Unsupervised Machine Learning

Umaa Rebbapragada, Ph.D.
Machine Learning and Instrument Autonomy Group

LSSTC Data Science Fellowship Program
Wednesday, November 7, 2018
Northwestern University

Research described in this presentation was carried out at the Jet Propulsion Laboratory under a Research and Technology Development Grant, under contract with the National Aeronautics and Space Administration.

Jet Propulsion Laboratory
California Institute of Technology

Outline

- Overview
- Unsupervised Learning Ingredients
- Clustering
- Anomaly Detection
- Summary
- Data for Day 2, 3 Hands-on Exercises

Overview

Unsupervised Learning

- Learning from data in absence of rewards (reinforcement learning) or labels (supervised learning)
- Major sub-types:
  - Clustering
  - Anomaly Detection
  - Dimensionality Reduction
  - Density Estimation

Unsupervised Learning

- Learning from data in absence of rewards (reinforcement learning) or labels (supervised learning)
- Major sub-types:
  - Clustering
  - Anomaly Detection
  - ~~Dimensionality Reduction~~
  - ~~Density Estimation~~

Unsupervised Learning

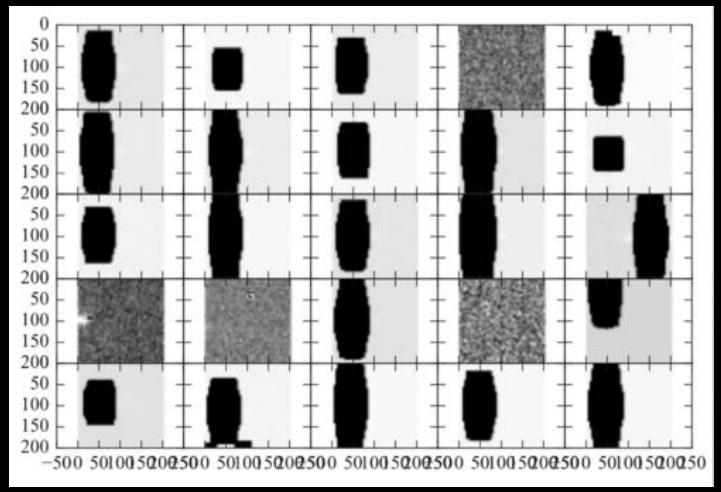
- Learning from data in absence of rewards (reinforcement learning) or labels (supervised learning)
- Major sub-types:
  - Clustering
  - Anomaly Detection
  - ~~Dimensionality Reduction~~
  - ~~Density Estimation~~

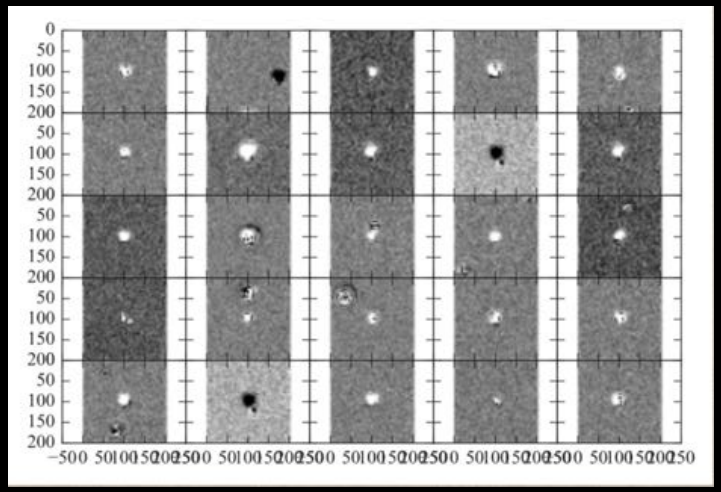
PCA, manifold learning (IsoMap)

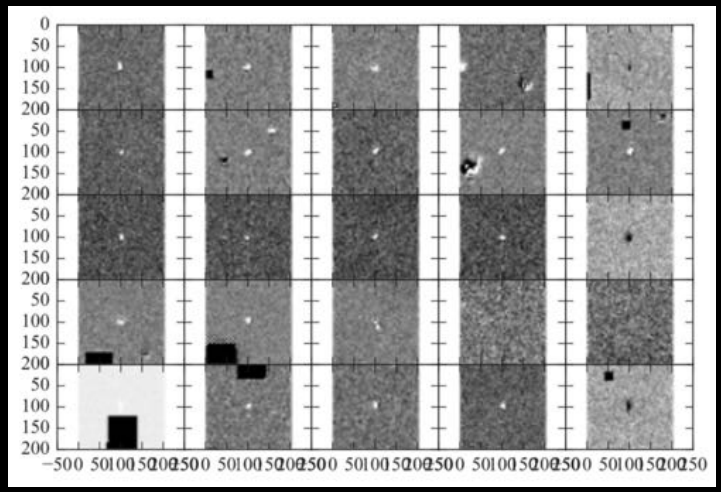
Finding an underlying probability density function

Clustering Example

Understanding Artifacts in ZTF Image Subtractions



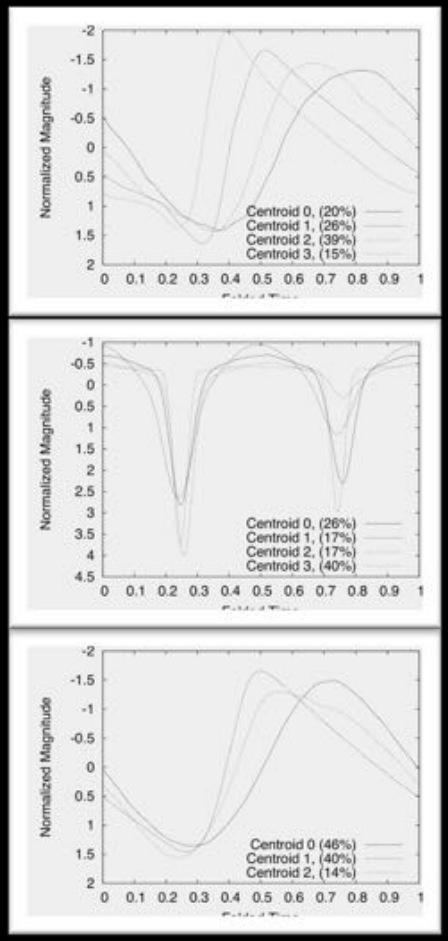




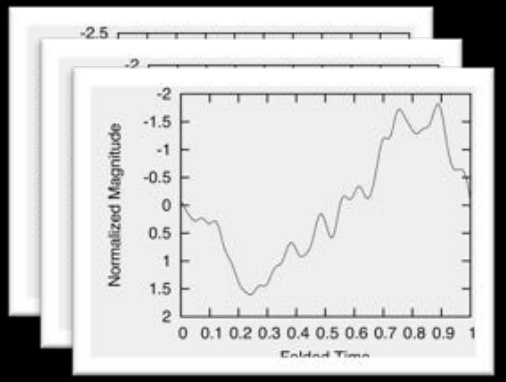
Clustering an early version of training data revealed major classes of artifacts.

Anomaly Detection

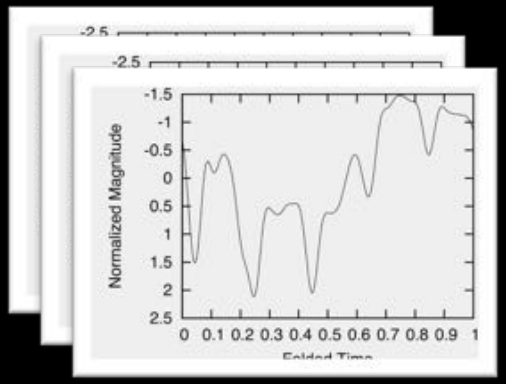
Finding Anomalous Lightcurves in Catalogs of Periodic Variables

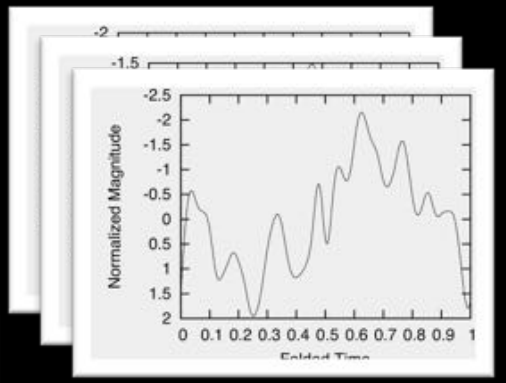


Cluster Centroids (examples of normality)



