

Data Exploration and Visualization

Data Science for Mobility



Readings

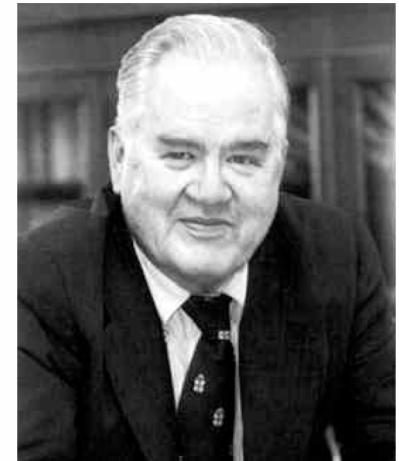
- “Data Mining and Analysis” – Chapter 2 & 3
- t-Distributed Stochastic Neighbor Embedding (t-SNE)
 - <http://lvdmaaten.github.io/tsne/>
 - <https://www.youtube.com/watch?v=RJVL80Gg3IA>
 - <http://alexanderfabisch.github.io/t-sne-in-scikit-learn.html>
 - <http://jmlr.csail.mit.edu/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>
 - <https://distill.pub/2016/misread-tsne/>
- Principal Component Analysis
 - “Data Mining and Analysis” – Chapter 7
 - http://sebastianraschka.com/Articles/2015_pca_in_3_steps.html
- Data Visualization
 - PhD Course by Daniele Loiacono

Before trying anything,
you should know your data!

What is Data Exploration?

- It is the preliminary exploration of the data aimed at identifying their most relevant characteristics
- What the key motivations?
 - Help to select the right tool for preprocessing and data mining
 - Exploit humans' abilities to recognize patterns not captured by automatic tools
- Related to Exploratory Data Analysis (EDA), created by statistician John Tukey

- “An approach of analyzing data to summarize their main characteristics without using a statistical model or having formulated a prior hypothesis.”
- “Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.”
- Seminal book is “Exploratory Data Analysis” by Tukey
- Nice online introduction in Chapter I of the NIST Engineering Statistics Handbook
<http://www.itl.nist.gov/div898/handbook/index.htm>
- http://en.wikipedia.org/wiki/Exploratory_data_analysis



- Exploratory Data Analysis (as originally defined by Tukey), was mainly focused on
 - Visualization
 - Clustering and anomaly detection
(viewed as exploratory techniques)
 - In data mining, clustering and anomaly detection are major areas of interest, and not thought of as just exploratory
- In this section, we focus on data exploration using
 - Summary statistics
 - Visualization
- We will come back later to data exploration when discussing clustering

Summary Statistics

- Numbers that summarize properties of the data
- Summarized properties include location, mean, spread, skewness, standard deviation, mode, percentiles, etc.
- Most summary statistics can be calculated in a single pass

- 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 an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

- 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

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

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

- For continuous data, the notion of a percentile is very 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 $x_{50\%}$

- Trimean

- It is the weighted mean of the first, second and third quartile

$$TM = \frac{x_{25} + 2x_{50} + x_{75}}{4}$$

- Truncated Mean

- Discards data above and below a certain percentile
 - For example, below the 5th percentile and above the 95th percentile

- Interquartile Mean

$$X_{IQM} = \frac{2}{n} \sum_{i=0.25n+1}^{0.75n} x_i$$

- 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

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

- However, this is also sensitive to outliers, so that other measures are often used

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

$$\text{MAD}(x) = \text{median}\left(\{|x_1 - \bar{x}|, \dots, |x_m - \bar{x}|\}\right)$$

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

- Given two attributes it measure how strongly one attribute implies the other, based on the available data
- Correlation does not imply causation
- Numerical Variables**
 - For two numerical variables, we can compute the correlation coefficient, Pearson's product moment coefficient
- Ordinal Variables**
 - We can compute Spearman rank correlation coefficient
- Categorical Variables**
 - For two categorical variables, A and B, we can compute χ^2 (chi-square) statistic test which tests the hypothesis that A and B are independent
- Binary Variables**
 - Compute Point-biserial correlation

- They are data objects that do not comply with the general behavior or model of the data, that is, values that appear as anomalous
- Most data mining methods consider outliers noise or exceptions.
- Outliers may be detected using
 - Statistical tests that assume a distribution or probability model for the data
 - Distance measures where objects that are a substantial distance from any other cluster are considered outliers
 - Deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group

- Outliers are typically filtered out by eliminating the data points containing them
- Winsorizing
 - A 10% Winsorizing, consider the 5th and 95th percentiles
 - Set the values below the 5th percentile to the 5th percentile itself
 - Set the values above the 95th percentile to the 95th percentile itself
- Trimming
 - Eliminate the outlier data values
- Note that, in some applications, outliers are the focus on the analysis. Fraud detection is a typical example of this.

Normalization

- When attributes have vastly different scales (e.g., age vs income), it is necessary to normalize them
- Range normalization

$$x'_i = \frac{x_i - \min_i x_i}{\max_i x_i - \min_i x_i}$$

- Standard Score Normalization

$$x'_i = \frac{x_i - \mu}{\sigma}$$

Data Visualization

Why visualize data?

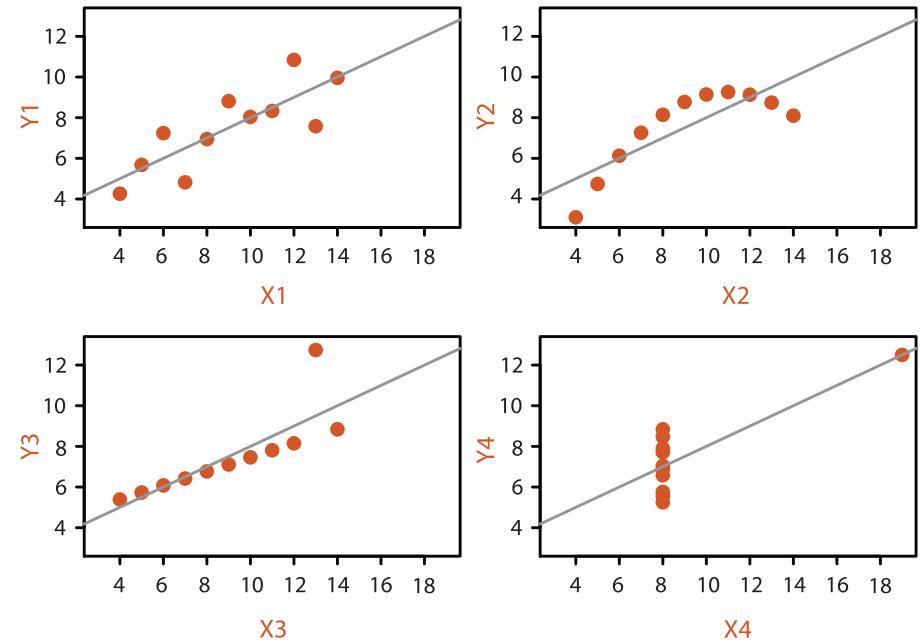
20

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

Why visualize data?

21

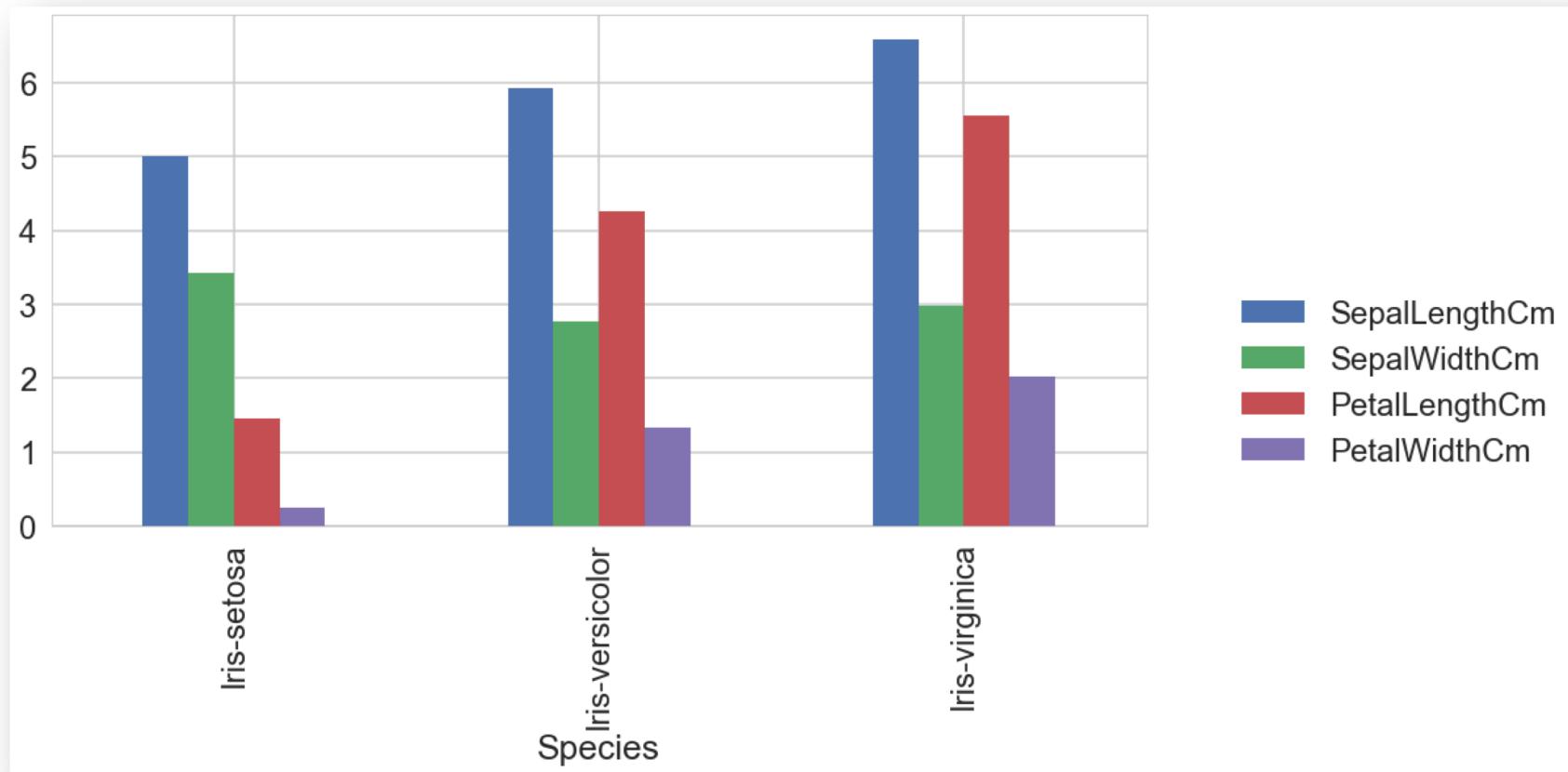
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



- 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
- Data visualization 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

Barplots

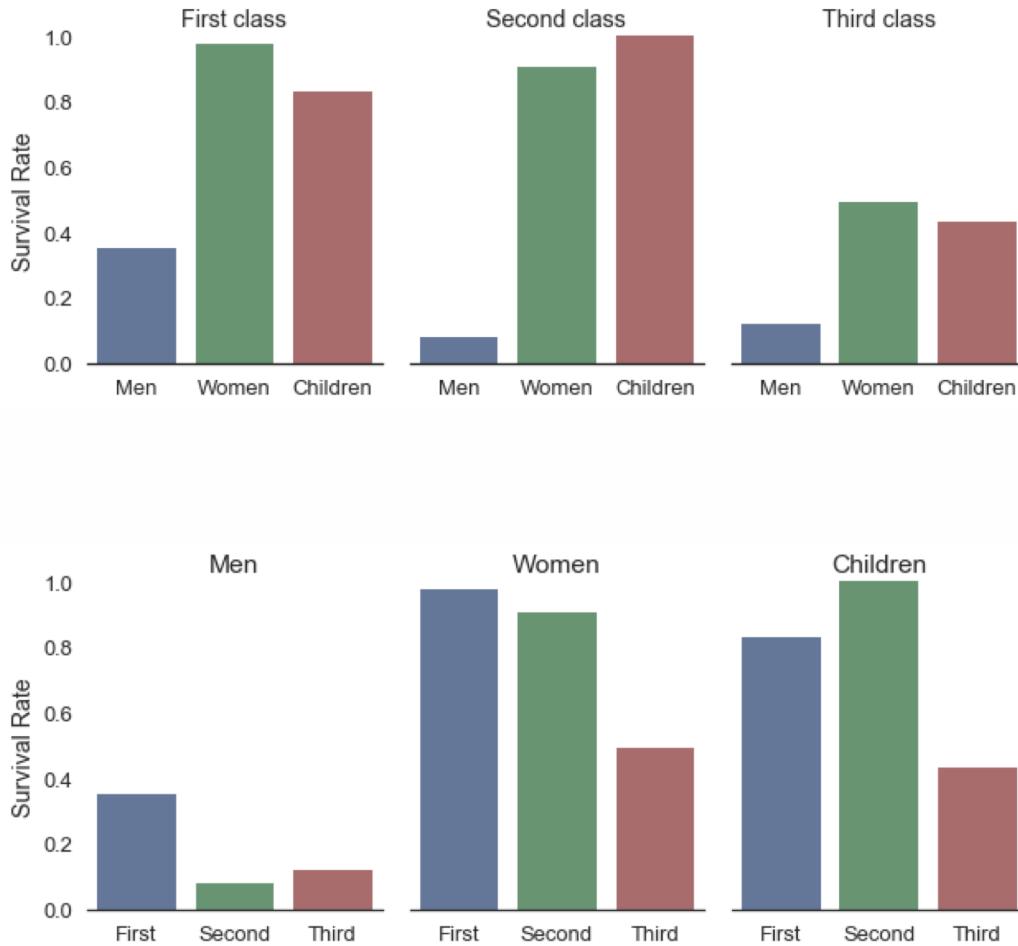
- They use horizontal or vertical bars to compare categories.
- One axis shows the compared categories, the other axis represents a discrete value
- Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulate effect (stacked bar graphs)



