Australian National University

COMP3430 / COMP8430 Data wrangling

Lecture 7: Data transformation, aggregation and reduction

(Lecturer: Thilina Ranbaduge)

Lecture outline

- Data transformation
  - Generalisation
  - Normalisation
  - Attribute/feature construction
- Data aggregation
- Data reduction
- Summary

Data transformation

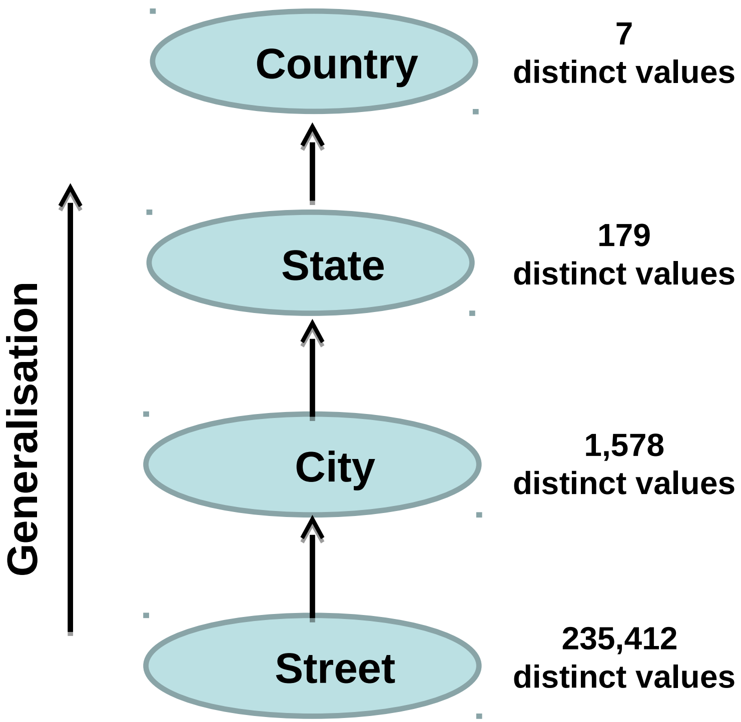
- Generalisation
  - Using concept hierarchy
- Normalisation
  - Scale data to fall within a small (specified) range
  - Min-max normalisation, z-score normalisation, decimal scaling, logarithm transformation
- Attribute/feature construction
  - New attributes constructed by applying a function on existing attributes

Generalisation (1)

- Based on concept hierarchy or value generalisation hierarchy
- Concept hierarchy – specifies ordering of attributes explicitly at the schema level (as discussed in the data warehousing lecture)
  - For example, Street < City < State < Country
- Value generalisation hierarchy – specifies a hierarchy for the values of an attribute by explicit data grouping
  - For example, {Dickson, Lyneham, Watson} < Canberra

Generalisation (2)

- Some concept hierarchies can be automatically generated
  - Based on the number of distinct values in each attribute
  - The attribute with the most distinct values is in the lowest level of the hierarchy
  - Day, month, year and time attributes are exception!



Normalisation (1)

- Min-max [0-1] normalisation
  - Subtracting the minimum value and dividing by the difference between maximum and minimum values
- Z-score normalisation
  - Subtracting the mean value and dividing by the standard deviation
- Robust normalisation
  - Subtracting the median value and dividing by the median absolute deviation

Normalisation (2)

