
# Data Mining Introduction

## Overview

AI+X
INSTITUTE FOR ARTIFICIAL INTELLIGENCE

# Administrative Details

- The text is a high-level overview of data mining.
- You can supplement this by papers discussed in the book. They will provide some details
- The WEKA Data mining tool is built in Java, but it is not necessary for a student to know the language to use it
- Discussions within Canvas: A forum just for this class for discussion
  - Students can ask and answer questions
- Instructor will try to step in where needed to correct a misconception or acknowledge a good answer

# What is Data Mining?

It can be described as “making sense of data”, or “intelligent data analysis”.

Understandable models of data may be extracted which help users make decisions.

New names: data analytics, data science

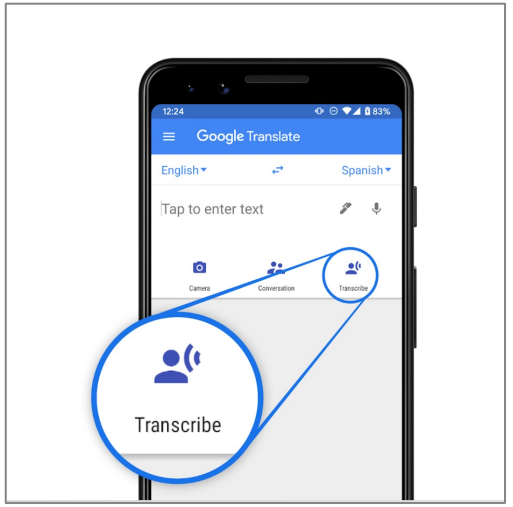

# What is Data Mining?

- A model which predicts the direction of a particular stock's price would be quite useful.
- There are models created by data mining which indicate whether credit for a purchase should be granted.

# Have you been affected by Data Mining?

Ever used a spam filter?
(Supervised learning)


You get suggested eBooks
(Unsupervised learning)




Google translate

# What is Data Mining?

- Data mining is used to find like users and use their preferences to suggest items to you.
- This is called Collaborative filtering.

# Data Mining

- Find patterns in data that provide insight or enable fast and accurate decision making
- Widely applicable, accurate patterns are needed for making decisions
  - Problem 1: most patterns are not interesting
  - Problem 2: patterns may be inexact (or spurious)
  - Problem 3: data may be garbled or missing
- Machine learning techniques identify patterns in data and provide many tools for data mining
- Of great interest are machine learning techniques that provide descriptions comprehensible by people

# Data Mining for Security

?

- Consider web pages on the Internet as nodes in a graph, can this structure be used to mine relationships?
- If so, could this be applied to database records in some way?

# Data Mining for Security

- Consider web pages on the Internet as nodes in a graph, can this structure be used to mine relationships?
  - Yes
- If so, could this be applied to database records in some way?
  - Yes

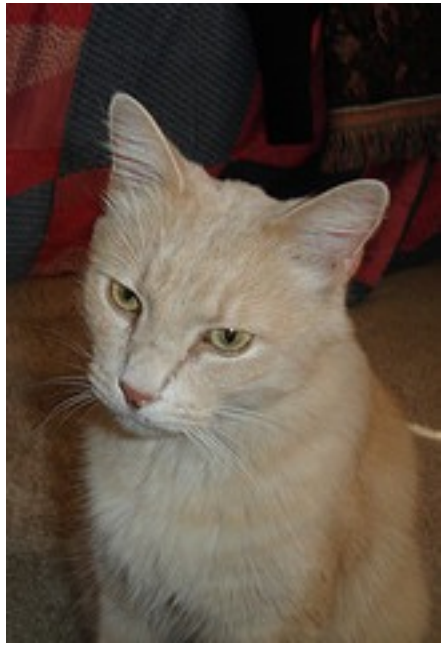
# Image Recognition – Lots of Data

Deep Neural Networks learned on million(s) of examples from ImageNet became dominant in categorizing images in 2016. There are (over) 1000 categories.

