

CMM510 Data Mining

Pamela Johnston (SoCET)

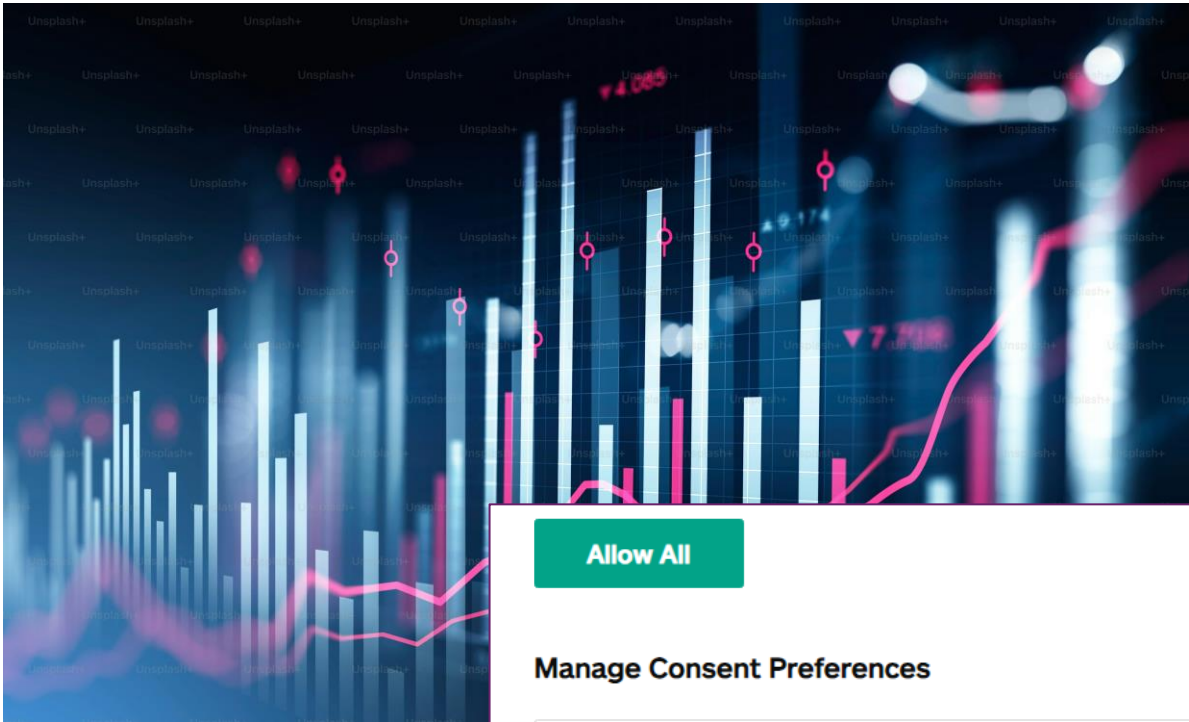
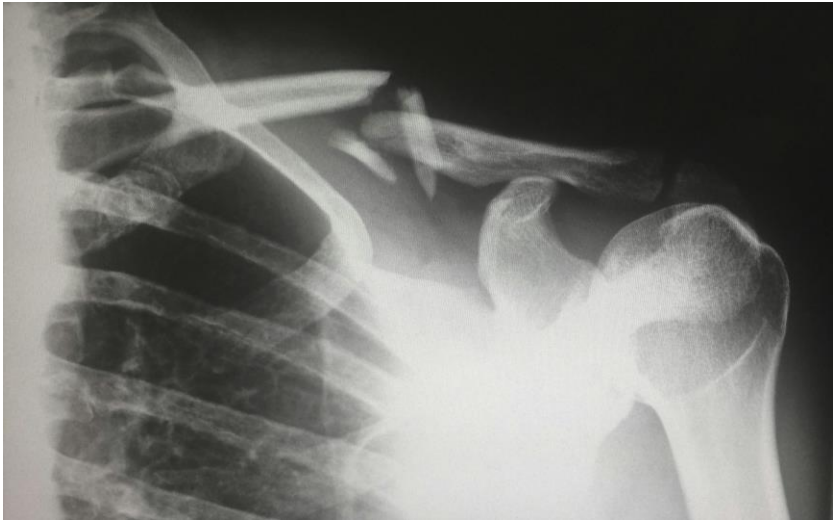
p.johnston2@rgu.ac.uk

Lecture 1: Introduction to Data Mining

Content

- Data mining
- Methodology
- Input and output
- Applications
- Ethical and professional issues
- Summary

Lots of data



Time	User full name	Affected user	Event context	Component	Event name	Description
4 September 2024, 3:05:30 PM	Pam Johnston	-	Study Area: [Module Study Area 2024/2025] CMM510 - data mining - Semester 1	System	Course viewed	The user with

