

Data Analysis Process : complete Lifecycle

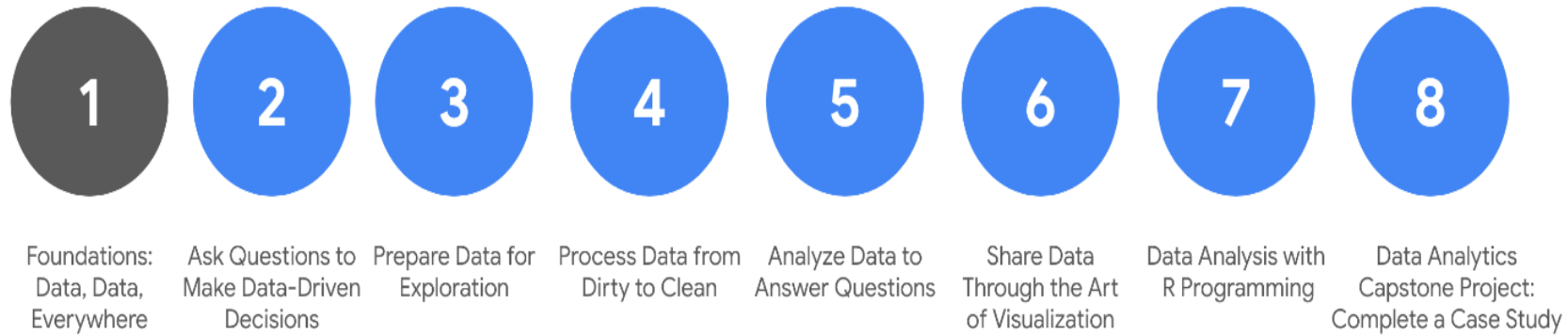
Dr. Sheetal Dhande-Dandge
Professor | Data Scientist
CSE-Dept
SIPNA COET



Data analysis life cycle

- **Data analysis life cycle**— the process of going from data to decision. Data goes through several phases as it gets created, consumed, tested, processed, and reused.
- With a life cycle model, all key team members can drive success by planning work both up front and at the end of the data analysis process.
- While the data analysis life cycle is well known among experts, there **isn't a single defined structure** of those phases.
- There might **not be one single architecture** that's uniformly followed by every data analysis expert, but there are some shared fundamentals in every data analysis process.

Data Analysis Process : complete Lifecycle



The process presented as part of the **Google Data Analytics** is one that will be valuable to you as you keep moving forward in your career:

Ask: Business Challenge/Objective/Question

Prepare: Data generation, collection, storage, and data management

Process: Data cleaning/data integrity

Analyze: Data exploration, visualization, and analysis

Share: Communicating and interpreting results

Act: Putting your insights to work to solve the problem

EMC's data analysis life cycle

- **EMC's data analysis life cycle**
- EMC Corporation's data analytics life cycle is cyclical with six steps:
 - Discovery
 - Pre-processing data
 - Model planning
 - Model building
 - Communicate results
 - Operationalize
- EMC Corporation is now Dell EMC. **This model, created by David Dietrich**, reflects the cyclical nature of real-world projects.
- The phases aren't static milestones; each step connects and leads to the next, and eventually repeats.
- Key questions help analysts test whether they have accomplished enough to move forward and ensure that teams have spent enough time on each of the phases and don't start modeling before the data is ready.
- It is a little different from the data analysis life cycle this program is based on, but it has some core ideas in common: the first phase is interested in discovering and asking questions; data has to be prepared before it can be analyzed and used; and then findings should be shared and acted on.

SAS's iterative life cycle

- An iterative life cycle was created by a company called **SAS**, a leading data analytics solutions provider. It can be used to produce repeatable, reliable, and predictive results:
 - Ask
 - Prepare
 - Explore
 - Model
 - Implement
 - Act
 - Evaluate
-
- The SAS model emphasizes the cyclical nature of their model by visualizing it as an infinity symbol.
 - Their life cycle has seven steps, many of which we have seen in the other models, like Ask, Prepare, Model, and Act. But this life cycle is also a little different; it includes a step after the act phase designed to help analysts evaluate their solutions and potentially return to the ask phase again.

