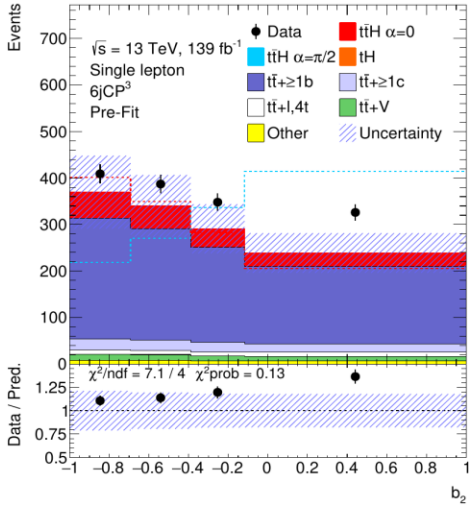
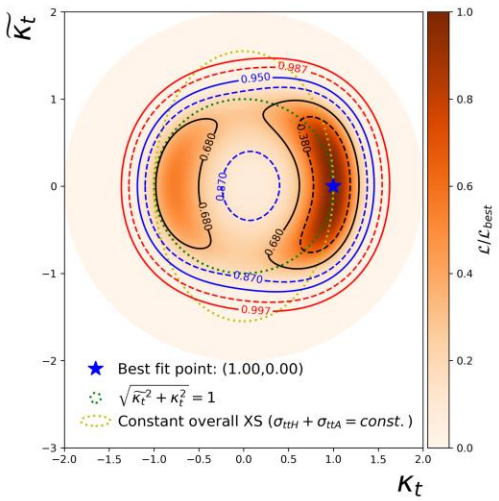
Data science in collaborative innovation projects

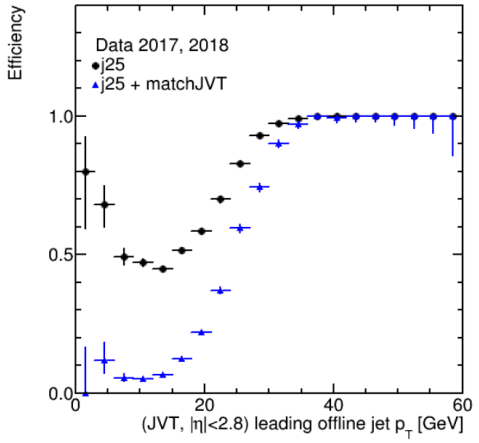
Emanuel Gouveia
DTx CoLab

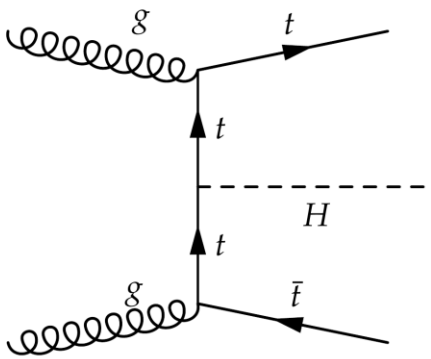
Data Science in Fundamental Physics and the bridge to industry & society
IGFAE, Santiago de Compostela
07/06/2024

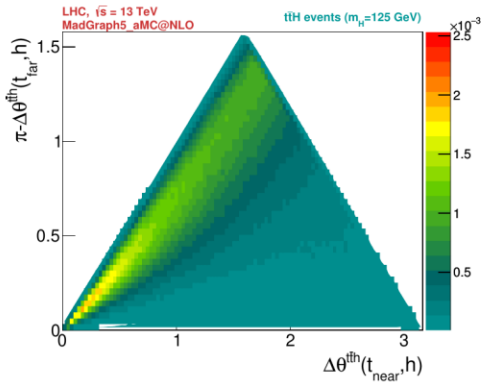
My background

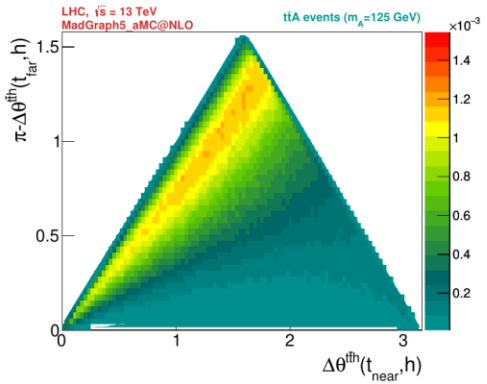












Universidade do Minho

LIP

ATLAS EXPERIMENT

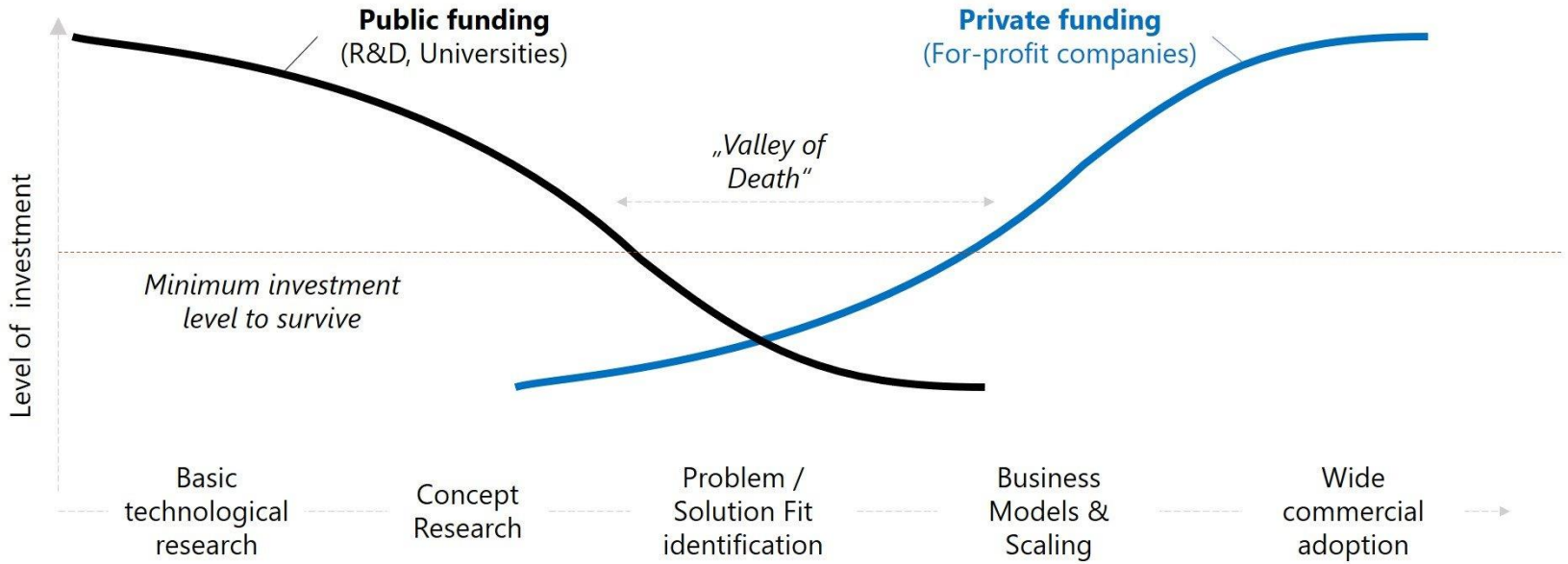
Digital
Transformation
CoLab

Experiencing
the Future

dtx-colab.pt

Rede de Laboratórios Colaborativos

Innovation Valley of Death



idea to value.com
The community for creativity and innovation

Associates

accenture
BOSCH
Tecnologia para a vida
Cachapuz
The best weigh
CEiiA
celoplás®
Centro de Computação Gráfica
cegid Primavera
SOCIETÀ COOPERATIVA
BILANCIAI
Strumenti e Tecnologie per Pesare
IKEA
INL
INTERNATIONAL IBERIAN NANOTECHNOLOGY
SIBS
AERnnova
NOS
piep
inovação em engenharia de polímeros
dstgroup
building culture
GRUPO
Simoldes
Plastic Division
TMG
AUTOMOTIVE
UNIVERSIDADE CATÓLICA PORTUGUESA
UNIVERSIDADE DE ÉVORA
Universidade do Minho
mobileum
Action driven by intelligence

