Anomaly and fraud detection with Machine Learning

Ahmed Rebai
Esprit Prépa & Esprit School of Business

Plan

- Anomaly/Fraud detection with machine learning/Deep learning
- Outlier detection general applications
- Applications to financial sector (banking, insurance): Towards Fintech
- Supervised vs. Unsupervised
- Algorithms and methods
- Conclusions Ensemble method (Isolation forest) and Density method (DBSCAN/HDBSCAN)

Anomaly/Fraud detection with machine learning/Deep learning

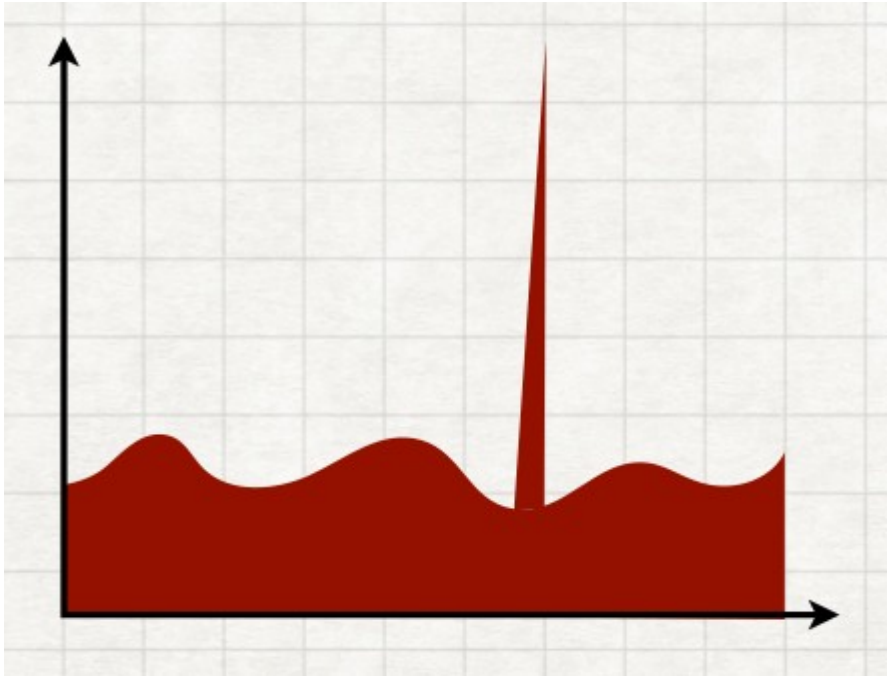
- Machine learning : data mining, predictive analytics, artificial intelligence (in practice : statistics and numerical methods for statistical analysis)

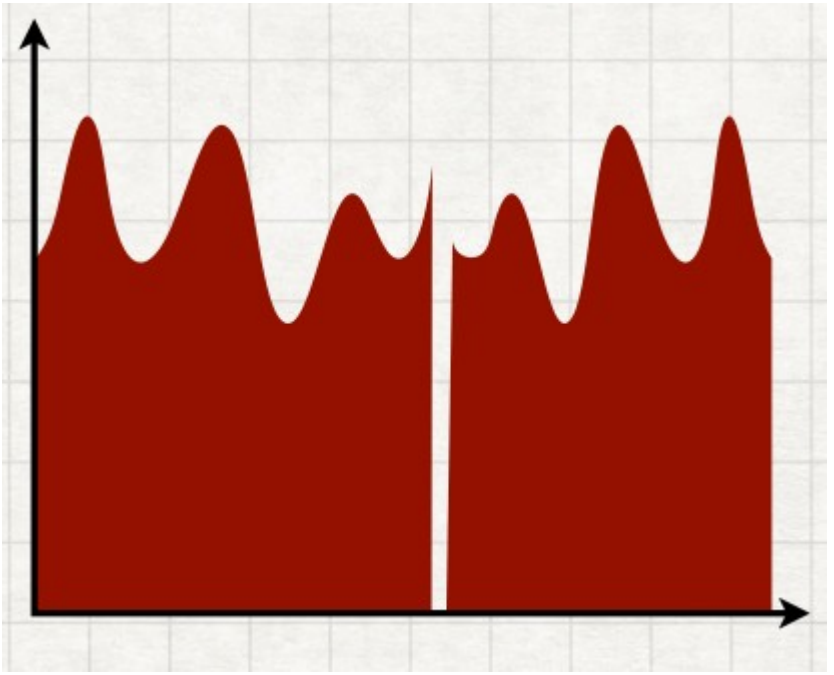
- Anomaly: deviation from “normal”/”expected”
  Anomaly=Outlier=Deviant or Unusual Data Point

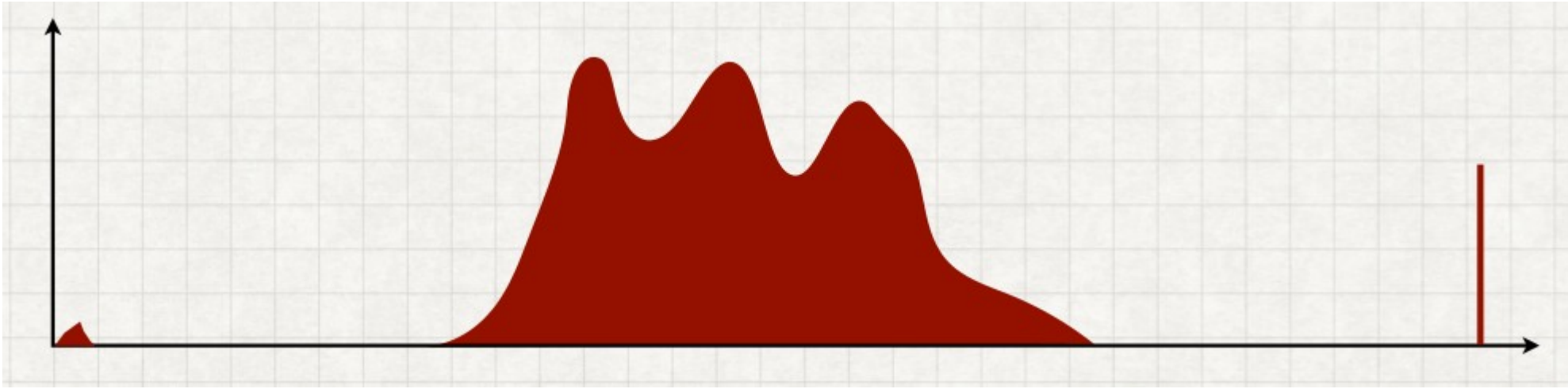
- Anomaly detection : detection of outlier events or observations: Detecting deviations from the expected pattern of a data set.

The real challenge in anomaly detection is to construct the right data model to separate outliers from noise and normal data.