Project-based data analytics life cycle

- A project-based data analytics life cycle has five simple steps:
 - Identifying the problem
 - Designing data requirements
 - Pre-processing data
 - Performing data analysis
 - Visualizing data
- This data analytics **project life cycle was developed by Vignesh Prajapati**. It doesn't include the sixth phase, or what we have been referring to as the Act phase. However, it still covers a lot of the same steps as the life cycles we have already described. It begins with identifying the problem, preparing and processing data before analysis, and ends with data visualization.

Big data analytics life cycle

- Authors **Thomas Erl, Wajid Khattak, and Paul Buhler** proposed a **big data analytics life cycle** in their book, **Big Data Fundamentals: Concepts, Drivers & Techniques**. Their life cycle suggests phases divided into nine steps:
 - Business case evaluation
 - Data identification
 - Data acquisition and filtering
 - Data extraction
 - Data validation and cleaning
 - Data aggregation and representation
 - Data analysis
 - Data visualization
 - Utilization of analysis results
- This life cycle appears to have three or four more steps than the previous life cycle models. But in reality, they have just broken down what we have been referring to as Prepare and Process into smaller steps. It emphasizes the individual tasks required for gathering, preparing, and cleaning data before the analysis phase.

How the data analysis process guides this program





Ask

- First up, the analysts needed to define what the project would look like and what would qualify as a successful result. So, to determine these things, they **asked** effective questions and collaborated with leaders and managers who were interested in the outcome of their people analysis. These were the kinds of questions they asked:
- What do you think new employees need to learn to be successful in their first year on the job?
- Have you gathered data from new employees before? If so, may we have access to the historical data?
- Do you believe managers with higher retention rates offer new employees something extra or unique?
- What do you suspect is a leading cause of dissatisfaction among new employees?
- By what percentage would you like employee retention to increase in the next fiscal year?



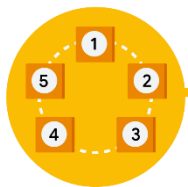
Act

- The last stage of the process for the team of analysts was to work with leaders within their company and decide how best to **implement changes and take actions** based on the findings. These were their recommendations:
- Standardize the hiring and evaluation process for employees based on the most efficient and transparent practices.
- Conduct the same survey annually and compare results with those from the previous year.
- A year later, the same survey was distributed to employees. Analysts anticipated that a comparison between the two sets of results would indicate that the action plan worked. Turns out, the changes improved the retention rate for new employees and the actions taken by leaders were successful!



Prepare

- It all started with solid **preparation**. The group built a timeline of three months and decided how they wanted to relay their progress to interested parties. Also during this step, the analysts identified what data they needed to achieve the successful result they identified in the previous step - in this case, the analysts chose to gather the data from an online survey of new employees. These were the things they did to prepare:
- They developed specific questions to ask about employee satisfaction with different business processes, such as hiring and onboarding, and their overall compensation.
- They established rules for who would have access to the data collected - in this case, anyone outside the group wouldn't have access to the raw data, but could view summarized or aggregated data. For example, an individual's compensation wouldn't be available, but salary ranges for groups of individuals would be viewable.
- They finalized what specific information would be gathered, and how best to present the data visually. The analysts brainstormed possible project- and data-related issues and how to avoid them.



Process

- The group sent the survey out. Great analysts know how to respect both their data and the people who provide it. Since employees provided the data, it was important to make sure all employees gave their consent to participate. The data analysts also made sure employees understood how their data would be **collected, stored, managed, and protected**. Collecting and using data ethically is one of the responsibilities of data analysts. In order to maintain confidentiality and protect and store the data effectively, these were the steps they took:
- They restricted access to the data to a limited number of analysts.
- They cleaned the data to make sure it was complete, correct, and relevant. Certain data was aggregated and summarized without revealing individual responses.
- They uploaded raw data to an internal data warehouse for an additional layer of security.



