INFO3406 Introduction to Data Analytics

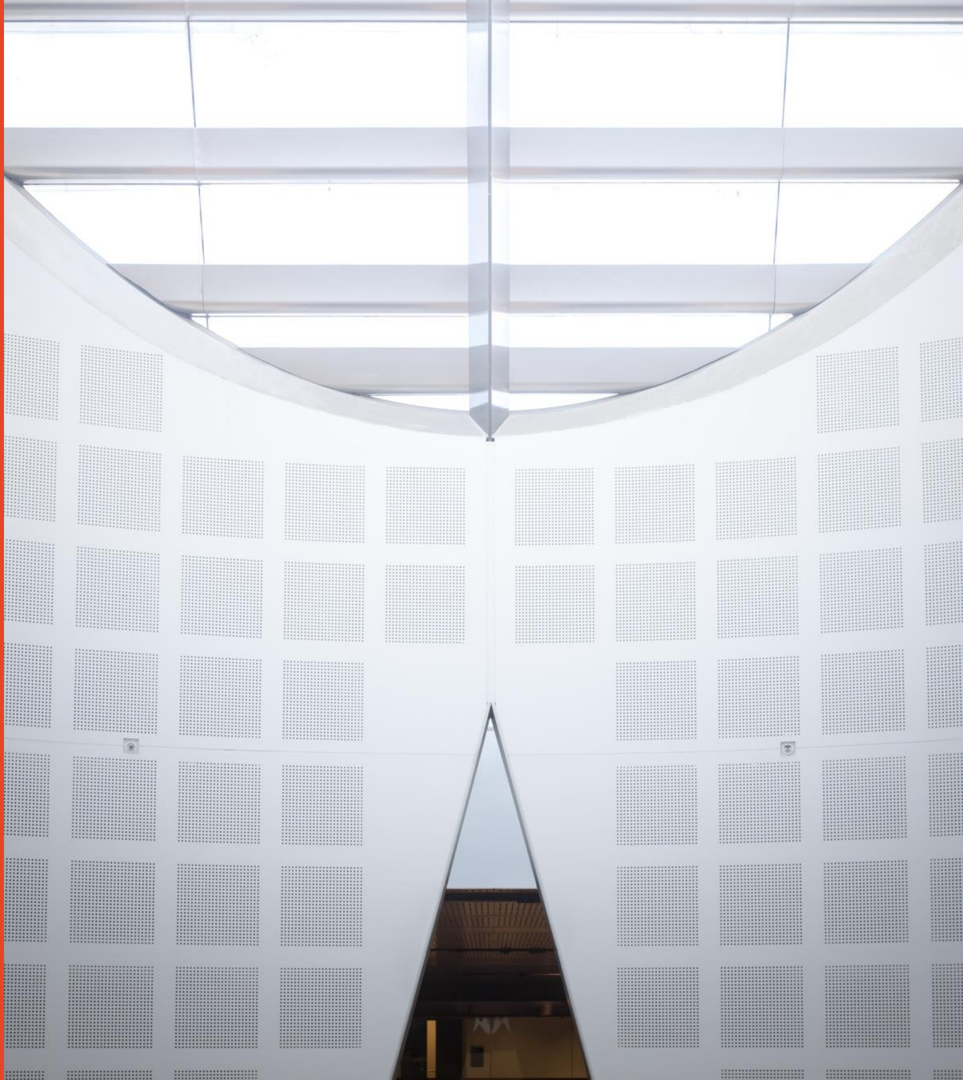
W1: Introduction

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 analytics

Perspectives and communication

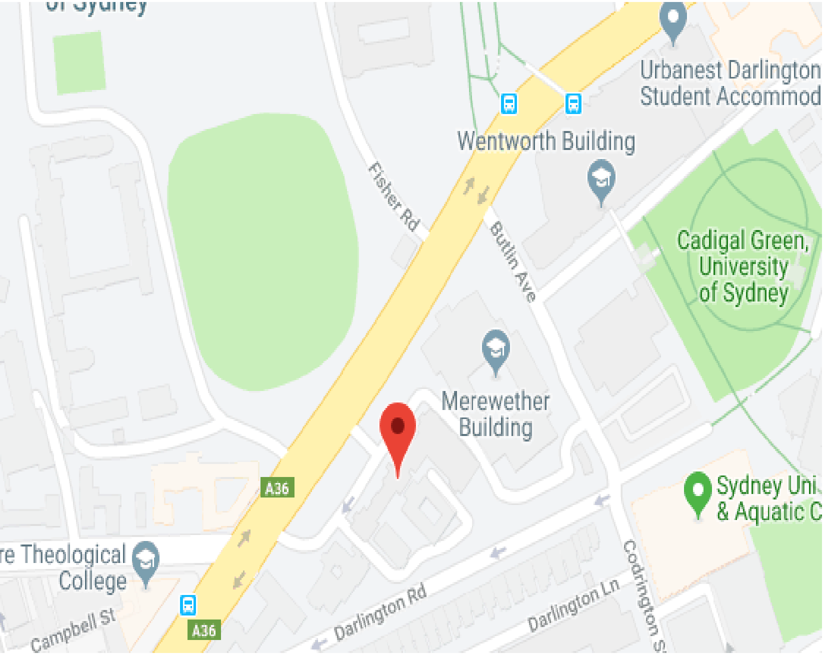
Doing data science requires

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

Chance to build key soft skills as well as technical skills

UNIT ARRANGEMENTS

Here you are



Institute building (Building H03)

- 1.5 hours Lecture

  Thursday, 11:00-12:30

  Institute Lecture Theatre 1
- One hour tutorial
  - Please check your lab allocations
- Extra three hours **optional** tutorial on Friday starting from week3
  - 3pm - 4pm, SIT lab 130B
  - 4pm - 5pm, SIT lab 114
  - 4pm - 5pm, SIT lab 115

