

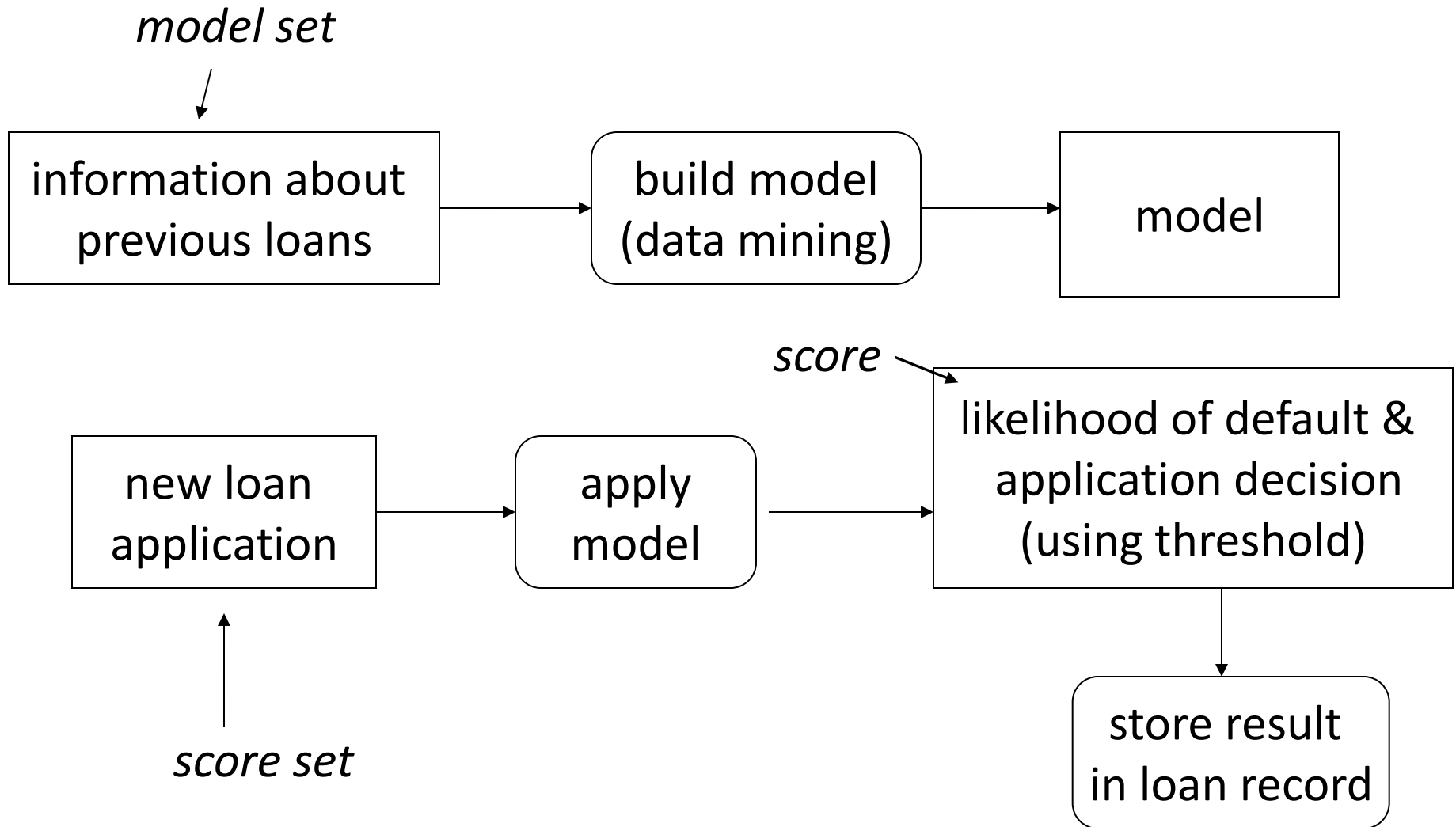
# Data Mining: An introduction

1. Some mining examples
2. What is data mining?
3. Data mining applications
4. The data mining cycle
5. Data sources

# 1. Mining examples

- Example 1: Suppose that a Bank wishes to assess loan applications (for risk)
- The Bank has information about previous loans (loan details, customer details and whether or not the loan defaulted)
- Using this information Data Mining tries to build a *(predictive) model* relating the loan & customer details to the target attribute *default*

# Example 1



# Mining examples

- Example 2: As a result of computerised checkouts & barcoding, supermarkets have *very* large volumes of data about shopping baskets
  - items purchased, time, date, day of the week, method of payment, loyalty card # (and hence shopper details)
- What can it do with this information?

