

An aerial photograph of a city skyline, likely New York City, taken from a high vantage point. The foreground is filled with numerous skyscrapers and buildings, their silhouettes and details softened by the distance. In the middle ground, the city extends towards a body of water, with the distinctive tower of the Empire State Building visible. The background shows a vast expanse of water and distant land under a dramatic, cloudy sky. The sky is filled with dark, heavy clouds, but a thin layer of lighter clouds near the horizon catches the low light of the setting or rising sun, creating a warm, golden glow that illuminates the city and the water. The overall mood is atmospheric and somewhat somber due to the dark clouds, yet hopeful due to the light on the horizon.

# Data Science

**Matthew Morris**

---

# UNIT 1 ORIENTATION, DEV ENVIRONMENT, PYTHON REVIEW, DATA SOURCE PRESENTATION

---

# Orientation

Introduction

Course Overview

Introduction to Data Science

Types of Data

*Matthew Morris*

Git: Morrisdata

MatthewMorris.DA@gmail.com



# Introduction

Name

Background in Data or reason for taking class

Something about you ie Hobby, fun fact,  
current projects, Favorite movies or books





# Course Overview



20 session grouped by 4 units, 3-unit demonstration projects, 1 final project.

## **UNIT 1 PYTHON FOUNDATIONS**

- Unit presentation: Python Technical Code Challenges

## **UNIT 2 EXPLORATORY DATA ANALYSIS AND HYPOTHESIS**

- EDA + Chipotle data
- Final presentation: Proposal + Dataset

## **UNIT 3 MODELS(Linear regression, KNN/ Classification, Logistic regression)**

- Final presentation: Initial EDA Brief
- Linear Regression and KNN Practice(Optional)

## **UNIT 4 TIME SERIES, NLP, MODEL EVALUATIONS**

- Final presentation Technical Report
- Final presentation Executive Report

# Course Outline

## UNIT 1 PYTHON FOUNDATIONS

- Welcome to Data Science
- Config Dev Environment
- Python Foundations
- Lab and Presentation

## UNIT 2 EDA AND HYPOTHESIS

- EDA in Pandas
- Working with data: APIs
- Data Visualization
- Statistics in Python
- Experiments and Hypothesis
- Lab and Presentation



## UNIT 3 MODELS(Linear Regression, KNN/ Classification, Logistic regression)

- Linear Regression
- Train-Test Split & Bias Variance
- KNN / Classification
- Logistic Regression
- Lab and Workshop

## UNIT 4 WRANGLING, TIME SERIES, & NLP

- Intro to Time Series
- Intro to Natural Language Processing
- Flex more models, SQL Data management, Data cleaning, more evals
- Review + Lab
- Presentations

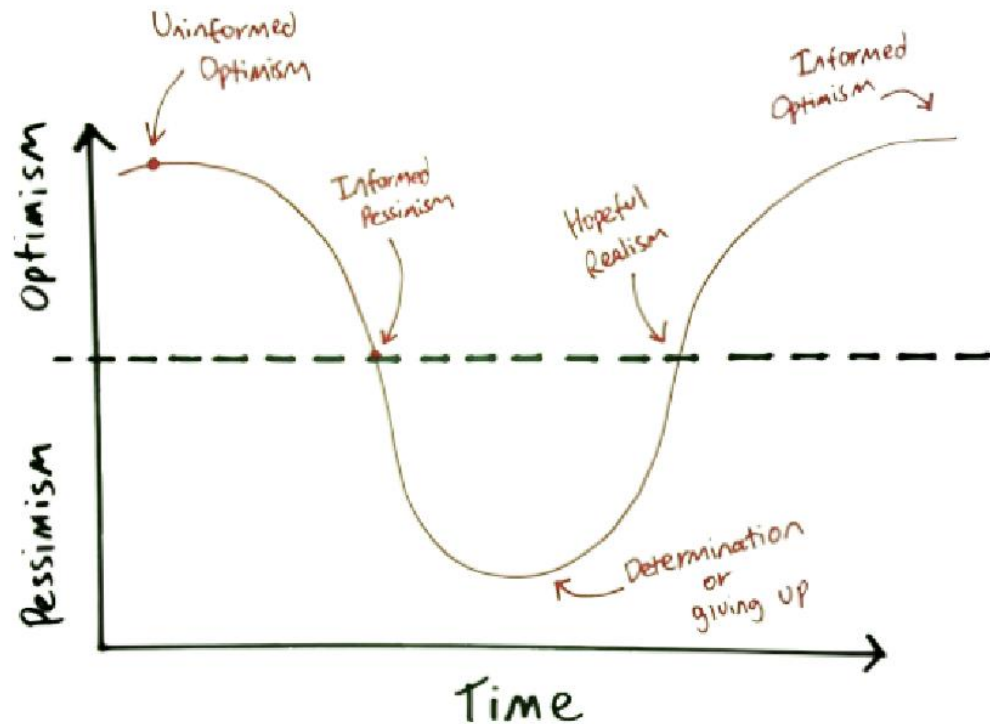
# Learning

10% skill

90% Never give up, Never surrender

Strategies for Class:

- Progress rating
- Road Map/Cheat sheets
- Mentor and Present



# Tips for success

**Study partners**, study groups increase retention, accountability, support, learning perspectives, networking, and can lead to friendships or fun.

**Practice skills on easy data sets:** Start small, find a data set that is easy and/or something that you like or fascinates you. Get the skills down and focus on harder data problems later.

**Teach others** a topic that is hard for you, a classmate, a partner, or a friend.

**Free time Fake out.** When you don't feel like doing a task, put a show on that you have seen 100 times, open your laptop and just start on the task, you will find it can turn into a few hours because well you have already seen that episode and you are not really working because you are watching TV.

**Set specific small goals** with quick deadlines. It is easier to complete several 1hr tasks then one 8-hour task.

**Procrastinate with 2 or 3 other tasks.** Rotate when you procrastinate.

**Be a learn it all** instead of a know it all. There is always something new to learn even with something you feel you have mastered.



# Tips for success

“I dedicate 1 to 3 hours a day, determined to master the material.”



