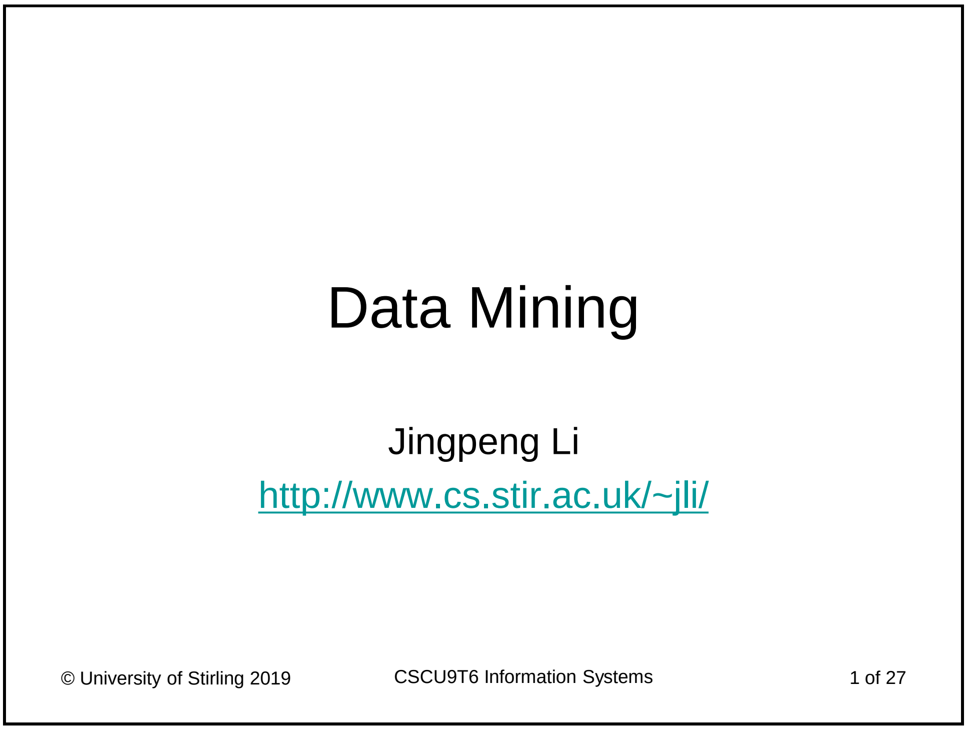
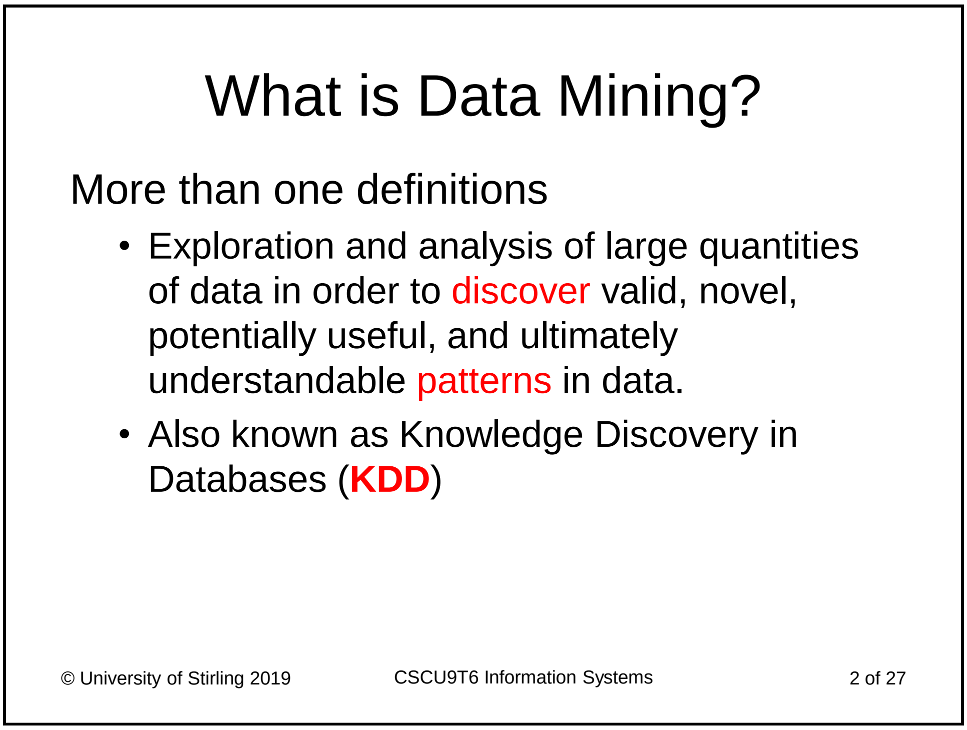


Data Mining

Jingpeng Li

http://www.cs.stir.ac.uk/~jli/



What is Data Mining?