Bar plot example

Perspective Can Dramatically Change

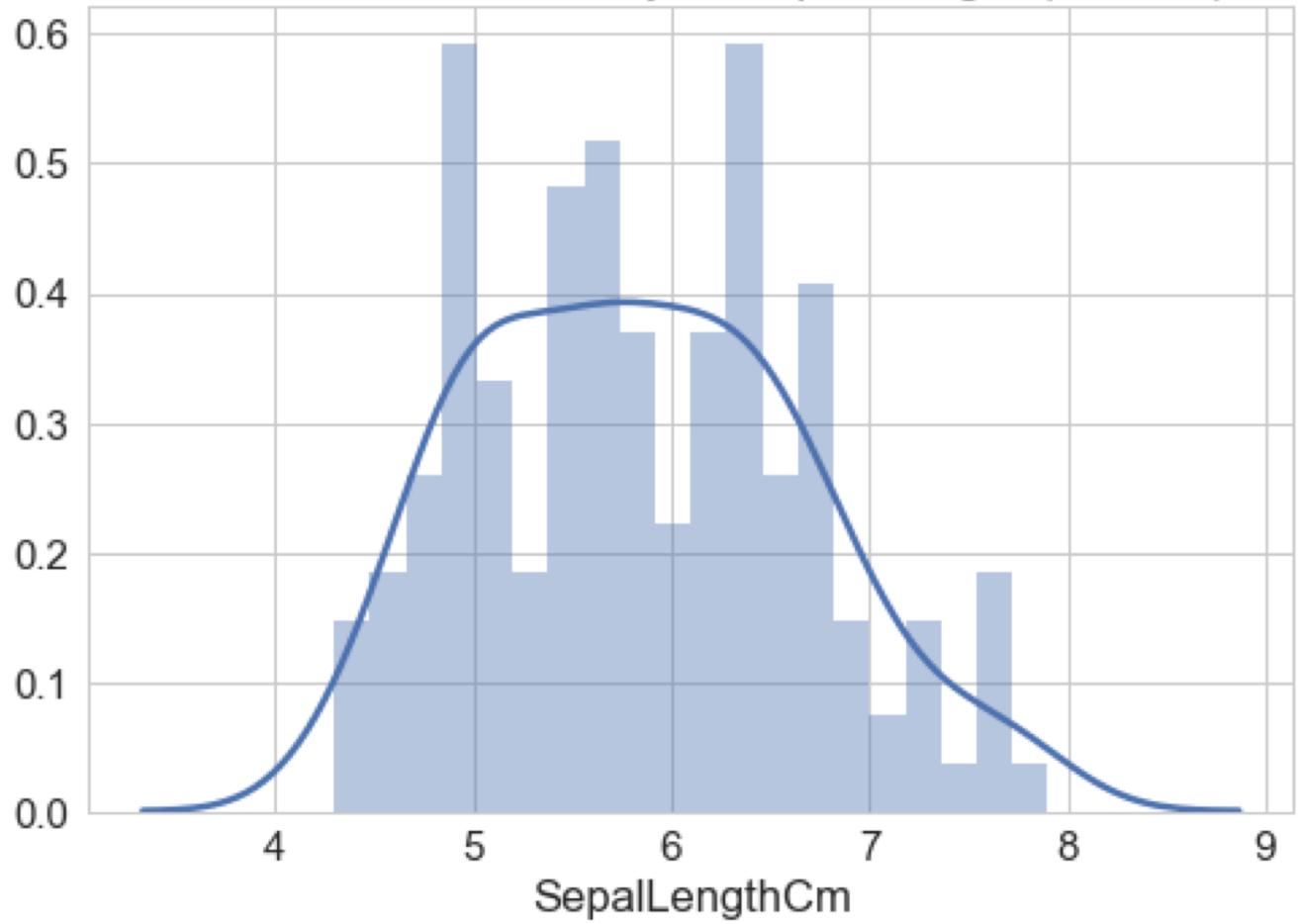
26



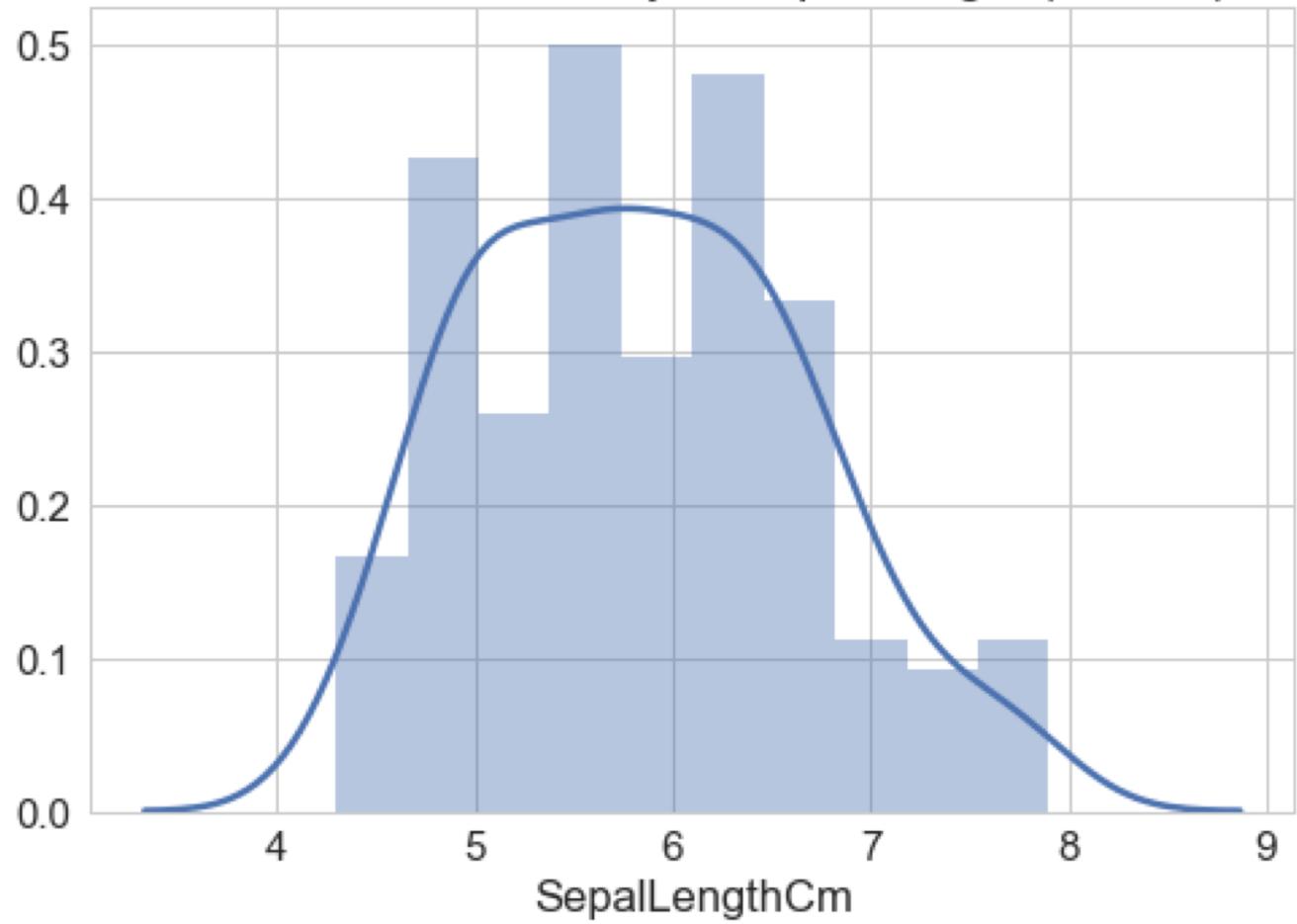
Histograms

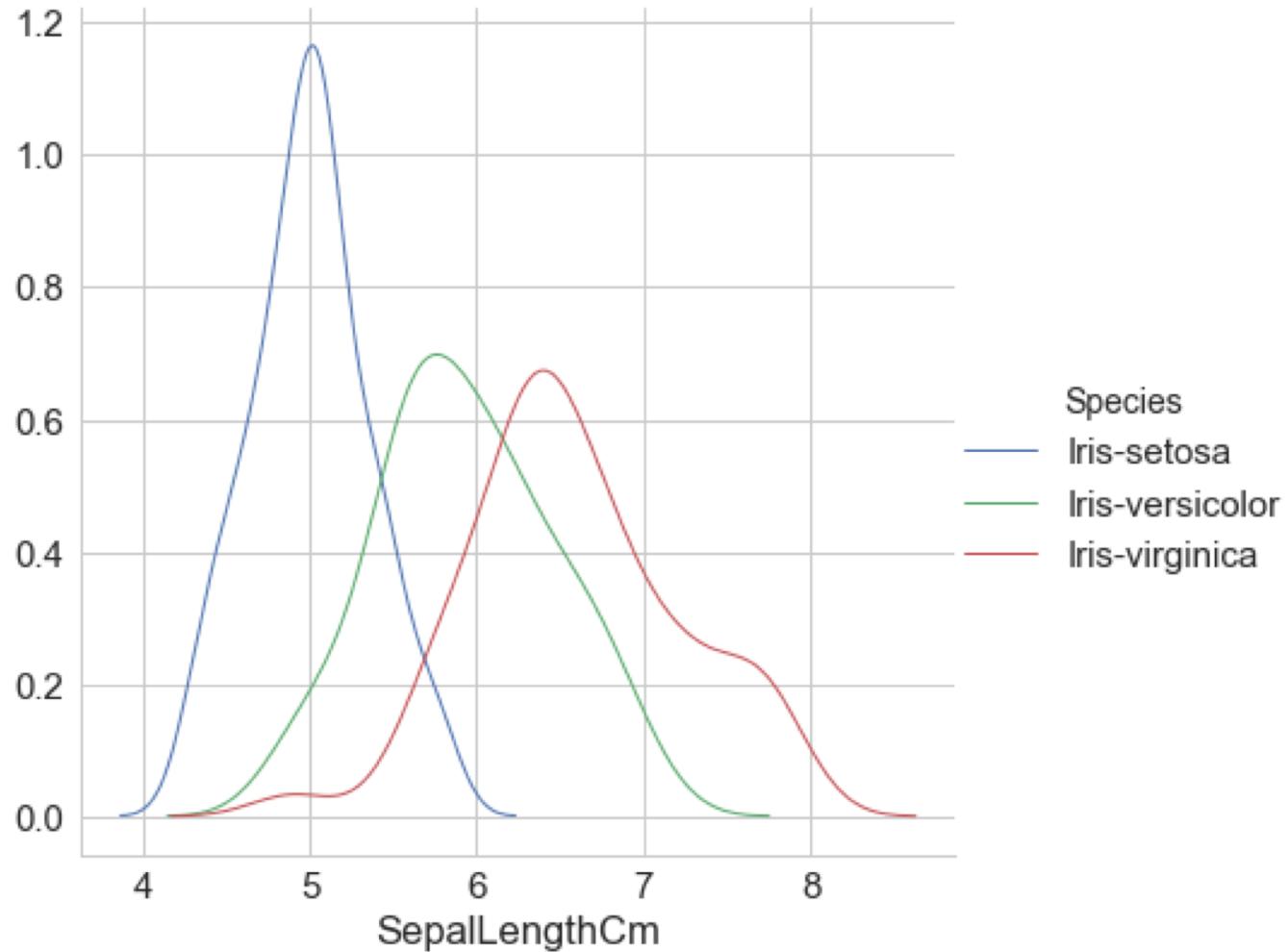
- They are a graphical representation of the distribution of data introduced by Karl Pearson in 1895
- They estimate the probability distribution of a continuous variables
- They are representations of tabulated frequencies depicted as adjacent rectangles, erected over discrete intervals (bins). Their areas are proportional to the frequency of the observations in the interval.
- The height of each bar indicates the number of objects
- Shape of histogram depends on the number of bins

Distribution and Density of Sepal Length (20 bins)