# Mining examples

- Understanding how products are related to each other and to customer characteristics

product1 → product2

productA → highly\_profitable

and then use this information for merchandising, sales promotions, targeted marketing, customer profiling

# Mining examples

- Example 3: Customer profiling aims to reduce the mass of customer information to a simple classification
  - steady customer, unprofitable customer, growing customer, long term profitable, dormant, likely churning, ...
    - *Churn* – cancellation of service
  - ... and then apply appropriate business strategies (depending on classification)

# Mining examples

- Customer profiling – alternatively we might try to identify new customer segments
  - segments contain customers with similar behaviour/characteristics
  - investigate more deeply their particular buying patterns & hence develop specific business strategies aimed at that particular segment

## 2. What is data mining?

Data mining is the process of exploration and analysis, by automatic *and/or* semi-automatic means, of *large* quantities of data to discover *meaningful, actionable* patterns and rules (that were previously unknown or unexpected)



# Data Mining – the prerequisites

- Relevant data sets are being produced
  - normally in large volumes
  - captured automatically, and stored in databases and warehouses
- Computing power is available
- The value of *hidden* information is increasingly recognised

# Mining vs. reporting

- Given a database containing information on property sales and rentals we might ask:
  - Retrieve the total annual revenue generated by property sales for each type of property (Flat or House) in each city
- Although such queries can indeed get very complex – they are providing (relatively shallow?) reporting
  - You get what you ask for

# Directed forms of data mining

- In directed mining we have a *prior* view of what we are trying to do/find - defined by a *target* variable
- Classification
  - Building a model that enables us to take a new record, and assign it to a class
  - e.g., for bank customers a target might be long-term profitability (low, medium, high)
  - The set of classes is *pre-defined*
  - Model set consists of pre-classified records (target value is already known)

# Classification

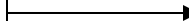
- So for long-term profitability of bank customers, the model set consists of pre-classified (present/historical) customer records consisting of
  - Customer data, and
  - A long-term profitability classification
- The resultant model will make a long-term profitability prediction for new customers ... with a certain degree of accuracy

# Classification

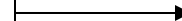
*model set*



pre-classified  
customer records



build model  
(data mining)

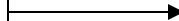


model

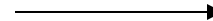
*score*



new customer  
record



apply  
model



long-term profitability  
(low, medium, high)



