GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

Dima Galat
Chief Computing Officer, Gronade

PRE-WORK REVIEW

- Bring a laptop with Anaconda installed. Scroll to your operating system version and click on the install button for Anaconda with Python 3.6.

- We will be using Jupyter Notebooks as the main IDE for the workshop. If you have installed Anaconda, then you are ready to go!

LEARNING OBJECTIVES

- Explain the field of data science, defining common roles & trends.
- Explore popular tools & resources to visualize, analyze, & model data.
- Recognize the types of problems that can be solved by data science.
- Apply the data science workflow to provide real world recommendations.
- Create a custom learning plan to build your data science skills after this workshop!

OPENING

ABOUT ME

- Here's a bit about me:

Name | Background | Fun Fact

ABOUT YOU

- Before we dive in, let’s talk a bit about you!

- Name
- What brings you to GA?
  - Current activities
  - Your Goals
- Fun fact about yourself!

OUR EXPECTATIONS

- You're ready to take charge of your learning experience.
- You're curious and excited about data science!

THE BIG PICTURE

- What we'll cover:
  - Why data science & what it can do for me?
  - Data science skills
  - Explore the Data Science Toolkit
  - Analyse data
  - Algorithms in action

THE BIG PICTURE

- Why this topic matters:
  - Data science is a sought-after skill
  - Using Python due to its increased popularity and simplicity

- Why this topic rocks:
  - Data science opens up a door to a variety of opportunities
  - Data science has been dubbed the “Sexiest job of the 21st century”!

WHAT IS DATA SCIENCE AND WHAT CAN IT DO FOR ME?

THE SEXIEST JOB OF THE 21ST CENTURY

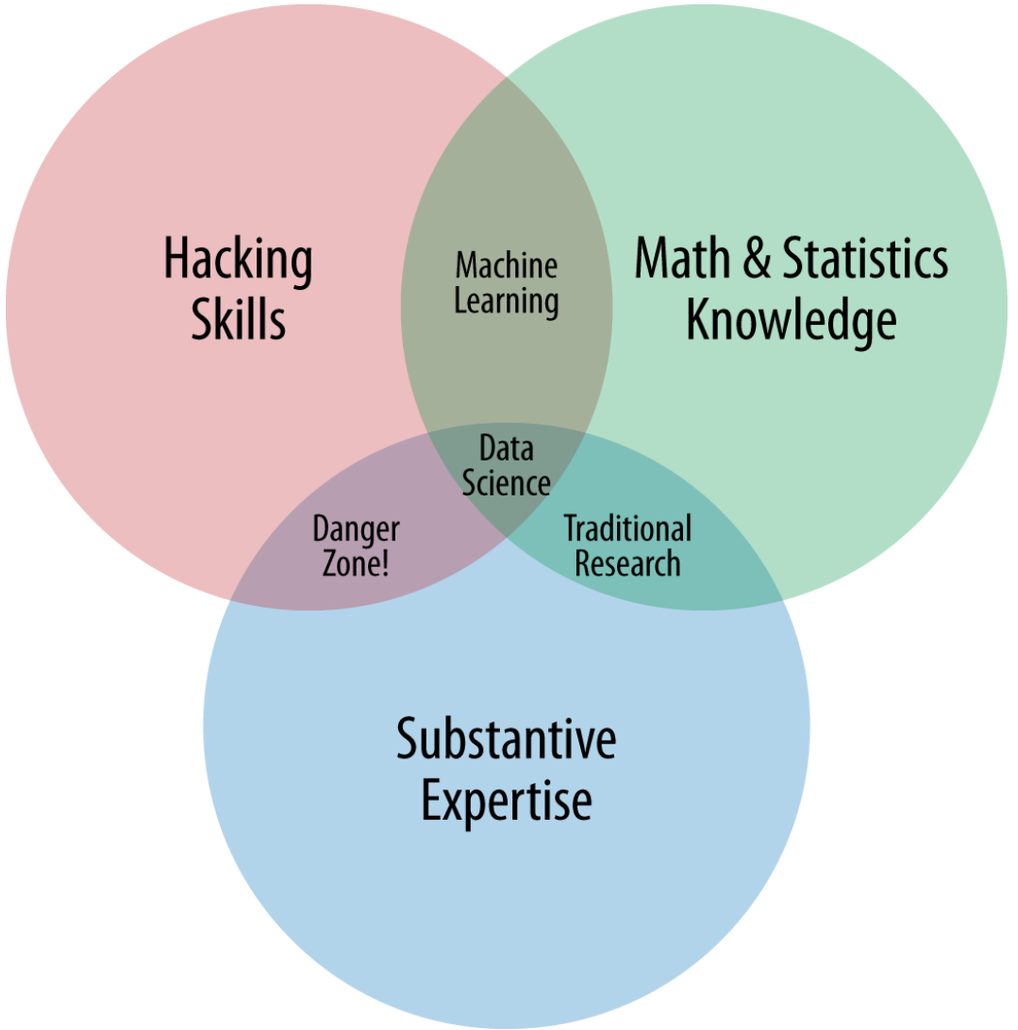
- Across industries and both large and small companies, the appetite for analytical insight is growing. Companies are increasingly looking to make sense of their data and for analytics to underpin their business decision making and deliver a tangible impact, increasing the demand for people who love data and want to solve problems with insight.

