Unsupervised Learning

Jose Martinez Heras

11/04/2018

Resources

Watch the video of this lecture

https://dlmultimedia.esa.int/download/public/videos/2048/04/011/4804_011_AR_EN.mp4

Watch the practical exercise video

https://dlmultimedia.esa.int/download/public/videos/2048/04/010/4804_010_AR_EN.mp4

Get presentation and additional resources on

https://github.com/jmartinezheras/2018-MachineLearning-Lectures-ESA

Outline for Unsupervised Learning

Session 5: Unsupervised Learning

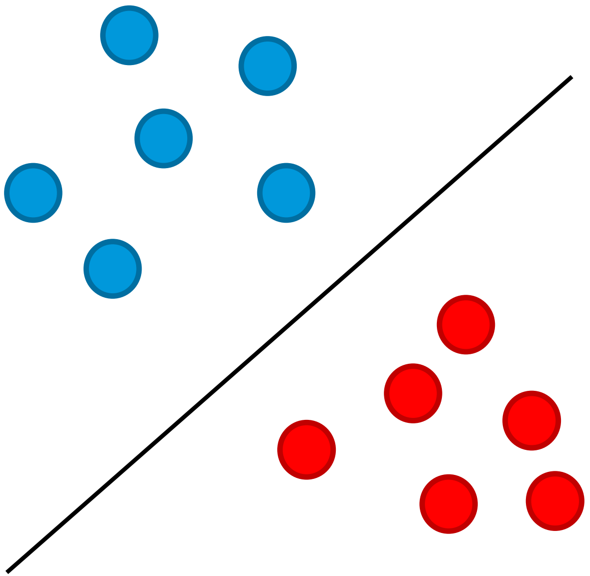
- Clustering
- Dimensionality Reduction: PCA, Auto-encoders

- Semi-supervised learning:
  - DrMUST
  - Novelty Detection

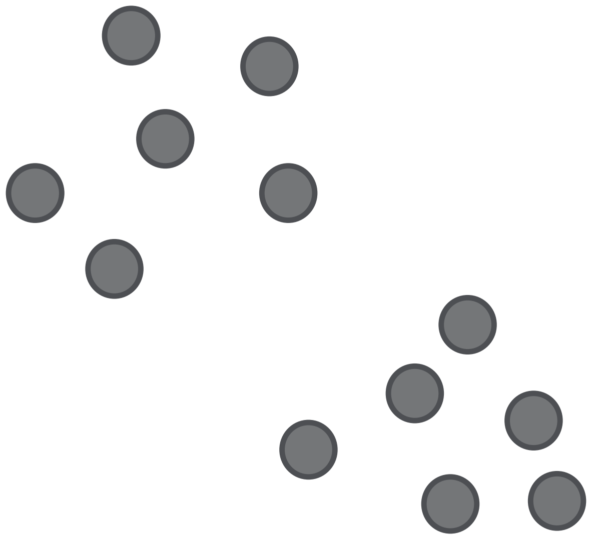
- Hands-on: group Dow Jones stocks according to their behavior

Unsupervised Learning

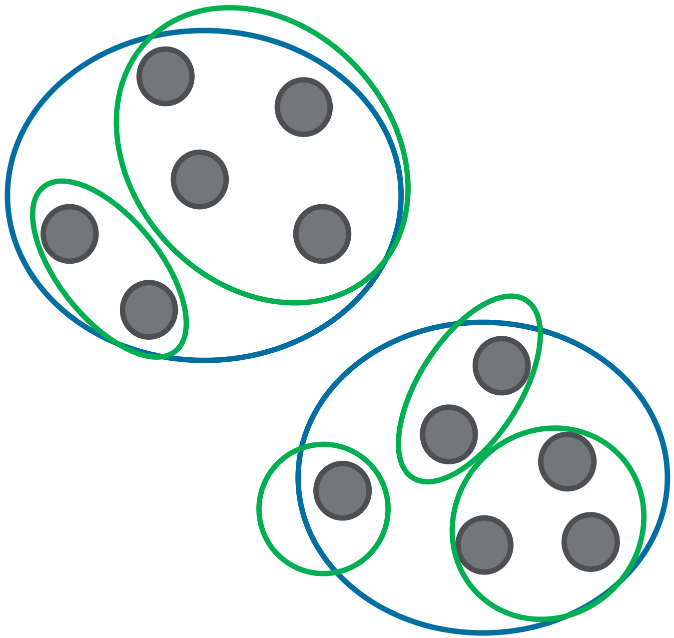
Supervised



Unsupervised



Clustering



Clustering is useful for understanding your data better.

Let’s say that we want to find the best grouping in 2 clusters

It’s a bit subjective

Let’s see some applications

Clustering applications – Customer Segmentation

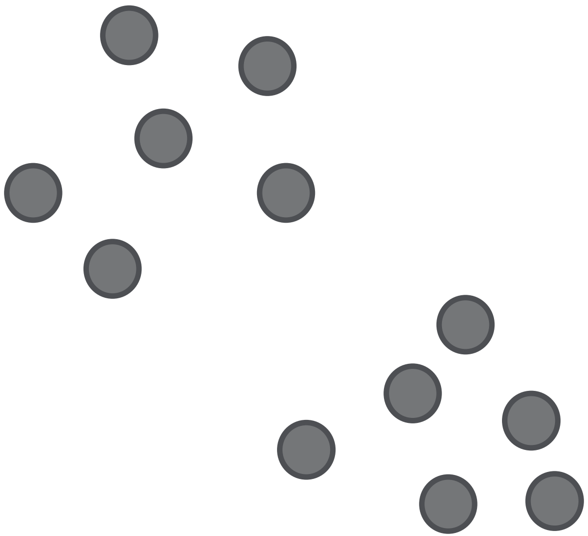
Image credit: https://select-statistics.co.uk/blog/customer-segmentation/

Clustering applications – Lossy Compression

Image credit: https://rpubs.com/yujingma45/155921

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

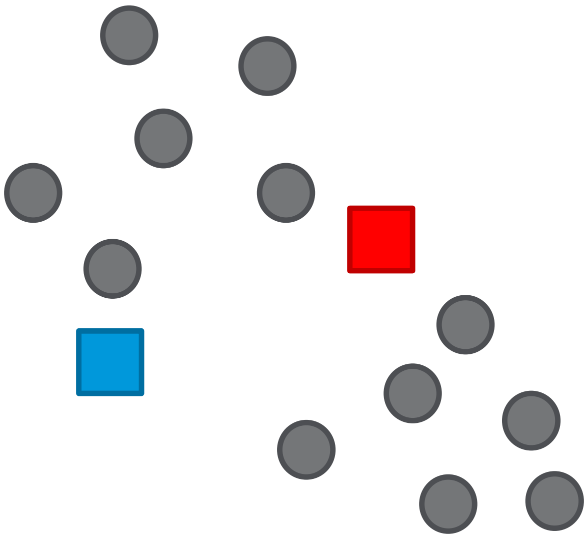
1. Initialize
2. Assign
3. Update

Let's run it for K=2

(2 clusters)

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

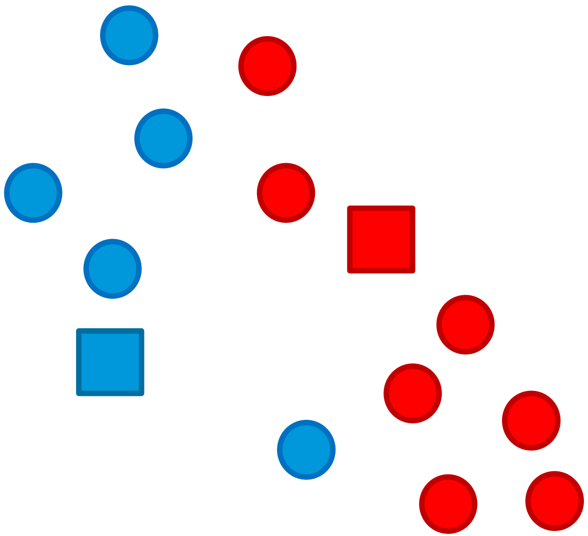
1. **Initialize**
2. Assign
3. Update

Let’s run it for K=2

(2 clusters)

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

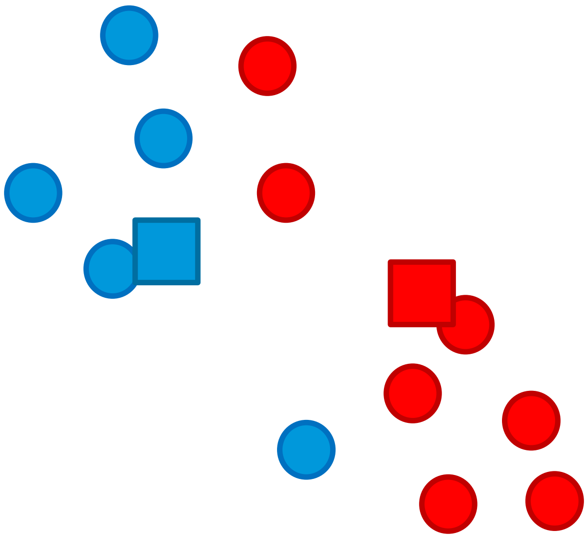
1. Initialize
2. **Assign**
3. Update

Let’s run it for K=2

(2 clusters)

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

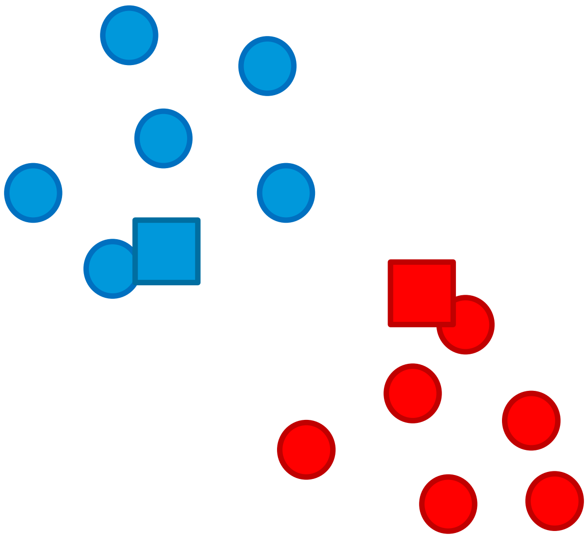
1. Initialize
2. Assign
3. **Update**

Let’s run it for K=2

(2 clusters)

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

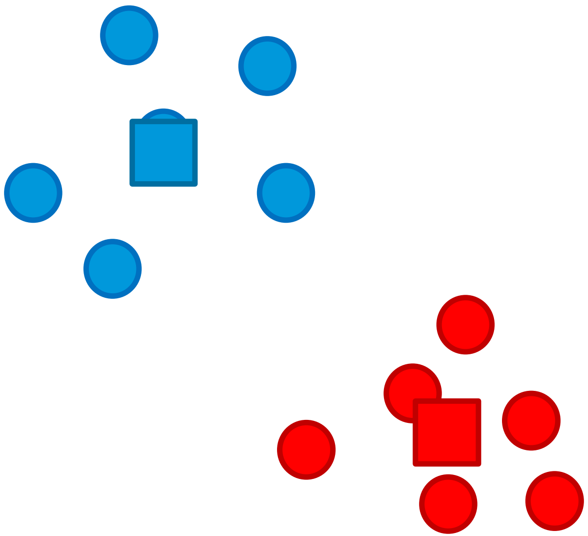
1. Initialize
2. **Assign**
3. Update

Let’s run it for K=2

(2 clusters)

