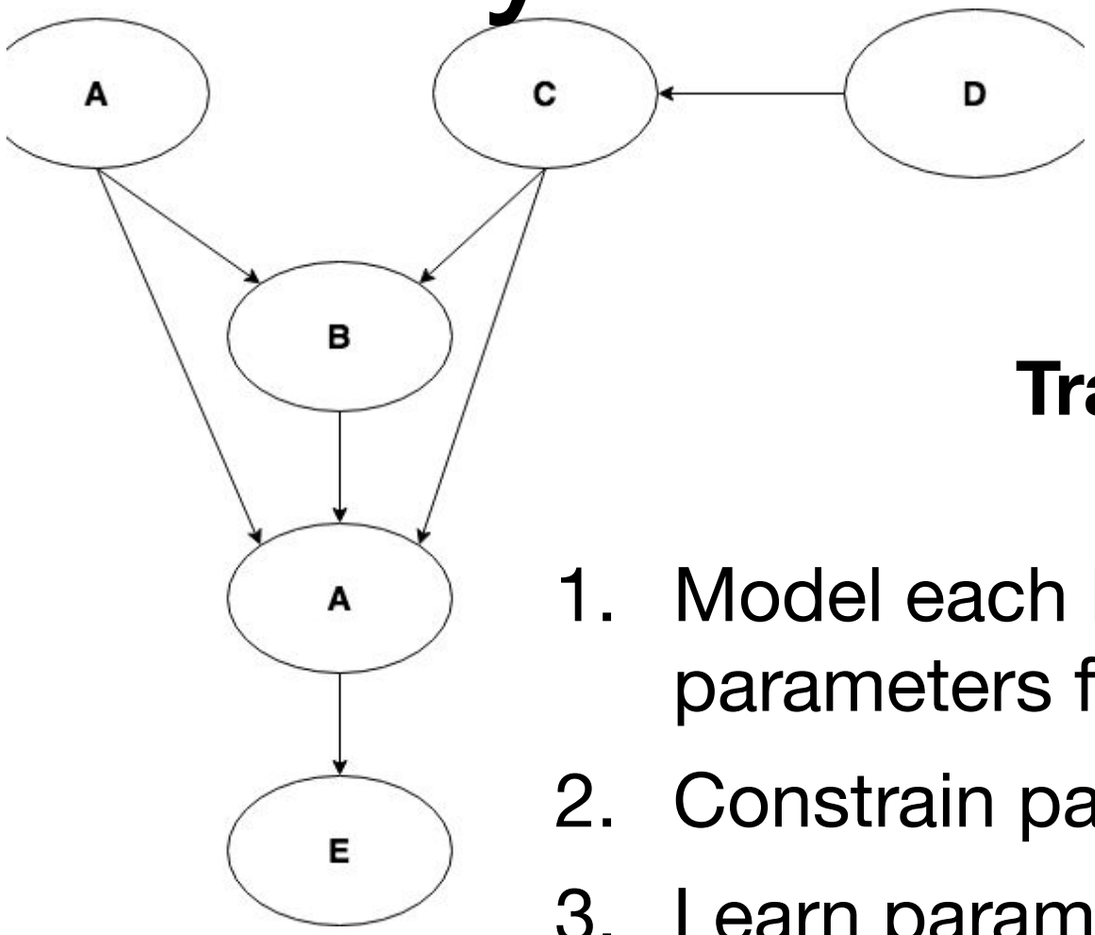


Statistical model of causal relationships between:

A. Alert data
B. Class
C. Redshift
D. Contextual info from cross-matching
E. Results from multiple, publicly-available (“targeted”) classifiers, (i.e., SNe classifier, variable stars classifier, …).