- Data Science: A set of tools and techniques used to extract useful information from data.
  - An interdisciplinary, problem-solving oriented subject.
  - The application of scientific techniques to practical problems.
  - A rapidly growing field.

Data Science

QUALITIES OF A DATA SCIENTIST

- Programming skills
- Math and Statistics knowledge
- Business acumen (substantive expertise)
- Plus: Communication skills

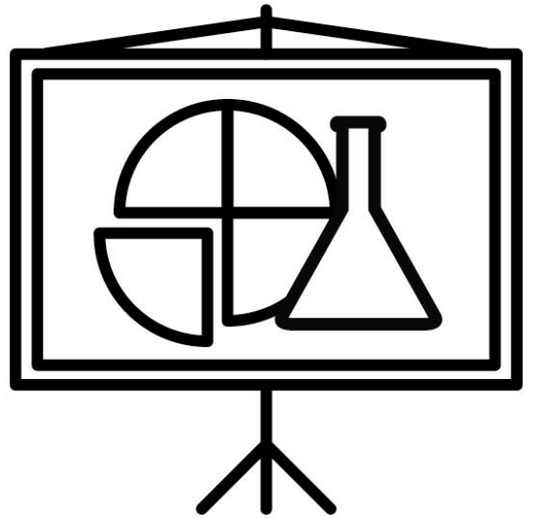


WHAT CAN DATA SCIENCE DO FOR ME?

WHAT CAN DATA SCIENCE DO FOR ME?

RESPONSIBILITIES OF A DATA SCIENTIST

- Understand the problem and ask the right questions
- Get data from various different sources and prepare it
- Using statistical analysis to understand relationships in data
- Using statistical models to determine the most likely outcomes
- Visualize data and write reports communicating results
- Create data products that deliver actionable insight



INDEPENDENT PRACTICE: SELF-ASSESSMENT

EXERCISE

DIRECTIONS

Please fill out this survey:

