



WORKSHOPS

Week 1 –Introduction

Danai Korre



PYTHON

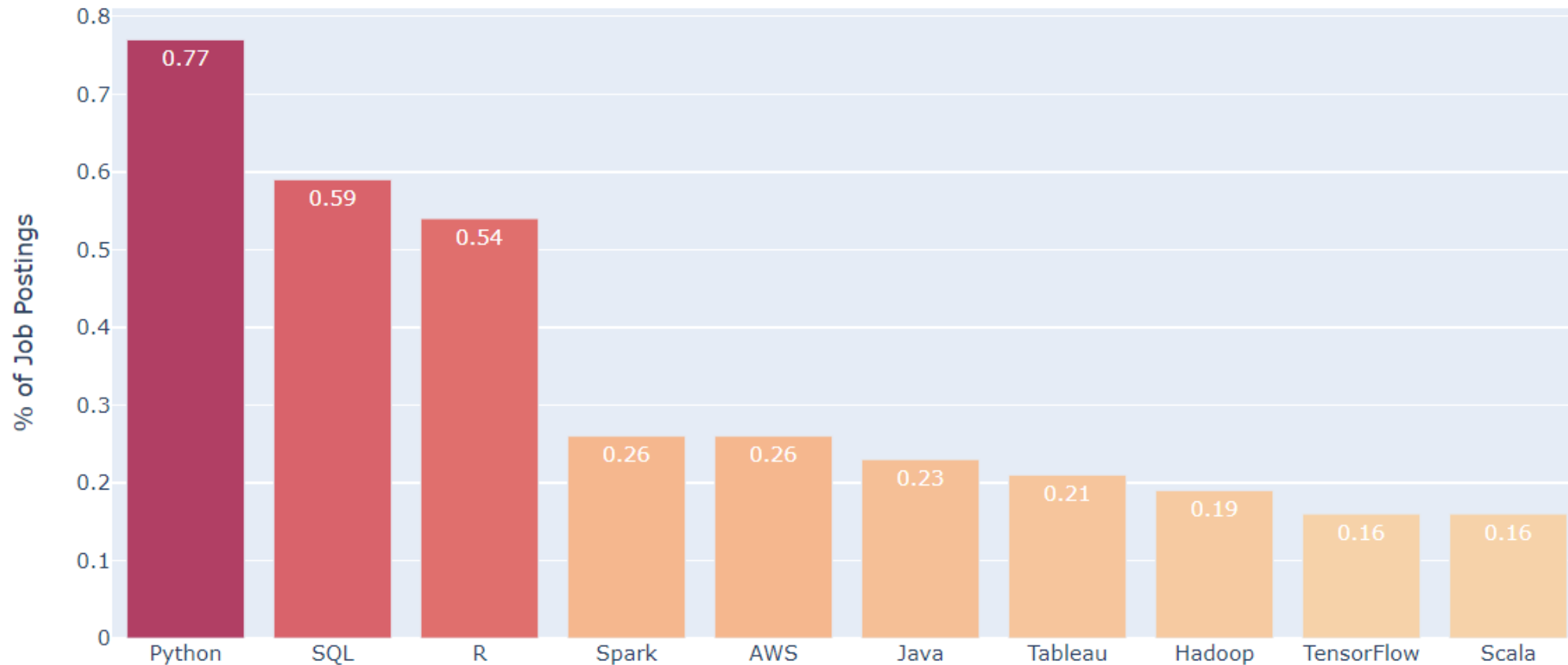
WHAT IS PYTHON?

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed [1].

THE MOST IN-DEMAND SKILLS FOR DATA SCIENTISTS IN 2021

RESULTS FROM WEBCRAPING OVER 15,000 DATA SCIENTIST JOB POSTINGS

10 Most In-Demand Data Science Skills in 2021



TOP SKILLS FOR DATA SCIENTISTS IN 2022

According to the Artificial Intelligence Report the top skills for data scientists in 2022 are:

- Data Wrangling / Data Carpentry/ Feature Engineering
- Writing SQL Queries & Building Data Pipelines
- Storytelling (i.e. Communication) to complement Data visualization
- Regression/Classification

WHY PYTHON?

- It is free and open-source software.
- It is well-documented and runs on all platforms.
- It has a large and constantly growing user-base which includes scientists.
- It is easier for novices to pick up than most other languages.

WHY PYTHON FOR DATA ANALYSIS?

Python has developed a large and active scientific computing and data analysis community. In the last 10 years, Python has become one of the most important languages for data science, machine learning, and general software development in academia and industry. In recent years, Python's improved support for libraries (such as pandas and scikit-learn) has made it a popular choice for data analysis tasks. Combined with Python's overall strength for general-purpose software engineering, it is an excellent option as a primary language for building data applications [2].

ANACONDA



WHAT IS ANACONDA?

Anaconda is a distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS [3].



