# Data, Uncertainty and Random Variables

## Informal Definition of Data

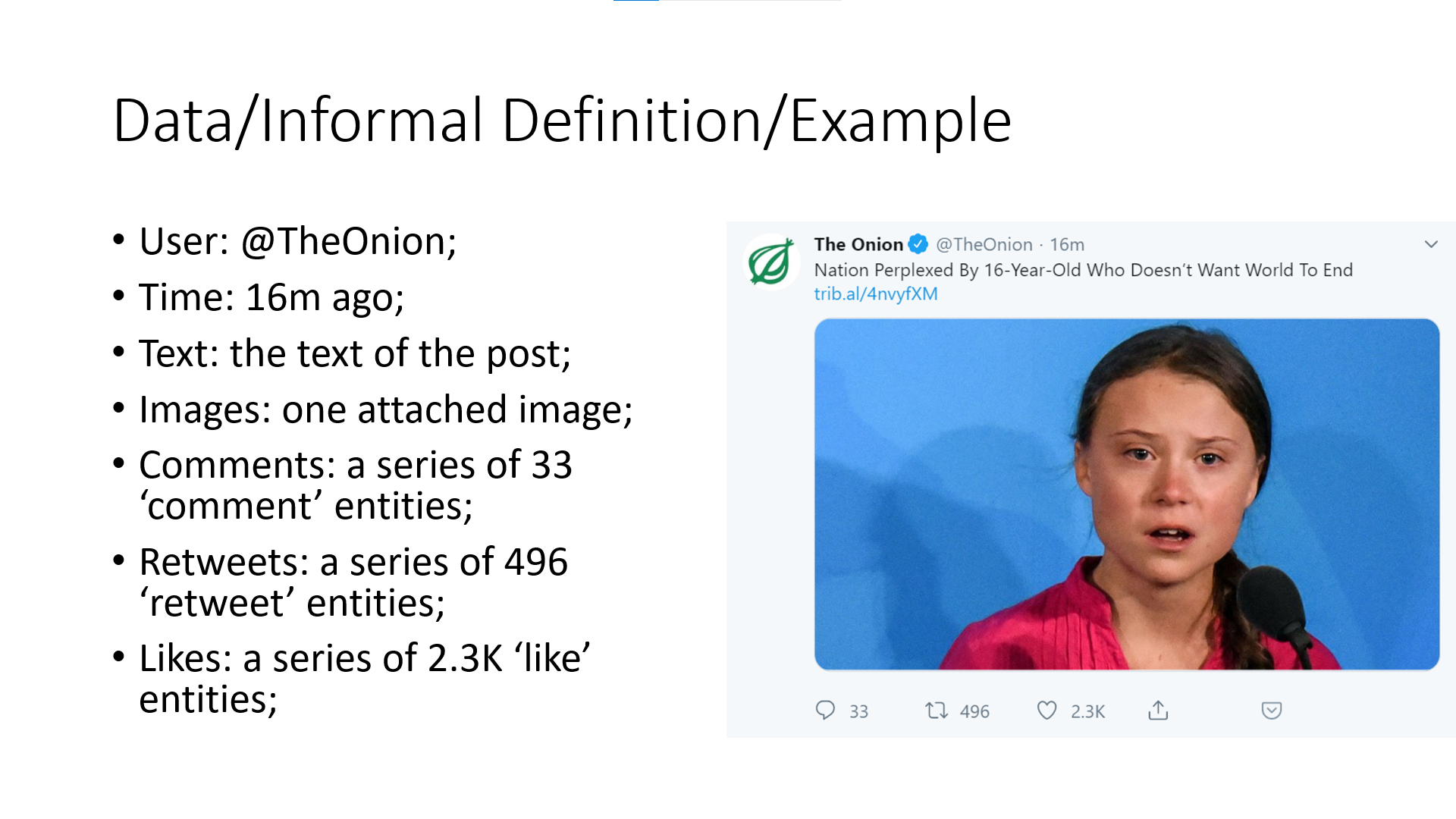
We have seen data social media is “all about the data” and about what we can do with this data. We will start by giving a definition of what we mean by ‘data’.

Following Wikipedia, we could say that:

‘Data is a set of values of subjects   
with respect to qualitative or quantitative variables’

Qualitative: properties that can be observed and cannot generally be measured with a numerical value (e.g., ‘color’, ‘sex’);

Quantitative: properties that can be quantified with a number (e.g., ‘height’, ‘weight’, ‘age’).



For example, a twitter post (a ‘tweet’) is data, i.e., a set of data with respect to different variables, such as user, time, text, images, comments, etc.

Some of these entities (e.g., comments) can be in turn a list of entities (e.g., a list of comments).

## Uncertainty & Importance of Probability Theory

We have seen that dealing with data means dealing with some variables assuming some values.

However, what can we say about what values will assume such variables?

In some cases, it is possible to predict the values of variables with perfect accuracy, given a set of initial conditions. For instance, think about a system of equations describing the speed of objects according to Newtonian laws.

In other cases, modeling the relationship between variables in a deterministic way is not possible.

We often say that this is due to uncertainty.

