University of Southampton

COMP6237 Data Mining Introduction Lecture

Zhiwu Huang

Zhiwu.Huang@soton.ac.uk

Lecturer (Assistant Professor) @ VLC of ECS

University of Southampton

Lecture slides available here:

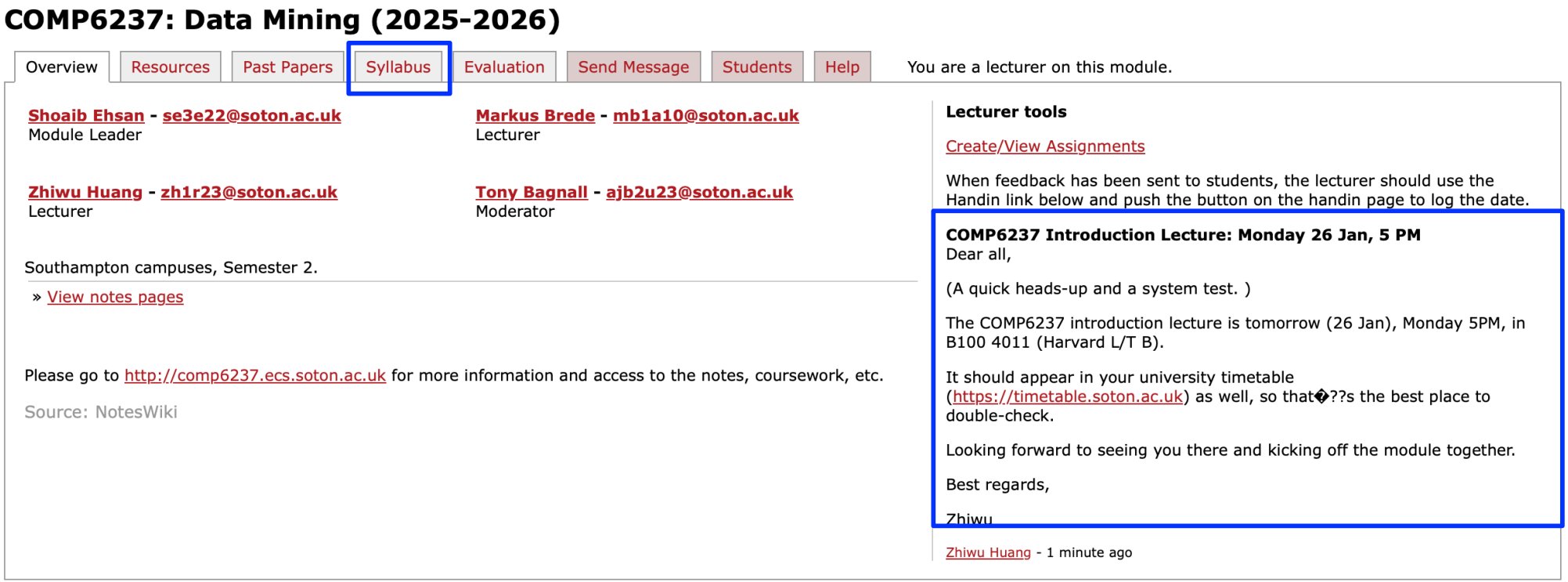
http://comp6237.ecs.soton.ac.uk/zh.html

Book time with Zhiwu Huang: Office Hour

(Thanks to Prof. Jonathon Hare and Dr. Jo Grundy for providing the lecture materials used to develop the slides.)

Module Overview

- **ECS module pages [syllabus, announcements]**
  - https://secure.ecs.soton.ac.uk:/module/comp6237



COMP6237: Data Mining (2025-2026)

Overview | Resources | Past Papers | Syllabus | Evaluation | Send Message | Students | Help

You are a lecturer on this module.

Shoaib Ehsan - **se3e22@soton.ac.uk**
Module Leader

Markus Brede - **mb1a10@soton.ac.uk**
Lecturer

Zhiwu Huang - **zh1r23@soton.ac.uk**
Lecturer

Tony Bagnall - **ajb2u23@soton.ac.uk**
Moderator

Southampton campuses, Semester 2.

» View notes pages

Please go to http://comp6237.ecs.soton.ac.uk for more information and access to the notes, coursework, etc.

Source: NotesWiki

Lecturer tools

Create/View Assignments

When feedback has been sent to students, the lecturer should use the Handin link below and push the button on the handin page to log the date.

COMP6237 Introduction Lecture: Monday 26 Jan, 5 PM
Dear all,

(A quick heads-up and a system test.)

The COMP6237 introduction lecture is tomorrow (26 Jan), Monday 5PM, in B100 4011 (Harvard L/T B).

It should appear in your university timetable (https://timetable.soton.ac.uk) as well, so that??s the best place to double-check.

Looking forward to seeing you there and kicking off the module together.

