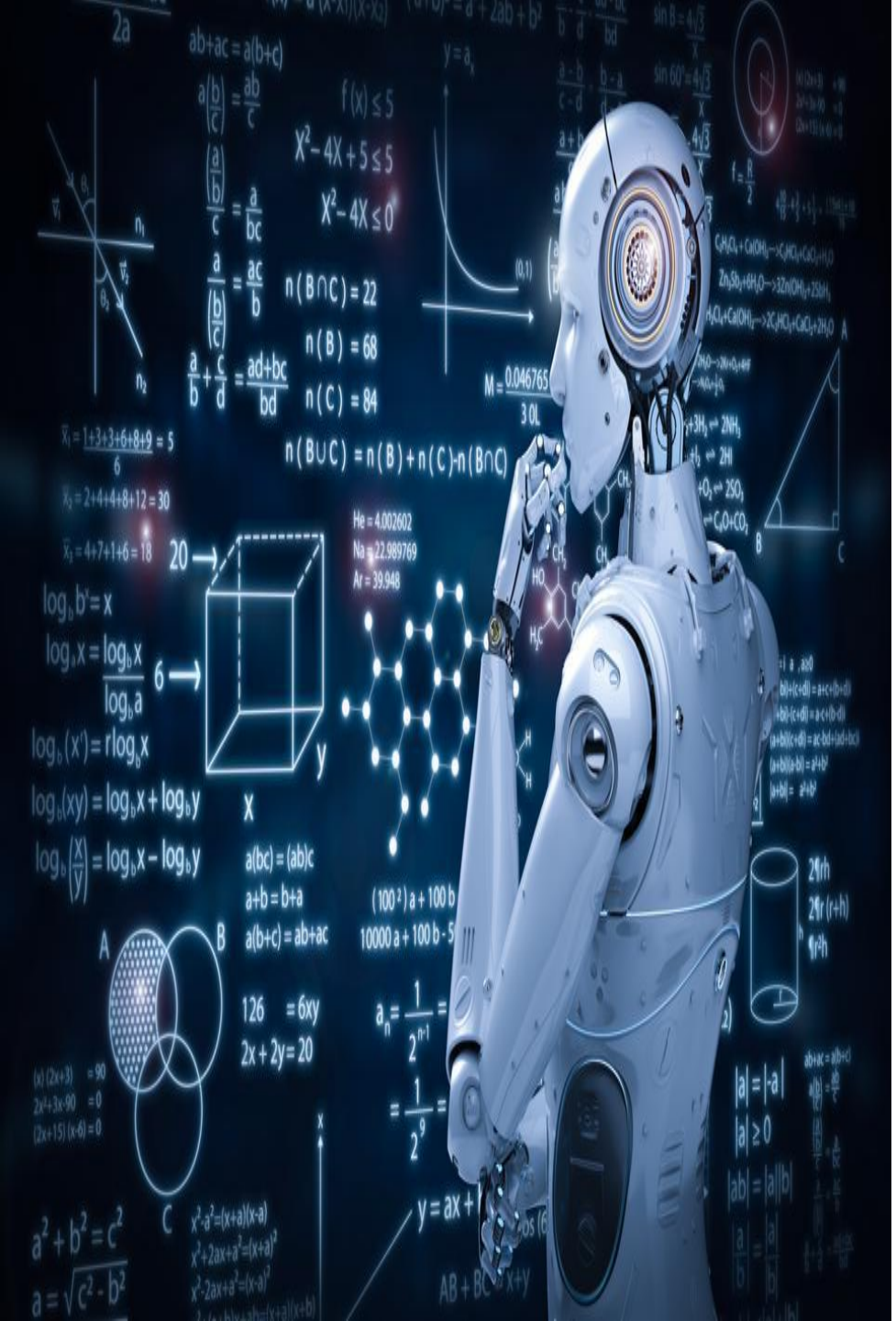


Practical Machine Learning

Day 4: Mar23 DBDA

Kiran Waghmare

Agenda

- Stages in Knowledge Extraction
- EDA
  - Summarize statistics
  - Measures
  - Visualization
- Modelling

KDD

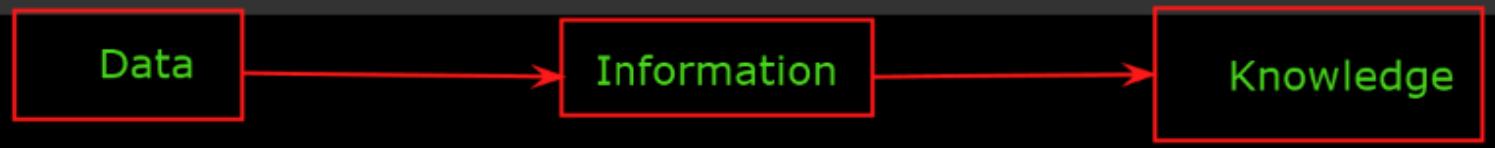
Date: 06/07/2023

Topics:

-Knowledge Extraction

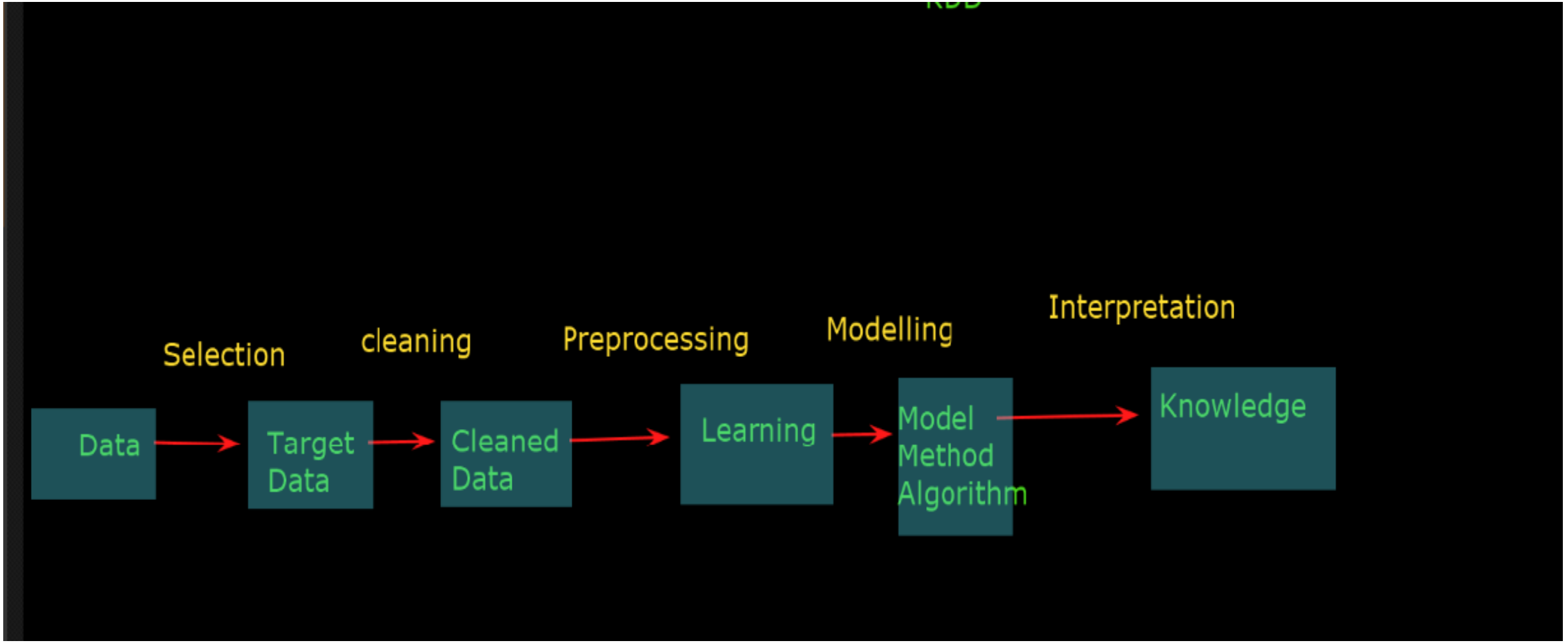
-EDA

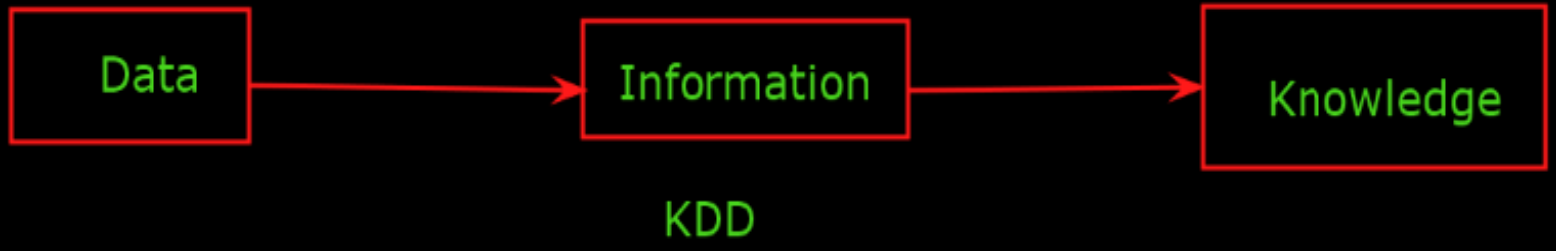
-Modelling



KDD

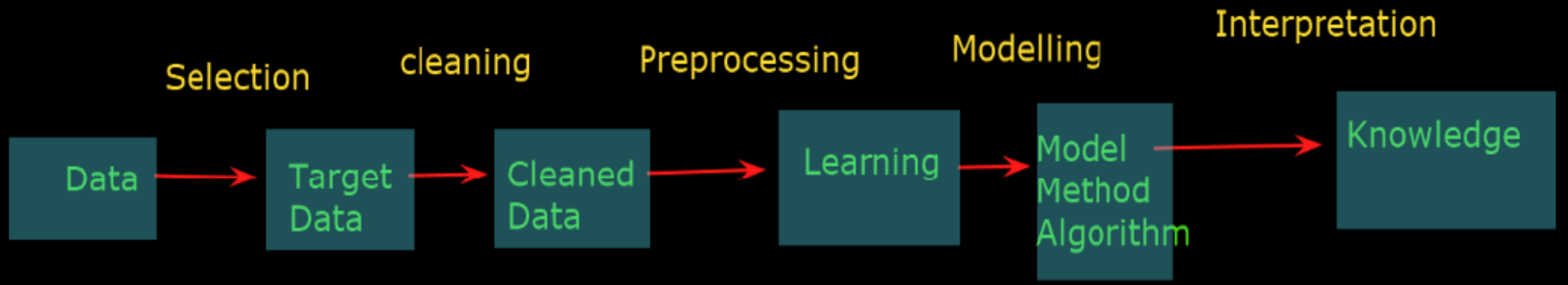
Steps in KDD Analysis



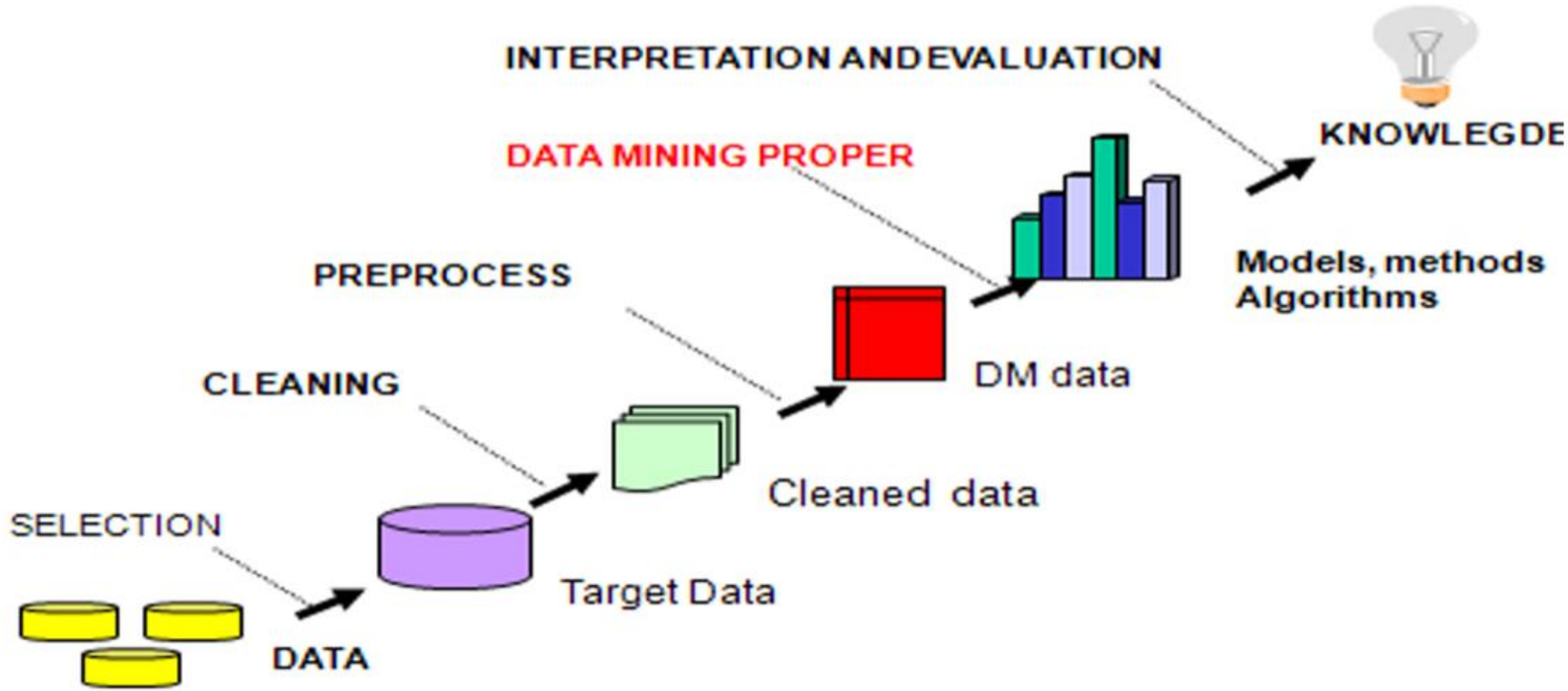


Challenges:

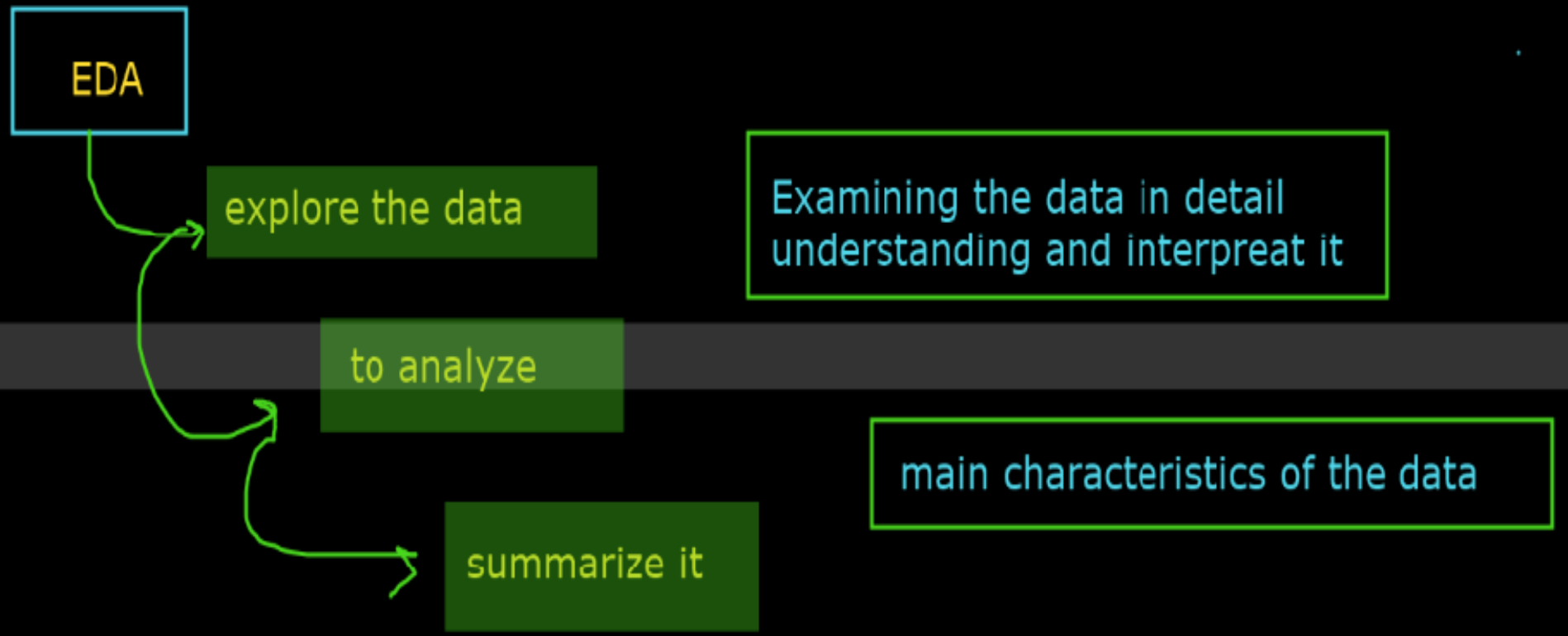
1. Large data
2. High dimensionality
3. Overfitting
4. Changibg data, missing data and noise
5. Domain knowledge
6. Understanding the patterns generated by model



DM- KDD process
(re-iterated if needed)