Anomalies found with respect to these cluster centroids

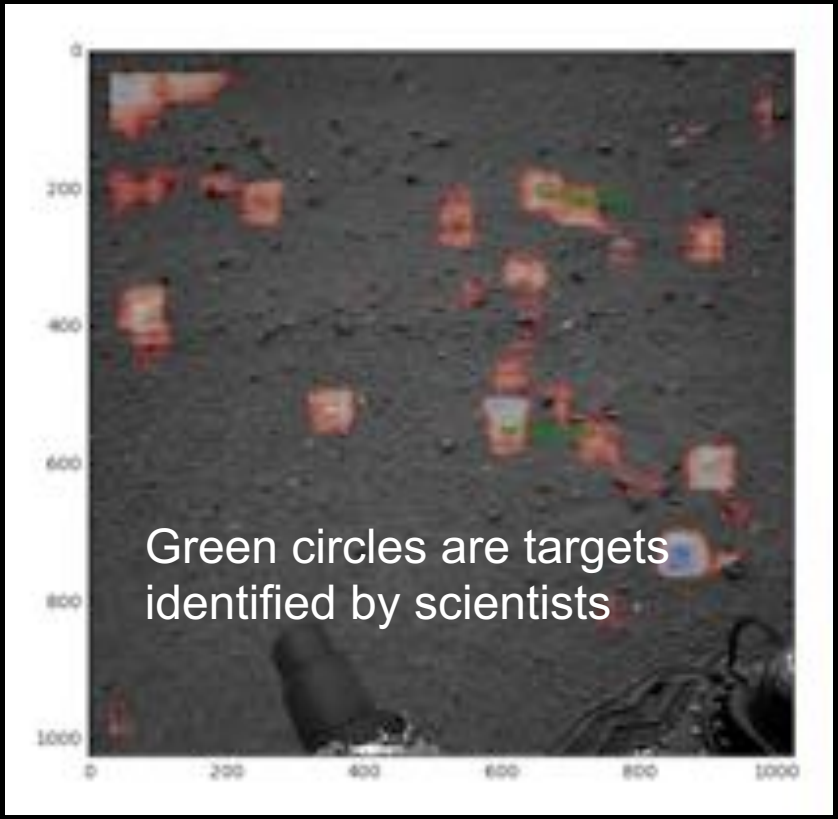


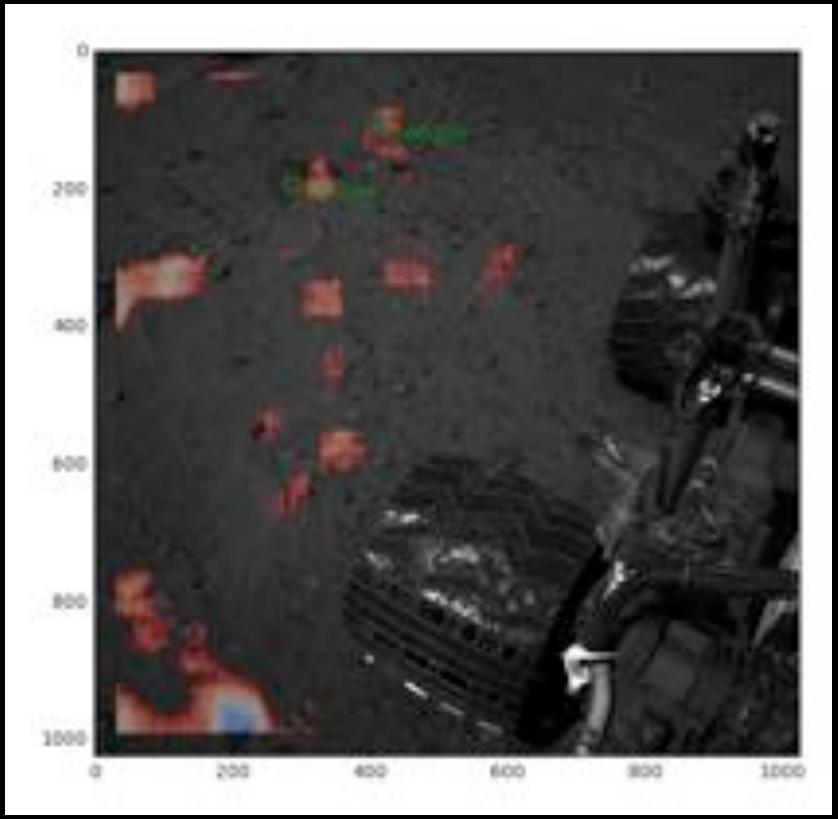


Novelty Detection on MSL Imagery

Navigation Camera Images

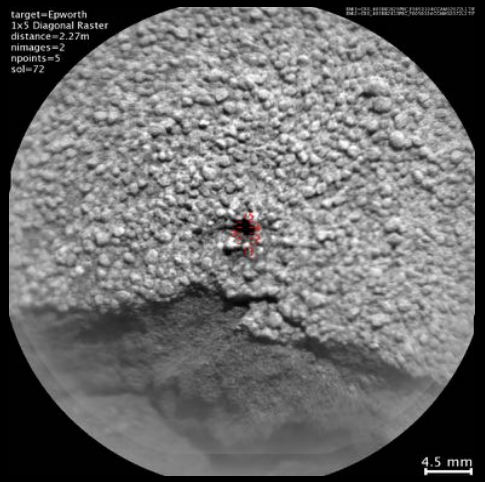
Anomalies identified using Isolation Forest

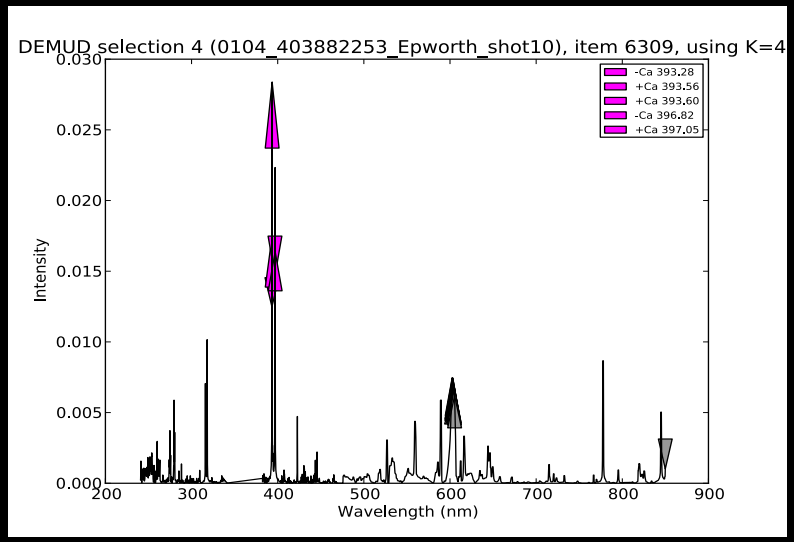




Dimensionality Reduction

Discovery via Eigenbasis Modeling of Uninteresting Data (DEMUD)





A DEMUD result (center) on ChemCam data taken on a soil sample at target Epworth (left). DEMUD found an unexpectedly high occurrence of Ca in this sample (magenta triangles), which turned out to correspond to a scientifically interesting detection of the mineral CaF (grey triangles).

DEMUD uses **singular value decomposition** to model normality in the dataset.

Unsupervised Learning Ingredients

Data

Columns

Rows

	# Pixels	Axis Length	Half Width	Median Flux	…
1	40	17.97	1.36	14.0	
2	49	16.77	2.00	13.0	
3	52	21.20	1.29	13.9	
4	92	32.42	0.86	24.2	
5	233	44.28	1.20	26.1	
6	61	13.25	1.37	170.3	
7	47	16.15	0.98	24.2	
8	120	25.71	1.01	119.7	
9	62	13.95	1.42	44.3	
10	180	29.09	1.35	19.9	
.					
.					
.					
N					

Data

Features

Examples

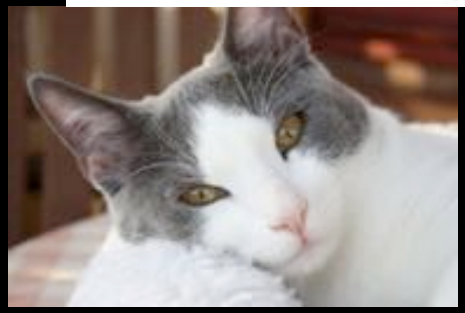
	# Pixels	Axis Length	Half Width	Median Flux	…
1	40	17.97	1.36	14.0	
2	49	16.77	2.00	13.0	
3	52	21.20	1.29	13.9	
4	92	32.42	0.86	24.2	
5	233	44.28	1.20	26.1	
6	61	13.25	1.37	170.3	
7	47	16.15	0.98	24.2	
8	120	25.71	1.01	119.7	
9	62	13.95	1.42	44.3	
10	180	29.09	1.35	19.9	
.					
.					
.					
N					

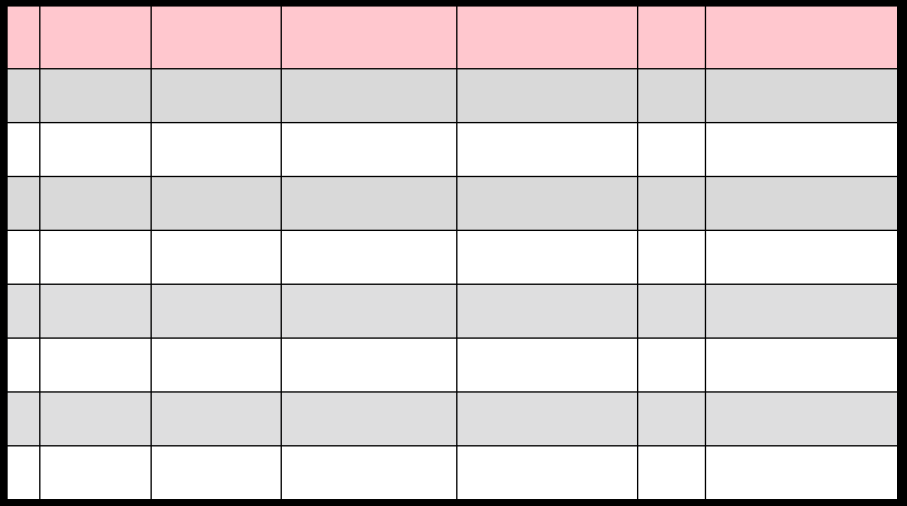
Representing Data

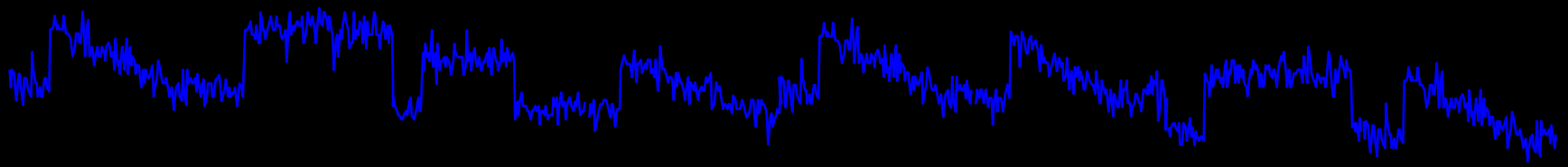
Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Sed at turpis vitae velit euismod aliquet. Pellentesque et arcu. Nullam venenatis gravida orci. Pellentesque et arcu. Nam pharetra. Vestibulum viverra varius enim.

Nam laoreet dui sed magna. Nunc in turpis ac lacus eleifend sagittis. Pellentesque ac turpis. Aliquam justo lectus, iaculis a, auctor sed, congue in, nisl. Aenean luctus vulputate turpis. Mauris urna sem, suscipit vitae, dignissim id, ultrices sed, nunc.