https://goo.gl/32zLUb

OUTCOME

Data for a learning plan discussion

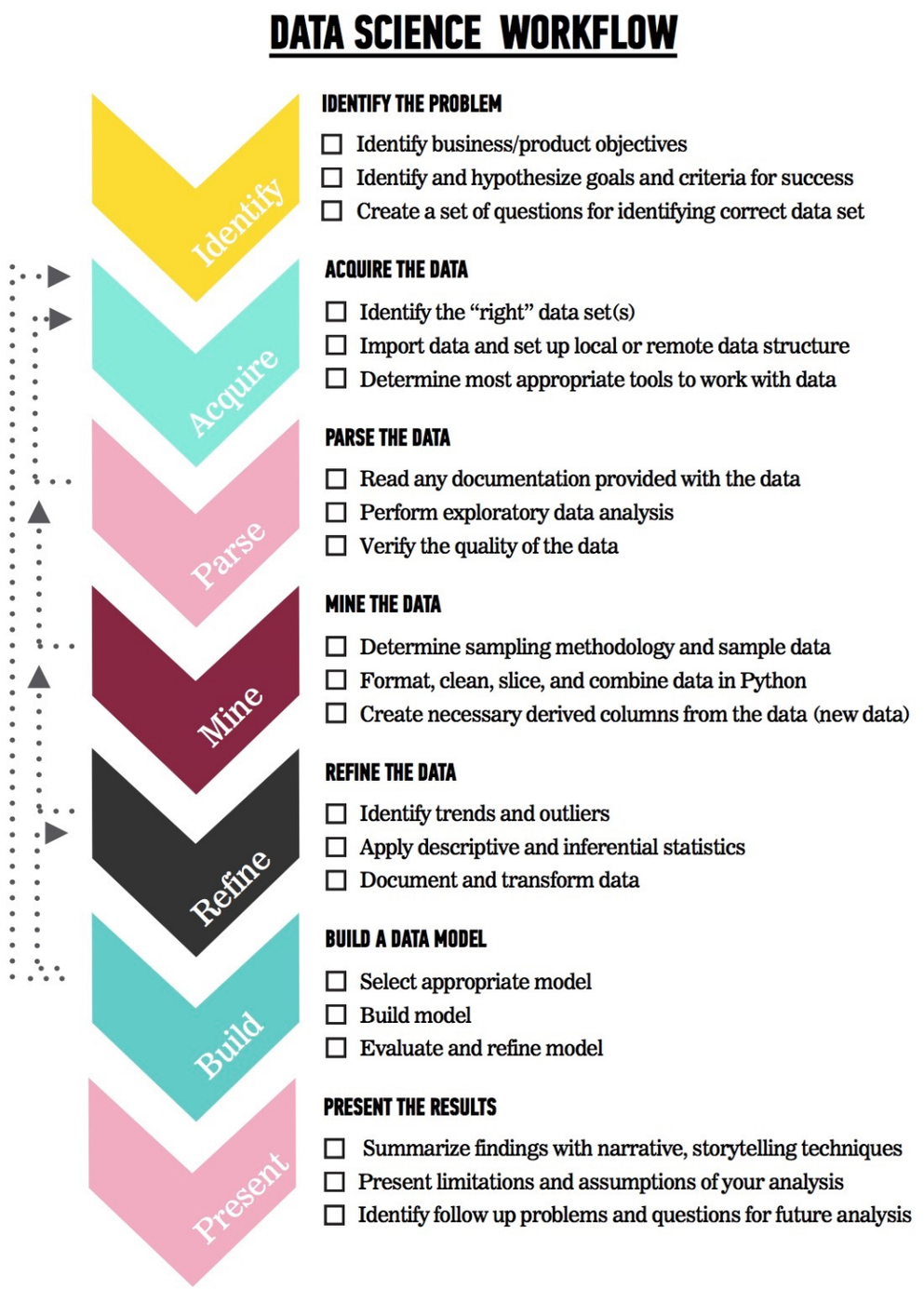
DEMO

VISUALIZING THE DATA SCIENCE WORKFLOW

THE DATA SCIENCE WORKFLOW

MAIN PHASES

- Identify the problem
- Acquire the data
- Parse the data
- Mine the data
- Refine the data
- Build a data model
- Present the results



YOUR TURN: VISUALIZING THE DATA SCIENCE WORKFLOW

EXERCISE

DIRECTIONS

You are a junior data scientist at Amazon. Your boss asks you about the leading indicators that a user will make a new online purchase. How would you go about solving this question?

1. Acquire Data: What could we do first here? What are some considerations we should make?
2. Parse Data: What do you think this means? Why is it important?
3. Mine and Refine: Is the raw data enough? What calculations/transformation do you recommend doing? How do you determine the presence of outliers?
4. Data model: What attributes would you include in the modeling stage? How do you know if the model is performing well?
5. Results: Who is your audience? What is the best way to present your results?

