

*COMP6237 Data Mining*

# Introduction to Data Mining

---

Jonathon Hare

[jsh2@ecs.soton.ac.uk](mailto:jsh2@ecs.soton.ac.uk)

Markus Brede

[mb8@ecs.soton.ac.uk](mailto:mb8@ecs.soton.ac.uk)

# Teaching Staff

---

- **Jonathon Hare**

- [jsh2@ecs.soton.ac.uk](mailto:jsh2@ecs.soton.ac.uk)
- 32/3023

- **Markus Brede**

- [mb8@ecs.soton.ac.uk](mailto:mb8@ecs.soton.ac.uk)
- 32/4033

# Module Overview

---

- Completely new module - this is the first time it has run!
- 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 24 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 5th May; presentations following that.
  - More info at the end of the lecture!
- Individual coursework
  - Set 15th Feb (week 4); due 17th 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.

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.”*

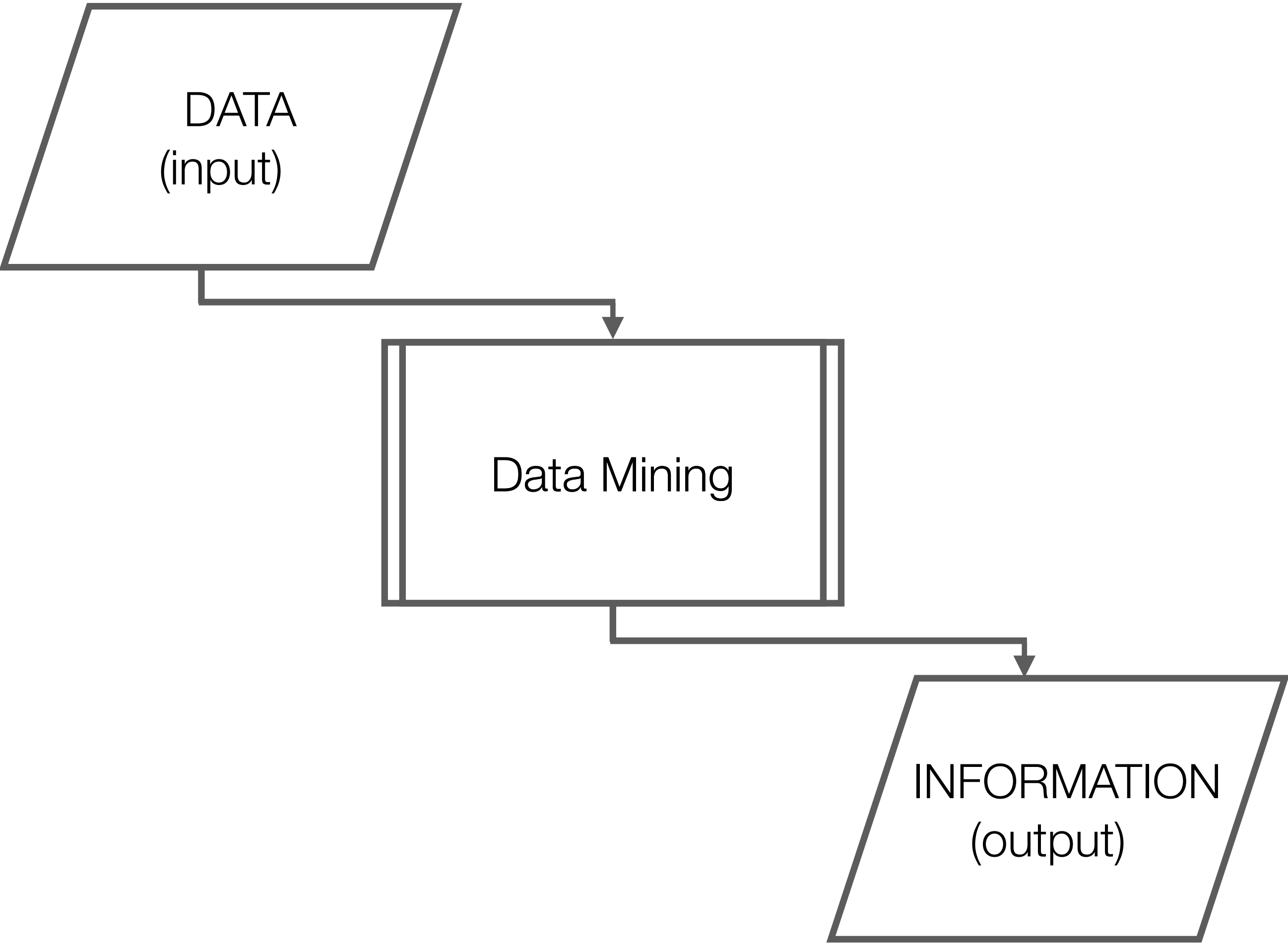
**–Wikipedia**



“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 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)

