Module 1

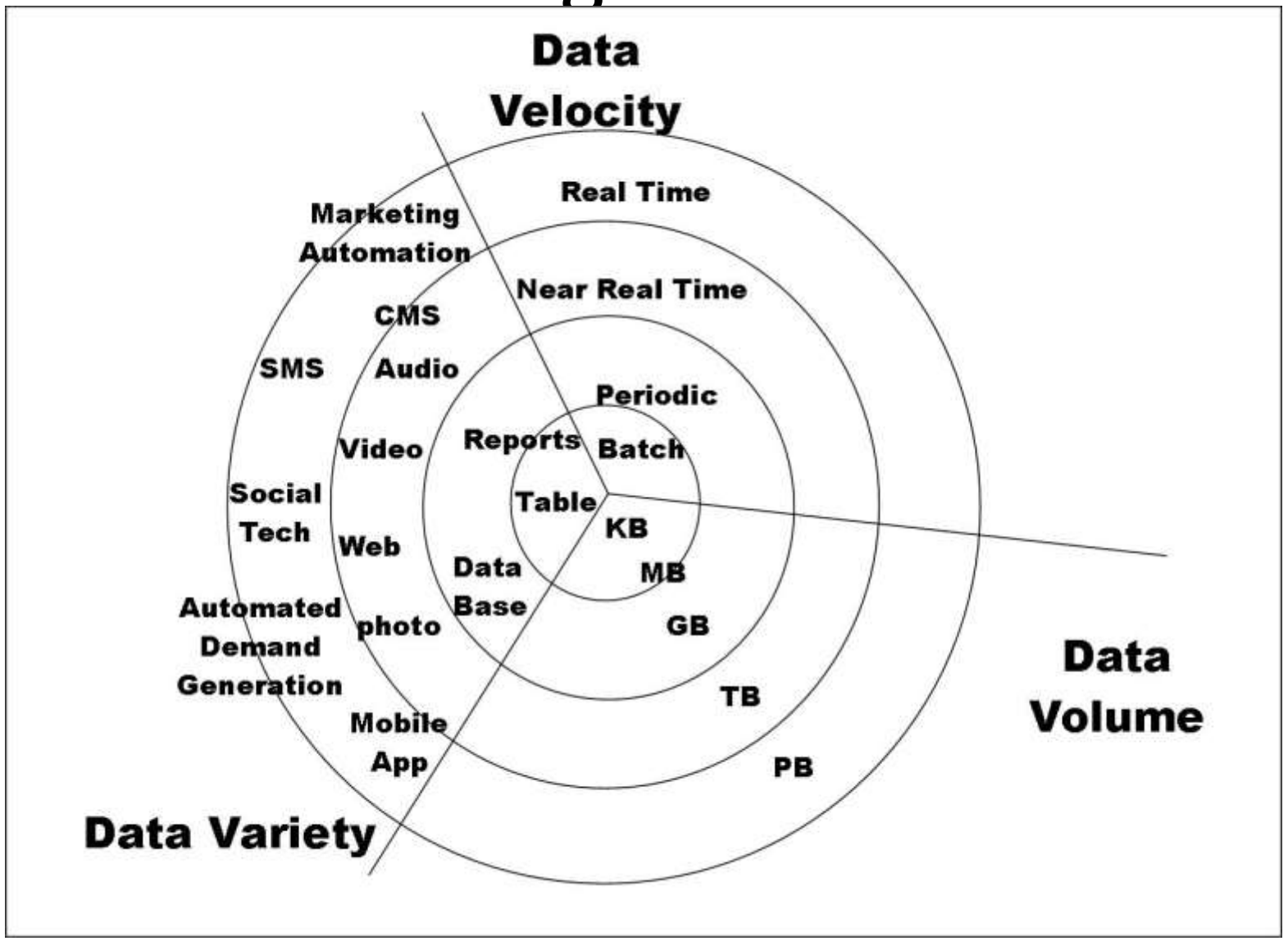
- ✓Data Science Context
- ✓Need for Data Science
- ✓What is Data Science
- ✓Data Science Process
- ✓Business Intelligence and Data Science
- ✓Prerequisites for a Data Scientist
- ✓Tools and Skills
- ✓required.

- ✓ Data, Big Data and Challenges
- ✓ Data Science
  - ✓ Introduction
  - ✓ Why Data Science
- ✓ Data Scientists
  - ✓ What do they do?
- ✓ Major/Concentration in Data Science
  - ✓ Tools to take.

Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - ✓Volume
    - ✓The size of the data
  - ✓Velocity
    - ✓The latency of data processing relative to the growing demand for interactivity
  - ✓Variety and Complexity
    - ✓the diversity of sources, formats, quality, structures.

