



Multi-view analysis of temporal data

Elena Tsiporkova

- Focusing on **data-driven AI**, i.e. addressing **real-world industrial challenges**
 - which **cannot be solved** via straightforward application of prediction/classification learning paradigms
 - and where data is **not immediately suitable** for consumption by traditional AI algorithms
- Proven track record in realizing data-driven proof-of-concepts in **diverse industrial application scenarios**



TEAM



21 scientists, 12 with PhD, 10 different nationalities, 6 women



driving industry by technology

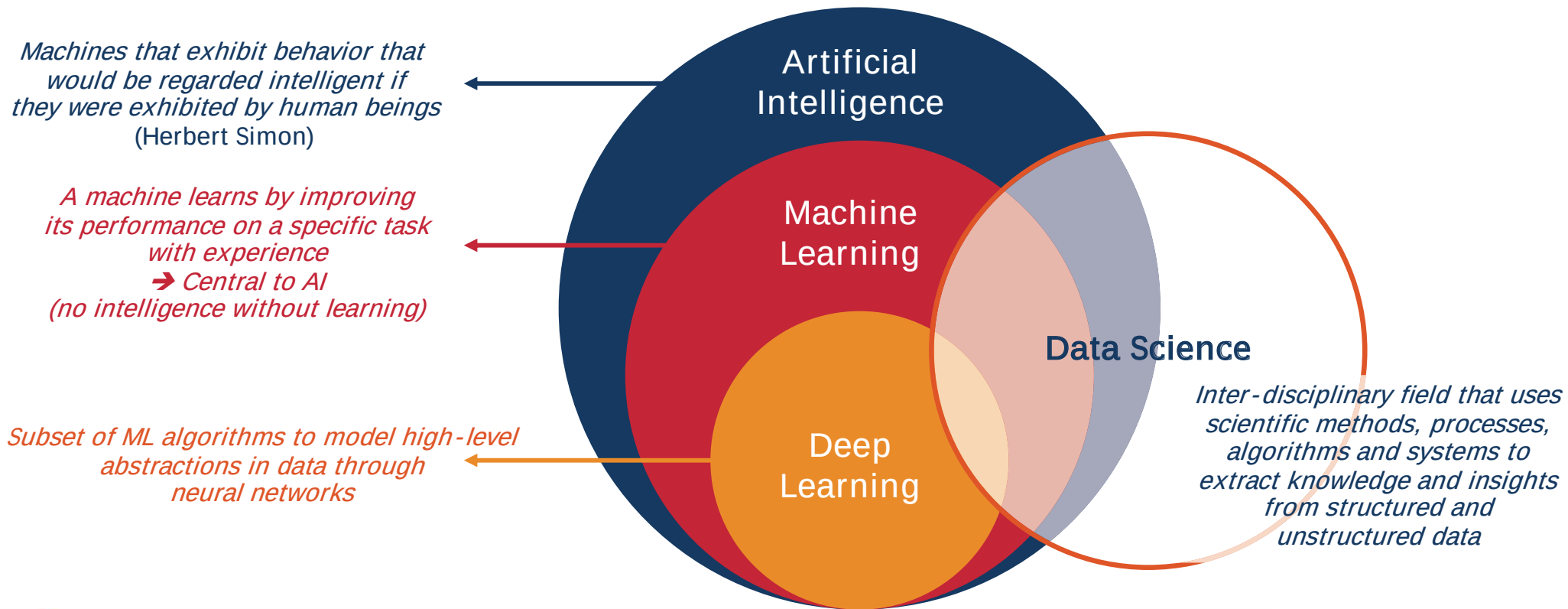
© sirris



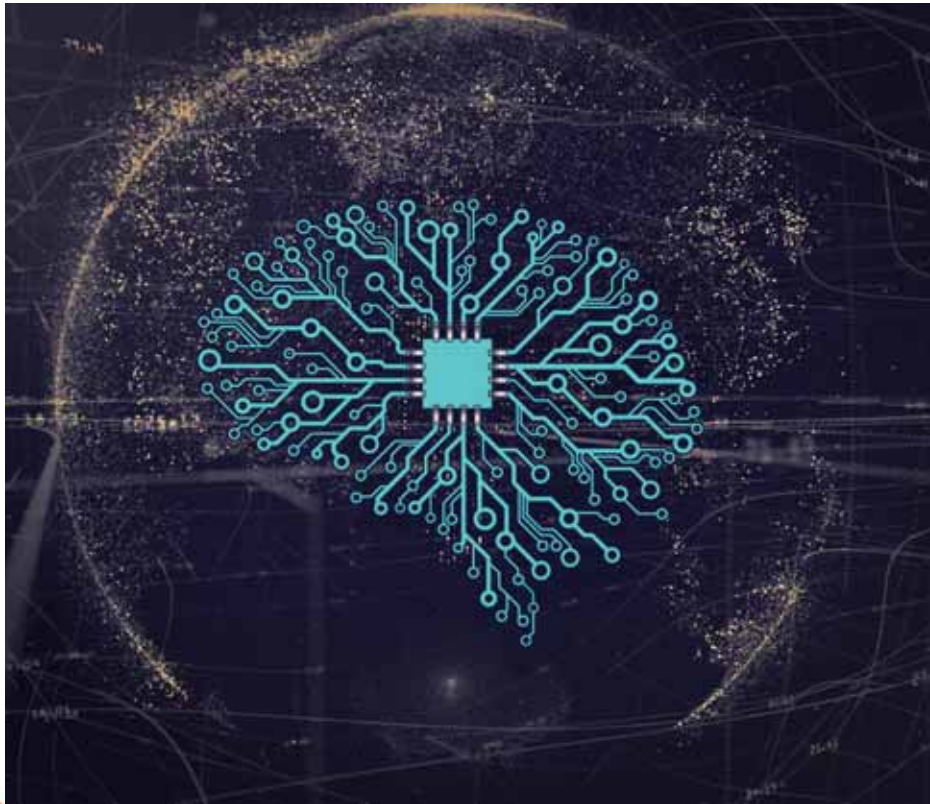
Outline

- Data science vs. AI
- Software vs. AI system
- The importance of data
- Improving the code vs. improving the data
- Challenges of real-world datasets
- Multi-view data analysis
 - Context-aware profiling
 - Multi-source integration

Data science vs. AI



The transformation of AI into a new "mega-technology".



- This new technology, was made possible primarily by
 - novel hardware
 - the availability of large amounts of data
- Deep Learning has played an overwhelming part in the commercial success of AI
 - seen today as a solution to almost any problem for which sufficient data is available

However, some disillusionment emerged

- Deep Learning does not build a true understanding of the problem
 - It merely captures and maps very complex statistical relationships
- Deep Learning is also characterized by the lack of
 - *Interpretability*
 - “black box”, the results are not understandable by humans
 - *Robustness*
 - high susceptibility to changes in the input signal
 - *Generalisability*
 - poor transferability to cases not covered by the data
 - *Fairness*
 - potential discrimination and biases infused by the data
 - *Data efficiency*
 - “big data hunger”, needs a lot of data to achieve high precision
- The expectations of industry and society at large for AI are not fully fulfilled



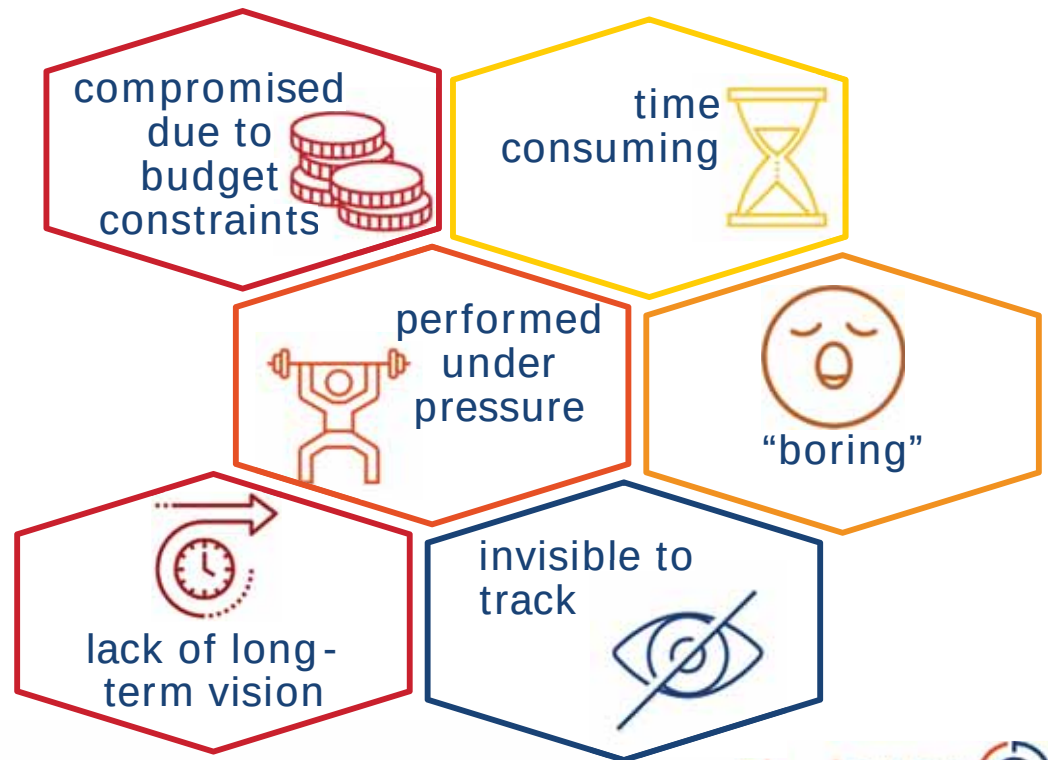
Software vs. AI system

- Traditional software is powered only by **code**
- AI systems are built using both
 - **code** implementing models & algorithms
 - **data** used to construct the models



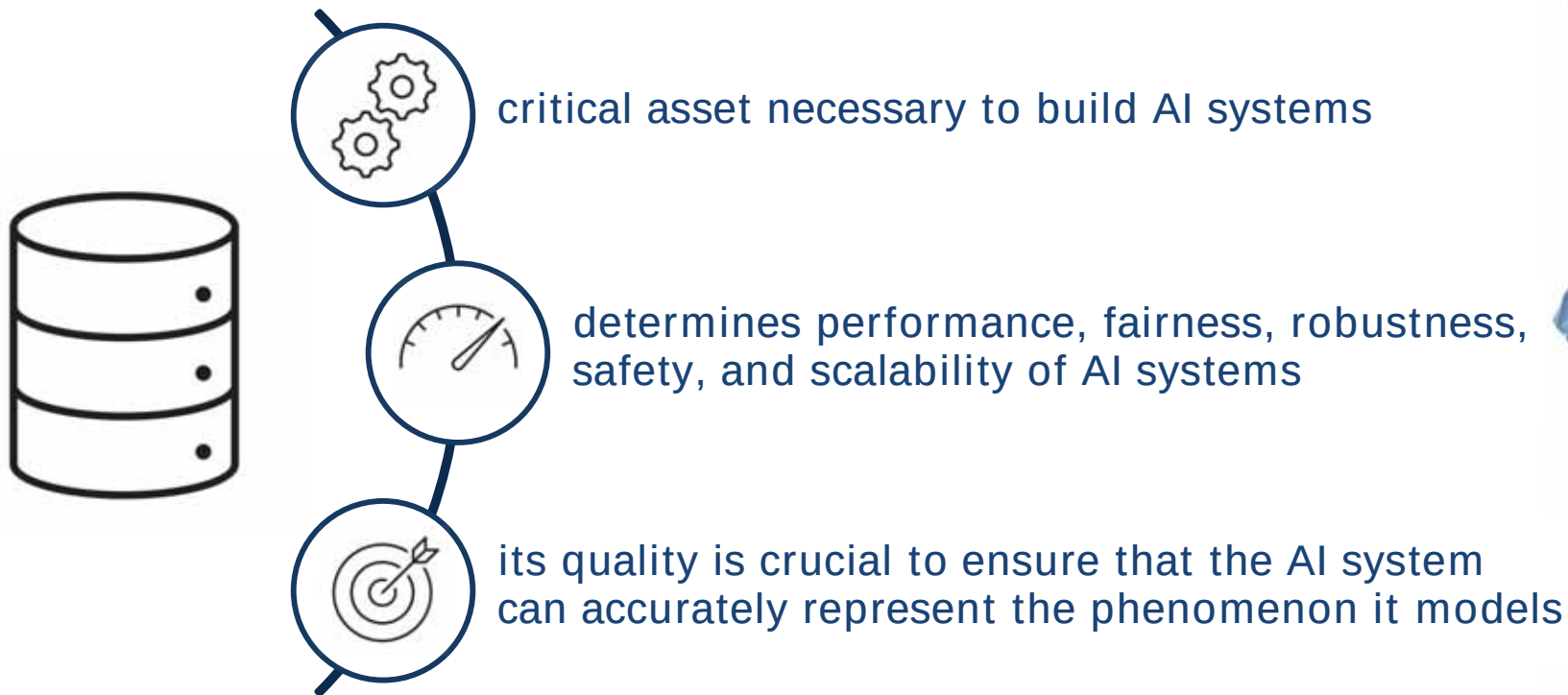
“Everyone wants to do the model work, not the data work”

