CS57300
PURDUE UNIVERSITY

SEPTEMBER 8, 2021

DATA MINING

PROJECT GUIDELINE IS OUT!

- Teamwork: 2-5 people
- Open topic
- Timeline:
  - September 26, 2021: project proposal due
  - October 31, 2021: project midterm report due
  - December 1, 6 & 8, 2021: project presentation
  - December 12, 2021: project report due
- Check the project guideline to see what needs to be included in each document/presentation you submit!

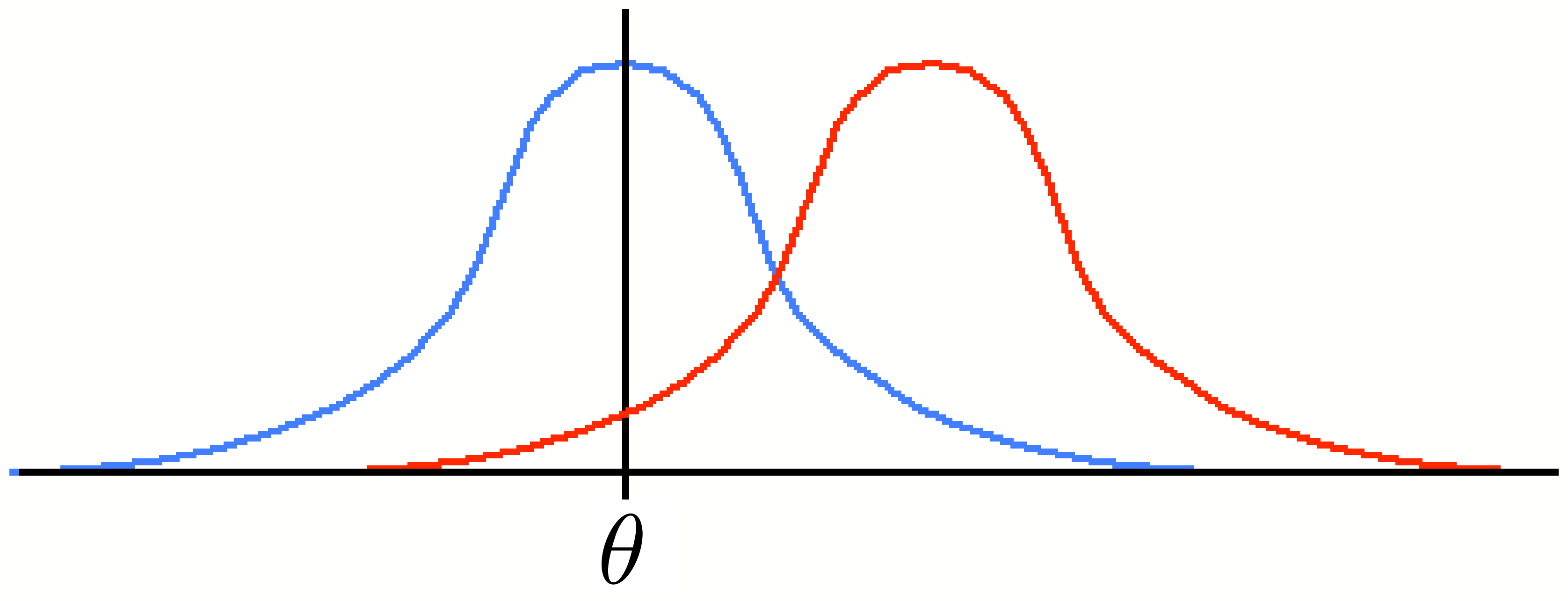
PROPERTIES OF ESTIMATORS

PROPERTIES OF ESTIMATORS

- Let $\hat{\theta}$ be an estimate for a population parameter θ
- Using different samples *D* will result in different estimates $\hat{\theta}_D$
- Thus $\hat{\theta}$ is a random variable with a distribution, mean, and variance
  - We can evaluate the quality of an estimator for θ based on the properties of the sampling distribution of $\hat{\theta}$

