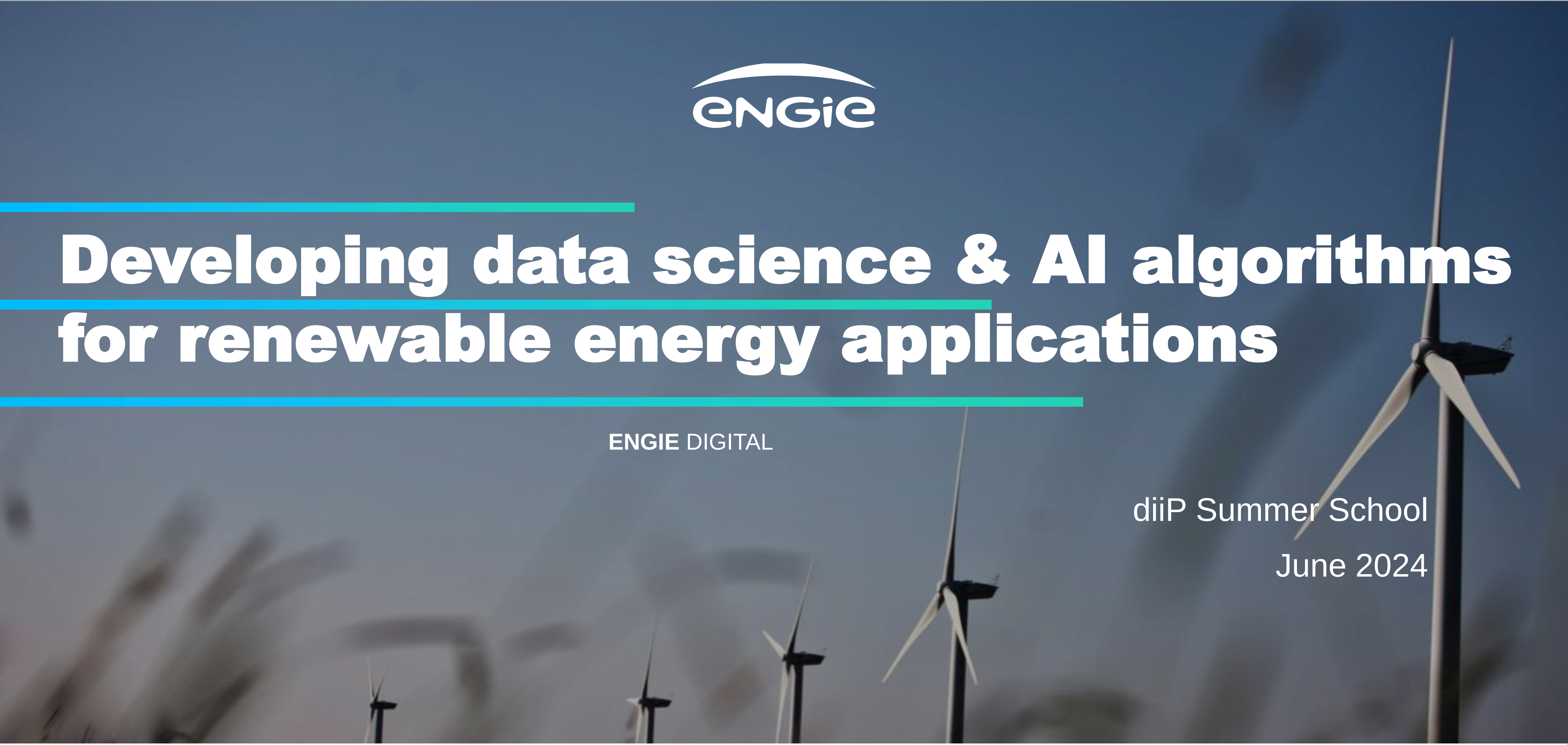


# Developing data science & AI algorithms for renewable energy applications

**ENGIE** DIGITAL

diiP Summer School

June 2024

RESTREINT | INTERNE | SECRET

# Who am I?

https://www.linkedin.com/in/paulponcet/

# Topics where data science & AI have proved useful at Engie

**Some business stakes (in a nutshell)**

- Predictive maintenance of industrial equipment
- Short-term forecasting of energy demand and energy production
- Clustering of sites/assets/customers
- Optimization / control of industrial assets
- Data understanding (esp. for unstructured data)
- Data quality / Data cleaning
- Losses and gains assessment
- Content classification
- Multi-agent systems in a GenAI context

# Topics where data science & AI have proved useful at Engie

**Some business stakes (in a nutshell)**

- Predictive maintenance of industrial equipment
- Short-term forecasting of energy demand and energy production
- Clustering of sites/assets/customers
- Optimization / control of industrial assets
- Data understanding (esp. for unstructured data)
- Data quality / Data cleaning
- Losses and gains assessment
- Content classification
- Multi-agent systems in a GenAI context

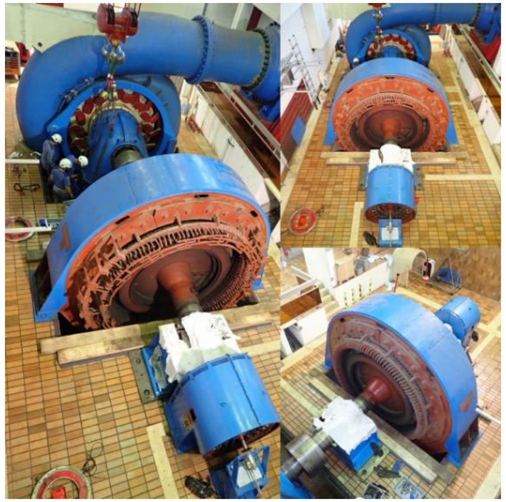
**Some scientific stakes (in a nutshell too)**

- Anomaly detection
- Multi-task learning and dimension reduction
- Causal analysis
- Transfer learning
- Online learning & dynamic models
- Data drift & concept drift detection
- Explanability
- Robust machine learning
- Frugal machine learning
- Multimodal learning
- Math. optimization & reinforcement learning

# Why is Predictive Maintenance at stake within Engie?

# We operate and maintain industrial assets

*Mont de la Grévière wind farm –*
*Engie’s photocenter*

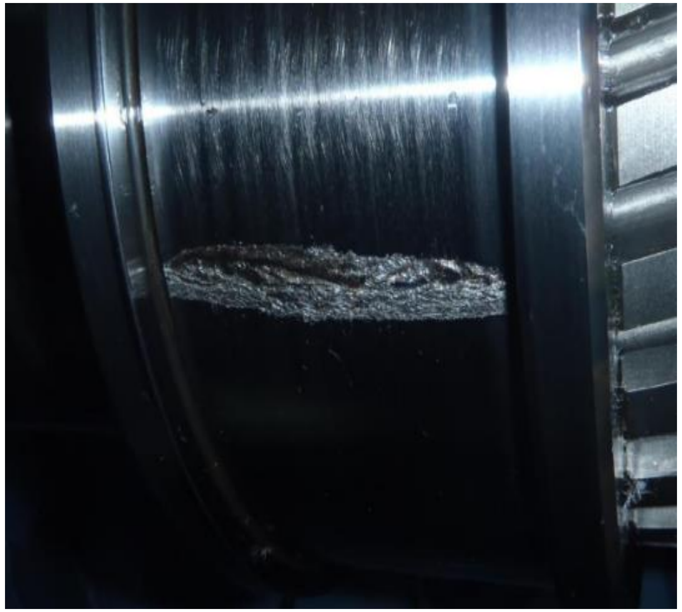


*Hydroelectric turbines –*
*Courtesy of Engie SHEM*

*Charleval PV farm –*
*Engie’s photocenter*

# Theses industrial assets may suffer from wear & tear and/or from abnormal degradation

*A wind turbine with broken blades – https://bit.ly/33JOyAf*



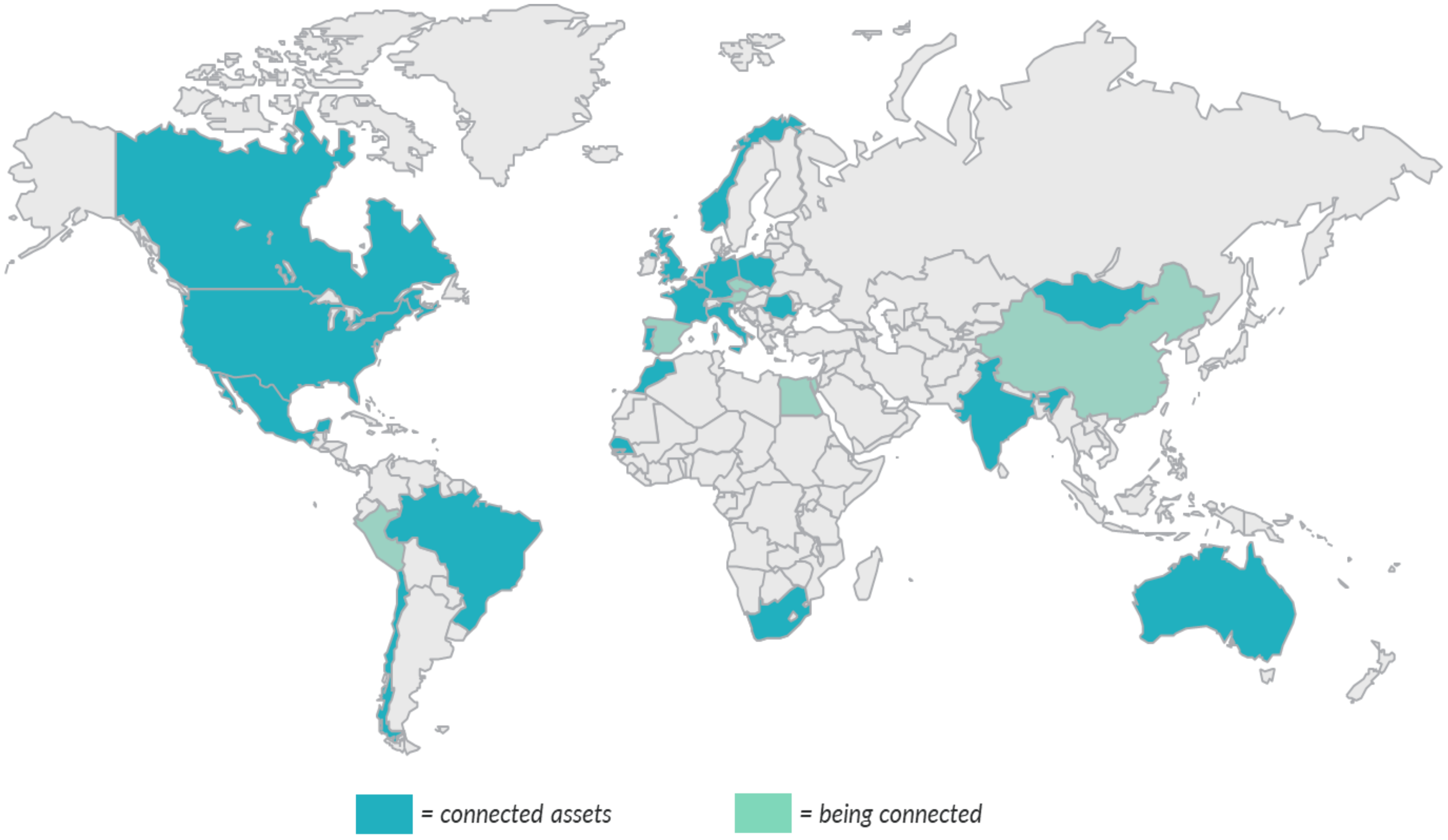
*A damaged bearing on the main shaft of a wind turbine – Courtesy of Engie Green*

# In this context, Engie created DARWIN, the Group software suite dedicated to Renewable Energies

**25** countries connected

**26 GW** monitored

**5** technologies addressed

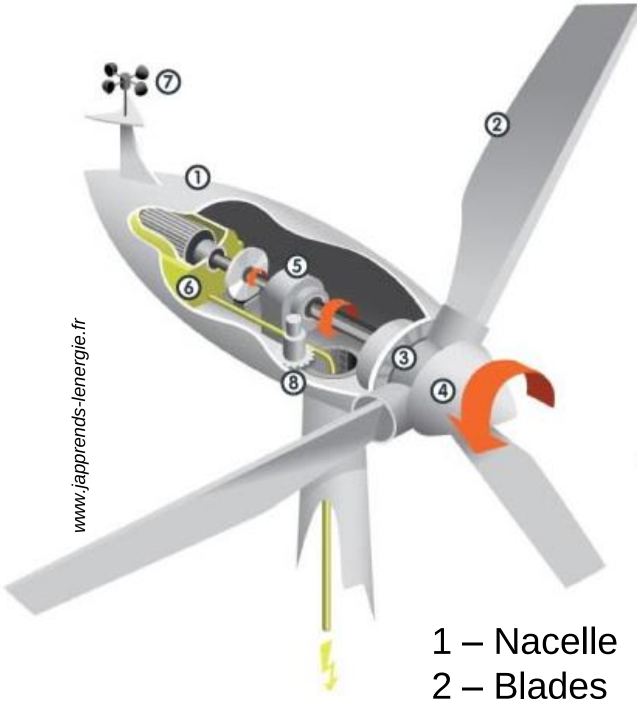

**Darwin monitors Engie's renewable energy assets, to improve their performance, reduce their unavailability, optimize operating costs.**

# As data scientists, we rely on time series acquired by the DARWIN system

We collect time series at the 1 second- and 10 minute- timestamps for each of our wind turbines.