**Uncertainty in a system can be due to different factors:**

* Inherent stochasticity in the system being modeled. For example, most interpretations of quantum mechanics describe the dynamics of subatomic particles as being probabilistic. We can also create theoretical scenarios that we postulate to have random dynamics, such as a hypothetical card game where we assume that the cards are truly shuﬄed into a random order.
* Incomplete observability. Even deterministic systems can appear stochastic when we cannot observe all the variables that drive the behavior of the system. For example, in the Monty Hall problem, a game show contestant is asked to choose between three doors and wins a prize held behind the chosen door. Two doors lead to a goat while a third leads to a car. The outcome given the contestant’s choice is deterministic, but from the contestant’s point of view, the outcome is uncertain.
* Incomplete modeling. When we use a model that must discard some of the information we have observed, the discarded information results in uncertainty in the model’s predictions. For example, suppose we build a robot that can exactly observe the location of every object around it. If the robot discretizes space when predicting the future location of these objects, then the discretization makes the robot immediately become uncertain about the precise position of objects: each object could be anywhere within the discrete cell that it was observed to occupy.

**Examples**

* Tossing a coin or rolling a dice: these kind of experiments are generally impossible to model in a deterministic way. This can be due to our limited ability to model the event (i.e., rolling a dice might be deterministic, but deriving a set of equations to determine the outcome might be too hard).
* Determining if a patient has a given pathology: different pathologies might have similar symptoms. Hence, observing some of them does not allow to determine with perfect accuracy if the patient has that pathology. In this case, uncertainty might arise from incomplete observability.

**Importance of Probability Theory**

Probability theory provides a consistent framework to work with uncertain events.

It allows to quantify and manipulate uncertainty with a set of axioms, as well as to derive new uncertain statements.

Probability theory is hence an important tool to work with data. We will start to revise the main concepts behind probability theory by talking about random variables.

## Random Variable

We discovered that ‘data’ has something to do with ‘variables’. When dealing with uncertain events, we need to use the concept of ‘random variables’. Informally (from wikipedia):

‘A random variable is a variable whose values   
depend on outcomes of a random phenomenon.’

A random variable is characterized by a set of possible values often called ‘probability space’ or ‘alphabet’ (this last term comes from information theory, where we often deal with sources emitting symbols from an alphabet, in which case the values of will be the symbols).

Random variables can be discrete or continuous. Discrete random variables can assume a finite number (or a countable infinite number) of possible values, whereas a continuous random variable is generally associated with a real number.

A random variable is generally denoted by a capital letter, such as .

Random variables can be scalar (e.g., X=1) or multi-dimensional (e.g., , or ).

A random variable is always related to some uncertain phenomenon which generates observations. This is often referred also as ‘experiment’. For instance, if a random variable takes the values of tossing a coin, then ‘tossing a coin’ is the underlying experiment.

**Example – discrete scalar**

* For example, may denote the outcome of tossing a coin;
* In this context, ‘tossing a coin’ is the random phenomenon characterizing the random variable;
* The space of possible values that can assume (the alphabet of ) can be defined as . In this example, is discrete;
* If tossing the coin, the outcome is , then the variable assumes the value .

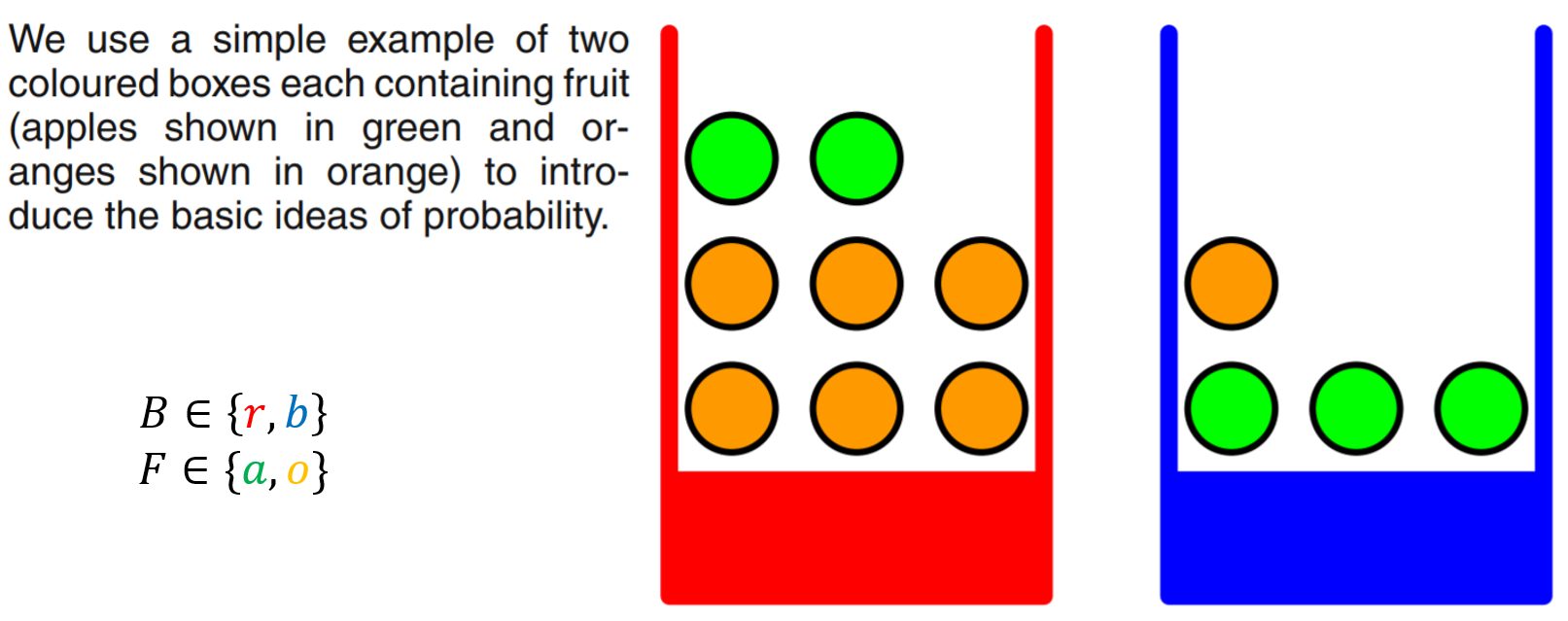
**Example – continuous scalar**

* may denote the height in meters of a student in this class;
* If we pick a student of heigh 1.75m, then ;

**Example – continuous multi-dimensional**

* may denote the latitude and longitude coordinates of a car in the world;
* Once we pick a car, we may have .