DELIVERABLE

Take 5 minutes to work with the person next to you and talk through answers to these questions. Then, we'll take 5 minutes to discuss this as a group.

LEADING INDICATORS THAT A USER WILL MAKE A NEW ONLINE PURCHASE

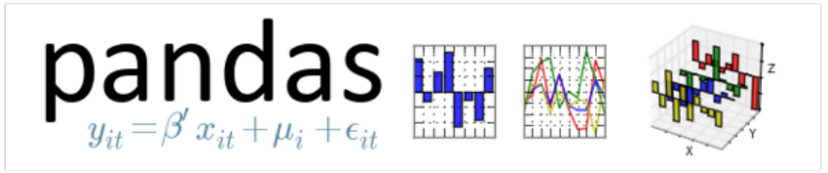
- User retention
- User actions within the product (potentially looking for external data)
- Extract aggregated values from raw data
  - How many times did a user share through Facebook within a week?
  - A month?
  - How often did they open up our emails?
- Look for patterns in data (for example distributions and correlations)
- Extract new meaning to predict if user would purchase again
- Share results using a Jupyter Notebook

GUIDED PRACTICE

EXPLORING THE DATA SCIENCE TOOLKIT

WHY PYTHON?

Python is can do more than other languages for data science, and has a great documentation and community support



PACKAGES

- Libraries of code written to solve particular set of problems
- Can be installed with: `conda install <package name>`

- Ever used Excel? How would you like working with data structured in a similar way, but without the irritation of formatting, long formula, and better graphics?
  - Try pandas!

- Does your application require the use of advanced mathematical functions or numerical operations with arrays, vectors or matrices?
  - Try SciPy (scientific Python).
  - Try NumPy (numerical Python).

INSTRUCTIONS (DROPBOX)

- We recommend using a Jupyter notebook for this practice.

The Dropbox link provided has a Zip file with the materials for the class.

1. Unzip the file downloaded in a known location in your file system
2. Locate the file called DataScience101_Part1_GuidedPractice.ipynb
3. Open Jupyter: Open a terminal
   - Mac: Using spotlight search for "Terminal"
   - Windows: Click the "Start" button and type "cmd"
   - In the terminal type: `` `jupyter notebook` ``
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the pandas library we introduced above.

INSTRUCTIONS (GITHUB)

- We recommend using a Jupyter notebook for this practice.
  To get a hold of the starter code, you'll need to download these materials.

  1. Visit this page:
     git clone https://github.com/dimagalat/data-science-101-cwe-materials.git
  2. Click on the "Clone or Download" button, and click "Download ZIP"
  3. Unzip the file downloaded in a known location in your file system
  4. Open Jupyter: Open a terminal
     - Mac: Using spotlight search for "Terminal"
     - Windows: Click the "Start" button and type "cmd"
     - In the terminal type: `` `jupyter notebook` ``
  5. Navigate to the folder where you have saved the file in step 1
  6. Open the file from the Jupyter interface
  7. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the pandas library we introduced above.

INDEPENDENT PRACTICE

ANALYZE SOME DATA!

NOW YOU TRY!

FLOWERS AND MORE

- You are a business intelligence manager at a fast moving startup that deals with flowers.

- You need to analyze some data for iris flowers of three different species.

- The business has received a sample data set with typical measures for the following three species for iris flowers…

IRIS DATA SET

- Famous data set analyzed by Ronald Fisher

- 50 samples of 3 different flower types:
  - Setosa
  - Virginica
  - Varsicolor

- 4 features:
  - Sepal: length and width
  - Petal: length and width

- Let us use Python to review some analytics that will help us differentiate these three species.

INSTRUCTIONS

- We recommend using a Jupyter notebook for this practice.

From the materials downloaded:

1. Unzip the file downloaded in a known location in your file system
2. Locate the file called DataScience101_Part1_IndPractice.ipynb
3. Open Jupyter: Open a terminal
   - Mac: Using spotlight search for "Terminal"
   - Windows: Click the "Start" button and type "cmd"
   - In the terminal type: `` `jupyter notebook` ``
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to type the commands we will cover below

- In this guided practice we are using a sample dataset, demonstrate how to carry out descriptive analytics using the pandas library we introduced above.

