

NOVA

IMS

Information
Management
School

Analysing Big Data

Introduction to Databricks, Python & Spark

Maria Almeida

malmeida@novaims.unl.pt

Niclas Frederic Sturm

nsturm@novaims.unl.pt

About me – Maria

- **Academic Path**
 - M.Sc. in Data Science and Engineering (Instituto Superior Técnico)
 - B.Sc. in Information Management (NOVA IMS)
- **Work**
 - Machine Learning Engineer @ Neuraspace (Space Traffic Management)

About me – Niclas

- **Academic Path**
 - Ph.D. Student in *Information Management* (Nova IMS)
 - M.Sc. in Business Analytics (Nova SBE)
 - B.Sc. in Economics & Ancient History (University of Heidelberg)
- **Research Interests**
 - Public Procurement
 - Network Science
 - Quantitative Methods in the Social Sciences and Humanities



General Purpose Cluster Computing Framework for Big Data



- Working in a notebook with Python (and soon PySpark)
- Using variables, operators, built in functions
- Controlling code flow with conditional statements and loops
- Using data types including lists, dictionaries, and tuples
- Defining and using both named functions and anonymous functions (lambda functions)

Databricks Community Edition



databricks®


- Community: free



Goals for the class

- 1. Creation of a Databricks account**
- 2. Introduction to Databricks**
 - Overview of Databricks and its features
 - Comparison of Databricks and Jupyter Notebook
 - Benefits of using Databricks
- 3. Uploading a File to Databricks**
- 4. Introduction to Distributed Processing**
 - Databricks Cluster
- 5. Introduction to Shell Commands**
 - Overview of shell commands and their importance in Big Data processing
 - How to use shell commands in Databricks


Databricks Community Edition

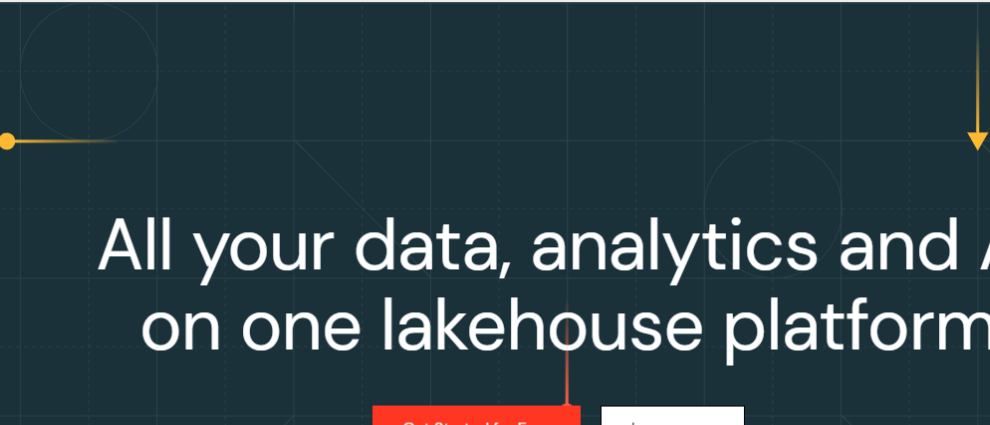

databricks

Platform
Solutions
Learn
Customers
Partners
Company

Try Databricks


Watch Demos
Contact Us
Login

**2021 Gartner reports:** From data warehousing to machine learning, Databricks is a Leader
Learn why the Databricks Lakehouse Platform is able to deliver on both data warehousing and machine learning use cases.
[Get the reports →](#)



All your data, analytics and AI on one lakehouse platform

[Get Started for Free](#)[Learn more](#)



Lakehouse

Data lakehouse: The best of both worlds

Lakehouse combines the reliability, performance and governance of data warehouses with the openness and flexibility of data lakes. Now you can

 <https://towardsdatascience.com/what-does-databricks-do-8a6c4ef9071b>

Databricks Community Edition

<https://databricks.com/try-databricks>



Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud.

- ✓ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ✓ Collaborate in your preferred language
Code in Python, R, Scala and SQL with coauthoring, automatic versioning, Git integrations and RBAC.
- ✓ 12x better price/performance than cloud data warehouses
See why over 7,000 customers worldwide rely on Databricks for all their workloads from BI to AI.



