MODULE 1: **INTRODUCTION**

Introduction: What is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Data science is the combination of statistics, mathematics, programming, and problem-solving; capturing data in ingenious (clear) ways; the ability to look at things differently; and the activity of cleansing, preparing, and aligning data.

The Data Science Lifecycle

Lifecycle consists of five distinct stages, each with its own tasks:

1. **Capture:** This stage involves gathering raw structured and unstructured data.
2. **Maintain:** This stage covers taking the raw data and putting it in a form that can be used.
3. **Process:** Data Mining, Clustering/Classification, Data Modeling, Data Summarization.
4. **Analyze:** Exploratory/Confirmatory, Predictive Analysis, Regression, and Text Mining. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.
5. **Communicate:** In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

Applications of Data Science

1. Healthcare

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

2. Gaming

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

3. Image Recognition

Identifying patterns is one of the most commonly known applications of data science.

4.Logistics

Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

6. Fraud Detection

Fraud detection comes to the next in the list of applications of data science. Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

Big Data and Data Science hype – and getting past the hype

Note: Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.

Data science enables companies not only to understand data from multiple sources but also to enhance decision making. As a result, data science is widely used in almost every industry, **including health care, finance, marketing, banking, city planning, and more**. If you are probably means you have something useful to contribute to making data science into a more legitimate field that has the power to have a positive impact on society. So, what is eyebrow-raising (shows surprise) about Big Data and data science? Let's count the ways:

There's a lack of definitions around the most basic terminology. What is "Big Data" anyway? What does "data science" mean? What is the relationship between Big Data and data science? Is data science the science of Big Data? Is data science only the stuff going on in companies like Google and Facebook and tech companies?

Why do many people refer to Big Data as crossing disciplines such as finance, tech, etc. and to data science as only taking place in tech? Just how big is big? Or is it just a relative term? **These terms are so ambiguous**; they're more or less meaningless.

There's a distinct lack of respect for the researchers in academia and industry labs who have been working on this kind of stuff for years, and whose work is based on decades of work by statisticians, computer scientists, mathematicians, engineers, and scientists of all types.

The hype is crazy—The longer the hype goes on, the more many of us will get turned off by it, and the harder it will be to see what's good underneath it all, if anything.

Statisticians already feel that they are studying and working on the "Science of Data." That's their bread and butter Although we will make the case that data science is not just a rebranding of statistics or machine learning but rather a field unto itself, the media often describes data science in a way that makes it sound like as if it's simply statistics or machine learning in the context of the tech industry.

People have said to us, "Anything that has to call itself a 'science' is probably isn't." Although there might

be truth in there, that doesn't mean that the term "data science" itself represents nothing, but of course what it represents may not be science but more of a craft (Create documents, which will make an impact).

Why now? – Datafication, Current landscape of perspectives, Skill sets

Data Science helps businesses to comprehend vast amounts of data from different sources, extract useful insights, and make better data-driven choices. Data Science is used extensively in several industrial fields, such as marketing, healthcare, finance, banking, and policy work.

It's not only the massiveness that makes all this new data interesting (or poses challenges). It's that the data itself, often in real time, becomes the building blocks of data *products. On the Internet, this means* Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, and so on.

Datafication can be defined as a process that "aims to transform most aspects of a business into quantifiable data (data that can be counted or measured in numerical values) that can be tracked, monitored, and analyzed.

Datafication is a process of "taking all aspects of life and turning them into data."

