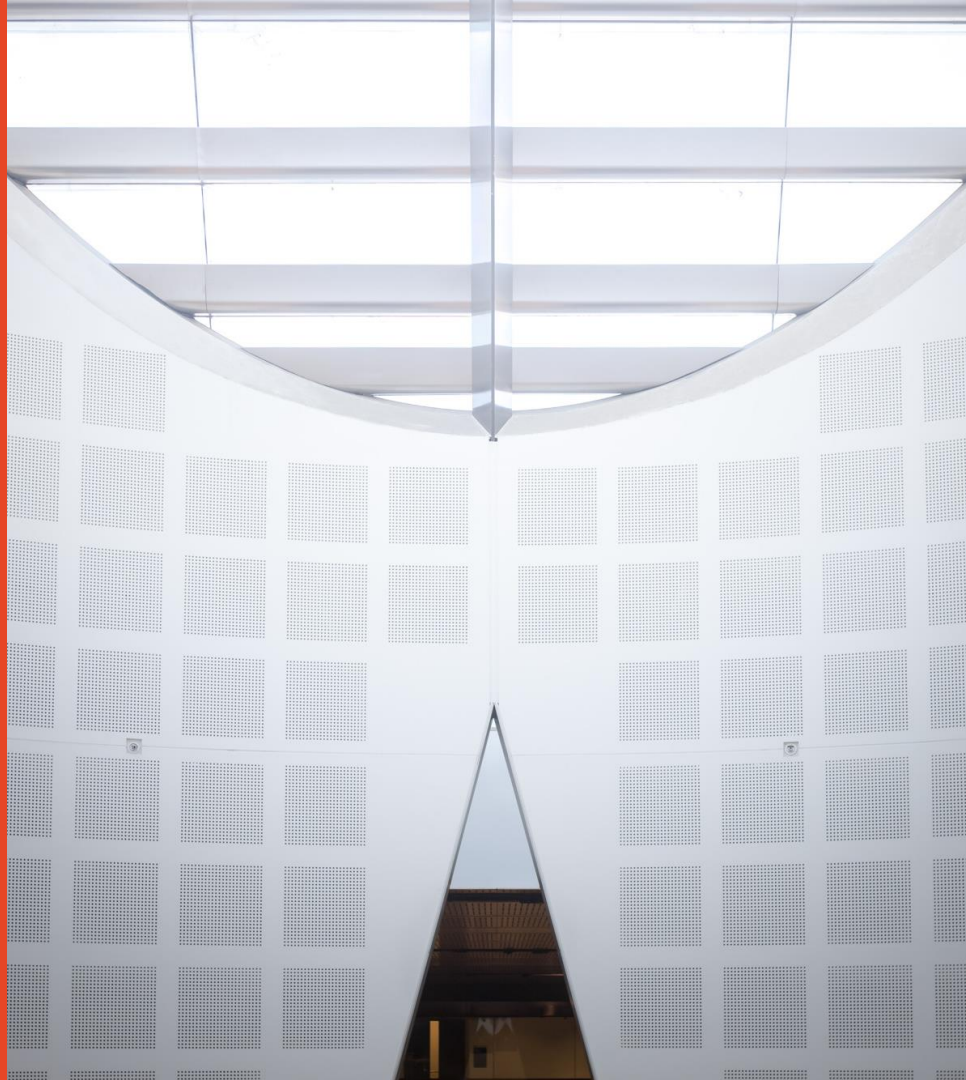
COMP5310: Principles of Data Science

W1: Introduction

Presented by
Dr Ali Anaissi
School of Information Technologies

THE UNIVERSITY OF SYDNEY



Curriculum at a glance

Whirlwind tour of:

- Data Exploration
- Data Engineering
- Data Mining & Machine Learning
- Making Decisions from Data

Focus on key activities of a data scientist

Perspectives and communication

Diverse cohort in this unit with:

- Honours degrees in non-quantitative disciplines
- Bachelors degrees in quantitative disciplines or IT
- Years of experience in industry

Doing data science requires

- Understanding application domain
- Learning, collaborating, communicating
- Product thinking

Chance to build key soft skills as well as technical skills

Questions and suggestions

We are very excited to be teaching this for the fourth year

Thank you for joining us!

