OE-Data Science 6OE371

COs

CO	Course Outcome Statement/s	Bloom's Taxonomy Level	Bloom's Taxonomy Description
CO1	acquaint core concepts and technologies in Data Science.	II	Understanding
CO2	illustrate various data collection and preprocessing techniques.	III	Applying
CO3	use visualization techniques to show relationship within datasets.	III	Applying
CO4	analyse possible relationship within large datasets and identify suitable prediction technique to solve real-world problems.	IV	Analyzing

Academic and Examination RR

Reference:

Point 04.04 Attendance (page no 21/47)

1. All students should attend the classes and expected to be regular (100% attendance) for all the courses. The attendance records of students should be maintained in WCE Moodle by the course teacher. The students should check their attendance in WCE Moodle regularly and should contact the respective course teacher for any discrepancy/grievance.
2. A maximum of 25% exemption in the attendance may be permitted for the approved leave of absence from class teacher/HoD for participating in co-curricular/extra-curricular activities/medical emergencies/reasons beyond the control of students. Students with more than 75% attendance shall not be imposed with any grade penalty.
3. The students with less than 75% attendance in theory course/s shall be liable for grade penalty as:

a. Students having attendance greater than or equal to 70% but less than 75% be allowed to appear for ESE in that course with maximum grade of BB.

b. Students having attendance greater than or equal to 60% but less than 70% be allowed to appear for ESE in that course with maximum grade of BC.

c. Students having attendance greater than or equal to 50% but less than 60% be allowed to appear for ESE in that course with maximum grade of CC.

d. Students having attendance less than 50% shall be awarded with XX grade in that course.

Syllabus

Module	Module Contents	Hours
I	**Module 1: Introduction to core concepts and technologies** Introduction, Terminology, data science process, data science toolkit, Types of data, Example applications	4
II	**Module 2 Data Collection and Management** Introduction, Sources of data, Data collection, Exploring and fixing data, Data storage and management, Using multiple data sources.	7
III	**Module 3 Data Pre-processing** Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.	8
IV	**Module 4 Data Visualization** Introduction, Types of data visualization, Data for visualization: Data types, Data encodings, Retinal variables, Mapping variables to encodings, visual encodings.	6
V	**Module 5 Data Analysis** Introduction, Terminology and concepts, Introduction to statistics, Central tendencies and distributions, Variance, Distribution properties and arithmetic, Samples/CLT, Correlation, Linear Regression, Least Squares, Residuals, Regression Inference, classification, classifiers.	8
VI	**Module 6 Recent trends** Recent trends in various data collection and analysis techniques, various visualization techniques, Case Study, application development methods used in data science.	6

Introduction to core concepts and technologies

- What is Data Science?
- Why Data Science?
- Components of DS
- Applications of Data Science.
- Data Science life cycle
- Data science toolkit
- Types of data

What is Data Science?

- Deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithms.
- Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization’s data.

