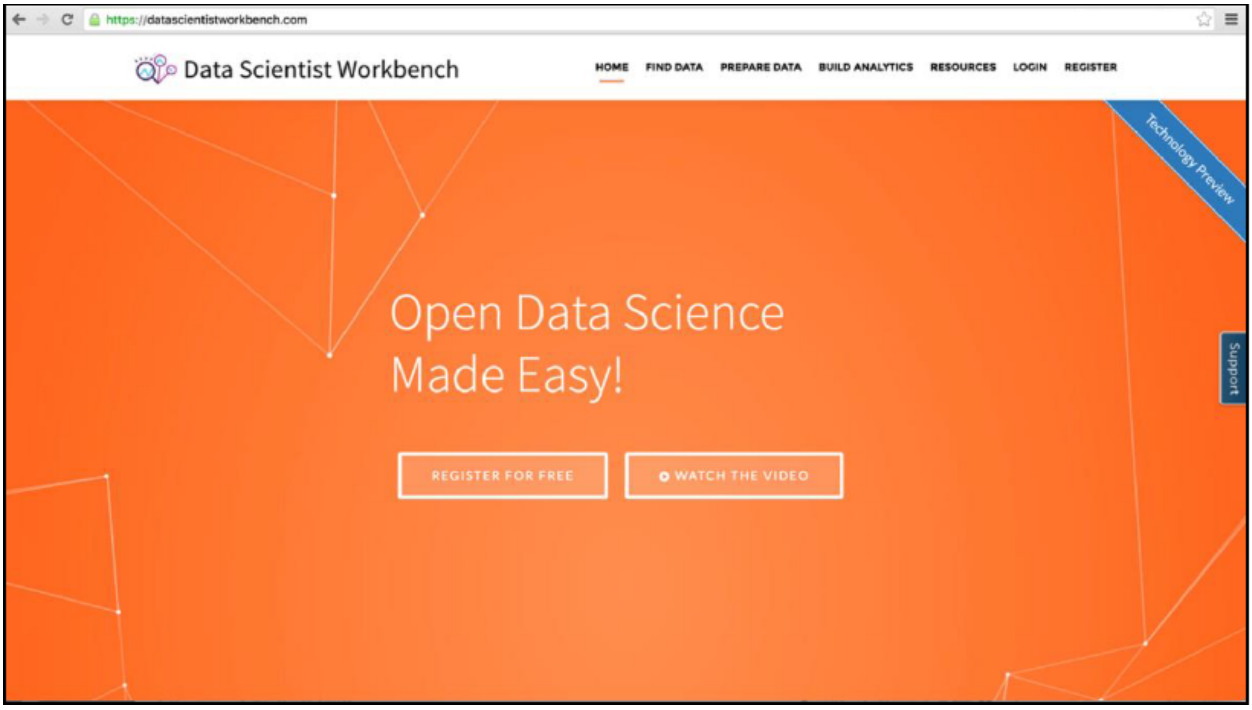
Lab: From Understanding to Preparation

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

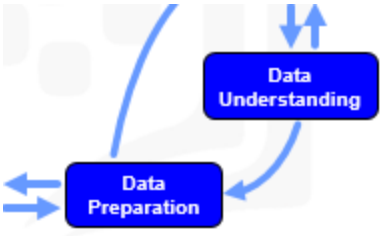
Lab 1 From understanding to preparation

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

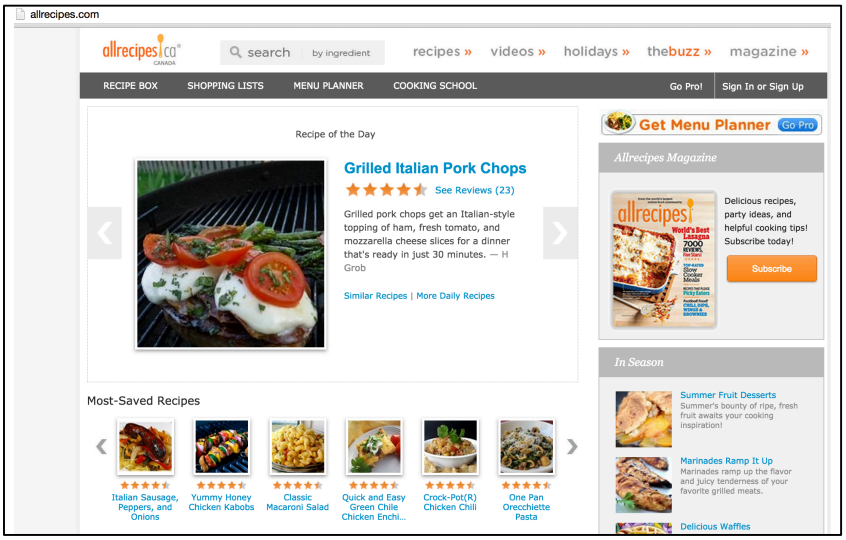
- What it means to “understand” data.
- What it means to “prepare” or “clean” data.
- Ways in which data is prepared.
- How to apply data understanding to any data science problem (cuisine, in this lab).
- How to apply data preparation to any data science problem (cuisine, in this lab).