store result  
in customer record

*score set*



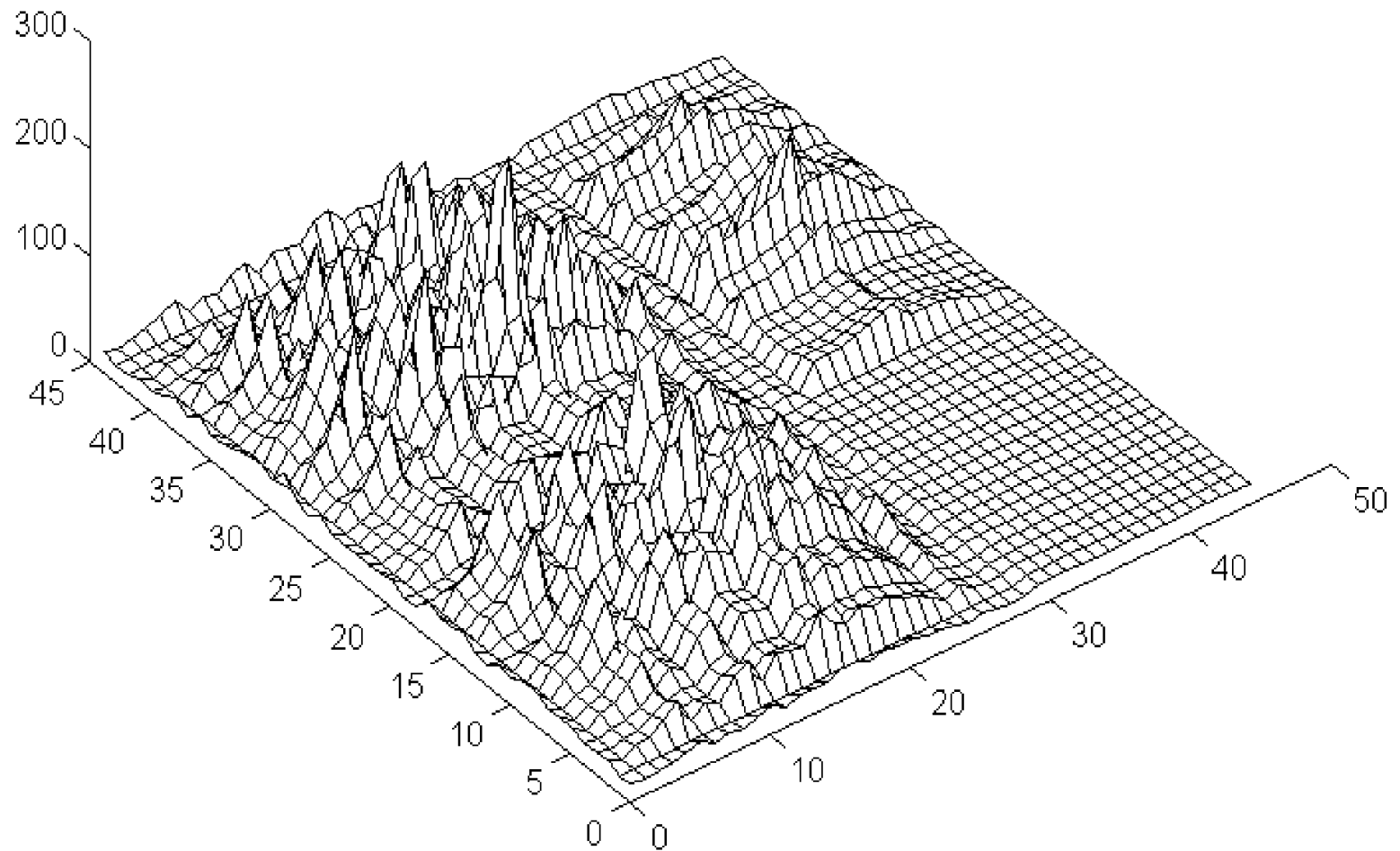
# Directed forms of mining

- Estimation
  - As in classification, but the target variable takes continuous (rather than discrete) values
  - e.g., likelihood of default
- Prediction
  - A form of classification/estimation in which the class/value will not be known until some future point in time
  - e.g., propensity to respond to a marketing campaign

# Undirected forms ...

- No prior view of what we are looking for; no target variable
- Association analysis/rules; Link analysis
  - Used to understand which things go together
- Clustering (or segmentation)
  - Clustering a group of individuals into subgroups that are (more) homogeneous: *No predefined classes*
- Description
  - to improve our understanding of the data
- Visualisation ... ditto

# Visualization





# Black vs. clear box techniques

- Sometimes we just want the answer
  - we don't care about the internal workings of the model - a black box technique is fine
  - e.g., credit card fraud
- Sometimes the inner details of the models give useful insights or explanations
  - clear box
  - e.g., loan approval/rejection

# 3. Data mining applications

- Retail/Marketing/CRM (Customer Relationship Mgt)
  - Analyzing sales
  - Customer acquisition, growth and retention
  - Identifying buying patterns of customers
    - merchandising, cross-selling, up-selling, designing product packages
  - Targeted marketing: predicting which customers will respond ...
  - Predicting when customers will want ...
  - Predicting the best next offer
  - Customer profiling/Personalisation

# Applications

- Banking
  - Detecting/understanding fraudulent credit card use
  - Predicting customers likely to change their credit card (churn)
  - Credit scoring
  - Identifying loyal customers, potentially profitable customers, ... Customer profiling
  - Marketing and sales; targeted marketing ...


# Applications

- The web
  - Providing recommendations
  - Click-stream analysis
  - Customer profiling etc etc
- Insurance
  - Claims analysis (fraud)
  - Predicting which customers will buy new policies
- Medicine
  - Identifying potentially successful drugs
  - Analyzing medical images

# Applications

- Predicting churn for mobile phone customers
- Library stock management
- Land use analysis (satellite image analysis)
- Astronomy (image analysis)
- Operational processing (e.g., stock control)
- Prediction (e.g., TV audiences, financial stock prices)
- Social network link analysis
- Terrorism detection
- Crime prevention via hotspot analysis
- ... and many, many, many more