Clustering – K-means

K-means is the most popular clustering algorithm



Algorithm

1. Initialize
2. Assign
3. **Update**

$$J = \frac{1}{m}\sum_{i=1}^{m}\left\|x_{(i)} - \mu_{c(i)}\right\|^2$$

Clustering Algorithms in scikit-learn

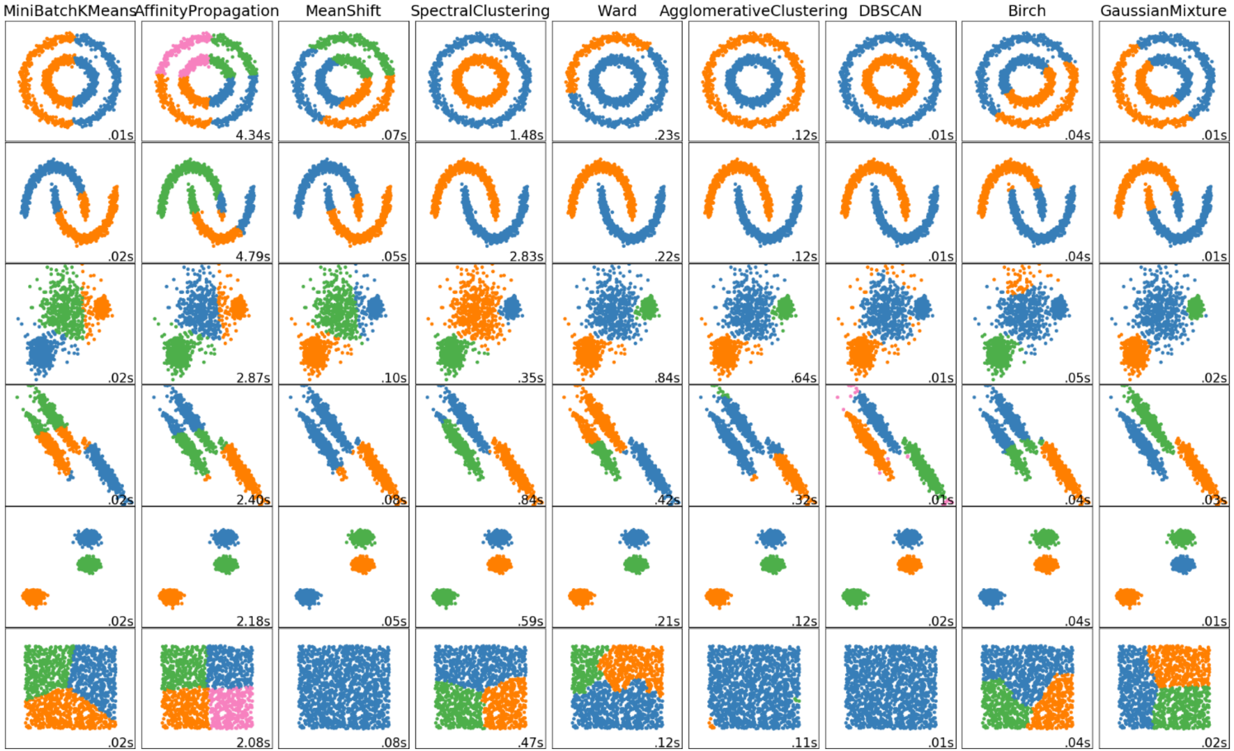


Image credit: http://scikit-learn.org/stable/modules/clustering.html

Clustering

- Allows to understand better your data by automatically finding groups
- Warning: Not that useful if you cannot tell which features are irrelevant

Imagine that instead of Price Sensitivity we would consider Eye Color

Since there is no target variable, we cannot figure out automatically what’s relevant

Image credit: https://select-statistics.co.uk/blog/customer-segmentation/

Dimensionality Reduction

Represent data in a lower dimension (fewer number of features)

Why is that useful?

- Machine Learning algorithms train faster

Dimensionality Reduction

Represent our data in a lower dimension (fewer number of features)

Why is that useful?

- Machine Learning algorithms train faster
- Get rid of redundant features

Women's Size Conversions

US Sizes	Euro Sizes	UK Sizes	Inches	CM
4	35	2	8.1875"	20.8
4.5	35	2.5	8.375"	21.3
5	35-36	3	8.5"	21.6
5.5	36	3.5	8.75"	22.2
6	36-37	4	8.875"	22.5
6.5	37	4.5	9.0625"	23
7	37-38	5	9.25"	23.5

Image credit: https://www.zappos.com/c/shoe-size-conversion

Dimensionality Reduction

Represent our data in a lower dimension (fewer number of features)

Why is that useful?

- Machine Learning algorithms train faster
- Get rid of redundant features
- Reduces the impact of the *curse of dimensionality*

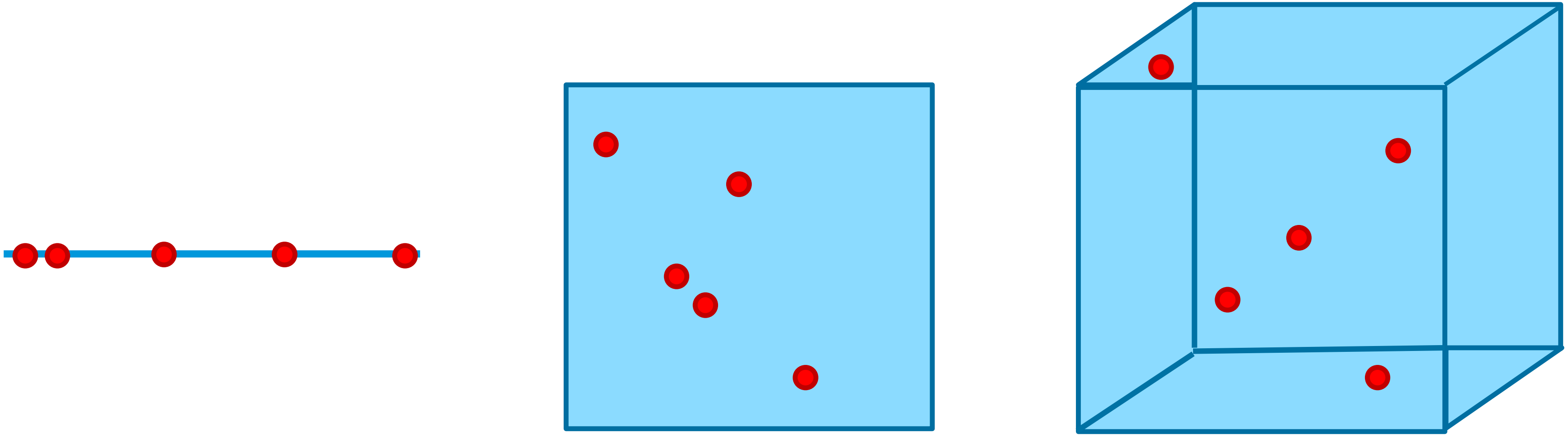
Women's Size Conversions

US Sizes	Euro Sizes	UK Sizes	Inches	CM
4	35	2	8.1875"	20.8
4.5	35	2.5	8.375"	21.3
5	35-36	3	8.5"	21.6
5.5	36	3.5	8.75"	22.2
6	36-37	4	8.875"	22.5
6.5	37	4.5	9.0625"	23
7	37-38	5	9.25"	23.5

Image credit: https://www.zappos.com/c/shoe-size-conversion

The Curse of Dimensionality

"When a measure such as a Euclidean distance is defined using many coordinates, there is **little difference in the distances** between different pairs of samples" [1]



[1] Wikipedia contributors, "Curse of dimensionality," *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Curse_of_dimensionality&oldid=831481596 (accessed April 3, 2018).

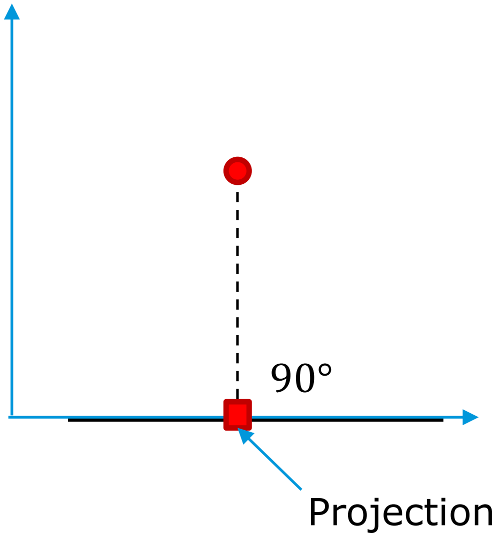
Dimensionality Reduction

- Principal Components Analysis (PCA)
- Auto-Enconders

- There are many other techniques

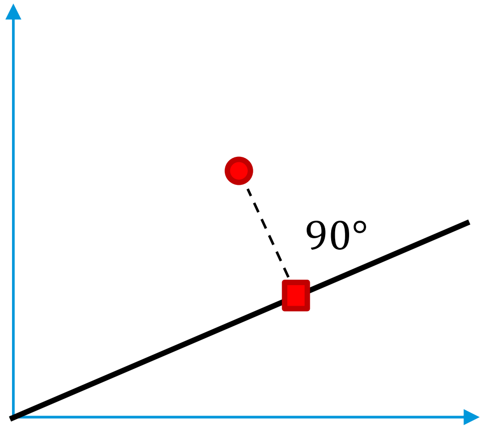
Principal Components Analysis - PCA

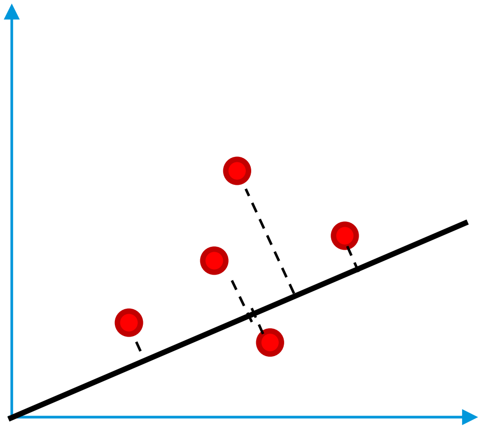
Projections



Principal Components Analysis - PCA

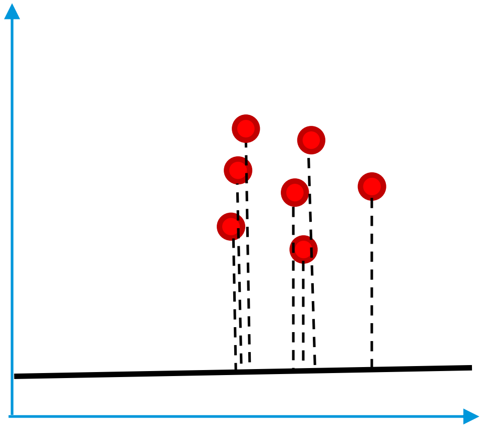
Projections



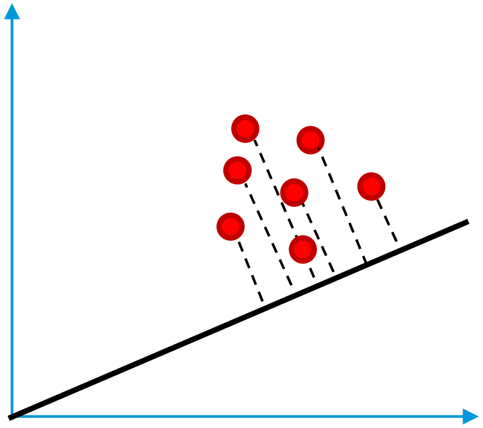


Principal Components Analysis - PCA

In which line is better to project?

