Pharma Manufacturing When Complexity Cannot be Linearly Managed or Solved via a Singular AI Model

by Toni
Manzano

Xavier AI Experts Network August 6, 2021

XAVIER HEALTH

Inspiring collaboration. Leading innovation. Making a difference.

Agenda

XAVIER HEALTH

- Introductions, Mission and Network Co-leaders: 5 minutes
- Insert title of presentation: 30 minutes
- Discussion and future topic selection: 25 minutes

XAVIER HEALTH

Manufacturing Science

The body of knowledge available for a specific product and process, including critical-to-quality product attributes and process parameters, process capability, manufacturing and process control technologies and quality systems infrastructure.

(Source: PhRMA Quality Technical Committee, 2003)

XAVIER HEALTH

PAT

(...) The applicant should demonstrate an enhanced knowledge of product performance over a range of material attributes, manufacturing process options and process parameters

(...) Real-time quality control, leading to a reduction of end-product release testing

(...) A monitoring program (e.g., full product testing at regular intervals) for verifying **multivariate prediction** models

(Source: ICH Q8 step 4, 2009)

Control Systems (by AI)

- Controllability: In order to be able to do whatever we want with the given dynamic system under control input, the system must be controllable.
- Observability: In order to see what is going on inside the system under observation, the system must be observable.

(Source: R. Kalman 1960)

CPV of the Future project

A collaboration between PDA, PQRI and AI Xavier

CPV of the Future

XAVIER HEALTH

A project supported by:

PQRI Product Quality Research Institute

PDA Parenteral Drug Association

Team

Agnès Hardy-Boyer *(CPV Team lead)*	Sanofi Pasteur	Head Global Validation CoE Corporate Quality
Mauro Giusti (PV Team lead)	Eli Lilly	Director, Technical Services/Mfg Sciences
Toni Manzano *(Execution lead)*	Aizon	CSO and Co-founder
Antonio Moreira *(Execution co-lead*	University of Maryland, Baltimore County	Vice Provost for Academic Affairs
Francisco Valero	Universidad Autònoma of Barcelona	Professor and head of department of BioChemical Engineering
Nilanjan Banerjee	University of Maryland, Baltimore County	Professor, Computer Science and Electrical Engineering
David Hubmayr	CSL Behring	Manager, Process Development & Breakthrough Technologies, R&D
Christophe Agut	Sanofi Pasteur	Head of Process Validation and Statistics Expertise
Mario Stassen (Regulatory expert)	AI Xavier University (AI in Operations Team)	BioPharma Regulatory expert
Matt Schmucki	AstraZeneca	Lean Coach and CPV Expert
Shereya Maiti	Bayer Pharmaceuticals	Senior Scientist
Joeri Van Wijngaarden	Aizon	Innovation Lab R&D Engineer and data scientist
Catarina Leitão	4Tune Engineering	CPV Expert

1. Demonstrate that AI is a good tool for CPV (Stage 3) during the early phase of development and define more robust process control and validation strategies

2. At short term, the biopharma manufacturing could add control parameters and new tools oriented to ensure the expected product quality.

3. At long term, the knowledge about the complex interactions based on AI leads to improve the scale-up and tech-transfer operations.

CPV of the Future

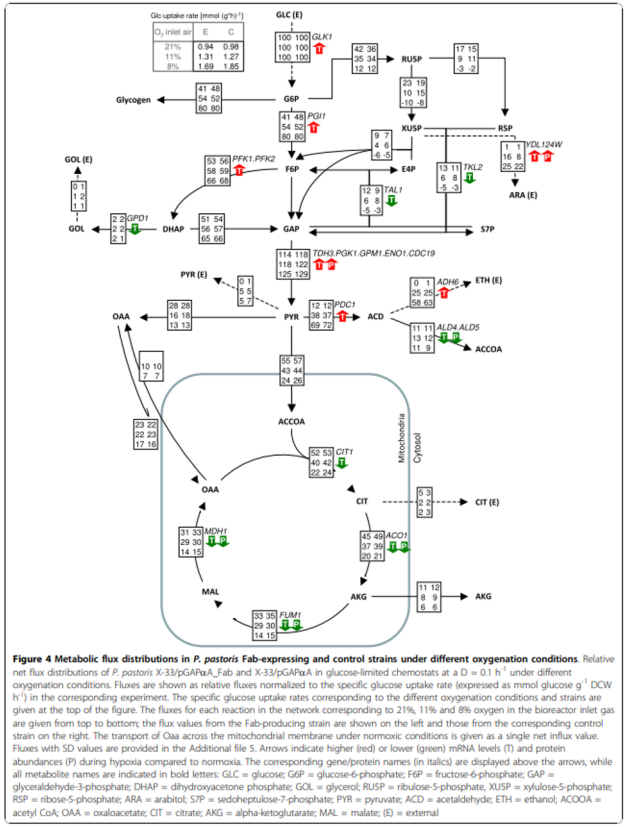


Figure 4 Metabolic flux distributions in *P. pastoris* Fab-expressing and control strains under different oxygenation conditions. Relative net flux distributions of *P. pastoris* X-33/pGAPαA_Fab and X-33/pGAPαA in glucose-limited chemostats at a D = 0.1 h^{-1} under different oxygenation conditions. Fluxes are shown as relative fluxes normalized to the specific glucose uptake rate (expressed as mmol glucose g^{-1} DCW h^{-1}) in the corresponding experiment. The specific glucose uptake rates corresponding to the different oxygenation conditions and strains are given at the top of the figure. The fluxes for each reaction in the network corresponding to 21%, 11% and 8% oxygen in the bioreactor inlet gas are given from top to bottom; the flux values from the Fab-producing strain are shown on the left and those from the corresponding control strain on the right. The transport of Oaa across the mitochondrial membrane under normoxic conditions is given as a single net influx value. Fluxes with SD values are provided in the Additional file 5. Arrows indicate higher (red) or lower (green) mRNA levels (T) and protein abundances (P) during hypoxia compared to normoxia. The corresponding gene/protein names (in italics) are displayed above the arrows, while all metabolite names are indicated in bold letters: GLC = glucose; G6P = glucose-6-phosphate; F6P = fructose-6-phosphate; GAP = glyceraldehyde-3-phosphate; DHAP = dihydroxyacetone phosphate; GOL = glycerol; RUSP = ribulose-5-phosphate, XUSP = xylulose-5-phosphate; RSP = ribose-5-phosphate; ARA = arabitol; S7P = sedoheptulose-7-phosphate; PYR = pyruvate; ACD = acetaldehyde; ETH = ethanol; ACOOA = acetyl CoA; OAA = oxaloacetate; CIT = citrate; AKG = alpha-ketoglutarate; MAL = malate; (E) = external

