GENERAL ASSEMBLY

DATA SCIENCE 101

LEARNING OBJECTIVES

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

PRE-WORK

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 2.7.

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

OPENING

WHO AM I?

Peter Goodin

What do I do : Lead Instructor – Data Science Immersive

What have I done : Ph.D Cognitive Neuroscience – Psychoneuroimmunology of Major Depressive Disorder, Consumer Neuroscience, Condom analysis.

What I'm interested in : Tech & machine learning for rehabilitation after brain damage

@peter_goodin

petergoodin

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!
- You've installed Anaconda with Python 2.7.

THE BIG PICTURE

- What we'll cover:
  - Why data science & what it can do for me?
  - Data science skills
  - Explore the Data Science Toolkit
  - Introduction to Algorithms
  - Basic and advanced data analysis

THE BIG PICTURE

- Why data science matters:
  - OECD ranks Australia's numeracy skills in the lower 50% of countries surveyed
  - Data science (programming, exploration, analysis, interpretation) is a sought-after skill in many industries

- Why it rocks:
  - Data science opens up a door to a variety of opportunities
  - Mix of creative and analytic skills required (never boring!)
  - Learning for life

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

WHAT IS DATA SCIENCE?

THE skill set of the information age.

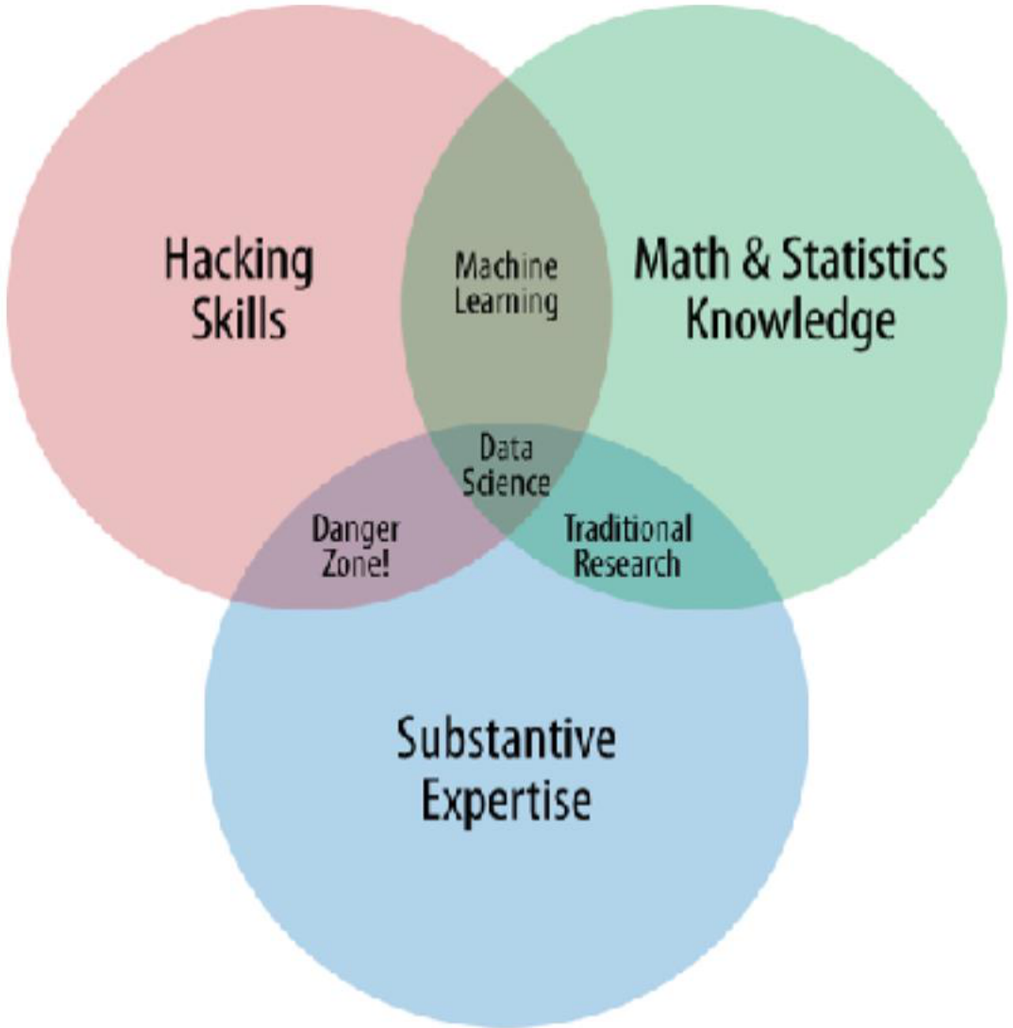
Data Science: A set of tools and techniques used to extract useful information from data.

