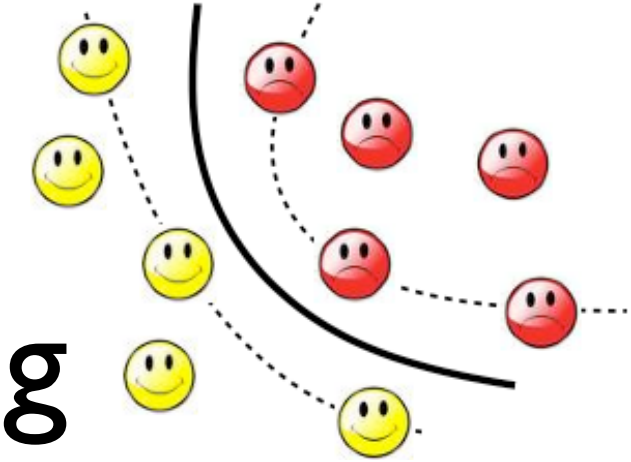
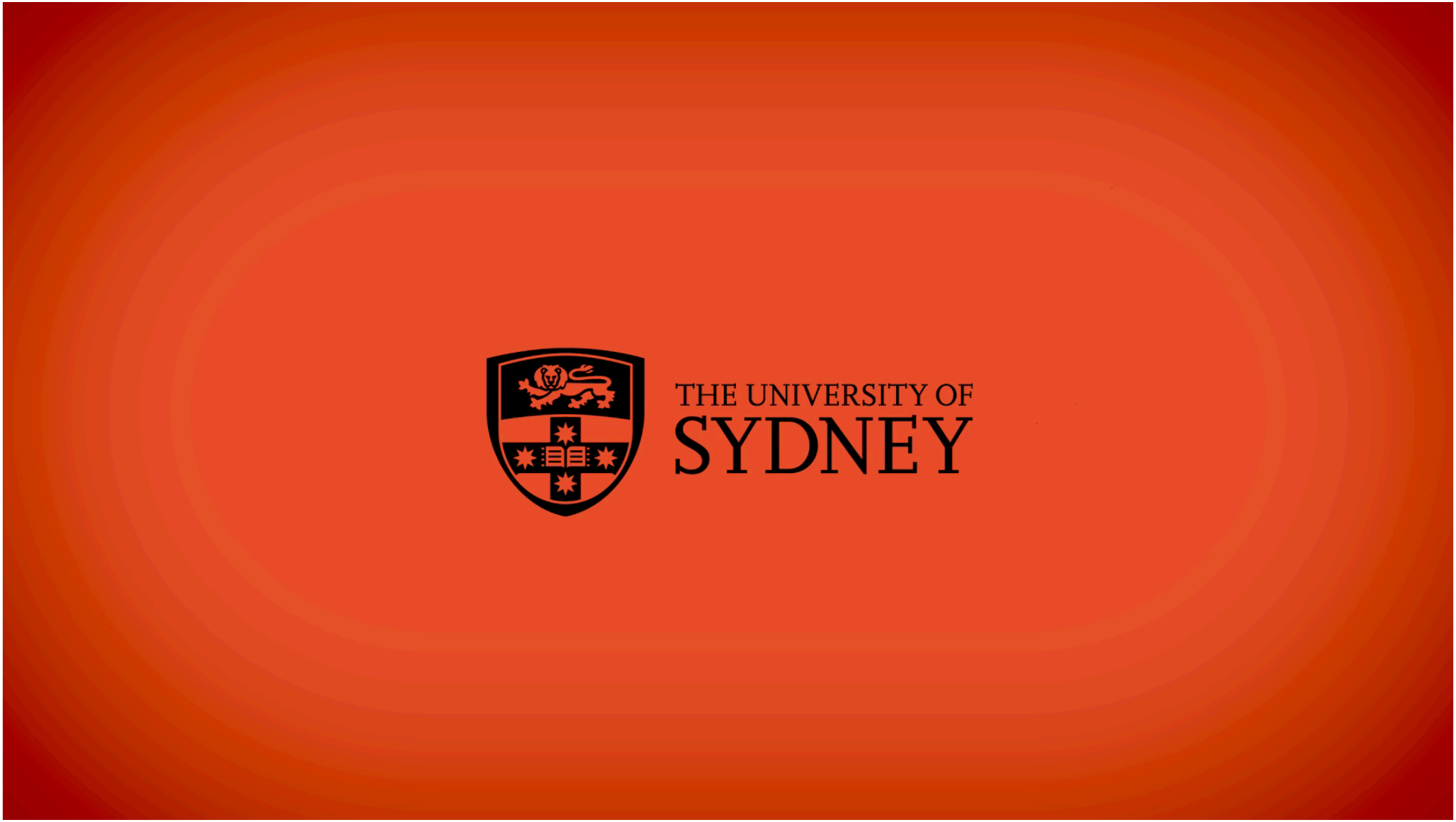


Machine Learning and Data Mining COMP 5318

School of Information Technologies

Fabio Ramos
Lionel Ott
Roman Marchant


THE UNIVERSITY OF
SYDNEY

WHS INDUCTION

School of Information Technologies

THE UNIVERSITY OF SYDNEY

General Housekeeping – Use of Labs

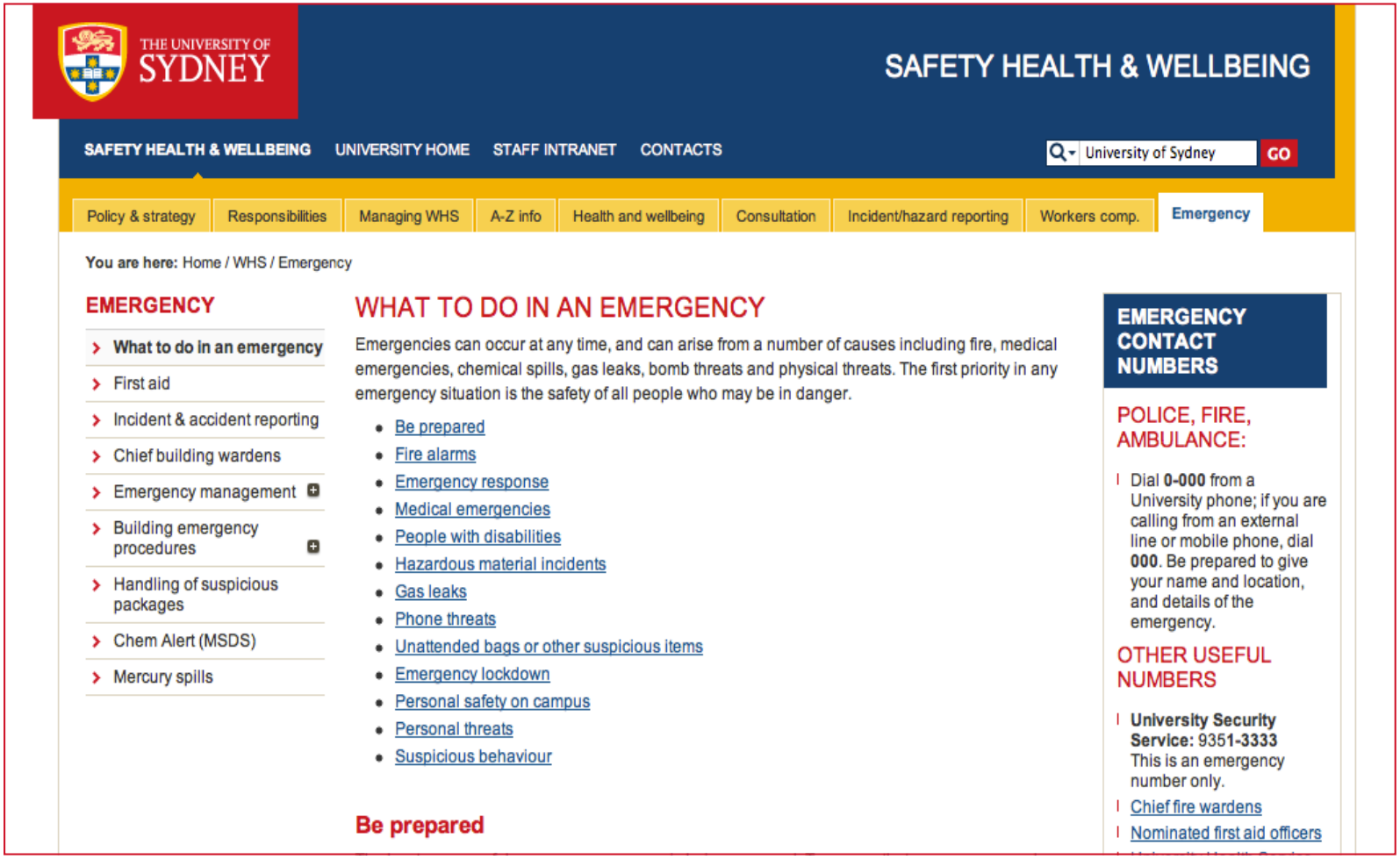
- Keep work area clean and orderly
- Remove trip hazards around desk area
- No food and drink near machines
- No smoking permitted within University buildings
- Do not unplug or move equipment without permission



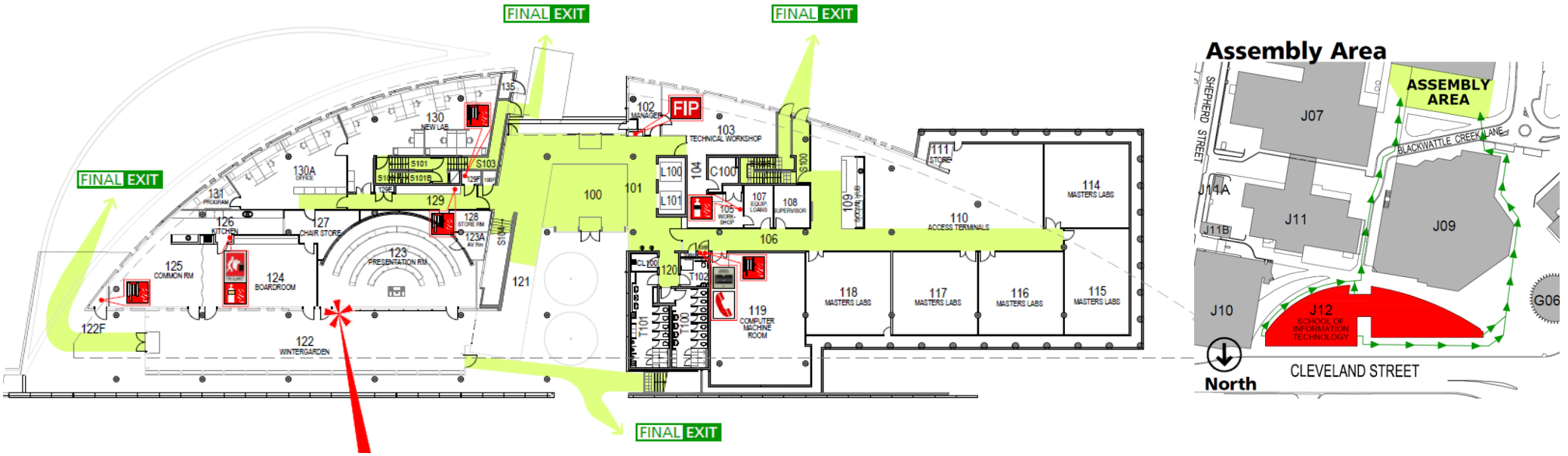


EMERGENCIES – Be prepared

www.sydney.edu.au/whs/emergency



WHERE IS YOUR CLOSEST SAFE EXIT ?