These time series provide us mostly with:

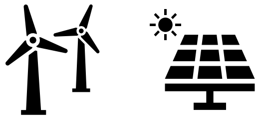
- local meteorological information (wind speed, wind direction, air temperature…)
- mechanical information (component temperatures inside the turbines, rotating speeds…)
- electrical information (active power, current, voltage…)




1 – Nacelle
2 – Blades
3 – Hub
4 – Rotor
5 – Gearbox
6 – Generator
7 – Anemometer and wind vane
8 – Yaw

# We gather a variety of data sources, types, frequencies and contents

**Assets**



400+ wind farms

200+ solar farms

**✪ Majority of timeseries**

**Meteorological, mechanical, electrical** and **control-loop** information

**✪ Different types of data**

**Dynamic, Static** and **Semi-static** data

✪ Variety of **Data sources, Data frequencies** across data sources**, Data contents**

**Key Figures**

**50-100 billion data point** / month

**~70 TB** / month

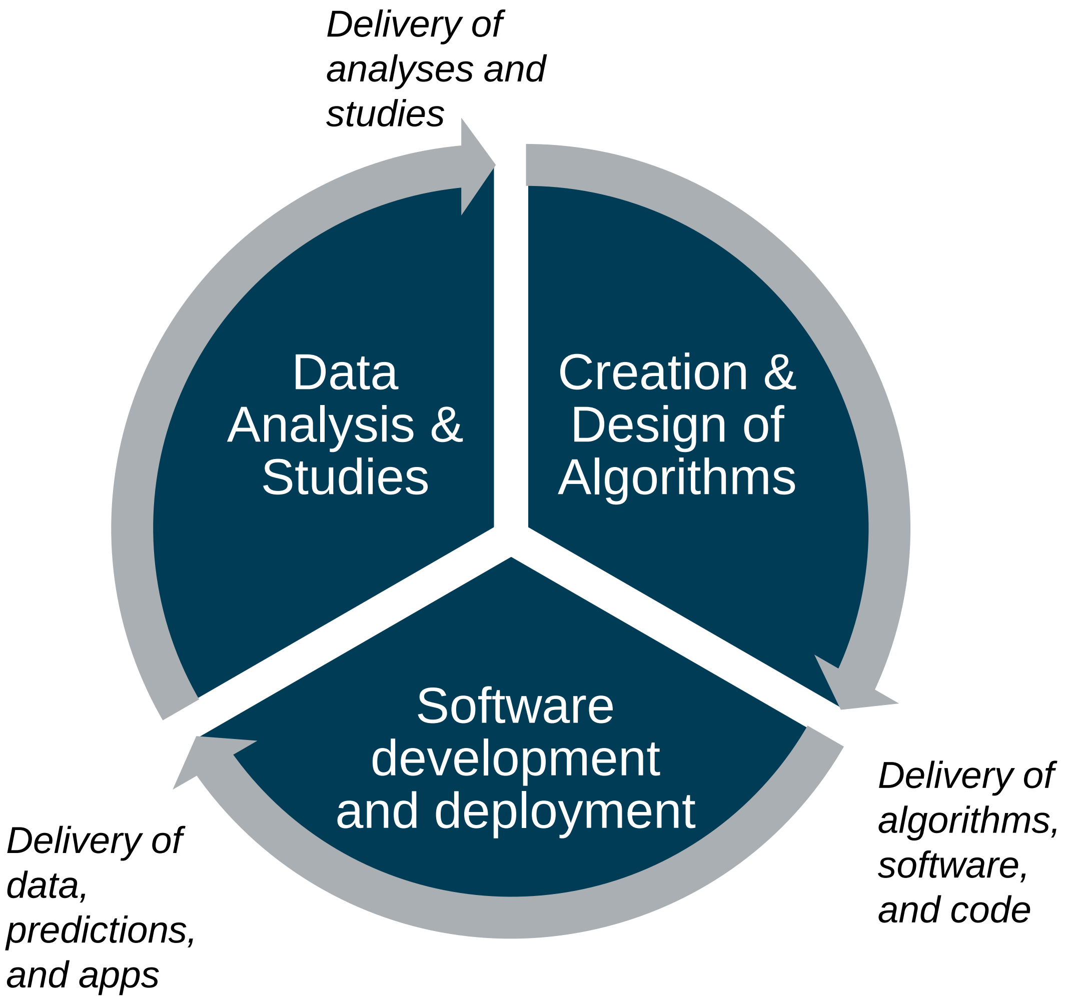
**Data granularity**:
1 sec / 10 min

**Retention period:**
Farm lifetime ~ 20 years

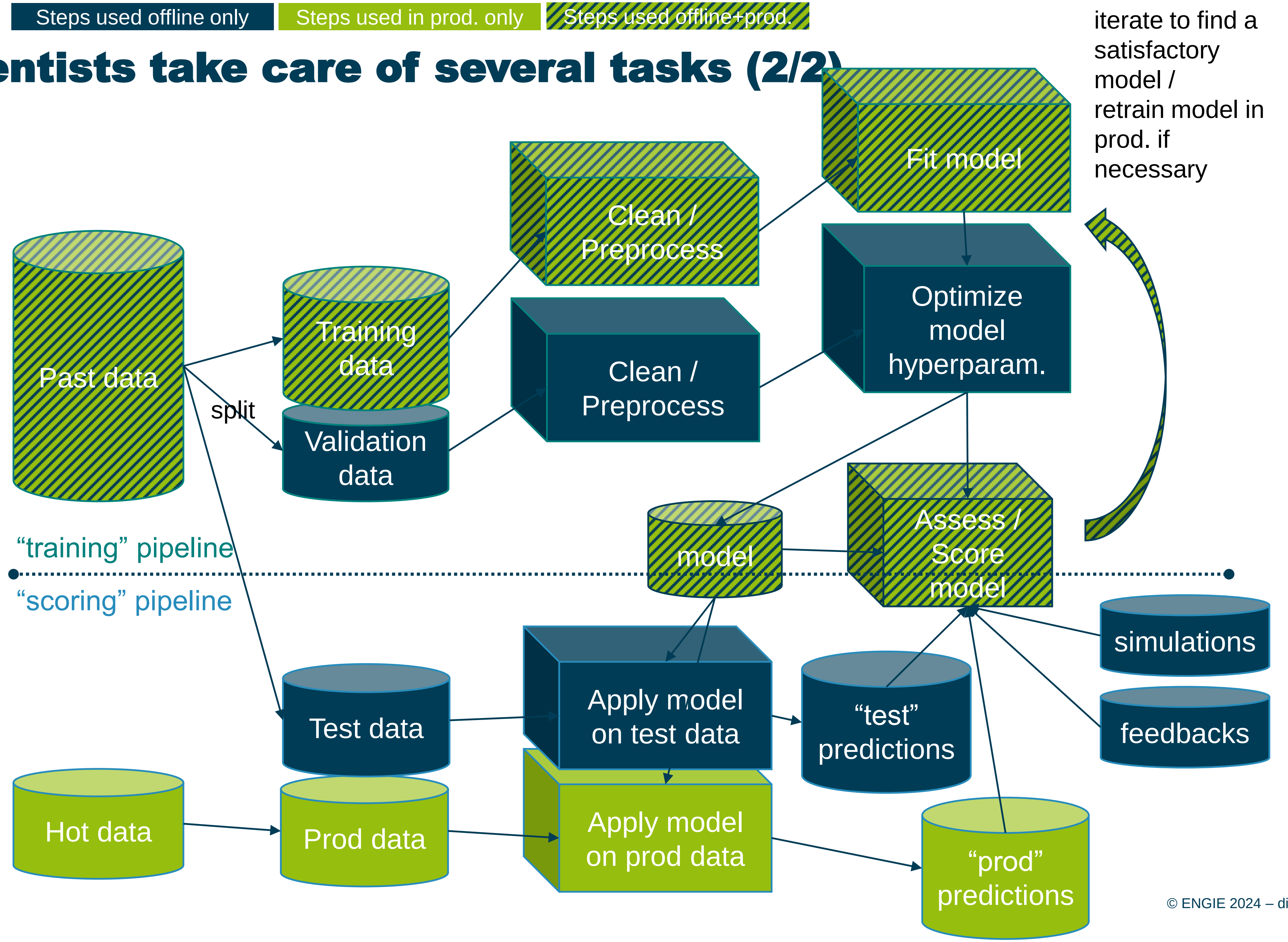
# Should data scientists care about the domain of application?

# Data scientists take care of several tasks (1/2)

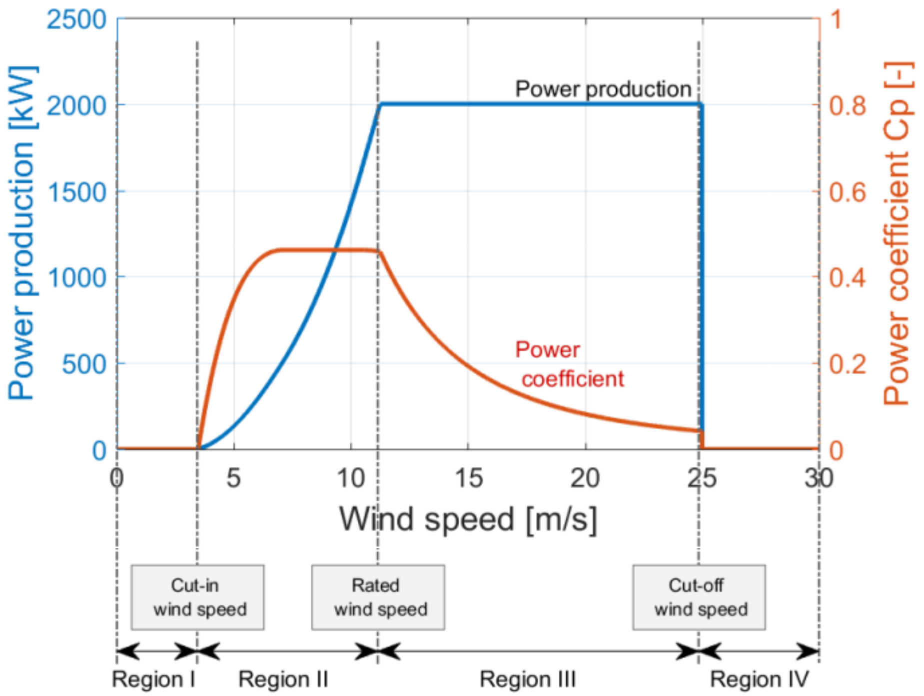
- Designing algorithms through research & prototyping phases
- Developing code and software with high quality standards
- Deploying these algorithms at scale
- Serving results of these algorithms through studies / web apps / reports
- Monitoring that everything works well
- Sharing documented software packages with other data scientists



