

IS 733 Lesson 2

Know your data

Announcements

- Homework 1 is out, posted at the course web page, due 2/16/2021
- Submit on Blackboard. Look for it under “Assignments”
- Either typed or handwritten + scanned answers are fine. There are often pen-and-paper calculations in this course. Consider using the CamScanner app to scan handwritten solutions.

Announcements

- Project handout is out (also on course webpage).
 - Start looking for teammates and thinking about a topic!
 - Blackboard has a “find other users” feature or use the class discussion forum to broadcast your novel ideas to seek for group member.
 - Groups formed by 2/16, proposal due 2/23!
- I have created synchronous class on Blackboard where you can join half an hour before each class (starting at 6:40 pm as class starts at 7:10 pm) and can chat with your peers. You could find teammates there.

Blackboard

- Use Blackboard discussion forum to ask questions about the course, instead of emailing me directly, so that everyone in the class can benefit from the answer
- Blackboard will also be factored into participation grades: two posts (questions, answers, comments) worth 1% of grade
- Blackboard will also be used for **announcements**, information sharing on the readings, and class **email** communications

Reminder: Sign up for Poll Everywhere

- Vote on polls at Pollev.com/nirmalyaroy910

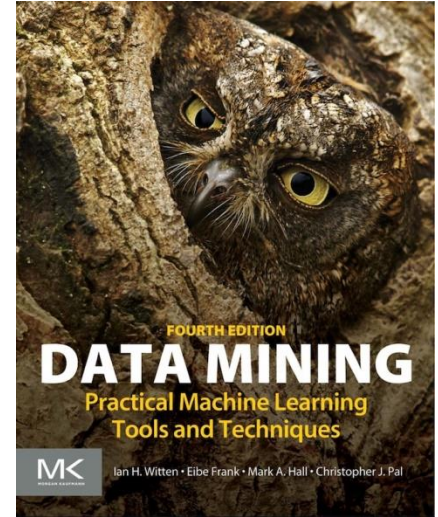
Reminder: Course Readings

- Required textbook:

Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition (Witten et al.)

See: <http://www.cs.waikato.ac.nz/ml/weka/book.html>

- Readings need to be completed **before each class**. We will do reading quizzes at the start of each class.
- It is **very important that you do the readings** so that we can make effective use of our limited lecture time together (a “flipped classroom” approach.)



Academic Integrity

- UMBC's policies on academic integrity will be strictly enforced
- **All of your work must be your own.**
- **Acknowledge** and **cite** source material in your papers or assignments.
- While you may verbally discuss assignments with your peers, **you may not copy or look at anyone else's written assignment work or code, or share your own solutions.**
- **Any exceptions will result in a zero on the assessment in question, and may lead to further disciplinary action.**

Academic Integrity

- “Cheating, fabrication, plagiarism, and helping others to commit these acts are all forms of academic dishonesty, and they are wrong.”
-(UMBC's academic integrity overview)
- "Students shall not submit as their own work **any work which has been prepared by others.**"
-(USM policy document)

The input to a typical machine learning scheme is a set of ____

concepts

concept descriptions

instances

attributes

The input to a typical machine learning scheme is a set of ____

concepts
concept descriptions
instances
attributes

In _____, groups of examples that belong together are sought.

Classification learning

Association learning

Clustering

Numeric prediction

In _____, groups of examples that belong together are sought.

Classification learning

Association learning

Clustering

Numeric prediction

The ARFF file format used by the Weka data mining system specifies which attribute is the class.

True

False

The ARFF file format used by the Weka data mining system specifies which attribute is the class.

True

False

Learning outcomes

By the end of the lesson, you should be able to:

- **Explain** what the typical inputs and outputs of data mining methods are, using appropriate terminology
- **Convert** raw data into the format required by typical machine learning methods
- **Discuss** several strategies for getting to know your data

Privacy and Machine Learning

- As individuals and consumers we benefit from ML systems trained on **OUR** data
 - **Internet search**
 - **Recommendations**
 - products, movies, music, news, restaurants, email recipients
 - **Mobile phones**
 - Autocorrect, speech recognition, Siri, ...



The cost is our privacy

Forbes / Tech

FEB 16, 2012 @ 11:02 AM 2,998,353 VIEWS

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill
FORBES STAFF

Welcome to The Not-So
Private Parts where
technology & privacy
collide

FULL BIO >

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target TGT -0.43%, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the [New York Times](#) how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a

