Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 6: Resolving data quality issues and data cleaning

(Lecturer: Peter Christen)

Lecture outline

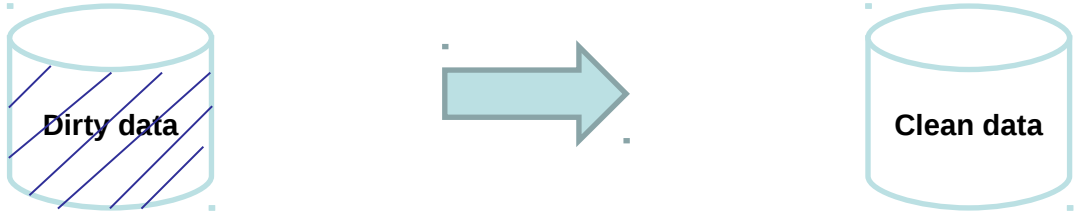
- Data quality issues
- Forms of data pre-processing
- An overview of data cleaning
  - Impute missing data
  - Smooth noisy data
  - Remove duplicate data
  - Resolve inconsistent data
- Summary

Data quality issues

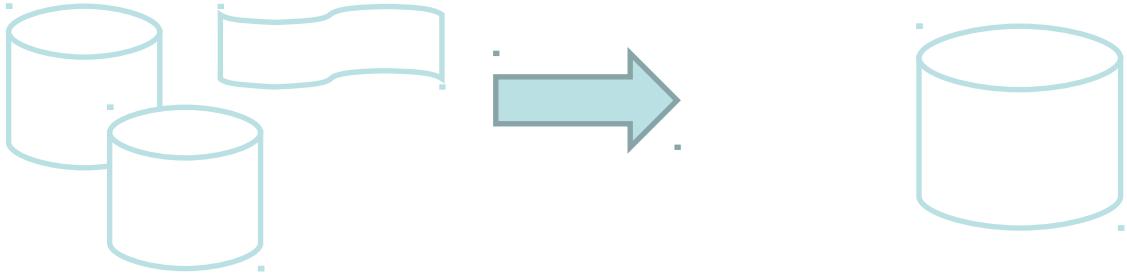
- Various causes of data errors:
  - Data entry errors / subjective judgment
  - Limited (computing) resources
  - Security / accessibility trade-off
  - Complex data, adaptive data
  - Volume of data
  - Redundant data
  - Multiple sources / distributed heterogeneous systems

