UNIT I INTRODUCTION TO DATA SCIENCE

Syllabus

Data Science Roles: *A Brief History, Data Science – Engineering – Analysis - Modelling/Inference, What Kind of Questions Can Data Science Solve? – Prerequisites - Problem Type, Structure of Data Science Team, Data Science Roles.*

Soft Skill for Data Science: *Statistician vs Data Scientist, Beyond Data and Analytics, Three Pillars of Knowledge, Data Science Project Cycle - Types of Data Science Projects - Problem Formulation and Project Planning Stage - Project Modelling Stage - Model Implementation and Post Production Stage - Project Cycle Summary, Common Mistakes in Data Science - Problem Formulation Stage - Project Planning Stage - Project Modelling Stage - Model Implementation and Post Production Stage - Summary of Common Mistakes.*

A Brief History of Data Science

- Interest in data science careers has surged recently.
- Data scientists have diverse backgrounds, making it hard to define the field.
- Media frequently discusses "Data Science," "Big Data," and "Artificial Intelligence."
- Outsiders often see data science as extracting useful information from data.
- Big data skills are essential, including Hadoop for processing large datasets and Spark for speeding up processing with machine learning functions.
- These skills are needed for large-scale computing but not for deriving insights.
- Handling big data required advanced computing skills, now easier with cloud computing.
- The size of data is less important than how it's used.
- The term "data science" dates back before 2004.
- Machine learning and big data existed before Google, contrary to media impressions.
- Many data science techniques are based on decades of work by statisticians, computer scientists, and mathematicians.
- Key historical developments:
    - Early 19th century: Legendre and Gauss developed the least squares method for linear regression, initially used by physicists.
    - 1936: Fisher introduced linear discriminant analysis.
    - 1940s: Logistic regression became widely used.
    - 1970s: Nelder and Wedderburn developed the generalized linear model (GLM).
    - 1984: Breiman introduced Classification and Regression Trees (CART).
    - 1990s: Ensemble techniques like bagging emerged.
    - 2001: Breiman developed the random forest model and discussed two cultures in statistical modelling:
        - ✓ Stochastic data models
        - ✓ Algorithmic models
- Algorithmic models like random forests, GBM, and deep learning handle large, complex data better than traditional models.
- Python is now more popular than R in data science due to its versatility.

- Since 2000, data analysis has shifted from traditional models to machine learning and deep learning.
- John Tukey identified four forces driving data analysis in 1962:
  - Theories of math and statistics
  - Advances in computers and display devices
  - Increasing amounts of data
  - Emphasis on quantification across disciplines
- These forces continue to drive data science today.
- The development of computers has enabled the use of complex algorithmic models.
- The internet and the internet of things have generated vast amounts of commercial data.
- Industries recognize the value of exploiting data.
- Data science will likely be crucial in commercial life for decades.
- Applications of data science are expanding rapidly, benefiting from digitized information and internet distribution.
- Today, data science is applied in business, health, biology, social science, politics, and more.

Data Science Role and Skill Tracks

- A well-known Chinese parable tells the story of blind men describing an elephant by touch, each having a different perception based on the part they touch (trunk, ear, leg, side, tail, tusk).
- A parable about blind men describing an elephant illustrates different perspectives based on experience.
- Data science is similar, with professionals having varied views based on their backgrounds.
- Many people call themselves “Data Scientists” without necessary qualifications.
- Understanding of data science varies widely.
- "We don’t see things as they are, we see them as we are." - Anais Nin.
- Data science has three main skill tracks: engineering, analysis, and modelling/inference.
- Each track has specific skills, and different combinations define different roles.
- Data engineering is crucial and often overlooked; it is like the unseen iceberg.
- Companies need data engineers before data scientists to handle and prepare data.
- For small, formatted datasets, simple files like CSV or spreadsheets may be enough.
- As data grows in volume, variety, and velocity, data engineering becomes more complex.

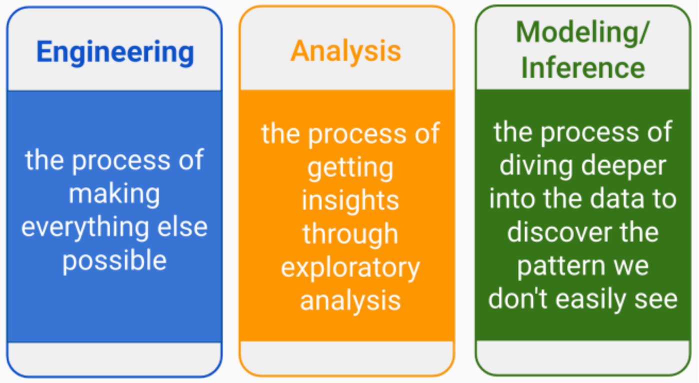

Engineering
the process of making everything else possible
Analysis
the process of getting insights through exploratory analysis
Modeling/ Inference
the process of diving deeper into the data to discover the pattern we don't easily see

FIGURE 1.1: Three tracks of data science

Engineering

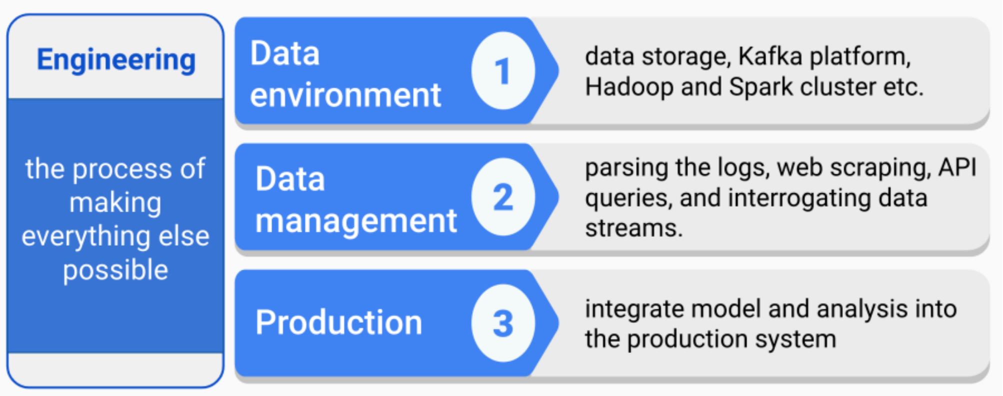

Engineering
the process of making everything else possible
Data environment
1
data storage, Kafka platform, Hadoop and Spark cluster etc.
Data management
2
parsing the logs, web scraping, API queries, and interrogating data streams.
Production
3
integrate model and analysis into the production system

FIGURE 1.2: Engineering track

- Foundation of Data Engineering:
  - Data engineering is essential for building data infrastructures and pipelines (Figure 1.2).
  - Traditionally involved major IT projects for local servers, including software, hardware, and ETL (extract, transform, and load) processes.
  - With cloud computing, storing and computing data is now primarily done on the cloud.
  - Modern data engineering focuses on software engineering with a primary emphasis on data flow, automation, modular code, and version control.