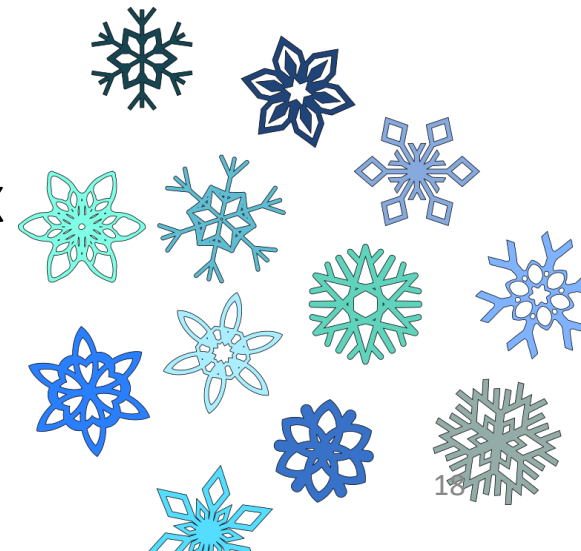


Target has got you in its aim

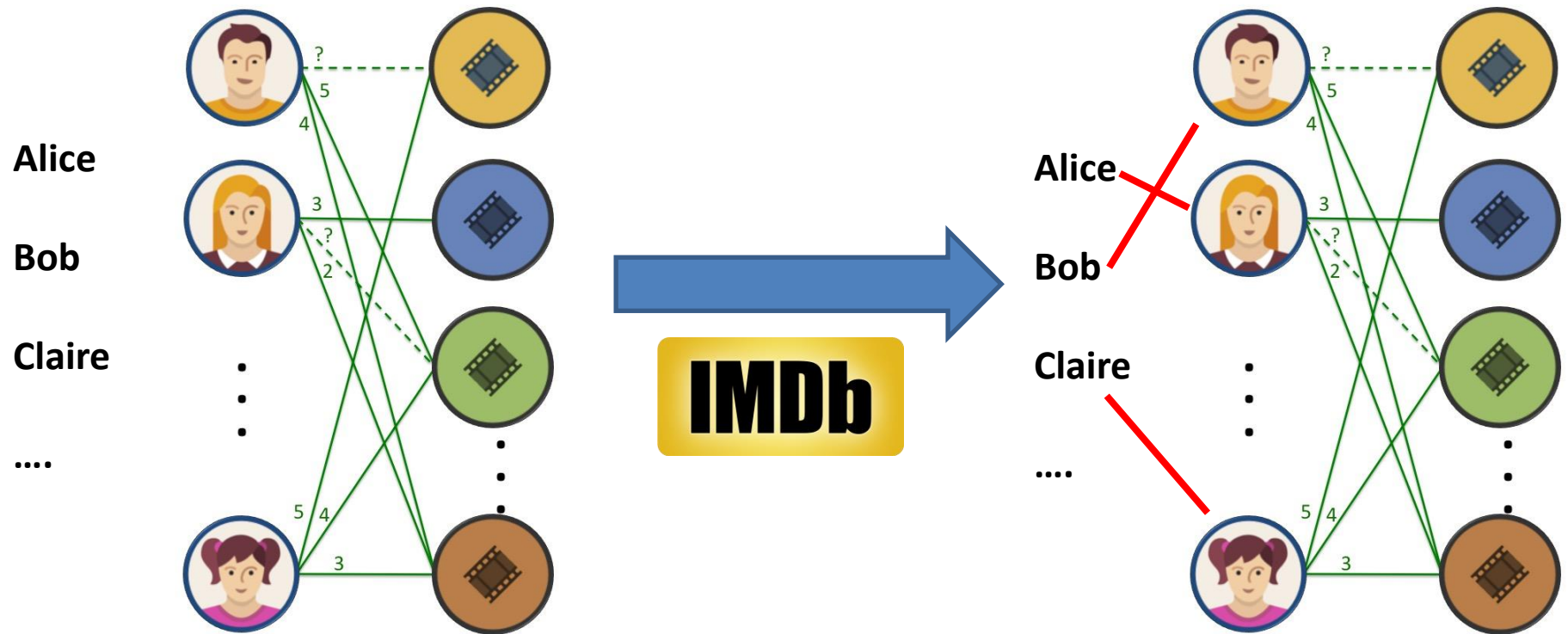
Data mining and ethics



- Ethical issues arise in practical applications, when using personal data from human beings
- Anonymizing data is difficult
 - 85% of Americans can be identified from just zip code, birth date and sex
 - As dimensions / attributes increase, pretty quickly you become unique



Anonymization Fails



Anonymized Netflix data + public IMDB data = identified Netflix data

Data mining and ethics



- 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
- Ensuring **fairness in machine learning/data mining** is now an active area of research

Think-pair-share

- Come up with the application of data mining that:
 1. Could make you (personally) the most money
 2. Could have the most benefit to the environment
 3. Could save the most lives
 4. Could do the most evil

Think-pair-share

- Come up with the application of data mining that:
 1. Could make you (personally) the most money

Stock trading, product recommendation, sales forecasting, targeted ads, monitoring energy consumption, crypto, finding good deals, ecommerce

Think-pair-share

- Come up with the application of data mining that:

Could have the most benefit to the environment

Clean energy, identifying harmful industries, quality, pollution monitoring, CO₂, gas/electricity, exploitation of resources, weather forecasting, identifying environmental issues, waste management, sustainability

Think-pair-share

- Come up with the application of data mining that:

Could save the most lives

Contact tracing, predictive medicine, traffic, covid factors, heart disease, crime prediction, weather, genomic info for personalized medicine, cancer, patient monitoring, population health, natural disasters

Think-pair-share

- Come up with the application of data mining that:

Could do the most evil

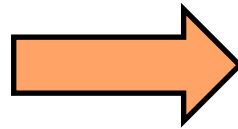
Govt tracking, genetic abuse, fraud, hacking, autonomous drones in military, manipulating elections, social media privacy, harmful data collection practices, racial profiling, start wars, medical data.