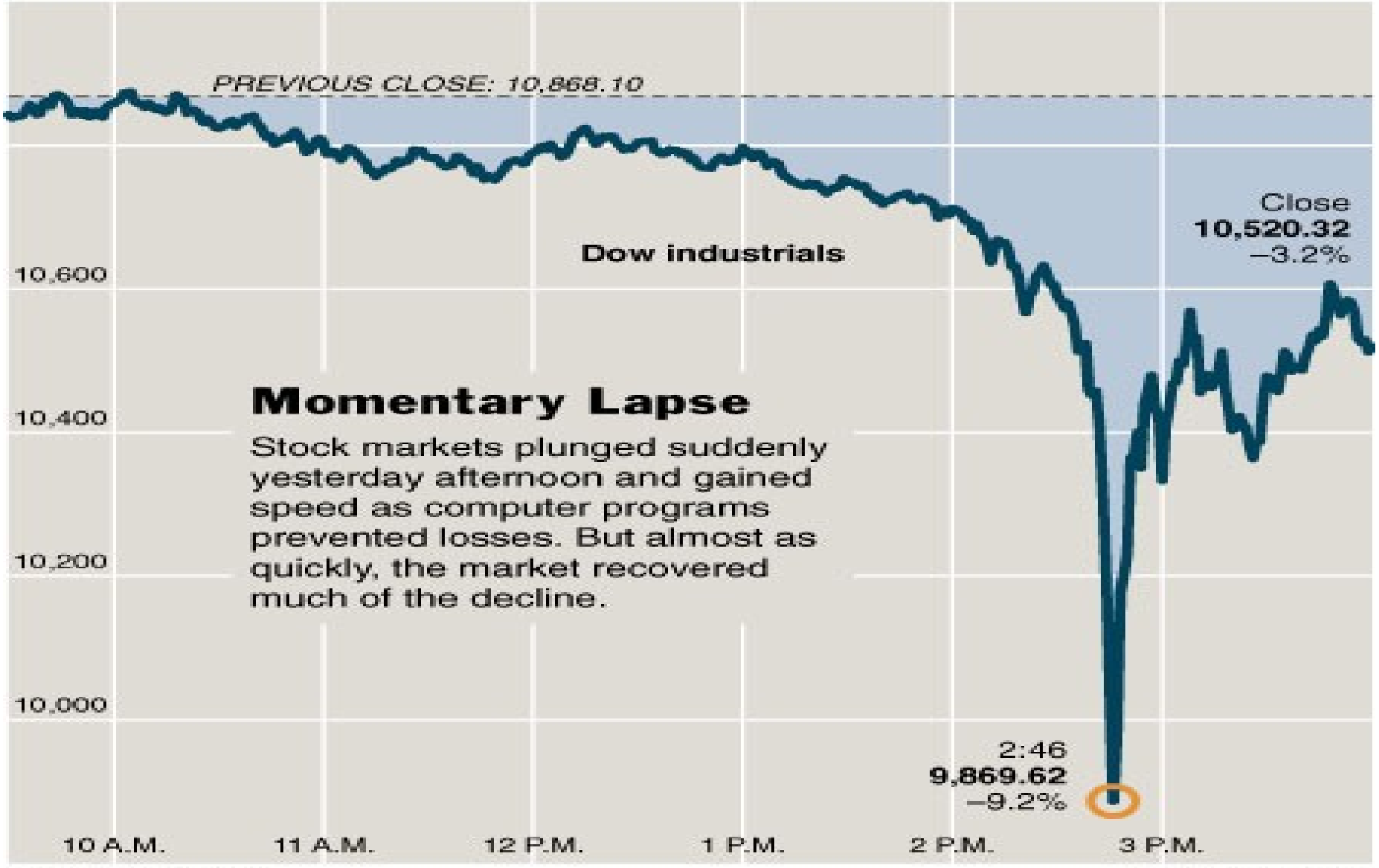
In one dimension: what anomalies look like?







