

ÉCOLE SUPÉRIEURE EN SCIENCES ET TECHNOLOGIES DE
L'INFORMATIQUE ET DU NUMÉRIQUE



FUNDAMENTALS OF DATA SCIENCE AND DATA MINING

CHAPTER I: INTRODUCTION TO DATA SCIENCE

Dr. Chemseddine Berbague

2024-2025

CONTENT

- Definition and Applications
- Basic Concepts
- Data Science Process.
- Importance in Decision Making process
 - Data Science Methodologies
 - Principal Tasks
- Tools and Technologies

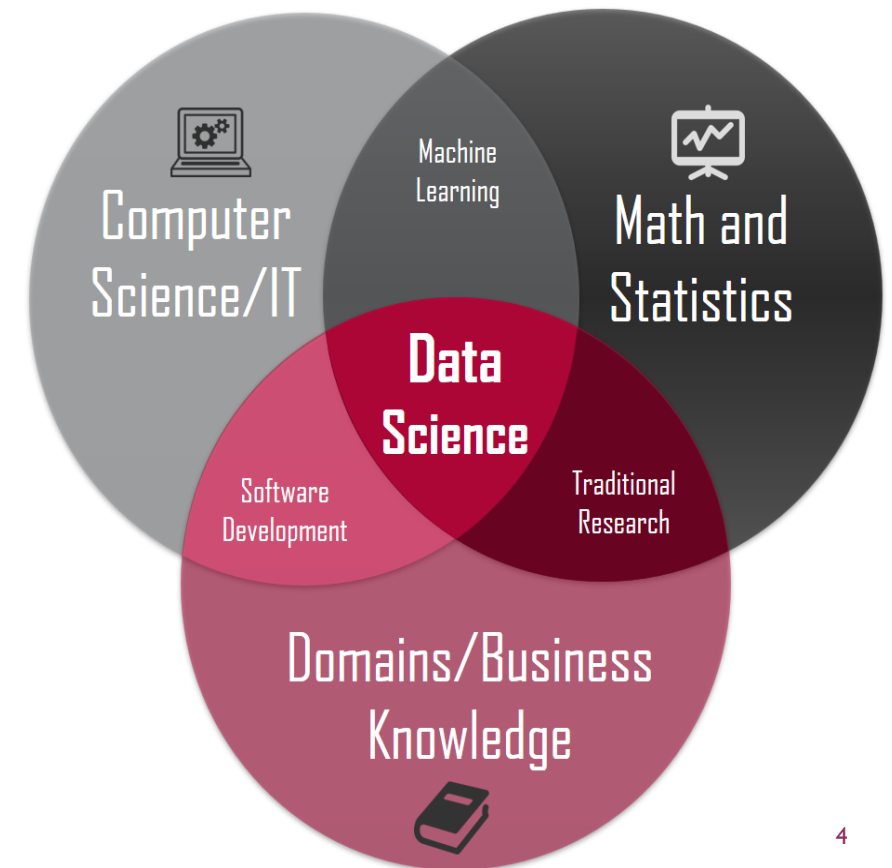
INTRODUCTION

*“...the role of data scientists is to extract value from data. Data scientists’ work helps improve how humans make decisions and how algorithms optimize outcomes. Through the **collection, analysis and interpretation** of data, data scientists extract empirically-based insights that augment and enhance how humans and algorithms work...”*

by Bob Hayes on October 23, 2015 in Big Data, Data Science

INTRODUCTION

- **Data science** adopts a collection of approaches, tools, and techniques to extract from raw data useful knowledge, or reusable patterns by firstly acquiring data, preparing data, modeling data, visualizing data, and deploying adequate tools.
- Most decision are made based on **partial information**, and **uncertain outcomes**. **Data science** allows to reduce the uncertainty by applying suitable computational tools.
- **Data science** provides the means to make precise, reliable, and quantitative arguments about any set of observations.



WHY DATA SCIENCE ?

- Harvard Business Review featured an article titled "Data Scientist: The Sexiest Job of the 21st Century" [1].
- More **conferences** are held annually focusing on **innovation** in the areas of **Data Science** and topics dealing with **Big Data**.
- **Data science** has big impact in solving real life problems such in **business** domain:
 - Empowering management and officers to make better decision.
 - Directing actions based on trends—which in turn help to define goals.
 - ...

What's the difference between data science and business intelligence ?

DATA DRIVEN DECISION MAKING

- **Data-driven decision making (DDDM):** Uses data to guide choices, replacing intuition with evidence-based insights.
- **Key benefits:**
 - **Accuracy:** Data uncovers patterns and improves predictions.
 - **Risk reduction:** Mitigates risks by highlighting potential issues.
 - **Efficiency:** Optimizes resource allocation and processes.
 - **Innovation:** Identifies new opportunities and trends.
- DDDM empowers organizations to make informed, strategic decisions that lead to better outcomes.

