Multivariate statistical methods for the analysis, monitoring and optimization of processes

Jay Liu

Dept. of Chemical Engineering

Pukyong National University

Credits

- Prof. John MacGregor @ McMaster University
  - “father” of multivariate statistics & industrial applications
  - Course outline and topics covered are similar to his course
- Mr. Kevin Dunn @ McMaster University/ConnectMV
  - Most of this lecture materials come from his
- Prosensus
  - Providing one-year ProMV academic license for free.
  - Providing materials for tutorials

1. Introduction

Extracting value from data

- Engineers can use large quantities of data:
  1. Improve process understanding
  2. Troubleshooting process problems
  3. Improving, optimizing and controlling processes
  4. Predictive modeling (inferential sensors)
  5. Process monitoring

We will come back to these in the next class

- Throughout this course,

  Data are collected in a table or matrix:

  - Each row represents an observation
  - Each column represents a variable

Convention in Multivariate statistics

	Variable 1	Variable 2	Variable 3
Object 1			
Object 2			
Object 3			
Object 4			

Data Characteristics

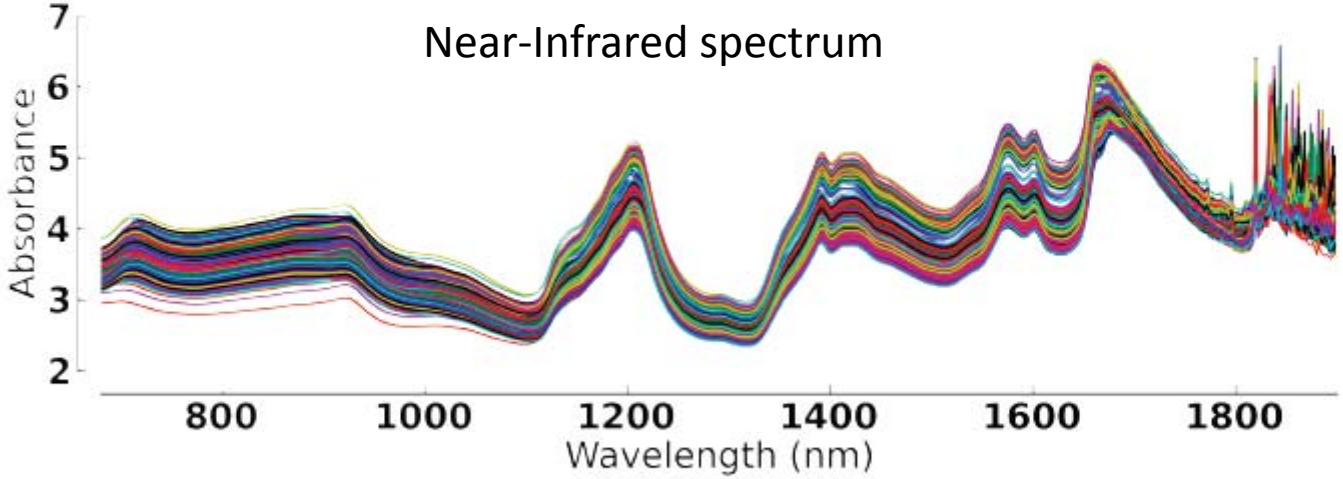
1920's to 1950's:

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

Data Characteristics

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

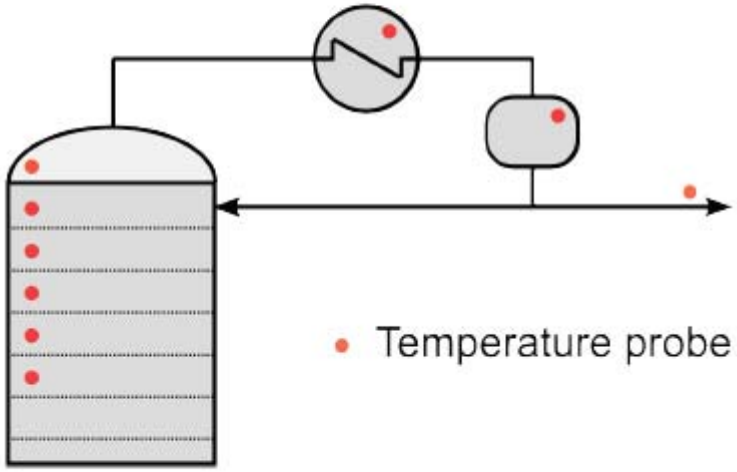
Near-Infrared spectrum



  - Cannot use MLR directly: $K \gg N$

Data Characteristics

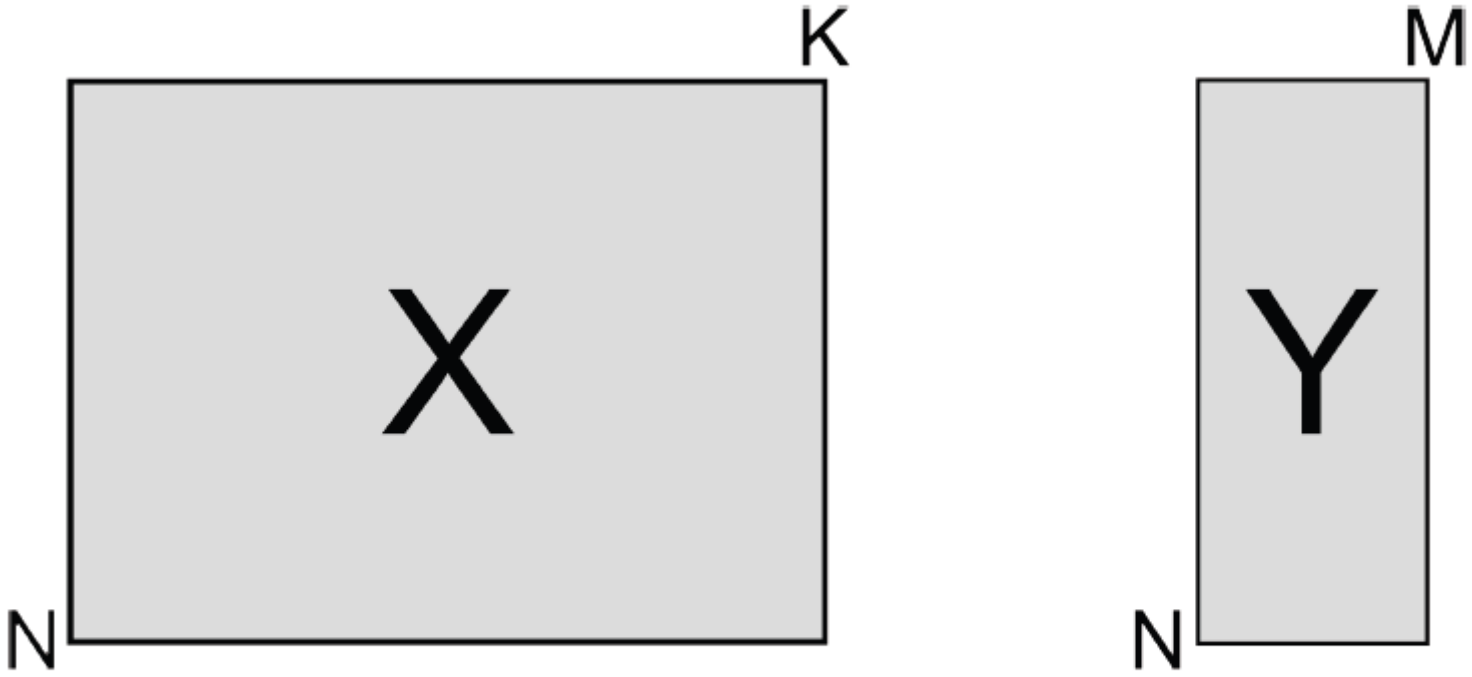
- **Large N and small K**
  - Refinery, most chemical plants
  - 2000 to 5000 variables (called tags) every second



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

Data Characteristics

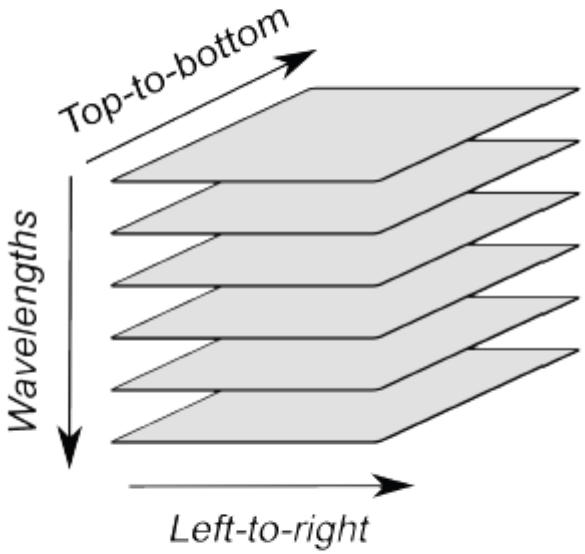
- **X and Y matrices**



- Predict one or more variables
- Could use MLR; fails for highly correlated data

Data Characteristics

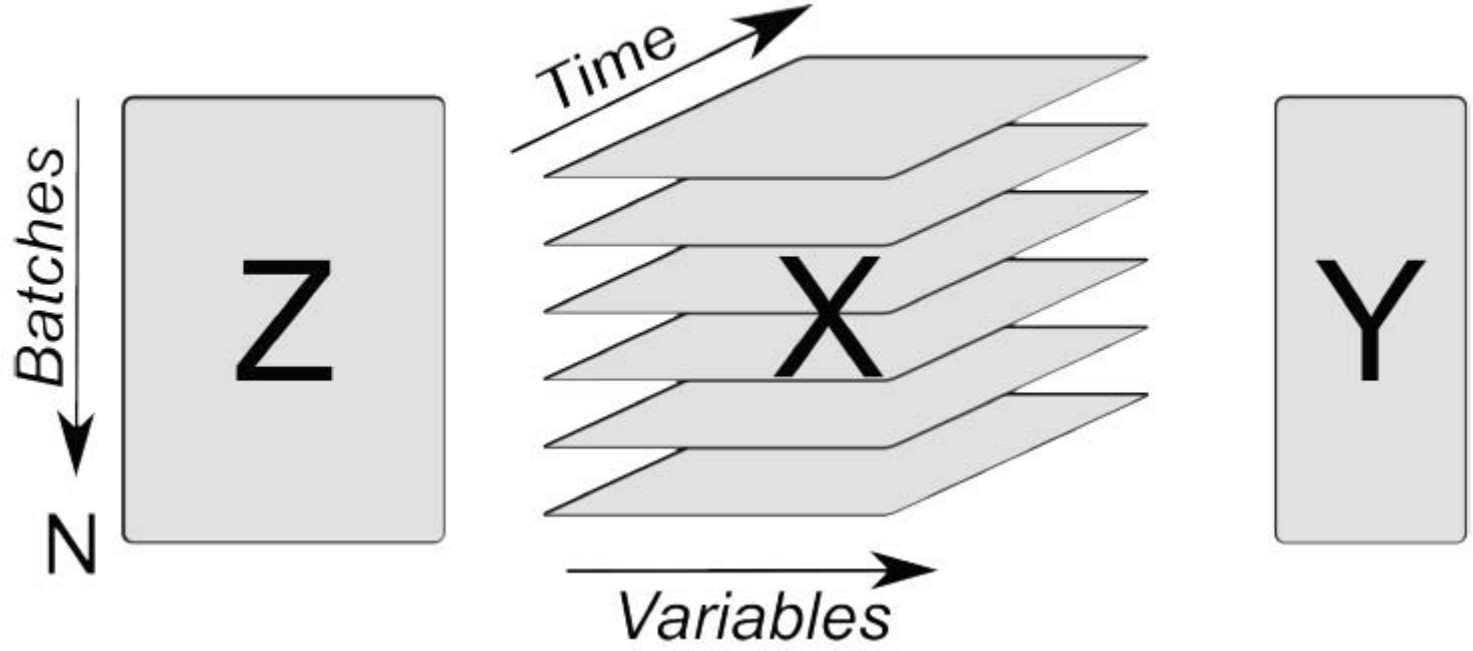
- **3D data sets and higher dimensions**
- Very common situation now
- Image data (medical imaging)



