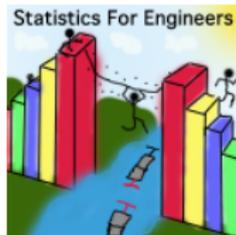


Statistics for Engineers

ChE 4C3 and 6C3



© Kevin Dunn, 2013

kevin.dunn@mcmaster.ca

<http://learnche.mcmaster.ca/4C3>

Overall revision number: 83 (April 2013)

Copyright, sharing, and attribution notice

This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License. To view a copy of this license, please visit

<http://creativecommons.org/licenses/by-sa/3.0/>



This license allows you:

- ▶ **to share** - to copy, distribute and transmit the work
- ▶ **to adapt** - but you must distribute the new result under the same or similar license to this one
- ▶ **commercialize** - you are allowed to use this work for commercial purposes
- ▶ **attribution** - but you must attribute the work as follows:
 - ▶ “Portions of this work are the copyright of Kevin Dunn”, *or*
 - ▶ “This work is the copyright of Kevin Dunn”

(when used without modification)

We appreciate:

- ▶ if you let us know about **any errors** in the slides
- ▶ **any suggestions to improve the notes**

All of the above can be done by writing to

`kevin.dunn@mcmaster.ca`

or anonymous messages can be sent to Kevin Dunn at

<http://learnche.mcmaster.ca/feedback-questions>

If reporting errors/updates, please quote the current revision number: 83

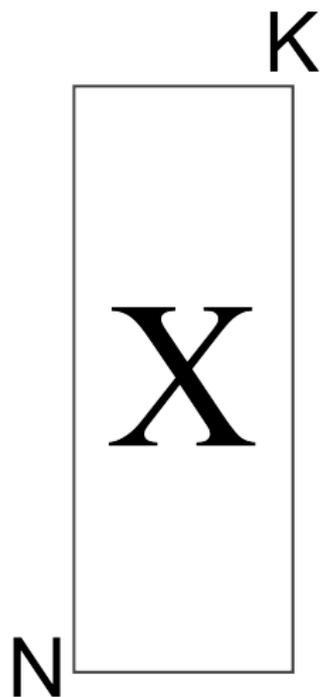
Please note that all material is provided “as-is” and no liability will be accepted for your usage of the material.

Extracting value from data

Engineers deal with **large quantities** of data from many **different sources**. We can do some some interesting *and profitable* things with these data:

1. **Improve process understanding**
 - ▶ Visualize it in interesting ways
 - ▶ Test if there are significant differences
 - ▶ Running designed experiments
2. **Troubleshooting process problems**
 - ▶ Patterns in monitoring charts/visualizations
 - ▶ Diagnose outliers in least squares models
3. **Predictive modelling**
 - ▶ LDPE physical properties (assignment 6)
 - ▶ LDPE melt index predicted from temperature
4. **Process monitoring**
 - ▶ Kappa number data set (assignment 4)
5. **Process optimization and improvement**
 - ▶ Designed experiments
 - ▶ Response surface methods

Types of data you will see

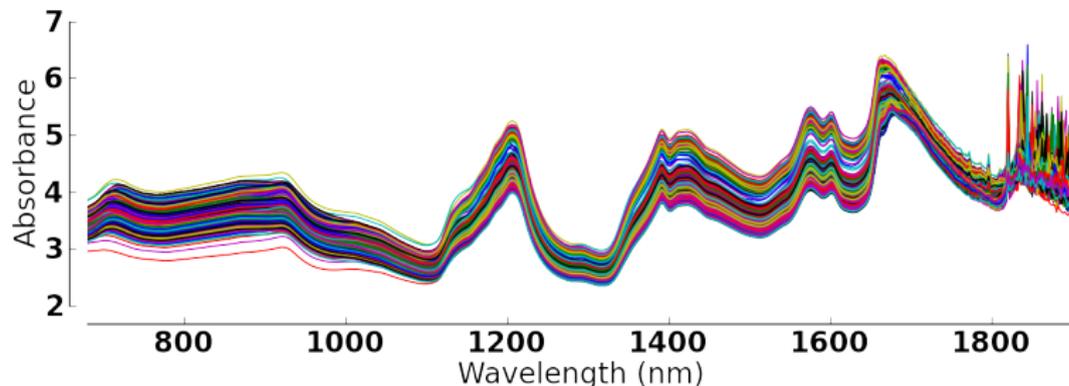


1920's to 1950's:

- ▶ small number of columns, K
- ▶ scatter plots (few combinations)
- ▶ time-series plots for each column
- ▶ Shewhart and EWMA charts
- ▶ multiple linear regression (MLR)
- ▶ carefully chose which columns to measure
 - ▶ independent
 - ▶ low error

Types of data you will see

- ▶ **Small N and small K**
 - ▶ expensive measurement, low frequency
 - ▶ use scatterplots, linear regression, *etc.*
- ▶ **Small N and large K**

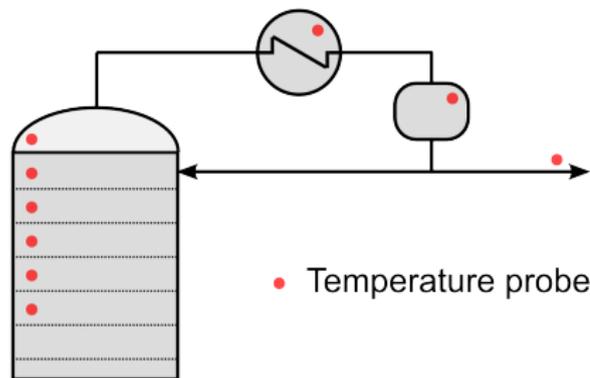


- ▶ Cannot use linear regression directly: because $k > n$

Types of data you will see

► Large N and small K

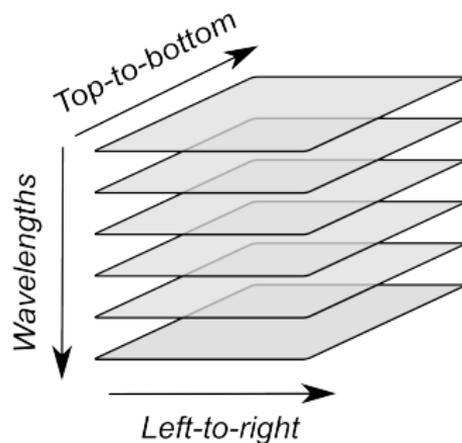
- Refinery, most chemical plants
- 2000 to 5000 variables (called tags) every second
- 50 to 100 Mb collected per second



