Data Analytics using Python (Part 1)

Stephen Goldrick & Cheng Zhang
Lecturer in Bioprocess Digitalisation

Contents

- Review Jupyter-Notebook
- Introduction into Multivariate Data Analysis
- Simple Plotting Exercise and correlation calculation
- Live Demo of data analysis

Contents

- Review Jupyter-Notebook

Please raise your hand if you not can access it?
Please raise your hand if you can accessed through Anaconda?

Please raise your hand if you can accessed through Google Co-Lab?

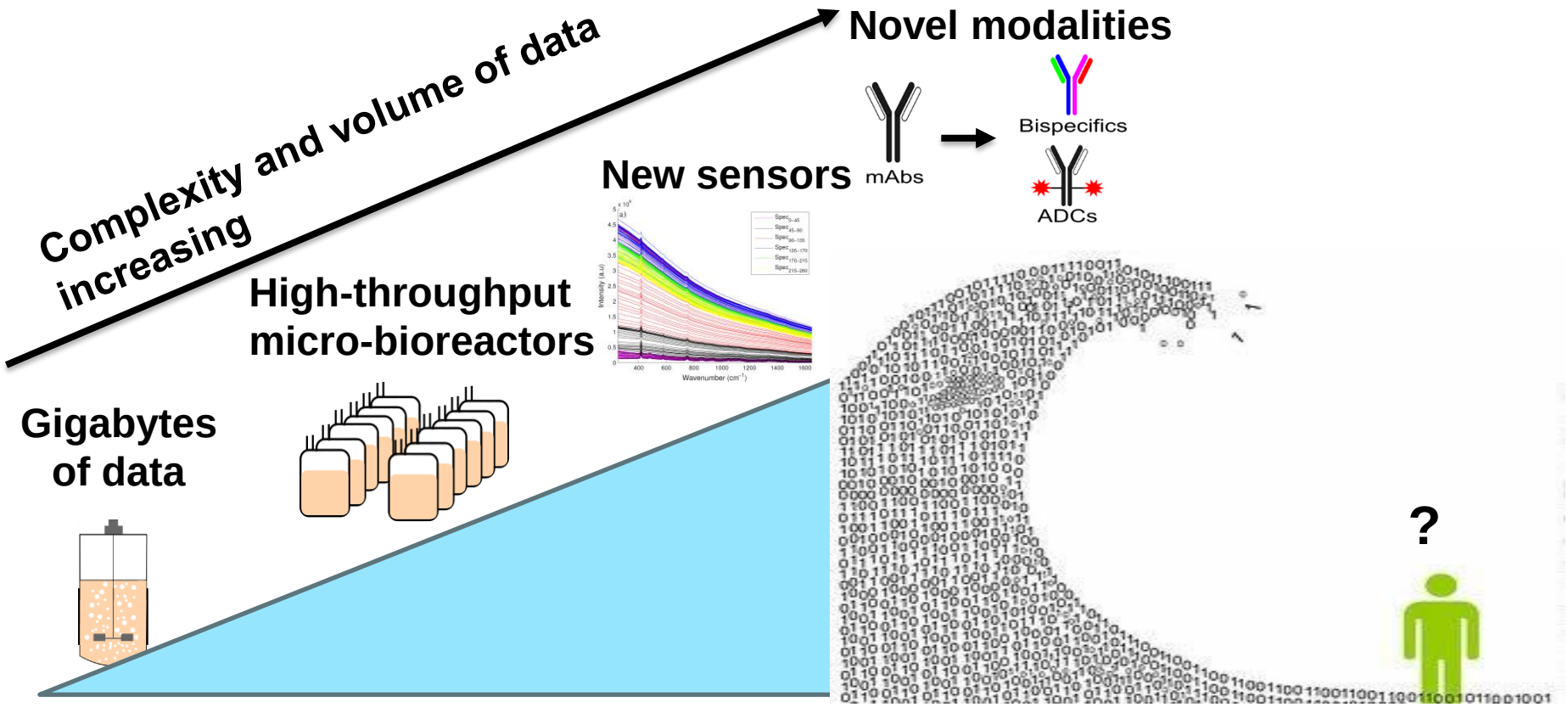
Overview of lecture series outline (Part 1)

- Introduction to Multivariate Data Analysis (MVDA)
- Understanding the Covariance and Correlation Matrix
  - Calculation of Covariance and Correlation matrix
- Overview of Python and Jupyter-notebook
  - Installation
  - Simple importation and plotting
  - Basic statistics
  - Sample Jupyter notebooks
  - Demonstration of advanced Jupyter-Notebook

Learning outcome from Python Data analytics lecture series Day 1

- To introduce the concepts and principles of effective and efficient multivariate data analysis
  - Calculation of the Covariance matrix
- How to crate and run a Jupyter-Notebook
  - Sample code for Plotting, correlation development

Data is getting bigger and more complicated



Unlock valuable insights -> Shorter timelines and cheaper drugs

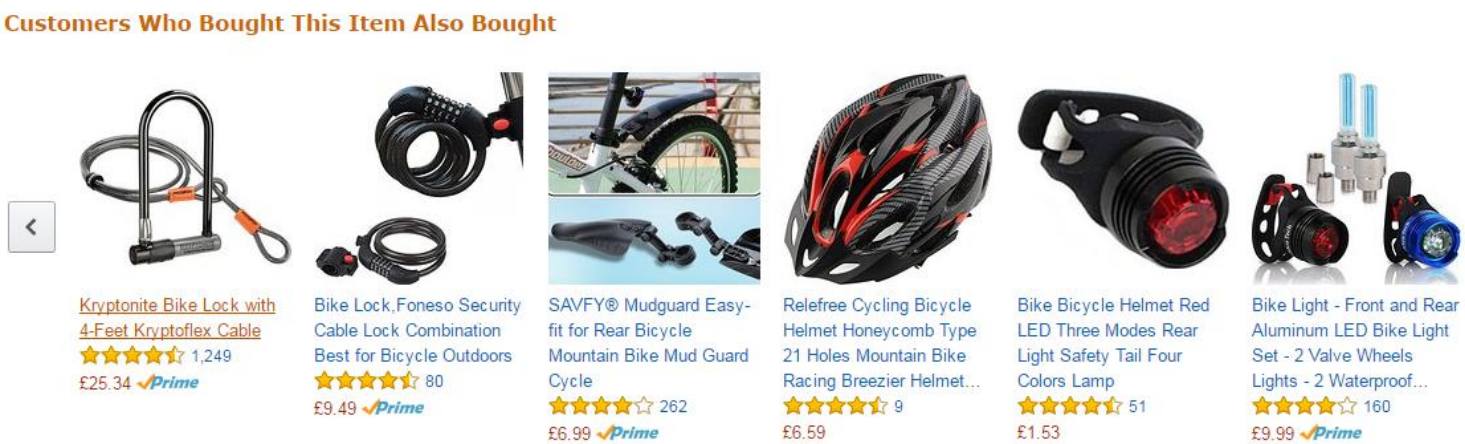
What is Multivariate Data Analysis (MVDA)?

- MVDA describes any statistical technique that is used to analyse data involving more than one variable
- MVDA reduces the dimensions (size) of large datasets, allowing for easier interpretation and enables identification of useful hidden correlations
- Typically most systems are described by more than one variable
  - Weather: wind, temperature, air pressure, humidity etc..
  - Batch fermentation: pH, temperature, DO_2, CO_2 flow rates, acid/base additions, osmolality, seeding density etc…

MVDA can improve our understanding of a complex process or system with multiple variables

Google (suggested searches), Amazon (suggested purchases), Facebook (suggested feeds) are all based on MVDA on a large scale

- Buy a **bike light** on Amazon:
  - Amazon's MVDA engine suggests



People who bought bike lights also bought bike locks, bike helmets etc

i.e Purchase of bike lights is highly correlated with purchase of other bike accessories

Spotify user recommendations

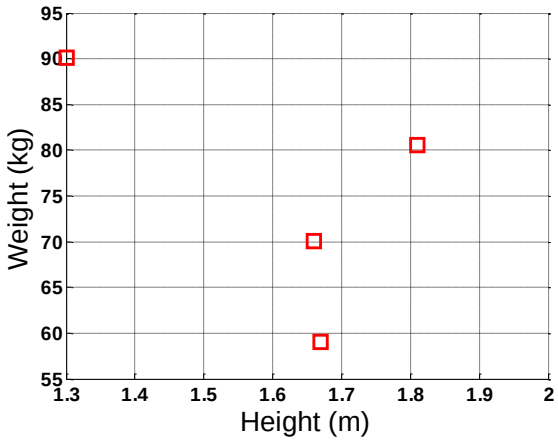
Person A loves listening to:

Spotify users love listening:
To both ***Taylor Swift***-*We are never getting back together* and ***Beyonce***-*Single Ladies*

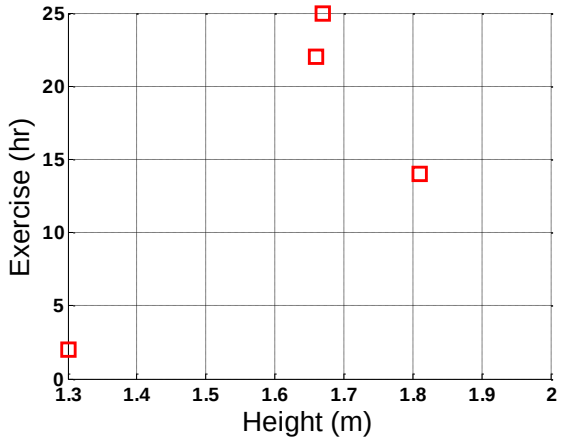
i.e: Two songs are highly **Correlated**

Spotify recommends to **Person A** to listen to ***Beyonce***-*Single ladies* based on Person A listening to ***Taylor Swift***-*We are never getting back together*. Recommendation based on the **High Correlation** between these songs.