Create your Databricks account

1/2

First name

Last Name

Email

Company

Title

Phone (Optional)

Country

- ☐ Yes, I would like to receive marketing communications regarding Databricks services, events and open source products. I understand I can update my preferences at any time.

Continue

1. Fill in the registration form (**hint: use your personal email**)
2. Click **"Get Started for free"**
3. Solve the verification enigma
4. Choose **"Get started with Community Edition"** (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Databricks Community Edition

<https://databricks.com/try-databricks>



Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud.

- ✓ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ✓ Collaborate in your preferred language
Code in Python, R, Scala and SQL with coauthoring, automatic versioning, Git integrations and RBAC.
- ✓ 12x better price/performance than cloud data warehouses
See why over 7,000 customers worldwide rely on Databricks for all their workloads from BI to AI.



Choose a cloud provider
2 / 2

Amazon Web Services

Microsoft Azure

Google Cloud Platform

Continue

By clicking "Get Started," you agree to the [Privacy Policy](#) and [Terms of Service](#).

Don't have a cloud account?

Community Edition is a limited Databricks environment for personal use and training.

Get started with Community Edition →

By clicking "Get started with Community Edition," you agree to the [Privacy Policy](#) and [Terms of Service](#).

1. Fill in the registration form (hint: use your personal email)
2. Click **"Get Started for free"**
3. Solve the verification enigma
4. Choose **"Get started with Community Edition"** (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Databricks Community Edition

<https://databricks.com/try-databricks>



Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud.

- ✓ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ✓ Collaborate in your preferred language
Code in Python, R, Scala and SQL with coauthoring, automatic versioning, Git integrations and RBAC.
- ✓ 12x better price/performance than cloud data warehouses
See why over 7,000 customers worldwide rely on Databricks for all their workloads from BI to AI.



Choose a cloud provider 2/2

Amazon Web Services

Microsoft

Google Cloud Platform

Continue

By clicking "Get Started," you agree to the [Privacy Policy](#) and [Terms of Service](#).

Don't have a cloud account?

Community Edition is a limited Databricks environment for personal use and training.

Get started with Community Edition →

By clicking "Get started with Community Edition," you agree to the [Privacy Policy](#) and [Terms of Service](#).

1. Fill in the registration form (hint: use your personal email)
2. Click **"Get Started for free"**
3. Solve the verification enigma
4. Choose **"Get started with Community Edition"** (in the bottom of the page)
5. When you receive the Welcome to Databricks email, click on the link to verify your email address
6. Reset your password. You are done!

Goals for the class

1. ~~Creation of a Databricks account~~

2. Introduction to Databricks

- Overview of Databricks and its features
- Comparison of Databricks and Jupyter Notebook
- Benefits of using Databricks

3. Uploading a File to Databricks

4. Introduction to Distributed Processing

- Databricks Cluster

5. Introduction to Shell Commands

- Overview of shell commands and their importance in Big Data processing
- How to use shell commands in Databricks

Databricks Overview

- **Databricks:** web-based platform to work with Spark.

Databricks	Jupyter
Cloud-based platform	Local application
For large datasets	For small datasets
Distributed processing	Single-node processing
Offers cluster management, data pipelines, ML	No features

Main Benefits of using Databricks:

- Scalability
- Collaboration;
- Integration with other cloud services

Goals for the class

1. ~~Creation of a Databricks account~~

2. ~~Introduction to Databricks~~

- ~~○ Overview of Databricks and its features~~
- ~~○ Comparison of Databricks and Jupyter Notebook~~
- ~~○ Benefits of using Databricks~~

3. **Uploading a File to Databricks**

4. **Introduction to Distributed Processing**

- Databricks Cluster

5. **Introduction to Shell Commands**

- Overview of shell commands and their importance in Big Data processing
- How to use shell commands in Databricks

Databricks Community Edition



Data Science & Engineering



Notebook

Create a new notebook for querying, data processing, and machine learning.

[Create a notebook](#)



Data import

Quickly import data, preview its schema, create a table, and query it in a notebook.

[Browse files](#)



Guide: Quickstart tutorial

Spin up a cluster, run queries on preloaded data, and display results in 5 minutes.