DATA DRIVEN DECISION MAKING

BENEFITS OF DATA-DRIVEN DECISION MAKING



Valuable
Insights



Continual
Growth



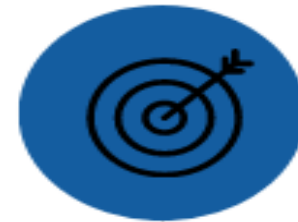
Improved
Program
Outcomes



Optimised
Operations

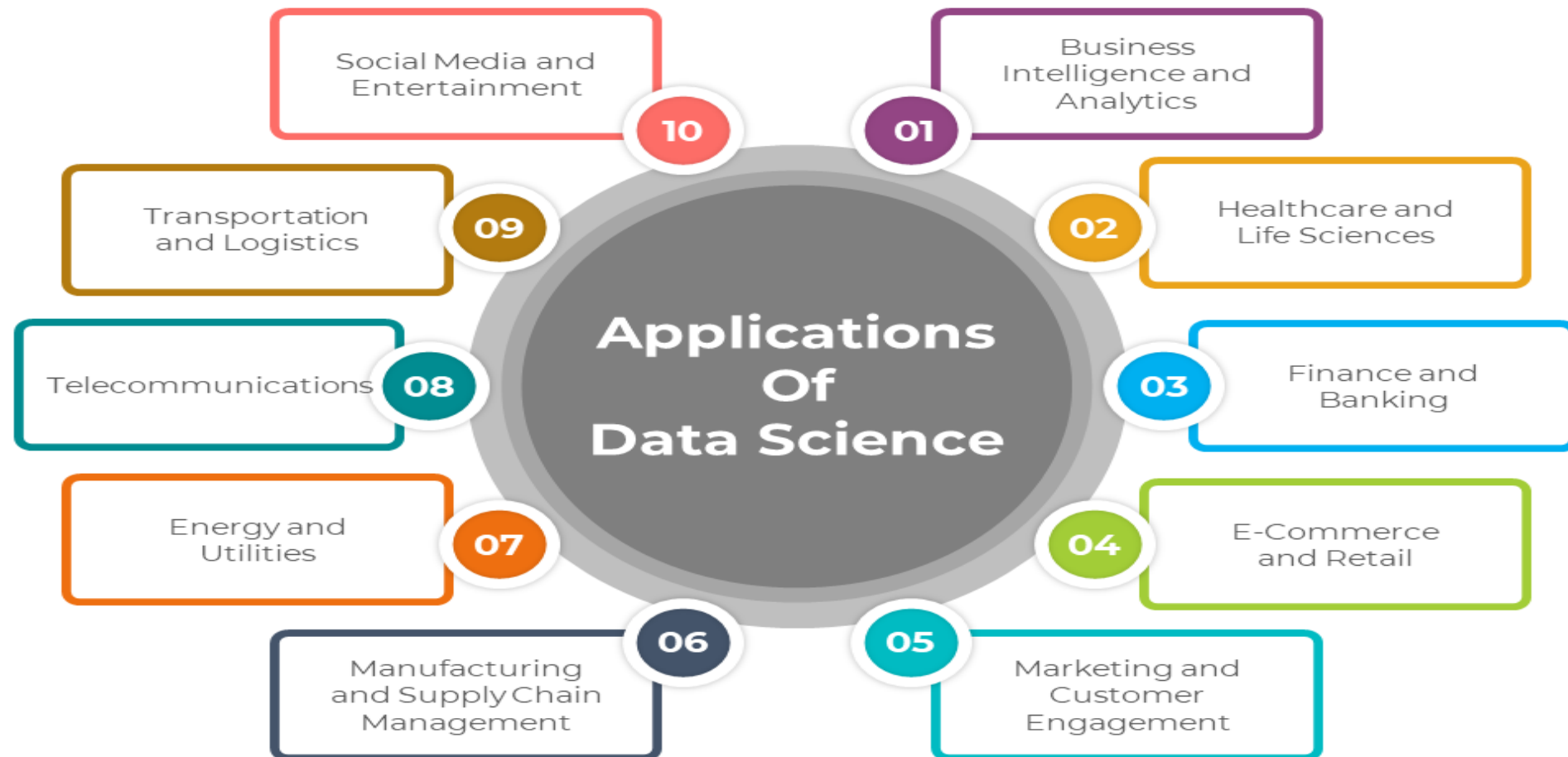


Prediction
Of Future
Trends



Actionable
Insights

APPLICATIONS OF DATA SCIENCE

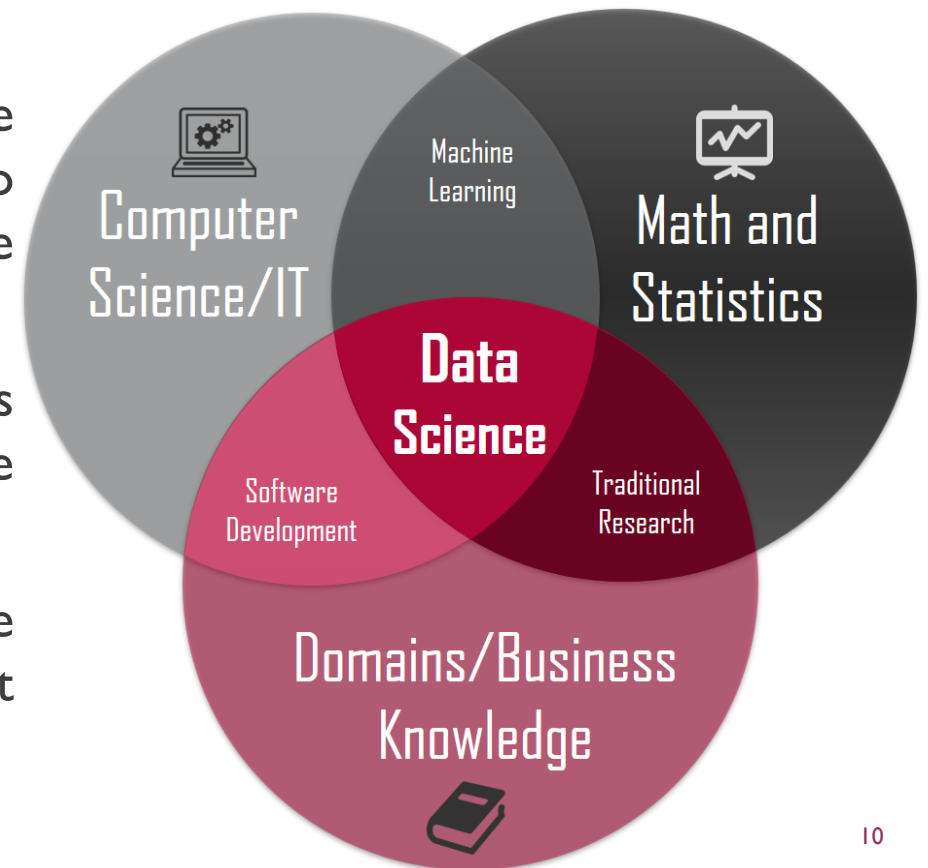


APPLICATIONS OF DATA SCIENCE

- **Healthcare:** Predicts outcomes, personalizes treatments, and optimizes operations.
- **Finance:** Detects fraud, manages risk, and improves financial decisions.
- **Marketing:** Analyzes customer behavior and creates targeted campaigns.
- **Retail:** Optimizes pricing, inventory, and product recommendations.
- **Transportation:** Improves logistics, route planning, and maintenance.
- **Entertainment:** Recommends content based on user preferences.

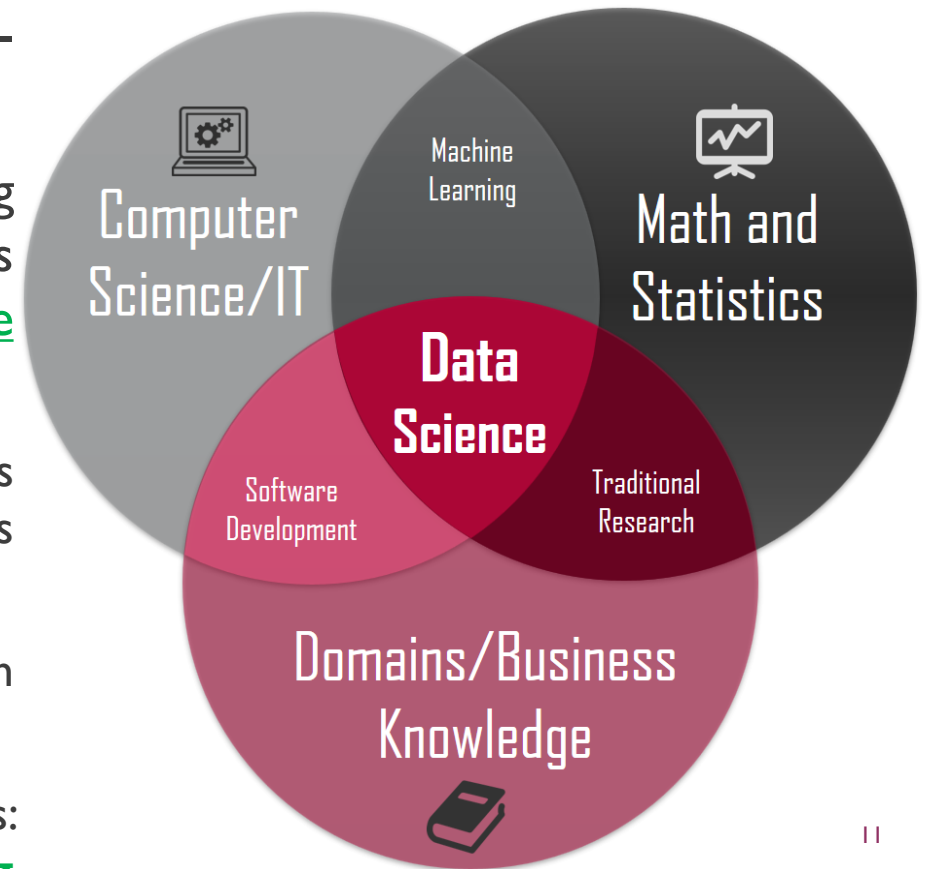
FUNDAMENTAL DEFINITIONS

- **Math and statistics** in data science involve concepts like probability, linear algebra, calculus, and statistical inference to analyze data patterns, model uncertainty, and make predictions.
- **Domain or business knowledge** in data science involves understanding the industry to interpret data accurately, solve real-world problems, and drive actionable insights.
- **Computer science and IT** in data science involve programming, algorithms, and data handling for efficient processing and scalable solutions.



FUNDAMENTAL DEFINITIONS

- **Data** is a set of descriptions of the world in form of texts, numbers stored in memory in a structured and non-structured forms.
- **Data science** extracts useful knowledge by applying computation tools on raw data. An effective data analysis requires different aspects such as: the exploration, the prediction, and the inference.
 - **Data science** requires some computational tools, whereas **Python**, and **R** are the most used languages used for this purpose.
 - **Data science** extends vocabulary, theories and findings from **statistics** field.
 - **Data science** adopts the same core inferential problems such as: **testing hypotheses, estimating confidence, and predicting unknown quantities ...etc.**

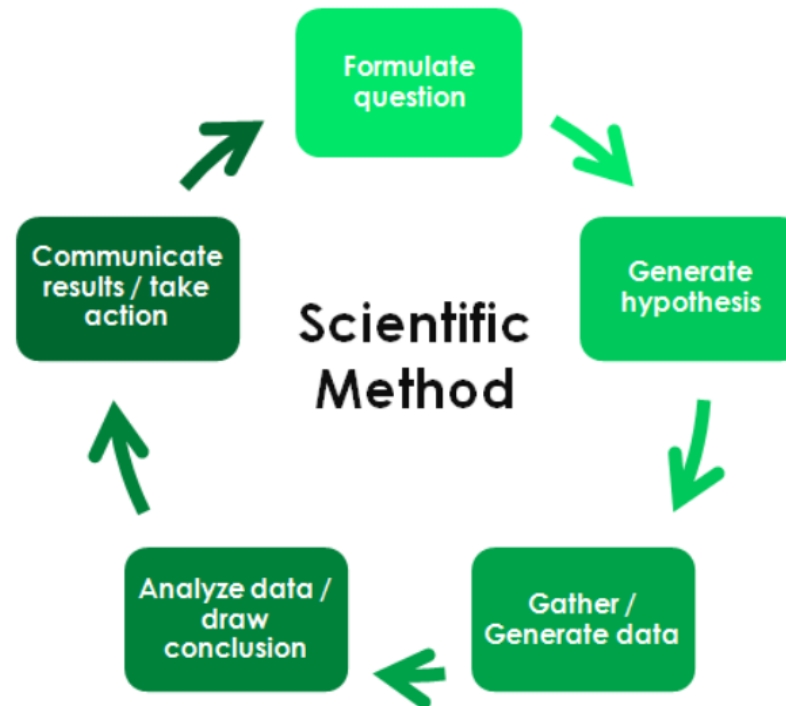


DATA SCIENCE PROCESS

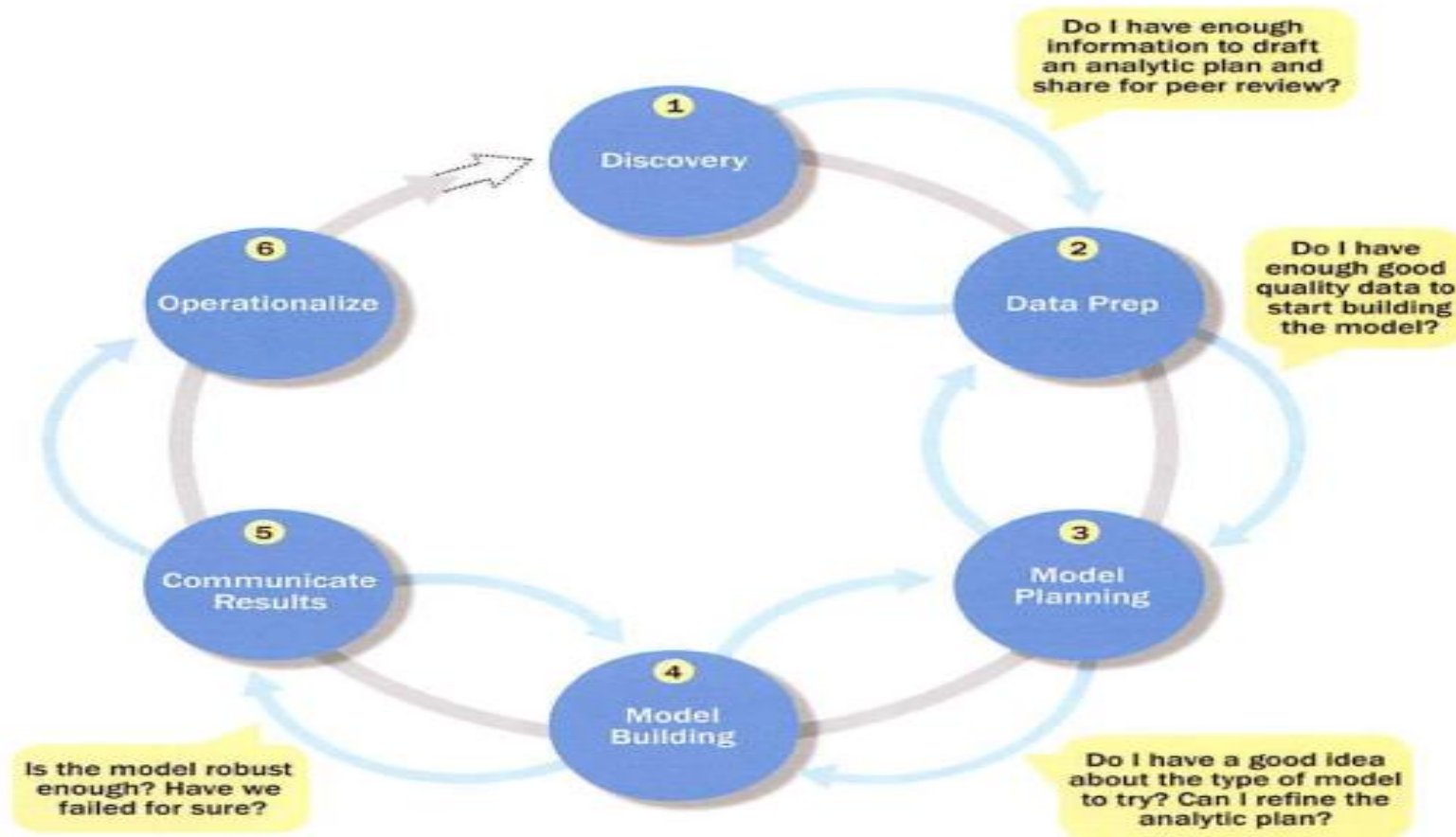
- A **data science project manager** follows a **process** to ensure the best practices for **spanning discovery** to project completion.
- These different guidelines and approaches were defined by consulting **expert data scientists** and **reviewing established approaches** from different disciplines:
 - A. Scientific method
 - B. CRISP-DM (**CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)
 - C. Tom Davenport's DELTA framework
 - D. Doug Hubbard's approach
 - E. MAD Skills" by Cohen et al.
 - F. TDSP by Microsoft



