UWE Bristol | University of the West of England

Presentation by

Dr. Phil Legg

Associate Professor in Cyber Security

Date: Autumn 2019

Security Data Analytics and Visualisation

3: Data Analytics

Recap

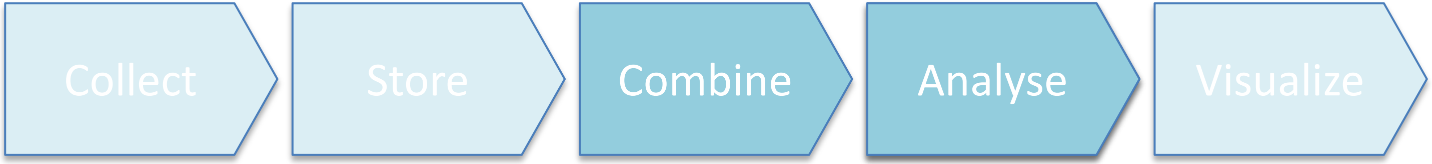
- What are the different stages of a data analytics pipeline?
- What are the different stages of a data science workflow?
  - Iterative processes between machine and user
  - "Human-in-the-Loop"
    - How do we make best use of the machine **and** the best use of the human?

- Using the Jupyter notebook environment and trying out a sample data exploration

- **Worksheet 1 available on Blackboard for lab sessions**

Data Analytics

Data analytics pipeline



Collect

Store

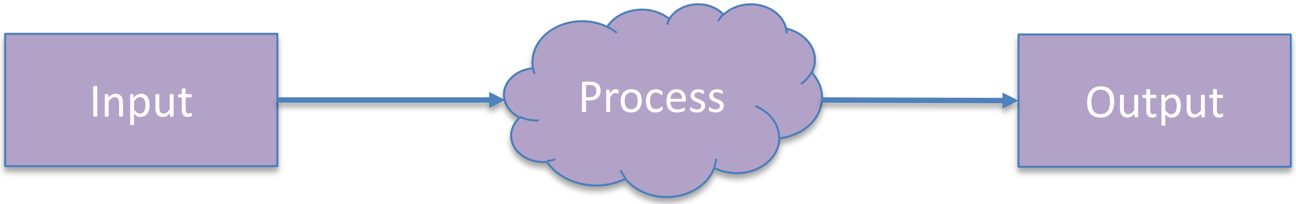
Combine

Analyse

Visualize

Think about:

- What is our input data? **How can we express our data as features?**
- What is the process / transform that we need to perform?
- What is our desired output?



Input

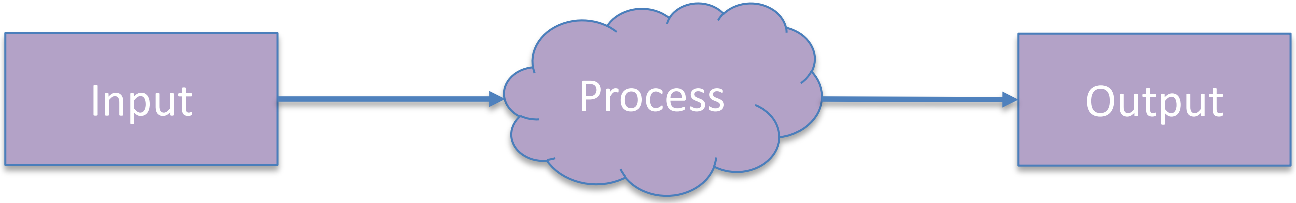
Process

Output

What are we trying to learn?

- Understand attributes of the data?
- Understand the trend of the data?

Supervised Machine learning is about learning how the machine can map input to output without explicitly programming or stating how this should be achieved by the process



Input

Process

Output

From Raw Data to Features

- What may we want to learn?

- How do we take raw data and curate features from this?
  - RAM → RAM usage per Hour?
  - E-mails → E-mails sent per Hour? #of new recipients per Day? #of words that match against a dictionary set?
  - PCAP → #unique_dest_IPs/Hour?, #unique_ports/Hour, #unique_src_dst_occurrences?
  - *Think about how to express raw data as a value over time to compare against…*
- Descriptive Statistics
- Ability to predict future observations - forecasting
- Structure of the data - can learn to separate normal from abnormal

Where to find data?