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

MEDICAL EMERGENCY

› If a person is seriously ill/injured:

1. **call an ambulance 0-000**
2. **notify the closest Nominated First Aid Officer**

   If unconscious– send for Automated External Defibrillator (AED)

   AED locations.

   NEAREST to SIT Building (J12)

   - Electrical Engineering Building, L2 (ground) near lifts
   - Seymour Centre, left of box office
   - Carried by all Security Patrol vehicles

3. **call Security -** 9351-3333
4. **Facilitate the arrival of Ambulance Staff (via Security)**

Nearest Medical Facility
University Health Service in Level 3, Wentworth Building

First Aid kit – SIT Building (J12)
kitchen area adjacent to Lab 110

School of IT Safety Contacts

CHIEF WARDEN
Name: Greg Ryan
Mobile:

FIRST AID OFFICERS

Name: Will Calleja
Location: 1 West
Phone: 9036 9706

Name: Katie Yang
Location: 2E-227
Phone: 9351 4918

Orally REPORT all INCIDENTS & HAZARDS to your SUPERVISOR

OR

Undergraduates: to Katie Yang 9351 4918

Coursework Postgraduates: to Cecille Faraizi 9351 6060

SIT School Manager: Shari Lee 9351 4158

Places

- Lecture: every Monday, 6 to 8pm, in Institute Lecture Theatre 1 (H03)

- Labs: every Monday, 8 to 9pm, in SIT (Labs 115, 116, 117, 130b)
  every Tuesday, 5 to 6pm, in SIT (Labs 114, 115, 116, 118)
  - More labs to open soon

- Do not miss classes, except for illness, emergencies, etc
- Get help from staff if you feel you are falling behind

Team

Lecturers: A/Prof Fabio Ramos (coordinator)

fabio.ramos@sydney.edu.au

Dr Lionel Ott (lecturer)

Dr Roman Marchant (lecturer)

Tutors: Philippe Morere (TA) philippe.morere@gmail.com

Rafael de Oliveira, Gilad Francis, Sheila Carceres, Kelvin Hsu, Prasad Cheema, Dongang Wang

Consultation: Monday, 2-3pm, SIT, Level 5, 547

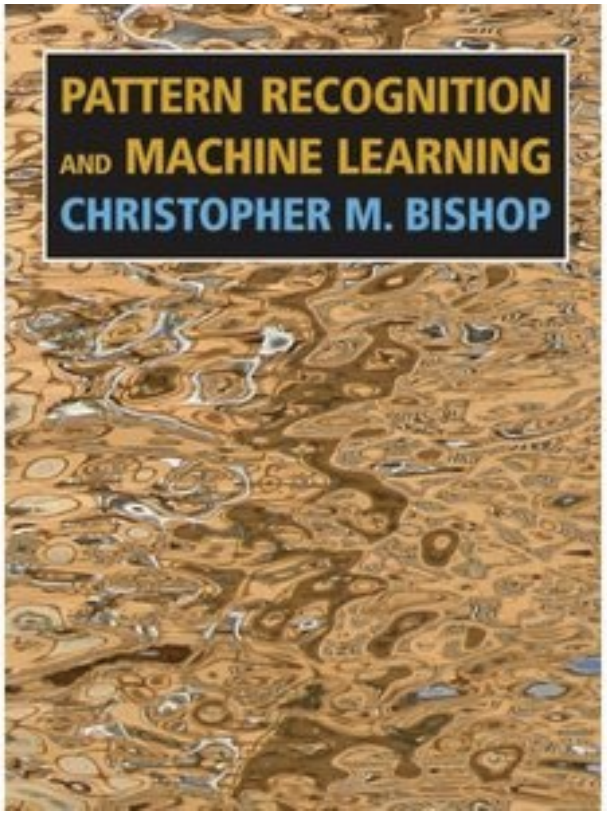
THE UNIVERSITY OF SYDNEY

Resources

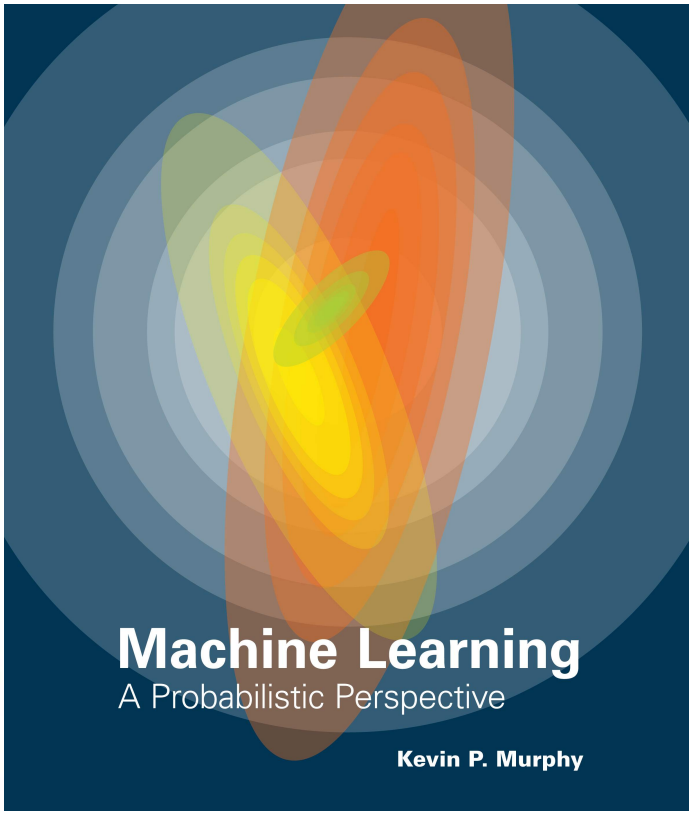
- eLearning
  - Login using Unikey and password
  - CUSP
    - Official schedule, list of learning outcomes, etc
  - Copies of slides
  - Lab instructions
  - Assignment instructions
  - Lecture videos
    - We intend to record the lectures, but the technology is not reliable
  - *Submit official assignments in eLearning*;
  - see your grades; etc
  - Discussion forum: on edstem, link in eLearning site

THE UNIVERSITY OF SYDNEY

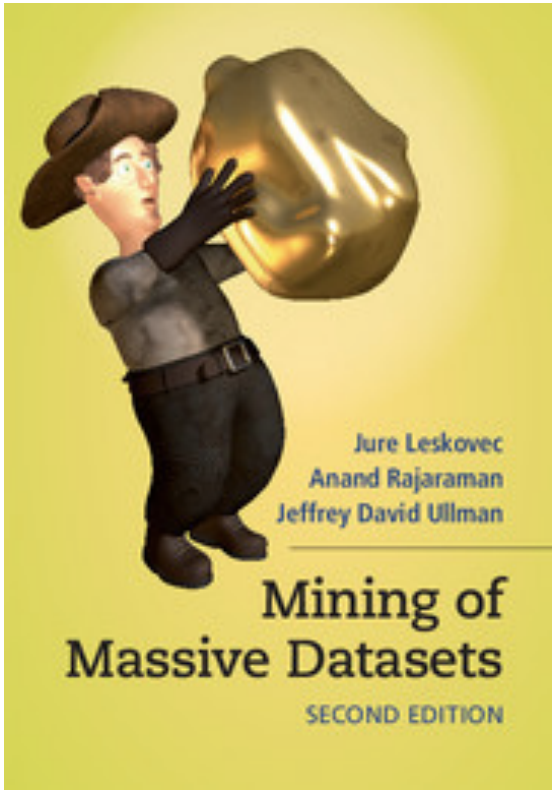
Best books available



Textbook



Textbook



Extra

THE UNIVERSITY OF SYDNEY

Assessment overview

- Quiz: **10%**
  - Week 4 (27/03) in-class, 6-8pm
  - Individual
  - Linear algebra + probability theory
- Assignment 1: **20%**
  - **8 May 2017,** 5:00pm
  - Groups of 2 or 3 students
  - Classification task
- Assignment 2: **20%**
  - **5 June 2017,** 5:00pm
  - Groups of 2 or 3 students
  - Method comparison and analysis
- **Final exam: 50%**
  - **June 2017 (date to be defined)**

THE UNIVERSITY OF SYDNEY

Assessment overview

In order to pass the course, the School requires at least 40% in the written exam, at least 40% in the other assessment components together and an overall final mark of 50 or more. This means that students who score less than 40% in the exam will fail the course regardless of their marks during the semester.