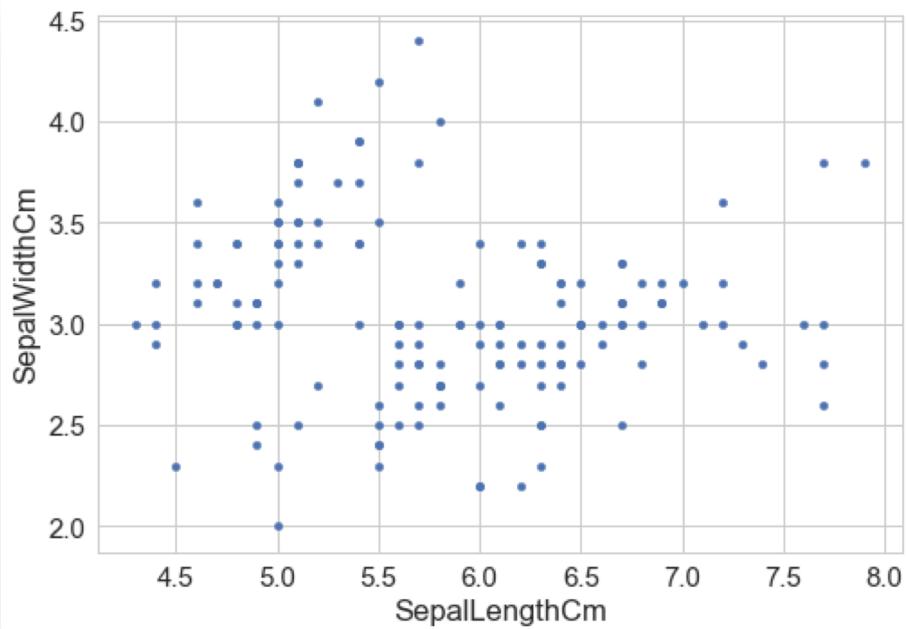


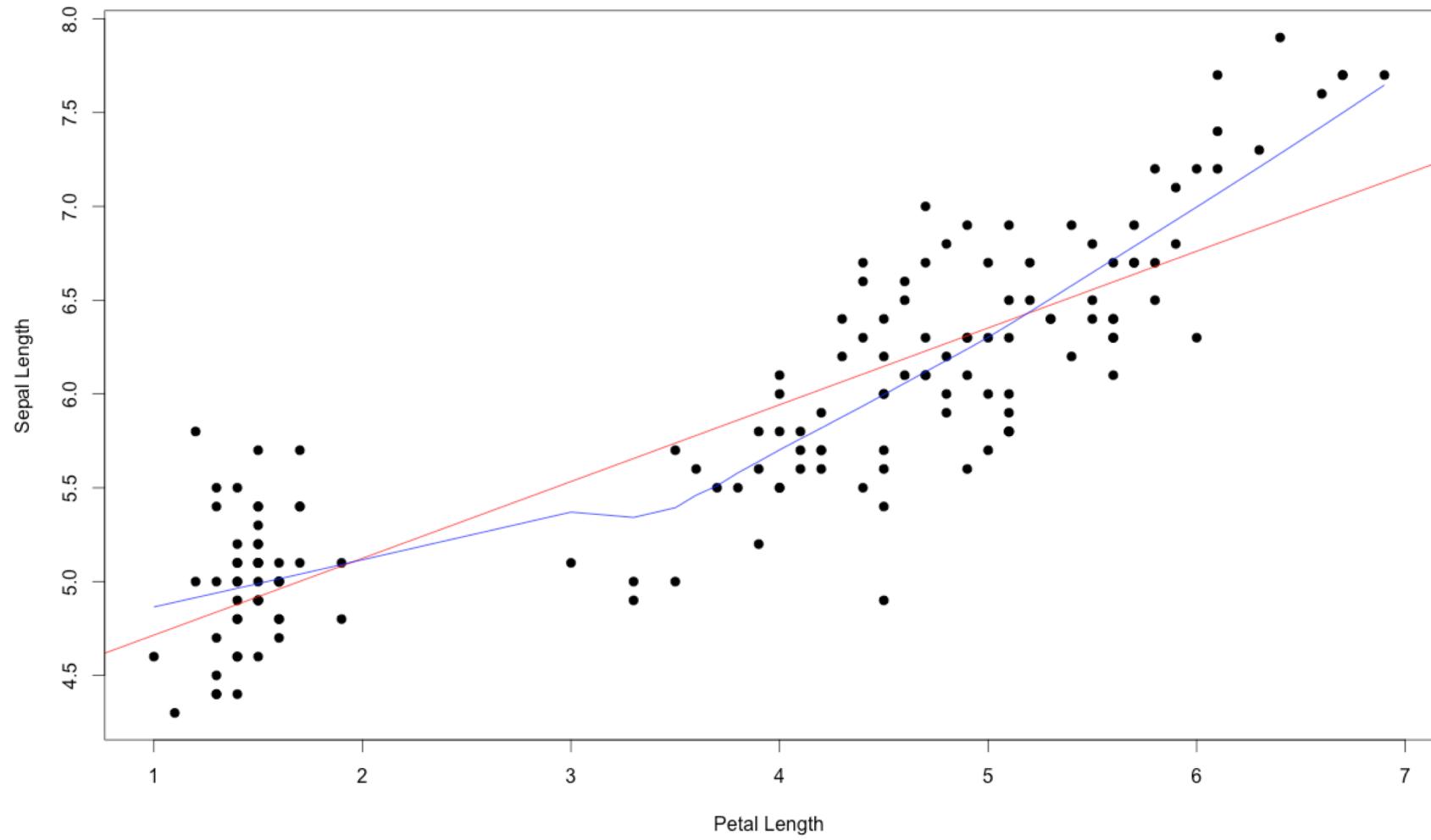
Distribution and Density of Sepal Length (10 bins)

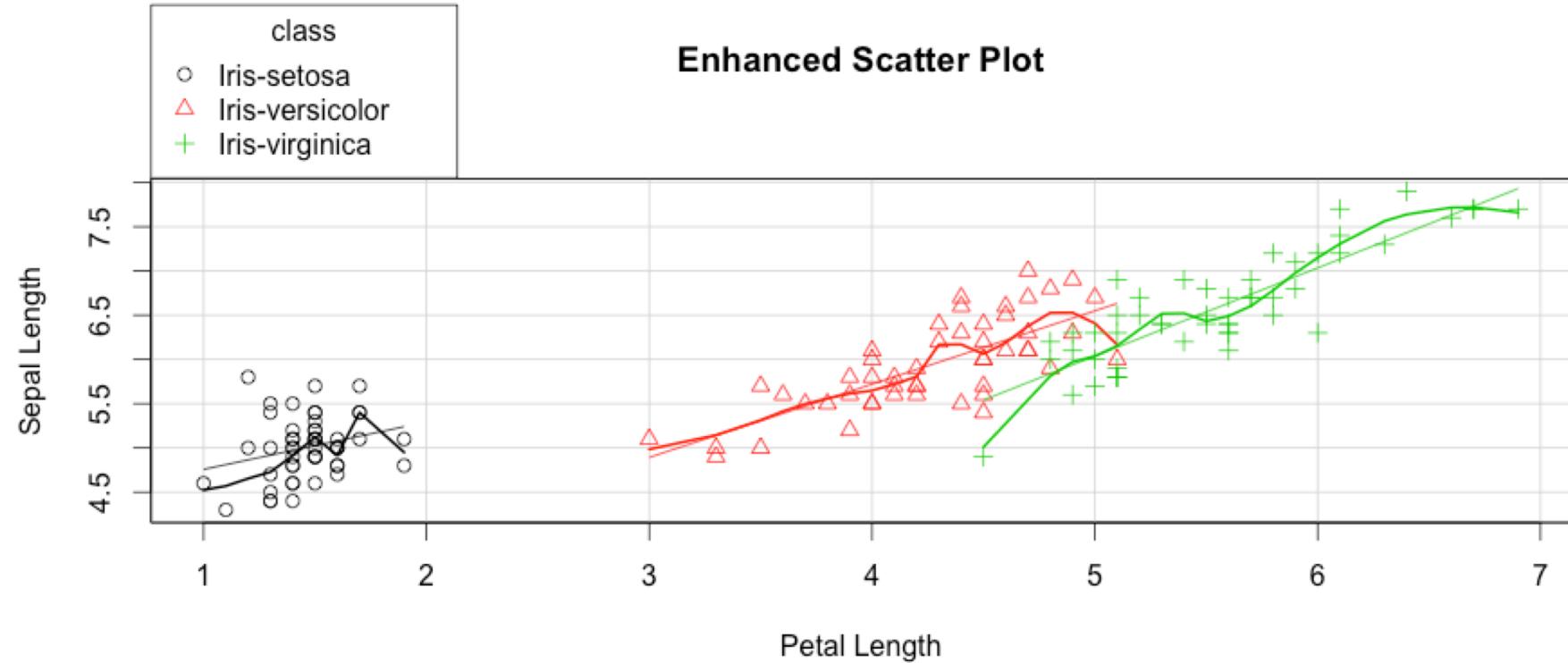


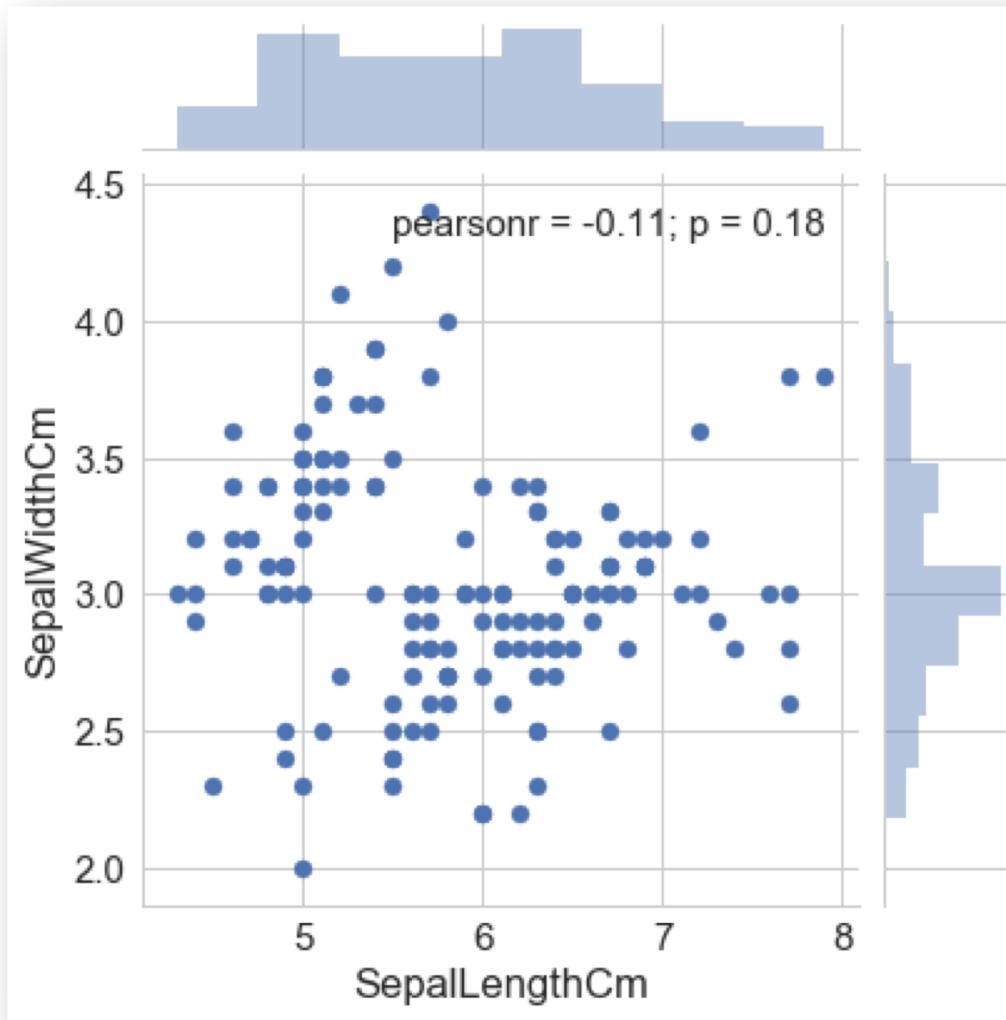


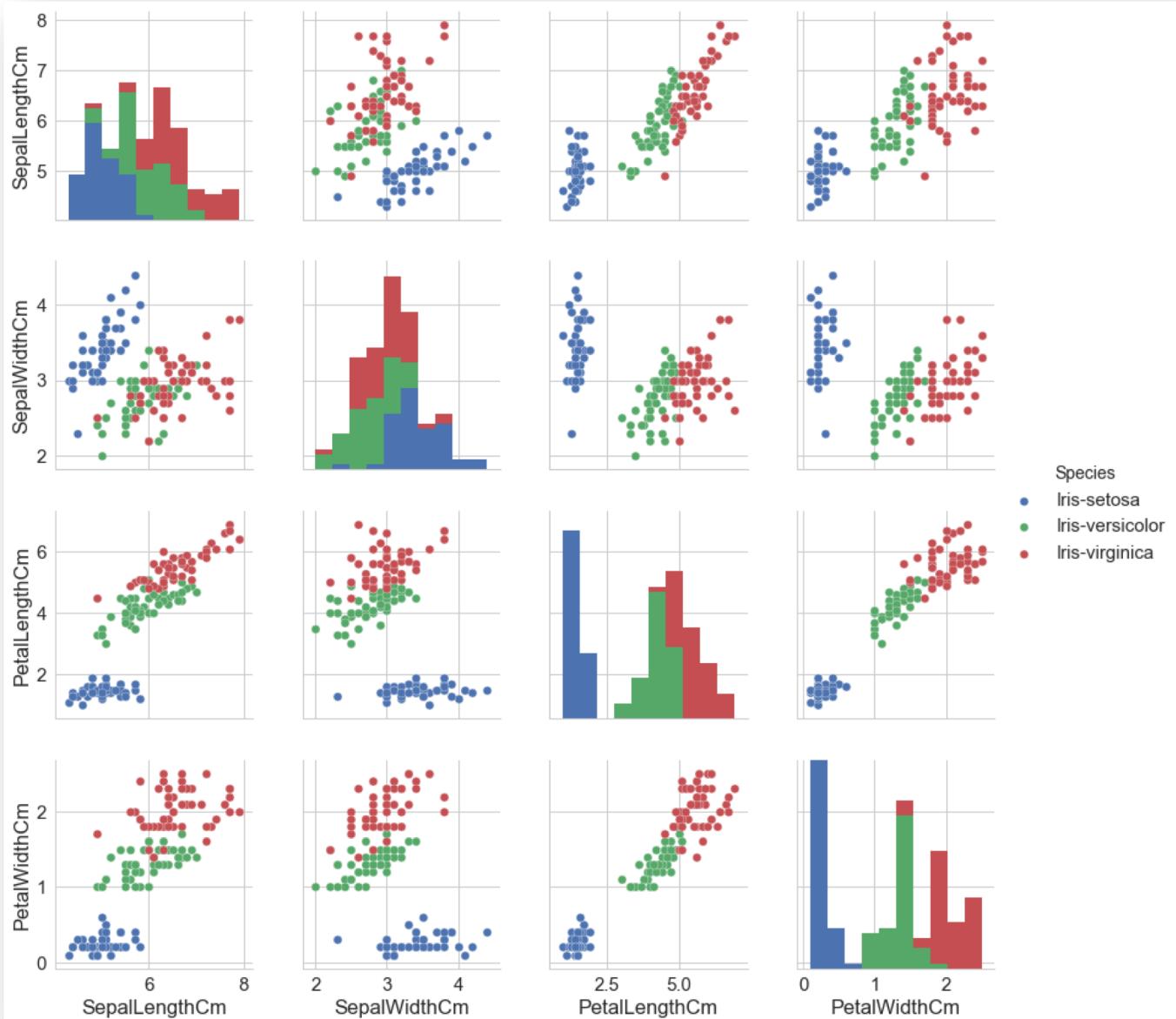
- Attributes values determine the position
- Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
- Often additional attributes can be displayed by using the size, shape, and color of the markers that represent the objects
- It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
- Examples: <http://www.statmethods.net/graphs/scatterplot.html>

