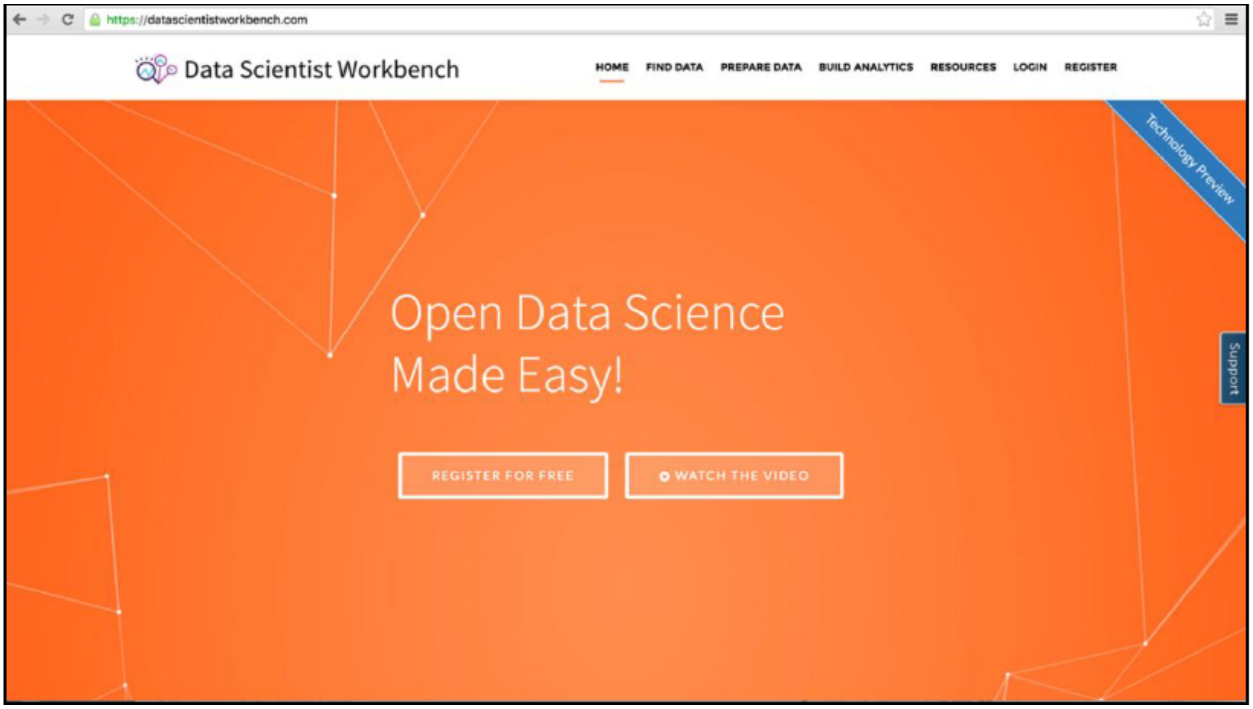
BIG DATA UNIVERSITY

Lab: From Problem to Approach

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.



Open Data Science
Made Easy!

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

Lab 1 Business understanding

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

- Why we're interested in data science.
- What a methodology is.
- Why data scientists need a methodology.
- What the first step of the data science methodology is.
- Two outstanding features of the data science methodology.
- How to apply business understanding to any data science problem (cuisine, in this lab).

Allow 30 minutes to complete this section of the lab.

1.1 Understanding the problem

__1. **This is the Data Science Methodology, a flowchart that begins from Business understanding:**

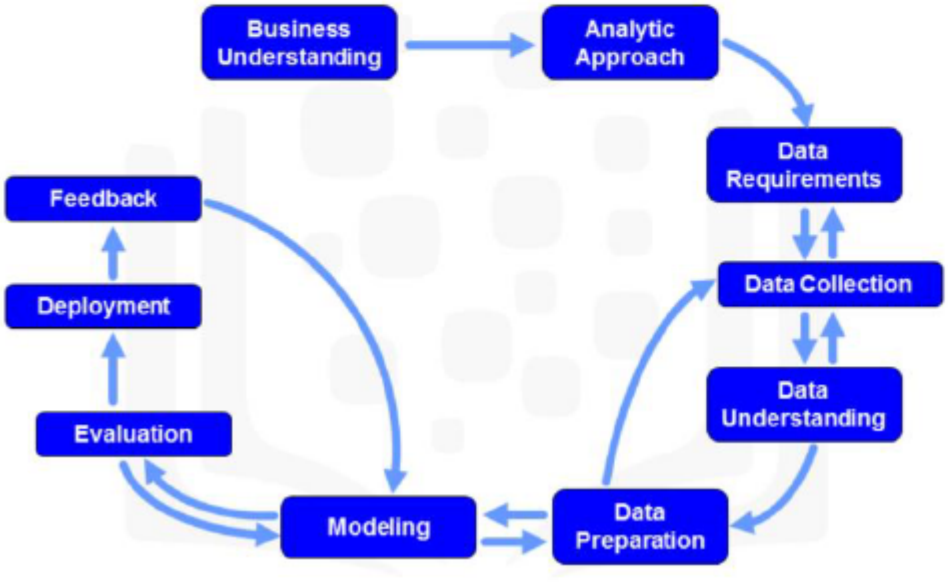
Business Understanding

What basic question does the Business understanding stage answer?