THE UNIVERSITY OF SYDNEY

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
  - practice and self-assess
- Students are responsible learners
  - Participate in classes, constructively
    - Respect for one another (criticise ideas, not people)
    - Humility: none of us knows it all; each of us knows valuable things
  - Check eLearning site at least once a week!
  - Notify academics whenever there are difficulties
  - Notify group partners honestly and promptly about difficulties

THE UNIVERSITY OF SYDNEY

Topics

Week	Date	Topic
1	6 March	Introduction, Admin matters, Research
2	13 March	Basic matrix analysis, SVD
3	20 March	Basic probability theory and Bayes' Rule
4	27 March	In-class Quiz
5	3 April	Classification and ROC curve
6	10 April	Logistic Regression and SVMs
7	24 April	Linear Regression
8	1 May	Clustering and EM
9	8 May	Latent Linear Models
10	15 May	Outlier detection
11	22 May	Neural Nets and large scale ML
12	29 May	Deep Learning
13	5 June	Industry talk and review

THE UNIVERSITY OF SYDNEY

Format of the lectures

- 10-15 min review from previous week
- 1h-1h15min of new content
- 5-10 min of examples
- 15-30 min of research topics

Assumed knowledge

- Linear algebra, calculus
- Basics of probability theory
- Programming skills

Labs: Python

- Python is a high-level programming language designed to enforce good coding practices.
- Interactive and very natural to use.
- Extremely versatile and excellent for prototyping.
- Great libraries for machine learning eg. scikit-learn, TensorFlow, Keras, Edward

www.python.org

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 (3 days)
  - http://sydney.edu.au/current_students/special_consideration/
- Also, notify the TA by email *as soon as anything begins to go wrong*
- There is a similar process if you need special arrangements eg for religious observance, military service, representative sports

Academic Integrity (University Policy)

- Please read the University policy on Academic Honesty carefully: http://sydney.edu.au/elearning/student/EI/academic_honesty.shtml
- All cases of academic dishonesty and plagiarism will be investigated
- There is a new process and a centralized University system and database
- Three types of offenses:
  - Plagiarism – when you copy from another student, website or other source. This includes copying the whole assignment or only a part of it.
  - Academic dishonesty – when you make your work available to another student to copy (the whole assignment or a part of it). There are other examples of academic dishonesty.
  - Misconduct - when you engage another person to complete your assignment (or a part of it), for payment or not. This is a very serious matter and the Policy requires that your case is forwarded to the University Registrar for investigation.

Academic Integrity (University Policy)

- The penalties are severe and include:
  1) a permanent record of academic dishonesty, plagiarism and misconduct in the University database and on your student file
  2) mark deduction, ranging from 0 for the assignment to Fail for the course
  3) expulsion from the University and cancelling of your student visa

- Do not confuse legitimate co-operation and cheating! You can discuss the assignment with another student, this is a legitimate collaboration, but you cannot complete the assignment together – everyone must write their own code or report, unless the assignment is group work.
- When there is copying between students, note that both students are penalised – the student who copies and the student who makes his/her work available for copying

Academic Integrity (University Policy)

- We will use the similarity detection software TurnItIn and MOSS to compare your assignments with these of other students (current and previous) and the Internet
  - Turnitin is for text documents: http://www.turnitin.com/en_us/higher-education
  - MOSS is for programming code: https://theory.stanford.edu/~aiken/moss/
- These tools are extremely good!
  - e.g. MOSS cannot be fooled by changing the names of the variables or changing the order of the conditions in `if-else` statements
- Examples of plagiarism in programming code:
  - http://www.upenn.edu/academicintegrity/ai_computercode.html

THE UNIVERSITY OF SYDNEY

Academic Integrity (University Policy)

- All these are cases of plagiarism and academic dishonesty we have seen in our school and the student excuses are not acceptable:
  - *I sat the test and then posted the questions and solutions to my friends whose test was later in the week. I only wanted to help them understand the concepts that are examinable.*
  - *I posted parts of my code on my web page (group discussion forum) because my solution was cool (or I wanted to help them). I didn't expect them to copy it.*
  - *I tried to do the assignment on my own but I had problems with the extension part that I couldn't fix, so I submitted my core part and his extension part. I didn't cheat.*
  - *I finished my assignment but my friend had family problems. I felt sorry for her, so I gave her my assignment as an example. She said she only wanted to have a look and promised not to copy it.*
  - *The test has finished but the tutor hasn't collected the papers yet. I showed my answer to my friend. I didn't expect him to copy it.*
  - *He is my best friend. I had no choice but to let him copy my assignment.*

Academic Integrity (University Policy)

- Plagiarism and any form of academic dishonesty will be dealt with, and the penalties are severe
- We use plagiarism detection systems such as MOSS and TurnItIn that are extremely good. If you cheat, the chances you will be caught are very high.
- If someone asks you to see or copy your assignment, or to complete the assignment instead of them, just say: *I can't do this. This is against the University policy. I will not risk my future by doing this*.

Be smart and don't risk your future by engaging in plagiarism and academic dishonesty!

THE UNIVERSITY OF SYDNEY

Assistance

- There are a wide range of support services available for students
- Please make contact, and get help
- You are not required to tell anyone else about this
- If you are willing to inform the unit coordinator or TA, they may be able to work with other support to reduce the impact on this unit
  - eg provide advice on which tasks are most significant

Disability Services

Do you have a disability?

- You may not think of yourself as having a 'disability' but the definition under the **Disability Discrimination Act** is broad and includes temporary or chronic medical conditions, physical or sensory disabilities, psychological conditions and learning disabilities.
- The types of disabilities we see include:

- anxiety, arthritis, asthma, asperger's disorder, ADHD, bipolar disorder, broken bones, cancer, cerebral palsy, chronic fatigue syndrome, crohn's disease, cystic fibrosis, depression, diabetes, dyslexia, epilepsy, hearing impairment, learning disability, mobility impairment, multiple sclerosis, post traumatic stress, schizophrenia , vision impairment, and much more.

- Students needing assistance must register with Disability Services –
  - it is advisable to do this as early as possible.
- http://sydney.edu.au/study/academic-support/disability-support.html

Other support

- 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 Islanders
  - 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/ or http://www.supra.net.au/
- Please make contact, and get help
- You are not required to tell anyone else about this
- If you are willing to inform the unit coordinator, they may be able to work with other support to reduce the impact on this unit
  - eg provide advice on which tasks are most significant

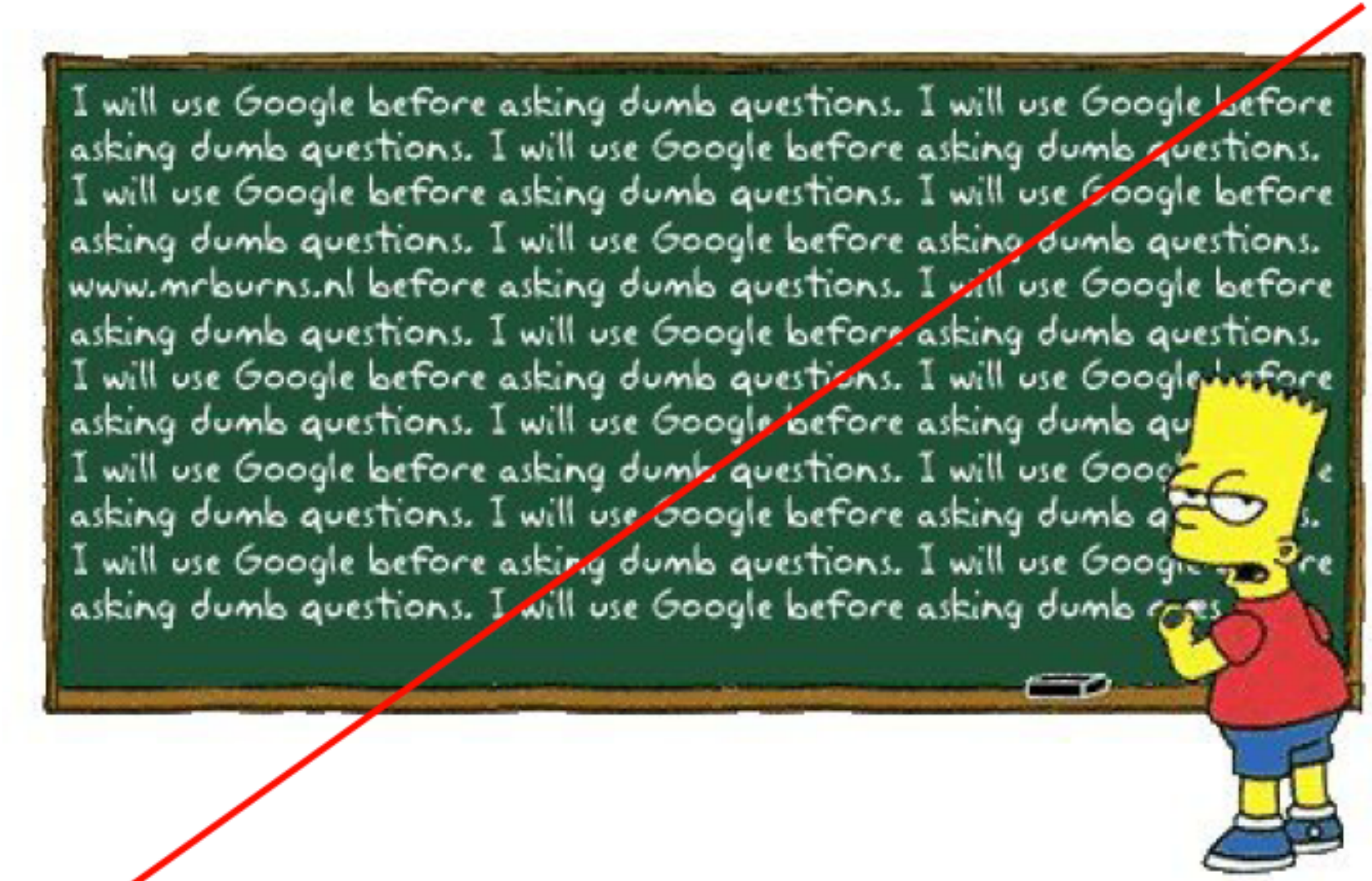
THE UNIVERSITY OF SYDNEY

Advice

- Metacognition
  - Pay attention to the learning outcomes in CUSP
  - Self-check that you are achieving each one
  - Think how each assessment task relates to these
- Time management
  - Watch the due dates
  - Start work early, submit early
- Networking and community-formation
  - Make friends and discuss ideas with them
  - Know your TA, tutor, lecturer, coordinator
  - Keep them informed, especially if you fall behind
    - Don’t wait to get help
- Enjoy the learning!

