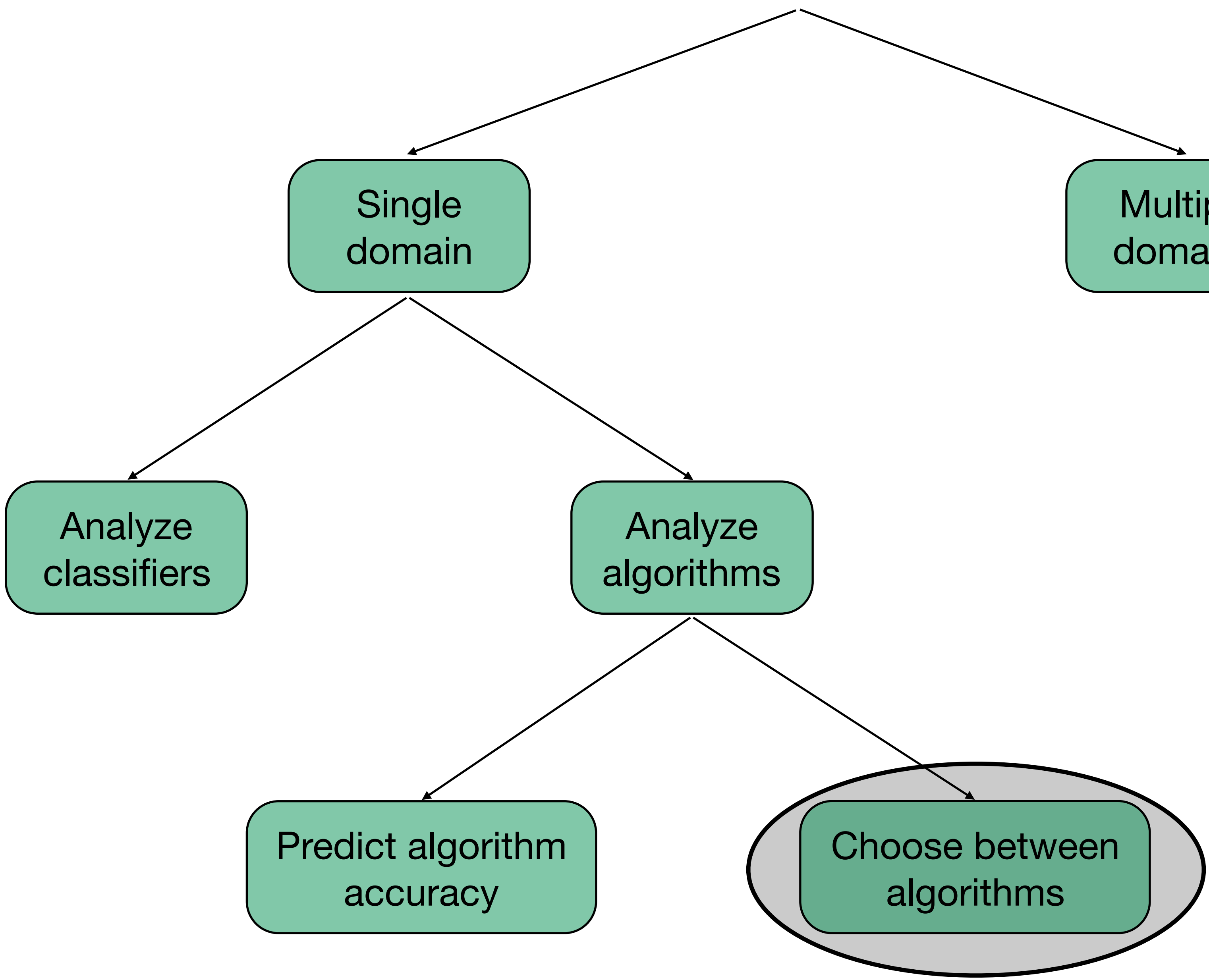
Data Mining & Machine Learning

CS37300
Purdue University

October 4, 2017

Predictive modeling: evaluation

Statistical questions in machine learning *(Dietterich '98)*



Algorithm comparison

- How to compare the performance of two learning algorithms A and B?
  - Estimate the expected value of the difference in errors, where expectation is over all datasets D of size N
  - In practice, we only have a limited sample D_0

