STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

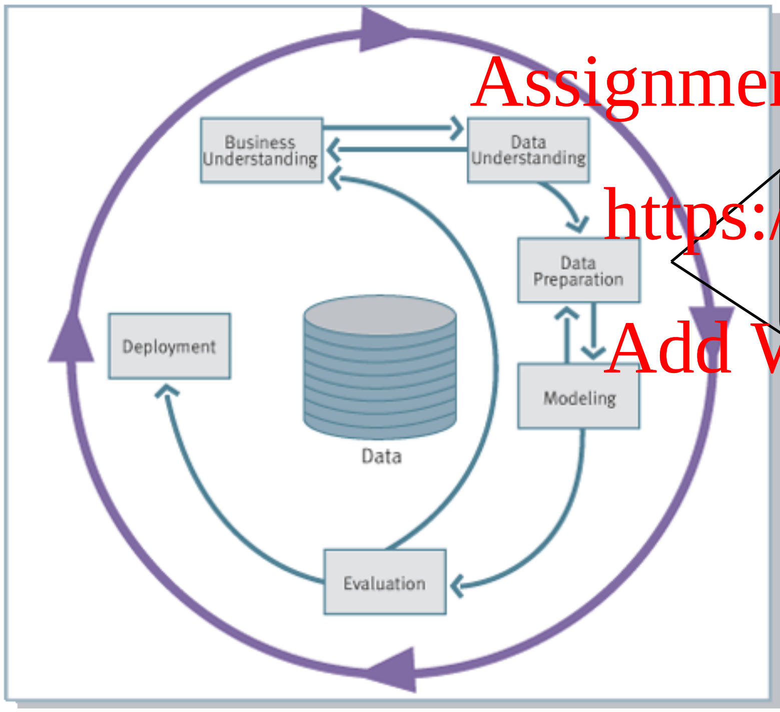
Machine Learning and Data Mining

Data management: generalized tools and techniques

Carlo Lipizzi
clipizzi@stevens.edu

SSE

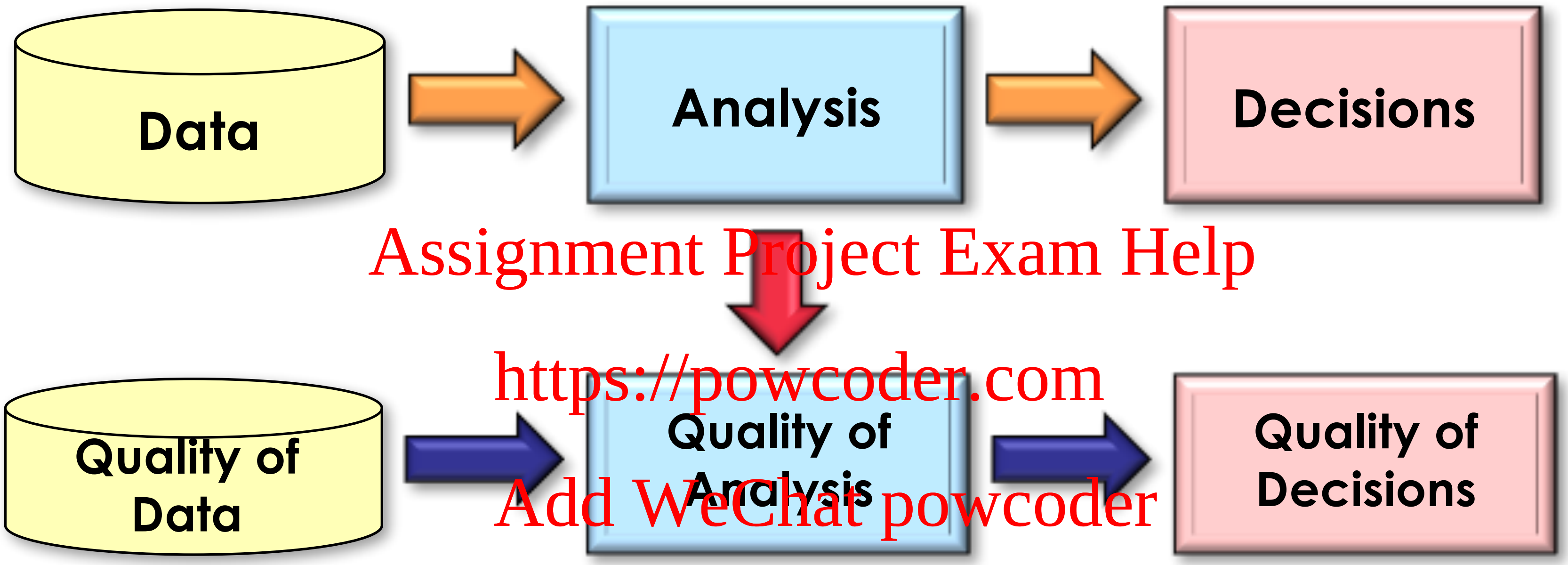
Knowledge Discovery Process, in practice