Most industrial contexts lack data-driven mindset



The importance of data is still not well understood

Data quality is essential to avoid the garbage in, garbage out problem



Improving the code vs. improving the data

- Andrew Ng (co-founder Coursera & deeplearning.ai), 06/04/2021:
 - *When a system isn't performing well, many teams instinctually try to improve the code. But for many practical applications, it's more effective instead to focus on improving the data, ...*
 - *If 80 percent of our work is data preparation, then why are we not ensuring data quality is of the utmost importance for a machine learning team ...*



Improving the data

Transformation into multiple views -> divide-and-conquer

- Many industrial assets are instrumented with monitoring systems generating *a vast amount of time series data*
- Making sense of such data is not always trivial
- High level of detail obscures interesting long-term trends and patterns



The transformation of data into multiple data representations/views
empowers the application of advanced analytics and reasoning

Real-world datasets are inherently complex in nature

- **Heterogeneous (multi-modal) data**

- Originating from the internal working of the asset (oil temperature, pressure, speed, vibration, . . .)
- Measured in the environment where the asset is operating (ambient temperature, wind speed, humidity, . . .)
- Set by an operator (machine configuration, alarm thresholds, . . .)
- Log records produced by the asset itself
- Vision-based data collected by dedicated devices
- Performance indicators (water flow, active power, electricity consumption, ...)
- ...



- **Multitude (multiple sources) of assets (e.g. fleets of vehicles, farms of wind turbines, portfolio of compressors, ...)**

- Parameters may differ in coverage, resolution, data quality, technical configuration, ...



Making sense of asset performance is challenging

- Multitude of influencing factors, which are not always known and often also highly interdependent
- Very often datasets do not contain ground truth
- Most datasets suffer from lack of good quality data
 - lack of statistical power
 - lack of standardization
 - ...
- Access to *a//* relevant data is difficult



Not all the influencing factors are known/measured

Optimizing the operational performance of wind turbines



No ground truth

Identifying root causes for underperforming solar parks

vegetation
growth



shading



soiling



others



How to approach analysis of real-world datasets?

Pool datasets together

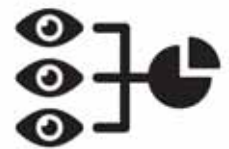
- not always feasible across different data types e.g. time series vs. images
 - does not exploit fully the richness of the different datasets
- may lead to information loss and mask source-specific features
 - e.g. not all assets might go through the same operating modes



Created by H Alberto Gongora
from the Noun Project

Multi-view approaches

- analyse separately different data representations (views)
- preserve the latent structure of the different data views
- benefit optimally from the extracted knowledge



Created by Adrian Coquet
from the Noun Project

Multi-view analysis methodology for temporal data

Multi-layered approach

Individual analysis layer

- Applies an individual data analysis per source, solely based on a subset (data view) of relevant parameters

Mediation analysis layer

- Represents the results from the previous layer by use of an alternative subset (data view) of parameters
- Allows to compare results across multiple contexts and/or data sources

Integration analysis layer

- Leverages the results from the previous analysis layers in order to derive explicit links between the different data views

Multi-view analysis methodology for temporal data

- Agnostic to the specific machine learning methods employed
- Two flavours
 - Context-aware profiling
 - Establishes a direct link between operating mode and prototypical behaviour
 - Allows to inspect behaviour for any specific context
 - Multi-source integration
 - Characterises prototypical behaviour across multiple data sources
 - Allows to inspect behaviour within and across data sources
- Potential is demonstrated in wind energy domain
 - SCADA data generated by a fleet of 4 wind turbines, located in La Haute Borne (France) and owned by Engie



Linking operational behaviour with performance

Relevant data views



Operation-related view

endogeneous parameters, i.e. originating from the *internal working* of the asset (e.g., oil temperature, pressure, speed, . . .)

- characterising the operational behaviour



Performance-related view

exogeneous parameters, i.e. originating from the *environment* where the asset is operating (e.g., ambient temperature, wind speed, humidity, . . .)

- characterising the operational context
- directly correlated to the observed performance

output parameters, i.e. resulting from the asset operation (e.g., produced energy, generated water flow, ...)

- directly measuring performance

Linking operational behaviour with performance

Multi-layered approach



Individual analysis layer
Operating Mode Identification

Identify characteristic behaviour (operating modes) within an individual source (asset e.g. wind turbine) based on the operation-related view (*endogeneous* parameters)



Mediation analysis layer
Performance Profiling

Characterise performance behaviour as a function of different operational contexts while exploiting the performance-related view (*exogeneous and output* parameters)



Integration analysis layer
Context-sensitive Performance Interpretation

Traceback derived context-specific performance profiles to a certain operational behaviour

Case study

- Data

- SCADA data from a fleet of 4 wind turbines, La Haute Borne, France (Engie), 2009 – 2017
- 31 sensors with a 10-minute interval:
 - Endogeneous (e.g. pitch angle, torque and gearbox oil temperature)
 - Exogeneous (e.g. wind speed, wind direction and outdoor temperature)
 - Output sensors (e.g. active power and reactive power)

- Goals

- Capture time windows with consistent turbine's behaviour and characterise its performance in a context-aware fashion
- Assess and benchmark a turbine's performance by comparing it to the other assets in the fleet
- Identify fleet-wide operating modes to allow for performance comparison and anomaly detection

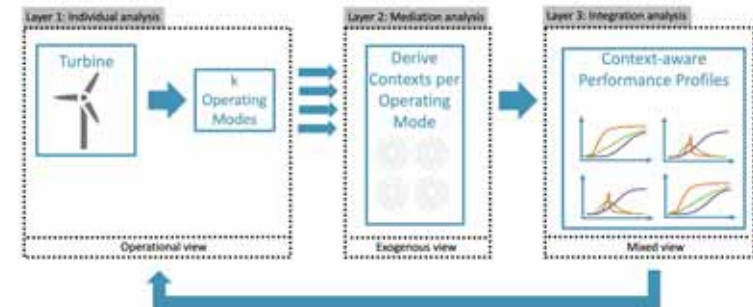


Linking operational behaviour with performance

Two flavours

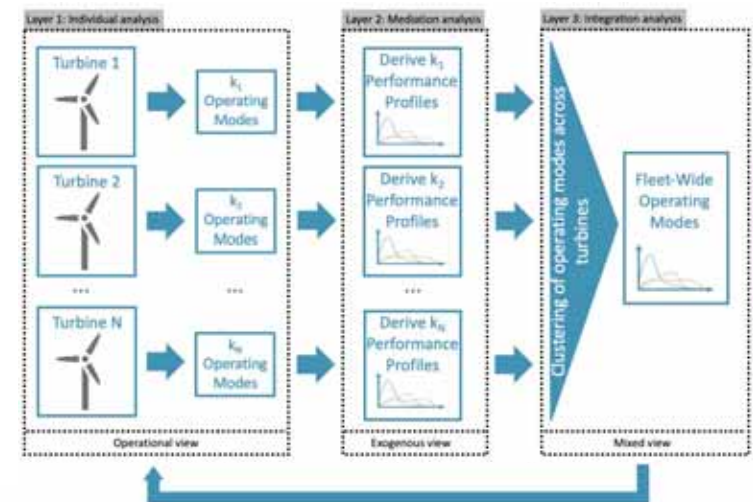
■ Context-aware profiling

- Establishes a direct link between operating mode and prototypical behaviour
- Allows to inspect behaviour for any specific context



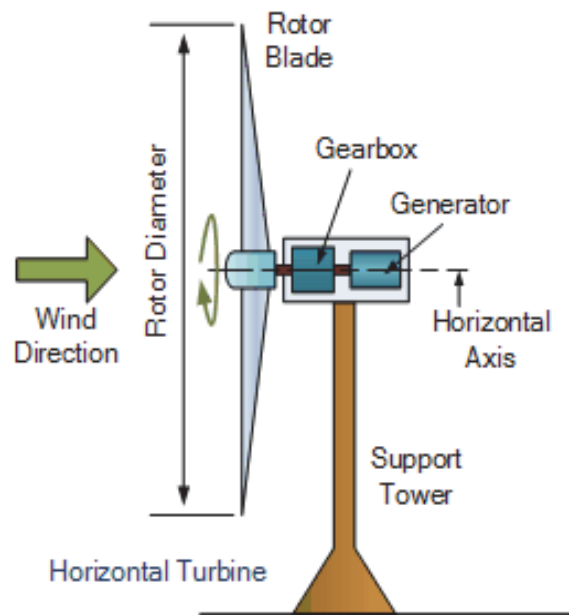
■ Multi-source integration

- Characterises prototypical behaviour across multiple data sources
- Allows to inspect behaviour within and across data sources