- 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
- Creativity & problem solving
- Plus: **Communication skills**



YOUR TURN: QUALITIES OF A DATA SCIENTIST AND YOU

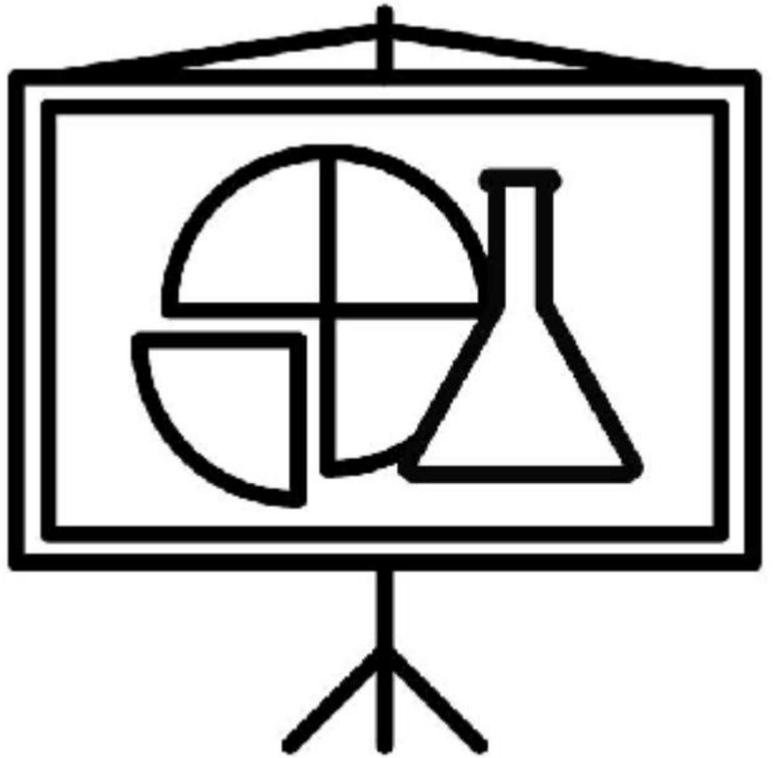
DIRECTIONS

Let's talk through the following questions together:

1. What do you think are the most important qualities for a data scientist?
2. Can you think of any other quality/skill we have not mentioned?