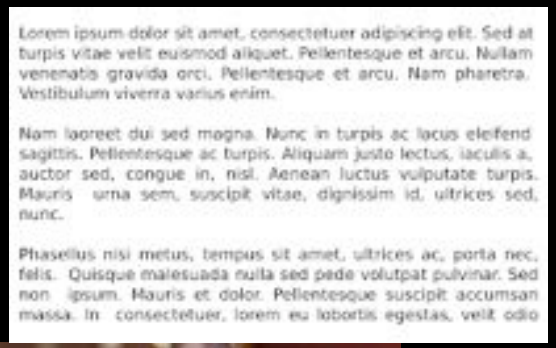
Phasellus nisi metus, tempus sit amet, ultrices ac, porta nec, felis. Quisque malesuada nulla sed pede volutpat pulvinar. Sed non ipsum. Mauris et dolor. Pellentesque suscipit accumsan massa. In consectetuer, lorem eu lobortis egestas, velit odio

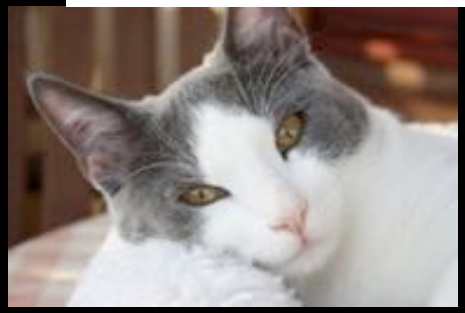






Representing Data

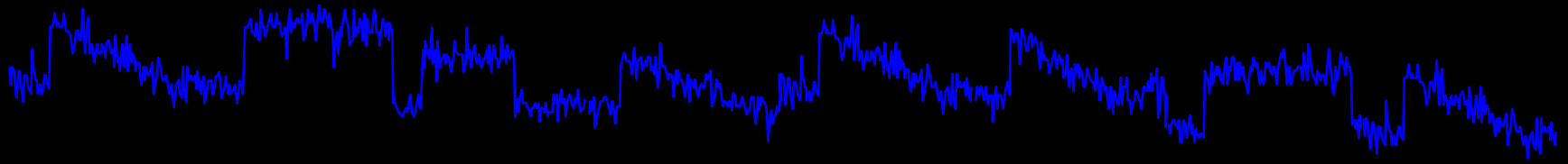




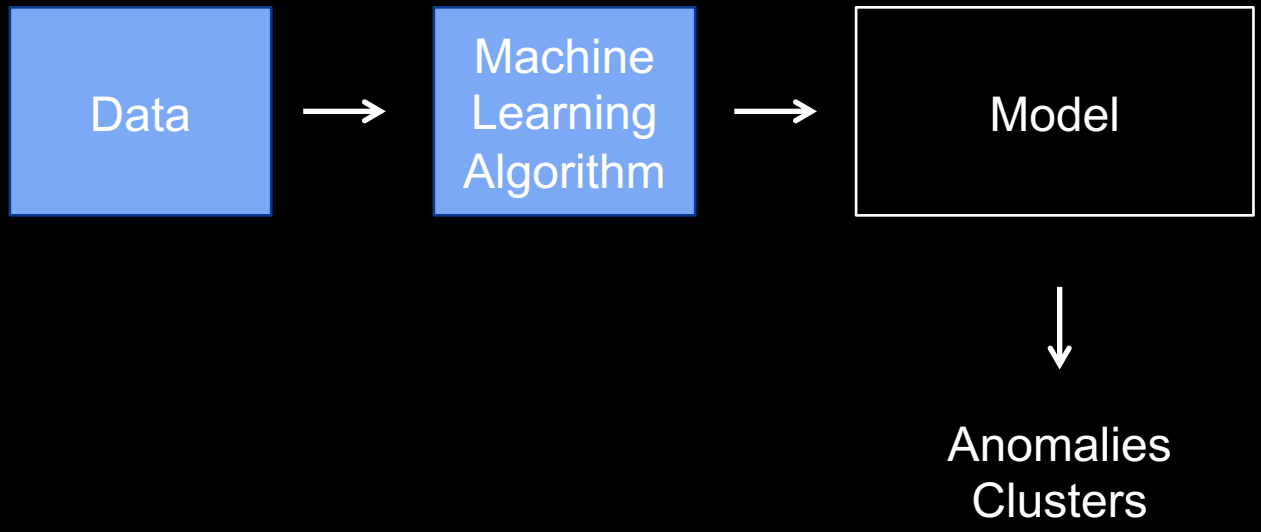