Please feel free to:

- Ask questions (we should know the answer or someone who does)
- Share thoughts and suggestions on how we can improve

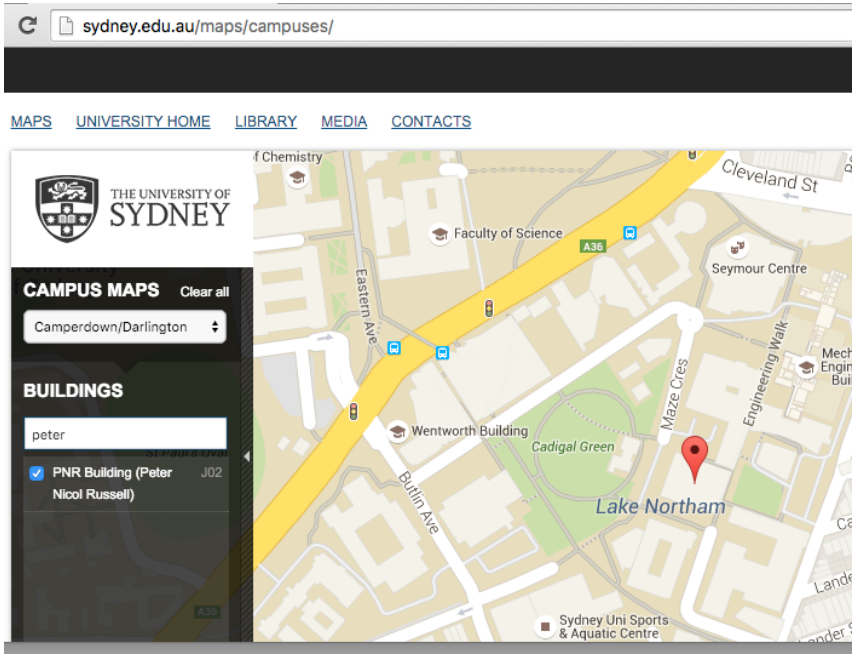
Questions about the MDS degree program or enrolments?

- **Keiko Narushima** (MDS admin officer), SIT Building, room 2E-229
- phone: 02 8627 0872 email: keiko.narushima@sydney.edu.au

UNIT ARRANGEMENTS

THE UNIVERSITY OF SYDNEY

Here you are



PNR (Building J02)

- Mixed lecture/lab
- Tuesdays, 6-9pm
- PNR Learning Studio 310 &311

Introducing Team

Lecturer

Dr Ali Anaissi

Unit Coordinator

Dr Ali Anaissi

SIT Building J12, Level 2
ali.anaissi@sydney.edu.au

Tutors

Seid Miad Zandavi

Omid Tavallaie

Raghavendra Chalapathy

Reza Behi

Mohsen Eskandari

Hossein Moeinzadeh

Hazel He

Mahdi Saki

Seyed Hashem Davarpanah

Usman Naseem

Resources

You will need a computer for exercises

- Please bring a laptop, else use a machine here

Google Sheets for spreadsheet exercises [week 2]

- Please create a Google account if you don't already have one!

Jupyter Hub accounts for Python/SQL exercises

- We will provide account details in week 3
- But we recommend you download Anaconda and PostgreSQL database on your PC

Textbooks and readings

Data Science from Scratch. Grus. O'Reilly Media. 2015.

- Available electronically through library.

Doing Data Science. O'Neill and Schutt. O'Reilly Media. 2015.

- Available electronically through library.

Learn Python and SQL with Grok

- Exercises will use Python from week 3
- We provide self-guided Python learning through Grok
- Grok learning modules are available now in Canvas under Assignments folder
- Please complete (sooner is better, week 5 at latest)

Find everything on Canvas

- The web site for this unit is on Canvas
- Use it to access contacts, schedule, readings, slides, etc
- Participate in Q&A with instructors and classmates

https://canvas.sydney.edu.au

ASSESSMENTS

Assessment

- 10%: Participation
- 10%: Project stage 1
- 20%: Project stage 2
- 5%: Project stage 3
- 55%: Final exam

Participation

Objective

Ensure everybody is keeping up.

Requirements

Submit code at end of each exercise

Complete Grok exercises (not marked)

Output

Code/spreadsheets from exercises

Marking

10% of overall mark

Project stage 1: Explore, Clean, Pitch

Objective

Explore a data set and define a research question based on research/business requirement.

Activities

Choose a data set

Explore, summarise and prepare data

Define problem, specify requirements

Output

2-page report summarizing problem analysis and proposal (plus code)

Marking

10% of overall mark (report and code)

Project stage 2 and 3: Experiment, Quantify, Report

Objective

Define an experimental framework and complete analysis/visualisation, data mining, machine learning, etc.