Big Data


Data
Velocity
Real Time
Near Real Time
Periodic
Batch
Marketing
Automation
CMS
SMS
Audio
Video
Reports
Table
Social
Tech
Web
KB
MB
GB
TB
PB
Data
Base
Automated
Demand
Generation
photo
Mobile
App
Data Variety
Data
Volume

Types of Data We Have

- ✓ Relational Data (Tables/Transaction/Legacy Data)
- ✓ Text Data (Web)
- ✓ Semi-structured Data (XML)
- ✓ Graph Data
- ✓ Social Network, Semantic Web (RDF), …
- ✓ Streaming Data
- ✓ You can afford to scan the data once

Data Science Context

- ✓ Data Science : Combination of multiple disciplines
- ✓ statistics, data analysis, and machine learning
- ✓ An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data
- ✓ Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data
- ✓ **Data science principles apply to all data – big and small**

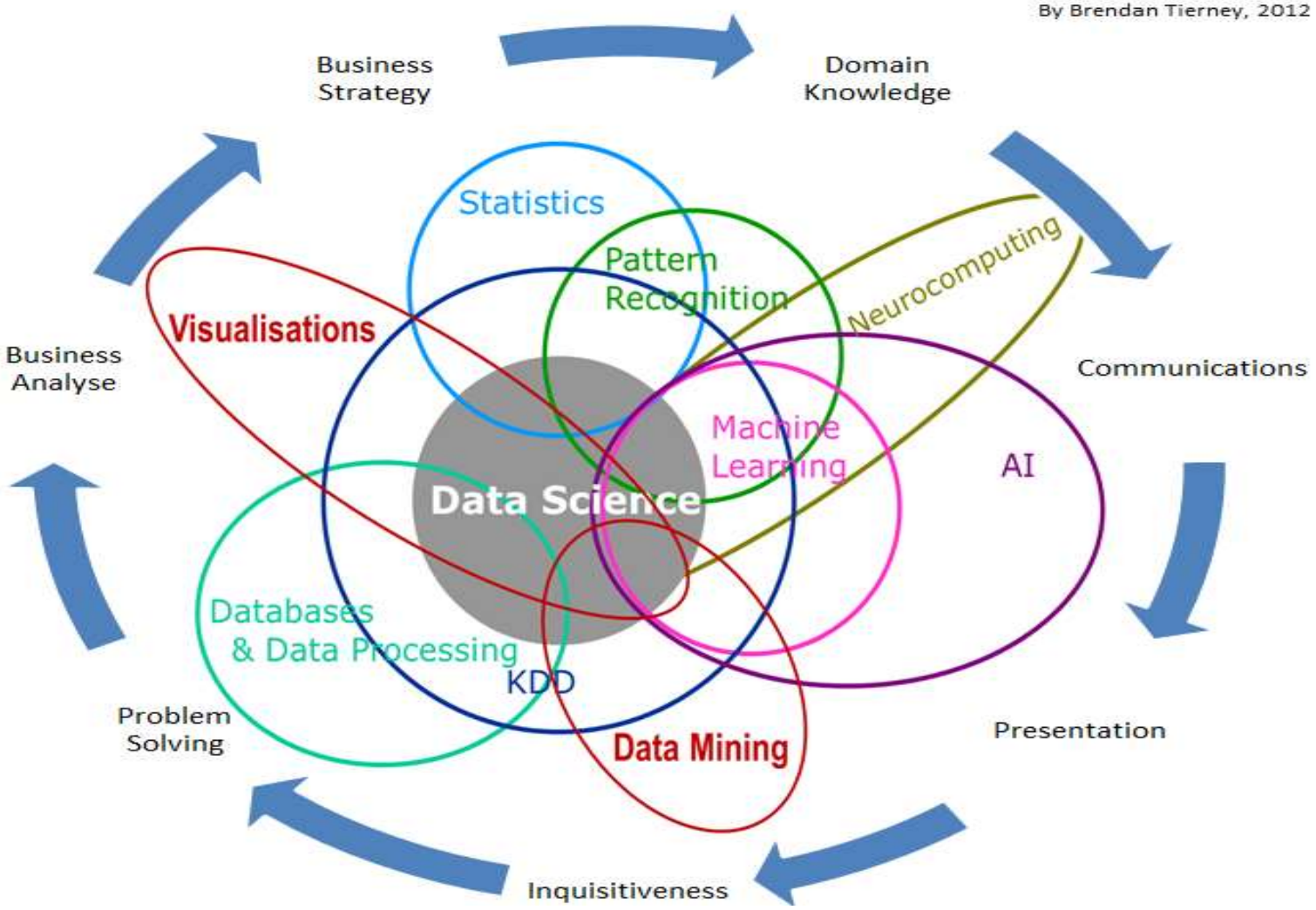
- ✓ To analyze data and extract knowledge and insights from it

Steps: data gathering, analysis and decision-making

- ✓ Finding patterns in data, through analysis, and make future predictions
- ✓ **Better decisions (should we choose A or B)**
- ✓ **Predictive analysis (what will happen next?)**
- ✓ **Pattern discoveries (find pattern, or maybe hidden information in the data)**