Allow All

Manage Consent Preferences

+ Strictly Necessary Cookies

Always Active

Confirm My Choices

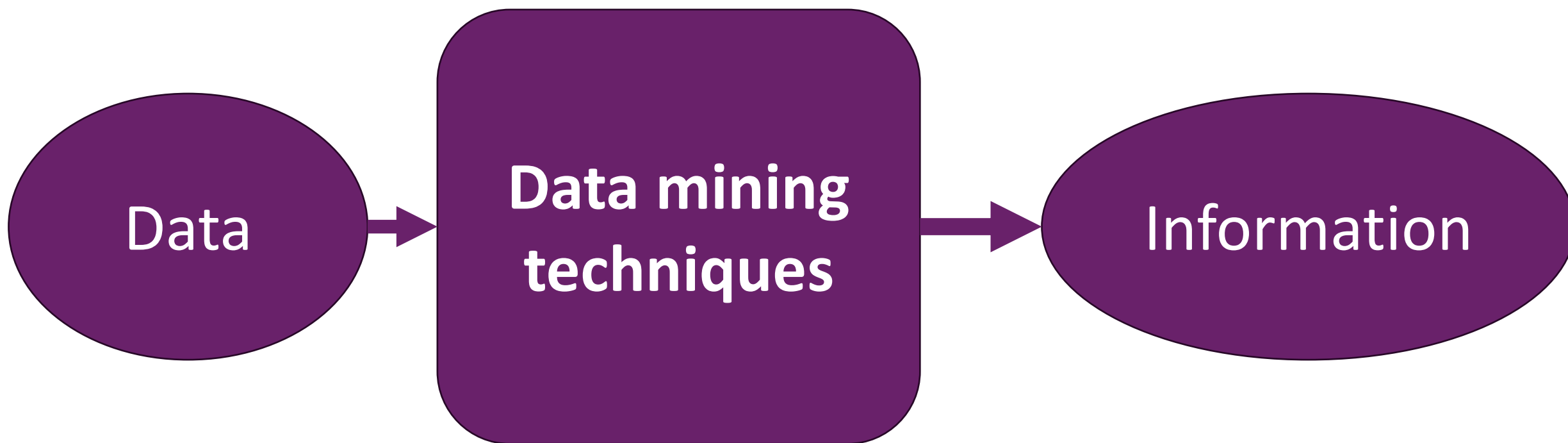
Powered by

onetrust

Data Banks

- Nowadays we collect vast amounts of data, e.g.
 - Shopping transactions
 - Bank transactions
 - Medical records
 - Web logs
 - Drilling information (bottom hole pressure, mud flow, porosity, permeability ...)
 - Pandemic data (positive cases, hospitalisations, deaths, countries, population ...)
 - Weather data
 - Smart meter data
- Raw data is not very useful
- Huge volume of data makes it difficult to handle.

“Data” != “Information”



Getting Information from Data

- Information is required in order to solve problems.
- Data can be a superb source of information.
- This may be difficult to extract due to the volume of data.
- BUT once extracted, we can get an understanding of the problem domain.
- E.g. Discovering fraudulent credit card use from transactions.
 - **Problem definition:** various data regarding the current transaction.
 - **Problem solution:** whether the current transaction is fraudulent or not.
 - **Information:** extracted from records of past transactions including whether they were fraudulent or not. How to determine fraudulent transactions.

Data mining

- **Data mining** is the process of extracting information which is implicitly stored in collections of data.
- Used to:
 - Solve new problems (e.g. detect credit card fraud)
 - Understand problems and their solutions (e.g. understand what situations may lead to fraud).
- Main challenges:
 - Work with large volumes of data
 - Distinguish between interesting and uninteresting information
 - Work with inaccurate and incomplete sets of data.

...

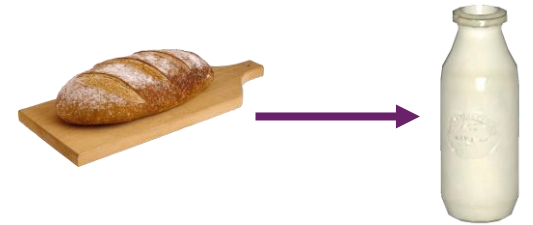
- Aim: find strong patterns in data
 - Pattern strength is related to prediction strength
- BUT
- Most patterns contained in data are not interesting
- Patterns may be
 - Not always true (inexact)
 - The result of chance (spurious)
- Missing data
- Inaccurate or erroneous data

Example

- Shopping

- Strong pattern – people who buy bread also buy milk

- But this is not interesting!



- Weaker pattern – men who buy nappies on a Friday also buy beer

- More interesting

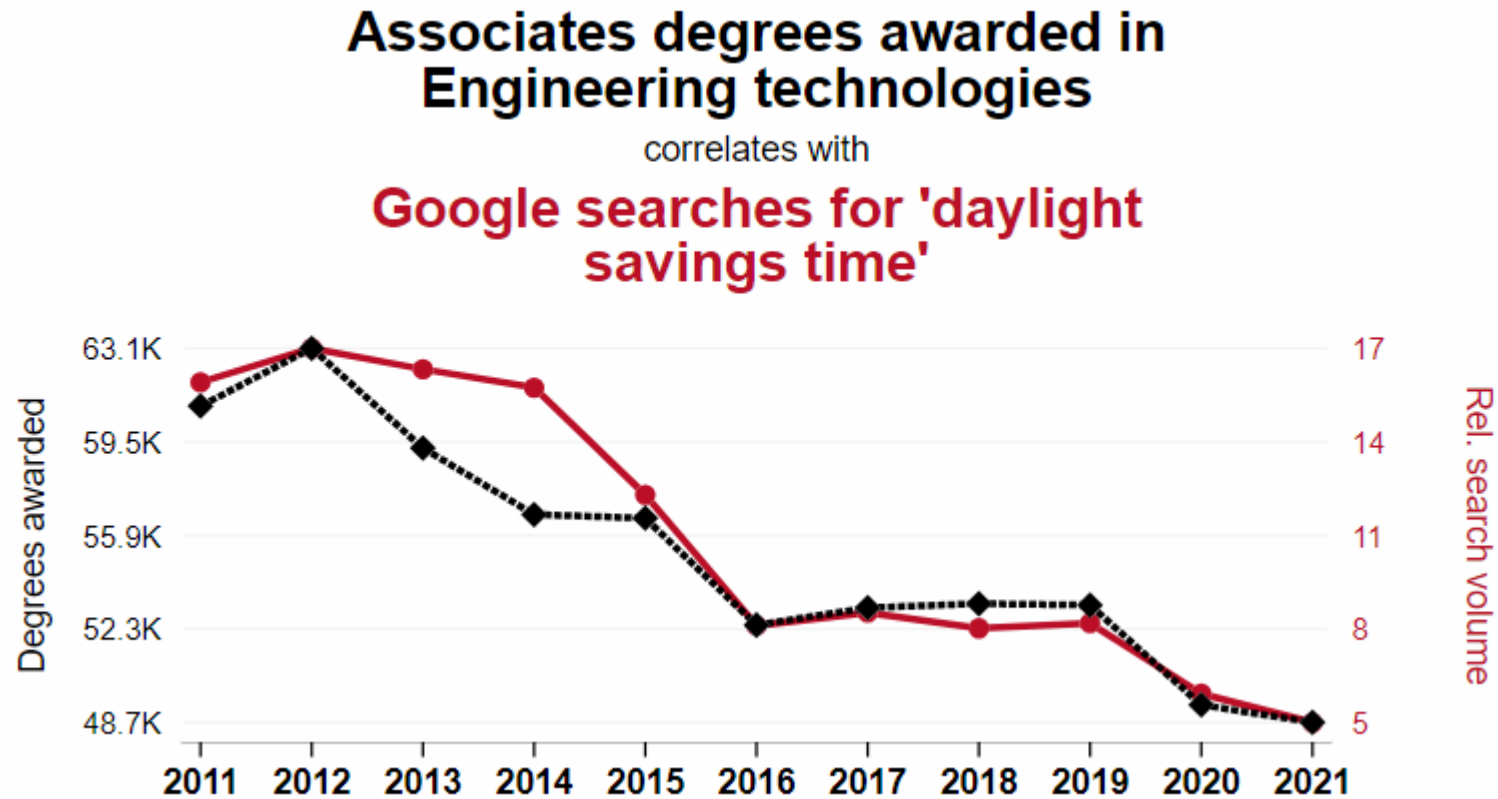
- Weaker – some men buy only nappies ...



