Introduction to the World of Data

Origin of Decisions

Decisions were always **data-driven**

Let's consider few situations that early civilizations might have faced

| When to stock up on firewood? |

Growth & Sustenance
of the community

| **Predict the weather** |

? inputs

- Geography
- Pressure
- Rainfall
- Temperature
- Humidity

| What is the best time to travel? |

| When to start cultivating a crop? |
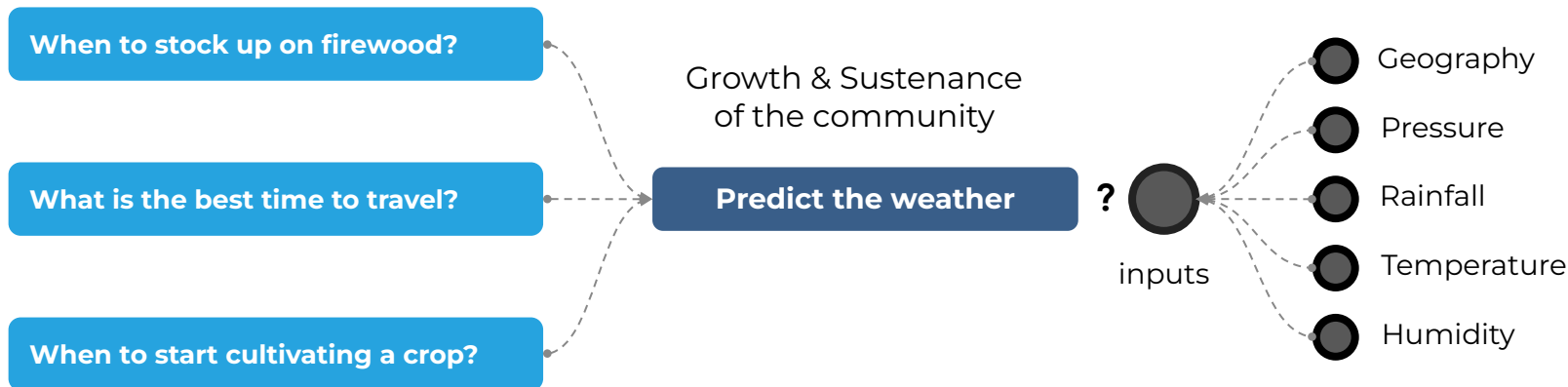
Decisions are made today by businesses the same way - but the methods have become more **accurate** and **faster** owing to the evolution of **statistical techniques** & **computing capabilities**

Paradigms in data science

Inferential	Computational	
Make predictions on population based on sample data	Leverage computational methods and technology to scale insight generation	Focus
Use statistical methods to draw conclusions / infer from data	Implement algorithms and computational methods to analyse data	Methods
Representativeness of data	Complexity of algorithms and cost of training large models	Limitations
1. Effectiveness of a new medication through randomized trial 2. Impact of a new policy on citizens	1. Weather forecasting based on historical and weather patterns 2. Optimize routing of vehicles to minimize costs	Examples

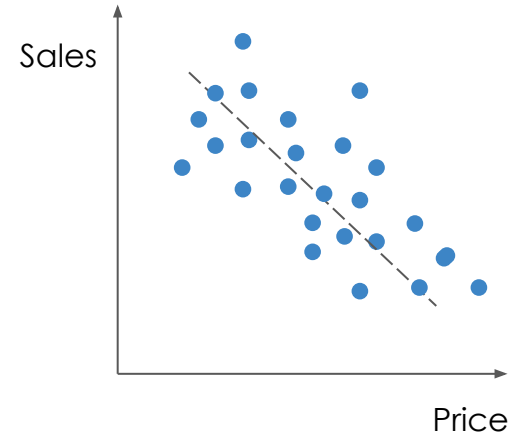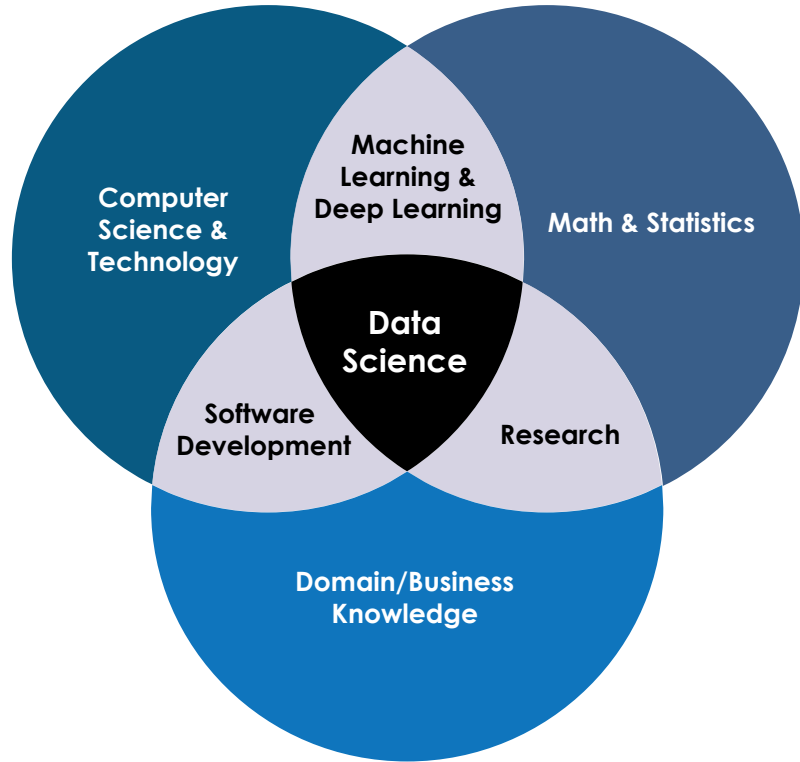# The need for math in Data Science

