Scaling up data streams for asteroid discovery

Jeremy Kubica

Google Pittsburgh

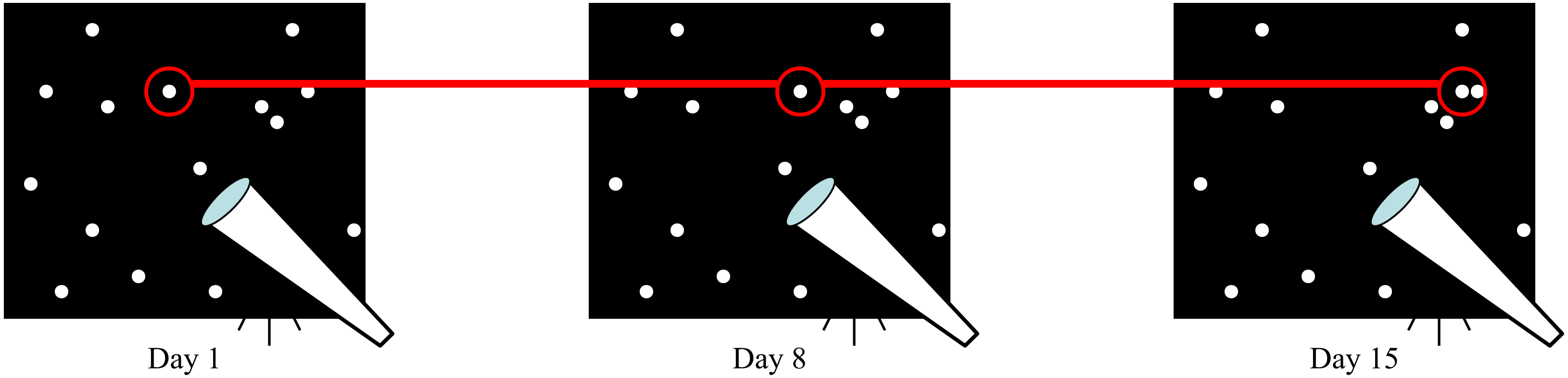
Overview

- Currently asteroid discovery is only using a portion of the data available.
- New data mining techniques can help:
  - Combine multiple noisy data sources and
  - Push into the noise to extract more signal from the current data.
  - **Drive new discoveries by allowing us to scale up the data streams.**

