Introduction to Machine Learning and Data Mining

COMP5318/COMP4318 Machine Learning and Data Mining

semester 1, 2023, week 1a

Irena Koprinska

Reference: Witten ch.1, Tan ch. 1

Outline

- Administrative matters
- Introduction to machine learning and data mining

Administrative matters

Welcome to COMP5318/COMP4318!

- There are >400 students currently enrolled in this course – this is a big course!
- Local and international, from various degrees
- Welcome to everyone!

Teaching team

- Unit coordinator and lecturer weeks 1-6 and 13
  - Associate Professor Irena Koprinska
  - Computer Science Building, room 450, irena.koprinska@sydney.edu.au
- Lecturer weeks 7-12
  - Associate Professor Nguyen Tran
  - Computer Science Building, room 428, nguyen.tran@sydney.edu.au
- Teaching assistants
  - Nicholas Rhodes (head TA) and Stephen McCloskey
  - Tung Anh Nguyen, Long Tan Le and Jiayu He (Assignment 2)
- Tutors
  - Nicholas Rhodes, Stephen McCloskey, Mashud Rana, Henry Weld, Christie Zhu, James Collins and Jiayu He

Timetable

- Lectures
  - 2 hours weekly, 6-8pm on Monday, start in week 1
  - Online, over Zoom – we are using Zoom “webinars”
  - The lectures will be recorded and available on Canvas in “Recorded lectures” on Tuesday morning (it takes several hours to process the recording)
- Tutorials (also called pracs or labs)
  - 1 hour weekly on Monday, Tuesday, Wednesday and Thursday; start in week 2
  - You need to attend 1 tutorial only as per your timetable
  - There are 2 types: online (RE) and face-to-face (CC) as per your enrolment
    - Online: via Zoom, live-streamed (Zoom “meetings”) on Monday 8-9
    - Face-to-face: in the School of Computer Science labs, level 1, or Madsen Building (Tuesday, Wednesday and Thursday)
    - One online tutorial will be recorded every week and made available on Canvas in “Recorded lectures” on Thursday after the last tutorial
  - Please attend your allocated tutorial – see the tutorial number in your timetable

Zoom webinars

- Zoom webinars are different from Zoom meetings

- You can ask questions using the “Q&A” tool, there is no “Chat”
- Tips:
  - Wait before you ask your question, don’t do it immediately
  - As we cover the material, many of the questions will be answered, so they become redundant
  - If you post questions all the time, this is distractive for the other students, as they can see activity on “Q&A”
  - Concentrate on one thing - the lecture, do not constantly switch between the slides and “Q&A”
  - We will stop for “Question time”, and you will be invited to ask questions
  - This will be a more efficient and effective way to learn

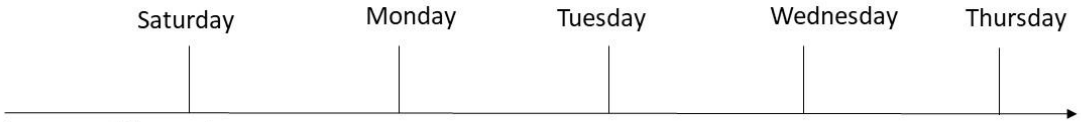
Course website

- The main place for this course is the Canvas website; we will use it for:
  - all teaching materials (unit outline, lecture slides, tutorial notes, tutorial solutions, assignments)
  - posting marks
- We will also use the discussion board Ed Discussion which is linked to Canvas

- Important document on Canvas: unit-outline-detailed.pdf – contains the most important information about this course

Course materials

- Lecture slides and tutorial notes will be available in advance on Saturday morning at 9am
- The lecture slides initially may not include the answers to questions and exercises that we will do at the lectures; the complete version with the answers will be uploaded after the lecture
- Tutorial solutions will be available on Thursday evening after the last tutorial, which finishes at 8pm



Lectures slides and tutorial notes available on Canvas (9am)

1. Attend lecture 6-8pm online, live-streamed.
2. Updated lecture slides (with answers) uploaded on Canvas at 8pm (after the lecture)
3. Lecture recording available the next day - Tuesday morning

1. Tutorial solutions available on Canvas on Thursday 8pm (after the last tutorial)
2. Tutorial recording available on Canvas on Thursday 8pm.

4. Attend tutorial (Monday, Tuesday, Wednesday or Thursday) as per your timetable. Online tutorials are on Monday and will be on Zoom live-streamed; face-to-face tutorials are on Tuesday, Wednesday and Thursday in the School of Computer Science labs or Madsen labs.

Discussion board

- We will use Ed Discussion, it is linked to Canvas

- Posting questions
  - Post your question on Ed instead of emailing them to us – this is beneficial for everyone
  - The question will be answered quicker
  - When it is answered, it is answered for everybody (and often many students have the same question)
  - If you are shy, you can ask your question anonymously!

