FIT1043 Lecture 12 Introduction to Data Science

Mahsa Salehi*

Faculty of Information Technology, Monash University

Semester 2, 2022

*with material from © Wray Buntine

Reminders

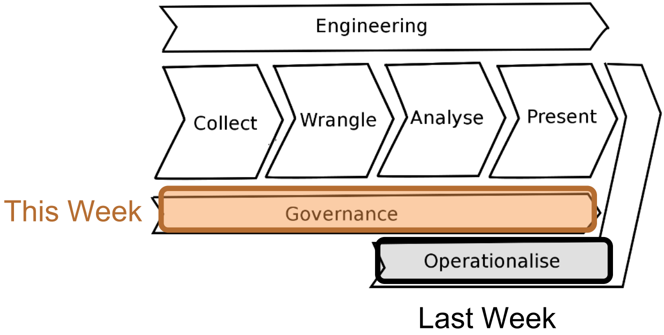
- SETU time: see SETU Unit Evaluation link in Moodle

- Final assignment due *Friday 21st October 2022, 11:55pm*
  - Use tutorials 9-10: shell/R commands to answer the questions
  - Copy/Paste your commands in your document
  - Do not include the questions into your assignments

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly quizzes
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Industry guest lecture	
12	Issues in data management/ Unit Review	Assignment 3

Our Standard Value Chain



Why Manage Data?

- The data is very valuable, data collection is usually time consuming and hard
- Large amount of data and documents are being generated with high growth rate
- Multiple sources of data (general business documents, ERP systems etc)

What is Data Management?

Data management is the development, execution and supervision of plans, policies, programs and practices that *control, protect, deliver and enhance the value of data and information assets.*

Data management is important

See “How to avoid a data management nightmare”, a video created by NYU Health Sciences Library (Youtube)

Data management in research

- Information privacy where human are involved is important

- See “Managing Research Data” from Digital Curation Centre in the UK

Issues in Data Management

overview of issues

Data management plan in an organisation

Deals with issues:

- integration and data warehousing [See Data Warehousing- an Overview (Youtube, 2:44-5:33)]
- replication and persistence
- standardising the vocabulary used across the organisation, e.g., job titles
- security

Privacy versus Confidentiality

- Privacy is (for our purposes) having control over how one shares oneself with others.
  - e.g. closing the blinds in your living room
- Confidentiality is information privacy, how information about an individual is treated and shared.
  - e.g. excluding others from viewing your search terms or browse history
- Security as the protection of data, preventing it from being improperly used
  - e.g. preventing hackers from stealing credit card data

Social media and the loss of confidentiality

- See: “The curly fry conundrum: Why social media ’likes’ say more than you might think” by Jennifer Golbeck (TED)
- Target is predicting which women is pregnant from their purchases
- many things can be predicted from Facebook “likes”
- Implicit data ::= data not explicitly stored but inferred with reasonable precision from available data

Confidentiality, cont.

- See: "*Empower consumers to control their privacy* in the Internet of Everything" by Carla Rudder (blog)

- For many apps or services, you must either accept their data sharing policies or you can't use their services fully

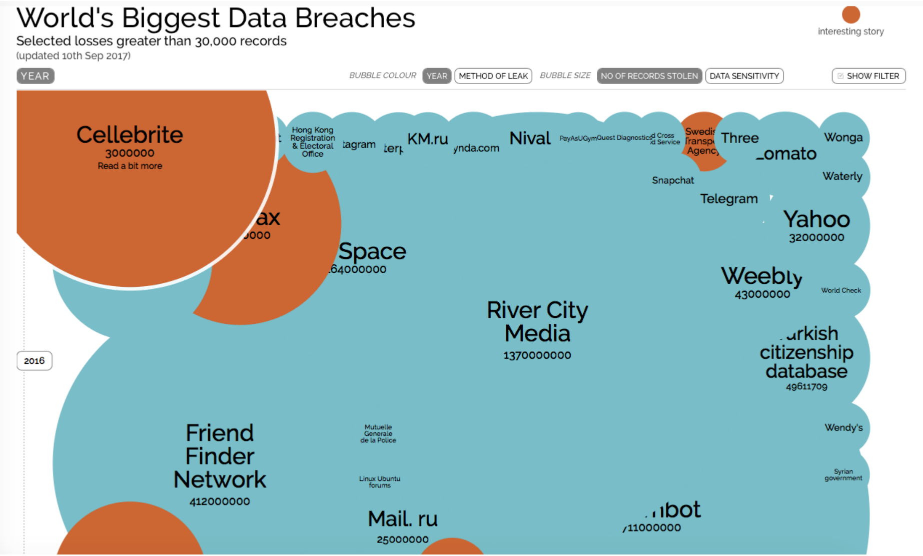
- There could be an agent to interact in a narrative form with individual consumers:
  - For instance the app might ask: 'Are you willing to share your health data with company X?'

Compliances and Regulations

- Ethics: the moral handling of data
- There should be regulations in place to ensure that confidentiality is protected
- The process of ensuring you meet regulations is called compliance

Compliances and Regulations Example

- PCI (Payment Card Industry) standard
  - Aims to reduce credit card fraud
    - By placing specific regulations, e.g., credit card information should be stored only in encrypted format.
  - Companies who handle credit card have to comply with PCI standards
  - Audit (validation of compliance) is done annually



World's Biggest Data Breaches

Selected losses greater than 30,000 records

(updated 10th Sep 2017)

Here is the link.

