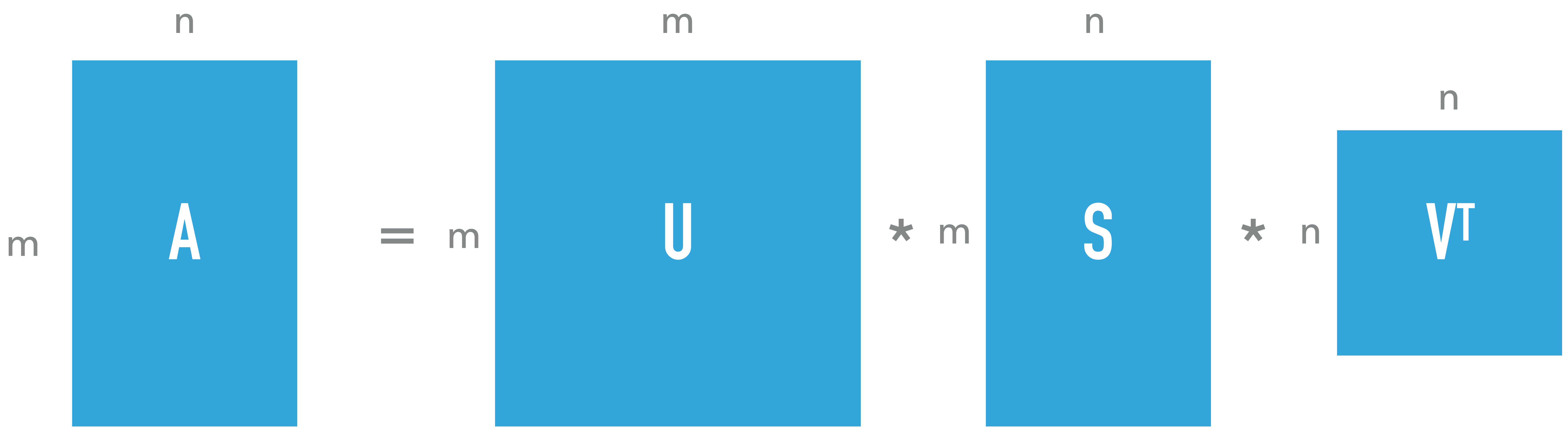
CS57300
PURDUE UNIVERSITY
SEPTEMBER 1, 2021

DATA MINING

SIGULAR VALUE DECOMPOSITION (SVD)

- A rectangular matrix *A* can be broken down into the product of three matrices: an orthogonal matrix *U*, a diagonal matrix *S*, and the transpose of an orthogonal matrix *V*.



SINGULAR VALUE DECOMPOSITION (SVD)

- Columns of U are eigenvectors of AA^T
- Columns of V are eigenvectors of A^TA
- Diagonal entries of S are the square roots of the non-zero eigenvalues of AA^T (as well as A^TA)

SINGULAR VALUE DECOMPOSITION (SVD)

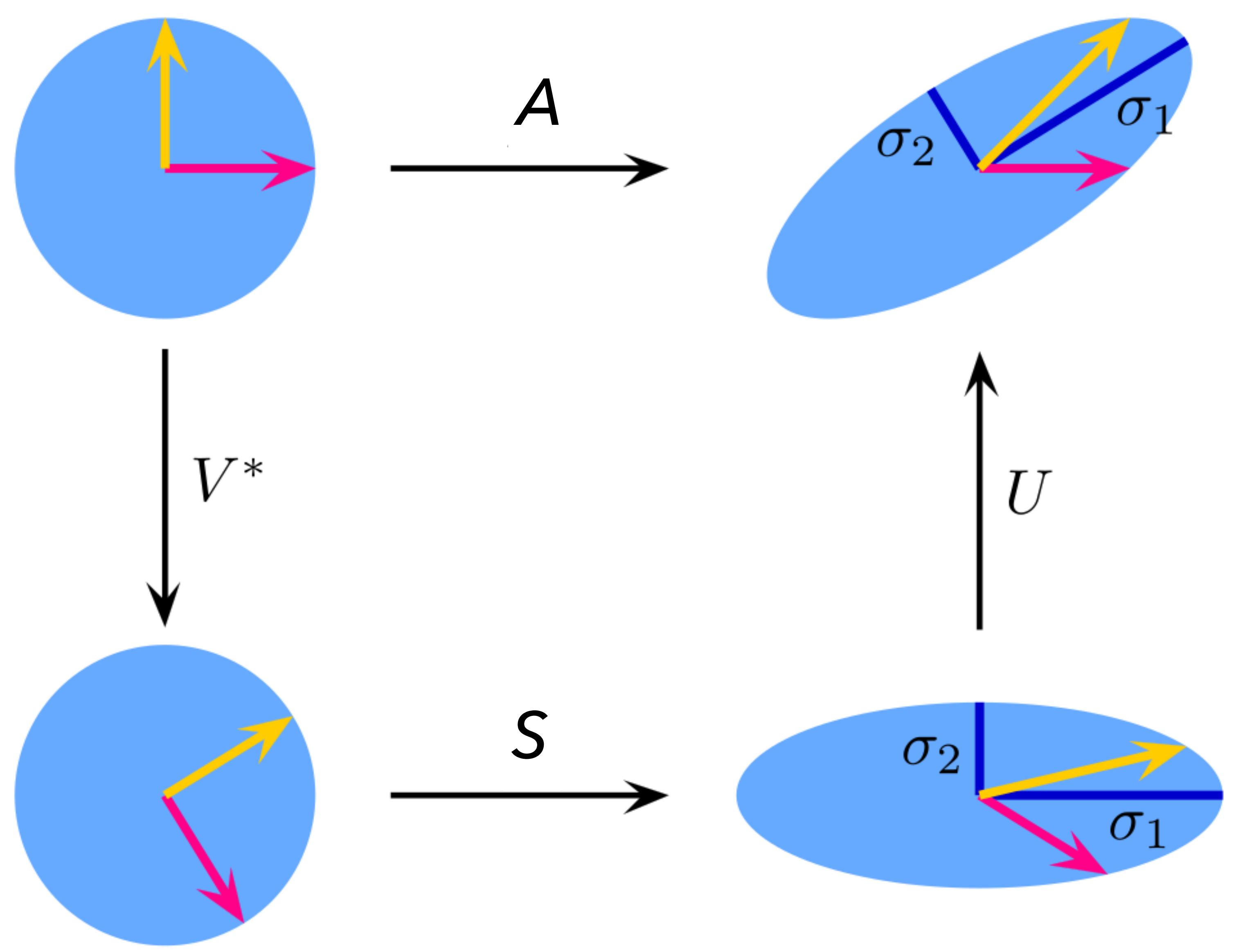
- $A = USV^T$; columns of U are eigenvectors of AA^T, columns of V are eigenvectors of A^TA, diagonal entries of S are the square roots of the non-zero eigenvalues of AA^T (as well as A^TA)

