UNIVERSITY OF TEXAS ARLINGTON
COLLEGE OF SCIENCE

DATA 1301
Introduction to Data Science
Logic and Probability Theory

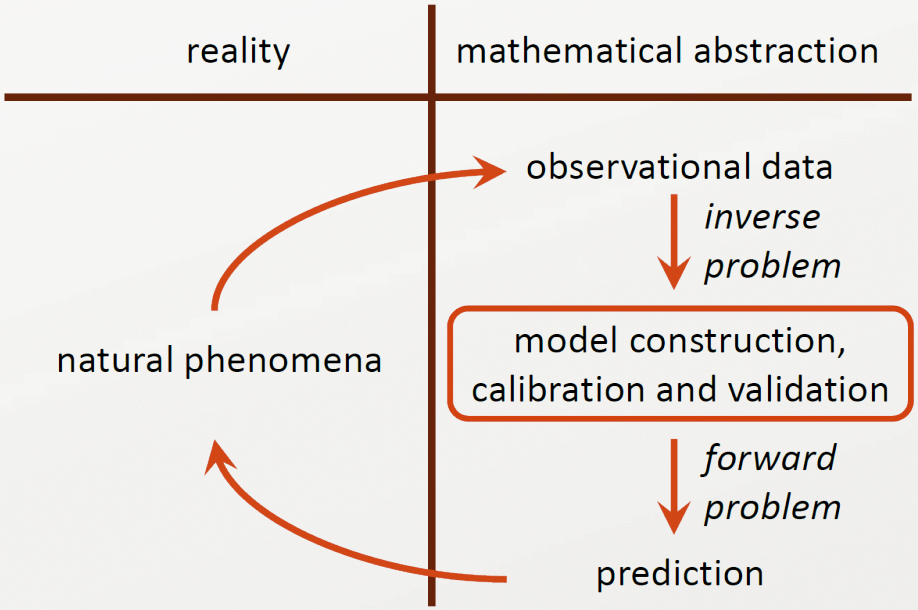
Amir Shahmoradi

Department of Physics / College of Science
Data Science Program / College of Science
The University of Texas
Arlington, Texas

The two classical pillars of science: Experiment and Theory

How do we make a scientific inference?

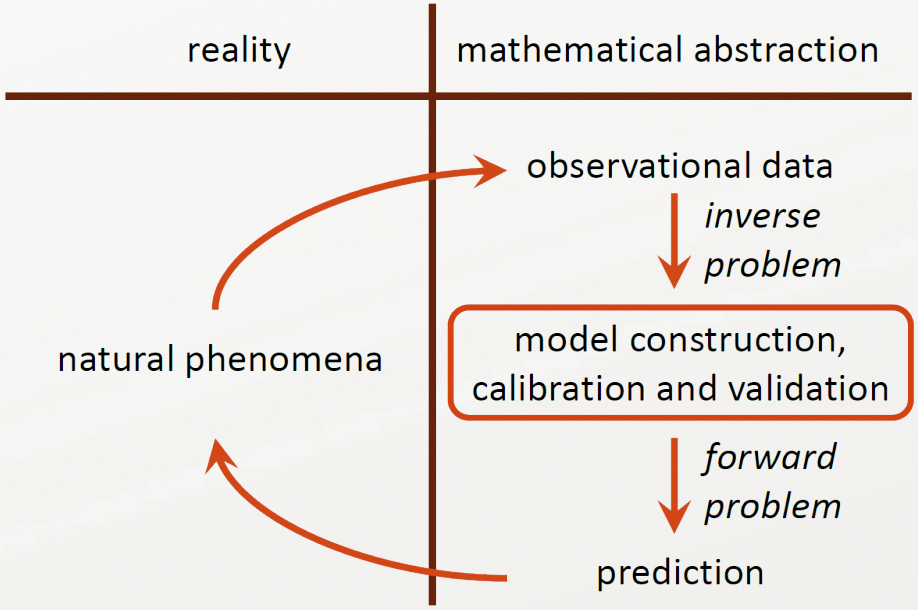
A very elementary depiction of the scientific method



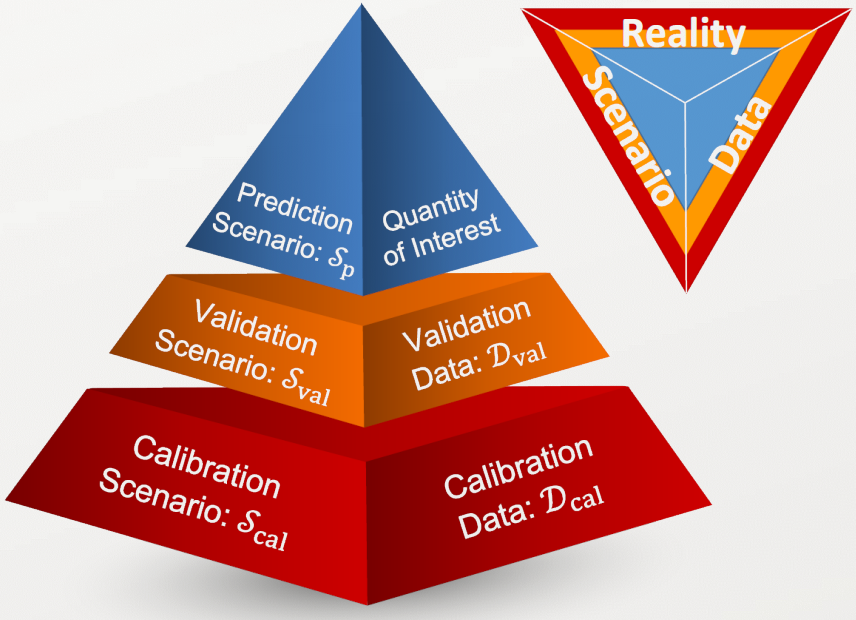
The two classical pillars of science: Experiment and Theory

How do we make a scientific inference?

A very elementary depiction of the scientific method



The prediction pyramid



The desiderata of Probability Theory

There is a set of properties that we **desire** to have in a theory of probability that we wish to construct now.

(I) Degrees of plausibility are represented by real numbers.

(II) Qualitative correspondence with common sense.

(III) Consistency.

- (I) If a conclusion can be reasoned out in more than one way, then every possible way must lead to the same result.

- (II) We must always consider all the evidence relevant to a question. We should not arbitrarily ignore some information, basing the conclusions only on what remains.

- (III) We must always represent equivalent states of knowledge by equivalent plausibility assignments. That is, if the state of knowledge is the same (except perhaps for labeling the propositions) in two problems, then it must assign the same plausibilities in both.

An example of correspondence with common sense

First, let's learn the conditional notation: A proposition (A) whose truth is conditioned on the truth of another proposition (B) is typically denoted by,

$$A|B$$

Second, **by convention**, we will assume that propositions with greater degree of plausibility correspond to greater real numbers.

Therefore,

$$(A|B) > (C|B)$$

An example of correspondence with common sense

First, let's learn the conditional notation: A proposition (A) whose truth is conditioned on the truth of another proposition (B) is typically denoted by,

$$A|B$$

Second, **by convention**, we will assume that propositions with greater degree of plausibility correspond to greater real numbers.

Therefore,

$$(A|B) > (C|B)$$

says that, given B, A is more plausible than C.

Now, what do we mean by "correspondence with common sense" ? Suppose,

$$B|C' > B|C$$

$$A|BC' = A|BC$$

Then, the desiderata of "correspondence with commonsense" requires us to have,

$$AB|C' \geq AB|C$$

An example of correspondence with common sense

To illustrate the principle of commonsense with an example, consider the following scenario,

$$
\begin{aligned}
A &\equiv \text{The probability that I will go to school today} \\
B &\equiv \text{The probability that it will rain today} \\
C &\equiv \text{The probability that the sky will be sunny today} \\
C' &\equiv \text{The probability that the sky will be cloudy today}
\end{aligned}
$$

Therefore,

$$B|C' > B|C$$

$$A|BC' = A|BC$$

Then, the desiderata of "correspondence with commonsense" requires us to have,

$$AB|C' \geq AB|C$$

Probability is real number whose range must be either [0,1] or [1,+infinity]

To illustrate the principle of commonsense with an example, consider the following scenario,

$$
\begin{aligned}
A &\equiv \text{The probability that I will go to school today} \\
B &\equiv \text{The probability that it will rain today} \\
C &\equiv \text{The probability that the sky will be sunny today} \\
C' &\equiv \text{The probability that the sky will be cloudy today}
\end{aligned}
$$

Therefore,

$$B|C' > B|C$$

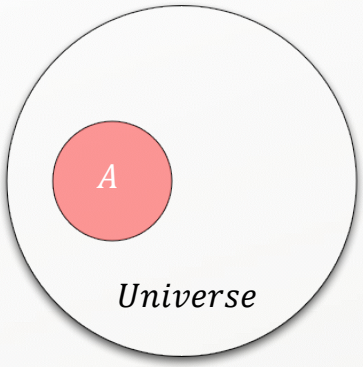
$$A|BC' = A|BC$$

Then, the desiderata of "correspondence with commonsense" requires us to have,

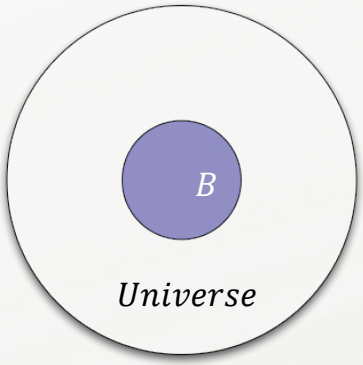
$$AB|C' \geq AB|C$$

Without going through proofs, we will state here that our three desiderates probability theory dictate that **probability is a real number whose range must be either [0,1] or [1,+infinity]**.

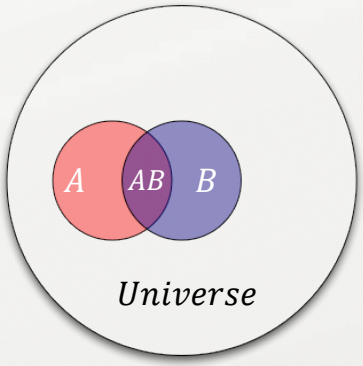
The product Rule (A prelude to the Bayesian Probability Theory)



