MODULE 1

INTRODUCTION TO DATA SCIENCE

Data Science

> What is Data Science?
> Why is the demand for data scientists growing?
> What data products are being created?
> What skills does a data scientist need?

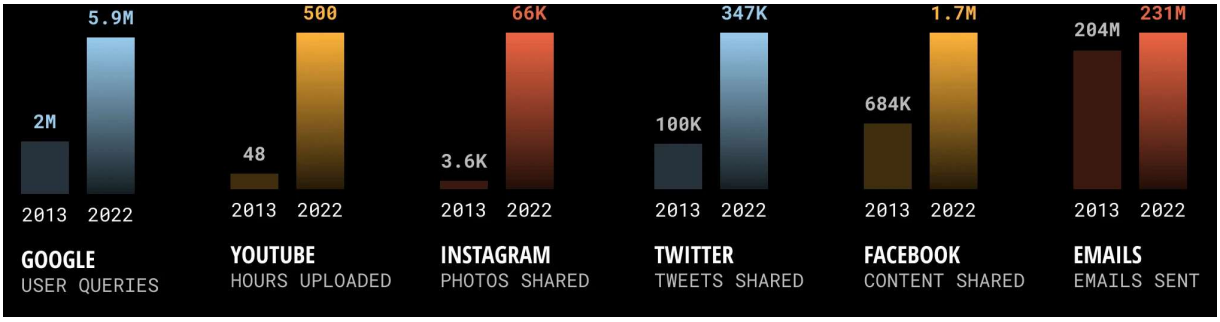
The Need for Data Science

- Huge amounts of data being generated
- Data generation much faster
  - Automation
  - Widespread Internet connectivity
  - User generated content
- Sample data
  - Financial transactions
  - Sensor networks
  - Application logs
  - Email and messaging
  - Social Media

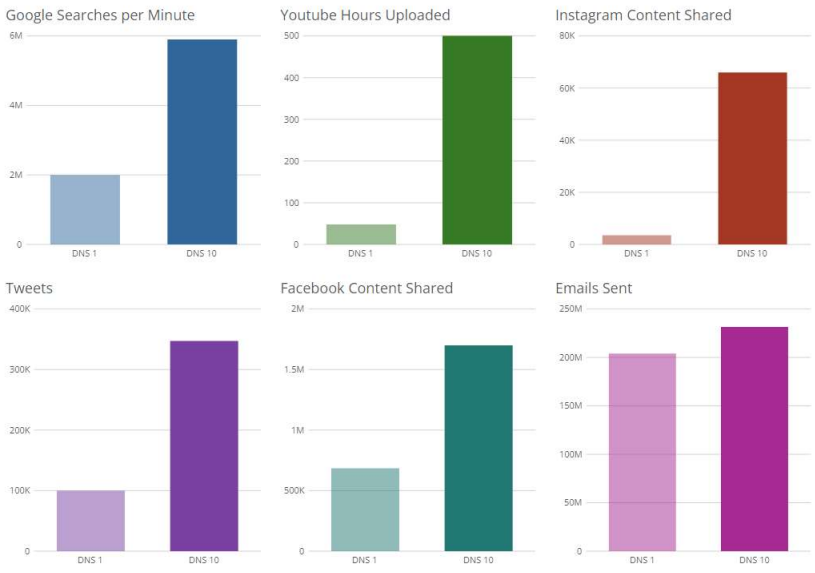
Data Never Sleeps

- This concept highlights the vast amounts of data
  - produced by individuals, businesses, and devices
  - every moment of every day
  - across various activities such as social media interactions, online transactions, digital communications, IoT device operations, and much more.

