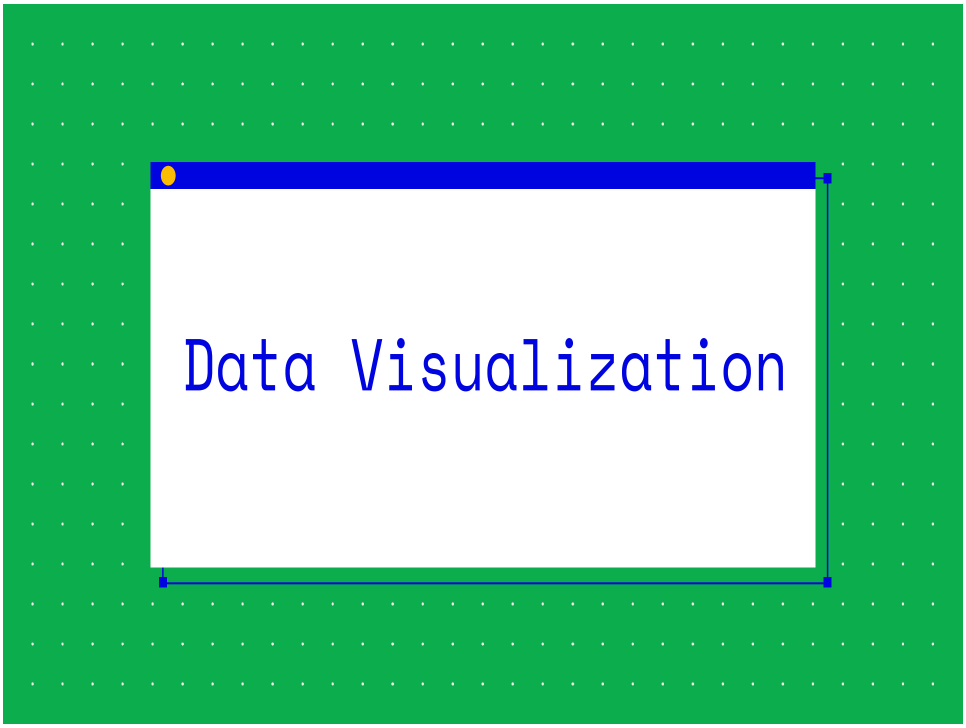
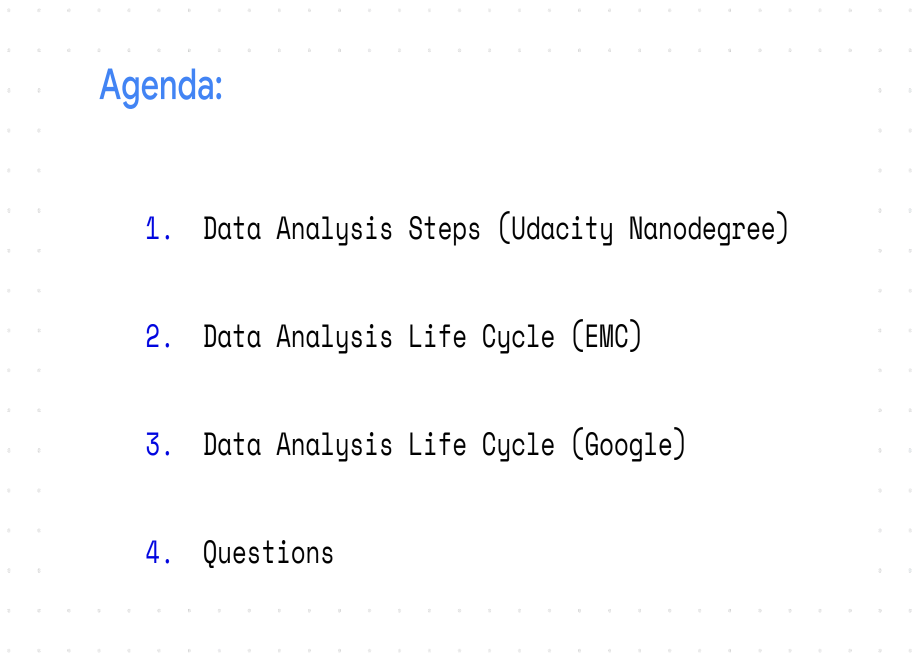


# Data Visualization



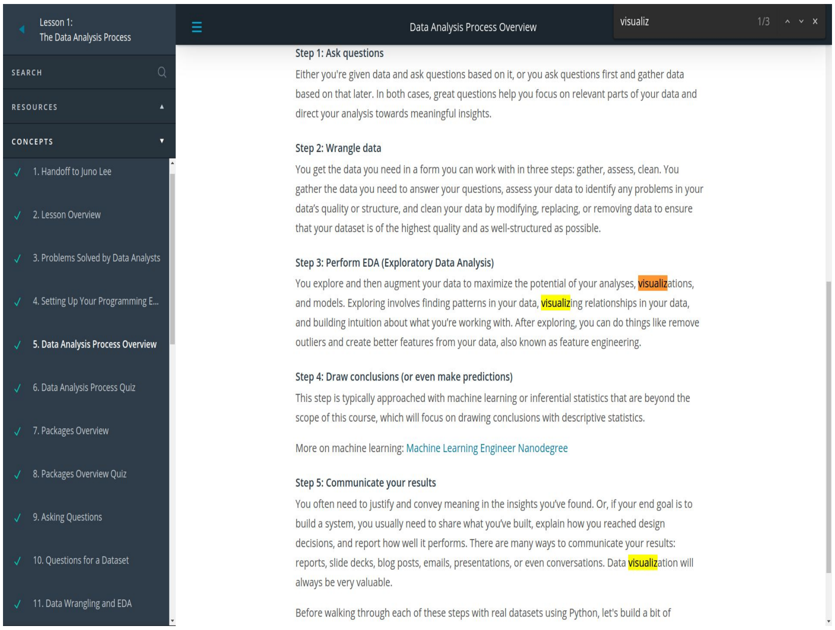
Agenda:

1. Data Analysis Steps (Udacity Nanodegree)
2. Data Analysis Life Cycle (EMC)
3. Data Analysis Life Cycle (Google)
4. Questions

Here is a quick overview of what we'll cover today.