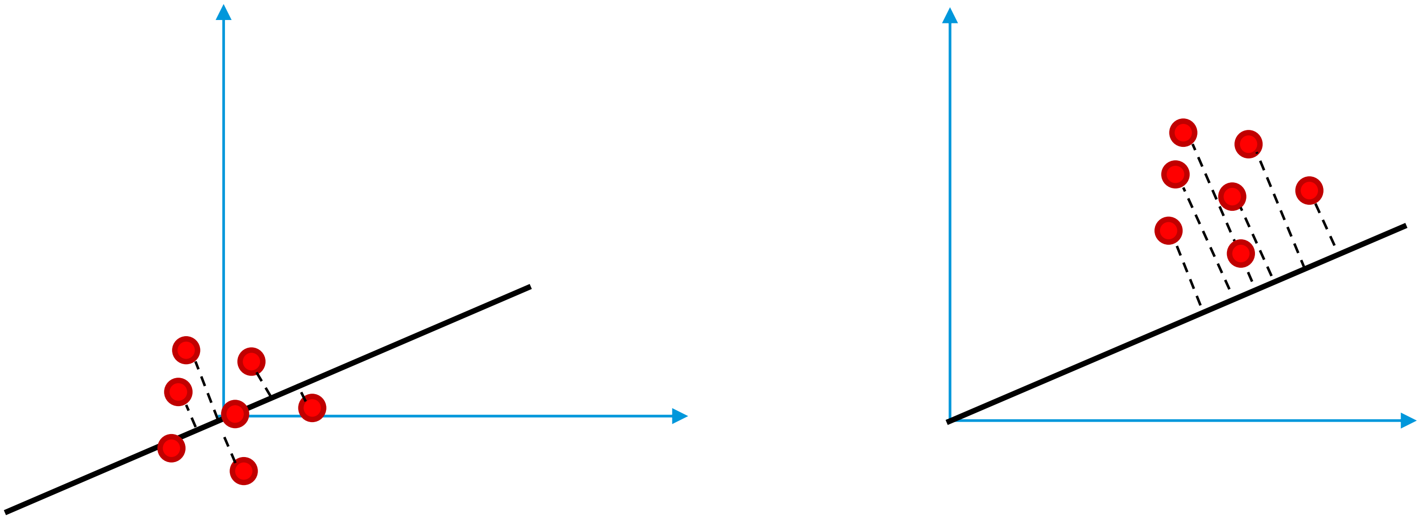


This line is better because it maximizes the variance of the projection



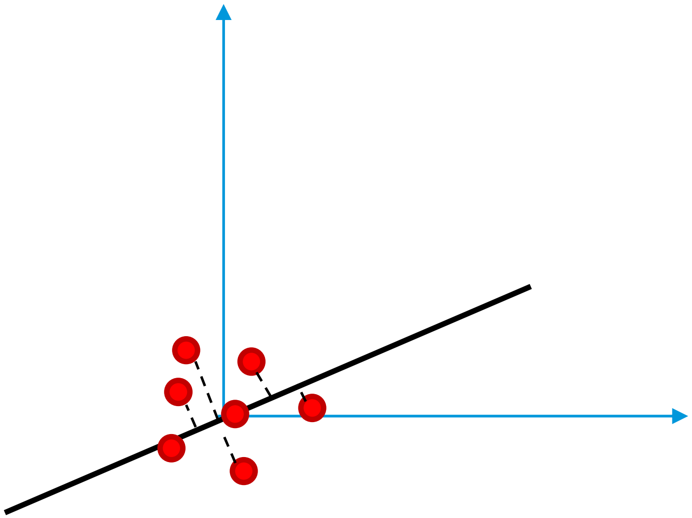
Principal Components Analysis - PCA

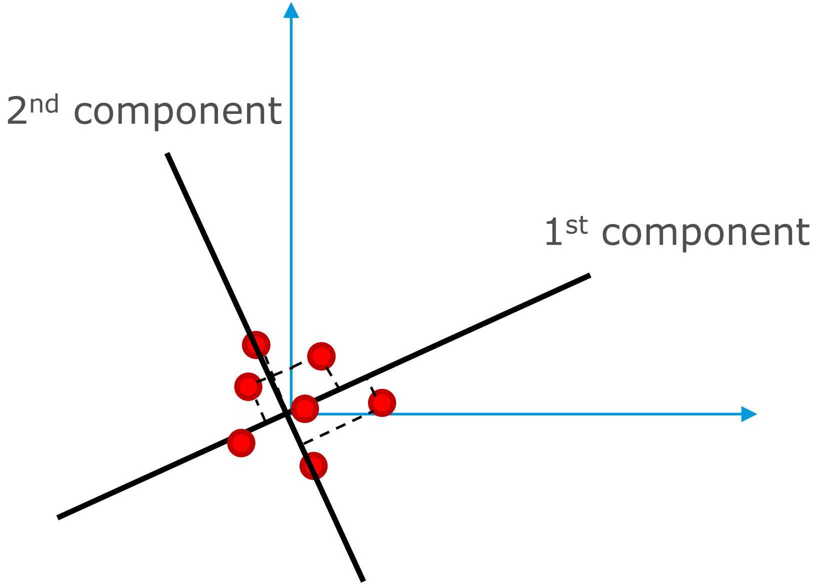
In practice, data is scaled so that all features are in the same scale



Principal Components Analysis - PCA

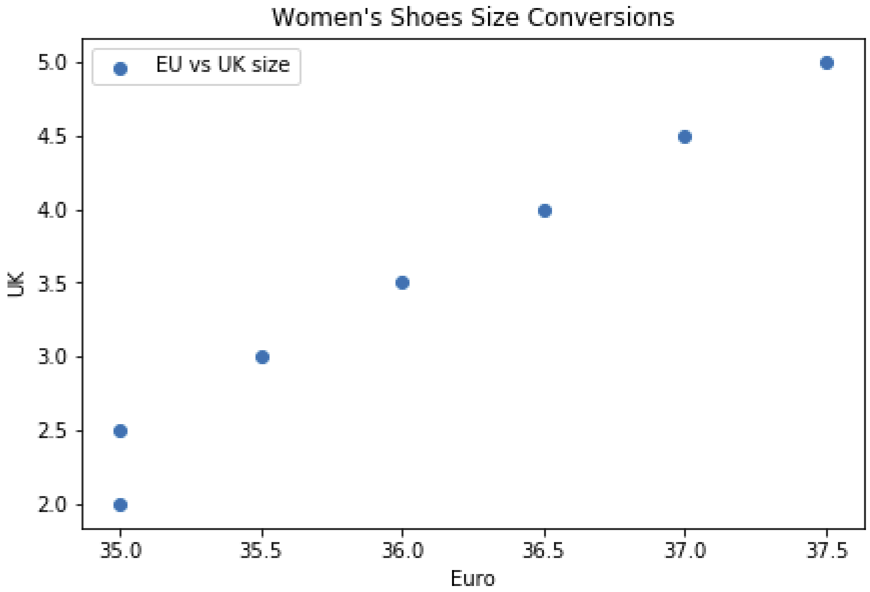
Each PCA Component is orthogonal to all the others





Principal Components Analysis - PCA

Example with women's shoes size conversions



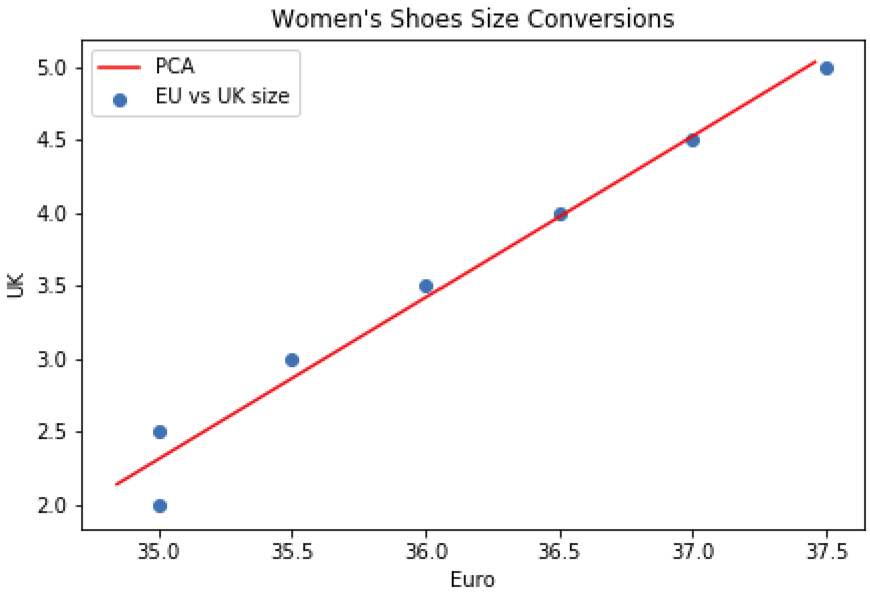
Women's Size Conversions

US Sizes	Euro Sizes	UK Sizes	Inches	CM
4	35	2	8.1875"	20.8
4.5	35	2.5	8.375"	21.3
5	35-36	3	8.5"	21.6
5.5	36	3.5	8.75"	22.2
6	36-37	4	8.875"	22.5
6.5	37	4.5	9.0625"	23
7	37-38	5	9.25"	23.5

Image credit: https://www.zappos.com/c/shoe-size-conversion

Principal Components Analysis - PCA

Example with women's shoes size conversions



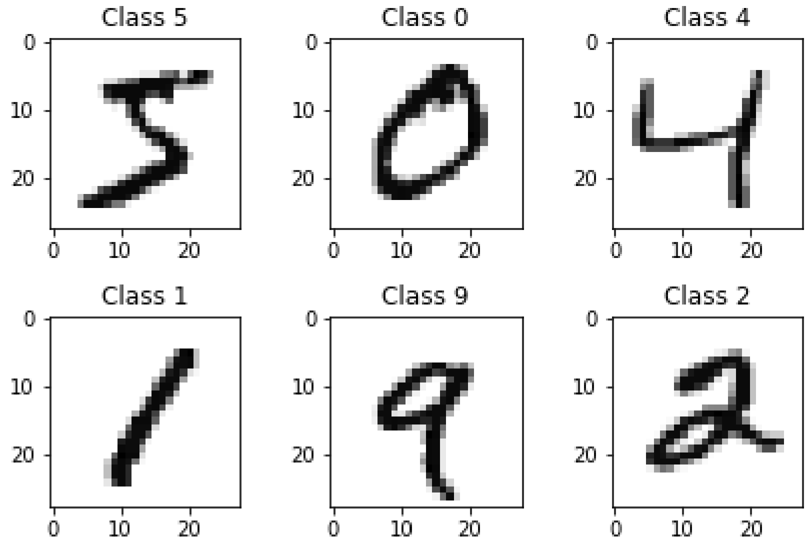
Women's Size Conversions

US Sizes	Euro Sizes	UK Sizes	Inches	CM
4	35	2	8.1875"	20.8
4.5	35	2.5	8.375"	21.3
5	35-36	3	8.5"	21.6
5.5	36	3.5	8.75"	22.2
6	36-37	4	8.875"	22.5
6.5	37	4.5	9.0625"	23
7	37-38	5	9.25"	23.5