[Start tutorial](#)

[Upgrade](#)

Recents

Name	Last viewed
There are no recents yet	

Documentation

Get started guide

This tutorial gets you going with Databricks Data Science & Engineering

Best practices

Get the best performance when using Databricks

Data guide

How to work with data in Databricks

[More documentation](#)

Release notes

Runtime release notes

[Databricks preview releases](#)

[Platform release notes](#)

[More release notes](#)

Blog posts

Building a Geospatial Lakehouse, Part 1

December 17, 2021

[Ray on Databricks](#)

November 19, 2021

[10 Powerful Features to Simplify Semi-structured Data Management in the Databricks Lakehouse](#)

November 11, 2021

[More blog posts](#)

Databricks Community Edition

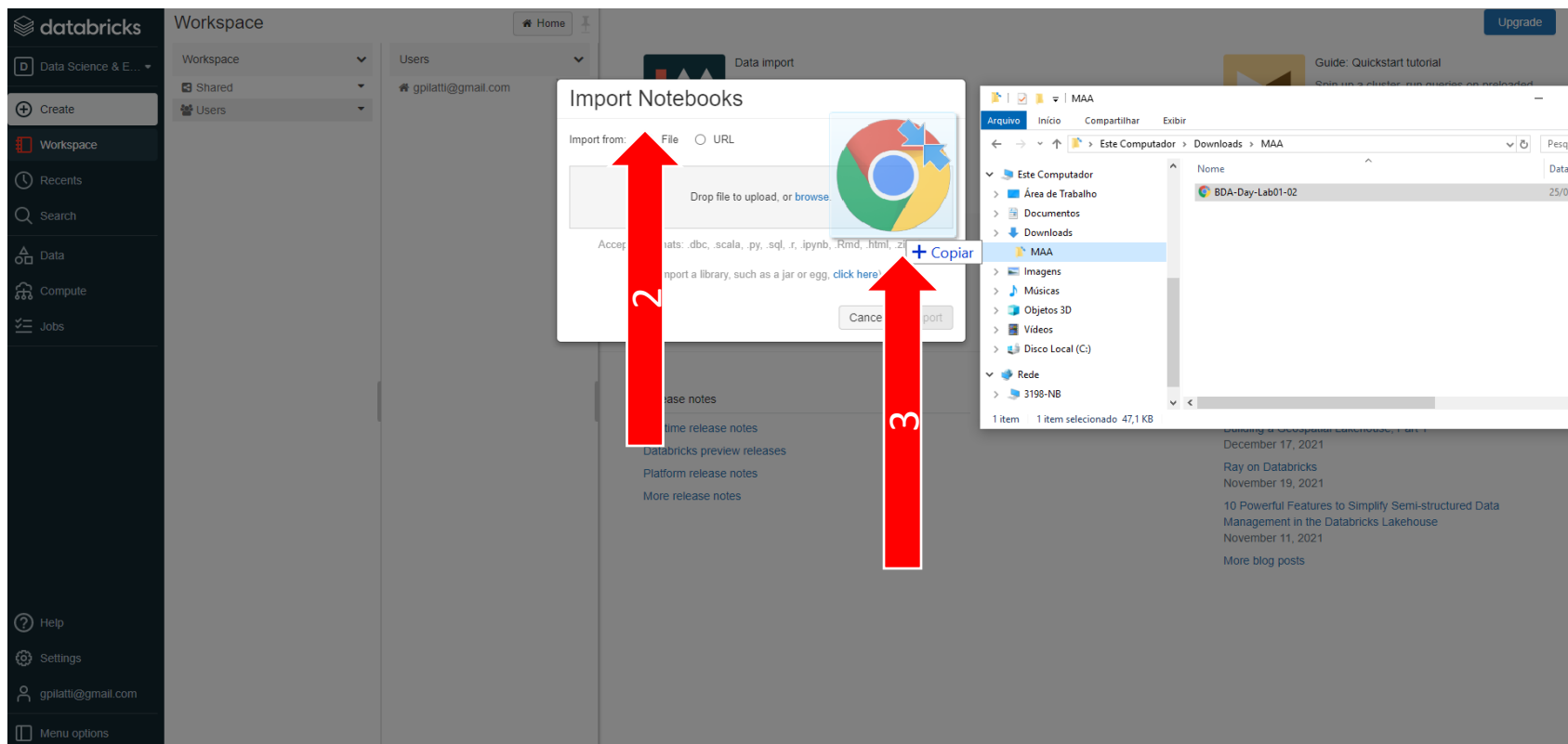
The screenshot shows the Databricks Community Edition interface. Four red arrows indicate the steps to import data:

- Arrow 1 points to the 'Workspace' button in the left sidebar.
- Arrow 2 points to the 'Users' dropdown menu in the top navigation bar.
- Arrow 3 points to the small inverted triangle next to the email address 'gpilatti@gmail.com' in the top navigation bar.
- Arrow 4 points to the 'Import' option in the dropdown menu that appears after clicking the triangle.