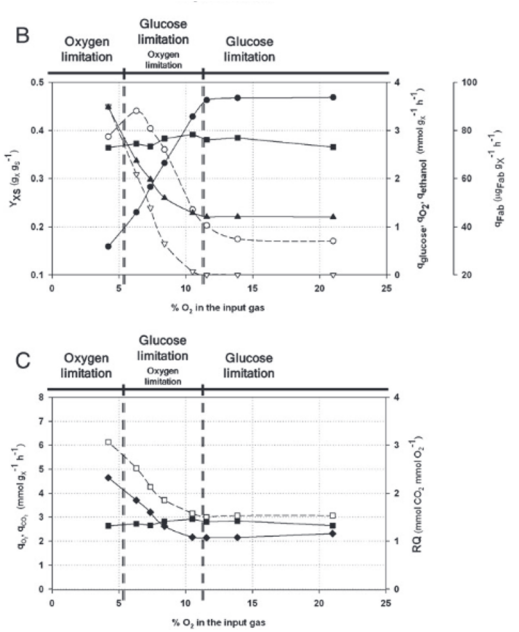
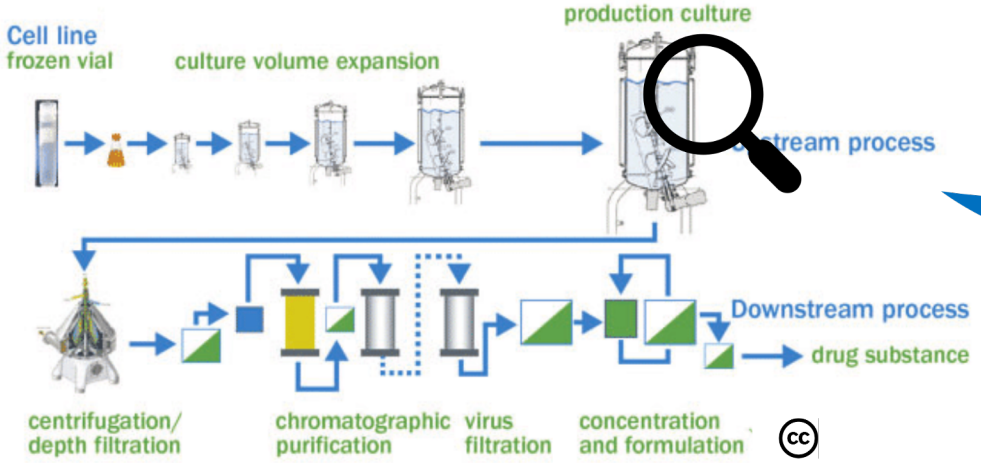
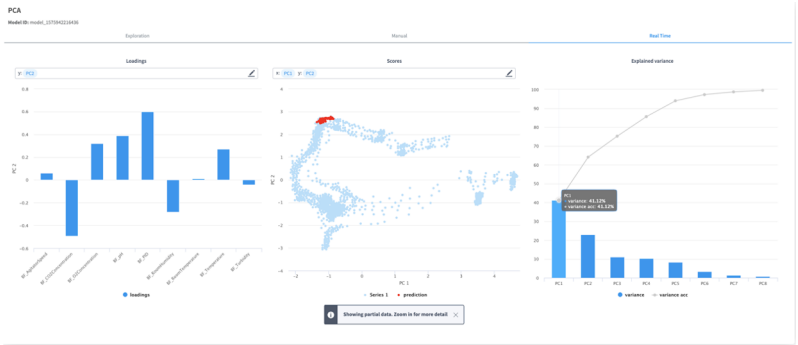


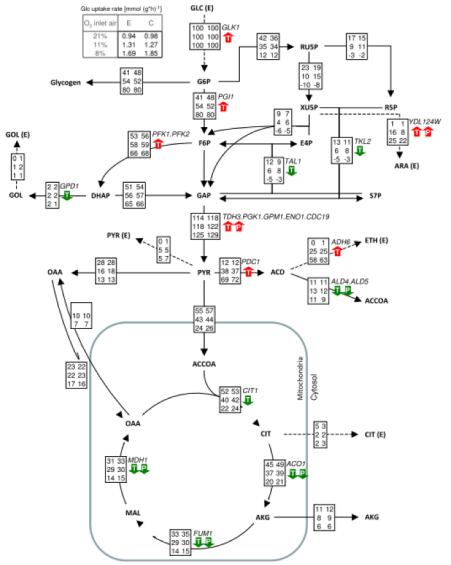
Figure 1. (A) Main cultivation parameters at different oxygen supply levels: dry cell weight (*DCW*, ●); 2F5 Fab titration (○); glucose concentration (▲); ethanol concentration (▽); and dissolved oxygen (*DO* or pO_2, ■). (B) Biomass yield and main specific rates of the cultivation at different molar fraction of oxygen in the inlet gas: biomass yield (Y_{XS}, ●); specific 2F5 Fab production rate (q_{Fab}, ○); specific glucose uptake rate ($q_{glucose}$, ▲); specific ethanol production rate ($q_{ethanol}$, ▽); and specific oxygen uptake rate (q_{O2},■). (C) Specific oxygen uptake rate (q_{O2}, ■); specific carbon dioxide production rate (q_{CO2}, □); and respiratory quotient (*RQ*, ◆) at different molar fraction of oxygen in the inlet gas.

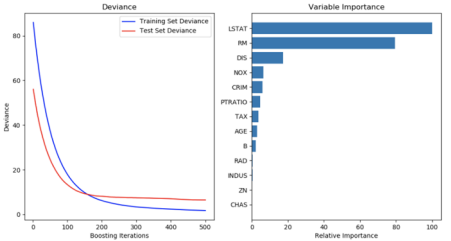
Getting Knowledge by Understanding Reality