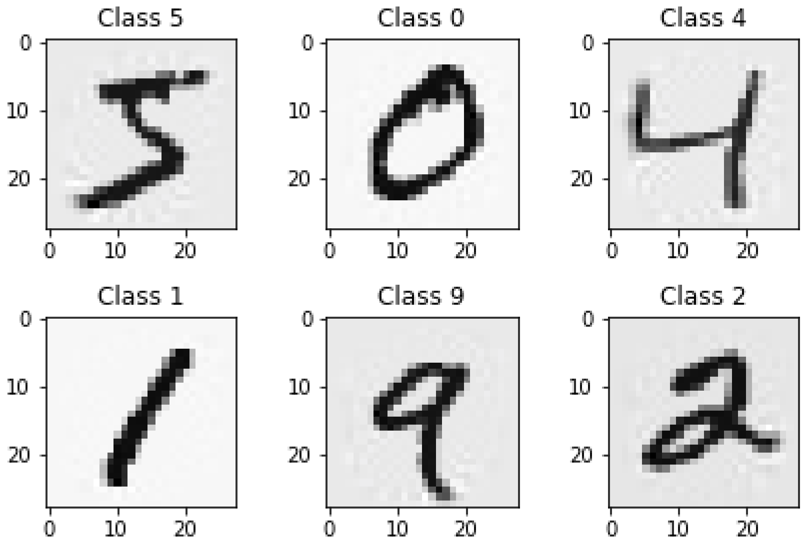
Image credit: https://www.zappos.com/c/shoe-size-conversion