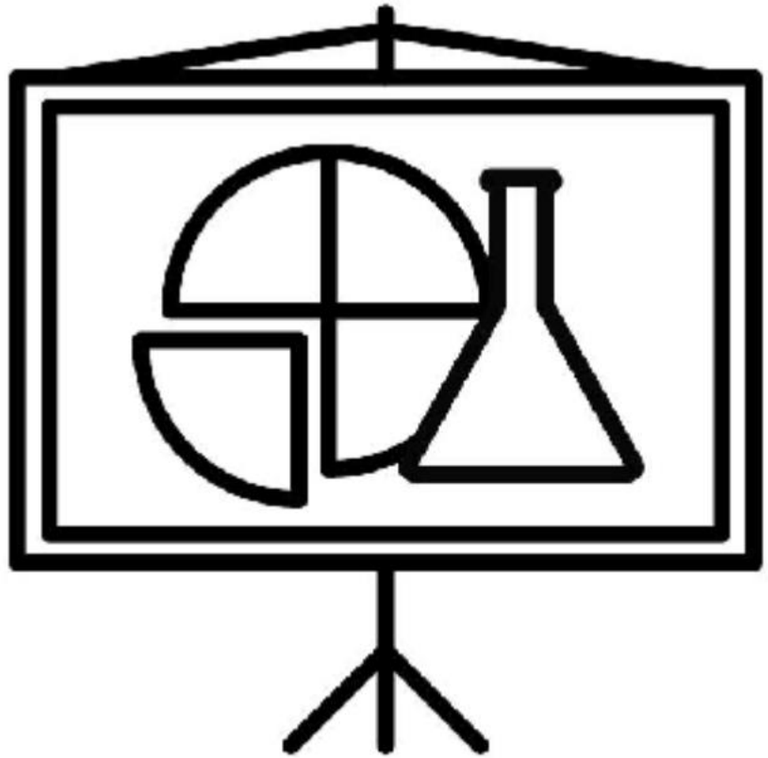
WHAT CAN DATA SCIENCE DO FOR ME?

- Ask better questions:
  - What is required?
  - How are results evaluated? (measures of success)
  - What do we currently know? (existing data)
  - What has happened? (descriptive analytics)
  - What will happen (if)? (predictive analytics)
  - What to do to achieve what we require? (insight)
- Define and test a hypothesis/run experiments.



WHAT CAN DATA SCIENCE DO FOR ME?

- Manipulate, sanitize, and wrangle data.
- Visualize data.
- Understand data relationships.
- Tell the machine how to learn from data.
- Create data products that deliver actionable insight.
- Tell relevant stories from data.



