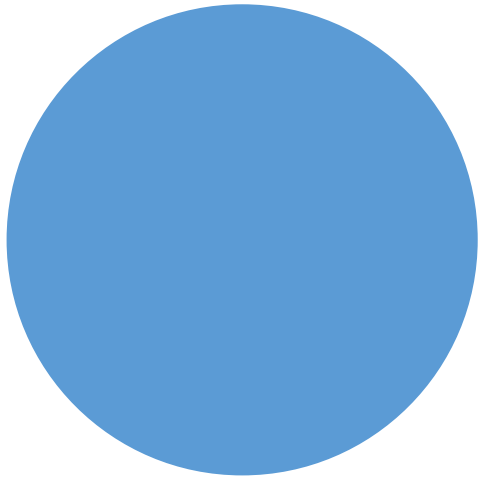
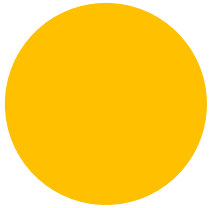



So you've clustered your data… now what?

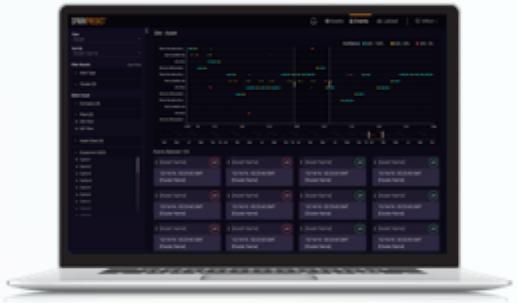
Kevin Gullikson

Staff Data Scientist
@SparkCognition


sparkcognition™

SparkCognition's Product Portfolio

SPARKPREDICT®



Anomaly Detection
- Sensor data ingestion
- Pattern recognition
- Normal v. abnormal behavior regimes

Failure Prediction
- Explainable to users
- Continuously learns

DEEPARMOR®

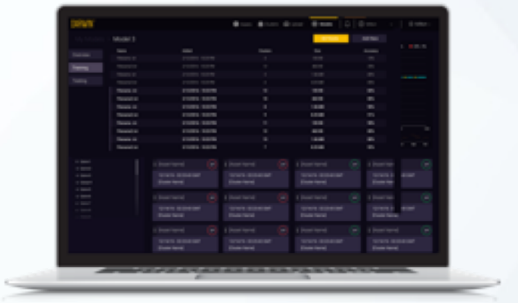
Endpoint Protection for IT/OT systems
- Detects multiple attack vectors and weaponized documents
- Prevents execution of malicious files
- Highest efficacy rates

DEEPNLP®

Automates Workflow
- Mine structured and unstructured data
- Enhance M&R task planning

Content Analytics
- Creates associations across management functions -- work orders, incident reports

DARWIN™



AI Building AI
- Accelerates AL/ML projects
- Increases model efficacy
- Augments DS teams

We're Hiring!