Anaconda at a glance 🧐

25M

Anaconda users

235

Countries & regions with
Anaconda users

2.4

Billion package
downloads in 2019

24,684

New packages added to
anaconda.org in 2019

192

Pounds of jellybeans
fueled our team last year

3,000+

Lego minifigs given
away at our events

JUPYTER NOTEBOOK



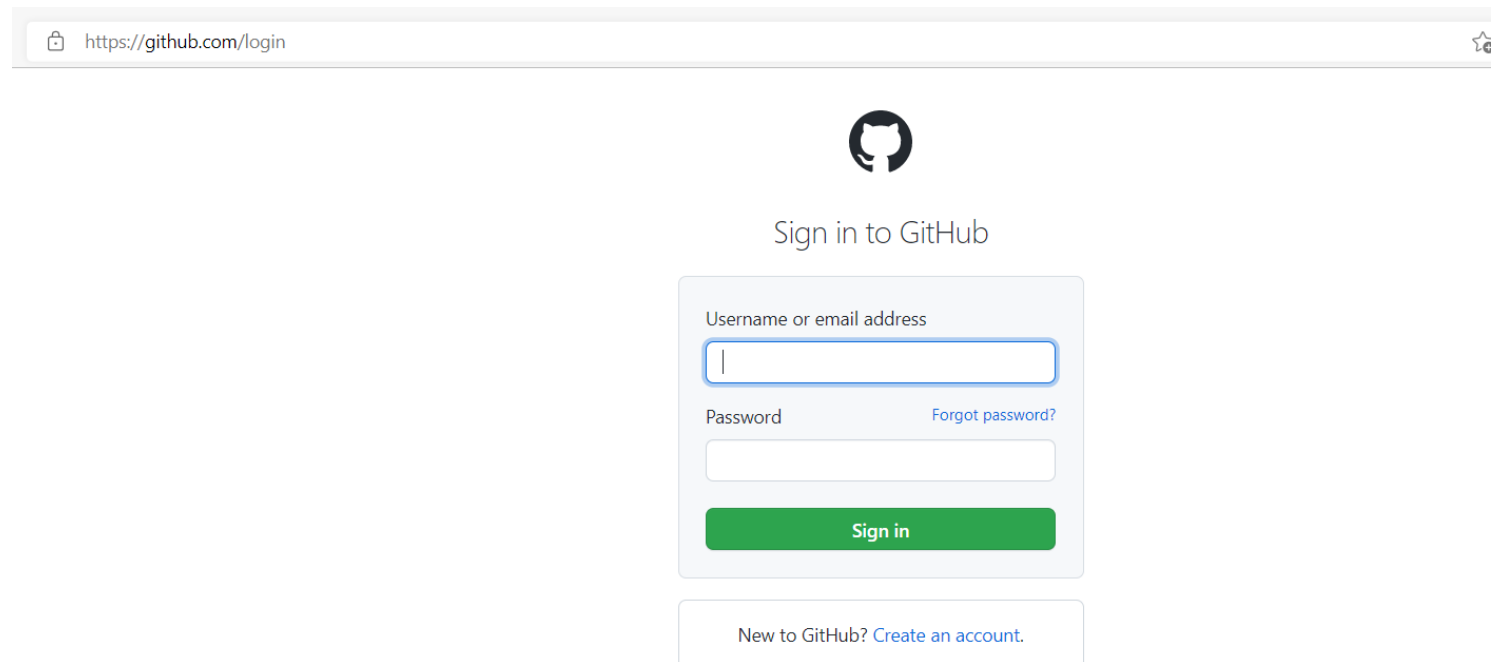
WHAT IS NOTEABLE?

The Noteable service is a cloud-based application providing access to Jupyter notebooks online. Noteable provides a central storage space to store and run Jupyter notebooks in a variety of languages.

The purpose of Noteable is to allow students and staff to access Jupyter notebooks at any time without the need for pre-installation which can be cumbersome and difficult for programming novices. Noteable is integrated with Learn to allow for a central launch point into a pre-set environment without the need for a separate login. Find more: <https://www.ed.ac.uk/information-services/learning-technology/noteable/about>

SET UP

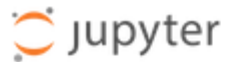
Step 1: Log in to your github account. If you don't have an account, please go to <https://github.com/> and sign up for a free account. It will be best if you used your personal email so you retain access to the material independent of your University account.



The screenshot shows the GitHub login interface. At the top, the browser address bar displays 'https://github.com/login'. Below this is the GitHub logo (Octocat) and the text 'Sign in to GitHub'. The login form consists of two input fields: 'Username or email address' and 'Password'. The 'Username or email address' field has a cursor in it. To the right of the password field is a link that says 'Forgot password?'. Below the input fields is a green 'Sign in' button. At the bottom of the form is a link that says 'New to GitHub? Create an account.'

SET UP

Step 2: Log in to Noteable
(<https://noteable.edina.ac.uk/launch>) using your University of Edinburgh student credentials.



Files

Running

Assignments

Select items to perform actions on them.

