StoneThree
The Future of Work. Now.™

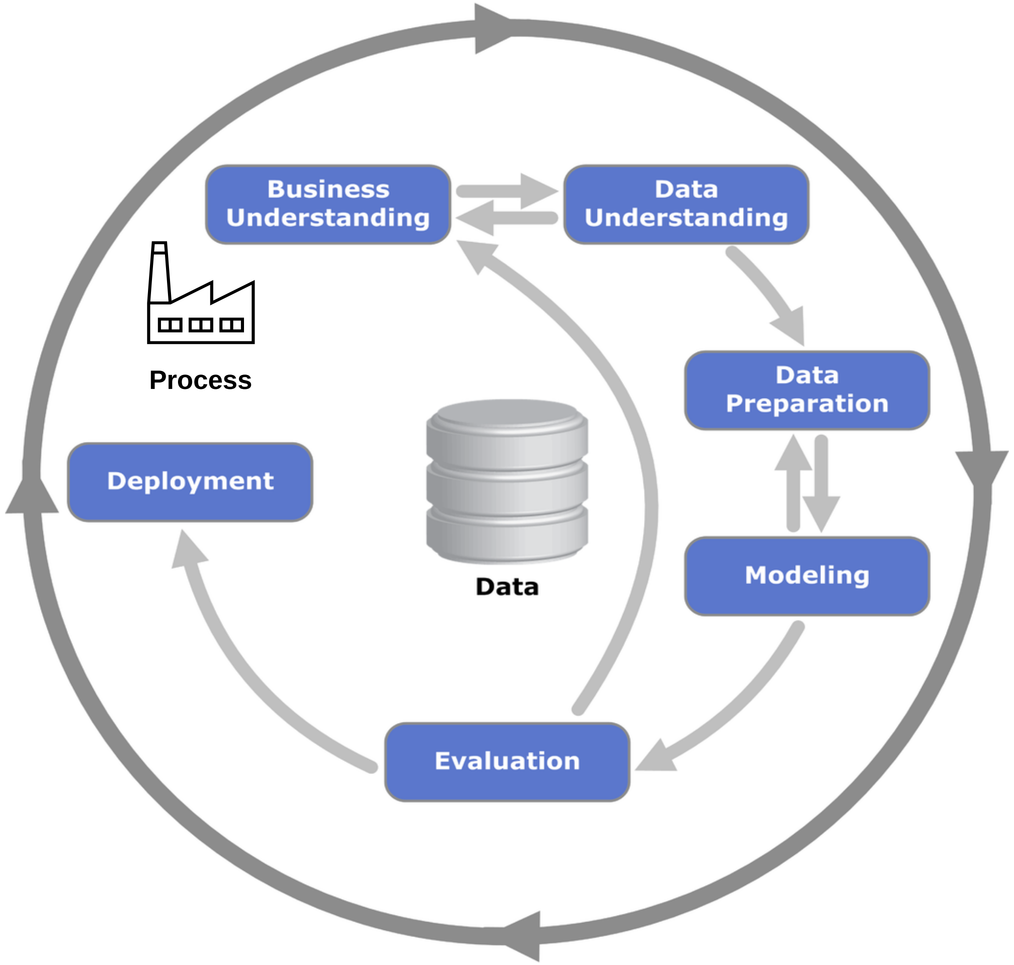
Exploratory Data Analysis

SACAC – 2025 (Lidia Auret, Tobi Louw)

Process data: Challenges and visualization

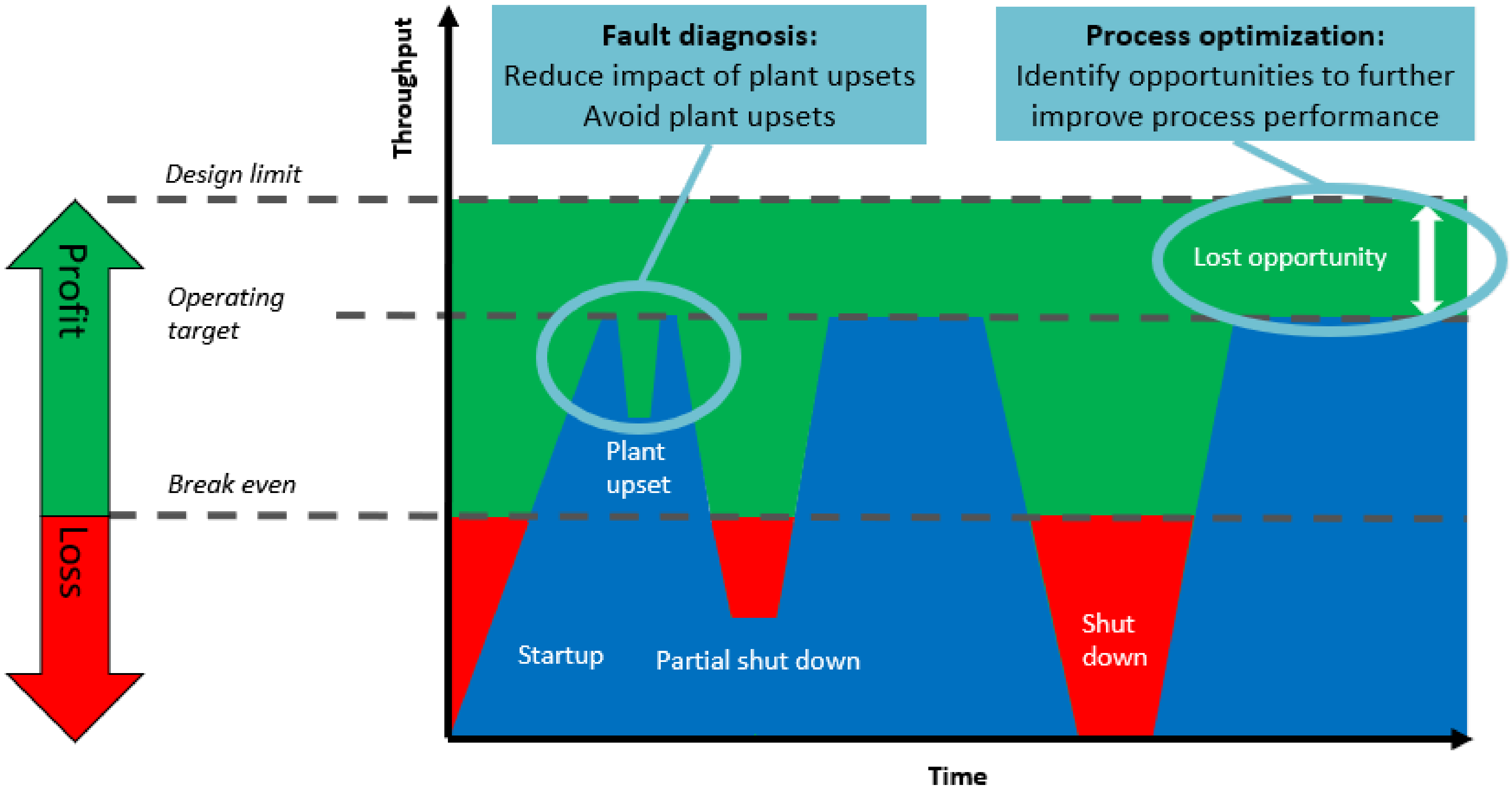
CRISP-DM

Cross-industry standard process for data mining



Process context

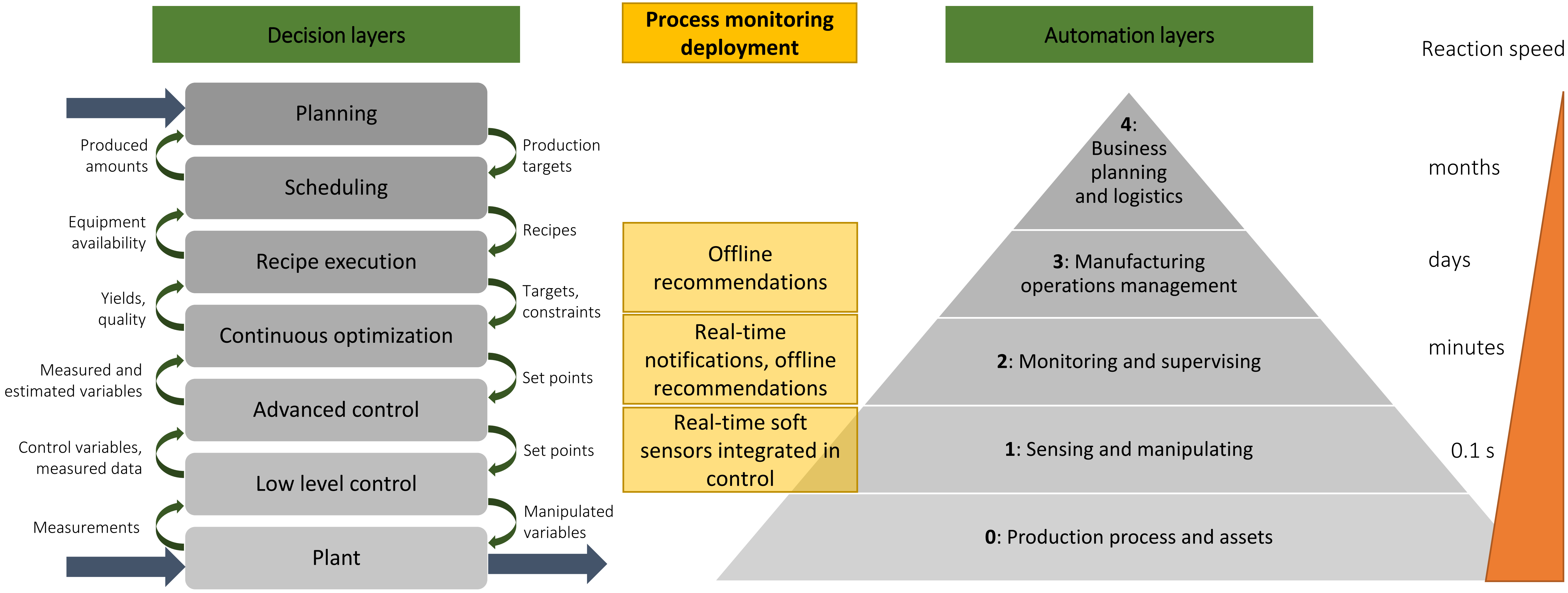
Process monitoring: Fault diagnosis and process optimization



Sand and Terwiesch, 2013. Closing the loops: An industrial perspective on the present and future impact of control. Euro J. Control. 19, 341-350.

Process context

Automation hierarchy

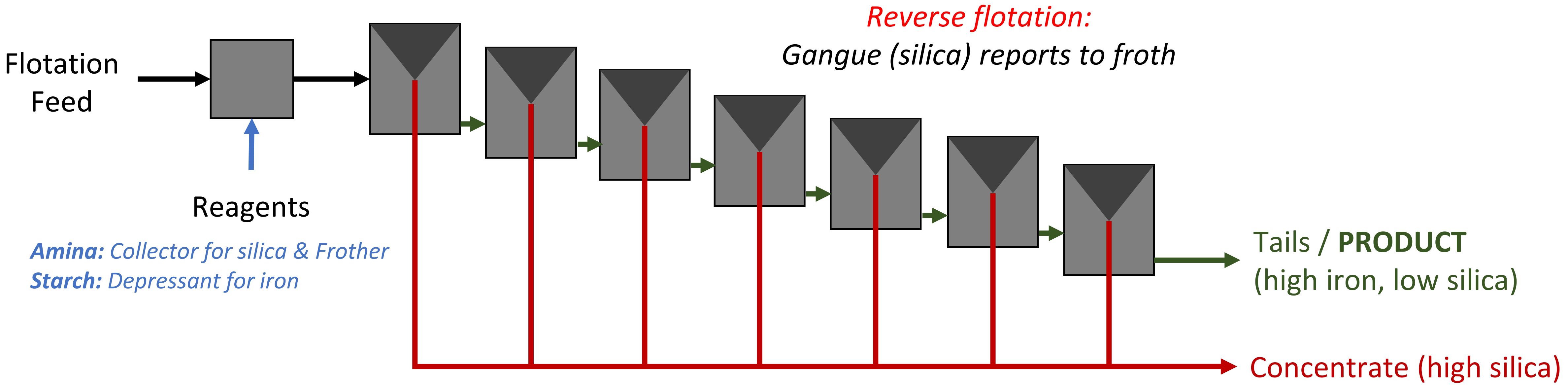


Isaksson, Harjunkoski and Sand, 2018. The impact of digitalization on the future of control and operations. Comp Chem Eng. 114, 122-149.

Process context

Case studies: Open access - Iron ore flotation

- **Key performance indicators:**
  - Impurity (silica content) in product; product rate; value in product (iron content); reagent use
- **Disturbances:**
  - Feed flow, feed density, feed composition
- **Decisions:**
  - Air flow, froth depth, reagent addition

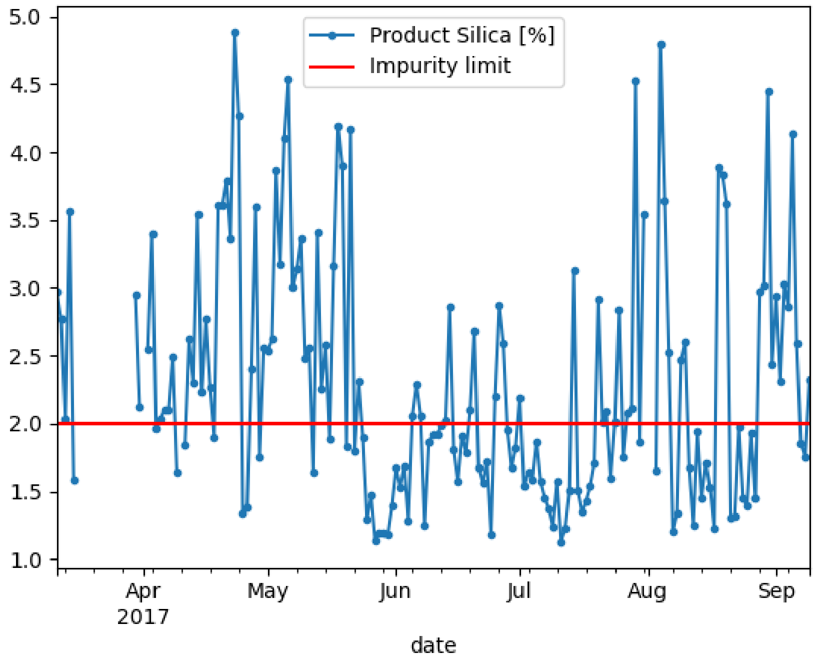


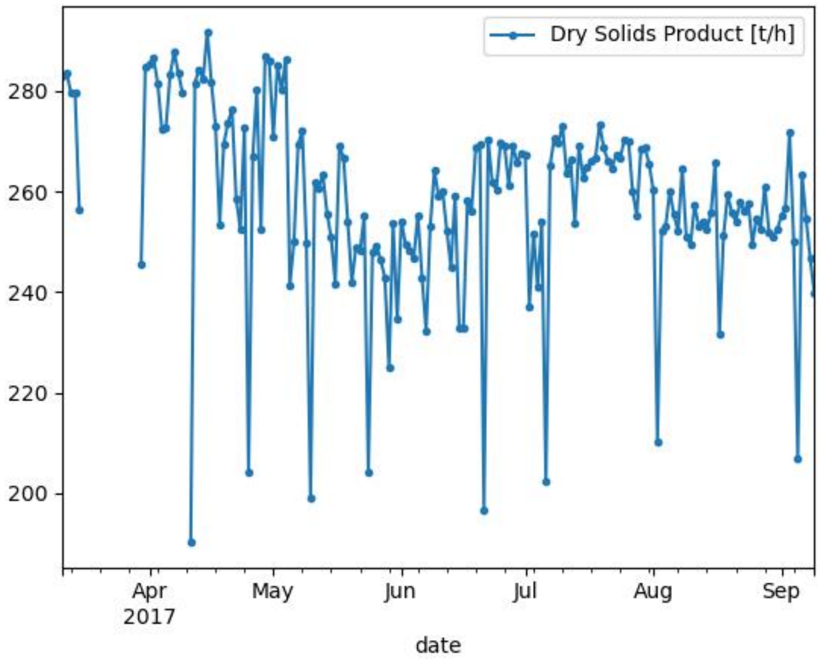
https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process?select=MiningProcess_Flotation_Plant_Database.csv