# Data scientists take care of several tasks (2/2)

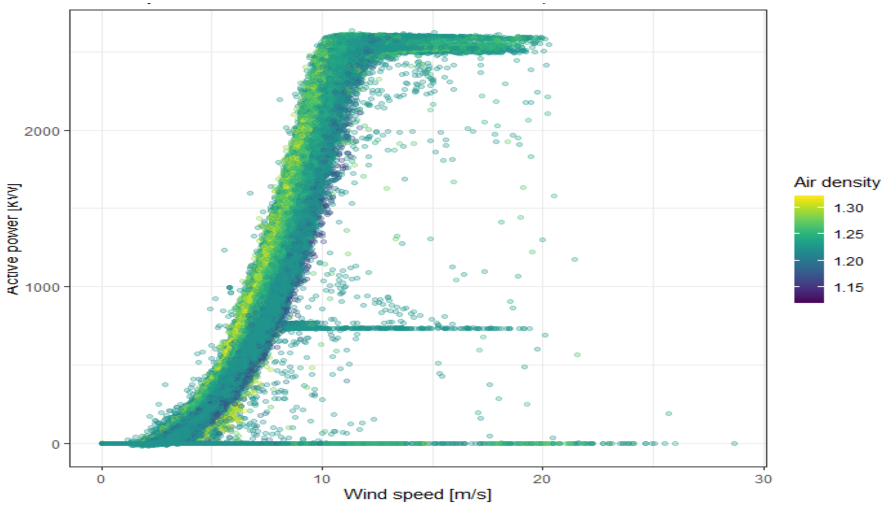


# Yet, there are some physical laws to consider, e.g., for feature engineering




Yves-Marie Saint-Drenan et al. (2019)

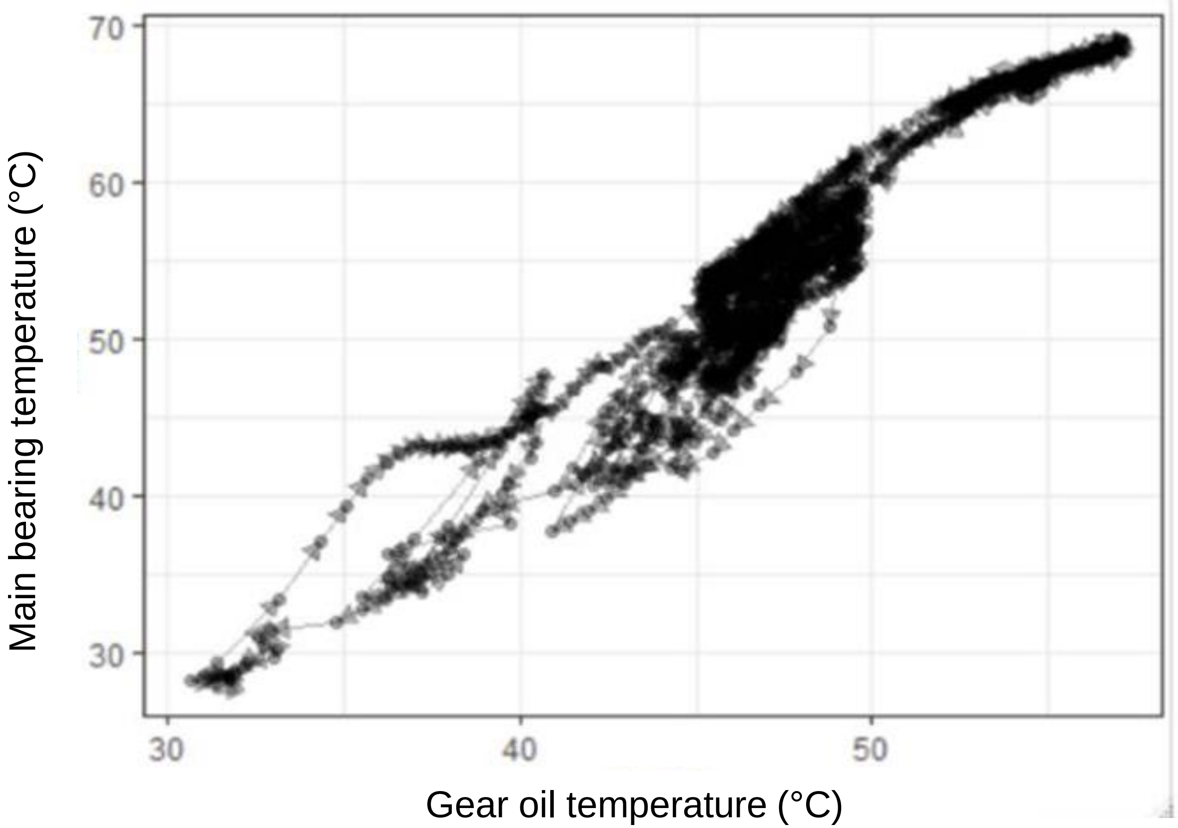
$$P_{WT} = \frac{1}{2}\rho A_{rotor} V_{WS}^{3} C_p(\lambda, \beta)$$




*Air density is known to have an influence on wind turbine performance.*

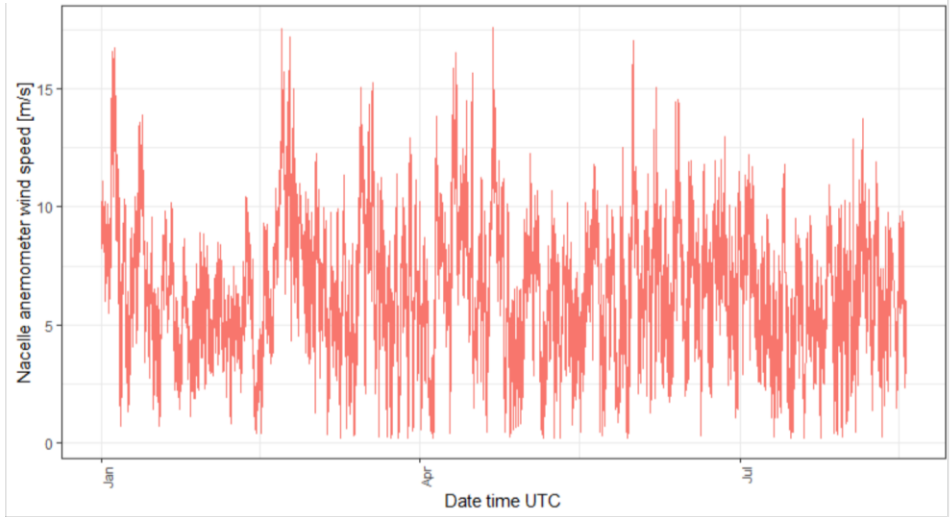
# Also, one should be warned that wind turbines are regulated machines (with control-command)

- Various control loops are at hand in a wind turbine.
- This breaks some “causal behaviors” often assumed by data scientists.
- Starting / Stopping phases may create additional hysteresis effects.

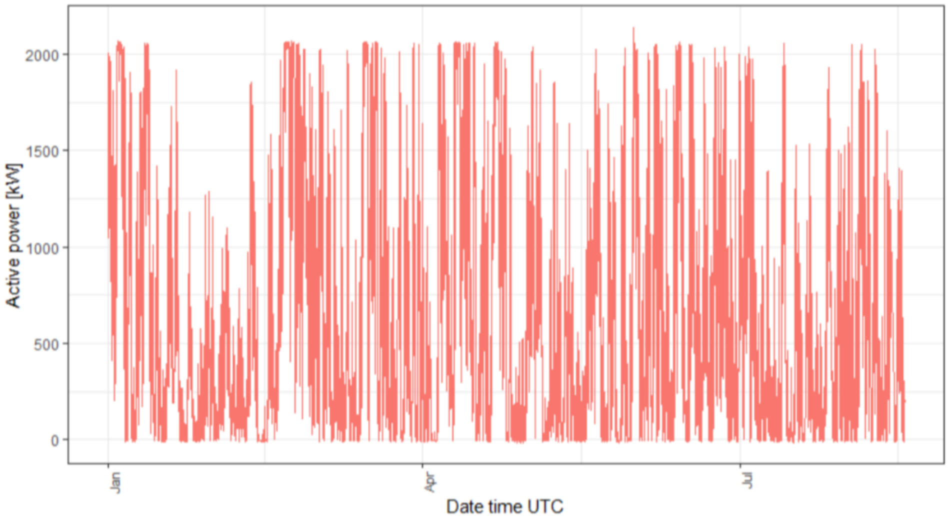


# Wind speed is quite non-stationary as a random process

A time series of **wind speeds** (measured every 10 min.)




A time series of **active powers** (measured every 10 min.)




- Wind speed is a non-stationary process with multi-seasonal effects that have impacts on the behavior of every component of a wind turbine.
- On top of that, the quality of wind speed measurement is **not well-known** (and is a never-ending concern in the wind business).

# Wind turbines behavior depends a lot on the surrounding environment

- trees / forests,
- other wind turbines,
- terrain rugosity,
- etc.

have an influence on the **turbulence and force of wind speed** received by a wind turbine.

*Photo by Christian Steiness / Vattenfall*
*(Horns Rev Offshore Wind Farm, Denmark)*

# A wind turbine is not just another industrial asset

*https://windeurope.org/*

# What is anomaly detection?

# What do we mean by “anomaly detection & diagnosis”?

⇒ Anomaly detection consists in a series of steps
⇒ A key consideration is that we don’t have the “real” anomalies to learn from

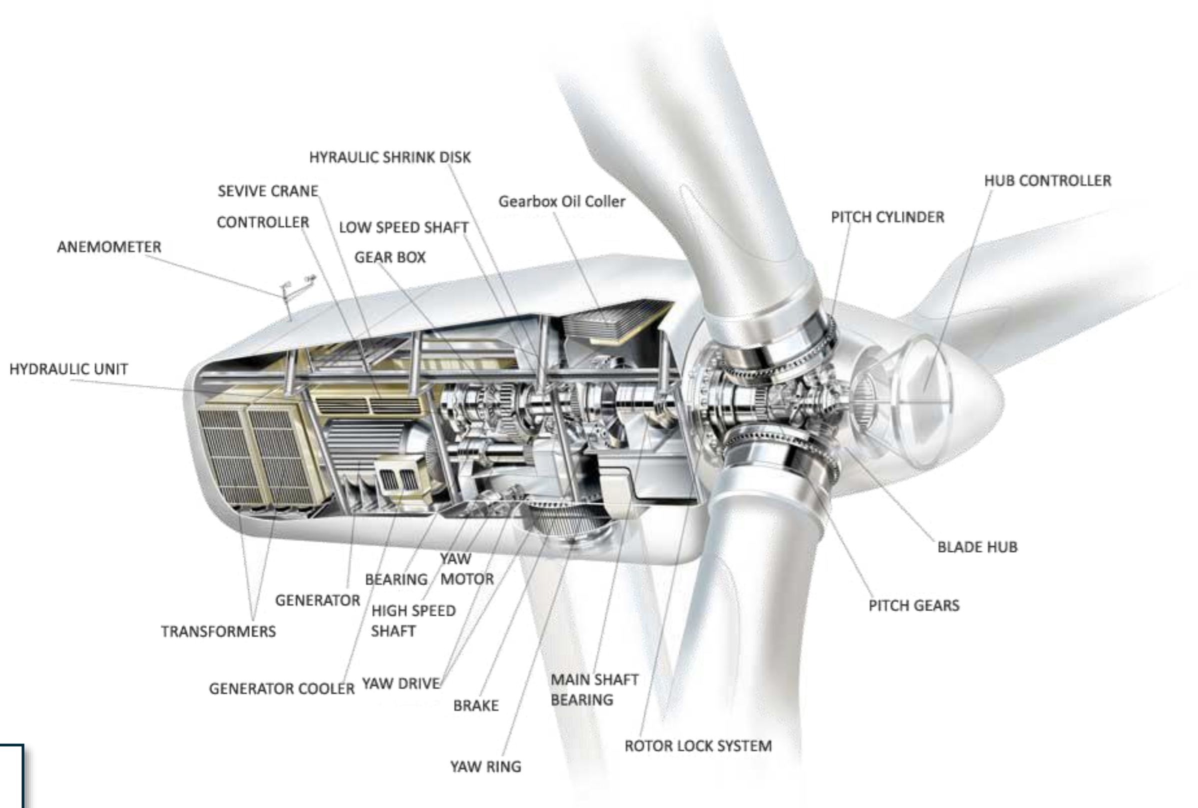
Detect anomalies / health & performance issues

Inform / Alert operators & maintainers

Provide diagnosis (root cause)