XAVIER HEALTH



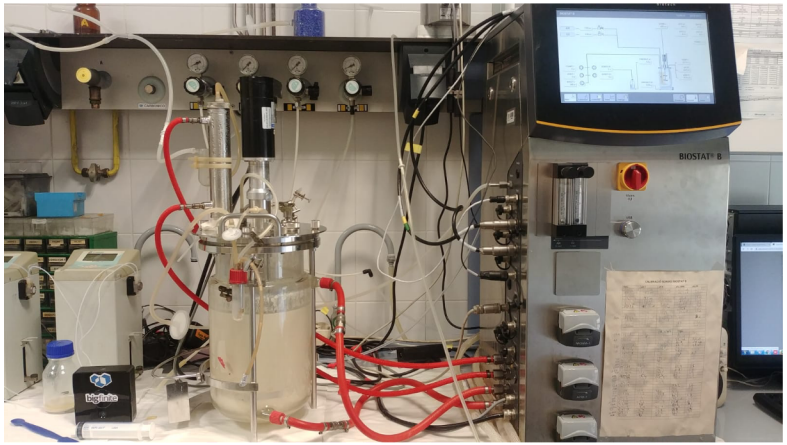


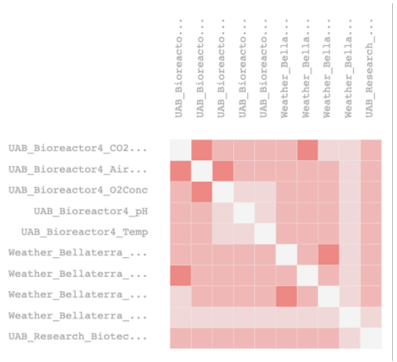


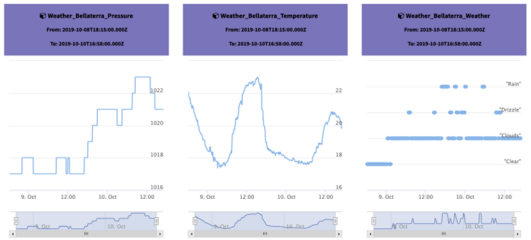


CPV of the Future: Batch Sequence-Based on DoE + AI

XAVIER HEALTH

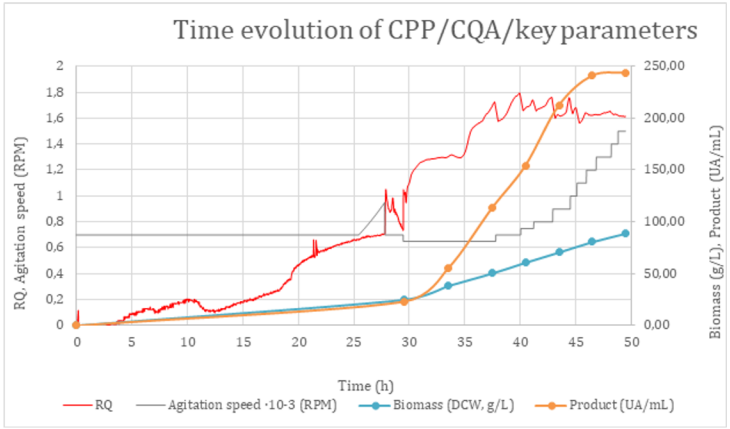




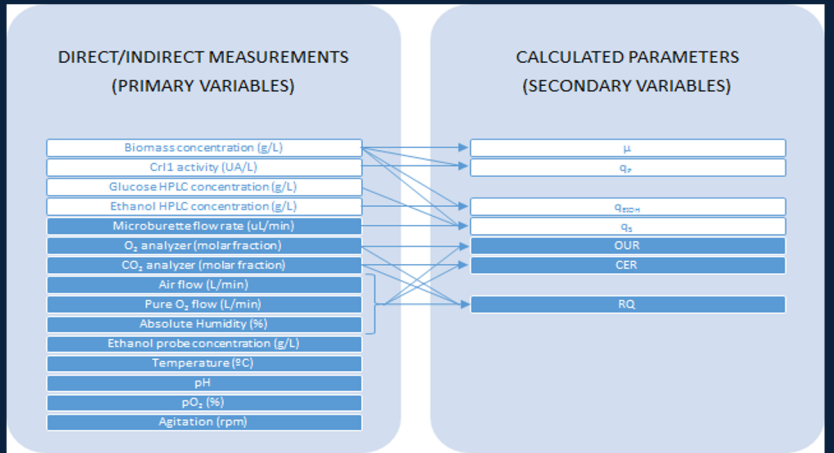


DoE, CQA, CPP

- 10 experimental runs performed (5 hypoxic and 5 normoxic)
- 11th experimental run for the validation of the AI model



	Normoxic			Hypoxic		
	Good	Average	Bad	Good	Average	Bad
Phase I (Batch)	FBHPX2 FBHPX5 !! FBHPX6 FBHPX9	FBHPX8		FBHPX3 FBHPX4 !! FBHPX10 FBHPX11	FBHPX7	
Phase II (Adaptation)	FBHPX2 FBHPX5 !! FBHPX6 FBHPX8	FBHPX9		FBHPX3 FBHPX4 !! FBHPX7 FBHPX10 FBHPX11		
Phase III (Early Fed Batch)	FBHPX2 FBHPX5 !! FBHPX6 FBHPX8 FBHPX9			FBHPX7	FBHPX4 !! FBHPX10 FBHPX11	FBHPX3
Phase IV (Later Fed Batch)	FBHPX5 !! FBHPX6 FBHPX8 FBHPX9	FBHPX2		FBHPX7 FBHPX10 FBHPX11		FBHPX3 FBHPX4 !!



The reality and the complexity

XAVIER HEALTH

Dependent variables:

- CER (carbon emission rate)
- OUR (oxygen uptake rate)
- RQ (respiratory quotient, defined as CER/OUR)

$$\frac{\Delta \log_2 J_i}{\Delta \log_2 J_i} = \frac{\Delta \log_2 [mRNA_i]}{\Delta \log_2 J_i} + \frac{\Delta \log_2 \left(\frac{V_{max,i}}{[mRNA_i]}\right)}{\Delta \log_2 J_i} + \frac{\Delta \log_2 g_i(X,K)}{\Delta \log_2 J_i}$$

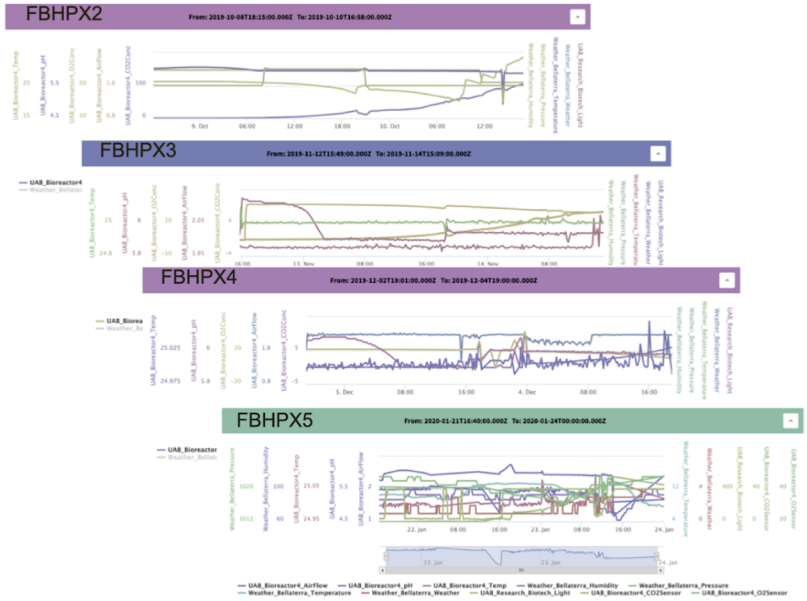
Independent variables:

- Ethanol concentration signal
- pO2
- Agitation speed (rpm)
- Temperature
- pH
- Microburette flow (substrate addition rate)
- Fraction of pure oxygen in the total in-flow (overall inflow is a combination of pure oxygen and air)

$$\frac{J_i}{e_i} = k_{i,cat} * \prod_j^n \left(\frac{\boldsymbol{X_j}}{\boldsymbol{P_j} + \boldsymbol{X_j}}\right) * h_i(\boldsymbol{Y}, \boldsymbol{Q}) * \left(1 - \frac{\boldsymbol{\Gamma_i}}{\boldsymbol{K_{i,eq}}}\right)$$

The reality and the complexity is not linear!!

