SARTORIUS

Simplifying Progress

Multivariate Data Analysis (MVDA) for the Beginner

Lennart Eriksson, Ph.D., Assoc. Prof.
Senior Lecturer and Principal Data Scientist

Born in Data Analytics

- Company founded in 1987 by Professor Svante Wold, in Umeå, Sweden
  - Originator of Chemometrics and the SIMCA® Methodology
- Patented technologies in Design of Experiments and Multivariate Data Analysis

- We help our customers bring high-quality products to market faster
- Part of Sartorius Stedim Biotech since April 2017
- Products like MODDE®, SIMCA® and SIMCA®-online
- Global strength with local presence

Business Growth Through the Entire Product Lifecycle

Active Dashboard
Interactive performance insight

Control Advisor
Avoid problems before they arise

MODDE®
Get it right from the start

SIMCA®
Turn data into growth

SIMCA®-online
Ensured manufacturing success

Umetrics® Suite

Education/training and consulting
Transferring expertise to you

Umetrics® Suite of Data Analytics Solution
MODDE® Design of Experiments Solution SIMCA® Multivariate Data Analysis Solution SIMCA®-online Solution

SARTORIUS

Contents

The Data Explosion

- More data than ever are measured

- How Much Data is Produced Every Day?
  - 1.7MB of data is created every second by every person during 2020
    https://techjury.net/blog/how-much-data-is-created-every-day/#gref

- Data is everywhere!
  - From production floor to your local grocery store and facebook!

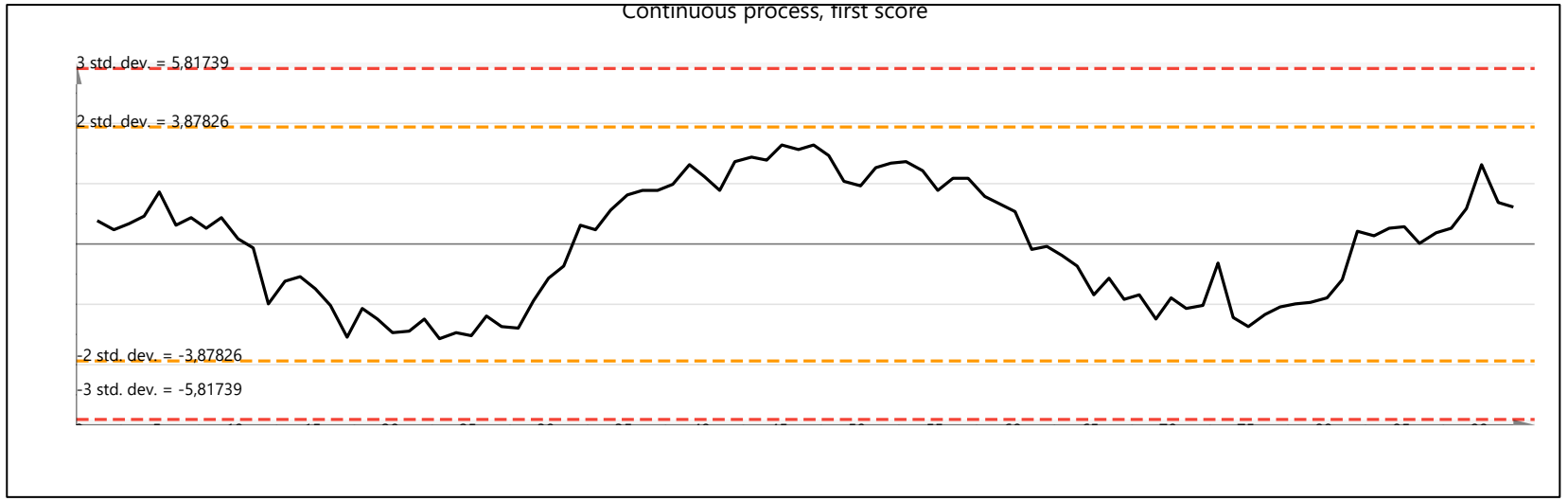
- How do we get *Value from data*®?

What is MVDA?

Is This Chart Familiar?

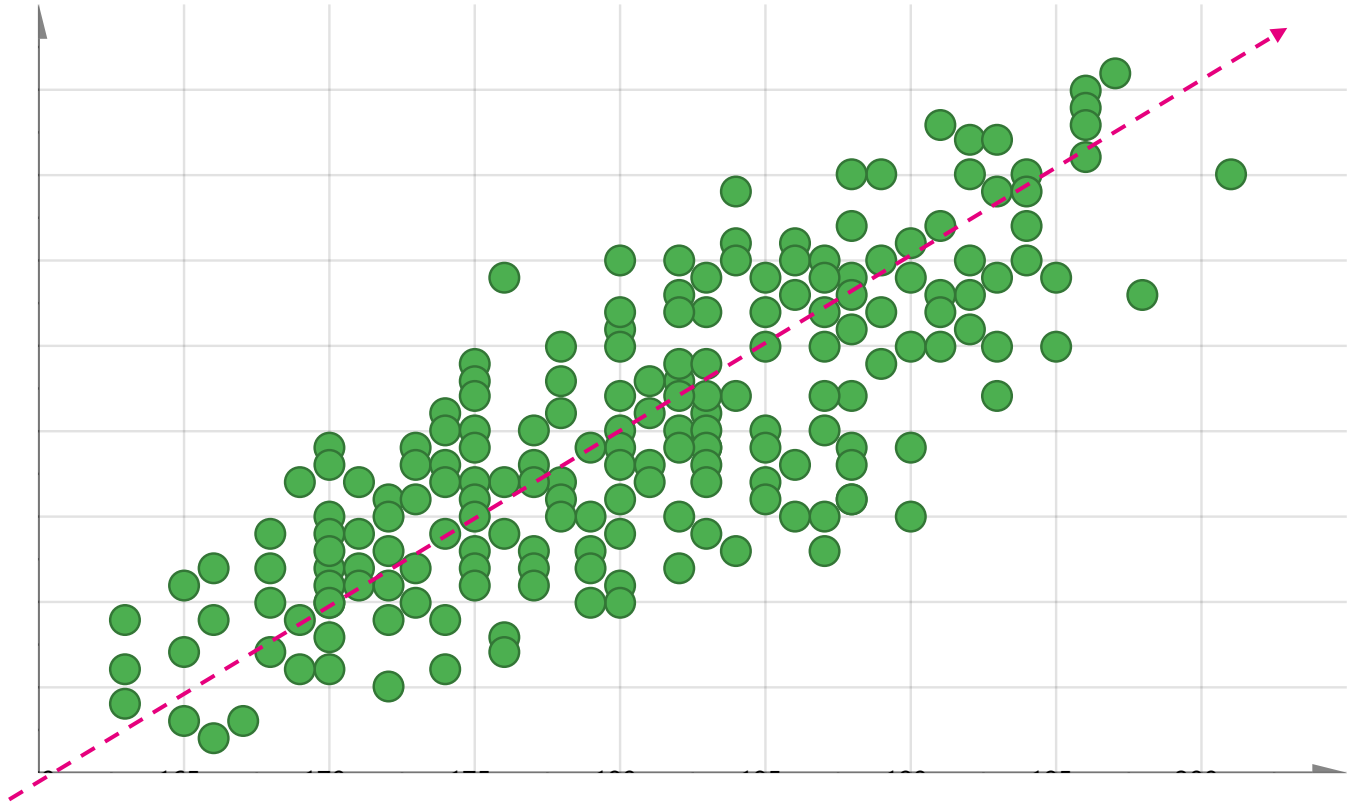
Dow Jones = x1*Merck + x2*J&J + x3*Pfizer + x4*DuPont + ...

So This Control Chart Is Easy to Understand....

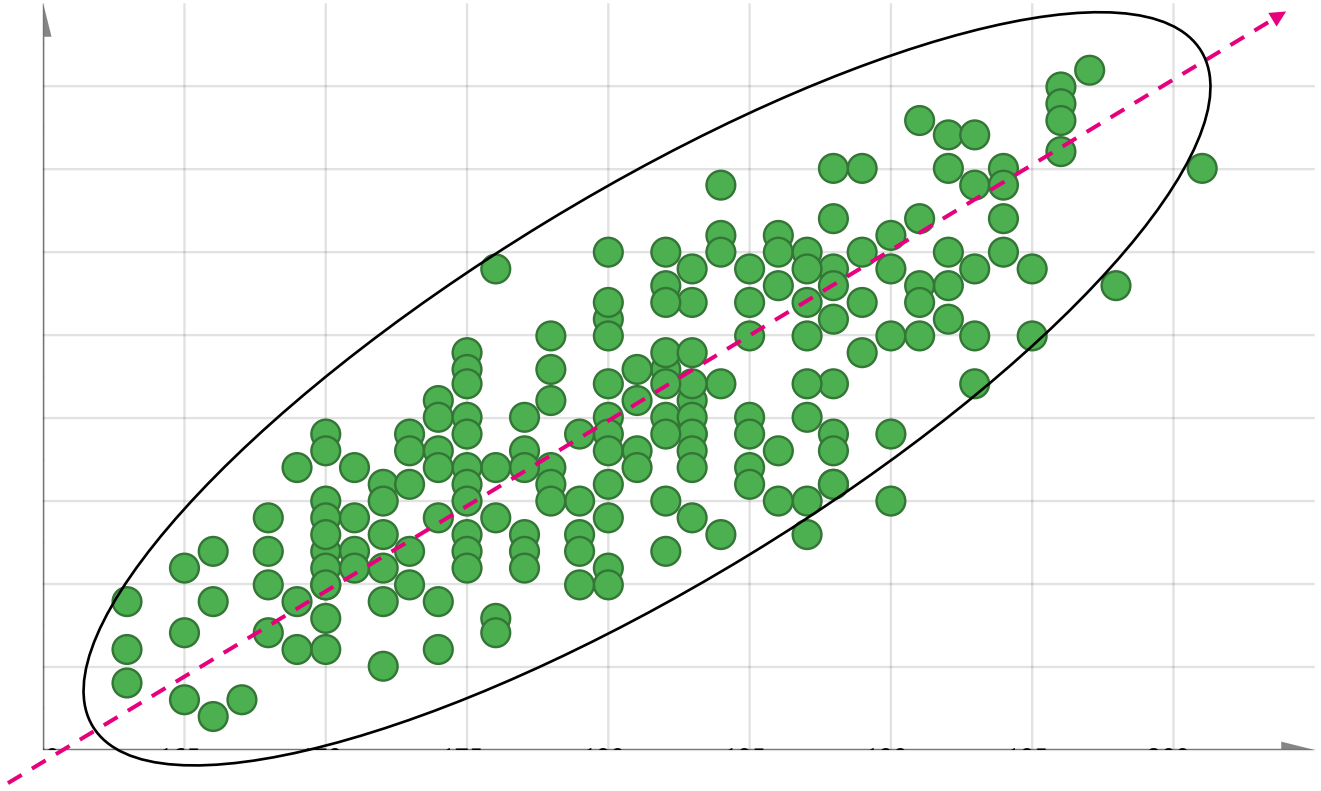


t[1] = x1*Temperature + x2*Pressure + x3*Speed + x4*pH + ...