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

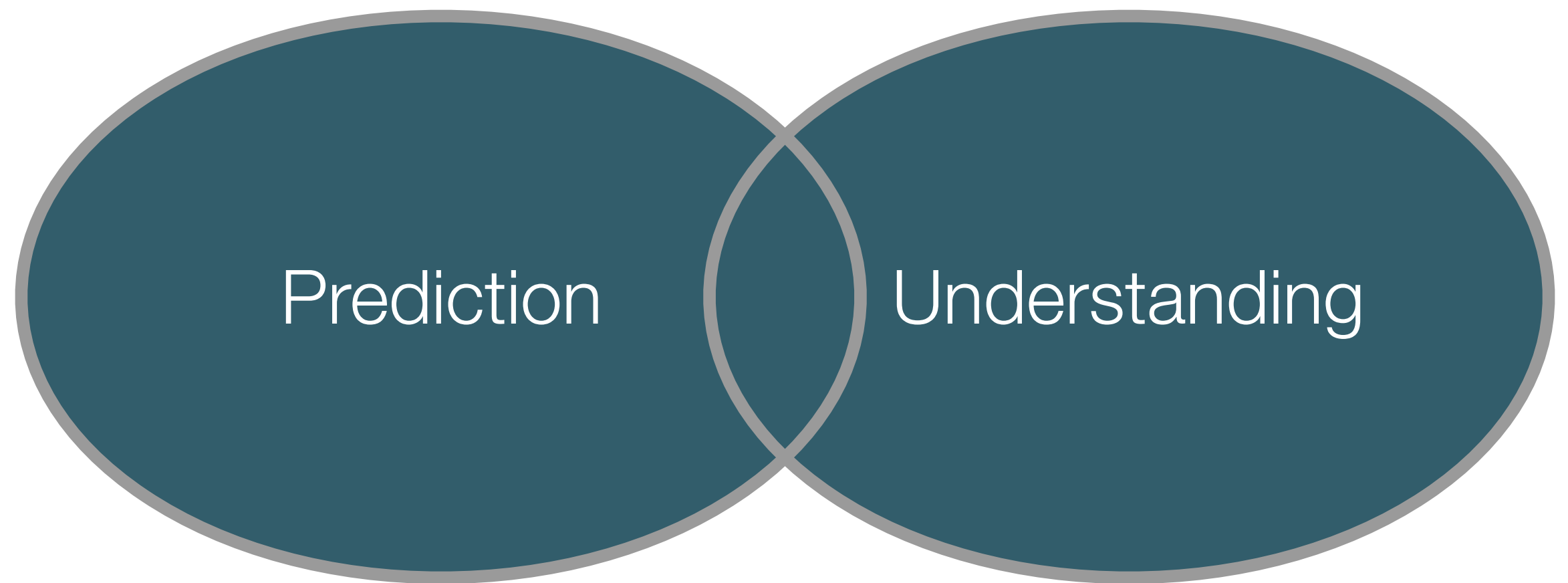
# What is Data Mining?