Ex: LinkedIn datafies professional networks

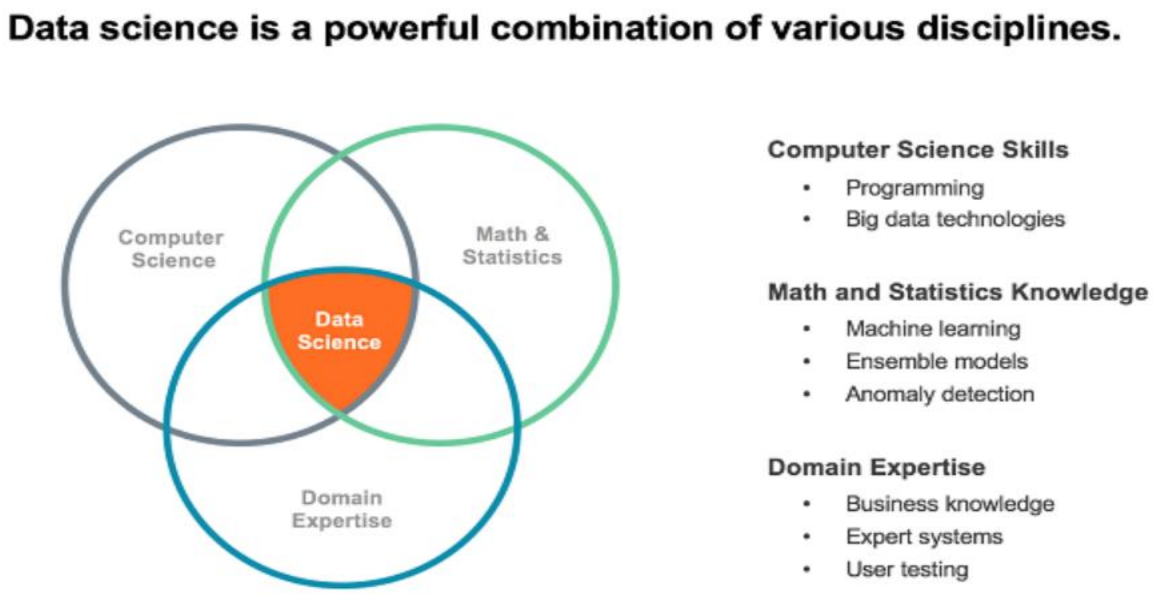
Datafication is an interesting concept and led us to consider its importance with respect to people's intentions about sharing their own data. We are being datafied , or rather our actions are, and when we "like" something online, we are intending to be datafied.

When we merely browse the Web, we are unintentionally, or at least passively, being datafied through cookies that we might or might not be aware of.

And when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors or cameras.

The Current Landscape

Data science is the process of extracting information, understanding and learning from raw data to inform decision making in a proactive and systematic fashion that can be generalized.



Data Science Jobs:

Data scientists need to be experts in computer science, statistics, communication, data visualization, and to have extensive domain ex-pertise**.**

Data Analyst bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses.

Data engineers focus on developing, deploying, managing, and optimizing the organization's data.

A data science profile need **skill levels** in the following domains:

- Computer science
- Math
- Statistics
- Machine learning
- Domain expertise
- Communication and presentation skills
- Data visualization

Needed Statistical Inference: Populations and samples

Population: A population is the entire group that you want to draw conclusion about.

Sample: Sample is the specific group that you will collect data from. Sample size is less than the size of population.

Generally, population refers to the people who live in a particular area at a specific time. But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc.

Figure: Population

If you had to collect the same data from a larger population, say the entire country of India, it would be impossible to draw reliable conclusions because of geographical and accessibility constraints, not to mention time and resource constraints. A lot of data would be missing or might be unreliable. Furthermore, due to accessibility issues, marginalized tribes or villages might not provide data at all, making the data biased towards certain regions or groups.

Samples are used when :

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size.

Statistical Inference is the process of using a sample to infer the properties of a population.

Consider N (Sample size)→used to represent the total number of observations in the population.

ALL → Population

For statistical inference N < ALL

Statistical Modeling:

Note: Data modeling is a process of creating a conceptual representation of data objects and their relationships to one another. The process of data modeling typically involves several steps, including requirements gathering, conceptual design, logical design, physical design, and implementation.

Before you get too involved with the data and start coding, it's useful to draw a picture of what you think the underlying process might be with your model. What comes first? What influences what? What causes what? What's a test of that?

But different people think in different ways. Some prefer to express these kinds of relationships in terms of math.

So, for example, if you have two columns of data, x and y, and you think there's a linear relationship, you'd write down **y=mx + b**

Other people prefer pictures and will first draw a diagram of data flow, possibly with arrows, showing how things affect other things or what happens over time.

Some techniques addressed under statistical modeling:

Regression analysis: Regression analysis is used to discover the connection between one or more independent variables and one or more dependent variables.

Time series analysis: Time series analysis is used to evaluate data that has been gathered over time. It is used to identify data trends, patterns, and seasonal fluctuations.

Cluster analysis: This technique is used to group comparable things.

Survival analysis: Survival analysis is used to assess time-to-event data, such as how long it takes for a patient to recover.

Decision trees: They are used to discover the most critical factors in a decision-making process.

Neural networks: Neural networks are used to simulate complicated interactions between variables. They are used in image recognition, natural language processing, among other things.

Probability Distribution

Note: Probability denotes the possibility of something happening. It is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 and 1. The definition of probability is the degree to which something is likely to occur. This fundamental theory of probability is also applied to probability distributions.

