COMP20008 Elements of Data Processing

Assessing Correlations

Announcements

- Exam study guide
  - Available on the “Exam” section of the LMS. Outlines what needs to be known for the exam, from the lectures and workshops so far

- Consultation session with Donia for python programming
  - Running on Monday 3 April: 11am-12pm Rm 10.22 Doug McDonell Building (10th floor)

Plan today

- Discuss about finding correlations between pairs of features in a dataset
  - Why useful and important
  - Pitfalls
  - Case study: genetic data
- Review methods for computing correlation
  - Euclidean distance
  - Pearson correlation
- Next week
  - Mutual information (another method to compute correlation)

What is Correlation?

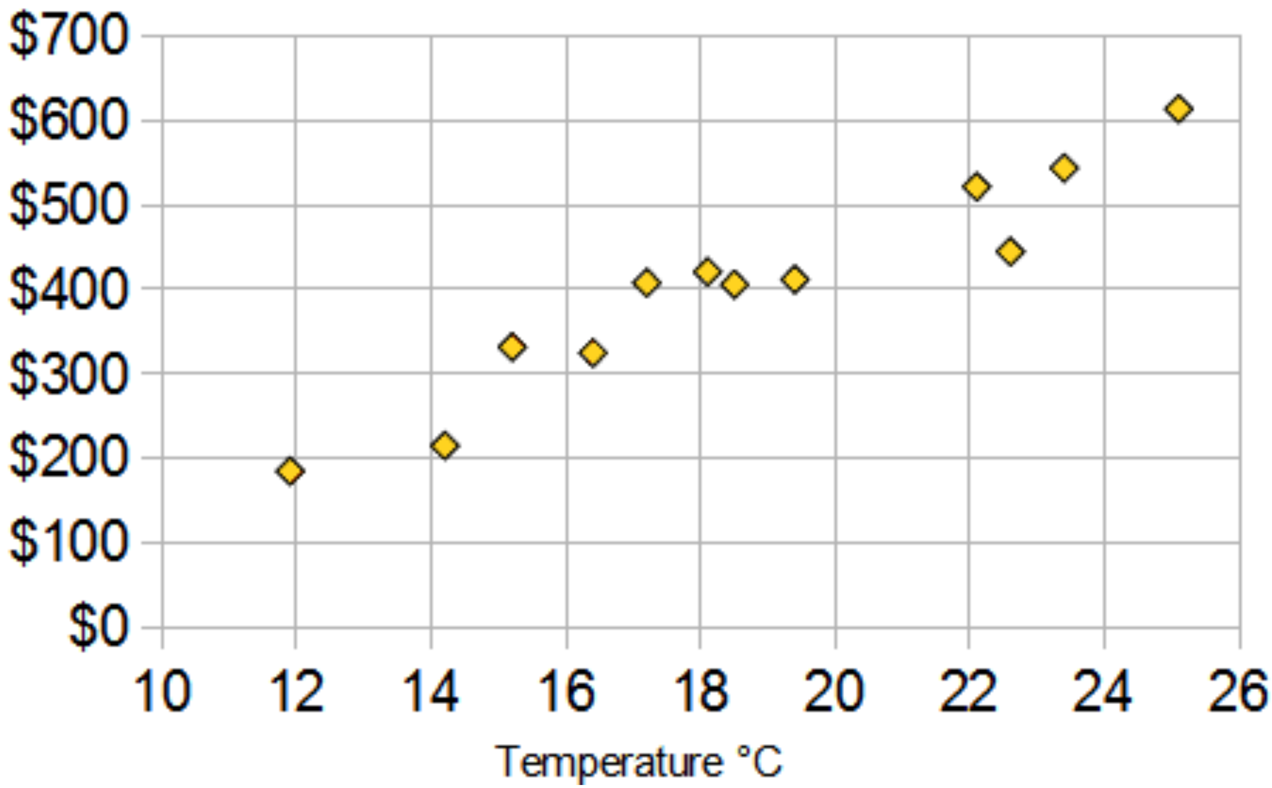
- Correlation is used to detect pairs of variables that might have some relationship

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

https://www.mathsisfun.com/data/correlation.html

What is Correlation?

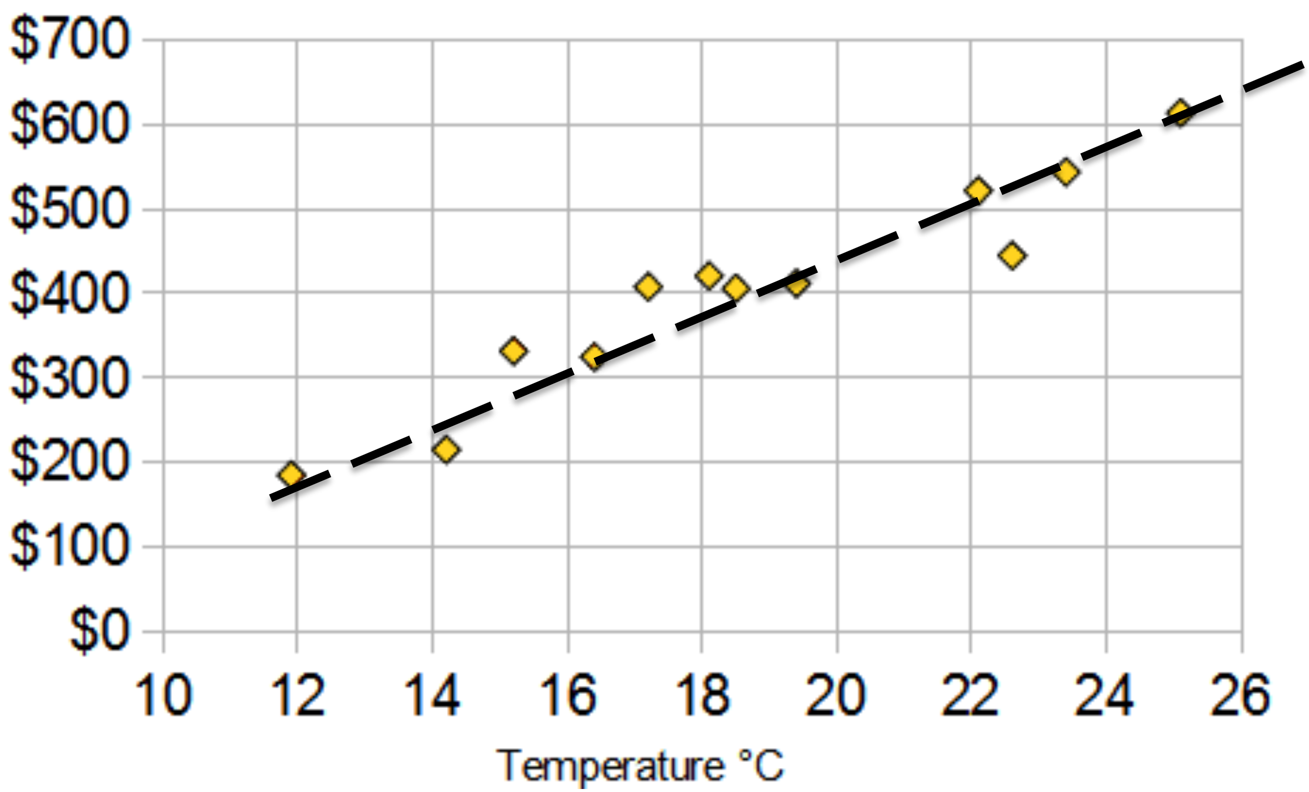
- Visually can be identified via inspecting scatter plots



https://www.mathsisfun.com/data/correlation.html

What is Correlation?

