Open Source is HOT

3 au 6 juillet 2017

SophiaConf

Le cycle azuréen de conférences Open Source

by TelecomValley
Animateur Azuréen du Numérique

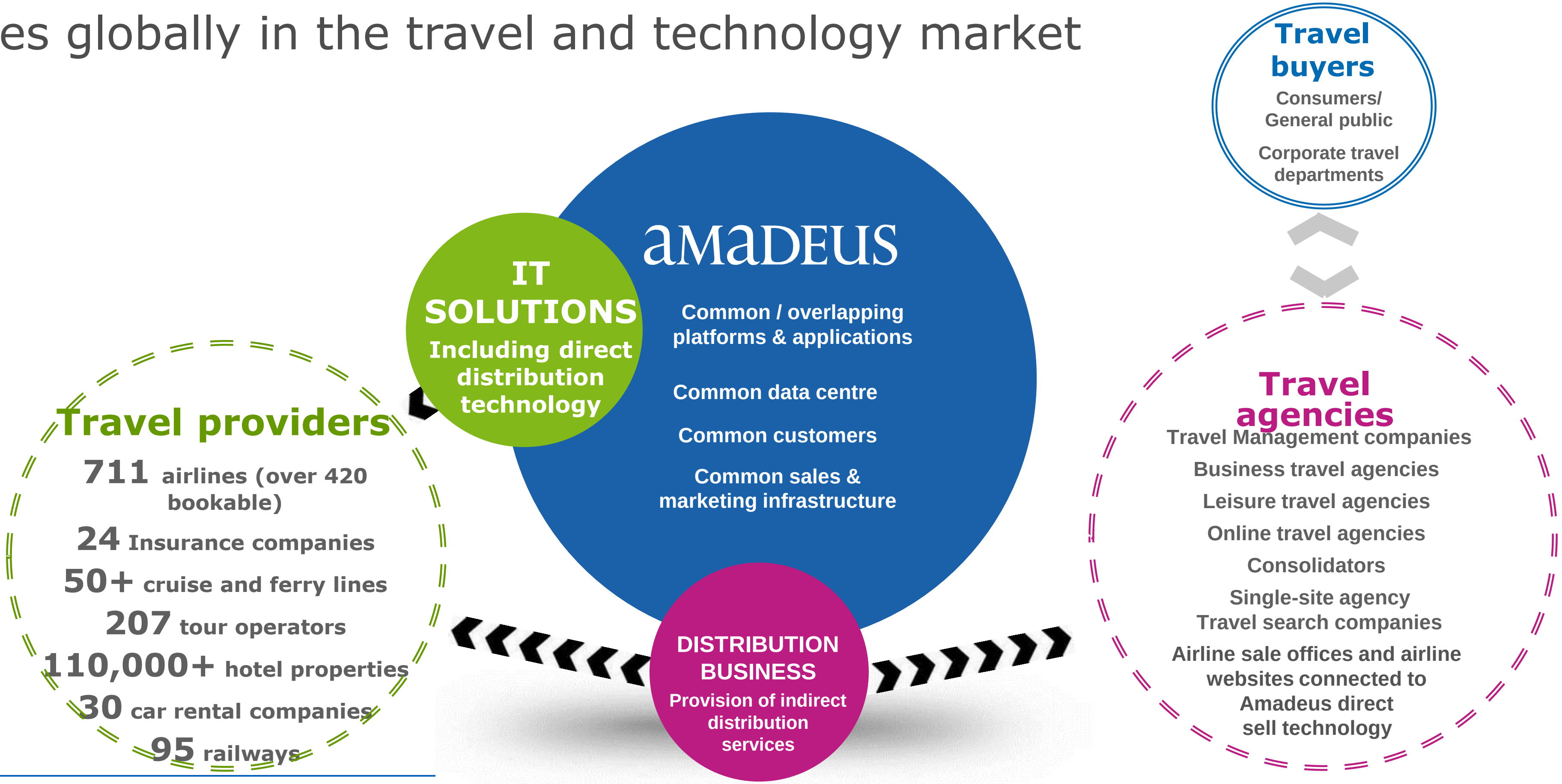
AMADEUS

# Anomaly Detection in airlines schedules

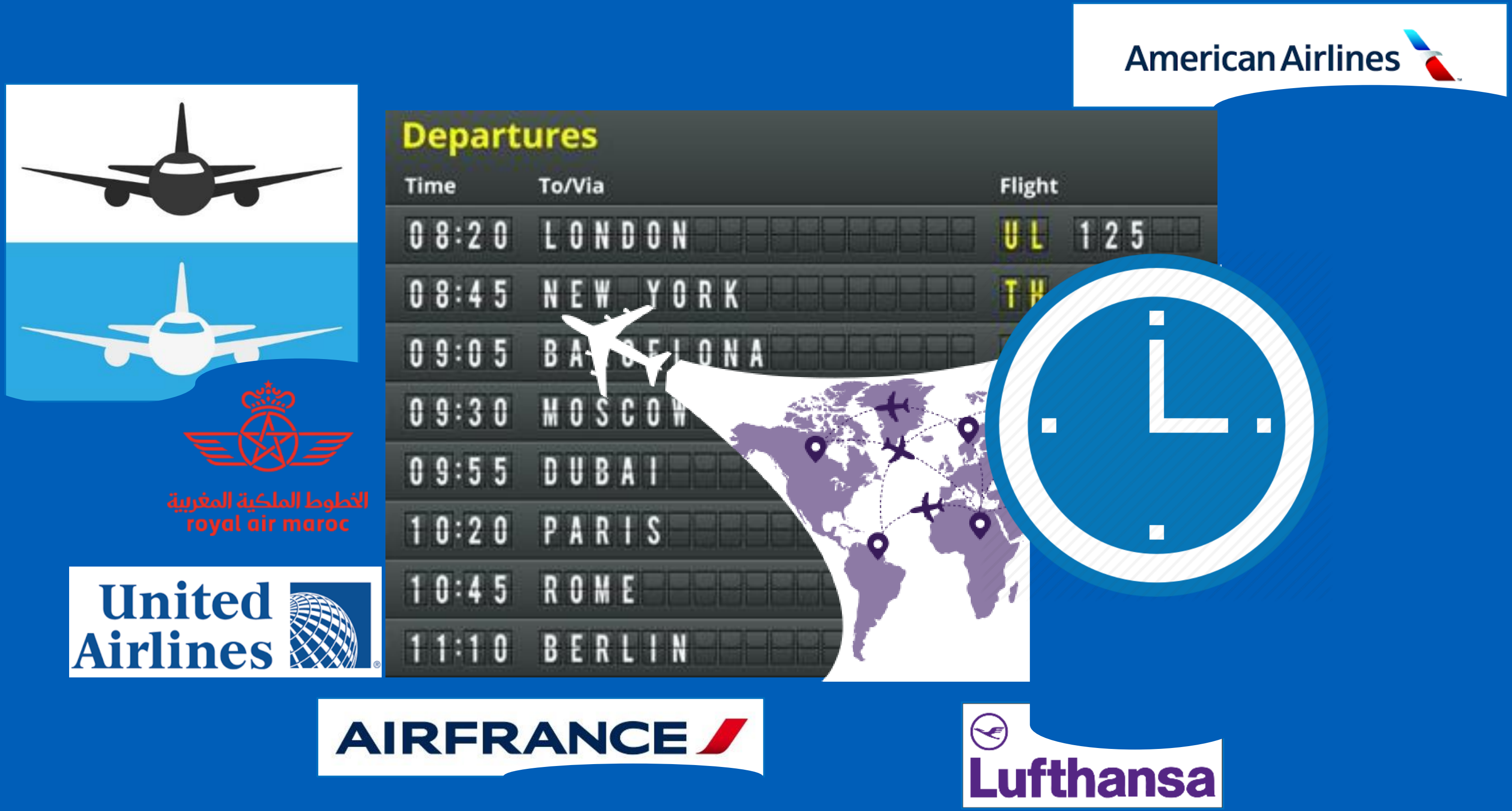

**Asmaa Fillatre**
**Data Scientist, Amadeus**

# AMADEUS PRESENTATION

1. IT company that develops business solutions for the travel and tourism industry
2. Operates globally in the travel and technology market