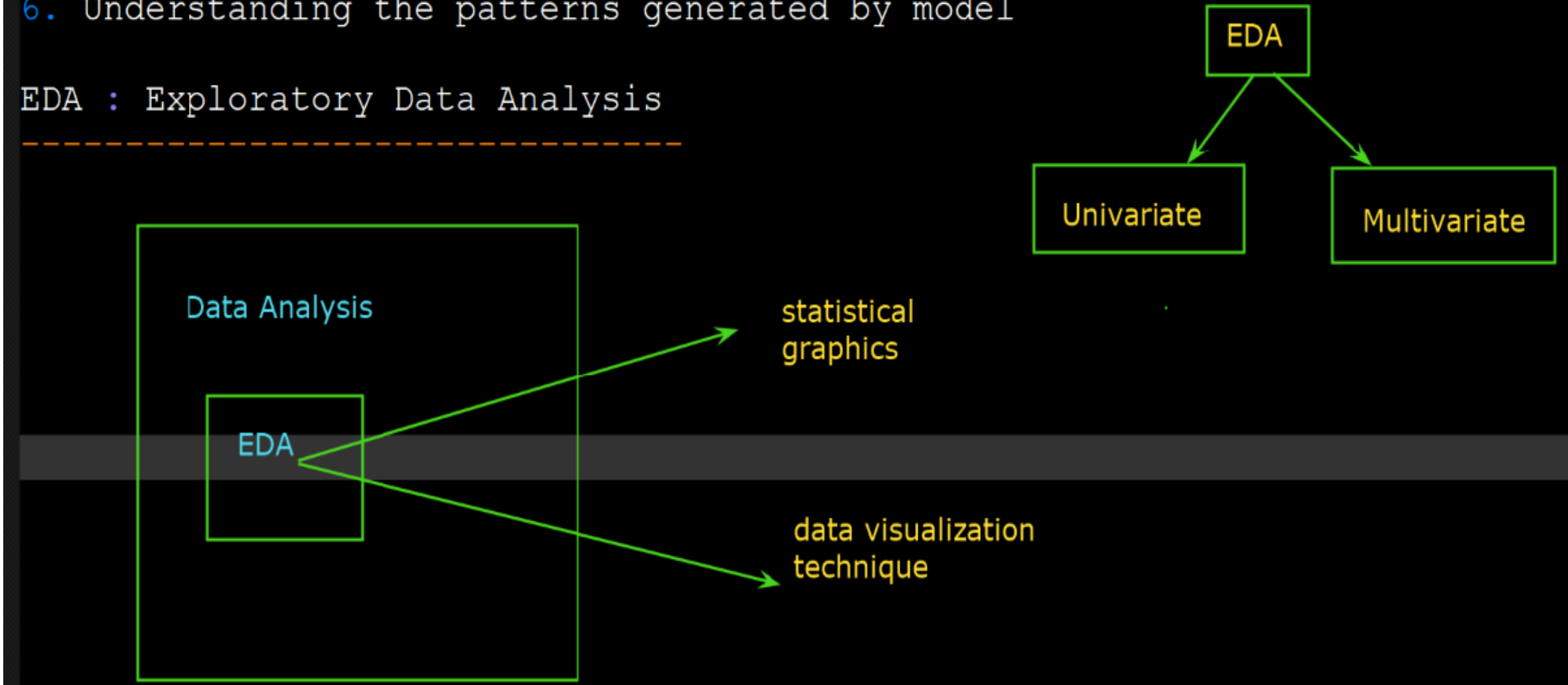


EDA : Exploratory Data Analysis



6. Understanding the patterns generated by model

EDA : Exploratory Data Analysis



Categorical Summary Statistics

- Summary statistics for a categorical feature:
  - Frequencies of different classes.
  - Mode: category that occurs most often.
  - Quantiles: categories that occur more than t times.

Population by year, by province and territory (Number)

	2014
Canada	**35,540.4**
Newfoundland and Labrador	527.0
Prince Edward Island	146.3
Nova Scotia	942.7
New Brunswick	753.9
Quebec	8,214.7
Ontario	13,678.7
Manitoba	1,282.0
Saskatchewan	1,125.4
Alberta	4,121.7
British Columbia	4,631.3
Yukon	36.5
Northwest Territories	43.6
Nunavut	36.6

Frequency: 13.3% of Canadian residents live in BC.
Mode: Ontario has largest number of residents (38.5%)
Quantile: 6 provinces have more than 1 million people.

Continuous Summary Statistics

- Measures of location for continuous features:
  - Mean: average value.
  - Median: value such that half points are larger/smaller.
  - Quantiles: value such that 'k' fraction of points are larger.

- Measures of spread for continuous features:
  - Range: minimum and maximum values.
  - Variance: measures how far values are from mean.
    - Square root of variance is "standard deviation".
  - Intequantile ranges: difference between quantiles.

Entropy as Measure of Randomness

- Another common summary statistic is entropy.
  - Entropy measures "randomness" of a set of variables.
    - Roughly, another measure of the "spread" of values.
    - Formally, "how many bits of information are encoded in the average example".
  - For a categorical variable that can take 'k' values, entropy is defined by:

$$\text{entropy} = -\sum_{c=1}^{k} p_c \log p_c$$

where p_c is the proportion of times you have value 'c'.

  - Low entropy means "very predictable".
  - High entropy means "very random".
  - Minimum value is 0, maximum value is log(k).
    - We use the convention that $0 \log 0 = 0$.

Distances and Similarities

- There are also summary statistics between features ‘x’ and ‘y’.
  - Hamming distance:
    - Number of elements in the vectors that aren’t equal.
  - Euclidean distance:
    - How far apart are the vectors?
  - Correlation:
    - Does one increase/decrease linearly as the other increases?
    - Between -1 and 1.