1. Data Environment:
   - Designing and setting up the environment to support data science workflows.
   - Includes cloud storage, Kafka platforms, Hadoop, and Spark clusters.
   - Varies by company based on data size, update frequency, analytics complexity, backend compatibility, and budget.
2. Data Management:
   - Involves automated data collection (e.g., parsing logs, web scraping, API queries, data stream interrogation).

- Includes constructing data schemas for analytics and modeling, ensuring data is correct, standardized, and documented.

3. Production:
    - Automating all data handling steps to integrate models or analysis into production systems.
    - Involves the entire pipeline from data access, preprocessing, modeling, to final deployment.
    - Requires monitoring the system with robust measures like error handling, fault tolerance, and graceful degradation to ensure smooth operation and user satisfaction.

Analysis

- Turning Raw Data into Insights:
- Fast and exploratory approach to uncover meaningful insights (Figure 1.3).
- Requires solid domain knowledge, efficient exploratory analysis, and compelling storytelling skills.

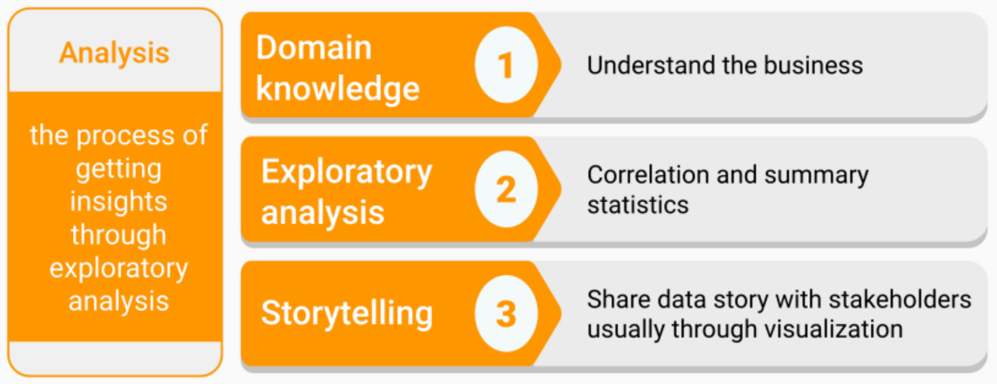

Analysis
the process of getting insights through exploratory analysis
Domain knowledge
1
Understand the business
Exploratory analysis
2
Correlation and summary statistics
Storytelling
3
Share data story with stakeholders usually through visualization

FIGURE 1.3: Analysis track

I. Domain Knowledge:
    - Understanding the organization or industry context is crucial.
    - Key questions include critical metrics, business questions, data types and representation, translating business needs into data problems, previous attempts and outcomes, accuracy-cost-time trade-offs, failure points, unaccounted factors, reasonable assumptions, and faulty ones.
    - Enables delivering results effectively and addressing the right problems with appropriate solutions.

2. Exploratory Analysis:
    - Focuses on exploration and discovery rather than rigorous conclusions.
    - Driven by correlation rather than causation, requiring extensive data examination.
    - Involves slicing and aggregating data to provide decision-makers with valuable insights.
    - Emphasizes avoiding overreaching conclusions beyond available data.

3. Storytelling:
    - Essential for communicating insights effectively to drive decision-making.
    - Involves data summarization, aggregation, and visualization.
    - Critical questions to address include identifying the audience, determining what information or actions to convey, and leveraging data to support key points.
    - Outputs typically include business-friendly reports or interactive dashboards.

Modeling/Inference

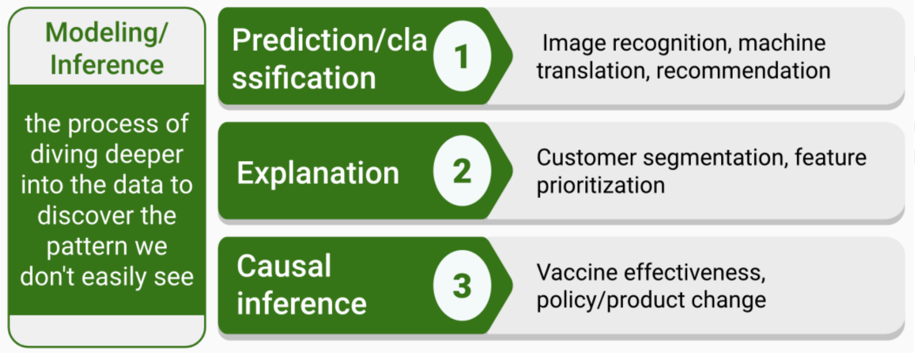

Modeling/ Inference
the process of diving deeper into the data to discover the pattern we don't easily see
Prediction/classification
1
Image recognition, machine translation, recommendation
Explanation
2
Customer segmentation, feature prioritization
Causal inference
3
Vaccine effectiveness, policy/product change

FIGURE 1.4: Modeling/inference track

- Overview of Modeling/Inference:
    - Focuses on uncovering hidden patterns within data that are not immediately apparent.
    - Often misunderstood; not exclusively reliant on complex machine learning models (Figure 1.4).

Three Key Objectives:

1. Prediction:
    - Forecasts based on historical data without necessarily understanding each variable's role.
    - Often utilizes black-box models like deep learning for tasks such as image recognition, machine translation, and recommendation systems.
    - These models operate primarily on associations rather than causal relationships.
2. Intervention:
    - Requires interpretability to understand the impact of changes.
    - Key question: "What happens if...?"
    - Example: Prioritizing features in product development.
    - Techniques include choice modeling and A/B testing.
3. Causal Inference:
    - Examines counterfactual scenarios when experiments are impractical.
    - Example: Exploring factors influencing divorce rates.
    - Uses methods like sample matching or pseudo-population creation.

Technical Methods:

- Supervised Learning:
  - Involves labeled data for training predictive models (e.g., regression and classification).
  - Optimization aims to minimize discrepancies between model outputs and actual responses.
- Unsupervised Learning:
  - Extracts patterns from data without predefined labels.
  - Functions include clustering, dimensionality reduction, and feature extraction.
- Customized Model Development:
- Tailors models to specific business problems or incorporates domain knowledge not covered by standard methods.
- Criteria for selecting techniques include data availability, scalability, and uniqueness of the problem context.

Criteria for Technique Selection:

- Is your data labeled? It is straightforward since supervised learning needs labeled data.
- Do you want to deploy your model at scale? There is a fundamental difference between building and deploying models. It is like the difference between making bread and making a bread machine. One is a baker who will mix and bake ingredients according to recipes to make a variety of bread. One is a machine builder who builds a machine to automate the process and produce bread at scale.
- Is your data easy to collect? One of the major sources of cost in deploying machine learning is collecting, preparing, and cleaning the data. Because model maintenance includes continuously collecting data to keep the model updated. If the data collection process requires too much human labor, the maintenance cost can be too high.
- Does your problem have a unique context? If so, you may not be able to find any off-the-shelf method that can directly apply to your question and need to customize the model.

