Dictionary

○ Look-up: O(1)

○ Length: O(1)

○ Insert: O(1)

○ Delete: O(1)

○ dictionary.keys(): O(n) - because a list is generated

○ Check if a key is in the dictionary: O(1)

List

○ Append: O(1)

○ Length: O(1)

○ Insert: O(n)

○ Delete: O(n)

○ Copy: O(n)

○ Sort: O(n log n)

○ Check if an item is in the list: O(n)

Total Time = Time per Iteration \* # of Iterations or Time per Call \* # of Calls

Linear search

○ Brute force search

○ List doesn’t have to be sorted

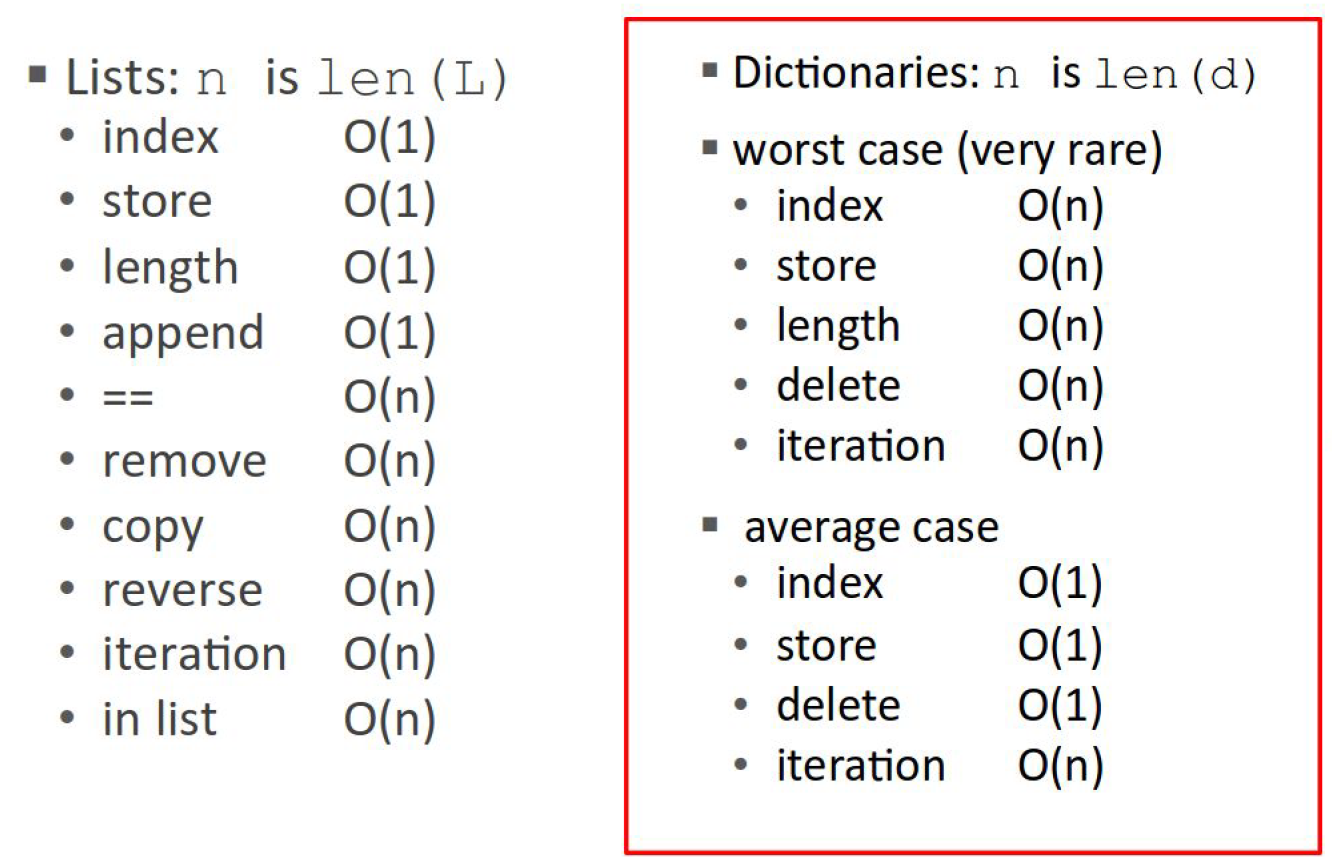
○ O(n)

● Bisection search

○ List must be sorted to give correct answer

○ O(log n)

Merge Sort ○ How many levels of the recursive tree? O(log n) ○ How much computation of each level of the tree? O(n) ○ Complexity? O(n log n)

If all keys hash to the same index Hash table large relative to number of keys Hash function good enough

A good hashing function should

● Be deterministic, i.e. it should not use any randomness (why?)