Introducing Team

Lecturer

Dr Ali Anaissi

Tutors

Seid Miad Zandavi, PhD Student

Claudio Diaz, PhD Student

Unit Coordinator

Dr Ali Anaissi

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

Typical lecture

- Introduce topic of the week
- 1st concept
  - Introduce topic/data/scenario
  - Exercise in Python (in lab)
- Other concepts (wash, rinse, repeat)
- Project progress and discussion
- Recap of topic and lessons learnt

Resources

Google Sheets for spreadsheet exercises [week 3]

- 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

PostgreSQL database

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 4
- We provide self-guided Python and SQL learning through Grok
- Please go to this link

https://canvas.sydney.edu.au/courses/4545/pages/data-analysis-skills#OLEO1300

- Enroll yourself in the following free courses:
  - Beginner Programming for Data Analysis (OLEO1306)
  - Managing and Analysing Data: Introduction to SQL (OLEO1300)

Please complete (sooner is better)

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
- 13%: Project stage 1
- 20%: Project stage 2
- 7%: Project stage 3
- 50%: Final exam

Participation

Objective

Ensure everybody is keeping up.

Requirements

Submit code at end of each exercise

Complete Grok exercises

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 summarising problem analysis and proposal (plus code)

Marking

13% 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

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

Presentation (2-3 min)

Marking

27% of overall mark

- 20% report and code
- 7% 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 Python, lecture material and readings

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

50% 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

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 10% of awarded marks per day
  - Maximum 7 days late, then 0 points
- Examples:
  - Work would have scored 60% and is 1 hour late: 54%
  - Work would have scored 70% and is 28 hours late: 56%
- 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."

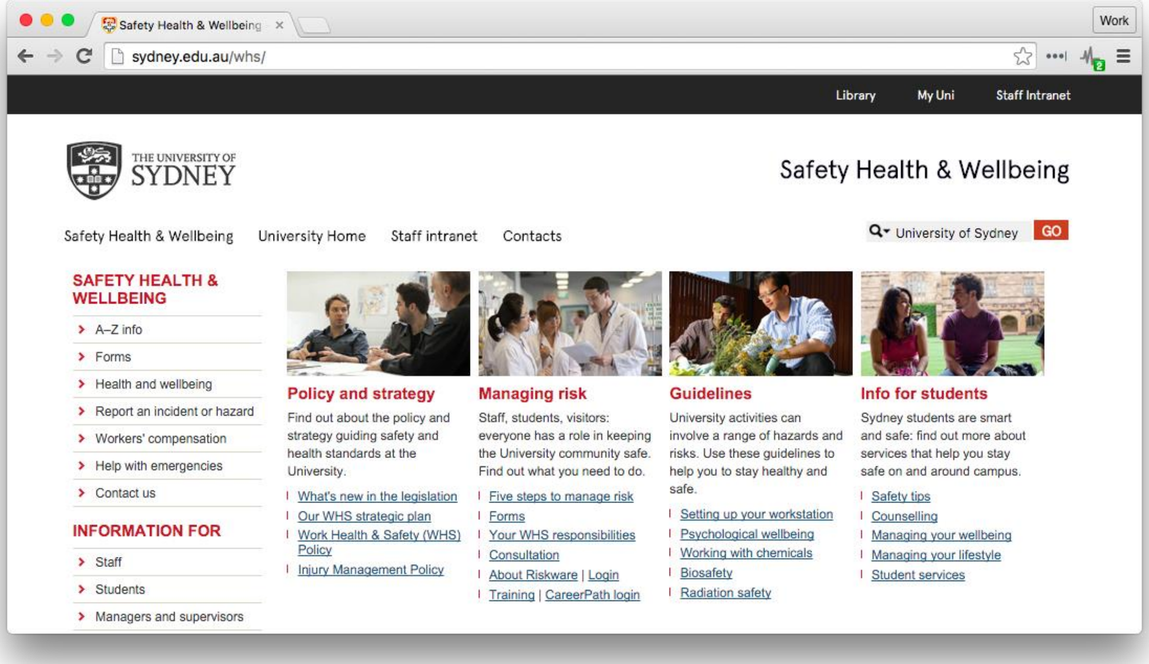
http://sydney.edu.au/elearning/student/EI/index.shtml

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

Health and safety information



Safety Health & Wellbeing

SAFETY HEALTH & WELLBEING

- A–Z info
- Forms
- Health and wellbeing
- Report an incident or hazard
- Workers' compensation
- Help with emergencies
- Contact us

INFORMATION FOR

- Staff
- Students
- Managers and supervisors

Policy and strategy

Find out about the policy and strategy guiding safety and health standards at the University.

- What's new in the legislation
- Our WHS strategic plan
- Work Health & Safety (WHS) Policy
- Injury Management Policy

Managing risk

Staff, students, visitors: everyone has a role in keeping the University community safe. Find out what you need to do.

- Five steps to manage risk
- Forms
- Your WHS responsibilities
- Consultation
- About Riskware | Login
- Training | CareerPath login

Guidelines

University activities can involve a range of hazards and risks. Use these guidelines to help you to stay healthy and safe.

- Setting up your workstation
- Psychological wellbeing
- Working with chemicals
- Biosafety
- Radiation safety

Info for students

Sydney students are smart and safe: find out more about services that help you stay safe on and around campus.

- Safety tips
- Counselling
- Managing your wellbeing
- Managing your lifestyle
- Student services