$$P(A) = \frac{A}{U}$$



$$P(B) = \frac{B}{U}$$



$$P(AB) = \frac{AB}{U}$$

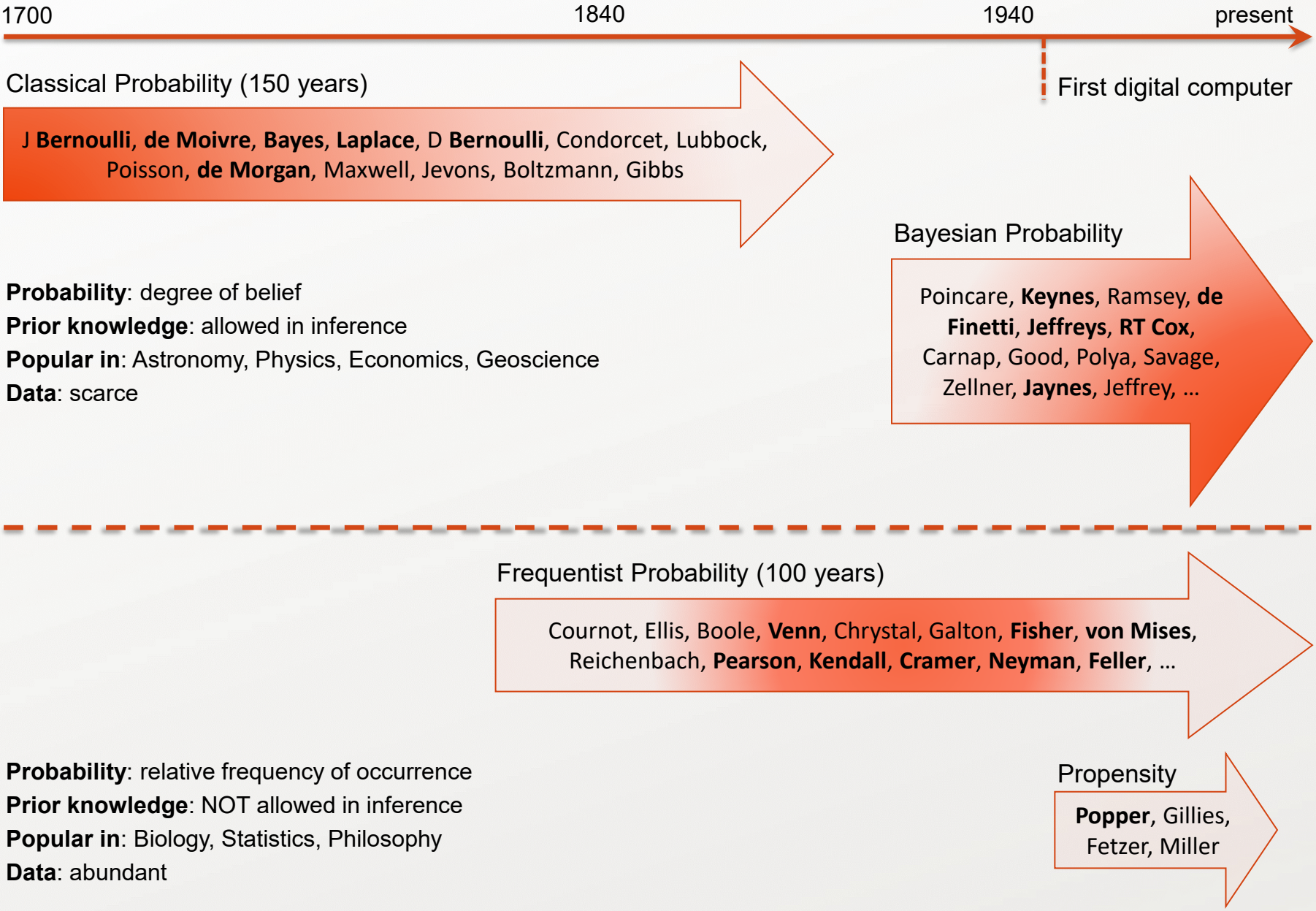
$$P(A|B) = \frac{AB}{B} = \frac{\frac{AB}{U}}{\frac{B}{U}} = \frac{P(AB)}{P(B)}$$

$$P(B|A) = \frac{AB}{A} = \frac{\frac{AB}{U}}{\frac{A}{U}} = \frac{P(AB)}{P(A)}$$

Bayes rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

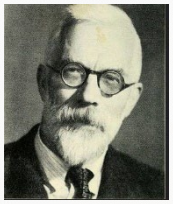
A Tug-of-War in the History of Probability Theory



Two Philosophically distinct approaches to Scientific inference

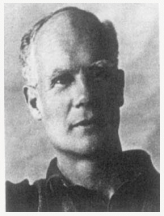
Frequentist Inference

Neyman–Pearson–Wald theory



Ronald Fisher
(1890 – 1962)
Statistician / Biologist

Jerzy Neyman
(1894 – 1981)
Statistician / Astronomer



Egon Pearson
(1895 – 1980)
Statistician

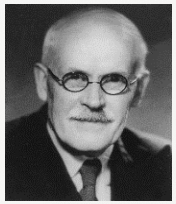
Abraham Wald
(1902 – 1950)
Statistician

No prior knowledge allowed.
Let the data speak for itself.
- R. A. Fisher

Bayesian Inference

Bayesian probability theory

Pierre Laplace
(1749 – 1827)
Astronomer / Mathematician



Harold Jeffreys
(1891 – 1989)
Astronomer / Geophysicist

Richard Cox
(1898 – 1991)
Physicist

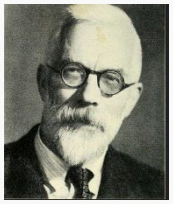
Edwin Jaynes
(1922 – 1998)
Physicist

Prior knowledge has a fundamental role
in inference along with Data
(via the Bayes' Rule)

Two Philosophically distinct approaches to Scientific inference

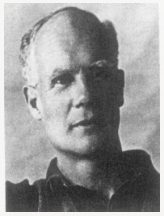
Frequentist Inference

Neyman–Pearson–Wald theory



Ronald Fisher
(1890 – 1962)
Statistician / Biologist

Jerzy Neyman
(1894 – 1981)
Statistician / Astronomer



Egon Pearson
(1895 – 1980)
Statistician

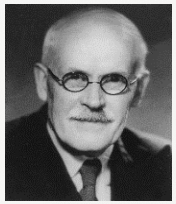
Abraham Wald
(1902 – 1950)
Statistician

No prior knowledge allowed.
Let the data speak for itself.
- R. A. Fisher

Bayesian Inference

Bayesian probability theory

Pierre Laplace
(1749 – 1827)
Astronomer / Mathematician



Harold Jeffreys
(1891 – 1989)
Astronomer / Geophysicist

Richard Cox
(1898 – 1991)
Physicist

Edwin Jaynes
(1922 – 1998)
Physicist

Bayes rule:

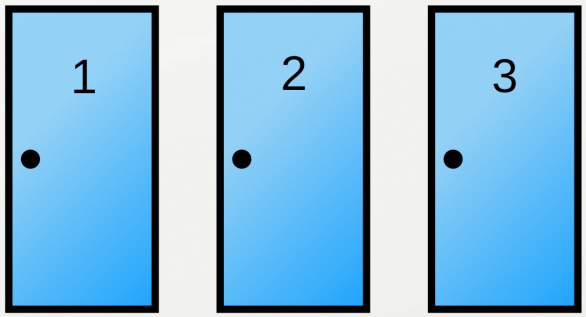
$$\underbrace{\pi(\boldsymbol{\theta}|\mathcal{D},M)}_{posterior} = \frac{\overbrace{\pi(\mathcal{D}|\boldsymbol{\theta},M)}^{likelihood}\ \overbrace{\pi(\boldsymbol{\theta}|M)}^{prior}}{\underbrace{\pi(\mathcal{D}|M)}_{evidence}}$$

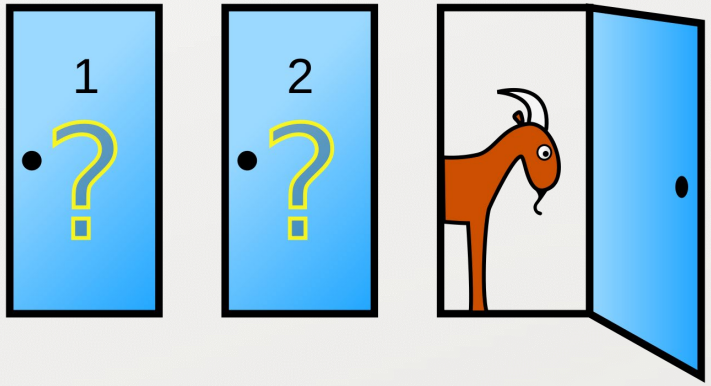
Digression: Bayes rule - The optimal method of inference

The Monty Hall Problem and Bayes Rule

Steve Selvin, 1975, "A problem in probability (letter to the editor)". *American Statistician*

Correct answer: The advantage of switching door, depends on your knowledge about the host's decision.





Digression: Bayes rule - The optimal method of inference

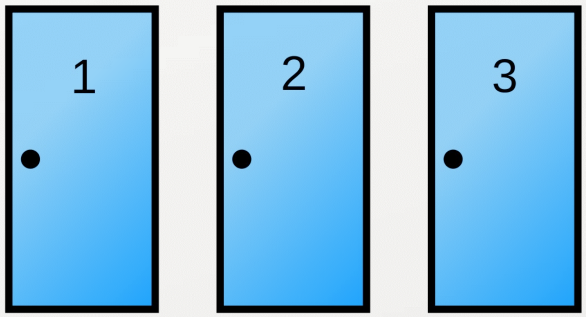
The Monty Hall Problem and Bayes Rule

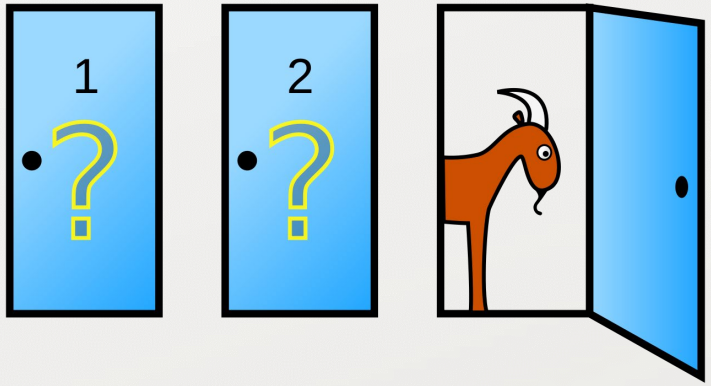
Steve Selvin, 1975, "A problem in probability (letter to the editor)". *American Statistician*

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the two others, goats. **You pick a door, say No. 1**, and **the host opens another door, say No. 3, which has a goat**. He then says to you, "Do you want to pick door No. 2?"

Question:
Is it to your advantage to switch your choice from door 1 to door 2?

Correct answer: The advantage of switching door, depends on your knowledge about the host's decision.



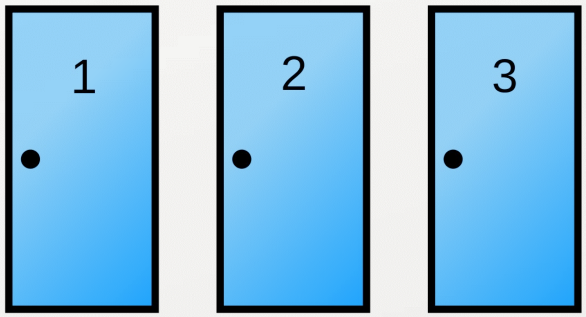


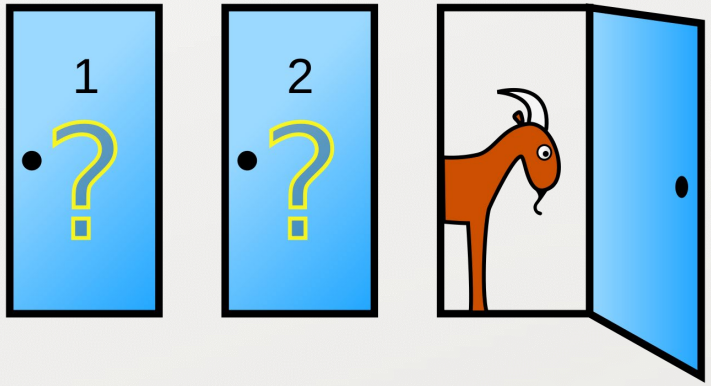
Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

You, the guest, choose door 1. The **Informed** host **consciously** opens door 3.

$$P(C2|H3, G1) = \frac{P(H3|C2, G1)\ P(C2)}{P(H3|G1)}$$





Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

You, the guest, choose door 1. The **Informed** host **consciously** opens door 3.