- Missing data – the gender of the shopper is unknown for some transactions.

- Inaccurate data – the gender of the shopper might have been entered incorrectly.

Spurious Correlations (Tyler Vigen)



Data Mining Requirements

Data

A (large) set of
past data
(labelled)

Machine learning

One or more
programs which
extract
relationships
(patterns)
between data, i.e.
information.

Evaluation

Correct?
Useful?

Machine Learning

- Used in data mining to obtain relationships (patterns) between data
- **Learning**
 - Capable of changing behaviour in order to perform better
- Learning from examples
 - **Training data:** examples used for learning
 - **[Validation data:** examples used for tuning parameters]
 - **Test data:** examples used to test learnt knowledge.

Supervised Data Mining

Classification

- this? or that?
- One of these
- Which category?
- Fraud or not?
- Spam or Ham?
- Faulty or not?
- Cat, dog or fish?

Regression

- How much tomorrow?
- How much, given this?
- House price estimate
- Share price prediction

Others





- What happens next?
- Best course of action?
- Online learning
- Reinforcement learning
- Metric learners

Types of Data Mining

- UNSUPERVISED (knowledge discovery)
 - **Association Rules:** find patterns in data
 - Purchasing habits in supermarkets
 - **Clustering:** groups data into clusters of similar cases
 - **Others:** e.g.
 - Summarisation: find compact definitions of data
 - Deviation Detection: detects changes from norm.

Evaluation

- How effective and efficient is the data mining model / output at classifying unseen data?
- The test data is used as 'unseen data'

Confusion matrix	Spam (Predicted)	Non-Spam (Predicted)	Accuracy
Spam (Actual)	27 	6 	81.81
Non-Spam (Actual)	10 	57 	85.07
Overall Accuracy			84

$$= 27 / (27 + 6)$$

$$= 57 / (57 + 10)$$

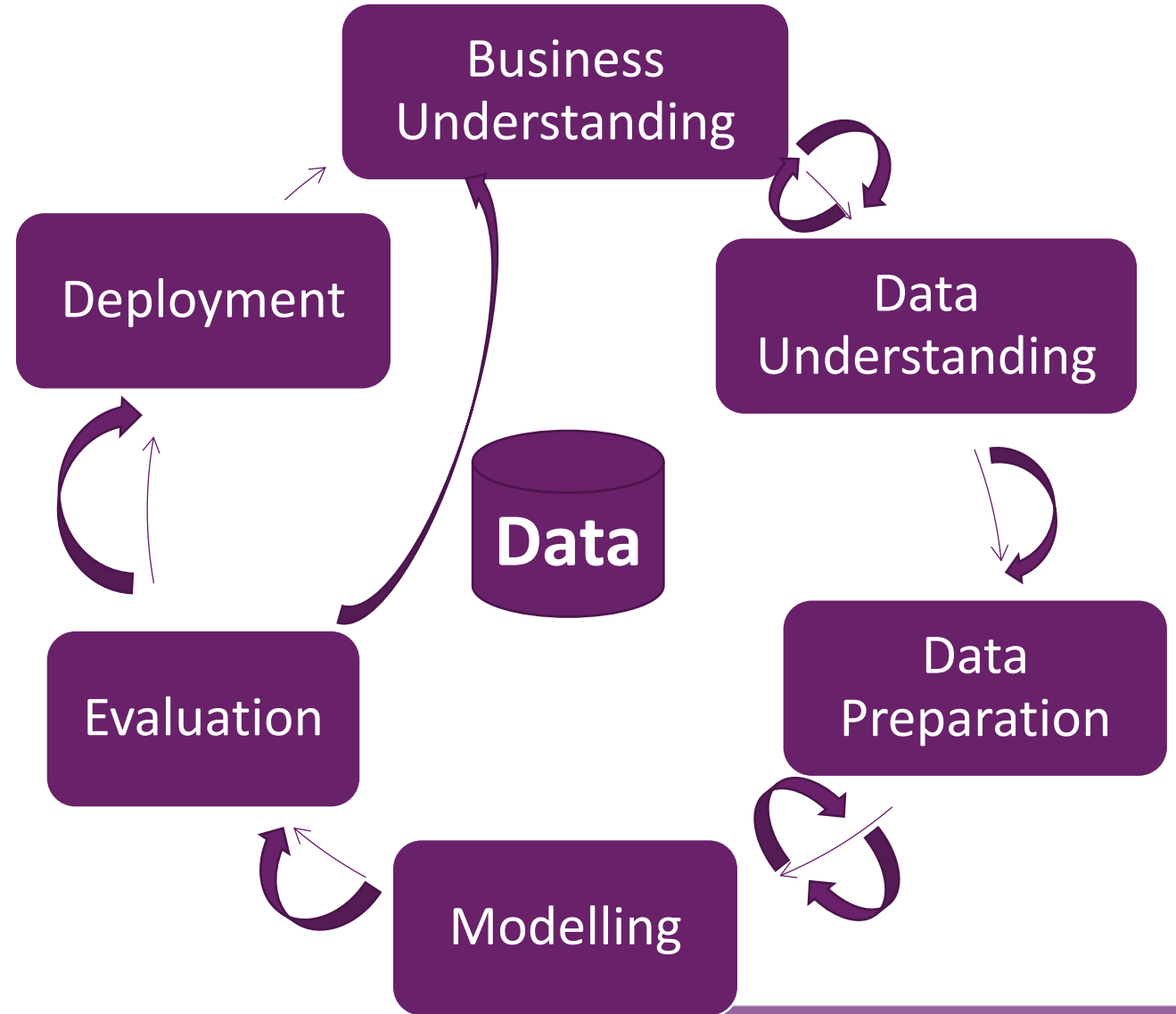
$$= (27 + 57) / (27 + 6 + 57 + 10)$$

Contents

- Data mining
- **Methodology**
- Input and output
- Applications
- Ethical and professional issues
- Summary

Methodology

- A popular methodology is the Cross Industry Standard Process for Data Mining (CRISP-DM)
- Agile methodology with cycle where
 - There is no strict sequence between stages
 - Movement between states is forward as well as backwards



Contents

- Data mining
- **Input and output**
- Applications
- Ethical and professional issues
- Summary

Input and output

Input:

Data about previous weather conditions and whether tennis was played

Outlook	Temp	Humidity	Windy	Play?
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Cloudy	Hot	High	No	Yes
Rainy	Mild	Normal	No	Yes

Output:

Result of learning applied to input data.