# Tips for success



1) Out of the units what is most interesting to you and what seems most challenging?

2) Which Study tips seem realistic to apply? Do you have additional study tips?

---

# Introduction to Data Science

What is Data Science and Why do we care?

What are the tools of a Data Scientist?

What is a Data Scientist workflow look like?

# What is Data Science

BI or Business Intelligence

Business Analyst

Data Analyst

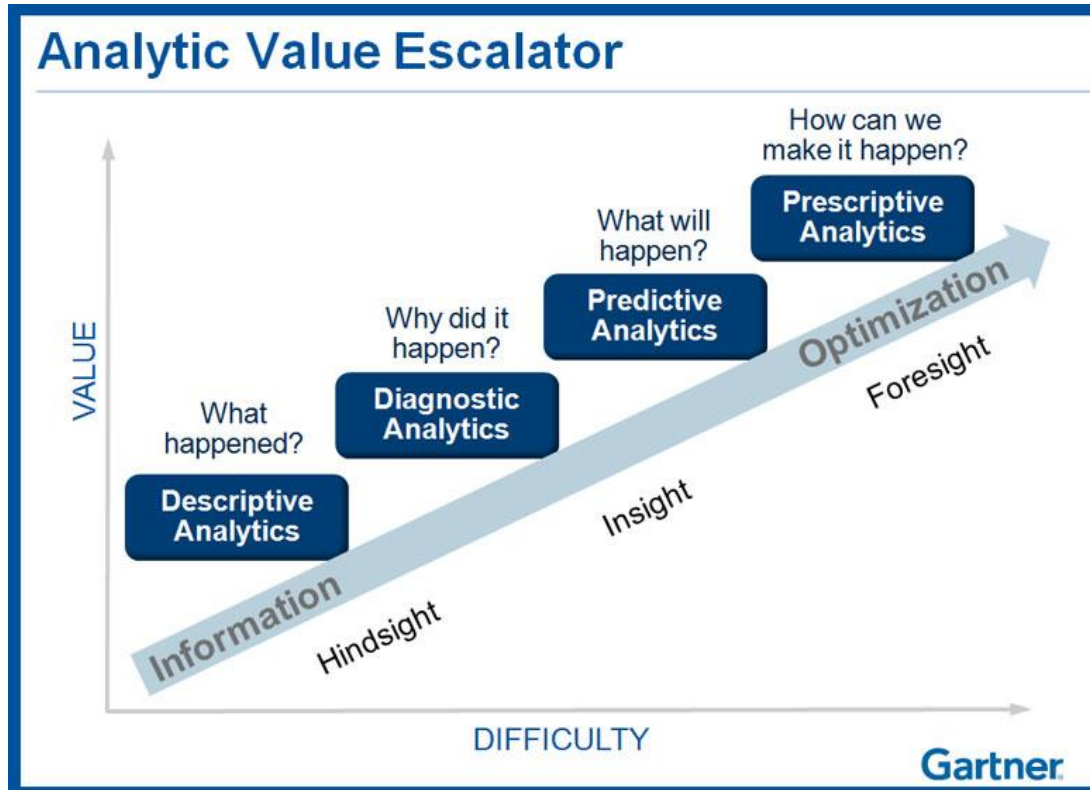
Data Scientist

Data Engineer

A good article on the difference between BI and Data Science

<https://www.itproportal.com/2016/08/18/10-differences-between-data-science-and-business-intelligence/>

# What is Data Science





# What is Data Science

“Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways.”

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data “in production”**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