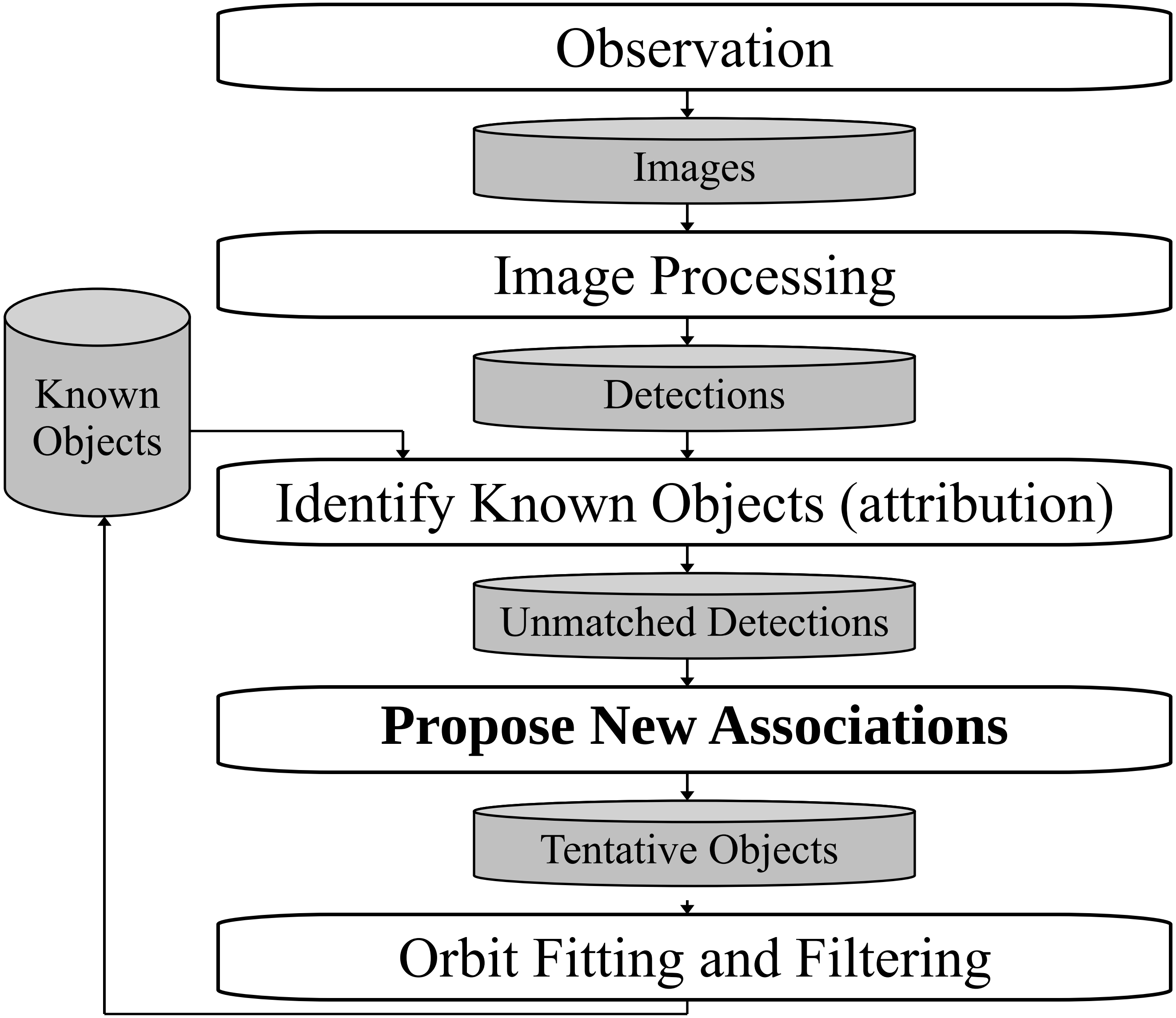
Asteroid Discovery

- Task: Asteroid discovery and tracking from images.
- Goals:
  - Associate individual *detections* in different images that correspond to the same true object.
  - Compute a trajectory or orbit for these objects.
- Find the “best” set of orbits or all orbits meeting some criteria:

$$\frac{1}{N}\sum(x_i - orbit(t_i)) < e \qquad \underset{orbit}{\arg\max}\, P(\overset{r}{x} \mid orbit)$$