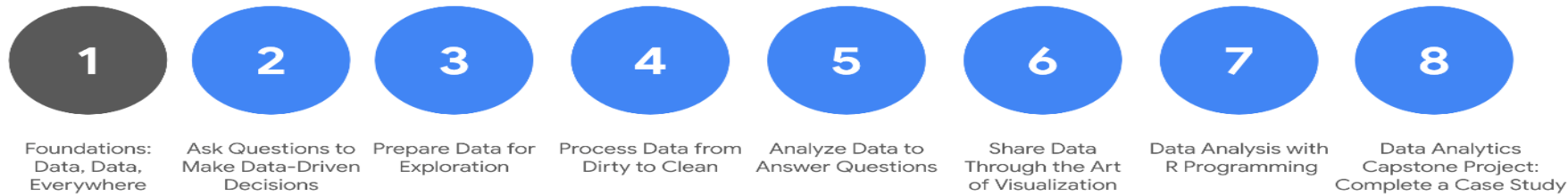
Analyze

- Then, the analysts did what they do best: analyze! From the completed surveys, the data analysts **discovered** that an employee's experience with certain processes was a key indicator of overall job satisfaction. These were their findings:
- Employees who experienced a long and complicated hiring process were most likely to leave the company.
- Employees who experienced an efficient and transparent evaluation and feedback process were most likely to remain with the company.
- The group knew it was important to **document** exactly what they found in the analysis, no matter what the results. To do otherwise would diminish trust in the survey process and reduce their ability to collect truthful data from employees in the future.



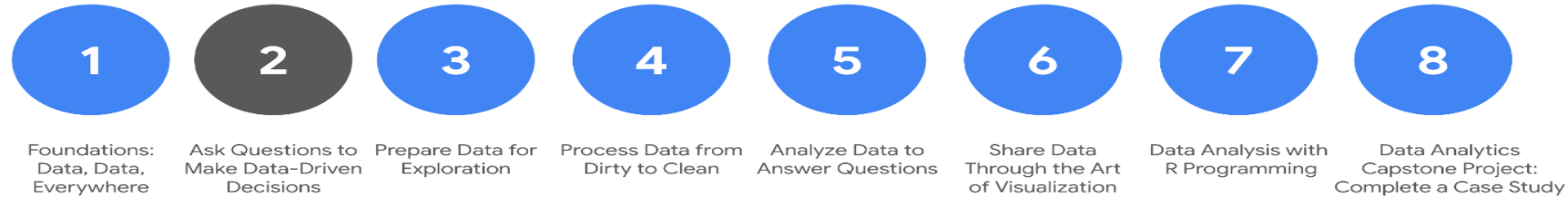
Share

- Just as they made sure the data was carefully protected, the analysts were also careful **sharing the report**. This is how they shared their findings:
- They shared the report with managers who met or exceeded the minimum number of direct reports with submitted responses to the survey.
- They presented the results to the managers to make sure they had the full picture.
- They asked the managers to personally deliver the results to their teams.
- This process gave managers an opportunity to **communicate the results** with the right context. As a result, they could have productive team conversations about next steps to improve employee engagement.



Foundations: Data, Data, Everywhere

- **Introducing data analytics:** Data helps us make decisions, in everyday life and in business.
- **Thinking analytically:** Data analysts balance many different roles in their work. You will also explore analytical thinking and how it relates to data-driven decision making.
- **Exploring the wonderful world of data:** Data has its own life cycle, and data analysts use an analysis process that cuts across and leverages this life cycle.
- **Setting up a data toolbox:** Spreadsheets, query languages, and data visualization tools are all a big part of a data analyst's job.
- **Discovering data career possibilities:** All kinds of businesses value the work that data analysts do. you will examine different types of businesses and the jobs and tasks that analysts do for them.
- **Completing the session Challenge:** At the end of this session, you will be able to put everything you have learned into perspective.



Ask Questions to Make Data-Driven Decisions

Asking effective questions: To do the job of a data analyst, you need to ask questions and problem-solve. you'll

- check out some common analysis problems and how analysts solve them. You'll also learn about effective questioning techniques that can help guide your analysis.

Making data-driven decisions: In analytics, data drives decision making. In this part of the course, you'll explore data of all kinds and its impact on decision making. You'll also learn how to share your data through reports and dashboards.

Mastering spreadsheet basics: Spreadsheets are an important data analytics tool. In this part of the course, you'll learn both why and how data analysts use spreadsheets in their work. You'll also explore how structured thinking can help analysts better understand problems and come up with solutions.

Always remembering the stakeholder: Successful data analysts learn to balance needs and expectations. In this part of the course, you'll learn strategies for managing the expectations of stakeholders while establishing clear communication with your team to achieve your objectives.







