Sargent Centre for Process Systems Engineering

Industrial Data Science Workshop (L1)

19-21 September 2023

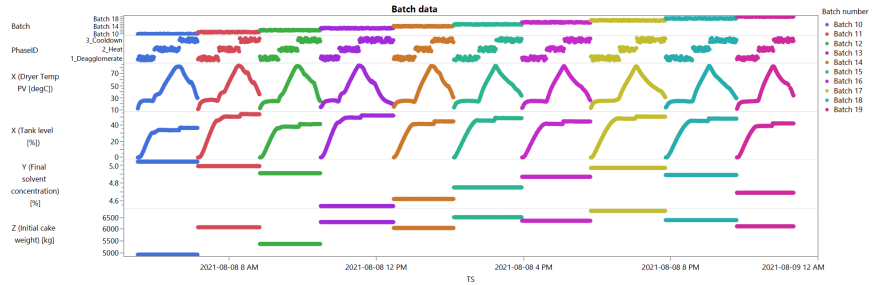
Dr. Mattia Vallerio
Manufacturing Excellence Site Manager at Solvay and
Advanced Process Control Specialist [in]

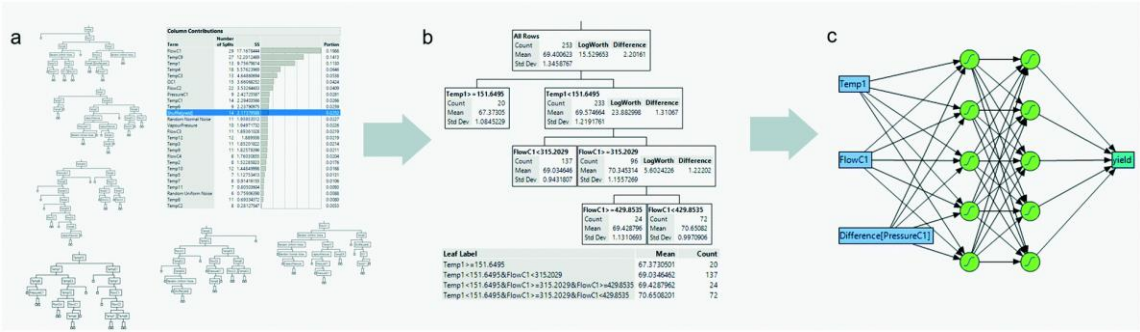
Dr. Carlos Perez
Industrial Data Scientist at Solvay and
Optimization Specialist [in]

Dr. Francisco Navarro
Sr. Data Science Training Lead at IFF and
Visiting Researcher at Imperial College London [in]

Agenda

- Day 1 – Industrial data science
  - Machine learning for process engineers
  - Industrial data and process control
  - Hands-on exercises
- Day 2 – Monitoring and screening
  - Identify relevant process changes
  - ML applied to batch processes
  - Hands-on exercises
- Day 3 – Modeling and understanding
  - Predictive modeling
  - Industrial examples
  - Hands-on exercises





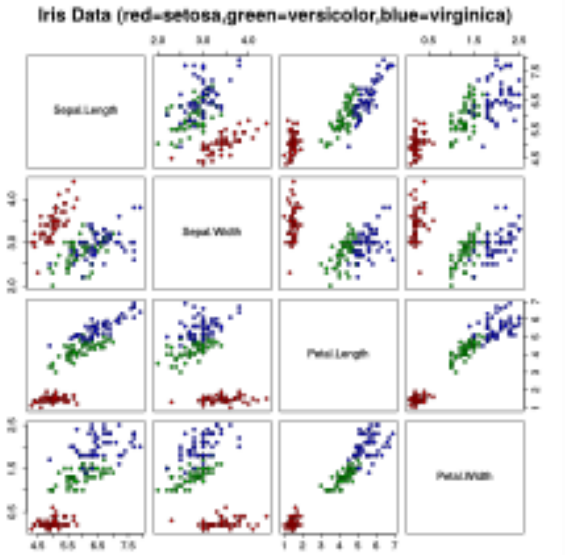
Sargent Centre for Process Systems Engineering

Day 1
Industrial Data Science

Manufacturing needs still are safety, efficiency, product quality & process reliability.

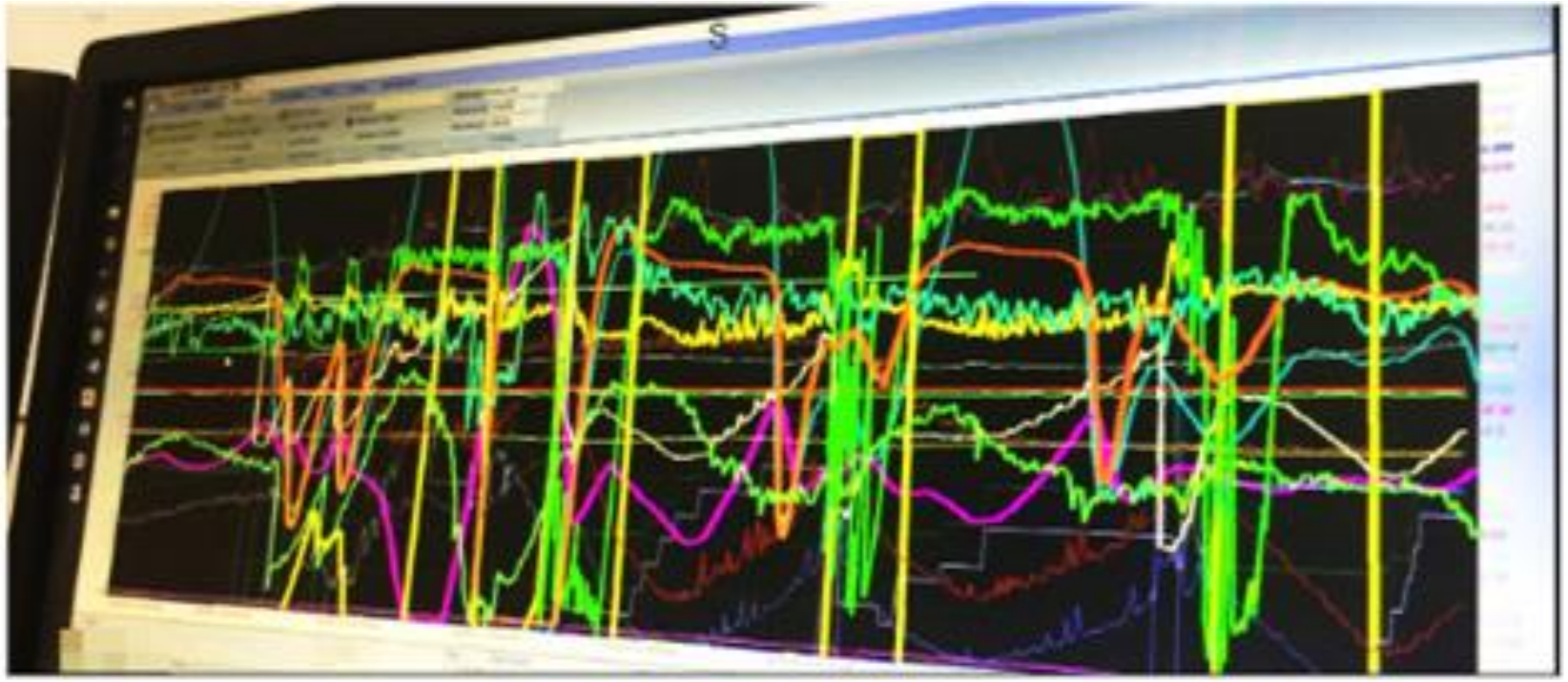
(and not video recommendations or predicting types of flowers)

Machine learning is (often) taught with irrelevant datasets



101 ML example:
Types of flowers

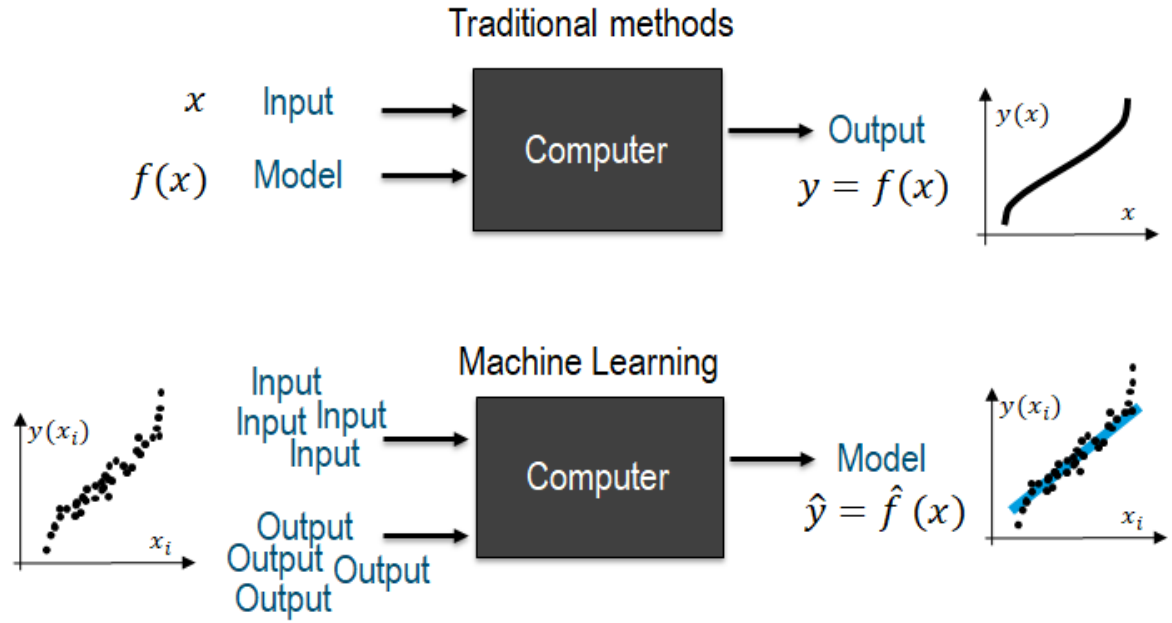
?

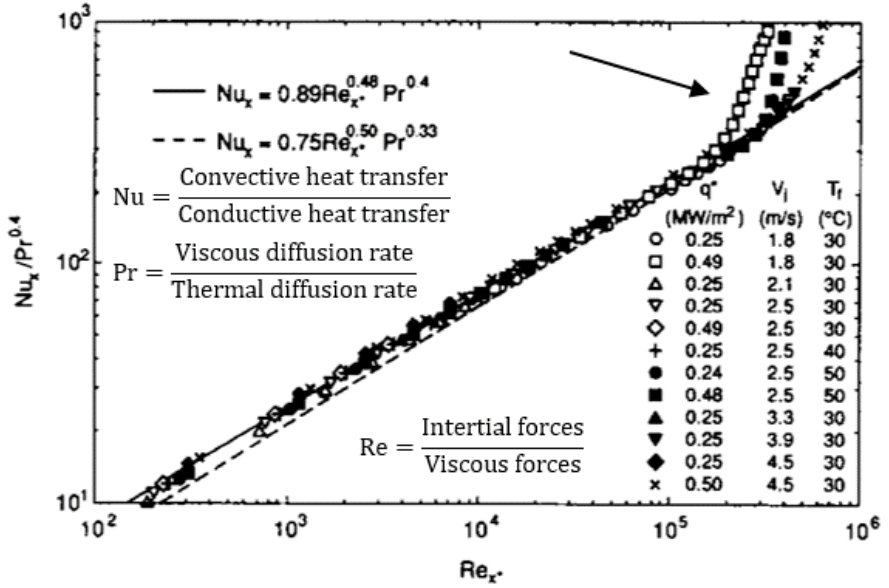


Tags in process historian (sensor data)

Image credit:
Scaling up the Use of Machine Learning in Chemical Process Industries (2023-EU-30MP-1346)

Good news, process engineers already know ‘machine learning’

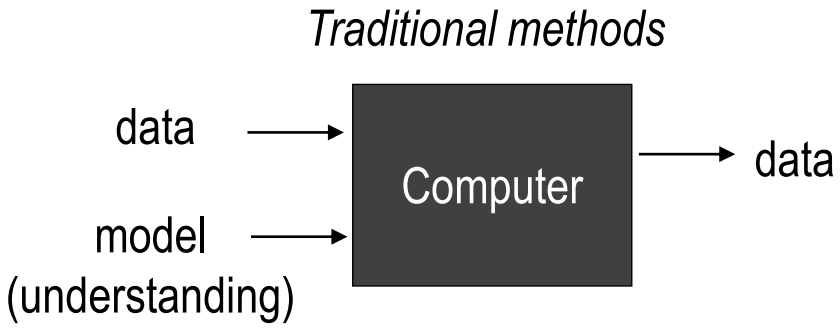


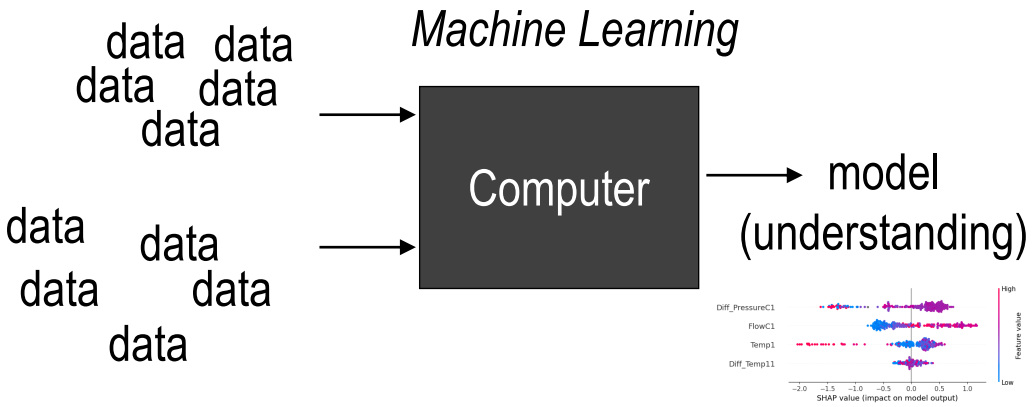


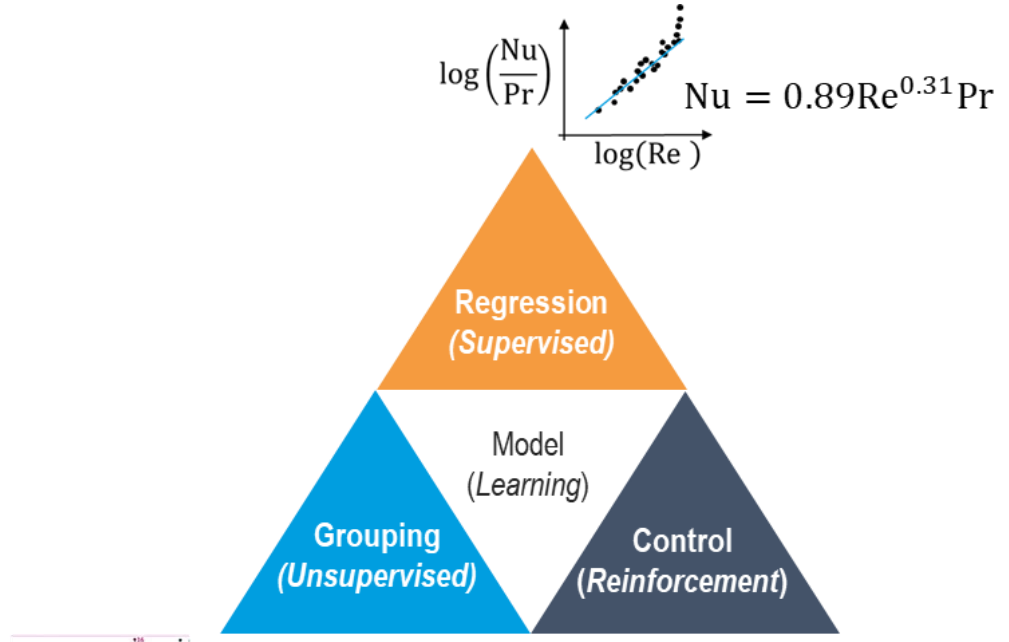
Traditionally, dimensionless numbers (Re, Pr, Nu, etc.,) and non-linear models were used to estimate heat or mass transfer coefficients from process data. In machine learning, these steps can be named feature engineering, feature selection, dimensionality reduction and regression (supervised learning). The risk of extrapolation is known, as models are specific to similar systems and operating conditions

Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

Process engineering and industrial data science in chemical manufacturing

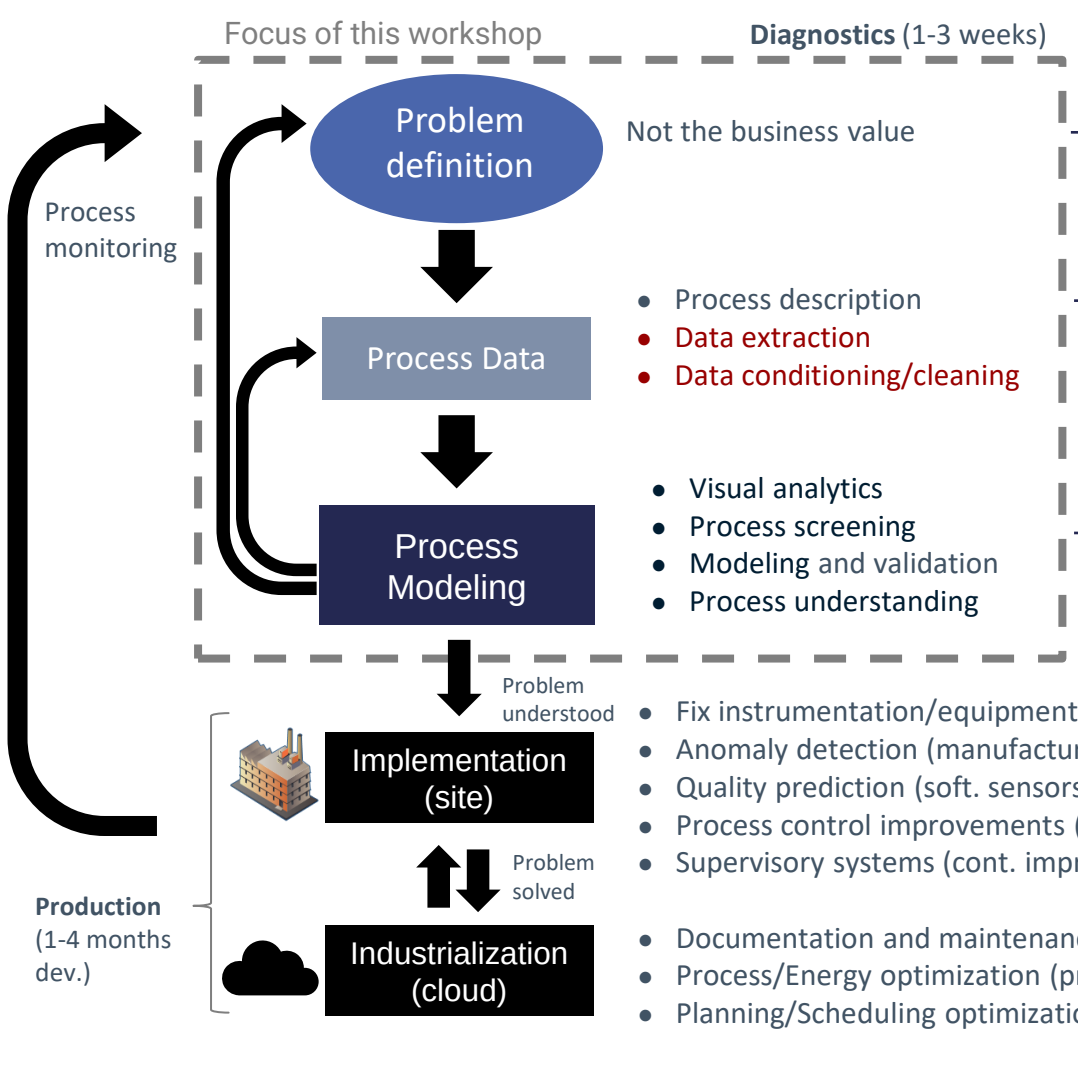






Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

Data-driven decisions require process knowledge (and time)



Scalability issues

Manual OEE tracking (lack of data integration and MES)… Spreadsheets

Data access/quality: historian (limited data access, lack of events or assets, data dictionaries), multiple data sources (e.g., lab data)… Missing OT configuration

Domain experts with limited time, no self-service access to data, with multiple tools and minimum access to training and support (e.g., process engineers cannot easily monitor processes and perform golden batch analysis)

Limited/slow access to resources to implement a solution: either via OPEX (cloud, MES, historians, APC/DCS) or CAPEX (sensors/actuators)

Long term sustainability of complex and rigid solutions. Lack of data ownership at the plant level. No IT/OT modernization.

ChatGPT agrees

Jan. 2023

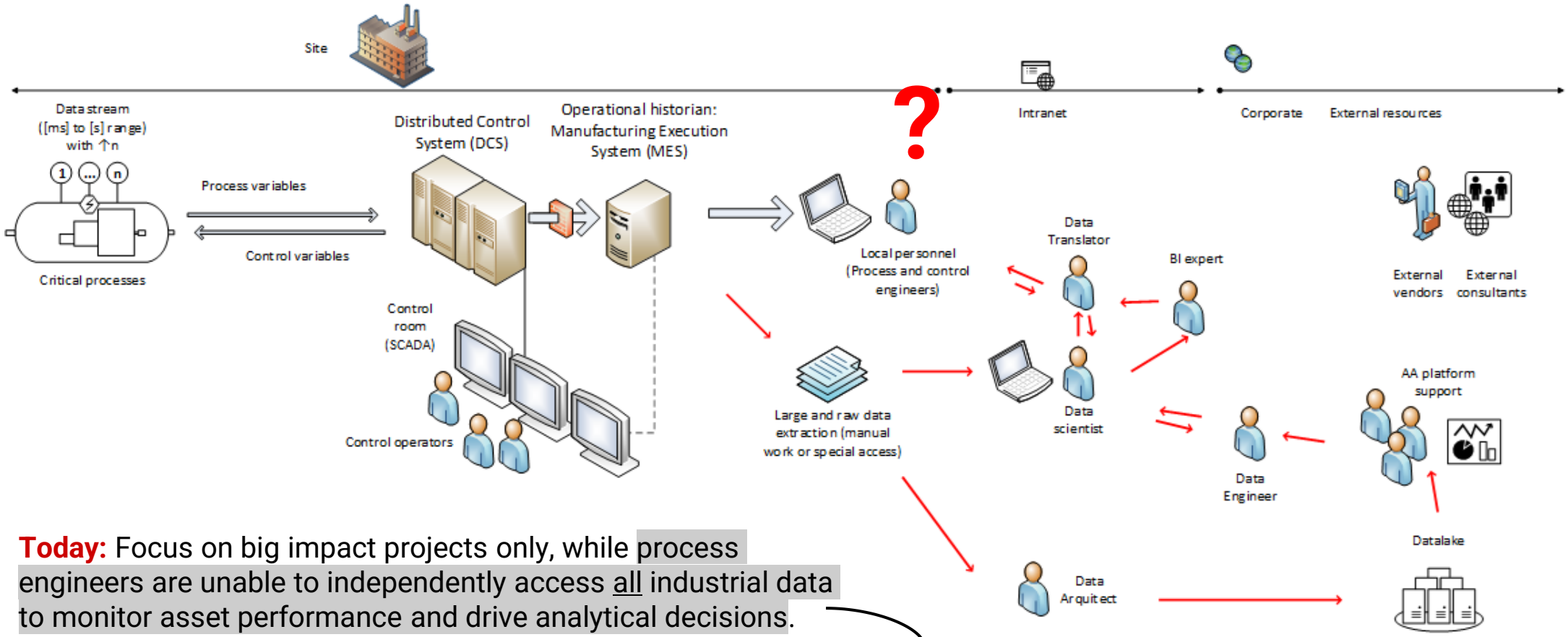
Could you describe, in one paragraph, how to use machine learning effectively in process and manufacturing industries so business impact is guaranteed?

To use machine learning effectively in process and manufacturing industries to guarantee business impact, it is important to start by clearly defining the problem or challenge that the business is facing, using relevant and high-quality data to train the model, selecting appropriate algorithms, monitoring performance, continuously improving, collaborating with experts, and communicating results to the relevant stakeholders. Additionally, process engineers should be involved in all stages of the process, from problem definition and data collection, to algorithm selection, model implementation, and performance monitoring, to ensure that the model is tailored to meet the specific needs of the manufacturing process and is delivering the desired results.

How not to scale a data-driven culture (number up approach)

Source:
Industrial data science – a review of machine learning applications for chemical and process industries

React. Chem. Eng., 2022, 7, 1471-1509



Today: Focus on big impact projects only, while process engineers are unable to independently access all industrial data to monitor asset performance and drive analytical decisions.

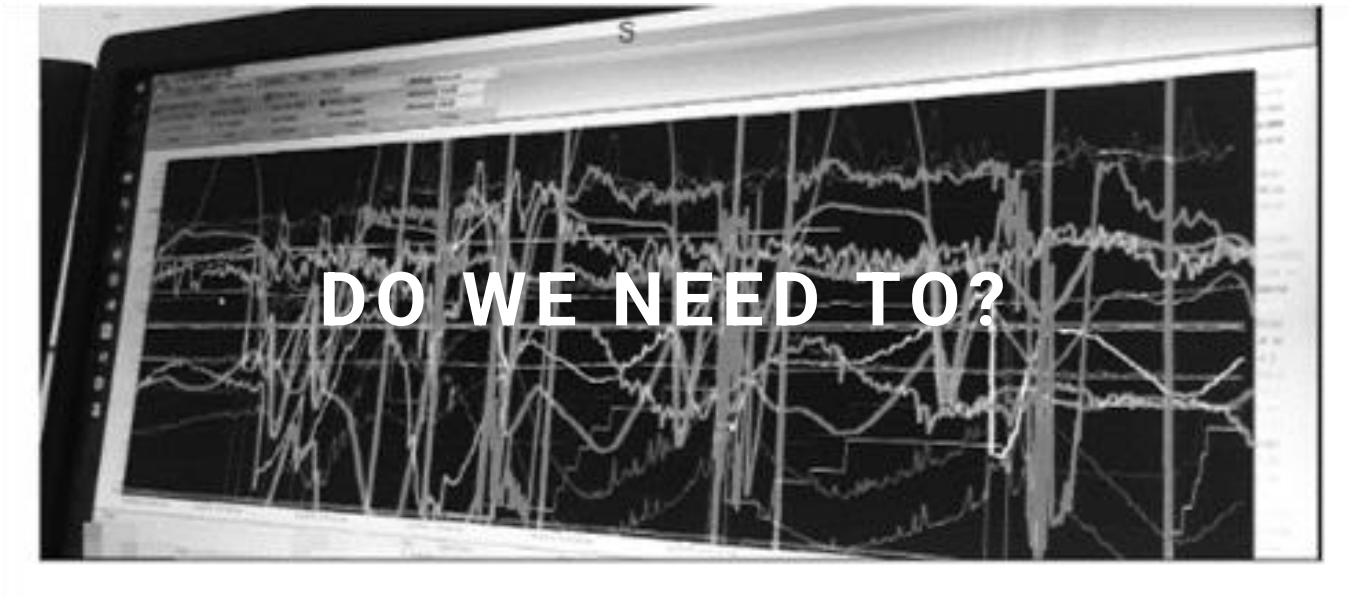
Digital and central teams delayed by:
1) domain expert availability
2) poor data quality
3) limited data access

Risk: Lack of data ownership at the plant level resulting in poor data quality: missing tags, unrecorded batch events, unstructered data (assets), siloed quality data…

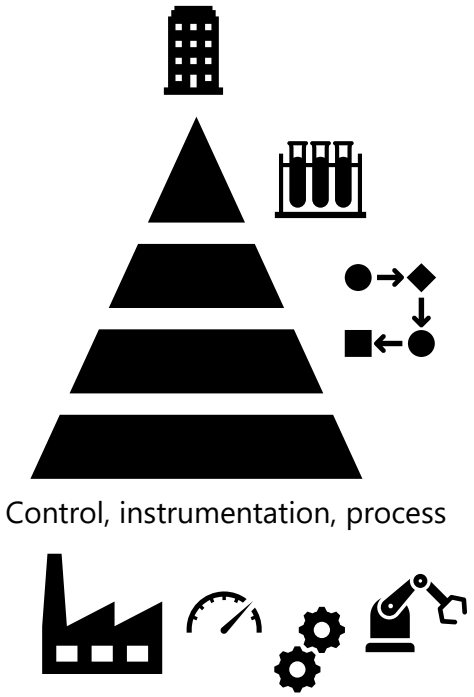
Small fraction of initatives are addressed (big impact projects only by specialized teams).

Complexity SELECTING DUPLICATING MAINTAINING **{industrial data}**

through multiple tools, systems and personas



There are two* types of process data



Tabular data

Rows are independent

Data integration (not data duplication)

Sensor data

Rows are time dependent

BatchID	LotID (PO number)	Grade (P1, P2, P3)	Quality (pass/no pass)	Yield [%]
...	...	...	...	...
...	...	...	...	...

(e.g., KPIs, LIMS and SAP data)

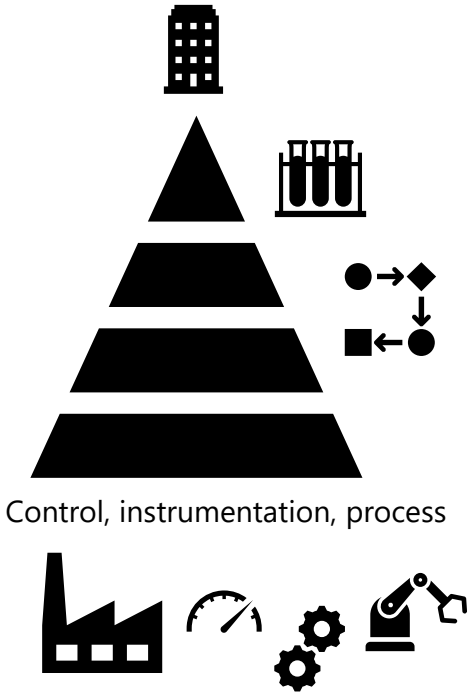
Datetime	PI_505.PV (Pressure) [bar]	TI_203.PV (Temperature) [degC]	FIC_101.MV (Valve position) [%]	FI_101.PV (Flowrate) [m3/min]
...	...	...	...	...
...	...	...	...	...

(e.g., tags in historians)

(*) Not *really* true, of course

There are many, many types of industrial analytics software

+info: g2.com, trustradius.com, capterra.com, gartner.com, **Insresearch.com**



Control, instrumentation, process

Tabular data

Rows are independent

Data integration (not data duplication)

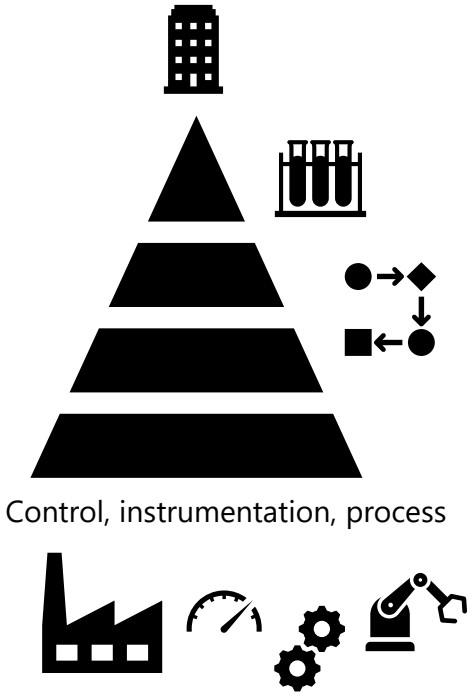
Sensor data

Rows are time dependent

Software integration (not duplication)

And many, many types of industrial databases

+info: g2.com, trustradius.com, capterra.com, gartner.com, **Insresearch.com**



Control, instrumentation, process

Tabular data

Rows are independent

Data integration (not data duplication)

Sensor data

Rows are time dependent

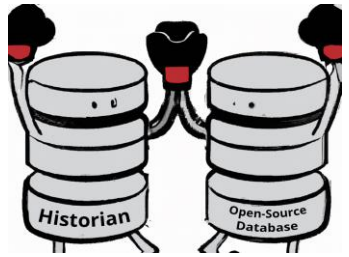
Azure aws Google BigQuery

ORACLE snowflake

MySQL Microsoft SQL Server

mongoDB. Timescale influxdb

Historians (OT) vs (IT) open-source databases



OSIsoft. aspentech Wonderware canary

SIEMENS GE Rockwell Automation

...

Data democratization tries to close the gap for industrial experts while covering areas with a majority of use cases

Six sigma analytics	Monitoring (anomaly detection)	Screening (descriptive)	Predictive analytics (modeling)	Optimization models (Energy, Process, Scheduling)

Excel

Daily data analysis and administrative tasks

(all process engineers, green belts)

Data democratization (no-code software)

Analysis done by operations experts allowing quick actionable insights and development of solutions

(digital process engineers, black belts, data translators, data scientists)

Python

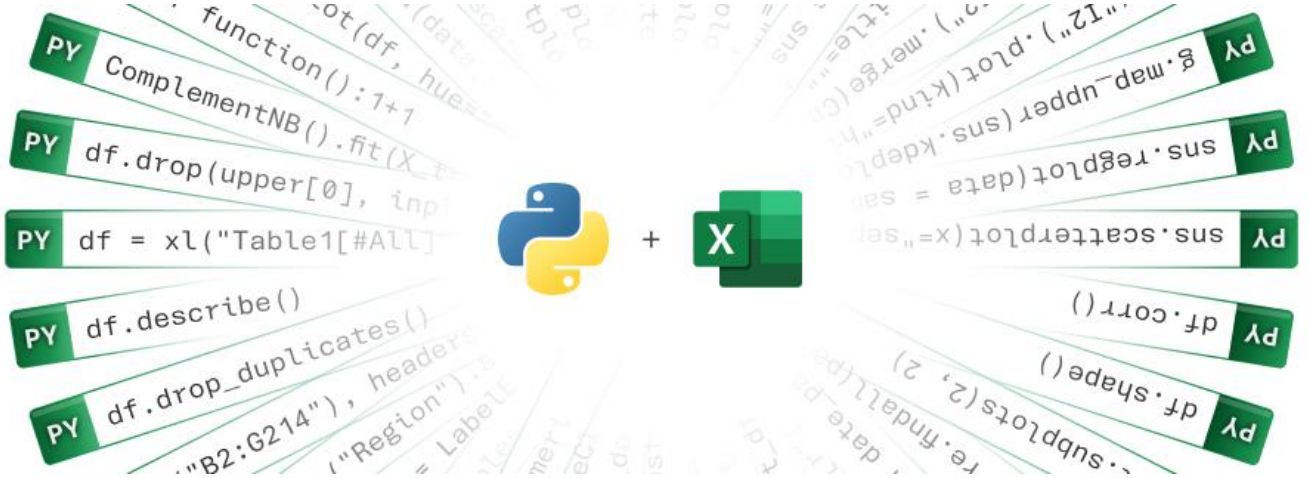
In cases with recurring and/or difficult modelling & optimization challenges, data science skills are crucial

(**data scientists, data engineers**)

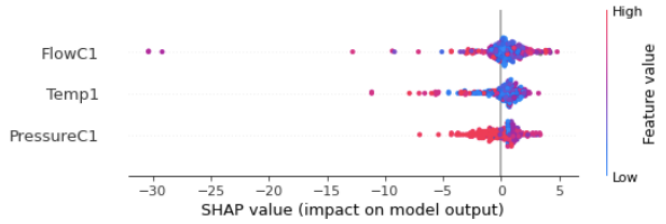
Number of use cases

Python in Excel

- The most popular data analysis tool integrated with the most popular programming language



https://techcommunity.microsoft.com/t5/excel-blog/announcing-python-in-excel-combining-the-power-of-python-and-the/ba-p/3893439



ML and Explainable AI integrated within Excel

Demo: screening processes using machine learning

Reaction Chemistry & Engineering

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal | View Issue

Check for updates

Cite this: *React. Chem. Eng.*, 2022, 7, 1471

Industrial data science – a review of machine learning applications for chemical and process industries†

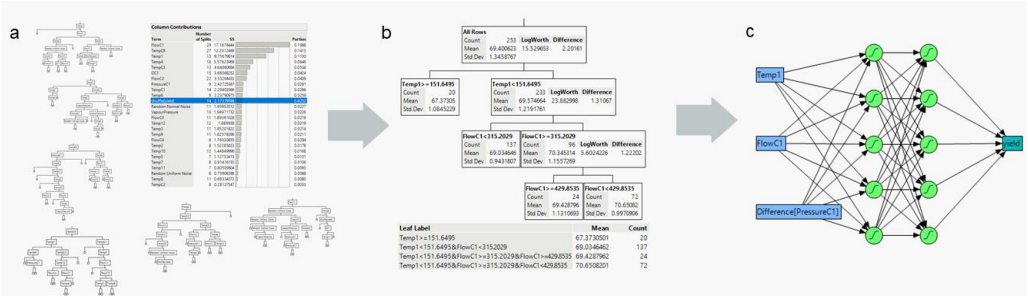
Max Mowbray, [a] Mattia Vallerio, [b] Carlos Perez-Galvan, [b] Dongda Zhang, [ac] Antonio Del Rio Chanona [c] and Francisco J. Navarro-Brull [*bc]

Received 1st December 2021,
Accepted 21st February 2022

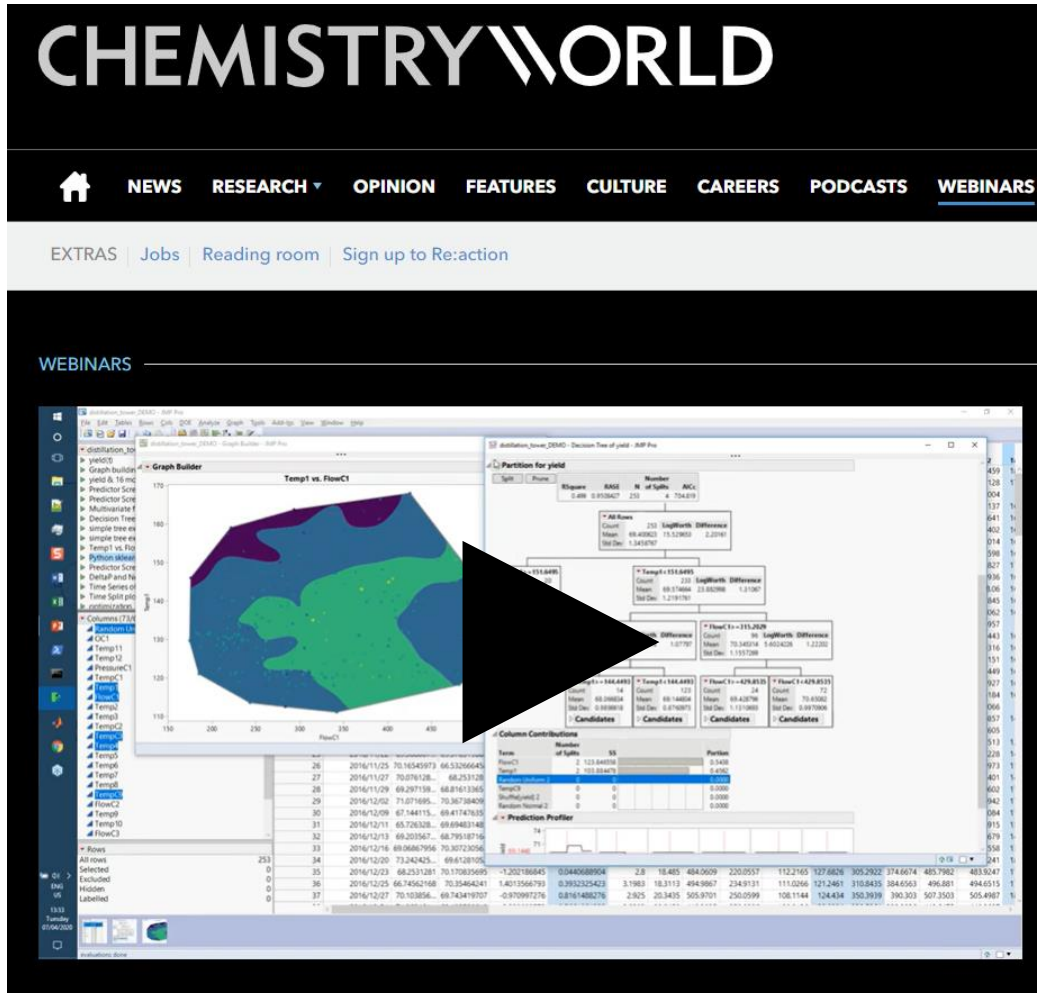
DOI: 10.1039/d1re00541c

rsc.li/reaction-engineering

In the literature, machine learning (ML) and artificial intelligence (AI) applications tend to start with examples that are irrelevant to process engineers (*e.g.* classification of images between cats and dogs, house pricing, types of flowers, *etc.*). However, process engineering principles are also based on pseudo-empirical correlations and heuristics, which are a form of ML. In this work, industrial data science fundamentals will be explained and linked with commonly-known examples in process engineering, followed by a review of industrial applications using state-of-art ML techniques.

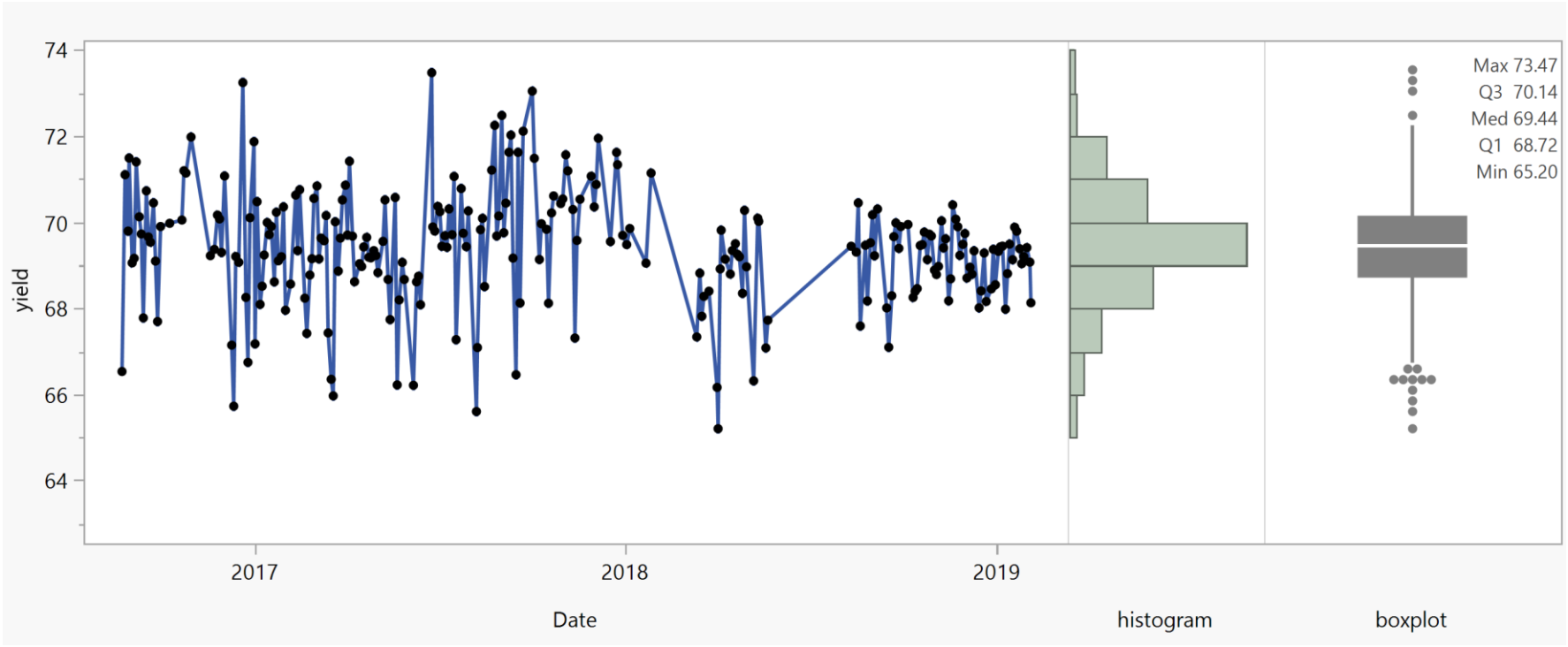


[Industrial data science – a review of machine learning applications for chemical and process industries]



[Industrial data science for chemical process perfection | Webinar | Chemistry World]

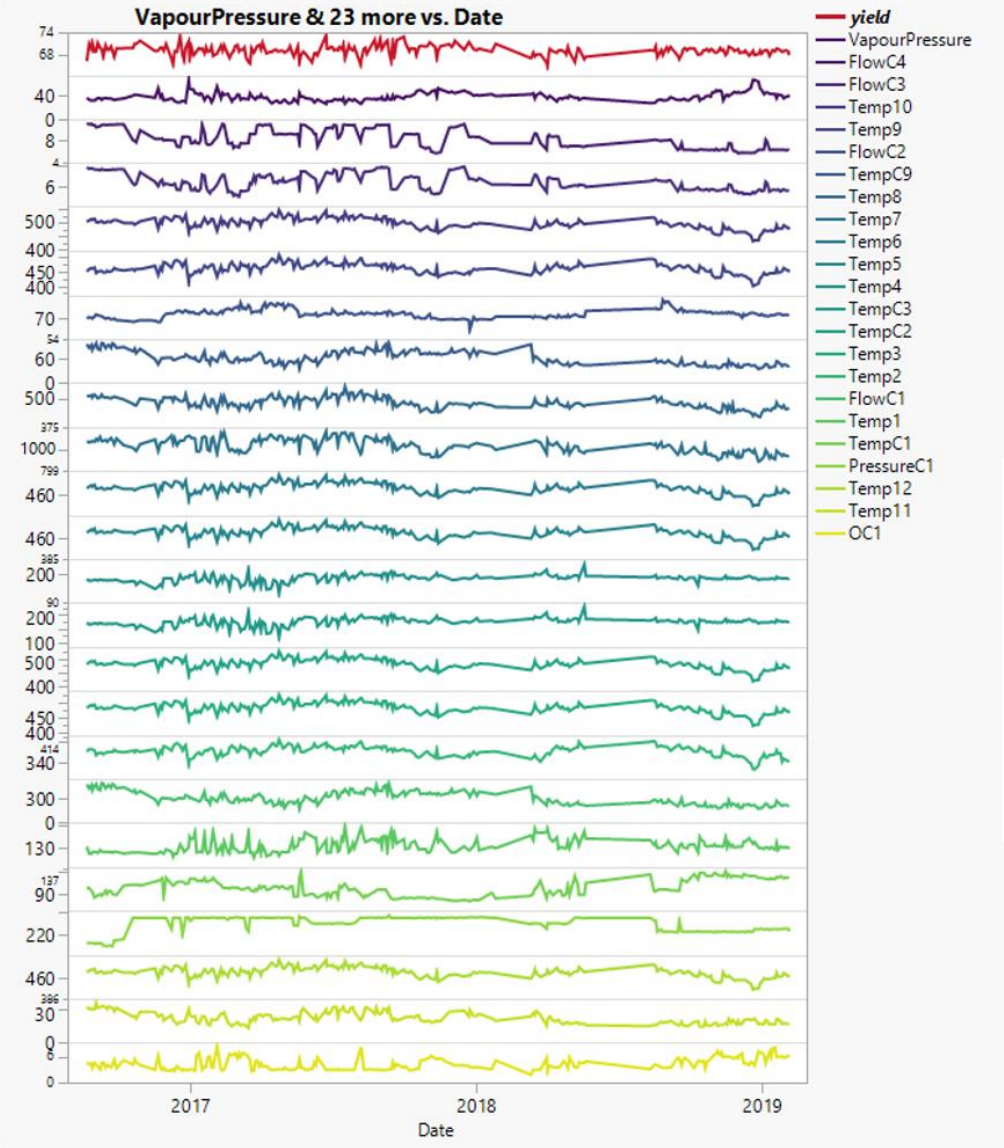
Distillation tower demo



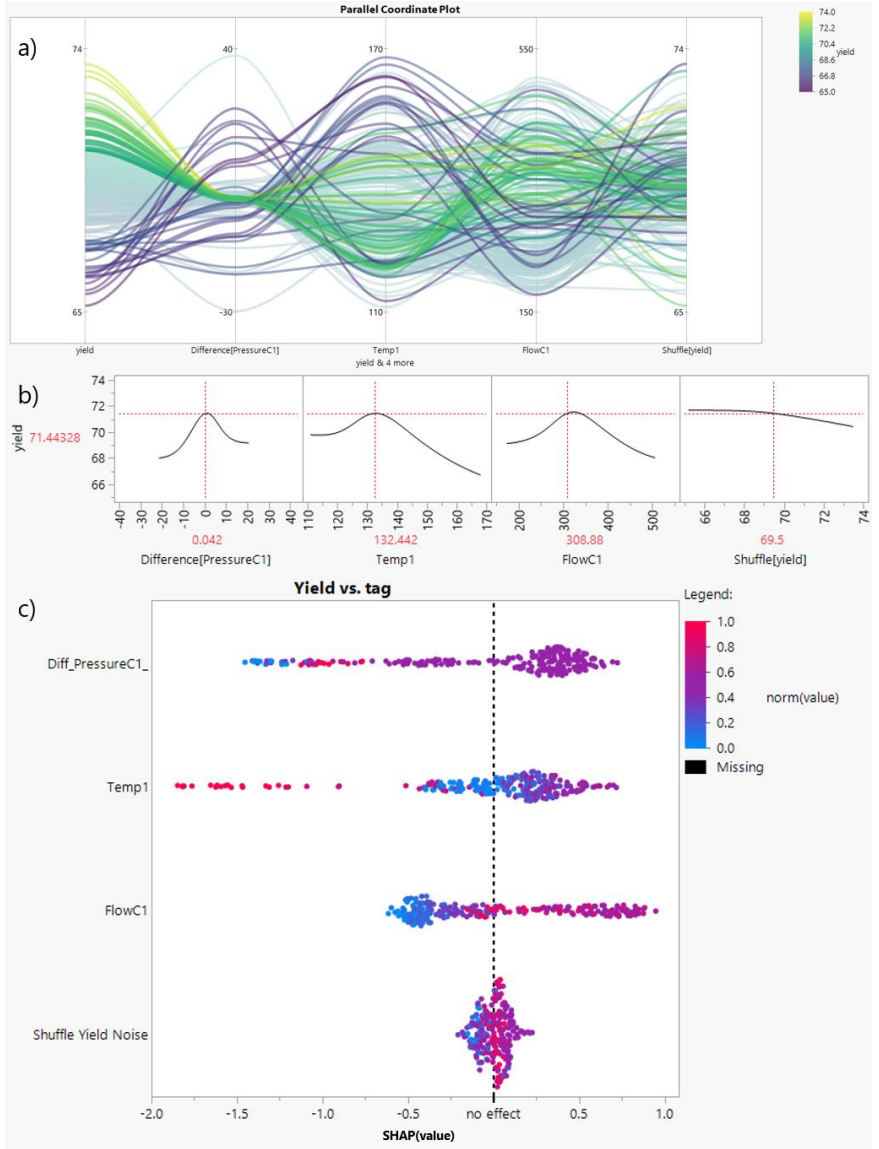
Process engineers from the site were able to stabilize the yield of a distillation tower but the baseline is now lower than before.

Distillation tower example

Live demo
https://www.chemistryworld.com/webinars/industrial-data-science-for-chemical-process-perfection/4011242.article



Automated screening of tags via correlation analysis (supervised learning)



Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Sargent Centre for Process Systems Engineering

Industrial Data Science

Industrial data and process control

What happens with all that data?

- Data is generated by sensors, automation systems, events, incidents, people, transactions...
- Several dedicated systems exist to handle this data
- A plant engineer needs to understand which data and systems are relevant for process performance improvements
- Data is not just numbers (P&ID, DCS, PFD, SOP, OEE, instruction manuals…)

10-100 staff

1k-10k sensors and actuators

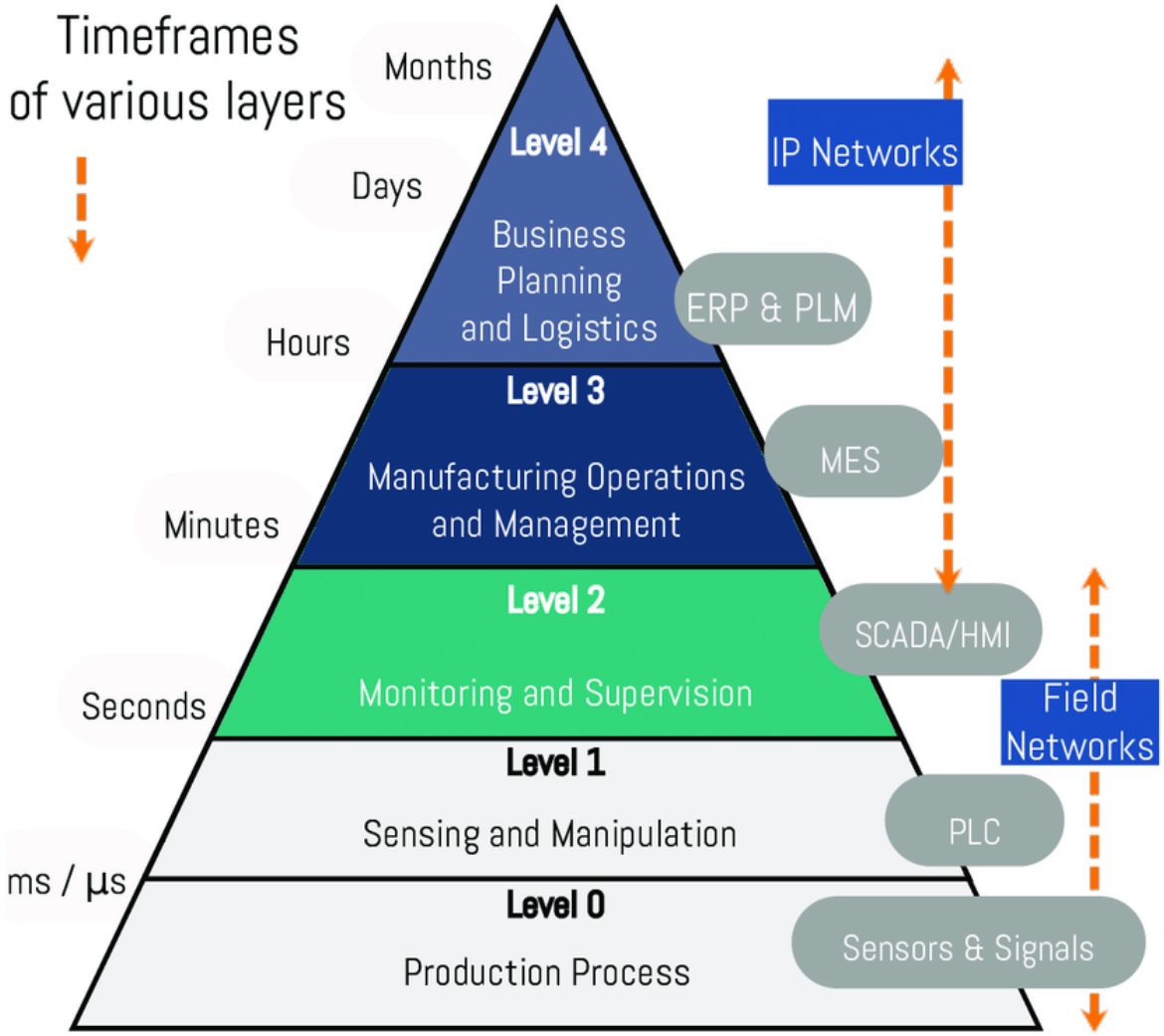
process control and automation

The Purdue model (Automation pyramid)

Source: Katti, Badarinath. (2020). Ontology-Based Approach to Decentralized Production Control in the Context of Cloud Manufacturing Execution Systems. 10.13140/RG.2.2.11486.46402.