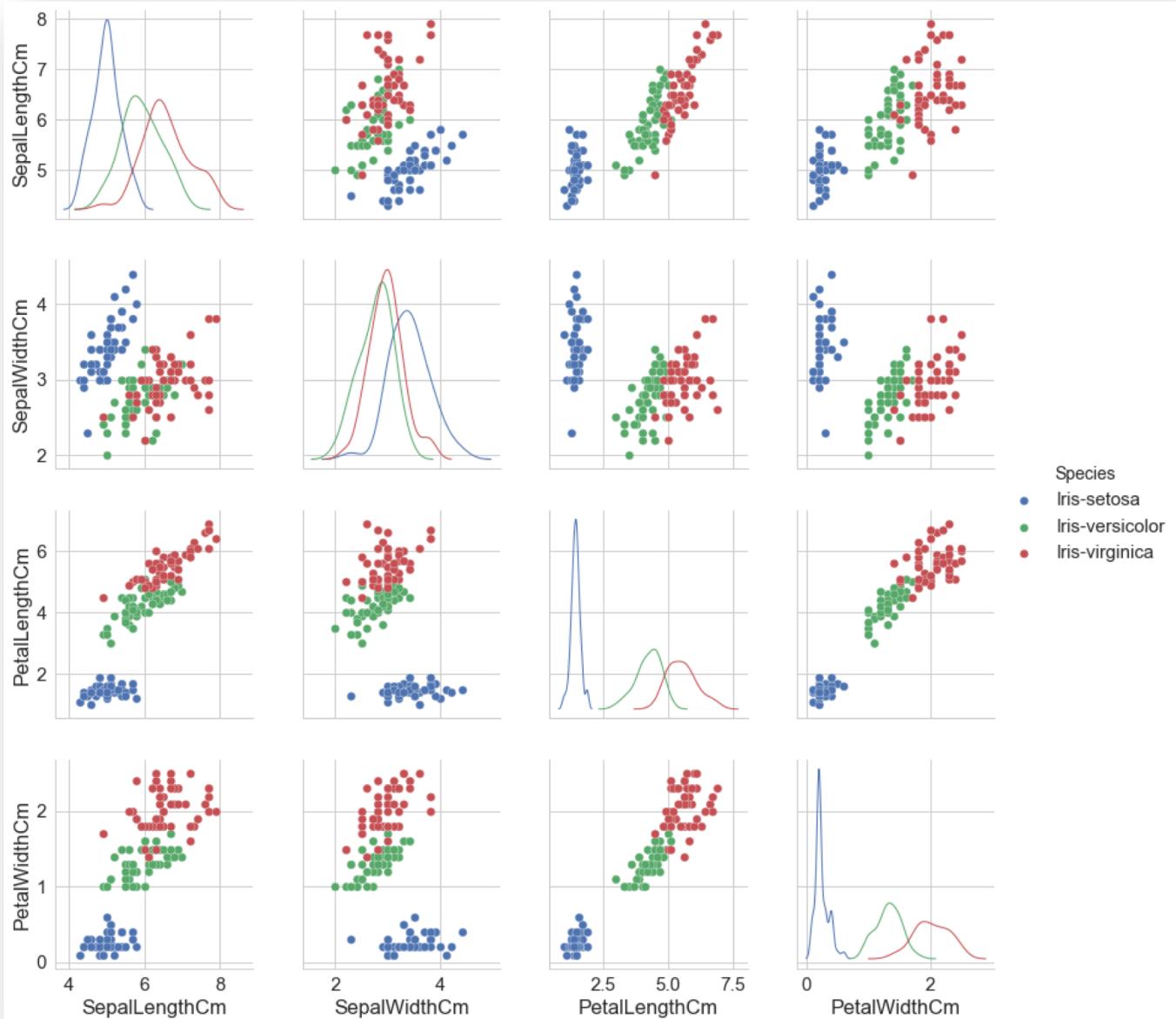




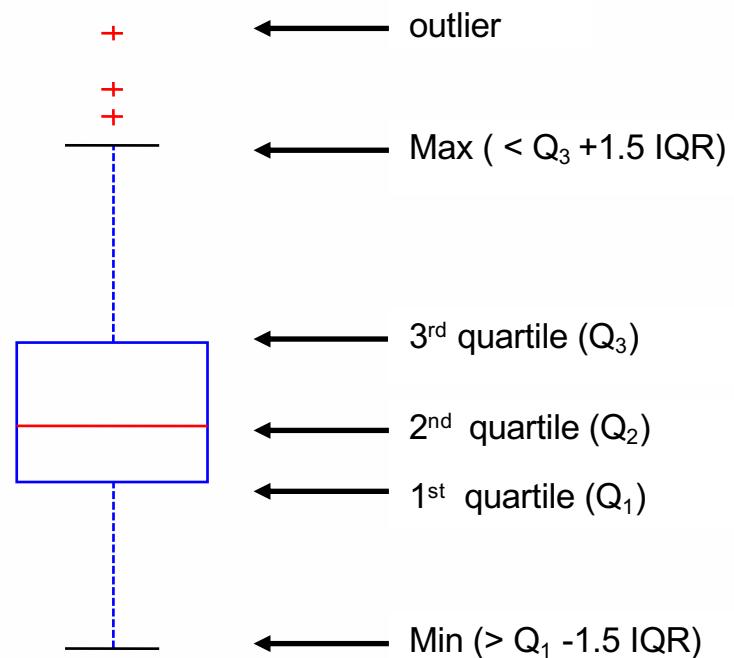






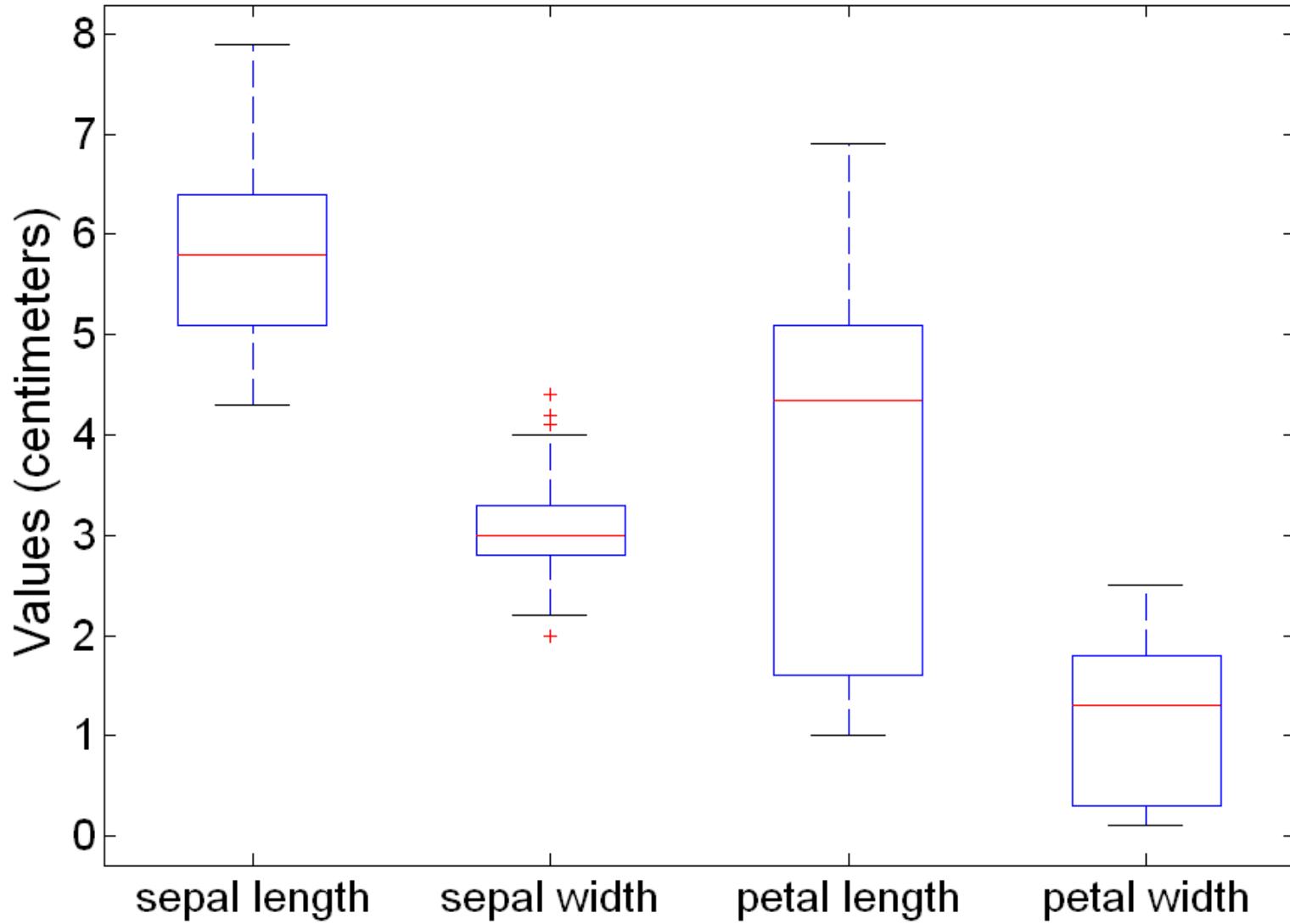


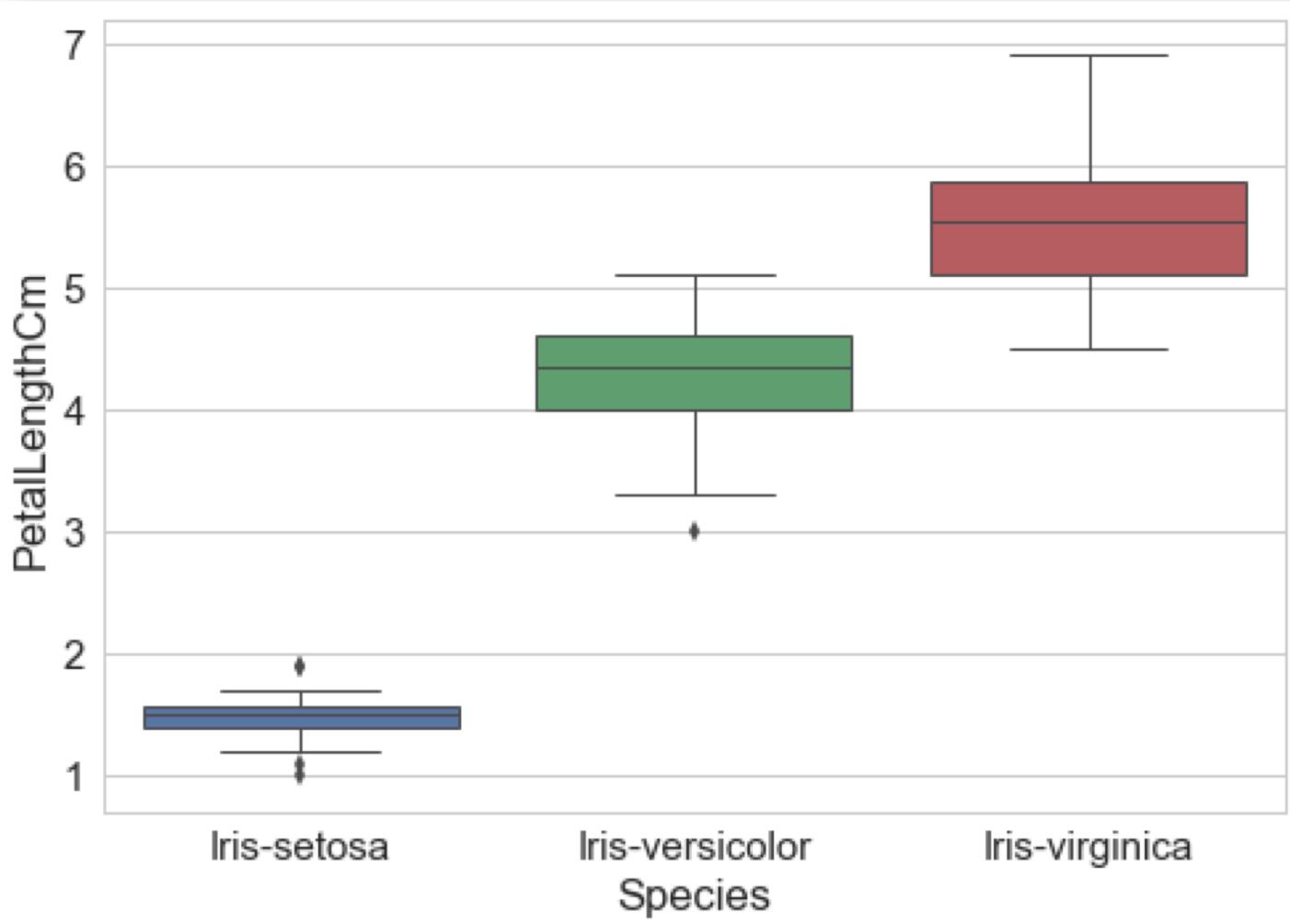
- Box Plots (invented by Tukey)
- Another way of displaying the distribution of data
- Following figure shows the basic part of a box plot
- IQR is the interquartile range or $Q_3 - Q_1$