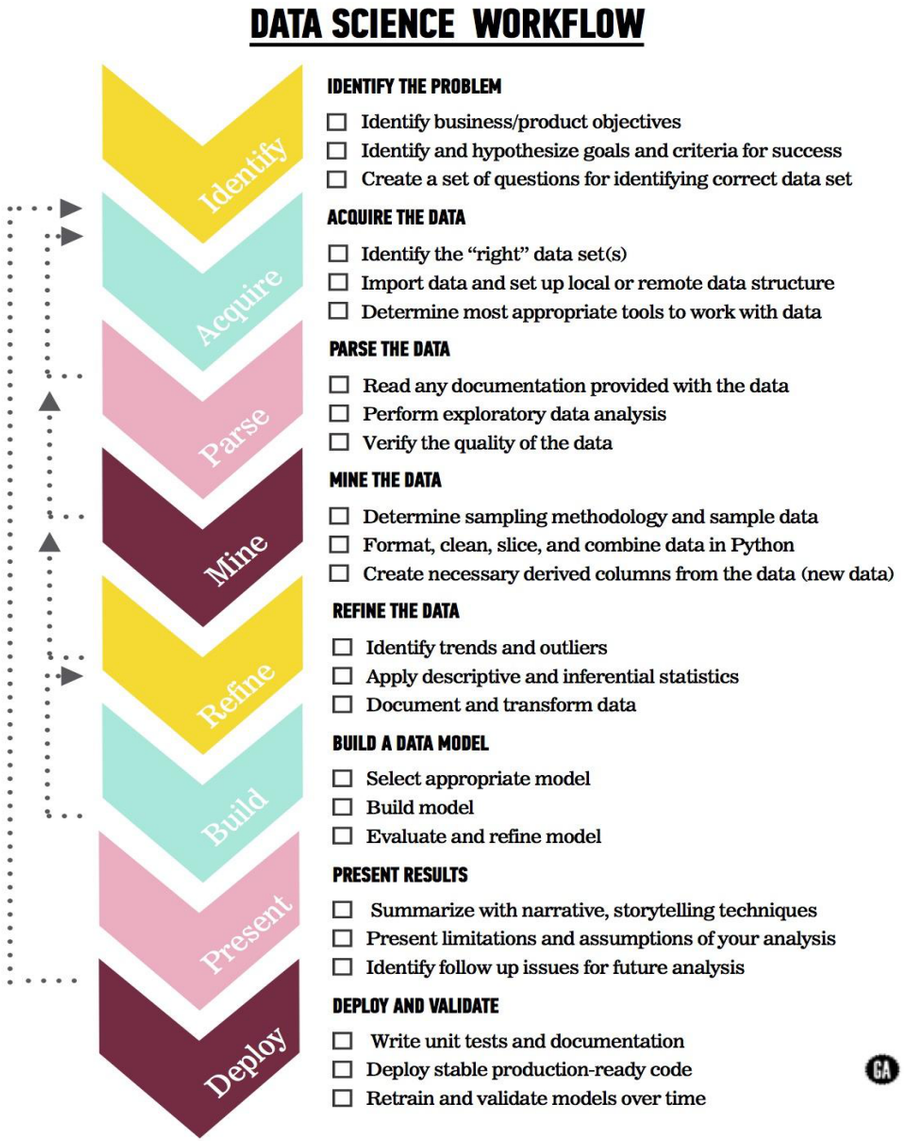
DEMO

VISUALIZING THE PASSENGER 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



GUIDED PRACTICE

EXPLORING THE DATA SCIENCE TOOLKIT

WHY PYTHON?

Python is:

- Great for rapid prototyping and full-stack commercial applications.
- A modern, elegant, object-oriented language.
- High level language, i.e., you can be more productive sooner.
- Well documented and has an established and growing community.
- Mature. In other words, Python has libraries that will help you with a lot of different tasks!

- Anaconda distribution comes with python, a large collection of libraries and extra features to help you get to the analysis sooner!

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).

PACKAGES

- Are you interested in using Python in a data science workflow and exploit the use of machine learning in your applications
  - Look no further than Scikit-learn.

- Are you tired of the boring-looking charts produced with Excel? Are you bored of looking for the right menu to move a label in your plot?
  - Take a look at the visuals offered by matplotlib & seaborn.

- Is your boss asking about significance testing and confidence intervals? Are you interested in descriptive statistics, statistical tests, plotting functions, and result statistics?
  - Well, statsmodels offers you that and more.

- All the data you require is available freely on the web but there is no download button and you need to scrape the website?
  - You can extract data from HTML using Beautiful soup.

ANALYSE SOME DATA!

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: https://github.com/petergoodin/ga_ds101
2. Click on the "Clone or Download" button, and click "Download ZIP"
3. Unzip the files 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 files in step 3
6. Open the file **iris_identify.ipynb**
7. Let's take a look at how data can help inform decisions

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

SCENARIO

FLOWERS AND MORE

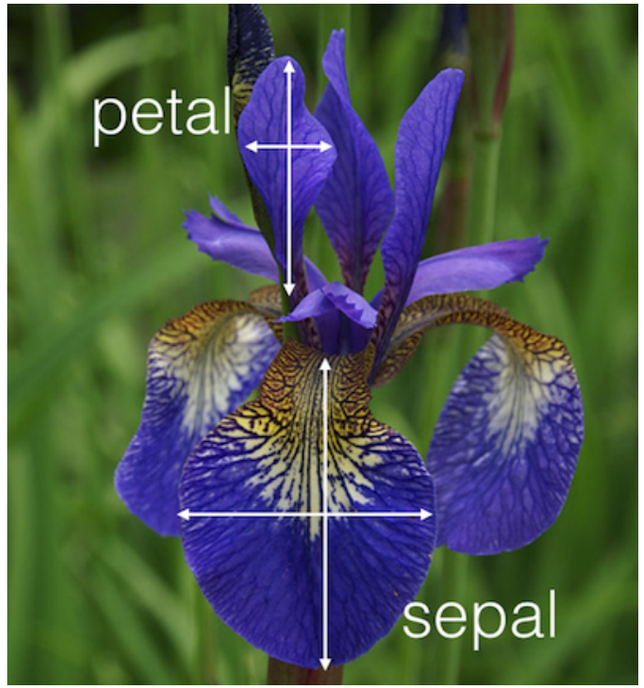
- You are a business intelligence manager at a fast moving startup that deals with flowers. Iris Mania is sweeping the world and certain species fetch upwards of $50,000 AU for a single flower!
- A new iris has just been delivered. It's species is not known and the resident florist is on holidays (typical).
- The business has a sample data set with typical measures for three species for iris flower.