Principal Components Analysis - PCA

MNIST dataset: original dimension 28x28 = 764



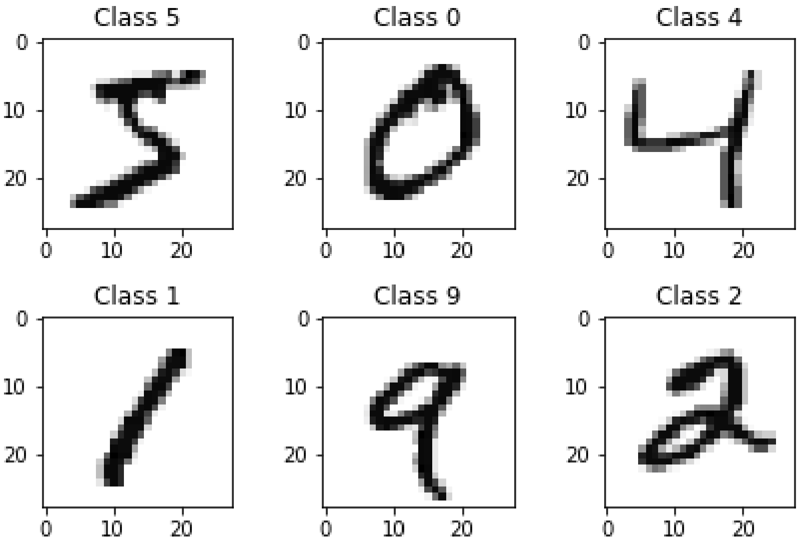
764 dimensions



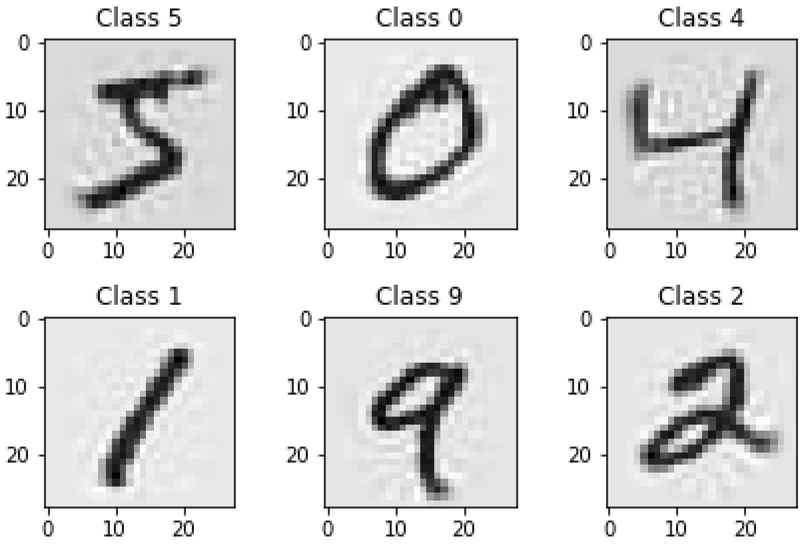
392 dimensions

Principal Components Analysis - PCA

MNIST dataset: original dimension 28x28 = 764



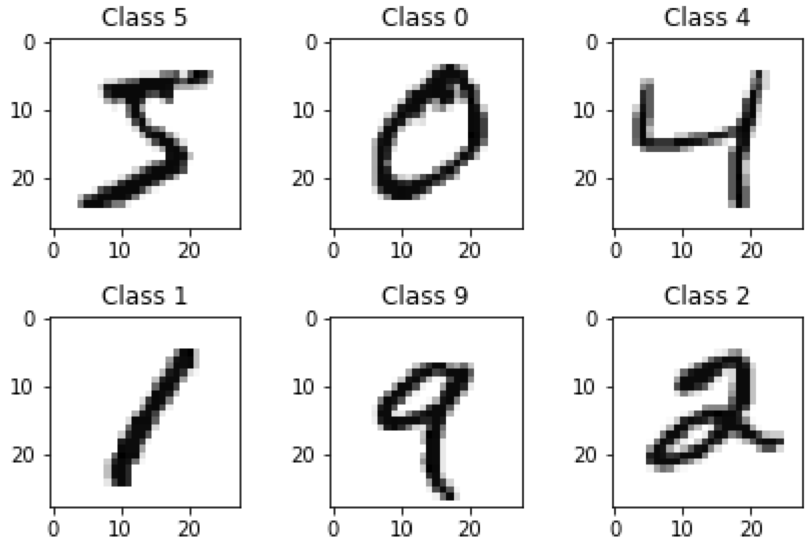
764 dimensions



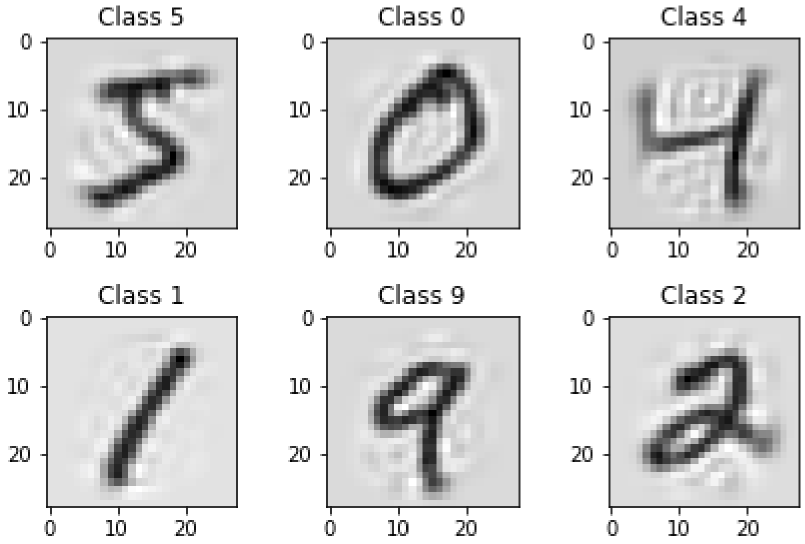
196 dimensions

Principal Components Analysis - PCA

MNIST dataset: original dimension 28x28 = 764



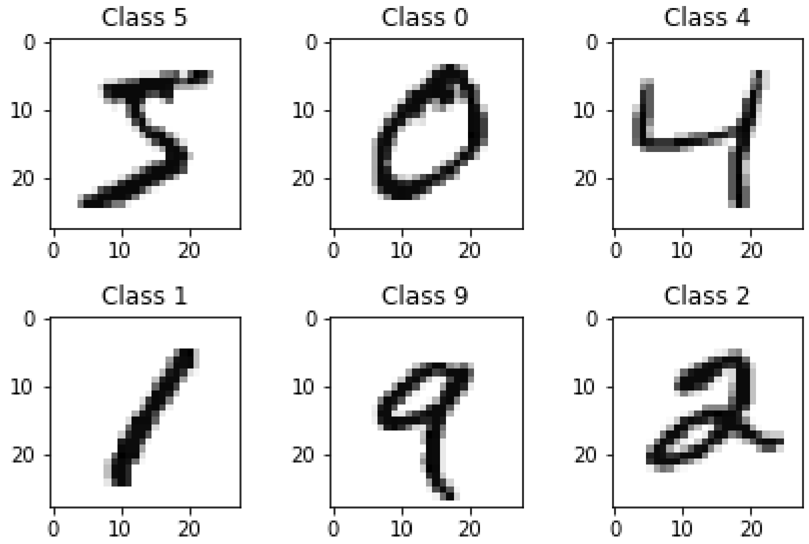
764 dimensions



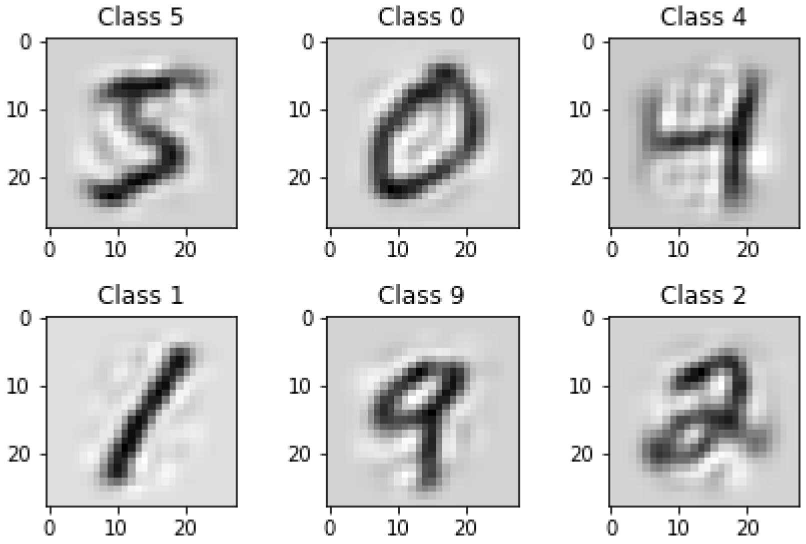
98 dimensions

Principal Components Analysis - PCA

MNIST dataset: original dimension 28x28 = 764



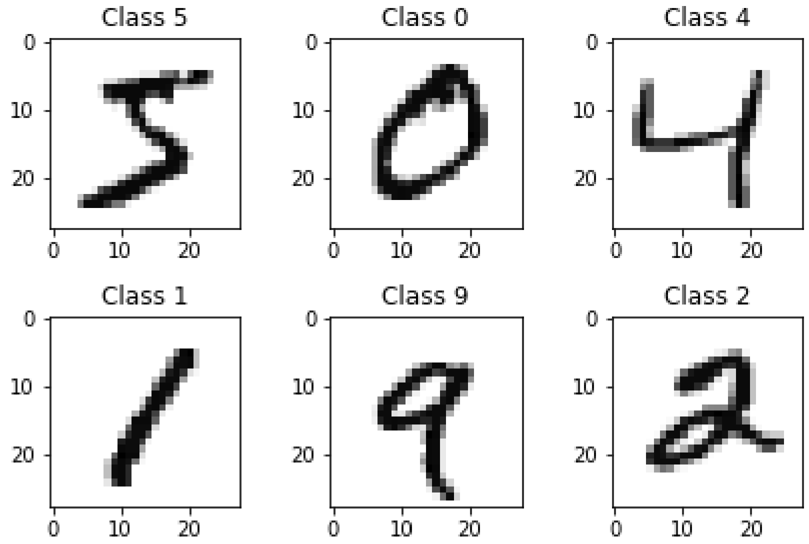
764 dimensions



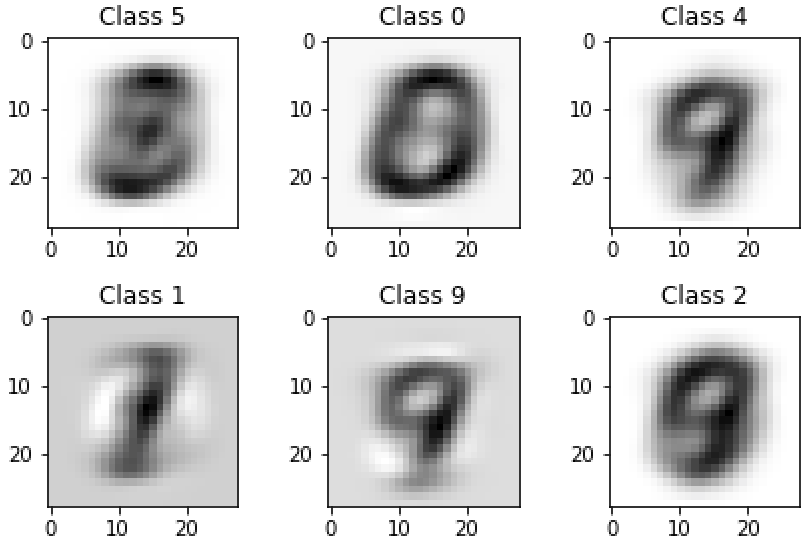
49 dimensions

Principal Components Analysis - PCA

MNIST dataset: original dimension 28x28 = 764



764 dimensions



2 dimensions

Auto-Encoders

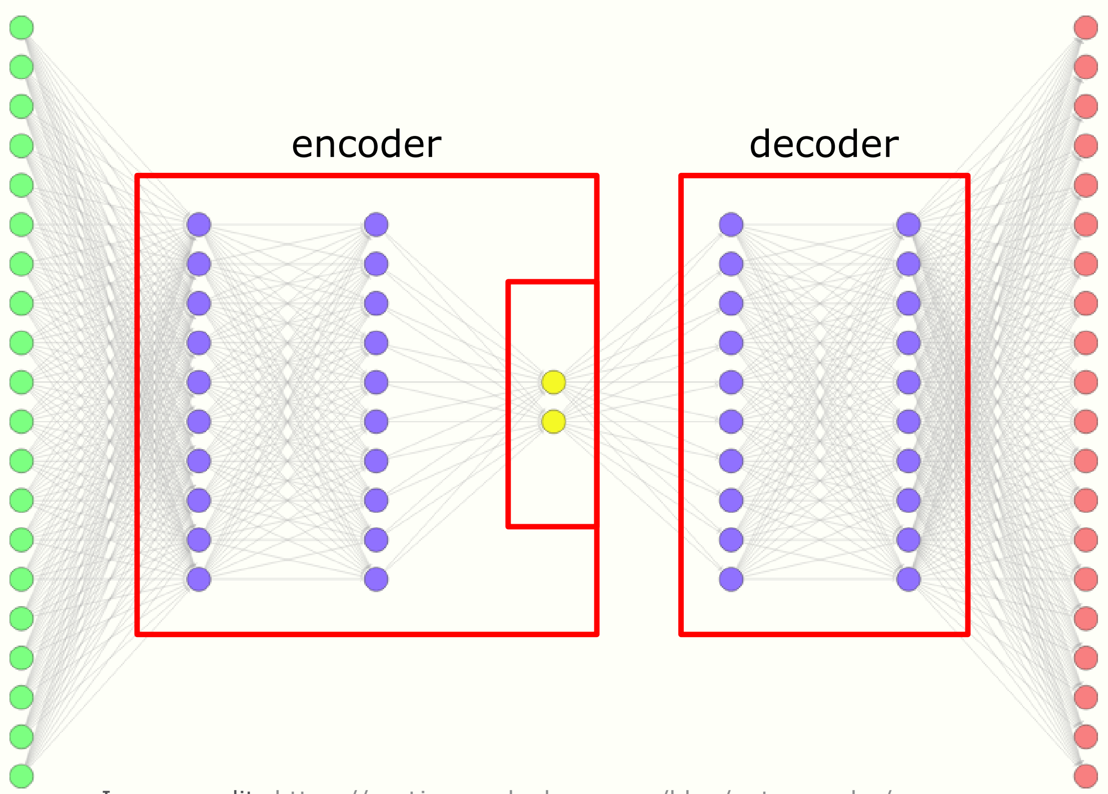
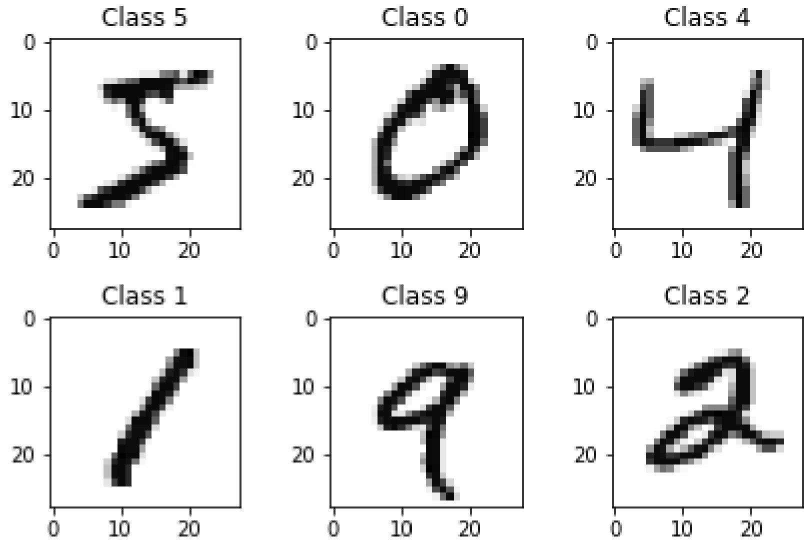


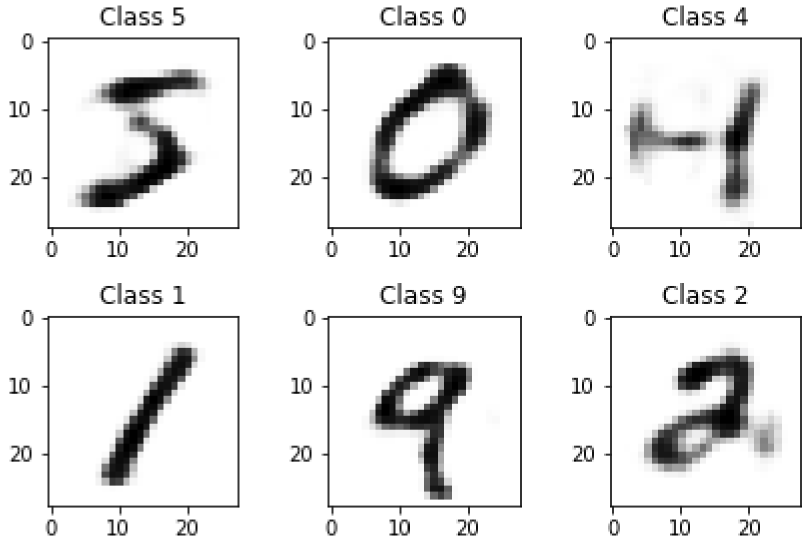
Image credit: https://gertjanvandenburg.com/blog/autoencoder/

Auto-Encoders

Auto-Encoder: 764 – $\boxed{50 – 50 –}\boxed{2}$ – $\boxed{50 – 50}$ – 764



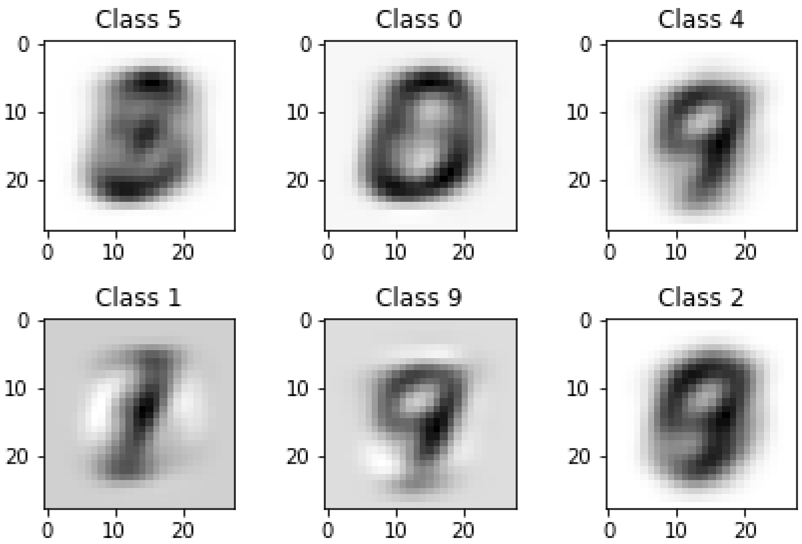
764 dimensions



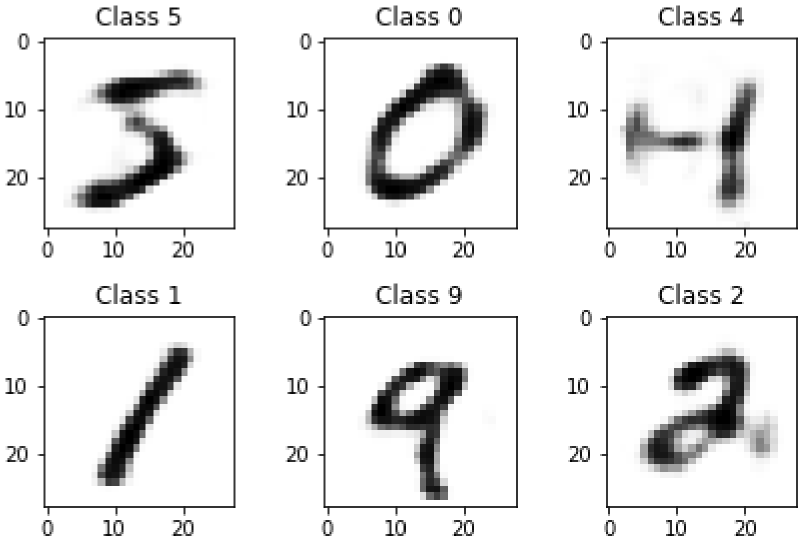
2 dimensions

PCA vs Auto-Encoders

Auto-Encoders capture better the non-linear relationships among the inputs



2 dimensions PCA



2 dimensions auto-encoder

Dimensionality Reduction

Represent our data in a lower dimension (fewer number of features)

Why is that useful?