---

- 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

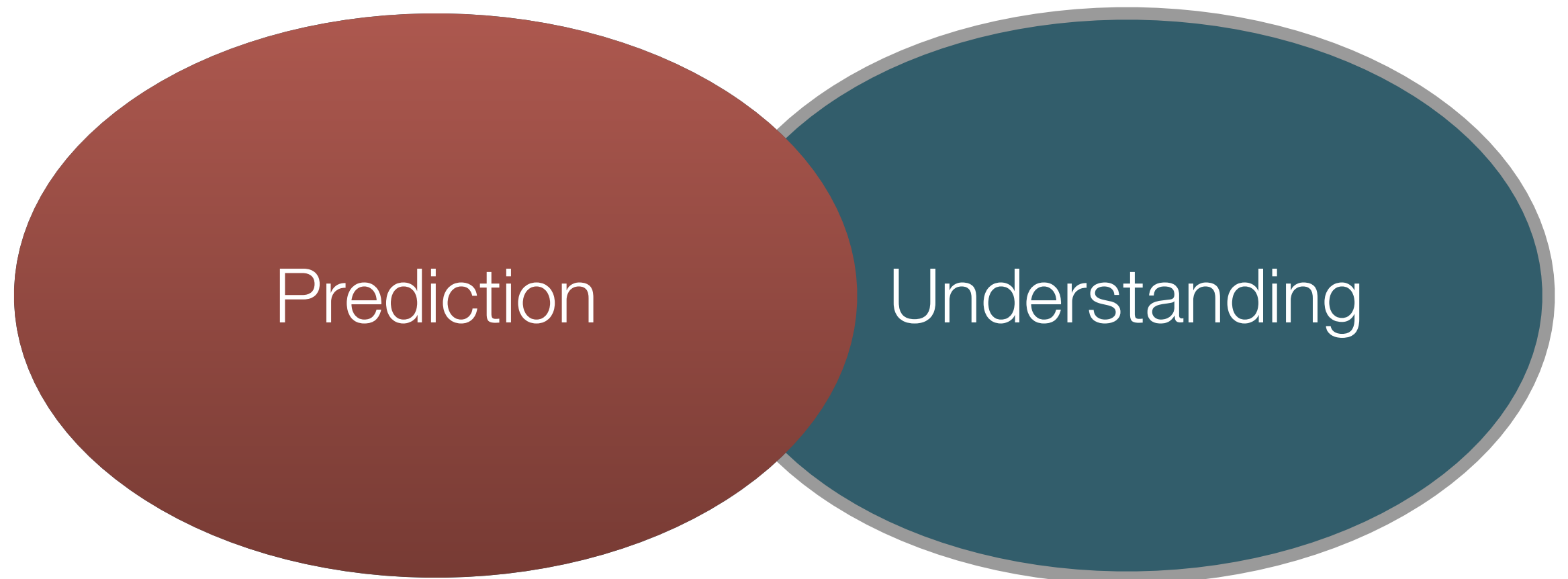
---





# Two complementary goals of data mining

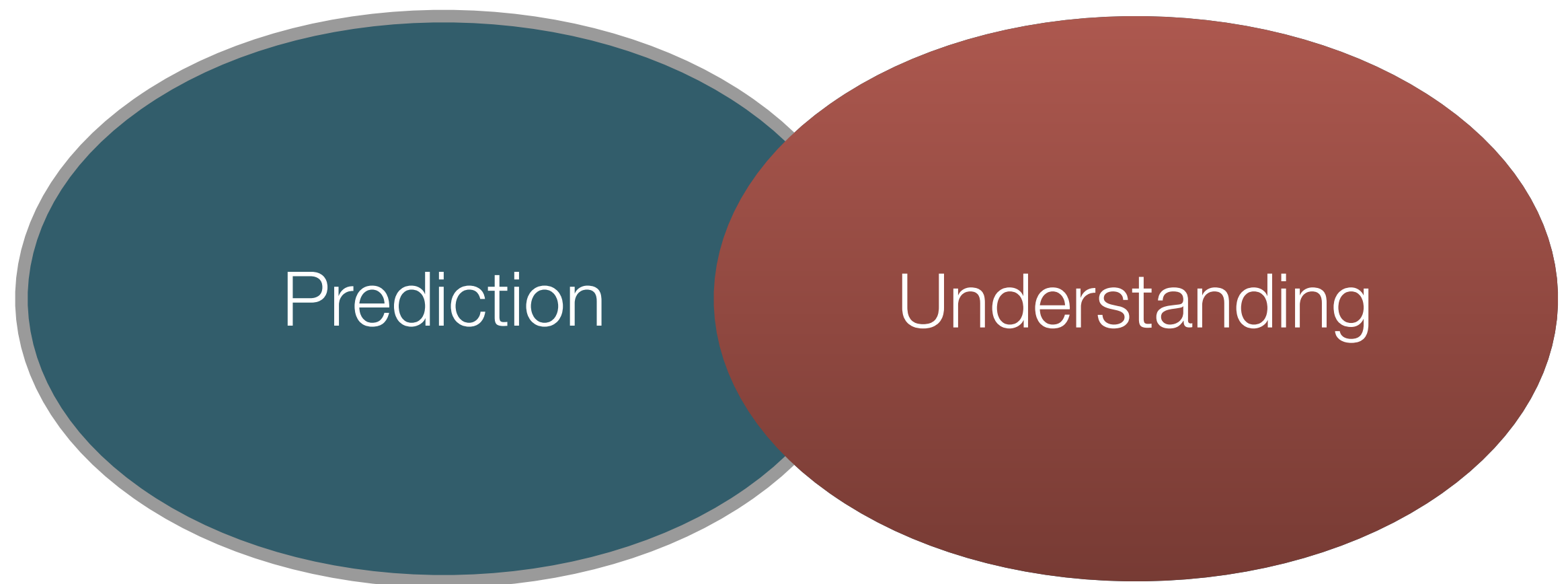
---



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?



MainWindow

File Home Others

Cut Copy Paste

Font

Number

Cells

Sales

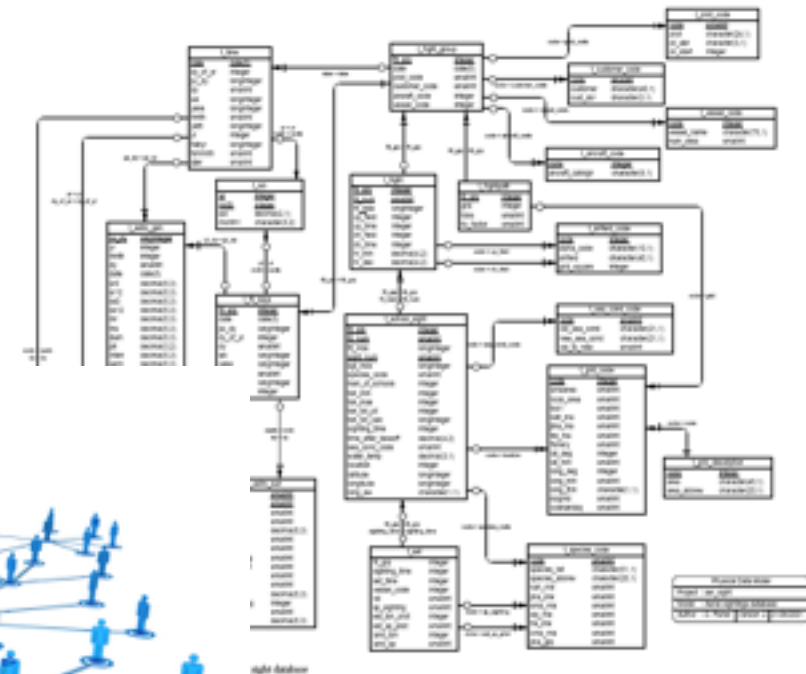
Period Starting:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,600.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 2	\$5,600.00	\$6,250.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 3	\$14,600.00	\$25,250.00	\$25,100.00	\$26,850.00	\$32,100.00	\$35,750.00	\$38,480.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$20,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$5,705.00	\$10,572.00	\$9,346.00	\$9,259.00	\$11,670.00	\$14,431.00	\$15,181.00
Product 4	\$16,850.00	\$17,890.00	\$18,830.00	\$19,290.00	\$20,890.00	\$21,439.00	\$22,364.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$20,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$7,955.00	\$3,212.00	\$3,076.00	\$1,700.00	\$1,460.00	\$1,120.00	\$9,065.00
Product 5	\$78,600.00	\$88,750.00	\$89,002.00	\$86,850.00	\$96,400.00	\$106,838.00	\$127,280.00
Budget	\$68,595.00	\$78,595.00	\$78,754.00	\$86,591.00	\$77,744.00	\$86,845.00	\$115,875.00
Over / (Under Budget)	\$10,005.00	\$10,155.00	\$10,248.00	\$1,259.00	\$18,656.00	\$20,000.00	\$11,405.00