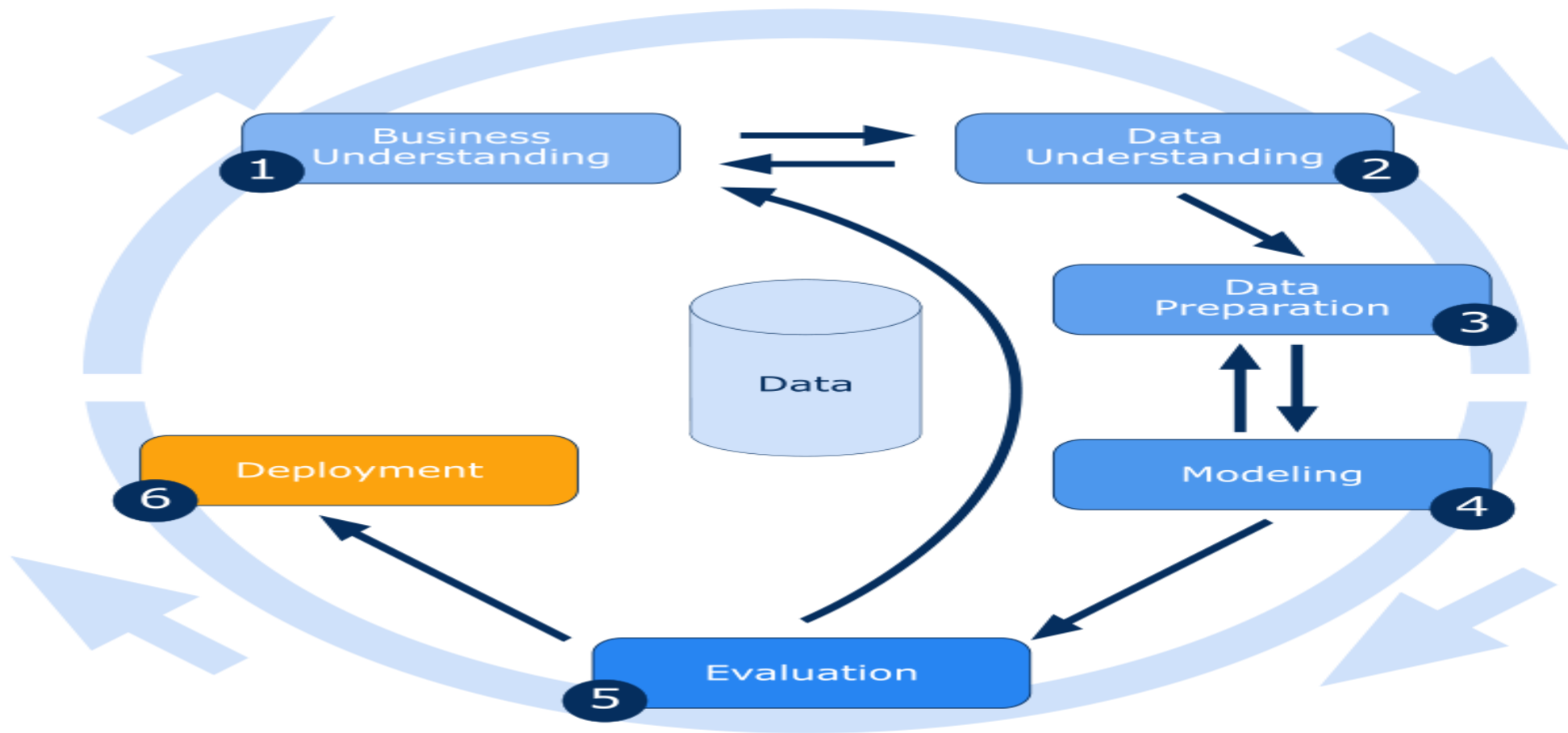
A. DATA SCIENCE APPROACH: SCIENTIFIC METHOD



A. DATA SCIENCE APPROACH: THE SCIENTIFIC METHOD IN MORE DETAILS



B. DATA SCIENCE APPROACH: CRISP-DM




CRISP-DM diagram process

B. DATA SCIENCE APPROACH: CRISP-DM

- CRISP-DM is one of the approaches adopted in data science projects and involves the next steps:
 1. Business understanding.
 2. Data acquisition and understanding.
 3. Modeling.
 4. Deployment.
 5. Customer acceptance.
- **CRISP-DM** has appeared in 1999 to standardize data mining process across industries, and is the most common used one for data science projects.
- Usually the **combination** of **CRISP-DM** and **agile** approaches achieve the best results.

C. DATA SCIENCE APPROACH: ANALYTICAL MATURITY MODEL (DELTA)



	DATA	ENTERPRISE	LEADERSHIP	TARGETS	ANALYSTS
STAGE 5 Analytical Competitors	Relentless search for new data and metrics	All key analytical resources centrally managed	Strong leadership passion for analytical competition	Analytics support the firm's distinctive capability and strategy	World-class professional analysts and attention to analytical amateurs
STAGE 4 Analytical Companies	Integrated, accurate, common data in central warehouse	Key data, technology and analysts are centralized or networked	Leadership support for analytical competence	Analytical activity centered on a few key domains	Highly capable analysts in central or networked organization
STAGE 3 Analytical Aspirations	Organization beginning to create centralized data repository	Early stages of an enterprise-wide approach	Leaders beginning to recognize importance of analytics	Analytical efforts coalescing behind a small set of targets	Influx of analysts in key target areas
STAGE 2 Localized Analytics	Data useable, but in functional or process silos	Islands of data, technology, and expertise	Only at the function or process level	Multiple disconnected targets that may not be strategically important	Isolated pockets of analysts with no communication
STAGE 1 Analytically Impaired	Inconsistent, poor quality, poorly organized	n/a	No awareness or interest	n/a	Few skills, and these attached to specific functions

- Developed in 2010 by Tom Davenport and al.
- It has 5 levels of maturity.

D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH

The screenshot displays the top portion of the Hubbard Decision Research website. At the top, a browser address bar shows the URL: [hubbardresearch.com/about/applied-information-economics/#:~:text=Applied%20Information%20Economics%20\(AIE\)%20is,on%20improving%20human%20e...](http://hubbardresearch.com/about/applied-information-economics/#:~:text=Applied%20Information%20Economics%20(AIE)%20is,on%20improving%20human%20e...). Below the address bar is a blue navigation bar containing the phone number (630) 858-2788, the email address info@hubbardresearch.com, and a 'Log In/Register' link. The main header area features the Hubbard Decision Research logo on the left and a series of navigation links on the right: 'Consulting', 'AIE Academy', 'About Us' (highlighted in blue), 'Big Decisions Blog', and 'Cont'. The background of the page is a dark blue gradient with faint, large white text that reads 'APPLIED INFORMATION ECONOMICS (AIE)'. The left side of the background image shows a blurred view of a computer screen displaying various numbers.

D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH


Hubbard Decision Research (HDR) is a consulting firm that helps organizations make measurably better decisions through the use of quantitative methods. Over its 20+ year history, HDR has helped firms discover that **there are no true “immeasurables”** and that they can make measurably better decisions.

Consultation

Training

About Us

D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH




Consulting ▾AIE Academy ▾About Us ▾

Home / Consulting

CONSULTING


Showing all 2 results

Default sorting ▾



DOUG HUBBARD PHONE CONSULTATION

\$500.00



HDR EXPERT PHONE CONSULTATION

\$325.00

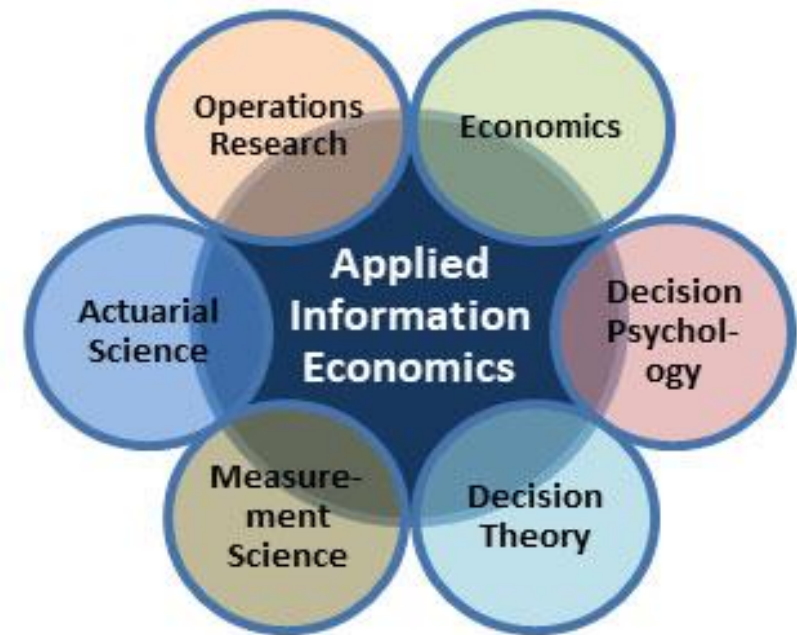
D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH

“ MEASURE WHAT MATTERS, MAKE BETTER DECISIONS”

“ A decision is only as good as the data and analysis on which it is based. Our proprietary quantitative analytical method, Applied Information Economics, provides a proven, scientific framework for measuring anything and using the results to make more informed – and better – decisions.”