RAMONES
CONCERT
WE ARE
DTX CoLAB

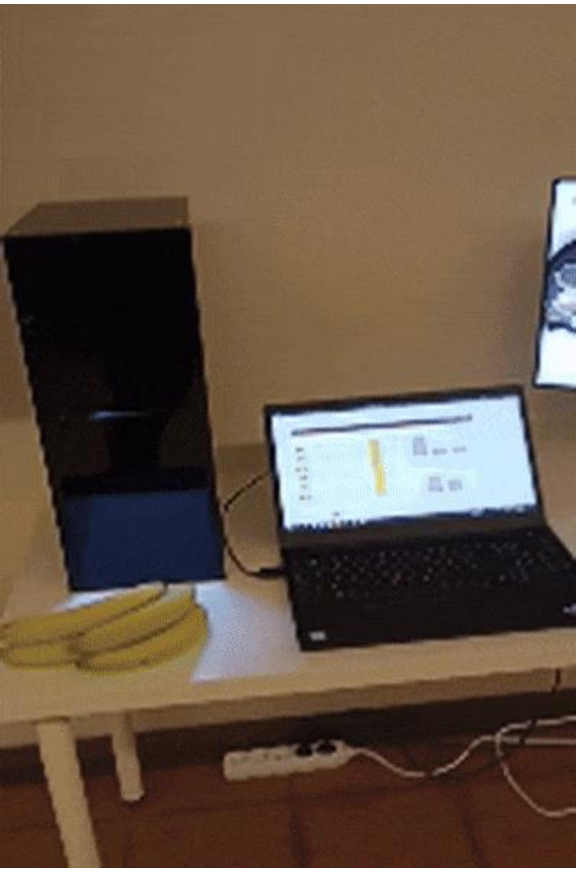
Software and Information Systems			Hardware and Sensors	Advanced Materials	Smart Manufacturing	Human Factors
Technological Innovation & Transfer Group **Data Science and Machine Learning**	Technological Innovation & Transfer Group **Data and Application Engineering**	Technological Innovation & Transfer Group **Computer Graphics and Vision**	Technological Innovation & Transfer Group **Embedded and Edge Computing**	Technological Innovation & Transfer Group **Functional and Sensitive Materials**	Technological Innovation & Transfer Group **Process and Equipment Simulation**	Technological Innovation & Transfer Group **Ergonomics and Engineering Psychology**
Future & Emergent Technologies Group						

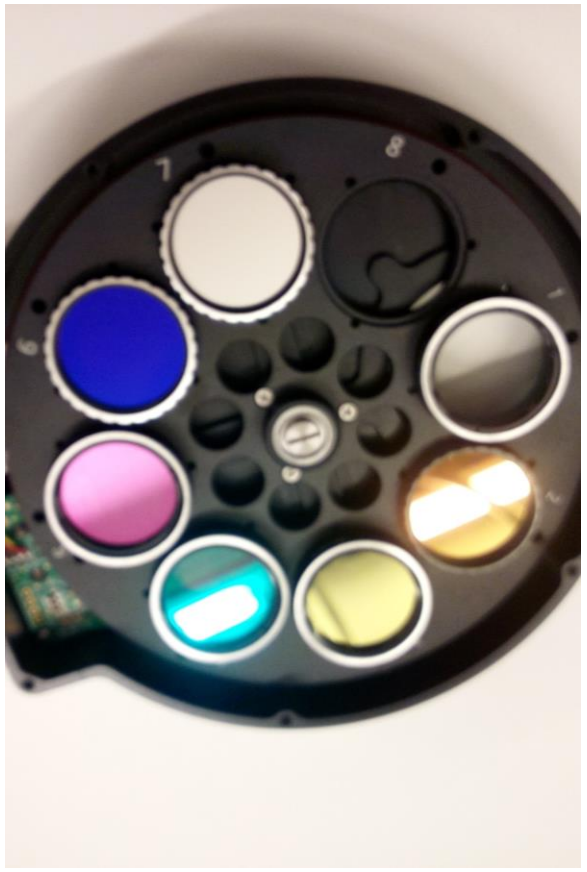
Data Science and Machine Learning Group

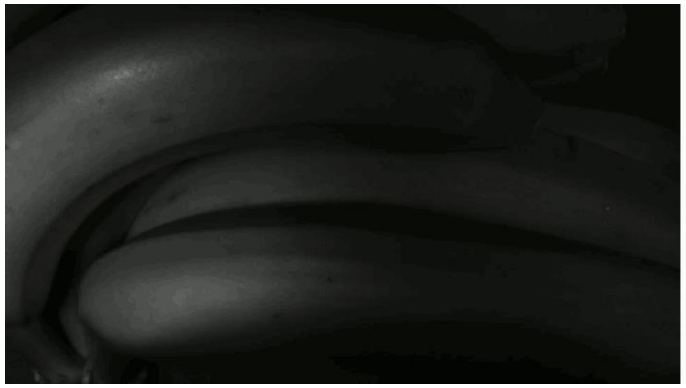
- 12 researchers
- 6 with PhD
- 5 with physics degree