If outlook = sunny **and** humidity = high
 then play = no

If outlook = rainy **and** wind = yes
 then play = no

If outlook = cloudy
 then play = yes

If humidity = normal
 then play = yes

If none of the above rules applies
 then play = yes

Input: instances (or past data examples)

- Instance: a single example of a concept
 - described by a set of attributes (features, columns)
- Input to learning algorithm
 - Set of instances (past observations / problems). E.g. examples of past weather conditions and whether tennis was played
 - Usually described as a single relation/flat file.
 - But could be text, images, etc (more difficult!).

<i>Outlook</i>	<i>Temp</i>	<i>Humidity</i>	<i>Windy</i>	<i>Play?</i>
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Cloudy	Hot	High	No	Yes
Rainy	Mild	Normal	No	Yes

Input (jargon alert!)

attributes or features

label

instances or samples

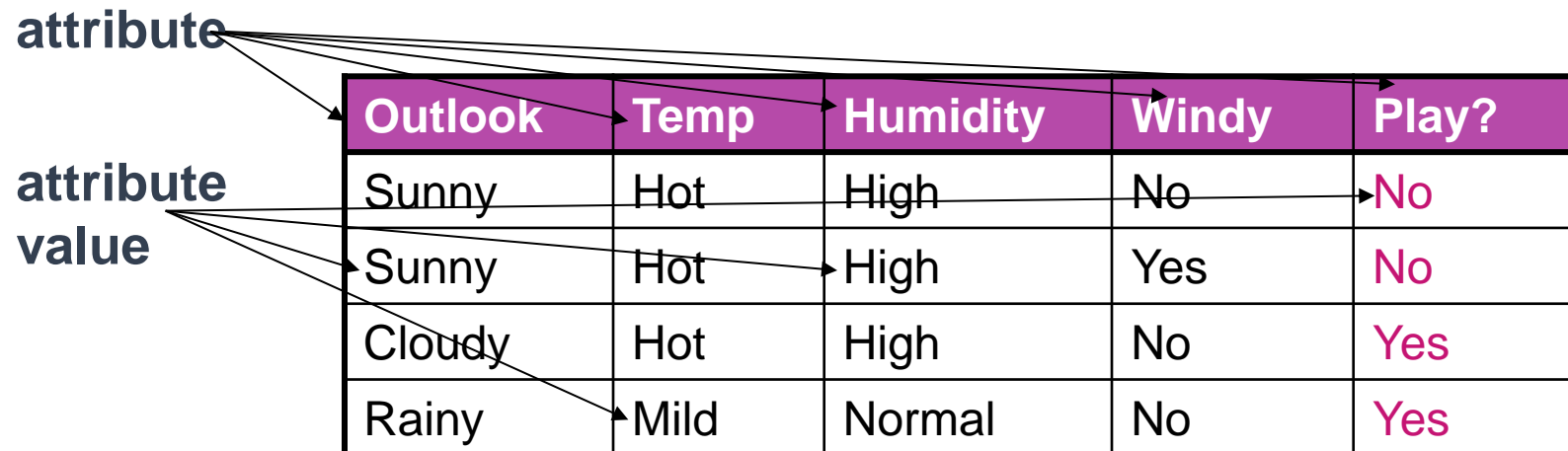
<i>Outlook</i>	<i>Temp</i>	<i>Humidity</i>	<i>Windy</i>	<i>Play?</i>
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Cloudy	Hot	High	No	Yes
Rainy	Mild	Normal	No	Yes

Attributes (or features)

- **Attribute (feature):** describes a specific characteristic of an instance
 - e.g. age, salary, ...
- Attributes are often predefined for a set of instances
 - an instance is described by its attribute values
 - e.g. 25, 20567, ...

attribute

attribute
value



Outlook	Temp	Humidity	Windy	Play?
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Cloudy	Hot	High	No	Yes
Rainy	Mild	Normal	No	Yes

Types of Attribute

- Nominal
 - Values are symbolic, e.g. desk, table, bed, wardrobe.
 - No (obvious) relation between nominal values.
 - Boolean attributes are a special case.
 - 0 and 1 or True and False
 - Also called categorical, enumerated or discrete.
- Ordinal
 - Values are ordered, e.g. small, medium, large, x-large
 - but difference between 2 values is not meaningful

Types of Attribute

- Interval
 - Quantities are ordered.
 - Measured in fixed equal units, years 2001, 2002, 2003, 2004.
 - Difference between values meaningful: $2005 - 2004$
 - Sum or product is not meaningful: $2005 + 2004$
- Ratio
 - Quantities include a natural zero.
 - Money: 0, 10, 100, 1000
 - Treated as real numbers because all mathematical operations are meaningful.

Qualitative or Quantitative?

- Weight
- IQ
- Distance
- Height
- Miles per gallon
- Calories per cake
- DOI (Digital Object index)?
- Phone number
- Colour
- Gender
- Name

Preparing the Input

- Need to obtain a dataset in 'correct format'.
- Missing data needs to be dealt with.
- Inaccurate data must be pre-processed.
- You will see the management of the above in module CMM535.

Output: examples

Outlook	Temp	Humidity	Wind	Play?
Sunny	Hot	High	No	No
Sunny	Hot	High	Yes	No
Cloudy	Hot	High	No	yes
Rainy	Mild	Normal	No	yes

Output info:

- May be
 - Complete, i.e. covers all possibilities
 - Incomplete
- Accuracy may be
 - 100%, i.e. works all the time
 - < 100%

Assuming 3 possible values for outlook, 3 for temperature, 2 for humidity and 2 for wind there are

$$3 * 3 * 2 * 2 = 36 \text{ possible combinations (or rows)}$$

Output: Decision list – 1st rule that fits is selected

- **If** *outlook* == sunny
 and *humidity* == high
 then *play* = no
- **If** *outlook* == rainy
 and *wind* == yes
 then *play* = no
- **If** *outlook* == cloudy
 then *play* = yes
- **If** *humidity* == normal
 then *play* = yes
- **If** none of the above rules applies
 then *play* = yes

Output: Decision list with numeric values

Outlook	Temp	Humidity	Wind	Play?
Sunny	85	85	No	No
Sunny	80	90	Yes	No
Cloudy	83	86	No	yes
Rainy	70	96	No	yes