- conditional independencies simplify the full joint probability to $P(A,B,C,D,E) = \Pi_{nodes} P(node \mid parents)$

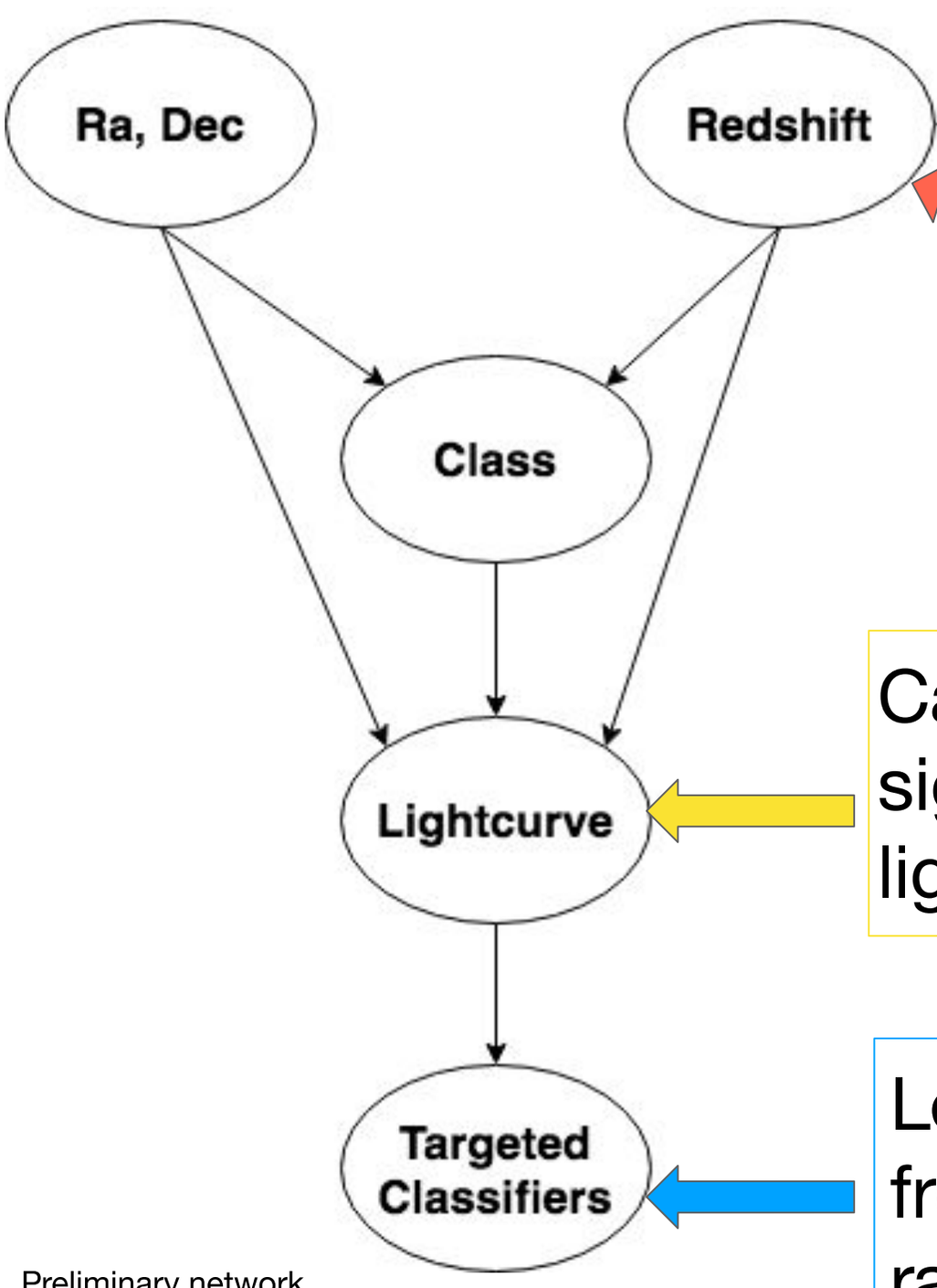
Bayesian Belief Network



Training and Prediction

1. Model each P(node | parents), leaving parameters free
2. Constrain parameters with priors
3. Learn parameters by maximizing P(data | params) x P(params) (unsupervised learning)
4. Use inference to make class (and redshift) determination

Bayesian Belief Network



Redshift separation allows for missing, multiple, uncertain, and/or unassociated cross matches.

