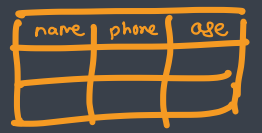
Data Analytics

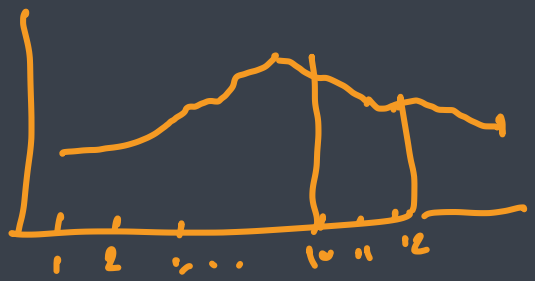
Data Analytics

- Data analytics is a broad term that defines the concept and practice (or, perhaps science and art) of all activities related to data → collection, organization .....
- Data that sits raw, as-is, has no value 1, 5, 8, 9 → raw data (meaningless)
- Instead, it's what you do with that data that provides value



→ meaningful information

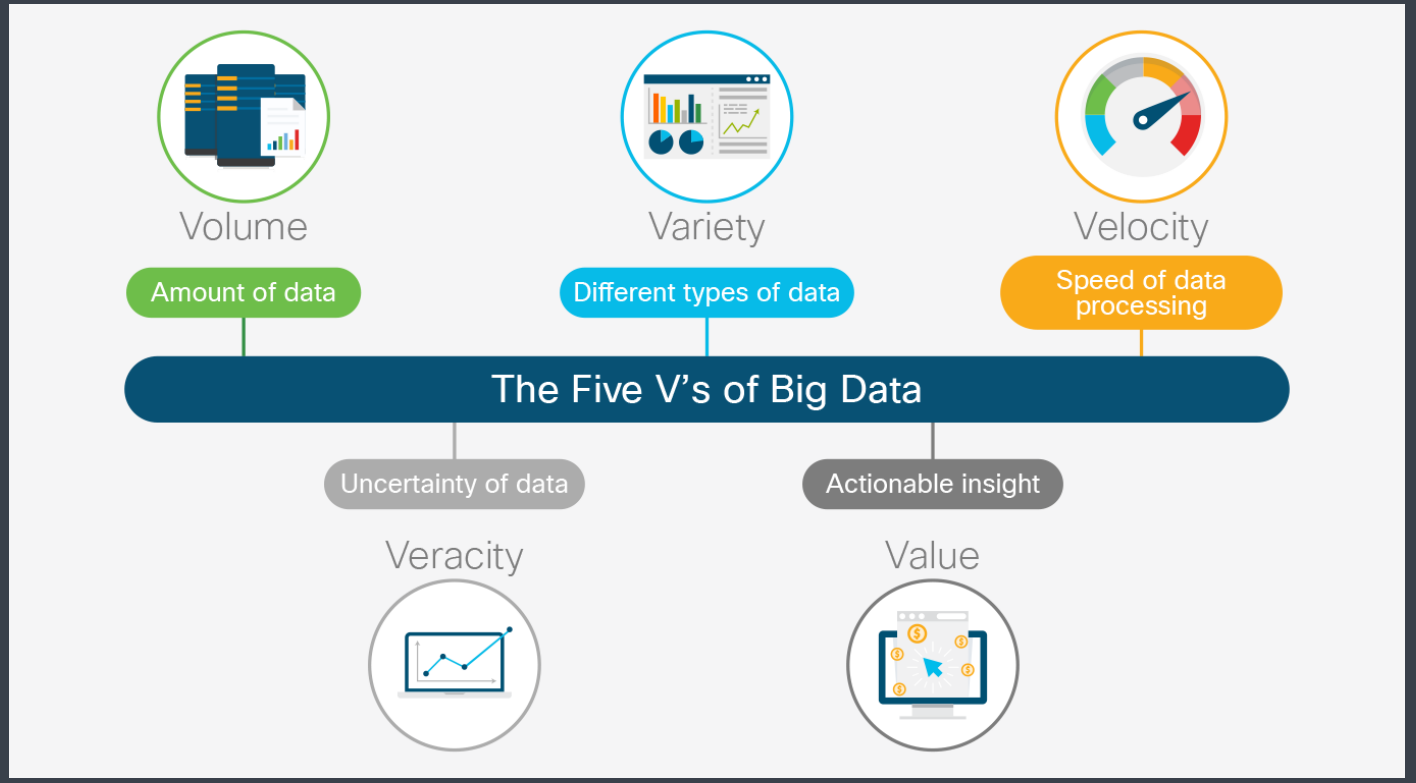
- Data analytics includes all the steps you take, both human- and machine-enabled, to discover, interpret, visualize, and tell the story of patterns in your data in order to drive business strategy and outcomes
- A successful data analytics practice can—should—provide a better strategy for where your business can go. When done well, data analytics can help you:
  - Find trends → patterns
  - Uncover opportunities
  - Predict actions, triggers, or events
  - Make decisions