# Image Recognition – Lots of Data

# Image Recognition – Lots of Data

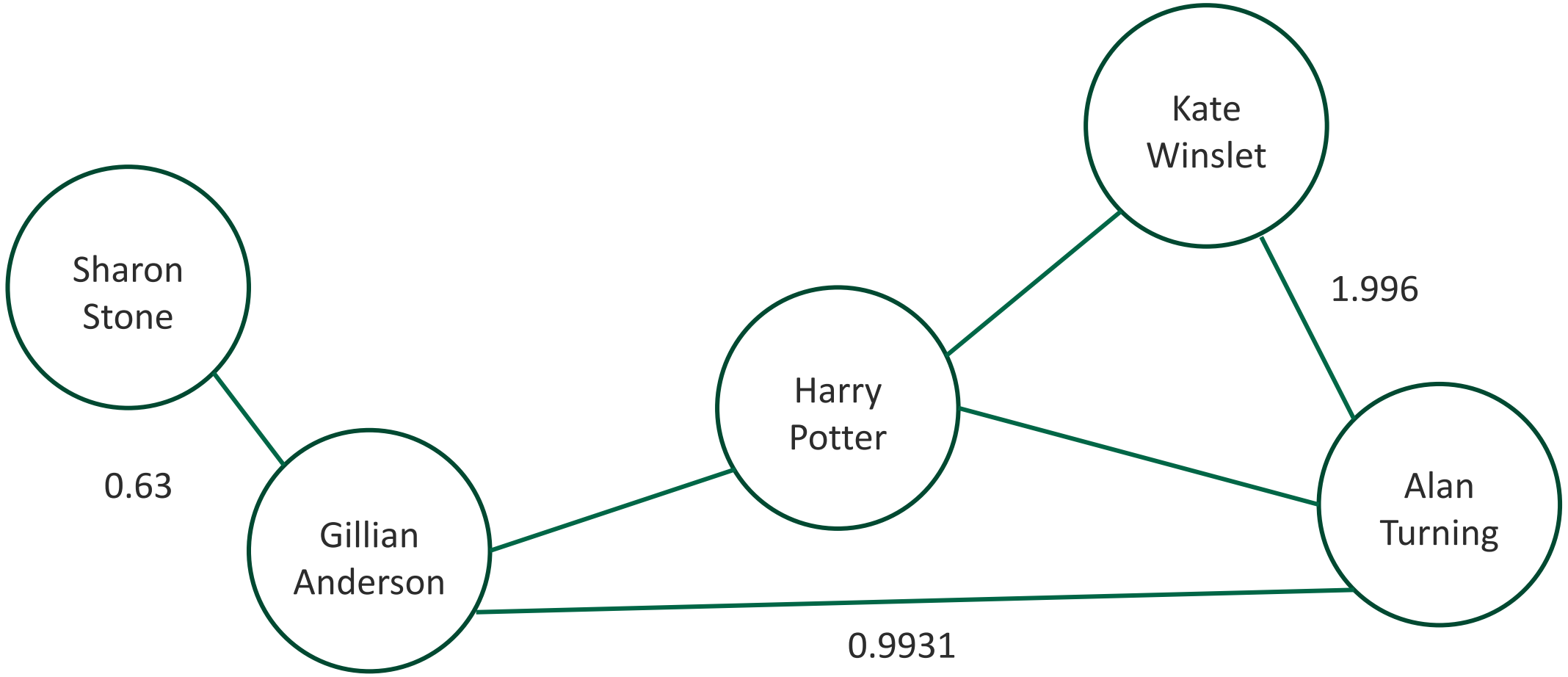
# Image Recognition – Lots of Data



# Data Mining = Deep Learning?

- With lots of data; deep learning performance must be shown inferior, today
- Deep learned models seem to work best on big data
  - The models have lots of parameters
- **No free lunch theorem** tell us other algorithms have value

# Web Mining for Relationships between Well-known People




Faloutsos, C, et.al. Fast Fast Discovery of Connection Subgraphs, pp. 118-127, KDD2004

# Course Perspective

- Data mining can be viewed in different ways:
  - An application of machine learning
  - An application of statistics
  - Visualization and one of the above
  - Some mixture of the above
- Can be thought of as search for the best model of the data (predictive or descriptive)

**In this course**

- Focus on data mining in the applied machine learning sense.
- We cover a number of machine learning approaches
- It is important to remember that the "***no free lunch theorem***" tells us that there is no machine learning algorithm that is the best for all data sets!

# Course Perspective

- For data mining problems with Images, Voice, or Text (among others) and lots of data, use
  - Neural Networks with two or more hidden layers
  - Deep Learning which has, since about 2012, been found to be most effective.
- Big data problems like image recognition or object recognition, voice recognition (Siri, Alexa, etc.), translation (Google translate) and more are generally addressed with a form of Deep Learning
- For data mining problems with small data sets or mixed nominal and numeric data non-neural networks can be fast to train and accurate
  - Deep Neural networks have MANY parameters to tune. Decision trees have few
  - Deep Neural Networks are sometimes hard to train. Rule learners and to some extent Support Vector Machines are generally pretty easy to train
- This class starts with pre-deep learning approaches and then moves on to cover Deep Learning with some theory

# Issues: Where does all data come from?

AT&T sees enough telephone calls each day that it cannot store the data.