- Data Science
- Software Engineering
- Director of Data Science
- UX
- DevOps

www.sparkcognition.com/careers

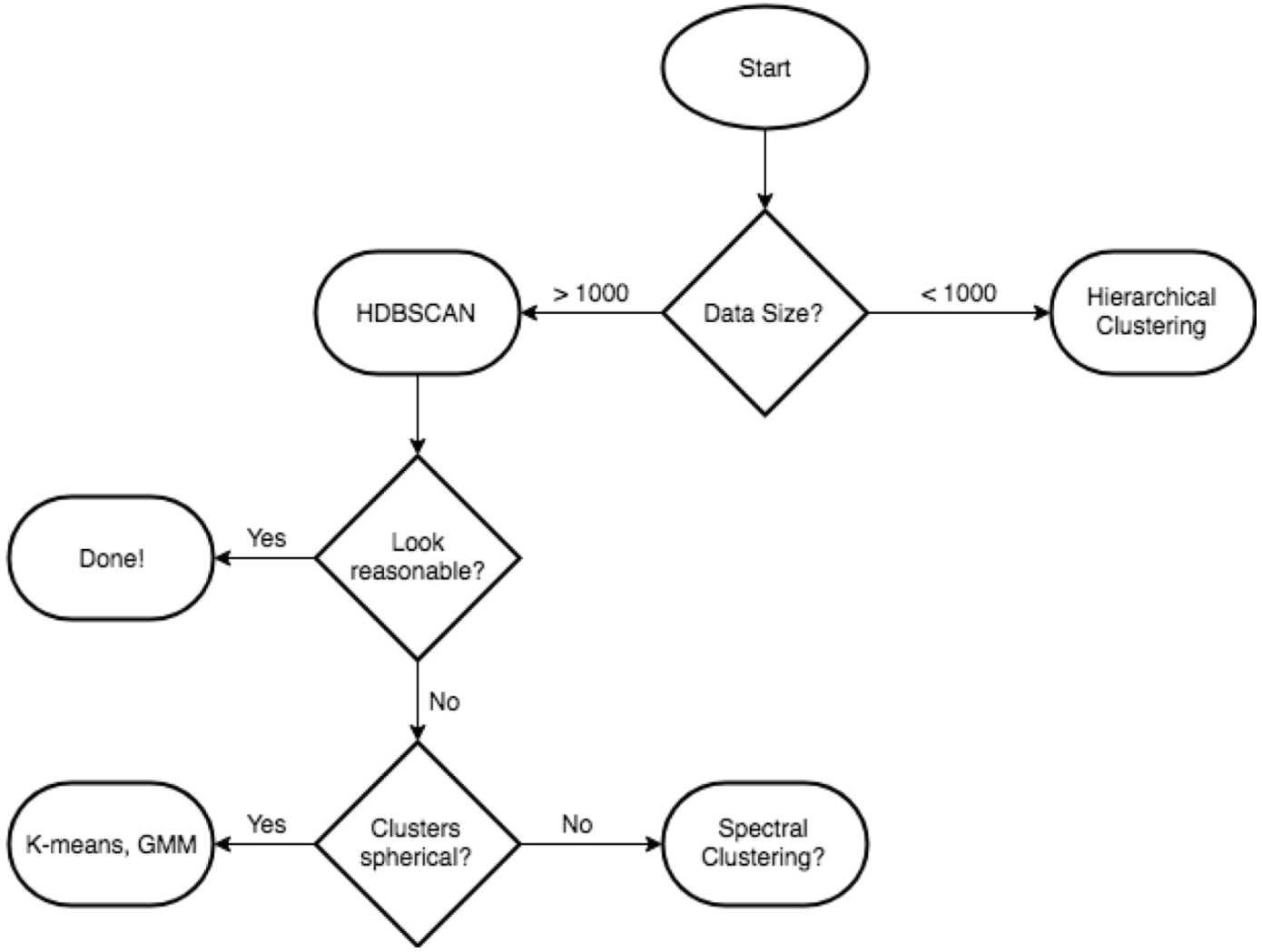
Overview

- Whirlwind tour of unsupervised learning, mostly clustering
  - Why are you clustering
  - What algorithms are there?
  - Anomaly Detection
- Interpreting clusters
- Tips for validating/scoring unsupervised models

“Why are you clustering?”

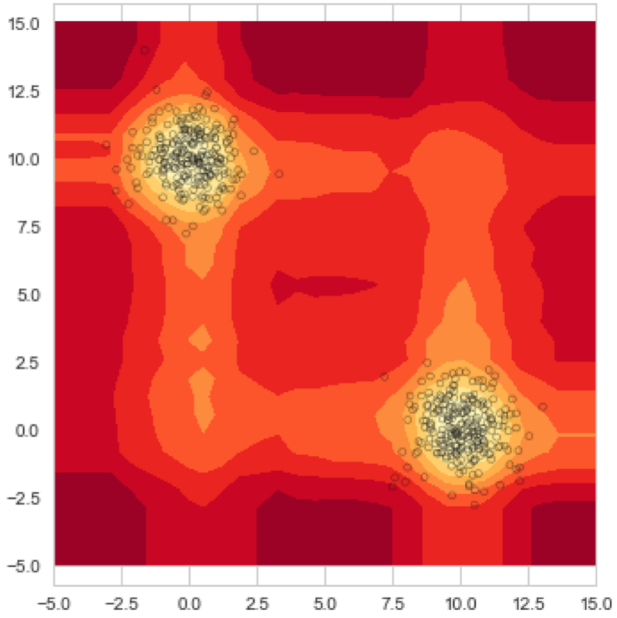
- Just exploring data ← Probably most common
- Want to build a separate supervised model for each cluster
- Want to use the cluster label as a feature in supervised model
- Want to decimate data in smart way

“What clustering algorithm should I use?”