35 temperatures, 5 to 10 flow rates, 10 pressures, 5 derived values

Types of data we deal with

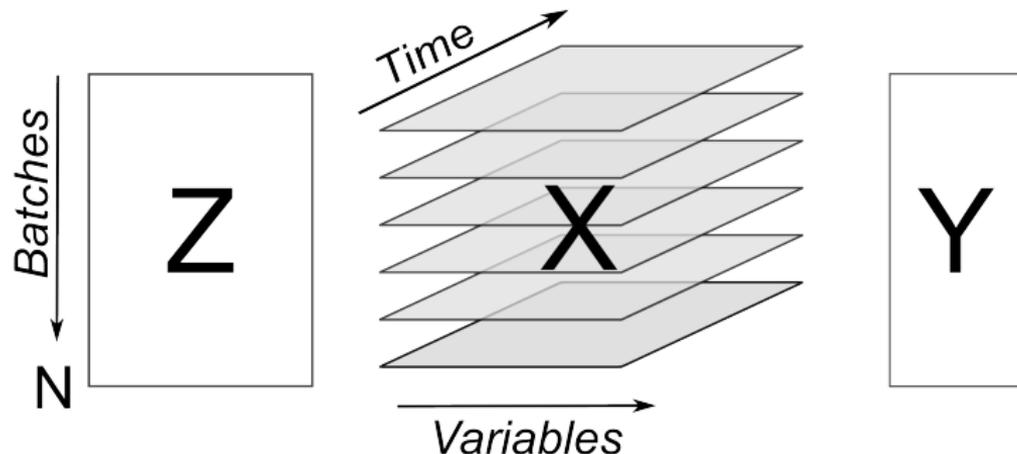
- ▶ **3D data sets and higher dimensions**
- ▶ Very common situation now
- ▶ Image data (medical imaging)



- ▶ 4th dimension: time
- ▶ Very high redundancy: neighbouring pixels are similar (spatially and in time)

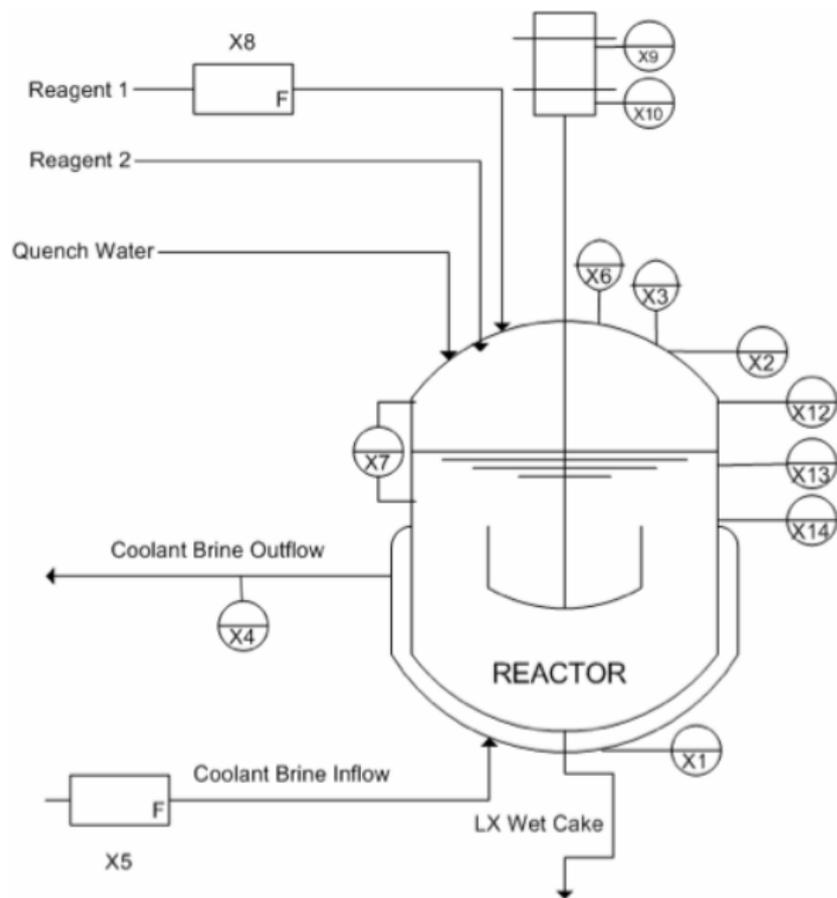
Types of data we deal with

- ▶ **Batch data sets**

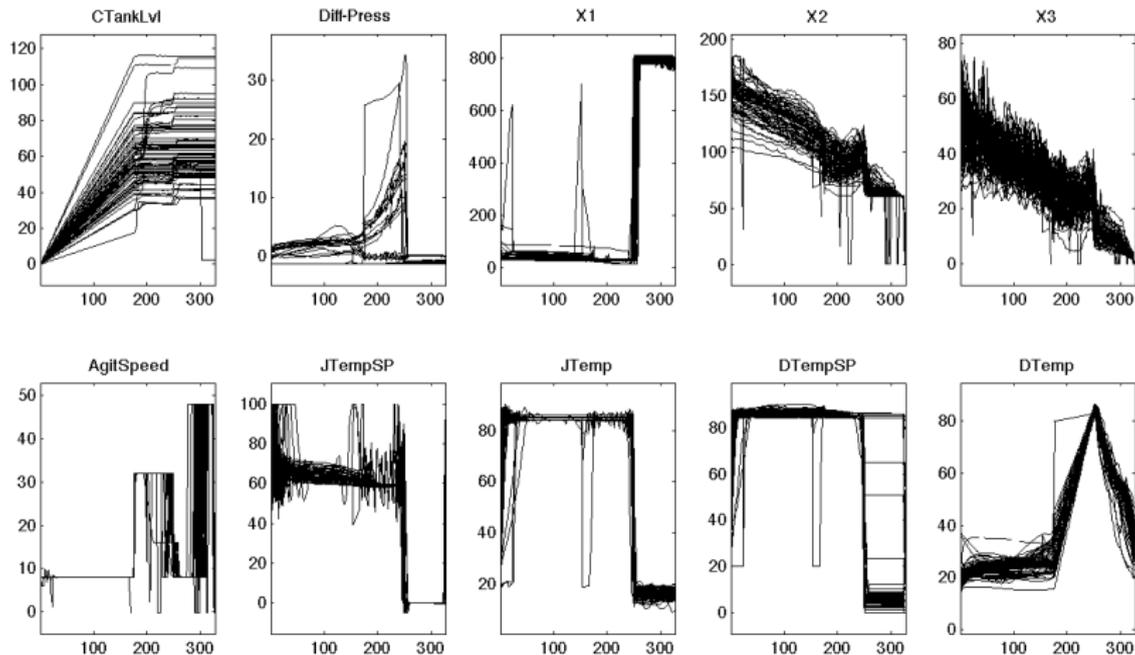


- ▶ This is an example of a **multiblock data set**
- ▶ Let's take a look at how the 3D cube of batch data arises ...