Can we achieve accurate results using a significantly reduced representation of the lightcurve?

Learns probability distributions of results from each targeted classifier on the full range of lightcurves expected in LSST.

Preliminary network

Costing Model

- Combining technical expertise with negotiation to make the most efficient use of the cost model for Cloud services.

- We will prioritize

  - Providing condensed, but information-rich streams to everyone

  - Setting up clear code instructions for how someone can run their own server

  - Consider joint funding models if additional external users would rather just pay money to contribute to a single main service.

- Apply for external public and private funding to support fuller public access.

- Dominant cost under is storage and Pub/Sub subscriptions. Could request discount in exchange for the data being publicly available. Because making it public is, in fact, the point, we are optimistic.

GCP Pricing: Conservative

- Without negotiated discounts, a conservative pricing is with 1,000 simultaneous users

BigQuery	Value	Unit	Price [USD]	Unit	Price/max year [USD]	Price/10 years** [USD]
Full AlertDB Storage	400	TB	0.02	/GB/month	98,304	491,520
Summary AlertDB Storage	4	TB	0.02	/GB/month	983	4,915
Full+Summary Query Access	1,000	slots	8,500	/500 slots/month	204,000	2,040,000
Pub/Sub*	1	TB / day	40	/TB	14,400	144,000
TOTAL					303,287	**2,680,435**

[*] Assume distribution of 100x compress stream to 1024 users
[**] For Storage: Price/10 years is 1/2 * (10 max years)

GCP Pricing: Optimistic

- Take storage at non-recent rate (> 90 days)
  400 TB/month of queries simultaneous users

BigQuery	Value	Unit	Price [USD]	Unit	Price/max year [USD]	Price/10 years** [USD]
Full AlertDB Storage	400	TB	0.01	/GB/month	49,152	245,760
Summary AlertDB Storage	4	TB	0.01	/GB/month	492	2,458
Full+Summary Query Access	400	TB	5	/TB	24,000	240,000
Pub/Sub*	1	TB / day	40	/TB	14,400	144,000
TOTAL					73,644	**632,218**

[*] Assume distribution of 100x compress stream to 1024 users
[**] For Storage: Price/10 years is 1/2 * (10 max years)

GCP Pricing: Public Data Set

- Convince Google to host AlertDB as Public Data Set.
  Pay for Large Access for 500 simultaneous users.

BigQuery	Value	Unit	Price [USD]	Unit	Price/max year [USD]	Price/10 years** [USD]
Full AlertDB Storage	400	TB	0	/GB/month	0	0
Summary AlertDB Storage	4	TB	0	/GB/month	0	0
Full+Summary Query Access	400	TB	5	/TB	24,000	240,000
Pub/Sub*	1	TB / day	10	/TB	3,600	36,000
TOTAL					24,000	**276,000**

[*] Assume distribution of 100x compress stream to 1024 users
[**] For Storage: Price/10 years is 1/2 * (10 max years)

Current Status

- Prototype backend:
  - Ingest Kafka stream from ZTF Archived into BigQuery
  - Pub/Sub module to publish and subscribe
- Prototype webpage interface:
  - Test Sandbox: https://ardent-cycling-243415.appspot.com/
- Interest from and Engagement with Google
- GCP grant of $5,000 GCP credits for 2020
- Google Software Architect design consulting in 2020

https://github.com/mwvgroup/Pitt-Google-Broker
https://pitt-broker.readthedocs.io/en/latest/index.html

Example Landing Page

Filters

Add to Watchlist

Object Page →

All Objects

ObjectID

Watchlist

ObjectID

Classification

Type

Confidence

>

Type to se...