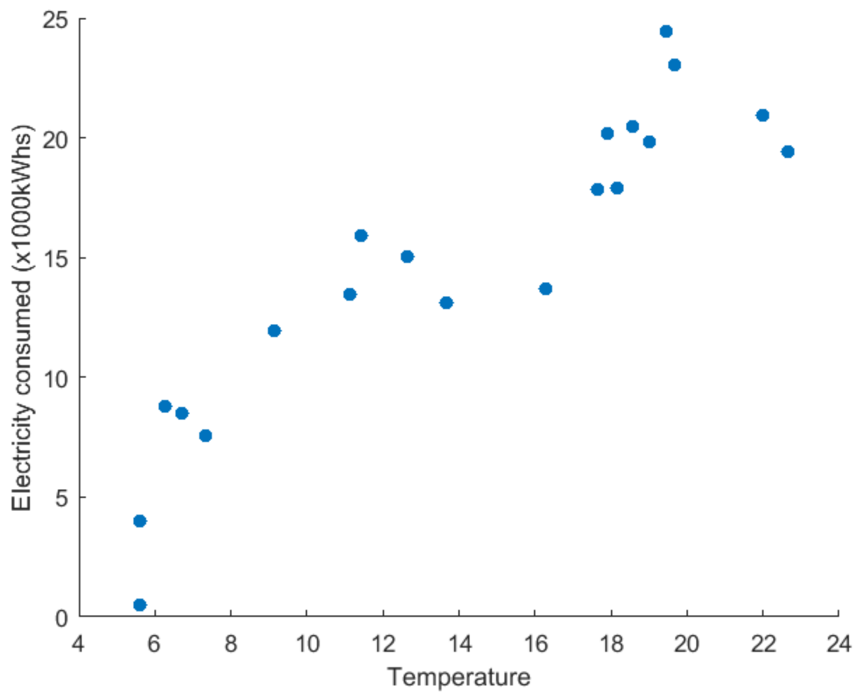
- Linear relations



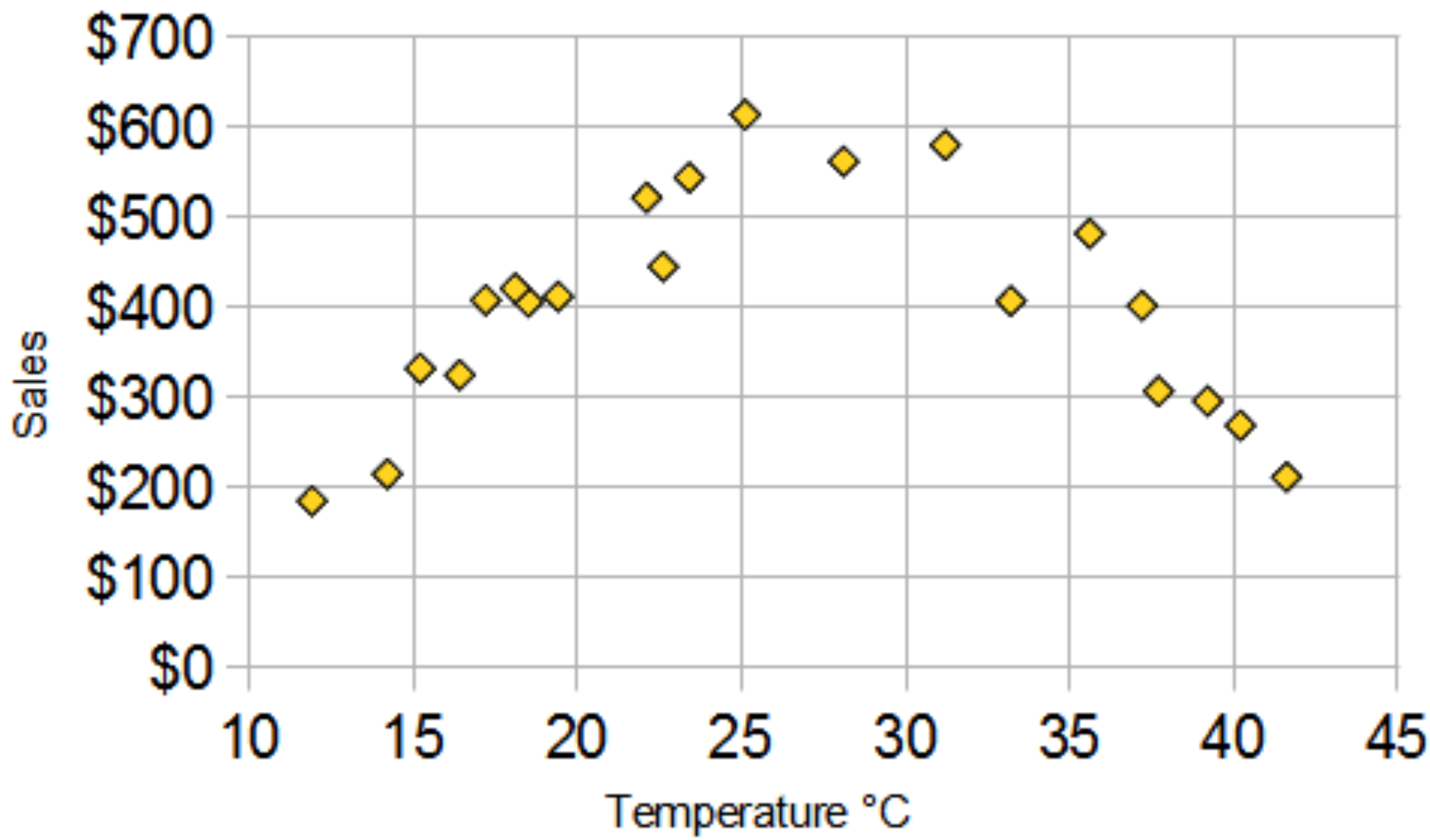
https://www.mathsisfun.com/data/correlation.html

Example of Correlated Variables

- Can hint at potential causal relationships (change in one variable is the result of change in the other)
- Business decision based on correlation: increase electricity production when temperature increases