Know your data

Data Mining

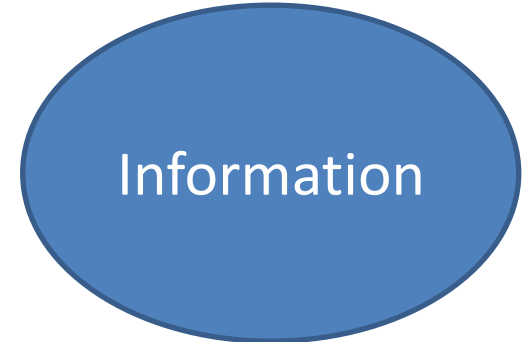


Complicated, noisy,
high-dimensional



Data Mining

Find patterns



Explore, understand, predict

Summary of what the next slides will cover

Input to data mining: concepts, instances, attributes

- Components of the input for learning
- What's a **concept**?
 - Classification, association, clustering, numeric prediction
- What's in an **example**?
 - Relations, flat files, recursion
- What's in an **attribute**?
 - Nominal, ordinal, interval, ratio
- **Preparing the input**
 - ARFF, sparse data, attributes, missing and inaccurate values, unbalanced data, getting to know your data

Components of the input

- **Concepts**: kinds of things that can be learned
 - Aim: intelligible and operational **concept description**
- **Instances**: the individual, independent examples of a concept to be learned
 - A.k.a. **examples, data points, samples, data objects**
 - More complicated forms of input with dependencies between examples are possible
- **Attributes**: measuring aspects of an instance
 - A.k.a. features, dimensions, covariates, variables
 - We will focus on nominal and numeric ones

What's a concept?

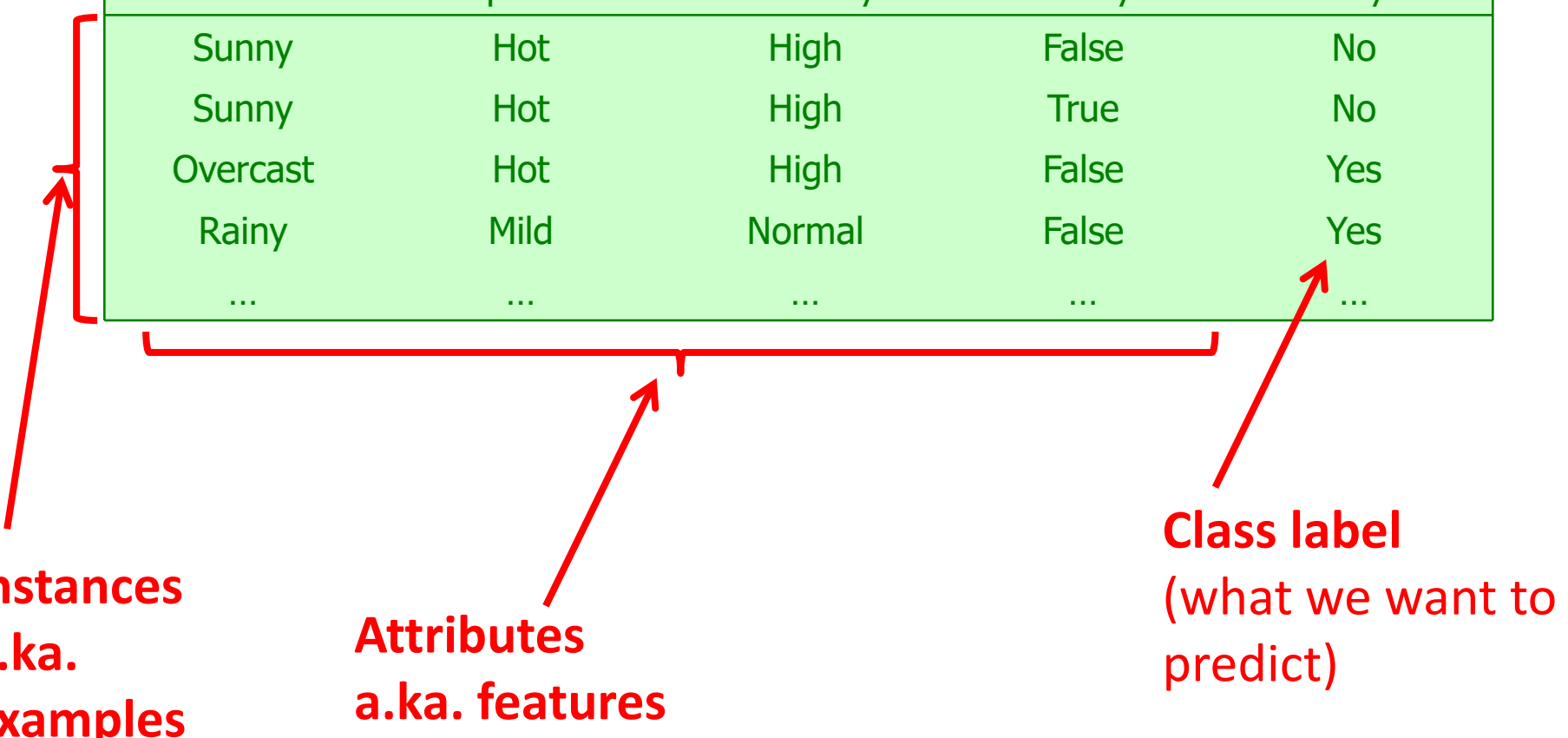
- **Concept**: thing to be learned
- **Concept description**: output of the learning scheme
- Styles of learning:
 - **Classification learning**:
predicting a discrete class
 - **Association learning**:
detecting associations between features
 - **Clustering**:
grouping similar instances into clusters
 - **Numeric prediction**:
predicting a numeric quantity