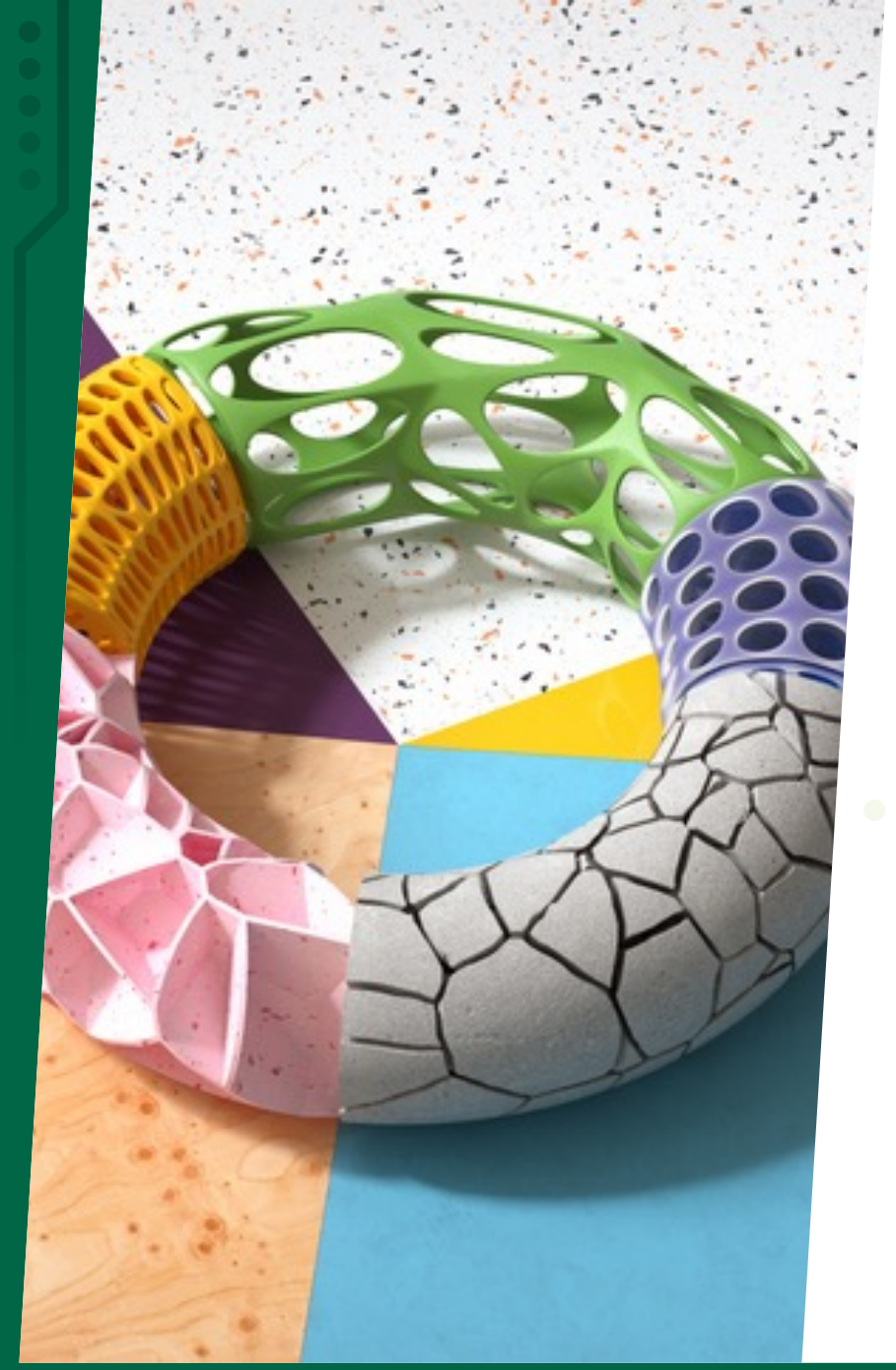
Biology contains a very rich and diverse set of genomic data. Protein folding (Alpha-Fold) is important.

There are many images posted to Facebook, Instagram, etc.

Lots of text documents exist on the web.

# Issues: What is data cleaning?


**Data cleaning:** the practice of removing errors, mislabeling, incorrect values, etc. from the data.



# Data Sets

- Our data will be made up of attributes which can be nominal or continuous or ordinal.
- Each attribute will be able to take on a set of values.
- One description of data mining via learning is to search through the representation space for the best model of the data.
- Our data may come with class labels, associated numeric values, or no labels at all.
- Unlabeled data may be utilized in building association rules or clustering (Unsupervised).
- Data with values associated with the attribute (feature) vector may be used in regression analysis (Supervised).

# Data Set Sizes

- There may be so much data that it must be treated in a streaming fashion, where incremental information is used.
- For very large data sets, we will discuss distributed data mining approaches.
- Deep learning is also applied to very large data.  You must (patiently) wait for a solution.

# Some Contact Lens Data

| *Age* | *Spectacle prescrip* | *Astigmatism* | *Tear production rate* | ***Lenses*** |
|---|---|---|---|---|
| young | myope | no | reduced | **none** |
| young | myope | no | normal | **soft** |
| young | myope | yes | reduced | **none** |
| young | myope | yes | normal | **hard** |
| young | hypermetrope | no | reduced | **none** |
| young | hypermetrope | no | normal | **soft** |
| young | hypermetrope | yes | reduced | **none** |
| young | hypermetrope | yes | normal | **hard** |

# Some Contact Lens Data

```
If tear production rate = reduced then recommendation = none.
If age = young and astigmatic = no and tear production rate = normal
   then recommendation = soft
If age = pre-presbyopic and astigmatic = no and tear production
   rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope and
   astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no and
   tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = young and astigmatic = yes and tear production rate =
    normal
   then recommendation = hard
If age = pre-presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
```