TITLE

BREAK

WHAT ARE ALGORITHMS, ANYWAY?

ACTIVITY: WHAT COMES TO MIND WHEN YOU HEAR THE WORD “ALGORITHM”?

DIRECTIONS

1. What do you think when you hear the word “algorithm”?
2. Can you give an example?
3. Do you use any algorithms in your every-day-life?

DELIVERABLE

Discussion with the class

A SET OF STEPS TO ACCOMPLISH A TASK

- Algorithms need to have their steps in the right order.

- When you write an algorithm, the order of the instructions is very important.

ALGORITHM

A SET OF STEPS TO ACCOMPLISH A TASK

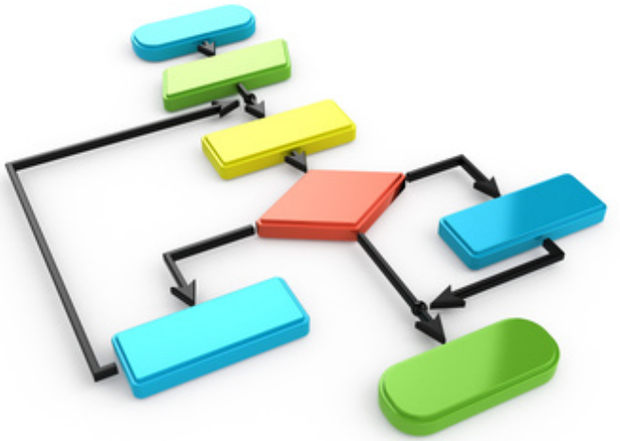
- Would you put on your shoes before you put on your socks?
- What if you put on your jacket before you put on your coat?

ALGORITHM

COMPUTER SCIENCE

- Algorithms are a formal way of describing precisely defined instructions.

- Computers *are very good* at carrying out series of precisely defined instructions.

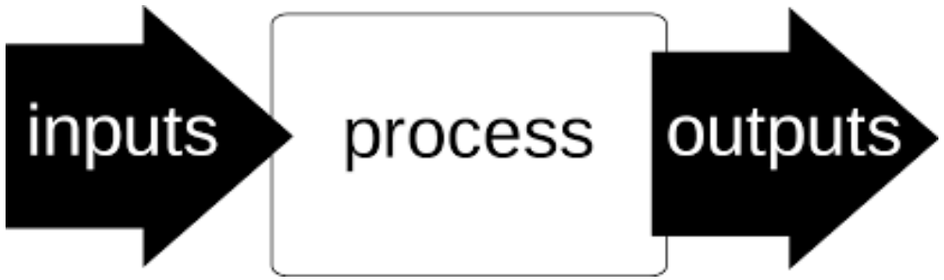


CRITERIA OF A GOOD ALGORITHM

- An algorithm is an **unambiguous** description that makes clear what has to be implemented.
  - “Bake until done” is ambiguous; “Choose a large number” is vague
- An algorithm should be guaranteed to terminate and produce a result, always stopping after a finite time. If an algorithm could potentially run forever, it wouldn’t be very useful because you might never get an answer!

CRITERIA OF A GOOD ALGORITHM

- We can condense some of this information as follows:



ACTIVITY: WHAT IS THE ALGORITHM FOR...?

EXERCISE

DIRECTIONS - PICK AN EVERY-DAY TASK (OR THINK OF YOUR OWN)

1. Making breakfast.
2. Commuting to work.
3. Making a cup of coffee.
4. Brushing teeth.

DELIVERABLE

Break down the steps into the smallest discrete, sequential items and think of the logical order in which things have to be done to achieve the task.

Discuss in your group and we will compare with the entire class afterwards

DEMO

ALGORITHMS IN ACTION

LET US SEE HOW TO WRITE AN ALGORITHM

- We will use Python to write our algorithm

Example:

- Problem: Given a list of positive numbers, return the largest number on the list.

- Inputs: A list ***L*** of positive numbers.
  The list must contain at least one number.

- Output: A number ***n***, which will be the largest number of the list.