Data Science Is Multidisciplinary

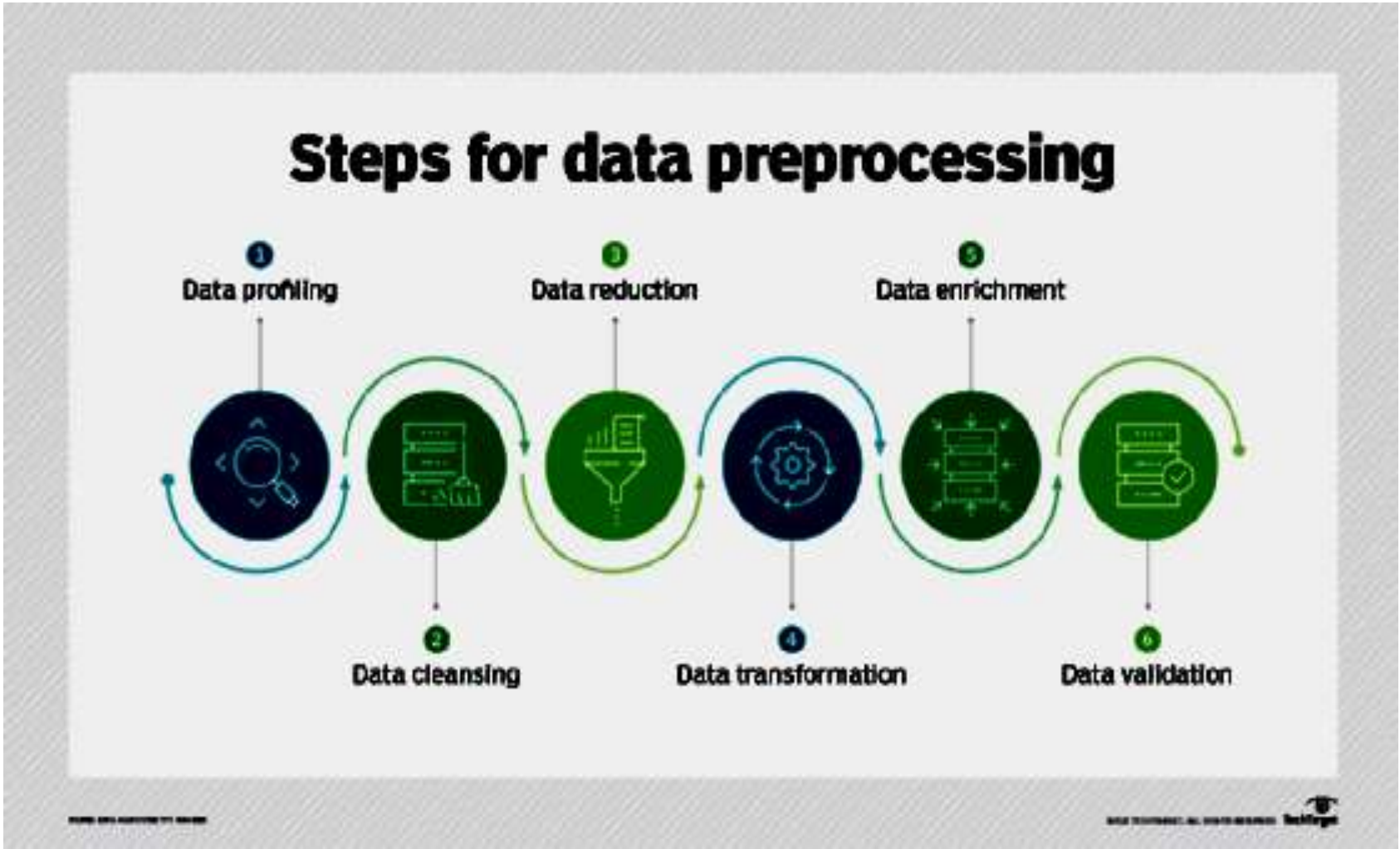
By Brendan Tierney, 2012



Need for Data Science

- ✓ For route planning: To discover the best routes to ship
- ✓ To foresee delays for flight/ship/train etc. (through predictive analysis)
- **Create promotional offers**
- **To find the best suited time to deliver goods**
- **To forecast the next years revenue for a company**
- **To analyze health benefit of training**
- **To predict who will win elections**

STEPS


Steps for data preprocessing
1
Data profiling
2
Data cleansing
3
Data reduction
4
Data transformation
5
Data enrichment
6
Data validation

Activity 1

List out 25 Real time Data Science Applications

- Choose one and Extract data

Need for Data Science

- ✓ Data science plays a crucial role in various industries and sectors
- ✓ Addressing numerous challenges and providing valuable insights that contribute to informed
- ✓ Decision-making and strategic planning. Here are some of the key reasons why data science is essential

Data-Driven Decision

- ✓ Data science enables organizations to make informed decisions based on data-driven
- ✓ Decision relying solely on intuition or experience.
- ✓ By analyzing large datasets businesses can identify trends, patterns, and correlations that help them understand customer behavior, market trends, and internal operations.