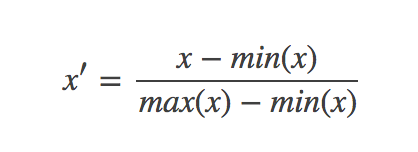
● Use the entire input in the hash computation

● Should map possible inputs to outputs uniformly (why?)

Feature scaling

● A way to standardize the range of features of data

● Allows us to have each number contribute equally to objective function

● Z-scaling

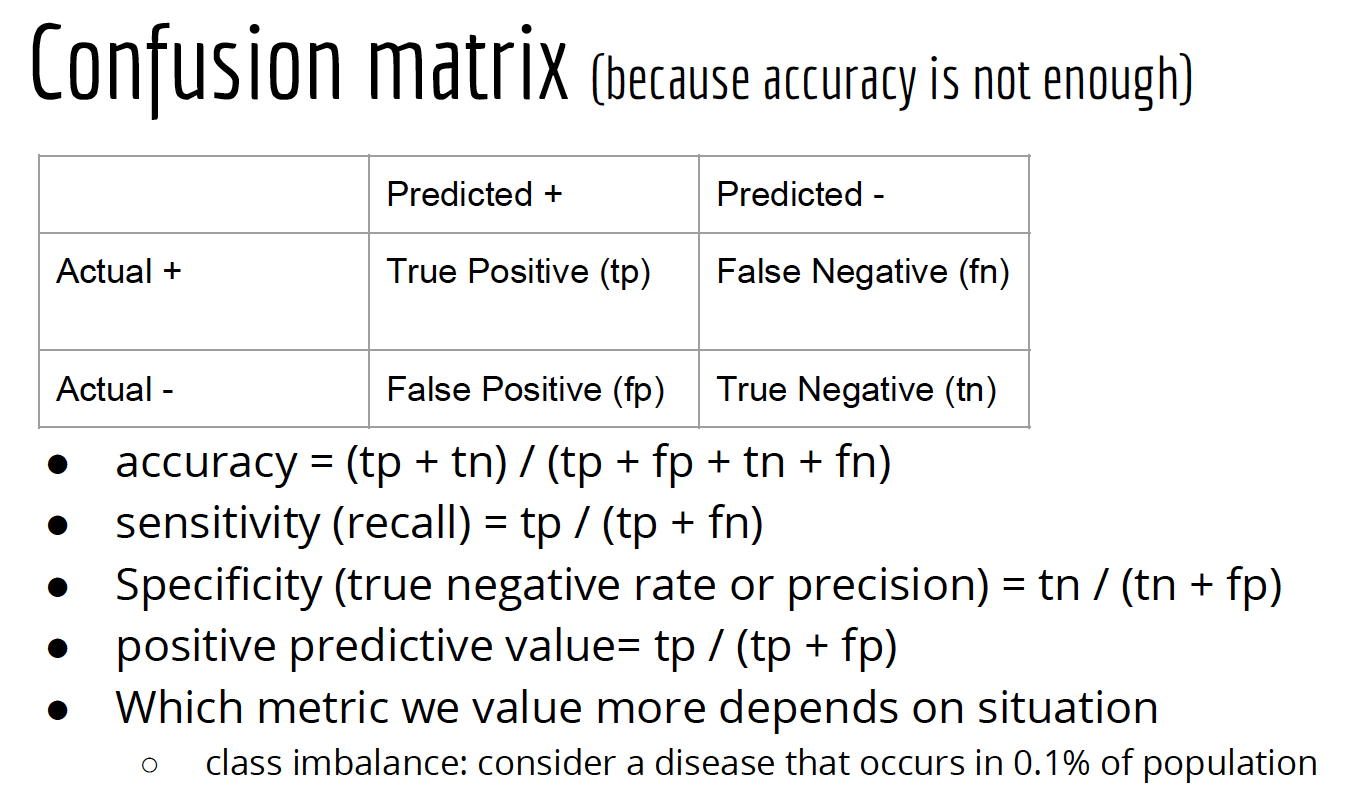
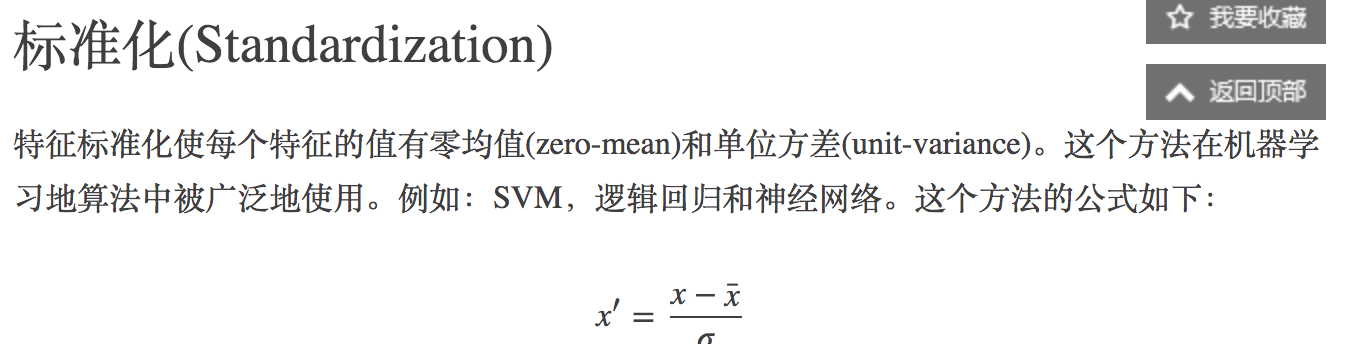
○ each feature has an average of 0 and a standard deviation of 1

