COMP5310: Principles of Data Science

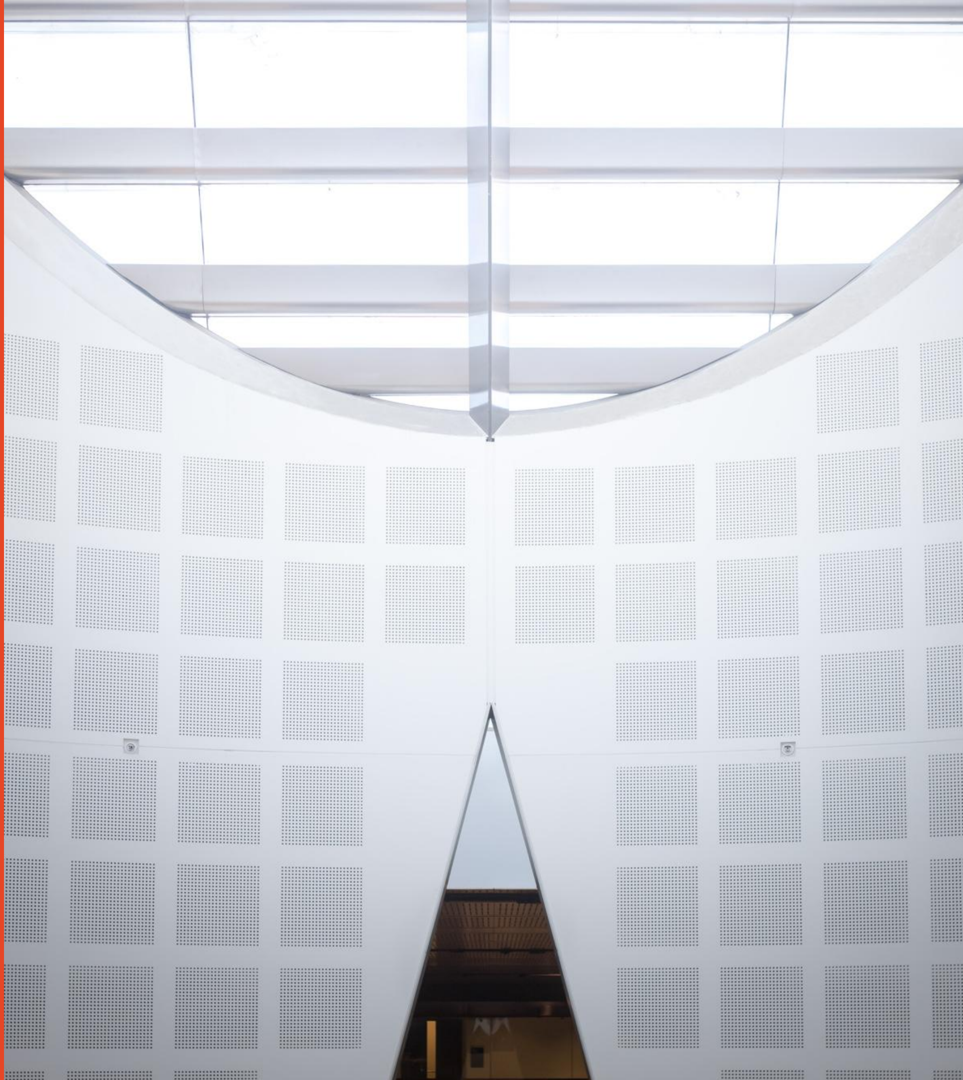
W1: Introduction

Presented by

Maryam Khanian

Based on slides by previous lecturers of this unit of study

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.

UNIT ARRANGEMENTS

COMP5310: 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
- W7: Data Mining: association rules
- W8: Data mining: clustering and dimensionality reduction
- W9: Machine learning: regression
- W10: Machine learning: classification
- W11: Unstructured data
- W12: Ethics in data science
- W13: Review

COMP5310: Places

- Lecture: Tuesday 5pm to 7pm
- Lab: depends on your timetable
  - Go to the lab you are scheduled for
  - If for some reason you missed it, you can attend a later lab session if there is space and the tutor agrees, but ask the tutor before taking a seat
- Do not miss class, except for illness, emergencies, etc
- Get help from staff if you feel you are falling behind

COMP5310: People

COMP5310: People – who to ask for what

- **EdStem Discussion Forum** (Canvas > Ed Discussion)
  - General questions about lectures, Python and SQL.
  - Content of lectures.
  - Technical questions about data science.