IT – Informational Technology

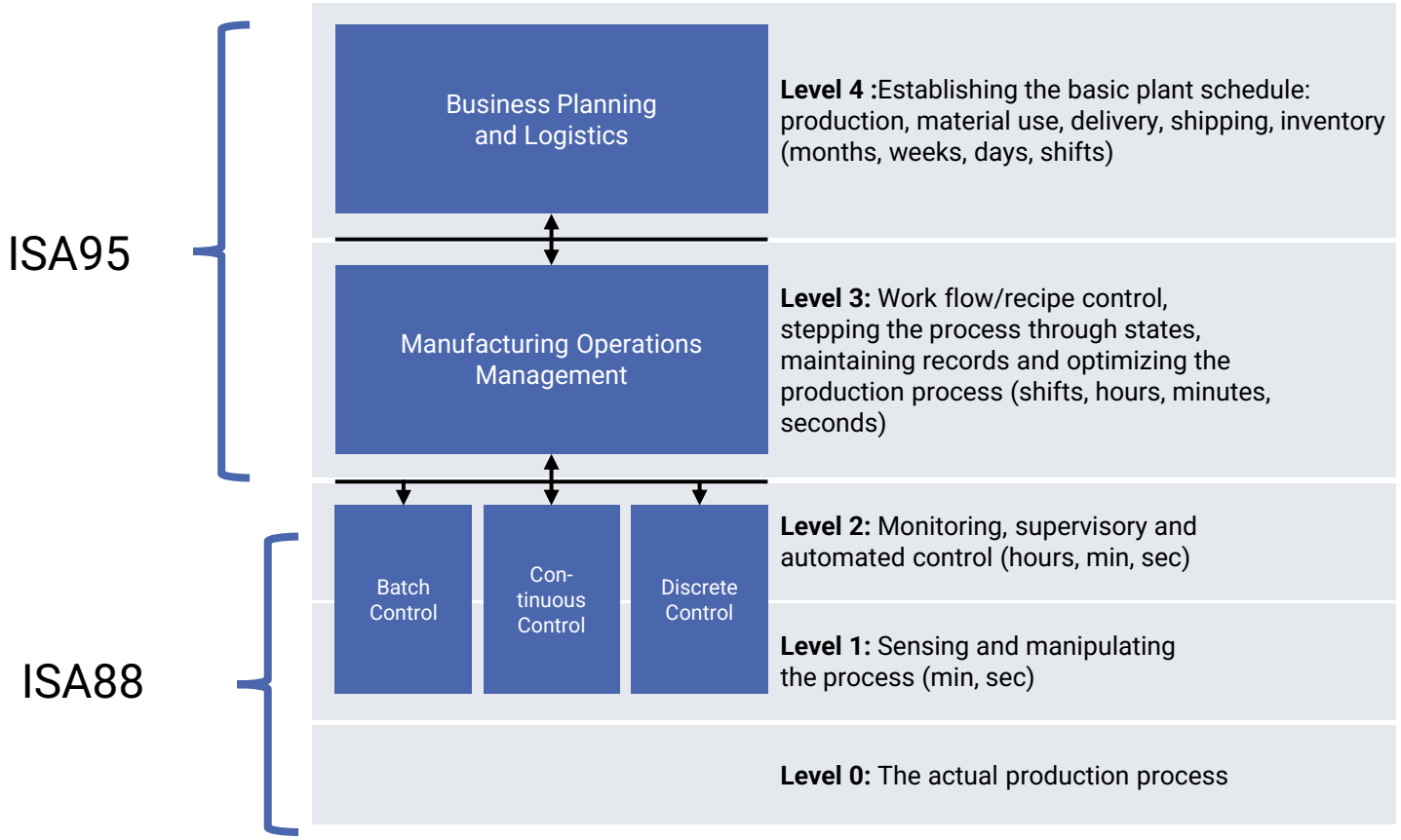
OT – Operational Technology



+info: https://learn.umh.app/lesson/introduction-into-it-ot-automation-pyramid/

Generic structures of process data

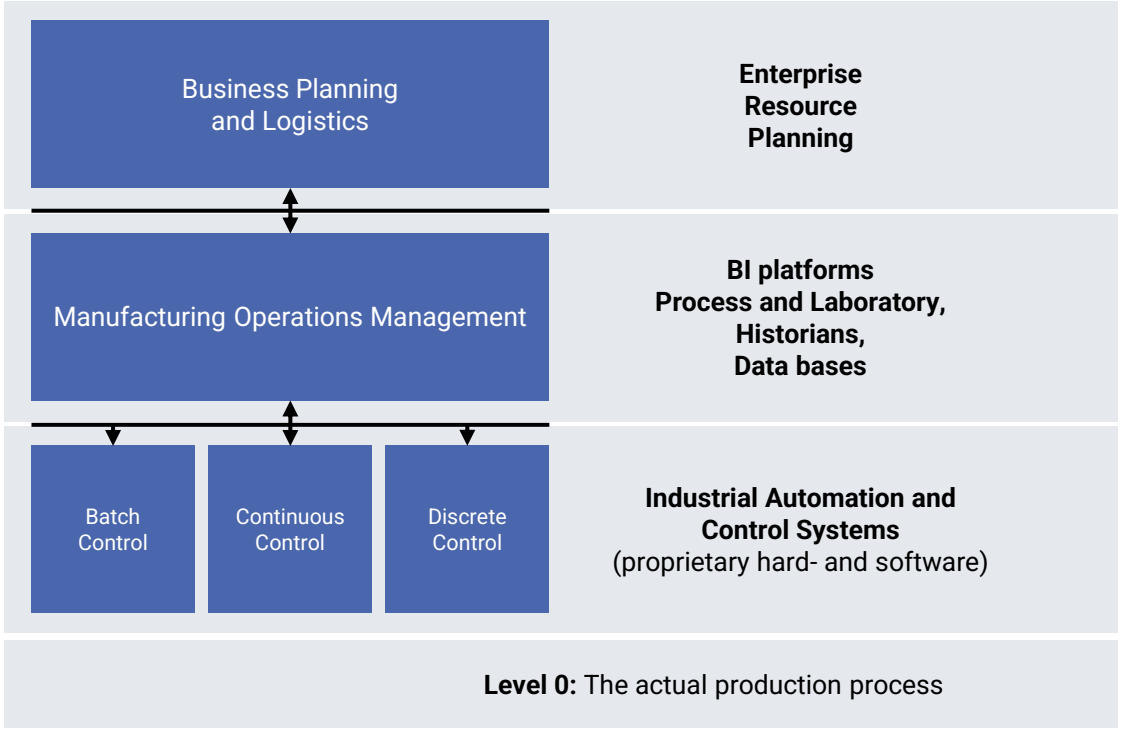
- Data for process improvement often involves sensor and actuator data, their processed or aggregated forms, and information from transactions based on this data
- Industry standards describe the relationships and uses (ISA95/ ISA88/IEC62264)
- The higher the level, the longer the time horizon and the coarser the data



IEC 62264

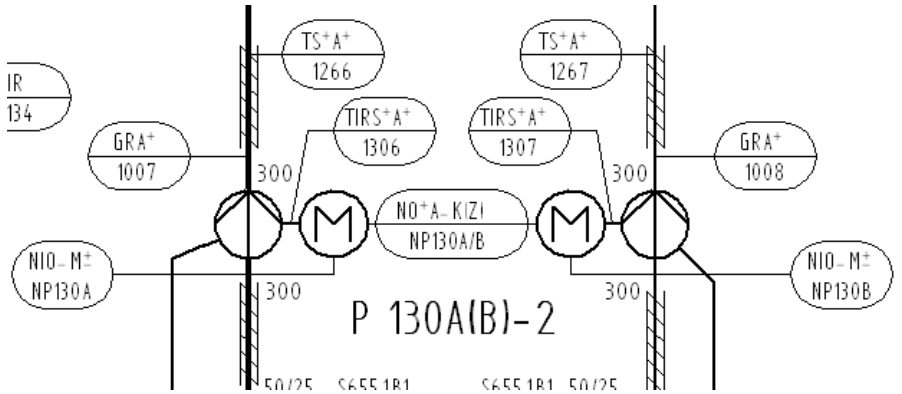
Data Systems Overview

- Depending on the function, proprietary or common IT solutions exist
- Particularly data in level 3 is relevant for Data Analytics, but data from other layers is used as well
- Data in level 3 is commonly stored in the form of times series data (trend data) with a sample frequency of ~minutes



Data identification level 1,2,3 – tag names

- Data in level 1,2 and 3 is often identified by a tag name: a combination of letters (that denote the function) number (that identify the particular sensor, actuator or control loop)

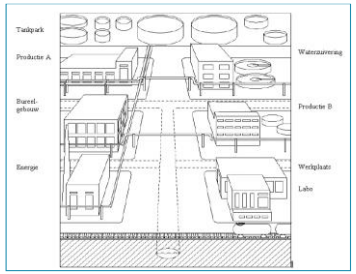


Letter	Measurement		Action
	1e letter	2e letter	
A			Alarm
C			Control
D	Density	Difference	
E	Electrical		
F	Flow	Ratio	
G	Position		
H	Manual action		
I			Indication
K	Time		
L	Level		
M			
N	Electrical consumer		

Letter	Measurement		Action
	1e letter	2e letter	
0			
P	Pressure		
Q	Quality	Integral	
R			Registration
S	Speed, frequency		Switch
T	Temperature		
U	Calculated property		
V	Viscosity		
W	Weight		
+			High level
-			Low level

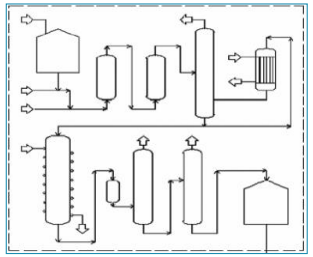
Data identification level 3,4 – functional location

- Data in level 4 (and to some extend in level 3) is identified by data base field names and functional location: a code that uniquely identifies all parts of a chemical plant



Production plant
E.g. steam cracker

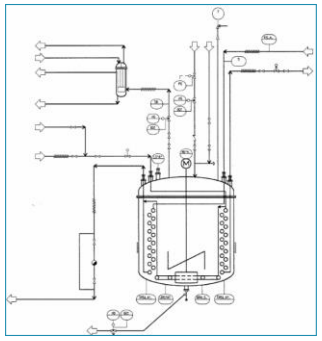
Specific part of the the production plant
E.g. synthesis and separation

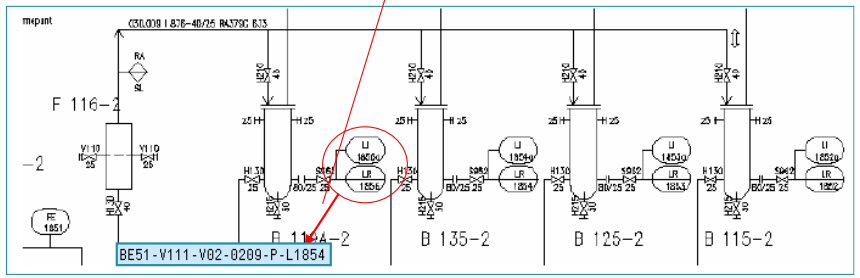


Site location

BE51	-	ANNN	-	ANN	-	XXXX	-	A	-	XXXXXXXXXXX

Specific unit operation
E.g. reaction

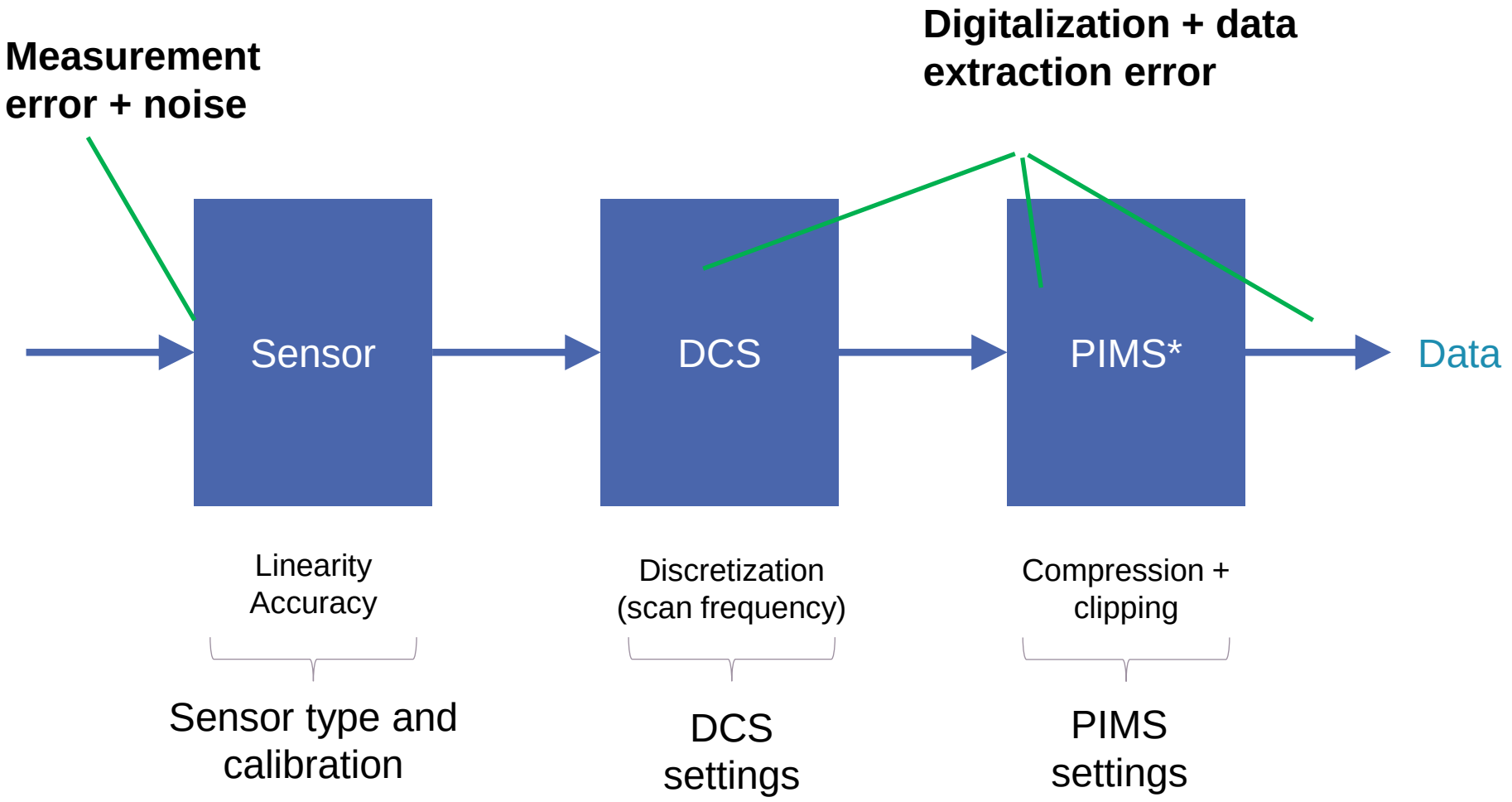




All the instrumentation and smaller equipments in the specific unit operation

Data quality assessment for chemical production data

Where are the errors coming from?

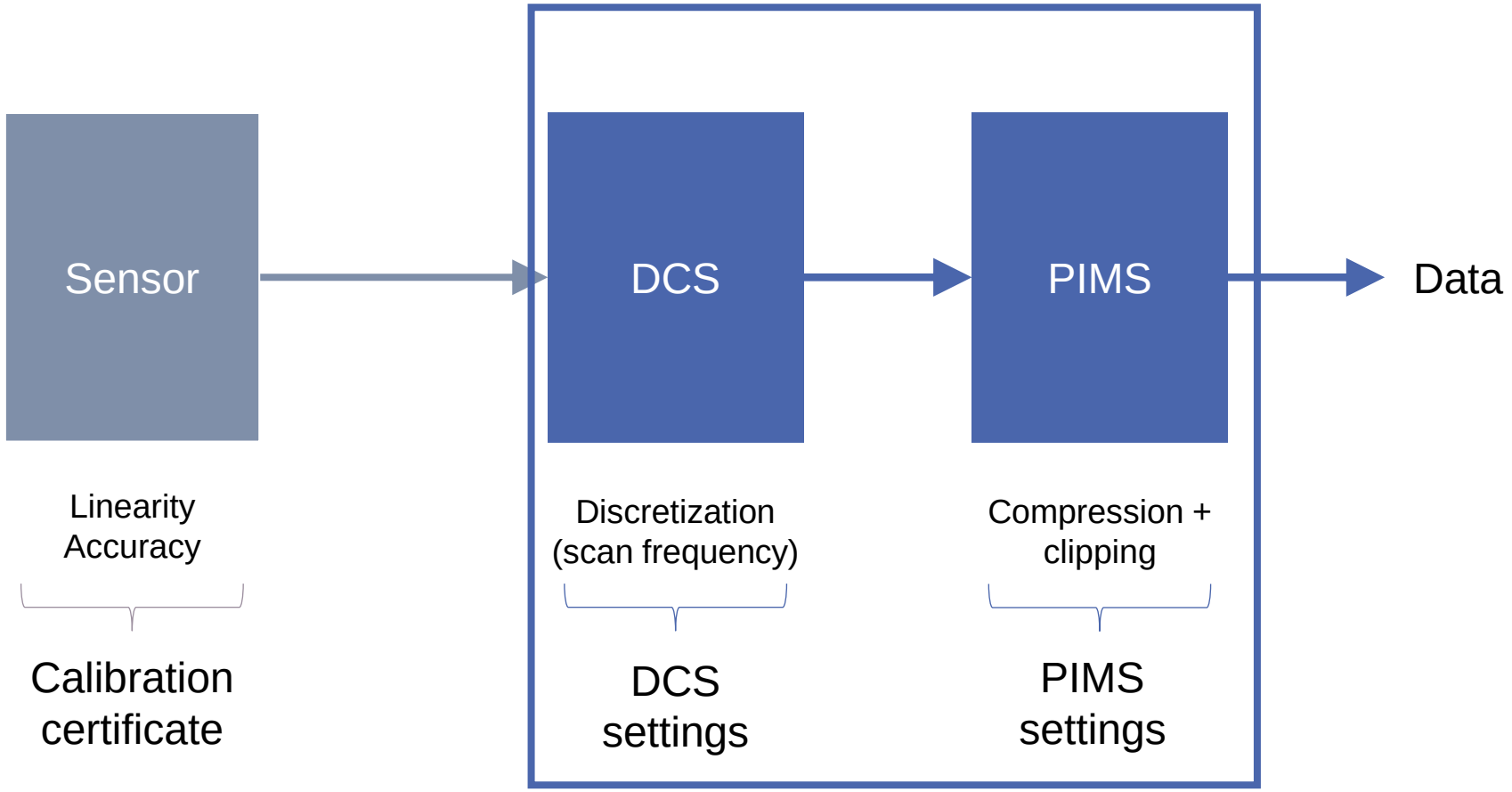


L1,L2: DCS – Distributed control system: all sensors are connected to this system which converts the instrument analog signal into a digital one

L2,L2: PIMS – Plant Information Management System: where data from the DCS system is stored, visualized and archived

Data quality assessement for PIMS data

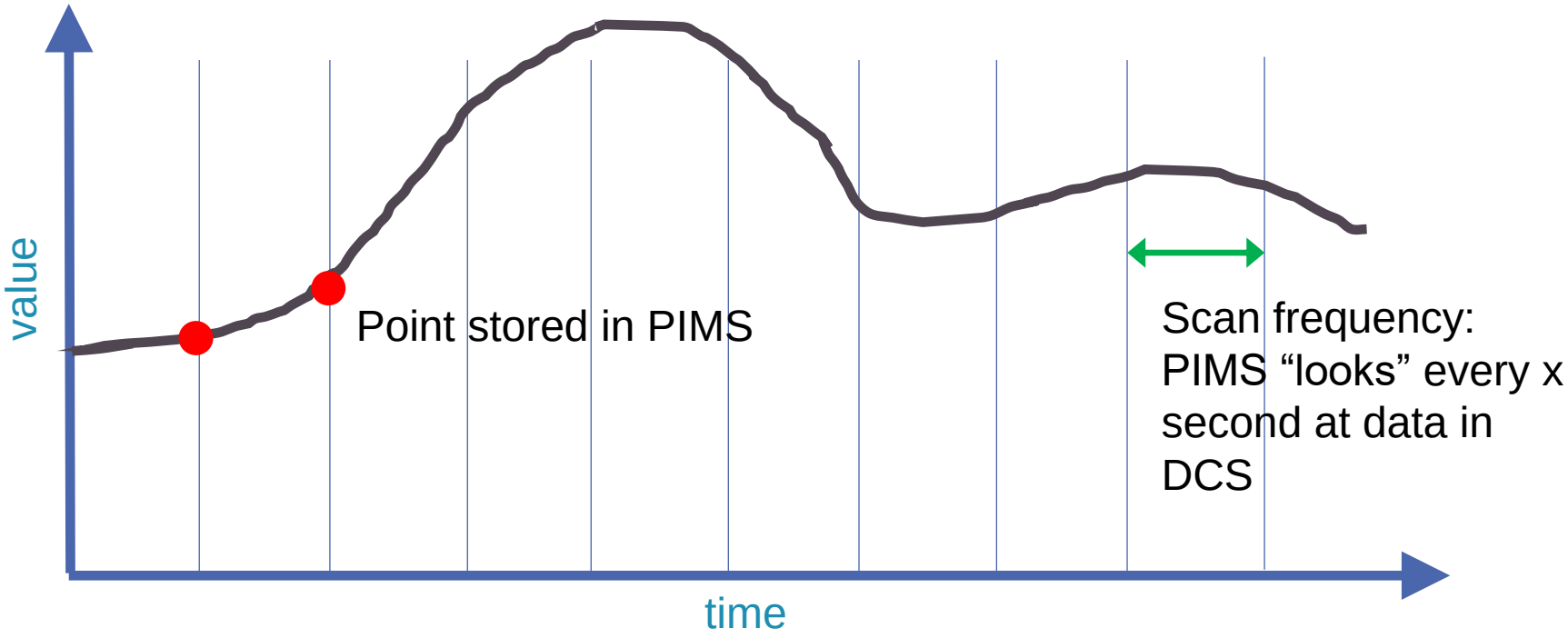
Determine the digitalization error



Data quality assessement for PIMS data
Digitalization error

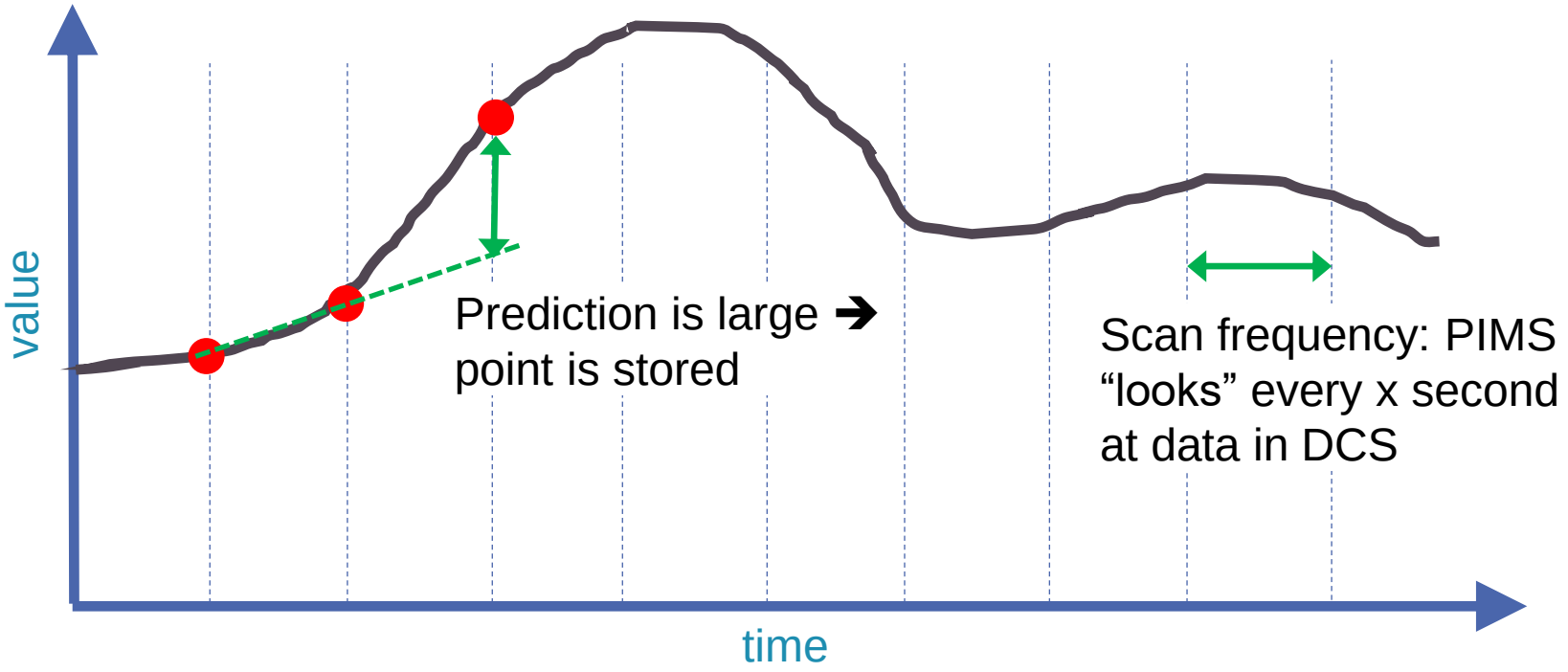
Measurement error
Sampling error

PIMS uses the scan frequency setting to determine when to look at the DCS data

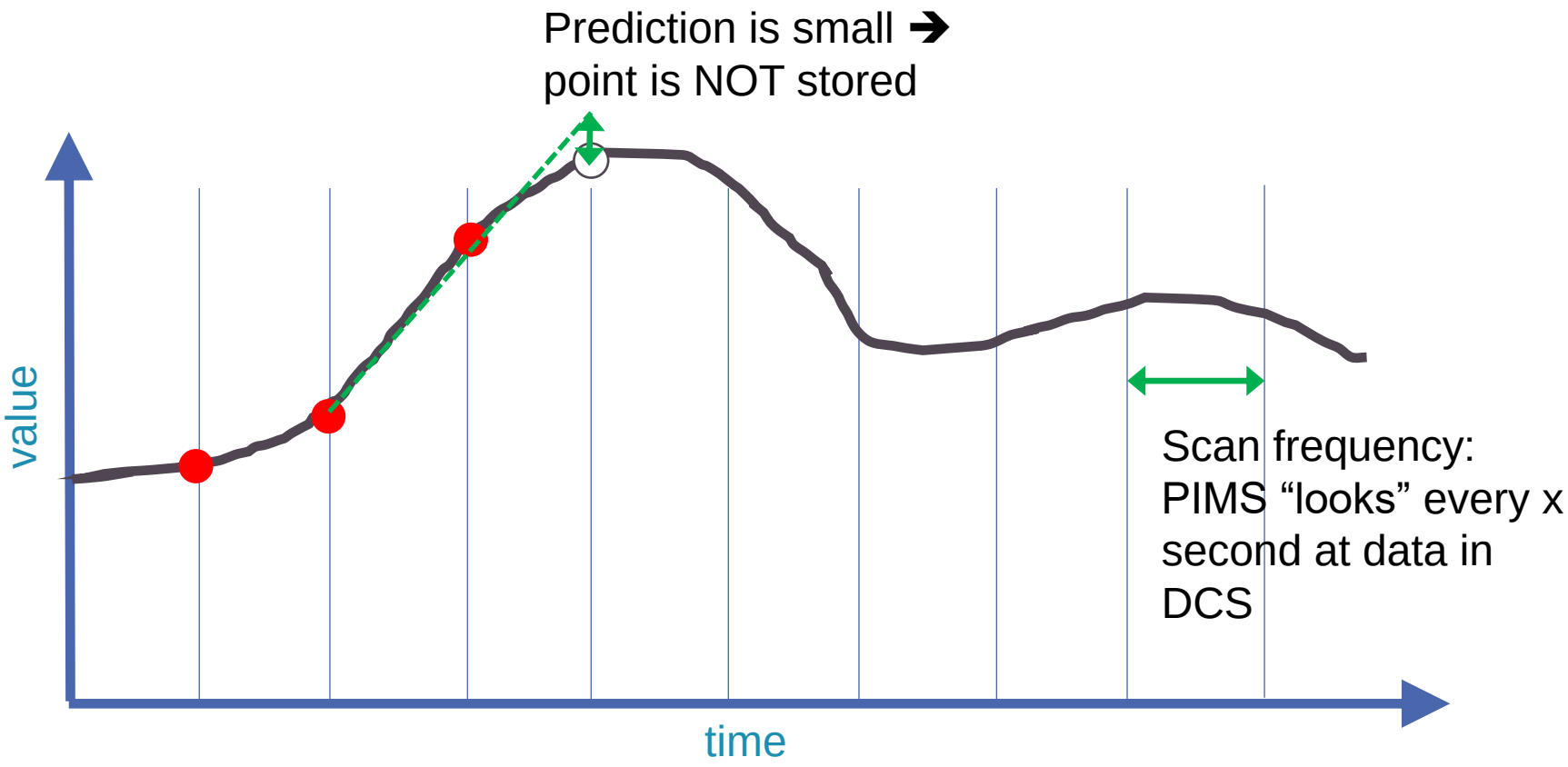


Data quality assessement for PIMS data
Digitalization error

For each new point, PIMS tries to predict the new point based on older points ➔ if the prediction error is too large (depending on compression settings) the new point is stored

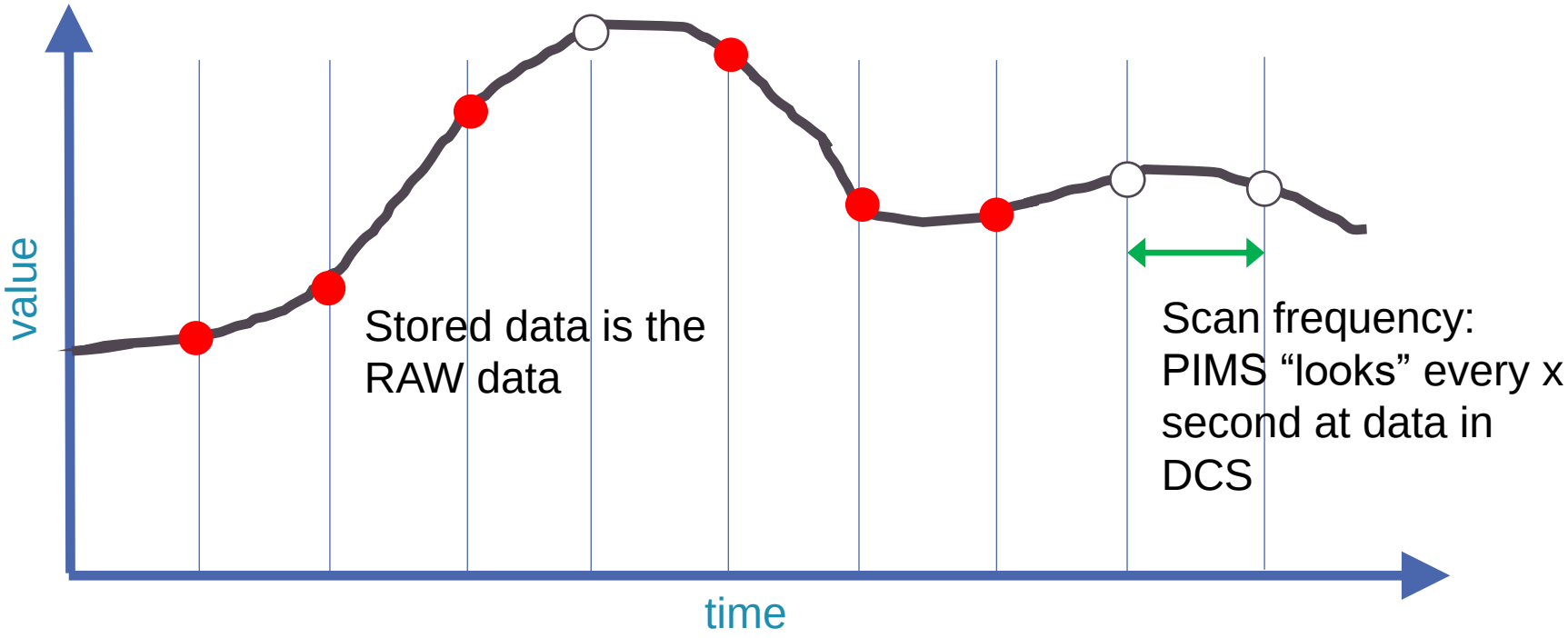


Data quality assessement for PIMS data
Digitalization error



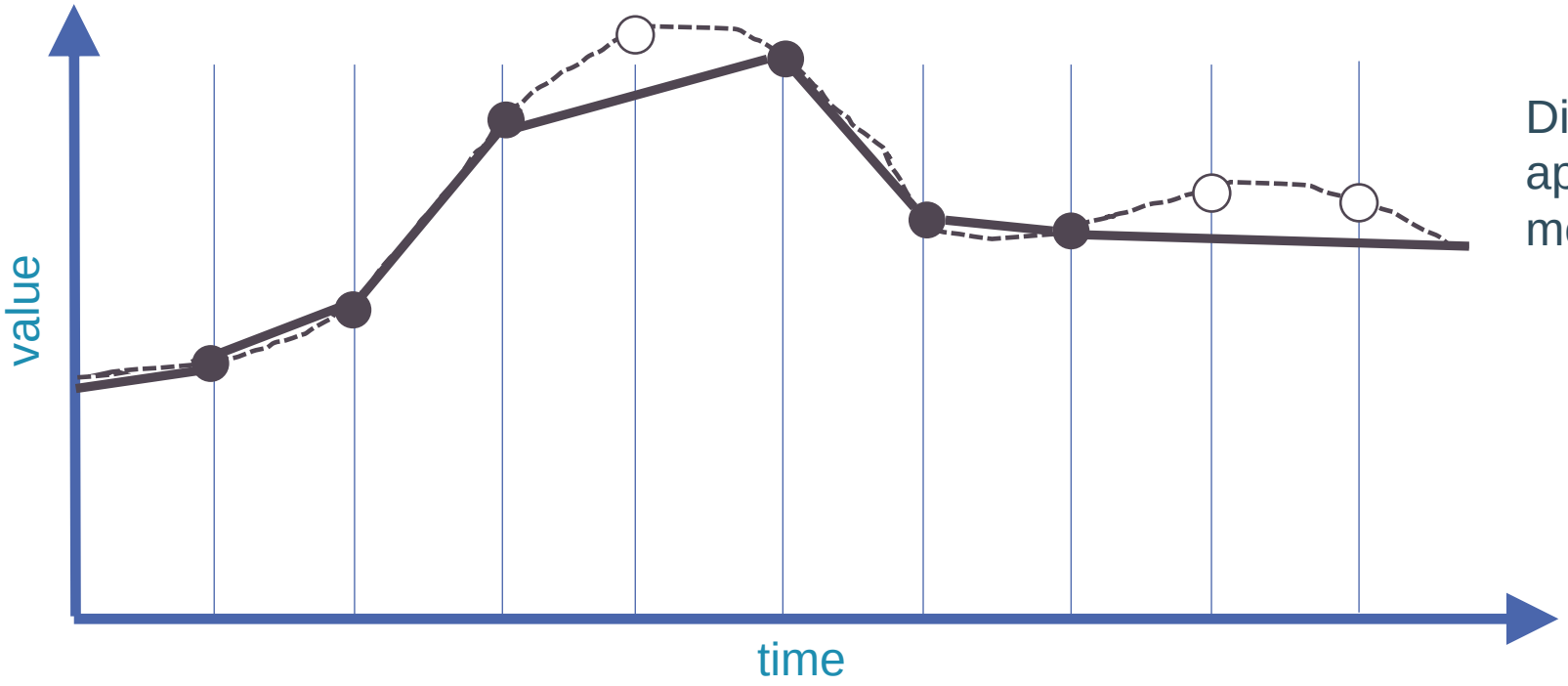
Data quality assessement for PIMS data
Digitalization error

Overall: not all scanned data points are stored (if compression is turned on)



Data quality assessement for PIMS data
Digitalization error

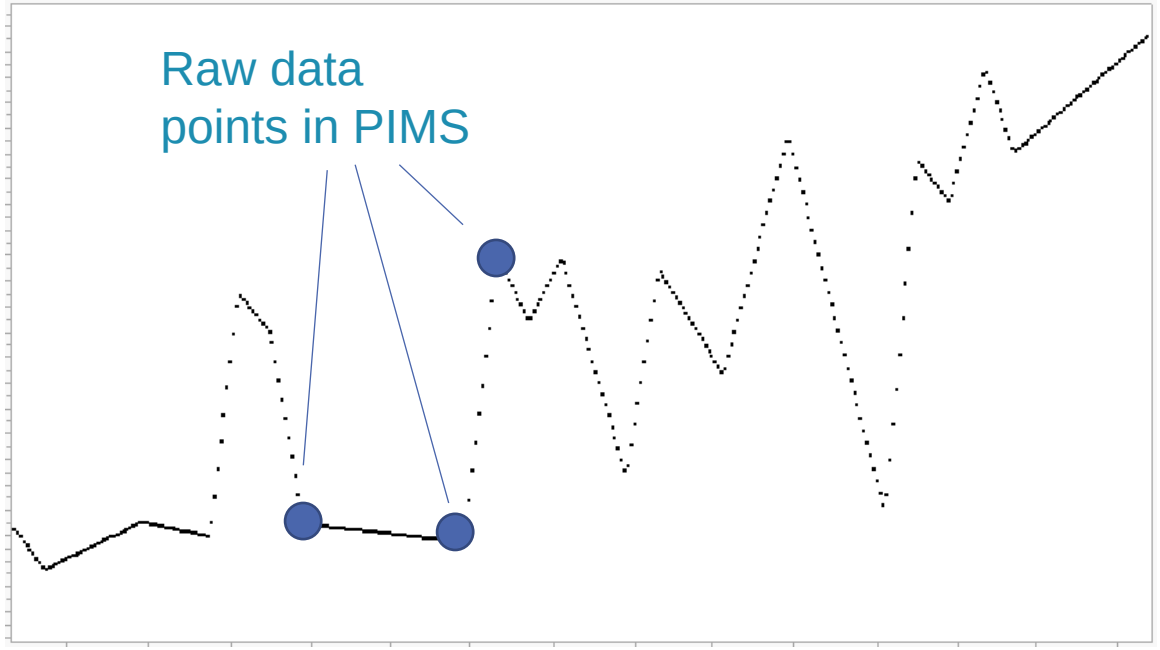
When retrieving the data (e.g. for data analysis) RAW data is interpolated to estimated missing data. Overall some detail is lost!



Different PIMS systems apply slightly different methods to store data.

Data quality assessement for PIMS data
Digitalization error – what if I ignore this?

- Risks
  - A lot of interpolated data that takes time to download from the system: data extraction takes much longer than needed.
  - Variations that happen faster than what is captured by PIMS are not recorded and cannot be analyzed ➔ the information is not in the data, but that does not mean it does not exist.



Example of data extracted with a timestamp lower than the actual time between raw values

Sargent Centre for Process Systems Engineering

Industrial Data Science

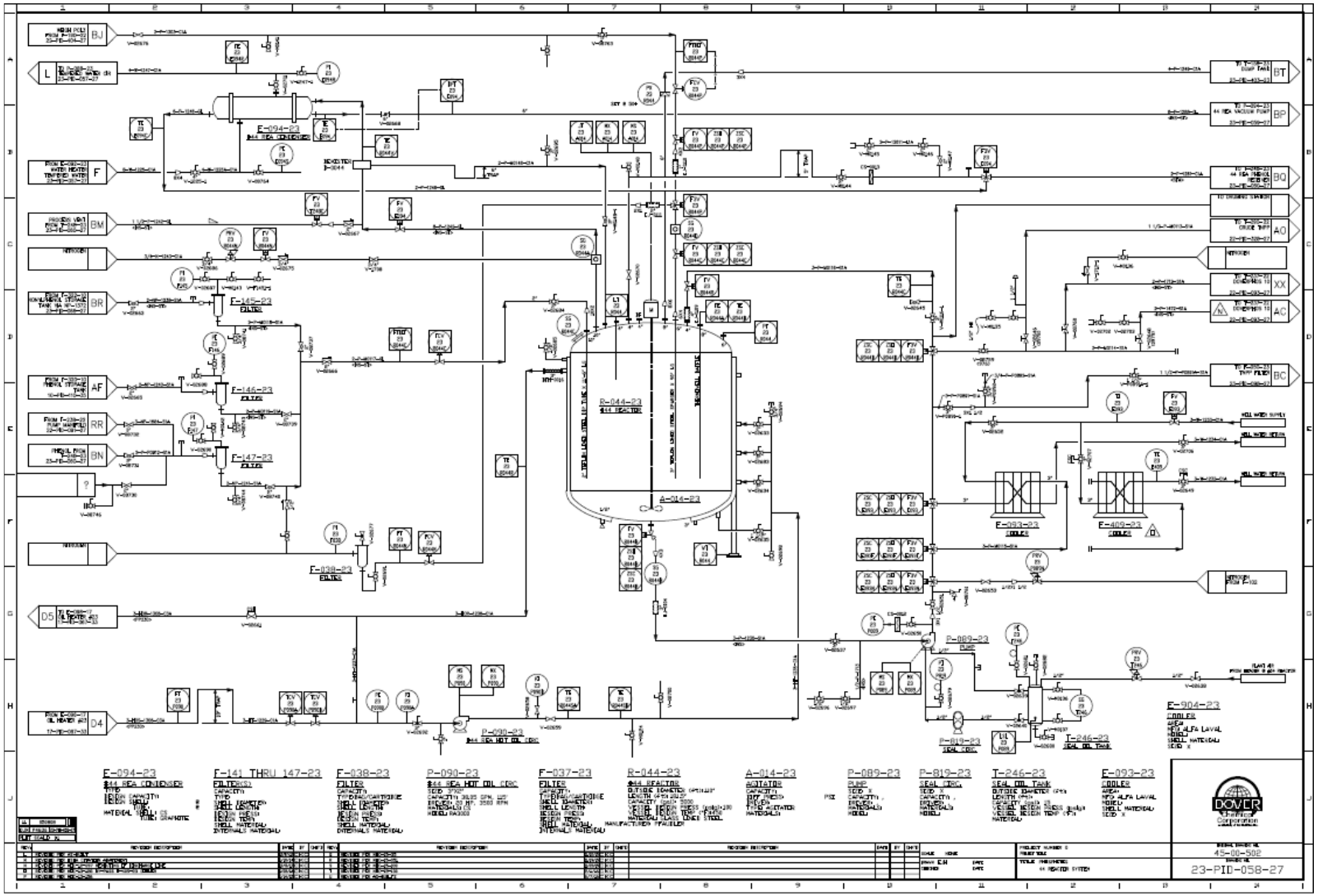
Industrial metadata: asset hierarchy, automation/batch sequences

Diagrams

- Block Flow Diagram (BFD)
  - Unit operations (blocks) connected by lines (streams)
- Process Flow Diagram (PFD)
  - Unit operations (icons) connected by numbered lines (streams)
  - Temps, Pressures, compositions, flows specified (e.g. Aspen)
- Piping and Instrumentation Diagram (P&ID)
  - Units, pipes, valves, pumps, exchangers, instruments, pipe specs
  - Control strategies (loops, interlocks, alarms, etc.)
- Isometric Diagram – scale drawing of process in 3D space
- Equipment Drawing – scale drawing of equipment + details

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

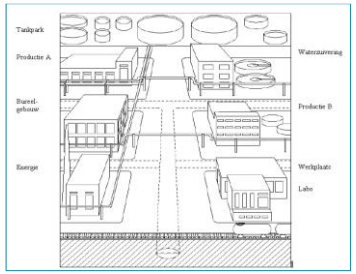
Actual Batch Reactor P&ID



Source: https://www.chemicalengineeringpractice.org/

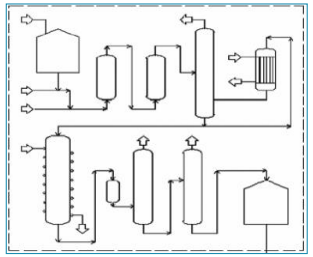
Data identification level 3,4 – functional location

- Data in level 4 (and to some extend in level 3) is identified by data base field names and functional location: a code that uniquely identifies all parts of a chemical plant



Production plant
E.g. steam cracker

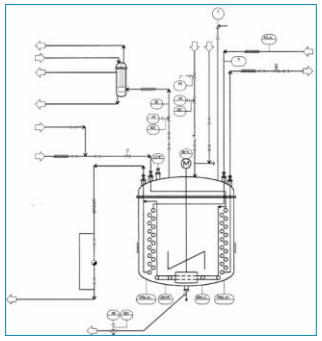
Specific part of the the production plant
E.g. synthesis and separation

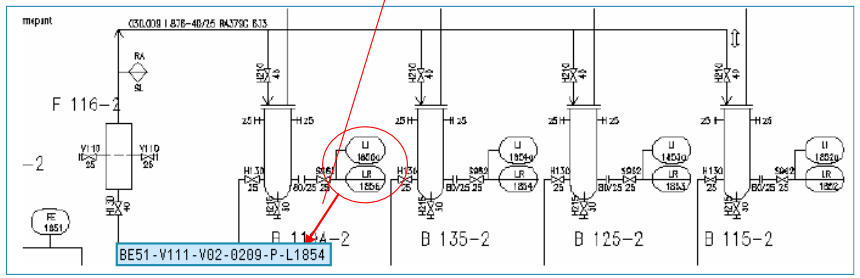


Site location

BE51	-	ANNN	-	ANN	-	XXXX	-	A	-	XXXXXXXXXXX

Specific unit operation
E.g. reaction





All the instrumentation and smaller equipments in the specific unit operation

Asset hierarchy

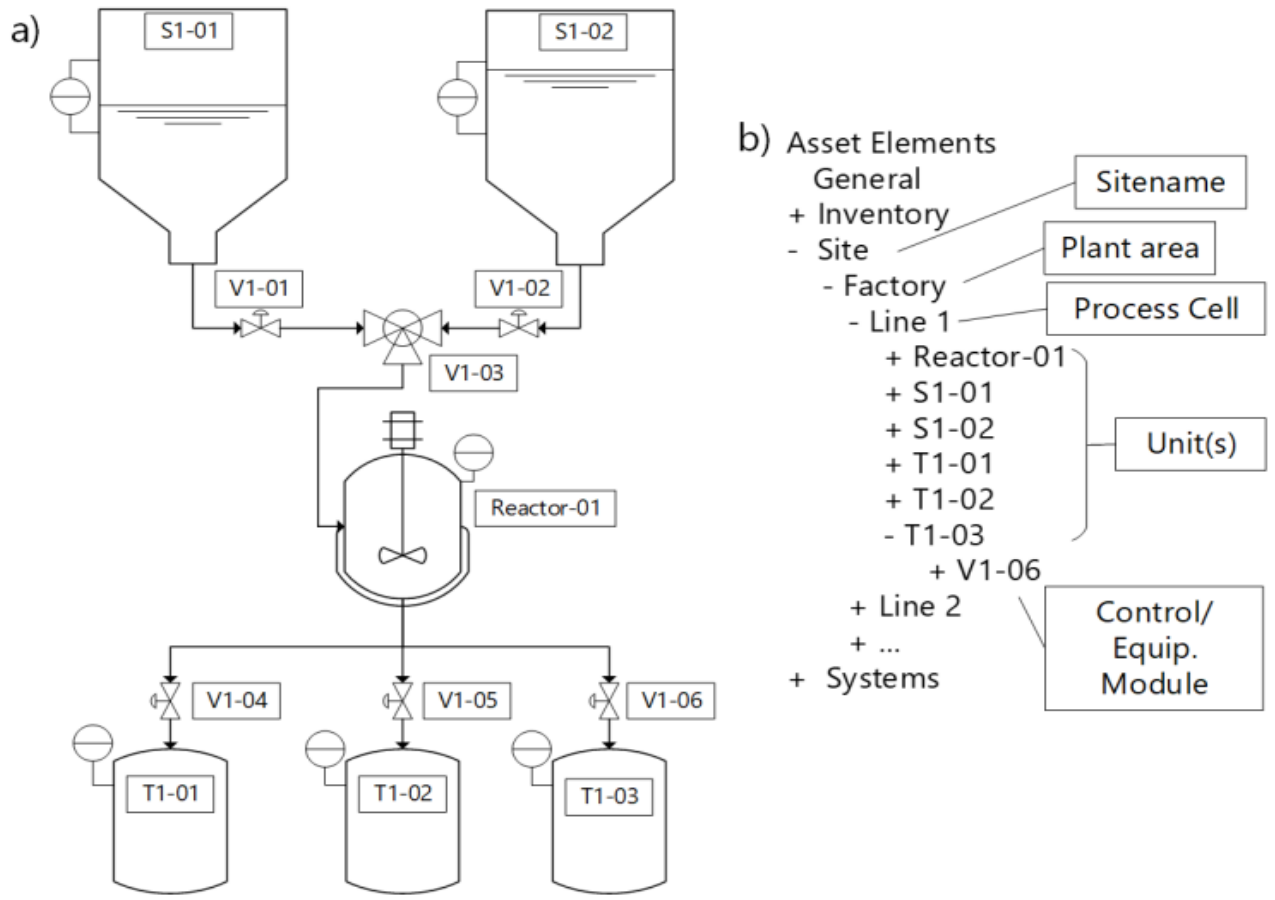


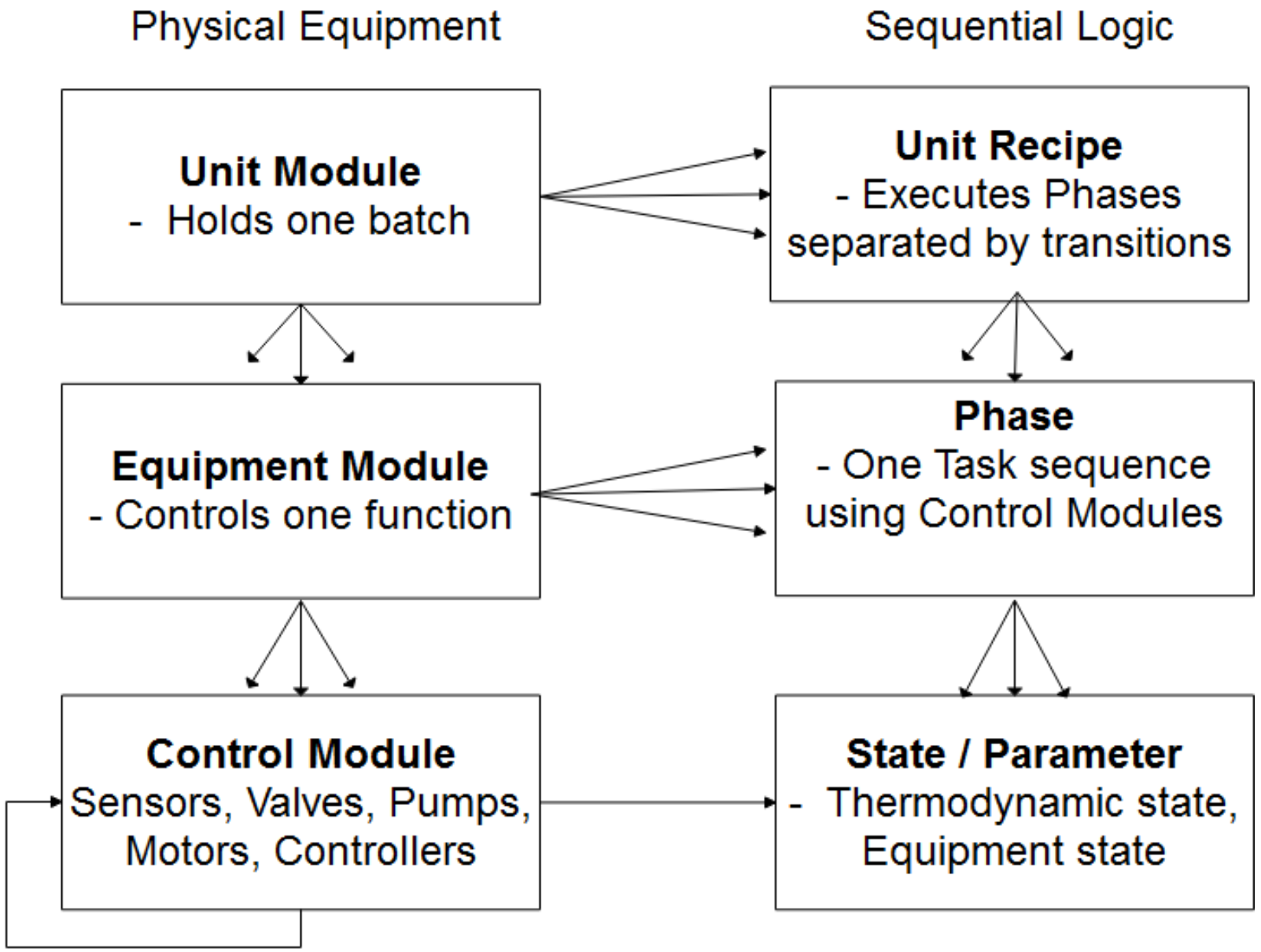
Figure 13: Simplified process (a) following ISA-88 to standardized asset hierarchy (b) in the plant.

https://try.seeq.dev/workbooks?t=TVlfRk9MREVS

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

S88 Model

- We will focus on bottom 3 layers
- Model also includes Process, Area, Plant, and Business layers
- Unit Recipes and Phases are Sequences: SFCs



Source: https://www.chemicalengineeringpractice.org/

S88 Layers – Control Modules / States

- Control Modules are the lowest level of the physical model
  - Most detailed – As shown on a P&ID, e.g.:

Control Module	Command	Feedback
Solenoid Valve	Open/Close	Open/Closed/In Transition
RTD	(alarm levels)	Process Temperature, alarm status
Variable Speed Pump	On/Off, Target Speed	Running/Stopped, Actual Speed
Level Control Loop	Auto/Manual, Target Level (setpoint)	Measured level, control valve position, target deviation alarm

- Control Modules measure the **Thermodynamic State**
- Control Modules manipulate the **Equipment State**
  - We also measure the Equipment State (e.g. motor/valve status)

Source: https://www.chemicalengineeringpractice.org/

S88 Layers – Equipment Modules / Phases

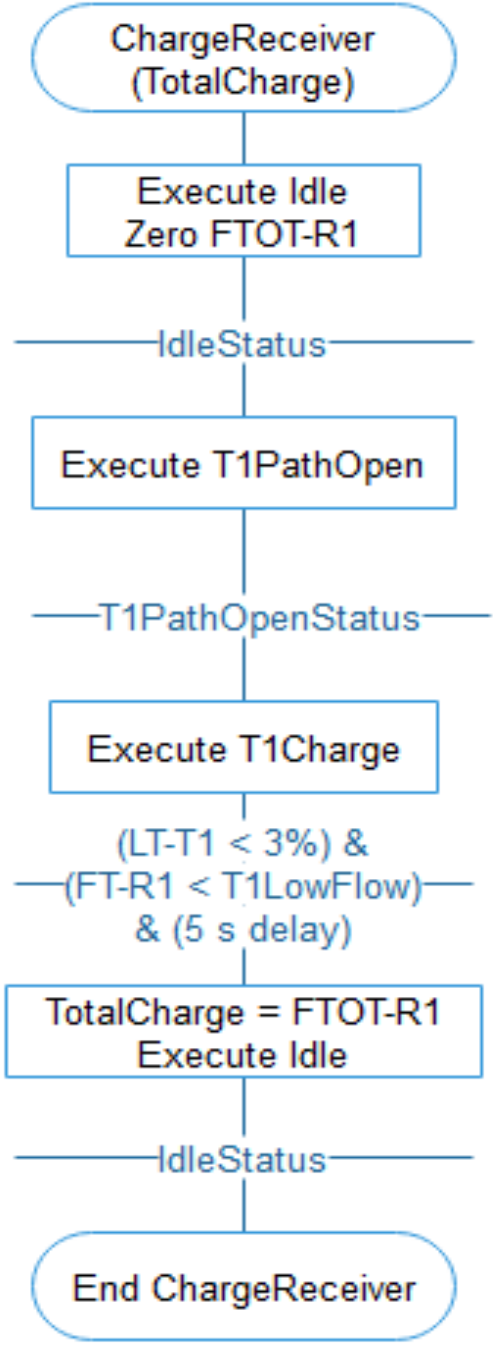
- Equipment Modules are a collection of Control Modules associated with one Function, e.g.
  - **Temperature Control**: Temperature sensors, control valves and block valves for steam/hot oil/cooling water, Temp. controllers
  - **Batch Charging**: Flow meters, pumps, valves, level meters
  - **Discharge/Recirculate**: Pumps and valves for directing flow out of and back to the reactor and storage
- Equipment Modules have one or more **Phases**, Sequences that manipulate control modules to carry out one task
  - **Phases** are the vocabulary of the **Unit Recipes**
  - **Control Module States** are the vocabulary of the **Phases**

Source: https://www.chemicalengineeringpractice.org/

Phases

- Phases manipulate Control Modules to Complete a Task

	Equipment States		
Device	**Idle**	**T1PathOpen**	**T1Charge**
SV-T1	Closed	Open	Open
SV-P1	Closed	Open	Open
P-1	Stopped	Stopped	Running
SV-T2	Closed	Closed	Closed
SV-P2	Closed	Closed	Closed
P-2	Stopped	Stopped	Stopped
FCV-R1	0%	100%	100%



Source: https://www.chemicalengineeringpractice.org/

Unit Modules / Unit Recipes

- **A Unit Modules holds one batch**
  - It is made up of many Equipment Modules
- It can execute one to many **Unit Recipes**
  - **Unit Recipes** are written with **Phases** as their vocabulary
  - A batch reactor would have one recipe for each product
  - A mixer or filter unit might only have one recipe
  - A unit recipe can also be a start up/shutdown sequence of a continous plant
- Unit Recipes resemble Chemist Recipes
  - Good Communication tool for R&D / Production / Engineering
  - Operators can track batches with dynamic displays of the SFC

Source: https://www.chemicalengineeringpractice.org/

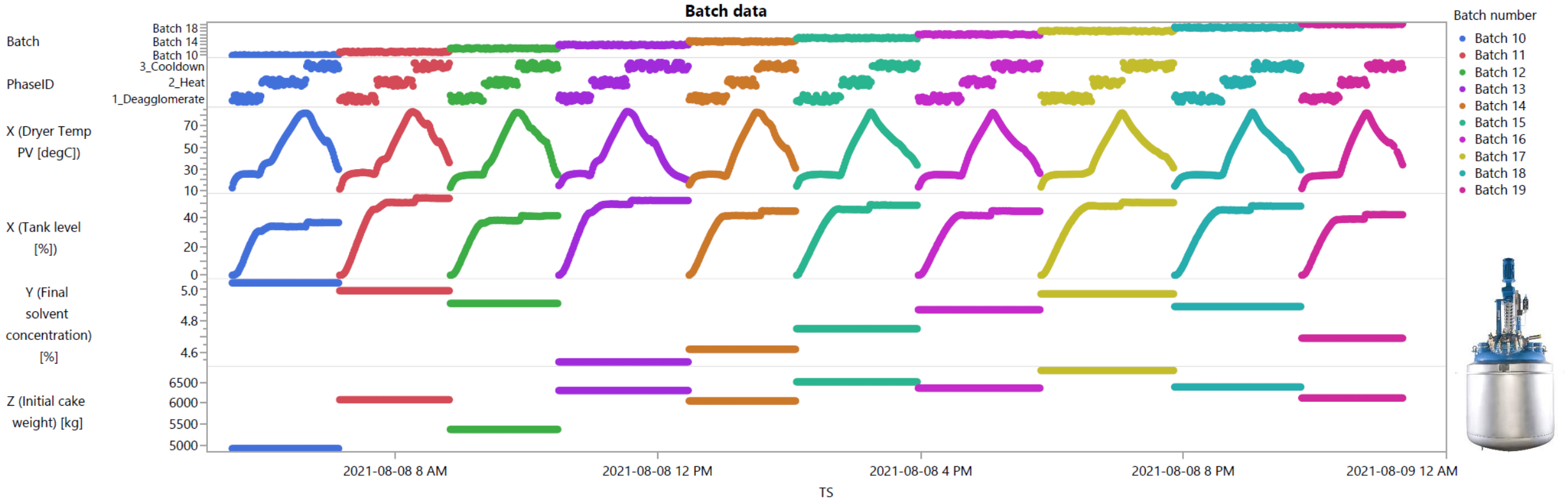
Master Recipes – Generic Unit Recipes

- Unit recipes and phases are specific to **one unit**
- **Master recipes** apply at any scale from lab to pilot to various production vessels
  - Phase parameters are **scaled** (usually to mass) or **unscaled**
  - Similar to extrinsic vs. intrinsic thermodynamic variables.