● Interpolation

○ map min to 0, max to 1, and linearly interpolate

调节比例（Rescaling）

这种方法是将数据的特征缩放到[0,1]或[-1,1]之间。缩放到什么范围取决于数据的性质。对于这种方法的公式如下：



Datasets and Validation

● Split into training, validation (to help choose parameter values), and

testing datasets.

● When we don’t have a huge amount of data we use cross-validation

○ Repeated Random Sampling - Randomly split data into testing and training datasets k

times. For each train its own model, and take the average of their error rates.

○ K-fold Cross-Validation - Divide data into N/k testing datasets. Train k models that each

use one of the N/k testing datasets and all the remaining data for training. Compute the

average error rate.

○ Leave-one-out - Same as above, but k = N



Supervised learning

● Each data point = (feature vector, annotation)

● Given: points + discrete labels (Classification)

○ Goal: predict labels for new points

● Given: points + real-valued scores (Regression)

○ Goal: predict scores for new points

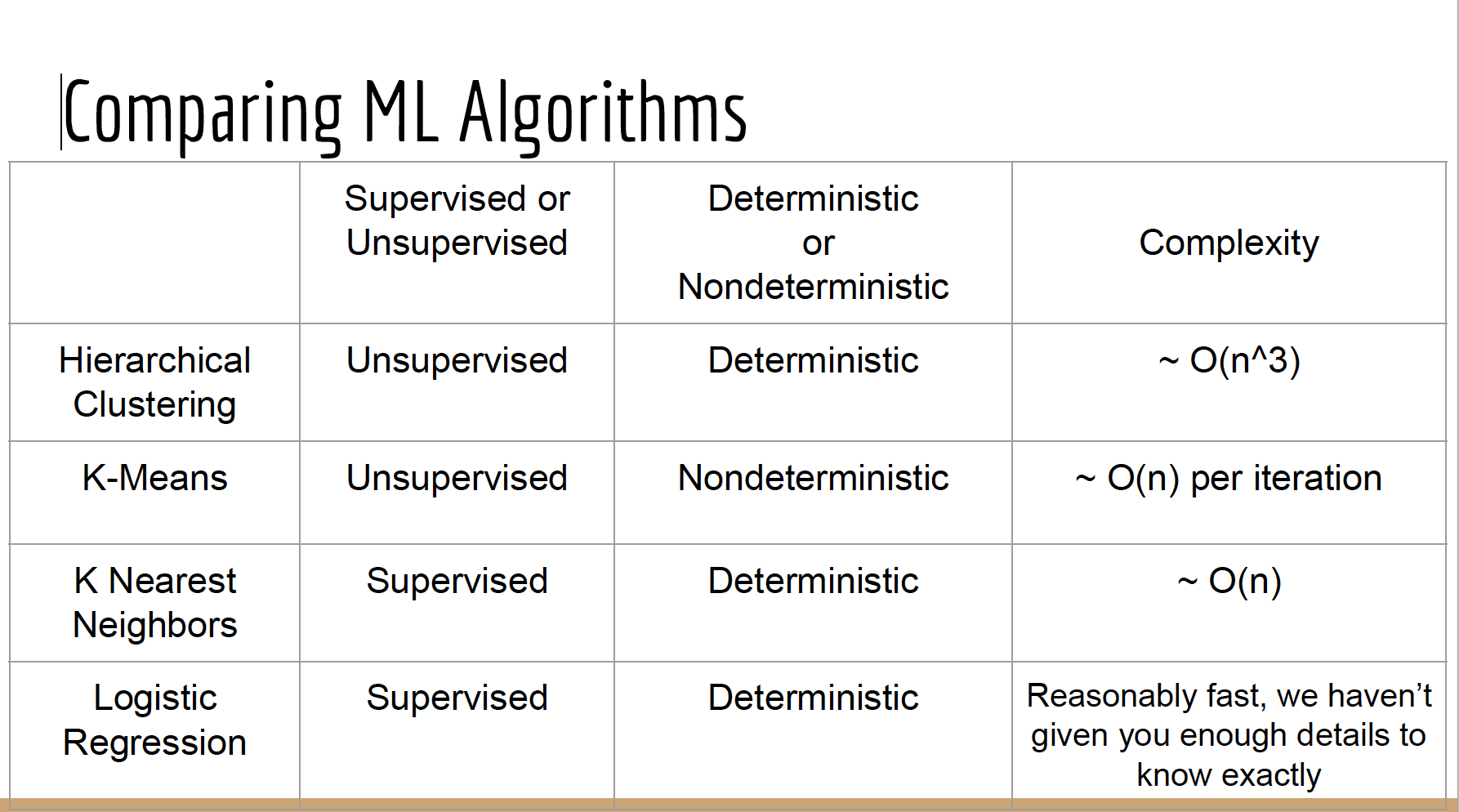
● Logistic Regression is useful when we want to predict