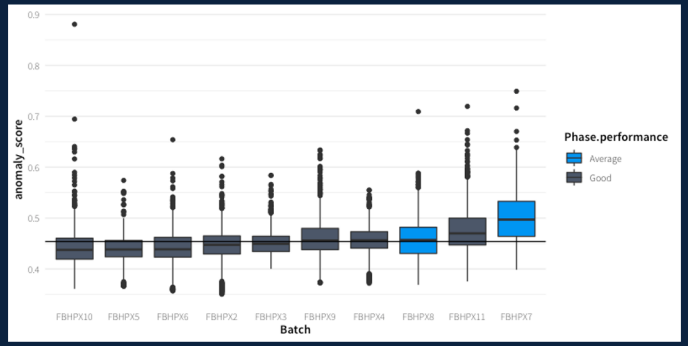
XAVIER HEALTH

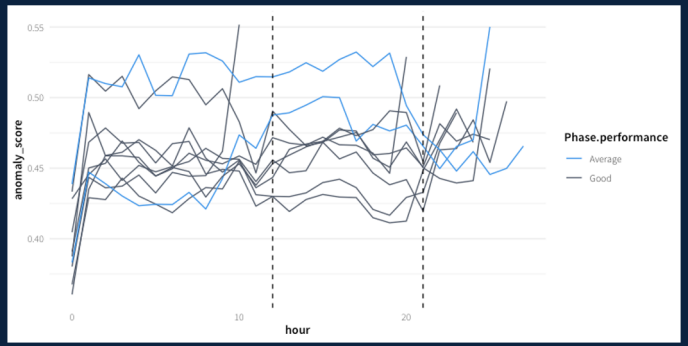


Anomaly detection in the Batch phase of the process

XAVIER HEALTH

- 9 process parameters included
- Because the state of the process is constantly changing, we need a model for every hour of the process
- Algorithm was able to isolate poor performing batches in agreement with SME evaluation.
- Different weights for different process parameters can give different anomaly results, giving the opportunity to tweak the model according to needs



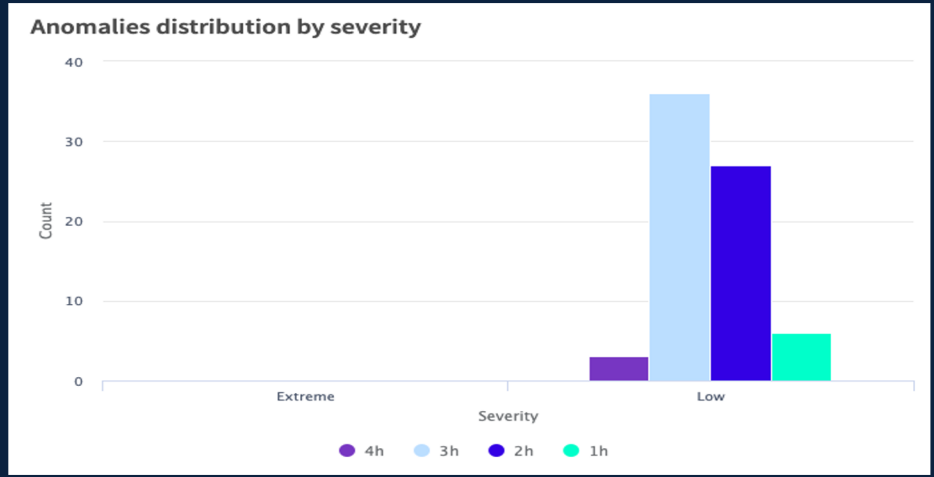


Application of the anomaly detection in in the cloud in real-time

One algorithm and one model per hour.

Added value:

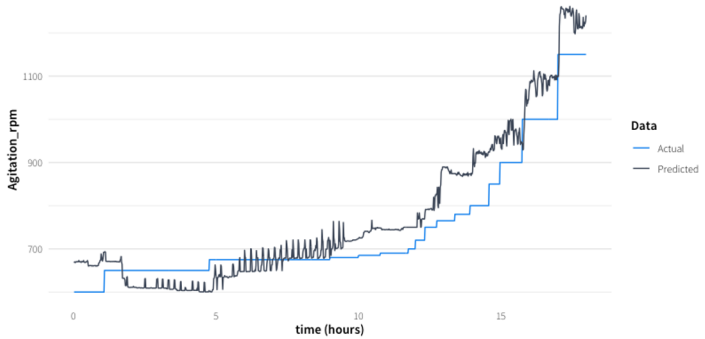
- Easy detection of anomalous batches in historical data + assess upcoming batches in real-time



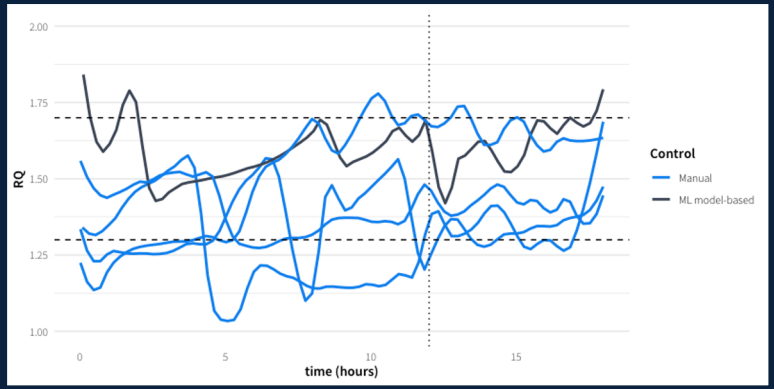
Final test: applying the model in real-time to guide the experiment

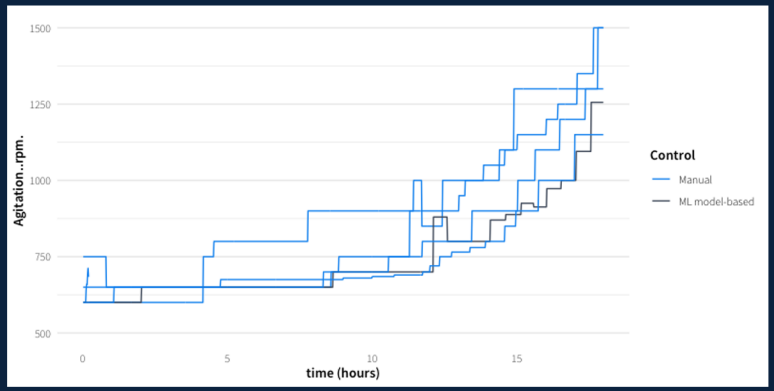
Manual action recommendation

- Task: developing a model for predicting the ideal agitation speed from a limited set of historical batches in early and later fed batch phases



The AI model-controlled batch performed equally well as manually controlled batches.