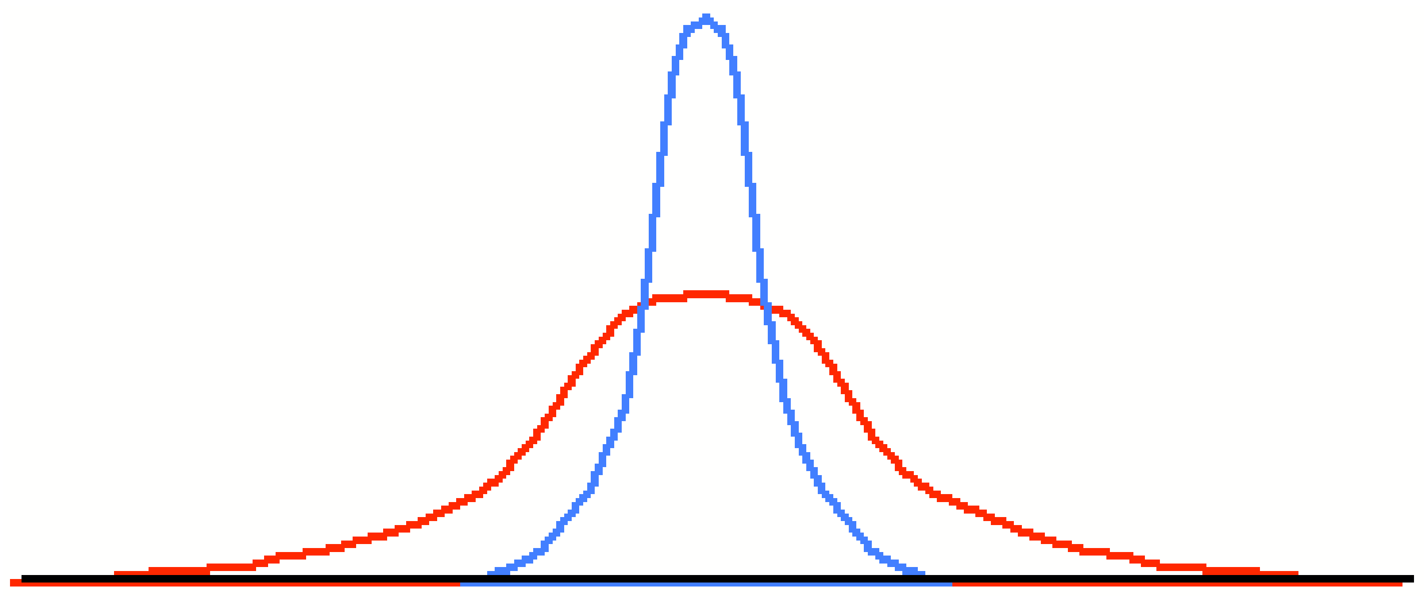
BIAS

- The best estimators produce values that center around the population parameter
- The **bias** of an estimator is defined as: $Bias(\hat{\theta}) = \underbrace{E[\hat{\theta}]}_{\text{Average estimated parameter}} - \underbrace{\theta}_{\text{True parameter in popul.}}$
- An estimator is unbiased if: $E[\hat{\theta}] - \theta = 0$



VARIANCE

- The best estimators produce values that differ only slightly from the population parameter
- The **variance** of an estimator is defined as: $Var(\hat{\theta}) = E[(\underbrace{\hat{\theta}}_{\text{Single parameter estimate}} - \underbrace{E[\hat{\theta}]}_{\text{Average estimated parameter}})^2]$
- Measures how sensitive the estimator is to different datasets
- Unbiased estimators with minimum variance are called *best unbiased estimators*



EXAMPLE

- Ignore data and declare that: $\hat{\theta} = 1.0$
- Estimate will not depend on data, thus: $Var(\hat{\theta}) = 0$
- However, in most cases this estimator will have a large bias (non-zero)

BIAS-VARIANCE TRADEOFF

- The mean-squared error (MSE) of $\hat{\theta}$ is:

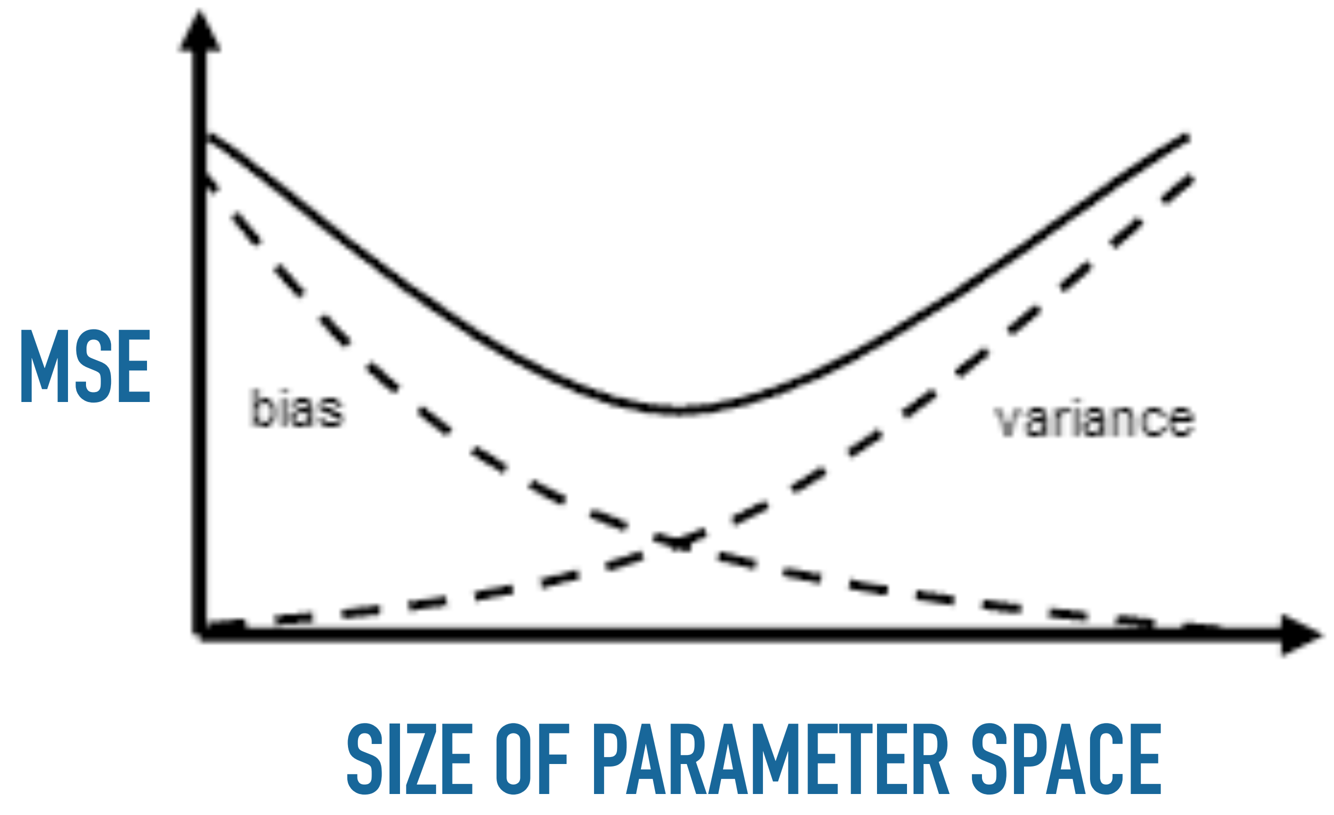
$$E[(\hat{\theta} - \theta)^2]$$

BIAS-VARIANCE TRADEOFF

- The mean-squared error (MSE) of $\hat{\theta}$ is:

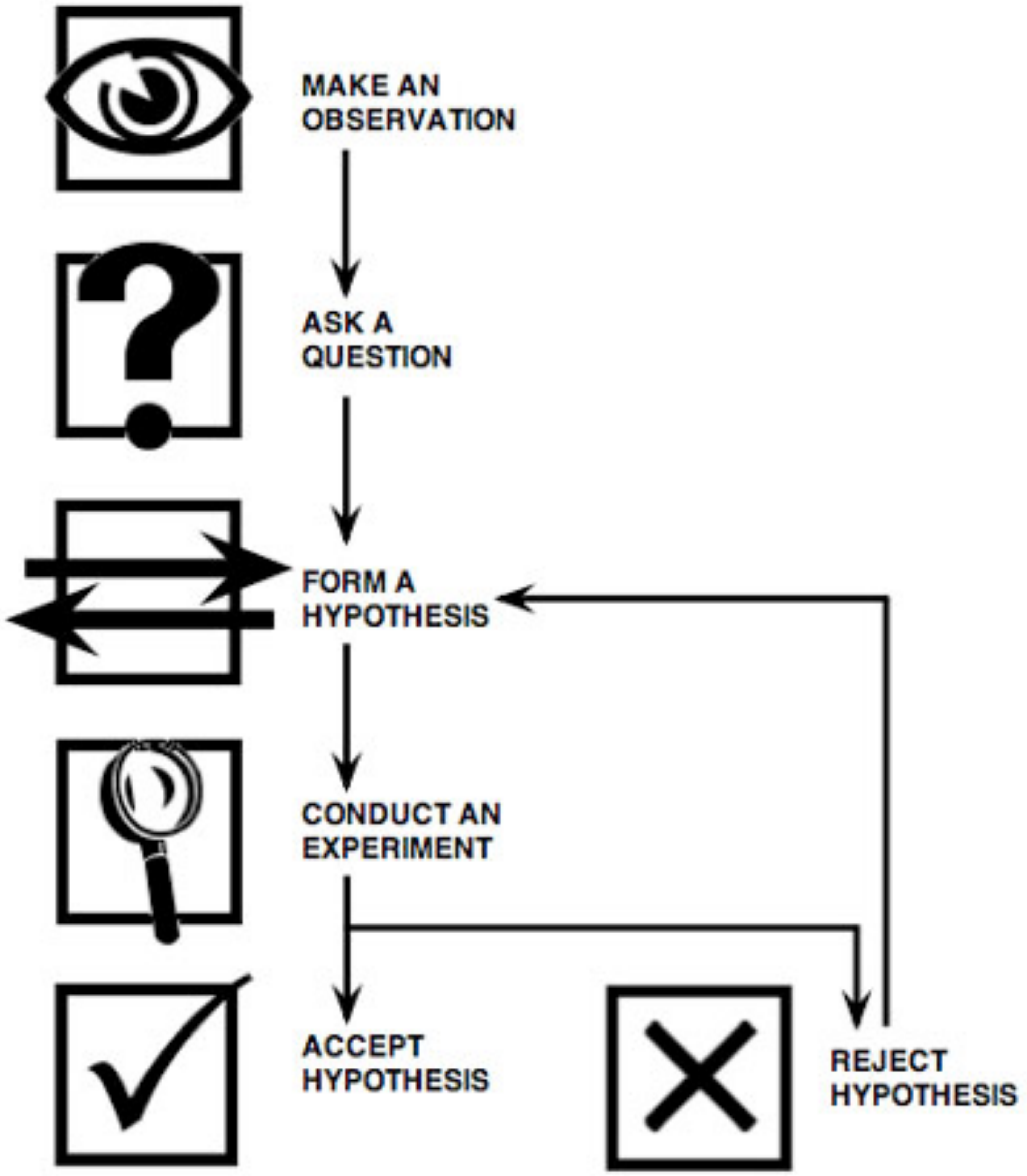
$$
\begin{aligned}
E[(\hat{\theta}-\theta)^2] &= E[(\hat{\theta}-E[\hat{\theta}]+E[\hat{\theta}]-\theta)^2] \\
&= \underbrace{(E[\hat{\theta}]-\theta)^2}_{\textbf{bias}} + \underbrace{E[(\hat{\theta}-E[\hat{\theta}])^2]}_{\textbf{variance}}
\end{aligned}
$$