Classification learning

- Example problems: weather data, contact lenses, irises, labor negotiations
- **Classification learning** is *supervised*
 - Scheme is provided with actual outcome
- Outcome is called the *class* of the example
- Measure success on fresh data for which class labels are known (*test data*)
- In practice success is often measured subjectively

Classification example: the weather problem

- Conditions for playing a certain game



The diagram shows a table with 5 columns and 6 rows. The first four columns are grouped by a red bracket labeled 'Attributes a.k.a. features'. The last column is pointed to by a red arrow labeled 'Class label (what we want to predict)'. A red arrow points to the first four rows, labeled 'Instances a.k.a. examples'. The table content is as follows:

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
...

**Instances
a.k.a.
examples**

**Attributes
a.k.a. features**

**Class label
(what we want to
predict)**

Association learning

- Can be applied if no class is specified and any kind of structure is considered “interesting”
- Difference to classification learning:
 - Can predict any attribute’s value, not just the class, and more than one attribute’s value at a time
 - Hence: far more association rules than classification rules
 - Thus: constraints are necessary, such as minimum coverage and minimum accuracy

Association rule mining example: Marketing and sales

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



Diaper



Beer

Clustering

- Finding groups of items that are similar
- Clustering is *unsupervised*
 - The class of an example is not known
- Success often measured subjectively

	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	Iris virginica
102	5.8	2.7	5.1	1.9	Iris virginica
...					

Numeric prediction

- Variant of classification learning where “class” is **numeric** (also called “**regression**”)
- Learning is supervised
 - Scheme is being provided with target value
- Measure success on test data

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
...

Think-pair-share: types of concept

Consider the following data science tasks.

- *Are the concepts to be learned best formulated as:* classification, association rule mining, clustering, or numeric prediction?
- *What data would you collect* to perform these tasks?
Suppose you have unlimited resources.
- Task 1: Predicting the winner of the 2024 presidential election
- Task 2: Choosing which coupons to offer at the checkout counter to customers of a pharmacy

What's in an example?

- **Instance:** specific type of example
 - Things to be classified, associated, or clustered
 - Individual, independent example of target concept
 - Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances/dataset
 - Represented as a single relation/flat file
- Rather restricted form of input
 - No relationships between objects
- Most common form in practical data mining

Types of Data Sets

- Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

- Graph and network

- World Wide Web
- Social or information networks
- Molecular Structures

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

- Spatial, image and multimedia:

- Spatial data: maps
- Image data:
- Video data:

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Data Matrix

- Data matrix

- n instances with p attributes

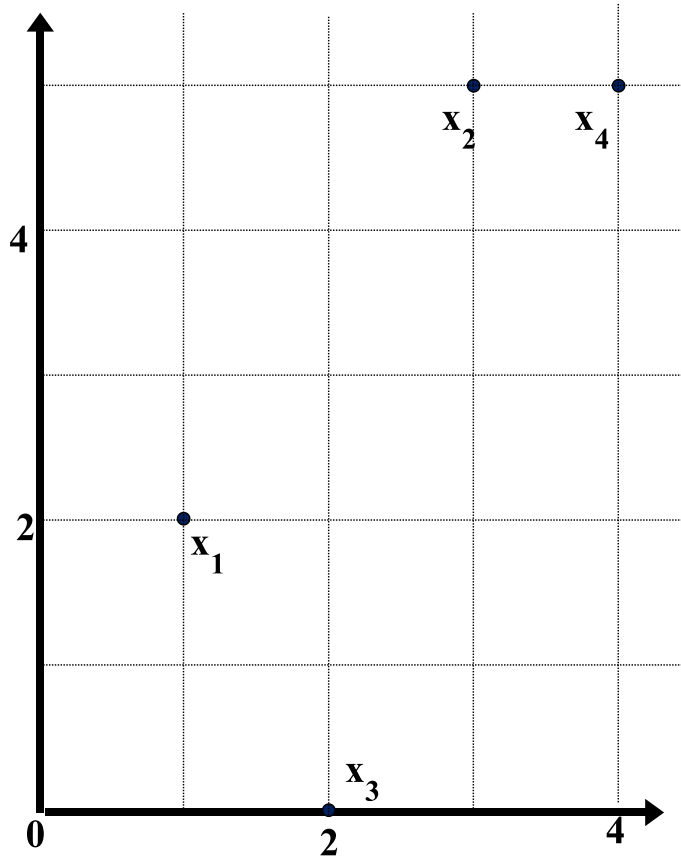
Attributes

Instances

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Need to convert data set to this form to apply standard machine learning algorithms!

Example: Data Matrix



Data Matrix

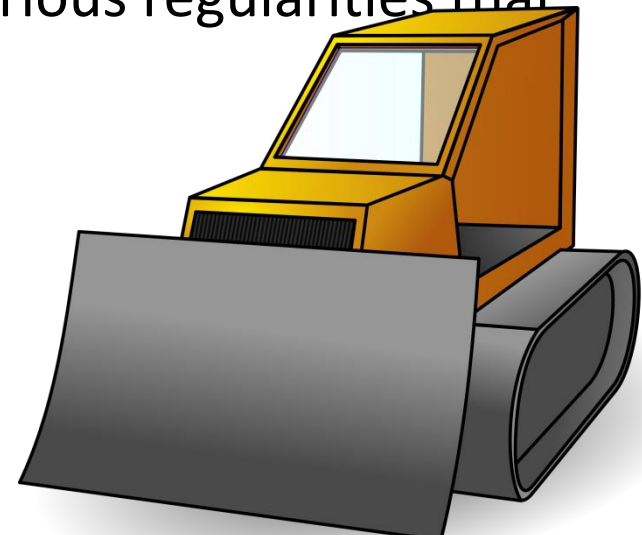
point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

Attributes

- **Attribute** (or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
 - *E.g., customer_ID, name, address*
- Types:
 - Nominal
 - Binary
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Generating a flat file data matrix

- Process of flattening called “denormalization”
 - Several relations are joined together to make one
- Possible with any finite set of finite relations
- Problematic: relationships without a pre-specified number of objects
 - Example: concept of *nuclear-family*
- Note that denormalization may produce spurious regularities that reflect the structure of the database
 - Example: “supplier” predicts “supplier address”



What's in an attribute?

- Each instance is described by a fixed predefined set of features, its “attributes”
- But: number of attributes may vary in practice
 - Possible solution: “irrelevant value” flag
- Related problem: existence of an attribute may depend of value of another one
- Possible attribute types (“levels of measurement”):
 - *Nominal*, *ordinal*, *interval* and *ratio*

Nominal levels of measurement

- Values are distinct symbols
 - Values themselves serve only as labels or names
 - *Nominal* comes from the Latin word for name
- Example: attribute “outlook” from weather data
 - Values: “sunny”, “overcast”, and “rainy”
- No relation is implied among nominal values (no ordering or distance measure)
- Only equality tests can be performed

Ordinal levels of measurement

- Impose order on values
- But: no distance between values defined
- Example:
attribute “temperature” in weather data
 - Values: “hot” > “mild” > “cool”
- Note: addition and subtraction don’t make sense
- Example rule:
temperature < hot \Rightarrow play = yes
- Distinction between nominal and ordinal not always clear (e.g., attribute “outlook”)

Interval quantities

- Interval quantities are not only ordered but measured in fixed and equal units
- Example 1: attribute “temperature” expressed in degrees Fahrenheit
- Example 2: attribute “year”
- Difference of two values makes sense
- Sum or product doesn’t make sense
 - Zero point is not defined!

Ratio quantities

- Ratio quantities are ones for which the measurement scheme defines a zero point
- Example: attribute “distance”
 - Distance between an object and itself is zero
- Ratio quantities are treated as real numbers
 - All mathematical operations are allowed
- But: is there an “inherently” defined zero point?
 - Answer depends on scientific knowledge (e.g., Fahrenheit knew no lower limit to temperature)

Attribute types used in practice

- Many data mining schemes accommodate just two levels of measurement: nominal and ordinal
- Others deal exclusively with ratio quantities
- Nominal attributes are also called “categorical”, “enumerated”, or “discrete”
 - But: “enumerated” and “discrete” imply order
- Special case: dichotomy (“Boolean” attribute)
- Ordinal attributes are sometimes coded as “numeric” or “continuous”
 - But: “continuous” implies mathematical continuity

How many of the following are ratio or interval attributes? Size of drinks available at a fast-food restaurant: {small, medium, large}, Price of drinks {\$1, \$2.50, \$3, \$4.25}, Percentage grades in a course: {0%, 1%, ..., 100%}, Letter grades in a course: {A, A-, B+, B, ..., F}, Customer satisfaction categories in a survey: {0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied}.