Data Challenges

Data Challenges

- It is a nontrivial exercise to turn data into information, in large part because of what's become known as the *five "V's" of big data*
- Volume
- Variety
- Velocity
- Veracity
- Value



Volume

- The volume of data has grown so much that traditional relational database management software running on monolithic servers is incapable of processing it

Variety

structured
semi-structured
unstructured

- The *variety* of data has increased. Structured data has been the norm, but is no longer. Unstructured data requires new approaches to collect, store, and process it into something useful.
- Our ability to process unstructured data changes the way we approach data analysis and application development. Previously, you would define your data structures and build an application to operate on the data. The lifetime of these kinds of applications was often measured in years.
- Now, data is collected in its original fidelity and explored for meaningful patterns. Applications are built to take advantage of these patterns, and then rebuilt when new patterns emerge. The lifetime of the applications is much shorter: months to days.

patterns → model

Velocity

- Velocity is the measure of how fast the data is coming in
- It has to be processed, often in real time, and stored in huge volume

Kafka

Veracity

- Veracity of data refers to how much you can trust it
- Traditional structured data goes through a validation process during which it's cleaned up, validated, and joined with related data
- This requires that you already know a lot about the data before processing it
- This approach does not work with unstructured data

Value

- Deriving value from the data is non-trivial, especially in light of the challenges presented by the other V's discussed above.
- Much of the progress made in compute, storage, and networking, and in distributed data processing frameworks is aimed at addressing these challenges with the goal of converting data into actionable insight

