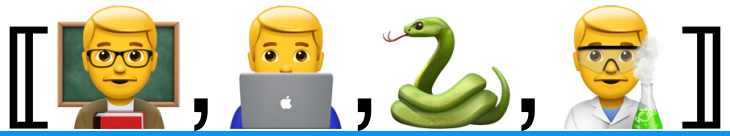
Lecture Notes for

Machine Learning in Python



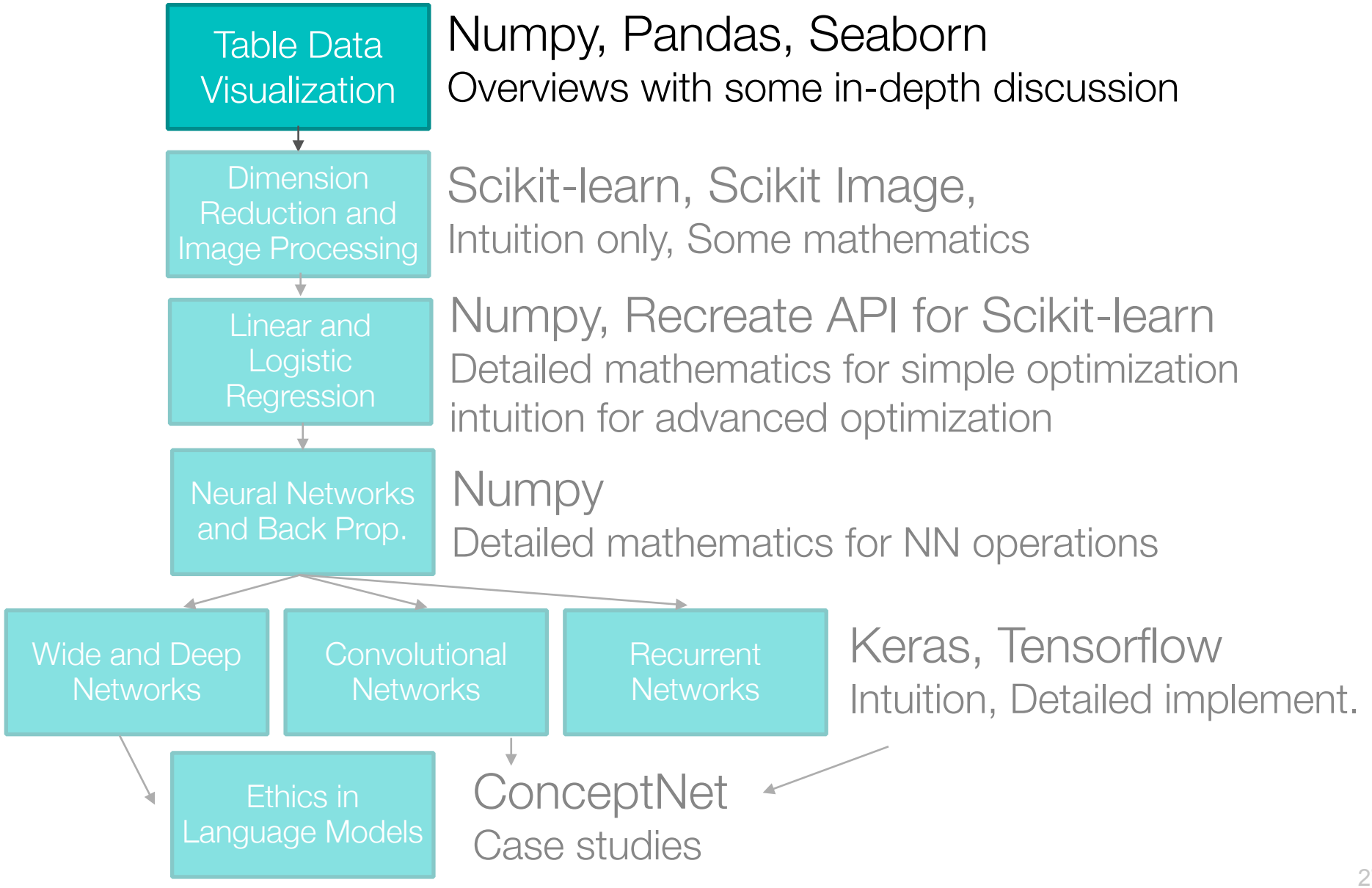
Professor Eric Larson

Visualization

Class Logistics and Agenda

- Logistics:
  - Lab One due very soon!
- Agenda:
  - Finish Visualization Demo
  - Town Hall
- Next Time:
  - Dimensionality Reduction
    - PCA
    - Sampling
  - Images