Best regards,

Zhiwu

Zhiwu Huang - 1 minute ago

Teaching Staff

COMP6237: Data Mining (2025-2026)

Overview | Resources | Past Papers | Syllabus | Evaluation | Send Message

Shoaib Ehsan - s.ehsan@soton.ac.uk
Module Leader B60/R2017

Markus Brede - Markus.Brede@soton.ac.uk
Lecturer B32/R4033

Zhiwu Huang - Zhiwu.Huang@soton.ac.uk
Lecturer B32/R3039

Tony Bagnall - **ajb2u23@soton.ac.uk**
Moderator

Southampton campuses, Semester 2.

https://secure.ecs.soton.ac.uk:/module/comp6237

Student Cohort

COMP6237: Data Mining (2025-2026)

Overview | Resources | Past Papers | Syllabus | Evaluation | Send Message | Students | Help | You are a lecturer on this module.

73 listed students.

DOWNLOAD AS CSV - please remember that this information is subject to data protection laws. If you are not sure what it can be used for, ask student services (SAA).

Cohorts

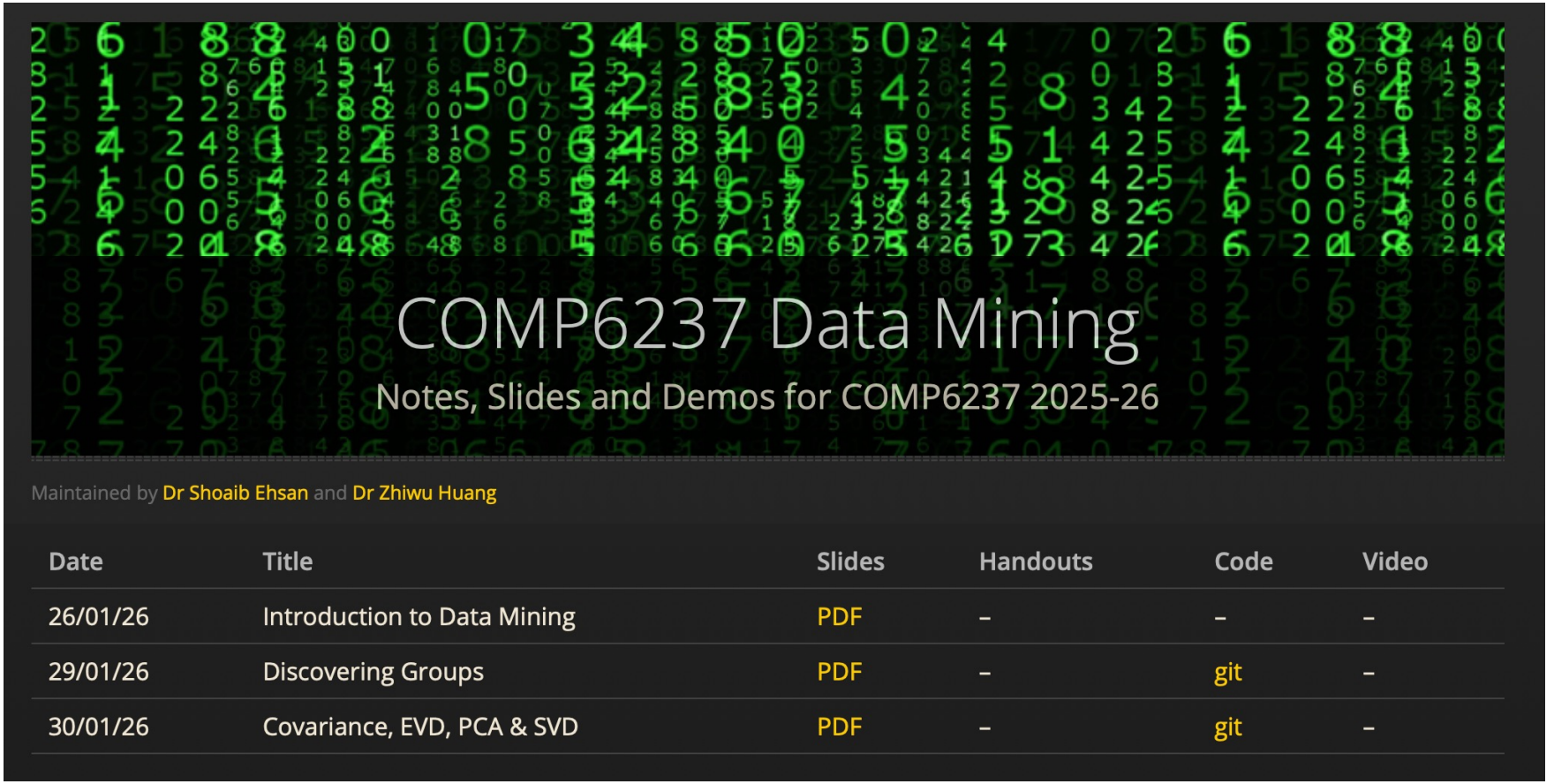
Show All entries Search:

Count	Year	Deg.	Degree	Prog.	Programme
1	05	6190	MEng Software Engineering with Industrial Studies		
1	02	4475	MSc Artificial Intelligence		
2	04	4439	Artifical Inteligence (MEng Electronic Engineering)		
3	04	4444	Artificial Intelligence (MEng Computer Science)		
4	01	4466	MSc Computer Science		
7	04	4443	Computer Science (MEng Computer Science)		
20	01	4475	MSc Artificial Intelligence		
35	01	6150	MSc Data Science		

https://secure.ecs.soton.ac.uk:/module/comp6237

Module Overview

- **Course website [notes, slides, recordings]**
  - https://comp6237.ecs.soton.ac.uk/zh.html

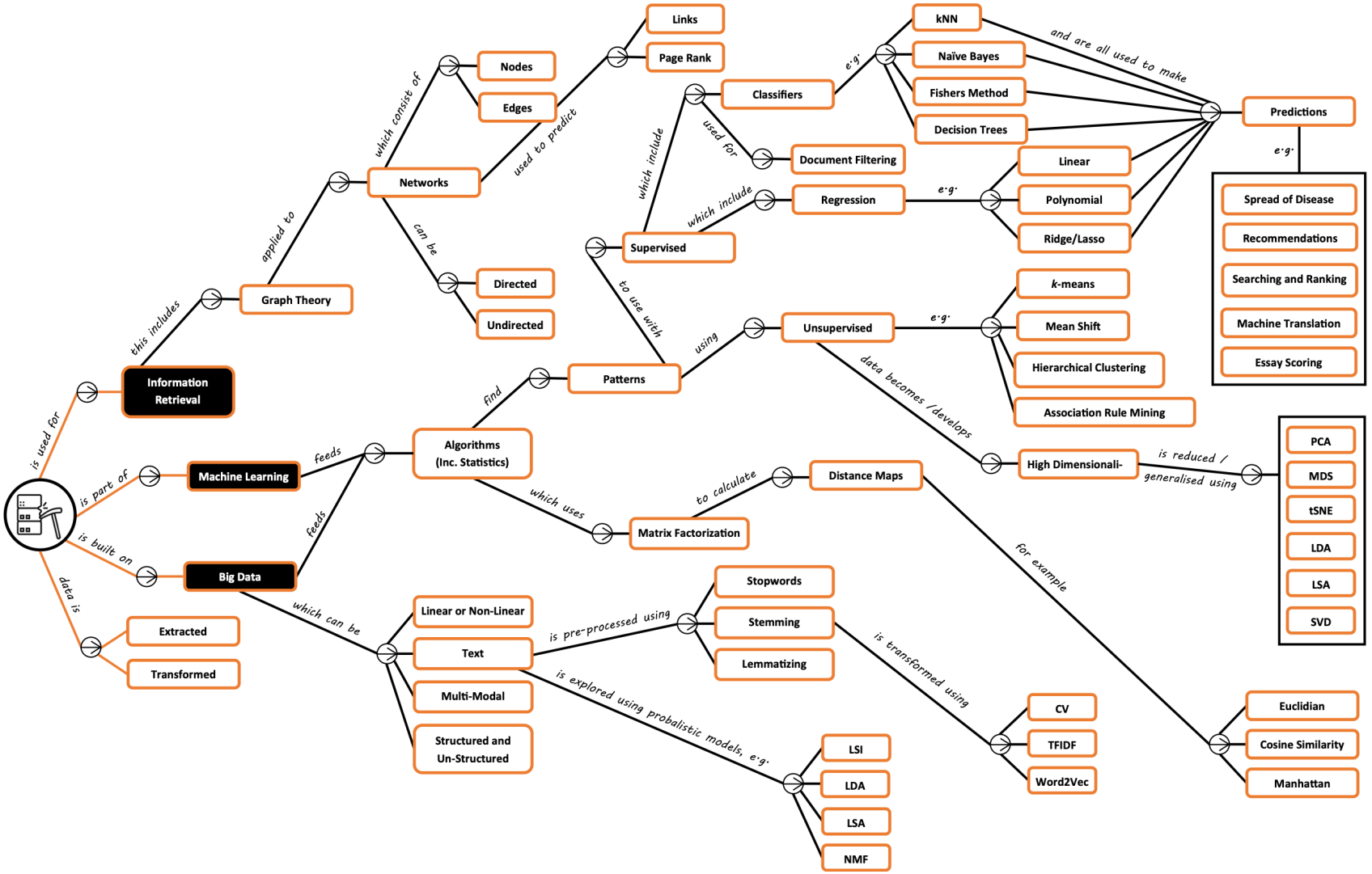


Module Overview

University of Southampton

- Developed by Prof Jon Hare & Dr Jo Grundy, run for the 9th time
- Created to fill a gap
  - **Data mining is almost synonymous with machine learning**
    - Inevitably have some overlap with machine learning modules e.g. COMP3222/COMP3223/COMP6245/COMP6208
      - Should be complementary and offer different views
    - **Slightly more applied pragmatic focus**
      - How do you work with real world data?
      - How do you solve real problems?