# 1 Airline Schedules



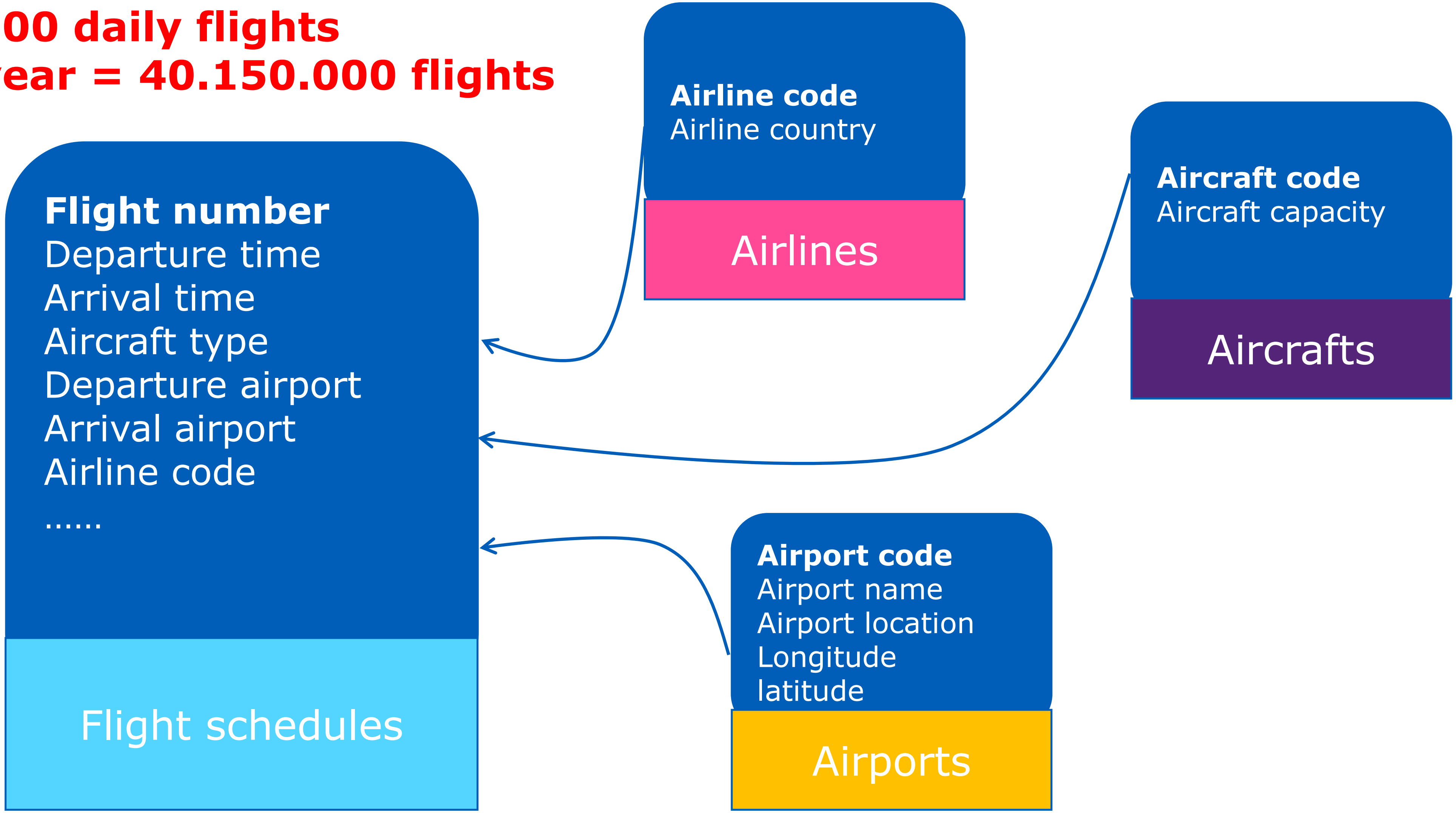
# Airline schedules




amadeus

# Airline schedules data

- **110.000 daily flights**
- **One year = 40.150.000 flights**



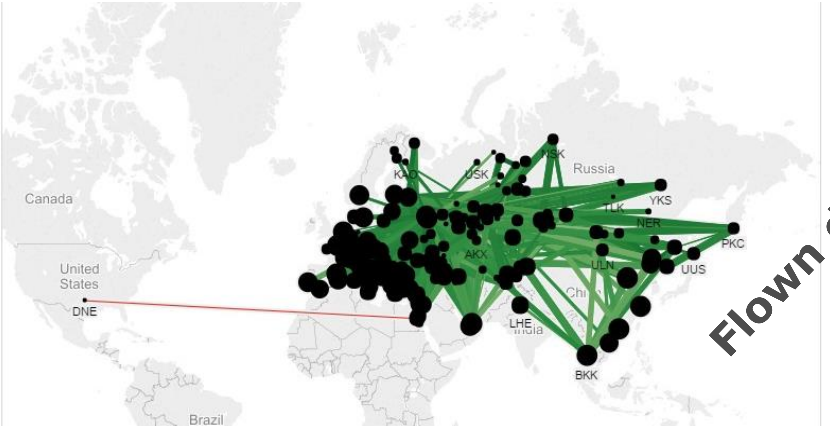
# Motivations

1. The airline schedules contain many errors.
2. It is important to identify outliers prior to modelling and analysis.
3. Detect anomalies automatically
4. Overcome the issue of non prior knowledge (no ground truth)

# Anomalies examples (1)

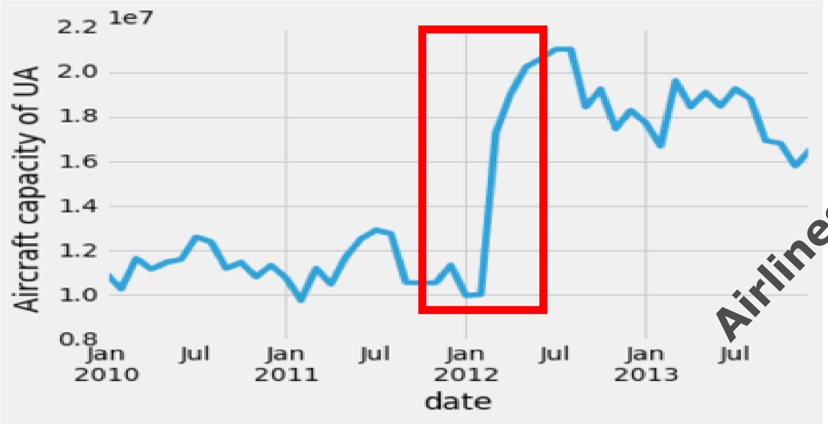
- Airlines use wrong IATA airport codes
- Airlines missing
- Merger between two companies
- Flown distance much higher than aircraft average
- Elapsed time/distance not appropriate
- New routes traffic
- Sports event (OG, *FIFA World Cup, etc)*
- …