- MSE measures systematic bias and random variance between estimate and population value
- Tradeoff: reducing bias tends to increase variance and vice versa



HYPOTHESIS TESTING

SCIENTIFIC METHOD



TYPES OF HYPOTHESES

Broad categories

- **Descriptive**: propositions that describe a characteristic of an object
- **Relational**: propositions that describe relationship between 2+ variables
- **Causal**: propositions that describe the effect of one variable on another

Specific characteristics

- **Non-directional**: an differential outcome is anticipated but the specific nature of it is not known (e.g., the tuning parameter will affect algorithm performance)
- **Directional**: a specific outcome is anticipated (e.g., the use of pruning will increase accuracy of models compared to no pruning)

Descriptive Hypothesis

Non-Directional Relational Hypothesis

Directional Relational Hypothesis

Directional Causal Hypothesis

Stronger

HYPOTHESES EXAMPLE

- The query response time is measured for a few different search engines
- Different hypotheses
  - **Descriptive:** The query response time for Google follows a normal distribution
  - **Non-directional relational:** The average response time for a new search engine, QuickSearch, is different from Google's average response time
  - **Directional relational:** The average response time of QuickSearch is shorter than that of Google's
  - **Directional causal:** The response time of QuickSearch is shorter than Google's because they cache results of more queries

HYPOTHESIS TESTING

- Statistical hypothesis test is a method used in statistics that tells you the likelihood of a specific result would happen by chance
- **Null hypothesis** (H_0):
  - Presumed true until statistical inference indicates otherwise; set up to be refuted by alternative
- **Alternative hypothesis** (H_1):
  - Rival hypothesis; that we conjecture is true
- Assuming the null hypothesis is true, what's the probability of getting a statistic that is at least as extreme as the statistic that was actually obtained through the data?

