Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 8: Data parsing and standardisation
(Lecturer: Thilina Ranbaduge)

Lecture outline

- Data types
- Processing multi-variate attributes
  - Parsing
  - Validating
  - Correction
  - Standardisation
- Segmentation methods
- Summary

Data types (1)

- Common data types include
  - String data (such as *first name*)
  - Numerical data
    - Continuous (such as *electricity price*)
    - Discrete (such as *age*)
  - Categorical data (such as *marital status*)
  - Ordinal data (such as *movie rating*)
  - Binary data (such as *smoking*)
  - Free-text data (such as *clinical notes*)

Data types (2)

- Other (complex) data types include
  - Data / time data (such as *date of birth*)
  - Geographical data (such as *location*)
  - Web data (such as *HTML table*)
  - Image data (such as *scanned receipt*)
  - Audio data (such as *recordings or songs*)
  - Video data (such as *Youtube video*)
  - Multi-variate data (such as *address*)

Processing multi-variate attribute (1)

- Multi-variate attributes contain values of multiple elements (features) in a single attribute
  - Examples: address, name, date and time of entry, telephone number
- Several steps for pre-processing such attributes:

Processing multi-variate attribute (2)

- Aims of pre-processing such attributes:
  - Segment into well defined fields
  - Remove unwanted characters and words
  - Remove punctuations and stop words
  - Correct misspellings
  - Verify if values are possible individually and in combination
  - Standardise values
  - Expand abbreviations
  - Replace nicknames
  - Convert into consistent upper/lower case

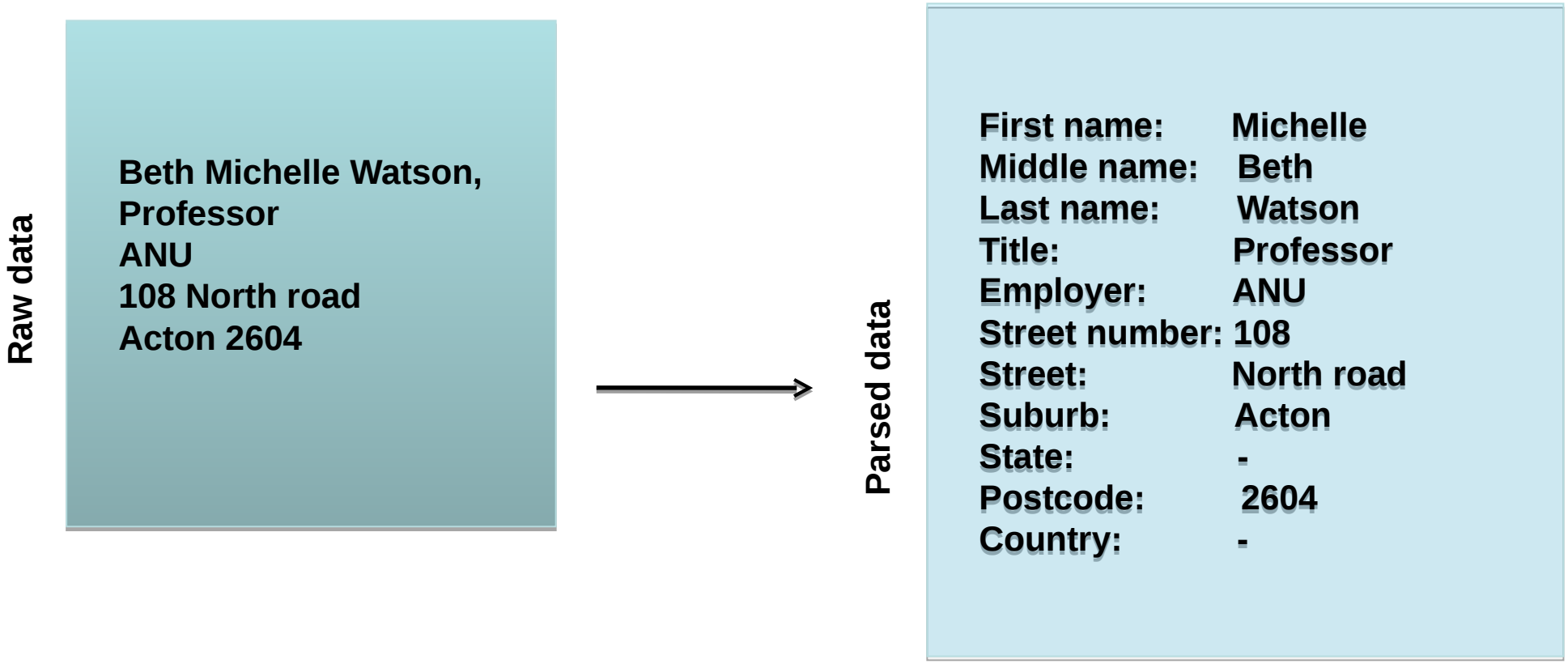
Processing multi-variate attribute (3)

- Multi-variate attributes are a special type of data
- Data parsing and standardisation is required to pre-process such attributes for improved data integration and analytics