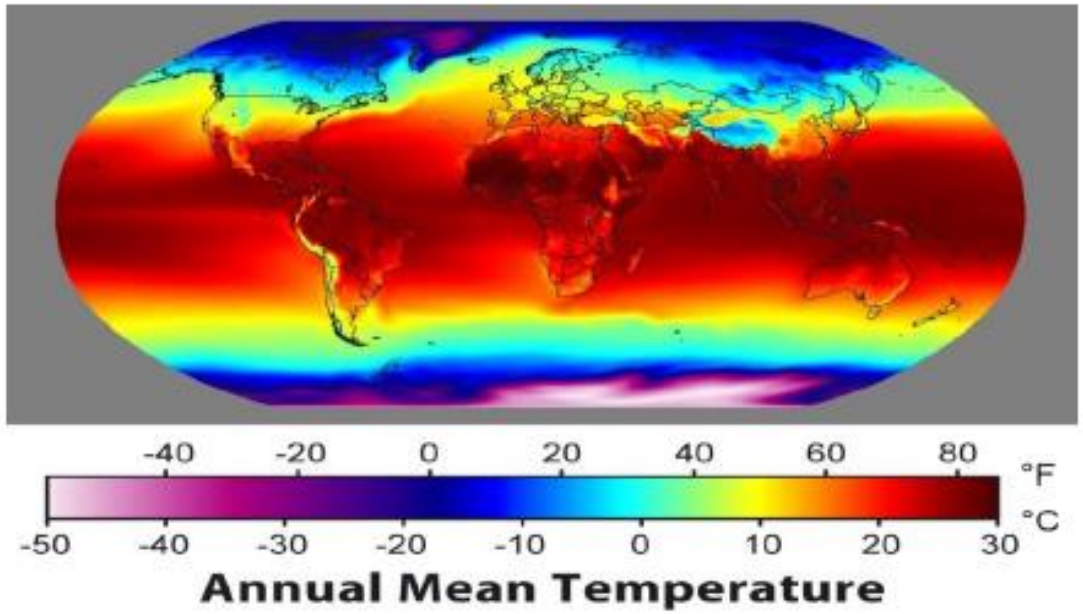
x	y
0	0
0	0
1	0
0	1
0	1
1	1
0	0
0	1
0	1

Visualization

- You can learn a lot from 2D plots of the data:
  - Patterns, trends, outliers, unusual patterns.

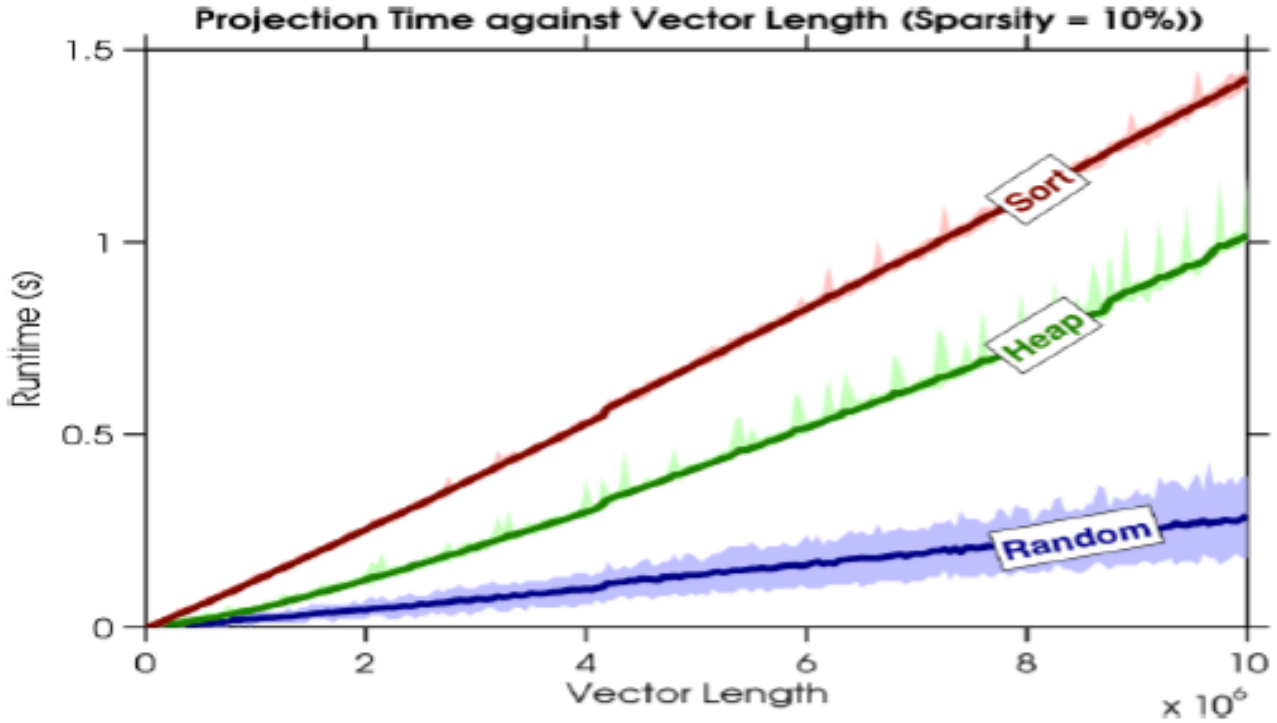
Lat	Long	Temp
0	0	30.1
0	1	29.8
0	2	29.9
0	3	30.1
0	4	29.9
…	…	…

