Data Preparation and Exploration

Lecture notes by Chandadevi Giri

HÖGSKOLAN I BORÅS

Part 1- Data Preparation

Learning goals

- Understand the importance of data preparation
- Different activities that are involved in data preparation
- To understand data quality issues and why it is need to be addressed
- Feature selections
- Feature transformation
- Dimensionality reduction
- Domain knowledge in data preparation

Data Terminology

Variables

Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Terms to Describe 'Variable'

- Attributes
- Feature
- Field
- Column

Other Names for 'Sample'

- Row
- Instance
- Record
- Observation

Data Types

Most common data types

Numerical Variables –

- Values are numbers (163.92, -0.4902, 2)
- Also called ‘quantitative’

- Height
- Score on an exam
- Change in stock price

Categorical Variables

- Values are labels, names, or categories
- Also called ‘qualitative’ or ‘nominal’

- Gender (M,F)
- Marital status (Single, Married)
- Type of customer (Active, Inactive)
- Product categories (Shirt, Tshirt etc)
- Color of an item (Red, Blue, Green)

Variables

- Feature
- Field
- Column

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Samples

- Instance
- Record
- Row
- Observation

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, χ^2 test
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests

Different Types of Data sets - example

Data that consists of a collection of records, each of which consists of a fixed set of attributes

Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

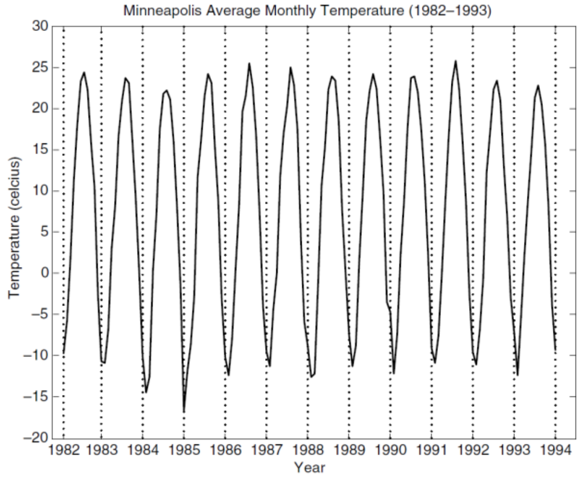
(a) Sequential transaction data.

Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Transaction data.

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Temperature time series.

Why Data Pre-Processing?

Garbage In —-> Garbage Out

Data Pre-Processing

- Data preparation can MAKE or BREAK a model's predictive ability
- How the predictors enter the model is important
- **Feature engineering** is how the predictors are encoded -> can have significant impact on model performance.
- Which feature engineering methods are the best?
    - It depends!

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Data compression
- **Data transformation**
  - Normalization

Preparing Data- The aim is to create data for analysis – consist of cleaning and formatting

Real-world data is messy!

Data preparation is very important for meaningful analysis.

Domain knowledge is required for addressing data quality issues effectively

Poor Data Quality → Poor Analysis Results

Preparing Data

Goal: Create data for analysis

Clean

Data quality issues
- Missing values
- Duplicate data
- Noise
- Outliers

Addressing Data Quality Issues

Some techniques:
- Remove data with missing values
- Merge duplicate records
- Generate best estimate for invalid values

Format
- Select features to use
- - Transform data

- Feature selection
  - Combing features
  - Adding/Removing features
- Feature transformation
  - Scaling
  - Dimensionality reduction

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

Missing values

Name	Age	Income
Angela	34	80
Sidney	--	56
Ratan	10	--
Kiril	68	--
Zhou	45	120

Removing Missing Data

Name	Age	Income
Angela	34	80
~~***Sidney***~~	~~--~~	~~***56***~~
~~***Ratan***~~	~~***10***~~	~~--~~
~~***Kiril***~~	~~***68***~~	~~--~~
Zhou	45	120

Imputing Missing Data

- Replace missing values with something reasonable

Name	Age	Income
Angela	34	80
Sidney	***50***	***56***
Ratan	***10***	***50***
Kiril	***68***	***50***
Zhou	45	120

Ways to Impute Missing Data

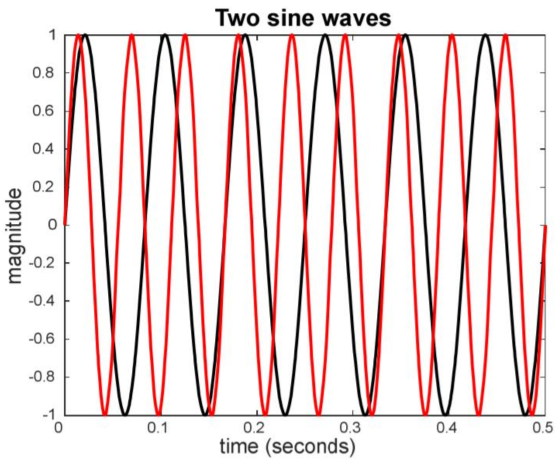
Replace missing value with

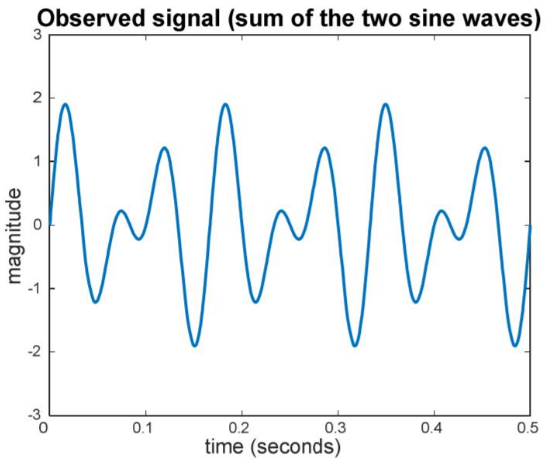
- Mean
- Median
- Most frequent
- Sensible value based on application

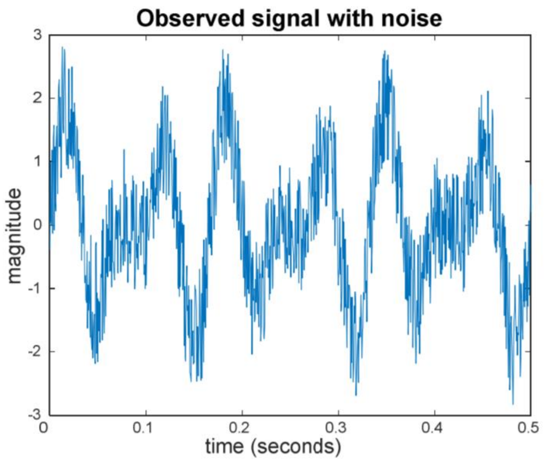
Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