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

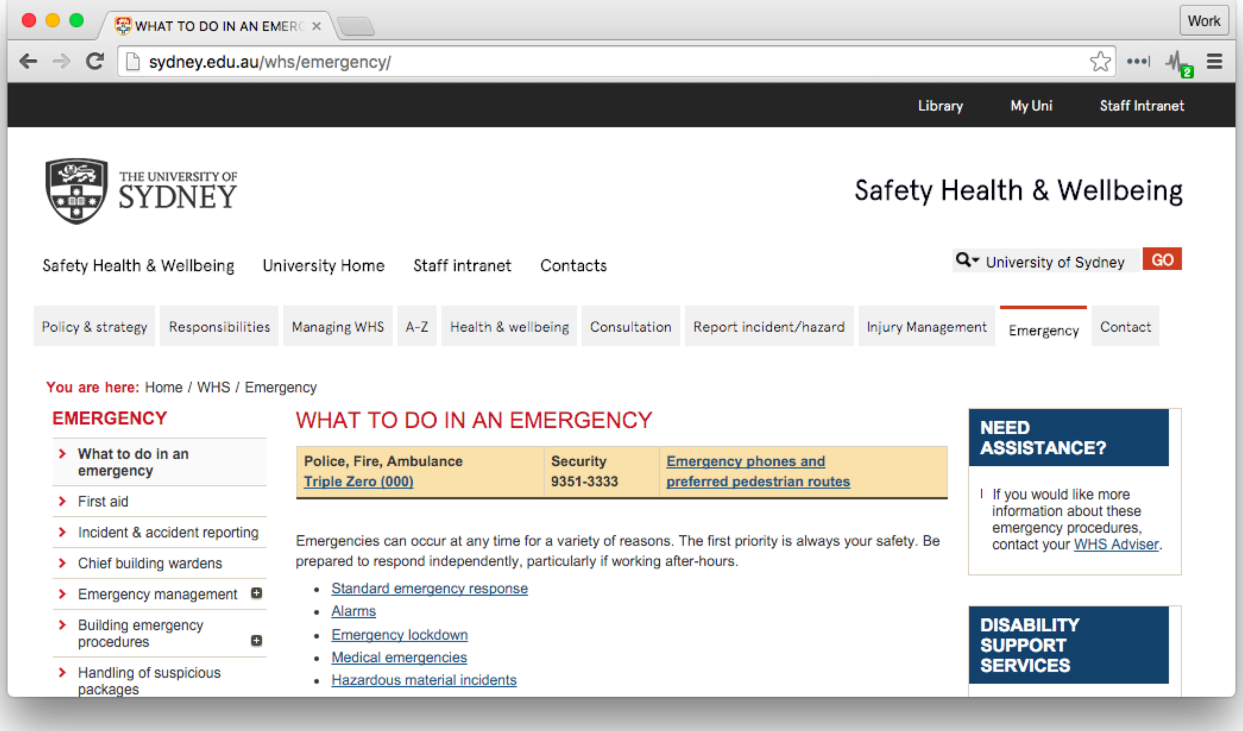
- Learning support
  http://sydney.edu.au/study/academic-support/learning-support.html

- 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



You are here: Home / WHS / Emergency

EMERGENCY

- What to do in an emergency
- First aid
- Incident & accident reporting
- Chief building wardens
- Emergency management
- Building emergency procedures
- Handling of suspicious packages

WHAT TO DO IN AN EMERGENCY

Police, Fire, Ambulance Triple Zero (000)	Security 9351-3333	Emergency phones and preferred pedestrian routes

Emergencies can occur at any time for a variety of reasons. The first priority is always your safety. Be prepared to respond independently, particularly if working after-hours.

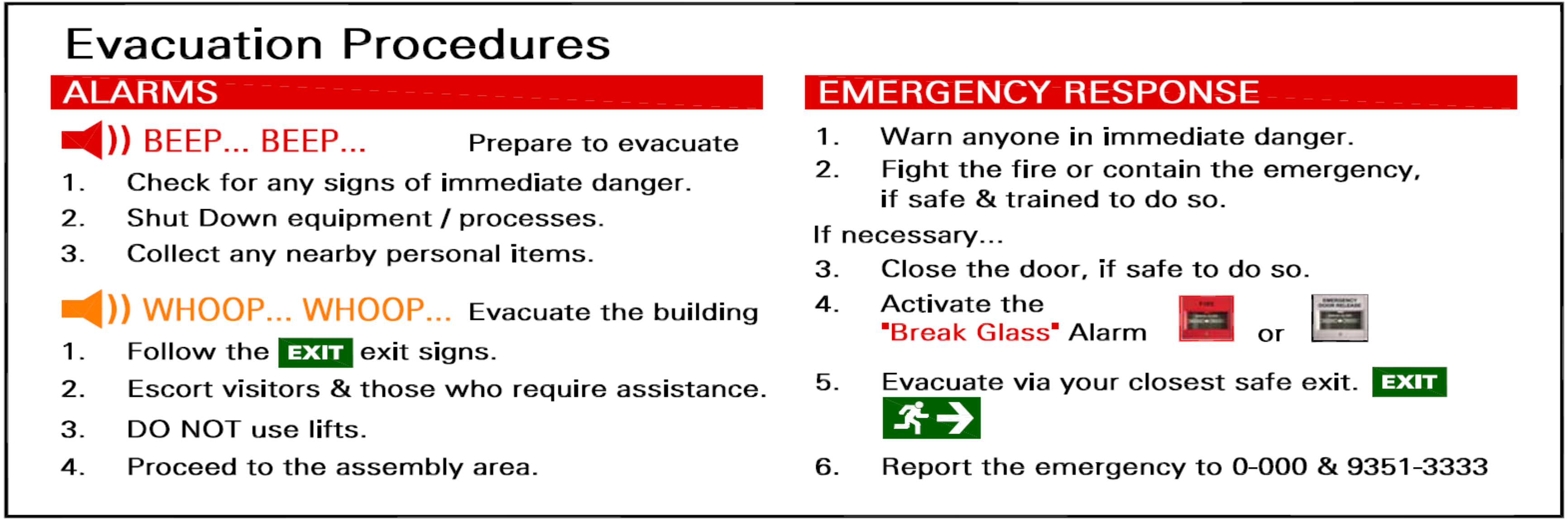
- Standard emergency response
- Alarms
- Emergency lockdown
- Medical emergencies
- Hazardous material incidents

NEED ASSISTANCE?

If you would like more information about these emergency procedures, contact your WHS Adviser.

DISABILITY SUPPORT SERVICES

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

Exercise: Survey of skills and interests

https://goo.gl/8rMhBB

(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 3 exercises!

WHAT IS DATA ANALYTICS?

Data analytics is the process of building intelligent systems to derive knowledge from data and make decisions

Data Analytics Skills

Data scientists/analysts help organisations:

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

Data

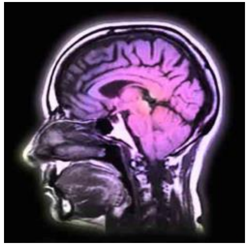
– How is the data generated?