# Data Analysis Steps (Udacity ND)

# Data Analysis Steps (From Udacity Nanodegree) - 1


Lesson 1:
The Data Analysis Process
SEARCH
RESOURCES
CONCEPTS
1. Handoff to Juno Lee
2. Lesson Overview
3. Problems Solved by Data Analysts
4. Setting Up Your Programming E...
5. Data Analysis Process Overview
6. Data Analysis Process Quiz
7. Packages Overview
8. Packages Overview Quiz
9. Asking Questions
10. Questions for a Dataset
11. Data Wrangling and EDA
Data Analysis Process Overview
visualiz
1/3

**Step 1: Ask questions**

Either you're given data and ask questions based on it, or you ask questions first and gather data based on that later. In both cases, great questions help you focus on relevant parts of your data and direct your analysis towards meaningful insights.

**Step 2: Wrangle data**

You get the data you need in a form you can work with in three steps: gather, assess, clean. You gather the data you need to answer your questions, assess your data to identify any problems in your data's quality or structure, and clean your data by modifying, replacing, or removing data to ensure that your dataset is of the highest quality and as well-structured as possible.

**Step 3: Perform EDA (Exploratory Data Analysis)**

You explore and then augment your data to maximize the potential of your analyses, visualizations, and models. Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you're working with. After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.

**Step 4: Draw conclusions (or even make predictions)**

This step is typically approached with machine learning or inferential statistics that are beyond the scope of this course, which will focus on drawing conclusions with descriptive statistics.

More on machine learning: Machine Learning Engineer Nanodegree

**Step 5: Communicate your results**

You often need to justify and convey meaning in the insights you've found. Or, if your end goal is to build a system, you usually need to share what you've built, explain how you reached design decisions, and report how well it performs. There are many ways to communicate your results: reports, slide decks, blog posts, emails, presentations, or even conversations. Data visualization will always be very valuable.

Before walking through each of these steps with real datasets using Python, let's build a bit of

# Data Analysis Steps (From Udacity Nanodegree) - 2

1. **Question**
2. **Wrangle**
3. **Explore**
4. **Draw Conclusions**
5. **Communicate**

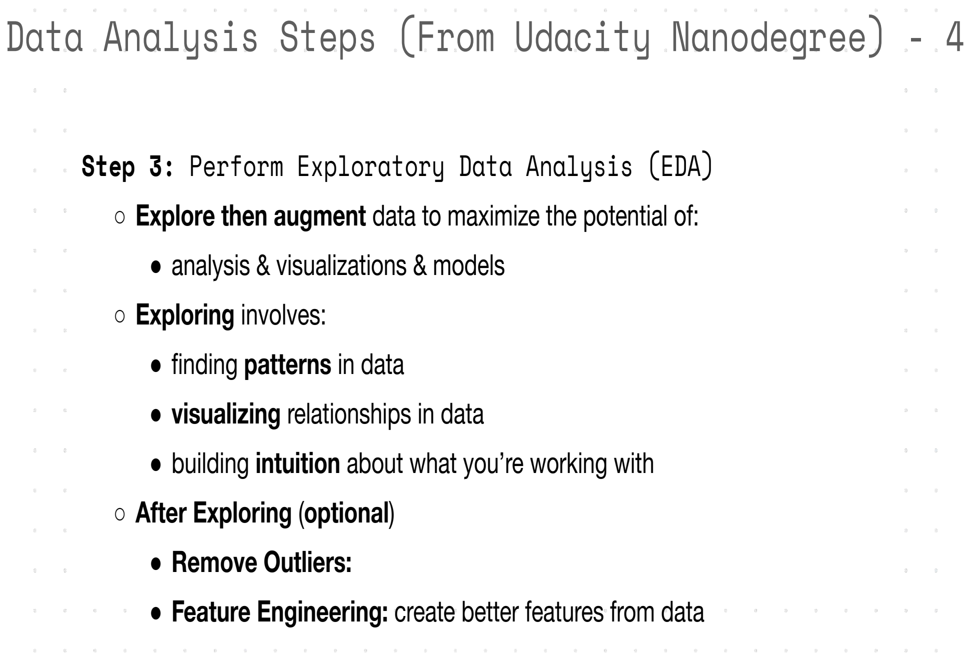
# Data Analysis Steps (From Udacity Nanodegree) - 3

**Step 1:** Ask Questions

- Given data then ask questions, or
- Ask questions then **gather** data

**Step 2:** Wrangle Data

a. **Gather** data to answer question
b. **Assess** data to identify any problems in your **data's quality** or structure
c. **Clean** data by modifying, replacing, or removing data



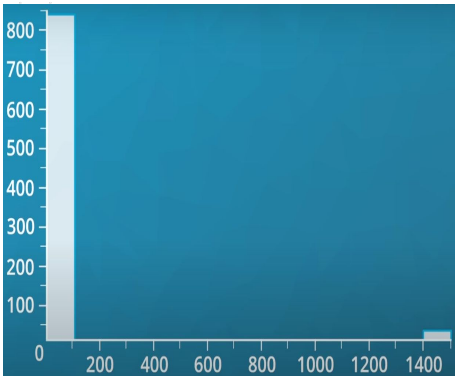
## Data Analysis Steps (From Udacity Nanodegree) - 4