Require confirmation

Prescribe actions + follow-up



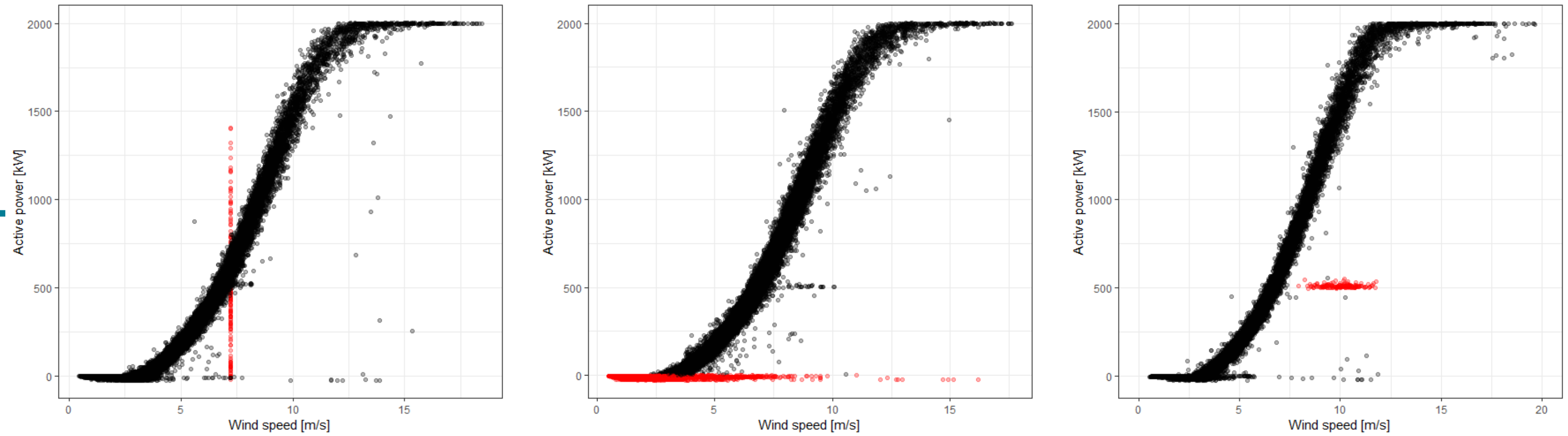

https://www.renewableenergyhub.co.uk/

# There are various terms for anomaly detection…

You may hear e.g., about:

- “Anomaly detection”
- “Outlier detection”
- “Early warning”
- “Novelty detection”
- “Pattern recognition”
- …

# What do we mean by “an anomaly”?

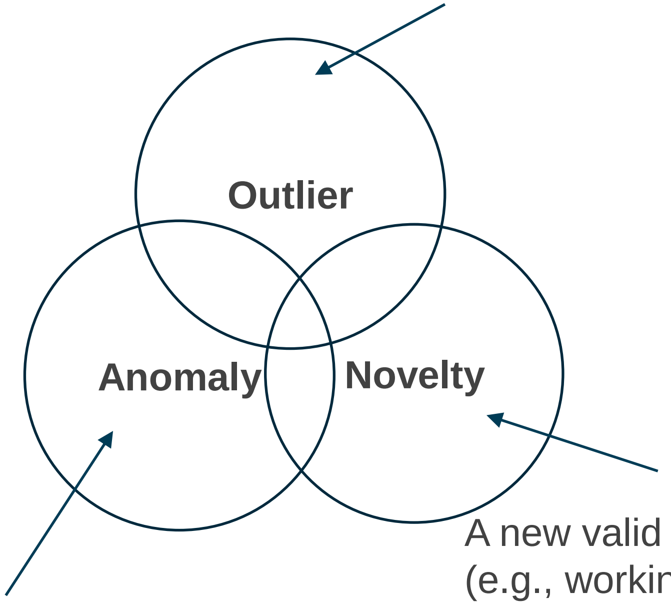



- <u>Anomaly</u>: what is not normal or not expected

- <u>Outlier</u>: what differs significantly from other observations

- <u>Novelty</u>: something new or unusual

- A fully automated anomaly detection is infeasible. In practice it’s automatic novelty & outlier detection that is often achieved



# There are various methodological approaches embedded in these software

- Rule-based approaches
- Physics-based approaches. Deviation from this model is an anomaly.
- Statistical analysis (without a learning phase)
- Machine learning based approaches that learn from historical data and build a “normality model”
- Contour-based approaches, also based on machine learning, but more of geometric and probabilistic flavor
- Ensemble approaches

# There are various methodological approaches embedded in these software

- Rule-based approaches
- Physics-based approaches. Deviation from this model is an anomaly.
- Statistical analysis (without a learning phase)
- Machine learning based approaches that learn from historical data and build a “normality model”
- Contour-based approaches, also based on machine learning, but more of geometric and probabilistic flavor
- Ensemble approaches

More expert domain knowledge

More AI

# There are algorithms & methods that support the abovementioned approaches

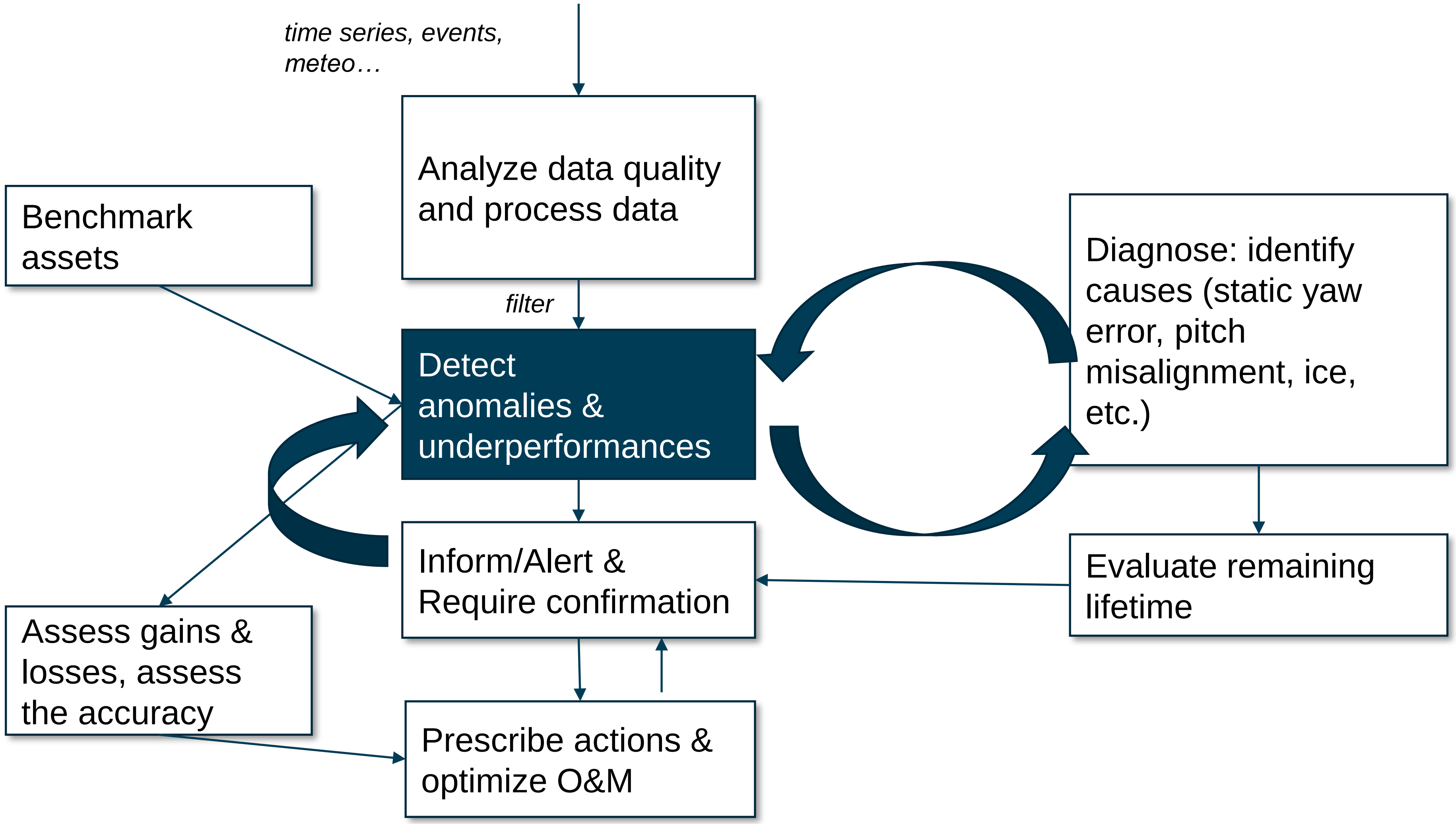
- Fuzzy logic models
- Digital twins
- (Semi-)Supervised models / Regression based models
- Probabilistic models / Bayesian networks
- Unsupervised models / Clustering models
- Deep learning & autoencoders
- …

# Be aware that each algorithm / method still has multiple variations

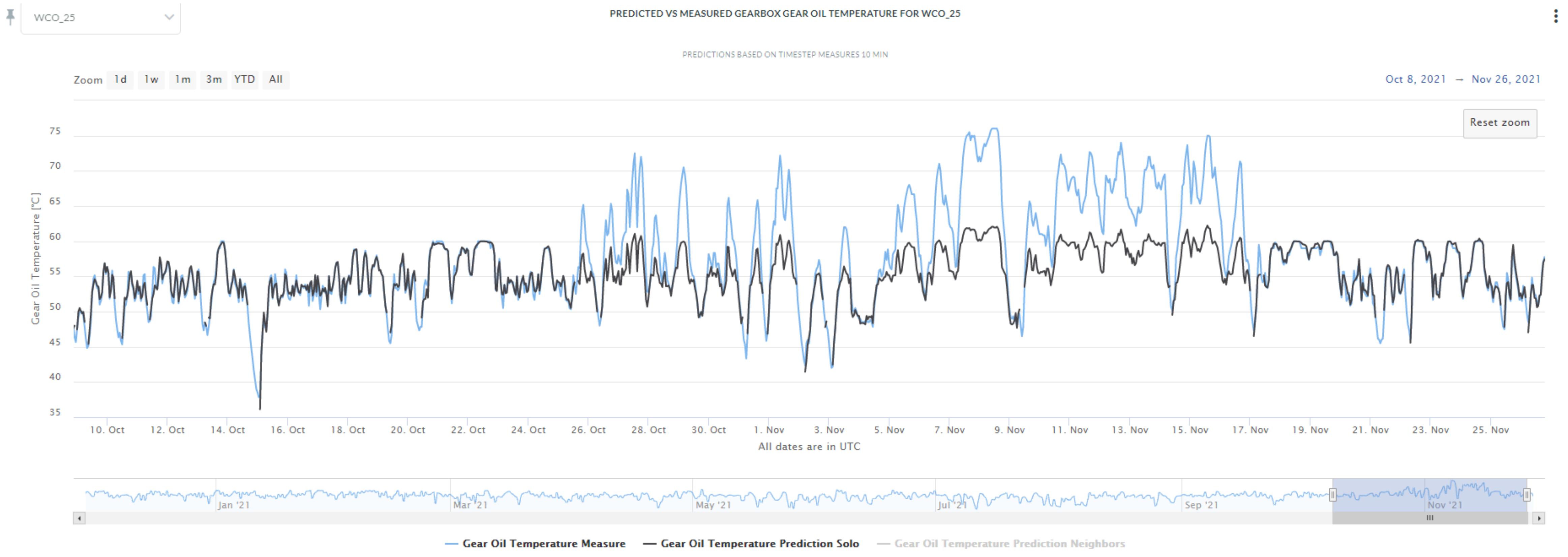
For instance, regarding regression-based models:

- What regression model to chose? (linear, nonlinear, GAM / nonparametric, random forest, neural network…)
- What numerical implementation of the regression model to chose?
- How to select the $x$ variables?
- Do we apply the regression model to a single asset? To multiple assets in one go?
- etc.