Data Never Sleeps

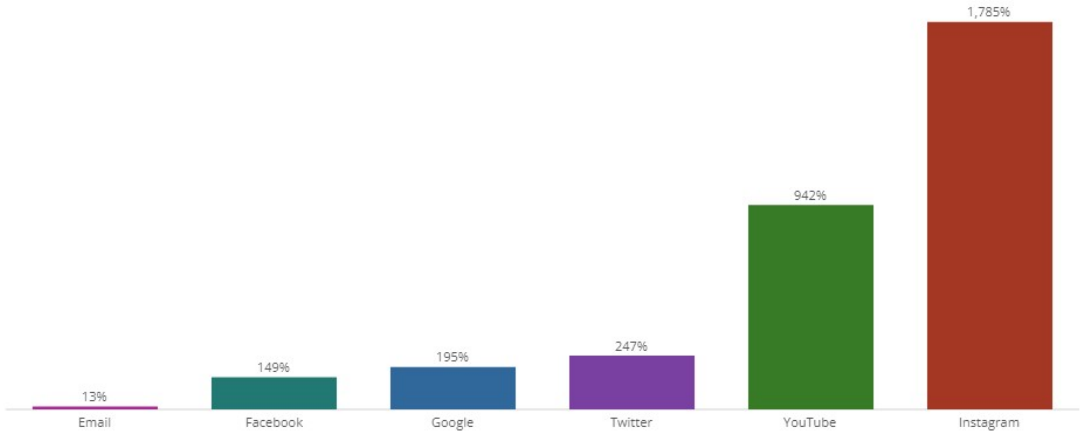


Data Never Sleeps



Data Never Sleeps

> Percentage Increase DNS1 to DNS10

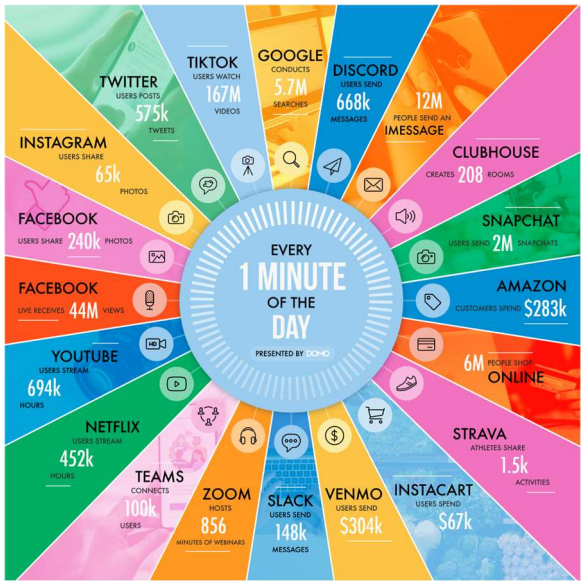


Data Never Sleeps

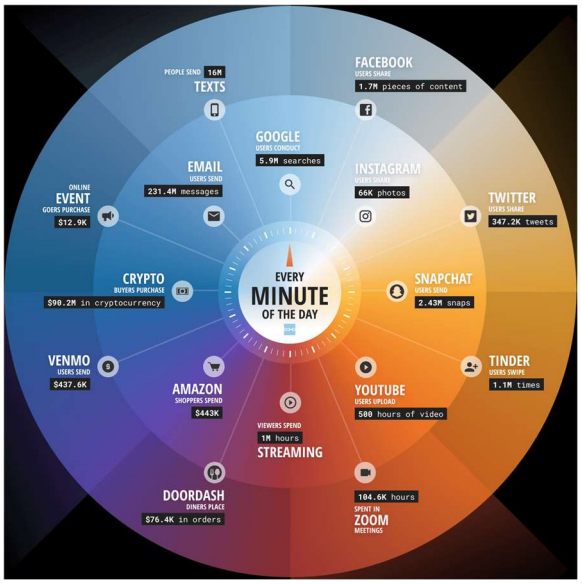
The idea behind "Data Never Sleeps" emphasizes several key points

> Continuous Data Generation
> The Importance of Real-Time Processing
> The Scale of Big Data
> Opportunities and Challenges
> The Need for Advanced Analytics and AI