Six problem types

Data analysts typically work with six problem types

- Data analytics is so much more than just plugging information into a platform to find insights.
- It is about solving problems. To get to the root of these problems and find practical solutions, there are lots of opportunities for creative thinking.
- No matter the problem, the first and most important step is understanding it.
- From there, it is good to take a problem-solver approach to your analysis to help you decide what information needs to be included, how you can transform the data, and how the data will be used.

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 







- **Making predictions**

- A company that wants to know the best advertising method to bring in new customers is an example of a problem requiring analysts to make predictions. Analysts with data on location, type of media, and number of new customers acquired as a result of past ads can't guarantee future results, but they can help predict the best placement of advertising to reach the target audience.

- Regression

- Classification

Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 







- **Categorizing things**

- An example of a problem requiring analysts to categorize things is a company's goal to improve customer satisfaction. Analysts might classify customer service calls based on certain keywords or scores. This could help identify top-performing customer service representatives or help correlate certain actions taken with higher customer satisfaction scores.

- Logistic Regression

- Decision tree







Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

- **Spotting something unusual**

- A company that sells smart watches that help people monitor their health would be interested in designing their software to spot something unusual. Analysts who have analyzed aggregated health data can help product developers determine the right algorithms to spot and set off alarms when certain data doesn't trend normally.







Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

- **Identifying themes**

- User experience (UX) designers might rely on analysts to analyze user interaction data. Similar to problems that require analysts to categorize things, usability improvement projects might require analysts to identify themes to help prioritize the right product features for improvement. Themes are most often used to help researchers explore certain aspects of data. In a user study, user beliefs, practices, and needs are examples of themes.
- By now you might be wondering if there is a difference between categorizing things and identifying themes. The best way to think about it is: categorizing things involves assigning items to categories; identifying themes takes those categories a step further by grouping them into broader themes.







Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

- **Discovering connections**

- A third-party logistics company working with another company to get shipments delivered to customers on time is a problem requiring analysts to discover connections. By analyzing the wait times at shipping hubs, analysts can determine the appropriate schedule changes to increase the number of on-time deliveries.







Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

- **Finding patterns**

- Minimizing downtime caused by machine failure is an example of a problem requiring analysts to find patterns in data. For example, by analyzing maintenance data, they might discover that most failures happen if regular maintenance is delayed by more than a 15-day window.

Six problem types

1. Making predictions 	2. Categorizing things 	3. Spotting something unusual 
4. Identifying themes 	5. Discovering connections 	6. Finding patterns 

• Key takeaway






- As you move through this session, you will develop a sharper eye for problems and you will practice thinking through the problem types when you begin your analysis. This method of problem solving will help you figure out solutions that meet the needs of all stakeholders.

More about SM ART questions

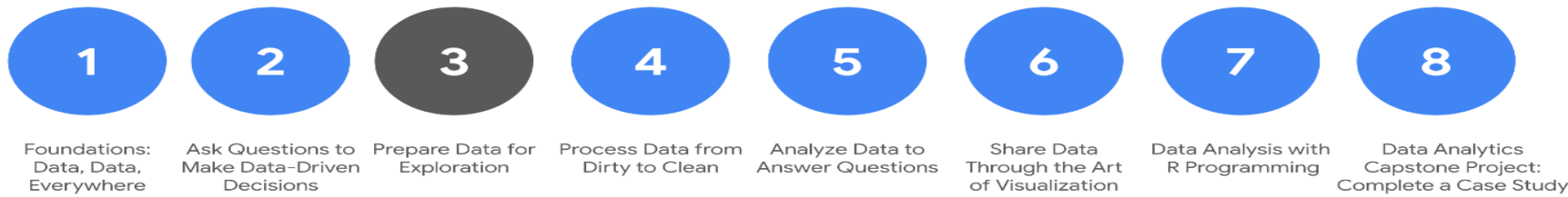
Asking the right questions can help spark the innovative ideas that so many businesses are hungry for these days.

The same goes for data analytics. No matter how much information you have or how advanced your tools are, your data won't tell you much if you don't start with the right questions. Think of it like a detective with tons of evidence who doesn't ask a key suspect about it. Coming up, you will learn more about how to ask highly effective questions, along with certain practices you want to avoid.

SMART

				
S-specific	M-easurable	A-ction-oriented	R-elevant	T-ime-bound
Is the question specific? Does it address the problem? Does it have context? Will it uncover a lot of the information you need?	Will the question give you answers that you can measure?	Will the answers provide information that helps you devise some type of action plan?	Is the question about the particular problem you are trying to solve?	Are the answers relevant to the specific time being studied?

