Data Science Overview

Premanand S

May 3, 2024


DISCLAIMER

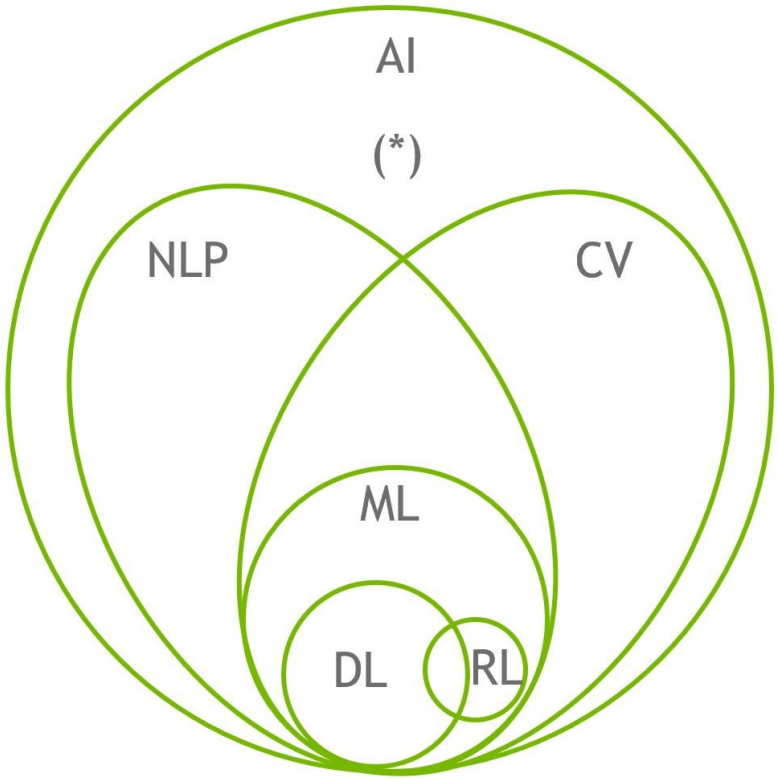
What is Technology & Why we need it?

Popular technologies, we came across

- Artificial Intelligence and Machine Learning
- Internet of Things (IoT)
- Blockchain
- 5G Technology
- Cybersecurity
- Quantum Computing
- Renewable Energy and Sustainability Technologies and many more...

Yesterday is HISTORY,
Tomorrow is MYSTERY,
Today is PRESENT!

Why do we need to know about these technologies?



AI = Artificial Intelligence
NLP=Natural Language Processing
CV=Computer Vision
ML=Machine Learning
DL=Deep Learning
RL=Reinforcement Learning

(*)=We would have more ellipses there (similar to NLP or CV) representing Robotics, Expert Systems, Speech, and Planning, Scheduling & Optimization systems. But it would look very messy. So, go ahead and imagine they are there too.