D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH

- Applied Information Economics (AIE) is a synthesis of techniques from **economics, actuarial science, and other mathematical methods** (next Figure). AIE employs methods that are proven by a large body of peer-reviewed academic research and empirical evidence on improving human expert judgments.



D. DATA SCIENCE APPROACH: DOUG HUBBARD'S APPROACH

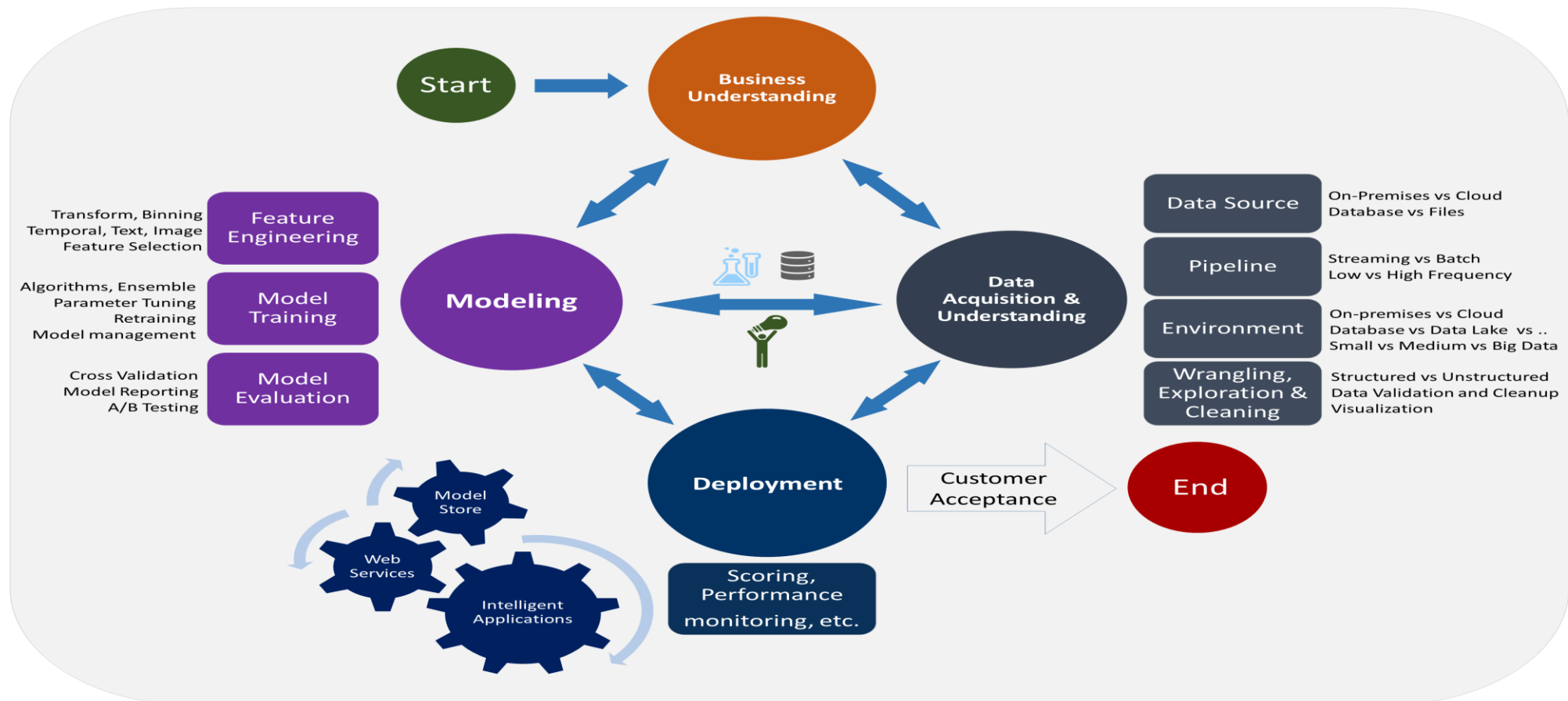
- Doug Hubbard is known for his approach to quantitative risk analysis and decision-making, particularly in the fields of **risk management and data-driven decision support**. His approach is outlined in his book "How to Measure Anything: Finding the Value of Intangibles in Business,“. Key principals of this approach are cited in next:
 - Define the Decision(s)
 - Model What We Know Now
 - Measure What Matters
 - Make Better Decisions

E. DATA SCIENCE APPROACH: TDSP

- The Team Data Science Process (**TDSP**) is a comprehensive framework developed by Microsoft for guiding data science projects.
- It provides a structured, end-to-end approach that encompasses various stages of a data science project, from data acquisition and understanding to model deployment.
- **TDSP** emphasizes collaboration among team members, documentation, and integration with Microsoft tools like Azure Machine Learning to streamline the data science workflow.
- It's designed to help data science teams effectively tackle real-world business problems by providing guidance and best practices throughout the entire project lifecycle.

E. DATA SCIENCE APPROACH: TDSP

Data Science Lifecycle



CRISP-DM diagram process

COMPARISON OF TDSP TO CRISP-DM

	TDSP	CRISP-DM
Scope	TDSP is a broader framework applicable on a wide range of data science projects, including machine learning, deep learning, big data, and more. It's designed to be flexible and adaptable to different project types and industries.	CRISP-DM was initially developed with a focus on data mining projects. While it has been adapted for broader data science use, its origins are in data mining, and it may be more rigid when applied to non-data mining projects.
Tools	TDSP is closely integrated with Microsoft tools and services, particularly Azure Machine Learning. It provides guidance on how to leverage these tools effectively throughout the data science project lifecycle.	CRISP-DM does not have specific tool integrations and is more tool-agnostic.
Process flow	TDSP typically consists of several phases, including Business Understanding, Data Acquisition and Understanding, Data Preparation, Modeling, Evaluation, Deployment, and Customer Acceptance.	CRISP-DM consists of six major phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.
Collaboration	TDSP places a strong emphasis on collaboration among team members, making it suitable for larger teams and enterprise settings.	CRISP-DM also encourages collaboration but may not provide as detailed guidance on team roles and responsibilities.
Documentation	TDSP provides extensive documentation templates and best practice guidance for each phase of the project, helping teams maintain good documentation and best practices throughout.	CRISP-DM provides less explicit documentation guidance.

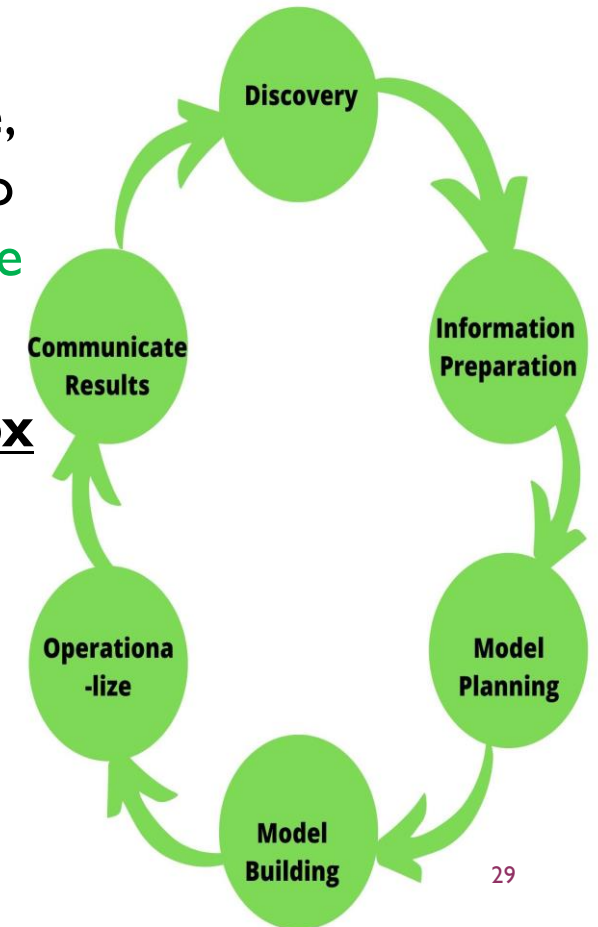
COMPARISON OF TDSP TO THE SCIENTIFIC METHOD





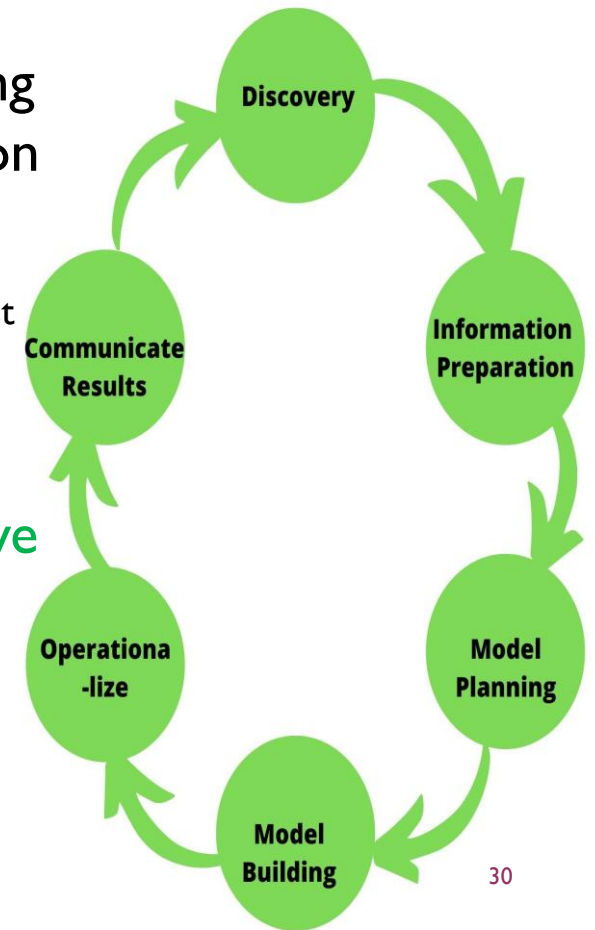
DATA SCIENCE PROCESS :A SUMMARY

- **Discovery:** the business team checks if any similar analytical problems were addressed previously, checks available resources such as data, time, technology, and people. Also, they put hypotheses, and propose a plan to proceed with the project including metrics to validate the success of the project.
- **Data preparation:** In this phase, the team collects data into one sandbox by extracting, loading, and transforming data from different resources. Additionally, the team should get familiar with the initial data.
- **Model planning:** In this phase, the working team tests the data by selecting among the available variables, the ones which describe more effectively the targeted problems, and checks different machine learning and statistical techniques which may be used to extract useful knowledge.