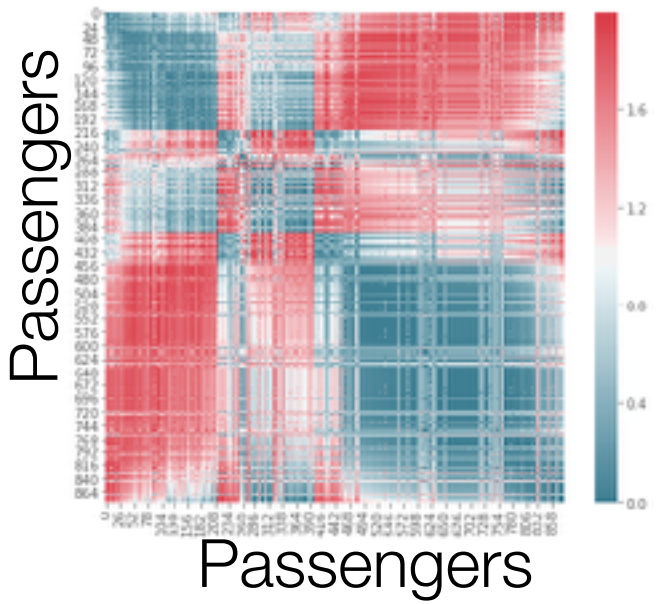
Class Overview, by topic

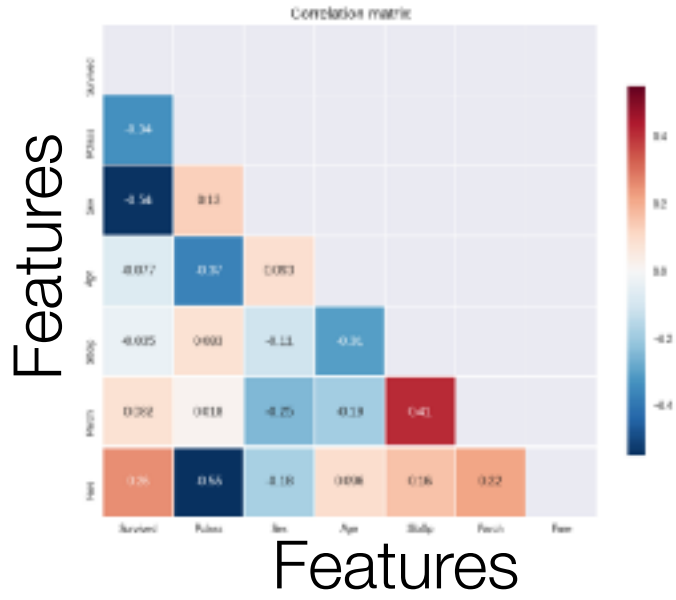


Self Test: What is the main difference in these plots?

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	2	3	male	Q	S
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	0	1	1	0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th	female	38.0	1	0	PC 17599	71.2833	C	0	0	0	0	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	0	1	0	0	1

A. These correlation plots are not different
B. Correlation taken is across rows (left) or columns (right)
C. The colormap in each is different
D. The left plot clusters, the right does not.