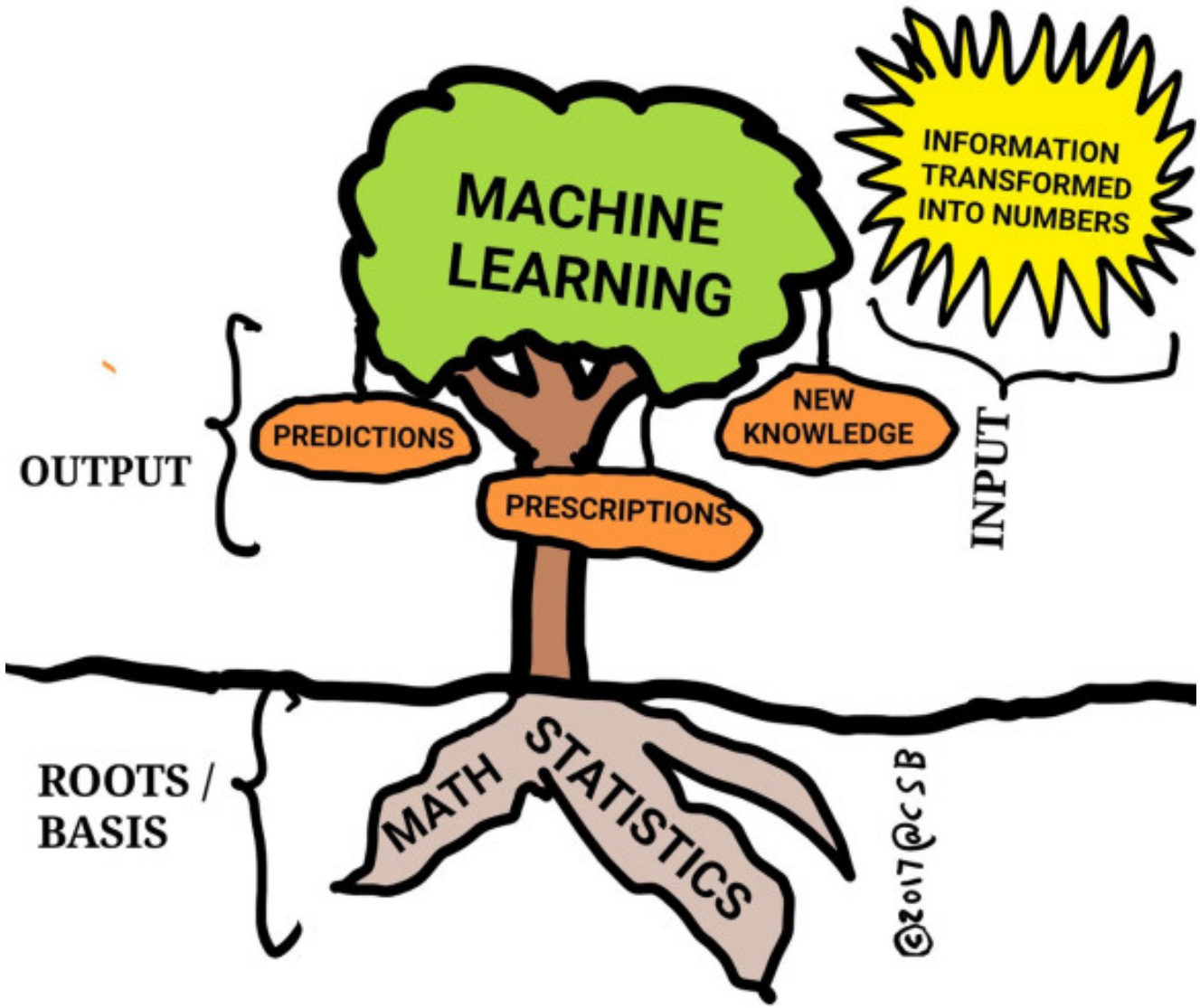
Data

- **Data** - Information
- **Source** - Any smart gadgets, Smart Phone (Apps), Sensors & many more...
- **Formats** - Numericals, Images, Audios, Videos & many more...

Every 1 minute

Any google users?

Statistics Vs Machine Learning



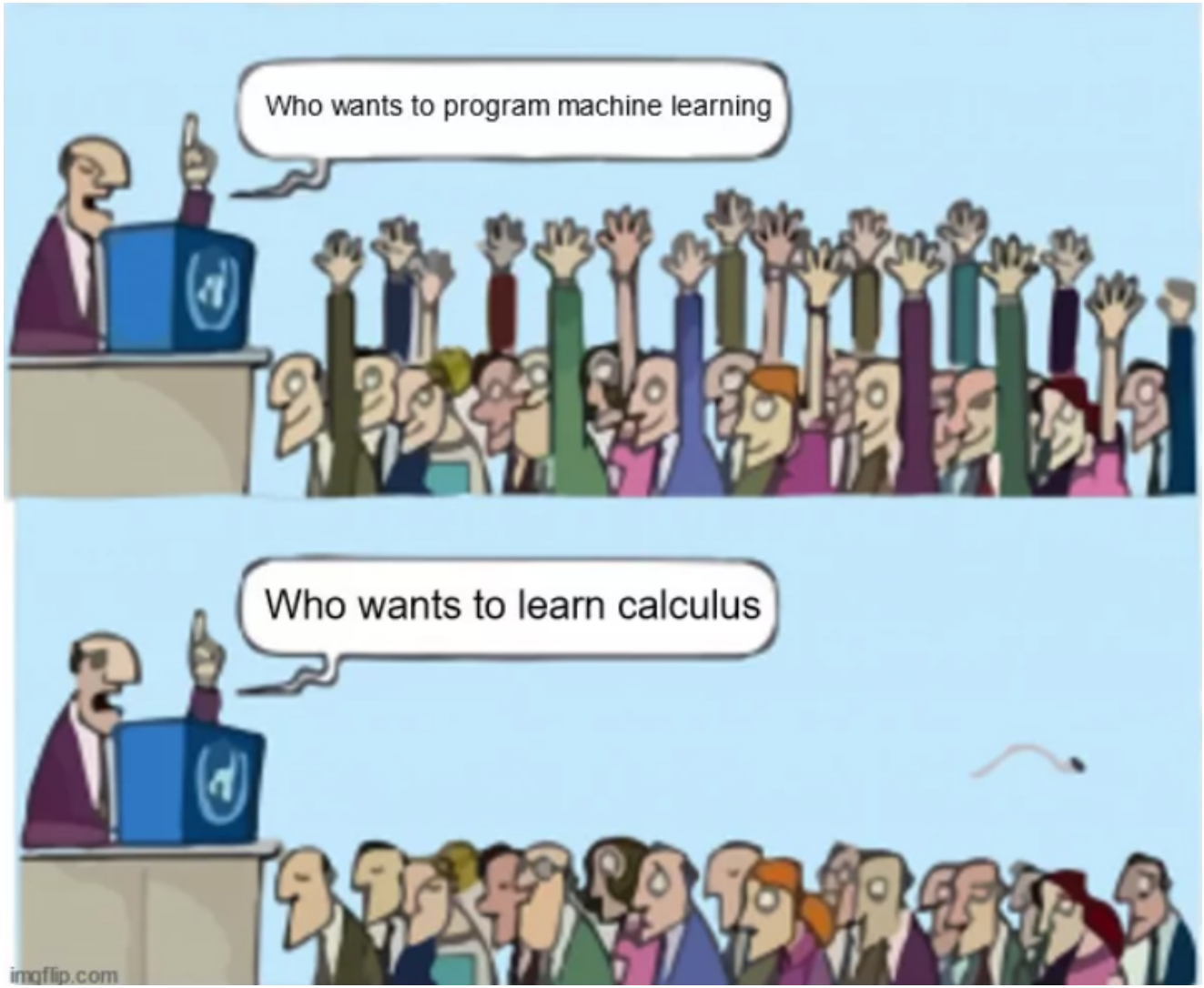
Statistics Vs Machine Learning

- **Statistics:** Problem - Data
- **Machine Learning:** Data - Solution

Why Statistics is needed?

- What features are important in raw data?
- What features can make a better model?
- How should we measure the performance of that model?
- What are already known outcomes and what we can achieve more?
- How can we fine-tune the model to make it more efficient?

Calculus Vs Machine Learning



Why Calculus is needed?

- When building a Machine Learning Model to train a robot/system to recognize whether a fruit is an apple or not, the model needs to learn from its mistakes to make better decisions.
- Local/Global minima
- Local/Global maxima
- Gradient Descent / Stochastic Gradient Descent
- Backpropagation
- Vanishing Gradient
- Exploring Gradient

Machine Learning - MEME (Today's Scenario)

Interviewer: What's your biggest strength?