THE UNIVERSITY OF SYDNEY

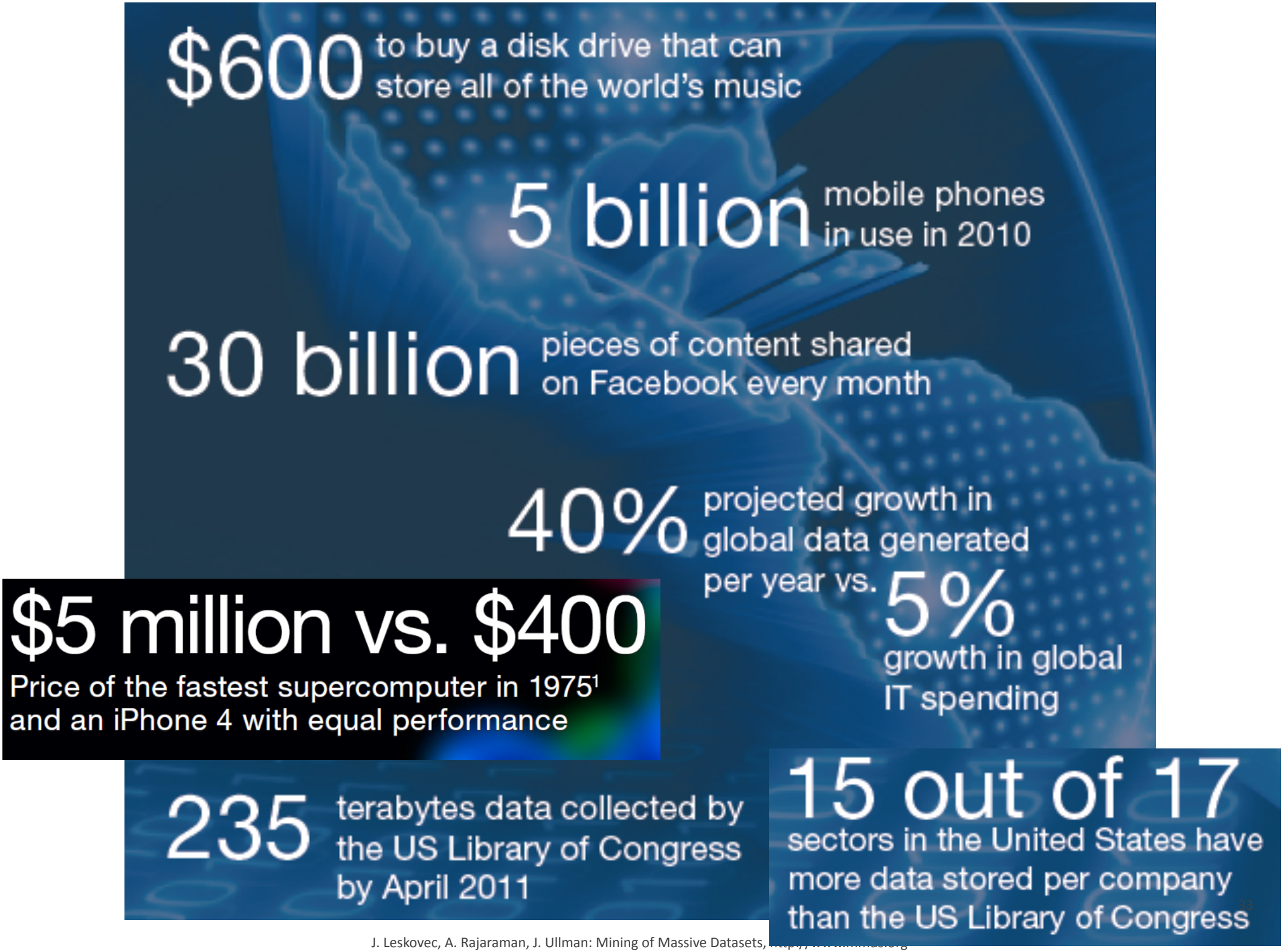
Ask questions sooner!



What is Machine Learning?

Informally: Knowledge discovery from data

Formally: The construction of a statistical model that is an underlying distribution from which the data is drawn from.


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
40% projected growth in global data generated per year vs. 5% growth in global IT spending
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17
sectors in the United States have more data stored per company than the US Library of Congress

THE UNIVERSITY OF SYDNEY

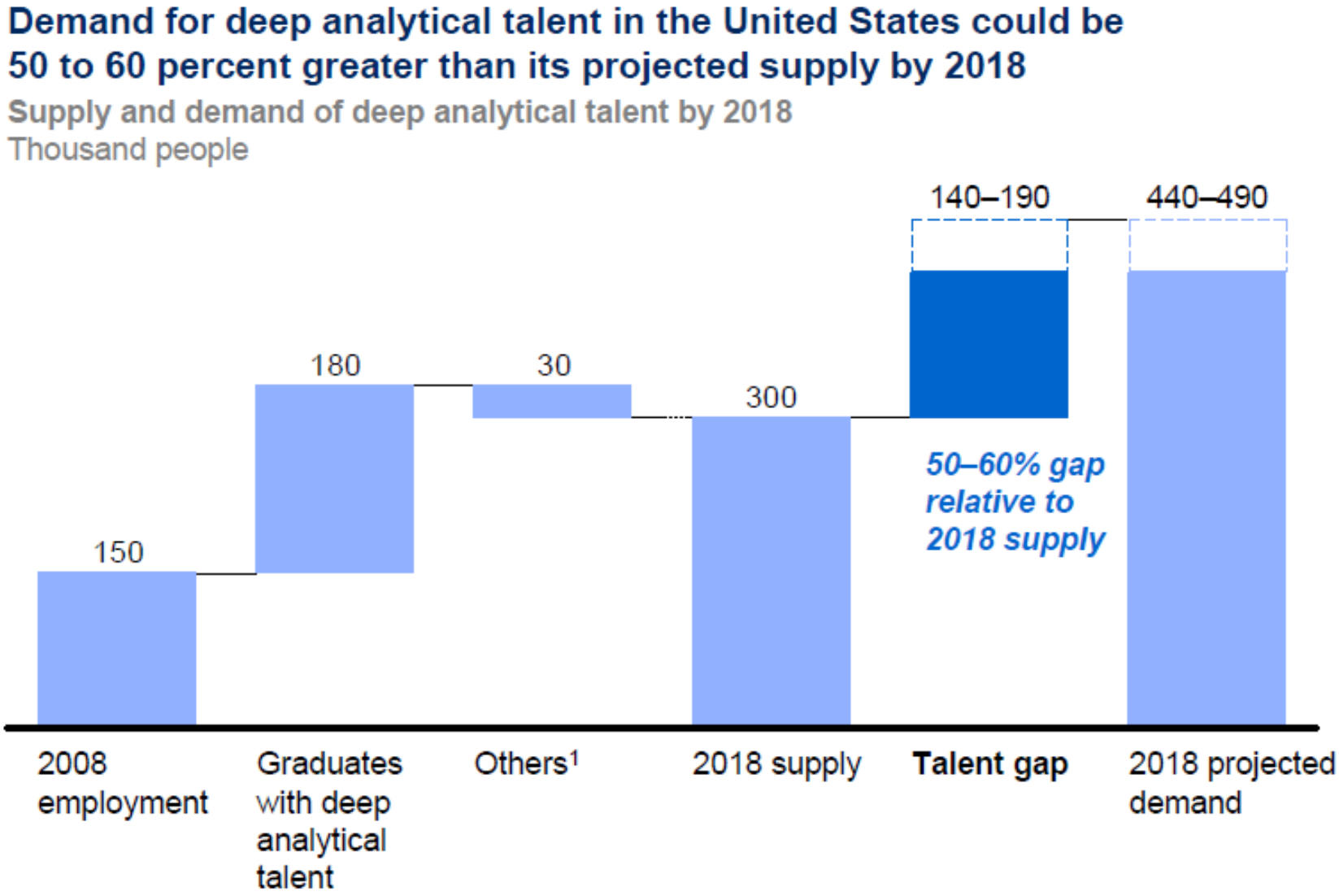
Data contains value and knowledge

ML / DM

- **But to extract the knowledge data needs to be**
  - **Stored**
  - **Managed**
  - **And ANALYSED ← this course**

Data Mining ≈ Big Data ≈ Statistics ≈ Machine Learning ≈ Data Science

Good news: Demand for Data Mining




Google


amazon.com


TOYOTA


SONY


SIEMENS


IBM
Research


Microsoft®
Research


hp
invent


HRI
Europe
Honda Research Institute


facebook
YAHOO!

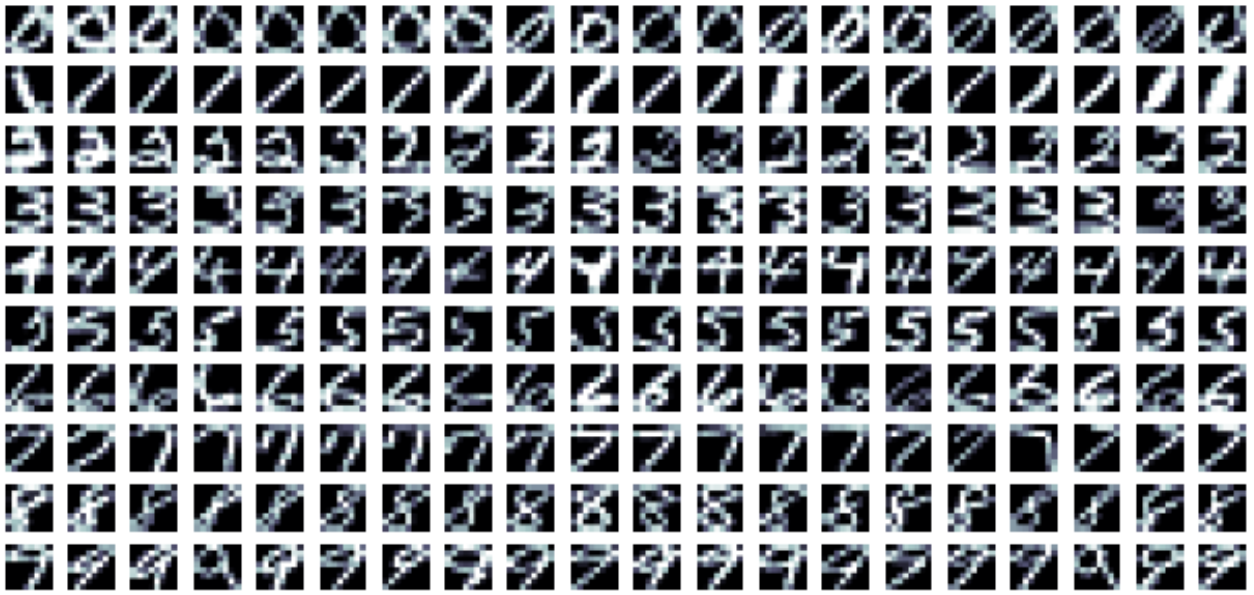

CommonwealthBank


RioTinto

THE UNIVERSITY OF SYDNEY

Speech recognition

Object and handwriting recognition



(NORB image from Yann LeCun)

Information retrieval


Google Search: Unsupervised Learning http://www.google.com/search?q=Unsupervised+Learning&sourceid=fir...