+GitRepo

Disk

Empty Trash

Upload

New ▾

☐ 0 ▾

/

Name ▾

Last Modified

File size

The notebook list is empty.

SET UP

Step 3: Click on the +GitRepo icon as shown bellow and add the github repository with the Jupyter Notebooks and data files necessary for the workshops.



FilesRunningAssignments

Select items to perform actions on them.

0

/

Enter the details of the Git Repository to clone:

Git Repository URL:

https://github.com/GSA/dat

Branch*:

<default>

Username*:

Password*:

* Optional

Clone

Cancel

+GitRepo

Disk

Empty Trash

Upload

New

Pull down a Git repository

Last Modified

File size

SET UP

Step 4:

Git Repository URL:

Branch: as is

Username: as is (your github username when repo is private)

Password: as is (your github password when repo is private)



FilesRunningAssignments

Select items to perform actions on them.

☐ 0

☐ /

Enter the details of the Git Repository to clone:

Git Repository URL:

https://github.com/GSA/data

Branch*:

<default>

Username*:

Password*:

* Optional

Clone

Cancel

+GitRepo

Disk

Empty Trash

Upload

New

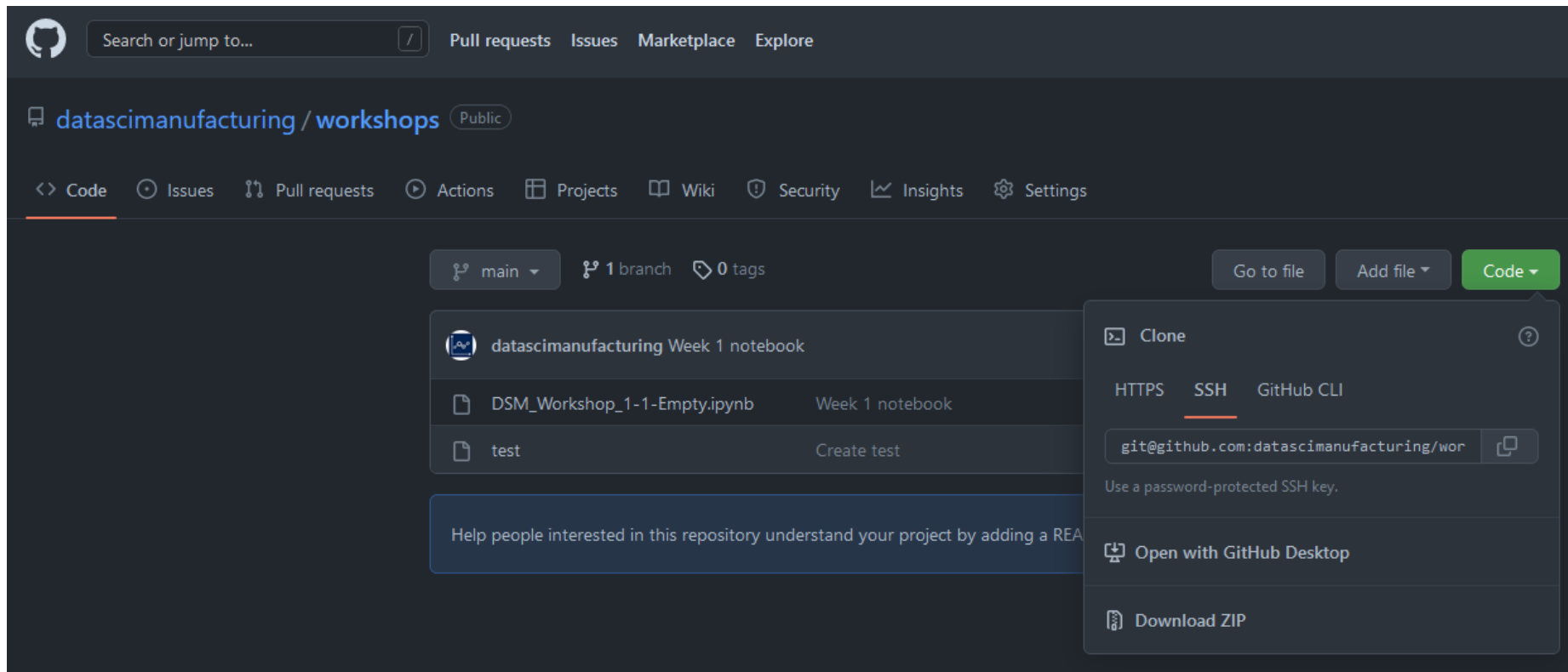
Pull down a Git repository

Last Modified

File size

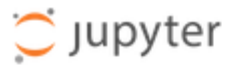
SET UP

You can get the repository URL by clicking on the Code button on the github repository