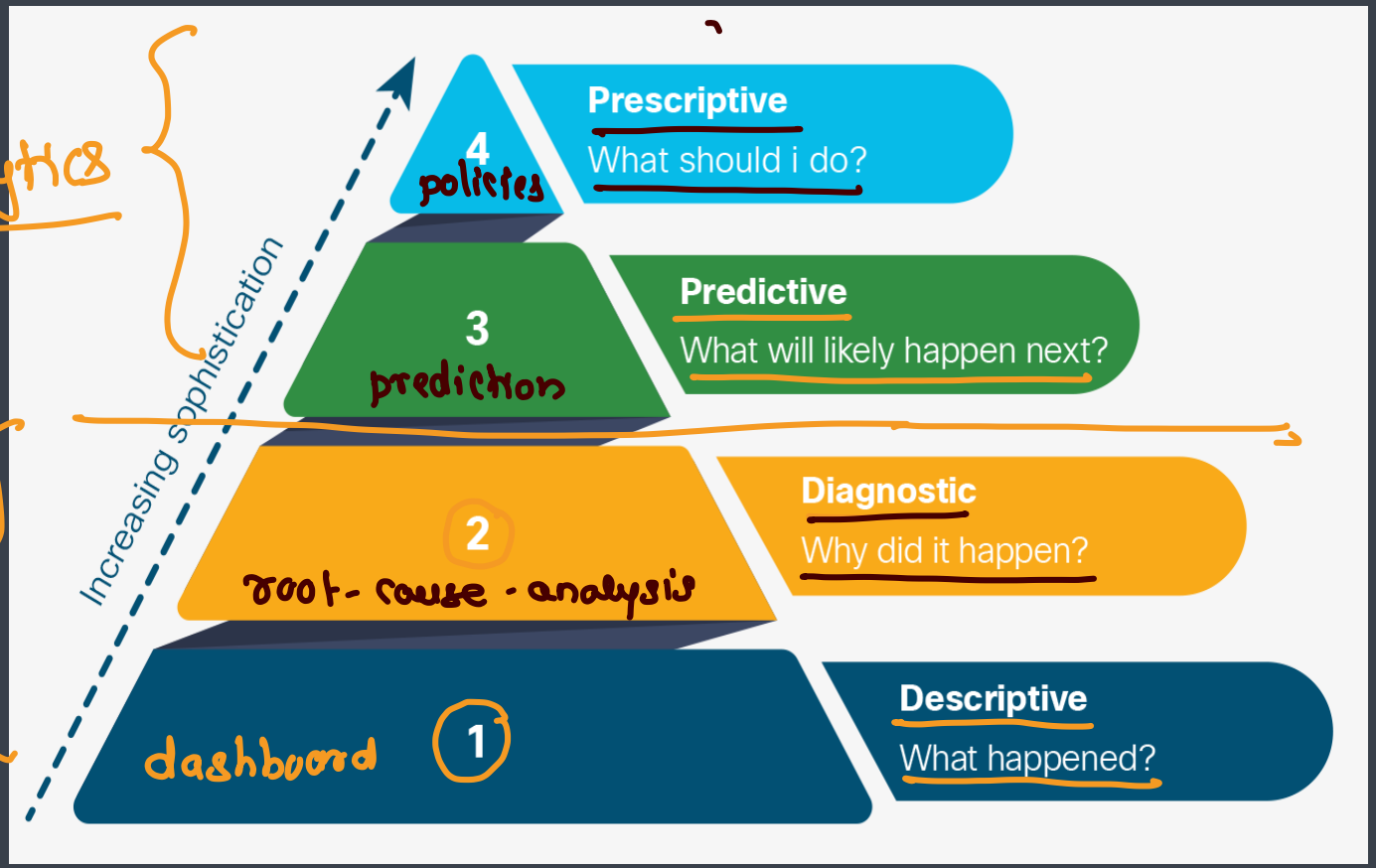
Types

Types of Data Analytics

- There are four types of analyses, ranging from relatively basic descriptions of what has happened to very sophisticated guidance on what actions to take

analysis

analytics



Descriptive Analytics: What Happened?

past data

- Descriptive analytics summarizes past raw data putting it into a form usable by humans
- Its emphasis is on helping the user understand what has already happened; for example, summarizing all sales data for all products in all regions. This analysis can be quite useful, allowing the user to view the data in many different ways.
- Note that the term "past" can be misleading → a seconds before data is also known as past data
- What happened one minute, or even one second ago is in the past
- Thus, real-time streaming data is often used in descriptive analytics → Kafka
- A common form for the analysis is the data displayed in a dashboard that updates as new data arrives

→ visual aspect of data

Diagnostic Analytics: Why Did It Happen?

- Diagnostic analytics are used for root-cause analysis to understand why something happened
- For example:
  - why sales were low in a particular region last month (the weather)
  - why sales for a particular product were low (high returns due to quality issues, or competitor pricing)
  - why a product is out of stock (supply chain issues or an unexpected spike in demand)
- Knowing what happened from descriptive analytics and why it happened from diagnostic analytics enables management to make informed decisions about what course of action to take

