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

Max Mowbray

max.mowbray@manchester.ac.uk

Department of Chemical Engineering

The University of Manchester

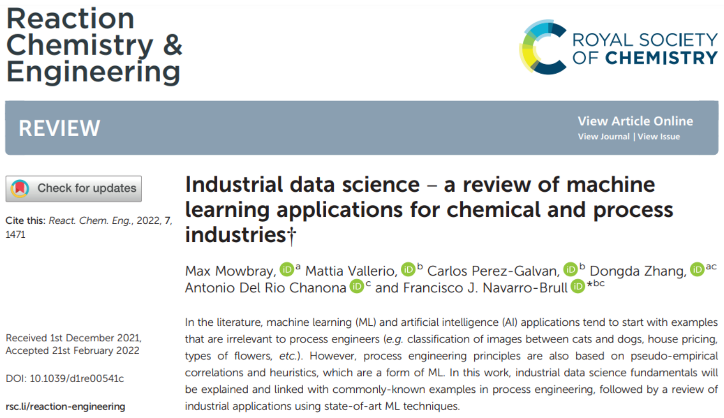
The University of Manchester

1. Overview

Overview

Work conducted collaboratively between The University of Manchester, Imperial College London and Solvay

- Review paper into major application and challenges of Machine Learning (ML) in process industries [1]


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

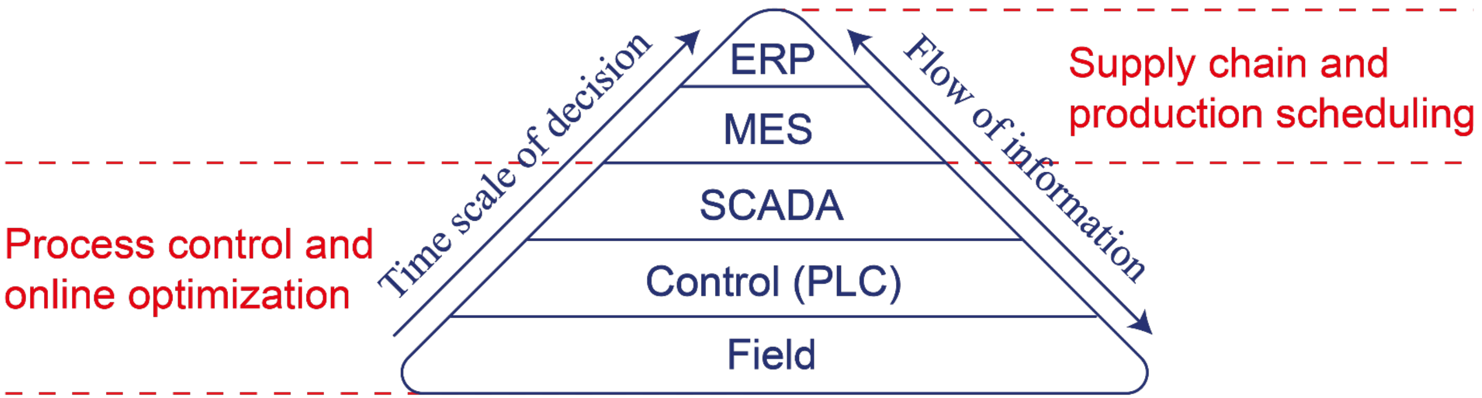
DOI: 10.1039/d1re00541c

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.

Considered contributions and challenges in application across the hierarchical control structure

- Focus on process level
- Paper provides discussion on upper-level decision functions

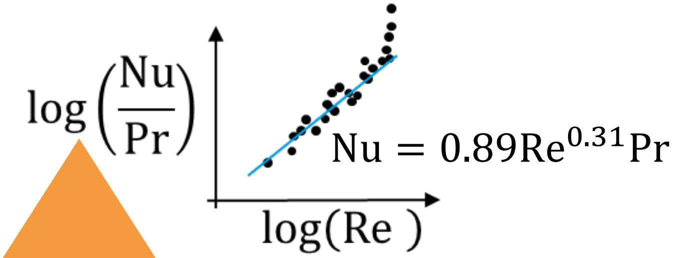


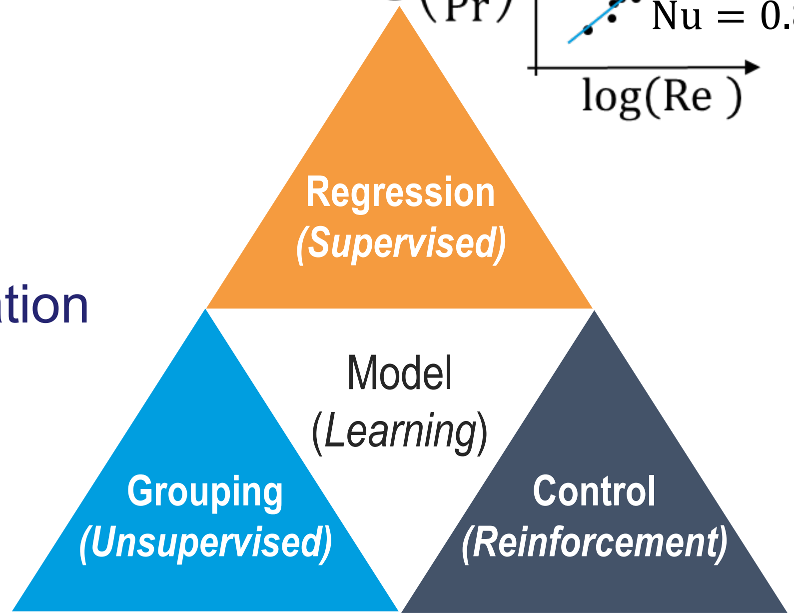
[1] Mowbray *et al.* (2022a)

Overview

Machine Learning has been widely applied within process systems, historically under a different disguise

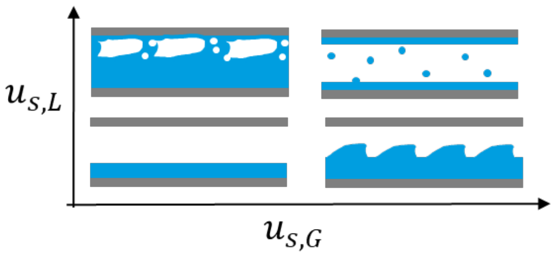
- Feature selection and engineering (dimensional analysis)
- Data pre-processing (signal processing) and partitioning
- Learning (statistical estimation and optimisation) and evaluation

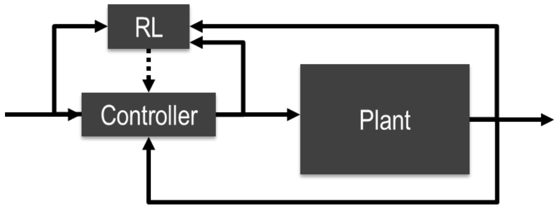




Process engineering cover all forms of analytics:

- Descriptive, diagnostic, predictive and prescriptive
- Increasingly flexible model classes, handling uncertainty and data visualisation techniques





Overview

Industrial data analytics and science work flow

Analytics

Not the business value

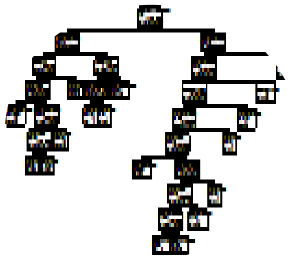
Problem definition
- Initial scope
- Process description

Process Data
- Data extraction
- Data quality (sensors)
- Data cleaning

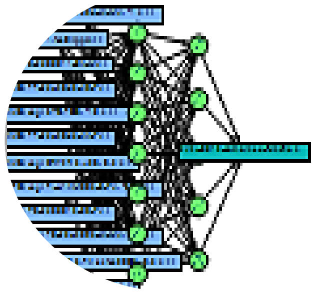
Process Modeling
- Visual analytics
- Process understanding
- Modeling and validation
- Optimization approach

Implementation + Industrialization
- Software sensor (quality prediction)
- Maintenance (anomaly detection)
- Process optimization (control)
- Planning and scheduling of production

