**Jupyter slideshow:** This notebook can be displayed as slides. To view it as a slideshow in your browser, type following in the console:

> jupyter nbconvert [this\_notebook.ipynb] --to slides --post serve

To toggle off the slideshow cell formatting, click the CellToolbar button, then View > Cell Toolbar > None.

For more help, check out [this tutorial](https://drive.google.com/open?id=17q01buf7YFuB4yF8cFjmnc_ZARQWOljGhlHs99i9X0c).



What is Data Science?[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#What-is-Data-Science?)

*Authors: Alexander Egorenkov (DC), Amy Roberts (NYC)*

Learning Objectives[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Learning-Objectives)

*After this lesson, you will be able to:*

* Describe the roles and components of a successful development environment.
* Define data science and the data science workflow.
* Apply the data science workflow to solve a task.
* Discuss common data science terminology and processes.

Lesson Guide[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Lesson-Guide)

* [Activity: Data Science in the Real World (5 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#ds-real-world)
* [How to Ask a Question (10 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#question)
* [Data Science Workflow Through Ames Data (20 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#dsfw)
* [Summary (5 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#summary1)
* [Common Machine Learning Definitions (15 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#common-ml-defs)
* [Activity: Quiz or Group (15 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#ml-activity)
* [Summary (5 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#summary2)
* [Course & Project Structure (5 min)](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#course-info)

Activity: Data Science in the Real World[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Activity:-Data-Science-in-the-Real-World)

* List five products or services that you think utilize data science.
* **Examples**
  + Providing movie recommendations on Netflix.
  + Making product suggestions on Amazon.
  + Offering election and sports coverage on the stats site FiveThirtyEight.
  + Calculating daily bet predictions on the fantasy sports site DraftKings.
  + Returning auto-translate and search results on Google.

How to Ask Good Questions[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#How-to-Ask-Good-Questions)

How Do Data Scientists Solve Problems?[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#How-Do-Data-Scientists-Solve-Problems?)

Most practitioners apply a version of the scientific method in order to logically deconstruct and analyze an issue. At General Assembly, we call this the data science workflow, which we've broken down into a series of steps.

This problem-solving framework will help you produce results that are reliable (so that your findings will be more accurate) and reproducible (so that others can follow your steps and achieve the same results).

Note that, depending on the problem, this process is not always linear. You may require lots of iteration and repetition before any conclusions can be drawn!

Asking a Good Question[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Asking-a-Good-Question)

Even though all data science projects have different general flows, they start in the same place: with a problem. From this problem statement arises questions; questions we will ask the data in order to gain more information so we can attempt to find a solution to that problem.

**Why do we need a good question?**

*“A problem well stated is half solved.”* — Charles Kettering

A good question:

* Sets you up for success as you begin analysis.
* Establishes the basis for reproducibility.
* Enables collaboration through clear goals.
  + It's hard to collaborate without a vision.

One way to approach formulating a question is through goal-setting via the SMART Goals Framework:

* **Specific**: The data set and key variables are clearly defined.
* **Measurable**: The type of analysis and major assumptions are articulated.
* **Attainable**: The question you are asking is feasible for your data set and not likely to be biased.
* **Reproducible**: Another person (or future you) can read and understand exactly how your analysis is performed.
* **Time-bound**: You clearly state the time period and population to which this analysis pertains.

Check for Understanding[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Check-for-Understanding)

* Can you make a goal for this class with a partner? Use the SMART Goals Framework to make sure that it's specific, measurable, attainable, reproducible, and time-bound.

What Are Some Common Questions Asked in Data Science?[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#What-Are-Some-Common-Questions-Asked-in-Data-Science?)

**Machine learning more or less asks the following questions:**

* Does X predict Y? (Where X is a set of data and y is an outcome.)
* Are there any distinct groups in our data?
* What are the key components of our data?
* Is one of our observations “weird”?

**From a business perspective, we can ask:**

* What is the likelihood that a customer will buy this product?
* Is this a good or bad review?
* How much demand will there be for my service tomorrow?
* Is this the cheapest way to deliver my goods?
* Is there a better way to segment my marketing strategies?
* What groups of products are customers purchasing together?
* Can we automate this simple yes/no decision?

*This list may seem limited, but we rewrite most questions to fit this form.*

**Instructor Note**: Budget about five minutes to walk through the workflow in general (frame, prepare, analyze, interpret, and communicate) and use the remaining 15 minutes to use the Ames data as a case study on how the workflow is applicable in the real world.