Real Example: Stock market : The May 6, 2010, Flash Crash at 2:45 pm

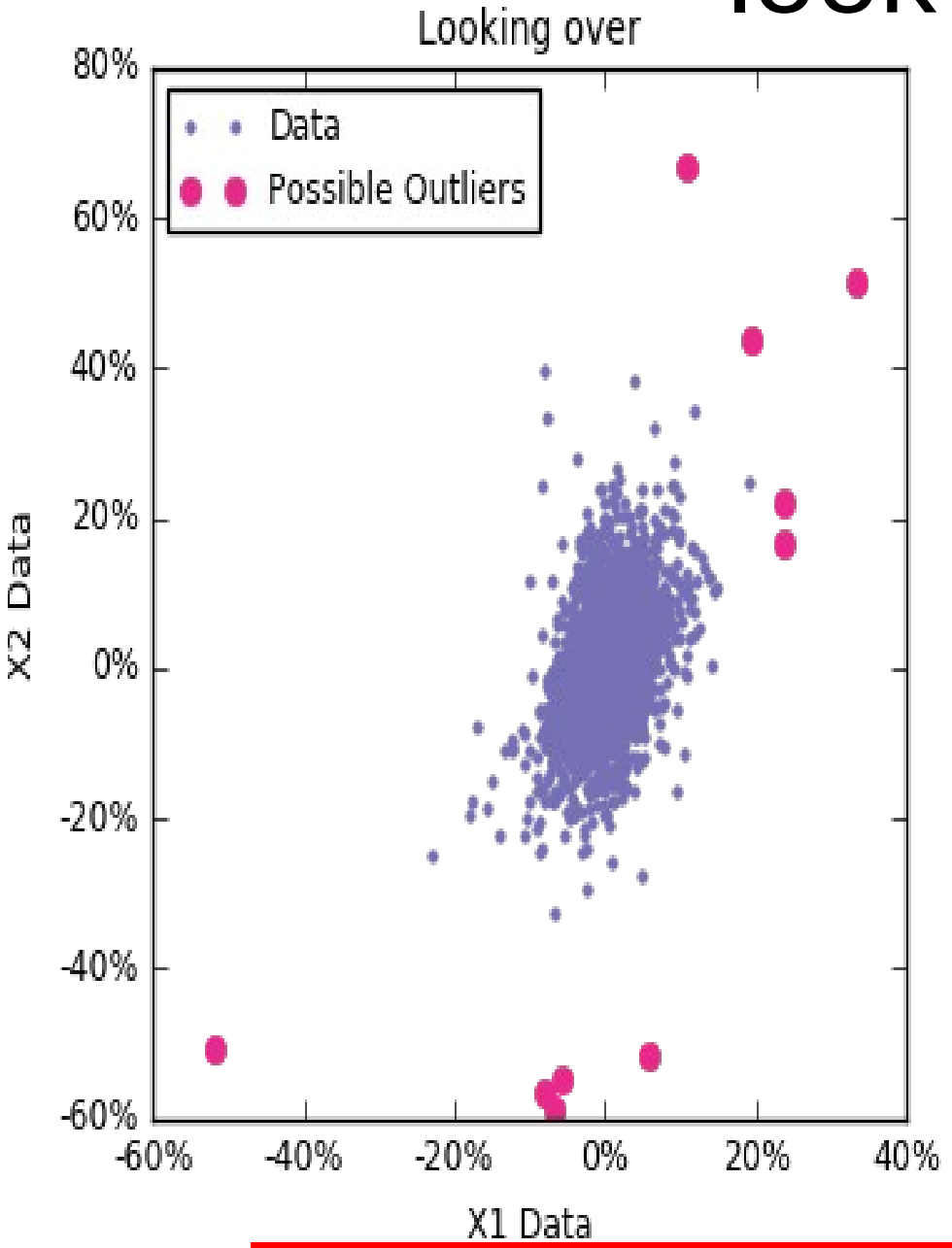


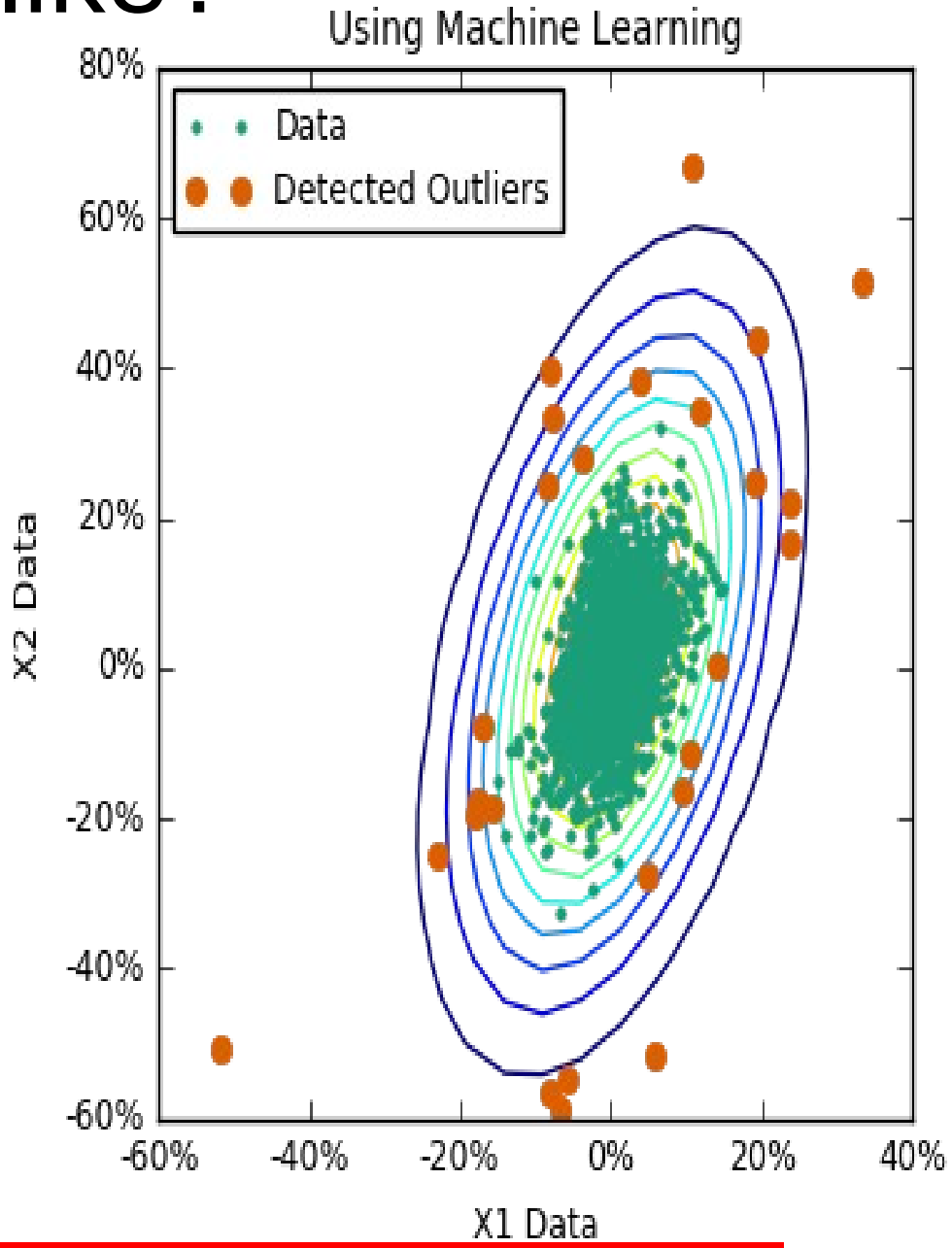
Real Example from the Stock market : A real financial fraud!

- On April 21, 2015, nearly five years after the incident, the U.S. Department of Justice laid "22 criminal counts, including fraud and market manipulation" against Navinder Singh Sarao, a trader. Among the charges included was the use of spoofing algorithms; just prior to the Flash Crash, he placed thousands of E-mini S&P 500 stock index futures contracts which he planned on canceling later.

Stock market : Flash Crash continues

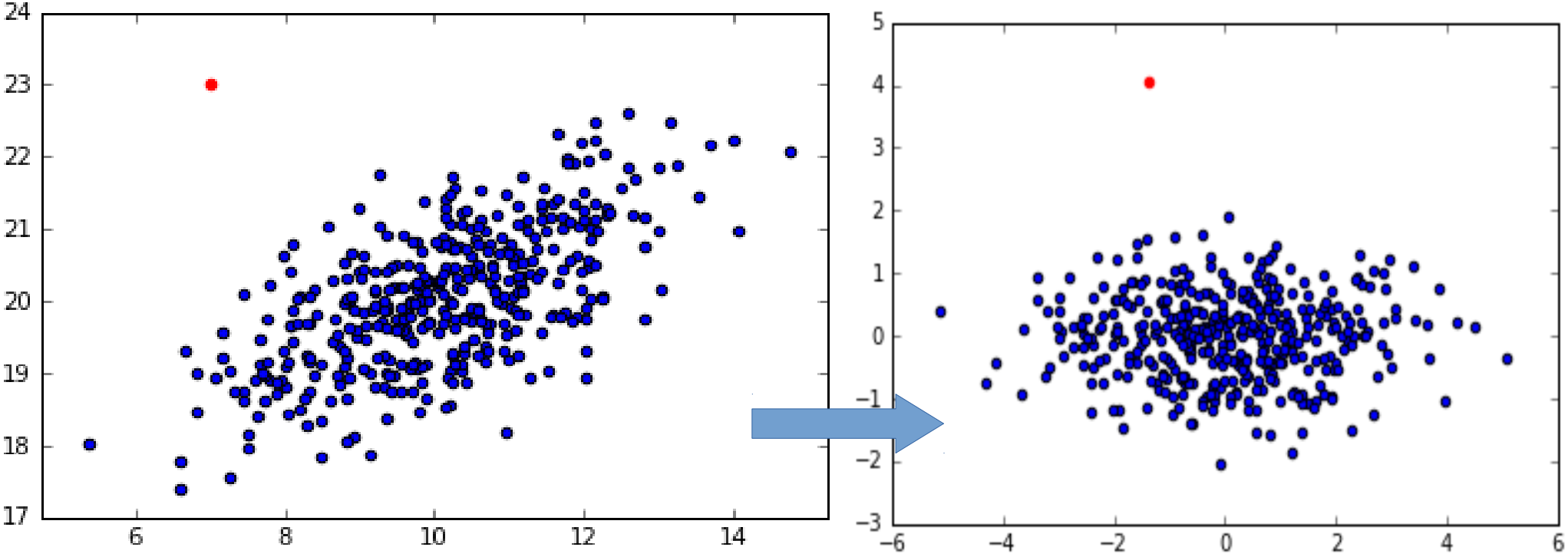
In two dimension: what anomalies look like?