Sales Product1 Product2 Product3 Product4 Product5 Conditional Formatting FormulaSupport

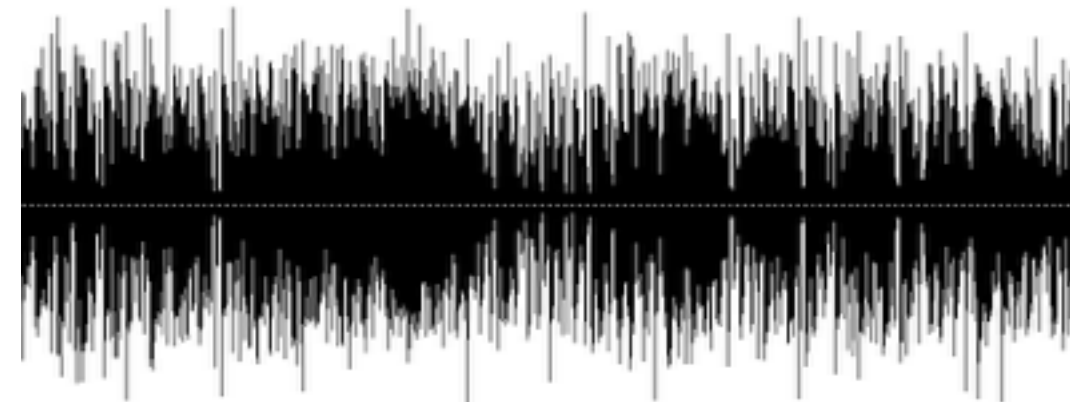
g back  
nd,  
is still  
e up  
17—  
en my  
bow  
an  
up his  
ere  
ding