Data Management and Data Science

- medical informatics: for predicting fungal infections from nursing notes, the team needs to abide by confidentiality and security

- internet advertising: what implicit and explicit data is stored about a user

Tutorial Week 12

In week 12 tutorial we investigate issues related to security and privacy of data.

- Legal requirements for companies dealing with sensitive user data.
- Example of private data (ENRON email corpus)
  - Very easy (with a couple of shell commands) to discover very sensitive information (mobile phone numbers, credit card information, etc.)
- Famous information leaks
  - Some very scary leaks ....
- Example website privacy policies:
  - What information is Google storing about you?
  - Why are they keeping that information?

Unit Review and Exam Information

The Exam

- Content of the Exam
  - What is examinable?
- Format of the Exam
  - What will the exam look like?

Content of the Exam

- Everything discussed in the lectures is examinable.

- Python, R, Unix Shell are examinable
  - **but** you do not need to memorise all the syntax!

- Content on Ed lesson provides a useful description of the content of the course
  - is not directly examinable, except where in slides

- Content of the tutorials explains concepts from the slides

Format of the Exam

What will the exam look like?

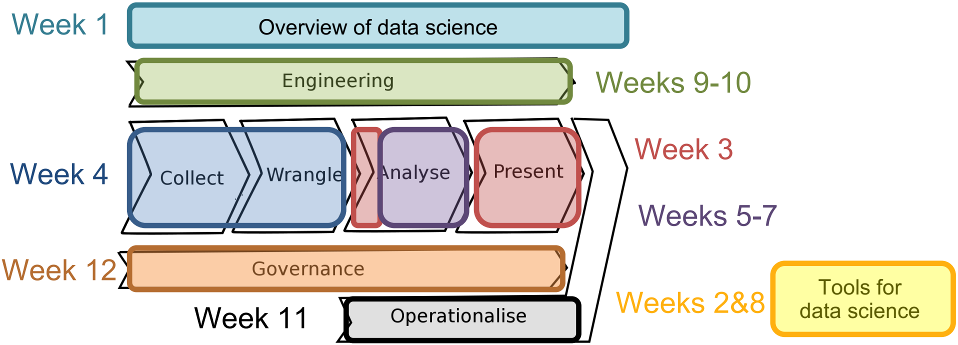
- Your Exam will be in an electronic format (i.e., e-Exam)
- Exam consists of two parts:
    - 15 ultiple-choice questions (1 mark each)
    - 25 short-answer questions (2 marks each)
- Exam duration: 2 and 10 minutes
- Close book
- No need to bring a calculator
- Sample questions available on Moodle, FLUX quesitons...

Unit

So, what did we cover in this unit?

- Quick overview of what we learnt

Our Standard Value Chain- Unit content



Week 1

- What is data science and
- Drew Conway’s Venn diagram
- Usefulness of machine learning
- Different components of a data science process
- Differentiate data science from other related disciplines

Week 2

- Essentials for coding in Python for data science
- Interpret given Python codes
- Why we study Python and its importance for data science

Week 3

- The importance/power of data visualization
- Approaches for data visualisation,
  - explain where each approach is appropriate to be used
- Concepts in descriptive statistics
- More sophisticated group-by operations in Python

Week 4

- Open data and linked open data
- How to access to new data sources through APIs
- How different APIs work
- Data quality problems in datasets
- Data wrangling commands in Python

Week 5

- What are models and predictive models
- Analyse predictive models in different examples
- How to evaluate predictive models
- How to estimate linear regression model
- linear regression and polynomial regression in Python

Week 6

- Overfitting and underfitting of different models
- Bias and variance trade-off
- “No Free Lunch Theorem”
- What are ensemble models

Week 7

- Differentiate between classification and regression models
- How decision trees and regression trees work
- How random forest works
- How k-means clustering works
- Confusion matrix and prediction accuracy
- Different classification metrics

Week 8

- Essentials for coding in R for data science
- Explain and interpret given R commands
- Apply R commands for data wrangling, visualisation, exploration and analysis

Week 9

- Characterising big data:
  - Volume, Velocity, Variety, Veracity
- What is metadata?
  - different types of metadata
- Growth laws related to big data:
  - Moore's law, Koomey's law, Bell's Law and Zimmerman's Law
- Introduction to Unix Shell commands for data science

Week 10

- Processing big data
  - different types of databases (SQL, graph, noSQL, etc.)
  - different types of processing (interactive, streaming, batch)
  - distributed processing (map-reduce, hadoop, spark, etc.)
- What is deep learning

Week 11

- Guest lecture from Microsoft: Data and Artificial Intelligence- An Industry Perspective

Microsoft

Pursuit Lead- Data & AI

Microsoft Services

Week 12

- Confidentiality and privacy
- Regulatory compliance
- Data management

- Tutorial:
  - Understanding Privacy, Legal Requirements and the Prevention of Information Leaks

- Phew! We’ve covered a lot of stuff in this unit!

Sample Exam

- A sample exam is available on Moodle

- Consultations will continue next week (Pre-exam consultations)

THE END

- I hope you’ve enjoyed the unit
- Do consider follow-on units, where you’ll learn the full stuff:
  - FIT2079 Data visualisation
  - FIT2086 Modelling for data analysis
  - FIT3152 Data analytics
  - FIT3181 Deep learning
  - more 3rd year units ...
- Best of luck for your revision and the exam!

"Our data analysis experts can't read your minds. You're going to reach your own decision regarding hiring us, even though it's the same decision we knew in advance you'd make."

"Sweetheart, my neural net predicts that you and I are 98.9% compatible. Will you be my Valentine?"