Figure 1. Rules for the contact lens data

# Some Contact Lens Data

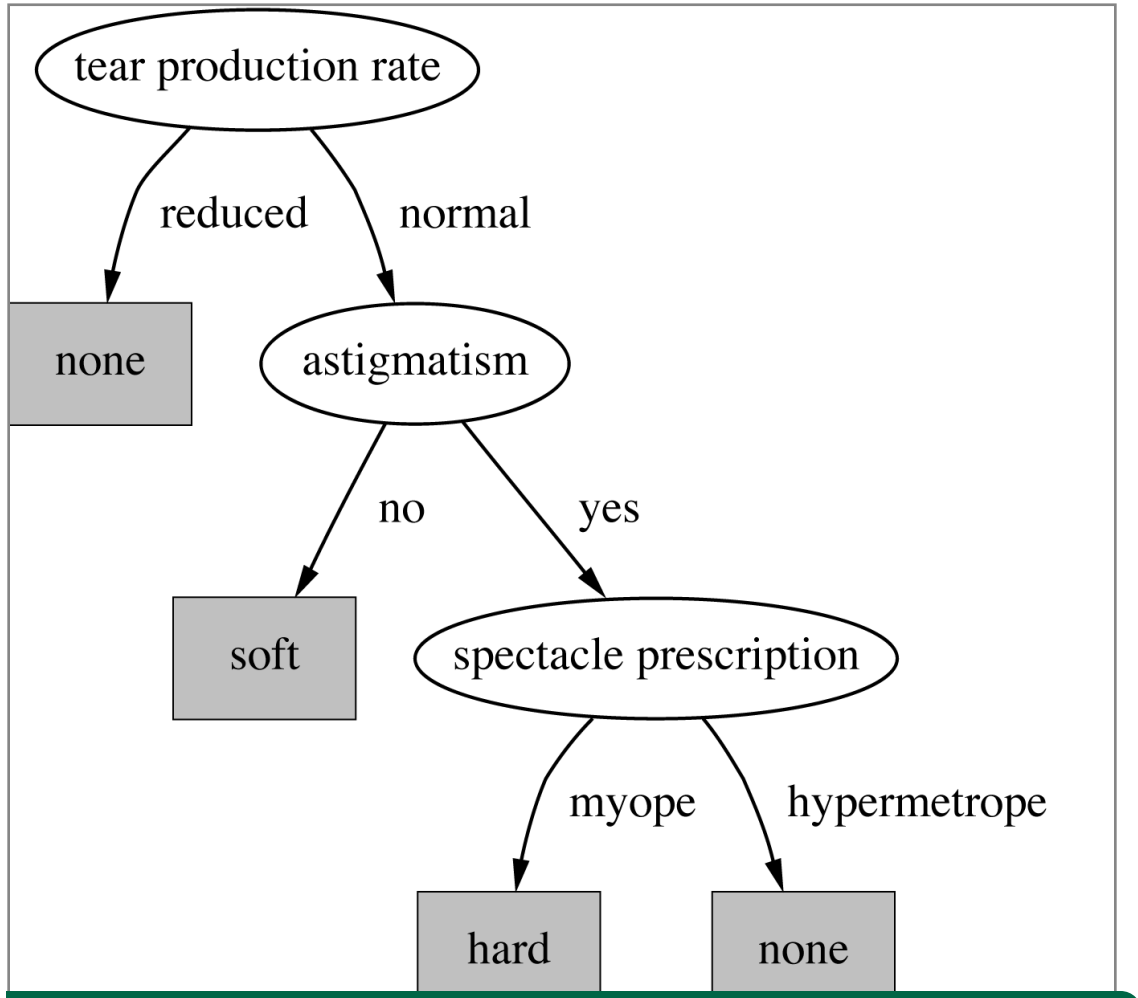



Figure 2. Decision tree for the contact lens data

# Representation Differences

The decision tree is easier to understand than the rule set.

However, simple decision trees are often less accurate in prediction tasks.

However, sometimes it is necessary to trade-off accuracy for understandability.

For instance, it may be necessary to explain why credit is not recommended to be granted to a customer by a Data mining system.

# Knowledge Check 1

An example of data mining is:

A. A loan officer looking at a person's credit history to decide on a loan.

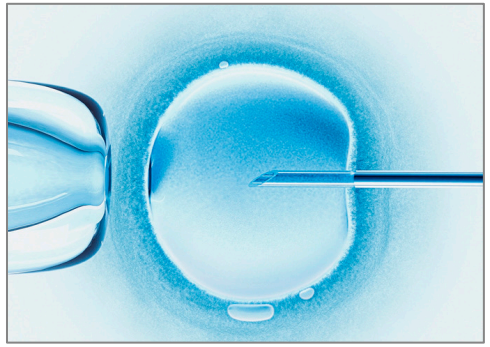
B. A company using data on all customers to suggest a product to a previous customer visiting their web site.

C. A spam filter that blocks addresses given by a user.

D. Google news grouping of web pages by subject.

# Information is Crucial (Supervised)

## Vitro Fertilization



- Given: embryos described by 60 features
- Problem: selection of embryos that will survive
- Data: historical records of embryos and outcome

## Cow Culling

- Given: cows described by 700 features
- Problem: selection of cows that should be culled
- Data: historical records and farmers' decisions

# Data Mining

Extracting information from data

- Implicit
- Previously unknown
- Potentially useful

Need programs that detect patterns and regularities in the data

# Machine Learning Techniques

Algorithms for acquiring structural descriptions from examples

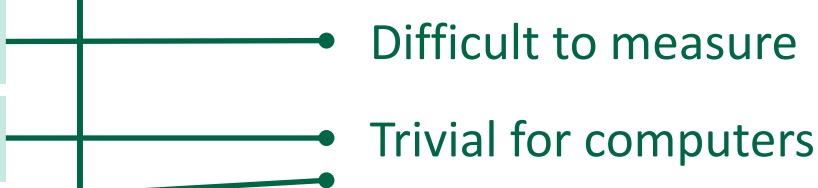
- Structural descriptions represent patterns explicitly
  - Can be used to predict outcome in new situation
  - Can be used to understand and explain how prediction is derived (may be even more important)

Methods originated from artificial intelligence, statistics, and research on databases.

# Can Machines really learn?

**Definitions of “learning” from dictionary**

- To get knowledge of by study, experience, or being taught
- To become aware by information or from observation
- To commit to memory
- To be informed of, ascertain; to receive instruction



# Can Machines really learn?

**Learning (Operational definition):** Things learn when they change their behavior in a way that makes them perform better in the future.

Does learning imply intention?

# The Weather Problem

Conditions for playing a certain game

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | Hot | High | False | No |
| Sunny | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Rainy | Mild | Normal | False | Yes |
| … | … | … | … | … |

```
If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes
```

# Classification vs. Association Rules

- **Classification rule** predicts value of a given attribute (the classification of an example).