Batch reactor



Typical batch trajectories



This is a small data set: 325 time points \times 10 tags \times 65 batches

Examples of interesting data sets: Millau Viaduct

In France: expressway connecting Paris and Barcelona



Examples of interesting data sets: Millau Viaduct



Examples of interesting data sets: Millau Viaduct



Examples of interesting data sets: Millau Viaduct

- ▶ Pylons, deck, masts have anemometers, accelerometers, inclinometers, extensometers, and temperature sensors
- ▶ Detect movement (micrometer level), monitor for oscillations, stress/strain
- ▶ Piezoelectric sensors gather traffic data: weight, speed, density of traffic
- ▶ Can distinguish between fourteen different types of vehicles
- ▶ 100 readings per second from the main pylon
- ▶ Data transmitted via ethernet and fibre optic cables

Other data sources

- ▶ Chemical plants are moving to wireless sensors and networks
- ▶ More and more data available **and accessible** to engineers than even before
 - ▶ Prior to about 2005: data recorded, but not easily available

Issues you will face with engineering data

- ▶ **1. Size of the data**
 - ▶ rows: we can deal with this
 - ▶ columns: $K(K - 1)/2$ pairs of scatterplots
- ▶ **2. Lack of independence** between columns (i.e. redundancy in data)
 - ▶ $\mathbf{X}^T \mathbf{X}$ becomes singular
 - ▶ make-shift approach: pick a reduced set of columns
- ▶ **3. Low signal to noise ratio**
 - ▶ aim to keep our processes constant
 - ▶ little signal and high noise
 - ▶ data collected is mostly uninformative: constant, noisy, has drift and error
 - ▶ Called “happenstance data”

Issues you will face with engineering data

4. Non-causal data

- ▶ Happenstance data is non-causal
 - ▶ Only see correlation effects
 - ▶ Good enough in many cases
- ▶ Opposite case: a designed experiment
 - ▶ cause-and-effect

Issues you will face with engineering data

- ▶ **5. Errors in the data**
 - ▶ Least squares: assumes no error in X
 - ▶ Not realistic in most cases
- ▶ **6. Missing data**

◇	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Ton_in	KR30_IN	KR40_IN	PARM	HS_1	TOTAVF	PAR	FAR	r_FAR	%Fe_FAR	%P_FAR	%Fe_malm
2	1	0	0	4.65	1.6	84.9314	5.275	25	0.0625	0.249377			
3	2	0	0	4.65	1.25	84.9314	4.775	10.75	0.0625	0.578035			
4	3	0	0	0	6.95	84.9314	-3.5125	7.8125	0.0625	0.793651			
5	4	0	0	0	0.2	84.9314	3.2375		0.0625	100			
6	5	0	0	0	1.4625	84.9314	1.975		0.0625	100			
7	6	0	0	0	4.1	89.6072	-0.6625	7.3125	0.0625	0.847458			
8	7	0	0	0	1.05	89.236	2.375	6.75	0.0625	0.917431			
9	8	0	0	0	1.4	89.926	2.025	6.3125	0.0625	0.980392			
10	9	0	0	0		90.2119	3.225	6.5	0.0625	0.952381			
11	10	0	0	0		90.3292	3.0125	5.75	0.0625	1.07527			
12	11	1271.81	275.813	190.875	62.975	90.4108	383.706	307.875	591.75	65.7774	66.2	0.24	47.9
13	12	1290.56	278.55	208.575	58.075	90.4108	384.281	314.625	601.5	65.657	66.2	0.24	47.9
14	13	1267.39	278.55	207.375	63.1875	90.4108	398.212	312.188	585.063	65.2062	66.2	0.24	47.9
15	14	1250.44	278.063	204.525	57.475	90.4108	384.556	298.625	576.75	65.886	66.2	0.24	47.9
16	15	1265.51	279.563	190.425	49.3125	90.4108	415.381	304.125	591.75	66.0527	66.2	0.24	47.9
17	16	1268.18	276.112	194.625	62.875	90.4108	403.706	310.875	592	65.5683	66.2	0.24	47.9
18	17	1284.3	272.55	211.275	58.675	90.4108	405.331	293.125	575.5	66.2541	66.2	0.24	47.9
19	18	1284.41	275.4	208.275	50.475	90.4108	420.531	304.125	580	65.6016	66.2	0.24	47.9
20	19	1272.79	274.35	207.525	62.675	90.4108	394.569	300.125	598.25	66.5925	66.2	0.24	47.9
21	20	1317.11	269.813	192.225	56.175	90.4108	409.681	311.125	579.75	65.0765	66.2	0.24	47.9
22	21	1273.16	264.712	195.375	49.5625	90.4108	405.469	291.625	585.25	66.7427	66.2	0.24	47.9
23	22	1048.39	213.113	166.538	32.5125	94.5204	339.569	248.563	491.688	66.4218			
24	23	1019.66	217.875	192.525	31.4125	94.5204	330.044	264.625	494.438	65.1379			
25	24	1049.06	222.375	183.525	35.7125	94.5204		258.875	479.438	64.9369			
26	25	1057.5	215.625	177.825	33.275	94.5204		238.125	535.25	69.2096			
27	26	1036.95	211.125	165.375	26.2125	94.5204		244.313	490.438	66.7489			
28	27	1057.95	218.025	168.488	35.2625	94.5204		237.063	485.938	67.2113			
29	28	1033.05	212.512	179.475	33.4125	94.5204		254.625	520.5	67.1505			
30	29	1041.38	212.475	170.738	24.5	94.5204		256.375	507.75	66.4485			
31	30	1062.49	217.613	170.625	32.4125	94.5204		242.563	477.188	66.2991	67.2	0.2	51.2
32	31	1024.8	218.063	178.425	43.4625	94.5204		261.125	505.5	65.9384	67.2	0.2	51.2
33	32	1070.74	215.063	165.337	38.5125	94.5204		248.563	501.75	66.8721	67.2	0.2	51.2
34	33	1054.65	216.075	176.025	31.4125	94.5204		249.563	523	67.6968	67.2	0.2	51.2
35	34	1072.05	214.725	166.238	41.075	94.5204		263.563	514.5	66.1258	67.2	0.2	51.2
36	35	1056.71	224.625	177.675	35.3125	94.5204		252.875	522.25	67.3762	67.2	0.2	51.2

Issues you will face with engineering data

▶ 7. Multiple data sources

- ▶ raw material properties (e.g. one block per material)
- ▶ NIR or UV-VIS spectra from the process
- ▶ Batch data from the system *or* Continuous data
- ▶ Final quality variables

These are called **multiblock data sets**.

Issues you will face with engineering data

Large data set: this is **not a problem**

It's not about the size of your data ... it's what you do with it.

- ▶ Many **rows**?
 - ▶ use a for loop
 - ▶ use parallel computing
 - ▶ Amazon EC2:
 - ▶ Simple CPU rent: \$0.06/hour
 - ▶ 23 GB memory, 4-core CPU, 1.7 TB storage, 64-bit: \$1.64/hour
- ▶ Many **columns**?
 - ▶ are they really all independent?
 - ▶ use latent variable methods