SET UP

Step 5: Your Noteable page should look like this

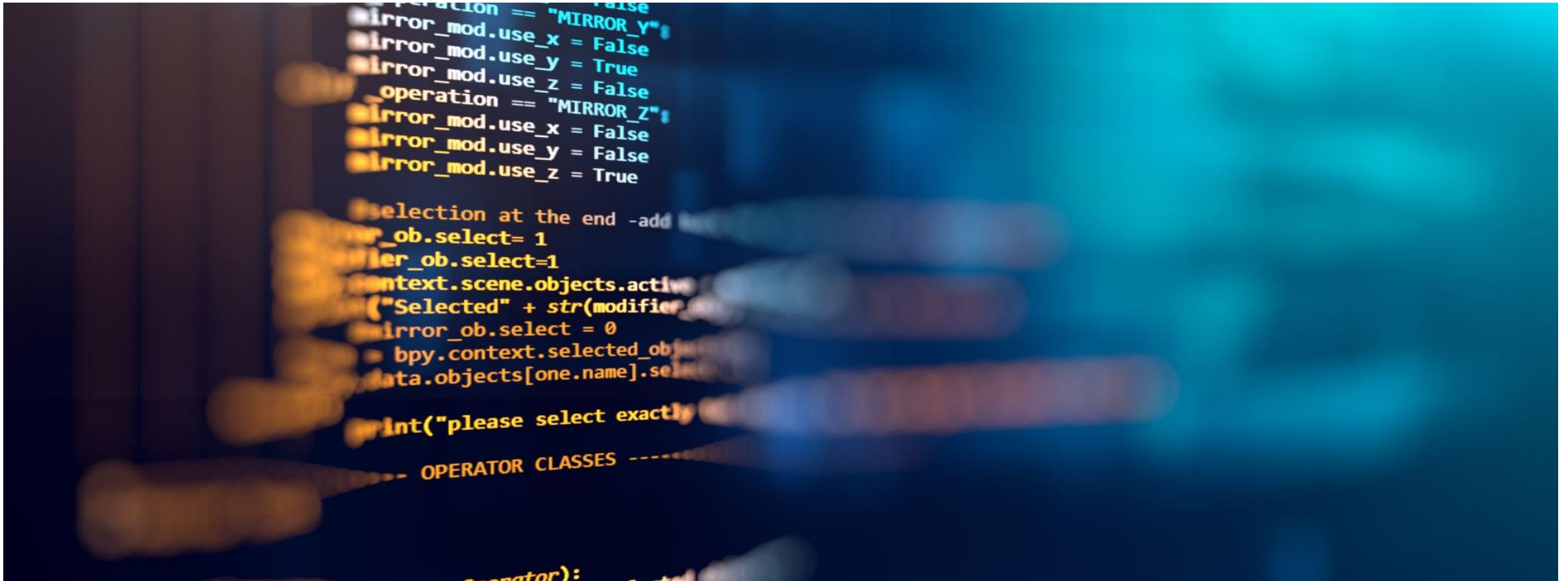


Files Running Assignments

Select items to perform actions on them.

+GitRepo Disk Empty Trash Upload New ↕

<input type="checkbox"/> 0	/ Workshops.git	Name ↓	Last Modified	File size
<input type="checkbox"/>	..		seconds ago	
<input type="checkbox"/>	Week_1		seconds ago	
<input type="checkbox"/>	LICENSE		seconds ago	1.52 kB