Common Skills Across Roles:

- Data Preprocessing:
  - The process of converting raw data into clean data is essential regardless of the role in the data science team. Data cleaning tends to be the least enjoyable part of anyone's job.
  - Data preprocessing includes tasks such as schema enforcement, repairing broken records, aggregating data, and ensuring data consistency.

Data Engineering Perspective:

- Data engineers are responsible for extracting data from various sources and storing it in a data lake. They ensure data integrity and format data for downstream analysis.

Data Analyst and Scientist's Role:

- Data analysts and scientists spend a significant amount of time cleaning and preparing data for analysis. They address issues such as data format discrepancies, missing values, and preprocessing variables tailored to specific models.

Focus Areas in Data Science:

- Most professionals specialize in one track, with a small number proficient in multiple tracks.

What Kind of Questions Can Data Science Solve?

- Prerequisites
  - Introduction:
    - Data science is not a universal solution; it has limitations that must be acknowledged.
    - Transparency and honesty are crucial when determining if data analytics can provide answers.
    - Negotiation helps align expectations with stakeholders regarding what data science can realistically achieve.
  - Specificity of Questions:
    - Questions must be specific to be effectively analyzed using data science methods.
    - Example 1: "How can I increase product sales?"
    - Example 2: "Is the new promotional tool boosting annual sales of P1197 in Iowa and Wisconsin?"
    - Specific questions enable clear identification of variables and facilitate focused analysis.
  - Data Requirements:
    - The accuracy and relevance of data are paramount; irrelevant or inaccurate data can compromise analysis.
    - Example: High accuracy is crucial for variables like "new promotion" and "sales of P1197."
    - Data quality is more critical than quantity, although more data generally improves model performance if quality is assured.
- Problem Type
  - Description:
    - Summarizes and explores data using descriptive statistics and visualization.
    - Essential for data preprocessing and understanding initial insights.
    - Example questions: "What is the distribution of annual income?" "Are there outliers in the dataset?"
  - Comparison:
    - Compares different groups within a dataset to identify differences or similarities.
    - Uses statistical tests like t-tests, ANOVA, or chi-square tests.

  - Example questions: "Are there demographic differences in customer satisfaction across different business districts?"
- Clustering:
  - Identifies natural groupings in data without predefined labels.
  - Common algorithms include K-Means and Hierarchical Clustering.
  - Example questions: "How many distinct customer segments exist based on historical purchase patterns?"
- Classification:
  - Predicts categorical labels based on training data.
  - Utilizes classifiers such as logistic regression, decision trees, or support vector machines.
  - Example questions: "Will a customer likely purchase our product based on their demographic data?"
- Regression:
  - Predicts numerical outcomes based on historical data patterns.
  - Utilizes regression techniques like linear regression, polynomial regression, or neural networks.
  - Example questions: "What will be the temperature tomorrow based on historical weather data?"
- Optimization:
  - Seeks the best possible solution by adjusting controllable variables within given constraints.
  - Uses optimization algorithms such as linear programming, genetic algorithms, or simulated annealing.
  - Example questions: "What is the optimal allocation of resources to maximize profit in a supply chain?"

Structure of Data Science Team

Introduction:

- Over the past decade, companies across various sectors have seen a rapid increase in the volume, complexity, and accessibility of data. This surge has surpassed traditional statistical analysis and business intelligence capabilities.
- To effectively leverage this big data, organizations must decide whether to establish an internal data science team as a core competency or outsource these capabilities. The decision hinges on the strategic importance of data-driven insights to the business.
- Becoming a data-driven organization requires a commitment to identifying data needs across departments, establishing robust data infrastructure, and standardizing analytical processes. Off-the-shelf solutions often fail to adapt adequately to specific business contexts, making internal teams preferable in most cases.

Organizational Models:

1. Standalone Team:
    a. Data science operates as an autonomous unit, typically reporting directly to senior leadership or the CEO.
    b. Advantages:
        i. Autonomy enables the team to prioritize and address critical business problems independently.
        ii. Promotes knowledge sharing and professional growth among data scientists.
        iii. Signals to employees that data is a primary asset, aiding in talent attraction and retention.
    c. Challenges:
        i. Risk of isolation; successful data science outcomes depend on collaboration with engineers, product managers, and other stakeholders.
        ii. Requires strong leadership to bridge communication gaps and foster cross-functional collaboration.
2. Embedded Model:
    a. Data science team members report to a senior manager within another department, often engineering or operations.
    b. Advantages:
        i. Closer alignment of data science efforts with specific business applications and operational needs.
        ii. Flexibility in deploying data science resources across different departments or projects.
    c. Challenges:
        i. Potential disconnect between data scientists' professional growth and their technical leadership within the host department.
        ii. Difficulty in retaining top talent due to perceived secondary status compared to standalone teams.
3. Integrated Team:
    a. No centralized data science team; individual departments hire their data science professionals as needed.
    b. Example: A marketing analytics team led by a manager with strong analytical skills.
    c. Advantages:
        i. Seamless integration of data science expertise with specific business functions, ensuring insights directly impact operations.
        ii. Data scientists are valued members of their respective teams, with clear career paths and growth opportunities.
        iii. Rapid implementation of data-driven decisions due to immediate application of insights.
    d. Challenges:

i. Limited cross-functional knowledge sharing and innovation across departments.
ii. Potential difficulty in relocating talent across different business functions.
iii. Risk of talent retention issues without a centralized career development path for data scientists.

Considerations:

- *Organizational Stage:* The optimal team structure depends on the company's growth stage and strategic priorities.
- *Business Impact:* Evaluate how critical data science insights are to achieving business objectives.
- *Talent Retention:* Consider the impact of team structure on attracting and retaining top data science talent.
- *Collaboration:* Emphasize the importance of collaboration between data scientists, engineers, product managers, and other stakeholders for effective data utilization.
- *Career Development:* Provide clear career paths and growth opportunities for data science professionals to ensure long-term retention and organizational commitment.