Google Web Images Groups News Froogle more »
Unsupervised Learning Search Advanced Search Preferences

Web Results 1 - 10 of about 150,000 for Unsupervised Learning. (0.27 seconds)

Mixture modelling, Clustering, Intrinsic classification ...
Mixture Modelling page. Welcome to David Dowe's clustering, mixture modelling and unsupervised learning page. Mixture modelling (or ...
www.csse.monash.edu.au/~dld/mixture.modelling.page.html - 26k - 4 Oct 2004 - Cached - Similar pages

ACL'99 Workshop -- Unsupervised Learning in Natural Language ...
PROGRAM. ACL'99 Workshop Unsupervised Learning in Natural Language Processing. University of Maryland June 21, 1999. Endorsed by SIGNLL. ...
www.aclweb.org/~kehler/unsup-acl-99.html - 5k - Cached - Similar pages

Unsupervised learning and Clustering
cgm.cs.mcgill.ca/~soss/cs644/projects/wljha/ - 1k - Cached - Similar pages

NIPS'98 Workshop - Integrating Supervised and Unsupervised ...
NIPS'98 Workshop "Integrating Supervised and Unsupervised Learning" Friday, December 4, 1998. ... 4:45-5:30, Theories of Unsupervised Learning and Missing Values. ...
www-2.cs.cmu.edu/~mccallum/supunsup/ - 7k - Cached - Similar pages

NIPS Tutorial 1999
Probabilistic Models for Unsupervised Learning Tutorial presented at the 1999 NIPS Conference by Zoubin Ghahramani and Sam Roweis. ...
www.gatsby.ucl.ac.uk/~zoubin/NIPStutorial.html - 4k - Cached - Similar pages

Gatsby Course: Unsupervised Learning : Homepage
Unsupervised Learning (Fall 2000). ... Syllabus (resources page): 10/10 1 - Introduction to Unsupervised Learning Geoff project: (ps, pdf). ...
www.gatsby.ucl.ac.uk/~quaid/course/ - 19k - Cached - Similar pages
[More results from www.gatsby.ucl.ac.uk]

[PDF] Unsupervised Learning of the Morphology of a Natural Language
File Format: PDF/Adobe Acrobat - View as HTML
Page 1. Page 2. Page 3. Page 4. Page 5. Page 6. Page 7. Page 8. Page 9. Page 10. Page 11. Page 12. Page 13. Page 14. Page 15. Page 16. Page 17. Page 18. Page 19 ...
acl.ldc.upenn.edu/J/J01/J01-2001.pdf - Similar pages

Unsupervised Learning - The MIT Press
... From Bradford Books: Unsupervised Learning Foundations of Neural Computation Edited by Geoffrey Hinton and Terrence J. Sejnowski Since its founding in 1989 by ...
mitpress.mit.edu/book-home.tcl?isbn=026258168X - 13k - Cached - Similar pages

[ps] Unsupervised Learning of Disambiguation Rules for Part of
File Format: Adobe PostScript - View as Text
Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging. Eric Brill. 1. ... It is possible to use unsupervised learning to train stochastic. ...
www.cs.jhu.edu/~brill/acl-wkshp.ps - Similar pages

The Unsupervised Learning Group (ULG) at UT Austin
The Unsupervised Learning Group (ULG). What ? The Unsupervised Learning Group (ULG) is a group of graduate students from the Computer ...
www.lans.ece.utexas.edu/ulg/ - 14k - Cached - Similar pages

Gooooooooooogle
Result Page: 1 2 3 4 5 6 7 8 9 10 Next

1 of 2 06/10/04 15:44

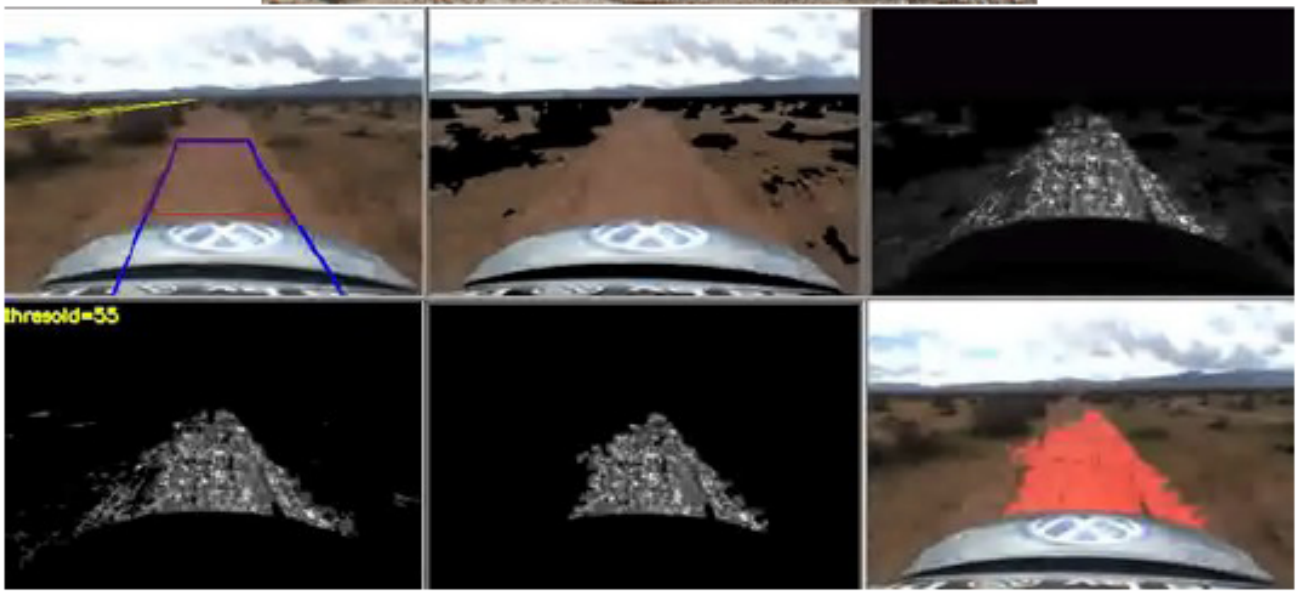
Web Pages

Retrieval
Categorisation
Clustering
Relations between pages

THE UNIVERSITY OF SYDNEY

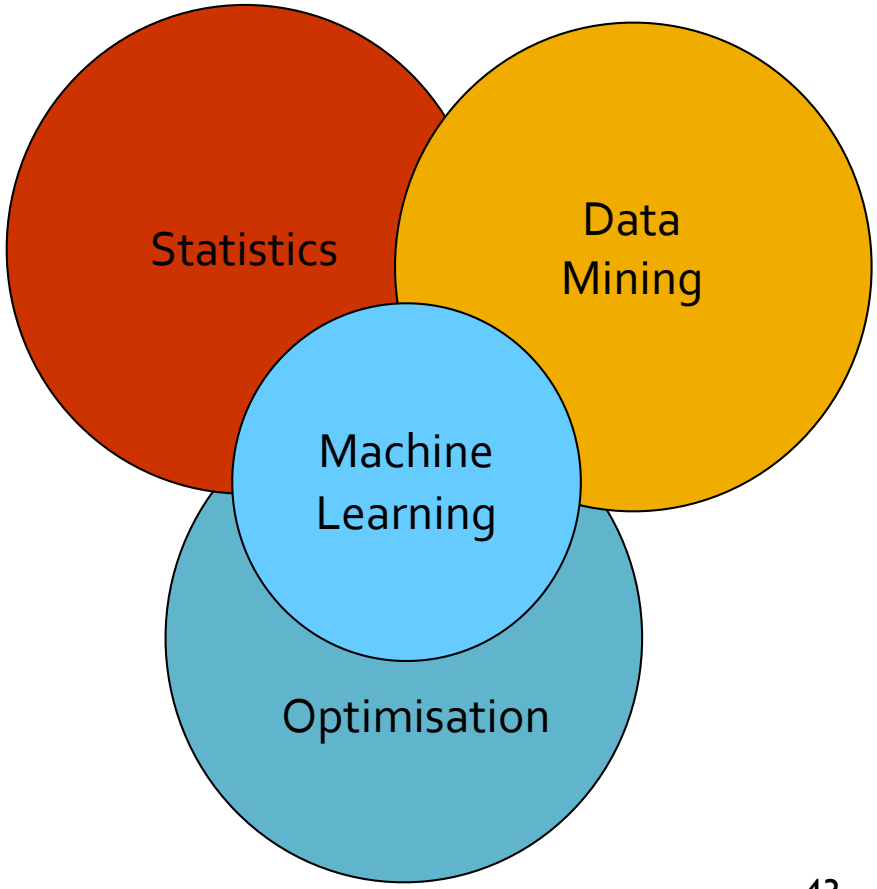
Financial prediction

Robotics



This Course: COMP5318

- **This course overlaps with statistics, artificial intelligence, databases but more stress on**
  - **Algorithms**
  - **Mathematical modelling**
  - Automation for handling **large data**



Machine Learning Problems

- Prediction
  - Classification and Regression
- Clustering, segmentation and summarisation
  - Find patterns in the data
- Outlier/anomaly detection
  - Find unusual patterns

