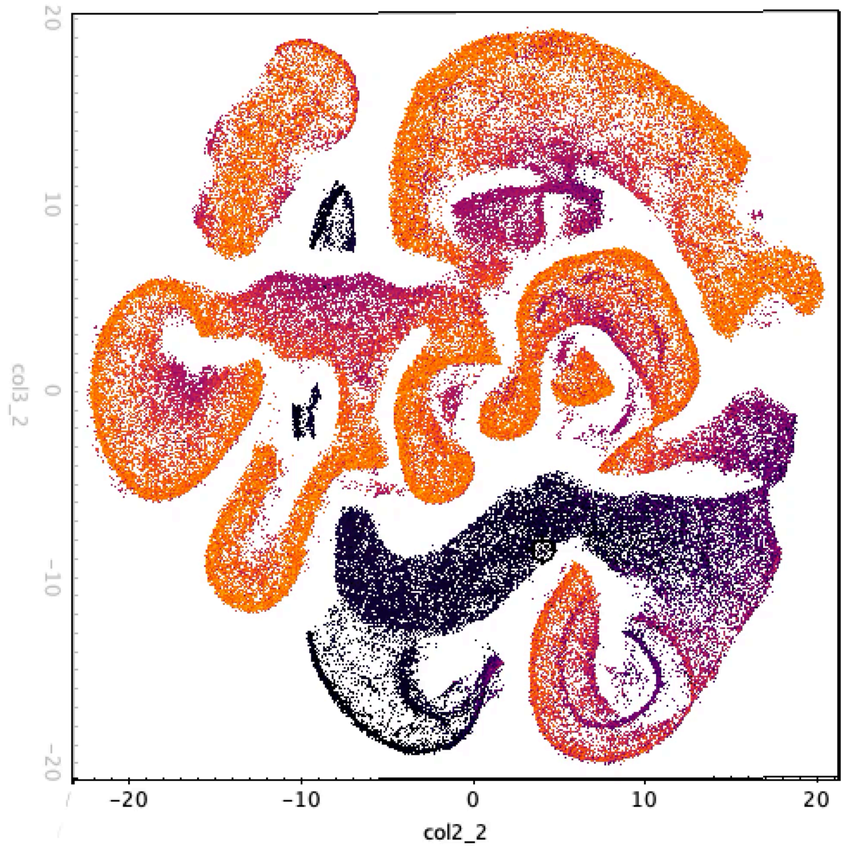
Harvesting outliers: data barriers to turn anomalies into discoveries