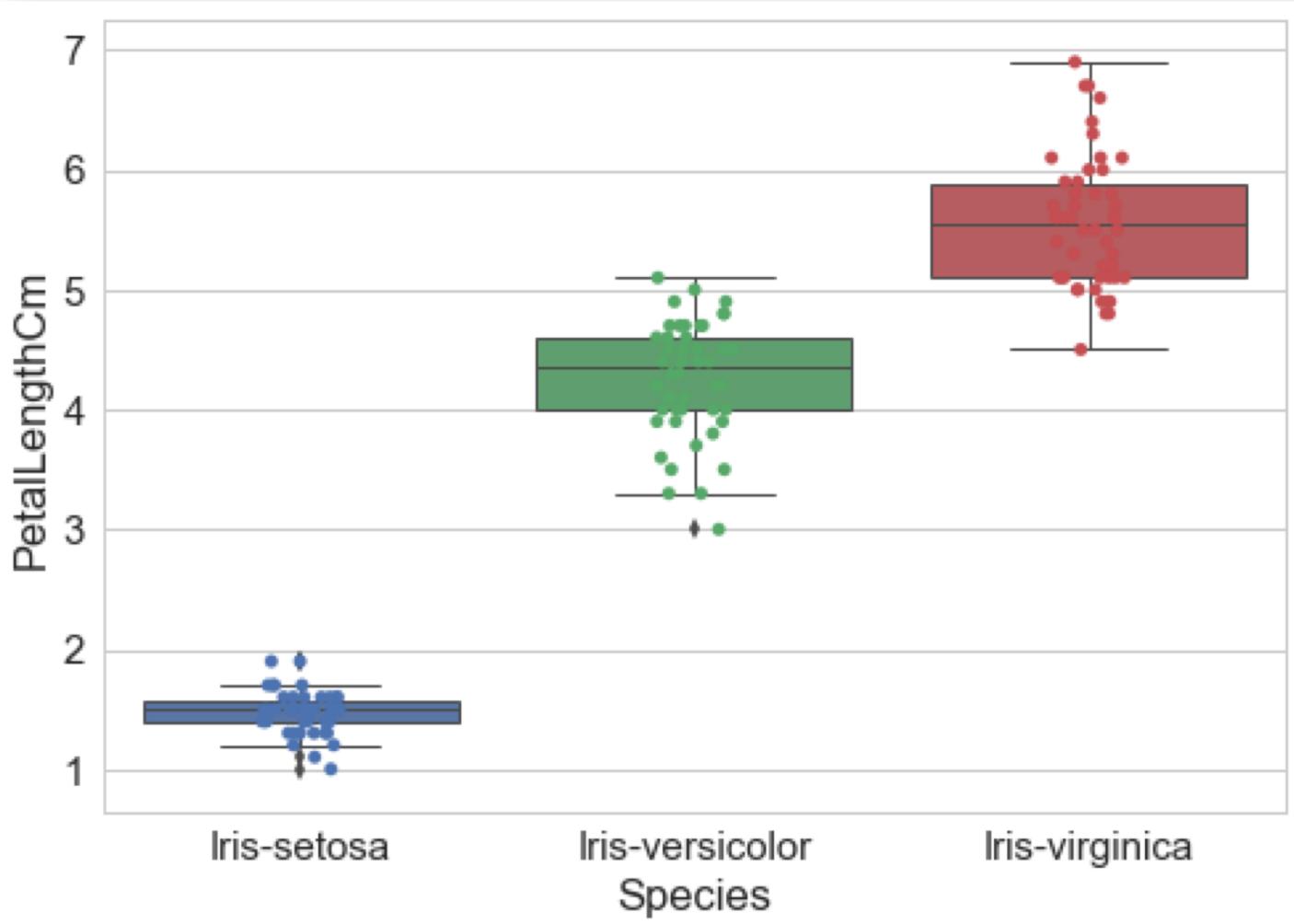


Box Plot Example

40

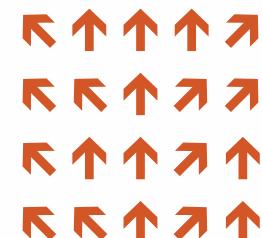
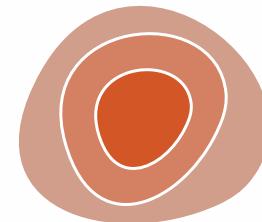






Visualizing Spatial Data

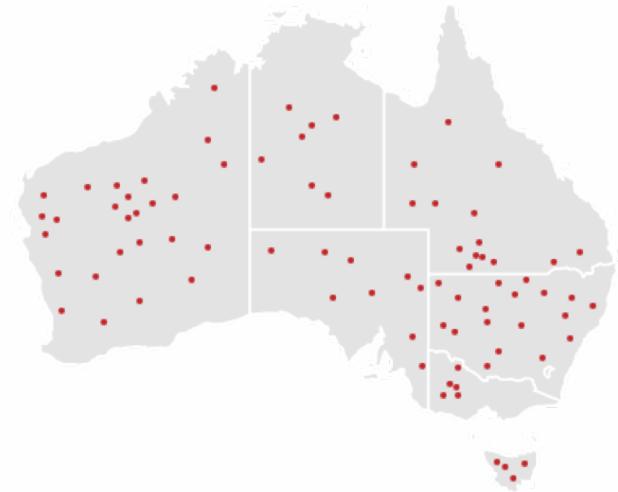
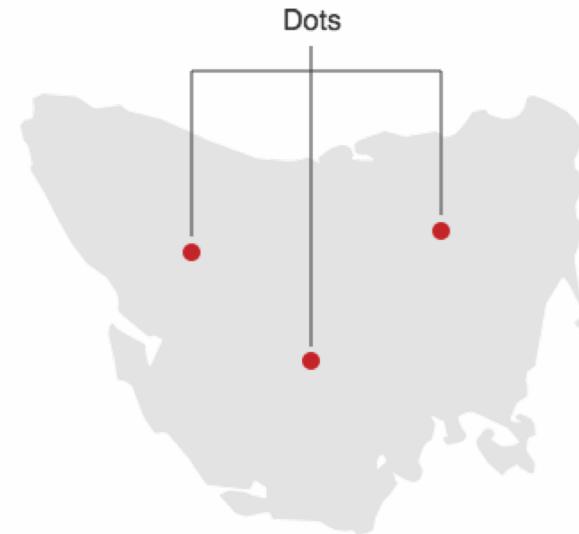
- Geometry
 - Geographic
 - Others
- Scalar fields (one value per cell)
 - Isocontours
 - Direct volume rendering
- Vector and tensor fields
(many values per cell)
 - Flow glyphs
 - Geometric (sparse seeds)
 - Textures (dense seeds)
 - Features (globally derived)



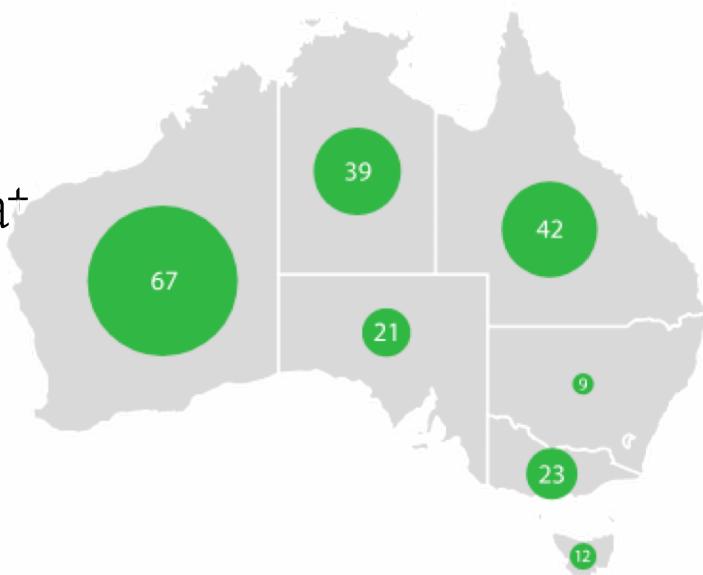
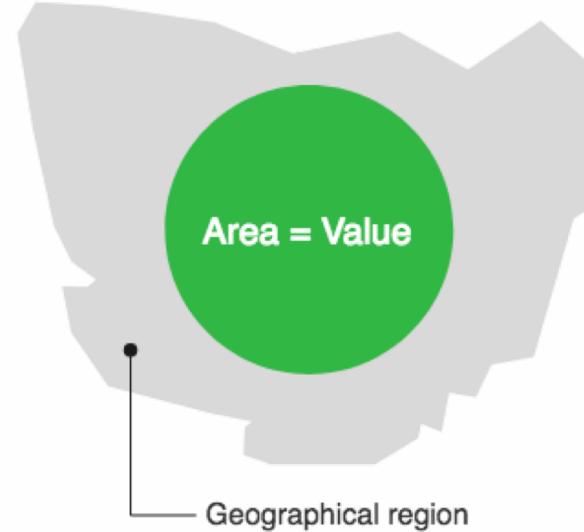
Geometry

Dot Maps

- What?
 - Geometry (position)
- Why?
 - Locate data in space
- Remarks
 - Scale up to hundreds of items
 - Color/shape can encode an additional categorical attribute (reduce scalability)

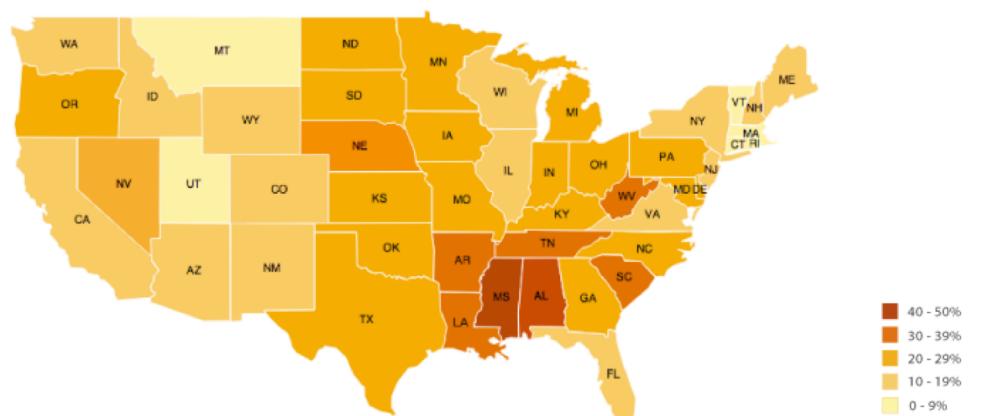
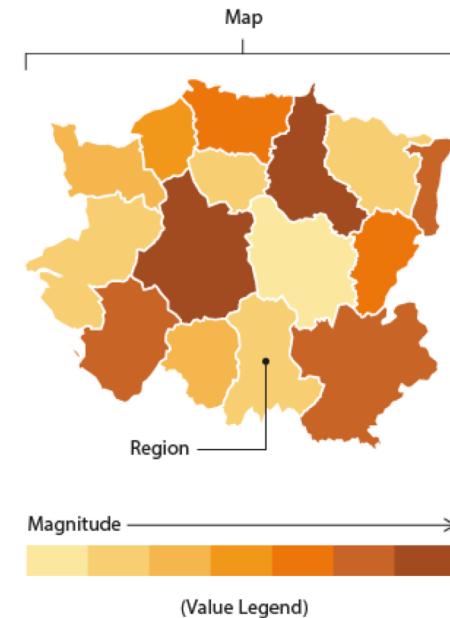


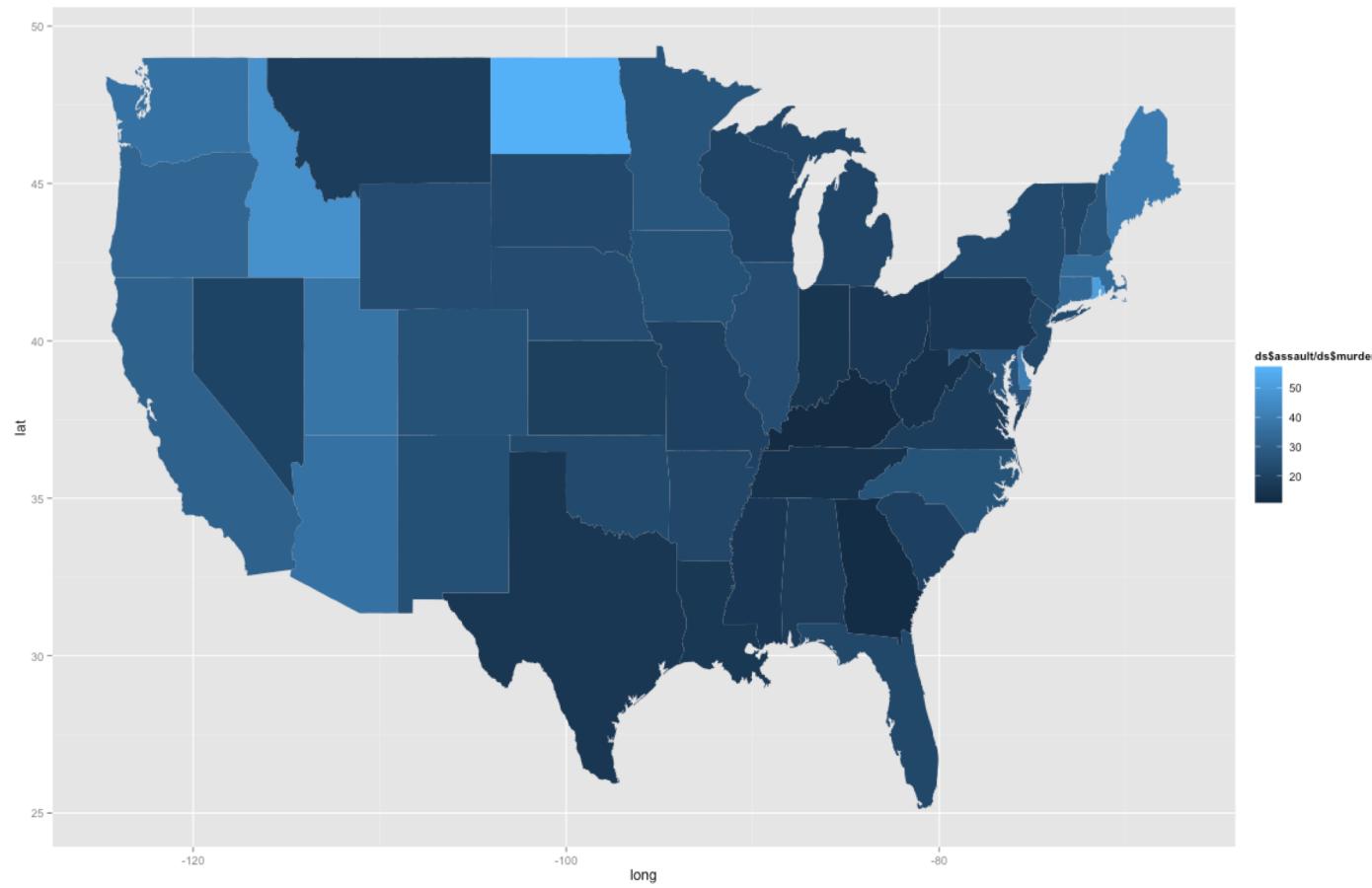
- What?
 - Geometry (position)
 - 1 quantitative attribute
- Why?
 - Locate data in space
 - Lookup and compare
- Remarks
 - Scale up to hundreds of items
 - Color can encode an additional category (interaction with size).