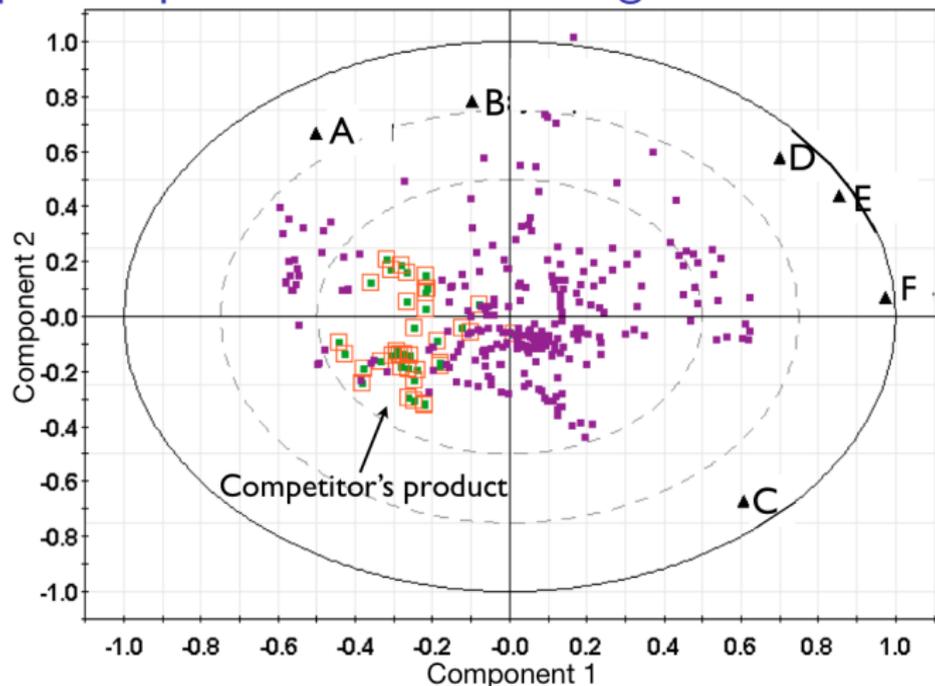
Issues you will face with engineering data

We require tools that:

- ▶ extract relevant **information** from data
- ▶ deal with missing data
- ▶ 3-D, 4-D and higher data sets
- ▶ combine data from different sources (same object)
- ▶ handle collinearity (low signal to noise ratio)
- ▶ handle error in recorded data

Latent variable methods meet these requirements.

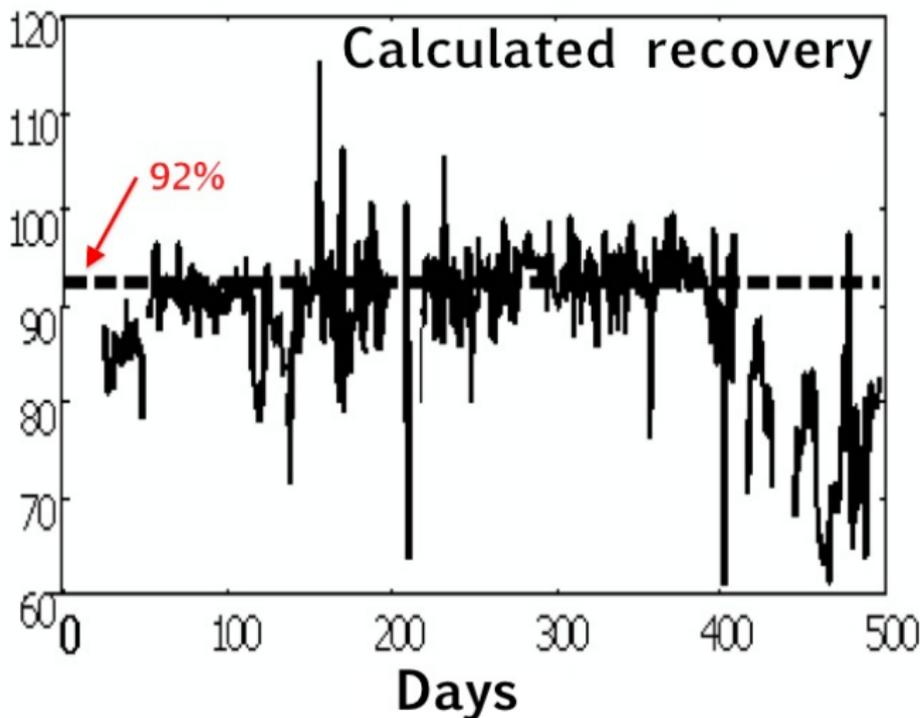
Improved process understanding



- ▶ Learn which variables are correlated
- ▶ Competitor has much less variability !
- ▶ Can we reproduce the competitor's product? Yes

LVM for troubleshooting

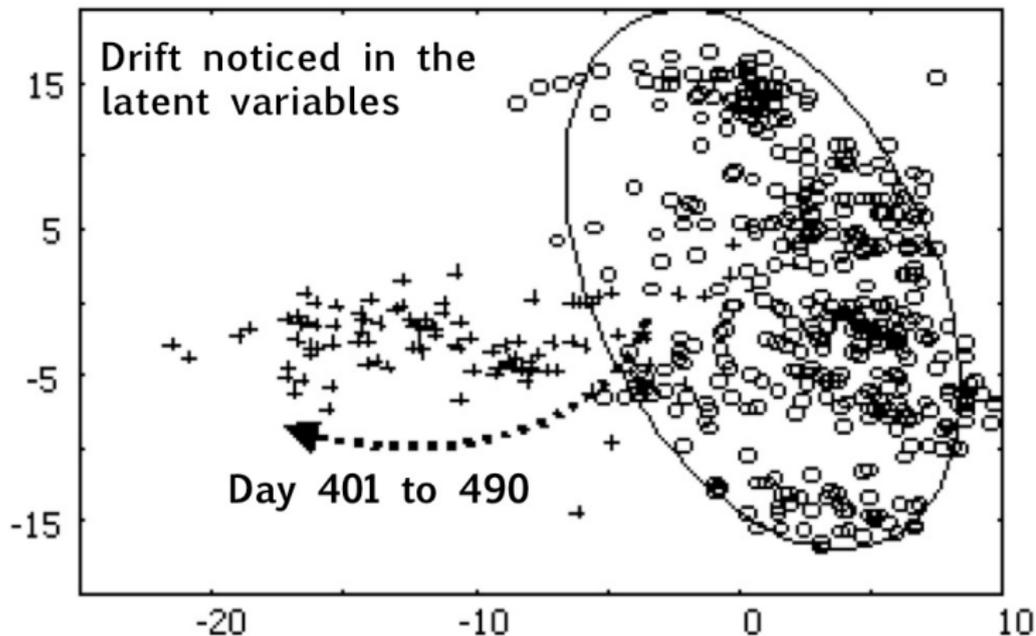
Reduction in monomer recovery \sim day 400. Target recovery is 92%



- ▶ Engineers looked at various time series plots for several weeks
- ▶ 100 days elapsed without finding the cause

LVM for troubleshooting

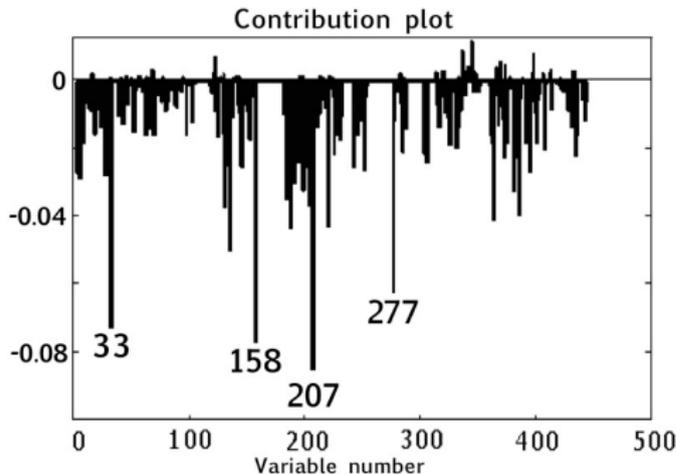
- ▶ A latent variable model with 2 variables was built
 - ▶ Compresses the 447 variables to 2 variables
 - ▶ Retains most of the information in t_1 and t_2



- ▶ Investigate the latent variables instead to see what changed ...