Data Preparation

Data Preparation estimated to take 70-80% of the time and effort

Data Processing Flow



- Types of Data Quality Problems:
  - Ambiguity
  - Uncertainty
  - Erroneous data values
  - Missing Values
  - Duplication
  - etc

Approaching Data Quality

We need a multi-disciplinary approach to attack data quality problems

- No one approach solves all problem

- **Process management**
  - Ensure proper procedures
- **Statistics**
  - Focus on analysis: find and repair anomalies in data.
- **Database**
  - Focus on relationships: ensure consistency.
- **Metadata / domain expertise**
  - What does it mean? Interpretation

Metadata

- **Data about the data**
- **Data types, domains, and constraints help, but are often not enough**
- **Interpretation of values**
  - Scale, units of measurement, meaning of labels
- **Interpretation of tables**
  - Frequency of refresh, associations, view definitions
- **Most work done for scientific databases**
  - Metadata can include programs for interpreting the data set

Process Management

Business processes which encourage data quality.

- Standardization of content and formats
- Enter data once, enter it correctly (incentives for sales, customer care)
- Automation
- Assign responsibility : data stewards
- End-to-end data audits and reviews
  - Transitions between organizations.
- Data Monitoring
- Data Publishing
- Feedback loops

Feedback Loops

Data processing systems are often thought of as open-loop systems

- Do your processing then throw the results over the fence?
- Computers don't make mistakes, do they?

Analogy to control systems: feedback loops

- Monitor the system to detect difference between actual and intended
- Feedback loop to correct the behavior of earlier components
- Of course, data processing systems are much more complicated than linear control systems

