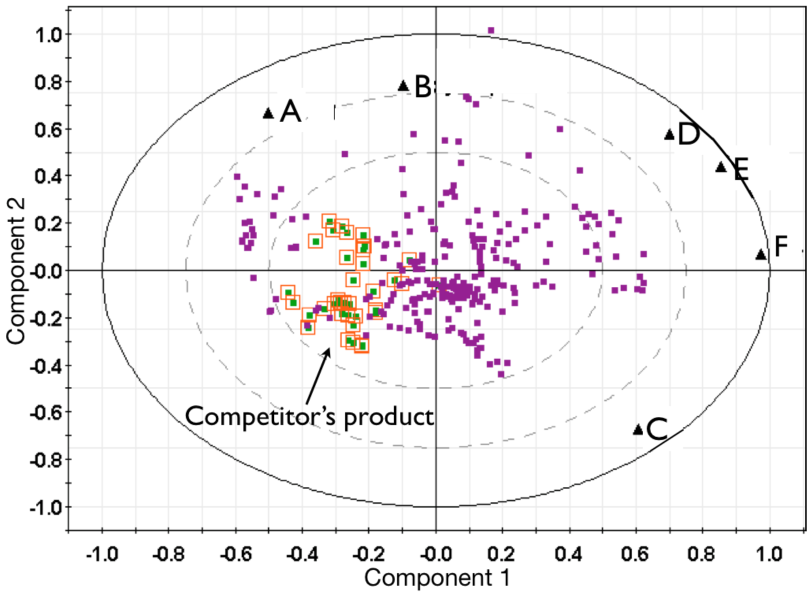
Chemical Engineering 4H03

Introduction Latent Variables

Jake Nease

McMaster University



Objectives

- Latent variables can be powerful modeling tools
  - What are they?
  - What are they used for?
  - How do we interpret them?
- How are latent variables calculated?
  - Computing a LV score from a known model
  - Geometric interpretation
- How do we **train** models to identify latent variables?
  - We'll bust out the math in the next section

Warm-Up

- Turn to your neighbour and try to answer these:
  - What do **you** interpret to be a latent variable?
  - Can you think of any examples from industry/university?

Basics of Latent Variables

The truth is out there

https://consequenceofsound.net/wp-content/uploads/2018/02/chris-carter-x-files-without-scully.jpg?quality=80

Definition

- A **LATENT VARIABLE (LV)** is defined as any variable that is not directly observed
  - Since it is not observed, it must be *constructed* based on measurements of other (often correlated) variables

- Example: your **health** is a latent variable
  - Blood pressure
  - Weight
  - Body proportions
  - Temperature
  - Bloodwork (cell counts *etc.*)
  - Living habits (drinking, exercise, smoking, sedentary…)

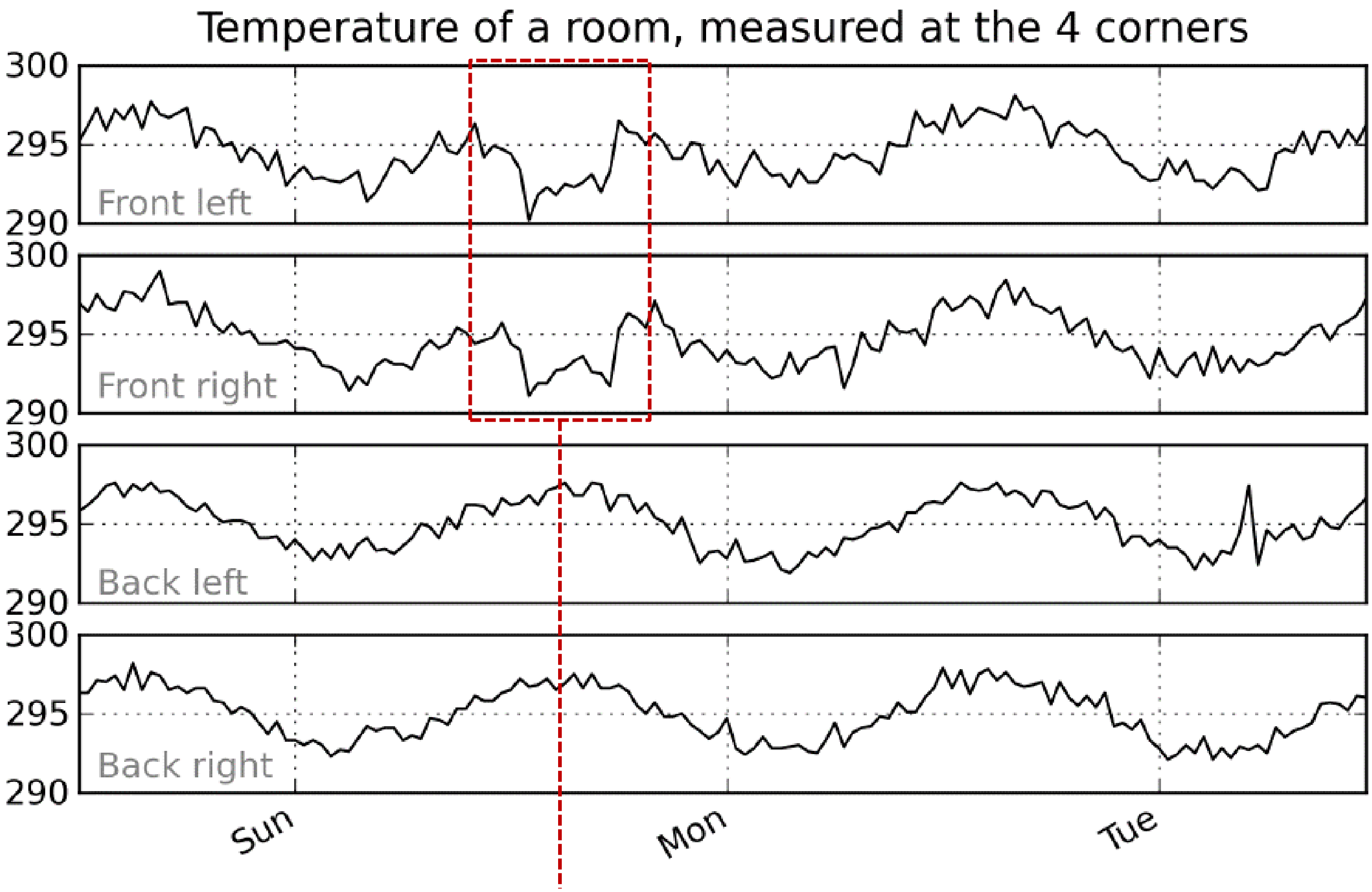
Fun fact – women are healthier than men!
I have a story about that…

- Can we combine these measurements?
  - We sure can! A doctor does this mentally

Definition from Young. *Handbook of Regression Methods*. CRC Press

LV Example

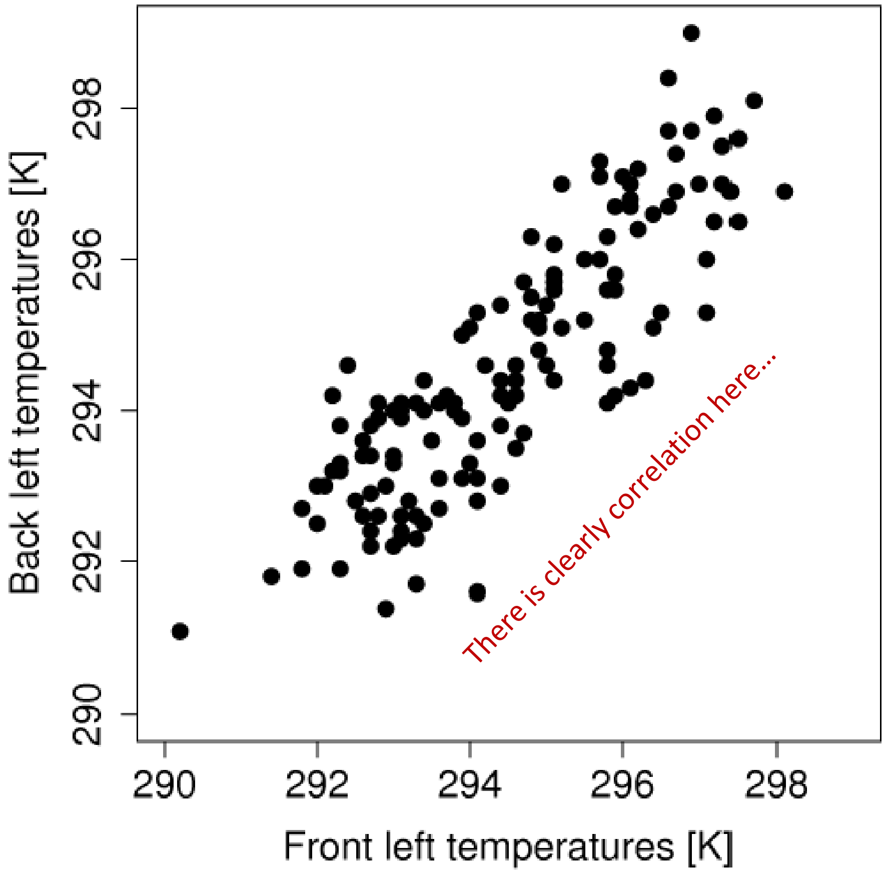
- Temperature in the room, measured at several points