How many of the following are ratio or interval attributes? Size of drinks available at a fast-food restaurant: {small, medium, large}, Price of drinks {\$1, \$2.50, \$3, \$4.25}, Percentage grades in a course: {0%, 1%, ..., 100%}, Letter grades in a course: {A, A-, B+, B, ..., F}, Customer satisfaction categories in a survey: {0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied}.

How many of the following are ratio-scaled attributes: Temperature in degrees Celsius, Temperature in degrees Kelvin, Calendar dates in years AD, Time measured from the Big Bang, Number of words in a document

1
2
3
4
5

How many of the following are ratio-scaled attributes: Temperature in degrees Celsius, Temperature in degrees Kelvin, Calendar dates in years AD, Time measured from the Big Bang, Number of words in a document

1
2
3
4
5

Preparing the input

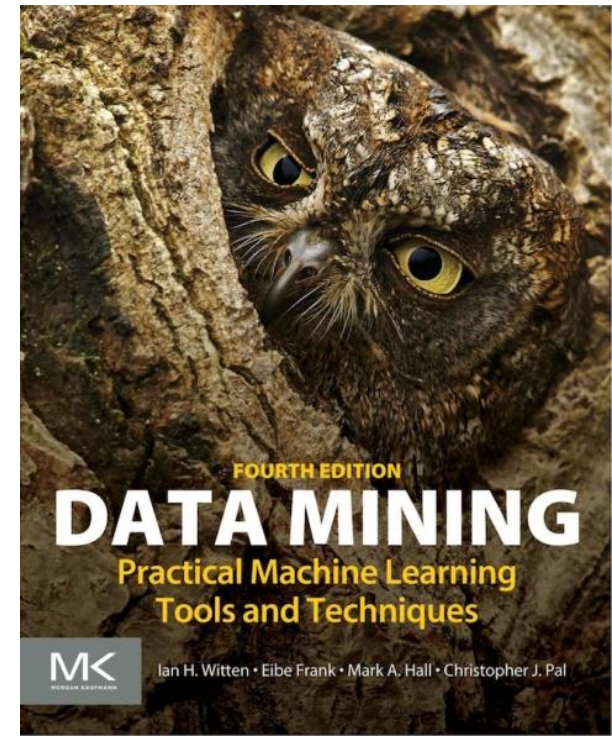
- Denormalization is not the only issue when data is prepared for learning
- Problem: different data sources (e.g., sales department, customer billing department, ...)
 - Differences: styles of record keeping, coding conventions, time periods, data aggregation, primary keys, types of errors
 - Data must be assembled, integrated, cleaned up
 - “Data warehouse”: consistent point of access
- External data may be required (“overlay data”)
- Critical: type and level of data aggregation

Weka

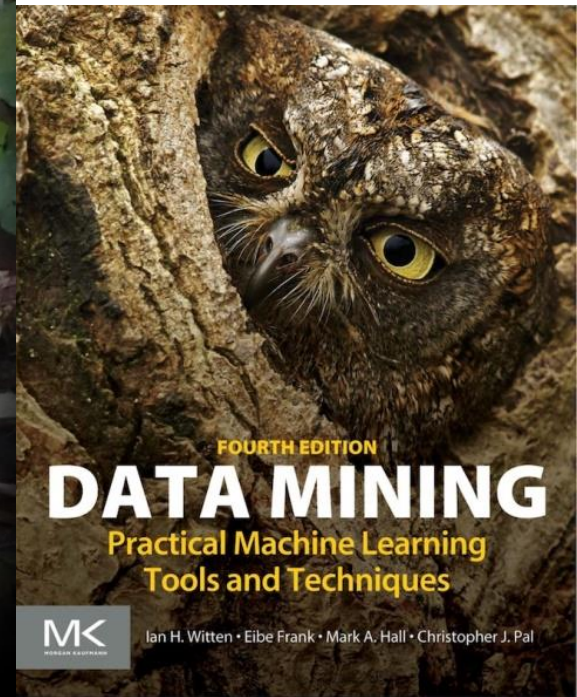
- Weka (Waikato Environment for Knowledge Analysis) is the data mining software we will focus on in this course.

Link: <https://www.cs.waikato.ac.nz/ml/weka/>

- The file format used by Weka is called ARFF
- Weka has a graphical interface and a java API



Weka – the native New Zealand bird!