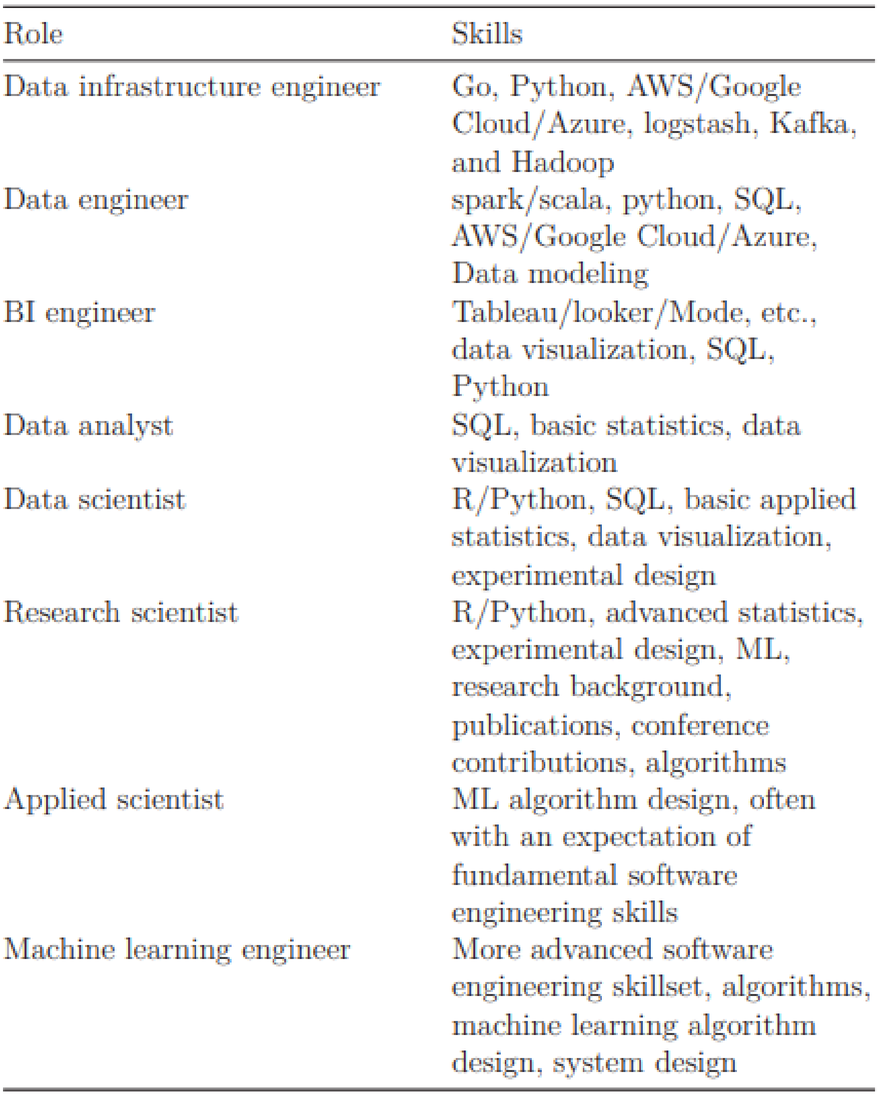
Conclusion:

- There is no one-size-fits-all approach to structuring data science teams. Each model offers unique advantages and challenges based on organizational context and strategic goals.
- Understanding where the data science team fits within the organization is crucial for maximizing its impact on business outcomes and fostering a data-driven culture.

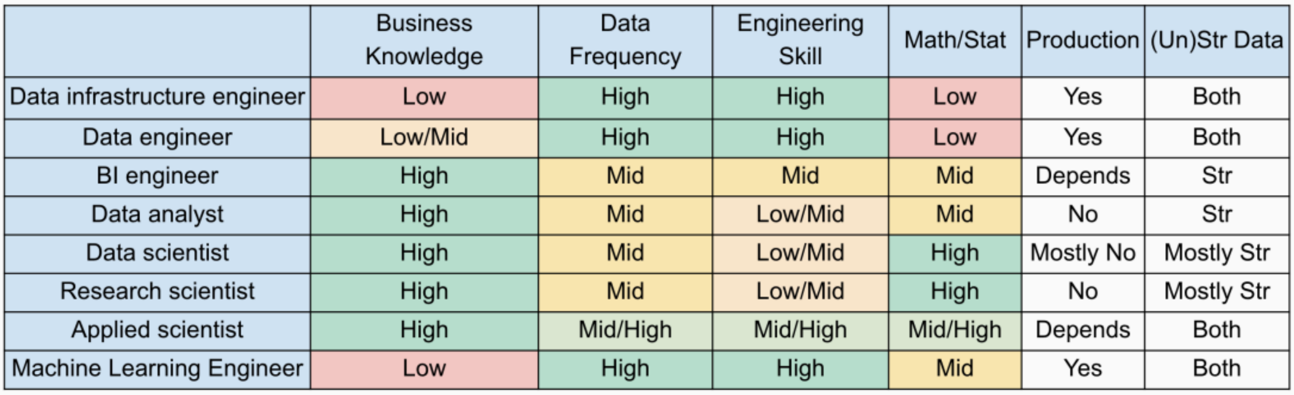
Data Science Roles

✦ Introduction:

- As businesses increasingly leverage data for decision-making, the field of data science has evolved, leading to specialized roles beyond the traditional "data scientist." Clear definitions of these roles are crucial for effective hiring and alignment with business objectives.



Role	Skills
Data infrastructure engineer	Go, Python, AWS/Google Cloud/Azure, logstash, Kafka, and Hadoop
Data engineer	spark/scala, python, SQL, AWS/Google Cloud/Azure, Data modeling
BI engineer	Tableau/looker/Mode, etc., data visualization, SQL, Python
Data analyst	SQL, basic statistics, data visualization
Data scientist	R/Python, SQL, basic applied statistics, data visualization, experimental design
Research scientist	R/Python, advanced statistics, experimental design, ML, research background, publications, conference contributions, algorithms
Applied scientist	ML algorithm design, often with an expectation of fundamental software engineering skills
Machine learning engineer	More advanced software engineering skillset, algorithms, machine learning algorithm design, system design



	Business Knowledge	Data Frequency	Engineering Skill	Math/Stat	Production	(Un)Str Data
Data infrastructure engineer	Low	High	High	Low	Yes	Both
Data engineer	Low/Mid	High	High	Low	Yes	Both
BI engineer	High	Mid	Mid	Mid	Depends	Str
Data analyst	High	Mid	Low/Mid	Mid	No	Str
Data scientist	High	Mid	Low/Mid	High	Mostly No	Mostly Str
Research scientist	High	Mid	Low/Mid	High	No	Mostly Str
Applied scientist	High	Mid/High	Mid/High	Mid/High	Depends	Both
Machine Learning Engineer	Low	High	High	Mid	Yes	Both

FIGURE 1.5: Different roles in data science and the skill requirements

✦ Overview of Data Science Roles:

1. Data Infrastructure Engineer:
    - *Responsibilities:* Works at the start of the data pipeline, ensuring smooth data ingestion and integration.

- *Skills:* Proficient in setting up data streaming with tools like Apache Kafka and integrating cloud services (AWS/GCP/Azure).
- *Focus:* Emphasizes data reliability and infrastructure setup rather than deep business insights.