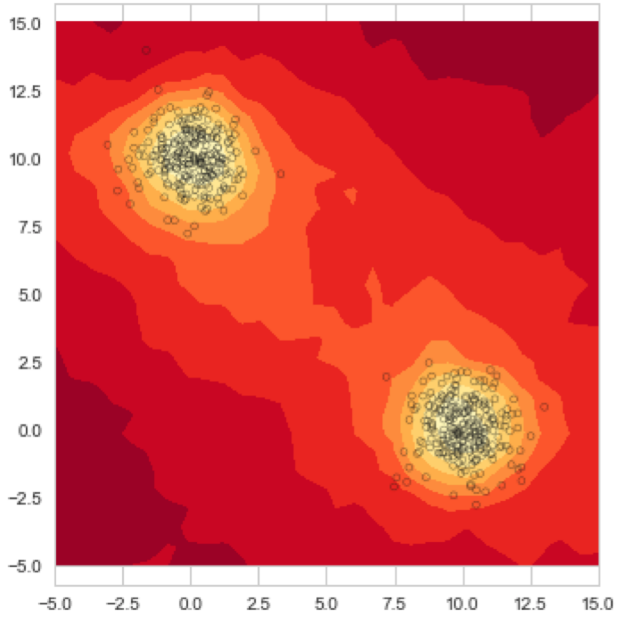


“What about anomalies?”

- Can influence clusters, usually negatively
  - Less so for hierarchical or density-based clustering
  - Should be run first and cluster on the rest
- My go-to: IsolationForest



Standard/scikit-learn



https://github.com/sahandha/eif

Now what?

Questions to ask after clustering:

- What makes each cluster different from the rest?
- What makes this cluster different from that other one?
- What clusters are similar to this one?
- Why is this anomaly an anomaly?
- Are my clusters meaningful?