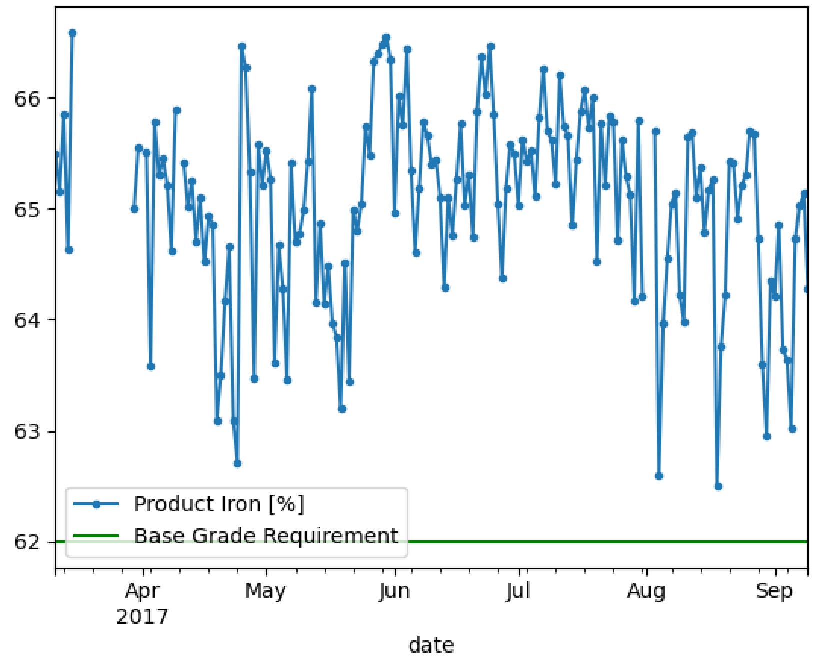
Process context

Case studies: Open access - Iron ore flotation

- **Key performance indicators:**
  - Impurity (silica content) in product; product rate; value in product (iron content); reagent use







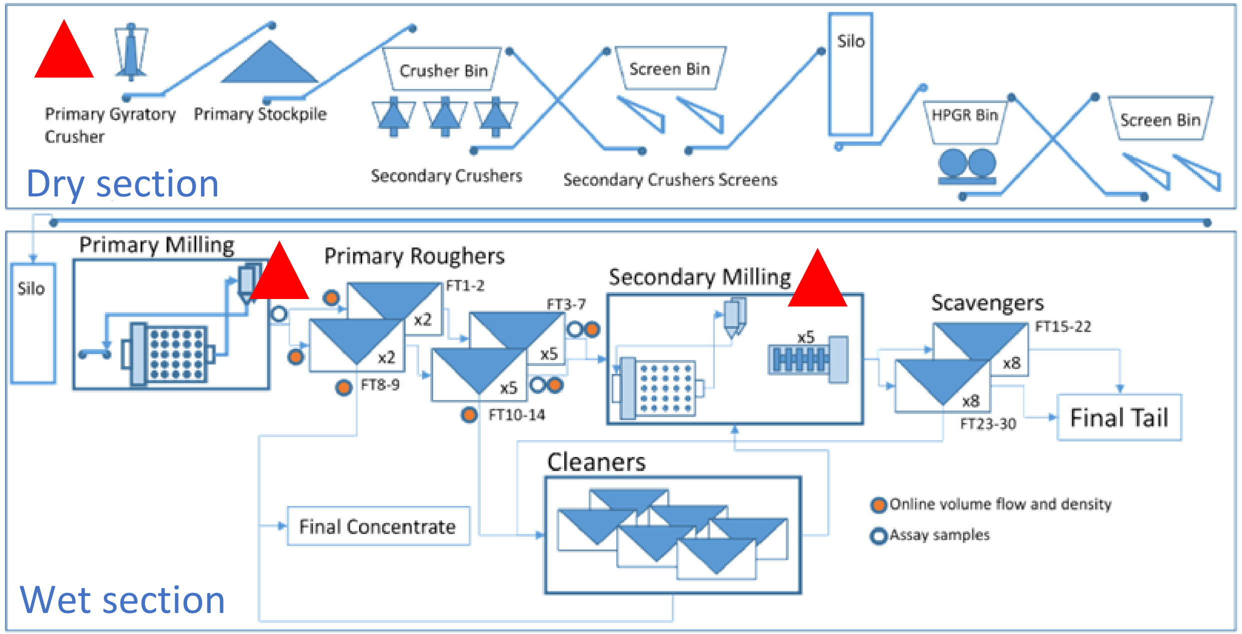
https://www.kaggle.com/edumagalhaes/quality-prediction-in-a-mining-process?select=MiningProcess_Flotation_Plant_Database.csv

Process data generation

Process plant online and offline data

Online data

Physical property sensors
(~ seconds)
E.g., volume flow rate, temperature, density, pressure
Image data
(~ seconds)
E.g., ore on conveyor belt, flotation froth



Data blind spots

Plant feed properties:
Feed grade, feed mineralization
Determines plant-wide performance; typically, least available data!

Liberation properties:
Grinding output: particle size distribution, flotation feed grade, flotation feed liberation; typically, low-frequency and/or low accuracy

Offline data

Laboratory data
(~ hours)
E.g., metal content, particle size distribution
Image data
(~ days)
E.g., microscopic grain size and colour
Text data
(~ days)
E.g., maintenance logs, reports
Mine planning data
(~ days)
E.g., modelled ore properties

Steyn and Sandrock, 2021. Causal model of an industrial platinum flotation circuit. Con Eng Prac. 109, 104736.

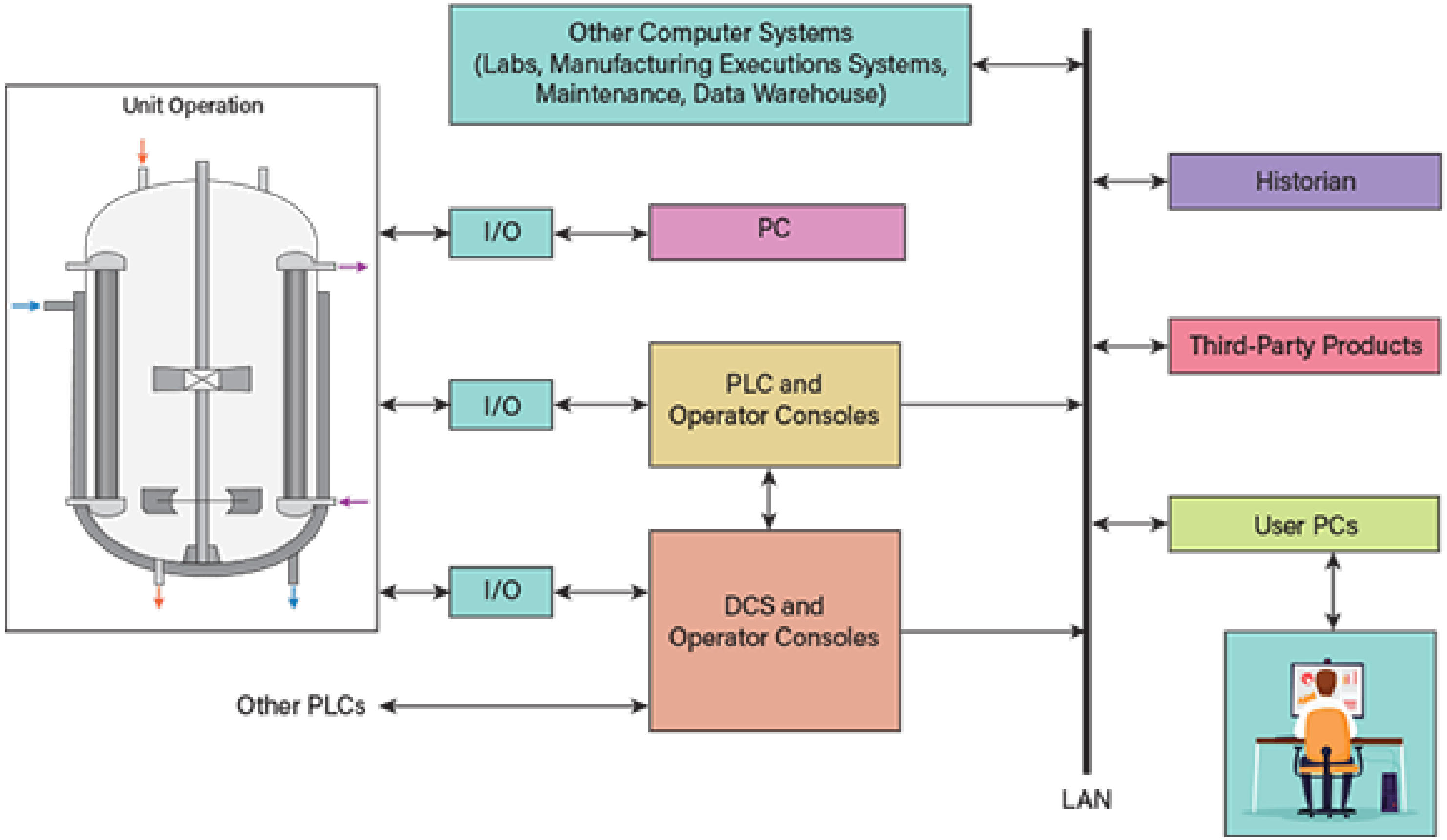
Process data generation

Control system data generation

I/O: input/output (sensors and final elements)

PLC: programmable logic controller (electronic, local focus, custom programs)

DCS: distributed control system (electronic, network, built-in control functions)



Cloud access: Remote monitoring and diagnosis

Alford and Buckbee, 2020, Industrial Process Control Systems: A New Approach to Education. AIChE.

Process data properties

Dynamic

- Plant does not operate at fixed values
- Random and systematic disturbances

Time-varying

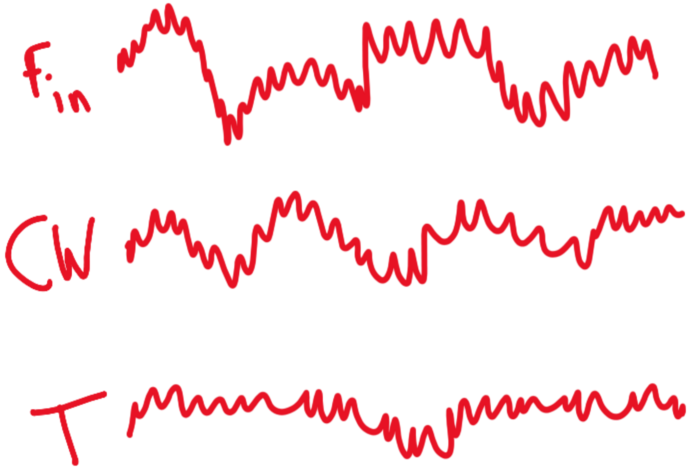
- Gradual changes in process parameters, e.g., due to degradation

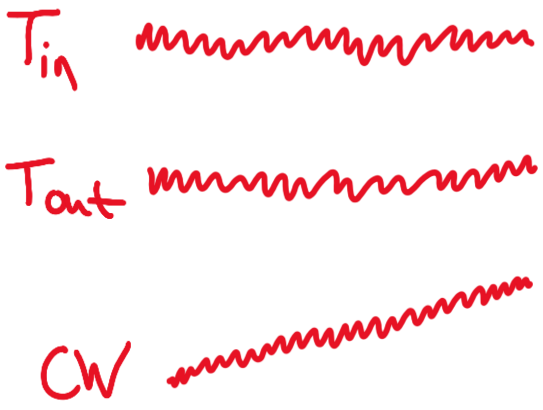
Batch vs continuous

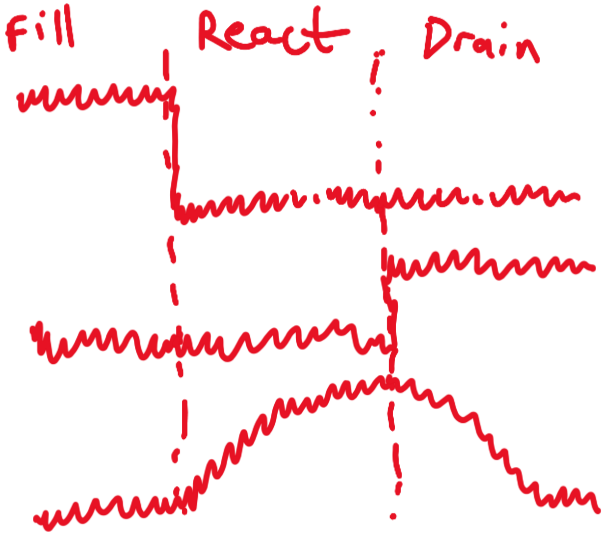
- Batch process = recipe executed over time

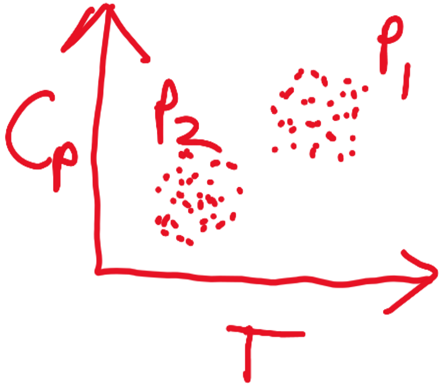
Multimode

- Switching between recipes changes distribution of data









Kumar and Flores-Cerrillo (2025) – Machine Learning in Python for Process Systems Engineering. Achieve operational excellence using process data.