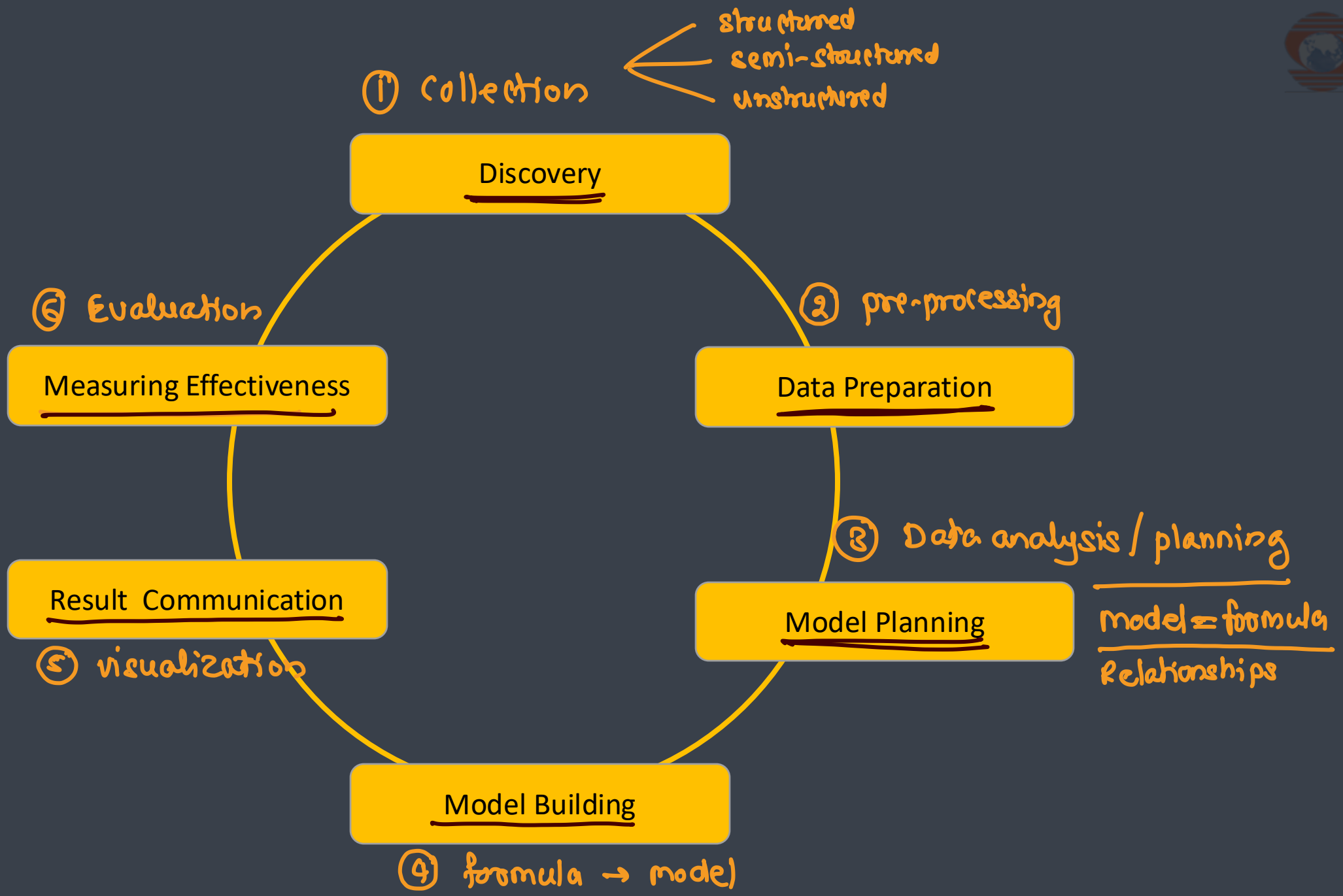
Predictive Analytics: What Will Likely Happen Next?

- Predictive analytics goes one step further
- By analysing patterns and trends in past data, analysts can make predictions about what could happen in the future, using this information to set goals and make plans
- Predictive analytics applies a variety of statistical algorithms to past data, often correlating it with past outcomes in an attempt to forecast what will happen next
- Examples of common uses include the following:
  - What-if analysis
  - Data mining
  - Root cause analysis
  - Forecasting
  - Monte Carlo simulation

Prescriptive Analytics: What Should I Do? — policies / strategies

- Prescriptive analytics is the next logical step after predictive analysis
- Prescriptive analytics is a combination of data, mathematical models and various business rules that are used to run simulations—often called scenario analysis—using various sets of assumptions
- By varying the assumptions and applying a variety of optimization techniques prescriptive analysis suggests what actions are likely to maximize a given business outcome
- In other words, prescriptive analysis explores several possible actions helping to identify what actions produce the best result