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

Data Characteristics

- **Batch data sets**

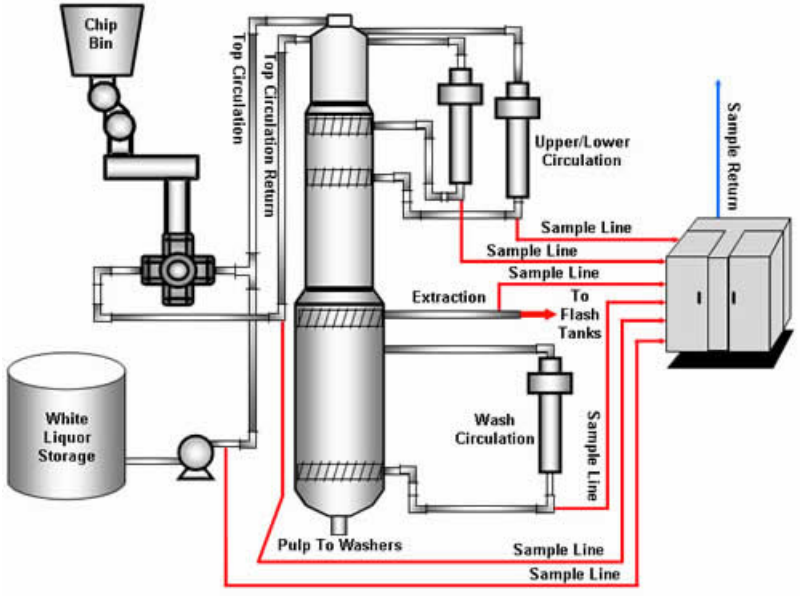


Data Analysis Purpose

- Summary & trouble-shooting
- Predictive modeling (Regression)
- Labeling or grading (Classification)
- Process & product design
- Scale-up & product transfer
- QSAR (Quantitative Structure Activity Relationships)
- And many more

Examples: trouble-shooting

- Undetected process changes @ pulp digester
  - Didn't know when & why the changes occurred
  - Data: 301 obs., 22 vars.
    - 22 time-series plots? 231 Scatter plots?

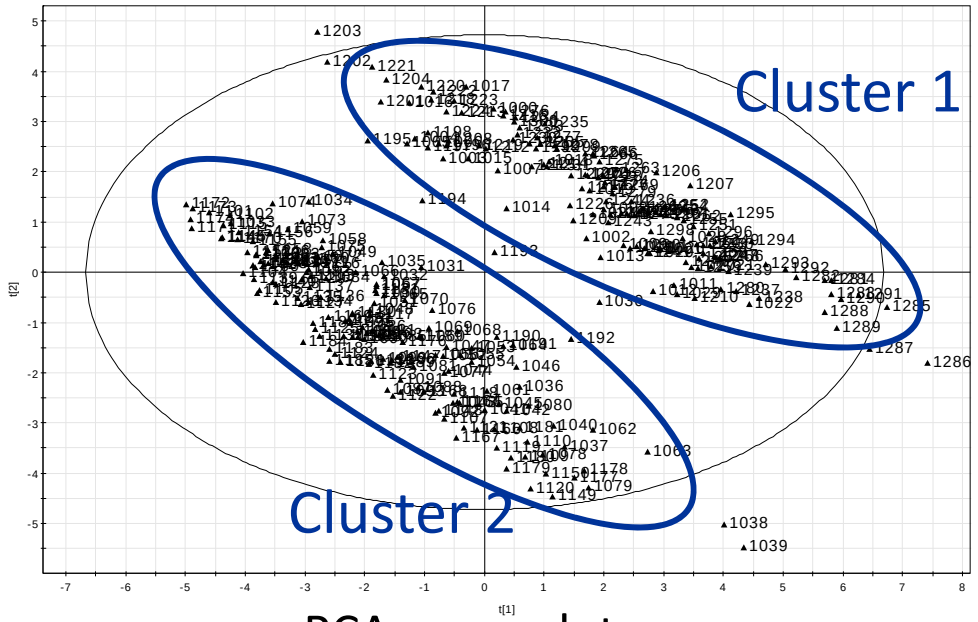


Examples: trouble-shooting

- PCA gives ONE excellent summary
  - Two distinct clusters in a t_1-t_2 score plot