- Data science is now wide-spread due to availability of data online.
- Kaggle, VAST Challenge, UC Irvine Machine Learning repo.
- Web scraping for datasets
- Internet-enabled data sensors

- Many of the activities we collect data about have existed before:
  - Technology is just making it easier (possible) to obtain and record data about such events

Cleaning Data

Data is messy

- Web scraping – data is messy, we need to parse this to get it clean
- System may expect data in particular format
  - Example – how should we represent a timestamp?
- How to ensure data is accurate when scraping?
- What routines are available to help us tidy things up before we store them?

Storing Data

Save my data

- Data can be saved in many ways – write it to a TXT file?
- How can we impose some structure on our data?
    - CSV, TSV (Comma/Tab separated values)
    - JSON (JavaScript Object Notation)
- These store data to files – however, to search them requires loading the file into memory.
    - What if our data is GBs large?
- Database offers a accessible and scalable storage approach
    - SQL (Structured Query Language)
    - MongoDB (Unstructured Document Database)
    - Hadoop (Distributed File System using Map-Reduce)

Pandas in Python

- Pandas is a powerful data analytics framework in Python, that supports many of the benefits of working with a database.
- For this course (and for your coursework), we will keep to using CSV files, and pandas for our data storage.
- Of course – other projects in the future may require you to think about the possible alternatives such as Hadoop, SQL, and MongoDB.

Machine Learning in Data Analytics

Machine Learning

- What can machine learning be used for?
  - Classification
  - Quantitative prediction
  - Inference
  - Exploration and discovery

Data Driven Security (p226-227)

Types of Machine Learning

Supervised Learning

- Instances of the data are labelled
- Given an input, learn what the mapping process should be to reach the appropriate output (e.g., classification)

Unsupervised Learning

- Instances of the data are not labelled
- Given an input, learn characteristics of the data to then identify similarities and dis-similarities (e.g., clustering)

Semi-Supervised Learning / Active Learning

- Human-centric forms of learning
- Semi-supervised may have a training set of a single known class (e.g., known normals)
- Active Learning may have very small training set, with the machine querying a human for labels that offer greatest benefit

4 Types of Anomalies

Point Anomaly

- a single point in a Time Series is anomalous compared to the rest of the data

Contextual Anomaly

- a data instance in a Time Series which is considered anomalous because of the context (for instance, low temperatures in summer in the northern hemisphere). In order to detect such anomalies, we need to have information on the context itself.

Subsequence Anomaly

- these are collective anomalies which are anomalous with regards to the rest of the data *even though* data points from the subsequence might not be considered anomalous

Time Series Anomaly

- Time Series which are anomalous with respect to a set (database) of anomalies

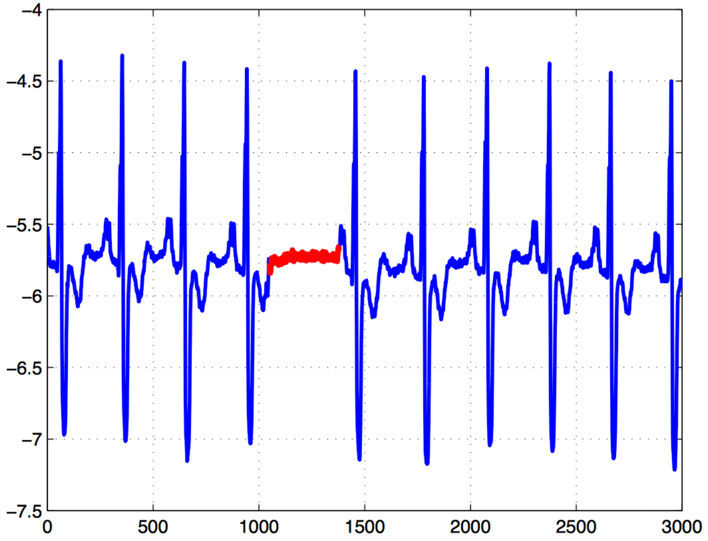
Anomaly Detection of Time Series, by Deepthi Cheboli, University of Minnesota, 2010.
Anomaly Detection: a survey, by Chandola and al. ACM Computing Surveys, 2009.
http://www.datasciencecentral.com/profiles/blogs/anomaly-detection-for-the-oxford-data-science-for-iot-course

Security is typically concerned with anomalies

If something looks ‘different’ to what is expected or what is deemed to be normal, then it could indicate some form of security breach.