Variable	Description	Scaled/Unscaled	Master Recipe	Unit Recipe (R-3)
Batch Size	Product mass	Scaled	1000	5500
A_Charge (kg)	Mass A Charged	Scaled	364	2000
B_Charge (kg)	Mass B Charged	Scaled	636	3500
React_Temp (˚C)	Rx Temperature	Unscaled	180	180
React_Wait (min)	Wait time	Unscaled	180	180
Strip_Press (mmHg)	Target pressure for Strip	Unscaled	50	50
Strip_Temp (˚C)	Target Strip tem[.	Unscaled	160	160

Source: https://www.chemicalengineeringpractice.org/

Industrial data and batch processes



https://try.seeq.dev/workbooks?t=TVlfRk9MREVS

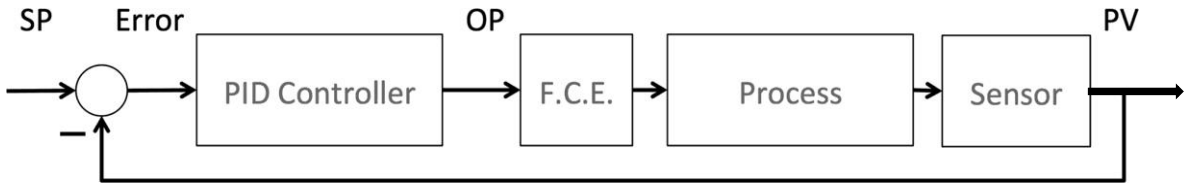
Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Sargent Centre for Process Systems Engineering

Industrial Data Science

Advanced Process Control

Control performance monitoring – Industrial datasets and applications



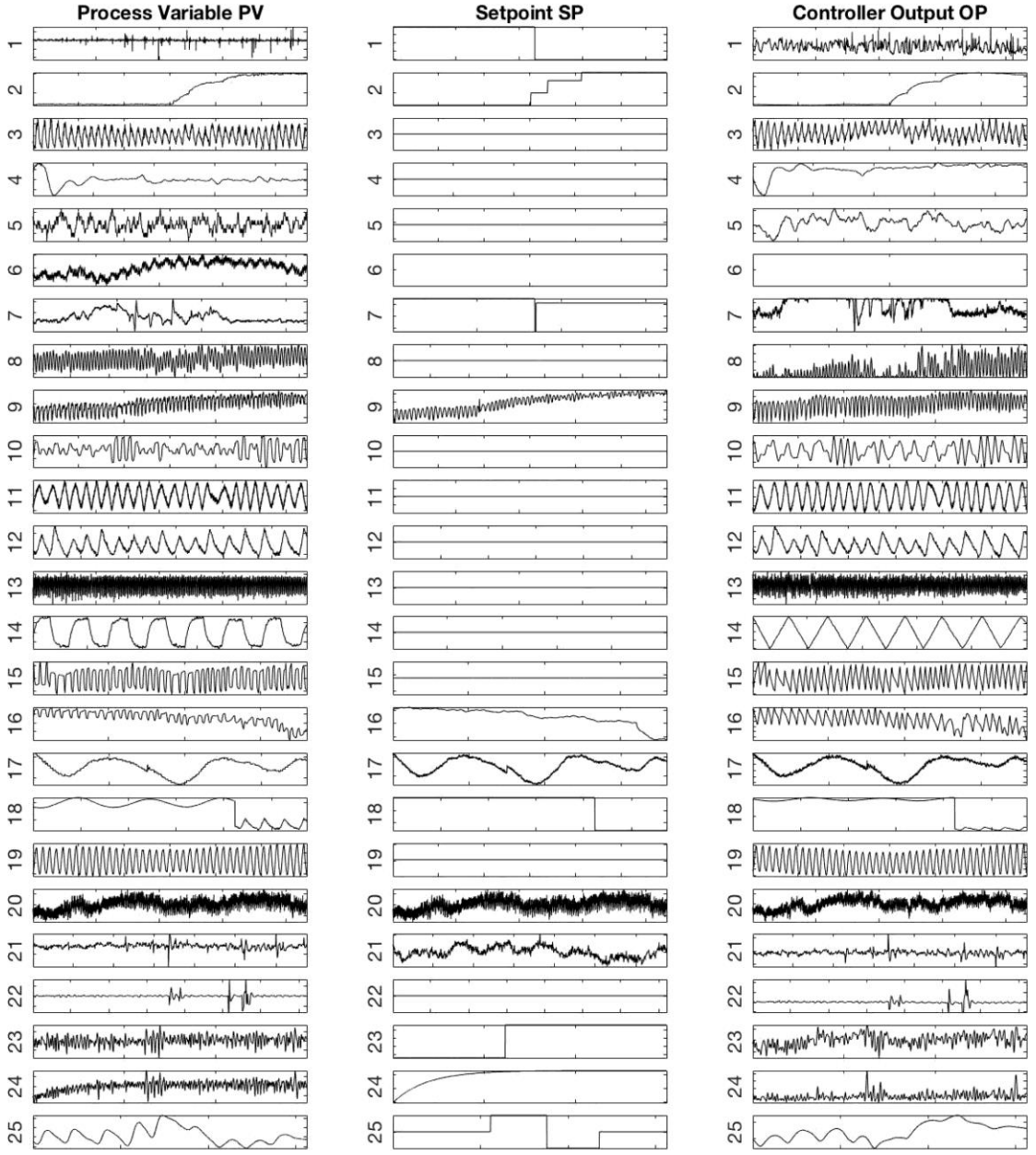
- incorrect tuning settings (SysID add-on)
- valve stiction (add-on)
- saturation (add-on)
- other actuator faults
- sensor faults
- other
- unknown

- Source: [Open Access] Industrial PID Control Loop Data Repository and Comparison of Fault Detection Methods | Industrial & Engineering Chemistry Research (acs.org)

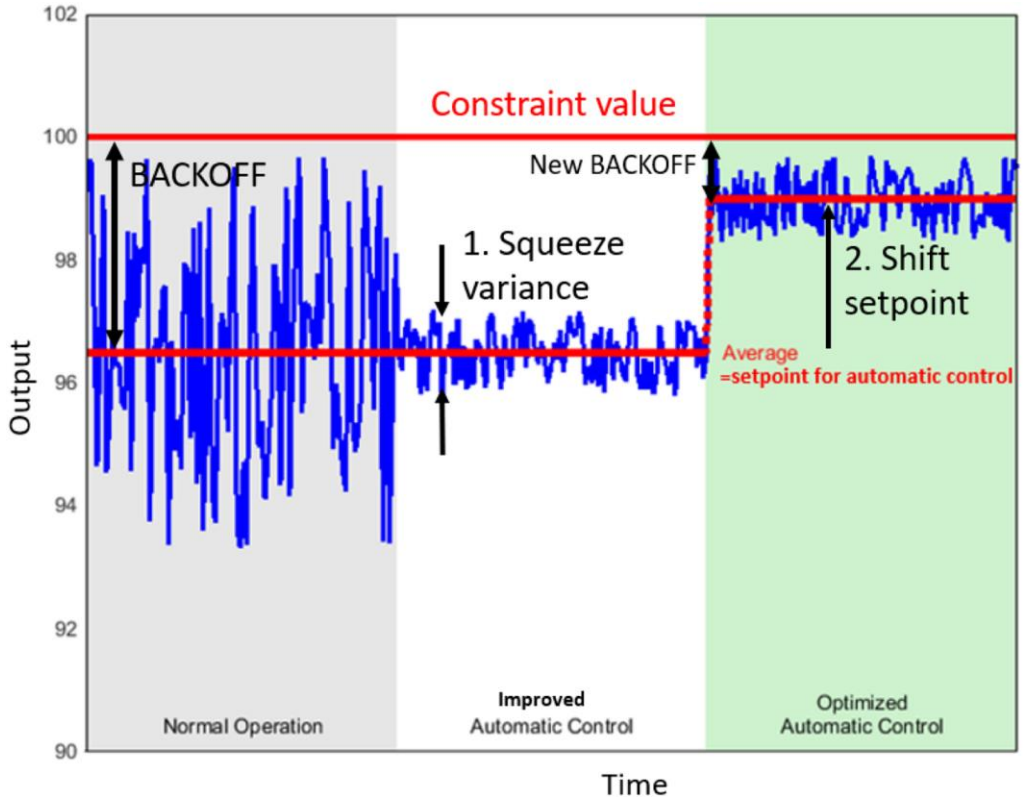
Other references

Industrial datasets and a tool for SISO control loops data visualization and analysis – ScienceDirect

Detection and Diagnosis of Stiction in Control Loops (ualberta.ca)

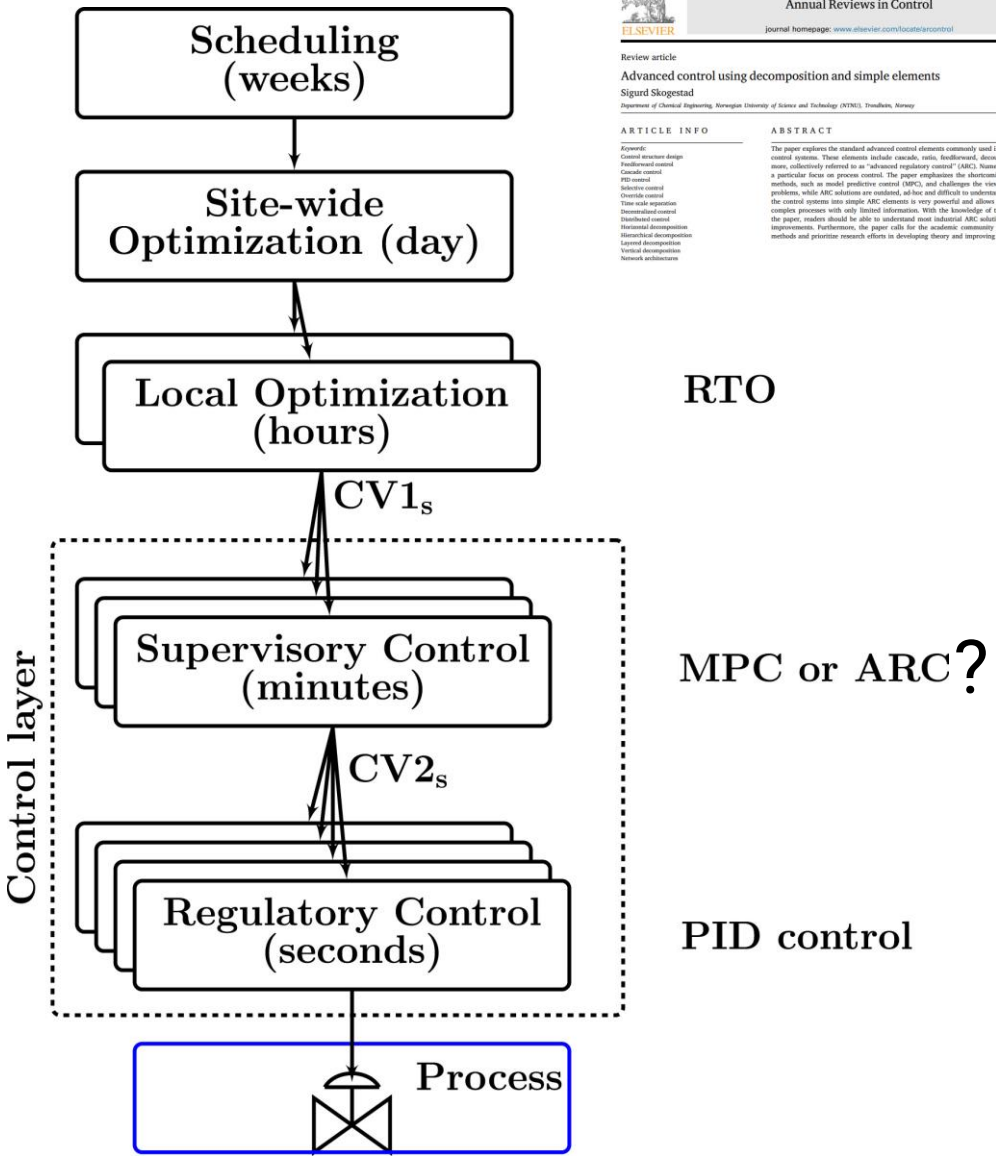


Advanced Process Control



Advanced control using decomposition and simple elements *S. Skogestad*

Annual Reviews in Control 56 (2023) 100903
https://www.sciencedirect.com/science/article/pii/S1367578823000676



Annual Reviews in Control

Review article

Advanced control using decomposition and simple elements

Sigurd Skogestad

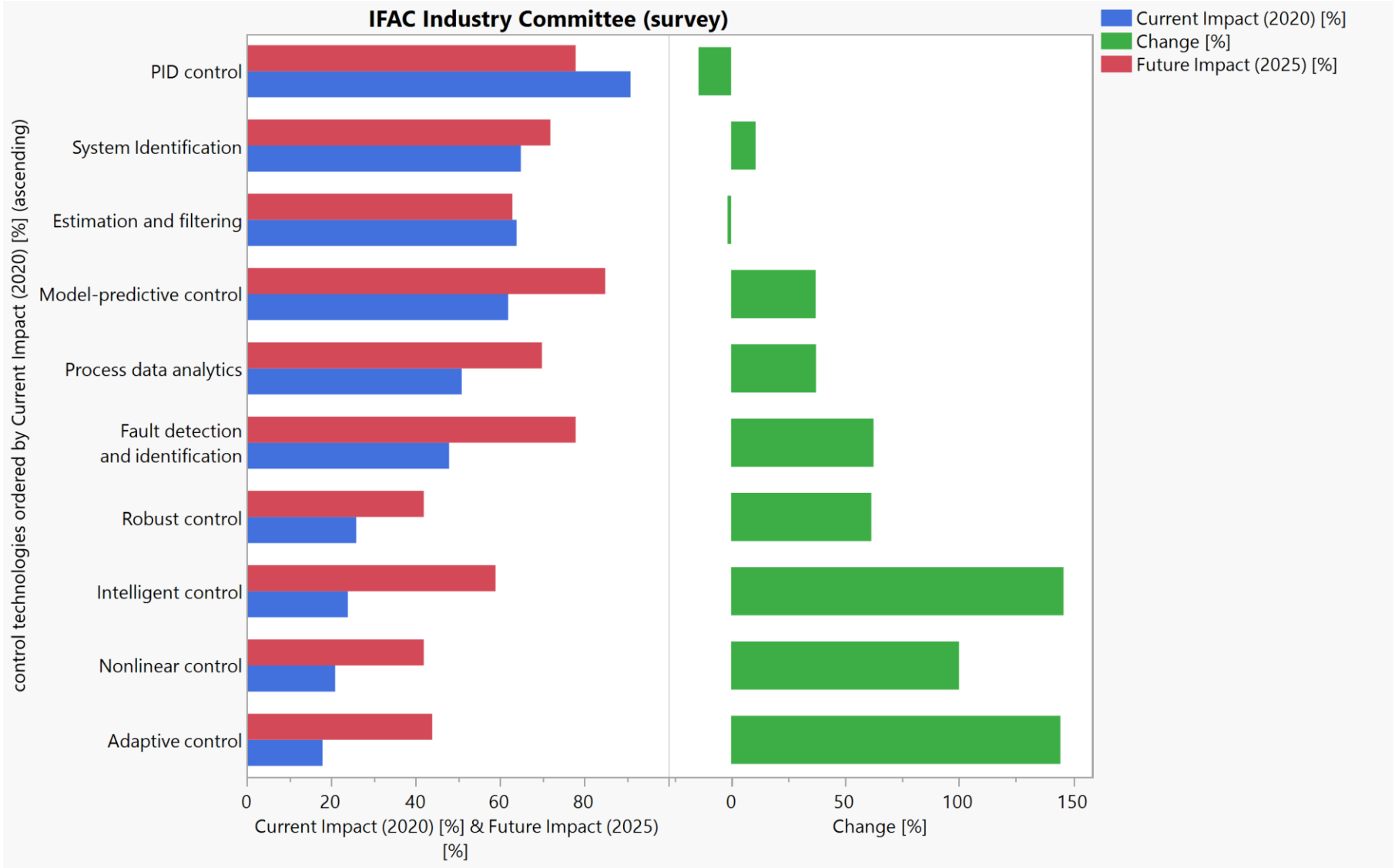
ARTICLE INFO

ABSTRACT

Control technologies and their impact (present and future)

- Industry engagement with control research: Perspective and messages

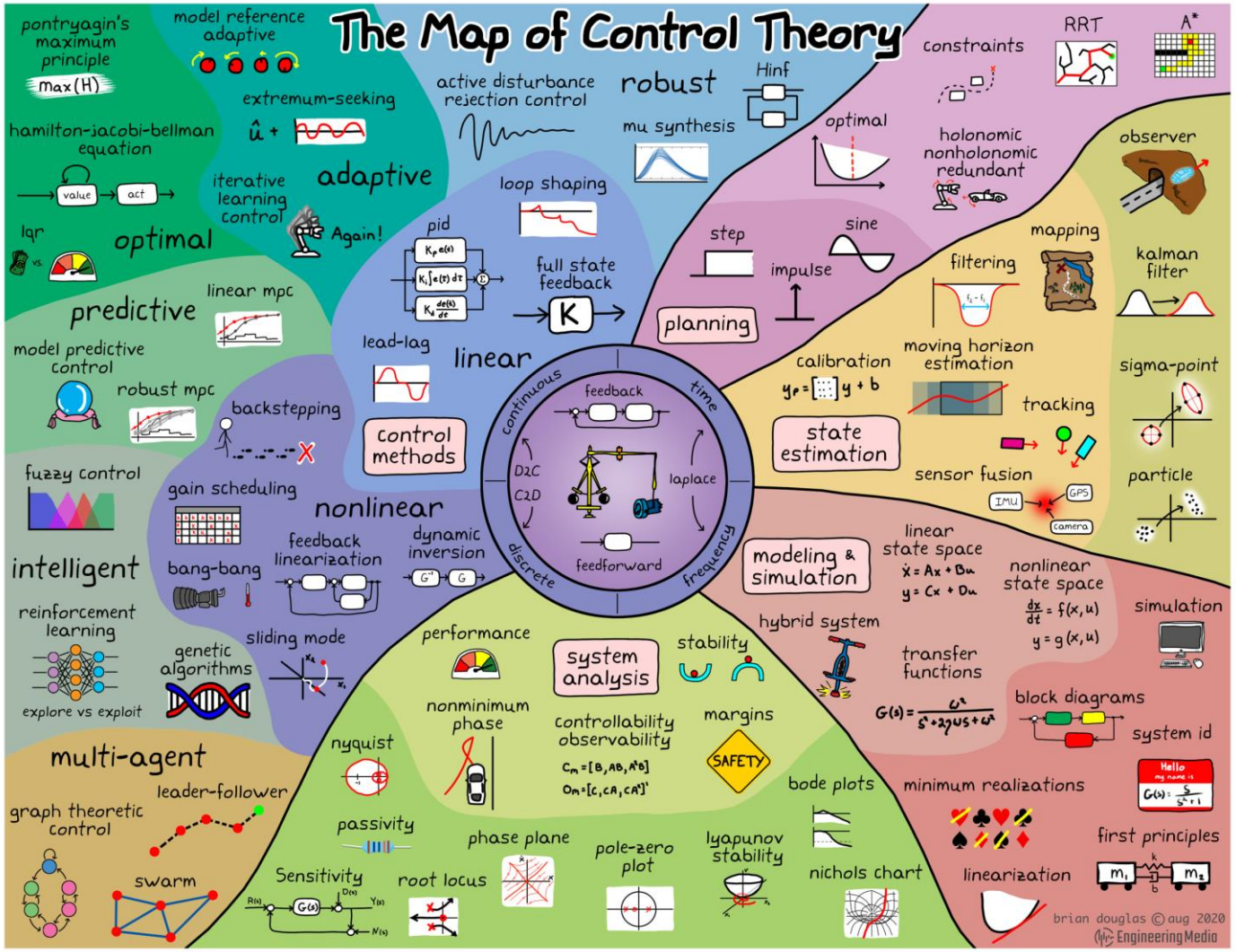
https://www.sciencedirect.com/science/article/pii/S1367578820300080



Percentage of survey respondents indicating whether a control technology had demonstrated ("Current Impact") or was likely to demonstrate over the next five years ("Future Impact") high impact in practice.

Online resources to learn process control

- https://www.chemicalengineeringpractice.org/ - Prof.
- https://engineeringmedia.com/ - Brian Douglas
- https://apm.byu.edu/prism/index.php/Site/OnlineCourses
- http://www.pc-education.mcmaster.ca/
- Process Control. A Practical Approach. Myke King
- Applying S88. Batch Control from a User's Perspective. Jim Parshall and Larry Lamb



Sargent Centre for Process Systems Engineering

Industrial Data Science

Data integration

The Purdue model (Automation pyramid)

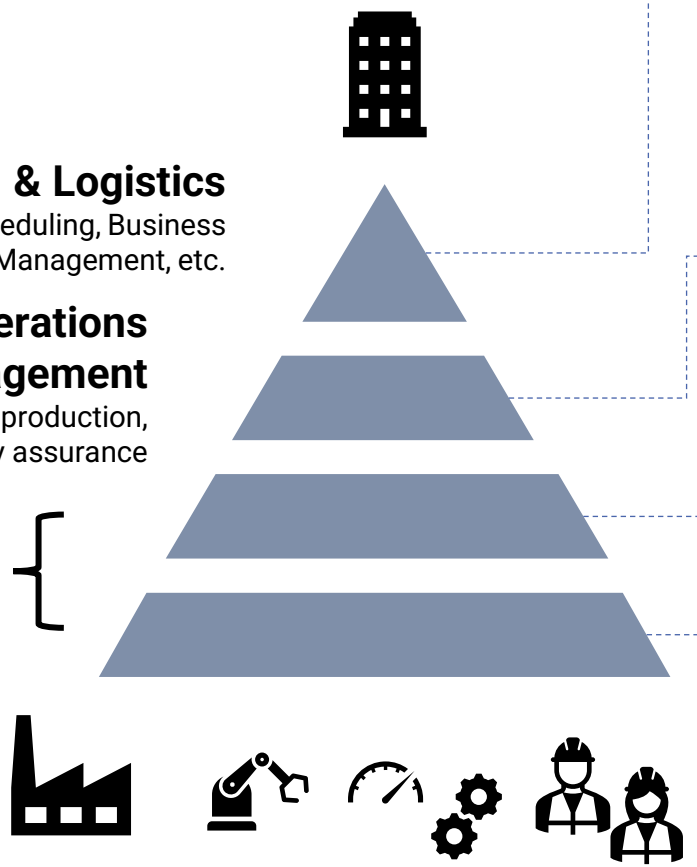
IT – Informational Technology

OT – Operational Technology

Business Planning & Logistics
Plant Production Scheduling, Business Management, etc.

Manufacturing Operations Management
Dispatching production, detailed production, scheduling, reliability assurance

Manufacturing Control
Basic control, supervisory control, process sensing, process manipulation.



4 - Establishing the basic plant schedule production, material use, delivery, and shipping
Determining inventory levels.

Time Frame: Months, weeks, days, shifts

3 - Work flow / recipe control to produce the desired end products. Maintaining records and optimizing the production process.

Time Frame: Shifts, hours, minutes, seconds

2 - Monitoring, supervisory control and automated control of the production process

1 - Sensing the production process, manipulating the production process

0 - The physical production process

LIMS, ERP & Sensor data

Enterprise Resource Planning (ERP)

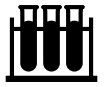
- Purchase Orders
- Finished Products and Raw Materials
- Suppliers and Inventory Levels

LIMS

- Quality of intermediate
- Specification Limits
- Control Limits

Sensor Data

- P, F, T
- P_{SP} , F_{SP} , T_{SP}
- Contextual data

Tabular data: KPIs  LIMS  ERP

BatchID	LotID (PO number)	Grade (P1, P2, P3)	Quality (pass/no pass)	Suplier
…	…	…	…	…
…	…	…	…	…

Rows are independent*

Data integration

(not data duplication)

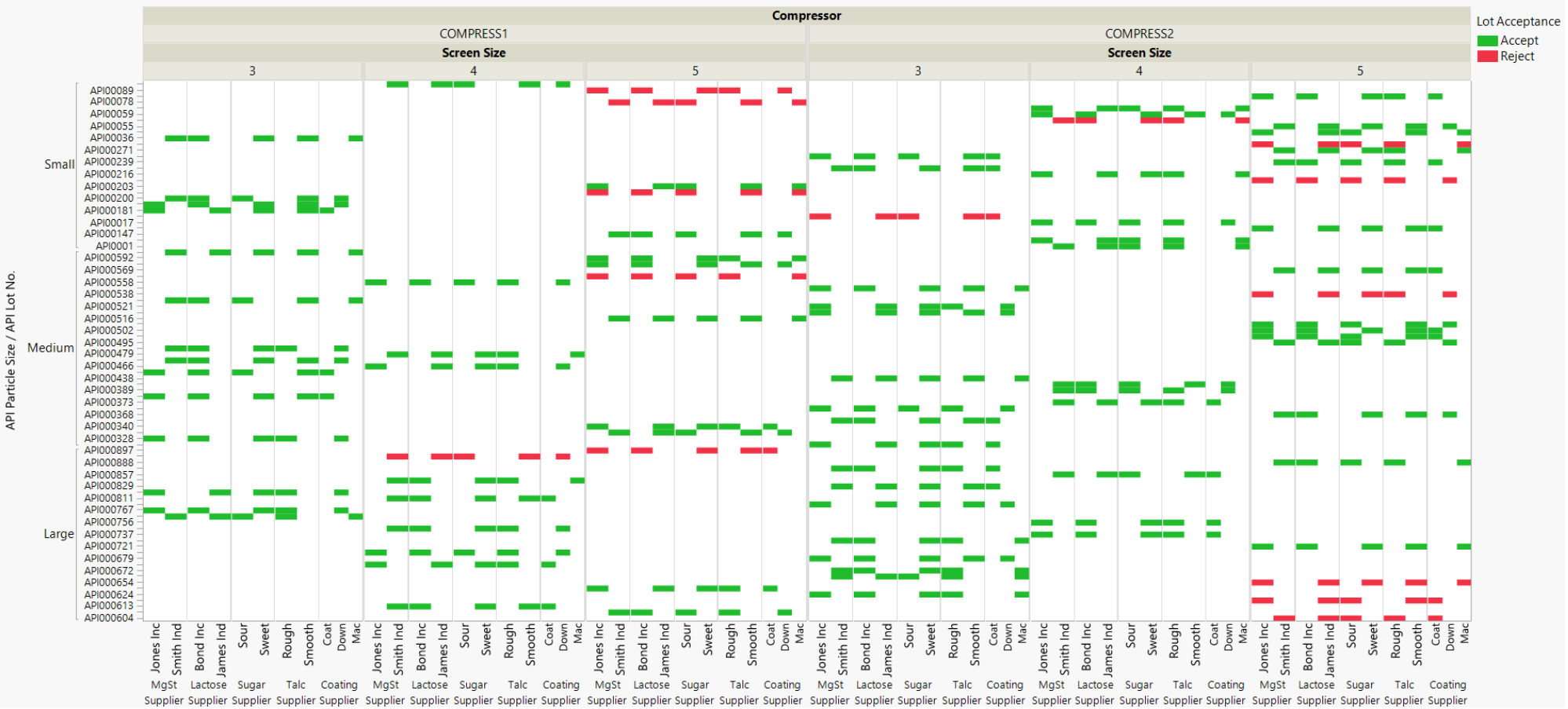
Sensor data: tags in historians …

Datetime	PI_505.PV (Pressure) [bar]	TI_203.PV (Temperature) [degC]	FIC_101.MV (Valve position) [%]	FI_101.PV (Flowrate) [m3/min]
…	…	…	…	…
…	…	…	…	…

Rows are <u>not</u> independent (sampling, dynamics…)

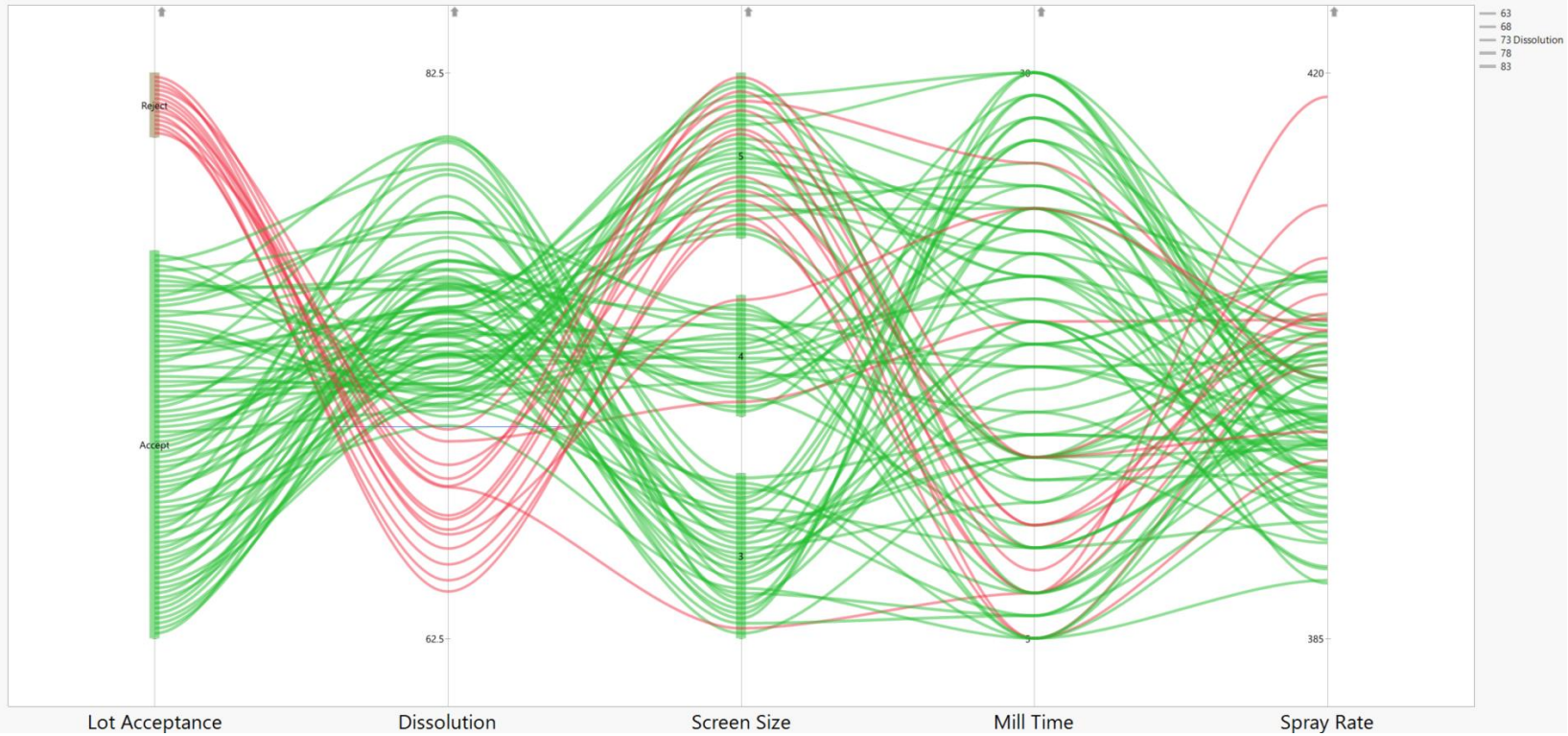
To leverage the full potential of the information captured by the business. The manufacturing data can be complemented with the operations management. This type of data integration can bring insights on information such as the supplier of the raw materials to investigate possible issues

ERP data example



ERP data contains multiple types of information. It is itself a consolidation of different types of data from different functions (controlling, operations, supply chain, quality, etc). Often, a good portion if the inputs to the system are manual. Among the challenges we have: finding out the latest up to date, name recoding (versioning), applying the right filters, different maturity depending on site.

Screening SAP-type data



Predictor	Dissolution Contribution	Portion	Rank
Screen Size	157.567	0.2760	1
Mill Time	125.876	0.2205	2
Spray Rate	61.262	0.1073	3
Blend Time	39.452	0.0691	4
Coating Viscosity	38.897	0.0681	5
Blend Speed	36.890	0.0646	6
Atom. Pressure	23.873	0.0418	7
Inlet Temp	15.500	0.0272	8
API Particle Size	12.864	0.0225	9
Force	10.962	0.0192	10
Lactose Supplier	10.167	0.0178	11
Exhaust Temp	9.853	0.0173	12
Talc Supplier	8.906	0.0156	13
Coating Supplier	5.972	0.0105	14
Sugar Supplier	5.556	0.0097	15
Compressor	4.585	0.0080	16
MgSt Supplier	2.629	0.0046	17

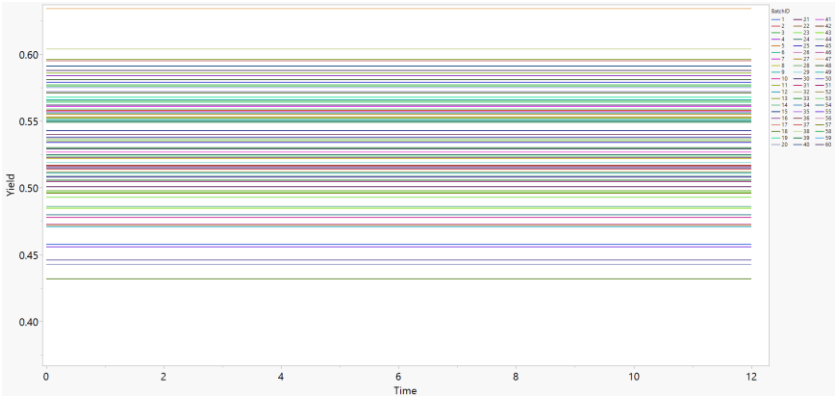
Since multiple types of information are maintained in ERP, also multiple data formats and types are available. Our tools need to cope with continuous and discrete data as well as with tabular and time series formats.

The right filtering conditions must be applied when using screening methods. Understanding of the discrete elements is key

Fermentation example (quality + process)

- Quality data per batch
- Measurement is taken at the end of the batch
- Repeatability and real sampling time must be verified

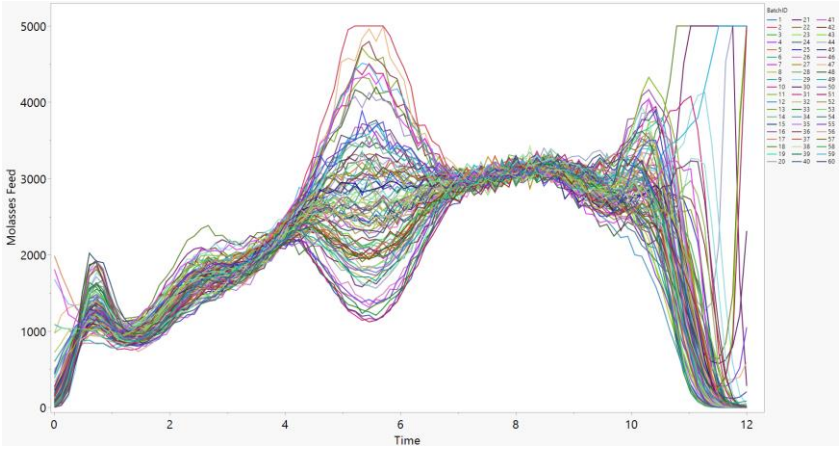
LIMS



+

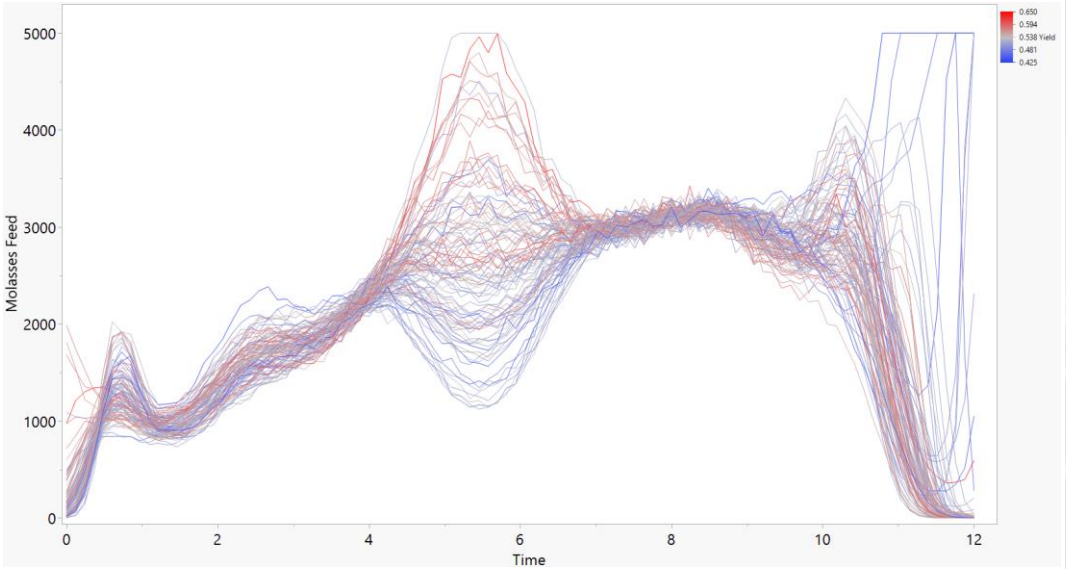
Sensor Data

- Process data with a profile per batch.
- A full profile of a given process variable belong s to each batch



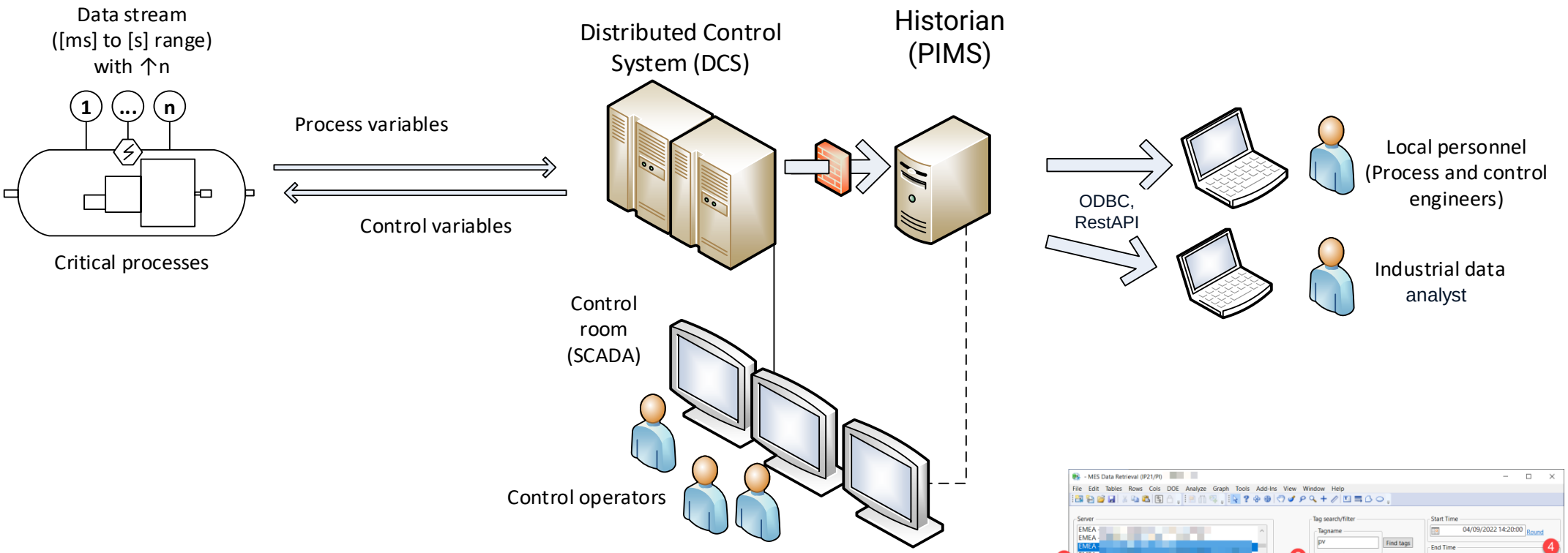
=

Integrated Data

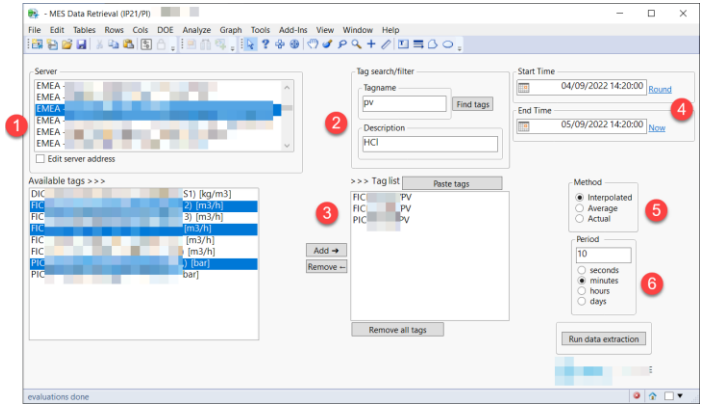


Join of process data and quality data. Color now belongs to quality. You need a unique batch ID and in-sample time

Industrial data (open-source) native connectors



tagreader 4.2.5



https://github.com/equinor/tagreader-python

https://github.com/industrial-data/JMP-MES-connector

Sargent Centre for Process Systems Engineering

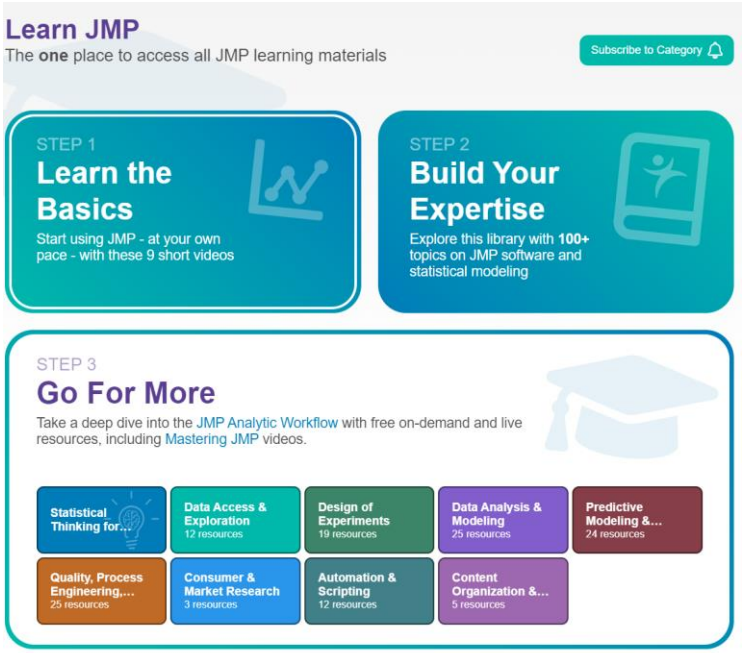
Industrial Data Science

Hands-on

First steps with JMP

Getting data into JMP, several options:

- ★ Directly from Excel (add-in) [1.5 min]
- ★ Excel file [3 min]
  - Text file
  - Multiple text files
  - Optional: External databases
  - IP21/PI data extraction



All learning material: https://community.jmp.com/t5/Learn-JMP/ct-p/learn-jmp

1-2 min tutorials: https://www.jmp.com/en_us/learning-library.html

Control charts: https://community.jmp.com/t5/Mastering-JMP/Analyzing-Improving-and-Controlling-Process-Stability-and/ta-p/500753

Training material for the Federal Government: www.jmp.com/fedgov

Online course (+certification): https://www.jmp.com/en_us/online-statistics-course.html

SPC training: https://community.jmp.com/t5/Learning-Center/JMP-Statistical-Process-Control-Course/ta-p/575330

Demos for visual analytics:

- ★ Graph builder for manufacturing & SAP data (28 min) [dataset I, II]
- ★ Process data and ML (tags, sensors) (45 min) [dataset]

Summary

- Machine learning concepts are not new for chemical engineers
  - Supervised learning → Regression
  - Unsupervised learning→ Statistical Process Control charts
- Process, control and automation knowledge is essential to ask relevant questions and interpret results in process industries
- Democratization of machine learning requires self-service access to all process data and analytics
- One of the main applications of machine learning is to understand chemical processes faster (e.g., selecting relevant sensors (tags) amongst noise)
- If the answer is obvious, question should be redefined. Example: data analysis points to mass or heat balance, which are already known. Target variable should be redefined to situations where assumptions do not hold (i.e., discrepancy models).

Sargent Centre for Process Systems Engineering

Day 2

Monitoring and screening process data

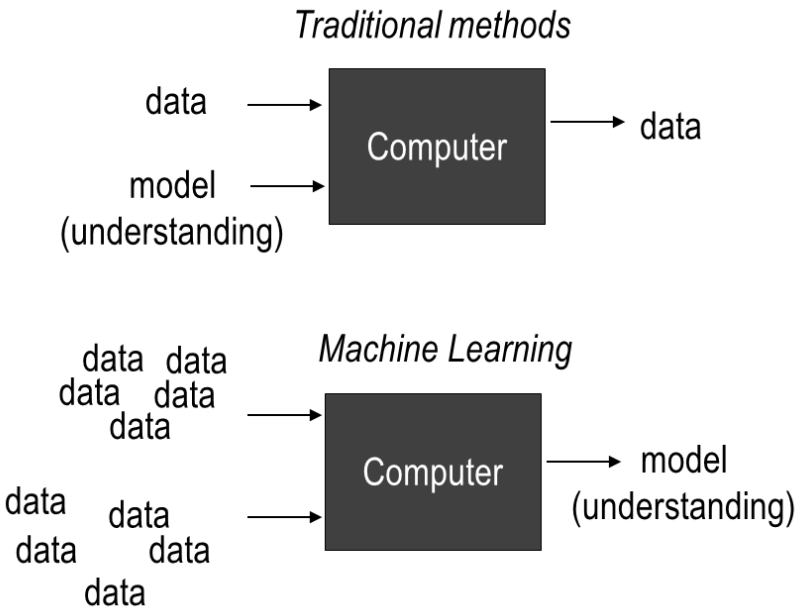
Business impact is measured in tones, not lines of code

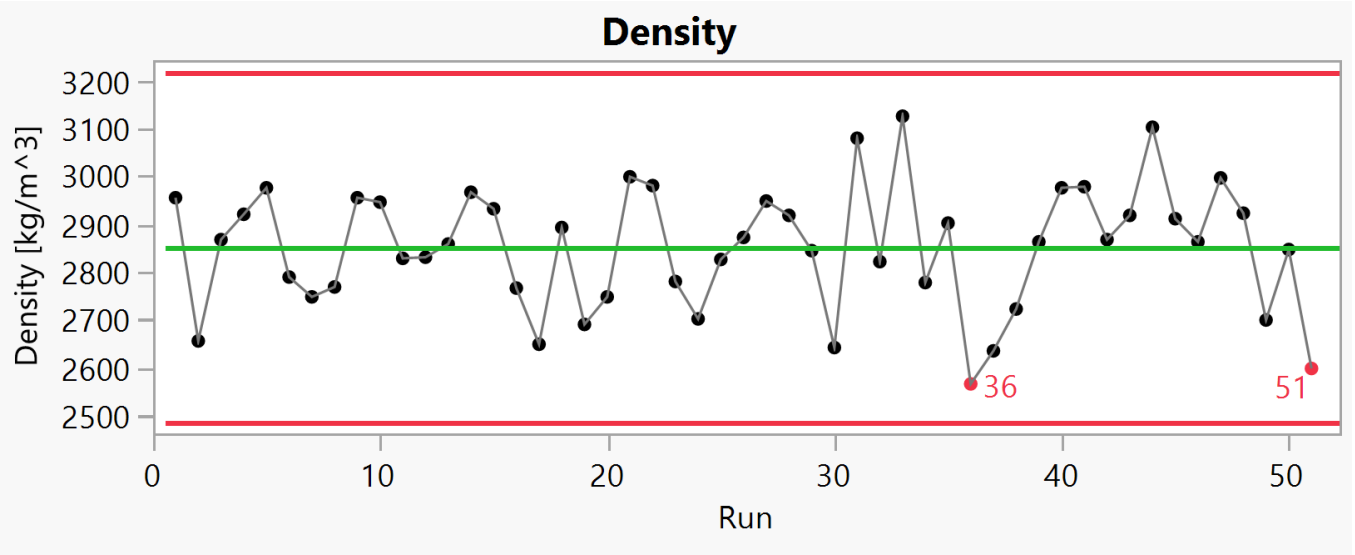
Sargent Centre for Process Systems Engineering

Monitoring and screening process data

Anomaly detection

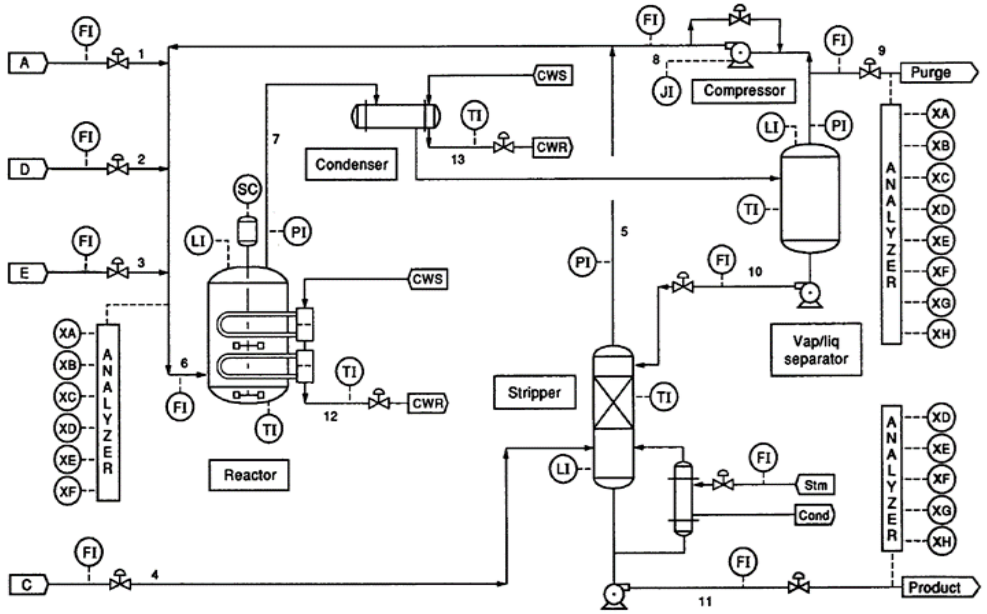
Control charts (Unsupervised learning)





Control charts are a form of unsupervised models where only input training data is given and a statistical model is built (mean and standard deviation, since controlled processes should follow a normal distribution). Finally, we can classify data points as in- or out-of-control.