2. Data Engineer:
   - *Responsibilities:* Transforms raw data from data lakes into structured formats suitable for analysis (data marts).
   - *Skills:* Designs data schemas, manages ETL processes using Hadoop/Spark, and ensures data quality and accessibility.
   - *Focus:* Bridges between data infrastructure and analytical needs, collaborates with data scientists on model deployment.
3. Business Intelligence (BI) Engineer:
   - *Responsibilities:* Develops and maintains automated dashboards and reporting pipelines.
   - *Skills:* Strong in SQL and data visualization tools, capable of writing production-level code for data pipelines.
   - *Focus:* Delivers actionable insights to business stakeholders through interactive dashboards and reports.
4. Data Analyst:
   - *Responsibilities:* Analyzes structured data sets to extract meaningful insights.
   - *Skills:* Proficient in SQL/R/Python for data manipulation and visualization.
   - *Focus:* Provides descriptive analytics and ad hoc reporting to support operational decision-making.
5. Data Scientist:
   - *Responsibilities:* Applies statistical analysis, machine learning, and experimental design to solve complex problems.
   - *Skills:* Expertise in predictive modeling, hypothesis testing, and data mining techniques.
   - *Focus:* Develops algorithms, conducts A/B testing, and translates data into actionable business insights.
6. Research Scientist:
   - *Responsibilities:* Conducts rigorous research, designs experiments, and investigates novel approaches to data analysis.
   - *Skills:* Strong background in scientific research methodologies and statistical analysis.
   - *Focus:* Publishes findings, contributes to academic or industry journals, and drives scientific innovation within the organization.
7. Applied Scientist:
   - *Responsibilities:* Bridges theoretical research with practical applications, implementing scalable solutions.
   - *Skills:* Applies scientific knowledge to real-world problems, proficient in coding and algorithm development.

- *Focus:* Translates research findings into actionable strategies, collaborates with cross-functional teams for solution implementation.

8. Machine Learning Engineer:
    - *Responsibilities:* Focuses on designing, implementing, and deploying machine learning models.
    - *Skills:* Strong software engineering skills, proficiency in model development, and deployment in production environments.
    - *Focus:* Works closely with data scientists to operationalize models, optimize algorithms, and ensure scalability and reliability.

✦ Conclusion:

- Each data science role plays a distinct part in the data lifecycle, from infrastructure management and data engineering to advanced analytics and research. Understanding these roles' nuances helps organizations build effective teams tailored to their data-driven goals and business needs.

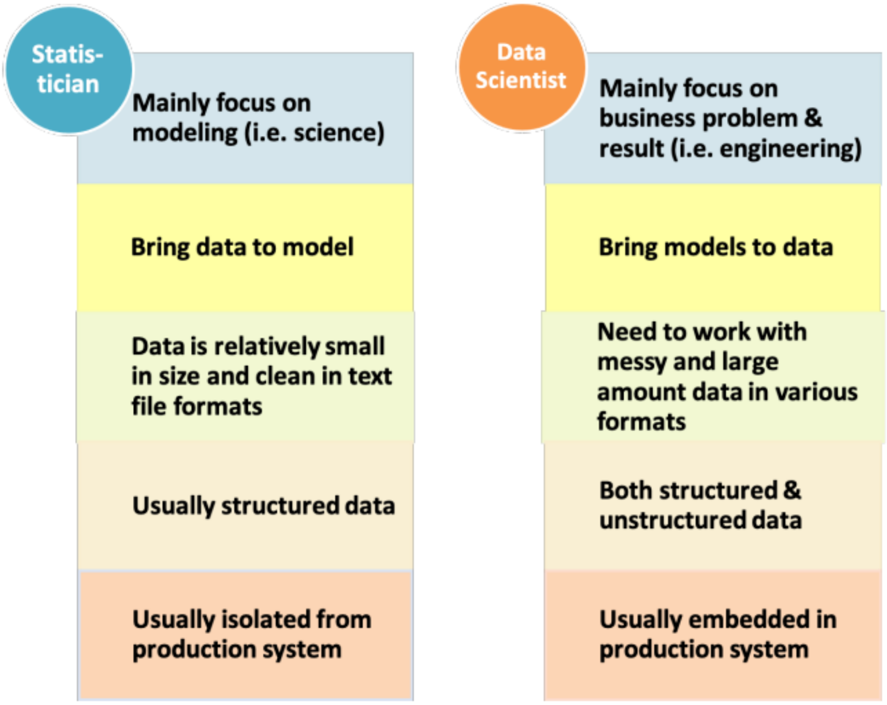
Comparison between Statistician and Data Scientist

✦ History and Career Longevity:

- Statistics as a scientific area can be traced back to 1749.
- Statisticians as a career have been around for hundreds of years with well-established theory and application.
- Data scientists have become an attractive career only in the last few years, driven by the increase in data size and variety beyond the traditional statistician's toolbox and the fast-growing computation power.

✦ Commonalities and Differences:

- Statisticians and data scientists have a lot in common, but there are also significant differences.



Statis-tician	Data Scientist
Mainly focus on modeling (i.e. science)	Mainly focus on business problem & result (i.e. engineering)
Bring data to model	Bring models to data
Data is relatively small in size and clean in text file formats	Need to work with messy and large amount data in various formats
Usually structured data	Both structured & unstructured data
Usually isolated from production system	Usually embedded in production system

FIGURE 2.1: Comparison of statistician and data scientist

- ✦ Nature of Data Handled:
  - Both statisticians and data scientists work closely with data.
  - Typical traditional statisticians work with well-formatted text files containing numbers (numerical variables) and labels (categorical variables).
  - The data size for statisticians is typically small enough to be loaded into a PC’s memory or saved on a PC’s hard disk.
- ✦ Data scientists deal with more varieties of data, including:
  - Well-formatted data stored in database systems with a size much larger than a PC’s memory or hard disk.
  - A huge amount of verbatim text, voice, image, and video data.
  - Real-time streaming data and other types of records.
- ✦ Statistical Inference and Modeling:
  - Statisticians have the unique power to make statistical inferences based on a small set of data.
  - Statisticians, especially in academia, spend most of their time developing models and don’t need to put much effort into data cleaning.
  - Due to the relative abundance of data recently, modeling is often a small part of the overall effort.
  - The active development by open-source communities has made fitting standard models not too far from button-pushing.
  - Data scientists in the industry spend a lot of time preprocessing and wrangling the data before feeding it into the model.
- ✦ Focus and Integration:

  - Data scientists often focus on delivering actionable results and sometimes need to deploy models to the production system.
  - The data available for model training for data scientists can be too large to be processed on a single computer.
  - Statisticians are usually not well integrated with the production system where data is obtained in real-time.
  - Data scientists are more embedded in the production system and closer to the data generation procedures.
- Summary:
  - Statisticians focus more on modeling and usually bring data to models.
  - Data scientists focus more on data and usually bring models to data.