Synthetic data generation

XAVIER HEALTH

The 3-step algorithm in a nutshell

Assumptions:

- The variables can be meaningfully split into dependent and independent (broadly referred to as “causes” and “consequences”)
- **No assumptions are made** regarding the relationship between the variables (no input from chemistry and physics knowledge), the algorithm is purely statistics and data driven, therefore can be applied to any variables without a priori knowledge.

1. Using experimental data, generate a supervised ML model for each dependent variable.

2. For each independent variable, generate synthetic data using a data-driven monte carlo algorithm.

3. For each dependent variable, apply the ML models to generate data using the generated independent variable data.

The 3-step algorithm

XAVIER HEALTH

Division of variables:

- Dependent variables: CER, OUR, RQ - key metabolic variables (CQAs)
- Independent variables: 7 control parameters (such as temperature, pH, metabolite concentration

Step 1: Development of ML models predicting CER, OUR and RQ from the independent parameters -- high accuracy models (MAPE < 5%)

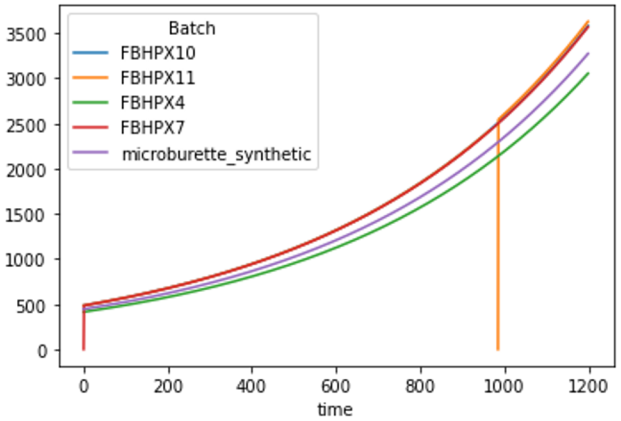
Product Quality = f(t)

CQA = $f(t = t_f)$ } ??

Step 2: two algorithms for independent data

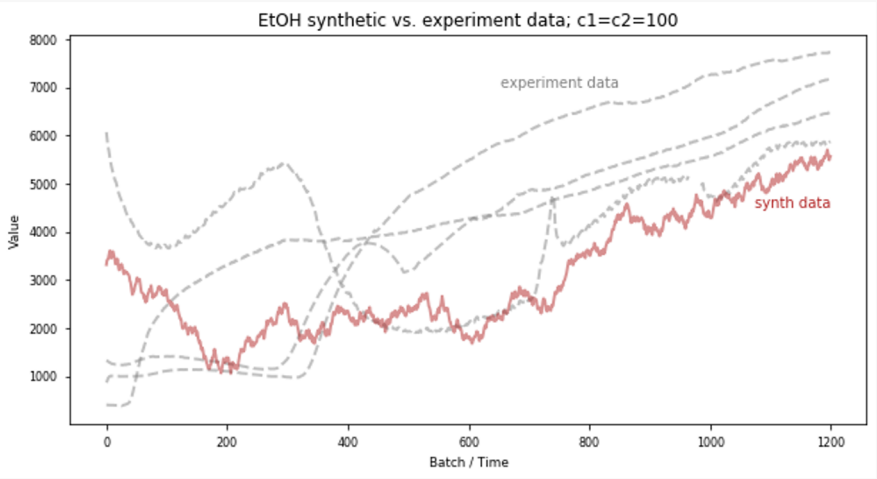
XAVIER HEALTH

Data that can be described by a function; example: **Microburette**



Microburette data is simulated using an exponential function with slight variation of the parameters

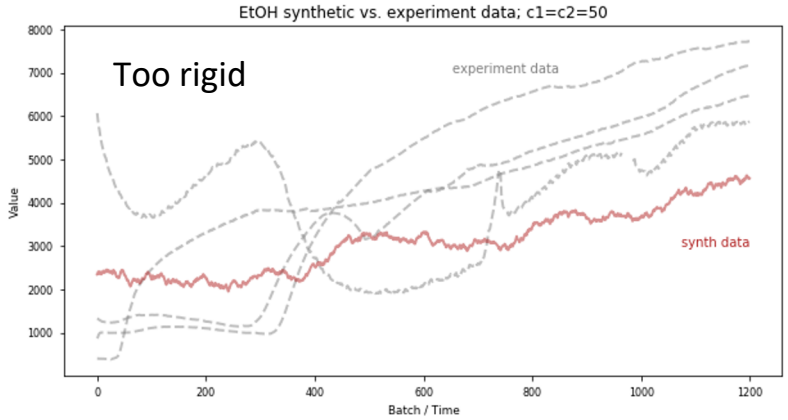
Data that **cannot** be described by a function; example: **Ethanol**

