(https://cognitiveclass.ai)

From Requirements to Collection

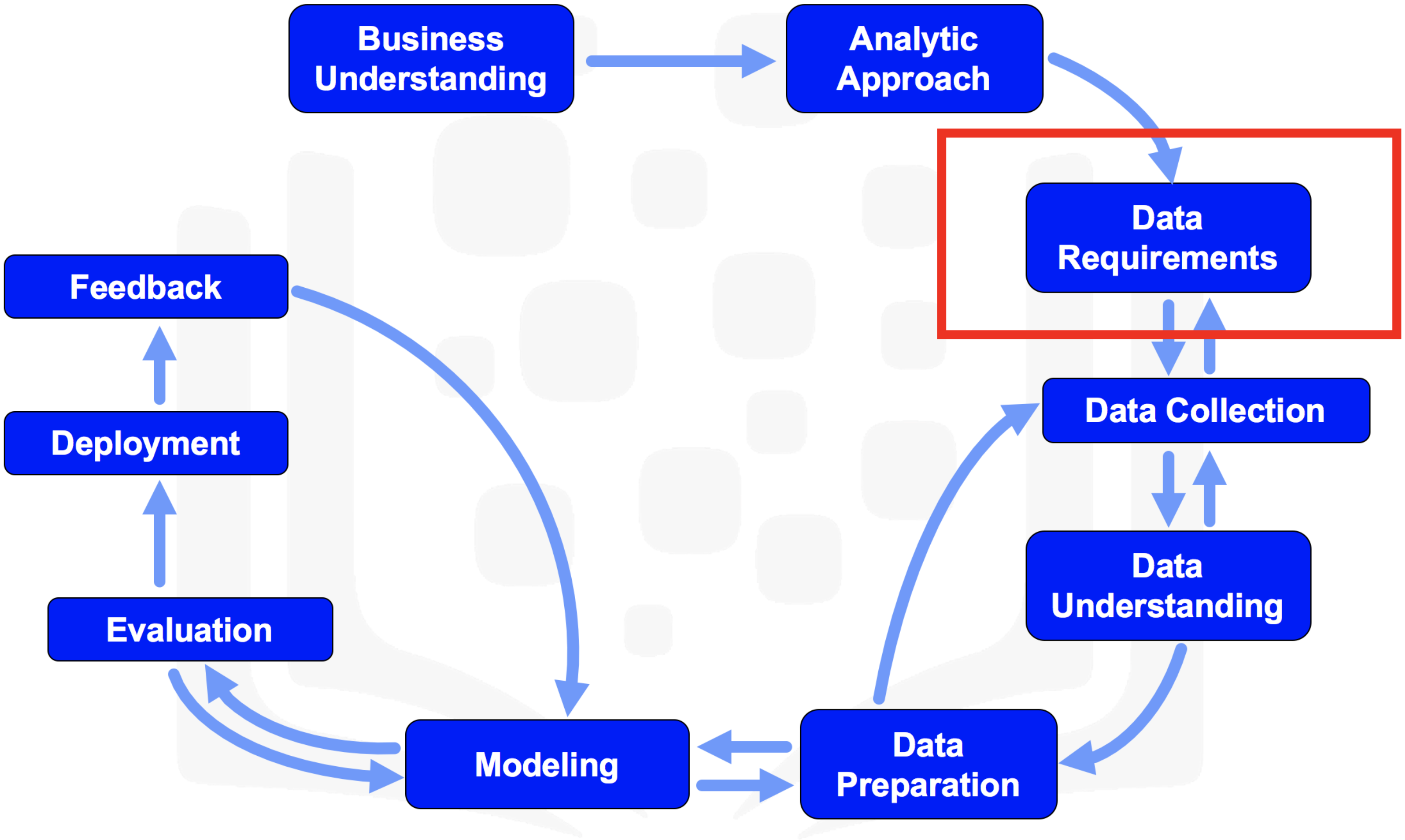
Introduction

In this lab, we will continue learning about the data science methodology, and focus on the **Data Requirements** and the **Data Collection** stages.