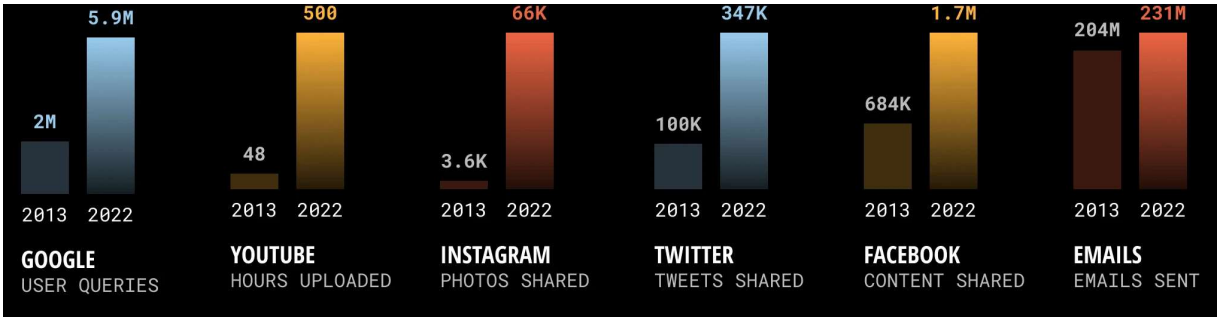
Data Never Sleeps



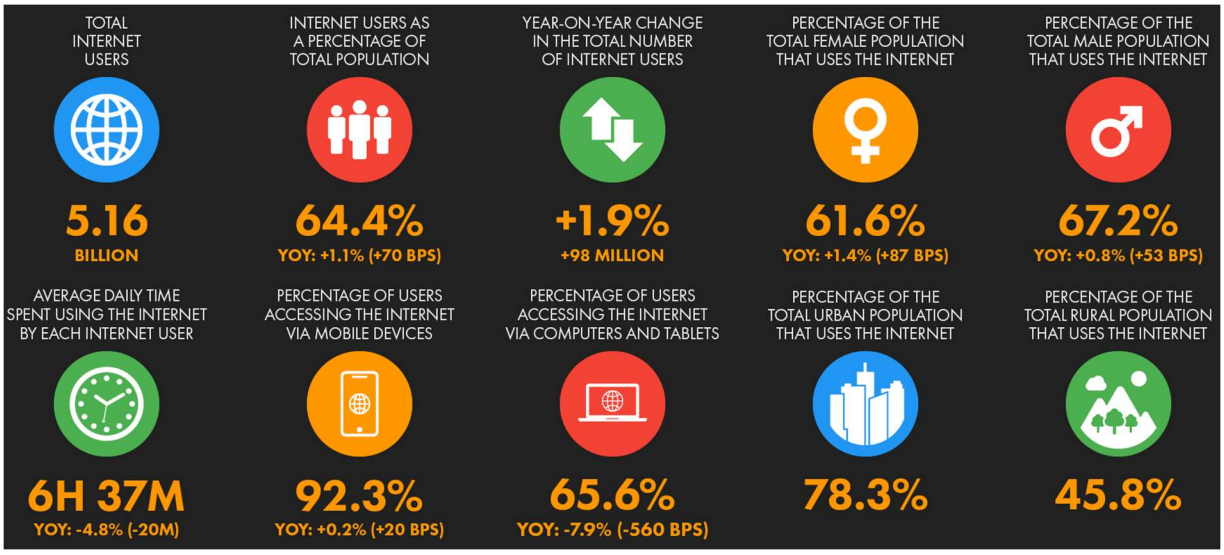
Data Never Sleeps



Data Never Sleeps

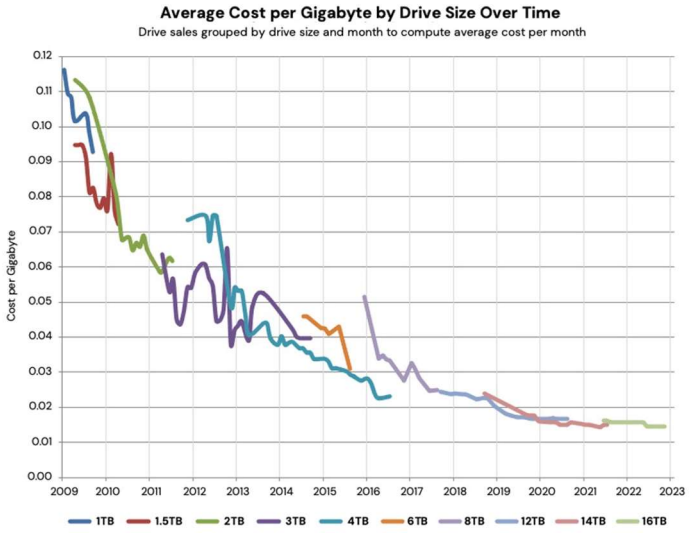


Essential Indicators of Internet Adoption and Use



Data Storage

> Data storage costs are falling



Data Science

> Interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from **structured** and **unstructured** data.

> The goal of data science

– To make informed decisions and predictions based on data.

Key Skills of a Data Scientist

- Statistical analysis and mathematics
  - for interpreting data.
- Machine learning
  - for predictive modeling and understanding complex data sets.
- Data visualization
  - for presenting data in a clear and effective manner.
- Programming skills (Python, R, SQL)
  - for processing and analyzing data.
- Domain knowledge
  - understanding the industry to make relevant conclusions.

DATA SCIENCE

DEFINITIONS

Data Science

Skill of extracting of knowledge from data

Using knowledge to predict the unknown

Improve business outcomes with the power of data

Employ techniques and theories drawn from broad areas of mathematics, statistics and information technology

Data Scientist

A practitioner of data science

Expertise in data engineering, analytics, statistics and business domain

Investigate complex business problems and ***use data*** to provide solutions

Entity

- A thing that exists about which we research and predict in data science
- Entity has a business context
- Customer of a business
- Patient at a hospital
  - The same person can be a patient and a customer, but the business context is different
- Car
  - Entities can be nonliving things

Characteristics

- Every entity has a set of characteristics. These are unique properties
- Properties too have a business context
- Customer
  - Age, income group, gender, education
- Patient
  - Age, blood pressure, weight, family history
- Car
  - Make, Model, Year, Engine, VIN

Environment

- Environment points to the eco-system in which the entity exists or functions