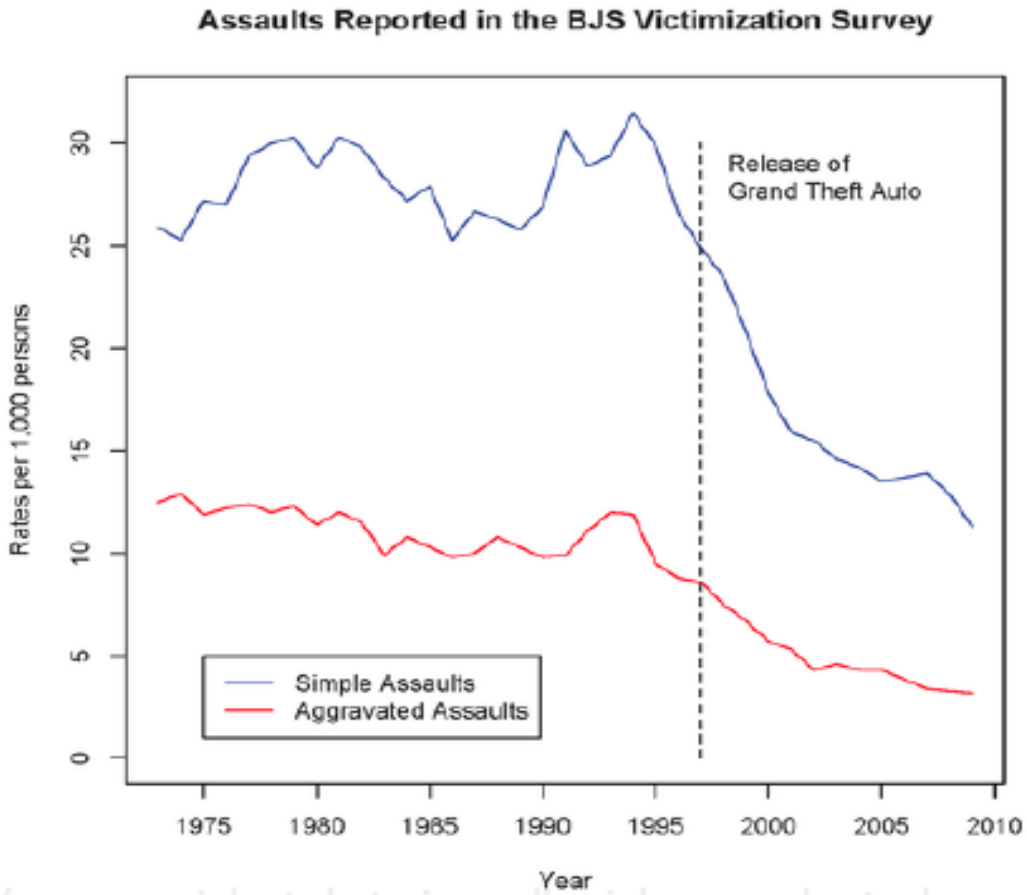
vs.



Basic Plot

- Visualize one variable as a function of another.



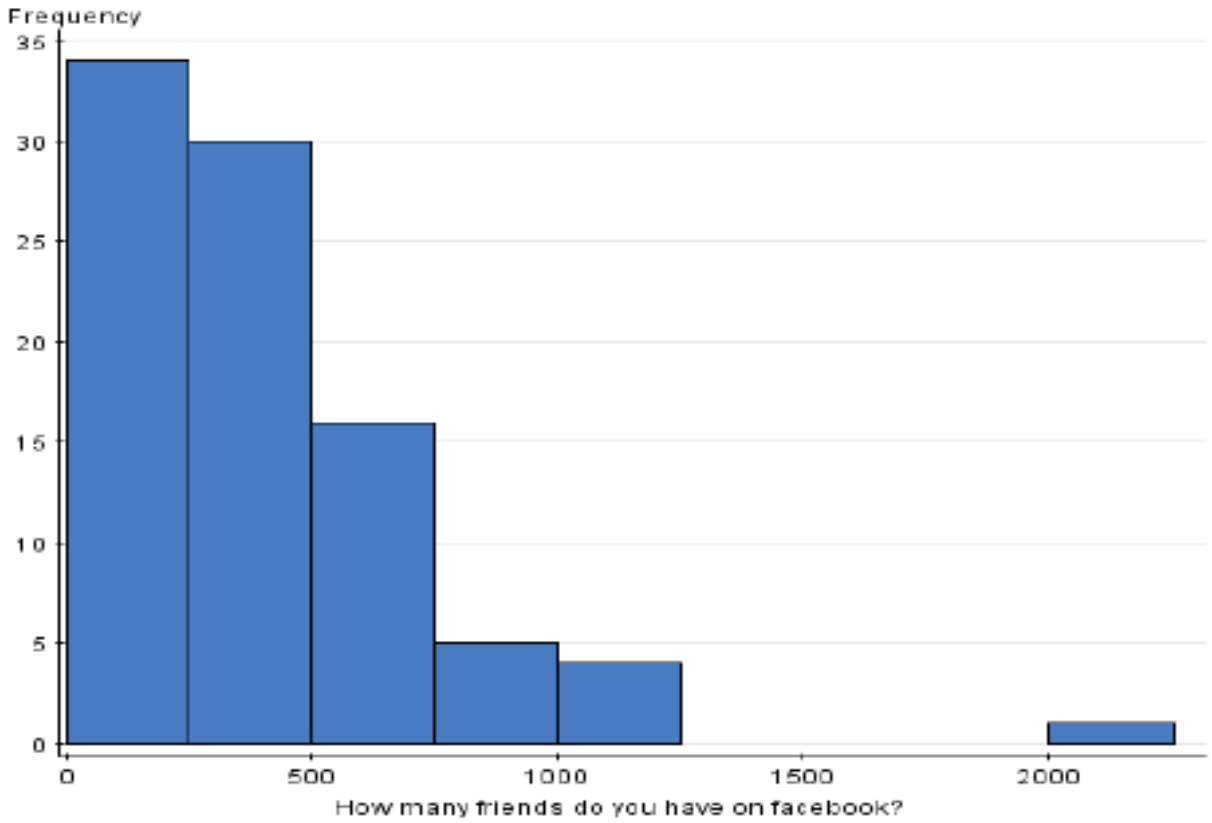


- Fun with plots.