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

- **Association rule** predicts value of arbitrary attribute (or combination).

```
If temperature = cool then humidity = normal
If humidity = normal and windy = false
   then play = yes
If outlook = sunny and play = no
   then humidity = high
If windy = false and play = no
   then outlook = sunny and humidity = high
```

# Knowledge Check 2

Association rules may have which attribute in its conclusion:

A. The class

B. Any

C. Only combinations of attributes

D. The class and any other

# Weather Data with Mixed Attributes

Some attributes have numeric values

| Outlook | Temperature | Humidity | Windy | Play |
|---|---|---|---|---|
| Sunny | 85 | 85 | False | No |
| Sunny | 80 | 90 | True | No |
| Overcast | 83 | 86 | False | Yes |
| Rainy | 75 | 80 | False | Yes |
| … | … | … | … | … |

```
If outlook = sunny and humidity > 83 then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity < 85 then play = yes
If none of the above then play = yes
```

# Classifying Iris Flowers

| | Sepal Length | Sepal Width | Petal Length | Petal Width | Type |
|---|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | Iris setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | Iris setosa |
| … | | | | | |
| 51 | 7.0 | 3.2 | 4.7 | 1.4 | Iris Versicolor |
| 52 | 6.4 | 3.2 | 4.5 | 1.5 | Iris Versicolor |
| … | | | | | |
| 101 | 6.3 | 3.3 | 6.0 | 2.5 | Virginica |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 | Virginica |
| … | | | | | |

```
If petal length < 2.45 then Iris setosa
If sepal width < 2.10 then Iris versicolor
...
```

# Predicting CPU Performance

Example: 209 different computer configurations

| | Cycle time (ns) | Main memory (Kb) | | Cache (Kb) | Channels | | Performance |
|---|---|---|---|---|---|---|---|
| | MYCT | MMIN | MMAX | CACH | CHMIN | CHMAX | PRP |
| 1 | 125 | 256 | 6000 | 256 | 16 | 128 | 198 |
| 2 | 29 | 8000 | 32000 | 32 | 8 | 32 | 269 |
| … | | | | | | | |
| 208 | 480 | 512 | 8000 | 32 | 0 | 0 | 67 |
| 209 | 480 | 1000 | 4000 | 0 | 0 | 0 | 45 |