Improved Efficiency and Productivity

- ✓ Through data analysis, organizations can identify inefficiencies in their processes and operations
- ✓ Allowing them to streamline workflows and optimize resource allocation.
- ✓ This leads to improved productivity, reduced costs, and better resource management.

Enhanced Customer Understanding and Personalization

- ✓ Data science helps businesses gain a deeper understanding of their customers' preferences, behaviors, and needs.
- ✓ By analyzing customer data, organizations can personalize their products, services, and marketing strategies.
- ✓ Leading to improved customer satisfaction and loyalty.

Predictive Analysis and Forecasting

- ✓ Data science facilitates predictive analysis,
- ✓ Allowing businesses to forecast future trends, market demands, and customer behavior.
- ✓ This capability enables organizations to proactively plan and adapt .
- ✓ Their strategies to meet changing market dynamics and customer preference.

Risk Management and Fraud Detection

- Data science helps in identifying and mitigating potential risks by analyzing
- Historical data and identifying patterns that may indicate potential risks or fraudulent activities.
- This is particularly crucial in industries such as finance, insurance, and cybersecurity.

Innovation and Product Development

- Data science can drive innovation by providing insights into market demands, emerging trends, and customer preferences.
- By understanding customer needs and preferences, businesses can develop
- New products and services that better meet the requirements of their target audience

Healthcare Advancements

- In the healthcare sector, data science contributes to advancements
- In disease prediction, personalized medicine, and the improvement of patient care.
- By analyzing large datasets, healthcare professionals can make more accurate
- Diagnoses, develop effective treatment plans, and improve overall patient outcomes.

Optimized marketing strategies

- ✓ Data science enables businesses to create targeted and effective marketing campaigns
- ✓ By analyzing consumer behavior, preferences, and responses to previous marketing efforts.
- ✓ Increased conversion rates, and improved return on investment (ROI).

Data science finds applications in a wide range of fields and industries

- ✓ Data science is used in businesses and industries for market analysis,
- ✓ Customer segmentation, demand forecasting, supply chain optimization, and customer relationship management.
- ✓ Helps businesses make data-driven decisions and develop effective strategies
- ✓ To enhance their operations and achieve their goals.
- ✓ Businesses can stay ahead of their competitors.
- ✓ Adapt to market changes more quickly, and capitalize on emerging opportunities.

Healthcare

- ✓ In healthcare, data science is employed for predictive analysis, patient diagnosis,
- ✓ Drug development, and personalized medicine.
- ✓ It helps healthcare professionals make accurate diagnoses,
- ✓ Identify potential health risks, and improve patient care and outcomes.

Finance and Banking

- ✓ Data science is utilized in the finance and banking sectors for risk management,
- ✓ Fraud detection, algorithmic trading, and customer analytics.
- ✓ It helps financial institutions make informed decisions, detect fraudulent activities,
- ✓ Provide personalized financial services to their customers

E-commerce and Retail

- ✓ Data science is used in e-commerce and retail industries for recommendation systems,
- ✓ Market basket analysis, and customer behavior prediction.
- ✓ It helps businesses provide personalized product recommendations.
- ✓ Optimize pricing strategies, and enhance
- ✓ The overall shopping experience for customers.

Big Data Sources

Manufacturing and Logistics

- ✓ Data science is employed in manufacturing and logistics for supply chain management, predictive maintenance, and quality control.
- ✓ It helps optimize production processes, predict equipment failures
- ✓ The timely and efficient delivery of products to customers.

Social Media and Marketing

- ✓ Data science is utilized in social media and marketing for sentiment analysis.
- ✓ targeted advertising, and customer engagement analysis.
- ✓ It helps businesses understand customer preferences, monitor brand perception.
- ✓ Develop effective marketing campaigns to reach the right audience.

Telecommunications

- ✓ In the telecommunications industry, data science is used for network
- ✓ Optimization, customer churn prediction, and service quality improvement.
- ✓ It helps telecommunications companies improve network performance, retain customers,
- ✓ Provide better services to their subscribers.

Education

- ✓ Data science is employed in education for personalized learning
- ✓ Student performance analysis, and curriculum development.
- ✓ It helps educators understand student behavior, identify learning patterns, and create tailored learning experiences for students.