- Maryam Khanian Najafabadi/ Sanket Srivastava (TA) /Michelle (Weiyi) Wang (TA)
  - Administrative questions.
  - Group work issues.
  - Special Consideration.
  - Rules and policies.
  - Illness and misadventure.

COMP5310: Resources

Log into Canvas with unikey/password

- Canvas > Modules: lab/lecture materials, readings.
- Canvas > Assignments: will be available in Week 3.
- Canvas > Recorded Lectures: (technology is not reliable).
- Canvas > Ed Discussion: discussion forum for general questions.
- Canvas > Ed Lessons: Python and SQL exercises.
- Official schedule, list of learning outcomes, etc.: https://sydney.edu.au/units/

COMP5310: Python and SQL material

- Tutorials from week 3 onwards will use Python and SQL
- Self-guided Python and SQL learning through Ed Lessons.
  - Please complete it by week 5

Canvas > Ed Lessons

COMP5310: Reference books

- **Data Science from Scratch.** Grus. O’Reilly Media. 2019.
  - Available electronically through library.

- **Doing Data Science.** O’Neill and Schutt. O’Reilly Media. 2015.
  - Available electronically through library.

COMP5310: Expectations

- Students attend scheduled classes and devote an *extra* 6-9 hrs. per week.
  - Doing assessments.
  - Preparing and reviewing for classes.
  - Revising and integrating the ideas.
  - Practicing and self-assessing.
- Students are responsible learners.
  - Participate in classes, constructively.
    - Respect for one another (criticize ideas, not people).
    - Humility: none of us knows it all; each of us knows valuable things.
  - Check Canvas site at least once a week!
  - Notify academics whenever there are difficulties.
  - Notify group partners honestly and promptly about difficulties.

ASSESSMENTS

Assessment

- The official syllabus is the authoritative source of assessment information.
  - https://www.sydney.edu.au/units/COMP5310/2025-S1C-NE-CC

- 15%: Assignment 1(Week 6)
- 25%: Assignment 2(Week 11)
- 60%: Final exam.

*Sydney time.

Assignment 1: Obtain data, clean it and load it.

Objective

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

Activities

- Choose a data set, clean it and load it.
- Define problem, specify requirements.

Output

- Group Report
  - **Individual Component**: Describe in detail any exploratory data analysis you performed which provided you relevant information to answer your research question.
  - **Group Component** : Discussion, Conclusion

Marking

- Based on both individual and group components.

Assignment 2: 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

- Progressive reports describing framework, analysis and conclusions (plus code).

Marking

- Based on both individual and group components.

Final exam

Objective

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

Format

- Written examination.
- Duration: 2 hours

Marking

- 60% of overall mark.
- Must get 40% on exam to pass unit per SCS policy.

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/ .
- Also, notify coordinator by email *as soon as anything begins to go wrong*.
- There is a similar process if you need special arrangements e.g., for religious observance, military service, representative sports.

Late submissions in COMP5310

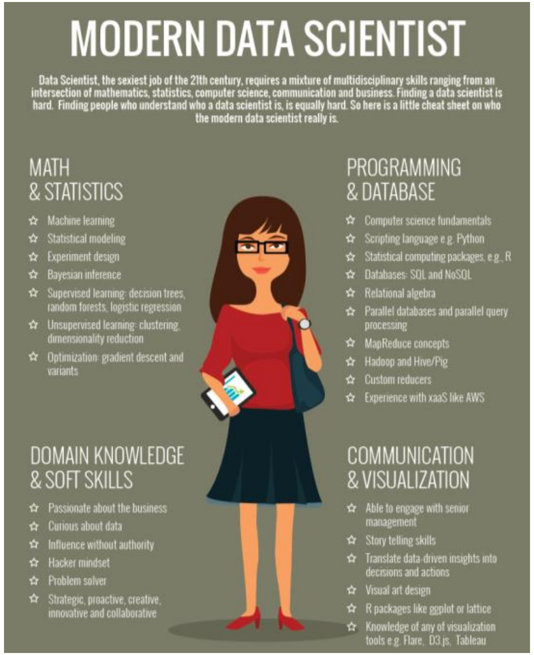
Suppose you hand in work after the deadline:

- If you have not been granted special consideration or arrangements:
  - A penalty of 5% of the maximum marks will be taken per calendar day late. After five days, a mark of zero will be awarded.
  - *E.g. An assignment that would normally get 9/10 and is 2 days late loses 10% of the full 10 marks, i.e. new mark = 8/10*
  - *E.g. An assignment that would normally get 5/10 and is 5 days late loses 25% of the full 10 marks, i.e. new mark = 2.5/10*
- **Warning:** submission sites get very slow near deadlines.
- Submit early

WHAT IS DATA SCIENCE?