Results with DBSCAN algorithm from sklearn.cluster

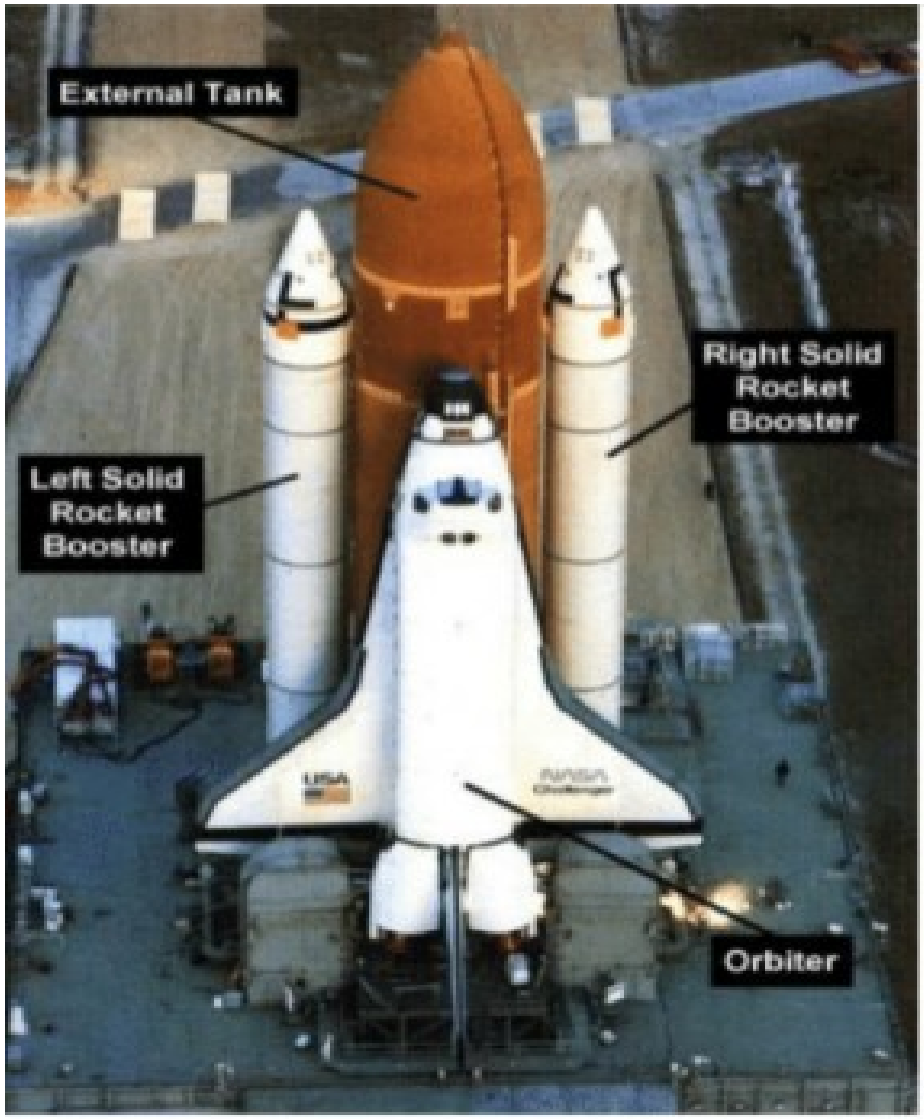
In two dimension: what anomalies look like?