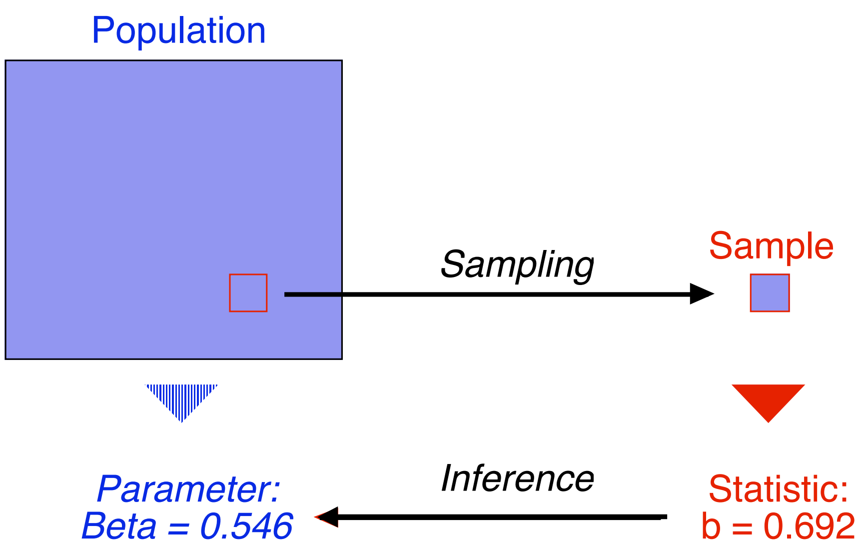
Populations and samples

Populations and samples

- In data mining we often work with a sample of data from the population of interest

- **Estimation** techniques allow inferences about population properties from sample data

- If we had the population we could **calculate** the properties of interest



Populations and samples

- Elementary units:
  - Entities (e.g., persons, objects, events) that meet a set of specified criteria
  - Example: All people who've purchased something from Walmart in the past month
- Population:
  - Aggregate of elementary units (i.e, all items of interest)
- Sampling:
  - Sub-group of the population
  - Serves as a reference group for estimating characteristics about the population and drawing conclusions

Sampling

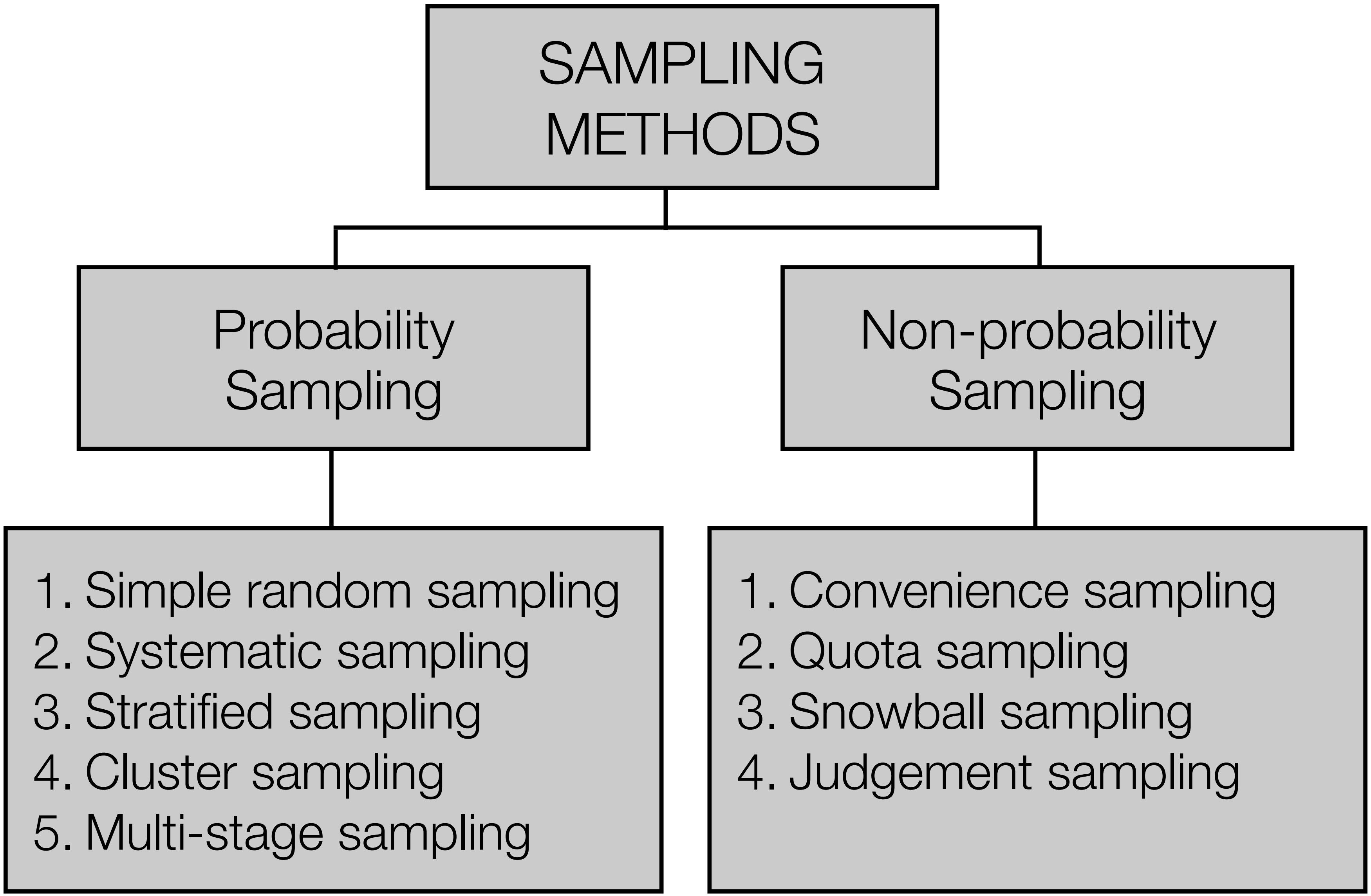
- Sampling is the main technique employed for data selection
    - It is often used for both the preliminary investigation of the data and the final data analysis
- Reasons to sample
    - Obtaining the entire set of data of interest is too expensive or time consuming
    - Processing the entire set of data of interest is too expensive or time consuming
    - Note: Even if you use an entire dataset for analysis, you should be aware of the sampling method that was used to gather the dataset