____

____

Looking at this diagram, we immediately spot two outstanding features of our methodology:



What are they?

- ____

  ____

- ____

  ____

__2. So, to get a business understanding of the food ingredients example, in summary:

- Every project starts with a business understanding. **True/False:** ____________

- The business sponsors who need the analytic solution play the most critical role in this stage by defining the problem, project objectives and solution requirements from a business perspective. This first stage lays the foundation for a successful resolution of the business problem. **True/False:** ____________

- To help guarantee the project's success, the sponsors should be involved throughout the project to:

  __i. ____

  __ii. ____

  __iii. ____

__3. Now let's illustrate this with an example of how this might play out, that is, how a business or research problem is first generated by observing the surroundings.

Can we predict what cuisine a food belongs to, based on the ingredients alone? *Circle Yes or No:*

By name?	**By appearance?**	**By smell?**
Yes No	Yes No	Yes No

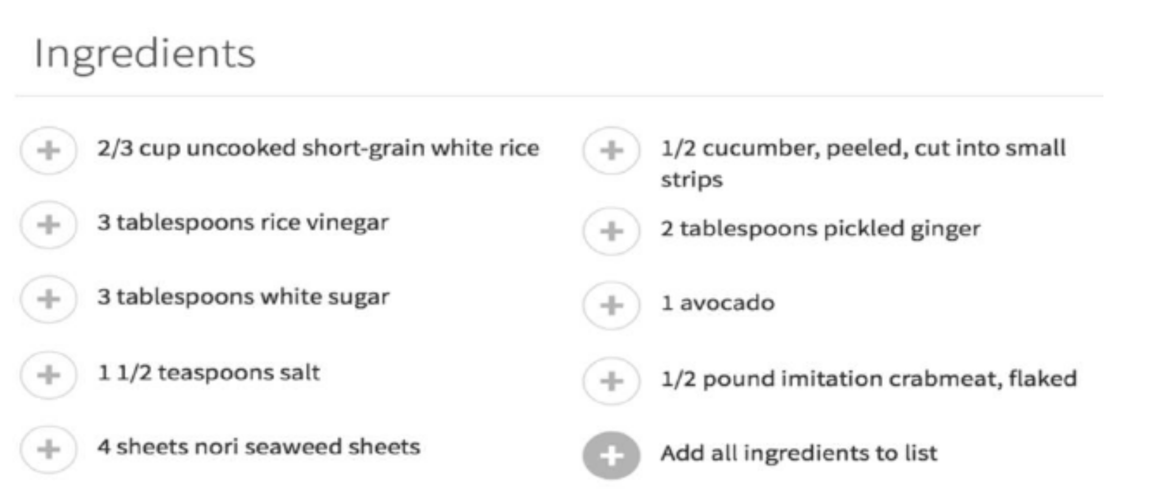
__4. **And then you start thinking about other kinds of food. What comes to mind when you see these dishes?**

Photograph by Avlxyz	
🤔	Do you ever wonder – how are we able to tell what kind of cuisine a food dish is, even if we've never seen it before? **List two reasons how you can tell.**
1.	2.

__5. **Now, we will try going through the methodology as we try to identify the cuisine based on data alone. What cuisine is this?**

2PM **4 minute** **BLT** **Beast**	Answer:

__*6.* **Based on the following ingredients, what cuisine would you say it is?**



Ingredients

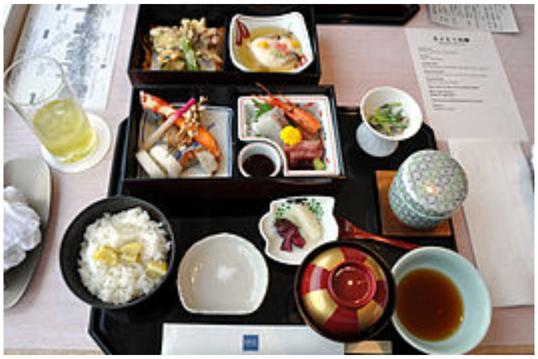
- 2/3 cup uncooked short-grain white rice
- 3 tablespoons rice vinegar
- 3 tablespoons white sugar
- 1 1/2 teaspoons salt
- 4 sheets nori seaweed sheets
- 1/2 cucumber, peeled, cut into small strips
- 2 tablespoons pickled ginger
- 1 avocado
- 1/2 pound imitation crabmeat, flaked