Rafael Martínez-Galarza

CENTER FOR ASTROPHYSICS
HARVARD & SMITHSONIAN

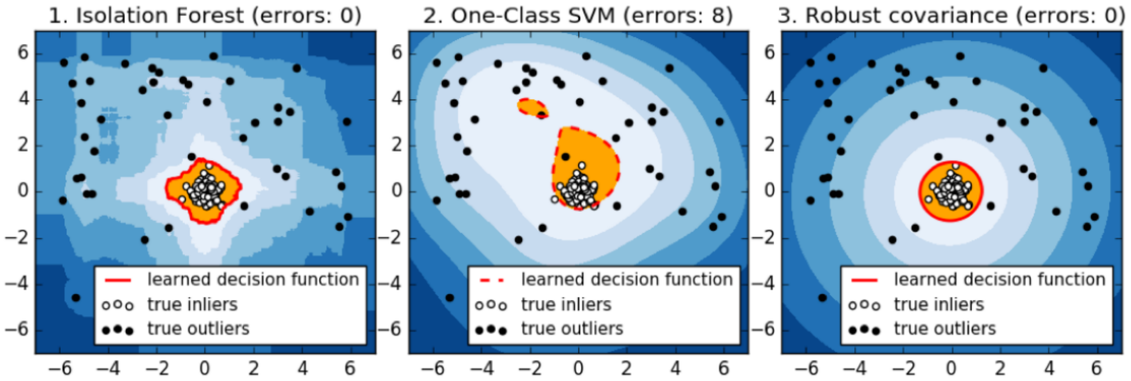


KG-ID session. IVOA. InterOp meeting. November 3, 2021

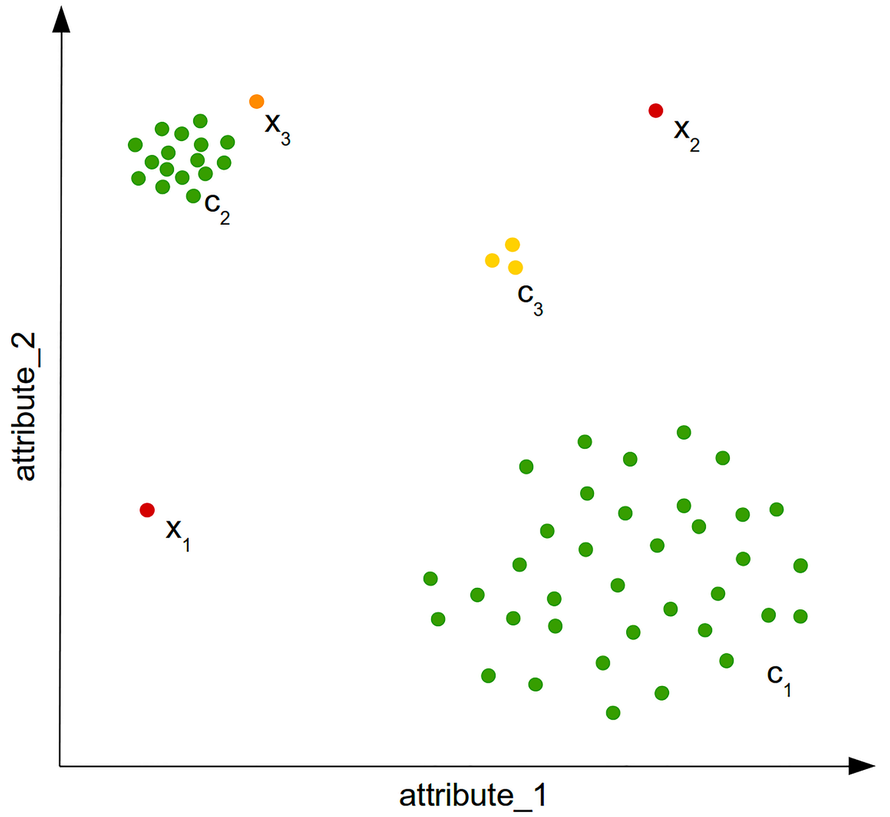
What is an anomaly?

"An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism"
-- Hawkins (1980)

Andreas C Mueller



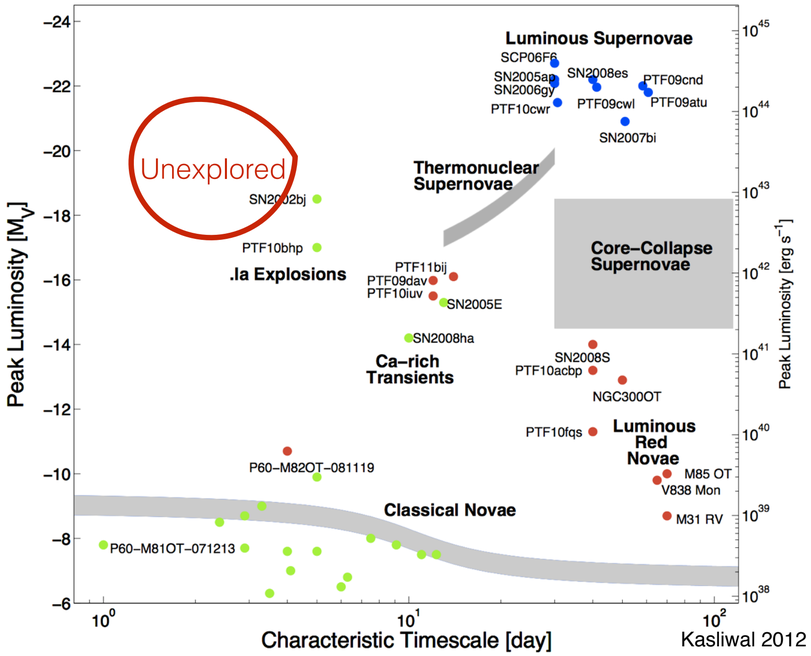
- "An outlying observation, or *outlier*, is one that appears to deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969)
- Anomalies are:
  - Different from the norm with respect to their features
  - Rare in a dataset compared to other instances
- **Anomalies are scientifically interesting objects: they are often not explained by current models, and could lead to new discoveries.**