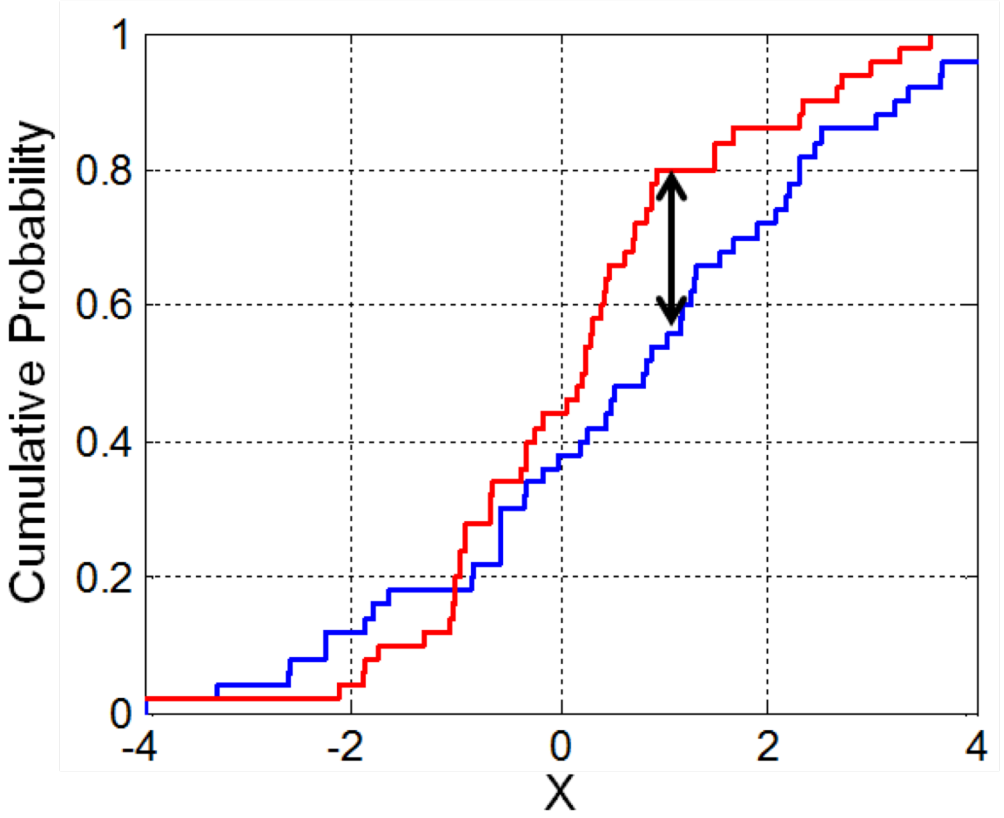
Statistical Distances

- A measure of how different two distributions are
- Features with larger distance between “this cluster” and “other clusters” are more important for “this cluster”
- Several options
  - Cohen’s d
  - KS distance
  - Wasserstein distance

Statistical Distances

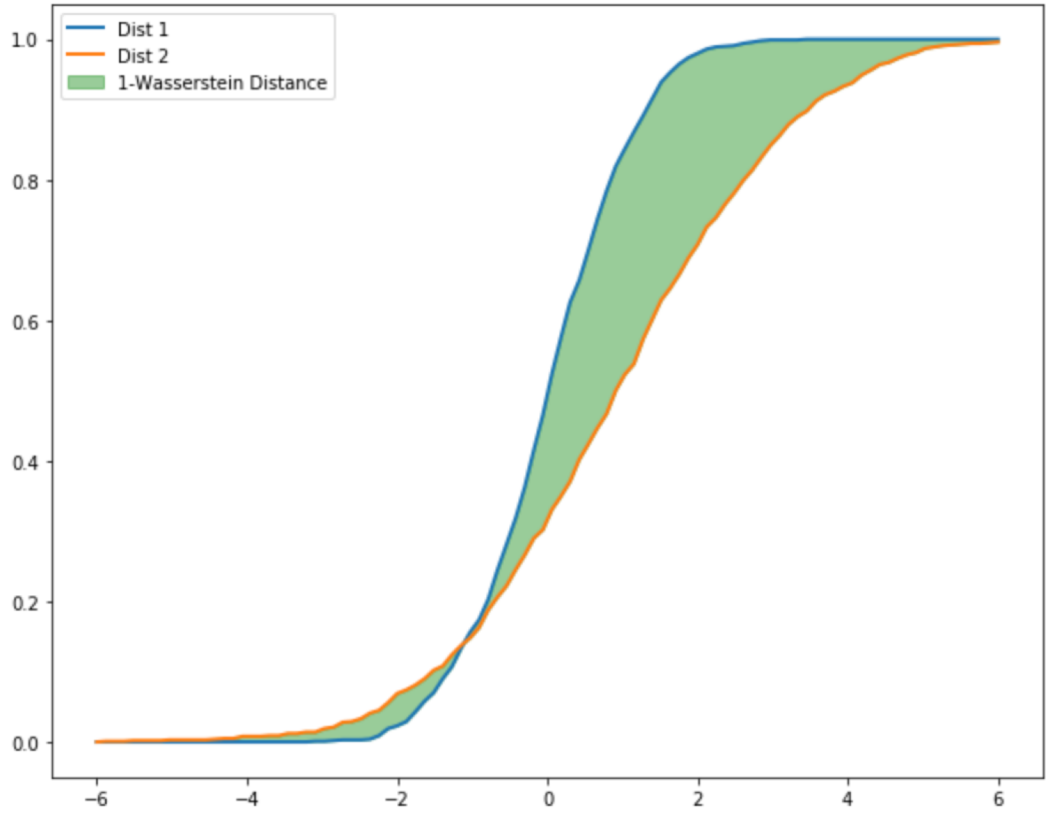
- A measure of how different two distributions are
- Features with larger distance between “this cluster” and “other clusters” are more important for “this cluster”
- Several options
  - ~~Cohen’s d~~ (only good for comparing two gaussians, which will almost never happen)
  - KS distance
  - Wasserstein distance
  - KL Divergence and related measures (good, but hard to compute)