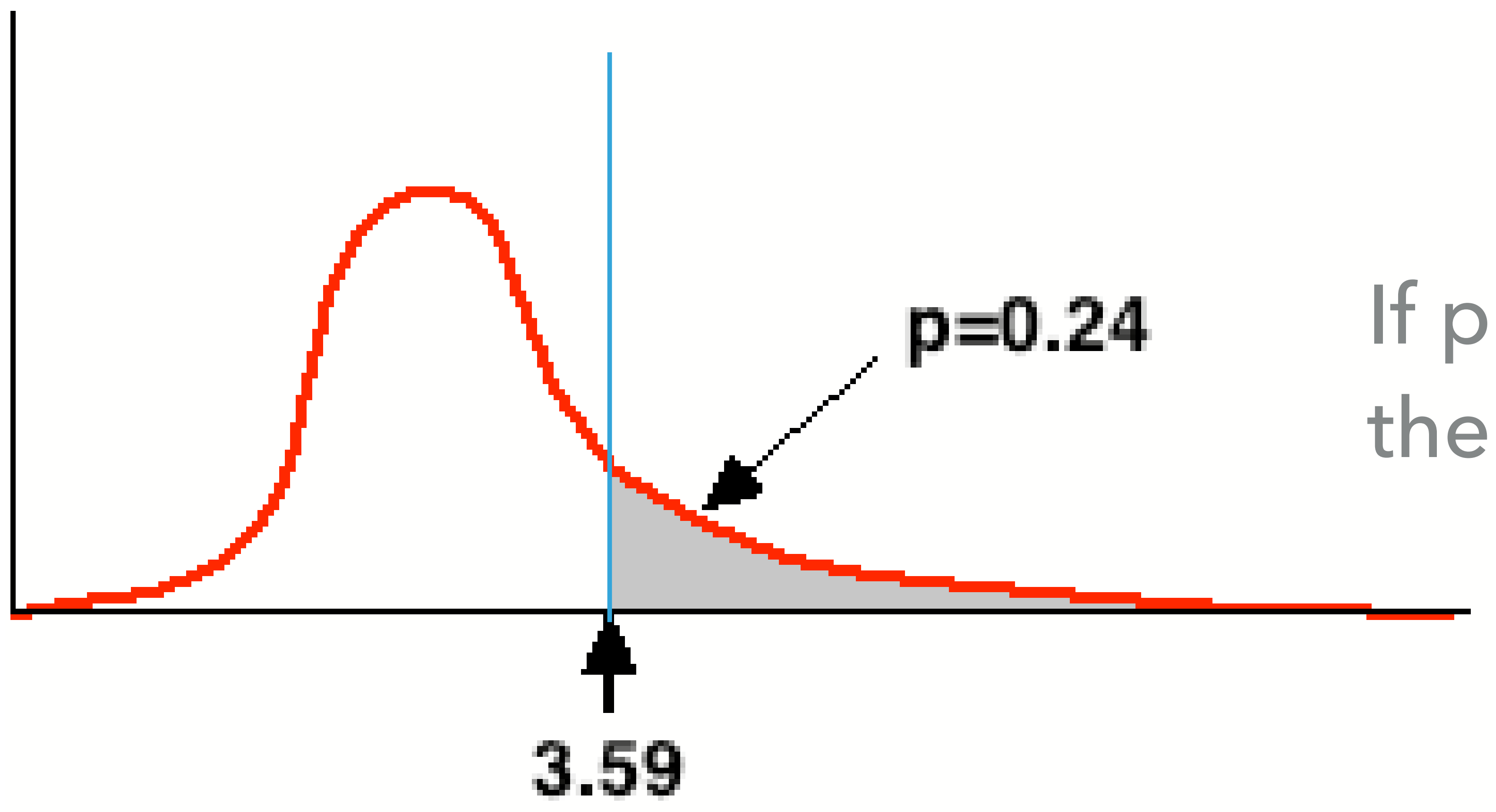
HYPOTHESIS TESTING STRATEGY

- Formulate null and alternative hypothesis
  - H_0: QuickSearch' mean response time = Google's mean response time
  - H_1: QuickSearch' mean response time ≠ Google's mean response time
- Gather a sample statistic (e.g., δ=difference of QuickSearch's and Google's mean response time)
- Determine the sampling distribution for the statistic under the null hypothesis
- Use the sampling distribution to calculate the probability of obtaining the observed value of δ, given H_0
  - If the probability is low, reject H_0 in favor of H_1