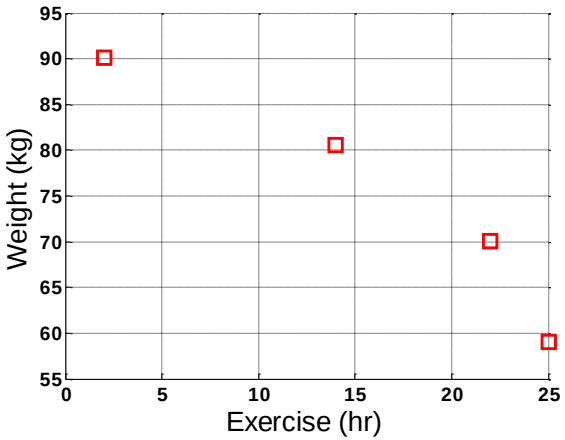
Multivariate analysis (two variables at a time)



No obvious correlation between Weight(kg) and Height(m)



No obvious correlation between Exercise(hr) and Height(m)



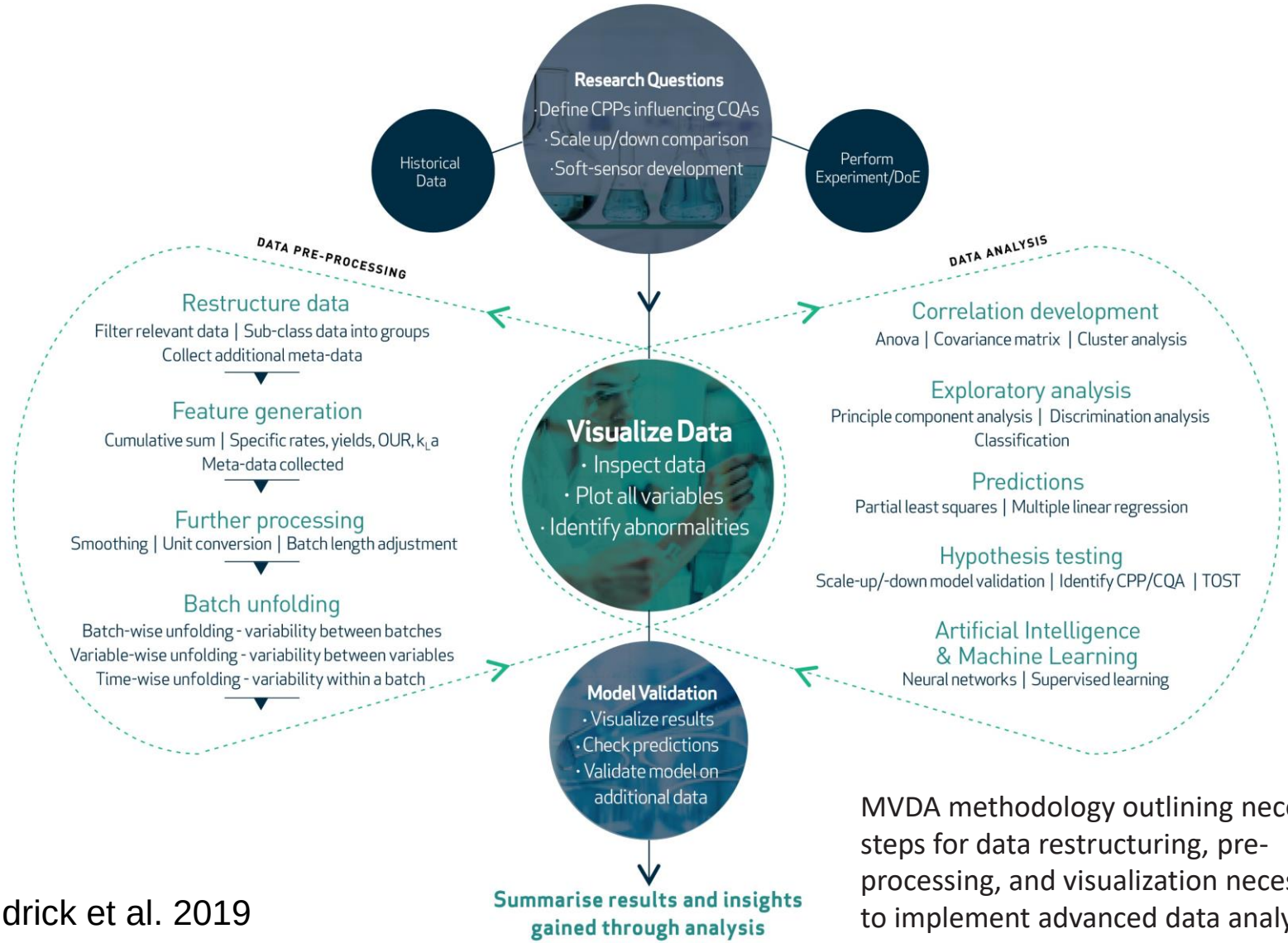
Correlation between Weight(kg) and Exercise(hr)
As exercise increases weight decreases
I.E Weight and Exercise are negatively **correlated**

<u>Question</u>
What if we want to compare 100, 1000, 10,000 variables?

<u>Solution</u>
Plot all possible variables against each other?
Very time consuming and difficult to interpret results

<u>Solution</u>
<u>Multivariate Data Analysis (MVDA)</u>

Data analysis methodology



Goldrick et al. 2019

MVDA methodology outlining necessary steps for data restructuring, pre-processing, and visualization necessary to implement advanced data analytics on complex biopharmaceutical data sets.

Different MVDA methods (there are 100's)

An example of some common MVDA techniques:

- **Linear regression (LR):** $y = \beta_1 X_1$
- **Multiple linear regression (MLR):** $y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
- **Covariance:** Analysis of relationships between variables $\mathrm{cov}(X_1, X_2)$
- **Discrimination/factor analysis (DA/FA):** Defines discrimination features related to associating variables different groups or clusters $DA(X)$
- **Principal Component Analysis (PCA):** Determines correlation between variables in the X-data structure $PCA(X)$
- **Partial Least Squares (PLS):** Determines relationships between the X variables to enable predictions of the Y variables $PLS(X, Y)$

Covariance – Basic statistics required

- **Mean($\bar{X}$)**: average point within dataset

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- **Standard deviation (std)**: A measure of the spread of data points in a dataset: i.e it is the average distance from mean of the data set to a point.

$$std = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$$

- **Variance (std²)**: squared standard deviation

$$\text{var} = std^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

- **Covariance (cov)**: measures the strength of correlation between two variables (X,Y) or more sets of variables

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

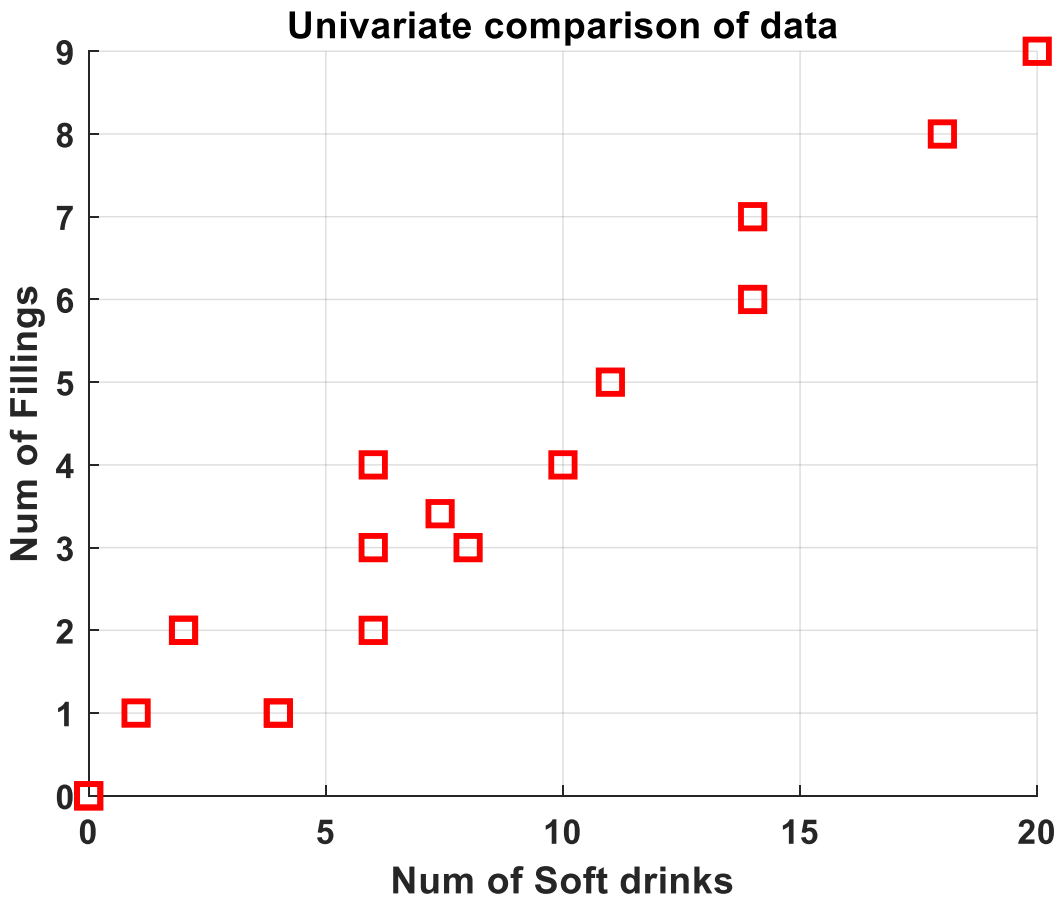
Consider a simple data set

- A survey was carried out involving 17 students, the students were asked two questions?
  - How many soft drinks they drink on average in a week
  - How many tooth fillings they have?

Num of soft drinks per week (#)	Num of tooth fillings (#)
6	2
14	6
20	9
10	4
4	1
0	0
11	5
6	3
8	3
14	7
18	8
2	2
6	4
0	0
2	2
1	1
4	1

Research question: Is the number of soft drinks consumed correlated to number of tooth fillings?

Clear positive correlation observed



Step 2 – Calculate basic statistics for data

Number of Soft drinks (X)	Number of Fillings (Y)	$(X-X_{mean})$	$(Y-Y_{mean})$	$[(X-X_{mean}))]^2$	$[(Y-Y_{mean})]^2$	$(X-X_{mean})(Y-Y_{mean})$
6.0	2.0	-1.4	-1.4	2.0	2.0	2.0
14.0	6.0	6.6	2.6	43.4	6.7	17.1
20.0	9.0	12.6	5.6	158.5	31.2	70.3
10.0	4.0	2.6	0.6	6.7	0.3	1.5
4.0	1.0	-3.4	-2.4	11.6	5.8	8.2
0.0	0.0	-7.4	-3.4	54.9	11.6	25.3
11.0	5.0	3.6	1.6	12.9	2.5	5.7
6.0	3.0	-1.4	-0.4	2.0	0.2	0.6
8.0	3.0	0.6	-0.4	0.3	0.2	-0.2
14.0	7.0	6.6	3.6	43.4	12.9	23.6
18.0	8.0	10.6	4.6	112.1	21.1	48.6
2.0	2.0	-5.4	-1.4	29.3	2.0	7.6
6.0	4.0	-1.4	0.6	2.0	0.3	-0.8
0.0	0.0	-7.4	-3.4	54.9	11.6	25.3
2.0	2.0	-5.4	-1.4	29.3	2.0	7.6
1.0	1.0	-6.4	-2.4	41.1	5.8	15.5
4.0	1.0	-3.4	-2.4	11.6	5.8	8.2
7.4	**3.4**		**Covariance**	**38.5**	**7.6**	**16.6**
X_{mean}	Y_{mean}			Cov(X,X)	Cov(Y,Y)	Cov(X,Y)