# What is Data Science



**Josh Wills**

@josh\_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite More

RETWEETS FAVORITES

907

418



12:55 PM - 3 May 2012



**Javier Nogales**

@fjnogales



Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET

1

FAVORITES

5



9:08 AM - 27 Jan 2014

# What is Data Science



**Josh Wills**

@josh\_wills



Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite More

RETWEETS FAVORITES

907

418



12:55 PM - 3 May 2012



**Javier Nogales**

@fjnogales



Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET

1

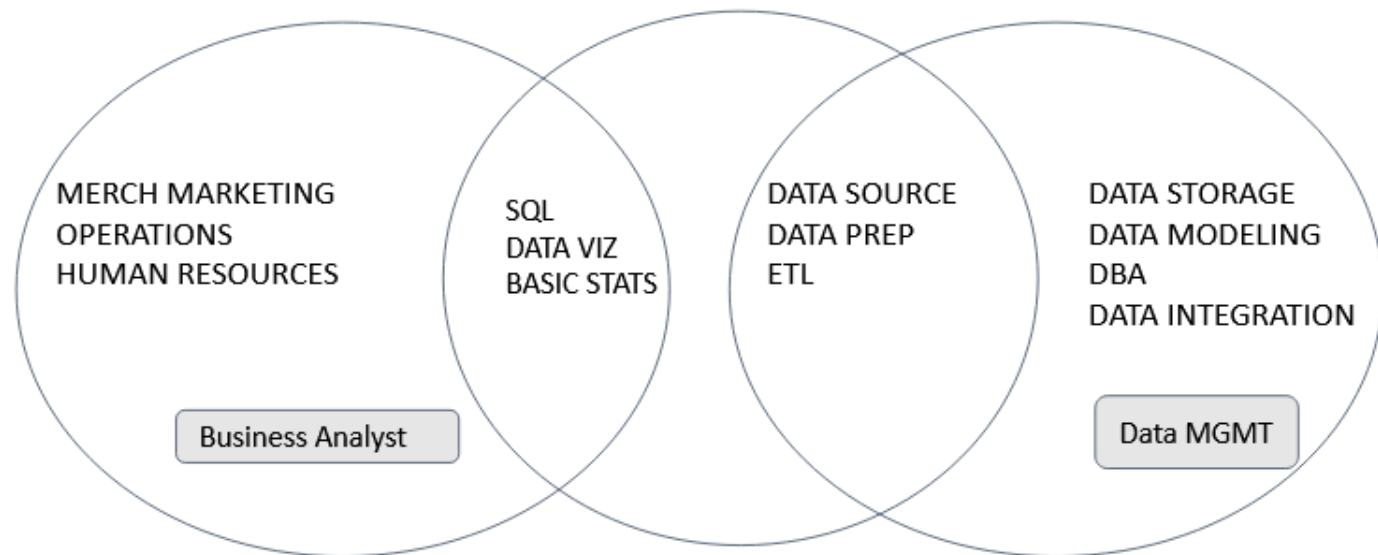
FAVORITES

5

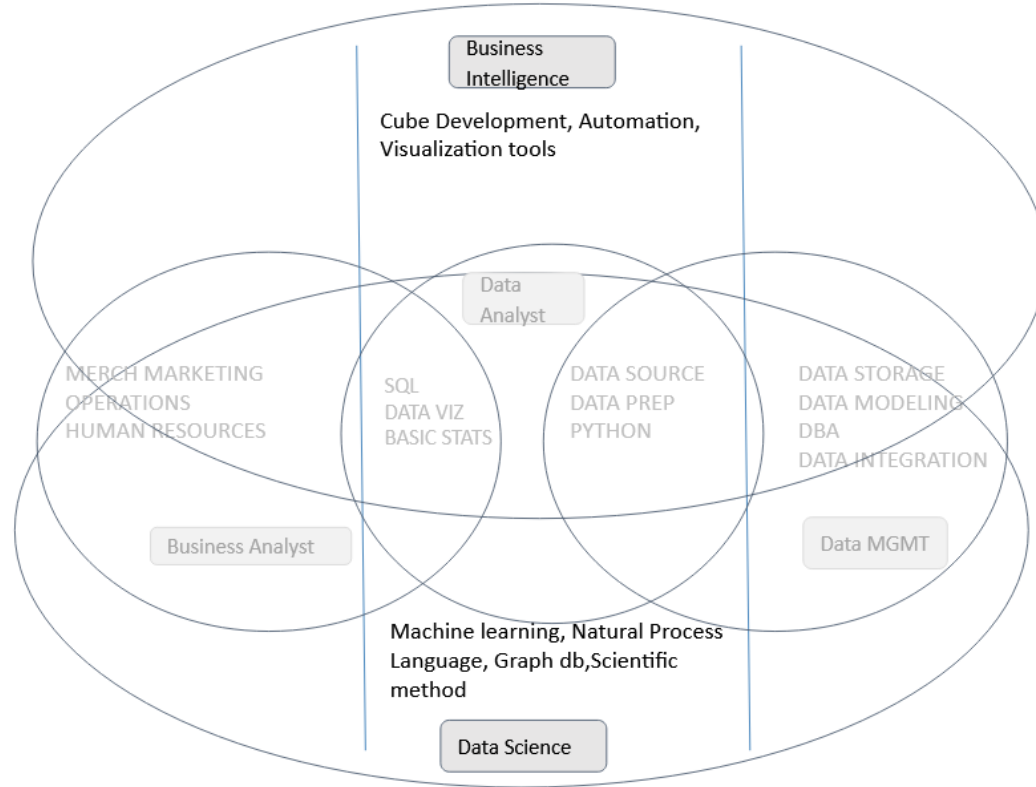


9:08 AM - 27 Jan 2014

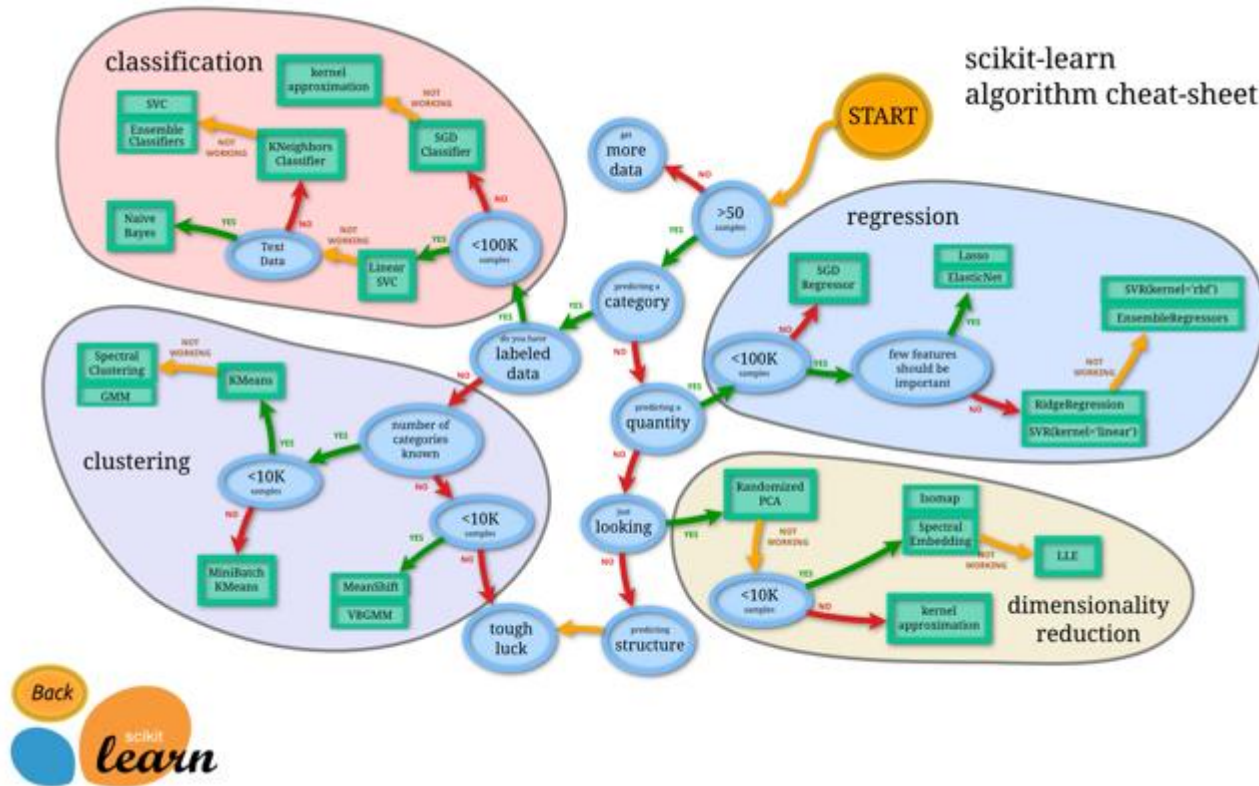
# What is Data Science



# What is Data Science

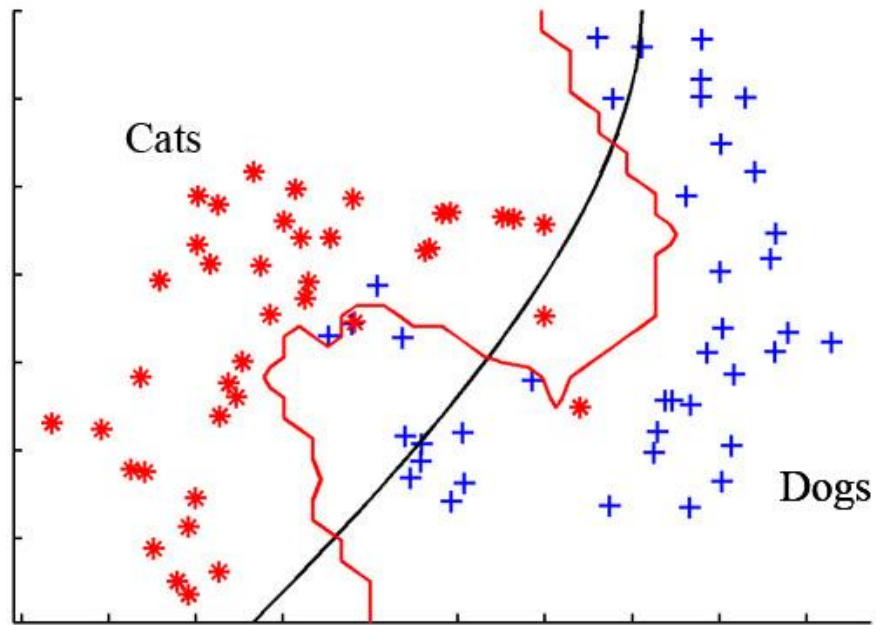


# What is Data Science

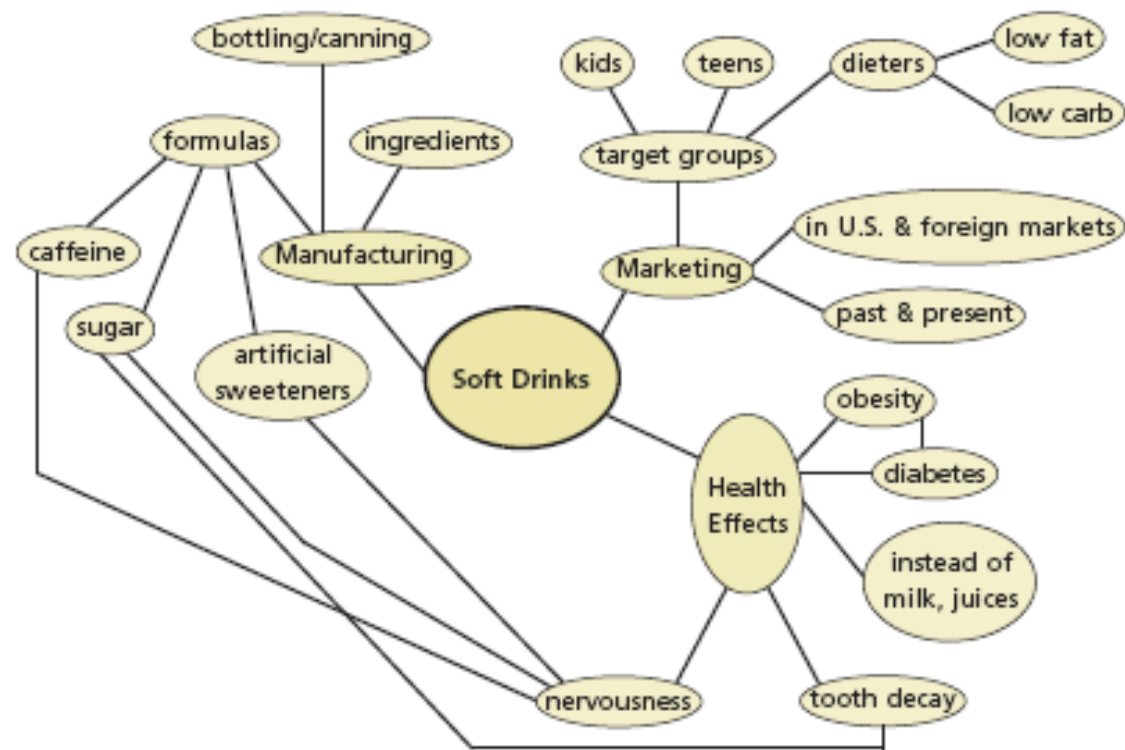




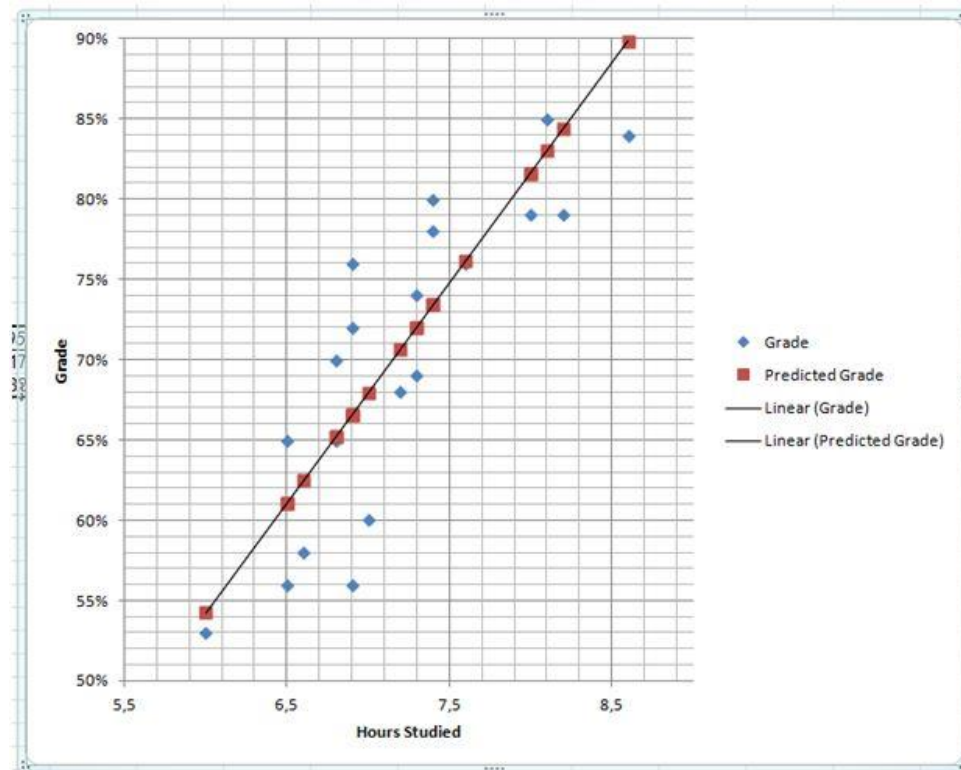
# What is Data Science



# What is Data Science



# What is Data Science



# What is Data Science



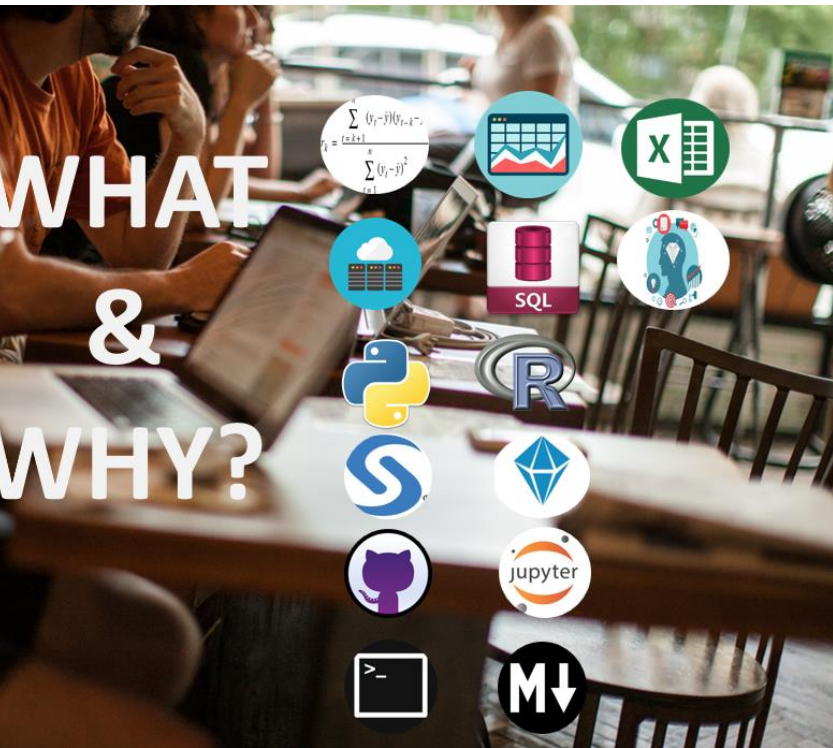
# What is Data Science

## EXAMPLES OF DATA SCIENCE IN ACTION

- Facebook facial recognition in photos
- Netflix/Amazon/Spotify recommendations
- Siri/Echo/Cortana voice recognition assistants