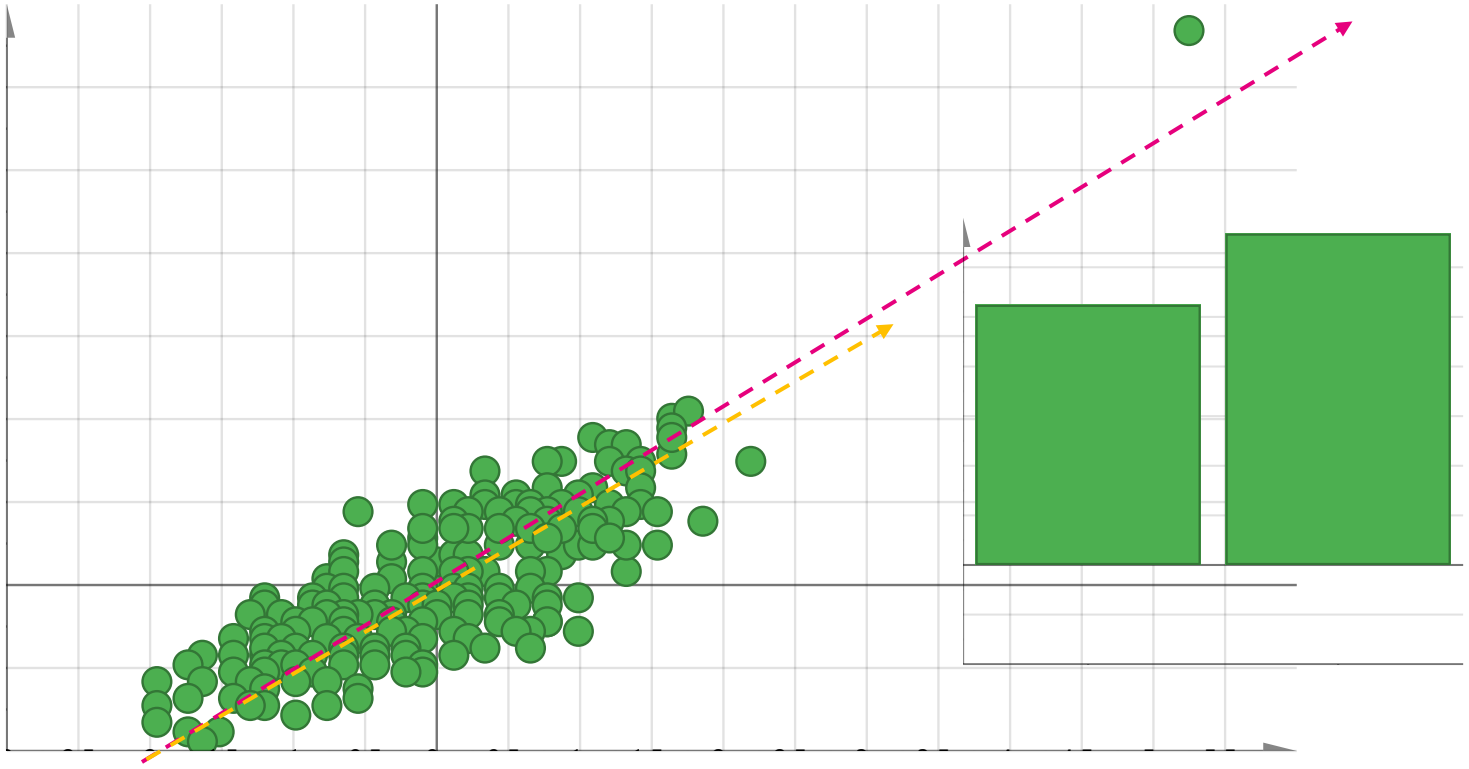
Let’s Play Soccer!



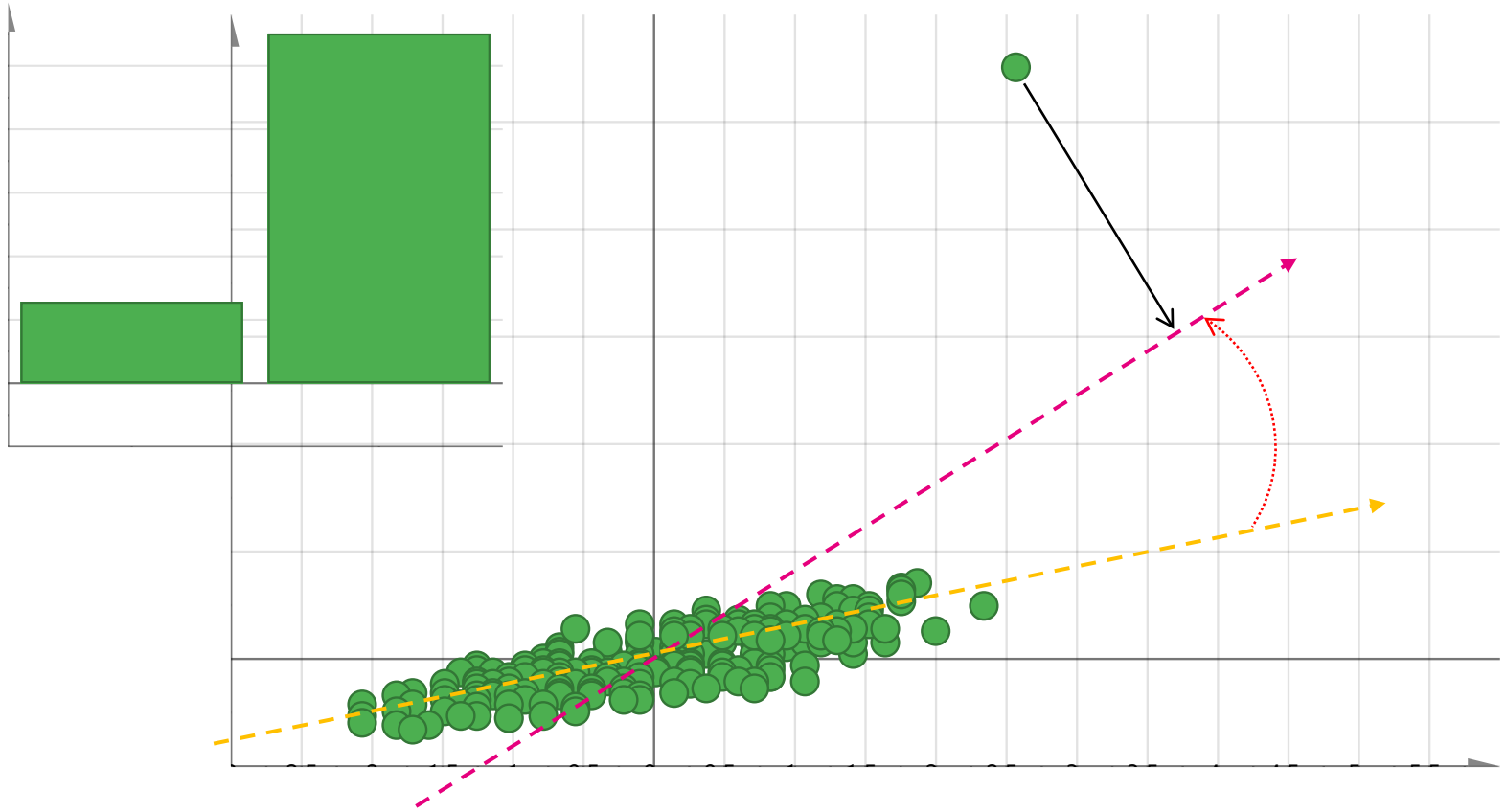
Let’s Play Soccer!