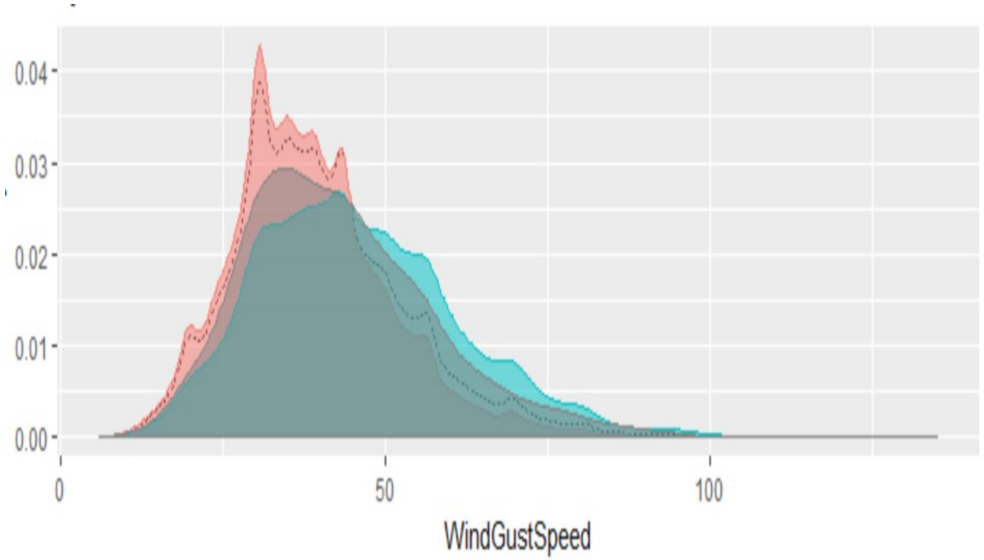
	smallest		*largest*		*median*										
Values	5	27	100	59	28	48	50	39	9	7	20	63	10	41	9
Min-max	0	0.23	1	0.57	0.24	0.45	0.47	0.36	0.04	0.02	0.16	0.61	0.05	0.38	0.04
Z-score	-1.13	-0.28	2.54	0.95	-0.24	0.53	0.61	0.18	-0.98	-1.06	-0.55	1.11	-0.94	0.26	-0.98
Robust	-1.08	-0.05	3.38	1.46	0	0.94	1.03	0.52	-0.89	-0.99	-0.37	1.64	-0.84	0.61	-0.89

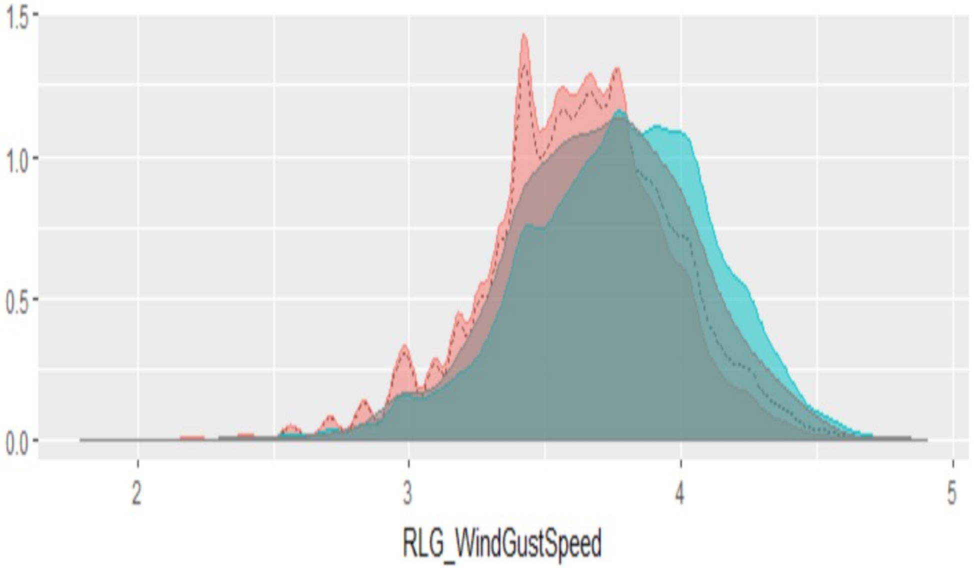
Normalisation (3)

- Logarithm normalisation
  - For attributes with skewed distribution (such as income)
  - Transforms a broad range of numeric values to a narrower range of numeric values
  - Useful when data have outliers with extremely large variance
  - For example, using a base 10 logarithm function, a list of income values [$10,000, $100,000, $150,000, $1,000,000] is transformed into [4, 5, 5.18, 6]

Normalisation (4)

- Logarithm normalisation on the WindGustSpeed attribute in the Rattle Weather data set





Attribute / feature selection (1)

- Reduce the number of features/attributes that are not significant for a certain data science project
- Select a minimum set of features/attributes such that
  - The probability of different classes or information gain given the values for these features is as close as possible given all the features
- Exponential number of choices
  - 2^d possible combinations of sub-features from *d* features

Attribute / feature selection (2)

- Step-wise forward selection
  - Best feature is selected first, then the next best feature condition to the first is selected, and so on
- Step-wise backward elimination
  - Repeatedly eliminate the least useful feature
- Combining forward selection and backward elimination
  - Repeatedly select best and eliminate worst features
- Decision-tree induction *(machine learning-based)*