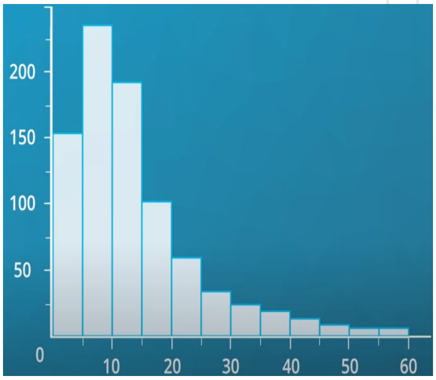
**Step 3:** Perform Exploratory Data Analysis (EDA)

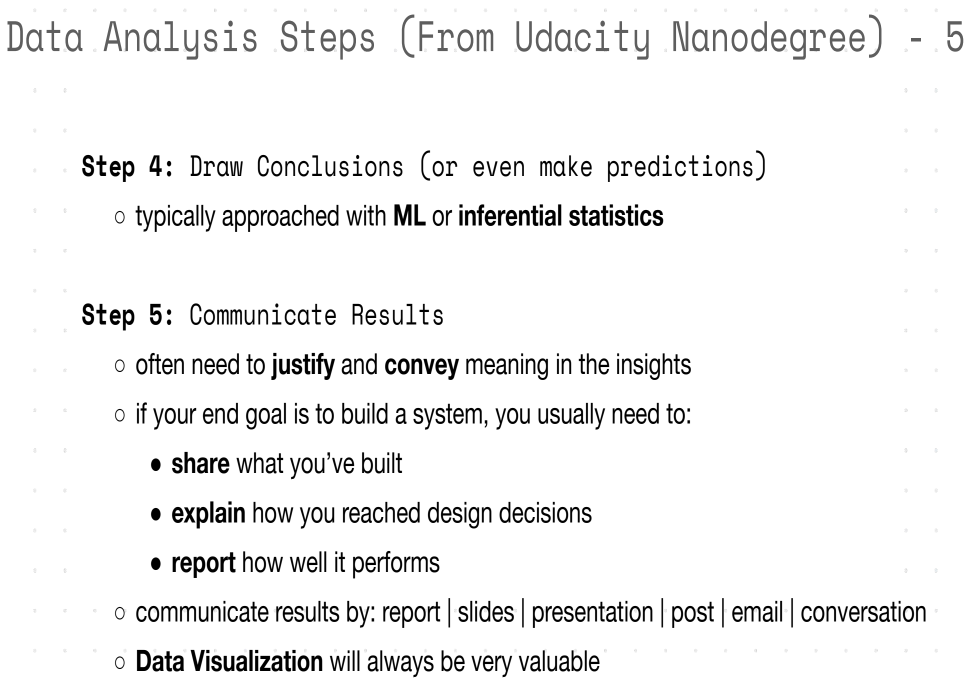
- **Explore then augment** data to maximize the potential of:
  - analysis & visualizations & models
- **Exploring** involves:
  - finding **patterns** in data
  - **visualizing** relationships in data
  - building **intuition** about what you’re working with
- **After Exploring** (**optional**)
  - **Remove Outliers:**
  - **Feature Engineering:** create better features from data

**Feature engineering** is the process of selecting, manipulating, and transforming raw data into features that can be used in ML.

# Remove Outliers







## Data Analysis Steps (From Udacity Nanodegree) - 5

**Step 4:** Draw Conclusions (or even make predictions)

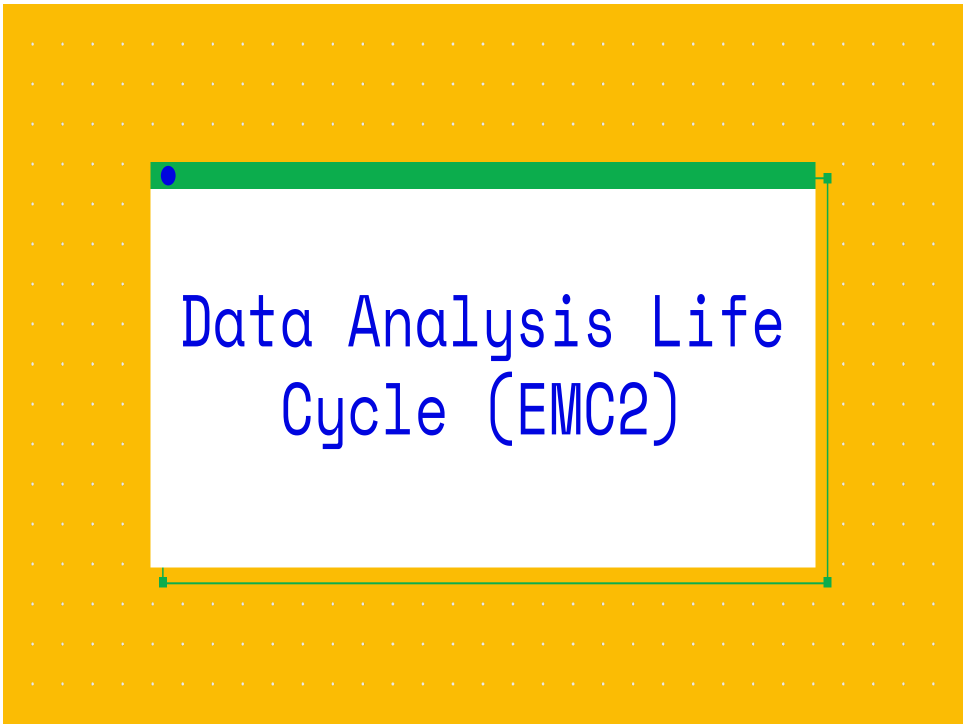
- typically approached with **ML** or **inferential statistics**