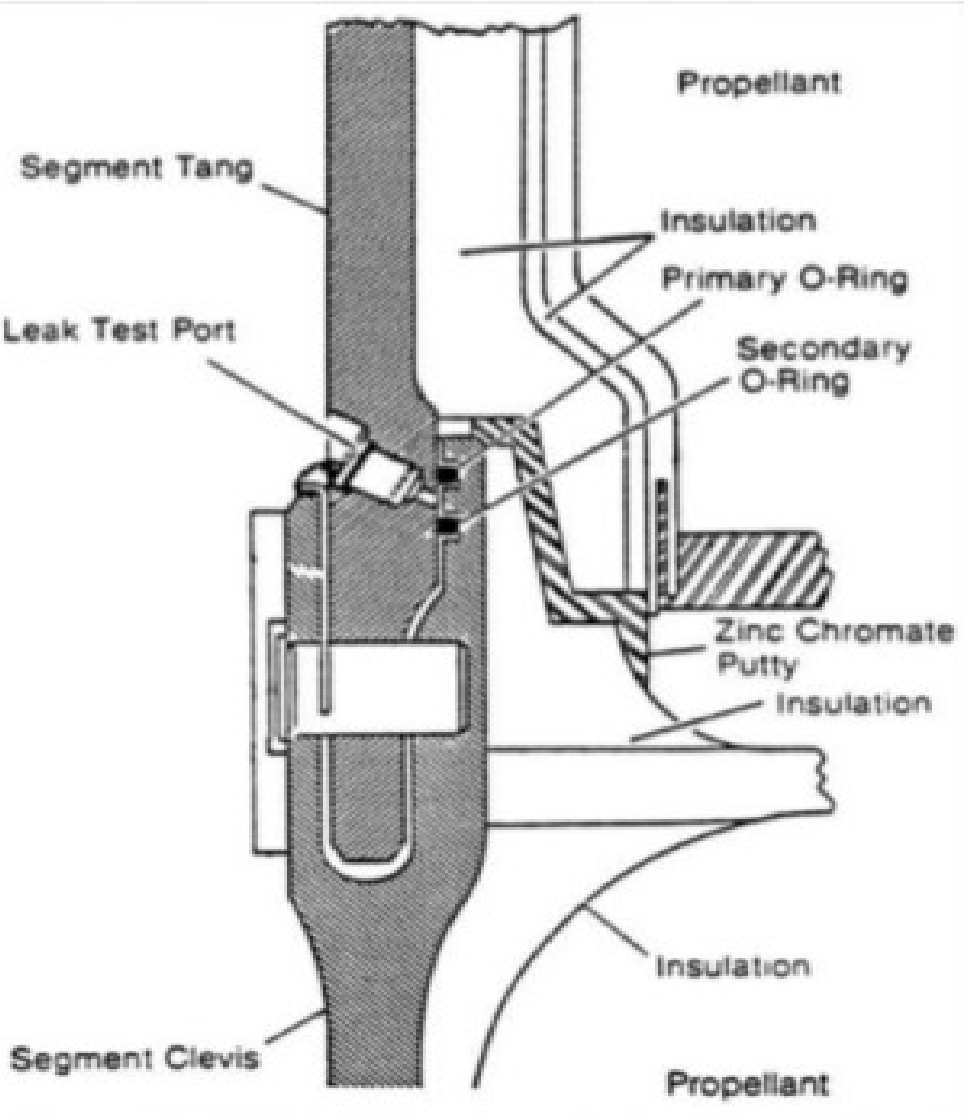


Results with 2D PCA method from sklear.decomposition
also you can do kernel PCA

Real Example: Rocket science

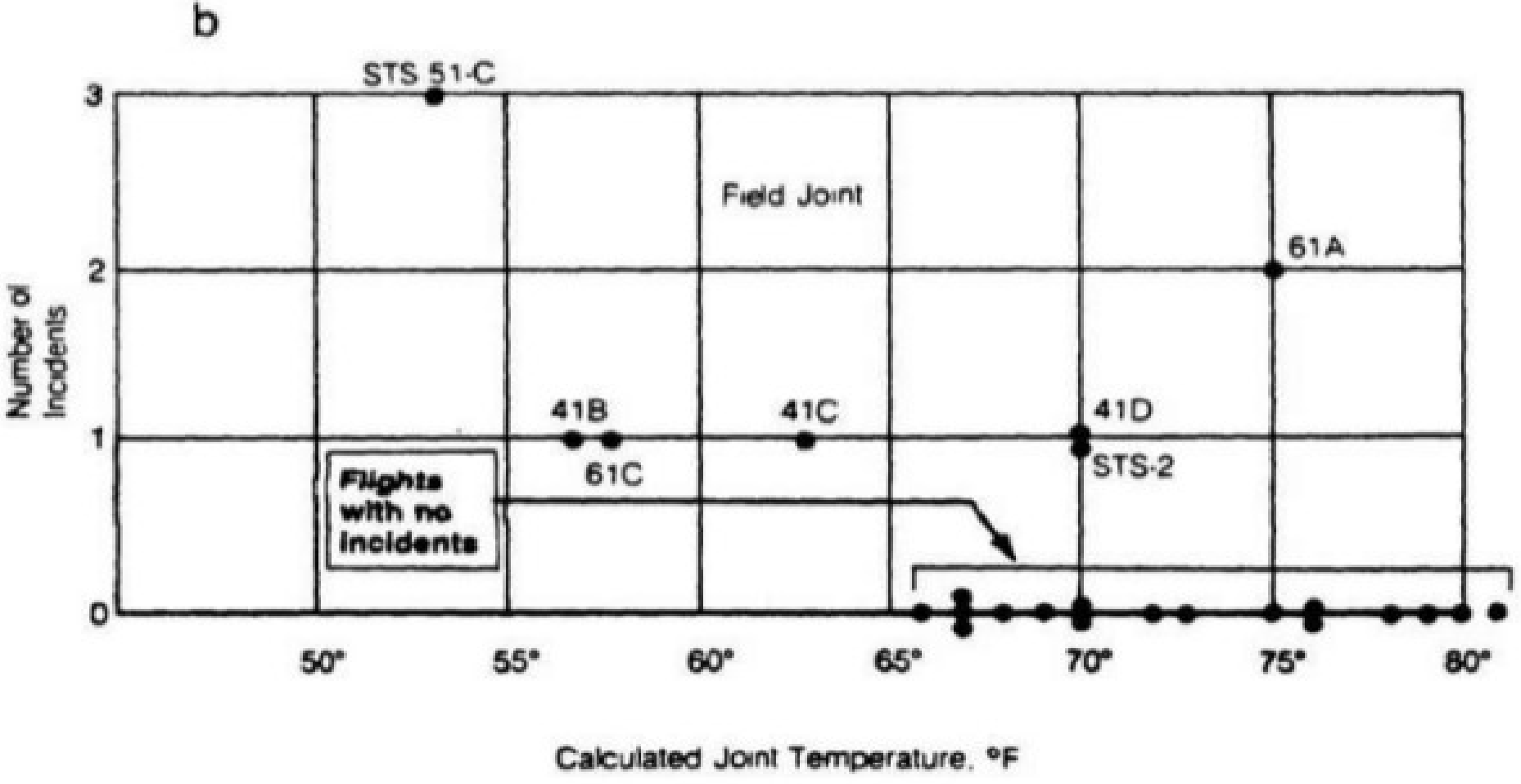
January 28 1986 Challenger spatial mission:
Dalal & al., 1989 Journal of the American Statistical Association