Anomaly detection example: Tennessee Eastman Plant



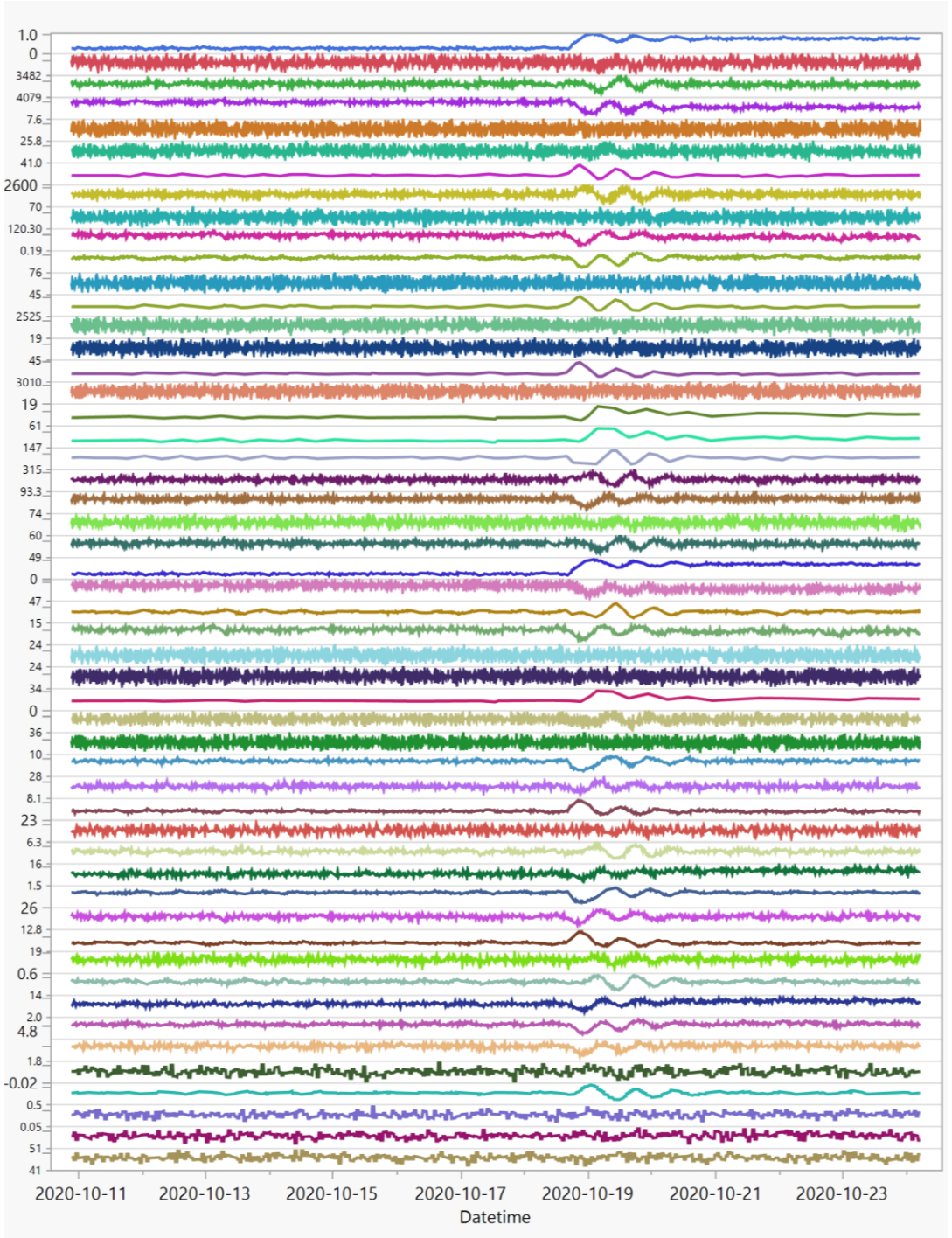
Tennessee Eastman plant has transitioned between two regimes, multiple measurements have changed.

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

PV

MV

Quality (online)

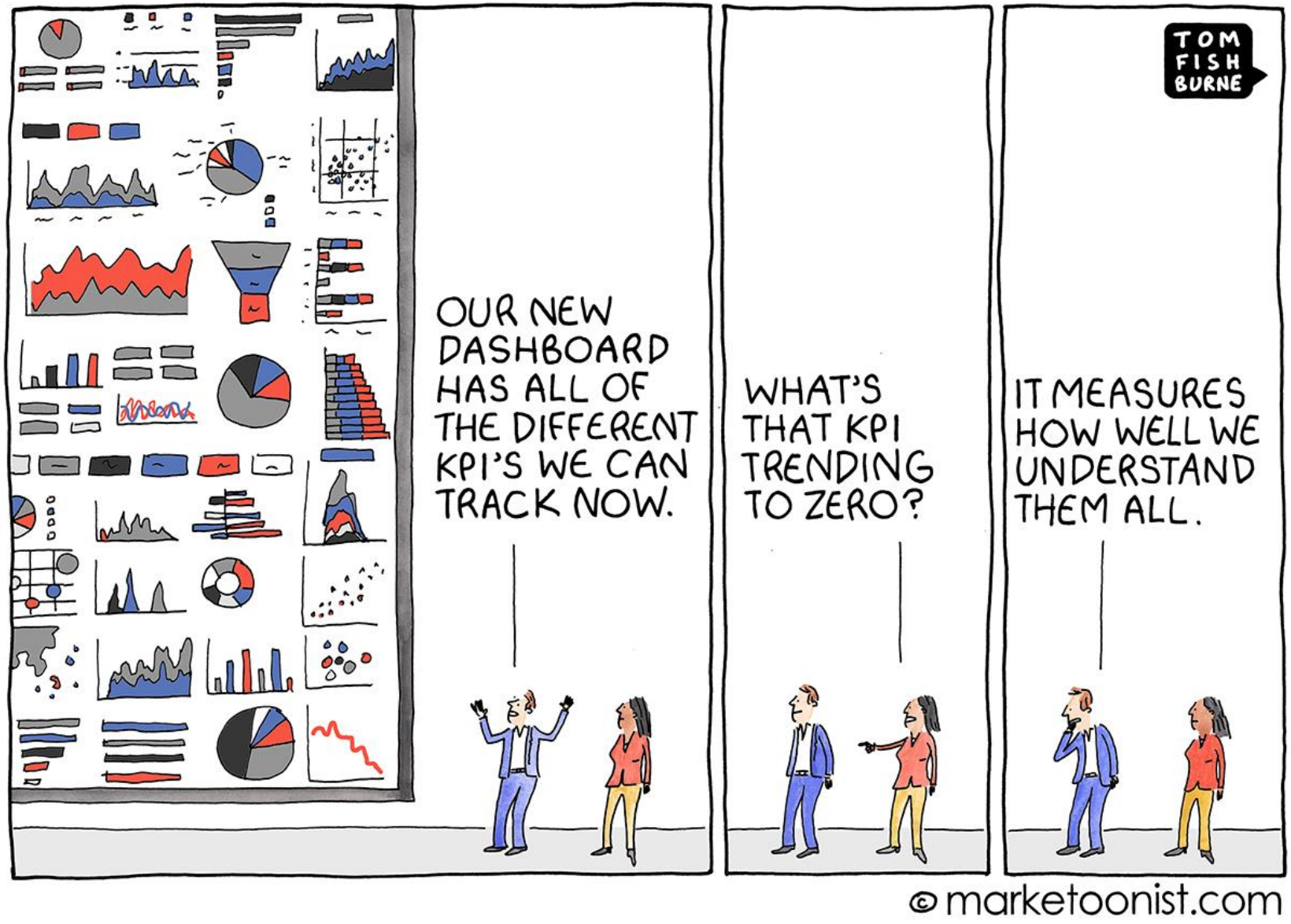


Sargent Centre for Process Systems Engineering

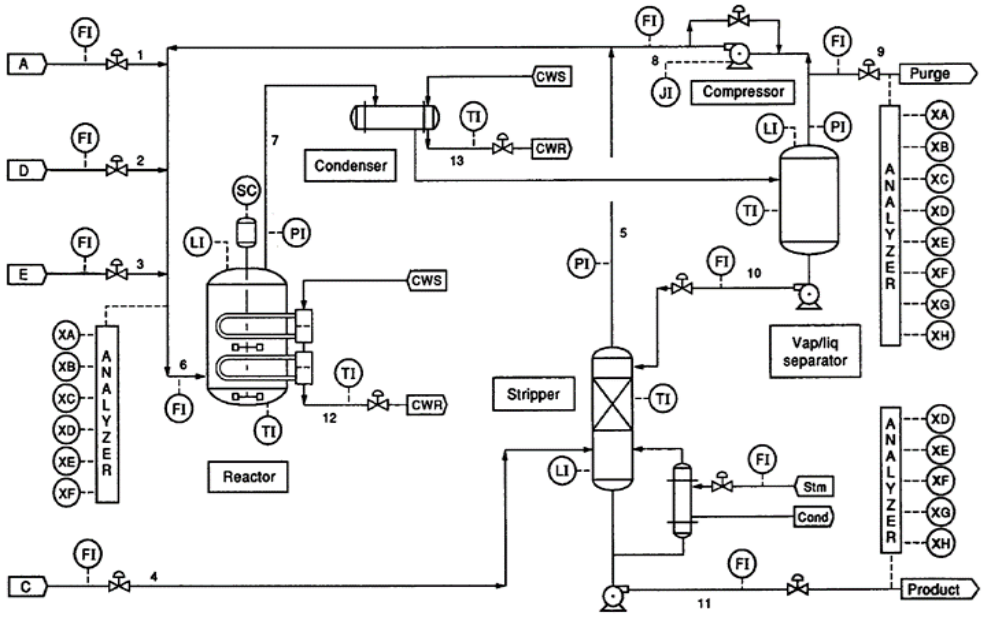
Imperial College London

UCL

Let’s check our KPIs



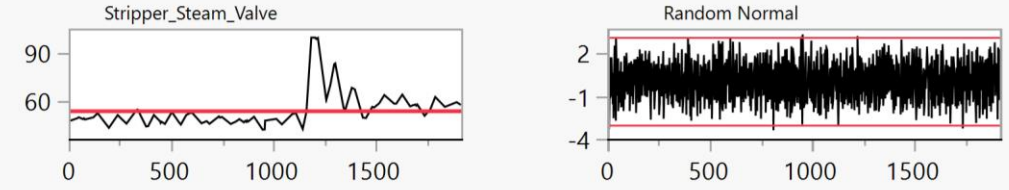
Anomaly detection (univariate analysis)



Univariate analysis can be used to identify major process changes, specially when interactive data filters are applied.

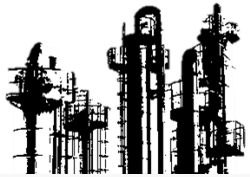
Column	Variability: Stability Index	Variability: Within Sigma	Variability: Overall Sigma	Summary: Mean	Summary: Count	Control: Alarm Rate
Stripper_Steam_Valve	44.56	0.23792	10.6022	54.0364	1920	0.97240
A_Feed_Flow_Stream_1	20.31	1.26077	25.608	45.47	1920	0.99688
Compressor_Recycle_Valve	4.93	0.2212	1.09144	22.1603	1920	0.35156
Purge_Valve_Stream_9	2.91	0.81064	2.36045	39.1521	1920	0.27448
Total_Feed_Flow_Stream_4	1.89	1.21906	2.30525	59.8505	1920	0.09635
E_Feed_Flow_Stream_3	1.85	0.36543	0.67572	53.8437	1920	0.08490
D_Feed_Flow_Stream_2	1.37	0.44069	0.60397	63.0226	1920	0.02500
Reactor_Cooling_Water_Flow	1.07	0.58194	0.62502	41.0625	1920	0.00885
Condensor_Cooling_Water_Flow	1.00	1.47692	1.47162	18.1105	1920	0.00365
Stripper_Liquid_Product_Flow_Stream_11	1.00	2.38078	2.37097	46.5732	1920	0.00156
Random Normal	0.99	1.01439	1.00409	0.04311	1920	0.00156
Separator_Pot_Liquid_Flow_Stream_10	0.97	2.94101	2.8672	38.0943	1920	0.00156

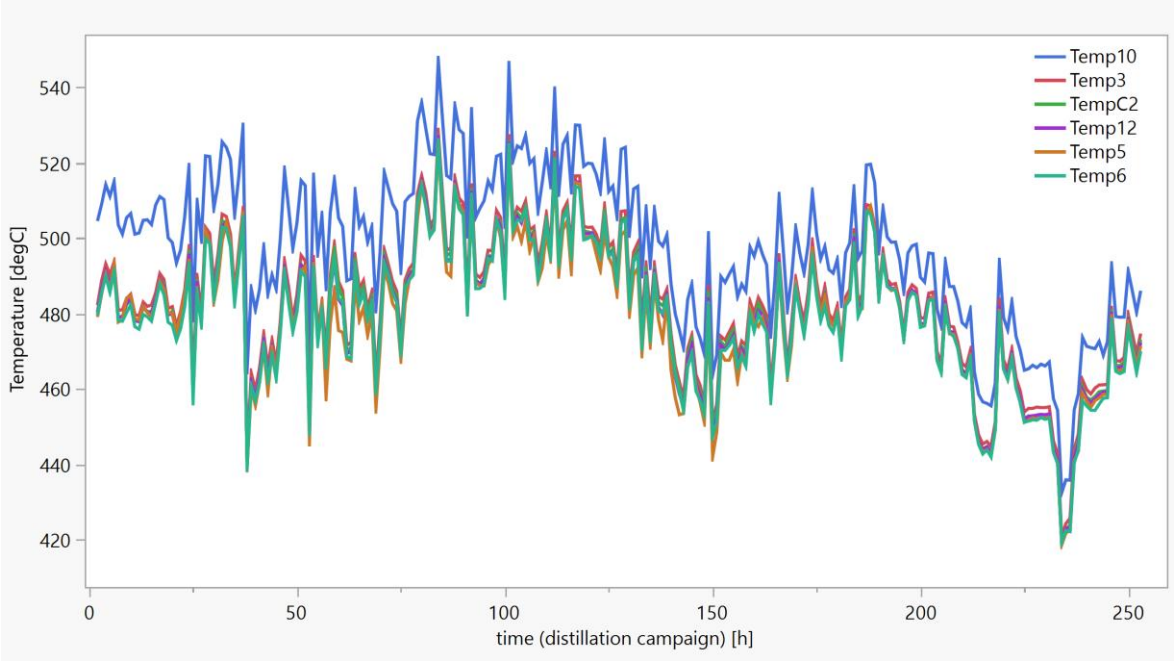
Charts for Selected

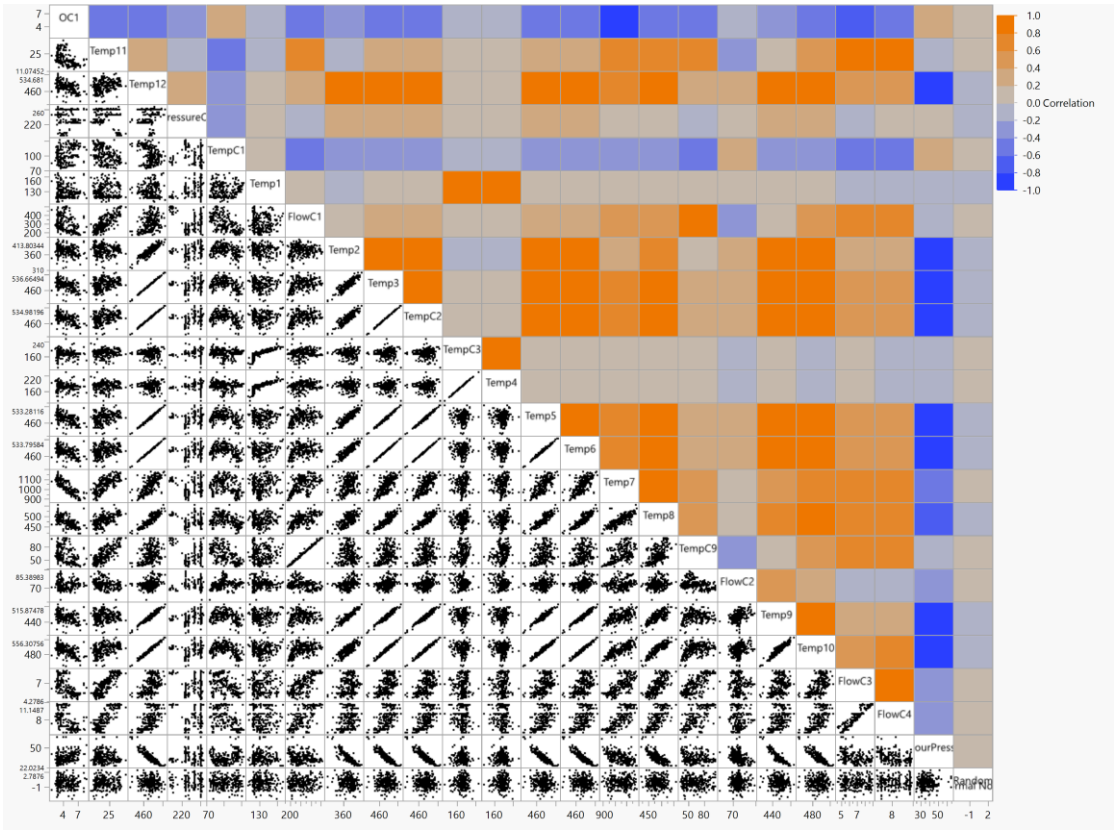


Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Redundant sensors





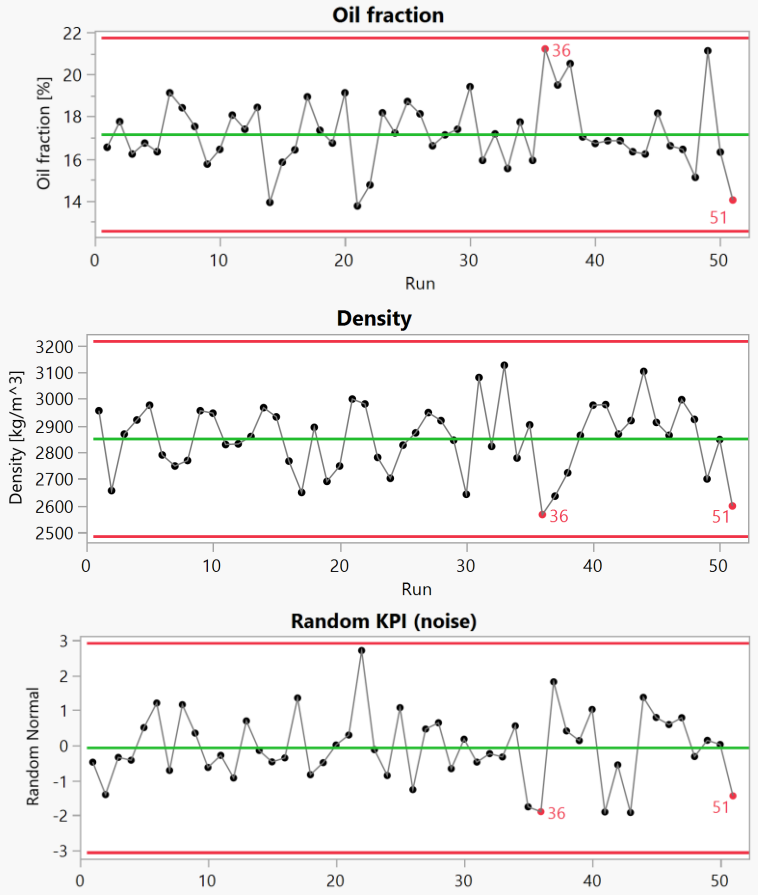


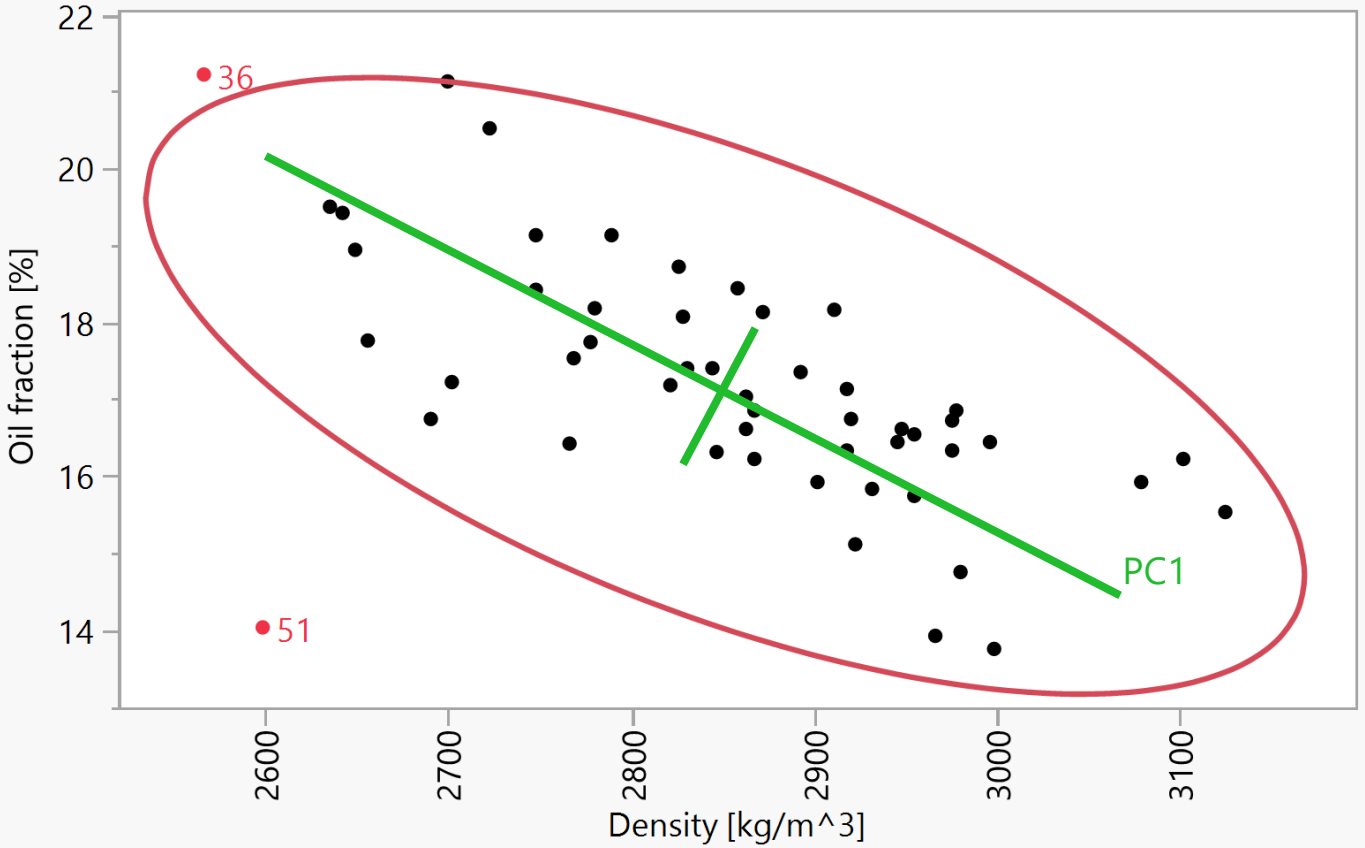
A common example of multivariate analysis is to monitor multiple redundant sensors. When several thermocouples are used to measure a critical temperature, these can be summarized by taking the average of all the sensor readings. The average is a linear combination of all these terms with equal weight. This way, the information is being reduced to one latent variable, the temperature we want to monitor. If a big variation exists between the average of the sensors and one thermocouple, in particular, an alert can be triggered. This reasoning is the same behind principal component analysis (PCA), and it has been widely used for multivariate process analysis, monitoring, and control.

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

+info:
https://learnche.org/pid/

Multivariate control charts (PCA)

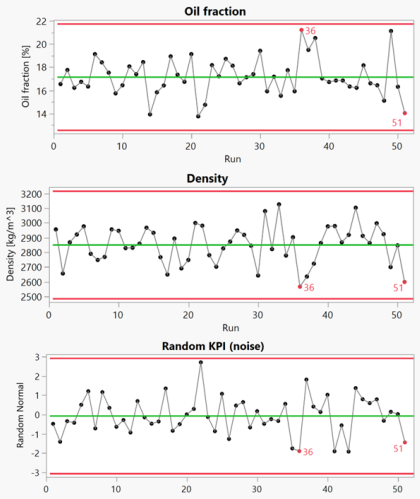


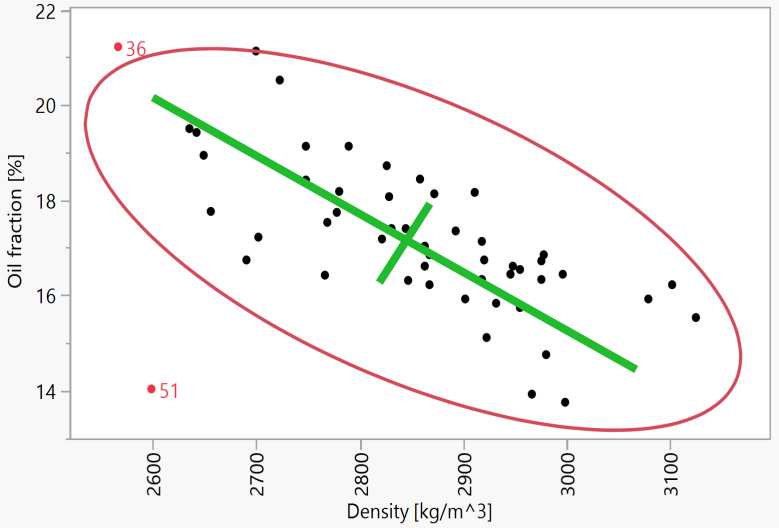


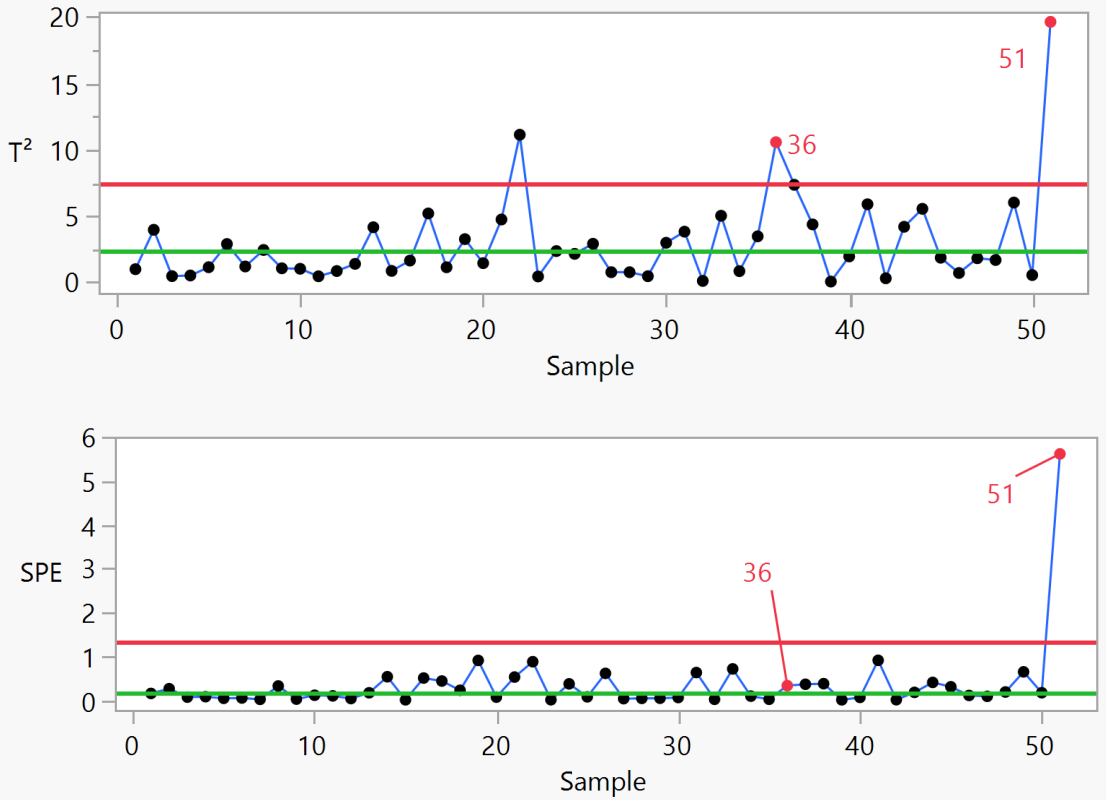
Highlighted samples appeared to be in control (individual control charts). Co-linerity among process variables can be used to detect multivariate anomalies.

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Multivariate control charts (T2 and SPE)



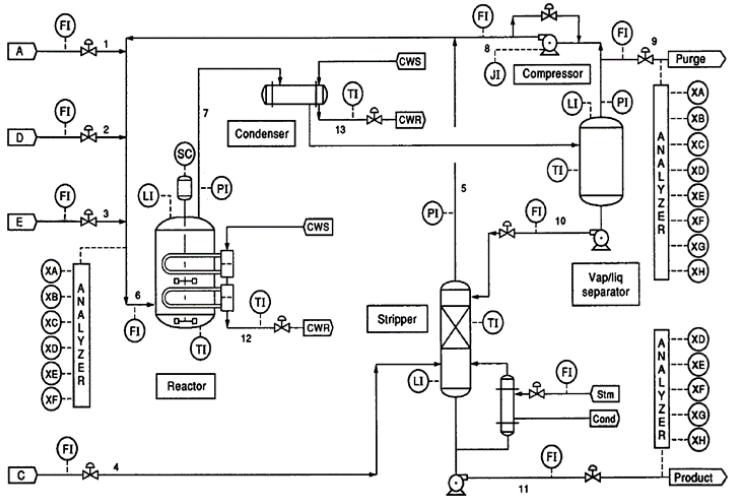


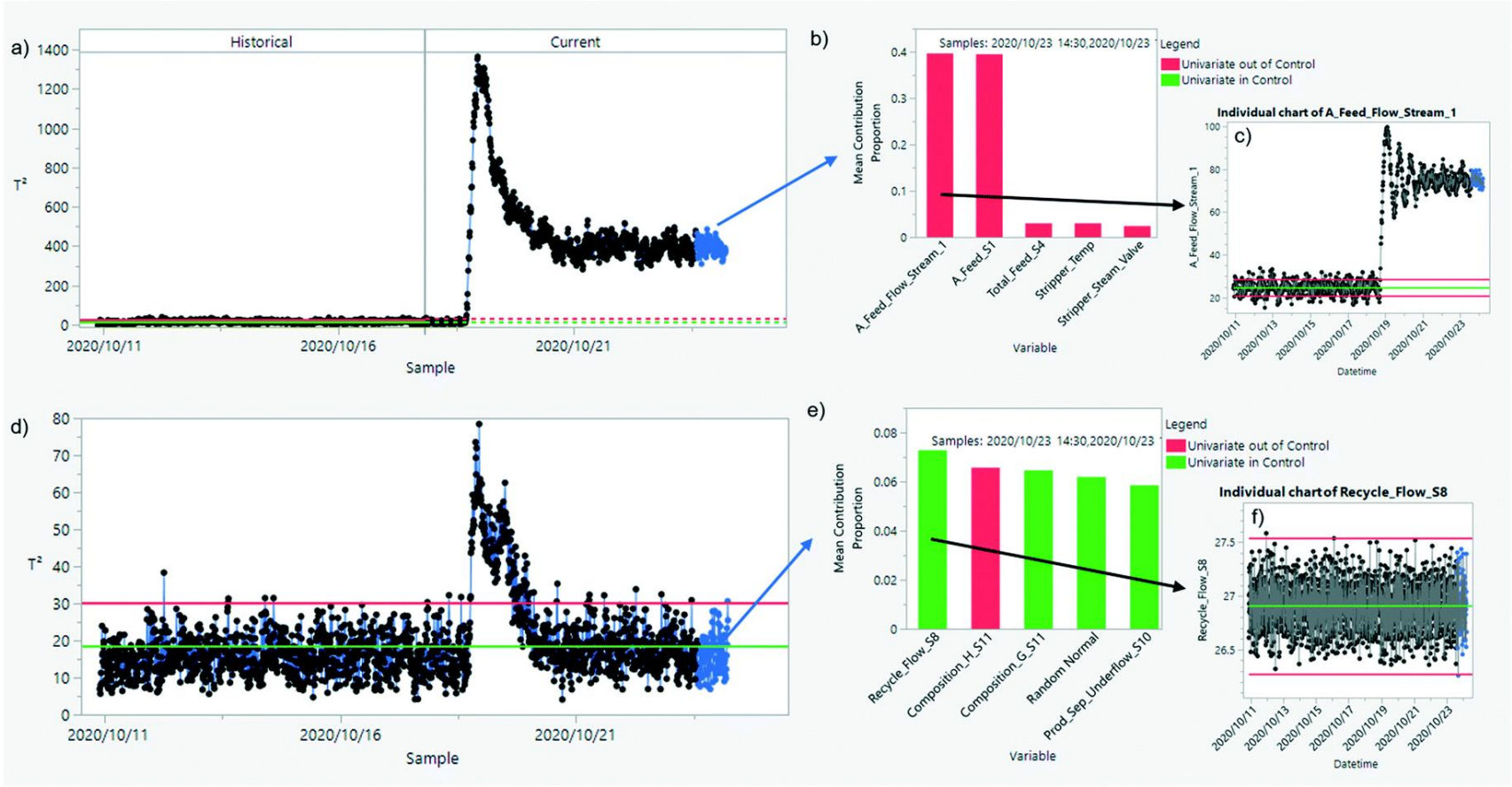


T2 and SPE are two popular metrics to identify outliers. T2 identifies points that are far from previous data while SPE focuses on those points that do not follow the co-linearity identified across redundant sensors

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Anomaly detection (multivariate analysis)



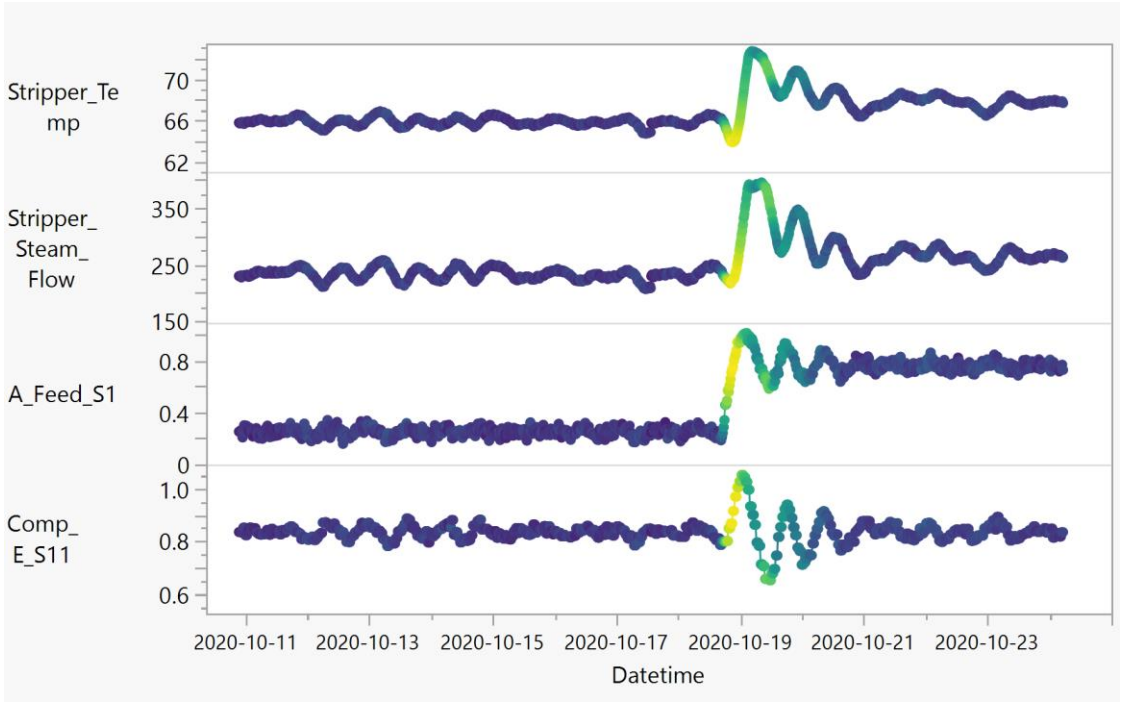


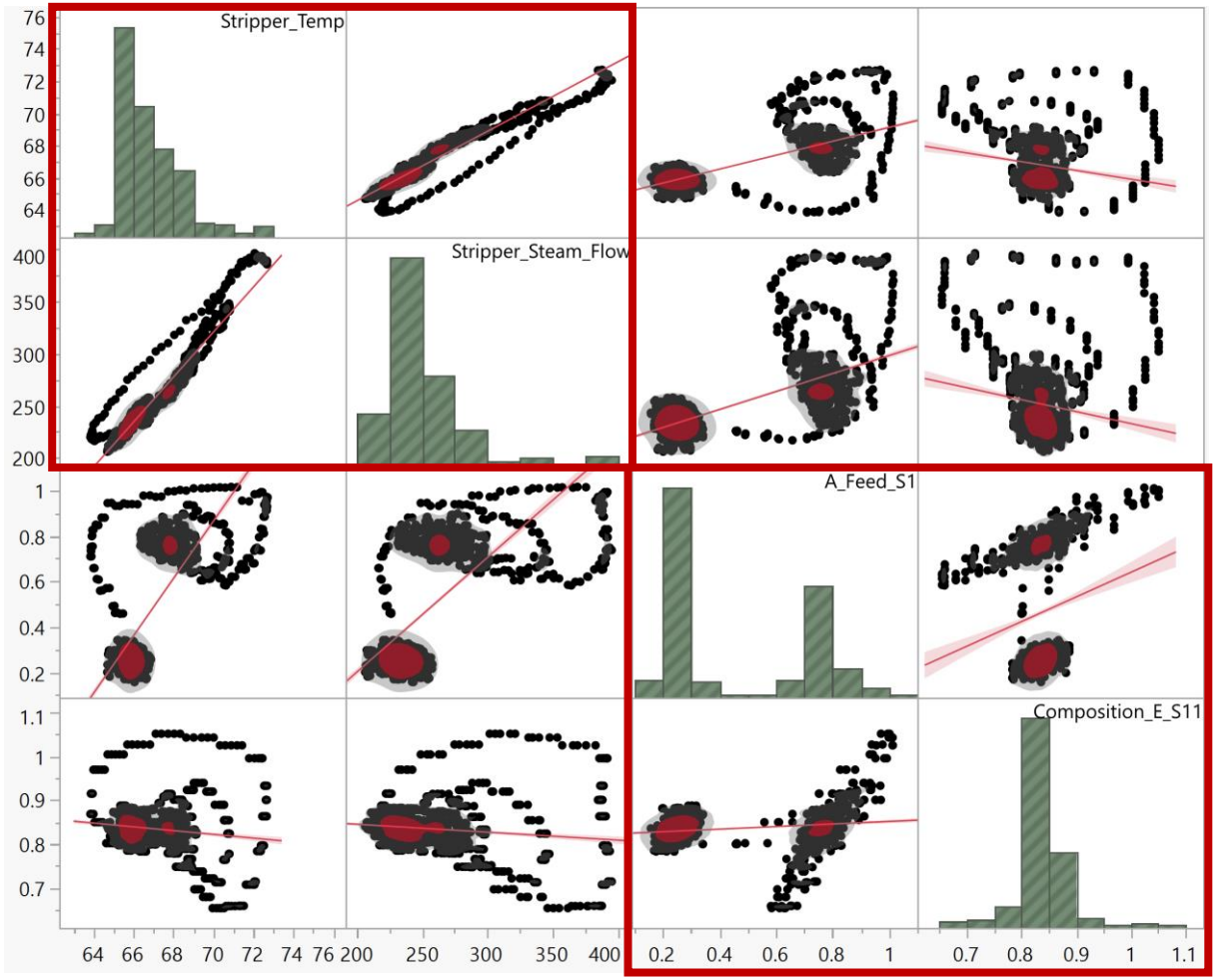
A transition between two steady-state regimes is detected using a multivariate control (PCA). If the model is built using historical data before the perturbation (a) the step changes in the feed flow of chemical A (b and c) are found in the current dataset for the points highlighted in blue. If all historical data is used to build the model (d) the contribution of recent data points in blue (e) shows signals close to random noise. The plant wide control in the simulation stabilizes the control loops and anomalies are only seen in the transition period, even though the plant is operating in a different state for chemical A.

Sargent Centre for Process Systems Engineering | Imperial College London | UCL

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Limitations of multivariate analysis with process data





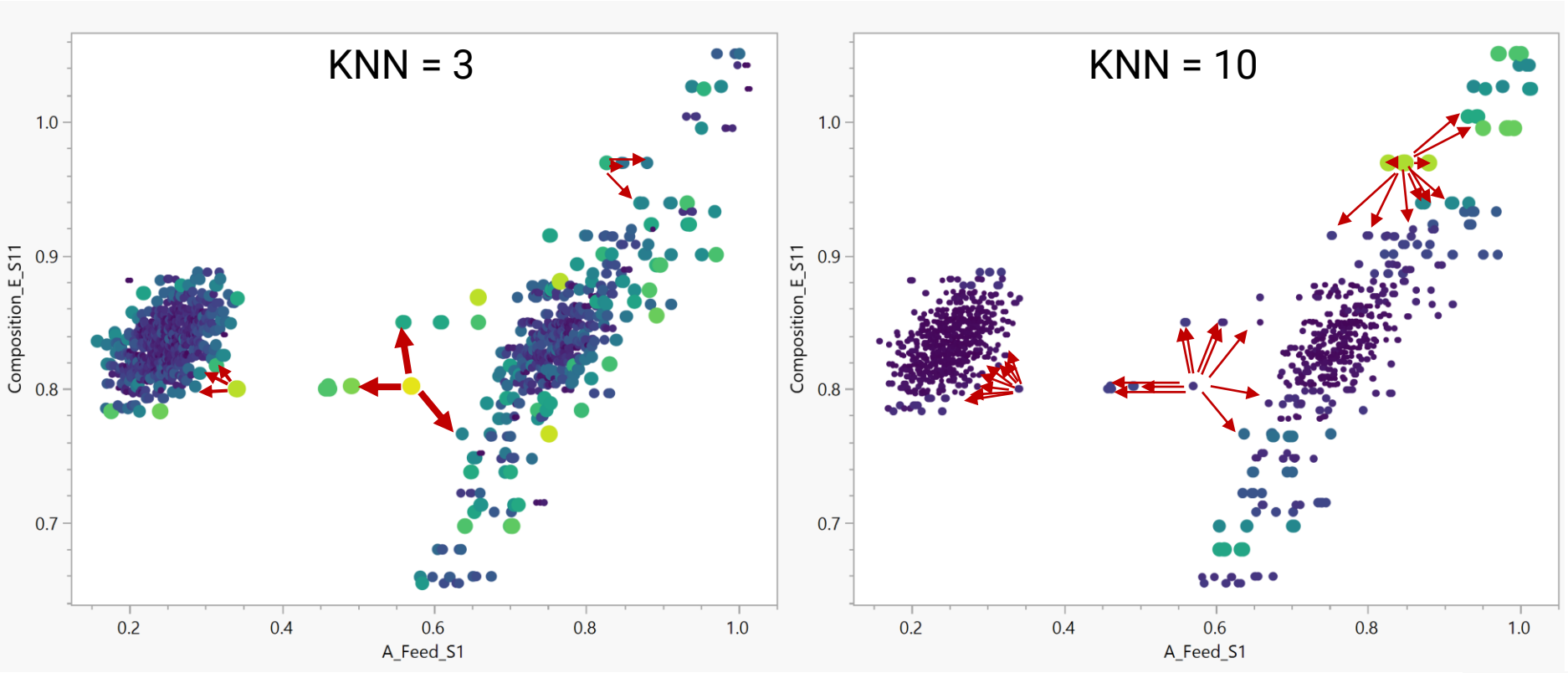
- System dynamics interfere with the analysis, requiring more robust statistics. KNN is another popular algorithm for anomaly detection, in the example above highlighting a process change.

- Steady state regimes can be identified by looking at the stability of main process variables (e.g., flows), or higher point-density areas (colored in red).

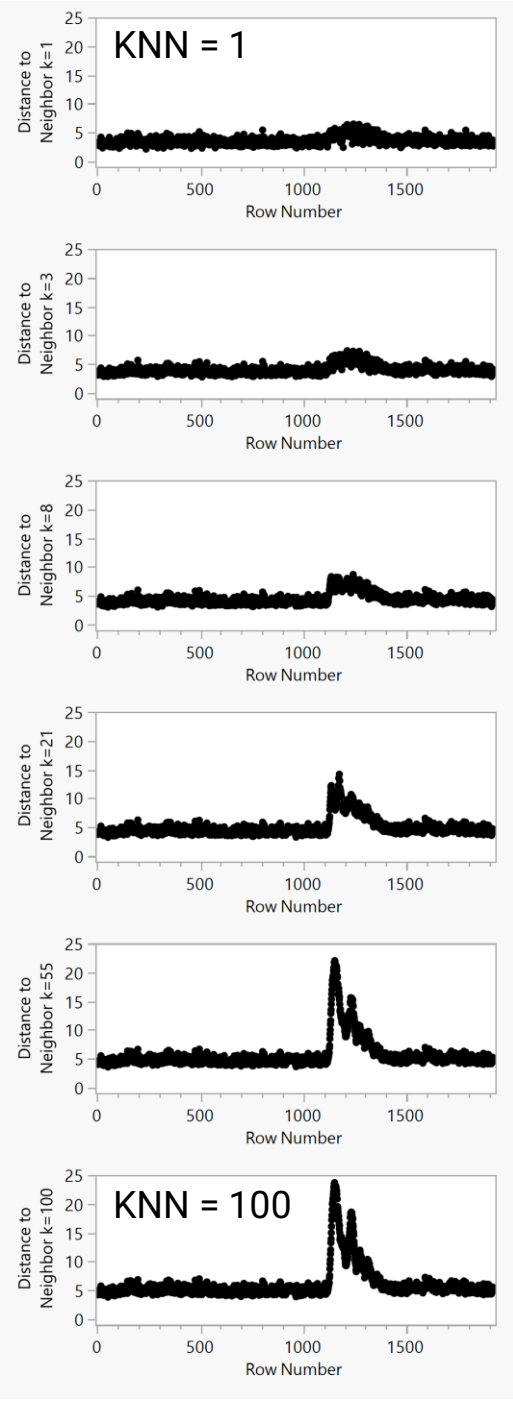
Sargent Centre for Process Systems Engineering | Imperial College London | UCL

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Anomaly detection (K Nearest Neighbors)



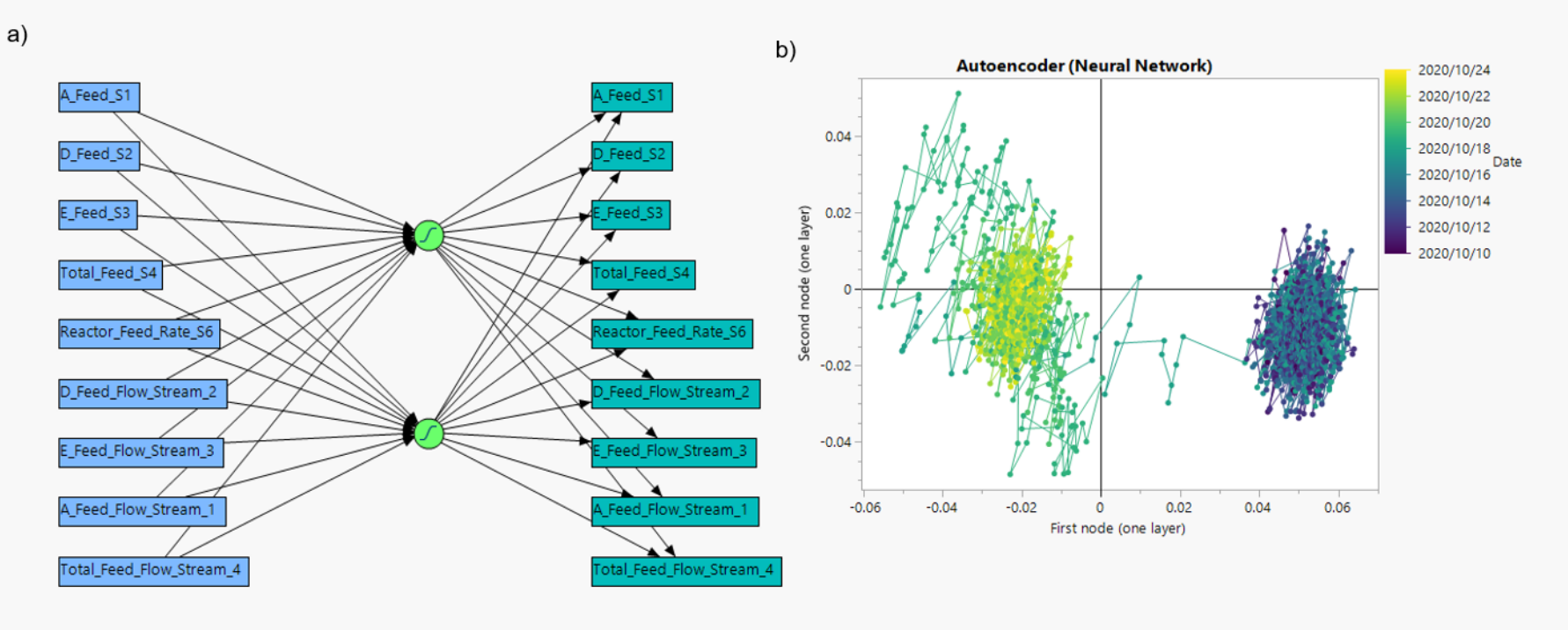
K Nearest Neighbor can identify an outlier based on distance. For each value of k, a Euclidean distance from each point to its kth nearest neighbor is calculated (color). A small value of k (left) can miss identifying points as outliers and a large value of k (right) can falsely classify points as outliers.



Sargent Centre for Process Systems Engineering | Imperial College London | UCL

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

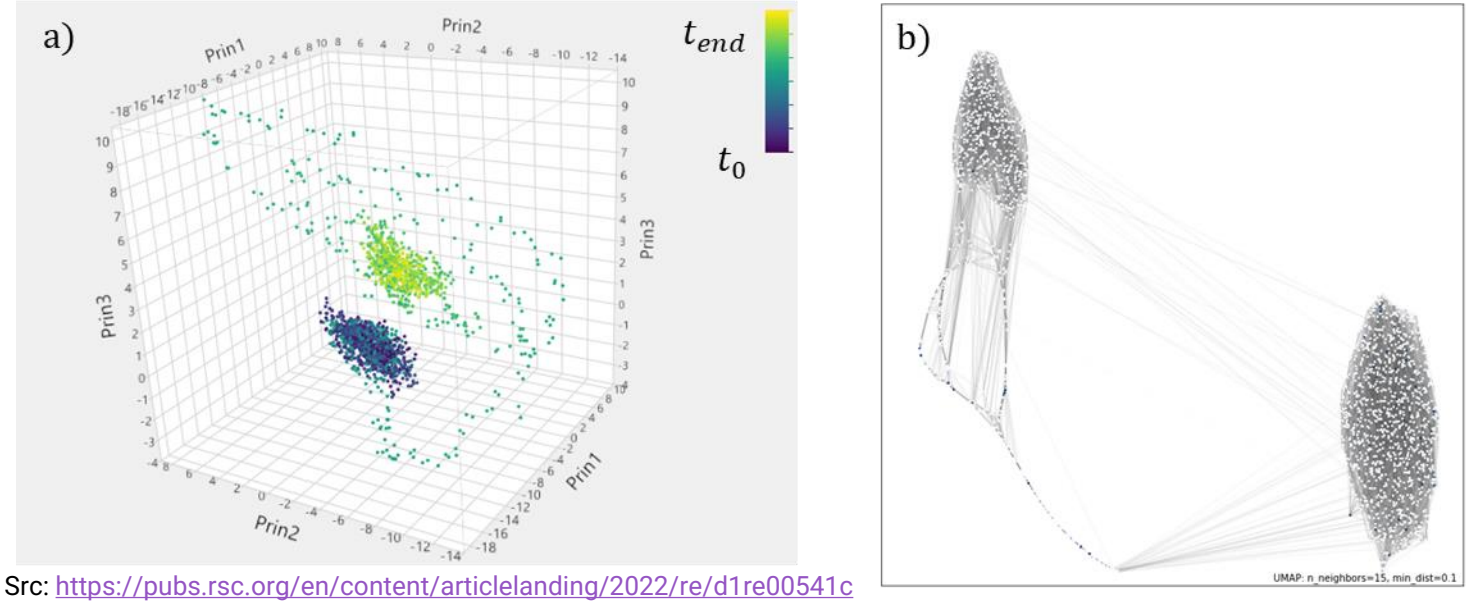
Anomaly detection (autoencoders)



Similar to PCA, autoencoders are neural networks (linear functions with saturation) that can reduce the dimensionality of the data. They achieved this by restricting the number of combinations in the middle layers. This way, correlated data (e.g., redundant measurements) can be reconstructed (a).

Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

Anomaly detection (UMAP)



Src: https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00541c

The transition between two steady-state regimes (Tennessee Eastman Plant) visualized with (a) PCA and (b) UMAP.

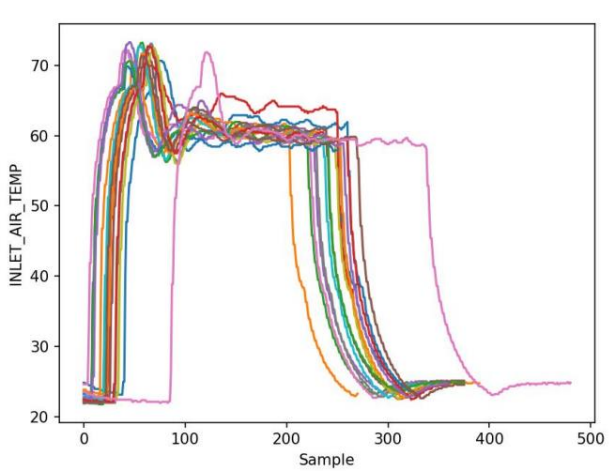
UMAP (Uniform Manifold Approximation and Projection) can outperform and bring new insights when compared with traditional methods. In this example, UMAP demonstrates better separation using only two dimensions, best local and global structure.

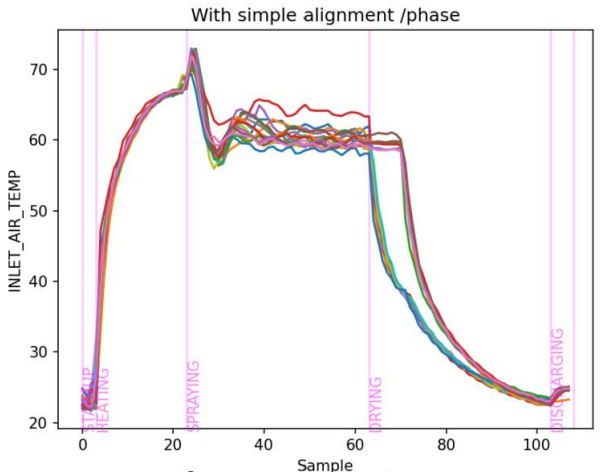
Detailed comparison: Dimensionality reduction for visualizing industrial chemical process data (Joswiak et al. 2019) - Dow

https://www.sciencedirect.com/science/article/abs/pii/S0967066119301728

Python package with applications to industrial batch data

PyPhiBatch – Multivariate Analysis of Batch Processes





Phi toolbox for multivariate analysis by Sal Garcia
https://github.com/salvadorgarciamunoz/pyphi

Years of industrial research made open-source:
Process Analytics Course at Sargent Centre

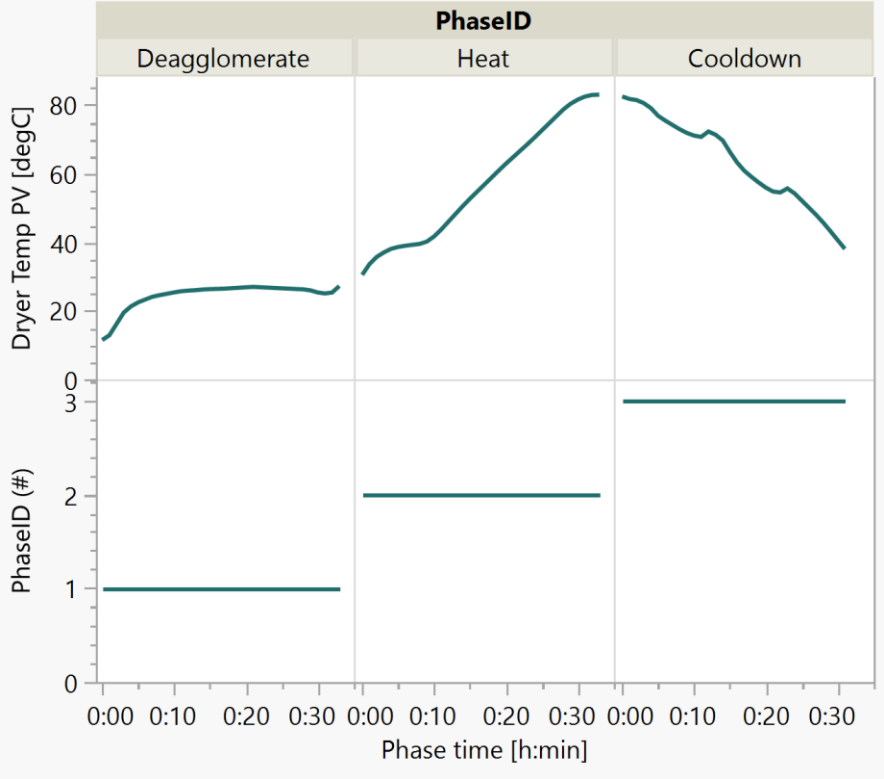
https://www.imperial.ac.uk/process-systems-engineering/courses-and-seminars/workshops-and-courses/process-analytics-course-/

Sargent Centre for Process Systems Engineering

Monitoring and screening process data

Batch process monitoring

Batch data example – Industrial dryer



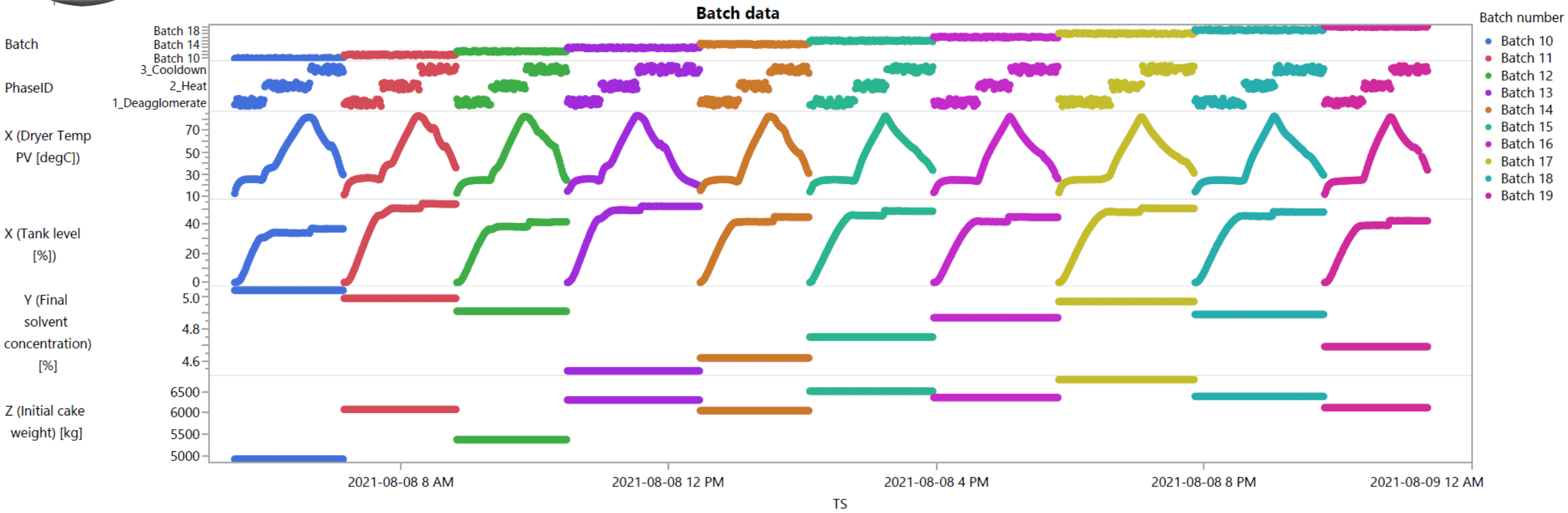
The dryer is charged with a variable amount of wet cake that evaporates and recollects a solvent material in an external tank. There are three distinct phases in the batch dataset:

- **Deagglomeration** phase, where the cake reacts at low agitation speed while the solvent is collected
- **Heating** phase, where the temperature of the cake is increased until its set-point
- **Cooling** phase, where the batch temperature is reduced before unloading

Dataset and original paper:
Troubleshooting of an Industrial Batch Process Using Multivariate Methods, Salvador Garcia et al., 2003
Ind. Eng. Chem. Res. 2003, 42, 15, 3592–3601

Batch data – time series view

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

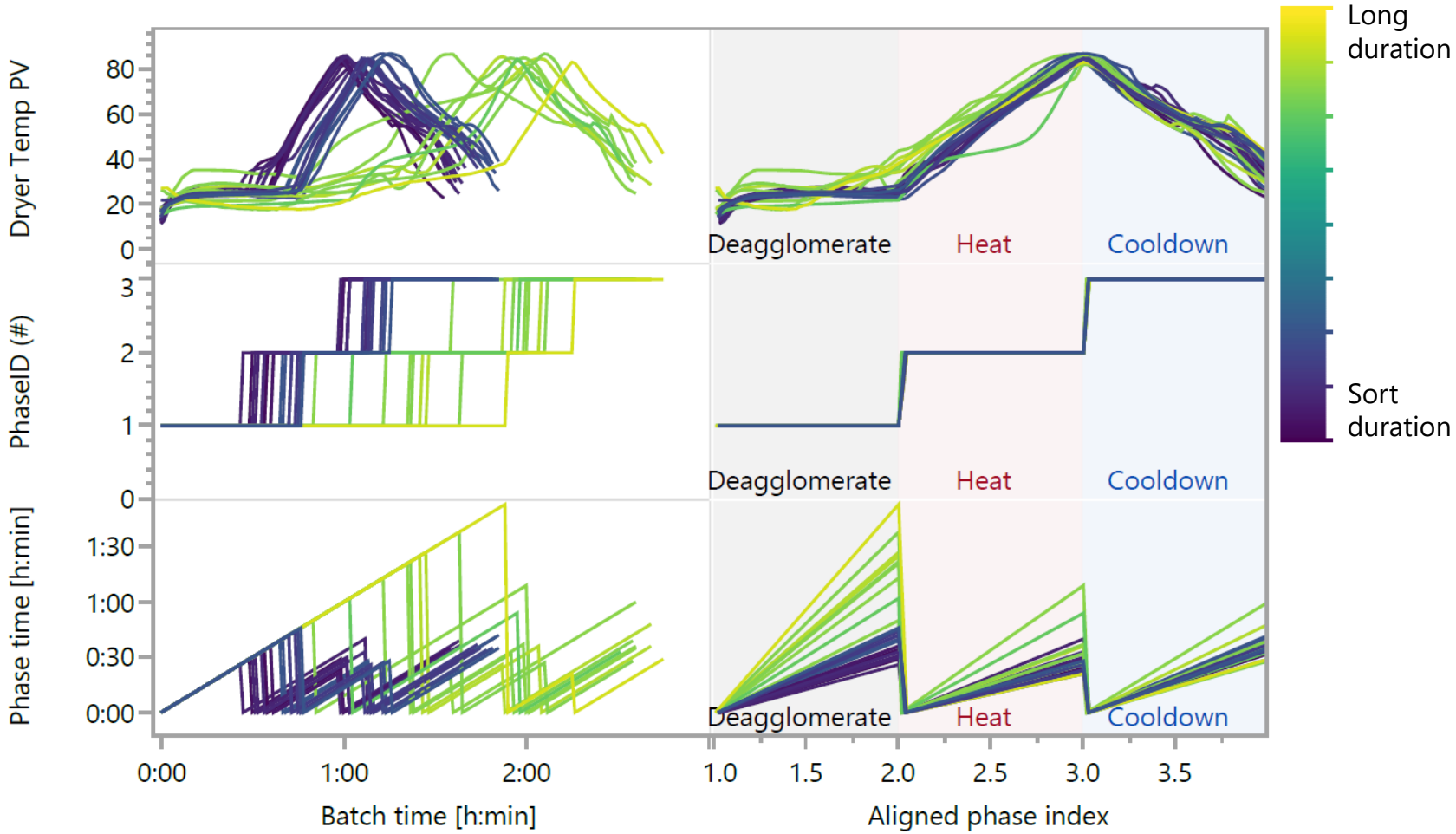


Batch processes have a combination of different data types. Event data which determines in which phase and in what batch the manufacturing process was. Time-varying data coming from sensors such as reactor temperature (c) which evolves through the process showing trends or trajectories (defined as inputs or X variables). Single measured values such as concentration, quantity, or quality for end products (named usually as target or Y variables) and initial conditions (defined as Z variables) that include raw materials properties.

Overlaying and aligning batch data

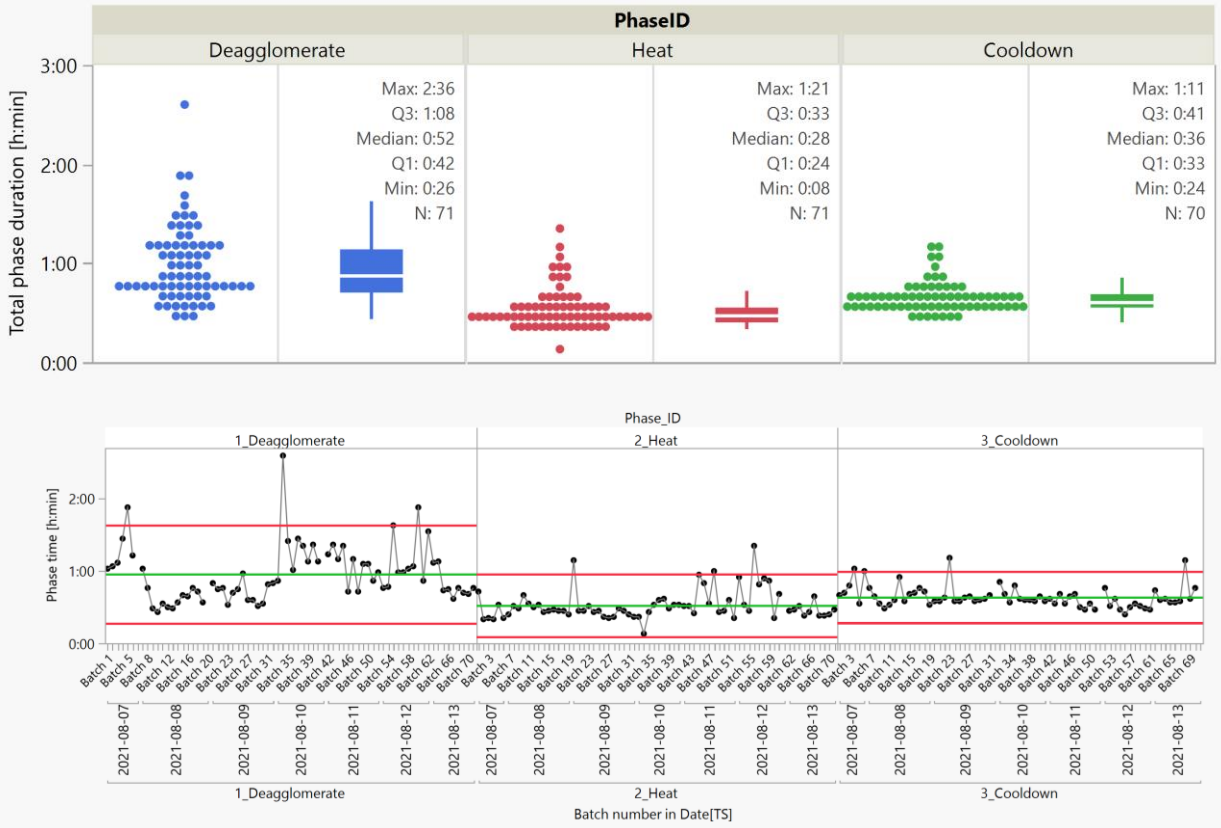
Overlay of the drying temperature profile and for a selection of historized batches colored by total batch duration.

Batches can be aligned using the automation data (phases or stages).



Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

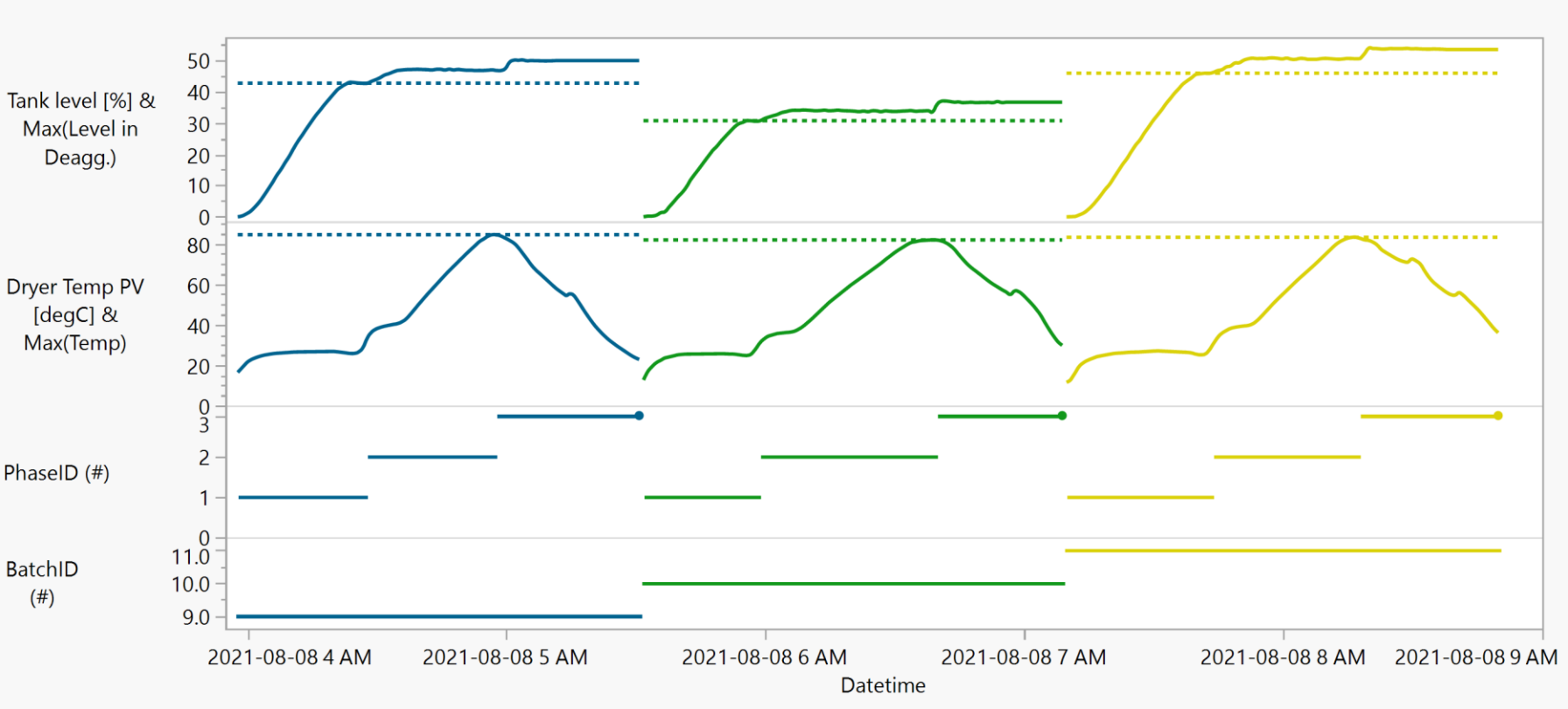
Time information is not lost after alignment



- Batch duration is one of the most important KPIs as it is directly related to productivity. Bar plots, distributions and control charts are often used to detect anomalies.

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

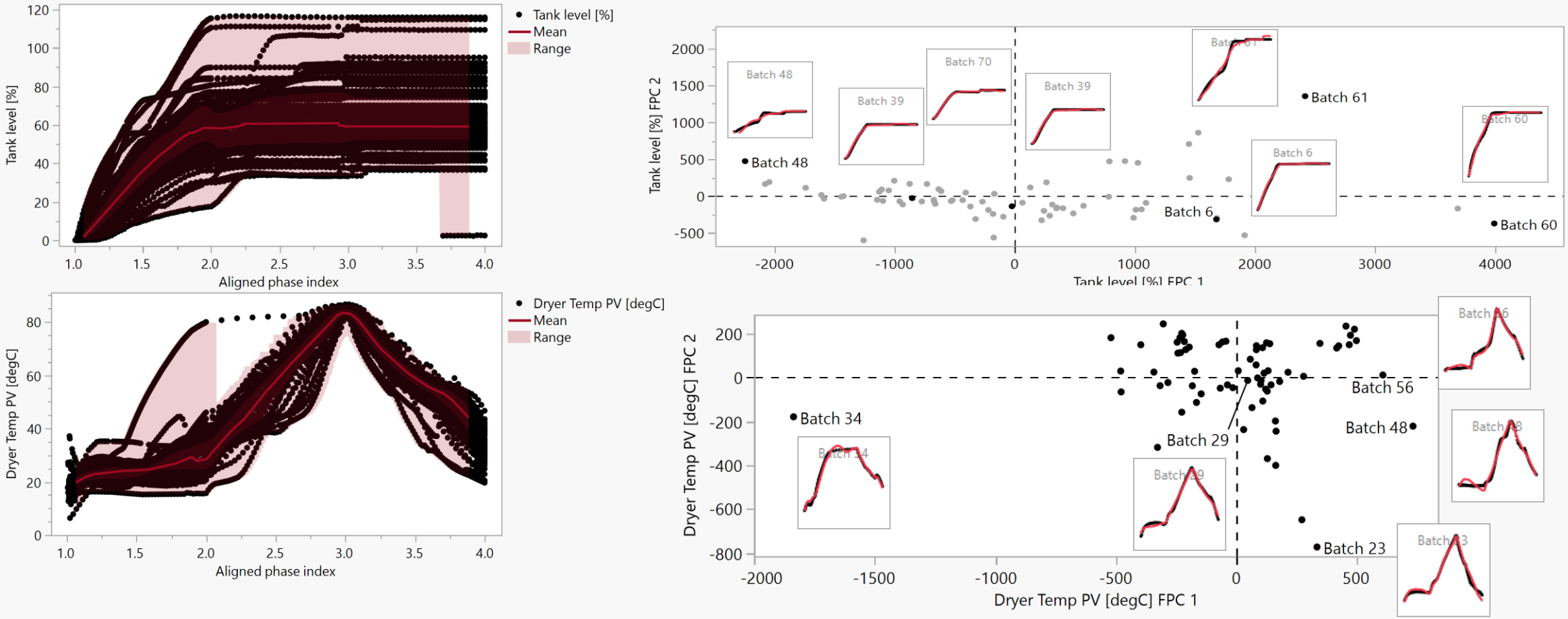
Batch monitoring can be done summarizing information into KPIs



Process and control knowledge is essential to summarize relevant batch information into useful batch statistics. These are often called KPIs, landmarks or fingerprints.

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Functional PCA finds batch variability without defining KPIs

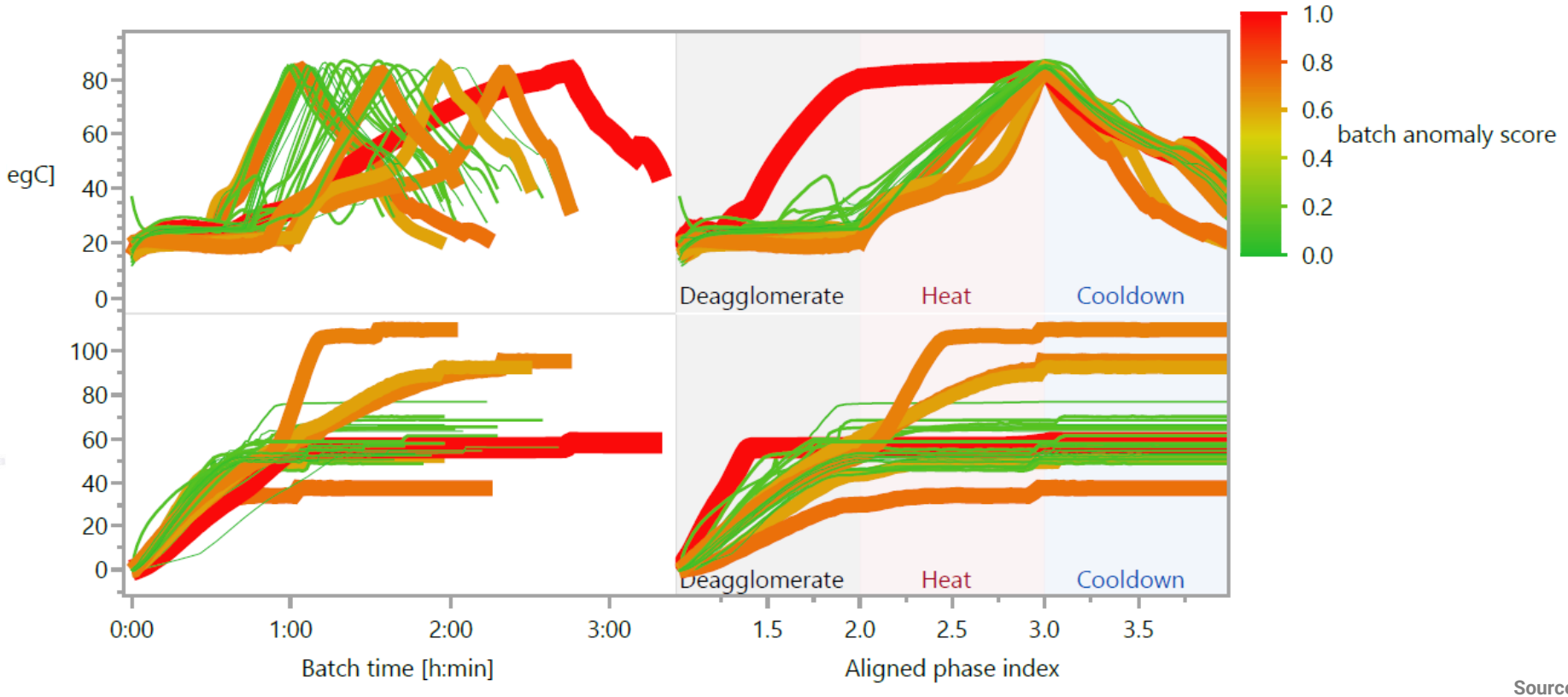


Without the need of asking for specific features, Functional PCA can capture and decompose the whole trajectories seen in batch processes.

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Sargent Centre for Process Systems Engineering
Imperial College London
UCL

Application: Batch monitoring (unsupervised learning)



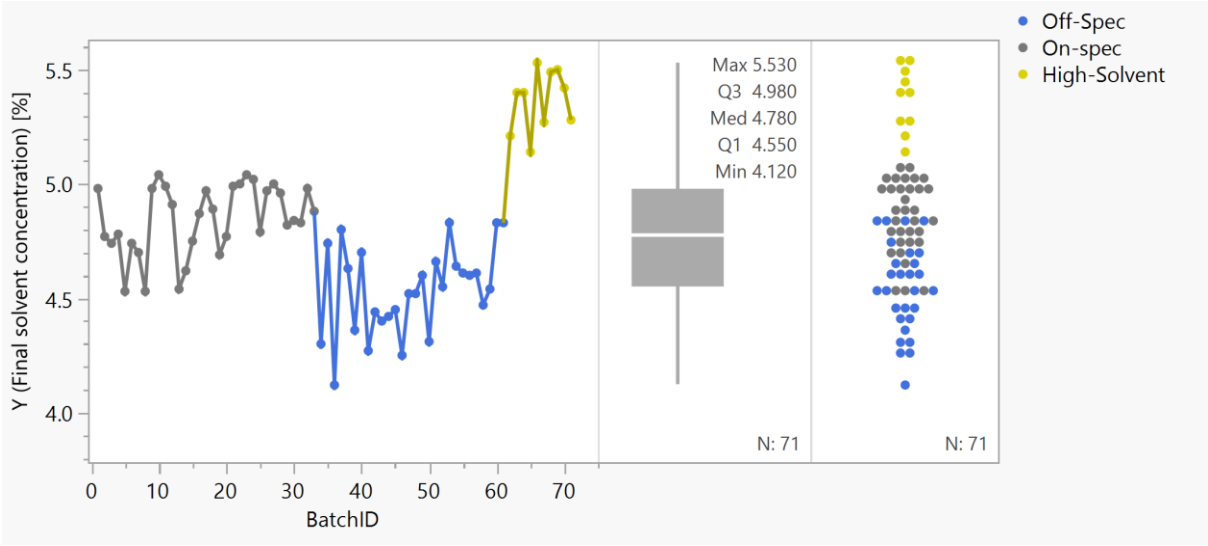
Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Sargent Centre for Process Systems Engineering

Monitoring and screening process data

Batch process screening

Screening of batch processes using process engineering



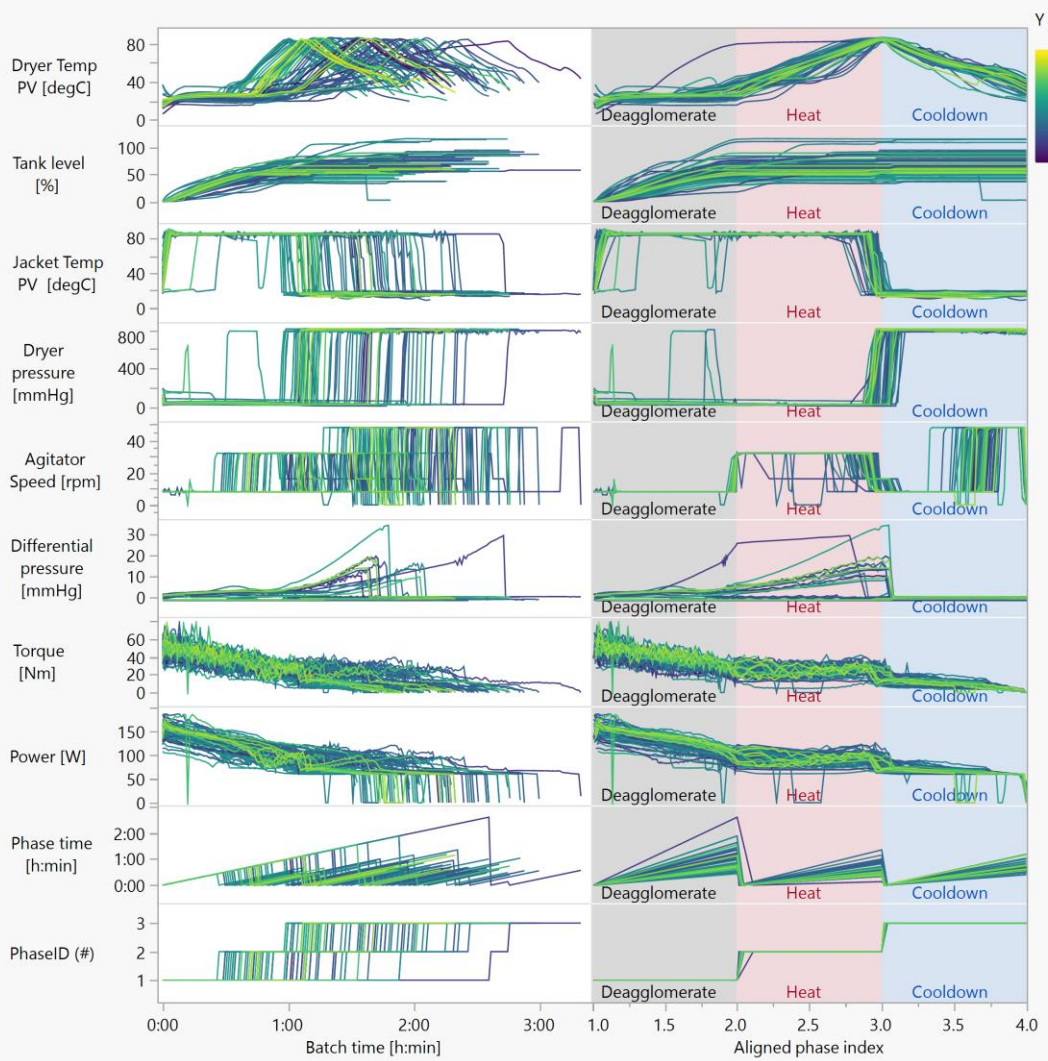
It is tempting to ask engineers for important KPIs that can explain solvent concentration, so a correlation analysis (supervised learning) can be done. Max pressures, temperatures rates, agitation power, jacket temperature, etc.

Do we need to?

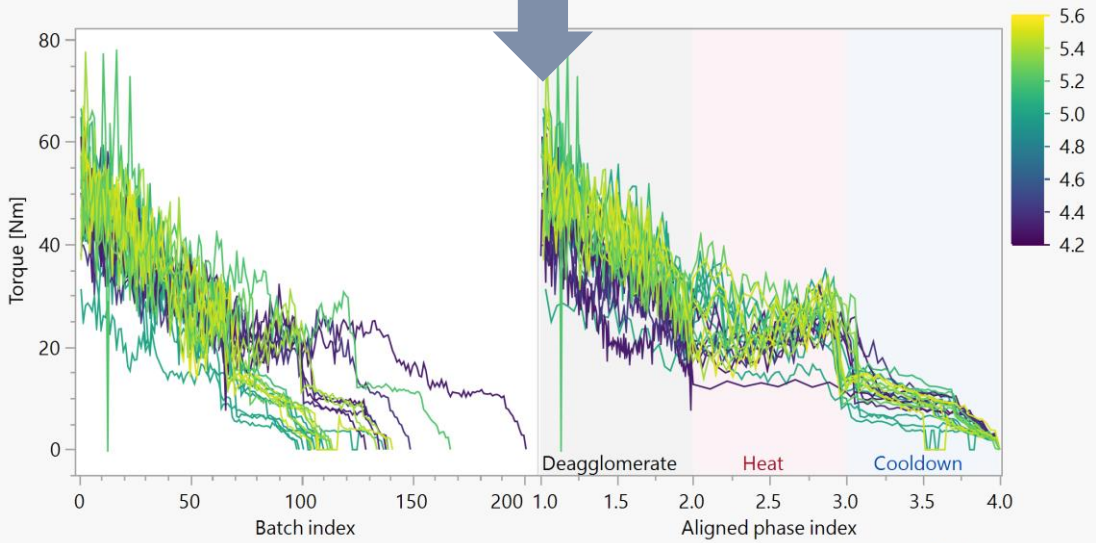
Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Screening of batch processes using feature engineering and variable selection

Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660



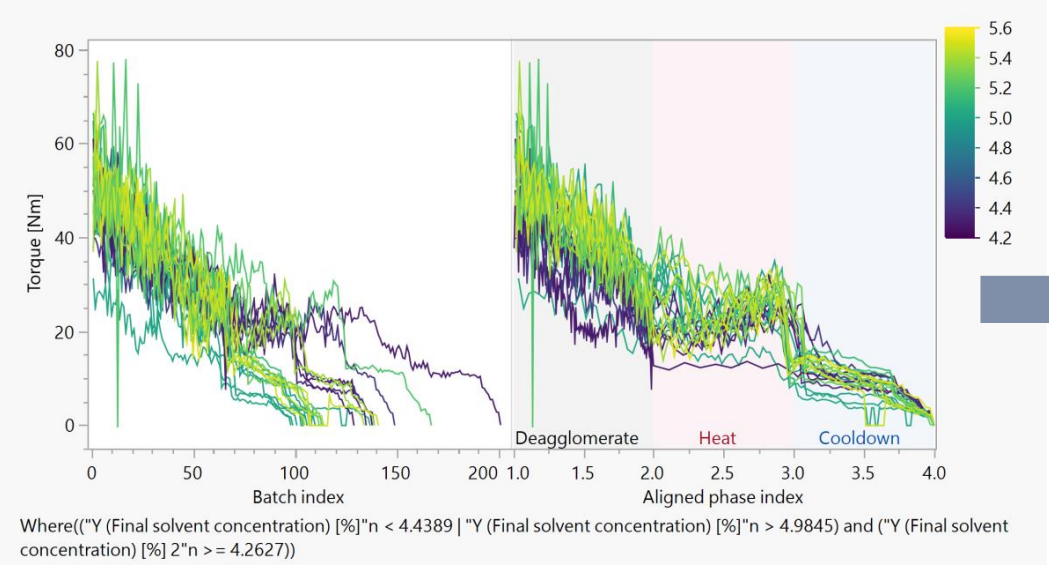
Predictor	Y (Final solvent concentration) [%] Contribution	Portion
CV of Torque [Nm], 2_Heat	0.284945	0.0740
Std Dev of Diff(Torque [Nm]), 1_Deagglomerate	0.141731	0.0368
CV of Diff(Dryer Temp PV [degC]), 1_Deagglomerate	0.115557	0.0300
Std Dev of Torque [Nm], 2_Heat	0.091500	0.0238
Mean of Torque [Nm], 1_Deagglomerate	0.080887	0.0210
Max of abs(Dryer Temp SP-PV) [degC], 1_Deagglomerate	0.080354	0.0209
Mean of Diff(Differential pressure [mmHg]), 2_Heat	0.072405	0.0188
CV of Random Uniform, 2_Heat	0.070162	0.0182

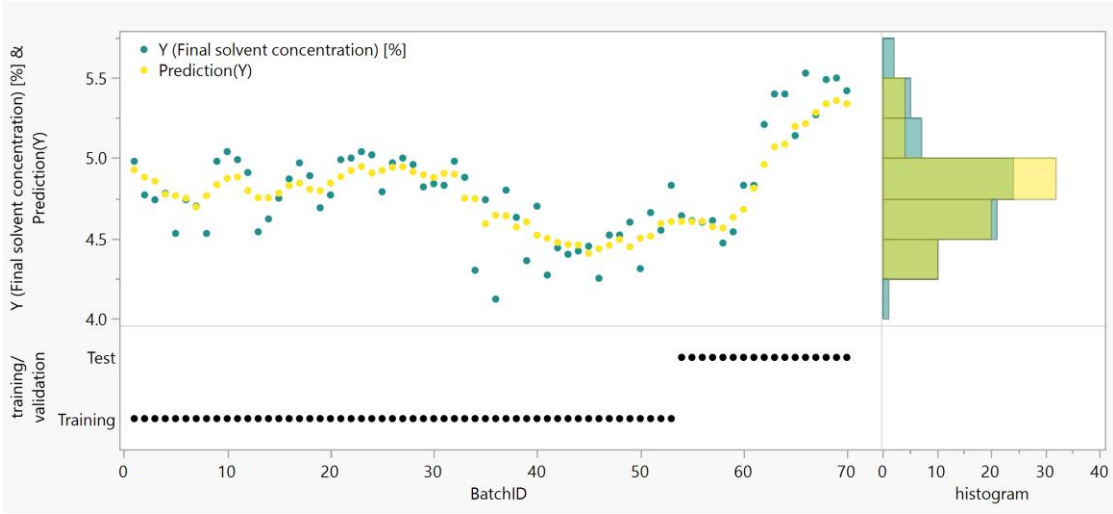


Where(("Y (Final solvent concentration) [%]"n < 4.4389 | "Y (Final solvent concentration) [%]"n > 4.9845) and ("Y (Final solvent concentration) [%] 2"n >= 4.2627))

Application: Viscosity/concentration prediction

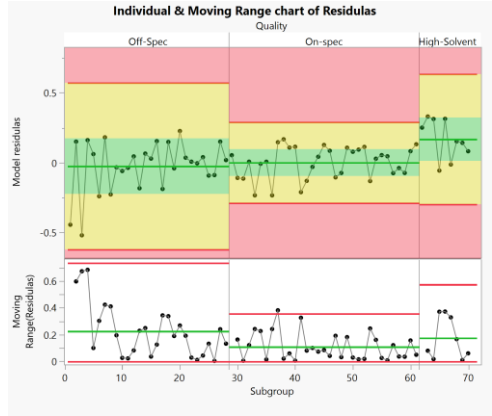
Source:
Industrial data science – a review of machine learning applications for chemical and process industries React. Chem. Eng., 2022, 7, 1471-1509





Agitation power correlates to solvent concentration due to changes of viscosity. In the literature, power consumption from re-circulating or rotary pumps and agitators are commonly used to predict quality measurements before they arrive from the laboratory.

Monitoring of residuals and adaptive models are often required to keep a model accuracy after deployment.



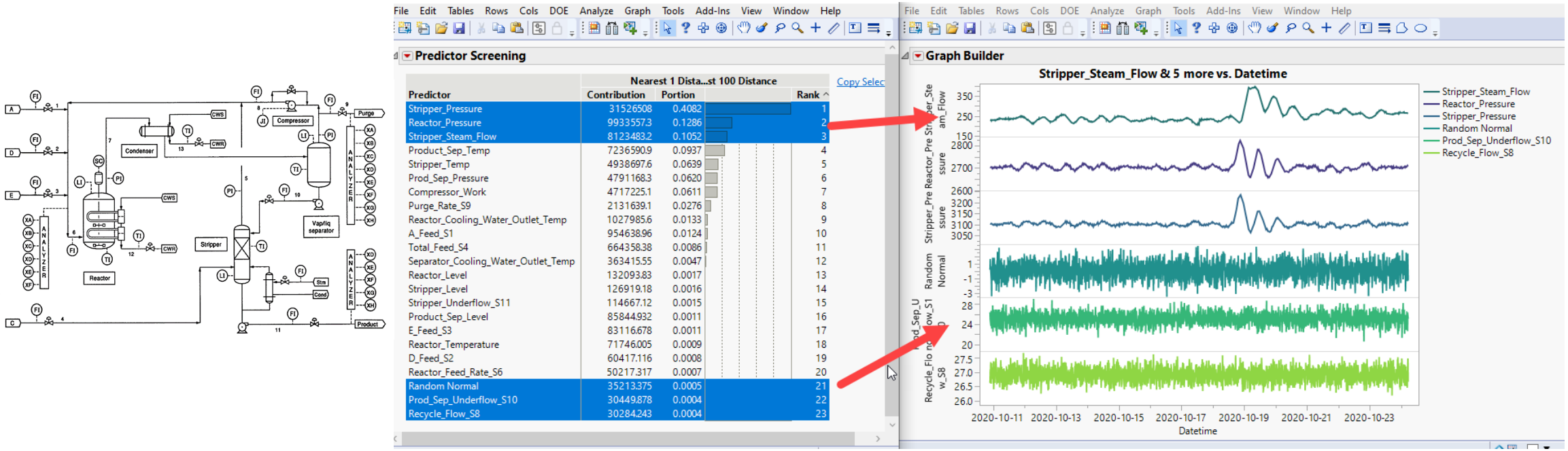
Source:
Industrial Data Science for Batch Manufacturing Processes
arXiv:2209.09660

Sargent Centre for Process Systems Engineering

Monitoring and screening process data

Hands-on

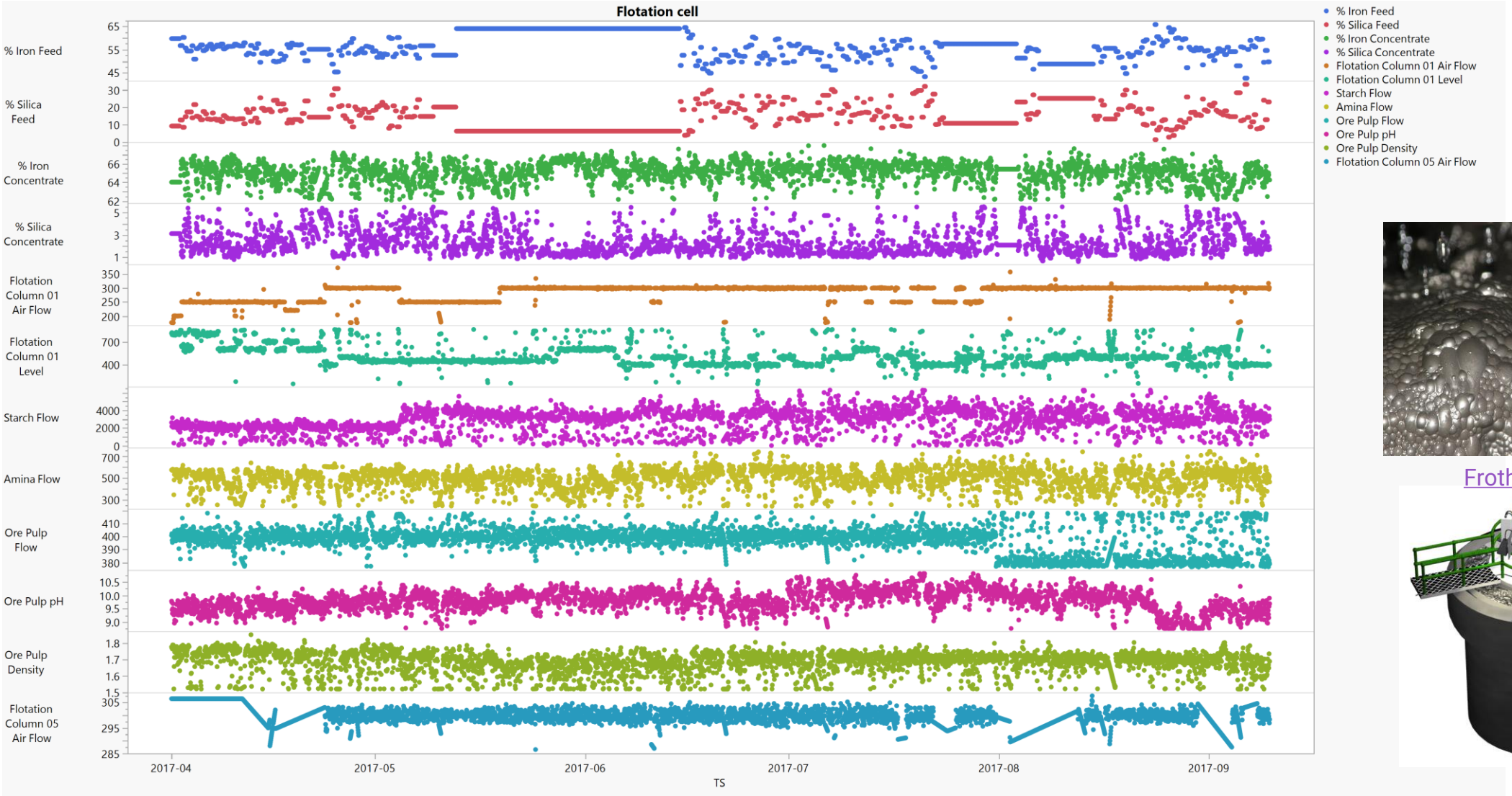
Predictor Explainer: Unsupervised monitoring



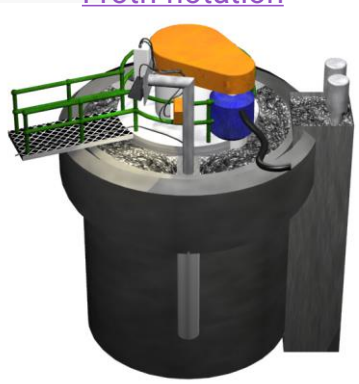
a. Frontend: JMP
b. Backend (optional): Python

https://github.com/industrial-data/predictor-explainer

Process data from mining industry

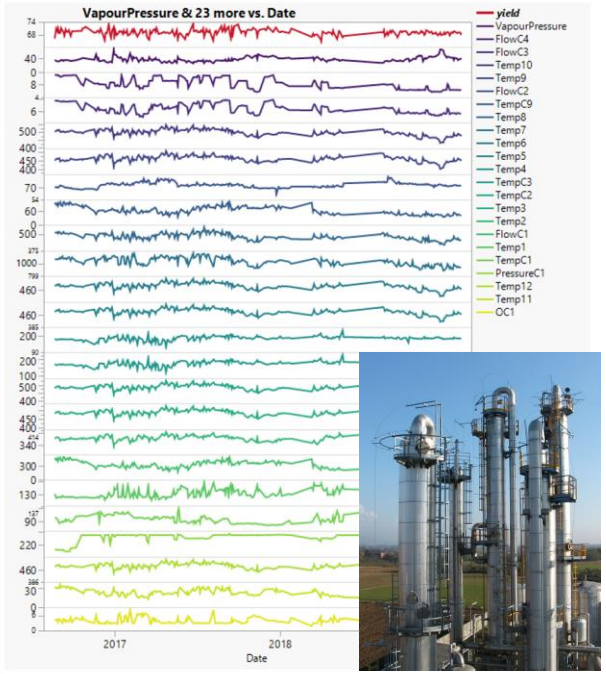


Froth flotation



https://www.kaggle.com/datasets/edumagalhaes/quality-prediction-in-a-mining-process

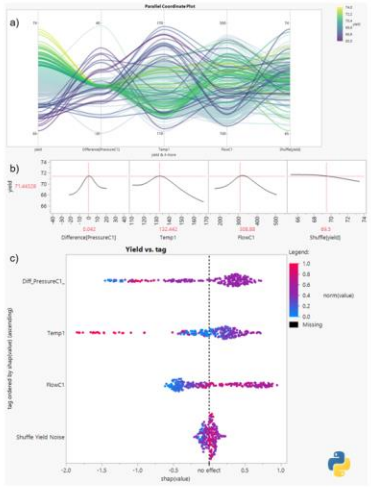
AutoML and ExplainableAI applied to process industries



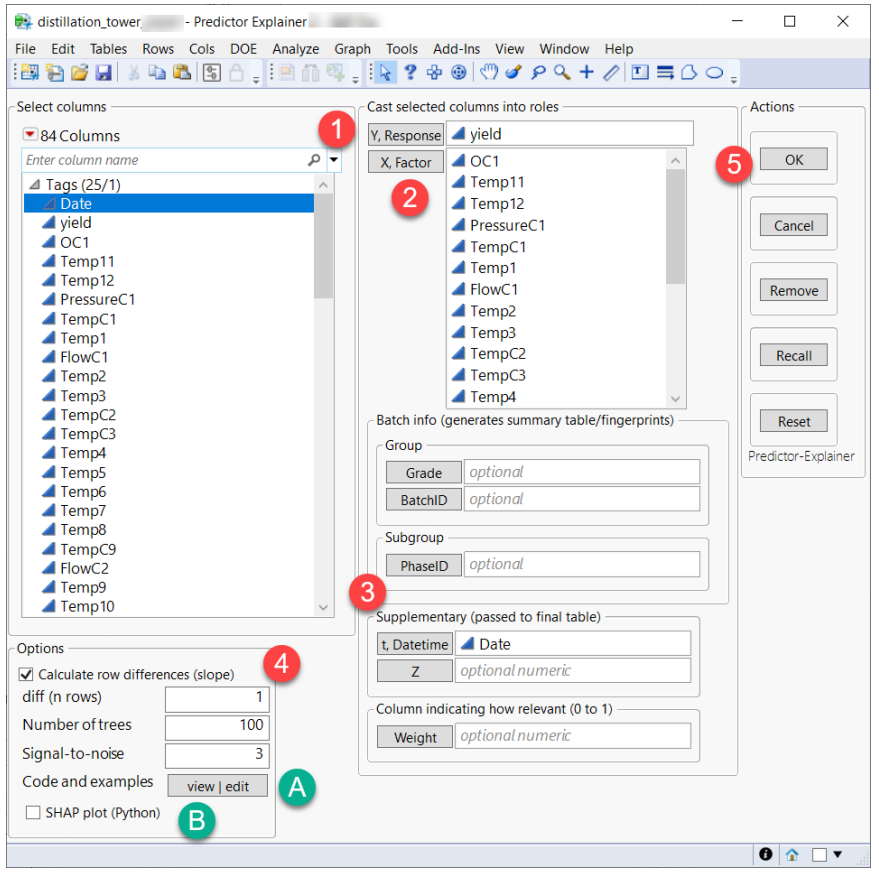
Predictor Explainer

jmp Statistical Discovery™ From SAS

Automated screening and selection of tags (sensors) using interpretable ML.



Ref:



a. Frontend: JMP
b. Backend (optional): Python

https://github.com/industrial-data/predictor-explainer

Similar project in pure Python:
https://explainerdashboard.readthedocs.io/

Summary

- In machine learning, something that doesn't change is irrelevant, both for anomaly and screening models. Example: a controlled reactor temperature won't appear in models as its variation will be minimal while measured disturbances and manipulated variables will.
- System dynamics are always present in process data, several methods exist to remove unstable periods (see BASF example in Day 3)
- Anomaly detection can be used to identify major process challenges
  - Univariate control charts
  - Multivariate control charts using Hotelling's T^2, SPE and KNN distances
- Batch alignment using automation triggers allows simpler batch anomaly detection
- Screening of batch processes can be automated using summary statistics and screening models with synthetic noise for variable selection
- Online estimation of quality parameters are a common application of batch predictive control (e.g., estimating viscosity as a function of agitation power). These are often called inferential sensors or software sensors and require continuous adaptation (re-calibration).