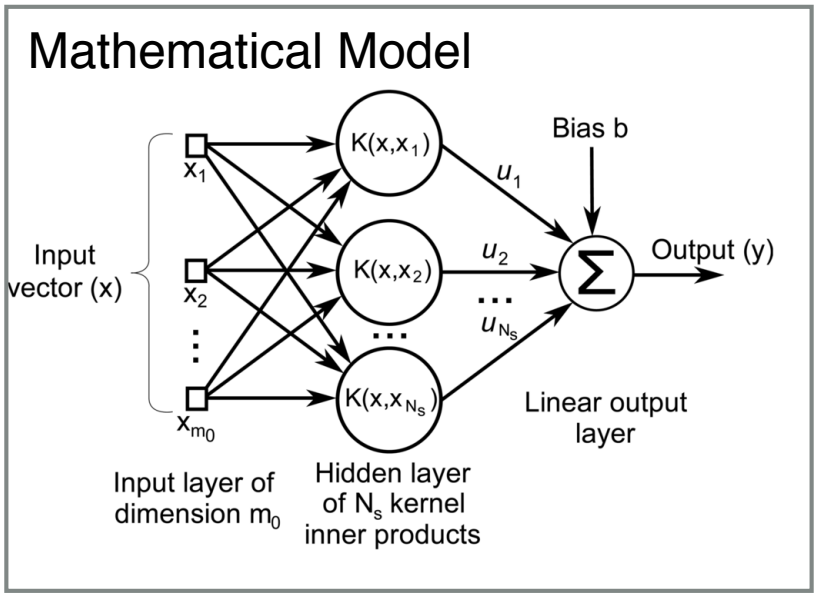
Elements of Machine Learning

THE UNIVERSITY OF SYDNEY

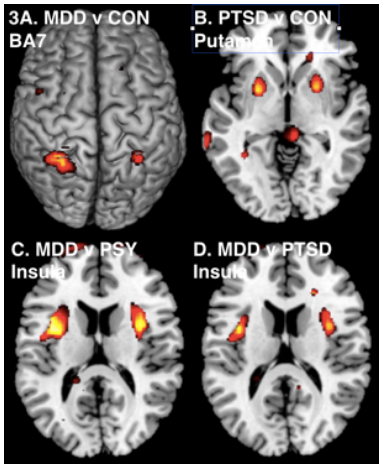
Group 1

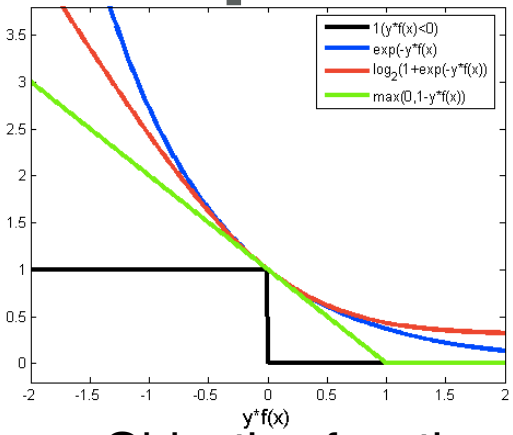
Group 2

Data

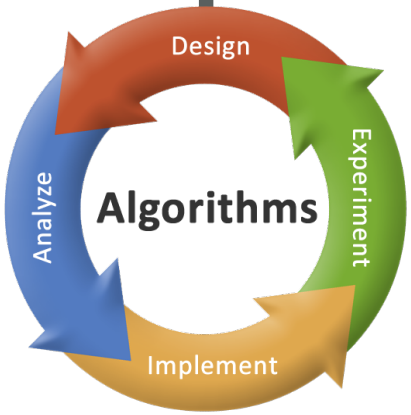


Predictions/Patterns





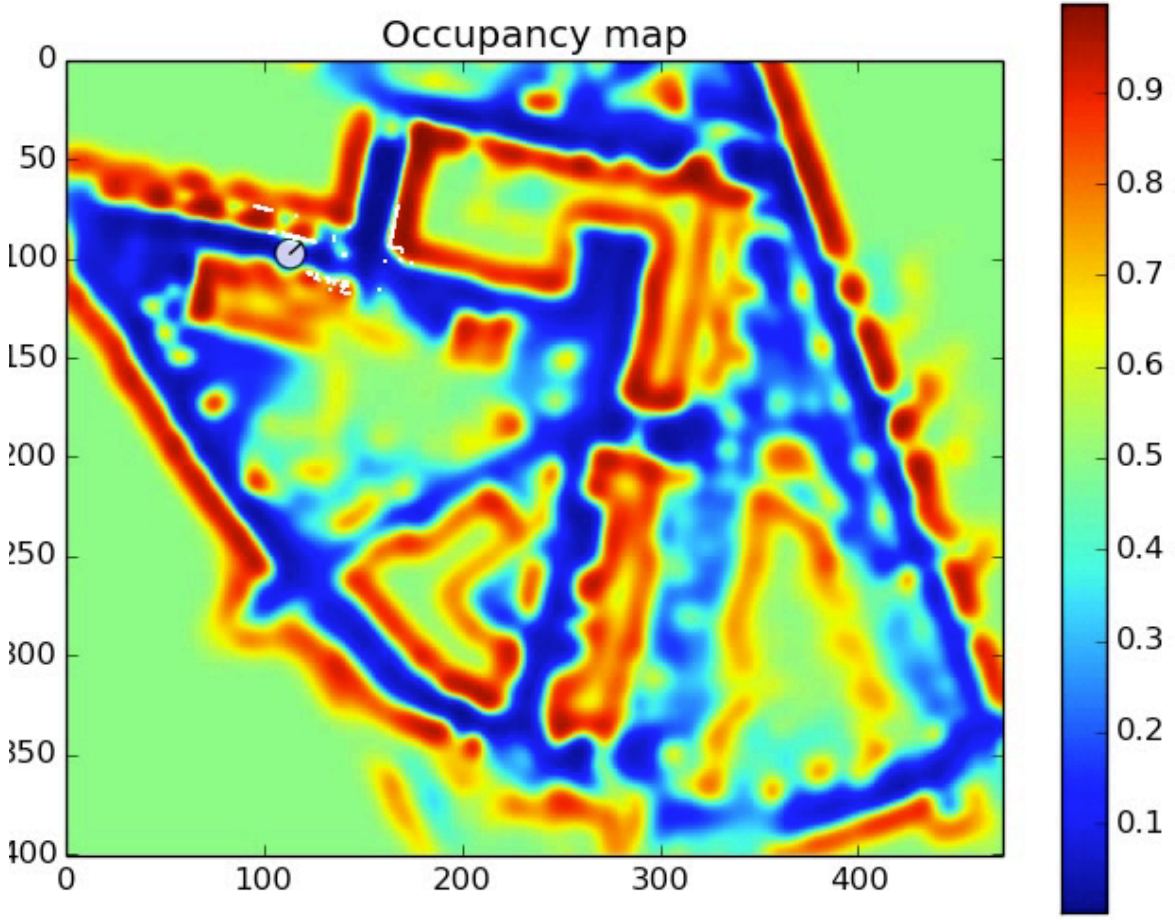
Objective function



THE UNIVERSITY OF SYDNEY

Classification

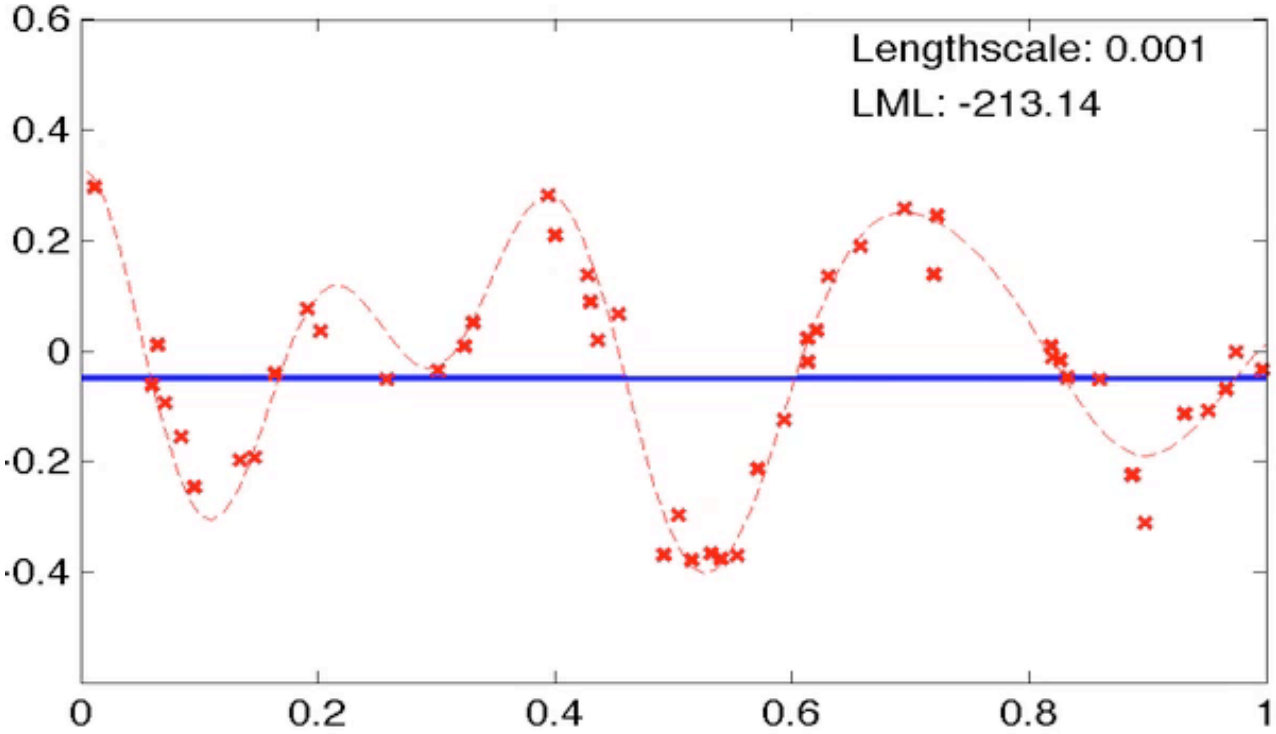
Aerial View



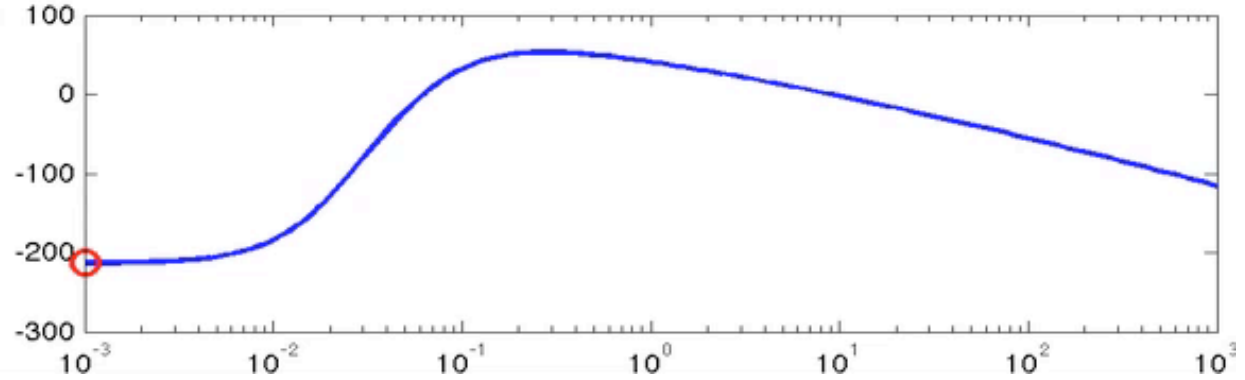