Kolmogorov-Smirnov Distance



$$KS(F, G) = max(|F(x) - G(x)|)$$

Wasserstein Distance



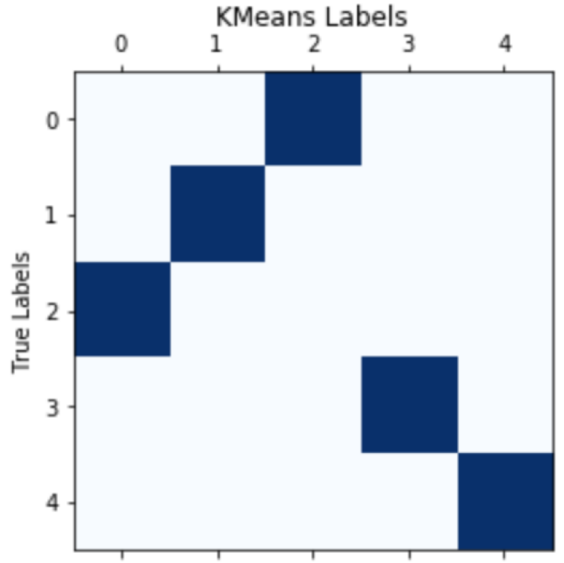
$$W_1(F,G) = \int_0^1 \left|F^{-1}(t) - G^{-1}(t)\right| dt$$

Toy Example

[92]:
```
X, true_labels = make_blobs(n_samples=10000, n_features=20, centers=5)
km = KMeans(n_clusters=5)
km_labels = km.fit_predict(X)

# Convert to dataframe for use later
datacols = [f'x{i+1}' for i in range(X.shape[1])]
data = pd.DataFrame(data=X, columns=datacols)
data['cluster'] = km_labels
```

[93]:
```
plt.matshow(confusion_matrix(true_labels, km_labels), cmap=plt.cm.Blues)
plt.ylabel('True Labels')
plt.title('KMeans Labels', pad=10);
```