Life Cycle


① Collection
structured
semi-structured
unstructured
Discovery
② pre-processing
Data Preparation
③ Data analysis / planning
model = formula
Relationships
Model Planning
④ formula → model
Model Building
⑤ visualization
Result Communication
⑥ Evaluation
Measuring Effectiveness

Data Discovery and Formation

- Everything begins with a defined goal. In this phase, you'll define your data's purpose and how to achieve it by the time you reach the end of the **data analytics lifecycle**.
- The initial stage consists of mapping out the potential use and requirement of data, such as
  - where the information is coming from
  - what story you want your data to convey
  - how your organization benefits from the incoming data
- Basically, as a data analysis expert, you'll need to focus on enterprise requirements related to data, rather than data itself
- Additionally, your work also includes assessing the tools and systems that are necessary to read, organize, and process all the incoming data
- Essential activities in this phase include structuring the business problem in the form of an analytics challenge and formulating the initial hypotheses (IHs) to test and start learning the data
- The subsequent phases are then based on achieving the goal that is drawn in this stage

Data Preparation and Processing

- This stage consists of everything that has anything to do with data
- In phase 2, the attention of experts moves from business requirements to information requirements
- The data preparation and processing step involve collecting, processing, and cleansing the accumulated data
- One of the essential parts of this phase is to make sure that the data you need is actually available to you for processing
- The earliest step of the data preparation phase is to collect valuable information and proceed with the **data analytics lifecycle** in a business ecosystem
- Data is collected using the below methods:
  - **Data Acquisition:** Accumulating information from external sources
  - **Data Entry:** Formulating recent data points using digital systems or manual data entry techniques within the enterprise
  - **Signal Reception:** Capturing information from digital devices, such as control systems and the Internet of Things

cleansing / pre-processing / data wrangling / data scrubing

1] Remove NA [missing] values
   - replace them with mean/mode
   - remove row containg NA
   → remove column containing NA

2] fix the data type

3] balance the data → upsampling, downsampling

4] scaling the data

5] conversion of qualitative to quantitative
   → label encoding, one hot encoding

6] shuffling data

7] Add columns required for processing

8] Drop unnecessary columns

Design a Model

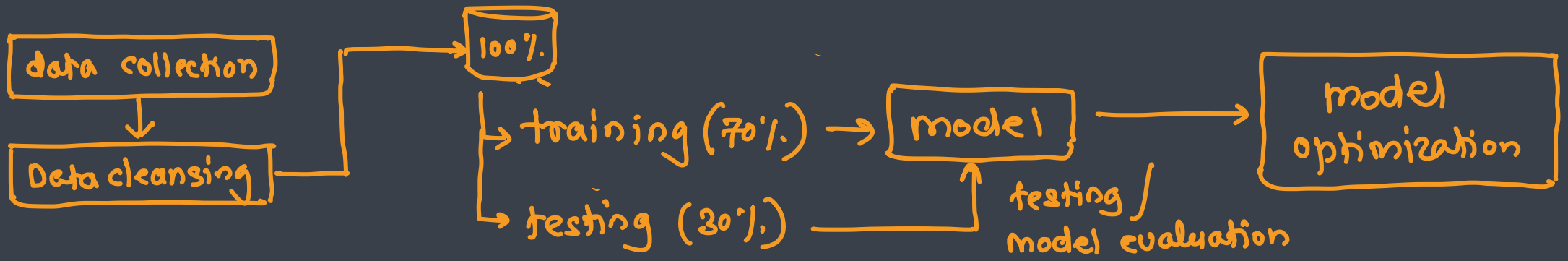
plan → data analysis
plan → variables
plan → language → python / R

- After mapping out your business goals and collecting a glut of data (structured, unstructured, or semi-structured), it is time to build a model that utilizes the data to achieve the goal
- **There are several techniques available to load data into the system and start studying it:**
  - ETL (Extract, Transform, and Load) transforms the data first using a set of business rules, before loading it into a sandbox
  - ELT (Extract, Load, and Transform) first loads raw data into the sandbox and then transform it
  - ETLT (Extract, Transform, Load, Transform) is a mixture; it has two transformation levels
- This step also includes the teamwork to determine the methods, techniques, and workflow to build the model in the subsequent phase
- The model's building initiates with identifying the relation between data points (→ correlation analysis) to select the key variables and eventually find a suitable model