Beyond Data and Analytics

- Roles in a Data Science Project:
  - Data scientists usually have a good sense of data and analytics, but data science projects are much more than that.
  - A data science project may involve people with different roles, especially in a large company:
    - The business owner or leader who identifies the business problem and value.
    - The data owner and computation resource/infrastructure owner from the IT department.
    - A dedicated policy owner to ensure the data and model are under model governance, security, and privacy guidelines and laws.
    - A dedicated engineering team to implement, maintain, and refresh the model.
    - A program manager to ensure the data science project fits into the overall technical program development and to coordinate all involved parties to set periodical tasks so that the project meets the preset milestones and results.
- Resource Allocation and Communication:
  - The entire team usually will have multiple rounds of discussion of resource allocation among groups (i.e., who pays for the data science project) at the beginning of the project and during the project.
  - Effective communication and in-depth domain knowledge about the business problem are essential requirements for a successful data scientist.
  - A data scientist may interact with people at various levels, from senior leaders who set the corporate strategies to front-line employees who do the daily work.
  - A data scientist needs to have the capability to view the problem from 10,000 feet above the ground and down to the detail to the very bottom.
  - To convert a business question into a data science problem, a data scientist needs to communicate using the language other people can understand and obtain the required information through formal and informal conversations.
- Project Cycle Involvement:

    - In the entire data science project cycle, including defining, planning, developing, and implementing, every step needs to get a data scientist involved to ensure the whole team can correctly determine the business problem and reasonably evaluate the business value and success.
    - Corporates are investing heavily in data science and machine learning, and there is a very high expectation of return for the investment.
- Realistic Goals and Timelines:
    - It is easy to set an unrealistic goal and inflated estimation for a data science project’s business impact.
    - The team’s data scientist should lead and navigate the discussions to ensure data and analytics, not wishful thinking, back the goal.
    - Many data science projects often over-promise in business value and are too optimistic on the timeline to delivery.
    - These projects eventually fail by not delivering the pre-set business impact within the promised timeline.
    - As data scientists, we need to identify these issues early and communicate with the entire team to ensure the project has a realistic deliverable and timeline.
- Collaboration with Data Owners and Infrastructure Team:
    - The data scientist team needs to work closely with data owners on different things, such as:
        - Identifying relevant internal and external data sources.
        - Evaluating the data’s quality and relevancy to the project.
        - Working closely with the infrastructure team to understand the computation resources (i.e., hardware and software) availability.
    - It is easy to create scalable computation resources through the cloud infrastructure for a data science project.
    - However, you need to evaluate the dedicated computation resources’ cost and make sure it fits the budget.
- Summary:
    - In summary, data science projects are much more than data and analytics.
    - A successful project requires a data scientist to lead many aspects of the project.

Three Pillars of Knowledge

- Analytics Knowledge and Toolsets:
    - A successful data scientist needs to have a strong technical background in data mining, statistics, and machine learning.
    - The in-depth understanding of modeling with insight about data enables a data scientist to convert a business problem to a data science problem.
    - Many chapters of this book are focusing on analytics knowledge and toolsets.
- Domain Knowledge and Collaboration:
    - A successful data scientist needs in-depth domain knowledge to understand the business problem well.

- For any data science project, the data scientist needs to collaborate with other team members.
- Communication and leadership skills are critical for data scientists during the entire project cycle, especially when there is only one scientist in the project.
- The scientist needs to decide the timeline and impact with uncertainty.

✦ (Big) Data Management and (New) IT Skills:

- The last pillar is about the computation environment and model implementation in a big data platform.
- It used to be the most difficult one for a data scientist with a statistics background (i.e., lacking computer science knowledge or programming skills).
- The good news is that with the rise of the big data platform in the cloud, it is easier for a statistician to overcome this barrier.
- The "Big Data Cloud Platform" chapter of this book will describe this pillar in detail.

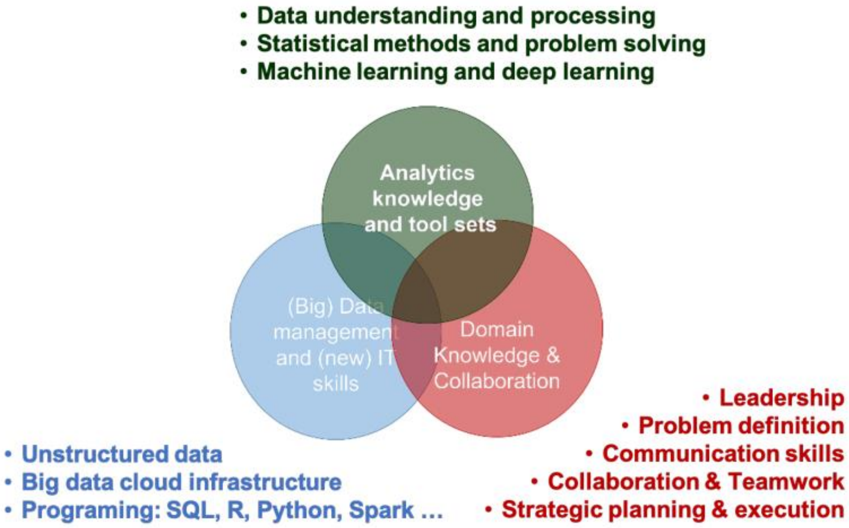

• Data understanding and processing
• Statistical methods and problem solving
• Machine learning and deep learning
Analytics knowledge and tool sets
(Big) Data management and (new) IT skills
Domain Knowledge & Collaboration
• Leadership
• Problem definition
• Communication skills
• Collaboration & Teamwork
• Strategic planning & execution
• Unstructured data
• Big data cloud infrastructure
• Programing: SQL, R, Python, Spark ...

FIGURE 2.2: Three pillars of knowledge

Data Science Project Cycle

A data science project has various stages. Many textbooks and blogs focus on one or two specific stages, and it is rare to see an end-to-end life cycle of a data science project. To get a good grasp of the end-to-end process requires years of real-world experience. Seeing a holistic picture of the whole cycle helps data scientists to better prepare for real-world applications. We will walk through the full project cycle in this section.

✦ Types of Data Science Projects

1. People often use data science projects to describe any project that uses data to solve a business problem, including traditional business analytics, data visualization, or machine learning modeling.
2. Here we limit our discussion of data science projects that involve data and some statistical or machine learning models and exclude basic analytics or visualization.

3. The business problem itself gives us the flavor of the project.
4. We can view data as the raw ingredient to start with, and the machine learning model makes the dish.
5. The types of data used and the final model development define the different kinds of data science projects.