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)}$$

$$\text{Covariance Matrix} = \begin{bmatrix} \mathrm{Cov(X,X)} & \mathrm{Cov(Y,X)} \\ \mathrm{Cov(X,Y)} & \mathrm{Cov(Y,Y)} \end{bmatrix} = \begin{bmatrix} 38.5 & 16.6 \\ 16.6 & 7.6 \end{bmatrix}$$

Step 3- Calculate the covariance matrix

- Covariance of two variables = $\begin{pmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{pmatrix}$
- Therefore covariance matrix of sample dataset: $\begin{pmatrix} 38.5 & 16.6 \\ 16.6 & 7.6 \end{pmatrix}$
- What does the covariance matrix tell us?

$$\begin{pmatrix} 38.5 & 16.6 \\ 16.6 & 7.6 \end{pmatrix}$$

COV(X,X) equals variance of X data i.e spread of X data from mean

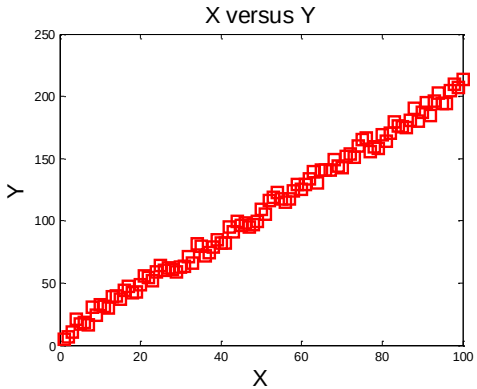
Cov(X,Y) = Cov(Y,X)

N.B Value of covariance between two variables [(cov(X,Y)] is not important only the sign is important

Cov(Y,X) is **<u>Positively</u>** correlated i.e an increase in number of softdrinks results in an increase in the number of fillings

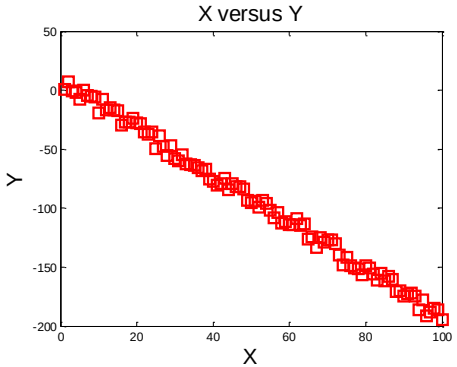
COV(Y,Y) equals variance of Y data i.e spread of Y data from mean

Covariance matrix [cov(x,y)]



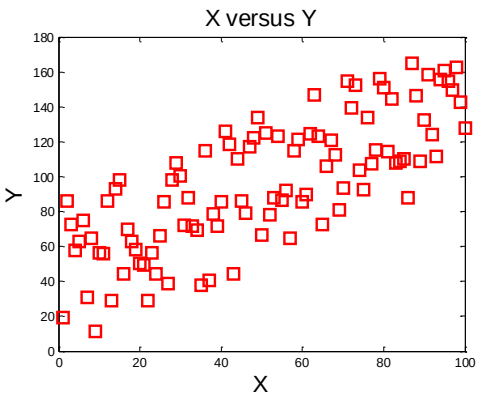
$$\mathrm{cov}(x,y)=\begin{pmatrix}840 & 1700\\ 1700 & 3400\end{pmatrix}$$

Strong positive correlation with X and Y



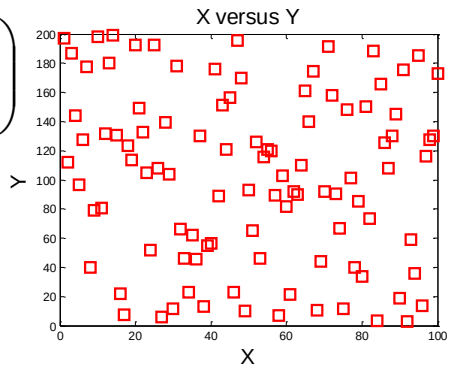
$$\mathrm{cov}(x,y)=\begin{pmatrix}840 & -1700\\ -1700 & 3400\end{pmatrix}$$

Strong negative correlation with X and Y



$$\mathrm{cov}(x,y)=\begin{pmatrix}810 & 1700\\ 1700 & 4300\end{pmatrix}$$

Moderate positive correlation with X and Y



$$\mathrm{cov}(x,y)=\begin{pmatrix}840 & -200\\ -200 & 3400\end{pmatrix}$$

No correlation between X and Y

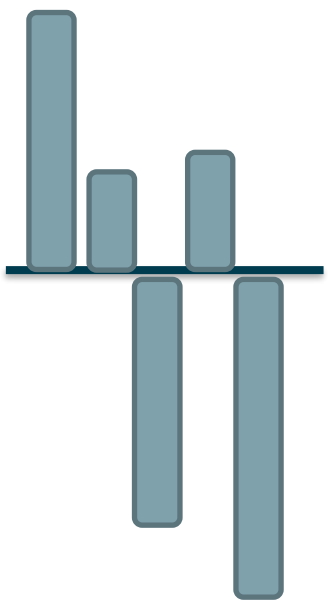
Covariance measures the strength of correlation between two variables

The importance of scaling

- Scaling can help transform the data to a suitable form for analysis.
  - Essential when analysing variables that are measured in different units. i.e Volume ranging from 10,000-20,000 Litres compared to pH ranging from 6.5 – 7.1
  - Allows variables with small variances to have equal weight during data analysis

Raw data

$$X$$

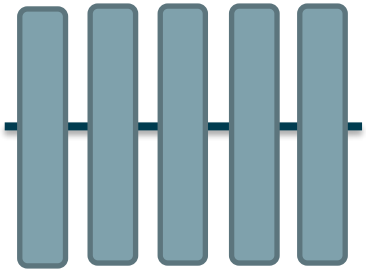


Mean centered data:

$$X - \bar{X}$$

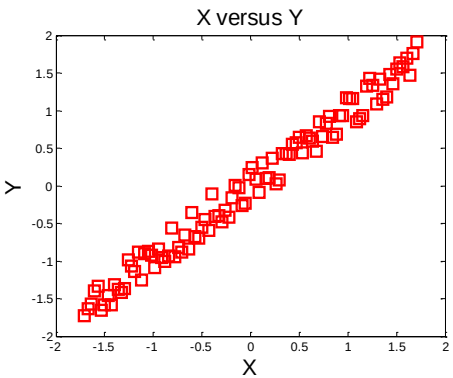
Mean centered data and scaled data:

$$\frac{X - \bar{X}}{std(X)}$$



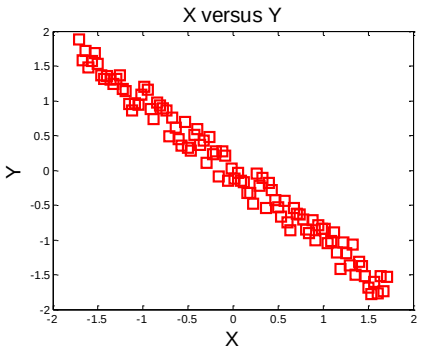
Also called normalising data or autoscaling

Scaling data before calculating Covariance matrix [cov(x,y)]



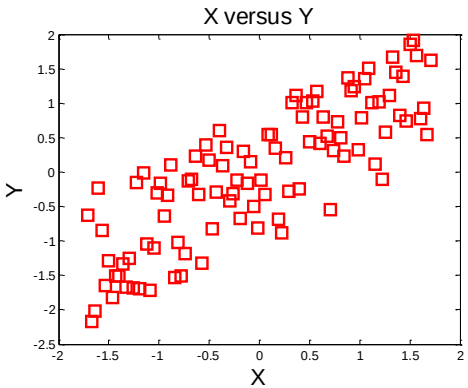
$$\mathrm{cov}(x,y) = \begin{pmatrix} 1 & 0.98 \\ 0.98 & 1 \end{pmatrix}$$

Strong positive correlation with X and Y



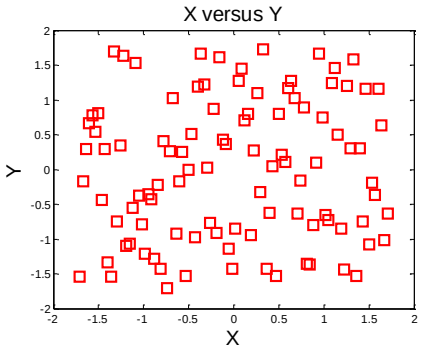
$$\mathrm{cov}(x,y) = \begin{pmatrix} 1 & -0.98 \\ -0.98 & 1 \end{pmatrix}$$

Strong negative correlation with X and Y



$$\mathrm{cov}(x,y) = \begin{pmatrix} 1 & 0.77 \\ 0.77 & 1 \end{pmatrix}$$

Moderate positive correlation with X and Y



$$\mathrm{cov}(x,y) = \begin{pmatrix} 1 & -0.01 \\ -0.01 & 1 \end{pmatrix}$$

No correlation between X and Y

Scaling data before calculating Covariance Matrix enables easier interpretation of results essential generates the Correlation matrix

i.E Correlation matrix is covariance of scaled data

Step 3- Calculate the correlation matrix

- Covariance of two variables = $\begin{pmatrix} Corr(X,X) & Corr(X,Y) \\ Corr(Y,X) & Corr(Y,Y) \end{pmatrix}$
- Therefore correlation matrix of sample dataset: $\begin{pmatrix} 1 & 0.97 \\ 0.97 & 1 \end{pmatrix}$
- What does the covariance matrix tell us?

$$\begin{pmatrix} 1 & 0.97 \\ 0.97 & 1 \end{pmatrix}$$

CORR(X,X) equals variance of X data i.e spread of X data from mean (= 1 as we normalised the data)

CORR(Y,X) is **<u>Positively</u>** correlated i.e an increase in number of softdrinks results in an increase in the number of fillings

Corr(X,Y) = Corr(Y,X)

N.B Value of correlation between two variables [(cov(X,Y)] is important and the sign indicated a positive or negative relationship

CORR(Y,Y) equals variance of Y data i.e spread of Y data from mean (= 1 as we normalised the data**)**

Example: Calculate Covariance matrix (Pen and Paper – example)

- A biopharmaceutical company has ran five cell culture experiments to identify the relationship between the pH set-point and the final mAb titre.
  - Using pen and paper generate the covariance matrix for this data set and discuss?

	pH set-point (-)	mAb titre (g L^{-1})
Cell culture 1	6.6	1.55
Cell culture 2	6.8	3.1
Cell culture 3	7	4.6
Cell culture 4	7.2	6.2

