Data Science, lesson 1

Welcome to Data Science

GENERAL ASSEMBLY

Meet Your Instructor

Vlada Rozova

vlada.rozova@generalassemb.ly

- Currently a PhD candidate at Macquarie Uni.
- Started my journey as a Data Scientist in a pharmaceutical company AstraZeneca.
- Adore sausage dogs!

Meet Your Teaching Assistant

Shel Zaroo

shel.zaroo@generalassemb.ly

- Local Data Science Instructor at GA
- Previously worked as Data Science Contractor
- Binge watch Crime thrillers

Your turn

About you

- Your name?
- Why Data Science?
- Fun fact about yourself!

Agenda

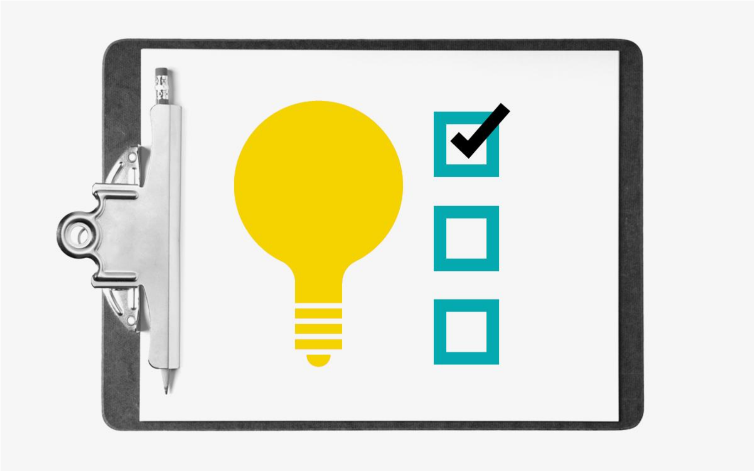
Here’s what we’ll be covering:

- Data Science Definitions
- Development Environment
- Python practice

Learning Objectives

After this lesson, you'll be able to:

- **Define** the Data Science Workflow and common Machine Learning concepts.
- **Discuss** the topics and goals of our course.
- **Set up** and confirm your development environment.
- **Use** types in Python correctly.
- **Create** basic functions in Python.