Position

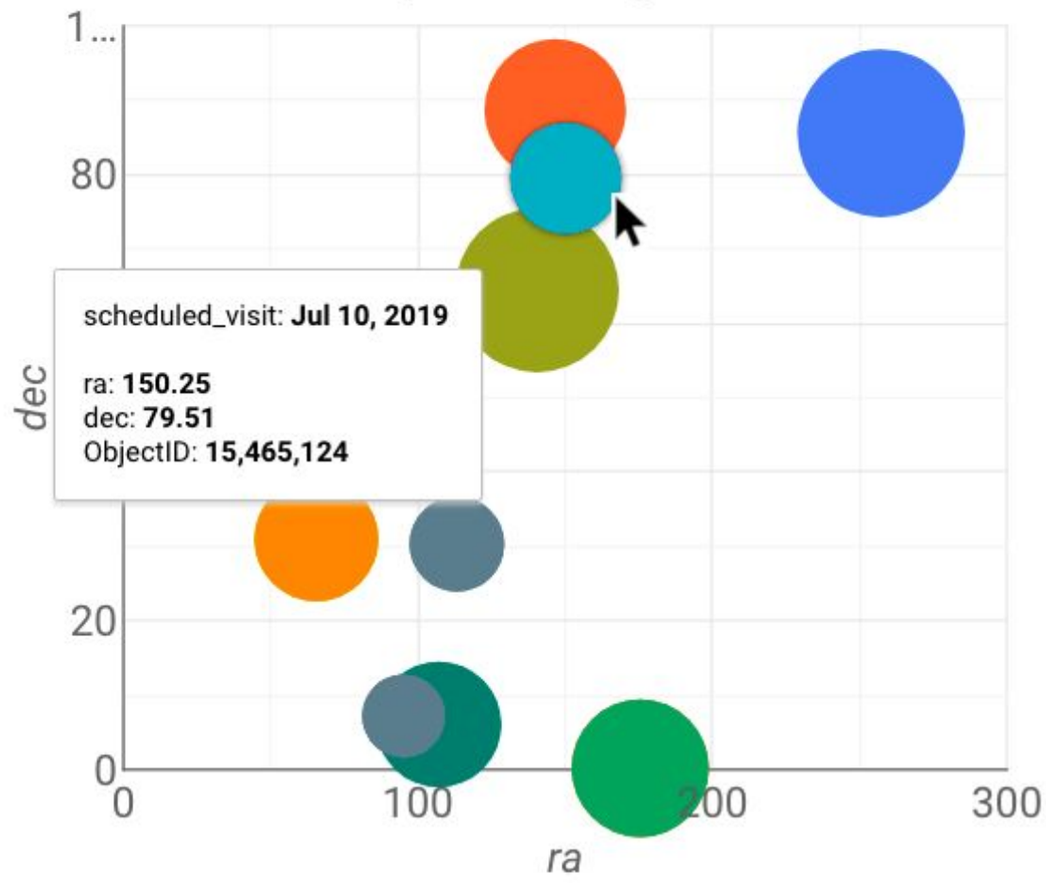
ra

BETWEEN ... AND ...

dec

BETWEEN ... AND ...