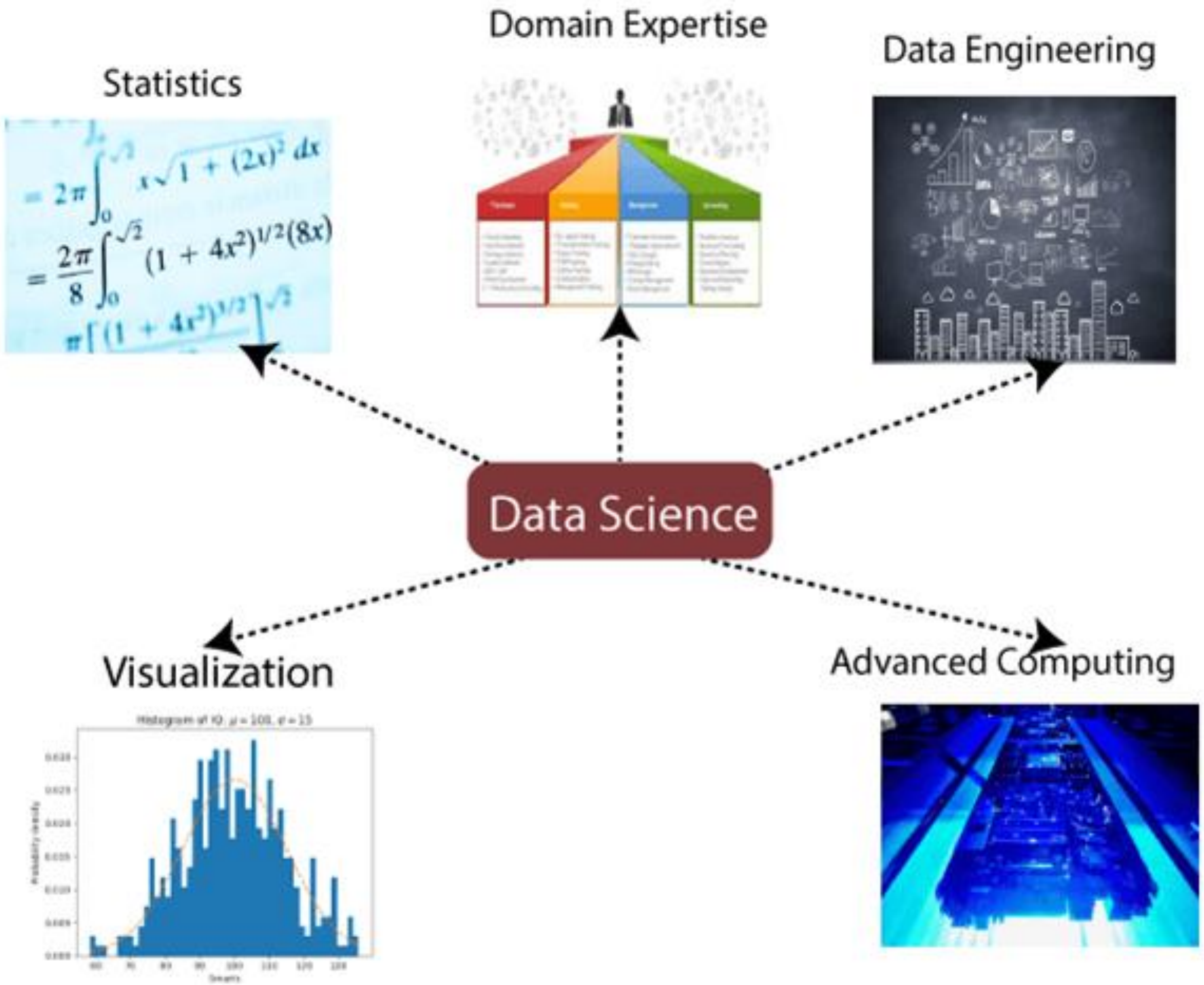
What is Data Science?

- Data science is the study of data to extract meaningful insights for business.
- Vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

Why Data Science?

- The purpose of data science is to find patterns.
- Data science enables businesses to interact with their customers.
- Products and businesses can better connect with their customers when they use.
- Industries can quickly examine their problems and successfully address them using data science.
- Depending on how data is used, can determine whether a product succeeds or fails.
- Giving management and officials the ability to foster new ideas.
- An improved user experience.

Components of DS



Domain Expertise

Statistics

Data Engineering

Data Science

Visualization

Advanced Computing

Applications of Data Science

- **1. Healthcare**
- **Predictive Analytics**: Predicting disease outbreaks, patient admission rates, and readmission probabilities.
- **Medical Imaging**: Enhancing image recognition for diagnosing diseases through techniques like deep learning.
- **Genomics**: Analyzing genomic data to understand genetic disorders and personalize medicine.

- **2. Finance**
- **Risk Management**: Assessing and predicting financial risks, fraud detection, and credit scoring.
- **Algorithmic Trading**: Using machine learning models to make trading decisions and optimize portfolios.
- **Customer Analytics**: Understanding customer behavior, segmentation, and improving customer service.

Applications of Data Science

- **3. Environmental Science**
- **Climate Modeling**: Predicting climate change patterns and assessing the impact of various factors on the environment.
- **Conservation**: Analyzing data from sensors and cameras to monitor wildlife and manage conservation efforts.
- **Renewable Energy**: Optimizing the production and distribution of renewable energy sources.