Example

Sales, provisioning, and billing for telecommunications service

– Many stages involving handoffs between organizations and databases

– Simplified picture

Transition between organizational boundaries is a common cause of problems

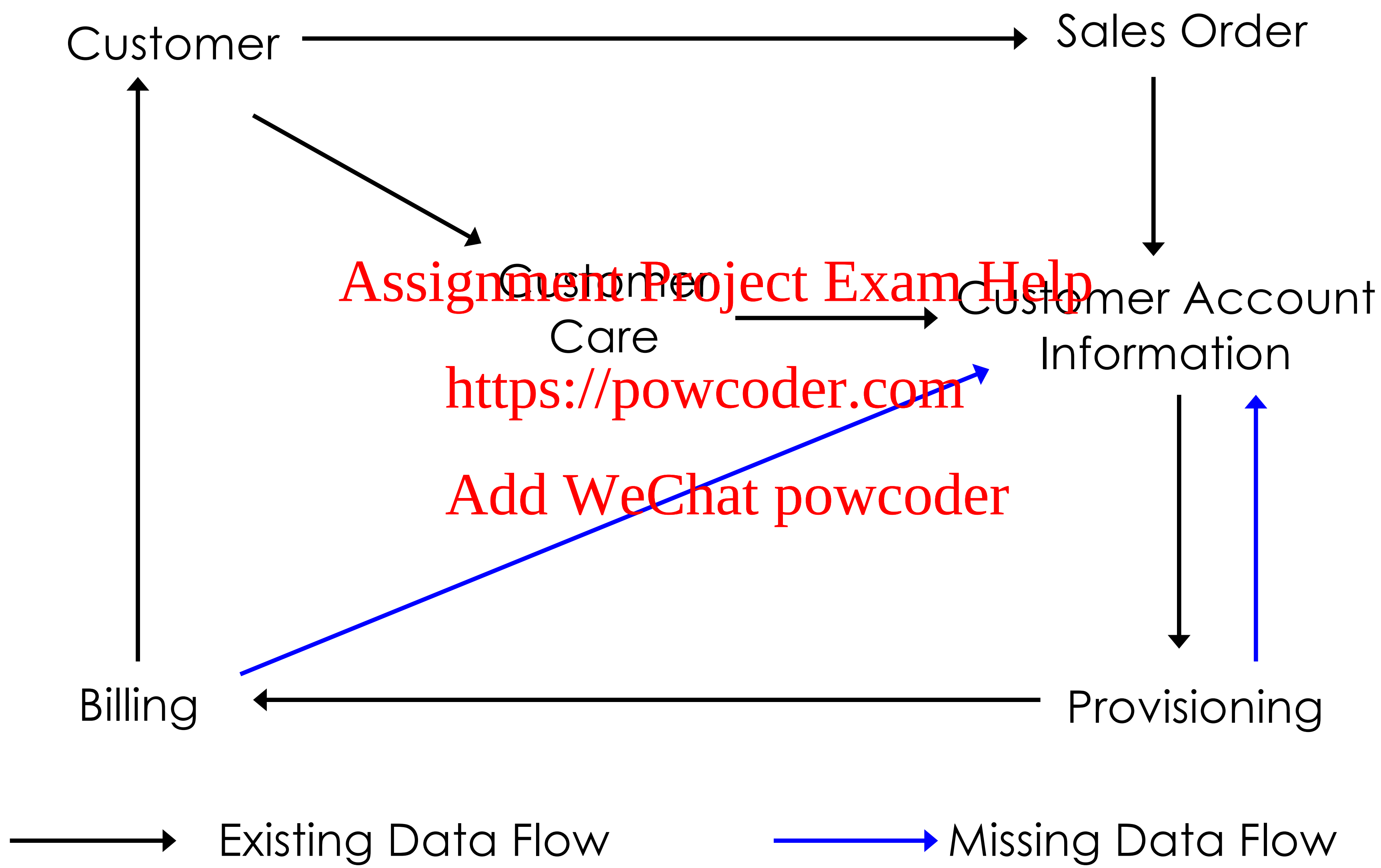
Natural feedback loops

– Customer complains if the bill is to high

Missing feedback loops

– No complaints if we undercharge

Example


Customer
Sales Order
Customer Care
Customer Account Information
Billing
Provisioning
Existing Data Flow
Missing Data Flow

Monitoring

Use data monitoring to add missing feedback loops
Methods:

- Data tracking / auditing
  - Follow a sample of transactions through the workflow.
  - Build secondary processing system to detect possible problems
- Reconciliation of incrementally updated databases with original sources.
- Mandated consistency with a Database of Record
- Feedback loop sync-up
- Data Publishing

Statistical Approaches

No explicit DQ methods

- Traditional statistical data collected from carefully designed experiments, often tied to analysis
- But, there are methods for finding anomalies and repairing data
- Existing methods can be adapted for DQ purposes

Four broad categories can be adapted for DQ

- Missing, incomplete, ambiguous or damaged data e.g. truncated, censored
- Suspicious or abnormal data e.g. outliers
- Testing for departure from models
- Goodness-of-fit

Statistics has two major chapters:

- **Descriptive Statistics**
  – Gives numerical and graphic procedures to summarize a collection of data in a clear and understandable way
- **Inferential statistics**
  – Provides procedures to draw inferences about a population from a sample

Descriptive Measures

- **Central Tendency measures**
  - They are computed to give a "center" around which the measurements in the data are distributed

- **Variation or Variability measures**

  - They describe "data spread" or how far away the measurements are from the center

- **Relative Standing measures**
  - They describe the relative position of specific measurements in the data