Not exhaustive list




Flown distance much higher than the aircraft average

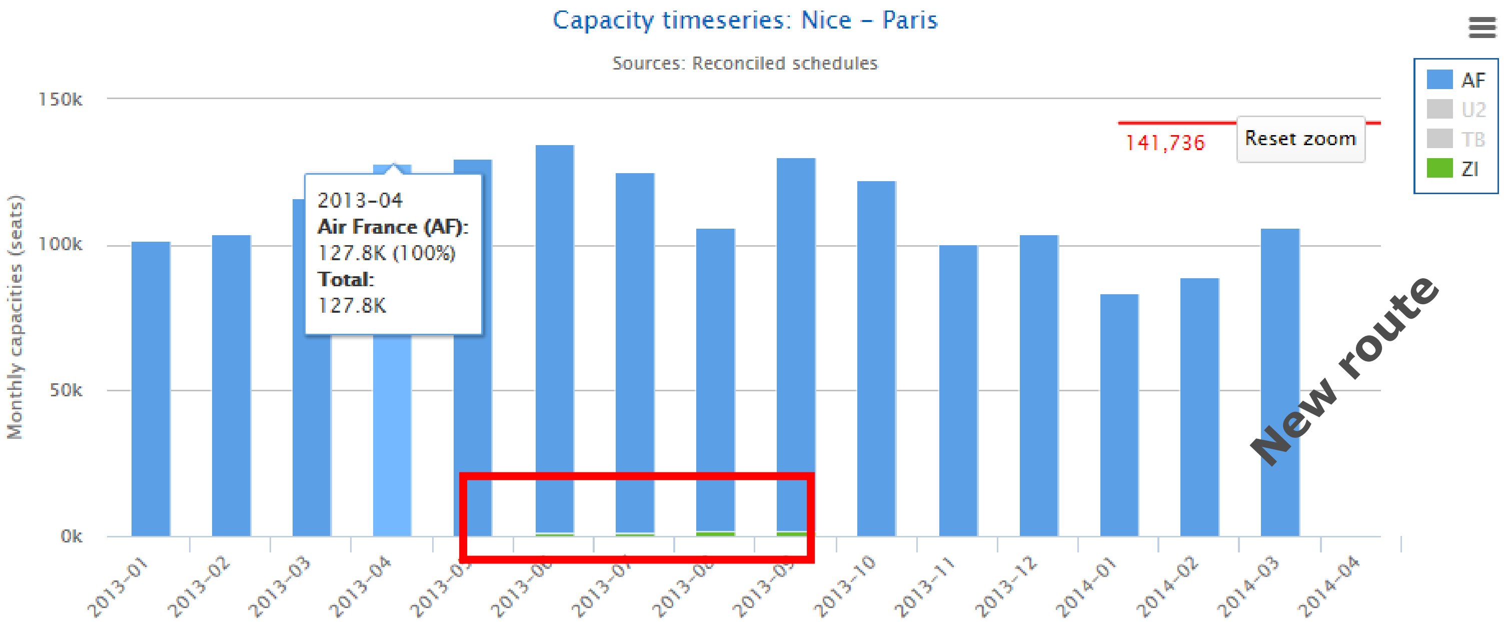
Flown distance error

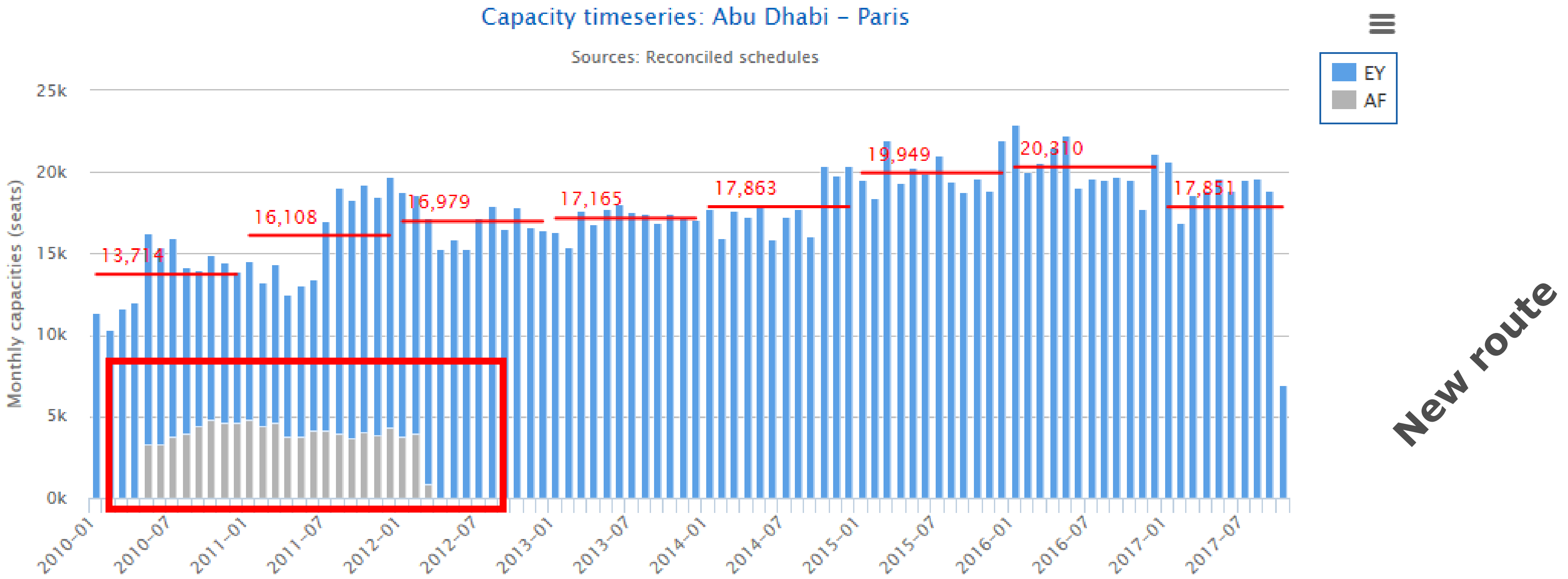



Sudden grow in monthly Aircraft capacity for United Airlines

Airlines merged

# Anomalies examples (2)



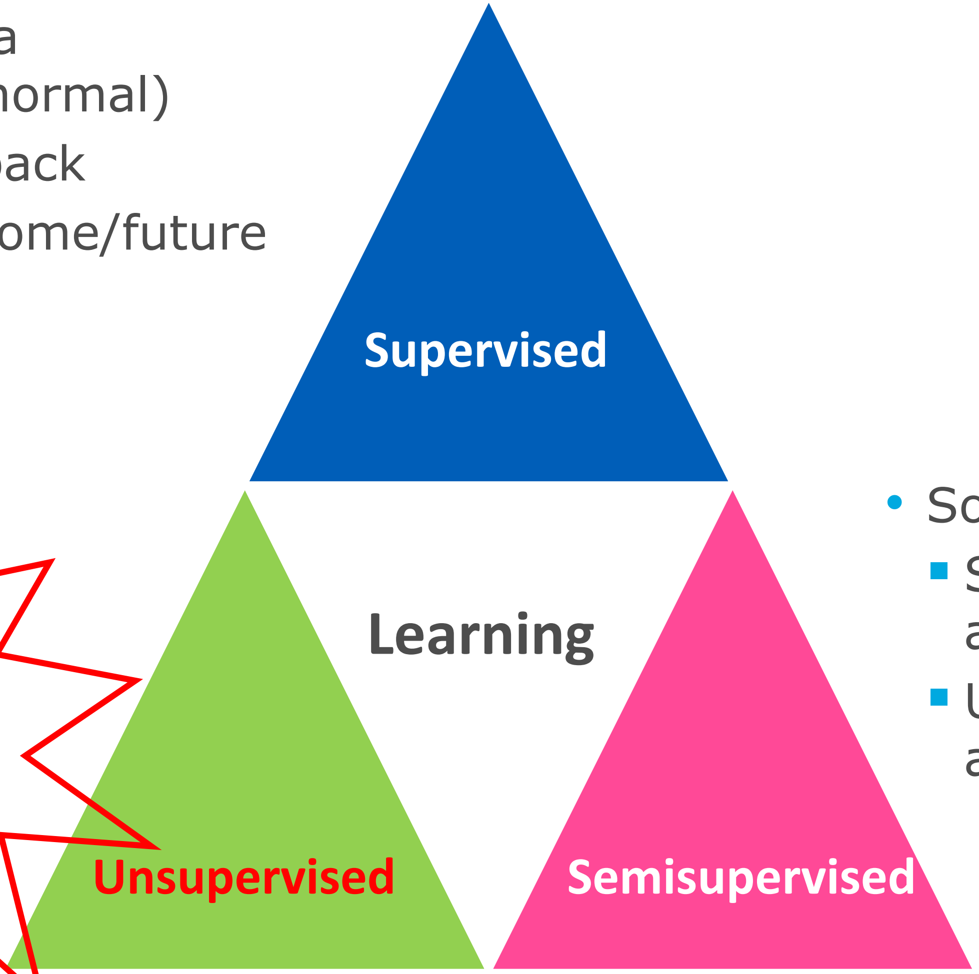



# 4 Unsupervised Anomaly detection

Goal: Process unlabelled data and detect anomalies

# Machine learning

- Labeled data (normal/abnormal)
- Direct feedback
- Predict outcome/future

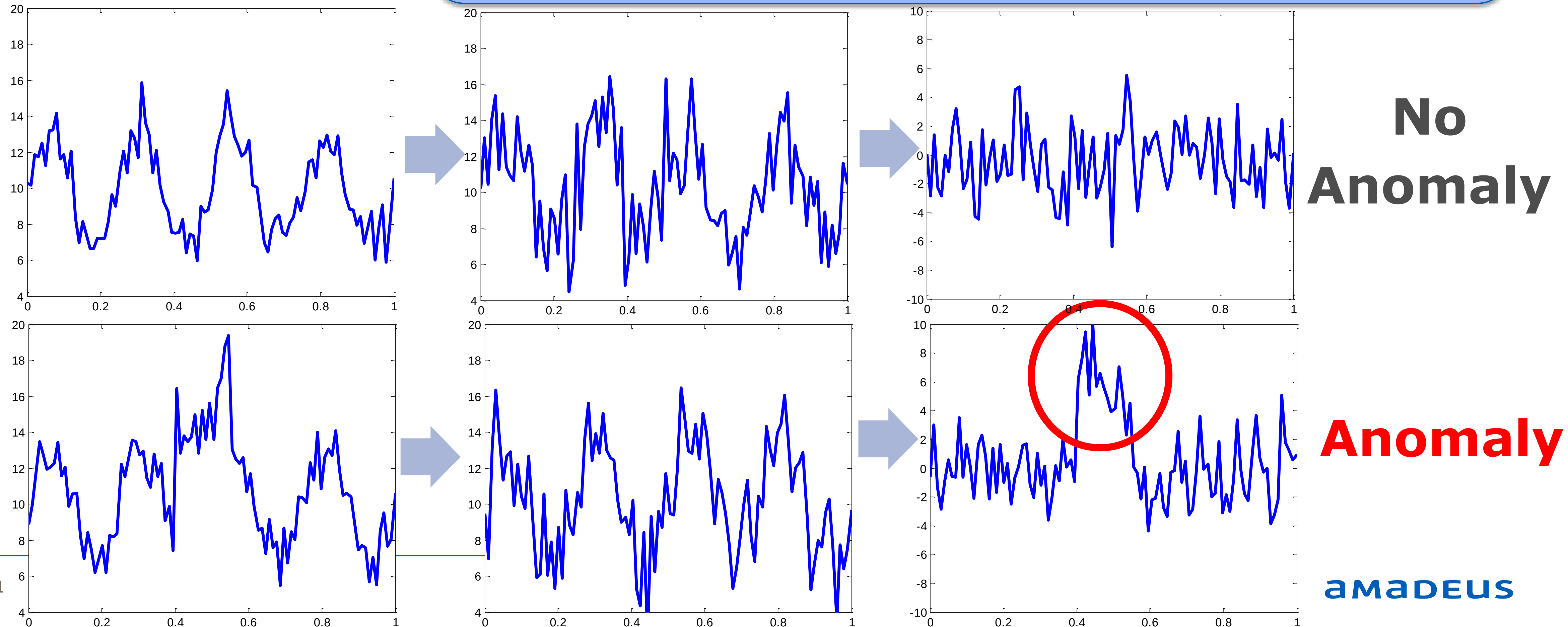



- No labels
- No feedback
- Find hidden structure

- Some labelled data :
  - Supervised learning + additional unlabelled data
  - Unsupervised learning + additional labelled data

# Residuals-based anomaly detection in three steps

Input data → Low-rank approximation → Residuals generation → Anomaly detection






amadeus

# Residual and Anomaly Detection

- Residual

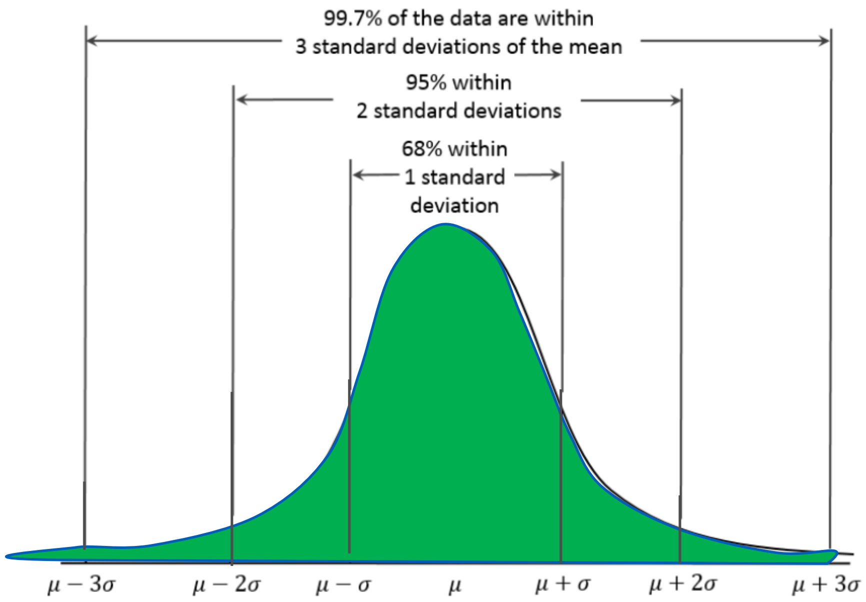
$$R_i = \text{Input} - \text{Reconstrucion}$$