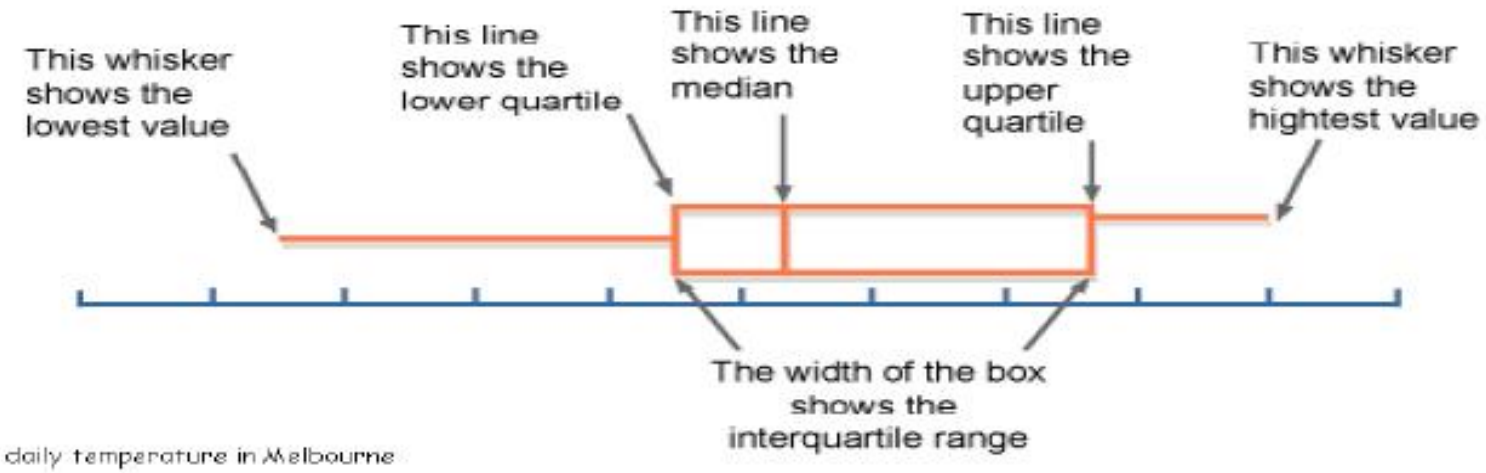
http://notunlikeresearch.typepad.com/something-not-unlike-rese/2011/01/more-on-violent-rhetoric-media-violence-and-actual-

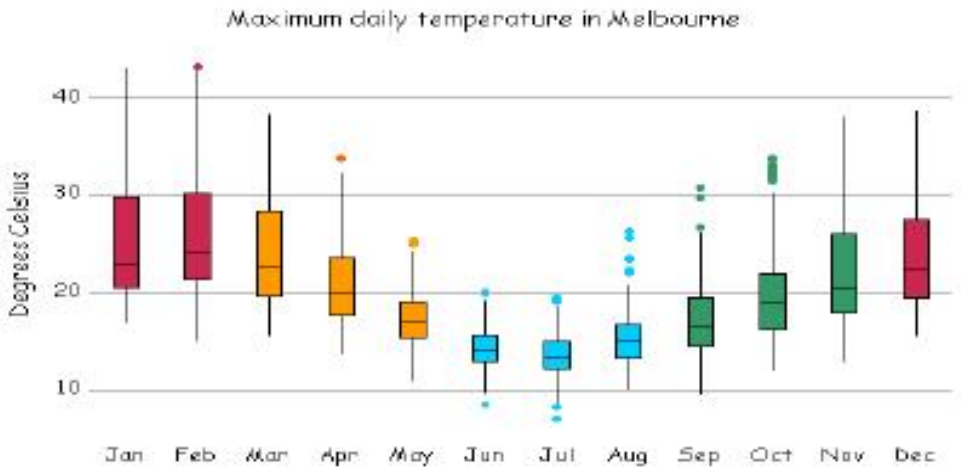
Histogram

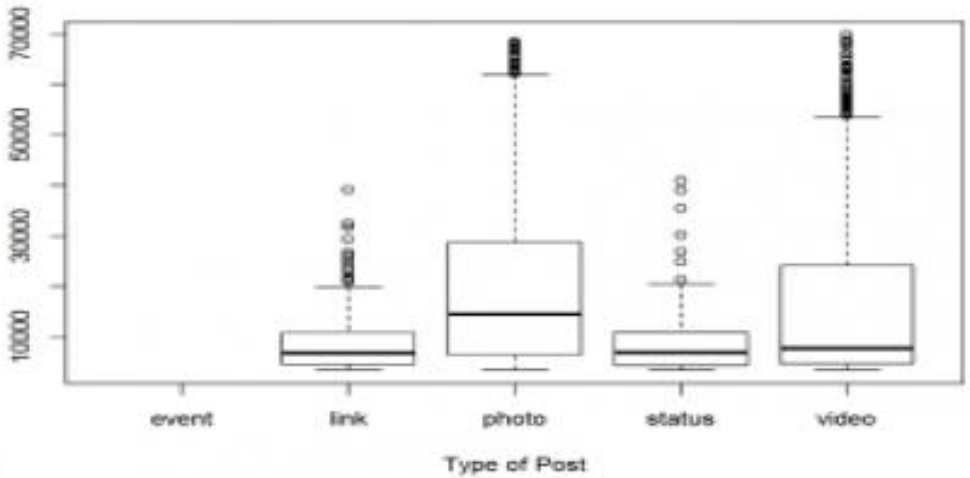
- Histograms display distribution of a variable.



Box Plot







http://www.bbc.co.uk/schools/gcsebitesize/maths/statistics/representingdata3hi

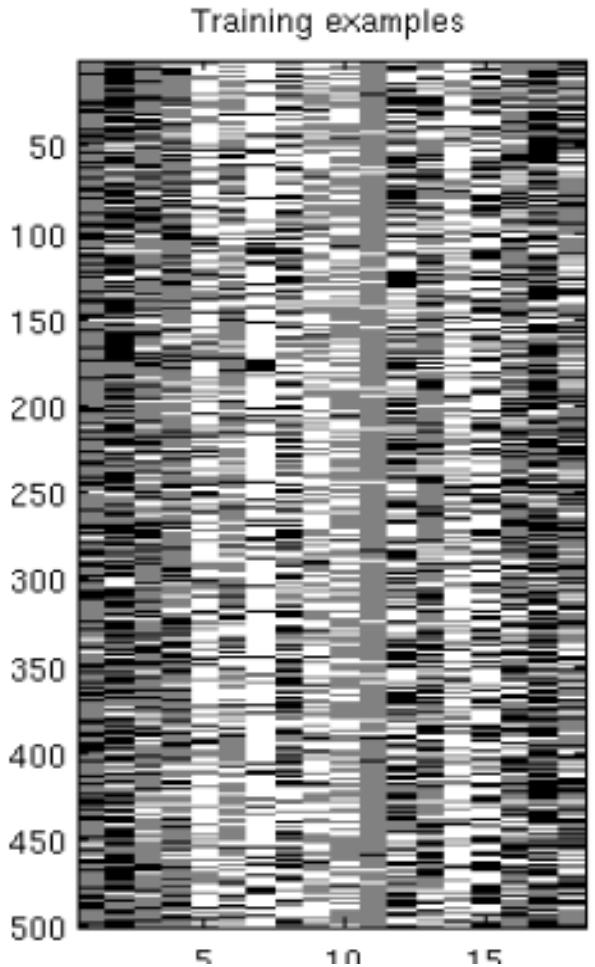
http://www.ccc.ms.unimelb.edu.au/whatisstatistics/weather.html

Box Plot

- Photo from CTV Olympic coverage in 2010:

Matrix Plot

- We can view (examples) x (features) data table as a picture:
  - "Matrix plot".
  - May be able to see trends in features.