The ARFF data format

```
%  
% ARFF file for weather data with some numeric features  
%  
@relation weather  
  
@attribute outlook {sunny, overcast, rainy}  
@attribute temperature numeric  
@attribute humidity numeric  
@attribute windy {true, false}  
@attribute play? {yes, no}  
  
@data  
sunny, 85, 85, false, no  
sunny, 80, 90, true, no  
overcast, 83, 86, false, yes  
...
```

Additional attribute types

- ARFF data format also supports *string* attributes:

```
@attribute description string
```

- Similar to nominal attributes but list of values is not pre-specified

- Additionally, it supports *date* attributes:

```
@attribute today date
```

- Uses the ISO-8601 combined date and time format *yyyy-MM-dd-THH:mm:ss*

Sparse data

- In some applications most attribute values are zero and storage requirements can be reduced
 - E.g.: word counts in a text categorization problem
- ARFF supports sparse data storage

```
0, 26, 0, 0, 0, 0, 63, 0, 0, 0, "class A"  
0, 0, 0, 42, 0, 0, 0, 0, 0, 0, "class B"
```

```
{1 26, 6 63, 10 "class A"}  
{3 42, 10 "class B"}
```

- This also works for nominal attributes (where the first value of the attribute corresponds to “zero”)
- Some learning algorithms work very efficiently with sparse data

Attribute types

- Interpretation of attribute types in an ARFF file depends on the learning scheme that is applied
 - Numeric attributes are interpreted as
 - ordinal scales if less-than and greater-than are used
 - ratio scales if distance calculations are performed (normalization/standardization may be required)
 - Note also that some instance-based schemes define a distance between nominal values (0 if values are equal, 1 otherwise)
- Background knowledge may be required for correct interpretation of data
 - E.g., consider integers in some given data file: nominal, ordinal, or ratio scale?

Nominal vs. ordinal

- Attribute “age” nominal

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

- Attribute “age” ordinal
(e.g. “young” < “pre-presbyopic” < “presbyopic”)

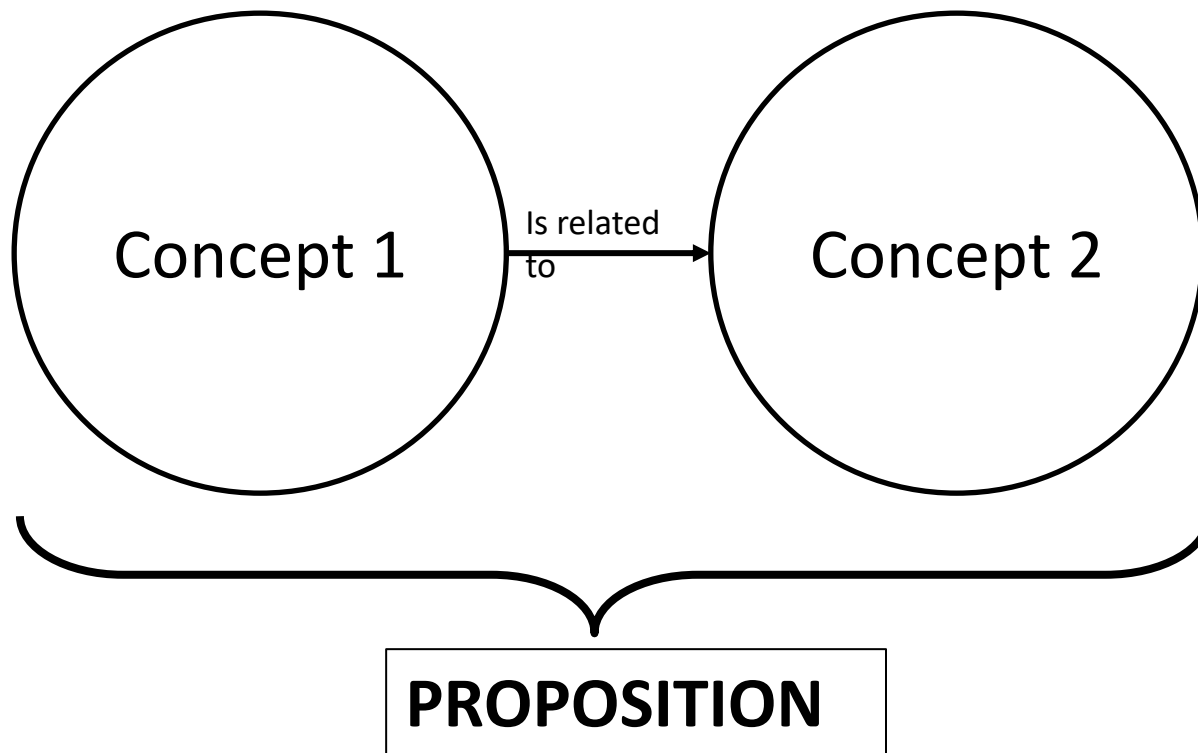
If age \leq pre-presbyopic and astigmatic = no
and tear production rate = normal
then recommendation = soft

