COMP20008
Elements of data processing

Semester 1 2020

Lecture 4: Data formats – cont. and pre-processing

ARFF format

- Weka is a popular open-source software package for machine learning

- It includes tools for visualizations, data analysis and predictive modelling
- Weka uses the Attribute Relation File Format (ARFF)
- ARFF files have two distinct sections. The first section is the **Header** information, which is followed the **Data** information.
- Declarations are case insensitive

ARFF header

- A dataset begins with a declaration of its name

  `@relation name`

- Followed by a list of all the attributes in the dataset of the form:

  `@attribute attribute_name specification`

- The specification may be of different types:

  `@attribute nominal_attribute {first_value, second_value}`
  `@attribute numeric_attribute`
  `@attribute string_attribute`
  `@attribute date_attribute date [<date-format>]`

ARFF data

- After this header information, the data is introduced by a tag: **`@data`**
- Lines that begin with a % are **comments**.

ARFF example


```
@relation weather

@attribute outlook { sunny, overcast, rainy }
@attribute temperature numeric
@attribute humidity numeric
@attribute windy { TRUE, FALSE }
@attribute play { yes, no }
@attribute recorded date "yyyy-MM-dd HH:mm:ss"

@data
sunny, 85, 85, FALSE, no, "2001-04-03 12:12:12"
sunny, 80, 90, TRUE, no, "2001-04-04 12:12:12"
overcast, 83, 86, FALSE, yes, "2001-04-05 12:12:12"
```

PDF format

- A format for presenting documents independently of application software / OS
- Introduced by Adobe in 1993, standardized in 2008
- Large amounts of useful data stored in PDF format (e.g. forms, invoices, etc.). Legacy data in particular.

Challenges of PDF format

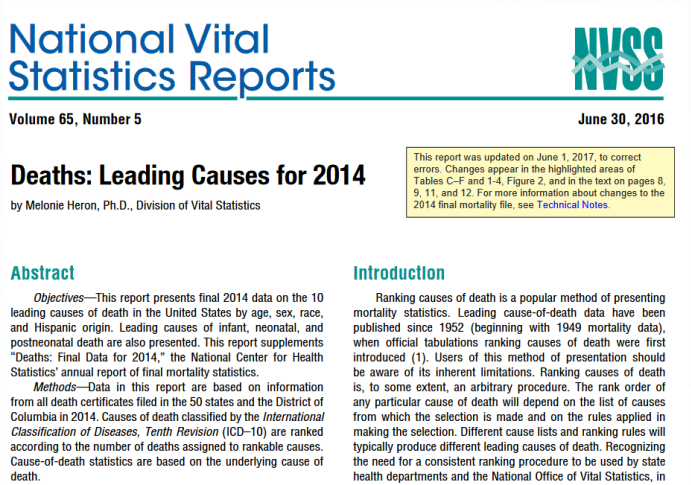
- Format is very flexible, limited consistency
- May contain different types of data; similar looking documents can be represented very differently
- Text may be stored as images, e.g., for scanned documents
- Data is unstructured
- As a result, automated extraction is difficult

- **May require substantial preprocessing to extract useable data**

Approaches to PDF data

Convert the PDF to text using a library (e.g. poppler, PDFMiner).

```
pdftotext nvsr65_05.pdf nvsr65_05.txt
```



National Vital Statistics Reports

Volume 65, Number 5 June 30, 2016

Deaths: Leading Causes for 2014

by Melonie Heron, Ph.D., Division of Vital Statistics

This report was updated on June 1, 2017, to correct errors. Changes appear in the highlighted areas of Tables C–F and 1-4, Figure 2, and in the text on pages 8, 9, 11, and 12. For more information about changes to the 2014 final mortality file, see Technical Notes.

Abstract

Objectives—This report presents final 2014 data on the 10 leading causes of death in the United States by age, sex, race, and Hispanic origin. Leading causes of infant, neonatal, and postneonatal death are also presented. This report supplements "Deaths: Final Data for 2014," the National Center for Health Statistics' annual report of final mortality statistics.