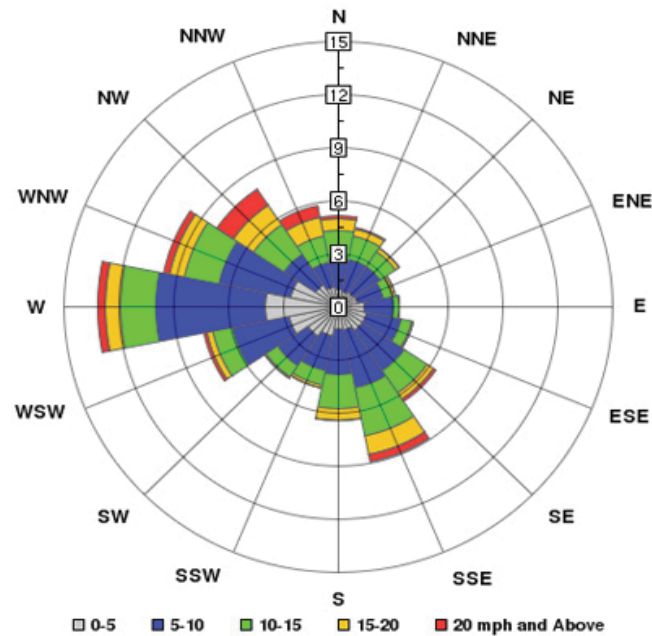


Understanding turbine performance

Wind Speed



Wind Direction

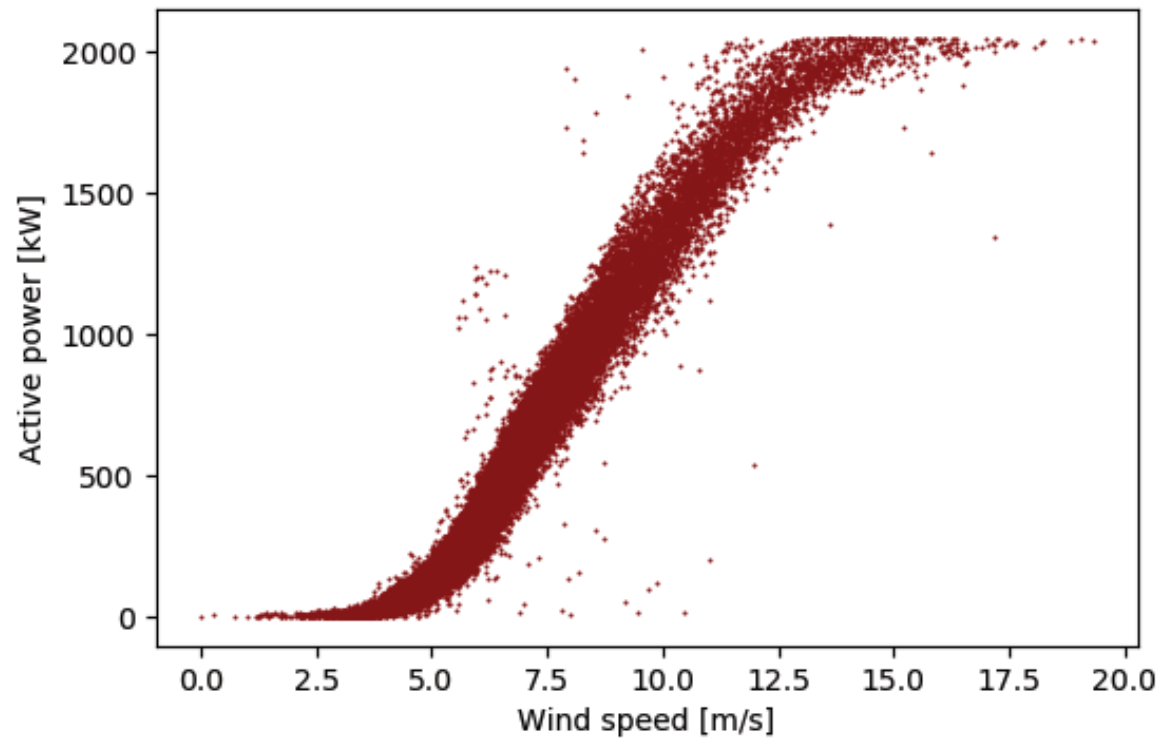


Temperature



Power curve of a wind turbine

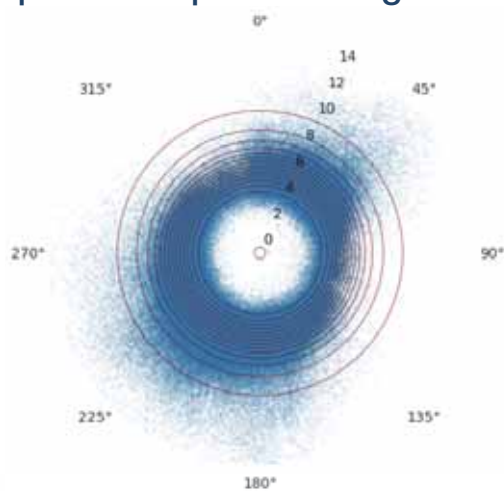
The MEASURE for wind turbine performance



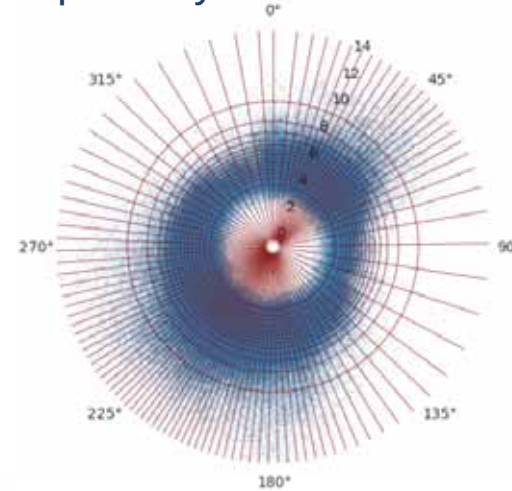
Performance-related view

Two-dimensional circular binning

- Considers the key parameters impacting production performance:
 - wind speed and wind direction
- Derives relatively homogeneous two-dimensional windows of operating conditions
- Benefits from the fact that wind direction is an angular variable
- Two subsequent steps treating each dimension separately:



1) wind speed bins: concentric circles

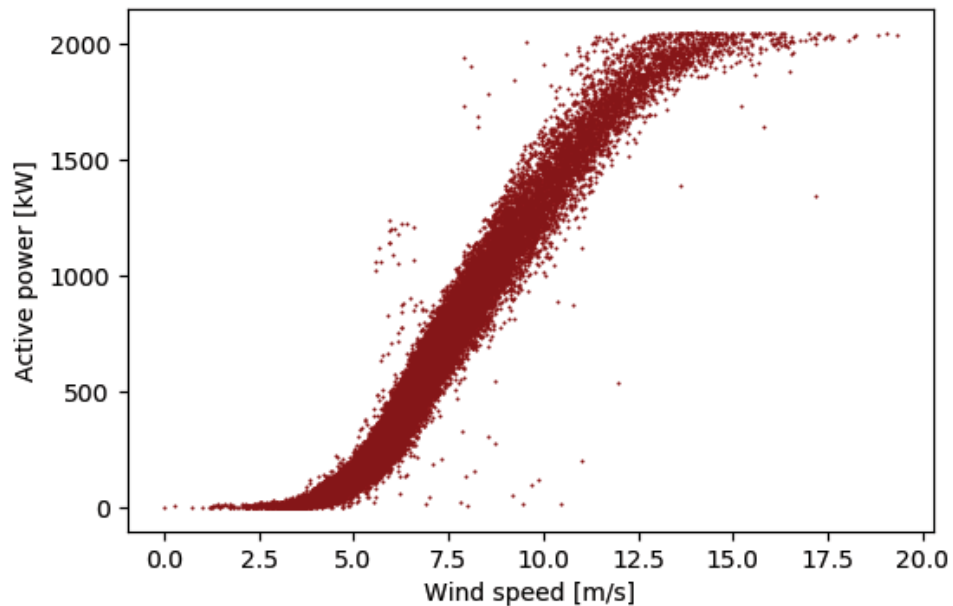


2) wind direction bins: sectors

Performance-related view

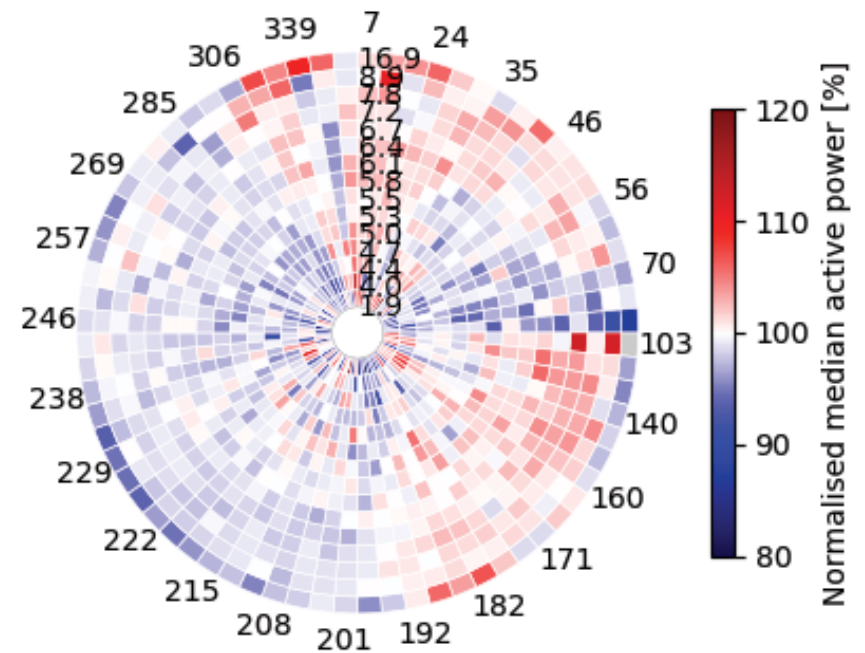
Power curve vs. circular power map

POWER CURVE



Performance expressed as a function of *only* wind speed

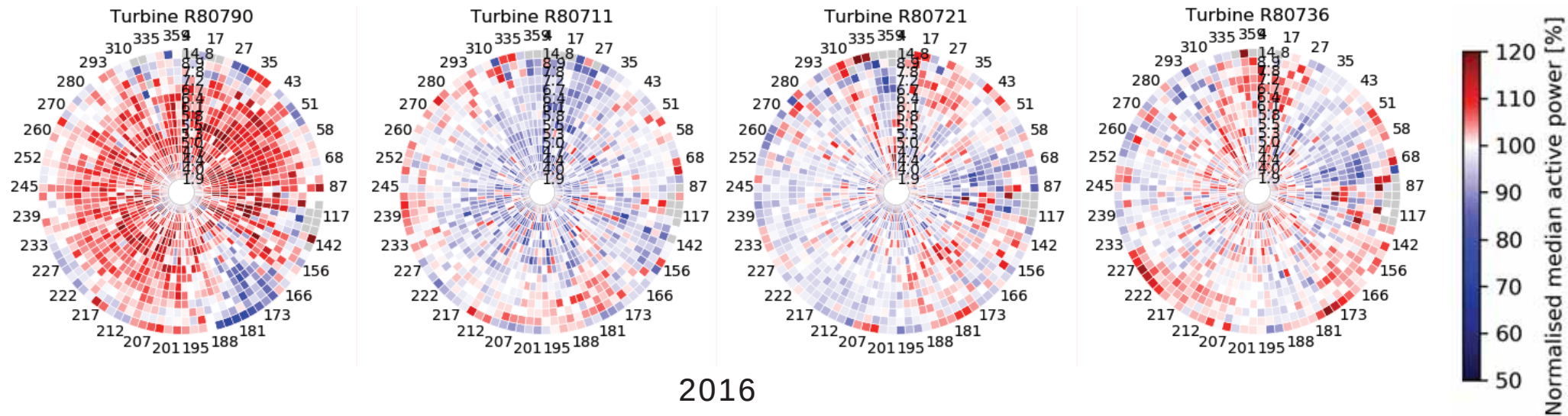
CIRCULAR POWER MAP



Mean production rates for different combinations of wind speed vs. wind direction

Performance-related view

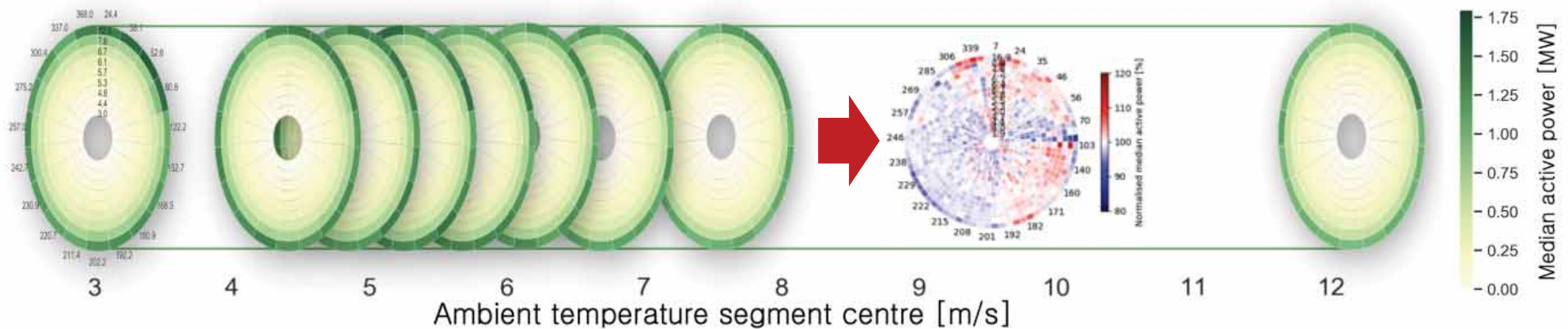
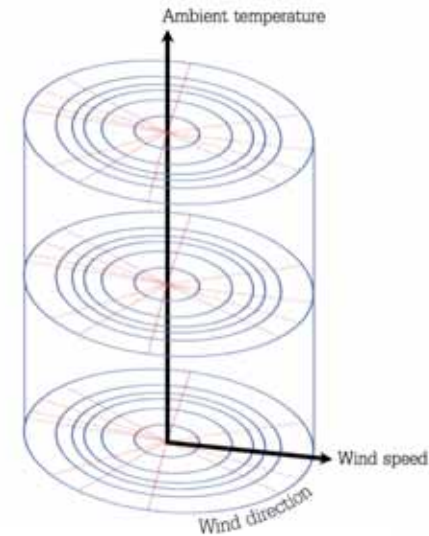
Fleet-wide performance comparison through circular power maps