Measures of Central Tendency

- **Mean**
  - Sum of all measurements divided by the number of measurements

- **Median**
  - A number such that at most half of the measurements are below it and at most half of the measurements are above it

- **Mode**
  - The most frequent measurement in the data

Example of Mean

Measurements	Deviation
x	**x - mean**
3	-1
5	1
5	1
1	-3
7	3
2	-2
6	2
7	3
0	-4
4	0
40	**0**

- **MEAN = 40/10 = 4**
- **Notice that the sum of the "deviations" is 0**
- **Notice that every single observation intervenes in the computation of the mean**

Excel Example

=AVERAGE(B72:B81)

Example of Median

Measurements	Measurements Ranked
x	x
3	0
5	1
5	2
1	3
7	4
2	5
6	5
7	6
0	7
4	7
40	40

- Median: (4+5)/2 = 4.5
- Notice that only the two central values are used in the computation
- The median is not sensible to extreme values

Excel Example

=MEDIAN(B72:B81)

Example of Mode

Measurements
x
3
5
5
1
7
2
6
7
0
4

- The mode in a list of numbers refers to the list of numbers that occur most frequently
- In this case the data have two modes: 5 and 7
- Both measurements are repeated twice

- Mode: 3
- Notice that it is possible for a dataset not to have any mode

Measurements
x
3
5
1
1
4
7
3
8
3

Excel Example

=MODE(B72:B81)

Maximum, Minimum, and Range

Excel	**Example:**
– =MIN(cellrange)	=MIN(D2:D81)
– =MAX(cellrange)	=MAX(D2:D81)

– There is no explicit command to find the range

– However, it can be easily calculated

– = MAX(D2:D81) - MIN(D2:D81)

Exercise – Companies Values

- **Data set:** *Companies1.xlsx*
  - 25 companies

- **For the 3 numeric variables calculate:**

  - Mean, Mode, Median, Max, Min, Range

- **Can you get any non explicit info from the values you calculated?**
- **Can you create any new variables to get more from your data? What is your goal?**

Variance

- Variance is the average of the **squared** differences from the Mean
- **Steps:**
  - Compute each deviation
  - Square each deviation
  - Sum all the squares
  - Divide by the data size (sample size) minus one: n-1

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Example of Variance

Measurements	Deviations	Square of deviations
x	x - mean	
3	-1	1
5	1	1
5	1	1
1	-3	9
7	3	9
2	-2	4
6	2	4
7	3	9
0	-4	16
4	0	0
40	0	54

- **Variance = 54/9 = 6**
- **It is a measure of "spread"**
- **Notice that the larger the deviations (positive or negative) the larger the variance**

Excel Example

=VAR.P(B72:B81)

Calculates variance based on the entire population

=VAR.S(B72:B81)