Introduction: The Data Science Workflow[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Introduction:-The-Data-Science-Workflow)

Throughout this course and for our projects, we'll be following a general workflow. This workflow will help you produce *reliable* and *reproducible* results.

* **Reliable**: Accurate findings.
* **Reproducible**: Others can follow your steps and achieve the same results.

Steps in the Data Science Workflow[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Steps-in-the-Data-Science-Workflow)

* **Frame**: Develop a hypothesis-driven approach to your analysis.
* **Prepare**: Select, import, explore, and clean your data.
* **Analyze**: Structure, visualize, and complete your analysis.
* **Interpret**: Derive recommendations and business decisions from your data.
* **Communicate**: Present (edited) insights from your data to different audiences.



Notes about GA's Data Workflow[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Notes-about-GA's-Data-Workflow)

*Remember, these steps are not hard-set rules; instead, think of them as problem-solving guidelines.*

* Some projects may not require every step.
* These steps are iterative; it's normal to go back and repeat certain steps a few times in a row.
* The process is cyclical; after completing the process, you may restart it on new findings.

Application: Data Science Workflow Through Ames Data[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Application:-Data-Science-Workflow-Through-Ames-Data)

Frame[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Frame)

* Identify the business/product objectives.
* Identify and hypothesize goals and criteria for success.
* Create a set of questions to help you identify the correct data set.

We work for a real estate company interested in using data science to determine the best properties to buy and resell. Specifically, your company would like to identify the characteristics of residential houses that estimate their sale price and the cost-effectiveness of doing renovations.

Identify the Business/Product Objectives[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Identify-the-Business/Product-Objectives)

The customer tells us their business goals are to accurately predict prices for houses (so that they can sell them for as large a profit as possible) and to identify which kinds of features in the housing market would be more likely to lead to foreclosure and other abnormal sales (which could represent more profitable sales for the company).

Identify and Hypothesize Goals and Criteria for Success[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Identify-and-Hypothesize-Goals-and-Criteria-for-Success)

Ultimately, the customer wants us to:

* Deliver a presentation to the real estate team.
* Write a business report discussing results, procedures used, and rationales.
* Build an API that provides estimated returns.

Create a Set of Questions to Help You Identify the Correct Data Set[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Create-a-Set-of-Questions-to-Help-You-Identify-the-Correct-Data-Set)

* Can you think of questions that would help this customer deliver on their business goals?
* What sort of features or columns would you want to see in the data?

**Instructor Note:** Use the check for understanding above to guide them to:

* "Fixed" characteristics, such as neighborhood, lot size, number of stories, etc.
* Sale prices over time.
* Optimizing for the difference between purchase price and sale price (bigger values are better).

**Ideal Data vs. Available Data**

Oftentimes, we'll start by identifying the *ideal data* we would want for a project.

Then, during the data acquisition phase, we'll learn about the limitations on the types of data actually available. We have to decide if these limitations will inhibit our ability to answer our question of interest or if we can work with what we have to find a reasonable and reliable answer.

For example, we provide a set of housing data for Ames, Iowa, which [includes](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/extra-materials/ames_data_documentation.txt):

* 20 continuous variables indicating square footage.
* 14 discrete variables indicating number of each room type.
* 46 categorical variables containing 2–28 classes each, e.g., street type (gravel/paved) and neighborhood (city district name).

**Optional Check**[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Optional-Check)

Take a moment to look through the data description. How closely does the set match the ideal data that you envisioned? Would it be sufficient for our purposes? What limitations does it have?

This is possibly the hardest step in the data science workflow. At this stage, it's common to realize that the problem you're trying to solve may not be solvable with the information available. The data could be incomplete, non-existant, or unable to meet the criteria necessary to answer your question.

That said, you now have a better feel for the data that's available and the information they could contain. You can now identify a new, answerable question that ultimately helps you solve or better understand your problem.

Data acquisition[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Data-acquisition)

* **What are some questions we should ask during the acquisition process?**
* Our Ames data set contains the following information:
  + [Ames Data Set Introduction PDF](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/extra-materials/ames.pdf) (from the "Journal of Statistics Education")
  + "Data set contains information from the Ames Assessor’s Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010."