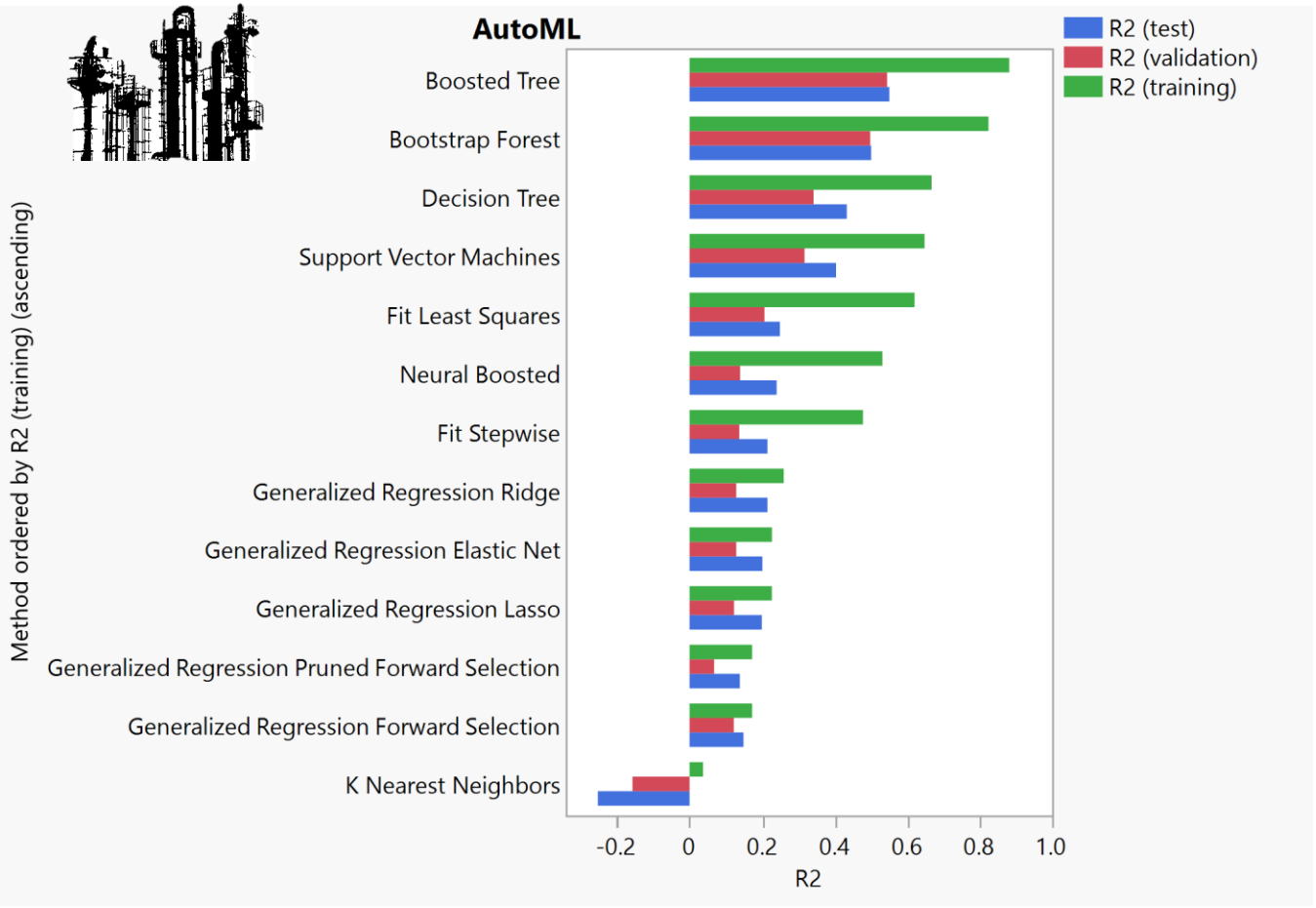
Sargent Centre for Process Systems Engineering

Modeling and understanding

All models are wrong, but some are useful

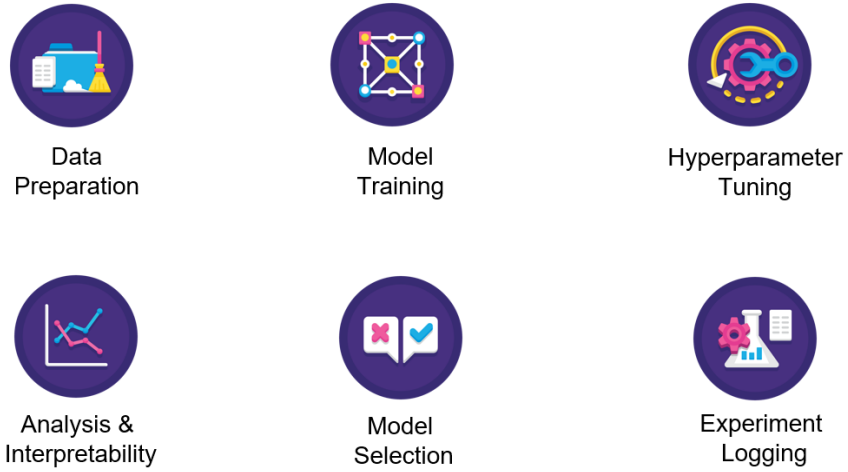
George Box

AutoML, trying all models and tunning parameters at once



JMP Pro Model Screening

https://pycaret.org/




```python
# Time Series Forecasting Functional API Example

# loading sample dataset
from pycaret.datasets import get_data
data = get_data('airline')

# init setup
from pycaret.time_series import *
s = setup(data, fh = 3, session_id = 123)

# model training and selection
best = compare_models()

# plot trained model
plot_model(best)

# predict on hold-out/test set
pred_holdout = predict_model(best)

# predict in unseen future
predictions = predict_model(best, fh = 36)

# save model
save_model(best, 'best_pipeline')
```

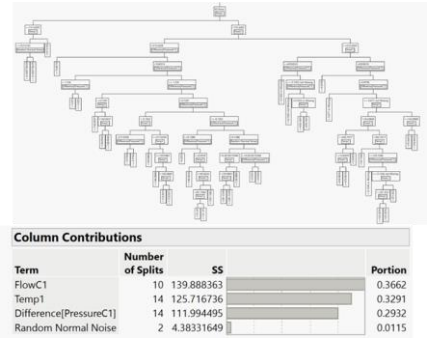
Careful optimizing the wrong problem

Possibly the most common error of a smart engineer is to optimize a thing that should not exist

Elon Musk

Modeling steps in practice

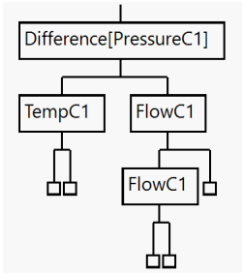
Variable selection

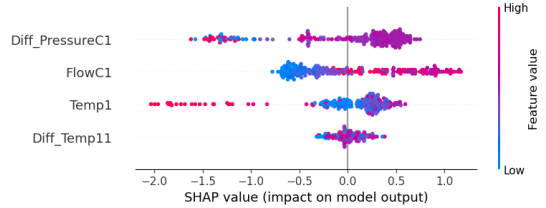


Column Contributions

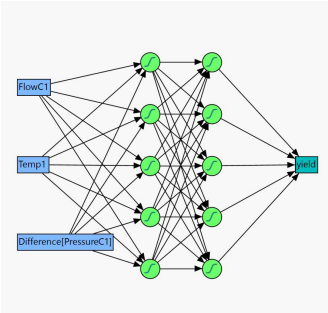
Term	Number of Splits	SS	Portion
FlowC1	10	139.888363	0.3662
Temp1	14	125.716736	0.3291
Difference[PressureC1]	14	111.994495	0.2932
Random Normal Noise	2	4.38331649	0.0115

Understanding



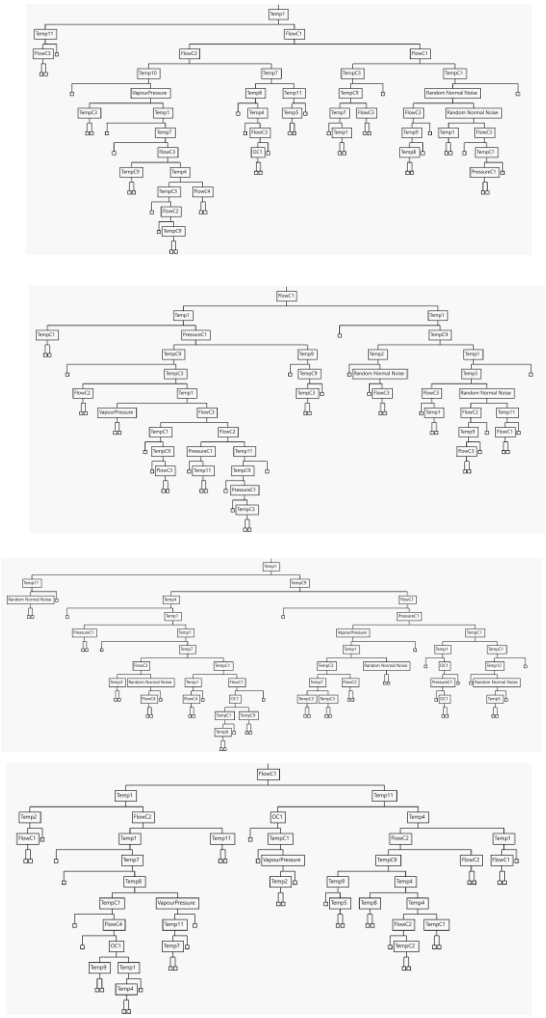


Prediction (if needed)



Variable selection

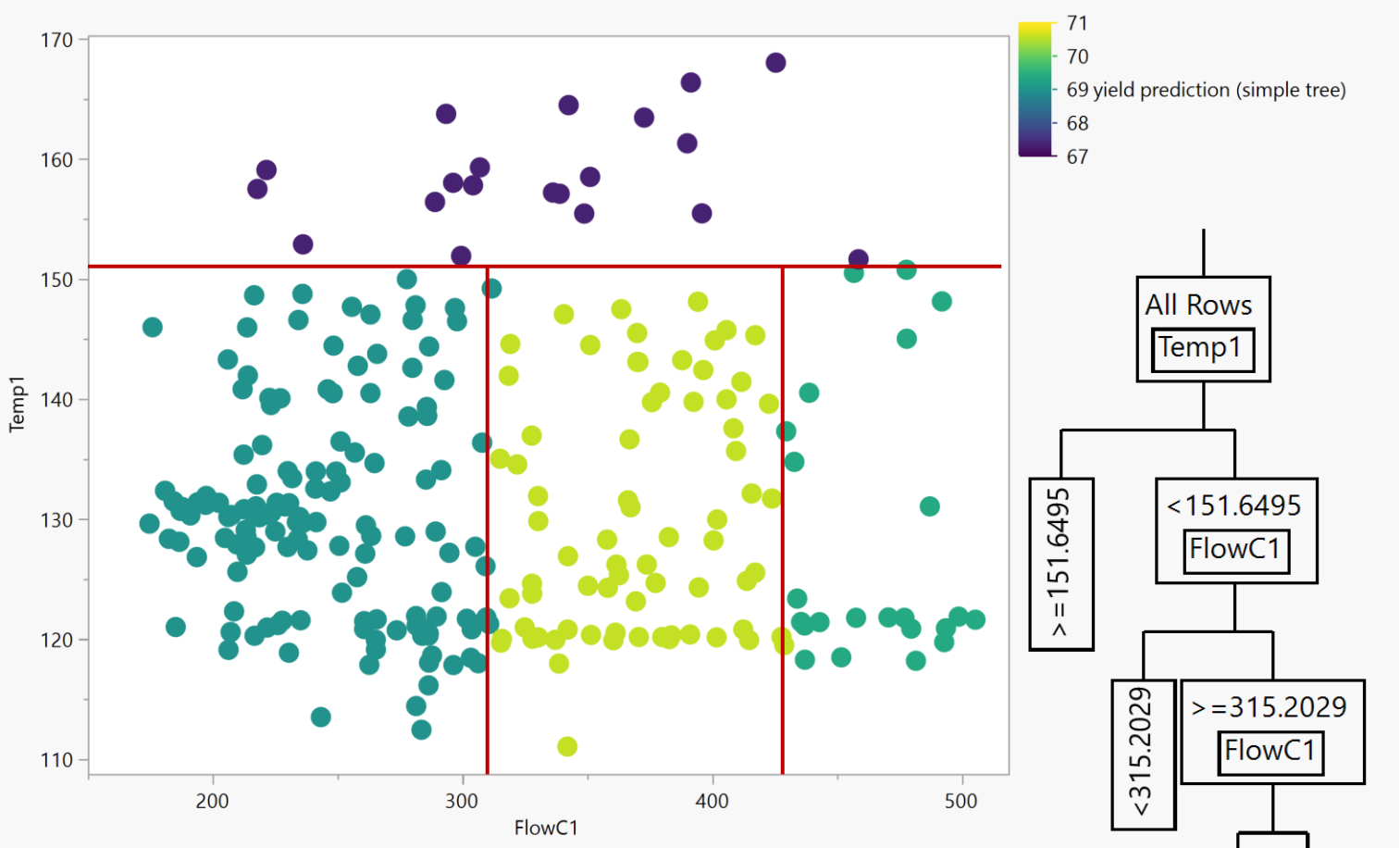
First, a **screening of variables** and selection of tags (sensors) using **tree-based models**. Many tags will end up being weakly correlated to the target, perhaps trying to explain noise in the yield measurement. By adding known noise additional tags, the selection of variables to keep in the model with a certain contribution is facilitated.



Term	Number of Splits	SS	Portion
Temp1	458	70.0036026	0.2916
FlowC1	303	53.8740494	0.2244
TempC9	269	24.6550721	0.1027
TempC3	169	10.5538236	0.0440
Temp4	146	9.47954992	0.0395
PressureC1	283	9.44016133	0.0393
Temp11	180	7.44144012	0.0310
OC1	206	6.19936447	0.0258
TempC1	236	5.81847667	0.0242
Random Normal Noise	238	5.16812352	0.0215
Temp7	166	4.41662431	0.0184
FlowC2	190	4.33130598	0.0180
FlowC3	190	4.32611297	0.0180
FlowC4	145	3.94285078	0.0164
Temp8	145	3.93741242	0.0164
Temp2	127	3.91257534	0.0163
VapourPressure	129	3.33040663	0.0139
Temp9	80	2.24753486	0.0094
Temp10	50	1.38220424	0.0058
Temp5	61	1.31617411	0.0055
TempC2	35	1.17638975	0.0049
Temp12	43	1.12349526	0.0047
Temp3	42	1.02621811	0.0043
Temp6	39	0.99931836	0.0042

Industrial data science – a review of machine learning applications for chemical and process industries

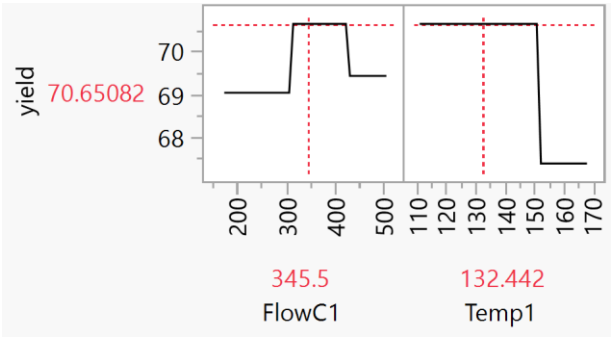
Partition trees, creating expert systems from data



Leaf Label	Mean	Count
Temp1>=151.6495	67.3730501	20
Temp1<151.6495&FlowC1<315.2029	69.0346462	137
Temp1<151.6495&FlowC1>=315.2029&FlowC1>=429.8535	69.4287962	24
Temp1<151.6495&FlowC1>=315.2029&FlowC1<429.8535	70.6508201	72

Tree-based models create simple if-then logics via data partitions that can better explain the target.

They are common for screening and model interpretation, as they can handle tags with different units, the presence of missing values, and outliers while uncovering non-linear relationships.

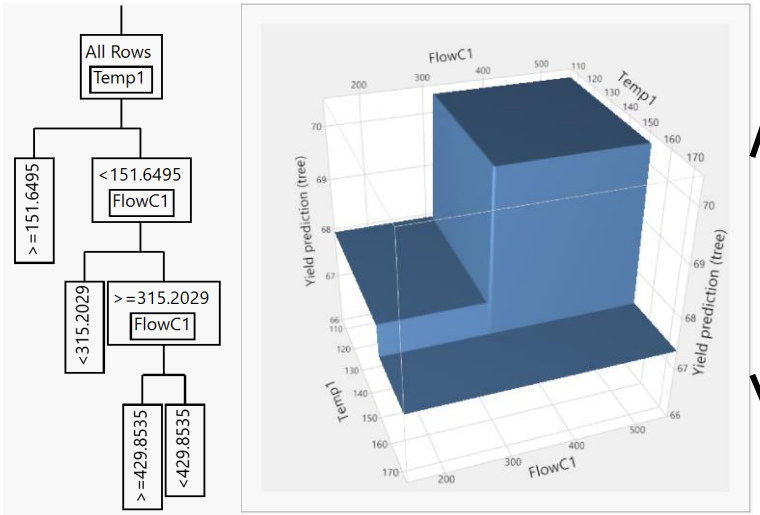


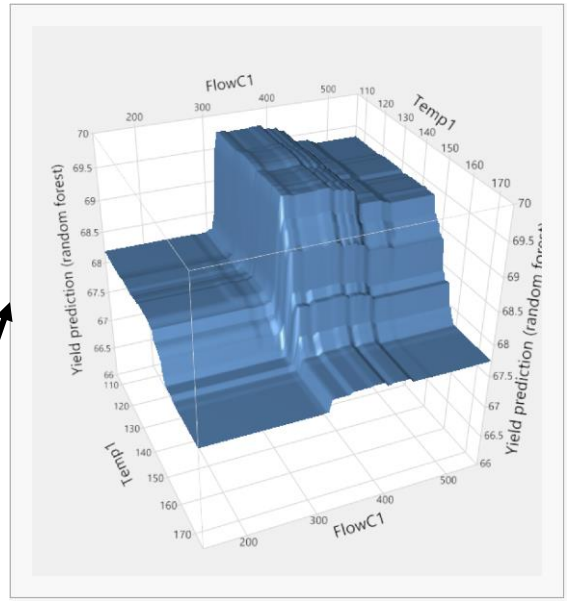
Industrial data science – a review of machine learning applications for chemical and process industries

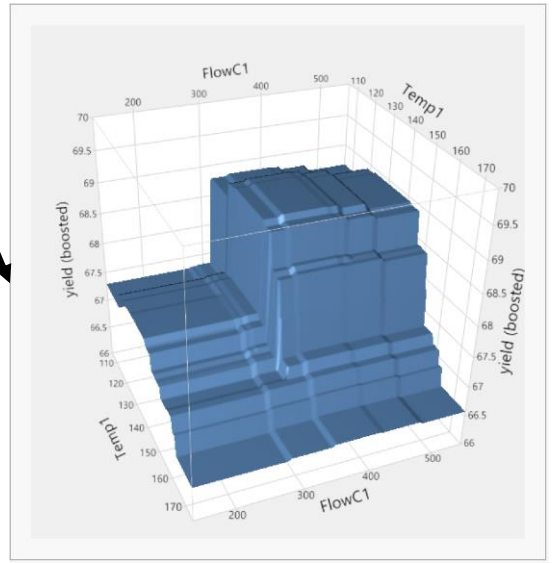
Tree-based models

Bootstrap tree
(Random forest)

Partition tree

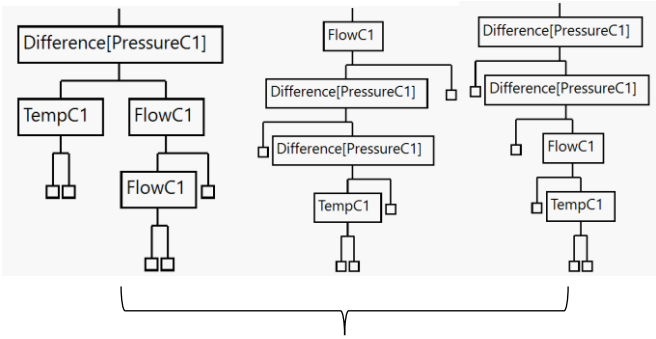






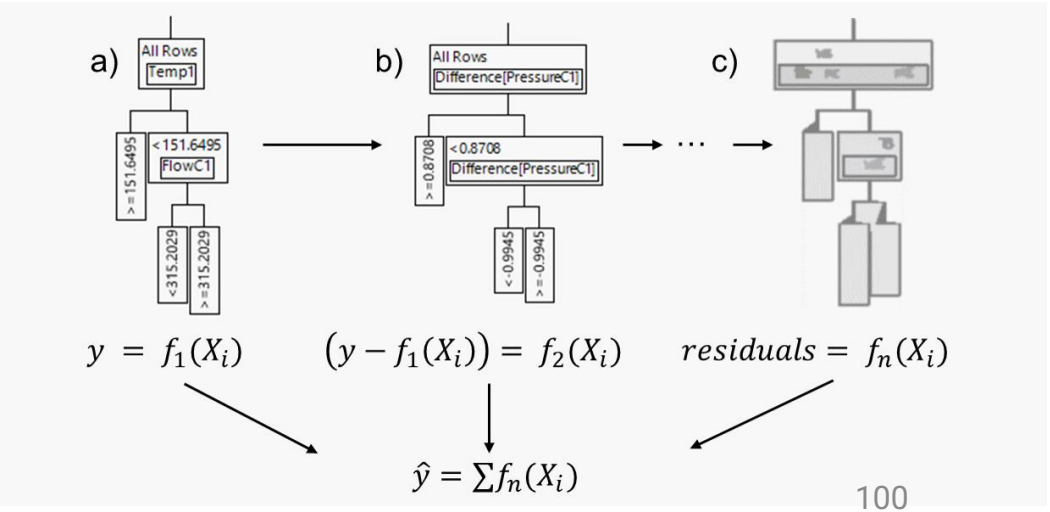
Boosted tree
(LightGBM, CatBoost, XGBoost…)

Average prediction of multiple trees after sampling



$$\hat{y} = \frac{\sum f_n(X_i)}{n}$$

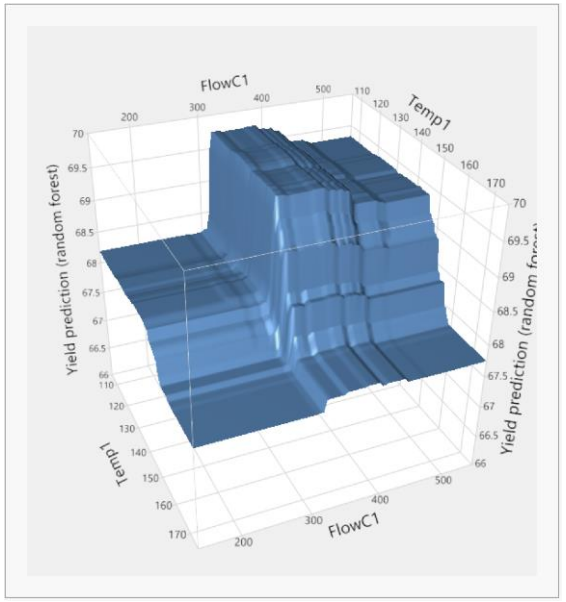
Sum of subsequent predictions



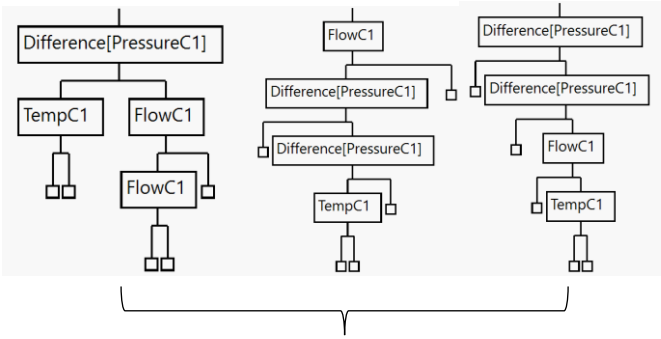
Tree-based models

Bootstrap tree
(Random forest)

- Faster
- Robust, resampling and averages
- **Overfits**



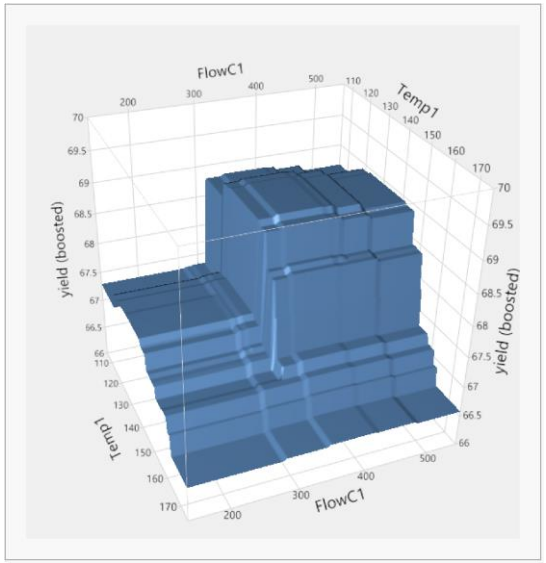
Average prediction of multiple trees after sampling (bagging)



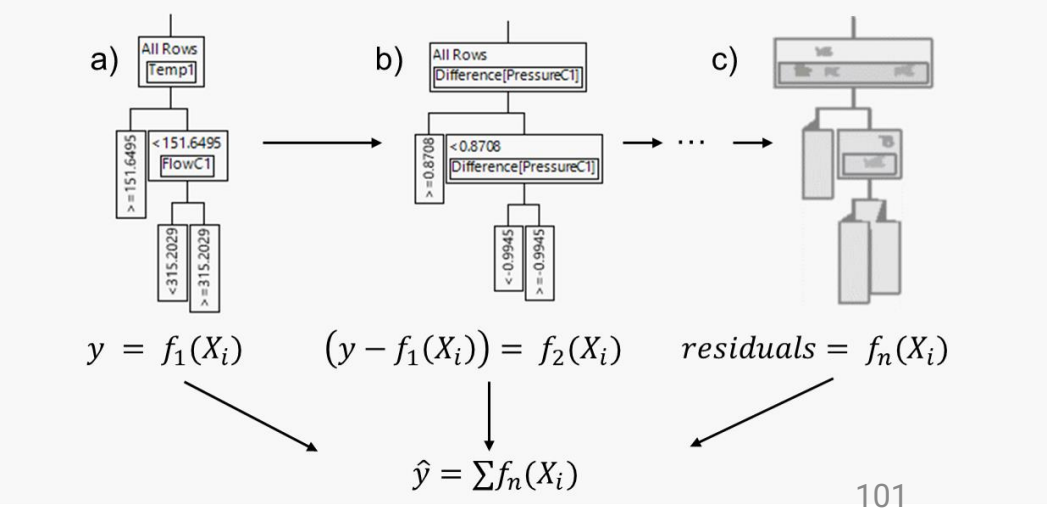
$$\hat{y} = \frac{\sum f_n(X_i)}{n}$$

Boosted tree
(LightGBM, CatBoost, XGboost,)

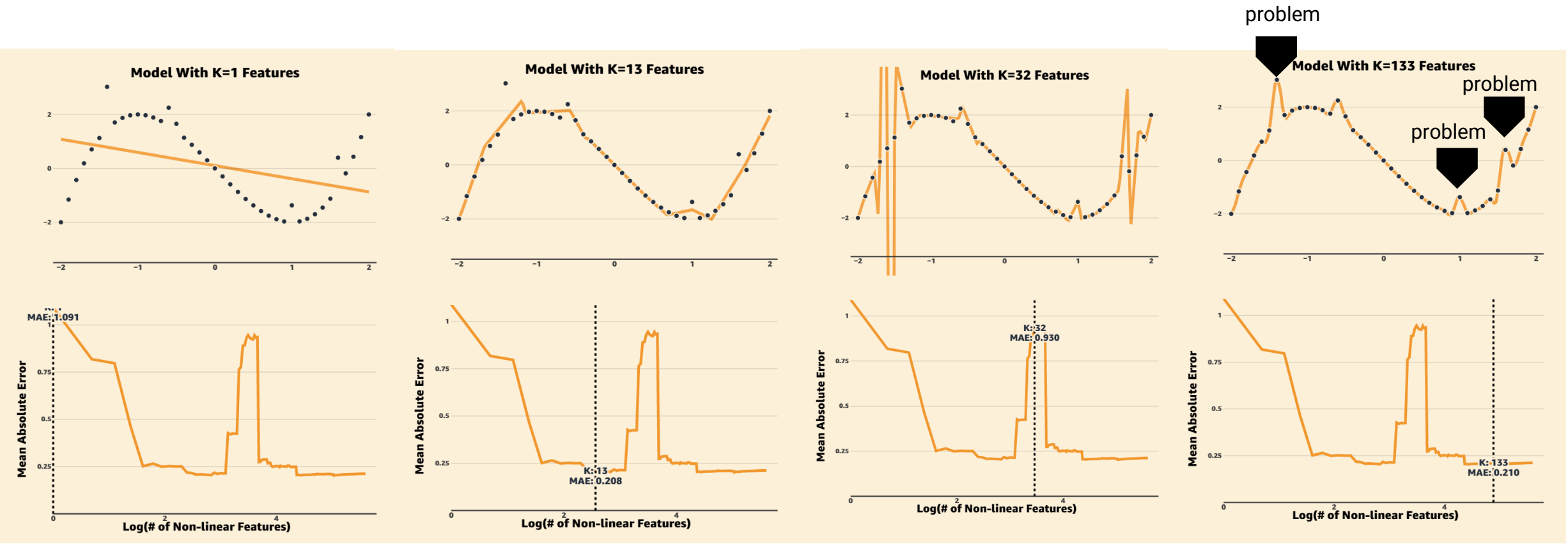
- Fast, specially LightGBM
- Handles known factors (used during the first trees)
- **Overfits**



Sum of subsequent predictions



Overfitting and the double descent

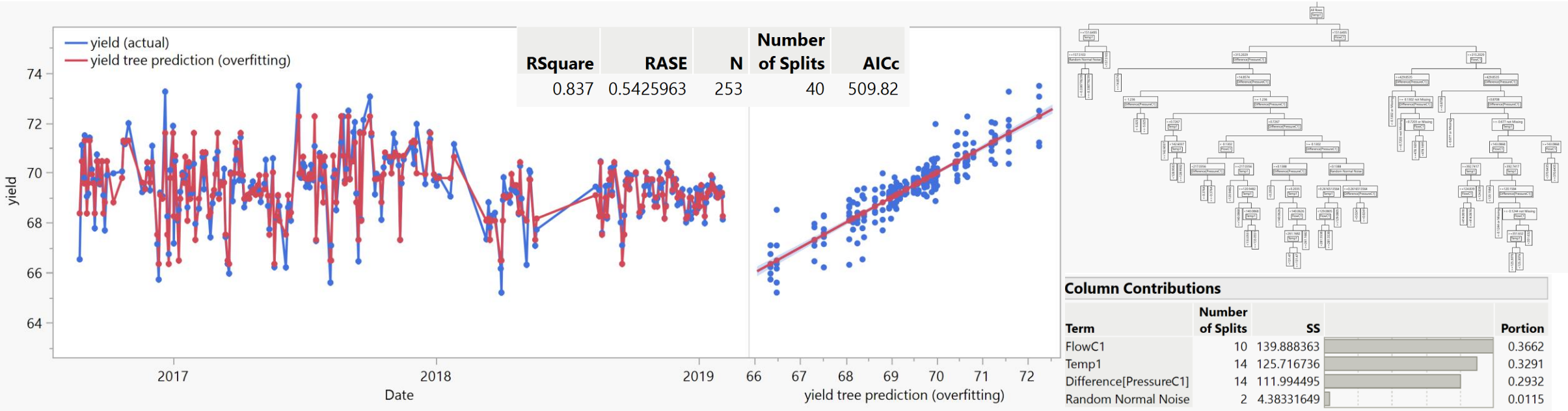


https://mlu-explain.github.io/double-descent/

MLU-EXPLAIN

Visual explanations of core machine learning concepts

Overfitting is a problem if model captures noise



RSquare	RASE	N	Number of Splits	AICc
0.837	0.5425963	253	40	509.82

Column Contributions

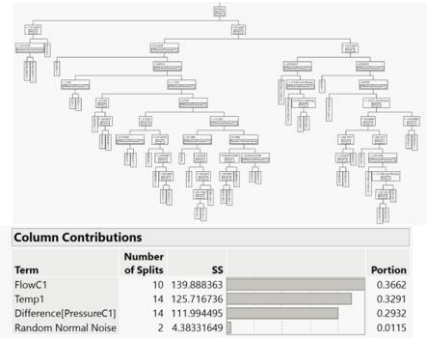
Term	Number of Splits	SS	Portion
FlowC1	10	139.888363	0.3662
Temp1	14	125.716736	0.3291
Difference[PressureC1]	14	111.994495	0.2932
Random Normal Noise	2	4.38331649	0.0115

- Data-driven models can adjust perfectly to data given enough parameters, even to noise

$$data = model + noise$$

Modeling steps in practice

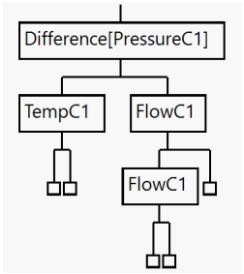
Variable selection

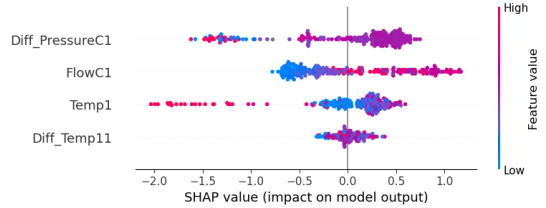


Column Contributions

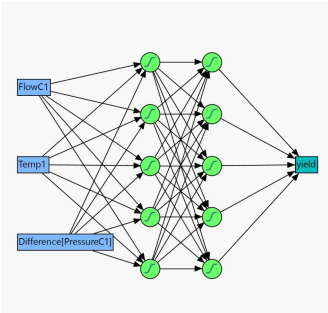
Term	Number of Splits	SS	Portion
FlowC1	10	139.888363	0.3662
Temp1	14	125.716736	0.3291
Difference[PressureC1]	14	111.994495	0.2932
Random Normal Noise	2	4.38331649	0.0115

Understanding

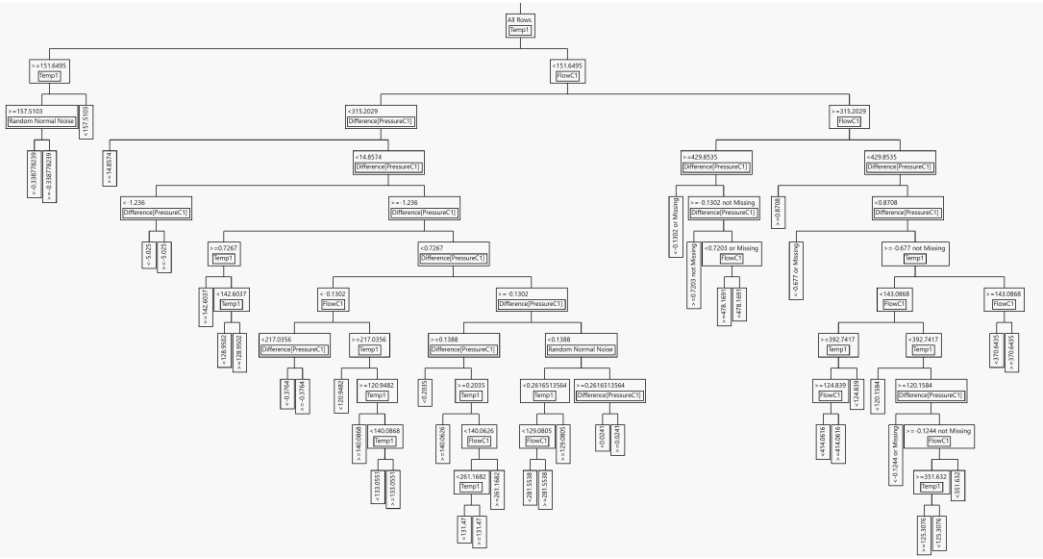


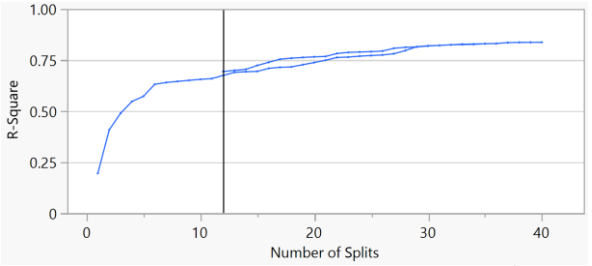


Prediction (if needed)

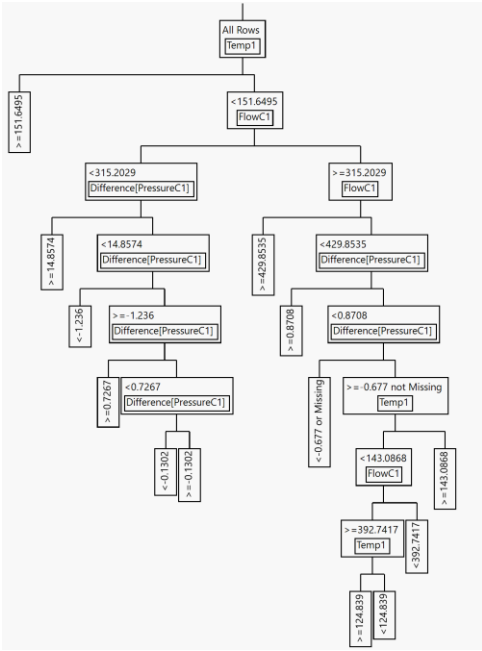


Tree will be simpler to interpret and implement





Simpler model
(less number of splits)

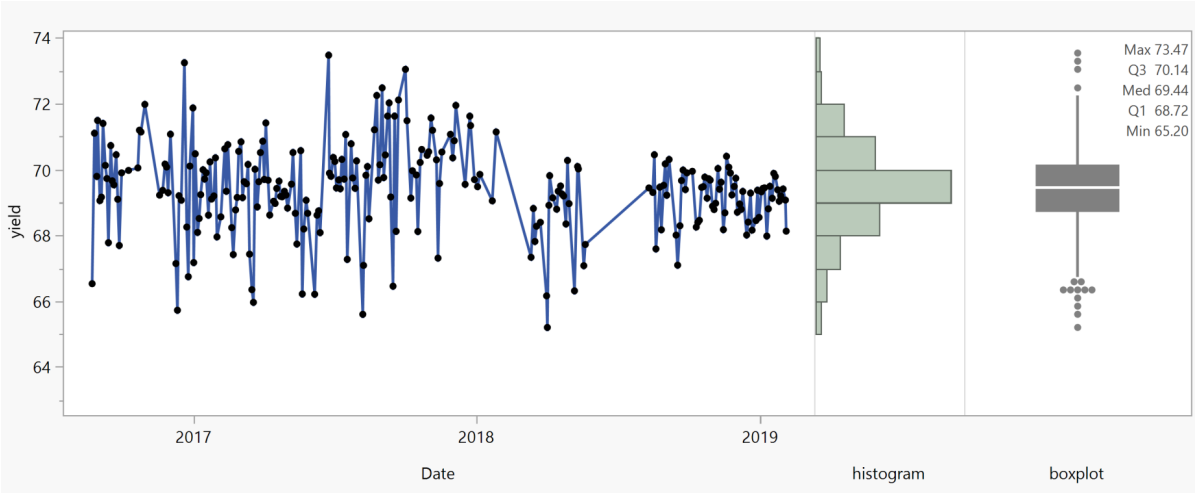


- Adding synthetic noise, a model can be used to describe the main effect of variables aside from common variation.

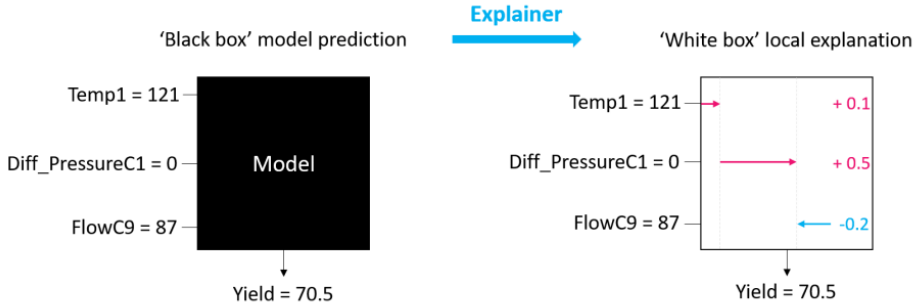
Term	Number of Splits	SS	Portion
FlowC1	3	126.590026	0.3997
Temp1	3	101.429523	0.3202
Difference[PressureC1]	6	88.7241233	0.2801
Random Normal Noise	0	0	0.0000

$$\boldsymbol{data} + \cancel{\boldsymbol{known\ noise}} = \boldsymbol{model} + \cancel{\boldsymbol{noise}}$$

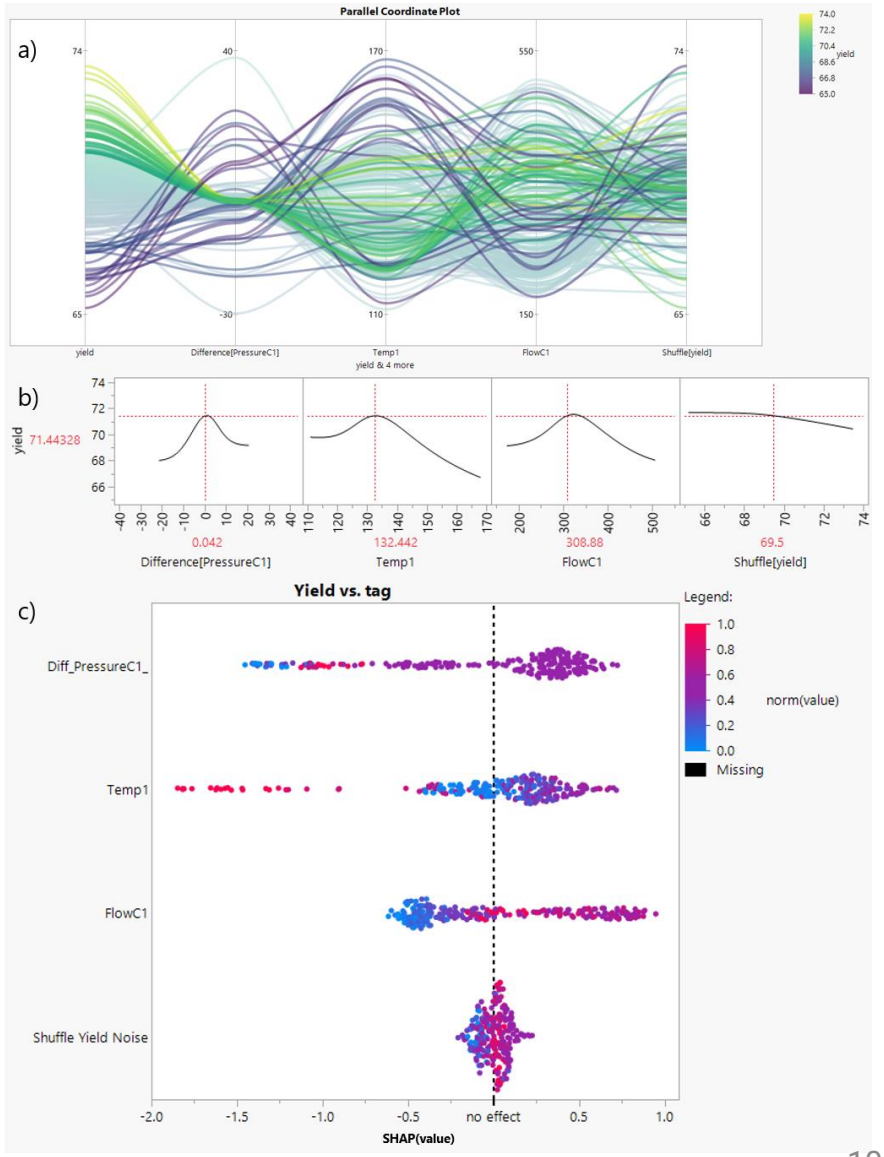
Understanding black boxes with SHAP (ExplainableAI)



Problem definition

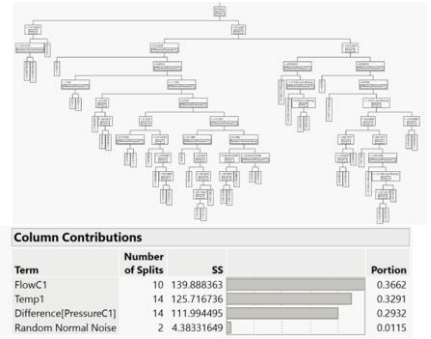


Problem understanding



Modeling steps in practice

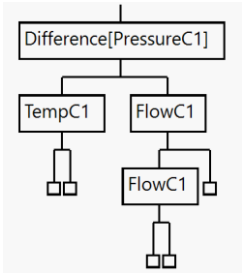
Variable selection

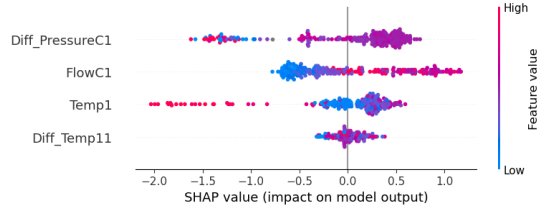


Column Contributions

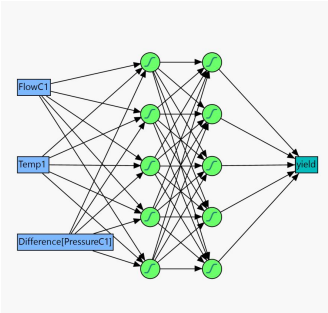
Term	Number of Splits	SS	Portion
FlowC1	10	139.888363	0.3662
Temp1	14	125.716736	0.3291
Difference[PressureC1]	14	111.994495	0.2932
Random Normal Noise	2	4.38331649	0.0115

Understanding





Prediction (if needed)



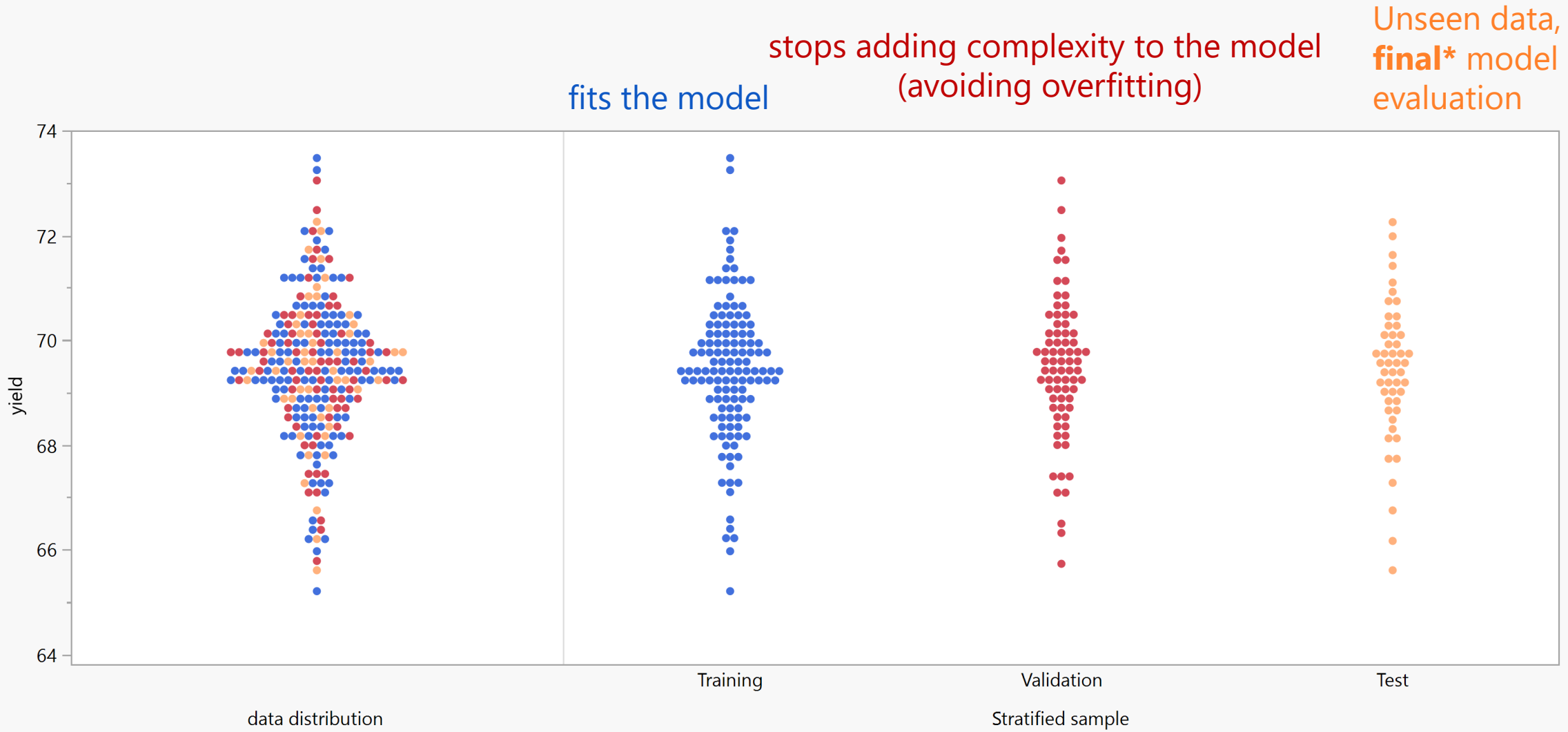
Prediction

It is difficult to make predictions, especially about the future

Anonymous*

(*) https://quoteinvestigator.com/2013/10/20/no-predict/

Better modeling by subsampling the data

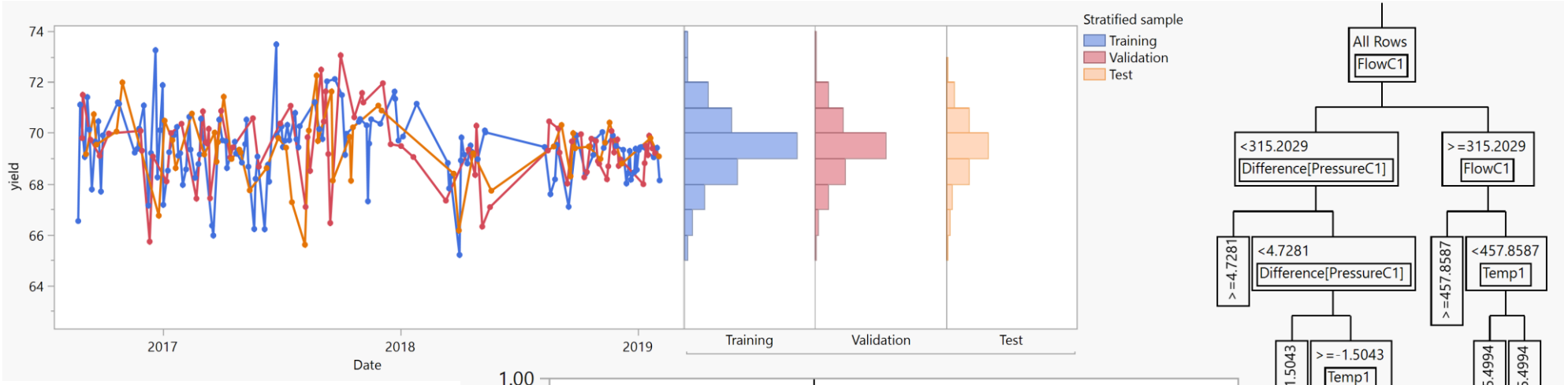


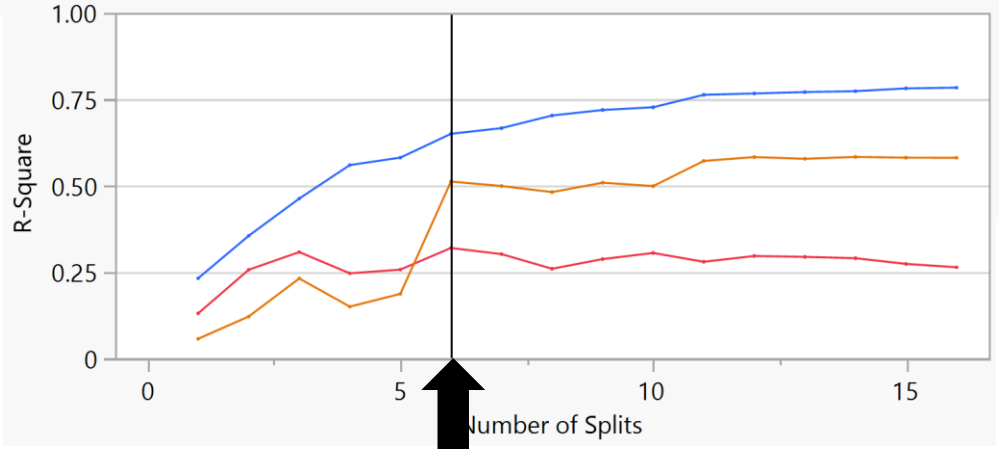
Training and validation splits can also be done multiple times (cross validation)

A rookie mistake? Train/validation/test

Training = fitting
Validation = overfitting
Test = unseen data

Stratified split (random split)



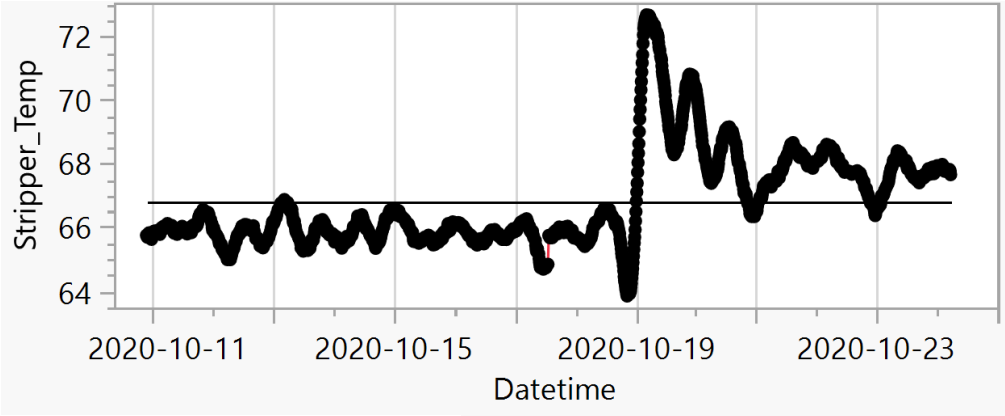


If sampling higher than process dynamics, data points will be very similar (autocorrelated).

