LHC instability detection and clustering

Speaker: Loïc COYLE

Acknowledgements:
M. Schenk, T. Pieloni, X. Buffat

EPFL

CERN

ADTObsBox

“The ADTObsBox is a very powerful computer system that was designed to receive a copy of the digital bunch-by-bunch transverse position data stream, analyze it online or offline, make it available to users outside of the ADT system, or to store it.”

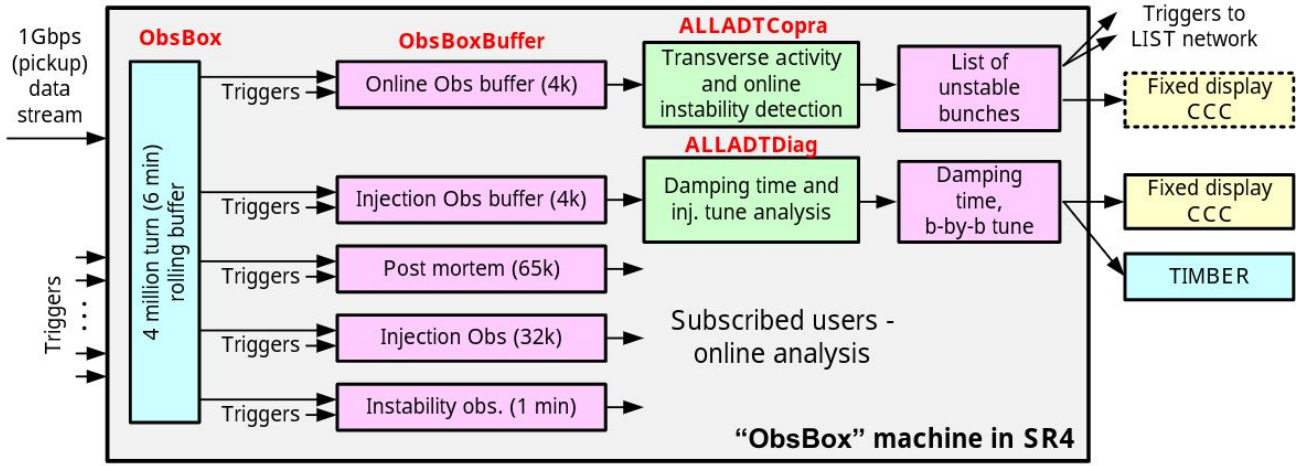
In short:

→ **rolling buffer & saves on trigger**

→ 65536 turns

→ **bunch by bunch**

→ transverse position data



Problem

The trigger is not very accurate :
Most of the data does not contain any instabilities.
→ instabilities make up less than 1% !

Large amount of data ~4 TB

Very little manually labeled data [1]
→ eliminates supervised learning methods

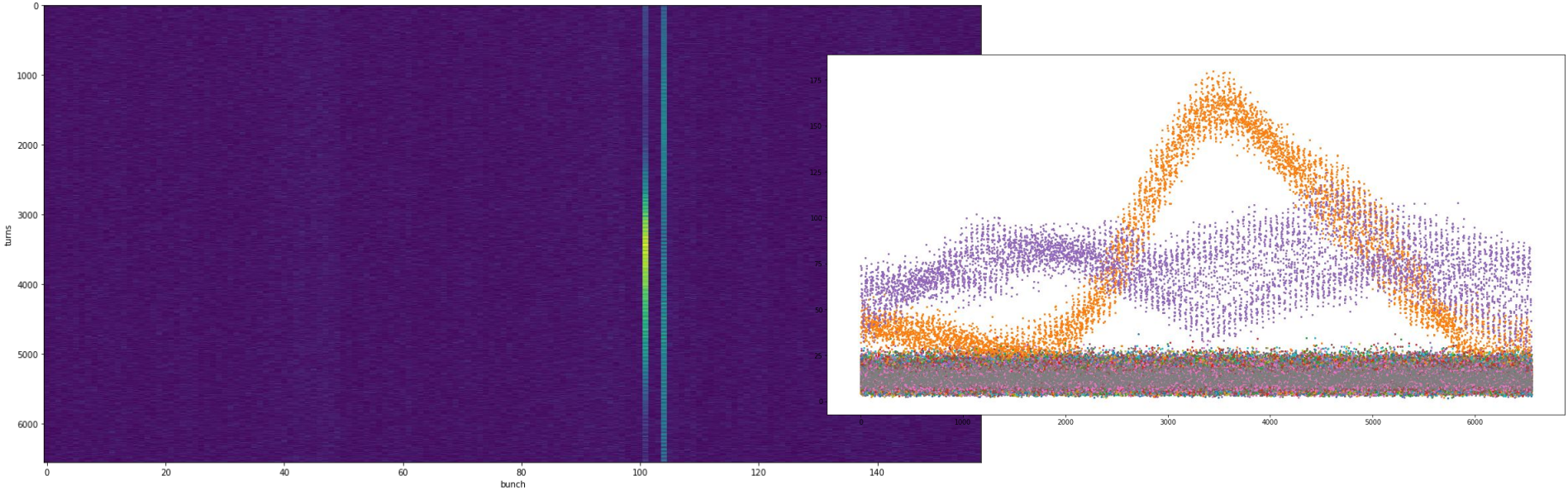
How to find/cluster instabilities at this scale ?