Let's look at some graphs

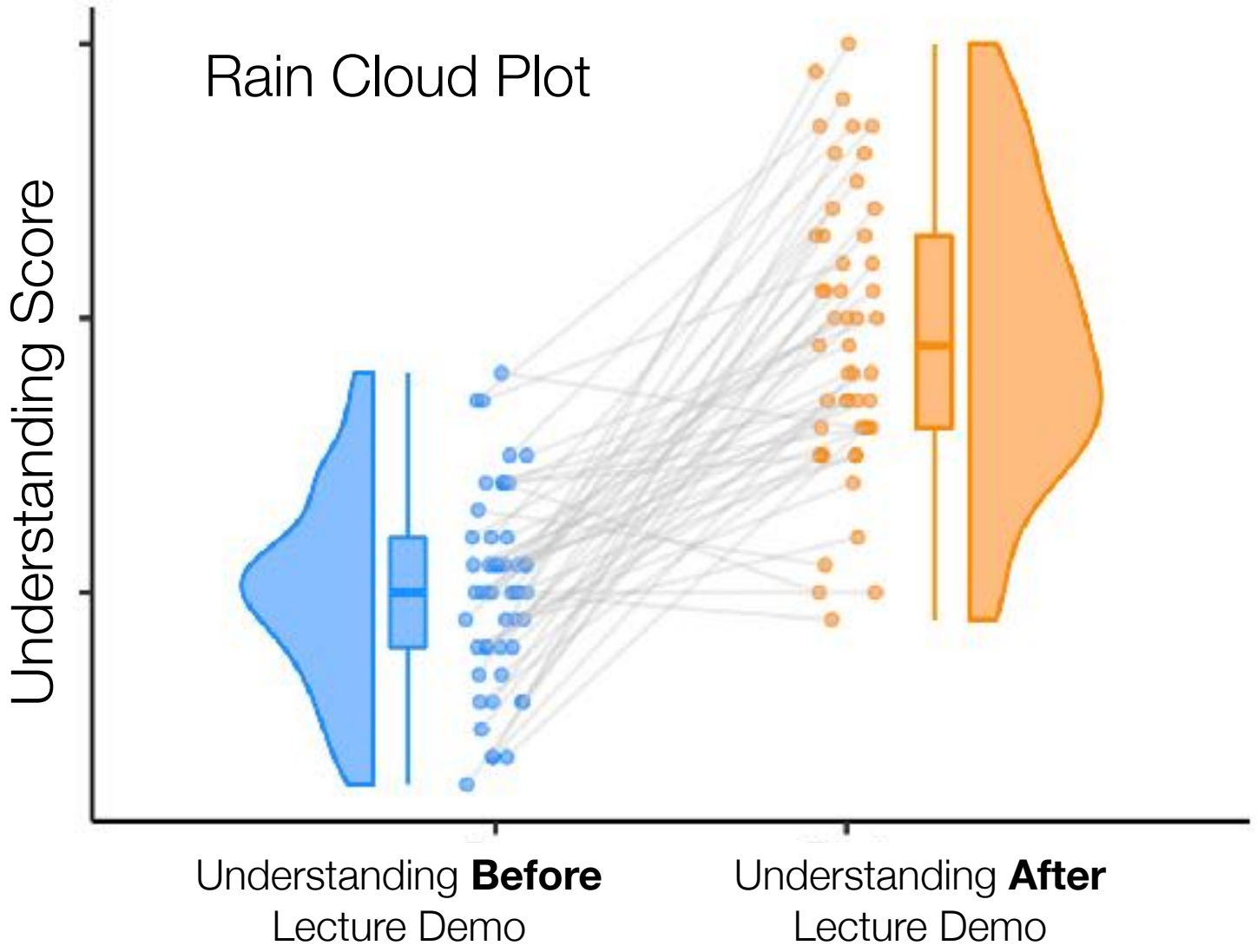
Demo

- You tell me what conclusions we are getting from these graphs
  - Histogram
  - KDE
  - HeatMaps and Correlation
  - Scatter and Scatter Matrix
  - Box / Violin / Swarm

`03.Data Visualization.ipynb`

Matplotlib
Seaborn
Plotly

Now you have visualization building blocks



Lab One: Town Hall

Supplemental Slides

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

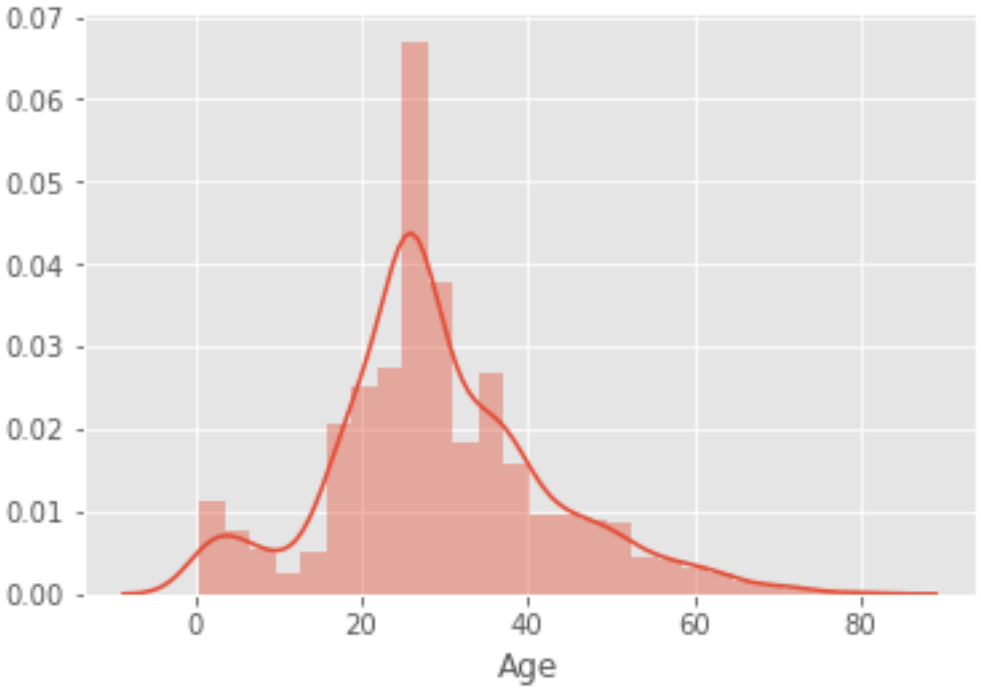
Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