- Environment is shared among entities. Multiple entities belong to the same environment
- Environment affects an entity's behavior
- Customer
  - Country, City, Workplace
- Patient
  - City, Climate
- Car
  - Use (City/Highway), Climate

Event

- A significant business activity in which an entity participates
- Events happen in the past
- Customer
  - Browsing, store visit, sales call
- Patient
  - Doctor visit, blood test
- Car
  - Smog test, comparison test

Behavior

- What an entity does during an event
- Entities may have different behaviors in different environments
- Customer
  - Phone call vs email, Clickstream, response to offers
- Patient
  - Nausea, light-headed, cramps
- Car
  - Skid, acceleration, stopping distances

Outcome

- The result of an activity deemed significant by business
- Outcome values be
  - Boolean (Yes/No, Pass/Fail)
  - Continuous (a numeric value)
  - Class (identification of type)
- Customer
  - Sale (Boolean), sale value (continuous)
- Patient
  - Blood pressure value (continuous), Diabetes type (class)
- Car
  - Car type (class), stopping distances (continuous), smog passed (Boolean)

Structured Data

- Attributes are labeled and distinctly visible
- Easily searchable and query able.
- Stored easily in tables

Unstructured Data

- Data is continuous text
- Attributes are not distinctly labeled. They are present within data.
- Querying is not easy

Semi-structured Data

> Mix of structured and unstructured
> Some attributes are distinctly labeled. Others are hidden within free text

DATA SCIENCE

LEARNING
DISCOVERING
KNOWLEDGE FROM DATA

Relationships

- Attributes in dataset exhibit relationships
- Relationships "model" the real world and have a logical "Explanation"
- For attributes A and B the relationships can be
  - When A occurs, B also c
  - When A occurs B does not occur
  - When A increases B also increases
  - When A increases B decreases
- Relationships can involve multiple attributes too
  - When A is present and B increases, C will decrease