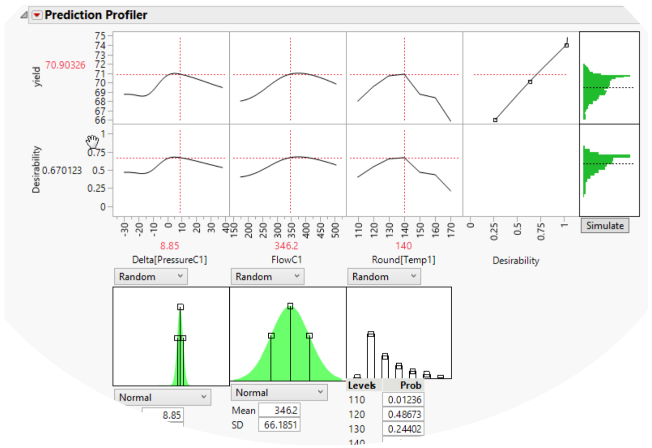
ML tools are key when knowledge is limited, but correlation is present



Decision tree model(s)



Neural Network



The University of Manchester

2. Case Study 1: Diagnosing yield variation in a distillation column

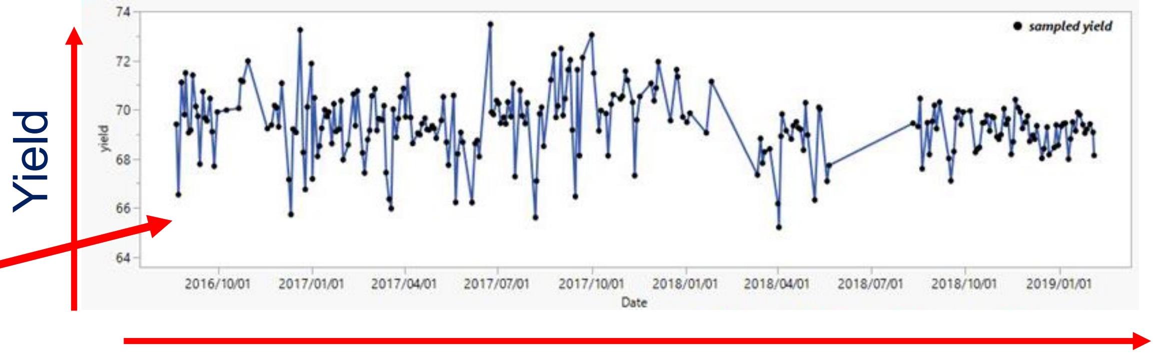
Diagnosing yield variation

Problem definition first consider trends in yield: variation driven by season or dynamics

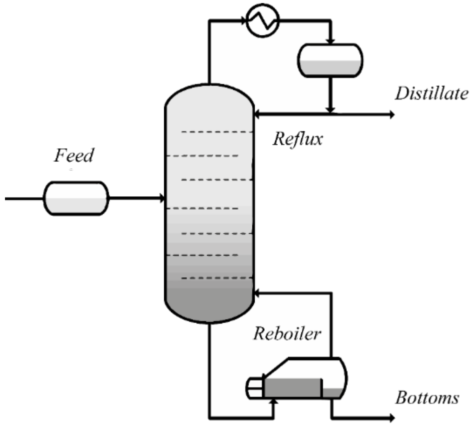
- No clear seasonal trend

High frequency,

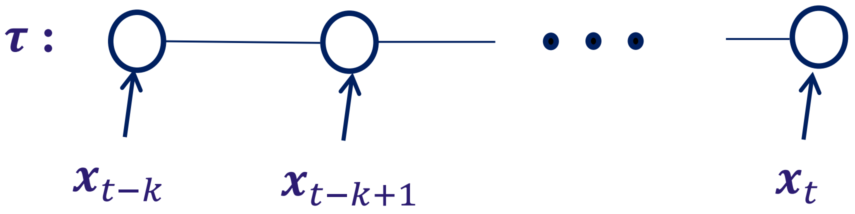
noisy variation

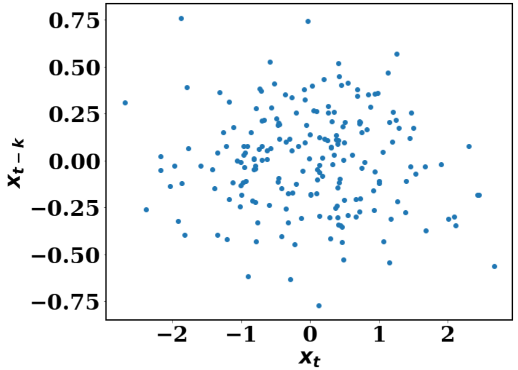


Time (months)

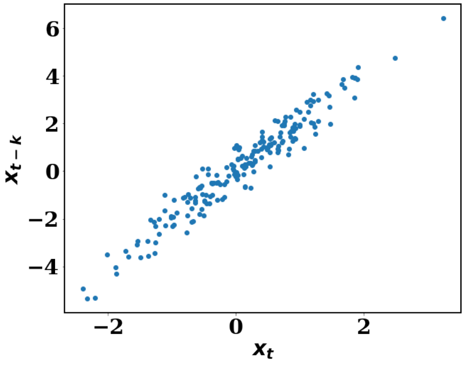


- Existing data collected at supposed steady state





History uncorrelated



History correlated

Diagnosing yield variation

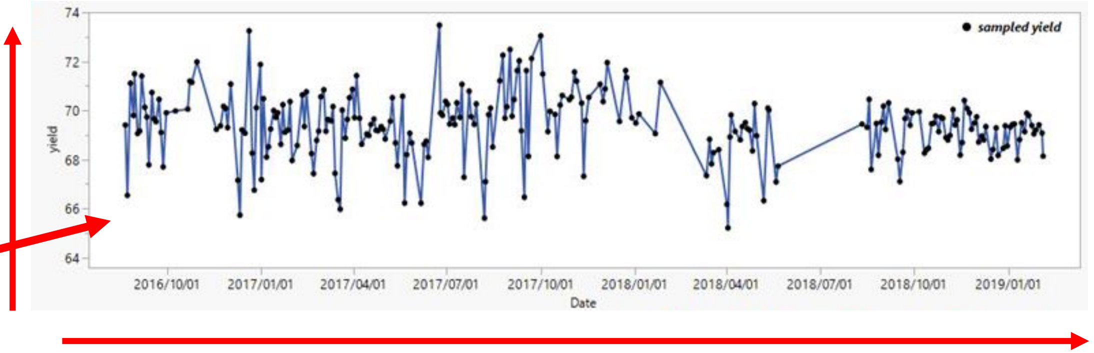
MANCHESTER 1824
The University of Manchester

Imperial College London

Problem definition first consider trends in yield: variation driven by season or dynamics

- No clear seasonal trend

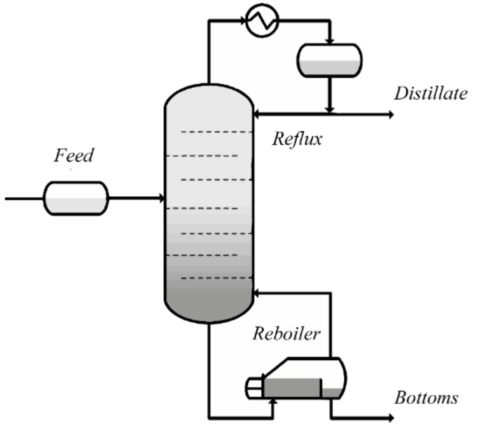
Yield



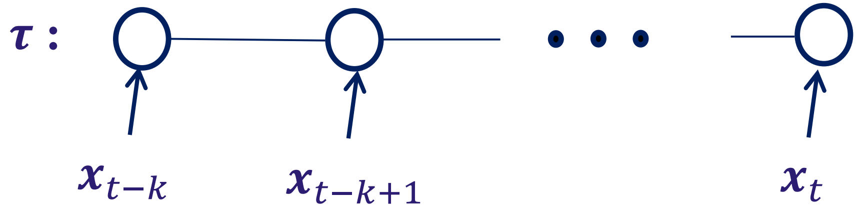
High frequency,

noisy variation

Time (months)

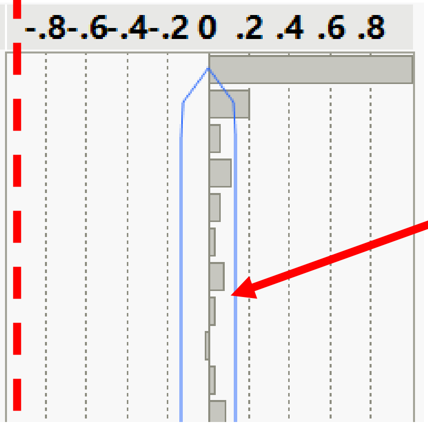


- Existing data collected at supposed steady state



Autocorrelation

Lag	AutoCorr
0	1.0000
1	0.1971
2	0.0593
3	0.1152
4	0.0558
5	0.0305
6	0.0732
7	0.0302
8	-0.0093
9	0.0355
10	0.0832