### Example



Let’s assume we have two colored boxes: one red and one blue.

Each box contains two kinds of fruits: apples and oranges.

We consider the experiment of randomly drawing some fruit from one of the two boxes. This happens in two stages:

* We first randomly pick one of the two boxes;
* Then we randomly pick one of the fruits in the box;
* After observing the type of the fruit, we replace it in the same box.

The outcome of the experiment can be characterized by two random variables:

* B represents the color of the box and can take values r (red) and b (blue).
* F represents the kind of fruit and can take values a (apple) and o (orange).
* If we pick an orange from the blue box, then the outcome of the experiment can be characterized by the values , ;
  1. As per our definition of data, the values , are ‘data’. Indeed, they represent a ‘bidimensional data point’ and describe the outcome of an experiment.
  2. An alternative view of having two random variables, is to have a single multi-dimensional variable .

## More Formal Definition of Data

We will define “data” as:

“The values assumed by a random variable”

**Example**

* For instance, if the outcome of tossing a coing is , then is data;
* It should be clear that the ‘data’ is the pair <random variable, value> and not just the value. Indeed alone would not be very useful (we don’t know which phenomenon it is related to), whereas can be useful, as we know that is the random variable describing the outcome of tossing a coin;
* In this example, the data is representing a fact: ‘I tossed a coin and the outcome was ’. This is also called ‘an event’;

**Example**

* We can characterize a tweet with the following random variabiles:
  1. : its textual content;
  2. : the attached images;
  3. : the date of publication;
* Once we pick a tweet, we may have:
  1. ;
  2. - a matrix representing an attached image;

# Probability

Since random variables are related to stochastic phenomena, we cannot say much about the outcome of a single phenomenon.

However, we expect to be able to characterize the class of experiments related to a random variable, to infer rules on what values the random variable is likely to take.

For instance, in the case of coin tossing, we can observe that, if I toss a coin for a large number of times, the number of heads will be roughly similar to the number of tails.

This kind of observations is useful, as it can give us a prior on what values we are likely to encounter and what are not.

To formally express such rules, we can define the concept of probability on a random variable.

Specifically, it is possible to assign a probability value to a given outcome. This is generally represented with a capital P:

* For instance, represents the probability of picking a blue box in the previous example;
* A probability is a number comprised between 0 and 1 which quantifies how likely we believe the event to be;
  + 0 means impossible;
  + 1 means certain;

When it is clear from the context which variable we are referring to, the probability can also be expressed simply as:

While it is intuitive to understand that a probability quantifies how likely that an event actually happens, we have not yet seen **how to assign probabilities to outcomes**. We will see that there are two main approaches to reason about probabilities: the frequentist approach and the Bayesian approach.

## Frequentist Approach to Probability

There are two main approaches to probability: frequentist and Bayesian.

**Frequentist approach**:

Probability theory was initially developed to analyze the frequency of events. For instance, it can be used to study events like drawing a certain hand of cards in a poker game. These events are repeatable and can be dealt with using frequencies. In this sense, when we say that an event has probability of occurring, it means that if we repeat the experiment infinitely many times, then a proportion of the repetitions would result in that outcome.

According to the frequentist approach, we can estimate probabilities by repeating an experiment for a large number of times and then computing:

* The number of trials: how many times we performed the experiment;
* The number of favorable outcomes: how many times the outcome of the experiment was favorable.

The probability is hence obtained by dividing the number of favorable outcomes by the number of trials.

For instance, let’s suppose we want to estimate the probability of obtaining a ‘head’ by tossing a coin. Let’s suppose we toss the coin 1000 times and obtain 499 heads and 501 tails. We can compute the probability of obtaining head as follows:

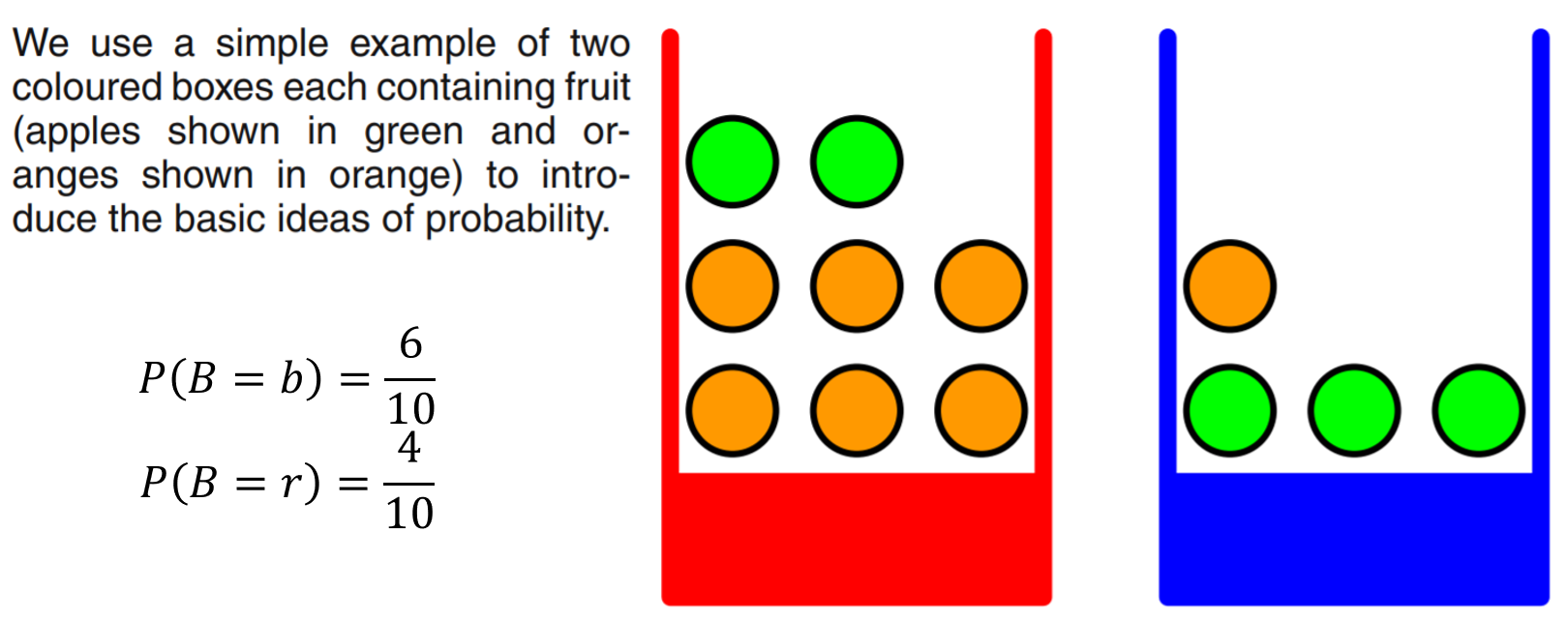
* Number of trials: 1000;
* Number of favorable outcomes: 499.

The probability of obtaining head will be 499/1000=0.49

**Examples:**

* The probability of obtaining ‘head’ when tossing a coin is 0.5. We know that because, if we toss a coin for a large number of times, about half of the times, we will obtain ‘head’;
* The probability of picking a red ball from a box with 40 red balls and 60 blue balls is 0.4. We know this because, if we repeat the experiment for a large number of times, we will observe that proportion.

### Example of Probability



Suppose we repeat this experiment for many times and observe that:

* We pick the red box 40% of the times;
* We pick the blue box 60% of the times;
* Once we selected a box, we are equally likely to select any of the fruits contained in it.

Using a frequentist approach, we can define the probabilities:

* This is done by using the formula:

## Joint probability

We can define **univariate** (= with respect to only one variable) probabilities and as we have seen in the previous examples.

However, in some cases, it is useful to define probabilities on more than one variable at the time. For instance, we could be interested in studying the probability of picking a given fruit from a given box. In this case, we would be interested in the **joint probability** .

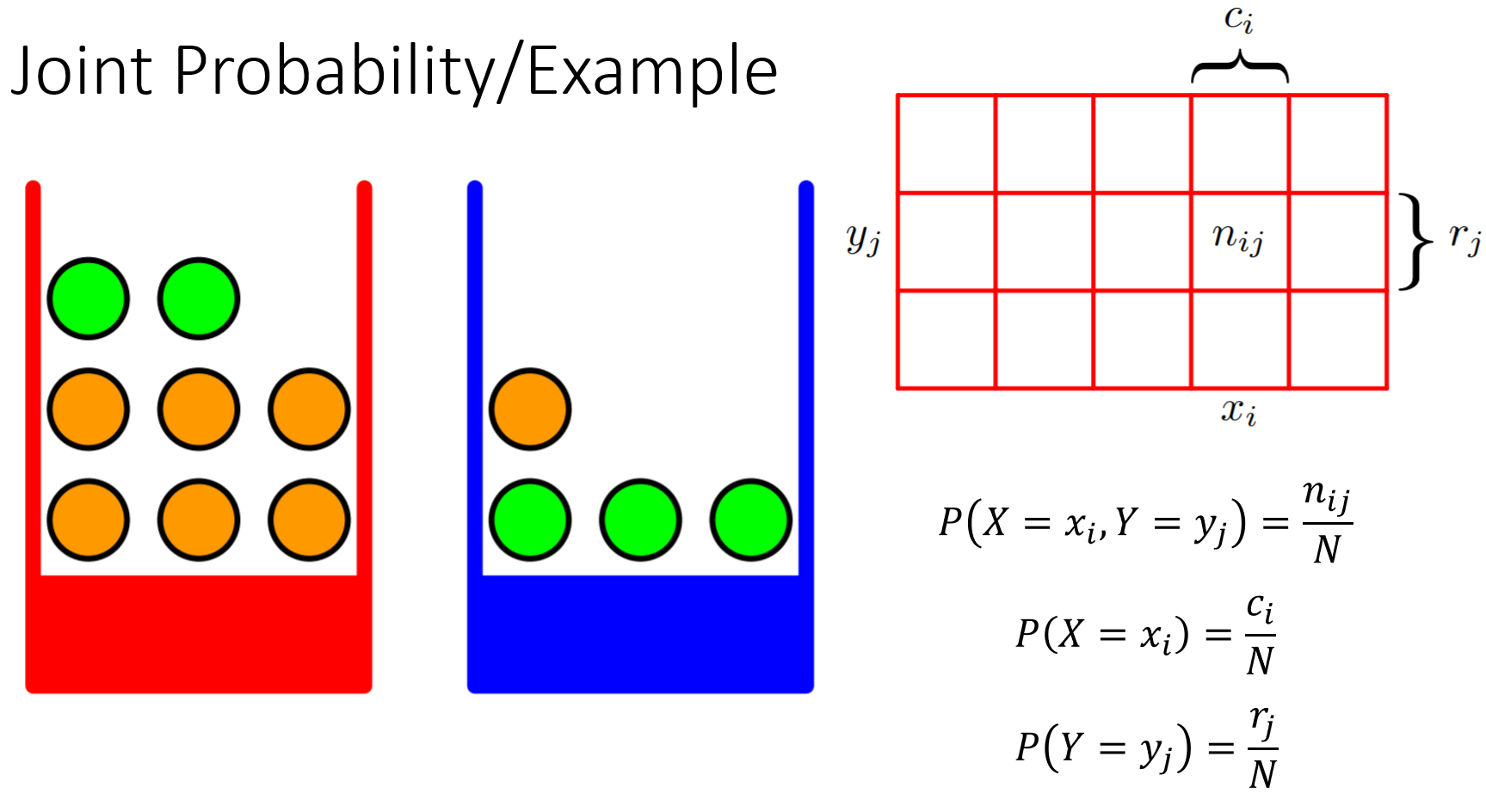
In general, we can have joint probabilities with arbitrary numbers of variable. For instance, .