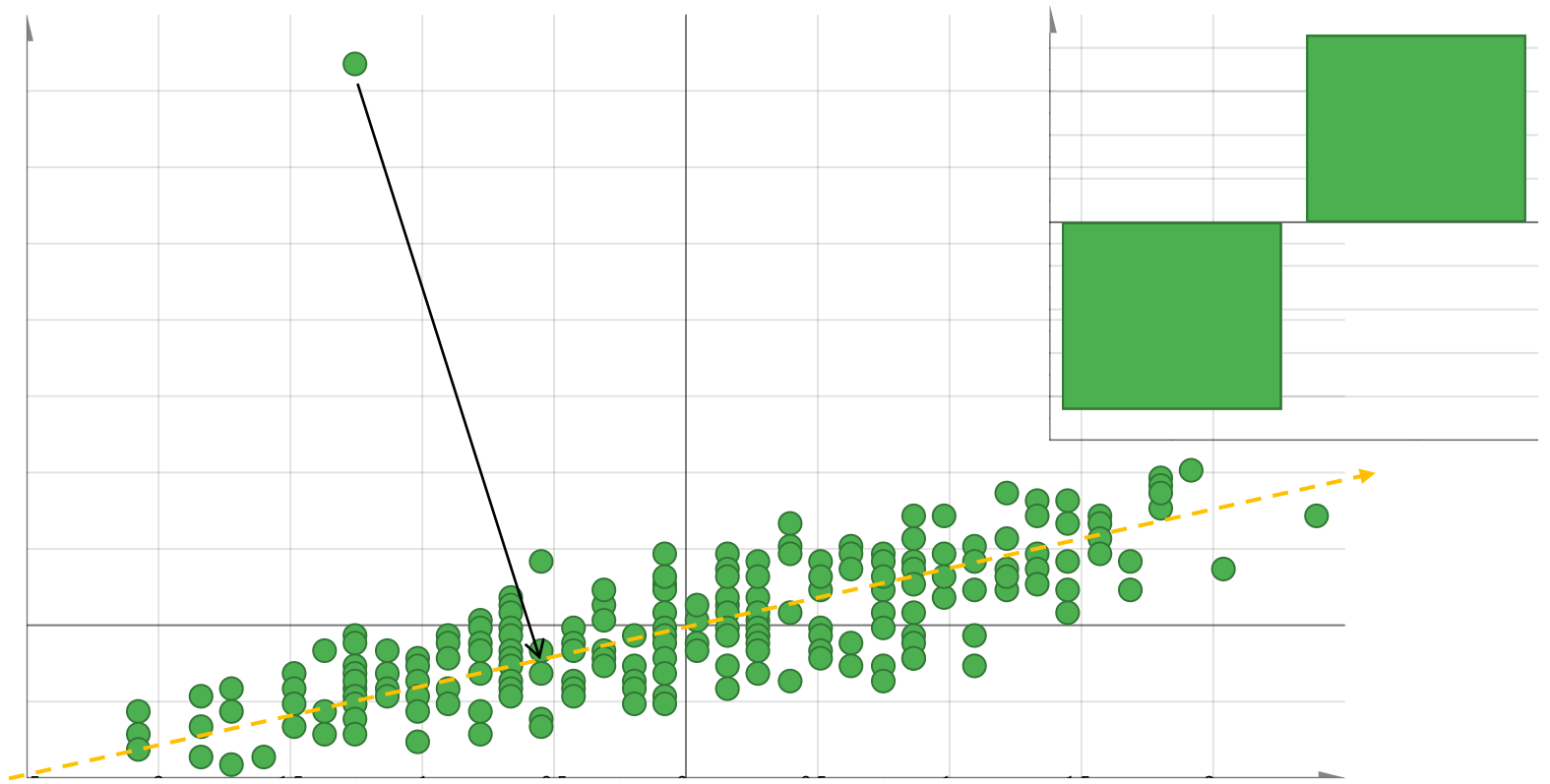
Adding a Basketball Player...?

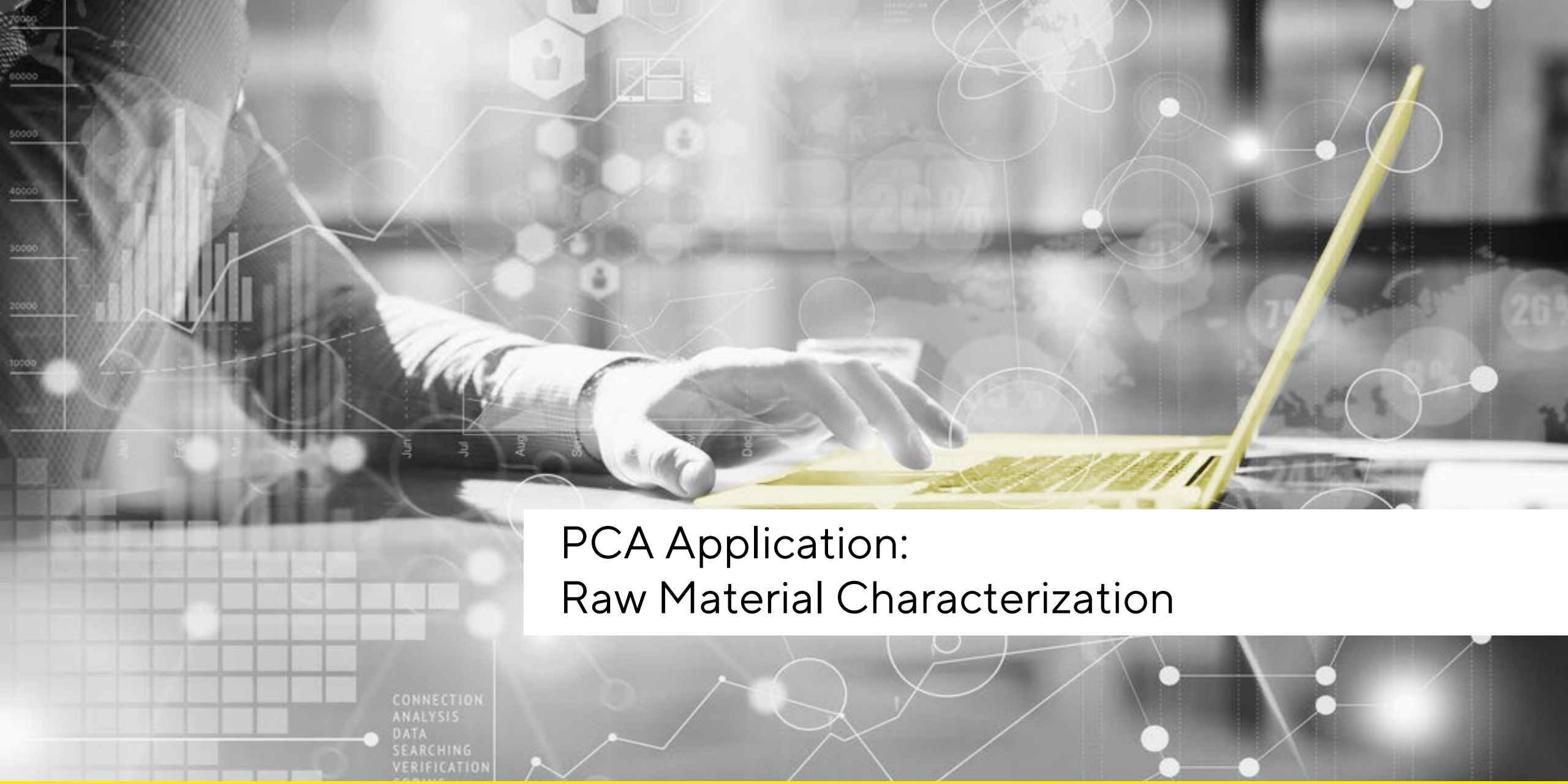


Adding a Sumo Wrestler…?



Adding a Gorilla...?





PCA Application:
Raw Material Characterization

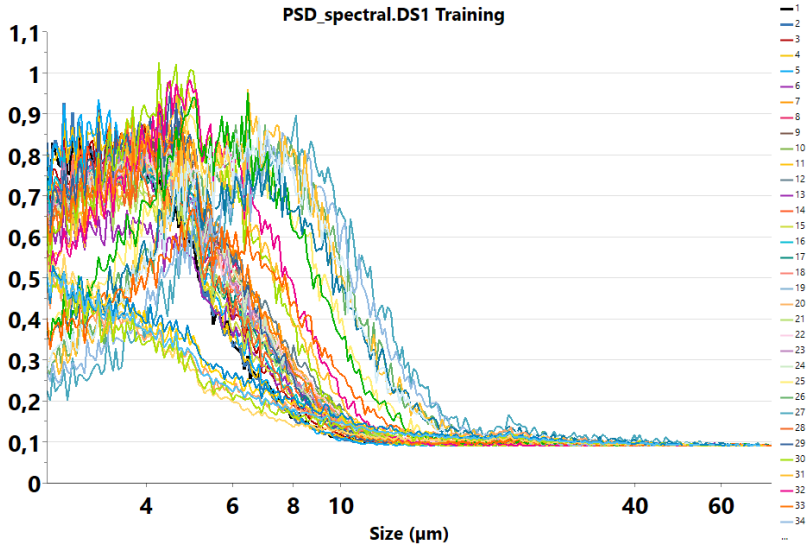
Introduction

- Variation in raw material properties can be one reason for problems during manufacturing.
- Such undesired variability may arise because different suppliers supply raw material that is not consistently the same over time, or that, within the batches of raw material from one single supplier, there is batch-to-batch variability.
- The objective of this presentation is to demonstrate the utility of PCA in the assessment of variations in raw material qualities.
- Great economical value if incoming batches of raw material can be classified as suitable or unsuitable for continued production.

Raw Material Characterization

- The particle size distribution (PSD) dataset comes from a pharmaceutical company and relates to all incoming batches during approximately two years of production.

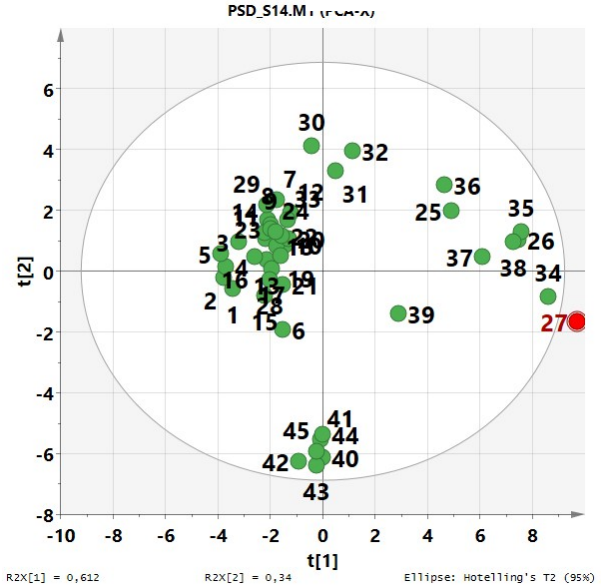
- There are in total 250+ variables corresponding to particle size intervals
  - Particle sizes range between 2.5–150 µm
  - Non-linear bin sizes; narrowest bins for the smallest particle sizes
    - Log transform of the X-axis in plots a possibility
  - The readout represents the number of particles in each bin (size interval).