Bag of Words
TFIDF

Pixel values,
SIFT, HoG,
histograms of visual words

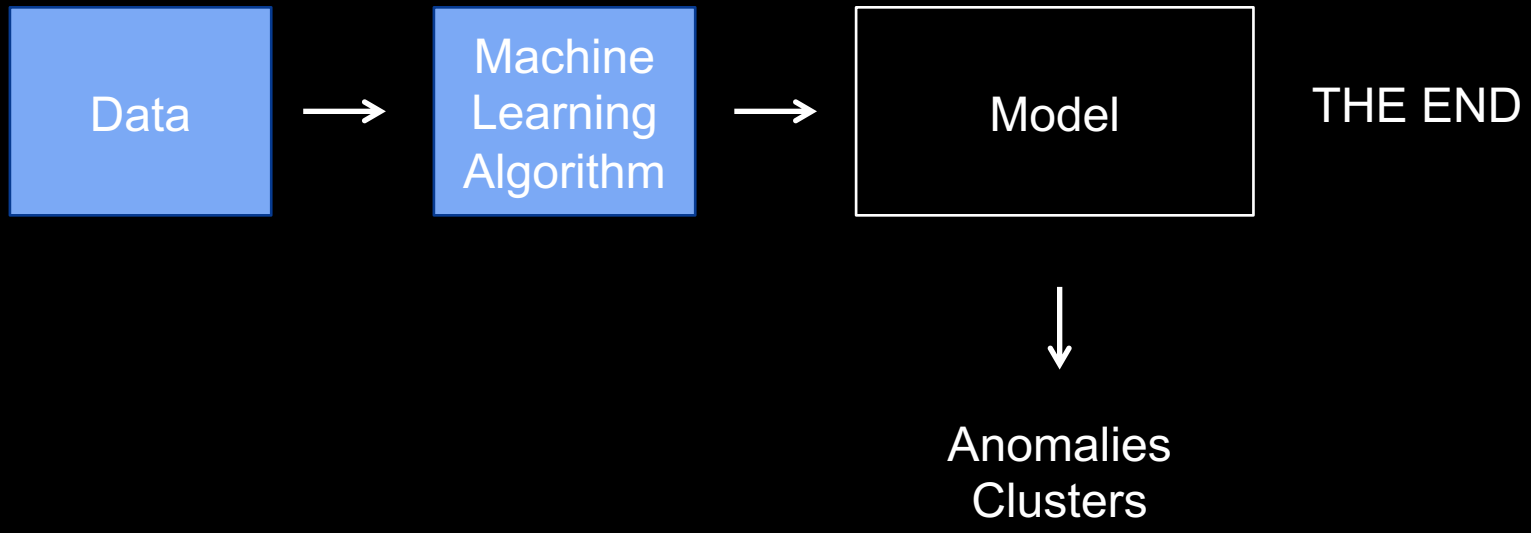
DFT, wavelets,
time series statistics



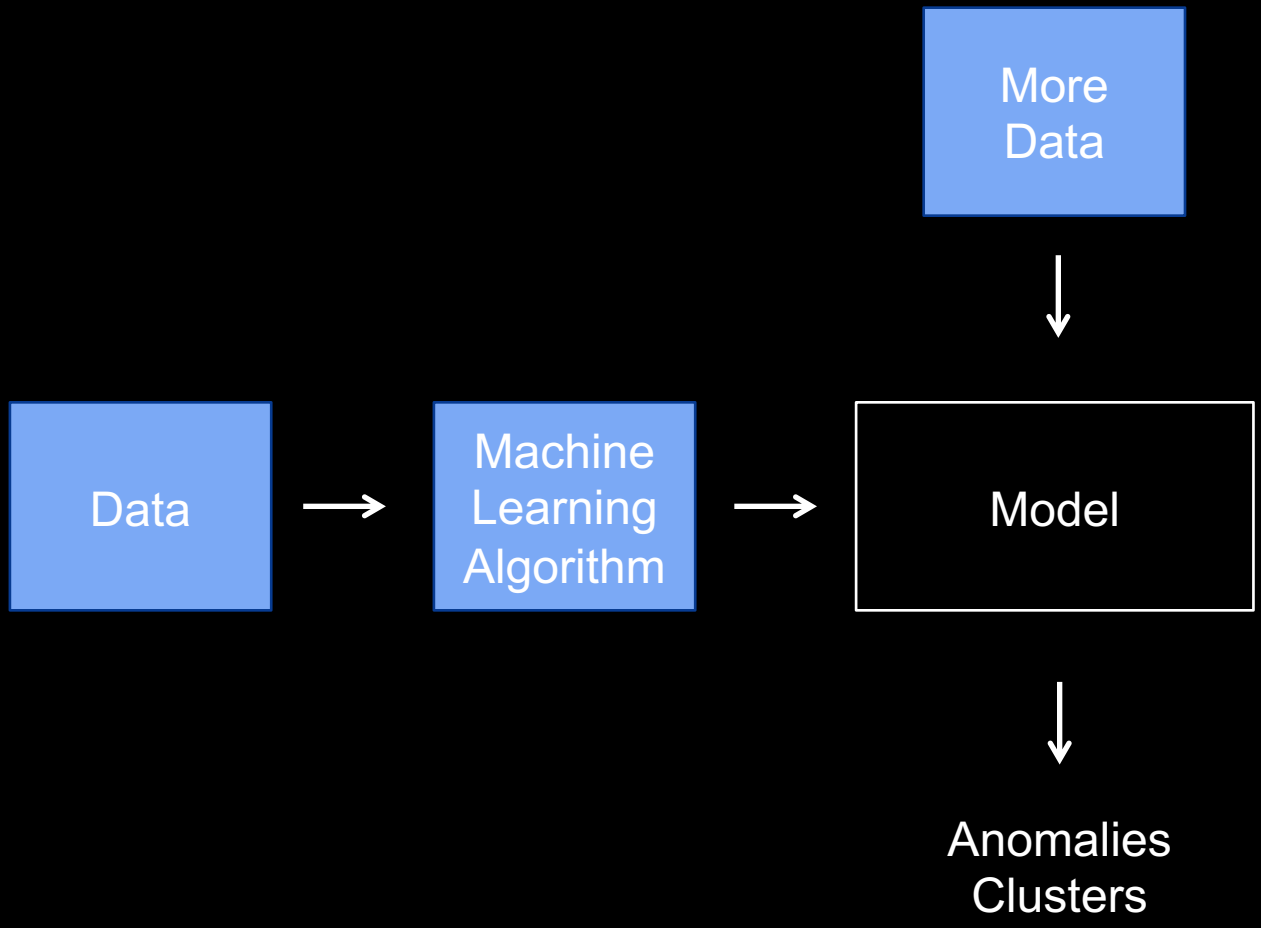
Unsupervised Learning



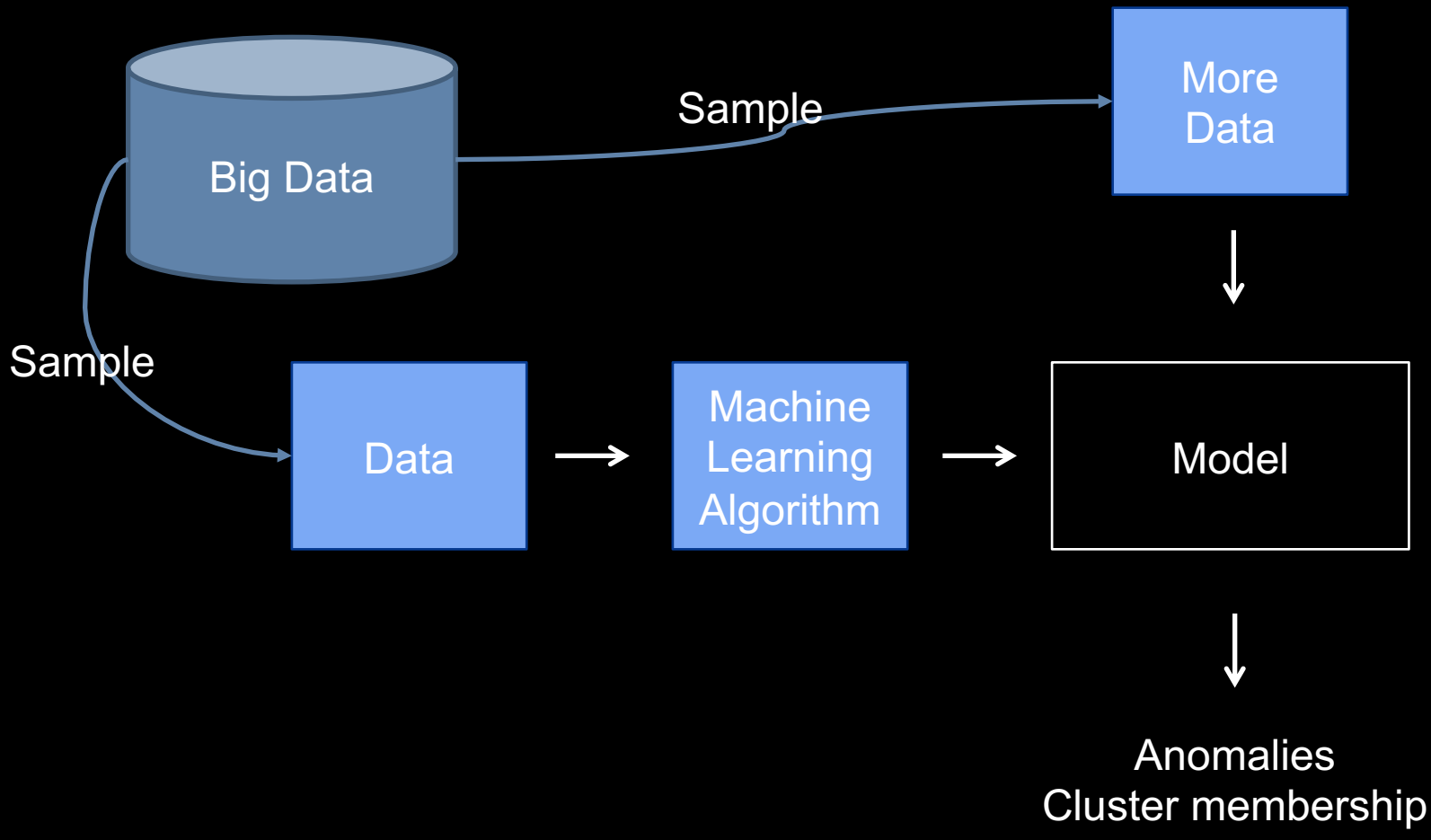
Unsupervised Learning



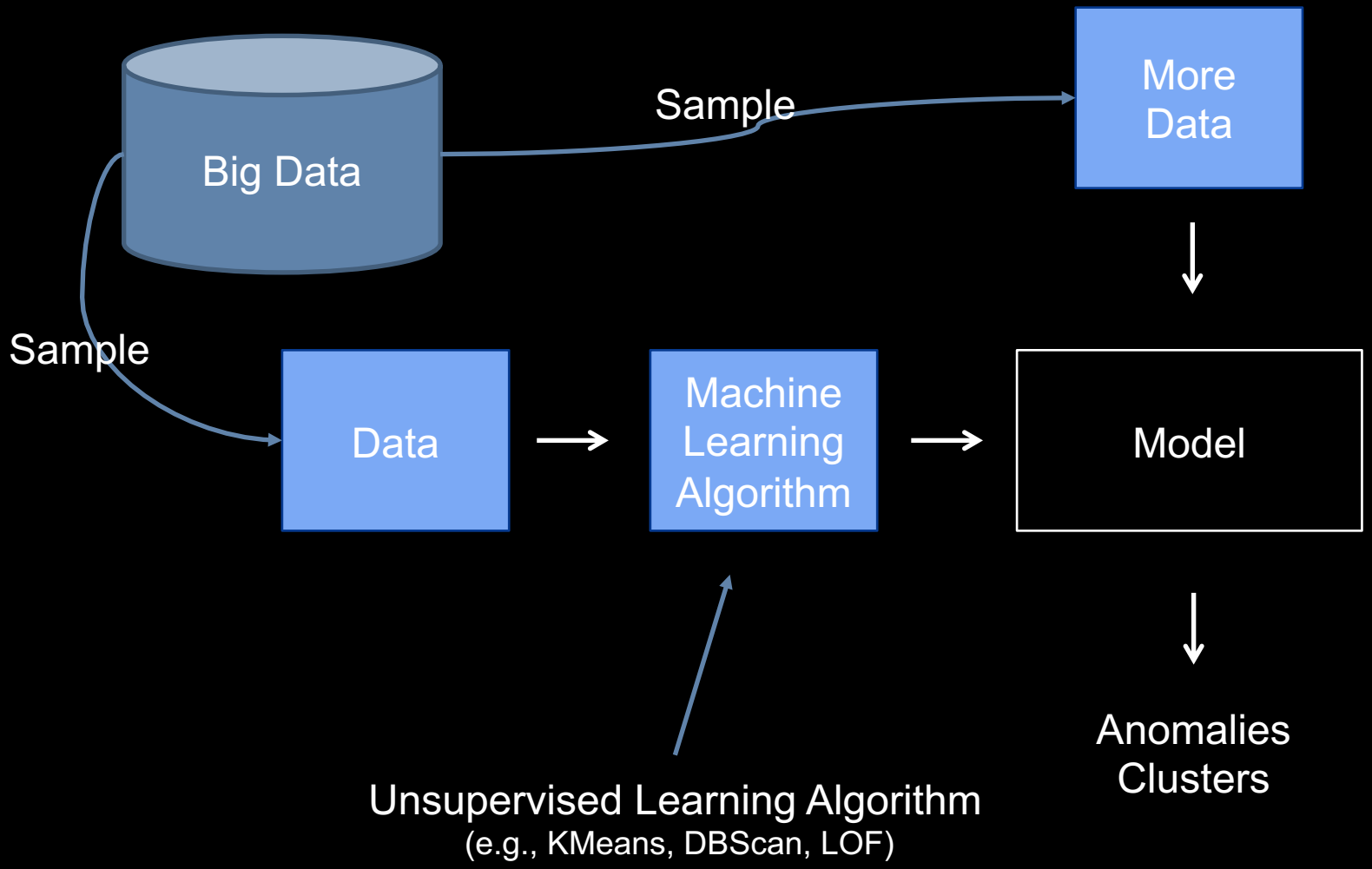
Unsupervised Learning



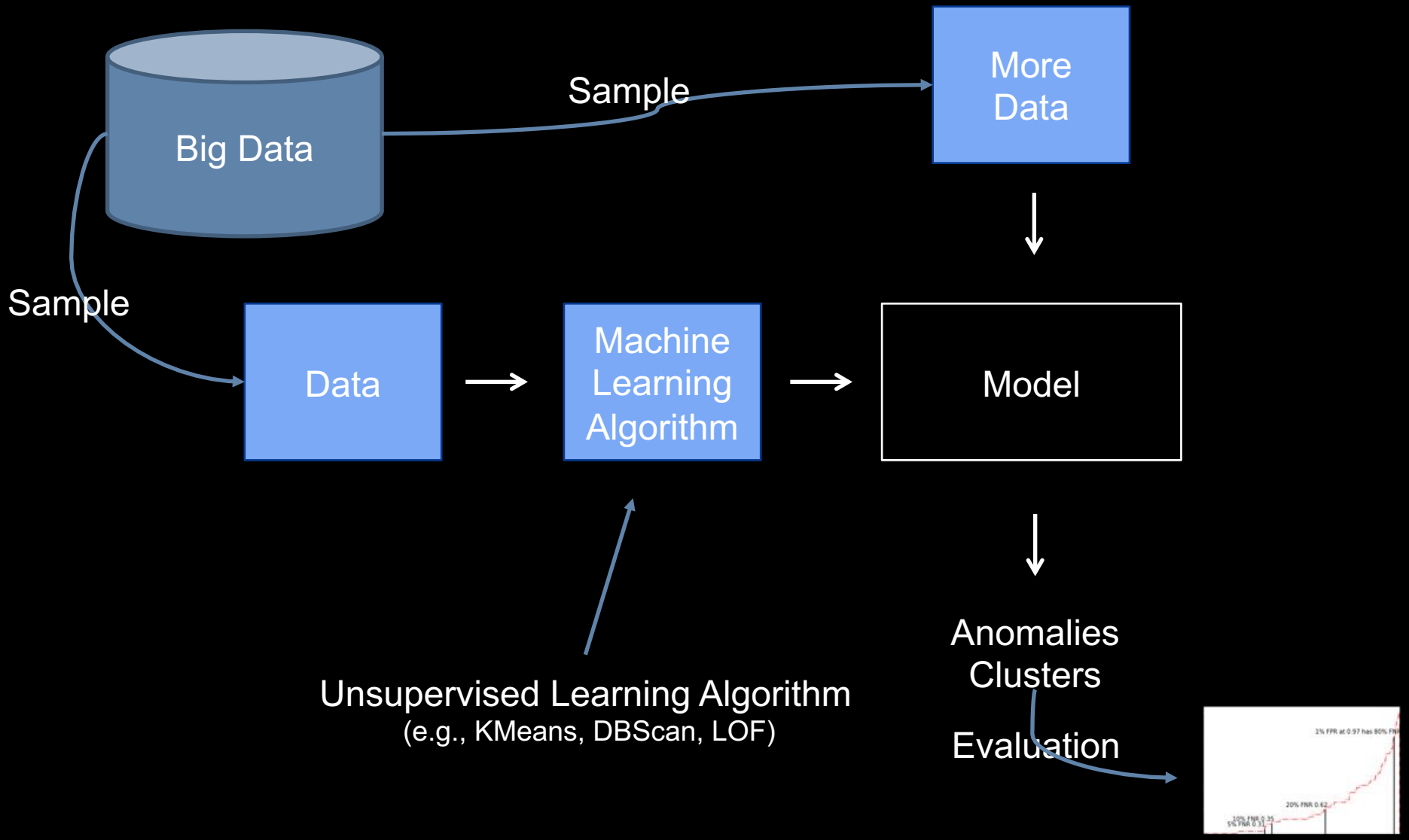
What are the Ingredients?