IRIS DATA SET

- Famous data set collected and analysed by Ronald Fisher.

- 50 samples of 3 different flower species:
  - Setosa (the valuable one)
  - Virginica
  - Varsicolor

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



How could we use our existing data to identify it?

DATA SCIENCE 101

ANALYSIS TAKE AWAY

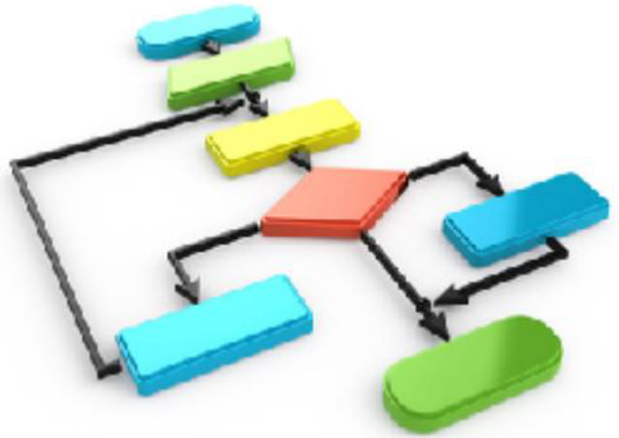
WHAT ARE ALGORITHMS ANYWAY?

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.



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 jumper?

CRITERIA OF A GOOD ALGORITHM

- An algorithm is an unambiguous description that makes clear what has to be implemented.
  - In a recipe, a step such as “Bake until done” is ambiguous because it doesn’t explain what “done” means.
  - In a computational algorithm, a step such as “Choose a large number” is vague: what does “large” mean to a computer?

CRITERIA OF A GOOD ALGORITHM

- An algorithm expects a defined set of inputs.
  - For example, it might require two numbers where both numbers are greater than zero. Or it might require a word, or a list of zero or more numbers.
- An algorithm produces a defined set of outputs.
  - It might output the larger of the two numbers, an all-uppercase version of a word, or a sorted version of the list of numbers.

CRITERIA OF A GOOD ALGORITHM

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!

DEMO

ALGORITHMS IN ACTION

LET US SEE HOW TO WRITE AN ALGORITHM

- We will use Python to write our algorithm

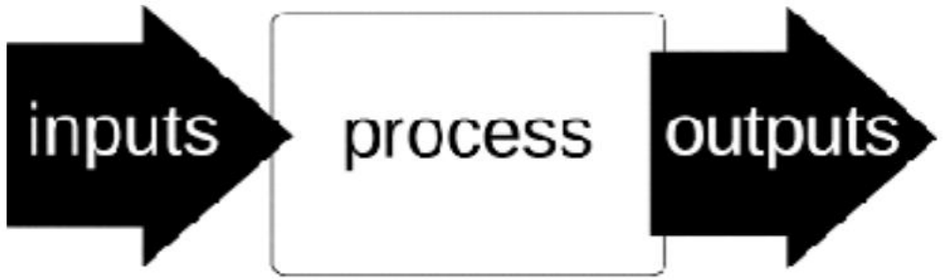
Example:

- 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.