Me: I'm an expert in machine learning.

Interviewer: What's 9 + 10?

Me: Its 3.

Interviewer: Not even close. It's 19.

Me: It's 16.

Interviewer: Wrong. Its still 19.

Me: It's 18.

Interviewer: No, it's 19.

Me: it's 19.

Interviewer: You're hired

Machine Learning - MEME (Today's Scenario)

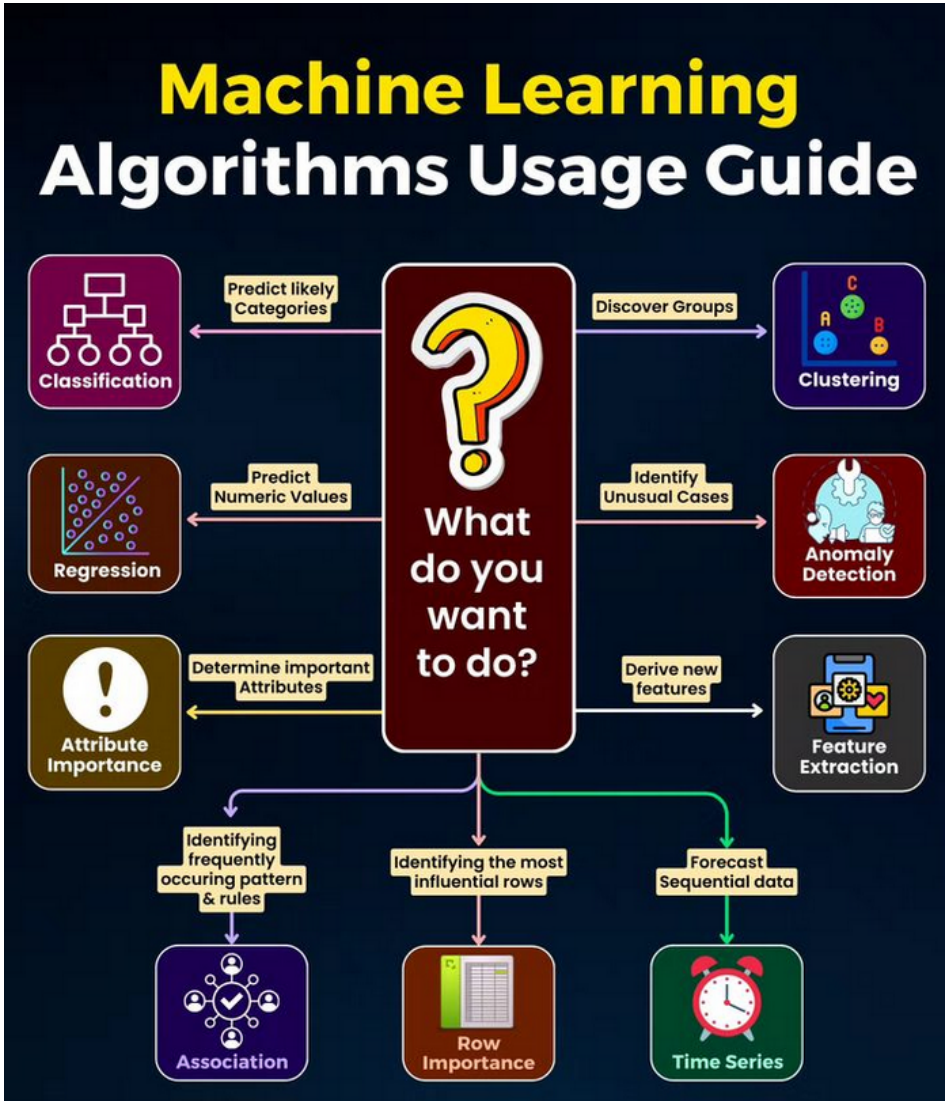
Albert Einstein: Insanity Is Doing the Same Thing Over and Over Again and Expecting Different Results

Machine learning:

Machine Learning - Actually

- Machine Learning is the field of study that gives computers, the ability to learn without being explicitly programmed - Arthur Samuel, 1959
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E - Tom Mitchell, 1997
- A baby learns to crawl, walk and then run. We are in the crawling stage when it comes to applying machine learning. - Dave Waters
- Mathematical toolbox used to solve the problems with data

What Machine Learning will do?



Machine Learning - Applications

- Facial Recognition and How it works in detail ?
- Speech to Text (Speech Recognition)
- Robot - SPOT, How it dance? - Reinforcement Learning
- Amazon, Flipkart - Recommended System & many more...

Machine Learning - Types

- Supervised Machine Learning - Train Me (Teacher-Student)
  - Classification
  - Regression
- Unsupervised Machine Learning - I am Self-Sufficient in learning
  - Clustering
  - Dimensionality Reduction
  - Association Rule
  - Anomaly Detection
- Reinforcement Learning - My life My rule
- Semi-Supervised Learning