Activities

Define experimental framework

Perform analysis or build tool

Describe evaluation and conclusions

Output

4-page report describing framework, analysis and conclusions (plus code)

Presentation (2-3/3-4 mins)

Marking

25% of overall mark

- 20% report and code
- 5% presentation

Final exam

Objective

Assess understanding of unit material, ability to frame data problems scientifically and critical thinking about claims made based on data

Activities

Answer questions about lecture materials

Practical excises and SQL queries

Describe an approach to answering a question with data

Critique a claim made based on data

Format

Written examination

Must get 40% on exam to pass unit per SIT policy

Marking

55% of overall mark

cap on final mark which cannot exceed exam mark by more than 10 marks

Lecture plan

- W1: Introductions and housekeeping
- W2: Data exploration (spreadsheets)
- W3: Data exploration (Python)
- W4: Cleaning and storing data
- W5: Querying and summarising data
- W6: Hypothesis testing
  Project stage 1 due
- W7: Data Mining - Association Rules and Dimensionality Reduction
- W8: Data Mining - Clustering
- W9: Machine Learning – Regression
- W10: Machine Learning – Classification
- W11: Unstructured Data
- W12: Information, actionable knowledge from data, and link to effective decision making.

  Project stage 2 and 3 due
- W13: Review
- ***Exam***

LATENESS AND PLAGIARISM

THE UNIVERSITY OF SYDNEY

Recipe for success

- Attend scheduled classes except for illness, emergency, etc
- Plan 6-9 hours per week for preparation, practice, project, etc
- Participate in classes and forums with respect and humility
- Submit assessments on time

- Let us know if any concerns, e.g., if you are falling behind

Special consideration (University policy)

- If your performance on assessments is affected by illness or misadventure
- Follow proper bureaucratic procedures
  - Have professional practitioner sign special USyd form
  - Submit application for special consideration online, upload scans
  - Note you have only a quite short deadline for applying
  - http://sydney.edu.au/current_students/special_consideration/
- Notify us by email *as soon as anything begins to go wrong*
- There is a similar process if you need special arrangements for religious observance, military service, representative sports, etc

Penalty for lateness

- If you have not been granted special consideration
  - Penalty is 5% of awarded marks per day
  - Maximum 10 days late, then 0 points
- Examples:
  - Work would have scored 60% and is 1 hour late: 57%
  - Work would have scored 70% and is 28 hours late: 63%
- Recommendation: submit early; submit often

Academic integrity (University policy)

"The University of Sydney is unequivocally opposed to, and intolerant of, plagiarism and academic dishonesty.

Academic dishonesty means seeking to obtain or obtaining academic advantage for oneself or for others (including in the assessment or publication of work) by dishonest or unfair means.

Plagiarism means presenting another person's work as one's own work by presenting, copying or reproducing it without appropriate acknowledgement of the source."

http://sydney.edu.au/elearning/student/EI/index.shtml

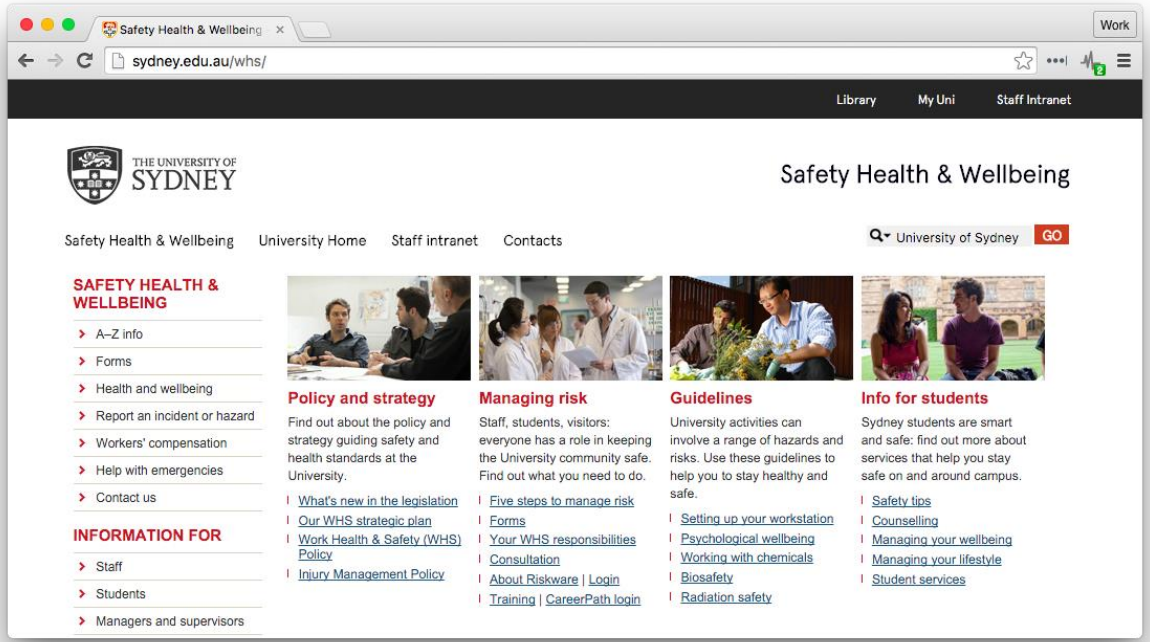
Academic integrity (University policy)