- Building art with Neural Networks - <https://github.com/jcjohnson/neural-style>
- Faceswap - <https://www.youtube.com/watch?v=UngUWA43q5o>
- Stock Market - <https://www.quantopian.com/> - building crowd source hedge funds
- Helping people  
<https://www.drivendata.org/> - who is a good bet to give money to  
<http://www.datakind.org/projects>
- Help find missing children - <http://www.datakind.org/projects/finding-30000-missing-children>
- Find correlations from sickness, grades, and attendance and try to find ways to improve them [http://coolculture.org/webfm\\_send/62](http://coolculture.org/webfm_send/62)
- Additional examples <https://www.kaggle.com/wiki/DataScienceUseCases>



# Data Science Toolkit





# Data Science Toolkit



Data storage, Understanding schemas, tables, fields, relational and non relational databases is a foundation of data analytics

ORACLE®



Amazon  
Redshift



Azure  
Synapse  
Analytics



# Data Science Toolkit



Business knowledge can include understanding: Knowing KPI's, Gather requirements, MetaData, Operational reports, Business acumen, communication and navigating politics and personalities of your business culture



Having a strong understanding of Lookup functions, string and numeric functions is necessary to understand the business and how the currently tackle problems.

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_t - k - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2}$$

Basic Statistics(Central Tendency) Understanding concepts is fine. Understanding long hand even better.



This is a must to understand the basic charts and graphs and be able to tell a story with them.



Structured Query Language: Unless someone is getting all of your data for you and cleaning it all for you, you will want to be proficient in SQL up to Advanced levels.

# Data Science Toolkit



Python is a general-purpose programming language. Allows you to give directions to a computer to tell it what to do.



R is a system for statistical computation and graphics.



SAS **SAS** (Statistical Analysis System) is a software suite developed by **SAS** Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.



SPSS Modeler IBM **SPSS Modeler** is a data mining and text analytics software application from IBM. It is used to build predictive models and conduct other analytic tasks

# Data Science Toolkit



Jupyter Notebooks allows you to create and share documents that contain live code, equations, visualizations and explanatory text.