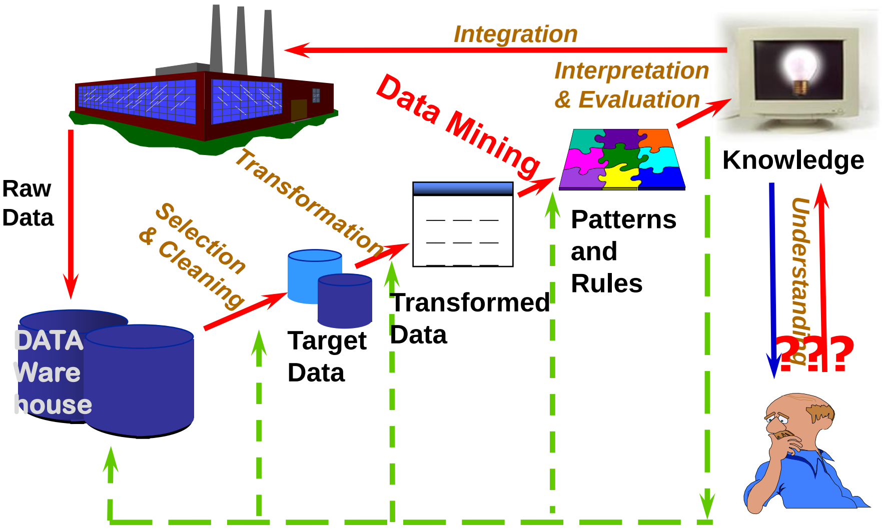
More than one definitions

- Exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.
- Also known as Knowledge Discovery in Databases (**KDD**)

What is Data Mining?

- Alternative to programming for certain types of task
- Functionality is determined by *data*, not programmed rules
- Data taken from events that we know about is used to help us guess about events that we don't know about
  - Predicting future events
  - Classifying current ones

Data Mining: the core of KDD



Some Data

Colour	Weight	Shape	Price	Fruit
Red	20	Round	3.92	Melon
Red	20	Oval	3.75	Melon
Red	10	Round	1.81	Apple
Red	25	Long	6.42	Melon
Red	25	Round	6.16	Melon
Yellow	25	Oval	5.83	Pineapple
Green	20	Oval	3.91	Melon
Green	25	Long	6.07	Melon
Yellow	25	Oval	5.87	Pineapple
Green	10	Long	1.28	Courgette

Some Questions

Colour	Weight	Shape	Price	Fruit
Red	20	Long	?	?
Yellow	25	Round	?	?
Green	12	Oval	?	?
Red	23	Long	?	?
Red	27	Round	?	?
Yellow	18	Oval	?	?
Green	20	Long	?	?
Green	5	Round	?	?
Yellow	5	Oval	?	?
Green	12	Long	?	?

The Task

- Build a system that can take the colour, weight and shape of an example fruit and tell us what price it should be and what kind of fruit it is most likely to be
- Based on no previous knowledge – just based on the data alone!
- Hence - no programming!!