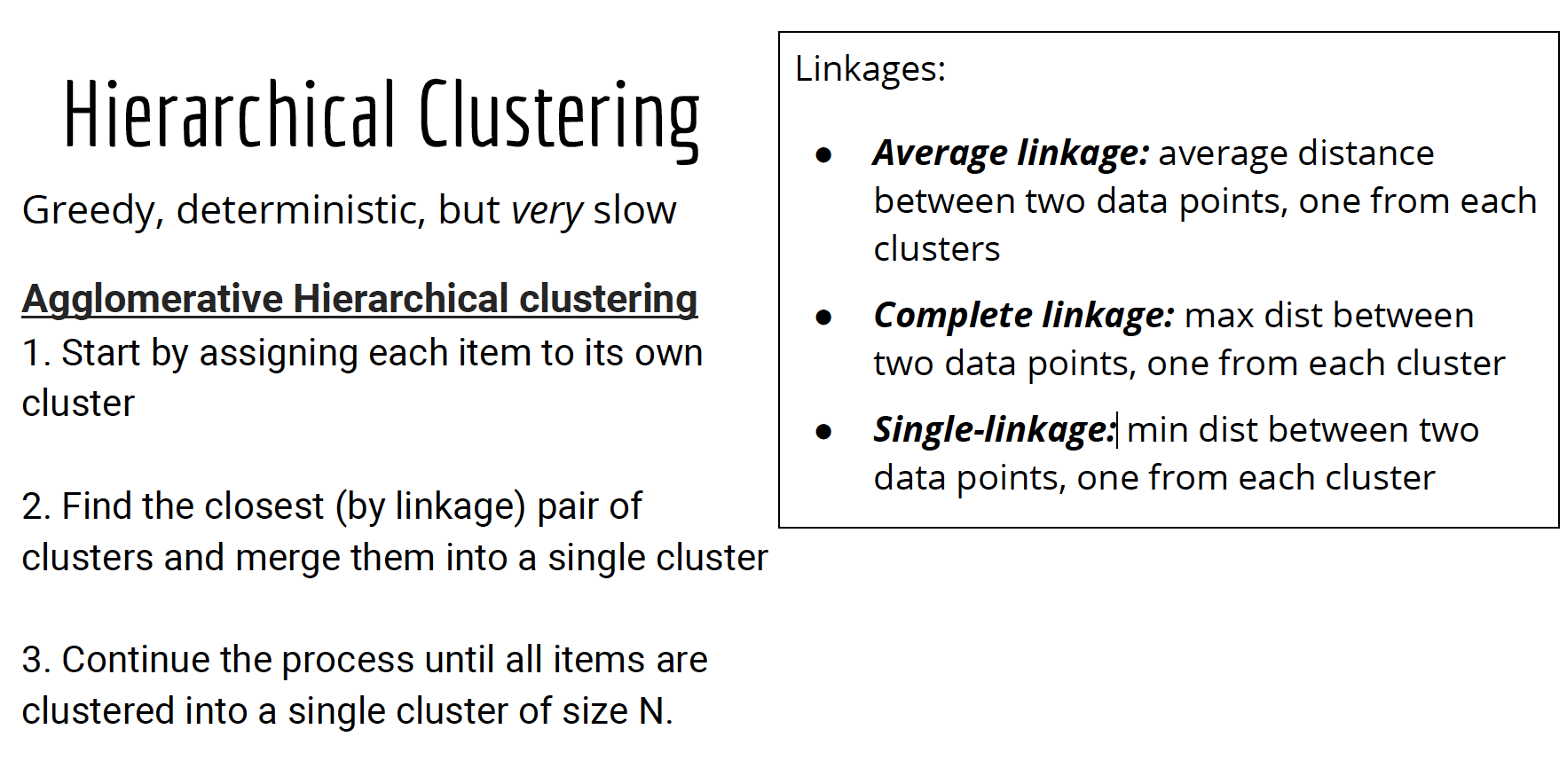
continuous values.