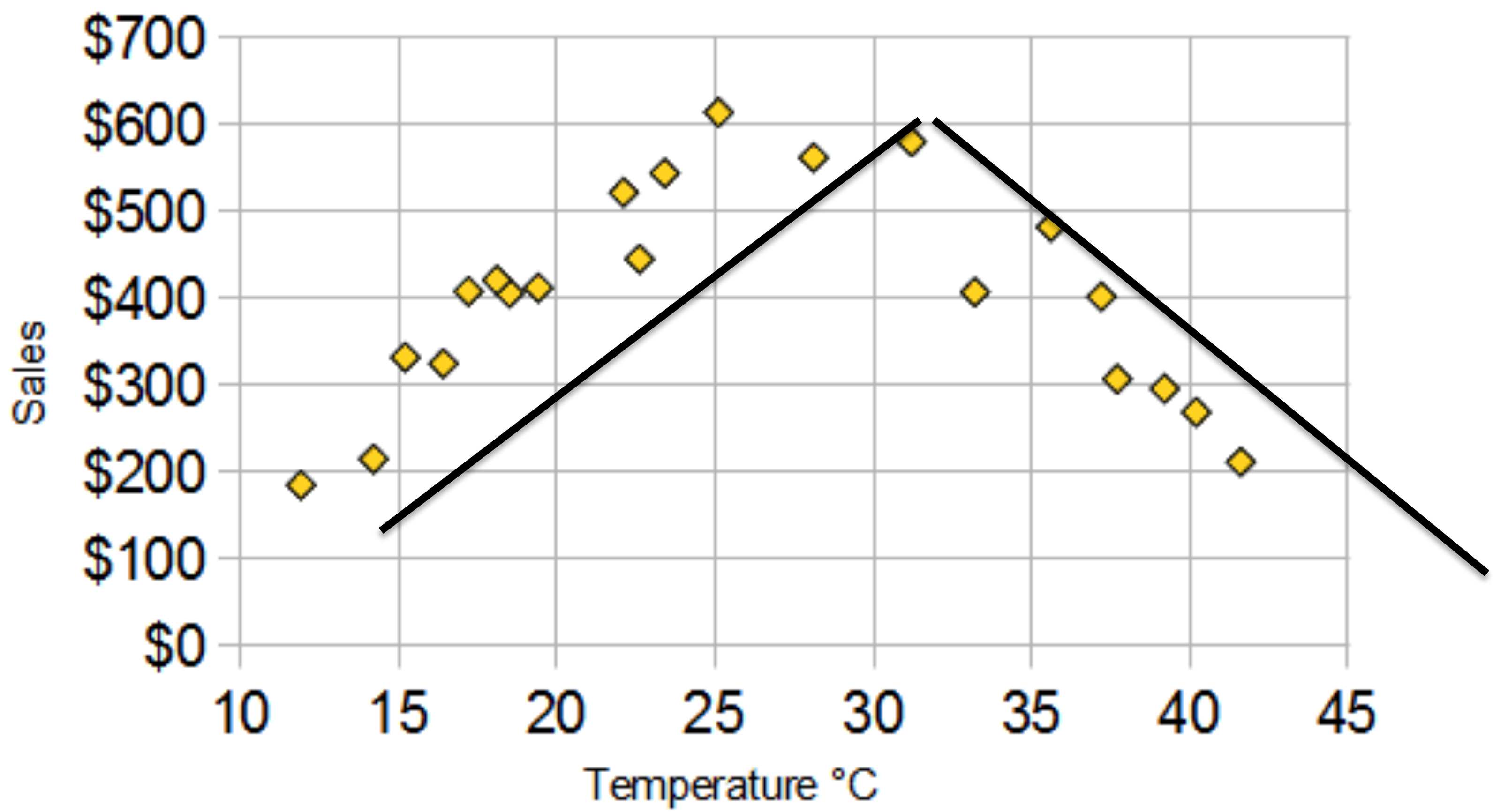
Example of non-linear correlation



It gets so hot that people aren't going near the shop, and **sales start dropping**

https://www.mathsisfun.com/data/correlation.html

Example of non-linear correlation



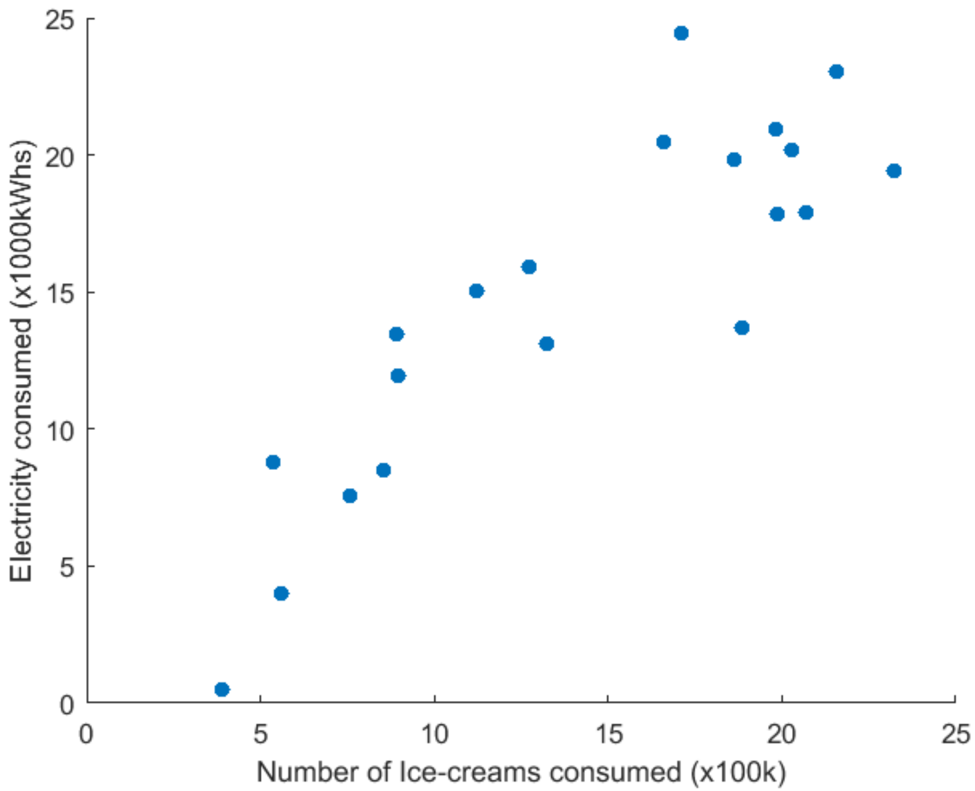
It gets so hot that people aren't going near the shop, and **sales start dropping**

https://www.mathsisfun.com/data/correlation.html

THE UNIVERSITY OF MELBOURNE

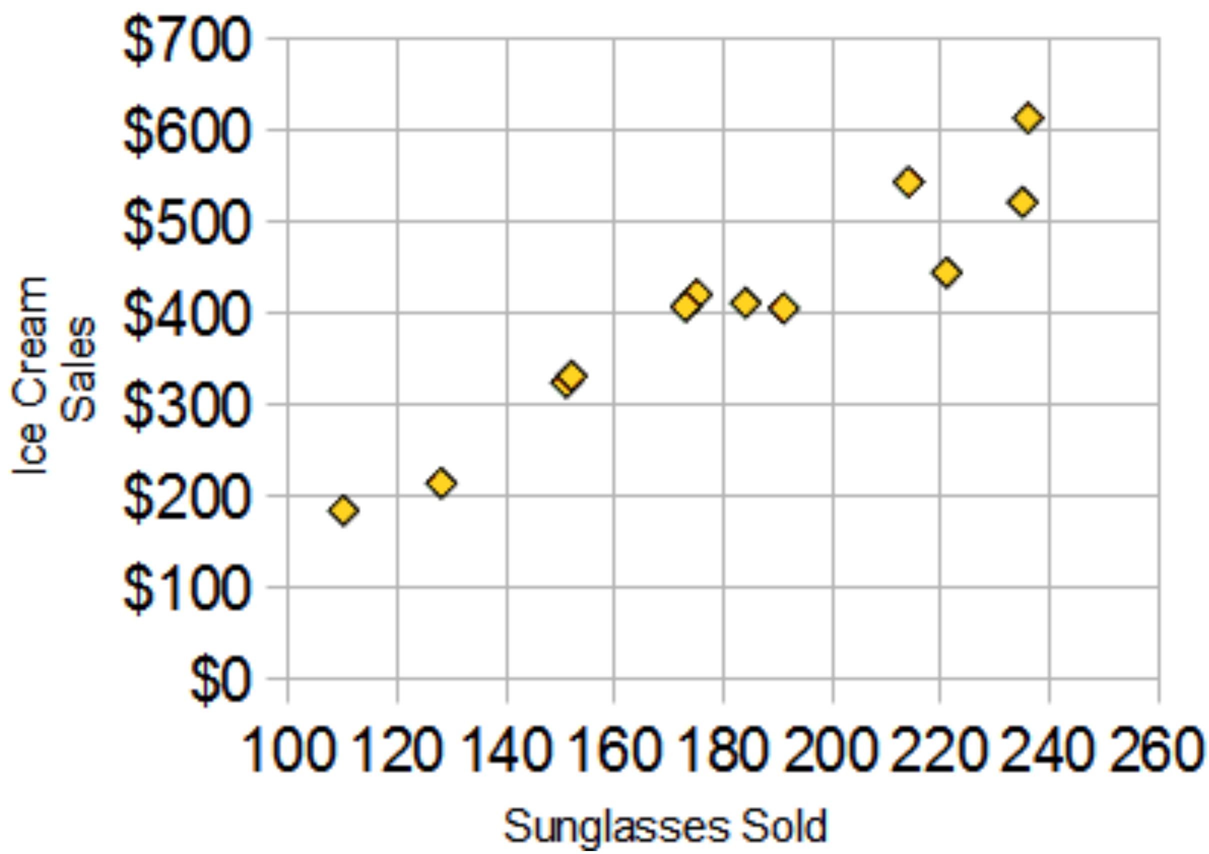
Example of Correlated Variables

- Correlation does not necessarily imply causality!



Example of Correlated Variables

- Correlation does not necessarily imply causality!