Lecture slides and recordings

- The lectures will be recorded and available on Canvas
  - It takes several hours to process the Zoom recording before it becomes available on Canvas -> lecture Monday evening, recording on Canvas -> Tuesday morning
- However, you may not need the recordings
  - My lecture slides are very detailed, with many examples
  - I put everything important on the slides, including updating the slides after the lecture to add the solutions/answers whenever applicable
  - My slides are self-content and intended to help you revise and catch-up quickly

Assessment overview

- Three components:
  1. Assignment 1 – 15% (week 7)
  2. Assignment 2 – 25% (week 11)
  3. Exam – 60%

Assessment overview - assignments

- Two assignments
  - Programming assignments using Python and its machine learning libraries
  - Given a problem, you need to apply machine learning algorithms to solve it

  - Assignment 1 (15%) - due Thursday week 7 (6 April); in groups of 2 students (no more than 2 students are allowed)
    - The due date is just before the public holiday (Good Friday, 7 April) and the mid-semester break
    - Computer program only
    - Submitted via Canvas

  - Assignment 2 (25%) - due Friday week 11 (12 May); in a group – the group size will be confirmed later
    - Computer program and report
    - Submitted via Canvas (code and report)

Assessment overview – assignments (2)

- Assignments are due at 11.59pm
- Late submissions – allowed up to 3 days late
  - Late penalty of 5% per day will apply
  - Assignments submitted more than 3 days late will not be accepted

- Important: Start working on the assignments as soon as possible, do not delay them until a few days before the deadline!
  - Programming assignments require time; even the ones that look simple, almost always require much more time than expected!
  - Submit early to avoid last minute problems

Assessment overview - exam

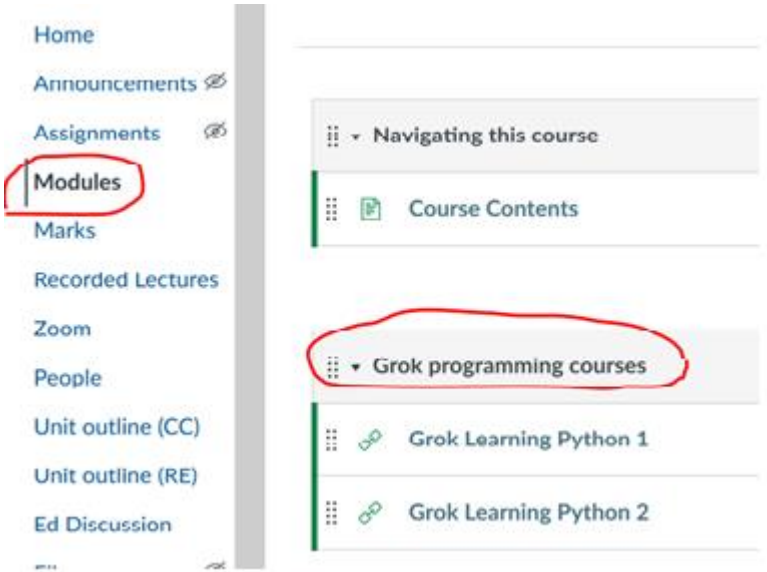
- Exam: 60% (individual), during the examination period
  - Supervised exam
  - The current advice from the University is: paper-based exam for CC students and online ProcrorU Live+ for RE students
  - Duration: 2 hours

- A minimum of 40% on the exam is required to pass the course – School of Computer Science policy. This means a minimum of 24 marks.
- More information about the exam will be provided in week 13, including sample exam questions

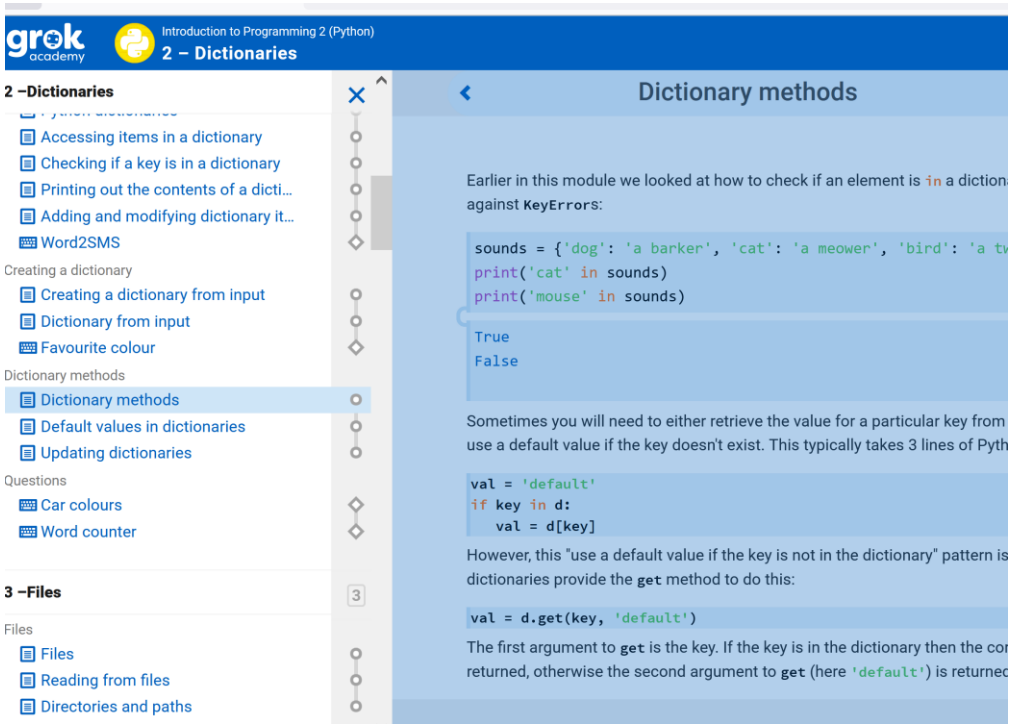
Assumed knowledge

- You need good programming skills for this course
- We expect that all students have a background in at least one programming language, preferably Python
- We will use Python and its machine learning libraries, e.g. sklearn
- If you don't know Python or haven't used it recently, you need to catch up
  - 1) In Canvas -> Modules, we have made available 2 Python programming courses offered by Grok Learning
  - 2) We have prepared a short Python refresher document – see Canvas
  - 3) There are many Python books and resources available on the web that you can use