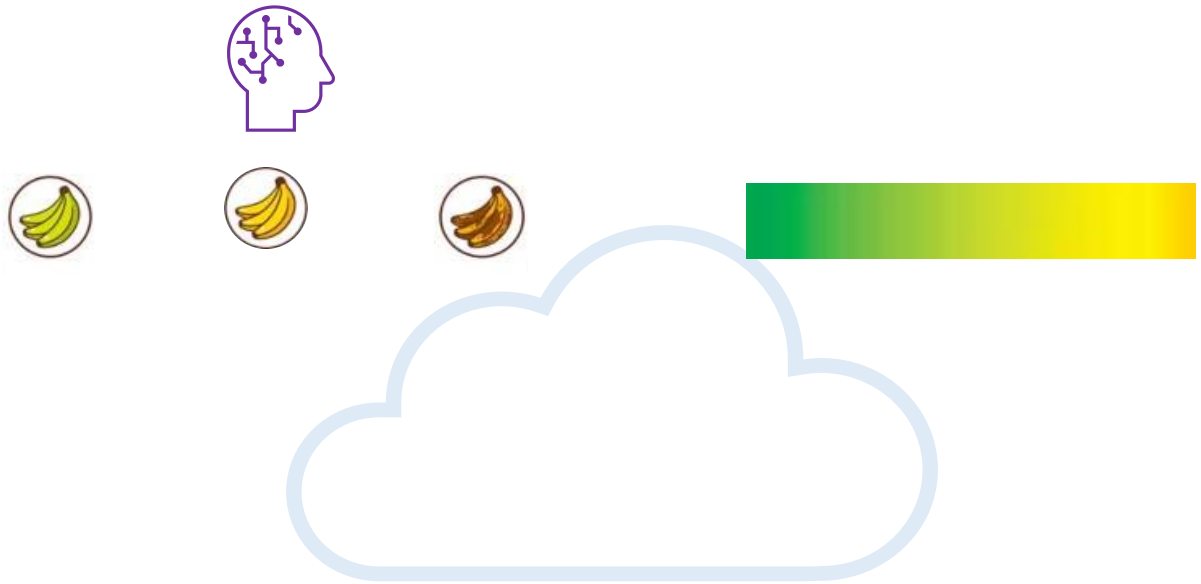
Projects

Automated optical inspection of perishable food









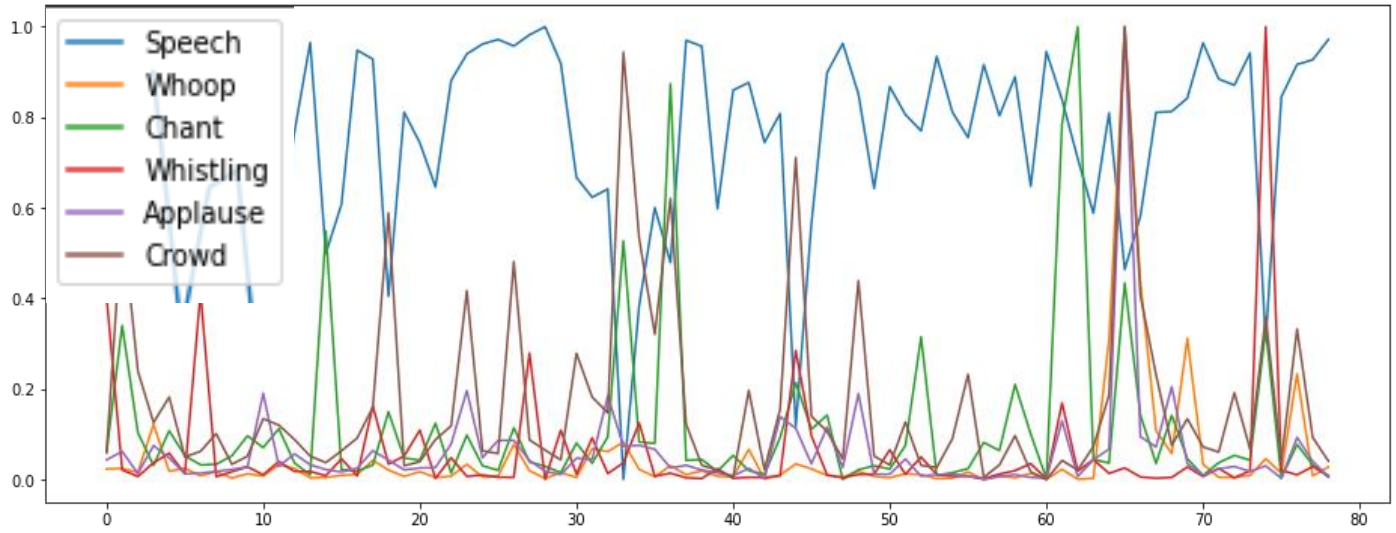
Highlight detection in audio of sports TV broadcasts

Highlight detection in audio of sports TV broadcasts

Pre-trained audio classifier

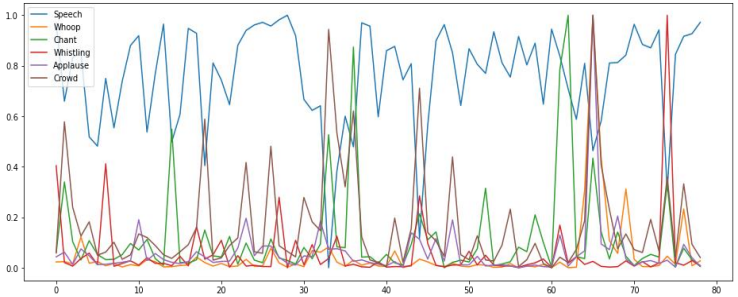
Highlight detection in audio of sports TV broadcasts

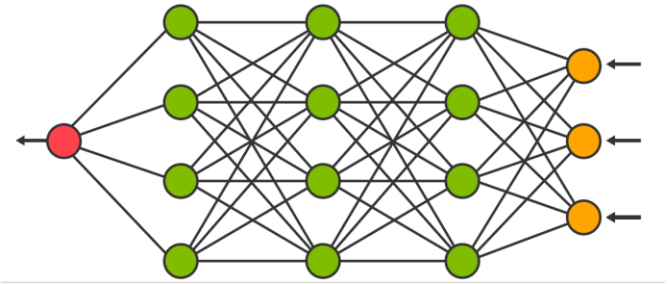
Pre-trained audio classifier



Highlight detection in audio of sports TV broadcasts

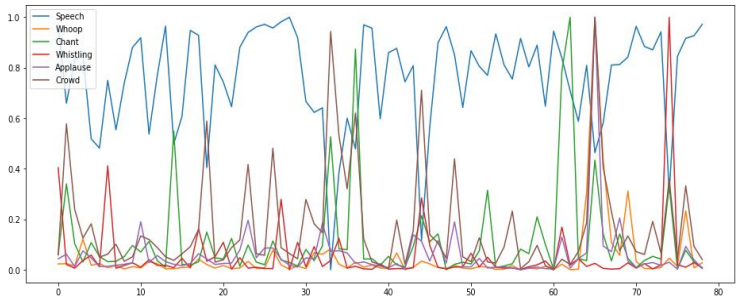
Pre-trained audio classifier

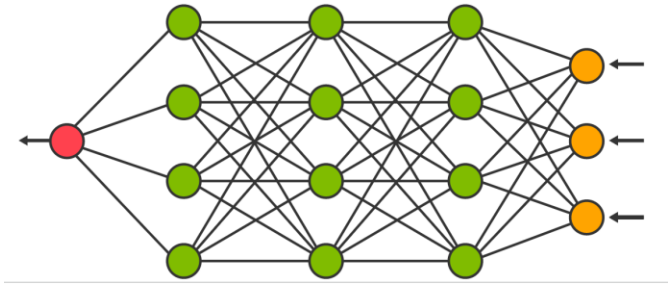


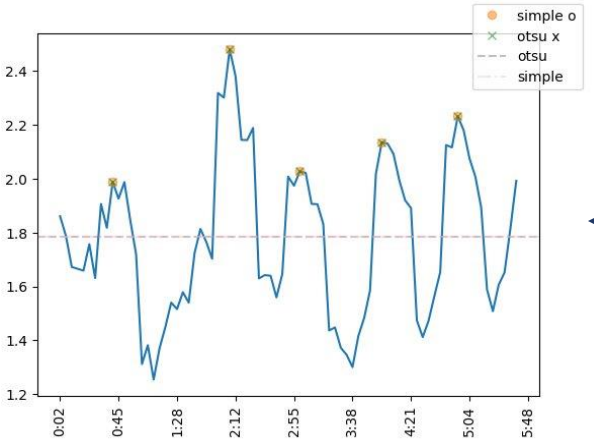


Highlight detection in audio of sports TV broadcasts

Pre-trained audio classifier







Highlight detection in audio of sports TV broadcasts

Automated speech recognition

Highlight detection in audio of sports TV broadcasts

Automated speech recognition


```
],
"18": [
    128.48,
    131.3,
    "Seems that an Ibrahimovic dropping off the front as he does a lot"
],
"19": [
    132.18,
    136.76,
    "When Cavani plays that allows Cavani to spin into space at centre-forward"
],
"20": [
    137.22,
    139.22,
    "Which can play the ball over the top for him"
```

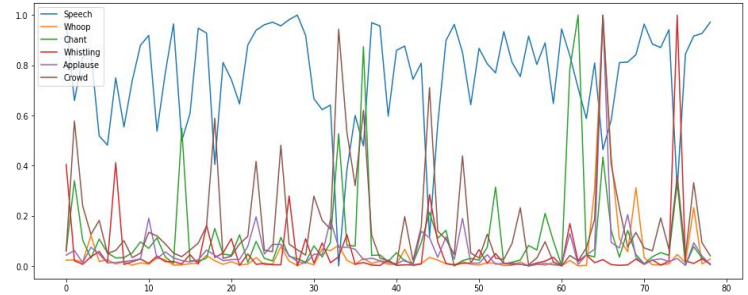
Highlight detection in audio of sports TV broadcasts