DATA: TYPES AND FORMATS

TYPES OF DATA

- Quantitative: numerical
- Qualitative: text

QUANTITATIVE

- Annual sales
- Profitability

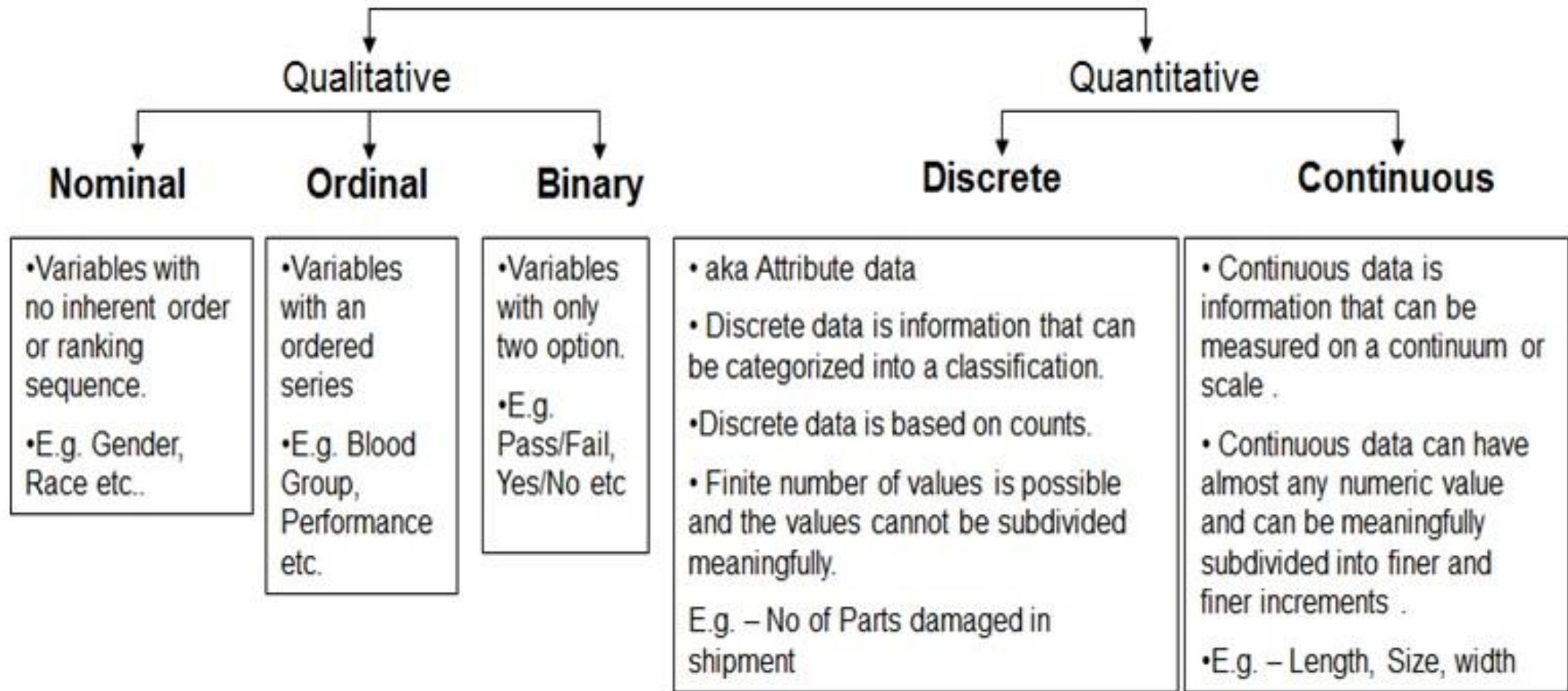
	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