Data Quality[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Data-Quality)

* **What are some questions we should ask when checking the data for quality?**
  + [Ames Data Set Documentation](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/extra-materials/ames_data_documentation.txt)

**Instructor Note**: During the **Framing** phases, guide students toward the following questions:

* Where are the data coming from?
* How do the data fit together?
* Are there enough data?
* Do our data appropriately align with the question/problem statement?
* Can the data set be trusted? How was it collected?
* Is this data set aggregated? Can we use the aggregation, or do we need to obtain it pre-aggregation?
* What are necessary resources, requirements, assumptions, and constraints?
* Can we import data from the web (Google Analytics, HTML, XML)?
* Can we import data from a file (CSV, XML, TXT, JSON)?
* Can we import data from a pre-existing database (SQL)?
* Can we set up local or remote data structures?
* What are the most appropriate tools for working with the data?
* Do these tools align with the format and size of the data set?

Prepare[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Prepare)

Often, we are given *secondary data*, or data that were collected previously. In these cases, we have to learn as much as possible about our data using tools like data dictionaries and source documentation to determine how the set was gathered.

Here's an example of a data dictionary:

| **Variable** | **Description** | **Type of Variable** |
| --- | --- | --- |
| Square Footage | Floating Point | Continuous |
| Street Type | 1 - Gravel, 2 - Paved | Categorical |
| Neighborhood | String, e.g., 'Tenderloin' | Categorical |
| Number of Bedrooms | Integer | Discrete |

**Common considerations when preparing our data include:**

* **Ensuring data is clearly defined and structured**
* **Check and clean data formatting as needed**

**Common considerations for cleaning include**:

* **Most data will** not **come perfectly clean and ready to use. Cleaning data is normally the most time-consuming task a data scientist faces.**

As you can see, the "Prepare" phase of the data science workflow encompasses several steps: the act of reviewing, indexing, and cleaning your data. This normally consumes a great deal of time!

**Instructor Note:** During the **Preparing** process, guide students toward the following steps:

* Reading any documentation provided with the data (e.g., the data dictionary above).
* Performing exploratory surface analysis via filtering, sorting, and simple visualizations.
* Describing the data structure and the information being collected.
* Exploring variables and data types via SELECT statements.
* Assessing preliminary outliers and trends.
* Verifying the quality of the data (feedback loop -> 1).
* Sampling the data and determining sampling methodology.
* Iterating and exploring outliers and null values via SELECT statements.
* Assessing qualitative versus quantitative data.
* Formatting and cleaning data in Python (e.g., dates, number signs, formatting).
* Defining how to appropriately address missing values (cleaning).
* Categorizing, manipulating, slicing, formating, and integrating data.
* Formatting and combining different data points, separate columns, etc.
* Determining most appropriate methods for aggregating, cleaning, etc.
* Creating the necessary derived columns from the data (new data).

Analyze[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Analyze)

As an example of basic statistics, Data scientists often check the mean, standard deviation, or specific frequency counts of their data. Statistics that we might expect for the earlier housing variables include:

| **Variable** | **Mean or Frequency (%)** |
| --- | --- |
| Square Footage | 2201.3 |
| Street Type - Gravel | 8% |
| Street Type - Paved | 92% |
| Number of Bedrooms | 1.8 |

**What sort of questions do these types of statistics allow us to answer? Why would we do this?**

Creating a Predictive Model[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Creating-a-Predictive-Model)

We generate predictive models based on the SMART goal we decided upon earlier. Typically, our interest is in predicting or guessing some sort of value we might be interested in (such as the housing price for a house given some set of fixed characteristics).

\**What are some other business goals we can support as data scientists for this realty company? What are some values we would like to guess?*

**What do you think are the steps for model building?**

*We'll be spending much of our time in this course on data analysis and predictive modeling.*

**Instructor Note:** For the **Analyzing** phase, guide students toward the following considerations:

* Identifying trends and outliers.
* Deciding how to deal with outliers — excluding, filtering, and communication.
* Applying descriptive and inferential statistics.
* Determining initial visualization techniques.
* Documenting and capturing knowledge.
* Choosing visualization techniques for different data types.
* Transforming data.

**Instructor Note:** For the **Interpreting** phase, guide students toward the following considerations:

* Selecting the appropriate model.
* Building the model.
* Evaluating and refining the model.
* Predicting outcomes and action items.