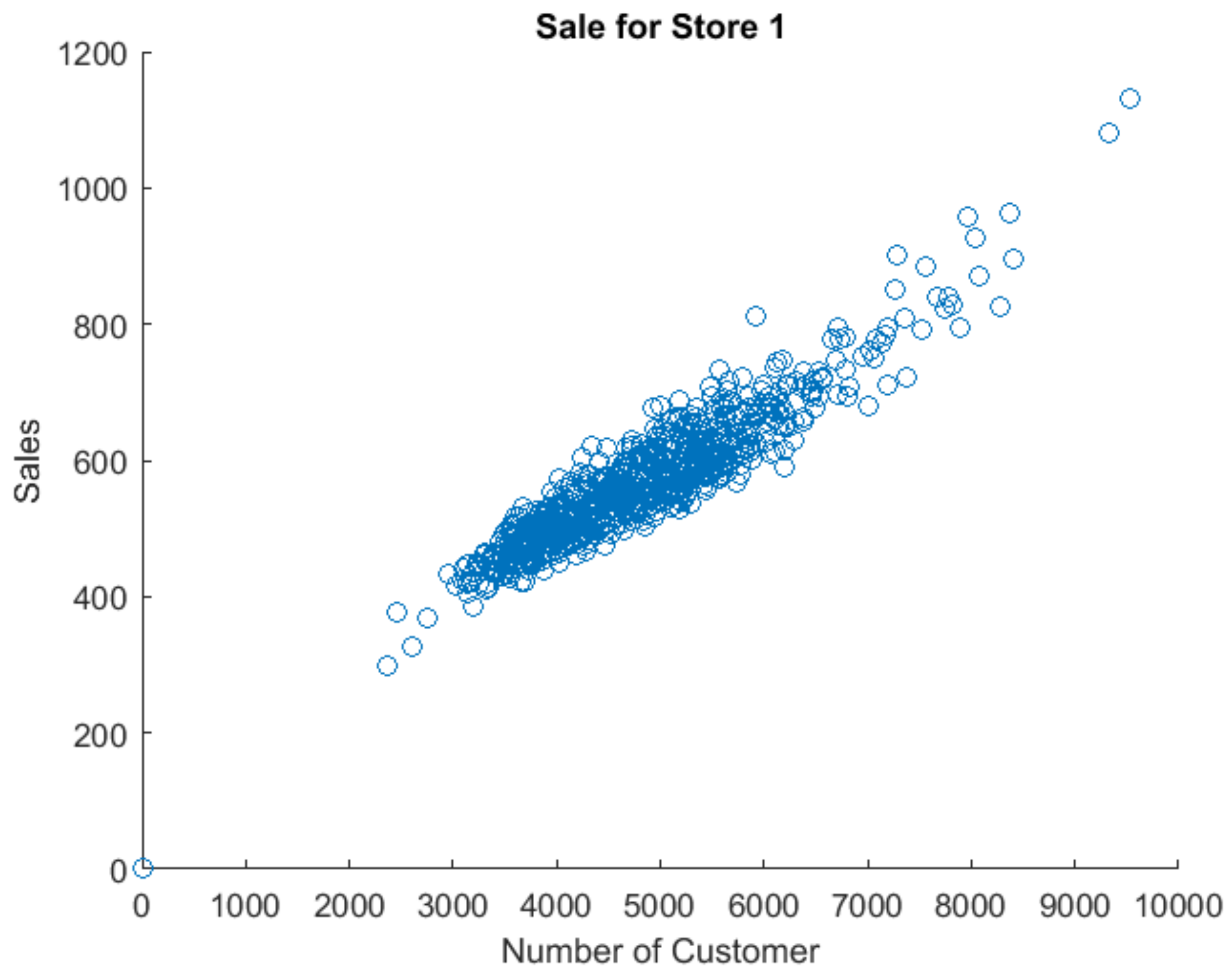


Example: Predicting Sales

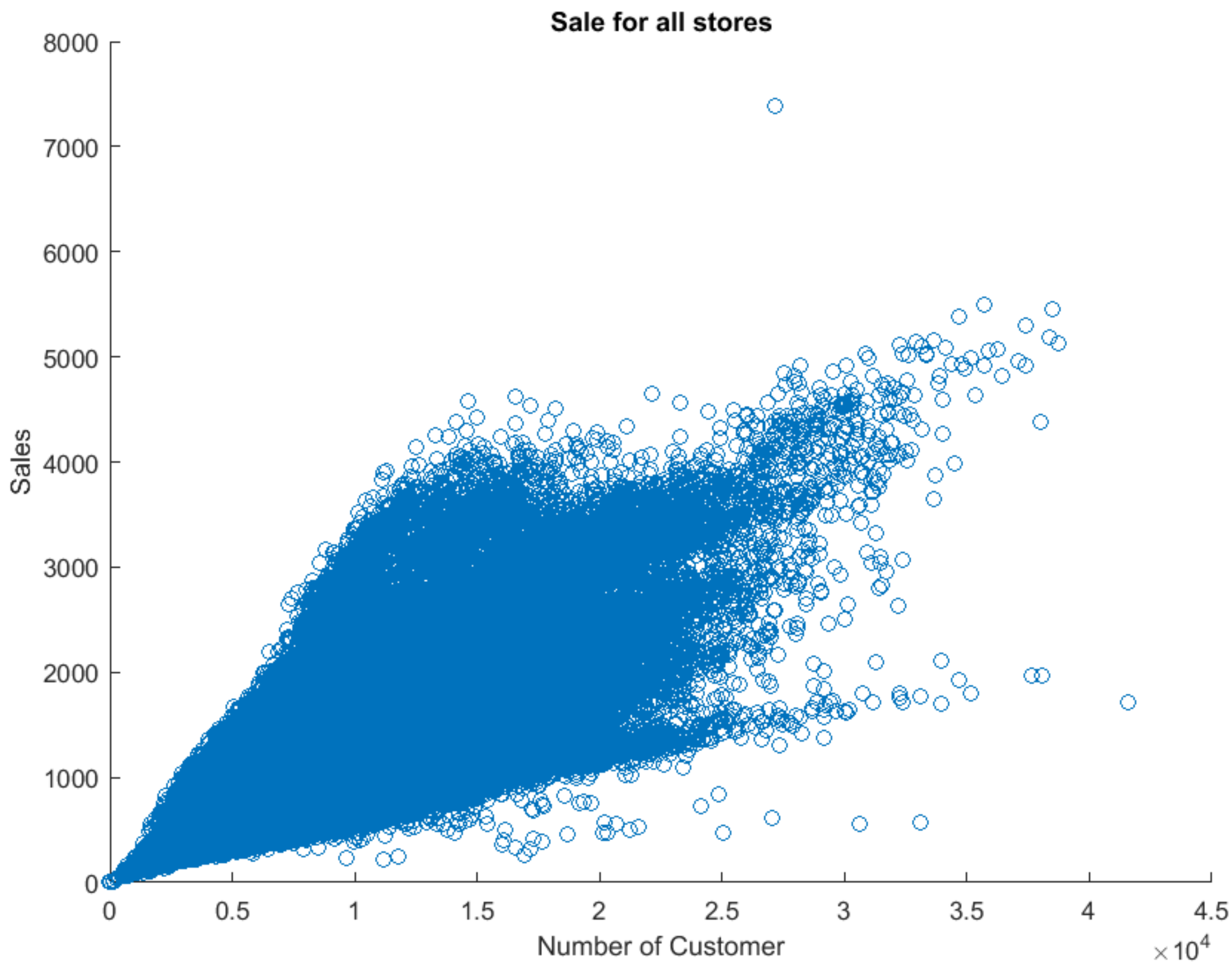
- https://www.kaggle.com/c/rossmann-store-sales/data