FALSE

● True or False: KNN is a deterministic algorithm. TRUE

● The results of hierarchical clustering depend on the linkage criteria used.

TRUE

● In hierarchical clustering there are O(n) iterations (combinations of

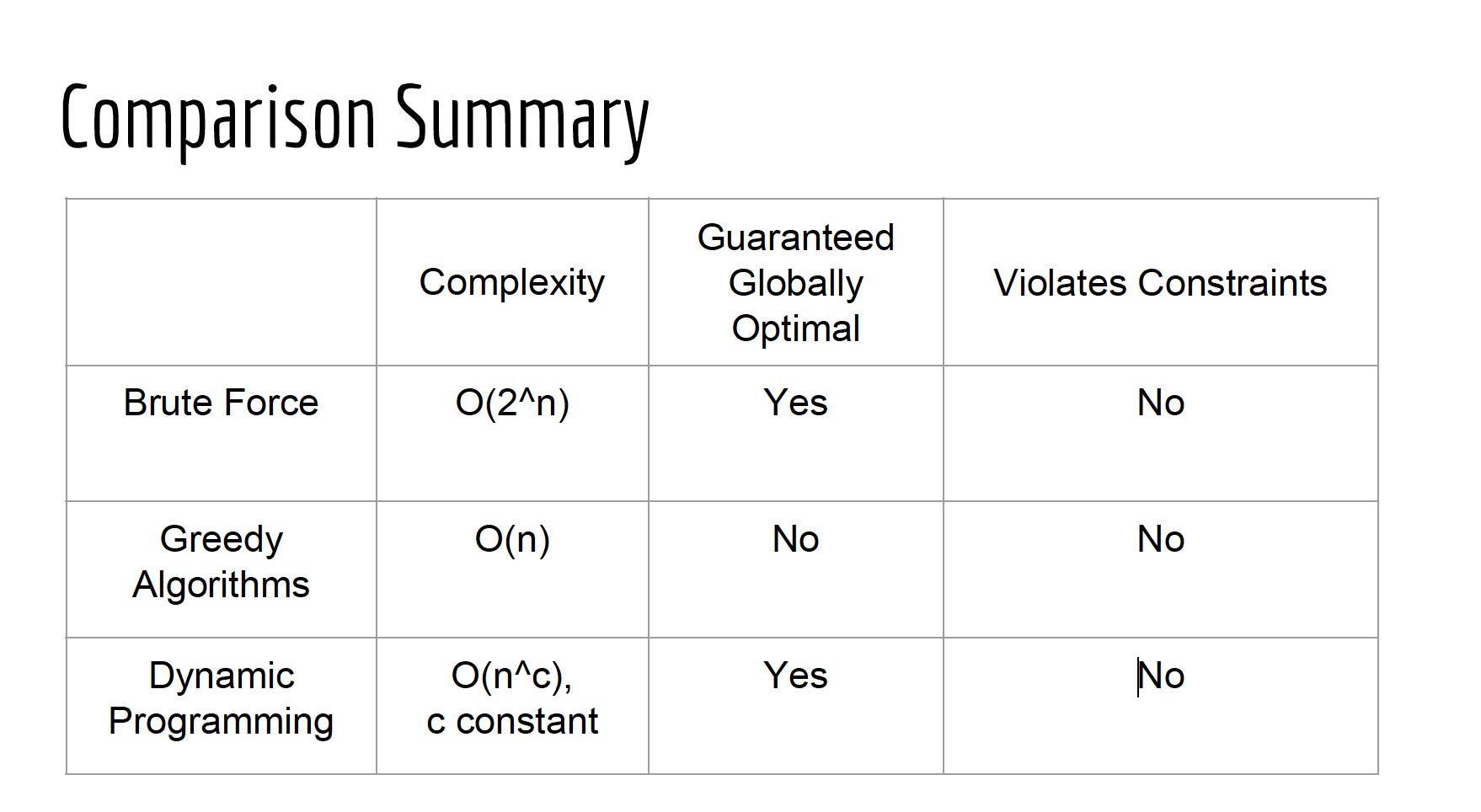
clusters) before the algorithm terminates. TRUE