Government and Public Policy

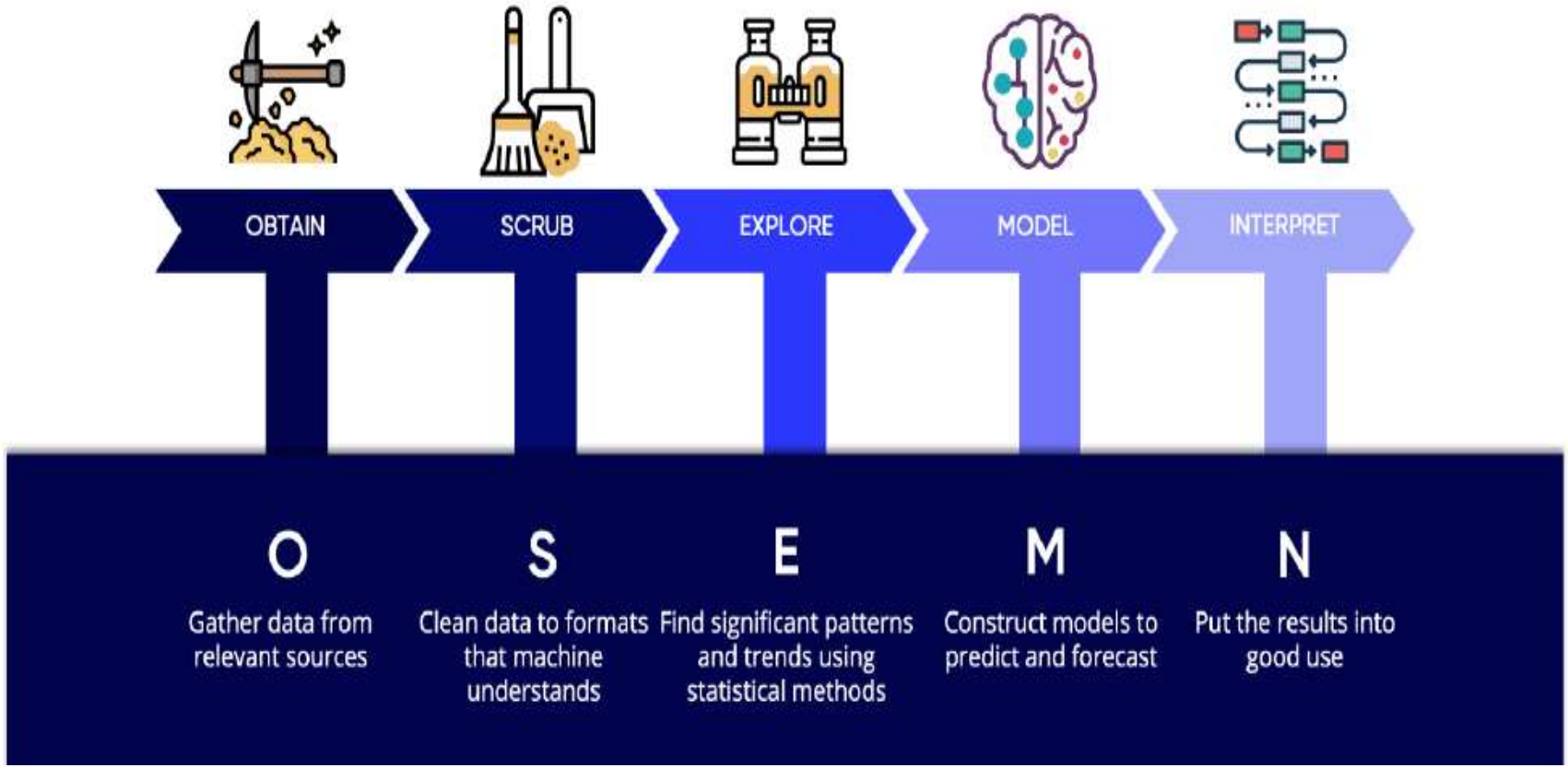
- ✓ Data science is utilized in the government sector for public health analysis, urban planning, and policy-making.
- ✓ It helps government agencies make data-driven decisions, develop effective public policies, and improve the overall well-being of cites .
- ✓ These are just a few examples of how data science is used in various industries and sectors
- ✓ To generate valuable insights, drive innovation, and make informed decisions

Data Science Process

- ✓ The data science process typically involves a series of steps.
- ✓ that enable data scientists to extract meaningful insights and valuable information from data.
- ✓ While the specific details of the process can vary depending on the project
- ✓ The organization, following is a generalized outline of the key steps involved in the data science process.

Data Science Process

LEAD



Problem Definition

- ✓ The first step is to clearly define the problem or question that the data analysis aims to address.
- ✓ This involves understanding the business context, defining project goals, and establishing key performance indicators (KPIs) to measure the success of the project.