‘Novelty’, ‘Outlier’, ‘Trend’, ‘Debunking’, ‘Forecasting’

Not all anomalies are ‘outliers’



We can see the red region is ‘different’ – yet the individual values here are not ‘outliers’.

Understanding context of the data, and the expected trend is crucial

Anomaly Detection of Time Series, by Deepthi Cheboli, University of Minnesota, 2010.
Anomaly Detection: a survey, by Chandola and al. ACM Computing Surveys, 2009.
http://www.datasciencecentral.com/profiles/blogs/anomaly-detection-for-the-oxford-data-science-for-iot-course

Statistical Measures

What is the **distribution** of our data?

How can we *describe* this distribution numerically?

- Maximum and Minimum (Range)
- Mean, Median, Mode (Mid-point)
- Standard deviation, Interquartile range (Spread)
- Kurtosis (Skewness)
- Correlation Coefficient (Comparison)

Sometimes referred to as ‘descriptive’ statistics

Statistical Measures

How can we compare and analyse these 4 datasets?

Data 1		Data 2		Data 3		Data 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Statistical Measures

How can we compare and analyse these 4 datasets?

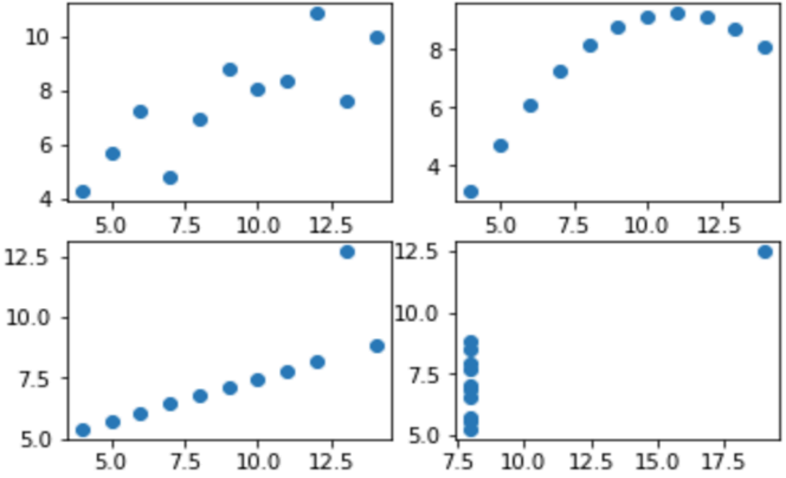
Data 1		Data 2		Data 3		Data 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Statistically, they are all equal!

Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	Y = 3.00 + 0.500 x

Statistical Measures

Data 1		Data 2		Data 3		Data 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
		14.0	8.10	14.0	8.84	8.0	7.04
		6.0	6.13	6.0	6.08	8.0	5.25
		4.0	3.10	4.0	5.39	19.0	12.50
		12.0	9.13	12.0	8.15	8.0	5.56
		7.0	7.26	7.0	6.42	8.0	7.91
		5.0	4.74	5.0	5.73	8.0	6.89



…Yet visually they are very different!
This is known as ***Anscombe's Quartet***.

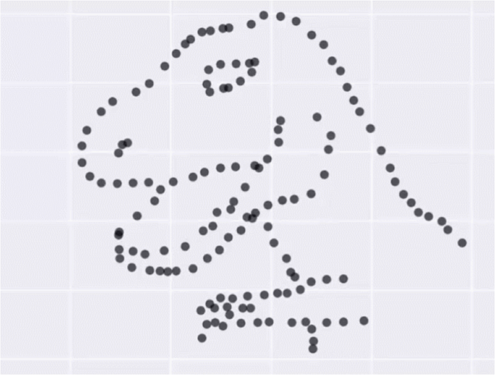
Statistical Measures

Statistical measures are important, and help to characterise properties of the data. However, they alone may not tell the whole story.