● The final results of k-means depends on the initial centroids TRUE

● Training an algorithm on data set A and testing it on the same data set A

is an example of supervised learning. FALSE

● Hierarchical and k-means clustering are both greedy algorithms. TRUE

All brute force algorithms are exponential. FALSE

● Dynamic programming can be used to reduce the order of algorithmic complexity of

sorting a list of integers to something below n log n, where n is the length of the list to

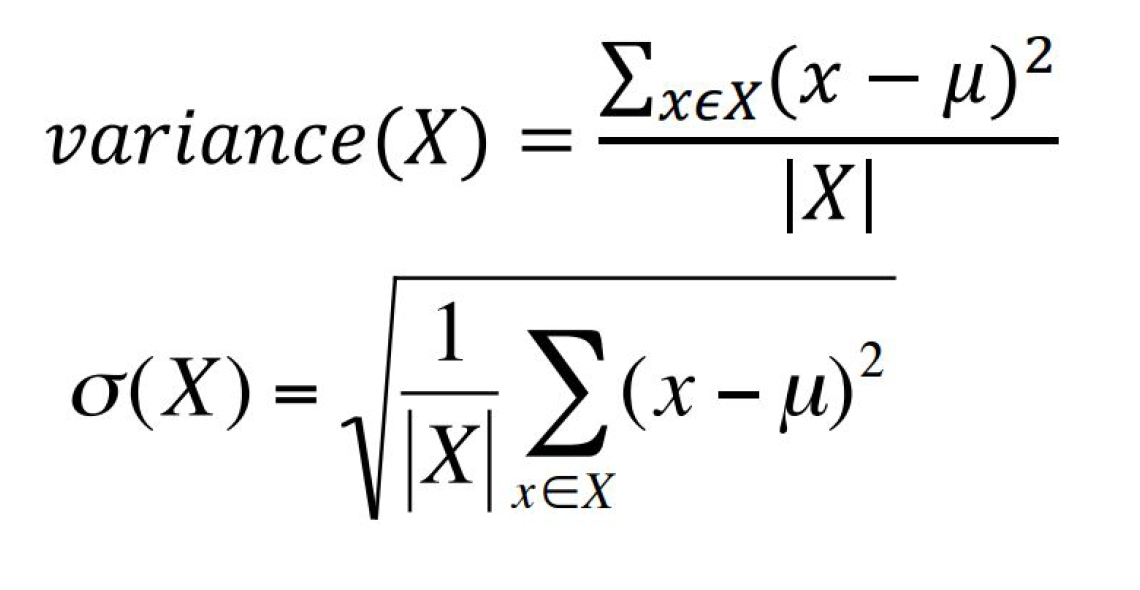
be sorted. FALSE

● Dynamic programming can be used to solve any problem that a greedy algorithm can

be used for. FALSE

Breadth vs. depth-first search

○ BFS (queue, FIFO): guaranteed shortest path for unweighted graphs

○ DFS (stack, LIFO): returns a path (not necessarily optimal)