    Cluster 1: obs.1000~1030, obs.1192~1300

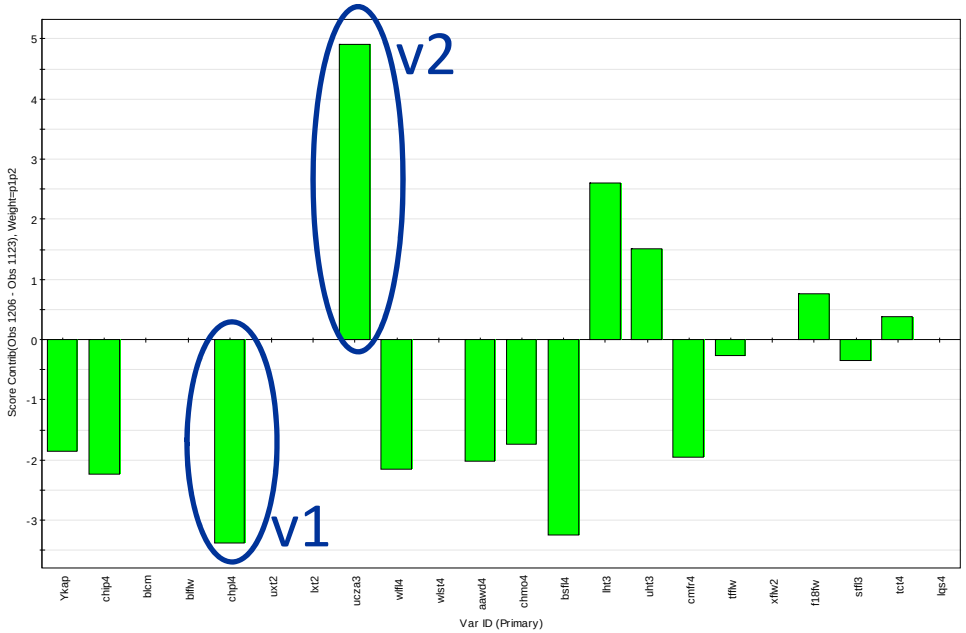
    Cluster 2: obs.1031~1191



PCA score plot

Examples: trouble-shooting

- PCA gives candidates of root-cause
  - During the period of cluster 2, v1 and v2 fluctuated while they were maintained constant level during the period of cluster 1.



PCA Contribution plot

Examples: predictive modeling

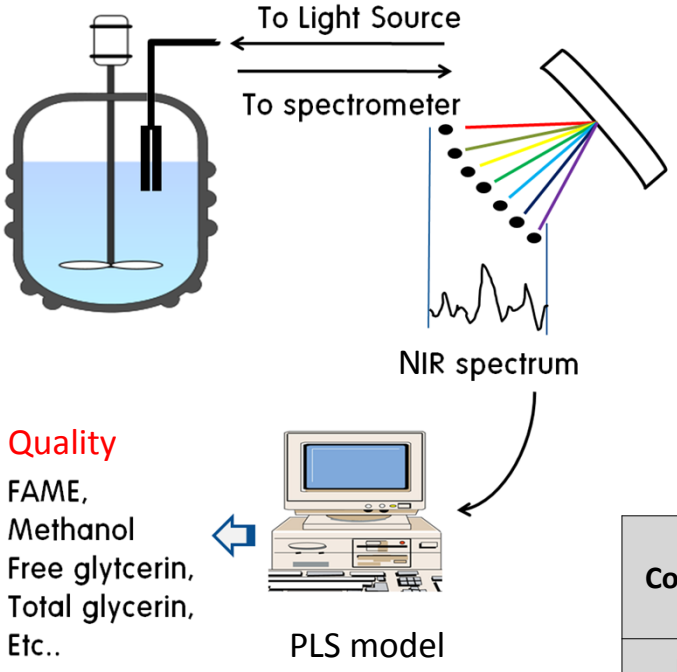
- Measurement of biodiesel quality
  - Spec. & methods given in ASTM

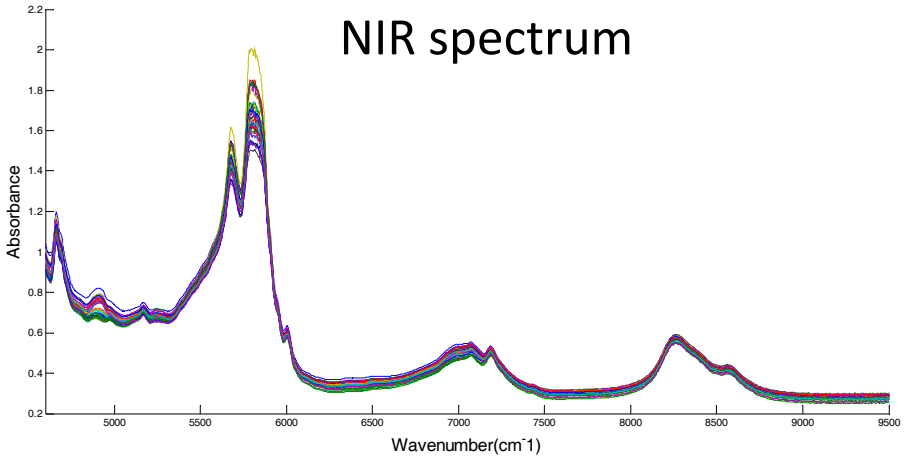
Property	ASTM Method	Limits	Units
Calcium & Magnesium, combined	EN 14538	5 maximum	ppm (ug/g)
Flash Point (closed cup)	**D 93**	**93 minimum**	**degrees C**
Alcohol Control (One of the following must be met)			
1. Methanol Content	EN14110	0.2 maximum	% volume
2. Flash Point	D93	130 minimum	Degrees C
Water & Sediment	**D 2709**	**0.05 maximum**	**% vol.**
Kinematic Viscosity, 40 C	D 445	1.9 - 6.0	$mm^2/sec.$
Sulfated Ash	D 874	0.02 maximum	% mass
Sulfur			
S 15 Grade	**D 5453**	**0.0015 max. (15)**	**% mass (ppm)**
S 500 Grade	**D 5453**	**0.05 max. (500)**	**% mass (ppm)**
Copper Strip Corrosion	D 130	No. 3 maximum	
Cetane	D 613	47 minimum	
Cloud Point	**D 2500**	**report**	**degrees C**
Carbon Residue 100% sample	D 4530*	0.05 maximum	% mass
Acid Number	**D 664**	**0.50 maximum**	**mg KOH/g**
Free Glycerin	**D 6584**	**0.020 maximum**	**% mass**
Total Glycerin	**D 6584**	**0.240 maximum**	**% mass**
Phosphorus Content	D 4951	0.001 maximum	% mass
Distillation, T90 AET	D 1160	360 maximum	degrees C
Sodium/Potassium, combined	EN 14538	5 maximum	ppm
Oxidation Stability	**EN 14112**	**3 minimum**	**hours**
Cold Soak Filtration	**Annex to D6751**	**360 maximum**	**seconds**
For use in temperatures below -12 C	**Annex to D6751**	**200 maximum**	**seconds**

