FIT1043 Introduction to Data Science

# Module 6: Data Curation and Management

Lecture 12

Monash University

# Reminders

- SETU time: see SETU Unit Evaluation link in Moodle
- Reminders:
  - Final assignment due this Sunday!
  - if you haven't tried to download the data yet ....
  - no tutorial this week

# Discussion: Privacy and Security

In last week's tutorial we investigated 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?
  - What control do they provide you with over the information they collect.

# Unit Schedule: This Week

| Module | Week | Content |
|---|---|---|
| **1.** | 1 | overview and look at projects |
| | 2 | (job) roles, and the impact |
| **2.** | 3 | data business models |
| | 4 | application areas and case studies |
| **3.** | 5 | characterising data and "big" data |
| | 6 | data sources and case studies |
| **4.** | 7 | resources and standards |
| | 8 | resources case studies |
| **5.** | 9 | data analysis theory |
| | 10 | data analysis process |
| **6.** | 11 | issues in data management |
| | 12 | **data management frameworks** |

# Unit Schedule: This Week

| Module | Week | Content |
|---|---|---|
| **1.** | 1 | overview and look at projects |
| | 2 | (job) roles, and the impact |
| **2.** | 3 | data business models |
| | 4 | application areas and case studies |
| **3.** | 5 | characterising data and "big" data |
| | 6 | data sources and case studies |
| **4.** | 7 | resources and standards |
| | 8 | resources case studies |
| **5.** | 9 | data analysis theory |
| | 10 | data analysis process |
| **6.** | 11 | issues in data management |
| | 12 | **EXAM INFO** |

# The Exam

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

# Content of the Exam

- Everything discussed in the lectures is examinable.
- That includes the "Brief Introduction to ..." slides:
  - only a few questions here
  - on Python, R, SAS, Unix Shell, Decision Trees
  - **but** you do not need to memorise all the syntax!
- Content linked from lecture slides is not **directly** examinable

  i.e. you **do not** need to learn everything that is linked too
  - **but** sometimes the definitions/explanations of the content discussed in the lectures *is given in the linked content*,
- Content on Alexandria 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 paper look like?

- Exam consists of two parts:
  - 50 multiple-choice questions (worth 50% of total mark)
  - 25 short-answer questions (worth 50% of total mark)
- Duration 3 hours
- Open book
- No need to bring a calculator
- Sample questions available on Moodle ...

# Unit

So, what did we cover in this unit?

- Quick overview of what we learnt

# Week 1

- What is data science?
- What is machine learning?
- What is big data?
- Data science process and data science value chain

# Week 2

- What does a data scientist do?
- What skills do they need?
- Impact data science is having
  - cloud services, effect on science, social good
- Tutorial
  - Investigated Motion charts as a data visualisation tool
  - Jobs in data science

# Week 3

- Data business models
- Analytics levels: Descriptive, Predictive and Prescriptive Analytics
- Modeling decision problems with Influence Diagrams
- Data business models:
  - information brokering services
  - information-based differentiation services
  - information-based delivery network services
  - data providers
- Introduction to Python for data science
- Tutorial
  - Modeling with influence diagrams
  - Getting familiar with Python

# Week 4

- Data science case studies
- Characterising them in terms of:
  - data sources
  - data volume, velocity, variety, veracity
  - software, analytics, processing
  - security, privacy
- Tutorial:
  - Visual analytics with SAS

# Week 5

- Characterising big data:
  - Volume, Velocity, Variety, Veracity, Variability, Visualisation, Value
- 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 R for data science
- Tutorial:
  - Exploratory analysis of big data in R

# Week 6

- Processing big data
  - different types of databases (SQL, semi-structured, graph, noSQL, etc.)
  - different types of processing (interactive, streaming, batch)
  - distributed processing (map-reduce, spark, etc.)
- Introduction to Unix Shell commands for data science
- Tutorial:
  - Manipulating large files in the shell
  - Understanding map-reduce

# Week 7

- Resources and the use of big data
- What is open data?
- What is data wrangling?
- Standards for publishing data and models
- Tutorial:
  - Wrangling with SAS, DataWrangler and Python

# Week 8

- Common tools used (Hadoop and related Apache tools)
- APIs and Software-as-a-Service
- Case studies
- Tutorial:
  - Wrangling big text data (from Twitter) using shell commands

# Week 9

- Types of data analysis:
  - prediction, prediction with unknown variables, clustering, forecasting, etc.
- Learning theory
  - error vs loss functions
  - linear and polynomial regression
  - overfitting due to overly complicated model / insufficient data
  - training and test split
  - signal to noise
  - ensembling multiple models
- Tutorial:
  - understanding learning theory though examples in Python

# Week 10

- Correlation vs Causation and the need for controlled experiments
- Imputing missing values
- Examples of analytic software
- Case studies
- Introduction to Decision/Regression trees
- Tutorial:
  - building predictive models with BigML

# Week 11

- Ethics and privacy
- Regulatory compliance
- What is Data Governance
- Data Management case studies
- Tutorial:
  - Understanding Privacy, Legal Requirements and the Prevention of Information Leaks

# Week 12

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

# THE END

- I hope you've enjoyed the unit
- No tutorial this week
- Do consider follow-on units, where you'll learn the full stuff:
  - FIT2079 Data visualisation
  - FIT2086 Modelling data
  - FIT3152 Data analytics
  - more 3rd year units ...

- Best of luck for your revision and the exam!