Goldstein & Uchida, 2016

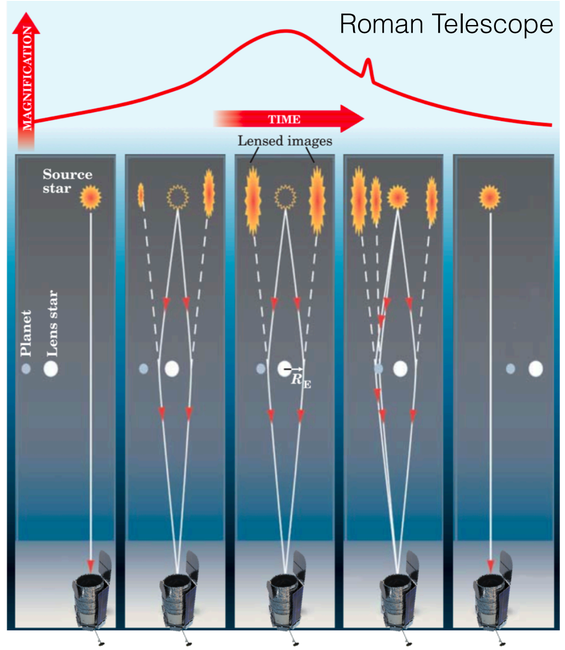
Why do we care about anomalies?

Exhibit A



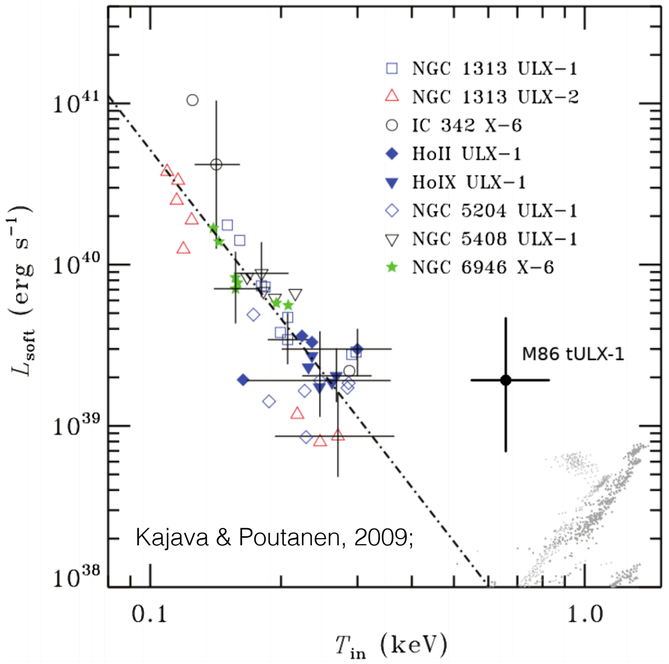
- New types of transients expected from last generation surveys.
- Fast and luminous transients of particular interest for gravitational wave astronomy.
- How do we identify those anomalies so that we can follow them up?

Exhibit B



- Anomalous light curves in micro-lensing surveys.
- Can reveal both planets and massive objects (stellar-mass BH) in orbit around MS stars.
- Of relevance for the Roman Telescope's microlensing survey