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Sampling …

- The key principle for effective sampling is the following:
    - Using a sample will work almost as well as using the entire data set, if the sample is **representative**
    - A sample is representative if it has approximately the same property (of interest) as the original set of data

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

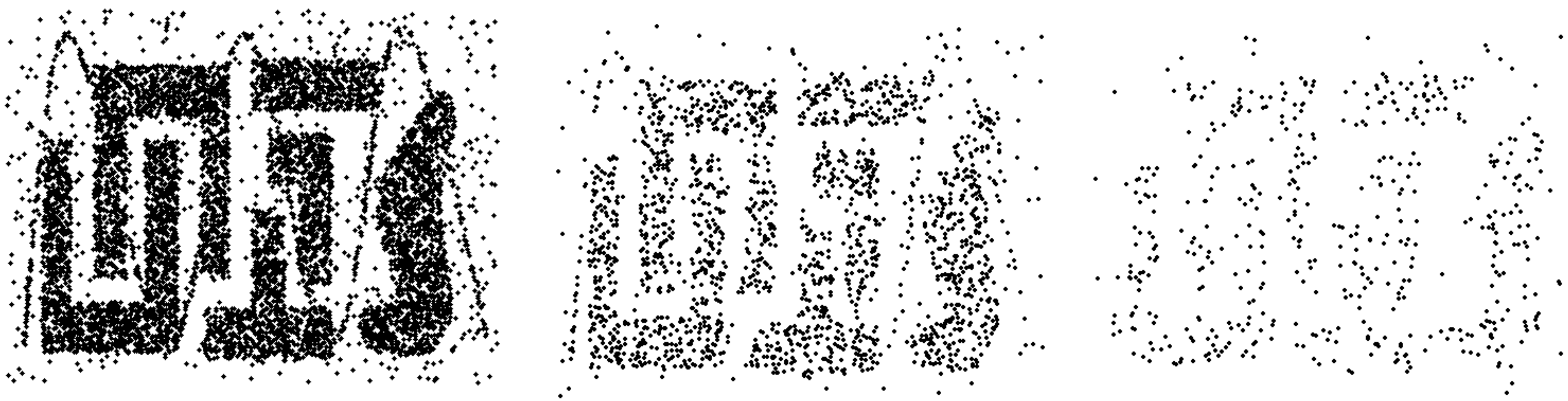

SAMPLING METHODS
Probability Sampling
Non-probability Sampling
1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multi-stage sampling
1. Convenience sampling
2. Quota sampling
3. Snowball sampling
4. Judgement sampling