- Residual normalization

$$Z_i = \frac{(R_i - \mu)}{\sigma}$$

- Residual thresholding

$$|Z_i| > 3$$




Three sigma rule


Any data sample outside the interval $[\mu - 3\sigma, \mu + 3\sigma]$ is considered to be potential **anomaly**



amadeus

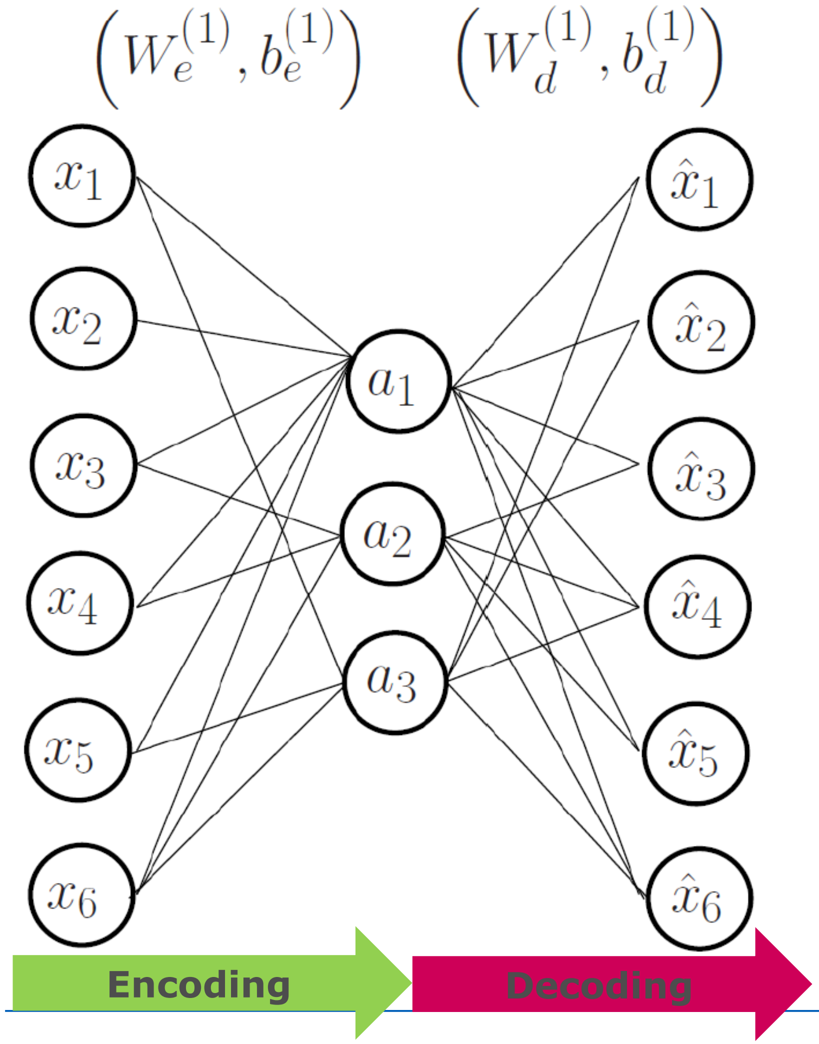
6

# Deep learning: Stacked Autoencoder

Goal: Learn the internal structure and features of the data itself

# Autoencoder

One hidden layer



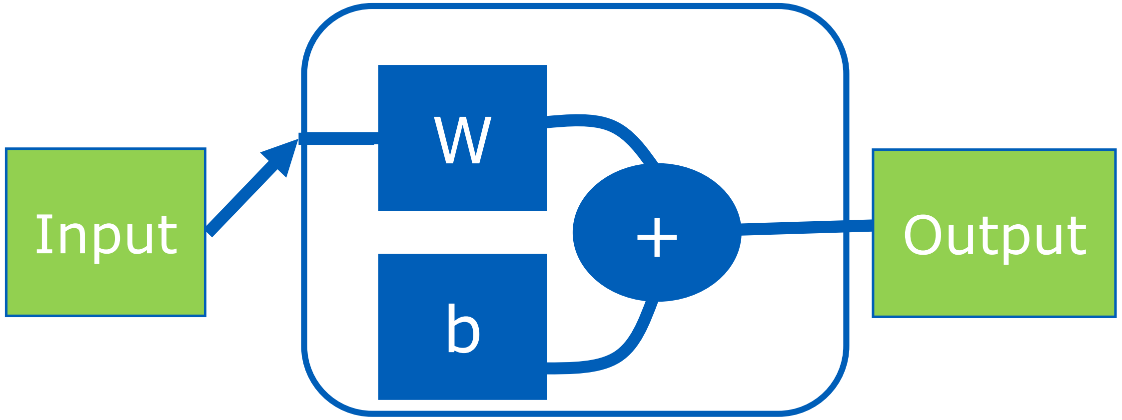

- Minimize $\|X - \hat{X}\|$ w.r.t. all $W_e^{(\ell)}, W_d^{(\ell)}$ and $b_e^{(\ell)}$, $b_d^{(\ell)}$
- Trained with **Backpropagation**
- Self-supervised technique
- Learn a meaningful representation of the data in some other dimensionality

$$a_i = f(z_i) \quad \text{where} \quad f(z) = \frac{1}{1+e^{-z}}, \forall z \in \mathbb{R}$$

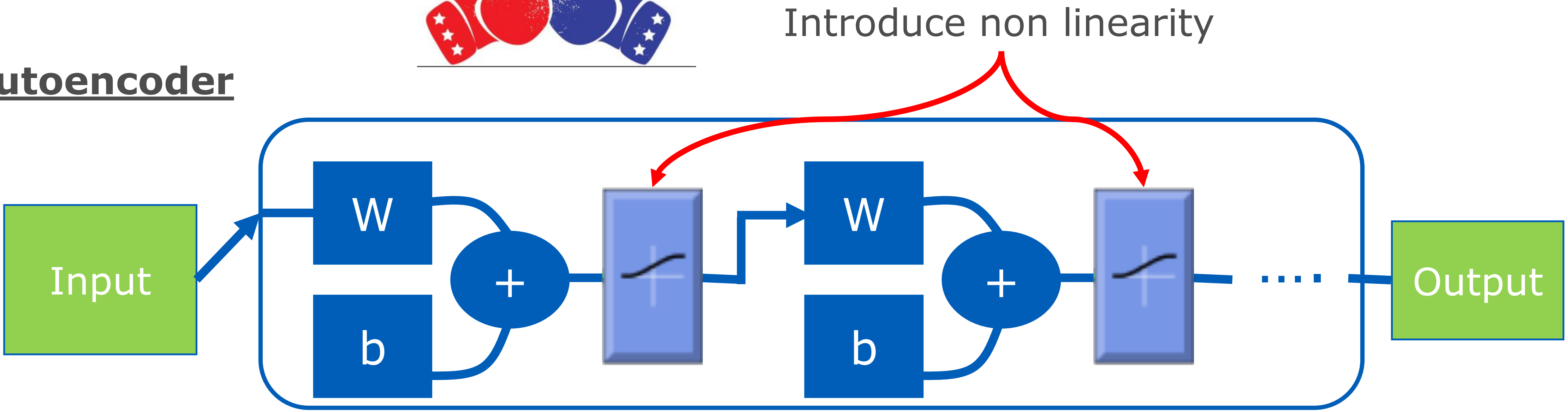
$$\text{and} \quad z_i = b_e^{(1)}[i] + \sum_{j=1}^{m} w_e^{(1)}[i,j]x_j$$

$$\hat{x}_i = f\left(b_d^{(1)}[i] + \sum_{j=1}^{k} w_d^{(1)}[i,j]a_j\right)$$

## PCA




## Autoencoder



# Deep Autoencoder or stacked autoencoder

## Cost function

$$J_{\text{sparse}}(W, B) = \frac{1}{2m}\sum_{i=1}^{m} ||\hat{X}_i - X_i||^2$$

Average sum of squared error

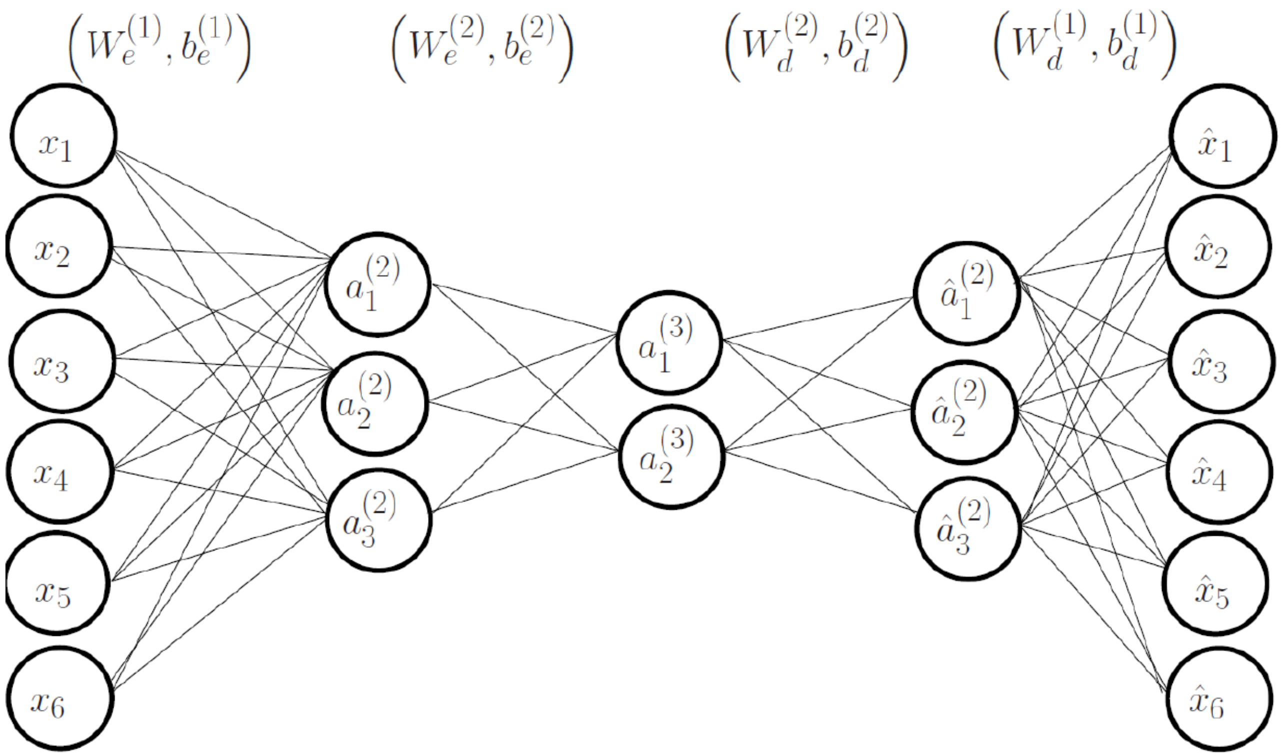
Regularization

$$+\frac{\lambda}{2}\sum_{\ell=1}^{L-1}\left(||W_e^{(\ell)}||_F^2 + ||W_d^{(\ell)}||_F^2\right)$$