REJECTING THE NULL HYPOTHESIS

α = selected significance level

Sampling
Distribution
Under H_0

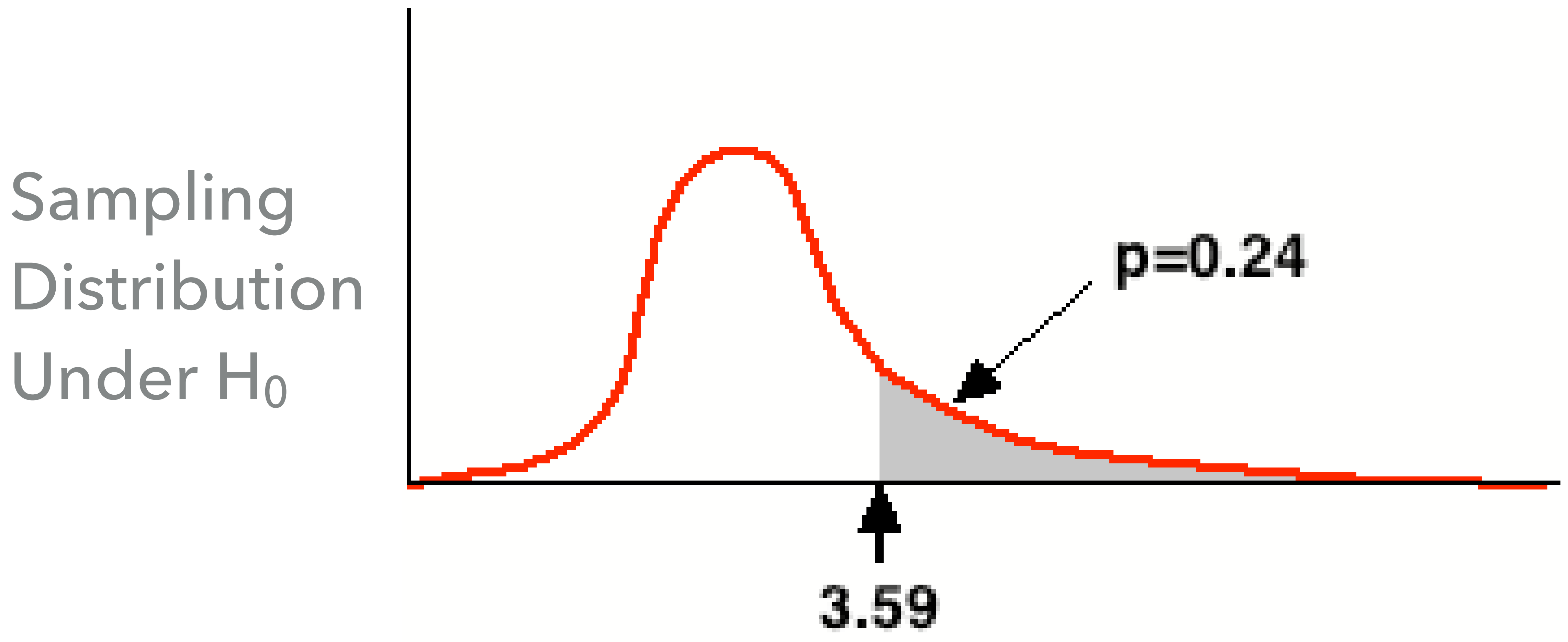


Observed value

If $p < 0.05$
then reject H_0

STATISTICAL SIGNIFICANCE

- A value of a statistic is **statistically significant** if it is unlikely to occur under the null hypothesis