Process data properties

Discrete/discontinuous

- Equipment switched on/off causing step changes

Nonlinear

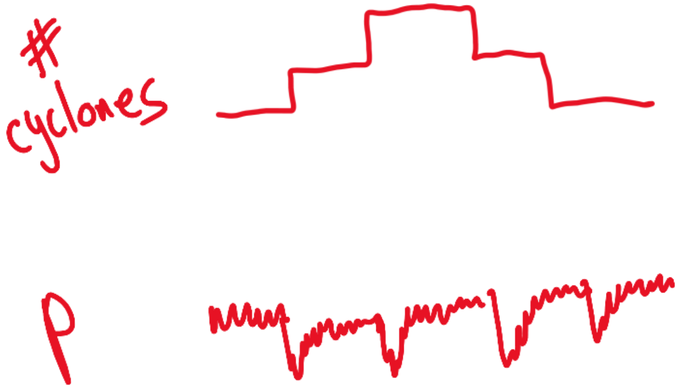
- Chemical and physical laws cause nonlinear relationships

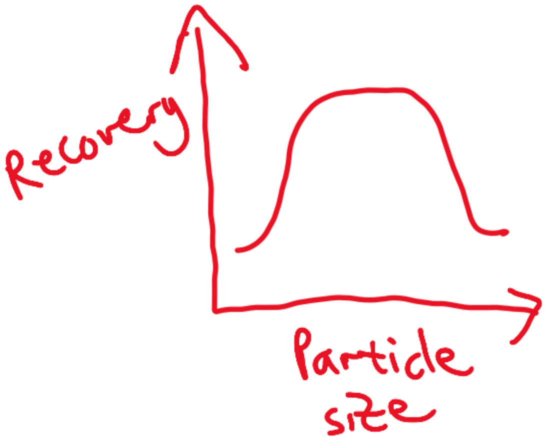
High dimensionality

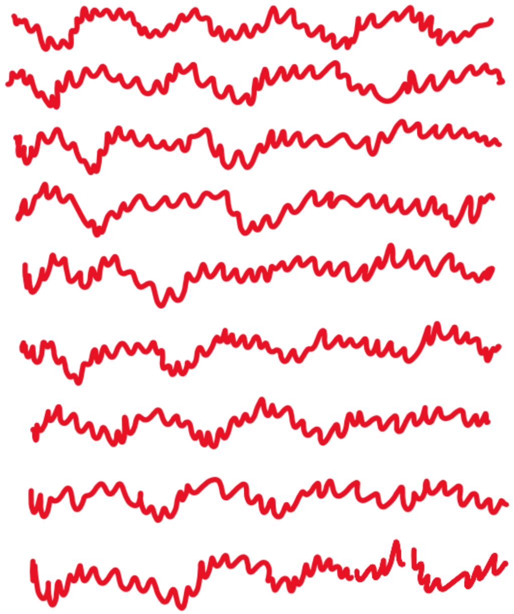
- Tens/hundreds/thousands of variables

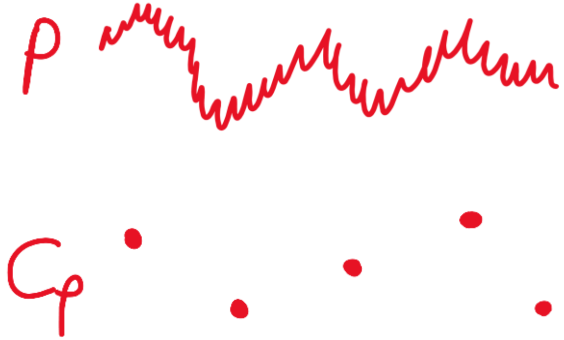
Multi-rate sampling

- Sampling frequency of measurements differ (seconds to days)









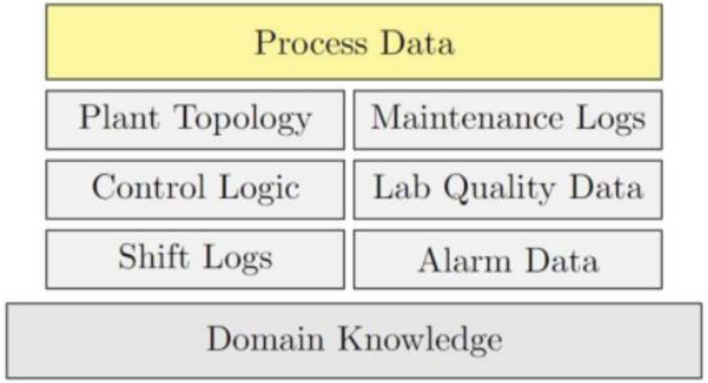
Kumar and Flores-Cerrillo (2025) – Machine Learning in Python for Process Systems Engineering. Achieve operational excellence using process data.

Process data challenges

Data retrieval and contextualization

Quantity vs quality

Industrial data comes in many formats from many sources, with additional context in terms of process layout and control system configuration



Hidden calculations: Not all data are direct measurements

B	C	D
Name	**ObjectType**	**exdesc**
PROCESS_YIELD	PIPoint	if 'FEED'>7000 then 100 * 'SIDEDRAW' / 'FEED' else 0

Lim, Elnawawi, Rippon, O'Connor, Gopulani (2023) – Data quality over quantity: Pitfalls and guidelines for process analytics. IFAC World Congress 2023.

Process data challenges

Data retrieval and contextualization

Quantity vs quality

Data quantity:
1 measurement per second
from one sensor =
32 million measurements per year

Data quality:
Continuous processes aim to operate at steady-state:
"data-rich but information poor"

Select appropriate operating regime for intended business use

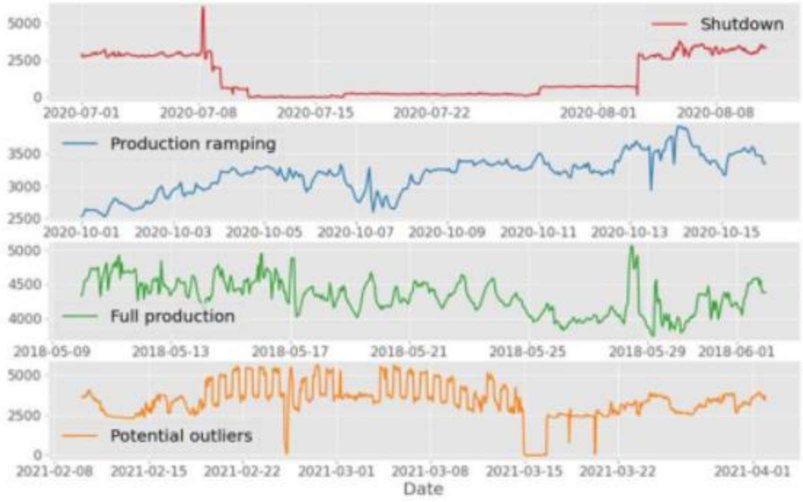


Figure 7a: Categories of operating data in one variable.

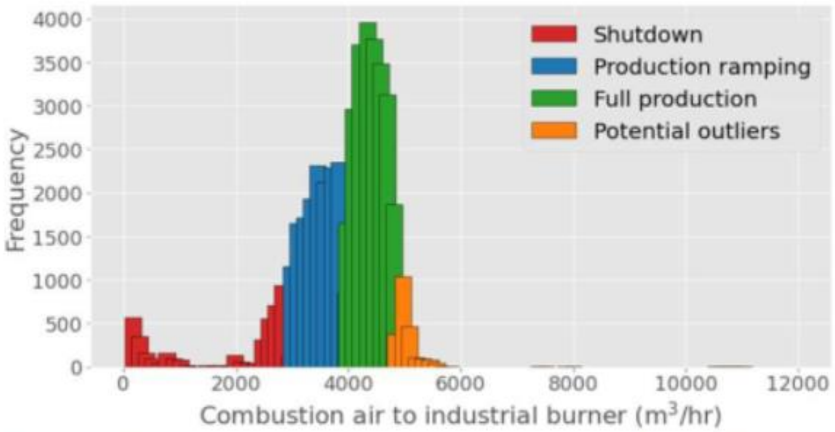
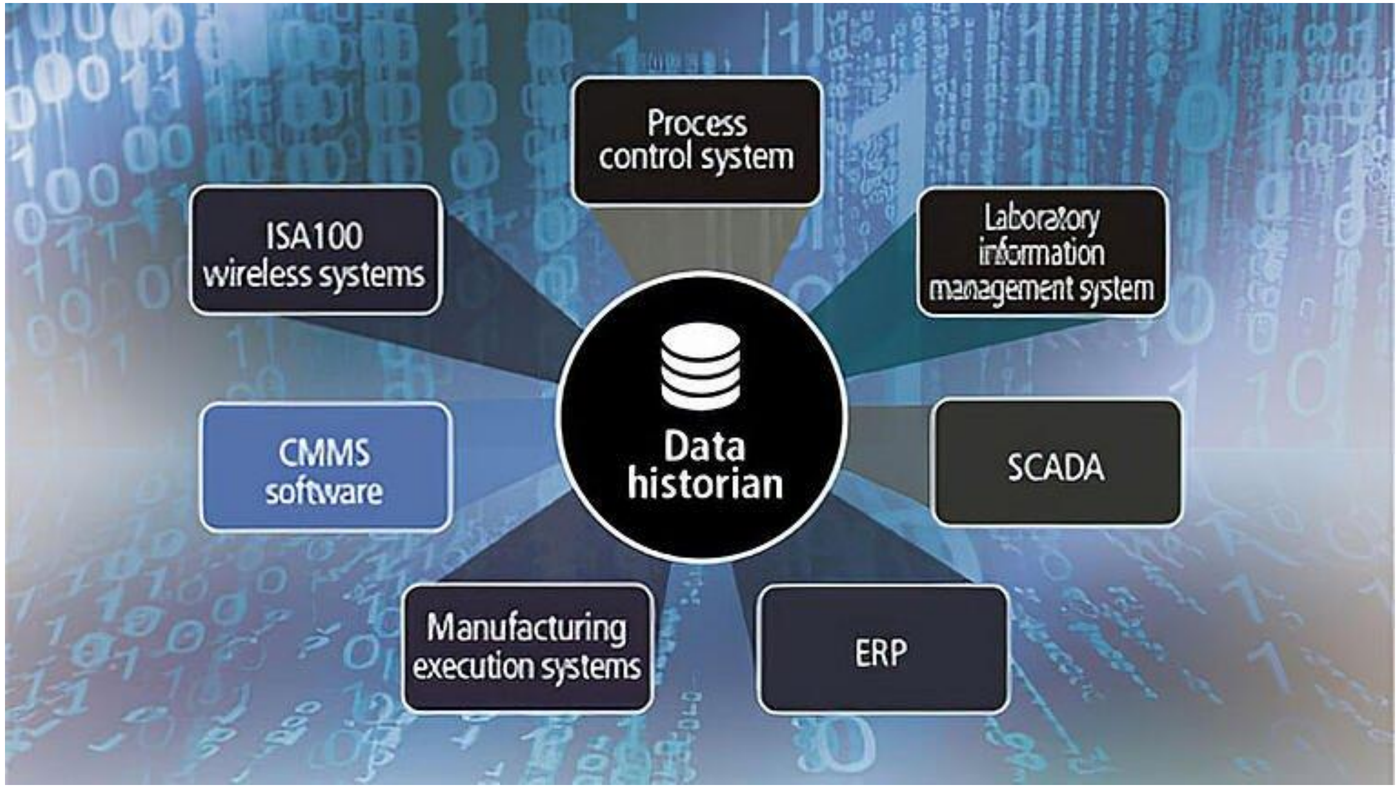


Figure 7b: Histogram of operating regimes from Fig. 8a.

Lim, Elnawawi, Rippon, O'Connor, Gopulani (2023) – Data quality over quantity: Pitfalls and guidelines for process analytics. IFAC World Congress 2023.

Data ingestion

- Various storage platforms and formats for process data
- Data historian collects, stores, and makes accessible data from various sources



https://blog.isa.org/process-historians-turn-industrial-data-actionable-information

Data ingestion

- CSV (comma-separated values) file is a common format (including as intermediary)
  - Delimited (comma, space, semicolon, etc., …)
  - Plain text

Delimiter

```
date,% Iron Feed,% Silica Feed,Starch Flow,Amina Flow,Ore Pulp Flow,Ore Pulp pH,Ore P
2017/03/10 01:02:00,55.2,16.98,3019.53,557.434,395.713,10.0664,1.74,249.214,457.396,7
2017/03/10 01:02:20,55.2,16.98,3024.41,563.965,397.383,10.0672,1.74,249.719,451.891,7
2017/03/10 01:02:40,55.2,16.98,3043.46,568.054,399.668,10.068,1.74,249.741,451.24,7,6
2017/03/10 01:03:00,55.2,16.98,3047.36,568.665,397.939,10.0689,1.74,249.917,452.441,7
2017/03/10 01:03:20,55.2,16.98,3033.69,558.167,400.254,10.0697,1.74,250.203,452.441,7
2017/03/10 01:03:40,55.2,16.98,3079.1,564.697,396.533,10.0705,1.74,250.73,444.384,7,6
2017/03/10 01:04:00,55.2,16.98,3127.79,566.467,392.9,10.0713,1.74,250.313,446.185,7,6
2017/03/10 01:04:20,55.2,16.98,3152.93,558.777,397.002,10.0722,1.74,249.895,445.985,7
```