Credit card swipes

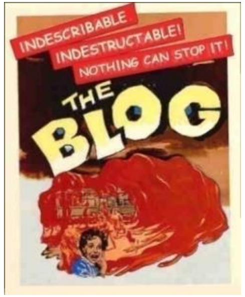
RFID tags

Digital video surveillance

E-mails



Radiology scans



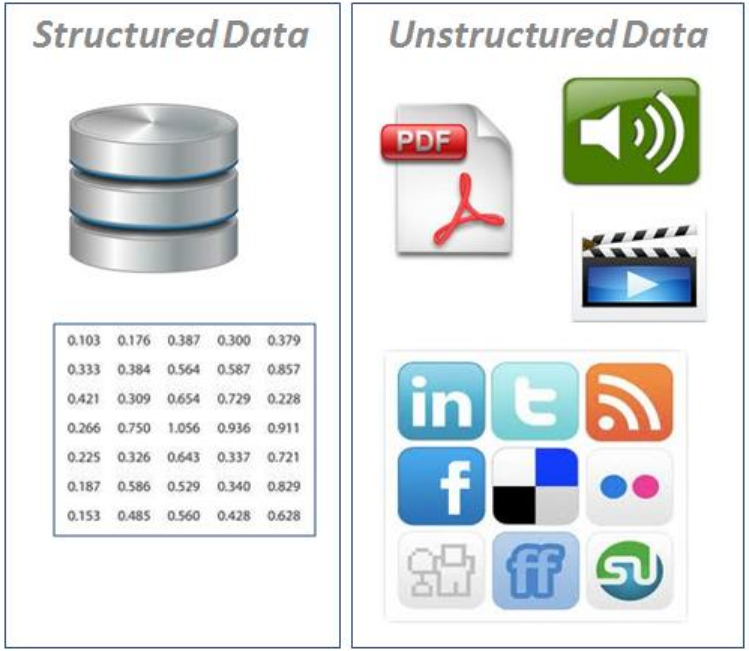
Blogs & Internet

Beacons & IoT

Other channels support

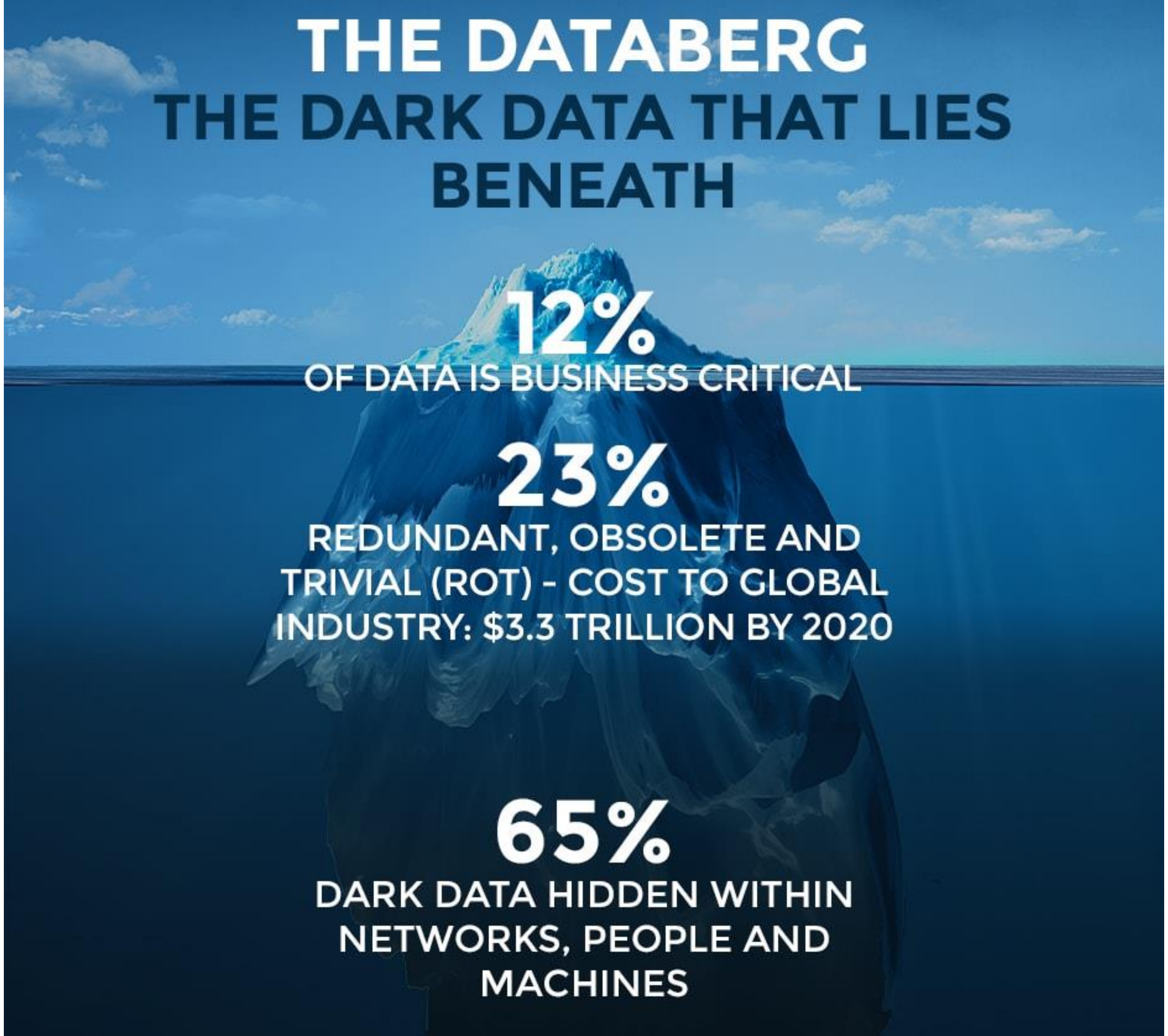
Data

– Types of data



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

https://www.laserfiche.com/content/uploads/2015/05/unstructured_data.png



THE DATABERG

THE DARK DATA THAT LIES BENEATH

12%

OF DATA IS BUSINESS CRITICAL

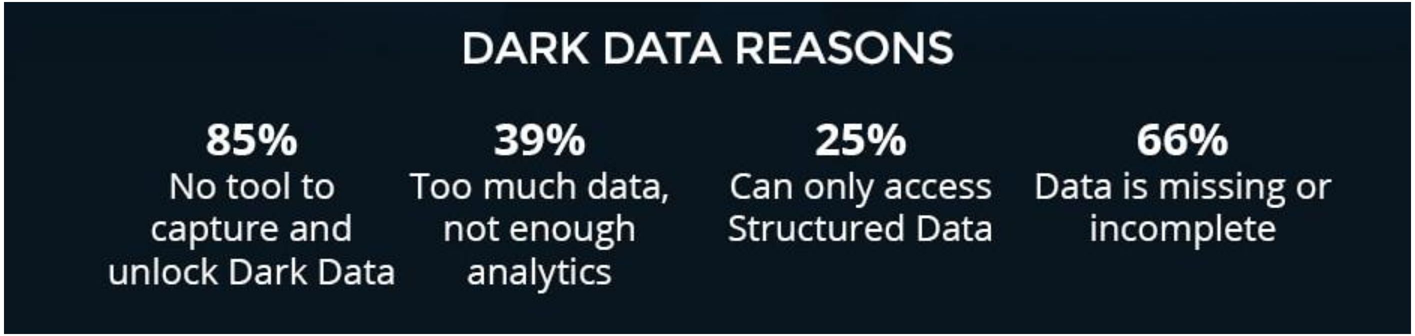
23%