1	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHolida
2	1	5	31/07/2015	5263	555	1	1	0	1
3	2	5	31/07/2015	6064	625	1	1	0	1
4	3	5	31/07/2015	8314	821	1	1	0	1
5	4	5	31/07/2015	13995	1498	1	1	0	1
6	5	5	31/07/2015	4822	559	1	1	0	1
7	6	5	31/07/2015	5651	589	1	1	0	1
8	7	5	31/07/2015	15344	1414	1	1	0	1
9	8	5	31/07/2015	8492	833	1	1	0	1
10	9	5	31/07/2015	8565	687	1	1	0	1
11	10	5	31/07/2015	7185	681	1	1	0	1
12	11	5	31/07/2015	10457	1236	1	1	0	1
13	12	5	31/07/2015	8959	962	1	1	0	1
14	13	5	31/07/2015	8821	568	1	1	0	0
15	14	5	31/07/2015	6544	710	1	1	0	1
16	15	5	31/07/2015	9191	766	1	1	0	1
17	16	5	31/07/2015	10231	979	1	1	0	1
18	17	5	31/07/2015	8430	946	1	1	0	1
19	18	5	31/07/2015	10071	936	1	1	0	1
20	19	5	31/07/2015	8234	718	1	1	0	1
21	20	5	31/07/2015	9593	974	1	1	0	0
22	21	5	31/07/2015	9515	682	1	1	0	1
23	22	5	31/07/2015	6566	633	1	1	0	0
24	23	5	31/07/2015	7273	560	1	1	0	1
25	24	5	31/07/2015	14190	1082	1	1	0	1
26	25	5	31/07/2015	14180	1586	1	1	0	1

Example: Predicting Sales



Example: Predicting Sales



Example: Predicting Sales

- Other correlations
  - Sales vs. holiday
  - Sales vs. day of the week
  - Sales vs. distance to competitors
  - Sales vs. average income in area

Example rank correlation

- *"If a university has a higher-ranked football team, then is it likely to have a higher-ranked basketball team?"*

Football ranking	University team
1	Melbourne
2	Monash
3	Sydney
4	New South Wales
5	Adelaide
6	Perth

Basketball ranking	University team
1	Sydney
2	Melbourne
3	Monash
4	New South Wales
5	Perth
6	Adelaide

Why is correlation important?

- Discover relationships
- One step towards discovering causality

A causes B

Examples:

Gene A causes lung cancer

- Feature ranking: select the best features for building better machine learning models

THE UNIVERSITY OF MELBOURNE

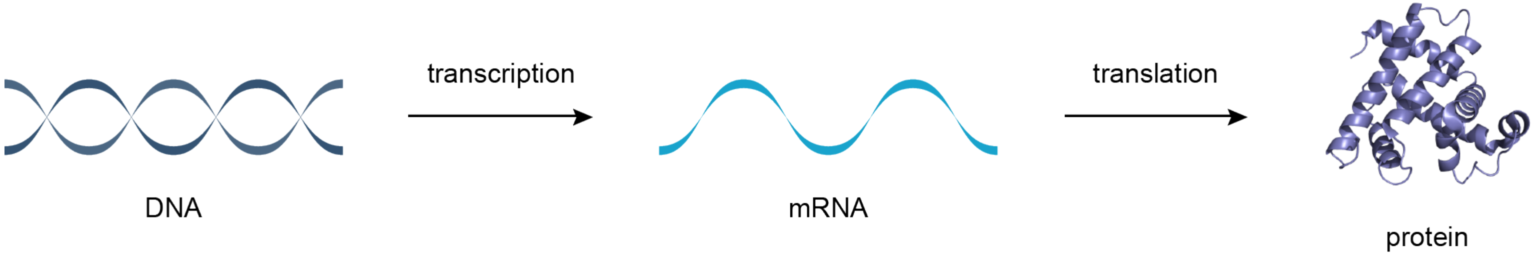
Case study: Microarrays data

- DNA Microarrays (Gene Chips)
- Measure genes' level of activity

https://en.wikipedia.org/wiki/Bio-MEMS

The Central Dogma of Molecular Biology

- DNA makes RNA makes proteins



- DNA contains multiple genes containing information to produce different types of proteins
- To much or too little proteins of certain type can cause diseases
- Gene chips can measure the amount or mRNA (a buffer for protein level) – activity level (expression level)

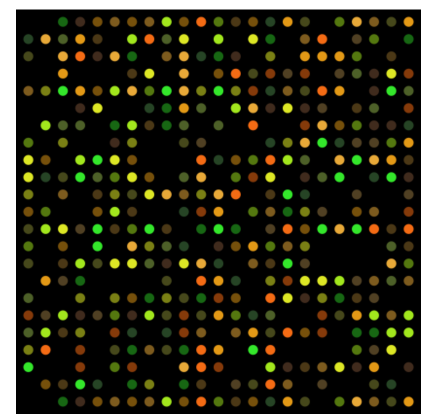
http://www.atdbio.com/content/14/Transcription-Translation-and-Replication

Microarray data

- Each chip contains thousands of tiny probes corresponding to the genes

(20k - 30k genes in humans)

	Gene 1	Gene 2	…	Gene 20K
Activity level	0.3	1.2	…	3.1



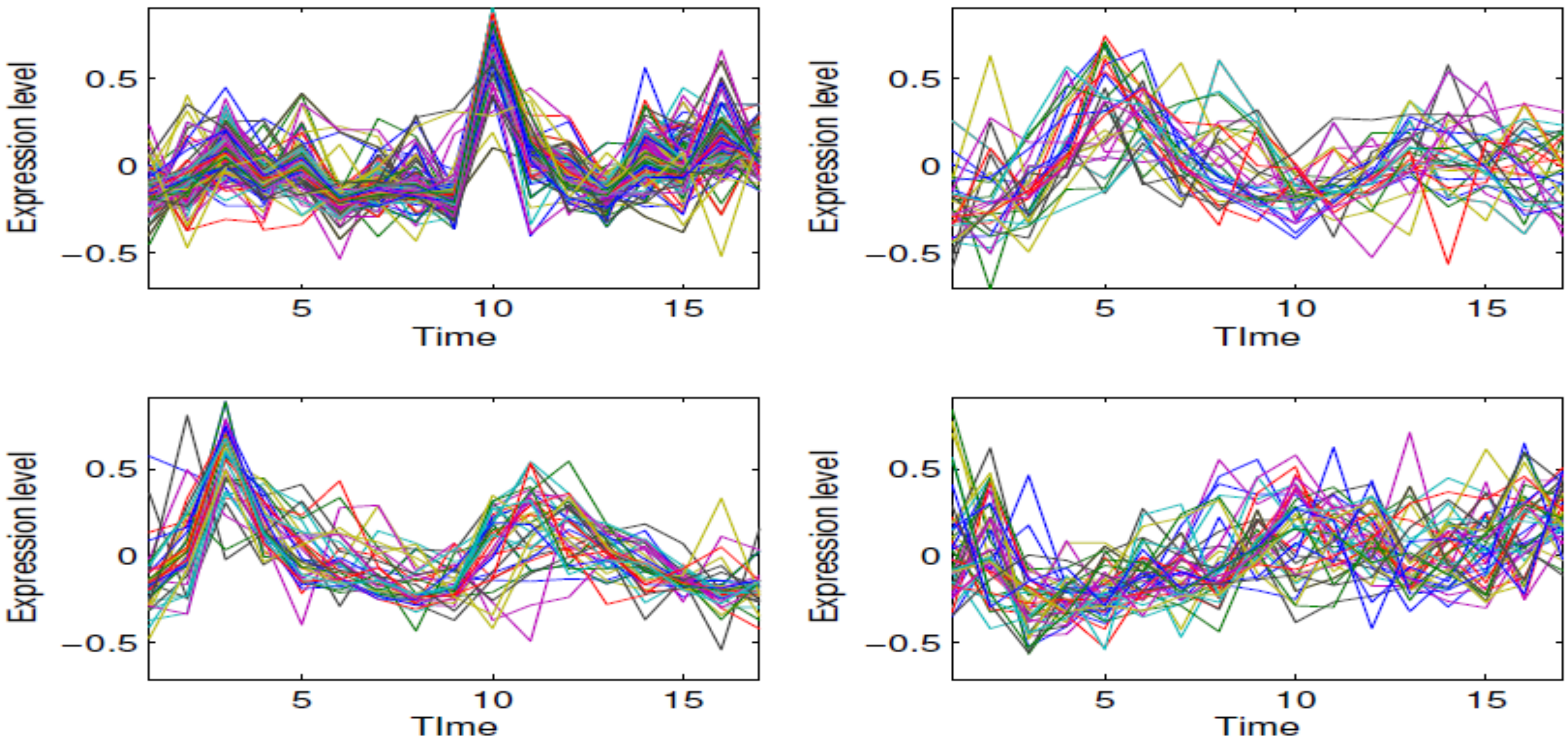
Microarray data from Multiple Conditions

	Gene 1	Gene 2	Gene 3	…	Gene n
Condition 1	2.3	1.1	0.3	…	2.1
Condition 2	3.2	0.2	1.2	…	1.1
Condition 3	1.9	3.8	2.7	…	0.2
…	…	…	…	…	…
Condition m	2.8	3.1	2.5	…	3.4

- Conditions:
  - different time points, same person, or
  - different people
- How correlation can help?