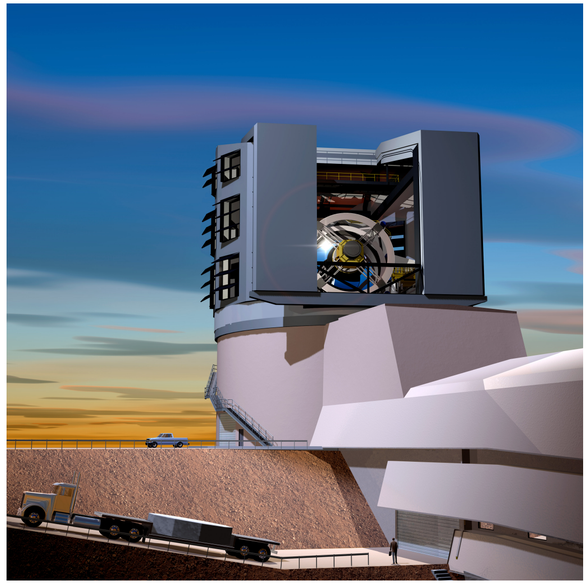
Exhibit C



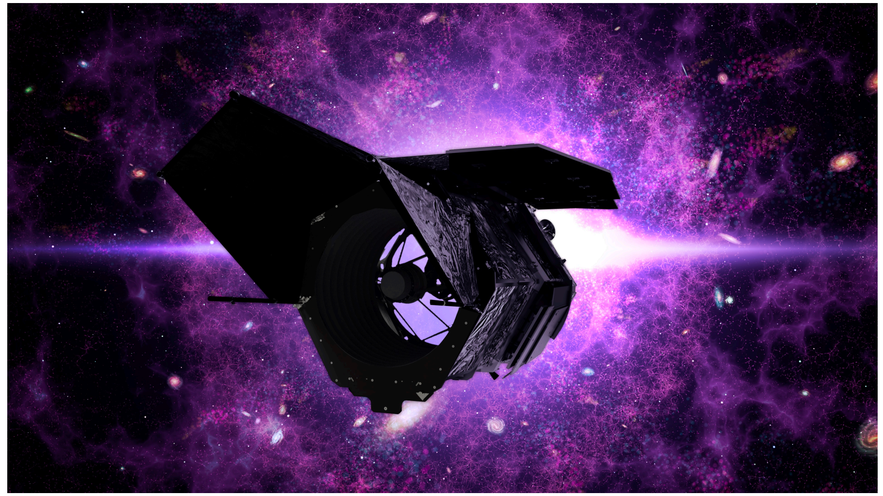
- High energy anomalies can reveal phase transition in X-ray binaries.
- This includes ULX that might signal IMBHs, as well as TDEs.
- Relevant for the eROSITA survey.