Attribute / feature construction

- A process of adding derived features to data (also known as constructive induction or attribute discovery)
- Construct new attributes/features based on existing attributes/features
  - Combining or splitting existing raw attributes into new one which have a higher predictive power
  - For example splitting date attribute into month and year attributes for monthly and annual processing
  - Generating new attribute on tax exclusive price values

Data aggregation

- Compiling and summarising data to prepare new aggregated data

- The aim is to get more information about particular groups based on specific attributes, such as age, income, and location
  - For example, aggregated phone usage of customers by age and location in a phone calling list data set

- Can also be aggregated from multiple sources

Data reduction

- Volume of data increases with the Big data growth
- A process of reducing data volume by choosing smaller forms of representation

- Parametric methods:
  - Construct model fitting the data, estimate model parameters, store only the parameters, and discard data
- Non-parametric methods:
  - Based on histograms, clustering, and sampling

Parametric methods (1)

- **Linear regression**: fit the data to a straight line ($Y=wX+b$), the regression coefficients w and b determine the line using the data
- **Multiple regression**: to transform to non-linear functions ($Y=b_0+b_1X_1+b_2X_2$)
- **Log-linear models**: approximate discrete multi-dimensional probability distributions

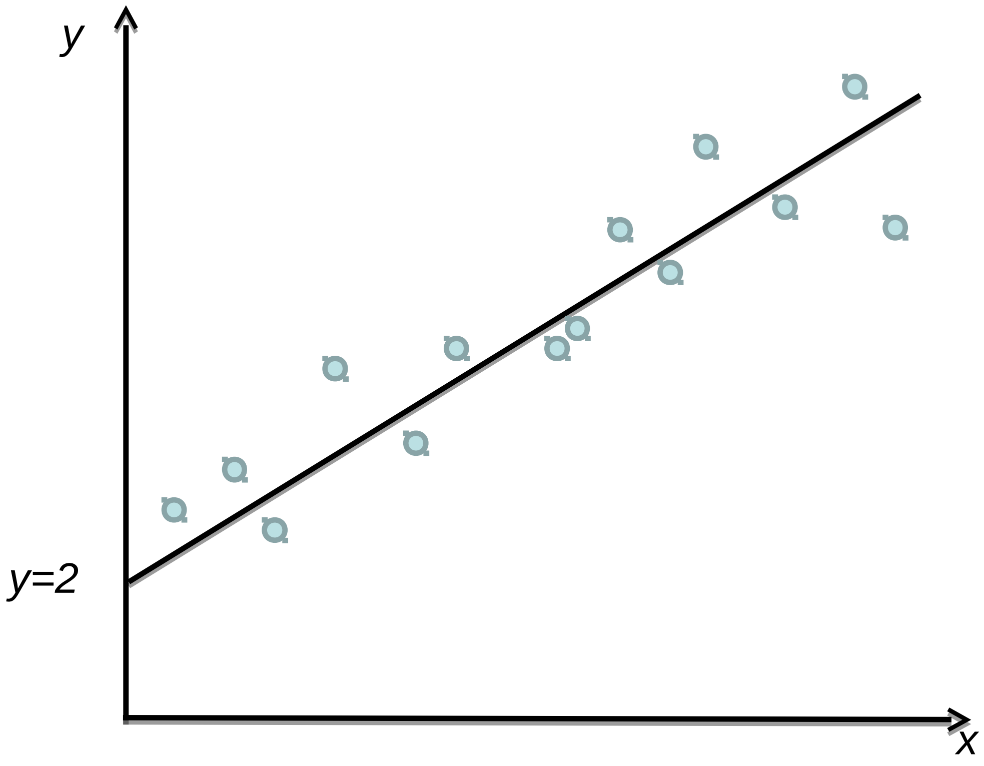
To be covered in more detail in the data mining course

Parametric methods (2)

Example:

Linear regression

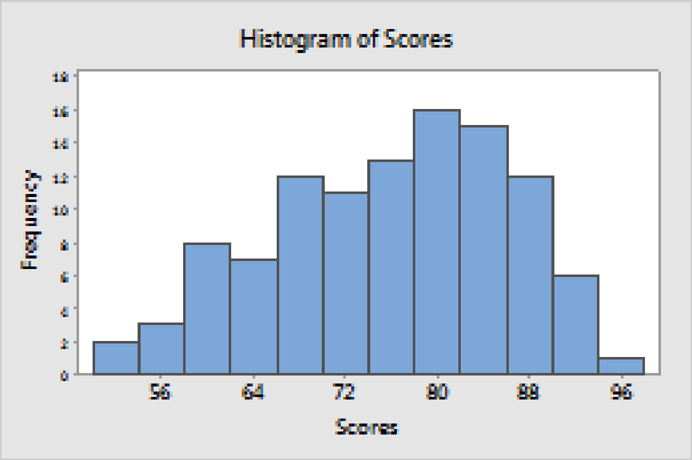
$$y = x + 2$$



To be covered in more detail in the data mining course

Histograms

- Binning:
  - Divides data into buckets and store summary for each bucket (total, average, median)
- Binning methods:
  - Equal width – with equal bin range
  - Equal frequency/depth – with equal bin frequency (same number of data points in each bin)



Clustering (1)

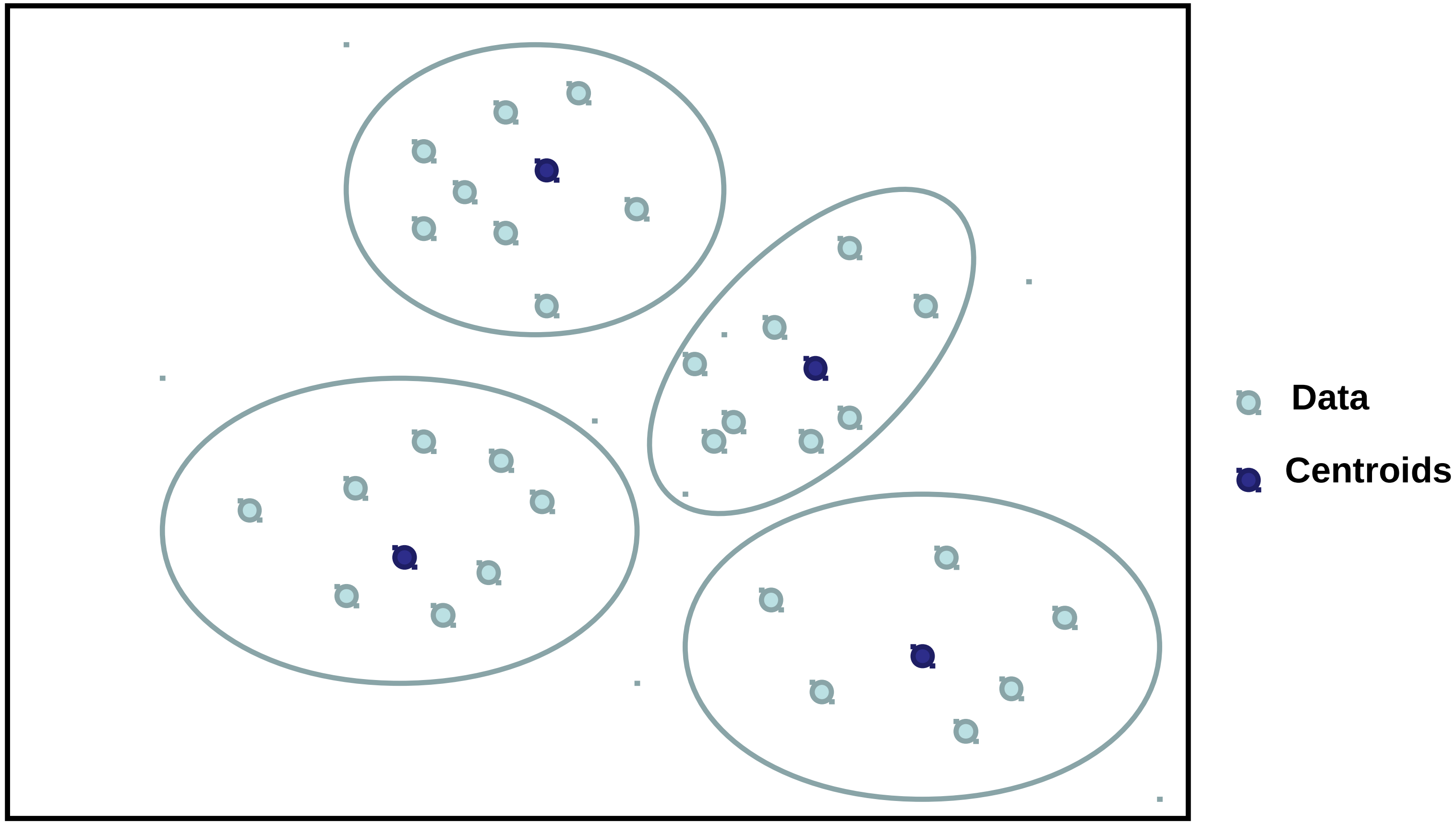
- Clustering:
  - Partition/group data into clusters based on similarity, and store only cluster representation (for example, centroid and diameter only)
- Clustering techniques:
  - **Centroid-based** - **K-means**: assigns data to the nearest cluster center (of k clusters), such that the squared distances from the center are minimised
  - **Connectivity-based** - **Hierarchical clustering**: data belong to a child cluster also belong to the parent cluster
  - **Density-based – DBSCAN**: Clusters data that satisfy a density criterion
  - **Distribution-based – Gaussian mixture models**: Models (iteratively optimized) data with a fixed number of Gaussian distributions