Plain text formatter that converts for use in html, used to create documentation within Jupyter

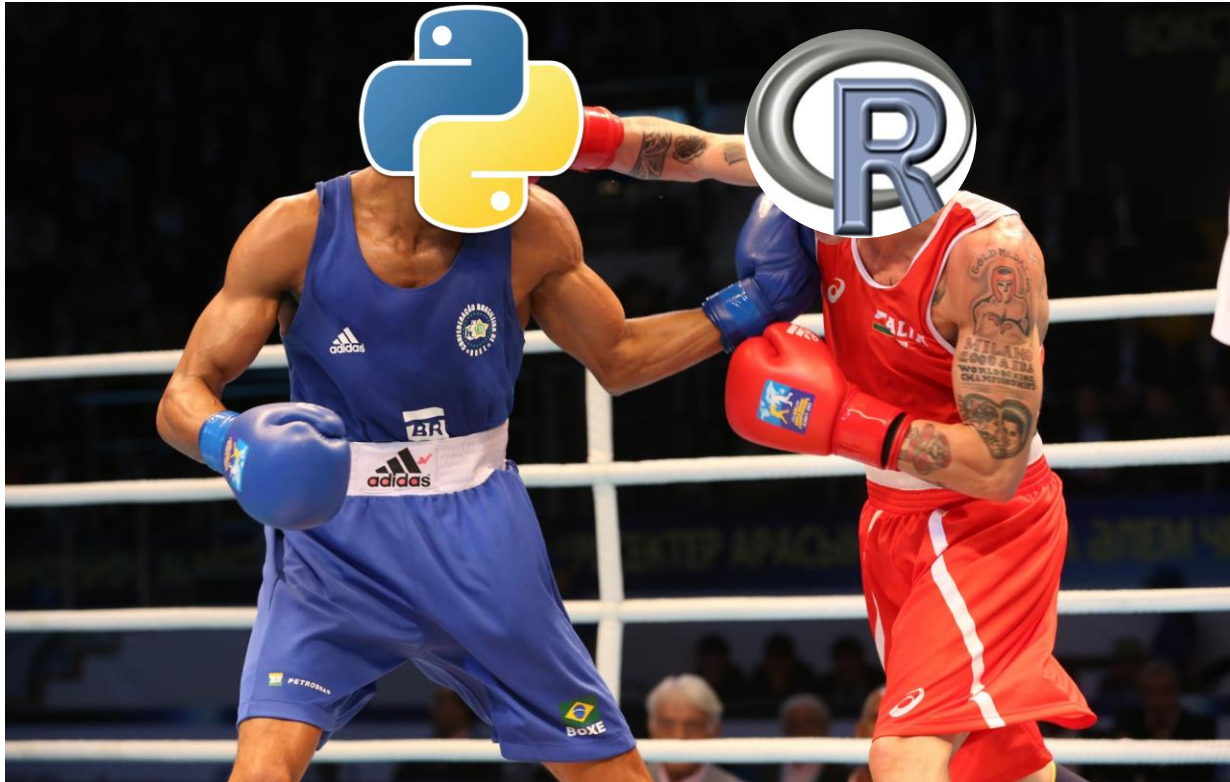


Purpose of git is to manage a project, or a set of files as they change over time. It allows for version control and collaboration.



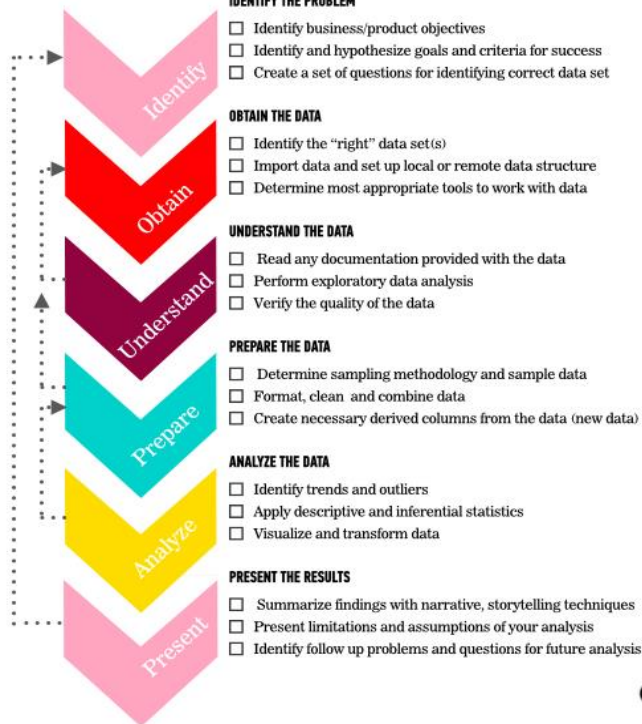
Command line is a user interface to a computers operating system. It allows you to navigate, manipulate and analyze files, data and more.

# Data Science Toolkit



# Data Science Workflow

## ANALYTICS WORKFLOW





# Data Science Workflow



- 1) What is the difference between a Data Analyst, Data Engineer, and a Data Scientist?
- 2) In your own words how would you describe the 4 primary algorithm groups in scikit learn?
- 3) What is the difference between R and Python? Which should you learn?
- 4) What is the difference between an Analytics workflow and a Data Science workflow?
- 5) When would you use Jupyter?

## Classroom Exercise

Create a Github account <https://github.com/>

Ensure you can view the home page for class  
<https://github.com/Morrisdata/DS>

Ensure you can view the current chapter  
[https://github.com/Morrisdata/DS/tree/master/01\\_Welcome\\_To\\_Data\\_Science](https://github.com/Morrisdata/DS/tree/master/01_Welcome_To_Data_Science)

Download or view in Github the current slide deck  
[https://github.com/Morrisdata/DS/blob/master/01\\_Welcome\\_To\\_Data\\_Science/GA%20Data%20Science%202024%20-%2001%20Welcome%20To%20Data%20Science.pdf](https://github.com/Morrisdata/DS/blob/master/01_Welcome_To_Data_Science/GA%20Data%20Science%202024%20-%2001%20Welcome%20To%20Data%20Science.pdf)

# Types of Data

Quantitative

Discrete and Continuous

Qualitative

Nominal and Ordinal

----Types of Data Formats----

Flat File

Text Documents

Time Series

Transactional Data

Relational Data

Spatio-Temporal Data

Image Data

Network Data

# Types of Data

Text Files – data conveyed in written or printed form. Surveys, Social Media Posts, Books, messages in chat. Can be used for word clouds sentiment analysis, give insights, however, may contain errors, bias