**Step 5:** Communicate Results

- often need to **justify** and **convey** meaning in the insights
- if your end goal is to build a system, you usually need to:
  - **share** what you've built
  - **explain** how you reached design decisions
  - **report** how well it performs
- communicate results by: report | slides | presentation | post | email | conversation
- **Data Visualization** will always be very valuable

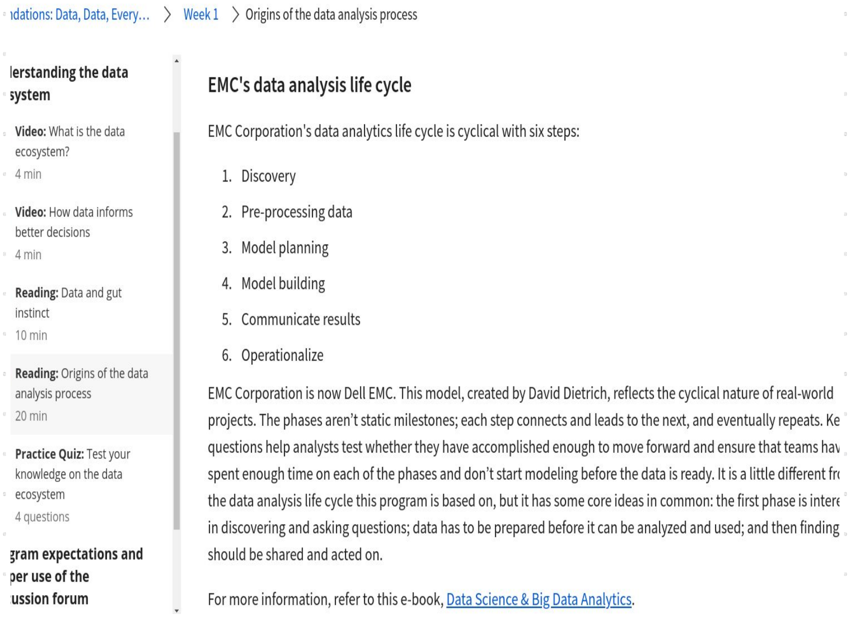
### Inferential statistics

- uses sample data to draw conclusions and make generalizations about a larger population



# Data Analysis Life Cycle (EMC2)

# Data Analysis Life Cycle (From EMC) - 1



ıdations: Data, Data, Every...
Week 1
Origins of the data analysis process
lerstanding the data
system
Video: What is the data ecosystem?
4 min
Video: How data informs better decisions
4 min
Reading: Data and gut instinct
10 min
Reading: Origins of the data analysis process
20 min
Practice Quiz: Test your knowledge on the data ecosystem
4 questions
gram expectations and
per use of the
ussion forum

## EMC's data analysis life cycle

EMC Corporation's data analytics life cycle is cyclical with six steps:

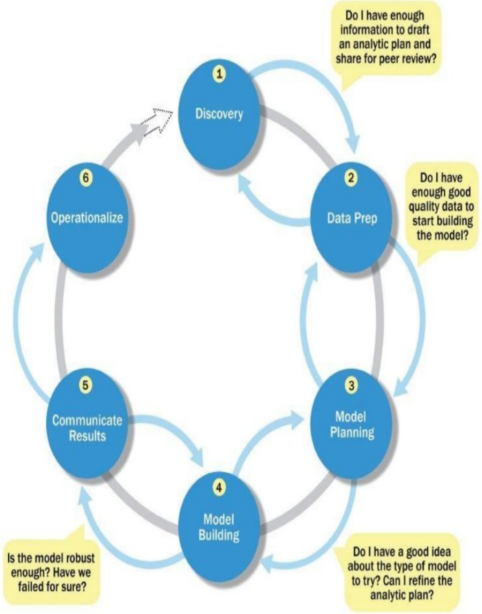
1. Discovery
2. Pre-processing data
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

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. Ke questions help analysts test whether they have accomplished enough to move forward and ensure that teams hav spent enough time on each of the phases and don't start modeling before the data is ready. It is a little different fr the data analysis life cycle this program is based on, but it has some core ideas in common: the first phase is intere in discovering and asking questions; data has to be prepared before it can be analyzed and used; and then finding should be shared and acted on.

For more information, refer to this e-book, Data Science & Big Data Analytics.

- https://www.coursera.org/learn/foundations-data/supplement/WWlrt/origins-of-the-data-analysis-process
- 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
    - don't start modeling before the data is ready
- It is different from other data analysis life cycles, but they have some core ideas in common:
    - first phase is interested in discovering/asking questions
    - data has to be prepared before it can be analyzed/used
    - findings should be shared and acted on