- Submitted work is compared against other work
  - Turnitin for textual tasks (through eLearning)
  - other systems for code
- Penalties for academic dishonesty or plagiarism can be severe
- Complete required self-education AHEM1001

HEALTH AND SAFETY

THE UNIVERSITY OF SYDNEY

Health and safety information



Disability services

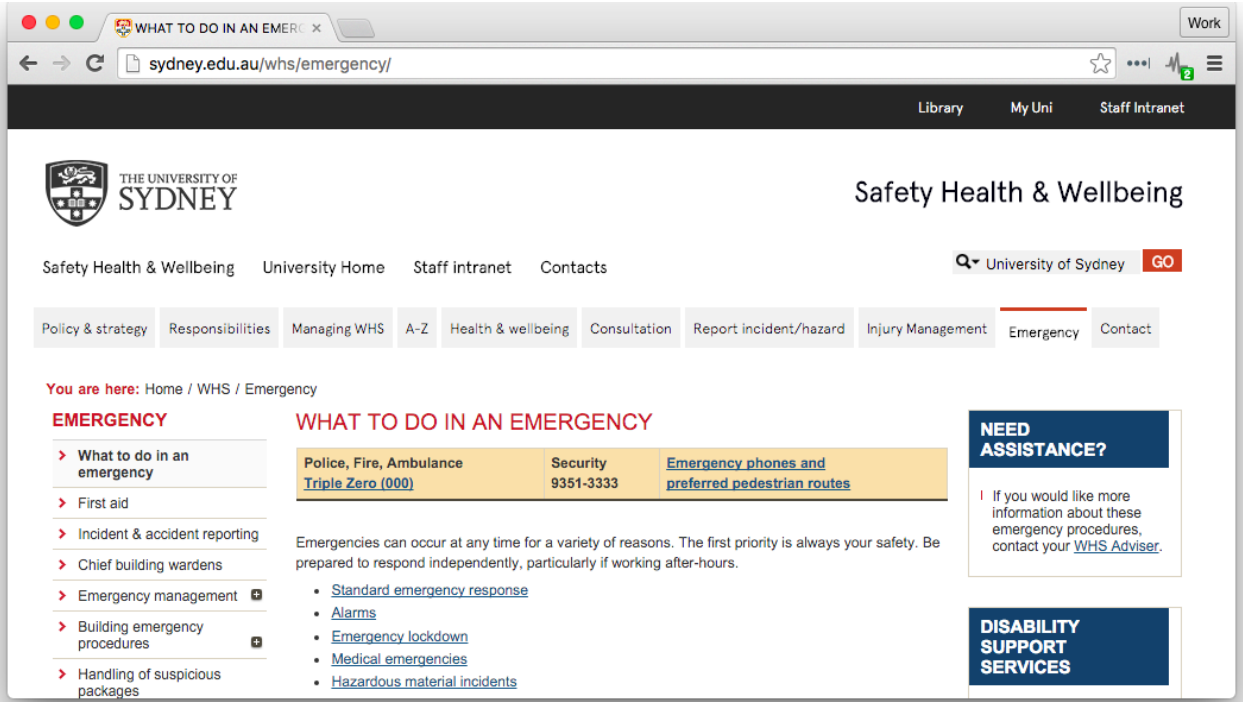
- Includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities
- Register with Disability Services early possible if you might need assistance

http://sydney.edu.au/study/academic-support/disability-support.html

Other support and services

- Learning support
  http://sydney.edu.au/study/academic-support/learning-support.html
- International students
  http://sydney.edu.au/study/academic-support/support-for-international-students.html
- Aboriginal and Torres Strait Islander students
  http://sydney.edu.au/study/academic-support/aboriginal-and-torres-strait-islander-support.html
- Student organisation (can represent you in academic appeals, etc)
  http://srcusyd.net.au/

Emergency information



Emergency evacuation

Evacuation Procedures

ALARMS

BEEP... BEEP... Prepare to evacuate

1. Check for any signs of immediate danger.
2. Shut Down equipment / processes.
3. Collect any nearby personal items.

WHOOP... WHOOP... Evacuate the building

1. Follow the EXIT exit signs.
2. Escort visitors & those who require assistance.
3. DO NOT use lifts.
4. Proceed to the assembly area.