THE UNIVERSITY OF SYDNEY

Regression

Problem

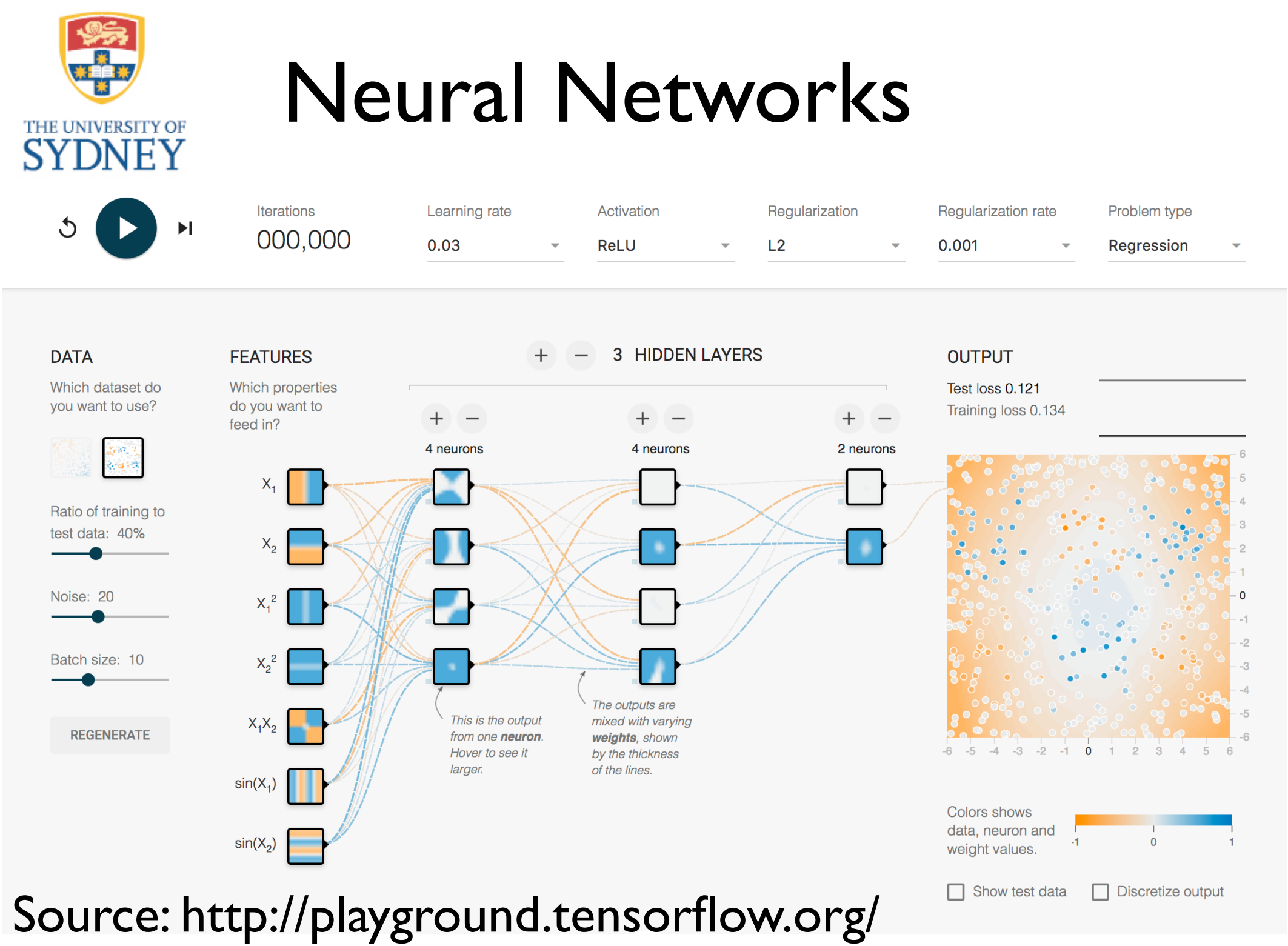


Objective



THE UNIVERSITY OF SYDNEY

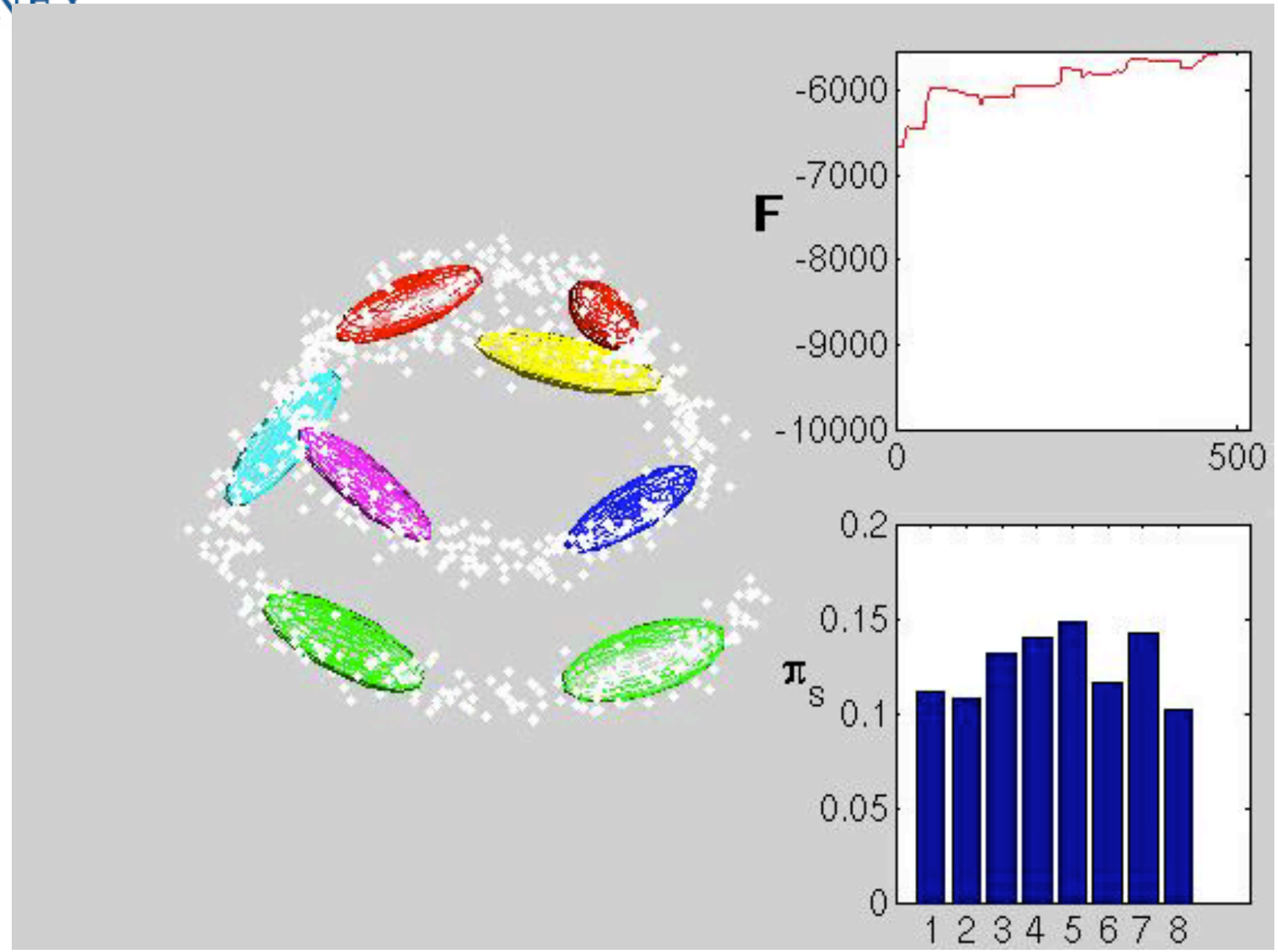
Neural Networks



Source: http://playground.tensorflow.org/

THE UNIVERSITY OF SYDNEY

Clustering



Common representation

IMAGE/ VIDEO

TEXT/ COMMENT

TIME SERIES

SYSTEM LOGS

NETWORK

TABULAR/ RATING

Is there a common way to represent data of different modalities ?