SINGULAR VALUE DECOMPOSITION (SVD)

- $A = USV^T$; columns of U are eigenvectors of AA^T, columns of V are eigenvectors of A^TA, diagonal entries of S are the square roots of the non-zero eigenvalues of AA^T (as well as A^TA)
  - $AA^T = USV^TVS^TU^T$
  - A^TA is a symmetric matrix, so the matrix composed of eigenvectors of A^TA (i.e., V) is orthogonal, thus $V^TV = I$
  - $AA^T = USS^TU^T$ (this is eigendecomposition of matrix AA^T).

SINGULAR VALUE DECOMPOSITION (SVD)

- Columns of U are eigenvectors of AA^T
- Columns of V are eigenvectors of A^TA
- Diagonal entries of S are the square roots of the non-zero eigenvalues of AA^T (as well as A^TA)
- Geometric interpretation:



DISTANCE MEASURES

REPRESENTING DATA IN EUCLIDEAN SPACE

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Many data mining techniques then use similarity/dissimilarity measures to characterize relationships between the instances

Height	Weight	Heart Rate	BP (Diastolic)	BP (Systolic)
1.79	80	70	73	112
1.60	51	73	69	105

METRIC PROPERTIES

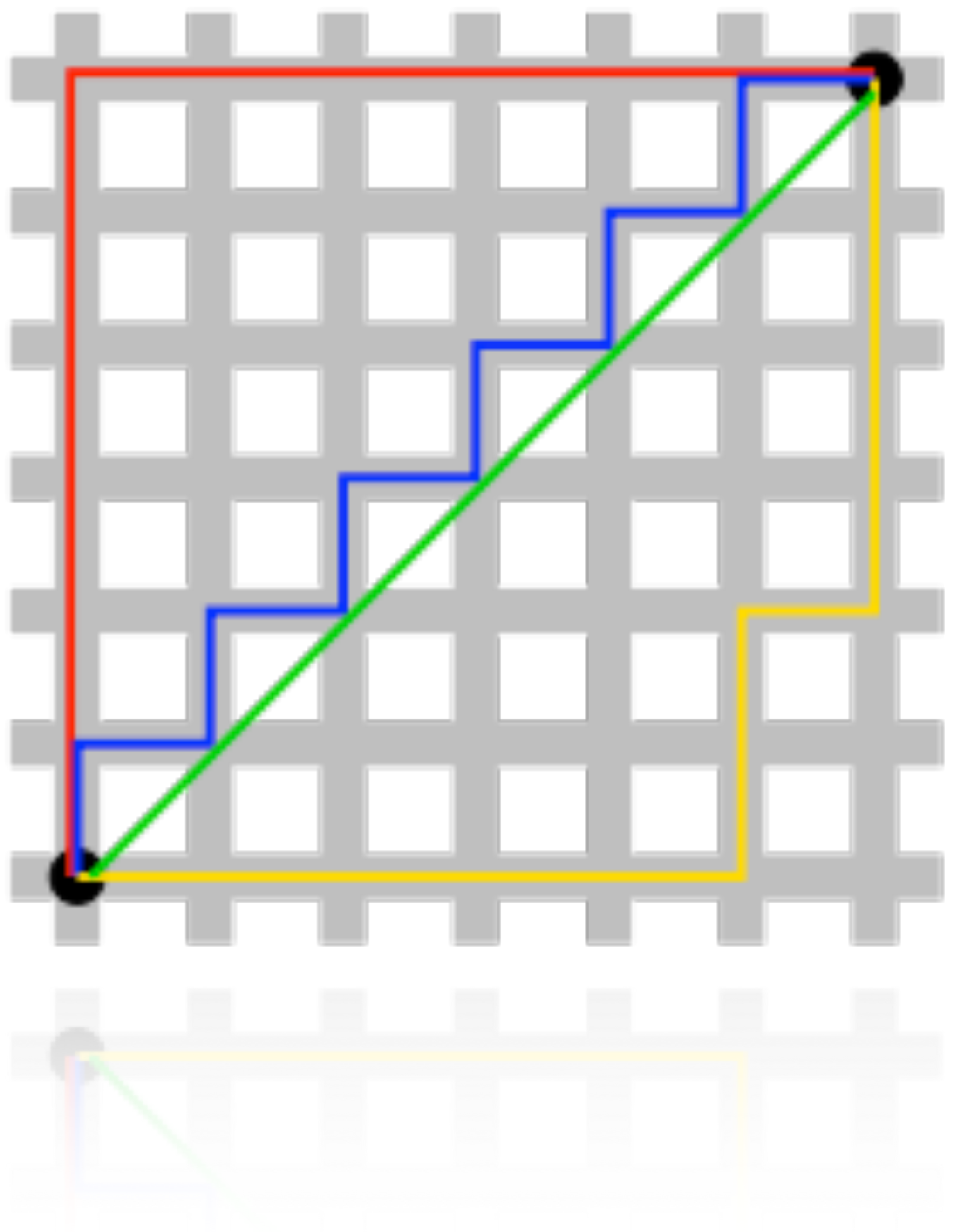
- A **metric** *d(x,y)* (or a distance function) is a function that satisfies the following properties:
  - $d(x,y) \geq 0$ for all x,y and $d(x,y)=0$ iff $x=y$ — **Positivity**
  - $d(x,y) = d(y,x)$ for all x,y — **Symmetry**
  - $d(x,y) \leq d(x,k)+d(k,y)$ for all x,y,k — **Triangle inequality**