- Example applications include:
  - Parse, standardise and validate mailing addresses of customers (for marketing)
  - Parse and standardise free-form text data elements (for example, customer reviews and opinions)
  - Standardise, validate, enhance, and enrich customer data

Data parsing

- Placement of various data elements into the appropriate fields



Data validating

- Once parsed, every field in every record needs to be audited for content
- An essential, but often overlooked step

- Identify records with no data, garbage data (punctuation signs and symbols) and inappropriate data
  - For example, Australian postcodes should contain 4 numeric characters
  - Incorrect postcode for a given suburb/town name
  - Missing state/territory and country

Data correction (1)

- Ensure that elements in the record fields are correct and sensible when related to other elements
  - For example, a suburb, state, and postcode all being not just correctly spelled and formatted in isolation, but correct and appropriate as part of a complete address
  - In the current example, suburb *Acton* and postcode *2604* are not correct together
  - Often requires external information and domain knowledge for this process of data correction (look-up tables)

Data correction (2)

Parsed data

First name:	Michelle
Middle name:	Beth
Last name:	Watson
Title:	Professor
Employer:	ANU
Street number:	108
Street:	North road
Suburb:	Acton
State:	-
Postcode:	2604
Country:	-

Corrected data

First name:	Michelle
Middle name:	Beth
Last name:	Watson
Title:	Professor
Employer:	ANU
Street number:	108
Street:	North road
Suburb:	Acton
State:	ACT
Postcode:	2601
Country:	Australia

Data standardisation (1)

- Once data have been corrected, the elements are standardised according to the criteria given
  - To further clean data by making them consistent
  - Employment of standards for elements
  - For example, *street*, *st*, and *str* are standardied as *ST*; road and *rd* are as *RD*; title P*rofessor* as *Prof*; country in 3 letters (*Australia* as *AUS*); state *Australian Capital territory* as *ACT*; or the employer name without abbreviation (*ANU* → A*ustralian National University*), etc.

Data standardisation (2)

Corrected data

First name:	Michelle
Middle name:	Beth
Last name:	Watson
Title:	Professor
Employer:	ANU
Street number:	108
Street:	North road
Suburb:	Acton
State:	ACT
Postcode:	2601
Country:	Australia

Standardised data

First name:	Michelle
Middle name:	Beth
Last name:	Watson
Title:	Prof
Employer:	Australian National University
Street number:	108
Street:	North RD
Suburb:	Acton
State:	ACT
Postcode:	2601
Country:	AUS

Segmentation methods (1)

- Parsing data requires segmenting values into separate elements
- Several methods for segmentation:
  - Rule-based: Manually developed or machine learning-based (using training data) rules for segmentation
  - Pattern matching languages: Regular expressions, for example, search for particular signatures in data for segmentation
  - Probabilistic methods: Hidden Markov models (HMM) and variations have been used for text segmentation in speech and natural language processing

Segmentation methods (2)