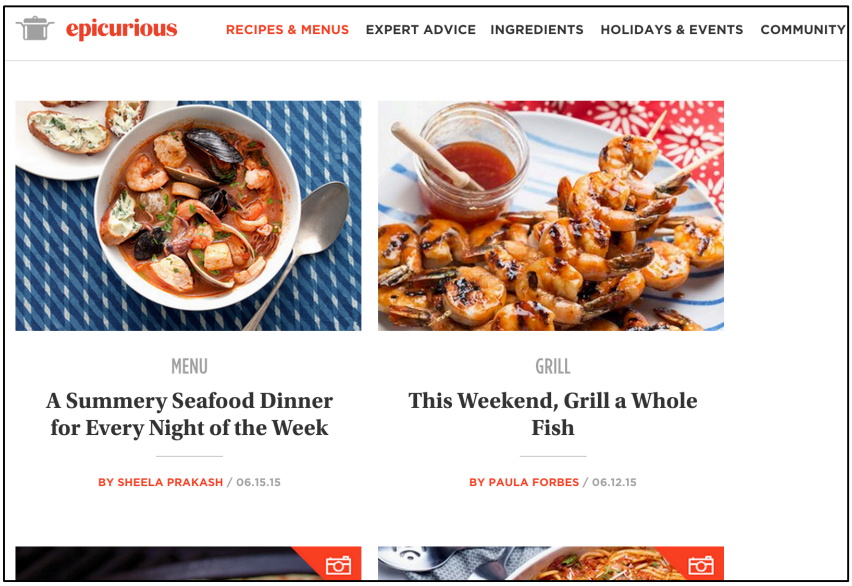
Allow 30 minutes to complete this section of the lab.

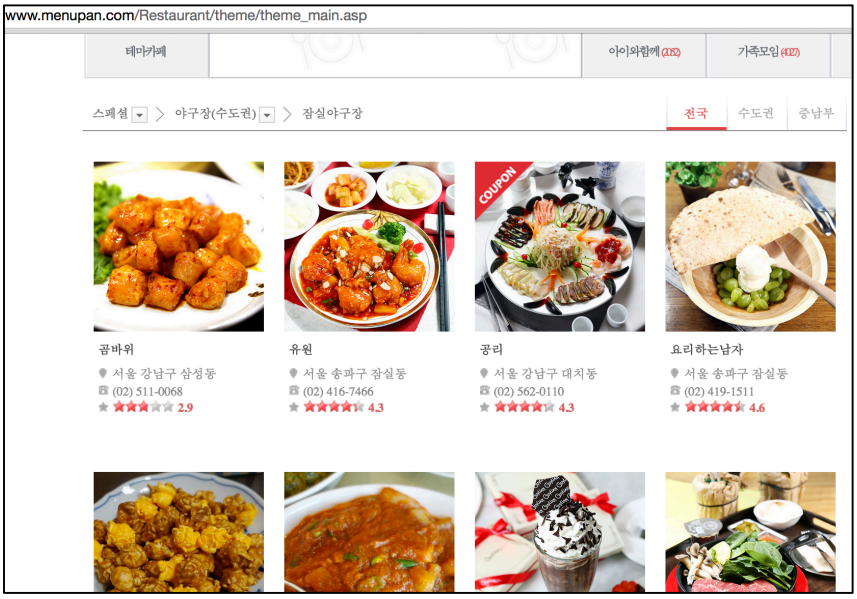
1.1 From data understanding to data preparation



__1. Web Scrape







The dataset was gathered from these websites

www.allrecipes.com
www.epicurious.com
www.menupan.com

We've combined the three datasets compiled by Yong-Yeol Ahn, and created the following dataset for this learning exercise.

Here's what the data looks like:

	country	almond	angelica	anise	anise_seed	apple	apple_brandy	apricot
1	Vietnamese	No	No	No	No	No	No	No
2	Vietnamese	No	No	No	No	No	No	No
3	Vietnamese	No	No	No	No	No	No	No
4	Vietnamese	No	No	No	No	No	No	No
5	Vietnamese	No	No	No	No	No	No	No
6	Vietnamese	No	No	No	No	No	No	No

Describe the data that you see:

_____

_____

_____

So now we have some goals.

__2. **So we have completed most of the data stages, and we're just about ready to get into the modeling stage.**

Data preparation encompasses all activities to construct the data set.

- **Data cleaning**
  - **Missing or invalid values**
  - **Eliminating duplicate rows**
  - **Formatting properly**
- **Combining multiple data sources**
- **Transforming data**
- **Feature engineering**
- **Text analysis**

Accelerate data preparation by automating common steps

Research data cleansing activities and document your findings

_____

_____

_____

_____

_____

_____

1.2 Summary

Congratulations! ...