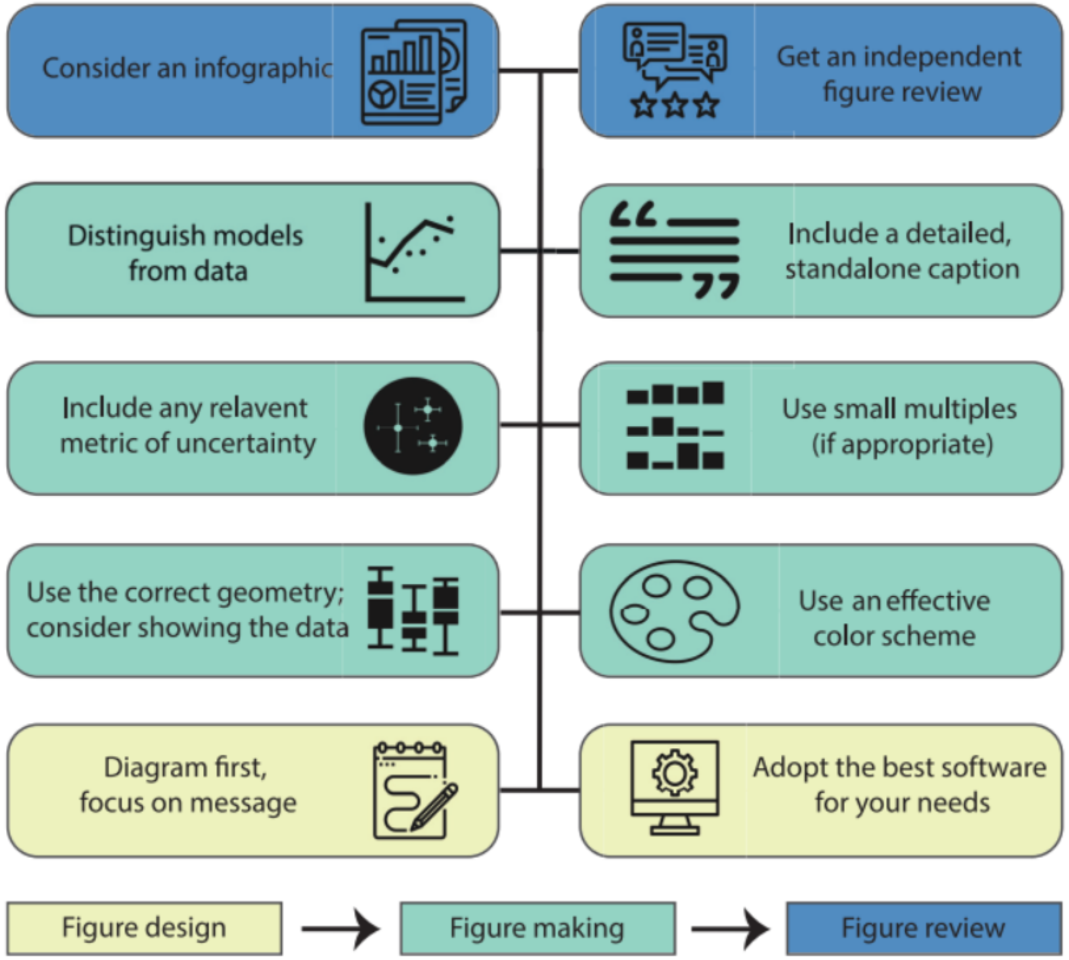
Delimiter

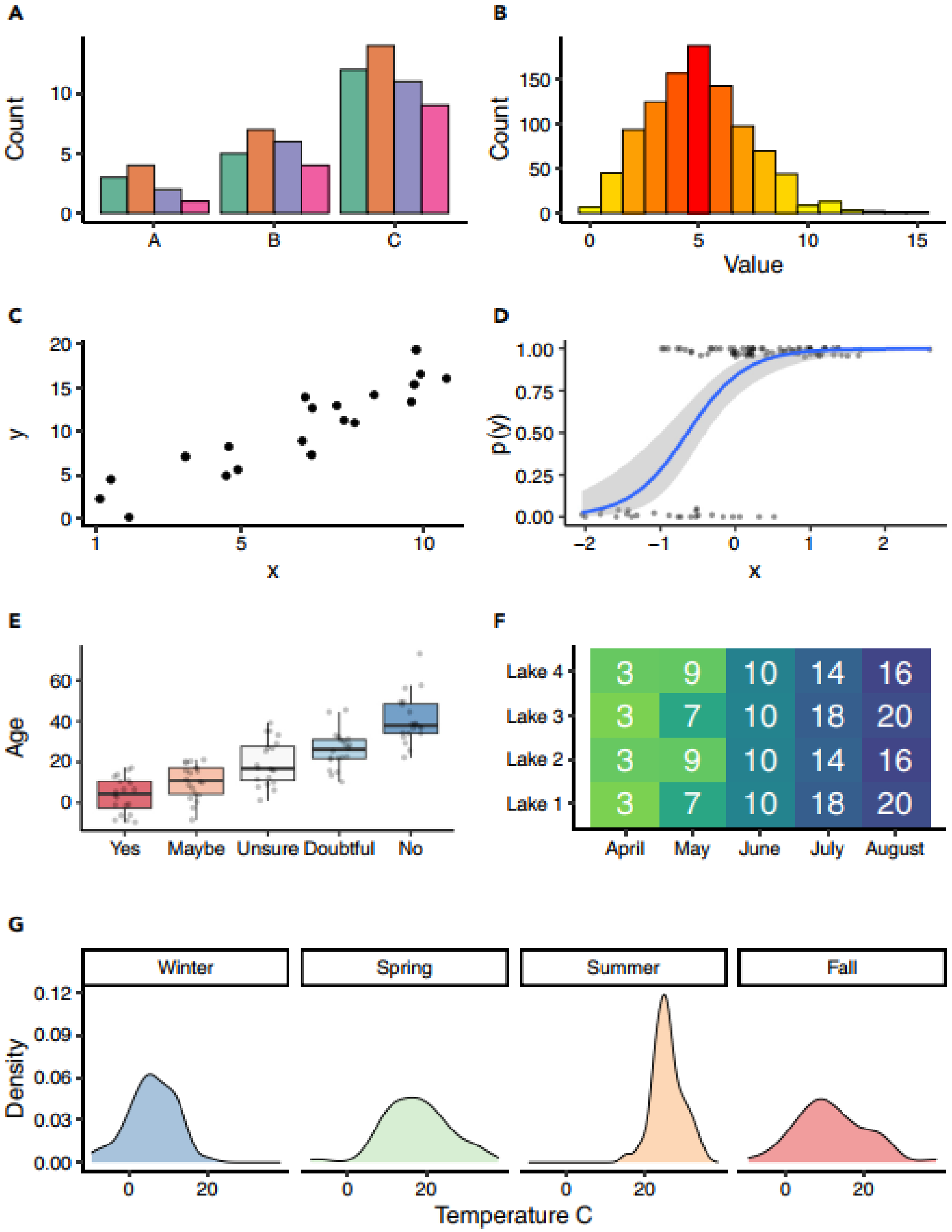
Decimal indicator

Plain text: No special/proprietary program required to open it

Data visualization

Data visualization: Principles



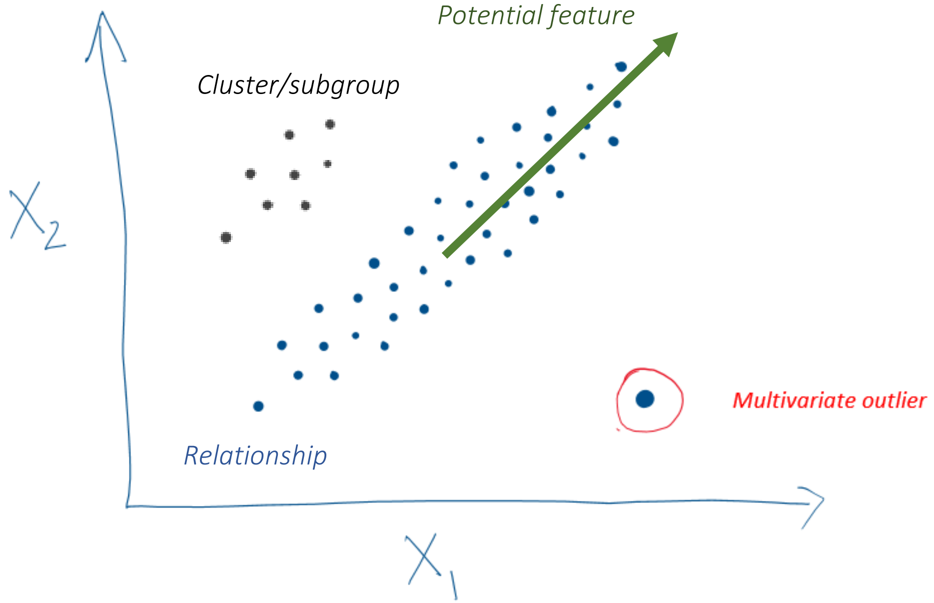


Midway (2020) Principles of effective data visualization.

Data visualization

Industrial processes produce large data sets

Data visualization aids humans to recognize patterns

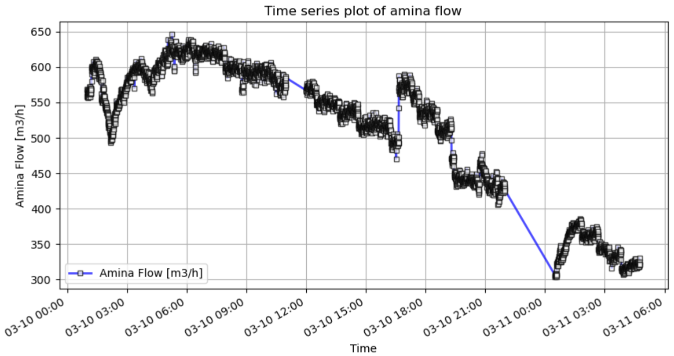


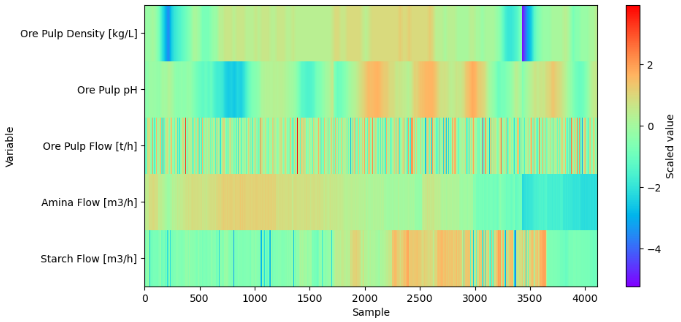
Interesting patterns:

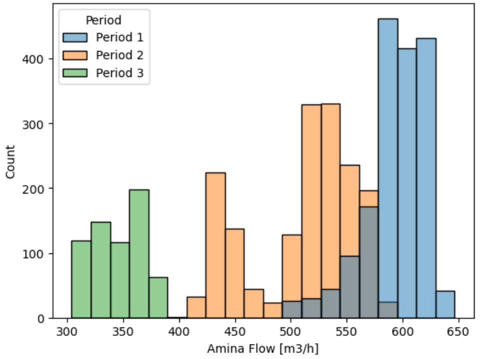
- Outliers
- Relationships
- Potential features
- Clusters/groups
- Noise levels
- Missing data prevalence

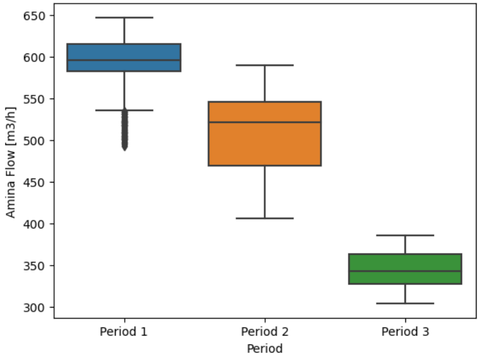
Data visualization

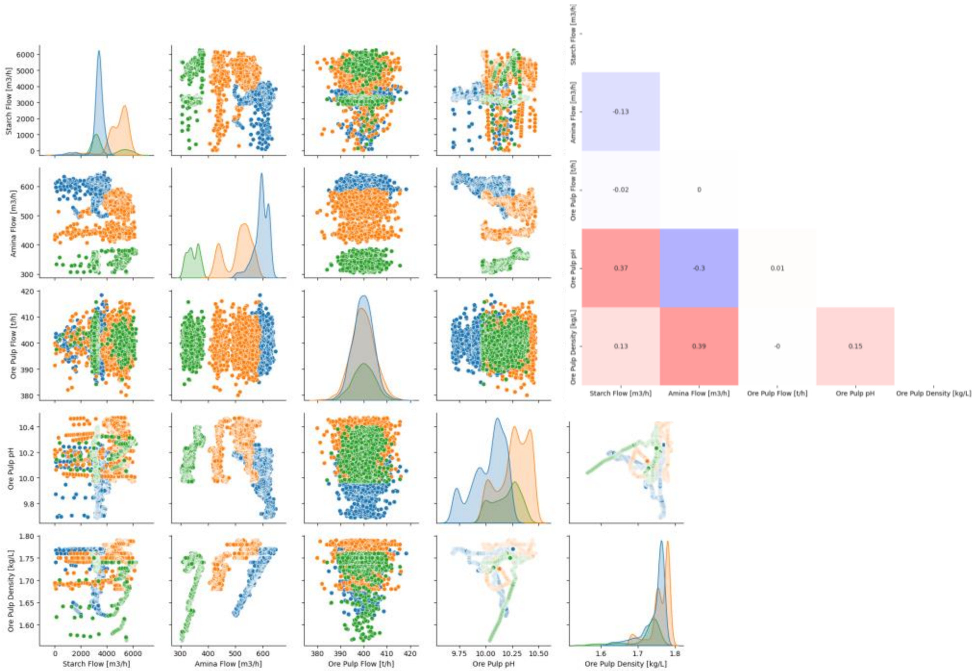
- **Time-related behaviour**
  - Time series
  - Heat maps
- **Distribution behaviour**
  - Histograms
  - Box plots
- **Relational behaviour**
  - Correlation plots
  - Scatter plots
  - Pair plots





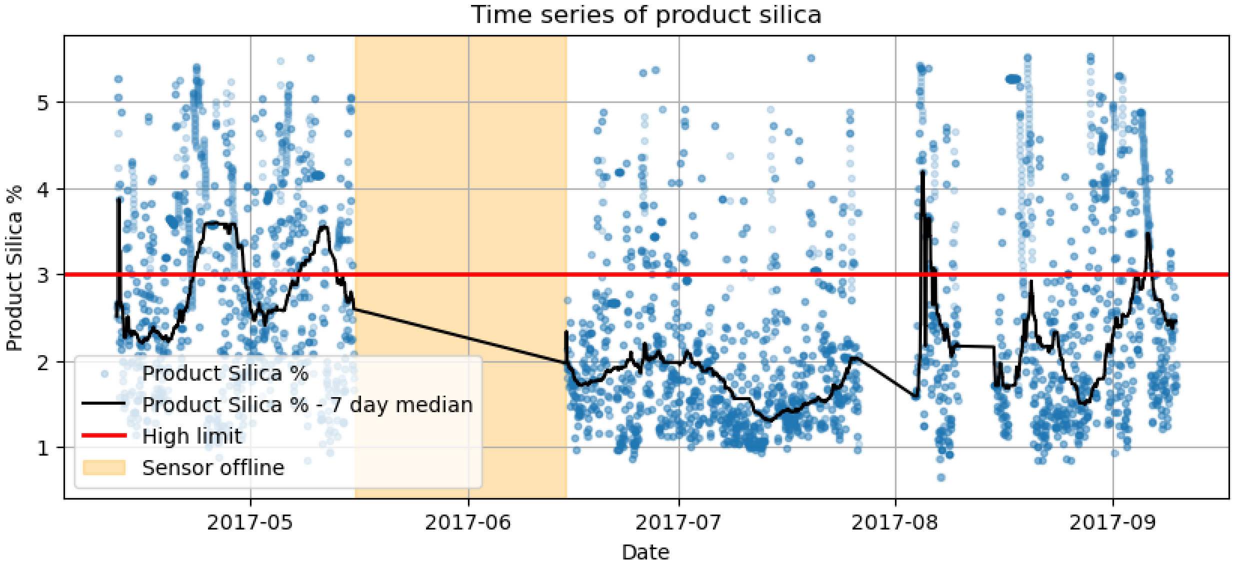






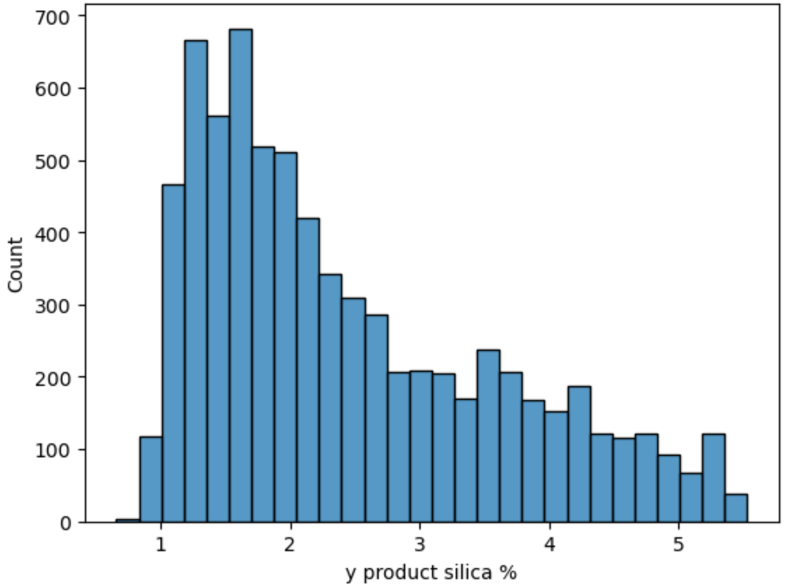
Data visualization

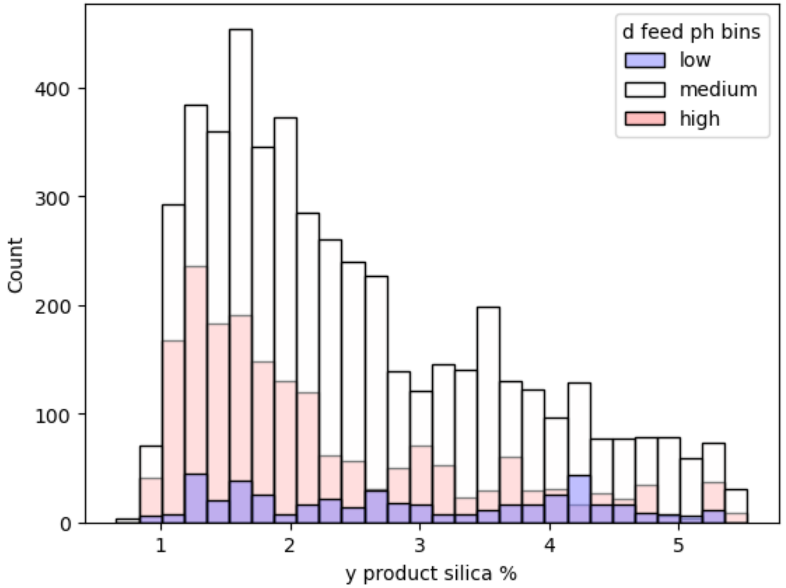
Time series plots
<u>Purpose:</u> Assess dynamic behaviour of process Identify outliers
<u>Construction:</u> X-axis: Time Y-axis: Values of one or more variables
<u>Interpretation:</u> Visual narrative of process changes Identify seasonality/periodicity Identify spikes/outliers Identify trends Assess noise levels