Forms of data pre-processing

Data cleaning



Data integration



Data transformation

-1	27	100	57	63

-0.01	0.27	1.0	0.57	0.63

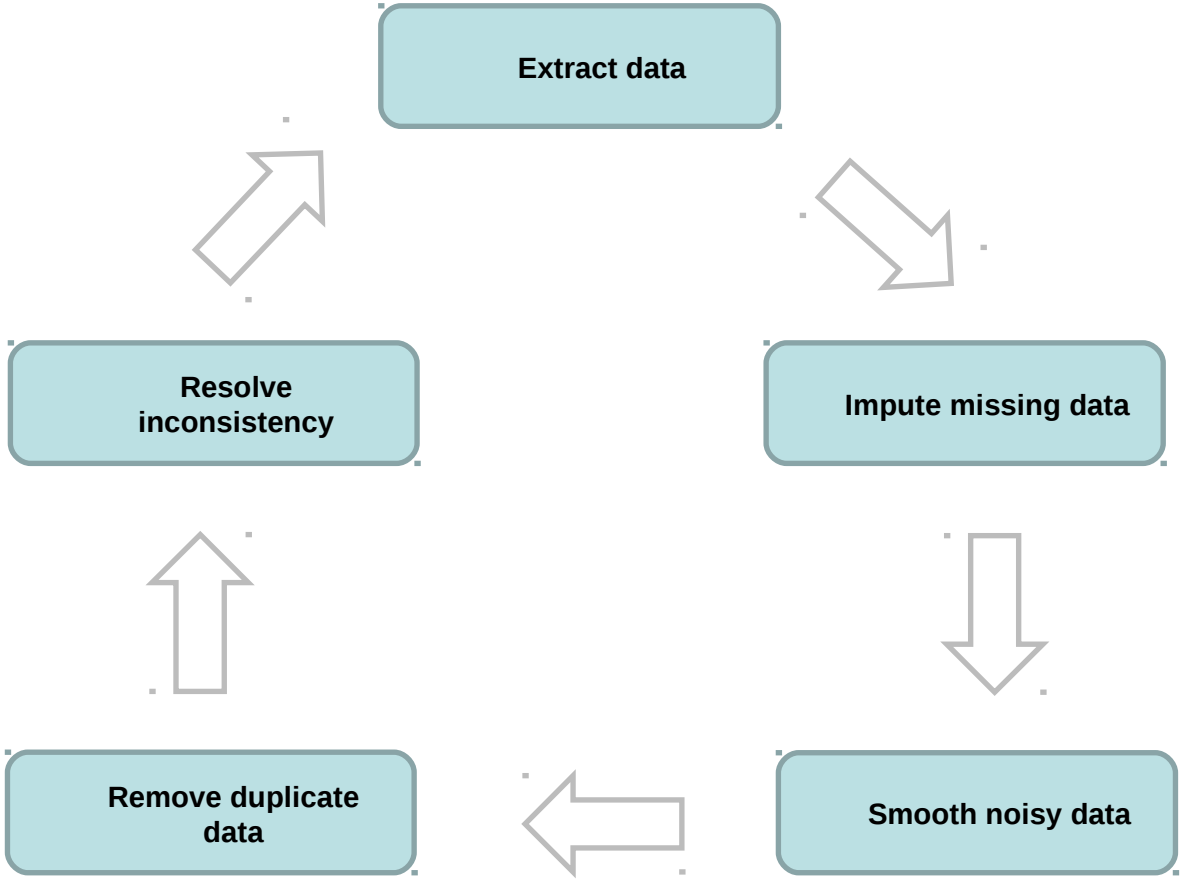
Data reduction

A1	A2	…..	A126
R1			
…… R8000			

A1	A2	…..	A100
R1			
…… R1000			

Data cleaning: An overview

- **A highly crucial data pre-processing step**
- Includes various tasks:
  - Dealing with missing data
  - Handling outliers and noisy data
  - Removing redundant and duplicate data
  - Resolving inconsistencies



Missing data

- One of the most common data quality issues is missing data

- Absence of attribute values due to various reasons
  - Equipment malfunction
  - Not entered due to misunderstanding
  - Not considered important during data entry
  - Deleted due to inconsistency with other values

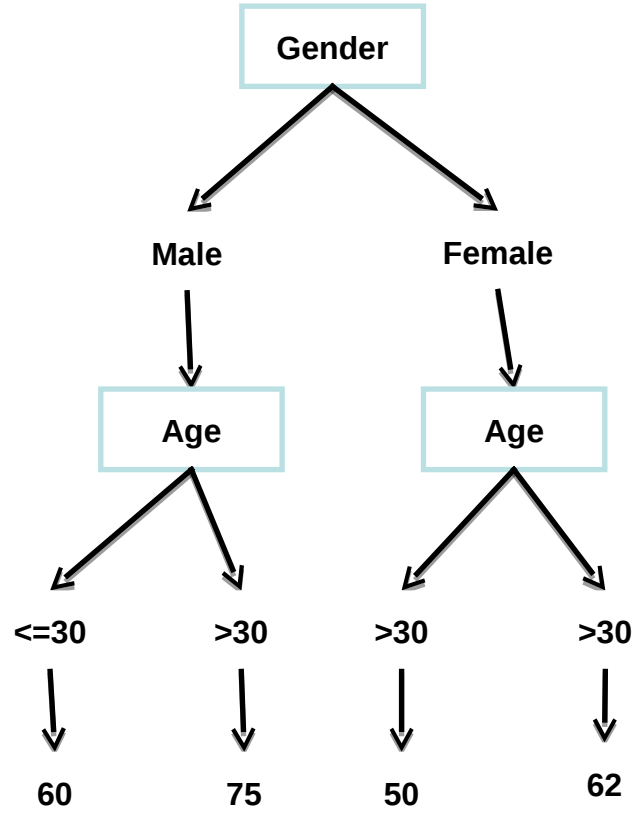
Impute missing data

- Manual imputation
  - Time consuming and infeasible
- Automatic imputation
  - Global constant (for example, N/A)
  - Mean attribute value
  - Mean value of all records belonging to the same class
  - Inference-based (for example, Bayesian or decision tree) – use data mining and machine learning to predict most likely values to impute

Automatic imputation

	Gender	Weight	Global Weight	Mean Weight	Group mean Weight
R1	M	65	65	65	65
R2	M	72	72	72	72
R3	F	54	54	54	54
R4	F	51	51	51	51
R5	M	?	N/A	64.8	73
R6	F	?	N/A	64.8	52.5
R7	M	82	82	82	82