Prior knowledge about the car being behind door 2 (C2):

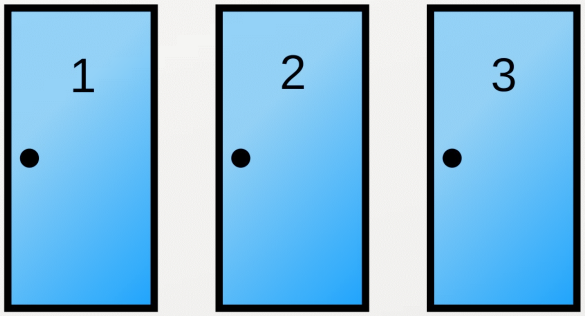
$$P(C2|G1) = P(C1|G1) = P(C3|G1) = P(C2) = \mathbf{\frac{1}{3}}$$

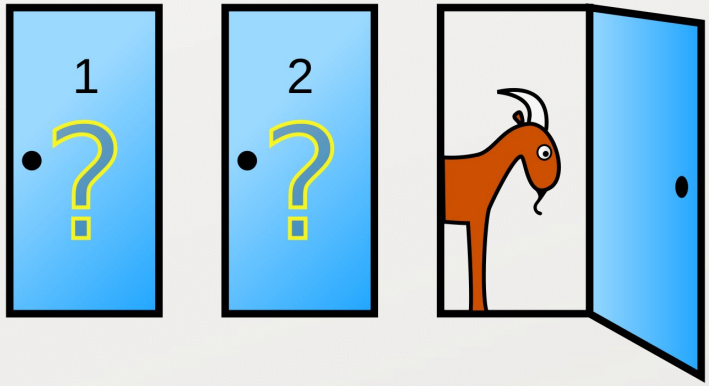
New knowledge: The informed host chooses door 3 (**H3**) (He knows that the car is behind door 2)

$$P(H3|C2, G1) = \mathbf{1} \qquad P(H3|G1) = \mathbf{\frac{1}{2}}$$

Update your knowledge about door C2 using Bayes rule:

$$P(C2|H3, G1) = \frac{P(H3|C2, G1)\ P(C2)}{P(H3|G1)} = \frac{1 \times 1/3}{1/2} = \mathbf{\frac{2}{3}}$$





Digression: Bayes rule - The optimal method of inference

Case 1: Informed Host: Knows where the car is and will not open the door that leads to car.

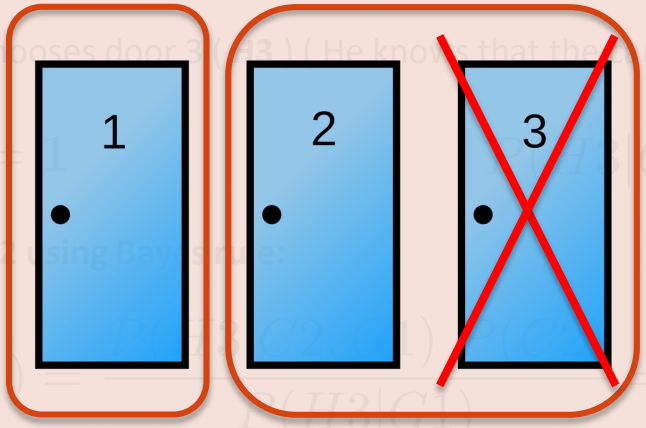
You, the guest, choose door 1. The **Informed** host **consciously** opens door 3 (knowing the car is behind #2).

Prior kn

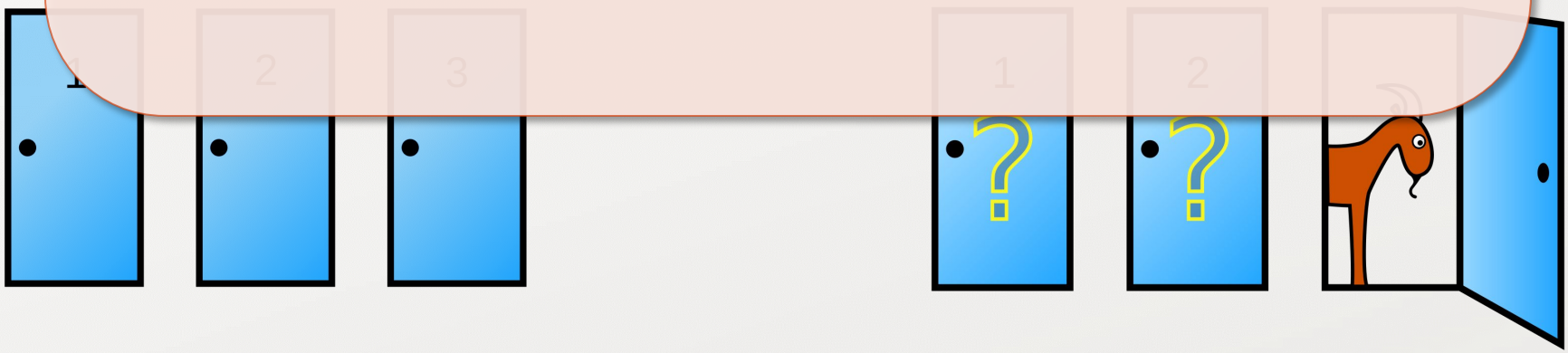
New k

Updat

$$P(C1) = \frac{1}{3}$$



$$P(C2 \cup C3) = \frac{2}{3}$$



Digression: Bayes rule - The optimal method of inference

Case 1: Uninformed Host

You, the guest, choose door 1. The **Uninformed** host **randomly** opens door 3 (knowing nothing a priori).

Prior knowledge about the car being behind door 2:

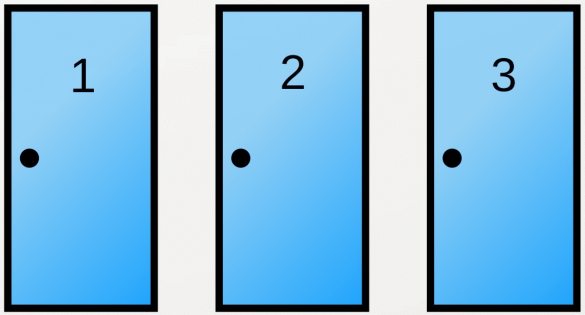
$$P(C2|G1) = P(C1|G1) = P(C3|G1) = P(C2) = \frac{\mathbf{1}}{\mathbf{3}}$$

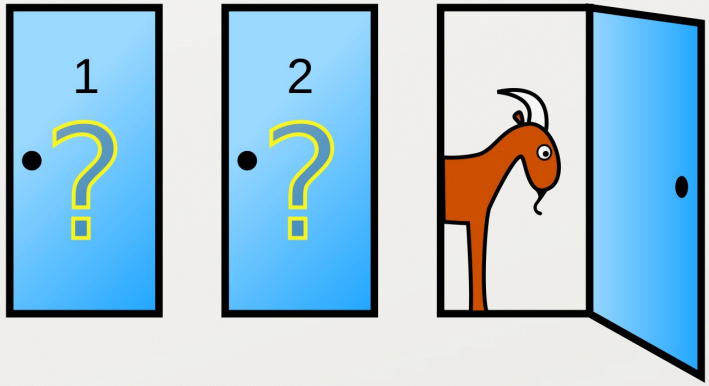
New knowledge: The **Un**informed host chooses door 3 (**H3**) (He does **not** know where the car is)

$$P(H3|C2, G1) = \frac{\mathbf{1}}{\mathbf{2}} \qquad P(H3|G1) = \frac{\mathbf{1}}{\mathbf{2}}$$

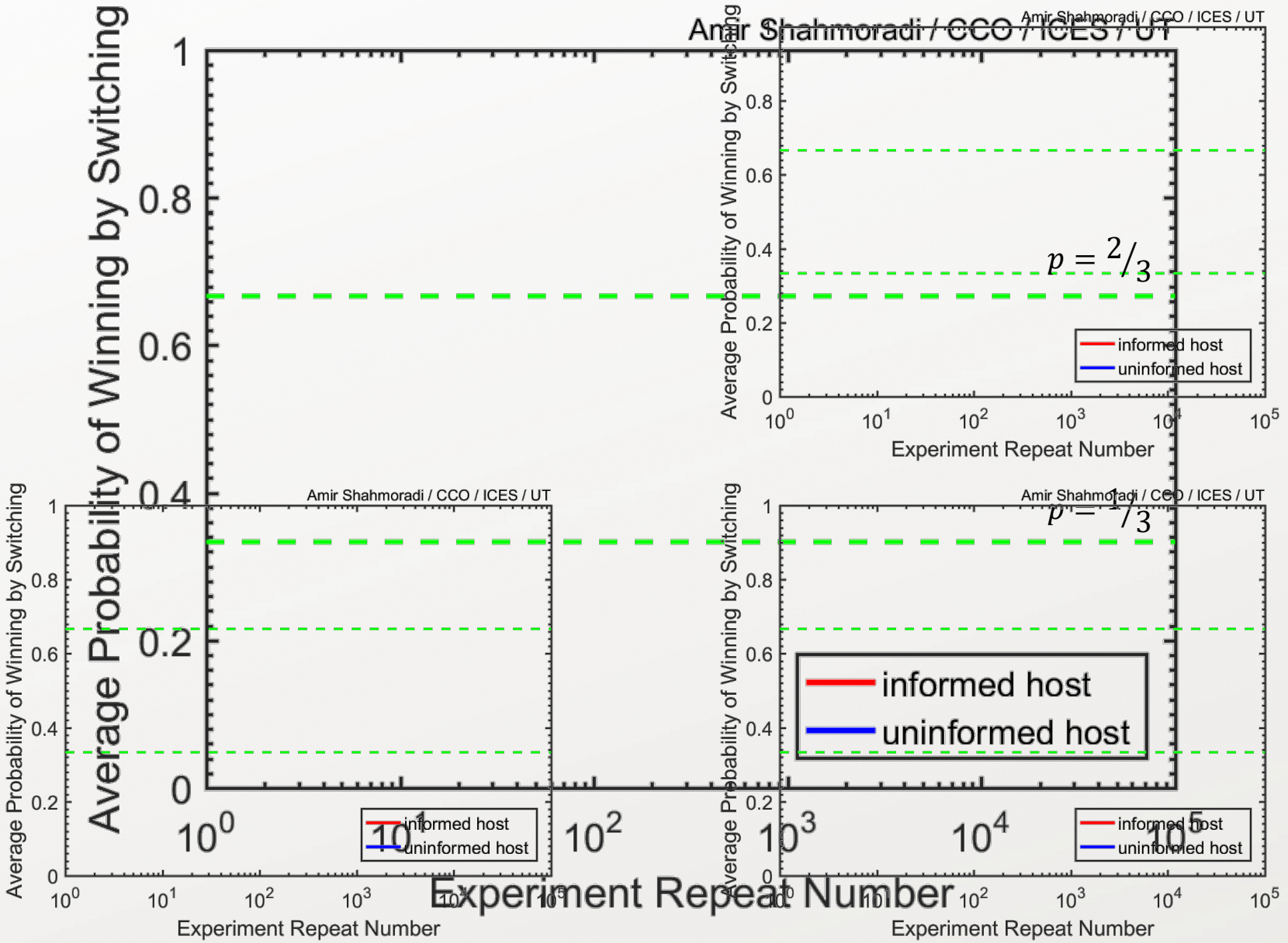
Update your knowledge about door C2 using Bayes rule:

$$P(C2|H3, G1) = \frac{P(H3|C2, G1)\ P(C2)}{P(H3|G1)} = \frac{1/2 \times 1/3}{1/2} = \frac{\mathbf{1}}{\mathbf{3}}$$





Digression: Bayes rule - The optimal method of inference



There is only one type of uncertainty in the world – epistemic

Suppose you have **blurred** vision.
You throw a die **once and** read your observation (possibly wrong reading).
What is the **type of uncertainty** in your observation?