$$Sales = 500 - 20 * Price$$

Probability & Descriptive Statistics

Bayes' Rule

"Determine probability of a hypothesis based on prior knowledge and new evidence."

Eg: Shopping patterns of 100 people in a retail store
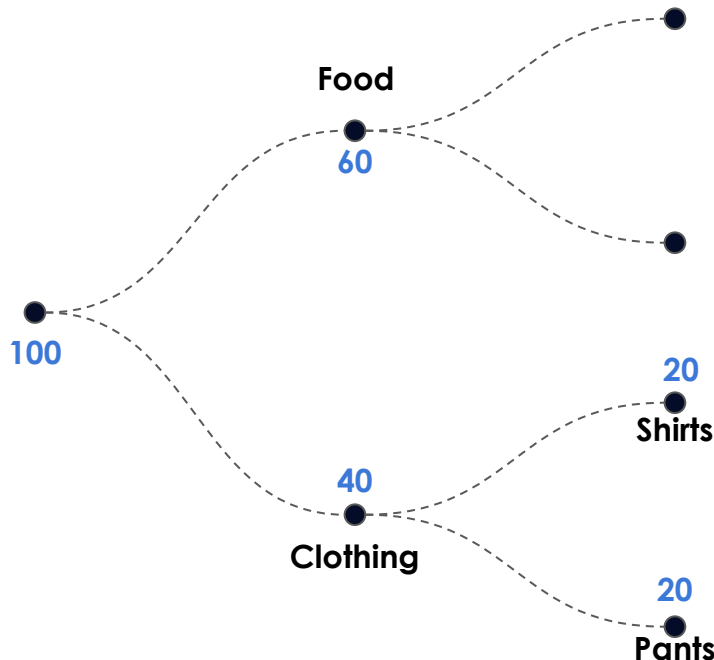
Probability of a new person to enter the food section?

Probability (Food) = 0.6

Probability of a person to enter the clothing section?

Probability (Clothing) = 0.4

Probability of a person in the shirt section, given, he is in the clothing section?

Probability (Shirt | Clothing) = 0.5

Food

60

100

20

Shirts

40

Clothing

20

Pants

Descriptive Statistics

"Describing features of a dataset by generating summaries about data samples "

Istanbul Stock Exchange National 100 Index
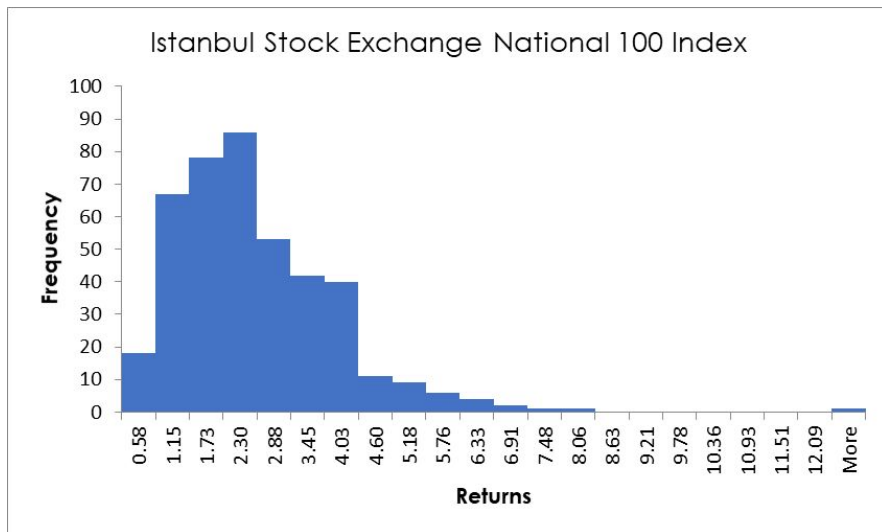
Average Return = Mean

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Risk = Standard Deviation

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

Descriptive Statistics

"Describing features of a dataset by generating summaries about data samples "

Happy Learning !