DIFFERENT TYPES OF METRICS

- Manhattan distance (L1) $d_M(x,y) = \sum_{i=1}^{p} |x_i - y_i|$
- Euclidean distance (L2) $d_E(x,y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$
  - Most common metric
  - Assumes dimensions are commensurate
- **Weighted** Euclidean distance

$$d_{WE}(x,y) = \sqrt{\sum_{i=1}^{p} w_i (x_i - y_i)^2}$$

  - Can weight variables by relative importance



STANDARDIZATION

- Normalization
  - Removes effect of scale
  - Divide each variable by its standard deviation
  - Weights all variables equally

$$x'_k = \frac{x_k - \bar{x}_k}{\hat{\sigma}_k}$$

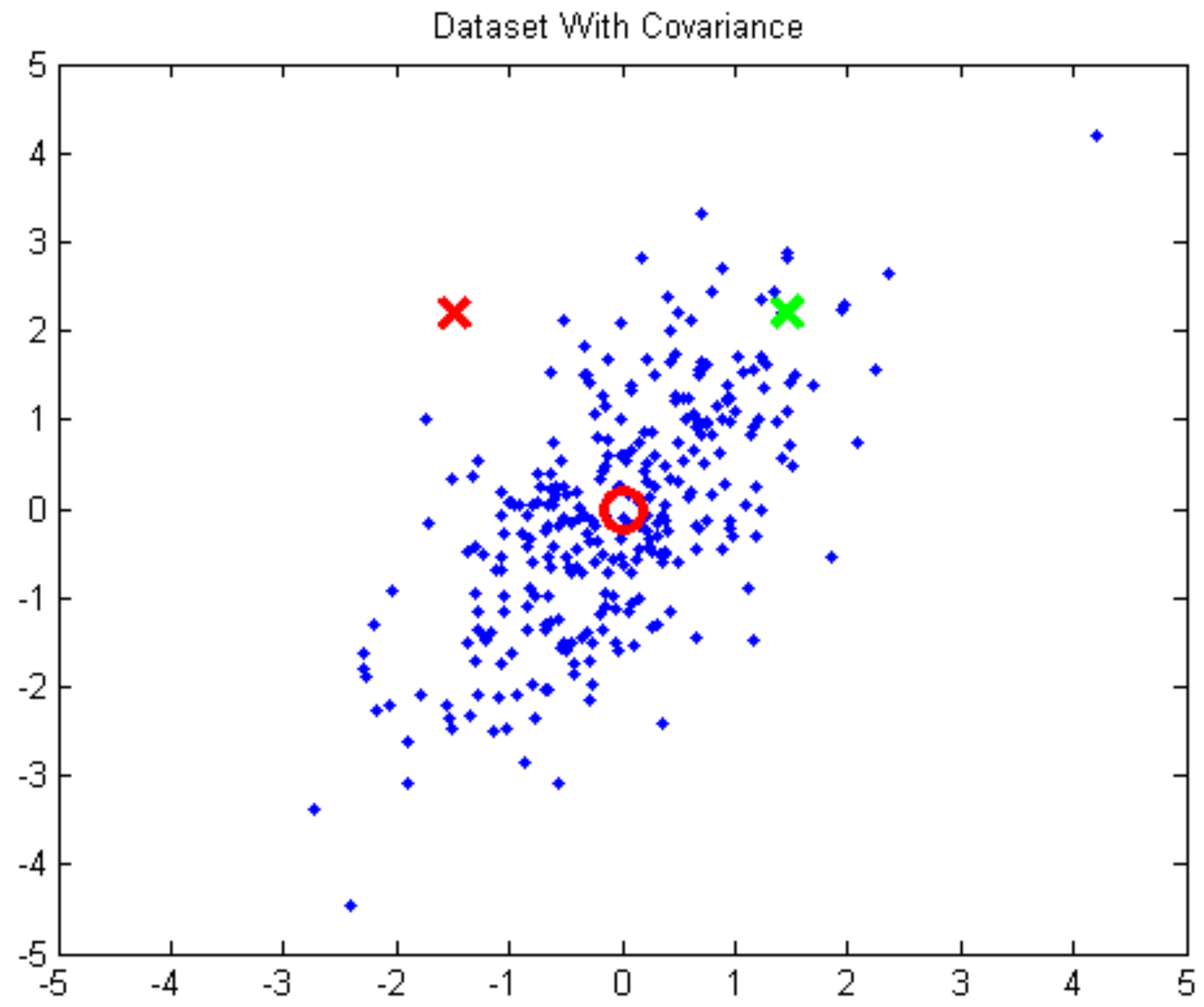
subtract mean

divide by stdev

$$d'_E(x, y) = \sqrt{\sum_{i=1}^{p} (x'_i - y'_i)^2}$$

CORRELATION AMONG VARIABLES

- Variables contribute independently to additive measure of distance
- May not be appropriate
  if variables are highly correlated
- Can standardize variables in a way that accounts for covariance



MAHALANOBIS DISTANCE

$$d_{MH}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y})}$$

pxp covariance matrix

- Automatically accounts for scaling
- Corrects for correlation between attributes
- Tradeoff:
  - Covariance matrix can be hard to estimate accurately
  - Memory and time complexity is quadratic rather than linear

DISTANCE MEASURES FOR BINARY DATA

- d(x,y) when items x and y are *p*-dimensional binary vectors
- Let n_{11} be the number of attributes where both items have value 1, etc.