QUALITATIVE

- Customer reviews
- Explicit description of a malfunction in evaluation reports

	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

TYPES OF DATA



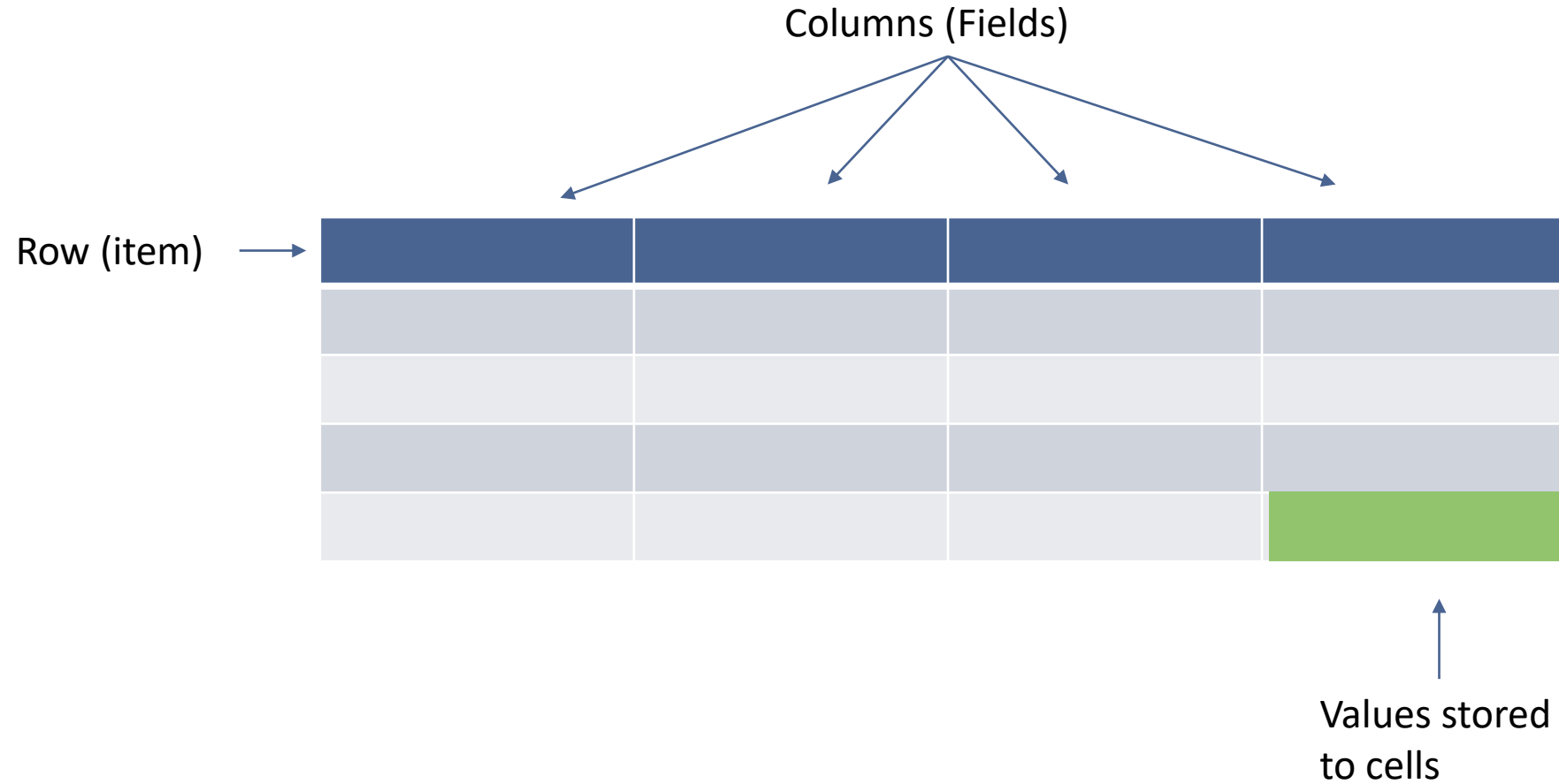
STRUCTURED DATA

- Tabular or spreadsheet-like data in which each column may be a different type (string, numeric, date, or otherwise). This includes most kinds of data commonly stored in relational databases or tab- or comma-delimited text files.
- Multidimensional arrays (matrices).
- Multiple tables of data interrelated by key columns (what would be primary or foreign keys for a SQL user).
- Evenly or unevenly spaced time series [1].

TABULAR AND CSV

- Tabular formats (ex. Excel)
- Comma-delimited text files, CSV
format

TABULAR DATA



TABULAR DATA

	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

CSV FORMAT

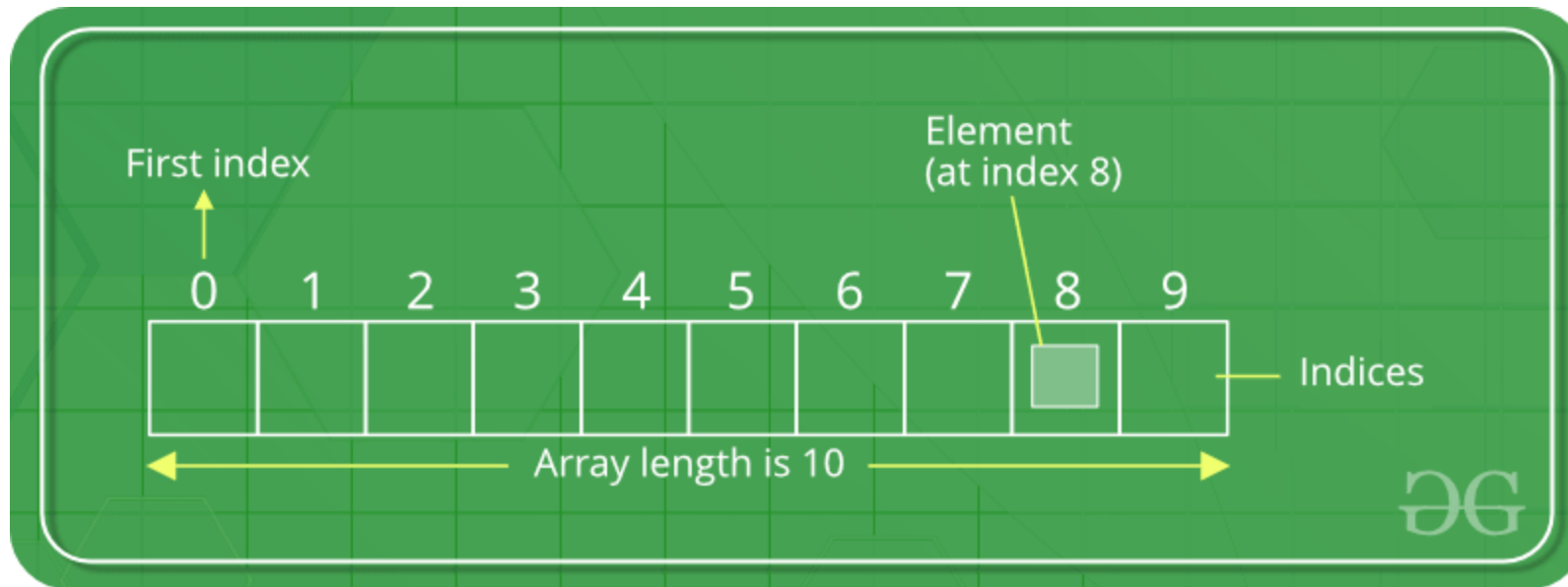
- CSV stands for Comma Separated Values
- Saves tabular information into a delimited text file with the series of values separated by commas
- It is lightweight and consumes less memory
- Each line of text is a single row
- It is human readable and can be opened using a text editor vs tabular data that are stored as binary files.