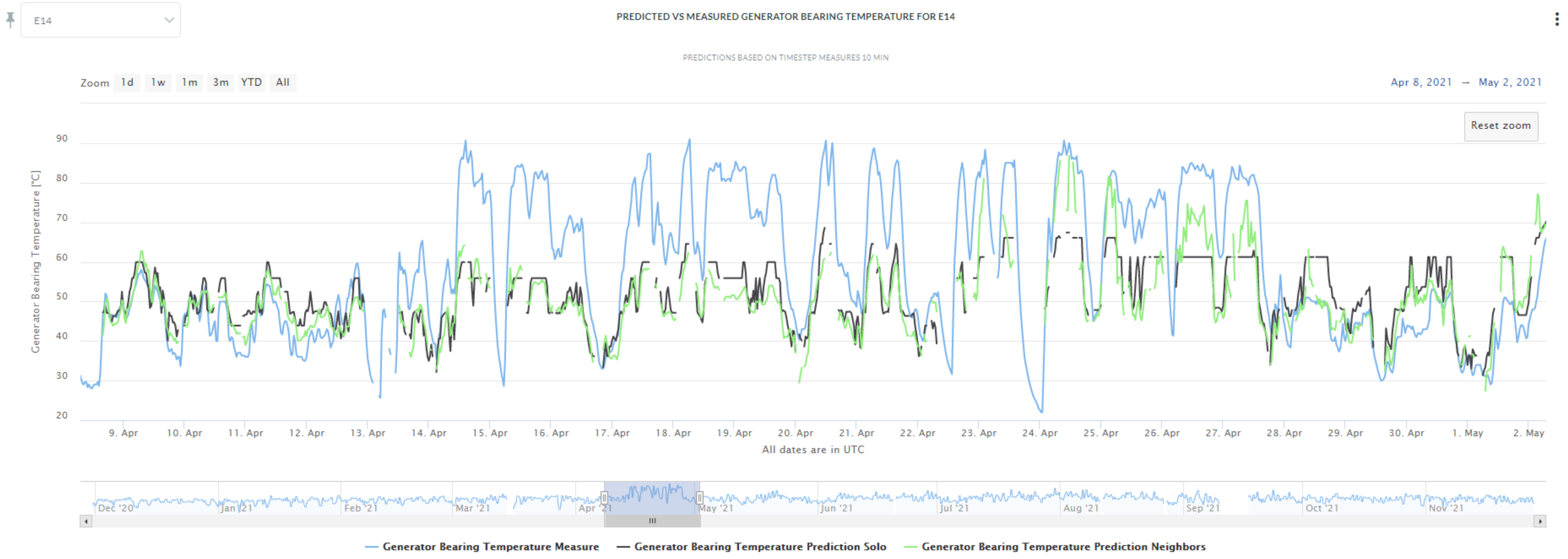
# We have been building up a system with several modules for detecting anomalies on renewable assets



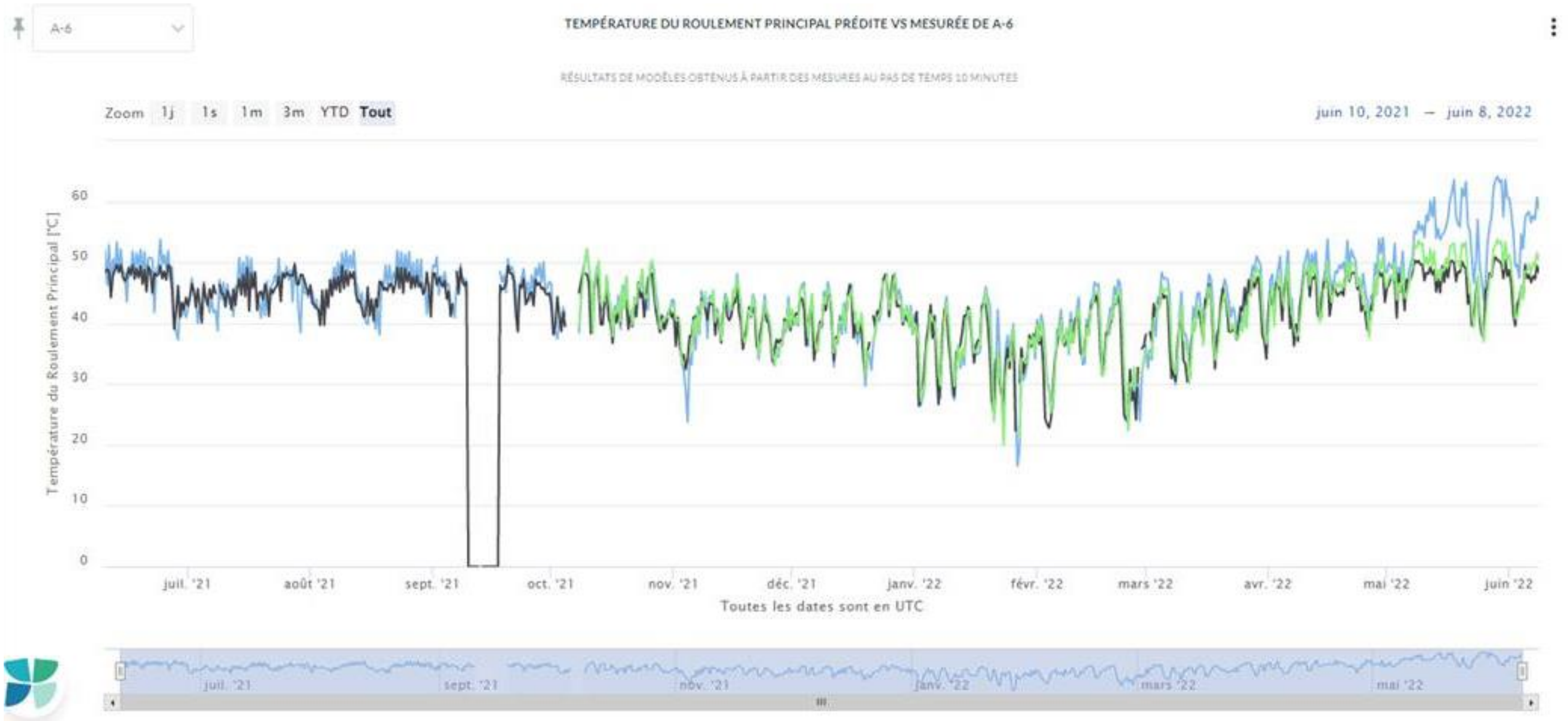
# Example of anomaly: gear oil temperature issue in October 2021



# Example of anomaly: generator bearing temperature issue in April 2021



# Example of anomaly: main bearing temperature issue in June 2022



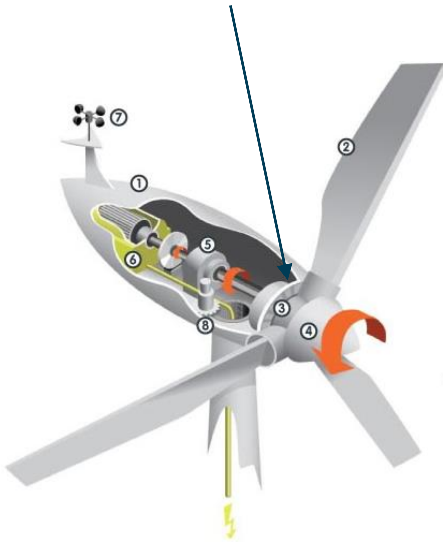
# Caveat: mostly, we do not have anomaly labels!

Some real anomalies are available:

- enough to validate our models against these real cases
- but **not enough** to build a learning model directly on these anomalies.

# Consequence of this absence of labels: we use regression models to be trained on wind turbine sensors

For instance, we are interested in anomalies in the main bearing of a wind turbine.




www.japprends-lenergie.fr

*A damaged main bearing of a wind turbine – Courtesy of Engie Green / Damien Bruyère*

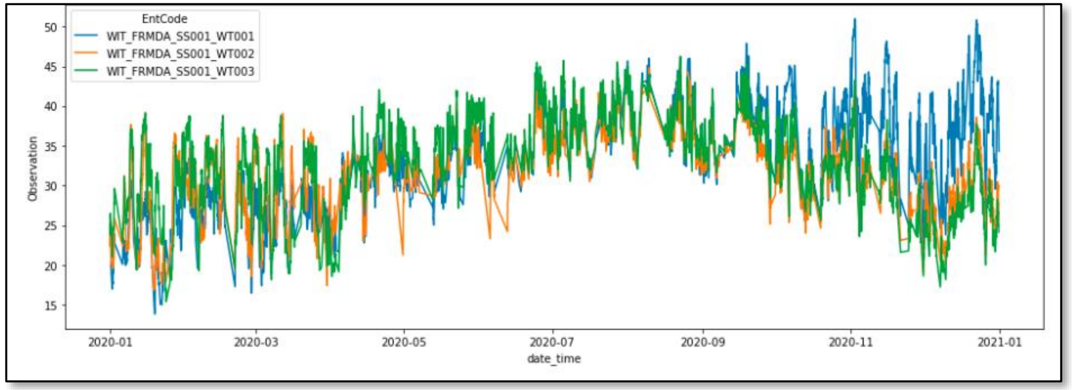
# Consequence of this absence of labels: we use regression models to be trained on wind turbine sensors

For instance, we are interested in anomalies in the main bearing of a wind turbine.

- So, we build a regression model on the main bearing temperature $y$

$$y_t = f(x_t) + \sigma(x'_t)\varepsilon_t$$

- To detect anomalies, we observe residuals $y_t - \hat{y}_t$ and the related prediction intervals deduced from the estimation of $\sigma(.)$




*An "obvious" example – Wind turbine WT001 has a very high main bearing temperature compared to wind turbines of the same farm*

# Consequence of this absence of labels: we use regression models to be trained on wind turbine sensors

For instance, we are interested in anomalies in the main bearing of a wind turbine.

- So, we build a regression model on the main bearing temperature $y$

$$y_t = f(x_t) + \sigma(x'_t)\varepsilon_t$$

- To detect anomalies, we observe residuals $y_t - \hat{y}_t$ and the related prediction intervals deduced from the estimation of $\sigma(.)$
- With this regression approach, we transform an unsupervised problem into a supervised one (in some way).
  - No free lunch: we must lose something, somewhere.

# Consequence of this absence of labels: we use regression models to be trained on wind turbine sensors

For instance, we are interested in anomalies in the main bearing of a wind turbine.

- So, we build a regression model on the main bearing temperature $y$

$$y_t = f(x_t) + \sigma(x'_t)\varepsilon_t$$

- To detect anomalies, we observe residuals $y_t - \hat{y}_t$ and the related prediction intervals deduced from the estimation of $\sigma(.)$
- With this regression approach, we transform an unsupervised problem into a supervised one (in some way).
  - No free lunch: we must lose something, somewhere.
- From the residuals (train + test phase), we build a function $g_\theta(\{y_t - \hat{y}_t\}_t)$ to trigger alerts
  - Here we need a good notion of distance between (possibly multivariate) probability distributions.

# Consequence of this absence of labels: we use regression models to be trained on wind turbine sensors

For instance, we are interested in anomalies in the main bearing of a wind turbine.

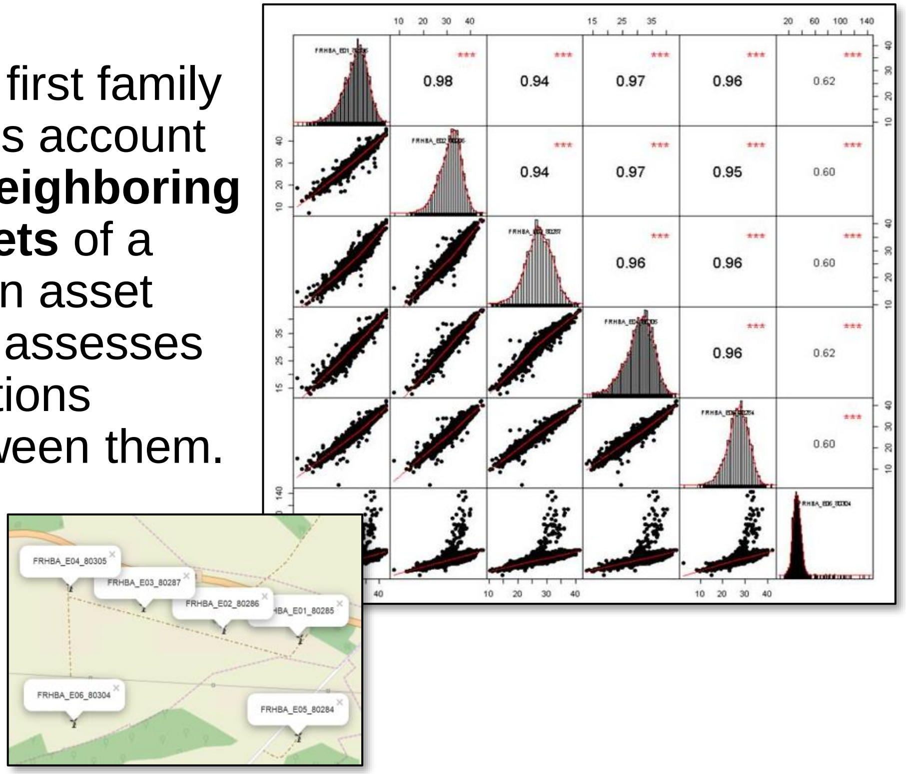
- So, we build a regression model on the main bearing temperature $y$