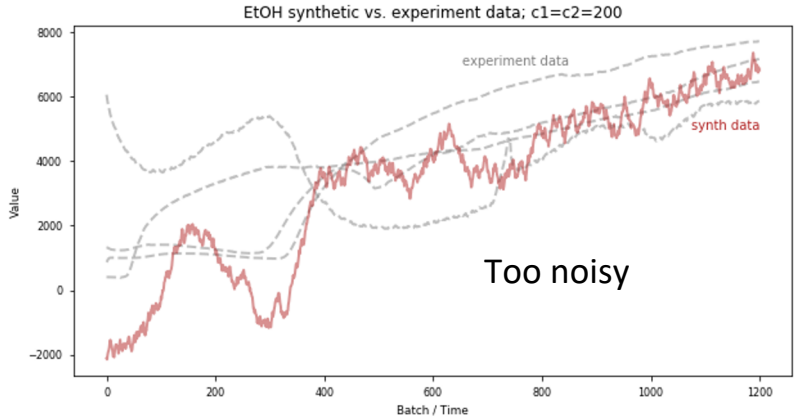


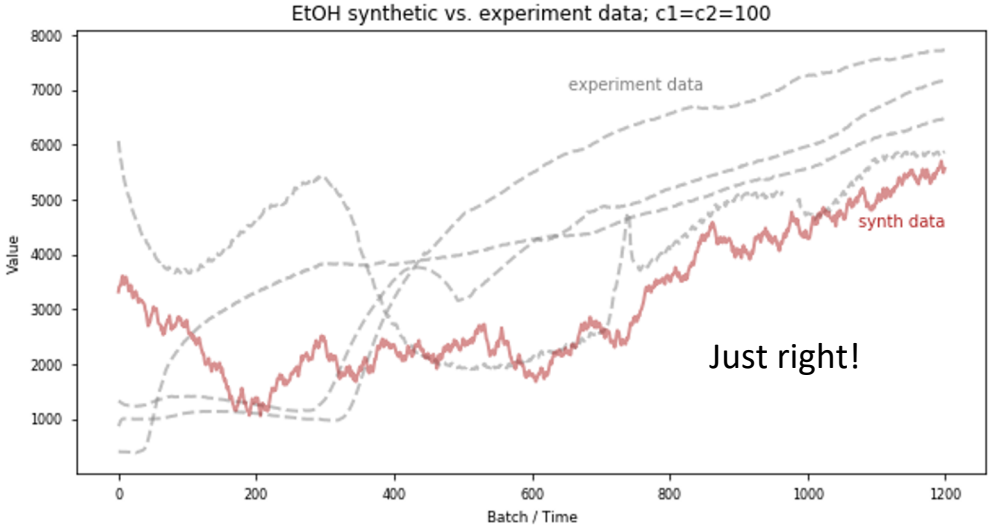
A custom monte carlo algorithm that samples the space in the (experimental) data

19

Step 3: The custom data generator.
2 parameters that determine the level of stiffness / flexibility



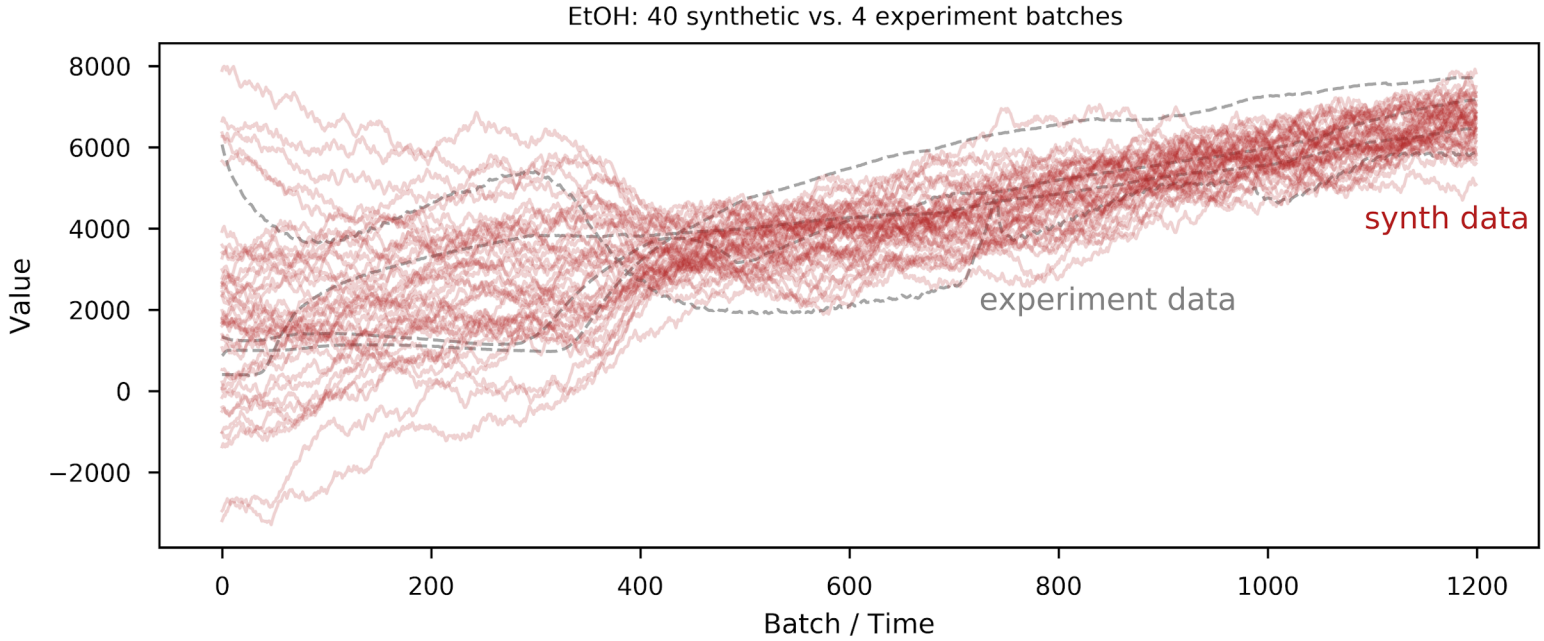




High-throughput generation of synthetic data

XAVIER HEALTH

With one run, we can generate an arbitrarily large number of synthetic batches



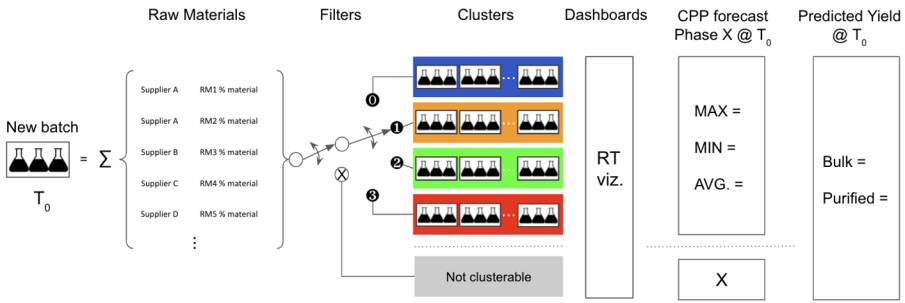
With

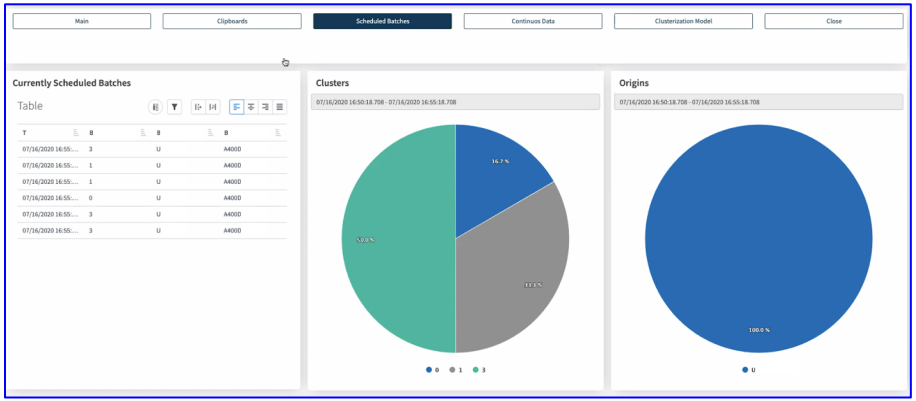
CPV of the Future project

A collaboration between PDA, PQRI and AI Xavier

Top 5 Plasma Company. *Yield Prediction*

XAVIER HEALTH





BUSINESS CHALLENGE

A concise and quantifiable problem statement defined by the customer

- Continuous decrease of yield in the last 4 years
- Wider process variability leading to abnormally low yields since 2 years ago

The customer was facing a highly multidimensional problem across several processes lasting between 7 and 9 days with great confusion about the root cause over univariate analysis.