Module Overview



- Around 26 lectures + additional tutorials
- Wide range of data mining topics

Module Overview

- Reading material
  - Toby Segaran. Programming Collective Intelligence: Building Smart Web 2.0 Applications. O'Reilly, 2007
  - Aurélien Géron. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. March 2017
  - J. Leskovec et al. Mining of Massive Datasets. Third Edition. Cambridge University Press. 2020
  - M. J. Zaki and W. Meira, Data Mining and Machine Learning: Fundamenal Concepts and Algorithms. Cambridge University Press. 2020.

Module Timetable

University of Southampton

Day	Time	Room
Monday	5 PM	B100 4011 (Harvard L/T B)
Tuesday	5 PM	B46 2003 (L/T B)
Thursday	12 PM	B46 2003 (L/T B)
Friday	3 PM	B46 2003 (L/T B)

Date	Semester Week	Lecturer(s)	Topic/Title
26-Jan	1	Zhiwu	Intro to Data Mining
29-Jan		Zhiwu	Finding Groups
30-Jan		Zhiwu	Covariance
02-Feb	2	Zhiwu	Embedding Data
03-Feb		Zhiwu	Search
06-Feb		Shoaib	Linear Regression I; Group CW set
09-Feb	3	Shoaib	Linear Regression II
12-Feb		Shoaib	Linear Regression Problem Sets

http://comp6237.ecs.soton.ac.uk/

Note: This may sometimes also change –we'll update you by email (check ECS module page)

Module Assessment

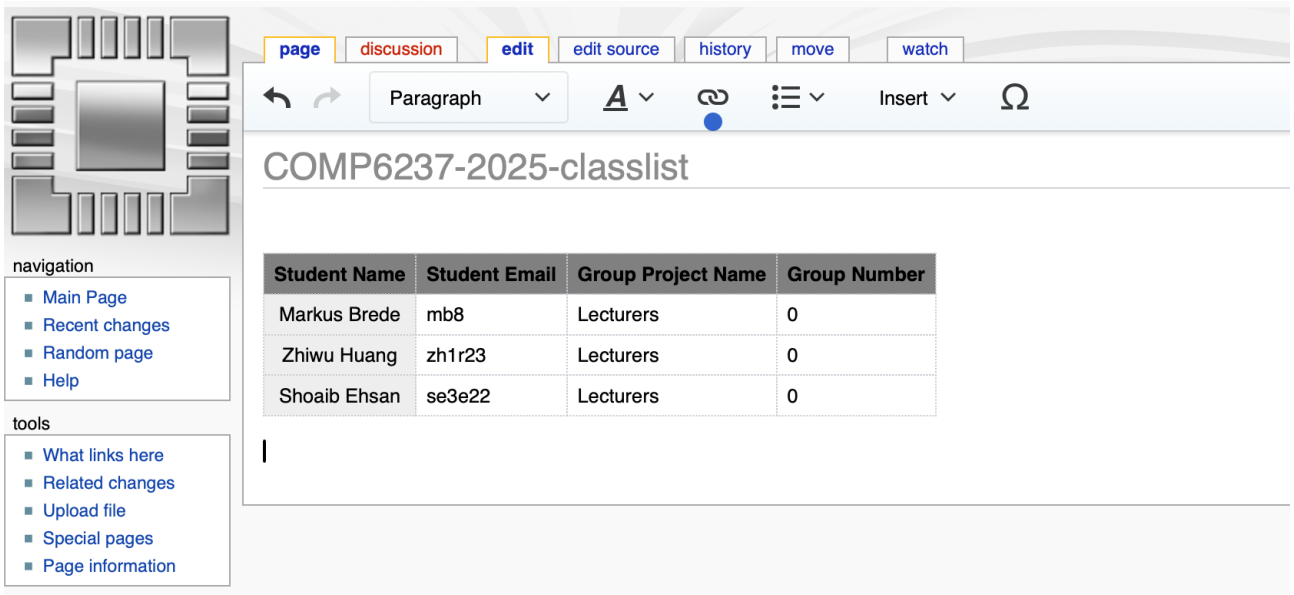
Assessment

Summative

Assessment method	Contribution to final mark
Group Coursework	30%
Final Exam	70%

Group Coursework

- Three Q & A sessions in **Week 3&4**; by that time we want you to have formed groups at https://secure.ecs.soton.ac.uk/student/wiki/w/COMP6237-2025-classlist



Student Name	Student Email	Group Project Name	Group Number
Markus Brede	mb8	Lecturers	0
Zhiwu Huang	zh1r23	Lecturers	0
Shoaib Ehsan	se3e22	Lecturers	0

- Four presentation sessions before Easter (**Week 8**)
- Report submission at the end of the term (**May 15**)

Final Exam

- **Computer-aided with only multiple-choice questions**
  - Shoaib (40 marks) + Zhiwu (40 Marks) + Markus (20 Marks)
  - 3 Lectures for revisions
  - Platform: https://moodle.ecs.soton.ac.uk

Please read the instructions below and wait on this page until an invigilator tells you to start. Press "Start Exam" when instructed to by an invigilator.