Automated speech recognition

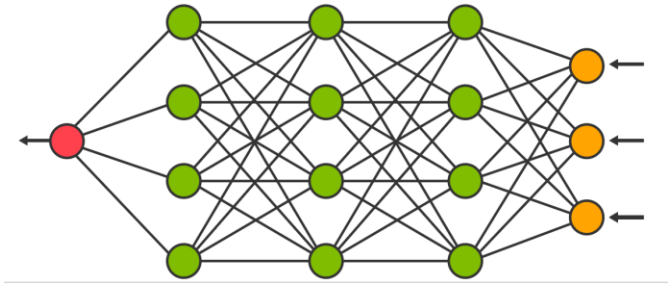

```
],
"18": [
    128.48,
    131.3,
    "Seems that an Ibrahimovic dropping off the front as he does a lot"
],
"19": [
    132.18,
    136.76,
    "When Cavani plays that allows Cavani to spin into space at centre-forward"
],
"20": [
    137.22,
    139.22,
    "Which can play the ball over the top for him"
```

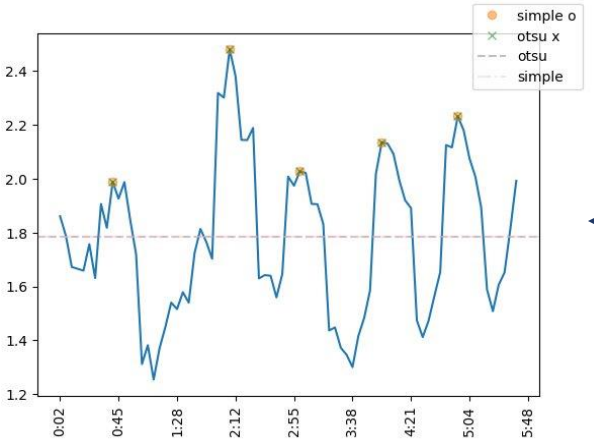

Zero-shot text classification

Highlight detection in audio of sports TV broadcasts



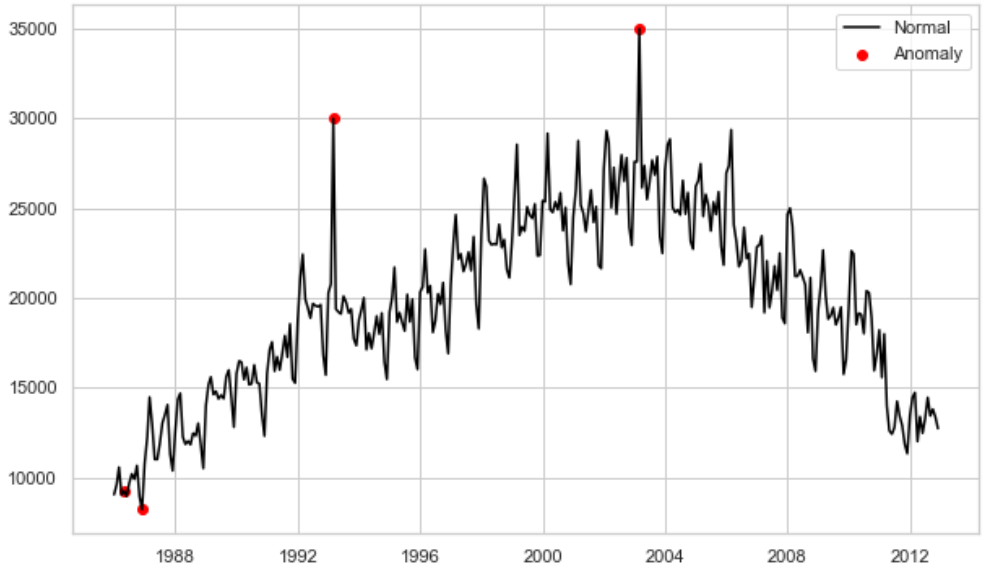
Zero-shot text classification



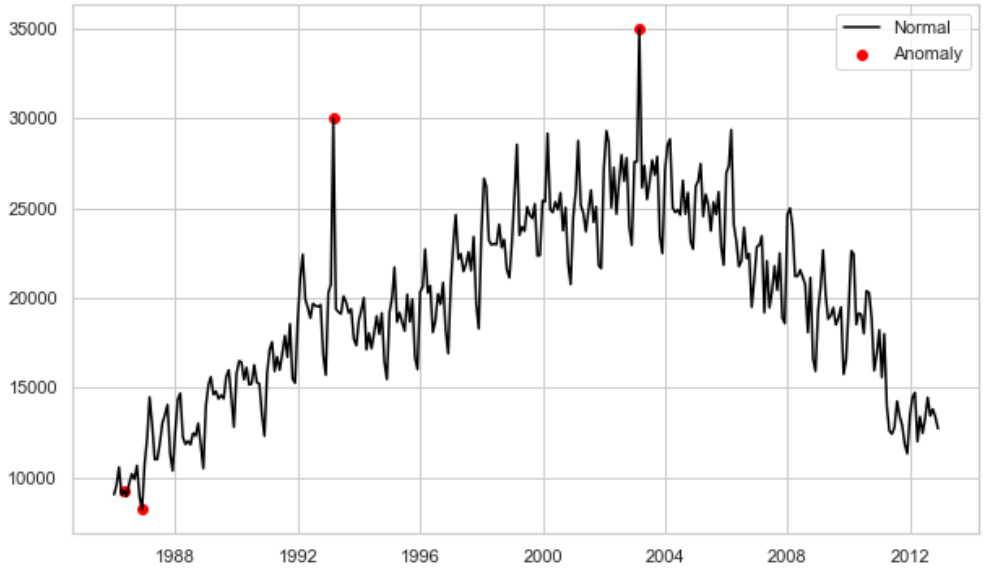


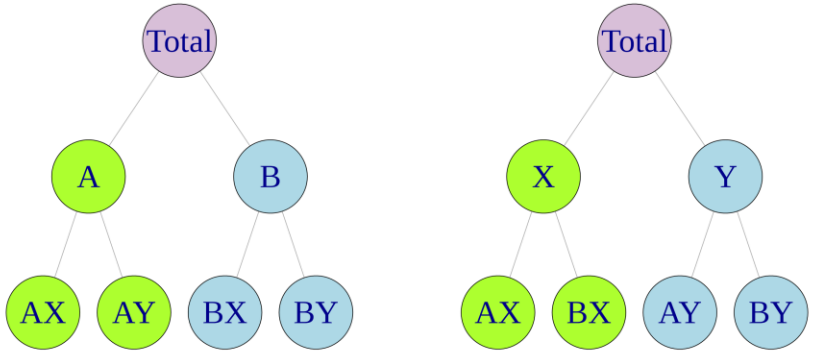
Anomaly detection in grouped time series