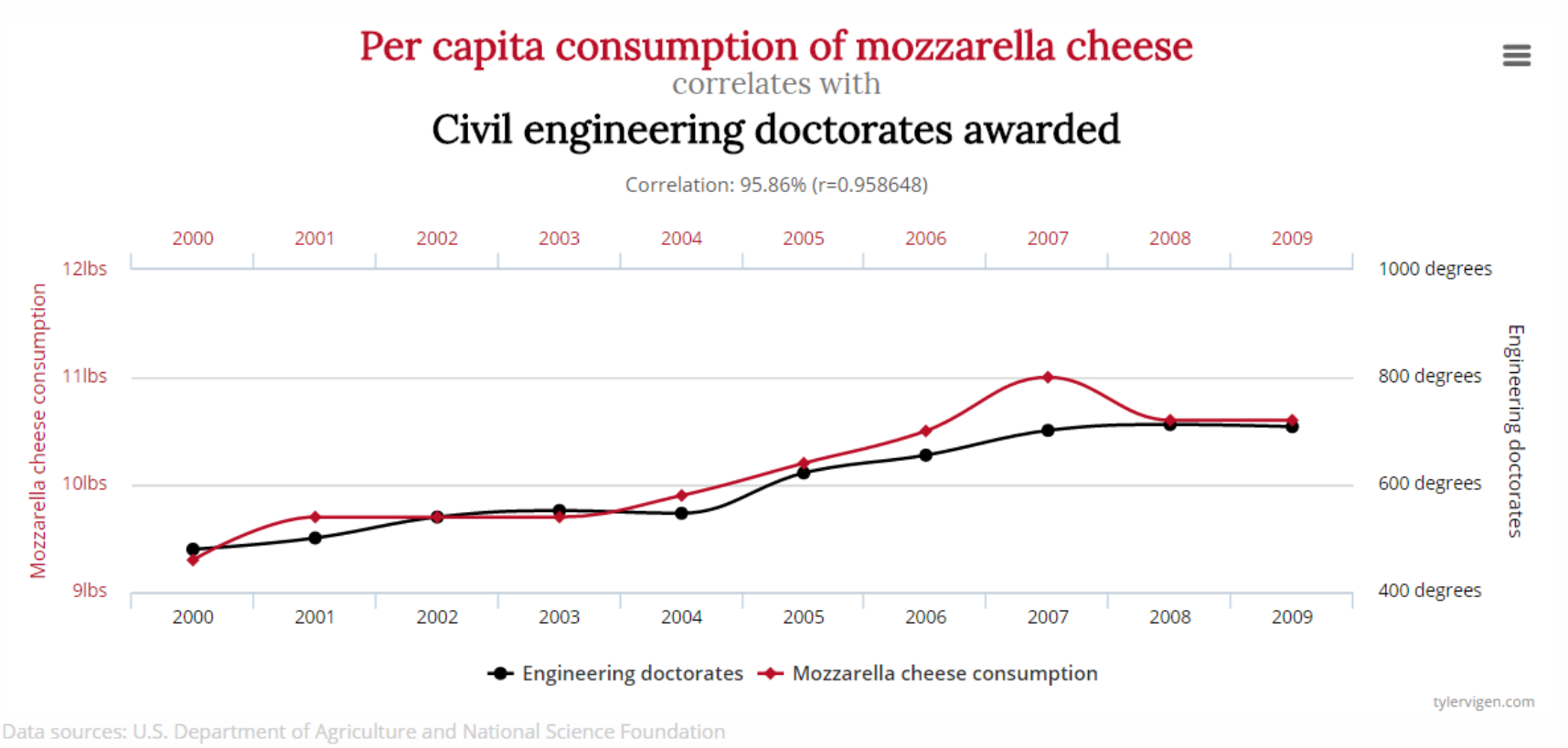
Covariance example: Solution

pH Set-point (-)	mAb Titre (g L^{-2})					
X	Y	X-Xmean	Y-Ymean	(X-Xmean)^2	(Y-Ymean)^2	(X-Xmean)(Y-Ymean)
6.6	1.55	-0.3	-2.3125	0.09	5.347656	0.69375
6.8	3.1	-0.1	-0.7625	0.01	0.581406	0.07625
7	4.6	0.1	0.7375	0.01	0.543906	0.07375
7.2	6.2	0.3	2.3375	0.09	5.463906	0.70125
6.9	**3.8625**			**0.066667**	**3.978958**	**0.515**

$$\mathrm{cov}(X,Y) = \begin{pmatrix} 0.66 & 0.515 \\ 0.515 & 3.979 \end{pmatrix}$$

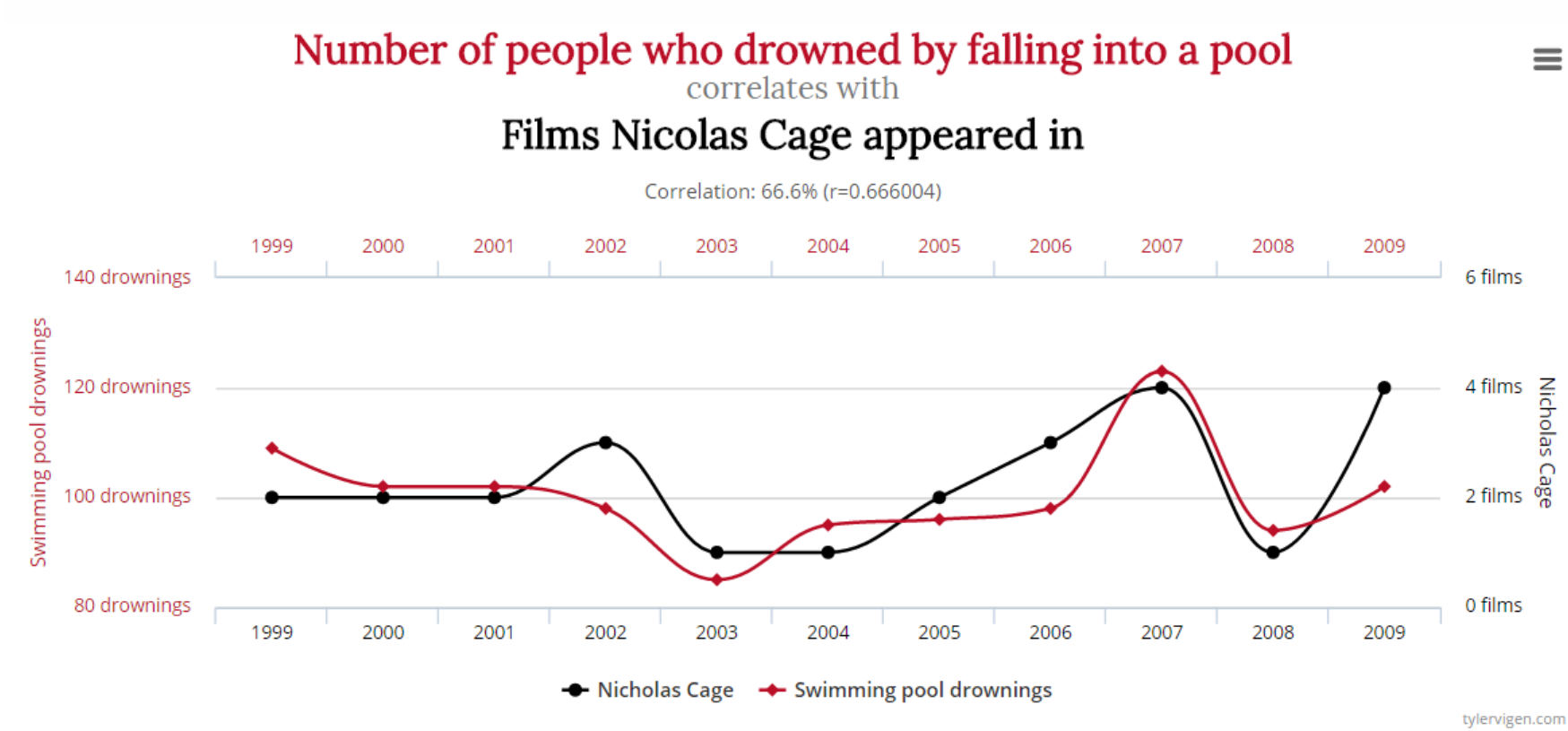
Correlation≠Causation

- Correlation does not imply causation
  - If two variables are correlated it does not always imply that changes in one will cause a change in the second
  - An understanding of the data set is crucial to ensure correct interpretation of the developed correlations



Correlation≠Causation

- Correlation does not imply causation
  - If two variables are correlated it does not always imply that changes in one will cause a change in the second
  - An understanding of the data set is crucial to ensure correct interpretation of the developed correlations

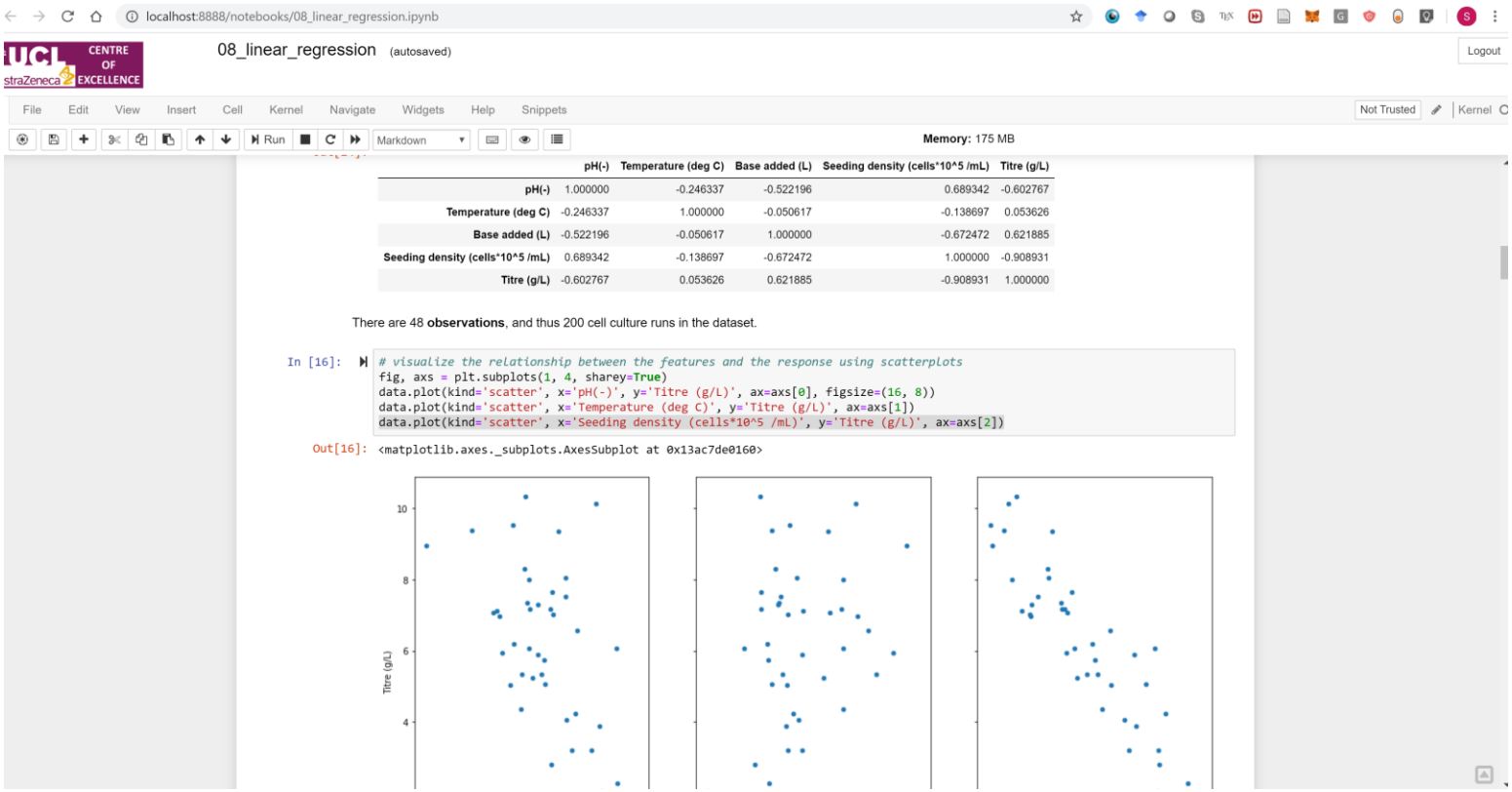


Introduction to Python and Jupyter Notebook

- Python one of the faster growing languages
- Ideal for data analysis
- Suitable for Big-Data, Industry 4.0 and Artificial intelligence application
- Open source so its free
- On-line community (Best in the world and growing)