REDUNDANT, OBSOLETE AND TRIVIAL (ROT) - COST TO GLOBAL INDUSTRY: $3.3 TRILLION BY 2020

65%

DARK DATA HIDDEN WITHIN NETWORKS, PEOPLE AND MACHINES

https://datumize.com/evolution-dark-data/



DARK DATA REASONS

85%
No tool to capture and unlock Dark Data

39%
Too much data, not enough analytics

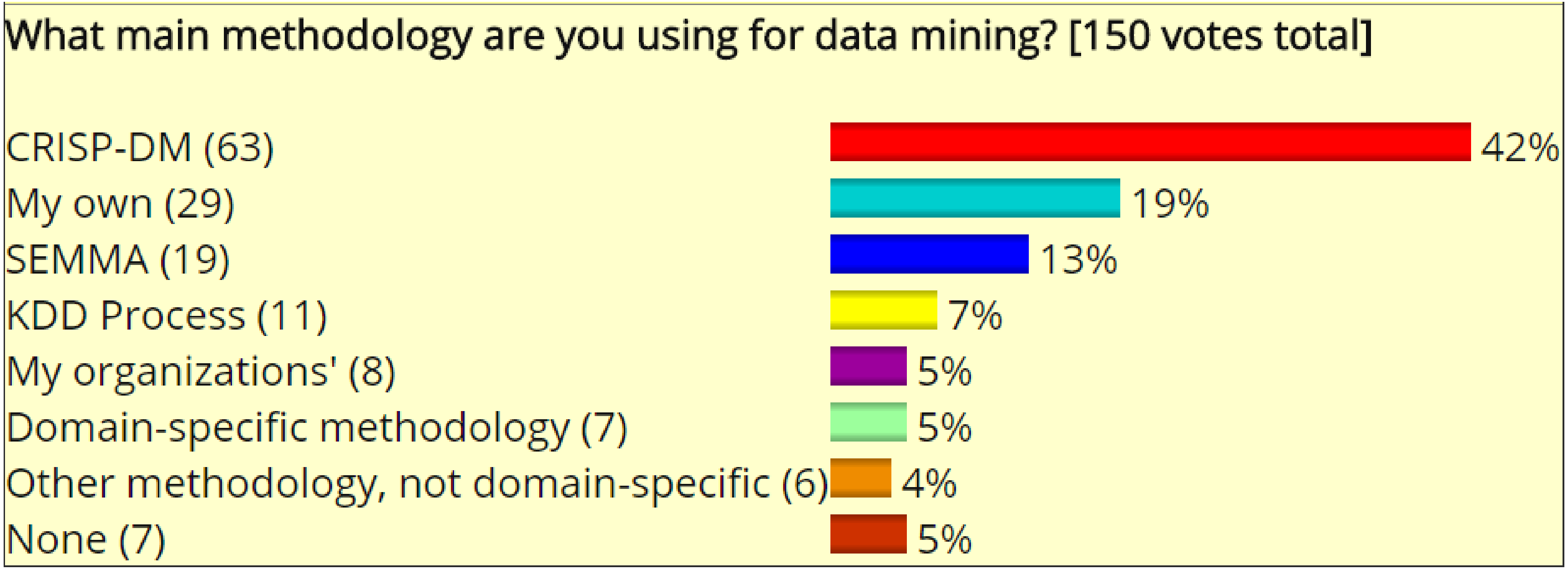
25%
Can only access Structured Data

66%
Data is missing or incomplete

Methodologies for Data Mining

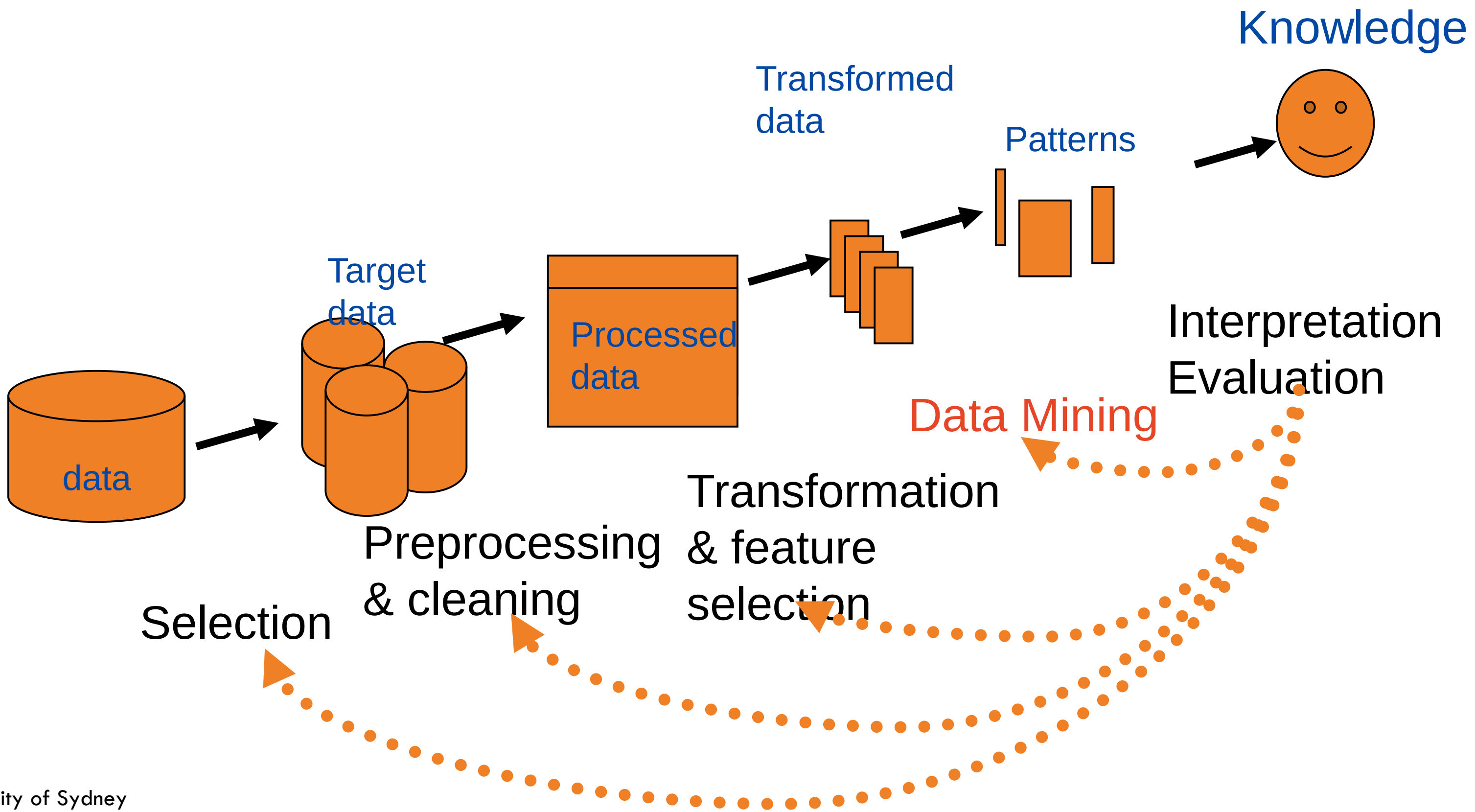
- Several methodologies have been developed, each with their own perspective.