[1] http://lhcinstability.web.cern.ch/lhcinstability/csv-to-html-table/2018.html

ADTObsBox

Raw beam amplitude data at a turn by turn and bunch by bunch resolution. → multivariate time series

Example: 07169_Inst_B1V_Q7_20180914_08h53m08s → 2 unstable bunches



Key steps

- Instability detection → filtering
- Clustering on the filtered dataset

Key steps - Filtering

Light preprocessing of Data

Extract features from the data → tsfresh [1] :

- Maximum/Minimum
- Mean
- Standard Deviation
- FFT coeffs
- CWT coeffs
- CID coefficient [2] (complexity coefficient)
- ...

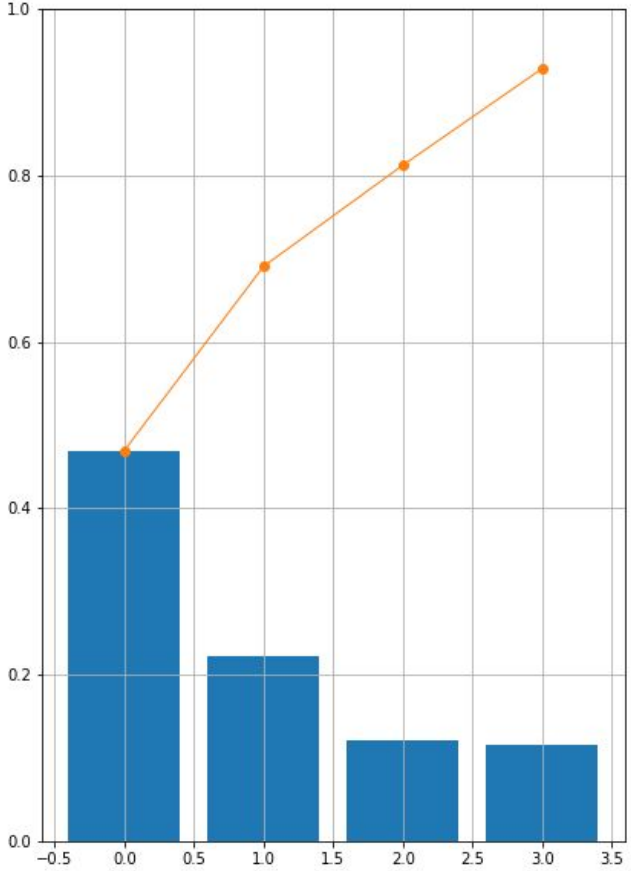
Dimensionality reduction : PCA on the extracted features