Costs too much time & money!

Examples: predictive modeling

- PLS gives ONE accurate multivariate calibration model for biodiesel & impurities.





Components	Biodiesel	MEOH	Free glycerin	Mono Glyceride	Di Glyceride	Tri Glyceride
R^2 value	0.999	0.994	0.995	0.996	0.994	0.994

Issues faced with engineering data

- **Size of the data**
  - rows: we can deal with this
  - columns: $K(K-1)/2$ pairs of scatterplots
- **Lack of independence**
  - $\mathbf{X}^T\mathbf{X}$ becomes singular
  - make-shift approach: pick a reduced set of columns
- **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 faced with engineering data

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

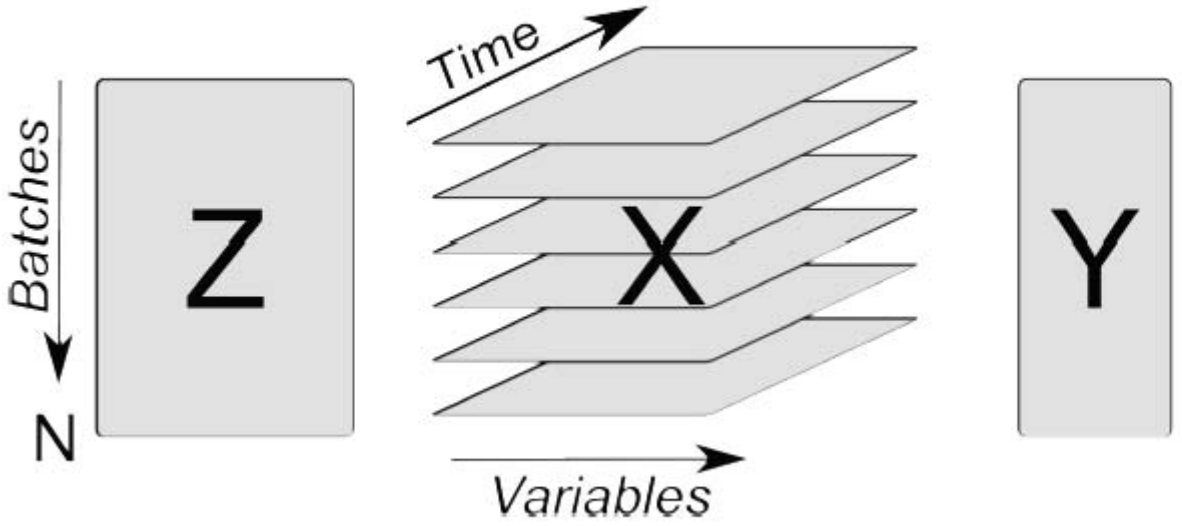
Issues faced with engineering data

- **Errors in the data**
  - Least squares: no error in X
- **Missing data** Can be up to 30% (or more!).

	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
37	36	1018.13	223.275	161.587	12.4	94.5204	250.225	222.125	577.563	72.2235			
38	37	1056.75	230.962	159.188	5.55	94.5204	276.144	231.563	565.5	70.948			

Issues faced with engineering data

- **Unaligned data**



Issues faced with engineering data

Tools that we require:

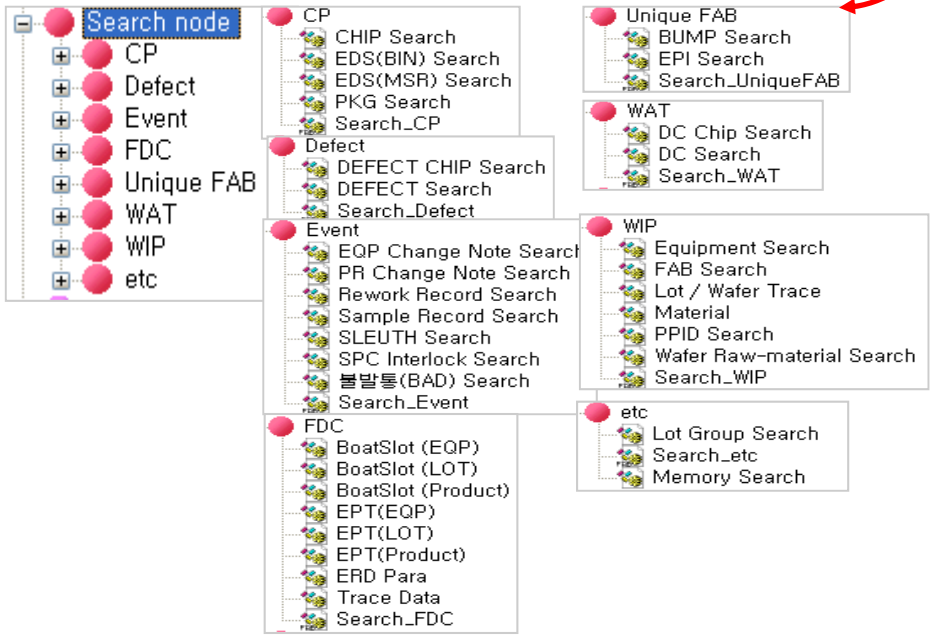
- extract relevant **information** from data
- deal with missing data
- 3-D, 4-D and higher data sets
- combine data from different sources (same object)
- handle collinearity (low signal to noise ratio)
- handle error in recorded data