# Data Analysis Life Cycle (From EMC) - 2

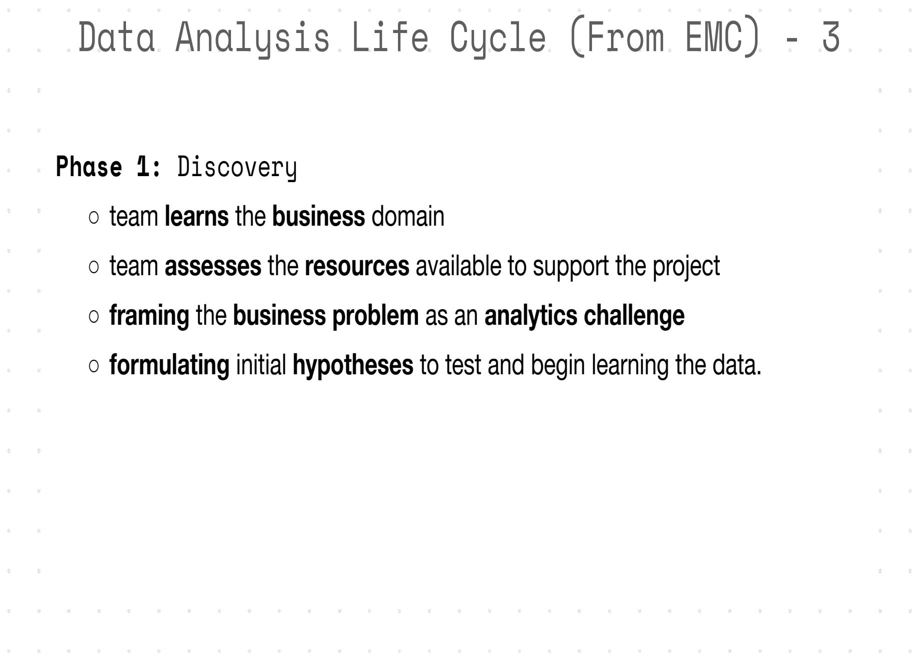


Do I have enough information to draft an analytic plan and share for peer review?

Do I have enough good quality data to start building the model?

Do I have a good idea about the type of model to try? Can I refine the analytic plan?

Is the model robust enough? Have we failed for sure?



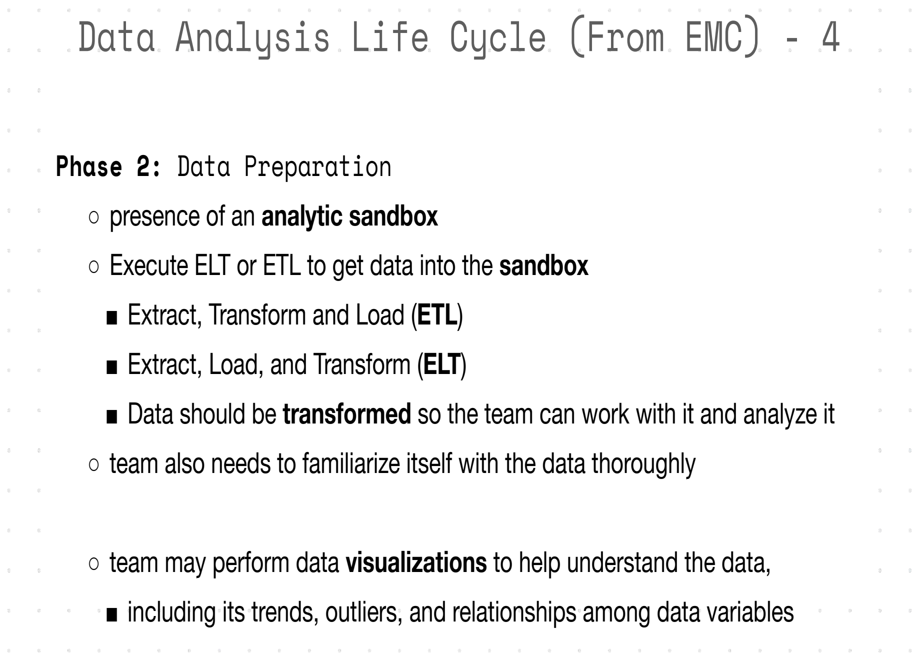
## Data Analysis Life Cycle (From EMC) - 3

**Phase 1:** Discovery

- team **learns** the **business** domain
- team **assesses** the **resources** available to support the project
- **framing** the **business problem** as an **analytics challenge**
- **formulating** initial **hypotheses** to test and begin learning the data.

**Phase 1: Discovery**

- team learns business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
- team assesses the resources available to support the project in terms of people, technology, time, and data.
- Important activities in this phase include
  - framing the business problem as an analytics challenge that can be addressed in subsequent phases and
  - formulating initial hypotheses (IHs) to test and begin learning the data.