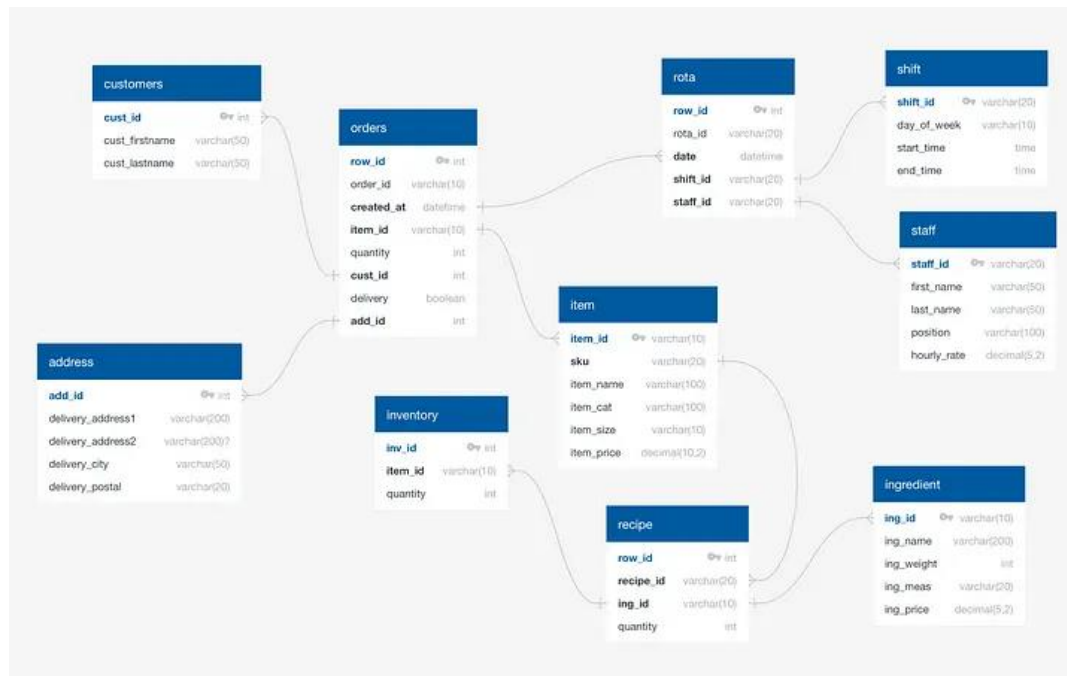
Flat File – rows and columns of data typically seen in .csv files. Single files that are not modeled or related.

Transactional Data– Date stamped events (logs): can be represented as time series.  
Customer bought a coffee at 12PM

Time Series – Similar to Transactional data, however, can be used to as successive measurements over time rather than over transaction. Number of coffees sold between 12 and 1pm or number of customers served between 12 and 1PM.

# Types of Data

Relational Data- stores tables that relate to on another example would be a customer table that relates to an item table that relates to a transaction table.



# Types of Data

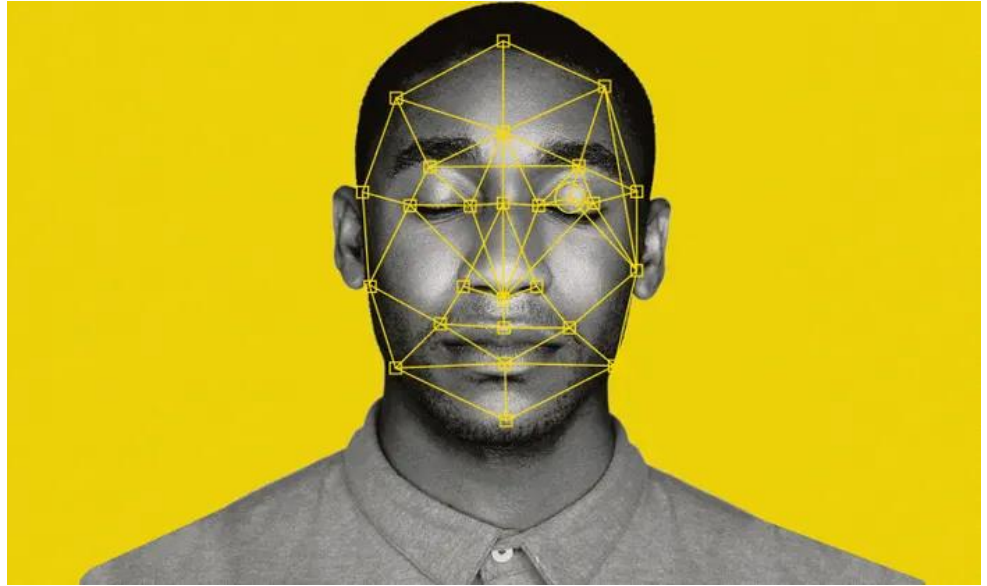
Spatio-Temporal Data – information about space and time. Examples include Weather patterns, predicting earthquakes, determining global warming, evaluation of traffic for city planning and more.





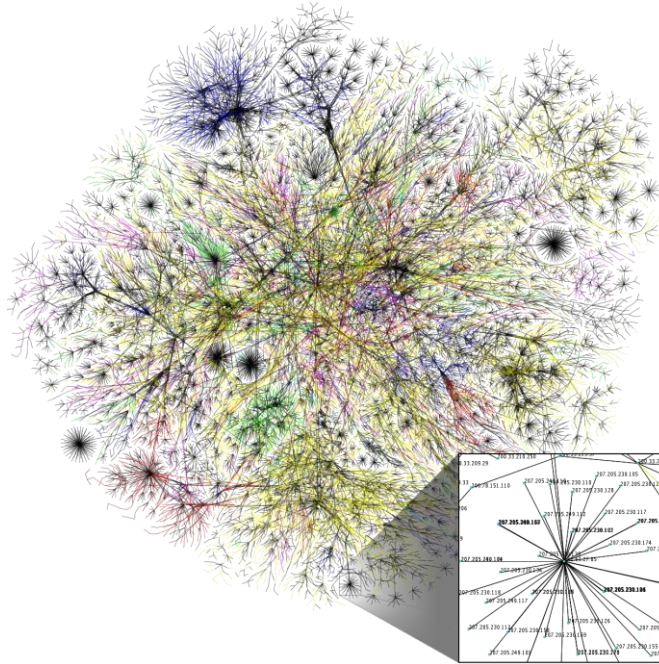
# Types of Data

Image Data: photographic or trace objects that represent the underlying pixal data of an area image element. Examples include face recognition, label classifications(dog vs cat), autonomous driving, Quality Control in manufacturing, target identification, threat assessment, Healthcare,



# Types of Data

Network Data : Data extracted from connected network devices and is limited to anonymized personal data.