A **probability distribution** is a statistical function that describes all the possible values and probabilities for a random variable within a given range. This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution.

Types of Probability Distribution

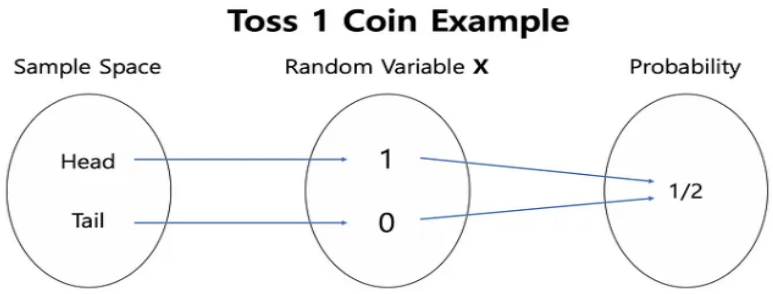
The probability distribution is divided into two parts:

1. Discrete Probability Distributions
2. Continuous Probability Distributions

A **discrete distribution** describes the probability of occurrence of each value of a discrete random variable. The number of spoiled apples out of 6 in your refrigerator can be an example of a discrete probability distribution.

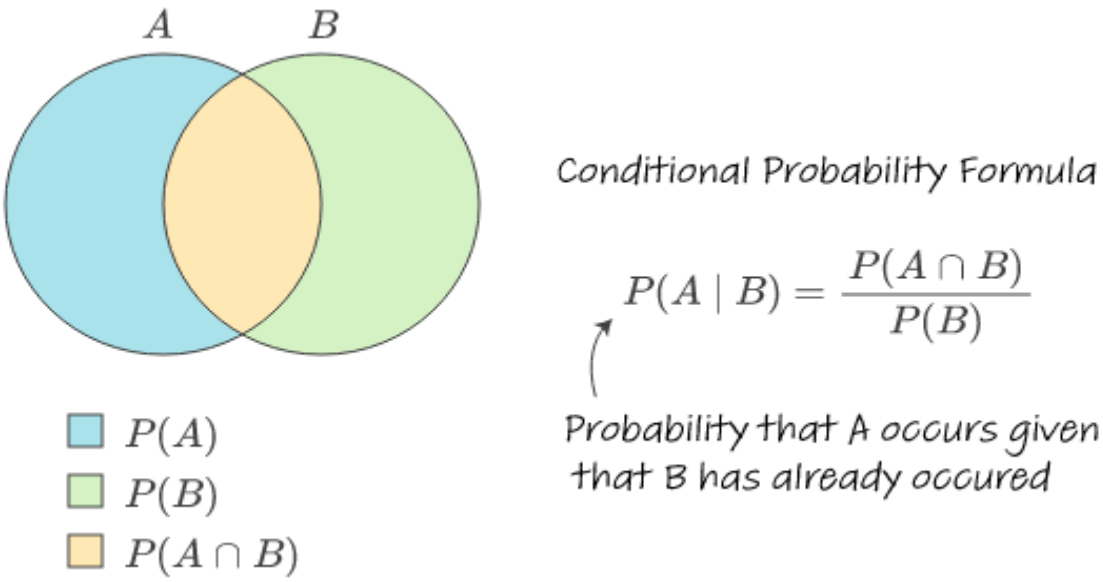
A **continuous distribution** describes the probabilities of a continuous random variable's possible values. A continuous random variable has an infinite and uncountable set of possible values.

Note: random variable is variable whose value is unknown or a function that assigns value to each of an experiment outcomes.



Conditional Probability

The probability of A given B is called the conditional probability and it is calculated using the formula P(A | B) = P(A ∩ B) / P(B) , when P(B) > 0.



Example:

Suppose we roll a balanced 6 sided die once. Consider the events A={1,2,3,4,5} and B={3,4,5,6}. What is the conditional probability of A, given B?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

P(A∩B)=3/6

P(B)=4/6

P(A|B) = 3/4

Joint Probability:

Joint probability is the product of the individual probabilities of independent events. Mathematically, P (A and B) = P(A) x P(B). The probability of A times the probability of B equals the joint probability of A and B happening at the same time.

MODULE 1: **INTRODUCTION**

Introduction: What is Data Science?

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Data science is the combination of statistics, mathematics, programming, and problem-solving; capturing data in ingenious (clear) ways; the ability to look at things differently; and the activity of cleansing, preparing, and aligning data.