Some Proper Examples

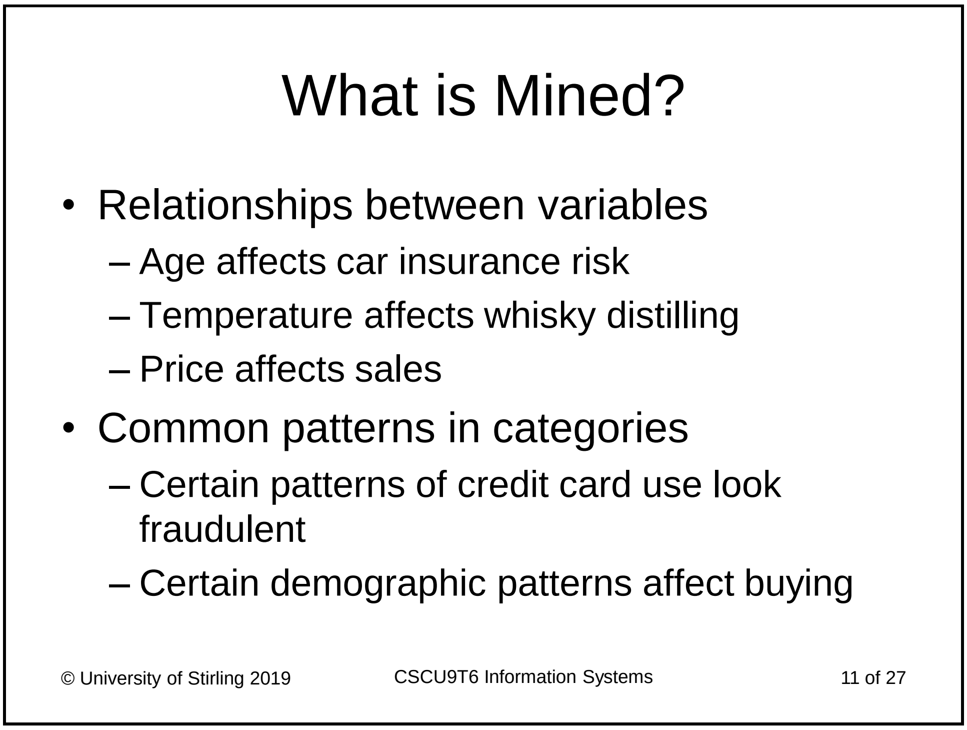
- Predicting what people will buy
- Spotting credit card or insurance fraud
- Targeting customers for advertising campaigns
- Predicting the price of stocks and shares or exchange rates
- Knowing when a cow is most fertile (no, really!)

Where Does Data Come From?

- Data reflects measurements of the real world
- It is usually a snapshot, or a sample, of the real world.
- Data mining assumes that whatever produced the data will continue to produce similar data in the future
- Data mining uses data from the past to predict behaviour in the future

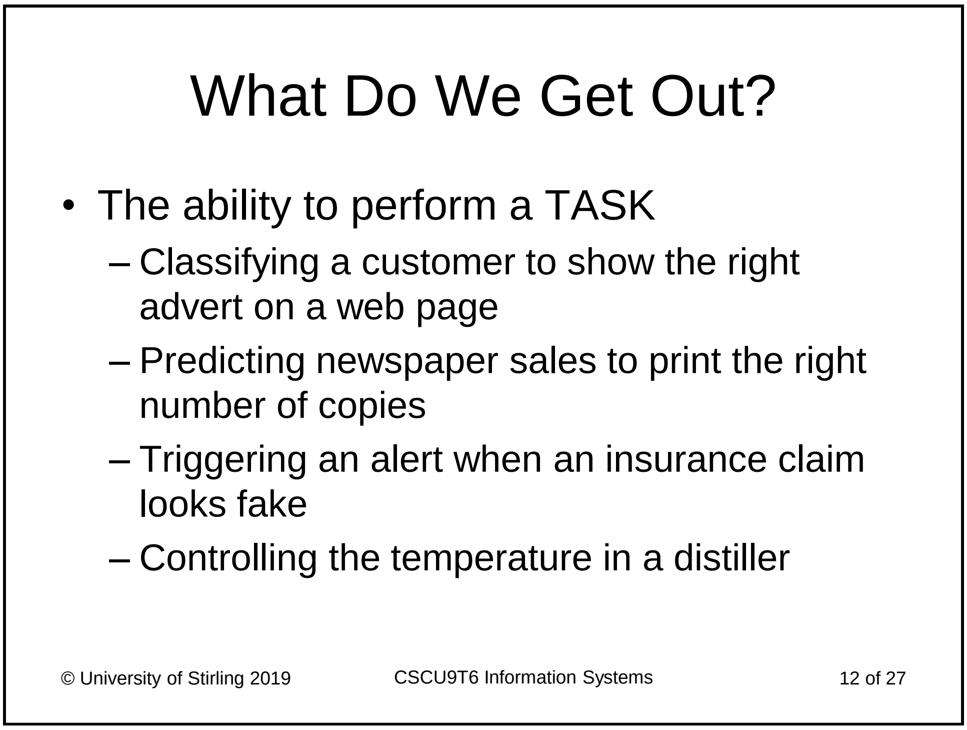
Problems With Data

- Errors – data might be measured wrong or entered wrong, or bits might be missing
- Quantity – Can you get enough data to represent the problem properly?
- Quality – Does the data measure the right things? Does it contain the information you need?
- Cost – Is it expensive (or even possible) to get the data you need?