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



right hand

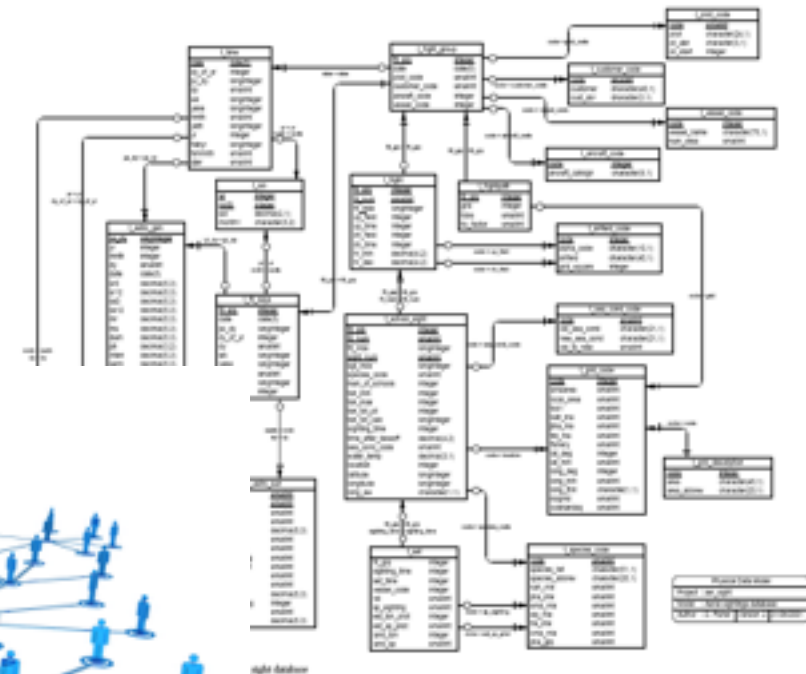




# Categorising data: Structured/unstructured



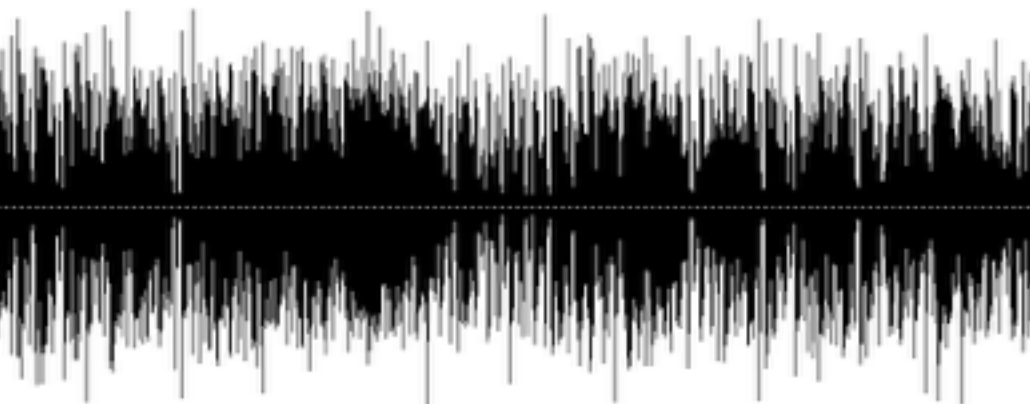
Sales							
Period Starting:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,600.00	\$6,350.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 2	\$5,600.00	\$6,350.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 3	\$14,600.00	\$25,350.00	\$25,100.00	\$26,850.00	\$32,100.00	\$35,750.00	\$38,480.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$20,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$5,705.00	\$10,672.00	\$9,346.00	\$9,259.00	\$11,670.00	\$14,431.00	\$15,181.00
Product 4	\$16,850.00	\$17,890.00	\$18,830.00	\$19,290.00	\$20,890.00	\$21,439.00	\$22,364.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$20,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$7,955.00	\$3,212.00	\$3,076.00	\$1,700.00	\$1,460.00	\$1,120.00	\$9,065.00
Product 5	\$78,600.00	\$88,750.00	\$89,000.00	\$85,850.00	\$96,400.00	\$106,830.00	\$127,280.00
Budget	\$68,595.00	\$78,595.00	\$78,754.00	\$85,591.00	\$77,744.00	\$86,845.00	\$115,875.00
Over / (Under Budget)	\$10,005.00	\$10,155.00	\$10,246.00	\$10,259.00	\$18,656.00	\$20,985.00	\$11,405.00



g back  
id,  
is still  
e up  
17—  
en my  
ibow  
ian  
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

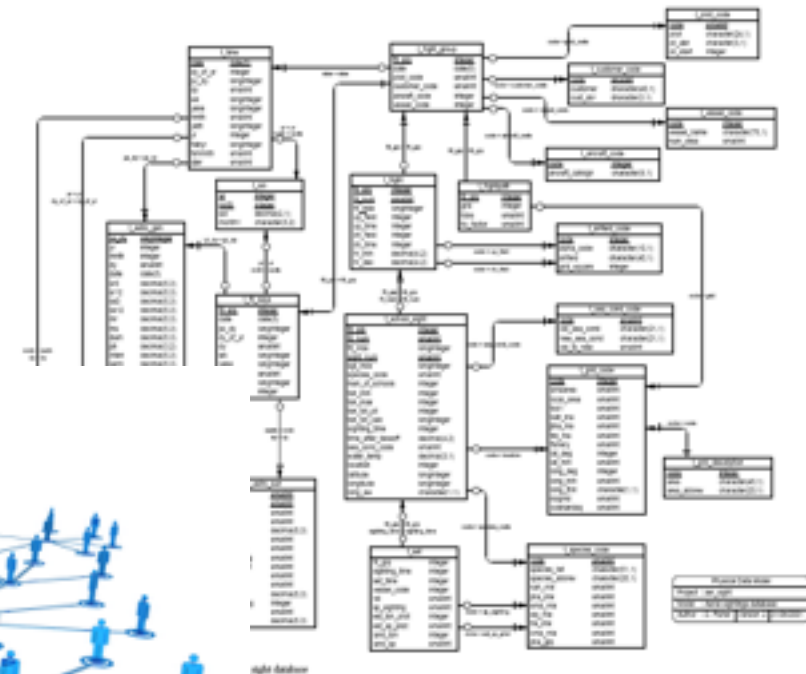




# Categorising data: Dynamic/static



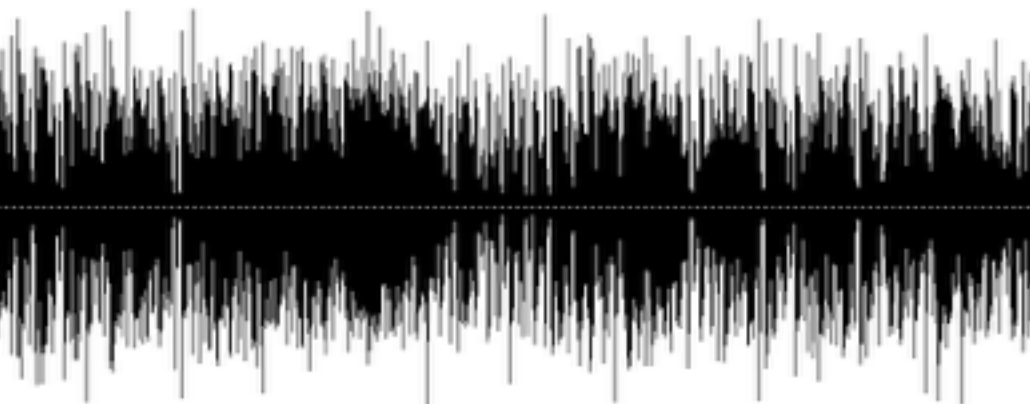
Sales							
Period Starting:	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,600.00	\$6,350.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 2	\$5,600.00	\$6,350.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 3	\$14,600.00	\$25,350.00	\$25,100.00	\$26,850.00	\$32,100.00	\$35,750.00	\$38,480.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$19,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$5,705.00	\$10,672.00	\$9,346.00	\$9,259.00	\$12,670.00	\$14,431.00	\$15,181.00
Product 4	\$16,850.00	\$17,890.00	\$18,830.00	\$19,290.00	\$20,890.00	\$21,439.00	\$22,364.00
Budget	\$8,895.00	\$14,678.00	\$15,754.00	\$17,591.00	\$19,430.00	\$21,319.00	\$23,299.00
Over / (Under Budget)	\$7,955.00	\$3,212.00	\$3,076.00	\$1,700.00	\$1,460.00	\$1,120.00	\$9,065.00
Product 5	\$78,600.00	\$88,750.00	\$89,000.00	\$85,850.00	\$96,400.00	\$106,830.00	\$127,280.00
Budget	\$68,595.00	\$78,595.00	\$78,754.00	\$85,591.00	\$77,744.00	\$86,845.00	\$115,875.00
Over / (Under Budget)	\$10,005.00	\$10,155.00	\$10,246.00	\$3,259.00	\$18,656.00	\$20,985.00	\$11,405.00