Methods—Data in this report are based on information from all death certificates filed in the 50 states and the District of Columbia in 2014. Causes of death classified by the *International Classification of Diseases, Tenth Revision* (ICD–10) are ranked according to the number of deaths assigned to rankable causes. Cause-of-death statistics are based on the underlying cause of death.

Introduction

Ranking causes of death is a popular method of presenting mortality statistics. Leading cause-of-death data have been published since 1952 (beginning with 1949 mortality data), when official tabulations ranking causes of death were first introduced (1). Users of this method of presentation should be aware of its inherent limitations. Ranking causes of death is, to some extent, an arbitrary procedure. The rank order of any particular cause of death will depend on the list of causes from which the selection is made and on the rules applied in making the selection. Different cause lists and ranking rules will typically produce different leading causes of death. Recognizing the need for a consistent ranking procedure to be used by state health departments and the National Office of Vital Statistics, in

National Vital
Statistics Reports
Volume 65, Number 5

June 30, 2016

Deaths: Leading Causes for 2014

by Melonie Heron, Ph.D., Division of Vital Statistics

Abstract
.
.
.

Approaches to PDF data

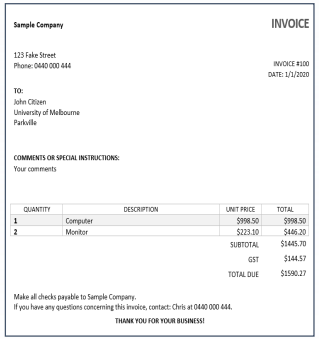
- Problem? Lose layout and formatting information
- Alternative: Convert PDF to HTML or XML using a tool like **`pdftohtml`**

- Example: http://pdftohtml.sourceforge.net/

- Retains layout and formatting, but resultant file not easy to parse

Approaches to PDF data

- Need to use Optical Character Recognition (OCR) software to process images into text



Sample Company **INVOICE**

123 Fake Street
Phone: 0440 000 444

INVOICE #100
DATE: 1/1/2020

TO:
John Citizen
University of Melbourne
Parkville

COMMENTS OR SPECIAL INSTRUCTIONS:
Your comments

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
1	Computer	$998.50	$998.50
2	Monitor	$223.10	$446.20
		SUBTOTAL	$1445.70
		GST	$144.57
		TOTAL DUE	$1590.27

Make all checks payable to Sample Company.
If you have any questions concerning this invoice, contact: Chris at 0440 000 444.

THANK YOU FOR YOUR BUSINESS!

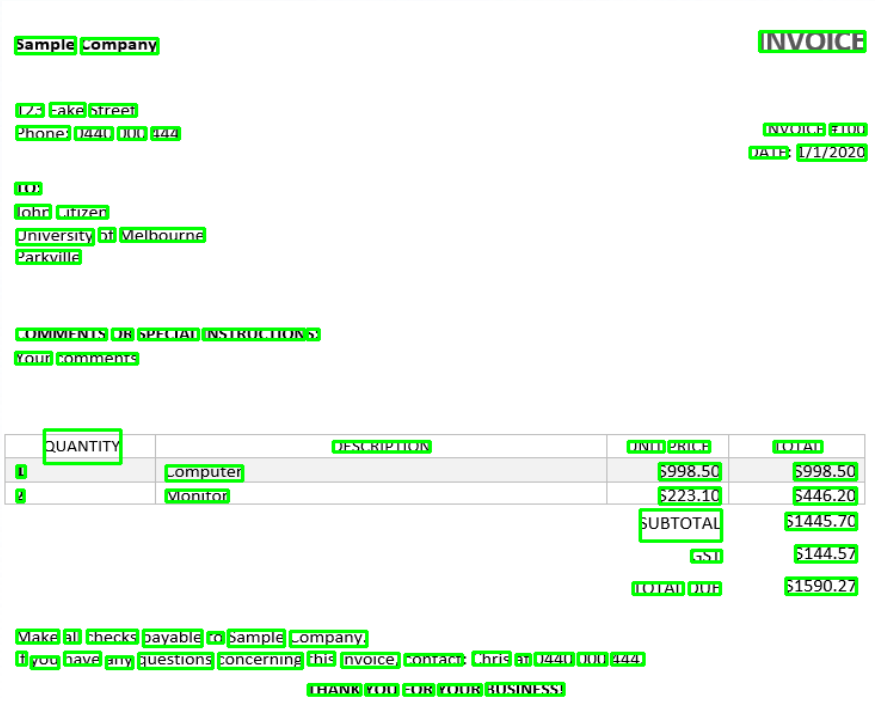
Sample Company INVOICE
123 Fake Street
Phone: 0440 000 444 INVOICE #100
DATE: 1/1/2020
To:
John Citizen
University of Melbourne
Parkville
COMMENTS OR SPECIAL INSTRUCTIONS:
Your comments