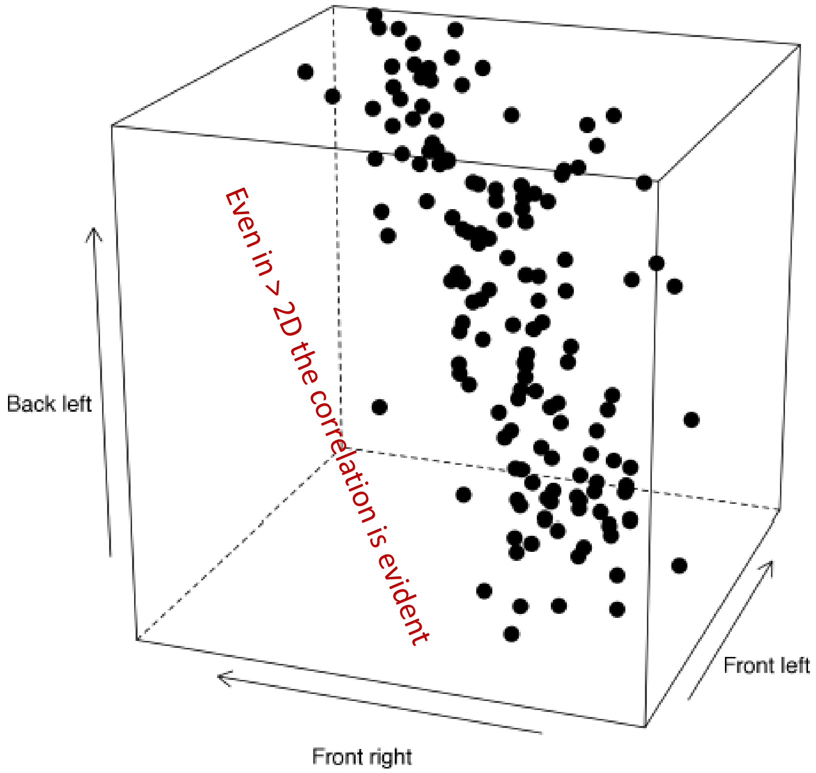
What's up with this? Use your visualization skills!

LV Example

- Temperature in the room, measured at several points



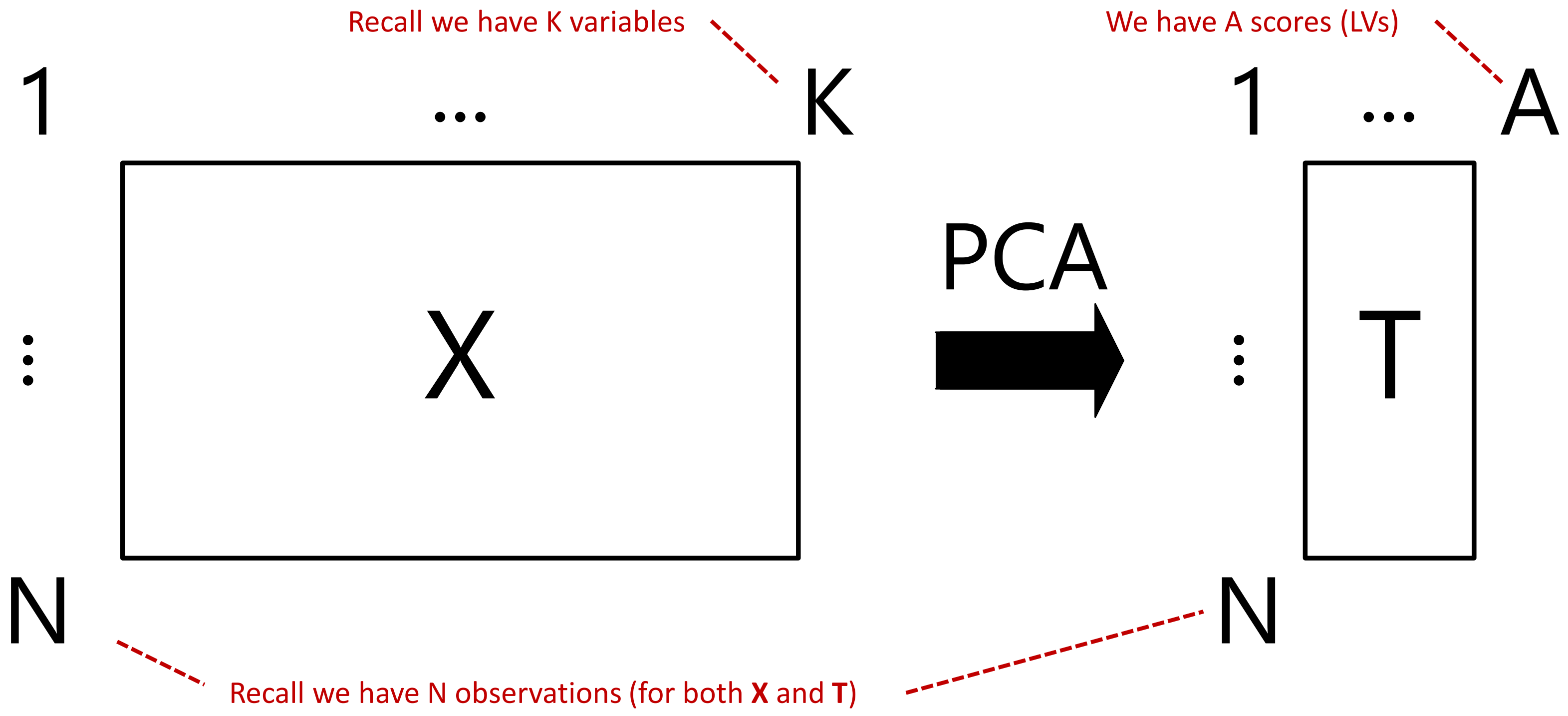
There is clearly correlation here…



Even in > 2D the correlation is evident

Principal Component Analysis (PCA)

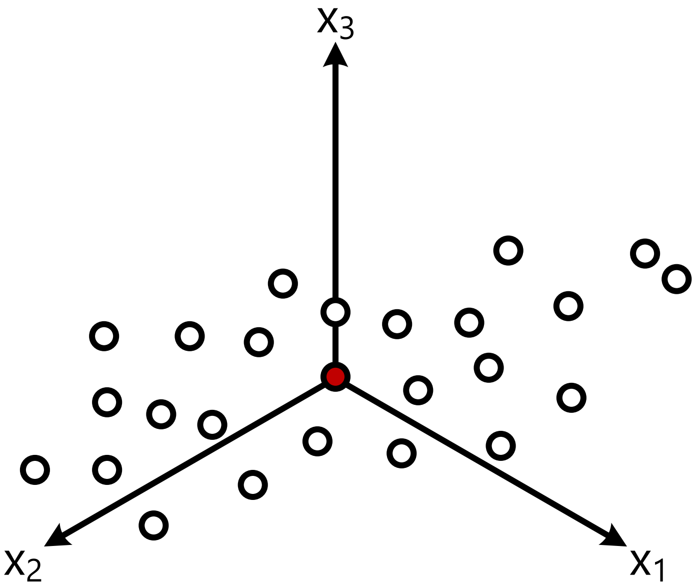
- **Mathematical Objective**
  - Find the best summary of data **X** using the **fewest** number of "summary variables"
  - These "summary variables" are known as the scores, **T**

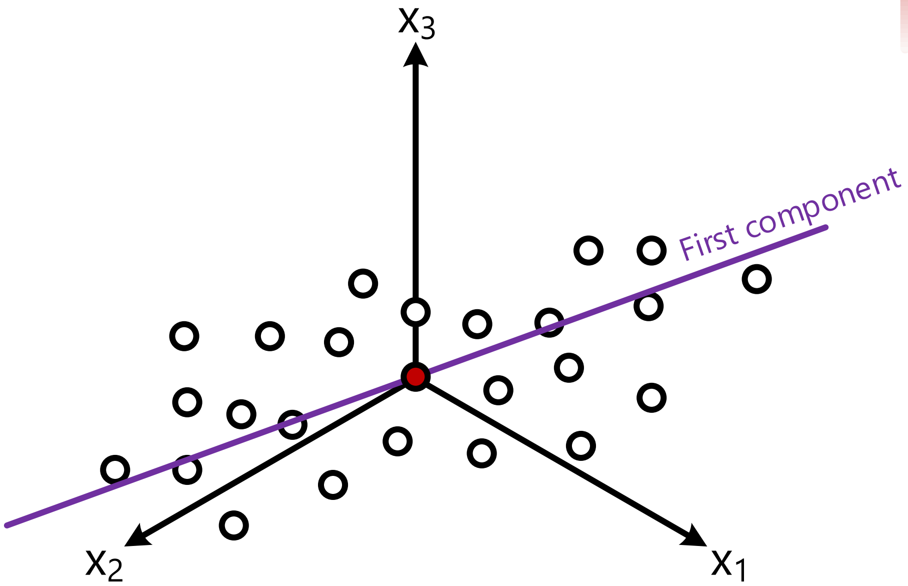

Recall we have K variables
We have A scores (LVs)
1
…
K
1
…
A
X
PCA
T
⋮
⋮
N
N
Recall we have N observations (for both **X** and **T**)

What Does PCA Do?