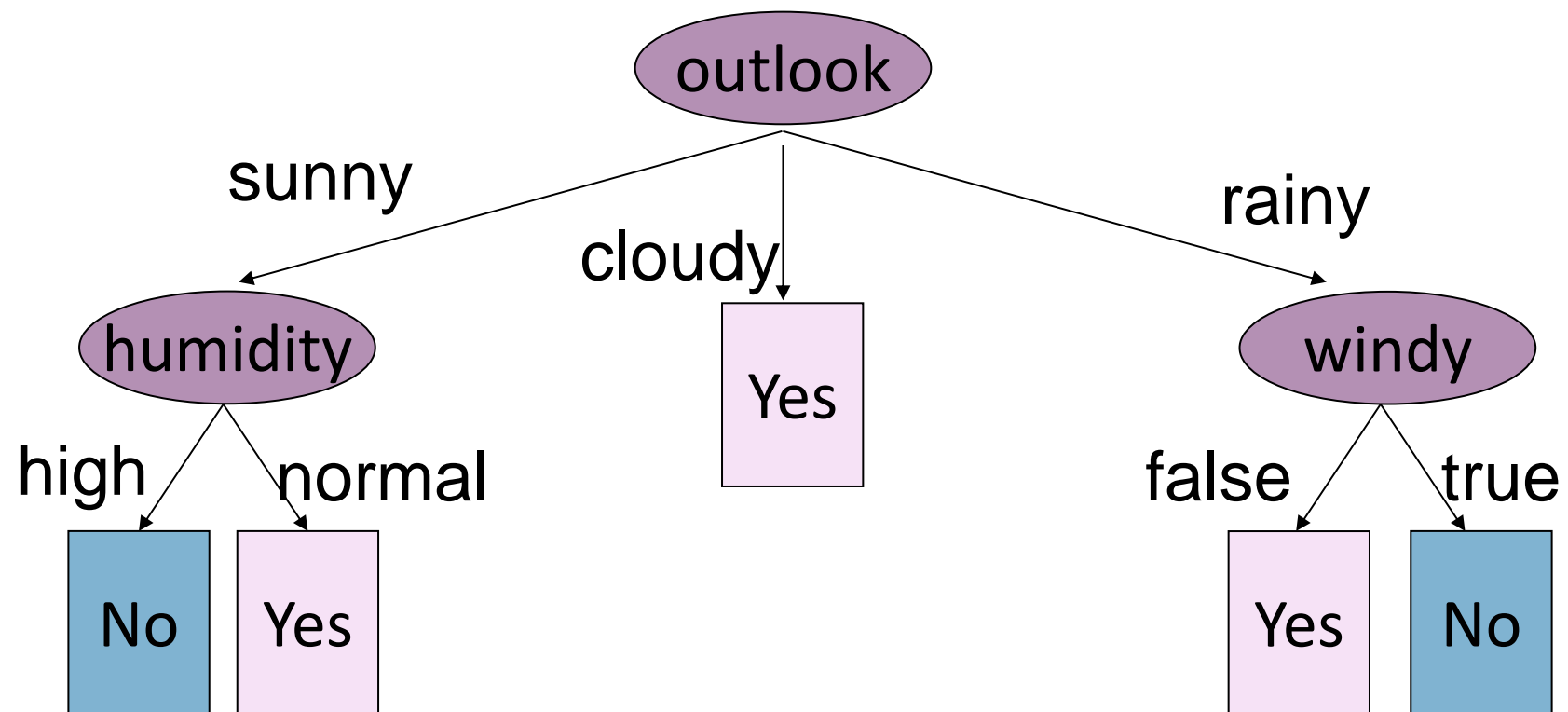
Requires inequalities to deal with numeric values. E.g.