Correlation analysis on Microarray data

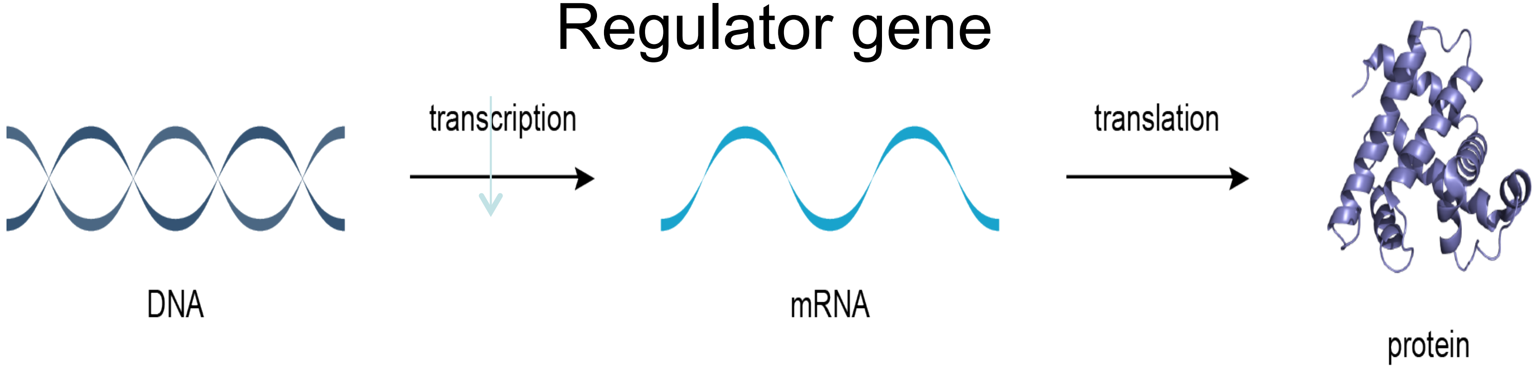
- Can reveal genes that exhibit similar patterns ➪ similar or related functions ➪ Discover functions of unknown genes