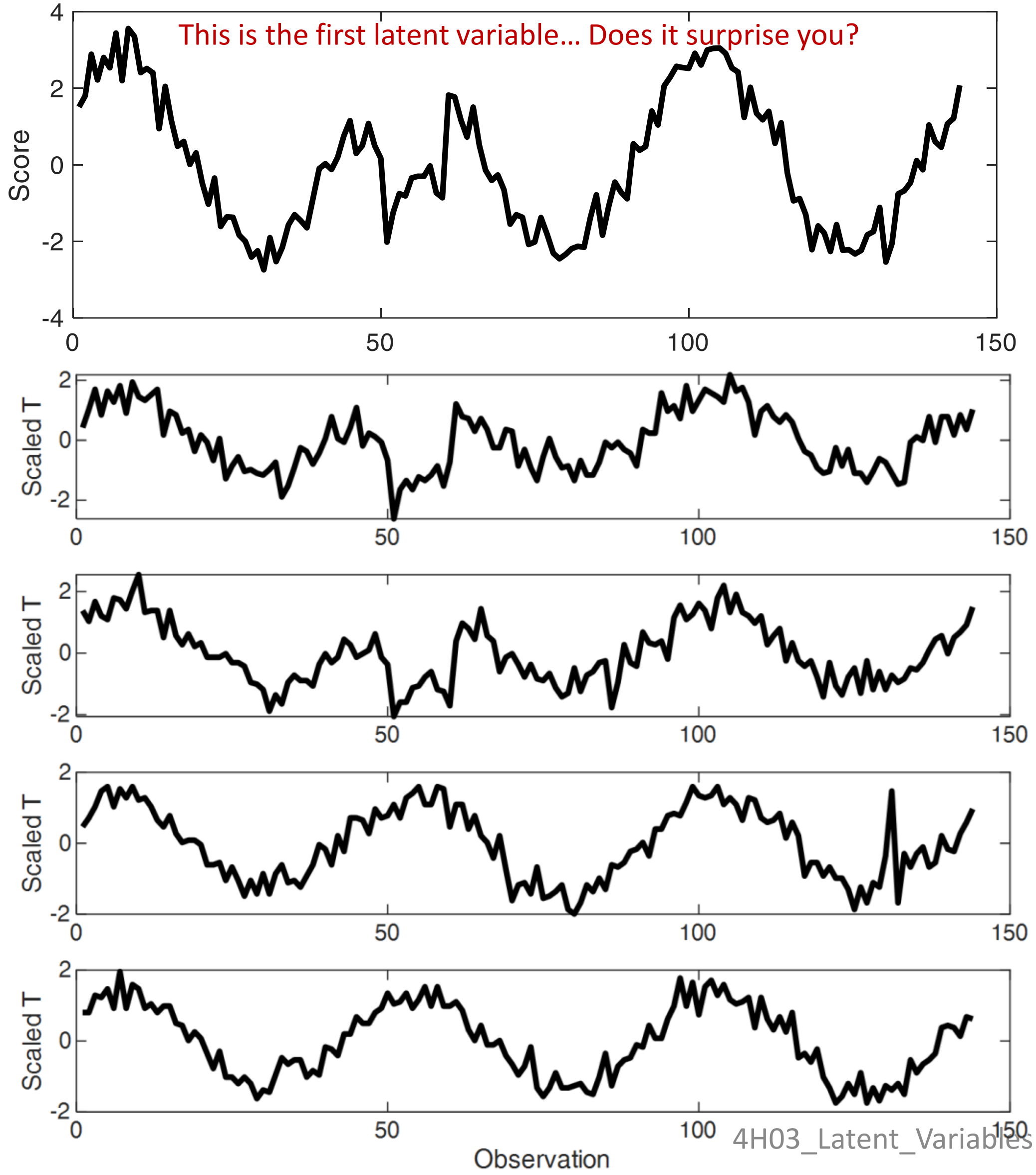
- It finds the directions that best explain variance
  - “Directions of greatest variance”
  - “Loadings → Scores”
  - “Components”
  - ‘Latent Variables”

- Component (LV) 1 explains the most variance. Adding further components exhibits diminishing returns but still adds to fidelity





PCA on Temperature Data



This is the first latent variable… Does it surprise you?

These are the same temperatures after **centering** and **scaling** the data (more on this in the next lecture set)

Calculating Scores

- FAR More on this later…
- Generally, a score (t) is computed as the product of an observation (x) and it's associated loadings (p) in the LV space
  - Effectively, the *loadings* are how much each measurement in x affect the result in t. In our example:

$$t_1 = 0.25x_1 + 0.25x_2 + 0.25x_3 + 0.25x_4$$

$$t_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} \qquad t_1 = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix} \begin{bmatrix} p_{1,1} \\ p_{2,1} \\ p_{3,1} \\ p_{4,1} \end{bmatrix}$$

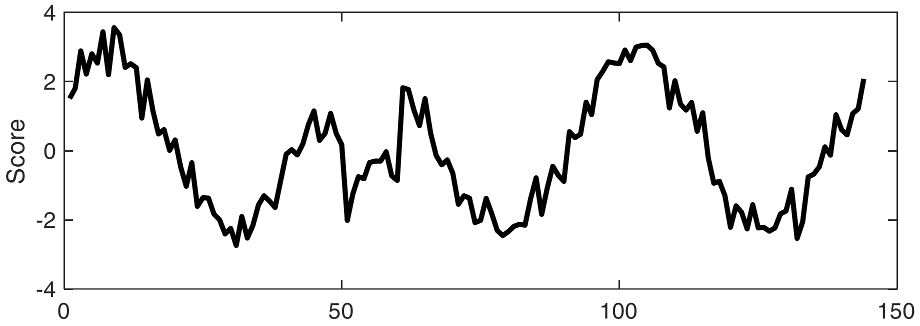
$$\boxed{t_1 = \boldsymbol{x}^T \boldsymbol{p}_1}$$

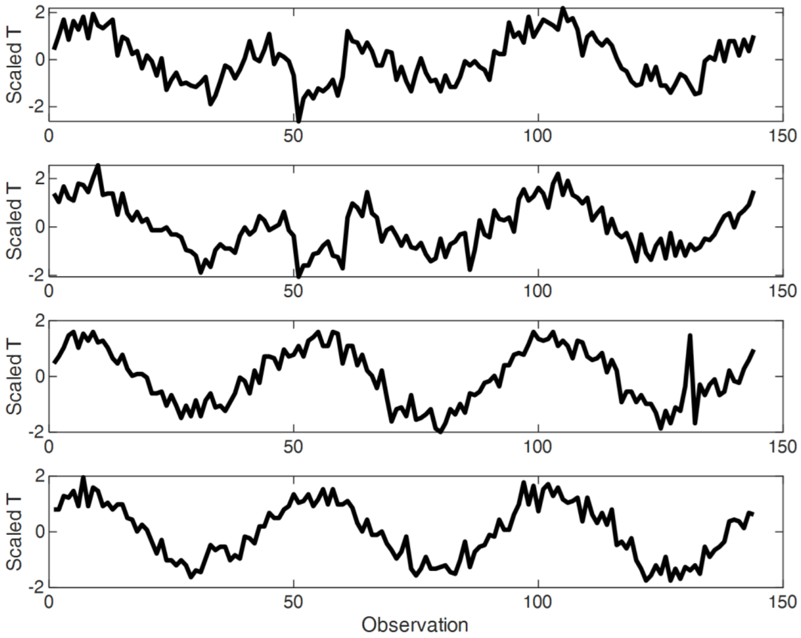
p_j is the **loading** vector of component *j*

$p_{i,j}$ is the loading (contribution) of x_i in the j^{th} component (latent variable)