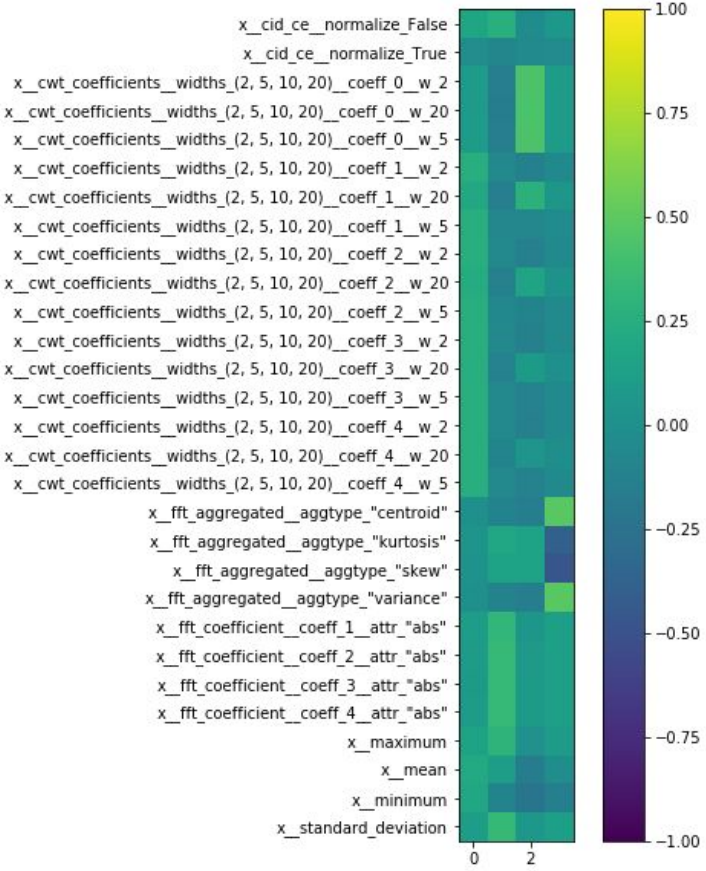
Find instabilities in PCA space

[1] https://github.com/blue-yonder/tsfresh
[2] Batista, Gustavo EAPA, et al (2014) Data Mining and Knowledge Discovery 28.3 (2014): 634-669.