- We will discuss three of them:
  - KDD Process
  - Sample, Explore, Modify, Model, and Assess (SEMMA)
  - Cross Industry Standard Process for Data Mining (CRISP-DM)

Methodology Poll Results from KD Nuggets

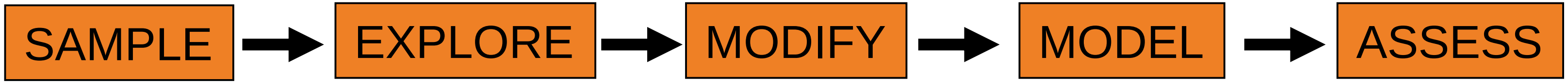


https://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

KDD Methodology



SEMMA Methodology



SAMPLE → EXPLORE → MODIFY → MODEL → ASSESS

Input data,
Sampling,
Data partition

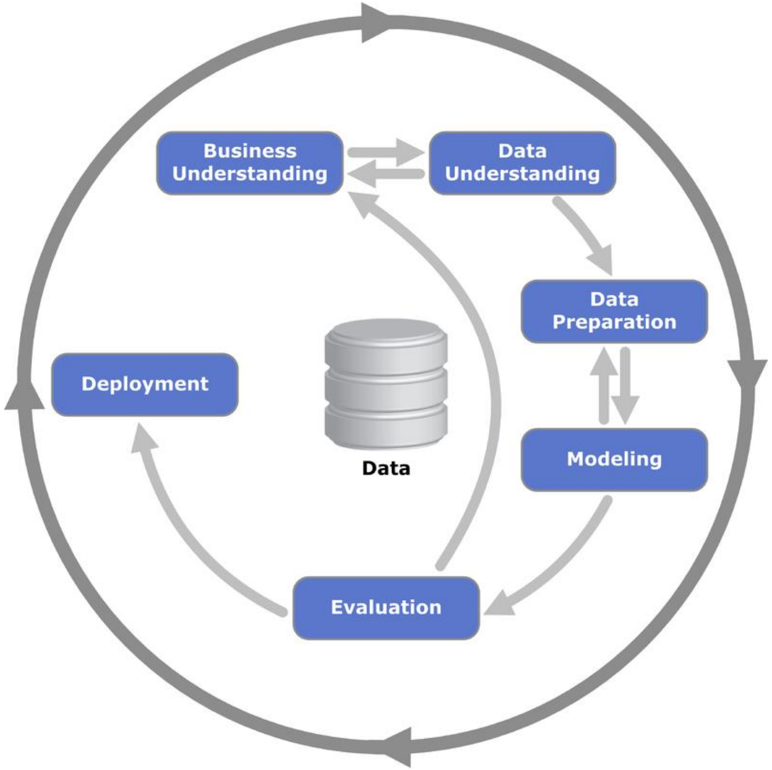
Distribution explorer,
Multiplot,
Insight,
Association,
Variable selection