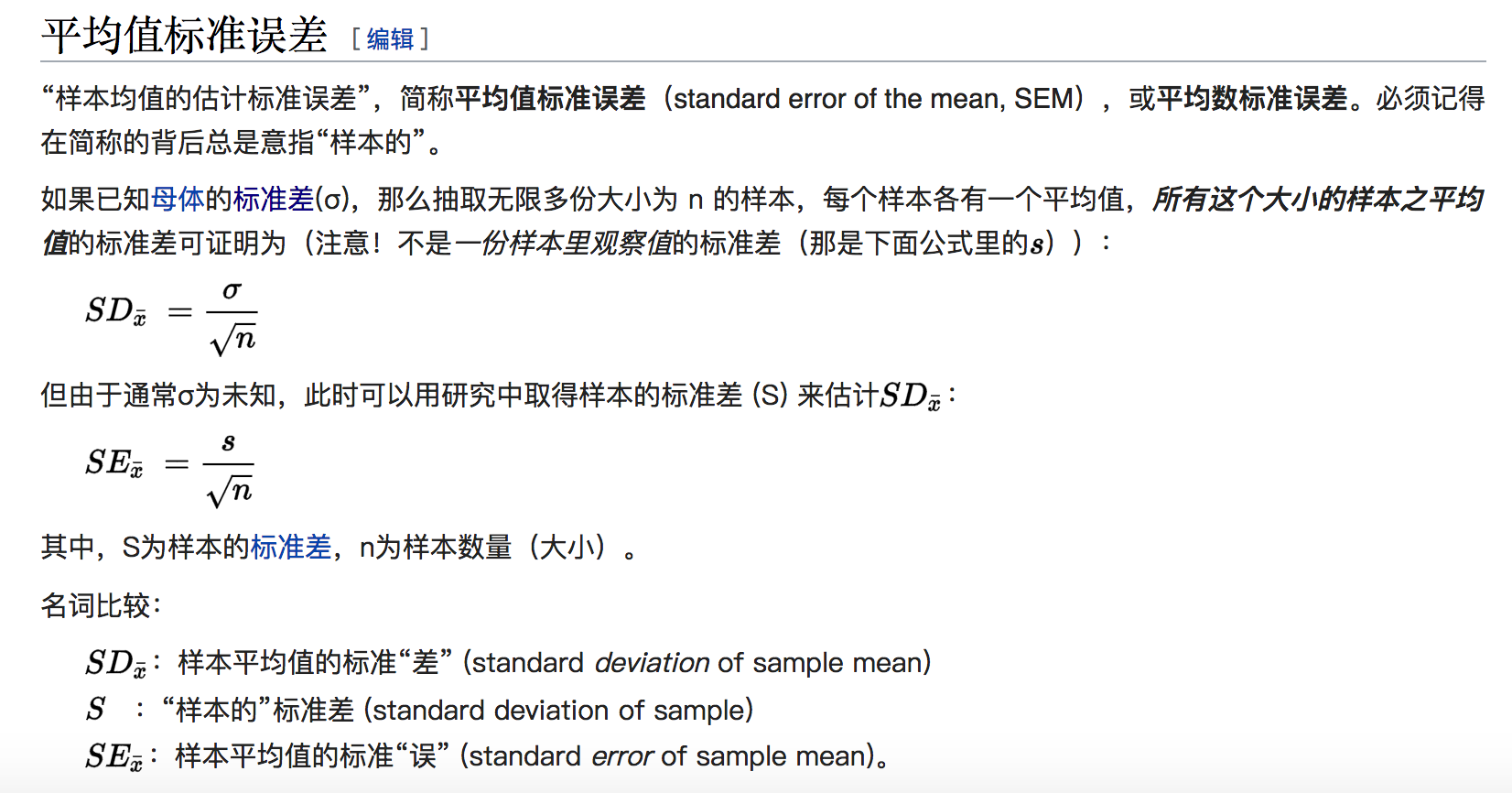
Standard Error of the Mean (SEM)

● SEM = σ/√n (σ is population s.d. -> unknown)

● Approximation

○ With one sample: s/√n

■ s: sample s.d. (s.d. of examples in this sample)

■ Reasonable approximation when n is large enough

○ Many samples (m): CLT

■ For large enough m and n, mean of these m samples will be

approximately normally distributed, with mean close to population

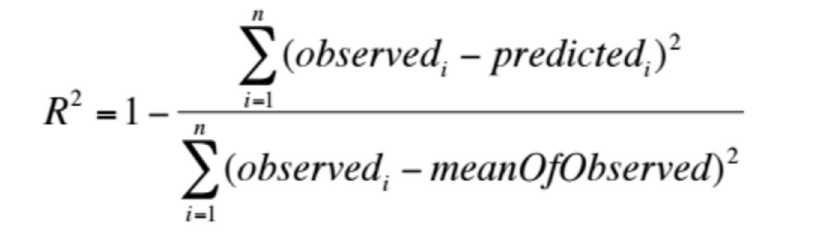
mean and s.d. close to σ/√n

■ Use s.d. of these samples as SEM

Central Limit Theorem

1) The means of the samples in a set of samples (the sample means) will be

approximately normally distributed,

2) This normal distribution will have a mean close to the mean of the

population, and

3) The variance of the sample means will be close to the variance of the

population divided by the sample size.