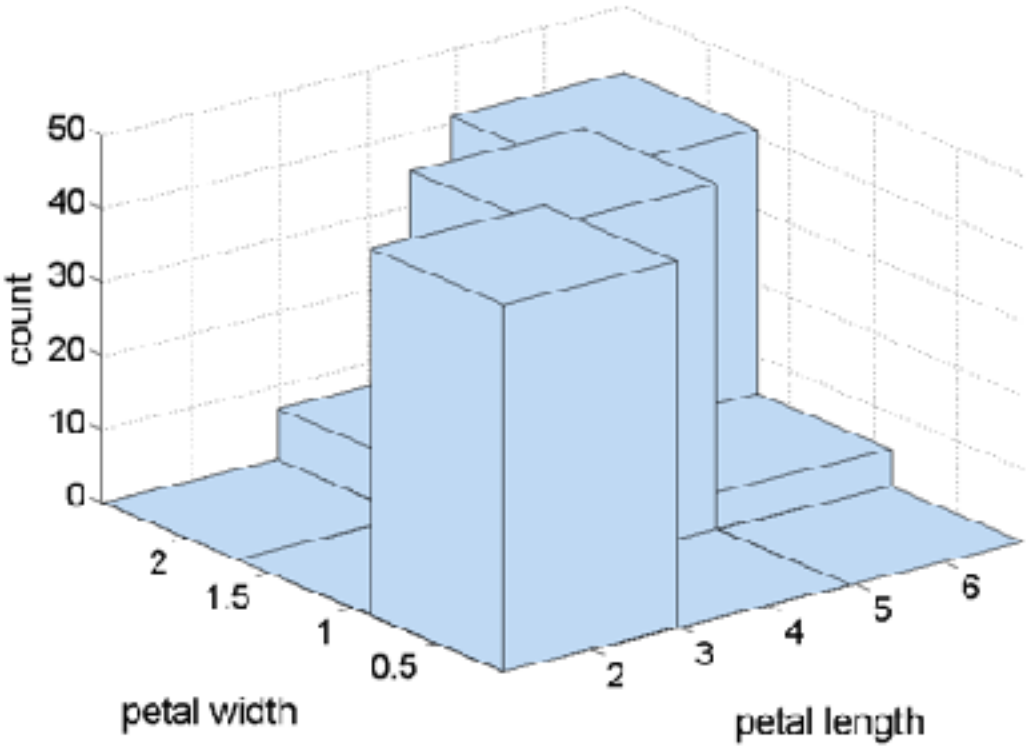
Visualization Techniques: Distributions

- 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.
- KDE
  - Add up Gaussian underneath each point value
  - STD of gaussian is equivalent to number of bins in histogram