EMERGENCY RESPONSE

1. Warn anyone in immediate danger.
2. Fight the fire or contain the emergency, if safe & trained to do so.

If necessary...

3. Close the door, if safe to do so.
4. Activate the "Break Glass" Alarm  or 
5. Evacuate via your closest safe exit. EXIT 
6. Report the emergency to 0-000 & 9351-3333

If a person is seriously ill or injured

- Call an ambulance 0-000
- Notify the closest Nominated First Aid Officer
- Call security 9351-3333

- Nearest medical facility:
  - University Health Service
  - Level 3, Wentworth Building
  - RPA Emergency

INTRODUCTIONS AND BACKGROUNDS

THE UNIVERSITY OF SYDNEY

Exercise: Getting to know your table

Brake into pairs of two

Take turns interviewing each other

- What is your name?
- What is your experience (industry, org unit, role, skills, etc)?
- What excites you about data science (research/business applications, career prospects, favorite data set or data science project)?

Exercise: Survey of skills and interests

https://goo.gl/BgVnjR

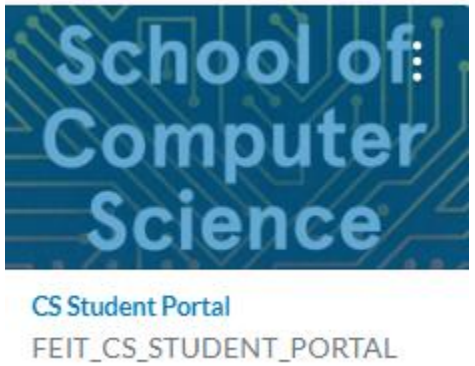
(link on Canvas)

Survey – Individual Responses

- What kind of role would you like (Data Engineer/Scientist, Analyst, etc)?
- What are the three most important data analytics skills?

We'll explore this data in week 2 exercises!

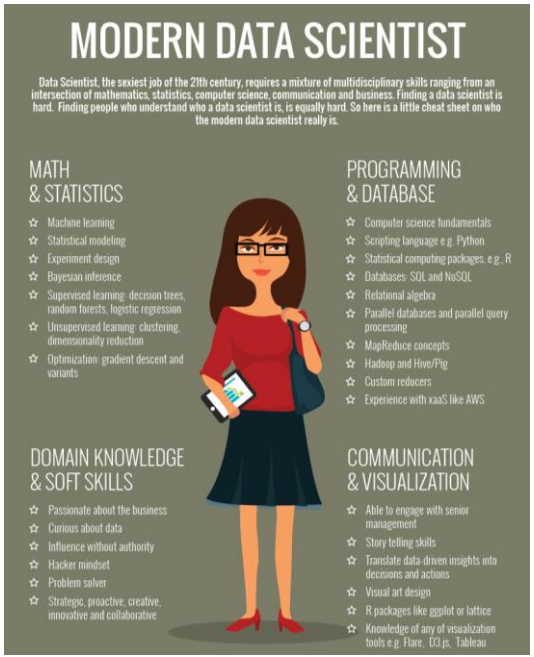
https://www.facebook.com/IT.Sydney.University/



WHAT IS DATA SCIENCE?

Data Scientists build intelligent systems to derive knowledge from data.

Data Science skills



http://www.marketingdistillery.com/2014/11/29/is-data-science-a-buzzword-modern-data-scientist-defined/

Data scientists help organisations:

- understand their data,
- ask meaningful questions,
- derive transformative insights,
- lead empirically grounded decision making.