$$y_t = f(x_t) + \sigma(x'_t)\varepsilon_t$$

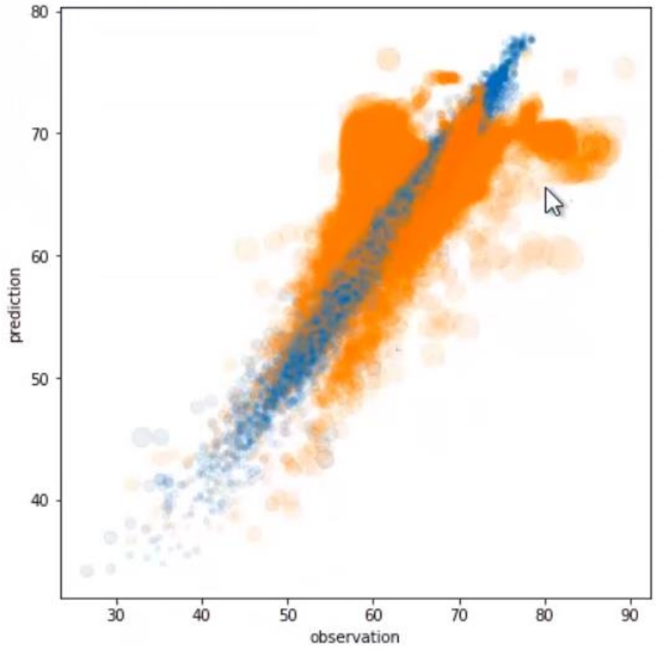
- To detect anomalies, we observe residuals $y_t - \hat{y}_t$ and the related prediction intervals deduced from the estimation of $\sigma(.)$
- With this regression approach, we transform an unsupervised problem into a supervised one (in some way).
  - No free lunch: we must lose something, somewhere.
- From the residuals (train + test phase), we build a function $g_\theta(\{y_t - \hat{y}_t\}_t)$ to trigger alerts
  - Here we need a good notion of distance between (possibly multivariate) probability distributions.
- With user feedbacks we may e.g., choose to retrain the model, disactivate the model, or update the parameter $\theta$.

# Two families of algorithms are currently enabled in ALPHEE to detect abnormal behaviors

- The first family takes account of **neighboring assets** of a given asset and assesses relations between them.




- The second family uses regression models to represent **causality relations** between sensors inside a given asset.



# A sound anomaly detection system may combine different approaches!

We notably combine:

- Contouring approaches (One-Class Classification) in R&D phase
- Rules for data cleaning
- Statistical analysis for outlier detection and data cleaning
- Physics for feature creation
- Machine learning for accuracy, self-learning and automation
- Rules again for verification and alerting

# A snapshot of machine learning challenges we face

# We face various machine learning challenges

- **The number of trained models to manage in production is increasing** – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)

# We face various machine learning challenges

- **The number of trained models to manage in production is increasing** – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)

Consider approaches with less models:

- Multi-assets models
- Models that are multivariate in Y
- Multi-task learning
- Contour-based approaches / One-class classification

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- **Dimension is becoming high**
  - Not thousands of variables, but sometimes a few hundreds

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- **Dimension is becoming high**
  - Not thousands of variables, but sometimes a few hundreds

Consider approaches to reduce dimension and/or select variables efficiently (and fast enough), but pay attention to **causality**

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- **Operators need predictions on the 1st day their wind farm is connected to Darwin** (no historical period to train on!)

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- **Operators need predictions on the 1st day their wind farm is connected to Darwin** (no historical period to train on!)

Consider transfer learning

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- **Alerts pushed to operators must not only take account of the short term** – the degradation of a wind turbine component can evolve over several months!

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- **Alerts pushed to operators must not only take account of the short term** – the degradation of a wind turbine component can evolve over several months!

Consider online learning / sequential algorithms

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- **We must retrain models at the right time** – not too often in an anomaly detection context

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- **We must retrain models at the right time** – not too often in an anomaly detection context

Consider methods to detect drifts in data and models / predictions

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- We must retrain models at the right time – not too often in an anomaly detection context
- **We want to understand the root cause of an anomaly as easily as possible**
  - distinguish between problems of data, problems of algorithms, problems in assets
  - find what component of the asset must be incriminated

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- We must retrain models at the right time – not too often in an anomaly detection context
- **We want to understand the root cause of an anomaly as easily as possible**
  - distinguish between problems of data, problems of algorithms, problems in assets
  - find what component of the asset must be incriminated

Enrich predictions with additional information, work on explanability

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- We must retrain models at the right time – not too often in an anomaly detection context
- We want to understand the root cause of an anomaly as easily as possible
  - distinguish between problems of data, problems of algorithms, problems in assets
  - find what component of the asset must be incriminated
- **We must make our models more reliable**
  - we would e.g., accept to reduce performance (MAE…) and improve stationarity of residuals or sensitivity to outliers

# We face various machine learning challenges

- The number of trained models to manage in production is increasing – it is costly and uneasy to monitor
  - (thousands of wind turbines) x (many models per turbine)
- Dimension is becoming high
  - Not thousands of variables, but sometimes a few hundreds
- Operators need predictions on the 1st day their wind farm is connected to Darwin (no historical period to train on!)
- Alerts pushed to operators must not only take account of the short term – the degradation of a wind turbine component can evolve over several months!
- We must retrain models at the right time – not too often in an anomaly detection context
- We want to understand the root cause of an anomaly as easily as possible
  - distinguish between problems of data, problems of algorithms, problems in assets
  - find what component of the asset must be incriminated
- **We must make our models more reliable**
  - we would e.g., accept to reduce performance (MAE…) and improve stationarity of residuals or sensitivity to outliers

Consider robust ML

# What is the role of a data scientist in the Industry?

Their role is to find the best compromise between:

- model accuracy / business relevance, and
- effort / cost, and
- simplicity / maintainability

# Q & A

**ENGIE** DIGITAL

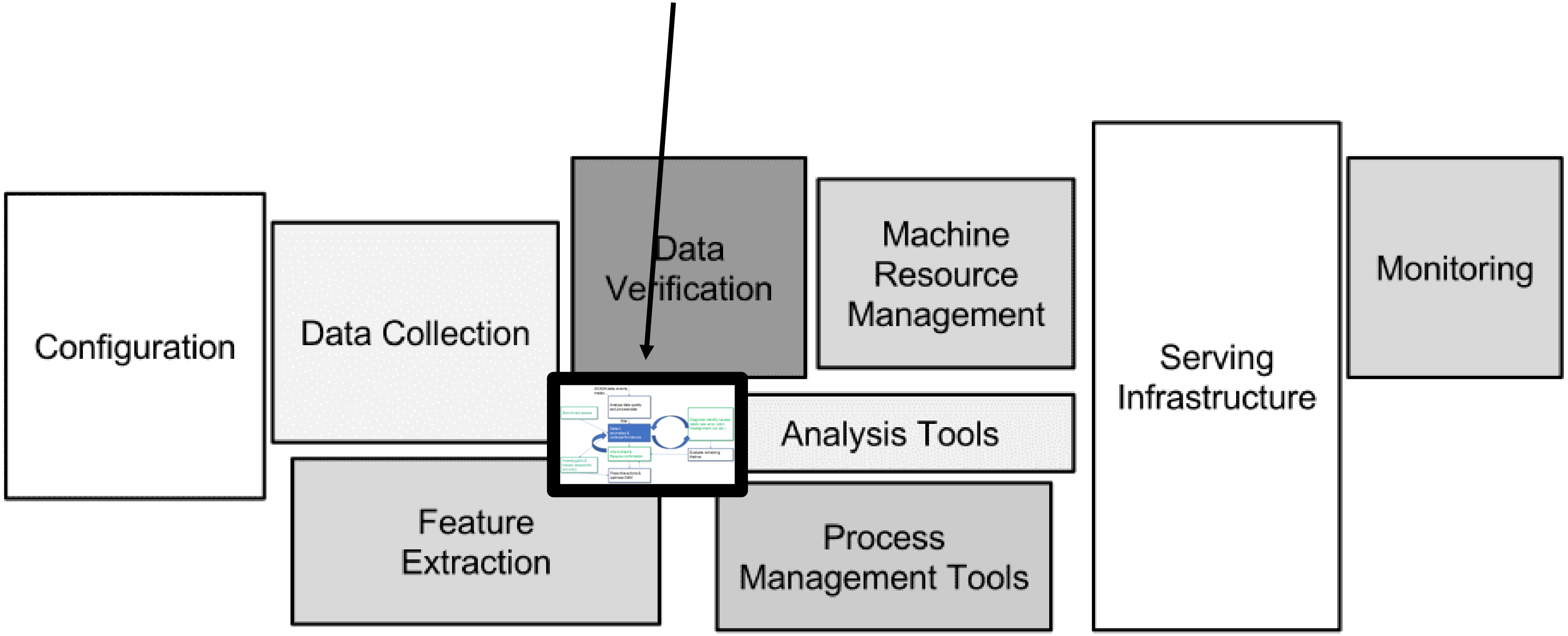
digital.engie.com

# Pros & Cons of the different approaches or methods

| | Pros | Cons | Other info |
|---|---|---|---|
| Rule-based | - Simple at the start<br>- Can embed domain knowledge<br>- Can work right from the asset connection | - No self-learning<br>- Difficult to maintain in the mid term | |
| Physics based | - Explainable<br>- Business experts like it (e.g., PVlib)<br>- Provides insights inaccessible to ML<br>- Not much modelling – physics-empirical laws are known<br>- May work (partly) right from the asset connection | - No self-learning<br>- May require a lot of static data / description of assets<br>- Difficult to maintain in the mid term<br>- Computer intensive | - For digital twins, a surrogate (ML-based) model is usually required for calculations / simulations are impossible to perform exhaustively |
| Statistical analysis | - Simple | - No self-learning<br>- Not sufficient to detect real anomalies | |
| Normality model with ML | - Data-based<br>- Self-learning<br>- Easy to maintain<br>- Evolutive / Can be combined with other approaches<br>- Easy to incorporate user feedback | - Data-based!<br>- Computer intensive if it involves deep learning | - Can be multivariate in Y<br>- Can be more or less black/white box |
| Probabilistic model with BN | - Data-based<br>- Self-learning<br>- Explainable / Helps in root cause analysis<br>- Multivariate in X/Y | - Data-based!<br>- Difficult to train<br>- Data must be discretized | |
| Agnostic or unsupervised model with ML | - Data-based<br>- Self-learning<br>- Suited to R&D / analyses / discovery phases<br>- Interested when limited domain knowledge is avail. | - Data-based!<br>- Uneasy to incorporate user feedback<br>- Might be hopeless in a nonstationary context – usual normality approaches might be needed at first anyway<br>- Computer intensive if it involves deep learning | - Is usually multivariate “in X/Y”<br>- There is usually no distinction between X and Y |

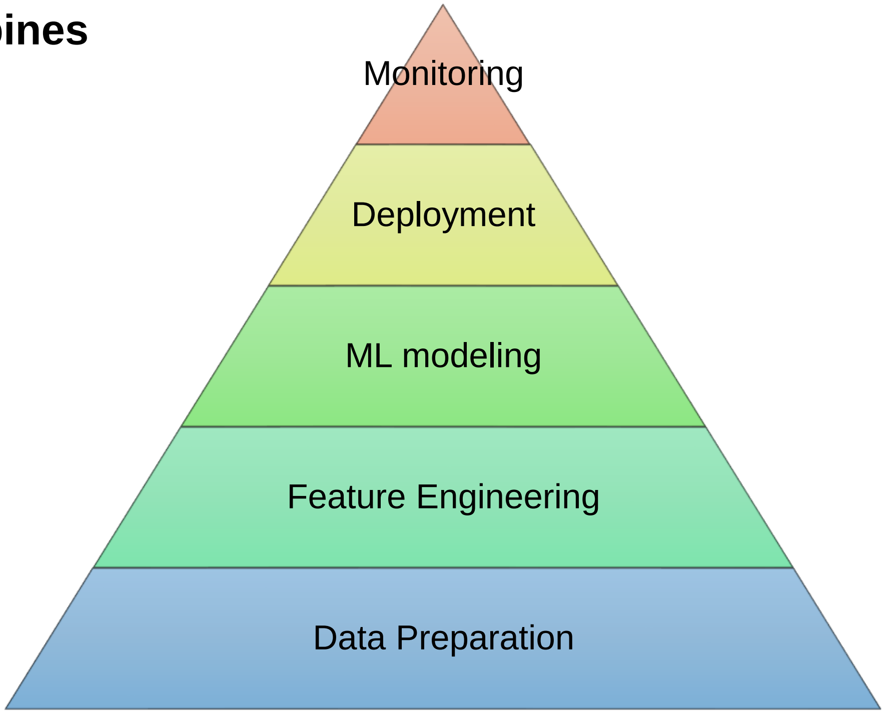
# The ML part is a small part of the whole picture

There is a tendency to talk a lot about the **Modeling part**, but this is just the tip of the iceberg!