$$n_{11} = \sum_{i}^{p} \mathbb{I}(x_i + y_i = 2)$$

	y=1	y=0
x=1	n_{11}	n_{10}
x=0	n_{01}	n_{00}

- Matching distance
  - Hamming distance normalized by number of bits

$$d_M(x, y) = 1 - \frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{10} + n_{01}}$$

- Jaccard distance
  - If we don't care about matches on zeros

$$d_M(x, y) = 1 - \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

POPULATIONS AND SAMPLES

ELEMENTARY UNITS, POPULATIONS, AND SAMPLES

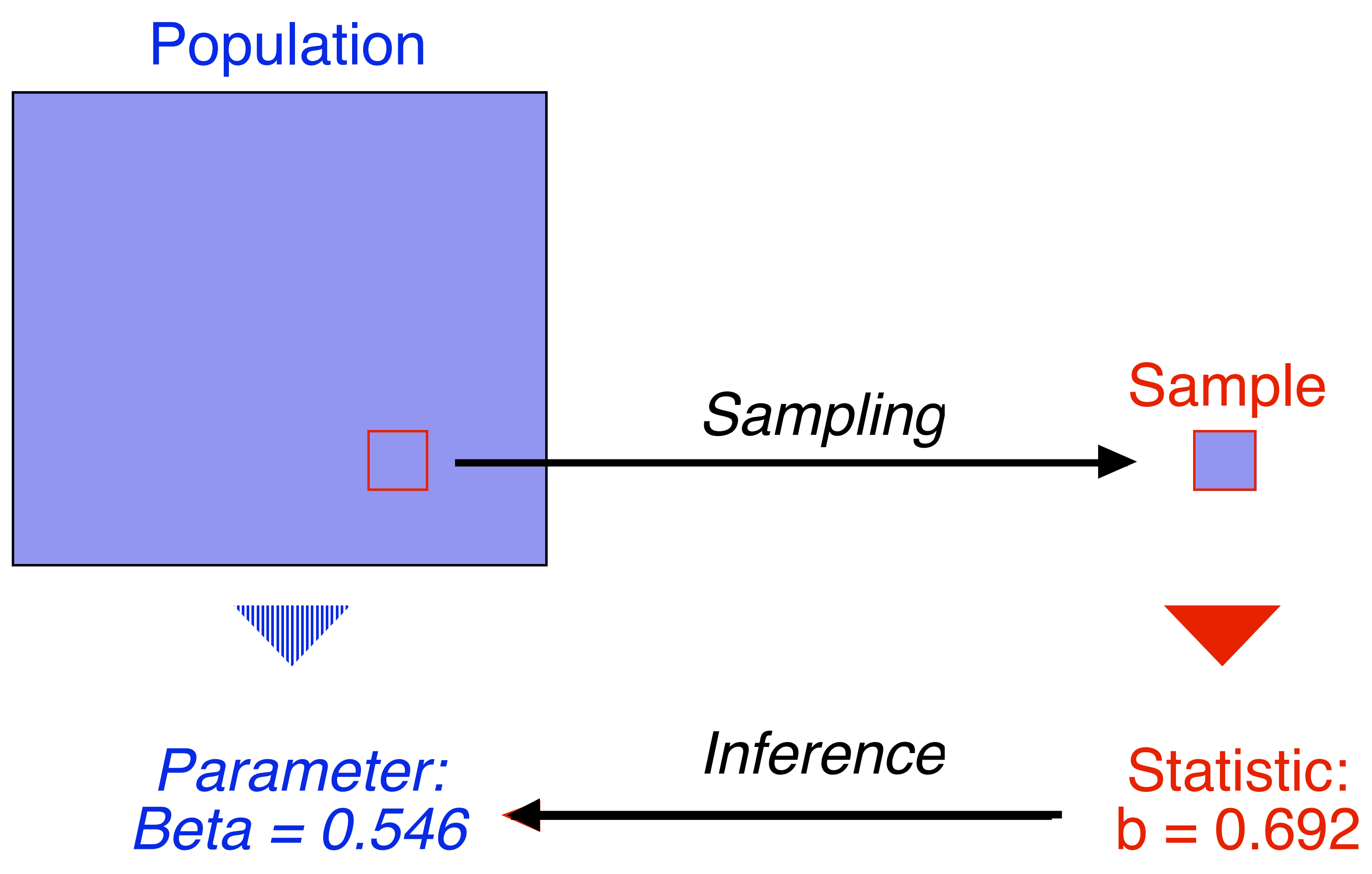
- Elementary units:
    - Entities (e.g., persons, objects, events) that meet a set of specified criteria
    - Example: A person who has purchased something from Walmart in the past month
- Population:
    - Aggregate of elementary units (i.e, all entities of interest)
- Sample:
    - Sub-group of the population

SAMPLING

- 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
- Sampling is the main technique employed for data selection