What are the Ingredients?



What are the Ingredients?



Ingredients Summarized

- Data Sampling
- Feature Representation
- Learning Algorithm
- Evaluation Metric

Ingredients Summarized

- Data Sampling - TOMORROW
- Feature Representation
- Learning Algorithm
- Evaluation Metric – Hand’s On Activity

Clustering

Types of Clustering

- Partitioning
- Density-based
- Hierarchical
- Model-based – Expectation Maximization (EM)

Partitioning

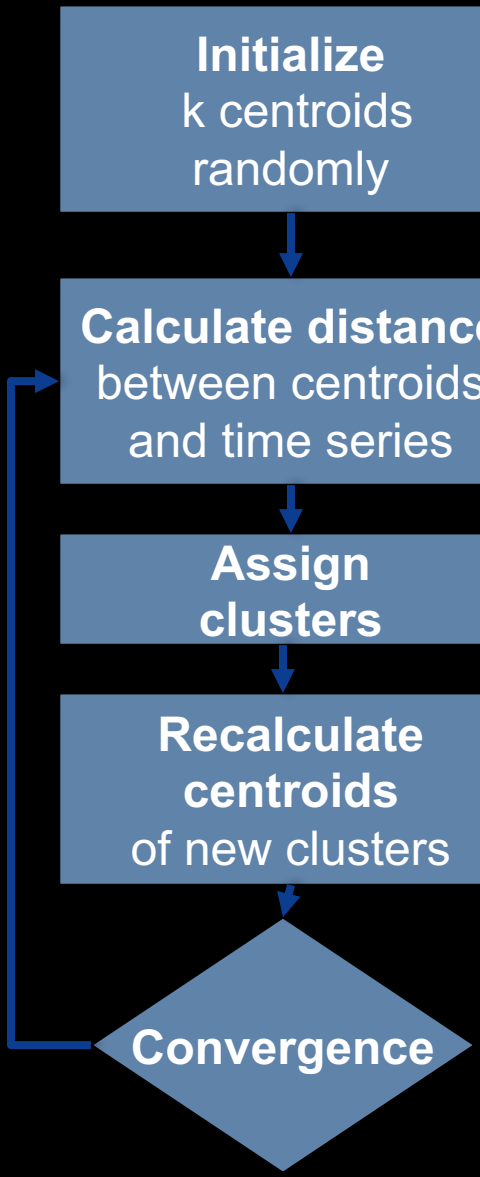
- K-means, K-medians
- Must pre-specify number of clusters
- Fast
- Requires a distance metric (e.g., Euclidean)

Curse of Dimensionality

Beware!

- High dimensional datasets are inherently sparse
- Examples are relatively equidistant, rendering clustering algorithms useless
- If dataset has ~10 or more features, I apply dimensionality reduction first

KMeans


Initialize
k centroids
randomly
Calculate distance
between centroids
and time series
Assign
clusters
Recalculate
centroids
of new clusters
Convergence

Convergence

- W are centroids
- C are clusters

$$E(W,C) = \sum_i \frac{1}{2} \min_{w \in W} (x_i - w)^2$$
$$= \sum_i \frac{1}{2} (x_i - w_{c(i)})^2$$

- E(W, C) decreases with each iteration of K-means
- K-means is proven to converge to a local optimum
- Initial centroid initialization may affect the final clusters
- Implicit assumption that data is Gaussian

Model Selection

- What's the ideal value of 'k' for a given dataset?
- What happens to error E if we set k = N, number of dataset examples?

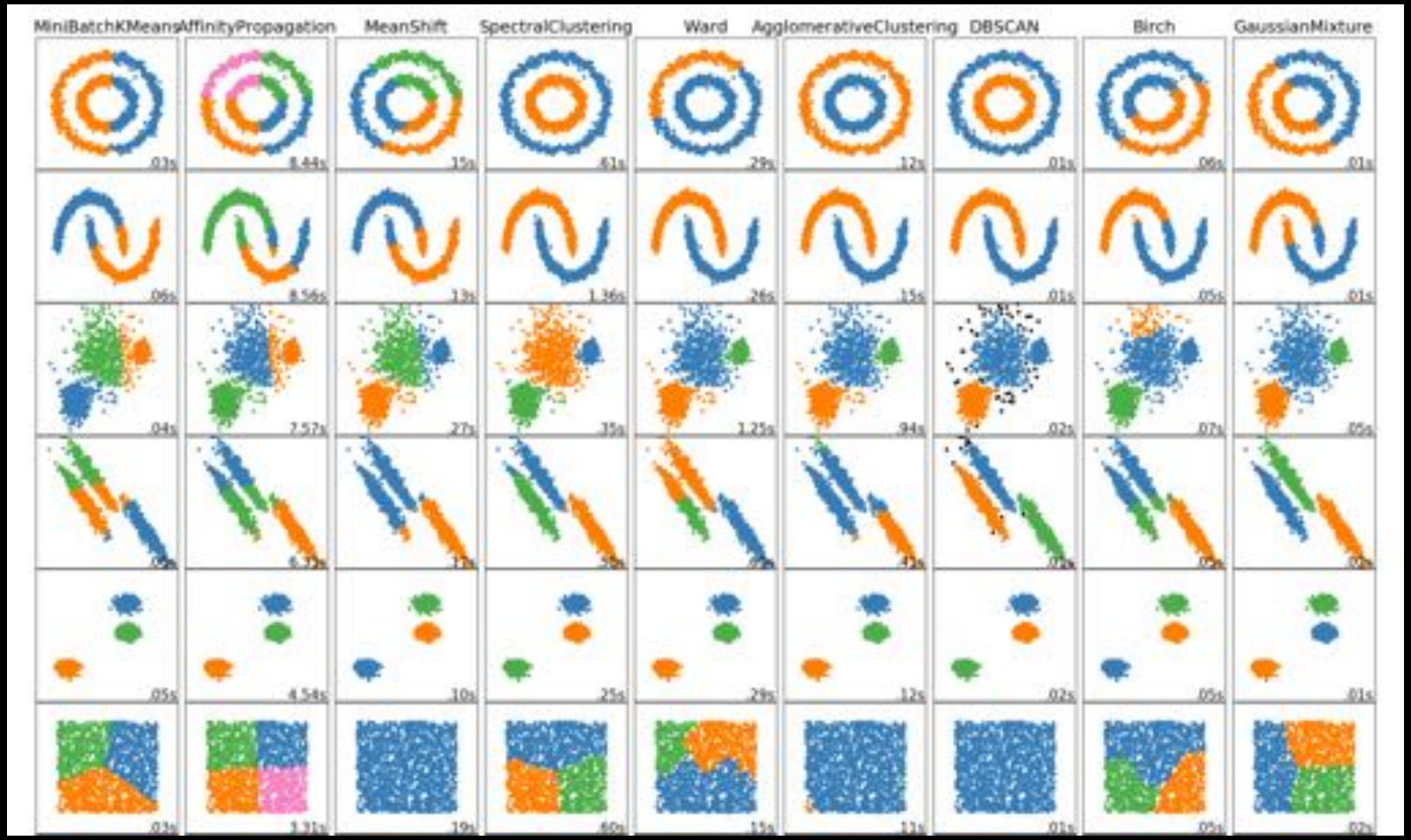
Model Selection

- What's the ideal value of 'k' for a given dataset?
- What happens to error E if we set k = N, number of dataset examples?
- Penalize large values of k (we'll see this concept again tomorrow)
  - Akaike Information Criterion
  - Bayesian Information Criterion
  - Pick k that minimize these

Other Types of Clustering

Density-based clustering: DBSCAN

Hierarchical Clustering: Birch, Ward, Agglomerative



https://scikit-learn.org/stable/modules/clustering.html#clustering

Anomaly Detection

Anomaly Detection

- Anomalies are typically found with respect to an unsupervised learning model
- N-sigma clipping is the simplest form of anomaly detection where a parametric model is fit to available data.