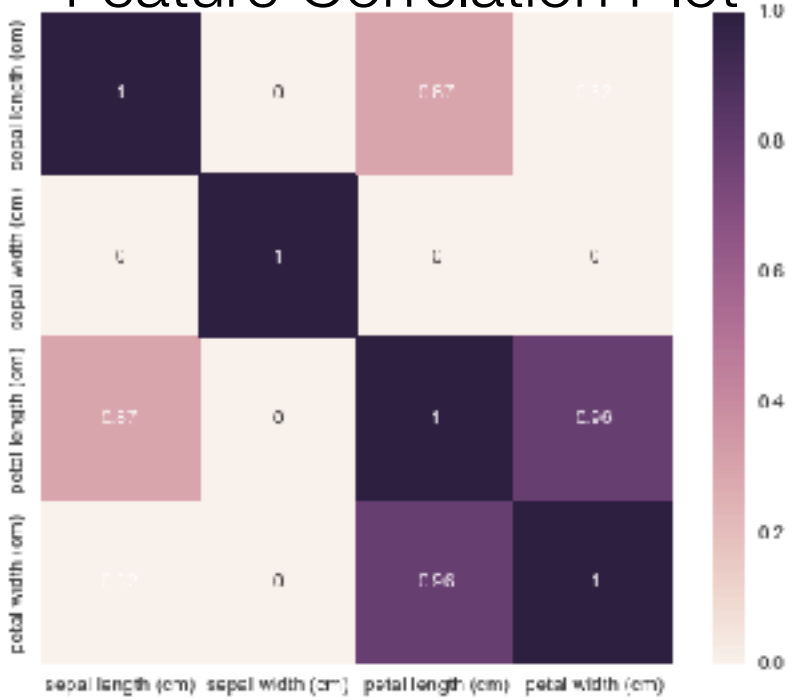


Two-Dimensional Distributions

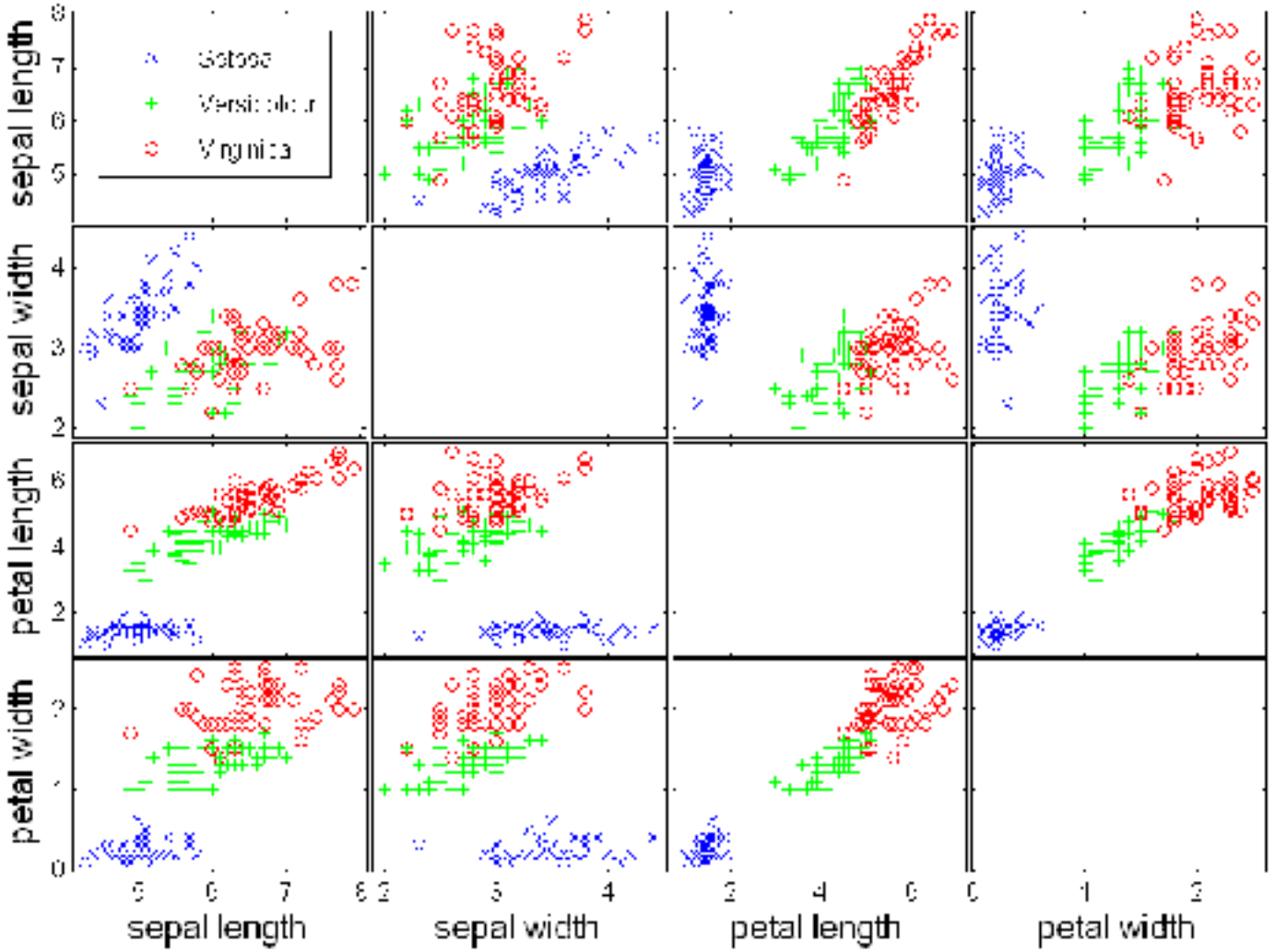
- Estimate the joint distribution of the values of two attributes
- Example: petal width and petal length
  - What does this tell us?