Datasets

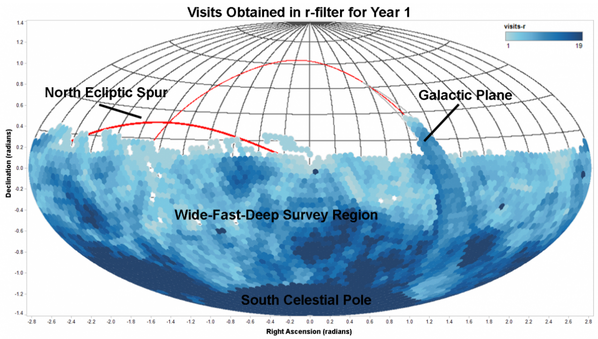
LSST

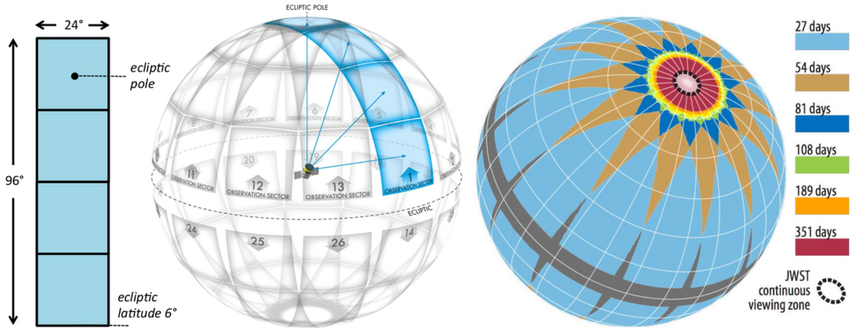


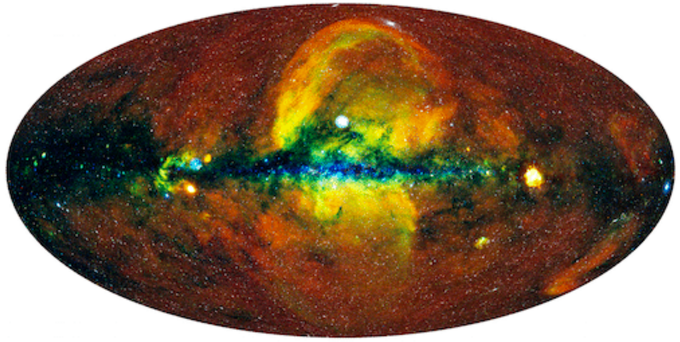
Kepler/TESS/Roman



Chandra, XMM, eROSITA



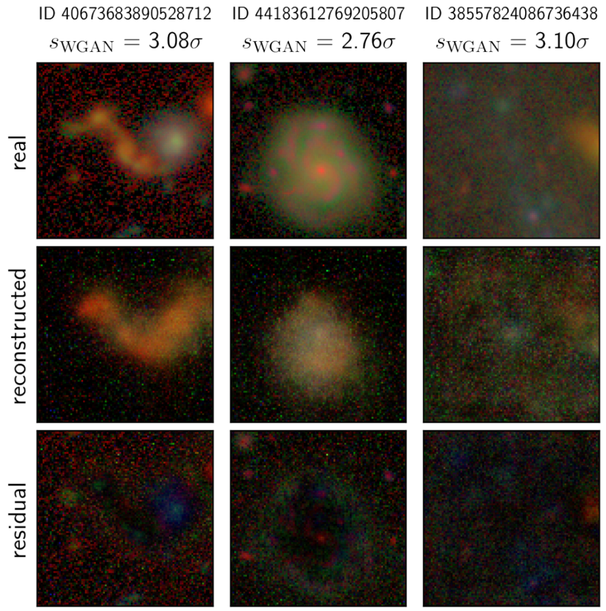




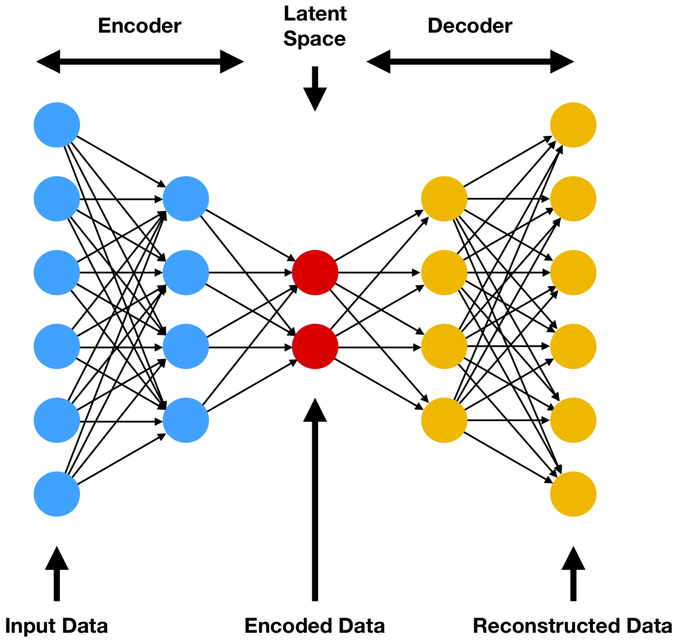
Unexplored discovery space. Majority of sources have not been classified, or in some cases even detected before.

Datasets with “meaningless” axes

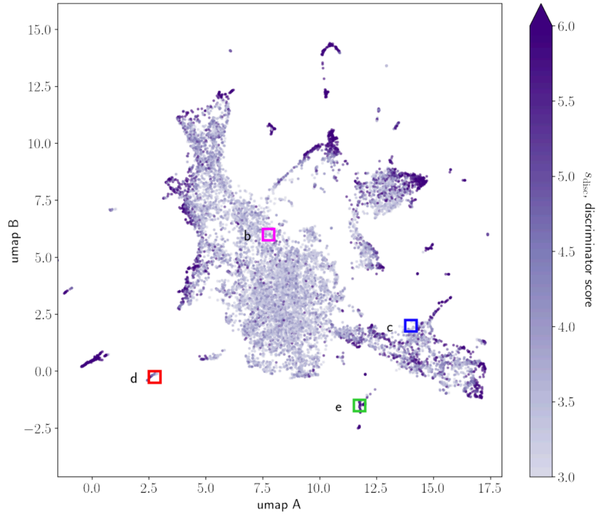
Visualizing anomalies requires dimensional reduction



Observable features
(In this case, pixels)



Embedded or latent
Features (we do not know
what they mean)