WHAT IS THE OUTPUT

- ALGORITHM

1. Set the variable `max` to 0.
2. For each number `x` in the list `L`, compare it to `max`.
   - If `x` is larger, set `max` to `x`.
3. `max` is now set to the largest number in the list.

HERE IT IS IN PYTHON

Python

```python
def find_max(L):
    max = 0
    for x in L:
        if x > max:
            max = x
    return max
```

ACTIVITY: DISCUSSION...?

DIRECTIONS

1. Does the algorithm above meet the criteria for a good algorithm?
    1. It is unambiguous?
    2. Does it have defined inputs and outputs?
    3. Is it guaranteed to terminate?
    4. Does it produce the correct results?

DELIVERABLE

Discuss in your group and we will compare with the entire class afterwards.

ALGORITHMS IN THE CONTEXT OF MACHINE LEARNING

- Machine learning is a subset of artificial intelligence. It is concerned with the construction and study of systems that can learn from data.

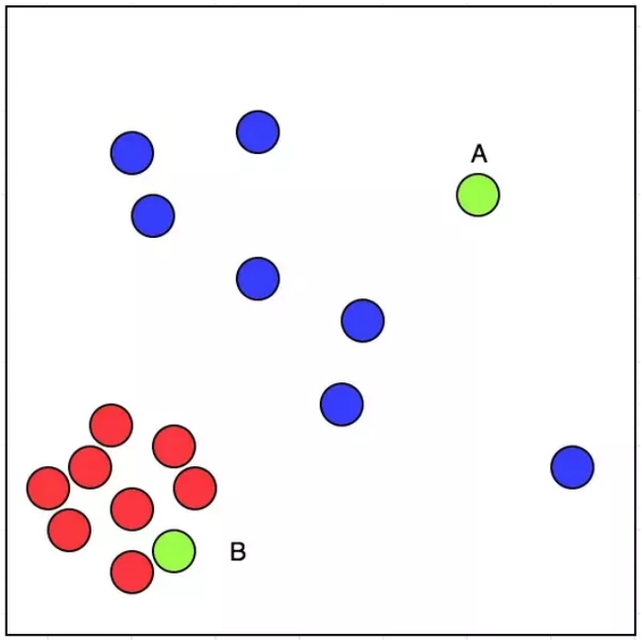
- The core of machine learning deals with representation and generalization.

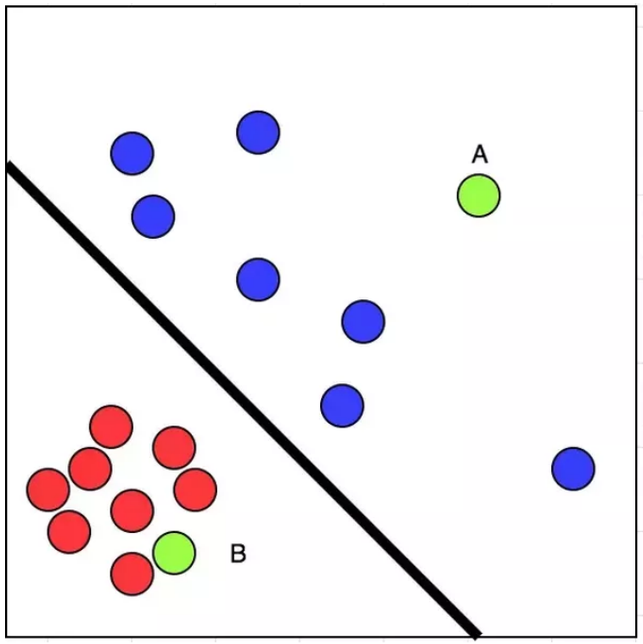
  - **Representation** – extracting structure from data
  - **Generalization** – making predictions from data

MACHINE LEARNING PROBLEMS

- **Supervised Machine Learning**: Making predictions (generalization)

- For example, suppose you want to predict whether someone will make make a purchase the week after they visit your site.