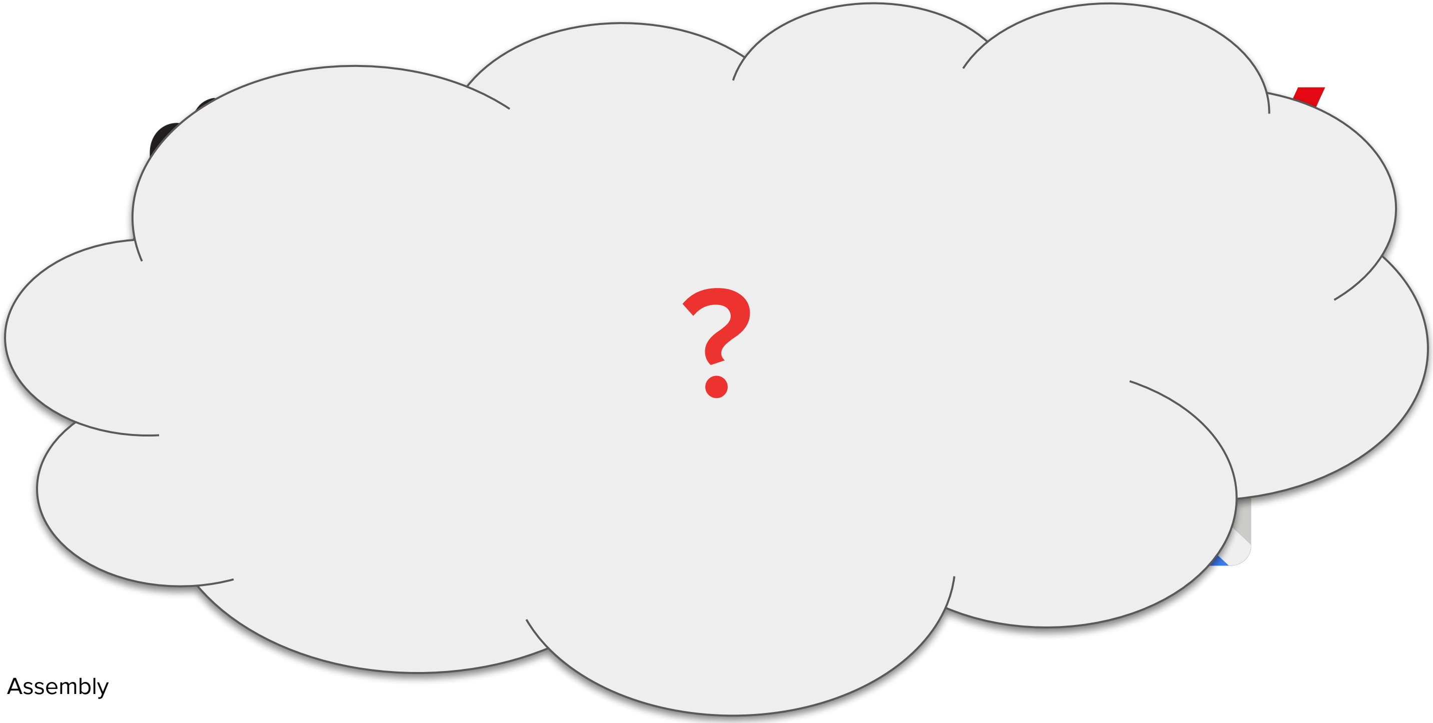


Data Science

Data Science definitions

Discussion:
Data Science in the Real World

List five products or services that you think utilise Data Science.



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 arise questions.
- These questions we will ask the data in order to gain more information so we can attempt to find a solution to that problem.
- Be specific! This will significantly speed up your analysis.

“

A problem well stated is half solved.

Charles Kettering

How to Ask a Good Question?

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.

What Are Some Common Questions Asked in Data Science?

Machine learning more or less asks the following questions:

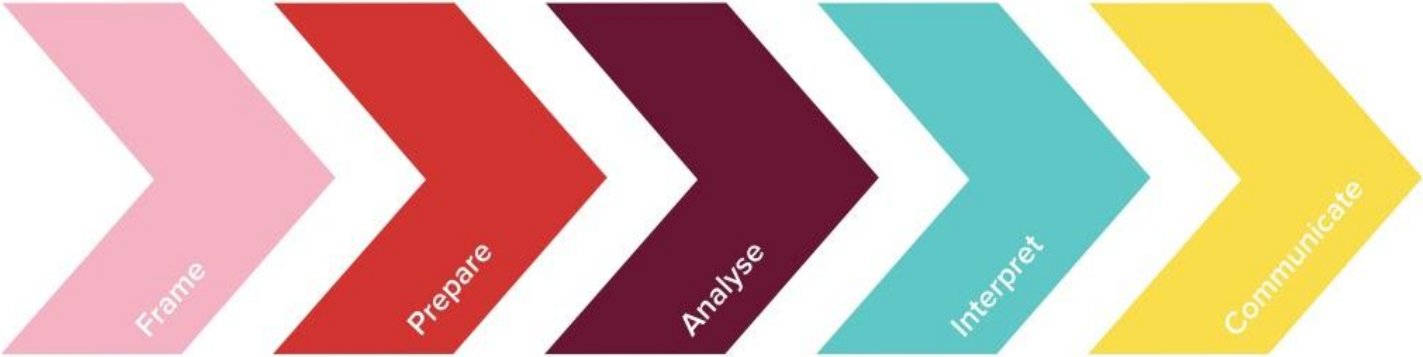
- Does X predict Y? (Where X is a set of data and y is an outcome.)
- Are there any distinct groups in our data?
- What are the key components of our data?
- Is one of our observations “weird”?

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?

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.