Calculates variance based on a sample

The standard deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

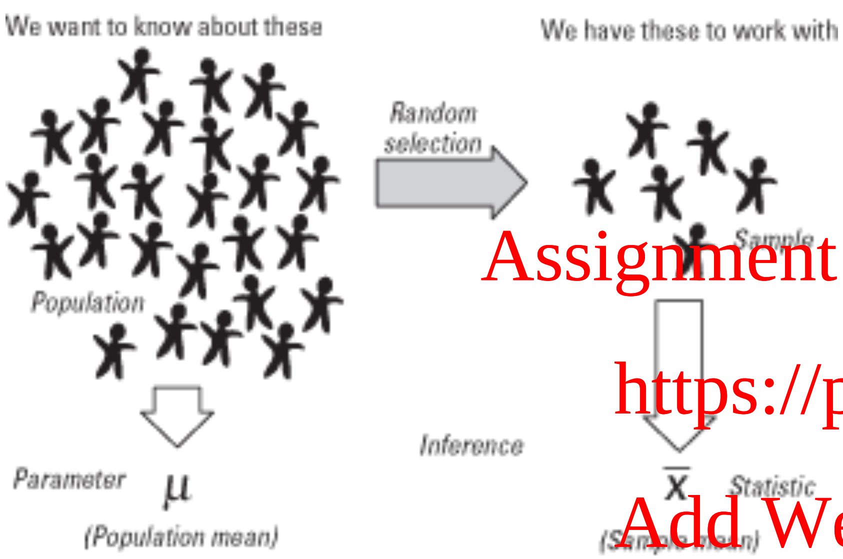
Population **Sample**

- It is defines as the square root of the variance
- In the previous example Variance = 6
- Standard deviation = Square root of the variance = Square root of 6 = 2.45
- We use n-1 instead N (Bessel's correction) to compensate the fact that x_i in Samples tend to be closer to their average

Excel Example

=STDEV.P(B72:B81)
=STDEV.S(B72:B81)

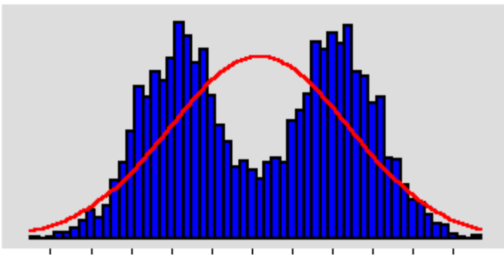
The standard deviation: Sample vs Population



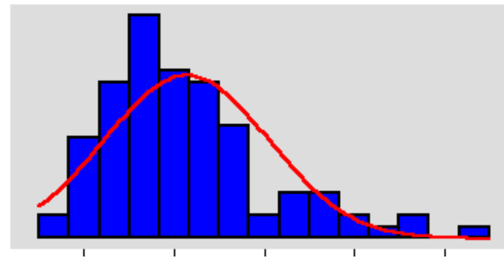
- The standard deviation is a measure of the spread of scores within a set of data
- Usually, we are interested in the standard deviation of a population. However, as we are often presented with data from a sample only, we can estimate the population standard deviation from a sample standard deviation.

- A population includes each element from the set of observations that can be made, while a sample consists only of observations drawn from the population
- These two standard deviations - sample and population standard deviations - are calculated differently. In statistics, we are usually presented with having to calculate sample standard deviations