This is a Computer Aided Assessment. Follow all instructions in the exam software.

SEMESTER 2 EXAMINATIONS 2024-2025

Data Mining

Duration: 120 mins (2 hours)

This paper contains 33 questions.

Answer all questions.

Only University approved calculators may be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct 'Word to Word' translation dictionary AND it contains no notes, additions or annotations.

What is Data Mining?

University of Southampton

"Data mining is **the process of extracting and finding patterns in massive data sets** involving methods at the intersection of machine learning, statistics, and database systems.

Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of **extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use.**"

–Wikipedia

University of Southampton

Why Do we Learn Data Mining?

- “Modern AI systems like ChatGPT are trained on massive amounts of data and can understand text, images, and tables. So why do we still need to learn data mining?”

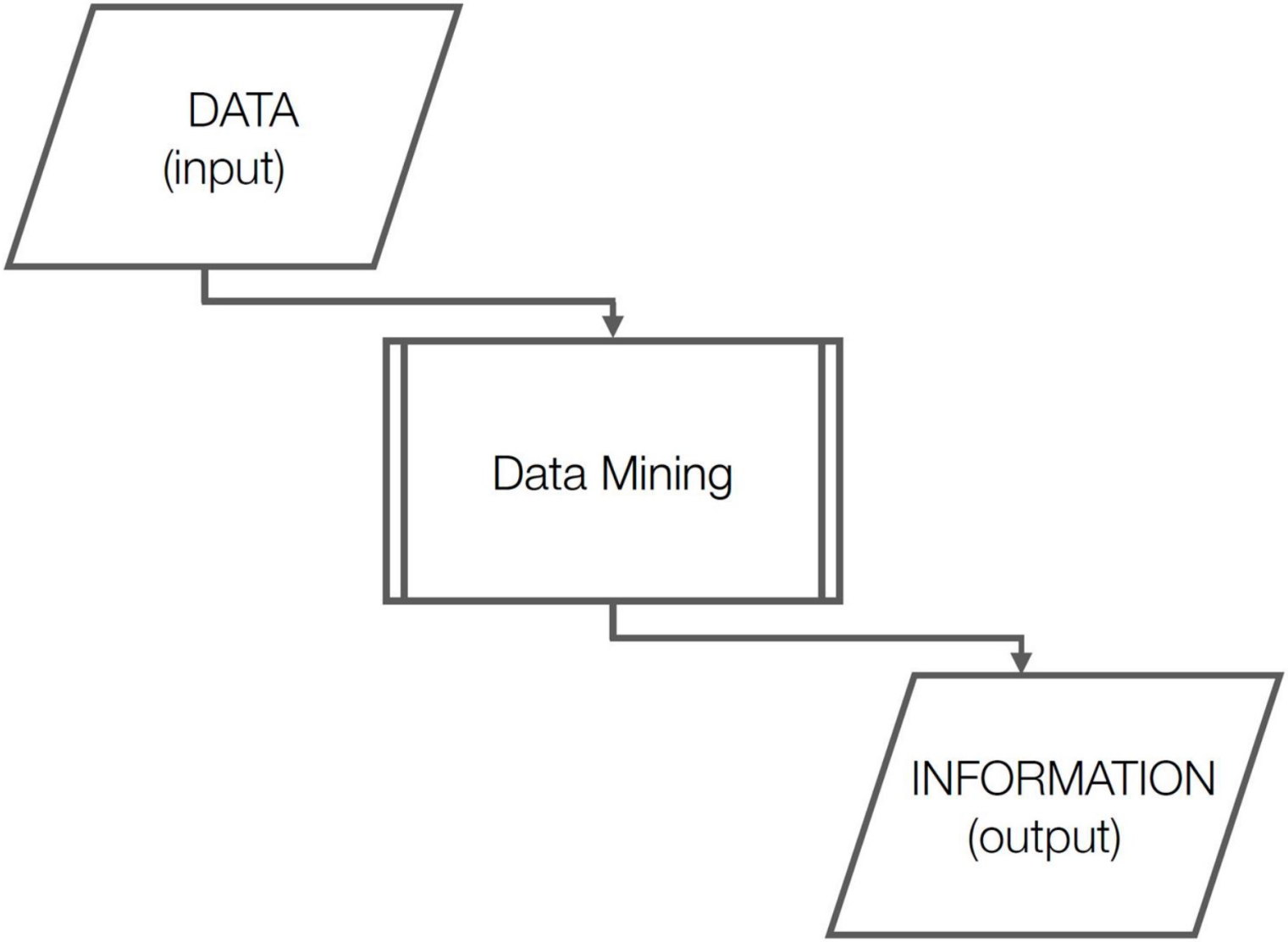
University of Southampton

Why Do we Learn Data Mining?

