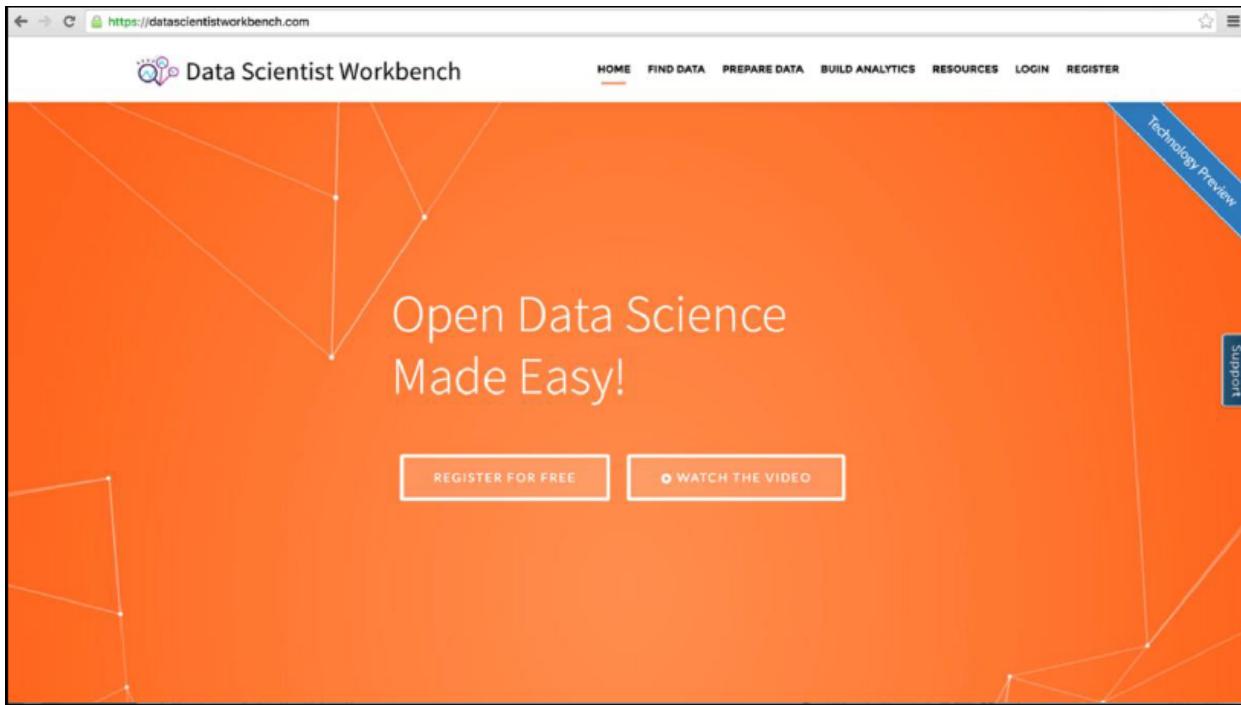




## Lab: From Understanding to Preparation

### *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 follow 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.



## Contents

<b>LAB 1</b>	<b>FROM UNDERSTANDING TO PREPARATION .....</b>	<b>3</b>
1.1	FROM DATA UNDERSTANDING TO DATA PREPARATION .....	4
1.2	SUMMARY .....	6



## Lab 1     From understanding to preparation

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

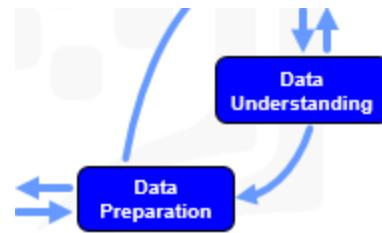
- What it means to “understand” data.
- What it means to “prepare” or “clean” data.
- Ways in which data is prepared.
- How to apply data understanding to any data science problem (cuisine, in this lab).
- How to apply data preparation to any data science problem (cuisine, in this lab).

Allow 30 minutes to complete this section of the lab.



# BIG DATA UNIVERSITY

## 1.1 From data understanding to data preparation



### 1. Web Scrape



# BIG DATA UNIVERSITY

www.menupan.com/Restaurant/theme/theme\_main.asp

The screenshot shows a website interface for a restaurant search. At the top, there are navigation links for '테마카페' (Theme Cafe), '아이와함께' (For Children), '기족모임' (Family Gathering), and language options '전국' (National), '수도권' (Metropolitan Area), and '충남부' (Chungnam). Below this, a breadcrumb navigation shows '스페셜 > 야구장(수도권) > 잠실야구장'. The main content area displays a grid of food items with small images, names, addresses, and ratings. The items include:

- 꼼바위 (Gombae): 서울 강남구 삼성동, (02) 511-0068, ★★★★☆ 2.9
- 유원 (Yuwon): 서울 송파구 잠실동, (02) 416-7466, ★★★★★ 4.3
- 꽁리 (Gongri): 서울 강남구 대치동, (02) 562-0110, ★★★★★ 4.3
- 요리하는남자 (Yoreohaneunnamja): 서울 송파구 잠실동, (02) 419-1511, ★★★★★ 4.6
- Other items shown in the grid include: 막창 (Makchang), 떡갈비 (Tteokgalbi), 그리고 커피 (And coffee), and 냄비 (Nimbi).

The dataset was gathered from these websites

[www.allrecipes.com](http://www.allrecipes.com)  
[www.epicurious.com](http://www.epicurious.com)  
[www.menupan.com](http://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! ...