Cross Industry Standard Process for Data Mining (CRISP-DM)

By Kenneth Jensen - Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610

Business Understanding Phase

- Business objective
  - Understand business processes
  - Associated costs/pain
- Assess situation
- Define the success criteria
- Data science goals
- Project plan
  - List assumptions and risk (technical/financial/business/ organizational) factors

Some example goals

- Farmer wants advice on what fertilizer to use, to maximize crop yield

- Bank wants to automatically flag some credit card purchases as potentially fraudulent, to delay payment till checks have been made

- Biologist wants to be able to find out which species of micro-organism are present in a location, given a list of protein fragments found in an environmental sample

Some example goals (cont'd)

- Doctor wants to determine whether a patient is likely to have a particular disease, given results of tests (none of which is perfect)

- Designer wants a car that brakes automatically when a pedestrian steps in front

Data Understanding Phase

- Collect Data
  - What are the data sources?
    - Original sources (these all will contain errors!):
      - sensors (measure the world)
      - surveys (ask people)
      - digital logs (track IT activities)
    - Secondary sources
      - other scholars, organizations, etc
      - data may already be summarized, transformed, cleaned, etc

Examples of datasets

- Census
  - raw data has individual level demographics etc
  - available summaries combine these into counts in a suburb etc
- Crop observations
  - many plantings, with many features (seed type, date, weather, soil, fertilizer etc), and resulting crop yields
- Credit card histories
  - lots of transactions of many users, with many features, some transactions were reported as fraudulent
- Medical records
  - lots of patients, their test results, diagnoses

Data Understanding Phase

- Data Description
  - Document data quality issues
  - Compute basic statistics
- Data Exploration
  - How is it structured? What is the meaning of the different features?
    - eg is temperature the daily maximum, monthly average, at some specific time? is income measured in actual dollars or inflation-adjusted ones?
  - Simple univariate data plots/distributions
  - Investigate attribute interactions
    - Can you find patterns connecting different features?
  - Data Quality Issues

Data Preparation Phase

- Integrate Data
  - Joining multiple data tables
  - Summarisation/aggregation of data
- Select Data
  - Attribute subset selection
    - Rationale for Inclusion/Exclusion
  - Data sampling
    - Training/Validation and Test sets

Data Preparation Phase (cont'd)

- Data Transformation
  - Using functions such as log
  - Factor/Principal Components analysis
  - Normalization/Discretization/Binarization
- Clean Data
  - Handling missing values/Outliers
- Data Construction
  - Derived Attributes

The Modelling Phase

- Select of the appropriate modelling technique
  - Dependent on
    - Data mining problem type
    - Output requirements
- Develop a testing regime
  - Sampling
    - Verify samples have similar characteristics and are representative of the population

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

The Modelling Phase (cont'd)

- Build Model
  - Choose initial parameter settings
  - Study model behaviour
    - Sensitivity analysis

- Assess the model
  - Beware of over-fitting
  - Investigate the error distribution
    - Identify segments of the state space where the model is less effective
  - Iteratively adjust parameter settings
    - Document reasons of these changes

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

Examples of Models

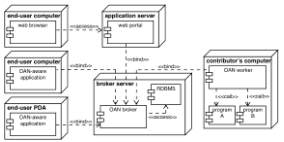
- Model to predict the purity of the environment based on carbon level (Regression prediction model)
- Model to classify a person whether he is cheating in his tax return or not (Classification prediction model).
- Model to find hidden patterns and association rules in the basket market analysis (Clustering or association rules).
- Model to detect anomalies or outliers such as spam emails (Classification prediction model).

The Evaluation Phase

- Validate Model
  - Human evaluation of results by domain experts
  - Evaluate usefulness of results from business perspective
    - Define control groups
    - Expected Return on Investment
- Review Process
- Determine next steps
  - Potential for deployment
  - Metrics for success of deployment

The Deployment Phase

- Knowledge Deployment is specific to objectives
  - Knowledge Presentation
  - Automated pre-processing of live data feeds
  - Generation of a report
    - Online/Offline
  - Monitoring and evaluation of effectiveness