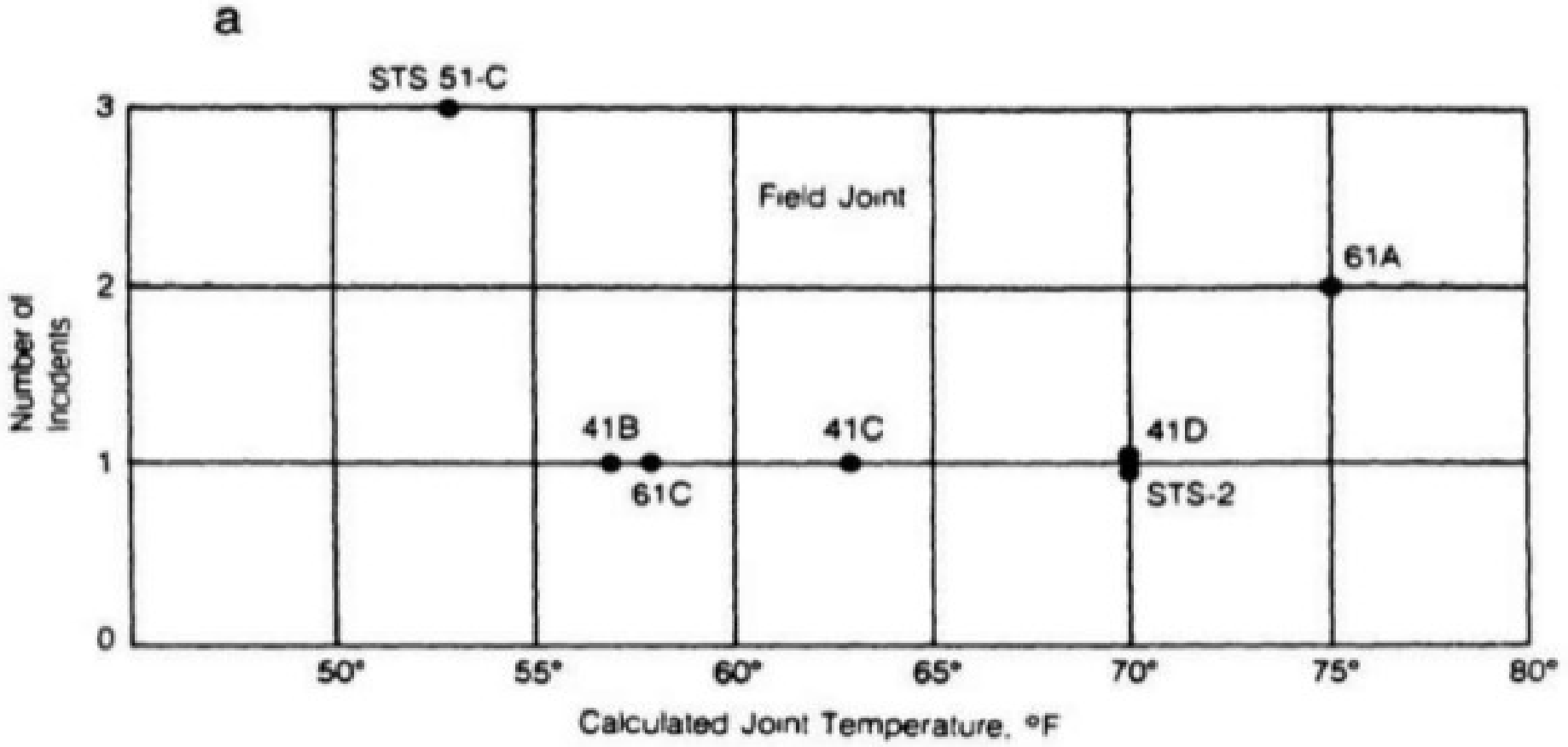


In two dimension: what anomalies look like?

Should have considered plot (b), which had data on 23 of the previous 24 flights.



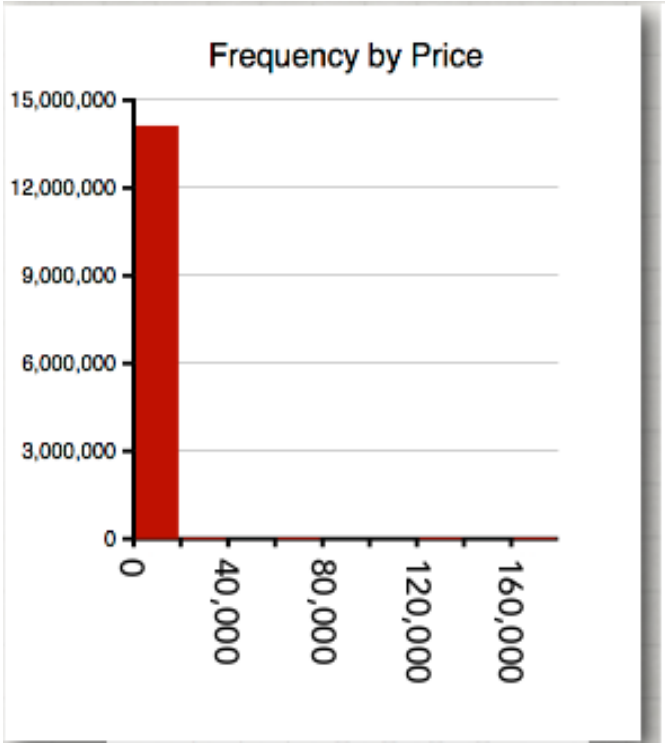
In two dimension: what anomalies look like?



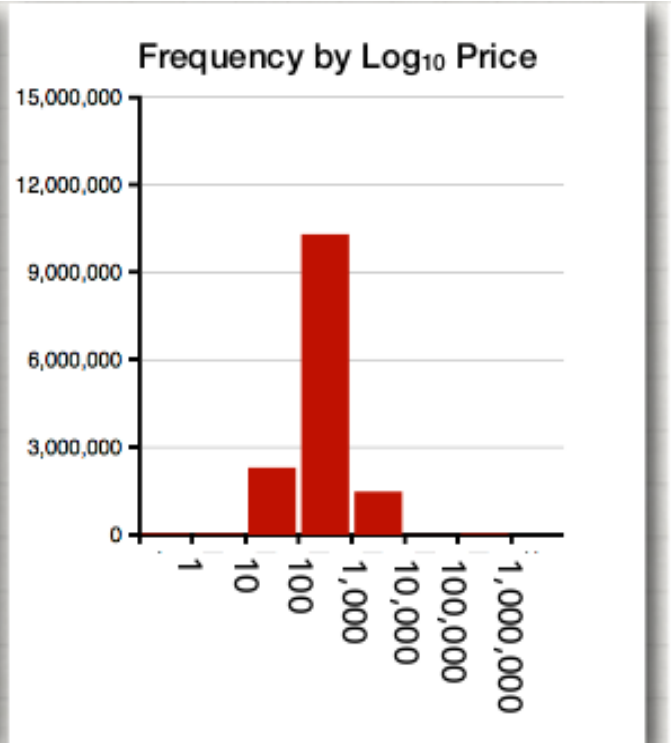
Based on plot (a), they concluded:

"Temperature data are not conclusive on predicting primary O-ring blowby."

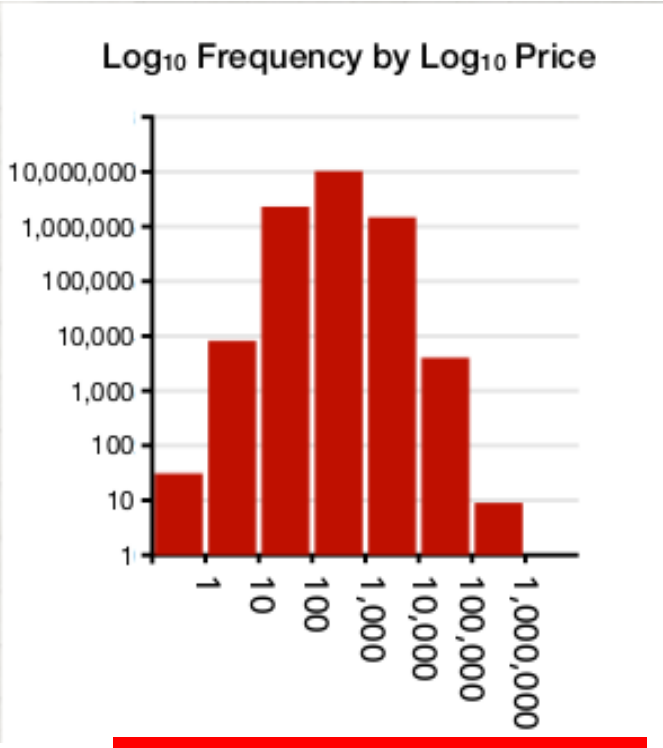
Using Logarithmic, log-log plots to display outliers can help