LVM for troubleshooting: contribution plot

Ask the LV model: what has changed between the two clusters of points

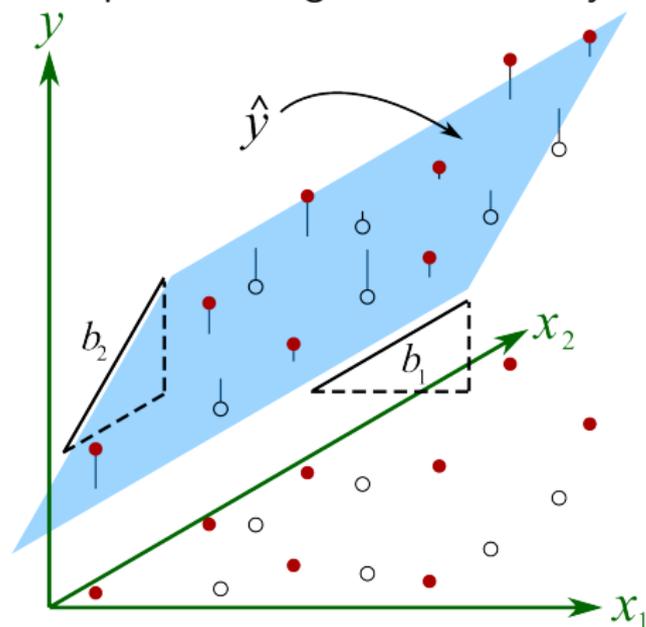


- ▶ **207**: temperature on tray 129 in distillation column #3
- ▶ **158**: a tag from distillation column #3
- ▶ **33** and **277**: related to concentration of feed A

- ▶ *Suggests*: bad temperature control on tray 129 when feed concentration of A is high
- ▶ Fixed the controller and recovery went back to normal

LVM for regression

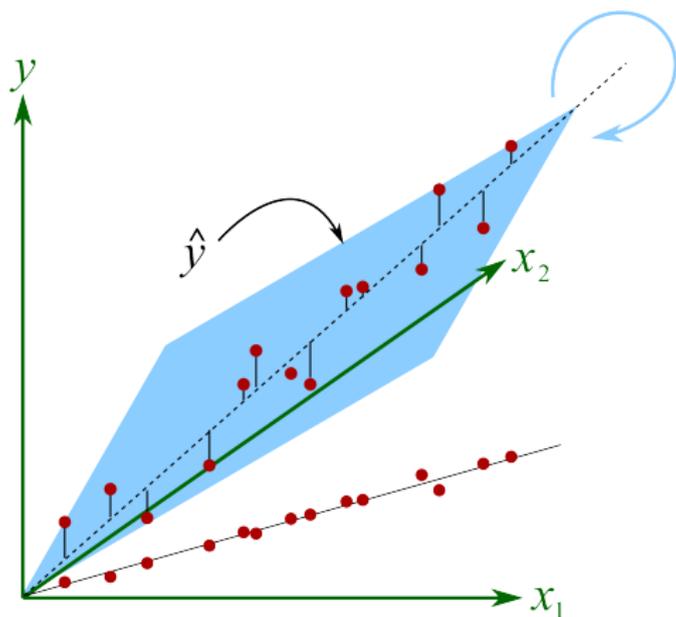
Multiple linear regression model: $y = b_1x_1 + b_2x_2$



- ▶ Stable estimates are desirable:
 - ▶ for learning about the process
 - ▶ for accurate predictions of \hat{y} in the future
- ▶ We get stable estimates for b_1 and b_2 when the plane is “well supported” by the measured points
- ▶ Think of DOE: we intentionally move to the corner points in x_1 and x_2 to fit the model

LVM for regression

- ▶ But what if the two x -variables are strongly correlated?



- ▶ The plane rotates for small changes in x -variables
- ▶ The slope coefficients change: can even change sign! (demo in class)
- ▶ Unstable slopes: cannot be interpreted reliably

LVM for regression

- ▶ What can we do about it?
- ▶ *Suboptimal solution*
 - ▶ Recognize that x_1 and x_2 are correlated
 - ▶ Choose either x_1 or x_2 in the model:
 - ▶ $y = b_0 + b_1x_1$
 - ▶ $y = b_0 + b_2x_2$
- ▶ Problems with correlated data
 - ▶ which variables do you choose to keep or throw out?
 - ▶ can I use an average of the two correlated variables?
 - ▶ how do you know what is “too strong correlation” before its problematic?
- ▶ Solution: don't select variables!

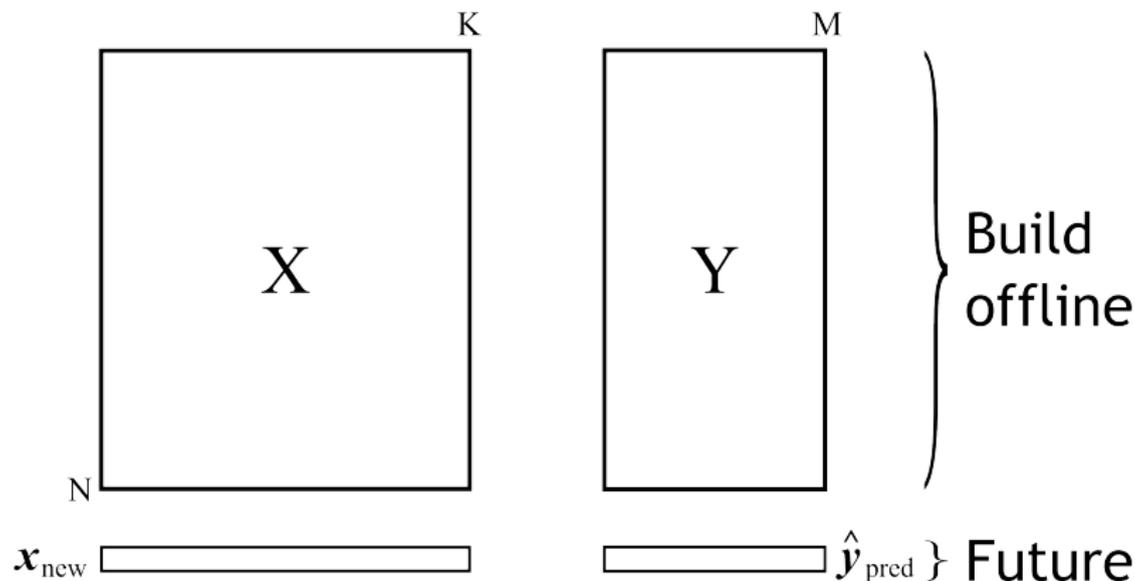
LVM for regression

Use projection to latent structures (PLS)