## 4. The data mining cycle

- Identify suitable business problem(s)
    - where data analysis might provide business value
  - Transform data into actionable results
    - using data mining
  - Act upon these results
  - Measure the impact of the actions
- 

# Some manual or semi-manual activities

- Choosing a suitable business problem
- Identifying, collecting & amalgamating relevant data
- Identifying relevant data cleaning
- Transforming data into an appropriate form for mining
- Preliminary data analysis
- Identifying appropriate mining techniques
- Modifying data mining tool parameters
- Evaluating data mining models
- Acting upon the mining results
- Measuring the impact of the action

# Identifying the right business problem

e.g.,

- Planning a new product
- Attracting – or retaining - customers
- Understanding patterns in sales figures
- Planning a marketing campaign
- ...

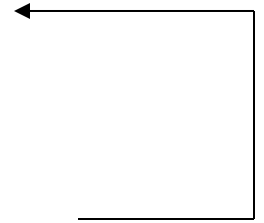


# ... the right business problem

- What is important to the business
  - products, customers, sales, geography, time, ...
- Which segments are of interest
- The relevant business rules
  - Data mining has a habit of finding known patterns
  - Business rules can direct the mining effort
- Is the required data available?
- Is a mining effort necessary?

# Transforming data into actionable results

- Obtain, validate & clean data
- Preliminary analysis
- Choose modelling technique
- Prepare the model set
- Build model & evaluate performance
- Pick “best” model(s) and apply to score set



# Acting upon the results

The results might suggest various “actions”

- Insight
- One-time results
- Remembered results
- Periodic predictions
- Real-time scoring
- Fixing data and business processes

# Measuring the impact of the action

- Often overlooked because of its long term value
  - -but feeds into the next cycle by highlighting what works and what doesn't
- E.g., If a mailshot is followed by a customer purchase – how do we know that the customer would not have made the purchase anyway?

## 5. Data sources

- Data for (traditional) data mining should be in a single table (or view)
- Each row should correspond to a single instance - a “unit of action”
  - The “unit of action” might be a loan, customer, insurance policy, insurance claim, basket, household, PC, inventory item
  - So we need to understand the interests/intentions of the business

# Data requirements

- Sufficient data to solve the problem
  - Sufficient attributes and sufficient records
- Representative of the target audience
  - time, geography, demographics, market
- Input data is needed for the score set too
- Quality
  - Accuracy; Consistency (integrity); Aggregated? Excessive summarisation; Confusing/inconsistent attribute names and data formats; Missing (null) values
- Legal issues

# Data sources

- Survey and product registration data
  - responders may not be a representative sample - they are self selecting
  - inaccurate and outdated data
- Service registration data, e.g., loyalty card
- External data sources ...
  - Deciding which may be challenging
  - Matching new, externally collected data to pre-existing customer DBs is challenging
    - Enrichment; aggregation

# Online transaction processing systems

## OLTP

- the obvious original source
  - assuming the OLTP captures the data
- data format may be a problem
  - OLTP are built for efficient operational processing
- data cleanliness might be a problem
  - particularly for non-essential data
- may embed business rules - that need to be understood



# Relational systems

- Consider a customer's transaction data
- Each transaction has a number of line items – each line item relates to a particular product
  - Normalisation would distribute this data across multiple tables
- For each week/month/quarter we might store
  - Number of transactions
  - Total value of transactions
  - Proportion of transactions by location
  - Proportion of transactions by day ... etc. etc.