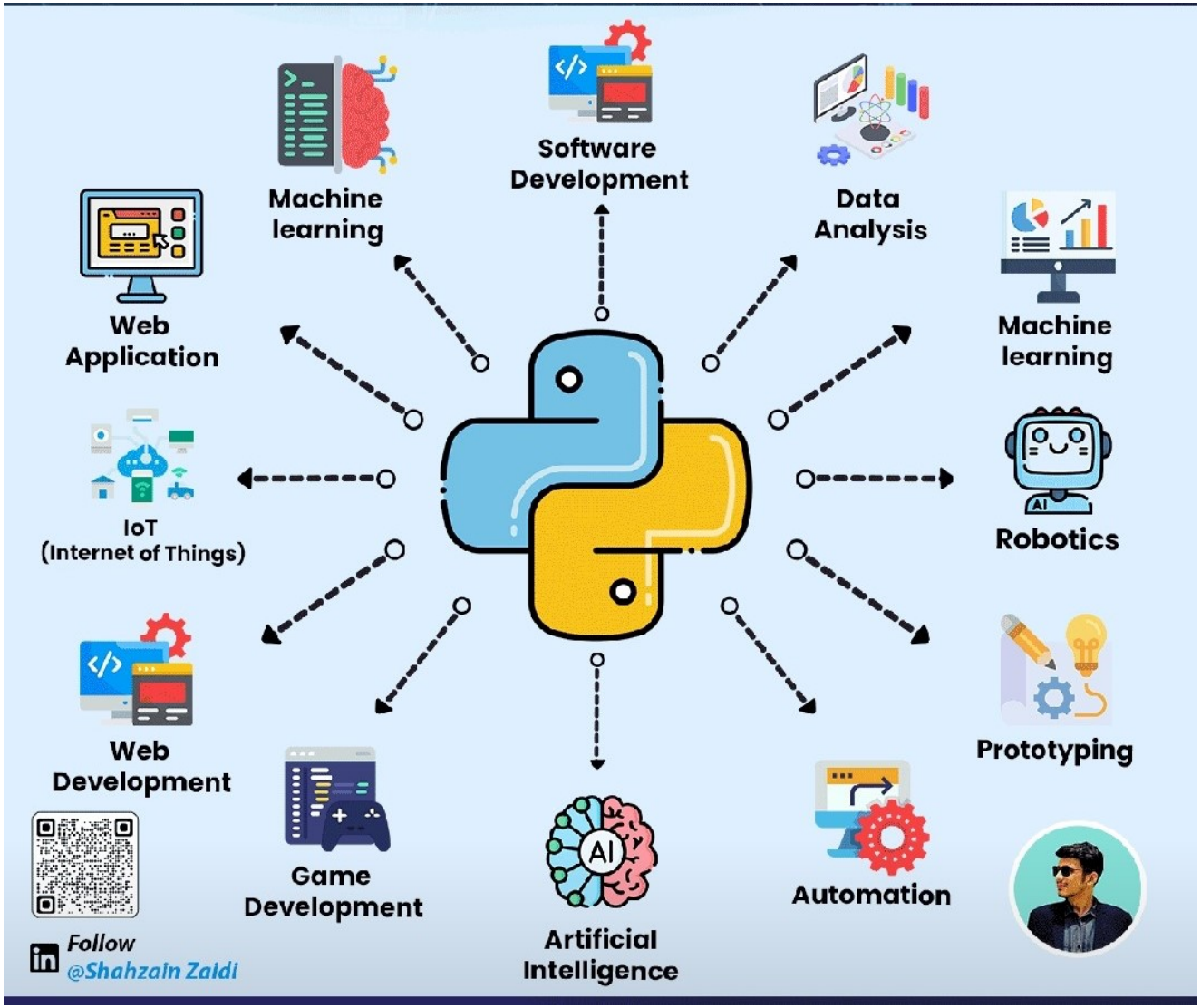
Machine Learning - Languages used

- Python – Wide range, data collection, modelling & Viz
- R - Statistics
- Julia - Numerical analysis & Scientific computing
- MATLAB - Mathematics & Statistics, User Interface, TB
- Java - Write once, Run anywhere, data analysis & mining
- Scala - Large amount of data, Java Bytecode, Java VM
- SQL - Structured data, Query database
- C / C++ - Codebase, Compiles quick
- JavaScript – rich interactive webpage, Viz
- WEKA – Data mining task, ML
- SAS – Statistical Data Analysis, Retrieve, Report, Analyze
- Excel- Non-Programmers, Business Analytics

Some front runners

Python	R	Julia
General Purpose	Statistical Analysis	Scientific Computing
Good	Good	Speed & Performance
Huge Community	Huge Community	Small Community
200k Libraries	15k Libraries	3k Libraries
In Billions	In Billions	13M downloads
Ijulia	-	Compile Just in time
Jupyter, Pycharm	R Studio	Juno IDE

Why Python?


Machine learning
Software Development
Data Analysis
Web Application
Machine learning
IoT (Internet of Things)
Robotics
Web Development
Prototyping
Game Development
Automation
AI
Artificial Intelligence
Follow @Shahzain Zaidi

Data Science - different IDE

- Anaconda Navigator
- Google Colab
- Visual Studio Code
- Thonny & many more...