- **4. Social Sciences**
- **Behavioral Analysis**: Studying social media and other data sources to understand human behavior and social trends.
- **Policy Making**: Using data to inform public policy decisions and evaluate the impact of policies.
- **Elections**: Analyzing voter behavior, predicting election outcomes, and detecting election fraud.

Applications of Data Science

- **5. Education**
- **Personalized Learning**: Creating adaptive learning systems that cater to individual student needs.
- **Student Performance**: Predicting student outcomes and identifying at-risk students for early intervention.
- **Curriculum Development**: Using data to optimize course content and teaching methods.

- **6. Agriculture**
- **Precision Farming**: Using data from sensors and drones to optimize crop yields and manage resources efficiently.
- **Supply Chain Management**: Analyzing data to improve the efficiency and sustainability of the agricultural supply chain.
- **Weather Prediction**: Forecasting weather conditions to help farmers make informed decisions.

Applications of Data Science

- **7. Sports**
- **Performance Analysis**: Using data to analyze and improve athlete performance and training regimens.
- **Injury Prevention**: Predicting and preventing injuries through data analysis.
- **Fan Engagement**: Understanding fan behavior and preferences to enhance the fan experience.

- **8. Marketing and Retail**
- **Customer Insights**: Analyzing customer data to understand preferences and improve marketing strategies.
- **Recommendation Systems**: Building recommendation engines to suggest products to customers.
- **Supply Chain Optimization**: Improving inventory management and logistics through data analysis.

Applications of Data Science

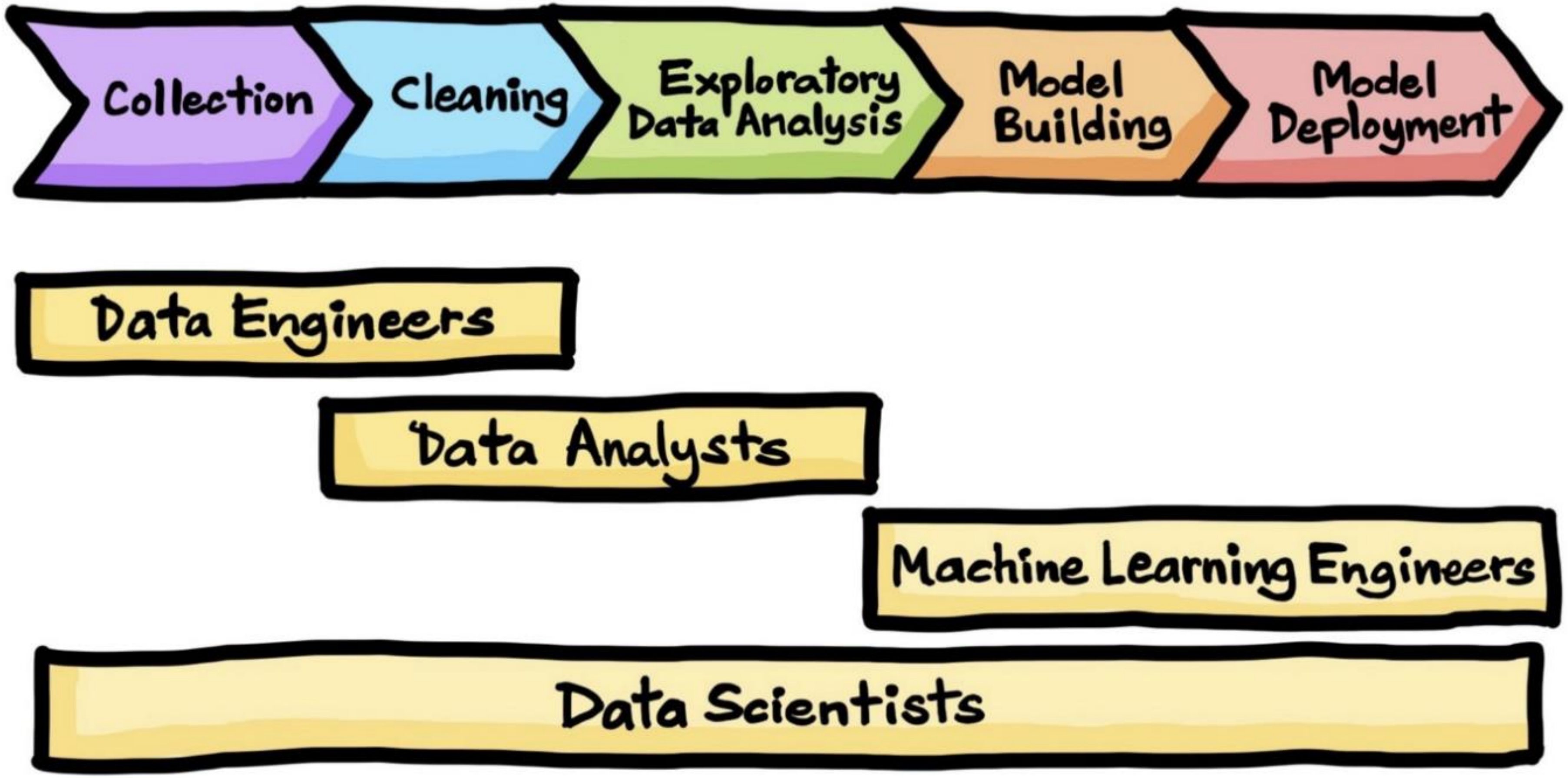
- **9. Transportation**
- **Traffic Management**: Analyzing traffic data to optimize flow and reduce congestion.
- **Autonomous Vehicles**: Developing self-driving cars using machine learning and data from sensors.
- **Logistics**: Enhancing route planning and delivery efficiency for logistics companies.