Anomaly detection in grouped time series



Anomaly detection in grouped time series





Anomaly detection in grouped time series

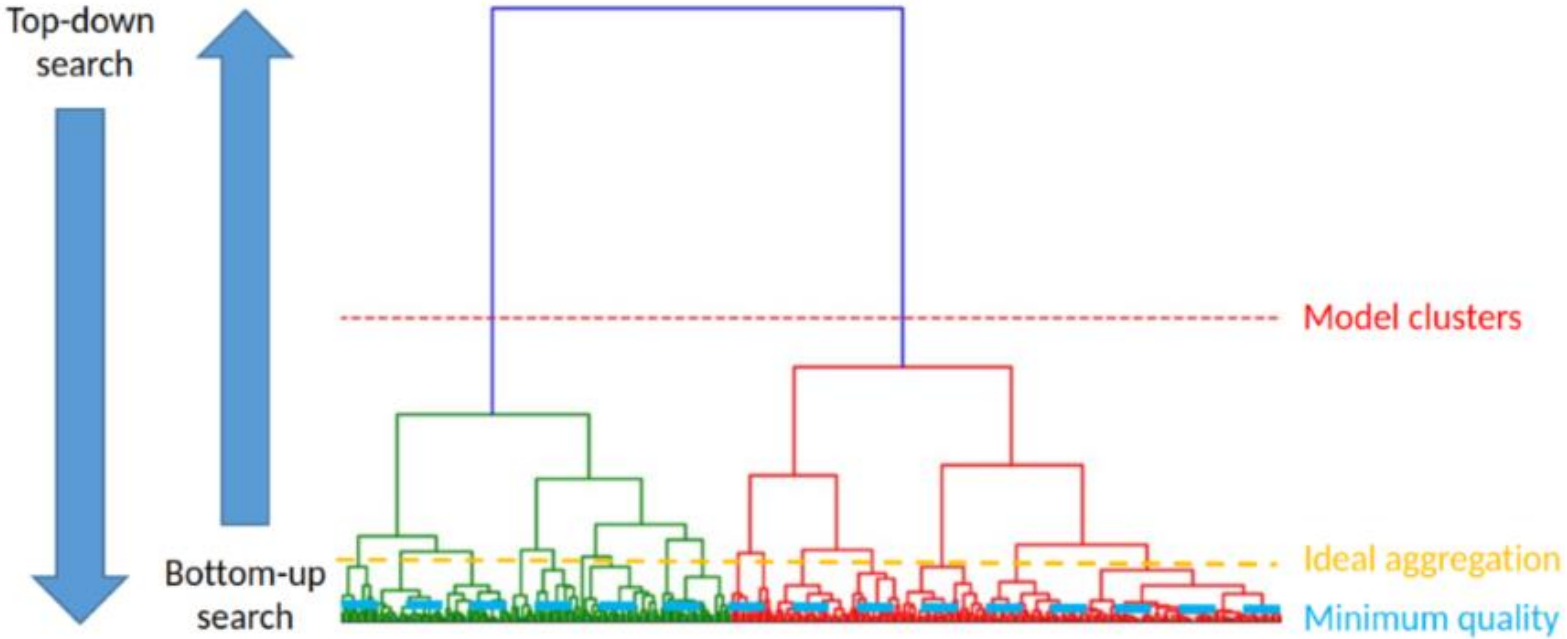
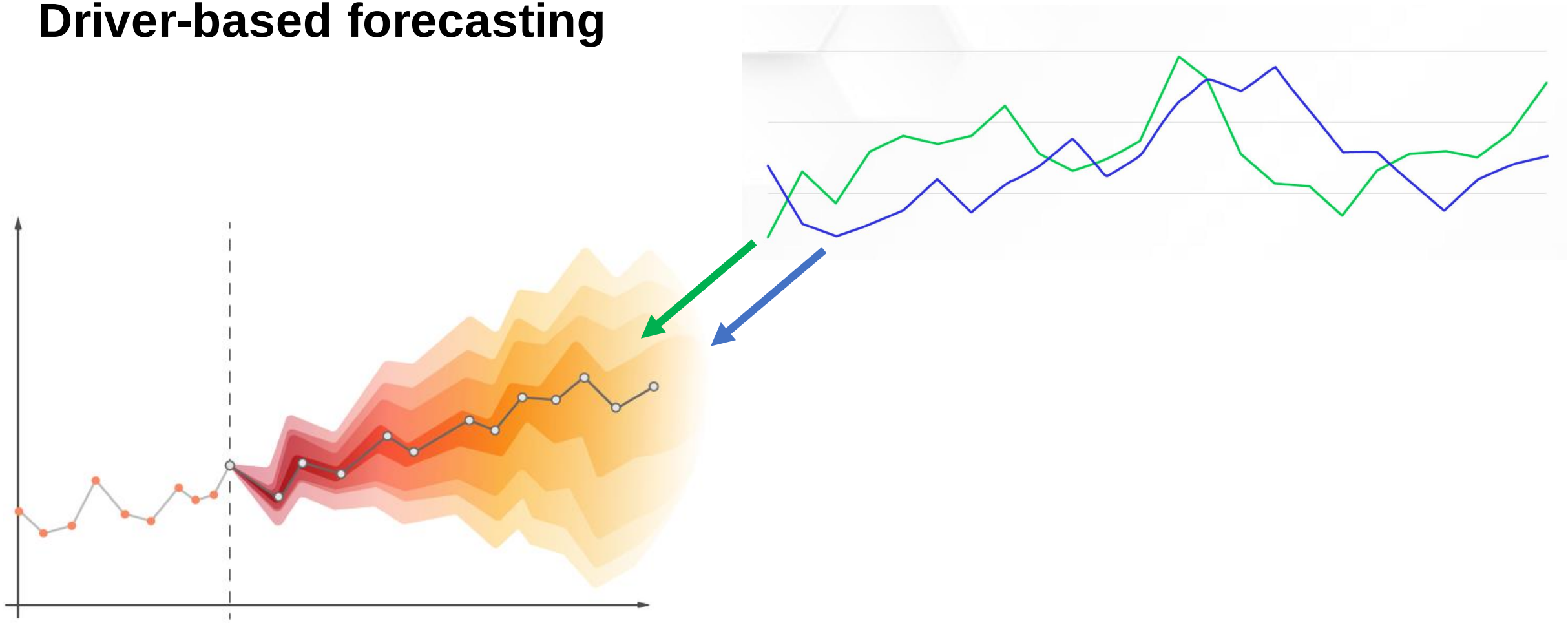
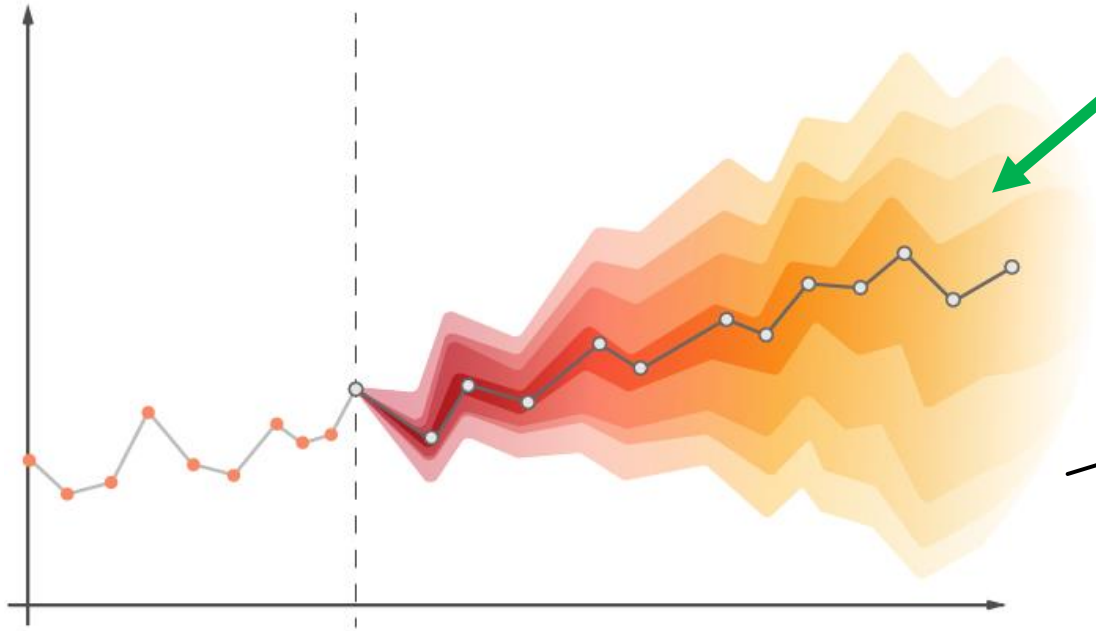