Anomalies / Outliers

- Example that is unusual with respect to the rest of the data

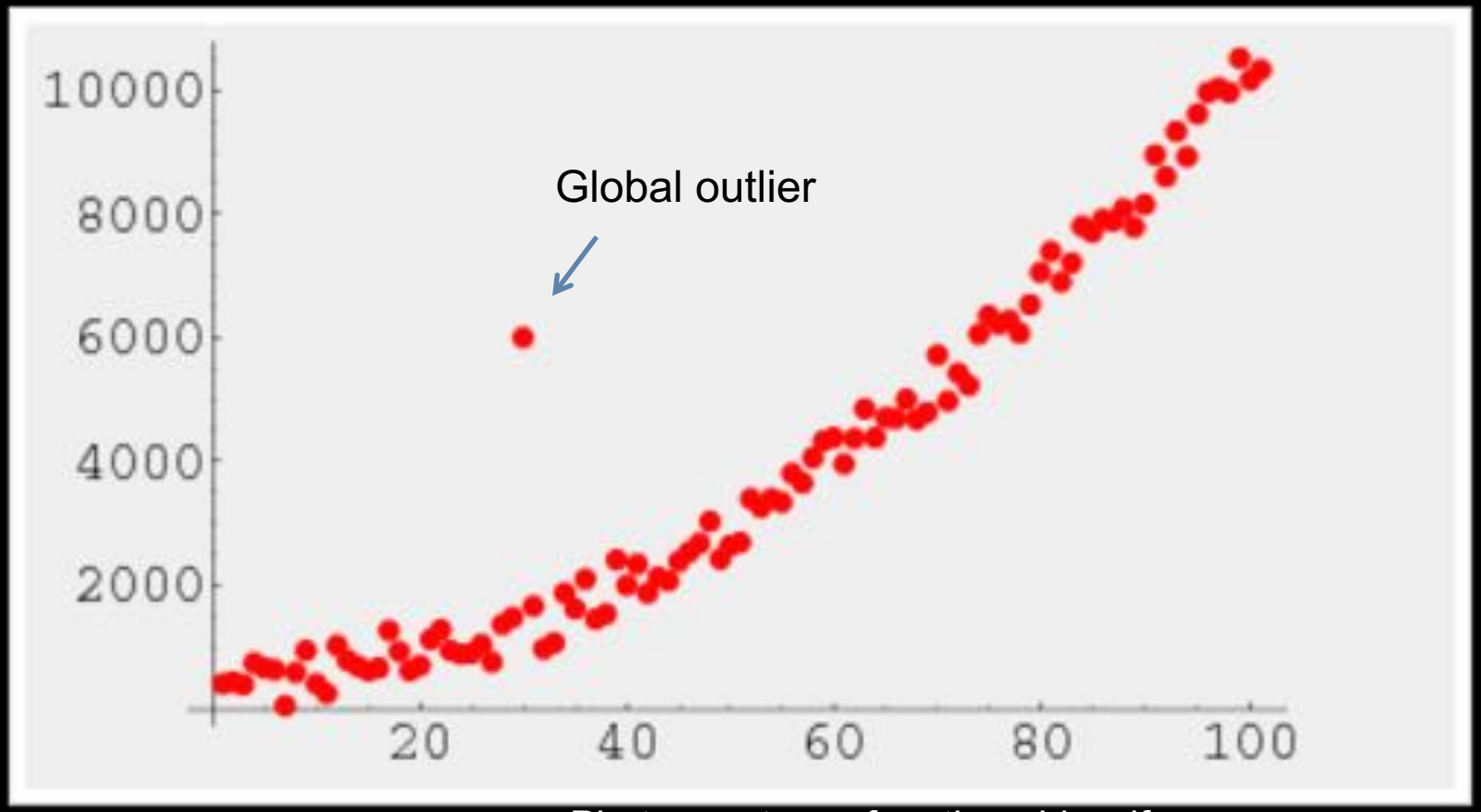
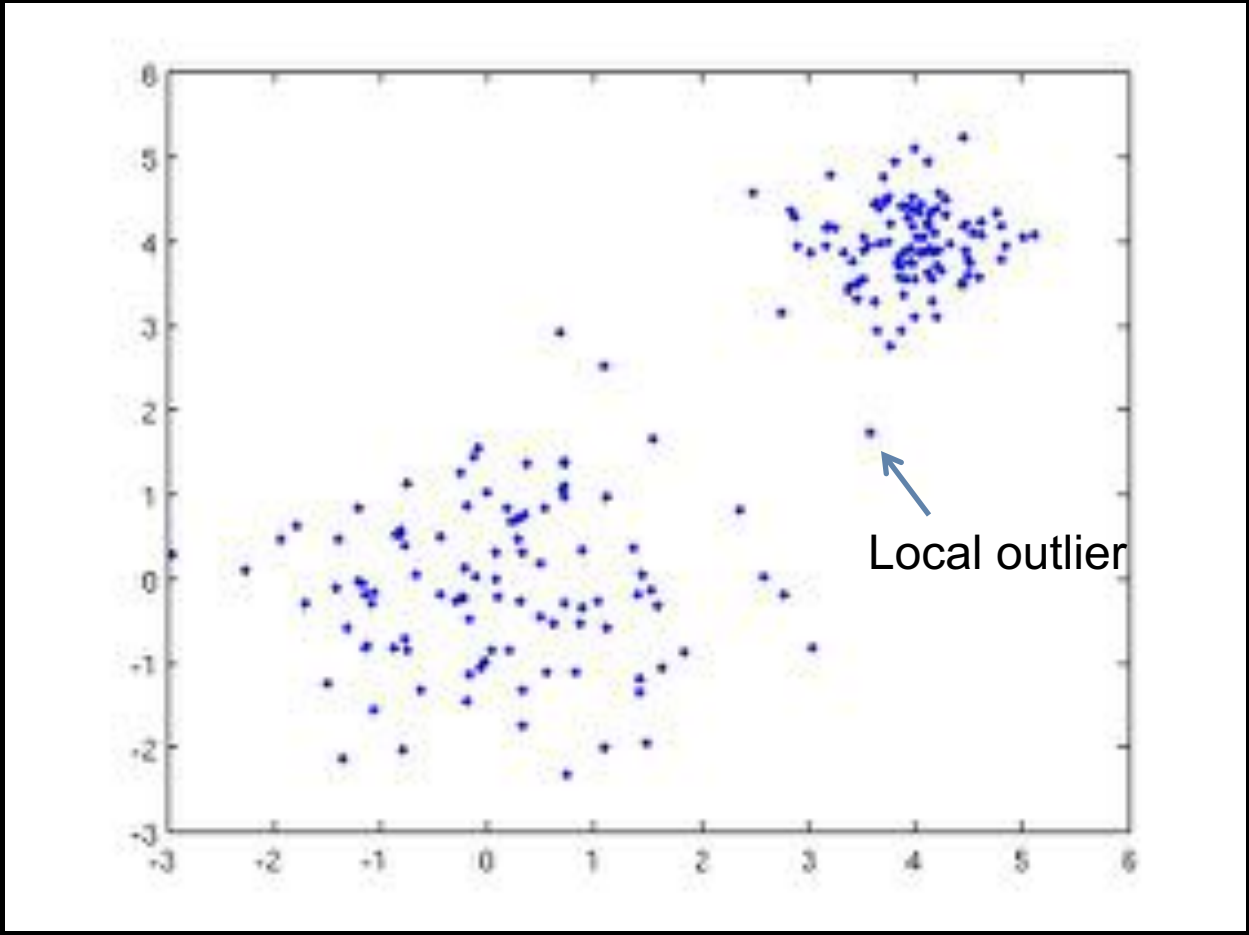


Photo courtesy of mathworld.wolfram.com

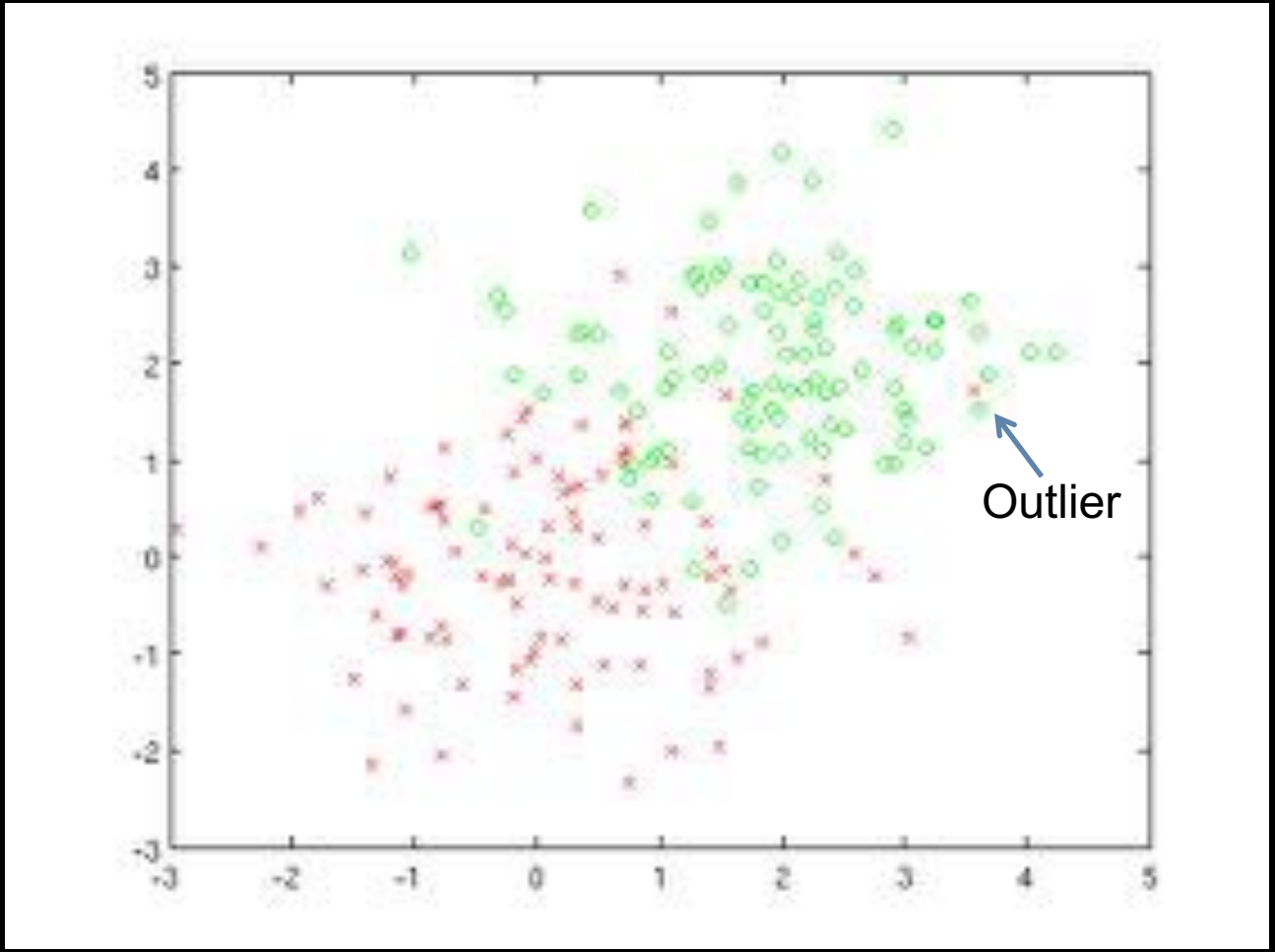
Local Outliers

- Example that is unusual with respect to a particular grouping of the data



Outliers in Labeled Data

- Example that is unusual with respect to its assigned class



Outlier: Positive Sense

- Outliers are indicative of scientifically interesting astrophysical anomalies
- Quasars were discovered by a scientist who followed up on anomalies in their data
- Anomaly detection can lead to scientific discovery

Source: Wikipedia.org/wiki/Quasar

Outlier: Negative Sense

- Non-astrophysical artifacts of telescope optics, image processing pipelines or non-detections
- Removed or modified

Summary

Key Takeaways

- Unsupervised learning constitutes learning without a target concept or reward
- Primary objective is data understanding
- K-means is fast clustering algorithm, but performs poorly in high dimensions and when data is not a mixture of Gaussians with constant variance. Be aware of its biases.
- Anomaly/Outlier detection is very subjective.