The Data Science Lifecycle

Lifecycle consists of five distinct stages, each with its own tasks:

1. **Capture:** This stage involves gathering raw structured and unstructured data.
2. **Maintain:** This stage covers taking the raw data and putting it in a form that can be used.
3. **Process:** Data Mining, Clustering/Classification, Data Modeling, Data Summarization.
4. **Analyze:** Exploratory/Confirmatory, Predictive Analysis, Regression, and Text Mining. Here is the real meat of the lifecycle. This stage involves performing the various analyses on the data.
5. **Communicate:** In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.

Applications of Data Science

1. Healthcare

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

2. Gaming

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

3. Image Recognition

Identifying patterns is one of the most commonly known applications of data science.

4.Logistics

Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

6. Fraud Detection

Fraud detection comes to the next in the list of applications of data science. Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

Big Data and Data Science hype – and getting past the hype

Note: Big data refers to significant volumes of data that cannot be processed effectively with the traditional applications that are currently used. The processing of big data begins with raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.

Data science enables companies not only to understand data from multiple sources but also to enhance decision making. As a result, data science is widely used in almost every industry, **including health care, finance, marketing, banking, city planning, and more**. If you are probably means you have something useful to contribute to making data science into a more legitimate field that has the power to have a positive impact on society. So, what is eyebrow-raising (shows surprise) about Big Data and data science? Let's count the ways:

There's a lack of definitions around the most basic terminology. What is "Big Data" anyway? What does "data science" mean? What is the relationship between Big Data and data science? Is data science the science of Big Data? Is data science only the stuff going on in companies like Google and Facebook and tech companies?

Why do many people refer to Big Data as crossing disciplines such as finance, tech, etc. and to data science as only taking place in tech? Just how big is big? Or is it just a relative term? **These terms are so ambiguous**; they're more or less meaningless.

There's a distinct lack of respect for the researchers in academia and industry labs who have been working on this kind of stuff for years, and whose work is based on decades of work by statisticians, computer scientists, mathematicians, engineers, and scientists of all types.

The hype is crazy—The longer the hype goes on, the more many of us will get turned off by it, and the harder it will be to see what's good underneath it all, if anything.

Statisticians already feel that they are studying and working on the "Science of Data." That's their bread and butter Although we will make the case that data science is not just a rebranding of statistics or machine learning but rather a field unto itself, the media often describes data science in a way that makes it sound like as if it's simply statistics or machine learning in the context of the tech industry.

People have said to us, "Anything that has to call itself a 'science' is probably isn't." Although there might

be truth in there, that doesn't mean that the term "data science" itself represents nothing, but of course what it represents may not be science but more of a craft (Create documents, which will make an impact).

Why now? – Datafication, Current landscape of perspectives, Skill sets

Data Science helps businesses to comprehend vast amounts of data from different sources, extract useful insights, and make better data-driven choices. Data Science is used extensively in several industrial fields, such as marketing, healthcare, finance, banking, and policy work.

It's not only the massiveness that makes all this new data interesting (or poses challenges). It's that the data itself, often in real time, becomes the building blocks of data *products. On the Internet, this means* Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations, and so on.

Datafication can be defined as a process that "aims to transform most aspects of a business into quantifiable data (data that can be counted or measured in numerical values) that can be tracked, monitored, and analyzed.

Datafication is a process of "taking all aspects of life and turning them into data."

