

COMP6237 Data Mining

Introduction to Data Mining

Jonathon Hare

jsh2@ecs.soton.ac.uk

Markus Brede

mb8@ecs.soton.ac.uk

Teaching Staff

- **Jonathon Hare**

- jsh2@ecs.soton.ac.uk
- 1/2003

- **Markus Brede**

- mb8@ecs.soton.ac.uk
- 32/4033

Module Overview

- Fairly new module - this is the third time it has run!
 - See feedback from last year
- Created to fill a gap:
 - Data mining is almost synonymous with applied machine learning
 - Inevitably some overlap in topics with COMP3206/COMP6208
 - Should be complementary & offer different views
 - Much more applied/pragmatic focus
 - How do you work with real world data?
 - How do you solve real problems?

Module Structure

- Around 28 lectures + additional tutorials
 - Wide range of data mining topics
- Assessment:
 - 50% 2 Hour Final Exam
 - 20% Individual Coursework
 - 30% Group Coursework

Coursework Timetable

- Group Coursework
 - Set today; report submission on the 16th May; presentations following that.
 - More info at the end of the lecture!
- Individual coursework
 - Set 21st Feb (week 4); due Friday 16th March (just before Easter break)

Resources

- Course web site (handouts, slides [inc interactive demos]):
 - <http://comp6237.ecs.soton.ac.uk>
- ECS Module pages (syllabus, announcements):
 - <https://secure.ecs.soton.ac.uk/module/comp6237>
- 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

What is “Data Mining”?

“Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

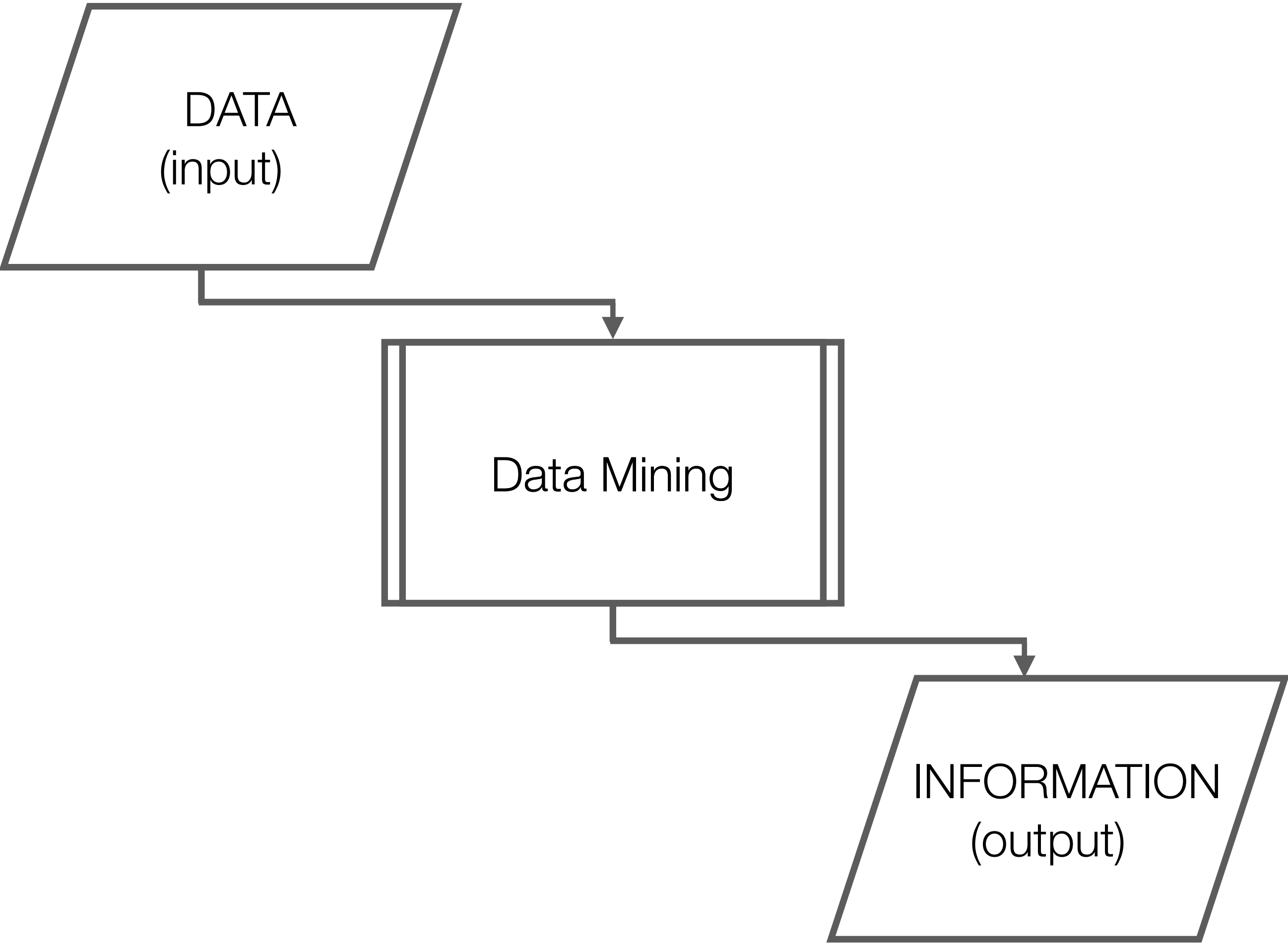
The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.”