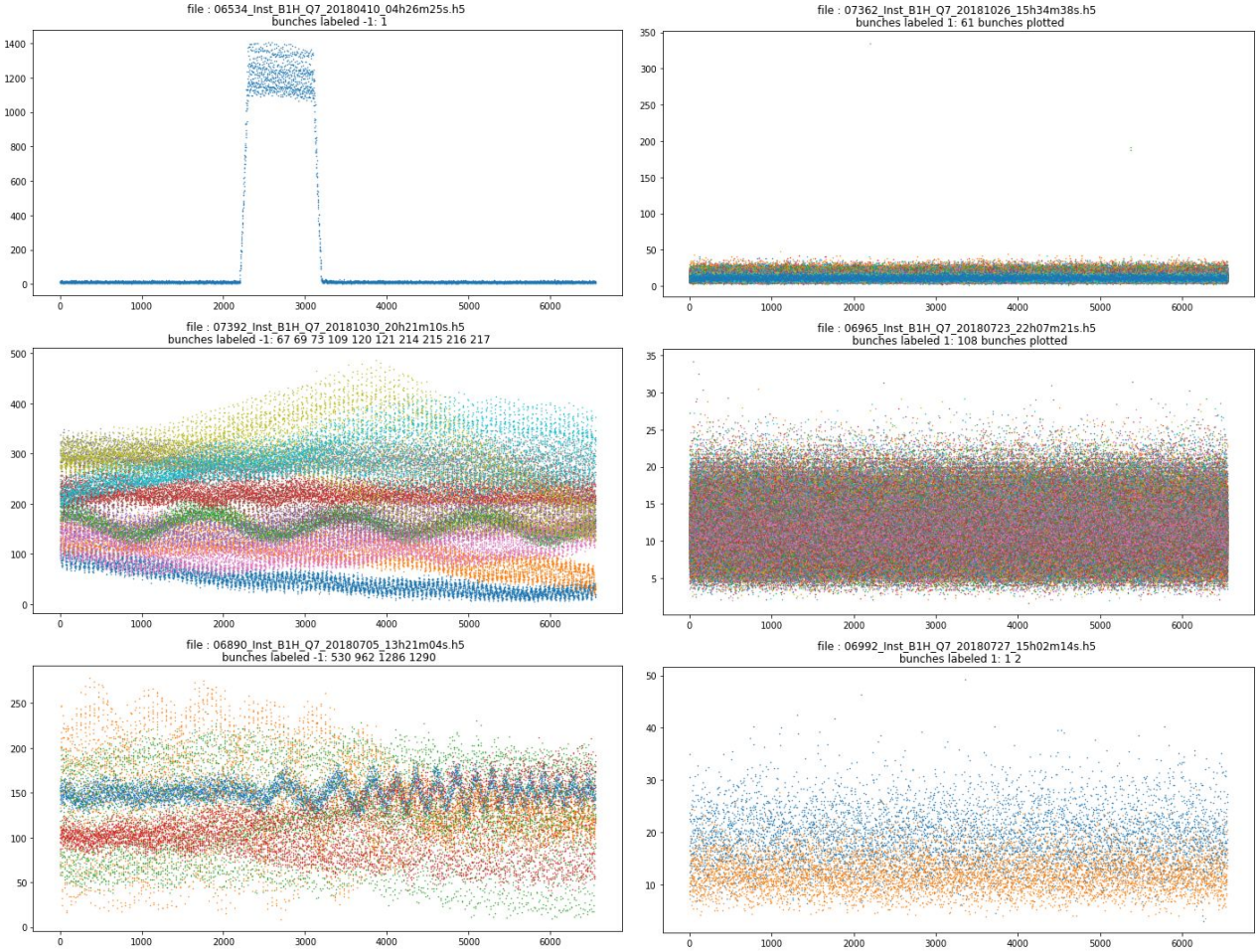
Principal Component Analysis

PCA vectors truncated to 4 components → ~93% variance explained





Classification results



Plots for a subset of the data, of the inlier/outlier bunches.

Left: outlier data
Right: inlier data

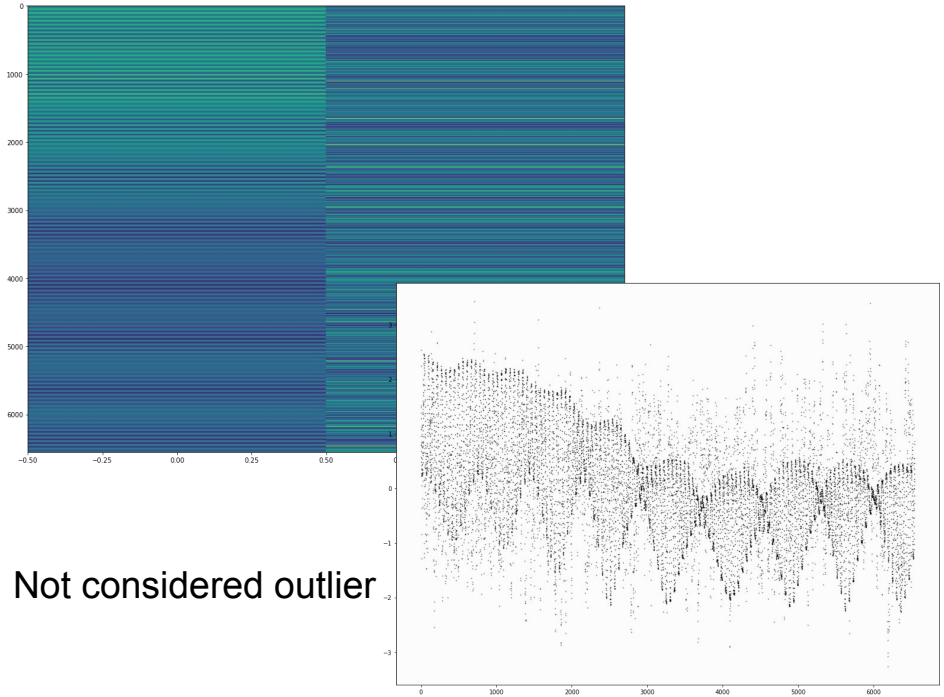
Classification avoids the stray points in **top right** plots.

The anomalies may not be instabilities but they are correctly identified as anomalous.

Comparison with instability table

One case from instability table:

Fill number	Fill type	Cycle phase	Date	Energy [TeV]	beam	plane	total nb of bunches	nb unstable bunches	Tag	other info
6561	Commissioning	FLATTOP	2018-04-15 18:11:24	6.5	1	H	2	1		(emittances are hypothetical, no meas avail.)