Linear histogram



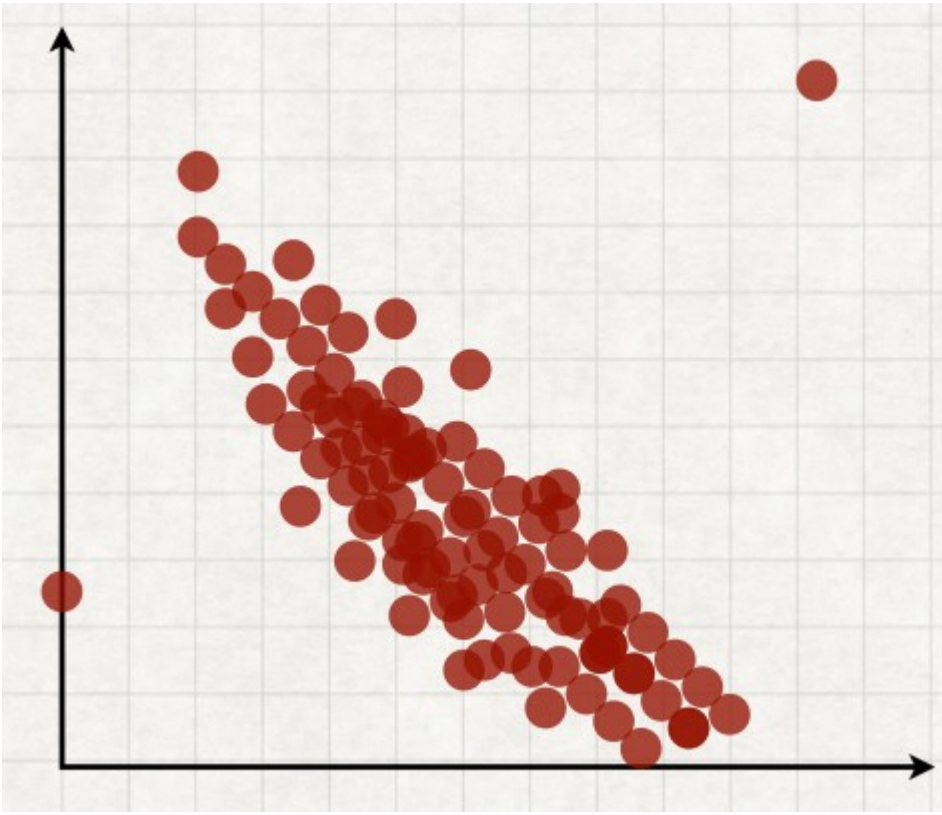
Frequency vs log10 Price



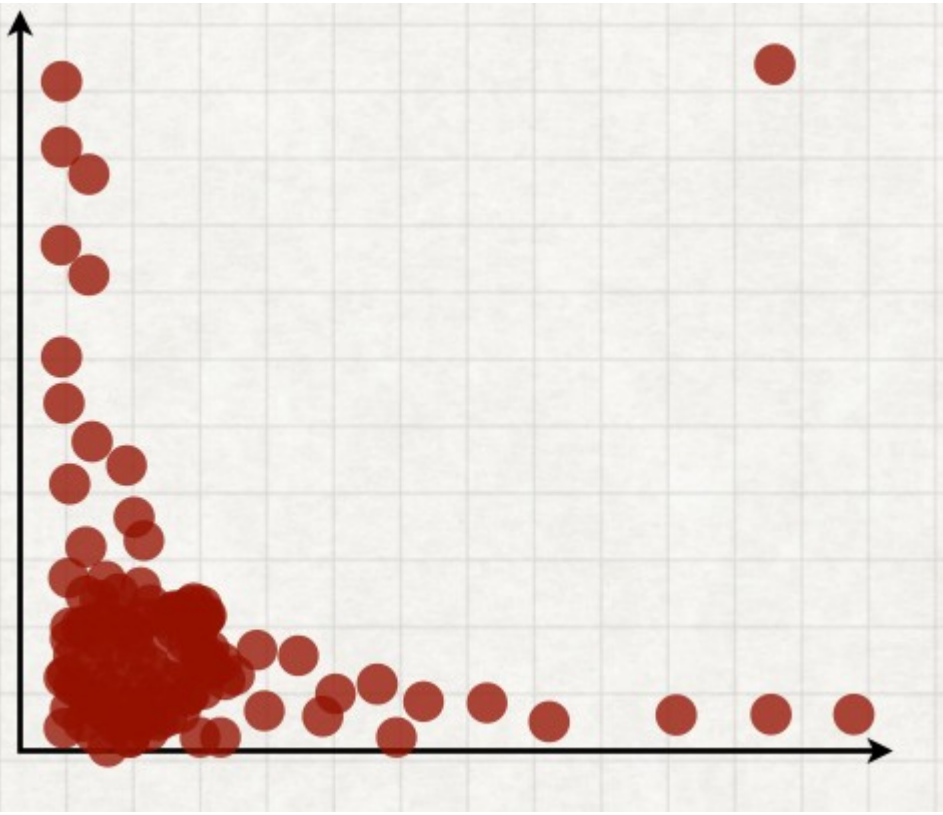
Log-log histogram log10 freq vs log10 price

pyplot.hist can "log" y axis for you with keyword argument log=True
Example : plt.hist(numpy.log10(data), log=True)

Using Logarithmic, log-log plots to display outliers can help



Linear plot



Log-log plot
matplotlib.pyplot.loglog

Outlier detection general applications

- Cyber security: Detect cyber-attacks on networks: more trusted connections
- Equipment failure (risk theory/théorie de la fiabilité) industrial sector
- Fraud detection
- Detecting cheaters in mobile gaming
- Preprocessing task for analysis or machine learning
- Reduce false declines then grow the revenue
- Detecting French regions where Le Pen’s scores at the 2017 presidential election deviate from predictions based on socio-economic variables (towardsdatascience)

Applications to financial sector (banking, insurance): Towards Fintech

- Retail bank: Credit card fraud
- Private bank: Market abuse, Anti-money laundering
- Investment bank: Market abuse (Flash Crash), Anti-money laundering
- Insurance companies: Fraudulent operations
- Central banks: detecting tax havens (panama papers)

Requires different approaches because Red flags are banking-type specific (specific business expertise)

Methodology?
Supervised learning vs. Unsupervised learning

Supervised vs. Unsupervised

- In credit card fraud detection one knows the target variable.

  How? Customers tell us. Can use supervised approach because the true class is self-revealing.

- In market abuse or money laundering detection we don’t really know classes (how money is laundered changes all the time and by the same criminal organisation)

Why supervised learning is difficult?

- Severe class imbalance : we estimate that 99.9% are trusted operations vs. 0.1% of fraudulent operations

- Problem during the train test split
- Solution you can use the stratification option in the train_test_split function of sklearn

And now why unsupervised is difficult?