–Wikipedia

“Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both”

**–Bill Palace, Anderson Graduate School of Management at UCLA,
1996**

DATA
(input)



```
graph TD; A[/DATA (input)/] --> B[Data Mining]; B --> C[/INFORMATION (output)/];
```

The diagram illustrates a three-step process. It begins with a parallelogram labeled 'DATA (input)' at the top left. A horizontal arrow points from this box to a central rectangular box labeled 'Data Mining'. This central box has a double-line border. From the bottom of the 'Data Mining' box, another horizontal arrow points to a final parallelogram labeled 'INFORMATION (output)' at the bottom right. The entire process is shown in a descending staircase layout.

Data Mining

INFORMATION
(output)

What is Data?

What is Data?

- Data is any sequence of one or more symbols given meaning by specific act(s) of interpretation.
- Data (or datum - a single unit of data) is not information.
- Data requires interpretation to become information.
- To translate data to information, there must be several known factors considered. The factors involved are determined by the creator of the data and the desired information.

What is Information?

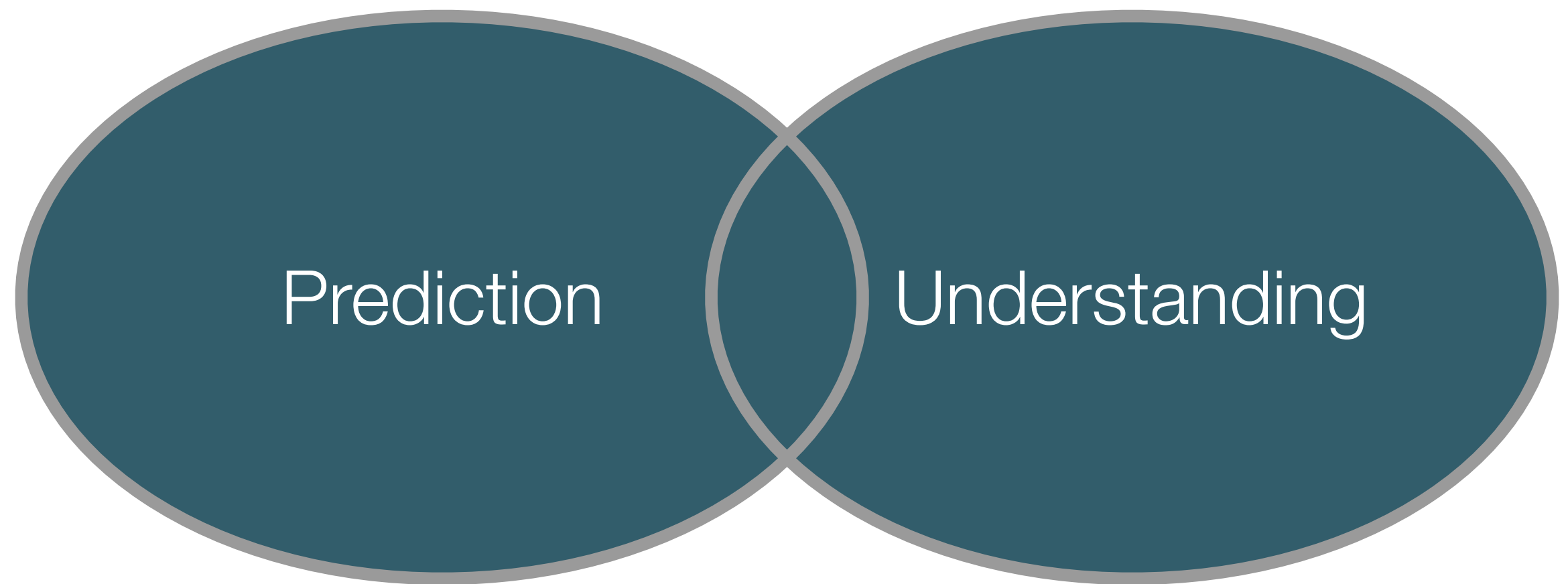
What is information?

- “Actionable knowledge”
- **Prediction**
 - Christoph Adami (Michigan State) defines information as: ‘the ability to make predictions with a likelihood better than chance’.
- **Understanding**
 - Making *sense* of the data