- ▶ handles missing values
- ▶ deals with correlated columns in X
- ▶ can be built simultaneously if more than 1 y -variable
- ▶ assumes error in the X 's, unlike least squares

Predictive modelling (inferential sensors)

- ▶ MLR has some serious disadvantages
 - ▶ Cannot handle missing data
 - ▶ Cannot handle strong correlations
 - ▶ MLR requires $N > K$
 - ▶ Only one y -variable at a time
 - ▶ Assumes no noise in \mathbf{X} , which is never true



Predictive modelling (inferential sensors)

- ▶ Inferential sensor = soft sensor
- ▶ Image data used as \mathbf{X}
- ▶ Snackfood example: <http://dx.doi.org/10.1021/ie020941f>
- ▶ Work by Honglu Yu
 - ▶ <http://digitalcommons.mcmaster.ca/opensdissertations/866/>

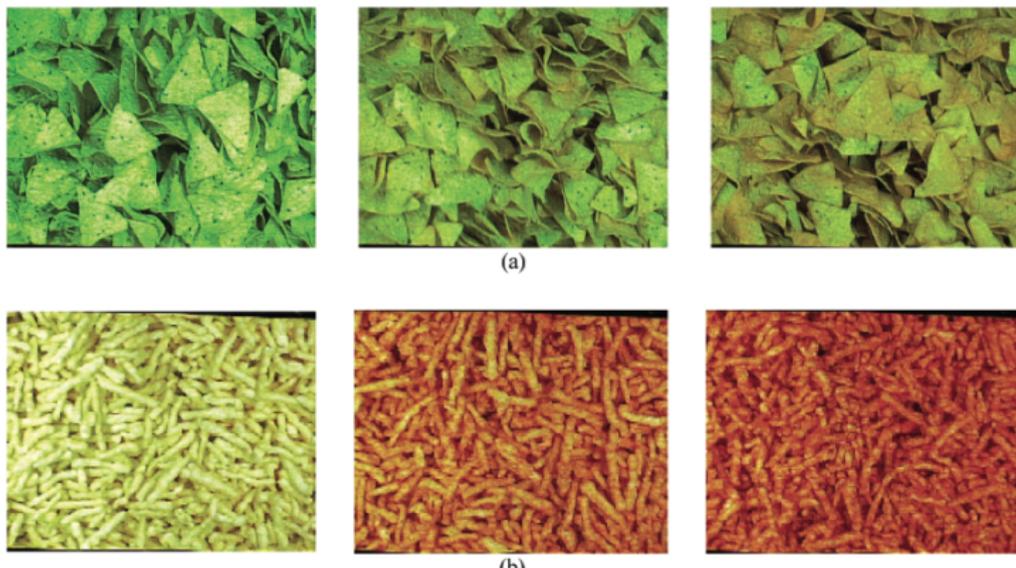


Figure 2 from the paper

Predictive modelling (inferential sensors)

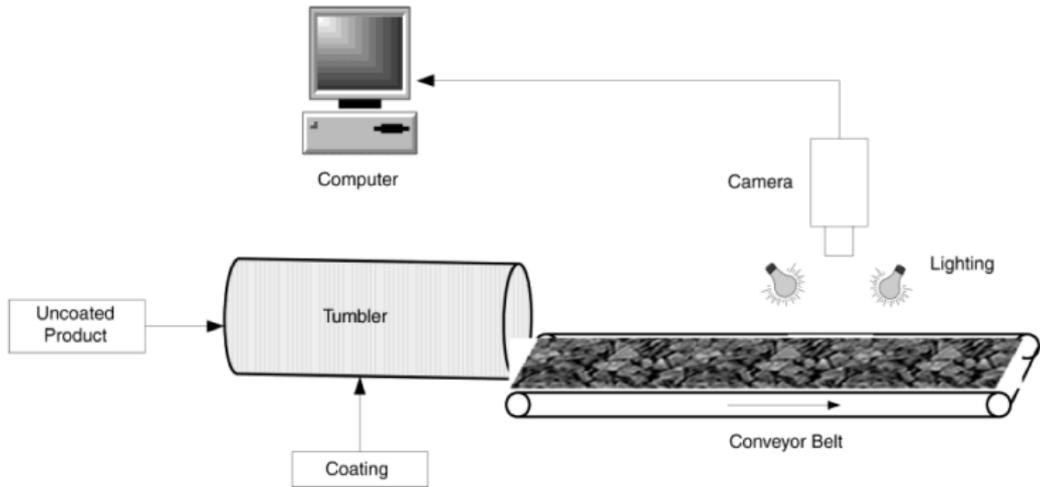


Figure 8. Schematic of the processes and imaging systems.

Figure 8 from the paper: <http://dx.doi.org/10.1021/ie020941f>

Predictive modelling (inferential sensors)

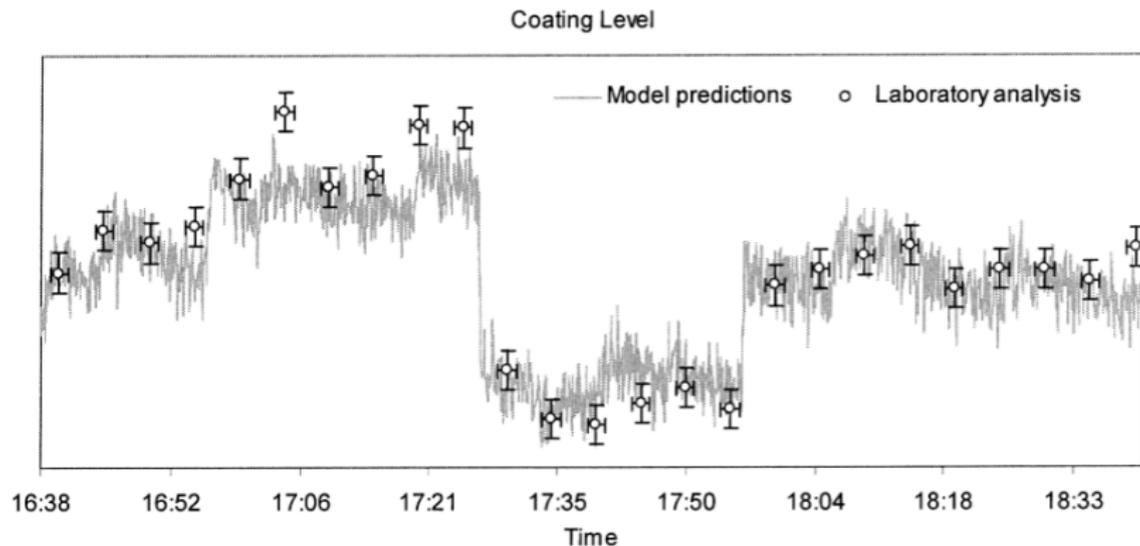
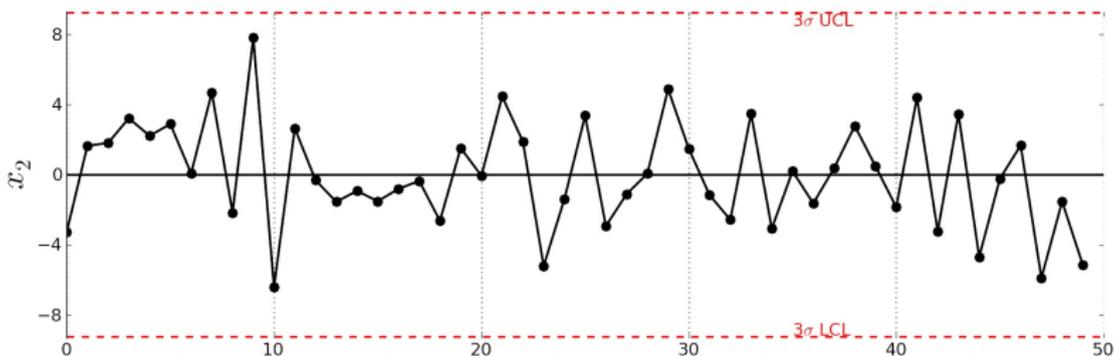
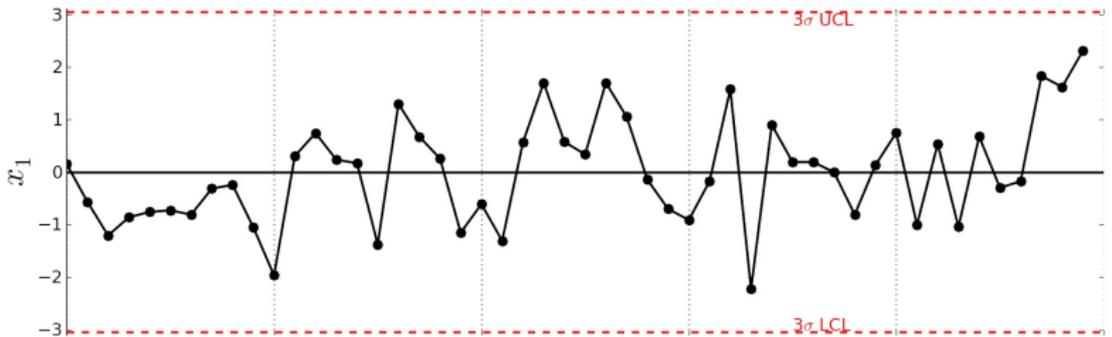