- Offline and Online Data

1. There are offline and online data.
2. Offline data are historical data stored in databases or data warehouses.
3. With the development of data storage techniques, the cost to store a large amount of data is low.
4. Offline data are versatile and rich in general (for example, websites may track and keep each user’s mouse position, click and typing information while the user is visiting the website).
5. The data is usually stored in a distributed system, and it can be extracted in batch to create features used in model training.
6. Online data are real-time information that flows to models to make automatic actions.
7. Real-time data can frequently change (for example, the keywords a customer is searching for can change at any given time).
8. Capturing and using real-time online data requires the integration of a machine learning model to the production infrastructure.
9. It used to be a steep learning curve for data scientists not familiar with computer engineering, but the cloud infrastructure makes it much more manageable.
10. Based on the offline and online data and model properties, we can separate data science projects into three different categories as described below.

- Offline Training and Offline Application

1. This type of data science project is for a specific business problem that needs to be solved once or multiple times.
2. The dynamic and disruptive nature of this type of business problem requires substantial work every time.
3. One example of such a project is “whether a brand-new business workflow is going to improve efficiency.”
4. In this case, we often use internal/external offline data and business insight to build models.
5. The final results are delivered as a report to answer the specific business question.
6. It is similar to the traditional business intelligence project but with more focus on data and models.
7. Sometimes the data size and model complexity are beyond the capacity of a single computer.
8. Then we need to use distributed storage and computation.
9. Since the model uses historical data, and the output is a report, there is no need for real-time execution.
10. Usually, there is no run-time constraint on the machine learning model unless the model runs beyond a reasonable time frame, such as a few days.

11. We can call this type of data science project “offline training, offline application” project.

- Offline Training and Online Application

1. Another type of data science project uses offline data for training and applies the trained model to real-time online data in the production environment.
2. For example, we can use historical data to train a personalized advertisement recommendation model that provides a real-time ad recommendation.
3. The model training uses historical offline data.
4. The trained model then takes customers’ online real-time data as input features and runs the model in real-time to provide an automatic action.
5. The model training is very similar to the “offline training, offline application” project.
6. But to put the trained model into production, there are specific requirements.
7. For example, as features used in the offline training have to be available online in real-time, the model’s online run-time has to be short enough without impacting user experience.
8. In most cases, data science projects in this category create continuous and scalable business value as the model could run millions of times a day.

- Online Training and Online Application

1. For some business problems, it is so dynamic that even yesterday’s data is out of date.
2. In this case, we can use online data to train the model and apply it in real-time.
3. We call this type of data science project “online training, online application.”
4. This type of data science project requires high automation and low latency.

✦ Problem Formulation and Project Planning Stage

1. A data-driven and fact-based planning stage is essential to ensure a successful data science project.
2. With the recent big data and data science hype, there is a high demand for data science projects to create business value across different business sectors.
3. Usually, the leaders of an organization are those who initiate the data science project proposals.
4. This top-down style of data science projects typically have high visibility with some human and computation resources pre-allocated.
5. However, it is crucial to understand the business problem first and align the goal across different teams, including:
   - the business team, which may include members from the business operation team, business analytics, insight, and metrics reporting team;
   - the technology team, which may include members from the database and data warehouse team, data engineering team, infrastructure team, core machine learning team, and software development team;
   - the project, program, and product management team depending on the scope of the data science project.
6. To start the conversation, we can ask everyone in the team the following questions:
   a. What are the pain points in the current business operation?

b. What data are available, and how is the quality and quantity of the data?
c. What might be the most significant impacts of a data science project?
d. Is there any positive or negative impact on other teams?
e. What computation resources are available for model training and model execution?
f. Can we define key metrics to compare and quantify business value?
g. Are there any data security, privacy, and legal concerns?
h. What are the desired milestones, checkpoints, and timeline?
i. Is the final application online or offline?
j. Are the data sources online or offline?

6. It is likely to have a series of intense meetings and heated discussions to frame the project reasonably.
7. After the planning stage, we should be able to define a set of key metrics related to the project, identify some offline and online data sources, request needed computation resources, draft a tentative timeline and milestones, and form a team of data scientists, data engineers, software developers, project managers, and members from the business operation.
8. Data scientists should play a significant role in these discussions.
9. If data scientists do not lead the project formulation and planning, the project may not catch the desired timeline and milestones.

✦ Project Modeling Stage

1. Even though we already set some strategies, milestones, and timelines at the problem formulation and project planning stage, data science projects are dynamic.
2. There could be uncertainties along the road.
3. As a data scientist, communicating any newly encountered difficulties or opportunities during the modeling stage to the entire team is essential to keep the data science project progress.
4. Data cleaning, data wrangling, and exploratory data analysis are great starting points toward modeling with the available data source identified at the planning stage.
5. Meanwhile, abstracting the business problem to be a set of statistical and machine learning problems is an iterative process.
6. Business problems can rarely be solved by using just one statistical or machine learning model.
7. Using a sequence of methods to decompose the business problem is one of the critical responsibilities for a senior data scientist.
8. The process requires iterative rounds of discussions with the business and data engineering team based on each iteration’s new learnings.
9. Each iteration includes both data-related and model-related parts.

▪ Data Related Part

1. Data cleaning, data preprocessing, and feature engineering are related procedures that aim to create usable variables or features for statistical and machine learning models.
2. A critical aspect of data-related procedures is to make sure the data source we are using is a good representation of the situation where the final trained model will be applied.

3. The same representation is rarely possible, and it is okay to start with a reasonable approximation.
4. A data scientist must be clear on the assumptions and communicate the limitations of biased data with the team and quantify its impact on the application.
5. In the data-related part, sometimes the available data is not relevant to the business problem we want to solve.
6. We have to collect more and relevant data before modeling.

- Model Related Part

1. There are different types of statistical and machine learning models, such as supervised learning, unsupervised learning, and causal inference.
2. For each type, there are various algorithms, libraries, or packages readily available.
3. To solve a business problem, we sometimes need to piece together a few methods at the model exploring and developing stage.
4. This stage also includes model training, validation, and testing to ensure the model works well in the production environment (i.e., it can be generalized well and not causing overfitting).
5. The model selection follows Occam's razor, choosing the simplest among a set of compatible models.
6. Before we try complicated models, it is good to get some benchmarks by additional business rules, common-sense decisions, or standard models (such as random forest for classification and regression problems).

✦ Model Implementation and Post-Production Stage

1. For offline application data science projects, the end product is often a detailed report with model results and output.
2. However, for online application projects, a trained model is just halfway from the finish line.
3. The offline data is stored and processed in a different environment from the online production environment.
4. Building the online data pipeline and implementing machine learning models in a production environment requires lots of additional work.
5. Even though recent advances in cloud infrastructure lower the barrier dramatically, it still takes effort to implement an offline model in the online production system.
6. Before we promote the model to production, there are two more steps to go:
    1. Shadow mode
    2. A/B testing
7. A shadow mode is like an observation period when the data pipeline and machine learning models run as fully functional, but we only record the model output without any actions. Some people call it proof of concept (POC).
8. During the shadow mode, people frequently check the data pipeline and model and detect bugs such as a timeout, missing features, version conflict (for example, Python 2 vs. Python 3), data type mismatch, etc.
9. Once the online model passes the shadow mode, A/B testing is the next stage.