Table of Contents

</div>

Data Requirements

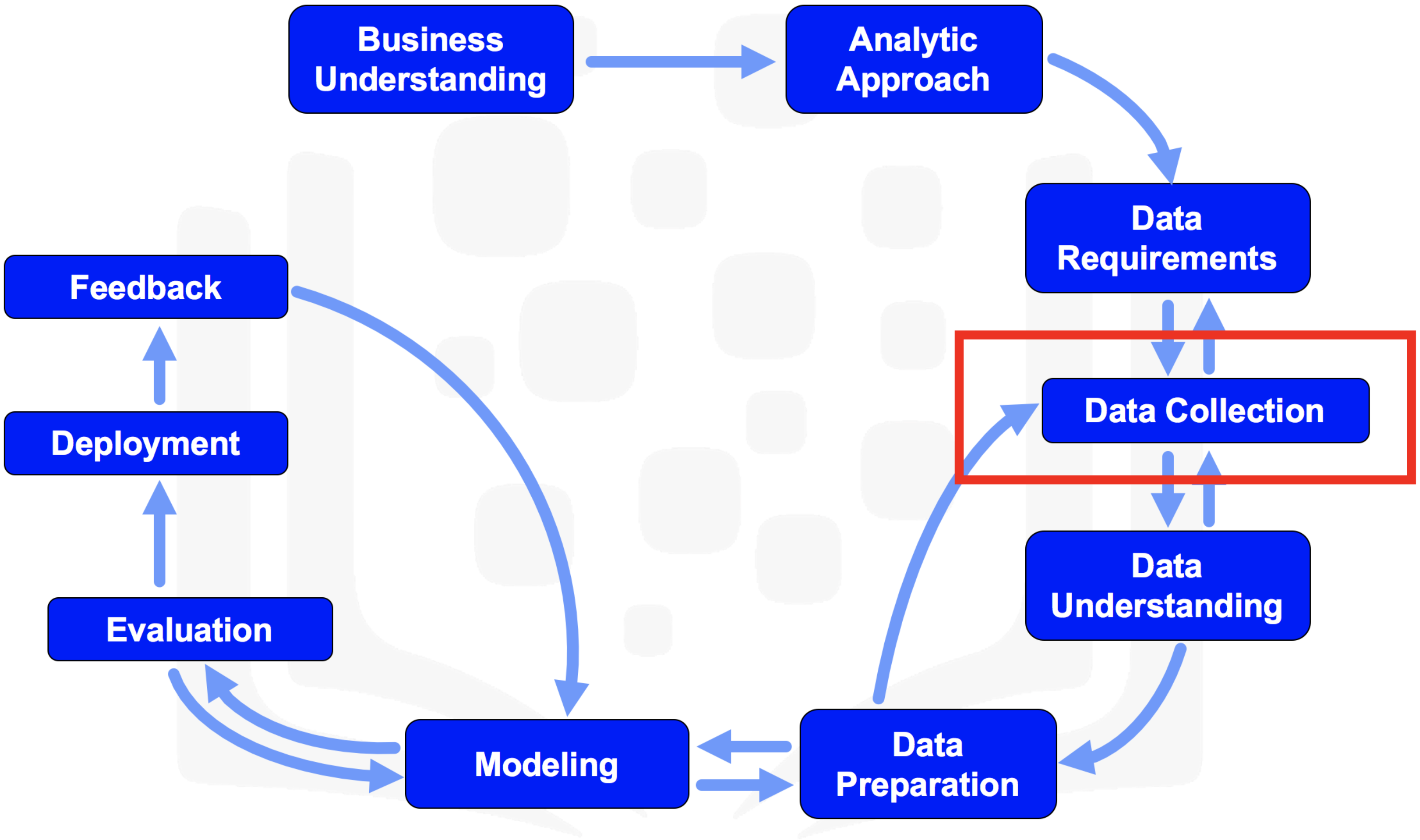

Business Understanding
Analytic Approach
Data Requirements
Data Collection
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Feedback

In the videos, we learned that 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 **From Problem to Approach Lab**, we determined that automating the process of determining the cuisine of a given recipe or dish is potentially possible using the ingredients of the recipe or the dish. In order to build a model, we need extensive data of different cuisines and recipes.