Linear regression function:

```
PRP = -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX +
      0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

# Data from Labor Negotiations

| Attribute | Type | 1 | 2 | 3 | … | 40 |
|---|---|---|---|---|---|---|
| Duration | (Number of years) | 1 | 2 | 3 | | 2 |
| Wage increase first year | Percentage | 2% | 4% | 4.3% | | 4.5 |
| Wage increase second year | Percentage | ? | 5% | 4.4% | | 4.0 |
| Wage increase third year | Percentage | ? | ? | ? | | ? |
| Cost of living adjustment | {none,tcf,tc} | none | tcf | ? | | none |
| Working hours per week | (Number of hours) | 28 | 35 | 38 | | 40 |
| Pension | {none,ret-allw, empl-cntr} | none | ? | ? | | ? |
| Standby pay | Percentage | ? | 13% | ? | | ? |
| Shift-work supplement | Percentage | ? | 5% | 4% | | 4 |
| Education allowance | {yes,no} | yes | ? | ? | | ? |
| Statutory holidays | (Number of days) | 11 | 15 | 12 | | 12 |
| Vacation | {below-avg,avg,gen} | avg | gen | gen | | avg |
| Long-term disability assistance | {yes,no} | no | ? | ? | | yes |
| Dental plan contribution | {none,half,full} | none | ? | full | | full |
| Bereavement assistance | {yes,no} | no | ? | ? | | yes |
| Health plan contribution | {none,half,full} | none | ? | full | | half |
| Acceptability of contract | {good,bad} | bad | good | good | | good |

# Data from Labor Negotiations

| Attribute | Type | 1 | 2 | 3 | … | 40 |
|---|---|---|---|---|---|---|
| Duration | (Number of years) | 1 | 2 | 3 | | 2 |
| Wage increase first year | Percentage | 2% | 4% | 4.3% | | 4.5 |
| Wage increase second year | Percentage | ? | 5% | 4.4% | | 4.0 |
| Wage increase third year | Percentage | ? | ? | ? | | ? |
| Cost of living adjustment | {none,tcf,tc} | none | tcf | ? | | none |
| Working hours per week | (Number of hours) | 28 | 35 | 38 | | 40 |
| Pension | {none,ret-allw, empl-cntr} | none | ? | ? | | ? |
| Standby pay | Percentage | ? | 13% | ? | | ? |
| Shift-work supplement | Percentage | ? | 5% | 4% | | 4 |
| Education allowance | {yes,no} | yes | ? | ? | | ? |
| Statutory holidays | (Number of days) | 11 | 15 | 12 | | 12 |
| Vacation | {below-avg,avg,gen} | avg | gen | gen | | avg |
| Long-term disability assistance | {yes,no} | no | ? | ? | | yes |
| Dental plan contribution | {none,half,full} | none | ? | full | | full |
| Bereavement assistance | {yes,no} | no | ? | ? | | yes |
| Health plan contribution | {none,half,full} | none | ? | full | | half |
| Acceptability of contract | {good,bad} | bad | good | good | | good |

Missing data

# Decision Trees for the Labor Data

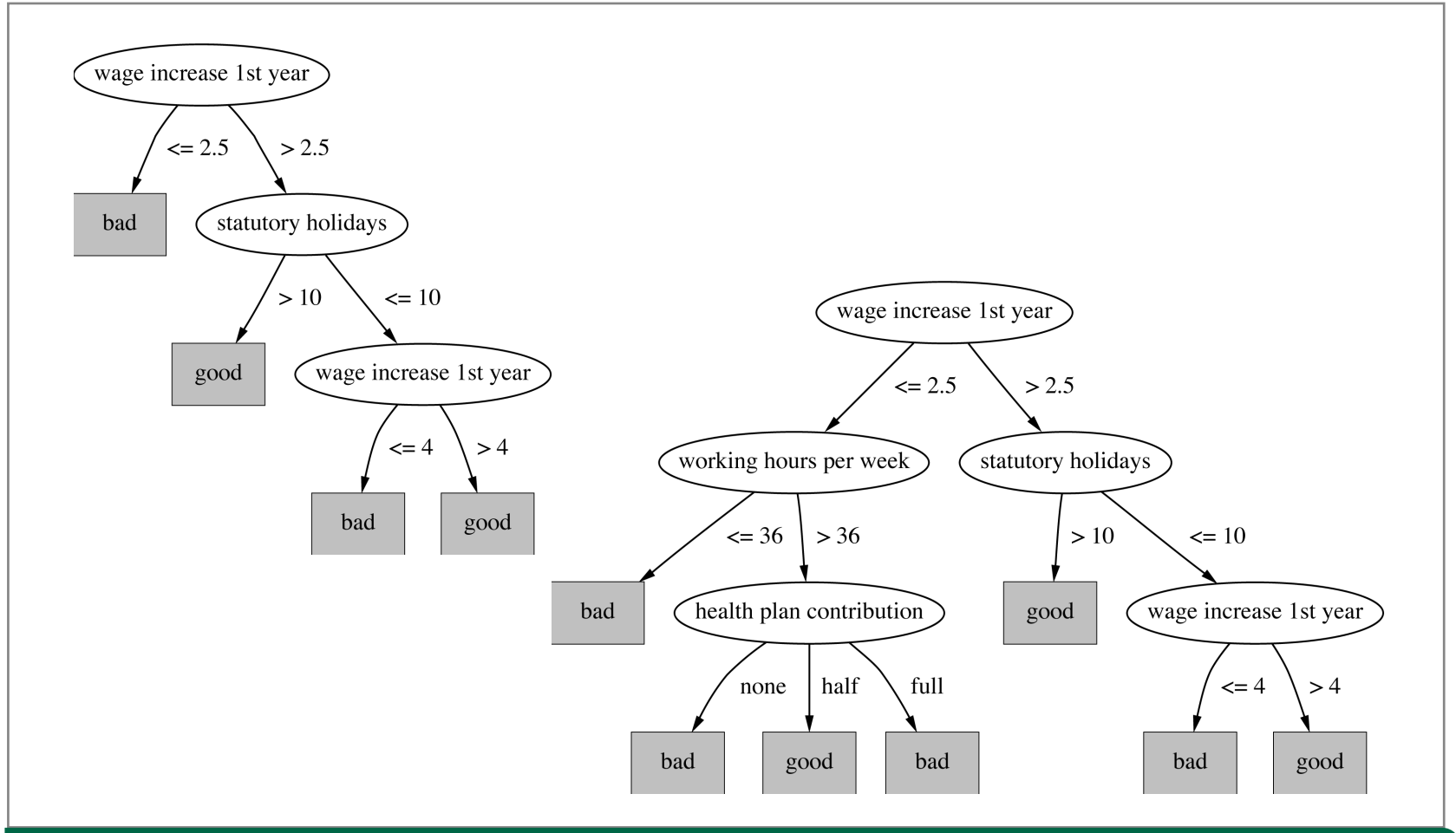



Figure 3. Decision tress for the labor data

# Soybean Classification

| | Attribute | Number of values | Sample Value |
|---|---|---|---|
| Environment | Time of Occurrence | 7 | July |
| | Precipitation | 3 | Above normal |
| … | | | |
| Seed | Condition | 2 | Normal |
| | Mold growth | 2 | Absent |
| … | | | |
| Fruit | Condition of fruit pods | 4 | Normal |
| | Fruit spots | 5 | ? |
| Leaves | Condition | 2 | Abnormal |
| | Leaf spot size | 3 | ? |
| … | | | |
| Stem | Condition | 2 | Abnormal |
| | Stem lodging | 2 | Yes |
| … | | | |
| Roots | Condition | 3 | Normal |
| Diagnosis | | 19 | Diaporthe stem canker |

# The Role of Domain Knowledge

```
If leaf condition is normal
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

```
If leaf malformation is absent
   and stem condition is abnormal
   and stem cankers is below soil line
   and canker lesion color is brown
then
   diagnosis is rhizoctonia root rot
```

But in this domain, “leaf condition is normal” implies “leaf malformation is absent”. Rules are the same.