What is Mined?

- Relationships between variables
  - Age affects car insurance risk
  - Temperature affects whisky distilling
  - Price affects sales
- Common patterns in categories
  - Certain patterns of credit card use look fraudulent
  - Certain demographic patterns affect buying



What Do We Get Out?

- The ability to perform a TASK
  - Classifying a customer to show the right advert on a web page
  - Predicting newspaper sales to print the right number of copies
  - Triggering an alert when an insurance claim looks fake
  - Controlling the temperature in a distiller

Techniques

- K – nearest neighbour
- Decision trees
- Neural networks
- K – means clustering
- Market basket analysis
- Logistic regression
- ARIMA

Advantages

- Capable of processing vast quantities of data that would be impossible to understand in any other way
- Allows computers to *learn* to perform tasks that we cannot program them to do
- Provides a method for embedding knowledge into processes

The Geography of Data

- Just as you wouldn’t dig a mine if you knew nothing about the geography of the earth,
- You shouldn’t do data mining without some understanding of the geography of data

Some Words

- Task – What it is you want the system to be able to do
  – Often defined in terms of inputs and outputs:
  – Take a description of an insurance claim as input. Output a decision on whether or not the claim looks fraudulent

Some Words - Variable

- Something that we can measure that varies!
  - E.g. Age, Height, Gender etc.
  - The variable is the building block of data mining
  - Once you have decided on a task, identifying the variables is the first thing to do
  - The inputs and outputs are defined in terms of variables
- Sometimes they are called ‘attributes’

Some Words - Value

- Variables can take one of a range (or set) of values
  - The variable ‘Gender’ can take values of ‘Male’ or ‘Female’
  - The variable ‘Age’ can take a number from 1 to 120

Some Words - Data

- Data *are* measurements of values associated with variables
- Data Point – A single example of whatever the data is measuring
  - For any given data point, each variable has one value, e.g. Gender=Male, Age=34, Marital Status=Single
  - A data point is often a single row in a file of data (the file is called the ‘Data Set’)

Some Words - Numeric

- Variables can be Numeric, that is taking numbers as values:
  - Continuous – E.g. Height. Any number in the range is acceptable
  - Discrete – E.g. Number of children – only certain values are acceptable: 3.2 children is not a valid value!!
- Values have a natural order: 3 > 2
- Ordinal values: values have an order
- Cardinal values: values count something

Some Words – Nominal

- Nominal variables (sometimes called Categorical or category labels)
- Used as qualities in inputs:
  - Gender
  - Colour
- Or labels in outputs:
  - Fruit name
- Values have no natural order:
  Apple > Pear makes no sense

Some Words – Data Model

- A Data Model is a representation of some aspect of the data, usually in a form that performs the chosen task
- The model will be built using a data mining technique, of which there are many
- Most models are represented as a set of parameters that are interpreted by software that implements a given technique

Learning

- Learning is the process of using data to build a model capable of performing a given task
- We talk of data mining models being trained
- They learn to generalise from specific data to general relationships

Inference

- Once you have a model, you can ask it questions
- This involves providing a new input pattern and the software producing an output pattern
- This output is a prediction or classification or measure of novelty …
- In any case, this is called *inference*

Topics Covered

- Data Preparation
- Classification
- Prediction
- Clustering & Association Rules
- Running a DM project
- Time Series Forecasting
- Visualisation

Objectives – What Will You Learn?

- What data mining is and what it is used for
- The types of task you can perform using data mining
- The uses for some data mining techniques (and a little of the technicalities of how they work)
- An understanding of what it means to build models and make predictions based on data

The Assignment

- There will be a assignment in this course, which will be given in the form of a consultancy task, imagining that you are data mining consultants
- You will use a software package called Weka, which we will also use in the practicals