The Grok Learning programming courses



- To help you learn Python or brush-up your skills
- Not compulsory, not assessable



Tutorials – Jupyter

- We will use Jupyter Notebook
  - A web-based interactive programming environment
  - See the document on Canvas on how to install it on your computer

- In summary - 3 important resources on Canvas related to the practical part of this course:
  - How to install Jupyter Notebook
  - Python refresher (short document)
  - Grok's Python courses (2 courses)

Tutorials

- During the tutorials we will do **practical exercises** using Python and its machine learning and neural network libraries
  - The exercises will be provided in a Jupyter Notebook format (.ipynb)
  - They are very detailed, we hope you will find them useful
  - Sometimes it may not be possible to finish all exercises during the 1-hour tutorial - you should do this at your own time
- For some of the weeks, we will also have **theoretical exercises -** paper-based exercises and calculations, testing your understanding of the algorithms
  - We will do some of them at the lectures, the rest should be done at your own time
  - Make sure that you do all theoretical exercises as they are similar in style to the exam questions
- The solutions for both types of exercises will be provided after the last tutorial (Thursday evening)

Textbooks

- **Textbooks:**
  - Pang-Ning Tan, Michael Steinbach, Anuj Karpathe and Vipin Kumar (2019). *Introduction to Data Mining*, 2nd edition. Pearson. (You can also use the previous edition)
  - Ian H. Witten, Eibe Frank, Mark Hall and Christopher J. Pal
  - *Data Mining - Practical Machine Learning Tools and Techniques*, 4th edition, Morgan Kaufmann, 2017 (You can also use the 3nd edition)
- **Books for the practical part using Python:**
  - Andreas C. Mueller and Sarah Guido (2016). *Introduction to Machine Learning with Python: a Guide for Data Scientists*, O'Reilly.
  - Aurelien Geron (2022). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow,* 2nd edition, O'Reilly. (You can also use the previous edition)
- All books are available from the library, both as hard copies and online versions, except Tan which is available as a hard copy only

Special Considerations (SC) due to Illness or Misadventure

- There is a centralized University system:
  http://sydney.edu.au/special-consideration
- Applications are submitted online, after login to “myUni”
- You are required to submit the SC form within 3 working days from the date when the assessment was due
- Applications are assessed by the University Student Administration Services (SAS) unit

University services, support and resources

Do you have a disability that impacts on your studies?

- You may not think of yourself as having a ‘disability’ but the definition under the Disability Discrimination Act (1992) 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, Autism, 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
- In order to get assistance, students need to register with Inclusion and Disability Services. It is advisable to do this as early as possible. Please contact us or review our website to find out more.
- Inclusion and Disability Services Office, sydney.edu.au/disability, phone: 02-8627-8422

COVID-19

- The University is following NSW Government and NSW Health guidance to prevent the spread of COVID-19, respiratory and other illnesses
- All staff, students and visitors are required to follow the University advice:
  - https://www.sydney.edu.au/covid-19/health-safety/keeping-our-campus-covid-safe.html
- This includes staying at home if unwell, isolating and testing, and not returning unless recovered or as advised by your health professional
- For more information about COVID-19, study arrangements and who to contact for help:
  - https://www.sydney.edu.au/covid-19/

COVID-19 support and care

- Stay at home if you are feeling unwell with any COVID-19 symptoms
- We don’t take attendance in this course, so there is no need to notify us and explain absences from tutorials
- If COVID-19 illness impacts assessment, use the usual mechanisms including simple extensions and special consideration to arrange reasonable adjustments
- Further information – Student wellbeing and support
  - https://www.sydney.edu.au/students/support.html

Health and safety advice

Student support

- Visit the Student life, wellbeing and support webpage to find out about the student services, resources and events available to support you while you study:
  - Health and wellbeing
  - Academic support
  - Personal support
  - Getting connected
  - https://www.sydney.edu.au/students/support.html

- Questions about getting started this semester? Come and visit us at a Welcome Hub!

Anderson Stuart
Welcome Hub

Carslaw West
Welcome Hub

Safer Communities Office

- Support and case management for people who have experienced sexual misconduct, domestic/family violence, bullying/harassment or issues relating to modern slavery
- Contact the Safer Communities office:
  - 8:30--17:30, Monday to Friday
  - phone: +61 2 8627 6808
  - email: safer-communities.officer@sydney.edu.au
  - campus: Level 5, Jane Foss Russell building

- Visit the website to make a complaint or disclosure of sexual misconduct to the University:
- https://www.sydney.edu.au/about-us/vision-and-values/safer-communities/report-sexual-misconduct.html

Tips for learning online

- Tips and guides on learning online and the tools we will use, refer to "Learning while off campus resources" in Canvas:
  - https://canvas.sydney.edu.au/courses/4901/pages/learning-while-off-campus

Emergency procedures on campus

- In the unlikely event of an emergency, we may need to evacuate the building
- If we need to evacuate, we will ask you to take your belongings and follow the green exit signs
- We will move a safe distance from the building and maintain physical distancing whilst waiting until the emergency is over
- In some circumstances, we might be asked to remain inside the building for our own safety
- More information is available at www.sydney.edu.au/emergency

Academic Honesty

Cheating and plagiarism

- Please read the University policy on Academic Honesty carefully:
  https://sydney.edu.au/students/academic-integrity.html
- All new students are required to complete the Academic Honesty Education Module (AHEM)
- All cases of academic dishonesty and plagiarism will be investigated
- There is 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/exam answer or only a part of it.
  - Academic dishonesty – when you make your work available to another student to copy (for assignments or exams). There are other examples of academic dishonesty.
  - Misconduct - when you engage another person to complete your assignment/exam (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.