MACHINE LEARNING PROBLEMS

- **Supervised Machine Learning**:

- Most of ML in business context is supervised, because a lot of business problems focus on finding chances of an event based on the past outcomes

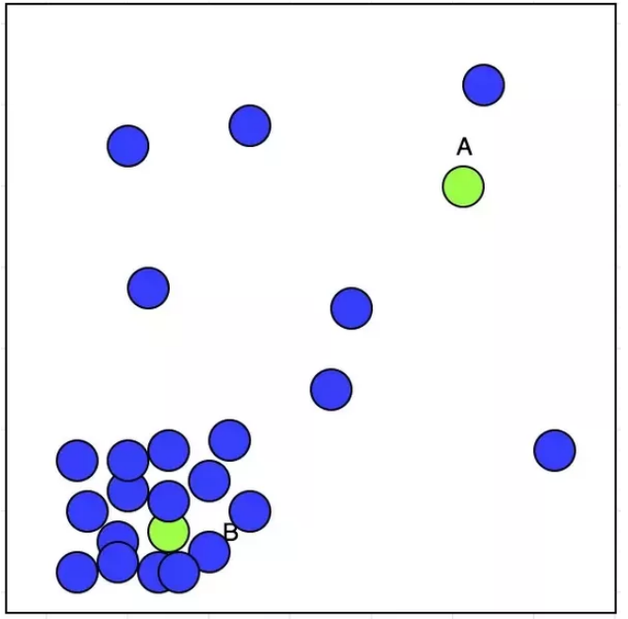
- If you know that an event will occur with a high likelihood, you can then take action and send a reminder or offer a discount. Amazon, Netflix, and others do this based on the history of their existing customers.

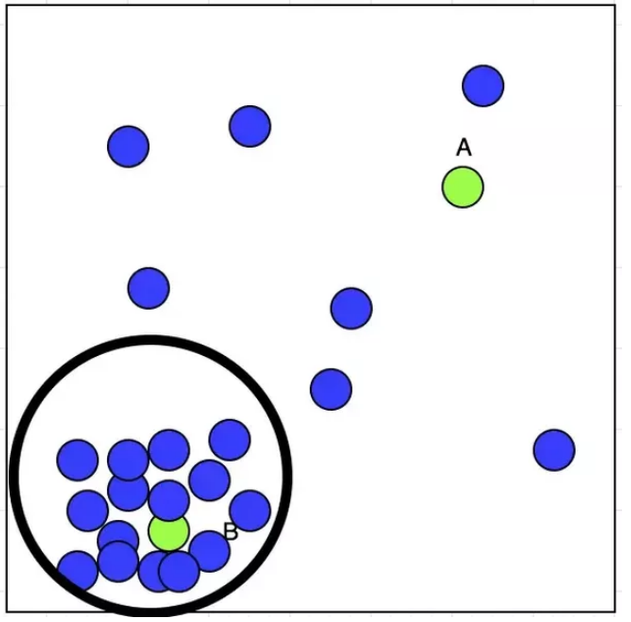
- Some examples of supervised learning algorithms include:
  - linear regression
  - decision trees
  - neural networks

MACHINE LEARNING PROBLEMS

- **Unsupervised Machine Learning**: Extracting structure (representation)

- For example, suppose you want to understand your customer base so that you can produce appropriate segments that you can target with your next marketing campaign.





MACHINE LEARNING PROBLEMS

- **Unsupervised Machine Learning**:

- Based on these attributes you can find similarities and differences that provide groupings (segments) of customers.

- You can then take action and make an offer or recommend a product specifically to these segments.

- Some unsupervised learning algorithms include:
  - clustering
  - anomaly detection
  - principal component analysis

INDEPENDENT PRACTICE

DATA SCIENCE: CASE STUDY

INSTRUCTIONS

- We recommend using a Jupyter notebook for this practice.

From the materials:

1. Unzip the file downloaded to a known location in your file system
2. Locate the file called DataScience101_Part2_DecisionTree.ipynb and the Iris dataset.
3. Open Jupyter: Open a terminal
   - Mac: Using spotlight search for "Terminal"
   - Windows: Click the "Start" button and type "cmd"
   - In the terminal type: `jupyter notebook`
4. Navigate to the folder where you have saved the file in step 1
5. Open the file from the Jupyter interface
6. Voilà, you are ready to follow this practice

- In this independent practice we are using the Iris data set to see how Python can help us construct a decision tree like the one we have discussed.