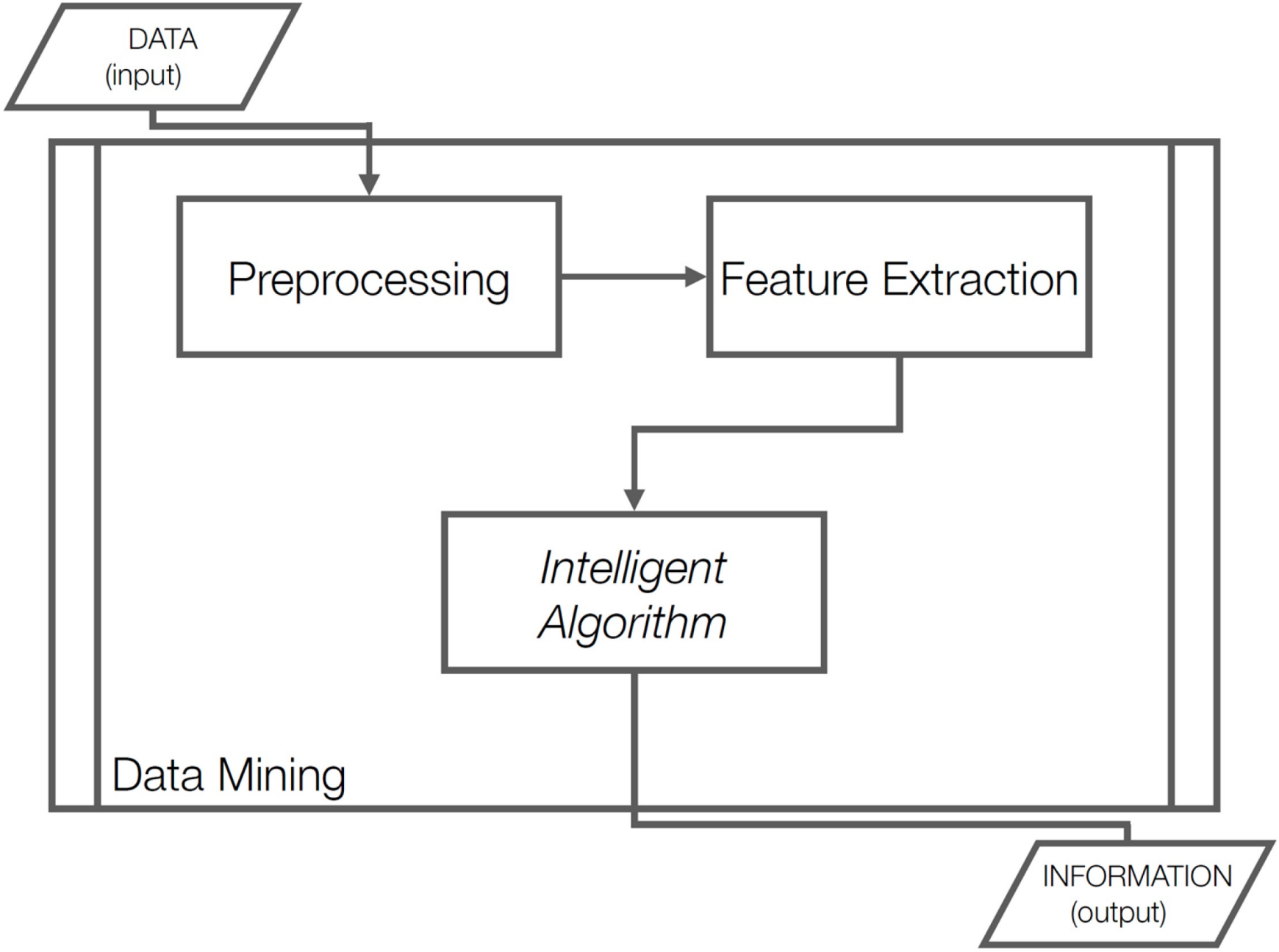
- You might ask ChatGPT: “Hi ChatGPT, this is my company’s customer data. Please use data-mining techniques to discover meaningful patterns, customer segments, and insights that are not immediately obvious from the raw data.”