Types of probability sampling

- Simple random sampling
  - There is an equal probability of selecting any particular item
- Sampling without replacement
  - As each item is selected, it is removed from the population
- Sampling with replacement
  - Items are not removed from the population as they are selected for the sample; the same item can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

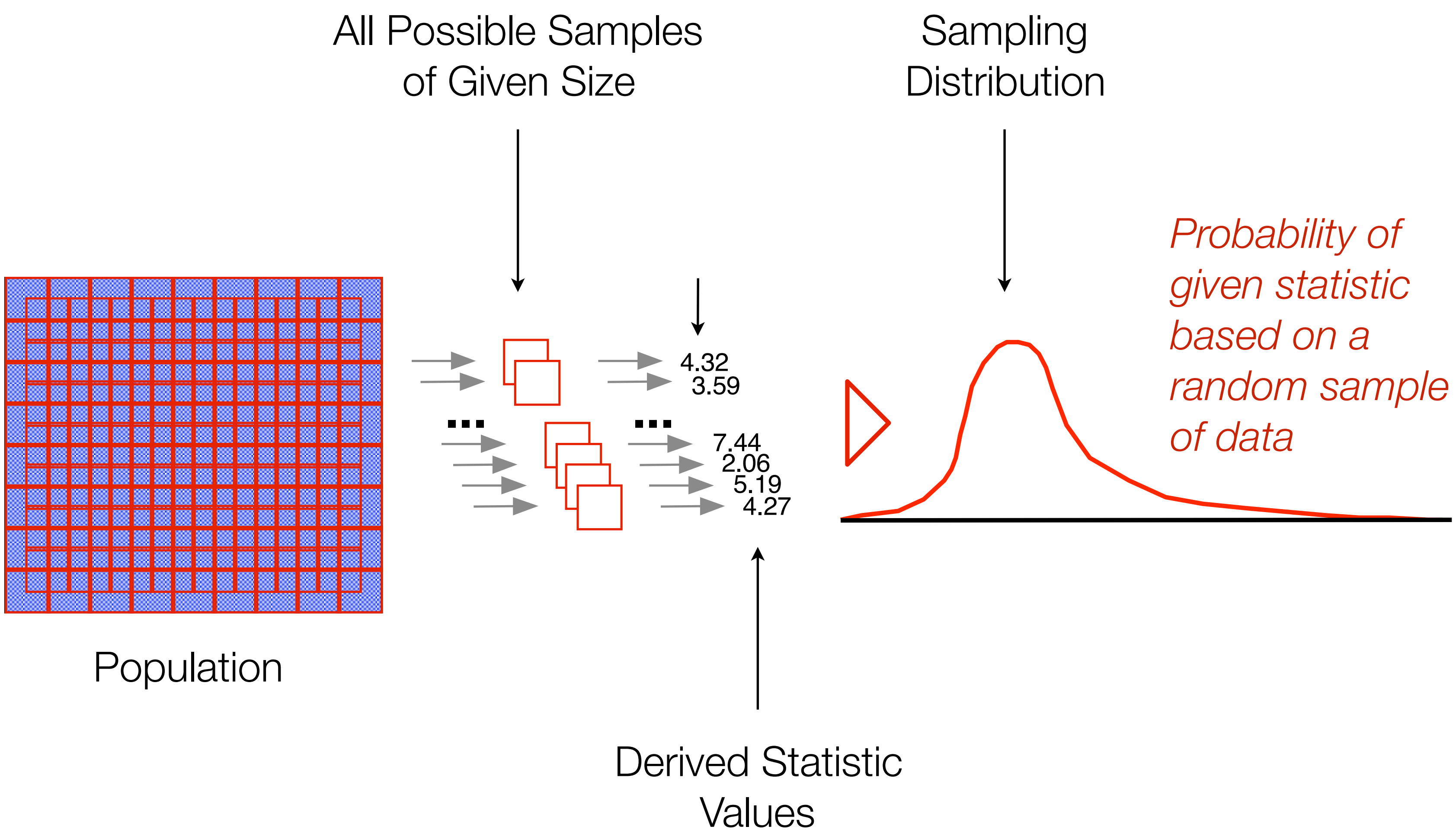
How does sample size affect learning?