Answer:

__7. What cuisine is this?

Ingredients Rice Seaweed Wasabi Soy Sauce	*Answer:*

__8. What cuisine is this?

 Photograph by cyclonebill (Magnus Manske)	*Answer:*

__9. Exploring ingredients: How is it that we can simply tell what cuisine some food dish belongs to by the ingredients alone? Are certain ingredients more represented by certain countries than others? What if we had data on all the ingredients and recipes in the world? How do cuisines of the world differ by the ingredients they use?
Try to list one ingredient that uniquely is identified with each country below.

Country	**Answer**
1 American	
2 British	
3 Canadian	
4 Chinese	
5 French	
6 Indian	
7 Italian	
8 Japanese	
9 Vietnamese	

Which country would you guess has the most ingredients? ____________________

__10. Of the many techniques and algorithms a Data Scientist should know, list 4 more:

1 Linear Regression

2 ____

3 ____

4 ____

5 ____

1.2 Summary

Congratulations! ...

Lab 2 Analytic approach

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

- How to approach a business problem analytically.
- What the second step of the data science methodology is.
- How to apply analytic approach to any data science problem (cuisine, in this lab).

Allow 30 minutes to complete this section of the lab.

1.3 Determining the analytic approach

__1. Why are we interested in data science?

Once the business problem has been clearly stated, the data scientist can define the **analytic approach** to solving the problem. This step entails expressing the problem in the context of statistical and machine-learning techniques, so that the organization can identify the most suitable ones for the desired outcome.

For example:

- If we want to predict a yes/no response to a promotional offer, then we may define the analytic approach as building, testing, and implementing a

  ______________________________ (which is a type of predictive modeling).

  Furthermore, if we need to see exactly how we arrived at a classification for a

  particular customer, then a _______________________ might be best because it shows the explicit decision path for everyone.

- Or, if we want to better understand the shopping and demographic patterns of

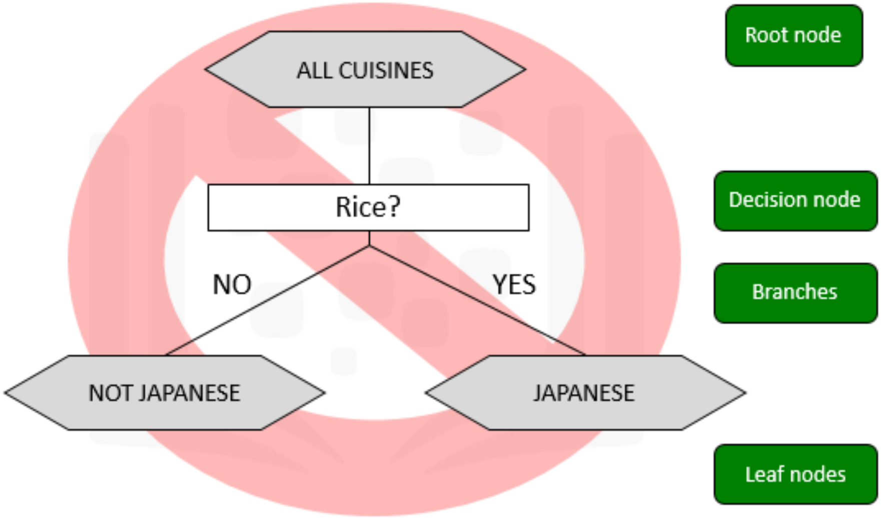
  our customers, then we may combine _________________________________ (which are descriptive modeling methods) to perform market basket analysis for one or more customer segments of interest.

What basic question does the Analytic approach stage answer?

____

____

__2. This is a decision tree that a naive person might create manually. Starting at the top with all the recipes for all the cuisines in the world, if a recipe contains teriyaki sauce, then this decision tree would classify it as Japanese cuisine. Otherwise, it would be "Not Japanese".



In technical terms, at the very top, you start off with your ____________ node, which is the data to be classified.