if *outlook* = sunny

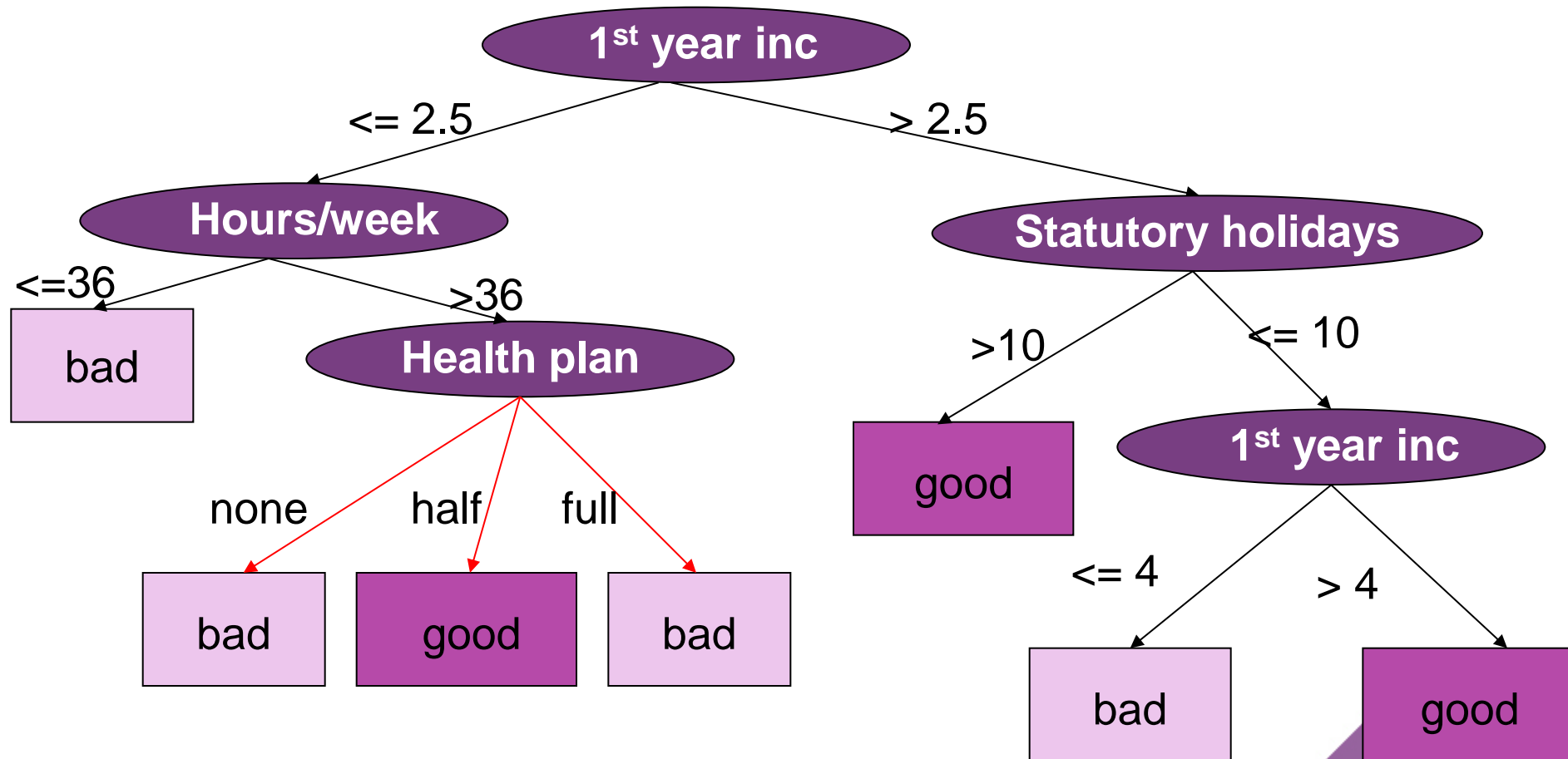
and *humidity* > 83

then *play* = no

Output: Decision tree

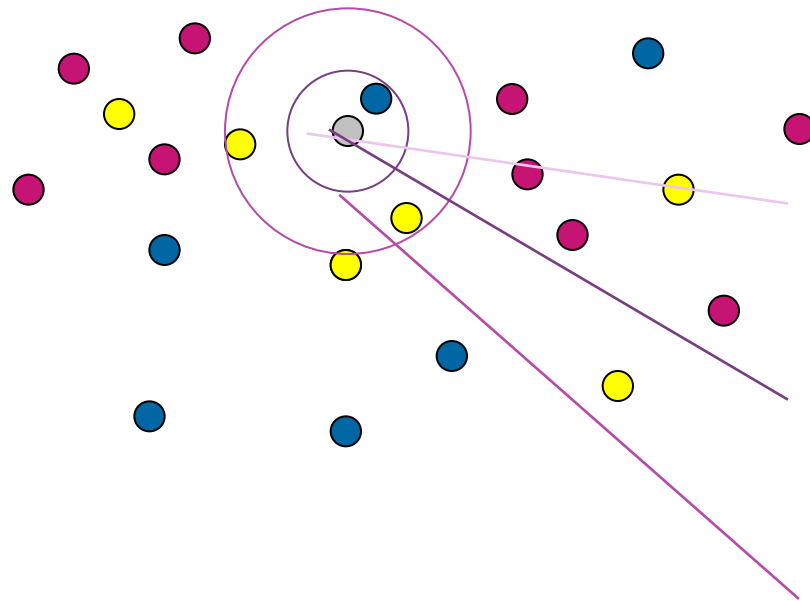


Decision Tree Example with numeric attributes



Output: instance-based

To solve problem reuse solution to most similar problem(s)



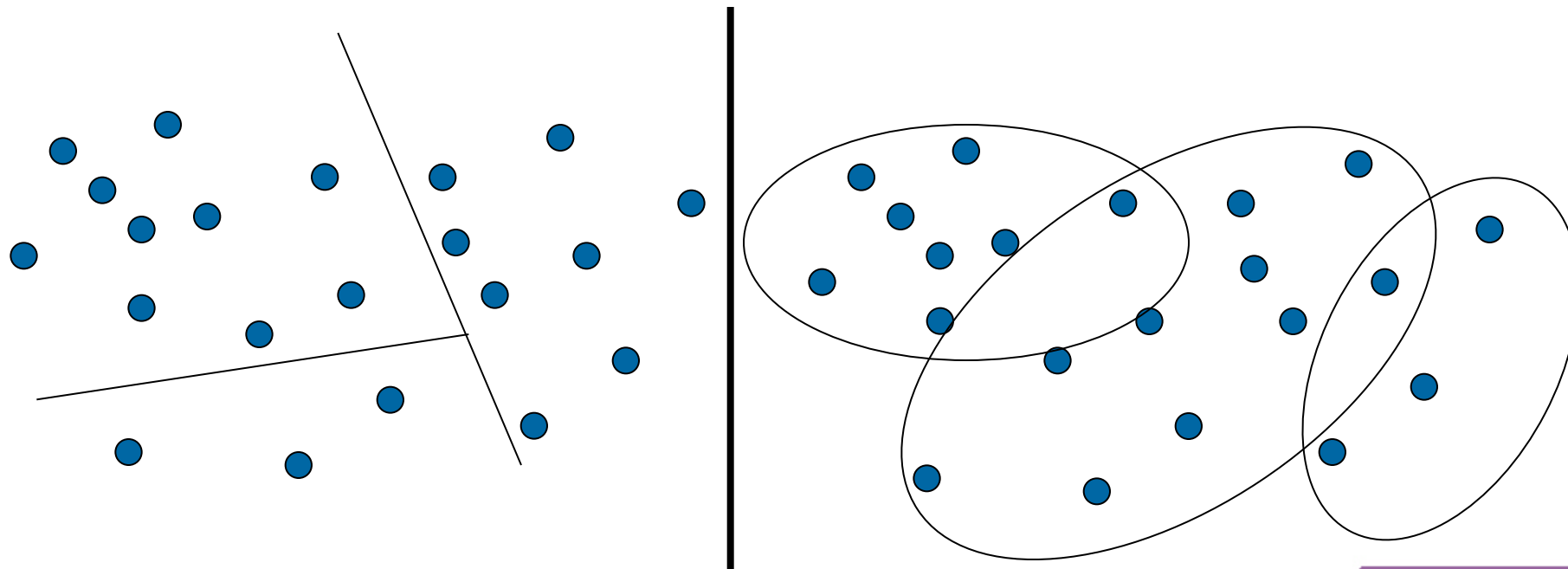
New problem (grey). Class?