g back  
nd,  
is still  
e up  
17—  
en my  
bow  
an  
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

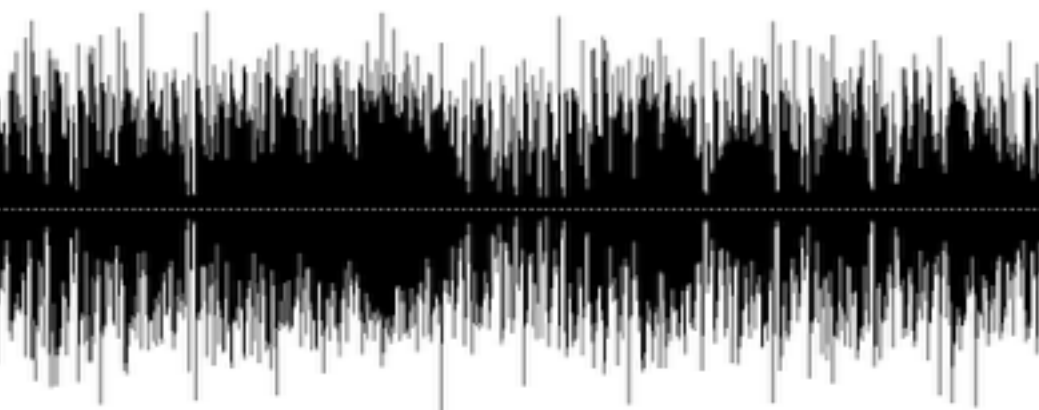
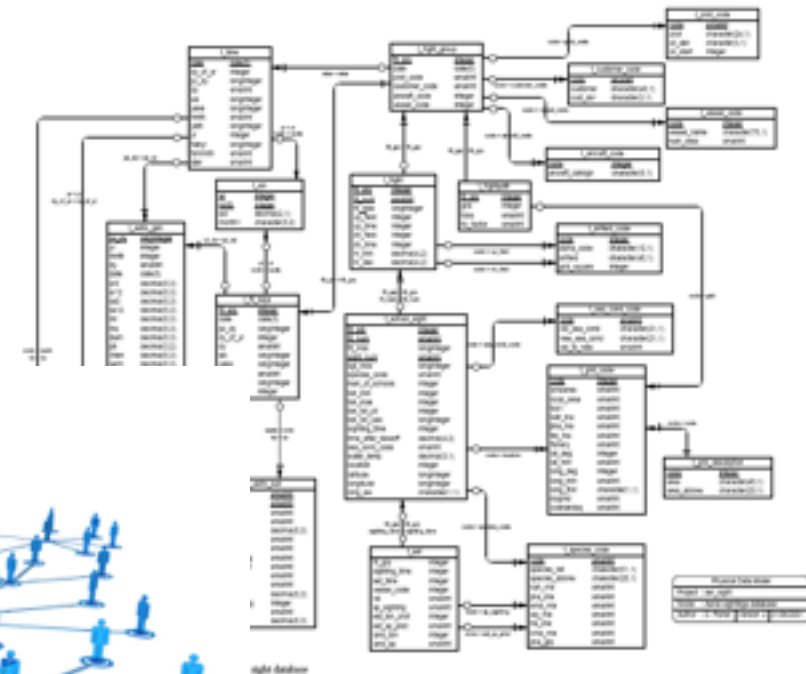