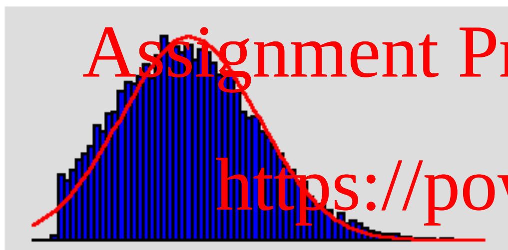
Shape – Patterns of Frequency



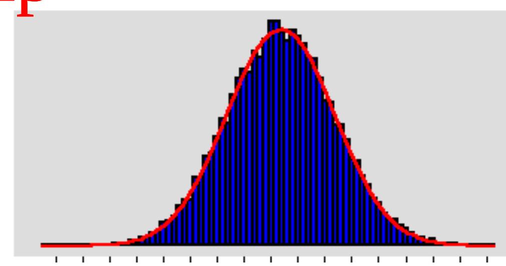
BIMOD



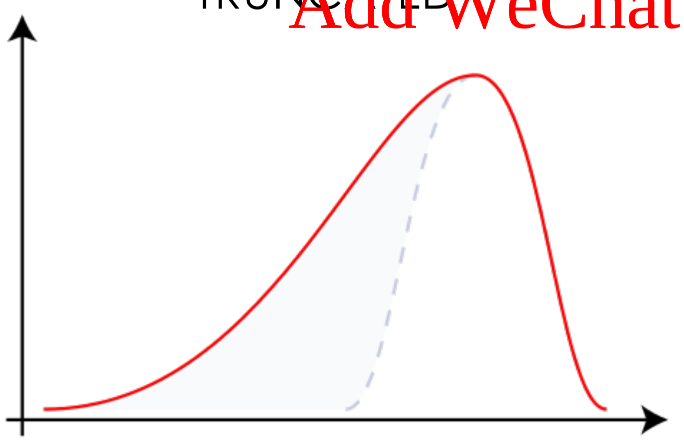
SKEWED



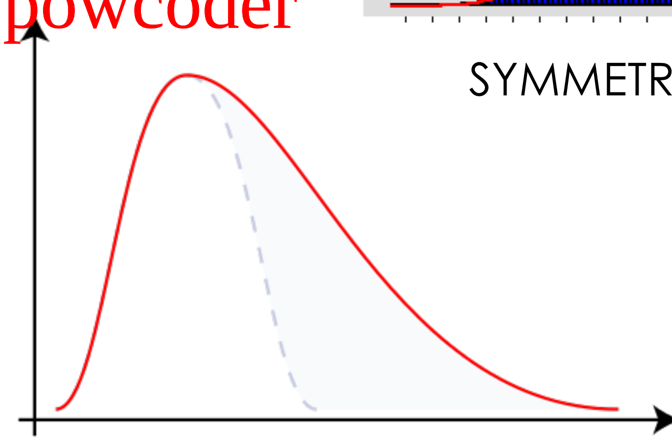
TRUNCATED



SYMMETRICAL



Negative Skew



Positive Skew

Percentiles

- The p^{th} percentile is a number such that at most p% of the measurements are below it and at most 100 – p percent of the data are above it
- Example, if in a certain data the 85th percentile is 340 means that 15% of the measurements in the data are above 340. It also means that 85% of the measurements are below 340
- Notice that the median is the 50th percentile

For any data

- At least 75% of the measurements differ from the mean less than twice the standard deviation

- At least 89% of the measurements differ from the mean less than three times the standard deviation

Note: This is a general property and it is called Tchebysheff's Inequality: Given a number $k >= 1$ and a population with n measurements, at least $1-1/k^2$ of the measurements will lie within k standard deviations of their mean. It is true for every dataset

Example of Tchebysheff's Inequality

- Suppose that for a certain data is : Mean = 20
- Standard deviation =3

- ***Bottom line: the rule guarantees that in any probability distribution, "nearly all" values are close to the mean***

Then:

- A least 75% of the measurements are between 14 and 26
- At least 89% of the measurements are between 11 and 29

Further Notes

- When the Mean is greater than the Median the data distribution is skewed to the Right

- When the Median is greater than the Mean the data distribution is skewed to the Left

- When Mean and Median are very close to each other the data distribution is approximately symmetric

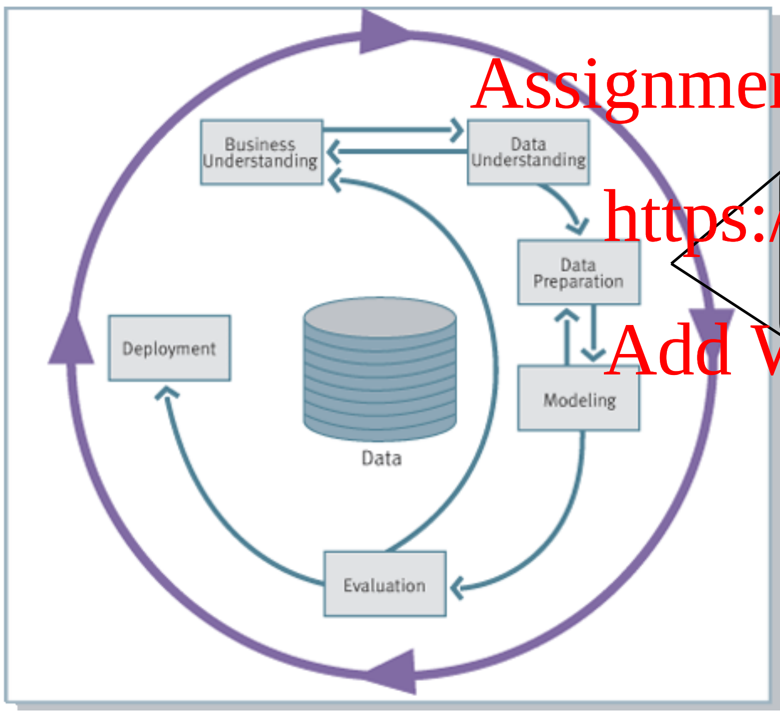
Exercise – Starting Salaries – 15'