Weight decay

$$+\beta\sum_{\ell=2}^{L}\sum_{j=1}^{s_\ell} KL(\rho||\hat{\rho}_j^{\ell}),$$

Sparsity Penality

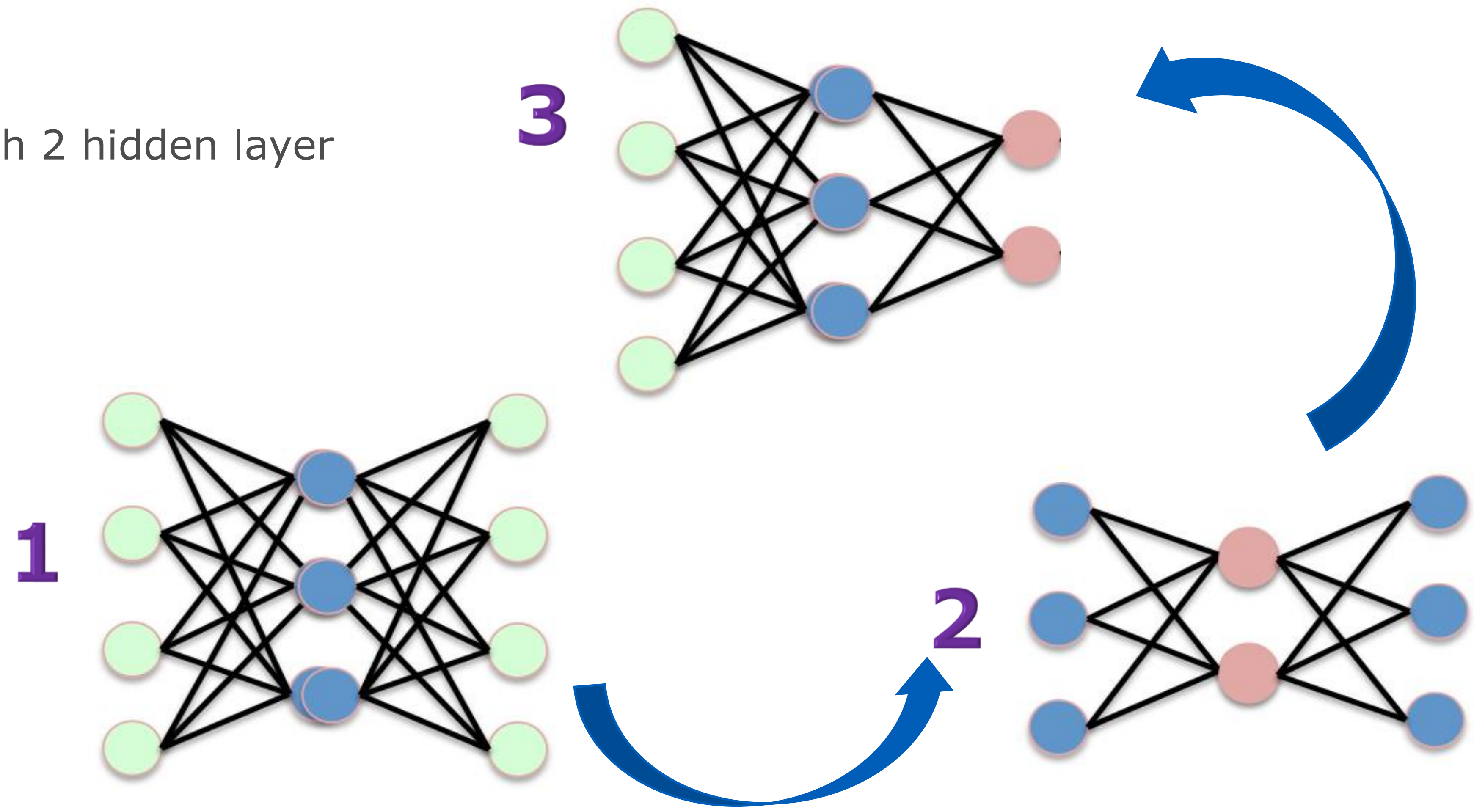



- Constraints on the activation $\hat{\rho}$ which should be close to $\rho$
- Regularization by $\lambda$