Closest solution to new problem is
“blue”

3-neighbour solution to new
problem is “yellow”

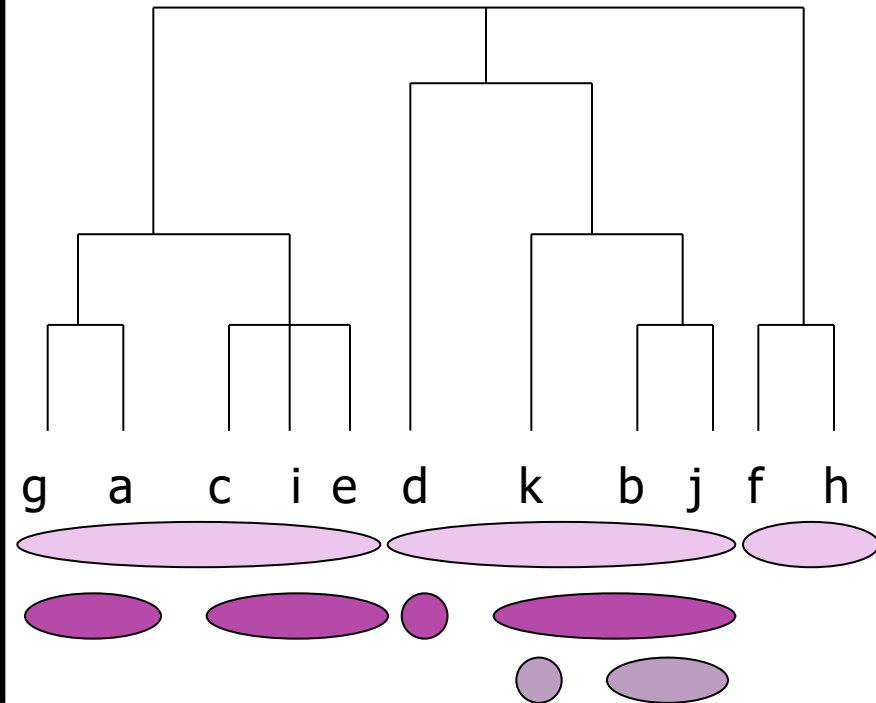
Output: Clusters

- Represent groups of instances which are similar
- Some allow overlapping clusters



Output: hierarchical clusters

instances	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
...			



Association rules

Unsupervised task => associate co-occurrences of values, e.g. in market basket data

customer	beer	nappies	bread	...
1	yes	no	yes	
2	yes	yes	no	
3	no	yes	yes	
4	no	no	no	

- Other application examples:
 - Amazon buying habits
 - Word usage in email or text communication

Output: Association Rules

- Association rule
 - **If beer == yes and crisps == no then nappy = yes**
 - **If beer == yes then nappy = yes and bread = no**

Different from

- **If outlook == sunny and windy == no then play = yes**
 - Predicted attribute changes [not always play]
- Like classification rules BUT
 - **used to infer the value of any attribute (not just class)**
 - **or a combination of attributes**

Contents

- Data mining
- Input and output
- **Applications**
- Ethical and professional issues
- Summary

Applications

- Automatic estimation of organisms in zooplankton samples
- Maintenance schedules of heavy machinery.
- Autoclave layout for aircraft parts
- Automated completion of repetitive forms
- Loan decision-making
- Image screening
- ...etc

Should an applicant get a loan?

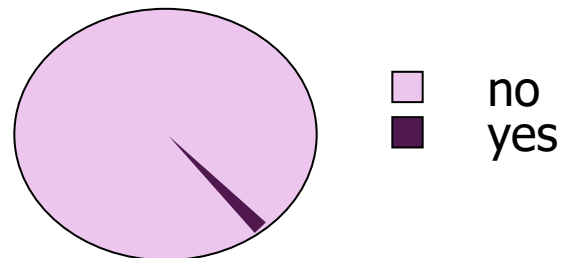
- Statistical model deals with 90% cases
- 10% cases referred to loan officers
- 50% referred cases are bad
- BUT referred customers generate money!!!
- Expert gets 50% of referred cases right
- Solution: use data mining to aid decision of borderline cases

Should an applicant get a loan?

- 1000 training examples
- 20 attributes
- Extracted rules accurately predict 70% referred cases
 - Much better (?) than human expert!
- Rules could be used to explain to customers the reasons for the company's decision.

Detecting Oil Spills from Images

- Data: radar satellite images
- Oil spills: dark regions with changing size and shape
- BUT weather conditions can also cause this effect!!!
- So spill detection is a specialised job.
- Problems:
 - very few training examples
 - data is not balanced (most dark areas are NOT spills)



Detecting Oil Spills from Images

- Normalised image used for extraction of dark regions
- 7 attributes used: *size, shape, area, intensity, sharpness and jaggedness of boundaries, proximity to other regions, info about background in vicinity of region.*
- Batch: regions from a specific image
- Adjustable false alarm rate required

Contents

- Data mining
- Methodology
- Examples: input and output
- Applications
- Ethical and professional issues
- Summary

Ethical and professional issues

- GDPR

- <https://gdpr-info.eu/> [accessed 04/09/2024]
- <https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/> [accessed 04/09/2024]

- The UK Government Data Ethics framework.

- <https://www.gov.uk/government/publications/data-ethics-framework> [accessed 04/09/2024]

- The BCS Code of Conduct.

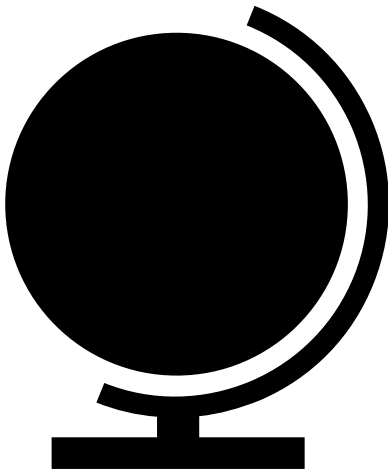
- <https://www.bcs.org/membership/become-a-member/bcs-code-of-conduct/> [accessed 04/09/2024]

- The UK Statistics Authority Code of Practice.

- <https://code.statisticsauthority.gov.uk/the-code/> [accessed 04/09/2024]

Data Protection

- GDPR describes how (personal) data should be used by organisations, businesses, the government and the general public. It includes
 - Data processing
 - Data movement



Ethical Issues

- How are ethical issues dealt with?
 - E.g. use applicant's sex, religion or race in order to decide whether to give a loan - unethical
 - BUT these same attributes are OK when used in medical application
- The use of data for certain applications may pose problems
 - E.g. postcode may be a strong indicator of an individual's race.
- Data collected for a particular reason should not be used (using data mining) for a completely different purpose without appropriate consent.
- Information mined may be surprising: red car owners are more likely to have problems paying their car loans in France.