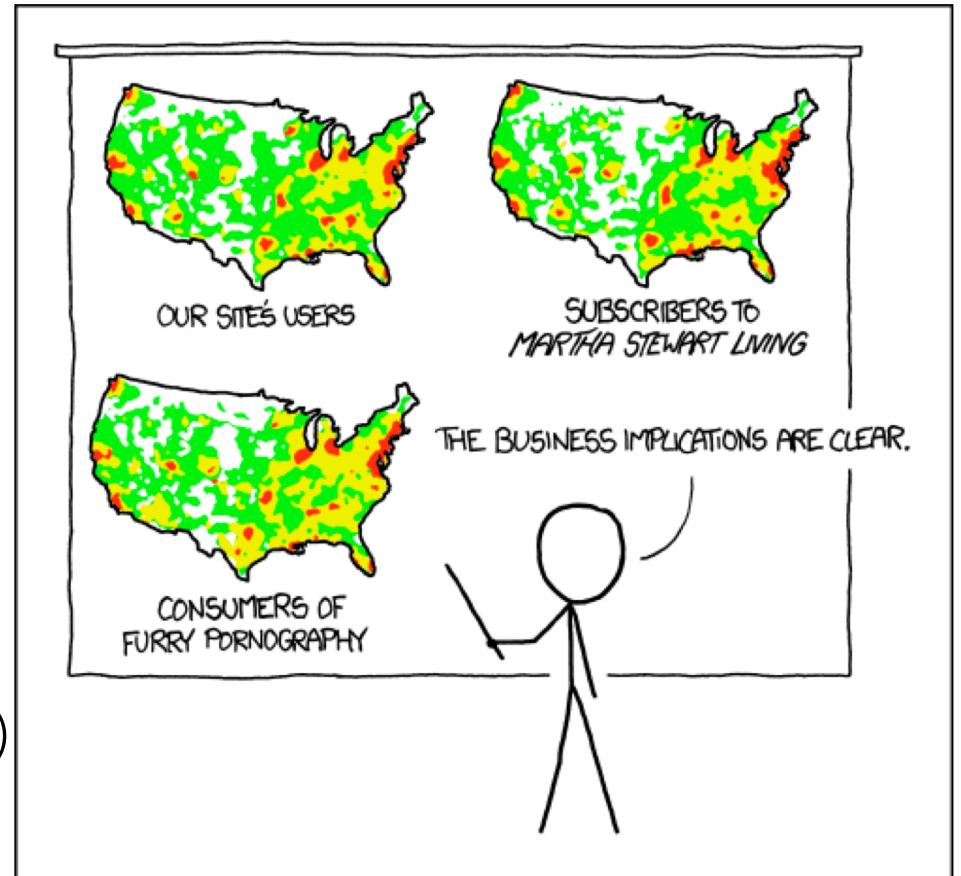
Choropleth Map

- What?
 - Geometry (position)
 - One quantitative attribute
- Why?
 - Locate data in space
 - Lookup and compare
- Remarks
 - Scale up to ~1000 items
 - Hue can encode an additional categorical attribute (better if binary)

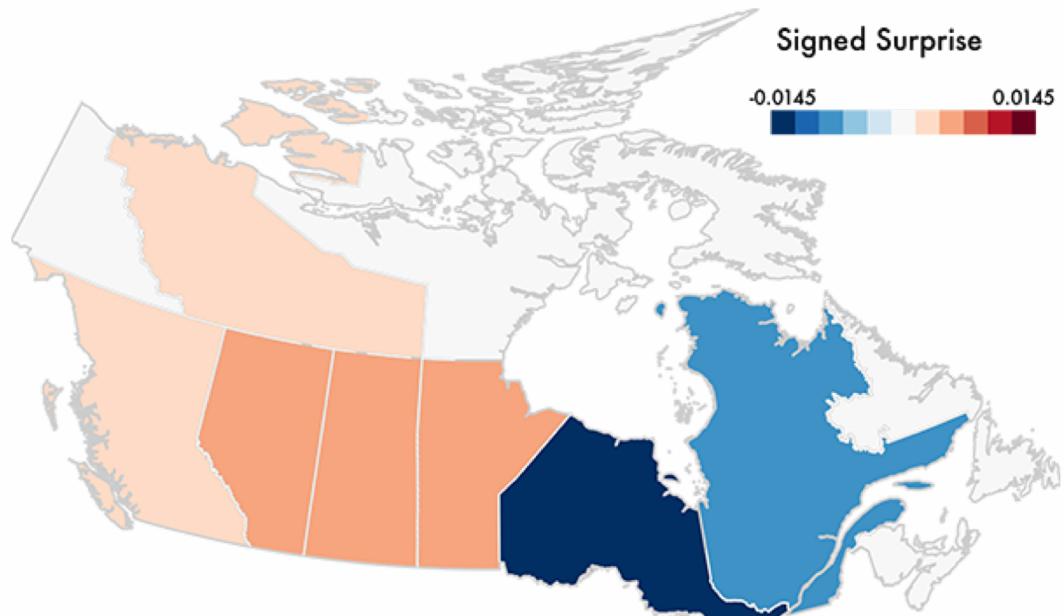




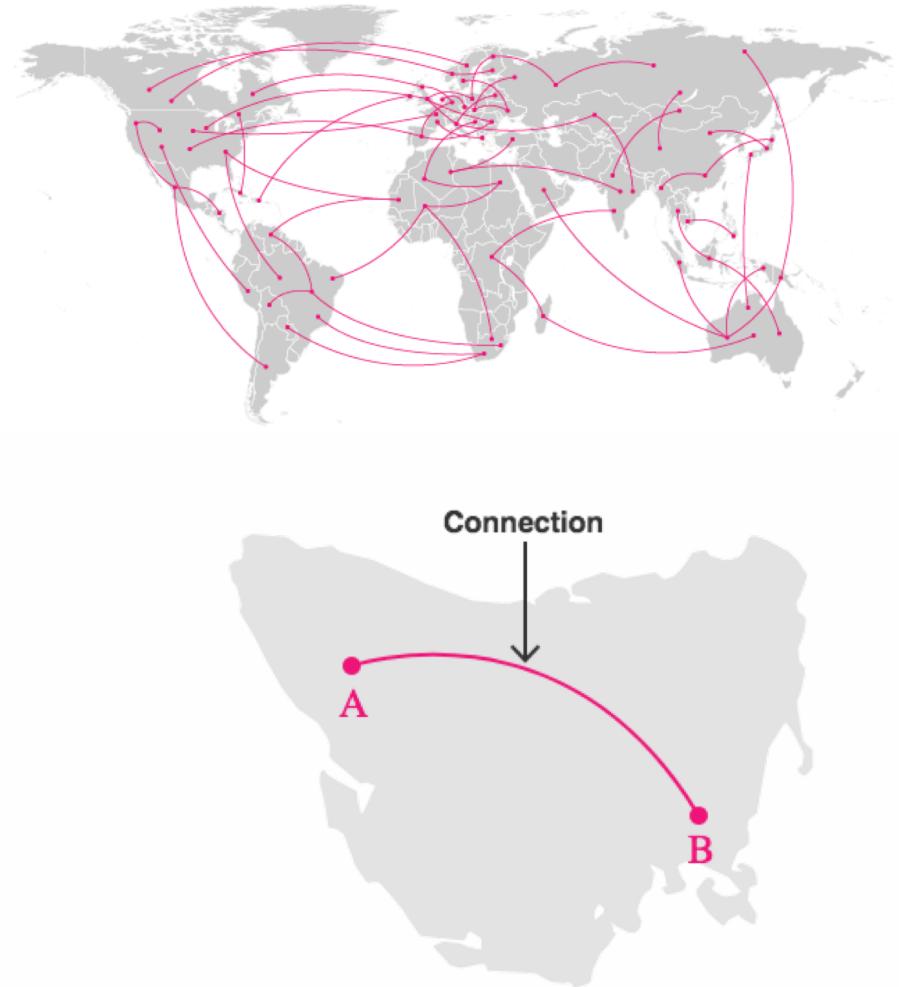
- Using absolute values is dangerous!
- Any map would show population distributionS
- How to deal with this?
 - Visualize per capita (relative)
 - Use statistical models
- See the example at
<https://xkcd.com/1138>



- What?
 - Geometry (position)
 - 1 quantitative attribute
 - 1 derivative attribute
- Why?
 - Locate data in space
 - Lookup and compare
- Remarks
 - The surprise is computed as a function prior and posterior probability of data distribution.
 - Prior probability is generated with a family of standard models.
- Illustration from <https://medium.com/@uwdata/surprise-maps-showing-the-unexpected-e92b67398865>



- What?
 - Network and positions
- Why?
 - Lookup path
 - Identify patterns
- Remarks
 - Size of links can encode an additional ordered attribute (3-4 bins at max)

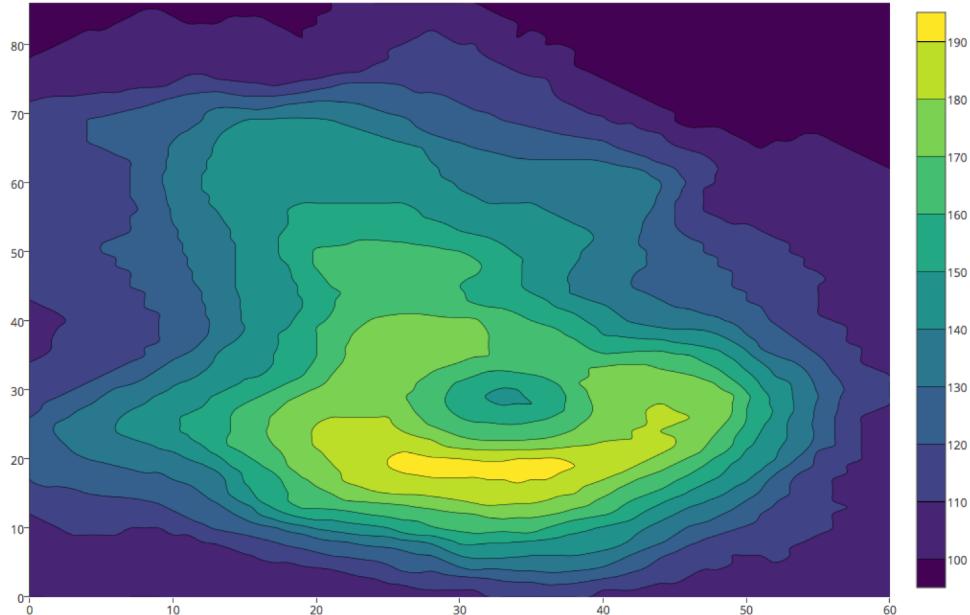


Scalar Fields

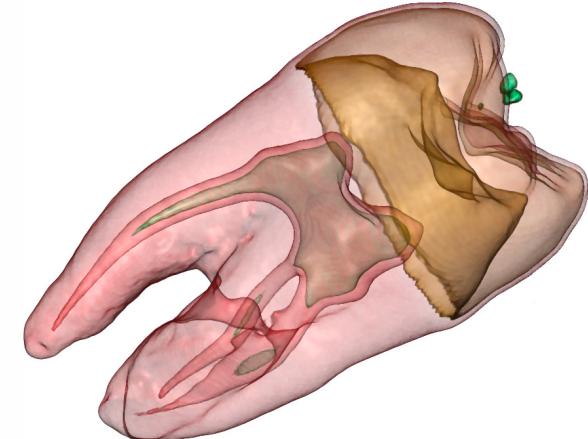
- What?
 - Geographic data
 - 1 quantitative attribute
 - derived positions
- Why?
 - Shape
- Remarks
 - The lines are computed from the values of scalar field
 - Area can be filled and color encoded



- What?
 - 2D spatial field
 - 1 quantitative attribute
 - derived geometry
- Why?
 - Shape and patterns
- Remarks
 - The lines are computed from the values of scalar field
 - Area can be empty or filled and color encoded