Performance-related view

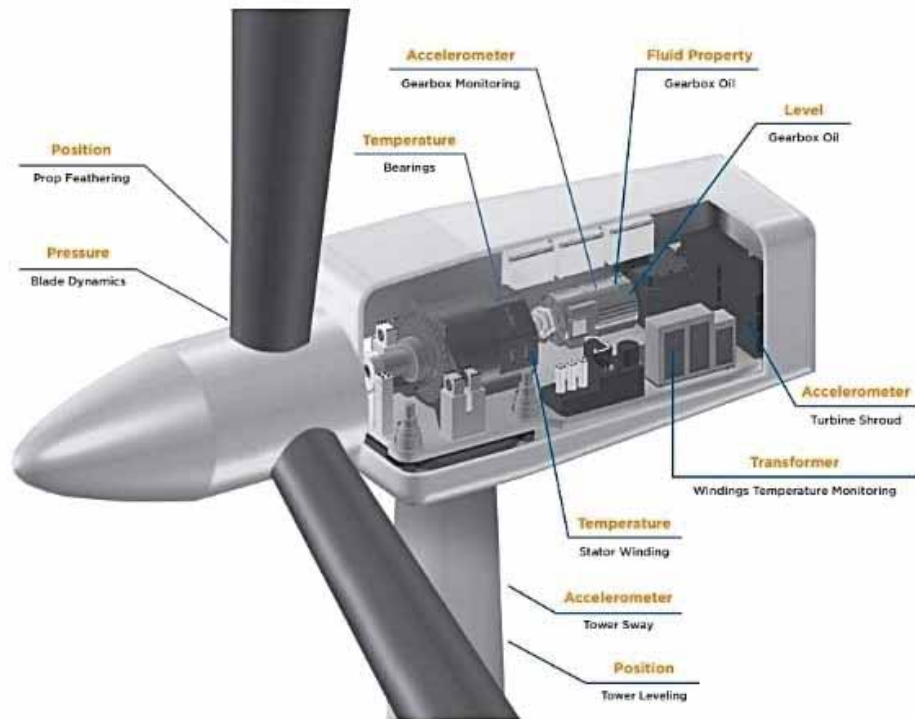
3D cylindrical binning

Adding the third exogeneous factor: ambient temperature

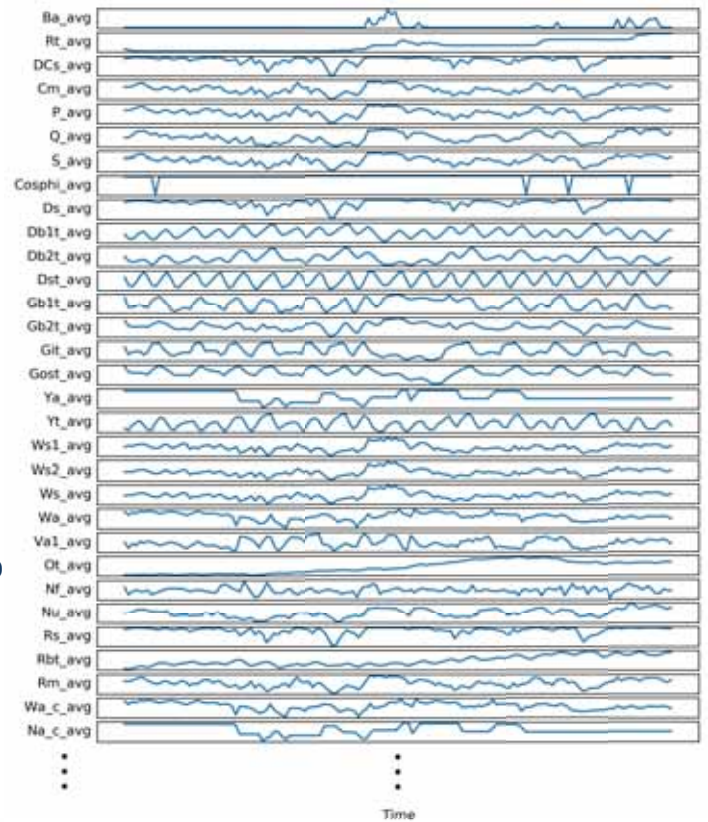


Operation-related view

The internal working of the turbine

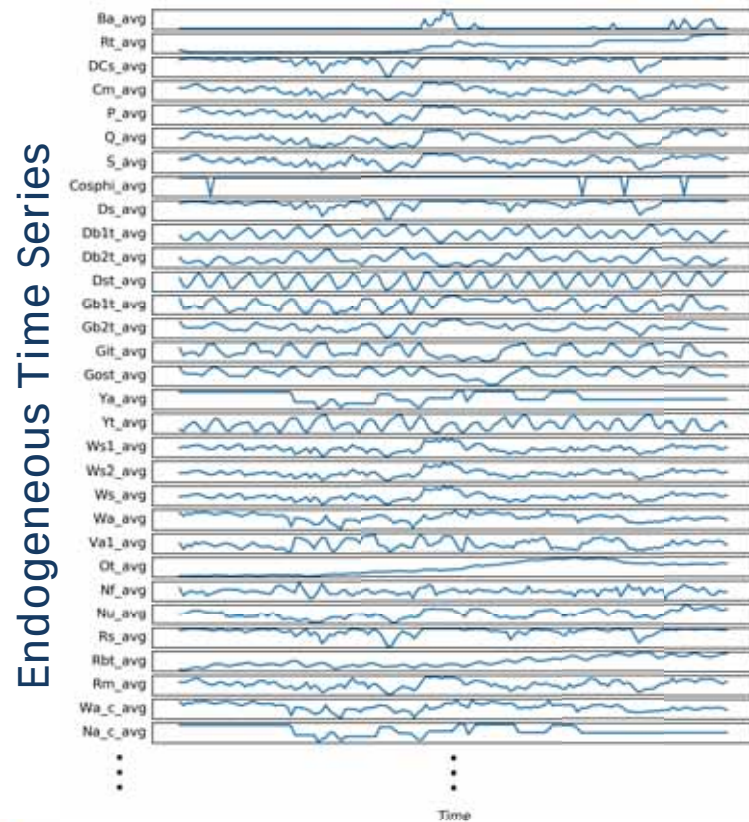


Endogeneous Time Series



Operation-related view

Operating mode identification

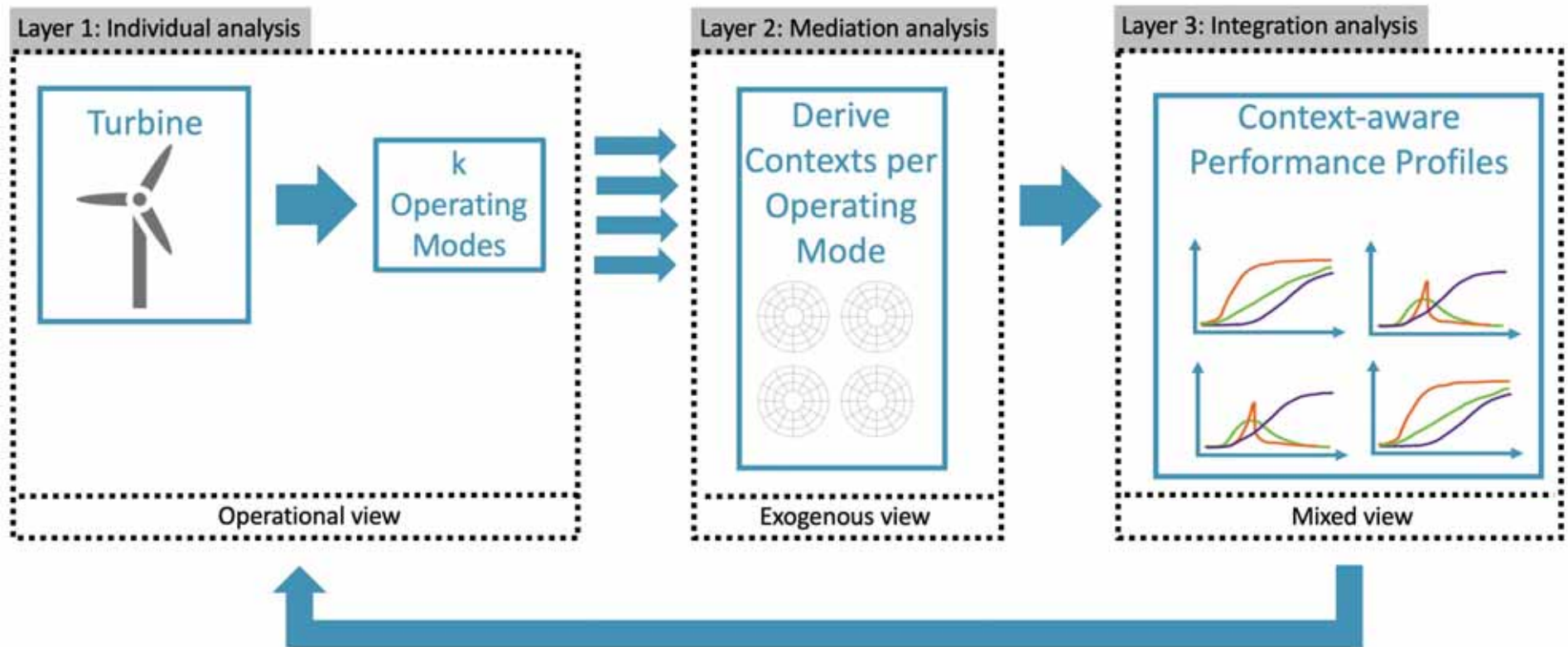


- Group timestamps for which the values of the endogeneous parameters relate to each other in a similar way
- Define a range of allowable values for each operational parameter
- Generate parametric characterisation of the operating modes



Operating Modes

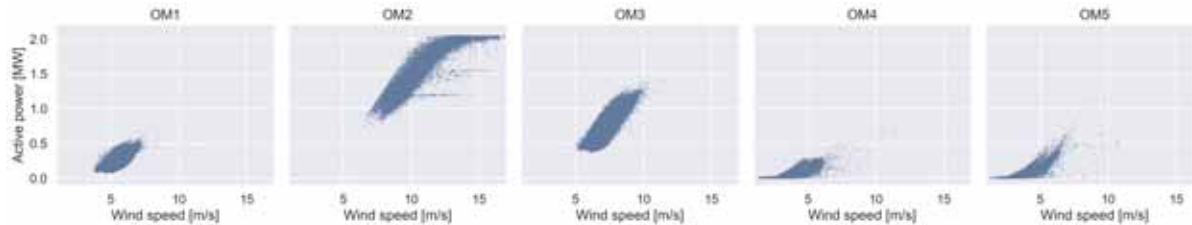
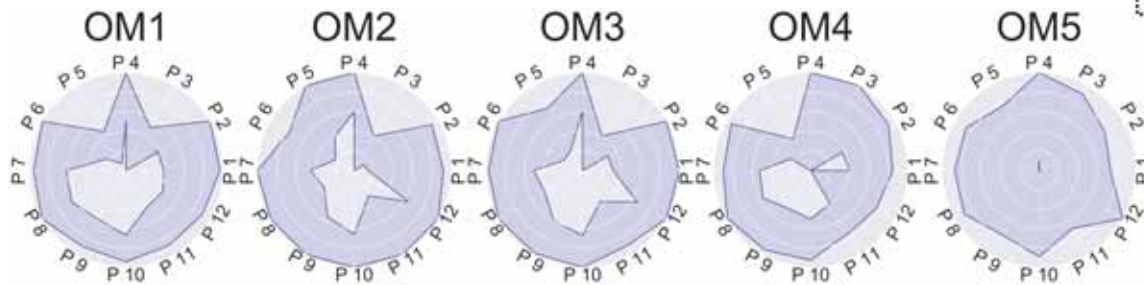
Context-aware profiling