For attributes, noise refers to modification of original values
– Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen
– The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise
- The magnitude and shape of the original signal is distorted







How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data quality issues

- Duplicate data
- Data inconsistency

Data set may include data objects that are duplicates, or almost duplicates of one another – Major issue when merging data from heterogeneous sources

Examples:

- Same person with multiple email addresses

id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

18

Correlation Analysis (Numeric Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

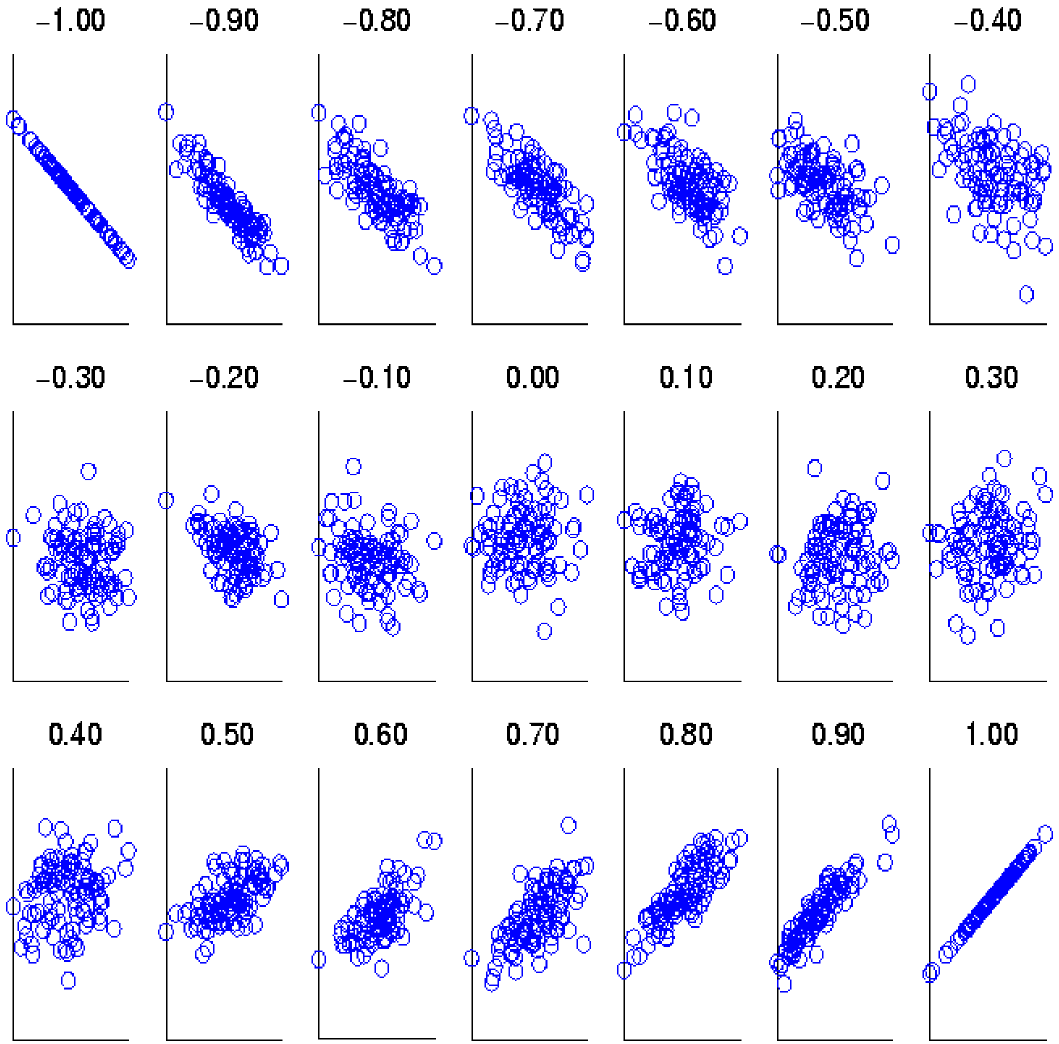
$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of data points,

$\bar{A}$, $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{AB} < 0$: negatively correlated

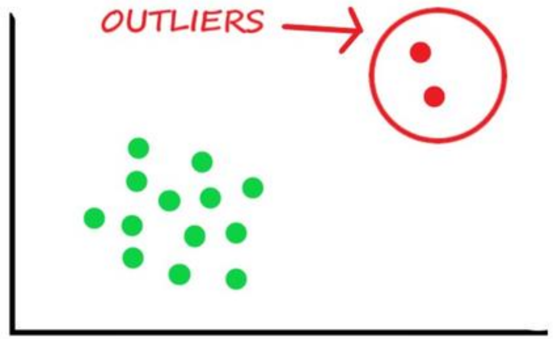
Visually Evaluating Correlation



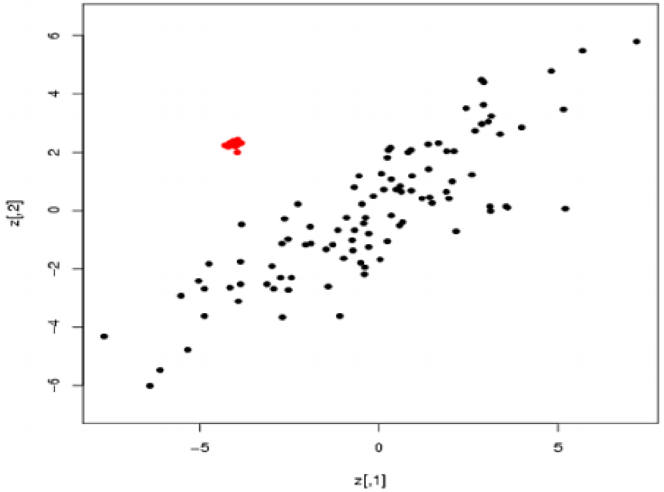
Scatter plots showing the similarity from –1 to 1.

Outliers

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Case 1: Outliers are noise that interferes with data analysis



Case 2: Outliers are the goal of our analysis

- Credit card fraud
- Intrusion detection

Aggregation

Combining two or more attributes (or objects) into a single attribute (or object)

Purpose

- Data reduction - reduce the number of attributes or objects
- Change of scale
  - Cities aggregated into regions, states, countries, etc.
  - Days aggregated into weeks, months, or years
- More “stable” data - aggregated data tends to have less variability

Feature engineering - how the predictors are encoded
Feature selection - the model will only include predictors that help maximize accuracy.

