The Data Science Process

Data Science Process

The Data Science Process

DEFINE THE GOAL

CLEAN THE DATA

FIND INSIGHTS AND VISUALIZE

ITTERATE

GET THE DATA

ENRICH THE DATA

DEPLOY MACHINE LEARNING

Source: *https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success*
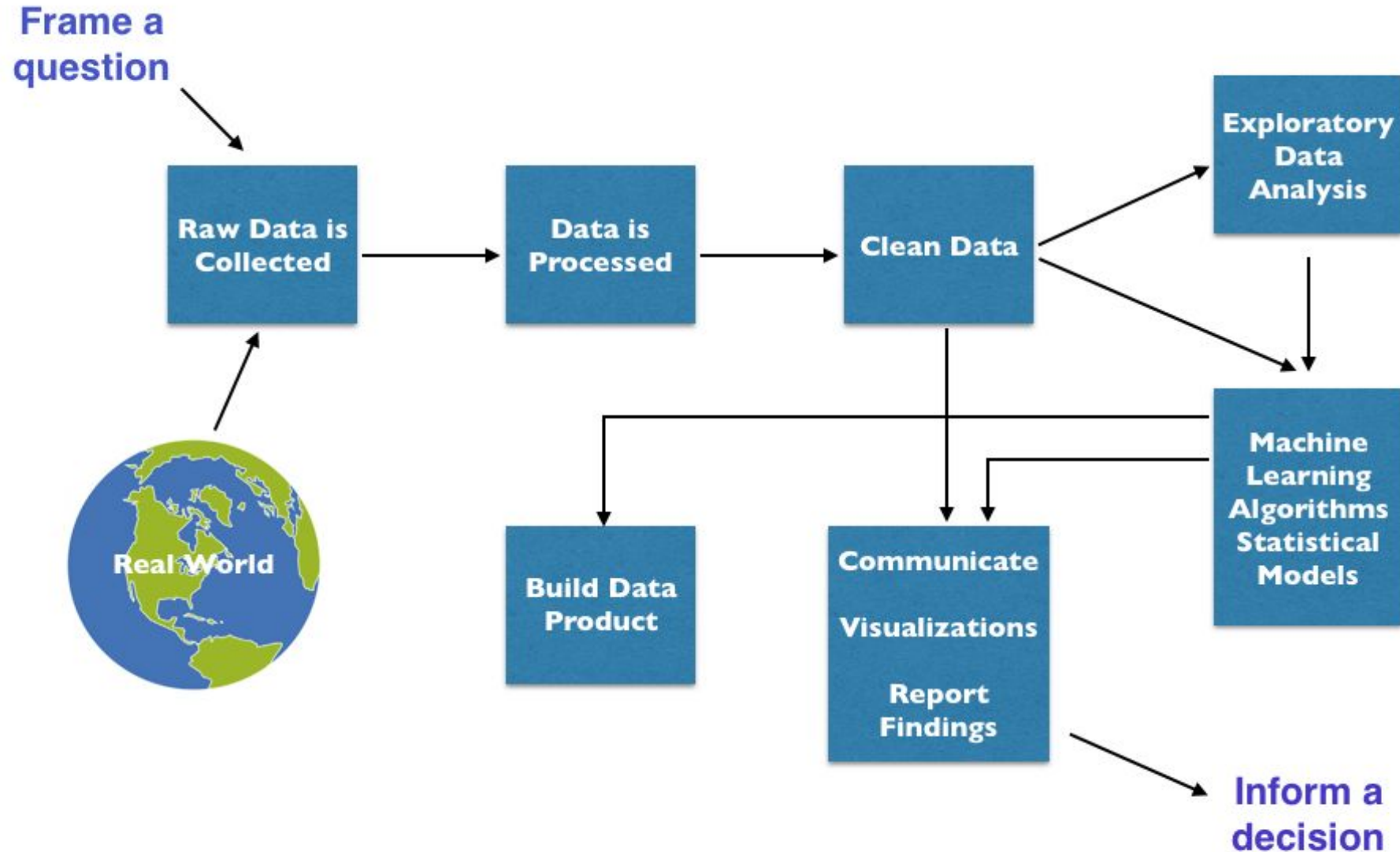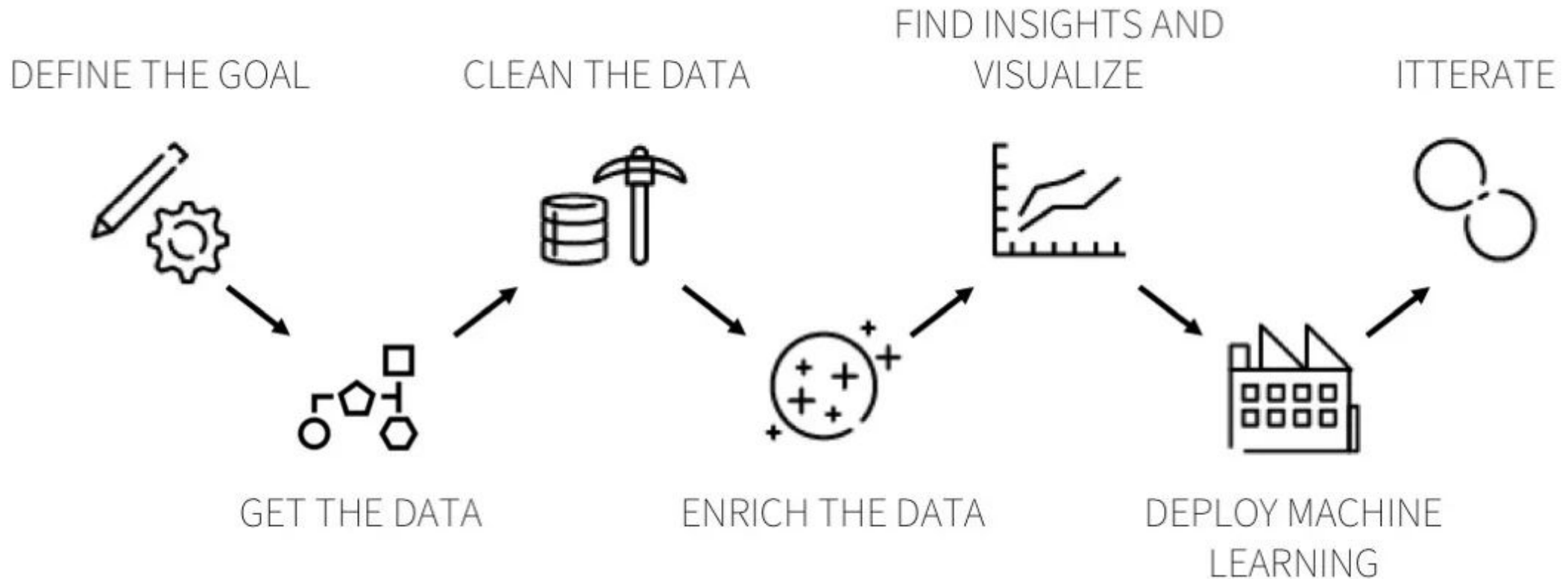
The Data Science Process