→ set: dependent variable, independent variable(s)

Model Building

- This step of **data analytics architecture** comprises developing data sets for testing, training, and production purposes
- The data analytics experts meticulously build and operate the model that they had designed in the previous step
- They rely on tools and several techniques like decision trees, regression techniques and neural networks for building and executing the model
- The experts also perform a trial run of the model to observe if the model corresponds to the datasets

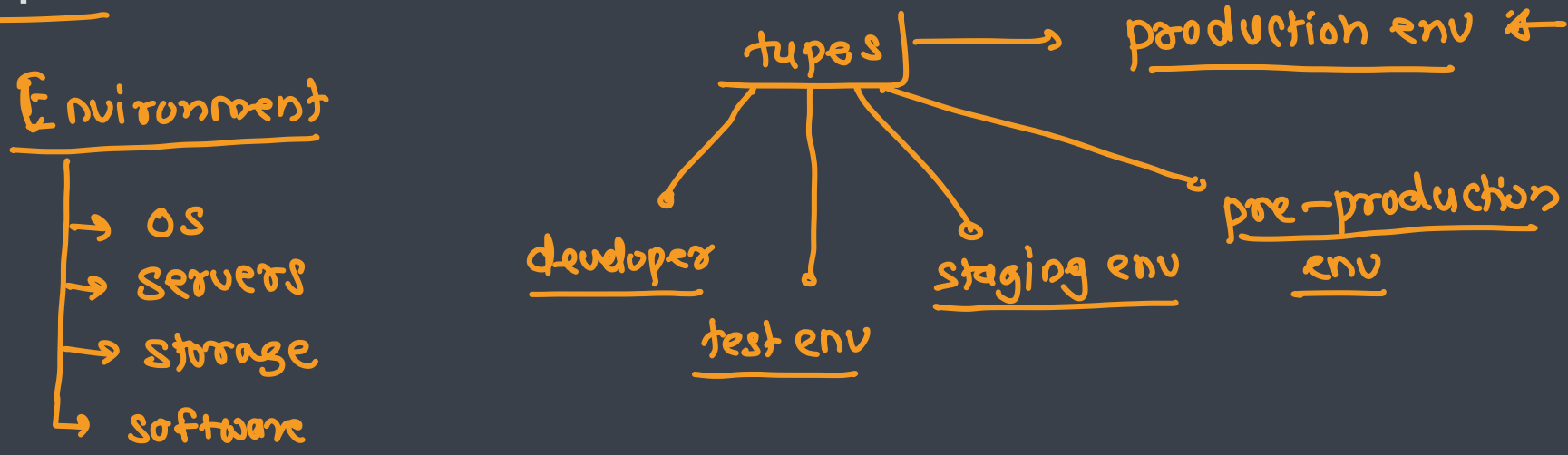


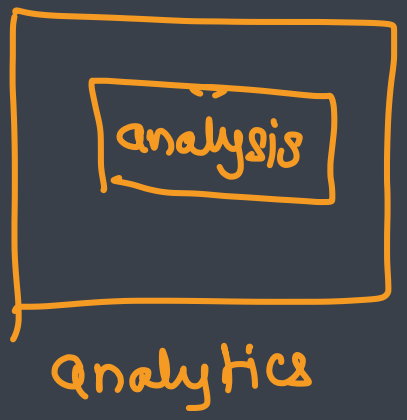
Result Communication and Publication

- Remember the goal you had set for your business in phase 1? Now is the time to check if those criteria are met by the tests you have run in the previous phase
- The communication step starts with a collaboration with major stakeholders to determine if the project results are a success or failure
- The project team is required to identify the key findings of the analysis, measure the business value associated with the result, and produce a narrative to summarise and convey the results to the stakeholders