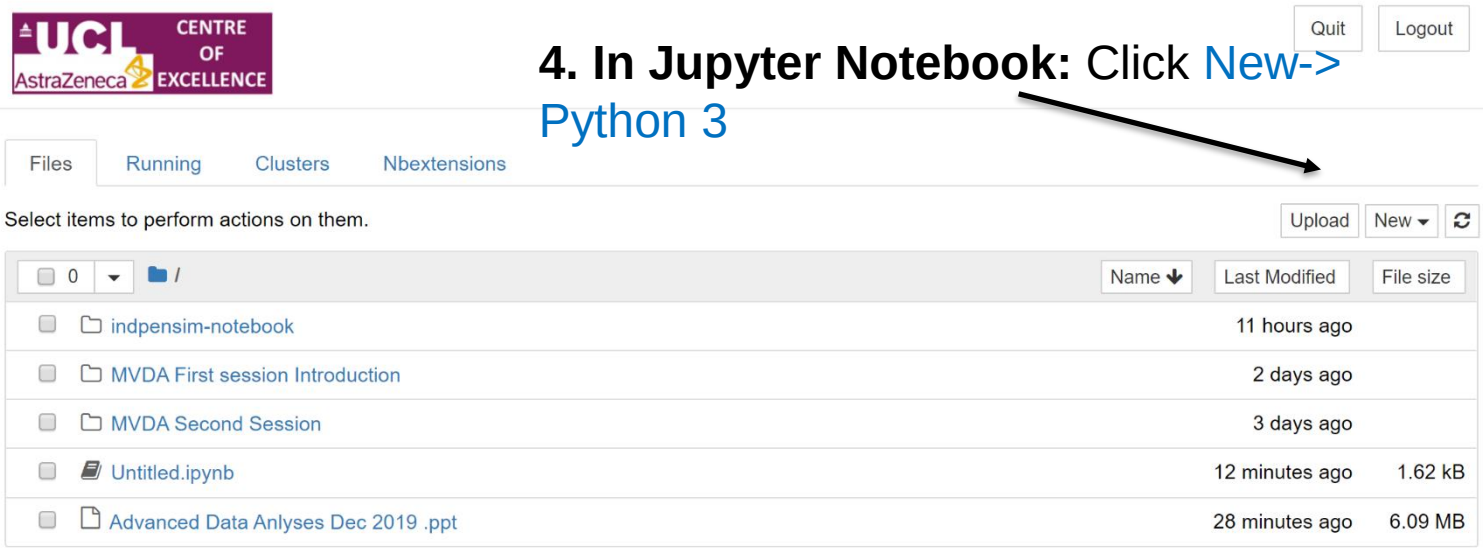
Jupyter- Notebook – front end for Python

- (Makes Python Look “Pretty”)
- Excellent tool to analysis and visualise your data
- Generate reports that are interactive



Jupyter-Notebook is essentially Excel on Steroids!!

Jupyter-Notebook



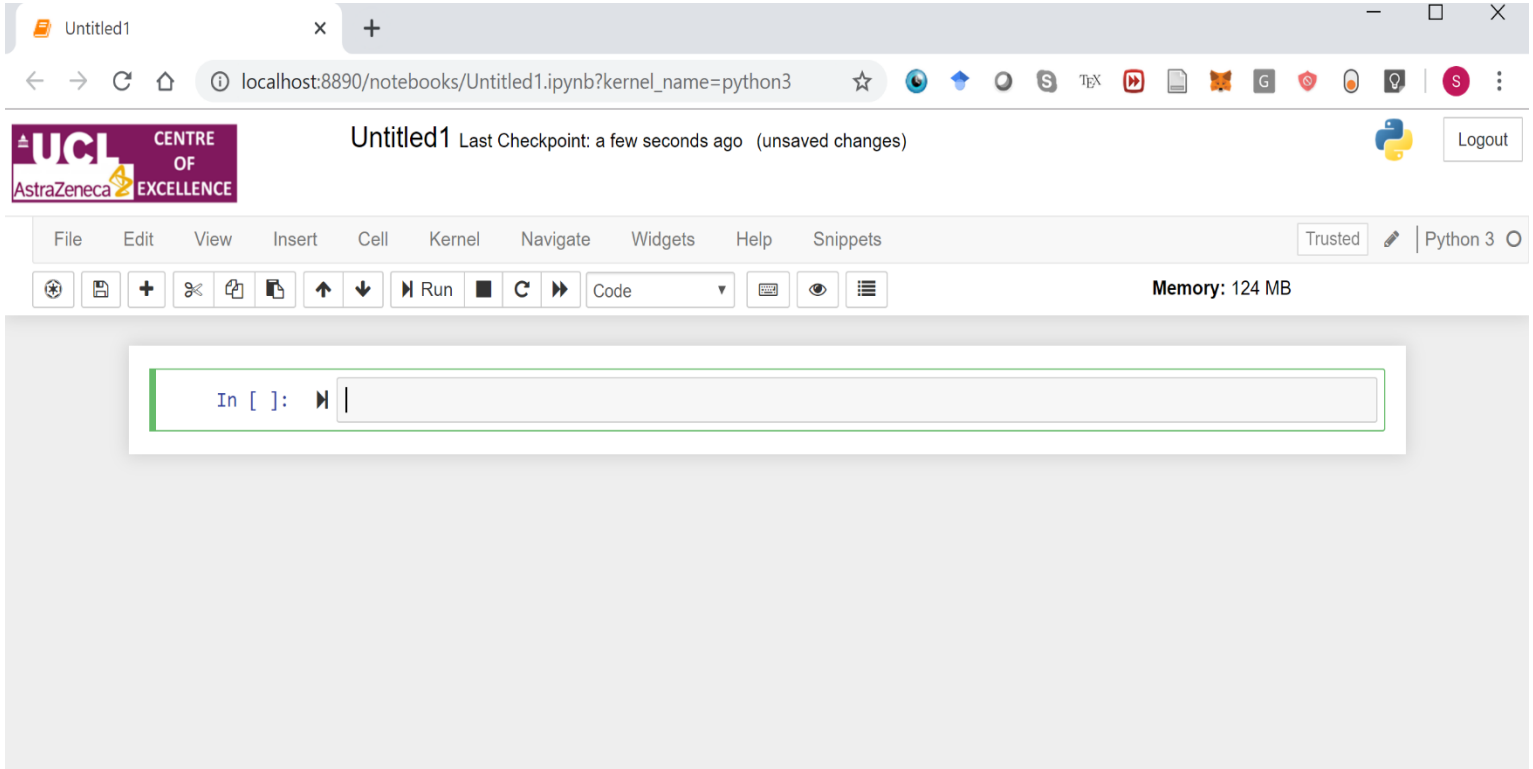
4. In Jupyter Notebook: Click New-> Python 3

Jupyter: can be used an interface to support multiple languages including Matlab

- Need to set up Kernel for each programming language you are interested in (Google it)

Ju – Pyt eR

Jupyter-Notebook



Example: Calculate Covariance matrix (Using Jupyter Notebook)

- A biopharmaceutical company has ran five cell culture experiments to identify the relationship between the pH set-point and the final mAb titre.
  - Using Jupyter Notebook generate the covariance matrix for this data set and discuss?
  - Using Jupyter Notebook generate the correlation matrix for this data set and discuss?

	Cell culture 1	Cell culture 2	Cell culture 3	Cell culture 4
pH set-point (-)	6.6	6.8	7.0	7.2
mAb titre (g L^{-1})	1.55	3.1	4.6	6.2

Under Kernel click

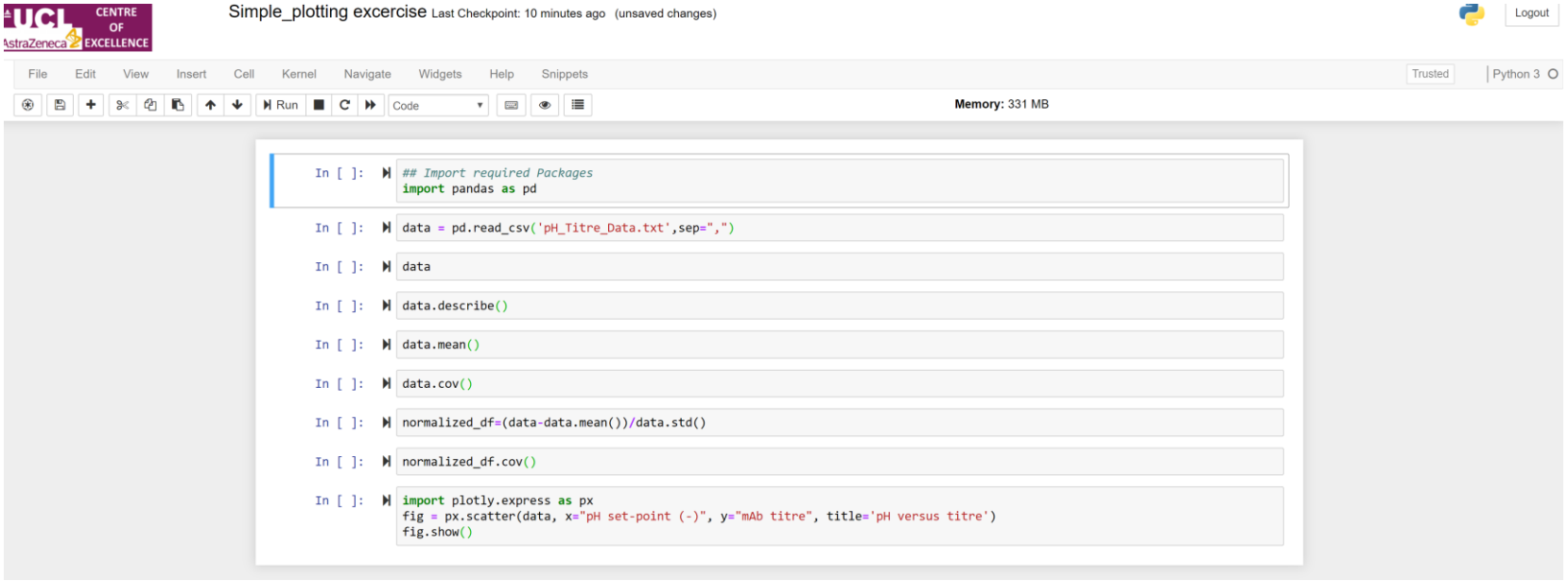
Restart & Clear output

Under Cells click

Run all

OR

(Shift Enter to run each cell individually)



Mostly likely you will an error

Modulenotfound: No Module named "plotly"

- Open anaconda prompt and type

  pip install plotly

  Or use a code cell in Jupyter-notebook

  pip install plotly

Click "Restart kernel and run all"

Python – Jupyter – Notebook

- Each cell can be run individually (ctrl Enter)
- Python packages need to imported before they are used

In [1]:
```
## Import required Packages
import pandas as pd
```

In [2]:
```
data = pd.read_csv('pH_Titre_Data.txt',sep=",")
```

Data will be saved as Dataframe called data

Name of file to import

Data is separated by comma

Pandas (pd) inbuilt function to read csv file
(File must be in the same folder as Jupyter-notebook!!)