- **10. Urban Planning**
- **Smart Cities**: Using data to manage urban infrastructure, optimize resource use, and improve quality of life.
- **Public Safety**: Analyzing crime data to improve law enforcement and public safety strategies.
- **Transportation Planning**: Designing and optimizing public transportation systems.

Applications of Data Science

- **11. Media and Entertainment**
- **Content Recommendation**: Using data to recommend movies, music, and other content to users.
- **Audience Analysis**: Understanding viewer behavior and preferences to tailor content.
- **Production Optimization**: Analyzing data to streamline production processes and reduce costs.

Data Science Life Cycle / process



Example

- **Simple Data Science Problem: Predicting Air Quality Index (AQI)**
- **Background**
- Air quality is a significant environmental and public health concern. Predicting the Air Quality Index (AQI) can help authorities and the public take timely actions to reduce exposure to harmful pollutants.
- **Problem Statement**
- Develop a predictive model to forecast the AQI for the next day using historical air quality data and meteorological data.
- **Data Sources**
- **Historical AQI Data**: Daily AQI values for various locations.
- **Meteorological Data**: Temperature, humidity, wind speed, wind direction, and precipitation.
- **Pollutant Data**: Concentrations of pollutants such as PM2.5, PM10, NO2, SO2, CO, and O3.
- **Geographical Data**: Location coordinates, altitude, and proximity to pollution sources like factories and highways.