Plagiarism detection

- For the assignments, we will use similarity detection software such as TurnItIn to compare your assignments with these of other students (current and previous) and the Internet
- These similarity detection tools are extremely good!

- Note: We always check all flagged cases manually, the decision is not taken by the similarity detection software – it is an academic decision
- There is always some baseline similarity depending on the assignment and other factors

Student excuses

- These are cases of plagiarism and academic dishonesty from our school
- The student excuses are not acceptable and both parties were penalized
- *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.*
- *He is my best friend. I had no choice but to let him copy my assignment.*
- *I posted parts of my code on my web page (group discussion) because my solution was cool (or I wanted to help them). I didn't expect them to copy.*

More about cheating

- Cheating at the exam? Don't even think about this!
- We had plagiarism cases in previous years (unsupervised exam)
- This year the exam is supervised – paper-based or ProctorU Live+

- Cheating at the assignments or exam - there are huge risks and penalties – it is not worth it!
- The stress of going through the investigation is immense. The investigation may take several months. Until the investigation is completed, your mark will not be finalised, you will not be able to enroll in other courses that have COMP5318 as a pre-requisite and your graduation may be delayed.
- It is not worth it. You will regret it all your life.
- You don't need to cheat - you can do well without cheating!

Cheating and plagiarism – key message

- Please do not confuse legitimate cooperation with cheating. You can discuss the assignment with other students, this is a legitimate collaboration, but you cannot complete the assignment together unless you are in the same group – every group must write their own code and report.
- The exam is individual – no collaboration is allowed
- Plagiarism and any form of academic dishonesty will be dealt with, and the penalties are severe
- We use plagiarism detection systems which 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 exam answers, or to complete the assignment or exam instead of them, just say: *I can't do this. This is against the University policy. I will not risk my reputation and future by doing this.*

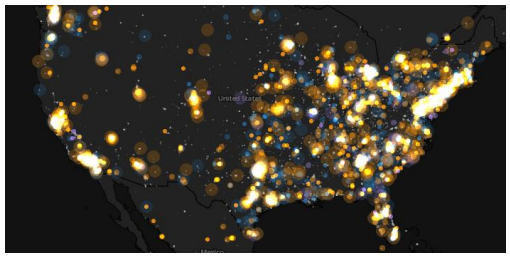
- **Be smart and don't risk your future by engaging in plagiarism and academic dishonesty!**

Introduction to Machine Learning and Data Mining

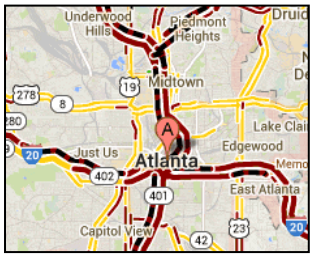
Data is everywhere

- 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
  - telephone call details, government statistics, traffic data

E-Commerce



Social Networking: Twitter



Traffic

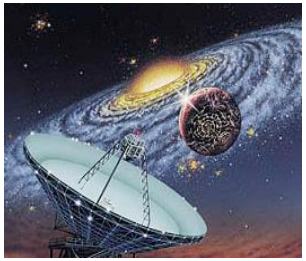
amazon.com

Google

facebook

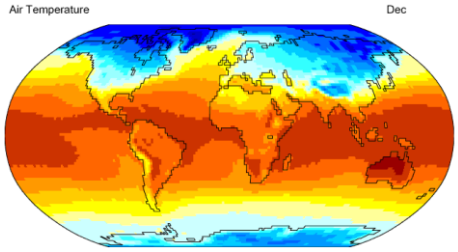
More examples

- Scientific data
  - telescopes scanning the skies
  - remote sensors on satellites
  - weather data