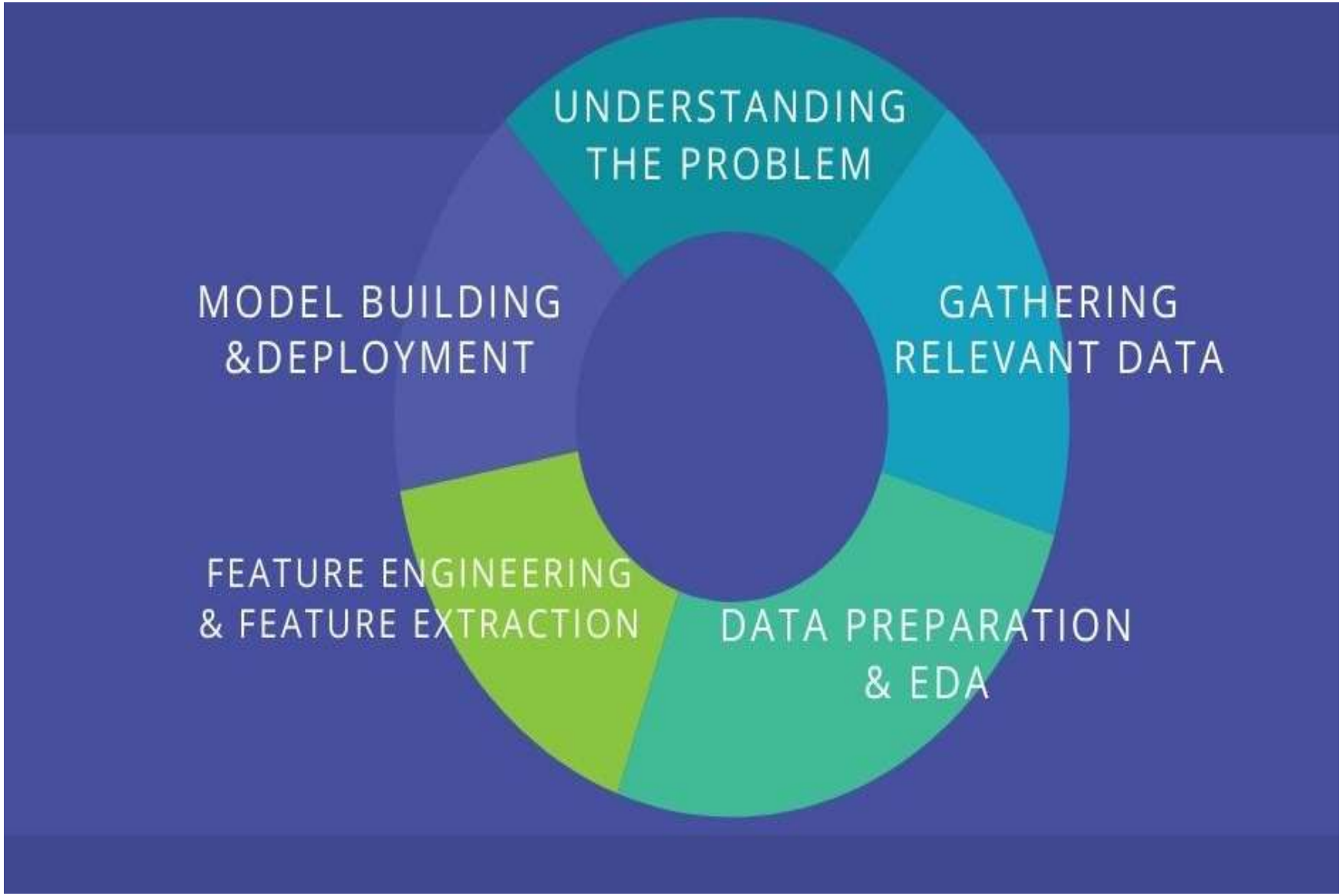
Data Collection

- ✓ In this step, relevant data is collected from various sources, which may include databases.
- ✓ APIs, data warehouses, or external data sets It's important to ensure
- ✓ That the data collected is relevant, reliable, and sufficient for addressing the defined problem.

Data Cleaning and Preprocessing

- ✓ Raw data often contains errors, missing values, outliers, and inconsistencies that need to be addressed before analysis.
- ✓ Data cleaning and preprocessing involve tasks such as handling missing data, removing duplicates, normalizing data, and dealing

Data Science Project Lifecycle



Pre Requisite for Data Scientist

Non Technical Prerequisite

- ✓ Curiosity -curiosity and ask various questions, then you can understand the business problem easily
- ✓ Critical Thinking -find multiple new ways to solve the problem with efficiency.
- ✓ Communication skills - you need to communicate it with the team

Technical Prerequisite

Machine learning - Concept of machine learning. Data science uses machine learning algorithms

Mathematical modeling - make fast mathematical calculations and predictions from the available data

Statistics - Basic understanding of statistics is required, such as mean, median, or standard deviation

Computer programming - one programming language is required. R, Python, Spark

Databases - Depth understanding of Databases such as SQL

Contrast: Databases

	Databases	Data Analytics
Data Value	**"Precious"**	**"Cheap"**
Data Volume	**Modest**	**Massive**
Examples	**Bank records, Personnel records, Census, Medical records**	**Online clicks, GPS logs, Tweets, Building sensor readings**
Priorities	**Consistency, Error recovery, Auditability**	**Speed, Availability, Query richness**
Structured	**Strongly (Schema)**	**Weakly or none (Text)**
Properties	**Transactions, ACID***	**CAP* theorem (2/3), eventual consistency**
Realizations	**SQL**	**NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,…**

ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

Contrast: Databases

Databases	Data Science
Querying the past	Querying the future

- ✓ Business intelligence (BI) is the transformation of raw data into meaningful and useful information for business analysis purposes.
- ✓ BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities

Contrast: Machine Learning

Machine Learning	Data Analytics
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	Take action!