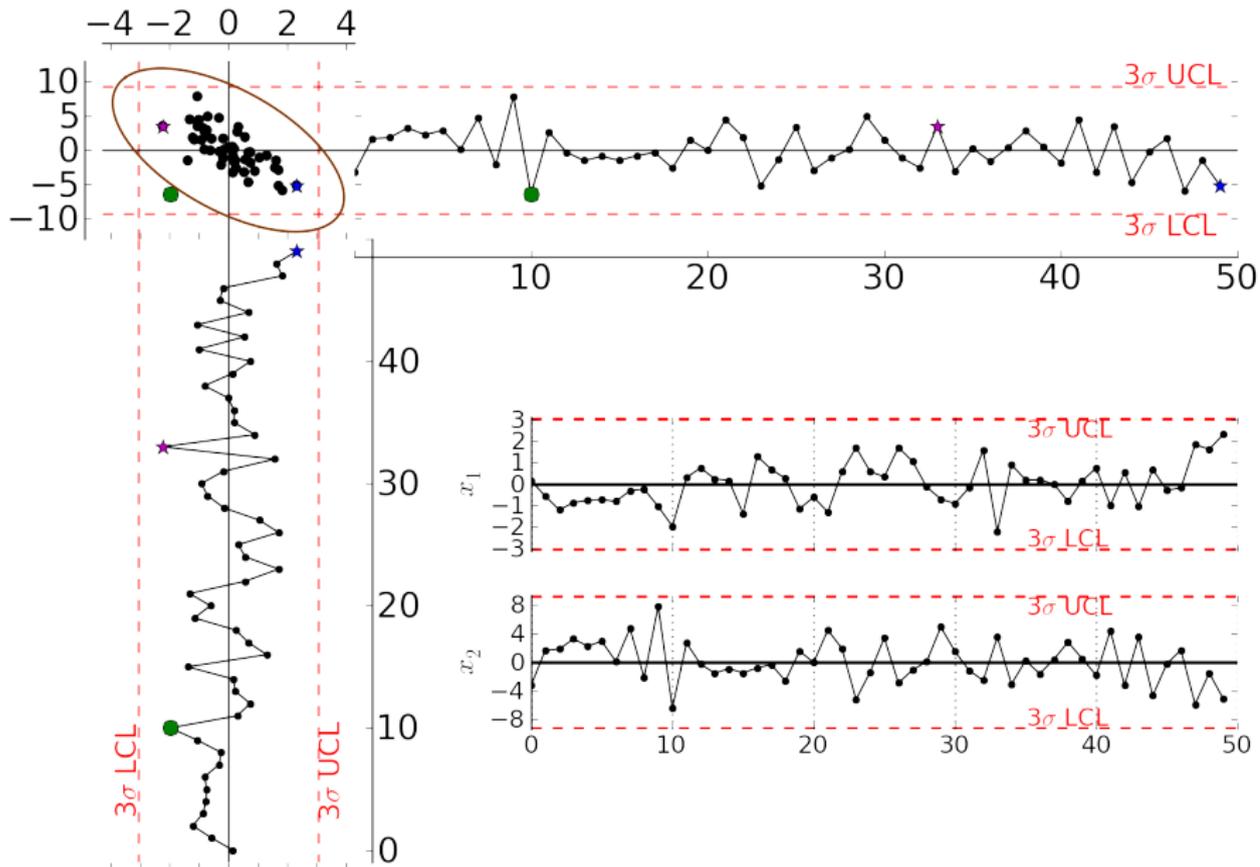
Figure 10 from the paper: <http://dx.doi.org/10.1021/ie020941f>

Why do we need latent variable methods?

- ▶ Shewhart chart for two variables, x_1 and x_2
 - ▶ e.g. final product quality from lab values



Why LVM? For process monitoring



Process monitoring

Monitoring with latent variables:

- ▶ We have K variables (tags)
- ▶ Reduce this to A scores (latent variables)
- ▶ Combine these A scores to a single value: Hotelling's T^2
- ▶ Errors: combined into a single value: SPE

Advantages:

- ▶ The scores are orthogonal
- ▶ Fewer scores than original variables
- ▶ Monitor anywhere that there is real-time data
- ▶ Don't have to wait for the lab's final measurement

Industrial case study: Dofasco

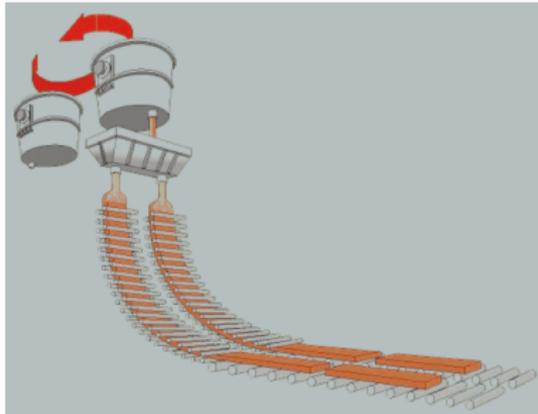
- ▶ ArcelorMittal in Hamilton (formerly called Dofasco) has used multivariate process monitoring tools since 1990's
- ▶ Over 80 latent variable applications used daily
- ▶ Most well known is their casting monitoring application, Caster SOS (Stable Operation Supervisor)
- ▶ It is a multivariate monitoring system