# Stacked Autoencoder training

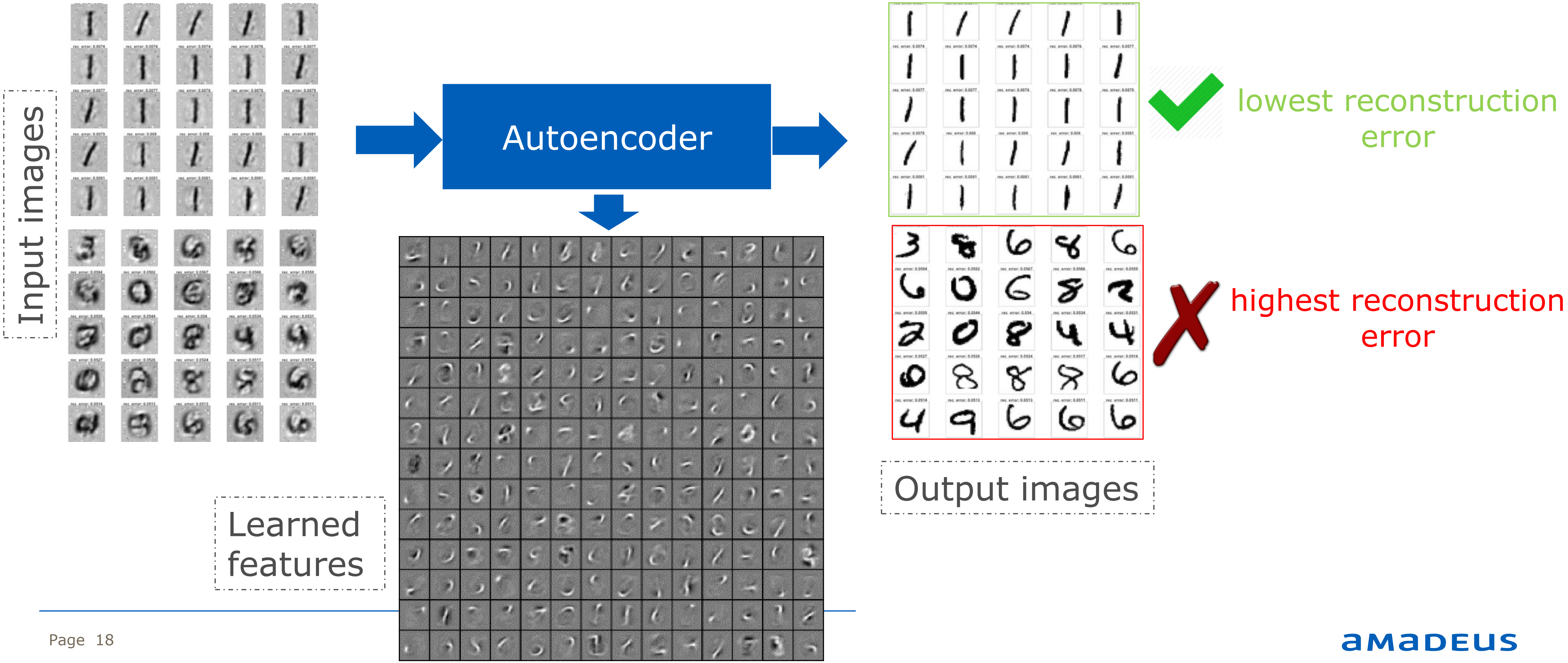
## Training one hidden layer at a time

Example with 2 hidden layer



# Hello world of deep learning

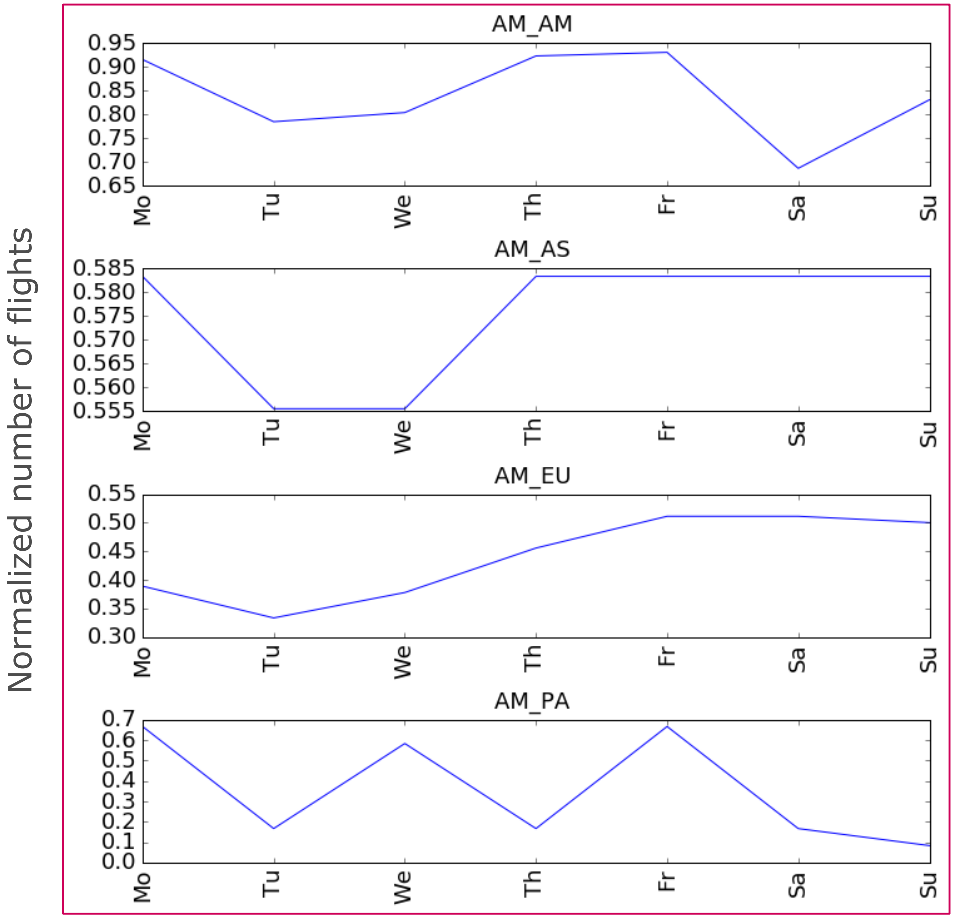
## Anomaly Detection on MNIST



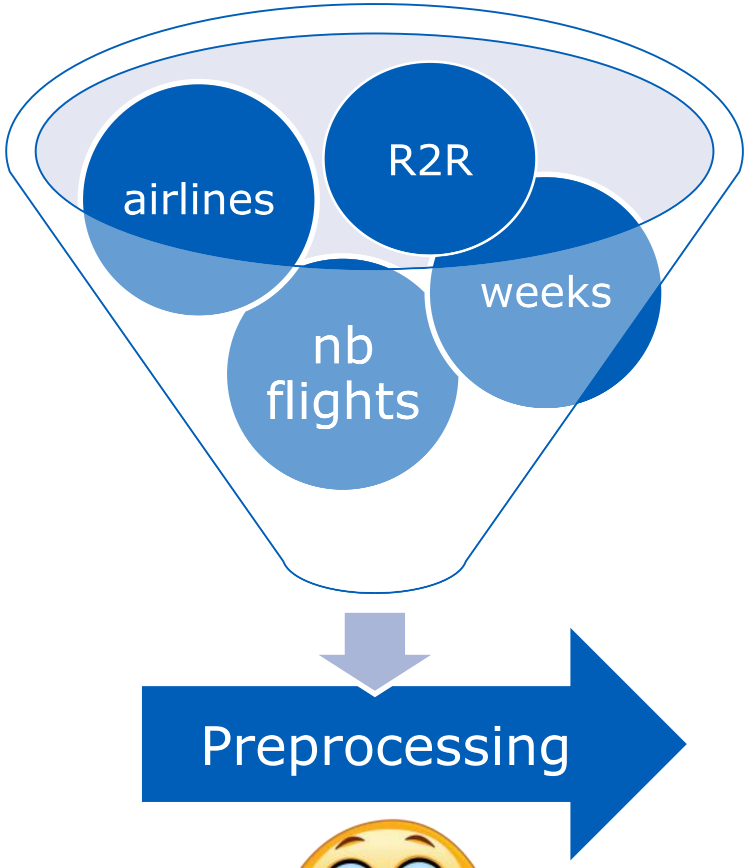
7

# Autoencoder based Anomaly detection for airlines schedules

# Raw data: multivariate time series

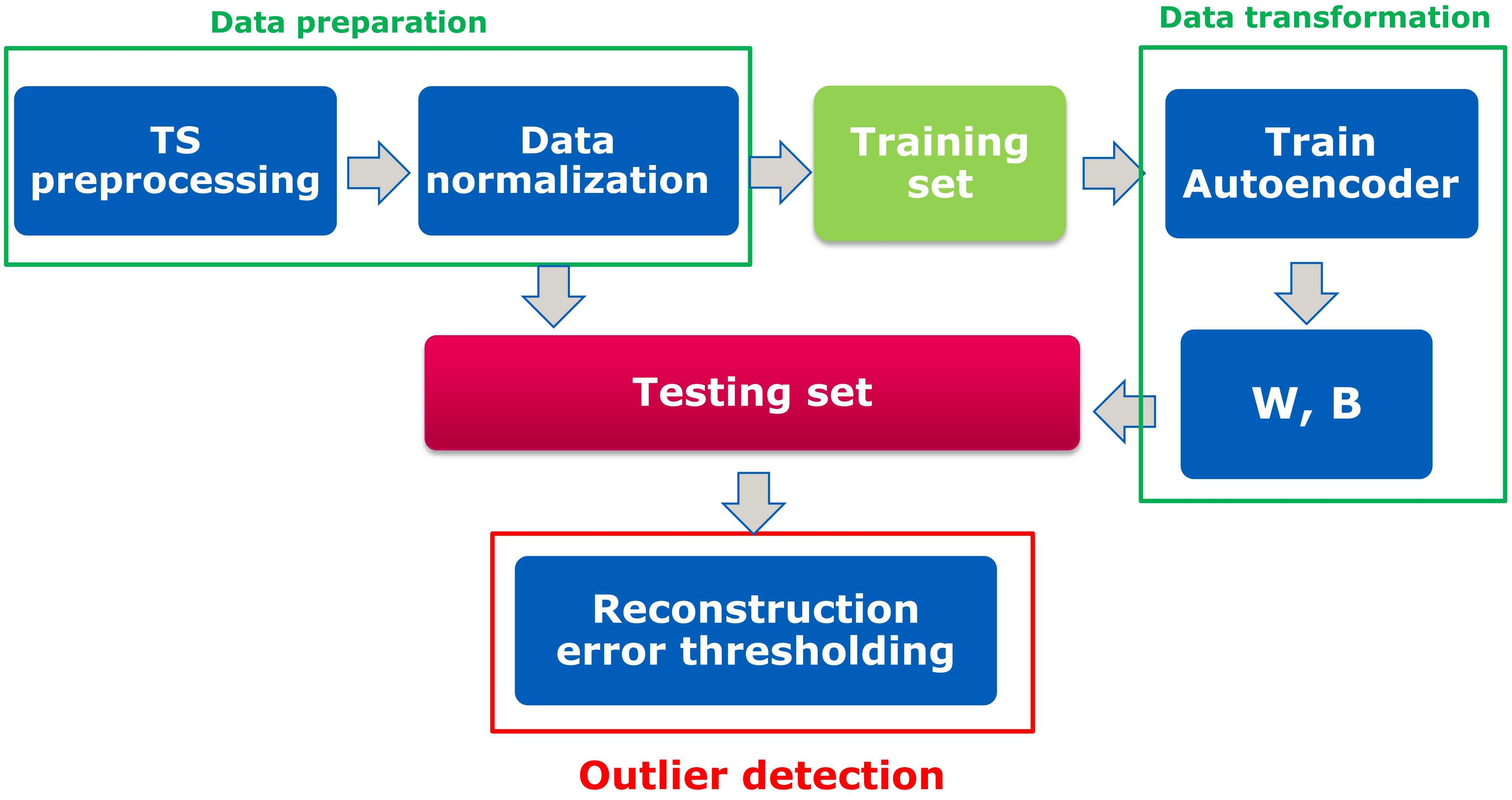


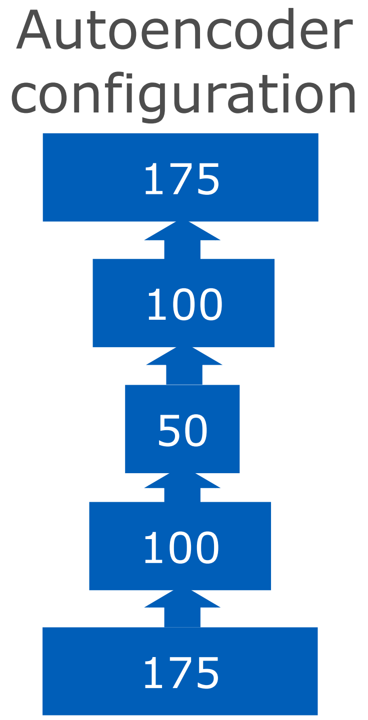

Some region to region time series 2012 week 10




- ✓ For more natural representations of data
- ✓ The Autoencoder can learn some patterns

# Autoencoder for time series – Anomaly detection



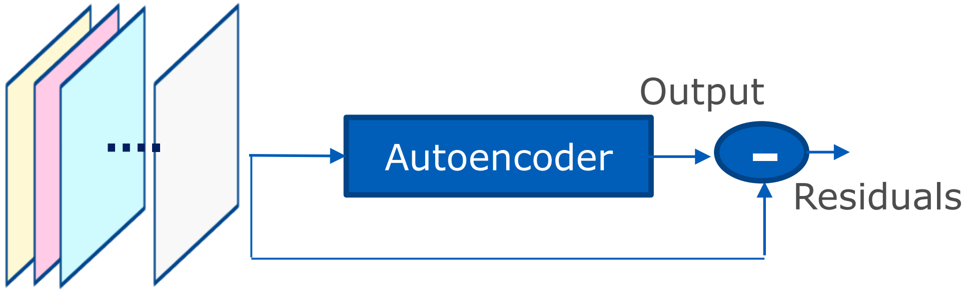

$+ \quad \beta, \lambda \; and \; \rho$

# 8 United Airlines (UA) schedules data processing

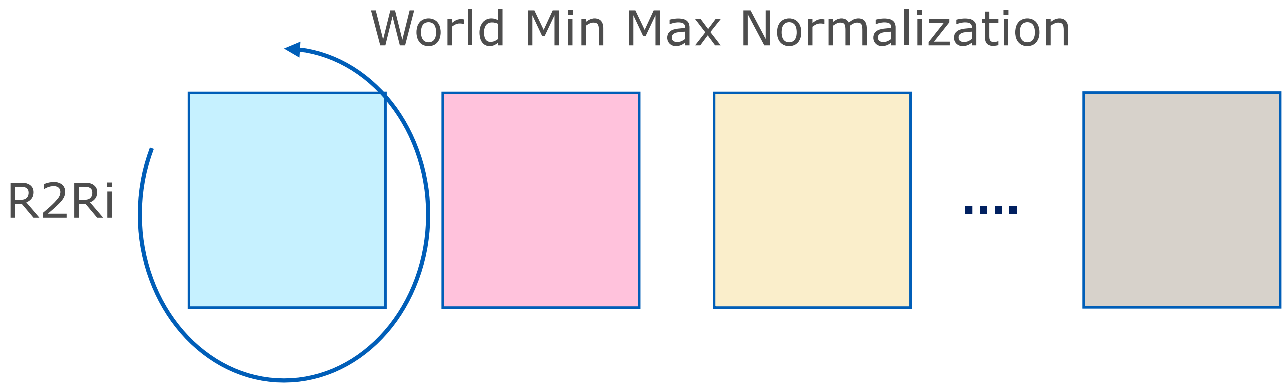
Goal: highlight how does the Autoencoder perform in practice

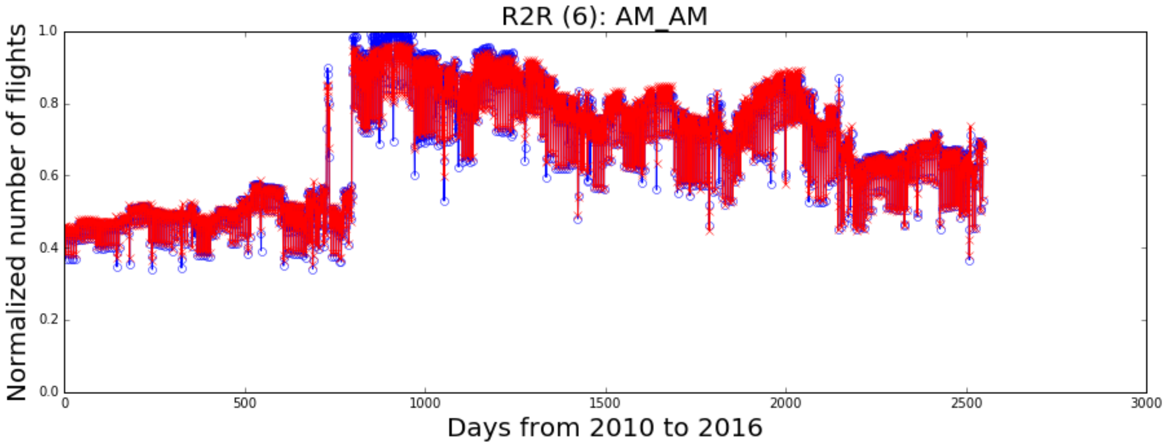
# UA anomaly detection (1)