Measuring of Effectiveness

- As your **data analytics lifecycle** draws to a conclusion, the final step is to provide a detailed report with key findings, coding, briefings, technical papers/ documents to the stakeholders
- Additionally, to measure the analysis's effectiveness, the data is moved to a live environment from the sandbox and monitored to observe if the results match the expected business goal
- If the findings are as per the objective, the reports and the results are finalized
- However, suppose the outcome deviates from the intent set out in phase 1 then you can move backward in the **data analytics lifecycle** to any of the previous phases to change your input and get a different output





Data Analysis

past data

What is it ?

understand

- **Data analysis** is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making
- The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis
- A simple example of Data analysis is whenever we take any decision in our day-to-day life is by thinking about what happened last time or what will happen by choosing that particular decision
- This is nothing but analyzing our past or future and making decisions based on it
- For that, we gather memories of our past or dreams of our future

Importance

- Data analysis can help businesses improve specific aspects about their products and services, as well as their overall brand image and customer experience → NLP → sentiment analysis
- Analyzed data reveals insights that tell you what your customers need and where you need to focus your efforts
- Instead of relying on intuition or experience, analyzing data provides solid evidence to support decisions
- Product teams, for example, often analyze customer feedback to understand how customers interact with their product, what they're frustrated with, and which new features they'd like to see. Then, they translate this insight into UX improvements, new features, and enhanced functionalities.
- Through data analysis, you can also detect the weaknesses and strengths of your competition, uncovering opportunities for improvement
- Analyze online reviews about your competition to answer questions like: what do customers love or hate about your competitors' products or services?

Data Analysis vs Data Analytics

Analytics	Analysis
Used in business to make the decisions from data driven → prediction	Used in business to analyze data and take some insights of it → patterns
Consists of data collection and inspect in general and it has one of more usage	Consists of defining a data, investigation, cleaning and transforming the data to give a meaningful outcome
Used to find anonymous patterns, customer preferences, market trends etc	Used to find meaning information necessary to understand the data

future predictions

past data

Data Analysis vs Reporting

Analysis	Reporting
Provides answers to any question or issue	Shows what had happened in the past to avoid inferences and help to get a feel for the data
Provides the information or the answer that is needed actually	Provides the data that is asked
A person is there for doing analysis and leading the complete analysis process	We can perform reporting using a tool and it generally does not involve any person in the analysis
Analysis is flexible	Reporting is inflexible

Data Analysis Techniques

Data Analysis Techniques

- Text Analysis
- Descriptive Analysis
- Inferential Analysis
- Diagnostic Analysis
- Predictive Analysis
- Prescriptive Analysis

Text Analysis

- Text analysis, also text analytics or text mining, uses machine learning with natural language processing (NLP) to organize unstructured text data so that it can be properly analyzed for valuable insights. Text analysis is a form of qualitative analysis that is concerned with more than just statistics and numerical values
- By transforming human language into machine-readable data, text analysis tools can sort text by topic, extract keywords, and read for emotion and intent. It tells us “What is happening” as specific, often subjective data. It offers more in-depth and targeted views into why something may be happening, or why something happened
- You can use text analysis to detect topics in customer feedback, for example, and understand which aspects of your brand are important to your customers
- Sentiment analysis is another approach to text analysis, used to analyze data and sort it as Positive, Negative, or Neutral to gain in-depth knowledge about how customers feel towards each aspect

Descriptive Analysis

- Descriptive data analysis provides the “What happened?” when analyzing quantitative data
- It is the most basic and most common form of data analysis concerned with describing, summarizing, and identifying patterns through calculations of existing data, like mean, median, mode, percentage, frequency, and range
- Descriptive analysis is usually the baseline from which other data analysis begins
- It is, no doubt, very useful for producing things like revenue reports and KPI dashboards
- However, as it is only concerned with statistical analysis and absolute numbers, it can’t provide the reason or motivation for why and how those numbers developed

Inferential Analysis

- Inferential analysis generalizes or hypothesizes about “What happened?” by comparing statistics from groups within an entire population: the population of a country, existing customer base, patients in a medical study, etc
- The most common methods for conducting inferential statistics are hypothesis tests and estimation theories.
- Inferential analysis is used widely in market research, to compare two variables in an attempt to reach a conclusion: money spent by female customers vs. male or among different age groups, for example. Or it can be used to survey a sample set of the population in an attempt to extrapolate information about the entire population. In this case it is necessary to properly calculate for a representative sample of the population.