Feature Correlation Plot

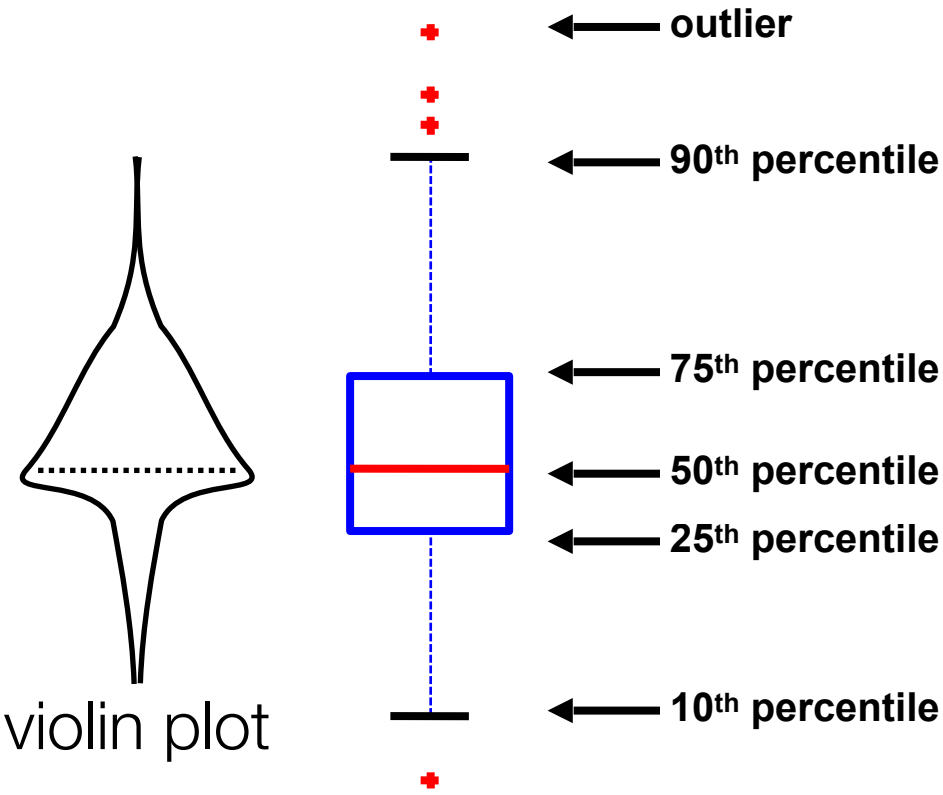


Scatter Plot Matrix Colored by Class



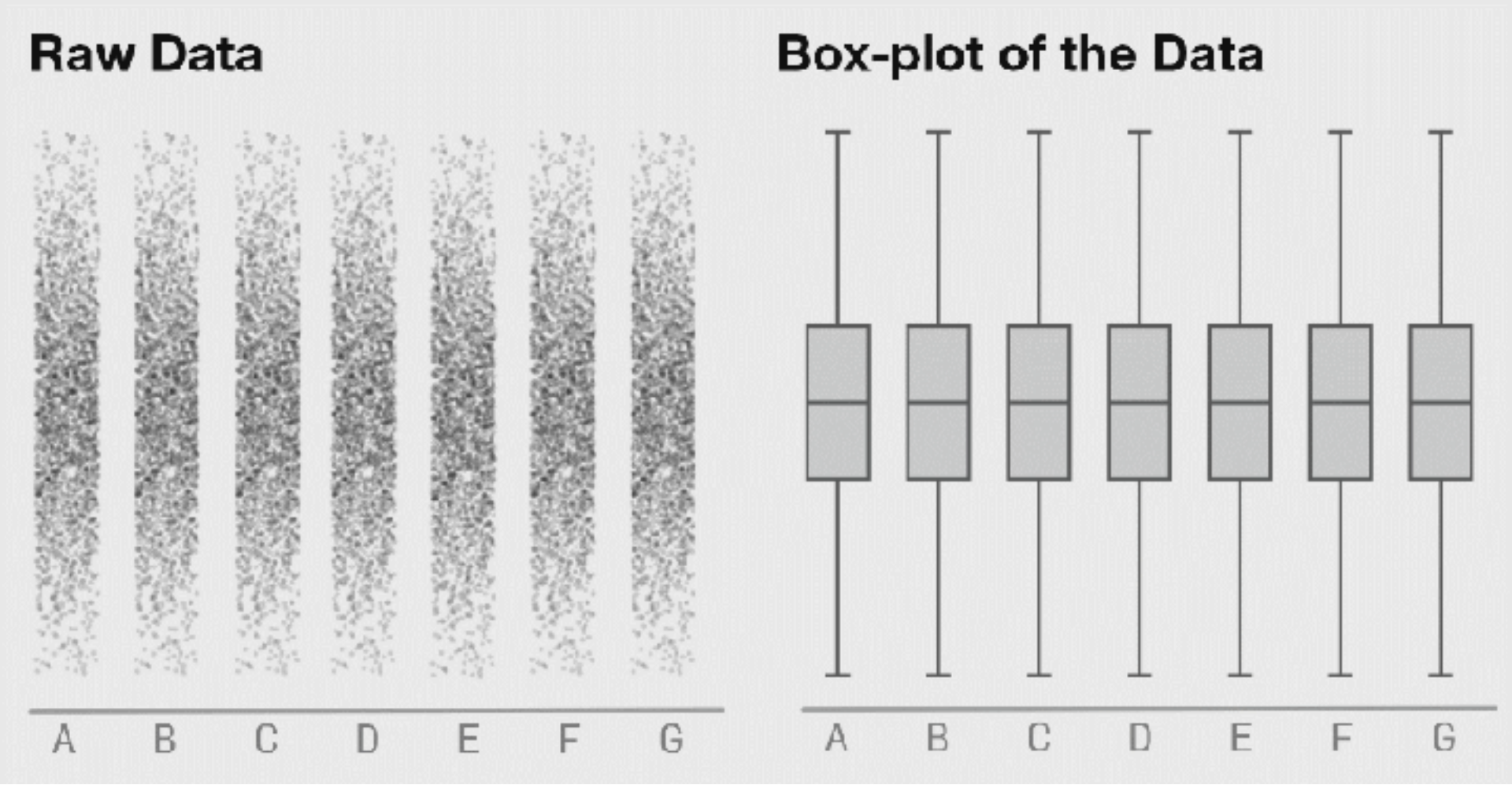
Visualization Techniques: Box Plots

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