SOLUTION

Multivariate analysis to determine relevant factors. Dependence test (Graphical Lasso) and Causality detection (Bayesian network), were inconclusive. PCA shown too many Principal Components and a myriad of latent process variables with poor dimensionality reduction

- Unsupervised learning model to identify patterns among different plasma composition origins in thousands of batches over the years
- Supervised learning model to predict yield for each upcoming batch
- Real-time Dashboarding

RESULTS

- Plasma composition origin explained more than 50% of the overall process variance.
- Batch clusterization was critical to yield prediction
- Only 2 process variables were sufficient to improve the yield by 4%

Yield decay in plasma purification process

XAVIER HEALTH

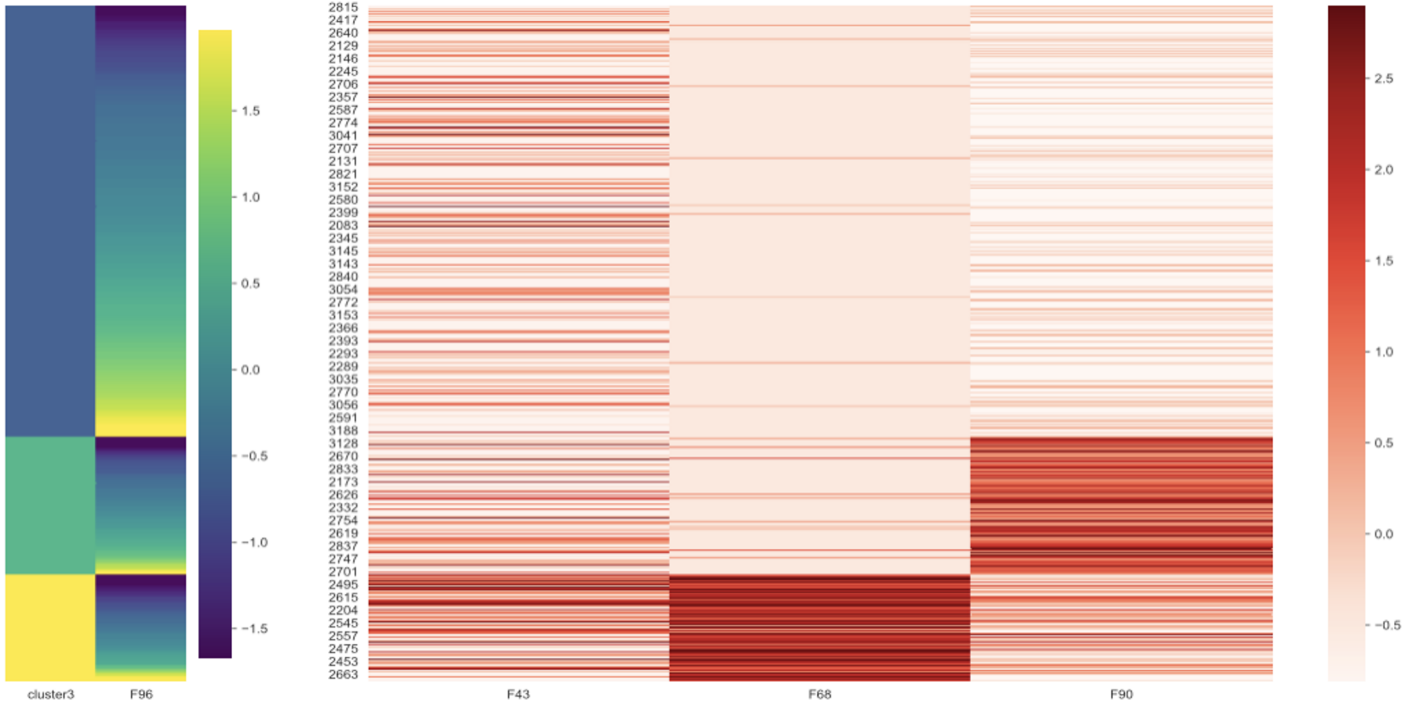
- **Problem**: reduction in the evolution of process yield over the course of 3 years, but only for early stage purification products
- **Goal**: identify root cause of decrease in mean yield and larger dispersion in the low-yield domain
- **Approach**:
  - Correlation between yield and CQA's → changes depending on the year
  - Univariate relationship over time with yield → no clear indications
  - Multivariate linear model (PLS) → Variance was not properly explained by model with origin parameters
  - Clusterization of batches by origin → four clear clusters emerge, two are related to decreasing yields
  - Stratified correlations by cluster → identify centrifugation & filtration CPP's to be involved
- **Blockers**:
  - Lack of data contextualization for off-process data
  - Low data integrity compliance (errors due to manual extraction)
  - No access to historical data

Targeted CPP Clustering

XAVIER HEALTH

Goal: Isolate certain sets of CPP's, employ clustering methodologies (K-means and hierarchical), and evaluate cluster results relative to yield (or other target variable)

Clustered 3 different hold time variables to determine best stage in process to hold processing



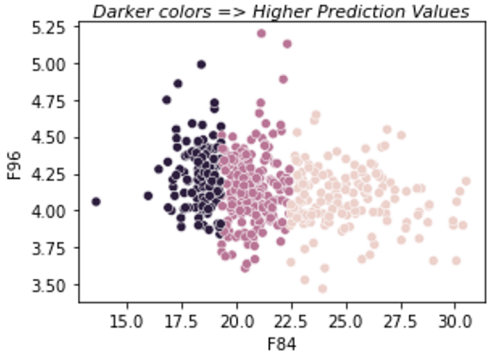
Decision Tree Analysis

Goal: Utilize shallow decision trees to determine optimal cutoff thresholds for CPPs

Initial conclusions:

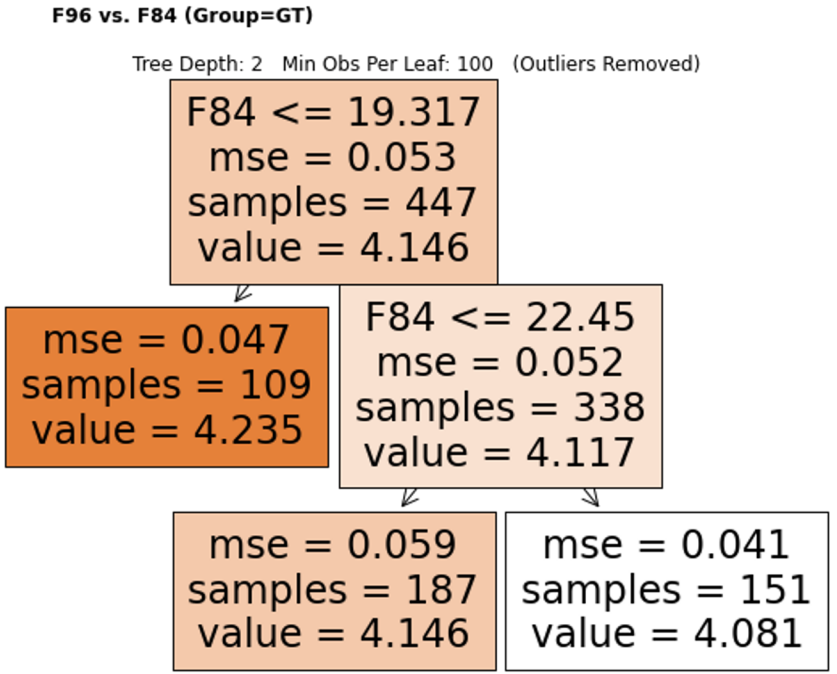
- Target = *Yield (g/L)*
- Input = *Process Time*

Shorter process times (<=19.3 hours) are associated with significantly higher yields



Top 5 Bins

Prediction	Min Value	Max Value
4.235	13.583	19.3
4.146	19.333	22.433
4.081	22.467	30.517



Rolling Correlations

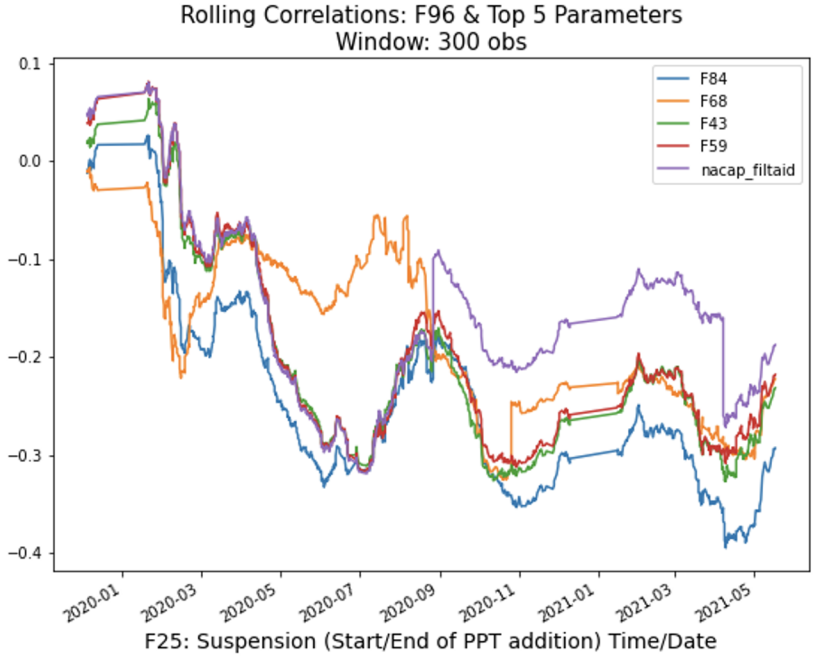
Goal: Look at historical rolling correlation windows to evaluate stability of correlations of many variables relative to a specified target. Average correlations relative to standard deviation of historical correlations are computed to create a “stability” metric (similar to a Sharpe ratio).

Result:

Evaluate all ‘duration’ variables relative to final yields. The most relevant [anonymized] variables were identified.

Avg. Correlations with F96

Var	mean	std
F84	-0.244	0.094
F68	-0.177	0.082
F43	-0.197	0.104
F59	-0.19	0.105
nacap_filtaid	-0.152	0.091
F38_sum_F59	-0.171	0.115
F90	-0.14	0.111
mic1_mic2	0.023	0.056
F38	0.024	0.094
F26	-0.001	0.04



Random Forest Modeling/Analysis

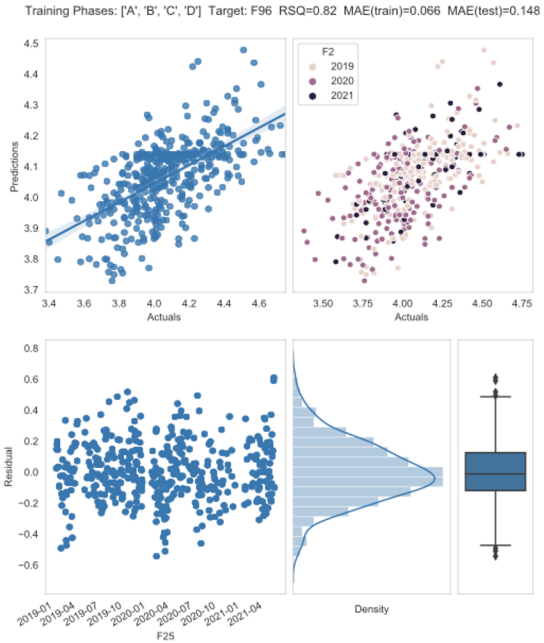
XAVIER HEALTH

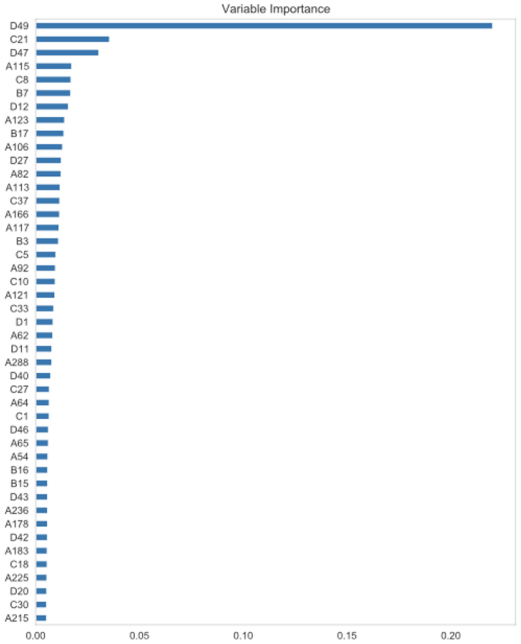
Goal: Implement Random Forests to create predictive models to be implemented in the platform and to discover 'important' features.

Result:

Predict final yield using all prior process phases in production (filtration, centrifugation, purification, etc).

- Rsq = 0.82
- D49 (anonymized attribute) is the most important feature





Thank you!

toni.manzano@aizon.ai