- **Data set:** *StartSalary.xlsx*
  - 12 datapoint

- **Calculate:**

  - Mean, Mode, Median, Standard Deviation, Sample Variance, Skewness, Max, Min, Range

Knowledge Discovery Process, in practice



Data Preparation

Data Preparation estimated to take 70-80% of the time and effort

Data Cleaning / Quality

- **Individual measurements**
  - Random noise in individual measurements
    - Outliers
    - Random data entry errors
    - Noise in label assignment (e.g., class labels in medical data sets)
    - can be corrected or smoothed out
  - Systematic errors
    - E.g.: all ages > 99 recorded as 99
    - More individuals aged 20, 30, 40, etc. than expected
- **Missing information**
  - Missing at random
    - Questions on a questionnaire that people randomly forget to fill in
  - Missing systematically
    - Questions that people don't want to answer
    - Patients who are too ill for a certain test

Handling Missing Data: 3 alternatives

- **Replace Missing Values with User-defined Constants**
  – Missing numeric values replaced with 0.0
  – Missing categorical values replaced with "Missing"
- **Replace Missing Values with Mode or Mean**
- **Replace Missing Values with Random Values**
  – Values randomly taken from underlying distribution
  – Method superior compared to mean substitution

Exercise on Handling Missing Data

- Examine *cars.txt* dataset containing records for 261 automobiles manufactured in 1970s and 1980s
- Examine the file and handle missing data
  - –Use one or more of the three alternative methods

Identifying Outliers

- Outliers are values that lie near extreme limits of data range
- Outliers may represent errors in data entry
- Certain statistical methods may produce unstable results
- Some data mining algorithms benefit from normalized data

Graphical Methods for Identifying Outliers

- A histogram examines values of numeric fields
- Gives us the possibility to identify the outliers and then decide what to do
- Multidimensional graphs could provide more insights

Exercise on Handling Outliers

- Examine *cars_full.txt* dataset containing full records for 261 automobiles manufactured in 1970s and 1980s
- Examine the file for outliers

Data Transformation - Normalization

- Variables tend to have ranges different from each other
- Some data mining algorithms adversely affected by differences in variable ranges
- Variables with greater ranges tend to have larger influence on data model's results
- Therefore, numeric field values should be normalized

Normalization – Min-Max

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

- Determines how much greater the selected field value is than minimum value for field

- Scales this difference by field's range

Min-Max - Example

- From the *cars* dataset, normalize the value for a vehicle taking 25 seconds to reach 60mph
- Max(time-to-60) = 25

$$X^* = \frac{X - \min(X)}{\max(X) - \min(X)} = \frac{25 - 8}{25 - 8} = 1.0$$

- Maximum field values have Min-max Normalization value = 1
- Min-max Normalization values range [0, 1]

Z-score Standardization

$$X^* = \frac{X - mean(X)}{SD(X)}$$

- Widely used in statistical analysis
- Takes difference between field value and field value mean
- Scales this difference by field's standard deviation

Z-score - Example

- Same: From the *cars* dataset, normalize the value for a vehicle taking 25 seconds to reach 60mph

$$X^* = \frac{X - mean(X)}{SD(X)} = \frac{8 - 15.548}{2.911} = 2.593$$

- Data values that lie below the mean have negative Z-score Standardization values

Z-score – Key points

- Z-score Standardization values typically range [-4, 4]
- Field values below field mean → negative Z-score Standardization values
- Field values equal to field mean → Z-score Standardization value = 0
- Field values above field mean → positive Z-score Standardization values

Exercise on Normalization

- Using the *cars_full.txt* dataset normalize the "time-to-60" values
- Use either Min-Max or Z-score method

Data Transformation – Data Reduction

– **Dimension Reduction**

- In general, incurs loss of information about x
- If dimensionality p is very large (e.g., 1000's), representing the data in a lower-dimensional space may make learning more reliable
  - e.g.: clustering example
    - 100 dimensional data
    - if cluster structure is only present in 2 of the dimensions, the others are just noise
    - if other 98 dimensions are just noise (relative to cluster structure), then clusters will be much easier to discover if we just focus on the 2d space
- Dimension reduction can also provide interpretation/insight (e.g.: for 2d visualization purposes)