Frequentist Inference

The uncertainty is due to my **lack of knowledge.**
I **can reduce uncertainty** with better vision**.**
Therefore, the **uncertainty** is **epistemic**.

Bayesian Inference

The uncertainty is due to my **lack of knowledge.**
I **can reduce uncertainty** with better vision**.**
Therefore, the **uncertainty** is **epistemic**.

There is only one type of uncertainty in the world – epistemic

Suppose you have **perfect** vision.
You throw a die **multiple times and** read your observations.
What is the **type of uncertainty (source of variability)** in your observations?

Frequentist Inference

The uncertainty is **inherent in the experiment.**
I **cannot reduce** uncertainty any further**.**
Therefore, the **uncertainty** is **aleatoric**.

Bayesian Inference

The uncertainty is due to my **lack of knowledge:**

1. Wrong / **inadequate** model.
2. Lack of sufficiently-detailed data which leads to inadequate model.

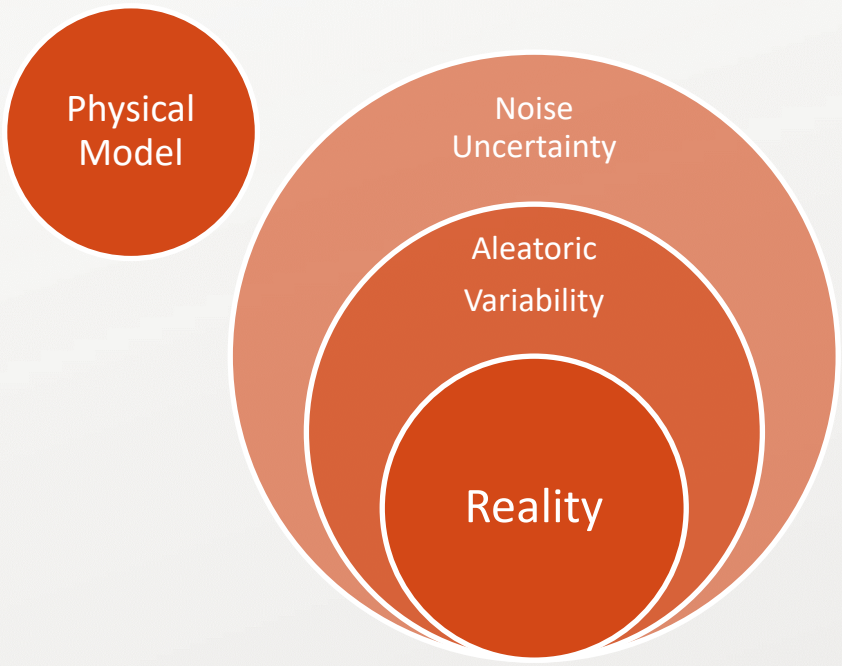
I **can reduce uncertainty** with better data / model**.**
Therefore, the **uncertainty** is **epistemic**.

There is only one type of uncertainty in the world – epistemic

Suppose you have **perfect** vision.
You throw a die **multiple times and** read your observations.
What is the **type of uncertainty (source of variability)** in your observations?

Frequentist Inference

The uncertainty is **inherent in the experiment.**
I **cannot reduce** uncertainty any further**.**
Therefore, the **uncertainty** is **aleatoric**.

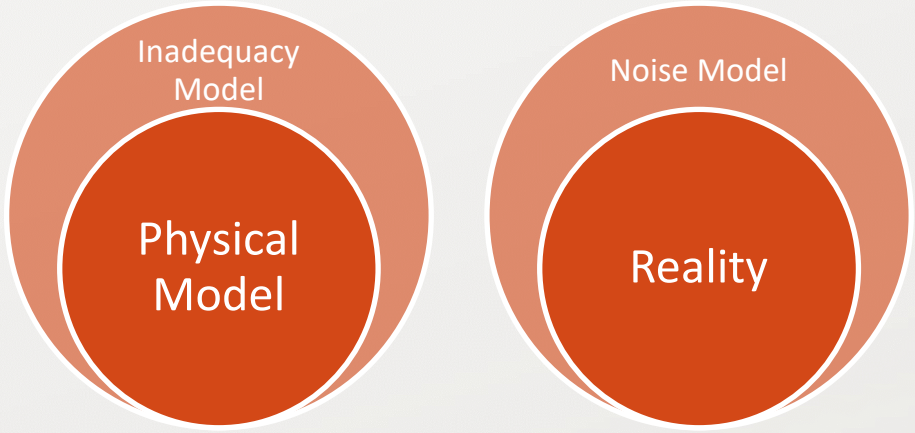


Bayesian Inference

The uncertainty is due to my **lack of knowledge:**

1. Wrong / **inadequate** model.
2. Lack of sufficiently-detailed data which leads to inadequate model.

I **can reduce uncertainty** with better data / model**.**
Therefore, the **uncertainty** is **epistemic**.



Observations, random variables, and the likelihood principle

Suppose we observe the following data of n events, each described by m variables,

$$\mathcal{D} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, ..., \boldsymbol{D}_n\}$$

$$m = 5 \text{ attributes, variables, features, or dimensions}$$

Data tuples

$DATA$	$\boldsymbol{X}$	$\boldsymbol{Y}$	$\boldsymbol{Z}$	$\boldsymbol{t}$	ϕ
$\boldsymbol{D_1}$	x_1	y_1	z_1	t_1	ϕ_1
$\boldsymbol{D_2}$	x_2	y_2	z_2	t_2	ϕ_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{D_n}$	x_n	y_n	z_n	t_n	ϕ_n

$\mathcal{D}$

Philosophical counterfactual assumption in statistical modeling

(John Stuart Mill, A system of Logic, Ratiocinative and Inductive, 1843)

Even though we have observed dataset $\mathcal{D}$, the experiment outcome could be different if repeated. In other words, $\boldsymbol{D}$ is a **random variable**; a real-valued function with a **sample space** as its domain.

What is a Data Object?

Data sets are made up of **data objects** (or **data points**). A data object represents an entity.

Examples:

- In a sales database, the objects may be customers, store items, and sales
- In a medical database, the objects may be patients
- In a university database, the objects may be students, professors, and courses.

Data objects are typically **described by attributes**.

Synonyms for **Data object**: **sample**, **example**, **instance**, **data point**, or **observation**.

If the data objects are stored in a database, they are also called **data tuples**. That is, the **rows** of a database **correspond to the data objects**, and the **columns correspond to the attributes**.

What is a Data Attribute?

An **attribute** is a data field, representing a **characteristic** or **feature** of a data object.

Common synonyms for attribute are,
- **dimension** (frequently used in data warehousing),
- **feature** (frequently used in Machine Learning),
- **Variable** (frequently used by statisticians),

Example:

- Attributes describing a customer object can include,
  - customer ID,
  - name,
  - address.

Observed values for a given attribute are frequently known as **observations**.

A set of attributes used to describe a given object is called an **attribute vector** (or **feature vector**).

The **distribution of data** involving **one attribute** (or variable) is called **univariate**.
The **distribution of data** involving **two attributes** (or variables) is called **bivariate**.
The **distribution of data** involving **multiple attributes** (or variables) is called **multivariate**.

What are the kinds of Data Attribute?

- **Nominal attributes**
  - Nominal means "**relating to names**". The values of a nominal attribute are **symbols** or **names of things**. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as **categorical**. The values do not have any meaningful order. In computer science, the values are also known as **enumerations**.
  - Example: hairColor, maritalStatus, gender, occupation, …
- **Binary Attributes**
  - A **binary attribute** is **a nominal attribute** with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as **Boolean** if the two states correspond to *true* and *false*.
  - Example: The "smoker" attribute in a patient data object.
- **Ordinal Attributes**
  - An ordinal attribute is an attribute with possible values that **have a meaningful order or ranking among them**, but the **magnitude between successive values is NOT known.**
  - Example:
    - Starbucks coffee cup size: small, medium, grande
    - School grading system: F,D,C,B,A
  - Ordinal attributes are useful for registering **subjective assessments of qualities** that cannot be measured objectively; thus, **ordinal attributes are often used in surveys for ratings**. For example, customer satisfaction in a survey can have the following ordinal categories: 0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and 4: very satisfied.

What are the kinds of Data Attribute?

- **Numeric Attributes**
  - A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values.
  - Example: room temperature, volume, air density, …
  - Numeric attributes can be **location** (**interval-scaled)** or **scale** (**ratio-scaled) attributes**.
    - **Location Attributes** do not have a starting value (or zero-point). For example, location of trees in a forest does not a reference point. The temperature in Celsius or Fahrenheit is also a location attribute.
    - **Scale Attributes**, unlike location attributes, do have a zeros point, for example, the time since an event, or the temperature in Kelvins.

Data Attribute classification based on the values

- **Continuous Attributes**
  - An attribute that can take an **uncountably infinite** number of possible values is **continuous**. Most numeric attributes are also continuous.
- **Discrete Attributes**
  - An attribute that can take only a **finite** or a **countably infinite** number of possible values is **discrete**. Integer numeric attributes are common examples of discrete variables. Ordinal and binary attributes are also discrete.

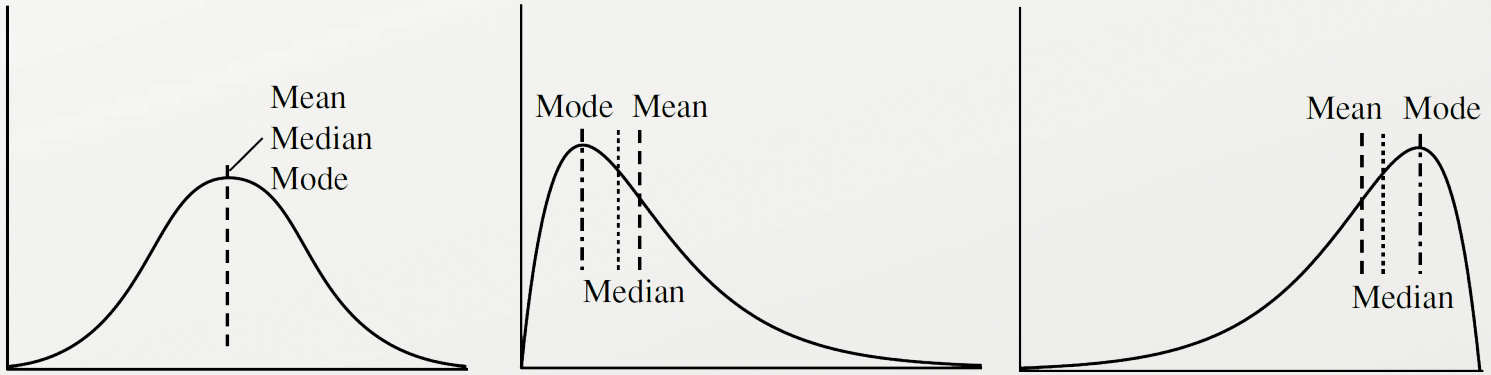
Basic Data Summarization Techniques

- **Measuring the Central Tendency (mean, median+)**
  - The most common and effective numeric measure of the "center" of a set of data is the (**arithmetic**) **mean**. The mean **statistic** is extremely useful method of summarizing **symmetric** data.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Arithmetic mean

  - When data is skewed or asymmetric, the **median** statistic provides a less biased summary of data. Median is simply the is the middle value in a set of ordered data values.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Basic Data Summarization Techniques