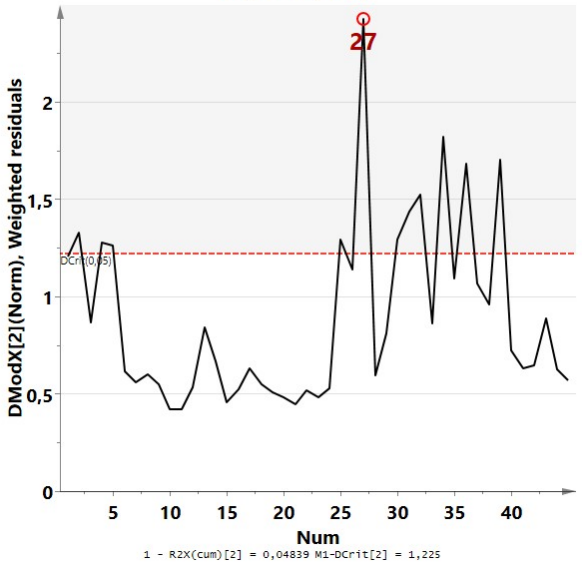


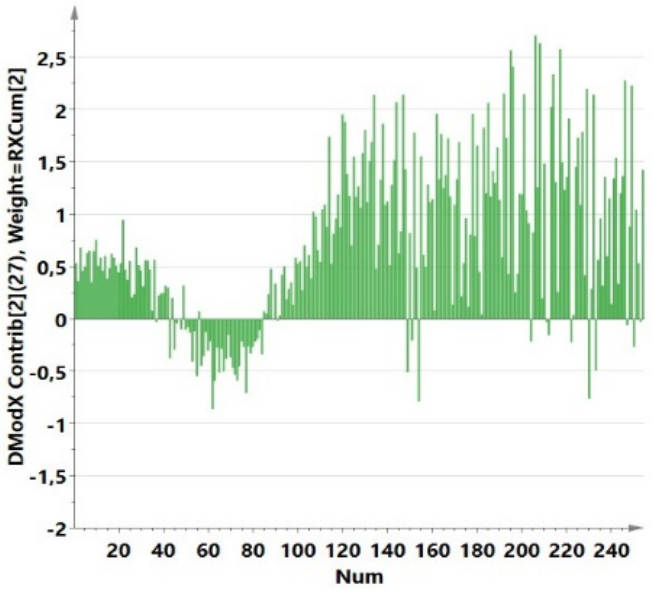
Can PSD curves be used for quality control of future batches?

Defining the Quality Control Model

- Two principal components with R2 and Q2 > 0.95
  - DModX suggests one moderate deviatior; can be interpreted with contribution plot

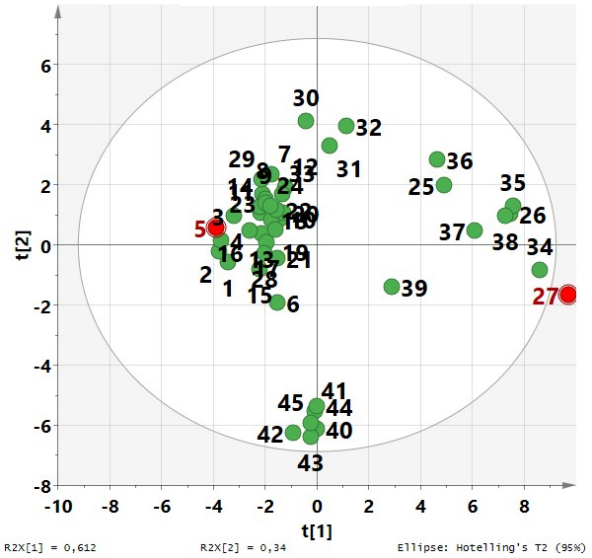


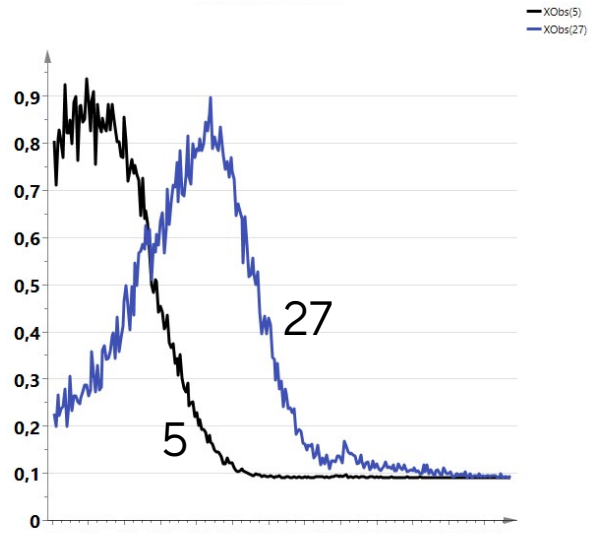




What Do the Components Mean?

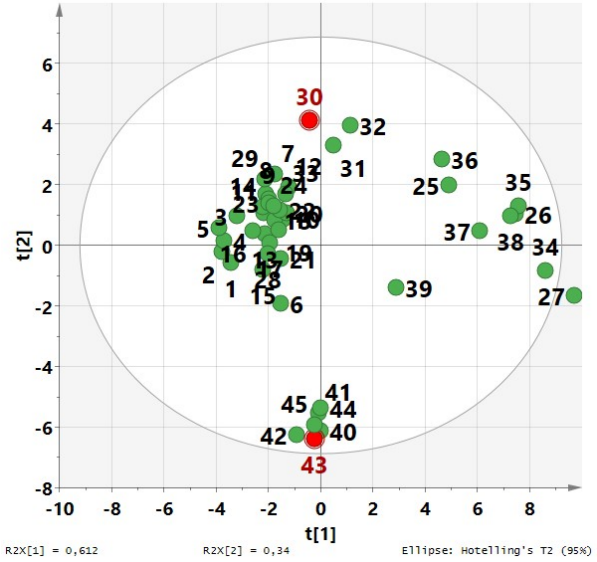
- Batches 5 and 27 span the first PC.
  - PC1 discriminates between the batches depending on the typical (average) particle size. Going from left to right (horizontal direction) in the score plot implies a transition from high proportion of small particles to high proportion of large particles.

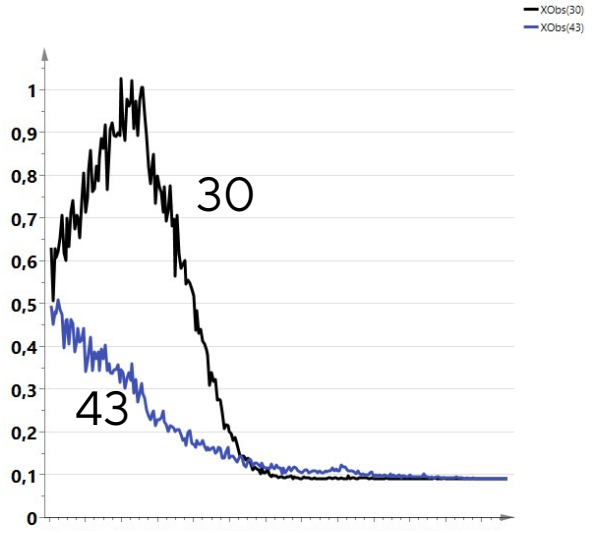




What Do the Components Mean?

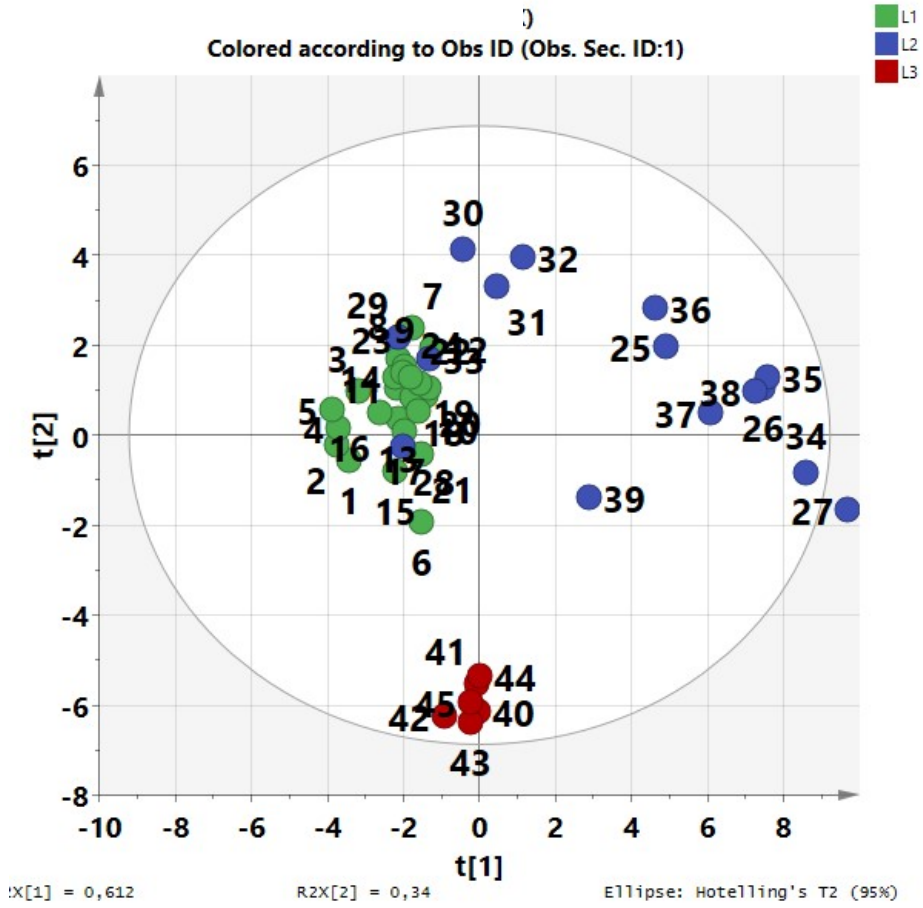
- Batches 43 and 30 span the second PC.
  - PC2 separates between the batches depending on the “peakedness” of the PSD curves. Going from bottom to top (vertical direction) in the score plot implies a transition from comparatively flat distributions to very sharp distributions.





