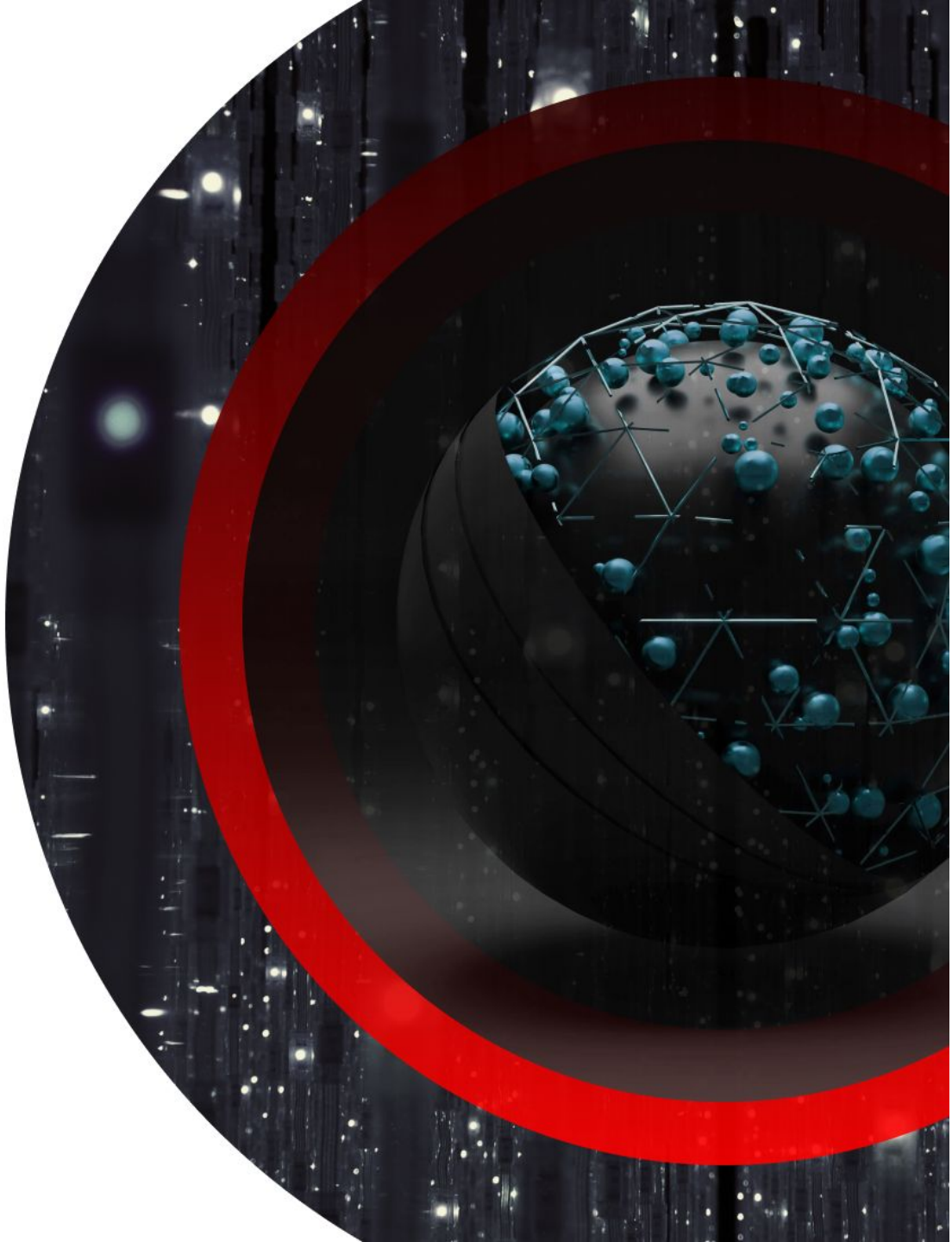
DATA SCIENCE AND ANALYTICS

Introductory Course



CLASS NORMS

COURTESY IN CLASS

Remaining on mute unless called on, exercising courtesy during breakout rooms, using the chat box for questions only

ATTENDANCE

100% attendance is expected and contributes to success in passing the course and the program

PARTICIPATION

Keeping an open mind in discussions and sharing experiences, making contributions during team assignments, submitting assignments in Canvas, and participating in discussion boards

USE OF CLASS RESOURCES

Follow along during the lecture with the lesson companion and download any in-class documents prior to class.