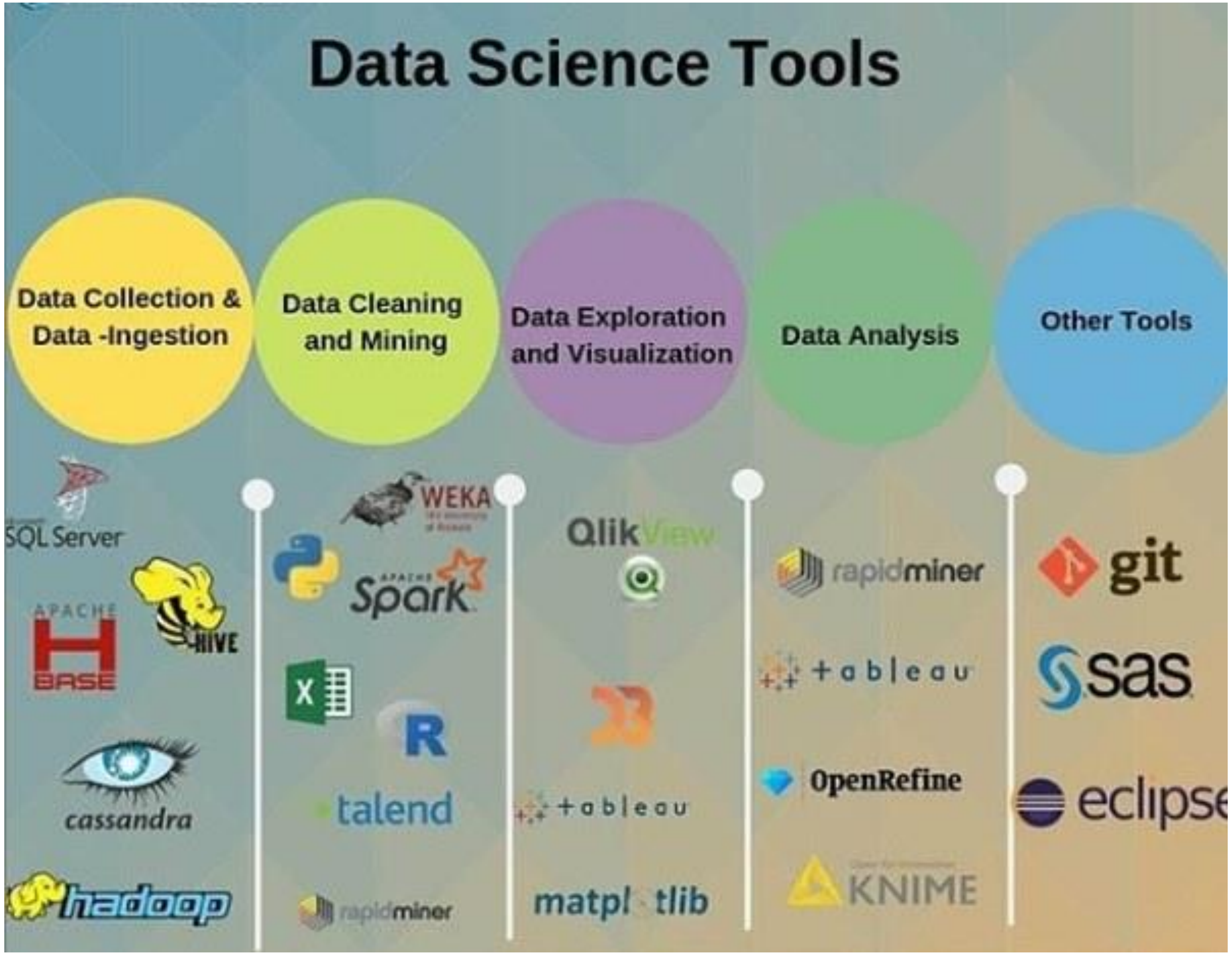
Example

- **Steps to Solve the Problem**
- **Data Collection and Integration**
  - Gather historical AQI data and meteorological data from government agencies, environmental monitoring stations, and weather services.
  - Integrate the data into a unified format for analysis.
- **Data Preprocessing**
  - Handle missing values through imputation or exclusion.
  - Normalize or standardize numerical features.
  - Encode categorical variables using techniques like one-hot encoding or label encoding.
  - Create time-based features such as the day of the week, month, and season.
- **Feature Engineering**
  - **Meteorological Features**: Daily average, maximum, and minimum temperature, humidity, wind speed, and wind direction.
  - **Pollutant Features**: Daily average concentrations of PM2.5, PM10, NO2, SO2, CO, and O3.
  - **Temporal Features**: Day of the week, month, season, and holidays (which might affect pollution levels).
  - **Geographical Features**: Location coordinates, altitude, proximity to pollution sources.
- **Model Selection**
  - Experiment with various models such as linear regression, decision trees, random forests, gradient boosting machines, and neural networks.
  - Use cross-validation to evaluate model performance.