significance level

$$\alpha = p(reject\ H_0 | H_0\ true) = p(type\ 1\ error)$$

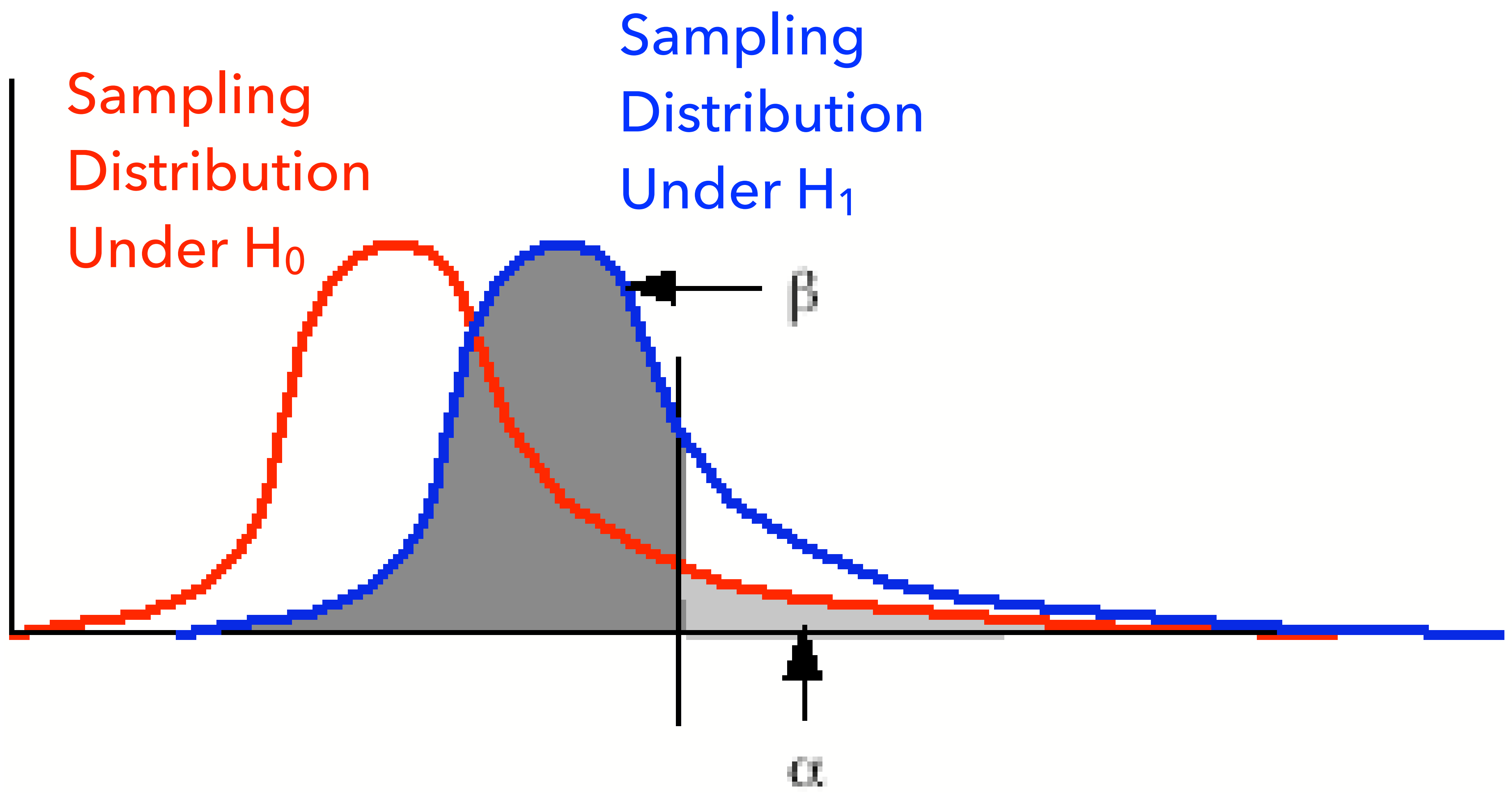
ERRORS

		Decision	
		Reject H_0	Don't reject H_0
Truth	H_0	*Type 1 error*	
	H_1		*Type 2 error*

- Type 1: null is rejected when it is true
    - E.g., conclude cancer drug increases life expectancy when in fact it doesn't
    - Generally considered to be most serious error
- Type 2: null is accepted when it is false
    - E.g., conclude that cancer drug does not increase life expectancy when in fact it does

STATISTICAL POWER

- Lack of statistical significance does not necessarily imply that H_0 is true
- Test could have low statistical power: $(1 - \beta)$ **portion of sampling distribution for alternative that is above threshold**



$$\beta = p(accept\ H_0 | H_0\ false) = p(type\ 2\ error)$$

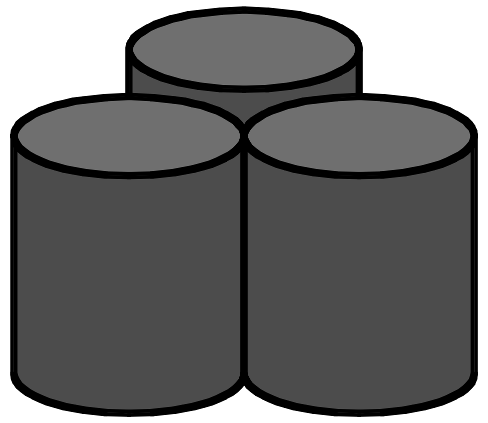
HOW TO INCREASE POWER

- Increase sample size
- Decrease sample variability
  - Matching, sample selection, control for confounding variables, increase precision of measurements
- Increase effect size
  - More extreme experimental conditions, avoid ceiling/floor effects
- Increase alpha (e.g., from 0.05 to 0.10, but this increases type 1 errors)

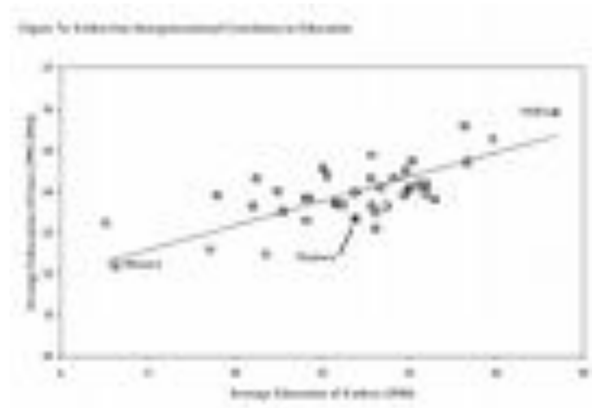
DATA AND MEASUREMENT

REFLECTING REAL WORLD THROUGH DATA

Real world



Data



Relationship in data

Relationship in real world

Goal: map domain entities to symbolic representations

WHAT IS DATA?

- Collection of entities and their attributes
- **Attribute**: property or characteristic of an entity (e.g., eye color, temperature)
- **Entity**: collection of attributes
  Aka: record, point, case, sample, object, or instance

Attributes

Entities

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

DISCRETE AND CONTINUOUS ATTRIBUTES

- Discrete
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, set of words in a collection of documents
  - Often represented as integer variables
- Continuous
  - Has real numbers as attribute values
  - Examples: temperature, height
  - Continuous attributes are typically represented as floating-point variables

Tan, Steinbach, Kumar. Introduction to Data Mining,

TABULAR DATA

- Collection of records, each of which consists of a fixed set of attributes

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

DOCUMENT DATA

- Each document is represented as a **term** vector, where each attribute records the number of times the term occurs in the document