	Jan	Feb	Mar	Apr	May	Jun	Jul
Product 1	\$5,600.00	\$6,350.00	\$5,100.00	\$6,850.00	\$8,600.00	\$8,850.00	\$12,180.00
Budget	\$4,790.00	\$5,678.00	\$4,754.00	\$5,591.00	\$7,744.00	\$8,645.00	\$10,875.00
Over / (Under Budget)	\$1,810.00	\$672.00	\$346.00	\$1,259.00	\$856.00	\$205.00	\$1,305.00
Product 2	\$5,600.00	\$6,350.00	\$18,400.00	\$16,830.00	\$19,290.00	\$17,690.00	\$15,600.00
Budget	\$4,195.00	\$5,678.00	\$9,754.00	\$9,168.00	\$18,747.50	\$14,327.00	\$14,898.50
Over / (Under Budget)	\$1,405.00	\$672.00	\$8,646.00	\$7,662.00	\$5,542.50	\$3,363.00	\$801.50
Product 3	\$14,600.00	\$25,350.00	\$25,100.00	\$26,850.00	\$32,100.00	\$35,750.00	\$38,480.00
Budget	\$5,895.00	\$3,678.00	\$5,754.00	\$7,591.00	\$9,430.00	\$11,319.00	\$13,299.00
Over / (Under Budget)	\$8,705.00	\$21,672.00	\$19,346.00	\$19,259.00	\$22,670.00	\$24,431.00	\$25,181.00
Product 4	\$16,850.00	\$17,690.00	\$18,830.00	\$19,290.00	\$20,690.00	\$21,439.00	\$22,364.00
Budget	\$8,895.00	\$14,678.00	\$7,754.00	\$6,591.00	\$9,430.00	\$11,319.00	\$13,299.00
Over / (Under Budget)	\$7,955.00	\$3,012.00	\$11,076.00	\$12,700.00	\$11,260.00	\$10,120.00	\$9,065.00
Product 5	\$78,600.00	\$88,750.00	\$89,000.00	\$86,850.00	\$96,400.00	\$106,830.00	\$127,280.00
Budget	\$68,595.00	\$78,595.00	\$78,754.00	\$86,591.00	\$77,744.00	\$86,845.00	\$115,875.00
Over / (Under Budget)	\$10,005.00	\$10,155.00	\$10,246.00	\$3,259.00	\$18,656.00	\$20,000.00	\$11,405.00

g back  
id,  
is still  
e up  
17—  
en my  
ibow  
ian  
up his  
ere  
ding

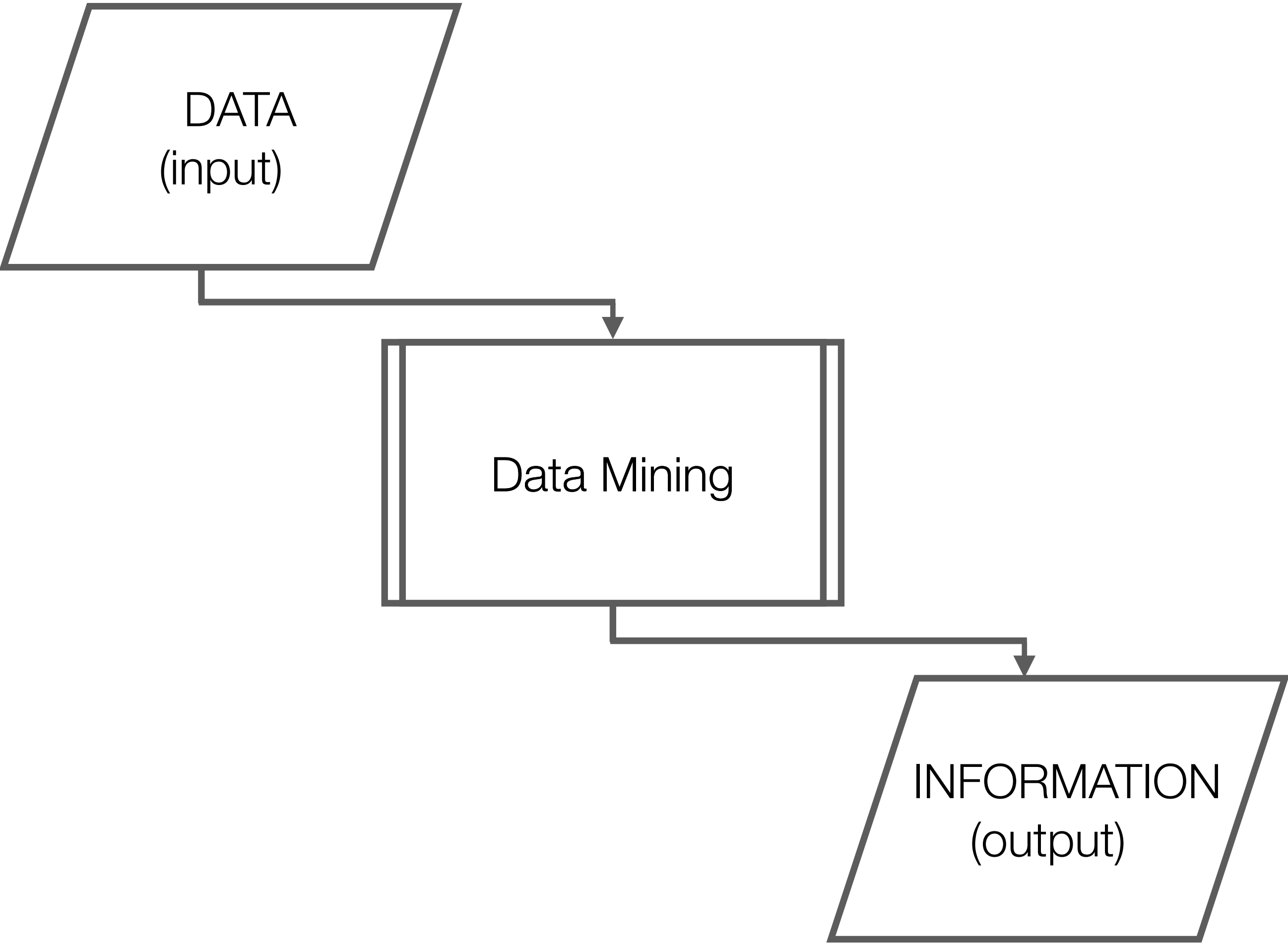
in that old sea-song that he sang  
so often afterwards:  
*'Fifteen men on the dead man's  
chest—Yo-ho-ho, and a bottle of  
rum!'* in the high, old tottering  
voice that seemed to have been  
tuned and broken at the capstan  
bars. Then he rapped on the door  
with a bit of stick like a handspike  
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)