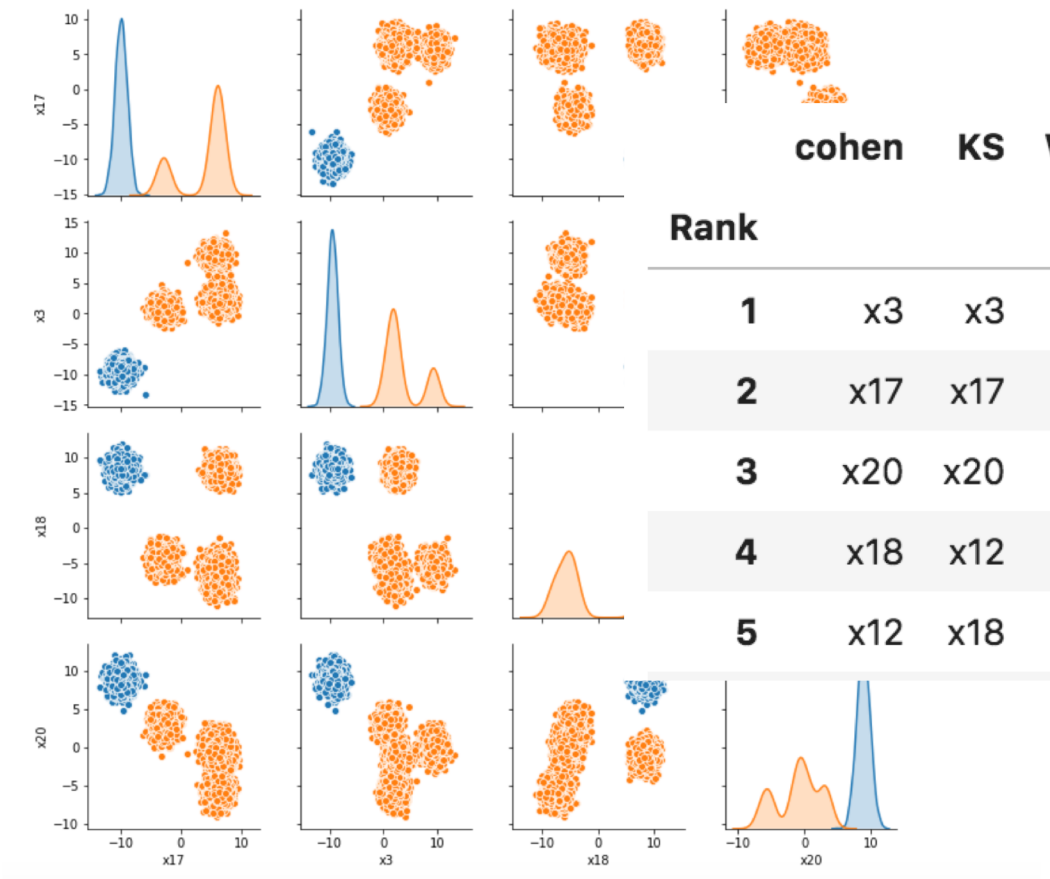


Toy Example Time!

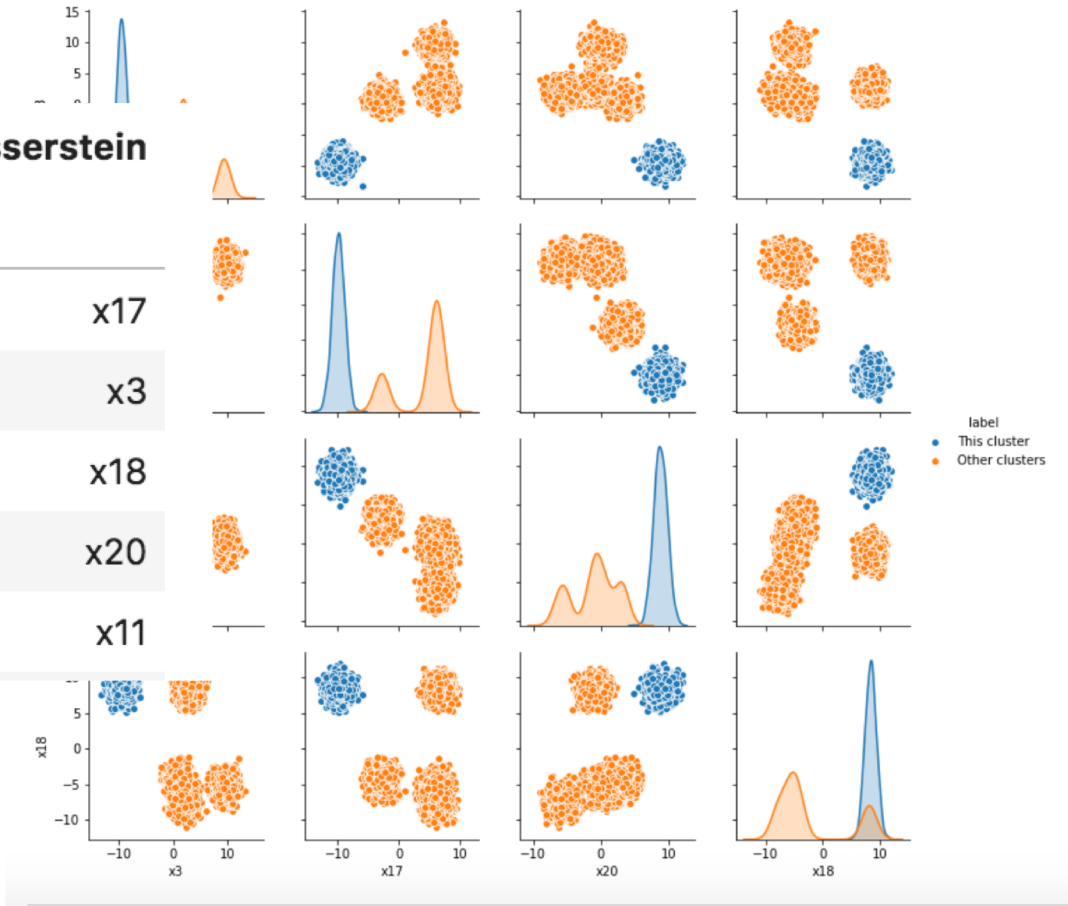
sklearn’s make_blobs with 20 features and 5 clustes