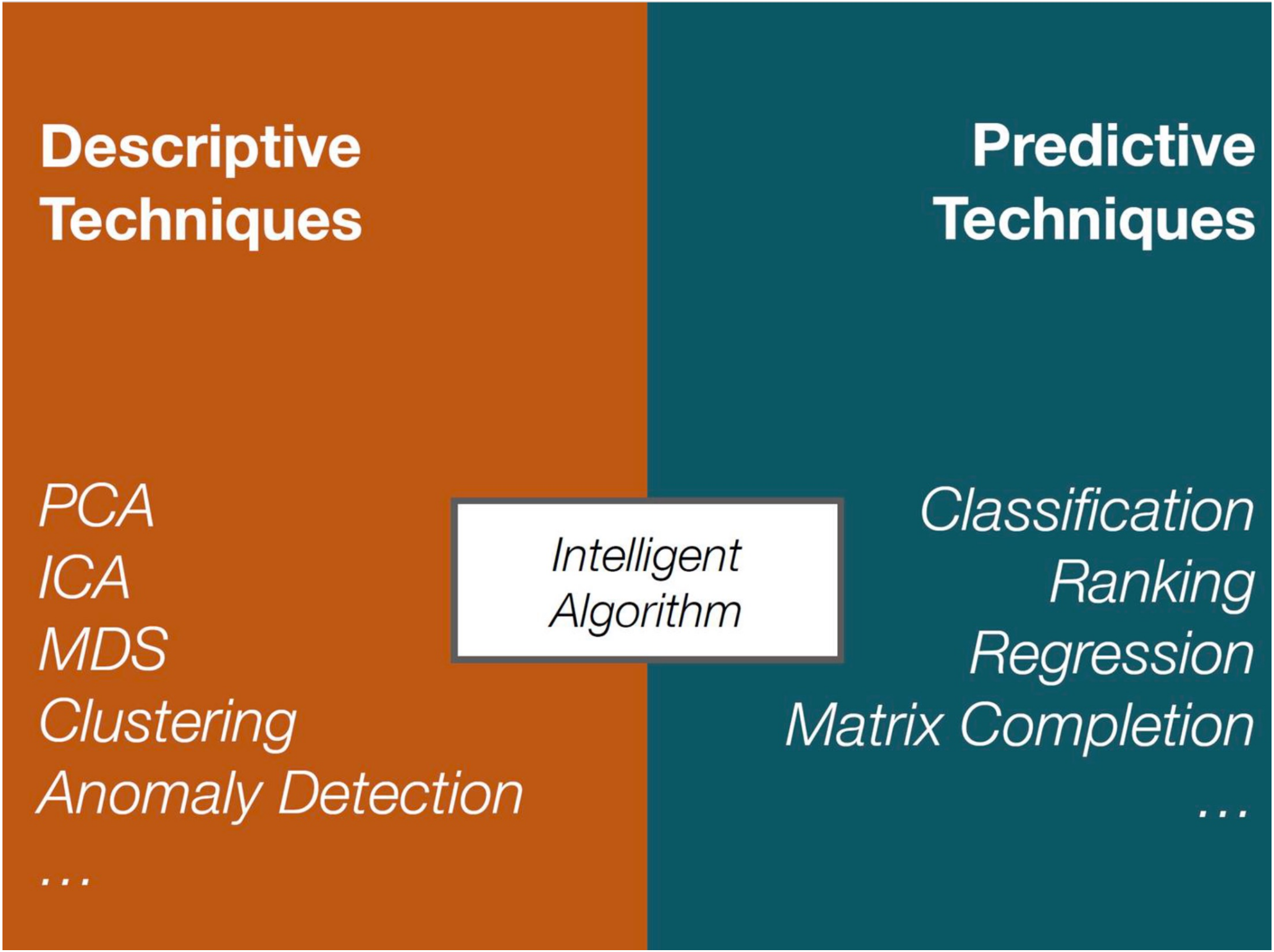
Task	Data Mining Skills (Student)	AI Capability	Why Students Still Need Data Mining
Problem definition	Turns real-world needs into data questions	Responds to given prompts	AI can’t choose the right problem
Data preparation	Cleans data, fixes errors, handles bias	Suggests generic steps	Bad data leads to wrong results
Pattern discovery	Evaluates significance and stability	Finds correlations	AI may find misleading patterns
Model evaluation	Selects proper metrics and validation	Computes metrics	Wrong metrics = wrong decisions
Interpretation	Links results to domain actions	Summarizes outputs	AI lacks real-world accountability

How Can We Do Data Mining?