Relationships: Examples

- Customer
  - As age goes up, spending capacity goes up (AGE & REVENUE)
  - Urban customers buy more internet bandwidth (LOCATION & BANDWIDTH)
- Patient
  - Older patient have more prevalence of Diabetes (AGE & DISEASE LEVEL)
  - Overweight patients typically have higher cholesterol level (WEIGHT & HDL)
- Car
  - Sports Cars have more insurance rates (TYPE & RATES)

Relationships

- Consistent vs Incidental Patterns in Data
- Correlations
- Signals and noise

What is Learning

- Learning implies learning about relationships
- It involves
  - Taking a domain
  - Understanding the attributes that represent the domain
  - Collecting data
  - Understanding relationships between the attributes
- Model is the outcome of learning

Model

- A simplified, approximated representation of a real world phenomenon
- Captures key attributes and their relationships
- Mathematical model
  - Represents relationships as an equation
- Blood pressure

  BP=56 + (AGE * .8)+(WEIGHT * .14)+(LDL * 0.09)
- Decision Tree model
  - Represents the outcome as a decision tree
- Accuracy of models depends on strength of relationships between attributes

Prediction

- A model can be used to predict unknown attributes

  BP=56 + (AGE * .8)+(WEIGHT * .14)+(LDL * 0.09)
- The above model represents the relationships between BP, AGE, WEIGHT and LDL
- If 3 of the 4 attributes are known, the model can be used to predict the 4^{th}
- The above equation can considered the prediction algorithm
- Relationships can be a lot more complex, leading to complex models and prediction algorithms.

Predictors and outcomes

- Outcomes are attributes that you want to predict
- Predictors are attributes that are used to predict outcomes
- Learning is all about building models that can be used to predict outcomes using the predictors

Example	Predictors	Outcomes
Customer	Age, Income Range, Location	Buy? Yes/No
Patient	Age, Blood Pressure, Weight	Diabetic?
Car	Cylinders, acceleration	Sports vs family

Humans vs Machines

- Humans understand relationships and predict all the time
- Humans can only handle finite amount of data
- Machines come into play when the number of entities and data about them are large
- There in comes machine learning, predictive analytics and data science

So what is Data Science?

- Picking a problem in a specified domain
- Understanding the problem domain (entities & attributes)
- Collect datasets that represent the entities
- Discover relationships (Learning)
  - When computers are used for this purpose, its called machine learning
- Build models that represent relationships
  - Uses past data where all predictors and outcomes are known
- Use models for predicting outcomes
  - Current/future data → predictors known, outcomes unknown

Data Science Example: Website Shopper

- Problem: Predict if the shopper will buy a smartphone
- Data: Past purchase history of shoppers
  - Shopper characteristics (age, gender, income etc)
  - Seasonal information
- Build Model
  - Decision model based on shopper and seasonal entities
  - Build every week
- Prediction
  - When a new shopper is browsing, predict if the shopper will buy
- Action
  - Offer Chat help

DATA SCIENCE

DATA SCIENCE USE CASES

Data Science Applications

> The use of data science is growing exponentially into and across multiple domains in business, science, finance, …

> Recent advances in computing power, open-source software and predictive algorithms have made it cost effective to effective to apply data science for commercial use

Data Science Applications

- Healthcare: Predicting disease outbreaks, personalized medicine.
- Finance: Fraud detection, risk management.
- E-commerce: Personalized customer experiences, inventory management.
- Transportation: Optimizing routes, predicting vehicle maintenance.
- Social Networking: Recommendation systems, targeted advertising.