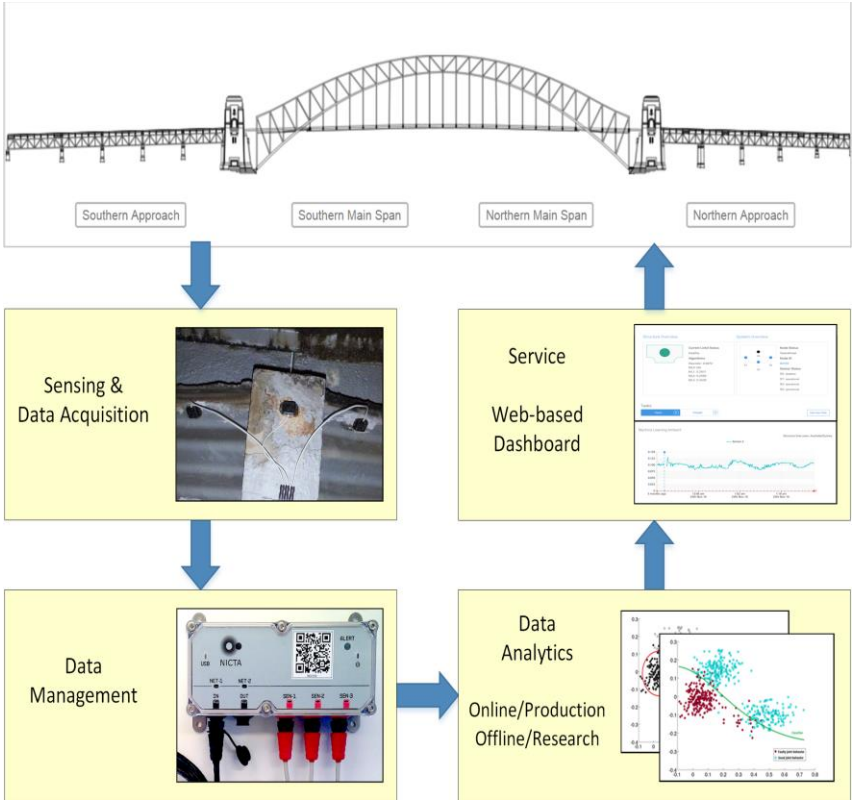
DATA SCIENCE PROJECTS

Example: Reducing costs through route optimisation

http://www.bloomberg.com/news/articles/2013-10-30/ups-uses-big-data-to-make-routes-more-efficient-save-gas

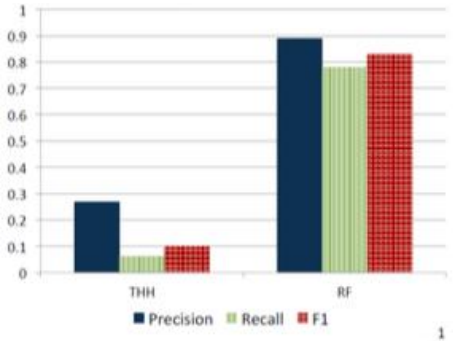
- Use customer, vehicle and delivery data
- 1 mile less per day for every driver saves $50 million p.a. in fuel, maintenance and time
- Less idling, e.g., by avoiding left turns, saved 1.6 million gallons of fuel in 2012

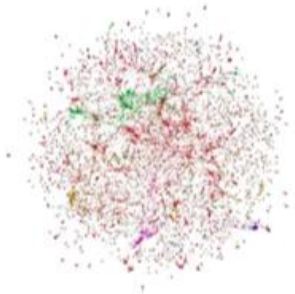
Example: Structural Health Monitoring (SHM)

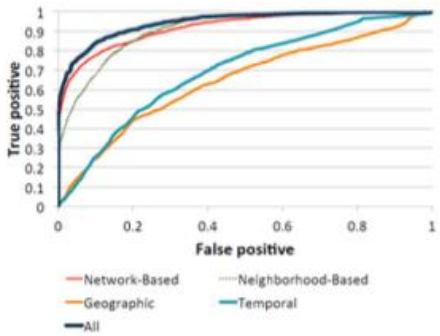


- Time-based maintenance:
  - Preventative maintenance schedules
  - Too early or too late
- SHM:
  - Condition-based maintenance using sensors
  - Data-driven approach establishes model from data, using machine learning techniques.