Ex: LinkedIn datafies professional networks

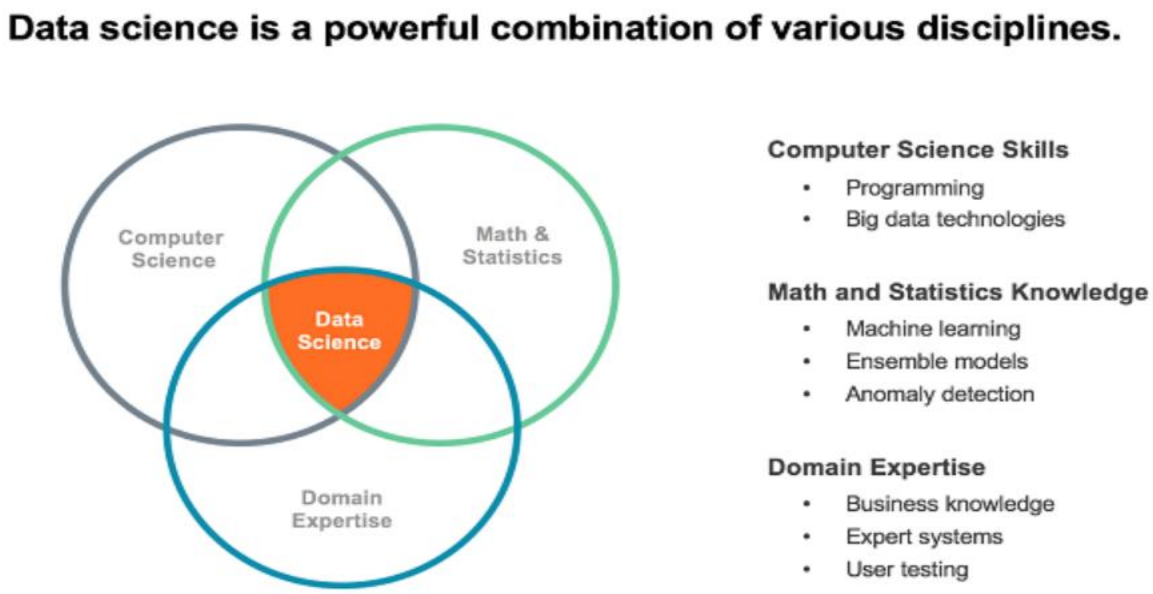
Datafication is an interesting concept and led us to consider its importance with respect to people's intentions about sharing their own data. We are being datafied , or rather our actions are, and when we "like" something online, we are intending to be datafied.

When we merely browse the Web, we are unintentionally, or at least passively, being datafied through cookies that we might or might not be aware of.

And when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors or cameras.

The Current Landscape

Data science is the process of extracting information, understanding and learning from raw data to inform decision making in a proactive and systematic fashion that can be generalized.



Data Science Jobs:

Data scientists need to be experts in computer science, statistics, communication, data visualization, and to have extensive domain ex-pertise**.**

Data Analyst bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses.

Data engineers focus on developing, deploying, managing, and optimizing the organization's data.

A data science profile need **skill levels** in the following domains:

- Computer science
- Math
- Statistics
- Machine learning
- Domain expertise
- Communication and presentation skills
- Data visualization

Needed Statistical Inference: Populations and samples

Population: A population is the entire group that you want to draw conclusion about.

Sample: Sample is the specific group that you will collect data from. Sample size is less than the size of population.

Generally, population refers to the people who live in a particular area at a specific time. But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc.

Figure: Population

If you had to collect the same data from a larger population, say the entire country of India, it would be impossible to draw reliable conclusions because of geographical and accessibility constraints, not to mention time and resource constraints. A lot of data would be missing or might be unreliable. Furthermore, due to accessibility issues, marginalized tribes or villages might not provide data at all, making the data biased towards certain regions or groups.

Samples are used when :

- The population is too large to collect data.
- The data collected is not reliable.
- The population is hypothetical and is unlimited in size.

Statistical Inference is the process of using a sample to infer the properties of a population.

Consider N (Sample size)→used to represent the total number of observations in the population.

ALL → Population

For statistical inference N < ALL

Statistical Modeling:

Note: Data modeling is a process of creating a conceptual representation of data objects and their relationships to one another. The process of data modeling typically involves several steps, including requirements gathering, conceptual design, logical design, physical design, and implementation.

Before you get too involved with the data and start coding, it's useful to draw a picture of what you think the underlying process might be with your model. What comes first? What influences what? What causes what? What's a test of that?

But different people think in different ways. Some prefer to express these kinds of relationships in terms of math.

So, for example, if you have two columns of data, x and y, and you think there's a linear relationship, you'd write down **y=mx + b**

Other people prefer pictures and will first draw a diagram of data flow, possibly with arrows, showing how things affect other things or what happens over time.

Some techniques addressed under statistical modeling:

Regression analysis: Regression analysis is used to discover the connection between one or more independent variables and one or more dependent variables.

Time series analysis: Time series analysis is used to evaluate data that has been gathered over time. It is used to identify data trends, patterns, and seasonal fluctuations.

Cluster analysis: This technique is used to group comparable things.

Survival analysis: Survival analysis is used to assess time-to-event data, such as how long it takes for a patient to recover.

Decision trees: They are used to discover the most critical factors in a decision-making process.

Neural networks: Neural networks are used to simulate complicated interactions between variables. They are used in image recognition, natural language processing, among other things.