Calculating Scores

- Workshop: Given the data and the first latent score, how do you think the **second** latent score will look?
  - Hint: recall that LVM tries to explain the greatest *variance*

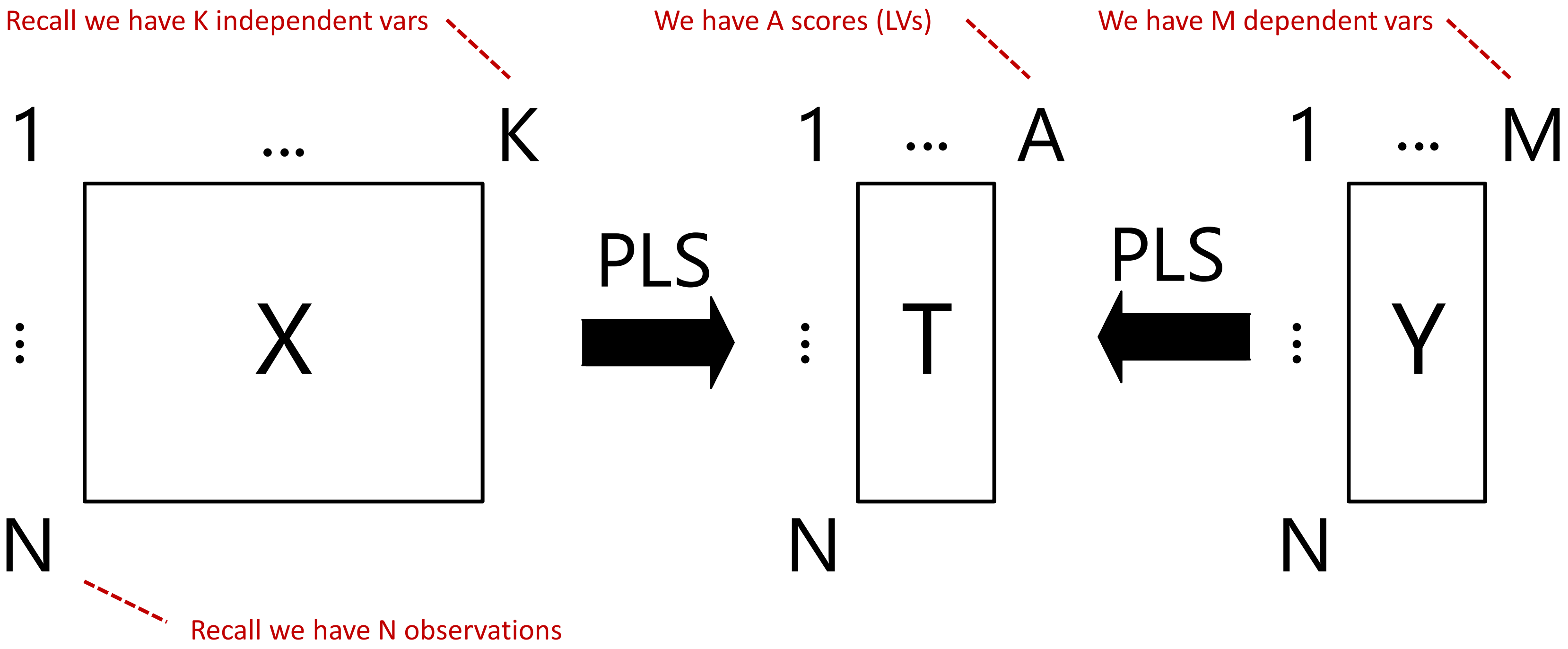




What is the next greatest source of variance in this data?

Projection of Latent Structures (PLS)

- **Mathematical Objective**
  - Find the best summary of data **X** AND the best summary of my data **Y** using a set of summary variables, **T**, so that **T** can also be used to **predict Y** given some values of **X**

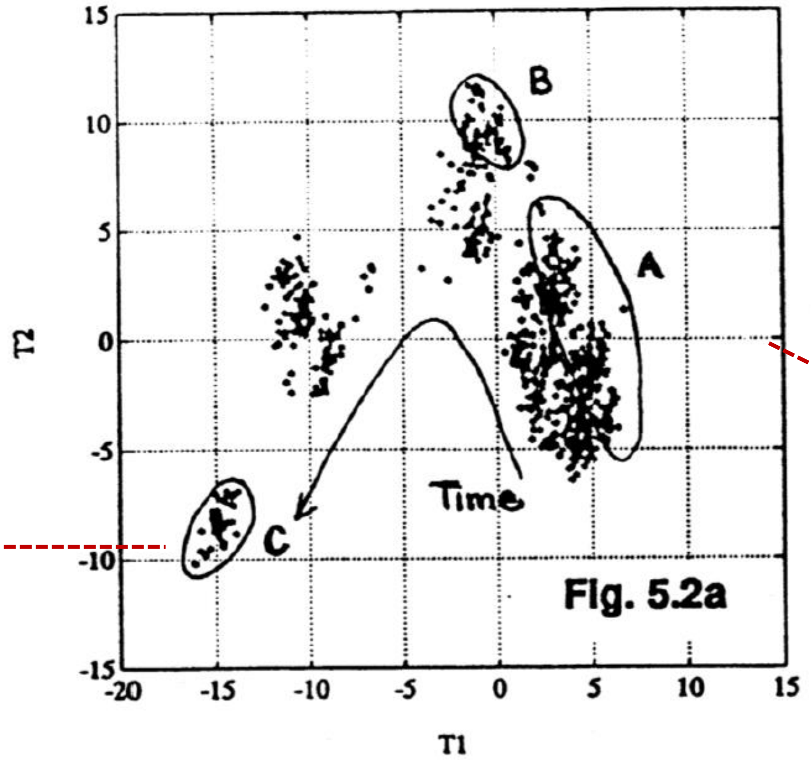


Applications of Latent Variables

Seeing is Believing

Learning from Data

- **Identifying process drift**
  - Performance of MANY variables in a chemical (or other) process can be visualized in a score plot, with each observation throughout time encoded to show trends



The LV scores are clearly drifting with time

Could be useful for equipment monitoring, utility/resource consumption monitoring…

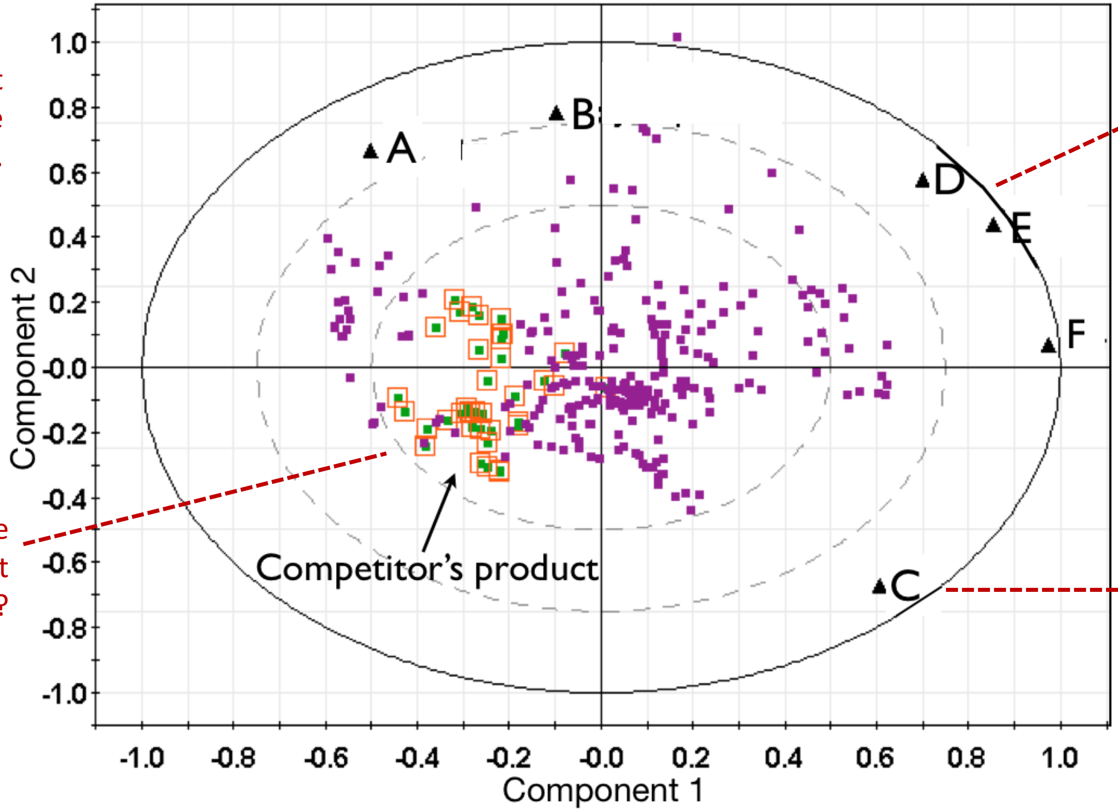
Also a terrific application of clustering (more on this later?)