Clustering (2)

Example:

Centroid-based clustering

(k-means with k=4)



To be covered in more detail in the data mining course

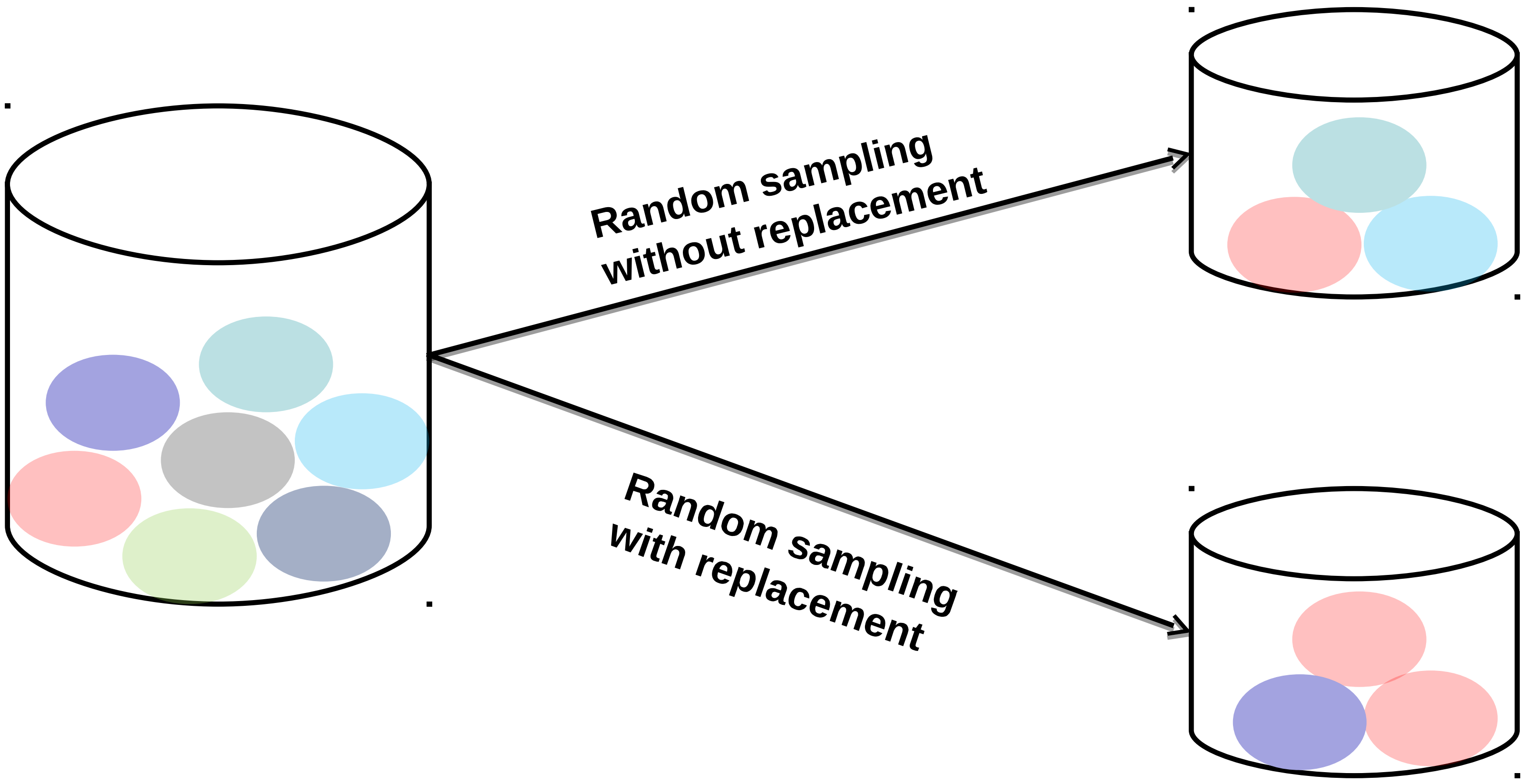
Sampling (1)