The interface also shows a 'Data import' section with a 'Quickstart tutorial' link and a 'Release notes' section with links to 'Runtime release notes', 'Databricks preview releases', 'Platform release notes', and 'More release notes'. There is also a 'Blog posts' section with links to 'Building a Geospatial Lakehouse, Part 1', 'Ray on Databricks', and '10 Powerful Features to Simplify Semi-structured Data Management in the Databricks Lakehouse'.

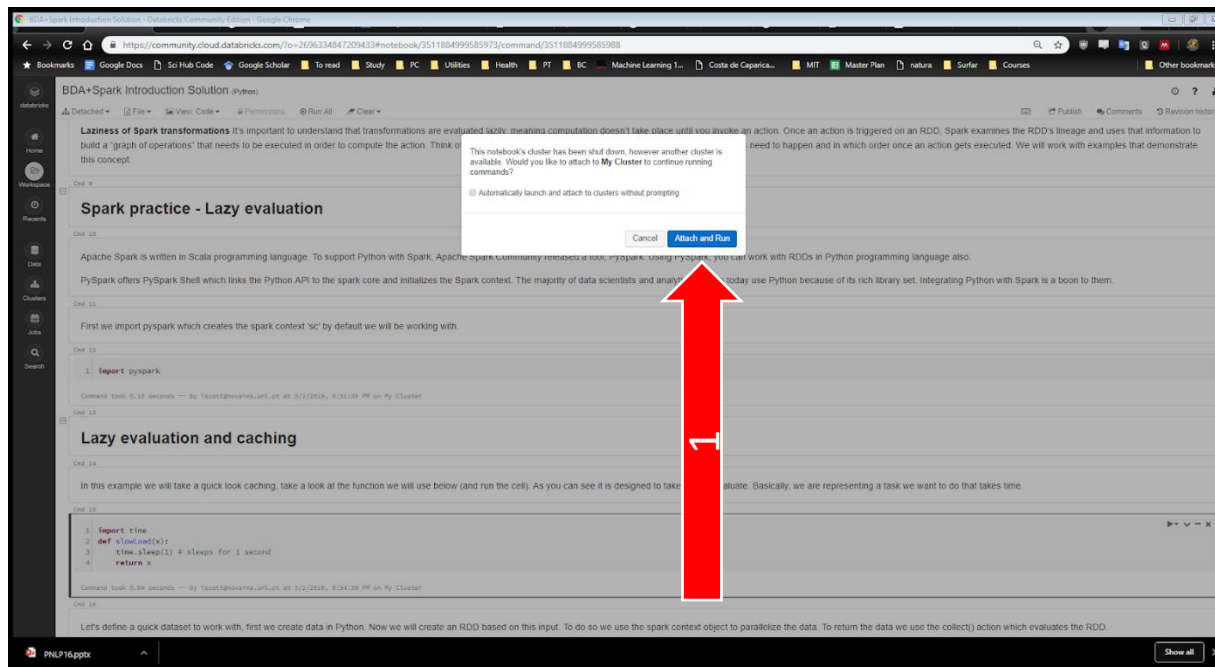
1. In Databricks press 'Workspace' on the left.
2. Click on 'Users'
3. Click on the small inverted triangle to the right of your email address
4. Select import

Databricks Community Edition



1. Click on the link to this week exercises on *Moodle* and download the notebook.
2. Choose Import from: File
3. Drag and drop the file you downloaded from *Moodle*
4. Click on Import

Databricks Community Edition



You should have a **notebook** filled with **exercises** to work with (the notebook will look slightly different to a Jupyter notebook in case that is what you are used to, you can still use Shift-Enter to run a cell).

You can now try to run a cell, you will see that you will be asked to **attach a cluster**, simply click "select resource" and "create resource".