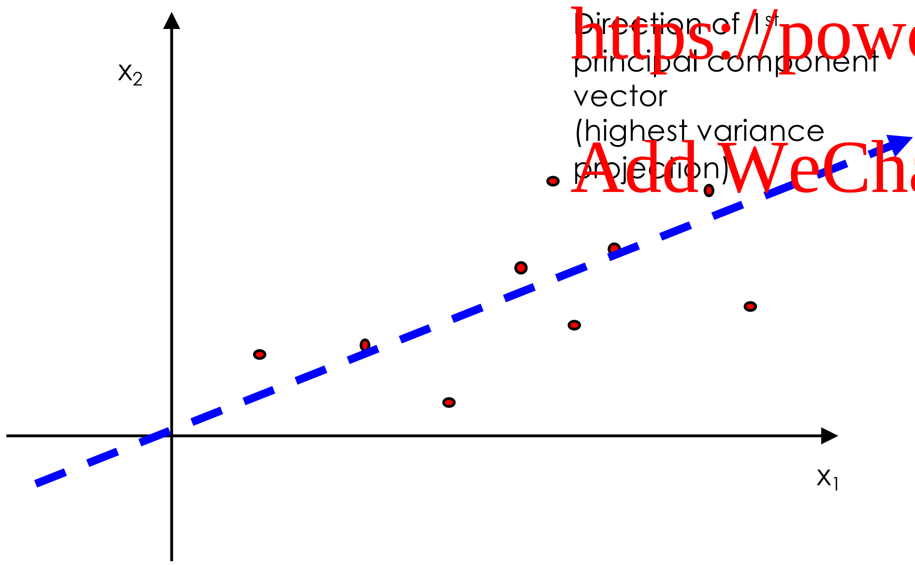
Data Reduction - Methods

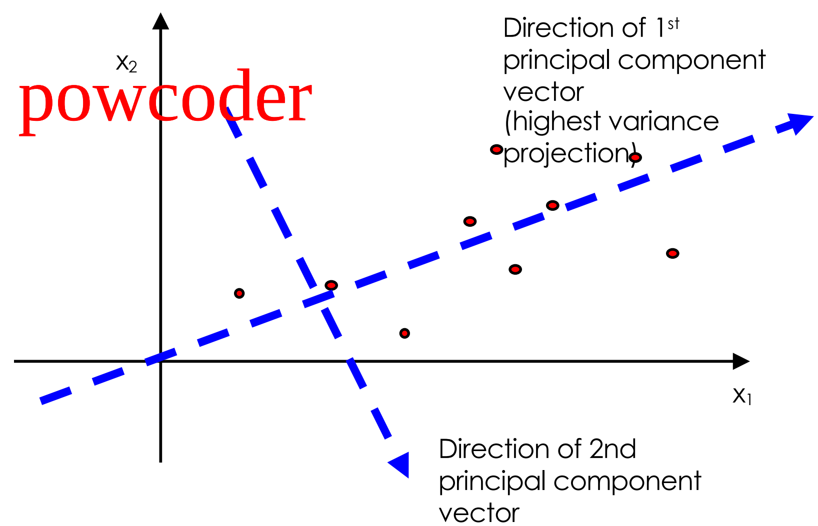
- **Sampling**
  - Choose a representative subset of the data

    Simple random sampling may be ok but beware of skewed variables
- **Principal Component**
  - One of several projection methods
  - Idea: Find a projection of your data in a lower dimension, that maximizes the amount of information retained

Data Reduction - Principal Component

Using orthogonal transformations, converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components





Data Reduction - Consolidation

- **Consolidating variables to create new logical variables**

  This is very domain-dependent and may create new insights on the data

- **In the *cars* dataset, creating the variable "hp/weight" can provide an indication of power/unit and make vehicles more comparable one to the other**