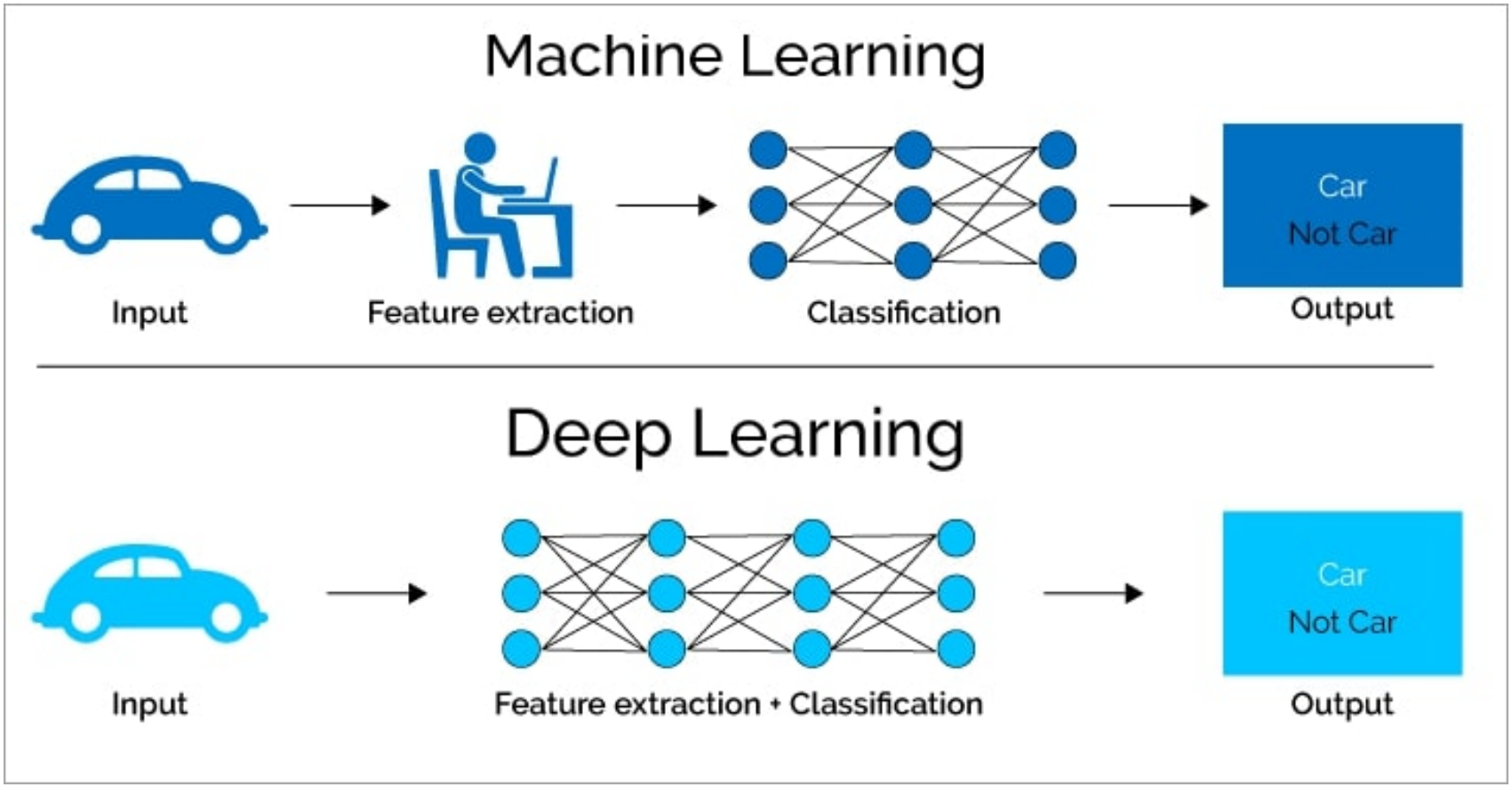
Data Science - Important Libraries

- Numpy, Pandas - Arrays
- Scikit-learn - Preprocessing, Modeling, Metrics
- Matplotlib, Seaborn - Visualization
- Tensorflow, Keras - Deep Learning

Need more clarity for ML Technology ?

I always believe, if any concept can be conveyed through animation, it won't forget easily.
This is the link: Machine Learning @Simplilearn.

Difference between Machine Learning and Deep Learning



Deep Learning

- Deep learning is a subset of machine learning.
- DL involves algorithms inspired by the structure and function of the human brain's neural networks.
- These algorithms attempt to learn layered representations of data, also known as neural networks, through a hierarchical structure of multiple layers.
- Each layer processes information in increasingly abstract ways, allowing the system to learn complex patterns and relationships within the data.