Adding/Combining Features

New features derived from existing features

Name	State
Angela	AK
Sidney	CA
Ratan	WA
Kiril	OR
Zhou	CA

Name	State	*In-State*
Angela	AK	***F***
Sidney	CA	***T***
Ratan	WA	***F***
Kiril	OR	***F***
Zhou	CA	***T***

Removing Features

Features that are very correlated

- Features with a lot of missing values
- Irrelevant features: ID, row number, etc.

Recoding Features

Examples

- Discretization: re-format continuous feature as discrete
- Customer's age => {teenager, young adult, adult, senior}

Feature Selection Summary

- Goal: Select smallest set of features that best captures data for application.
- Domain knowledge is important
- aka 'feature engineering'

Two ways to scale your Data

- Normalization: putting each observation on a relative scale between the values of 0 and 1

  Value of Observation / Sum of all observations in variable

- Standardization: Rescaling data so that it has zero mean and unit variance

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$

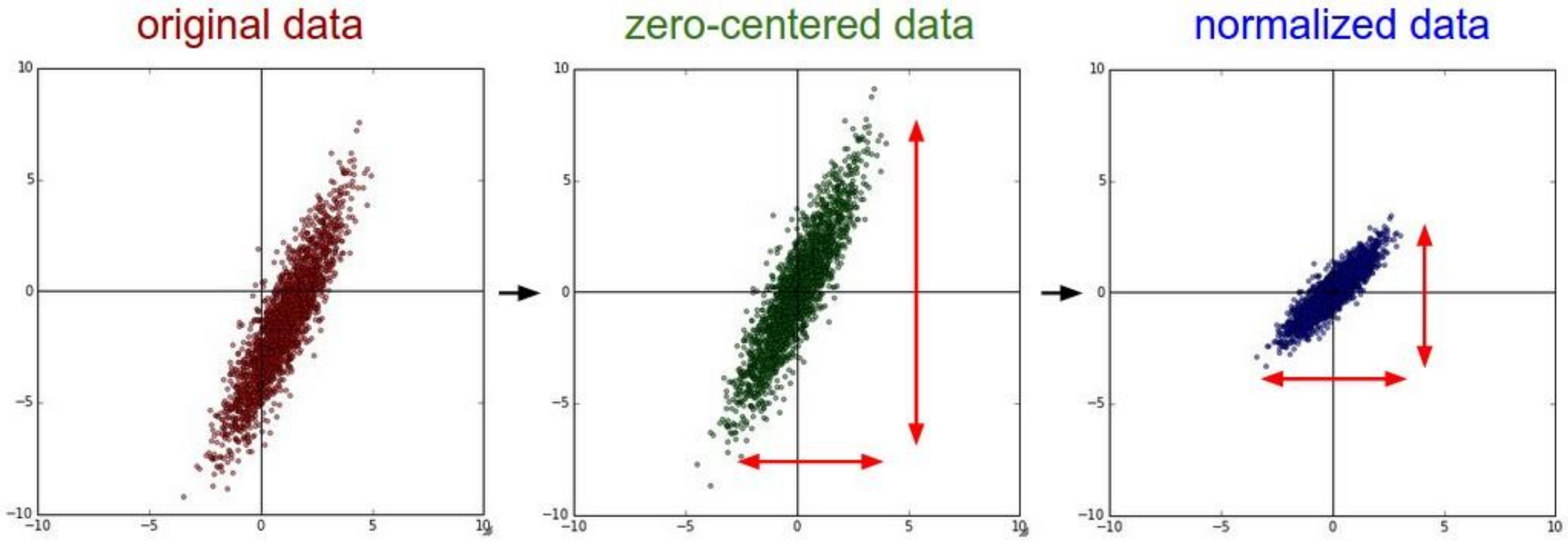
- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

Centering and Scaling

- The average predictor value is subtracted from all the values.
- Centering: the predictor has a zero mean
- Scaling: each value of the predictor variable is divided by its standard deviation.



Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization

Part 2 - Data Exploration

How to get to know data?

- Calculate statistics
- Use aggregations
- Use visualizations
- Use predictive and descriptive models to provide insights

What is data exploration?

A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

- Related to the area of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - A nice online introduction can be found in Chapter 1 of the NIST Engineering Statistics Handbook

  http://www.itl.nist.gov/div898/handbook/index.htm

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques
  - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory

- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Iris Sample Data Set

- Many of the exploratory data techniques are illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
  - Summarized properties include frequency, location and spread
    - Examples: location - mean
spread - standard deviation
  - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the
  data set
  - For example, given the attribute ‘gender’ and a representative population of people, the gender ‘female’ occurs about 50% of the time.
- The mode of a an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the *p*th percentile is a value x_p of x such that *p*% of the observed values of x are less than x_p.

- For instance, the 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread: Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

- However, this is also sensitive to outliers, so that other measures are often used.
  - Average Absolute Deviation
  - Mean Absolute Deviation

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

$$\text{MAD}(x) = median\Big(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\Big)$$

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Aggregations

- An effective way of investigating relationships among attributes
- How?
  - Databases (SQL): GROUP BY with aggregating functions (SUM, COUNT etc)
  - KNIME: GroupBy or Pivoting nodes
    - 02→02→06-09
    - Combine with e.g. row filtering or binning to look at specific subsets or subgroups
  - Python: pandas
    - https://jakevdp.github.io/PythonDataScienceHandbook/03.08-aggregation-and-grouping.html