## Data Analysis Life Cycle (From EMC) - 4

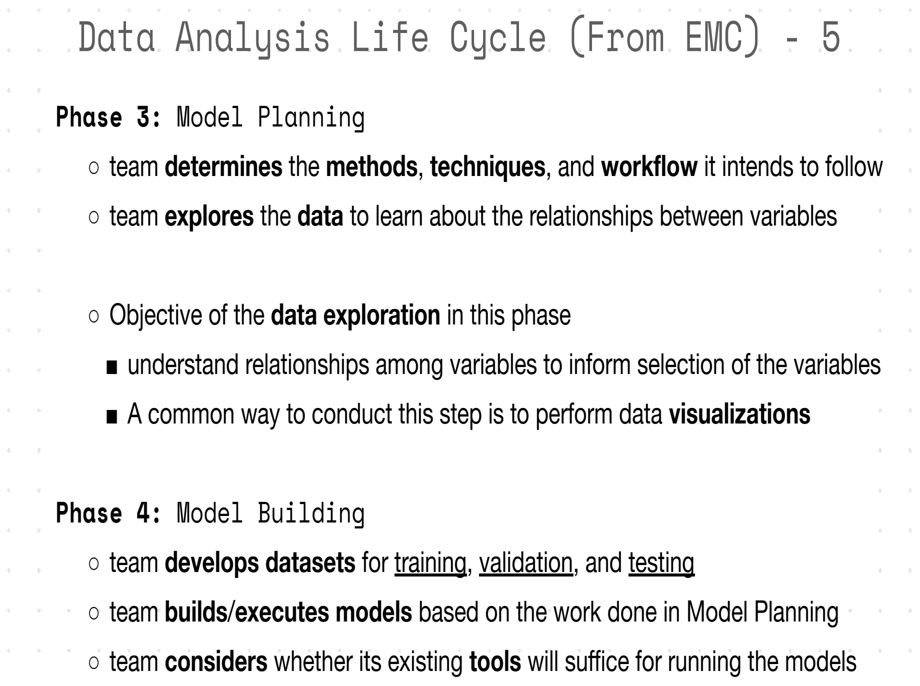
**Phase 2:** Data Preparation

- presence of an **analytic sandbox**
- Execute ELT or ETL to get data into the **sandbox**
  - Extract, Transform and Load (**ETL**)
  - Extract, Load, and Transform (**ELT**)
  - Data should be **transformed** so the team can work with it and analyze it
- team also needs to familiarize itself with the data thoroughly

- team may perform data **visualizations** to help understand the data,
  - including its trends, outliers, and relationships among data variables

### Phase 2: Data Preparation

- Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
- The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox.
- The ELT and ETL are sometimes abbreviated as ETLT.
- Data should be transformed in the ETLT process so the team can work with it and analyze it.



Data Analysis Life Cycle (From EMC) - 5

**Phase 3:** Model Planning

- team **determines** the **methods**, **techniques**, and **workflow** it intends to follow
- team **explores** the **data** to learn about the relationships between variables

- Objective of the **data exploration** in this phase
  - understand relationships among variables to inform selection of the variables
  - A common way to conduct this step is to perform data **visualizations**

**Phase 4:** Model Building

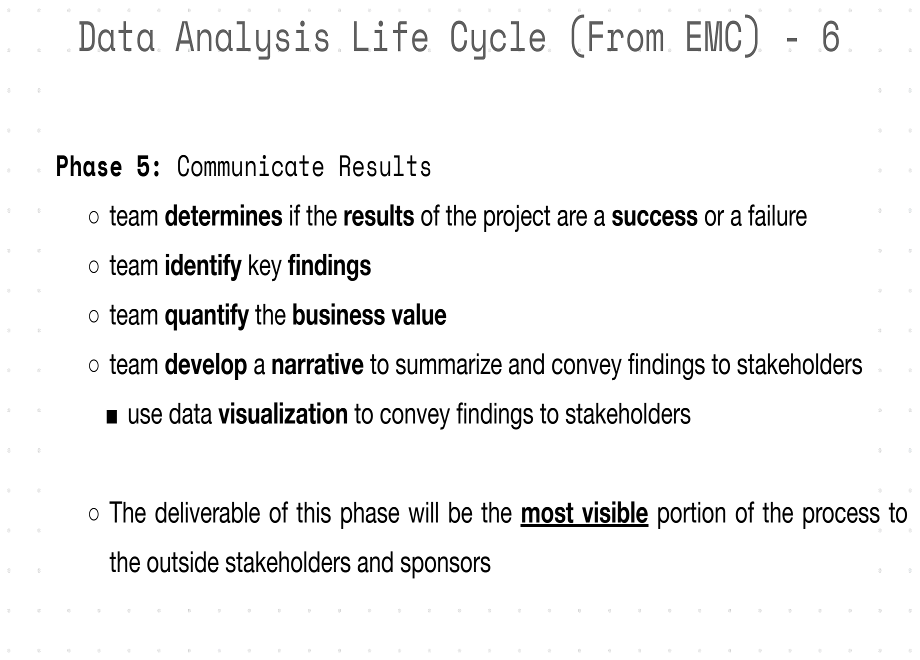
- team **develops datasets** for training, validation, and testing
- team **builds/executes models** based on the work done in Model Planning
- team **considers** whether its existing **tools** will suffice for running the models

## Phase 3: Model Planning

- Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

### Phase 4: Model building

- In Phase 4, the team builds and executes models based on the work done in the model planning phase.
- The team also considers whether its existing tools will suffice for running the models, or if it will need a more robust environment for executing models and workflows (for example, fast hardware and parallel processing).