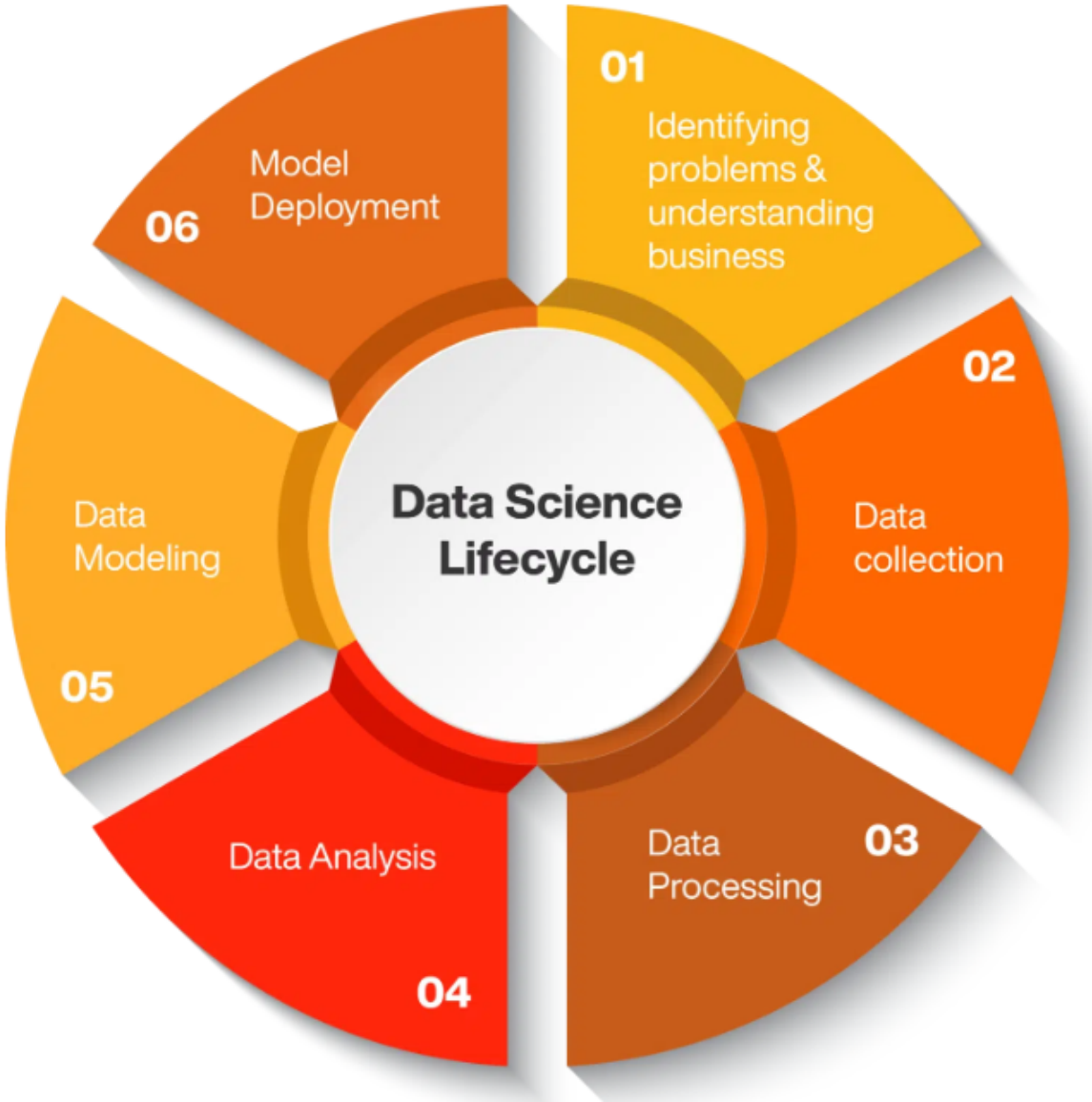
Use case of Deep Learning

- CNN - Image Classification, Object Detection, Image Segmentation
- RNN - Sequential data like time series data, text data, and speech data
- Autoencoders - dimensionality reduction, data denoising, and feature learning
- Generative tasks, such as image generation, style transfer, and data synthesis
- Transformers - Natural language processing tasks, such as language translation, text generation, and language understanding
- Deep Belief Networks - feature learning and dimensionality reduction
- Attention Mechanism - machine translation and natural language understanding

Not convinced, need more clarity for DL ?

I always believe, that if any concept can be conveyed through animation, it won't be forgotten easily.
This is the link: Deep Learning @Simplilearn.

Data Science Life Cycle



Importance of Data Science

- Decision Making
- Predicitve Capabilities
- Operational Efficiency
- Personalization and Targetting
- Innovation
- Risk Management
- Handling Big data

Python and its importance

- Ease of Learning and Use
- Rich Ecosystem of Libraries
- Community and Support
- Versatility
- Integration and Scalability
- Adaptability in Advanced Fields

Data Science Roles and Responsibilities

Role	Primary Responsibilities	Required Skills
Data Scientist	Analyze and interpret complex data, perform predictive modeling, and communicate findings.	Statistical analysis, Python/R, machine learning.
Data Analyst	Analyze data, generate reports, create dashboards, visualize data to support decisions.	SQL, data visualization (Tableau, PowerBI), basic statistics.
Data Engineer	Build and maintain scalable data systems, integrate data, ensure compliance.	Programming, big data tech, data pipeline tools.
Machine Learning Engineer	Design and deploy ML models, optimize performance and scalability.	Python, ML algorithms, TensorFlow/PyTorch.

Data Science Roles and Responsibilities (cont.)

Role	Primary Responsibilities	Required Skills
Data Architect	Design data architecture, develop new and optimize existing systems.	Database design, data warehousing, data modeling tools.
Business Intelligence Developer	Develop BI solutions, create dashboards and reports, analyze data.	BI tools, database knowledge, data warehousing.
Statistician	Apply statistical methods to analyze data and draw conclusions.	Statistical software, strong mathematical skills.
Quantitative Analyst / Data Modeler	Develop models for financial decision-making.	Quantitative analysis, financial theory, programming.
AI Specialist	Develop AI systems that simulate human intelligence processes.	AI principles, machine learning, programming languages.
Data Steward	Manage data assets, ensure data quality and compliance.	Data management, data protection laws, organizational skills.

Various application of Data Science

- Healthcare
- Finance
- E-Commerce
- Entertainment and Media
- Telecommunication
- Transporation and Logistics
- Public Sector
- Education
- Real Estate

Trending domains for ML and DL

- Healthcare Diagnosis
- Education
- Speech Processing (Natural Language Processing)
- Agriculture
- Digital Marketing
- Robotics

Data Science in Spotify

Data Science in Spotify

- Recommended System
- Natural Language Processing
- User Segmentation and clustering
- A/B Testing
- Music Categorization and Tagging
- Marketing and Advertising
- Challenges: Data Scalability, User Privacy
- Impact: User Engagement, Subscription Growth, Customer Satisfaction

Data Science in LinkedIN

Data Science in LinkedIN

- Recommended System - People You May Know (PYMK), Job Recommendations
- Natural Language Processing - Skill Extraction and Matching, Content Recommendation
- User Segmentation and clustering - User Behavior Analysis
- A/B Testing - Feature Optimization
- Economic Graph Analysis - Insights and Trends
- Marketing and Advertising - Ad Targeting and Performance Measurement
- Challenges: Data Privacy, Scalability
- Impact: User Engagement, Job Placement Efficiency, Advertising Revenue

Data Science in Uber

Data Science in Uber

- Demand Forecasting and Supply Positioning - Advanced Modeling, Heat Maps for driver
- Dynamic Pricing (Surge Pricing) - Algorithmic Pricing, User Acceptance Modeling
- Route and Traffic Optimization- Mapping Technology, Traffic pattern analysis
- Safety and Security Features - Risk assessment
- Customer and Driver Experience - Rating system, support ticket analysis
- Operational Efficiency - Driver onboarding and background check, Resource allocation
- Challenges: Data Privacy, Algorithm bias
- Impact: Increased efficiency, Market adoption, User Satisfication