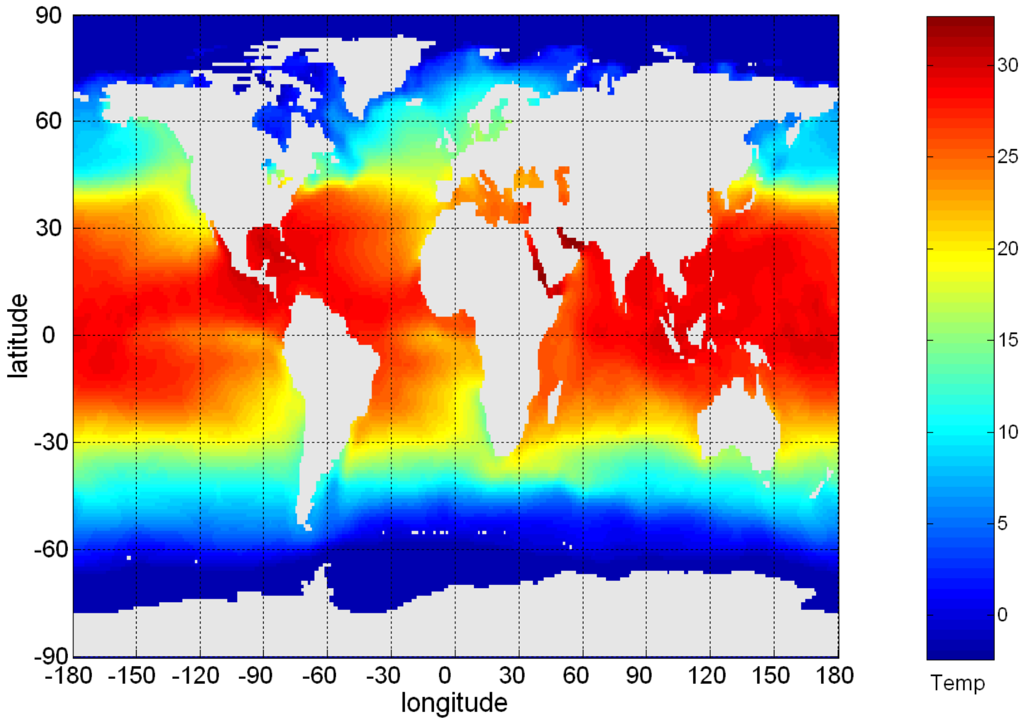
Visualization

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.

- Visualization of data is one of the most powerful and appealing techniques for data exploration.
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure



Representation

- Is the mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Arrangement

- Is the placement of visual elements within a display
- Can make a large difference in how easy it is to understand the data
- Example:

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

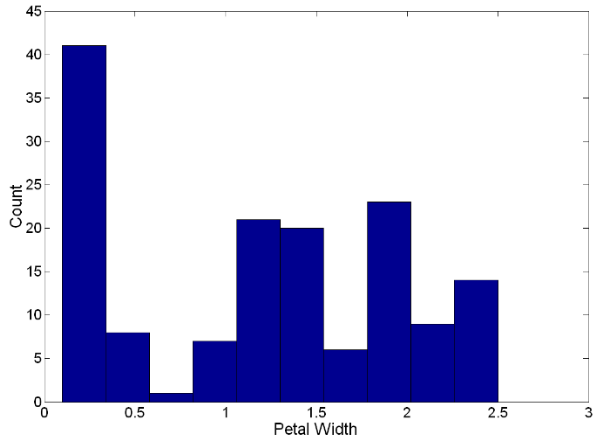
	6	1	3	2	5	4
4						
2						
6						
8						
5						
3						
9						
1						
7						

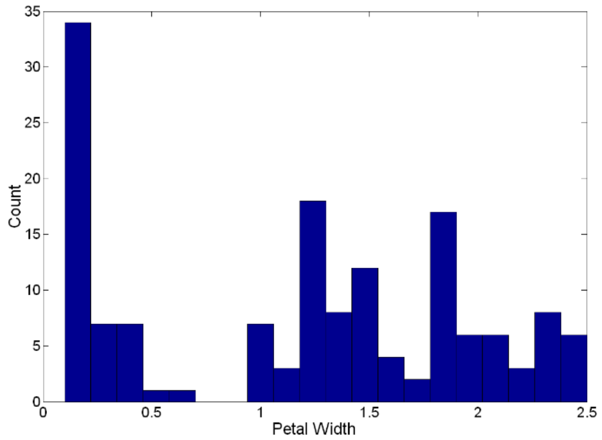
Selection

- Is the elimination or the de-emphasis of certain objects and attributes
- Selection may involve choosing a subset of attributes
  - Dimensionality reduction is often used to reduce the number of dimensions to two or three
  - Alternatively, pairs of attributes can be considered
- Selection may also involve choosing a subset of objects
  - A region of the screen can only show a limited number of points
  - Can sample, but want to preserve points in sparse areas

Visualization Techniques: Histograms

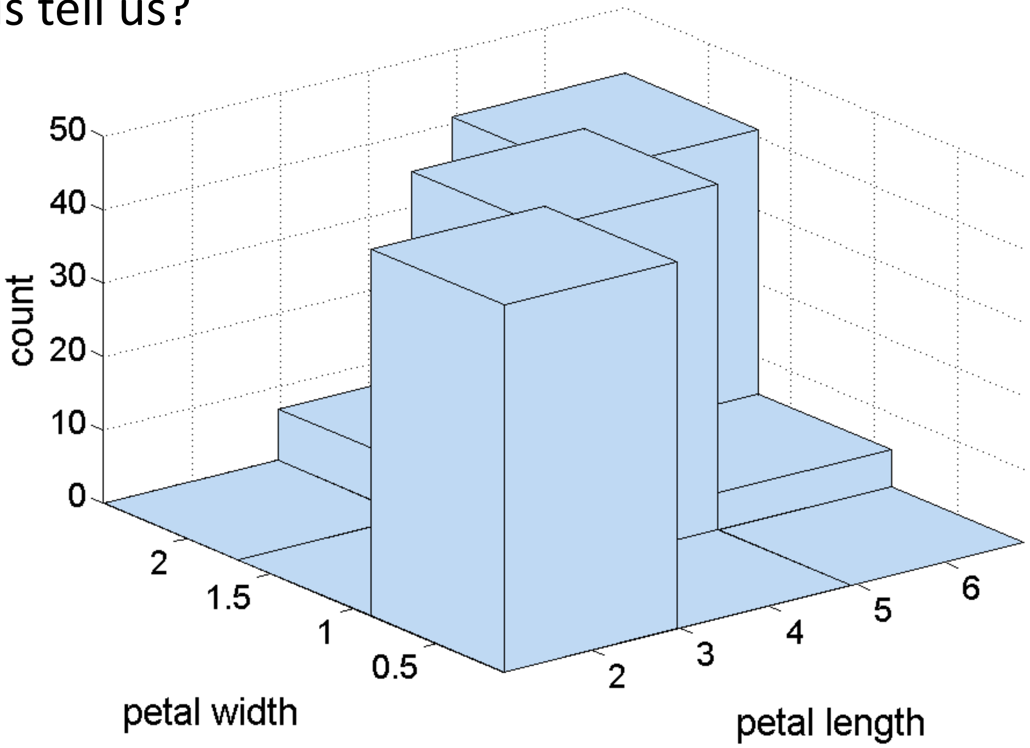
- Histogram
  - Usually shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