Develop a hypothesis-driven approach to your analysis

Select, import, explore, and clean your data

Structure, visualise, and complete your analysis

Make recommendations and business decisions from your data

Present insights from your data to different audiences

The workflow is an **iterative** process, not necessarily a linear one!

Group Exercise:

Data Science Workflow through Ames data

15 minutes

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.

Group Exercise:

Data Science Workflow through Ames data

15 minutes

Frame

1. Identify the business/product objectives.
2. Identify and hypothesize goals and criteria for success.
3. Create a set of questions to help you identify the correct data set.
4. Ideal data vs. available data.
5. What are some questions we should ask during the acquisition process?
6. What are some questions we should ask when checking the data for quality?

Group Exercise:

Data Science Workflow through Ames data

15 minutes

Prepare

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.

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

Group Exercise:

Data Science Workflow through Ames data

15 minutes

Analyse

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 real estate company? What are some values we would like to guess?
- What do you think are the steps for model building?

Group Exercise:

Data Science Workflow through Ames data

15 minutes

Interpret

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?

Group Exercise:

Data Science Workflow through Ames data

15 minutes

Communicate

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."

Fantastic example: research presented by Nate Silver's FiveThirtyEight blog.

Group Exercise:

Data Science Workflow through Ames data

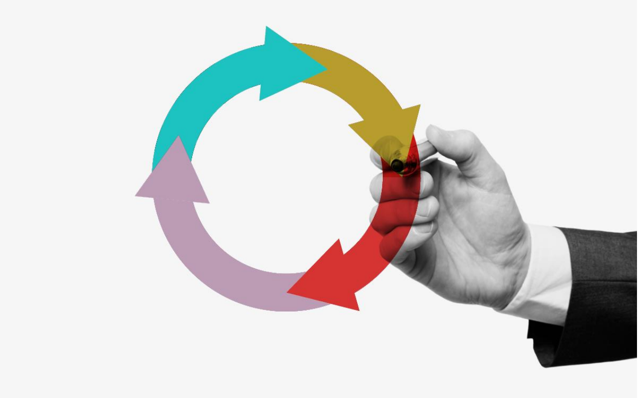
15 minutes

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.

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.