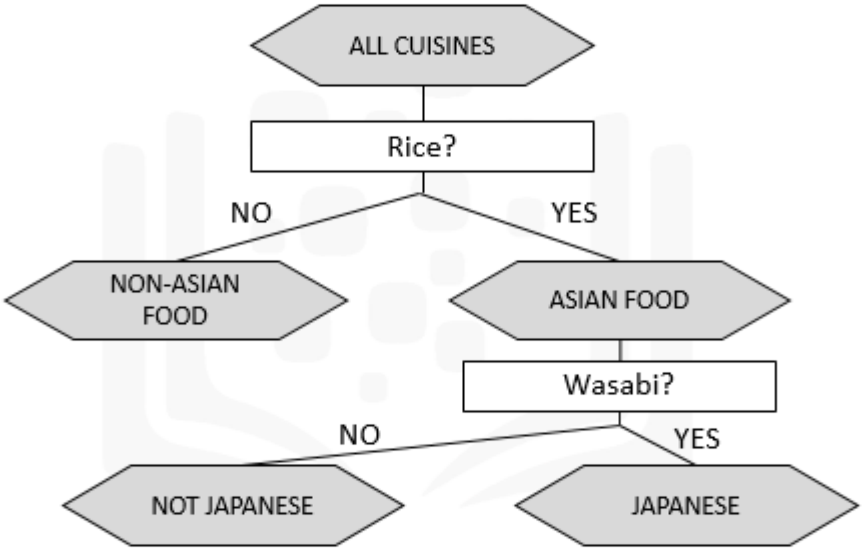
- Then you come down to a ______________ node, which indicates that a decision is to be made on an attribute.
- This decision node splits into _____________ that indicate the possible outcomes of the decision.
- Then, at the very bottom of your decision tree, you end with your ____________ nodes (terminal nodes). If you have more intermediate decision nodes before reaching the leaf nodes, those are called internal nodes (split nodes).

Is this a good decision tree? Yes / No Why? ________________________________

There are other cuisines in the world that use rice in their food. So how might we be able to come up with a more accurate decision tree?

__3. **Using a simple example, this demonstrates what decision trees are. Is the following decision tree more accurate in terms of how you can classify a food dish based on its ingredients?**

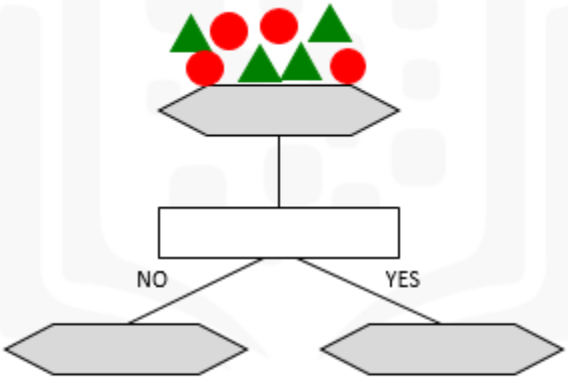
Yes / No / Maybe: ______________________



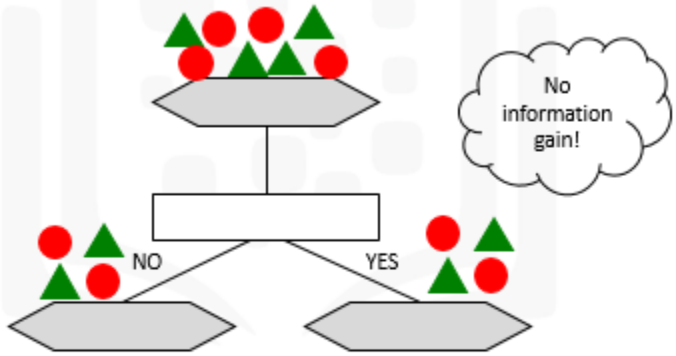
__4. **Understanding decision trees (DTs) – Research Only Exercise**

- DTs are built using recursive partitioning to classify the data.
- The algorithm chooses the most predictive feature to split the data.
- "Predictiveness" is based on decrease in entropy (gain in information) or "impurity".

Suppose that our data is comprised of green triangles and red circles:



This would be considered the optimal model for classifying the data into a node for green triangles and a node for red circles. Each of the classes in the leaf nodes are completely pure – that is, it only contains data of the same class. Information has been gained through the loss of entropy, or “impurity”.



Above is an example of the worst-case scenario that the model could produce.

At the decision nodes, regardless of which branch you take, you end up with the same result in both leaf nodes.

No information is gained from adding this decision, so it might as well _not_ have a tree.

Here are some characteristics of decision trees.

Pros	Cons
Easy to interpret	Easy to overfit or underfit the model
Can handle numeric or categorical features	Cannot model interactions between features
Can handle missing data	Large trees can be difficult to interpret
Uses only the most important features	
Can be used on very large or small data	

A tree stops growing at a node when…

- pure or nearly pure
- no remaining variables on which to further subset the data
- the tree has grown to a preselected size limit

1.4 Summary

Congratulations! ...