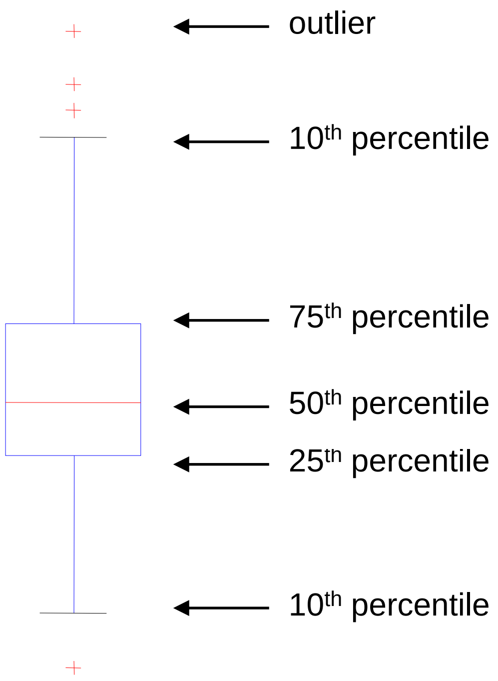
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



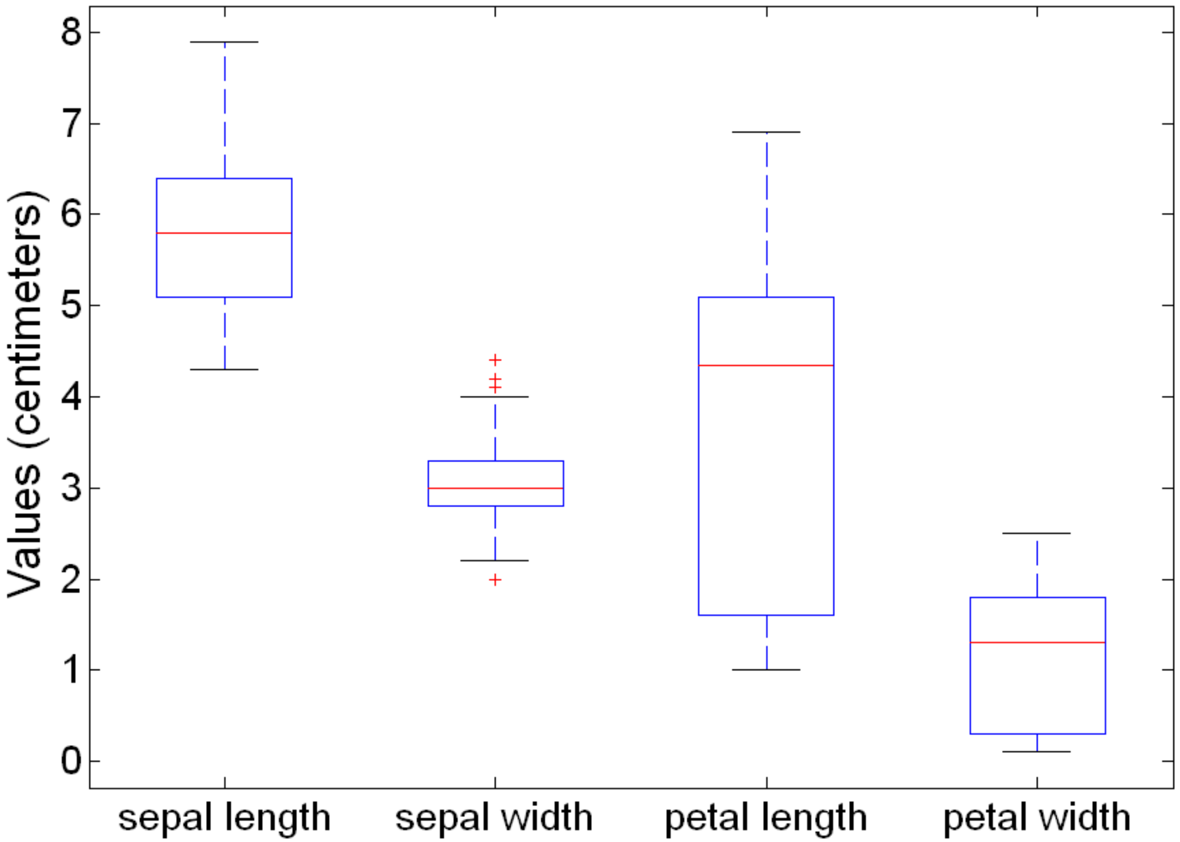
Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Another way of displaying the distribution of data
  - Following figure shows the basic part of a box plot



Example of Box Plots

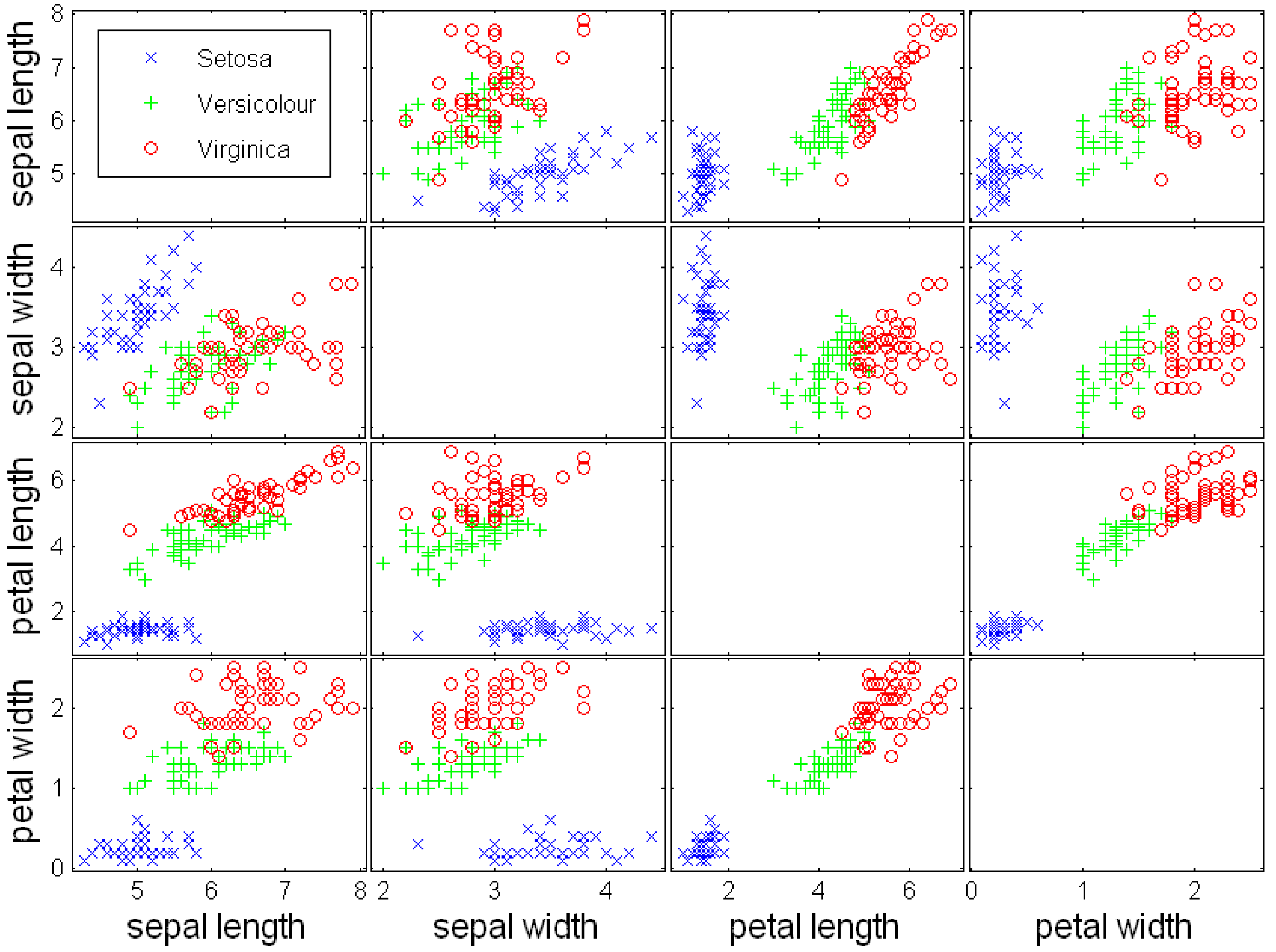
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- Scatter plots
    - Attribute values determine the position
    - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
    - Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
    - It is useful to have arrays of scatter plots
        - Can compactly summarize the relationships of several pairs of attributes
        - See example on the next slide