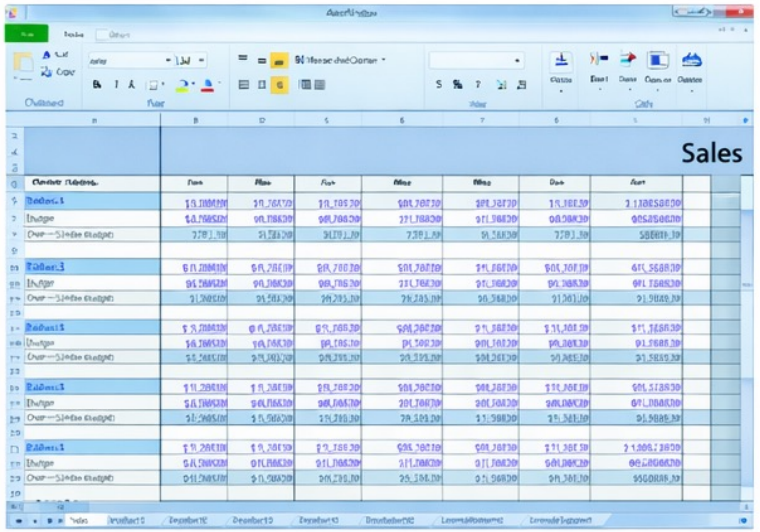


How Can We Do Data Mining?



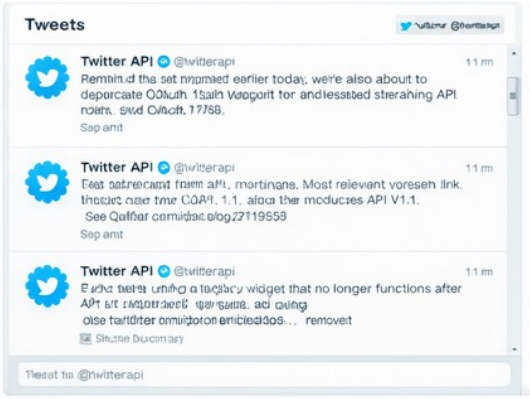

Descriptive Techniques
Predictive Techniques
PCA
ICA
MDS
Clustering
Anomaly Detection
...
Intelligent Algorithm
Classification
Ranking
Regression
Matrix Completion
...

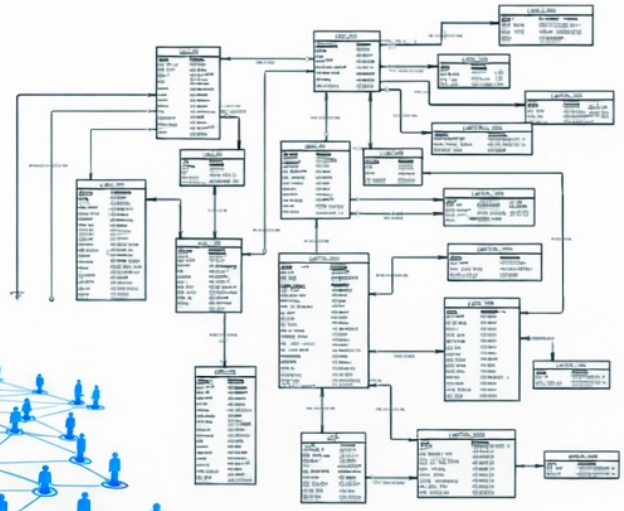
What Data Can We Mine?



in that old sea-song that he sang
so often afterwards:
'Fifteen men on the dead man's chest-yo-ho-ho, and a bottle of rum!"
in the high, old tottering voice that
seemed to have been tuned and
broken at the capstan bars. Then
he rapped on the door with a bit of
stick like a handspike that he carrie,
d, and when my father appeared, a
glass of rum. This, when it was us e

berth
the cri
the ba
and he
here a
plain
eggs is
what
and h
moug
the tir





The Plan for the Next 12 Weeks

University of Southampton

.You will learn to solve real-world problems – e.g.:

- Recommender systems
- Market Basket Analysis
- Document filtering and spam detection
- Duplicate document detection
- Link prediction
- Community detection
- Ranking search results
- Social network analysis

.You will also learn various tools & techniques - e.g.:

- Linear algebra (SVD, Eigendecomposition & PCA, NNMF, etc.)
- Optimisation (e.g. stochastic gradient descent)
- Dynamic programming (frequent itemsets)
- Hashing (LSH, Sketching, Bloom Filters)
- Statistics of regression analysis
- Information theory
- Network theory

The Group Coursework

University of Southampton

- You need to form groups
  - Target size is ~~4~~ 6 (**strictly**)
  - As a group, you need to choose a data mining problem to work on
    - (You'll need to train and evaluate models and compare their performance [possibly against approaches from others])

- Come along to the slots in week 3 to discuss your ideas for problems to work on with us

- Enter your team name and team members on the student wiki:

https://secure.ecs.soton.ac.uk/student/wiki/w/COMP6237-2025-classlist

Key Date

- Each team needs to submit a 1-page project brief by the end of week 4 (**20th of Feb**).
- Before Easter groups must present their idea and approaches to the class.
  - ❖ Teams should be prepared to present in the first slot; to ensure fairness we will pick teams at random
- Teams must submit a conference paper by **4pm on May 15**.