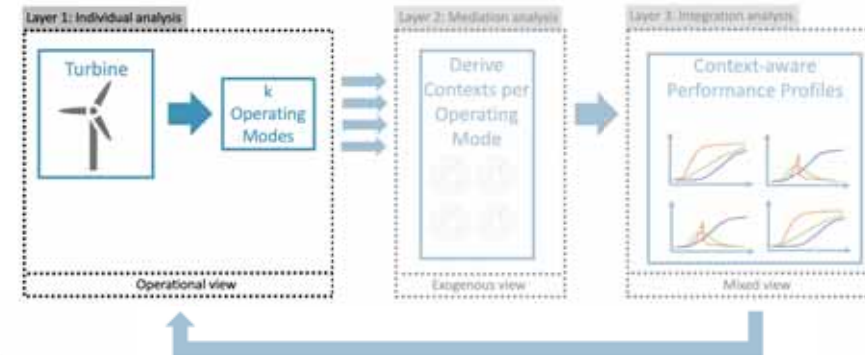
Context-aware profiling

Individual Analysis Layer

Individual operating modes



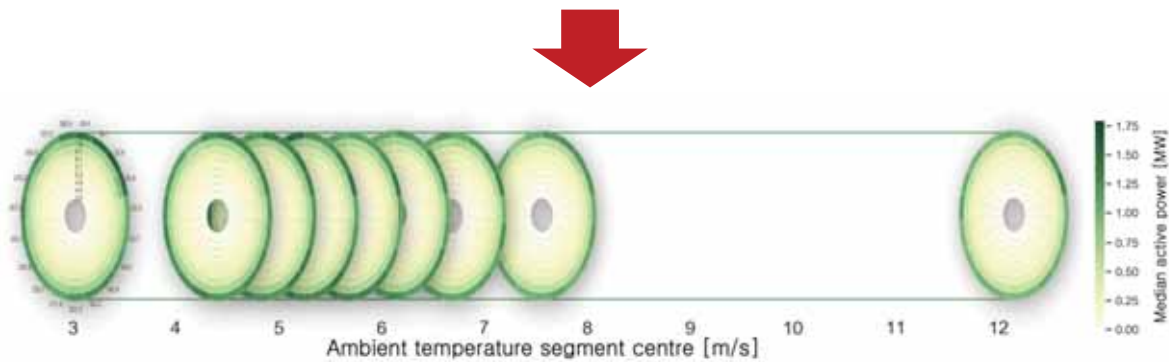
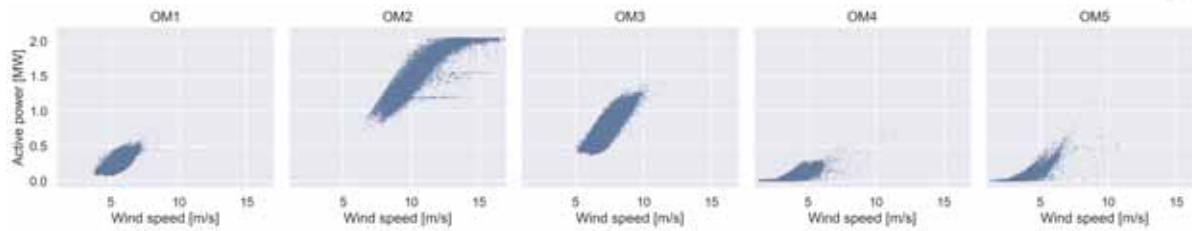
Performance profiles expressed via power curves



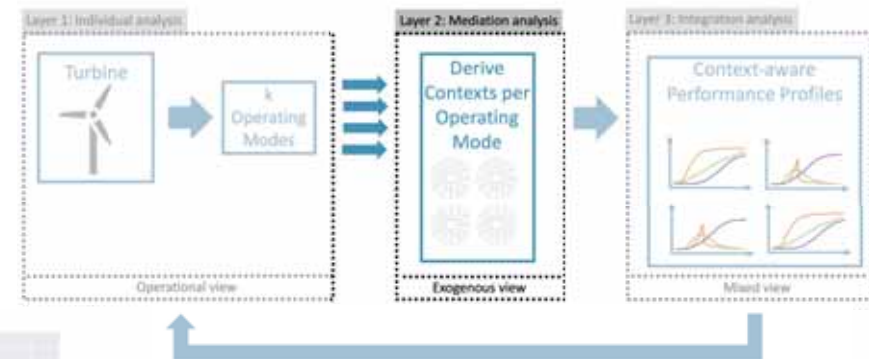
Context-aware profiling

Mediation Analysis Layer

Performance profiles expressed via power curves

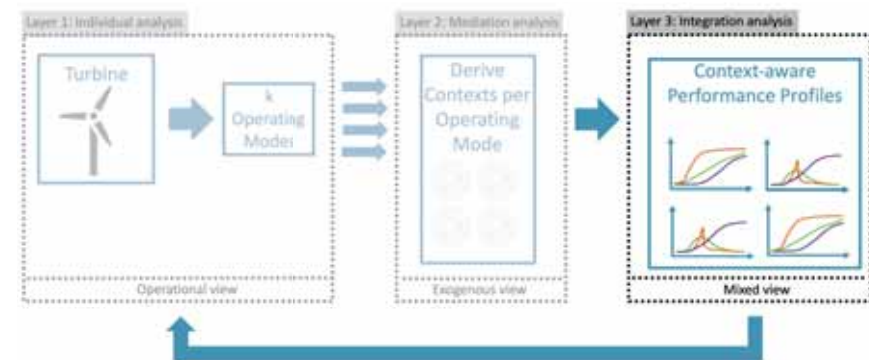


3D (cylindrical) binning per operating mode



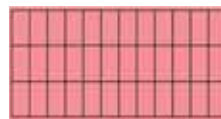
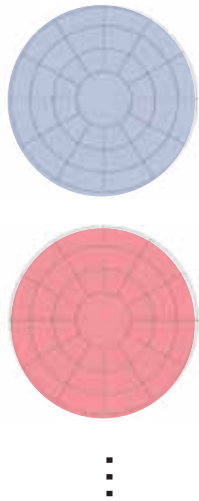
Context-aware profiling

Integration Analysis Layer



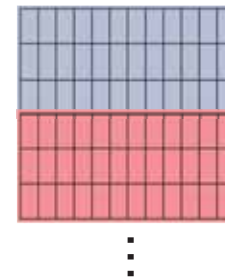
Construct performance matrices using the cylindrical binning segmentation

1. Unroll each power map (cylindrical bin) into a matrix



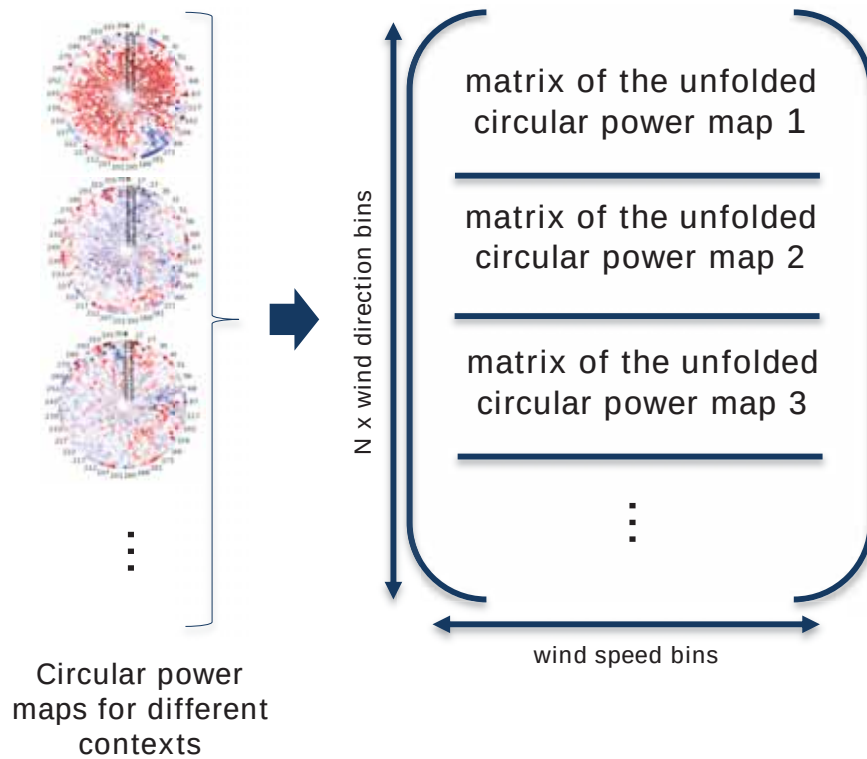
⋮

2. Obtain an overall performance matrix by stacking all the individual power map matrices vertically



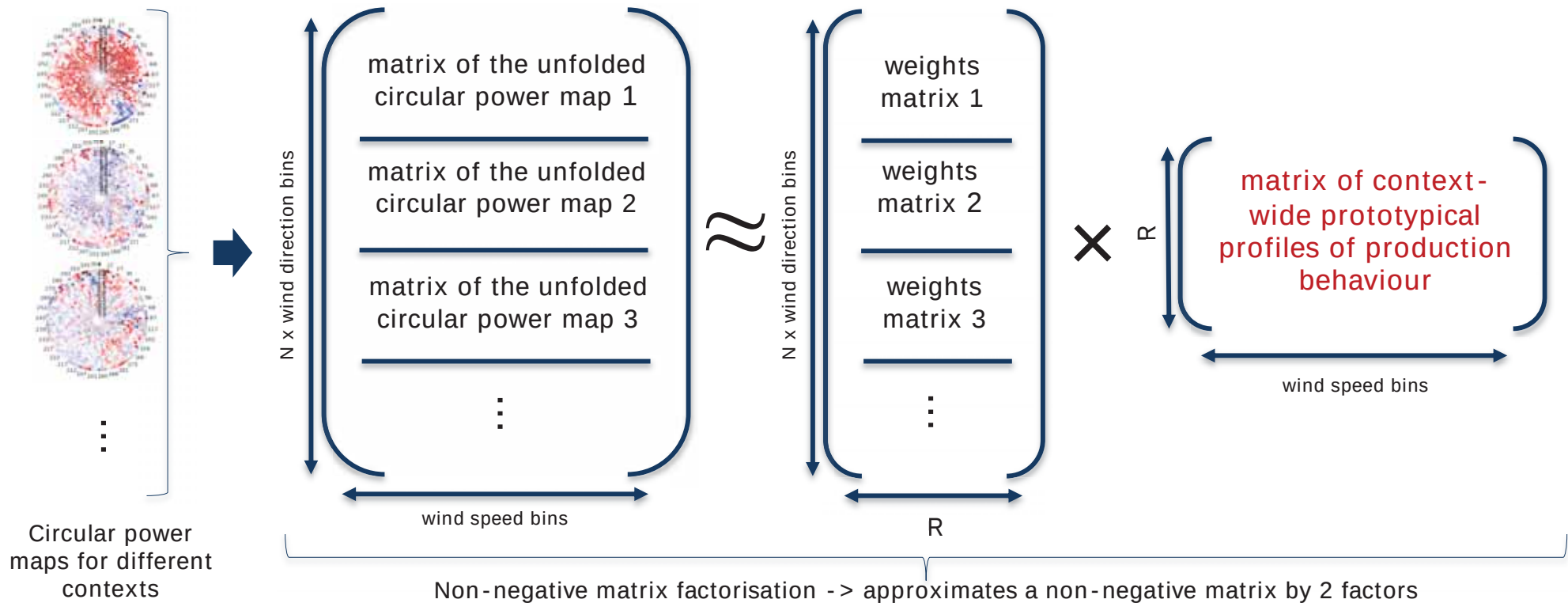
Context-aware profiling

Prototypical profiling via Non-negative Matrix Factorisation (NMF)



Context-aware profiling

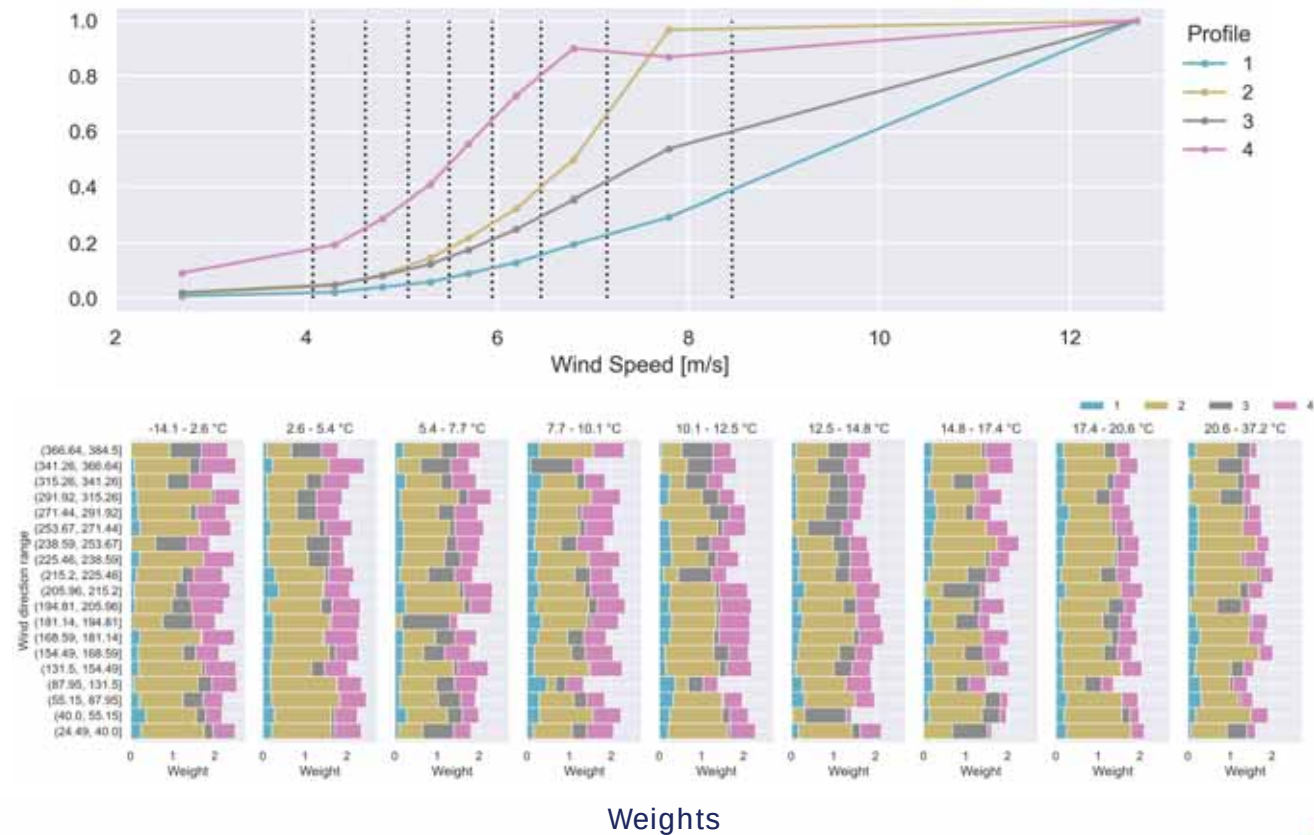
Prototypical profiling via Non-negative Matrix Factorisation (NMF)