Transform variable,
Filter outliers,
Clustering

Regression,
Tree,
Neural Network,
Ensemble

Assessment,
Score,
Report

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 mining goals
- Project plan
  - List assumptions and risk (technical/financial/business/ organisational) 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
- 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
    - requirements for data preparation
  - 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
    - Missing Values
      - Understand its source: Missing vs Null values
    - Strange Distributions

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

- 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

- 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

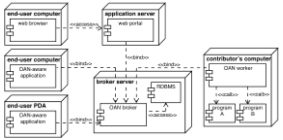
- Model to predict the purity of the environment based on carbon level
- Model to classify a person whether he is cheating in his tax return or not.
- Model to find hidden patterns and association rules in he basket market analysis
- Model to detect anomalies or outliers such as spam emails

The Evaluation Phase

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

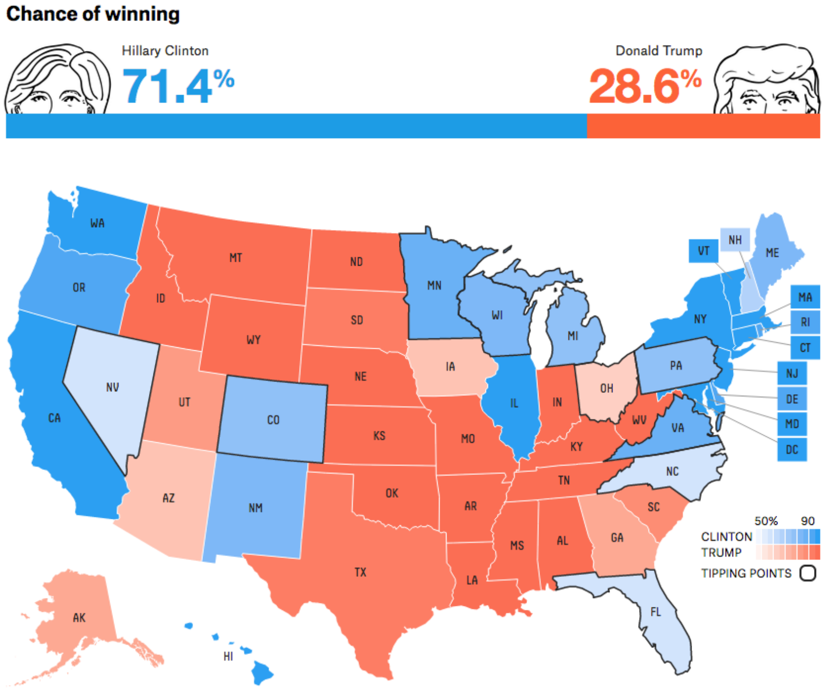
The Deployment Phase

- Knowledge Deployment is specific to objectives
  - Knowledge Presentation
  - Deployment within Scoring Engines and Integration with the current IT infrastructure
    - Automated pre-processing of live data feeds
  - Generation of a report
    - Online/Offline
  - Monitoring and evaluation of effectiveness



DATA ANALYTICS PROJECTS

Prediction of election outcomes - 2016



https://projects.fivethirtyeight.com/2016-election-forecast/

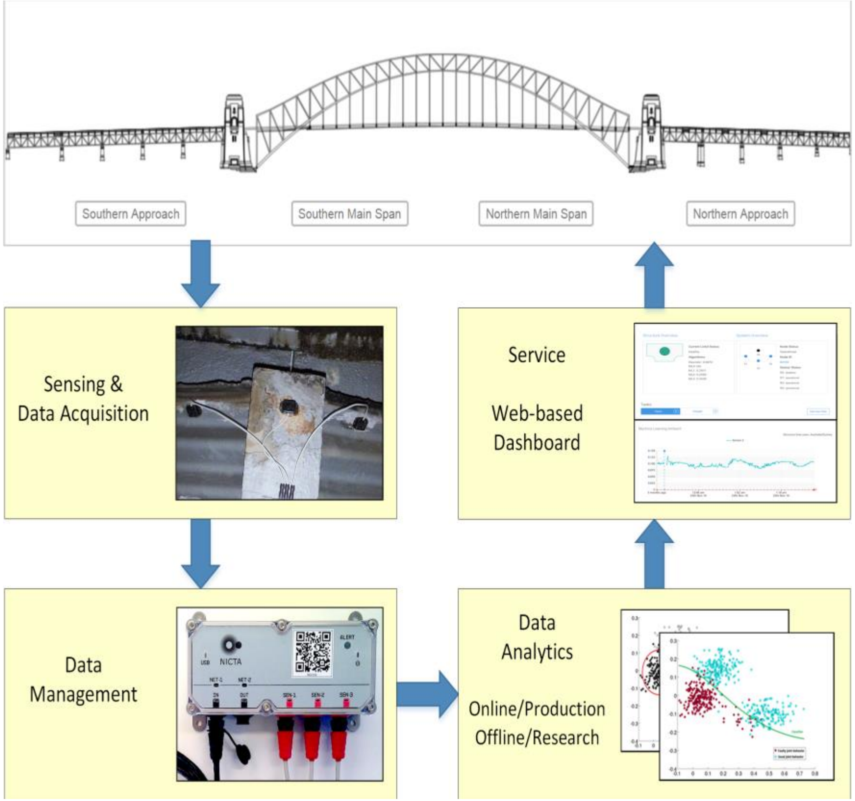
- National polls are a bad predictor of election outcomes
- 538 accounts for:
  - electoral vote allocation
  - weighting pollsters
  - decaying average
  - etc.