PROGRAM PATH

1 Introductory Course

2 SQL and Databases

3 Statistics and Probability

4 Data Storytelling

Milestone 1: Building and Presenting Data Stories

5 Python Programming

6 Data Wrangling

7 Visual Communications

8 Advanced SQL Programming

Milestone 2: Data Integration, Preparation, Reporting, and Presentation

9 Business Intelligence

10 Big Data

11 Machine Learning

12 Applied AI

Milestone 3: Capstone Project: Delivering Insights and Presentations

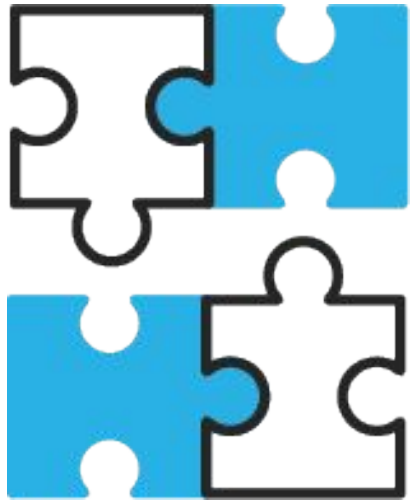
INTRODUCTORY COURSE PATH

			1 INTRODUCTION TO DATA SCIENCE AND ANALYTICS
2 COMPUTING PRIMER	**3** PROGRAMMING CONCEPTS	**4** DISCOVERING AND CURATING DATA	**5** STRUCTURING AND ANALYZING DATA
6 CLEANING AND ENRICHING DATA	**7** VALIDATING AND PRESENTING DATA	**8** **INTRODUCTION TO DATA SCIENCE PROJECTS**	**9** ASSESSMENT NIGHT

LESSON 1: INTRODUCTION TO GITHUB

LESSON OUTLINE

- Define GitHub and version control.
- Compare Git and GitHub.
- Use the GitHub GUI to manage files.
- Build a data analytics portfolio.



DISCUSSION

Imagine you are collaborating with five other people on a single document that can't be shared. Each of you is working separately but often on the same sections. You plan to merge all the changes after you have each finished working.

What are some of the challenges you might face as you merge the final document?

LESSON GOALS

WHAT ARE THE GOALS?

Define GitHub and version control.

Compare Git and GitHub.

Use the GitHub GUI to manage files.

Build a data analytics portfolio.

WHY ARE THEY IMPORTANT?

Data professionals use GitHub to store files and data.

It is particularly useful for managing collaborations.

YOUR TAKE

What are your goals for this lesson?