Demo : ARFF file for Iris

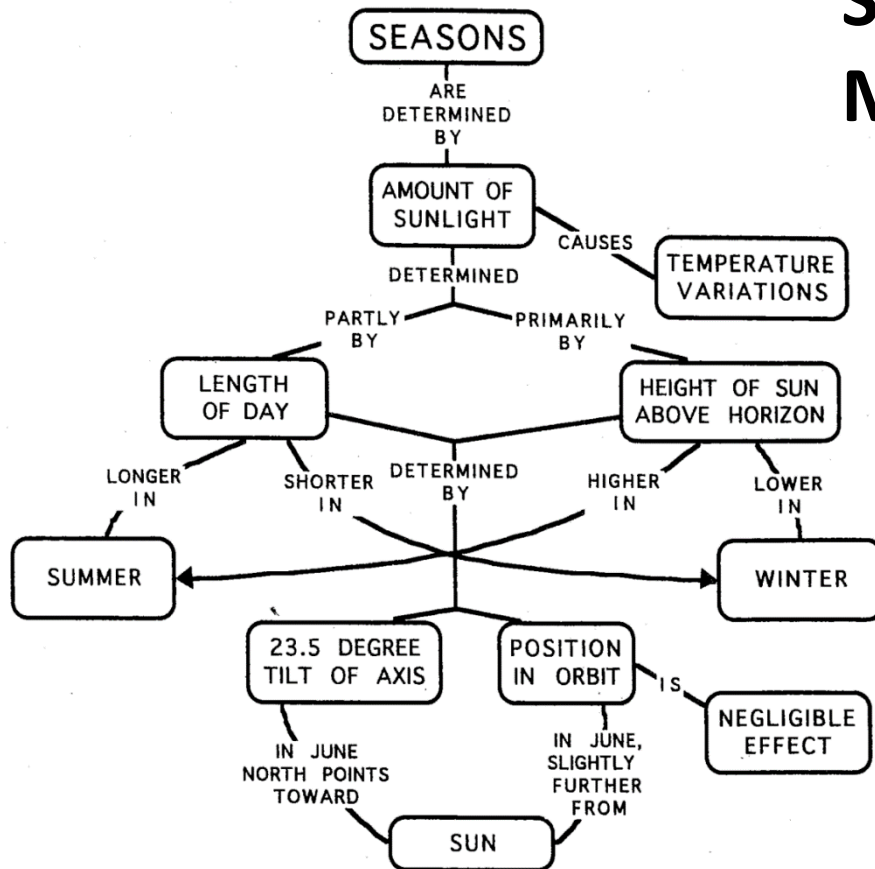
- Open ARFF file in text editor
 - look at comments, header, data.
- Open the file in WEKA Explorer
 - Identify attributes, attribute types
 - Number of instances

concept map /'kähn,sept 'map/ (noun)

a graphic representation of knowledge
that shows how different concepts are related
and how they combine to form propositions



Sample Concept Map



General

Respond to focal question, e.g.:
What determines the seasons?

Specific

Exercise: Concept Map

- In groups of 4-5, on pen and paper please **create a concept map** for the concepts covered so far:

data mining, supervised learning, instance, numeric attribute, class label, ordinal attribute, clustering, concept numeric prediction, data, WEKA, attribute, ratio attribute, concept description, machine learning, regression, ARFF format, classification, relational data, unsupervised learning, data matrix, nominal attribute, interval attribute, association learning, denormalization

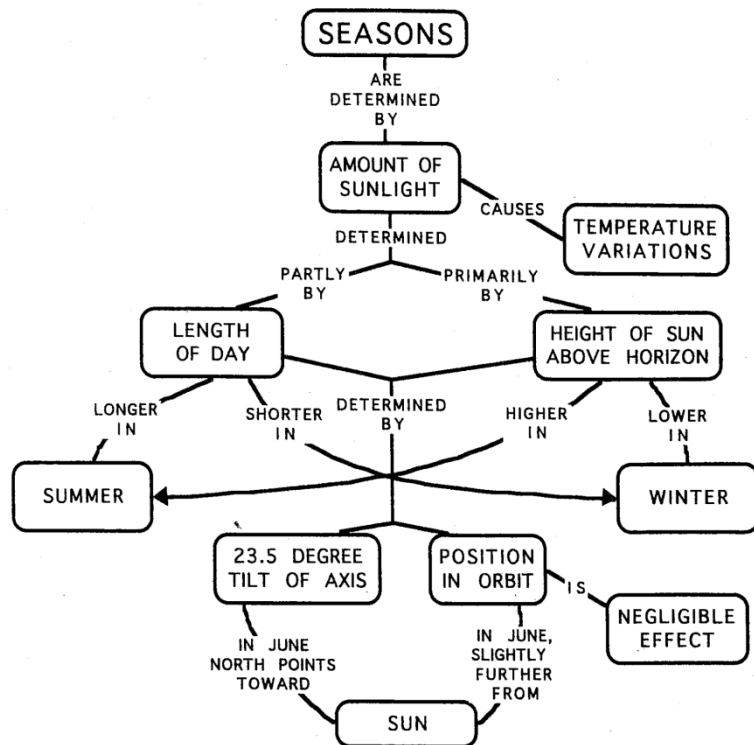


FIG. 7.4. A concept map showing the key ideas needed to understand why we have seasons. Many people fail to understand the effect of the inclination of the earth on its axis as the primary cause for summer and winter in both hemispheres.