- **Measuring the dispersion of data (variance, interquartile range)**
  - **Variance**
    - The most common and effective numeric measure of the "dispersion" of the values of an attribute in a dataset is the (square root of) **variance**. Variance is an extremely useful statistic for quantifying uncertainty in data.
    - Let $x_1, x_2, \ldots, x_N$ be a set of observations for a given numeric attribute, X. Then, variance is simply average of the dispersion (distance-squared) of data from its center (mean $\boldsymbol{\mu}$).
    - Just like the mean of an attribute, variance is also sensitive to the presence of outliers. As such, other measures of the spread of data are also commonly used when there is a suspicion for the presence of outliers.
    - The square-root of variance has the special name **Standard Deviation** in statistics and is one of the most effective measures of spread of data **around the mean**. As such, **it should be used and reported only when mean is used to quantify the central tendency of data.**

Basic Data Summarization Techniques

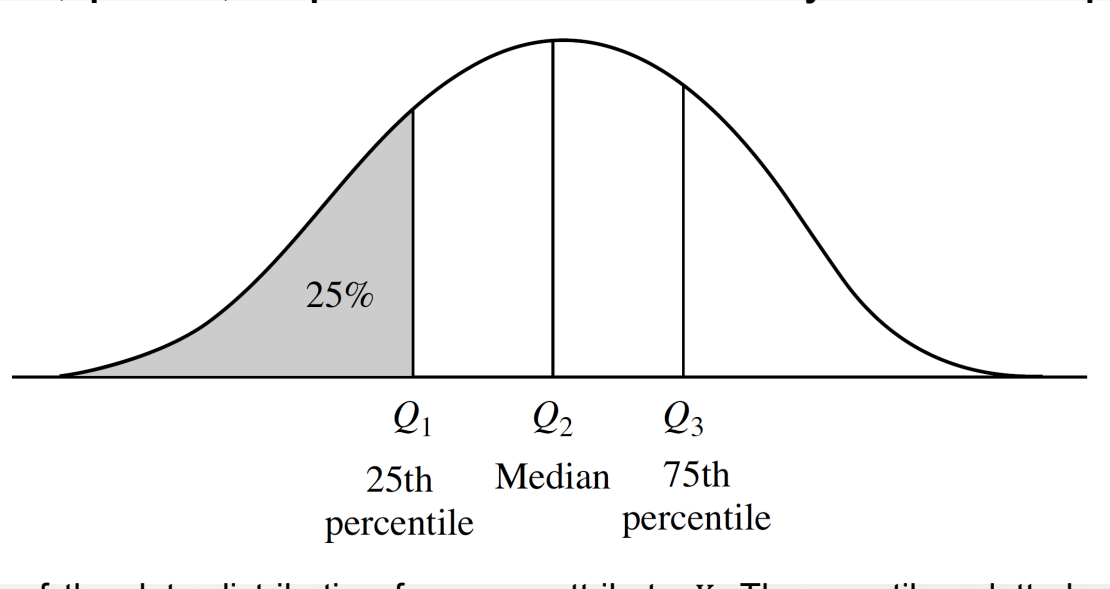
- **Measuring the dispersion of data (variance, interquartile range)**
  - **Variance**
    - The most common and effective numeric measure of the "dispersion" of the values of an attribute in a dataset is the (square root of) **variance**. Variance is an extremely useful statistic for quantifying uncertainty in data.
    - Let $x_1, x_2, \dots, x_N$ be a set of observations for a given numeric attribute, X. Then, variance is simply average of the dispersion (distance-squared) of data from its center (mean $\boldsymbol{\mu}$).
    - Just like the mean of an attribute, variance is also sensitive to the presence of outliers. As such, other measures of the spread of data are also commonly used when there is a suspicion for the presence of outliers.
    - The square-root of variance has the special name **Standard Deviation** in statistics and is one of the most effective measures of spread of data **around the mean**. As such, **it should be used and reported only when mean is used to quantify the central tendency of data.**
    - **Remark**: An observation is unlikely to be more than several standard deviations away from the mean. Mathematically, using Chebyshev's inequality, it can be shown that at least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the observations are no more than k standard deviations from the mean. In other words, at most $\frac{1}{k^2}$ fraction of the observations can be farther than k standard deviations from the mean.

Basic Data Summarization Techniques

- **Measuring the dispersion of data (variance, interquartile range)**
  - **Range**
    - Let $x_1, x_2, \ldots, x_N$ be a set of observations for a given numeric attribute, X. The **range** of the set is the difference between the largest (**max()**) and smallest (**min()**) values.
    - Example:
      - The range of [-10,0,3,1,-15,22.5] is [-15, 22.5].
    - The range statistic is a special case of a more general summary statistic named **Inter-quantile Range**.
  - **Quantile**
    - Suppose that the data for attribute X are sorted in increasing numeric order. Imagine that we can pick certain data points so as to split the data distribution into equal-size consecutive sets. These data points are called **quantiles**. Quantiles are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets. Sometimes, there may not be data values of X that divide the data into exactly equal-sized subsets. The kth q-quantile for a given data distribution is the value x such that at most $k = q$ of the data values are less than x and at most $(q - k)/q$ of the data values are more than x, where k is an integer such that $0 < k < q$. There are $q - 1$ q-quantiles.

Basic Data Summarization Techniques

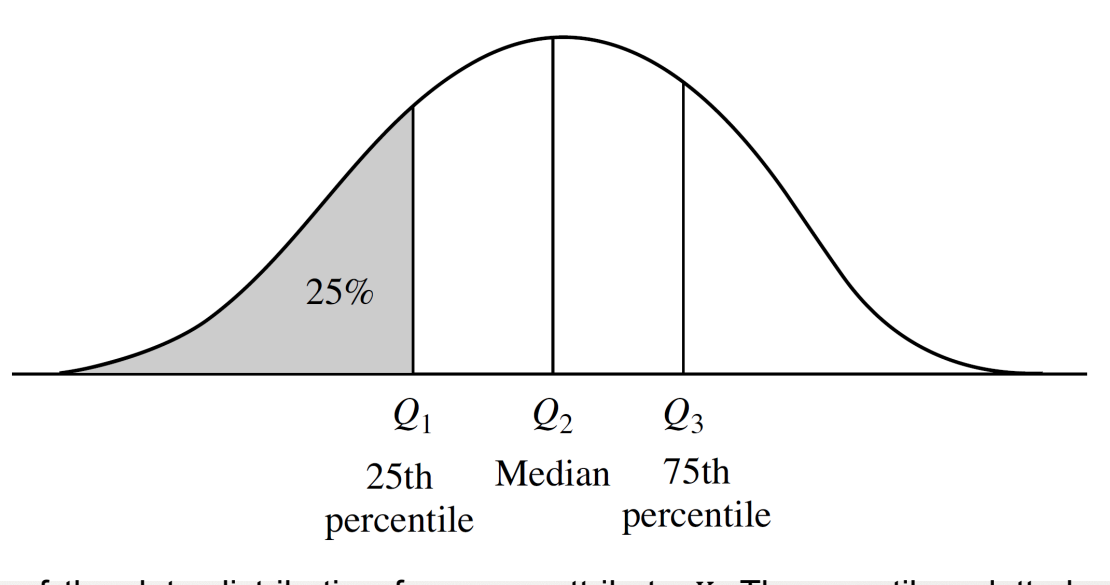
- **Measuring the dispersion of data (variance, interquartile range)**
  - **Quantile**
    - The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median. The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles**. The 100-quantiles are more commonly referred to as **percentiles**; they divide the data distribution into 100 equal-sized consecutive sets. The **median**, **quartiles**, and **percentiles** are **the most widely used forms of quantiles**.



An example illustration of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The **second quartile** corresponds to the median. The quartiles give an indication of a distribution's center, spread, and shape. The **first quartile**, denoted by $\boldsymbol{Q1}$, is the 25th percentile. It cuts off the lowest 25% of the data. The **third quartile**, denoted by $\boldsymbol{Q3}$, is the 75th percentile; it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

Basic Data Summarization Techniques

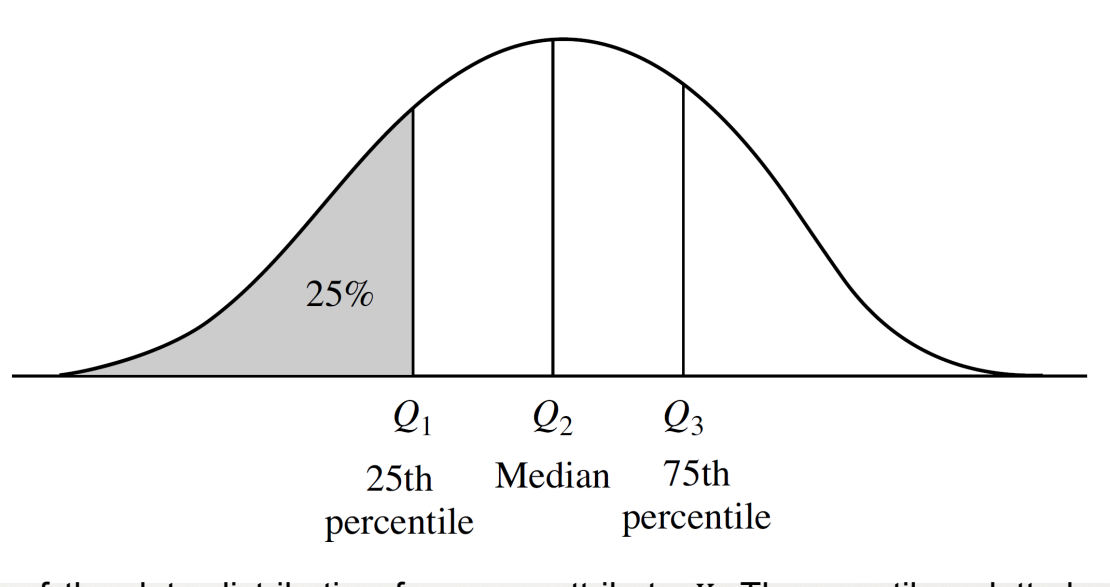
- **Measuring the dispersion of data (variance, interquartile range)**
  - **Interquartile Range (IQR)**
    - The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the **interquartile range (IQR)** and is defined as $IQR = Q3 - Q1$.
    - IQR statistic is less sensitive to the presence of outliers, unlike variance.



An example illustration of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median. The quartiles give an indication of a distribution's center, spread, and shape. The first quartile, denoted by $\boldsymbol{Q1}$, is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by $\boldsymbol{Q3}$, is the 75th percentile; it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

Basic Data Summarization Techniques

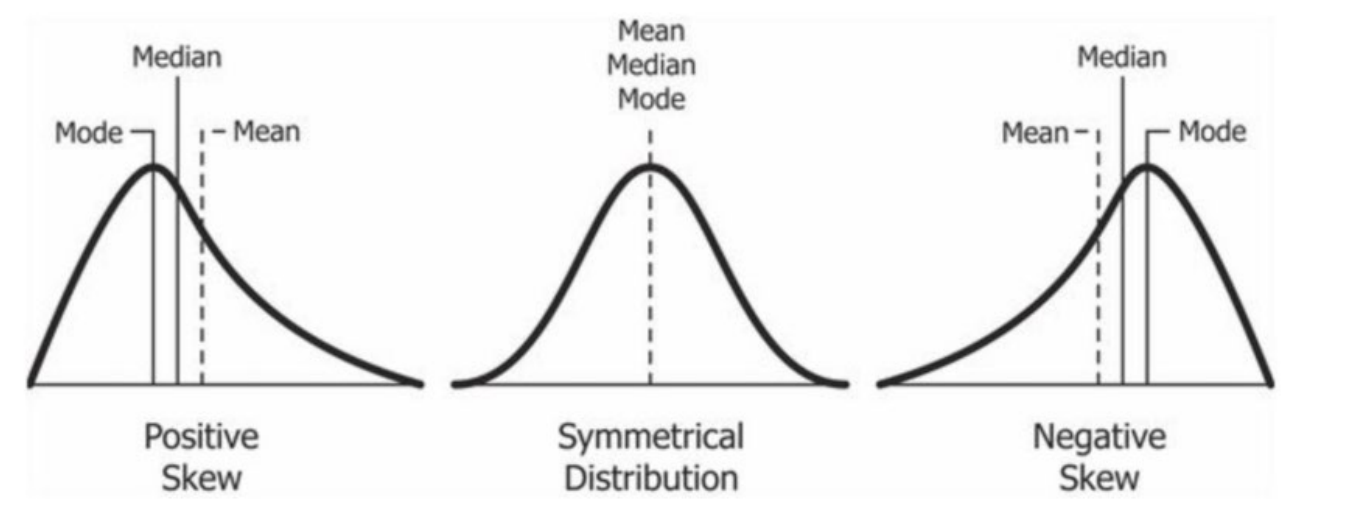
- **Measuring the dispersion of data (variance, interquartile range)**
  - **The five-number summary**
    - When data is highly skewed, even the interquartile range can become ineffective summary of data. In such a case, it is more sensible to report the interquartile range of the attribute along with the median (the value at 50%-quantile), as well as the minimum and the maximum of the values. The five-number summary is best illustrated by a special type of visualization known as the **Box-Plot** as we shall see later on in this course.