Joint probabilities are symmetric, i.e., .

We should note that, when dealing with multiple unidimensional variables, we can always define a new multi-variate variable comprising all of them:

* .

### Example



We can see the concept of joint probability in the context of the examples of the two boxes;

We have seen how to define the univariate probability over the whole probability space of ;

However, we could be interested in the probability of both variables jointly: , i.e., the joint probability of B and F.

To ‘measure’ the joint probability, we could repeat the experiment for many times and observe the outcomes.

We could then build a ‘contingency table’ which keeps track of how many times we observed a given combination.

In the slide, we show an example for two general random variables and (they could be B and F):

* represents the number of times we observed and ;
  1. represents the number of times assumed the value , independently from the value assumed by ;
  2. represents the number of times assumed the value , independently from the value assumed by ;

We can measure the joint probability with a frequentist approach using the formula:

where is the total number of trials (i.e., the sum of all the values in the matrix).

Also, we note that we can define the probabilities of X and Y as follow:

## Sum Rule (Marginal Probability)

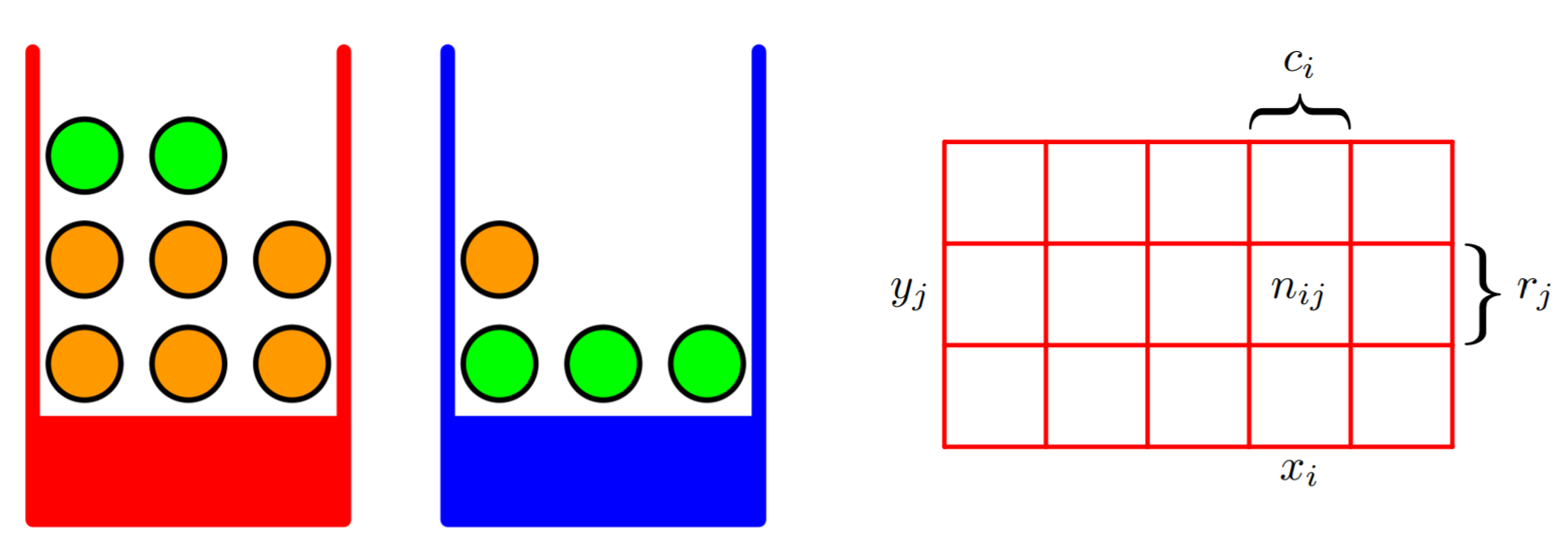
Sometimes we can compute probabilities on a set of variables, but we are interested on a probability on a subset of them.