DOING DATA ANALYTICS

Jeff Hammerbacher's Model

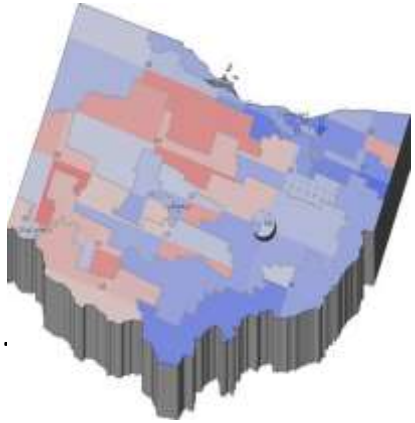
1. Identify problem

2. Instrument data sources

3. Collect data

4. Prepare data (integrate, transform, clean, filter, aggregate)

5. Build model



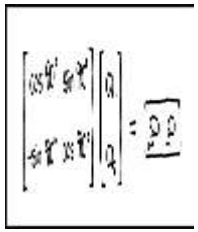
6. Evaluate model

7. Communicate resul

Data Scientist's Practice

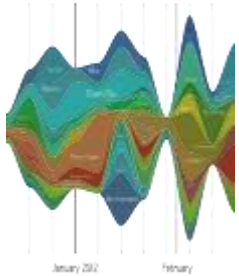
Digging Around in Data

Clean, prep



Hypothesize Model

Large Scale Exploitation



Evaluate Interpret

Data Analytics tools

- **Candela -** web visualization components.
- **Charted -C**reate line graphs or bar charts from CSV files and Google spreadsheets.
- **Datawrapper -** mobile-friendly tool,create charts and reports in seconds
- **Google Data Studio**
- **Goolge Charts**
- **Leaflet**
- **Myheatmap**
- **Open heatmap**
- **Palladio**
- Rawgraphs
- **Tableau Public**
- **Timeline**
- **Chartist.js**
- **Colorbrewer**
- **Ploty**
- **Polymaps**
- Weave
- Dygraphs
- Ganttpro

Open Source Big Data Tools

- ✓Apache Hadoop
- ✓Apache Storm
- ✓Apache Spark
- ✓Apache Cassandra
- ✓Mongo DB
- ✓R Programming Environment
- ✓Neo 4J
- ✓**Apache SAMOA**

What's Hard about Data Analytics

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

TOOLS AND SKILLS

- Data scientists need to be proficient in various tools and possess a diverse skill set
- To effectively handle and analyze complex data sets.
- You need some of the essential tools and skills to become a data scientist

Missing Data

There is no good way to deal with missing data

Reason : Why Data Goes Missing?

- ✓ **Missing at Random (MAR):**

  Propensity for a data point to be missing is not related to the missing data, s related to some of the observed data.

- ✓ **Missing Completely at Random (MCAR):**

  certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

- ✓ **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value

(e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value

(e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

Missing Value Treatment

- ✓ Missing data in the training data set can reduce the power / fit of a model or can lead to a biased model because we have not analyzed the behavior and relationship with other variables correctly. It can lead to wrong prediction or classification.

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%

Why Data's are having missing values

Data Extraction

- ✓ problems with extraction process.
- ✓ Double-check for correct data with data Owners.
- ✓ Some hashing procedures can also be used to make sure data extraction is correct.
- ✓ Errors at data extraction stage -Easy to find,Corrected easily as well.

Data collection:

- ✓ Occur at time of data collection and are harder to correct.