USE SAMPLES FOR ESTIMATION

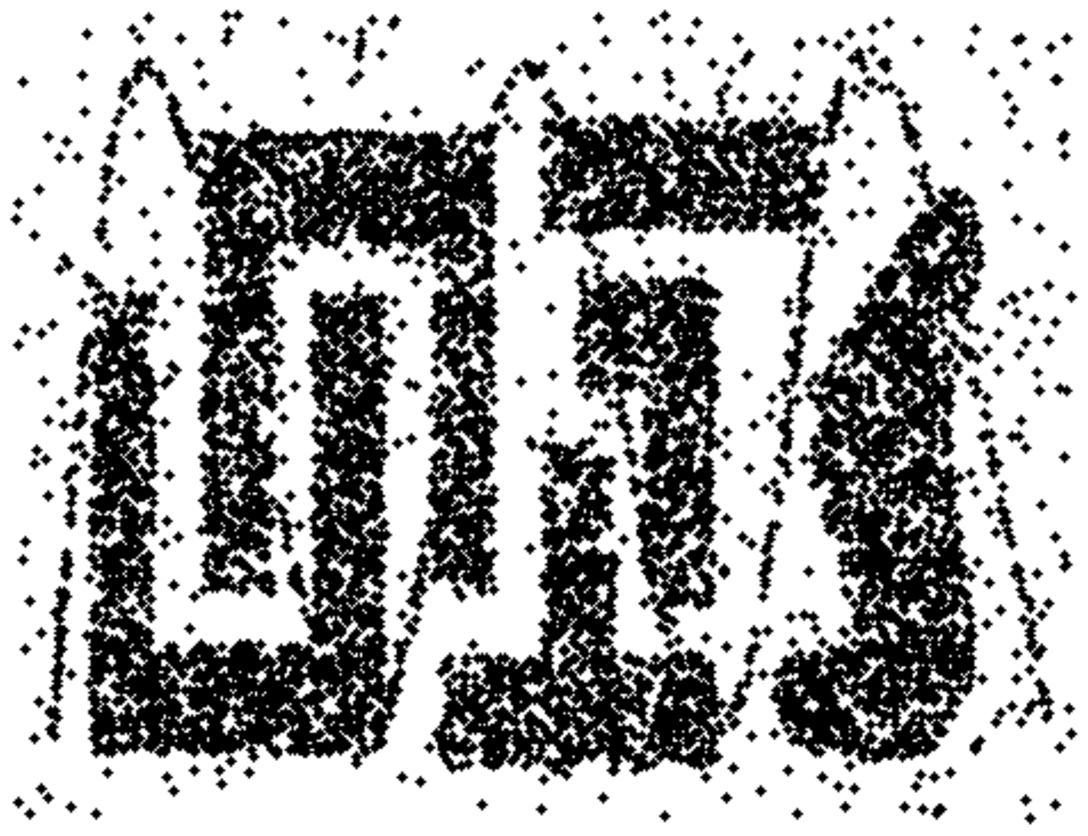
- In data mining we often work with a sample of data from the population of interest
- If we had the population we could calculate the properties of interest
- Sample serves as a reference group for **estimating** characteristics about the population and drawing conclusions