Data Scientists build intelligent systems to derive knowledge from data.

Data Science skills



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ☆ Machine learning
- ☆ Statistical modeling
- ☆ Experiment design
- ☆ Bayesian inference
- ☆ Supervised learning: decision trees, random forests, logistic regression
- ☆ Unsupervised learning: clustering, dimensionality reduction
- ☆ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ☆ Computer science fundamentals
- ☆ Scripting language e.g. Python
- ☆ Statistical computing packages, e.g., R
- ☆ Databases: SQL and NoSQL
- ☆ Relational algebra
- ☆ Parallel databases and parallel query processing
- ☆ MapReduce concepts
- ☆ Hadoop and Hive/Pig
- ☆ Custom reducers
- ☆ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ☆ Passionate about the business
- ☆ Curious about data
- ☆ Influence without authority
- ☆ Hacker mindset
- ☆ Problem solver
- ☆ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

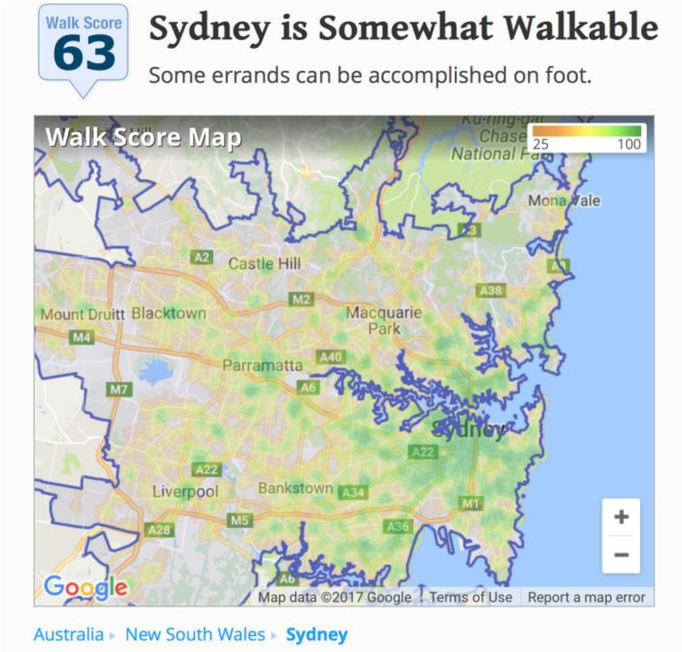
- ☆ Able to engage with senior management
- ☆ Story telling skills
- ☆ Translate data-driven insights into decisions and actions
- ☆ Visual art design
- ☆ R packages like ggplot or lattice
- ☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

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.

Example: Urban & Transport Planning, Public Health



Walk Score 63

Sydney is Somewhat Walkable

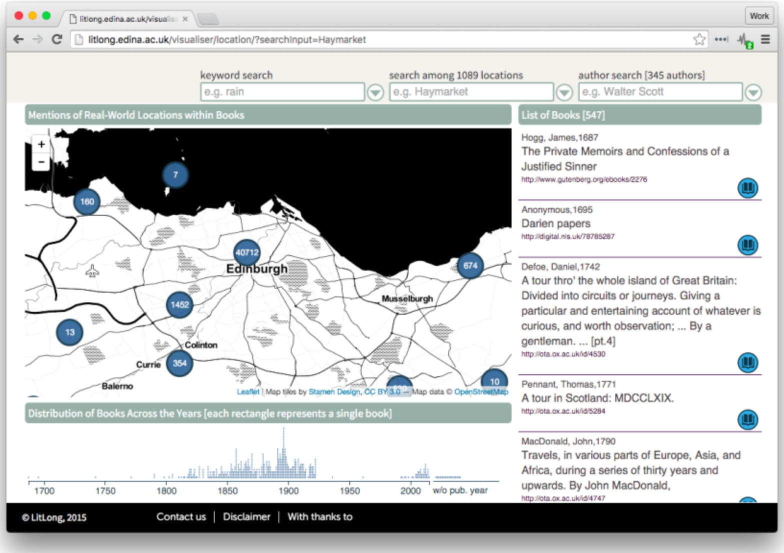
Some errands can be accomplished on foot.