- Machine Learning algorithms train faster
- Get rid of redundant features
- Reduces the impact of the *curse of dimensionality*
- Can also be used for lossy data compression
- Low dimensions are useful to visualize compressed representations

- Prototype: anomaly detection (by comparing reconstruction from real)

Semi-Supervised Learning

We can make some assumptions about data

Anomaly Investigation

- We know when the anomaly was
- We know when it was nominal

DrMUST

Anomaly Detection

- We know when it was nominal

Novelty Detection

DrMUST – Anomaly Investigation

What's an anomaly?

- We didn't want this to happen
- We wanted this to happen but it didn't happen

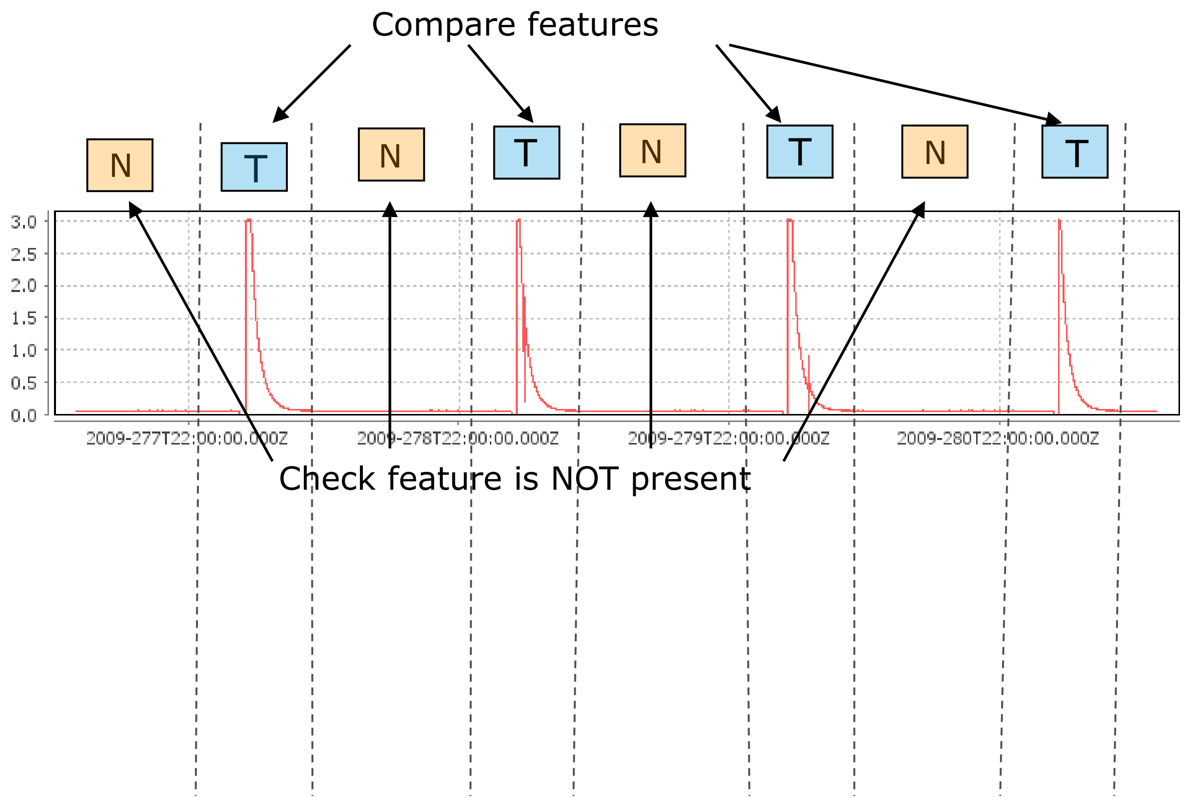
DrMUST

- Other parameters may hold the key to understand why the anomaly happened
- Check all 20,000+ TM parameters (and TCs and Events) for insights

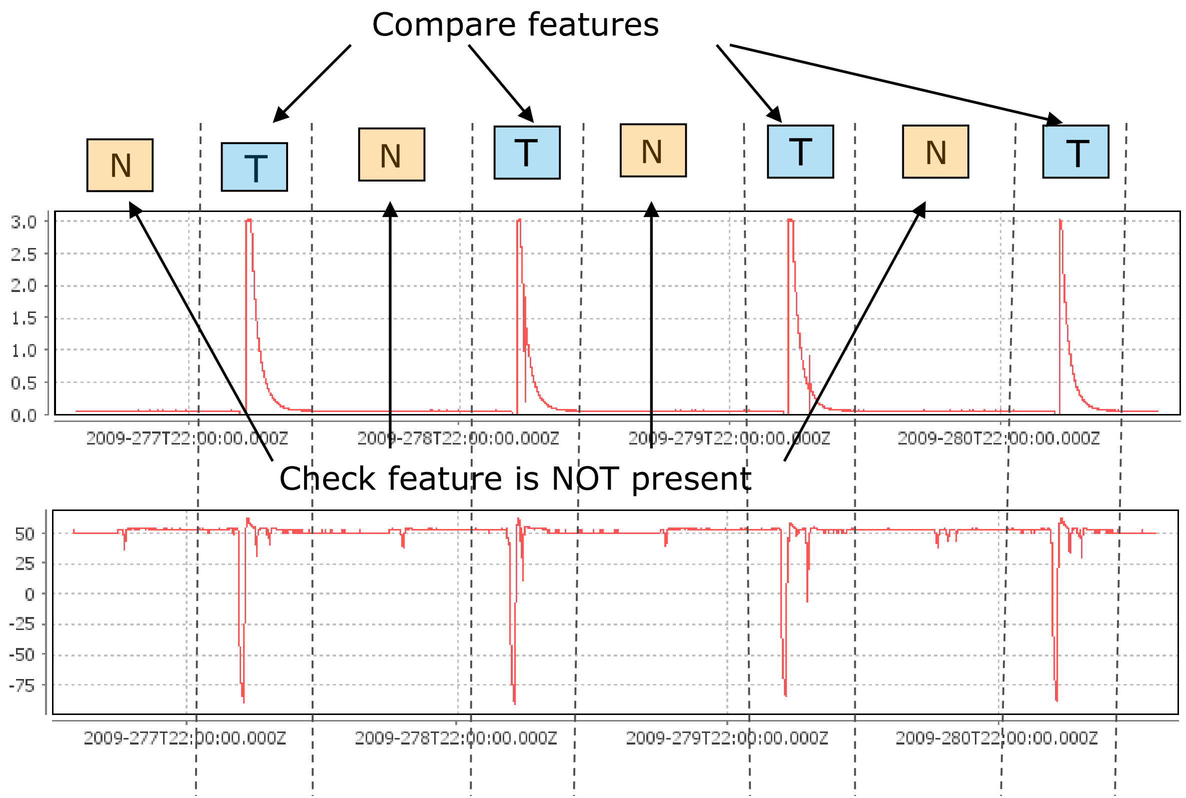
DrMUST – Anomaly Investigation

Assumption: parameters involved in the anomaly behave differently during nominal and anomaly periods