Interpret[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Interpret)

Develop Recommendations and Decisions[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Develop-Recommendations-and-Decisions)

**Now that you have a model, what are some things you should check?**

**Now that you have a model, can you convert your model's finding into a conclusion or next step for your employer?**

**Instructor Note:** For things to check after a model, guide students toward the following questions:

* Did you reject or fail to reject your hypotheses?
  + What does this mean for your project?
  + What does this mean for your client?
* Were your questions answered?
  + Which ones?
  + What do you need to do to answer the ones that weren't?
* Do your findings support any business recommendations, actions, or decisions?
  + Is there further supportive analysis?
  + How do your data support these recommendations?

**Instructor Note:** For the **Communication** phase, guide students toward the following questions:

* Reaching a conclusion:
  + Seek guidance/interaction with subject matter experts (SMEs).
  + If those are not available, check with the data — are you coming to reasonable conclusions and predictions given what you've seen?
  + Do the next steps that you envision have any dependencies or corollary steps?
* What are some conclusions you can draw?
  + Conclusion: "Customers from large companies were twice as likely to place another order with Planet Express than customers from small companies."
  + Recommendation: "We should target more large companies to use our delivery service."
  + Conclusion: "Other than size of company, I found no significant evidence that any other feature affected the odds of customers reusing our delivery service."

Communicate[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Communicate)

Share the Results of Your Analysis[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Share-the-Results-of-Your-Analysis)

Presentations are a critical part of your analysis. It doesn't matter how brilliant your model is or how illuminating your findings are — without effective communication, your work will not be used.

The most basic form of a data science presentation should include a simple sentence that describes your results:

*"Customers from large companies had twice (CI 1.9, 2.1) the odds for placing another order with Planet Express compared to customers from small companies."*

Data science presentations can also be far more complex and exciting, like some of the [research presented by Nate Silver's FiveThirtyEight blog](http://fivethirtyeight.com/burrito/#brackets-view).

When crafting a presentation, always consider your audience and make sure to practice your presentation beforehand. Consider the types of questions people might ask or — better yet — test your presentation on a few people and pay attention to their response. Clarify and refine your presentation accordingly.

Make sure to consider your needs and goals, as well as those of your audience. A presentation created for your fellow data scientists will be vastly different than a presentation intended for executives trying to make a business decision.

**A Note About Iteration**

Iteration is an important part of *every step* in the data science workflow. At any given point in the process, you may find yourself repeating or going back and redoing steps in order to better understand your data, clarify your model, and refine your presentation.

**What are some things you may want to redo or iterate over after presenting your findings?**

**Instructor Note:** During iteration process, guide students to the following steps:

* Identifying follow-up problems and questions for future analysis.
* Creating a visually effective summary or report.
* Adding more features.
* Considering the needs of different stakeholders and how your report might be changed for them.
* Identifying the limitations of your analysis.
* Identifying relationships between visualizations.

Summary[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Summary)

1) **Crafting good questions is key.**

* Without a thoughtful, targeted, and SMART question, it can be difficult to create an effective model. 2) **Use the data science workflow to iteratively develop solutions.**
* **Frame**: Develop a hypothesis-driven approach to your analysis.
* **Prepare**: Select, import, explore, and clean your data.
* **Analyze**: Structure, visualize, and complete your analysis.
* **Interpret**: Derive recommendations and business decisions from your data.
* **Communicate**: Present (edited) insights from your data to different audiences. 3) **Informed by your past work, continue to refine your findings and models.**
* While the data science workflow may appear to be linear, we consistently return to past steps to implement new findings

Introduction: Machine Learning[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Introduction:-Machine-Learning)

Examples of Machine Learning[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Examples-of-Machine-Learning)

**Instructor Note**: This is a good section in which to provide your own work (or side project) experience as well! These are just a couple of options:

* [Google Quick Draw](https://quickdraw.withgoogle.com/)
* [Deep Dream Generator](https://deepdreamgenerator.com/)
* Add your own!

Common Machine Learning Definitions[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Common-Machine-Learning-Definitions)

There are two main categories of machine learning: supervised learning and unsupervised learning.

**Supervised learning (a.k.a., “predictive modeling”):**  
*Classification and regression*

* Predicts an outcome based on input data.
  + Example: Predicts whether an email is spam or ham.
* Attempts to generalize.
* Requires past data on the element we want to predict (the target).