Example: Preventative policing





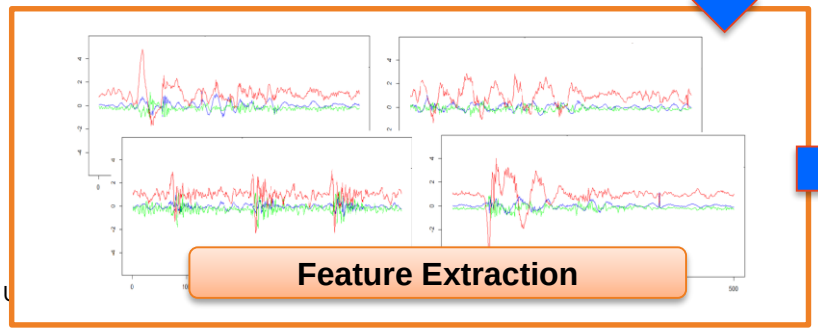


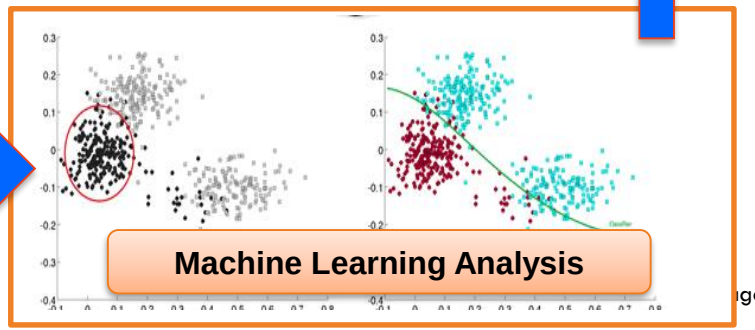
- Given social network from arrest records, geographic, temporal data
- Predict whether a person is likely to be involved in crime
- Chicago police using to issue preemptive warnings:

"We're watching you"

http://arxiv.org/pdf/1508.03965v1.pdf

Example: Road Condition Assessment from Vehicle-mounted Sensor





WHERE DO I GET DATA?

Source Example: UCI Machine Learning Repository Datasets

About

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

URL

https://archive.ics.uci.edu/ml/datasets.html

Data sets

- Classification
  - Breast Cancer
  - Diabetes
  - Letter Recognition …etc.
- Regression
  - Forest Fires
  - Buzz in social media ..etc.
- Clustering
  - Bag of Words
  - Sponge …etc.

Source Example: Kaggle Datasets

About

Kaggle is an online platform for data science competitions. Some data sets are publicly available.

URL

https://www.kaggle.com/datasets

Data sets

- Amazon fine food reviews
- Health insurance marketplace
- World food facts
- Ocean ship logbooks
- Reddit comments
- Hillary Clinton's emails
- GOP debate Twitter sentiment
- NIPS 2015 papers

Source Example: AIHW Data

About

Australian Institute of Health & Welfare collects data that provide insight into the health and wellbeing of the multifaceted Australian population.

URL

http://www.aihw.gov.au/data-by-subject/

Data sets

- Alcohol, Tobacco & Drugs
- Cancer
- Children's health
- Height & weight
- Hospitals
- Indigenous health
- Mental health
- Lots more!

Source Example: Reddit comments

About

Reddit is a social news web site that functions like an online bulletin board.

URL

https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment

Data sets

- 1.7 billion public comments

REVIEW

W1 Review: Introductions and housekeeping

Objective

Housekeeping; Learn about backgrounds and goals; Define data science.

Lecture

- Welcome, introductions
- Unit overview, assessment, resources
- Learning Python with Grok
- Discuss definitions/scope of data science

Readings

- Data Science from Scratch: Ch 1
- Is being a data scientist really the best job in America?
- 8 skills you need to be a data scientist

Exercises

- Introductions / interviews
- Interests / definitions

TODO in W1

- Grok Python modules 1-3
- Fill out & submit background survey
- Choose possible project data

Formulating a COMP5310 project (Stage 1 & 2)

- By next week:
  - Identify possible problems and data sets
  - Think about questions the data can answer

- Other possible data sets…

Source Example: Yahoo Webscope

About

The Yahoo Webscope program is a reference library of data sets for non-commercial use by academics.

URL

http://webscope.sandbox.yahoo.com/

Data sets

- 13.5 TB of user interaction data
- Search engine query logs
- Q&A forum data
- Query entity disambiguation

Source Example: GovHack Data

About

GovHack is an annual event that brings people together to innovate with open government data. They list many data sets from Australia and New Zealand.

URL

http://portal.govhack.org/datasets.html

https://data.gov.au/

Data sets

- ABC news and TV archives
- Australian census data
- Labour, industry, transport data
- Health and welfare data
- Various CSIRO data sets
- Finance, IP, geoscience, archives, etc

NEXT TIME

Next week: Data exploration with spreadsheets

Objective

Use interactive tools to explore a new data set quickly.

Lecture

- Data types, cleaning, preprocessing
- Descriptive statistics, e.g., mean, stddev, median
- Descriptive visualisation, e.g., scatterplots, histograms

Readings

- Data Science from Scratch: Ch 2-3

Exercises

- Google Sheets: Visualisation
- Google Sheets: Descriptive stats

TODO for W2

- Grok Python modules 1-3
- Make sure you answered today's background survey
- Explore project data
- **GET YOUR GOOGLE ACCOUNT!**

Project and discussion time

Time for you to talk
to tutors, instructors and each other
about data sets, data exploration
and possible research questions.


THE UNIVERSITY OF
SYDNEY