Data for Hands-On Exercises

Zwicky Transient Facility (ZTF)

- The Zwicky Transient Facility (ZTF) had first light in 2017
- ZTF will use a new camera with a 47 square degree field of view mounted on the Samuel Oschin 48-inch Schmidt telescope at Palomar Observatory
- Scans more than 3750 square degrees an hour to a depth of 20.5 mag
- ZTF conducts nightly searches for rare and exotic transients.
- Repeat imaging of the Northern sky (including the Galactic Plane) will produce a photometric variability catalog, ideal for studies of variable stars, binaries, AGN, and asteroids.

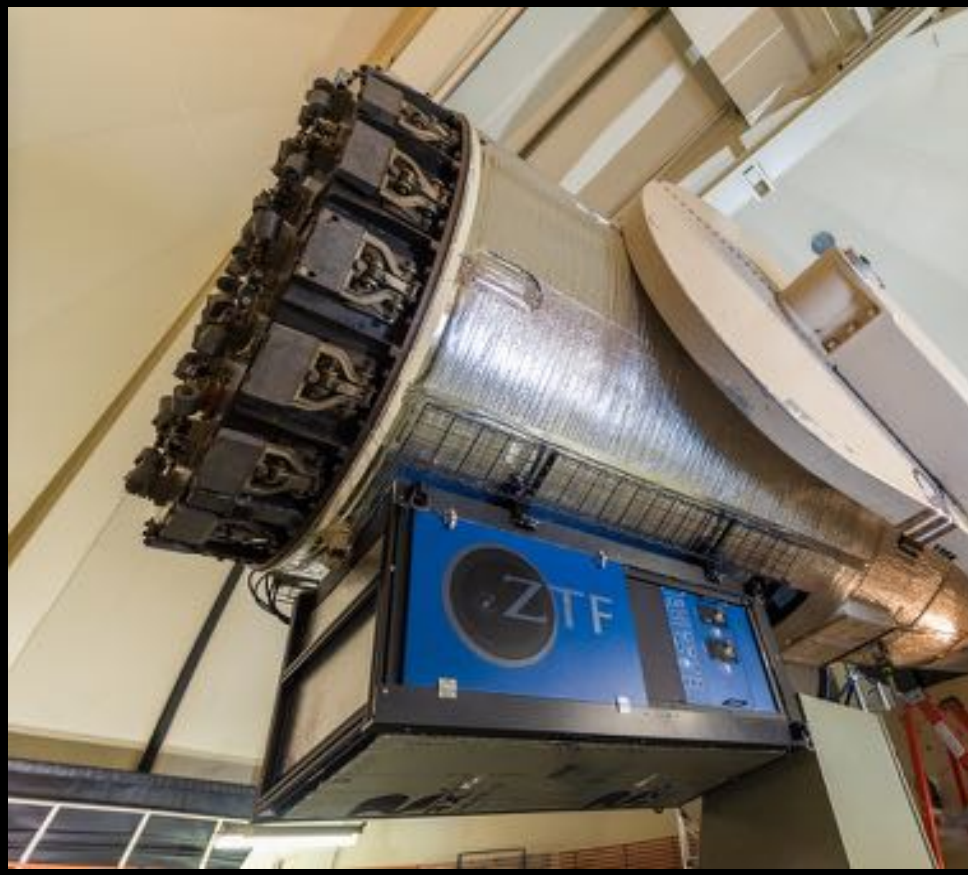
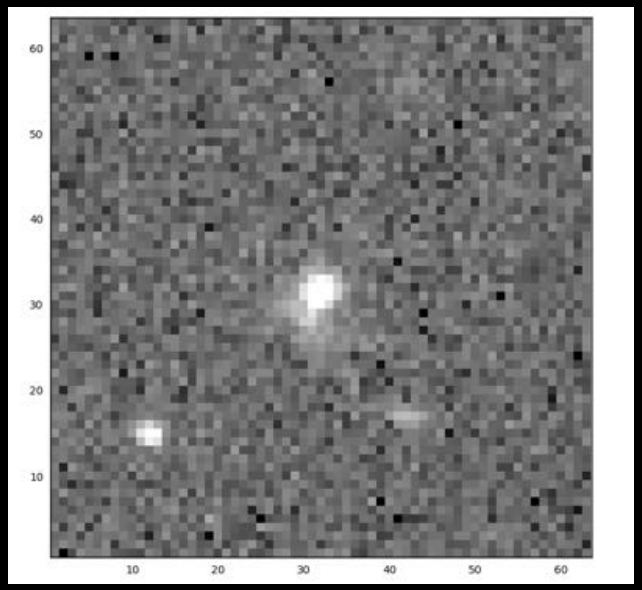


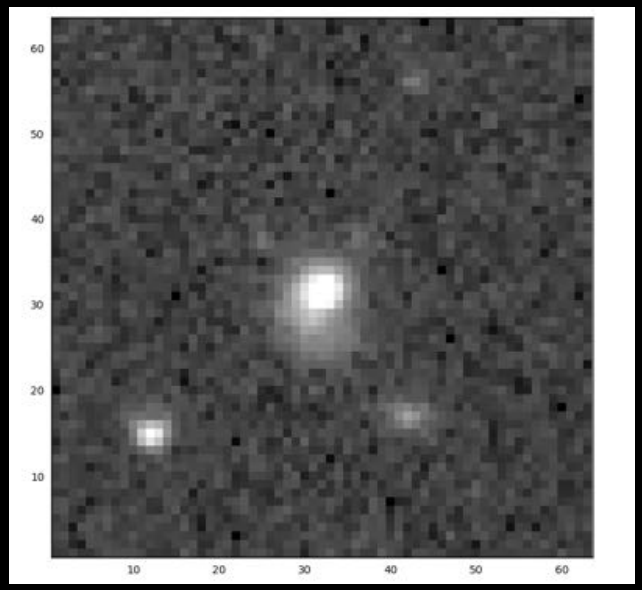
Image Subtraction

- Tool of discovery for PTF, Dark Energy Survey (DES), Skymapper, and the Large Synoptic Survey Telescope (LSST)
- Robust to crowded fields, high spatially-varying backgrounds

Science



Reference



Subtracted

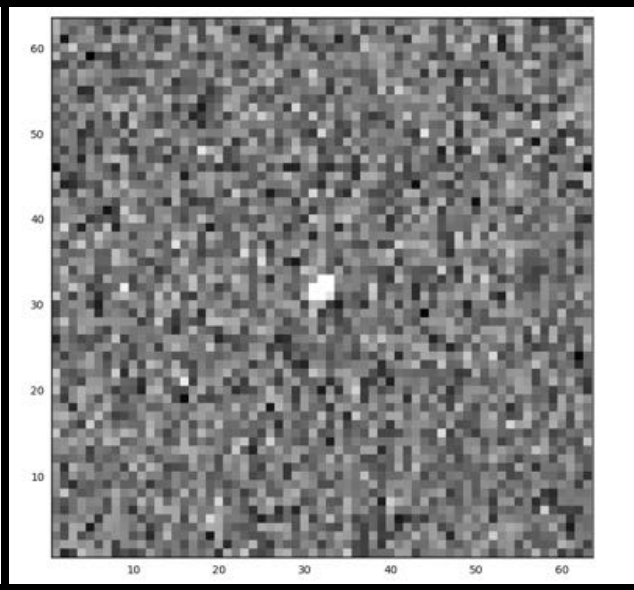
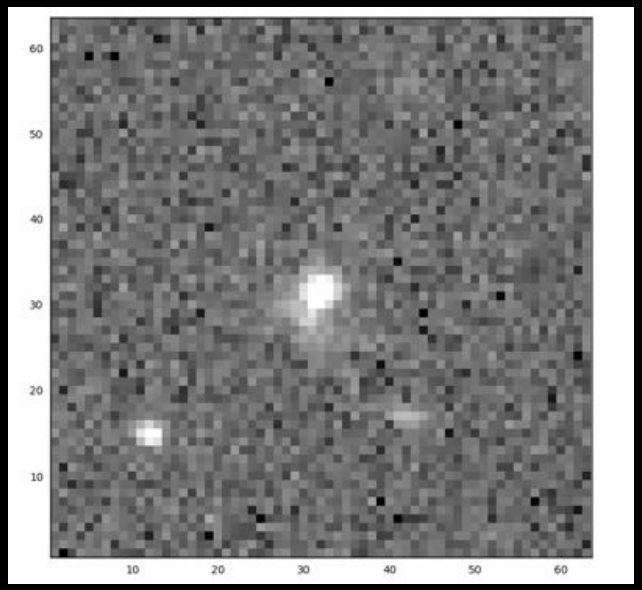
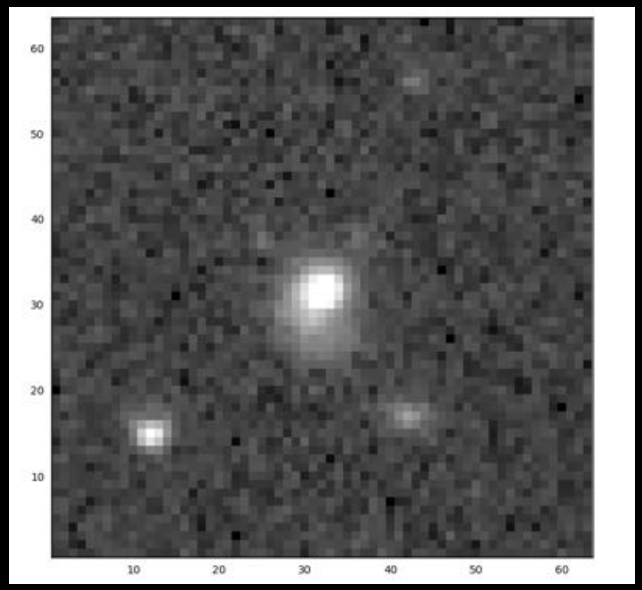