**Unsupervised learning:**  
*Clustering and dimensionality reduction*

* Extracts structure from data.
  + Example: Segmenting grocery store shoppers into “clusters” that exhibit similar behaviors.
* Attempts to represent.
* **Does not require** past data on the element we want to predict.

Oftentimes, we may combine both types of machine learning in a project to reduce the cost of data collection by learning a better representation. This is referred to as transfer learning.

Unsupervised learning tends to present more difficult problems because its goals are amorphous. Supervised learning has goals that are almost too clear and can lead people into the trap of optimizing metrics without considering business value.

Supervised Learning[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Supervised-Learning)

Supervised learning tends to be the most frequent type of work that data scientists do and will be the main focus of this course. How does supervised learning work?

1) We train a **machine learning model** (more on that shortly) using **labeled data** (the "response" label from earlier).

- The “machine learning model” learns some kind of relationship between the features and the response.

2) We make predictions on **new data** for which the response is unknown.

The primary goal of supervised learning is to build a model that “generalizes” — i.e., accurately predicts the **future** rather than the **past**!

Practice: Classification vs. Regression[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Practice:-Classification-vs.-Regression)

There are two categories of supervised learning:

**Regression**

* The outcome we are trying to predict is a continuous value.
  + **Can you think of anything we would want to predict like this?**

**Classification**

* The outcome we are trying to predict is categorical (i.e., it comes in one of a set number of classes).
  + **Can you think of anything that we would want to predict like this?**

The type of supervised learning problem has nothing to do with the features; only the response matters!

**Instructor Note:** Examples of regression targets include price, blood pressure, temperature, etc.

**Instructor Note:** Examples of classification include spam/ham, cancer class of tissue sample, etc.

Unsupervised Learning[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Unsupervised-Learning)

Common Types of Unsupervised Learning[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Common-Types-of-Unsupervised-Learning)

* **Clustering:** Groups “similar” data points together.
* **Dimensionality reduction:** Reduce the dimensionality of a data set by extracting features that capture most of the variance in the data.

**Steps for Clustering**

Imagine that we had a bunch of coins we wanted to automatically split into groups. An unsupervised learning technique would involve the following steps:

1) Clustering the coins based on “similarity" — this could be through the size, the material, or the language on the coins.   
2) Inspecting the grouping that the algorithm found.

Hopefully this would put the coins into sets of related groups.

**Steps for Dimensionality Reduction**

Imagine that we had a huge amount of features related to those coins — country of origin, size, weight, mass, density, condition, chemical makeup, etc. Moreover, say that we had thousands or (in some cases, millions) of different features. Not all of these features are helpful, however! Unsupervised learning can help us by grouping features together automatically. It would involve the following:

1) The unsupervised learning technique groups or combines features that are similar or do the same thing into a smaller set, leading to a set of new features that's smaller in size.

Hopefully, the algorithm would recognize something like.

*masssize* =*density* masssize=density

Here, density could take the place of two different features from before.

Sometimes unsupervised learning is used as a “preprocessing” step for supervised learning. (Can you guess how?)

**Instructor Note:** It can be used as a preprocessing step to provide a new set of features for a structured machine learning technique down the line.

Examples[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Examples)

**Supervised Learning: Coin Classifier**

* **Observations:** Coins.
* **Features:** Size and mass.
* **Response or target variable:** Hand-labeled coin type.
* Train a machine learning model using labeled data.
  + The model learns the relationship between the features and the coin type.
* Make predictions on new data for which the response is unknown.
  + Give the model a new coin and it will predict the coin type automatically.

**Unsupervised Learning: Types of Customers at a Bar**

* **Observations:** Customers.
* **Features:** Drink purchases, people they interact with, etc.
* **Response or target variable:** There isn’t one — instead, we group similar customers together.

<a id = 'algorithm'></a>

Algorithms[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Algorithms)

Regardless of whether it's supervised or unsupervised, the underlying engine driving a machine learning model is an algorithm. These algorithms are used to help identify trends, represent said trends, and explain the overall variance of the data.

Let's say we are a real estate agent looking to price a house using only its square footage. We know there are other features that can highly influence this outcome, but we are only focusing on square footage for now. We know that, as square footage increases, so does price. At this point, you may be thinking that a simple algebra equation could be useful; one that helps us price the house by its square footage.