Build genetic networks

- Genes do not act in isolation: they control each other or work together

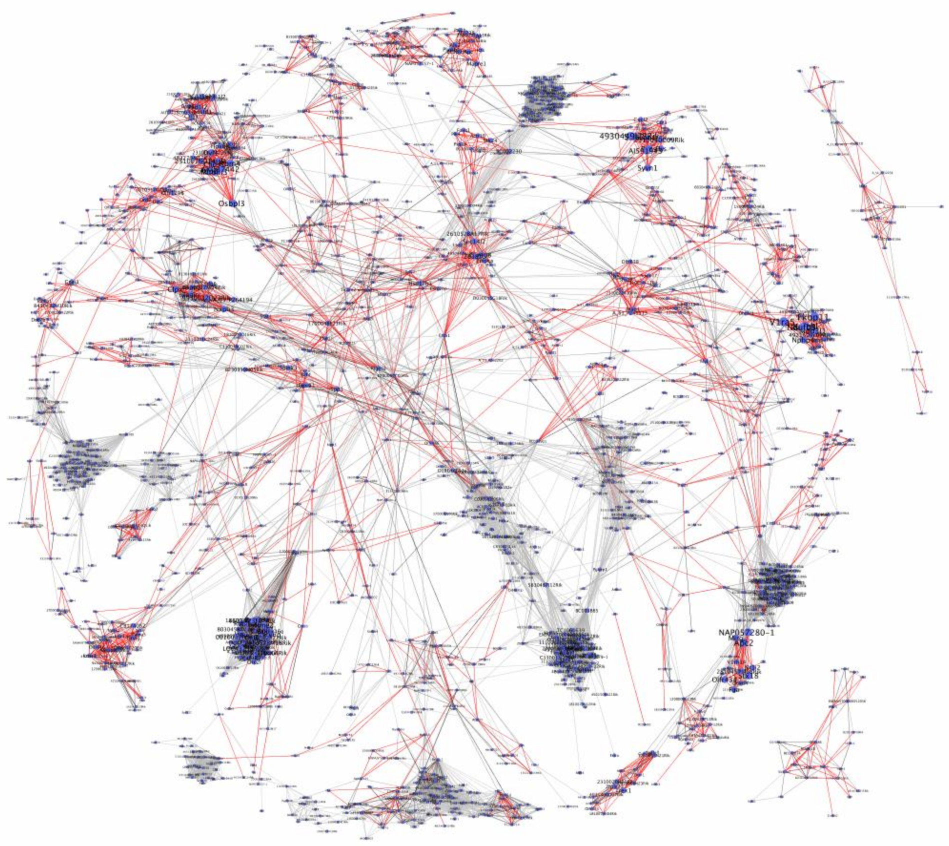
Regulator gene



- Gene A controls the activity level of Gene B: causality relationship

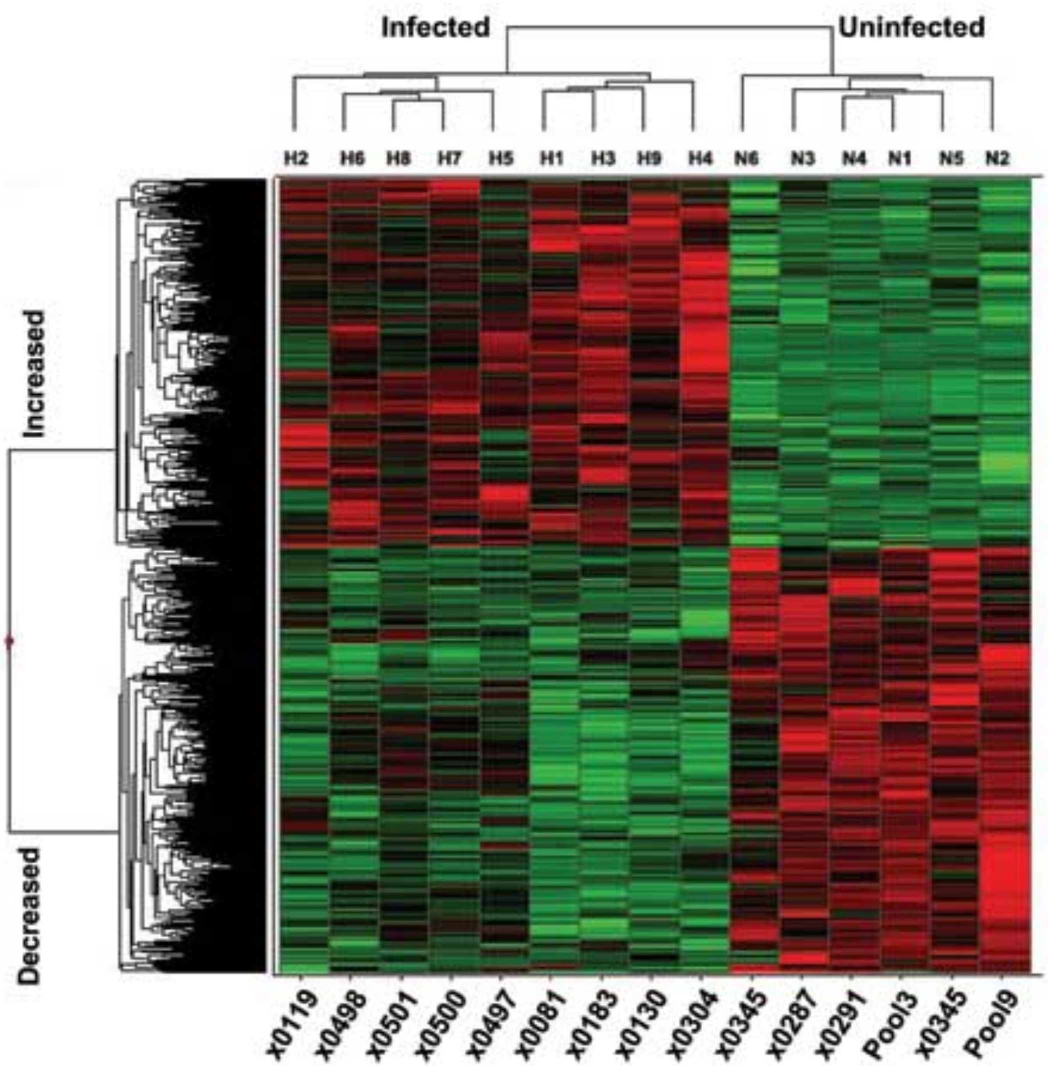
Genetic network

- Connect genes with high correlation



http://www.john.ranola.org/?page_id=116

Discover genes that are relevant to a disease



http://compbio.pbworks.com/w/page/16252903/Microarray

Correlation and Feature Ranking

- Why is correlation important?
  – Discover relationships, causality
  – Select the best features for building better machine learning models
- Measure of correlations:
  – Euclidean distance
  – Pearson coefficient
  – Mutual Information
- Case study: Microarrays data
  – What is it? How to collect?
  – How to build genetic networks from correlation
  – Which genes cause skin cancer?

Notation

	Gene 1	Gene 2	Gene 3	…	Gene n
Person 1	2.3	1.1	0.3	…	2.1
Person 2	3.2	0.2	1.2	…	1.1
Person 3	1.9	3.8	2.7	…	0.2
…	…	…	…	…	…
Person m	2.8	3.1	2.5	…	3.4

- Data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$
- Each object can be either a row or a column
- Each row $\mathbf{x}_i \in \mathbb{R}^n$ represents a person
- Each column $\mathbf{x}^j \in \mathbb{R}^m$ represents a gene

Euclidean distance

- Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
- **Euclidean distance**

$$
\begin{aligned}
d(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \\
&= \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\end{aligned}
$$

Length of the line segment connecting $\mathbf{x}$ and $\mathbf{y}$

- **Squared Euclidean distance**

$$
\begin{aligned}
d^2(\mathbf{x}, \mathbf{y}) &= (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 \\
&= ||x - y||^2 = 2 - 2\langle x, y \rangle
\end{aligned}
$$

L2 norm: $||\mathbf{x}|| = \sqrt{\sum_{i=1}^{n} x_i^2}$

Inner product: $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$

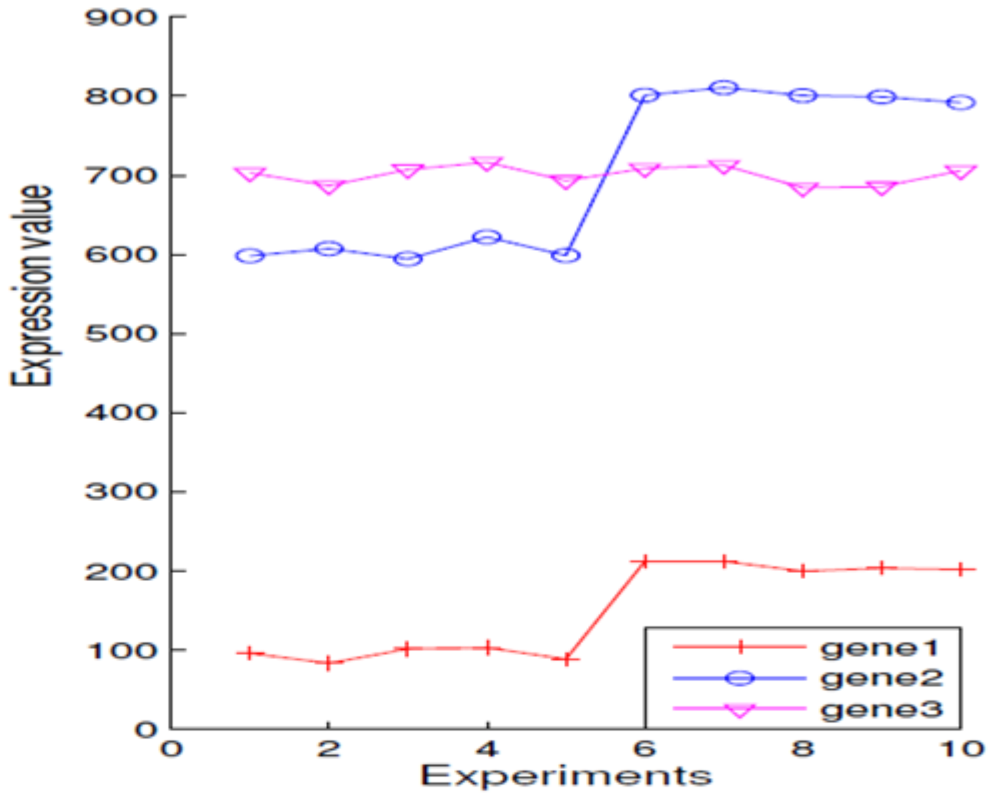
Problem of Euclidean distance

- Objects can be represented with different measure scales

	Day 1	Day 2	Day 3	…	Day m
Temperature	20	22	16	…	33
#Ice-creams	50223	55223	45098	…	78008
#Electricity	102034	105332	88900	…	154008

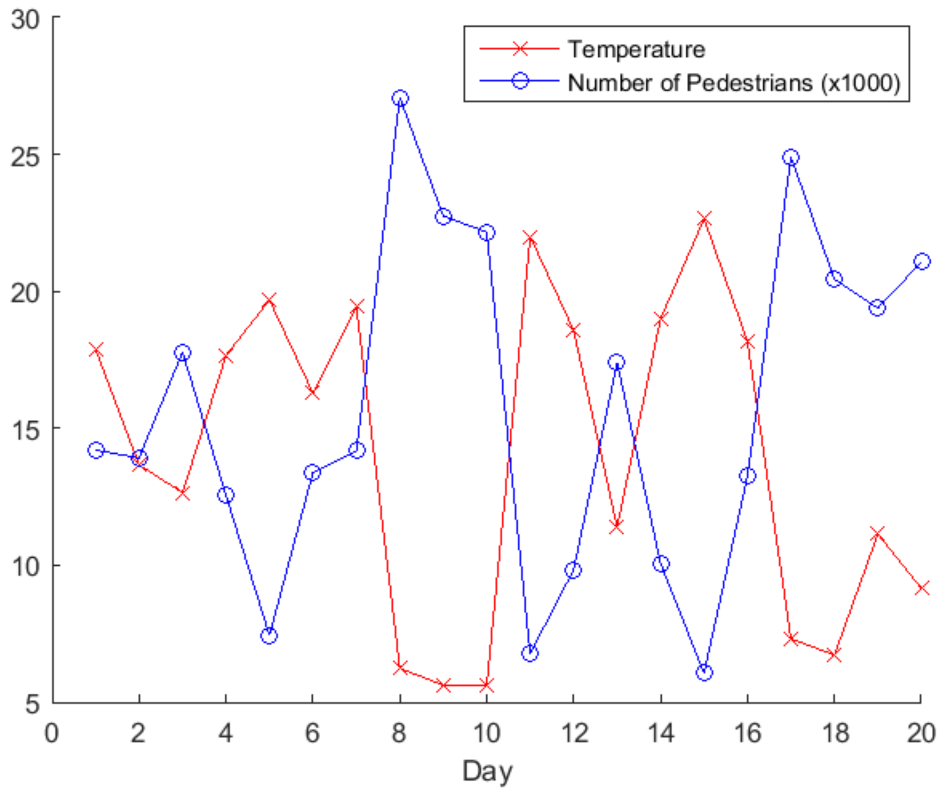
- Euclidean distance: does not give a clear intuition about how well variables are correlated