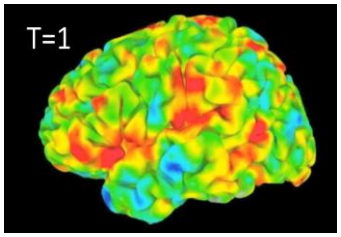


Sky survey data

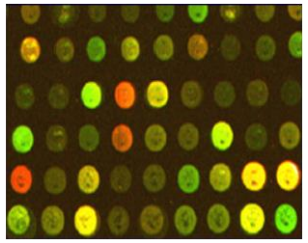
Sensor networks



Temperature



fMRI brain data



Gene expression data

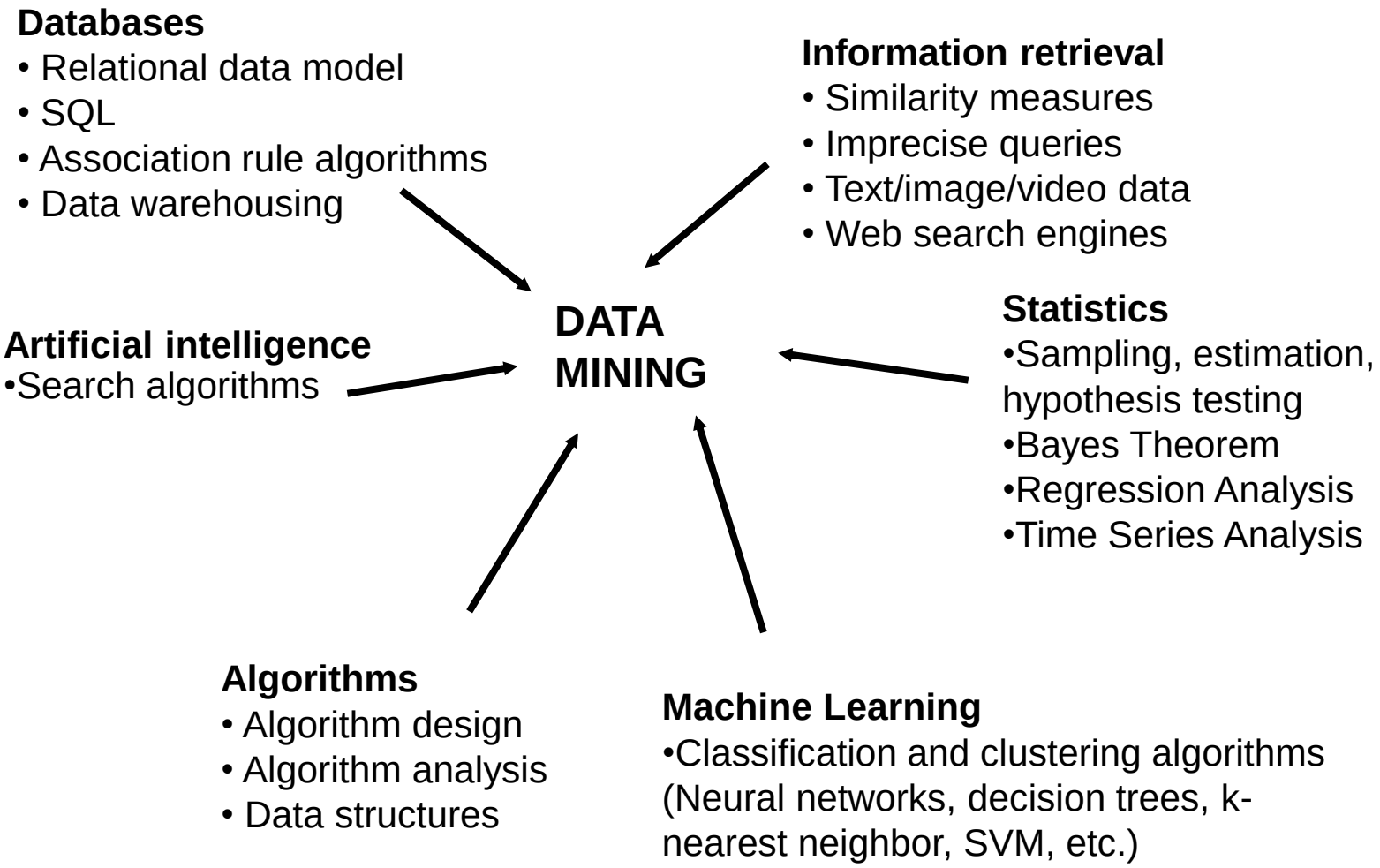
  - medical records and scans
  - biological data (high-throughput) – cytometry, gene expression

From data to knowledge – ML and DM

- Current trend: Gather whatever data you can, whenever and wherever possible!
- Expectation: it will be useful either for the purpose being collected or another purpose, not yet envisioned
- However, raw data is useless – need for methods to automatically extract knowledge (useful patterns) from it
- Machine Learning (ML) and Data Mining (DM) are concerned with finding patterns in data
  - These patterns should be meaningful, useful and actionable
  - The process is automatic or semi-automatic

- ML vs DM
  - ML is a core part of Artificial Intelligence
  - Most of the algorithms used for DM have been developed in ML
  - DM deals with large and multidimensional data, ML not necessarily
  - DM can be seen as applied ML – we use ML algorithms to do DM

DM origins


Databases
• Relational data model
• SQL
• Association rule algorithms
• Data warehousing
Information retrieval
• Similarity measures
• Imprecise queries
• Text/image/video data
• Web search engines
Artificial intelligence
•Search algorithms
DATA MINING
Statistics
•Sampling, estimation, hypothesis testing
•Bayes Theorem
•Regression Analysis
•Time Series Analysis
Algorithms
• Algorithm design
• Algorithm analysis
• Data structures
Machine Learning
•Classification and clustering algorithms (Neural networks, decision trees, k-nearest neighbor, SVM, etc.)

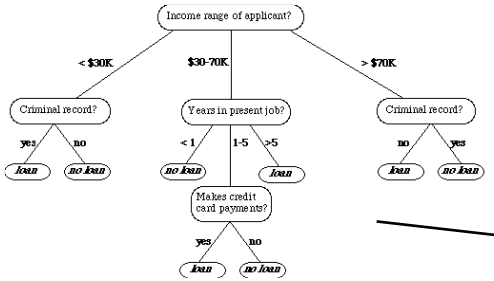
ML and DM tasks

Main tasks

- 2 main types of tasks:
  - Supervised learning - classification and regression
  - Unsupervised learning – clustering
- Other:
  - Association rule mining
  - Reinforcement learning
  - Outlier detection

- We will cover algorithms for supervised and unsupervised learning, and also for reinforcement learning (briefly in week 12)

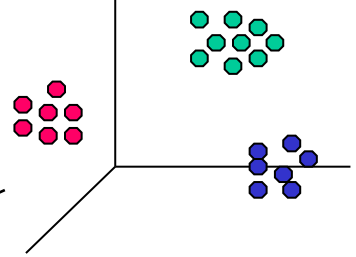
Main tasks - diagram



Supervised learning
(classification and regression)

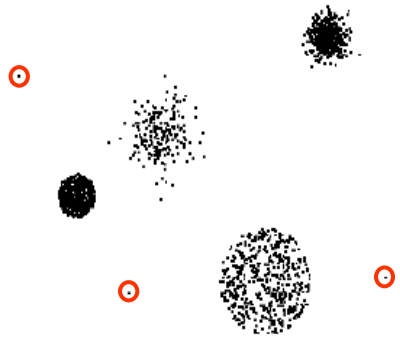
Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