Not considered outlier

```
/nfs/cs-ccr-adtobsnfs/lhc_adtobsbox_data/6561/
instability_data/06561_Inst_B1H_Q7_20180415_18
h12m30s.h5
     Predicted unstable bunches: 1
     Number of bunches: 2
```

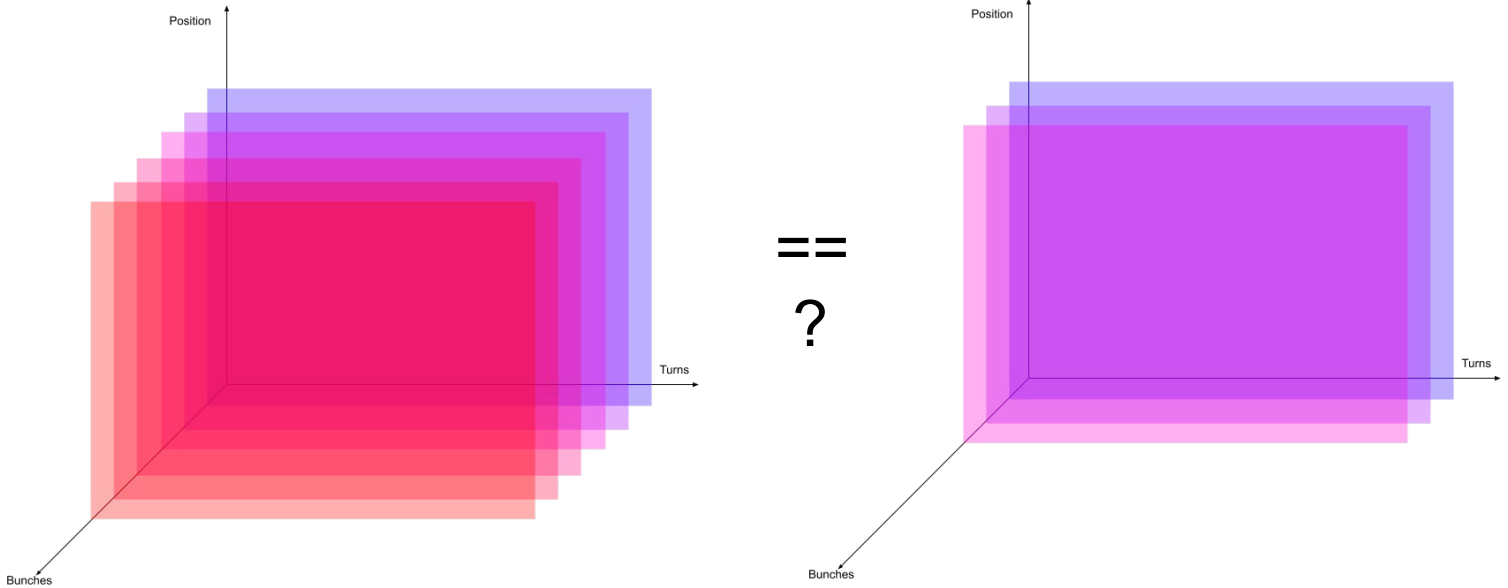
Finds correct number of unstable bunches in next file, ~1 min after the table's date entry.
→ instability table data columns doesn't line up with timestamps on files.
→ need to tune the isolation forest's parameters.
→ makes quantitatively measuring the accuracy the anomaly detection challenging.

In most cases finds instability in or around table entry.

Outlier Classification

The **challenge**: classify different types of instabilities.
The **problem**: multivariate time series (number of bunches) with differing number of dimensions (bunches).

Consider each bunch as independent, figure out how to extend to multi-bunch later.



Classification within the outliers

Time series classification, proof of concept at the **bunch level (univariate)**.