Dynamics not clearly driving variation

$c << 1$ history not correlated (IID assumption)

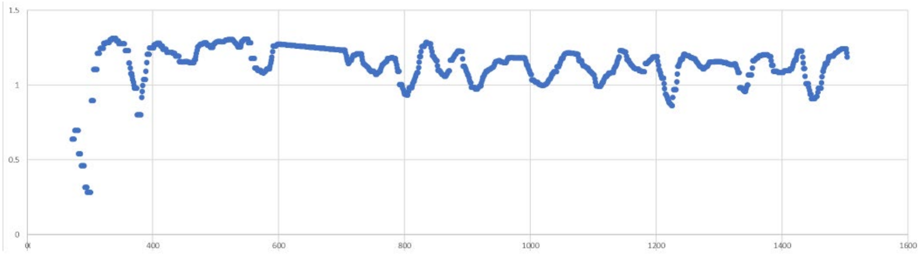
MANCHESTER 1824
The University of Manchester

Imperial College London

Diagnosing yield variation

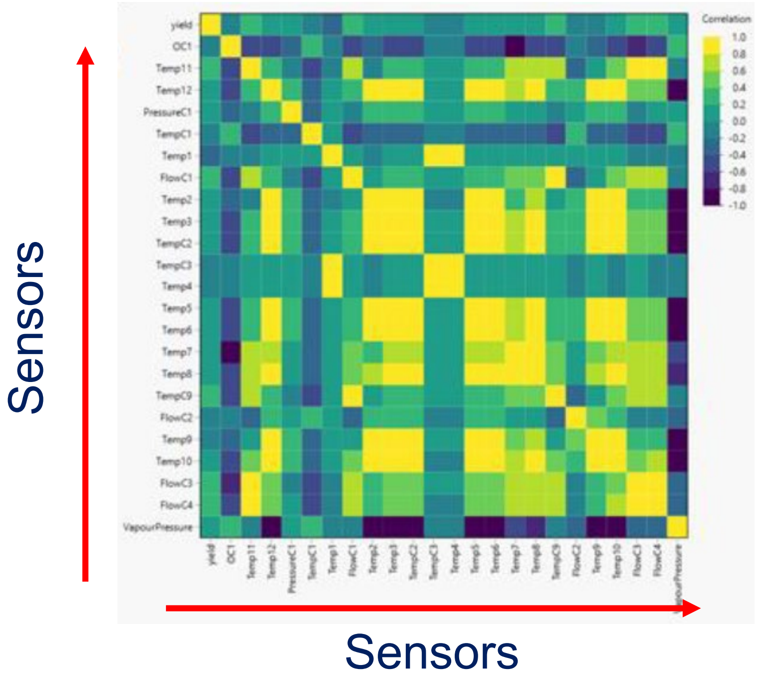
Data pre-processing; first understand the data at hand

- Data is high dimensional sensor measurements
  - Quantify correlation between two sensors’ data



Data visualisation powerful to quickly understand the data

- Pearson’s correlation coefficient
- 2D visualisation of very high dimensional data
  - The more yellow the pixel the higher the correlation
- Mutlicollinear data causes issues in model construction
  - Dimensionality reduction



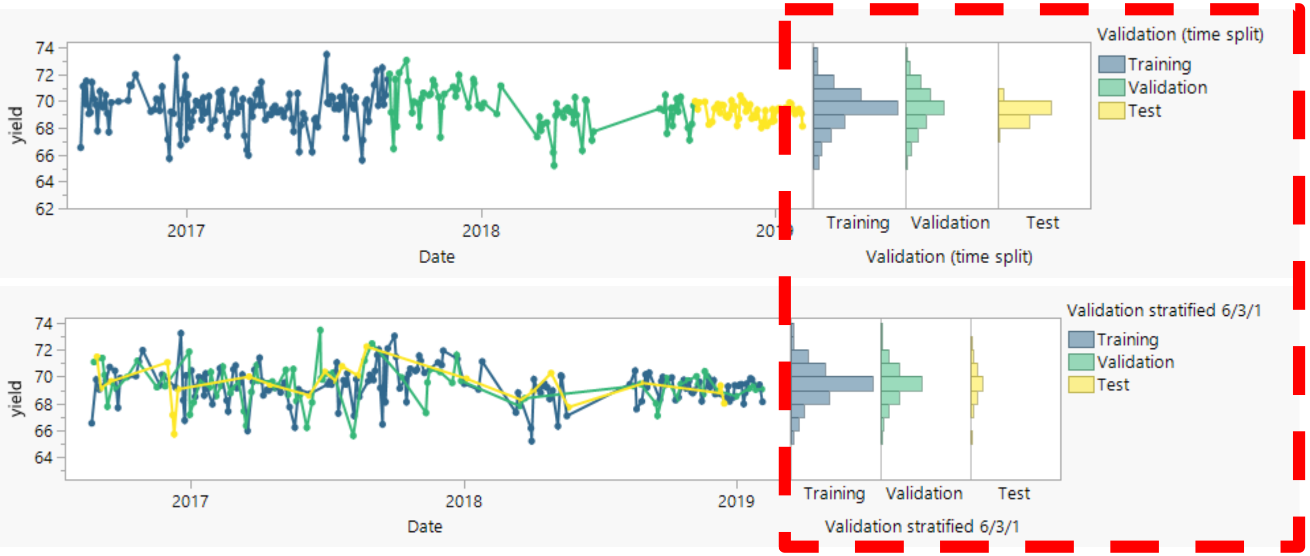
MANCHESTER 1824
The University of Manchester

Diagnosing yield variation

Imperial College London

Data pre-processing: data partitioning strategy is key for ensuring adequate validation of model

- Training
- Validation
- Test



- Subsampling time-series considers the data to I.I.D. (dynamics are not driving variability, rarely the case!)
- K-fold cross validation designs can ensure against mismatch in data distributions

Data processing: k-fold cross validation

- K-fold cross validation:

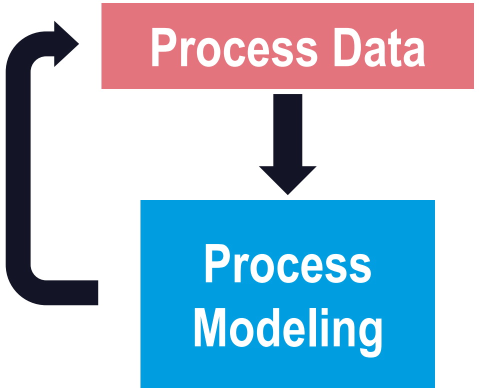
D 1 | D 2 | D 3 | D 4 | D 5 | D 6

Training Set

Validation Set

Diagnosing yield variation

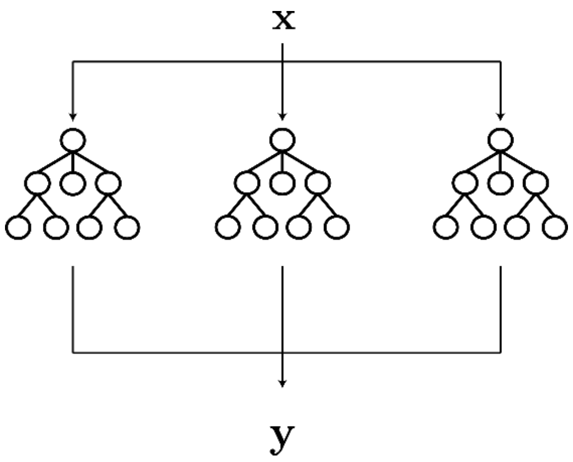
Generate knowledge by modelling; screen the data to select variables



Identifying significant variables may be an iterative process:

1. Leverage feature importance techniques with model structure cross validation…

Random forests: robust and interpretable models



Knowledge generation

Uniform noise

Column Contributions

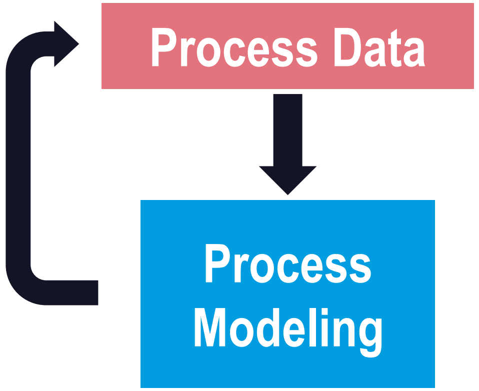
Term	Number of Splits	SS	Portion
FlowC1	29	17.1678444	0.1988
TempC9	27	12.2012469	0.1413
Temp1	13	9.75679814	0.1130
Temp4	18	5.57623969	0.0646
TempC3	13	4.64860694	0.0538
OC1	15	3.66068252	0.0424
FlowC2	22	3.53264433	0.0409
PressureC1	9	2.42725587	0.0281
TempC1	14	2.29403566	0.0266
Temp6	9	2.23790975	0.0259
Shuffle[yield]	14	2.17379588	0.0252
Random Normal Noise	11	1.95953512	0.0227
VapourPressure	10	1.94971732	0.0226
FlowC3	11	1.89361028	0.0219
Temp12	12	1.889938	0.0219
Temp3	11	1.85201822	0.0214
Temp9	11	1.82578396	0.0211
FlowC4	8	1.76030855	0.0204
Temp2	8	1.52285823	0.0176
Temp10	12	1.44849998	0.0168
Temp5	7	1.12753413	0.0131
Temp7	8	0.91416155	0.0106
Temp11	7	0.80503604	0.0093
Random Uniform Noise	6	0.75906398	0.0088
Temp8	11	0.69334372	0.0080
TempC2	8	0.28127547	0.0033

Variables (temp., flow, pressure) significant to prediction

Importance

Diagnosing yield variation

Generate knowledge by modelling; screen the data to select variables



Identifying significant variables may be an iterative process:

2. Repeat and analyse new model

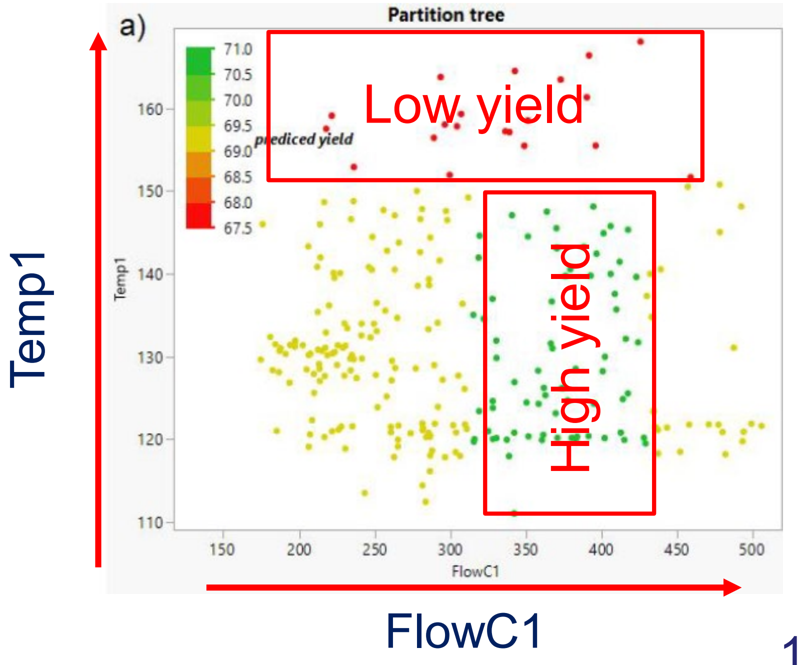
Updated random forest model: feature importance

Column Contributions

Term	Number of Splits	SS	Portion
FlowC1	2	123.846558	0.5811
Temp1	1	89.2786595	0.4189
Shuffle[yield]	0	0	0.0000
OC1	0	0	0.0000
PressureC1	0	0	0.0000
Random Uniform Noise	0	0	0.0000
TempC9	0	0	0.0000
Temp6	0	0	0.0000
Random Normal Noise	0	0	0.0000
FlowC2	0	0	0.0000
TempC1	0	0	0.0000
TempC3	0	0	0.0000
Temp4	0	0	0.0000

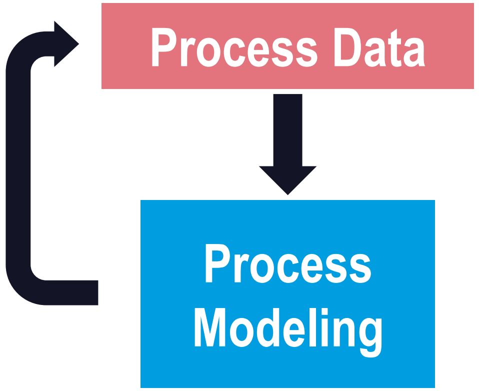
Flow and Temp. important

Knowledge generation



Diagnosing yield variation

Generate knowledge by modelling; screen the data to select variables



Identifying significant variables may be an iterative process:

2. Repeat and analyse new model

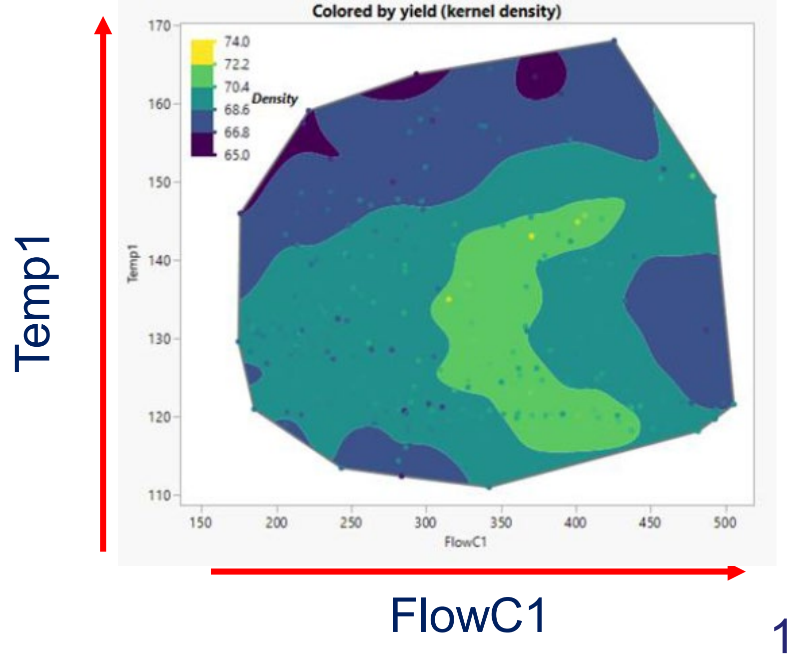
Updated random forest model: feature importance

Column Contributions

Term	Number of Splits	SS	Portion
FlowC1	2	123.846558	0.5811
Temp1	1	89.2786595	0.4189
Shuffle[yield]	0	0	0.0000
OC1	0	0	0.0000
PressureC1	0	0	0.0000
Random Uniform Noise	0	0	0.0000
TempC9	0	0	0.0000
Temp6	0	0	0.0000
Random Normal Noise	0	0	0.0000
FlowC2	0	0	0.0000
TempC1	0	0	0.0000
TempC3	0	0	0.0000
Temp4	0	0	0.0000

Flow and Temp. important

Knowledge generation



Diagnosing yield variation

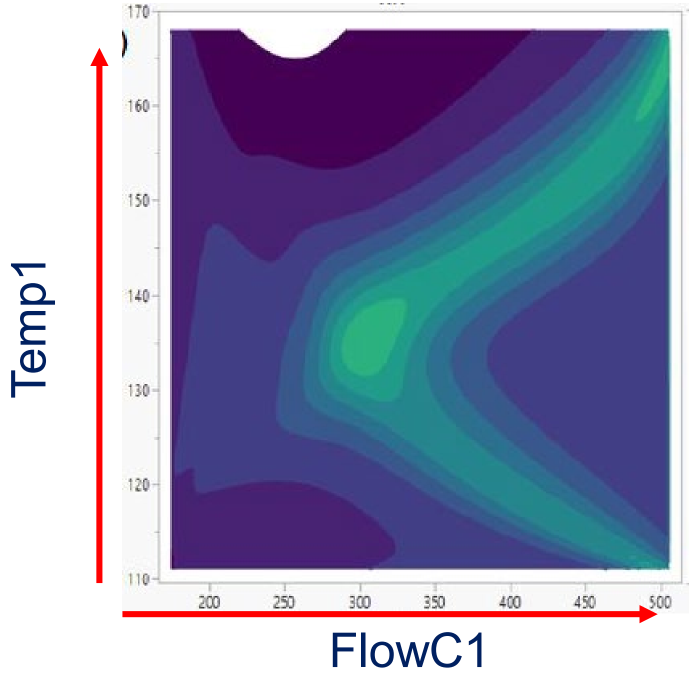
Random forests are interpretable and good at screening, but what about predictive accuracy?