An example illustration of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median. The quartiles give an indication of a distribution's center, spread, and shape. The first quartile, denoted by $\boldsymbol{Q1}$, is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by $\boldsymbol{Q3}$, is the 75th percentile; it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

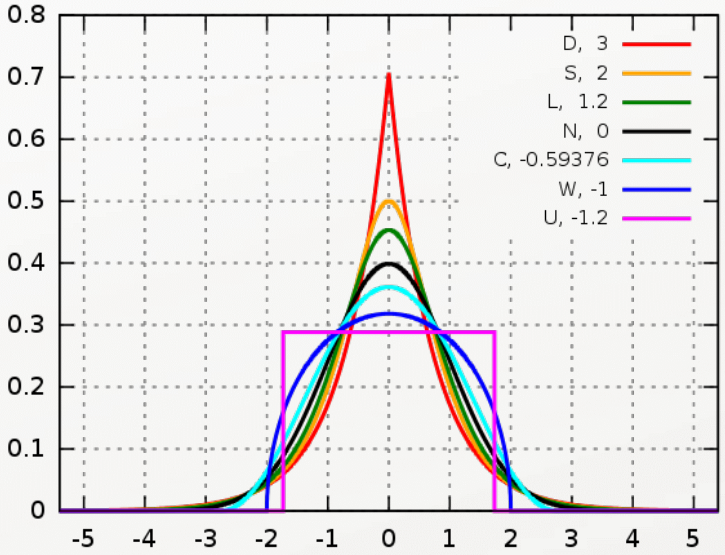
Basic Data Summarization Techniques (optional)

- **Measuring the asymmetry of data (skewness)**
  - **Skewness**
    - The most common measure of the asymmetry in the values of an attribute of a dataset is the **skewness**.
    - An attribute with **positive skewness** has a **few extremely large positive values** relative to the central tendency of data as measured by attribute's mean. Its histogram has a long tail toward large positive values.
    - An attribute with **negative skewness** has a **few extremely large negative values** relative to the central tendency of data as measured by attribute's mean. Its histogram has a long tail toward large negative values.
    - Skewness can be negative or positive, just like the **mean**.

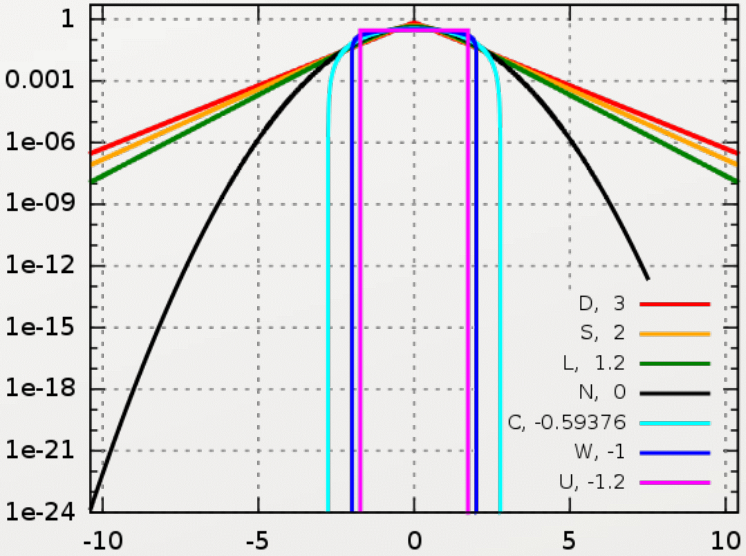


Basic Data Summarization Techniques (optional)

- **Measuring the tailedness of data (kurtosis)**
  - **(Excess-) Kurtosis**
    - The most common measure of the (heavy-)**tailedness** of the values of an attribute of a dataset is the **Kurtosis**. Kurtosis is, by definition, a **positive number**, just like the variance.
    - Since the **Normal distribution** has a kurtosis of **3**, Kurtosis is commonly normalized to 3 to obtain the **excess kurtosis** (with respect to the Normal distribution).
    - **Higher kurtosis** corresponds to **greater extremity of deviations** (or outliers) compared to the center of data.
    - An attribute with **zero excess kurtosis** are called **mesokurtic** (for example, Normally distributed).
    - A distribution with **positive excess kurtosis** is called **leptokurtic**, meaning that more outliers are present in the tails of data.
    - A distribution with **negative excess kurtosis** is called **platykurtic**, meaning that it has thinner tails (and fewer outliers in the tails).

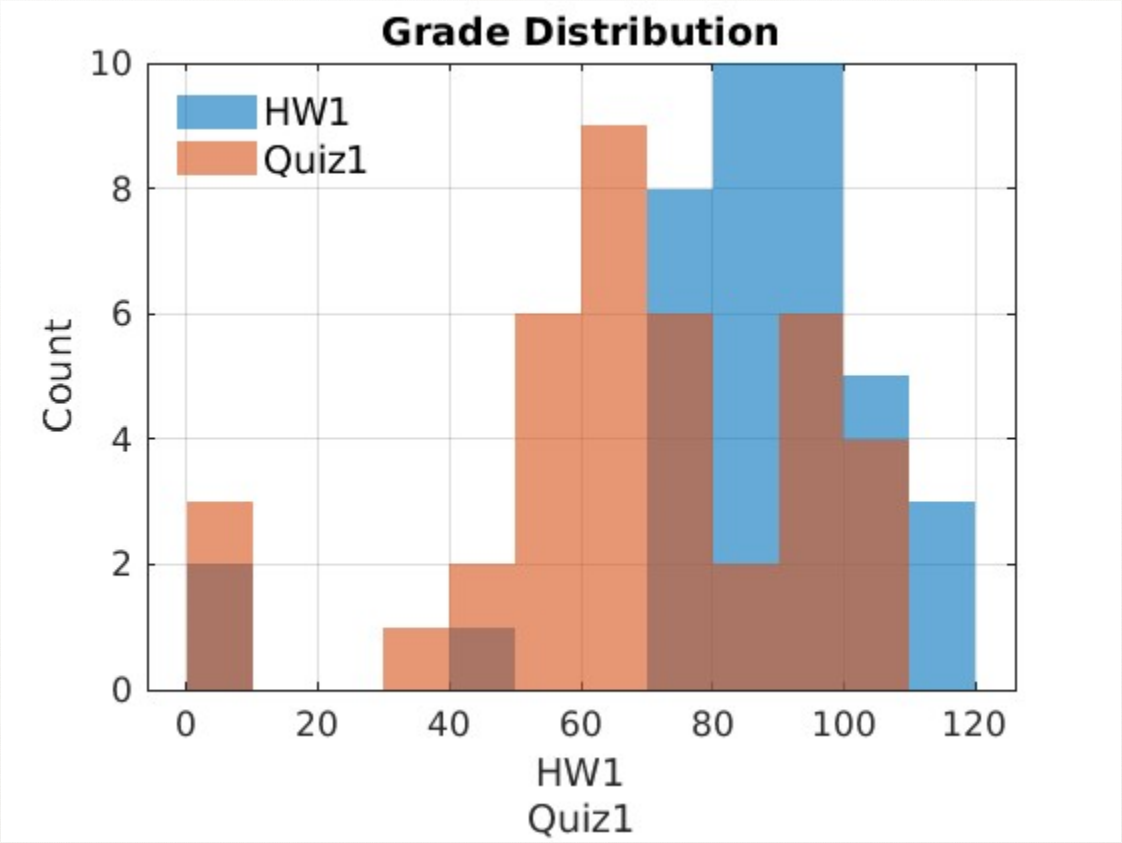


Probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis.



Logarithms of probability density functions for selected distributions with mean 0, variance 1 and different excess kurtosis.

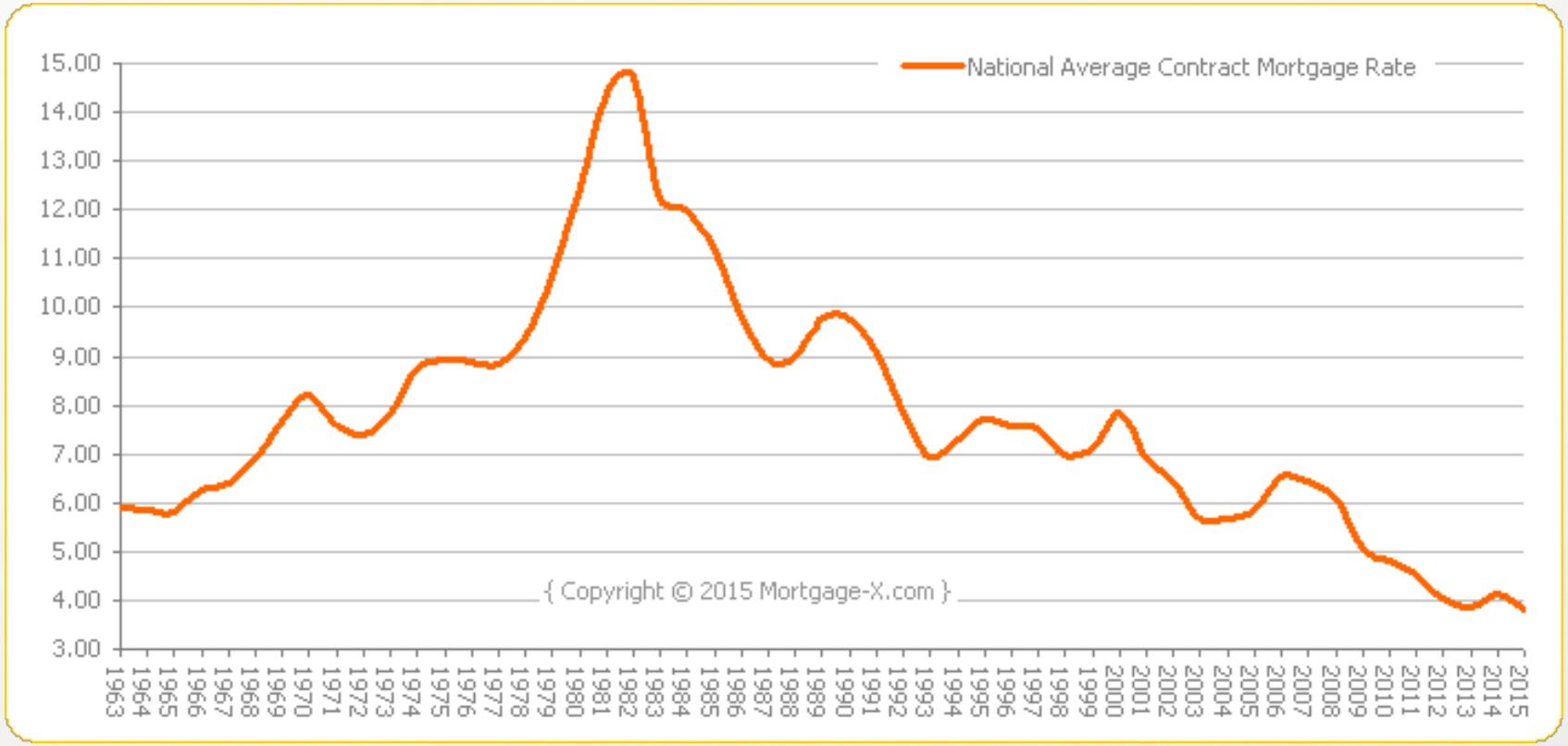
Basic Data Summarization Techniques



Time Series Data

A **time series** is a series of observations that are sequentially ordered. Typically, this order is determined by the times of observations.