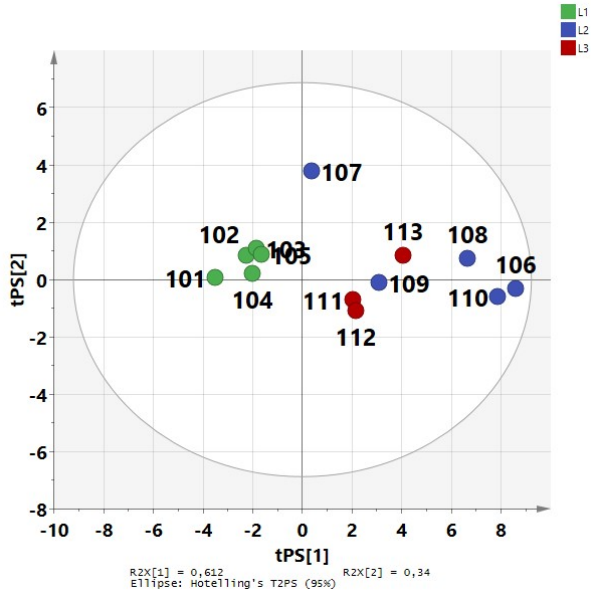
What About the Vendors?

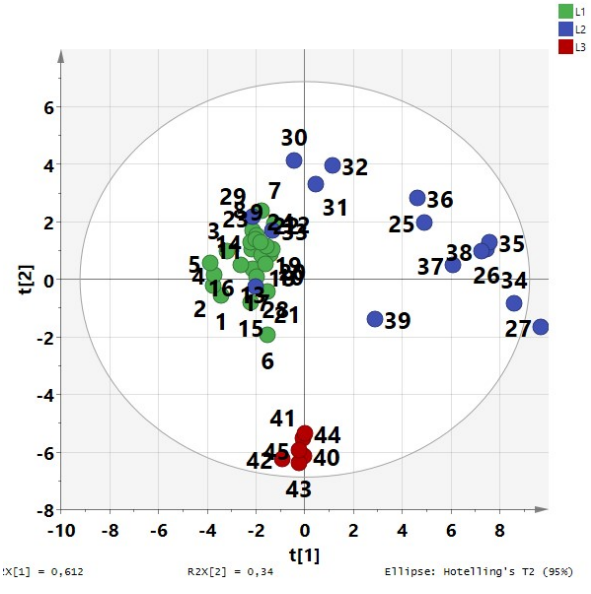
- Suppliers L1 and L3 provide the most homogeneous and consistent starting material with little quality variation.
  - However, there is a systematic difference between L1 and L3.
- L2 is the vendor that has the largest variations in raw material quality.



Are the Suppliers Shipping Consistent Qualities?

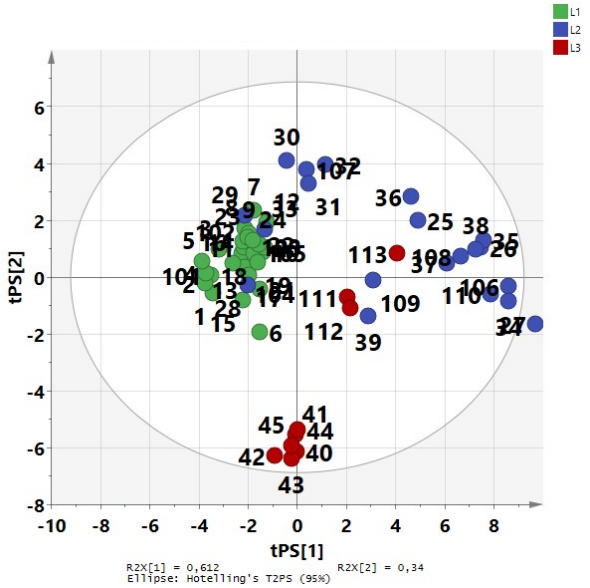
- L1 YES !
- L2 YES, but still with larger variability than L1 !
- L3 NO ! Huge shift !

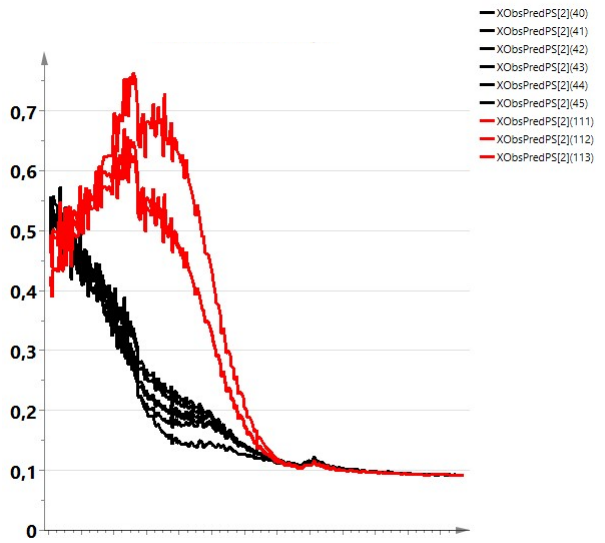




Why the Quality Change for L3?

- Co-display the training set and prediction set in the same scatter plot
  - Use Plot Xobs option to visualize change for L3
  - Significant shift in shape of PSD curves





The Value

- Reliable quality control model established
- It uncovered the new batches from the L3 supplier as being very different compared with the earlier L3 batches
- The main reason being that the fresh batches comprised much higher proportions of larger particles

SIMCA® modelling for Ambr® cell culture development

SARTORIUS

Background

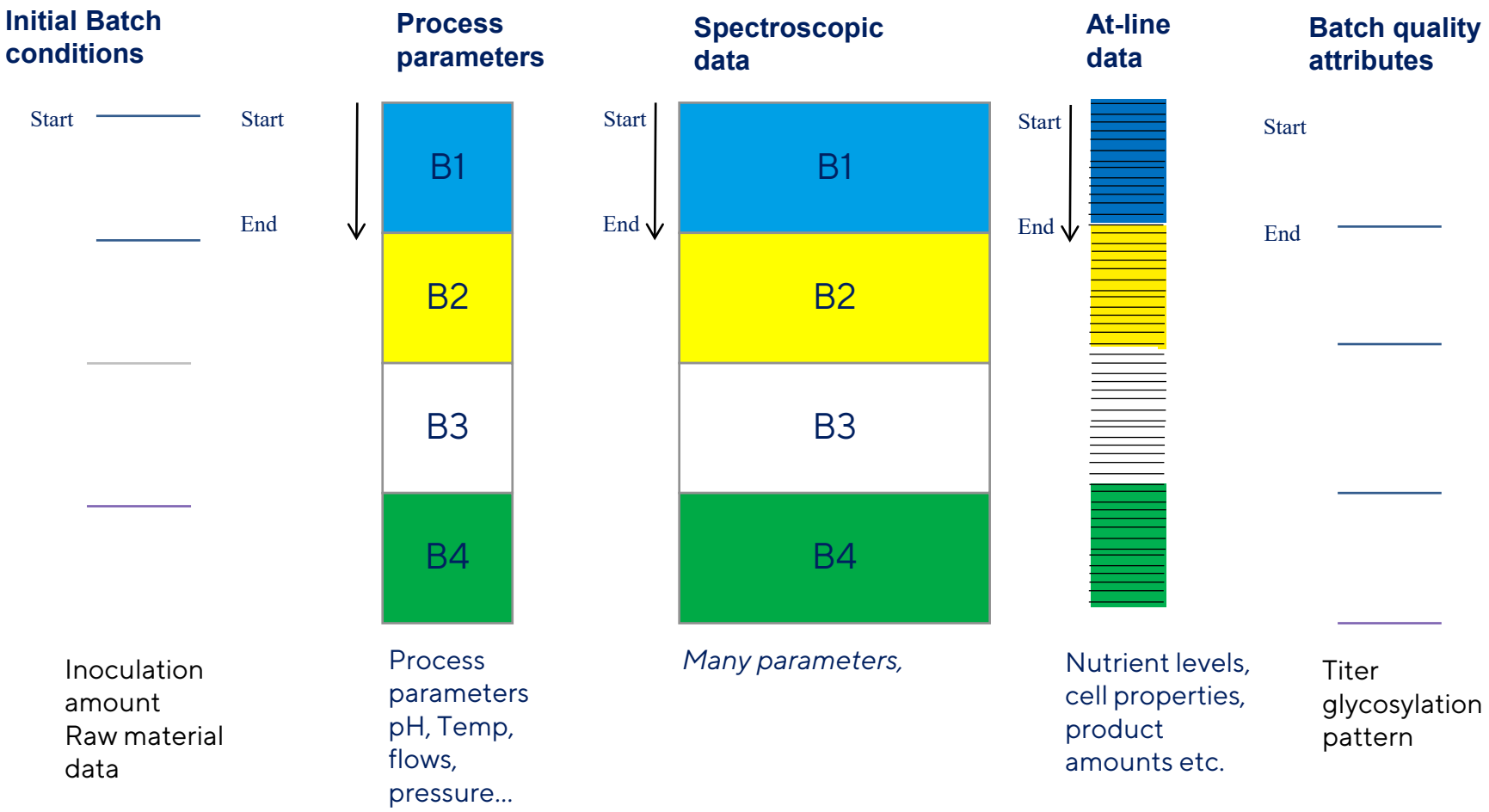
Ambr® 15 cell culture and Ambr® 250 high throughput systems will generate large amount of data spread over different data tables, which creates challenges in analyzing the data.