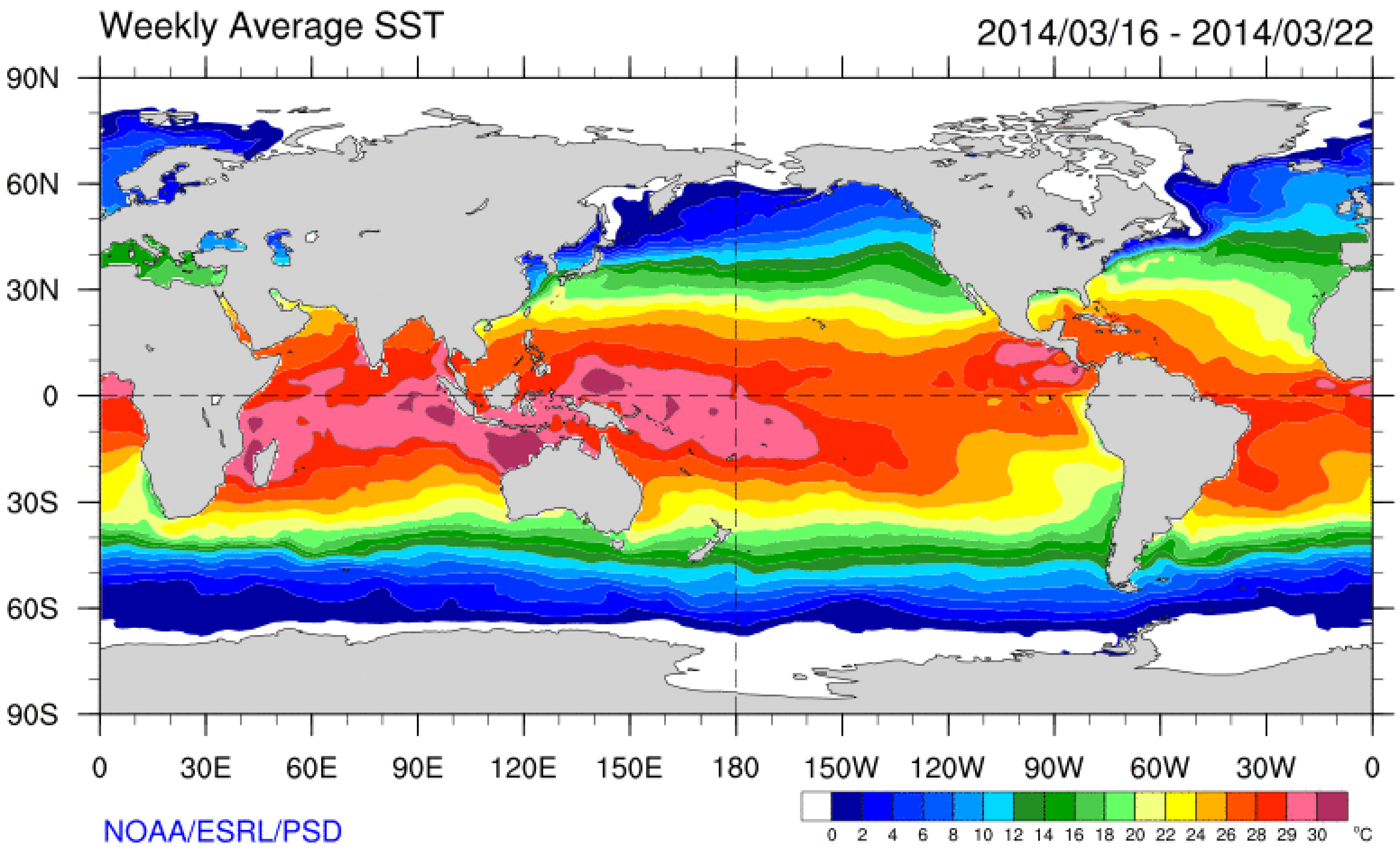
Scatter Matrix of Iris Attributes



Visualization Techniques: Contour Plots

- Contour plots
  - Useful when a continuous attribute is measured on a spatial grid
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values
  - The most common example is contour maps of elevation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide

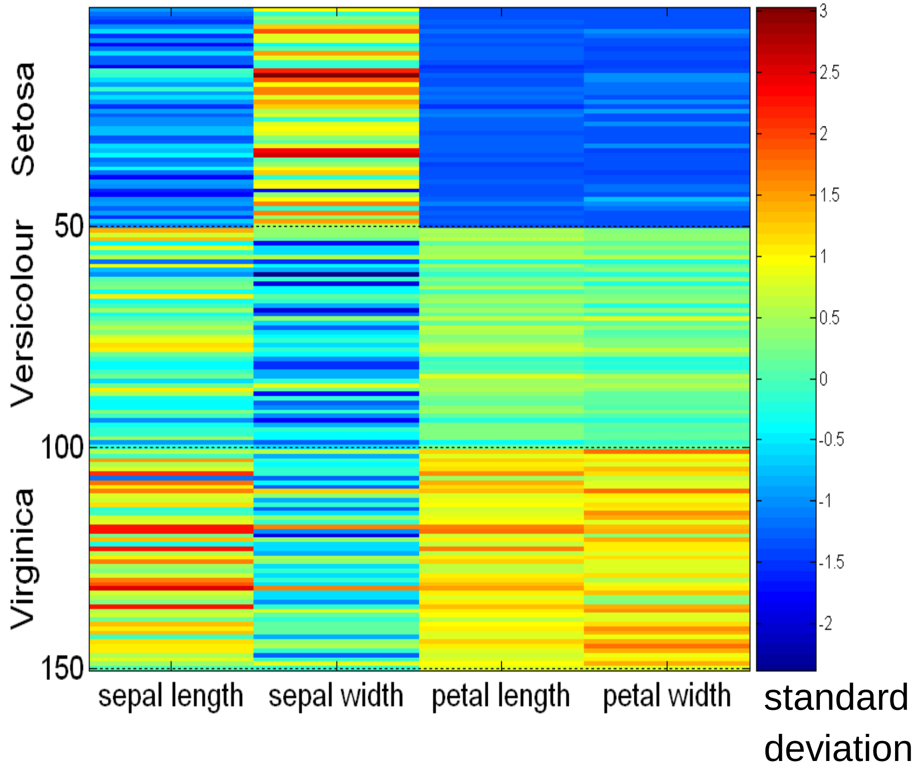
Contour Plot Example: SST March, 2014



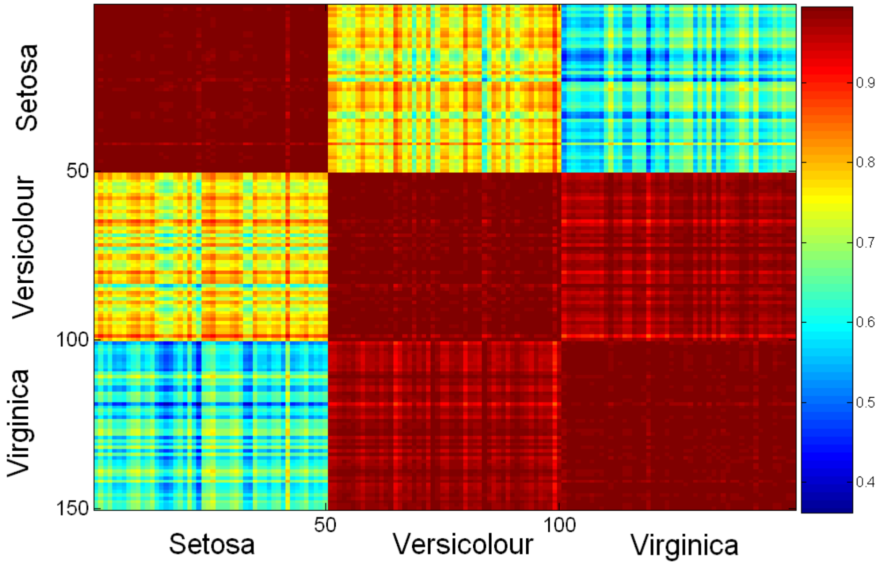
Visualization Techniques: Matrix Plots

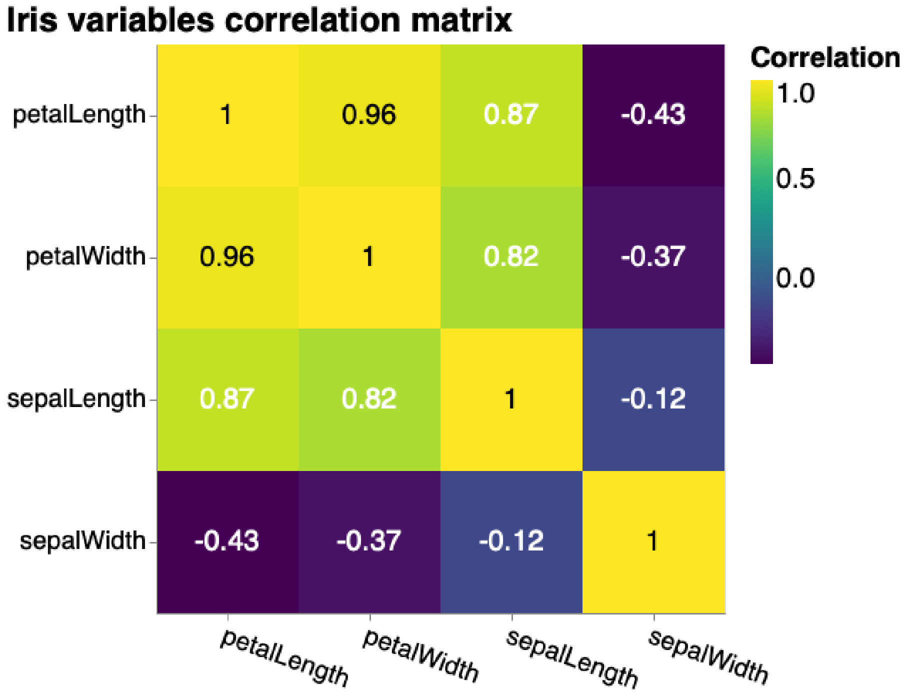
- Matrix plots
  - Can plot the data matrix
  - This can be useful when objects are sorted according to class
  - Typically, the attributes are normalized to prevent one attribute from dominating the plot
  - Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects
  - Examples of matrix plots are presented on the next two slides