Model will be overfitting without you knowing

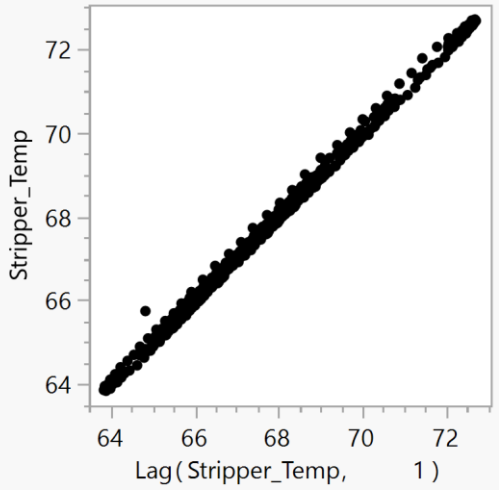
Training stops when validation performance decays

High autocorrelation



As time series, sensor data often autocorrelation (higher sampling rates). Historians can also interpolate data if high compression was configured (data quality)

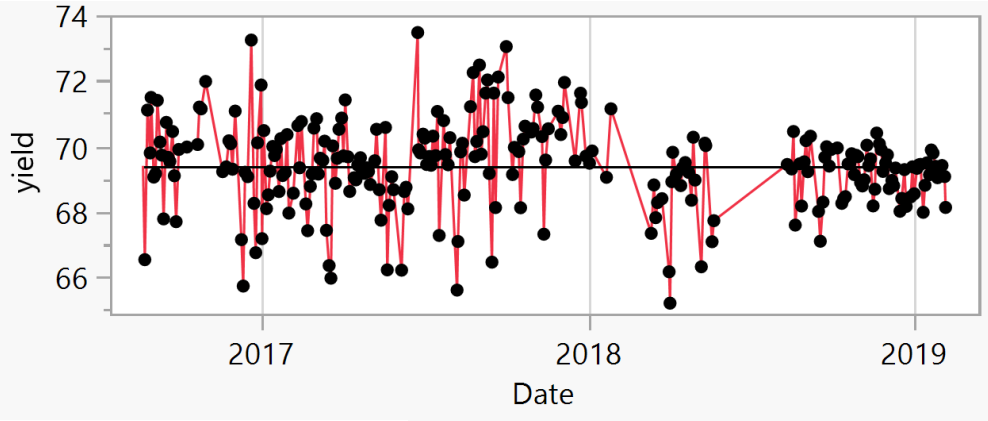
Lag Plot



Time Series Basic Diagnostics

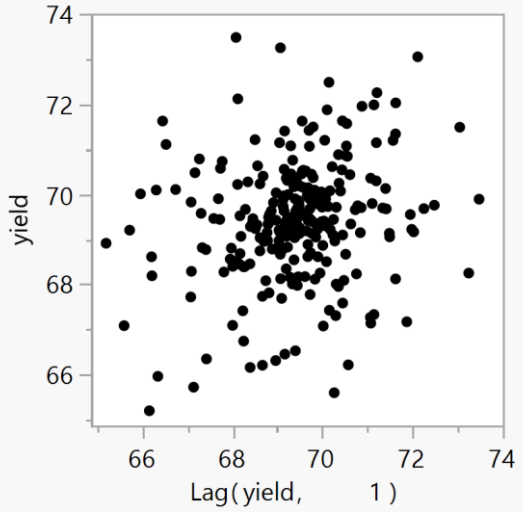
Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.9988		1918.29	<.0001*	1	0.9988	
2	0.9962		3827.83	<.0001*	2	-0.5363	
3	0.9929		5725.59	<.0001*	3	-0.0549	
4	0.9890		7609.44	<.0001*	4	-0.0597	
5	0.9845		9477.01	<.0001*	5	-0.1442	
6	0.9791		11325.3	<.0001*	6	-0.1914	
7	0.9728		13150.8	<.0001*	7	-0.1709	
8	0.9656		14950.2	<.0001*	8	-0.1307	
9	0.9575		16720.5	<.0001*	9	-0.0913	
10	0.9486		18459.2	<.0001*	10	-0.0828	
11	0.9390		20163.7	<.0001*	11	-0.0731	
12	0.9287		21831.9	<.0001*	12	-0.0416	
13	0.9178		23461.9	<.0001*	13	-0.0284	
14	0.9062		25051.8	<.0001*	14	0.0220	
15	0.8942		26600.6	<.0001*	15	0.0309	
16	0.8816		28107.0	<.0001*	16	0.0264	
17	0.8687		29570.3	<.0001*	17	0.0460	
18	0.8553		30989.8	<.0001*	18	0.0473	
19	0.8417		32365.0	<.0001*	19	0.0459	
20	0.8277		33695.7	<.0001*	20	0.0509	
21	0.8136		34981.9	<.0001*	21	0.0482	
22	0.7993		36224.0	<.0001*	22	0.0541	
23	0.7849		37422.3	<.0001*	23	0.0418	
24	0.7704		38577.5	<.0001*	24	0.0503	
25	0.7560		39690.3	<.0001*	25	0.0503	

No autocorrelation



Distillation column example sampled only when process reached steady state, hence yield rows are not correlated to each other

Lag Plot

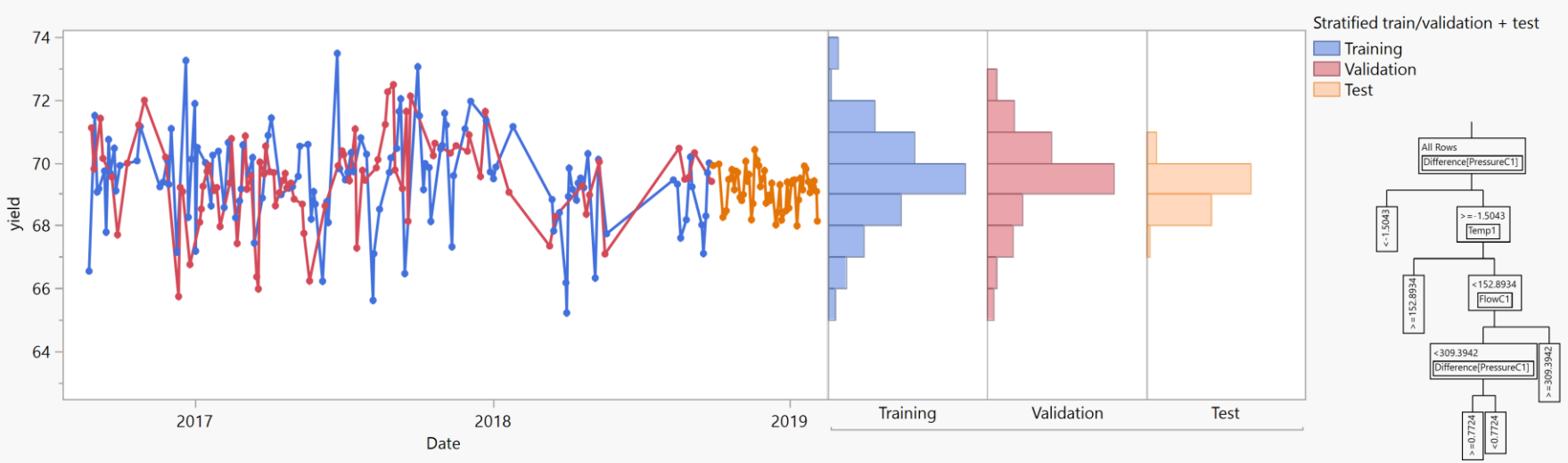


Time Series Basic Diagnostics

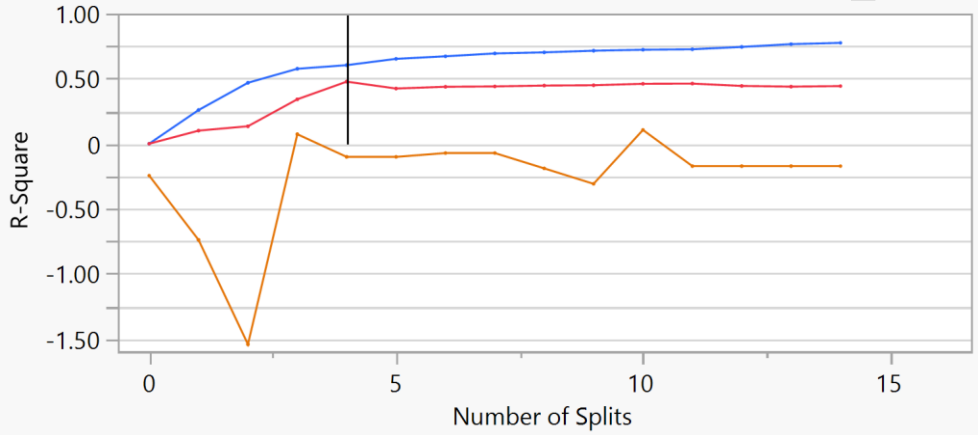
Lag	AutoCorr	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value	Lag	Partial	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
0	1.0000		.	.	0	1.0000	
1	0.1971		9.9490	0.0016*	1	0.1971	
2	0.0593		10.8526	0.0044*	2	0.0213	
3	0.1152		14.2758	0.0026*	3	0.1036	
4	0.0558		15.0835	0.0045*	4	0.0141	
5	0.0305		15.3251	0.0091*	5	0.0116	
6	0.0732		16.7242	0.0104*	6	0.0556	
7	0.0302		16.9629	0.0176*	7	-0.0014	
8	-0.0093		16.9857	0.0303*	8	-0.0229	
9	0.0355		17.3189	0.0439*	9	0.0292	
10	0.0832		19.1558	0.0383*	10	0.0690	
11	0.0746		20.6398	0.0373*	11	0.0496	
12	0.0061		20.6498	0.0558	12	-0.0299	
13	0.0183		20.7399	0.0782	13	0.0033	
14	-0.0003		20.7400	0.1085	14	-0.0186	
15	0.0934		23.1041	0.0820	15	0.0976	
16	0.0319		23.3804	0.1039	16	-0.0147	
17	0.0823		25.2318	0.0896	17	0.0750	
18	-0.0233		25.3802	0.1148	18	-0.0719	
19	0.0157		25.4485	0.1463	19	0.0262	
20	0.0506		26.1586	0.1606	20	0.0230	
21	0.0607		27.1841	0.1648	21	0.0386	
22	0.1495		33.4282	0.0561	22	0.1341	
23	-0.0890		35.6479	0.0448*	23	-0.1691	
24	0.0410		36.1214	0.0534	24	0.0894	
25	0.0436		36.6594	0.0622	25	-0.0205	

Keeping most recent data as test

Stratified split + test



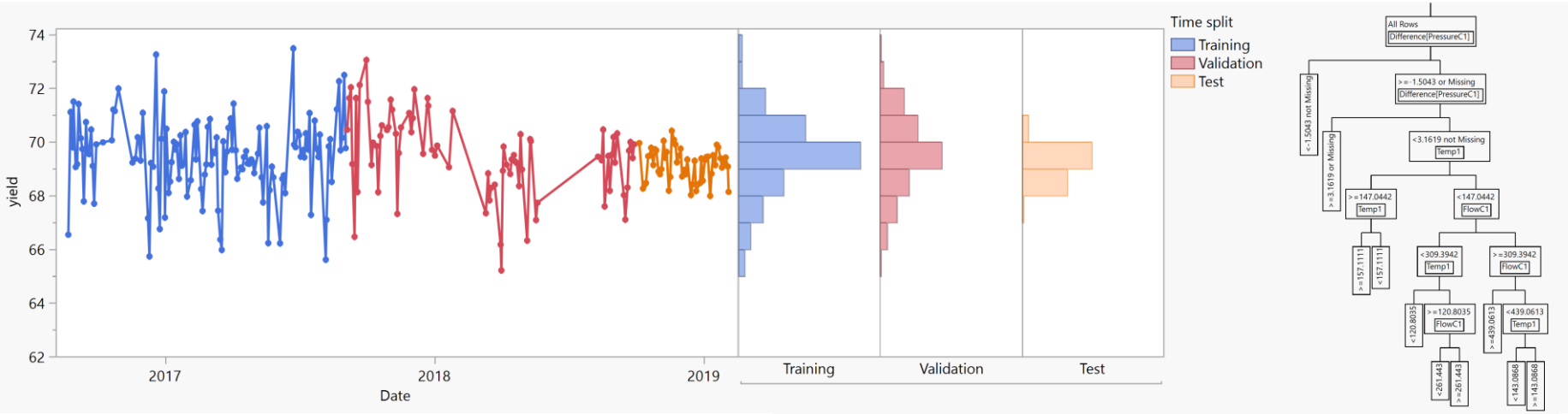
Model describes training/validation data, but it won't predict well with recent data (see test results and their distribution)

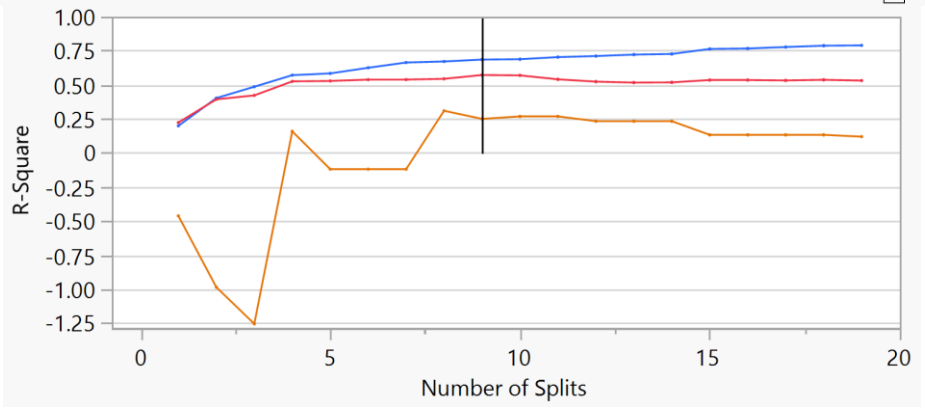


Time split, usually a better approach

Training = fitting
Validation = overfitting
Test = unseen data

Time-split (or cut point)



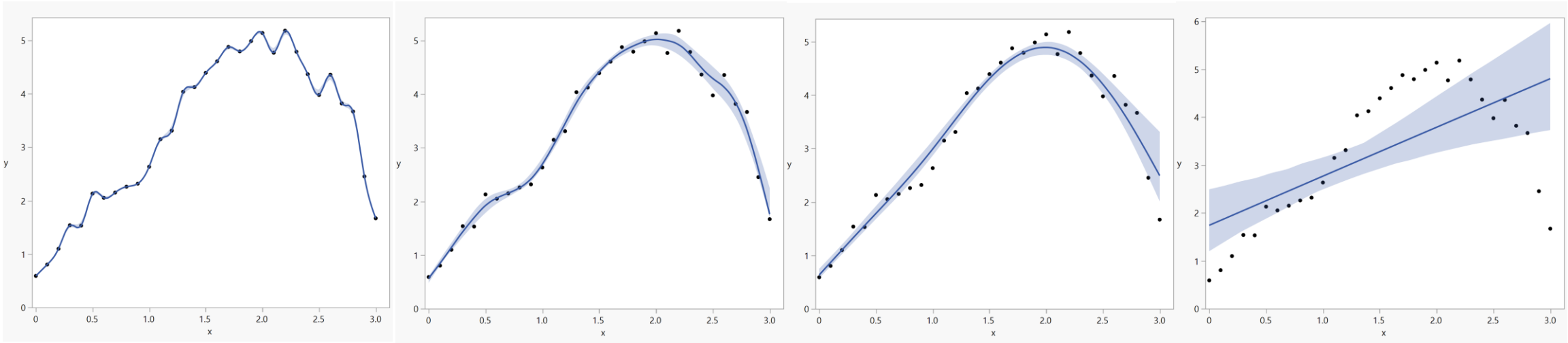


Model describes training/validation data, but it won't predict well with recent data if process has changed over time. **Use your test data** wisely **and** try to obtain **simple models**.

An accuracy-simplicity tradeoff (penalized regression)

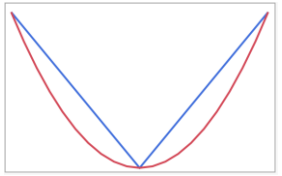
overfitting

underfitting



$$Model: \ \hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \ ...$$

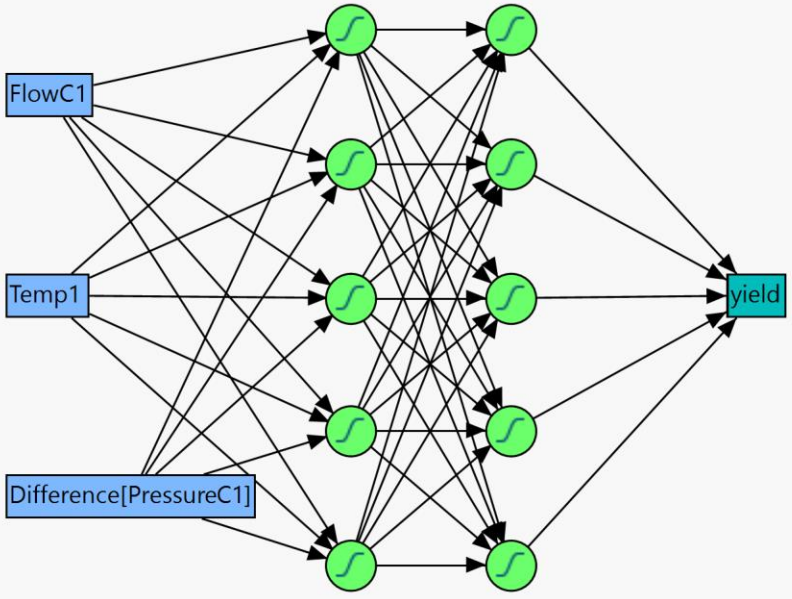
$$Minimize: \quad \sum (y - \hat{y})^2 + \lambda \sum |\beta_j|$$

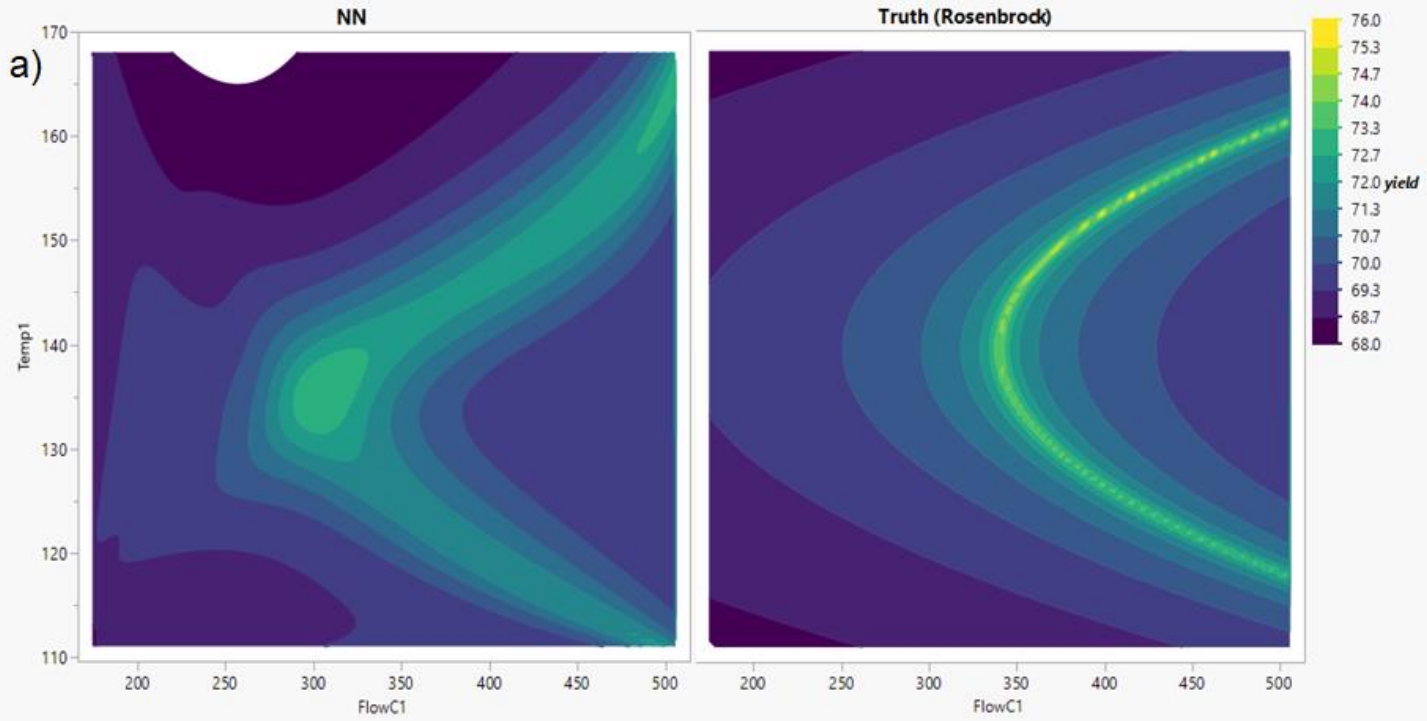


Lasso regression minimizes prediction error as well as the model coefficients, being able to eliminate them entirely. Other penalization techniques exist (e.g., ridge regression, BIC, AIC…)
How to select the best one? Don't use all your data ☺

Complex models can approximate anything

- By combining multiple nodes and layers, Neural Networks can be used to capture intricate non-linear correlations within datasets, so what are they?

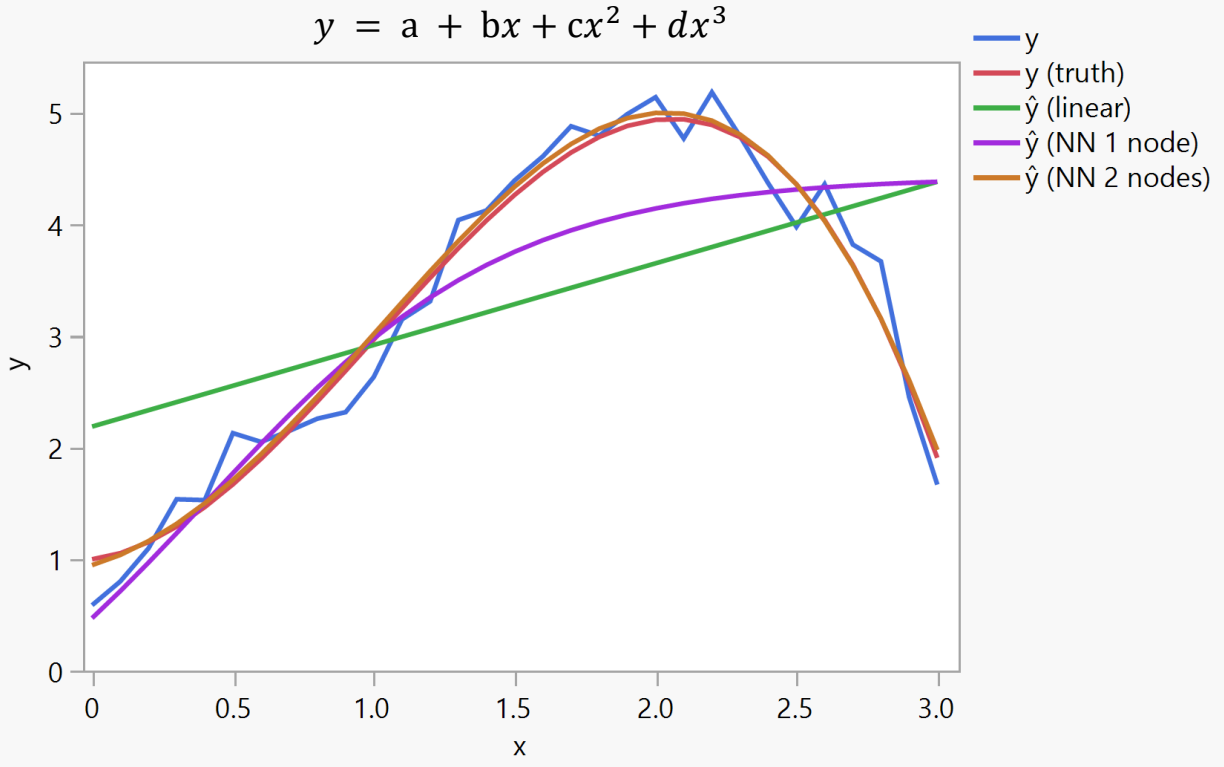


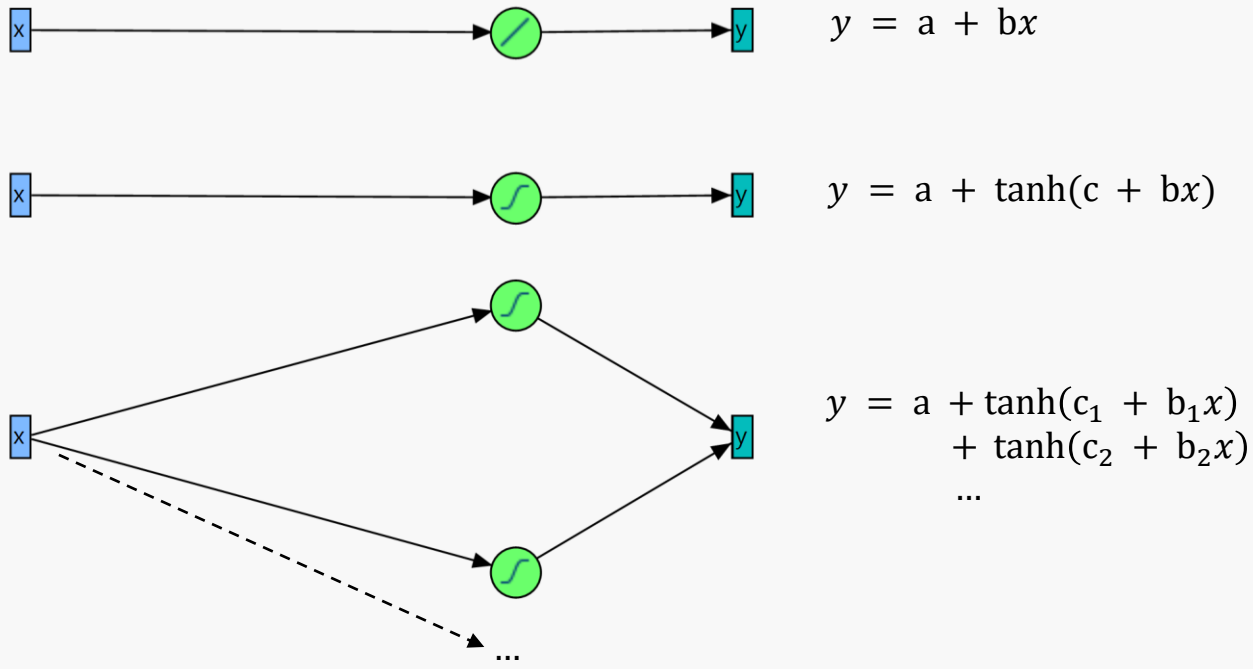


Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

Neural Networks are function approximators

- Like polynomials, Neural Networks approximate functions by increasing the number of non-linear elements (sigmoid-like functions call nodes or activation functions)

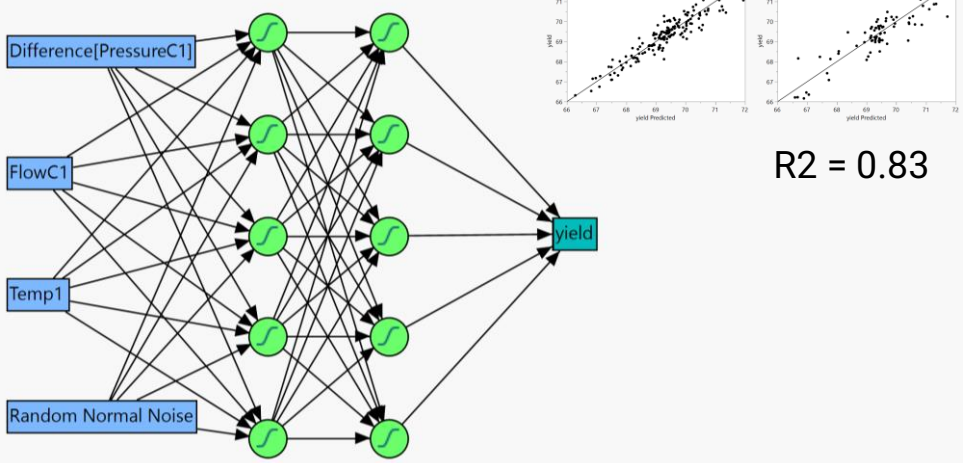




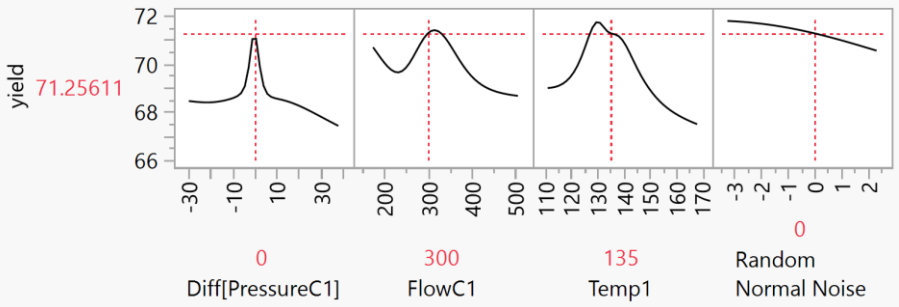
Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

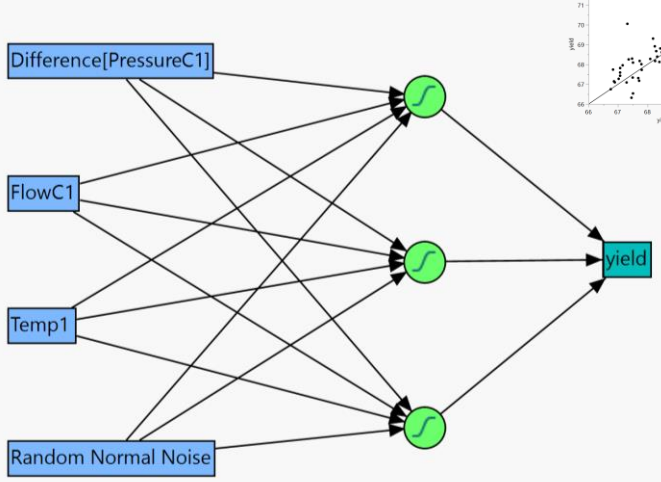
You probably don’t need deep learning

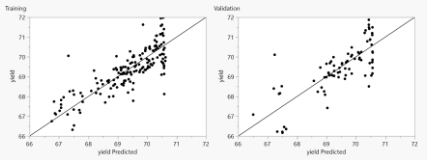
- Their can also fit the noise in the data, adding noise to see variable importance is recommend to detect overfitting and discard less impactful parameters.



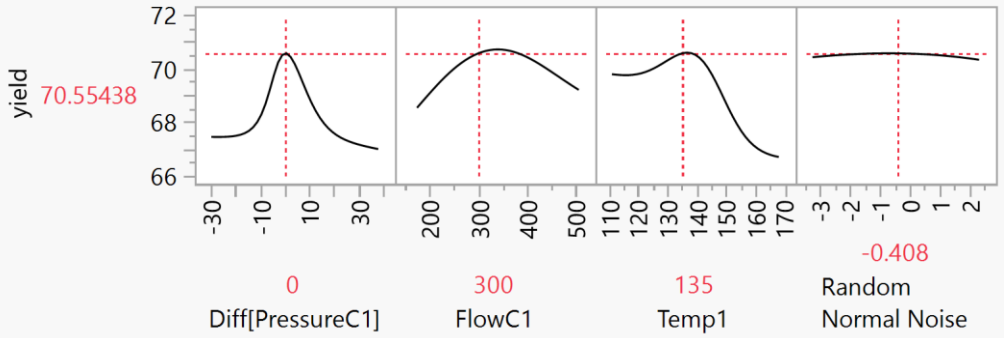
R2 = 0.83







R2 = 0.62

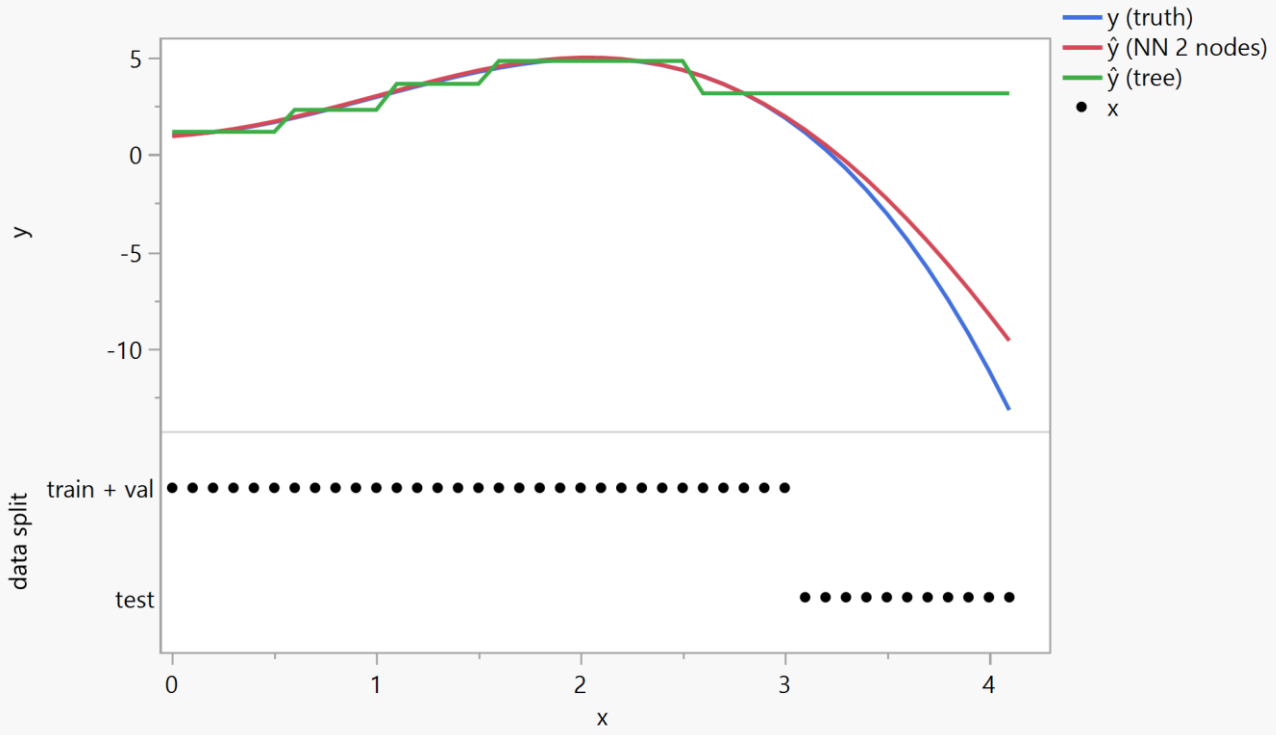


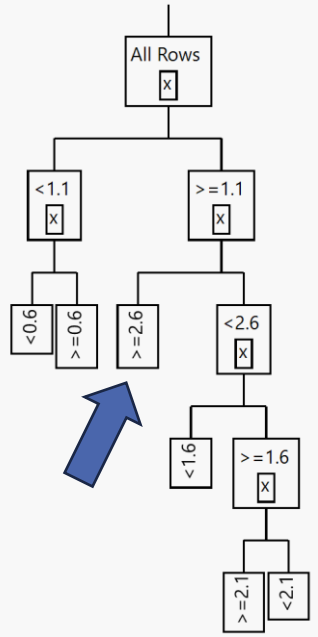
Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

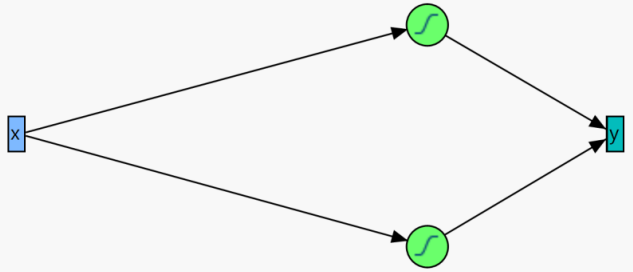
Interpolation vs extrapolation

- Like polynomials, Neural Networks extrapolate outside training data
- In practice, Neural Networks require anomaly detection (unsupervised learning) to know how far from known data the model is
- Tree-based models flatten outside training data

$$y = a + bx + cx^2 + dx^3$$





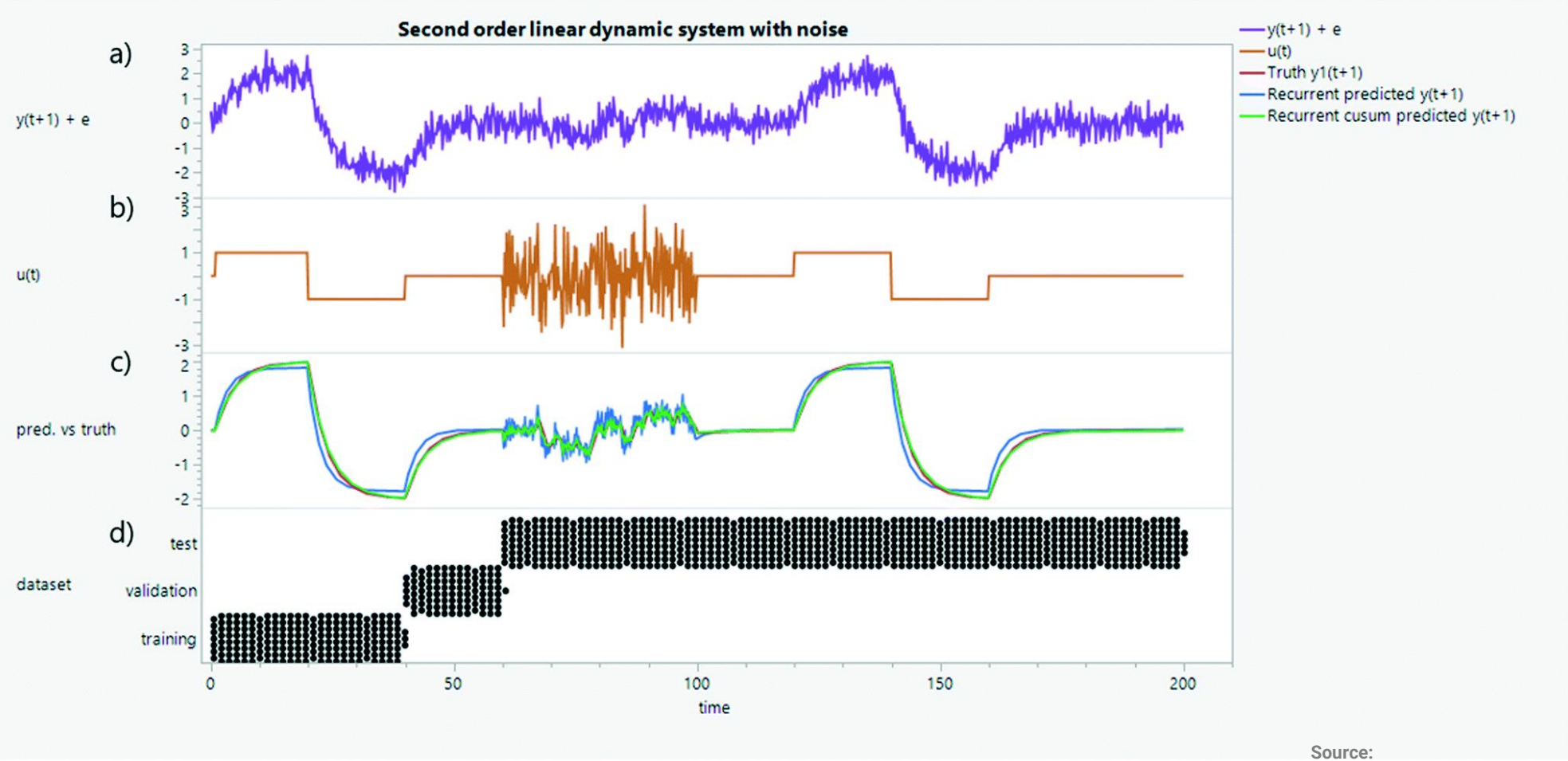


$$y = a + \tanh(c_1 + b_1 x) + \tanh(c_2 + b_2 x)$$

Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

Application: System Identification

- Predicting one step ahead



Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

Summary

- AutoML compares multiple models and optimizes their tuning parameters, so you don’t have to
- Prediction is only important after problem is understood, focus on descriptive models first
- Tree-based models are best for screening and understanding. SHAP is a common approach to explain them
- Neural networks can produce more accurate results at higher risk of overfitting and extrapolation
- Use synthetic noise and data splits for better/simpler models
- We aware of process dynamics (autocorrelation) when splitting data, and the field of system identification for one-step ahead model predictions.

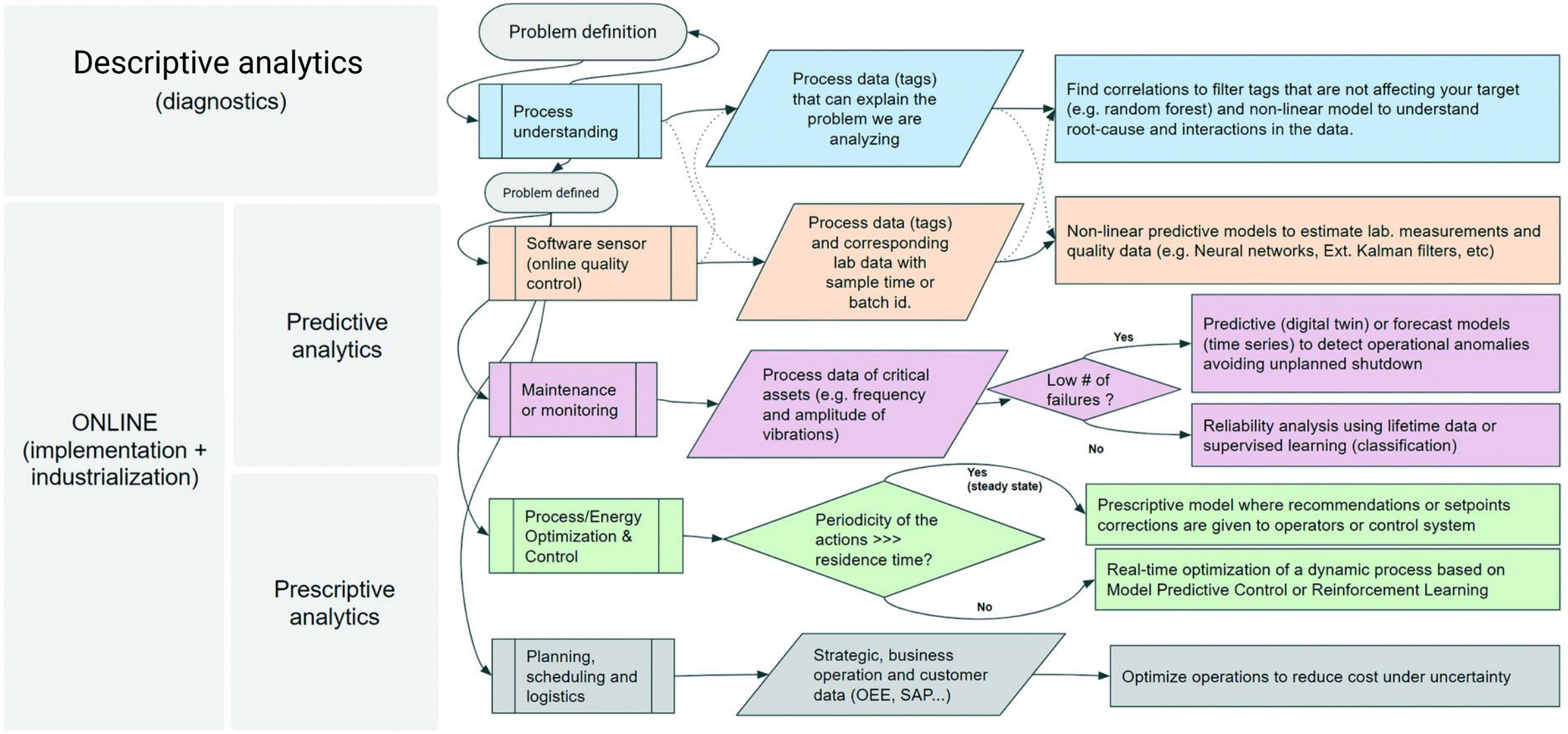
Sargent Centre for Process Systems Engineering

Industrial applications

All models are wrong, but some are useful

George Box

Industrial data science applications



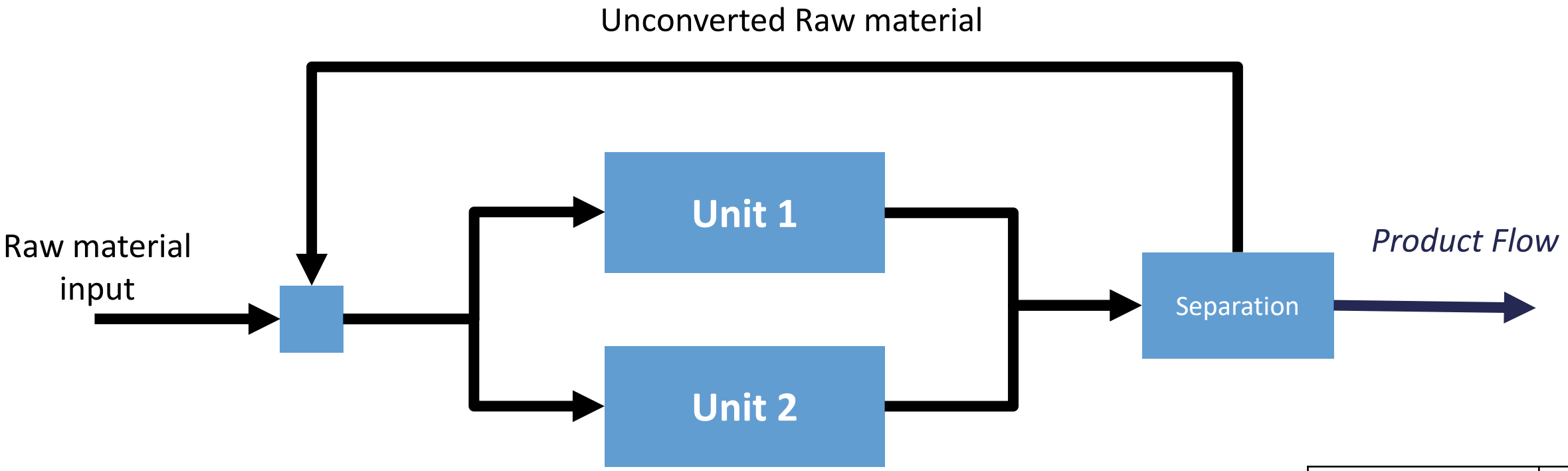
Source:
Industrial data science – a review of machine learning applications for chemical and process industries
React. Chem. Eng., 2022, 7, 1471-1509

BASF

https://community.jmp.com/t5/Discovery-Summit-Europe-2018/Discovering-hidden-relationships-in-production-data-EU2018-113/ta-p/51283

https://community.jmp.com/t5/Discovery-Summit-Europe-2019/DOE-for-World-Scale-Manufacturing-Processes-Can-We-Do-Better/ta-p/184245

The problem



$$Yield = \frac{Product\ Flow}{Raw\ Material\ Input}$$

$$Ratio = \frac{Unit\ 1}{Unit\ 2}$$

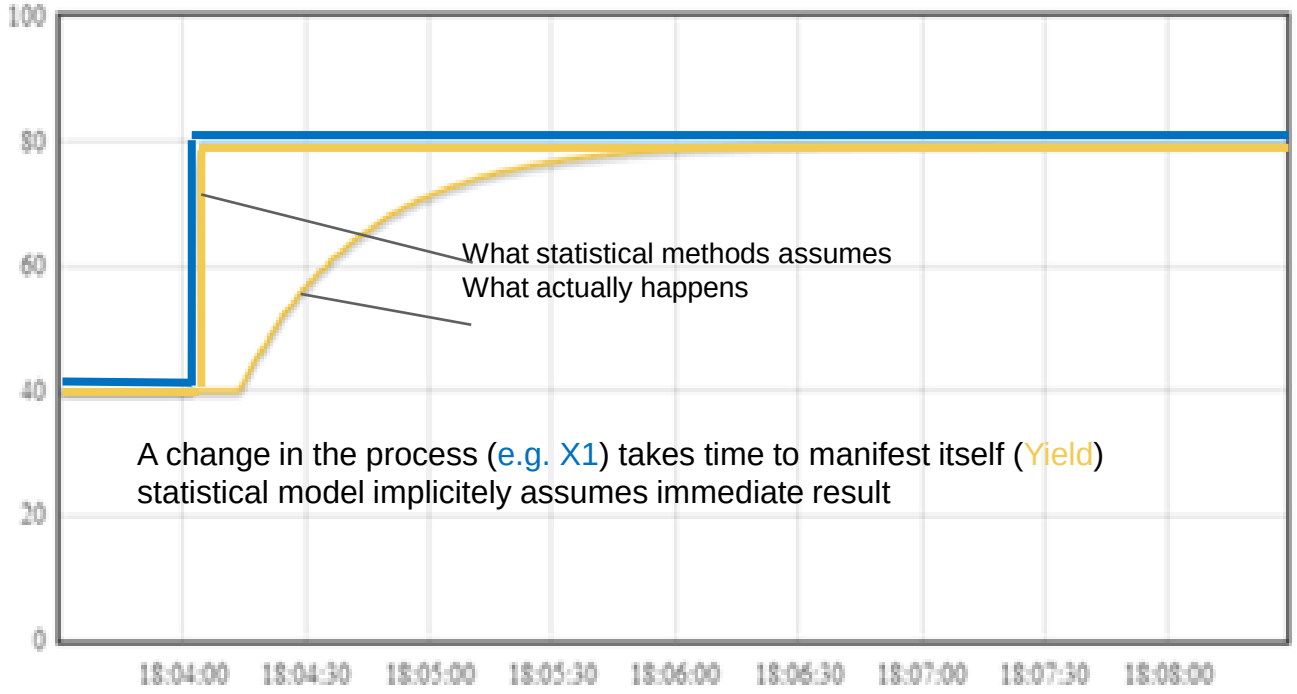
$$Conversion = \frac{Product\ Flow}{Raw\ Material\ Input + Unconverted\ Raw\ material}$$

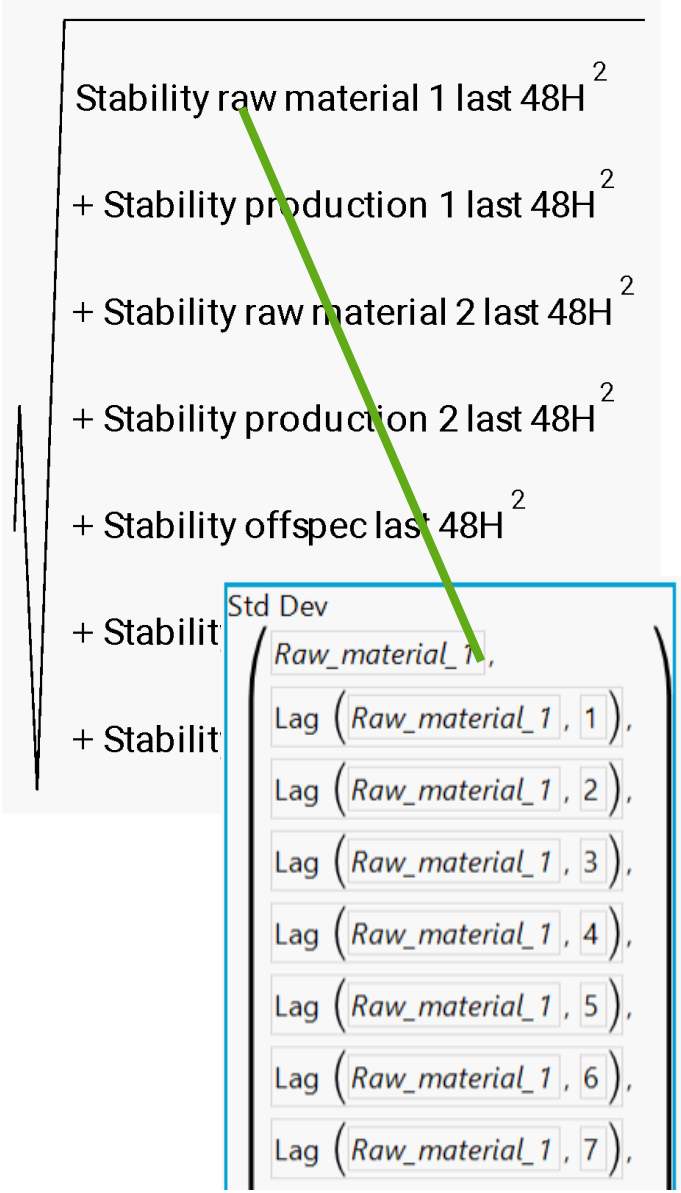
Yield	Y
Product Flow	X
Conversion	X
Ratio	X

Data analysis in continous plant

Do we need non steady-state data?