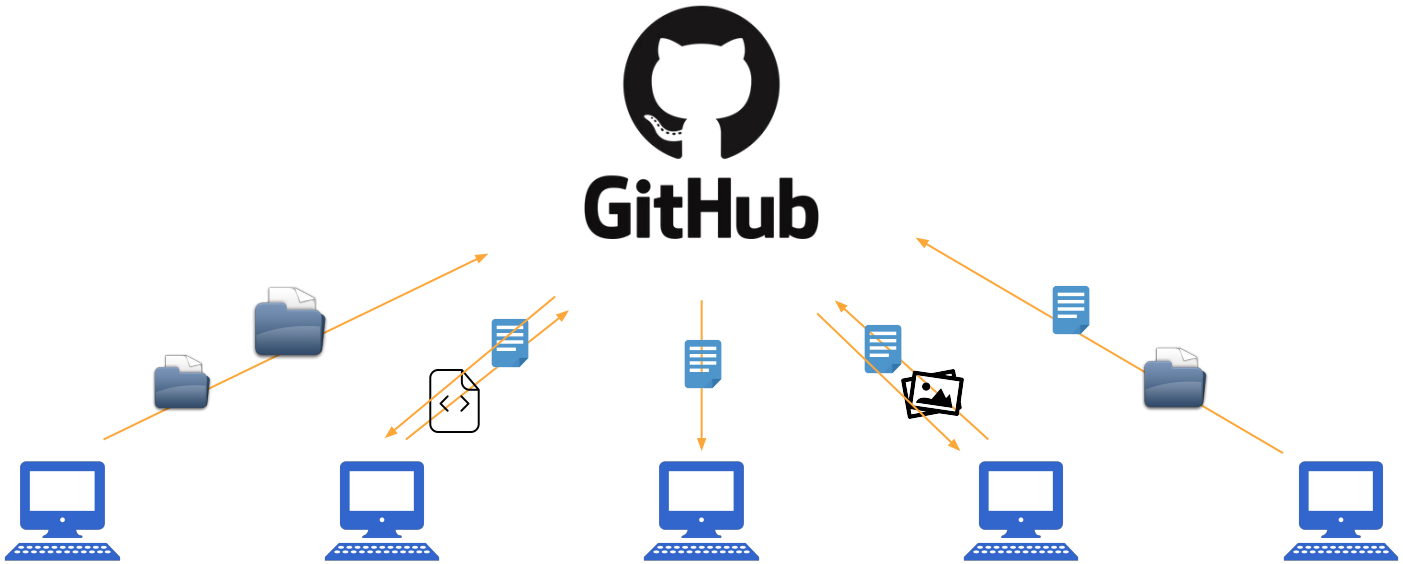
Share three things you know about the lesson topic and two things you want to know about the lesson topic.

WHAT IS GITHUB?

"GitHub is a website and cloud-based service that helps developers store and manage their code, as well as track and control changes to their code."
—Kinsta

GitHub.com

HOW DOES GITHUB WORK?

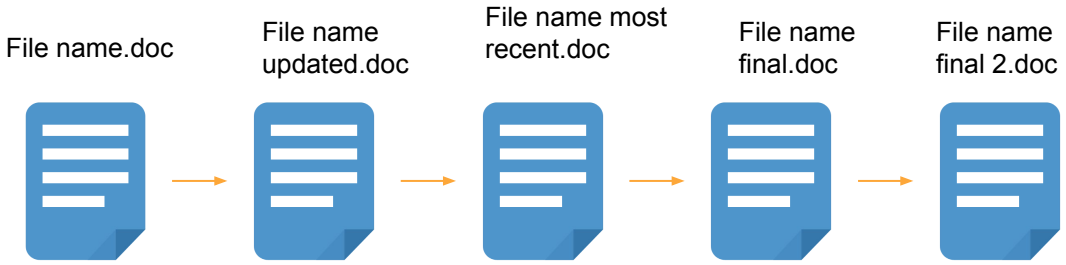


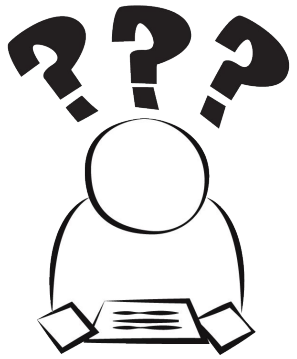
From: BSCW, Wikimedia Commons, FindIcons, OnlineWebFonts

VERSION CONTROL

"Version control helps developers track and manage changes to a software project's code."—Kinsta

Managing files without version control:





From: Cliparts.co

WHAT IS GIT?

"Git is software for tracking changes in any set of files…"

For more information, visit:
https://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