Visualization of the Iris Data Matrix



Visualization of the Iris Correlation Matrix

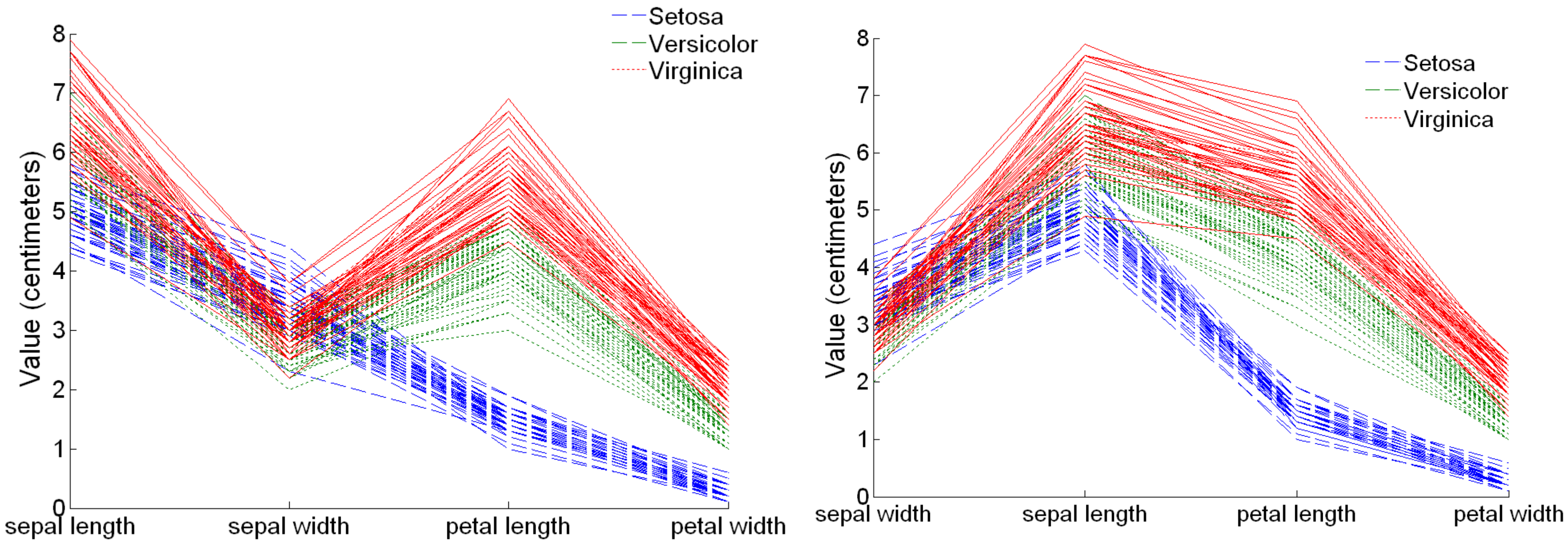




Visualization Techniques: Parallel Coordinates

- Parallel Coordinates
  - Used to plot the attribute values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some attributes
  - Ordering of attributes is important in seeing such groupings

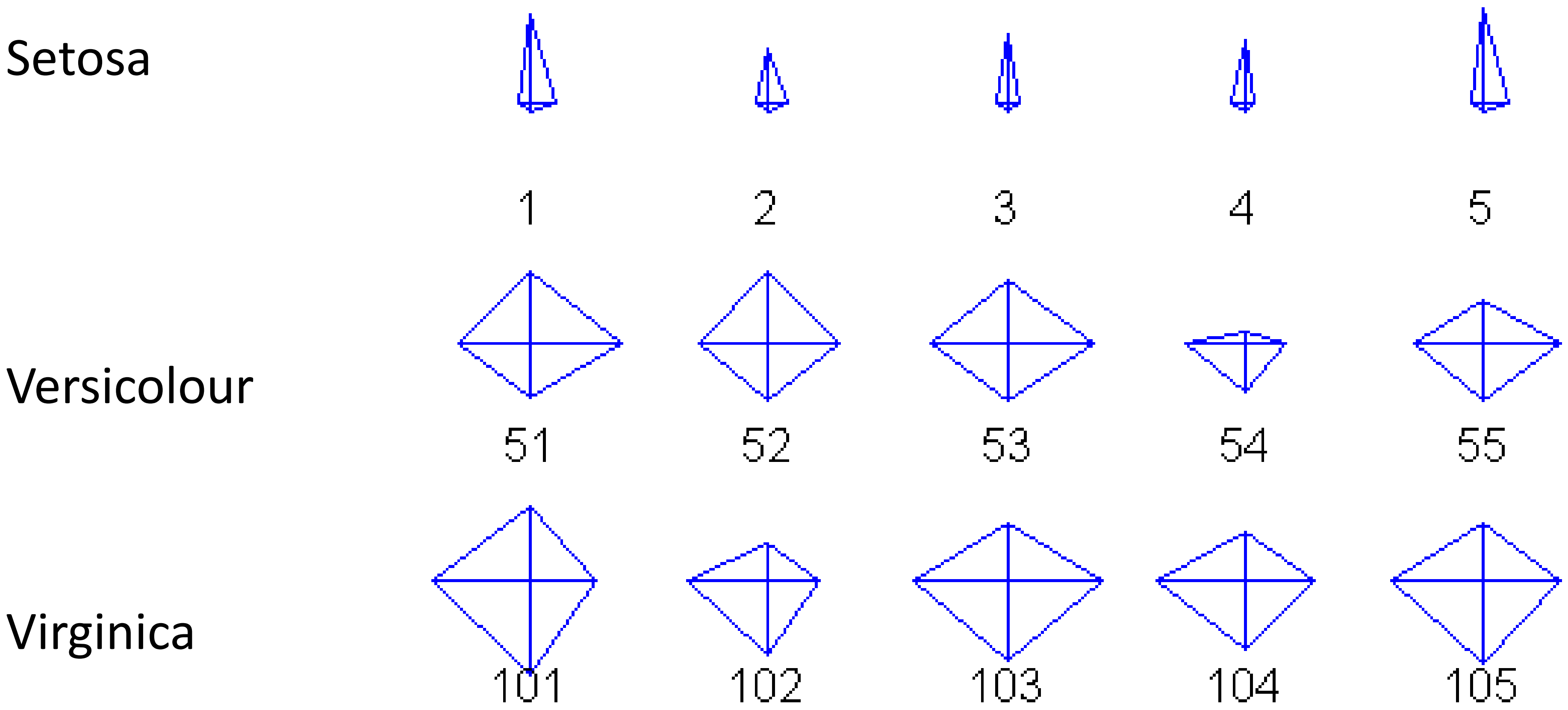
Parallel Coordinates Plots for Iris Data



Other Visualization Techniques

- Star Plots (Radar Plots in KNIME)
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon

Star (Radar) Plots for Iris Data



Interpretable Machine Learning

- Learning from machine learning can be an effective way of exploring data
- Some techniques produce more interpretable models than others
  - Decision trees – Both the entire model and individual paths
  - Linear/Logistic Regression - Slopes
  - Bagging Ensembles - Attribute Statistics

Questions

Magnus.bengtsson@hb.se