- Here's an example that breaks down the thought process of turning a problem question into one or more SMART questions using the SMART method: **What features do people look for when buying a new car?**
- Specific:** Does the question focus on a particular car feature?
- Measurable:** Does the question include a feature rating system?
- Action-oriented:** Does the question influence creation of different or new feature packages?
- Relevant:** Does the question identify which features make or break a potential car purchase?
- Time-bound:** Does the question validate data on the most popular features from the last three years?
- Questions should be **open-ended**. This is the best way to get responses that will help you accurately qualify or disqualify potential solutions to your specific problem. So, based on the thought process, possible SMART questions might be:
 - On a scale of 1-10 (with 10 being the most important) how important is your car having four-wheel drive?
 - What are the top five features you would like to see in a car package?
 - What features, if included with four-wheel drive, would make you more inclined to buy the car?
 - How much more would you pay for a car with four-wheel drive?
 - Has four-wheel drive become more or less popular in the last three years?



Prepare Data for Exploration

- **Understanding data types and structures:** We all generate lots of data in our daily lives. In this part of the session, you will check out how we generate data and how analysts decide which data to collect for analysis. You'll also learn about structured and unstructured data, data types, and data formats as you start thinking about how to prepare your data for exploration.
- **Understanding bias, credibility, privacy, ethics, and access:** When data analysts work with data, they always check that the data is unbiased and credible. In this part of the course, you will learn how to identify different types of bias in data and how to ensure credibility in your data. You will also explore open data and the relationship between and importance of data ethics and data privacy.
- **Databases: Where data lives:** When you are analyzing data, you will access much of the data from a database. It's where data lives. In this part of the course, you will learn all about databases, including how to access them and extract, filter, and sort the data they contain. You will also check out metadata to discover the different types and how analysts use them.
- **Organizing and protecting your data:** Good organization skills are a big part of most types of work, and data analytics is no different. In this part of the course, you will learn the best practices for organizing data and keeping it secure. You will also learn how analysts use file naming conventions to help them keep their work organized.
- **Engaging in the data community (optional):** Having a strong online presence can be a big help for job seekers of all kinds. In this part of the course, you will explore how to manage your online presence. You will also discover the benefits of networking with other data analytics professionals.

Use the flowchart below if data collection relies heavily on how much time you have:

How the data will be collected

Decide if you will collect the data using your own resources or receive (and possibly purchase it) from another party. Data that you collect yourself is called first-party data.

Data sources

If you don't collect the data using your own resources, you might get data from second-party or third-party data providers. Second-party data is collected directly by another group and then sold. Third-party data is sold by a provider that didn't collect the data themselves. Third-party data might come from a number of different sources.

Solving your business problem

Datasets can show a lot of interesting information. But be sure to choose data that can actually help solve your problem question. For example, if you are analyzing trends over time, make sure you use time series data — in other words, data that includes dates.