Data visualization

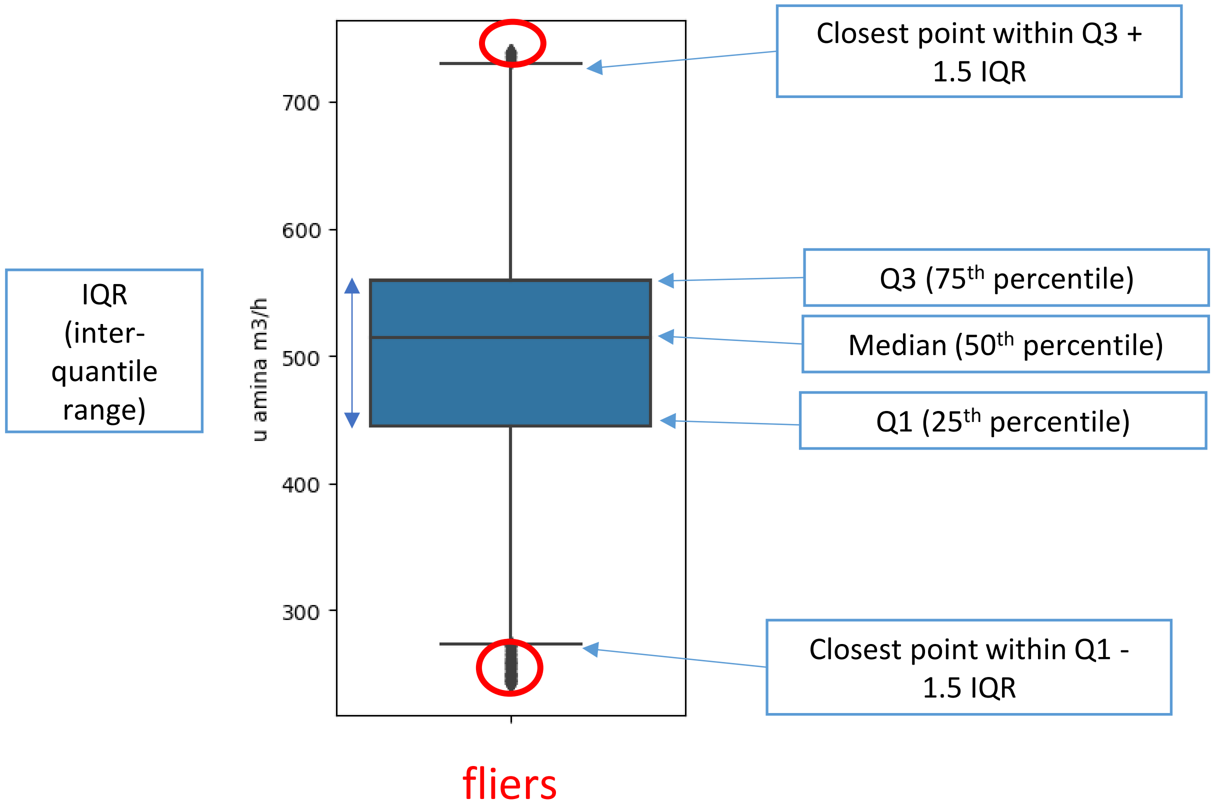
Distribution plots: Histograms
<u>Purpose:</u> Assess spread and operating modes of process
<u>Construction:</u> X-axis: Value ranges of one variable Y-axis: Frequency of occurrence of value range
<u>Interpretation:</u> Visual summary of process variability Indicate spread, symmetry Indicate grouping, extreme values





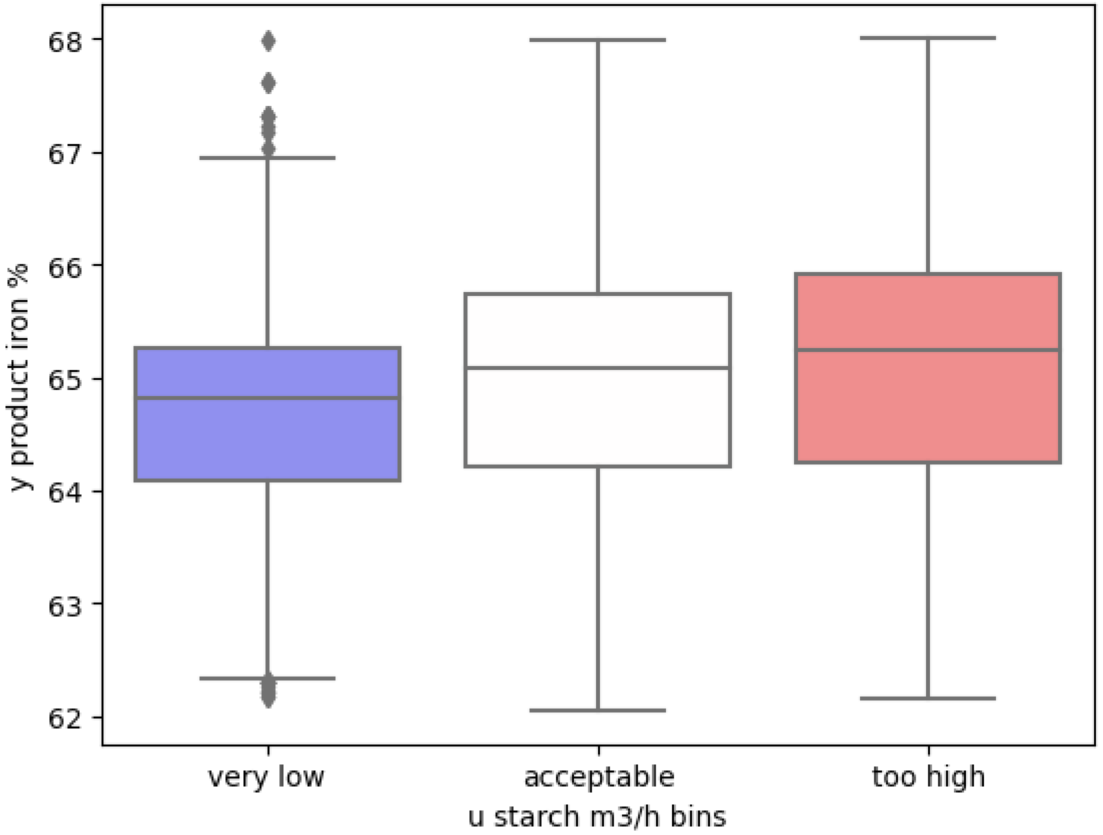
Data visualization

Distribution plots: Box-and-whisker plots
<u>Purpose:</u> Assess spread of process
<u>Construction:</u> X-axis: Categorical indicator / group Y-axis: Distribution statistics (5-number summary)
<u>Interpretation:</u> Visual summary of process variability Indicate spread, symmetry Indicate grouping, extreme values



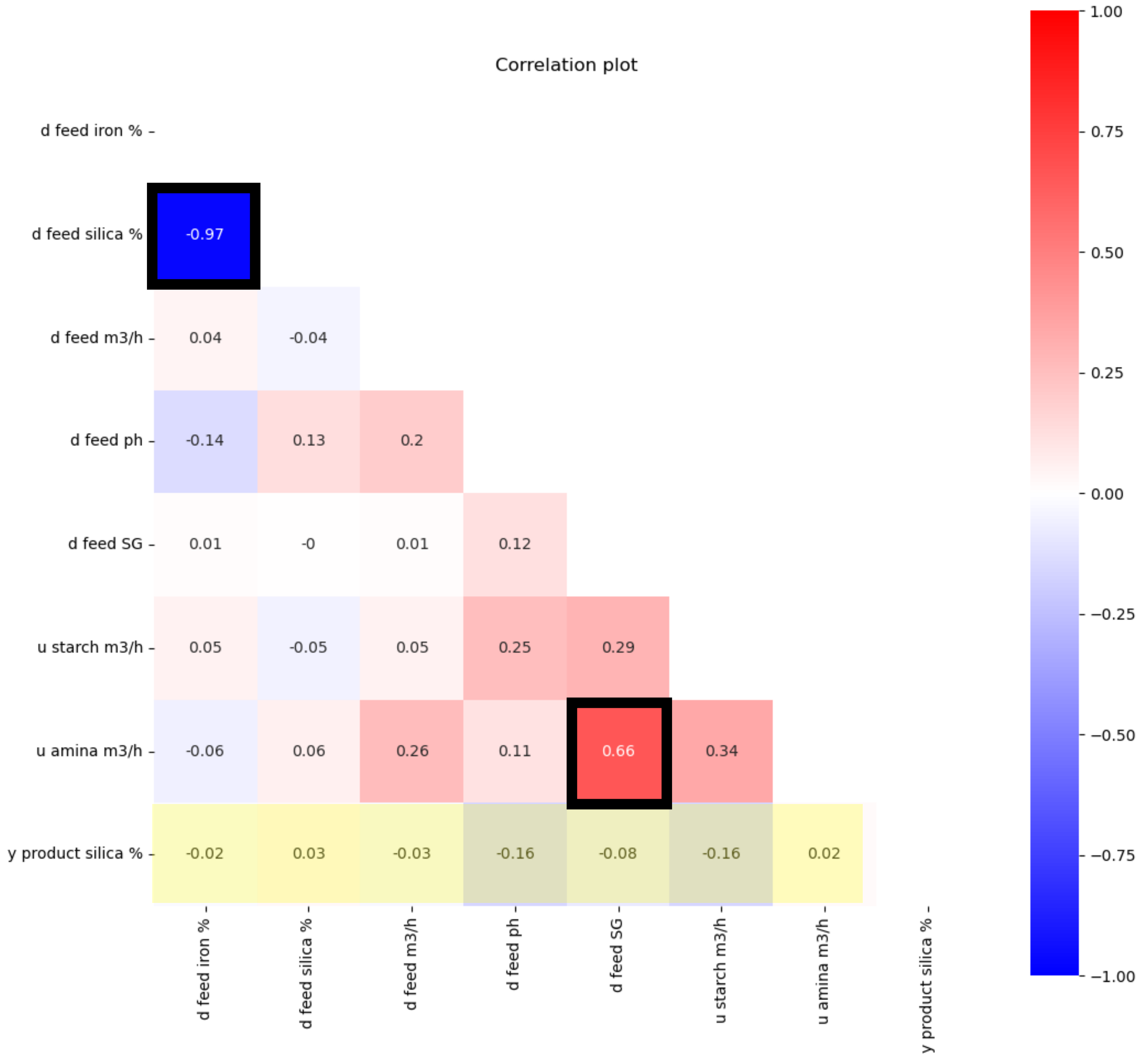
Data visualization

Distribution plots: Box-and-whisker plots
<u>Purpose:</u> Assess spread of process
<u>Construction:</u> X-axis: Categorical indicator / group Y-axis: Distribution statistics (5-number summary)
<u>Interpretation:</u> Visual summary of process variability Indicate spread, symmetry Indicate grouping, extreme values



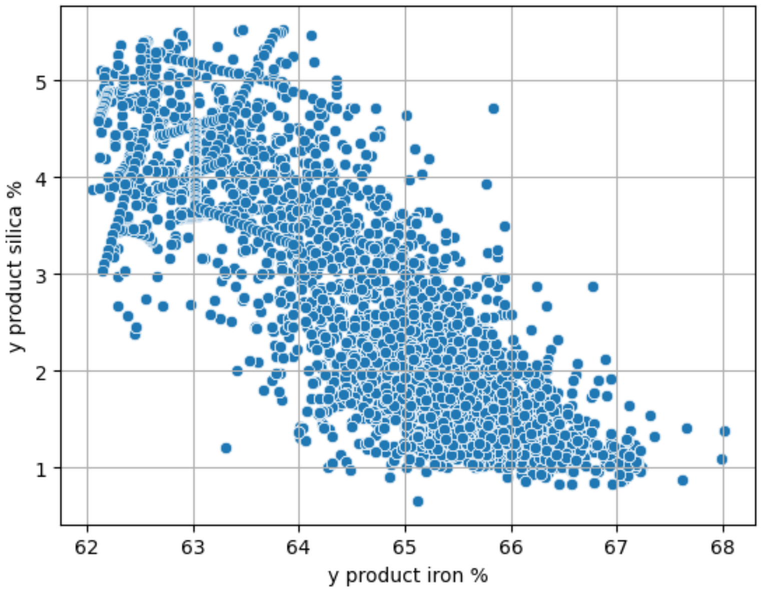
Data visualization

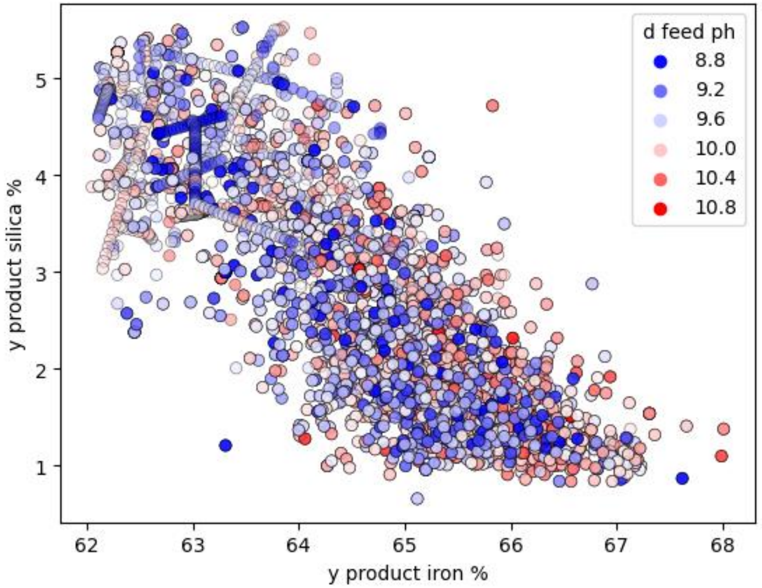
Relationship plots: Correlation heatmap
<u>Purpose:</u> Assess relationships between many variables Highlight pairs for further investigation
<u>Construction:</u> X-axis: (multiple) Selection of variables Y-axis: (multiple) Selection of variables
<u>Interpretation:</u> Succinct visual summary of pairwise variable relationships Note: Correlation is not equal to causation! Identify correspondence (none, positive, negative)



Data visualization

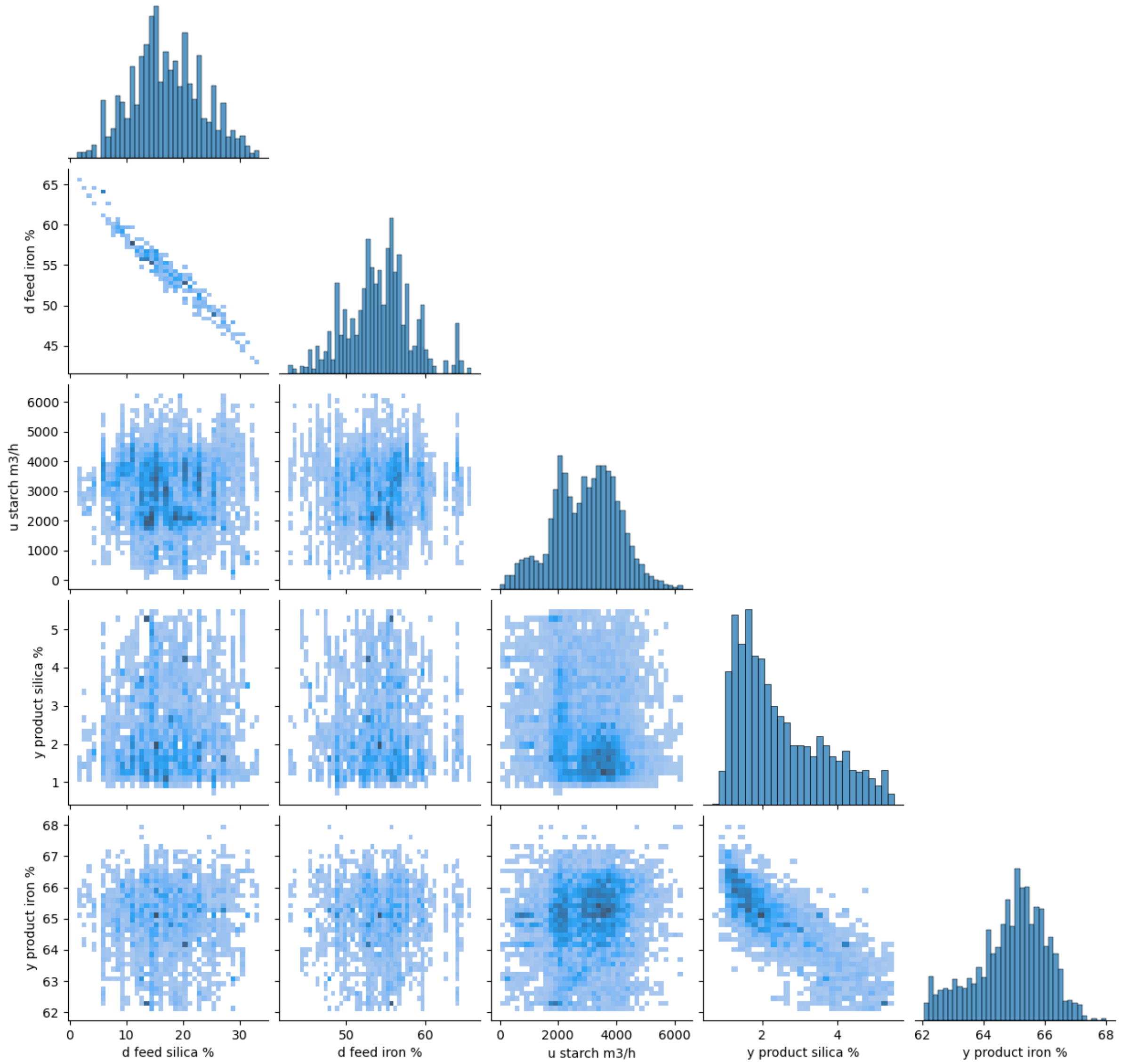
Relationship plots: Scatter plots
<u>Purpose:</u> Assess relationship between variables
<u>Construction:</u> X-axis: Variable 1 Y-axis: Variable 2
<u>Interpretation:</u> Visual summary of pairwise variable relationship Note: Correlation is not equal to causation! Identify correspondence (none, positive, negative) Identify linearity Identify groups/clusters