Figures courtesy of ConnectMV

Learning from Data

- **Which variables are correlated?**
  - Can visualize variability
  - Can see variables that behave “together”
  - My competitor has higher prices/market share. Why?

Let’s take a moment to discuss what we are looking at here…



These Δs are actually a visualization of the **loadings** as they contribute to the **scores**!!

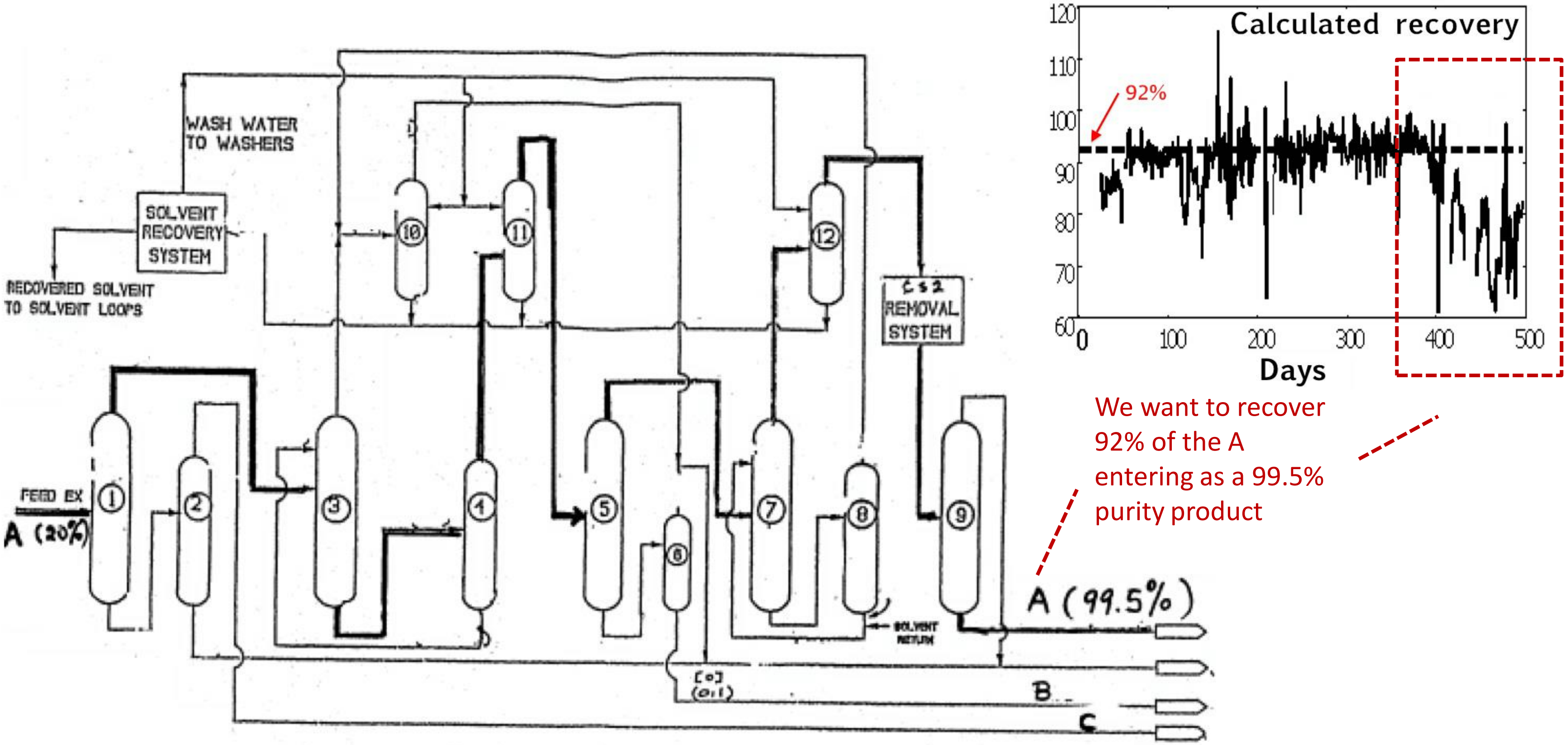
If I want to replicate my competitor, what can I do?

How would you describe the relationship of variable C to the others?

Figures courtesy of ConnectMV

LVM for Troubleshooting

- **Why is my process not meeting recovery targets?**
  - ~ 450 tags measured for 500 days of operation
  - After ~ 400 days, recovery fell below targets

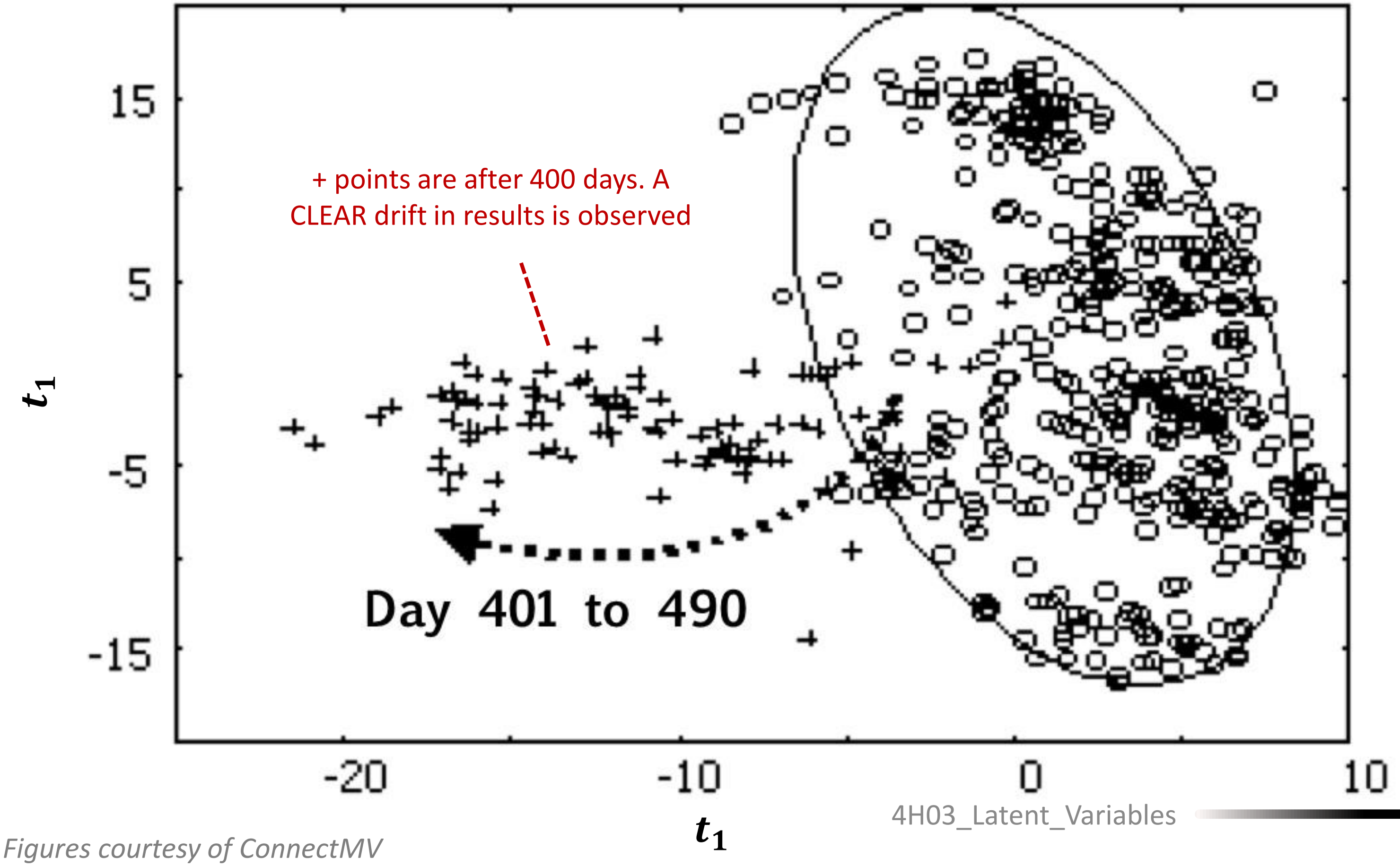


We want to recover 92% of the A entering as a 99.5% purity product

Figures courtesy of ConnectMV

LVM for Troubleshooting

- Trained a LV model with two variables
  - Compressed ~450 variables to **two**
  - A lot of information was retained