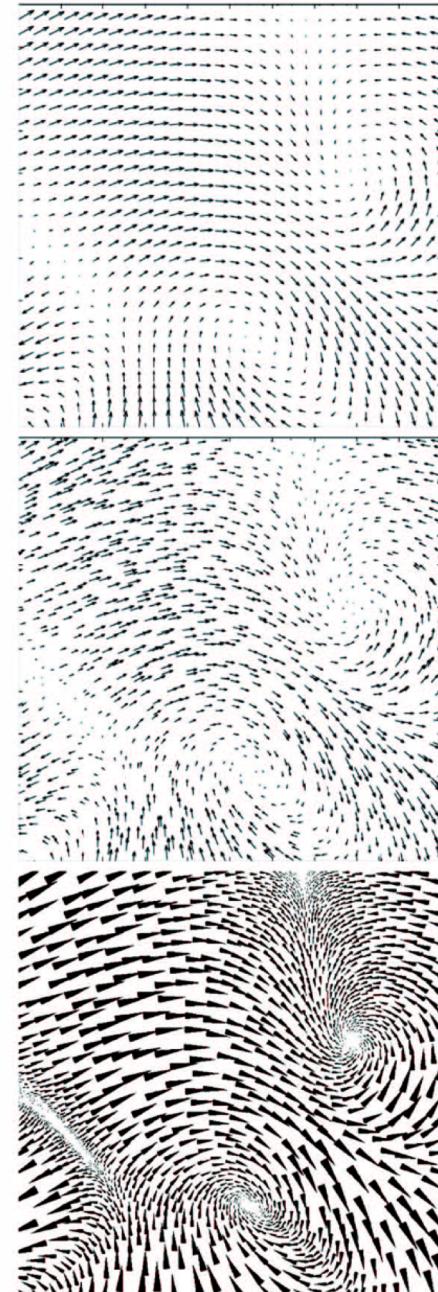
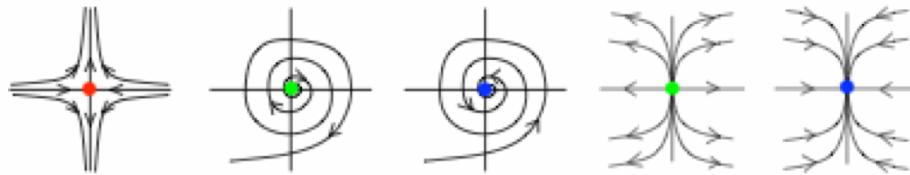


- What?
 - 3D spatial scalar field
 - 1 quantitative attribute
 - derived geometry
- Why?
 - Shape
- Remarks
 - Tree of isosurfaces: positions computed for specific values of the scalar field

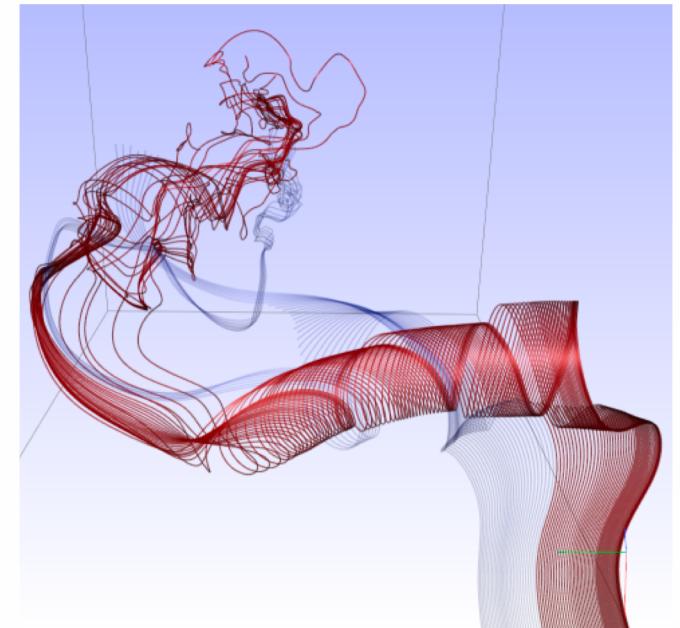


Vector Field

- What?
 - 2D vector fields
- Why?
 - Shape and patterns
 - Identify critical points
- Remarks
 - Different glyphs can be used to represent vectors
 - Density of grid and jittering

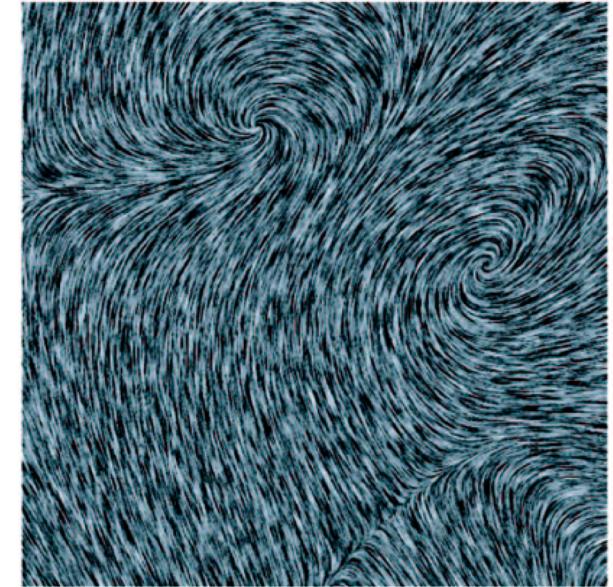


- What?
 - 2D/3D vectorial field
 - derived geometry
- Why?
 - Shape and patterns
 - Identify critical points
- Remarks
 - Seeding strategy affects the outcome
 - Usage of clustering and color coding improves readability

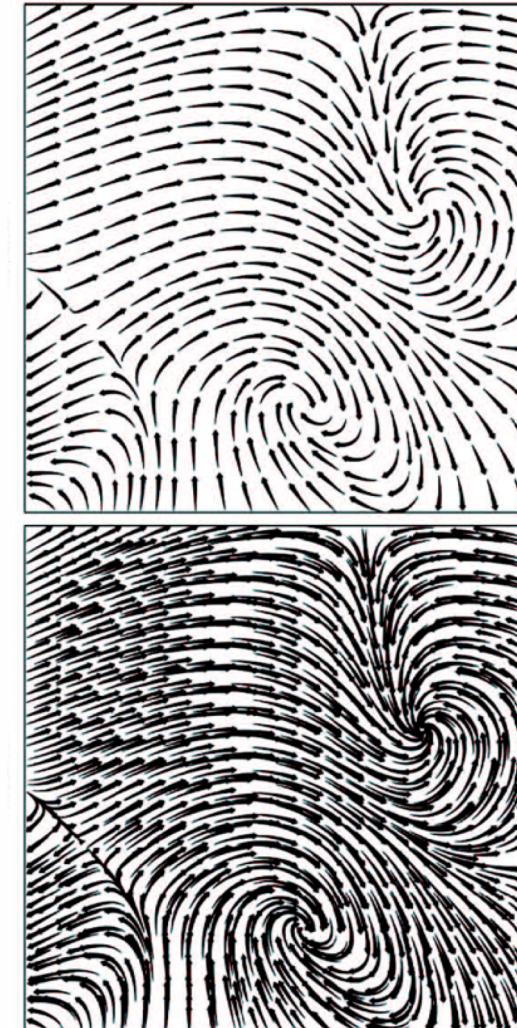


Texture Flow

- What?
 - 2D vectorial field
- Why?
 - Shape and patterns
 - Identify critical points
- Remarks
 - Similar to glyphs flow,
but computes the flow of
a continuous distribution of particles

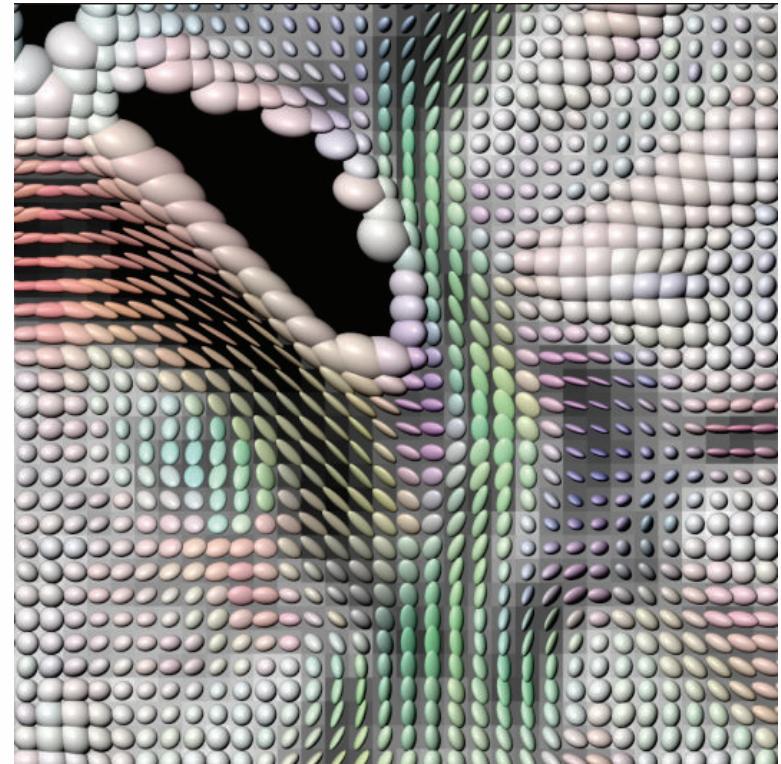


- What?
 - 2D vector fields
- Why?
 - Shape and patterns
 - Identify critical points
- Remarks
 - Similar to glyphs flow
 - But seeding is based on global computing strategy to identify areas with similar behaviors



Tensor Fields

- What?
 - 2D/3D tensor field
- Why?
 - Shape and patterns
- Remarks
 - Use shape, orientation, color, opacity to represents field tensors

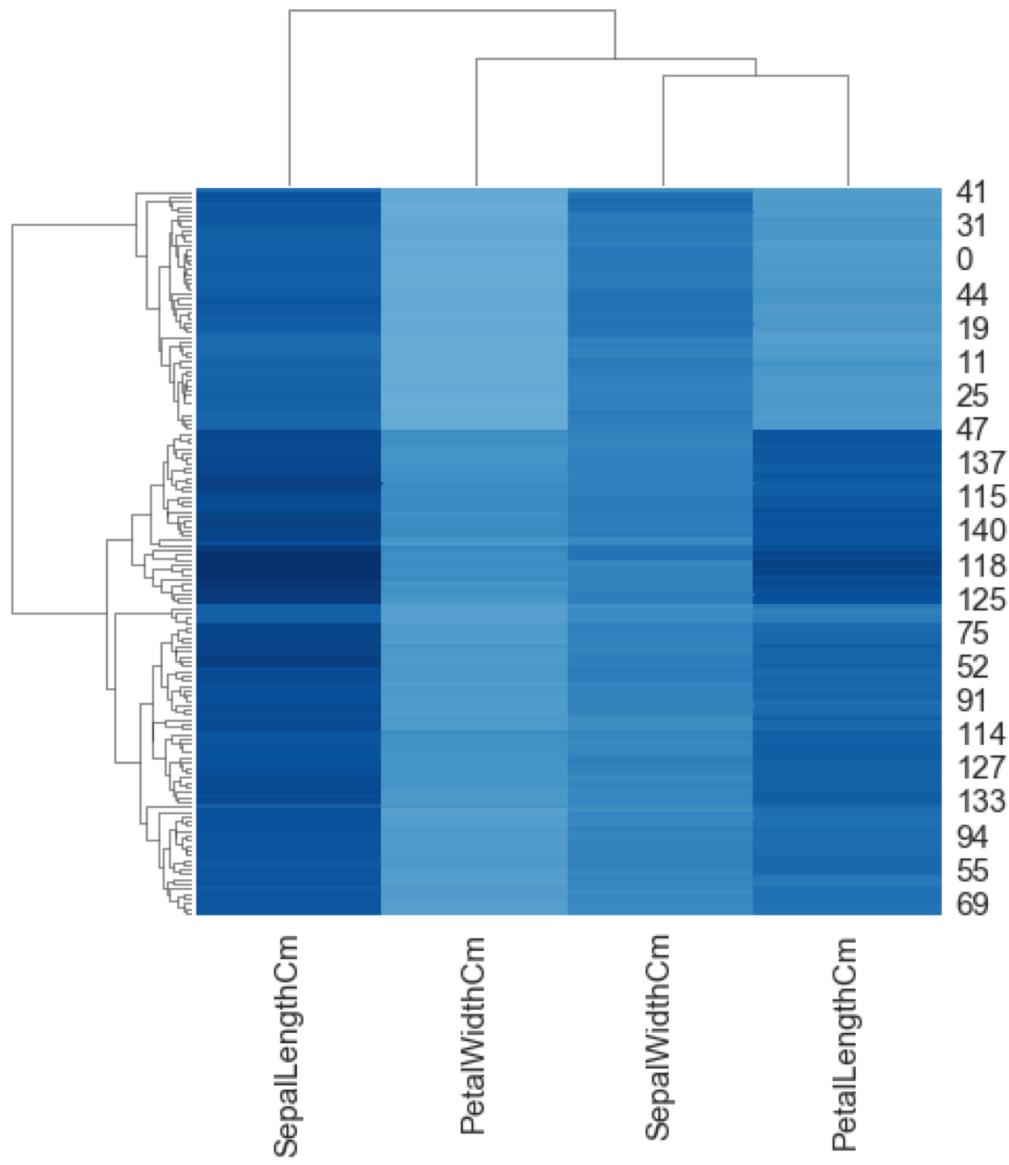


More than Two Dimensions at Once

Two main approaches

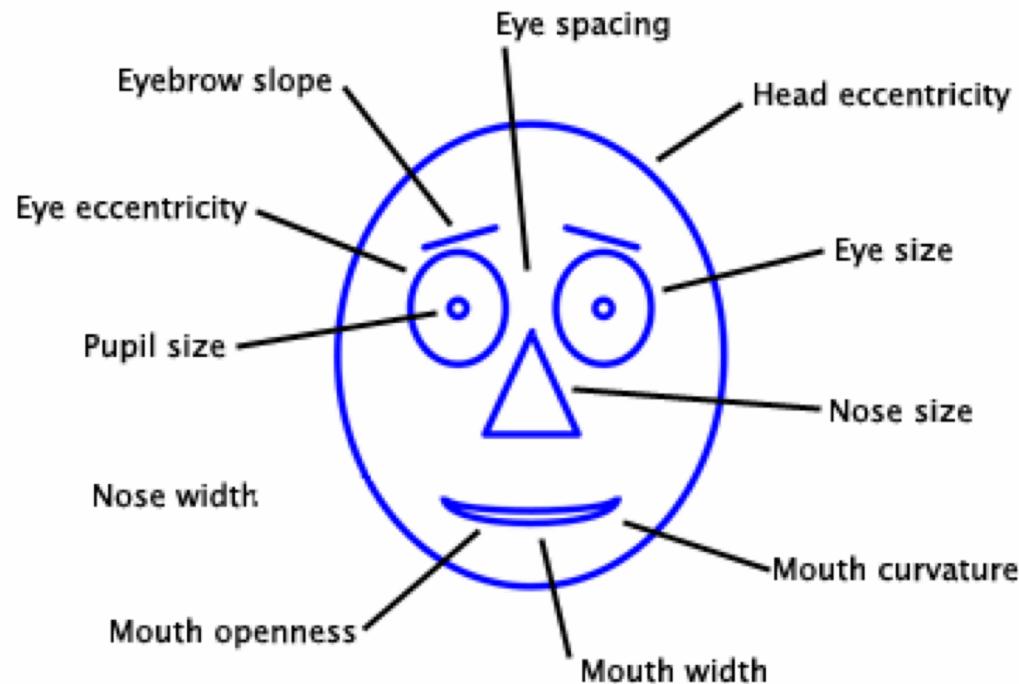
Visualize all the dimensions at once
(e.g., Chernoff faces and starplots)