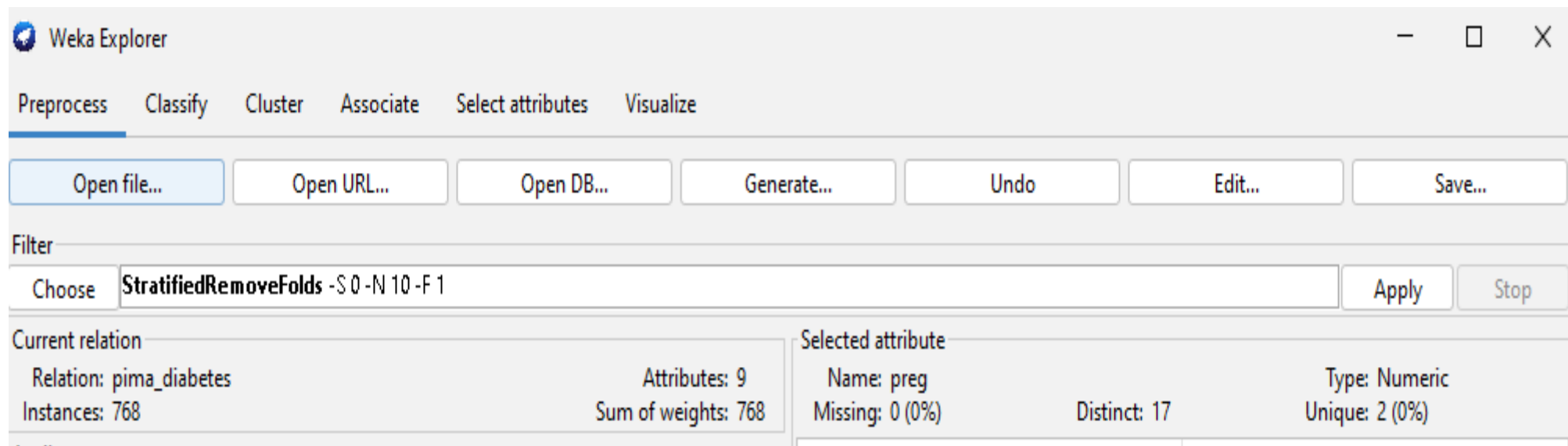


DATA SCIENCE PROCESS :A SUMMARY

- **Model building:** Plans from previous step are ran in this phase, by looking to the suitable execution environment (such as storage and computation resources).
- **Communicate results:** The team **validates** the objectives setting in the 1st phase, by checking the results obtained in the previous phase. In this phase, the team **should identify the key findings**, and **prepare summaries**, illustrations to **communicate and explain** the results, or put an **alternative plan** to proceed with the analytical project.
- **Operationalize:** Final results, validated models and technical files are delivered in this phase. Applications based on the constructed models may be deployed and integrated within the information system for exploitation.



DATA SCIENCE: MAIN TASKS



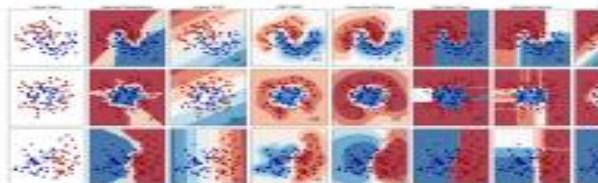
DATA SCIENCE: MAIN TASKS

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...



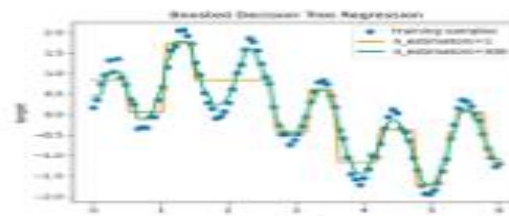
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...



Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes.

Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



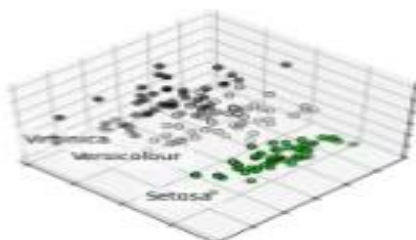
Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, increased efficiency.

Algorithms: PCA, feature selection, non-negative matrix factorization, and more...



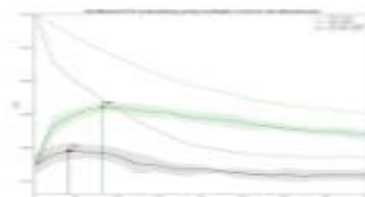
Examples

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning.

Algorithms: grid search, cross validation, metrics, and more...



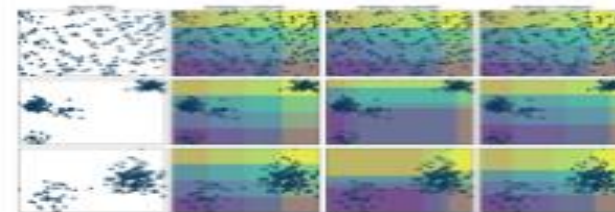
Examples

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.

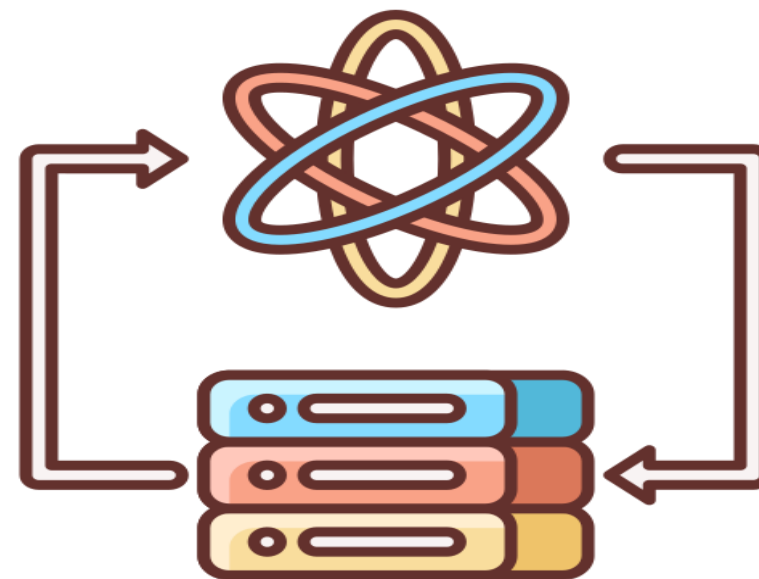
Algorithms: preprocessing, feature extraction, and more...



Examples

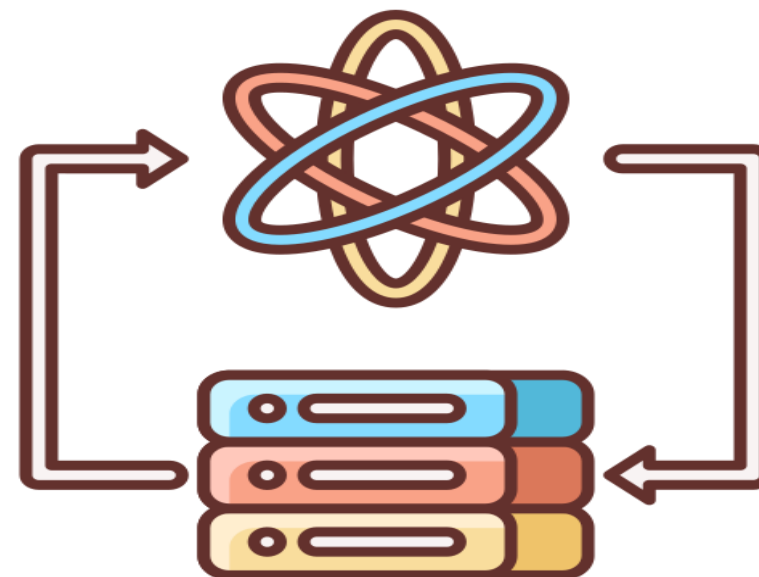
DATA SCIENCE TASKS: DEFINE OBJECTIVES

- **Business Understanding:** Clearly define the problem and expected outcomes.
- **Project Scope:** Identify key deliverables and involved stakeholders.
- **Data Availability:** Ensure relevant data is accessible and sufficient.
- **Performance Metrics:** Set measurable success criteria (e.g., accuracy, cost savings).
- **Constraints:** Consider time, budget, and resource limitations. Long-term Impact: Align with future goals and scalability.