Data Science in Education

- Personalized Learning
- Predictive Analysis
- Enhancing Teacher Performance
- Improving Learning Management System
- Curriculum Development
- Enhancing research
- Security and Compliance
- Admissions and Enrollment Analysis

Data Science in Customer - STARBUCKS

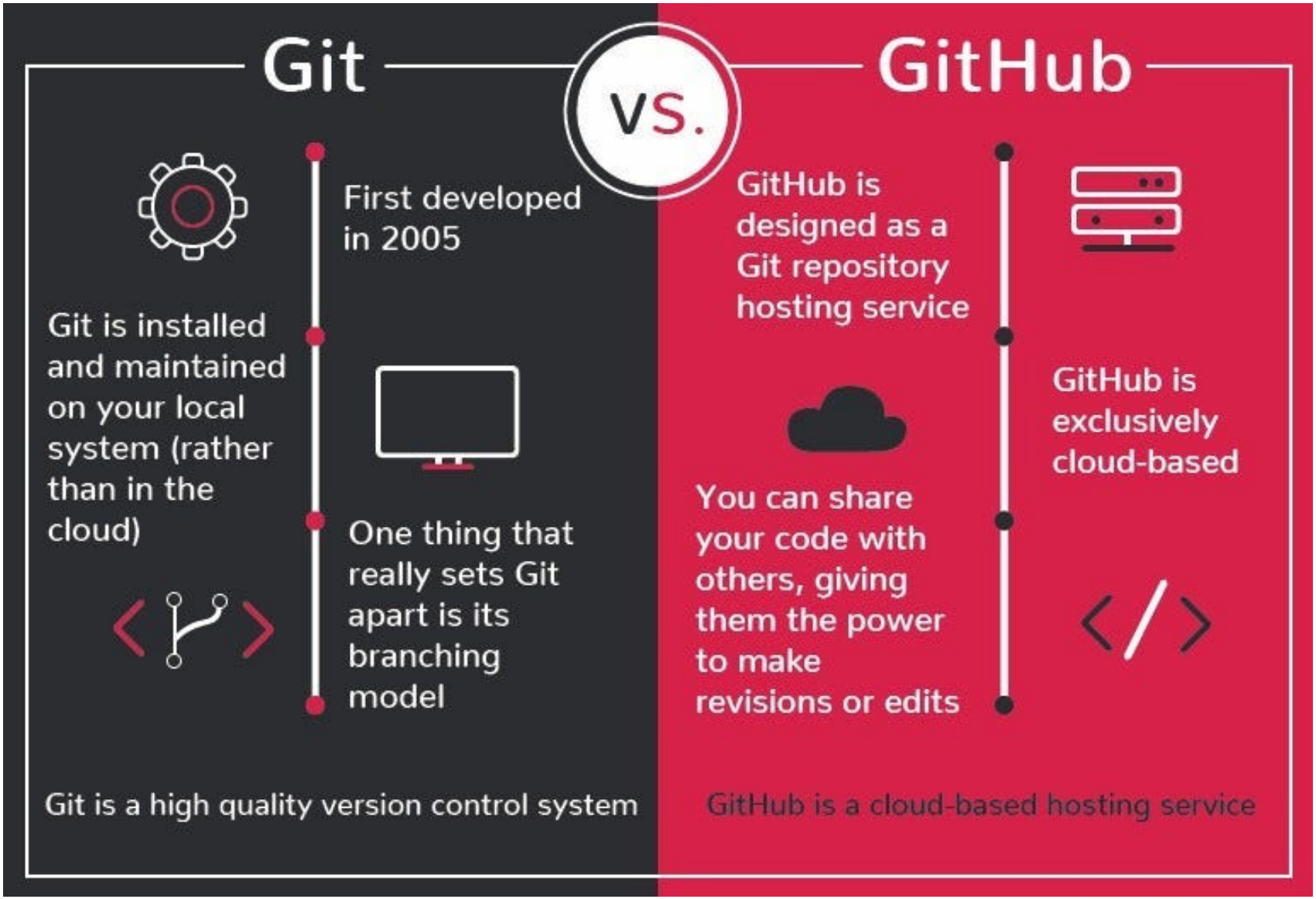
Data Science in Customer - STARBUCKS

- Personalized Marketing
- Optimal Store Location
- Inventory Management
- Customer Experience Management
- Supply Chain Optimization
- Challenges: Data Privacy, Integration of Offline and Online Data
- Impact: Enhanced Customer Engagement, Improved Sales, Strategic Expansion