> *"...make **both** calculations **and** graphs. Both sorts of output should be studied; each will contribute to understanding."*
>
> *F. J. Anscombe, 1973*

*(A good argument for why to **visualise** your data!*

But more on that next week...)



Another great example is the **Datasaurus Dozen** – see here

Machine Learning Techniques

Linear Regression

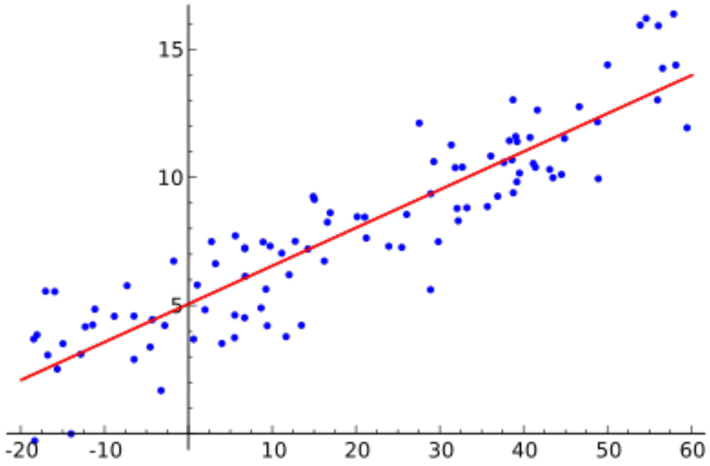
Learning a general representation of our data so that we can express as a function (e.g., line)

$$\mathbf{y} = m\,\mathbf{x} + c$$

If I observe a new x value, I can predict approximately what the corresponding y value would be.

e.g., If my system is running 20 processes (x), what would be the expected RAM usage for a uninfected system (y)?

Mean Squared Error (MSE) can be used to measure how close the line fits the points

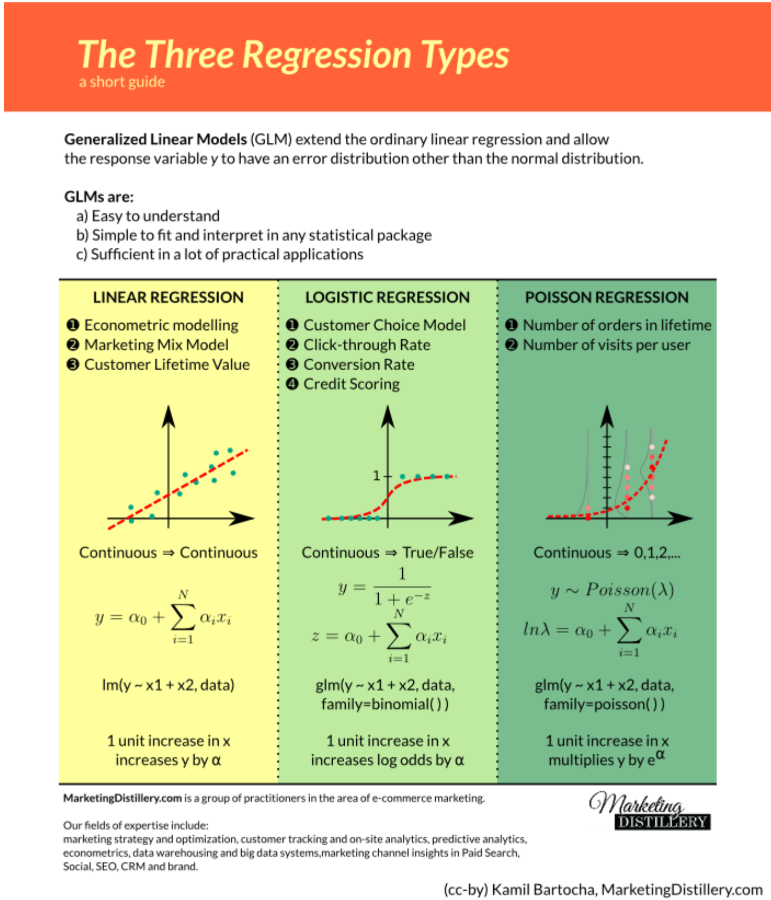


Regression Tools

Linear Regression gives a continuous value for ***y***

What if we just want a binary decision? (e.g., is this a concern or not?)