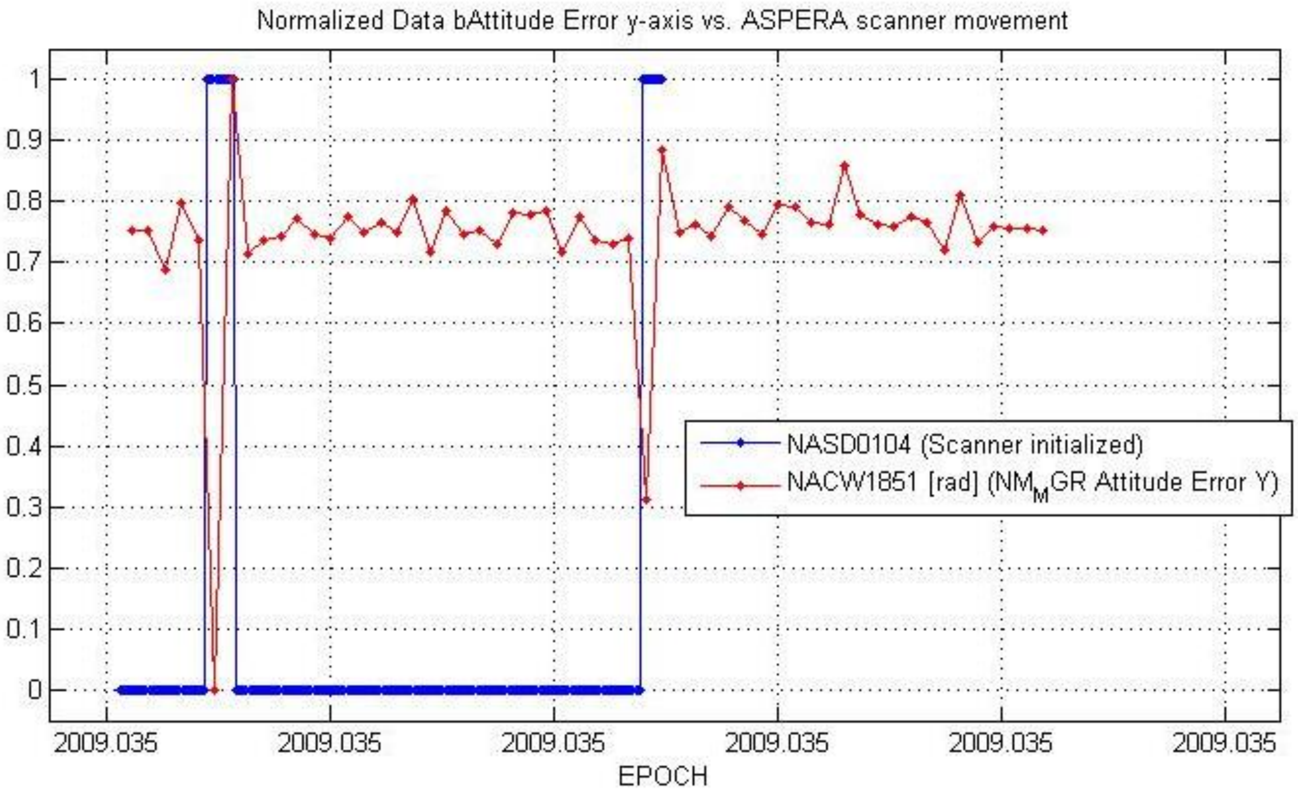
DrMUST – Anomaly Investigation



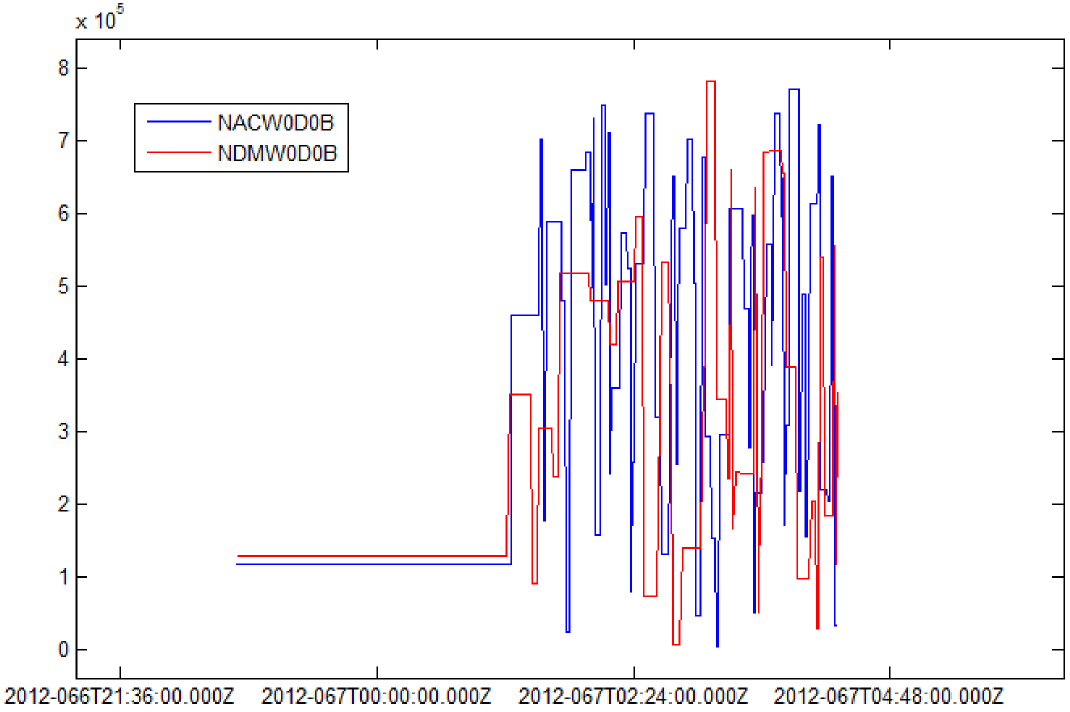
DrMUST – Anomaly Investigation



DrMUST - Examples



DrMUST - Characterization



Number of bit errors detected & corrected

In addition of some star tracker parameters affected

Novelty Detection

- Goal: automatically detect anomalies – as early as possible

What's an anomaly?

- We didn't want this to happen
- We wanted this to happen but it didn't happen

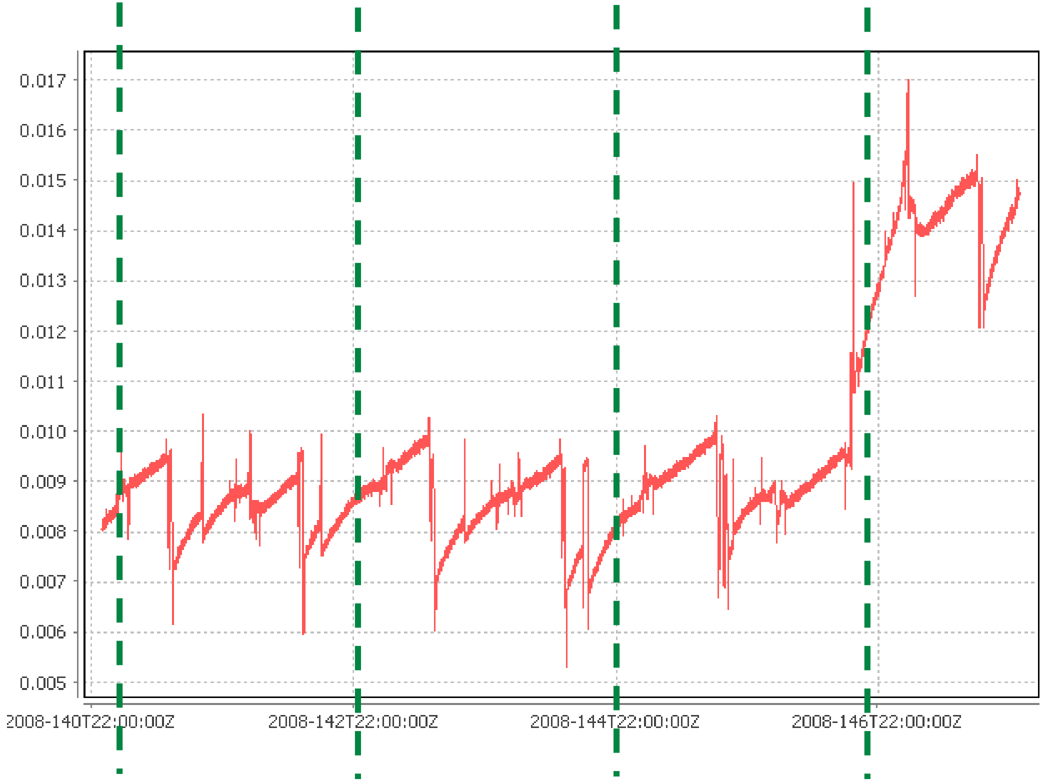
But … we don't know what you wanted …

- Anomaly Detection → Novelty Detection
- Novelty Detection looks for unusual behaviour
    - unusual behaviour is often the signature of an anomaly in the way to happen

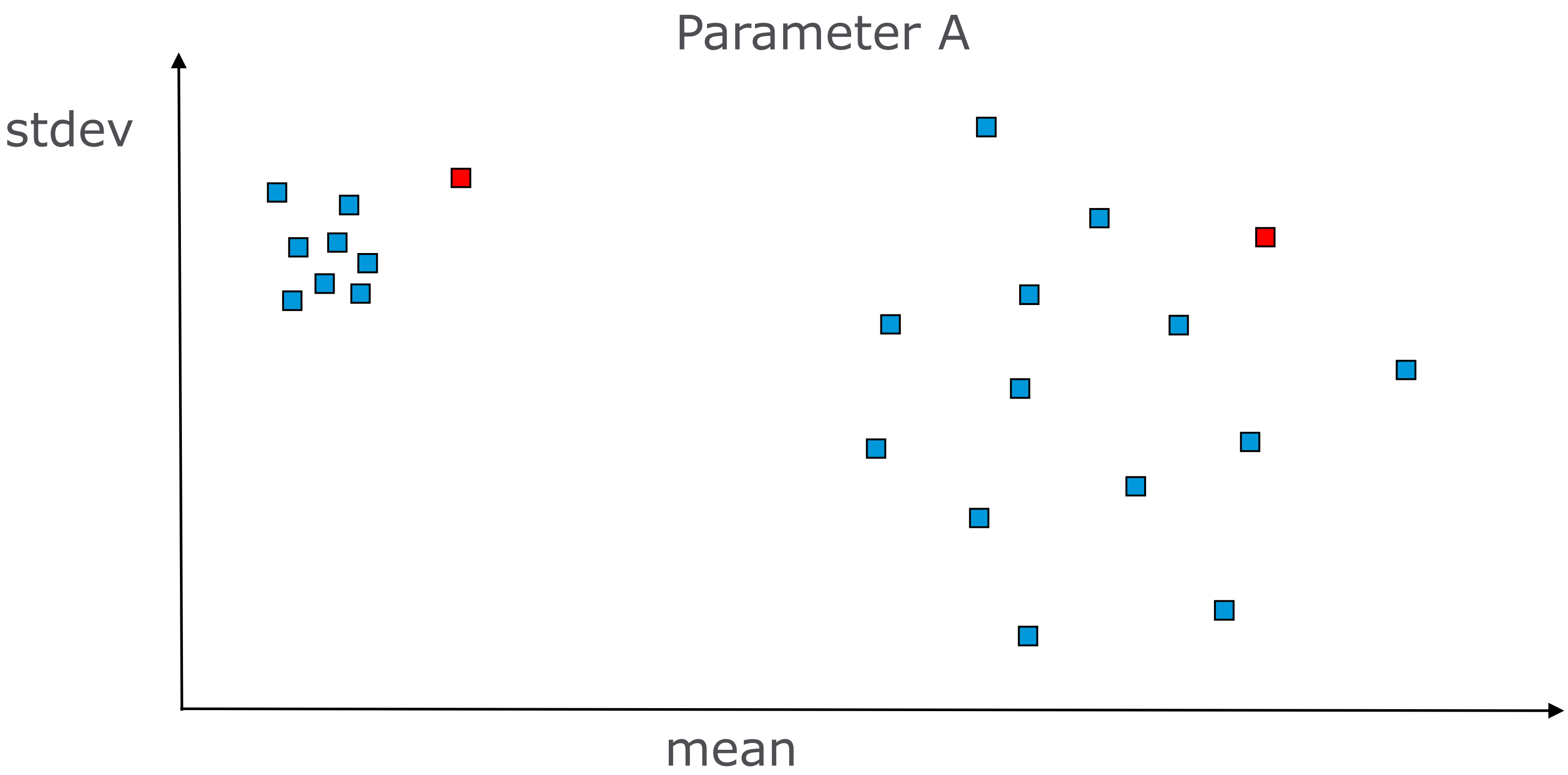
Novelty Detection

Split parameters time series in periods (e.g. 1 day) and extract 4 features:

- Mean
- Standard Deviation
- Maximum
- Minimum



Novelty Detection



Novelty Detection – Local Outlier Probabilities

1. Makes no assumptions on how nominal behaviour should be
2. Takes into account that a parameter can have different nominal behaviours
3. Takes into account density (no distance threshold required)
4. Outlier probability allows for comparison among different parameters