Inference-based



Outliers and noisy data

- Random error or variance in the data
- Incorrect values and errors occur due to
  - Faulty data collection instruments
  - Data entry problems
  - Data transmission problems
  - Technology limitation
  - Misunderstanding of required data
- Depending upon application outliers are important
  - For example fraud detection or national security

Smooth noisy data

- Binning
  - Sort data and partition into equal-frequency bins
  - Smooth by bin means, bin median, bin boundaries
- Regression
  - Smooth by fitting data to regression functions
- Clustering
  - Identify outliers not belonging to clusters
- Manual inspection (active learning) of possible outliers

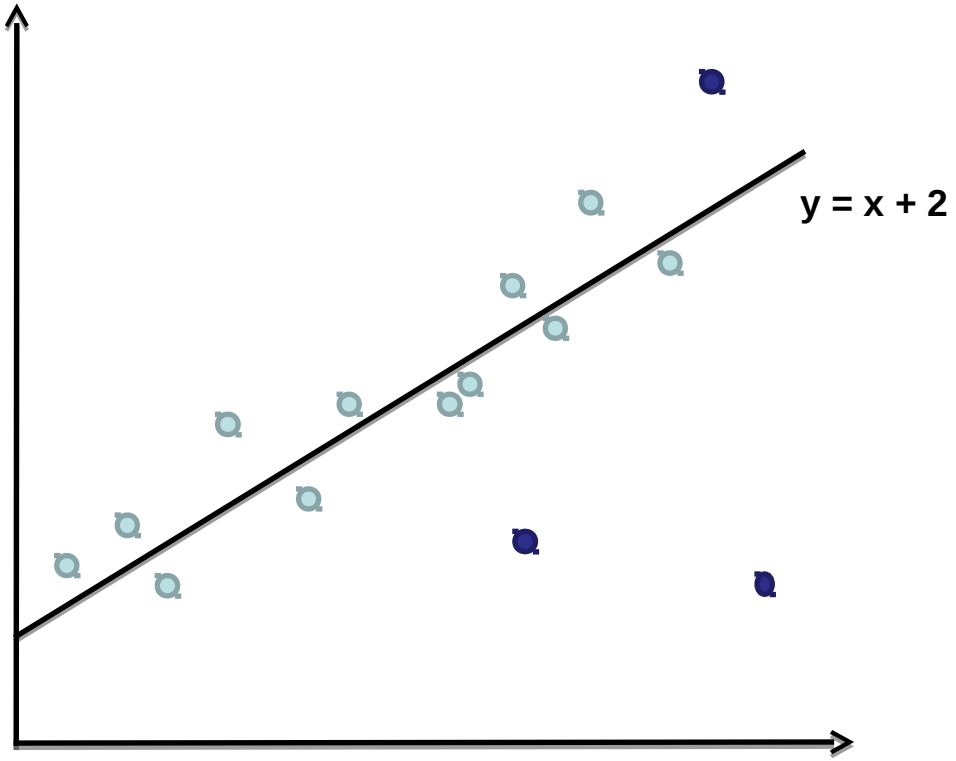
Binning (1)

- Equal-width / distance
  - Divide the range into N intervals of equal size
  - Width of intervals = (max-min)/N
  - Skewed data is not handled well
- Equal-depth / frequency
  - Divide the range into N intervals of (approximately) same number of samples
  - Suitable for skewed data distributions

Binning (2)