Asteroid Tracking Pipeline

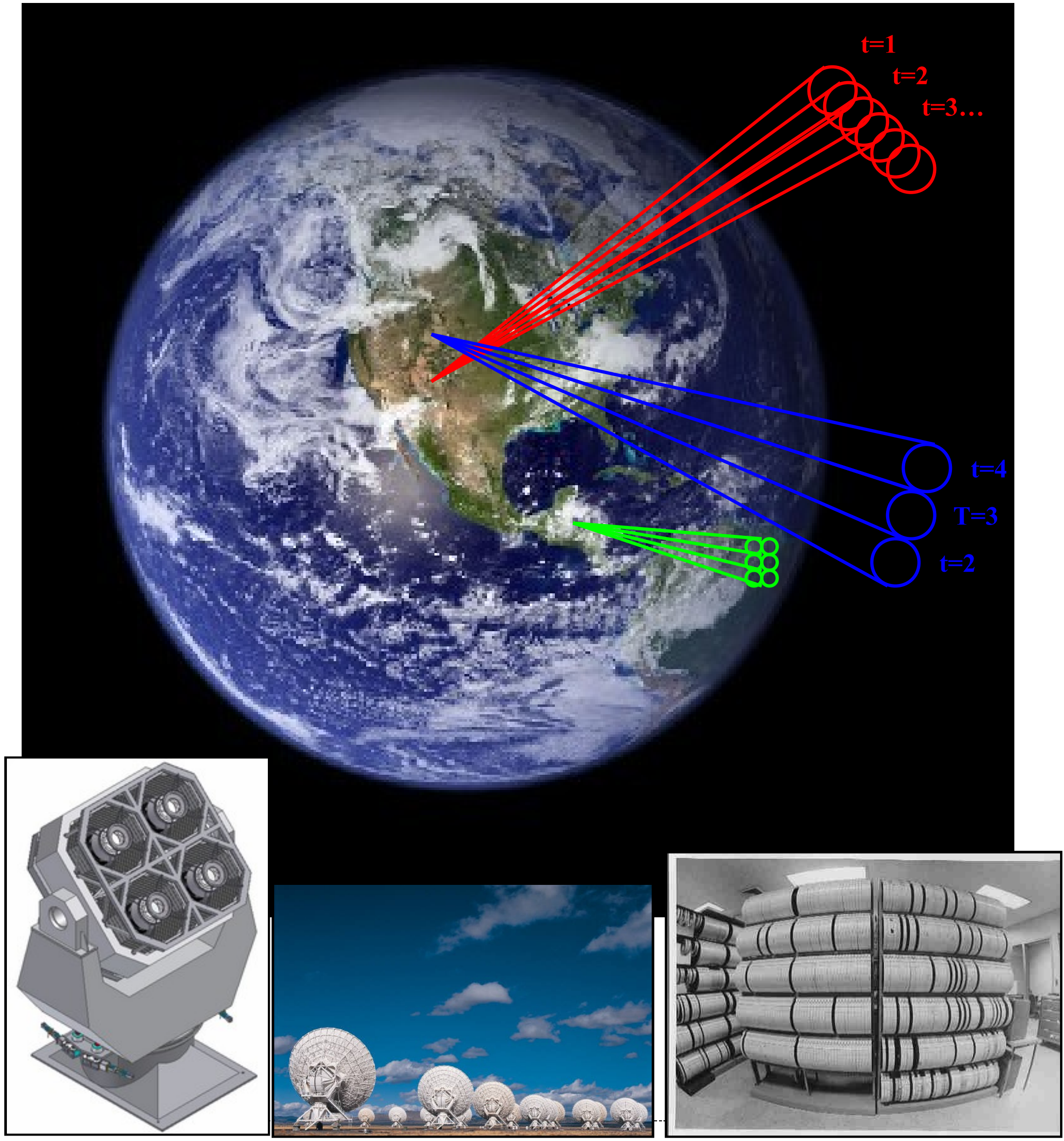


Using **more** data from survey images

With the efficient algorithmic techniques we can make better use of the data:

- – Singleton observations
- – Negative “observations”
- – Stacked images
- – Larger gaps in time

Combining Massive Data Sets



- **Combine data from many surveys including historical.**
- **Good News**: Increase coverage (chance of seeing an asteroid).
- **Bad News**: Replaced a massive data steam with multiple massive data sets.
- This is a promise of the NVO.

Long Tail Data Streams

We can go further and augment the deep, systematic coverage of the surveys with long tail data sources.
Examples: Amateur astronomers, Mars rover.

Non-survey data sources:

- Can provide additional coverage and breadth.
- Can provide a source of “lucky” supporting detections.
- Cannot go “deep” for faint objects.
- Have very noisy data with many unknown parameters (e.g. camera).
- Have uncoordinated schedules.

Pushing into the Noise

Current asteroid linkage pipelines start by extracting significant detections - throwing away large amounts of potential data.

- Pushing into the noise:
  - We can push into fainter detections: 3 sigma → 5+ sigma
  - Push the tracking to the raw pixels.
- Data explosion - non-linear scaling.
- Massive noise.

Challenges

- Combining terabyte data streams:
  - A new 10x in scalability.
  - In memory algorithms become infeasible.
- Unreliable data:
  - Noisy,
  - Incomplete features (e.g. colors)
  - Irregular (and unplanned) observation cadence,
  - Heterogeneous observation (instrument) parameters.
- **Core challenge: How can we best make use of a vast amount of highly unreliable data.**

Promises

- **Key promises: Much more data and better signals from each piece of data.**
- Allow us to push into the noise - finding fainter and further objects.
- Provides additional “lucky” supporting observations.
- Provides better coverage than current survey cadence.

Machine Learning Directions

- **Effectively scaling up the data streams will require new data mining advances.**
- Statistical models to push through the noise.
- Online probabilistic noise models to capture (undocumented) instrument, environment effects.
- Highly efficient algorithms, including: online, streaming, and distributed algorithms.