In particular, we have seen that, given the contingency table related to the two variables X and Y, we can estimate both the joint probability and the probability of each of the variables: and .

To compute these probabilities, we can use the **sum rule of probability**:

The act of computing from is also known as marginalization and is also named **marginal probability**.

### Frequentist Derivation



We can see how the sum rule of probability can be derived from the simple example of the contingency table seen before;

Indeed, given the contingency table between variables X and Y, we have:

* : the number of trials in which is obtained by summing the numbers in the i-th colum;
* : the number of trials in which is obtained by summing th enumbers in the j-th row;

Therefore:

* ;
* ;

## Conditional Probability

In many cases, we are interested in the the probability of some event, given that some other event happened.

This is called **conditional** **probability** and is denoted as and read as ‘P of X=y given that Y=y’. In this context, is the condition, and we are interested in studying the probability of X only in the cases in which the condition is verified.

For instance, in the case of the two boxes, we could be interested in , that is to say, what is the probability of picking an apple, given that we know that we are drawing from the blue box?

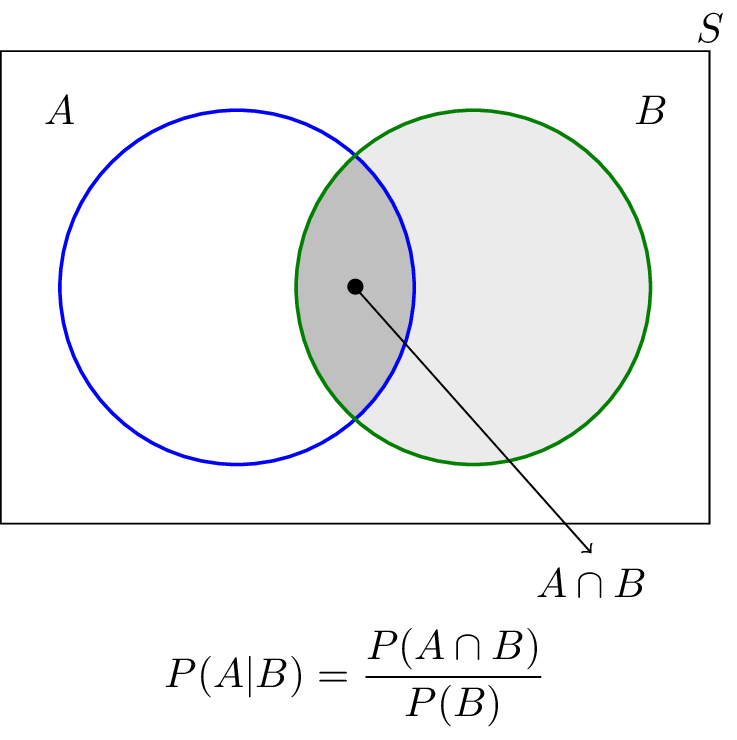
It should be noted that, in general .

We can define the conditional probability as follows:

The conditional probability is defined only when , that is, we cannot define a probability conditioned on an event that never happens.

In terms of sets, this can be seen as follows:

Where measures the probability that events A and B happen together, whereas represents the probability that B happens alone. The probability measures how likely it is that A happens, knowing that B has happened. We can see this graphically in terms of sets as follows:



## Product Rule (Factorization)

We can see the definition of conditional probability:

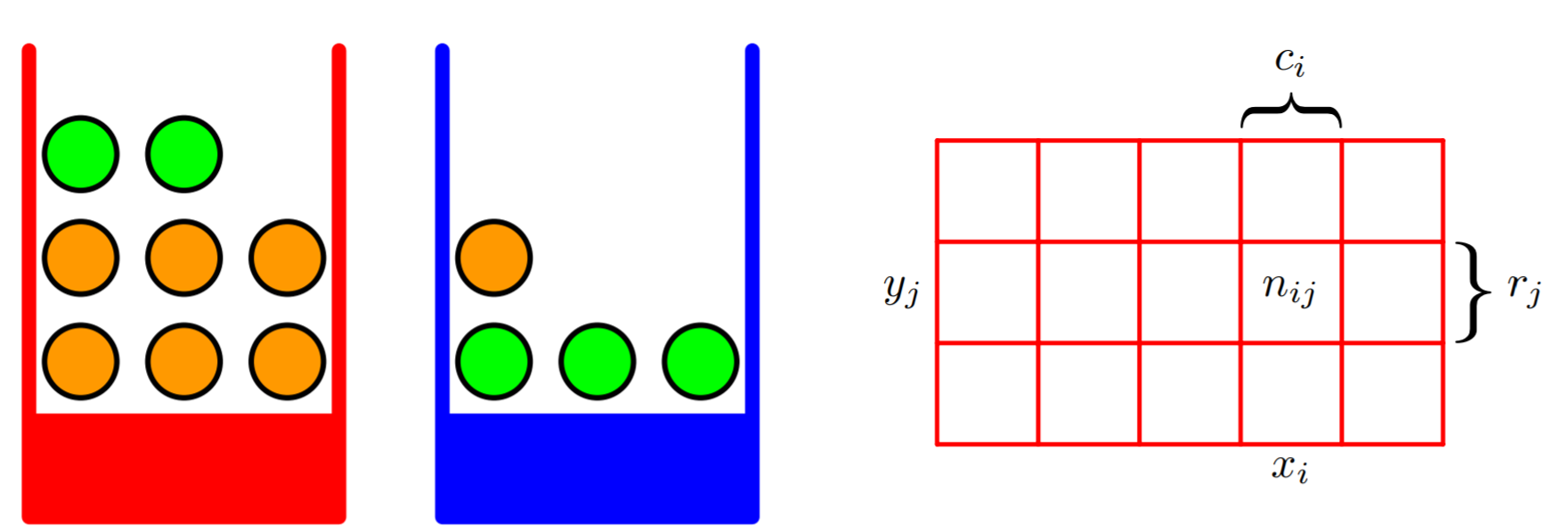
As follows:

which is often referred to as **the product rule**.