- In addition, the data tables will be in difference frequencies which will add additional challenges
  - On-line Process data can be extracted every 10 minutes or faster
  - BioPAT® Spectro every 6 hours or faster
  - Daily data, At line data every day
  - Initial conditions and Process Outputs and CQA data will have one row od data per Batch
- SIMCA® is a very good tool to analyze all of these data tables one by one
- SIMCA is also the ideal tool to analyze all of these data tables **jointly** to get a more **holistic view** of all the data
- SIMCA modelling will give deeper process understanding from your Ambr experiments
  - leading to better optimization overall

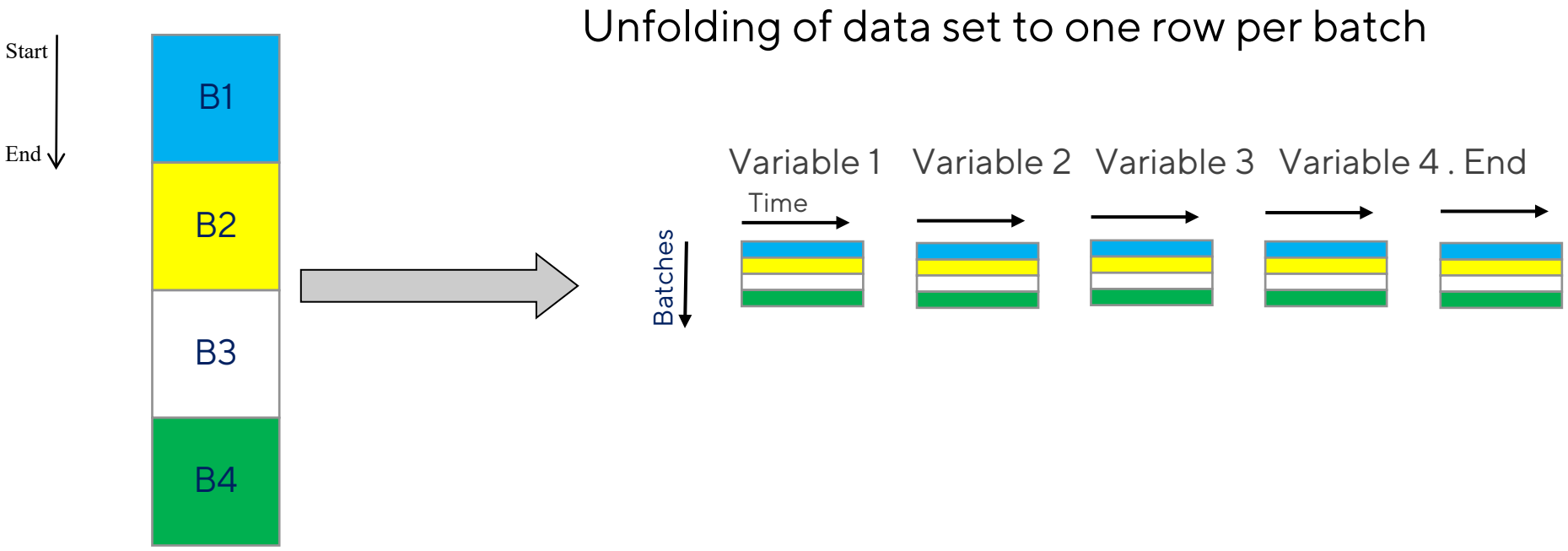
Experimental Set-Up

- Background
  - Fourteen (14) fed-batch cell culture processes were performed for process development objectives in Ambr 250 HT. The experiments involved processes performed under different conditions. As common practice the process performance metabolite measurements and cell characteristics were registered once per day.
- Data
  - Evolution data collected once per day. Also initial batch conditions and the process output are available. The process output will be used when a batch level model is generated.
- Scope
  - Show how a multivariate model for the batch evolution is generated from an excel table.
  - Demonstrate how to visualize the data in two ways, Batch Evolution mode (BEM) and Batch Level mode (BLM)
  - Identify odd batches, outliers
  - At the end we will use the evolution data predict the final titer and what are the drivers for higher titer

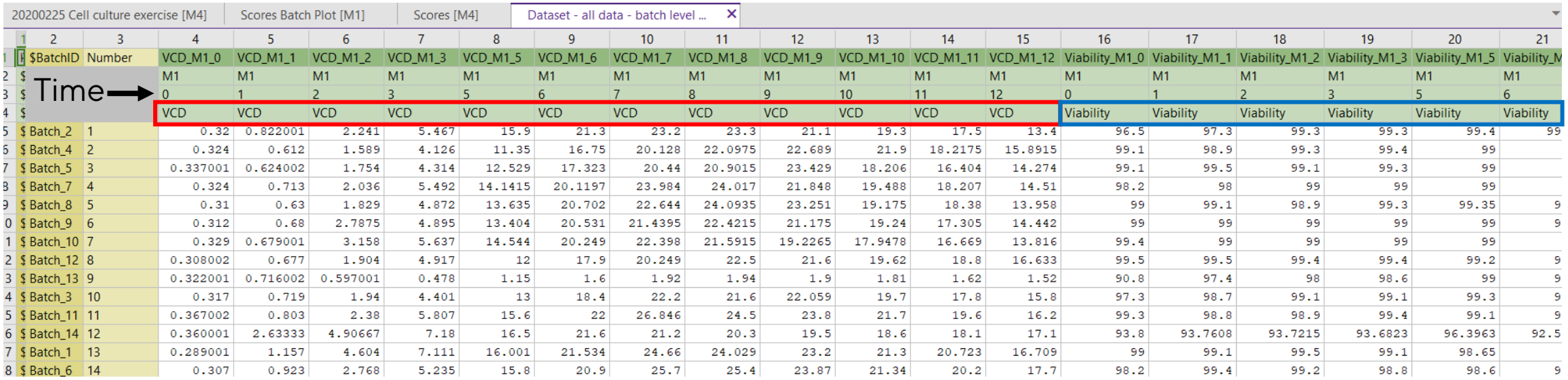
Data Structure (Batch Evolution Model, BEM)



Summarizing a Batch as One Point. Basis for Batch Level Model (BLM)