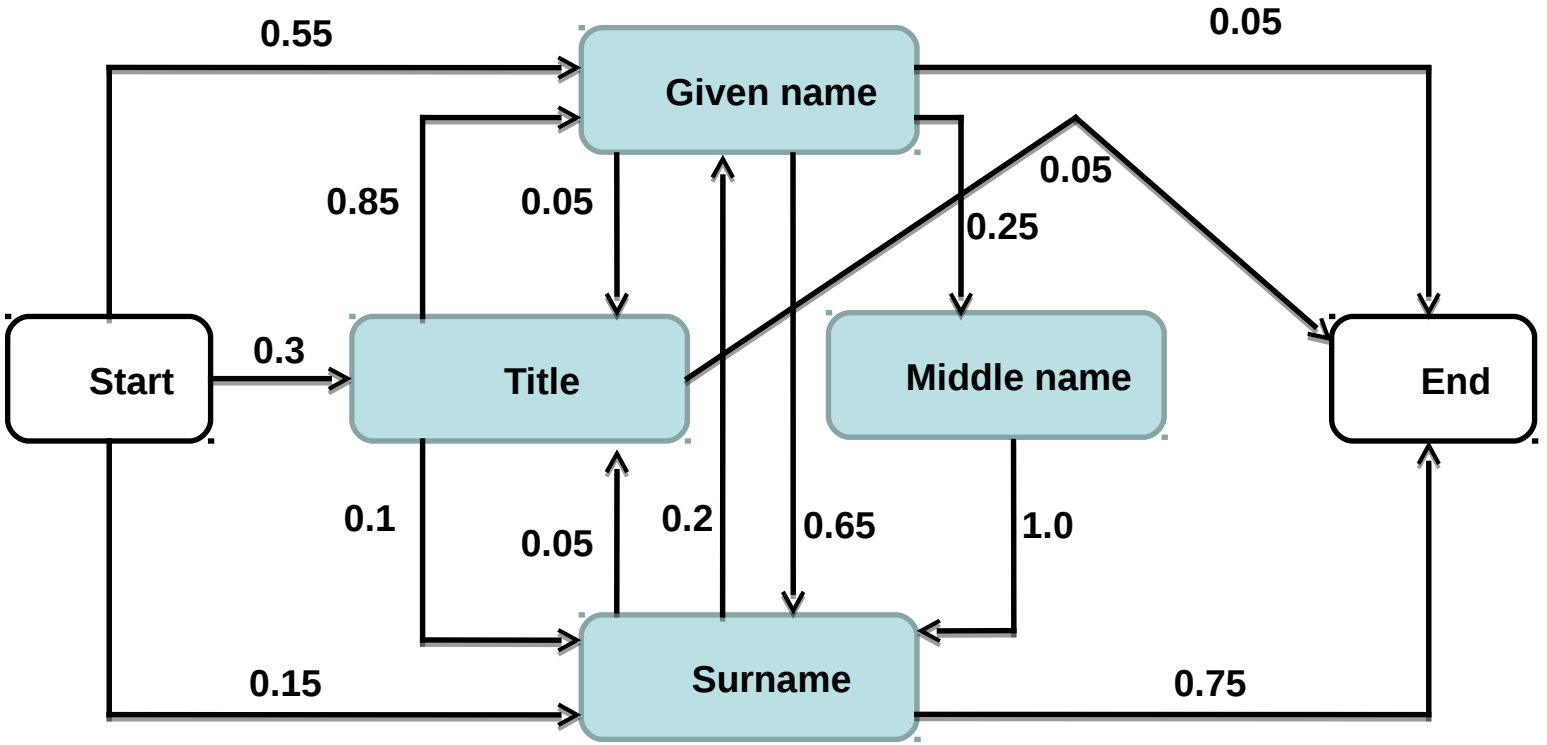
- Hidden Markov model
  - A probabilistic finite state machine that consists of
    - A set of hidden states
    - Transition edges between these states
    - A finite dictionary of discrete observation symbols
  - Transition probability: Each edge is associated with a transition probability (which sum to 1.0 for a given state)
  - Observation probability: Each state emits observation symbols from the dictionary with a certain probability

    (which also sum to 1.0 for a given state)

Segmentation methods (3)

An example HMM for names:

- Starts with a given name with probability 0.55
- Followed by surname with probability 0.65 and by a middle name with probability 0.25



Segmentation methods (4)

Transition probabilities

	To					
From	**Start**	**Title**	**GName**	**MName**	**SName**	**End**
Start	-	0.3	0.55	0.0	0.15	-
Title	-	0.0	0.85	0.0	0.1	0.05
GName	-	0.05	0.0	0.25	0.65	0.05
Mname	-	0.0	0.0	0.0	1.0	0.0
Sname	-	0.05	0.2	0.0	0.0	0.75
End	-	-	-	-	-	-

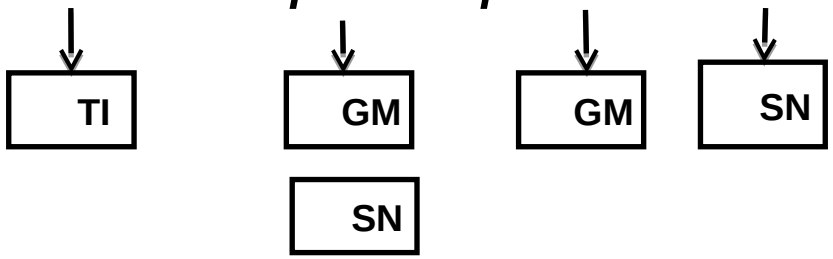
Observation probabilities

	State					
Observation	**Start**	**Title**	**GName**	**MName**	**SName**	**End**
TI	-	0.96	0.01	0.01	0.01	-
GM	-	0.01	0.35	0.33	0.15	-
GF	-	0.01	0.35	0.27	0.14	-
SN	-	0.01	0.09	0.14	0.45	-
UN	-	0.01	0.2	0.25	0.25	-

TI – title, GM – male given name, GF – female given name, SN – surname, and UN – unknown words

Segmentation methods (5)

- Example: *professor peter paul miller*



- Two possible paths through the HMM:
  - **Path 1**: Start → Title(TI) → GName(GM) → MName(GM) → SName(SN) → End
    p = 0.3 x 0.96 x 0.85 x **0.35** x 0.25 x 0.33 x 1.0 x 0.45 x 0.75 = 0.002385
  - **Path 2**: Start → Title(TI) → GName(SN) → MName(GM) → SName(SN) → End
    p = 0.3 x 0.96 x 0.85 x **0.09** x 0.25 x 0.33 x 1.0. x 0.45 x 0.75 = 0.000613
- Path 1 has the highest probability *p*, and therefore the corresponding segmentation is selected

Summary

- Pre-processing multi-variate attributes consists of several steps:
  - Data parsing (segmentation)
  - Data validation and correction
  - Data standardisation
- Crucial for personal data that often contain names and addresses, but is challenging
- Has received attention in several research areas (data cleaning, data matching, and natural language processing)