Context-aware profiling

Performance profiles vs. environmental context

Normalised prototypical performance profiles



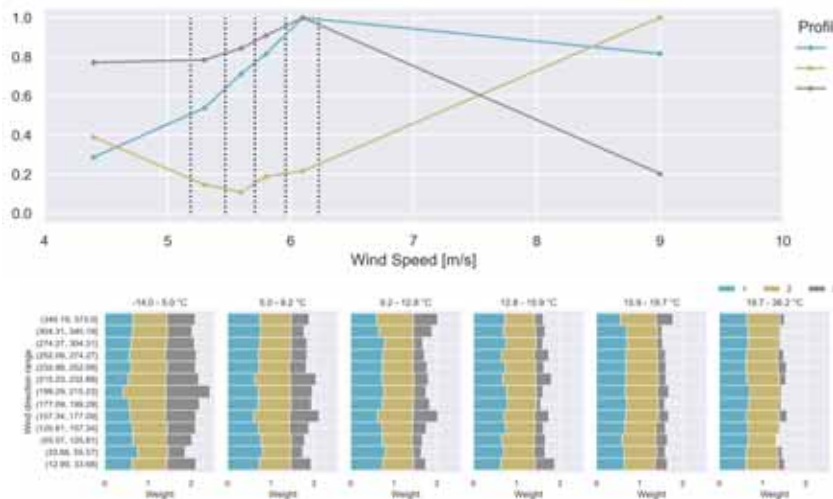
Context-aware profiling

Comparative analysis across the operating modes

OM1: low wind speeds, hardly any production

- profile 2 is the most dominant one
- profile 3 disappears for higher temperatures
- ...

Normalised prototypical performance profiles

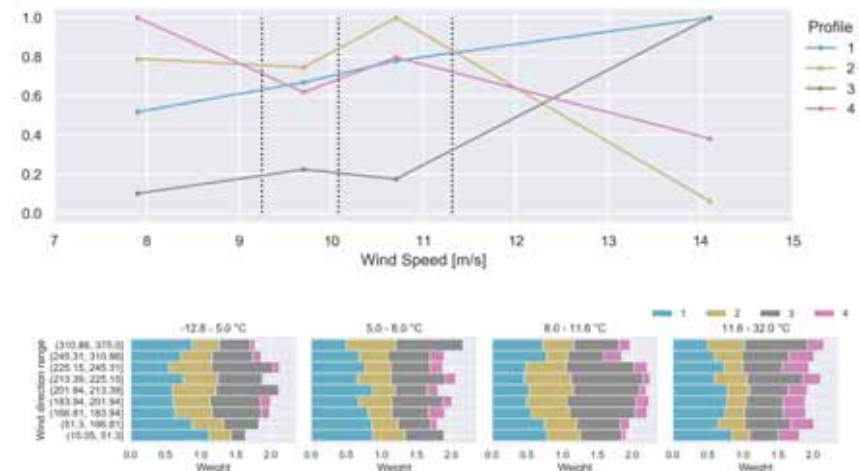


Weights

OM2: an optimal wind flow, high production

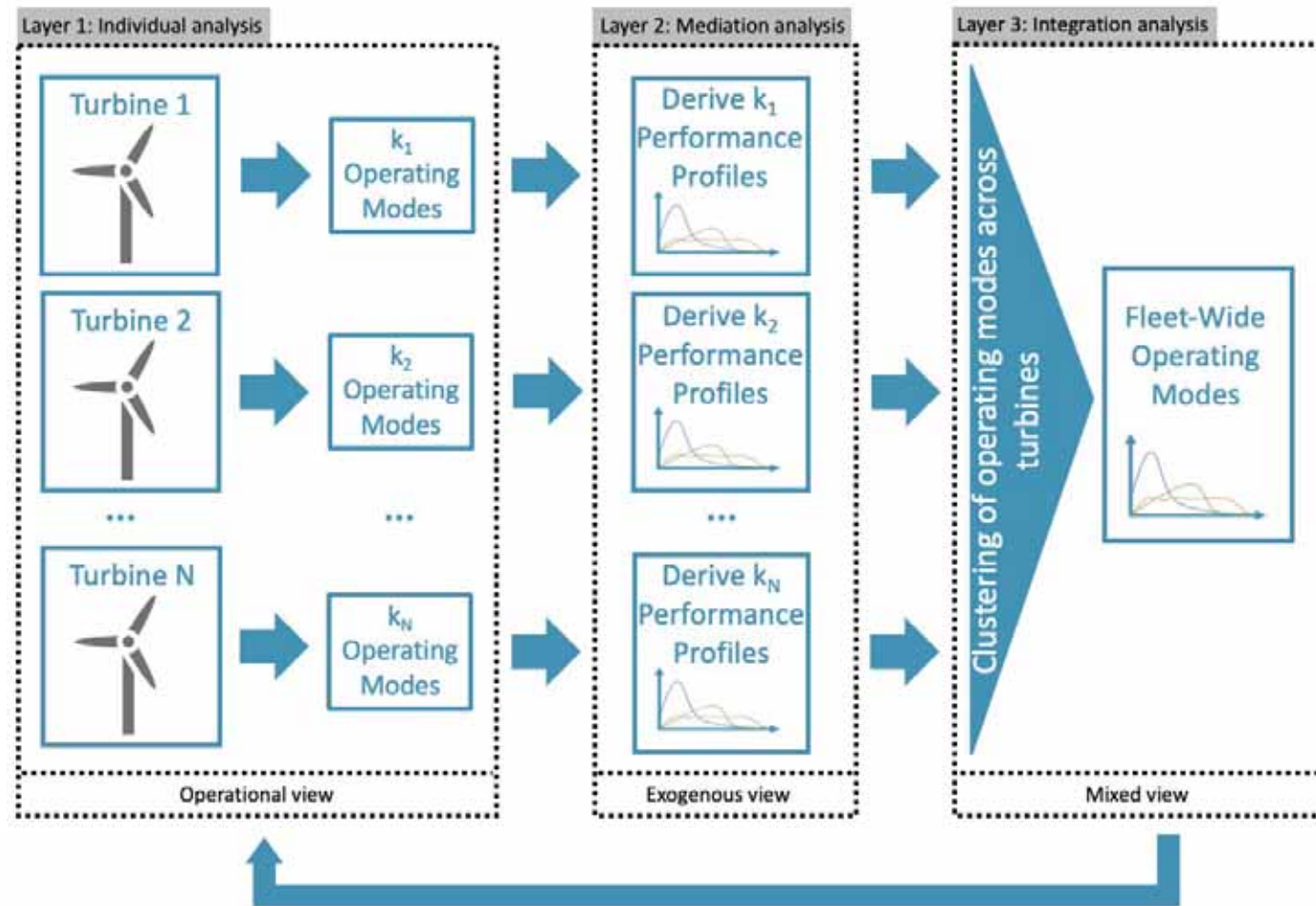
- profiles 2 and 4 are active in different temperature ranges
- ...

Normalised prototypical performance profiles



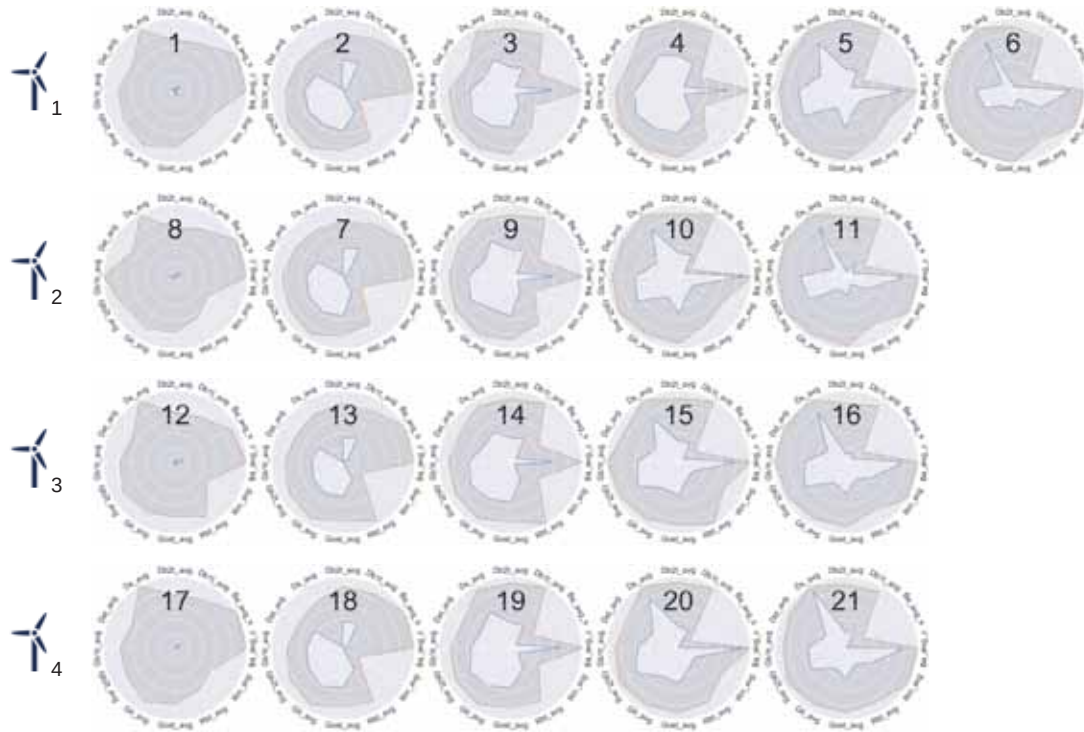
Weights

Multi-source integration

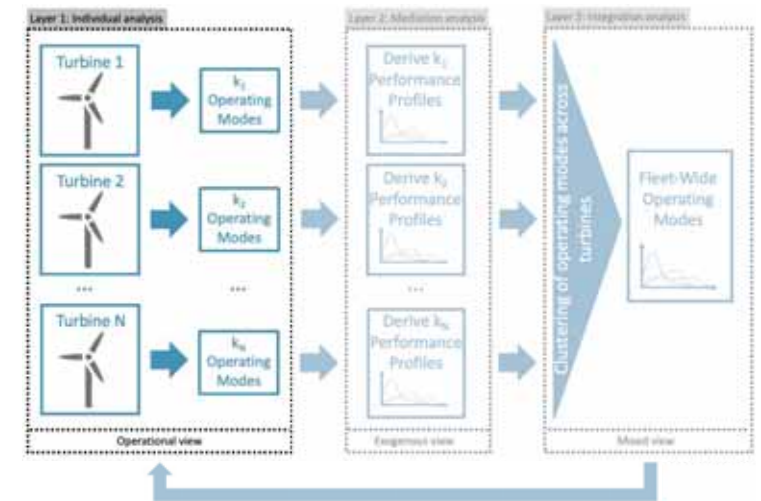


Multi-source integration

Individual Analysis Layer



Individual operating modes



Multi-source integration

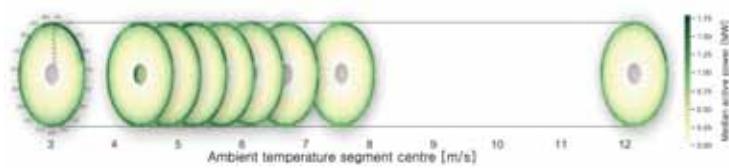
Mediation Analysis Layer

Derive an alternative representation for each operating mode in terms of expected performance