WHAT IS THE OUTPUT

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

ALGORITHM



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
```

ALGORITHMS IN THE CONTEXT OF MACHINE LEARNING

- Machine learning is concerned with the construction and study of systems that can learn from data.
- The core of machine learning deals with supervised and and non-supervised problems.
- Supervised – Learn from labelled data (dog, cat, yes, no temperature Tuesday given previous examples of temperature)
- Non-supervised – Learn without labels (some internal criteria)

SUPERVISED METHODS

- Further distinction in supervised methods:

  - Classification:
    Sort unknown data into classes (apple / orange, sale / no sale, healthy / sick) based on previously seen data

  - Prediction:
    Predict an unknown value from data based on previously seen data (temperature, price, weight)

UNSUPERVISED METHODS

- Finding patterns in the data without previous examples
  - Clustering like with like (red things with red things, animals with animals,

WAIT! ISN’T THAT LIKE CLASSIFICATION?!

- Nope! Classification uses previous data to base the decision.

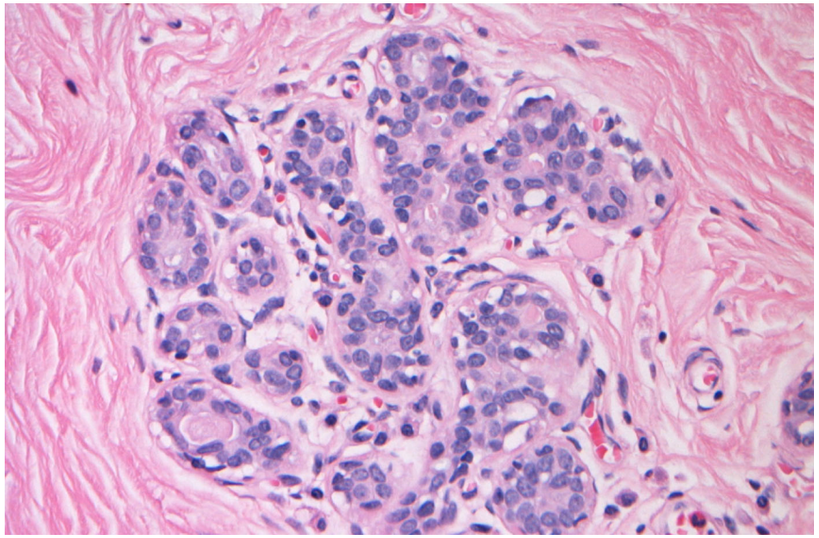
Cluster makes it up as it goes along.

ANALYSE SOME MORE DATA!

SCENARIO

CANCER CLASSIFICATION

- Using measures from biopsied breast tissue, is it possible to classify non-cancerous (benign) and cancerous (metastatic) cases?



Wisconsin Diagnostic Breast Cancer

- 596 samples and two categories:
  - Benign (357)
  - Metastatic (212)

- 30 features (Derived from 10 initial measurements)

- How accurately can non-cancerous and cancerous cases be identified?

- Does one algorithm perform better than another?

INSTRUCTIONS

From the materials:

1. Open a terminal and reopen `jupyter notebook`
2. Navigate to the folder downloaded from https://github.com/petergoodin/ga_ds101
3. Open the file cancer_classify.ipynb
4. Lets see how algorithms can be used to help save lives.

- In this final guided practise we will apply basic transformations and two different algorithms to determine whether cancerous tissue has spread or not.

DATA SCIENCE 101

ANALYSIS TAKE AWAY

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)
- Knowledgeable 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?

Want to be a better programmer?

Work on these:

- Continue learning Python syntax on sites like Codecademy or Code School.
- 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:

- 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)

Mining of Massive Datasets (free PDF)

Think Stats (free PDF or HTML)

MOOCS

- Andrew Ng’s Machine Learning Class on Coursera
- MIT’s Artificial Intelligence course
- Johns Hopkins' Data Analysis Methods
- Cal Tech’s Learning from Data course

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

DATA SCIENCE 101

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

EXIT TICKETS

DATA SCIENCE 101

DON’T FORGET TO FILL OUT YOUR EXIT TICKET

DATA SCIENCE 101

THANKS FOR COMING!