10. During A/B testing, all the incoming observations are randomly separated into two groups: control and treatment.
11. The control group will skip the machine learning model, while the treatment group is going through the machine learning model.
12. After that, people monitor a list of pre-defined key metrics during a specific period to compare the control and treatment groups.
13. The differences in these metrics determine whether the machine learning model provides business value or not.
14. Real applications can be complicated. For example, there can be multiple treatment groups, or hundreds, even thousands of A/B testing running by different teams at any given time in the same production environment.
15. Once the A/B testing shows that the model provides significant business value, we can put it into full production.
16. It is ideal that the model runs as expected and continues to offer scalable values.
17. However, the business can change, and a machine learning model that works now can break tomorrow, and features available now may not be available tomorrow.
18. We need a monitoring system to notify us when one or multiple features change.
19. When the model performance degrades below a pre-defined level, we need to fine-tune the parameters and thresholds, re-train the model with more recent data, add or remove features to improve model performance.
20. Eventually, any model will fail or retire at some time with a pre-defined model retirement plan.

✦ Project Cycle Summary

1. Data science end-to-end project cycle is a complicated process that requires close collaboration among many teams.
2. The data scientist, maybe the only scientist in the team, has to lead the planning discussion and model development based on data available and communicate key assumptions and uncertainties.
3. A data science project may fail at any stage, and a clear end-to-end cycle view of the project helps avoid some mistakes.

Common Mistakes in Data Science

▪ Common Systematic Mistakes in Data Science Projects:

- Many textbooks and online blogs focus on technical mistakes about machine learning models, algorithms, or theories, such as detecting outliers and overfitting.
- It is important to avoid these technical mistakes.
- There are common systematic mistakes across data science projects that are rarely discussed in textbooks.

✦ Problem Formulation Stage

▪ Challenges in Problem Formulation:

- The most challenging part of a data science project is problem formulation.

- Data science projects stem from pain points of the business.
- The draft version of the project's goal is relatively vague without much quantification or is the gut feeling of the leadership team.
- Often there are multiple teams involved in the initial project formulation stage, and they have different views.
- It is easy to have misalignment across teams, such as resource allocation, milestone deliverables, and timeline.

▪ Common Mistakes in Problem Formulation:

- Data science team members with technical backgrounds sometimes are not even invited to the initial discussion at the problem formulation stage.
- A lot of resources are spent on solving the wrong problem, the number one systematic common mistake in data science.
- People over-promise about business value all the time, another common mistake that will fail the project at the beginning.
- Leaders across many industries often have unrealistic high expectations of data science, especially during enterprise transformation.
- Unrealistic expectations are based on assumptions that are way off the chart without checking data availability, data quality, computation resource, and current best practices in the field.
- Project leaders sometimes ignore the data-driven voice of the data science team during the problem formulation stage.

▪ Solutions to Avoid Mistakes in Problem Formulation:

- Formulating a business problem into the right data science project requires an in-depth understanding of the business context, data availability and quality, computation infrastructure, and methodology to leverage the data to quantify business value.
- Having a strong data science leader with a broad technical background helps avoid mistakes.
- Letting data scientists coordinate and drive the problem formulation and set realistic goals based on data and business context.

✦ Project Planning Stage

▪ Challenges in Project Planning:

- Once the data science project is formulated correctly with a reasonable expectation of business value, the next step is to plan the project by allocating resources, setting up milestones and timelines, and defining deliverables.
- Project managers coordinate different teams involved in the project and use agile project management tools similar to those in software development.

▪ Common Mistakes in Project Planning:

- The project management team may not have experience with data science projects and hence fail to account for the uncertainties at the planning stage.
- People are often too optimistic about the timeline, not realizing that data exploration and data preparation may take 60% to 80% of the total time for a given data science project.
- People assume enough data is available without considering its quality, leading to being too optimistic about data availability and quality.
- There are unexpected efforts to bring the right and relevant data for a specific data science project.

▪ Solutions to Avoid Mistakes in Project Planning:

- Account for the "unexpected" work at the planning stage.
- Educate other team members and the leadership team about the time-consuming nature of data preprocessing and feature engineering.

✦ Project Modeling Stage

▪ Challenges in Project Modeling:

- Start looking at the data and fitting models.

▪ Common Mistakes in Project Modeling:

- Using unrepresentative data where the model trained using historical data may not generalize to the future.
- There is always a problem with biased or unrepresentative data.
- Overfitting and obsession with complicated models, sometimes leading to using the simplest among a set of compatible models with similar results.
- People are sometimes obsessed with complicated models instead of using simpler models that are better to generalize.
- If there is a fundamental gap between data and the business problem, the data scientist must make the tough decision to unplug the project.
- Data science projects usually have high visibility and may be initiated by senior leadership, leading to taking too long to fail.

▪ Solutions to Avoid Mistakes in Project Modeling:

- Use data closer to the situation where the model will apply and quantify the impact of model output in production.
- Use simpler models that are better to generalize and provide consistent business value.
- Prevent failing projects early to put valuable resources into other promising projects.

✦ Model Implementation and Post-Production Stage

▪ Challenges in Model Implementation and Post-Production:

- Implement the model after finding a model that works great for training and testing data, which can be alien work for data scientists without software engineering experience.

- ▪ Common Mistakes in Model Implementation and Post-Production:
- Missing shadow mode and A/B testing, assuming model performance at model training/testing stays the same in the production environment.
- Model performance nearly never performs the same in the production environment.
- People usually focus on model performance without paying too much attention to model execution time.
- Lack of computation capacity, engineering resources, or non-tech culture and environment leading to failure to scale in real-time applications.
- Missing necessary online checkups, leading to deteriorating model performance over time.

- ▪ Solutions to Avoid Mistakes in Model Implementation and Post-Production:
- Machine learning models in production should always go through shadow mode and A/B testing to evaluate performance.
- Set a monitoring dashboard and automatic alarms, create model tuning, re-training, and retirement plans.
- Ensure feature availability is crucial to running a real-time model.
- Regular checkups during the entire life of the model cycle from implementation to retirement.

✦ Summary of Common Mistakes

- The data science project is a combination of art, science, and engineering.
- A data science project may fail in different ways.
- The data science project can provide significant business value if we put data and business context at the center of the project.
- Get familiar with the data science project cycle and proactively identify and avoid these potential mistakes.
- Summary of the mistakes:
  - Solving the wrong problem
  - Overpromise on business value
  - Too optimistic about the timeline
  - Too optimistic about data availability and quality
  - Unrepresentative data
  - Overfitting and obsession with complicated models
  - Take too long to fail
  - Missing A/B testing
  - Fail to scale in real-time applications
  - Missing necessary online checkup