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

TRANSACTION DATA

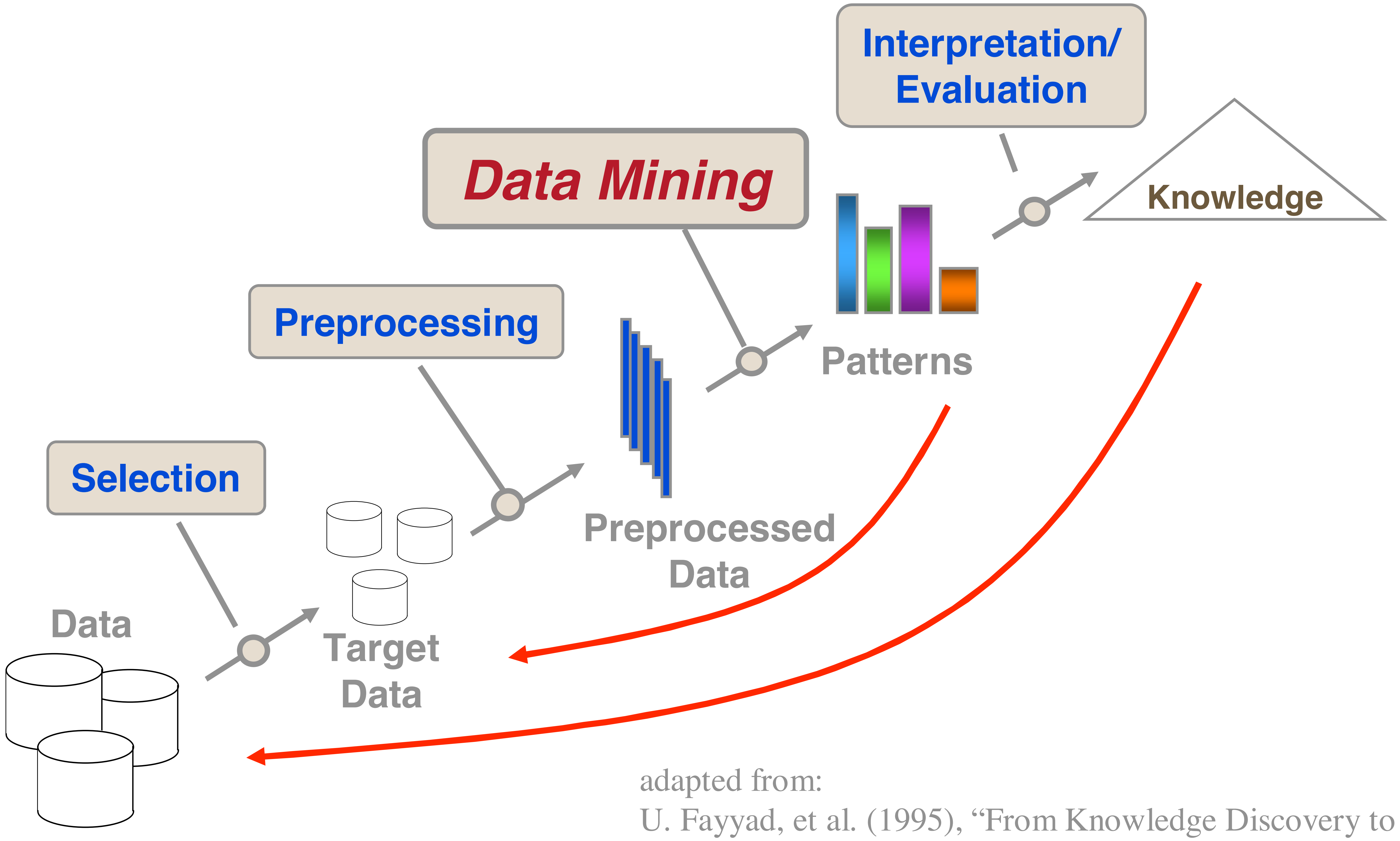
- Each record corresponds to a transaction that involves a set of items
- E.g., in a grocery store purchase, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items

Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b,d,e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

ELEMENTS OF DATA MINING ALGORITHMS

DATA MINING PROCESS



adapted from:
U. Fayyad, et al. (1995), “From Knowledge Discovery to Data Mining: An Overview,” Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

OVERVIEW

- Task specification
- Knowledge representation
- Learning technique
  - Search + scoring
- Prediction and/or interpretation

OVERVIEW

- **Task specification**
- Knowledge representation
- Learning technique
  - Search + scoring
- Prediction and/or interpretation

TASK SPECIFICATION

- Objective of the person who is analyzing the data
- Description of the characteristics of the analysis and desired result

EXPLORATORY DATA ANALYSIS

- Goal
  - Interact with data without clear objective
  - Summarize the main characteristics of the data
- Techniques
  - Mostly visualization