Python – Jupyter – Notebook

- *data* saved as Dataframe
  - Rows are observations
  - Columns are variables

```
data
```

	pH set-point (-)	mAb titre
0	6.6	1.55
1	6.8	3.10
2	7.0	4.60
3	7.2	6.20

Python – Jupyter – Notebook

- *.describe() for* Dataframe
  - Prints out basic stats for data

```
data.describe()
```

6]:

	pH set-point (-)	mAb titre
count	4.000000	4.000000
mean	6.900000	3.862500
std	0.258199	1.994733
min	6.600000	1.550000
25%	6.750000	2.712500
50%	6.900000	3.850000
75%	7.050000	5.000000
max	7.200000	6.200000

Python – Jupyter – Notebook

- Dataframe enables simple calculations of useful statistics

```
data.mean()
```

```
pH set-point (-)    6.9000
mAb titre           3.8625
dtype: float64
```

```
data.cov()
```

	pH set-point (-)	mAb titre
pH set-point (-)	0.066667	0.515000
mAb titre	0.515000	3.978958

Python – Jupyter – Notebook

- How to normalise the data and then calculate correlation matrix

```
normalized_df=(data-data.mean())/data.std()
```

```
normalized_df.cov()
```

	pH set-point (-)	mAb titre
pH set-point (-)	1.000000	0.999927
mAb titre	0.999927	1.000000

Consider a simple data set

- A survey was carried out involving 17 students, the students were asked two questions?
  - How many soft drinks they drink on average in a week
  - How many tooth fillings they have?

Research question: Is the number of soft drinks consumed correlated to number of tooth fillings?

Answer this question using Jupyter-Notebook data is saved as Soft_drinks_filling_data.txt

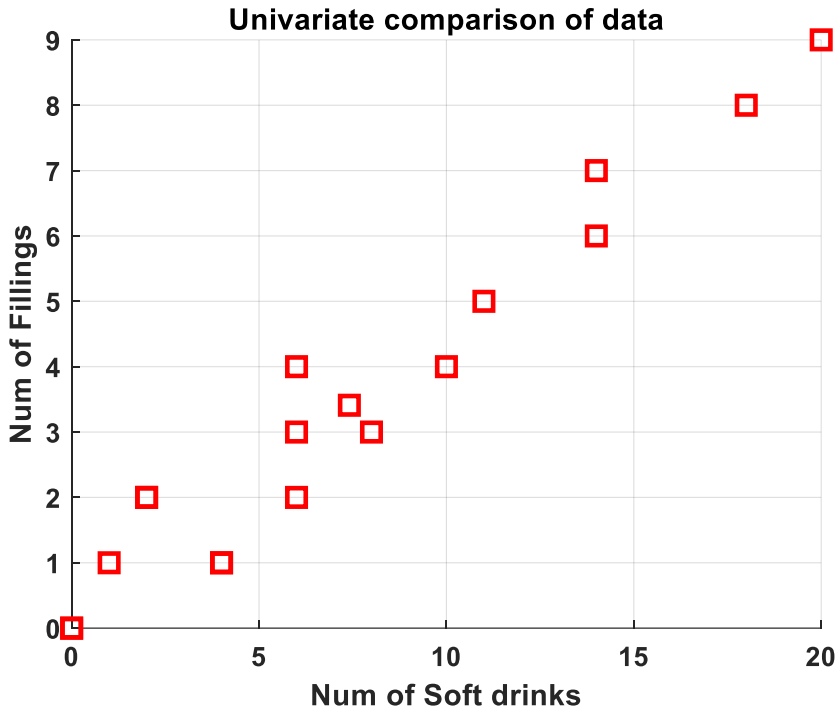
Num of soft drinks per week (#)	Num of tooth fillings (#)
6	2
14	6
20	9
10	4
4	1
0	0
11	5
6	3
8	3
14	7
18	8
2	2
6	4
0	0
2	2
1	1
4	1

Python example

- Import data file called “Soft_drinks_filling_data.txt ”
- Calculate Covariance and Correlation Matrix of data of given data set
- Plot the raw data

Num of soft drinks per week (#)	Num of tooth fillings (#)
6	2
14	6
20	9
10	4
4	1
0	0
11	5
6	3
8	3
14	7
18	8
2	2
6	4
0	0
2	2
1	1
4	1

Correlation coefficient easier to understand



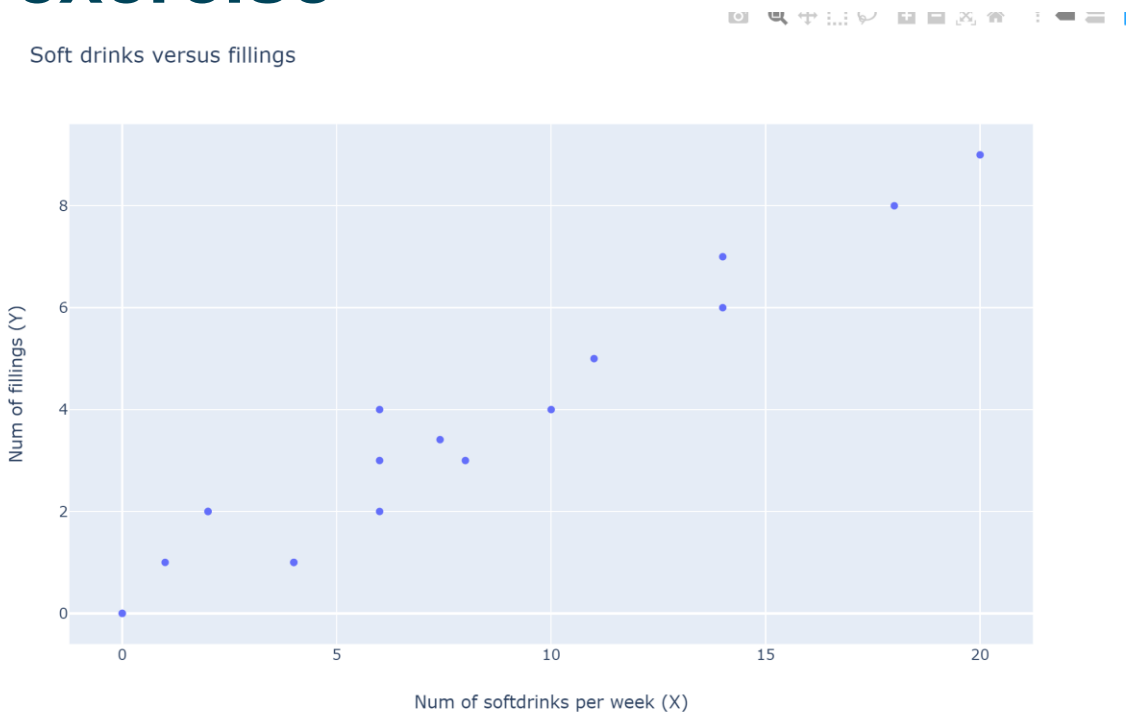

```
data.cov()
```

	Num of softdrinks per week (X)	Num of fillings (Y)
Num of softdrinks per week (X)	36.242215	15.653979
Num of fillings (Y)	15.653979	7.183391

```
normalized_df.cov()
```

	Num of softdrinks per week (X)	Num of fillings (Y)
Num of softdrinks per week (X)	1.000000	0.970181
Num of fillings (Y)	0.970181	1.000000

Python exercise




```
data.cov()
```

	Num of softdrinks per week (X)	Num of fillings (Y)
Num of softdrinks per week (X)	36.242215	15.653979
Num of fillings (Y)	15.653979	7.183391

```
normalized_df=(data-data.mean())/data.std()
```

```
normalized_df.cov()
```

	Num of softdrinks per week (X)	Num of fillings (Y)
Num of softdrinks per week (X)	1.000000	0.970181
Num of fillings (Y)	0.970181	1.000000

Live Demo of Data analysis for a Bioreactor problem

Multivariate data analysis: 48 batch records

- **Big data (Available on Moodle)**
  - Download and unzip the folder on desktop
  - Open text file called "Bioreactor_data_headers_v1.csv"
- **Overview of data set**
  - 48 mammalian cell culture batches were operated at different temperatures and pH set points and initial RPMS
  - Final Titre and Aggregation was recorded for each batch and classed as
    - 1: Good Batches for Aggregation below 5 g/L (Classification = 1)
    - 2: Average Batches Aggregation between 3-5 g/L g/L (Classification = 2 in Column 6)
    - 3: Bad Batches for aggregation over 8 g/L (Classification = 3)

pH	Temp	RPM	Titre	Aggregation	Classification
6.2309	23.781	2433.8	0.39391	12.133	3
6.319	26.805	2614.2	0.59547	5.614	2
6.1867	13.402	2396.9	-1.1285	10.902	3
6.2093	37.635	2202.6	0.96203	10.18	3
5.9788	27.495	2442.6	-0.19477	8.5016	3
6.0208	32.504	2094.8	2.9491	10.453	3
5.9951	41.044	2018.5	3.0823	10.63	3
6.0619	34.635	2244.5	2.3326	10.755	3
6.2883	31.368	2047.3	1.8792	9.3396	3
6.0482	28.663	2010.7	2.6717	9.5146	3
6.7819	32.798	2063	4.5927	6.777	2
6.5332	38.517	1980.3	4.4523	5.7471	2
6.0466	24.293	1972	2.4619	7.2148	2
6.2408	34.082	1992.5	3.3688	7.9356	3
6.4596	39.85	1770	1.9887	9.2769	3
6.1928	25.061	1721.4	3.9809	5.6065	2
6.4492	39.173	1569.8	4.5535	6.021	2

- Data set = [48x6] i.e 24 batches and 5 variables

Cell culture MVDA example

Research question:

Can PCA identify the process conditions that influences final aggregation and Titre?