Identifying the required data fulfills the data requirements stage of the data science methodology.

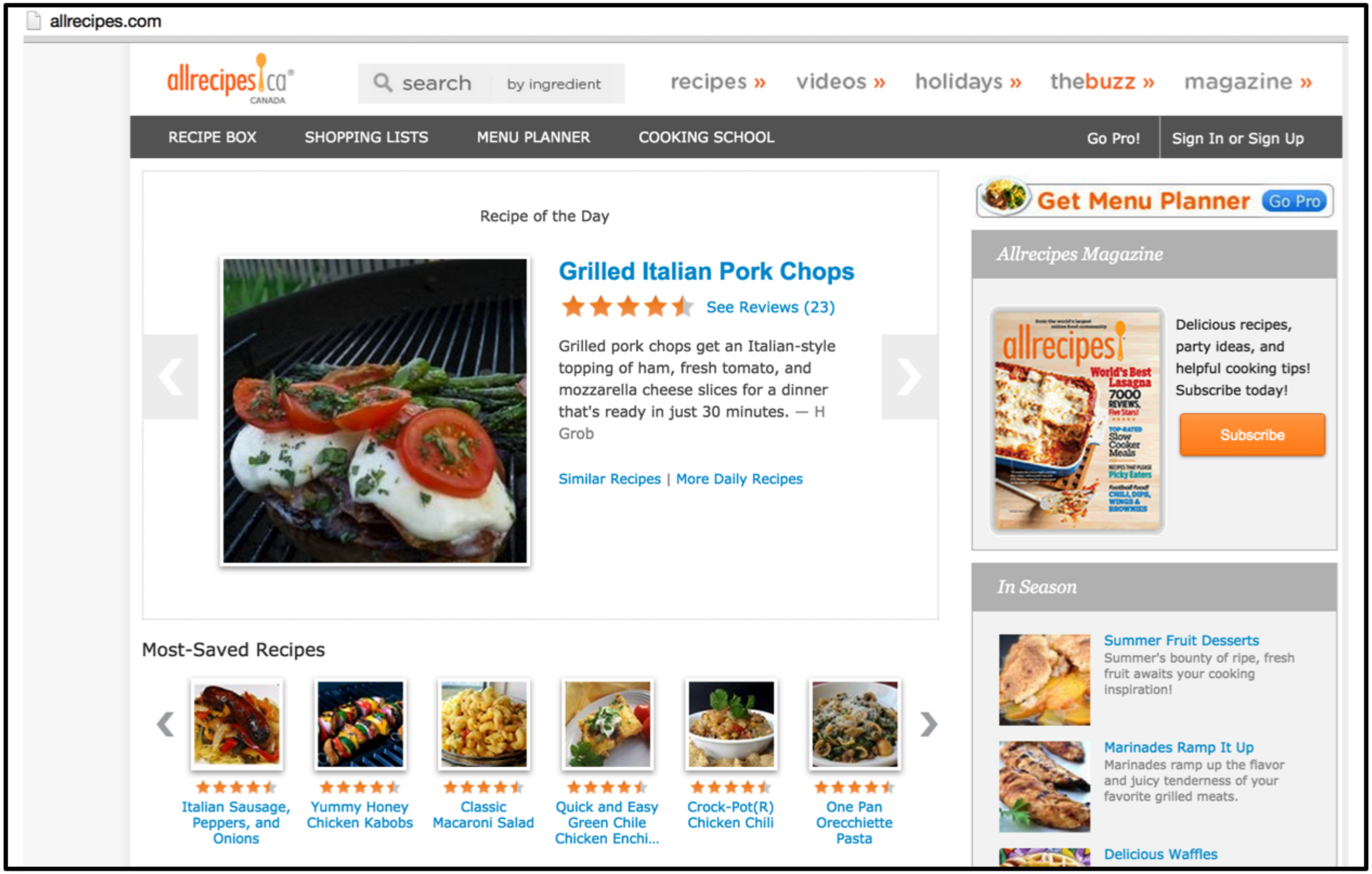
Data Collection



In the initial data collection stage, data scientists identify and gather the available data resources. These can be in the form of structured, unstructured, and even semi-structured data relevant to the problem domain.