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



Southern Approach

Southern Main Span

Northern Main Span

Northern Approach

Sensing & Data Acquisition

Service

Web-based Dashboard

Data Management

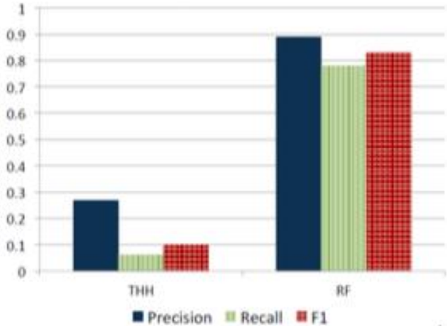
Data Analytics

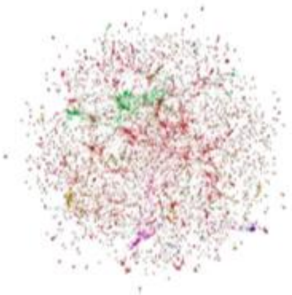
Online/Production

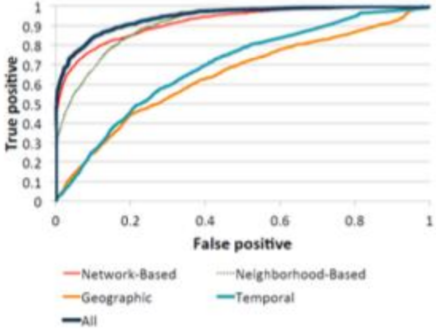
Offline/Research

- 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





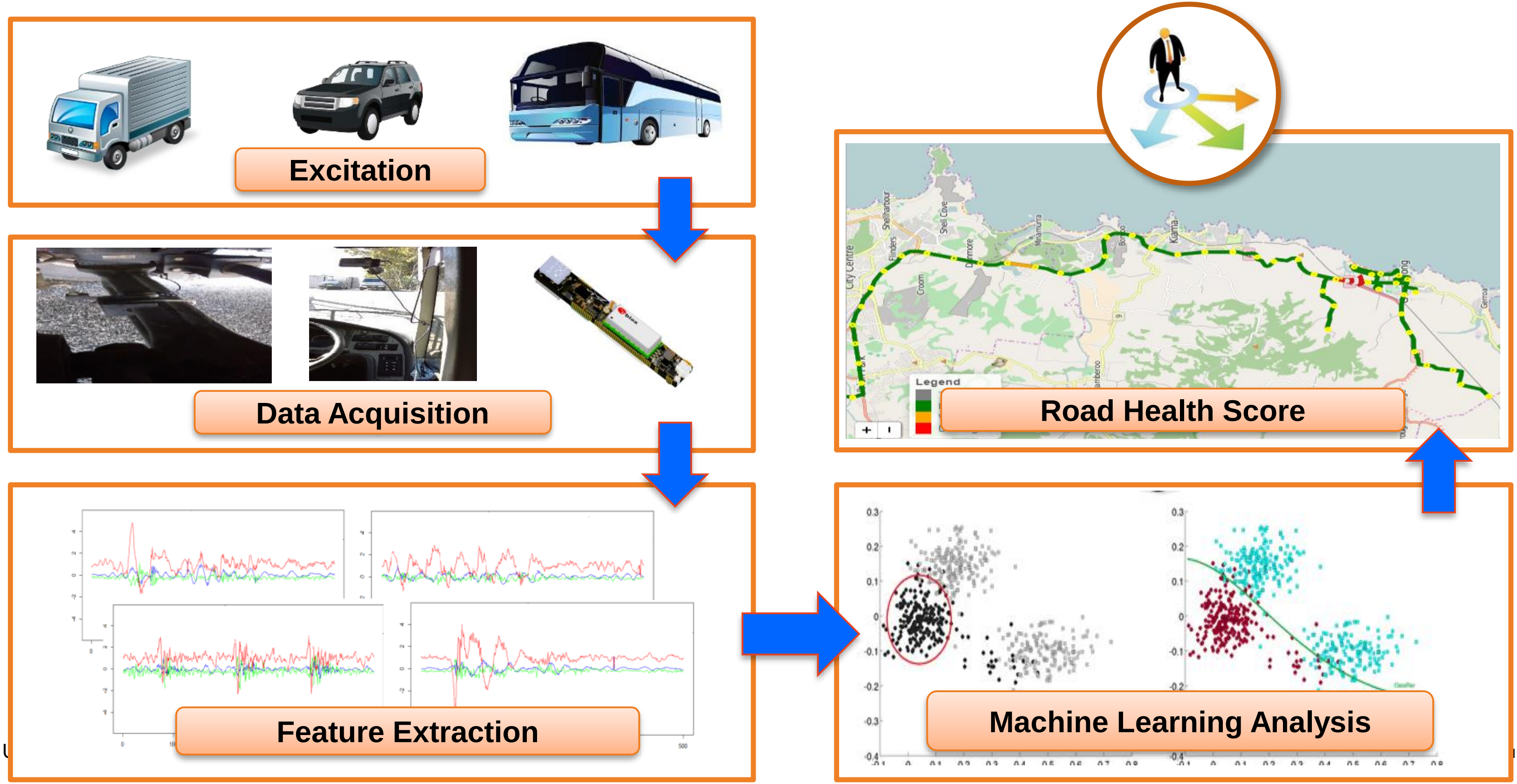


http://arxiv.org/pdf/1508.03965v1.pdf

- 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"

Example: Road Condition Assessment from Vehicle-mounted Sensor



WHERE DO I GET DATA?

Datasets

- Governments Open Data
  - AU: http://data.gov.au/dataset
  - Chile: http://datos.gob.cl/
  - US: https://www.data.gov/
  - India: https://data.gov.in/

- Kdnuggets: http://www.kdnuggets.com/datasets/index.html
- UCL MLR: https://archive.ics.uci.edu/ml/datasets.html

- When selecting, be aware of
- Type of data
  - Unstructured
  - Semi Structure
  - Structure
- Lenght

Project

You will have to use the CRISP-DM methodology to achieve a complete process

Project Stage 1: Obtain data, clean it and load and summarize.
Business Understanding
Data Understanding

- Collect Data
- Data Description
- Data Exploration

Project Stage 2: Develop and test a predictive model.
Modeling
Evaluation

Project Stage 3: presentation of results.

Project

- The Project is summative
  - Each stage depends of the other
- The report has to explain everything and every step taken, the problems you encountered and how you solved them. Be detailed.
- In the report is recommended to use figures and plots to help understanding
- Write down in the report the source of every information you use (remember plagiarism)

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 INFO3406 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: 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 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

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!

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!**