The product rule allows to compute joint probabilities starting from conditional probabilities and marginal probabilities. This is useful because measuring joint probabilities generally involves creating large tables, whereas conditional and marginal probabilities might be easier to derive.

This operation of expressing a joint probability in terms of two factors is known as factorization.

### Frequentist Derivation



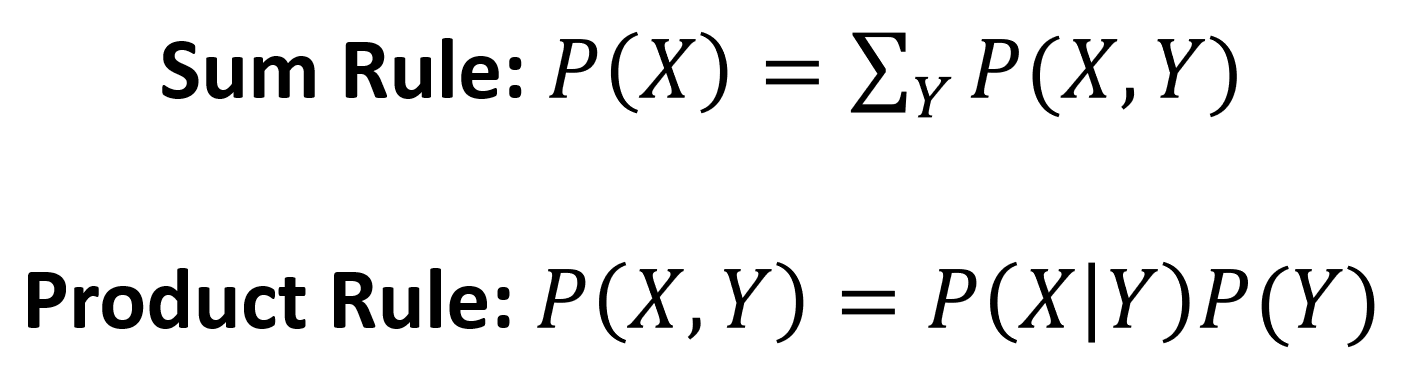
Also in this case, we can derive the product rule from our example;

In particular, the conditional probability can be computed as follows in a frequentist way:

Since we are restricting our probabilities to the cases in which , the total number of trials to be consider, is the number of trials in which , which is ;

We note that:

## The Fundamental Rules of Probability



To summarize, there are two fundamental rules of probability:

* Sum rule: allows to perform marginalization;
* Product rule: allows to factorize probability (i.e., breaking them into factors);

These two rules are the foundations of the framework of probability and allow to reason about random variables and uncertain statements.

All other rules and theorems (including Bayes’ rule) are derived from these two statements.

## The Chain Rule of Conditional Probabilities

When dealing with multiple variables, the product rule can be applied in an iterative fashion, thus obtaining the ‘chain rule’ of conditional probabilities.

For instance:

Since:

We obtain:

Since joint probabilities are symmetric, we could equally obtain:

This rule can be formalized as follows:

## Bayes’ Theorem

Given two variables A and B, from the product rule, we obtain:

Since joint probabilities are symmetric, we have:

Which implies:

;

This last expression is known as Bayes’ Theorem (or Bayes’ rule).

Using the sum rule, the denominator can be seen in terms of the quantities appearing in the numerator:

* We can see the denominator as a normalization constant required to make sure that is a valid probability.

The Bayes’ rule can be used to ‘invert’ conditional proabilities, i.e., to turn probabilities of the kind into probabilities of the kind ;

The different terms on the Bayes’ rule have names:

* is called ‘the prior’;
* is called ‘the likelihood’;
* is called ‘the evidence’;
* is called ‘the posterior’;
* We will get back to these names with examples in the following slides.

### Bayesian Approach to Probability

Some events are not related to repeatable experiments, but they are still characterized by uncertainty. Examples of such events are ‘will the sun extinguish in 4.5 billion years?’ or ‘given the observation of some symptoms, does the patient have a given pathology’. These events are not related to repeatable experiments. Yet, they are characterized by uncertainty. Also, observing more data might help reduce uncertainty. For instance, if a new symptom is observed, we might be less uncertain about one outcome or another.