- Sampling:
  - Generate a small sample to represent the whole dataset
- Choose a representative subset of the data
- Sampling methods:
  - Simple random sampling does not perform well on skewed data (for example, only a few people with high salary)
  - Stratified sampling is an adaptive sampling method that divides the data into groups (known as *strata*) and a probability sample is drawn from each group

Sampling (2)

Example:

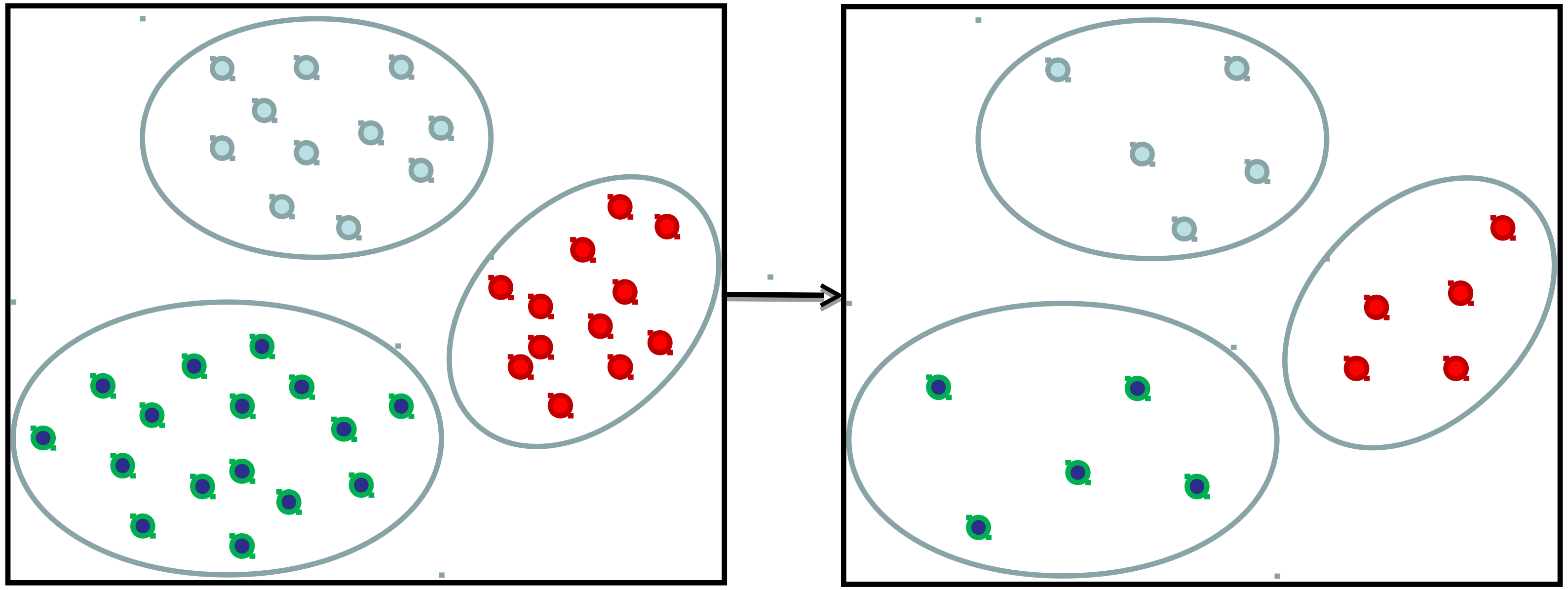
Random sampling



Sampling (3)

Example:

Stratified sampling (sample 5 data points per group / cluster)



Summary

- Data transformation, aggregation, and reduction are being used in data science applications to improve effectiveness and quality of data analysis and mining
- Data pre-processing includes:
  - Data cleaning, transformation, aggregation, and reduction
  - Data standardisation and parsing (will be covered tin lecture 8)
  - Data integration (will be covered later in the course)
- Various methods have been developed for data pre-processing, but this is still an active area of research