From: git-scm.com

3 WAYS TO MOVE FILES TO/FROM GITHUB

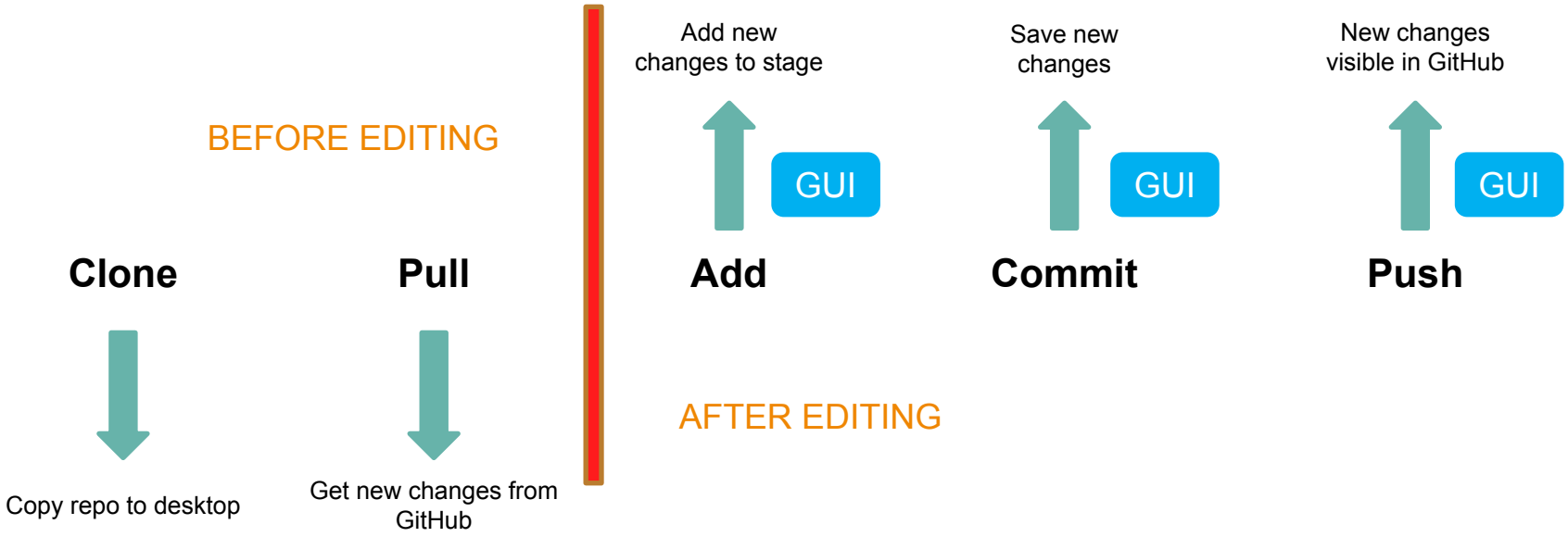
- GitHub GUI
  GUI: Graphical user interface