Recently, we sold a house whose square footage was 2,500 for about $285,000. If we apply this information to a normal linear equation — *Y*=*mx*+*b* Y=mx+b — we can create a simple *algorithm* to help us predict a house.

285,000=2,500*x*+*b* 285,000=2,500x+b

*x*=114,*b*=0 x=114,b=0

*The Y intercept has been omitted for this example.*

Final Algorithm[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Final-Algorithm)

*Price*=114*x* Price=114x

This is an example of a model built with the intent of predicting price. The algorithm is simple and built off of limited information. Typically, our models will be more complex, and we'll consider a greater amount of prior data to help us develop a final algorithm.

Algorithm Training[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Algorithm-Training)

In our example, we used previously known information to find our coefficients. This action is also referred to as "training." But, let's make something clear:

* Model building would be the task of constructing an actual algorithm.
  + This is the linear model of *Y*=*mx*+*b* Y=mx+b .
* Training involves figuring out the coefficient and the Y intercept the model uses for *our intended purpose*.
  + The coefficients uncovered via training were *m*=114 m=114 and *b*=0 b=0 .

Partner Activity[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Partner-Activity)

Partner with someone sitting next to you and tackle the following challenge! We'll spend a few minutes working together and then check in with one or two groups to see what they've come up with.

You'll need the [Ames Data Set documentation](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/extra-materials/ames_data_documentation.txt) and, optionally, the [Ames Data Set Introduction PDF](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/extra-materials/ames.pdf):

**Your Task**

With a partner, sketch out answers to the following:

* What is a potential target in your data for a regression model?
* What is a potential target in your data for a classification model?
* Could unsupervised learning be used within this data? How so?

Once you've done so:

* Together, pick one of your targets and sketch out what a data science workflow would look like for that question. Don't forget to identify what you think would be most important during each step:
  + **Frame**: Develop a hypothesis-driven approach to your analysis.
  + **Prepare**: Select, import, explore, and clean your data.
  + **Analyze**: Structure, visualize, and complete your analysis.
  + **Interpret**: Derive recommendations and business decisions from your data.
  + **Communicate**: Present (edited) insights from your data to different audiences.

Feel free to avoid getting too specific — this exercise is meant to help you practice approaching a problem the way a data scientist would.

Conclusion[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Conclusion)

Check to see if you can answer the following questions easily:

* What is data science?
* What is the data science workflow?
* What is the difference between supervised and unsupervised learning?
* What is the difference between regression and classification?
* What is an algorithm?

Course Information[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Course-Information)

GA offers a special learning environment.[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#GA-offers-a-special-learning-environment.)

* What you should know: GA is a global community of individuals and organizations empowered to pursue the work we love.
* Who we are: Meet your instructional team.
* How to provide feedback: exit tickets, mid-course survey, and end-of-course survey. We want to hear from you!

Road to Success[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Road-to-Success)

* The emotional cycle of change: This course is fast and covers a lot of material. There will be times when you may feel discouraged or overwhelmed, but don't give up - this is natural (and part of the design). By the end of the course, you'll feel more confident in your ability to define problems, analyze data, and prototype solutions.
* Student learning responsibility: Our lessons cover topic foundations, but there is always more to learn! You are responsible for your learning experience - but don't get overwhelmed! Instead, just make sure you follow along, practice as much as possible, and ask questions.
* GA requirements: Show up. Be on time. Participate. Submit your projects. Allow yourself to struggle. Read the docs. Have fun!
* Q/A.

Course Outline and Project Due Dates[¶](http://localhost:8888/nbconvert/html/Desktop/SCHOOL_WORK/Week%201/welcome-to-data-science-master/course_info_definitions.ipynb?download=false#Course-Outline-and-Project-Due-Dates)

General Assembly's part-time Data Science materials are organized into **four** units.

| **Unit** | **Title** | **Topics Covered** | **Length** |
| --- | --- | --- | --- |
| Unit 1 | Foundations | Python Syntax, Development Environment | Lessons 1–4 |
| Unit 2 | Working with Data | Stats Review, Visualization, & EDA | Lessons 5–9 |
| Unit 3 | Data Modeling | Regression, Classification, & KNN | Lessons 10–14 |
| Unit 4 | Applications | Decision Trees, NLP, & Flex Topics | Lessons 15–19 |

**Instructor Note:** If there is time, briefly walk through the entire course-info repository with your students. If not, refer them to it for class information.