## Data Analysis Life Cycle (From EMC) - 6

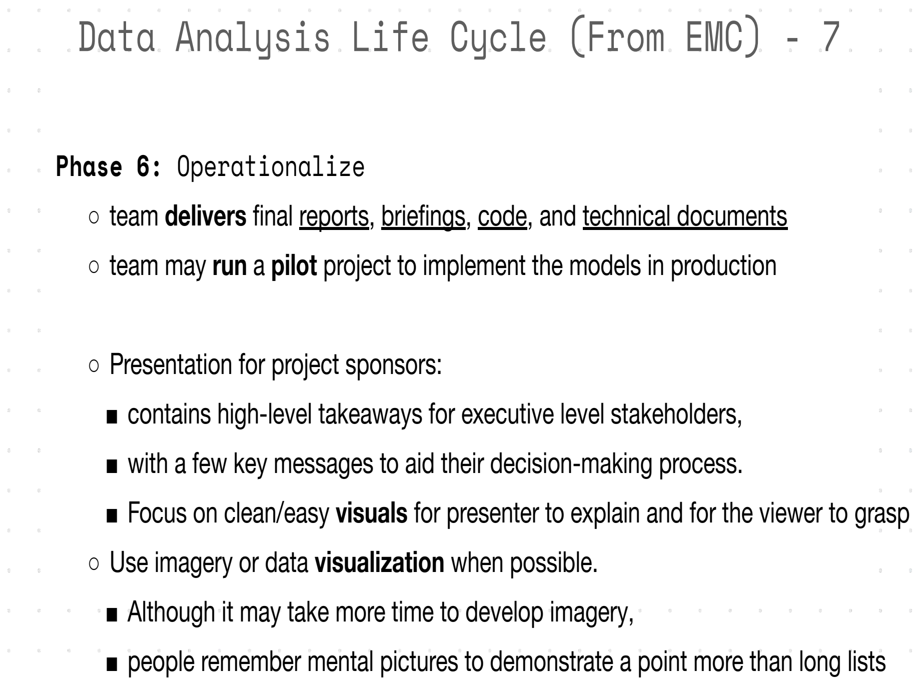
**Phase 5:** Communicate Results

- team **determines** if the **results** of the project are a **success** or a failure
- team **identify** key **findings**
- team **quantify** the **business value**
- team **develop** a **narrative** to summarize and convey findings to stakeholders
  - use data **visualization** to convey findings to stakeholders

- The deliverable of this phase will be the **most visible** portion of the process to the outside stakeholders and sponsors

### Phase 5: Communicate Results

- In Phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in Phase 1.
- The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.



## Data Analysis Life Cycle (From EMC) - 7

**Phase 6:** Operationalize

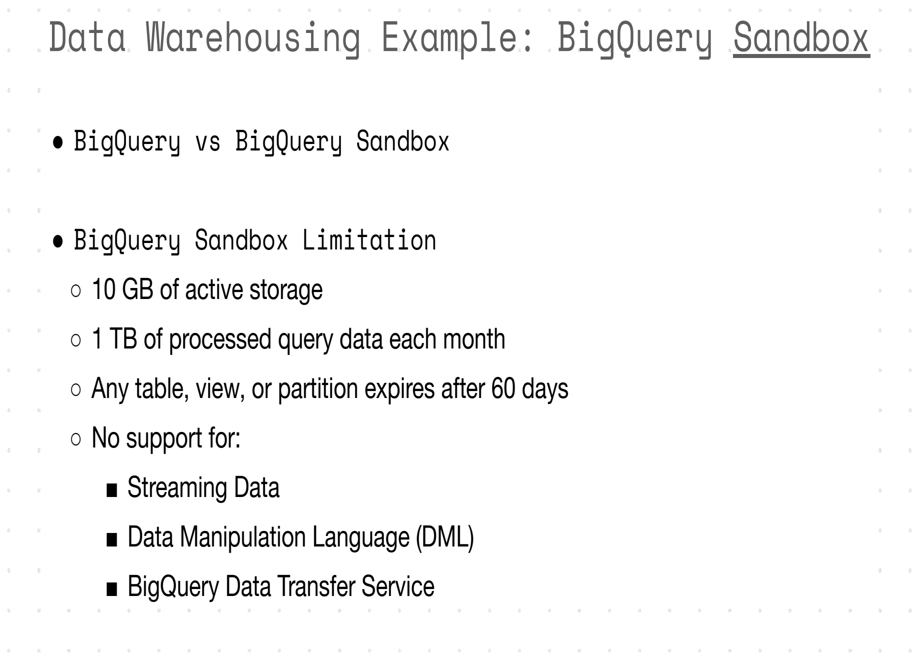
- team **delivers** final reports, briefings, code, and technical documents
- team may **run** a **pilot** project to implement the models in production

- Presentation for project sponsors:
  - contains high-level takeaways for executive level stakeholders,
  - with a few key messages to aid their decision-making process.
  - Focus on clean/easy **visuals** for presenter to explain and for the viewer to grasp
- Use imagery or data **visualization** when possible.
  - Although it may take more time to develop imagery,
  - people remember mental pictures to demonstrate a point more than long lists

### Phase 6: Operationalize

- In Phase 6, the team delivers final reports, briefings, code, and technical documents.
- In addition, the team may run a pilot project to implement the models in a production environment.

**A pilot project** is a small-scale, short-term trial or experiment designed to test the feasibility, effectiveness, and potential outcomes of a larger project before full-scale implementation.



## Data Warehousing Example: BigQuery Sandbox

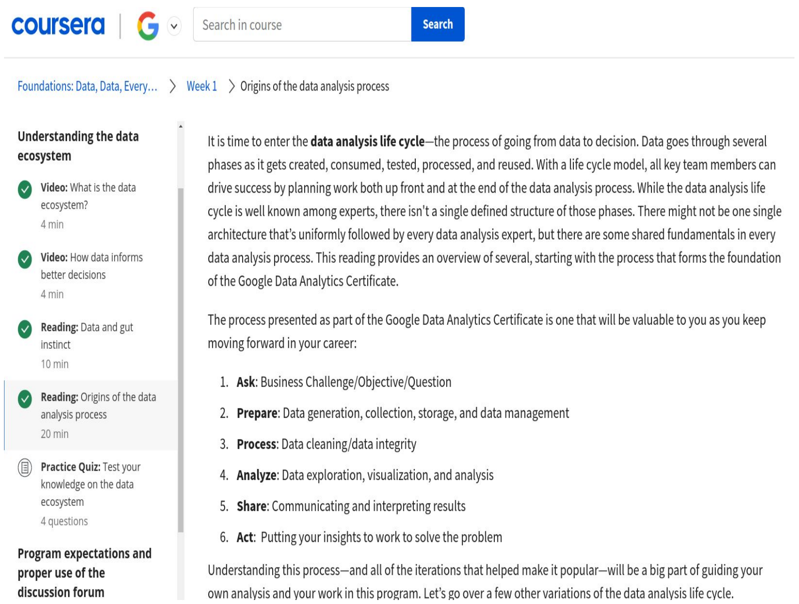
- BigQuery vs BigQuery Sandbox

- BigQuery Sandbox Limitation
  - 10 GB of active storage
  - 1 TB of processed query data each month
  - Any table, view, or partition expires after 60 days
  - No support for:
    - Streaming Data
    - Data Manipulation Language (DML)
    - BigQuery Data Transfer Service

https://cloud.google.com/bigquery/docs/sandbox

# Data Analysis Life Cycle (Google)

# Data Analysis Life Cycle (From Google Data Analytics Certificate) - 1



Foundations: Data, Data, Every... > Week 1 > Origins of the data analysis process