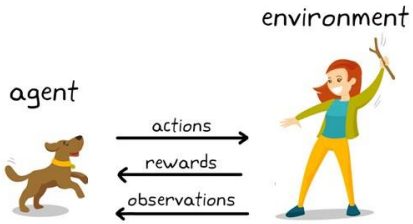
Unsupervised learning
(clustering)

Association rule mining



Anomaly
detection

Reinforcement
learning



Topics covered

Week	Date	Topic	Lecturer
1	20 February	Administrative matters and course overview. Introduction to machine learning and data mining. Data: cleaning, pre-processing and similarity measures.	Irena
2	27 February	Nearest neighbour. Rule-based algorithms.	Irena
3	6 March	Linear regression. Logistic regression. Overfitting and regularization.	Irena
4	13 March	Naïve Bayes. Evaluating machine learning methods. Assignment 1 out (Monday – 13 March)	Irena
5	20 March	Decision trees. Ensembles.	Irena
6	27 March	Support vector machines. Kernels. Dimensionality reduction.	Irena
7	3 April	Neural networks - perceptrons and multilayer perceptrons. Assignment 1 due (Thursday – 6 April) Assignment 2 out (Thursday – 6 April)	Nguyen
		Mid-semester break	
8	10 April	Deep neural networks: convolutional and recurrent.	Nguyen
9	24 April	Clustering I: Partitional, model-based and hierarchical.	Nguyen
10	1 May	Clustering II: Density-based and grid-based. Evaluating clustering results.	Nguyen
11	8 May	Markov models. Assignment 2 due (Friday – 12 May)	Nguyen
12	15 May	Reinforcement learning.	Nguyen
13	22 May	Guest lecture. Revision.	Irena

Supervised learning

- Given: a set of pre-classified (labelled) examples {x,y}
  - x – input vector, y - target output
- Task: learn a function (classifier, model) that maps x->y and can be used predictively
  - i.e. to predict the value of y given the values of x for new, unseen examples
- Why is it called supervised?

- Two types of supervised learning
  - Classification: the variable to be predicted is categorical (i.e. its values belong to a pre-specified, finite set of possibilities)
  - Regression: the variable to be predicted is numeric

Classification - example

input vector, with 3 features

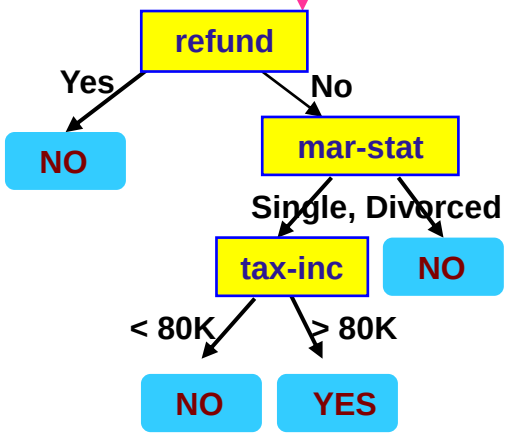
target class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

training data

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

new data



predict the class

Classifier

Step 1: Create the classifier
Step 2: Use it predictively on new data

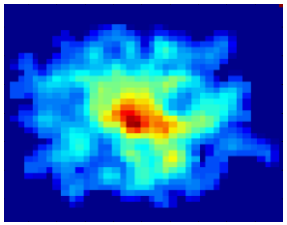
More classification examples (1)

- Ex. 2: Fraud detection in credit card transactions
  - Data about customers and their transactions
    - previous credit card transactions
    - what they typically buy and when
    - demographic and socio-economic information - age, education, income, etc.
  - Label previous transactions as fraud or fair
  - Build a classifier to detect fraud transactions on new data (for new transactions of the same customer or for new customers)

More classification examples (2)

- Ex. 3: Direct marketing – find a set of customers likely to buy a product
  - Given a user, is she/he likely to buy a product, e.g. a new mobile phone?
  - Data about
    - the user – phone usage, demographic and lifestyle
    - previous similar products – what are the characteristics of the customers who decided to buy and who didn't – extract features
  - 2 classes {buy, don't buy} – build a classifier

- Ex. 4: Sky survey cataloging
  - Task: Predict the class (star or galaxy) of sky objects
  - Data: images from an observatory
  - Dataset: 72 million stars, 20 million galaxies
  - From Fayyad et. al. *Advances in Knowledge Discovery and Data Mining*