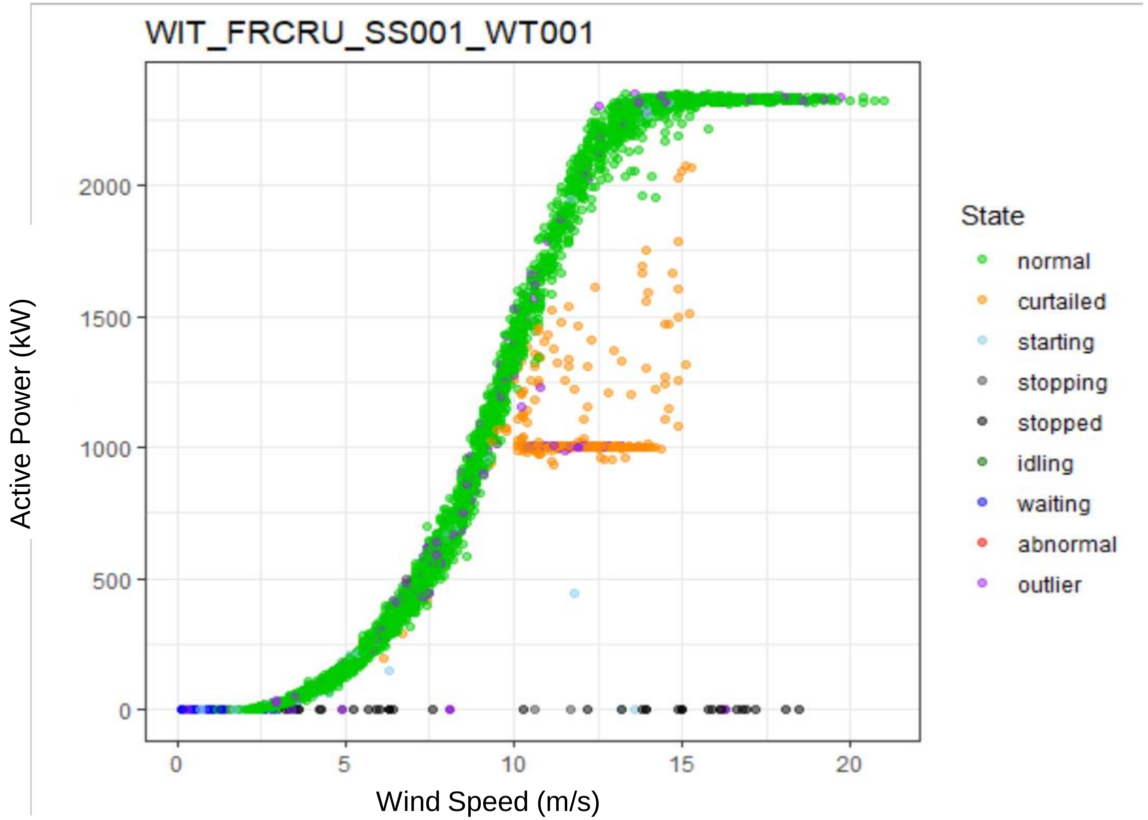
# ALPHEE is Darwin’s data science software dedicated to predictive maintenance on renewable assets

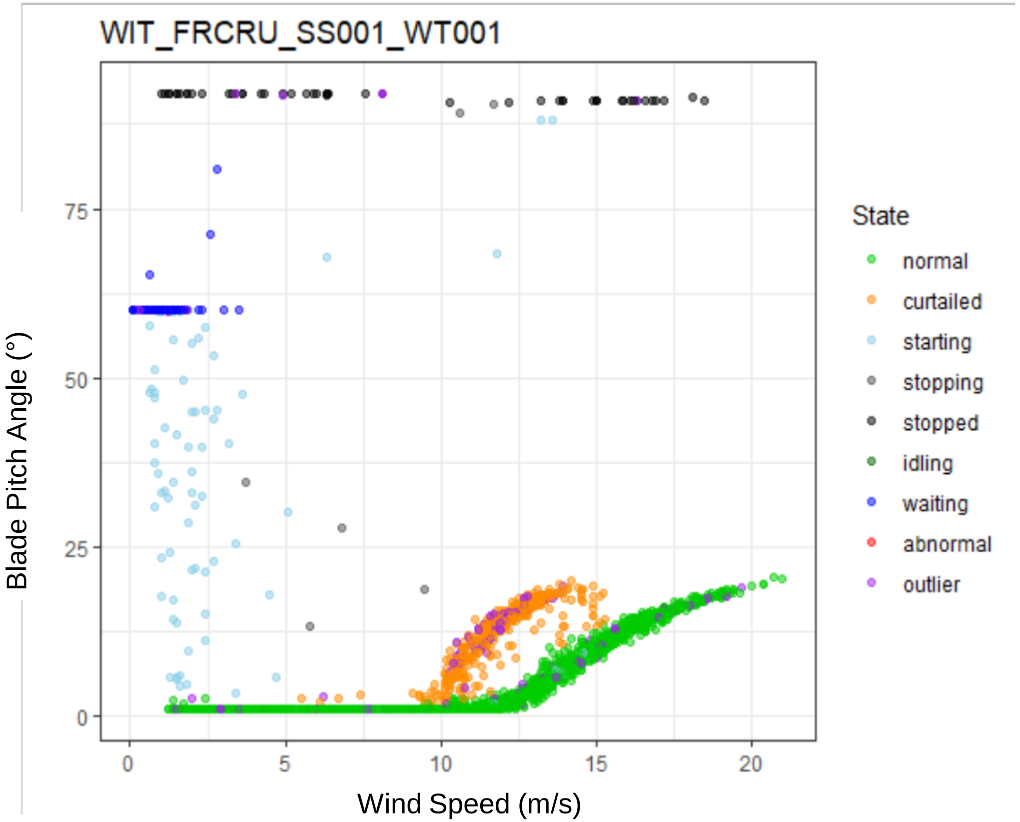
**ALPHEE** is a professional data science software developed since 2016. Its main purpose is to industrialize algorithms on underperformance and anomaly detection in **wind turbines** and **solar PV farms**.



# A portfolio of solutions from detection to diagnosis… still to be developed

# Detecting the operating state of wind turbines

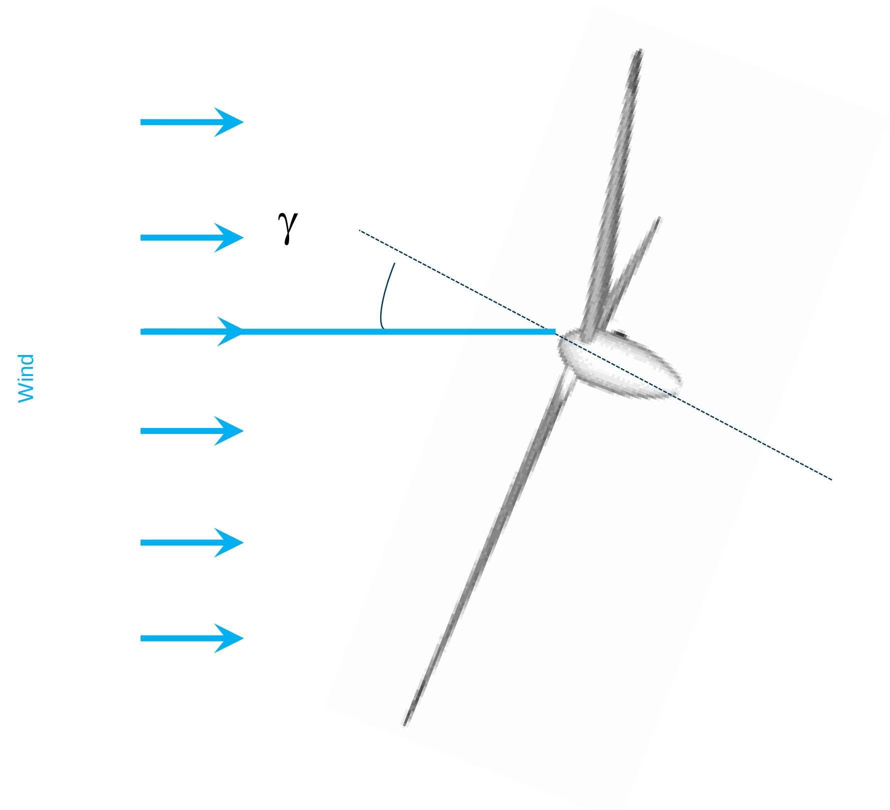


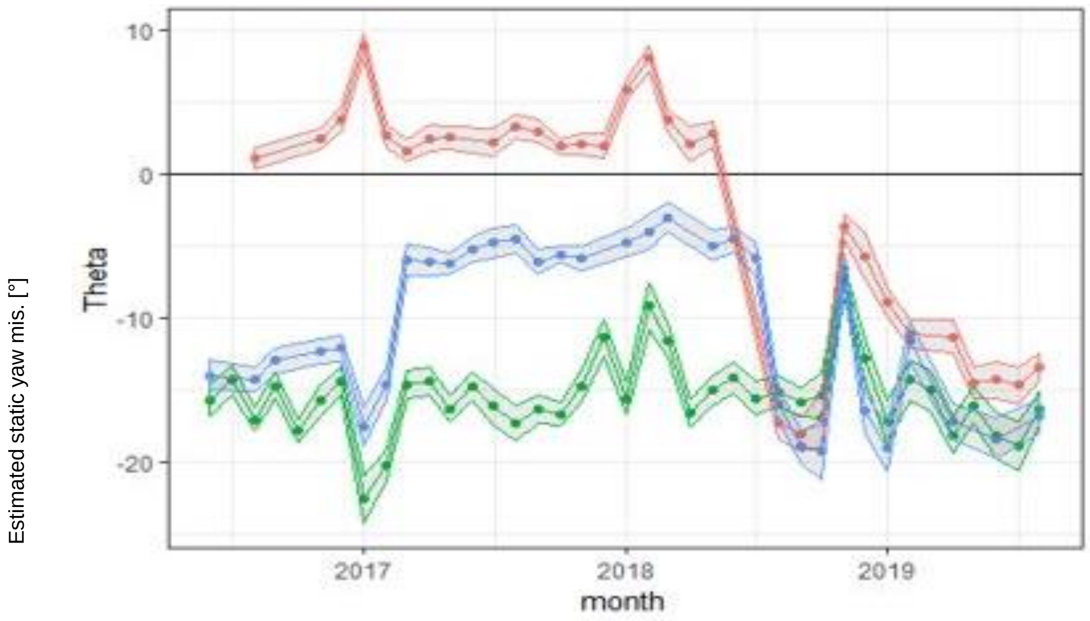

*A wind turbine and its operating states detected.*

**Clustering algorithm**

# Estimating static yaw misalignment (a misalignment of 10° is worth 3% of lost energy production)




*At time t, the turbine may be misalignment by an angle $\gamma$ with respect to wind direction. If this angle is nonzero in average, then we talk about a static yaw misalignment.*