CONCLUSION

REVIEW & RECAP

- In this workshop, we've covered the following topics:
  - Why data science?
  - What can data science do for me?
  - What is the data science workflow?
  - How to analyze and visualize data using Python
  - Define the role of algorithms and their relationship with machine learning
  - Demonstrate how these concepts can be applied to make predictions

LEARNING PLAN

Evaluate your data science skills! How confident are you with:

- Programming skills (Python or R)
- Knowledgable in algebra and statistics (analyzing and modeling data)
- Business acumen (how to work with stakeholders)
- Industry expertise (for the type of field you're working within)
- Communication skills (visualize data, tell stories)

WHAT SHOULD YOU DO NEXT?

Refer back to your earlier self-assessment:

1. Which skills do you want to improve first? Which ones are you most interested in learning about?

2. Rank these and identify the top three focus areas.

3. For each focus area, identify at least *one* possible resource and a related goal.

WHAT SHOULD YOU DO NEXT?

Want to be a better programmer?

Work on these:

- Continue learning Python syntax on sites like Codecademy or Code School; Leet Code and Hacker Rank
- Already know R? Work on comparing the two.
- Interested in other frameworks? Try Spark!

WHAT SHOULD YOU DO NEXT?

Want to brush-up on your math and statistics skills?

Have a look at these:

- Data Analysis with Open Source Tools, P. K. Jannert
- Pattern Recognition and Machine Learning, C. Bishop
- Data Science and Analytics with Python, J Rogel-Salazar
- An Introduction to Statistical Learning with Applications in R (free PDF)
- Elements of Statistical Learning (free PDF)

WHAT SHOULD YOU DO NEXT?

Concerned about business acumen & communication skills?

Have a look at these:

- Data Science for Business, F. Provost and T. Fawcett
- Storytelling with Data: A Data Visualization Guide for Business Professionals, C. Nussbaumer Knaflic

WANT MORE?

General Assembly offers courses in data science!

Check out our:

- Part-time Data Science Course
- Data Science Immersive Course

ADDITIONAL RESOURCES

BOOKS

- Data Analysis with Open Source Tools, P. K. Jannert
- Data Science for Business, F. Provost and T. Fawcett
- Pattern Recognition and Machine Learning, C. Bishop
- Data Science and Analytics with Python, J Rogel-Salazar
- An Introduction to Statistical Learning with Applications in R (free PDF)
- Elements of Statistical Learning (free PDF)
- Think Stats (free PDF or HTML)
- Mining of Massive Datasets (free PDF)

MOOCS

- Andrew Ng’s Machine Learning Class on Coursera link
- MIT’s Artificial Intelligence course link
- Johns Hopkins' Data Analysis Methods link
- Cal Tech’s Learning from Data course link
- Fast.ai’s Computational Linear Algebra for coders link

AGGREGATORS

- DataTau: Like Hacker News, but for data
- MachineLearning on reddit: Very active subreddit
- Quora's Machine Learning section: Lots of interesting Q&A
- Quora's Data Science topic FAQ
- KDnuggets: Data mining news, jobs, classes and more

SOCIAL

- Hillary Mason (@hmason): Data Scientist in Residence at Accel and Scientist Emeritus at bitly.
- Dj Patil (@dpatil): VP of Product at RelateIQ.
- Jeff Hammerbacher (@hackingdata): Founder and Chief Scientist at Cloudera and Assistant Professor at the Icahn School of Medicine at Mount Sinai.
- J Rogel-Salazar (@quantum_tunnel): Data scientist at IBM and GA instructor
- Peter Skomoroch (@peteskomoroch): Equity Partner at Data Collective, former Principal Data Scientist at LinkedIn.
- Drew Conway (@drewconway): Head of Data at Project Florida

Q&A

DATA SCIENCE 101

BIT.LY/INTROSURVEYSYD