Upcoming Visits



Alerts

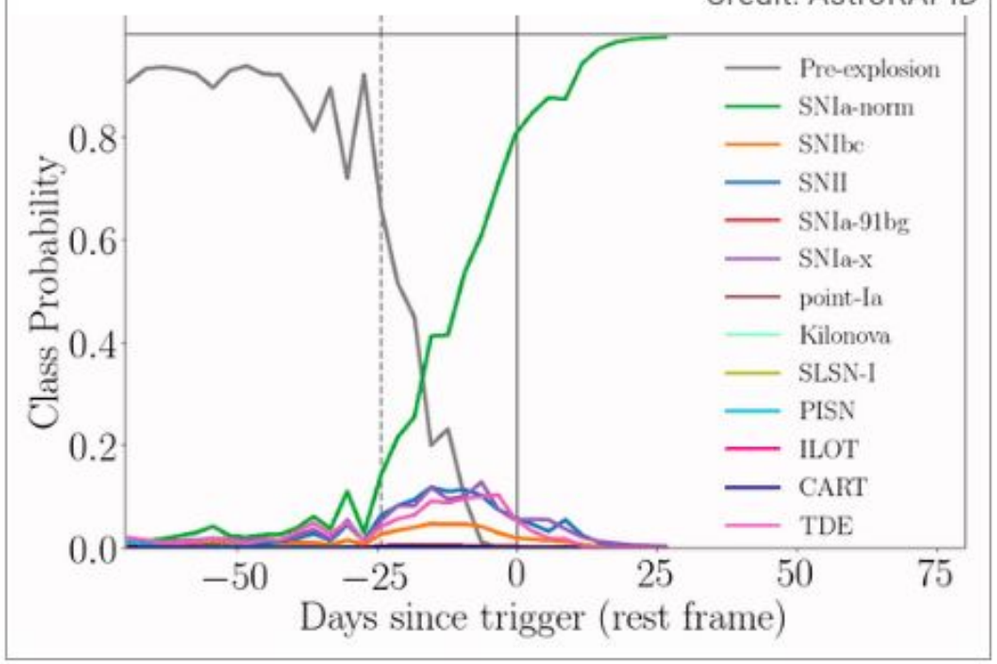
	ObjectID	visitID	visitDate	ra	dec	flux	fluxBand	pm	parallax
1.	8617405	48250083	Jul 17, 2019	95.2187355	7.2155175	0.199	6500	283	0.025
2.	15465124	83966235	Jul 17, 2019	150.2495306	79.5114188	0.821	8000	409	0.036

Example Object Page

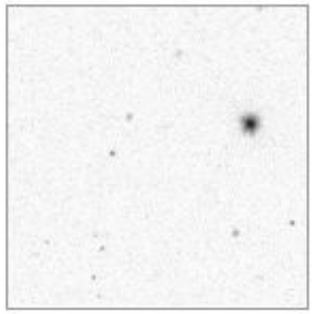
Search Object: Watchlist: ObjectID | Object ID: ObjectID: 19703395 (1) | Visit Date: visitDate | Add to Watchlist

Upcoming Visits
- Jul 23, 2019
- Aug 4, 2019
- Aug 8, 2019

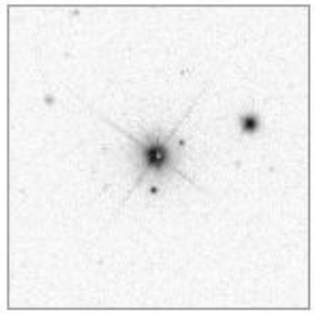
Classification



Template



Difference Image



Photometry

Cross Match

	catalog	xID	type	relat	relat_conf	ra	dec	gmag	rmag	imag	zmag
1.	GAIA	87516608643	star	self	91	107.0612705	6.0743112	13.94219971	13.22000027	12.83300018	12.58500004
2.	GAIA	86280984582	star	binary_comp	89	107.0612705	6.0743112	13.94219971	13.22000027	12.83300018	12.58500004

Visit Alerts

	ObjectID	visitDate	visitID	ra	dec	pm	parallax	flux	fluxBand
1.	19703395	Jul 7, 2019	48250083	107.0612705	6.0743112	39	0.015	0.921	6500

Discussion Questions

- Our raison d’être is essentially to collaborate and make this whole process easier for both consumers and developers.
- We aim to provide a cloud-based solution and framework
- We aim to provide streams to the world:
  raw, annotated, filtered
- We will have all of the postage stamps available. We haven’t thought about how to support processing of these images within GCP.