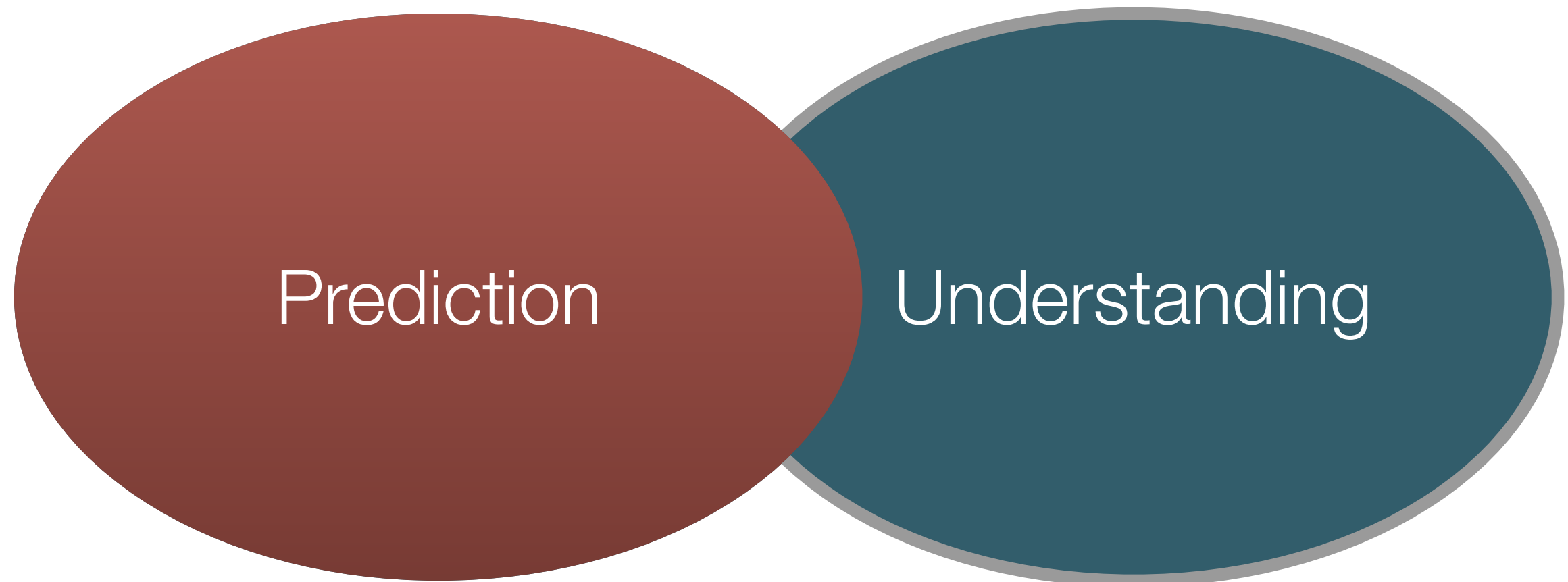
What is Data Mining?

- Given lots of data
- **Discover patterns and models that are:**
 - **Valid:** hold on new data with some certainty
 - **Useful:** should be possible to act on the item
 - **Unexpected:** non-obvious to the system
 - **Understandable:** humans should be able to interpret the pattern

Two complementary goals of data mining

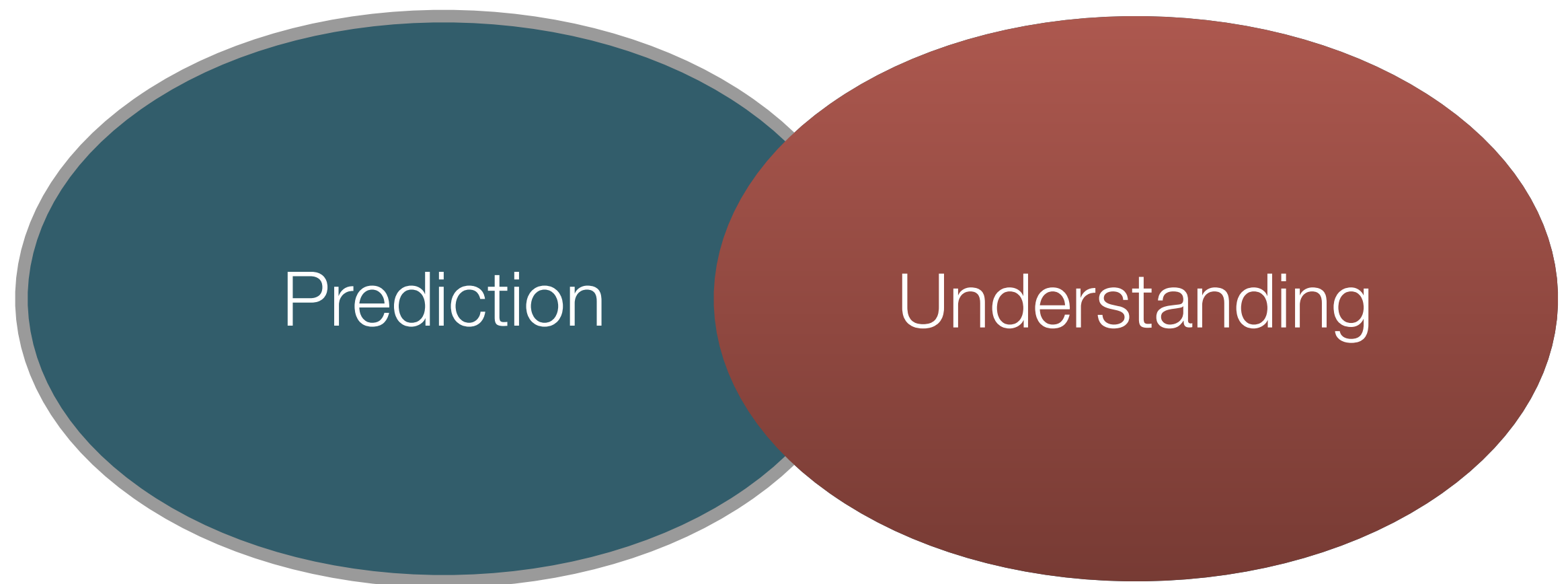


Two complementary goals of data mining



Use some variables to predict unknown
or future values of other variables

Two complementary goals of data mining



Find human-interpretable patterns that describe the data

What kinds of data are we interested in mining?