Step 1:

- Frame a Question

- Define a Goal

- What is the business need driving this project?

The Data Science Process

Raw Data is Collected/Get the Data

- Connect to a Database (eg. SQL)

- Use an API (Application Programming Interface)

- Scrape from the Web

- Find Open Data (eg. UNData, data.nashville.gov)

The Data Science Process

Data is Processed/Clean Data/Enrich Data

- Checking for missing values/improperly formatted values/duplicates

- Extracting new variables

- Combining/Merging Multiple Data Sources

- Consolidating/Filtering

- Transforming/Tidying (Pivot or Melt)

The Data Science Process

Exploratory Analysis

- Summary Statistics

- Groupby + Aggregation (Split-Apply-Combine)

- Exploratory Plots

- Analyzing Distributions - Univariate and Bivariate

- Looking for Outliers/Anomalies

- Look for and Analyze Patterns

- Ask and (attempt to) Answer Questions about the Data

The Data Science Process

Machine Learning Algorithms/Models

- Uncovering trends/patterns/groupings in the data (unsupervised)

- How can we make predictions on new data? (supervised)

This will be the focus of the later portions of the bootcamp.

The Data Science Process

Visualize and Communicate

- Create Effective Visualizations to Communicate your Findings
- Data Storytelling/Presentation

Data Question 1