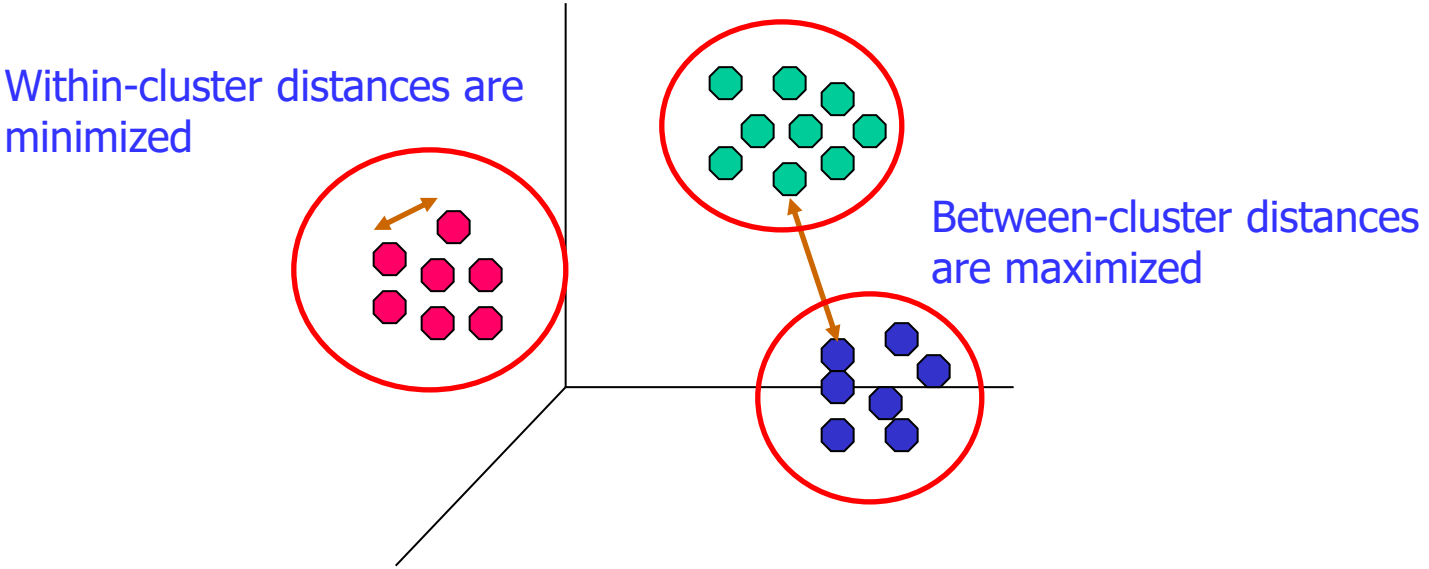


Regression - examples

- Predict the electricity demand
  - Data: previous electricity demand, weather data and weather forecast data for the future days
  - Important to prevent blackouts and ensure reliable supply of electricity; also important for the economical and efficient operation of the electricity grid and for supporting the electricity market participants
  - Short and long term predictions - for the next few hours, next day, next week etc.; every 5 min, 30 min, 60 min, etc.
- Predict the exchange rate of AUD
  - Data from previous days, economical indicators, political events
- Predict retirement savings
  - Data: current savings and market indicators
- Predict the house prices in Sydney in 2030
- Predict the stock market index
- Predict wind velocity based on temperature, humidity, pressure

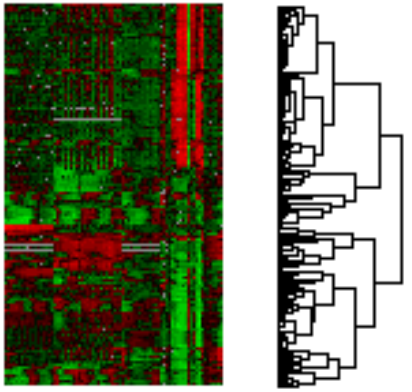
Unsupervised learning

- Given: a set of examples containing only input vectors x (no target outputs y)
- Task: group (cluster) the examples into a finite number of clusters, so that the examples
  - from each cluster are similar to each other
  - from different clusters are dissimilar to each other



Clustering - examples

- Ex.1: Targeted marketing
  - Segment customers into groups with distinct characteristics and use this knowledge to develop targeted marketing campaigns
  - (targeted campaigns are cheaper than mass-campaigns)
- Ex. 2: Customer loyalty
  - Analyse customer behavior and find groups of customer who are likely to defect, e.g. to another medical insurance, electricity or phone company
- Ex. 3: Gene clustering
  - Find genes with similar structure and functionality – important for understanding diseases and finding effective treatments
  - Data: microarray – from thousands of genes, analysed simultaneously



Clustering – examples (2)

- Ex. 3: Document clustering
  - Find groups of documents that are similar to each other based on their content
  - Applications:
    - Patent documents assessment: group similar patent documents to make the evaluation of a new patent document easier
    - Personalized news recommendations

- Ex. 4: Clustering for understanding eating habits and dietary patterns of a particular cohorts (e.g. of young Australians)
  - Group 1: People who skip breakfast, care about weight, do not exercise regularly; eat high protein, low fat and high sugar diet; eat out because they enjoy the social aspect; snack after dinner
  - Group 2: …
  - Use this knowledge to promote good eating habits and changes in government policies

Association rule mining

- Find combinations of items that occur together
- Also called *market-basket analysis*
- Assumes transaction data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Potato Chips, Milk
4	Beer, Bread, Potato Chips, Milk
5	Coke, Potato Chips, Milk

Rules discovered:
{Milk} --> {Coke}
{Potato Chips, Milk} --> {Beer}

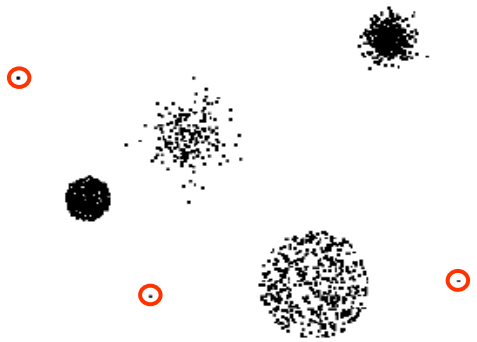
- Sequential version of association rule mining: find *frequent sequences* in data

Association rule mining - examples

- Business and marketing
  - Rules are used for sales promotion, shelf management and inventory management
  - E.g. sales promotion: services purchased together by telecommunication customers (e.g. broad band Internet, call forwarding, etc.) help determine how to bundle these services together to maximize revenue
- Telecommunication alarm diagnosis
  - Find combination of alarms that occur together frequently in the same time period
- Insurance
  - Unusual combinations of insurance claims can be a sign of a fraud
- Medical informatics
  - Find combination of patient symptoms and test results associated with certain diseases
  - Medical histories can give indications of complications based on combinations of treatments