Australia ▸ New South Wales ▸ **Sydney**

http://www.walkscore.com/research/

- Integration of data about road and public transport network with data about population, services, restaurants, amenities etc.
- Summarising *Walkability Score overlayed* on map visualisation
- Prediction of impact of new developments
- API for use in 3rd party apps, eg. supporting real estate agents

Example: Mapping literary references


litlong.edina.ac.uk/visualise ×
Work
litlong.edina.ac.uk/visualiser/location/?searchInput=Haymarket
keyword search
e.g. rain
search among 1089 locations
e.g. Haymarket
author search [345 authors]
e.g. Walter Scott
Mentions of Real-World Locations within Books
+
−
7
160
40712
Edinburgh
674
Musselburgh
1452
13
Colinton
Currie
354
Balerno
10
Leaflet | Map tiles by Stamen Design, CC BY 3.0 — Map data © OpenStreetMap
Distribution of Books Across the Years [each rectangle represents a single book]
1700
1750
1800
1850
1900
1950
2000
w/o pub. year
© LitLong, 2015
Contact us | Disclaimer | With thanks to

List of Books [547]

Hogg, James,1687
The Private Memoirs and Confessions of a Justified Sinner
http://www.gutenberg.org/ebooks/2276

Anonymous,1695
Darien papers
http://digital.nls.uk/78785287

Defoe, Daniel,1742
A tour thro' the whole island of Great Britain: Divided into circuits or journeys. Giving a particular and entertaining account of whatever is curious, and worth observation; ... By a gentleman. ... [pt.4]
http://ota.ox.ac.uk/id/4530

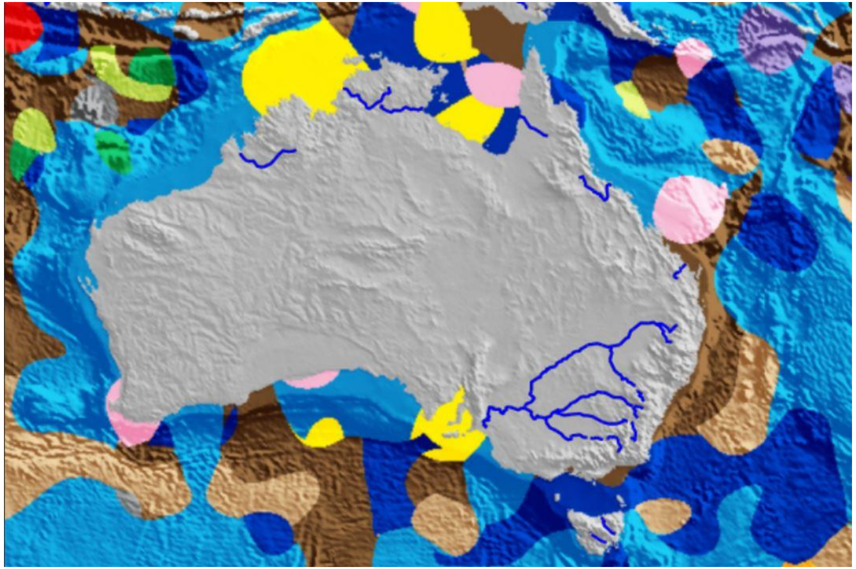
Pennant, Thomas,1771
A tour in Scotland: MDCCLXIX.
http://ota.ox.ac.uk/id/5284

MacDonald, John,1790
Travels, in various parts of Europe, Asia, and Africa, during a series of thirty years and upwards. By John MacDonald,
http://ota.ox.ac.uk/id/4747

– Identify and resolve location mentions in literature
– Overlay references on map visualisation
– Keyword, location and author search