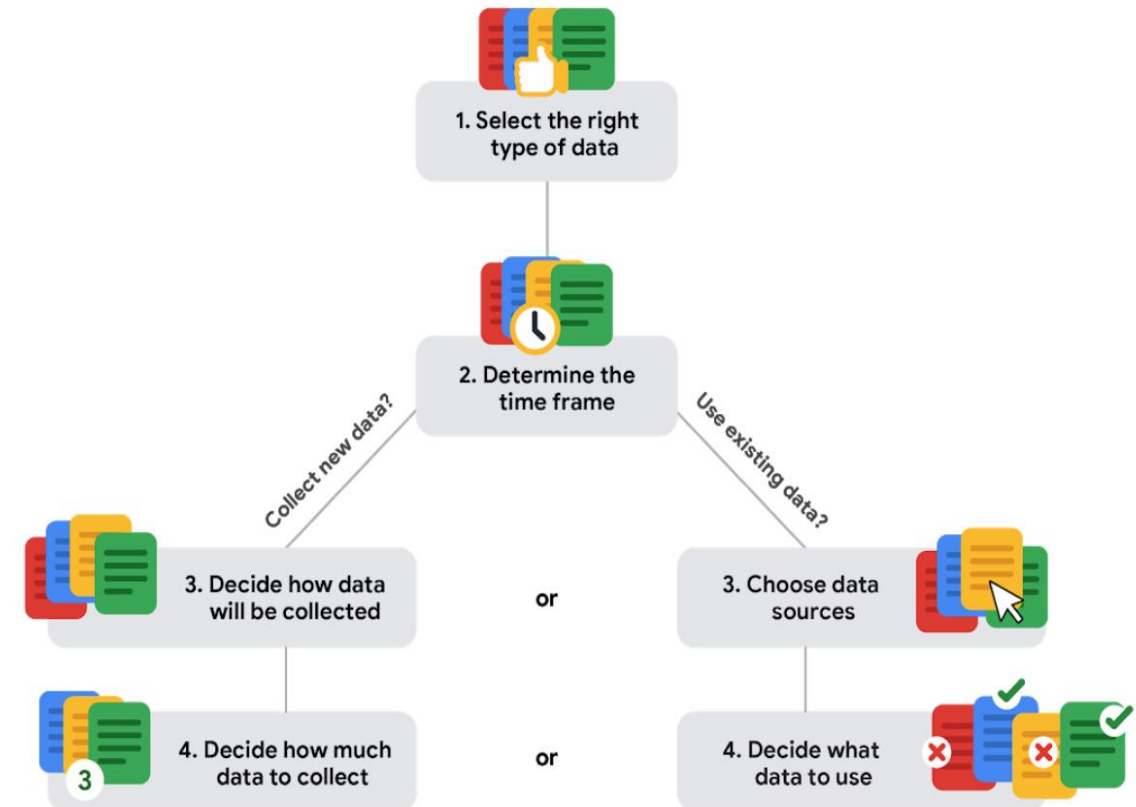
How much data to collect

If you are collecting your own data, make reasonable decisions about sample size. A random sample from existing data might be fine for some projects. Other projects might need more strategic data collection to focus on certain criteria. Each project has its own needs.