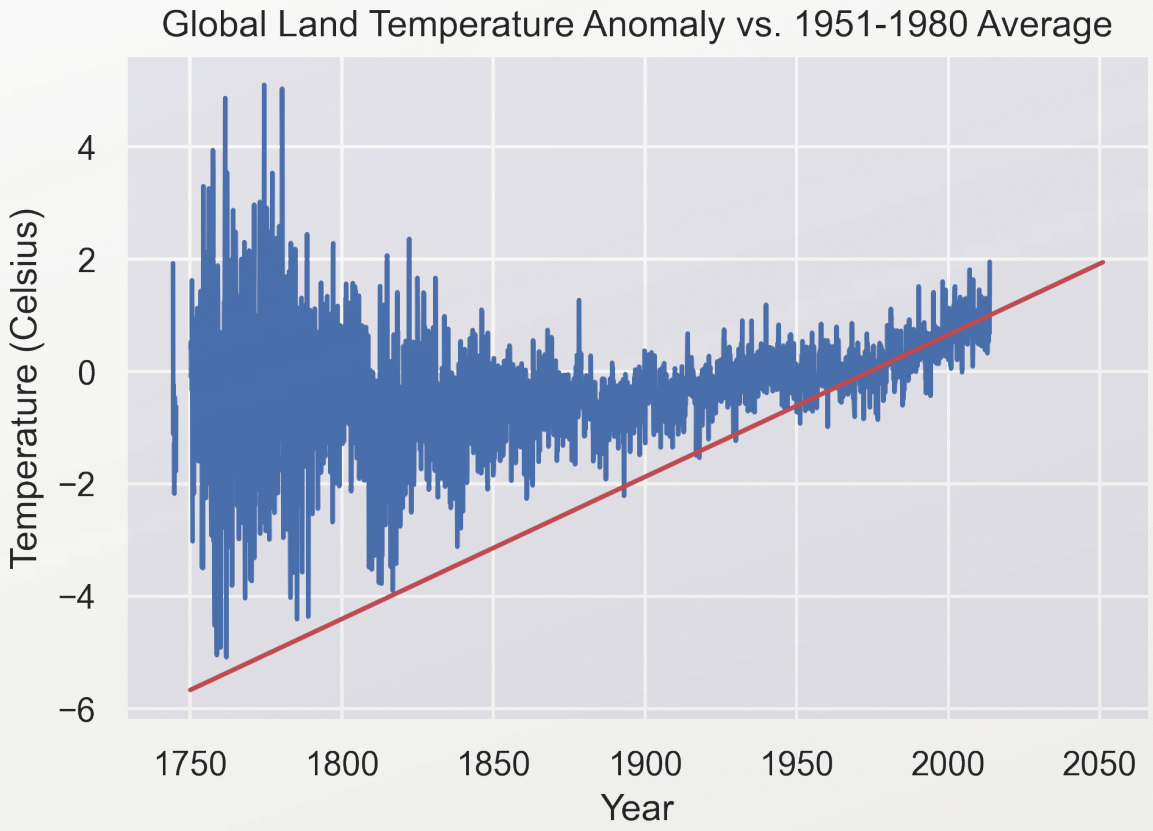
Example Time Series data: The history of US Mortgage rates



Time Series Data

A **time series** is a series of observations that are sequentially ordered. Typically, this is order is determined by the times of observations.

Example Time Series data: Global temperature increase with a **linear** hypothesis



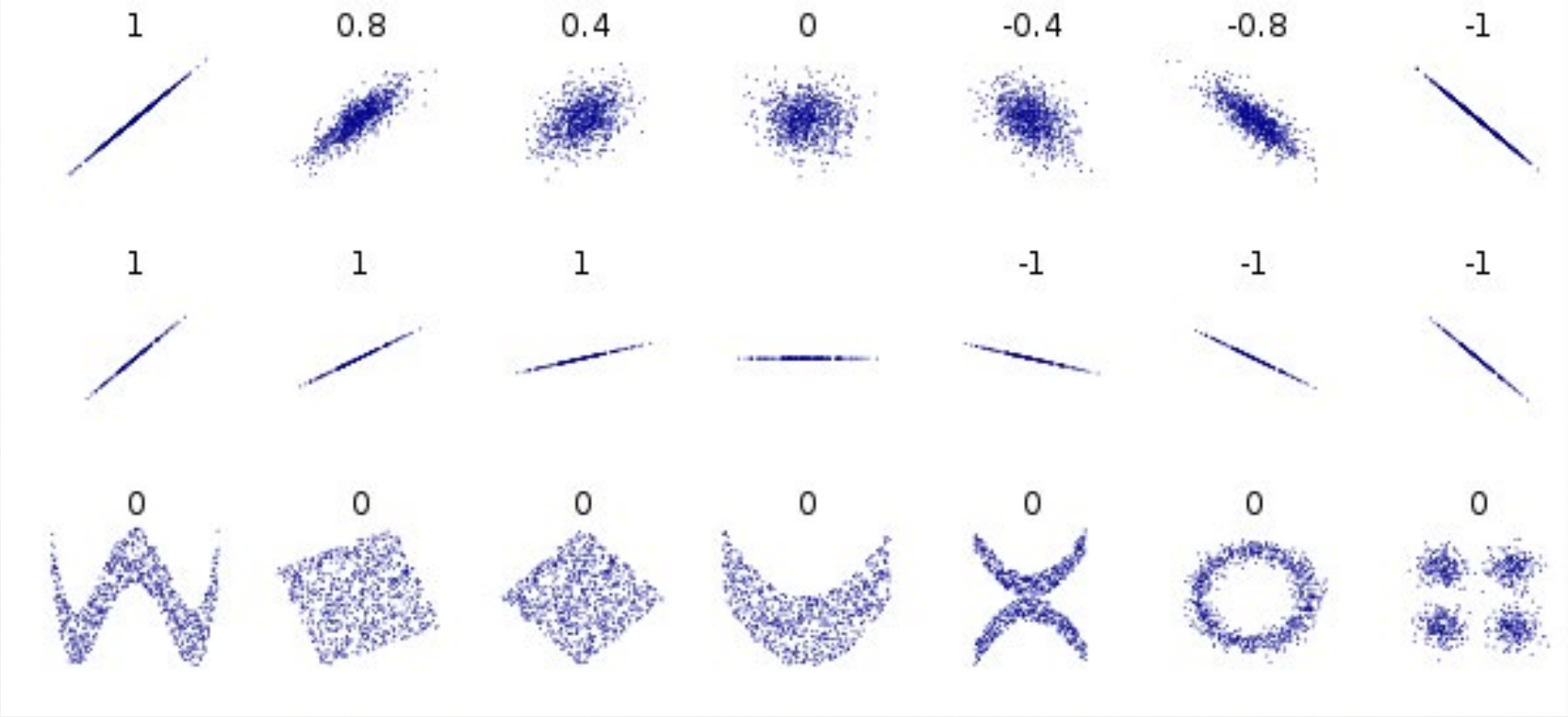
Correlation, Dependence, and Causation

Correlation is a statistical concept, typically quantified by a real number between -1 and 1, that represents the amount of co-variation of two random variables with each other.

NOTE: Correlation does NOT necessitate causation.

NOTE: Lack of (linear) correlation does NOT mean independence.

Correlations could be linear or non-linear

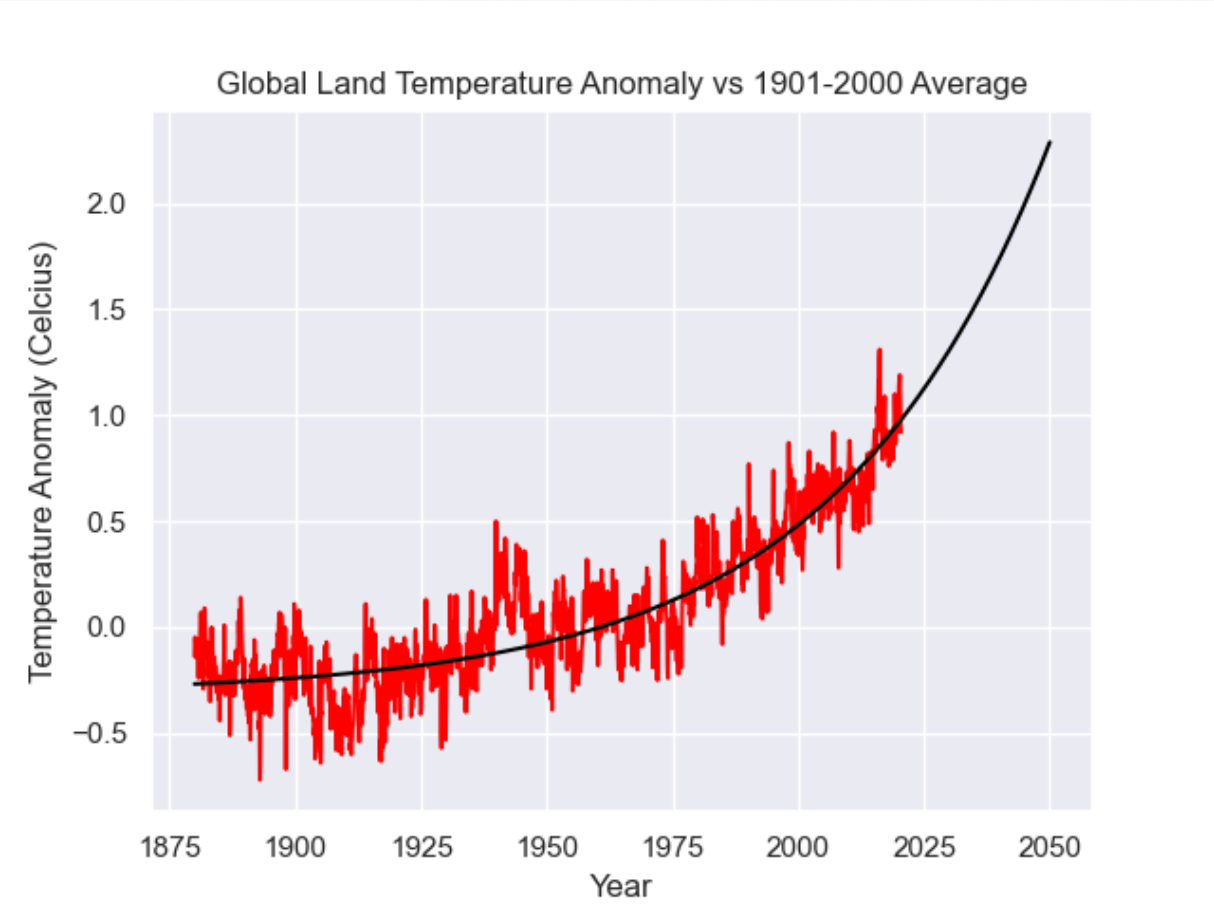


Correlation, Dependence, and Causation

Correlation is a statistical concept, typically quantified by a real number between -1 and 1, that represents the amount of co-variation of two random variables with each other.