http://litlong.org/

Example: Mapping seafloor geology with SVM

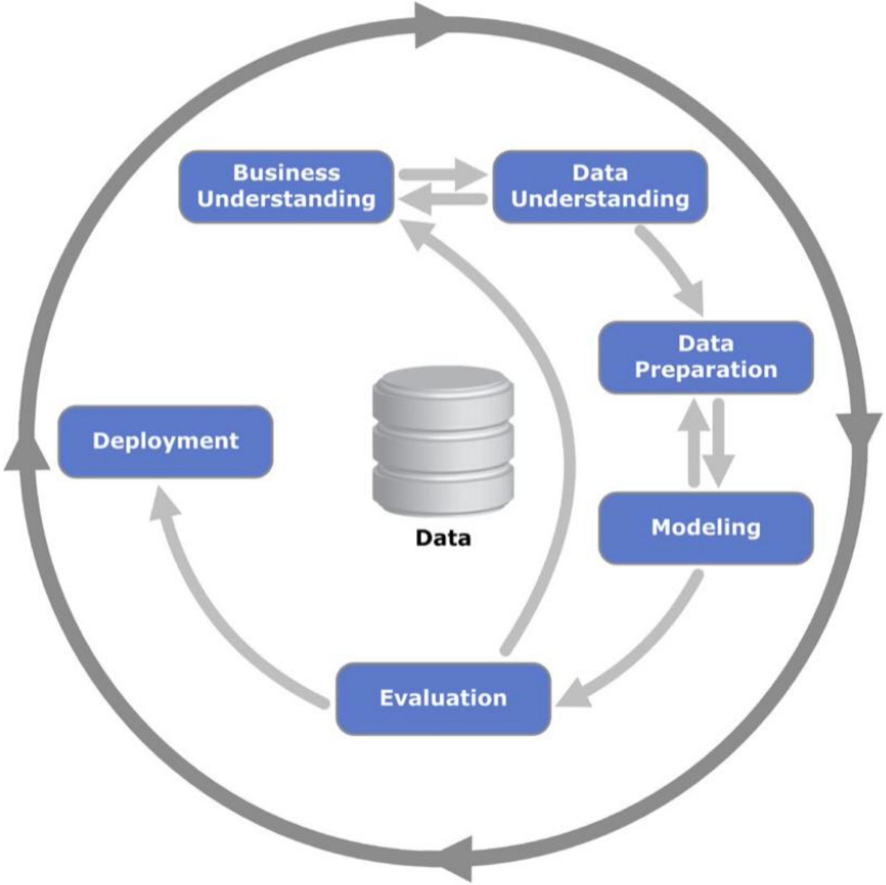


http://portal.gplates.org/#SEAFLOOR

- Use descriptions from 14,500 samples collected from 1950-present
- Predict sediment in unobserved regions using support vector machine

DATA SCIENCE WORKFLOW

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

The DM process

1) Business understanding

- Investigating the business objectives and requirements
- Deciding whether DM can be applied to meet them
- Determining what kind of data can be collected to build a deployable model

2) Data understanding

- Get an initial dataset; is it suitable for further processing?
- If the data quality is poor, collect more data
- Gain insights from data and review the objective – can DM be applied?

The DM process

3) Data preparation - preprocessing the data, so that ML algorithms can be applied. This involves cleaning and various transformations:

- Cleaning: data in real world is:
  - Incomplete, e.g. missing values   lacking attribute values e.g., occupation=“ ”
  - Noisy, e.g. containing errors or outliers   Salary=“-10”
  - Inconsistent, e.g. in codes, names
    - Age=“27” Birthday=“03/07/1997”

  Fill in missing values, smooth noisy data, identify outliers and remove them, resolve inconsistencies
- Transformation – convert to common format; transform to new format; perform normalization, dimensionality reduction and feature selection

The DM process

4) Modelling – building ML models, e.g. a prediction model

3) and 4) go hand-in-hand and there are many iterations, e.g. the model informs the use of different preprocessing – e.g. use different feature selection and dimensionality reduction, build a model again

The DM process

5) Evaluation – very important

- How good is the performance? E.g. accuracy, F1 measure, etc.
- Are the patterns meaningful and useful, or just reflecting spurious regularities?
- If the performance is poor, reconsider the project and return to step 1)
- If the performance is good -> deploy it in practice

6) Deployment

- Typically requires integration into a larger software system by software engineers
- May be necessary to re-implement the model in a different programming language

DATA SCIENCE WORKFLOW

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 factors (technical/financial/business/organizational).

Goal examples

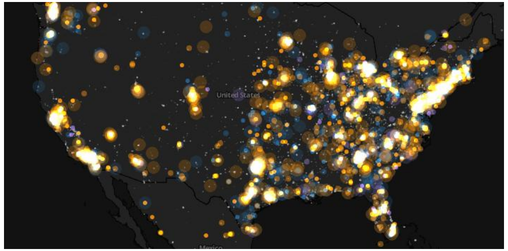
- Farmer wants advice on what fertilizer to use to maximise crop yield.
- Bank wants to automatically flag some credit card purchases as potentially fraudulent to delay payment until checks have been made.
- Biologist wants to be able to find out which species of micro-organisms are present in a location given a list of protein fragments found in an environmental sample.