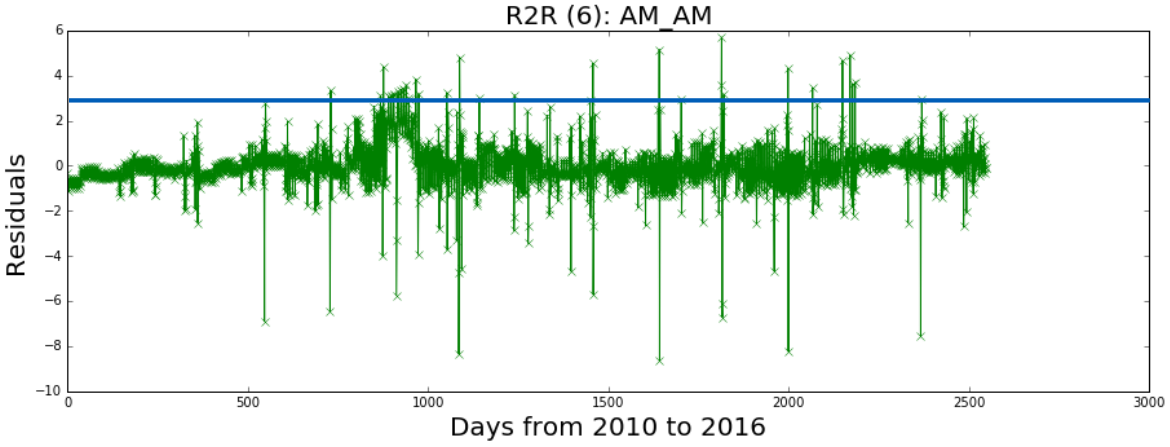
## World normalization of Input data




2010 to 2016
Input from UA



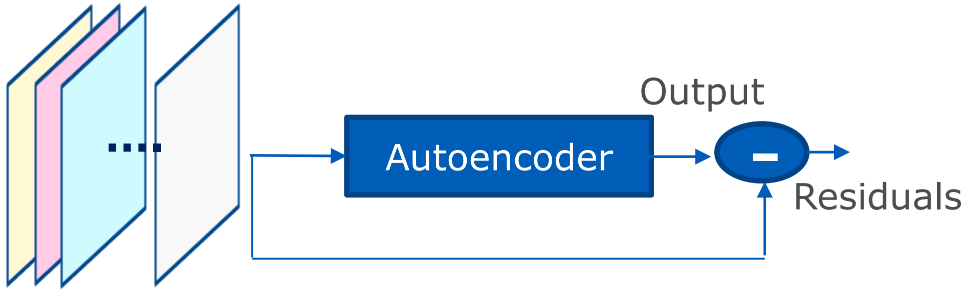






amadeus

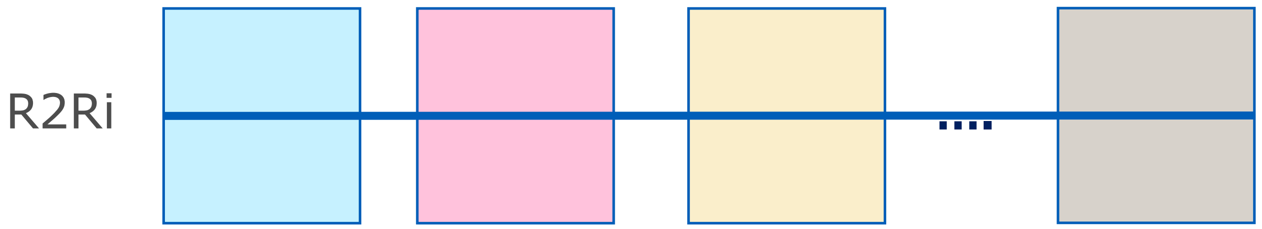
# UA anomaly detection (2)

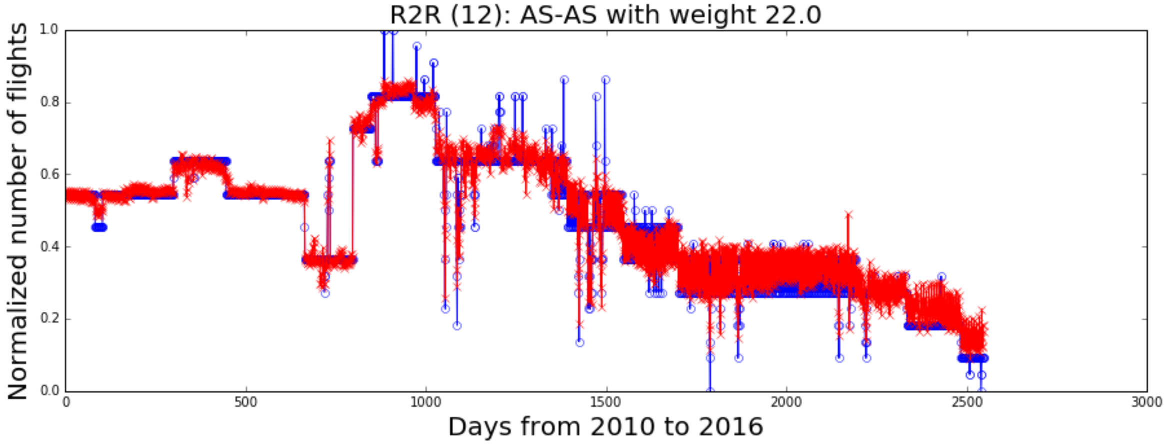
## Regional normalization of Input data

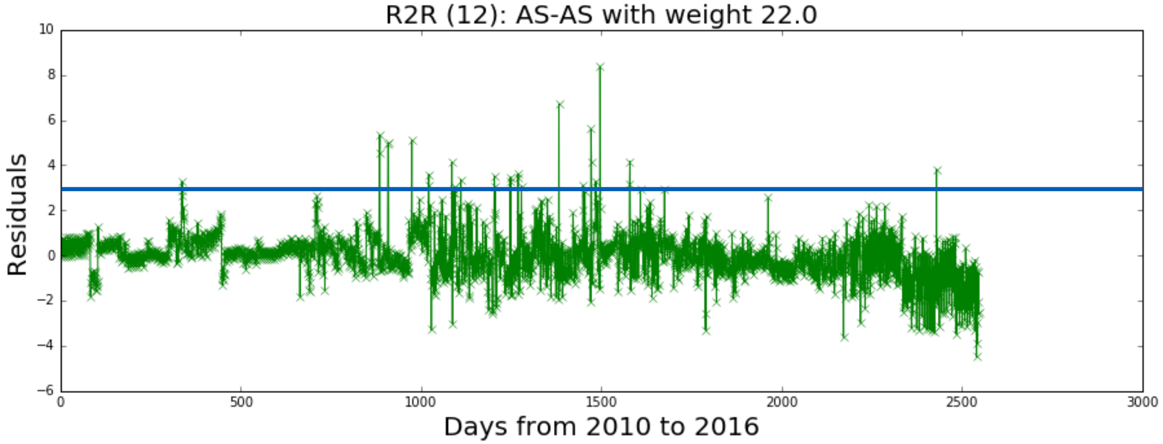



2010 to 2016
Input from UA

Min Max Normalization per region



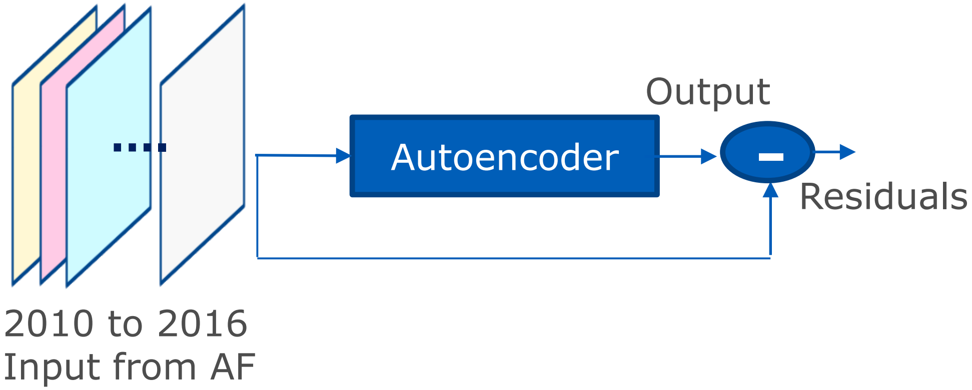
8

# Air France (AF)

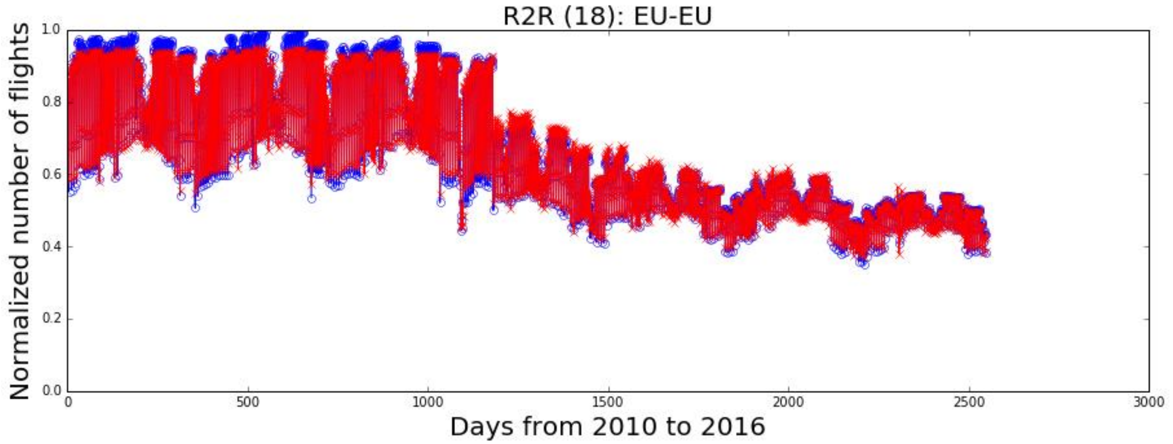
Goal: highlight how does the Autoencoder perform in practice