- Git Bash
  Has many aliases: shell, bash, command prompt, terminal

- GitHub Desktop
  A desktop app

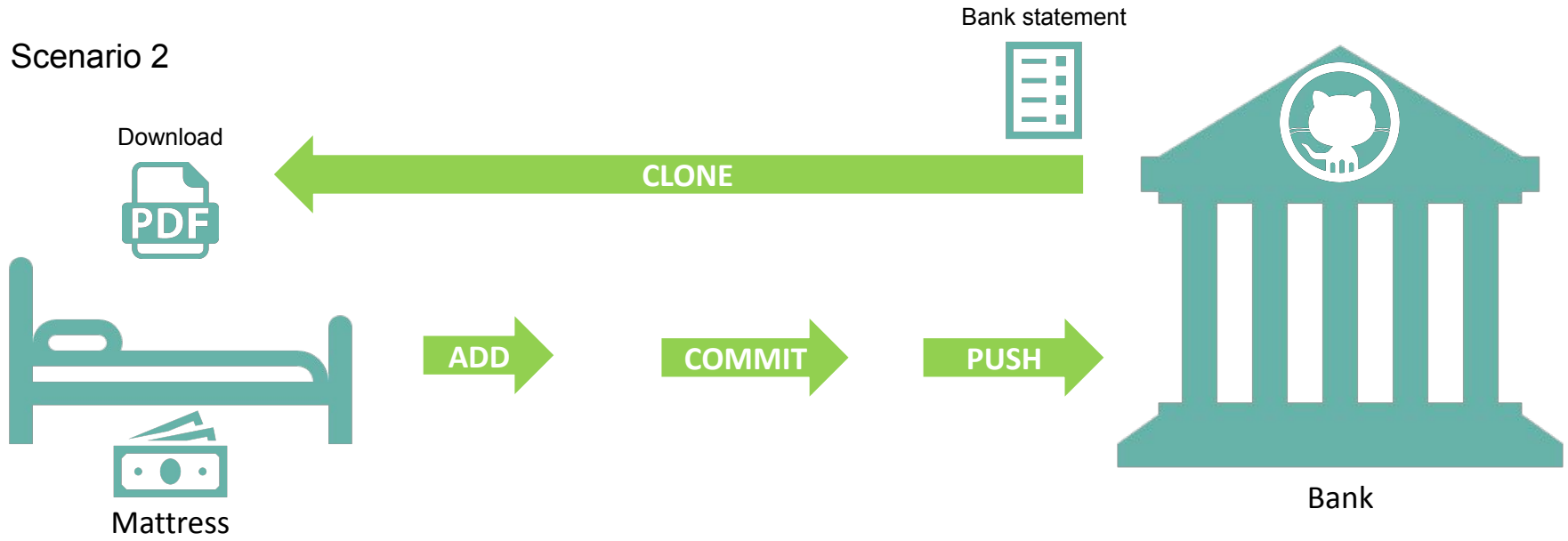
From: IoT One, Lifewire, GitHub

THE GITHUB METHOD



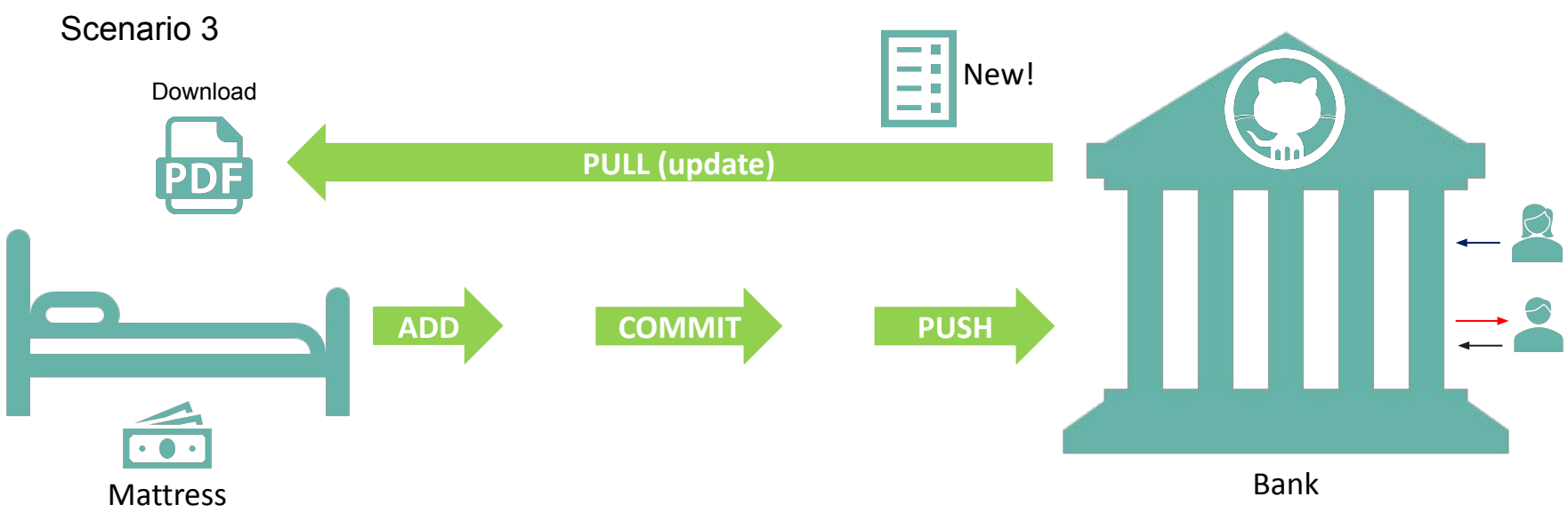
THE GITHUB METHOD: BANK METAPHOR

THE GITHUB METHOD: BANK METAPHOR



From: OnlineWebFonts

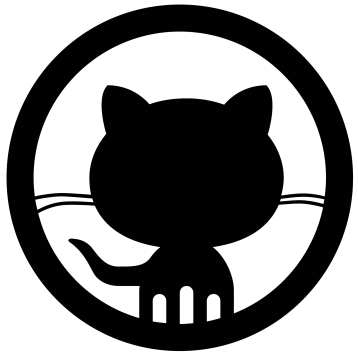
THE GITHUB METHOD: BANK METAPHOR



From: OnlineWebFonts

THE GITHUB METHOD DEMO (OPTIONAL)