A note about iteration

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

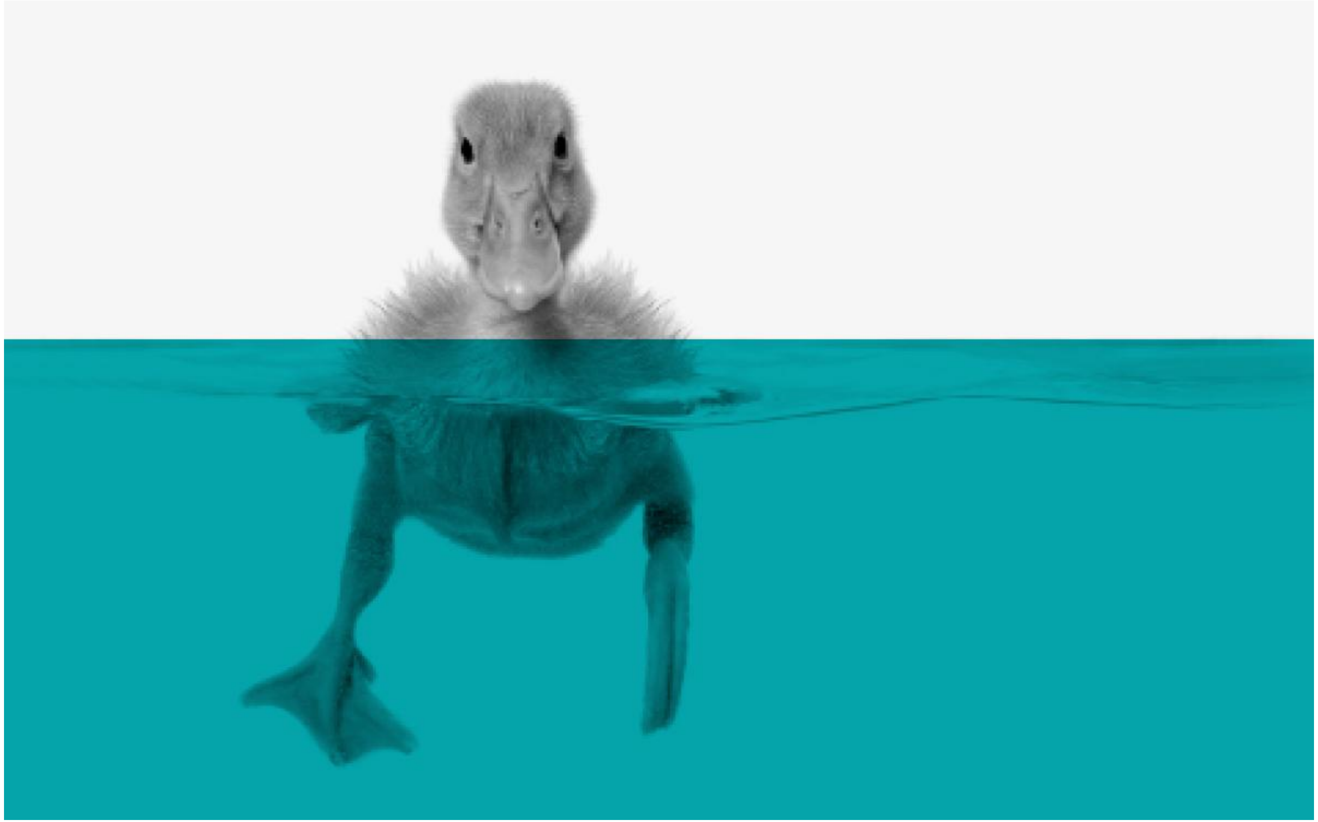
Computers Out: Signing up for GitHub

2 minutes

BEFORE WE BEGIN...

If you have not already done so please create an account on GitHub and share your username on Slack.

Introduction to Machine Learning



Supervised learning:

Classification and regression

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

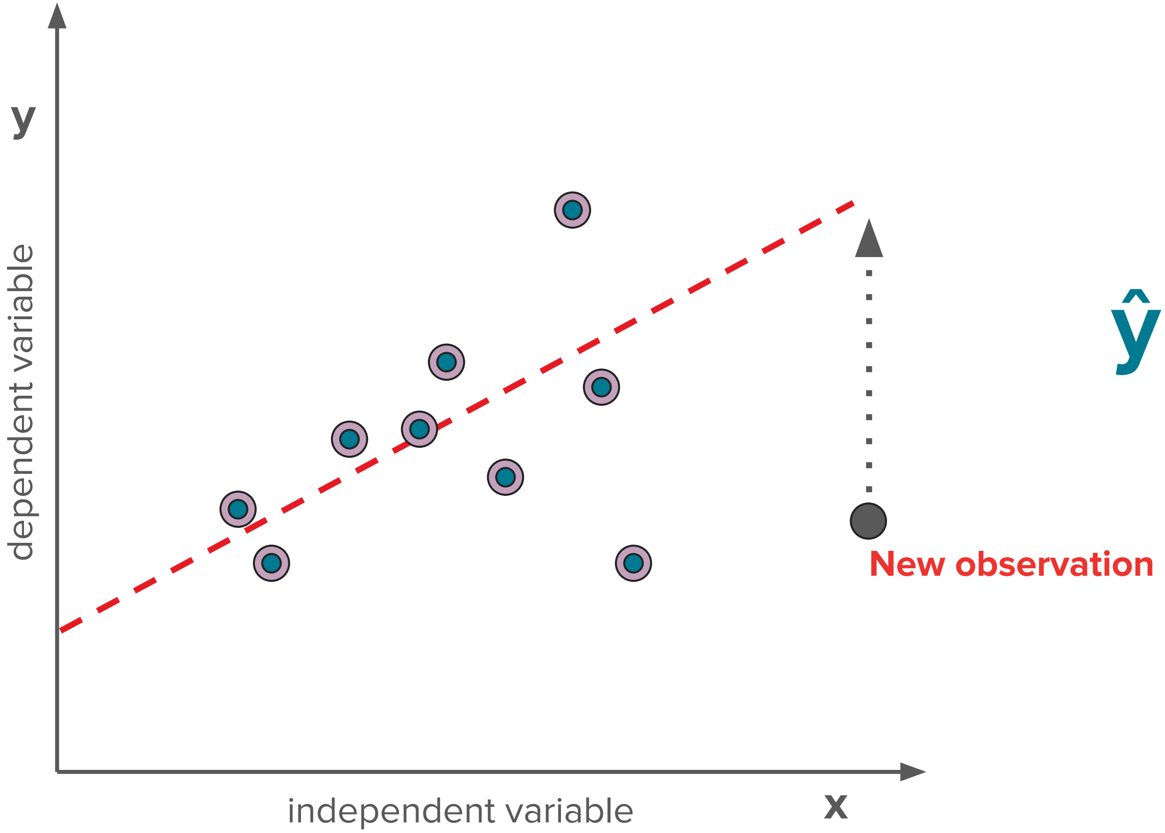
Unsupervised learning:

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.

More details here

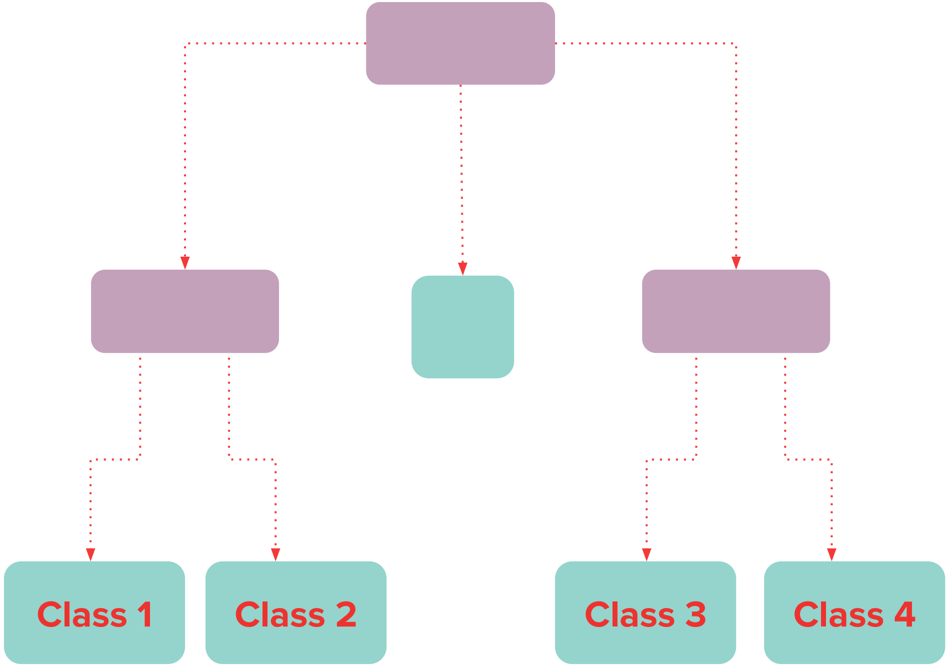
Linear Regression Model



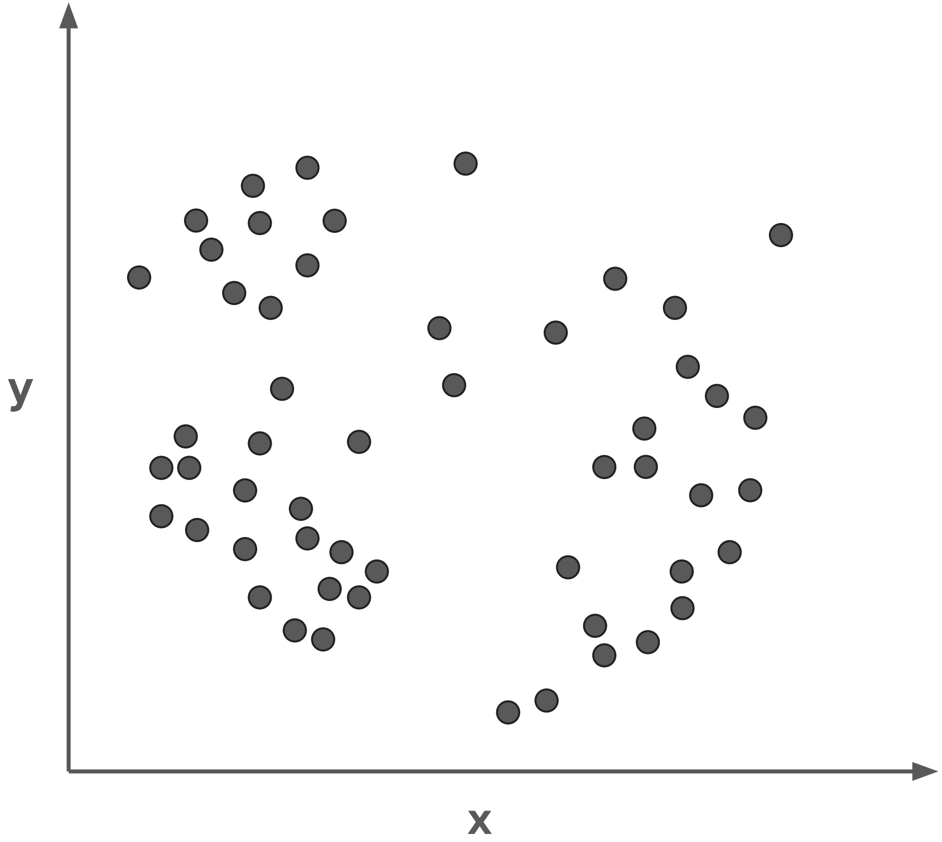
New observation

$$\hat{y} = b_0 + b_1 * x_1$$

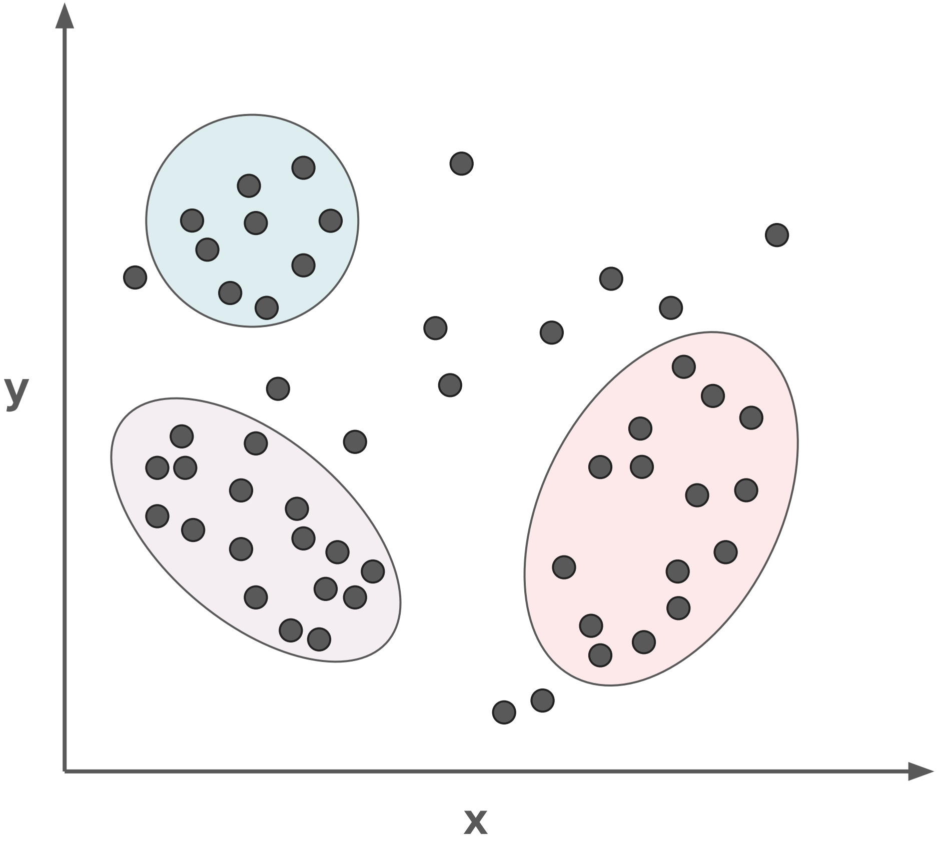
Decision Tree Classification Model



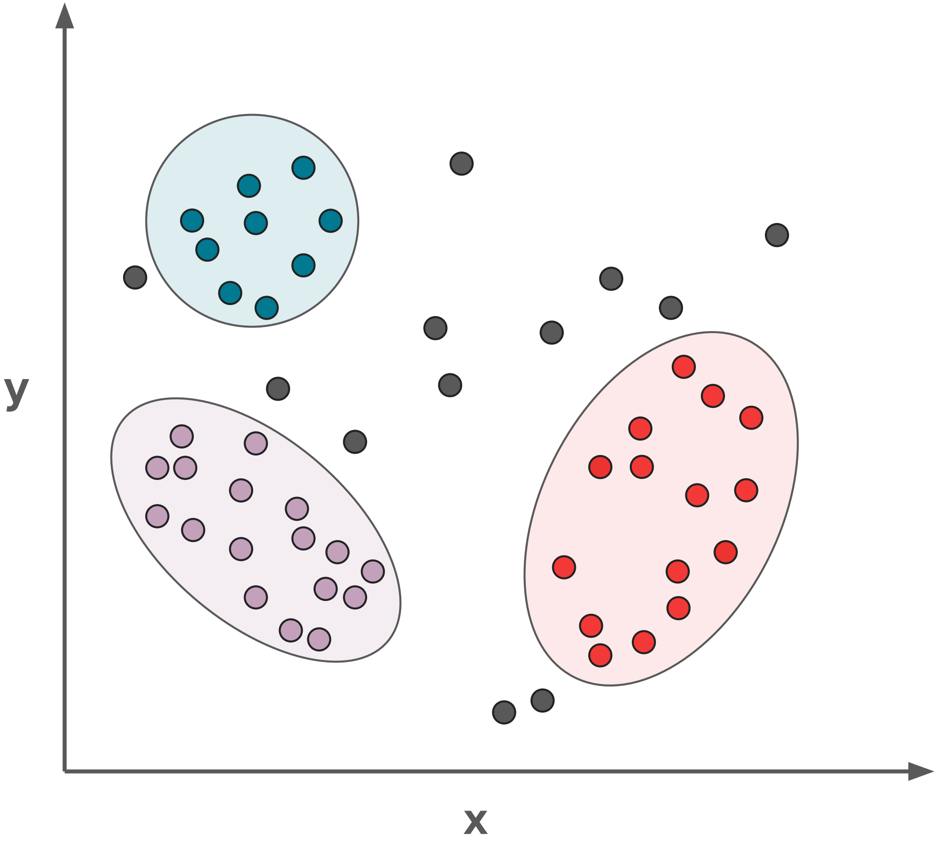
Cluster Model - Process - Step 1



Cluster Model - Process - Step 2



Cluster Model - Process - Step 3 - Labeled



Cluster A
Cluster B
Cluster C

KNOWLEDGE CHECK

Imagine in a grocery store you put a bag of veggies on a scale and it automatically tells you these are carrots. This task is an example of:

a. Regression

b. Classification

KNOWLEDGE CHECK

You are working at a winery and trying to predict the price of the new wine based on various conditions (region, weather, quality of soil, etc). This is an example of:

a. Regression

b. Classification

Computers Out: Installation check

15 minutes

Installation check:

1. Git bash shell for Windows
2. Anaconda installation
3. Git configuration

Ames dataset:

1. Documentation
2. Introduction PDF
3. ames_housing.ipynb

Partner Exercise:

Application of Machine Learning

10 minutes

You'll need the Ames Data Set documentation and, optionally, the Ames Data Set Introduction 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?