# Fielded Applications

The result of learning of the learning method itself is deployed in practical applications:

- Processing loan applications
- Screening images for oil slicks
- Electricity supply forecasting
- Diagnosis of machine faults
- Marketing and sales
- Google Translate – Deep Learning
- Autoclave layout for aircraft parts
- Automatic classification of sky objects
- Automated completion of repetitive forms
- Text retrieval

# Processing Loan Application


- Given: Questionnaire with financial and personal information
- Question: Should money be lent?
- Simple statistical method covers 90% of cases
- Borderline cases referred to loan officers
  - 50% of accepted borderline cases defaulted
- Solution: Reject all borderline cases?
  - No! Borderline cases most active customers

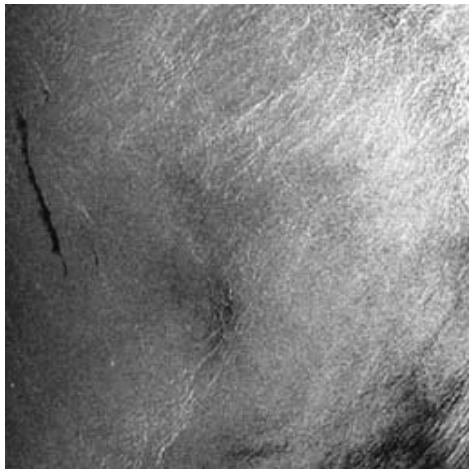
# Processing Loan Application

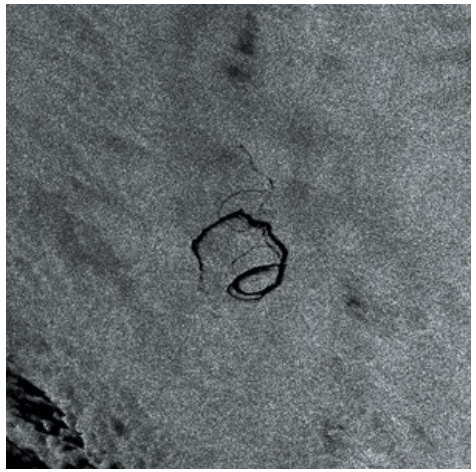
## Entering Machine Learning

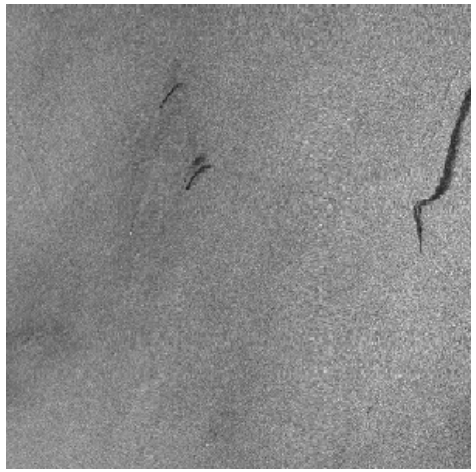
- 1000 training examples of borderline cases
- 20 attributes:
  - Age
  - Years with current employer
  - Years at current address
  - Years with the bank
  - Other credit cards possessed,…
- Learned rules: correct on 70% of cases
  - Human experts only 50%
- Rules could be used to explain decisions to customers

# Screening Images

- Given: radar satellite images of coastal waters
- Problem: detect oil slicks in those images
- Oil slicks appear as dark regions with changing size and shape
- Not easy: lookalike dark regions can be caused by weather conditions (e.g. high wind)
- Expensive process requiring highly trained personnel







# Screening Images

## Entering Machine Learning

- Extract dark regions from normalized image
- Attributes:
  - Size of region
  - Shape, area
  - Intensity
  - Sharpness and jaggedness of boundaries
  - Proximity of other regions
  - Info about background
- Constraints:
  - Few training examples—oil slicks are rare!
  - Unbalanced data: most dark regions aren’t slicks
  - Regions from same image form a batch
  - Requirement: adjustable false-alarm rate

# Load Forecasting

- Electricity supply companies need forecast of future demand for power
- Forecasts of min/max load for each hour -> significant savings
- Given: manually constructed load model that assumes “normal” climatic conditions
- Problem: adjust for weather conditions
- Static model consists of
  - Base load for the year
  - Load periodicity over the year
  - Effect of holidays

# Load Forecasting

## Entering Machine Learning

- Prediction corrected using “most similar” days
- Attributes:
  - Temperature
  - Humidity
  - Wind speed
  - Cloud cover readings
  - Difference between actual load and predicted load
- Average difference among three “most similar” days added to static model
- Linear regression coefficients form attribute weights in similarity function

# Marketing and Sales I

Companies precisely record massive amounts of marketing and sales data.

Applications

- Customer loyalty
  - Identifying customers that are likely to defect by detecting changes in their behavior (e.g., banks/phone companies)
- Special offers
  - Identifying profitable customers (e.g., reliable owners of credit cards that need extra money during the holiday season.)

# Marketing and Sales II

## Market basket analysis

- Association techniques find groups of items that tend to occur together in a transaction (used to analyze checkout data)

## Historical analysis of purchasing patterns

## Identifying prospective customers

- Focusing promotional mailouts (targeted campaigns are cheaper than mass-marketed ones)