8000 Points **2000 Points** **500 Points**

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Sampling distributions

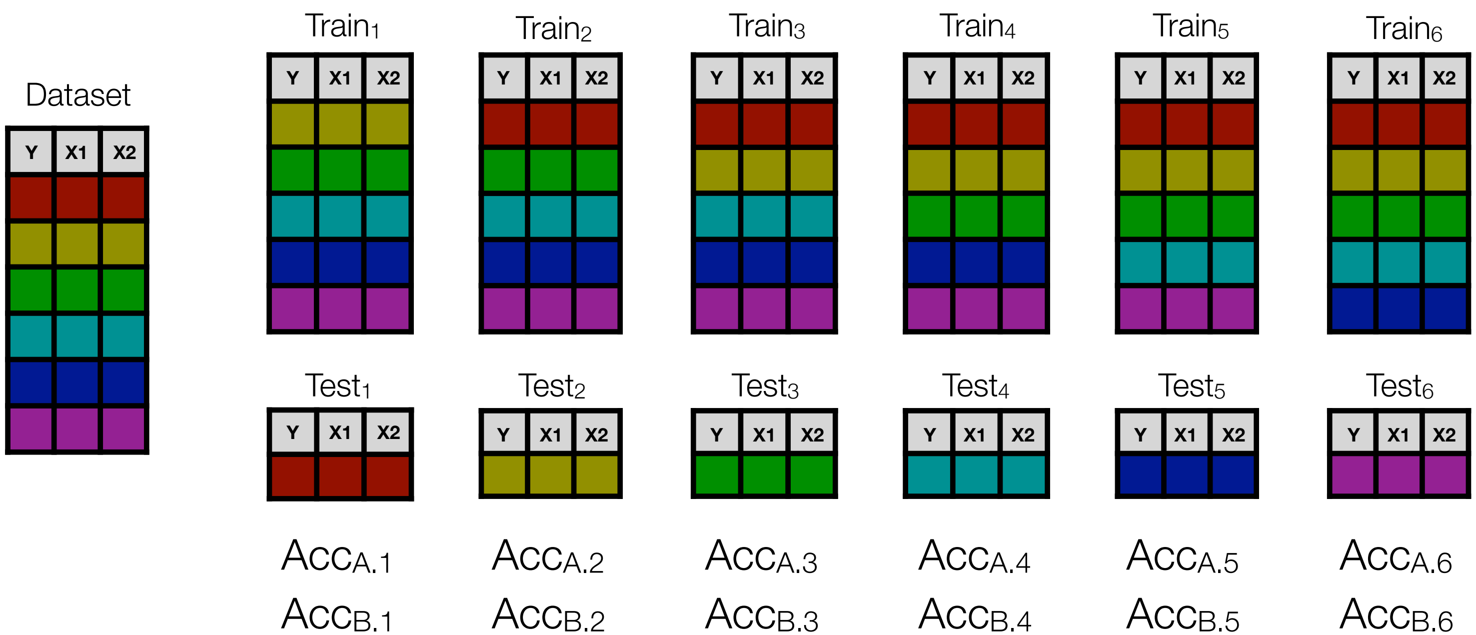


Let’s return to algorithm comparison...

- How to compare the performance of two learning algorithms A and B?

Comparing algorithms A and B

- Use k-fold cross-validation to get k estimates of error for M_A and M_B



- Set of errors estimated over the test set folds provides empirical estimate of sampling distribution
- Mean is estimate of expected error

Assessing significance

- Use **Bayesian hypothesis tests (via simulations)** to assess whether the two distributions of errors are statistically different from each other