What are the main factors that result in high/low titres and Aggregation?

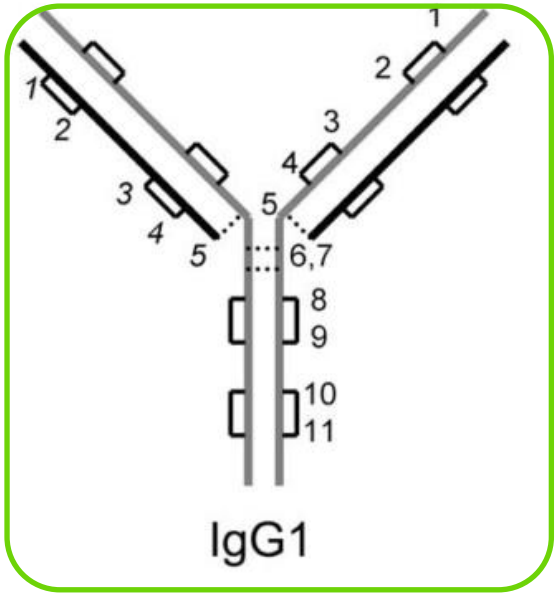
Are there any obvious groups in the data set?

Case Study:
Multivariate Data Analysis (MVDA) to help determine product quality issues on mammalian cell culture

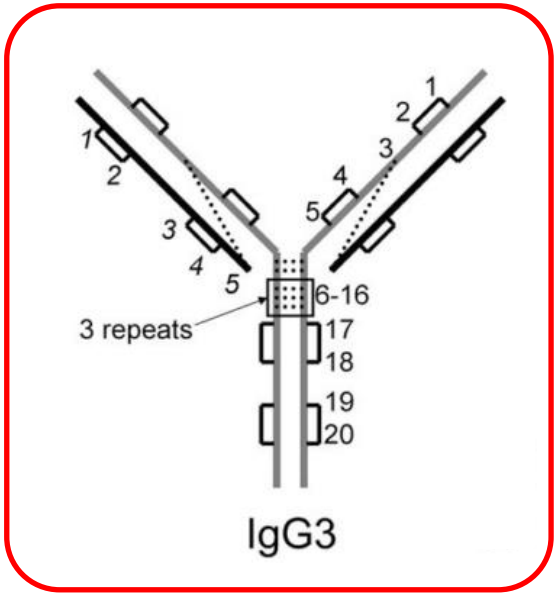
Application of MVDA for root cause determination of mAb heterogeneities on an industrial process

- A trisulfide bond (TSB) was detected on a novel recombinant antibody-peptide fusion expressed in mammalian cell culture during R&D for one of MedImmune's primary drug candidates

Correct folding



Incorrect folding - TSB



What process changes result in TSB formation?

Design of Experiment to investigate product heterogeneities

DoE Design: 3-Level Fractional Factorial (43 Cell culture runs)

Factors Manipulated:

Temperature: 34, 35.5, 37°C

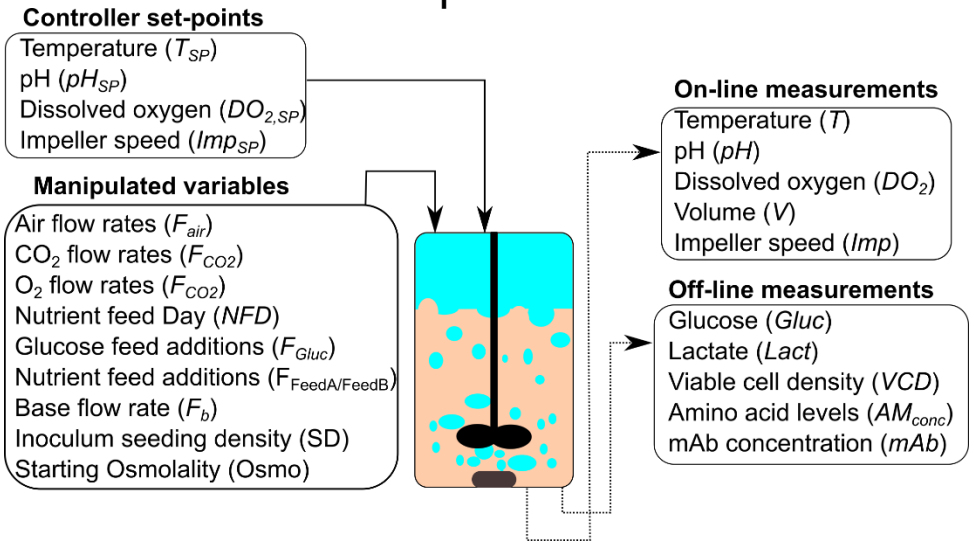
pH: 6.8-7.2

Initial Nutrient Feed Day: Day -1,2,3,4

Nutrient Feed Volume: 80-120%

Seeding Density: 50-150%

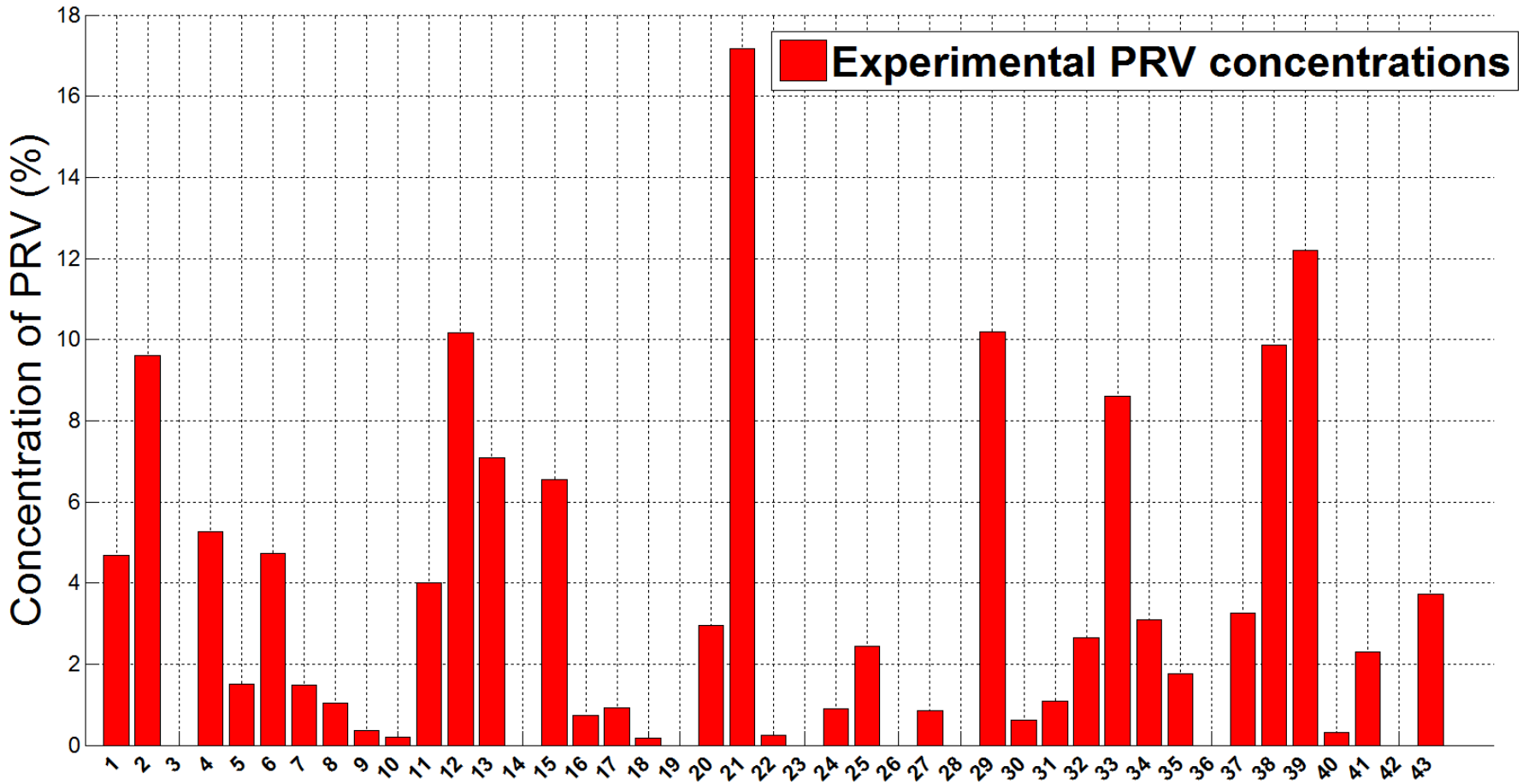
Complex process with multiple variables



- Difficult challenge
  - Consider a single variable (pH) recorded every 10 seconds for 14 days for each vessel
  - Data size equals 1 x 8640 x 14 = 1,209,600 data points
  - 25 variables therefore 25 x 1,209,600 = Massive Data set

