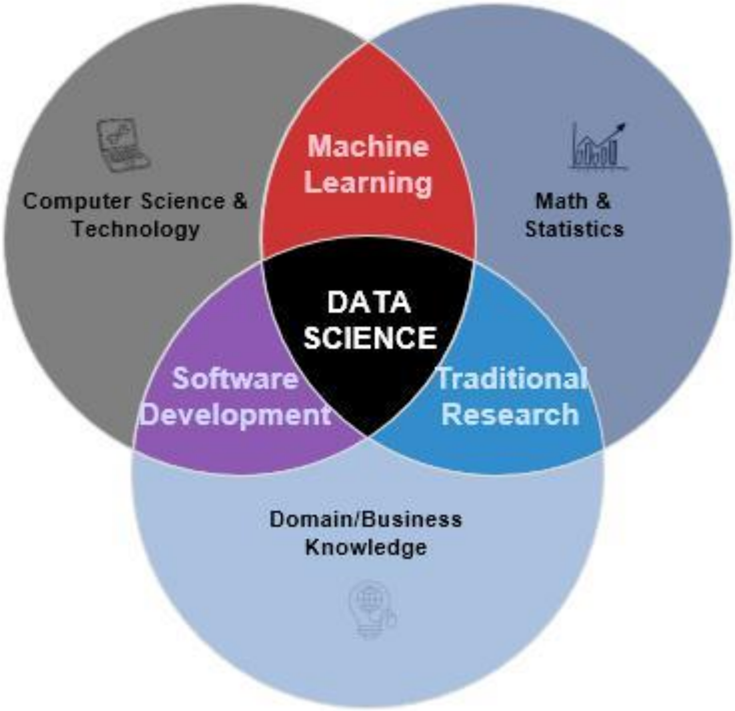
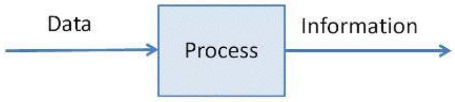
Introduction to Data Science and Probability

What is Data Science?

- Data Science is the science of extracting hidden patterns from large data sets
- Hidden patterns can appear in form of trends, cycles, associations, rules, groups etc. in the data
- Data sets usually refer to large volume of cleansed, structured data prepared for the analysis
- Science refers to the statistical tools and techniques employed to understand the data and reliability of the identified patterns.
    - That part of statistics which is used to understand the data is called descriptive statistics. Descriptive statistics give vital insights into the data in terms of central values, spread and distribution shape of the data
    - The part of statistics which is used to establish the reliability of the potential patterns identified, is called inferential statistics
- This can easily be represented pictorially by following *venn* diagram



Data Vs Information:



What is Data?

Data is footprint of happenings – natural or human driven. Traditionally footprints used to be physical – in form of records in files or registers or receipts. Today majority of these footprints are digital – your browsing behavior, purchasing behavior etc. Data is the raw material that has information contained within but it has a lot of noise in it as well. When processed correctly, it can give us meaningful information.

- Ticket sales on a band on tour
- Survey data: Different companies collect data by survey to know the opinion of people about their product.

What is Information?

Information is processed data. Information is basically the data plus the meaning of what the data was collected for minus the noise that got collected unintentionally.

Example:

- Sales report by region and venue - tells us which venue is most profitable.
- Survey Reports and Results: Survey data is summarized into reports/information to present to management of the company

Key Differences:

1. Data is the input and information is the output
2. Data is unprocessed records but information is processed data which has been made sense of

The Ascendance of data

We live in a world that's drowning in data. Websites track every user's every click. Your smartphone is building up a record of your location and speed every second of every day.

“Quantified selfers” wear pedometers-on-steroids that are ever recording their heart rates, movement habits, diet, and sleep patterns. Smart cars collect driving habits, smart homes collect living habits, and smart marketers collect purchasing habits. The internet itself represents a huge graph of knowledge that contains (among other things) an enormous cross-referenced encyclopedia; domain-specific databases about movies, music, sports results, pinball machines, memes etc.

Probability

Probability is simply how likely something is to happen. Whenever we’re unsure about the outcome of an event, we can talk about the probabilities of certain outcomes—how likely they are. The analysis of events governed by probability is called statistics.

The best example for understanding probability is flipping a coin:

There are two possible outcomes—heads or tails.
What’s the probability of the coin landing on Heads?

Let’s call probability of coin landing on Head is P(H). You might intuitively know that the likelihood is half/half, or 50%. But how do we work that out?

PROBABILITY OF AN EVENT = (# OF WAYS IT CAN HAPPEN) / (TOTAL NUMBER OF OUTCOMES)