Web Scraping of Online Food Recipes

A researcher named Yong-Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredients) from three different websites, namely:



allrecipes.com

Recipe of the Day

Grilled Italian Pork Chops

See Reviews (23)

Grilled pork chops get an Italian-style topping of ham, fresh tomato, and mozzarella cheese slices for a dinner that's ready in just 30 minutes. — H Grob

Similar Recipes | More Daily Recipes

Allrecipes Magazine

Delicious recipes, party ideas, and helpful cooking tips! Subscribe today!

Most-Saved Recipes

Italian Sausage, Peppers, and Onions

Yummy Honey Chicken Kabobs

Classic Macaroni Salad

Quick and Easy Green Chile Chicken Enchi...

Crock-Pot(R) Chicken Chili

One Pan Orecchiette Pasta

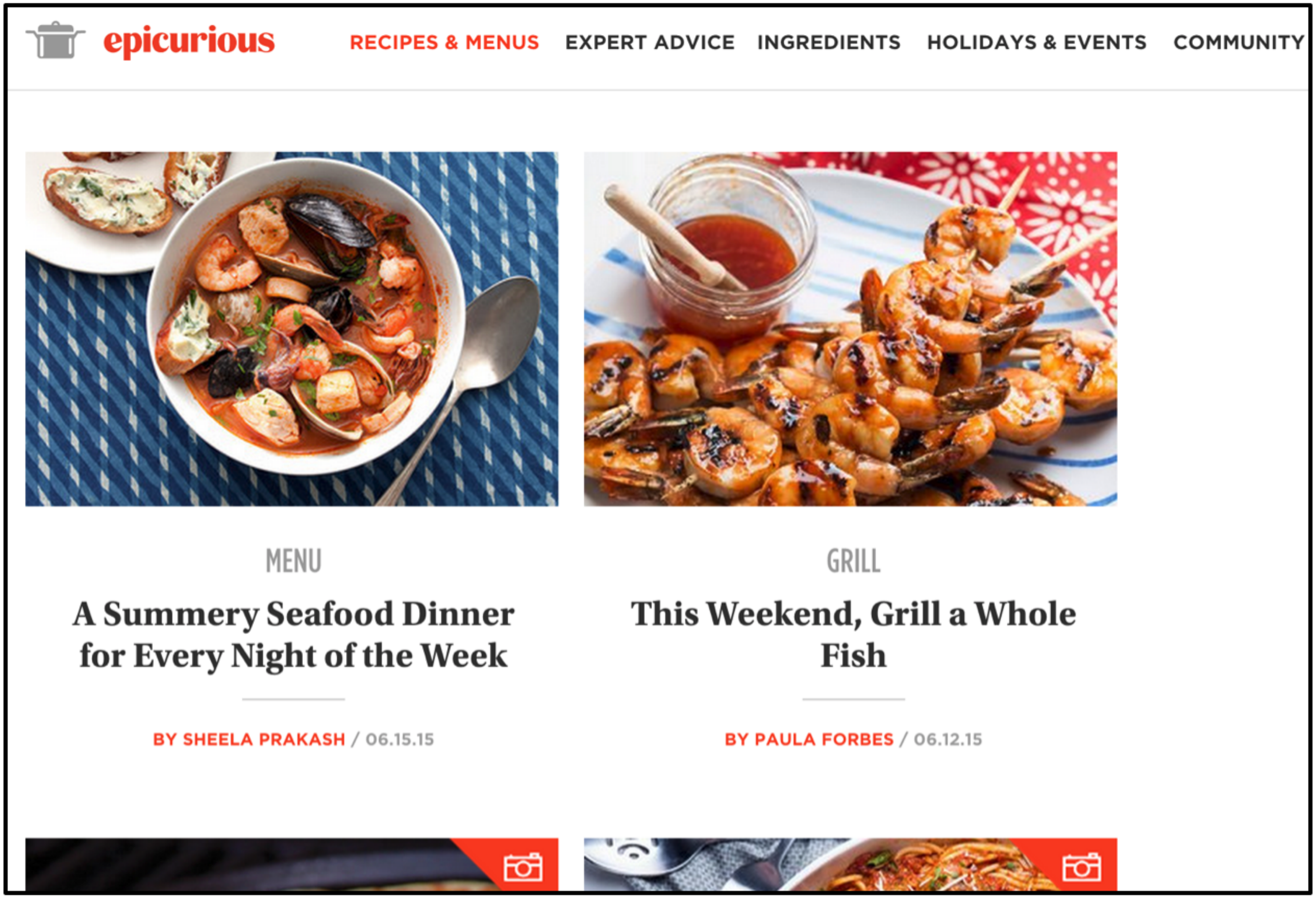
In Season

Summer Fruit Desserts
Summer's bounty of ripe, fresh fruit awaits your cooking inspiration!

Marinades Ramp It Up
Marinades ramp up the flavor and juicy tenderness of your favorite grilled meats.