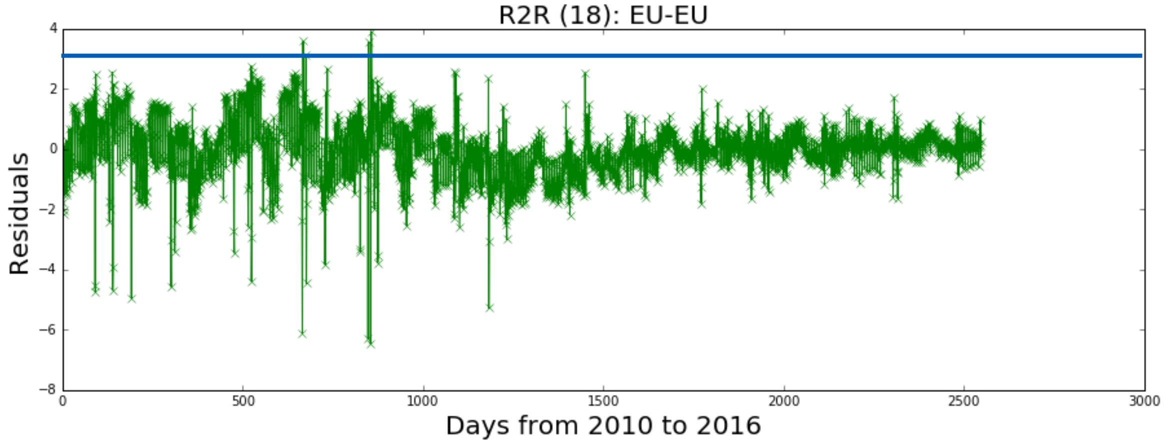
# AF anomaly detection (1)

World normalization of Input data




World Min Max Normalization

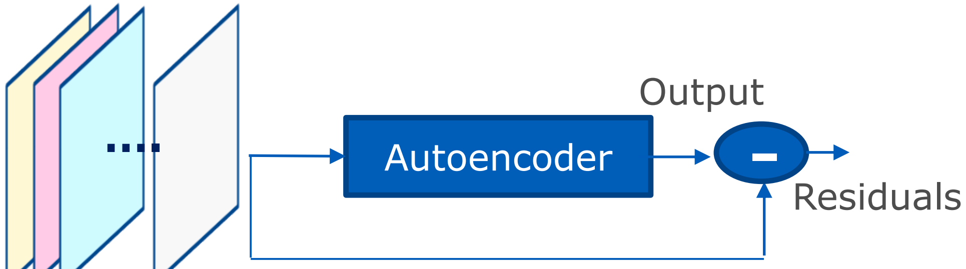






amadeus

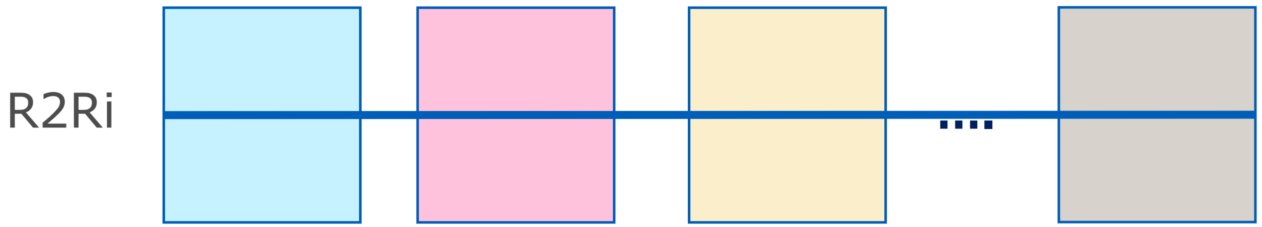
# AF anomaly detection (1)

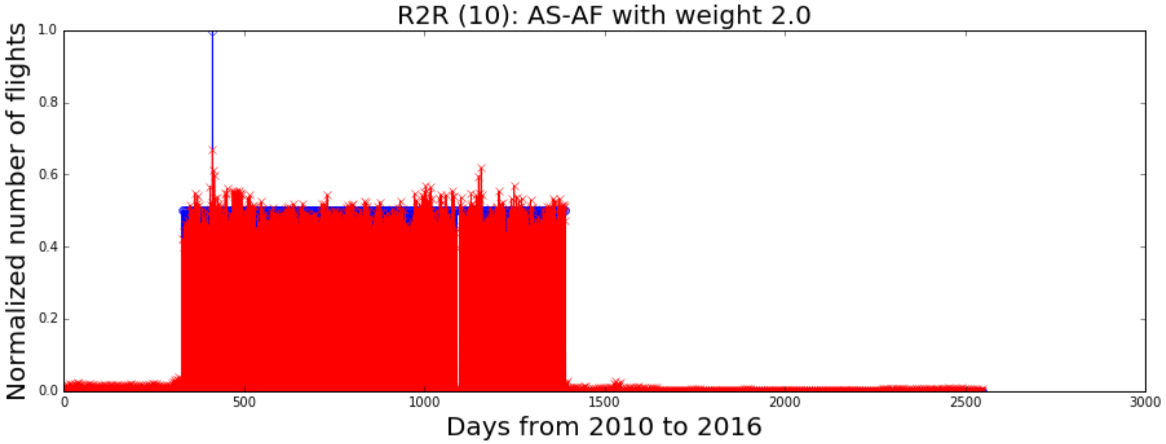
## Regional normalization of Input data

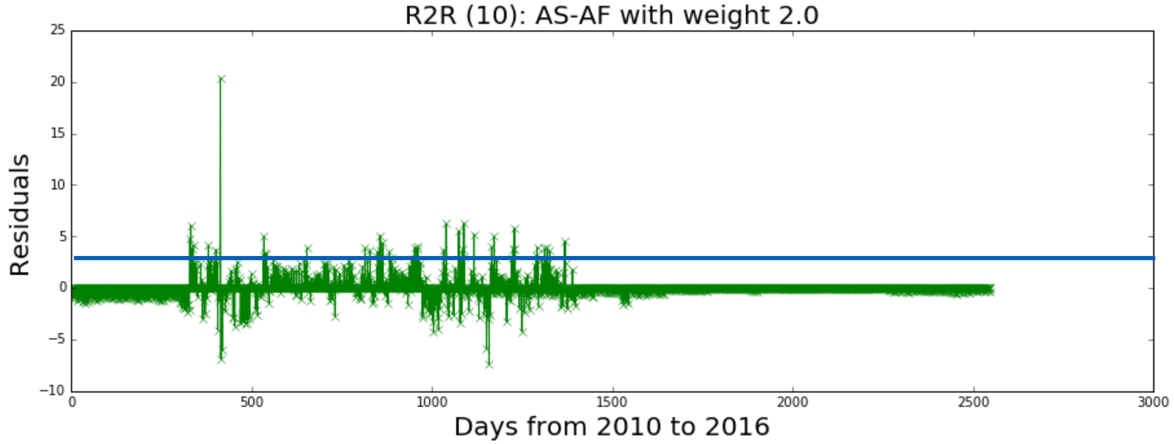



2010 to 2016
Input from AF

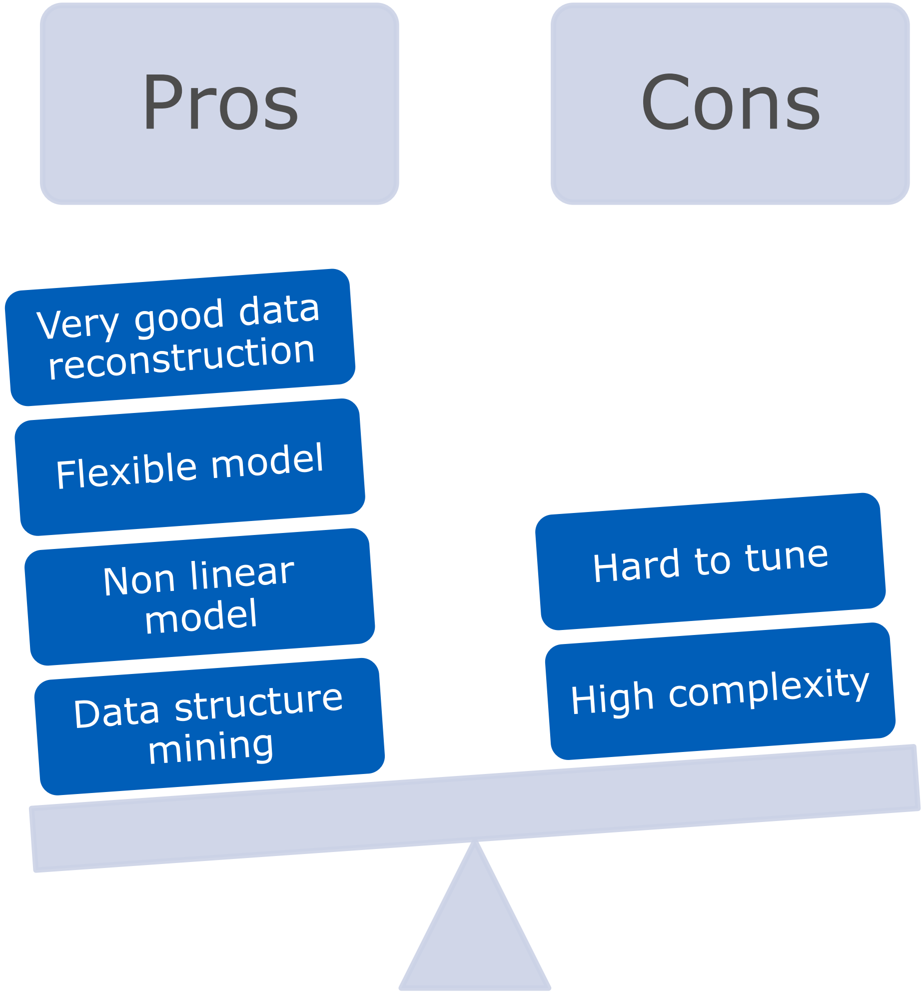
Min Max Normalization per region



# Autoencoder pros and cons



# Conclusion

- Unsupervised machine learning (no ground truth)
  - Well adapted to the absence of labels
  - Hard to interpret: the review process of outliers relies on domain experts
- Deep learning/feature engineering

Thanks for your attention