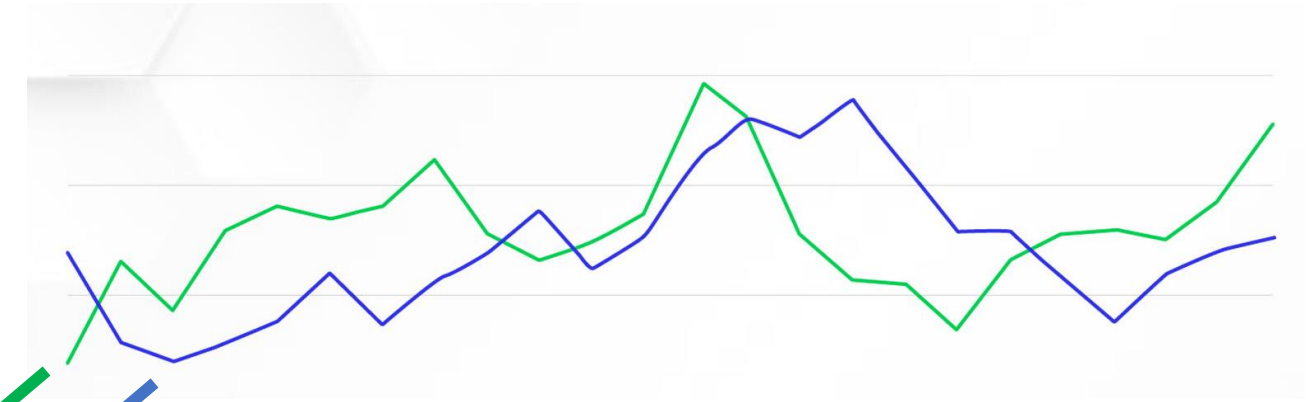
*Figure 2 - Dendrogram representing the established hierarchical structure. The search for the level of aggregation can be bottom-up or top-down. The search is informed by the **quality criteria, expected anomaly detection performance and computational cost**, enabling to find the ideal model clusters and corresponding time-series (aggregated or not).*