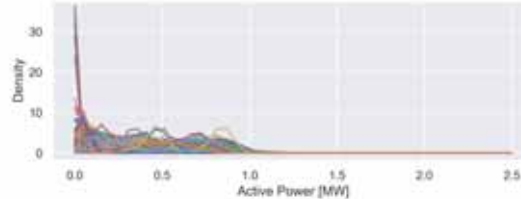
Operating Mode



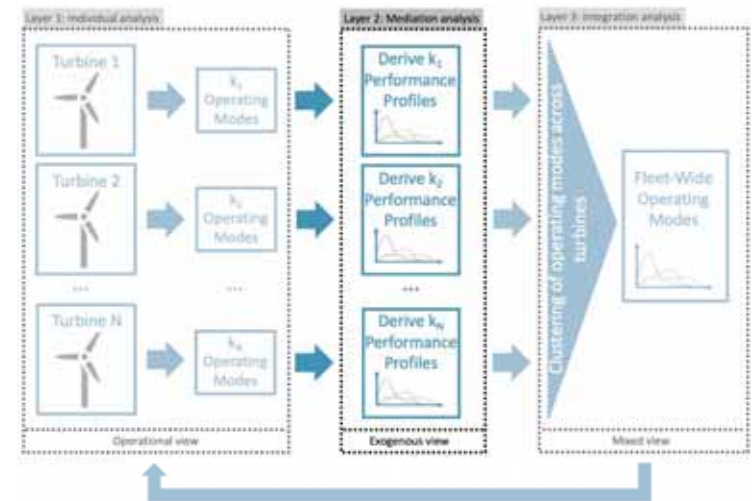
3D (cylindrical) binning per operating mode



Individual probability distributions



Mixture probability distribution

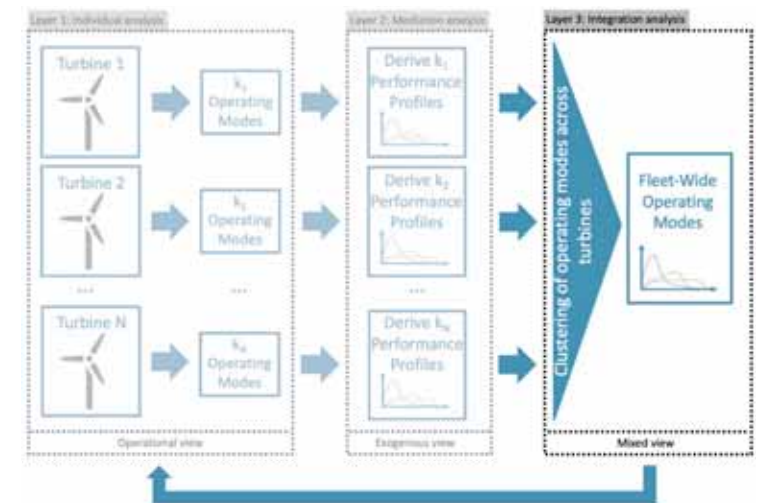
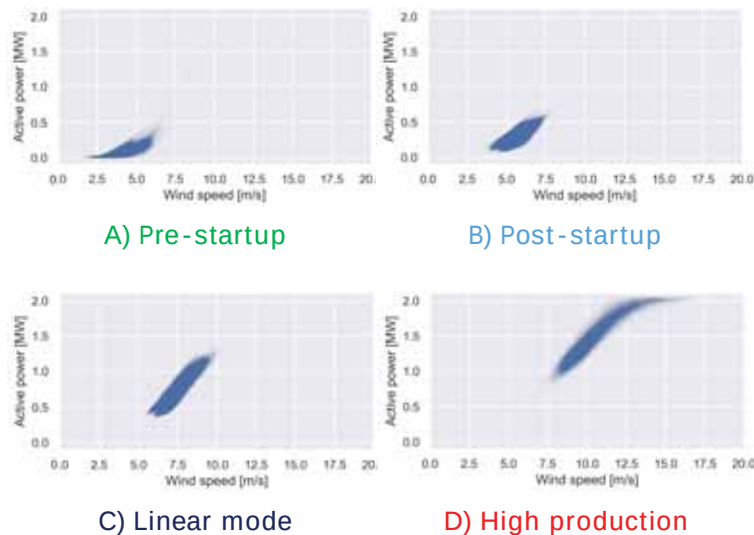


Multi-source integration

Integration Analysis Layer

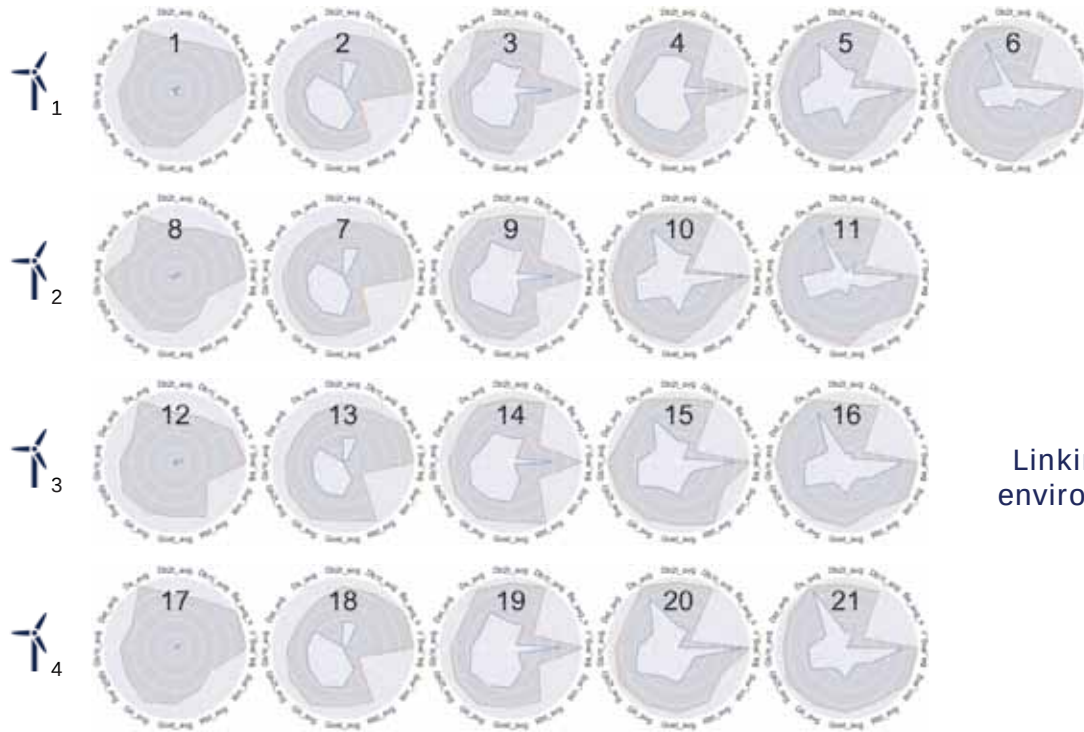
Derive fleet-wide performance profiles by leveraging the results from the previous layers

- Pool together the obtained performance profiles (mixture distributions) per individual operating mode
- Derive *fleet-wide (composite) performance profiles* by subjecting the pool of profiles to some *clustering*



Multi-source integration

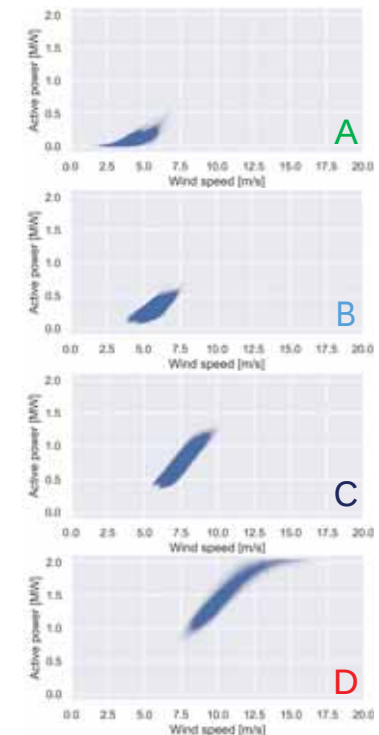
Individual operating modes vs. fleet-wide performance profiles



Individual operating modes



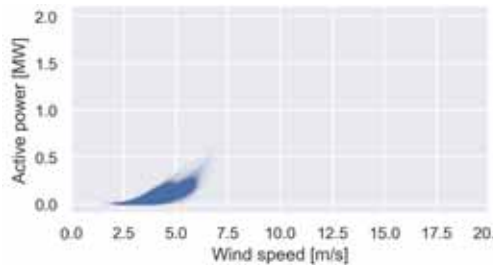
Linking operational with
environmental parameters



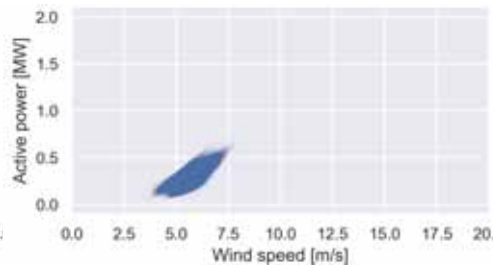
Fleet-wide performance profiles

Multi-source integration

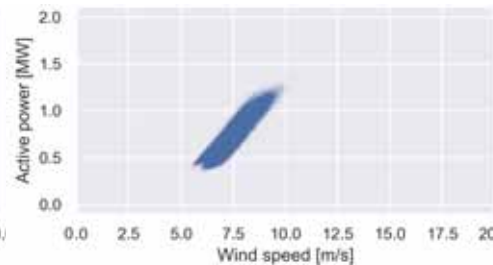
Converting the fleet data into a letter code