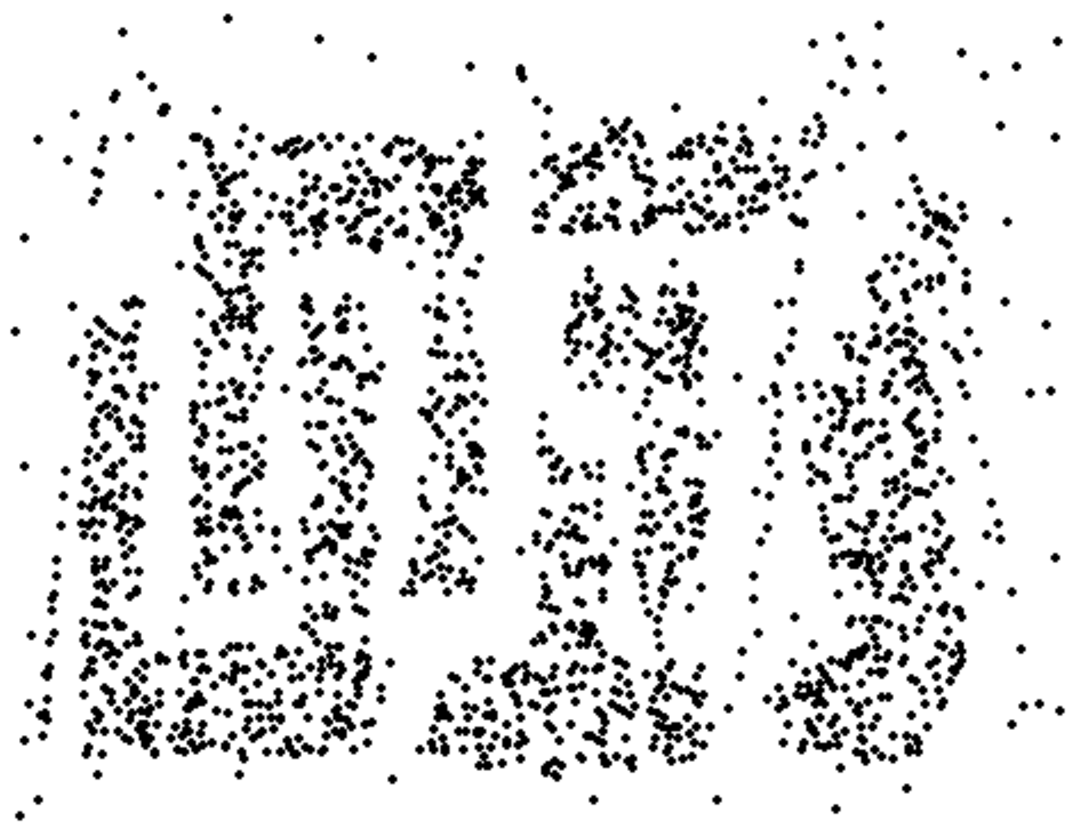
PRINCIPLE FOR EFFECTIVE SAMPLING

- 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

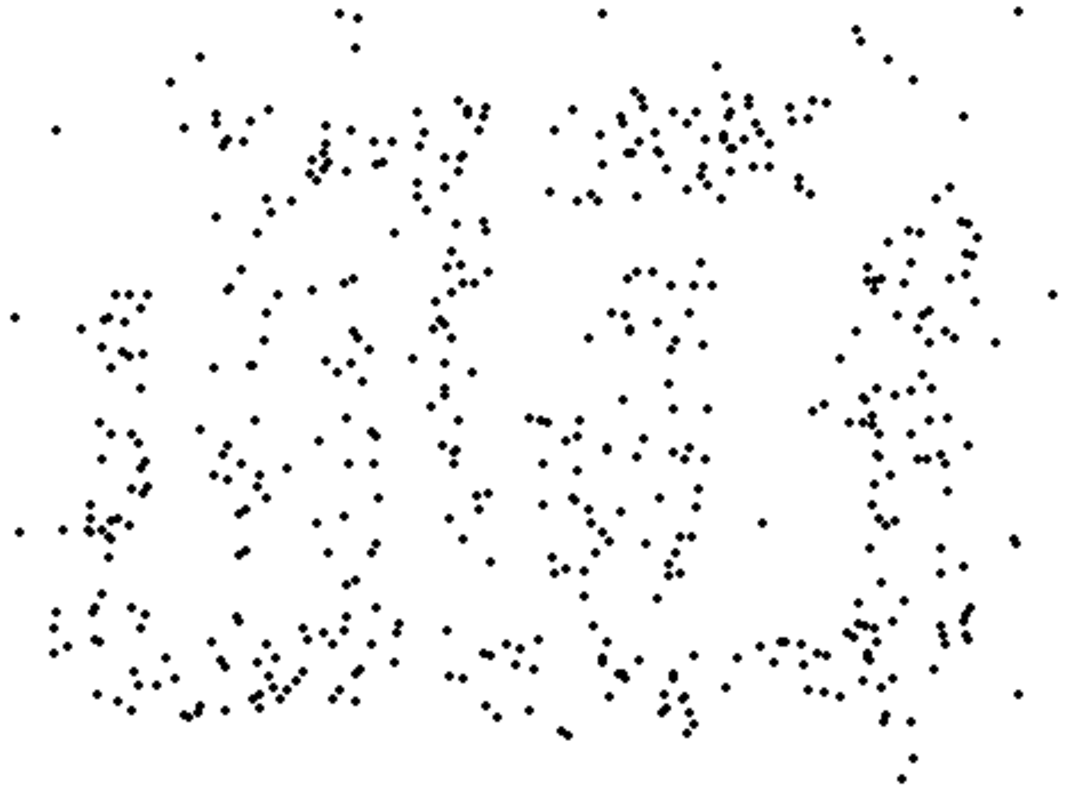
SAMPLE SIZE



8000 Points



2000 Points



500 Points

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

SAMPLING METHODS

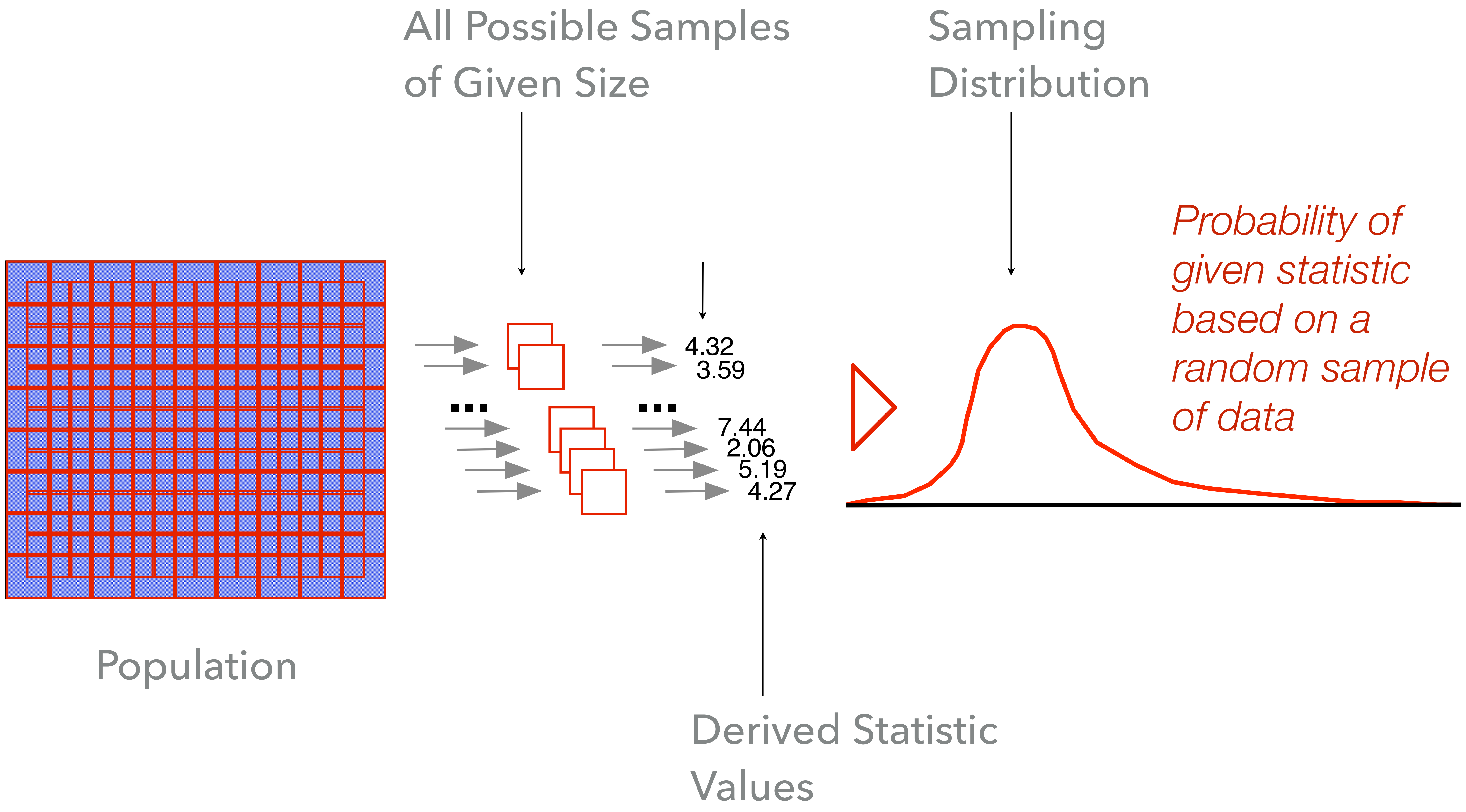
PROBABILITY SAMPLING

1. SIMPLE RANDOM SAMPLING
2. SYSTEMATIC SAMPLING
3. STRATIFIED SAMPLING
4. CLUSTER SAMPLING
5. MULTI-STAGE SAMPLING

NON-PROBABILITY SAMPLING

1. CONVENIENCE SAMPLING
2. QUOTA SAMPLING
3. SNOWBALL SAMPLING
4. JUDGEMENT SAMPLING

SAMPLING DISTRIBUTIONS



STATISTICAL INFERENCE

STATISTICAL INFERENCE

- Infer properties of an unknown distribution with sample data generated from that distribution
- Parameter estimation
  - Infer the value of a population parameter based on a sample statistic (e.g., estimate the mean)
- Hypothesis testing
  - Infer the answer to a question about a population parameter based on a sample statistic (e.g., is the mean non-zero?)