CSV FORMAT EXAMPLE

```
CollectorID,StartDate,EndDate,I found Moneyworld confusing to use.  
92742972,10/26/2016 14:10:35,10/26/2016 14:12:43,Slightly Disagree,  
92742972,10/26/2016 10:03:07,10/26/2016 11:21:47,Strongly Disagree,  
92742972,10/25/2016 16:56:13,10/25/2016 17:14:13,Disagree,Disagree,  
92742972,10/25/2016 14:42:38,10/25/2016 15:34:03,Disagree,Strongly  
92742972,10/25/2016 11:40:06,10/25/2016 12:23:13,Disagree,Disagree,  
92742972,10/25/2016 09:59:18,10/25/2016 10:22:46,Slightly Agree,Dis  
92742972,10/24/2016 17:12:38,10/24/2016 17:28:29,Slightly Agree,Neu  
92742972,10/24/2016 15:45:40,10/24/2016 16:30:08,Slightly Disagree,  
92742972,10/24/2016 14:19:51,10/24/2016 14:22:06,Strongly Disagree,  
93508248,10/24/2016 11:01:33,10/24/2016 11:03:43,Strongly Disagree,
```

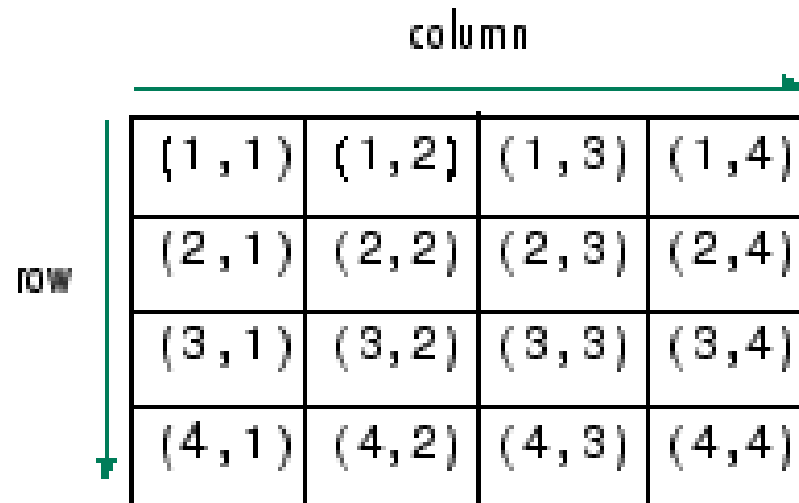

ARRAYS

- An array is a special variable, which can hold more than one value at a time.



MULTIDIMENSIONAL ARRAYS

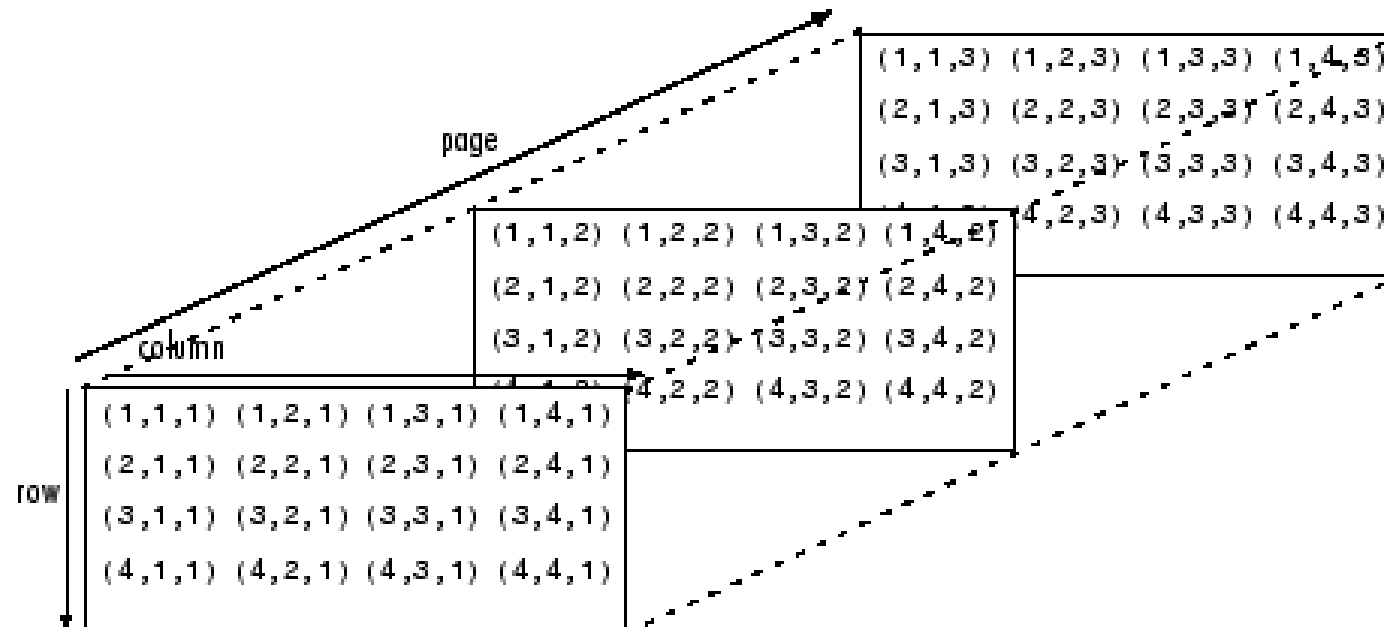
- Multidimensional arrays are data arrays with more than two dimensions. Each element is defined by two subscripts, the row index and the column index.