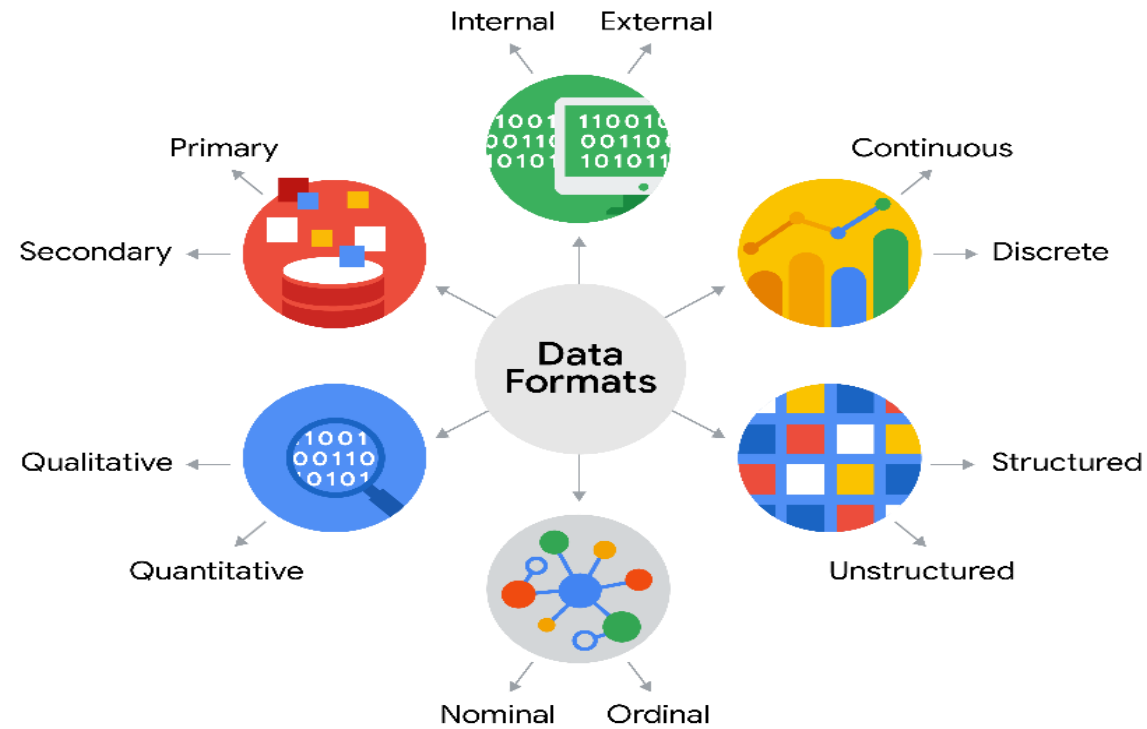
Time frame

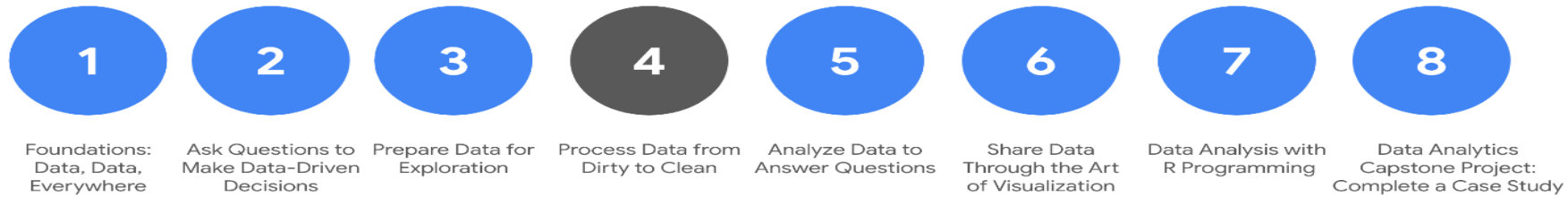
If you are collecting your own data, decide how long you will need to collect it, especially if you are tracking trends over a long period of time. If you need an immediate answer, you might not have time to collect new data. In this case, you would need to use historical data that already exists.

Data collection considerations



Data formats in practice

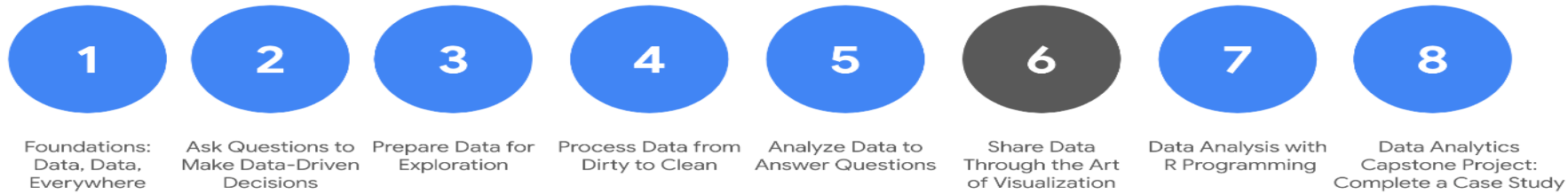




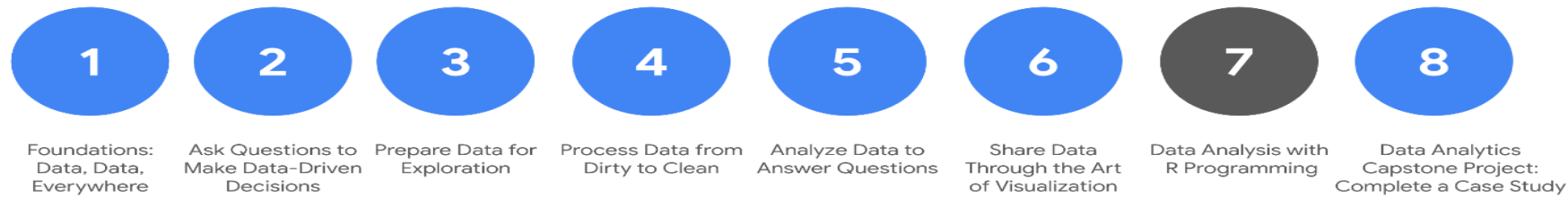
- **Ensuring data integrity.** Data integrity is necessary to ensure a successful analysis. In this part of the session, you will explore methods and steps that analysts take to check data for integrity. This includes knowing what to do when you have an insufficient amount of data. You will also learn about sample size, avoiding sample bias, and using random samples. All of these measures also help to ensure a successful data analysis.
- **Understanding clean data.** Every data analyst wants clean data to work with when performing an analysis. In this part of the session, you will learn the difference between clean and dirty data. You will practice data cleaning techniques in spreadsheets and other tools.
- **Cleaning data using SQL.** Knowing a variety of ways to clean data can make an analyst's job much easier. In this part of the course, you will use SQL to clean data from databases. You will explore how SQL queries and functions can be used to clean and transform your data before an analysis.
- **Verifying and reporting cleaning results.** Cleaning data is an important step in the data analysis process. In this part of the course, you will verify that data is clean and report data cleaning results. With verified clean data, you will be ready for the next step in the data analysis process.