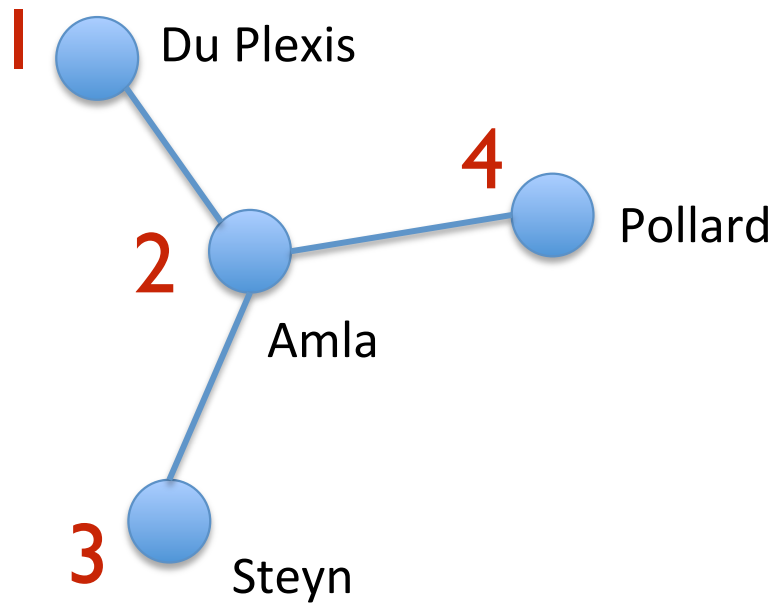
Text to matrix

- Document- Word Matrix
- Document 1:“AACCBBAAA”
- Document 2:“CCAABBDD”

$$\begin{pmatrix} A & B & C & D \\ 5 & 2 & 2 & 0 \\ 2 & 2 & 2 & 2 \end{pmatrix}$$

THE UNIVERSITY OF SYDNEY

Network data



$$\begin{array}{c} \\ \text{Nodes} \end{array} \begin{array}{c} \begin{array}{cccc} \text{Nodes} & & & \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Image data

www.sydney.visitorsbureau.com.au

700 x 500

4	45	6
6	12	33
22	17	44

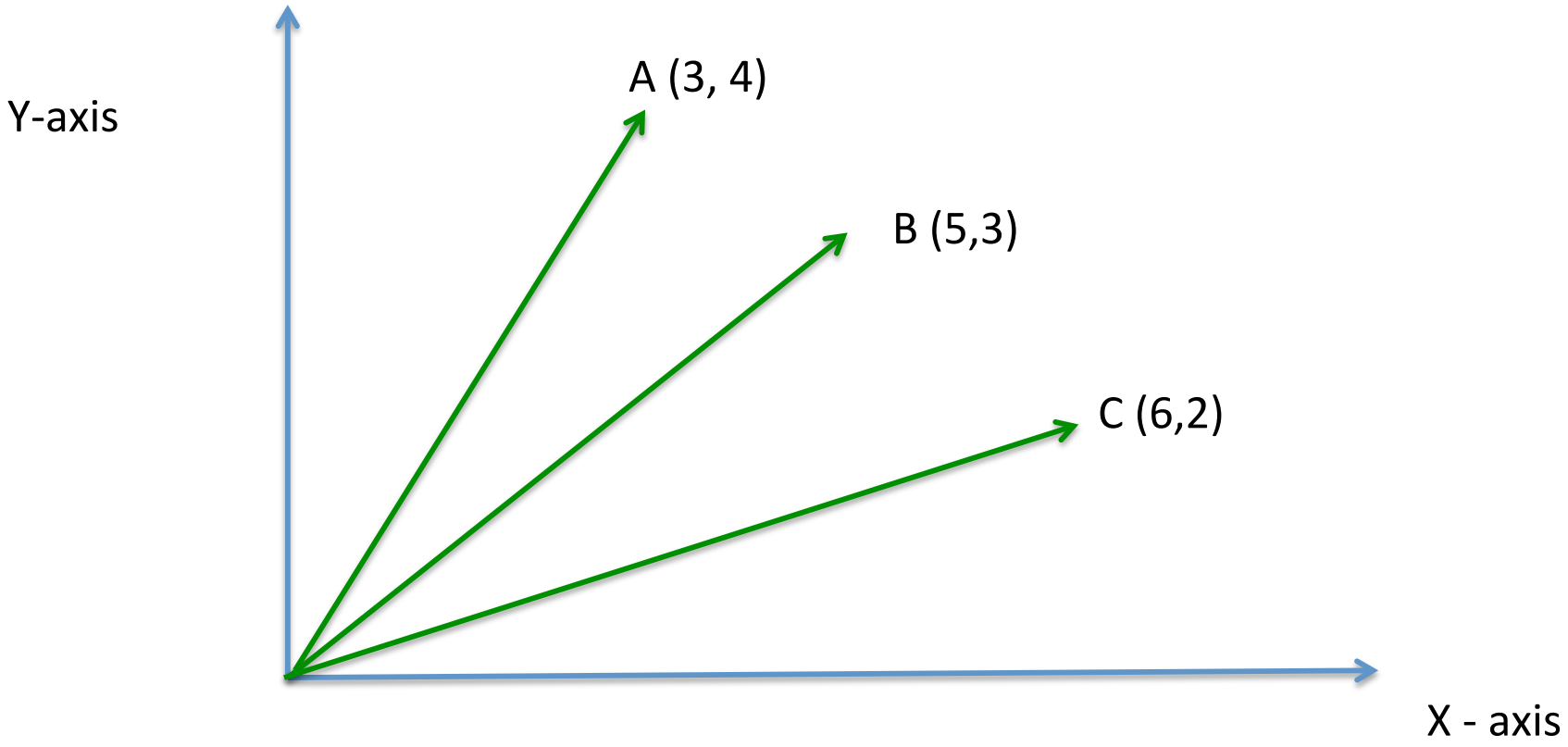
4	45	6	6	12	33	22	17	44

Similarity Computation

- We can now represent most data types as a matrix.
- A special case of a matrix is a vector.
- Now lets compute similarities with these objects.

Similarity Computation

How can we quantify similarity between A, B and C ?



THE UNIVERSITY OF SYDNEY

Similarity Computation

- Dot product
  $x = (x_1, x_2, \ldots, x_n)$; $y = (y_1, y_2, \ldots, y_n)$
  $x.y = (x_1y_1 + x_2y_2 + \ldots + x_ny_n)$

- Norm (length) of a vector
  $||x|| = (x.x)^{1/2} = (x_1x_1 + x_2x_2 + \ldots + x_nx_n)^{1/2}$

Similarity Computation

- The similarity between two vectors x and y is given by

  – sim(x,y) = x.y/(||x|| ||y||)

THE UNIVERSITY OF SYDNEY

Example

- Let x = <3,1,2,4>; y = <1,2,1,2>
- Step 1: Compute the dot-product
  - x.y = 3.1 + 1.2 + 2.1 + 4.2 = 15
- Step 2: Compute length of x vector
  - $||x|| = (3^2 + 1^2 + 2^2 + 4^2)^{0.5} = 5.477$
  - ||y|| = 3.162
- sim(x,y) = x.y/(||x|| ||y||) = 0.8660

Properties

- When is sim(x,y) = 0 ?
- When is sim(x,y) = 1 ?
- Can sim(x,y) < 0 ?
- Can sim(x,y) > 1 ?

THE UNIVERSITY OF SYDNEY

Image search engine

Object Recognition

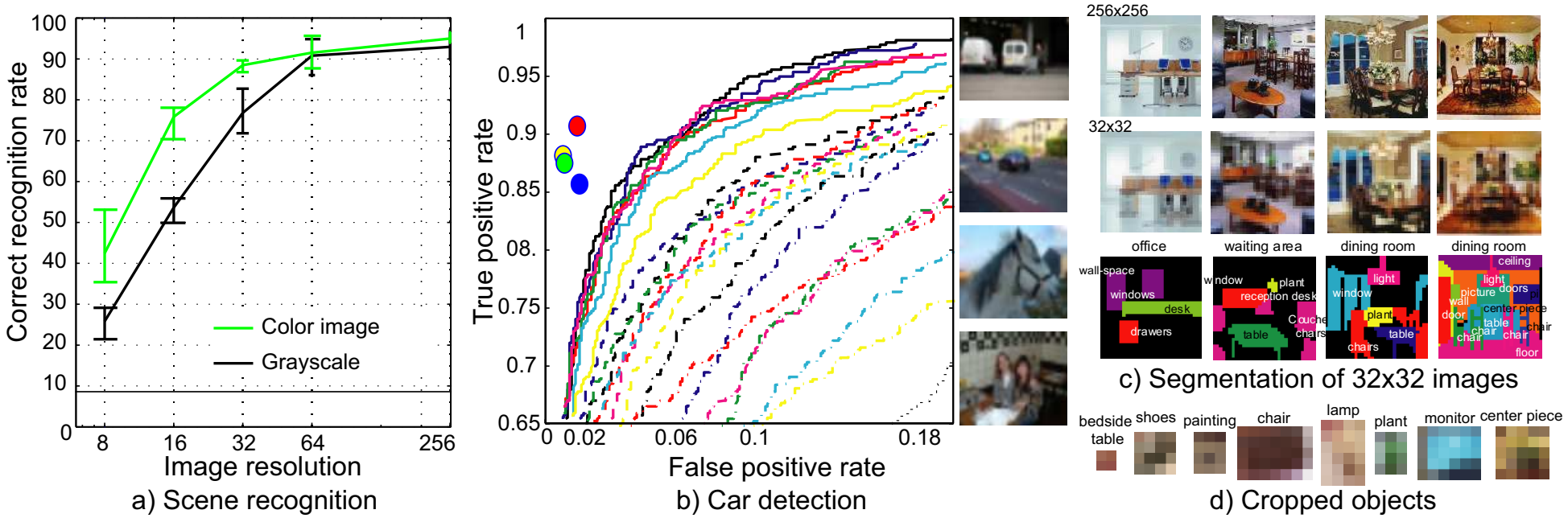


Fig. 1. a) Human performance on scene recognition as a function of resolution. The green and black curves show the performance on color and gray-scale images respectively. For color 32×32 images the performance only drops by 7% relative to full resolution, despite having 1/64th of the pixels. b) Car detection task on the PASCAL 2006 test dataset. The colored dots show the performance of four human subjects classifying tiny versions of the test data. The ROC curves of the best vision algorithms (running on full resolution images) are shown for comparison. All lie below the performance of humans on the tiny images, which rely on none of the high-resolution cues exploited by the computer vision algorithms. c) Humans can correctly recognize and segment objects at very low resolutions, even when the objects in isolation can not be recognized (d).

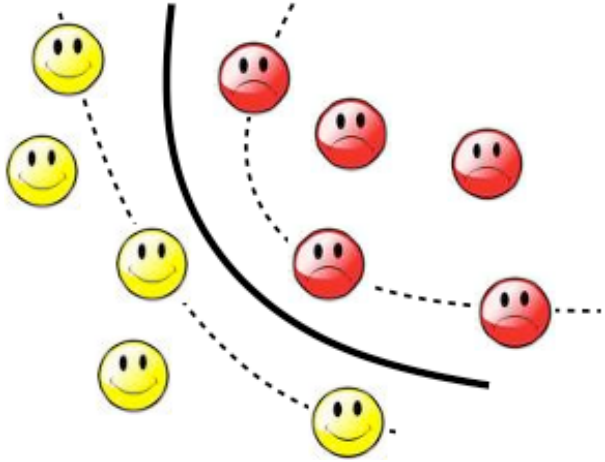
Torralba et al. 80 million tiny images: a large dataset for non-parametric object and scene recognition, PAMI 2008

Data Mining and Turing Test

Source: BBC Radio 4 - Analysis Podcast

THE UNIVERSITY OF SYDNEY

YouTube channels



My research