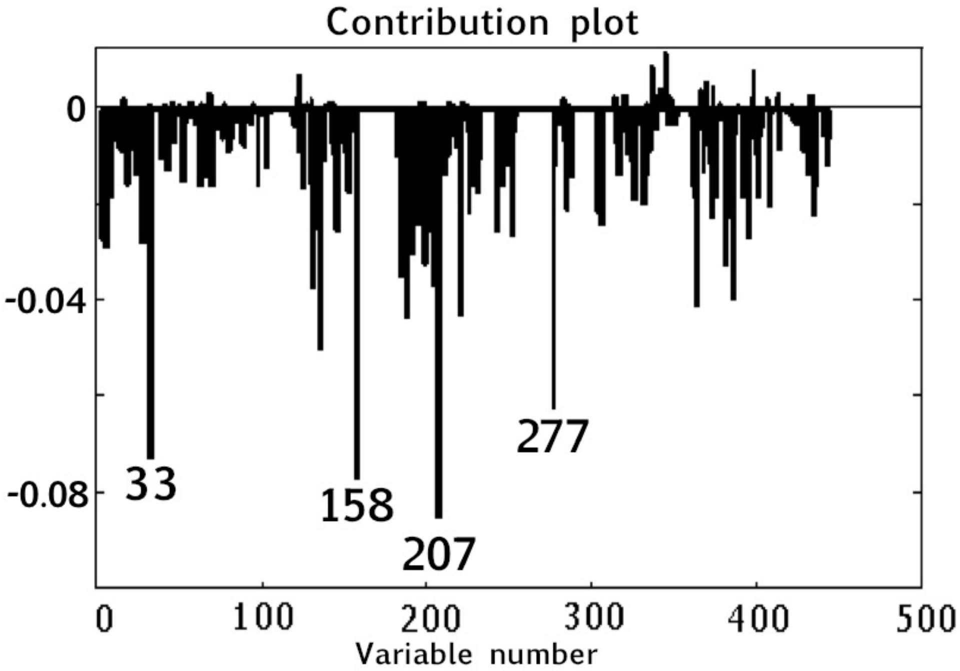


+ points are after 400 days. A CLEAR drift in results is observed

Figures courtesy of ConnectMV

LVM for Troubleshooting

- The question becomes… **What causes LOW t_1 scores**?
  - Examine the **loadings** (p) via a contribution plot
  - HIGH loadings might flag variables that are making t_1 drop!



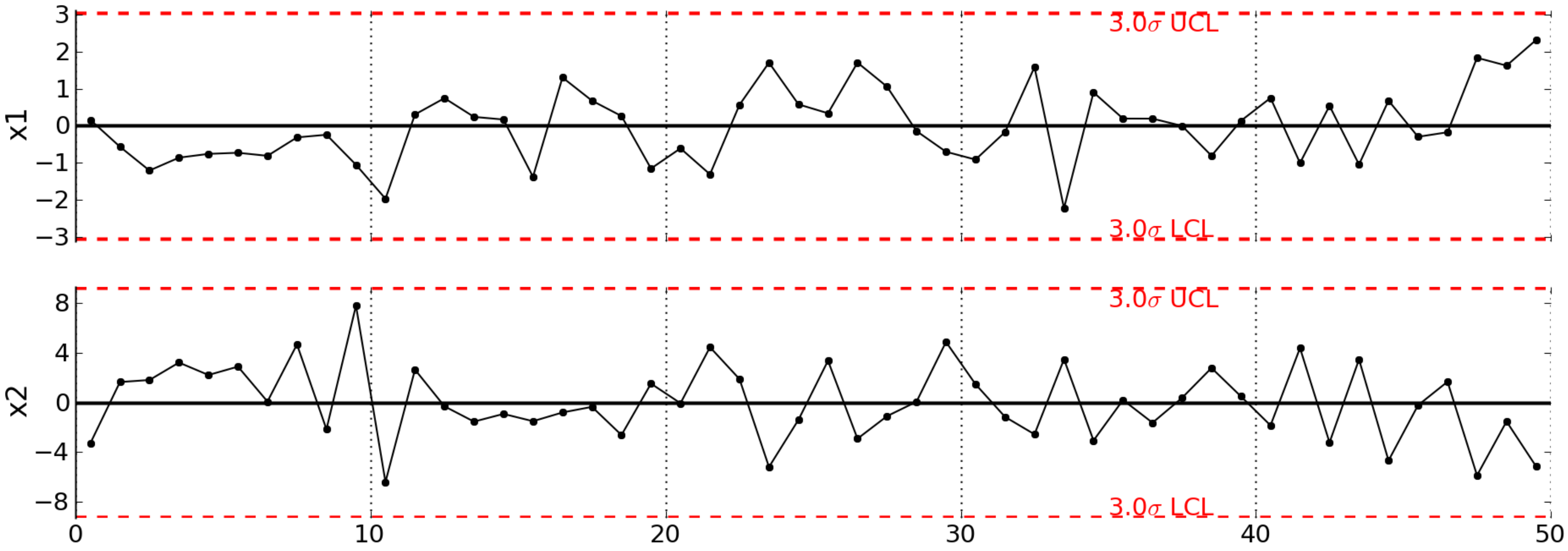
- **207**: temperature on a tray near bottom of column 3
- **158**: another process measurement from column 3
- **33** and 277: related to feed concentration of component A targeted for recovery

- Suggests bad temperature control in column 3 when feed concentration is high
  - Fixed controller (sensor drift), process returned to normal

Figures courtesy of ConnectMV

LVM for Process Monitoring

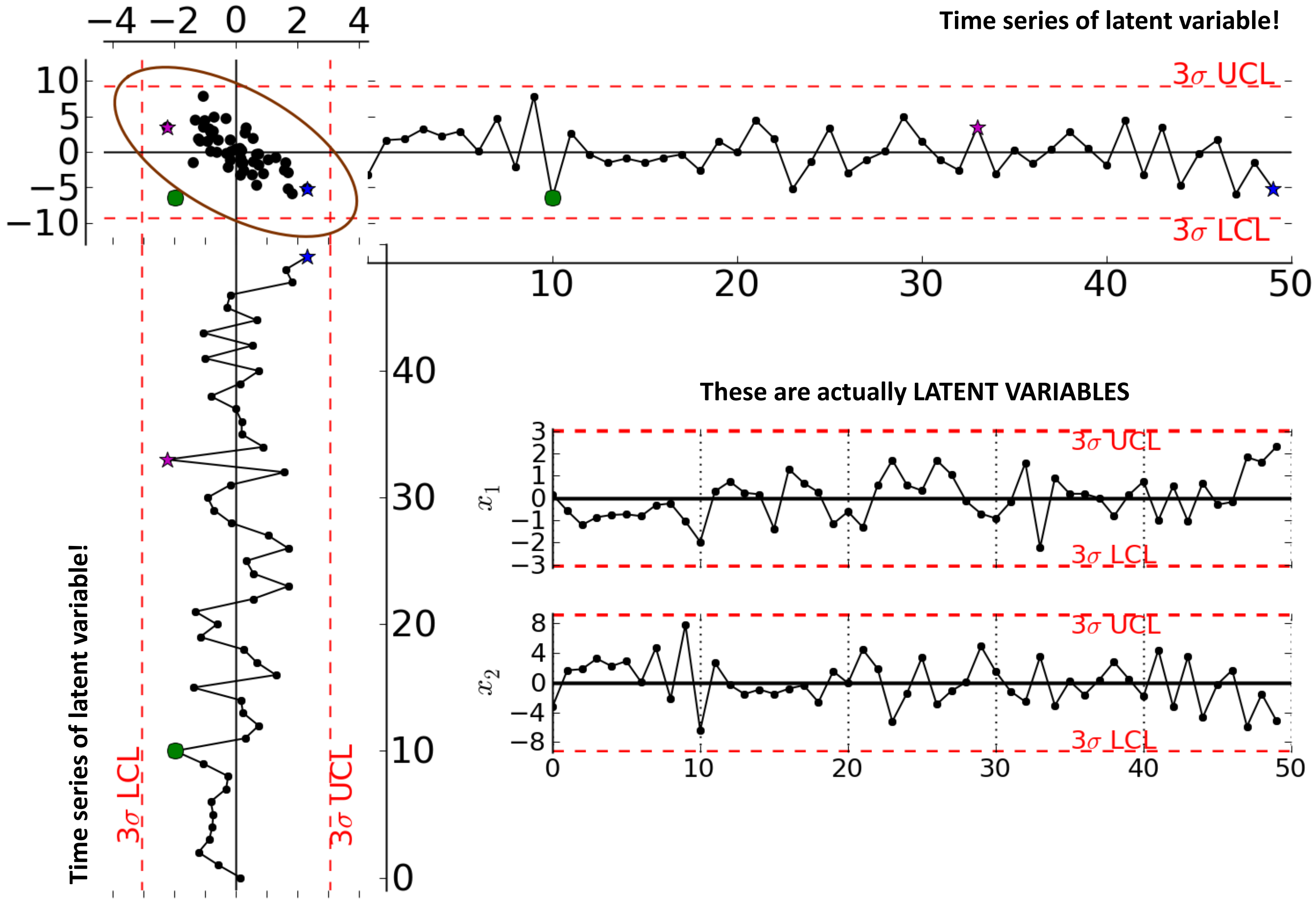
- Any variable can be monitored (T, P, vibration…)
  - Example for two variables:
  - Called "soft sensors"