Personality Analysis

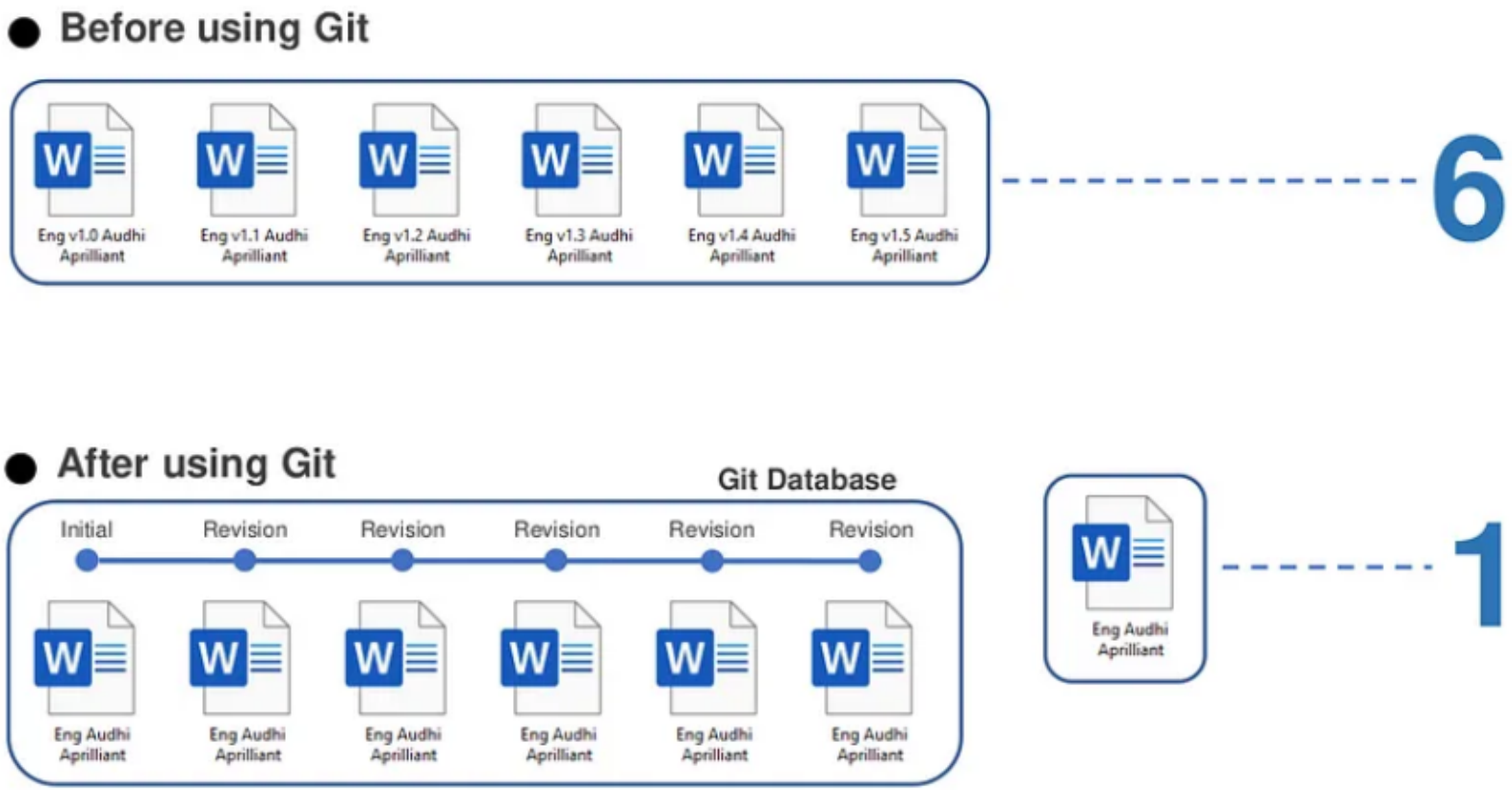
- Applying statistical and machine learning techniques to understand and predict human behaviors based on personality traits
- Data Collection
- Feature Extraction
- Modelling Personality
- Application of Personality Analysis
- Ethical Considerations
- Challenges - Accuracy, Bias, Interpretability
- Future?

Git and Github



Git and Github

- Git and GitHub are essential tools in modern software development, source code management, and version control.



Git

- Git is a distributed version control system designed to handle everything from small to very large projects with speed and efficiency.
- Distributed Architecture: Every developer's working copy of the code is also a repository that can contain the full history of all changes.
- Data Integrity: Git uses a data model that ensures the cryptographic integrity of every part of your project. Each versioned file and commit is checksummed and retrieved by its checksum at checkout.
- Branching and Merging: Git's branching features are robust and performant. These features facilitate various workflows, such as feature-based, task-based, or environment-specific branches.

Git Commands

- git init: Initializes a new Git repository.
- git clone: Copies an existing Git repository from another server.
- git add: Adds files to the staging area before committing.
- git commit: Records snapshots of the staging area (essentially, it takes the files as they are in the staging area and stores that snapshot permanently in your Git directory).
- git push: Updates the remote repository with any commits made locally to a branch.
- git pull: Fetches and merges any commits from the tracking remote branch.
- git branch: Lists, creates, or deletes branches.
- git checkout: Switches branches or restores working tree files.
- git merge: Combines multiple sequences of commits into one unified history.

Github

- GitHub is a web-based hosting service for version control using Git.
- It offers all of the distributed version control and source code management (SCM) functionality of Git plus its own features.
- Repository Hosting: GitHub provides a web-based graphical interface. It also provides access control and several collaboration features, such as a wikis and basic task management tools for every project.
- Pull Requests: Facilitates the discussion and review of code before it is merged into the main codebase.
- Forking: Copying a repository from one user's account to another, enabling you to make changes without affecting the original repository.
- GitHub Actions: Enables automation of workflows directly in your GitHub repository, from testing to deployment.
- Social Networking: Like any other social network, GitHub allows users to follow each other, rate each other's work, receive updates for specific projects, and communicate publicly or privately.

Tips to improve Data Science coding & knowledge

- Open GitHub repository & and start coding from scratch for different dataset Github link - Details about Github - Hit me!
- Try to participate, competitions in Kaggle website for major attractions. Eg: Abhishek Thakur (Approaching (Almost) Any Machine Learning Problem) - Kaggle link
- Try to follow worlds top scientist, R & D, some reowned personalities in the field of Data science, Machine Learning and Deep Learning for their work and tips - Linkedin link

To be in trend, we need to think differently.

Mail me: er.anandprem@gmail.com
Ring me: +91 73586 79961
Follow me: LinkTree

Learning gives Creativity,
Creativity leads to Thinking,
Thinking provides Knowledge,
and
Knowledge makes you Great
- Dr APJ Abdul Kalam