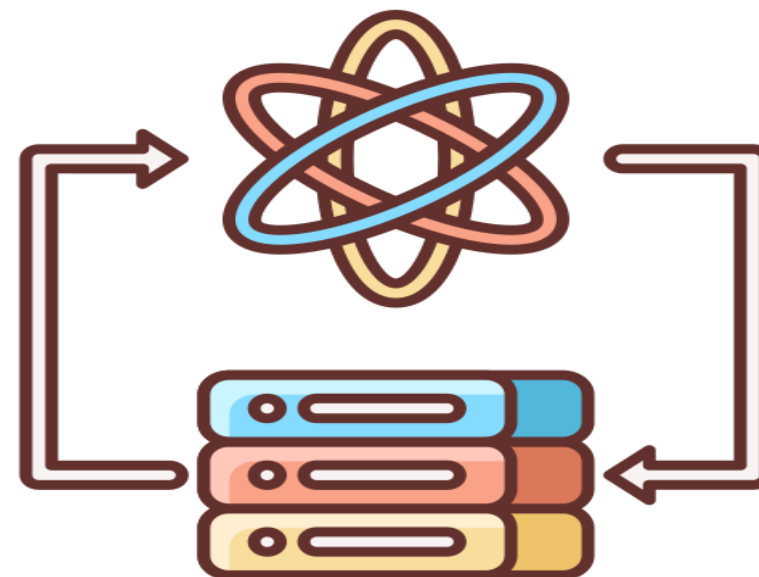
DATA SCIENCE TASKS: DEFINE OBJECTIVES

- **Increase Customer Retention:** Build a predictive model to identify customers likely to churn.
- **Optimize Marketing Campaigns:** Use data segmentation to target the most responsive customer groups.
- **Predict Equipment Failure:** Develop a maintenance schedule based on sensor data and predictive analytics.
- **Fraud Detection:** Identify suspicious patterns in transactions to prevent fraud.
- **Improve Product Recommendations:** Enhance personalization in a recommendation system using user behavior data.



DATA SCIENCE TASKS: DEFINE OBJECTIVES

- Tools used to identify project objectives may include:
 - **SWOT Analysis:** Helps identify project strengths, weaknesses, opportunities, and threats.
 - **SMART Framework:** Ensures objectives are Specific, Measurable, Achievable, Relevant, and Time-bound.
 - **OKR (Objectives and Key Results):** Aligns goals with measurable outcomes.
 - **CRISP-DM:** A structured approach to planning and organizing data science projects.
 - **Project Management Tools:** Tools like Trello, Asana, or Jira for tracking objectives, tasks, and timelines.



DATA SCIENCE TASKS: DEFINE OBJECTIVES

■ Here, we have an example of using SWOT analysis to define DS objectives:

■ **Strengths**

- Rich historical data: 5 years of equipment sensor data and maintenance records
- Domain expertise: Experienced maintenance engineers on the team

■ **Weaknesses**

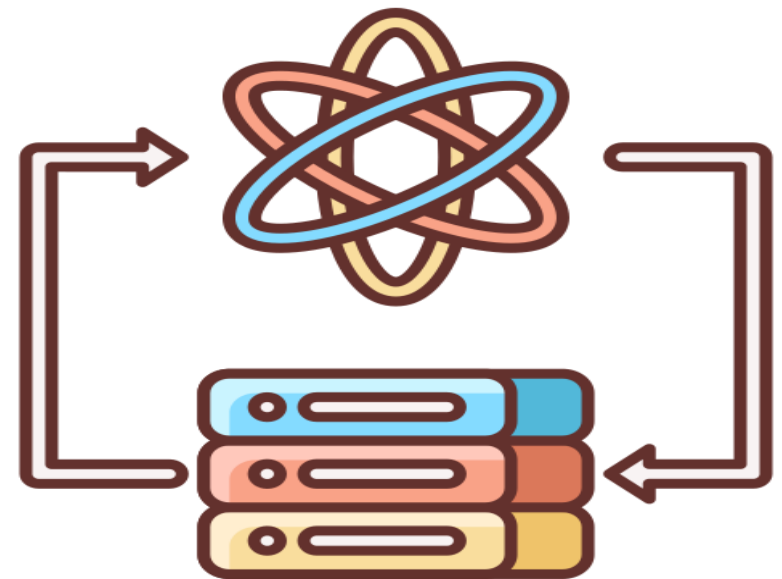
- Data quality issues: 15% of historical data has missing values
- Limited data science expertise: Only two data scientists on the team
- Resistance to change: Some staff skeptical about adopting new maintenance procedures

■ **Opportunities**

- Growing market for smart manufacturing solutions
- Academic research: Recent advancements in time-series forecasting techniques
- Regulatory incentives: Government grants available for energy-efficient manufacturing projects

■ **Threats Cybersecurity risks:**

- Increased vulnerability due to connected systems
- Competitor advancements: Two major competitors already implementing predictive maintenance
- Economic uncertainty: Potential budget cuts in case of market



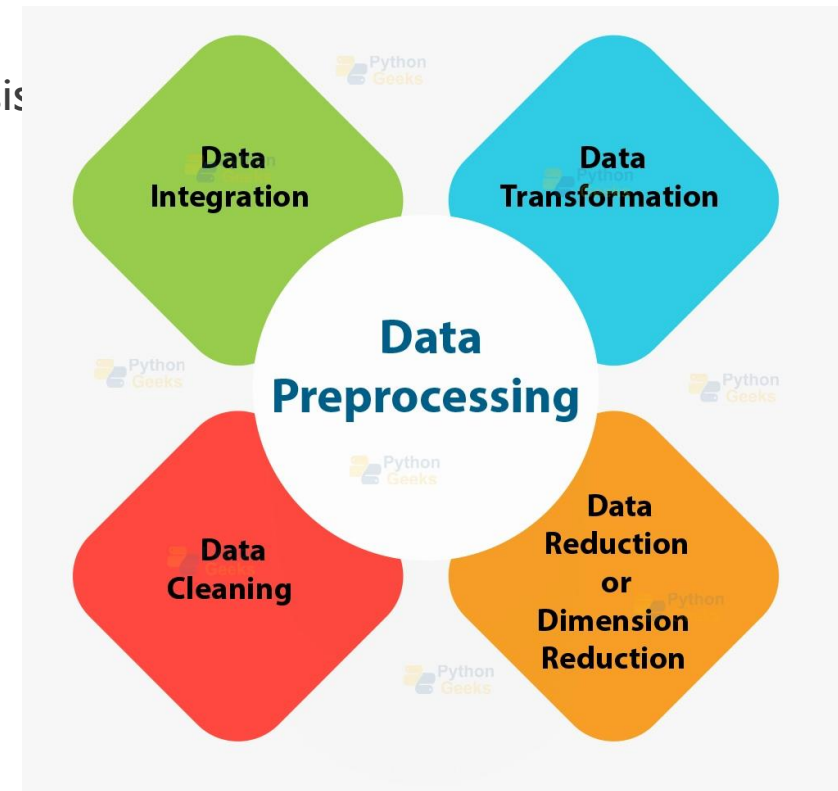
DATA SCIENCE TASKS: RETRIEVE THE DATA

- **Data retrieval** refers to the process of extracting relevant information from a database or data source for analysis or use in applications. Different techniques and tools are employed depending on the structure and type of data, the complexity of queries, and the intended use.
- **Techniques for Data Retrieval** include SQL queries, NoSQL retrieval, web scraping, APIs, search engines, ETL processes.



DATA SCIENCE TASKS: PREPROCESS THE DATA

- **Definition:** Transforming raw data into a clean, structured format for analysis.
- **Techniques:**
 - **Data Cleaning:** Handle missing values, remove duplicates.
 - **Data Transformation:** Scaling, encoding.
 - **Feature Engineering:** Create new features.
 - **Dimensionality Reduction:** Use PCA, t-SNE.
 - **Data Integration:** Combine sources.
 - **Outlier Detection:** Identify and manage outliers.



DATA SCIENCE TASKS: ANALYTICAL EXPLORATORY

■ Definition:

- EDA is the process of analyzing and visualizing data to uncover patterns, trends, relationships, and anomalies before formal modeling.

■ Objectives:

- Understand data structure and distribution.
- Identify patterns, correlations, and anomalies.
- Generate hypotheses and insights for further analysis.

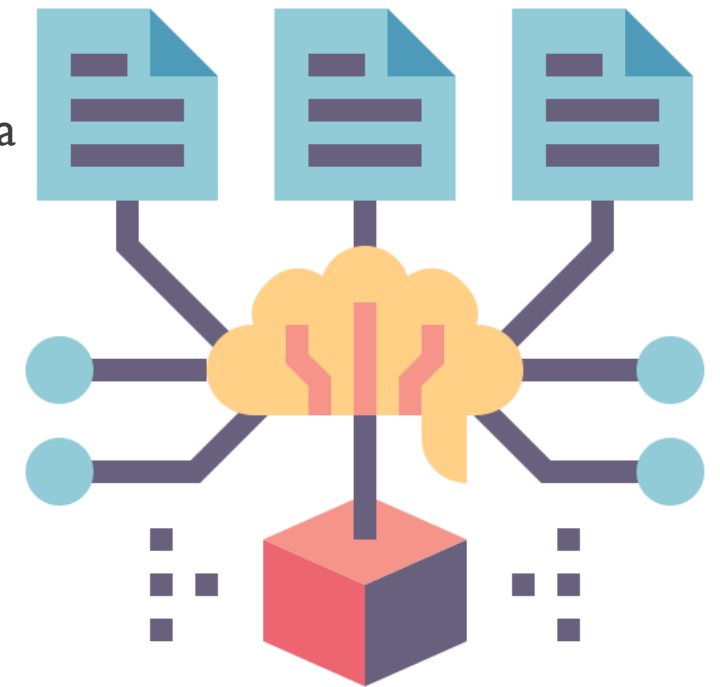
■ Main Techniques:

- Descriptive Statistics: Mean, median, standard deviation.
- Data Visualization: Histograms, scatter plots, box plots.
- Correlation Analysis: Pearson/Spearman correlations.
- Outlier Detection: Box plots, Z-scores.
- Missing Data Analysis: Identify and handle missing values.