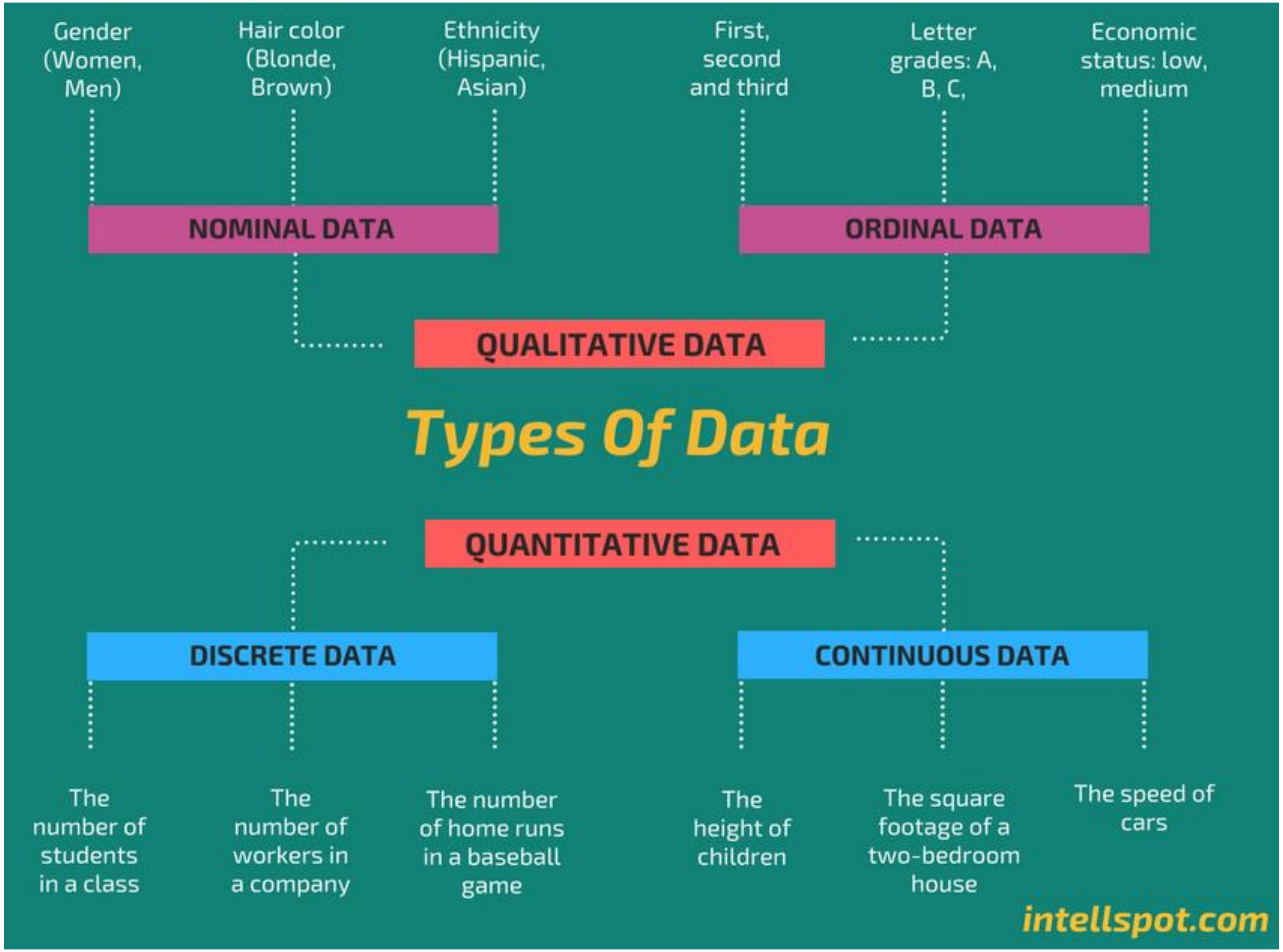
Example

- **Model Training**
  - Split data into training and test sets.
  - Train multiple models and fine-tune hyperparameters using techniques like grid search or random search.
  - Use evaluation metrics like mean absolute error (MAE), mean squared error (MSE), and R-squared to assess model performance.
- **Model Evaluation**
  - Evaluate models on the test set to check for overfitting.
  - Perform additional validation on a separate validation set if available.
  - Interpret model outputs and feature importance to understand the key predictors of AQI.
- **Deployment**
  - Deploy the model to provide daily AQI forecasts for specific locations.
  - Integrate the model with a web or mobile application for real-time predictions.
  - Provide actionable insights and recommendations to users, such as reducing outdoor activities on high AQI days.
- **Monitoring and Maintenance**
  - Continuously monitor model performance and update it with new data.
  - Retrain the model periodically to adapt to changing environmental conditions and pollution patterns.
  - Collect feedback from users and adjust the model to improve accuracy and usability.