Matrix Plot

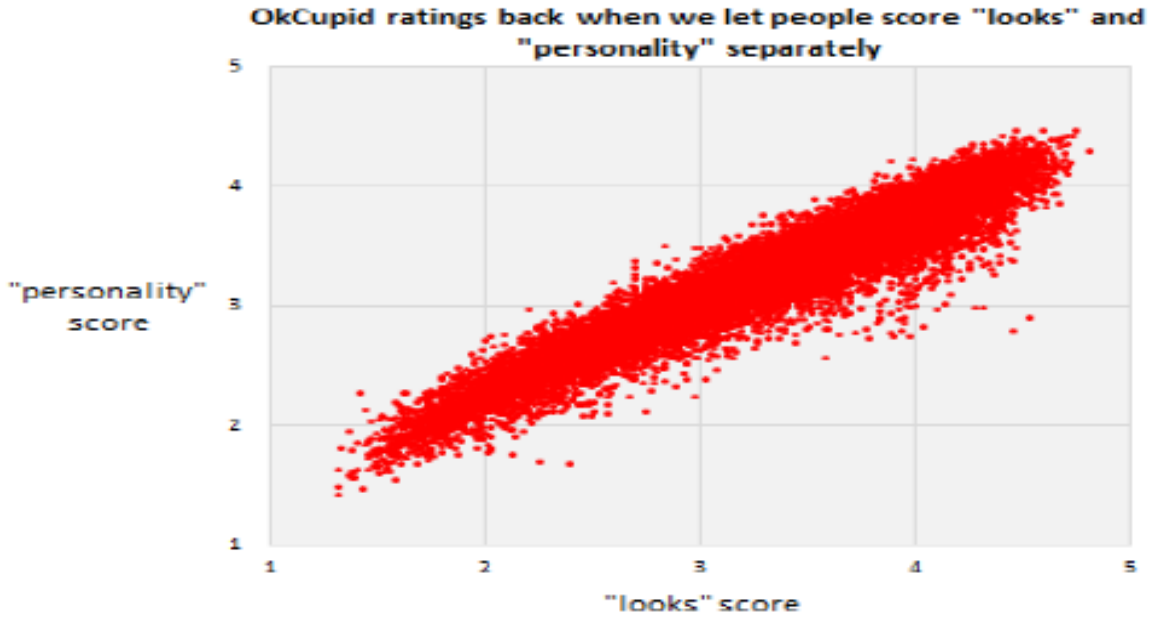
- A matrix plot of all similarities (or distances) between features:
  - Colour used to catch attention.

	BTC	ETH	XRP	XEM	ETC	LTC	DASH	XMR
BTC	1.00	0.61	0.36	0.51	0.60	0.56	0.55	0.66
ETH	0.61	1.00	0.28	0.49	0.68	0.43	0.70	0.64
XRP	0.36	0.28	1.00	0.48	0.08	0.35	0.40	0.44
XEM	0.51	0.49	0.48	1.00	0.40	0.43	0.47	0.52
ETC	0.60	0.68	0.08	0.40	1.00	0.47	0.56	0.53
LTC	0.56	0.43	0.35	0.43	0.47	1.00	0.59	0.67
DASH	0.55	0.70	0.40	0.47	0.56	0.59	1.00	0.74
XMR	0.66	0.64	0.44	0.52	0.53	0.67	0.74	1.00

"Correlation plot"

Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.

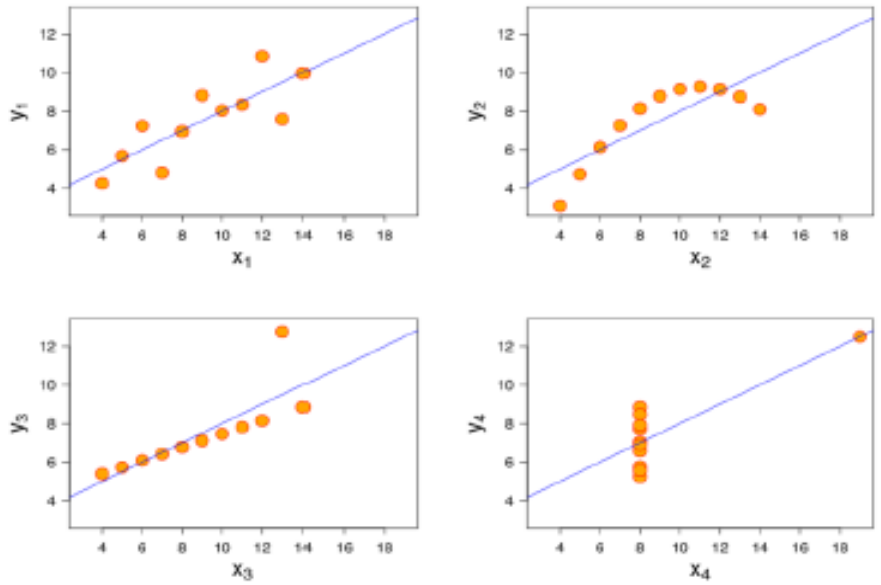


http://cdn.okccdn.com/blog/humanexperiments/looks-v-personality.png

- Shows correlation between “personality” score and “looks” score.

Scatterplot

- Look at distribution of two features:
  - Feature 1 on x-axis.
  - Feature 2 on y-axis.
  - Basically a “plot without lines” between the points.



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

- Shows correlation between “personality” score and “looks” score.
- But scatterplots let you see more complicated patterns.