1. Clone your data analytics portfolio in GitHub.
2. Change a file.
3. Change into your new repo in Git Bash.
4. Add the new changes to the Git *stage*.
5. Commit your changes to Git.
6. Push your changes to GitHub.



From: PlusPNG

DISCUSSION

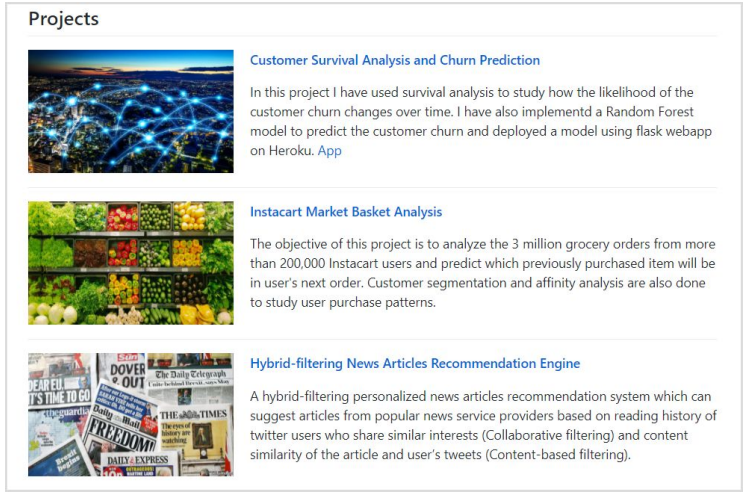
What are the five main commands you will use repeatedly in GitHub?

GITHUB AS A PORTFOLIO

A place to store your projects

Examples:

- Data-Analyst-Portfolio by mkumar7
- data_science_portfolio by melvfnz
- Data Science Portfolio by Arch Desai



From: GitHub

1.8.1 ACTIVITY: GETTING STARTED IN GITHUB

- In this activity, we will: sign in to Github, create a new repo, edit the ***README.md*** file, and add two files to our ***data-analytics-portfolio*** repo.
- Follow the directions in the 1.8.1 Activity document to complete the assignment.

REVIEW AND WRAP-UP

- Defined GitHub and version control
- Compared Git and GitHub
- Used the GitHub GUI to manage files
- Built a data analytics portfolio

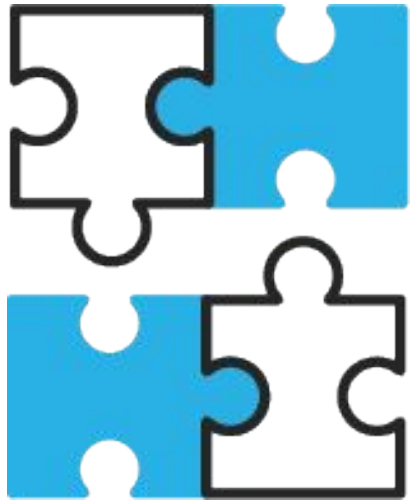
QUESTIONS?

BREAK TIME

LESSON 2: THE DATA SCIENCE WORKFLOW

LESSON OUTLINE

Become familiar with the data science workflow.