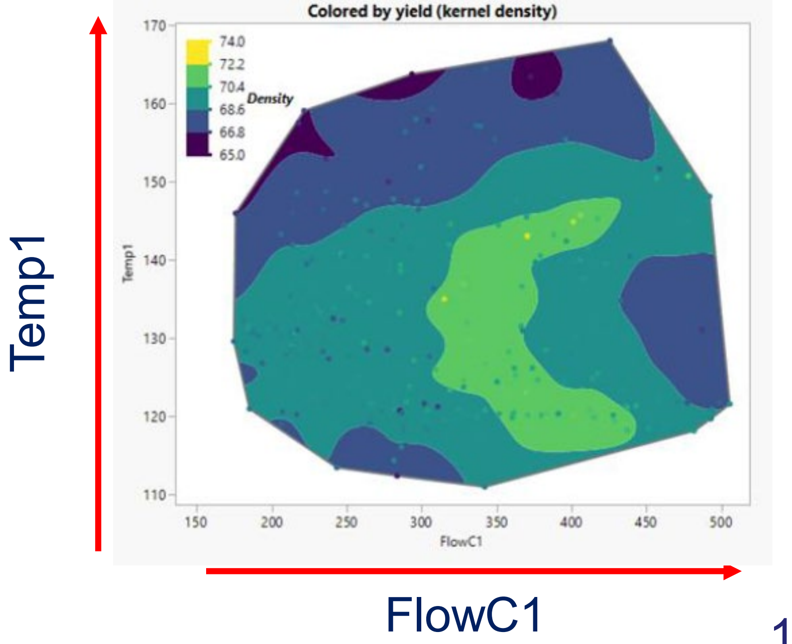
Process Modeling

- Having identified important sensors, we can search over model classes and structures

Manual, random or automated search

- Neural networks are a go-to
- Improved generalisation accuracy
- Smooth relationship identified





MANCHESTER 1824
The University of Manchester

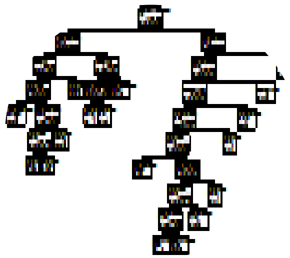
Summary: Case Study 1

Imperial College London

Yield variation: diagnostic summary

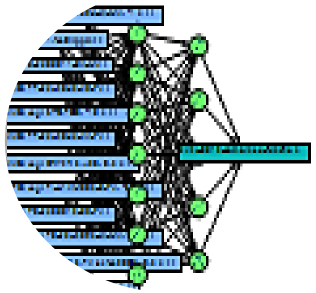
- Two step approach:

Important process variables



Decision tree model(s)

[JMP Add ins]



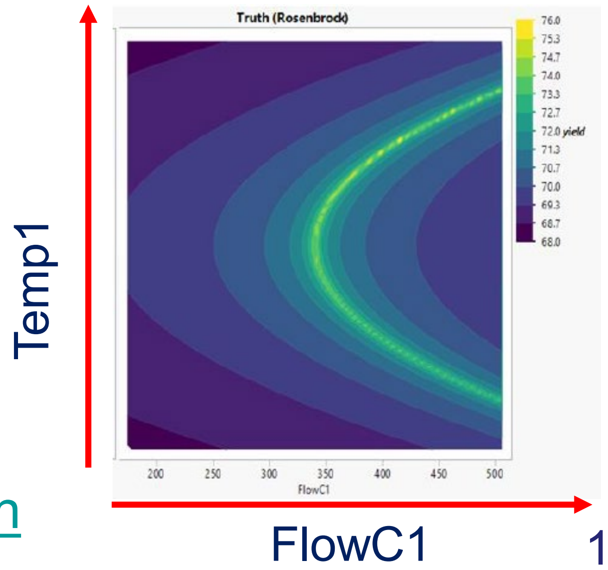
Neural Network

Flexible function approximation

Synthetic dataset with known ground truth and assumed steady state

- Conceptual demonstration of correlation analysis and screening
- Future work to identify dynamics
- **Real world data key to further exploration of ML**

Implementation + Industrialization



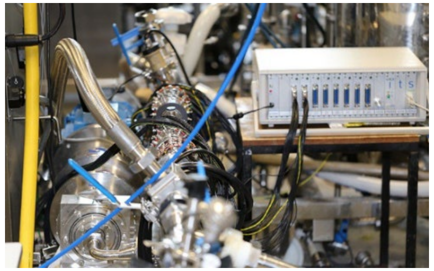
[Application 1: demonstration]

3. Case Study 2: Data-driven soft-sensing of product quality in batch processing

Soft sensing

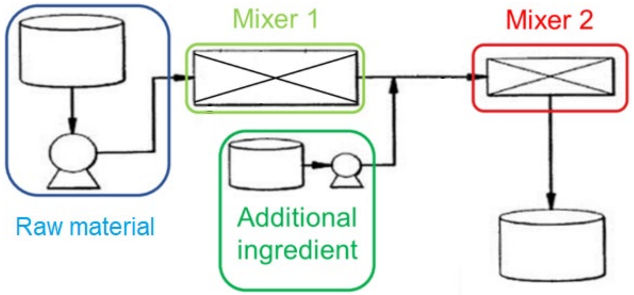
Predictive analytics for estimating viscosity in a batch process handling non-Newtonian fluid [4]

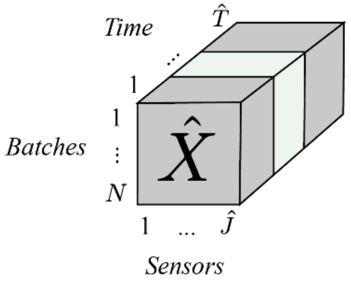
- Real-time PAT is expensive and difficult to retrofit
- Offline measurement slow and has standard error



Emulsification process (batch processes) for personal care product manufacturing

- 28 online sensors (temperature, pressure, flowrate)
- Sensor data recorded once per second (in total 2 hours)
- Available datasets: 30 batches × 28 sensors × 7673 time steps





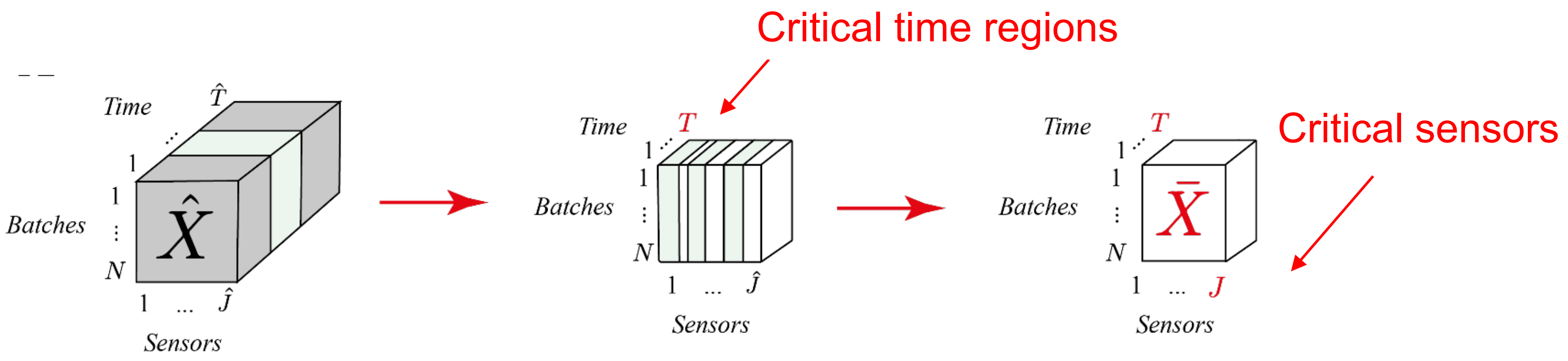
[4] Mowbray *et al.* (2022b)