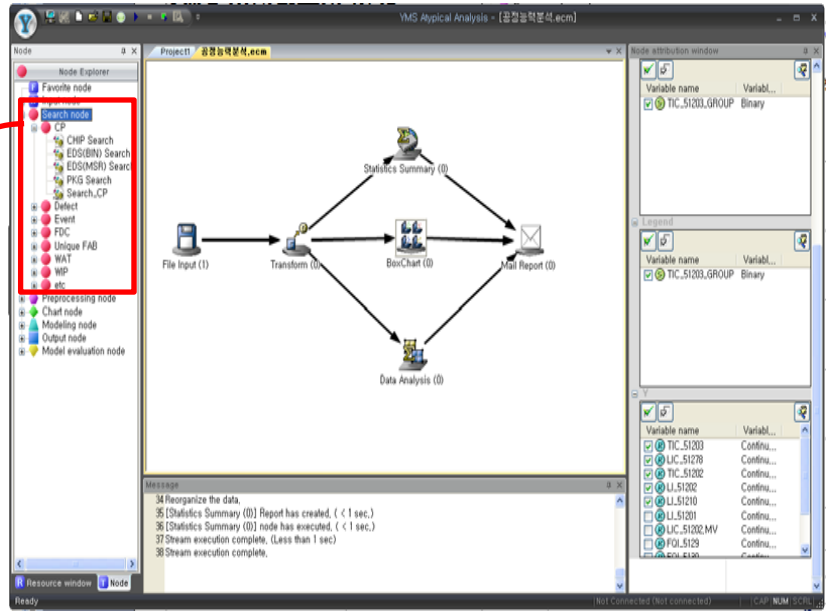
Latent variable methods are a suitable tool that meet these requirements.

Examples of large data sets

- Typical fab processes
  - Data size: order of terabytes in a few days

Databases in a typical fab process





Database merging in a data-mining package

What is a latent variable?

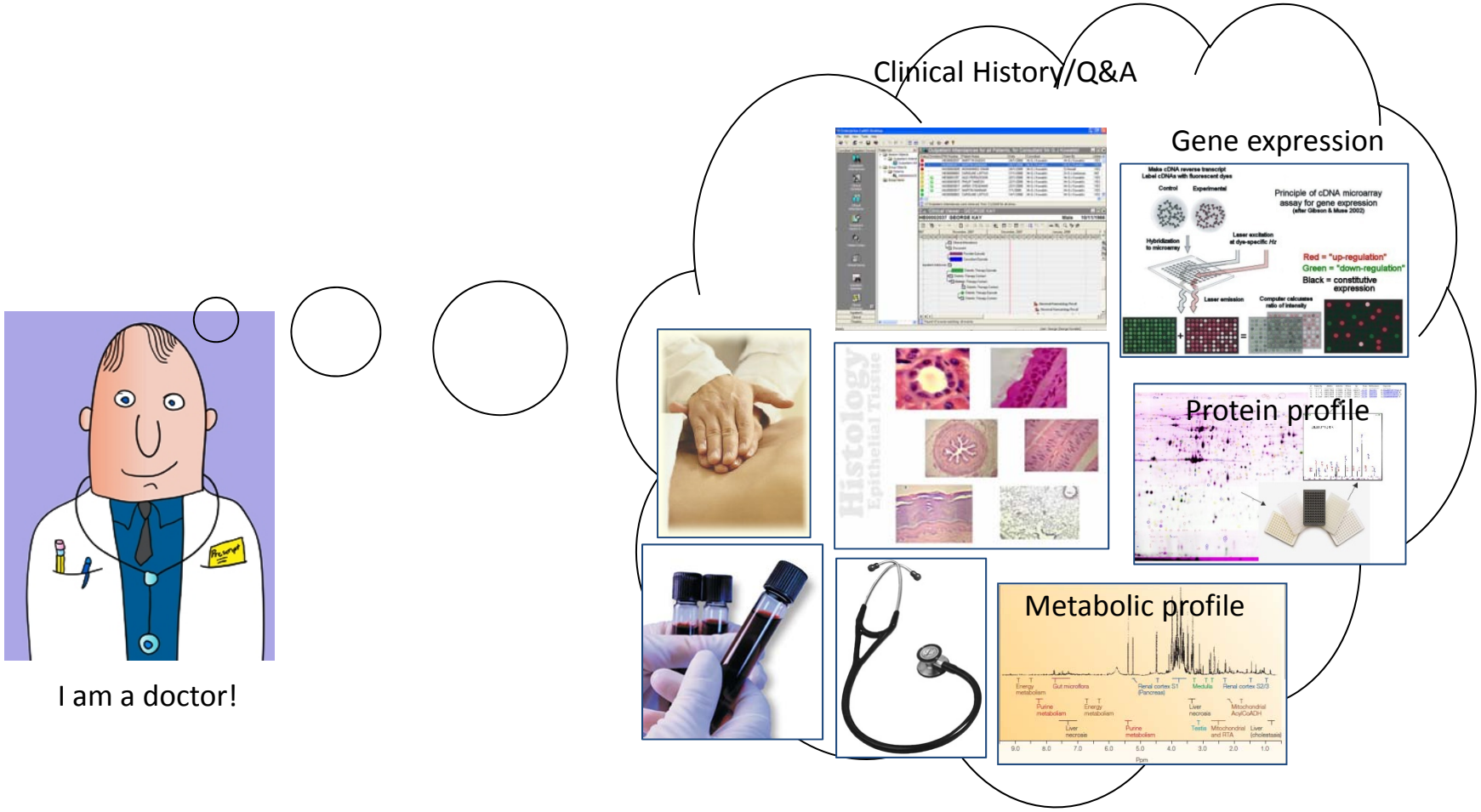
- Fortunately (& unfortunately at the same time), all variables are not independent.
  - Redundant images of few ***"latent"*** variables