LESSON GOALS

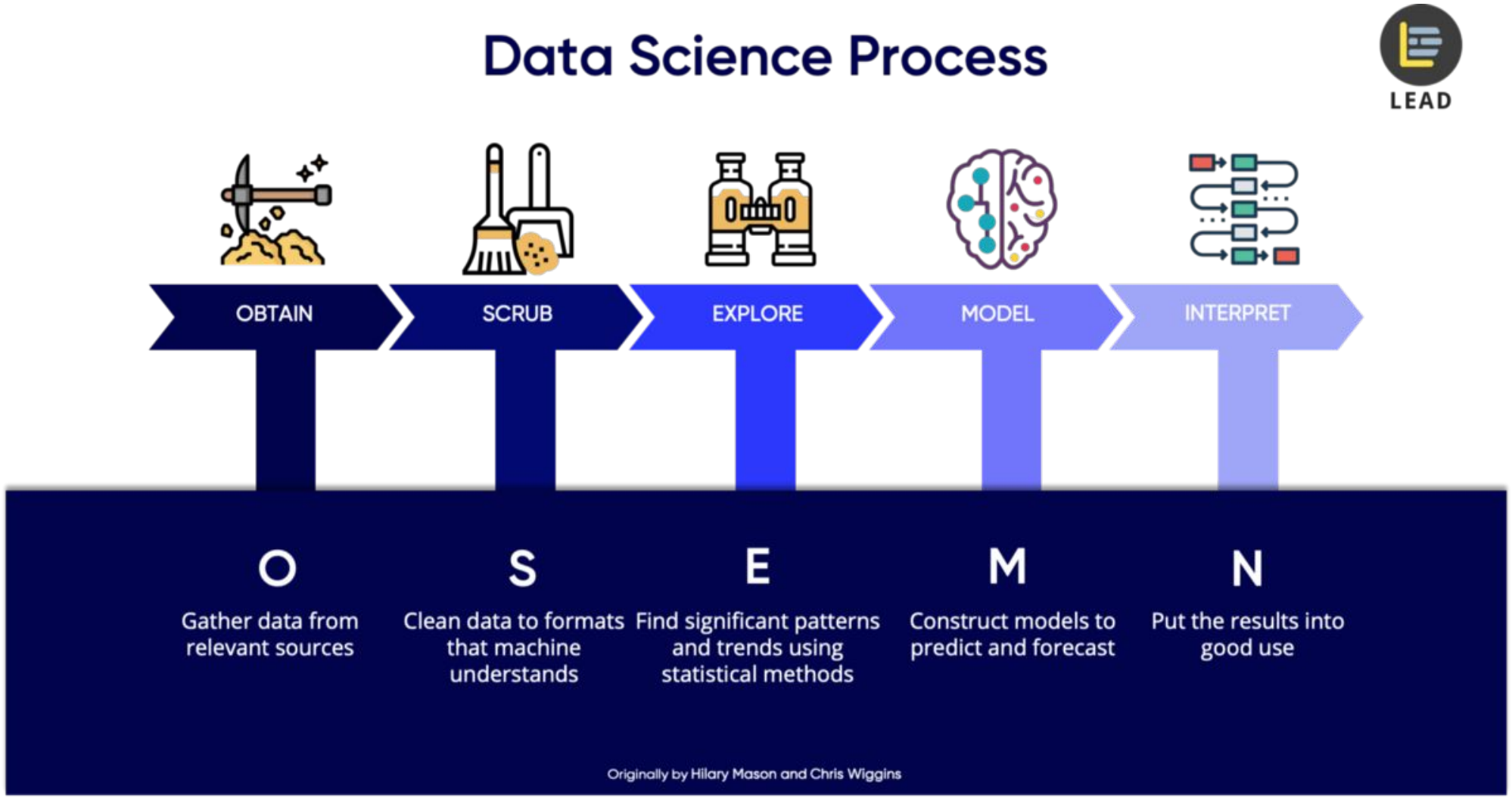
WHAT ARE THE GOALS?

Become familiar with the data science workflow.

WHY ARE THEY IMPORTANT?

Although every data analysis is different, you'll walk through the same set of steps each time on your way to discovering meaning in your data.

Once you know the basics of examining a data set, you can apply those skills to any project.


Data Science Process
LEAD
OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
O
Gather data from relevant sources
S
Clean data to formats that machine understands
E
Find significant patterns and trends using statistical methods
M
Construct models to predict and forecast
N
Put the results into good use
Originally by Hilary Mason and Chris Wiggins

OBTAIN

- Obtaining or acquiring data can include scraping it from the web or gathering it from a source while considering the privacy implications of the information.
- Some common databases for data projects are Kaggle, FiveThirtyEight, and the US Census Bureau.
- Keep in mind the question you want to answer with the project, and evaluate whether the data can provide those answers.