Data visualization

Relationship plots: Pair plots
<u>Purpose:</u> Assess relationships between many variables
<u>Construction:</u> X-axis: (multiple) Variable 1 to final Y-axis: (multiple) Variable 1 to final
<u>Interpretation:</u> Visual summary of pairwise variable relationships Note: Correlation is not equal to causation! Identify correspondence (none, positive, negative) Identify linearity Identify groups/clusters



Diagonal plots = often histograms

Data cleaning

Context of data cleaning

Data pre-processing	Model training	Model evaluation	Model tuning
• Get right data in the right form	• Estimate model parameters	• Check model's generalization	• Adjust model hyperparameters

Data cleaning
- De-noising
- Variable selection
- Outlier handling
- Missing data handling

Data transformation
- Centering and scaling
- Feature extraction
- Feature engineering

Iterative

Regularization

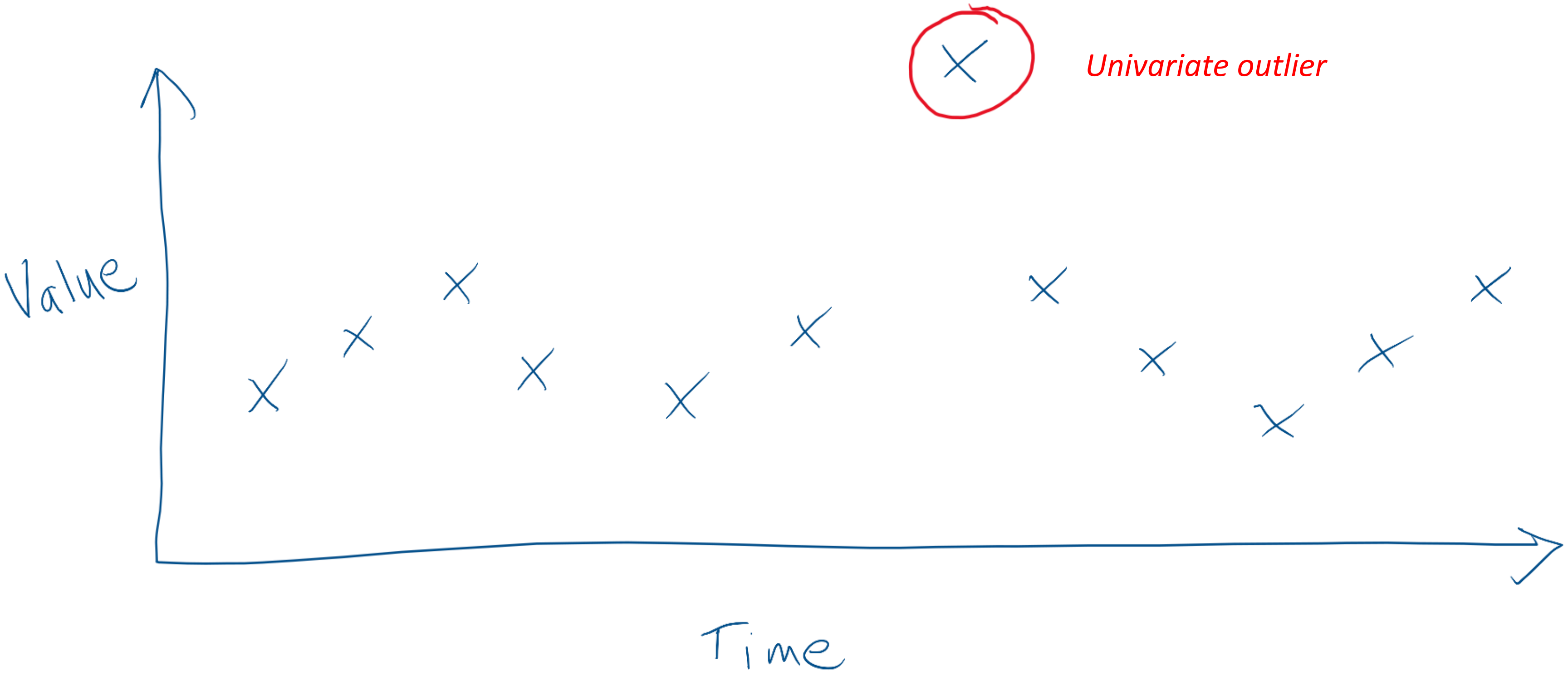
Training/validation/testing data

K-fold cross validation

Model interpretation

Data cleaning motivation

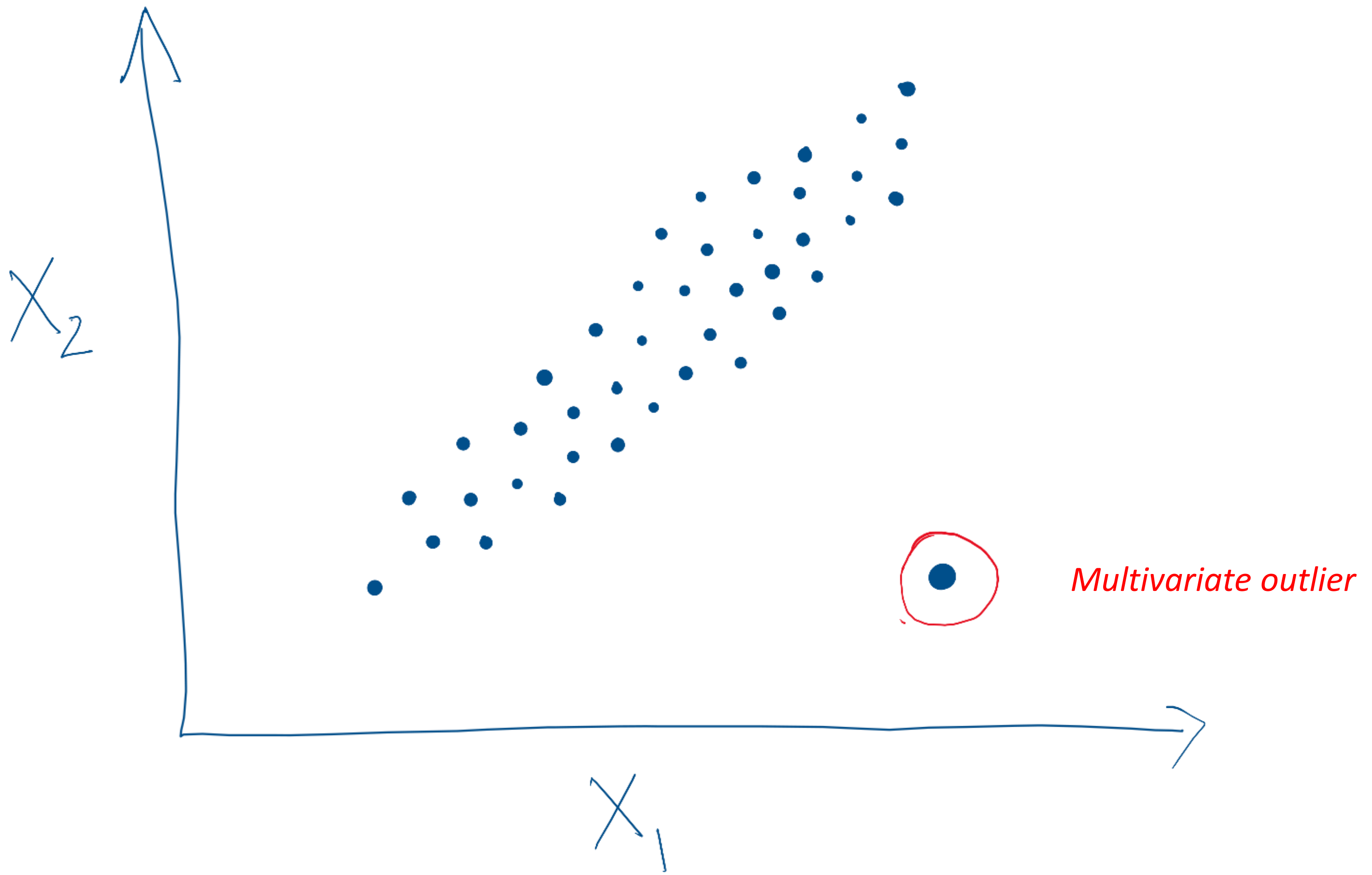
Outliers
Definition
Observations that do not show consistent behaviour with rest of data set from a statistical perspective
Causes
Sensor malfunction Inappropriate missing data handling



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Data cleaning motivation

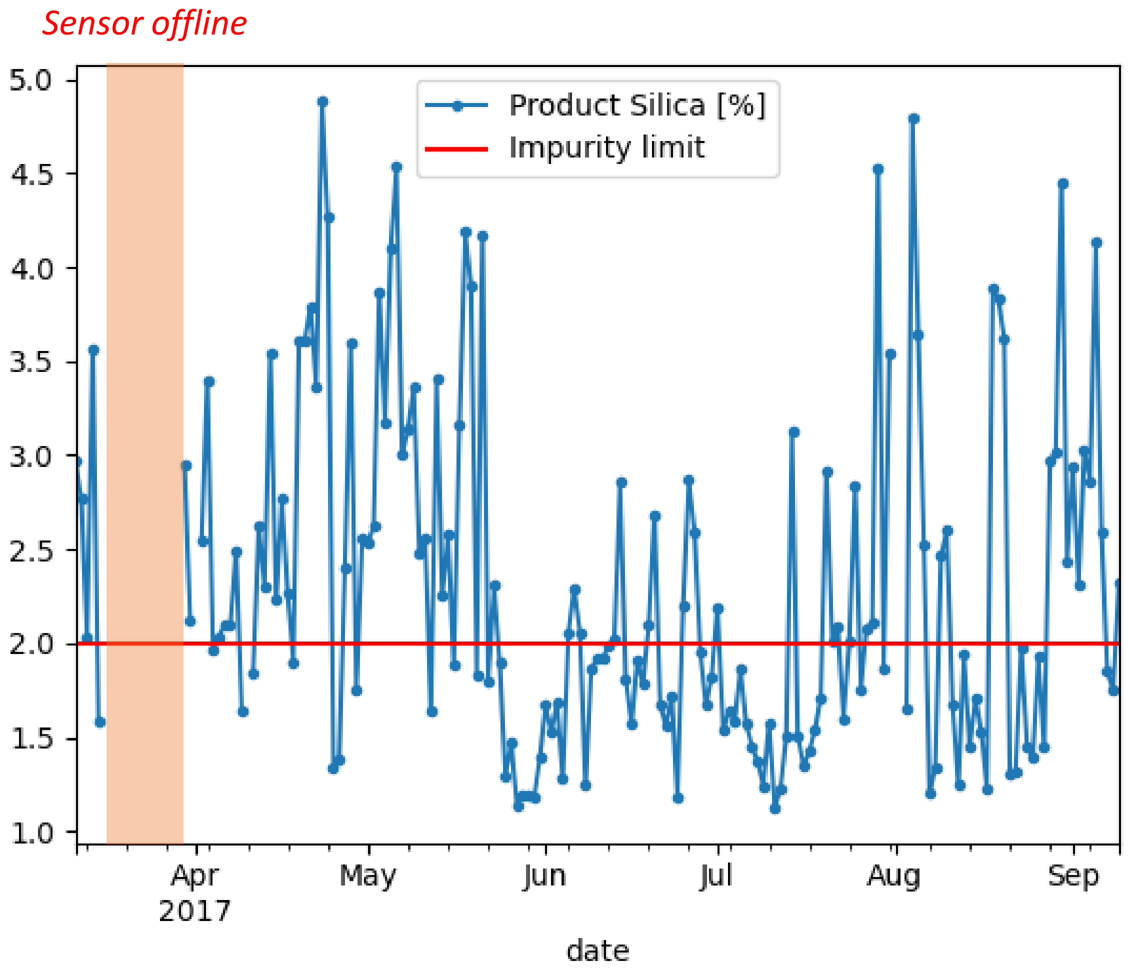
Outliers
Definition
Observations that do not show consistent behaviour with rest of data set from a statistical perspective
Causes
Sensor malfunction Inappropriate missing data handling



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Data cleaning motivation

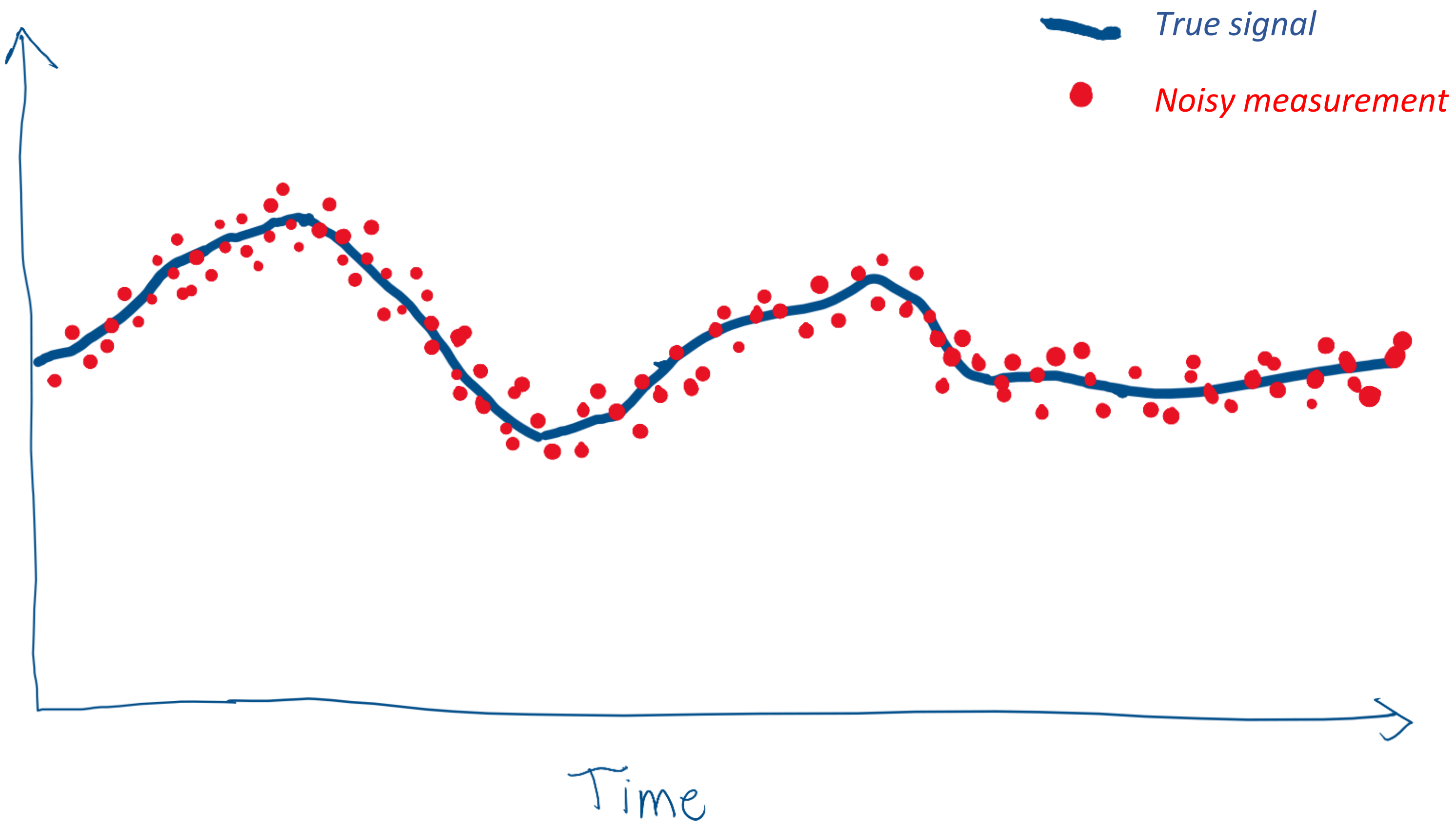
Missing data
Definition
Entries in data set that have no connection with the real state of the process
Causes
Sensor failure Fault in process unit Outlier removal Sampling rate