Rather than fitting a straight line, we can fit other functions (e.g., sigmoid function), that can be used for decision boundaries.



The Three Regression Types

a short guide

Generalized Linear Models (GLM) extend the ordinary linear regression and allow the response variable y to have an error distribution other than the normal distribution.

GLMs are:

a) Easy to understand
b) Simple to fit and interpret in any statistical package
c) Sufficient in a lot of practical applications

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
❶ Econometric modelling ❷ Marketing Mix Model ❸ Customer Lifetime Value	❶ Customer Choice Model ❷ Click-through Rate ❸ Conversion Rate ❹ Credit Scoring	❶ Number of orders in lifetime ❷ Number of visits per user
Continuous ⇒ Continuous	Continuous ⇒ True/False	Continuous ⇒ 0,1,2,...
$y = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$	$y = \frac{1}{1+e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$	$y \sim Poisson(\lambda)$ $ln\lambda = \alpha_0 + \sum_{i=1}^{N} \alpha_i x_i$
lm(y ~ x1 + x2, data)	glm(y ~ x1 + x2, data, family=binomial( ))	glm(y ~ x1 + x2, data, family=poisson( ))
1 unit increase in x increases y by α	1 unit increase in x increases log odds by α	1 unit increase in x multiplies y by e^{α}

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing.

Our fields of expertise include:
marketing strategy and optimization, customer tracking and on-site analytics, predictive analytics, econometrics, data warehousing and big data systems,marketing channel insights in Paid Search, Social, SEO, CRM and brand.

Marketing DISTILLERY

(cc-by) Kamil Bartocha, MarketingDistillery.com

Clustering

Unsupervised learning – identify clusters of similar instances

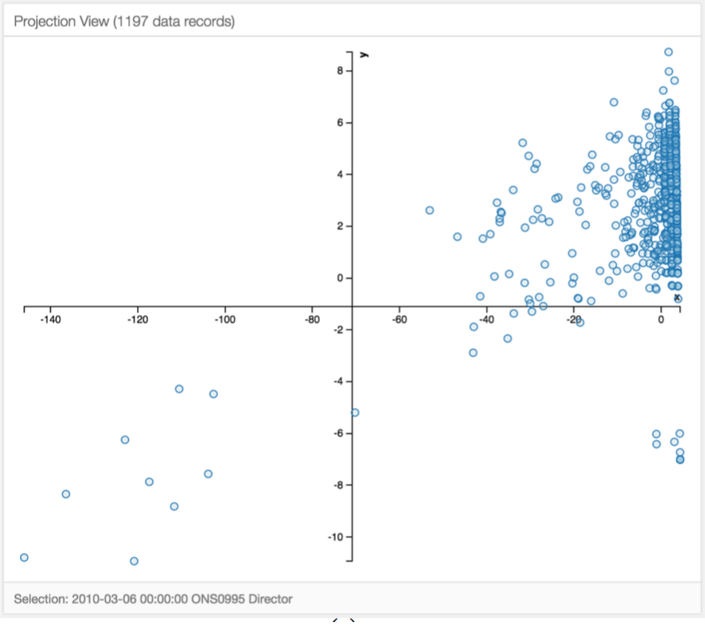
- e.g., k-Means Clustering

Data is typically in a high-dimensional space (i.e., we have many different attributes)

– how can we reduce this to something observable (i.e., 2D or 3D?)

Dimensionality Reduction

- Principle Component Analysis (PCA)
- T-Distributed Stochastic Neighbor Embedding (T-SNE)



Projection View (1197 data records)

Selection: 2010-03-06 00:00:00 ONS0995 Director

Neural Network

Popular technique that is suitable for learning how to map from inputs to outputs.