**Understanding the data ecosystem**

✓ **Video:** What is the data ecosystem?
4 min

✓ **Video:** How data informs better decisions
4 min

✓ **Reading:** Data and gut instinct
10 min

✓ **Reading:** Origins of the data analysis process
20 min

**Practice Quiz:** Test your knowledge on the data ecosystem
4 questions

**Program expectations and proper use of the discussion forum**

It is time to enter the **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. This reading provides an overview of several, starting with the process that forms the foundation of the Google Data Analytics Certificate.

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

1. **Ask**: Business Challenge/Objective/Question
2. **Prepare**: Data generation, collection, storage, and data management
3. **Process**: Data cleaning/data integrity
4. **Analyze**: Data exploration, visualization, and analysis
5. **Share**: Communicating and interpreting results
6. **Act**: Putting your insights to work to solve the problem

Understanding this process—and all of the iterations that helped make it popular—will be a big part of guiding your own analysis and your work in this program. Let's go over a few other variations of the data analysis life cycle.

https://www.coursera.org/learn/foundations-data/supplement/WWlrt/origins-of-the-data-analysis-process

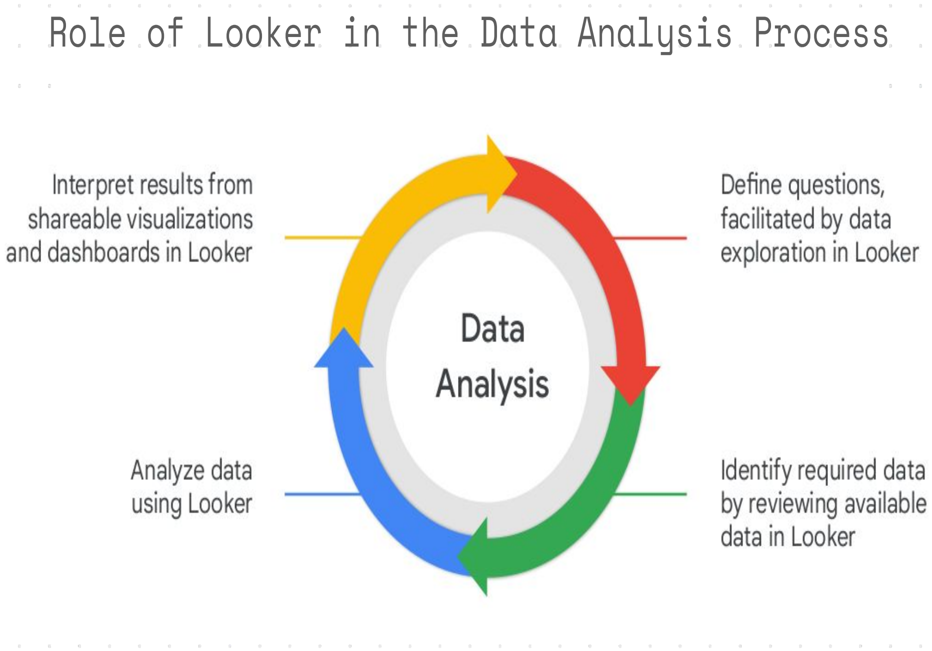
1. **Ask**: Business Challenge/Objective/Question
2. **Prepare:** Data generation, collection, storage, and data management
3. **Process:** Data cleaning / data integrity
4. **Analyze:** Data exploration, visualization, and analysis
5. **Share:** Communicating and interpreting results
6. **Act**: Putting your insights to work to solve the problem

https://www.sap.com/mena/products/technology-platform/what-is-data-management.html

**Data management** is the practice of collecting, organizing, managing, and accessing data to support productivity, efficiency, and decision-making.

# Data Analysis Life Cycle
# (From Google Data Analytics Certificate) - 2

1. **Ask**
2. **Prepare**
3. **Process**
4. **Analyse**
5. **Share**
6. **Act**



Role of Looker in the Data Analysis Process

Data Analysis
Define questions, facilitated by data exploration in Looker
Identify required data by reviewing available data in Looker
Analyze data using Looker
Interpret results from shareable visualizations and dashboards in Looker

1. **Define questions**. Identify what questions need to be answered using your data.
2. **Identify required data.** Determine the specific dimensions and measures you will need to answer those questions.
3. **Analyze data.** Explore the dimension and/or measure relationships via tables and visualizations. This exploration of your data should empower you to take some kind of action or make some kind of decision with regard to your work.
4. **Interpret the results.** Glean actionable insights from your analyzed data.

https://cloud.google.com/looker
Looker is an enterprise platform for BI, data applications, and embedded analytics that helps you explore and share insights in real time.

# Questions

# Links

https://github.com/fcai-b/dv

# References

1. https://www.udacity.com/course/data-analyst-nanodegree--nd002
    - Udacity Nanodegree
2. https://www.coursera.org/learn/foundations-data
    - Google Data Analytics Professional Certificate - 1st Course
3. https://cloud.google.com/bigquery/docs/sandbox
    - BigQuery Sandbox