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Data cleaning motivation

Noise
Definition
True process signal contaminated with high frequency noise
Causes
Electronic interference Vibrations Optical interference

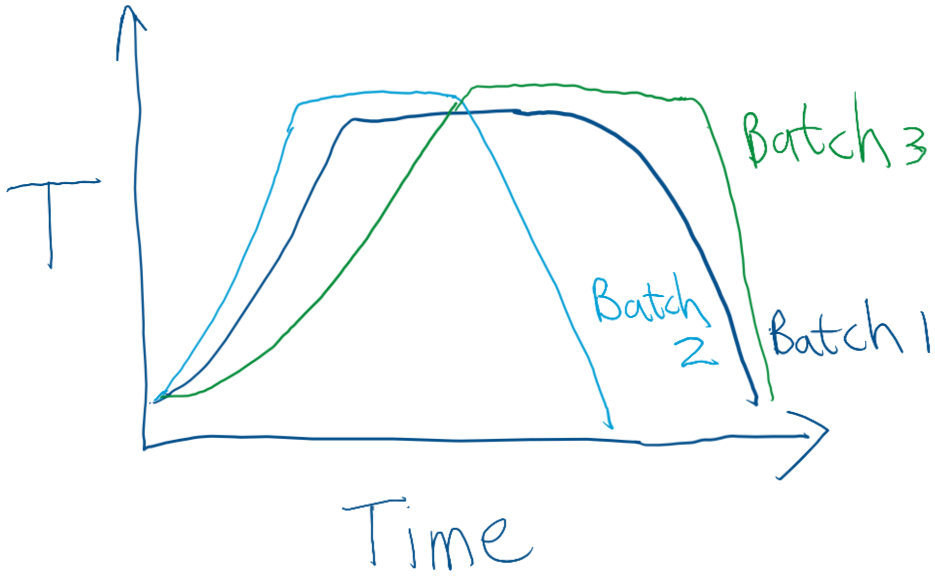


Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

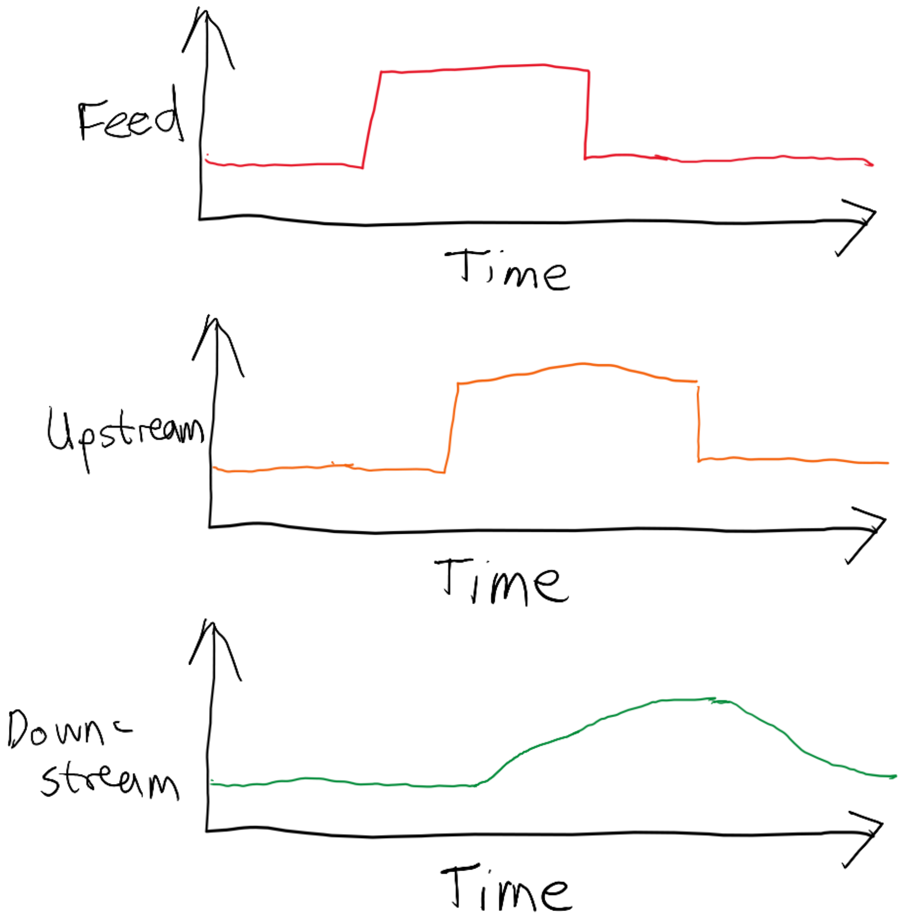
Data cleaning motivation

Time misalignment
Definition
Batch-to-batch mismatch of data OR cause-effect mismatch of continuous data
Causes
Varying batch durations Transport delays Process unit residence time Instrumentation delay

Batch misalignment



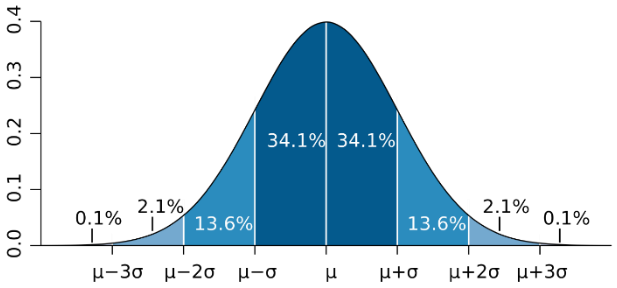
Process delays



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Outlier detection

- Knowledge-based outlier detection
- Statistical outlier detection



Knowledge-based outlier detection

Process knowledge provides insight in terms of minimum and maximum allowable values

E.g., negative values for flow not possible
E.g., if goal is to model behaviour of process unit under acceptable operating conditions, then extreme operating conditions can be considered as outliers

Statistical outlier detection

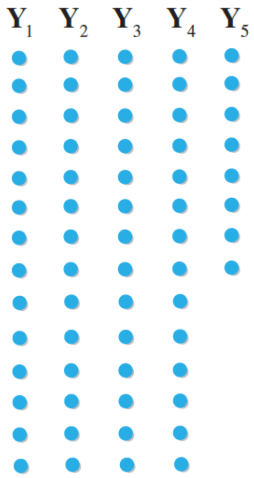
Univariate detection: **3σ rule**
Given:
- Measurement observation x_k
- Sample mean $\bar{x}$ (approximation of μ)
- Sample standard deviation (approximation of σ)

Rule:
If $|x_k - \bar{x}| > 3s$
Then x_k is an **outlier**

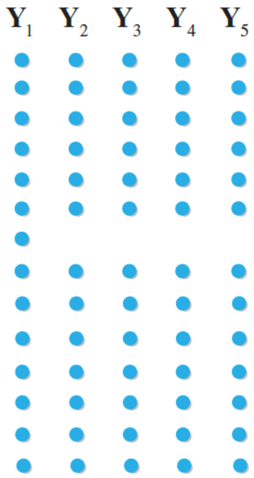
Missing data

- A data point is missing if no value is reported for a specific time stamp for a specific variable
- Understanding the cause of missing data (random or not) is important

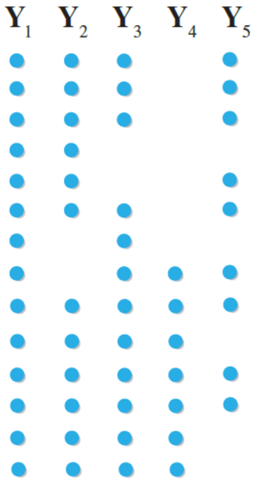
One variable with missing values (e.g., single sensor failure)



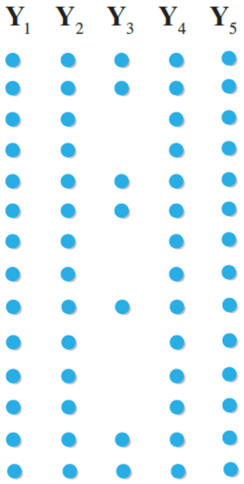
Associated variables with missing values for same time stamps (e.g., fault in process unit)



Irregular missing values (e.g., outlier removal, sensor malfunction)



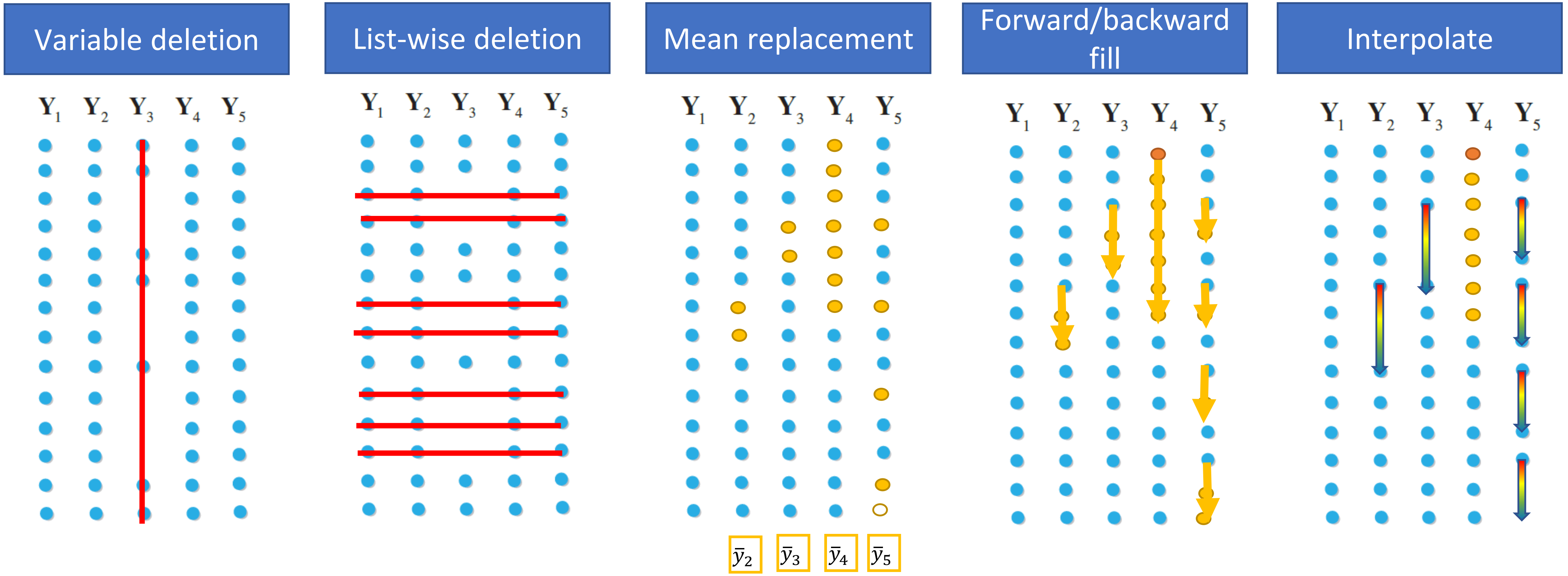
One variable with regular patter of missing values (e.g., multi-rate sampling)



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Missing data

- Missing data handling and imputation

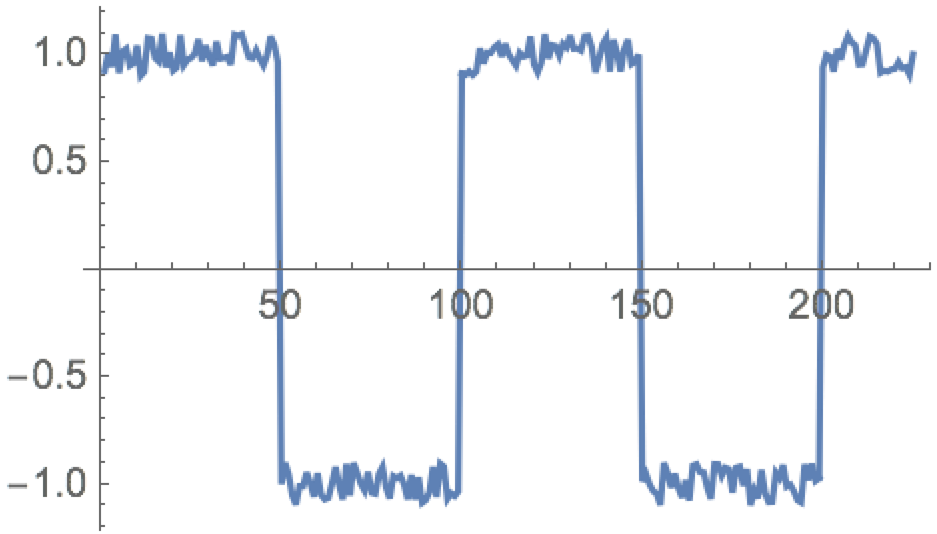


Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

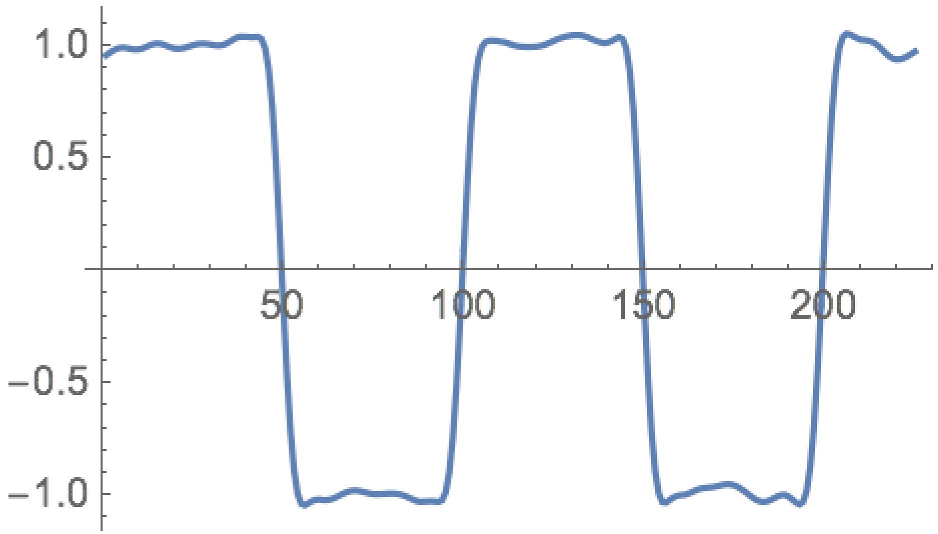
Noise removal

- Sensor measurements are subject to high frequency noise
- Filtering aims to remove high frequency noise while preserving low frequency signal