- Sever class overlap : money laundering is mixed with legal financial activity, especially in Investment banks
- Uncertainty around the data model
- The complexity of data
- The huge volume of data (time complexity)

- Next we will discuss the data models
- To avoid time complexity: use dask, numba (used to speed numpy), ray and the newly module modin (used to speed pandas) + demonstration

Algorithms

- We can differentiate between three methods:

  - Distance based algorithms (similarity)

  K-NN for classification

  K-Means for clustering

  - Density-based (fitting a density)

  DBSCAN and HDBSCAN

  Local outlier factor (LOF)

  - Parametric

  Gaussian mixture models (GMM)

  Single class SVM

  Extreme value theory: Tukey outlier labeling

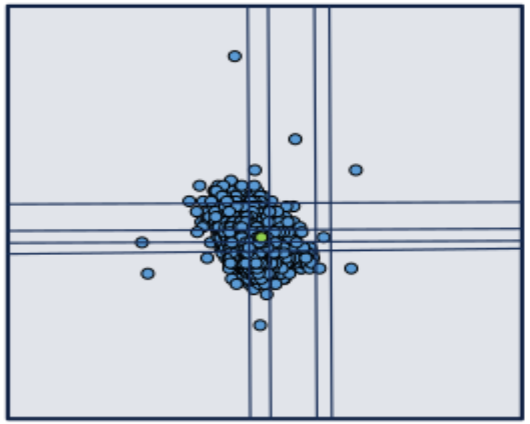
Tree ensemble algorithm : Isolation forest (used in Credit-swiss)

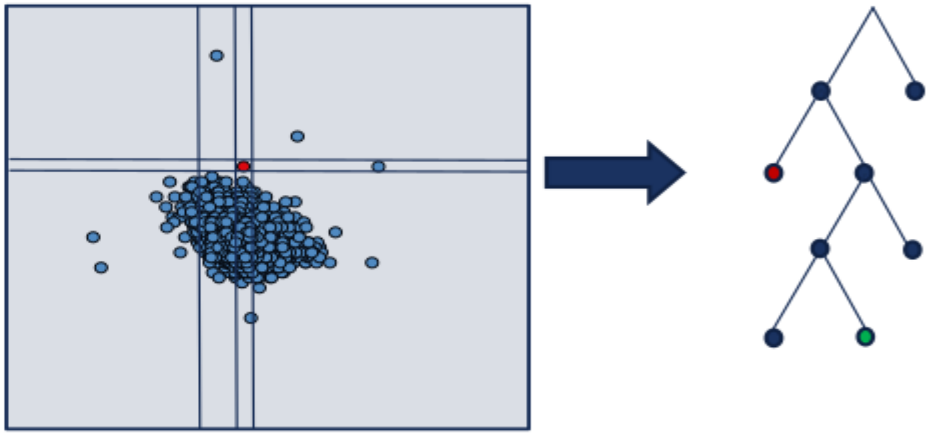
(F. T. Liu, et al., Isolation Forest, Data Mining, 2008. ICDM'08, Eighth IEEE International Conference)

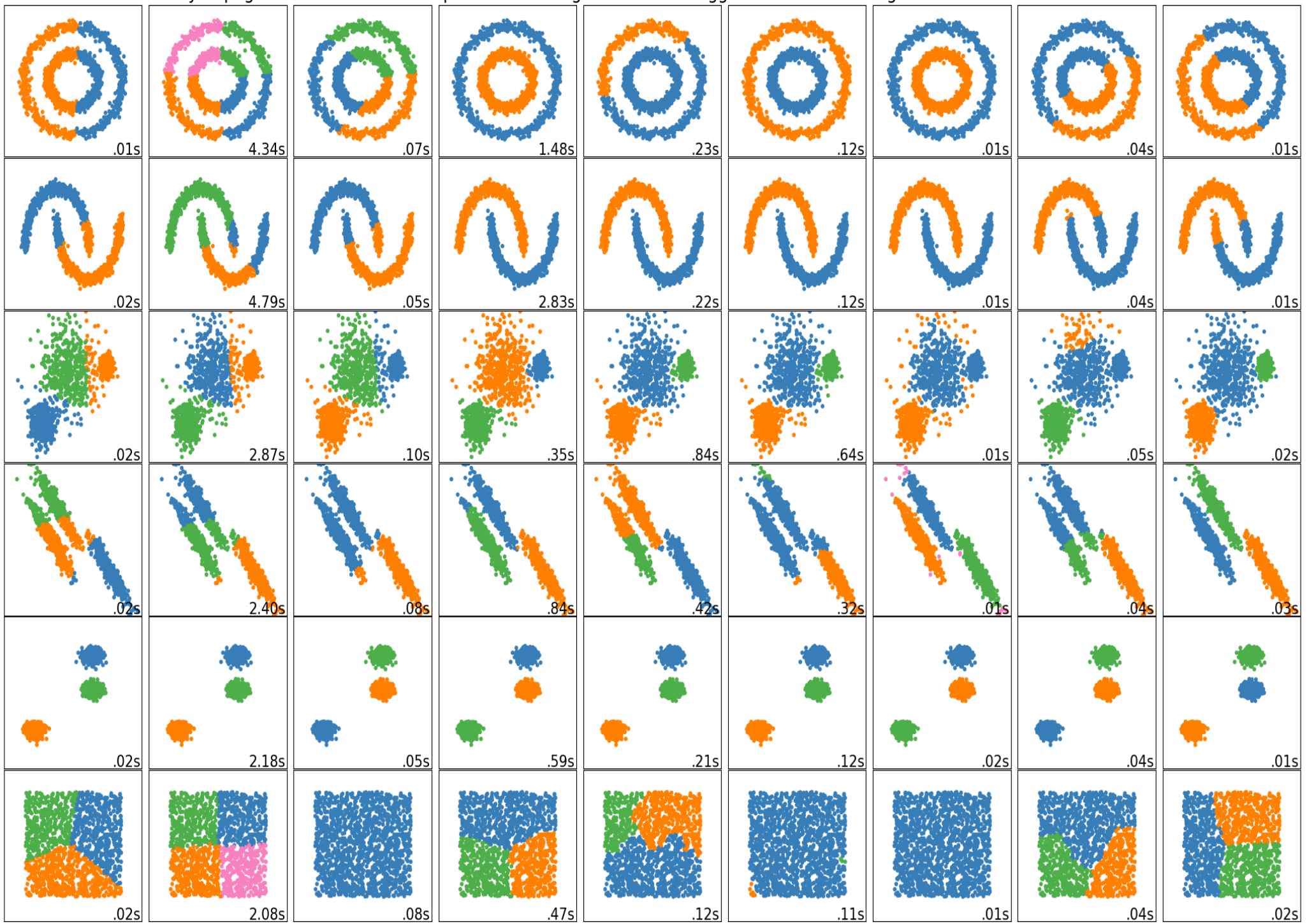
from sklearn.ensemble import IsolationForest

Ensemble regressor uses the concept of isolation to explain/separate-away anomalies

- No point based distance calculation
- Instead I.F. builds an ensemble of random trees for a given data set and anomalies are points with the shortest average path length






MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s

Unsupervised learning for anomaly detection (conclusion)

- Unsupervised learning is all about finding structure in data
- Techniques: Clustering (K-means, spectral clustering)
- Principle Components Analysis
- Support Vector machine
- **Autoencoder Deep Neural Networks : DNN autoencoder anomaly detection (exotic)**
- Filtering, Sequential Bayesian Filtering
- Gaussian mixture Model clustering via EM
- **LightGBM (Exotic)**