# The Data Mining Process

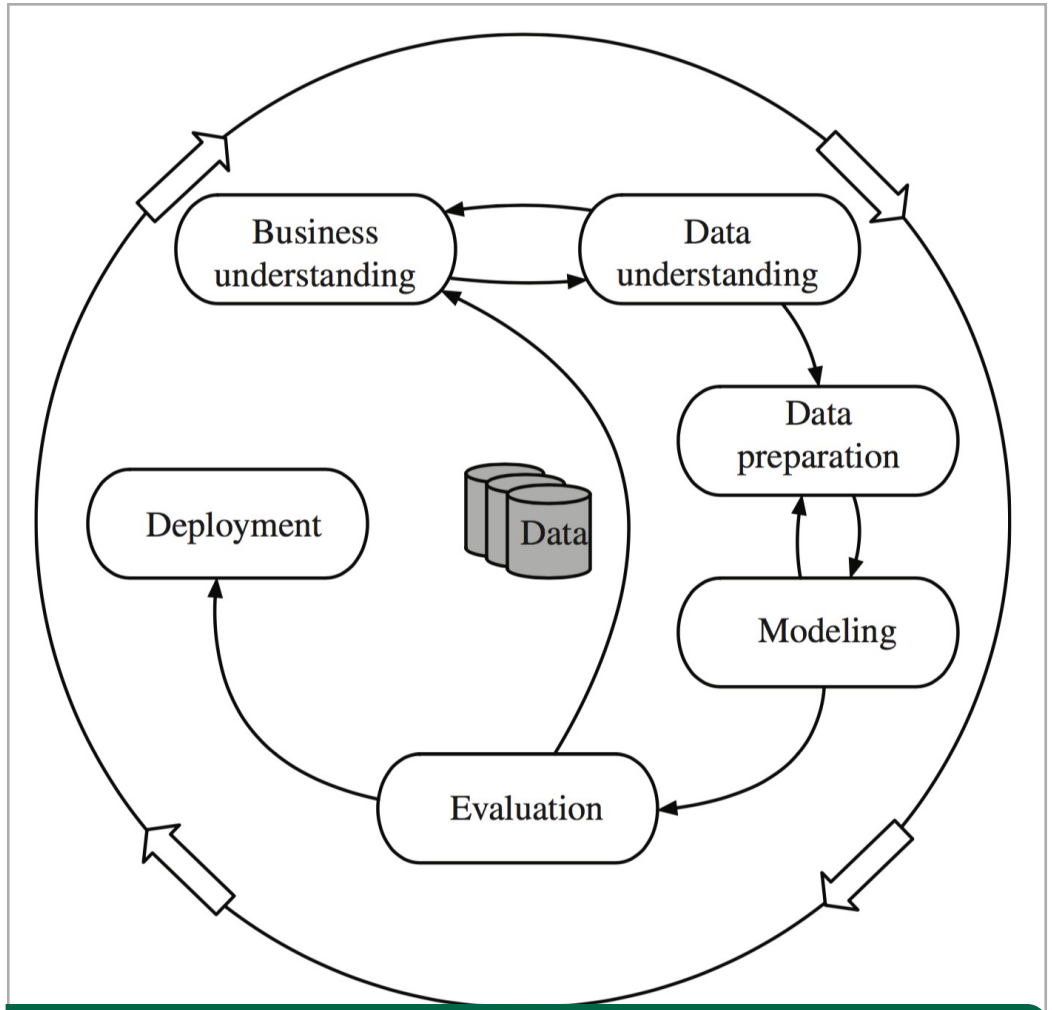



Figure 4. The data mining process

# Machine Learning and Statistics

- Historical difference (grossly oversimplified):
  - Statistics: testing hypotheses
  - Machine learning: finding the right hypothesis
- But: huge overlap
  - Decision trees (C4.5 and CART)
  - Nearest-neighbor methods
- Today: perspectives have converged
  - Most ML algorithms employ statistical techniques

# Generalization as Search

- Inductive learning: find a concept description that fits the data
- Example: rule sets as description language
  - Enormous, but finite, search space
- Simple solution:
  - Enumerate the concept space
  - Eliminate descriptions that do not fit examples
  - Surviving descriptions contain target concept

# Enumerating the Concept Space

- Search space for weather problem
  - 4 x 4 x 3 x 3 x 2 = 288 possible combinations
  - With 14 rules -> $2.7 \times 10^{34}$ possible rule sets
- Solution: greedy directed search in the space

# Enumerating the Concept Space

- Search space for weather problem
  - 4 x 4 x 3 x 3 x 2 = 288 possible combinations
  - With 14 rules -> $2.7 \times 10^{34}$ possible rule sets
- Solution: greedy directed search in the space
- Other practical problems
- More than one description may survive
  - Language is unable to describe target concept
  - Or data contains noise

# Bias

**Important decisions in learning systems**

- Concept description language
- Order in which the space is searched
- Way that overfitting to the particular training data is avoided.

**These form the “bias” of the search:**

- Language bias
- Search bias
- Overfitting-avoidance bias

# Language Bias

Is language bias universal or does it restrict what can learned?

- Universal language can express arbitrary subsets of examples.
- If language includes logical or (“disjunction”), it is universal.
- Example: rule sets
- Domain knowledge can be used to exclude some concept descriptions a priori from the search.

# Search Bias

## Search heuristic

- "Greedy" search: performing the best single step
- "Beam search": keep several alternatives

## Direction of search

- General-to-specific
  - E.g., specializing a rule by adding conditions
- Specific-to-general
  - E.g., generalizing an individual instance into a rule

# Overfitting-avoidance Bias

Can be seen as a form of search bias

Modified evaluation criterion

- E.g., balancing simplicity and number of errors

Modified search strategy

- E.g., pruning (simplifying a description)
  - Pre-pruning: stops at a simple description before search proceeds to an overly complex one
  - Post-pruning: generates a complex description first and simplifies it afterwards


# Data Mining Ethics I

- Ethical issues arise in practical applications
- Data mining often used to discriminate
  - E.g., loan applications: using some information (e.g., sex, religion, race) is unethical
- Ethical situation depends on application
  - E.g., same information ok in medical application
- Attributes may contain problematic information
  - E.g., area code may correlate with race

# Data Mining Ethics II

- Important questions:
  - Who is permitted access to the data?
  - For what purpose was the data collected?
  - What kind of conclusions can be legitimately drawn from it?
- Caveats must be attached to results.
- Purely statistical arguments are never sufficient!
- Are resources put to good use?

You have reached the end of the lecture.