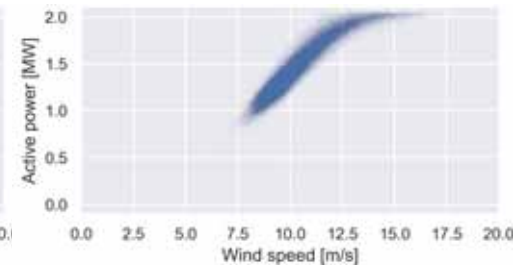
A) Pre-startup



B) Post-startup



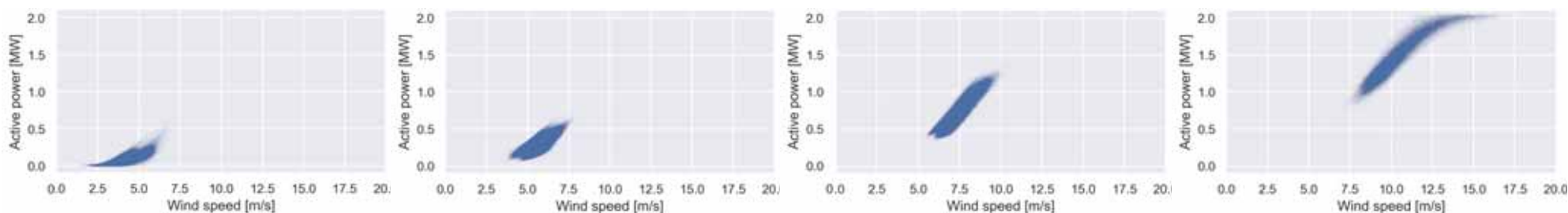
C) Linear mode



D) High production

Multi-source integration

Converting the fleet data into a letter code



A) Pre-startup

B) Post-startup

C) Linear mode

D) High production



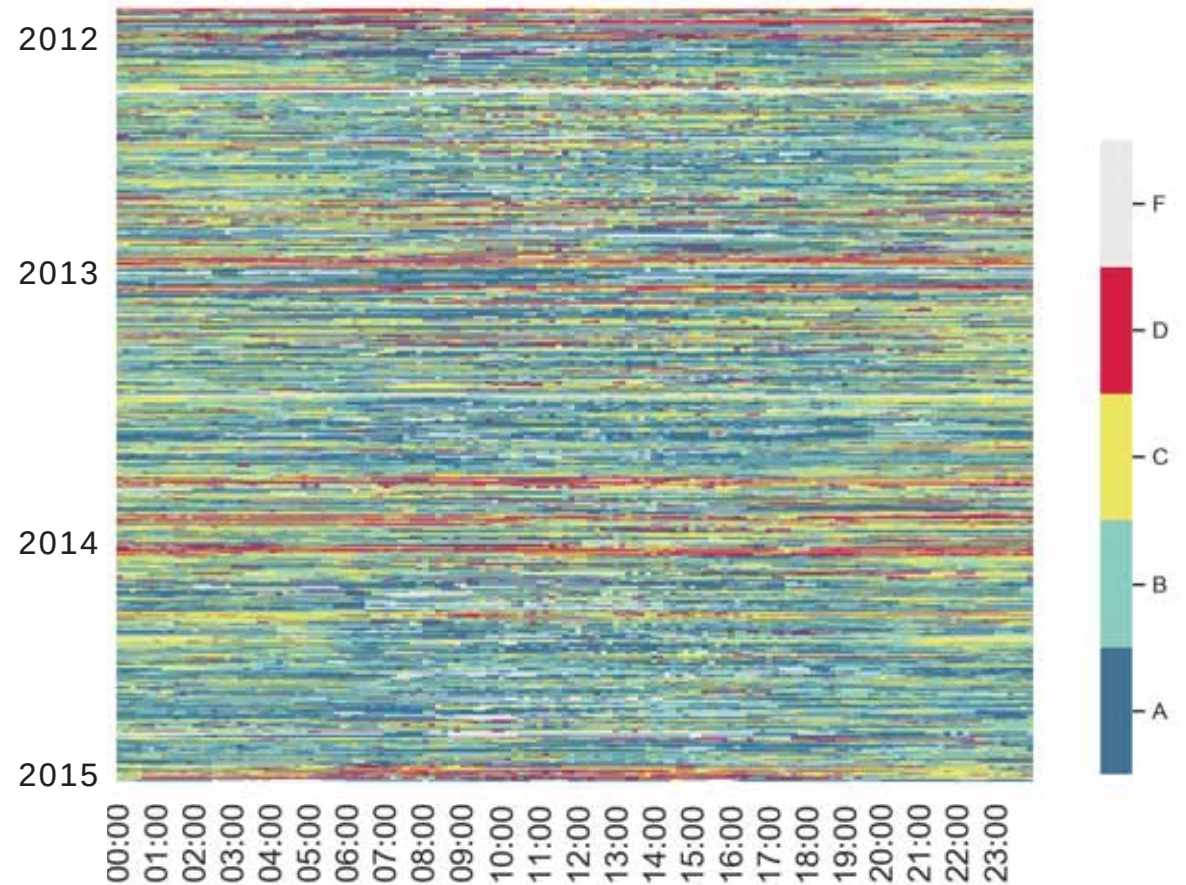
	... D D B D C C D A A A D B D C A D D A D A D D A D D C C C A A C C D D D A C D D C B D C D D D D A A D ...
	... C C A D C A C A A C C A A A C D D A A C C A A A C A A C A A A A A A C A A A D D D C B C B A ...
	... A C A A C C C A C D C C C C A C A C C A D D A C A C A C D C A D D C C C A A A A A A B D D D C ...
	... C A A C D C C A C C C D C D D A C D B C A C B C B D D C C A C D A A C D C D A C C A D D B A A A D A A A ...

Letter Code -> turbine DNA

Multi-source integration

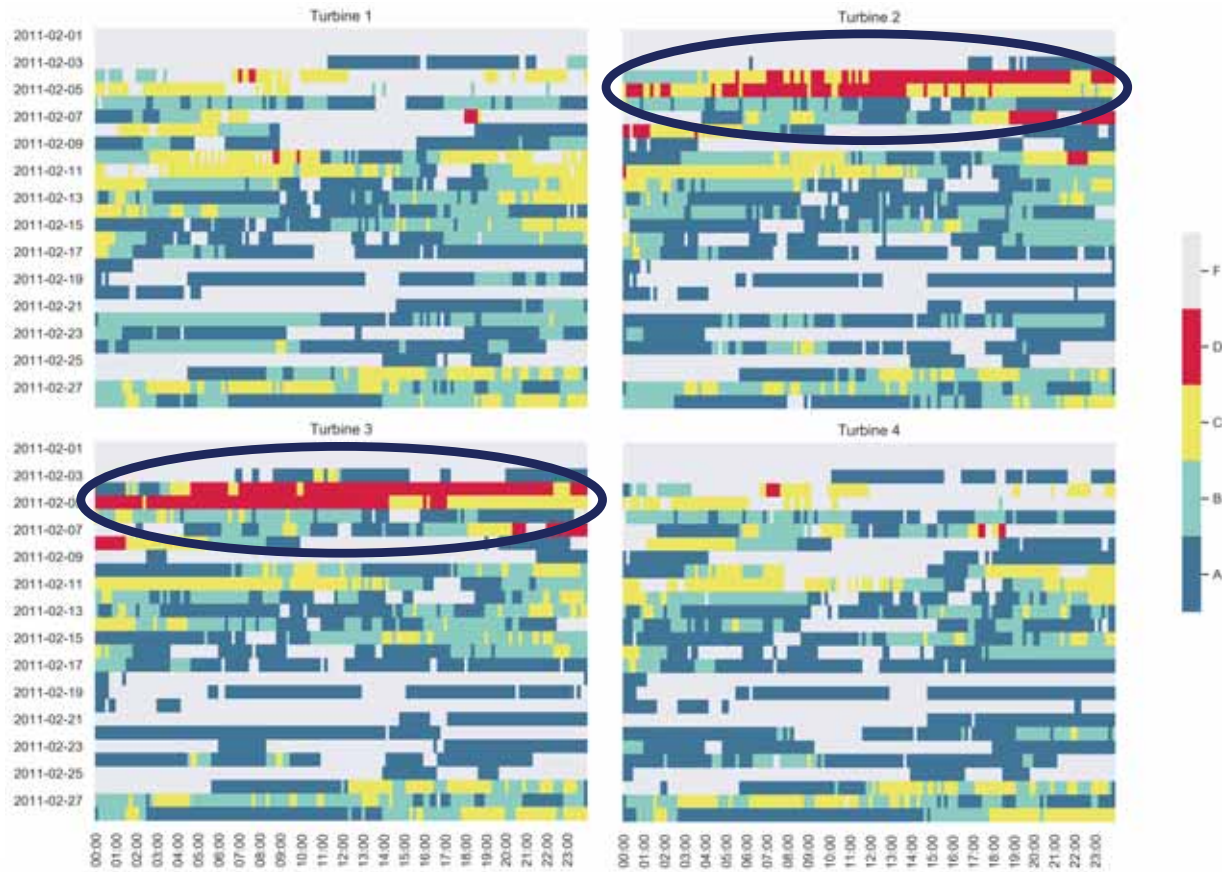
Temporal behaviour characterisation through label maps

- ✓ Operating mode D (*high production*) occurs frequently during the year transitions
- ✓ Operating mode C (*linear mode*) occurs more often at night



Multi-source integration

Cross-fleet performance comparison through label maps

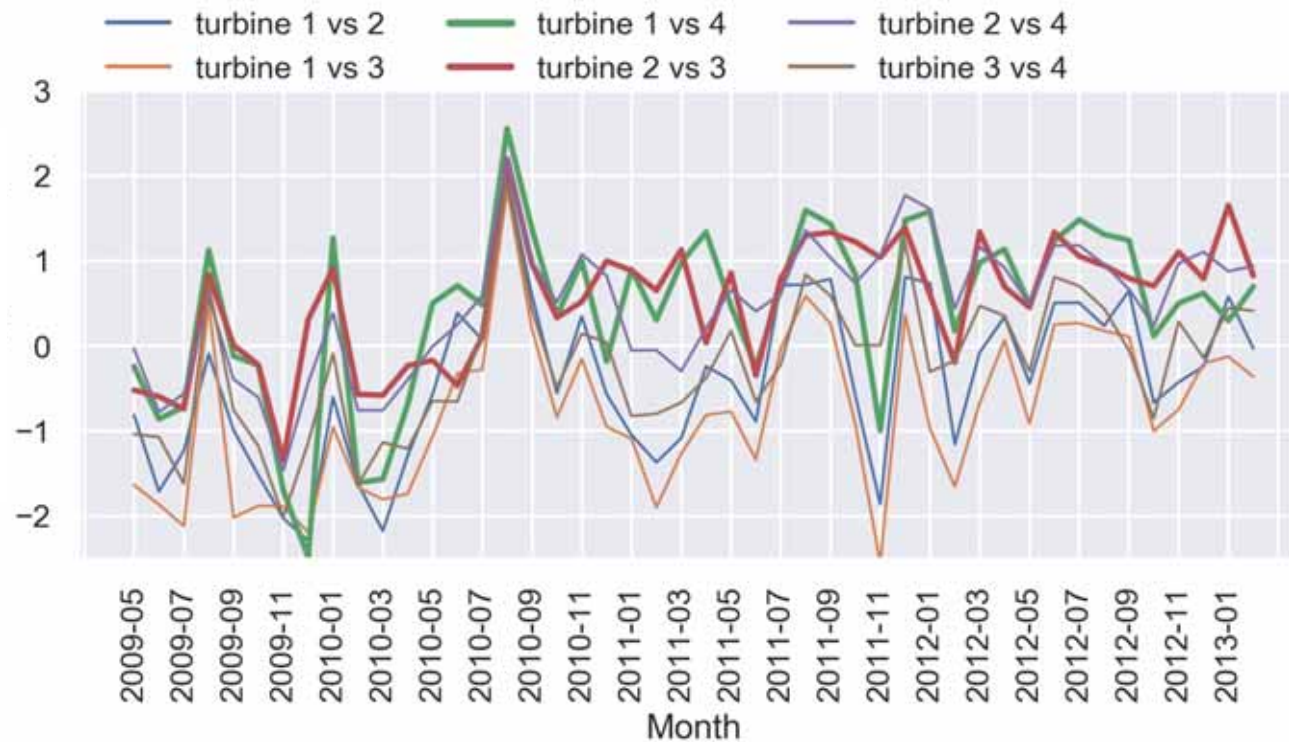


February 2011

Multi-source integration

Estimating similarity between turbines using DNA alignment algorithm

Pairwise sequence alignment scores between the turbines calculated with the Smith-Waterman algorithm using a rolling time window of 1 month



outage turbine 4

Conclusions

Real-world time series data



Created by Vector Portal

- Real-world datasets capture very rich and diverse information in the form of time series about the operation and performance of industrial assets
- However, the inherent multi-source nature of the real-world data makes it challenging to directly integrate the different time series data without information loss
- Well thought data preparation and discretisation are essential prerequisites for the successful application of advanced analytics on time series data

Conclusions

Multi-view analysis of temporal data



Created by Vector Portal

- We demonstrated the potential of a multi-view analysis approach empowering
 - the integration of heterogeneous real-world datasets originating from multiple source
 - the establishment of a direct link between operating mode and prototypical performance
- The proposed methodological workflow is generic in nature
 - can be applied in any context where the available data can be split in operation-related and performance-related views
- Potential use cases
 - Visual inspection for detecting trends
 - Continuous anomaly detection
 - Production forecasting
 - Process optimisation by understanding the influence of different contexts

References

- M. Dhont, E. Tsiorkova & V. Boeva. **Layered Integration Approach for Multi-view Analysis of Temporal Data.** *In: Lemaire V., Malinowski S., Bagnall A., Guyet T., Tavenard R., Ifrim G. (eds) Advanced Analytics and Learning on Temporal Data. AALTD 2020. Lecture Notes in Computer Science, vol 12588. Springer, 2020.*
- M. Dhont, E. Tsiorkova, T. Tourwé & N. González -Deleito. **Visual Analytics for Extracting Trends from Spatio-temporal Data.** *In: Lemaire V., Malinowski S., Bagnall A., Guyet T., Tavenard R., Ifrim G. (eds) Advanced Analytics and Learning on Temporal Data. AALTD 2020. Lecture Notes in Computer Science, vol 12588. Springer, 2020.*
- M. Dhont, E. Tsiorkova & N. González -Deleito. **Deriving Spatio-temporal Trajectory Fingerprints from Mobility Data using Non-Negative Matrix Factorisation.** *16th International Workshop on Spatial and Spatiotemporal Data Mining (SSTD-21), IEEE International Conference on Data Mining (ICDM).*
- M. Dhont, E. Tsiorkova & V. Boeva. **Performance Profiling of Operating Modes via Multi-View Analysis Using Non-Negative Matrix Factorisation.** *Springer-Verlag, Recent Advancements in Multi-View Data Analytics, 2021.*
- M. Dhont, E. Tsiorkova & V. Boeva. **Advanced Discretisation and Visualisation Methods for Performance Profiling of Wind Turbines.** *Energies, special issue: Wind Turbine Monitoring through Operation Data Analysis, 2021.*

My co-authors



Michiel Dhont
Data scientist & PhD student
Sirris & VUB



Veselka Boeva
Professor of Computer Science
Blekinge Institute of Technology
Sweden