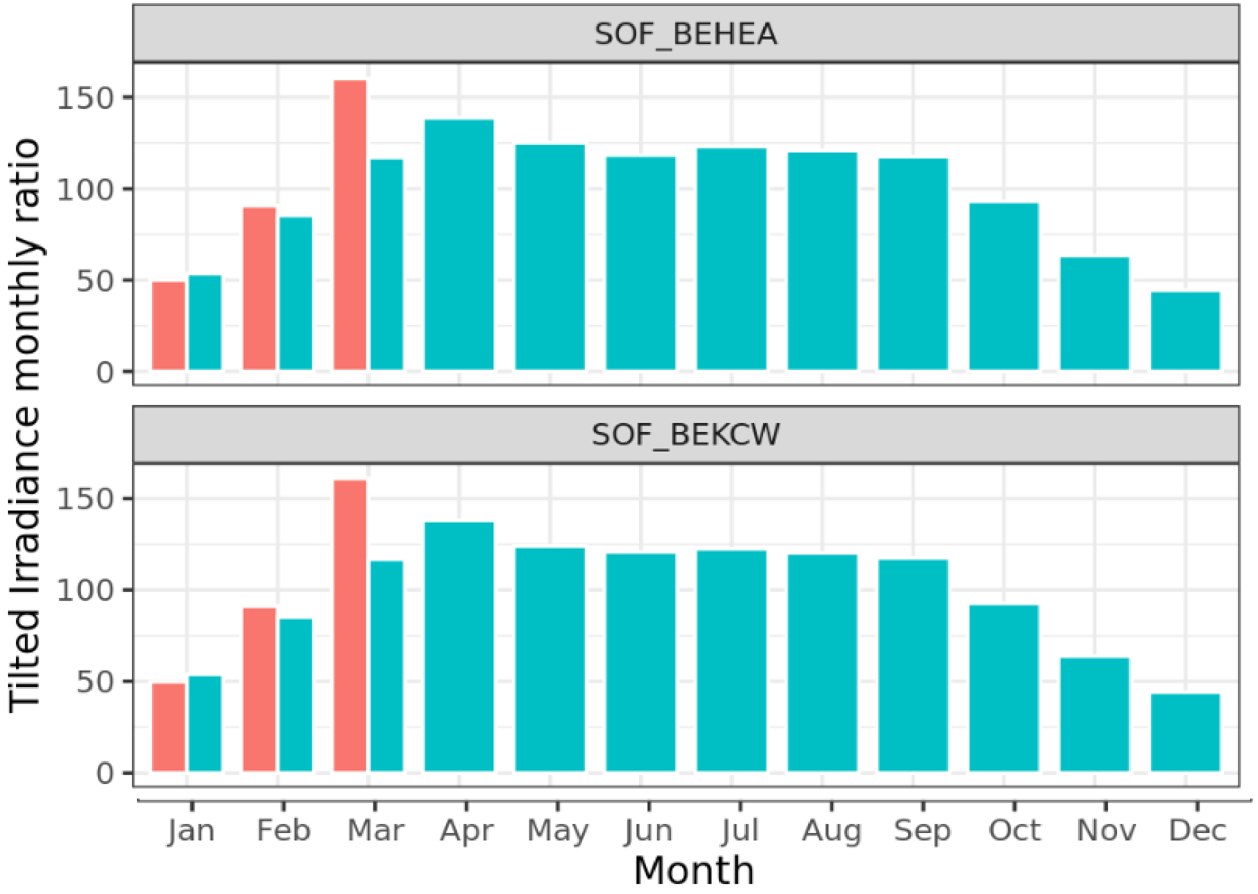

*Applying one of our algorithms on 3 wind turbines of the same wind farm reveals changes of static yaw misalignment over time.*

**Statistical estimation**

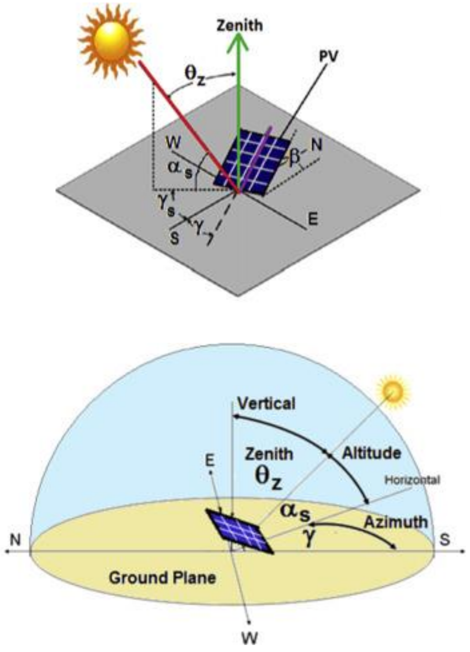
# Assessing meteo effect on wind & solar production

Every month we observe that production is different from expectations;

$\Rightarrow$ What part of these differences can be explained by solar & wind resource variations?




*Two solar farms and their solar indices calculated on the long term (blue) and short term (red)*



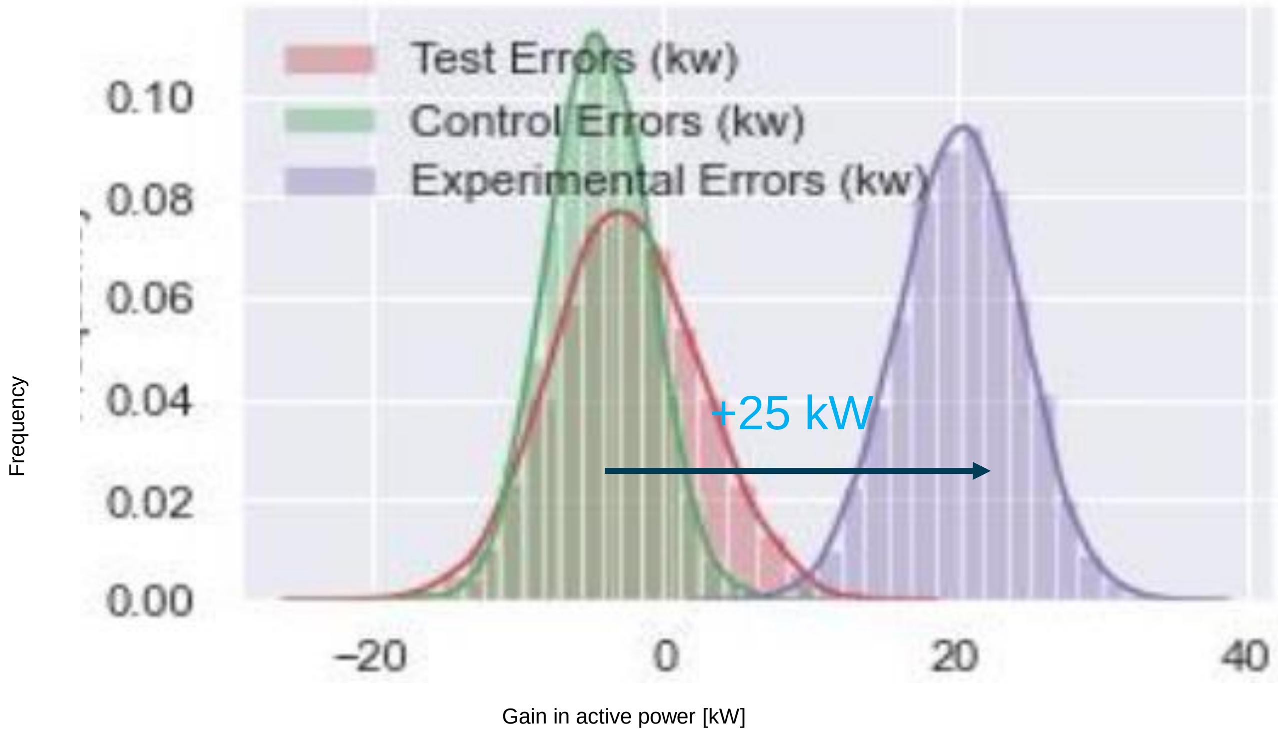

**Physics based algorithm**

# Detecting ice on wind turbines

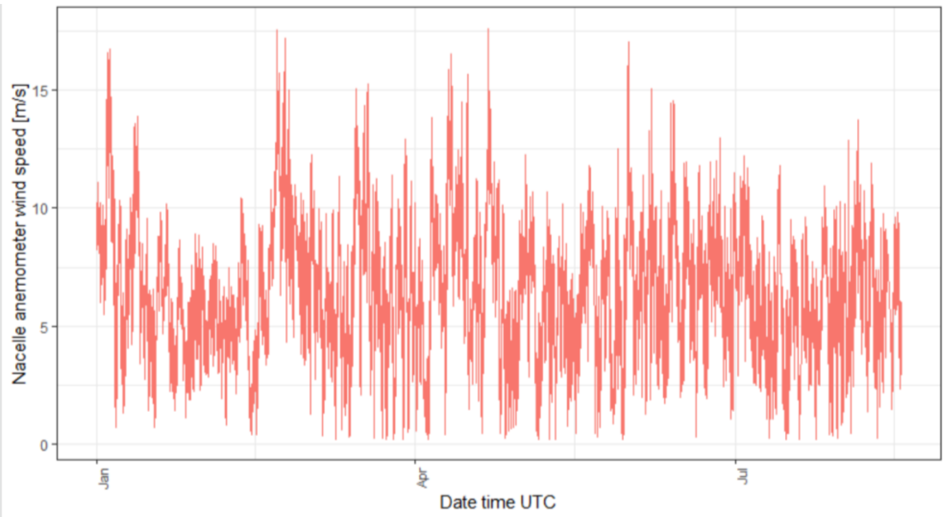
windpowerengineering.com

**Supervised classification**

# Estimating small production gains after a maintenance action



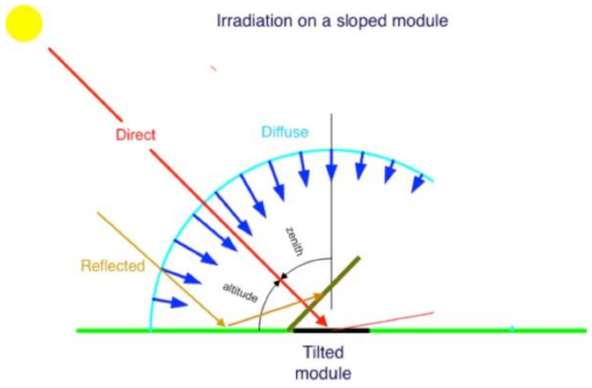

*The gain detected by our algorithm for this retrofit is around +25 kW of additional power produced in average.*




*Assessing such gains is uneasy, since wind speed is a highly non-stationary process with multi-seasonal effects*

**Non-linear regression**

# Creating a digital twin for solar assets




- Clearness index:

$$\text{Clearness index} = \frac{G}{G_{\text{cst}} \cos(z)}$$

- Splitting coefficient / Transposition law:

$$k = 0.952 - 1.041 \exp(-\exp(2.3 - 4.702 \text{ Clearness index}))$$

- Tilted irradiance, in W/m²:

$$G_{\text{tilted}} = \frac{1 + \cos(\text{tilt angle})}{2} k\, G + \cos(\text{AOI}) \,\text{airmass}(z)(1 - k)\, G$$

- Theoretical solar power, in kW:

$$P = \begin{cases} P_{\text{peak}} \dfrac{G_{\text{tilted}}}{1000} \left(1 + \dfrac{\delta}{100}\left(T - T_{\text{avg}}\right)\right) & \text{if } G_{\text{tilted}} > G_{\min} \\ 0 & \text{otherwise} \end{cases}$$

where:

- $G$ is the global horizontal irradiance as given by ERA5 weather data, in W/m²
- $G_{\text{cst}}$ is the solar constant, equal to 1376 W/m²
- $z$ is the sun zenith angle, in radians; it depends on latitude, longitude, date, time of the day
- AOI is the angle of incidence of sun beams on the PV panel, in radians; it relies on a purely trigonometric formula and depends on panel azimuth, panel tilt angle, and sun position
- airmass(z) is the air mass coefficient (unitless), which defines the direct optical path length through the Earth's atmosphere
- $P_{\text{peak}}$ is the peak power of the solar farm, in kW
- $\delta$ is the temperature coefficient of power, in 1/(deg. Celsius), with a default value of -39.
- $T$ is the air temperature at 2 meters as given by ERA5 weather data, in deg. Celsius
- $T_{\text{avg}}$ represents an average temperature, in deg. Celsius, with a default value of 25°C
- $G_{\min}$ is the minimum irradiance below which solar power is zero, in W/m², with a default value of 20 W/m²

# Estimating energy lost by wind turbines

- When an asset (wind turbine…) is stopped, it does not produce energy. How much energy has been lost?

- Estimating energy losses is usually needed for:
  - reporting losses to the management and to the PERFORM database in a unified way;
  - communicating KPIs to shareholders;
  - identifying main losses and their causes, and prioritizing O&M actions;
  - valuing business interruptions to be discussed with the insurer in case of claims.