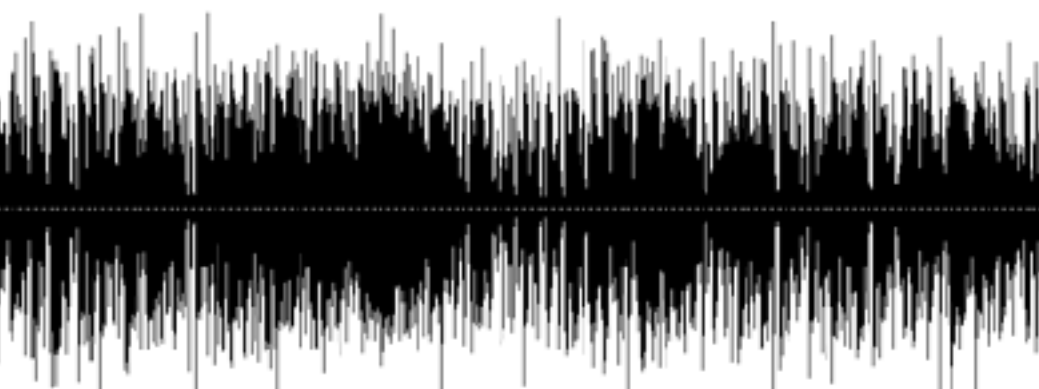
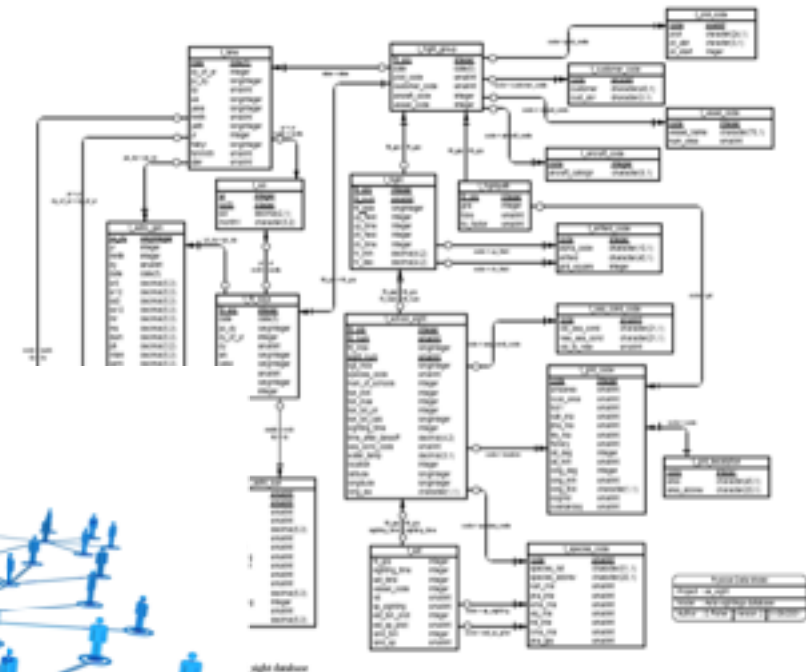


	Jan	Feb	Mar	Apr	May	Jan	Jul
Product 1	\$5,000.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,000.00	\$8,054.00	\$12,100.00
Budget	\$4,700.00	\$5,478.00	\$4,754.00	\$5,501.00	\$7,744.00	\$8,641.00	\$11,976.00
Over / (Under Budget)	\$300.00	\$772.00	\$346.00	\$1,349.00	\$2,256.00	\$-587.00	\$1,124.00
Product 2	\$5,000.00	\$6,250.00	\$5,400.00	\$6,100.00	\$7,700.00	\$7,694.00	\$11,600.00
Budget	\$4,595.00	\$5,478.00	\$4,754.00	\$5,501.00	\$7,744.00	\$8,641.00	\$11,976.00
Over / (Under Budget)	\$405.00	\$772.00	\$646.00	\$599.00	\$-44.00	\$-947.00	\$-376.00
Product 3	\$14,000.00	\$16,250.00	\$15,100.00	\$16,850.00	\$20,000.00	\$20,764.00	\$30,400.00
Budget	\$13,295.00	\$15,478.00	\$14,754.00	\$15,701.00	\$19,834.00	\$21,311.00	\$31,209.00
Over / (Under Budget)	\$705.00	\$772.00	\$346.00	\$1,149.00	\$2,166.00	\$-547.00	\$919.00
Product 4	\$16,000.00	\$17,250.00	\$16,200.00	\$17,900.00	\$21,000.00	\$21,404.00	\$32,800.00
Budget	\$15,000.00	\$16,478.00	\$15,754.00	\$16,701.00	\$20,834.00	\$21,811.00	\$32,209.00
Over / (Under Budget)	\$1,000.00	\$772.00	\$446.00	\$1,199.00	\$1,166.00	\$-407.00	\$591.00
Product 5	\$78,000.00	\$88,250.00	\$89,000.00	\$96,850.00	\$96,000.00	\$106,834.00	\$121,200.00
Budget	\$68,595.00	\$78,478.00	\$78,754.00	\$86,501.00	\$77,744.00	\$89,841.00	\$111,976.00
Over / (Under Budget)	\$9,405.00	\$9,772.00	\$10,246.00	\$10,349.00	\$18,256.00	\$17,000.00	\$9,224.00

g back
nd,
is still
e up
17—
en my
bow
man
up his
ere
ding

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 carried, and when my fa-
ther appeared, called roughly for
a glass of rum. This, when it was