Low frequency signal +
High frequency noise



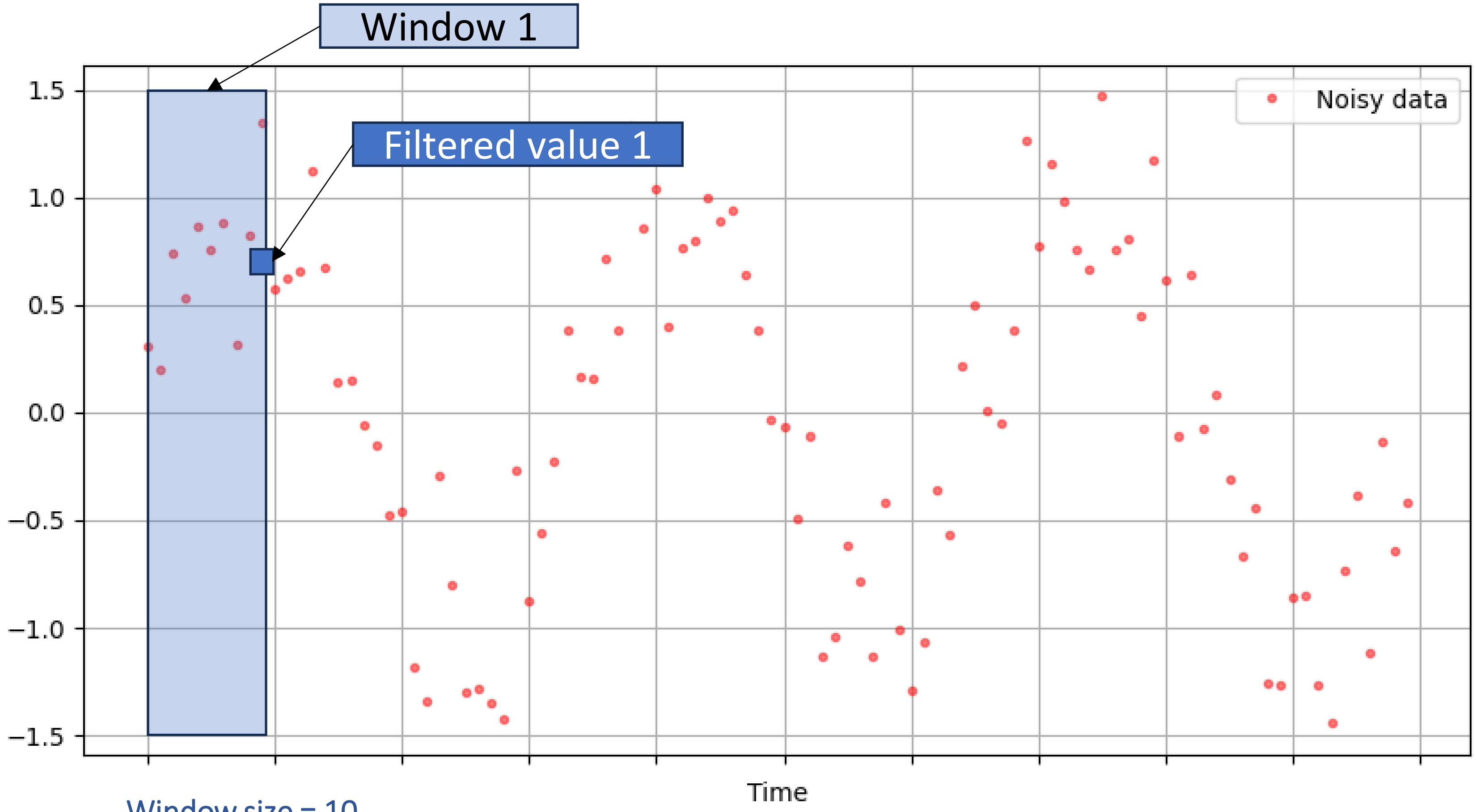
Filtered measurement



Xu et al. (2015) Data cleaning in the process industries. Reviews in Chemical Engineering.

Noise removal

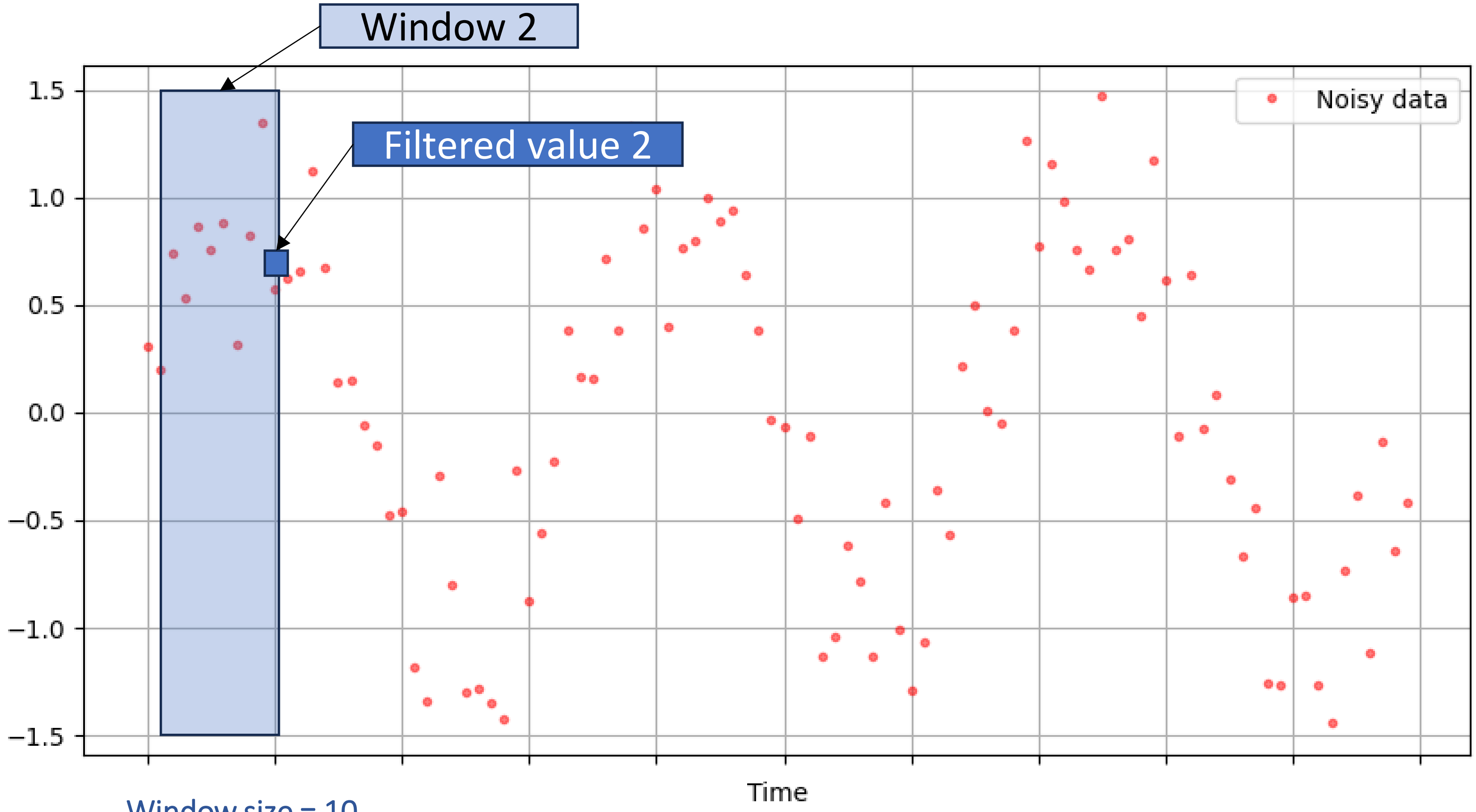
- Rolling window noise removal



Window size = 10
Calculation = median

Noise removal

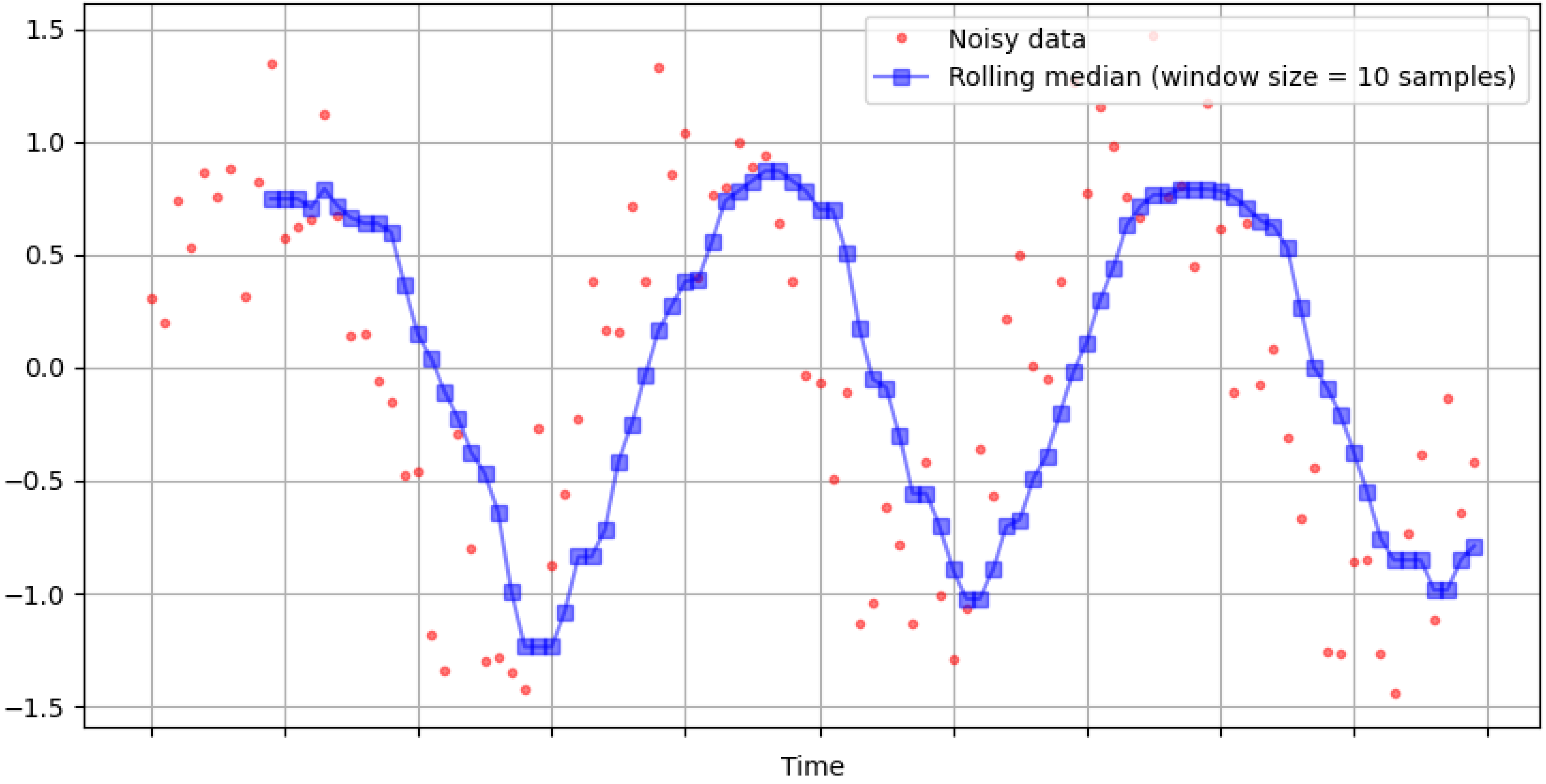
- Rolling window noise removal



Window size = 10
Calculation = median

Noise removal

- Rolling window noise removal



Noise removal

- Moving average filter

$$y_j = \frac{\sum_{i=0}^{N-1} x_{j-i}}{N}$$

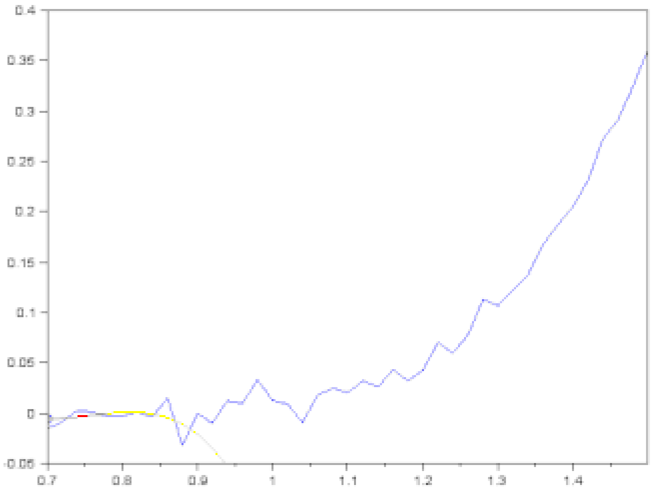
- Exponentially weighted moving average filter

$$y_j = \alpha x_j + (1-\alpha) y_{j-1}$$

- Savitzky-Golay filter

$$y_j = \sum_{i=\frac{1-m}{2}}^{\frac{m-1}{2}} C_i y_{j+i}$$

E.g.: $m = 5: C_i = -\frac{3}{35}, \frac{12}{35}, \frac{17}{35}, \frac{12}{35}, -\frac{3}{35}$



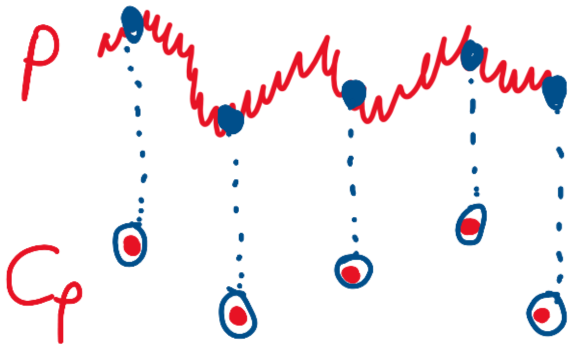
[Savitzky-Golay filter (Wikipedia)]

Resampling

Addressing mismatch in sampling frequencies

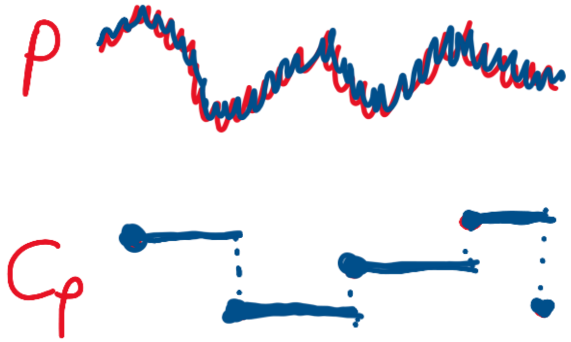
- To model relationship between variables, their sampling frequency should be similar
- Some <u>easy-to-measure</u> properties are available at high frequency (e.g., flow measurements at second intervals)
- Some <u>hard-to-measure</u> properties are available at low frequency (e.g., assays at day intervals)

Downsampling | **Upsampling**



<u>easy-to-measure</u>

<u>hard-to-measure</u>



StoneThree

The Future of Work. Now.™

+27 21 851 3123

info@stonethree.com

24 Gardner Williams Avenue

Paardevlei Somerset West

South Africa 7130

www.stonethree.com