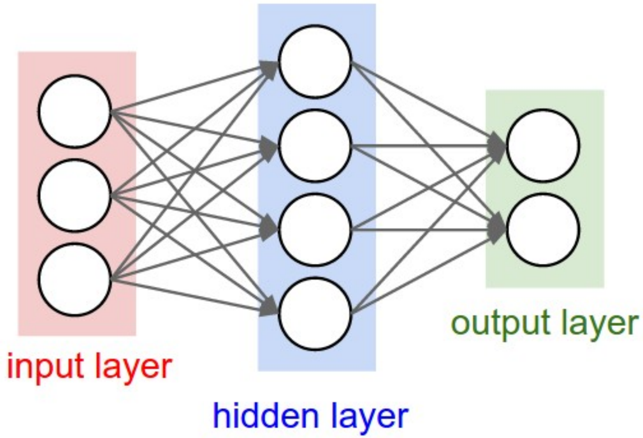
Network learns what should be at the hidden layer for transforming input to output

Image classification

- Input: Image pixels
- Output: Image type (e.g., person, cat, dog)

Anomaly detection

- Input: Activity features (e.g., counts)
- Output: Probability of anomaly



Learn by Example

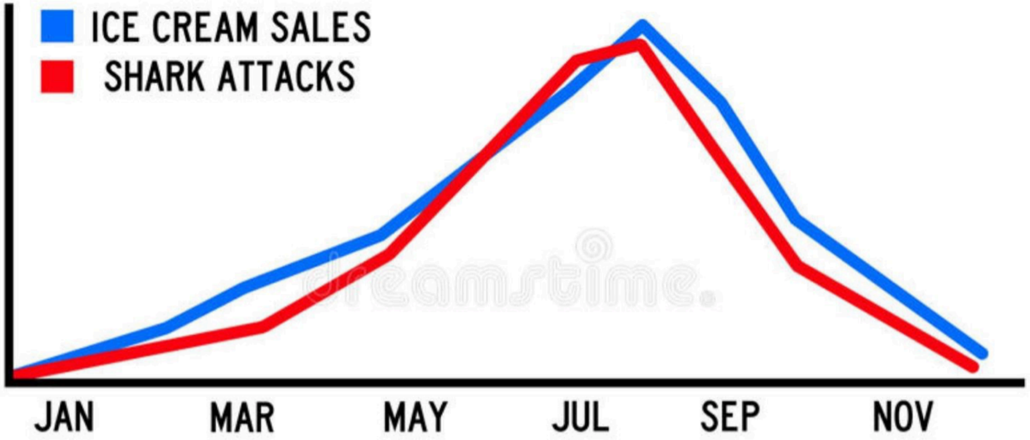
- For now, we are just introducing the concepts of regression, clustering, and neural networks

- We can take an input, and learn a process, to map our data to a desired output.
  - Output may be a group (clustering), may be a continuous value (regression), or maybe a more complex function that is either class or continuous (neural network)

- Next week we will continue to explore these methods and learn more of how they actually work
  - *Jupyter Notebook examples will be available*

Correlation != Causation

- Think about what you data is telling you
- *Does it really make sense?*



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

Takeaway

- Data analysis is complex – there is no one solution for analysing a problem
- Feature Engineering is a fundamental challenge – how to express raw data as a value over time to make useful comparisons with future observations?
- Statistics can give excellent descriptive information about the data – but these are not always unique for a set of data however
- Machine learning techniques can help to identify trends and characteristics that may be present - important to know what the appropriate tool to use is
- Anomalies are probably the most commonly used characteristic in security investigations - however not all anomalies are outliers, it depends what you are looking for!
    - *"… if you look hard enough you'll always find something – even if there's nothing there …"* [*https://towardsdatascience.com/the-hidden-risk-of-ai-and-big-data-3332d77dfa6*](https://towardsdatascience.com/the-hidden-risk-of-ai-and-big-data-3332d77dfa6)