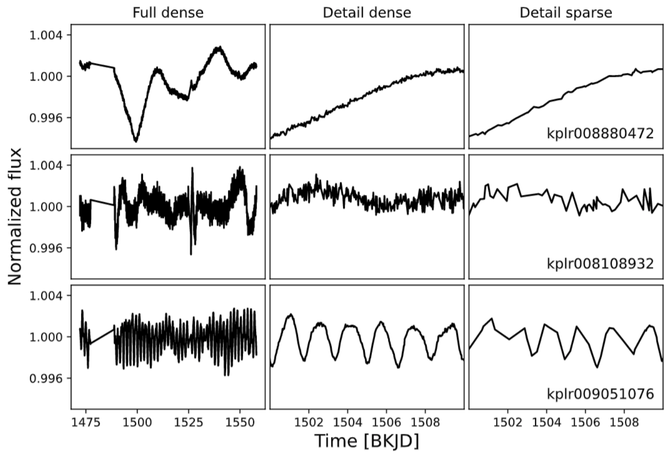
Visualization in
Lower dimension

Storey-Fisher et al. 2021

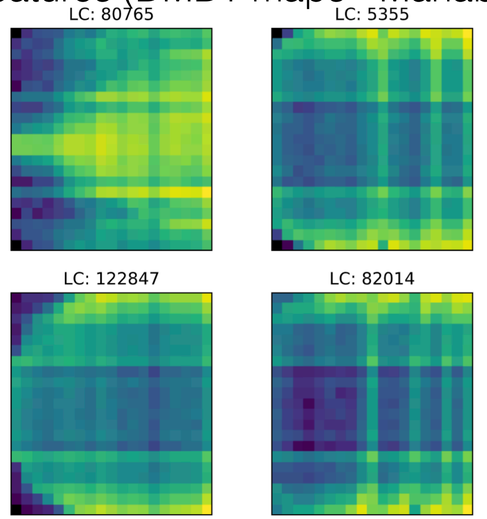
Given a LC of interest, can we find similar objects?

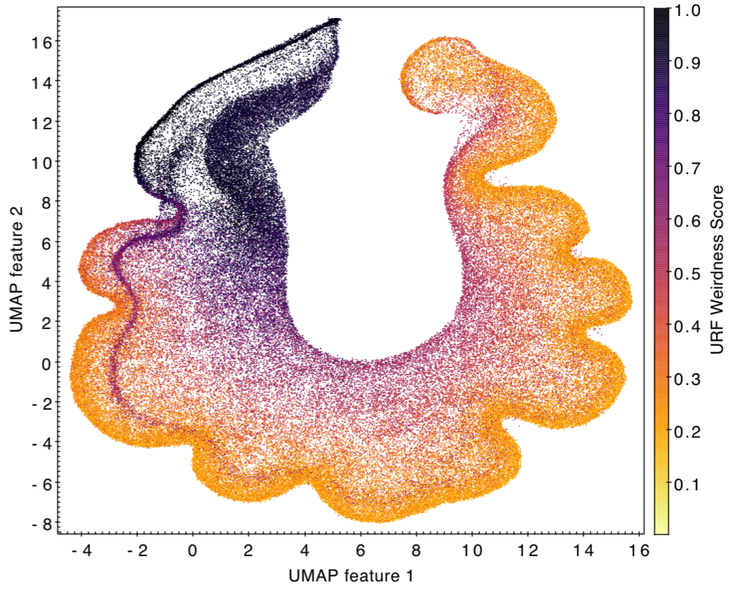
- Dimensionality reduction + DMDT maps
- t-SNE and UMAP embeddings correlated with independently computed URF score.
- We use the low-dim space of these embeddings to break the degeneracy of the 1-D URF score.

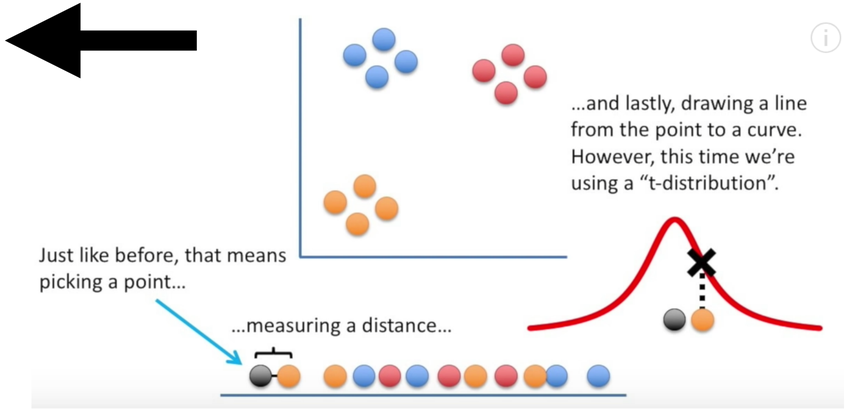
A Kepler dataset (pre-processed)



Extract features (DMDT maps - Mahabal et al.)