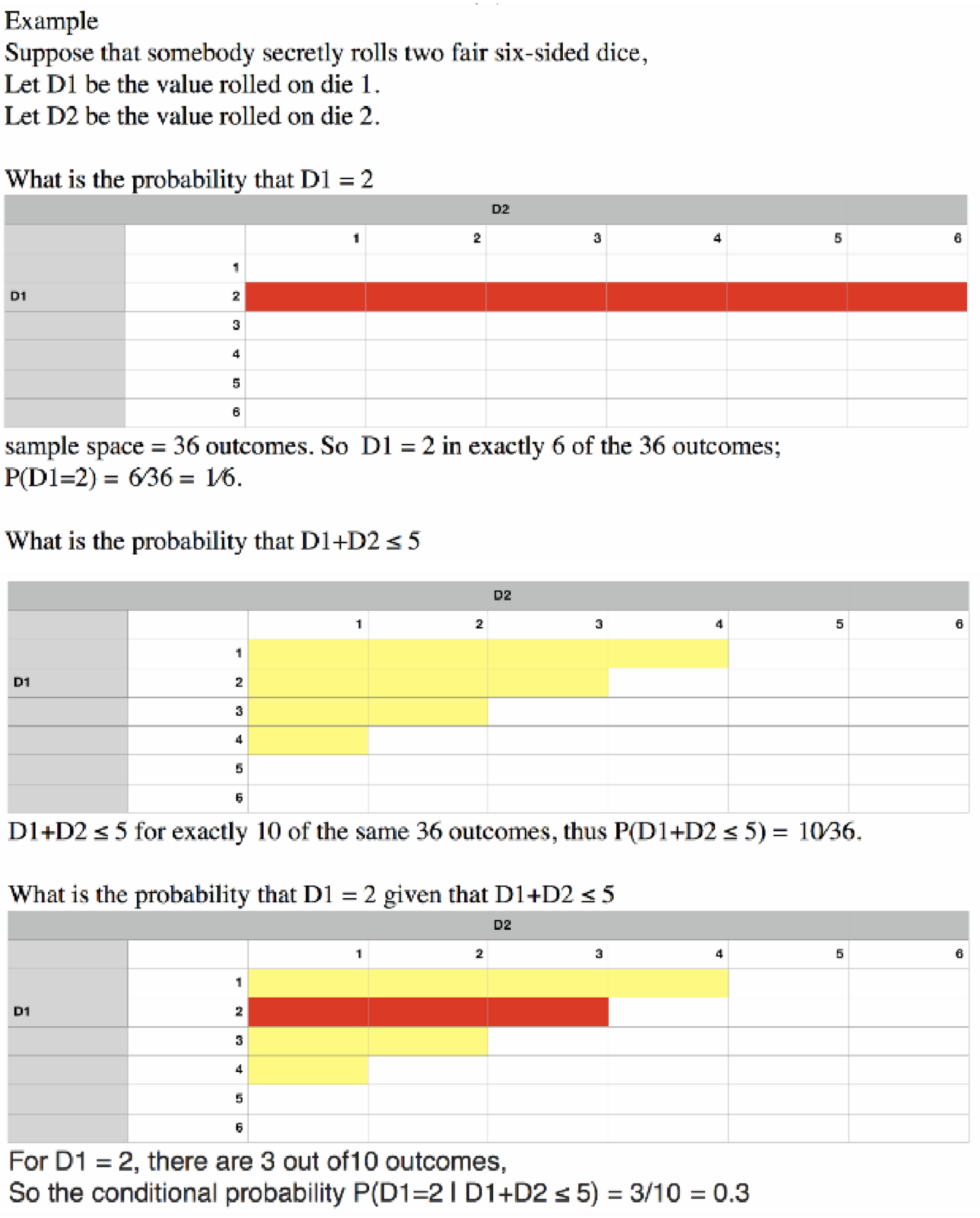
P(A) = (# OF WAYS A CAN HAPPEN) / (TOTAL NUMBER OF OUTCOMES)

So, in this case, # ways coin can land on Head = 1 and total number of possible outcomes = 2 (Head or Tails) and so P(H) = ½ = 50%

- The probability of an event can only be between 0 and 1 and can also be written as a percentage.
- The probability of event A is often written as P(A).
- If P(A)>P(B), then event A has a higher chance of occurring than event B.
- If P(A) = P(B), then events A and B are equally likely to occur.

Conditional Probability

Conditional Probability is a measure of the probability of an event given that (by assumption, presumption, assertion or evidence) another event has already occurred. If the event of interest is A and the event B is known or assumed to have occurred, “the conditional probability of A given B”, is usually written as P(A|B).



Example
Suppose that somebody secretly rolls two fair six-sided dice,
Let D1 be the value rolled on die 1.
Let D2 be the value rolled on die 2.

What is the probability that D1 = 2

		D2					
		1	2	3	4	5	6
	1						
D1	2	🟥	🟥	🟥	🟥	🟥	🟥
	3						
	4						
	5						
	6						

sample space = 36 outcomes. So D1 = 2 in exactly 6 of the 36 outcomes;
P(D1=2) = 6⁄36 = 1⁄6.

What is the probability that D1+D2 ≤ 5

		D2					
		1	2	3	4	5	6
	1	🟨	🟨	🟨	🟨		
D1	2	🟨	🟨	🟨			
	3	🟨	🟨				
	4	🟨					
	5						
	6						

D1+D2 ≤ 5 for exactly 10 of the same 36 outcomes, thus P(D1+D2 ≤ 5) = 10⁄36.

What is the probability that D1 = 2 given that D1+D2 ≤ 5

		D2					
		1	2	3	4	5	6
	1	🟨	🟨	🟨	🟨		
D1	2	🟥	🟥	🟥			
	3	🟨	🟨				
	4	🟨					
	5						
	6						

For D1 = 2, there are 3 out of10 outcomes,
So the conditional probability P(D1=2 I D1+D2 ≤ 5) = 3/10 = 0.3

Independence

Two events are said to be independent of each other, if the probability that one event occurs in no way affects the probability of the other event occurring, or in other words if we have observation about one event it doesn’t affect the probability of the other. For Independent Events A and B below is true

$$P(A,B) = P(A) * P(B) \quad where \quad P(A) \neq 0 \quad and \quad P(B) \neq 0$$

$$P(A \mid B) = P(A) \quad \text{and} \quad P(B \mid A) = P(A)$$

Example

Let’s say you rolled a die and flipped a coin. The probability of getting any number face on the dice is no way influences the probability of getting a head or a tail on the coin.

Understanding of probability is fundamental to making meaningful inferences and predictions using data science. All different analytical techniques eventually get down to a probability – what is the probability of this ‘observation’ having happened by chance OR what is the probability that there is actually a relationship between variable A and B OR what is the probability that certain event will happen?