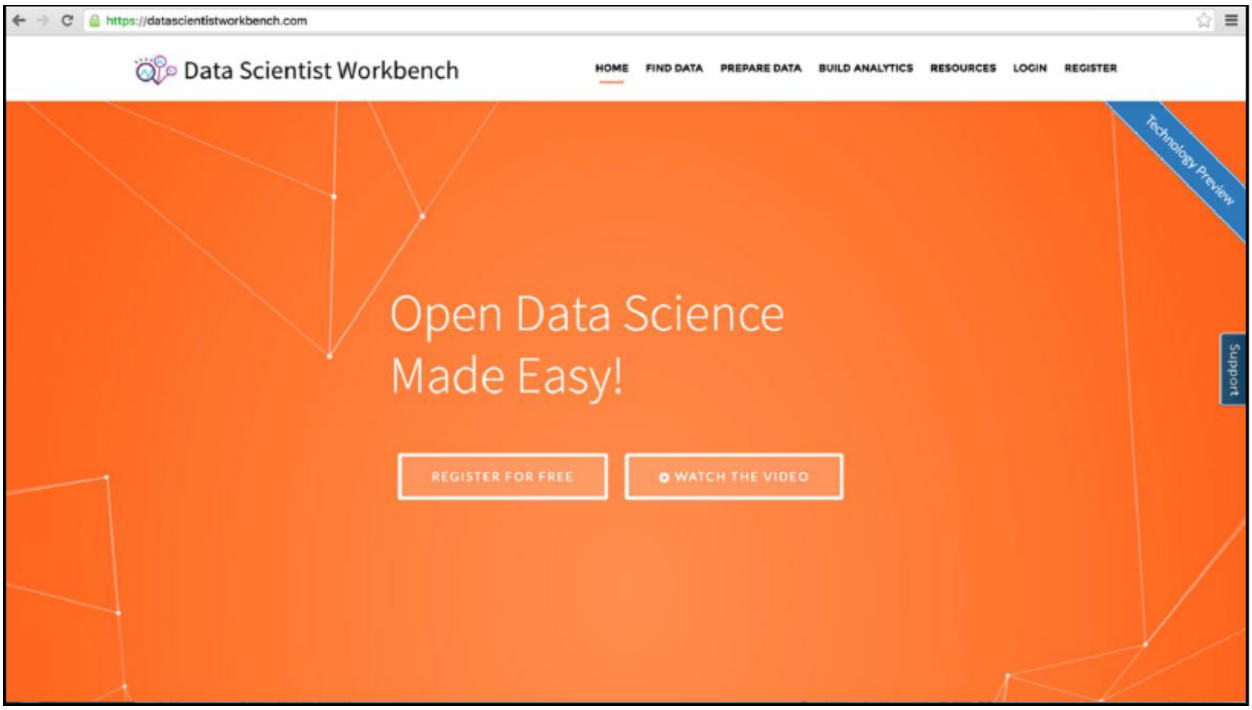
Lab: From Requirements to Collection

Hands-on exercises using data on food ingredients

Normally this hands-on lab would be done using the Data Scientist Workbench (DSWB), which provides a great way of organizing your models, displaying the output, and showing visualizations all in the same interface. However, since this is a fundamentals course without prerequisites, you might not yet be familiar with this tool. Therefore, this lab is mainly theoretical, describing the process you would following if you were utilizing the workbench.



If in future, you are so inclined to learn more about Data Science and the Data Scientist Workbench, you can start by taking the **Data Science Hands-on with Open-Source Tools** course and then come back to this course to complete the two optional hands-on labs that use the DSWB.

BIG DATA UNIVERSITY

Contents

BIG DATA UNIVERSITY

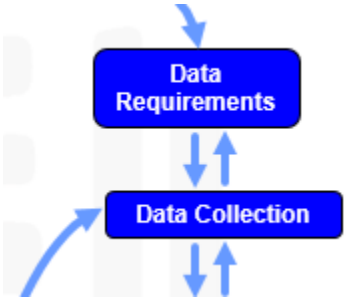
Lab 1 From requirements to collection

After completing this hands-on lab, you will know:

- What data requirements are?
- What occurs during data collection?
- What data understanding is?
- How to apply data requirements to any data science problem (cuisine, in this lab).
- How to apply data collection to any data science problem (cuisine, in this lab).

Allow 30 minutes to complete this section of the lab.

1.1 From data requirements to data collection



__1. The chosen analytic approach determines the _____

- Content, formats, representations

__2. Initial _____ is performed when:

- Available data resources (structured, unstructured, semi-structured) relevant to the problem domain
- Decide whether to obtain less-accessible data elements
- Revise data requirements or collect more data, if needed

The chosen analytic approach determines the data requirements.

- Specifically, the analytic methods to be used require certain data content, formats and representations, guided by domain knowledge.

In the initial data collection stage, data scientists identify and gather the available data resources—structured, unstructured and semi-structured—relevant to the problem domain.

- Typically, they must choose whether to make additional investments to obtain less-accessible data elements.
- It may be best to defer the investment decision until more is known about the data and the model.
- If there are gaps in data collection, the data scientist may have to revise the data requirements accordingly and collect new and/or more data.
- While data sampling and sub-setting are still important, today's high-performance platforms and in-database analytic functionality let data scientists use much larger data sets containing much or even all of the available data.

    - By incorporating more data, predictive models may be better able to represent rare events such as disease incidence or part failure.

After the initial data collection, data scientists typically use descriptive statistics and visualization techniques to gain data understanding. They need to:

- Understand the data content,
- Assess data quality, and
- Discover initial insights about the data.
- Additional data collection may be necessary to fill any gaps that may be found during this stage.

__3. Web Scrape of Online Food Recipes

A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (countries and ingredients) from three different websites and created a dataset. We will use an adapted version of his recipes dataset to explore how cuisines of the world differ based on their ingredients -- using decision trees.

For more information on Yong-Yeol Ahn and his research, please visit: www.yongyeol.com. You can also read his paper on "Flavor Network and the Principles of Food Pairing".

1.2 Summary

Congratulations! ...

BIG DATA UNIVERSITY

Lab 2 OPTIONAL: Data Understanding & Data Preparation using Data Science Workbench

After completing this hands-on lab, you will know:

- How to apply data understanding using the Data Scientist Workbench.
- How to apply data preparation using the Data Scientist Workbench.

Allow 30 minutes to complete this section of the lab.

1.3 Data understanding and data preparation

A. Login to your Data Scientist workbench account at my.datascientistworkbench.com. If you have not signed up yet, to datascientistworkbench.com and watch the video; then click the SIGN UP FOR FREE button to create your account.

B. Upload the Jupyter Notebook "*DS0103ENData Science Methodology 101 Module 2 Hands-on Lab using R in DSWB*" to your data folder to your Data Scientist Workbench.

C. Click on **Jupyter Notebook** and follow the instructions in the uploaded notebook (from step B): execute each of the code cells in the Notebook and review the results.

D. These are the steps within the Jupyter Notebook:

1. Import recipes.csv into R.
2. Data understanding & data preparation
    i. What does the data look like?
        1. look at the data
        2. What can we tell about the data? Summarize, visualize it
        3. Is the data clean? If not? Clean the data (i.e., inconsistent country names)
    ii. Which cuisines are most similar to each other? (k-means clustering)

1.4 Summary

Congratulations! ...