- Analyse Data to Answer Questions
- **Organizing data to begin analysis.** Organizing data makes the data easier to use in an analysis. In this part of the course, you will learn the importance of organizing your data with sorting and filtering. You will explore organizing data in both spreadsheets and with SQL queries and temporary tables.
- **Formatting and adjusting your data.** As you move closer to analyzing your data, you will want to have the data formatted and ready to go. In this part of the course, you will learn all about converting and formatting data, including how to use SQL queries to combine data. You will also discover the value of feedback and support from your colleagues and how it can lead to new insights that you can apply to your work.
- **Aggregating data for analysis.** During an analysis, you might need to combine data to gain insights and complete business objectives. In this part of the course, you will explore the functions, procedures, and syntax to combine, or aggregate data. You will learn how to combine data within multiple cells in spreadsheets, and within multiple database tables using SQL queries.
- **Performing data calculations.** Calculations are one of the more common tasks that data analysts perform during an analysis. In this part of the course, you will explore formulas, functions, and pivot tables in spreadsheets and SQL queries. All of these are used in data calculations. You will also learn about the benefits of using SQL to manage temporary database tables.



- **Data visualization:** Data visualization is in many ways the culmination of the data analysis process. In this part of the course, you will be introduced to the concepts involved in data visualization. You will learn about accessibility, design thinking, and other factors that play a role in visualizing the data in your analysis.
- **Data visualizations with Tableau:** Tableau is a tool that can help analysts create effective data visualizations. In this part of the course, you will learn all about Tableau and its uses. You will also explore the importance of creativity and clarity while visualizing your findings appropriately.
- **Stories about your data:** Connecting your objective with your data through insights is essential to good data storytelling. In this part of the course, you will learn about data-driven stories and their attributes. You will also gain an understanding of how to use Tableau to create dashboards and dashboard filters.
- **Developing presentations and slideshows:** In this part of the course, you will discover how to give an effective presentation about your data analysis. You will consider all aspects of your analysis when creating a presentation and learn how to use multiple data sources in the data visualizations you will share. In addition, you will learn how to anticipate potential limitations and questions that might arise and how to provide useful answers to stakeholders.



- **Data Analysis with R Programming**

- RStudio—an integrated developer environment (IDE) for R that you will use to create advanced data visualizations with lots of detail. R makes it easier to present your data with beautiful, artistic style. A few other advantages of R include its:
 - **Popularity:** R is frequently used for data analysis
 - **Tools:** R has a convenient library of ready-to-use tools for data cleaning and analysis
 - **Focus:** R was created with statistics in mind; data analysts can conveniently use a rich library of statistical routines
 - **Adaptability:** R adapts well for use in both machine learning and data analysis projects
 - **Availability:** R is an open source programming language

The R-versus-Python debate

Languages	R	Python
Common features	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support	- Open-source - Data stored in data frames - Formulas and functions readily available - Community for code development and support
Unique advantages	- Data manipulation, data visualization, and statistics packages - "Scalpel" approach to data: <i>find packages to do what you want with the data</i>	- Easy syntax for machine learning needs - Integrates with cloud platforms like Google Cloud, Amazon Web Services, and Azure
Unique challenges	- Inconsistent naming conventions make it harder for beginners to select the right functions - Methods for handling variables may be a little complex for beginners to understand	- Many more decisions for beginners to make about data input/output, structure, variables, packages, and objects - "Swiss army knife" approach to data: <i>figure out a way to do what you want with the data</i>

- For more information on comparing R and Python, refer to these resources:
- [R versus Python, a comprehensive guide for data professionals](#): This article is written by a data professional with extensive experience using both languages and provides a detailed comparison.
- [R versus Python, an objective comparison](#): This article provides a comparison of the languages using examples of code use.
- [R versus Python: What's the best language for data science?](#): This blog article provides RStudio's perspective on the R vs. Python debate.

THANK YOU