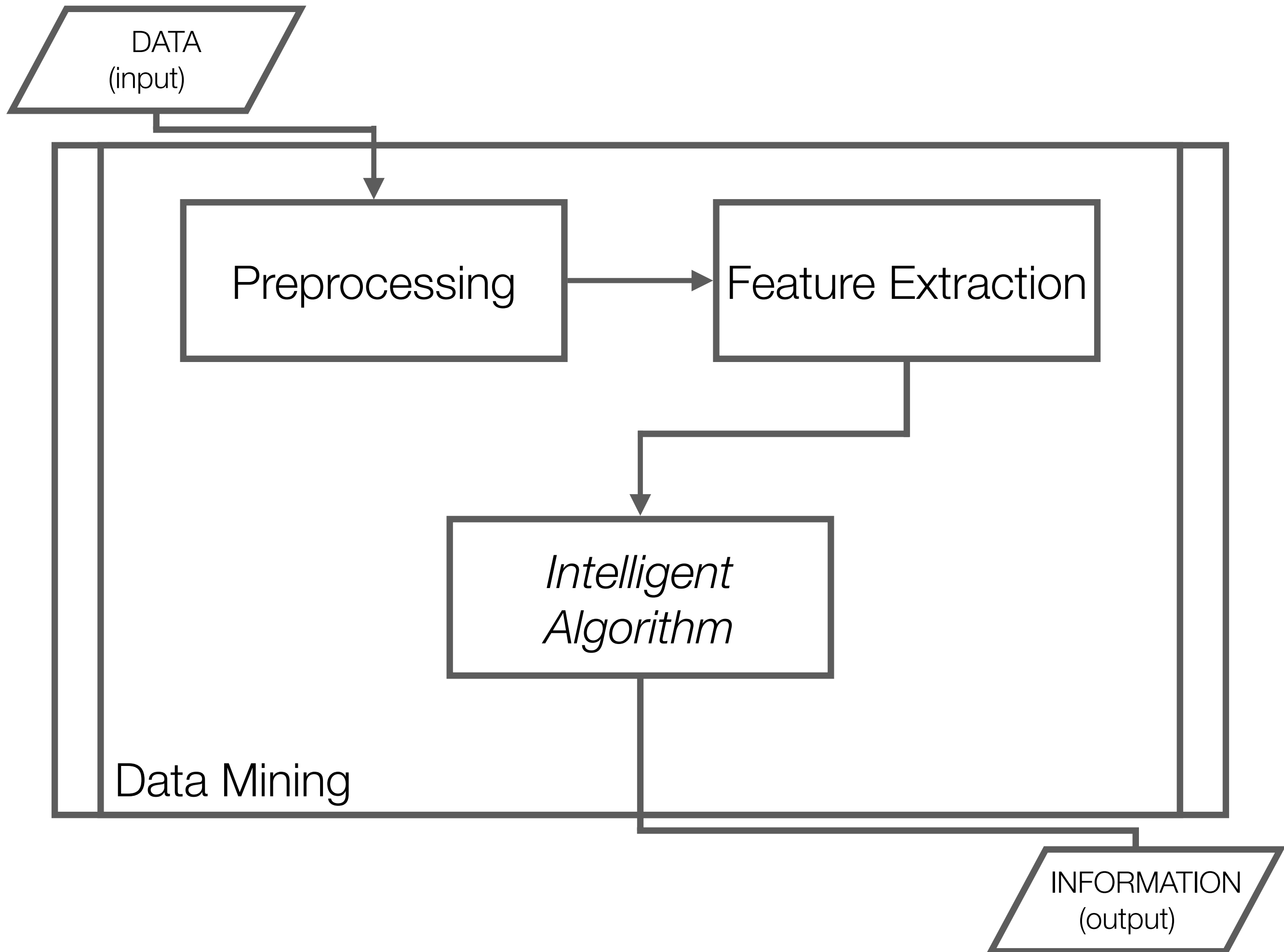
```
graph TD; A[/DATA (input)/] --> B[Data Mining]; B --> C[/INFORMATION (output)/];
```

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 5-6 people
- As a group you need to chose 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 tutorial slot this 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 12th Feb.
  - On the 19th Feb each team must pitch their project to the class (2 minutes to pitch; 3 for Q&A)
  - Teams must submit a conference paper by 4PM on the 5th May
  - Projects will be presented in the lecture/tutorial slots in week 11 & 12
    - Presentation timetable to be published at a later date