Fraud Detection

- Credit Card Frauds exhibit patterns in transactions
- Historical Transaction Data are used to identify fraudulent patterns build models
- Each new transaction is then given a fraud score based on the model
- Action taken for high scored transactions

Recommendations

- Items exhibit patterns on how they are brought together
  - Cell phone and accessories
  - Books
- Patterns used to build affinity scores between items
- When one item is brought, items with high affinity scores to that item are recommended

Scoring of Callers and Agents

- Past interactions used to score callers based on their value or type
- Agents are scored based their ability to sell or handle a specific type of problem
- The right callers and then matched with the right agents to optimize business outcomes
- Call recordings analyzed using machine learning to grade quality of call and outcome

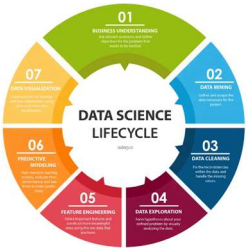
Predicting Disease Outbreaks

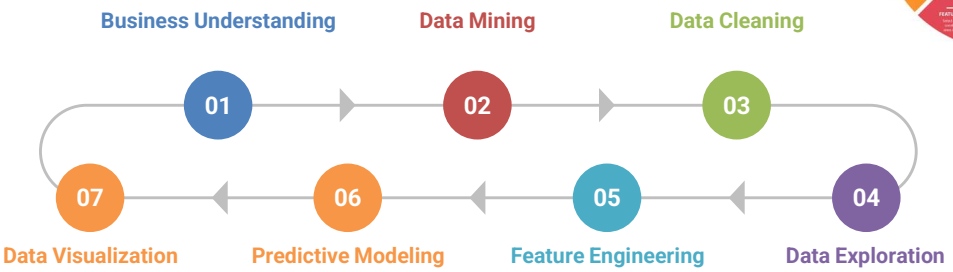
- Dataset collection from public domains like Google searches, Twitter feeds etc
- Data linked with location information and disease patterns to build outbreak forecasting models
- Model used to track potential outbreaks and take preventive actions

DATA SCIENCE

PROJECT LIFECYCLE

Data Science Project Lifecycle





Business Understanding

- **Objective Identification**
  - Understand and define the business problem. Identify the goals and objectives of the project.
- **Requirements Gathering**
  - Collect requirements, expectations, and constraints from stakeholders.

Business Understanding

- **Objective Identification**
  - Understand and define the business problem. Identify the goals and objectives of the project.
- **Requirements Gathering**
  - Collect requirements, expectations, and constraints from stakeholders.

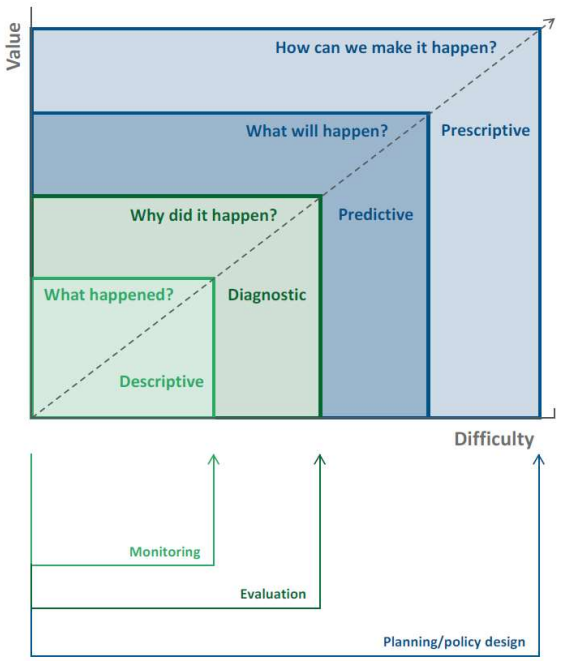
DATA SCIENCE

ANALYTICS AND PREDICTIONS

Types of Analytics

Type of Analytics	Description
Descriptive	Understand what happened
Exploratory	Find out why something is happening
Inferential	Understand a population from a sample
Predictive	Forecast what is going to happen
Causal	What happens to one variable when you change another
Deep	Use of advanced techniques to understand large and multi-source datasets

Types of Analytics



Goals of EDA

- Understand the predictors and targets in the data set
  - Spreads
  - Correlations
- Uncover the patterns and trends
- Find key variables and eliminate unwanted variables
- Detect outliers
- Validate previous data ingestion processes for possible mistakes
- Test assumptions and hypothesis

Tools used for EDA

- Correlation matrices
- Boxplots
- Scatterplots
- Principal Component Analysis
- Histograms