**Examples:**

* The probability of a patient to have a flu given the observation of their temperature;
* The probability that the average temperature of the planet will increase by 1° next year if we reduce the emission of CO2 by a given amount

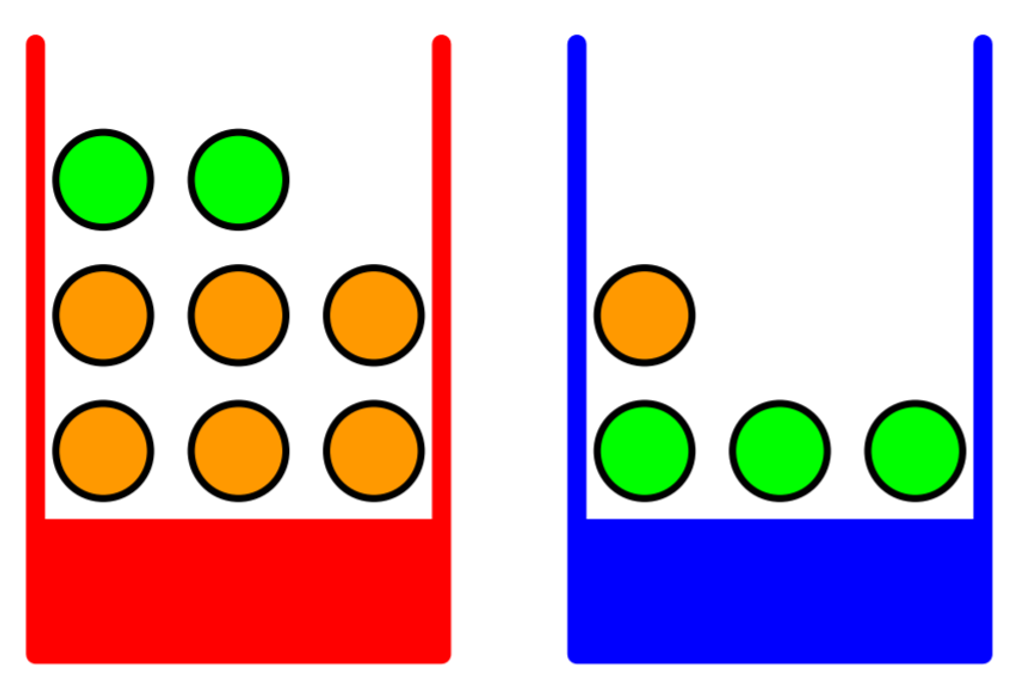
#### Degree of belief

An important concept related to Bayesian probability is the one of “degree of belief”. Since the experiments are not repeatable, to quantify probability we cannot count the number of favorable outcomes. Instead, we ‘inject’ our belief into the system. Specifically, this is done using the Bayes’ rule, which provide the tools to insert our degree of belief in a consistent way. We will talk about the Bayes’ rule later.

**Example:**

Let’s suppose we toss a coin three times and we obtain three heads. Now, three is not a large number and hence, if we try to apply the frequentist approach, we wrongly estimate that the probability of getting head is 1, while the probability of getting tail is zero. By using the Bayesian approach, we can insert our belief that the coin is fair by introducing a “prior probability” that the probability of getting head is 0.5.

## Example of Probability Manipulation



Let’s get back to our example of the two boxes and let’s suppose that, by repeating several trials, we discovered the following probabilities (in a frequentist way):

We also focused on a given box and performed different trials, observing the following proportions:

We shall note that these probabilities are normalized such that:

We can now use the rules we have seen before to answer questions such as:

1. What is the overall probability of choosing an apple?
2. What is the probability of picking a red box, given that we have drawn an orange?

**What is the overall probability of choosing an apple?**

To answer this question, we need to find . We note that, by the sum rule:

We also observe that the joint probabilities can be recovered using the product rule:

Hence, our probability can be found using the formula:

From the definition of probability, we have:

Hence:

**What is the probability of picking a red box, given that we have drawn an orange?**

To answer this question, we need to find the conditional probability . To find this probability, we need to ‘invert’ our conditional probabilities using the Bayes’ rule:

**Interpretation of Bayes’ rule:**

We can give the following interpretation of Bayes’ rule. If somebody asked us what is probability of picking the red box, we would have said it is . We call this ‘prior probability’ as it is our **best guess given that no additional constrain is provided**. Once we are told that we have picked an orange, our **belief** on the identity of the box can be refined. In particular, we know that the red box contains more oranges than the blue one, so we should assign probabilities accordingly. We use the Bayes’ theorem to obtain the conditional probability as done above. We call this the ‘posterior probability’, as this is our updated estimate, once we know the identity of the fruit. To refine our guess, we use the ‘likelihood’ which tells us how many oranges are in the red box and the ‘evidence’ which tells us how many oranges there are at all.

### Excercise

Suppose that we have three colored boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities P(r)=0.2, P(b)=0.2, P(g)=0.6, and a piece of fruit is extracted from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

## Independence and Conditional Independence

**Independence**

Two variables and are **said to be independent** if and only if:

It should be noted that the expression above **is generally not true** as we cannot always assume that two variables are independent.

Independence can be denoted as:

**Examples**

Intuitively, two variables are independent if the values of one of them do not affect the values of the other one:

* Weight and height of a person are **not independent**. Indeed, taller people are usually heavier;
* Height and richness are **independent**, as the richness does not depend on the height of a person;

**Conditional independence**

Two random variales and are said to be **conditionally independent** given a random variable if:

Conditional independence can be denoted as:

**Example**

* Height and vocabulary are not independent: taller people are usually older, and hence they have a more sophisticated vocabulary (they know more words). However, if we condition on age, they become independent. Indeed, among people of the same age, height should not influence vocabulary. Hence, height and vocabularyare **conditionally independent** with respect to age;

## References

* Parts of chapter 1 of [1];
* Most of chapter 3 of [2];
* Some parts of chapter 1 of [3].

[1] Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006. <https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf>

[2] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016. <https://www.deeplearningbook.org/>

[3] Cover, Thomas M., and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.