Soft sensing

The University of Manchester

CAFE4DM THE CENTRE IN ADVANCED FLUID ENGINEERING FOR DIGITAL MANUFACTURING

Challenge 1: Identifying the spatio-temporal trajectories that influence product quality



Projection to latent structures (PLS)

$$z_1 = \overbrace{\theta_{1,1} \cdot x_1 + \theta_{1,2} \cdot x_2 + \cdots + \theta_{1,n}}^{\text{Importance of each variable}} \cdot x_n$$

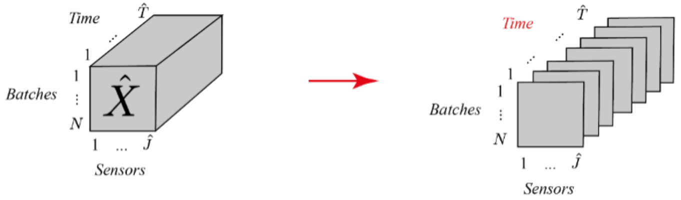
$$y = a_1 \cdot z_1 + e$$

- Screen data to find correlations relevant to prediction

Solution: Loadings analysis of PLS models [2]

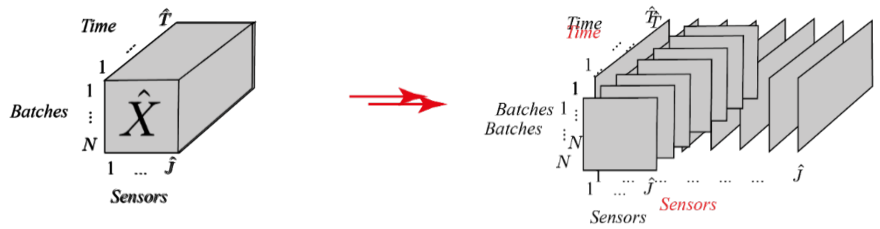
- Construct a PLS model for each sensor in time, $\theta^j \in \mathbb{R}^{n_z \times \hat{T}}$

- Construct a PLS model for the sensors at each timestep, $\theta^t \in \mathbb{R}^{n_z \times \hat{J}}$

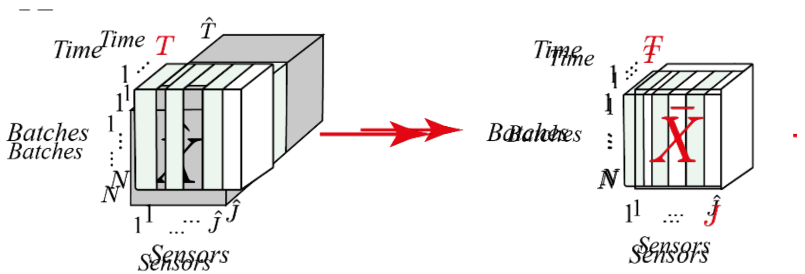


Soft sensing

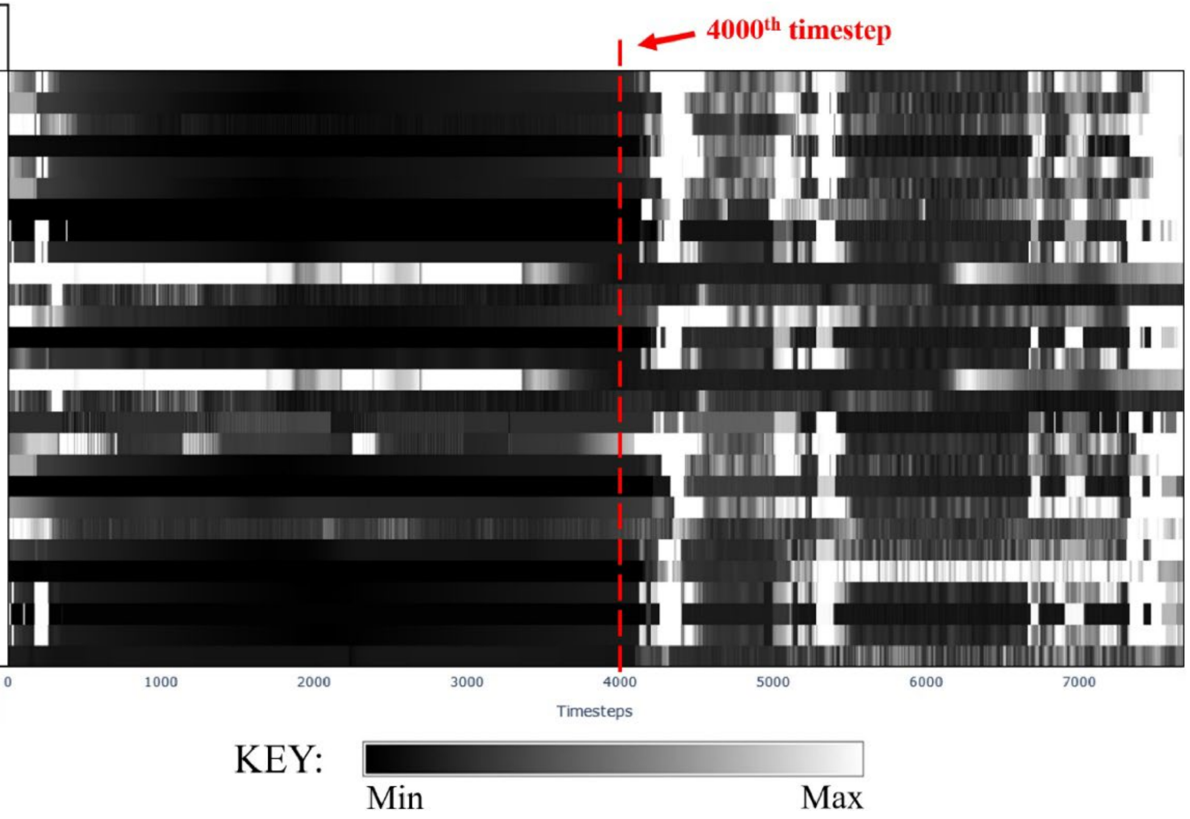
Challenge 1: Identifying the spatio-temporal trajectories that influence viscosity



- Analyse the loadings of the PLS models
- Heat maps are interpretable



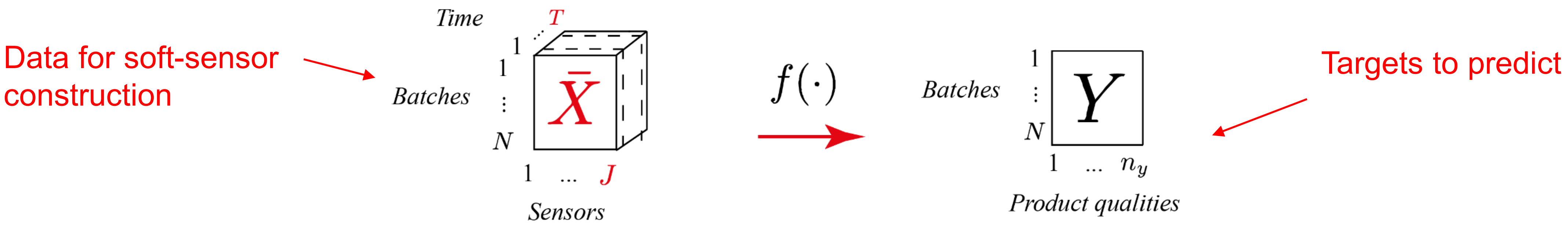
Sensor	Components needed for MAPE <0.5%
1	18
2	17
3	18
4	16
5	18
6	17
7	19
8	16
9	19
10	27
11	13
12	29
13	17
14	18
15	27
16	13
17	21
18	21
19	17
20	23
21	20
22	17
23	17
24	17
25	18
26	19
27	18
28	13