Figures courtesy of ConnectMV

LVM for Process Monitoring

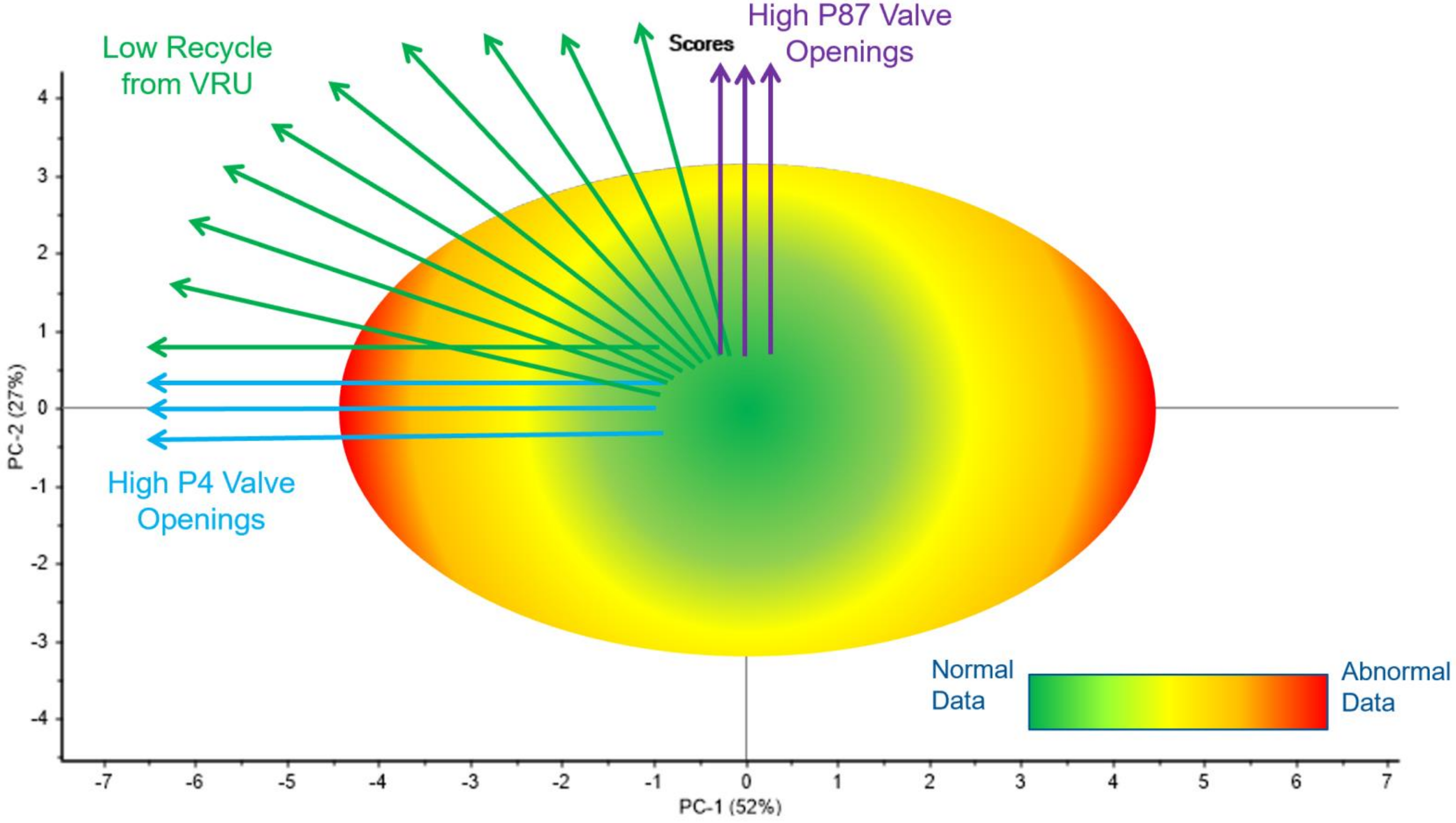
- Can visualize SCORES and search for deviations



Figures courtesy of ConnectMV

LVM for Process Monitoring

- Wonderful example from Sasha Korp!
  - McMaster ChE student on internship at Suncor
  - Monitoring process variables related to **venting incidents**

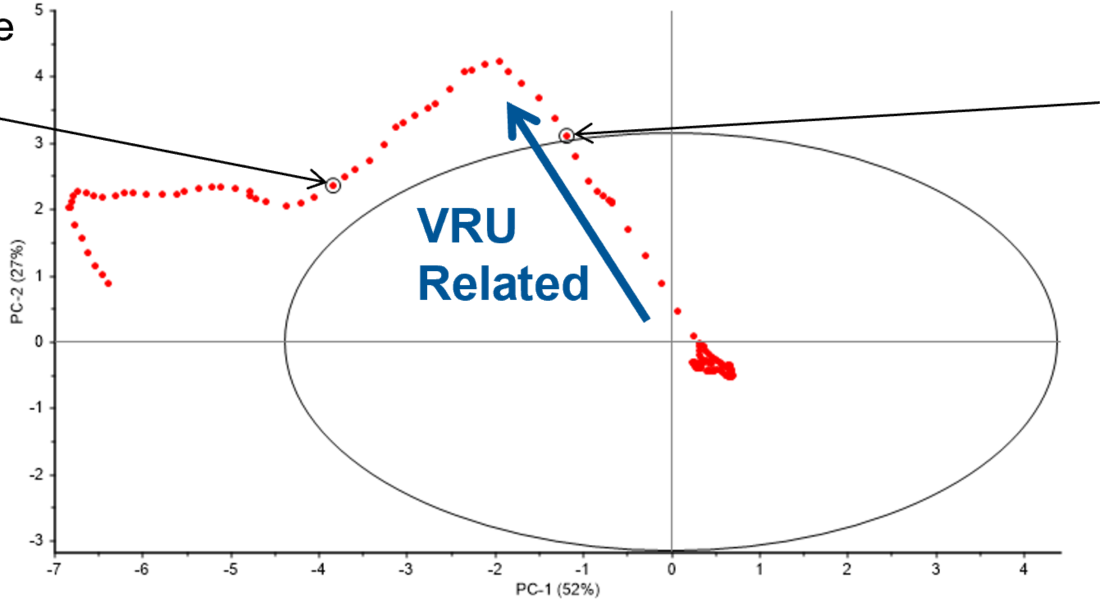


Figures courtesy of Sasha Korp

LVM for Process Monitoring

- Wonderful example from Sasha Korp!
  - Process variables monitored in 99% confidence interval
  - Process deviated from confidence interval
  - 3 minutes later, venting incident was experienced!

Venting valve opens
15-Feb
3:21:30 AM



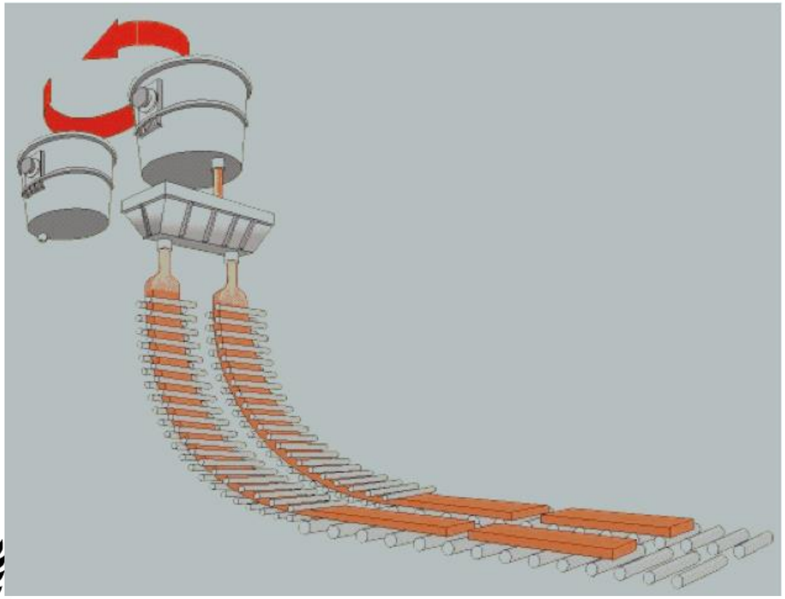
Outside 99%
Confidence Interval
15-Feb 3:18:20 AM

Figures courtesy of Sasha Korp

LVM for Process Monitoring

- ArcelorMittal Dofasco has used LVM process monitoring tools since the 90s

- Most well known is the casting monitoring application
  - Caster SOS (stability operation supervisor)
  - A multivariate monitoring system in disguise!