Data science toolkit



Types of Data


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Qualitative data/Categorical data

- Qualitative or Categorical Data describes the object under consideration using a finite set of discrete classes.
- It means that this type of data can’t be counted or measured easily using numbers and therefore divided into categories.
- The gender of a person (male, female, or others) is a good example of this data type.
- These are usually extracted from audio, images, or text medium.
- Another example can be of a smartphone brand that provides information about the current rating, the color of the phone, category of the phone, and so on.
- All this information can be categorized as Qualitative data.

Nominal

- These are the set of values that don’t possess a natural ordering.
- e.g The color of a smartphone as we can’t compare one color with others.
- It is not possible to state that ‘Red’ is greater than ‘Blue’.
- The gender of a person where we can’t differentiate between male, female, or others.
- Nominal data types in statistics are not quantifiable and cannot be measured through numerical units.
- Nominal types of statistical data are valuable while conducting qualitative research as it extends freedom of opinion to subjects.

Ordinal

- These types of values have a natural ordering while maintaining their class of values.
- e.g If we consider the size of a clothing brand then we can easily sort them according to their name tag in the order of small < medium < large.
- The grading system while marking candidates in a test can also be considered as an ordinal data type where A+ is definitely better than B grade.
- These categories help us deciding which encoding strategy can be applied to which type of data.
- Data encoding for Qualitative data is important because machine learning models can't handle these values directly and needed to be converted to numerical types as the models are mathematical in nature.
- For nominal data type where there is no comparison among the categories, one-hot encoding can be applied which is similar to binary coding considering there are in less number and for the ordinal data type, label encoding can be applied which is a form of integer encoding.

Quantitative Data Type

- This data type tries to quantify things and it does by considering numerical values that make it countable in nature.
- The price of a smartphone, discount offered, number of ratings on a product, the frequency of processor of a smartphone, or ram of that particular phone, all these things fall under the category of Quantitative data types.
- The key thing is that there can be an infinite number of values a feature can take.
- For instance, the price of a smartphone can vary from x amount to any value and it can be further broken down based on fractional values.

- **Interval-scaled attributes**
- Type of numerical attribute where the difference between two values is meaningful.
- The term "interval scale" refers to an ordered series of numbers where the difference between the values is consistent, but the zero point is not truly meaningful.
- Interval variables can be added and subtracted, providing meaningful results.
- For example, temperature, as measured in degrees Celsius or Fahrenheit, is an interval variable.
- If it is 20 degrees today and 30 degrees tomorrow, it is correct to say that tomorrow is 10 degrees hotter than today.

- **Ratio-scaled attribute**
- Ratio variables are a type of numerical attribute where the difference between two values is meaningful and there is a true "zero" point, which denotes the absence of the quantity.
- This zero point allows for the comparison of values through multiplication or division, unlike interval-scaled attributes.
- Examples of ratio variables include age, salary, and height.
- In these examples, a value of 0 signifies the absence of the quantity: 0 years old means no age or not born yet, a salary of $0 means no income, and a height of 0 cm signifies no height.
- If person A is 20 years old and person B is 40 years old, it's correct to say that person B is twice as old as person A.

Discrete

- Discrete data is a type of numerical data that only takes specific or 'discrete' values and cannot be meaningfully subdivided into smaller increments. This often corresponds to items or events that are countable.
- Examples of discrete data include:
- The number of pets a person has. You can have 2 dogs or 3 dogs, but it doesn't make sense to have 2.7 dogs.
- The number of cars in a parking lot. You can have 10, 20, or 30 cars, but not 22.5 cars.
- The number of students in a class. You can't have a fraction of a student.