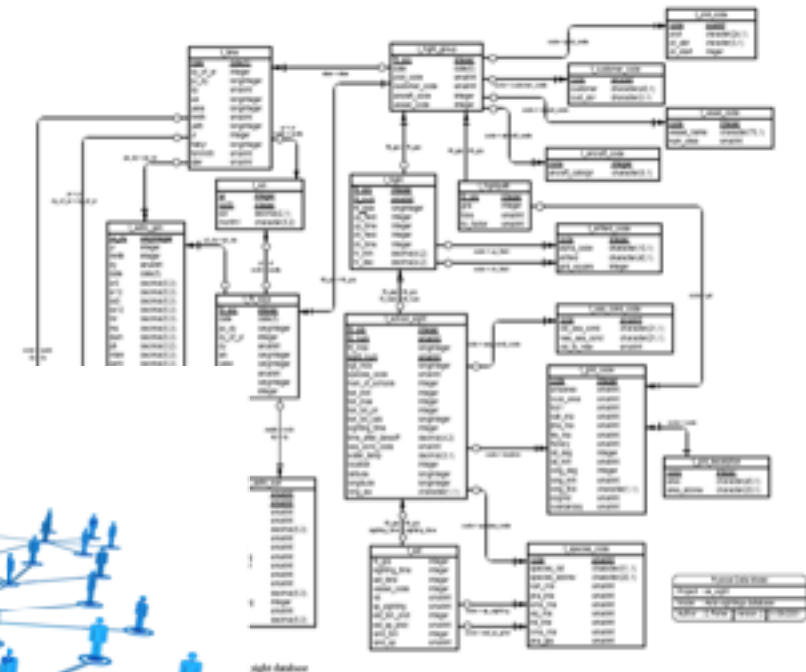
berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre



Categorising data: Structured/unstructured



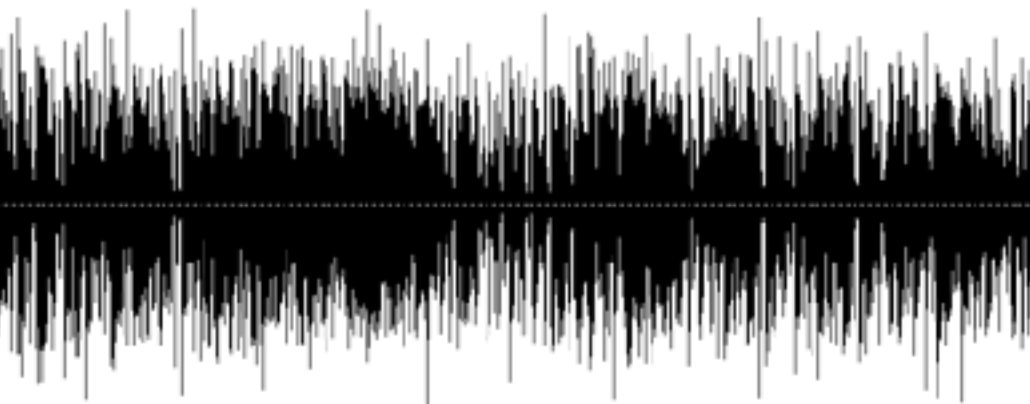
Sales							
Period Ending:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,500.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,300.00	\$8,050.00	\$12,100.00
Budget	\$4,700.00	\$5,475.00	\$4,750.00	\$5,500.00	\$7,744.00	\$8,640.00	\$10,975.00
Over / (Under Budget)	\$800.00	\$775.00	\$350.00	\$1,350.00	\$555.60	\$-590.00	\$1,125.00
Product 2	\$5,500.00	\$6,250.00	\$5,400.00	\$6,100.00	\$7,700.00	\$7,600.00	\$11,600.00
Budget	\$4,500.00	\$5,475.00	\$4,750.00	\$5,500.00	\$7,744.00	\$8,640.00	\$10,975.00
Over / (Under Budget)	\$1,000.00	\$775.00	\$650.00	\$600.00	\$-44.00	\$-1,040.00	\$625.00
Product 3	\$14,000.00	\$16,250.00	\$15,100.00	\$16,850.00	\$20,100.00	\$19,700.00	\$28,400.00
Budget	\$12,000.00	\$13,475.00	\$12,750.00	\$14,500.00	\$17,744.00	\$18,640.00	\$23,200.00
Over / (Under Budget)	\$2,000.00	\$2,775.00	\$2,350.00	\$2,350.00	\$2,356.00	\$1,060.00	\$5,200.00
Product 4	\$16,000.00	\$17,250.00	\$16,000.00	\$17,000.00	\$20,000.00	\$19,400.00	\$27,800.00
Budget	\$14,000.00	\$15,475.00	\$14,750.00	\$16,500.00	\$19,744.00	\$18,640.00	\$23,200.00
Over / (Under Budget)	\$2,000.00	\$1,775.00	\$1,250.00	\$500.00	\$2,256.00	\$760.00	\$4,600.00
Product 5	\$78,000.00	\$88,250.00	\$89,000.00	\$95,850.00	\$96,800.00	\$106,830.00	\$121,200.00
Budget	\$68,500.00	\$78,500.00	\$78,750.00	\$85,500.00	\$97,744.00	\$99,840.00	\$115,975.00
Over / (Under Budget)	\$9,500.00	\$9,750.00	\$10,250.00	\$10,350.00	\$19,056.00	\$7,990.00	\$5,225.00



g back
nd,
is still
e up
17—
en my
brow
nan
up his
ere
ding

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 carried, and when my fa-
ther appeared, called roughly for
a glass of rum. This, when it was