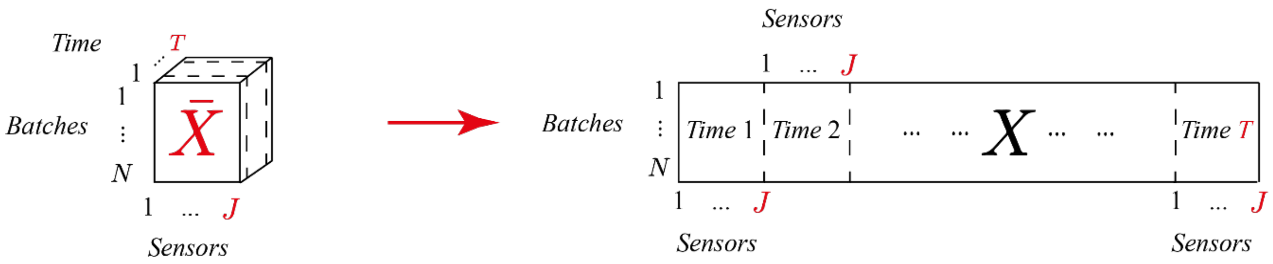
Critical time region: 5-15 minute period from 2 hours ; Critical sensors: 8 from 28 sensors

Soft sensing

Challenge 2: Feature extraction and dimensionality reduction



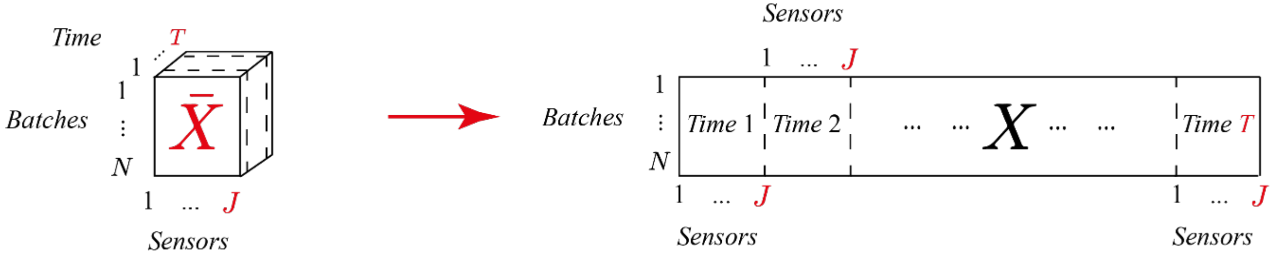
Partial Solution: Multiway methods



- Representation that allows for extraction of important information

Soft sensing

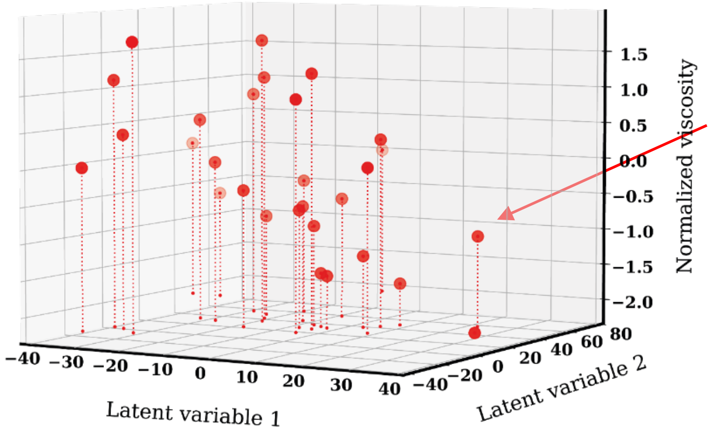
Challenge 2: Feature extraction and dimensionality reduction



Solution: Feature extraction using multiway PLS or autoencoders*

- Reduced dimension representation of spatio-temporal trajectories correlated to end-product quality
- Can now identify a map to product quality, $f: \mathbb{R}^{n_z} \rightarrow \mathbb{R}$

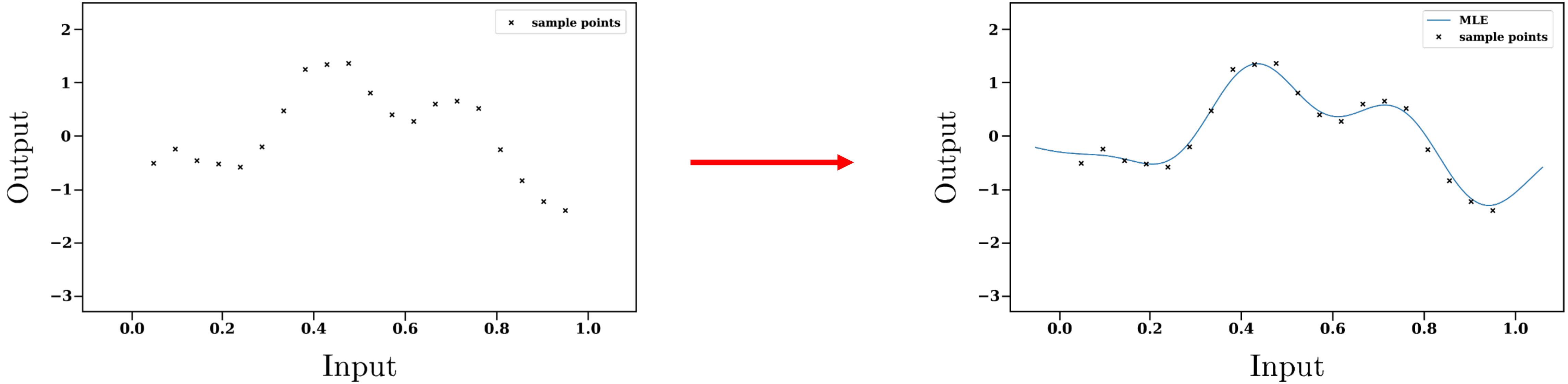
Number of latent variables enters model structure selection problem



Extracted features expressed in linear subspace

Soft sensing

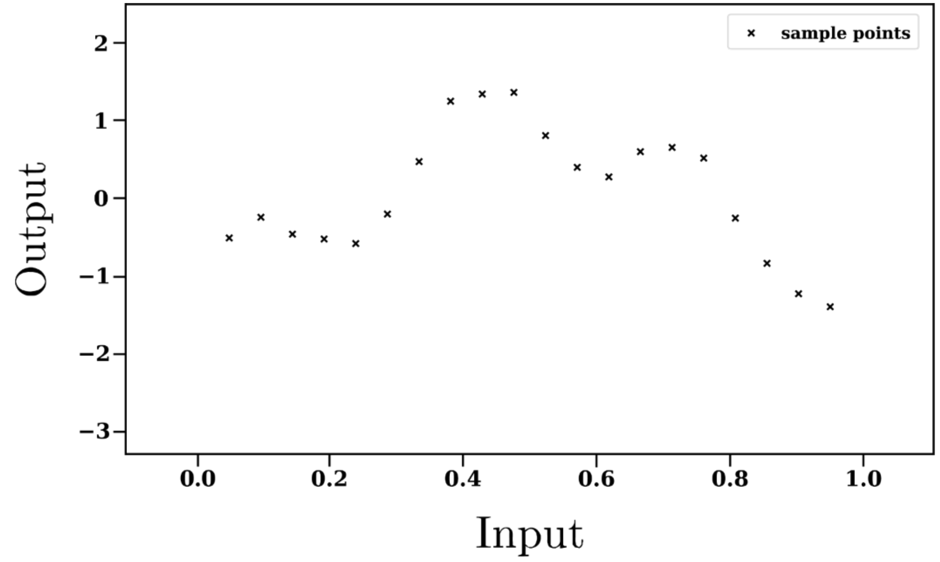
Challenge 3: Expressing nonlinearity and prediction uncertainty