Diagnostic Analysis

- Diagnostic analysis, also known as root cause analysis, aims to answer “Why did 'X' happen?”. It uses insights from statistical analysis to attempt to understand the cause or reason behind statistics, by identifying patterns or deviations within the data to answer for why.
- Diagnostic analysis can be helpful to understand customer behavior, to find out which marketing campaigns actually increase sales, for example. Or let’s say you notice a sudden decrease in customer complaints: Why did this happen?
- Perhaps you fired a certain employee or hired new ones. Maybe you have a new online interface or added a particular product feature. Diagnostic analysis can help calculate the correlation between these possible causes and existing data points.

Predictive Analysis

- Predictive analysis uses known data to postulate about future events. It is concerned with “What is likely to happen.” Used in sales analysis, it often combines demographic data and purchase data with other data points to predict the actions of customers.
- For example, as the demographics of a certain area change, this will affect the ability of certain businesses to exist there. Or as the salary of a certain customer increases, theoretically, they will be able to buy more of your products.
- There is often a lot of extrapolative guesswork involved in predictive analysis, but the more data points you have on a given demographic or individual customer, the more accurate the prediction is likely to be.

Prescriptive Analysis

- Prescriptive analysis is the most advanced form of analysis, as it combines all of your data and analytics, then outputs a model prescription: What action to take. Prescriptive analysis works to analyze multiple scenarios, predict the outcome of each, and decide which is the best course of action based on the findings
- Artificial intelligence is an example of prescriptive analysis that's at the cutting edge of data analysis. AI allows for prescriptive analysis that can ingest and break down massive amounts of data and effectively teach itself how to use the information and make its own informed decisions
- AI used to require huge computing power, making it difficult for businesses to implement. However, with the rise of AI data analysis software, there are many exciting options available

Steps

How to Analyse your Data?

- **Define Your Goals**
  - Setting clear objectives is key and will help determine the type of data that you'll need to collect and analyze

- **Collect Your Data**
  - Data is everywhere, and you'll want to bring it all into one place ready for analysis
  - Whether you're collecting quantitative or qualitative data, Excel is a great platform for storing your data, or you could connect data sources directly to your analysis tools via APIs and integrations

- **Clean Your Data**
  - It's likely that unstructured data will need to be cleaned before analyzing it to gain more accurate results
  - Get rid of the noise, like special characters, punctuation marks, stopwords (and, too, she, they), HTML tags, duplicates, etc.

How to Analyse your Data?

- **Integrate Data Analysis Tools**
  - Most tools can easily be integrated via APIs and one-click integrations
  - Once connected, your data can run freely through data analysis tools

- **Analyse Your Data**
  - Now that you've connected data analysis tools, you'll need to choose the analysis type you want to perform

- **Visualize Your Data**
  - Dashboards are a great way to aggregate your data, and make it easy to spot trends and patterns
  - Some data analysis tools have in-built dashboards or you can connect to your existing BI tools

- **Draw Conclusions**
  - Gain actionable insights and make data-based decisions by digging into your data from every angle

Tools

Data Analysis Tools

- **MonkeyLearn**
  - No-code machine learning platform that provides a full suite of text analysis tools and a robust API
  - Easily build custom machine learning models in a point and click interface

- **KNIME**
  - Open-source platform for building advanced machine learning solutions, and visualizing data

- **RapidMiner**
  - For data analytics teams that want to tackle challenging tasks and handle large amounts of data

- **Microsoft Excel**
  - Filter, organize, and visualize quantitative data
  - The perfect tool for performing simple data analysis. Explore common functions and formulas for data analysis in Excel

Data Analysis Tools

- **Tableau**
  - A powerful analytics and data visualization platform
  - Connect all your data and create interactive dashboards that update in real-time
- **R**
  - A free software environment for statistical computing and graphics
- **Python**
  - The preferred programming language for machine learning
  - Use it to build data analysis solutions for various use cases