Visualization Techniques: Box Plots

- Box Plots

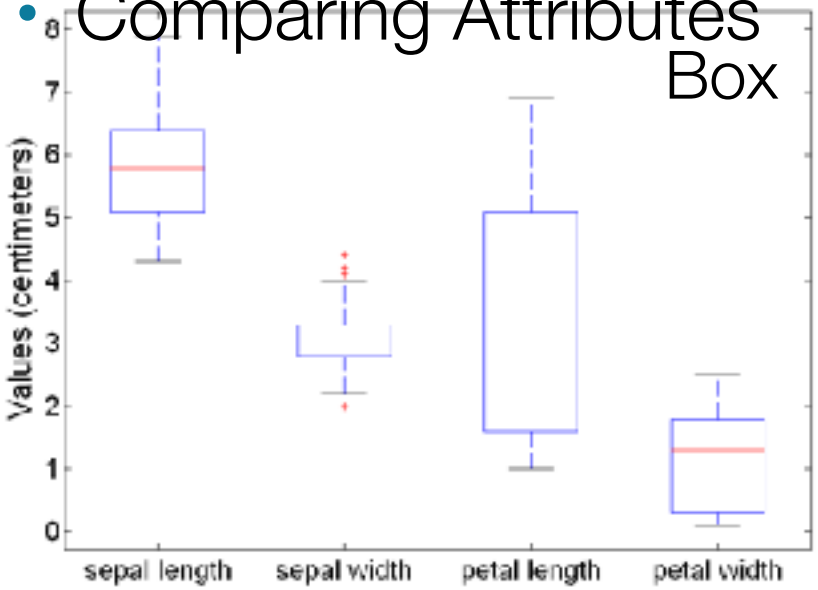


https://www.reddit.com/r/dataisbeautiful/comments/6ssmzc/be_wary_of_boxplots_they_could_be_hiding/