QUANTITY DESCRIPTION UNIT PRICE TOTAL
1 Computer $998.50 $998.50
2 Monitor $223.10 $446.20
SUBTOTAL $1445.70
GST $144.57
TOTAL DUE $1590.27
Make all checks payable to Sample Company.
If you have any questions concerning this invoice, contact: Chris at 0440 000 444,
THANK YOU FOR YOUR BUSINESS!

OCR: Optical Character Recognition

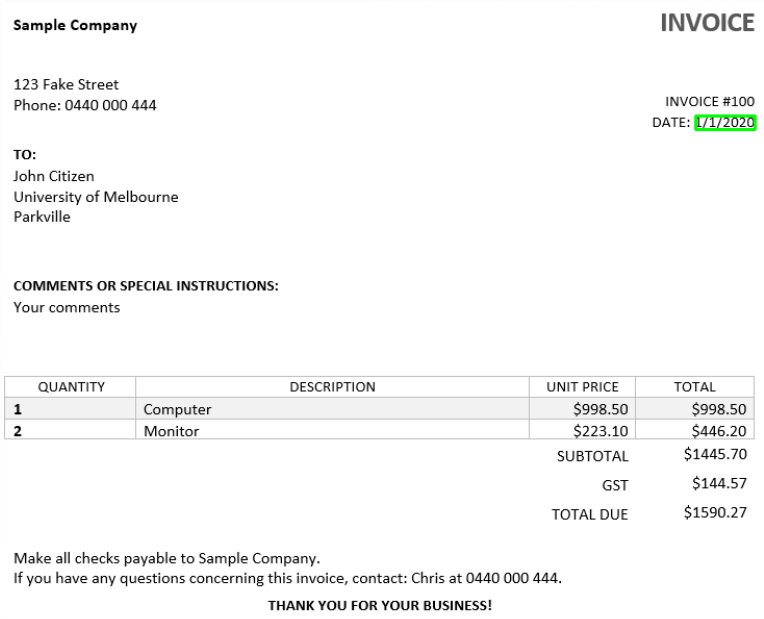
- The roots of modern OCR technology date back over 100 years, but OCR remains an active research area
- Modern OCR implementations rely on machine learning
- Many OCR packages available:
  - Tesseract (first developed by HP in the 1980s, currently open source and sponsored by Google). A python wrapper, `pytesseract` is also available
  - Many commercial packages (e.g. Adobe Acrobat, ABBYY)

Approaches to PDF data



Can identify the location of text on PDFs

Approaches to PDF data



Sample Company

INVOICE

123 Fake Street
Phone: 0440 000 444

INVOICE #100
DATE: 1/1/2020

TO:
John Citizen
University of Melbourne
Parkville

COMMENTS OR SPECIAL INSTRUCTIONS:
Your comments

QUANTITY	DESCRIPTION	UNIT PRICE	TOTAL
1	Computer	$998.50	$998.50
2	Monitor	$223.10	$446.20
		SUBTOTAL	$1445.70
		GST	$144.57
		TOTAL DUE	$1590.27

Make all checks payable to Sample Company.
If you have any questions concerning this invoice, contact: Chris at 0440 000 444.

THANK YOU FOR YOUR BUSINESS!

And identify important data by location and/or pattern matching:

Why is pre-processing needed?