Driver-based forecasting



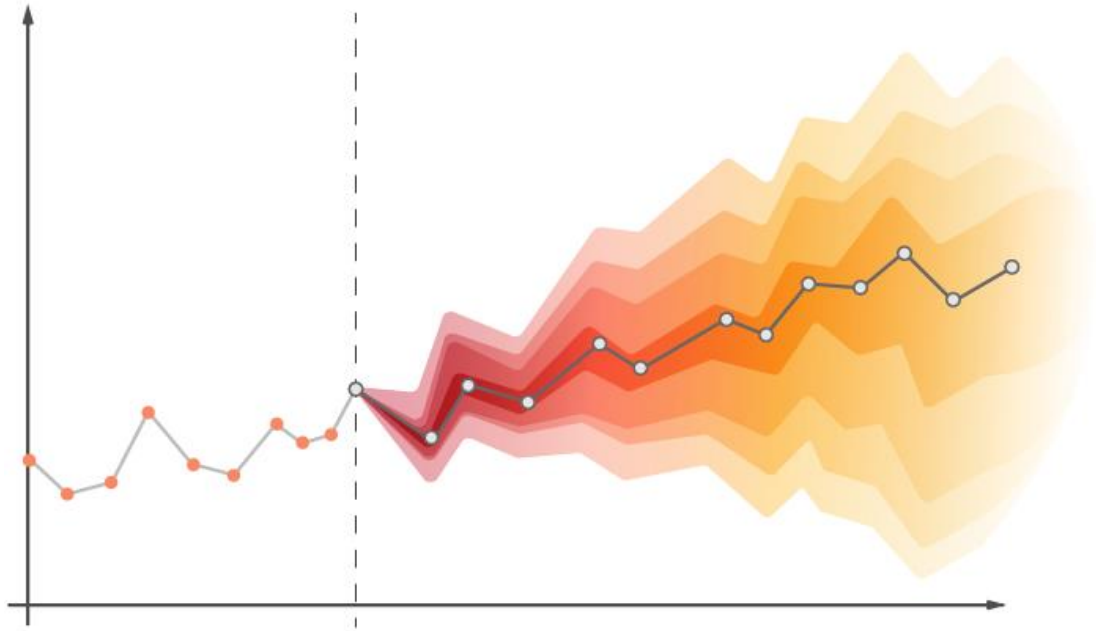
Driver-based forecasting

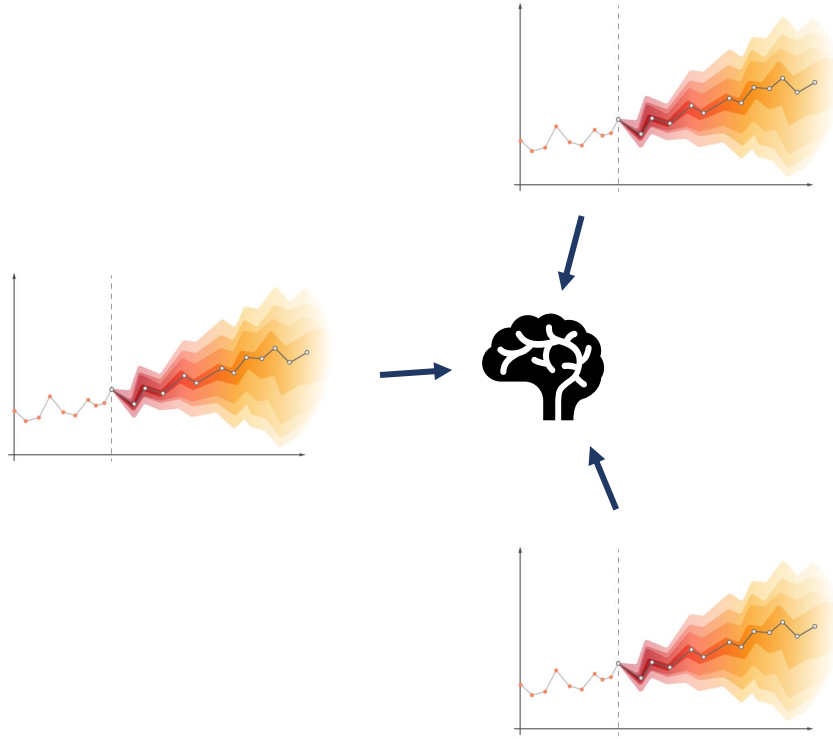




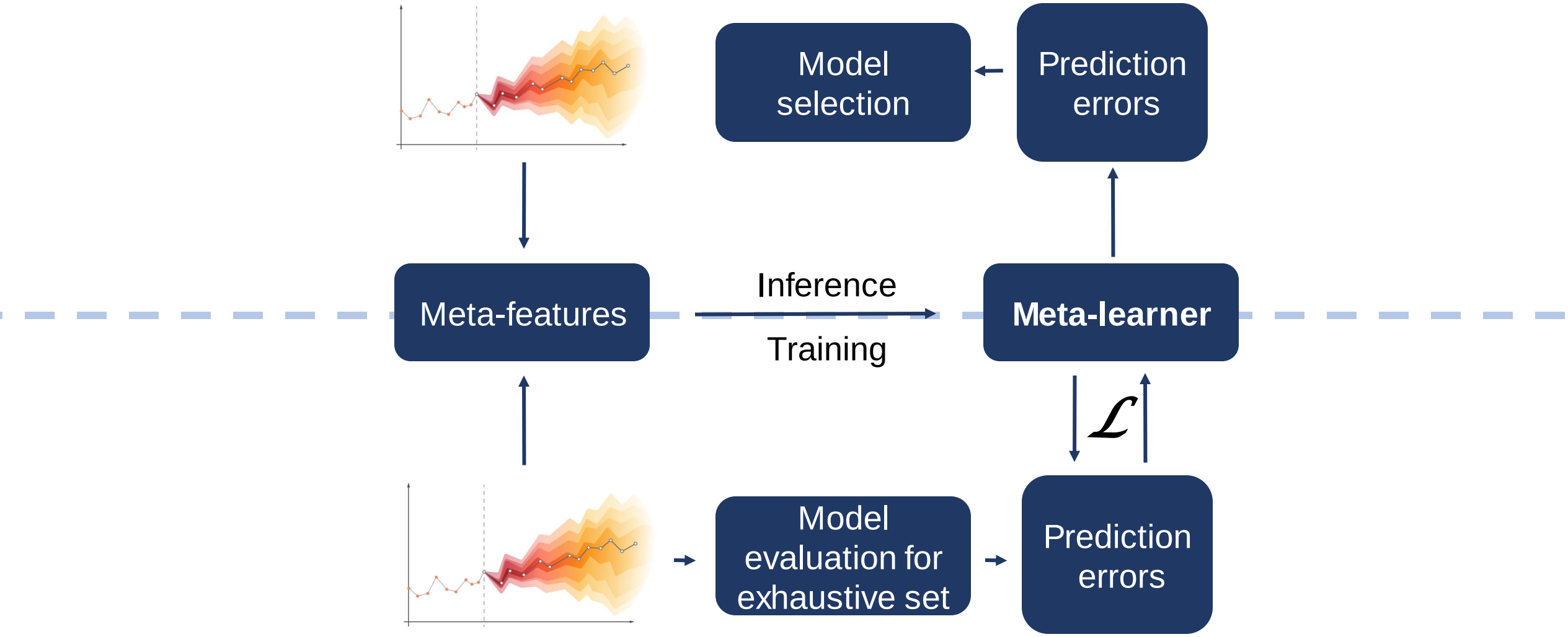
$$C_t = A_{t\text{-}lag} + g(B,C) + noise$$