Outlier detection

- Outliers are examples that are significantly different than the others (i.e. are far away from the others)
- In statistics an outlier is typically defined as an example that differs more than 3 standard deviations from the mean of all examples
- Detecting outliers is important for 2 reasons:
  - Outliers are noise and should be removed before data analysis
  - Outliers are the goal of our DM analysis – to detect unusual behavior, e.g. credit card fraud detection or intrusion detection



Outliers should be considered carefully

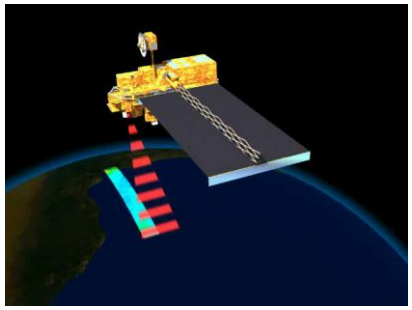
- An example where outliers were mistakenly removed

- Detecting the ozone hole in 1985:
  - Data collected by the British Antarctic Survey showed 10% drop in the ozone concentration for Antarctica
  - However, data collected by a satellite (Nimbus 7) did not show this drop

  - Why?
  - The satellite correctly recorded the ozone concentration but the values were so low that they were treated as outliers by the computer program and discarded!

Outlier detection - examples

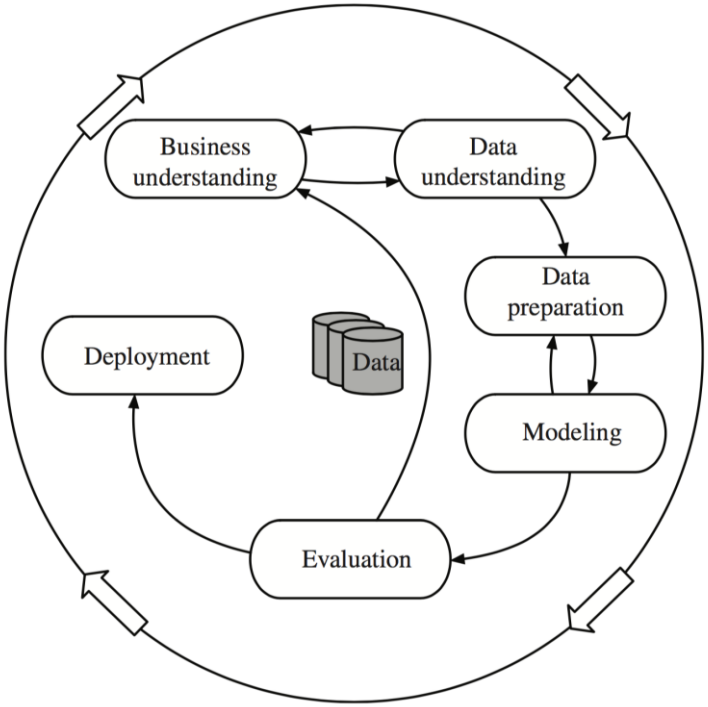
- The goal is to detect significant deviations from normal behavior
  - Fraud detection – e.g. deviation from typical behavior in credit card usage
  - Intrusion detection – monitoring computers and networks for unusual behavior
  - Hurricanes, floods, heat waves and fires prediction – atypical events with significant effect on humans
  - Health care – unusual symptoms or test results may indicate potential health problems and should be investigated
  - Detecting changes in the global forest cover



The DM process

1) Business understanding

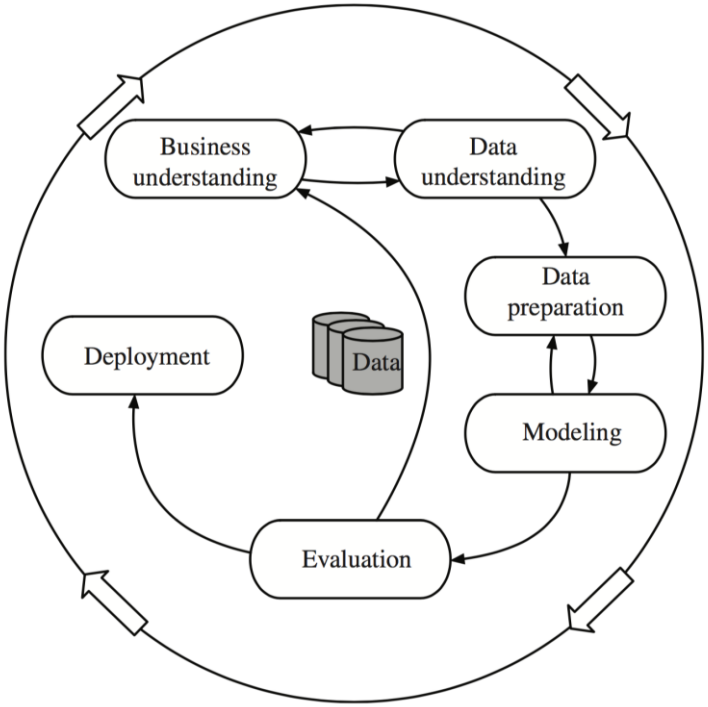
- Investigating the business objectives and requirements
- Deciding whether DM can be applied to meet them
- Determining what kind of data can be collected to build a deployable model



The DM process (2)

2) Data understanding

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



The DM process (3)

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
  - Noisy, e.g. containing errors or outliers
  - Inconsistent, e.g. in codes, names

  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

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 (4)

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