THE UNIVERSITY OF MELBOURNE

Name	Age	Date of Birth
“Henry”	20.2	20 years ago
Katherine	Forty-one	20/11/66
Michelle	37	5/20/79
Oscar@!!	“5”	13th Feb. 2019
-	42	-
Mike___Moore	669	-
巴拉克奥巴马	52	1961年8月4日

Data Pre-processing

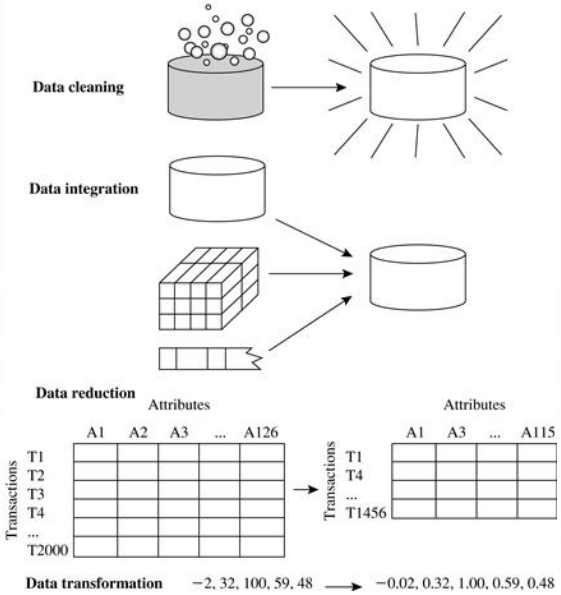
Why is pre-processing needed?

Measuring data quality

- Accuracy
  - Correct or wrong, accurate or not
- Completeness
  - Not recorded, unavailable
- Consistency
  - E.g. discrepancies in representation
- Timeliness
  - Updated in a timely way
- Believability
  - Do I trust the data is true?
- Interpretability
  - How easily can I understand the data?

	Date of Birth
1	20 years ago
2	20/11/66
3	5/20/79
4	13th Feb. 2019
5	-
6	-
7	1961年8月4日

Major data preprocessing activities



Data mining concepts and techniques, Han et al 2012

Noisy data

- Truncated fields (exceeded 80 characters limit)
- Text incorrectly split across cells (e.g. separator issues)
- Salary=“-5”
- Some causes
  - Imprecise instruments
  - Data entry issues
  - Data transmission issues

Inconsistent data

- Different naming representations (“Melbourne University” versus “University of Melbourne”) or (“three” versus “3”)
- Different date formats (“3/4/2016” versus “3rd April 2016”)
- Age=20, Birthdate=“1/1/1971”
- Two students with the same student id
- Outliers (e.g. 62,72,75,75,78,80,82,84,86,87,87,89,89,90,999)
  - No good if it is list of ages of hospital patients
  - Might be ok for a listing of people number of contacts on Linkedin though
  - Can use automated techniques, but also need domain knowledge

Disguised missing data

- Everyone's birthday is January 1st?
- Email address is xx@xx.com
- Adriaans and Zantige
  - *"Recently, a colleague rented a car in the USA. Since he was Dutch, his post-code did not fit the fields of the computer program. The car hire representative suggested that she use the zip code of the rental office instead."*
- How to handle
  - Look for "unusual" or suspicious values in the dataset, using knowledge about the domain

Missing or incomplete data

- Lacking feature values
  - Name=“”
  - Age=null
- Some types of missing data (Rubin 1976)
  - Missing completely at random: Data are missing independently of observed and unobserved data; e.g. coin flipping to decide whether or not to answer an exam question.
  - Missing not completely at random

  I create a dataset by surveying the class about how healthy they feel. What is the meaning of missing values for those who don’t respond?

Missing completely at random

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Missing data are **MCAR** when the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself.

Missing not at random

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Data are **MNAR** when the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.

For example, when data are missing on IQ and only the people with low IQ values have missing observations for this variable.

Causes of missing data

- Malfunction of equipment (e.g. sensors)
- Not recorded due to misunderstanding
- May not be considered important at time of entry
- Deliberate

Data cleaning – process

- Many tools exist (Google Refine, Kettle, Talend, …)
  - Data scrubbing
  - Data discrepancy detection
  - Data auditing
  - ETL (Extract Transform Load) tools: users specify transformations via a graphical interface
- Our emphasis will be to understand some of the methods employed by typical tools
- Domain knowledge is important

A practical example

https://www.youtube.com/watch?v=B70J_H_zAWM&feature=emb_logo