# Types of Data

**Quantitative[Discrete, Continuous Data]:** Expressed by a number and measured by numerical variables, How many, how much, how often.

Discrete Data – Number of employees at a company, Number of movies on streaming platform

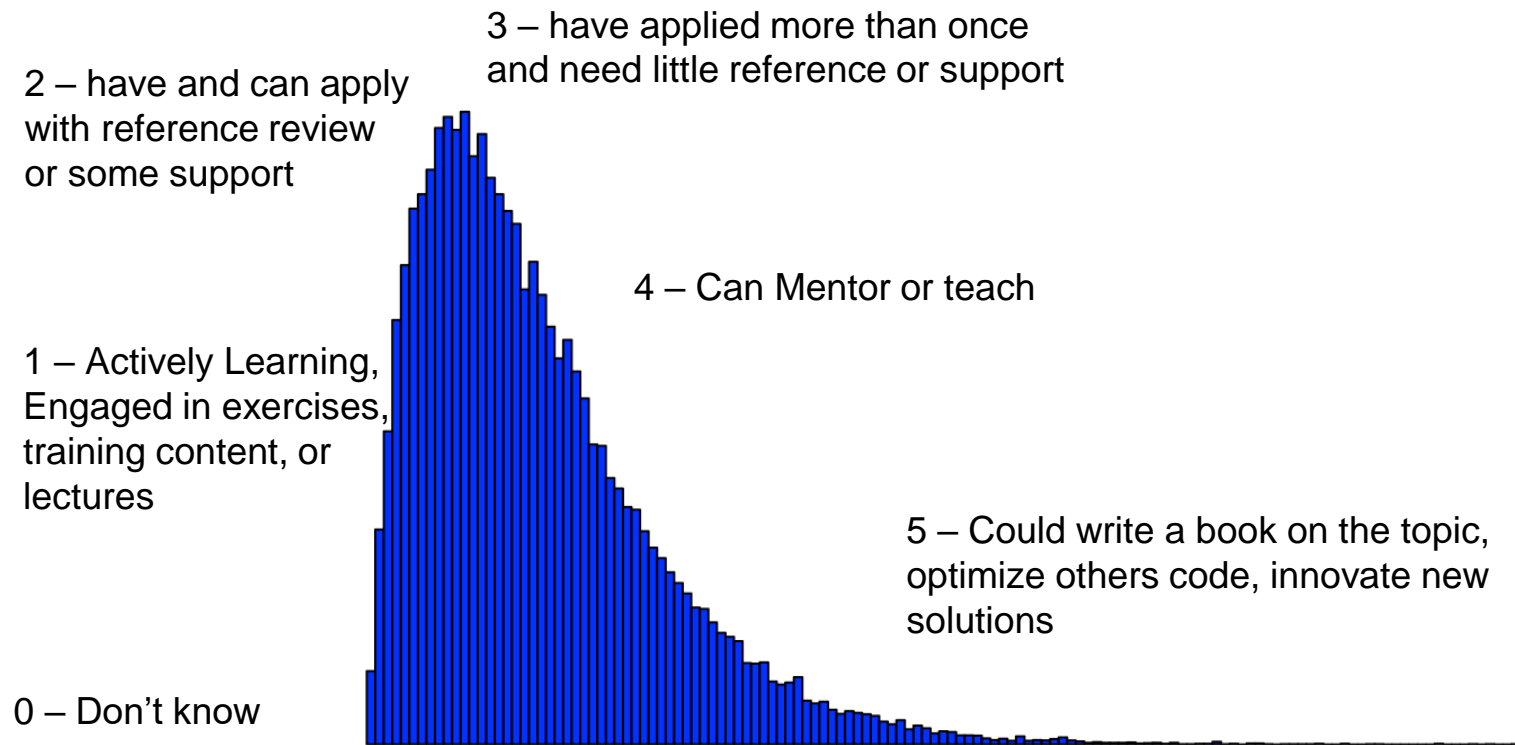
Continuous Data – Tenure of employees at a company, Length of movies on streaming platform

**Qualitative/Categorical[Nominal Data Ordinal Data]:** cannot be expressed as a number and cannot be measured. Qualitative consists of words, symbols, or pictures. What is? Who is? Where is?

Ordinal: Place numbers in order but math cannot be done on the order – Heat setting on stove; low, medium, high

Nominal: Used for Labeling variables – Type of food cooked: Thia, Mexican, Indian

# Tips for success



- 1) 8 data types were described, name as many as you can recall.
- 2) What is the difference between Qualitative and Quantitative?
- 3) What is the difference between Ordinal and Nominal?
- 4) Are Ordinal and Nominal considered Quantitative or Qualitative?
- 5) What is the difference Discreet and continuous?
- 6) Are Discreet and Continuous Quantitative or Qualitative?

## Classroom Exercise

1. Download the titanic dataset
2. Identify a question
3. What are observations made about the data?
4. What next steps would you make?

<https://github.com/Morrisdata/data/blob/main/titanic.xls>

## Bonus Content

Github is an in-depth versioning and collaboration tool. While we do not cover it extensively in class it is a good tool to be familiar with. Organization such as Microsoft and Amazon have their own versions of a Git environment. Using Github as a way to understand and practice is not a bad idea.

<https://docs.github.com/en/get-started/quickstart/hello-world>

# Tasks

- Familiarize with Git content
- Complete any Python pre-work content
- Connect with Students
- Set up folder structures for ease of Content access
- Consider the final project and a data set you are interested in

*Matthew Morris*

*Git: Morrisdata*

*MatthewMorris.DA@gmail.com*