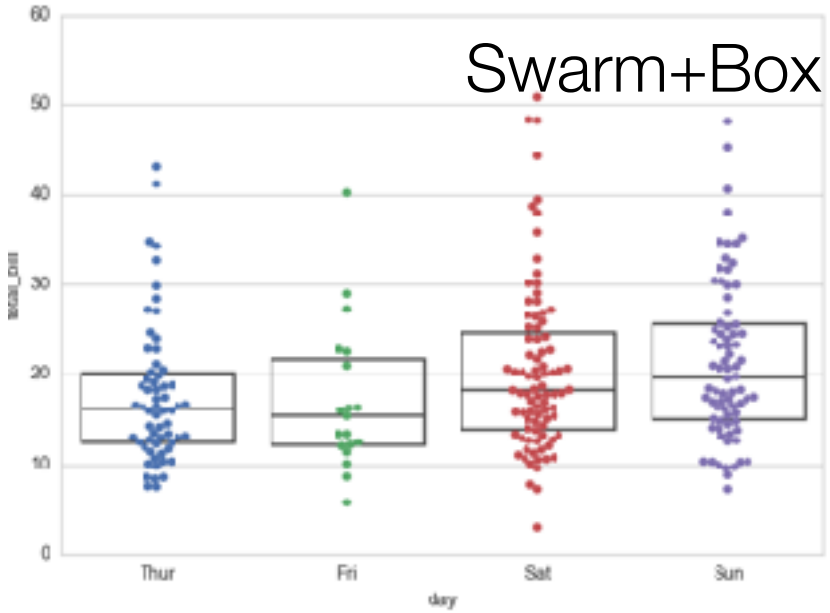
Example: Comparing Attributes

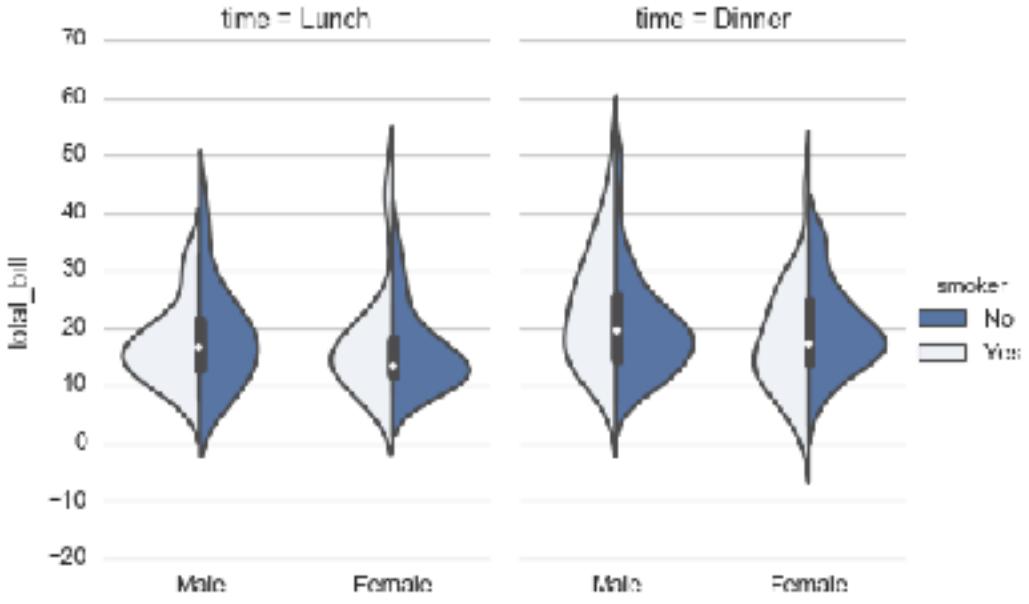
- Comparing Attributes

Box



Swarm+Box





Mixed Violin + Box