Methods to treat Missing Values

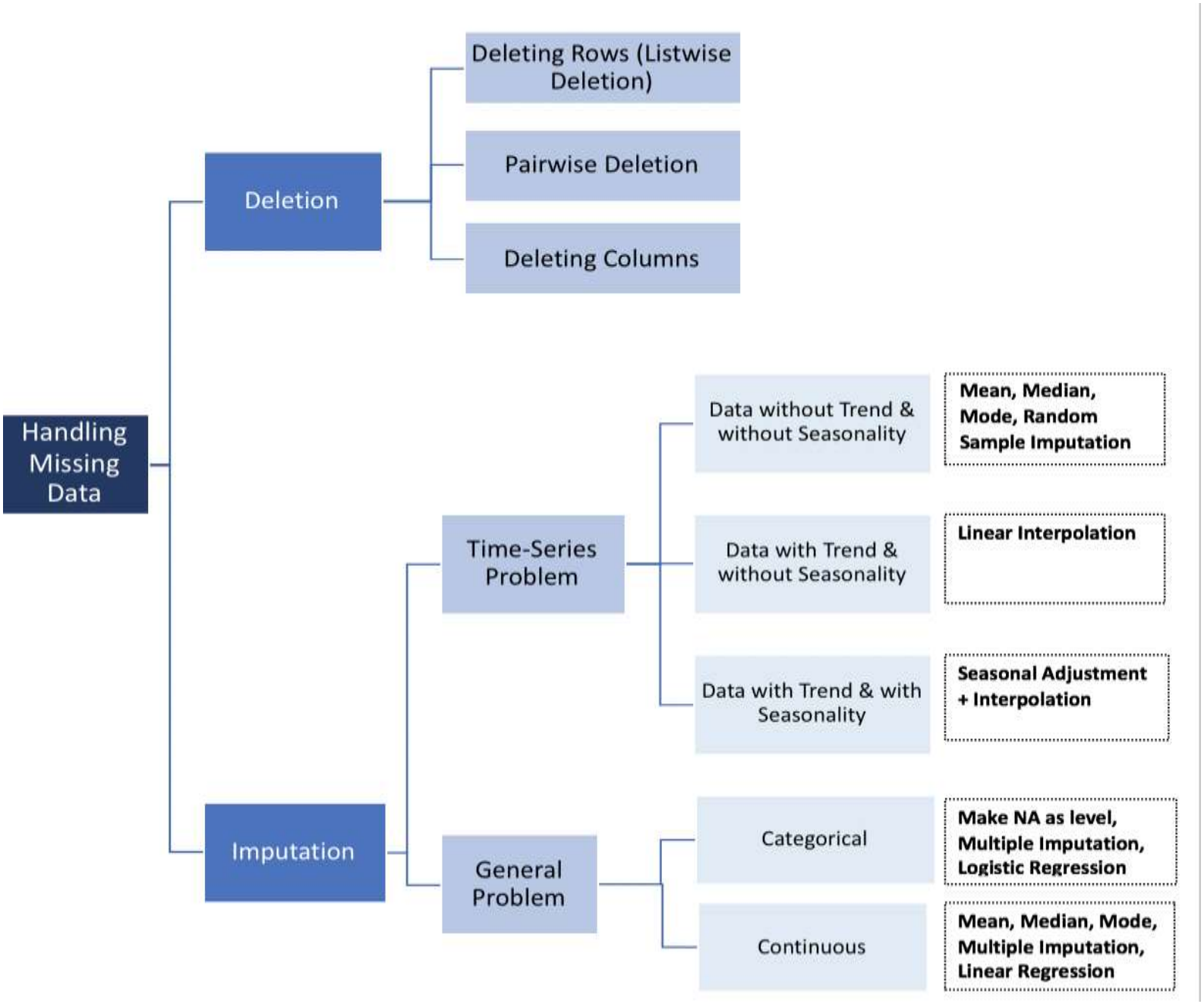
Deletion: List Wise Deletion and Pair Wise Deletion

List wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	.	280
M	33	332
M	.	272
F	25	.
M	29	326
	26	259
M	32	297


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Different Ways: Compensate for Missing Values In a Dataset (Data Imputation with examples)

- Many real-world datasets may contain missing values for various reasons. They are often encoded as NaNs, blanks or any other placeholders
- Training a model with a dataset that has a lot of missing values can drastically impact the machine learning model's quality
- algorithms such as *scikit-learn estimators* assume that all values are numerical and have and hold meaningful value.

Do Nothing: That's an easy one.

- You just let the algorithm handle the missing data.
- Some algorithms can factor in the missing values and learn the best imputation values for the missing data based on the training loss reduction .
- Others have the option to just ignore them .
- However, other algorithms will panic and throw an error complaining about the missing values.
- In that case, you will need to handle the missing data and clean it before feeding it to the algorithm.

▪ Imputation Using (Mean/Median) Values:

Calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others.

It can only be used with numeric data.
Easy and fast.
Wo

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Imputation Using (Most Frequent) or (Zero/Constant) Values

- Replacing missing data with the most frequent values within each column.
- Works with categorical features (strings or numerical representations
- Replaces the missing values with either zero or any constant value you specify

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Imputation Using k-NN:

- The *k* nearest neighbours is an algorithm that is used for simple classification.
- The algorithm uses '**feature similarity**' to predict the values of any new data points.
- This means that the new point is assigned a value based on how closely it resembles the points in the training set.
- This can be very useful in making predictions about the missing values by finding the *k's* closest neighbours to the observation with missing data and then imputing them based on the non-missing values in the neighbourhood
- Can be much more accurate than the mean, median or most frequent imputation methods (It depends on the dataset).

- Computationally expensive. KNN works by storing the whole training dataset in memory.

Imputation Using Multivariate Imputation by Chained Equation (MICE)

- This type of imputation works by filling the missing data multiple times.
- Multiple Imputations (MIs) are much better than a single imputation as it measures the uncertainty of the missing values in a better way.
- The chained equations approach is also very flexible and can handle different variables of different data types

Imputation Using Deep Learning (Datawig):

- This method works very well with categorical and non-numerical features.
- It is a library that learns Machine Learning models using Deep Neural Networks to impute missing values in a dataframe

List of R Packages

- ✓MICE
- ✓Amelia
- ✓missForest
- ✓Hmisc
- ✓mi