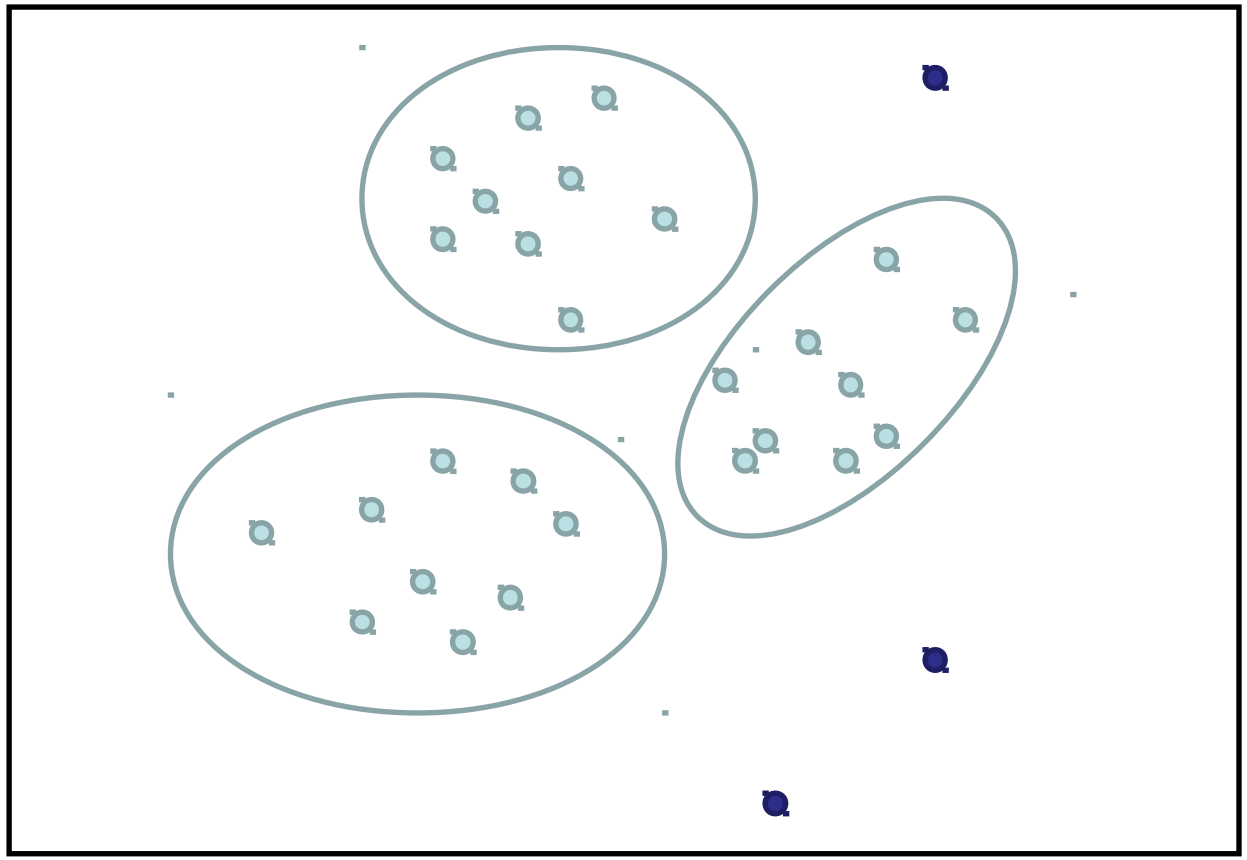
Values	5	27	100	59	28	48	50	39	9	7	20	63	10	41	9
Bins equal-frequency	5	7	9	9	10	20	27	28	39	41	48	50	59	63	100
Smooth by bin means	8	8	8	8	8	31	31	31	31	31	64	64	64	64	64
Smooth by bin medians	9	9	9	9	9	28	28	28	28	28	59	59	59	59	59
Smooth by bin boundaries	5	5	10	10	10	20	20	20	41	41	48	48	48	48	100

Regression



To be covered in more detail in the data mining course

Clustering



Data

Outliers

To be covered in more detail in the data mining course

Redundant data

- Duplicate records occur within a single data source, or when combining multiple sources
  - The same entity/object might have different values in an attribute
  - One attribute may be a derived attribute in another database
  - Attribute values of the same object entered in different time

- Redundant attributes can be identified by correlation analysis
- Redundant records can be identified by deduplication or data integration (more about this later in the course)

Identifying redundant attributes (1)

- Correlation analysis
  - Numerical attributes (A and B) using Pearson coefficient

$$r = \frac{\sum(A - A_{mean})(B - B_{mean})}{(n-1)\, A_{std}\, B_{std}}$$

  - Categorical attributes (A and B) using chi-square test

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

  - The higher the value the stronger the correlation

Identifying redundant attributes (1)

	Cancer		No cancer		Sum (row)
	Observed	Expected	Observed	Expected	
Smoking	250	90	200	360	450
Not smoking	50	210	1000	840	1050
Sum (column)	300		1200		1500

$$x^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Therefore, smoking and cancer are highly correlated

Inconsistent data

- Different formats, codes, and standards across different sources (even within a single source)
- Resolving using external reference data
  - Lookup tables
    - E.g. Sydney, NSW, 7000 -> Sydney, NSW, 2000
  - Rules
    - Male or 0 -> M
    - Female or 1 -> F

Summary

- Data cleaning is a crucial data pre-processing step
- The data cleaning cycle includes several tasks:
  - Handling missing values, smoothing noisy data, removing redundant values, and resolving inconsistencies
- Directions of future developments in data cleaning:
  - Efficient data cleaning tools for Big data, automated data cleaning, and interactive data cleaning