DATA SCIENCE TASKS: MODEL CONSTRUCTION

- **Definition:** Creating a mathematical representation of real-world phenomena using data to predict outcomes and identify patterns.
- **Key Steps:** Problem definition, data collection, data preparation, model selection, model training, model evaluation, model tuning, and deployment.
- **Importance:** Facilitates data-driven decision-making and predictive analytics.
- **Challenges:** Managing incomplete data, overfitting, and ensuring model interpretability.



DATA SCIENCE TASKS: RESULTS DELIVERY AND DEPLOYMENT

- **Results Delivery**

- **Data Visualization:** Create dashboards and visual reports using tools like Tableau, Power BI, or libraries in Python (e.g., Matplotlib, Seaborn) to present findings clearly.
- **Documentation:** Provide comprehensive documentation of methodologies, data sources, and assumptions made during the analysis to ensure transparency.
- **Presentations:** Prepare presentations tailored to your audience (technical vs. non-technical) to communicate results and recommendations effectively.
- **Stakeholder Feedback:** Engage stakeholders to gather feedback on findings and ensure alignment with their needs.

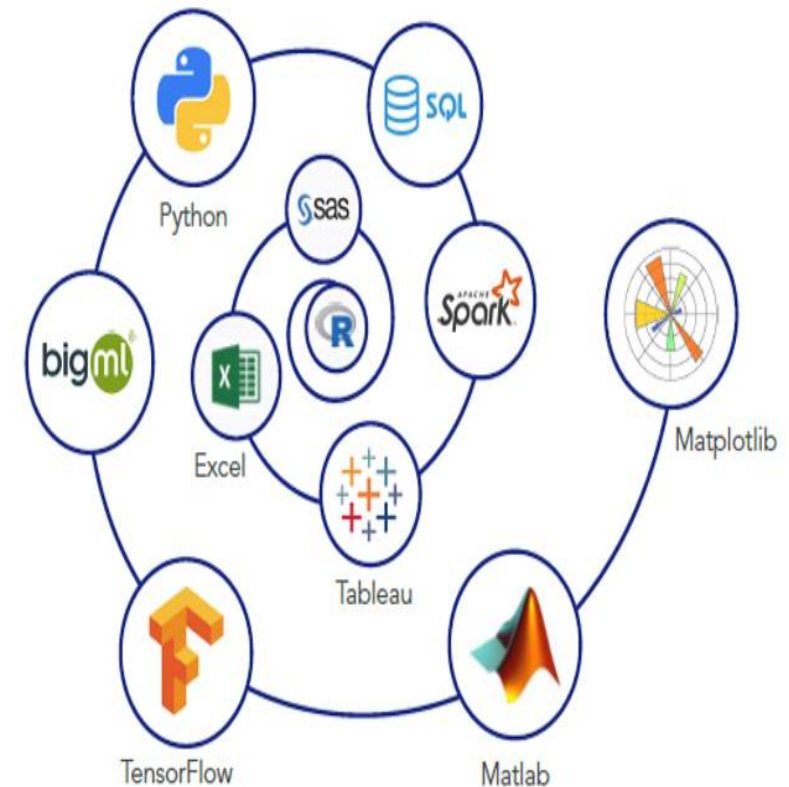
DATA SCIENCE TASKS: RESULTS DELIVERY AND DEPLOYMENT

- **Model Deployment**

- **Model Packaging:** Use tools like Docker to containerize models for consistent deployment across different environments.
- **API Development:** Develop APIs (e.g., using Flask or FastAPI) to allow applications to interact with the model, enabling real-time predictions.
- **Monitoring and Maintenance:** Implement monitoring tools (e.g., Prometheus, Grafana) to track model performance in production and set up alerts for degradation.
- **Version Control:** Use version control systems (e.g., Git) to manage code and model changes over time, allowing for easier rollbacks and updates.
- **Automation:** Consider CI/CD pipelines to automate the deployment process, ensuring smooth updates and consistent model performance.

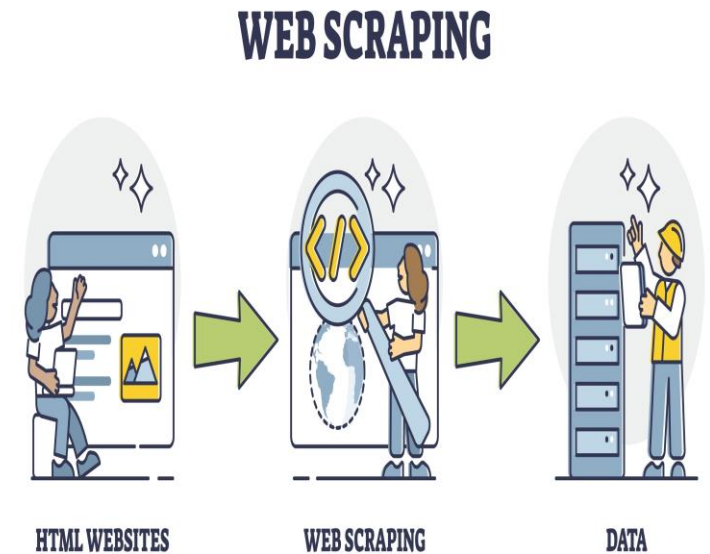
DATA SCIENCE ECOSYSTEM

- **Data Sources:** Databases, APIs, Cloud Storage
- **Processing:** Pandas, Apache Spark, Hadoop
Analysis: R, Python (Pandas, Numpy)
- **Machine Learning:** Scikit-learn, TensorFlow, PyTorch
- **Visualization:** Matplotlib, Seaborn, Tableau
- **Deployment:** Flask, Docker, Kubernetes
- **Collaboration:** Git, Jupyter Notebooks



DATA SCIENCE TOOLS

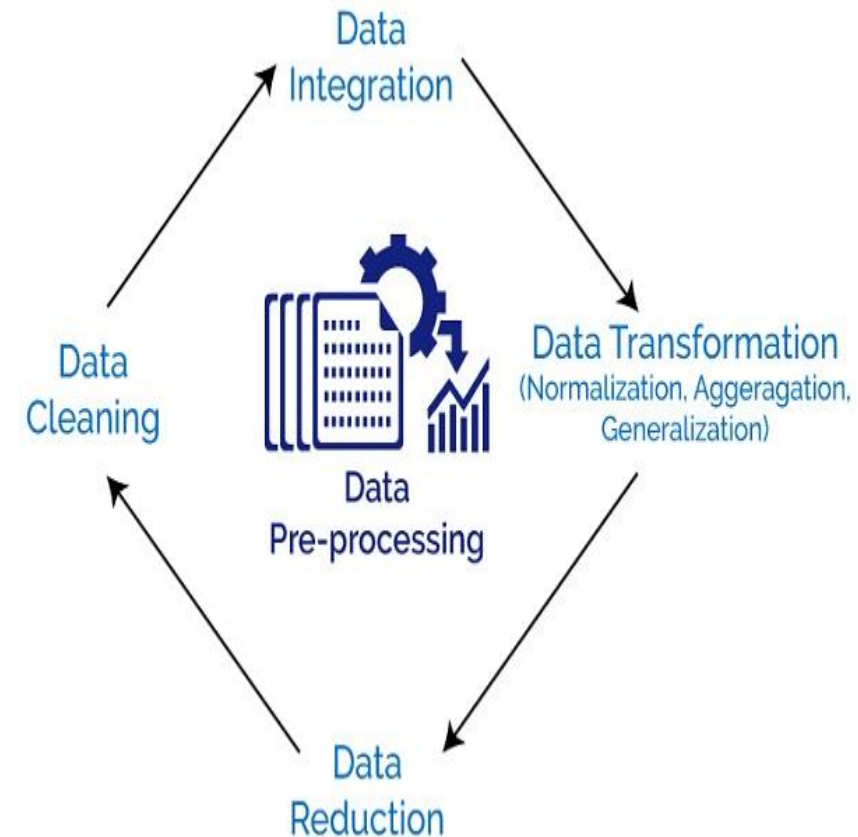
- **Data Collection APIs:** Tools like Postman for testing APIs and libraries such as requests in Python for data extraction.
- **Web Scraping:** Tools like BeautifulSoup, Scrapy, and Selenium for gathering data from web pages.
- **Database Management Systems:** SQL databases (MySQL, PostgreSQL) and NoSQL databases (MongoDB, Cassandra) for structured and unstructured data storage.



DATA SCIENCE TOOLS

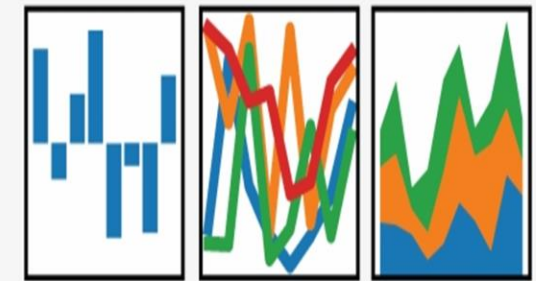
- **Data Preparation and Cleaning**

- **Pandas:** A powerful Python library for data manipulation and analysis, offering data structures like DataFrames.
- **Dplyr:** An R package for data manipulation, allowing for filtering, grouping, and summarizing data.
- **OpenRefine:** A tool for cleaning messy data, transforming it from one format to another, and exploring large datasets.



DATA SCIENCE TOOLS

- **Exploratory Data Analysis (EDA):** is an essential step in the data analysis process, allowing analysts to summarize and understand datasets, identify patterns, and detect anomalies. Key tools used in EDA include:
 - **Pandas:** Pandas, users can easily compute descriptive statistics (mean, median, standard deviation), perform data cleaning and transformation, and create pivot tables for data summarization.
 - **NumPy:** is fundamental package for numerical computing in Python, NumPy provides support for arrays and matrices, along with a collection of mathematical functions to operate on these data structures.
 - **SciPy:** provides additional functionality for scientific and technical computing. It includes modules for optimization, integration, interpolation, eigenvalue problems, and statistical tests. It is particularly useful for performing statistical analyses, including t-tests, ANOVA, and other hypothesis testing to assess relationships and differences between groups.