SCRUB (CLEAN)

- Data comes in many different formats.
- Cleaning or scrubbing the data includes preparing the data to be analyzed.
- This step can take the most time depending on the state of the data set.
- Common techniques include changing data types, removing unnecessary columns, and handling missing values.

From: Medium

EXPLORE

- Exploratory data analysis (EDA) includes using preliminary **visualizations** and other summarizing tools to get a sense of the data.
- Once you get a sense of the information in the data set, you can think about possible **predictions** to make.
- A few first steps include finding the average of numeric columns, calculating the frequency of categories, seeing how many rows there are, and determining the **association** between numeric columns using a scatter plot.

MODEL

- Models are how we use data to make predictions and forecasts.
- **Supervised learning** uses data where the “answer” is known and labeled. (The upcoming project will use supervised learning.)
- **Unsupervised learning** uses unlabeled data where there is no clear “answer” that we are trying to predict. In unsupervised learning, we attempt to find hidden patterns through grouping or clustering.

MODEL: LOGISTIC REGRESSION

- There are many different machine learning models (you'll learn about many of them later in the machine learning course). One of the simpler models is called a logistic regression.
- "In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead, or healthy/sick."
  —Wikipedia

INTERPRET

- Put the results to good use.
- Take predictions and insights and communicate them to stakeholders.
- Document patterns and abnormalities.

INTERVIEW QUESTION

What is the main difference between supervised and unsupervised learning?

YOUR TAKE

- What was one thing you learned?
- What is something you would like to learn more about?

REVIEW AND WRAP-UP

Became familiar with the data science workflow

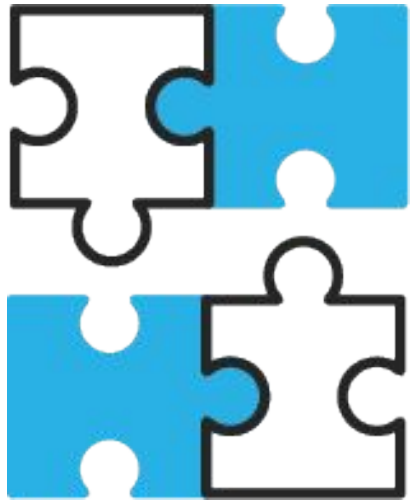
QUESTIONS?

BREAK TIME

LESSON 3: PROJECT: TITANIC SURVIVORS

LESSON OUTLINE

Analyze a real-world data set using the steps of the data science workflow.



LESSON GOALS

WHAT ARE THE GOALS?

Analyze a real-world data set using the steps of the data science workflow.

WHY ARE THEY IMPORTANT?

At the end of the lesson, you'll have finished your first analysis as a data scientist!

DATA: TITANIC

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered *unsinkable* RMS Titanic sank after colliding with an iceberg.

Unfortunately, there weren't enough lifeboats for everyone on board, resulting in the death of 1,502 people out of 2,224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.
—Kaggle

DATA: TITANIC

The data set has columns with information about the Titanic passengers and whether they survived in the **survival** column.

Use the features of the passengers to predict whether they survived.

This will provide insights on which passengers were more likely to survive the sinking.

DATA DICTIONARY

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

DISCUSSION

- What do you notice about the data set?
- Which columns do you think affect the passengers' survival?
- How might we need to clean or modify before running a model?

1.8.3 ACTIVITY

- You will complete the 1.8.3 Activity in the ***1.8.3 Activity.ipynb*** file.
- A file that ends in .ipynb is called a Jupyter Notebook file (note the “py” in Jupyter, which refers to Python).
- We will be able to open and run this file on a web service called Google Colaboratory.
- See the instructions in your module companion for getting started with Colab and completing the activity.

YOUR TAKE

- What was one thing you learned during the activity?
- What is something you would like to learn more about?

REVIEW AND WRAP-UP

Today, you learned about:

- How to analyze a real-world data set

QUESTIONS?

NEXT STEPS

- [x] Assigned Activities
- [x] Reminders