Figures courtesy of ConnectMV

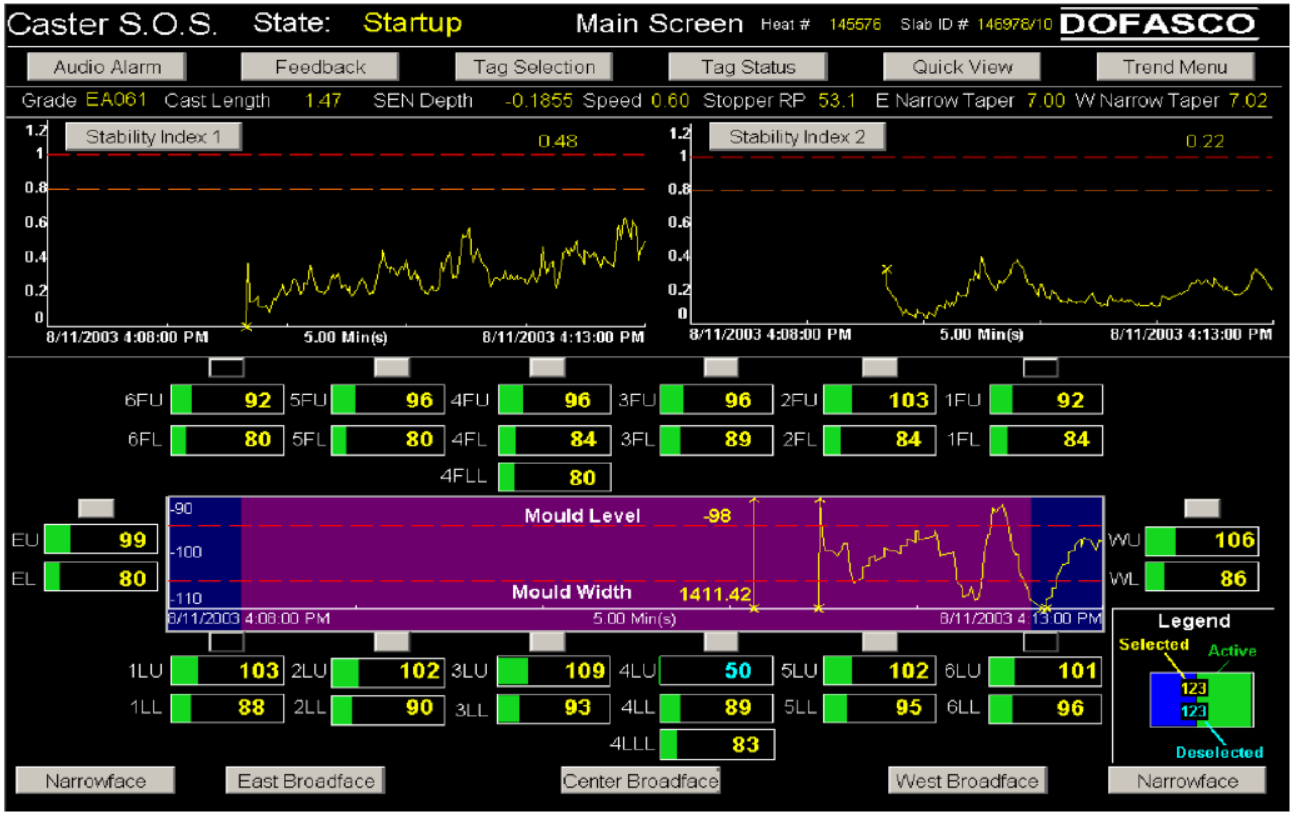
LVM for Process Monitoring

- Improper cooling times can cause **breakouts**
  - Outer shell ruptures, splashing liquid metal all over!
  - A huge safety and production concern ($200,000+)

Figures courtesy of ConnectMV

LVM for Process Monitoring

- Process monitoring software creates timeseries plots of so-called **stability indexes**
  - But really, these stability indexes are just LVs known to contribute strongly to a higher chance of breakout!



Caster S.O.S. State: Startup Main Screen Heat # 145576 Slab ID # 146978/10 DOFASCO

Grade EA061 Cast Length 1.47 SEN Depth -0.1855 Speed 0.60 Stopper RP 53.1 E Narrow Taper 7.00 W Narrow Taper 7.02

Stability Index 1 0.48

Stability Index 2 0.22

6FU 92 5FU 96 4FU 96 3FU 96 2FU 103 1FU 92

6FL 80 5FL 80 4FL 84 3FL 89 2FL 84 1FL 84

4FLL 80

EU 99

EL 80

Mould Level -98

Mould Width 1411.42

WU 106

WL 86

1LU 103 2LU 102 3LU 109 4LU 50 5LU 102 6LU 101

1LL 88 2LL 90 3LL 93 4LL 89 5LL 95 6LL 96

4LLL 83

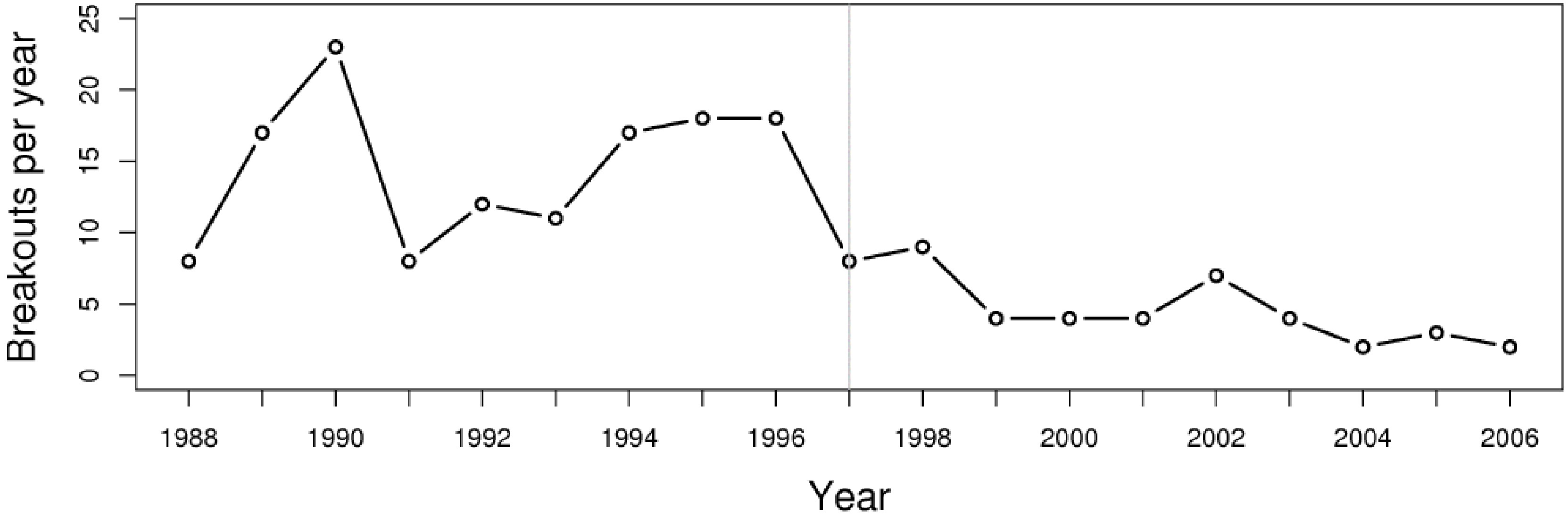
Tell your printer I'm sorry…

Contains ALARM limits! When alarms sound, contributions show to help operator understand what to change to reduce breakout potential

Figures courtesy of ConnectMV

LVM for Process Monitoring

- Implemented in 1997, data available to 2006
  - SIGNIFICANT reduction in breakouts due to better operator preparedness and much simpler monitoring system
  - Over $1M saved in first year alone



Figures courtesy of ConnectMV

Additional Applications

- Literature is FULL of great LVM applications
  - Personality classifications
  - Snack food coatings
  - Sensors to predict food spoilage
  - Forecasting electricity demand

- Lots of wonderful literature available
  - Review of LVMs for process control

Figures courtesy of ConnectMV

Final Words

- There are many applications of LVMs in engineering
  - Improved understanding
  - Troubleshooting
  - Soft sensors/predictive modeling
  - Process monitoring
  - Reverse engineering