Summarizing a Batch as One Point

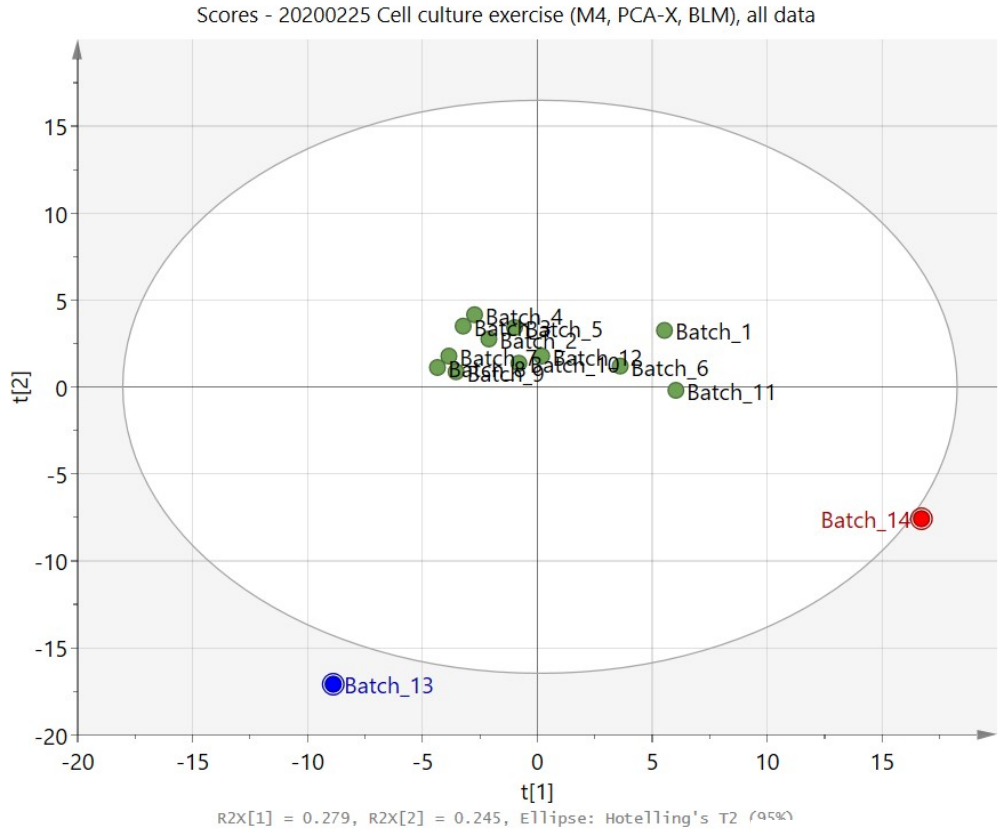


20200225 Cell culture exercise [M4] | Scores Batch Plot [M1] | Scores [M4] | Dataset - all data - batch level ...

Time →

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	$BatchID	Number	VCD_M1_0	VCD_M1_1	VCD_M1_2	VCD_M1_3	VCD_M1_5	VCD_M1_6	VCD_M1_7	VCD_M1_8	VCD_M1_9	VCD_M1_10	VCD_M1_11	VCD_M1_12	Viability_M1_0	Viability_M1_1	Viability_M1_2	Viability_M1_3	Viability_M1_5	Viability_M
2	$		M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1
3	$		0	1	2	3	5	6	7	8	9	10	11	12	0	1	2	3	5	6
4	$		VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	Viability	Viability	Viability	Viability	Viability	Viability
5	$ Batch_2	1	0.32	0.822001	2.241	5.467	15.9	21.3	23.2	23.3	21.1	19.3	17.5	13.4	96.5	97.3	99.3	99.3	99.4	99
6	$ Batch_4	2	0.324	0.612	1.589	4.126	11.35	16.75	20.128	22.0975	22.689	21.9	18.2175	15.8915	99.1	98.9	99.3	99.4	99	
7	$ Batch_5	3	0.337001	0.624002	1.754	4.314	12.529	17.323	20.44	20.9015	23.429	18.206	16.404	14.274	99.1	99.5	99.1	99.3	99	
8	$ Batch_7	4	0.324	0.713	2.036	5.492	14.1415	20.1197	23.984	24.017	21.848	19.488	18.207	14.51	98.2	98	99	99	99	
9	$ Batch_8	5	0.31	0.63	1.829	4.872	13.635	20.702	22.644	24.0935	23.251	19.175	18.38	13.958	99	99.1	98.9	99.3	99.35	9
10	$ Batch_9	6	0.312	0.68	2.7875	4.895	13.404	20.531	21.4395	22.4215	21.175	19.24	17.305	14.442	99	99	99	99	99	9
11	$ Batch_10	7	0.329	0.679001	3.158	5.637	14.544	20.249	22.398	21.5915	19.2265	17.9478	16.669	13.816	99.4	99	99	99	99	
12	$ Batch_12	8	0.308002	0.677	1.904	4.917	12	17.9	20.249	22.5	21.6	19.62	18.8	16.633	99.5	99.5	99.4	99.4	99.2	9
13	$ Batch_13	9	0.322001	0.716002	0.597001	0.478	1.15	1.6	1.92	1.94	1.9	1.81	1.62	1.52	90.8	97.4	98	98.6	99	9
14	$ Batch_3	10	0.317	0.719	1.94	4.401	13	18.4	22.2	21.6	22.059	19.7	17.8	15.8	97.3	98.7	99.1	99.1	99.3	9
15	$ Batch_11	11	0.367002	0.803	2.38	5.807	15.6	22	26.846	24.5	23.8	21.7	19.6	16.2	99.3	98.8	98.9	99.4	99.1	9
16	$ Batch_14	12	0.360001	2.63333	4.90667	7.18	16.5	21.6	21.2	20.3	19.5	18.6	18.1	17.1	93.8	93.7608	93.7215	93.6823	96.3963	92.5
17	$ Batch_1	13	0.289001	1.157	4.604	7.111	16.001	21.534	24.66	24.029	23.2	21.3	20.723	16.709	99	99.1	99.5	99.1	98.65	
18	$ Batch_6	14	0.307	0.923	2.768	5.235	15.8	20.9	25.7	25.4	23.87	21.34	20.2	17.7	98.2	99.4	99.2	98.8	98.6	9

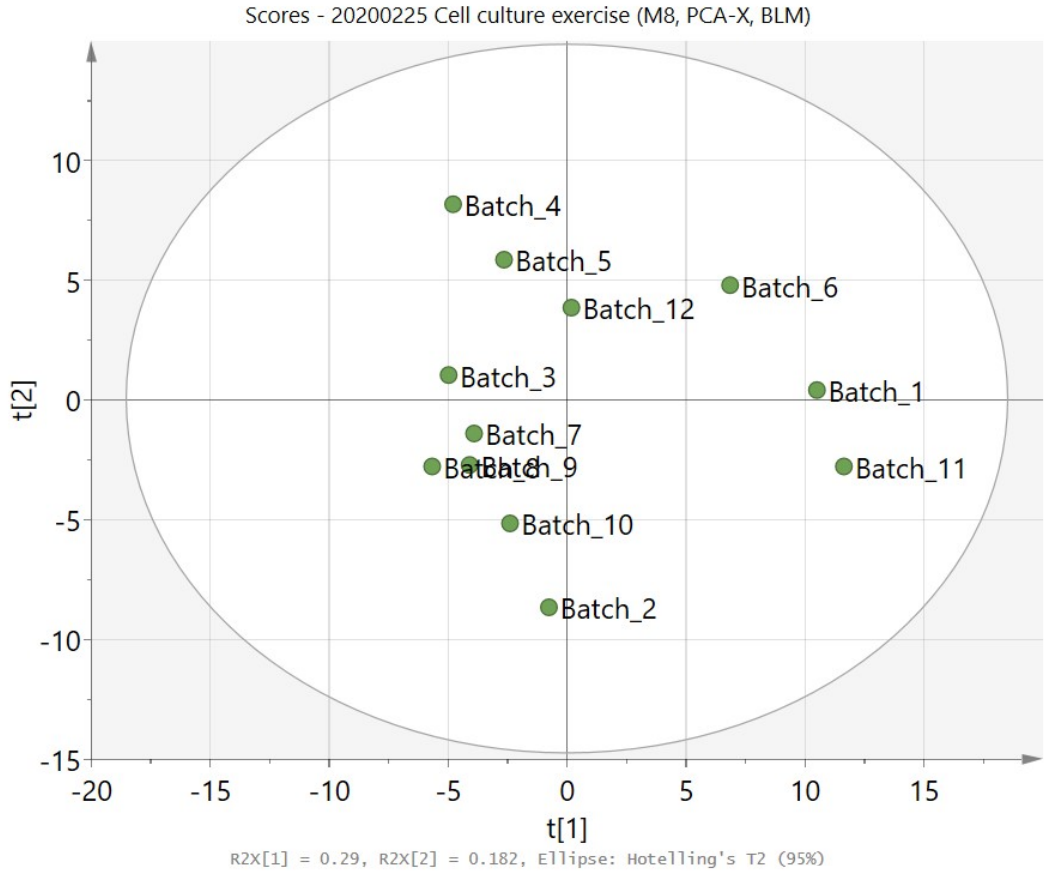
Interpretation of BLM using PCA. Batch 13 and Batch 14 Deviates.



Potential criteria to look for:

1. Are there batches out-side the 95% ellipse
2. Are there clusters visible (weak cluster batch *1,6,11)*
3. Use contribution plot to investigate difference

BLM 12 Batches, Overview using PCA.