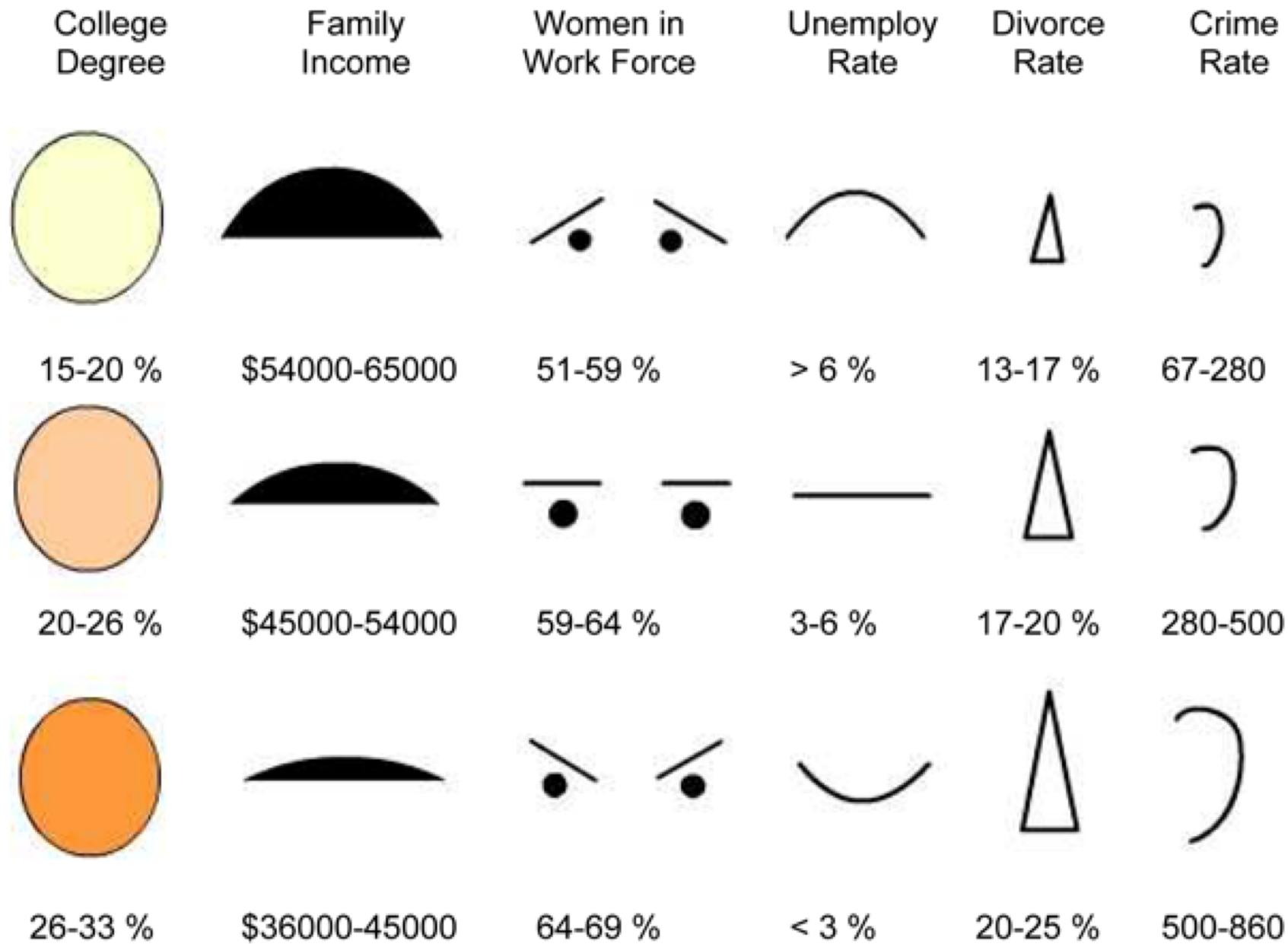
Project the data into a smaller space
and visualize the the projected data



- **Chernoff Faces**
 - Approach created by Herman Chernoff
 - Associates each attribute with a characteristic of a face
 - The values of each attribute determine the appearance of the corresponding facial characteristic
 - Each example becomes a separate face
 - Relies on human's ability to distinguish faces
- **Star Plots**
 - Similar approach that uses axes radiate from a central point
 - The line connecting the values of an object is a polygon

- More dimensions are plotted using several features of human faces





Mazda RX4



Mazda RX4 Wag



Datsun 710



Hornet 4 Drive



Hornet Sportabout



Valiant



Duster 360



Merc 240D



Merc 230



Merc 280



Merc 280C



Merc 450SE



Merc 450SL



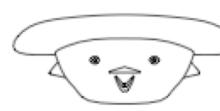
Merc 450SLC



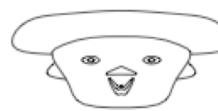
Cadillac Fleetwood



Lincoln Continental



Chrysler Imperial



Fiat 128



Honda Civic



Toyota Corolla



Toyota Corona



Dodge Challenger



AMC Javelin



Camaro Z28



Pontiac Firebird



Fiat X1-9



Porsche 914-2



Lotus Europa



Ford Pantera L



Ferrari Dino



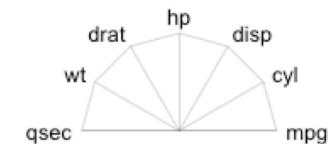
Maserati Bora



Volvo 142E



Motor Trend Cars : stars(*, full = F)



Motor Trend Cars



Mazda RX4



Mazda RX4 Wag



Datsun 710



Hornet 4 Drive



Hornet Sportabout



Valiant



Duster 360



Merc 240D



Merc 230



Merc 280



Merc 280C



Merc 450SE



Merc 450SL



Merc 450SLC



Cadillac Fleetwood



Lincoln Continental



Chrysler Imperial



Fiat 128



Honda Civic



Toyota Corolla



Toyota Corona



Dodge Challenger



AMC Javelin



Camaro Z28



Pontiac Firebird



Fiat X1-9



Porsche 914-2



Lotus Europa



Ford Pantera L



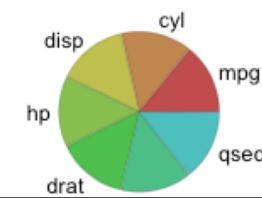
Ferrari Dino



Maserati Bora



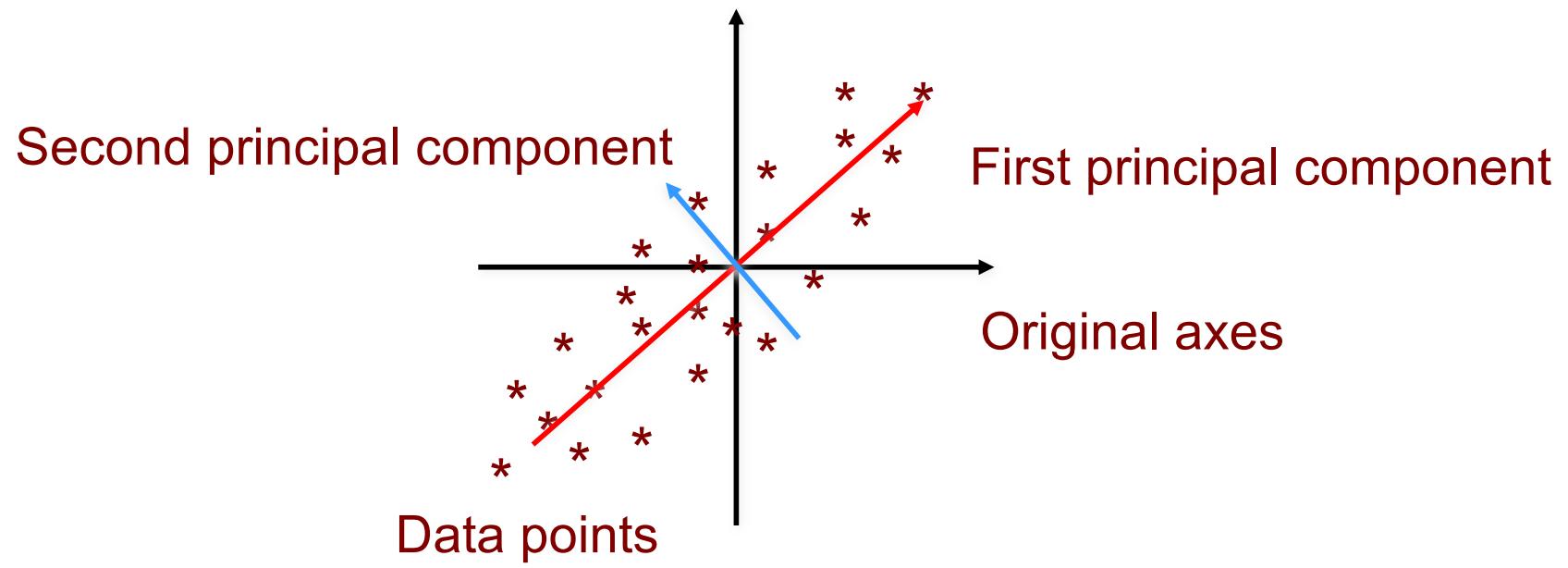
Volvo 142E



Principal Component Analysis

- Typically applied to reduce the number of dimensions of data (feature selection)
- The goal of PCA is to find a projection that captures the largest amount of variation in data
- Given N data vectors from n -dimensions, find $k < n$ orthogonal vectors (the principal components) that can be used to represent data
- Works for numeric data only and it is affected by scale so data usually need to be rescaled before applying PCA

- Normalize input data
- Compute k orthonormal (unit) vectors, i.e., principal components
- Each input data (vector) is a linear combination of the k principal component vectors
- The principal components are sorted in order of decreasing “significance” or strength
- Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data





t Distributed Stochastic Neighbor Embedding

- Nonlinear dimensionality reduction technique used to map high-dimensional data into two or three dimensions
- Maps data points from the original data into “map points” of a 2D/3D map space
- Converts similarities between data points to joint probabilities and models high-dimensional points into map points so that the position of the map points aims at conserving the structure of the data
- Similar data points are modeled by nearby map points while dissimilar data points are modeled by distant map points

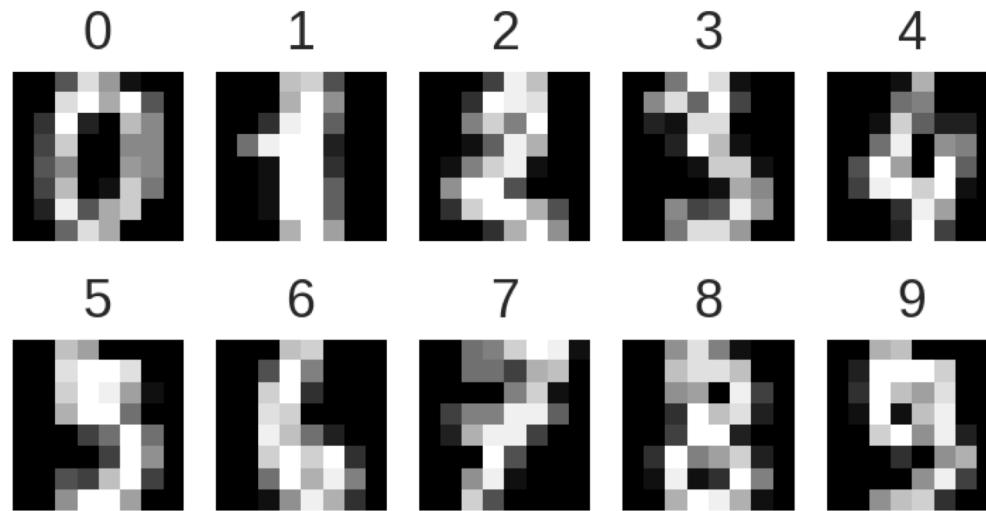
- The t-SNE algorithm has two main steps
 - 1. Define a probability distribution over pairs of high-dimensional data points so that:
 - Similar data points have a high probability of being picked
 - Dissimilar points have an extremely small probability of being picked
 - 2. Define a similar distribution over the points in the map space
 - Minimize the Kullback–Leibler divergence between the two distributions with respect to the locations of the map points
 - To minimize the score, it applies gradient descent

- Assume that map points are all connected with springs.
- The stiffness of a spring connecting two points depends on the mismatch between the similarity of the two data points and the similarity of the two map points
- Let the system evolve according to the laws of physics. If two map points are far apart while the data points are close, they are attracted together. If they are nearby while the data points are dissimilar, they are repelled.
- The final mapping is obtained when the equilibrium is reached.

- Optical Recognition of Handwritten Digits Data Set from the UCI machine learning repository

<http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits>

- Contains 1797 images with 8x8 pixels each



- The parameter perplexity says (loosely) how to balance attention between local and global aspects of the data
- Different initializations will lead to different results
- Should be applied to data with a "reasonable" number of dimensions (e.g. 30-50)
- If the data have more dimensions, another dimensionality reduction algorithm should be applied

Visualization of retweets during
the 2013 Boston Marathon on April 15, 2013

obtained using Gephi and the Twitter API

Run the notebooks for this lecture