Your health

- No single measurement of "health"
  - blood pressure
  - cholesterol
  - weight
  - waist, hip (waist:hip ratio)
  - blood sugar
  - temperature, *etc*
- Combine these in some way? Trained doctor does this mentally.

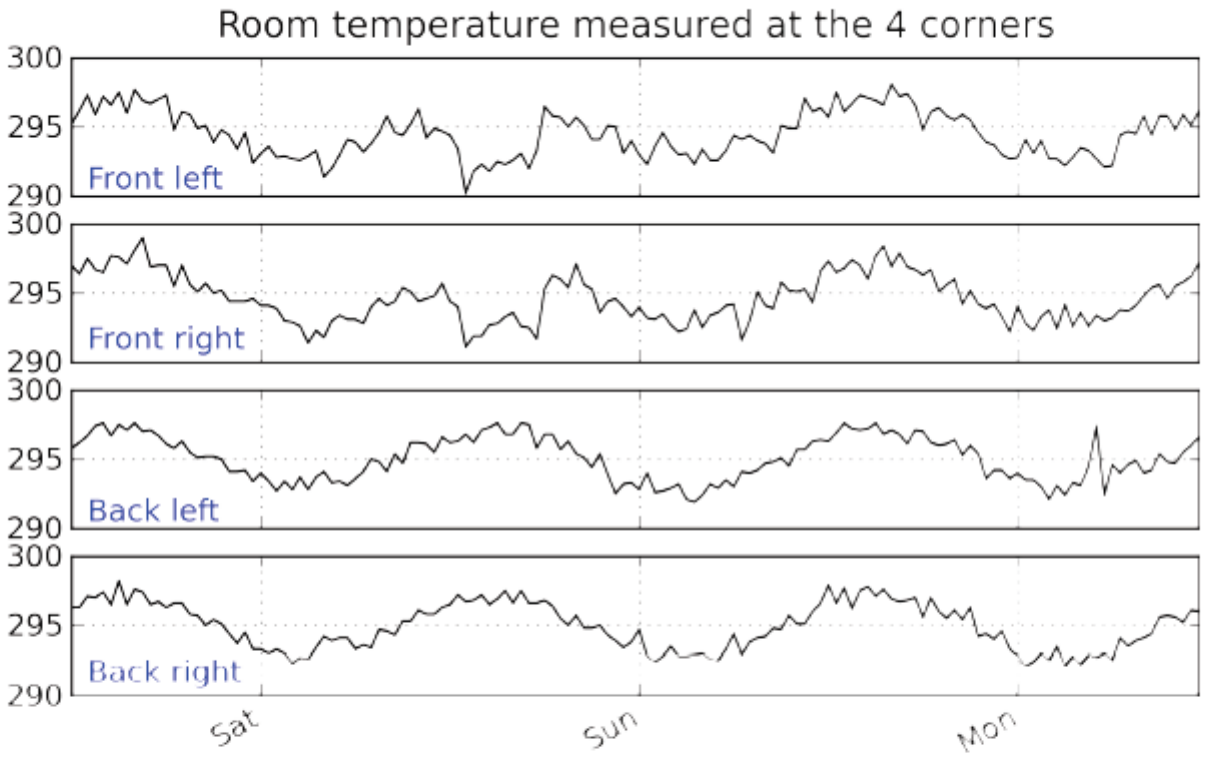
Health is a latent (hidden) variable

What is a latent variable?



What is a latent variable?

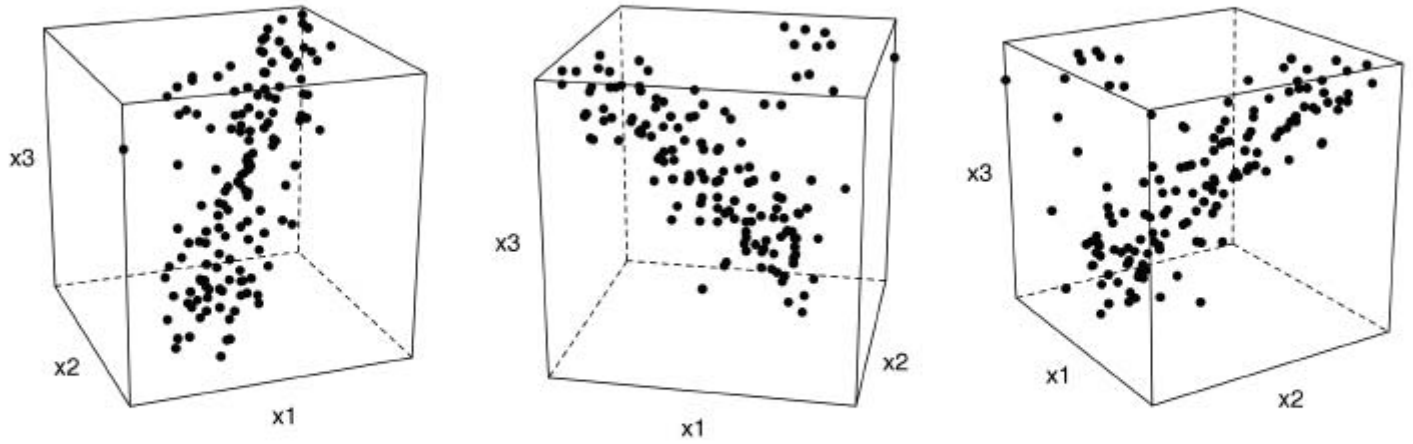
Temperature in this room



- What drives the movement up and down?
- Correlation of thermometers with the driving force.