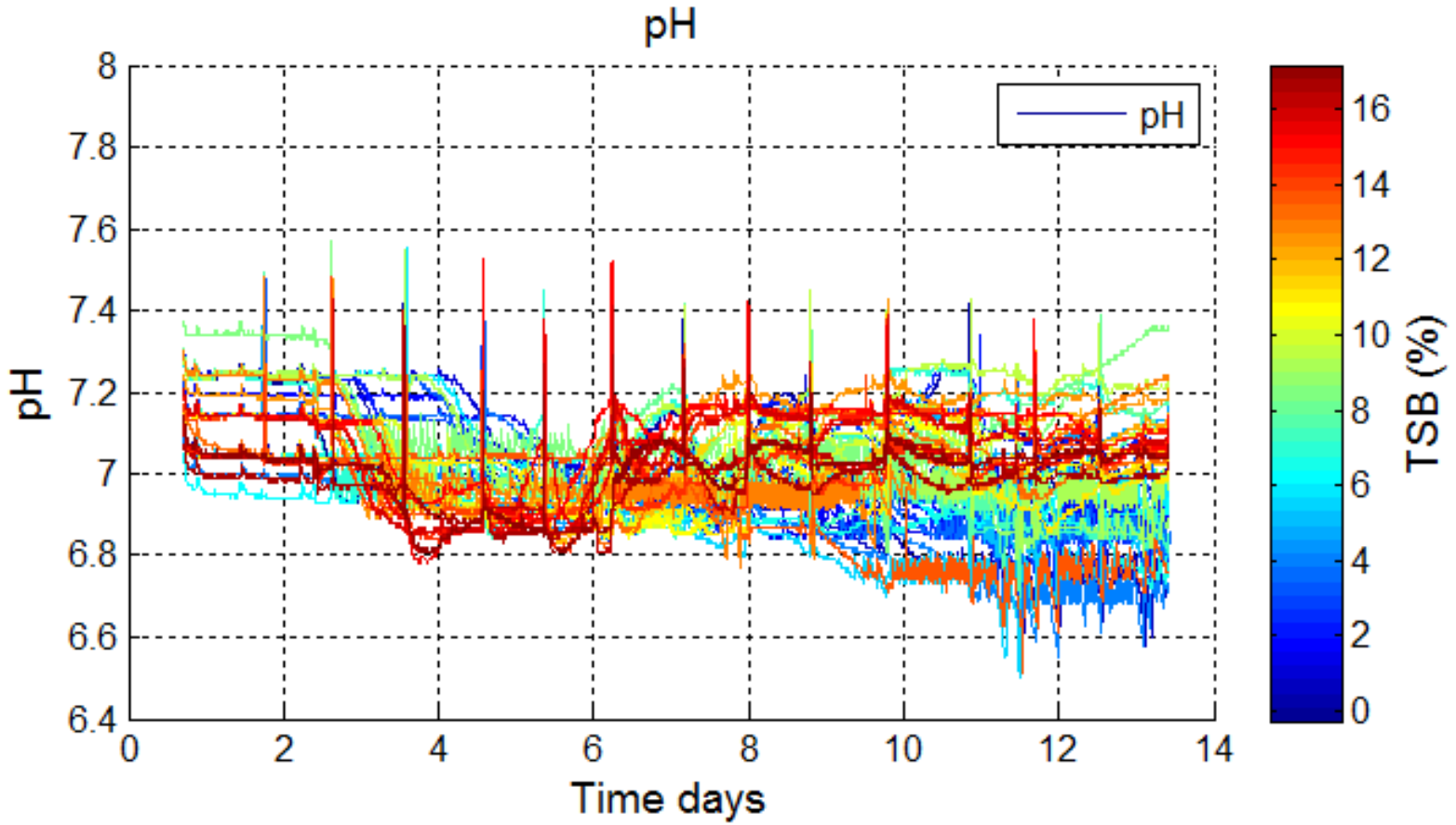
MVDA is necessary to analysis this complex data set

Concentration of TSB for 43 culture runs



What are the key process variables that are driving high TSB concentrations in these fermentations?

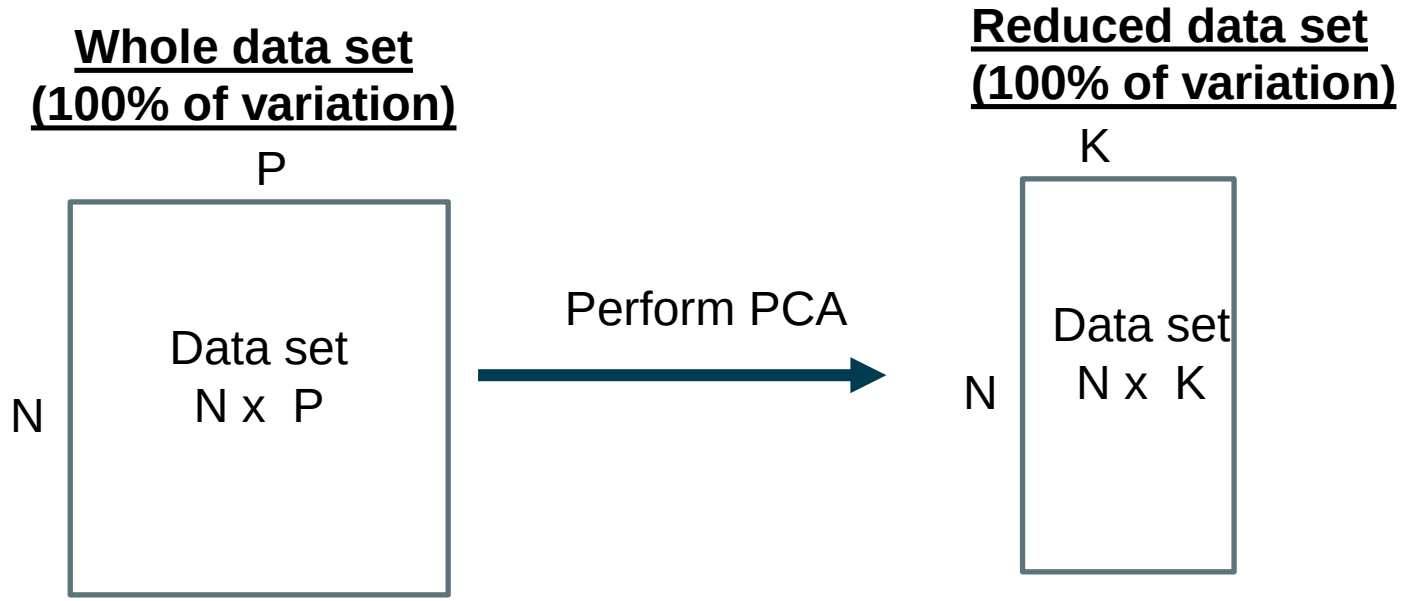
Do we need MVDA to analyse this data set?



Analysing one factor at a time is inefficient and can lead to misleading conclusions

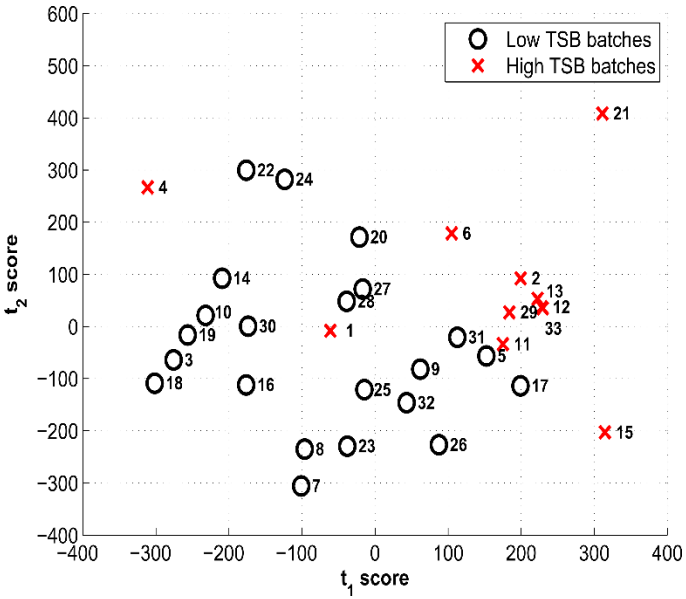
Complex data set analysed through Principal Component Analysis (PCA)

- PCA is technique that is suitable to analysis complex data sets by reducing the dimensionality of the data set
  - Essentially summarising the main sources of variability through newly generated vectors (components)

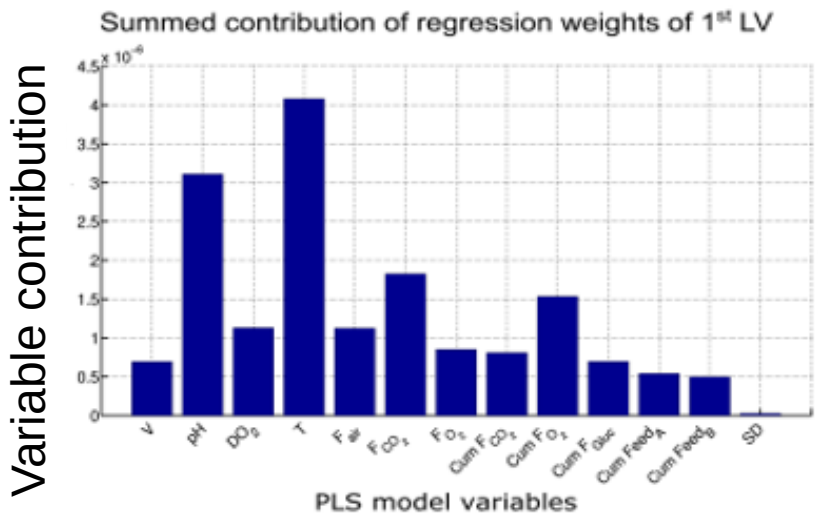


Principal Component Analysis of TSB problem

1st and 2nd principal components



Summary of main variables contributing to 1st principle component

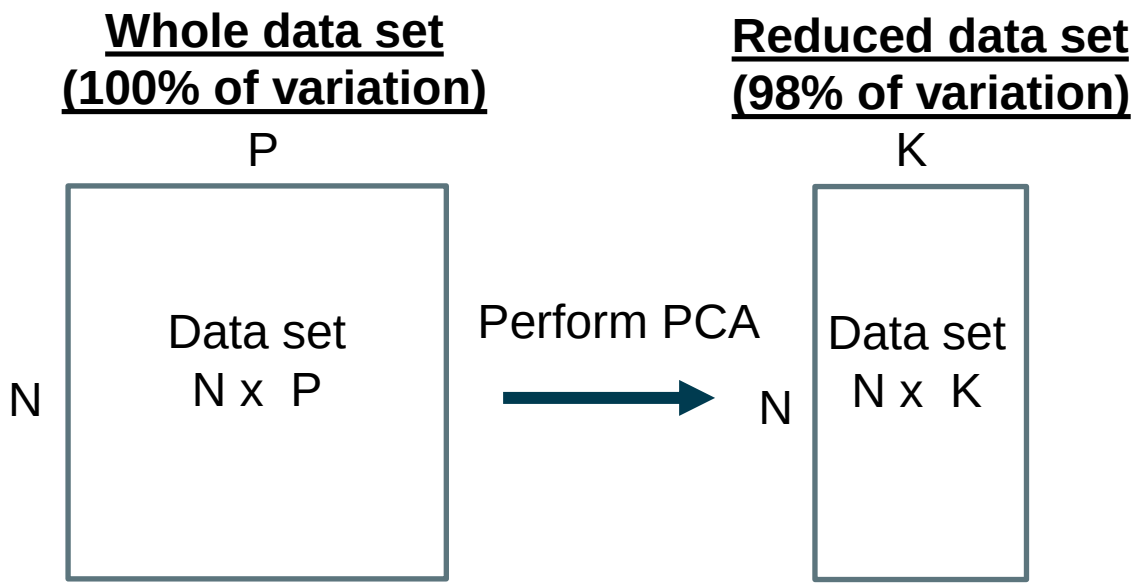


<u>Why is MVDA important?</u>

All on-line, off-line variables and initial conditions summarised into two graphs, allowing for easy interpretation of this complex data set

MVDA – Principal Component Analysis

- Principal component analysis (PCA) is the workhorse of the majority of MVDA techniques
- PCA is technique that is suitable to analysis complex data sets by reducing the dimensionality of the data set and enabling patterns to be identified
  - Essentially summarising the main sources of variability through newly generated vectors (called components)



Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set.