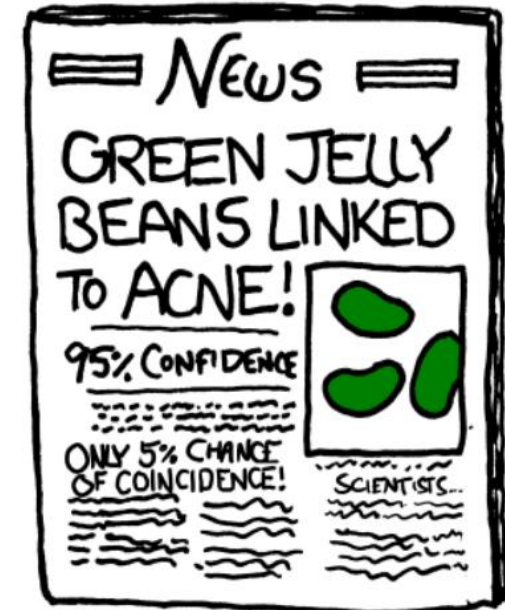
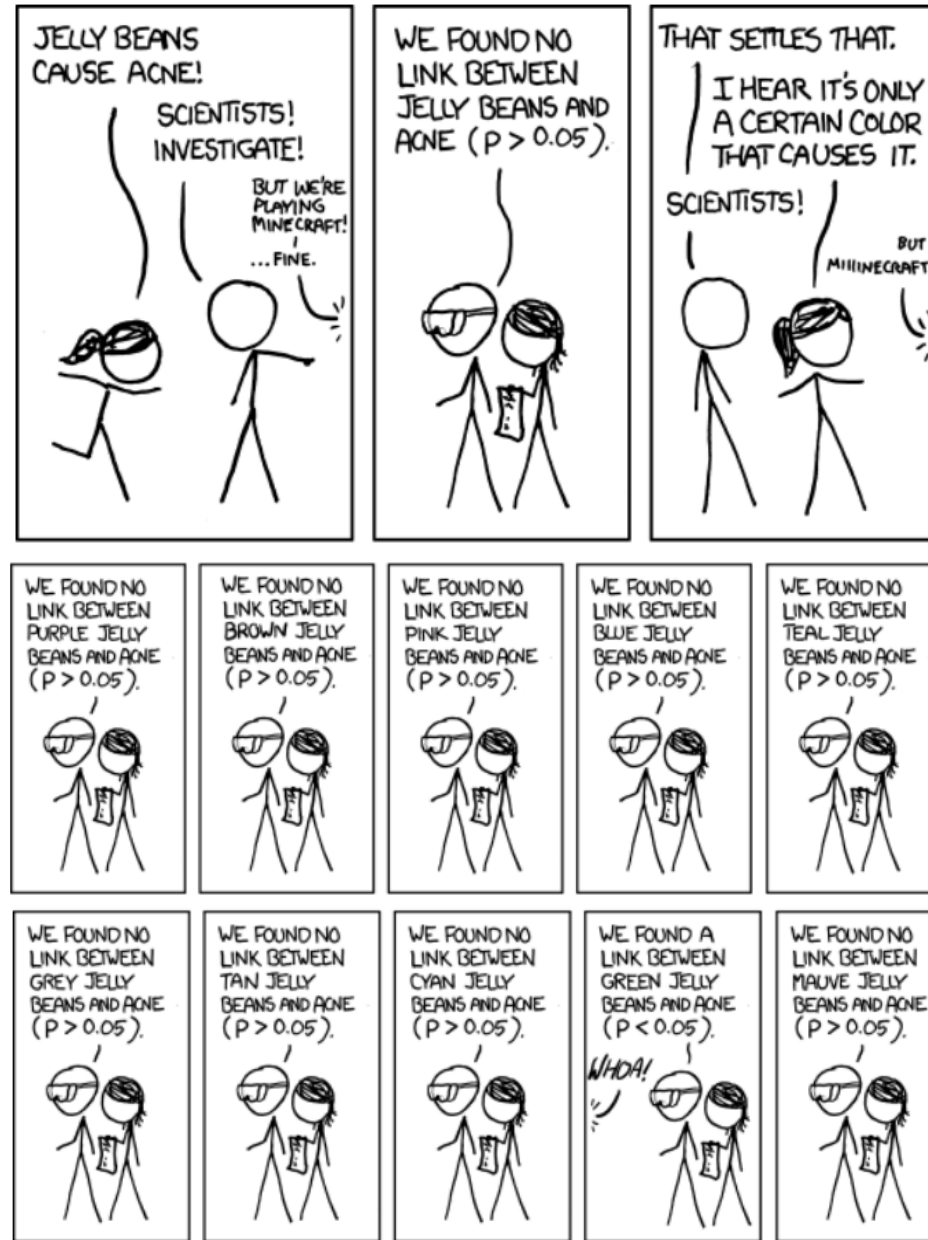
Ethical issues

- Anonymisation of data
 - Does NOT guarantee data is “anonymous”
 - E.g. Staff satisfaction questionnaire which asks for race and position
 - There may be only one person of that race with that position
 - E.g. 85% Americans identified by postcode, birth date and gender
 - In the UK, postcode and car model may be enough to identify a person even if car model is “common”.

Ethical issues

- Output from data mining must be carefully considered
 - Arguments purely based on statistics are not sufficient
 - Caveats should be put on conclusions
 - Includes p-hacking!

Significant



[xkcd: Significant](#)

The data ethics framework

1. Start with clear user need and public benefit
2. Be aware of relevant legislation and codes of practice
3. Use data that is proportionate to the user need
4. Understand the limitations of the data
5. Ensure robust practices and work within your skillset
6. Make your work transparent and be accountable
7. Embed data use responsibly

The data ethics workbook

- *“Should be completed collectively by practitioners, data governance or information assurance specialists, and subject matter experts like service staff or policy professionals”*
- Also decide how often to reassess the project with respect to the framework principles.
- See questions to be answered at
 - <https://www.gov.uk/government/publications/data-ethics-workbook/data-ethics-workbook> [accessed 04/09/2024]

BCS professional conduct

- Principles
 - Make IT for everyone
 - Show what you know, learn what you don't
 - Respect the organisation or the individual you work for
 - Keep IT real, keep IT professional, pass IT on.

Statistics code of practice

- 3 pillars:
 - Trustworthiness
 - Quality
 - Value
- 3 cross-cutting themes:
 - Collaboration
 - Coherence
 - Transparency

Contents

- Data mining
- Input and output
- Data mining and machine learning
- Applications
- Ethical issues
- **Summary**

Summary

- Very valuable information can be extracted from data
- Relies on a large set of examples and machine learning techniques.
- Methodology is often agile, e.g. CRISP-DM
- Format of input and output constrain what can be learnt.
- Wide range of applications.
- Ethical issues restrict use of data for certain purposes.