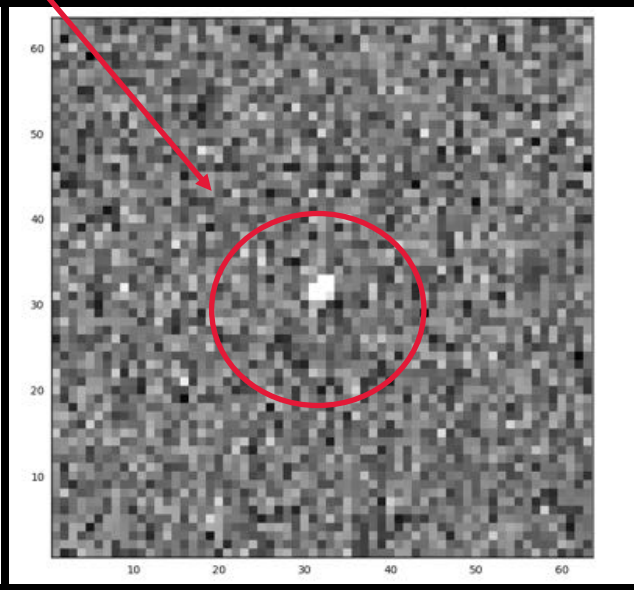
Image Subtraction

- Tool of discovery for PTF, Dark Energy Survey (DES), Skymapper, and the Large Synoptic Survey Telescope (LSST)
- Robust to crowded fields, high spatially-varying backgrounds

Science | Reference | candidate | Subtracted

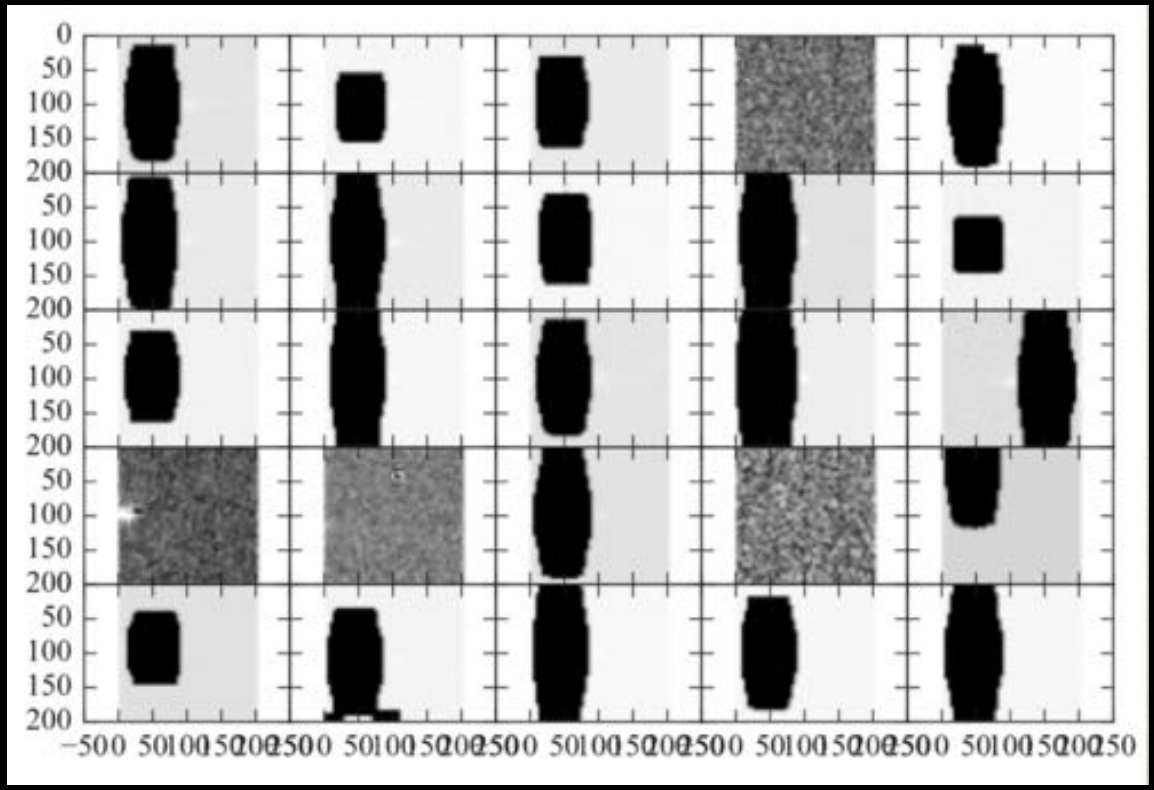






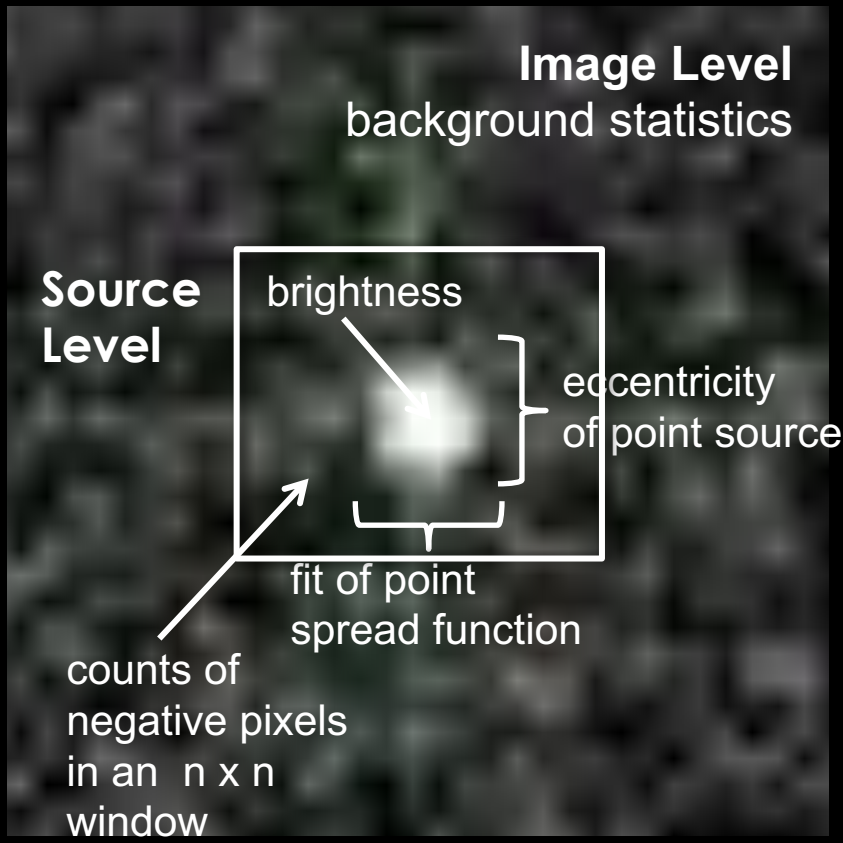
Subtraction is Challenging

- Image subtraction requires astrometric alignment, flux-scaling, fitting of point-spread function (PSF) to both science and reference images

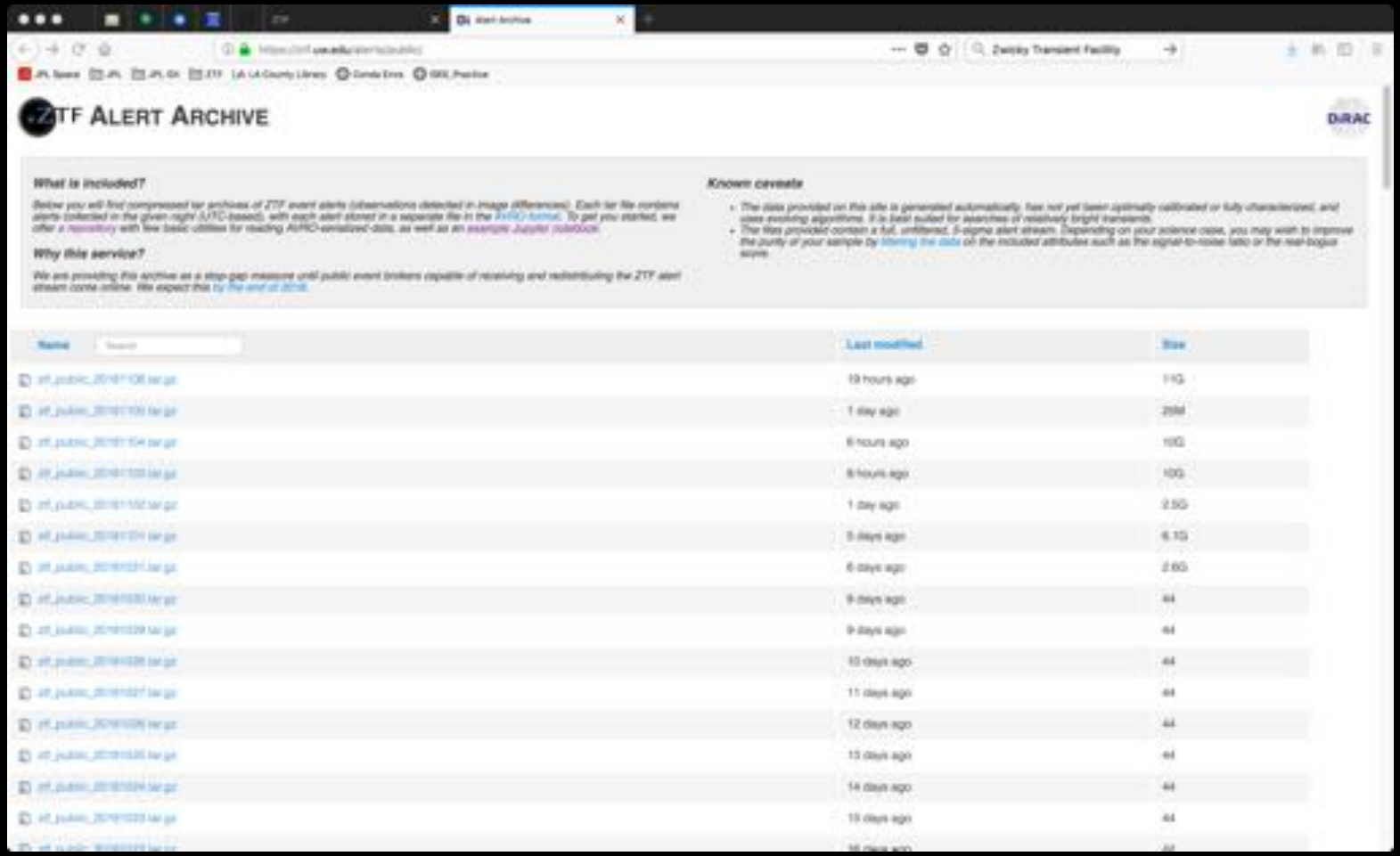


Feature Representation

- Features are a by product of image subtraction, source finding, and photometry.



ZTF Public Alerts



Jet Propulsion Laboratory
California Institute of Technology

jpl.nasa.gov