Time series **distance metric** ? Dynamic Time Warping [1]

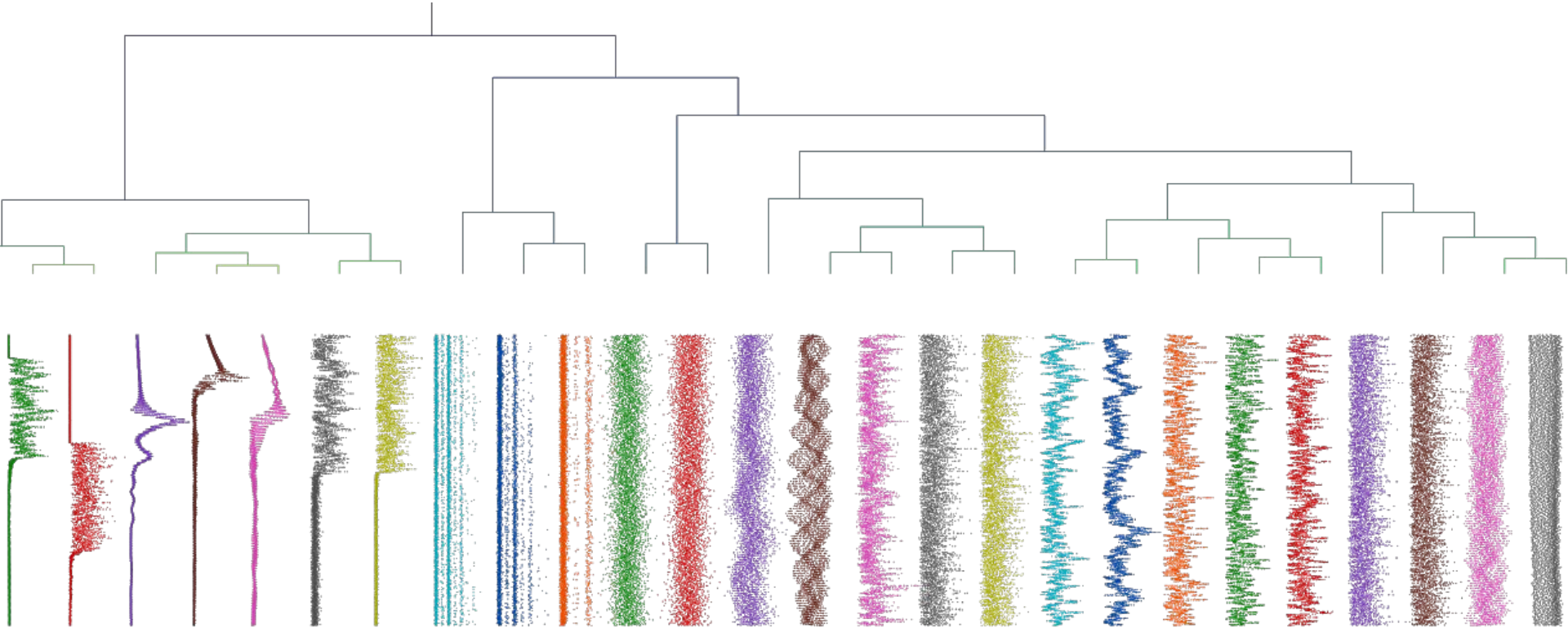
→ implementations: FastDTW [2] & dtaidistance [3]

→ **distance matrix** of the outlying time series'

Clustering using **Hierarchical Clustering**

[1] R. Bellman and R. Kalaba, "On adaptive control processes," Automatic Control, IRE Transactions on, vol. 4, no. 2, pp. 1–9, 1959.
[2] https://github.com/slaypni/fastdtw
[3] https://github.com/wannesm/dtaidistance

Linkage Tree



Full plot: https://cernbox.cern.ch/index.php/s/F6m2LQlVVBvCK79 or https://imgur.com/a/jeDk8ts

Conclusion

OBsBox:

- Anomaly detection for instability detection ~working
- Some preliminary (univariate) time series clustering ~working
- Proof of concept seems to produce coherent results
- Improvement:
  - Upgrade the anomaly detection step, no feature extraction
  - Find a way to get an accuracy metric → tuning
  - Figure out where to cut the dendrogram → defines the clusters

How to make the anomaly detection online compatible ?

- Model input should be the ~raw ObsBox data stream
  - Recurrent Neural Network → LSTM?
  - *Time Series Anomaly Detection Using Convolutional Neural Networks and Transfer Learning* → U-Net (basically a Convolutional AE)