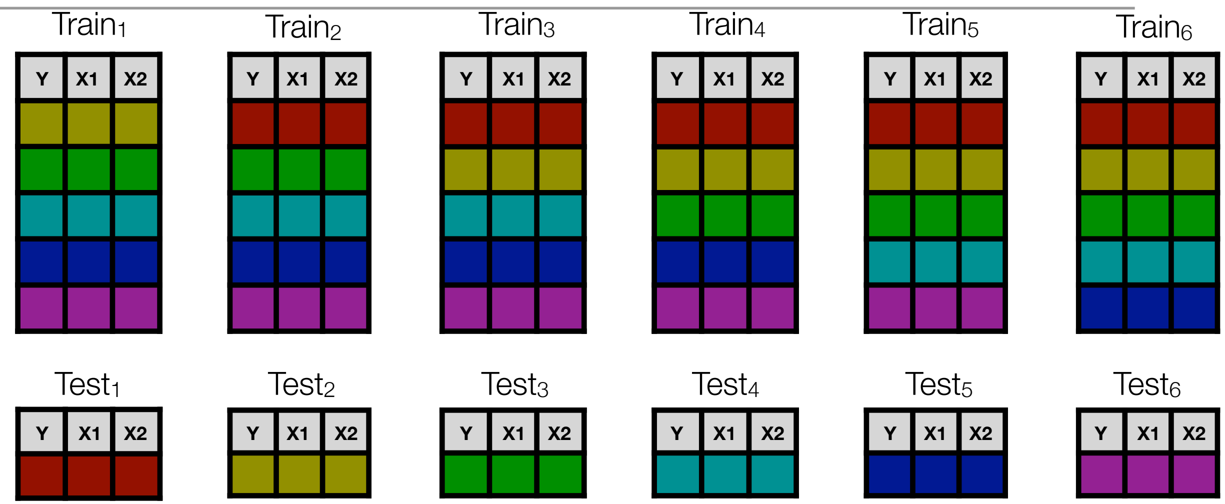
- What is the hypothesis? We want to check whether accuracy of A is better than B

  - Hypothesis: E[X] > 0?

    - E[X] > 0, but algo A loses 99% of the time and wins big in the 1% it wins…

  - Better hypothesis
    $H_0 := P[A \text{ wins } B] > 0.5$
    $:= P[X_7 > 0] > 0.5$

  - How can we test this hypothesis?



$Acc_{A.1}$	$Acc_{A.2}$	$Acc_{A.3}$	$Acc_{A.4}$	$Acc_{A.5}$	$Acc_{A.6}$
$Acc_{B.1}$	$Acc_{B.2}$	$Acc_{B.3}$	$Acc_{B.4}$	$Acc_{B.5}$	$Acc_{B.6}$

Compute the differences:

$X_1 = Acc_{A.1} - Acc_{B.1}$

$X_2 = Acc_{A.2} - Acc_{B.2}$

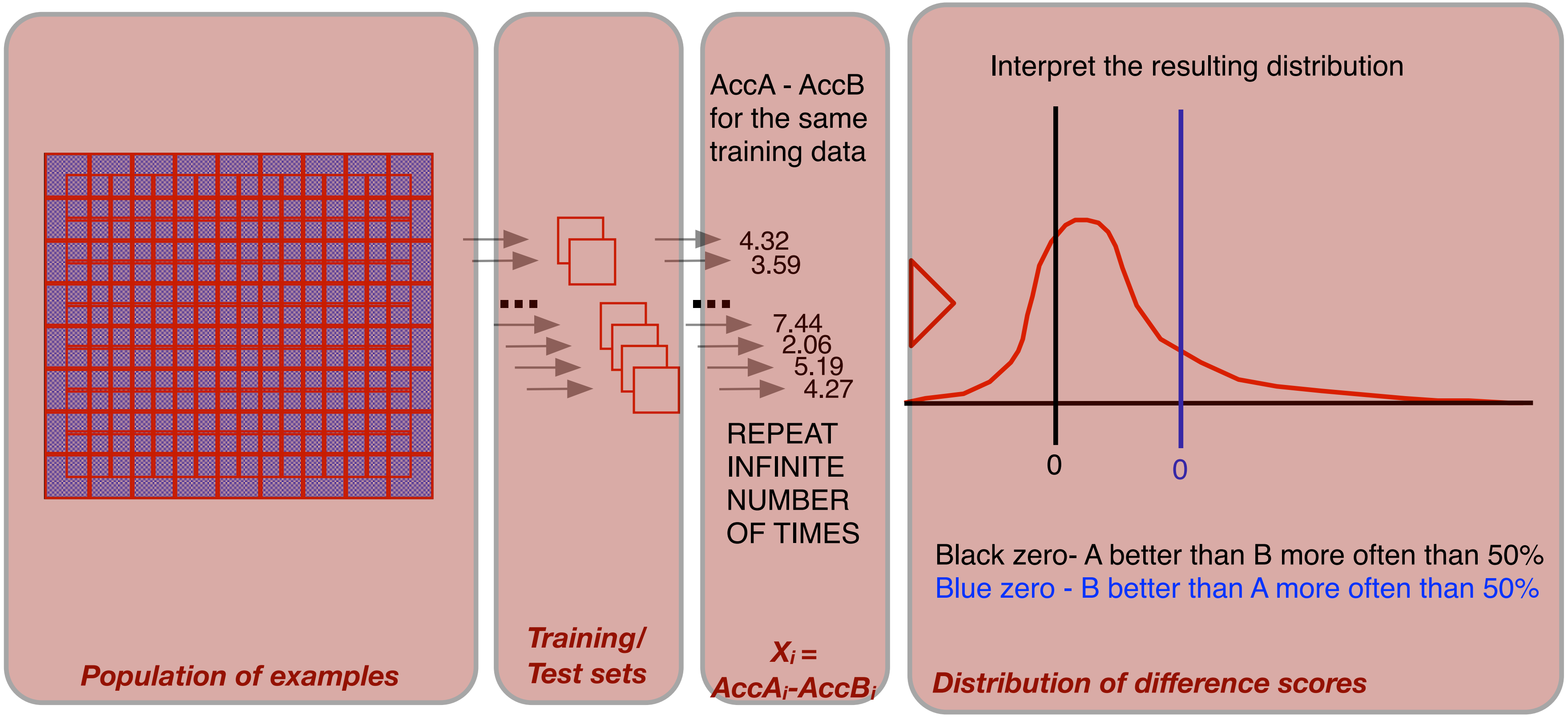
$X_3 = Acc_{A.3} - Acc_{B.3}$

$X_4 = Acc_{A.4} - Acc_{B.4}$

$X_5 = Acc_{A.5} - Acc_{B.5}$

$X_6 = Acc_{A.6} - Acc_{B.6}$

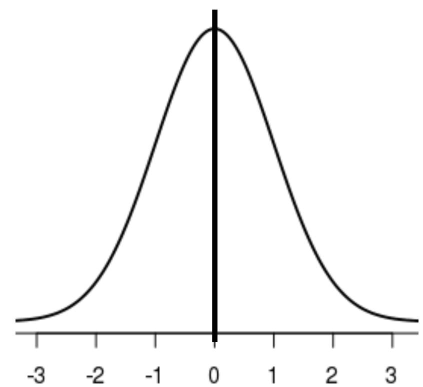
Sampling distributions



Q: How can we do something similar with finite data?

A: Simulations

How can we perform these tests?



- Assume X_i ~ Normal(μ,1)
  - Assume μ ~ Normal(0,1)
- We need to evaluate $P[H_0] := P[\mu > 0 \mid X_1, \ldots, X_6]$
- We will evaluate $P[\mu > 0 \mid X_1, \ldots, X_6]$ by simulating from $P[\mu \mid X_1, \ldots, X_6]$
  - How to get $P[\mu \mid X_1, \ldots, X_6]$?
  - By Bayes rule $P[\mu|X_1, \ldots, X_6] = \dfrac{P[X_1, \ldots, X_6|\mu]P[\mu]}{P[X_1, \ldots, X_6]}$
  - Notes conjugate.pdf show $P[\mu \mid X_1, X_2, \ldots, X_6]$ has a closed-form equation:

$$P[\mu|X_1, \ldots, X_n] = \text{Normal}\left(\underbrace{\frac{\sum_{i=1}^n X_i}{n+1}}_{\text{Average}}, \underbrace{\frac{1}{n+1}}_{\text{Variance}}\right)$$

Estimating P[μ > 0 | X_1, X_2,…, X_6] via simulation

\# K is the number of rounds in the simulation

$$\text{count}_{\hat{\mu}_r > 0} \leftarrow 0$$

- For r in 1 to K:
  - $\hat{\mu}_r = \text{Sample\_Normal}\left(\frac{\sum_{i=1}^{n} X_i}{n+1}, \frac{1}{n+1}\right)$

    Average Variance (careful, in python this is the standard dev.)

  - $\text{count}_{\hat{\mu}_r > 0} \leftarrow \text{count}_{\hat{\mu}_r > 0} + \mathbf{1}(\hat{\mu}_r > 0)$

\#Estimate of P[μ > 0 | X_1, …, X_n]

$$P[\mu > 0 | X_1, \ldots, X_n] \approx \frac{\text{count}_{\hat{\mu}_r > 0}}{K}$$