Probability Distribution

Note: Probability denotes the possibility of something happening. It is a mathematical concept that predicts how likely events are to occur. The probability values are expressed between 0 and 1. The definition of probability is the degree to which something is likely to occur. This fundamental theory of probability is also applied to probability distributions.

A **probability distribution** is a statistical function that describes all the possible values and probabilities for a random variable within a given range. This range will be bound by the minimum and maximum possible values, but where the possible value would be plotted on the probability distribution.

Types of Probability Distribution

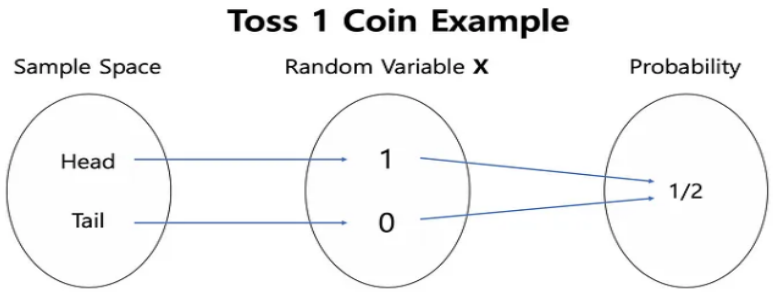
The probability distribution is divided into two parts:

1. Discrete Probability Distributions
2. Continuous Probability Distributions

A **discrete distribution** describes the probability of occurrence of each value of a discrete random variable. The number of spoiled apples out of 6 in your refrigerator can be an example of a discrete probability distribution.

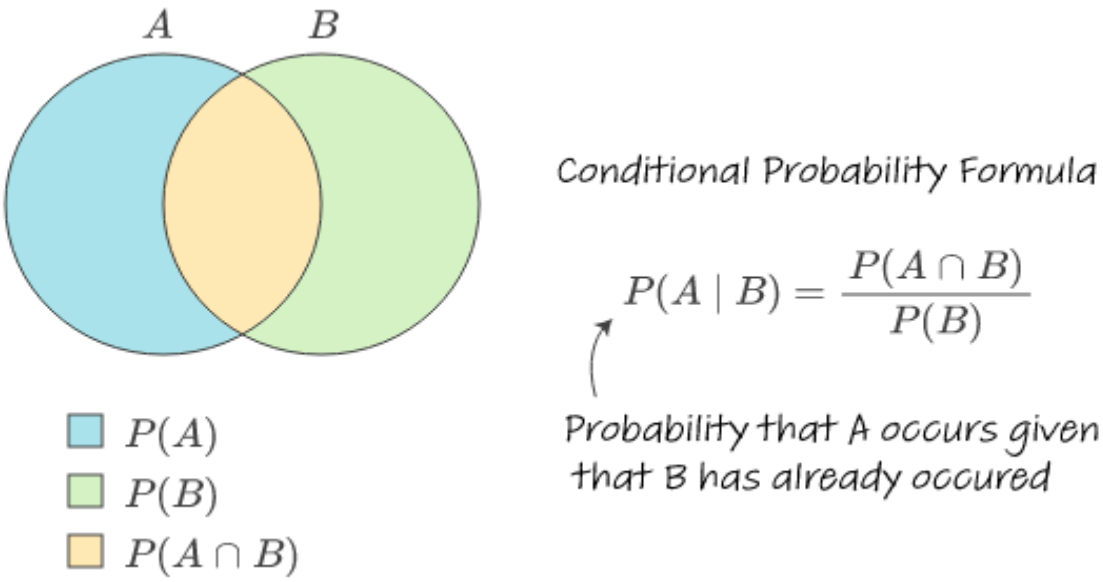
A **continuous distribution** describes the probabilities of a continuous random variable's possible values. A continuous random variable has an infinite and uncountable set of possible values.

Note: random variable is variable whose value is unknown or a function that assigns value to each of an experiment outcomes.



Conditional Probability

The probability of A given B is called the conditional probability and it is calculated using the formula P(A | B) = P(A ∩ B) / P(B) , when P(B) > 0.



Example:

Suppose we roll a balanced 6 sided die once. Consider the events A={1,2,3,4,5} and B={3,4,5,6}. What is the conditional probability of A, given B?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

P(A∩B)=3/6

P(B)=4/6

P(A|B) = 3/4

Joint Probability:

Joint probability is the product of the individual probabilities of independent events. Mathematically, P (A and B) = P(A) x P(B). The probability of A times the probability of B equals the joint probability of A and B happening at the same time.