“Empirical rule” allows us to relate

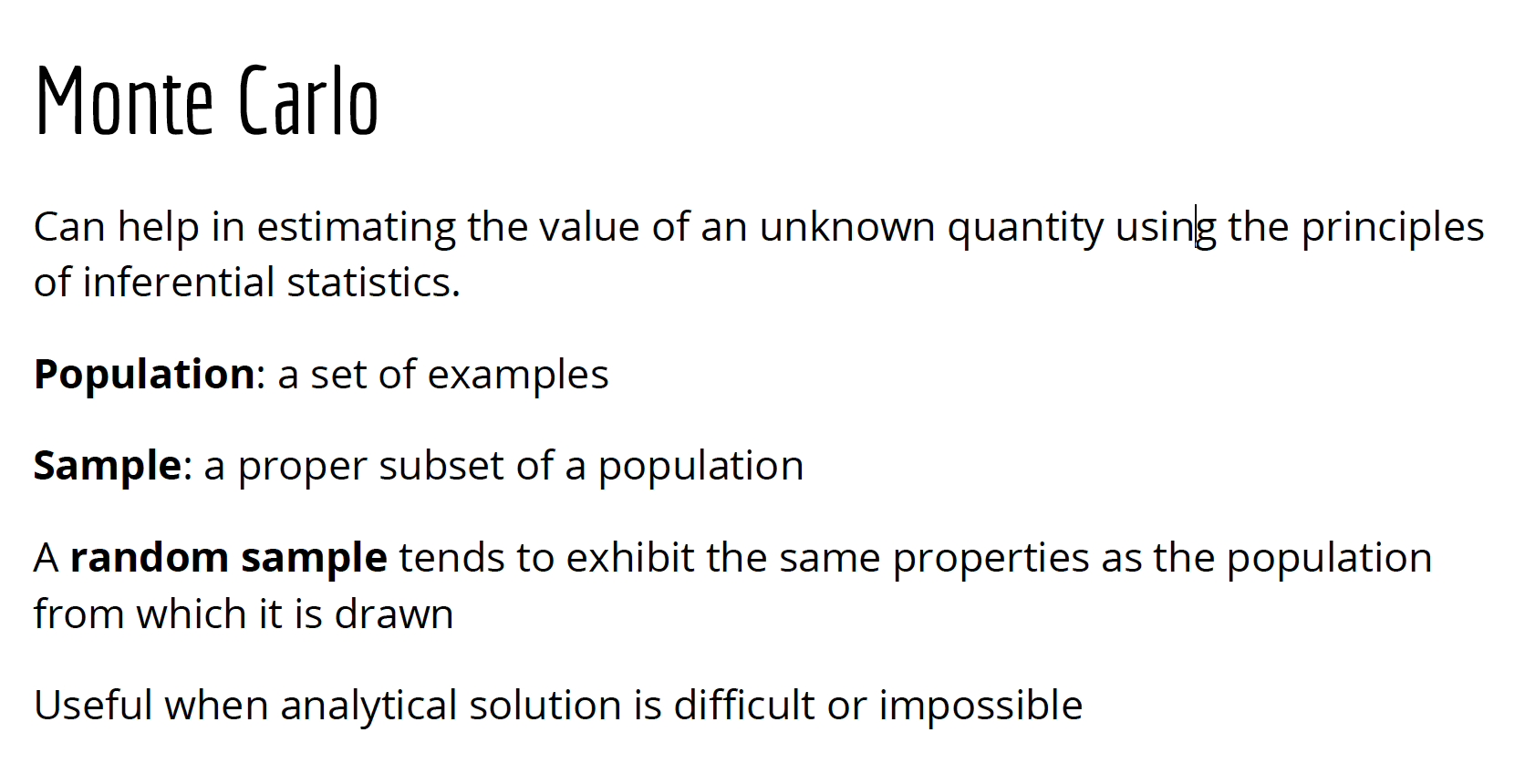
data points to overall structure

- 68-95-99.7

● The standard deviation fully defines a uniform distribution FALSE.

● An unfair coin with p(heads) = 3/4 is tossed 10 times. What is the probability

that exactly 6 heads will occur? 10c6 \* (3/4)^6 \* (1/4)^4



Lies, damned lies, and statistics

● Be careful in making conclusions from your data or reading others’

conclusion

● Truncating y-axis

● Non-representative Sampling

○ Land-lines

○ Course evaluations

○ Survivor’s Bias