$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

pandas

DATA SCIENCE TOOLS

- **Machine Learning and Modeling**

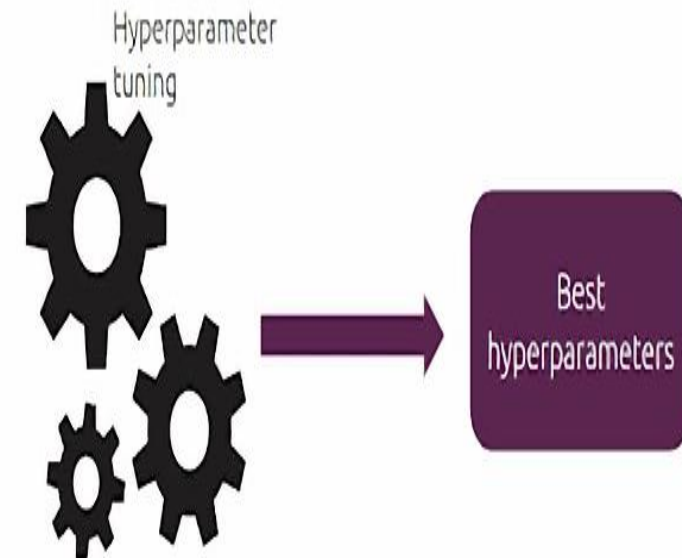
- **Scikit-learn:** A Python library that provides simple and efficient tools for data mining and machine learning, including classification, regression, and clustering.
- **TensorFlow:** An open-source deep learning framework developed by Google, widely used for building and training neural networks.
- **PyTorch:** A popular deep learning library that offers dynamic computation graphs and is widely used in research and industry.



DATA SCIENCE TOOLS

■ Model Evaluation and Tuning

- **MLflow**: An open-source platform for managing the machine learning lifecycle, including experimentation, reproducibility, and deployment.
- **GridSearchCV**: A Scikit-learn tool for hyperparameter tuning that helps in finding the optimal parameters for a model through cross-validation.
- **SHAP and LIME**: Tools for model interpretability that explain the output of machine learning models.



DATA SCIENCE TOOLS

- **Data Visualization**

- **Plotly:** A Python library for creating interactive plots and dashboards that can be easily shared online.
- **Bokeh:** A Python interactive visualization library that targets modern web browsers for presentation and can handle large datasets.
- **ggplot2:** An R package based on the Grammar of Graphics, enabling the creation of complex multi-layered visualizations.



DATA SCIENCE TOOLS

- **Deployment and Monitoring**

- **Flask/FastAPI:** Python web frameworks for building RESTful APIs to serve machine learning models.
- **Docker:** A tool for creating, deploying, and managing containerized applications, ensuring consistency across environments.
- **Kubernetes:** An orchestration platform for automating the deployment, scaling, and management of containerized applications.



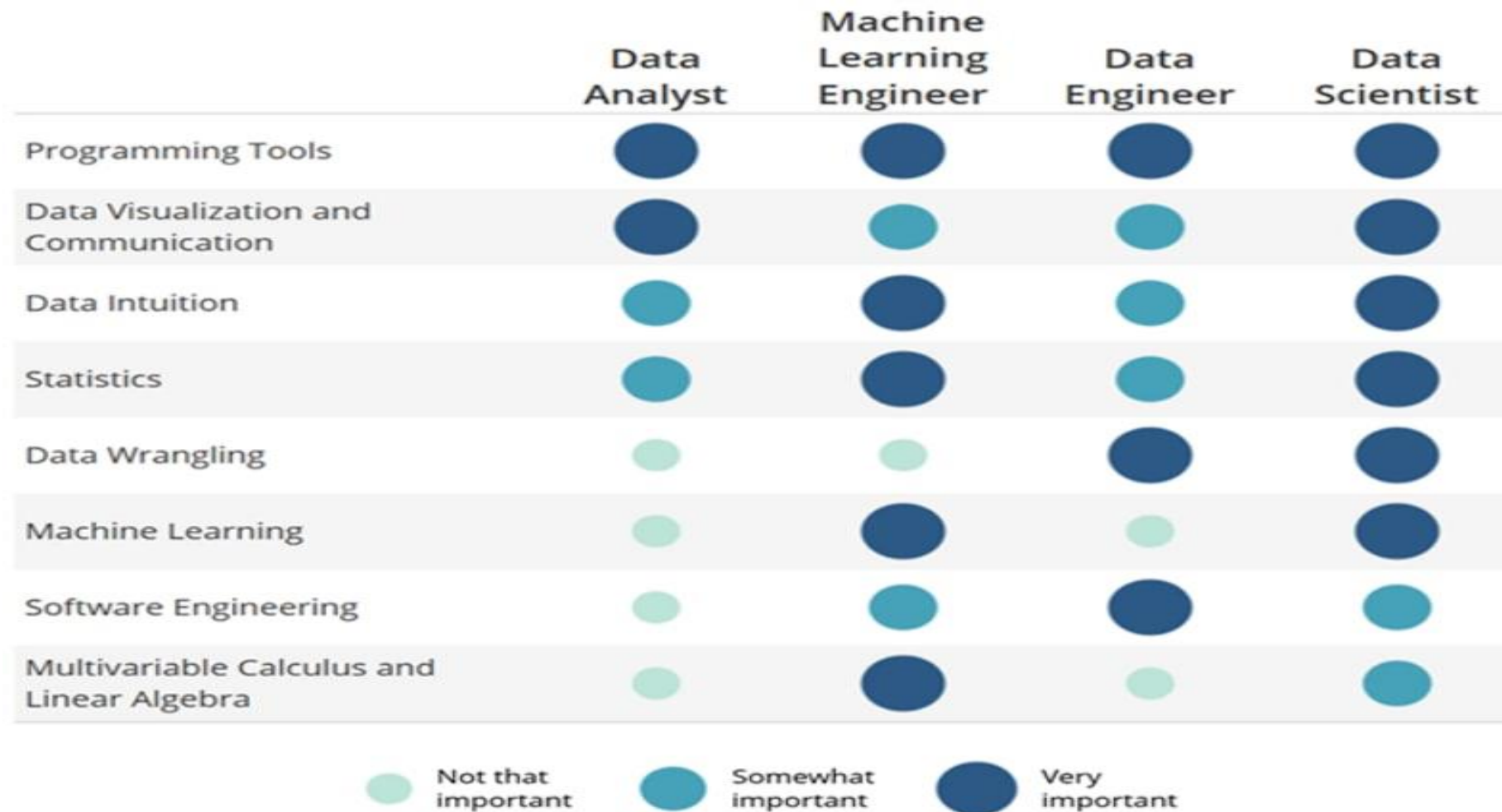
DATA SCIENCE TOOLS

- **Collaboration and Documentation**

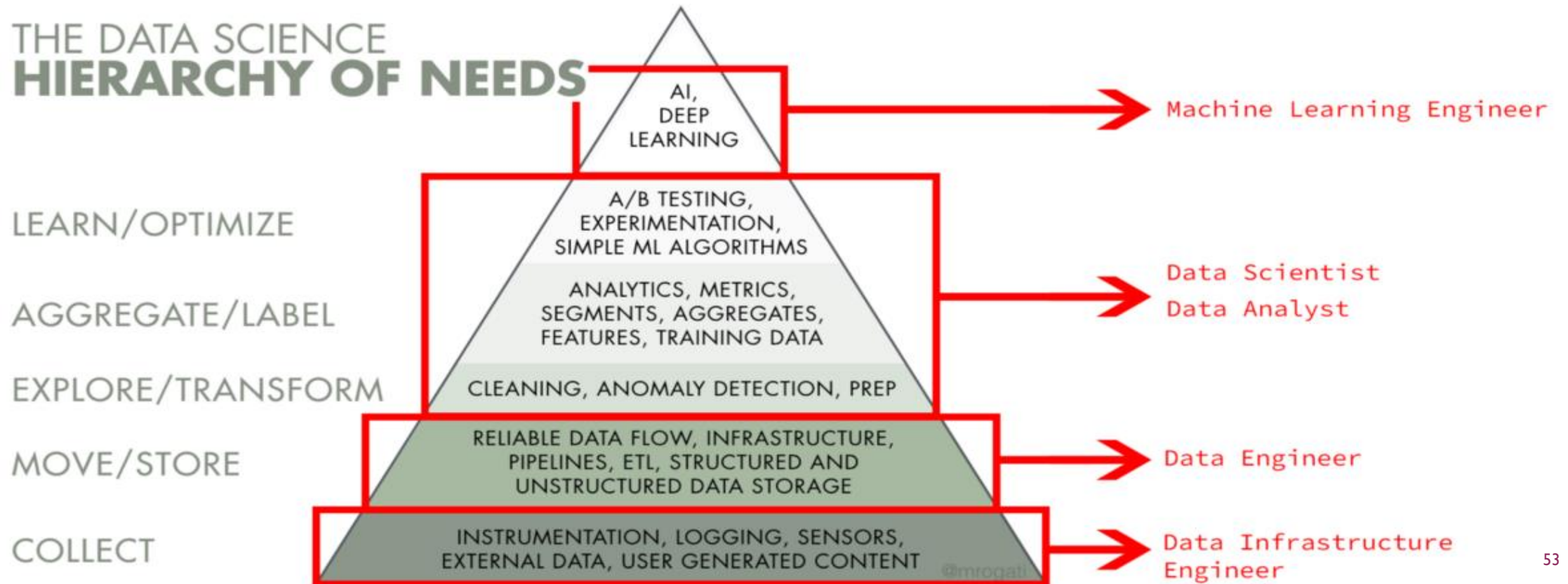
- **Jupyter Notebooks:** An open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.
- **Google Colab:** A cloud-based Jupyter notebook environment that provides free access to GPUs for running Python code and collaborating on data science projects.
- **Git:** A version control system that helps track changes in code and collaborate with other data scientists.



DATA SCIENCE ROLES



DATA SCIENCE ROLES



CONCLUSION

1. What is the difference between data science and business intelligence ?
2. Prepare a presentation holding details about the scientific method process ? And answer how it servers us in data science ?
3. Compare the scientific method to the CRISP-DM citing advantages of each one ?
4. By discussing the scientific method, several questions naturally arise, including:
 1. What is the difference between Treatment group and Control group ?
 2. Explain randomized controlled experiment in research and compare it to randomized controlled trial.
 3. What do we mean by an experimental design ?