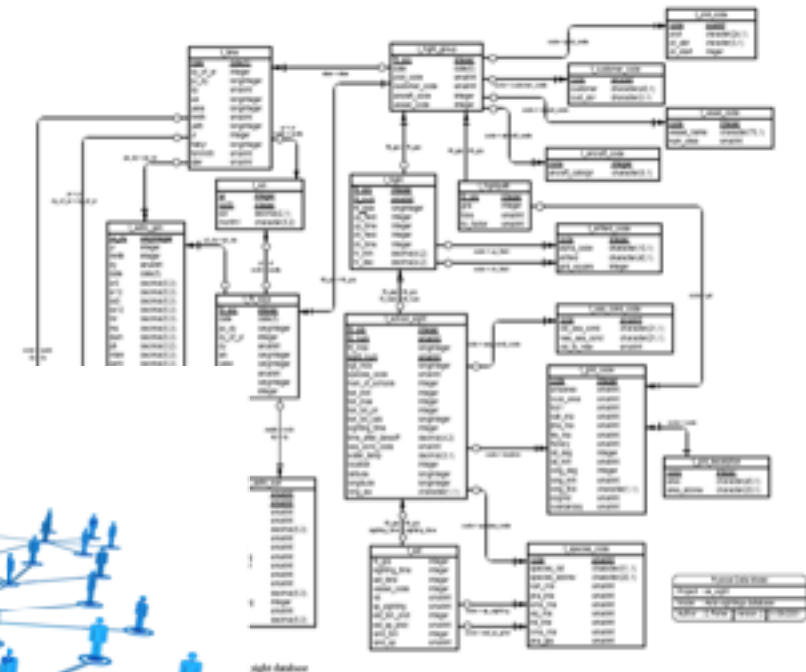
berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre



Categorising data: Dynamic/static



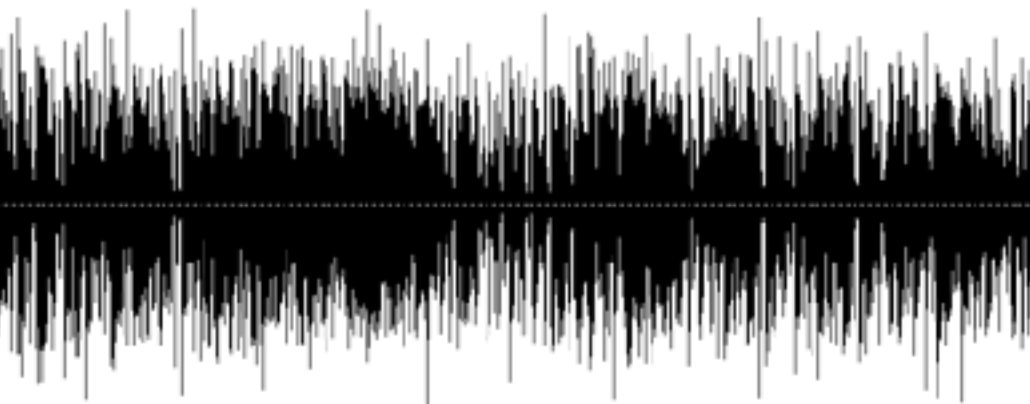
Sales							
Period Ending:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,000.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,000.00	\$8,050.00	\$12,100.00
Budget	\$4,700.00	\$5,678.00	\$4,754.00	\$6,501.00	\$7,744.00	\$8,641.00	\$11,976.00
Over / (Under Budget)	\$300.00	\$572.00	\$346.00	\$1,349.00	\$2,256.00	\$1,409.00	\$1,124.00
Product 2	\$5,000.00	\$6,250.00	\$5,400.00	\$6,100.00	\$7,800.00	\$7,600.00	\$11,600.00
Budget	\$4,995.00	\$5,678.00	\$4,754.00	\$6,500.00	\$7,743.00	\$8,640.00	\$11,975.00
Over / (Under Budget)	\$505.00	\$572.00	\$646.00	\$600.00	\$2,057.00	\$960.00	\$1,625.00
Product 3	\$14,000.00	\$16,250.00	\$15,100.00	\$16,850.00	\$18,000.00	\$18,700.00	\$26,400.00
Budget	\$13,800.00	\$15,678.00	\$14,754.00	\$16,201.00	\$17,444.00	\$18,311.00	\$25,209.00
Over / (Under Budget)	\$200.00	\$572.00	\$346.00	\$649.00	\$2,256.00	\$3,389.00	\$1,191.00
Product 4	\$16,000.00	\$17,250.00	\$16,200.00	\$17,900.00	\$19,000.00	\$19,700.00	\$28,800.00
Budget	\$15,800.00	\$16,678.00	\$15,754.00	\$17,501.00	\$18,744.00	\$19,611.00	\$27,209.00
Over / (Under Budget)	\$200.00	\$572.00	\$446.00	\$399.00	\$2,256.00	\$1,089.00	\$1,591.00
Product 5	\$78,000.00	\$88,250.00	\$89,000.00	\$96,850.00	\$98,000.00	\$106,800.00	\$121,200.00
Budget	\$68,500.00	\$78,500.00	\$78,754.00	\$88,501.00	\$97,744.00	\$109,841.00	\$111,976.00
Over / (Under Budget)	\$9,500.00	\$9,750.00	\$10,246.00	\$8,349.00	\$1,256.00	\$16,959.00	\$10,224.00



g back
nd,
is still
e up
17—
en my
brow
nan
up his
ere
ding

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 carried, and when my fa-
ther appeared, called roughly for
a glass of rum. This, when it was