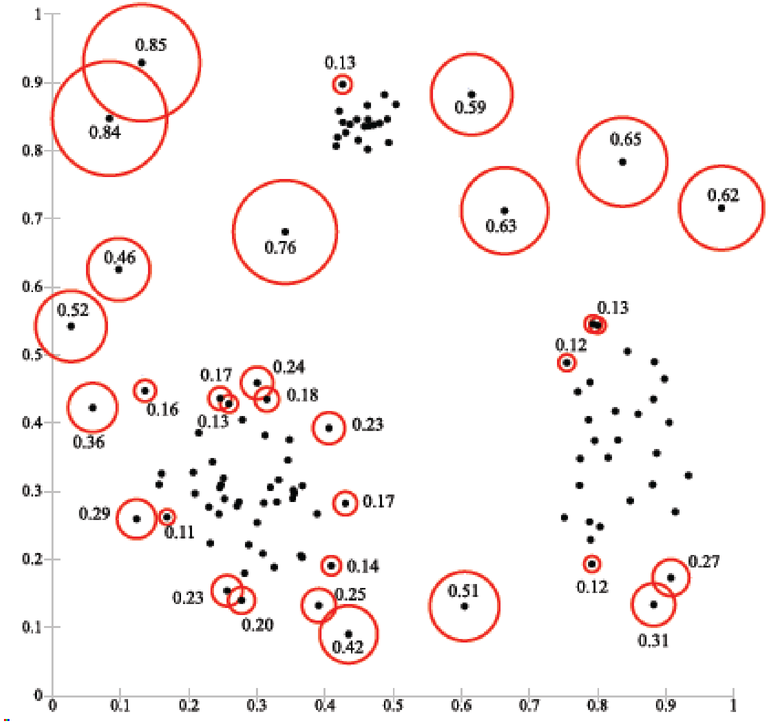
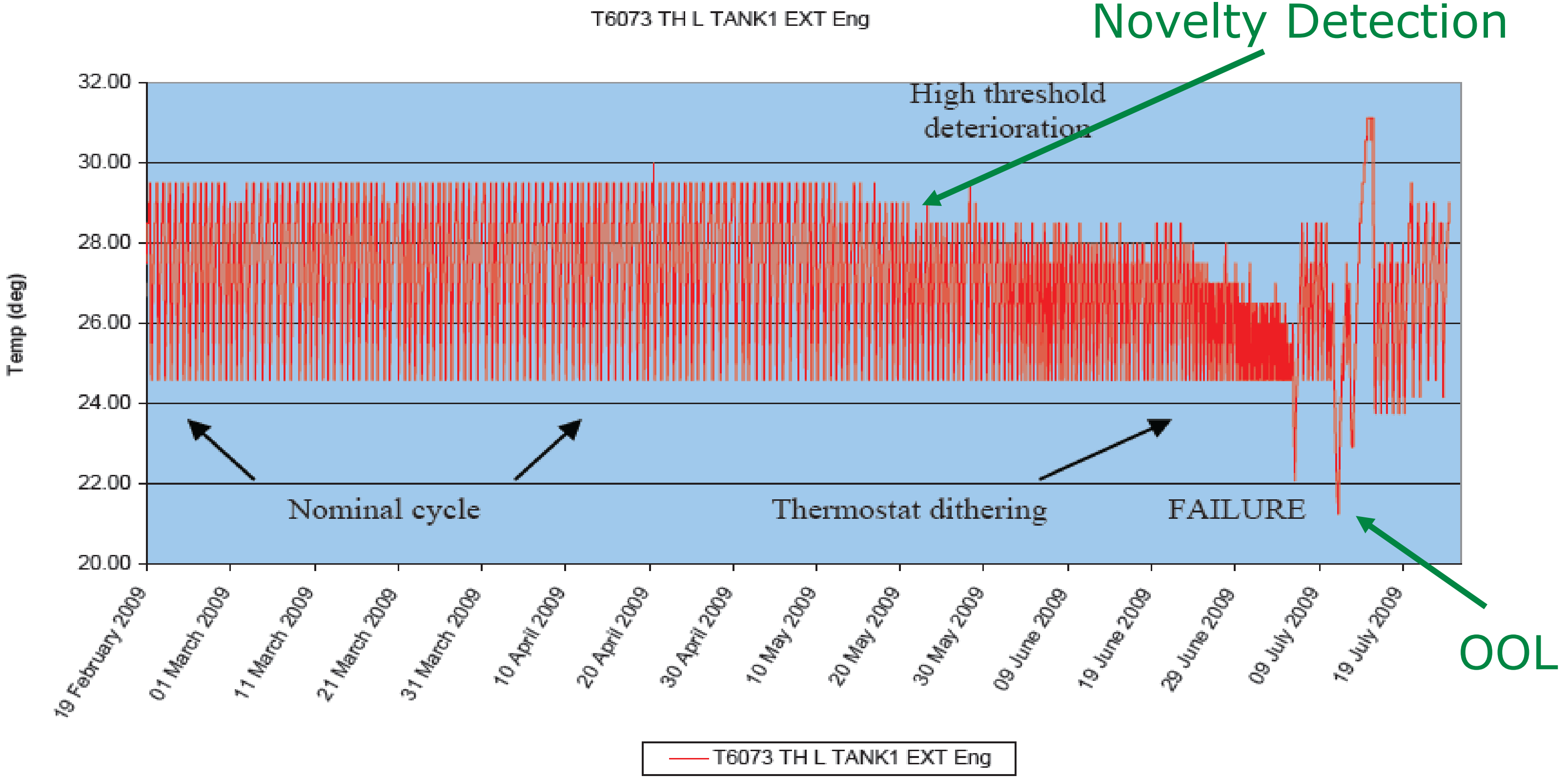


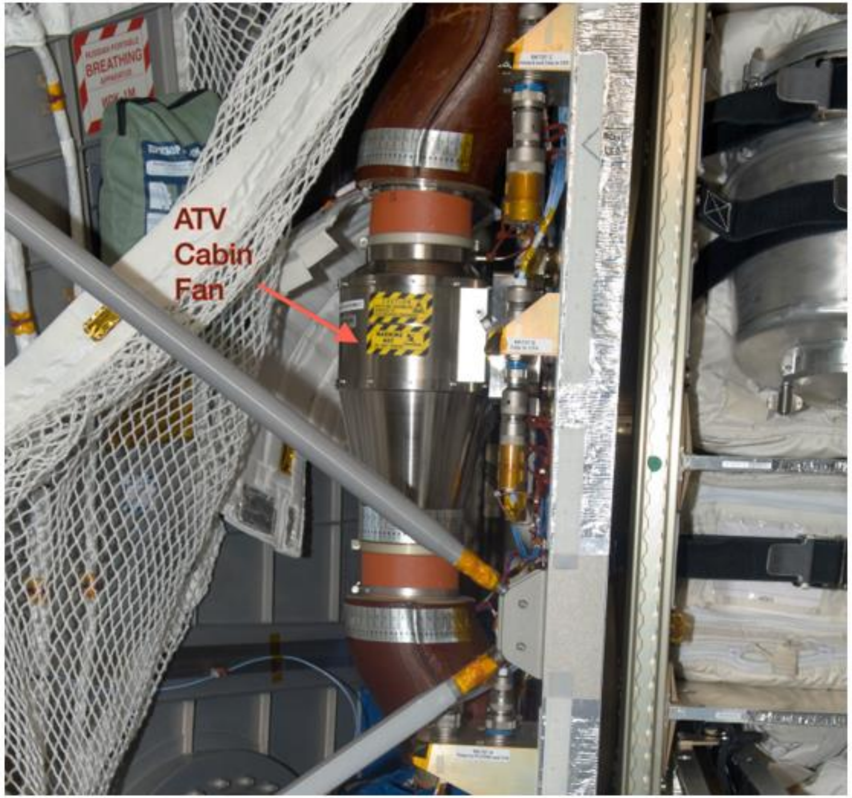
Image credit: Kriegel, Hans-Peter, Peer Kröger, Erich Schubert, and Arthur Zimek. "LoOP: local outlier probabilities." In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1649-1652. ACM, 2009.

Novelty Detection – Examples



XMM TANK1 EXT Eng

Novelty Detection - Examples



ATV Cabin Fan

The Novelty Detection system found novel behaviour, without any a-priori knowledge, beginning at the same period as the engineering experts had stated that the cabin fan was starting to show suspect behaviour, several days in advance of the failure. The Novelty Detection system managed to detect the fan failure as soon as it failed (as verified by the flight control team in Toulouse and the engineering experts from industry). We have seen the smoke detectors and the fan are tightly linked. So, this early warning was really telling that an anomaly in the fan was about to occur!

Credit: Marcus Deus Da Silva

Summary

Unsupervised Learning

- Clustering
- Principal Components Analysis
- Auto-Encoders

Semi-supervised Learning

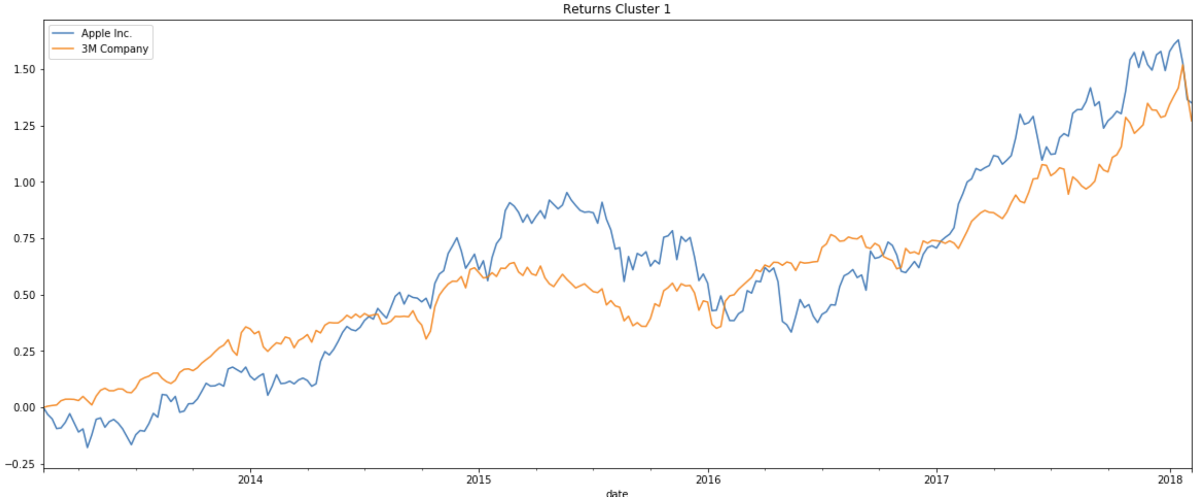
- DrMUST
- Novelty Detection

Materials: Slides, Code, Videos

Available on the Data Analytics ESA connect community

url: https://connect.esa.int/communities/community/data-analytics

Hands on: group Dow Jones stocks according to their behavior



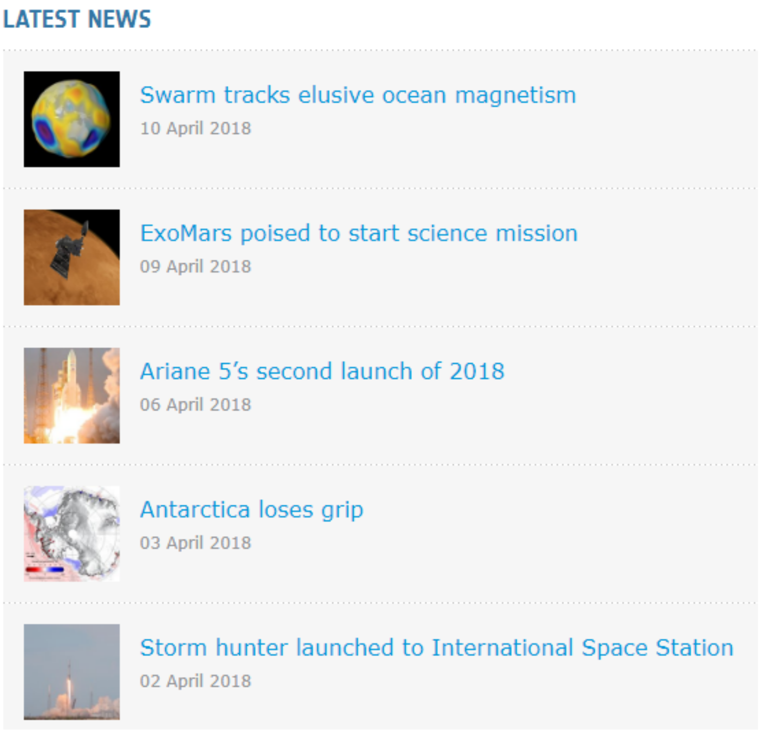
What is next?

April 26th 16:00 – Press Room

Session 6: Text Mining

- Text representation
- Topic Extraction
- Machine Learning with Text

- Hands on: predict the number of views of ESA News articles



http://www.esa.int/Our_Activities/Space_News

Resources

Watch the video of this lecture

https://dlmultimedia.esa.int/download/public/videos/2048/04/011/4804_011_AR_EN.mp4

Watch the practical exercise video

https://dlmultimedia.esa.int/download/public/videos/2048/04/010/4804_010_AR_EN.mp4

Get presentation and additional resources on

https://github.com/jmartinezheras/2018-MachineLearning-Lectures-ESA

Thank you

Data Analytics Team for Operations (DATO)

Jose Martinez Heras

LinkedIn: https://www.linkedin.com/in/josemartinezheras/