Data is everywhere

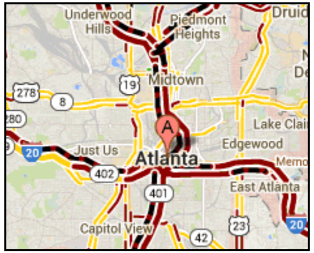
- Data explosion – society produces and stores huge amounts of data
  - Due to automated data collection tools and sensors, mature database technology, cheaper and more powerful computers
  - Sources: business, science, medicine, economics, environment, web, etc.
- Examples:
  - purchase data – supermarket, department stores, online stores – e.g. Amazon handles millions of visits a day
  - bank/credit card usage data
  - web data – Google, Facebook; other social networking sites

Sky survey data

E-Commerce



Social Networking: Twitter



Traffic

2

Data Understanding Phase

Collect Data

- What are the data sources?
  - Original sources (these will contain errors!):
    - Sensors (measure the world).
    - Surveys (ask people).
    - Digital logs (track IT activities).
  - Secondary sources:
    - Other scholars, organisations, etc.
    - Data may already be summarized, transformed, cleaned, etc.

Dataset examples

Census

- Raw data has individual level demographics.
- Available summaries combine these into counts in a region, suburb, city, etc.

Crop observations

- Many plantings with many features (seed type, date, weather, soil, fertilizer, etc.) and 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, etc.

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?
    - e.g., Is temperature the daily maximum, monthly at some specific time?
    - e.g., Is income measured in actual dollars or inflation-adjusted dollars?
- Simple univariate data plots/distributions.
- Investigate attribute interactions.
  - Can you find patterns connecting different features?

3 Data Preparation Phase

- **Integrate data**
  - Joining multiple data sources.
  - Summarisation/aggregation of data.
- **Select data**
  - Attribute subset selection.
    - Rationale for inclusion/exclusion.
  - Data sampling.
    - Training/validation and test sets.
- **Transform data**
  - Using functions such as log.
  - Principal components analysis.
  - Normalisation, discretisation or binarization.
- **Clean data**
  - Handling missing values/outliers.
- **Construct data**
  - Derived attributes.

DATA SCIENCE WORKFLOW

Example data sources

THE UNIVERSITY OF SYDNEY

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: Crowdflower Data for Everyone

About

Crowdflower is an online platform for crowdsourcing data and annotation. Some data sets are released to the public.

URL

http://www.crowdflower.com/data-for-everyone

Data sets

- Clothing pattern identification
- Relevancy of terms to disaster relief
- Economic news tone and relevance
- Police-involved fatalities
- Wikipedia image classification
- Image classification: people and food
- Biomedical image modality
- Academy Award demographics

Source Example: AWS Large Data Sets

About

Big data sets hosted on Amazon Web Services.

URL

https://aws.amazon.com/public-data-sets

Data sets

- Landsat (satellite imagery of Earth)
- NEXRAD (real-time/archival weather)
- NASA NEX (earth science collection)
- Common Crawl (5 billion web pages)
- US Census (1980, 1990 and 2000)
- Several genome 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: Reddit comments

About

Reddit is a social news website 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

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

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!

DATA SCIENCE WORKFLOW

4 Modelling Phase

- **Select an appropriate modelling technique**
  - Depends on:
    - Problem type.
    - Output requirements.
- **Develop a testing regime**
  - Sampling.
    - Verify samples have similar characteristics and are representative of the population.
- **Build model**
  - Choose initial parameter settings.
  - Study model behaviour.
    - Sensitivity analysis.
- **Assess model**
  - Beware of over-fitting.
  - Investigate the error distribution.
    - Identify segments where the model is less effective.
- **Iteratively adjust parameter settings**
  - Document reasons of these changes.

Model Examples

- Model to predict the purity of the environment based on carbon level (regression prediction model).
- Model to classify a person as whether is cheating on 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).

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 (ROI).
- **Review process**
- **Determine next steps**
  - Potential for deployment.
  - Metrics for success of deployment.

6

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.

REVIEW

W1 Review: Introductions and housekeeping

Objective

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

Lecture

- Welcome, introductions.
- Unit overview, assessment, resources.
- Discuss definitions/scope of data science.

Readings

- Data Science from Scratch: Ch 1.

Tutorial

- Install Anaconda and PostgreSQL.

TO-DO in W1

- Ed Lessons Python modules 1-3.
- Organise into project groups.