Delicious Waffles

www.allrecipes.com



MENU

A Summery Seafood Dinner for Every Night of the Week

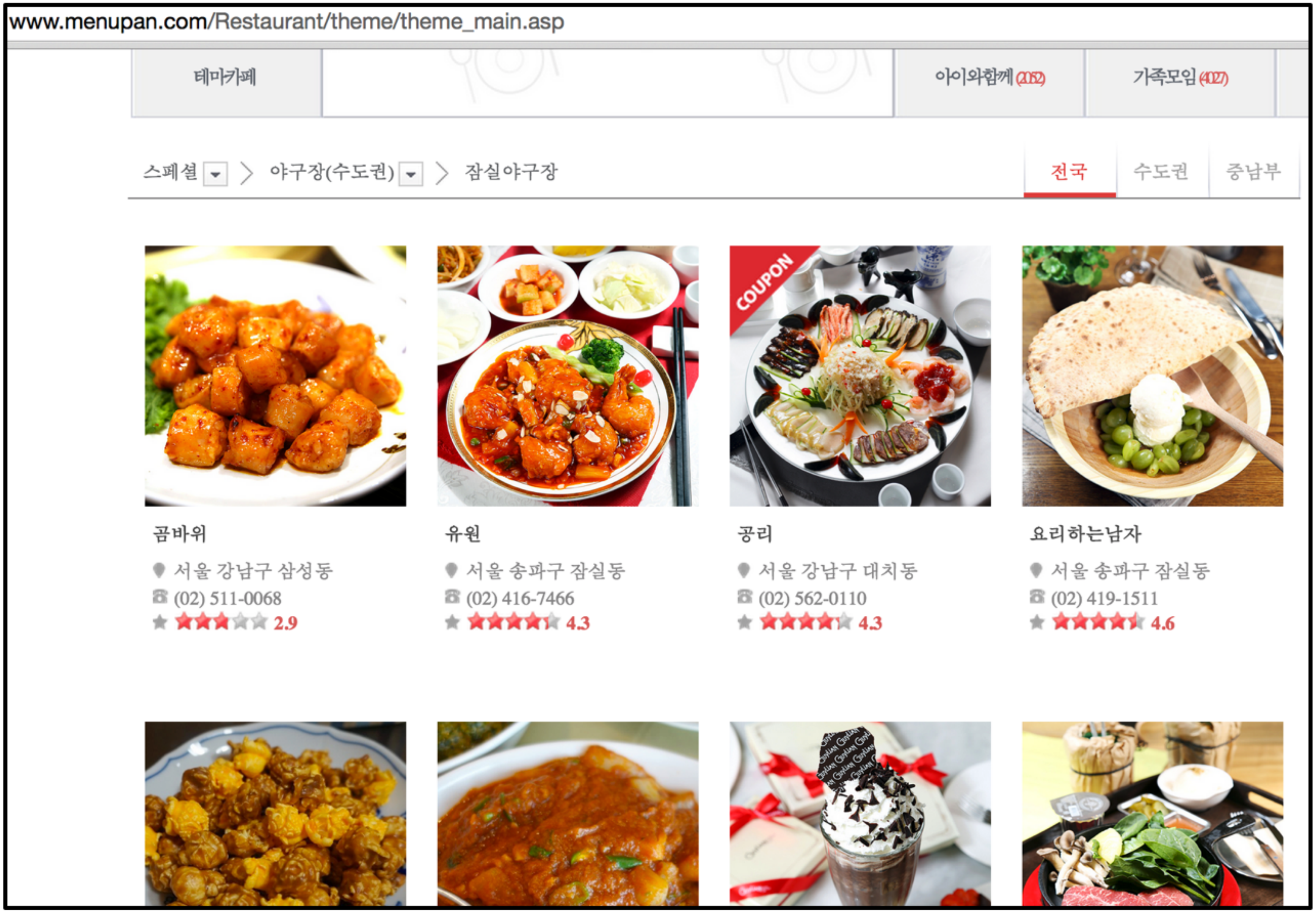
BY SHEELA PRAKASH / 06.15.15

GRILL

This Weekend, Grill a Whole Fish

BY PAULA FORBES / 06.12.15

www.epicurious.com



곰바위
서울 강남구 삼성동
☎ (02) 511-0068
★ 2.9

유원
서울 송파구 잠실동
☎ (02) 416-7466
★ 4.3

공리
서울 강남구 대치동
☎ (02) 562-0110
★ 4.3

요리하는남자
서울 송파구 잠실동
☎ (02) 419-1511
★ 4.6

www.menupan.com

For more information on Yong-Yeol Ahn and his research, you can read his paper on Flavor Network and the Principles of Food Pairing (http://yongyeol.com/papers/ahn-flavornet-2011.pdf).

Luckily, we will not need to carry out any data collection as the data that we need to meet the goal defined in the business understanding stage is readily available.

We have already acquired the data and placed it on an IBM server. Let's download the data and take a look at it.

Important note:: Please note that you are not expected to know how to program in R. The following code is meant to illustrate the stage of data collection, so it is totally fine if you do not understand the individual lines of code. We have a full course on programming in R, R101 (http://cocl.us/RP0101EN_DS0103EN_LAB2_R), so please feel free to complete the course if you are interested in learning how to program in R.

Using this notebook:

To run any of the following cells of code, you can type **Shift + Enter** to excute the code in a cell.

Get the version of R installed.

```
In [1]:
```

'R version 3.5.1 (2018-07-02)'

Download the data from the IBM server.

```
In [2]:
```

```