- Dynamic effects: after changing the process, it takes up to 48 hours to get to a new steady state condition (and often another change is made within that time seldom at steady state)

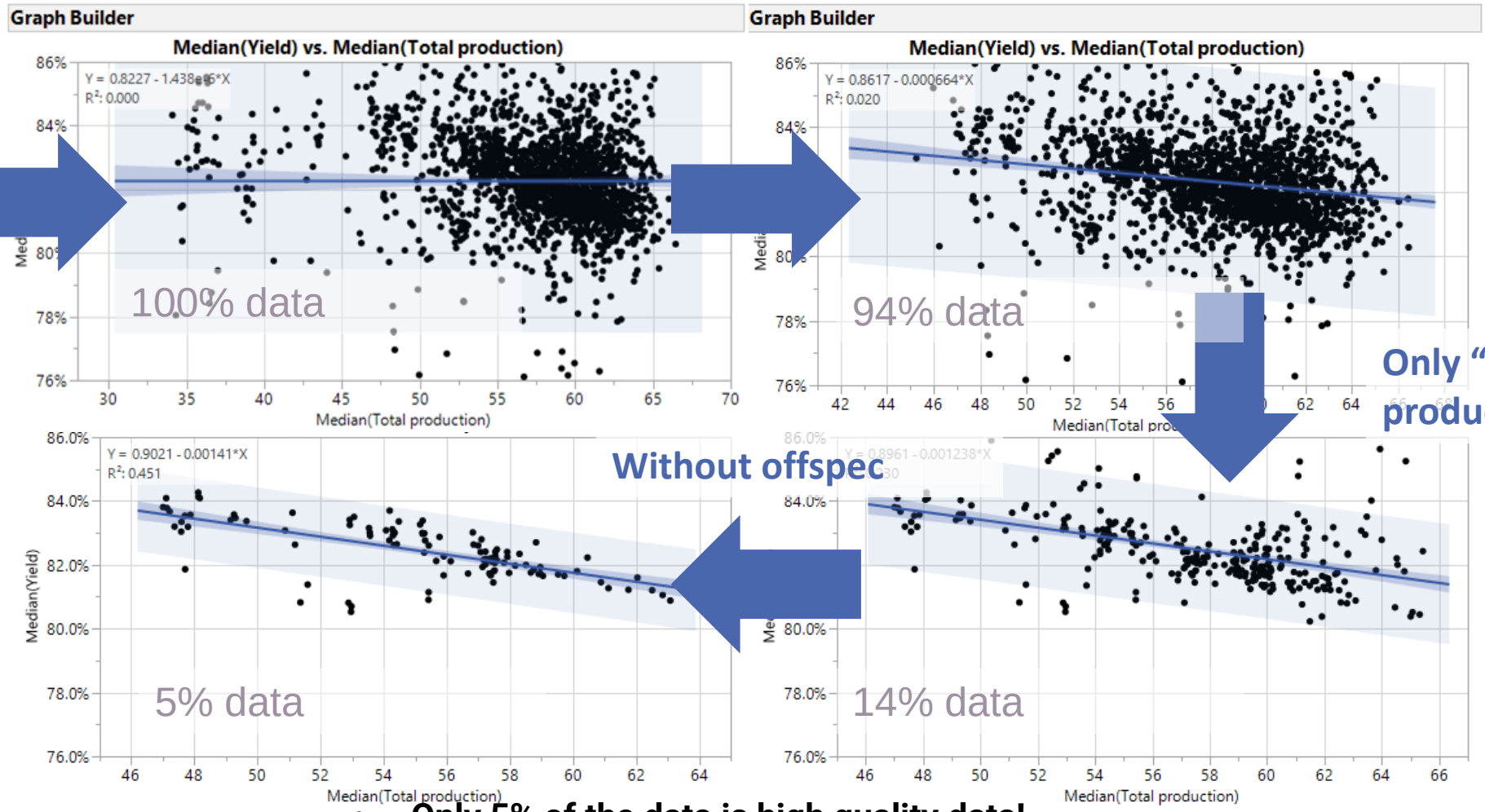




Data cleaning is key

Only medium to high production volumes

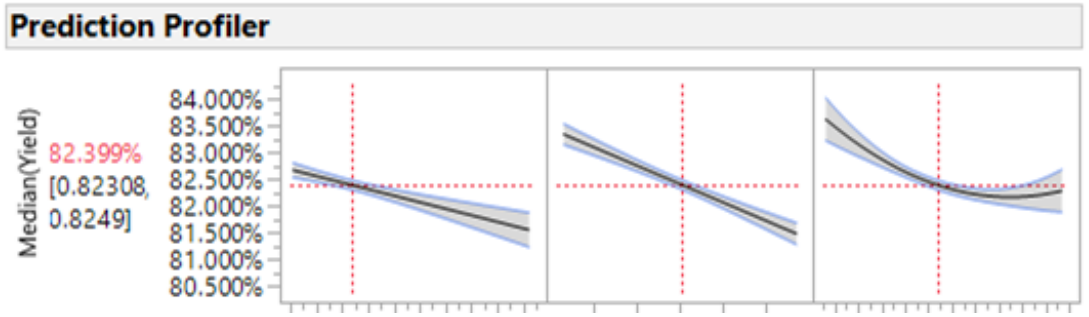
Data: 24H average values

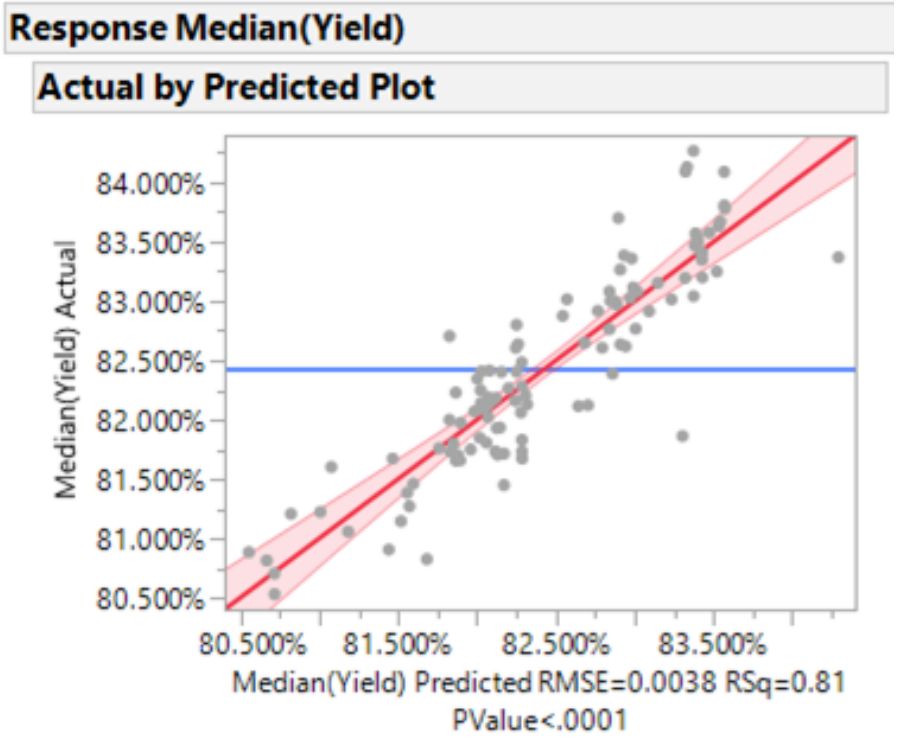


- **Only 5% of the data is high quality data!**

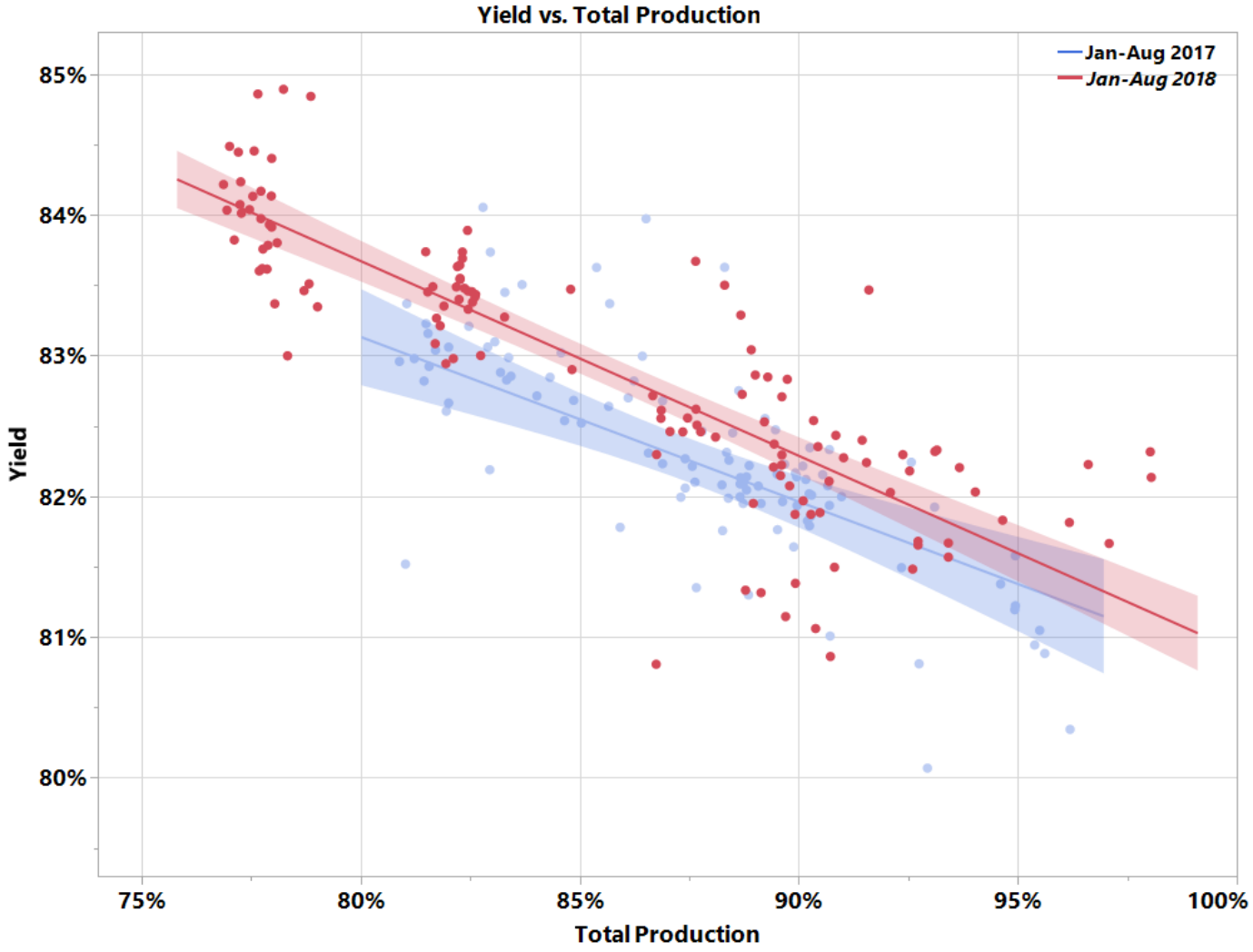
A regression model is enough

- OLS model on high quality data
- Key process variables can be explained by process expert
- Impact of key process variables can be accurately estimated (including an interaction and a non linear effect)
- Result
  - Optimal settings for yield (at a certain load)
  - Prediction of expected yield (an detection of deviations)





What benefit did we achieve



Ambition at the start of the project

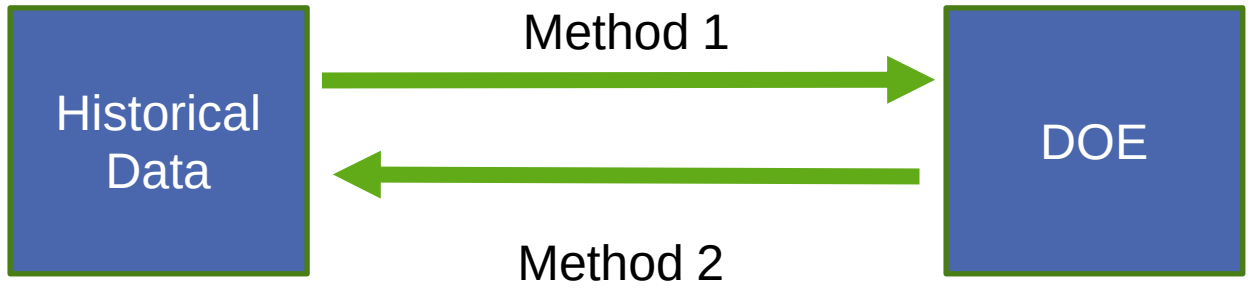
- 0.3% yield improvement (0.4 M€/y)

Result

- Total Yield improvement of 1.1% (1.4 M€/y)
  (Jan-Aug 2018 vs Jan-Aug 2017)
- Similar benefit for 2019 wrt 2017

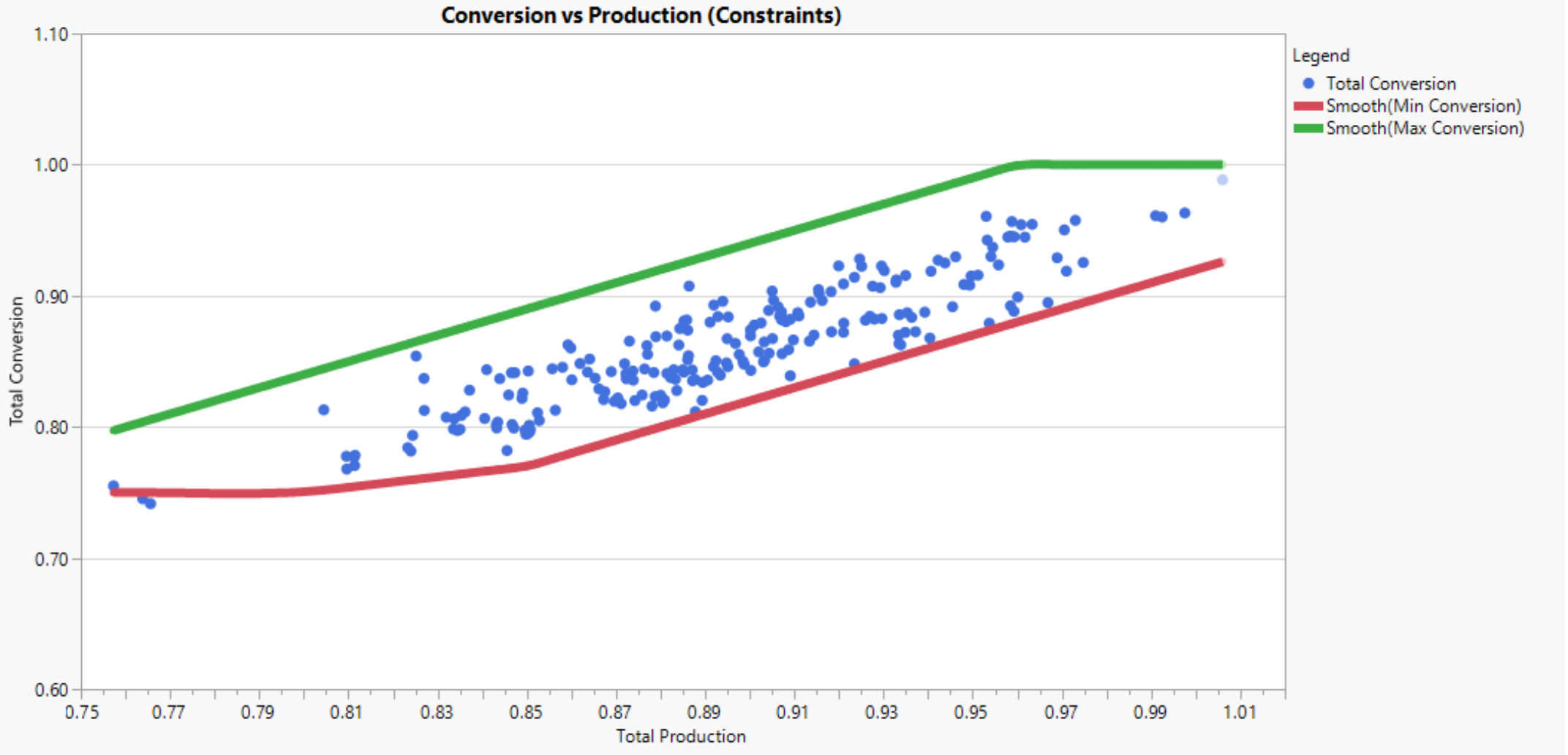
How can we further improve our insights?
a cost-effective DOE on an industrial process?

- The idea … leverage as much as possible historical data:
- Method 1: Augment "most reliable" historical datapoints with Space Filling Design
- Method 2: Reduce experiment cost of Advanced Design by leveraging historical data

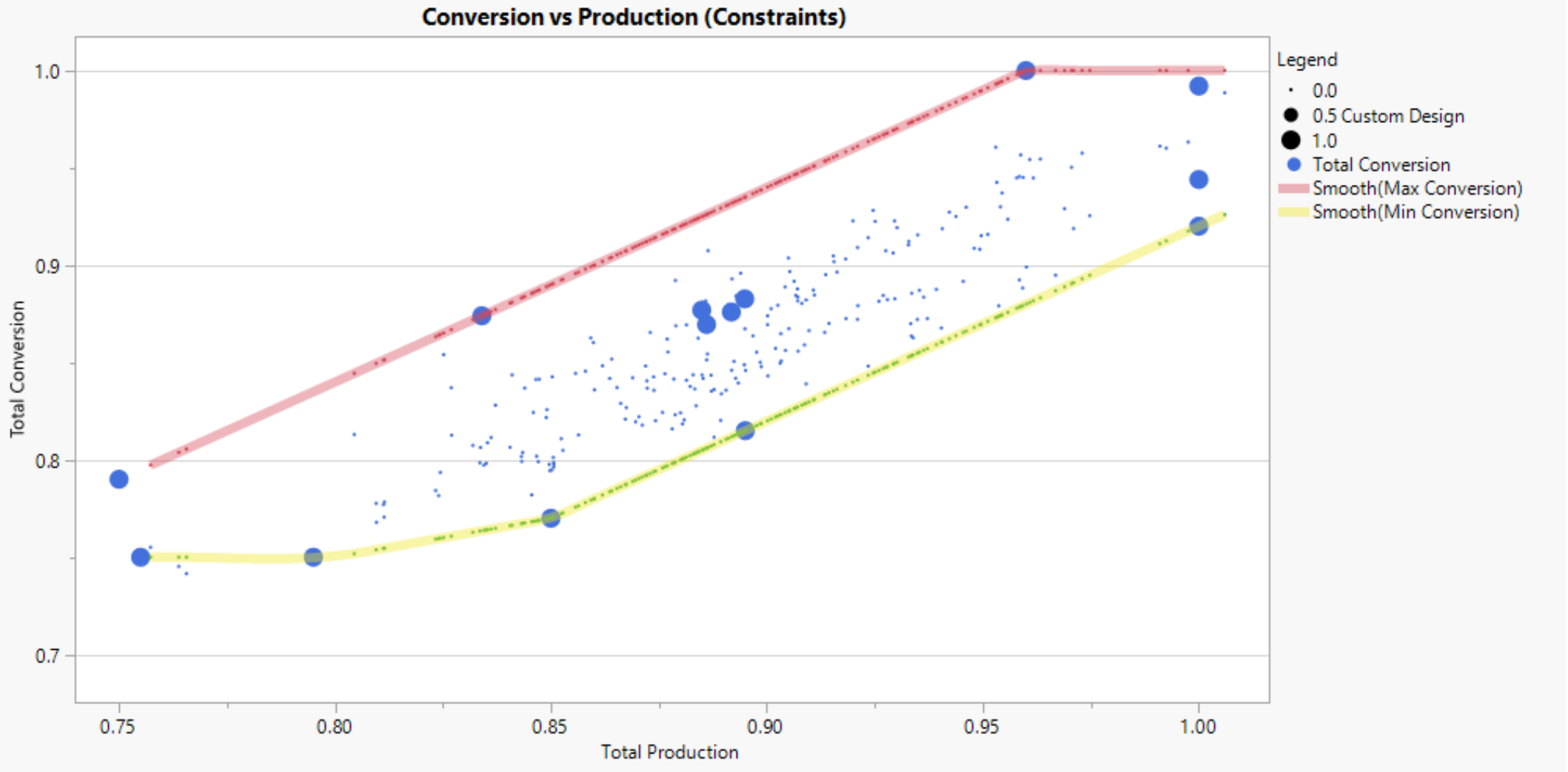


Are we missing information?

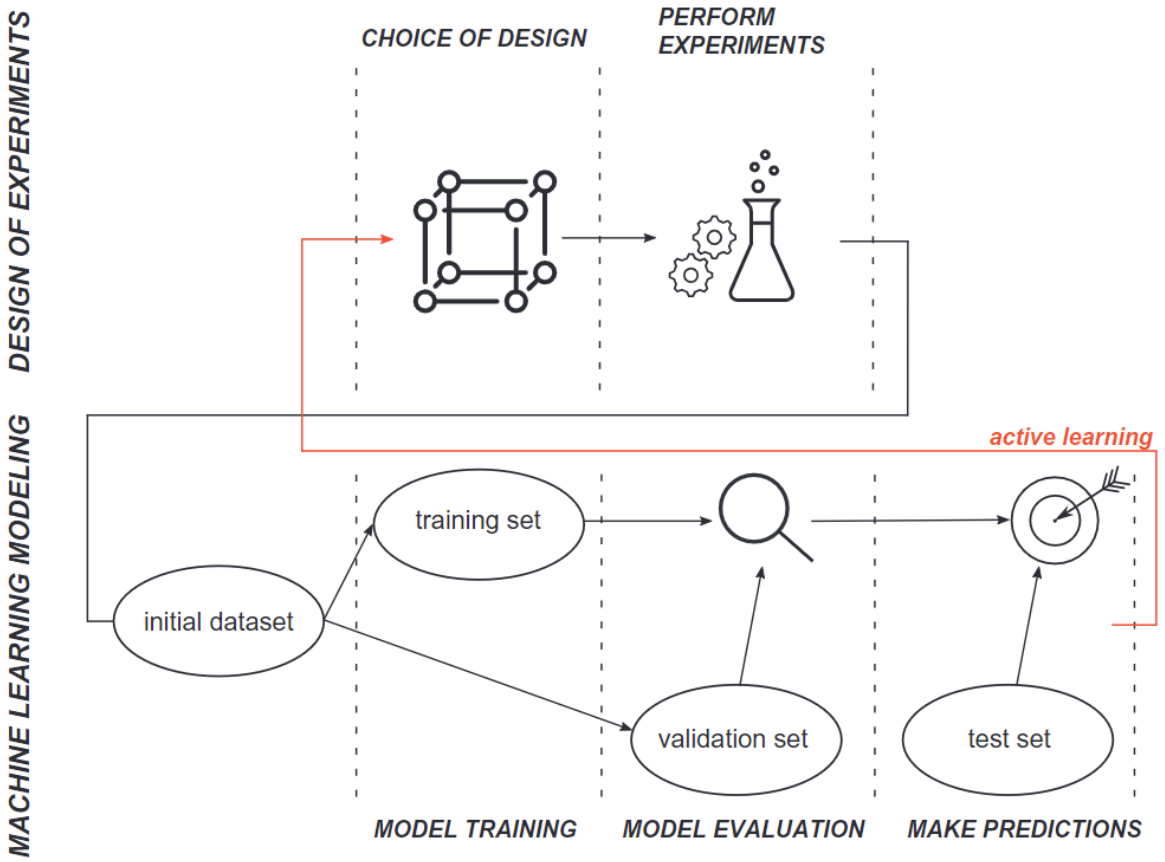
Historical data does not cover the feasible space evenly



Space Filling DOE with historical data as previous experiment



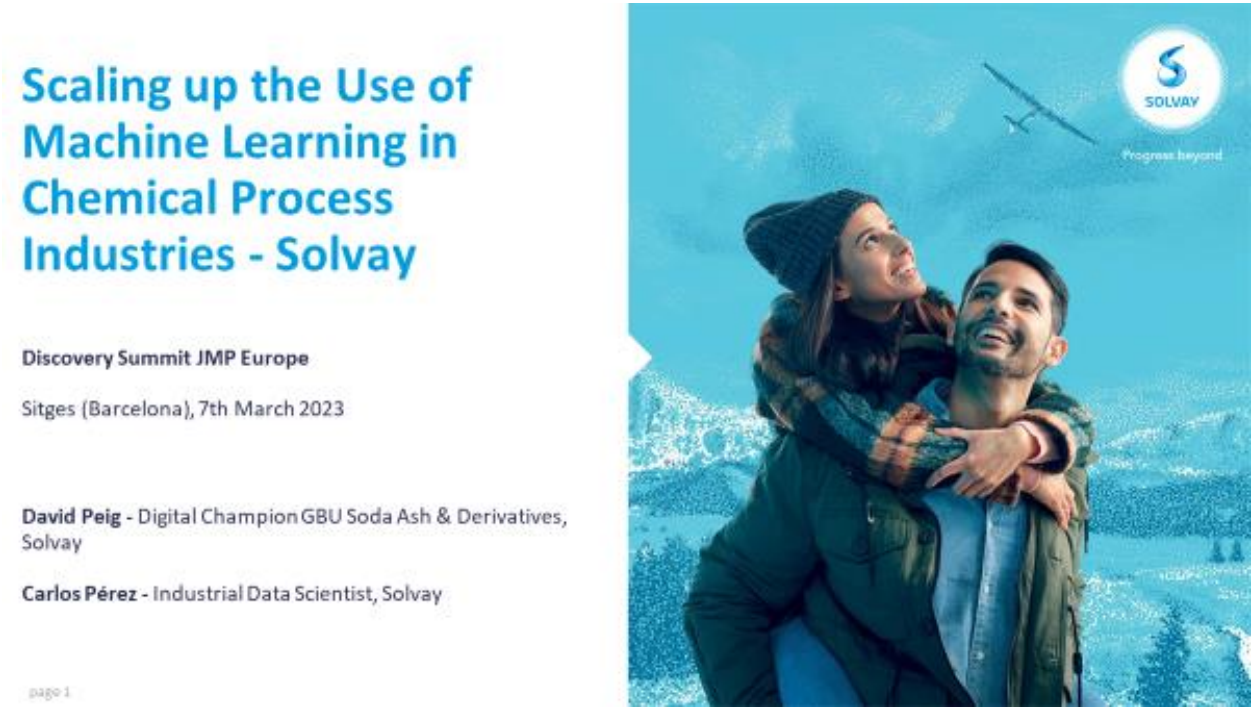
Where to learn more about DoE and Machine Learning



Source:
Design of Experiments and machine learning for product innovation: A systematic literature review Qual Reliab Engng Int.2022;38:1131–1156
Design choice and machine learning model performances Qual Reliab Engng Int.2022;38:3357–3378

SOLVAY

https://www.solvay.com/en/brands/soda-solvay



Scaling up the Use of Machine Learning in Chemical Process Industries - Solvay

Discovery Summit JMP Europe

Sitges (Barcelona), 7th March 2023

David Peig - Digital Champion GBU Soda Ash & Derivatives, Solvay

Carlos Pérez - Industrial Data Scientist, Solvay

page 1

https://community.jmp.com/t5/Discovery-Summit-Europe-2023/Scaling-up-the-Use-of-Machine-Learning-in-Chemical-Process/ta-p/572644

Data-driven approach improve plant performance

Problem

Historically low productivity across site

$$Na^{+}_{(aq)} + CO_{2\,(g)} + NH_{3\,(g)} + H_2O_{\,(l)} \rightarrow NaHCO_{3\,(s)} + NH_4^{+}{}_{(aq)}$$

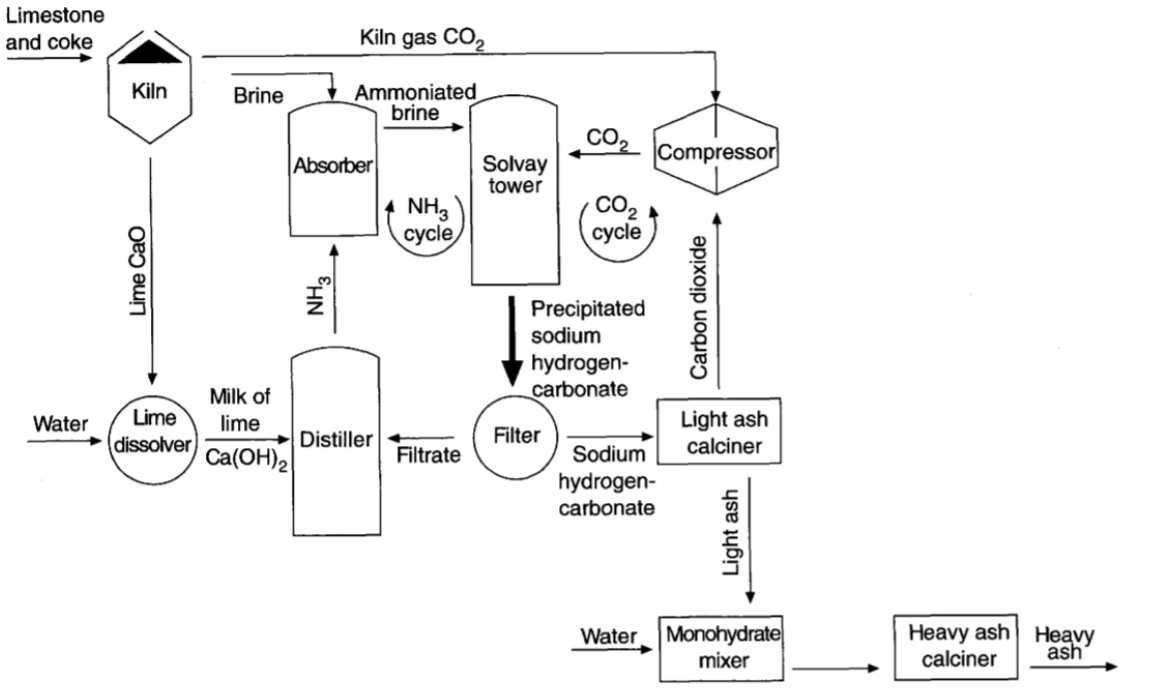


Fig 2 A simplified flow diagram for the Solvay process showing the ammonia and carbon dioxide cycles

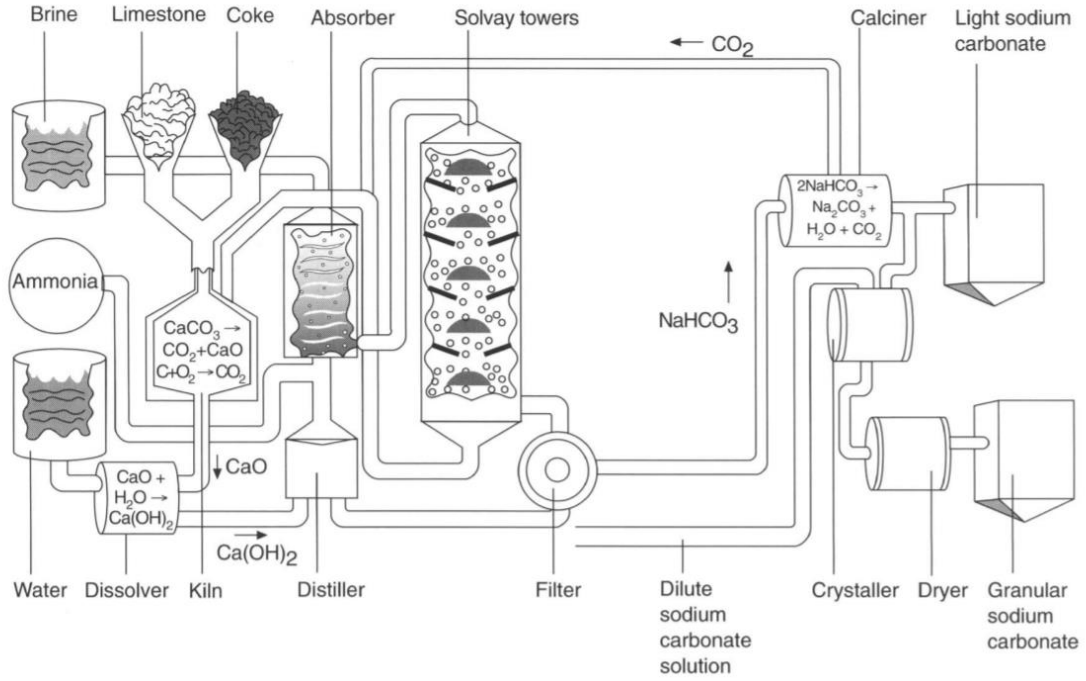


Fig 3 The Solvay process

Three different products are formed from this process, light sodium carbonate (light ash), granular sodium carbonate (heavy ash) and refined sodium hydrogencarbonate.

The main by-product is calcium chloride. A little of this can be sold for use in refrigeration, curing concrete and as a suspension in oil drilling. The bulk of it is disposed of in the nearby river Weaver.

Source: https://edu.rsc.org/resources/manufacturing-sodium-carbonate-by-the-solvay-process/2215.article

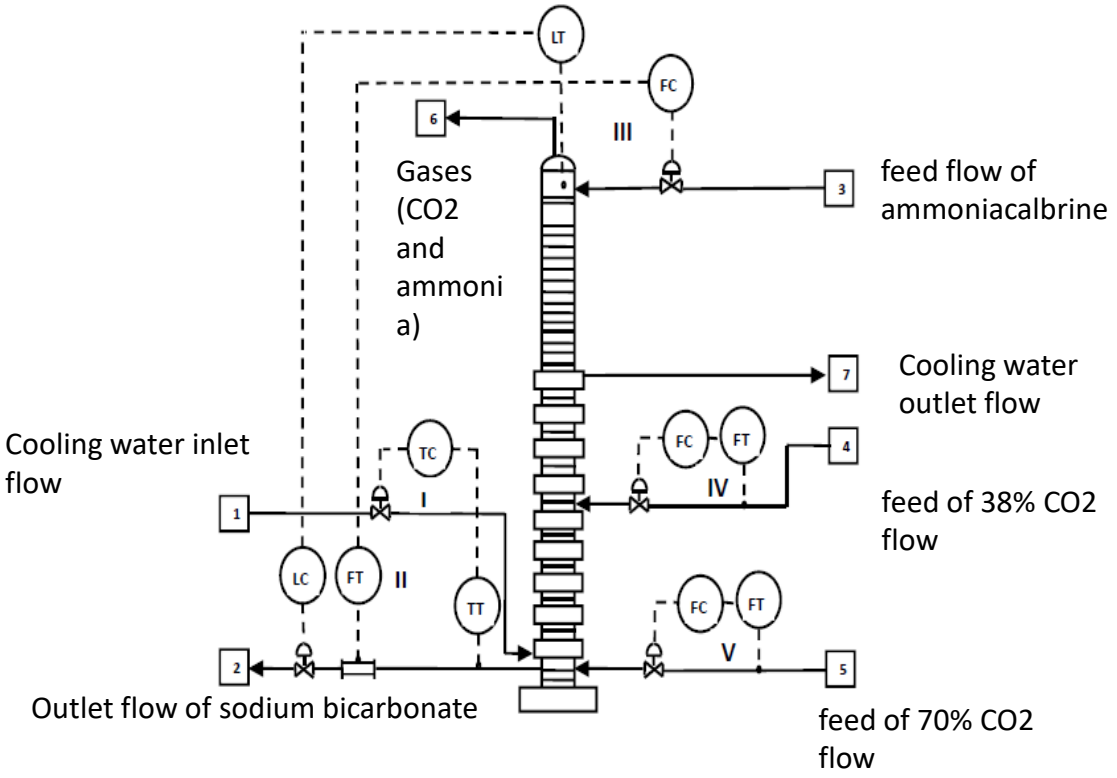
Data-driven approach improve plant performance

Context

- Plant structure across sector makes difficult to track the performance of individual equipment to track root-cause
- There is a lack of measurements for key process variables

What was done?

- Screening of multiple variables and select the most important ones to explain the variability of the target (yield)
- To visualize the long-term variability of the target and its relationship between the most important variables
- Data-driven predictive model on the ratios of key variables
- Optimization on data-driven model to recommend best setpoint across equipment



Source: http://www.chem.ubbcluj.ro/~studiachemia/issues/chemia2017_4/tom2/01Cristea_221_229.pdf

Debottlenecking of limekilns

Problem

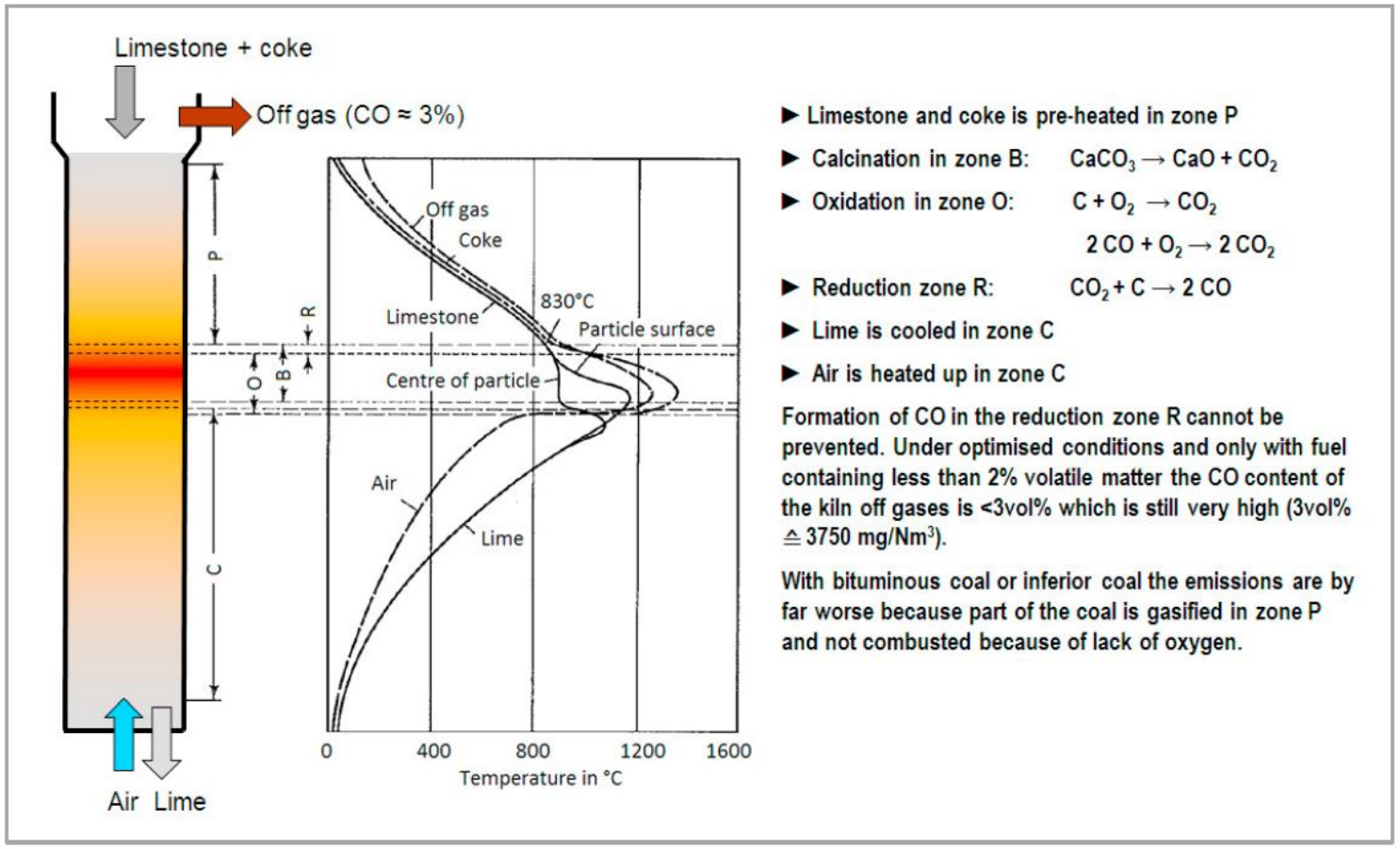
Energy-intensive process cannot effectively respond to the changes in raw materials, fuel variability, and setpoint of the plant leading to losses in productivity

Source: https://www.sciencedirect.com/science/article/pii/S1876610217327248

Debottlenecking of limekilns

Context

- The cooking of the limestone has a temperature profile that is optimal when stoichiometry is satisfied
- Measurement of the key process variables is challenging.

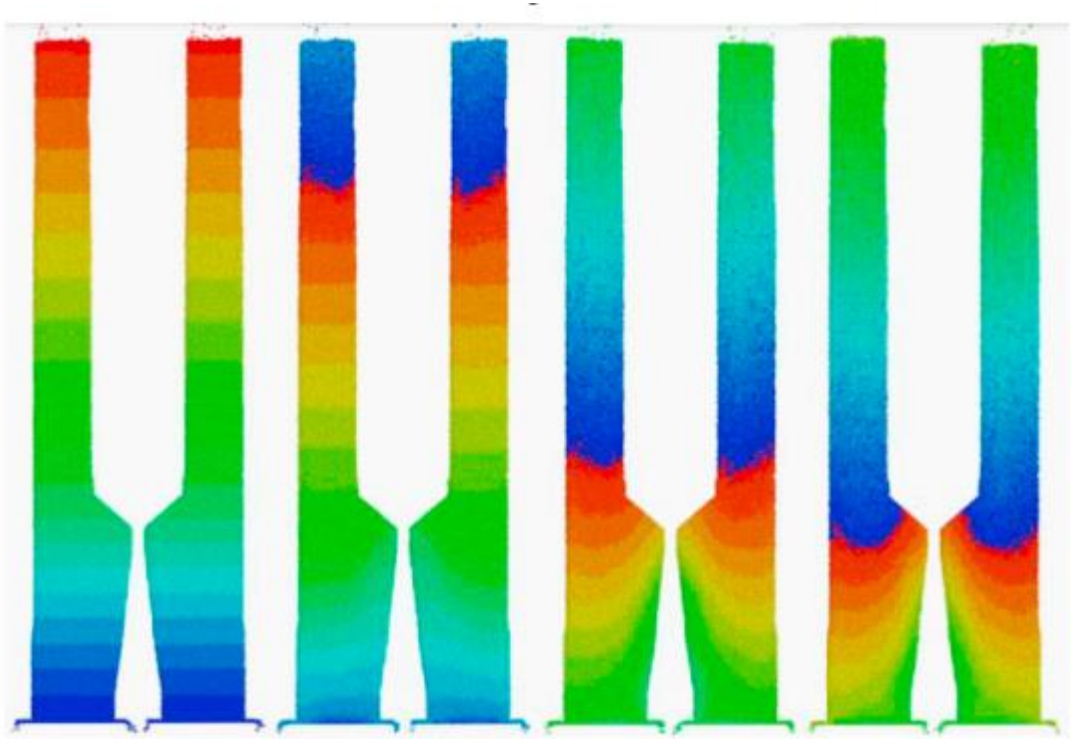


Source: https://www.sciencedirect.com/science/article/pii/S1876610217327248

Debottlenecking of limekilns

What was done?

- New measurement of key process variables allow us to understand behaviour of kiln
- Screening of process variables confirms main contributions of variability
- Soft sensor provides useful KPI
- New control system design takes into account the insights from the data

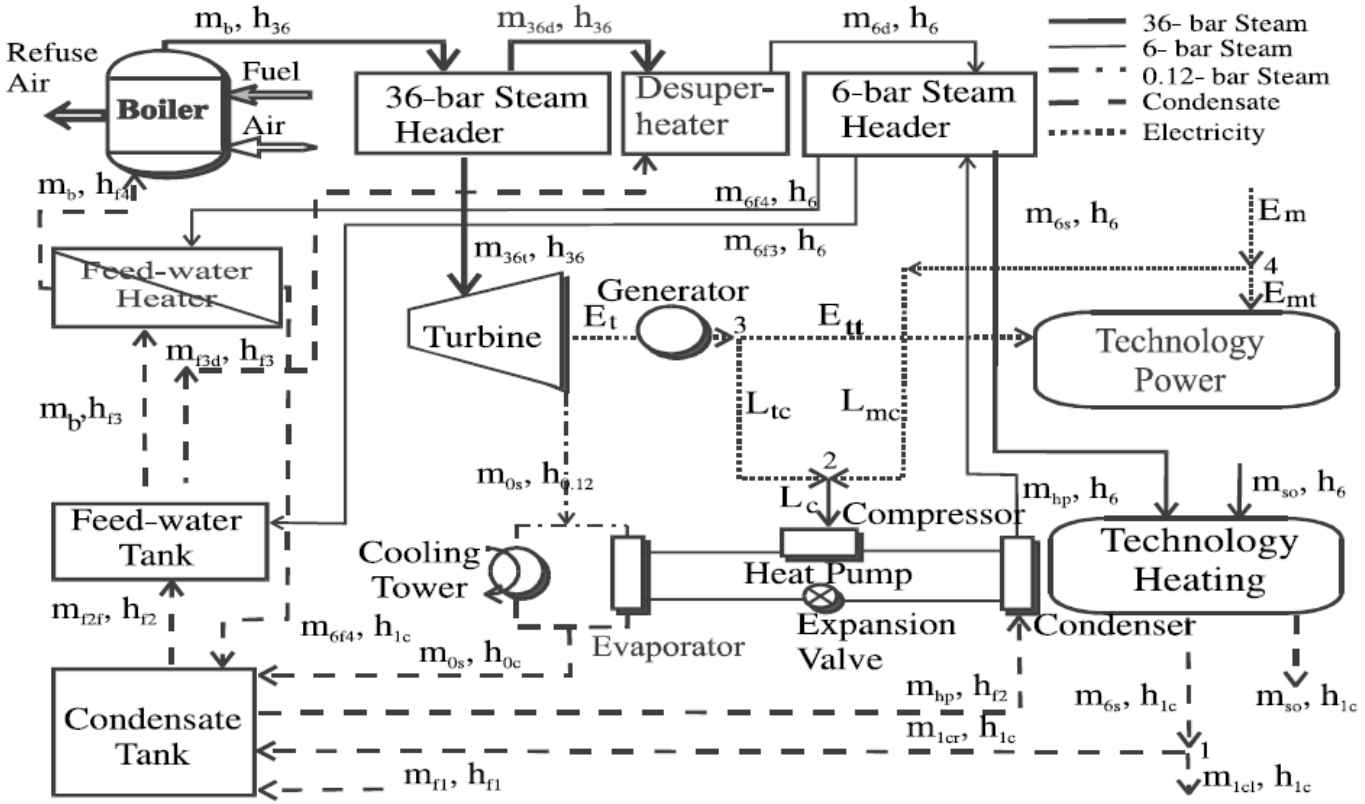


Source: https://www.sciencedirect.com/science/article/pii/S1876610217327248

Leveraging data-driven model in optimization of steam networks

Problem

Energy-intensive steam network cannot effectively respond to the changes in prices and demand causing financial losses and negative environmental impact.

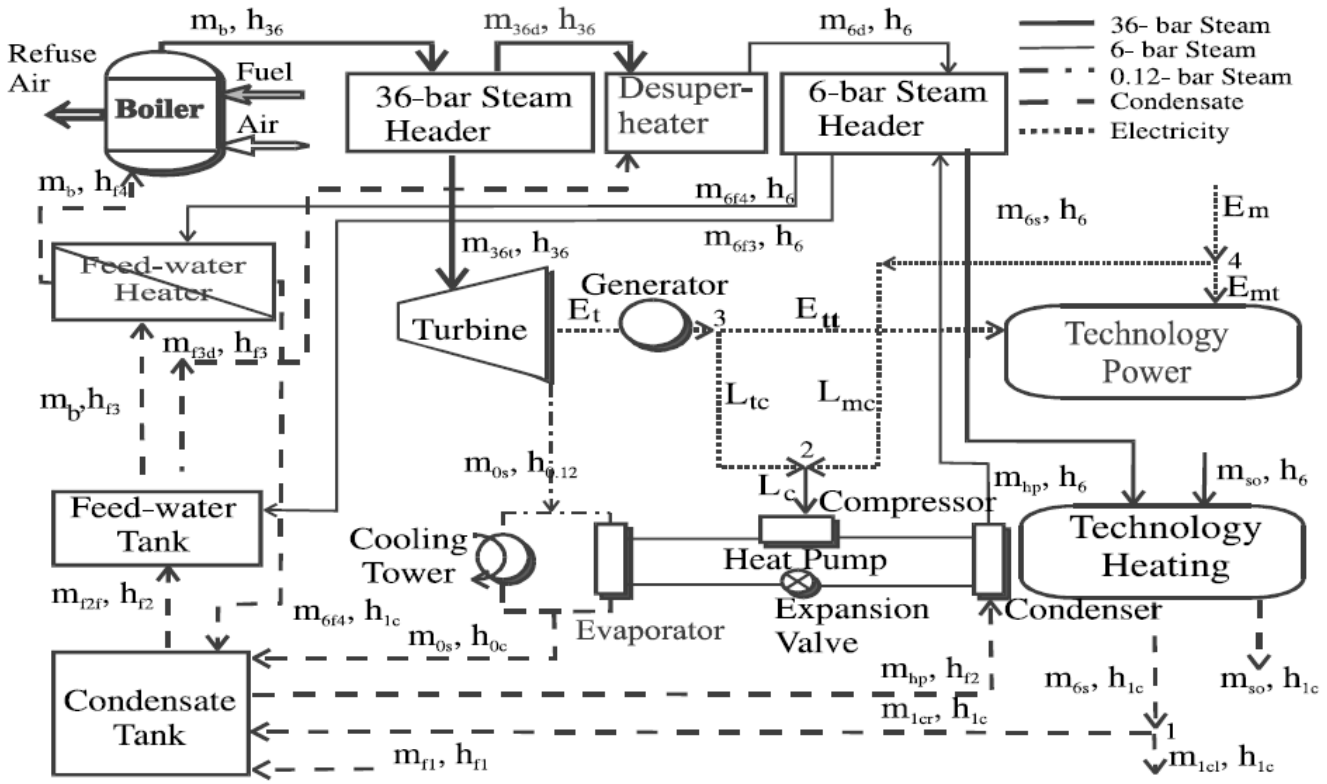


Source: https://www.sciencedirect.com/science/article/pii/S0196890401000280

Leveraging data-driven model in optimization of steam networks

Context

- Steam networks are widely used in the manufacturing environment
- Producers, consumers, electricity generators, letdown valves and cogeneration are some of the elements that need to work together to obtain the best network configuration.
- Equipment behavior can be approximated using data-driven models to account for main effects

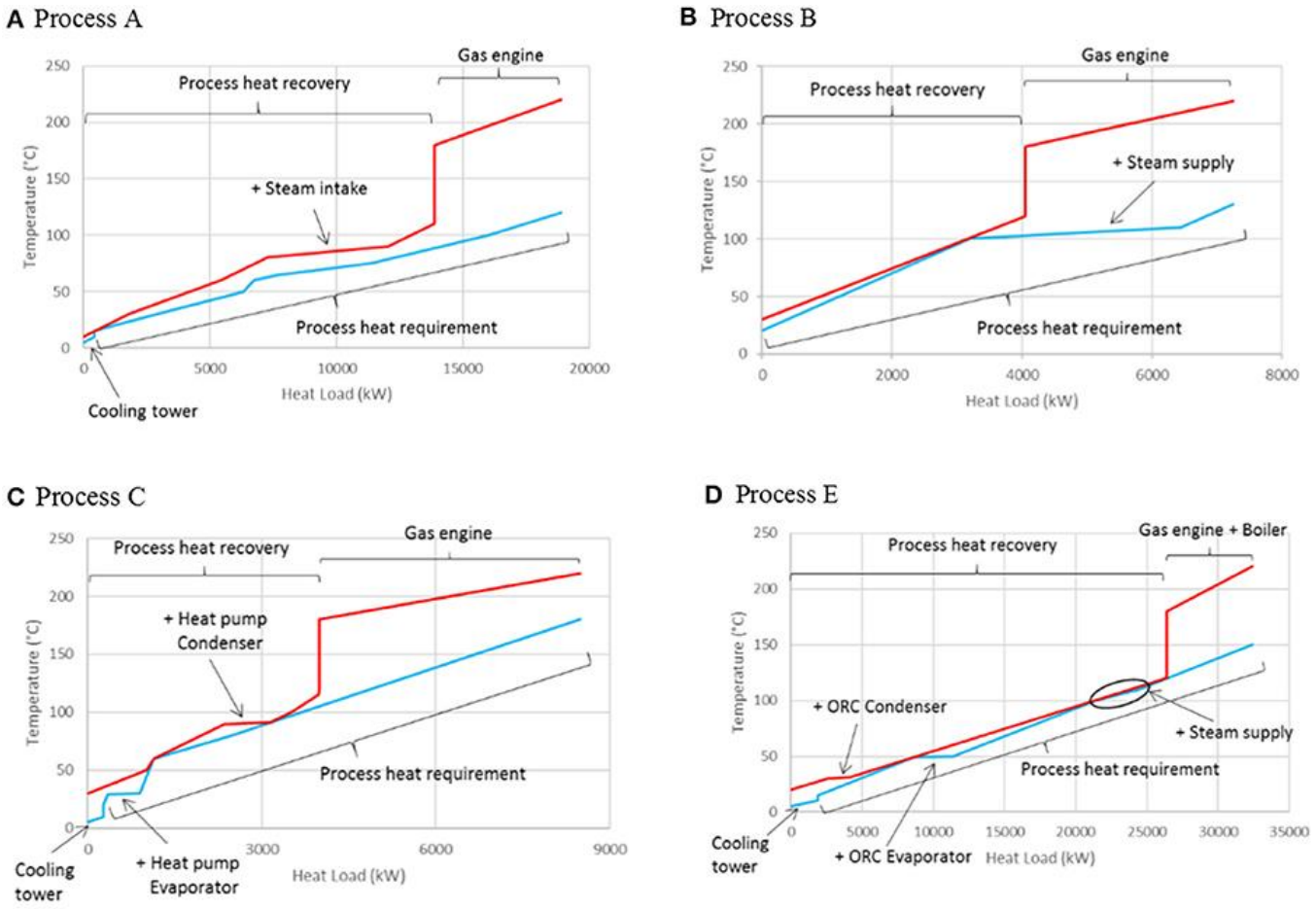


Source: https://www.sciencedirect.com/science/article/pii/S0196890401000280

Leveraging data-driven model in optimization of steam networks

What was done?

- Full description of objective and constrains was done to define an optimization problem as in the literature
- Plant gets recommendations to minimize cost an environmental impact



Source: https://www.frontiersin.org/articles/10.3389/fenrg.2020.00049/full

Troubleshooting in evaporator

Problem

Production losses in evaporators with chemical additives not helping

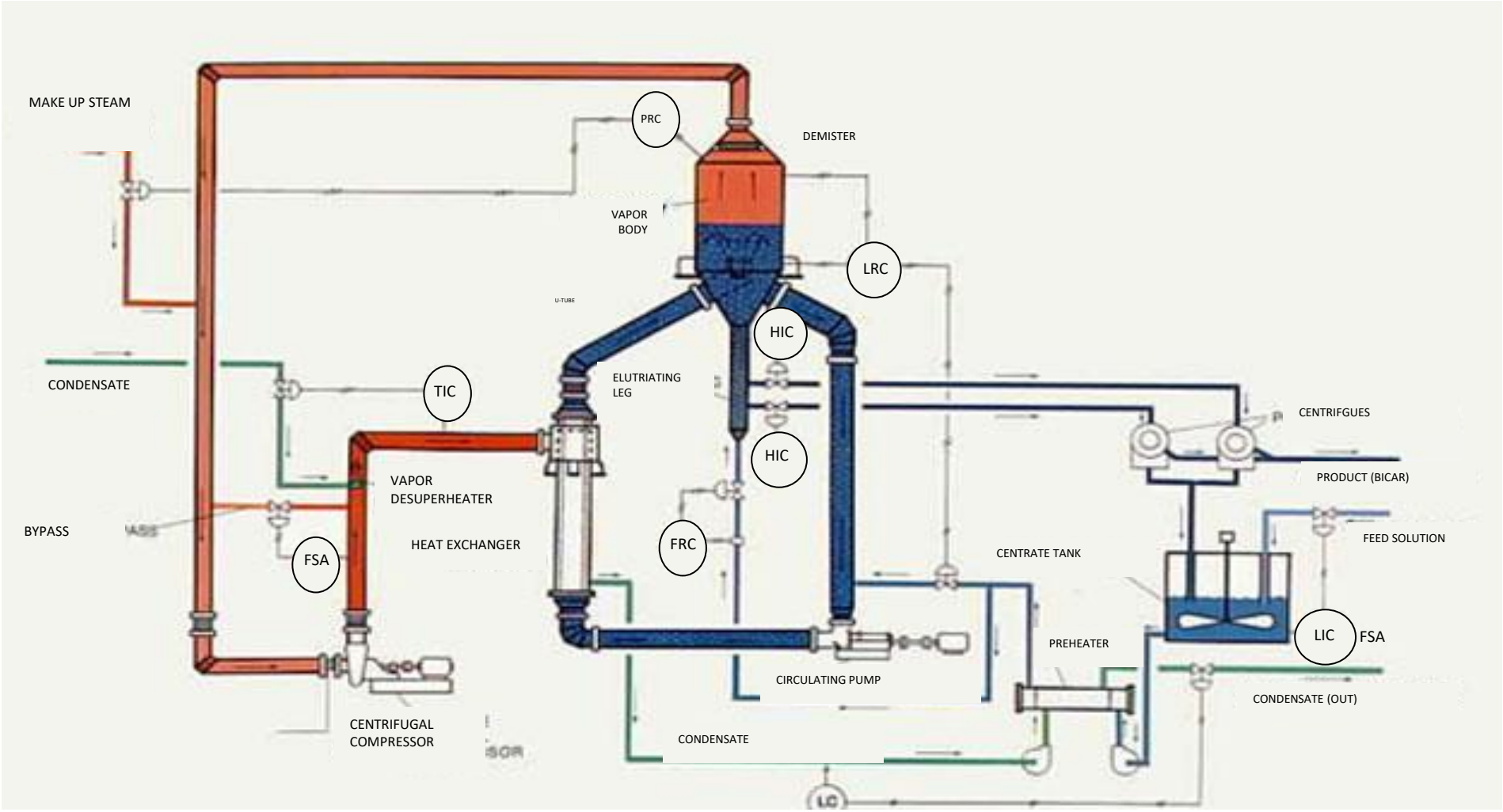
Recurring events, main KPI subject to operator’s opinion

Source:
https://www.whiting.ca/evaporators-whiting-equipment-canada-inc/
https://www.veoliawatertech.com/en/expertise/industries-we-serve/soda-ash

Troubleshooting in evaporator

Solution

- Screening found key contributors to process anomalies leading the elimination of events
- Control strategy was implemented to automate the addition of additives



Source: https://www.whiting.ca/evaporators-whiting-equipment-canada-inc/

Sargent Centre for Process Systems Engineering

Thank you!

Dr. Mattia Vallerio
Manufacturing Excellence Site Manager at Solvay and
Advanced Process Control Specialist [in]

Dr. Carlos Perez
Industrial Data Scientist at Solvay and
Optimization Specialist [in]

Dr. Francisco Navarro
Sr. Data Science Training Lead at IFF and
Visiting Researcher at Imperial College London [in]