Continuous

- Continuous data is a type of numerical data that can take on any value within a certain range.
- Continuous data can be meaningfully subdivided into finer and finer increments, depending on the precision of the measurement system.
- Examples of continuous data include:
- The height of people. You can be 170.18 cm or 170.19 cm tall, or any height in between.
- The time it takes to run a marathon. It could be 3 hours, 45 minutes, 30.2 seconds, or 3 hours, 45 minutes, 30.3 seconds, or any time in between.
- The weight of a bag of apples. It could be 1.5 kg, 1.51 kg, 1.515 kg, and so on, depending on how precise your scale is.

Exercise

Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.

- A person’s age.
- A person’s race.
- Age groupings (baby, toddler, adolescent, teenager, adult, elderly).
- Clothing brand.
- A person’s IQ score.
- Temperature in degrees Celsius.
- The amount of mercury in a tuna fish.

Exercise

- **Select the measurement scale Nominal, Ordinal, Interval or Ratio for each scenario.**
- Temperature in degrees Kelvin.
- Eye color.
- Year in school (freshman, sophomore, junior, senior).
- The weight of a hummingbird.
- The height of a building.
- The amount of iron in a person’s blood.
- A person’s gender.
- A person’s race.

Exercise

- **State which type of variable each is, qualitative or quantitative?**
- A person’s age.
- A person’s gender.
- The amount of mercury in a tuna fish.
- The weight of an elephant.
- Temperature in degrees Fahrenheit.
- **State which type of variable each is, qualitative or quantitative?**
- The height of a giraffe.
- A person’s race.
- Hair color.
- A person’s ethnicity.
- Year in school (freshman, sophomore, junior, senior).

Exercise

- **State whether the variable is discrete or continuous.**
- A person’s weight.
- The height of a building.
- A person’s age.
- The number of floors of a skyscraper.
- The number of clothing items available for purchase.
- **State whether the variable is discrete or continuous.**
- Temperature in degrees Celsius.
- The number of cars for sale at a car dealership.
- The time it takes to run a marathon.
- The amount of mercury in a tuna fish.
- The weight of a hummingbird.

Real life applications of data science

- **PERSONALIZING TREATMENT PLANS**
- Oncora's software uses machine learning to create personalized recommendations for current cancer patients based on data from past ones. Healthcare facilities using the company's platform include UT Health San Antonio and Scripps Health. Their radiology team collaborated with Oncora data scientists to mine 15 years' worth of data on diagnoses, treatment plans, outcomes and side effects from more than 50,000 cancer records. Based on this data, Oncora's algorithm learned to suggest personalized chemotherapy and radiation regimens.

Real life applications of data science

- **OPTIMIZING FOOD DELIVERY**
- The data scientists at UberEats have a fairly simple goal: getting hot food delivered quickly. Making that happen across the country though, takes machine learning, advanced statistical modeling and staff meteorologists. In order to optimize the full delivery process, the team has to predict how every possible variable — from storms to holiday rushes — will impact traffic and cooking time.

Real life applications of data science

- **TRACKING PHYSICAL DATA FOR ATHLETES**
- WHOOP makes wearable devices that track athletes' physical data like resting heart rate, sleep cycle and respiratory rate. The goal is to help athletes understand when to push their training and when to rest — and to make sure they're taking the necessary steps to get the most out of their body. Professional athletes like Olympic sprinter Gabby Thomas, Olympic golfer Nelly Korda and PGA golfer Nick Watney are among the WHOOPS' users, according to the company's website.

Real life applications of data science

- **SUGGESTING FRIENDS ON FACEBOOK**
- Meta’s Facebook platform, of course, uses data science in various ways, but one of its buzzier data-driven features is the “People You May Know” sidebar, which appears on the social network’s home screen. Often creepily prescient, it’s based on a user’s friend list, the people they’ve been tagged with in photos and where they’ve worked and gone to school. It’s also based on “really good math,” according to the *Washington Post* — specifically, a type of data science known as network science, which essentially forecasts the growth of a user’s social network based on the growth of similar users’ networks.