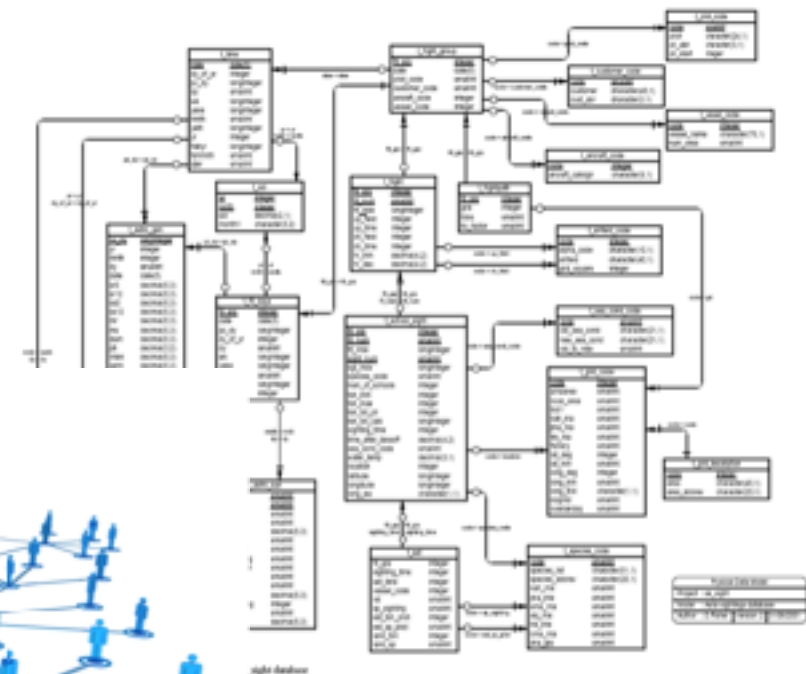
berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre



Categorising *data mining*: Unimodal/multimodal



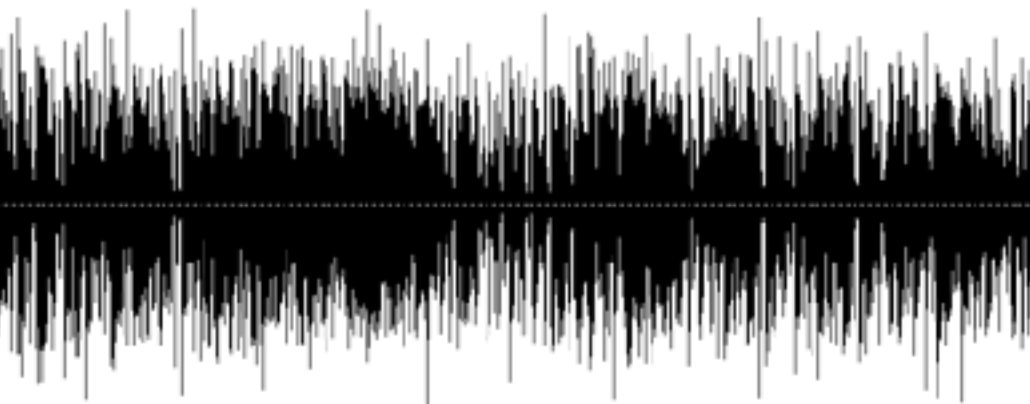
Sales							
Period Ending:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,000.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,000.00	\$8,054.00	\$12,100.00
Budget	\$4,700.00	\$5,478.00	\$4,754.00	\$5,501.00	\$7,744.00	\$8,641.00	\$11,976.00
Over / (Under Budget)	\$300.00	\$772.00	\$346.00	\$1,349.00	\$2,256.00	\$1,203.00	\$1,124.00
Product 2	\$5,000.00	\$6,250.00	\$5,400.00	\$6,100.00	\$19,100.00	\$17,604.00	\$11,600.00
Budget	\$4,500.00	\$5,478.00	\$4,754.00	\$5,501.00	\$18,100.00	\$17,321.00	\$11,808.00
Over / (Under Budget)	\$500.00	\$772.00	\$646.00	\$599.00	\$1,000.00	\$283.00	\$792.00
Product 3	\$14,000.00	\$16,250.00	\$15,100.00	\$16,850.00	\$22,100.00	\$26,764.00	\$36,400.00
Budget	\$13,500.00	\$15,478.00	\$14,754.00	\$17,201.00	\$20,834.00	\$21,311.00	\$31,208.00
Over / (Under Budget)	\$500.00	\$772.00	\$346.00	\$1,649.00	\$1,266.00	\$5,453.00	\$5,192.00
Product 4	\$16,000.00	\$17,250.00	\$16,200.00	\$17,900.00	\$20,000.00	\$21,404.00	\$24,800.00
Budget	\$15,500.00	\$16,478.00	\$15,754.00	\$16,701.00	\$19,834.00	\$21,211.00	\$23,208.00
Over / (Under Budget)	\$500.00	\$772.00	\$446.00	\$1,199.00	\$1,166.00	\$1,193.00	\$1,592.00
Product 5	\$78,000.00	\$88,250.00	\$89,000.00	\$95,850.00	\$96,000.00	\$106,834.00	\$121,200.00
Budget	\$68,500.00	\$78,595.00	\$78,754.00	\$85,501.00	\$77,744.00	\$89,841.00	\$111,976.00
Over / (Under Budget)	\$9,500.00	\$9,655.00	\$10,246.00	\$10,349.00	\$18,256.00	\$16,993.00	\$9,224.00



g back
nd,
is still
e up
17—
en my
brow
nan
up his
ere
ding

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 carried, and when my fa-
ther appeared, called roughly for
a glass of rum. This, when it was

berth f
he crie
the bar
and he
here a
plain n
eggs is
up the
What y
mough
see wh
he thre



What is the *typical* data mining pipeline?