Problem of Euclidean distance



- Cannot discover variables with similar behaviours/dynamics but at different scale

Problem of Euclidean distance



- Cannot discover variables with similar behaviours/dynamics but in the opposite direction (negative correlation)

Pearson's correlation coefficient (r)

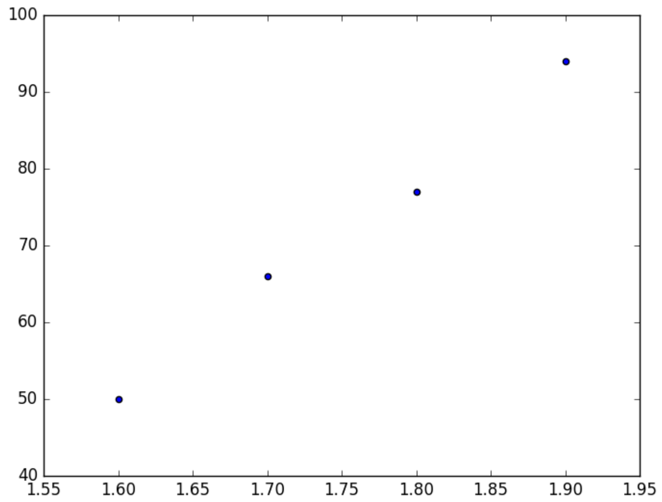
$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Sample means $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- Range within [-1,1]:
  - 1 for perfect positive linear correlation
  - -1 for perfect negative linear correlation
  - 0 means no correlation
  - Absolute value |r| indicates strength of linear correlation

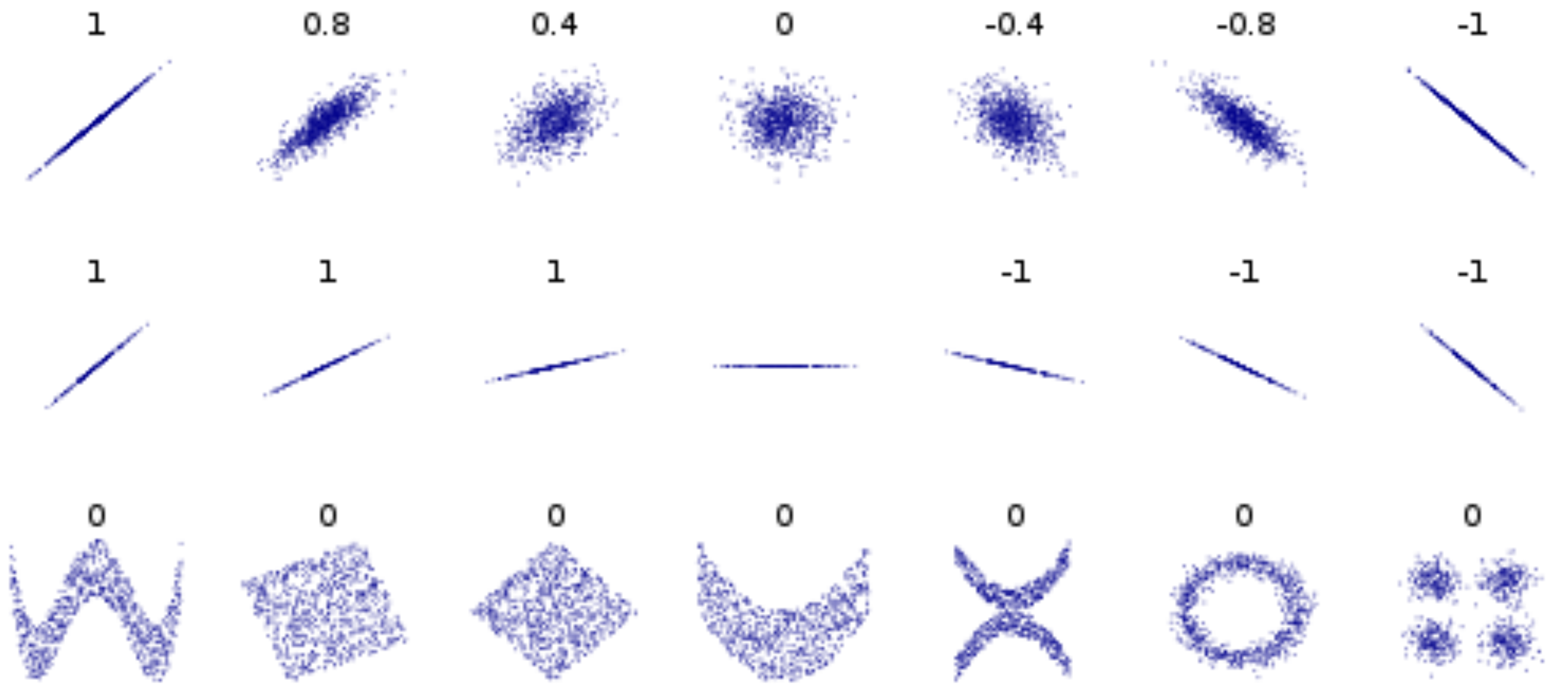
Pearson coefficient example

Height	Weight
1.6	50
1.7	66
1.8	77
1.9	94

$$x_1 = 1.6, x_2 = 1.7, x_3 = 1.8, x_4 = 1.9$$
$$y_1 = 50, y_2 = 66, y_3 = 77, y_4 = 94$$
$$\bar{x} = (1.6 + 1.7 + 1.8 + 1.9)/4$$
$$\bar{y} = (50 + 66 + 77 + 94)/4$$
$$r = 0.997$$



Examples



https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient

Interpreting Pearson correlation values

- In general it depends on your domain of application. Jacob Cohen has suggested
  - 0.5 is large
  - 0.3-0.5 is moderate
  - 0.1-0.3 is small
  - less than 0.1 is trivial

Properties of Pearson's correlation

- Range within [-1,1]
- Scale invariant: $r(x,y)= r(x, Ky)$
- Location invariant: $r(x,y)= r(x, K+y)$
- Can only detect linear relationships

$$y = a.x + b + \text{noise}$$

Cannot detect non-linear relationship

$$y = \sin(x) + \text{noise}$$

Notes

- Interactive correlation calculator
  - http://www.bc.edu/research/intasc/library/correlation.shtml
- Correlation <> Causality

http://tylervigen.com/spurious-correlations

- Google trend correlation

Points you should know from today

- be able to explain why identifying correlations is useful for data wrangling/analysis
- understand what is correlation between a pair of features
- understand how correlation can be identified using visualisation
- understand the concept of a linear relation, versus a non linear relation for a pair of features
- understand why the concept of correlation is important, where it is used and understand why correlation is not the same as causation
- understand the use of Euclidean distance for computing correlation between two features and its advantages/ disadvantages

Point to know – cont.

- understand the use of Pearson correlation coefficient for computing correlation between two features and its advantages/ disadvantages
- understand the meaning of the variables in the Pearson correlation coefficient formula and how they can be calculated. Be able to compute this coefficient on a simple pair of features. The formula for this coefficient will be provided on the exam.
- be able to interpret the meaning of a computed Pearson correlation coefficient
- understand the advantages and disadvantages of using the Pearson correlation coefficient for assessing the degree of relationship between two features