PARAMETER ESTIMATION

- Infer the value of population parameters (θ) from data
- θ can take values in the parameter space Θ
- Frequentist approach
  - Population parameters are fixed but unknown
  - Data is a random sample drawn from population
  - Use maximum likelihood estimation (MLE)
- Bayesian approach
  - Parameters are random variables with a distribution of possible values
  - Data is fixed and known, provides evidence for different parameter values
  - Use maximum aposteriori estimation (MAP)

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

- Suppose we have a set of data $X = \{x_i\}_{i=1}^{N}$ independently drawn from the population
- The maximum likelihood estimation finds the parameter values that maximize the likelihood of observing the data

$$
\begin{aligned}
\theta_{MLE} &= \arg\max_{\theta} P(X|\theta) \\
&= \arg\max_{\theta} \prod_{i} P(x_i|\theta) \\
&= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta) \\
&= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)
\end{aligned}
$$

MAXIMUM A-POSTERIORI ESTIMATION (MAP)

- Suppose we have a set of data $X = \{x_i\}_{i=1}^{N}$ independently drawn from the population, and the prior distribution for the parameter is $P(\theta)$
- The maximum a-posteriori estimation finds the mode of the posterior distribution of the parameters

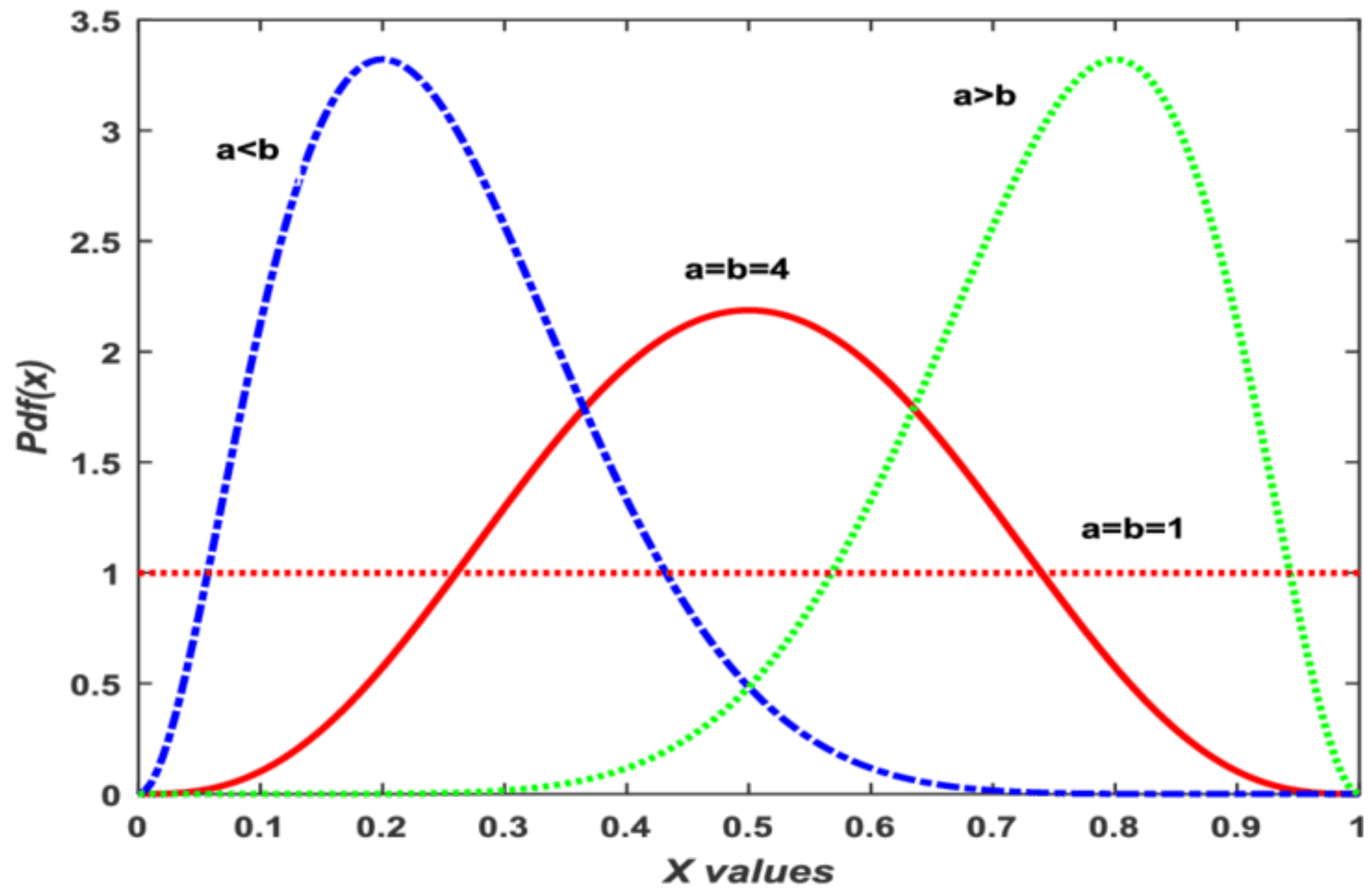
$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} P(X|\theta)P(\theta) \\ &= \arg\max_{\theta} \log P(X|\theta)P(\theta) \\ &= \arg\max_{\theta} \log \prod_{i} P(x_i|\theta)P(\theta) \\ &= \arg\max_{\theta} \sum_{i} \log P(x_i|\theta)P(\theta)\end{aligned}$$