Dofasco case study: slabs of steel



All screenshots with permission of Dr. John MacGregor

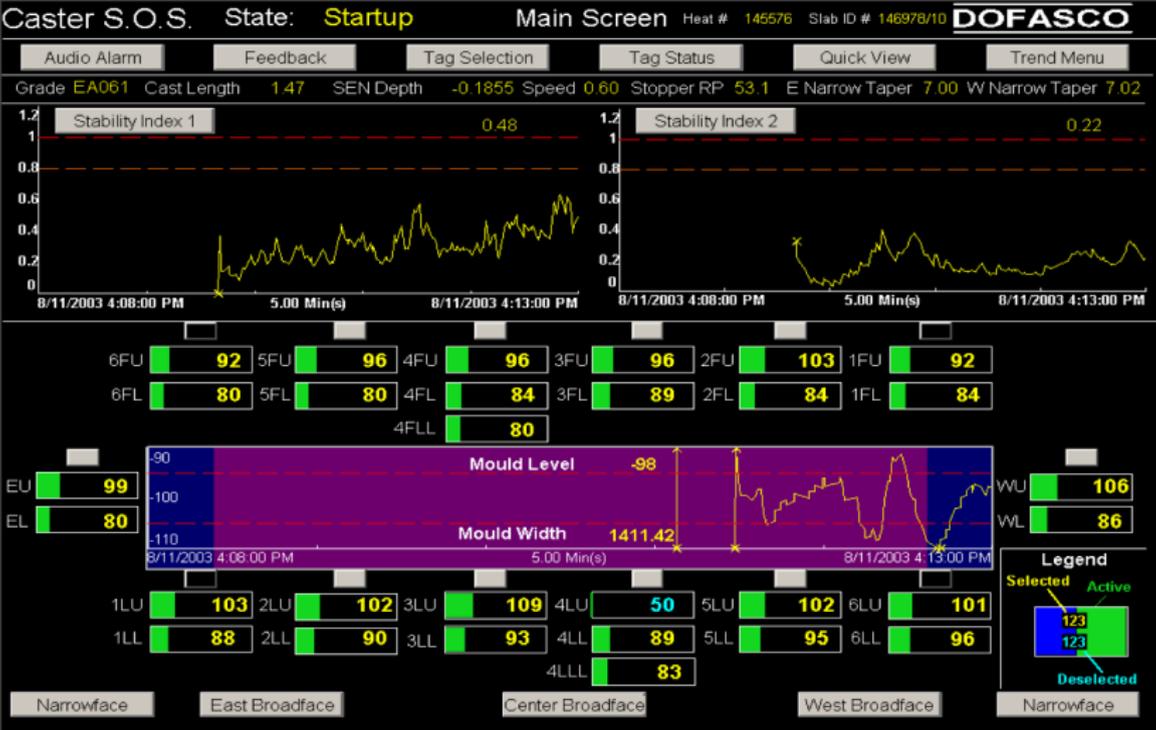
Dofasco case study: casting



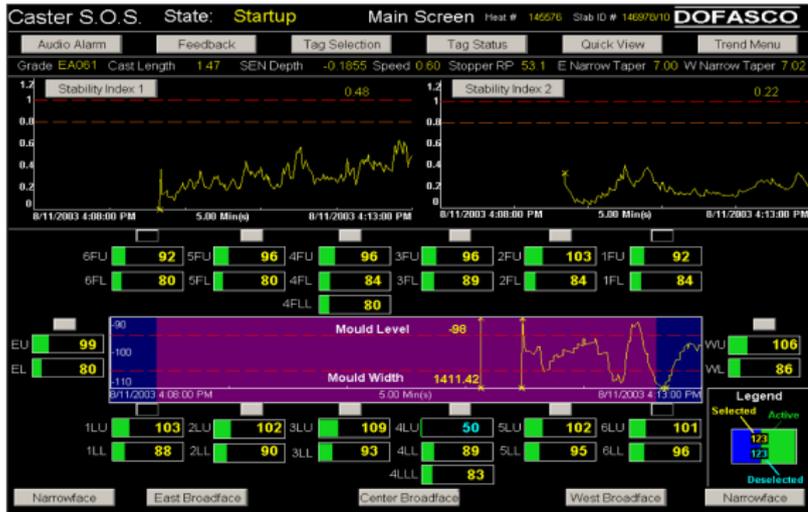
Dofasco case study: breakout



Dofasco case study: monitoring for breakouts



Dofasco case study: monitoring for breakouts



- ▶ Stability Index 1 and 2:
 - ▶ Hotelling's T^2
 - ▶ SPE
- ▶ When alarm: shows contribution plots
- ▶ Shows real-time raw data, as operator requires it

Dofasco case study: an alarm

Caster S.O.S.
Stability Index 1 1.02
Stability Index 1
Stability Index 2 0.30
DOFASCO

Main Screen
Stability Index 1
Stability Index 2
Tag Selection
Tag Status
Quick View
Trend Menu

Stability Index 1 01/19/2003 9:50:05 PM
 1.2
 1.02
 0.8
 0.6
 0.4
 0
 -240 500000 -600

Miscellaneous

 Cast Speed
 SEN Depth
 Stopper Rod Pos.
 Negative Strip
 Carbon Eqv.
 Carbon %

Tundish

 Net Weight
 Arg Flow
 Temperature
 Superheat

Top 5 Stability Index 1 influences

(1) Stopper Strok Remain	
(2) Loose Upper 1 TC	
(3) Loose Lower 6 TC	
(4) Loose Upper 6 TC	
(5) Fixed Lower 6 TC	

<input type="checkbox"/>					
6FU	5FU	4FU	3FU	2FU	1FU
6FL	5FL	4FL	3FL	2FL	1FL

E Narrow Taper
 E Width Chg Spd

Width
 Level
 Osc Frequency
 Osc Stroke

W Narrow Taper
 W Width Chg Spd

<input type="checkbox"/>					
1LU	2LU	3LU	4LU	5LU	6LU
1LL	2LL	3LL	4LL	5LL	6LL

Heat Removal

 Loose
 Fixed
 East
 West

Mold Water Delta Temp

 Fixed
 Loose
 East
 West

Mold Water

 Inlet Temp.

Mold Water Flow

 Fixed
 Loose
 East
 West

45

Summary: Extracting value from data

1. **Improve process understanding**
 - ▶ Competitor example
2. **Troubleshooting process problems**
 - ▶ Monomer example
3. **Predictive modelling (inferential sensors)**
 - ▶ Snackfood example
4. **Process monitoring**
 - ▶ Dofasco example
5. **Optimizing and improving processes**
 - ▶ DOE with RSM: excellent tools, even on modern data systems
 - ▶ Can combine latent variable methods with DOE