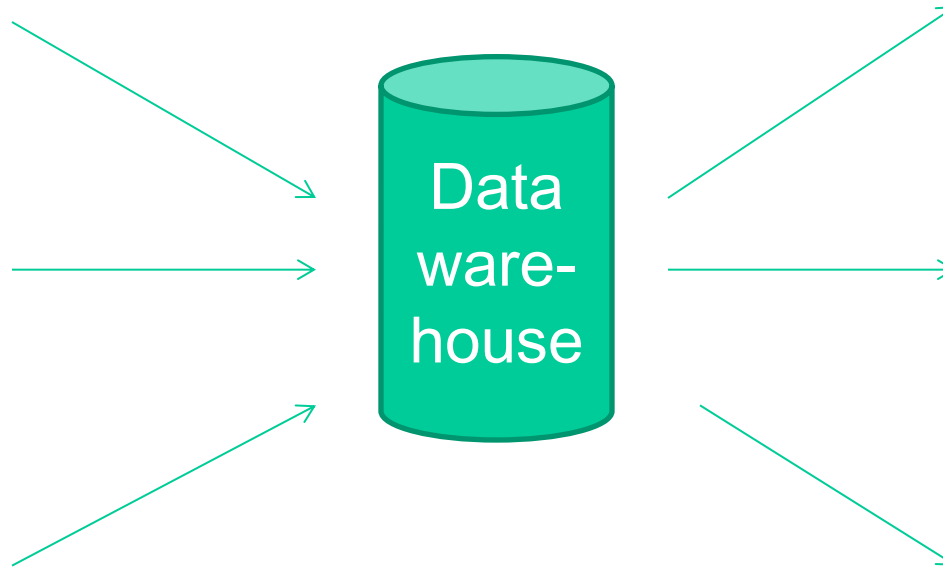
Derived  
variables

There is  
nothing  
definitive  
about this –  
need to think  
about the  
intended  
purpose

# Data warehouses



....



Reporting



OLAP



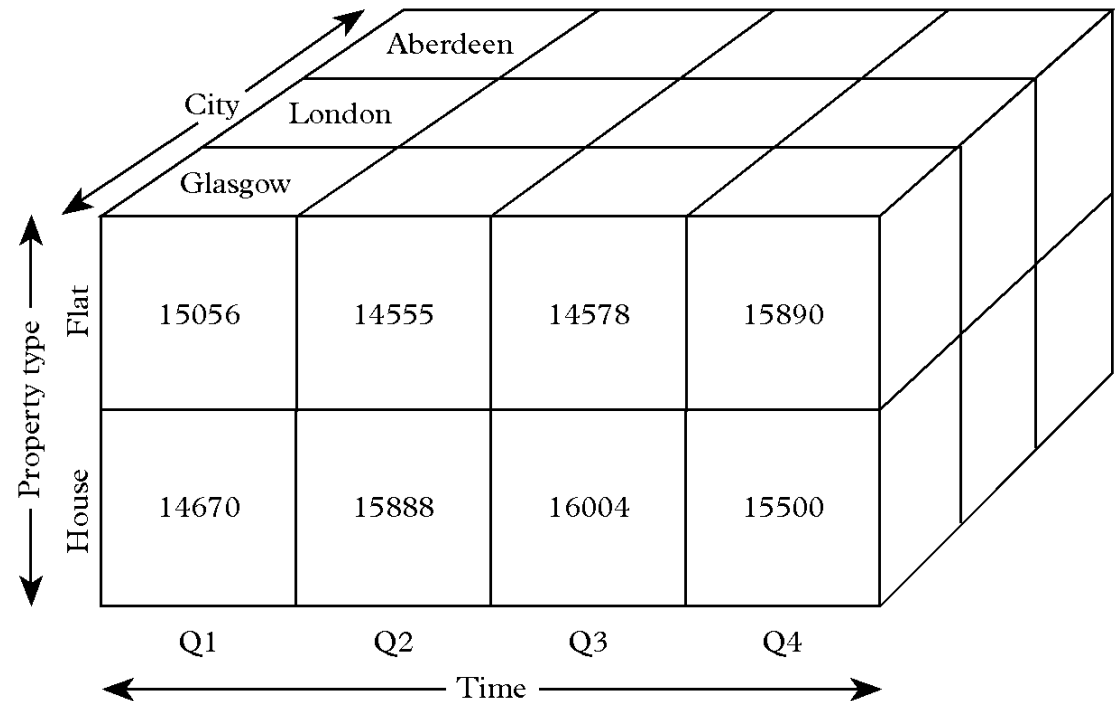
Data Mining

# Data warehouses

- A subject-oriented, integrated, time-stamped and non-volatile collection of data to support management decision-making
  - usually clean
  - may not contain the right data
    - too broad ... data marts
    - summarised

# OLAP

Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....	.....	.....	.....
.....	.....	.....	.....



**Total annual revenue generated by property sales for each type of property (Flat or House) in each city**

# OLAP

- Complex query and “what-if” analysis of multi-dimensional data (time, city, property-type)
  - roll-up and drill-down
- OLAP results can be used as reference data in mining
  - E.g., average sales by region
- OLAP good for data exploration/preliminary analysis
- Conversely: Mining results are candidates for insertion in OLAP systems
- OLAP may not contain the right data content
  - OLAP reports may not need the customer field and hence this field may be omitted
  - OLAP systems might contain summarised data