[1] "Done!"
```

Read the data into an R dataframe and name it **recipes**.

```
In [3]:
```

Show the first few rows.

In [4]:

A data.frame: 6 × 384

	country	almond	angelica	anise	anise_seed	a
	<fct>	<fct>	<fct>	<fct>	<fct>	<
1	Vietnamese	No	No	No	No	
2	Vietnamese	No	No	No	No	
3	Vietnamese	No	No	No	No	
4	Vietnamese	No	No	No	No	
5	Vietnamese	No	No	No	No	
6	Vietnamese	No	No	No	No	

Get the dimensions of the dataframe.

In [5]:

```
57691
```

In [6]:

```
384
```

So our dataset consists of 57,691 recipes. Each row represents a recipe, and for each recipe, the corresponding cuisine is documented as well as whether 384 ingredients exist in the recipe or not, beginning with almond and ending with zucchini.

Now that the data collection stage is complete, data scientists typically use descriptive statistics and visualization techniques to better understand the data and get acquainted with it. Data scientists, essentially, explore the data to:

- understand its content,
- assess its quality,
- discover any interesting preliminary insights, and,
- determine whether additional data is necessary to fill any gaps in the data.

Thank you for completing this lab!

This notebook was created by Alex Aklson (https://www.linkedin.com/in/aklson/). I hope you found this lab session interesting. Feel free to contact me if you have any questions!

This notebook is part of the free course on **Cognitive Class** called *Data Science Methodology*. If you accessed this notebook outside the course, you can take this free self-paced course, online by clicking here (http://cocl.us/DS0103EN_LAB2_R).