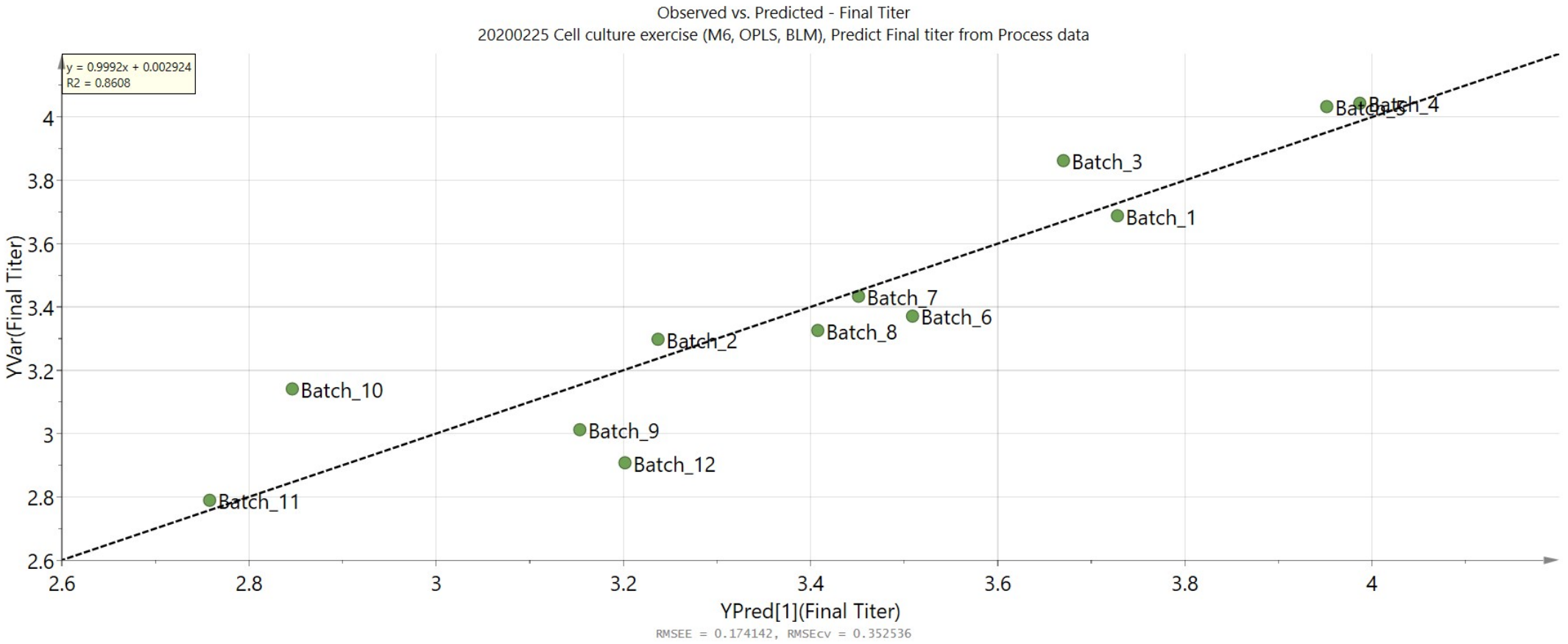


BLM 12 Batches Prediction of Titer

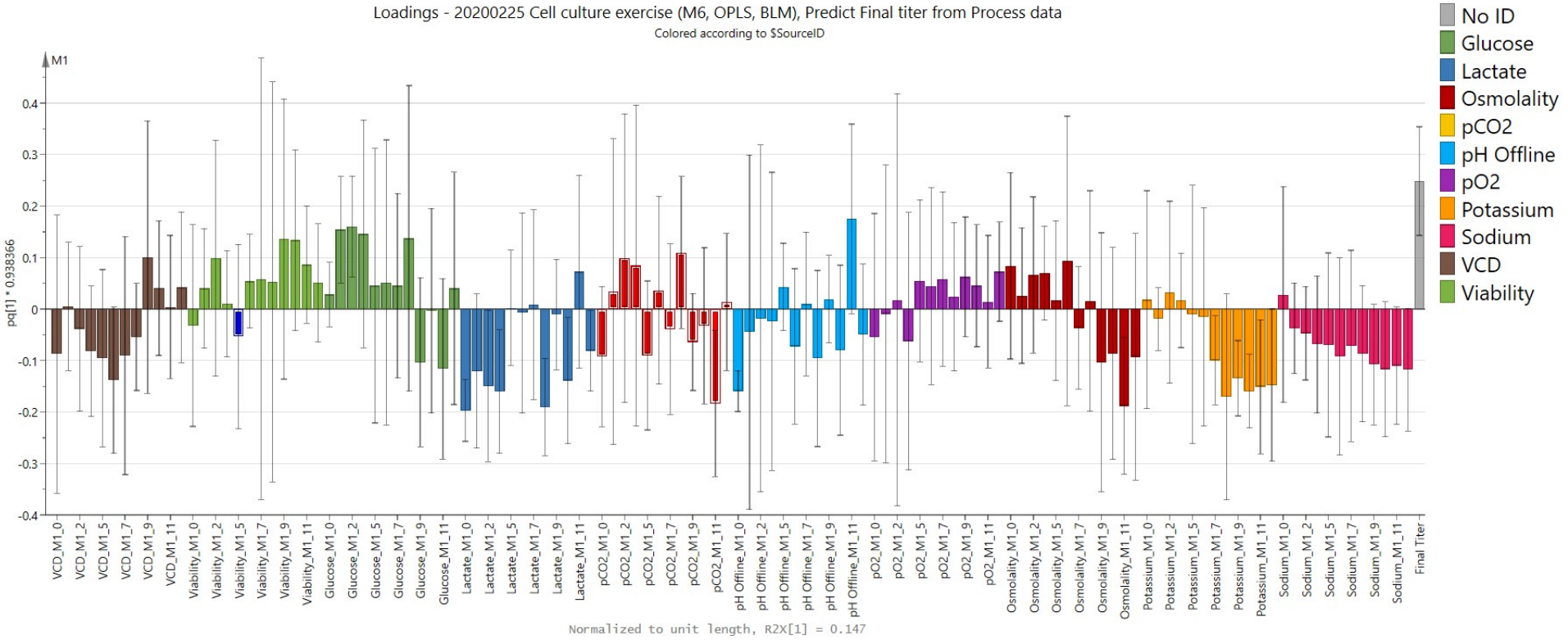
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Primary ID	$BatchID	Number	VCD_M1_0	VCD_M1_1	VCD_M1_2	VCD_M1_3	VCD_M1_5	VCD_M1_6	VCD_M1_7	VCD_M1_8	VCD_M1_9	VCD_M1_10	VCD_M1_11	VCD_M1_12	Viability_M1_0	Viability
2	$PhaseID			M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1	M1
3	$MaturityID			0	1	2	3	5	6	7	8	9	10	11	12	0	1
4	$SourceID			VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	VCD	Viability	Viability
5	$Batch_2	Batch_2	1	0.32	0.822001	2.241	5.467	15.9	21.3	23.2	23.3	21.1	19.3	17.5	13.4	96.5	
6	$Batch_4	Batch_4	2	0.324	0.612	1.589	4.126	11.35	16.75	20.128	22.0975	22.689	21.9	18.2175	15.8915	99.1	
7	$Batch_5	Batch_5	3	0.337001	0.624002	1.754	4.314	12.529	17.323	20.44	20.9015	23.429	18.206	16.404	14.274	99.1	
8	$Batch_7	Batch_7	4	0.324	0.713	2.036	5.492	14.1415	20.1197	23.984	24.017	21.848	19.488	18.207	14.51	98.2	
9	$Batch_8	Batch_8	5	0.31	0.63	1.829	4.872	13.635	20.702	22.644	24.0935	23.251	19.175	18.38	13.958	99	
10	$Batch_9	Batch_9	6	0.312	0.68	2.7875	4.895	13.404	20.531	21.4395	22.4215	21.175	19.24	17.305	14.442	99	
11	$Batch_10	Batch_10	7	0.329	0.679001	3.158	5.637	14.544	20.249	22.398	21.5915	19.2265	17.9478	16.669	13.816	99.4	
12	$Batch_12	Batch_12	8	0.308002	0.677	1.904	4.917	12	17.9	20.249	22.5	21.6	19.62	18.8	16.633	99.5	
13	$Batch_3	Batch_3	10	0.317	0.719	1.94	4.401	13	18.4	22.2	21.6	22.059	19.7	17.8	15.8	97.3	
14	$Batch_11	Batch_11	11	0.367002	0.803	2.38	5.807	15.6	22	26.846	24.5	23.8	21.7	19.6	16.2	99.3	
15	$Batch_1	Batch_1	13	0.289001	1.157	4.604	7.111	16.001	21.534	24.66	24.029	23.2	21.3	20.723	16.709	99	
16	$Batch_6	Batch_6	14	0.307	0.923	2.768	5.235	15.8	20.9	25.7	25.4	23.87	21.34	20.2	17.7	98.2	

$BatchID	DS3.Final Titer
Batch_1	3.6895
Batch_2	3.29939
Batch_3	3.86063
Batch_4	4.04247
Batch_5	4.03154
Batch_6	3.37
Batch_7	3.43495
Batch_8	3.32718
Batch_9	3.01178
Batch_10	3.14113
Batch_11	2.78789
Batch_12	2.90643

Observed vs Predicted



Loadings, Low Lactate, Potassium and Sodium Correlate With High Titer



Demo

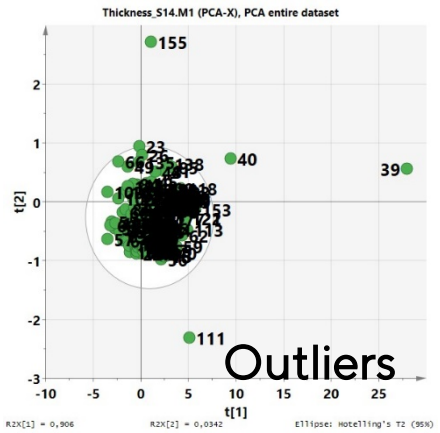
Conclusions

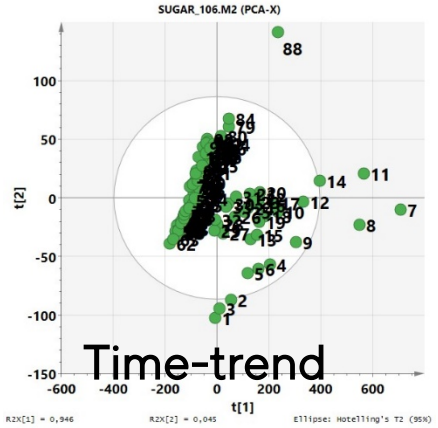
- PCA is a method to summarize the data
  - Recommended method to start a multivariate investigation
- The scores are the new variables summarizing the original ones
- Plotting the scores gives a picture of the observations as projected into the model plane
  - Outliers in scores should be scrutinized
- The loadings show the importance of the variables and help us understand (interpret) the scores
- Residuals give information on:
  - How well the scores summarize the variables
  - Which variables are well explained
  - How far the observations are from the model plane

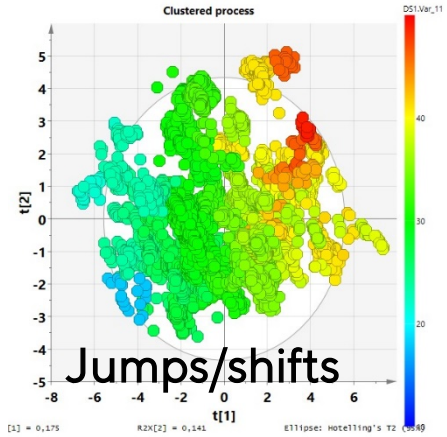
Conclusion

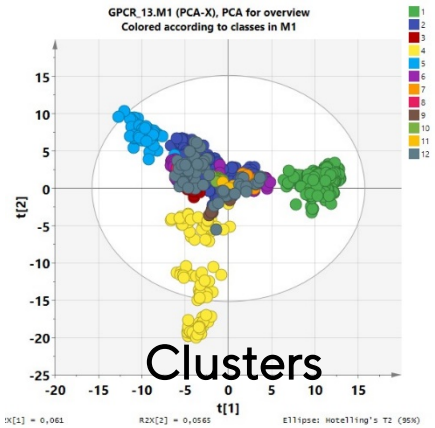
- The big value the presented methods can bring to you & your business is
- Deeper process understanding from your Ambr experiments
  - leading to better optimisation overall
- Accelerate your development workflow by able to get a more holistic view on all your data and thereby making more informed decisions
- In this example the measurable improvement is titre
  - and an understanding of what parameters that correlate to a higher titre
  - other examples have been prediction of Glycosylation pattern to select the best cell line
  - prediction of charge profile, charge variants

Summary: What Can It Look Like and What Does It Mean?



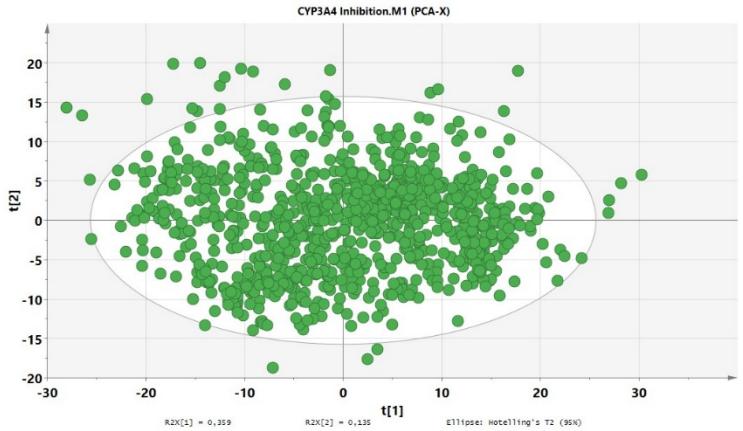






Ideal case to strive for: ***Fairly even scatter of points!***

- Model not disturbed by outliers
- Homogenous mechanism
- Reliable interpretation of parameters & limits as many points participate in model building



Upcoming Webinars (https://www.sartorius.com/en/company/exhibition-conferences)