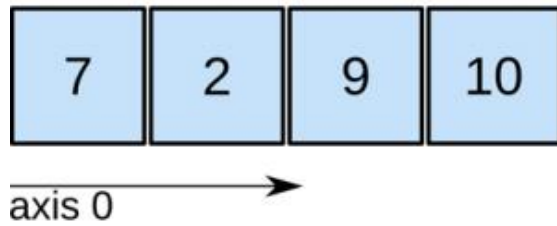
	column			
row	(1, 1)	(1, 2)	(1, 3)	(1, 4)
	(2, 1)	(2, 2)	(2, 3)	(2, 4)
	(3, 1)	(3, 2)	(3, 3)	(3, 4)
	(4, 1)	(4, 2)	(4, 3)	(4, 4)

MULTIDIMENSIONAL ARRAYS

- Multidimensional arrays are an extension of 2-D matrices and use additional subscripts for indexing. A 3-D array, for example, uses three subscripts. The first two are just like a matrix, but the third dimension represents pages or sheets of elements [2].

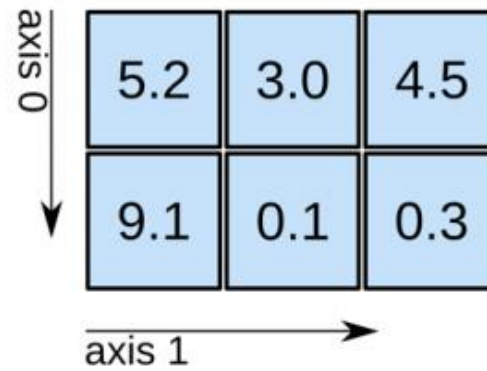


1D array



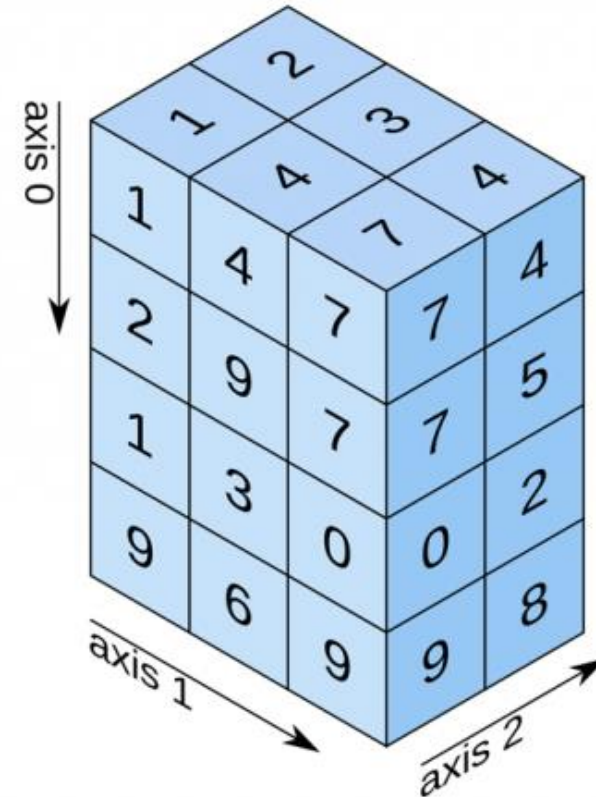
shape: (4,)

2D array



shape: (2, 3)

3D array



shape: (4, 3, 2)

REFERENCES

1. Rosett C.M., Hagerty A. (2021) Data Wrangling. In: Introducing HR Analytics with Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-67626-1_13
2. Stefanski, R., Sinha, V. and Poddar, A., 2022. *Data Wrangling in 6 Steps: An Analyst's Guide For Creating Useful Data*. [online] Learn | Hevo. Available at: <https://hevodata.com/learn/data-wrangling/#s2>
3. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M. and Emmert-Streib, F., 2021. Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing. *Frontiers in Artificial Intelligence*, 4.

REFERENCES

1. Python.org. 2022. *What is Python? Executive Summary*. [online] Available at: <<https://www.python.org/doc/essays/blurb/>>
2. McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.
3. En.wikipedia.org. 2022. *Anaconda (Python distribution)* - Wikipedia. [online] Available at: [https://en.wikipedia.org/wiki/Anaconda_\(Python_distribution\)](https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)).



RESOURCES

1. <https://docs.anaconda.com/anacondaorg/faq/#what-is-anaconda-inc>
2. <https://docs.anaconda.com/anacondaorg/glossary/#cloud-glossary-cloud>