Wasserstein Top Features

KS Top Features



Rank	cohen	KS	Wasserstein
1	x3	x3	x17
2	x17	x17	x3
3	x20	x20	x18
4	x18	x12	x20
5	x12	x18	x11



Questions to ask after clustering:

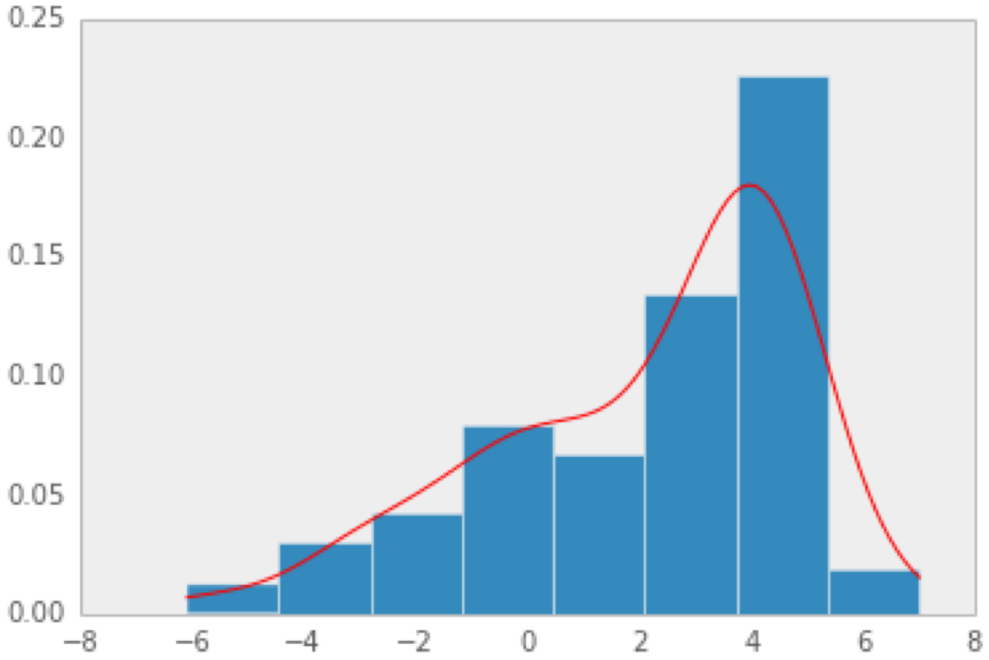
- ~~What makes each cluster different from the rest?~~
- What makes this cluster different from that other one?
  - Just change your “reference” data from everything else to the cluster you are comparing too
- What clusters are similar to this one?
  - The most different features are least different (have small distance)

$$\sum_{f \in topfeatures} d_f(x[label = 1], x[label = 2])$$

- Why is this anomaly an anomaly?
- Are my clusters meaningful?

Anomaly Feature Importance

- You want P(this value | distribution of this feature)
- KDE →
  - Need to pick bandwidth
  - Won't rank well when outside of training data range
- Treat as dirac delta PDF and use Wasserstein distance



Are my clusters meaningful?

- Things like silhouette score and metrics you may have seen on scikit-learn assume spherical clusters
- Best answer: what matters for your analysis?
  - Are you clustering to identify bad states in machinery? Then score based on how well your clusters point to the problems
  - Are you trying to segment a population? The defining characteristics of clusters should “make sense”
- Slightly more general answer:
  - All clusters should be well-separated from the rest of the data, at least in the top few features

Takeaways

Clustering is easy now. Interpreting and validating results is black magic

Use statistical distances to find defining characteristics of clusters

Use those top features to validate/score your model