Driver-based forecasting

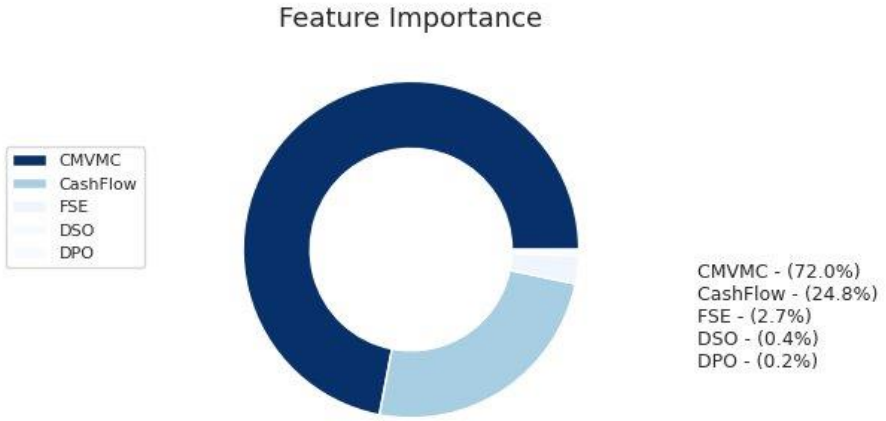


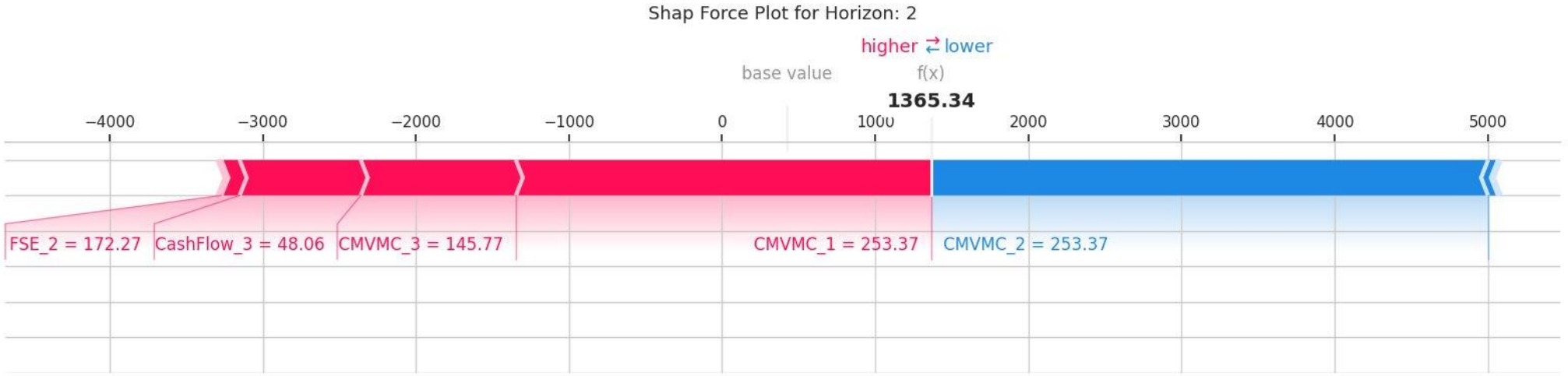


Driver-based forecasting

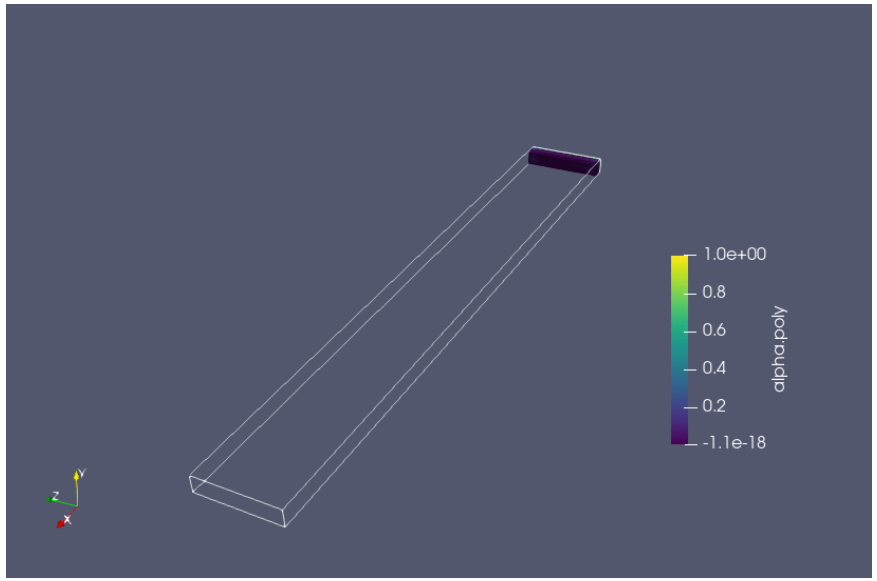


Driver-based forecasting

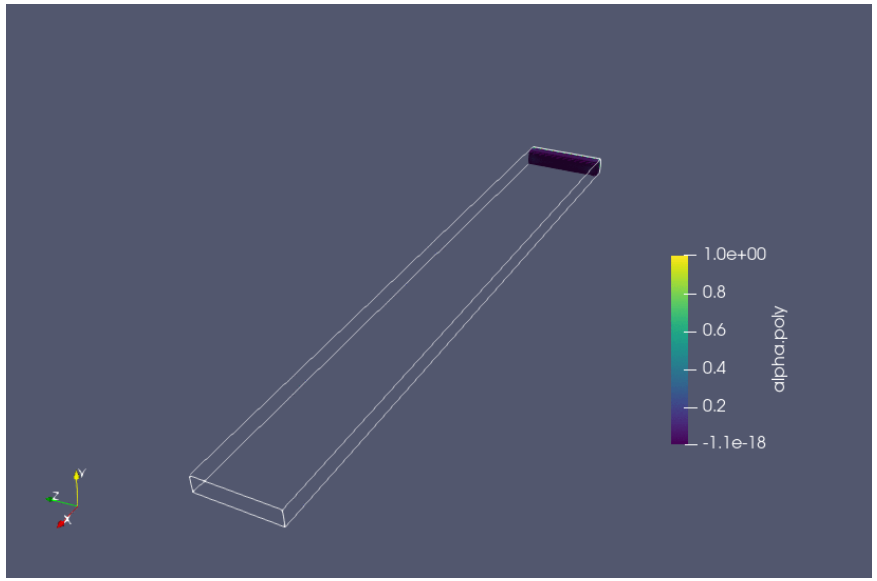




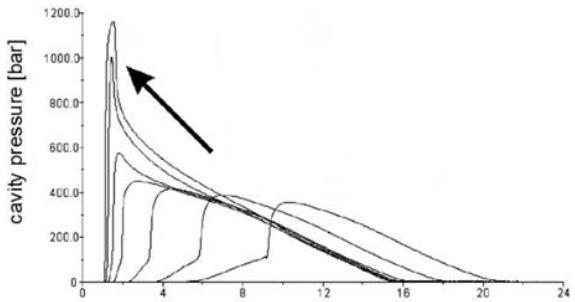
Industrial process control



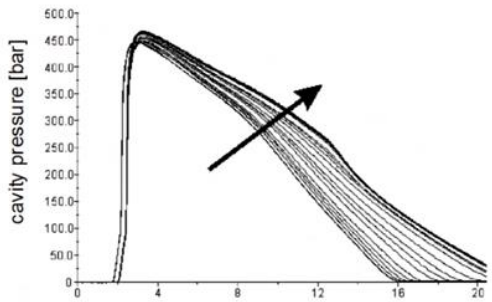
Industrial process control



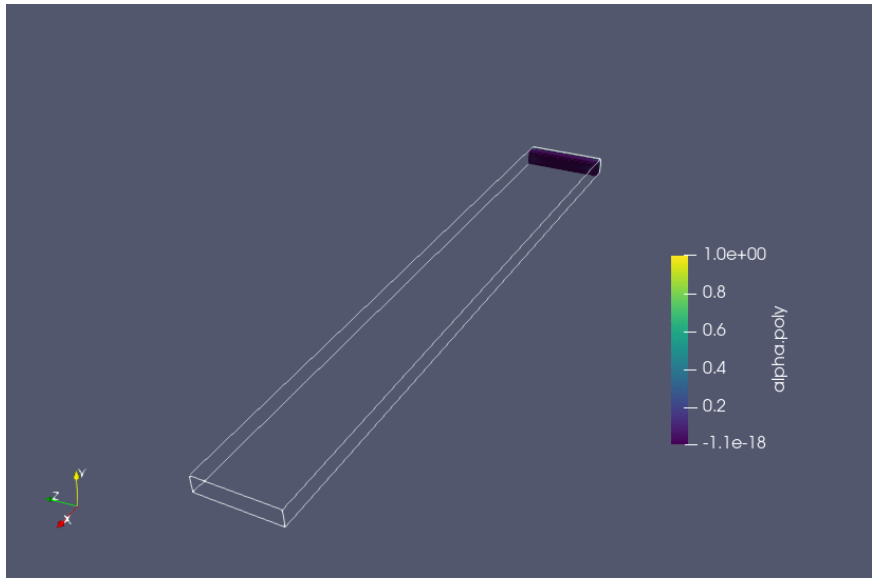
Injection Speed



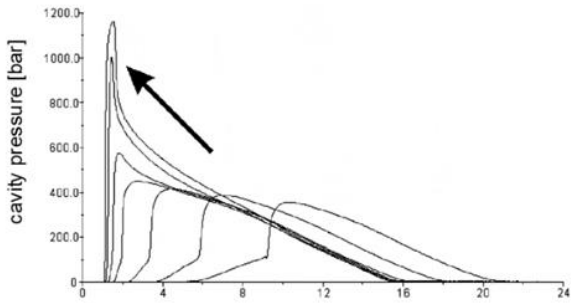
Mold Temperature



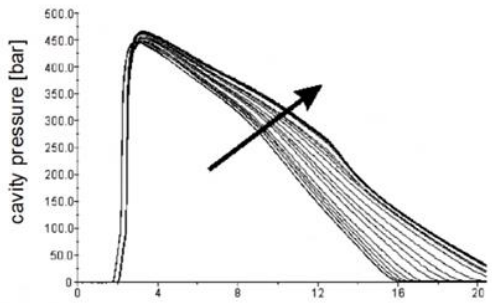
Industrial process control

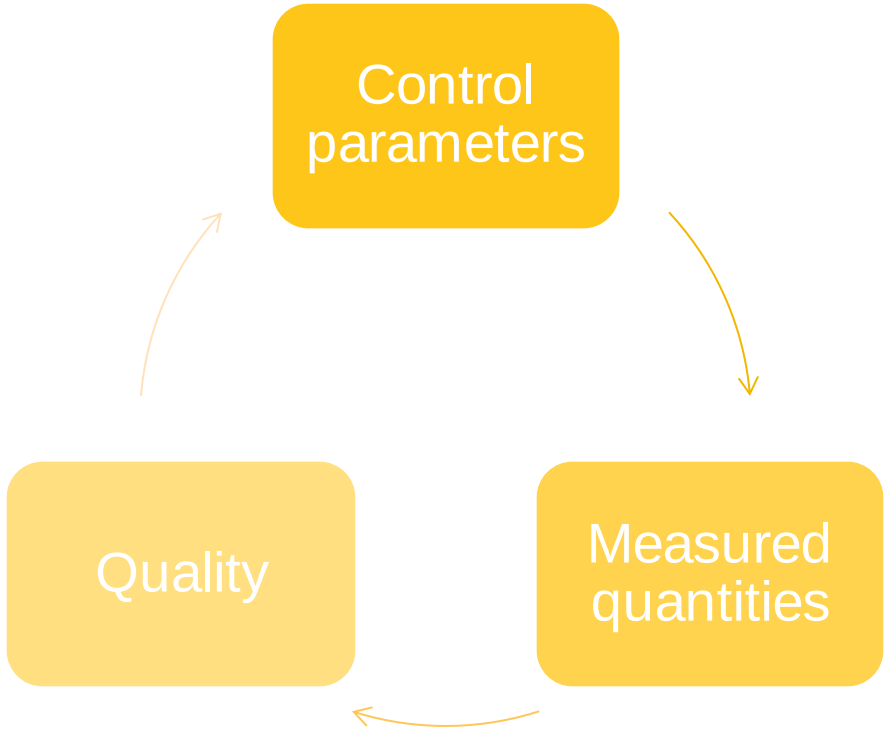


Injection Speed



Mold Temperature





Final remarks

Final remarks

- Interface organizations like DTx provide unique environment to grow as data scientist
  - Diverse interesting problems from industry
  - Exposure to network of clients/associates
  - Focus on economic value creation
  - Still with more room for experimentation than in companies
- How physics training can differentiate you in data science
  - Not too focused on particular tools
  - Intuition to look for what can be neglected and to make approximations
  - Confidence to come up with new approaches from first principles, or to adapt established methods

Final remarks

- Potential pitfalls:
  - Physicists have a talent for finding first-principle reasons to conclude that something is not worth trying
  - "The data is not good/enough" - some times it's true, other times it's a lazy response. Consider injecting domain knowledge or using transfer learning, data augmentation, dimensionality reduction, …
  - Ignoring the domain knowledge. As a data scientist, you will not become a domain expert for every project, but being curious about the domain will only give you an advantage

emanuel.gouveia@dtx-colab.pt