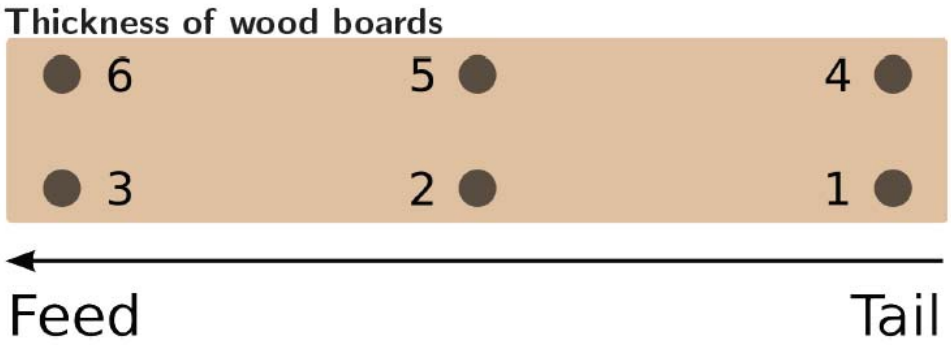
What is a latent variable?

Temperature in this room: geometrically



- Each measurement is one point
- Rotation

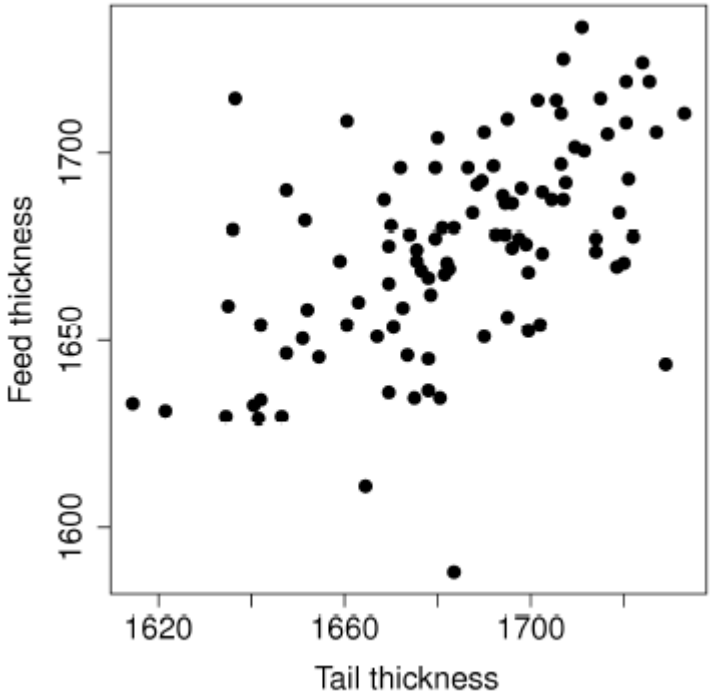
What is a latent variable?

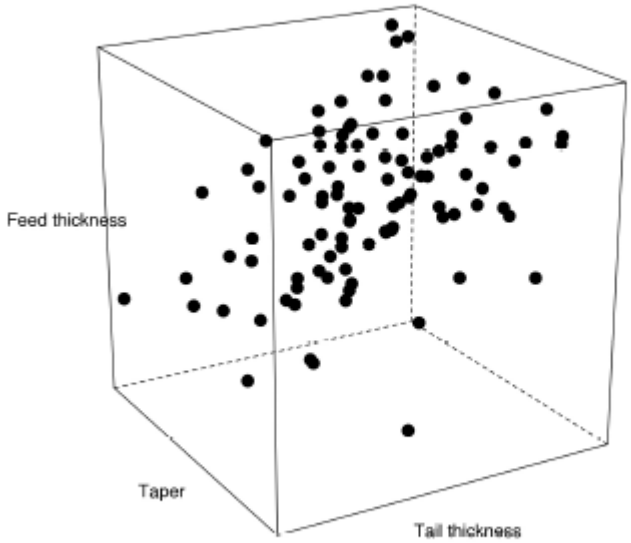


- x_1 = average tail thickness: average of thickness 1 and 4
- x_2 = average feed thickness: average of thickness 3 and 6
- x_3 = average taper: average of thickness 1, 2 and 3 subtracted from average thickness 4, 5, and 6

$$\underbrace{\mathbf{X}_{\text{raw}}}_{100\times 3}$$

What is a latent variable?





1. The fact that the entire board is thicker or thinner is captured by the feed and tail thickness measurements. These measurements are correlated with whatever physical phenomenon causes that average thickness to increase or decrease (e.g. spacing of the saw blades).
2. The third measurement, taper of the board, is capturing a different phenomenon in the system; possibly caused by how much the blades are skewed out of alignment.

Latent variable methods

(Multivariate statistical methods = latent variable methods)

- Principal component analysis (PCA)
  - a.k.a Karhunen-Loève transform or Hotelling transform
- Factor analysis
- Fisher's discriminant analysis
- Independent component analysis
- Principal component regression
- Partial least squares (or projection to latent structures)
- …..