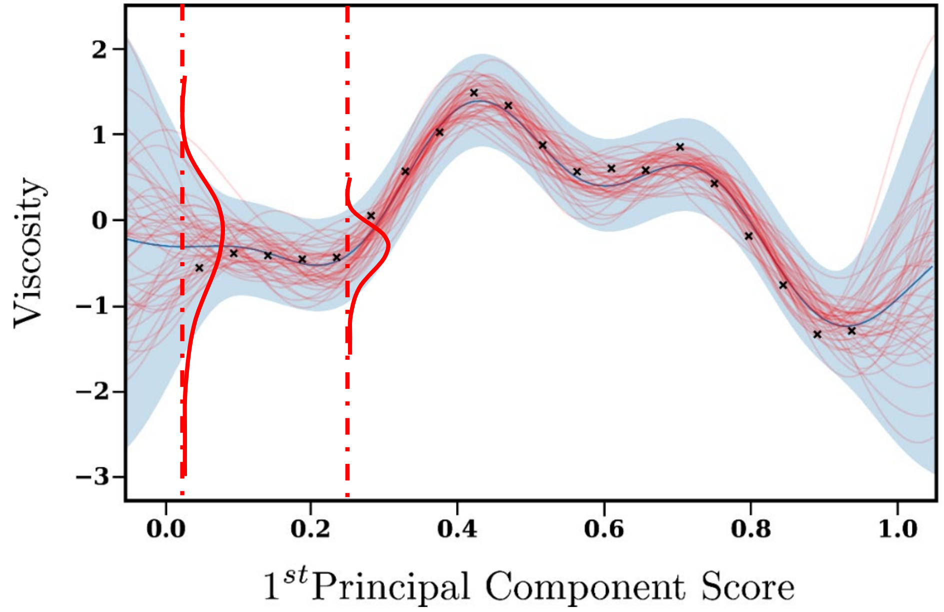


- Classical regression practice identifies deterministic models
- Our data has moderate noise in the measurement

Soft sensing

Challenge 3: Expressing nonlinearity and prediction uncertainty





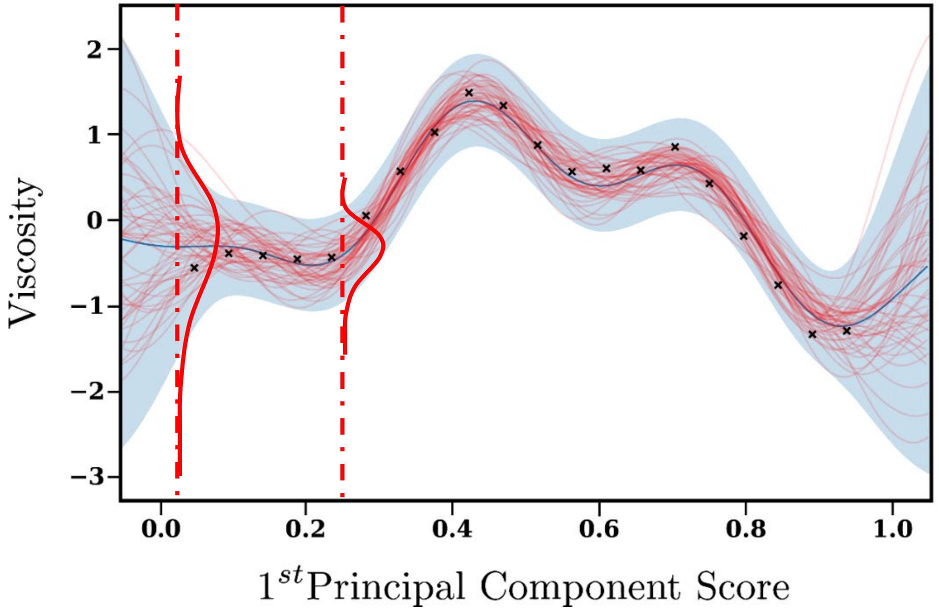
- Classical regression practice identifies deterministic models
- Our data has moderate noise in the measurement

Soft sensing

Challenge 3: Expressing nonlinearity and prediction uncertainty

Gaussian Processes (GPs):

- $f_{GP}(\boldsymbol{z}) \sim GP\left(m(\boldsymbol{z}), k(\boldsymbol{z}, \boldsymbol{z}')\right)$
- Exploit statistical relationships in data
  - Bayes' Rule: $y_i \sim p(y_i | \boldsymbol{z}_i{}^*, \mathcal{D}) = \mathcal{N}(\bar{\mu}, \Sigma)$
- Uncertainty reflects data variation and lack of information



Soft sensing

Model structure selection via cross-validation and subsequent model testing

Dataset	Season	Use	Batches
A	1	Train	30
B	2	Test	16

1. Cross validation on Dataset A to identify model structure for each class

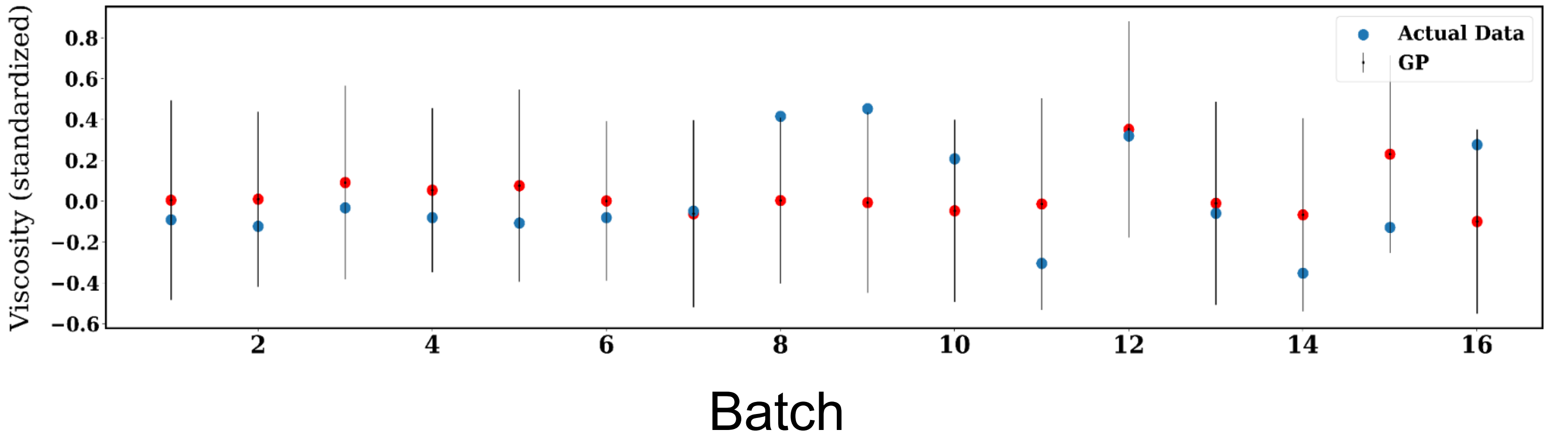
→

2. Test on Dataset B to evaluate prediction **same variant** but **different season**

Cross-validation and test results

- GP performs well in validation with average error of 10%
- Uncertainty estimate covers residual – indicating it is reliable

Prediction plots for GP on dataset B



Summary: Case Study 2

Data visualisation is an effective step to analyse historical datasets

- Screen critical time region and sensors (***knowledge informed*** dimensionality reduction)

Probabilistic machine learning methods are excellent for soft-sensor design

- A high accuracy soft-sensor provides avenue to monitor
- Reliable uncertainty estimates to guide process engineers

Potential for industrial application

- Fast prediction online of critical product quality vs slow offline measurement
- Investigating methodologies to transfer soft-sensors between processes.

4. Conclusions

Conclusion

Here we present an intuitive data focused framework for problem solving

- Problem definition; data processing; modelling; implementation

Machine Learning tools can be used for descriptive, diagnostic and predictive analysis

- Correlation analysis for quick screening of tags (sensors)
- Mapping identification for steady state behaviour as well as spatio-temporal trajectories to final product qualities
- Future work will consider methodology for identification of dynamic behaviour

The litmus test of Machine Learning is practical implementation to real processes and data

References

[1] Mowbray, M., Vallerio, M., Perez-Galvan, C., Zhang, D., Chanona, A. D. R., & Navarro-Brull, F. J. (2022a). Industrial data science–a review of machine learning applications for chemical and process industries. Reaction Chemistry & Engineering.

[2] Beck, D. A., Carothers, J. M., Subramanian, V. R., & Pfaendtner, J. (2016). Data science: Accelerating innovation and discovery in chemical engineering. *AIChE Journal*, *62*(5), 1402-1416.

[3] Shang, C., & You, F. (2019). Data analytics and machine learning for smart process manufacturing: recent advances and perspectives in the big data era. Engineering, 5(6), 1010-1016.

[4] Mowbray, M., Kay, H., Kay, S., Caetano, P. C., Hicks, A., Mendoza, C., ... & Zhang, D. (2022b). Probabilistic machine learning based soft-sensors for product quality prediction in batch processes. Chemometrics and Intelligent Laboratory Systems, 228, 104616.

Thank you for listening!


@MaxMowbray3


@DongdaZhang


@AntonioE89


Phil Martin


Francisco Navarro-Brull