MLE VS. MAP EXAMPLE

- Flip a coin for *N* times and observe *n* heads; what's the probability of seeing the head if tossing the coin once?
- Likelihood of observing the data: $P(D\,|\,\theta) = \binom{N}{n} \theta^n (1-\theta)^{N-n}$
  - The number of heads observed follows a binomial distribution
- Maximum likelihood estimation:

$$\theta_{MLE} = \text{argmax}_\theta P(D\,|\,\theta) = \frac{n}{N}$$

MLE VS. MAP EXAMPLE



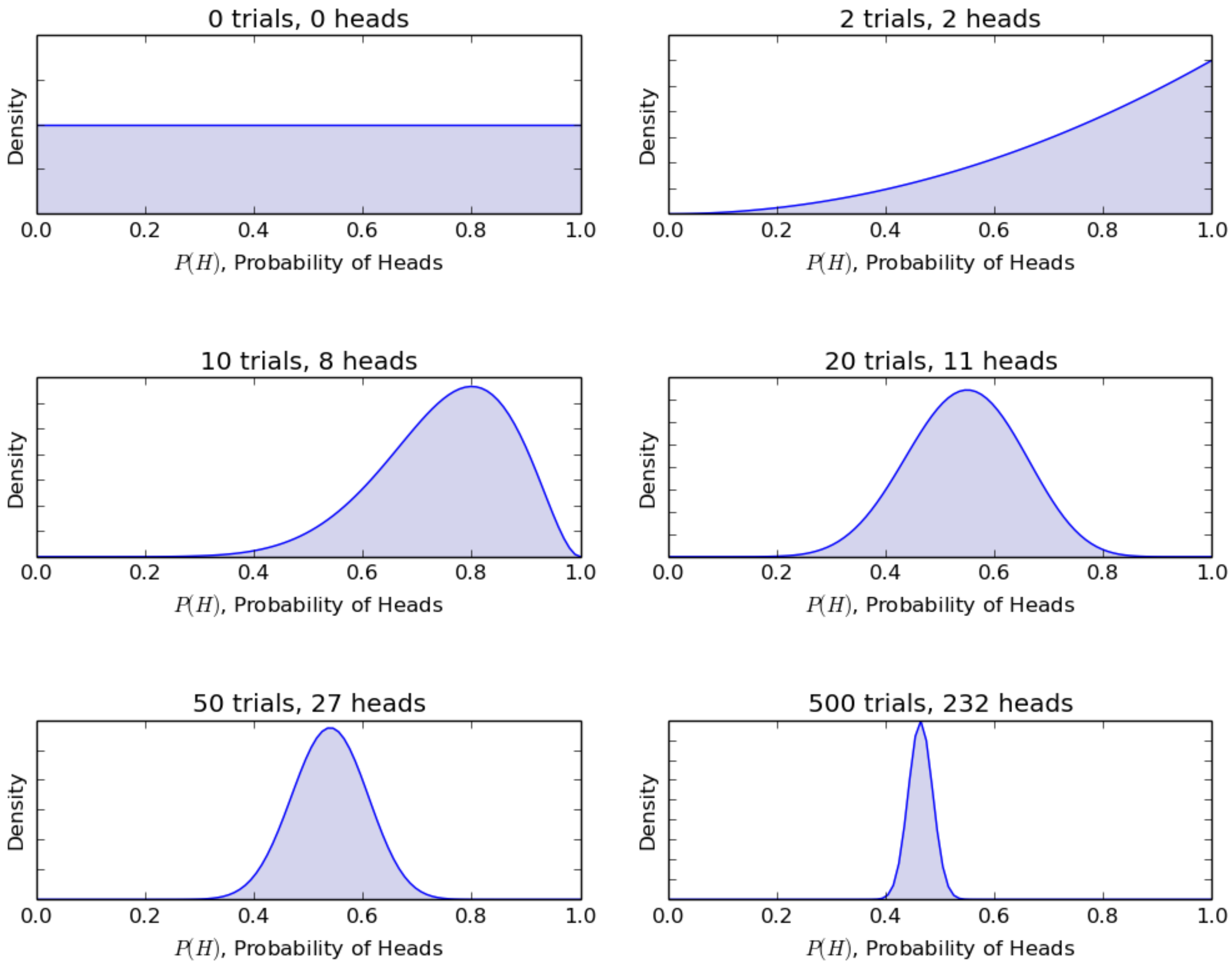
- Maximum a-posteriori estimation:
  - Suppose the prior is a Beta distribution

$$P(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \sim Beta(a,b), \text{where } B(a,b) = \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx$$

  - Then, the posterior is:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{\binom{N}{n}(\theta^{a+n-1}(1-\theta)^{b+N-n-1}/B(a,b))}{\int_0^1 \binom{N}{n}(\theta^{a+n-1}(1-\theta)^{b+N-n-1}/B(a,b))d\theta}$$

$$\sim Beta(a+n, b+N-n)$$


0 trials, 0 heads
2 trials, 2 heads
10 trials, 8 heads
20 trials, 11 heads
50 trials, 27 heads
500 trials, 232 heads
Density
0.0
0.2
0.4
0.6
0.8
1.0
$P(H)$, Probability of Heads

MLE VS. MAP EXAMPLE

- Maximum a-posteriori estimation:

$$\begin{aligned}\theta_{MAP} &= \text{argmax}_\theta P(\theta \mid D) \\ &= \text{argmax}_\theta Beta(a+n, b+N-n) \\ &= \frac{a+n-1}{a+b+N-2}\end{aligned}$$

- Notice that in this example, the posterior distribution is in the same probability distribution family as the prior distribution.
  - The prior and posterior are called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood function.
  - The beta distribution is a conjugate prior to the binomial likelihood.