StatQuest

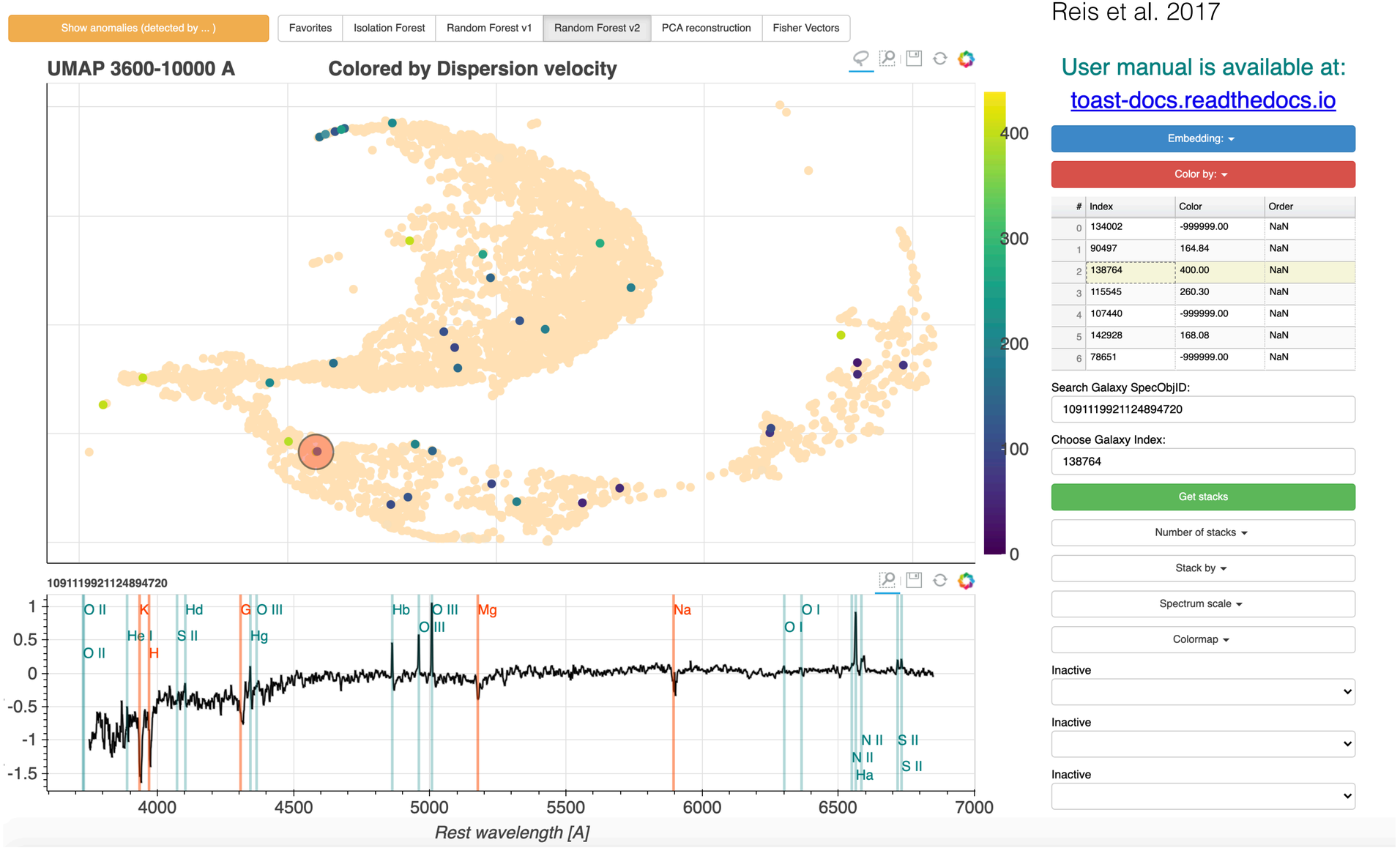
Martínez-Galarza+ 2021

Making sense of anomalies

https://galaxyportal.space/

Reis et al. 2017

User manual is available at:
toast-docs.readthedocs.io



Some ideas/questions

- What kind of data infrastructures should we consider to support anomaly investigations?
- Ideally we want to visualize how the “embeddings” relate to physical quantities, so that domain knowledge experts can select regions of the “embedded” space where anomalies are likely.
- From yesterday’s CSP panel, Gregory Dubois-Felsmann (IPAC): How to better use data that we don’t know can be useful to our research?
- Now: how do we use features whose meaning is not intuitive in order to make discoveries? How can IVOA help?