Deliverables

- 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.

Discussion:

Important Takeaways

5 minutes

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?

Data Science

Course Info

Course Information

Road to Success

- 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. Fill out exit tickets. **Have fun!**

Course Outline

General Assembly's part-time Data Science course consists of 20 classes organized into **four** units.

During the next 10 weeks we will cover the following topics:

- Python Syntax and Development Environment;
- Statistics, Data Visualisation techniques and Exploratory Data Analysis;
- Machine Learning Basics including models for both Regression and Classification;
- Advanced topics such as Decision Trees & Random Forest, Clustering, Natural Language Processing, and Time Series Forecasting.

Let’s go through the detailed course overview.

Data Science

Break

Data Science

Python Fundamentals

Computers Out:

Python fundamentals

1. Download `intro_to_python_fundamentals.ipynb` from Slack and move it to your Desktop.
2. Open your Terminal and type: `jupyter notebook`
3. Your browser will automatically open. Navigate to Desktop and click on `intro_to_python_fundamentals.ipynb` to start Jupyter Notebook.

Data Science

Conclusion

Discussion:

Important Takeaways

5 minutes

Let’s Debrief

- What did you learn in this lesson?
- What was challenging about this topic?
- What questions do you still have?
- What are some next steps you could take with this information?

Discussion:

How Are We Doing?

10 minutes

Share Out

- What have we accomplished so far? Where are we going next?
- What is one thing you've taken away from today's lessons so far?
- What do you feel good about? Where do you need more information?

Finish That Sentence

What are your biggest takeaways from today?

"Something that really got me thinking is…"

"The best thing I got out of today is…"

"I discovered…"

"I still want to learn about…"

"I was surprised that…"

Data Science

Q&A

Ask Me Anything!

Data Science

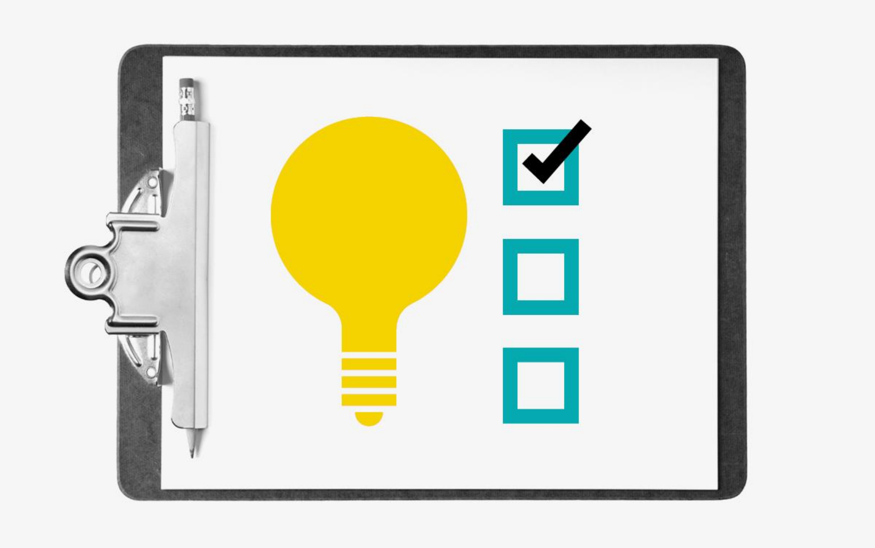
Next Steps

Before Next Class:

A. Review Data Science Workflow
B. Think of other examples of Supervised vs Unsupervised Learning
C. Practice your Python!

A Few Good References

- Data Science Interview Guide
- Kaggle Learn
- More on how to use GitHub



See you next time!

Thank you!

GENERAL ASSEMBLY