NOTE: Correlation does NOT necessitate causation.

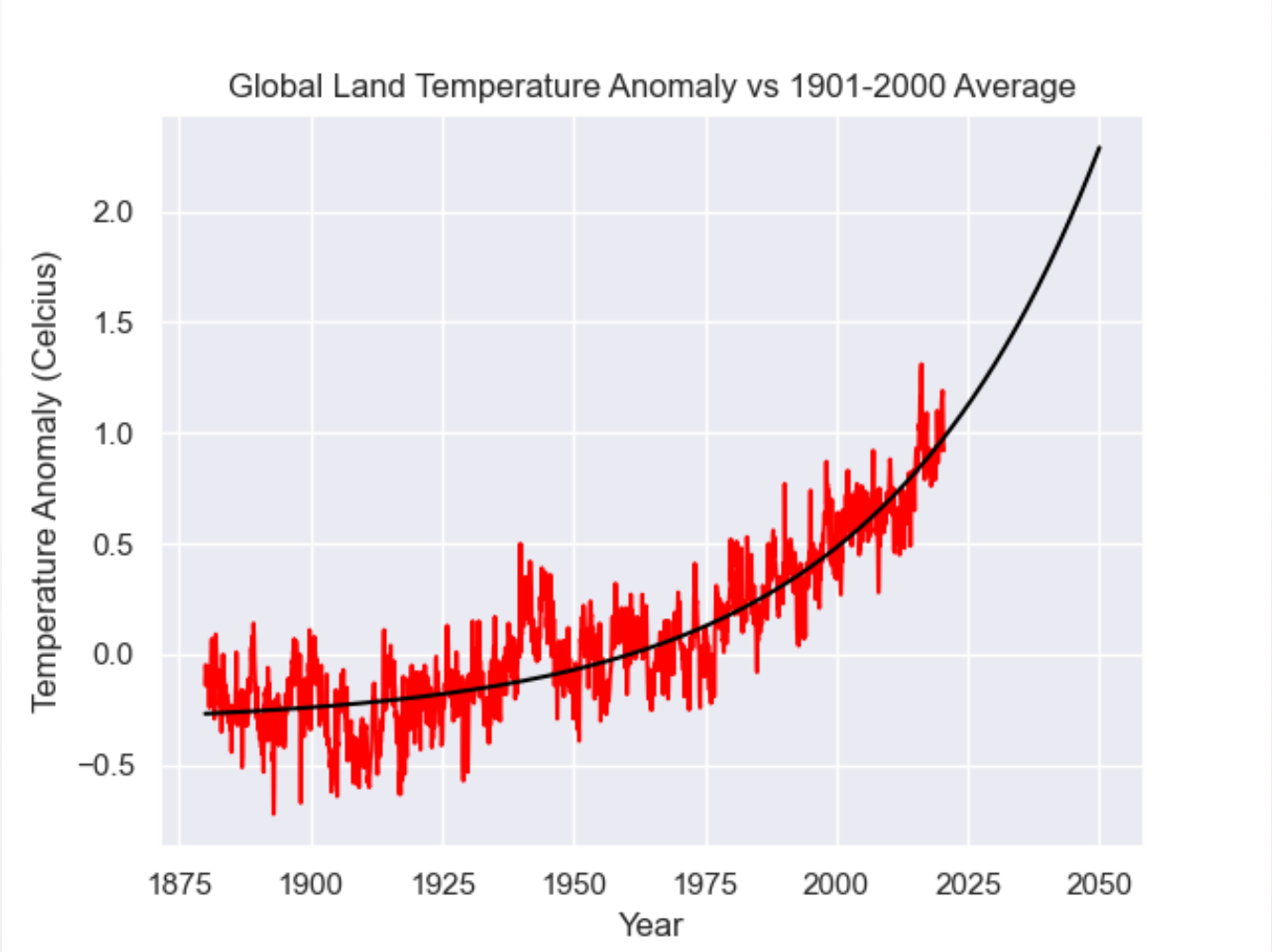
Correlations could be linear or non-linear



Time Series Data

A **time series** is a series of observations that are sequentially ordered. Typically, this is order is determined by the times of observations.

Example Time Series data: Global temperature increase with an **exponential** hypothesis

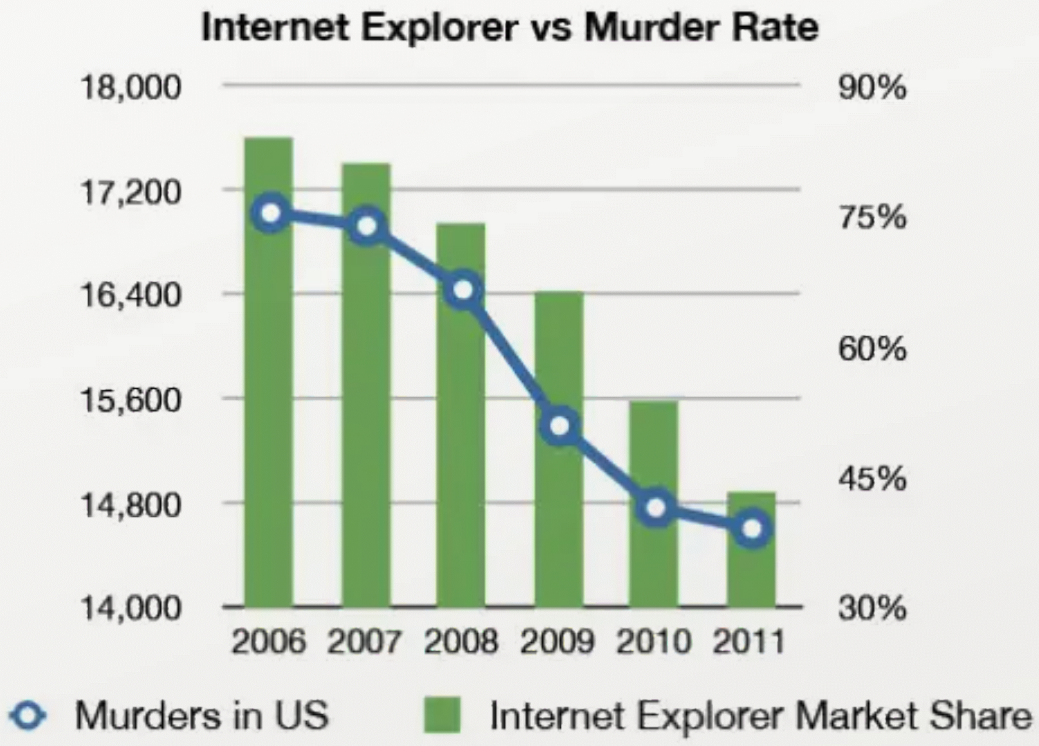


Correlation, Dependence, and Causation

Correlation is a statistical concept, typically quantified by a real number between -1 and 1, that represents the amount of co-variation of two random variables with each other.

NOTE: Correlation does NOT necessitate causation.

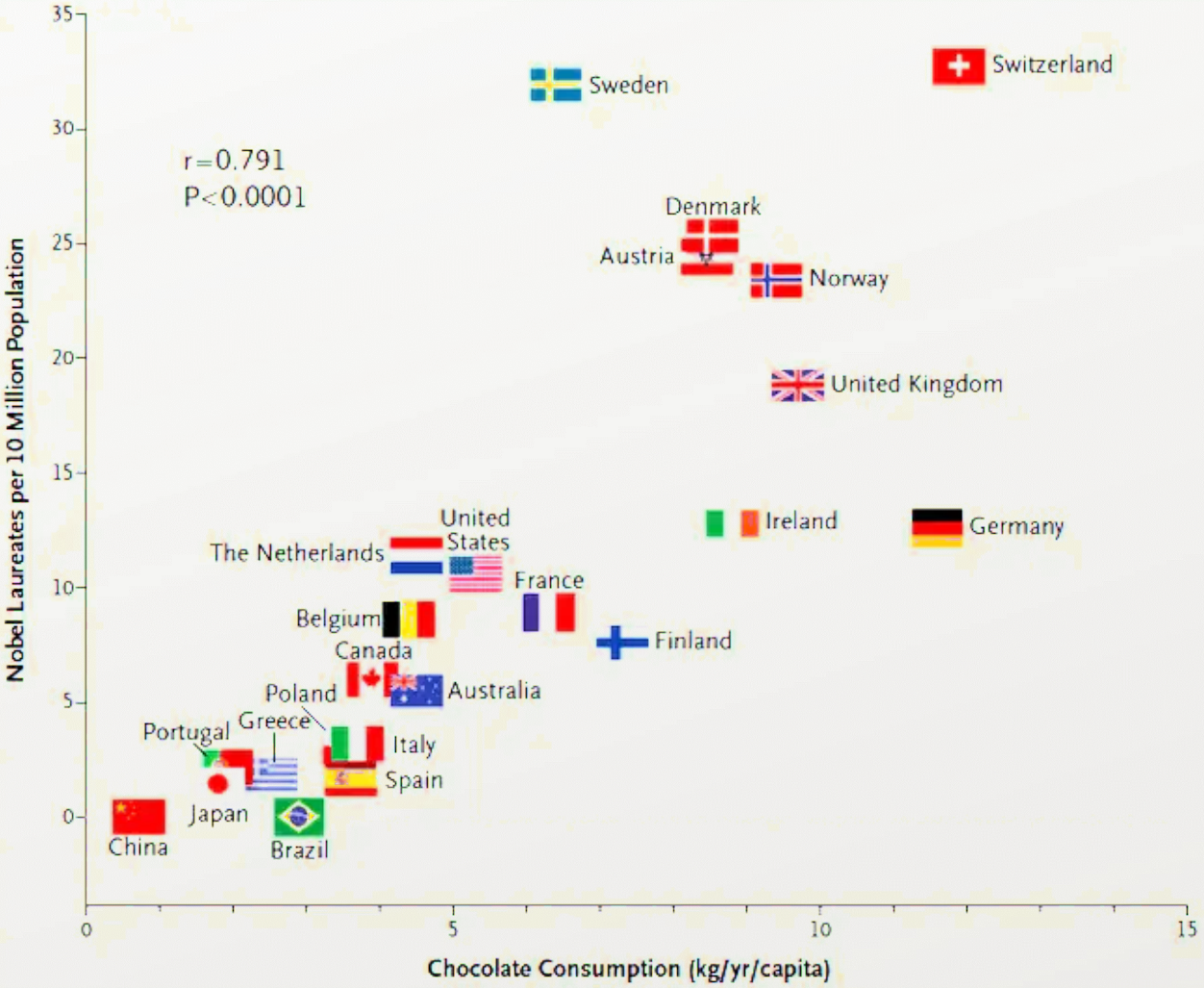
Examples of correlation but no causation:
Internet Explorer usage and Murder Rate in the US



Correlation, Dependence, and Causation

Correlation does NOT necessitate causation.

Eating more chocolate leads to winning more Nobel Prizes! **Wrong**



Correlation, Dependence, and Causation

Correlation does NOT necessitate causation.

Drinking more milk leads to winning more Nobel Prizes! **Wrong**