DATA
(input)

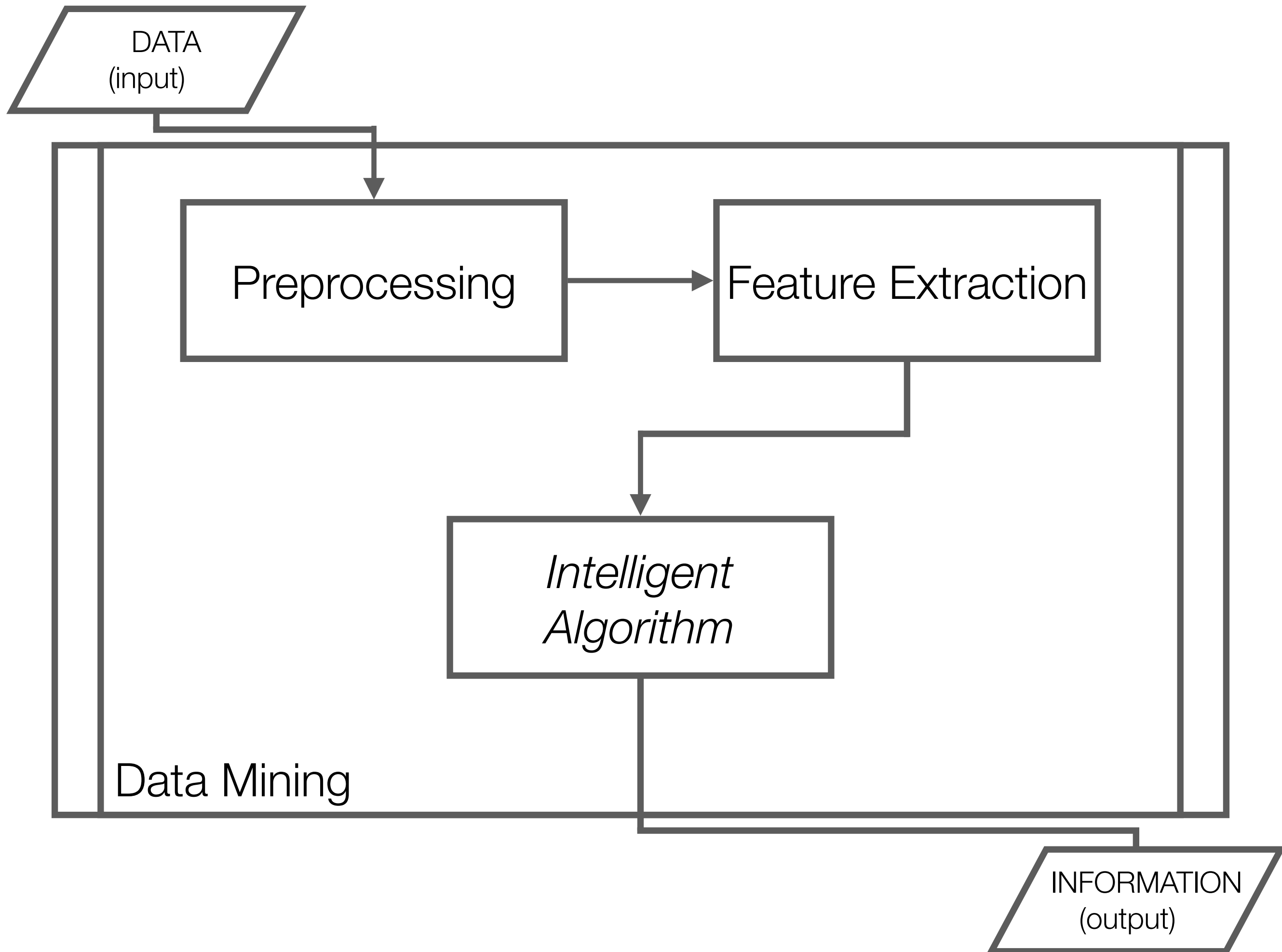


```
graph TD; A[/DATA (input)/] --> B[Data Mining]; B --> C[/INFORMATION (output)/];
```

The diagram illustrates a three-step process. It begins with an input stage labeled 'DATA (input)' in a parallelogram. An arrow leads from this stage to a central rectangular box labeled 'Data Mining'. From the 'Data Mining' box, another arrow points to the final stage, 'INFORMATION (output)', which is also in a parallelogram. The flow is linear and sequential, showing the transformation of raw data into useful information through the process of data mining.

Data Mining

INFORMATION
(output)



Descriptive Techniques

PCA

ICA

MDS

Clustering

Anomaly Detection

...

*Intelligent
Algorithm*

Predictive Techniques

Classification

Ranking

Regression

Matrix Completion

...

The plan for the next 12 weeks

...we're going to look at a range of topics...

- **You will learn to solve real-world problems** - e.g.:
 - Recommender systems
 - Market Basket Analysis
 - Document filtering and spam detection
 - Duplicate document detection
- **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)

The Group Coursework

- You need to form groups
- Target size is 6 (+/- 1) people
- As a group you need to choose a **predictive** 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 Tuesday slot next week to discuss your ideas for problems to work on with us

- Key dates:
 - Each team must submit a 1-page project brief by the end of the day on the 16th Feb.
 - In week 7 teams must present their idea and approaches to the class.
 - Teams should be prepared to present in the first slot; to ensure fairness we'll pick teams at random.
 - Teams must submit a conference paper by 4PM on the 16th May