Databricks Community Edition

merged (Python)

My Cluster

g Data Analytics - Lab Week 1

Python exercises are adapted from the course:

<https://github.com/rajathkumarp/Python-Lectures> where you can find more exercises if you wish to review, and from <https://www.practicepython.org/> a great resource for simple exercises and solutions.

Quick Python review

This week we still with reviewing how to work with python with a focus on the elements we will need when using PySpark. If you haven't used a notebook like this before you can type code into the cells below and then execute it. Typically, I will provide sc to just run the add excercises for you to apply your understanding and solve problems.

Variables

A name that is used to denote something or a value is called a variable. In python, variables can be declared and values can be assigned to it as follows, Run the cell below by clicking on it and pressing Shift and Enter together or selecting run (the little g) the cell.

When we run a cell for the first time in databricks it will need to setup an online computing cluster for us to work with, this can take 1-2 minutes, so just be patient. After the cluster is running cells should execute quickly.

If you leave your notebook for a while it may **detach** from the **cluster**, giving you an error when you try to run cells. In this case you just need to reselect your cluster at the top left and the notebook will reattach to it.

Goals for the class

1. ~~Creation of a Databricks account~~

2. ~~Introduction to Databricks~~

- ~~○ Overview of Databricks and its features~~
- ~~○ Comparison of Databricks and Jupyter Notebook~~
- ~~○ Benefits of using Databricks~~

3. ~~Uploading a File to Databricks~~

4. Introduction to Distributed Processing

- Databricks Cluster

5. Introduction to Shell Commands

- Overview of shell commands and their importance in Big Data processing
- How to use shell commands in Databricks

Databricks Community Edition - Cluster

Clusters / My Cluster

My Cluster [Edit] [Clone] [Restart] [Terminate] [Delete]

Configuration Notebooks (1) Libraries Event log Spark UI Driver Logs Metrics Apps Spark cluster UI - Master

Databricks Runtime Version

9.1 LTS (includes Apache Spark 3.1.2, Scala 2.12)

Driver type

Community Optimized 15.3 GB Memory, 2 Cores, 1 DBU

Instance


Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please upgrade your Databricks subscription.

Instances Spark JDBC/ODBC Permissions

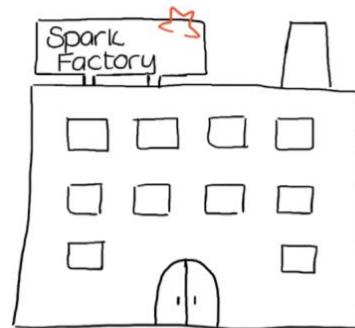
Availability zone

us-west-2c

APACHE **Spark**

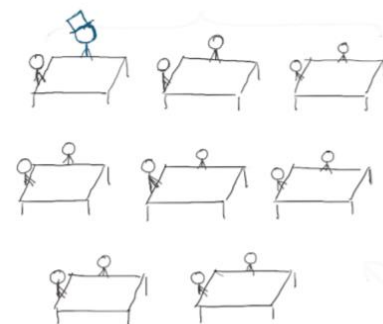


Outside view



Cohesive unit

Under the hood



- **Set of computation resources** that **perform the data workloads** ran in Databricks (commands in notebooks, commands run from BI tools, ETLs, and so on).
- Consists of **multiple nodes** (individual machines) that operate on the workloads in parallel.
- There is **one driver node for every cluster**, which is the one that delegates tasks and oversees the execution of the specific workload.
- Spark uses **RAM** to **store data in memory** for fast processing, but it also uses **distributed memory** to **scale out** across multiple nodes in a cluster.
- Spark's in-memory computing model allows for **faster processing** of **large datasets**, but it also requires careful management of memory resources to avoid running out of memory.

Goals for the class

1. ~~Creation of a Databricks account~~

2. ~~Introduction to Databricks~~

- ~~○ Overview of Databricks and its features~~
- ~~○ Comparison of Databricks and Jupyter Notebook~~
- ~~○ Benefits of using Databricks~~

3. ~~Uploading a File to Databricks~~

4. ~~Introduction to Distributed Processing~~

- ~~○ Databricks Cluster~~

5. Introduction to Shell Commands

- Overview of shell commands and their importance in Big Data processing
- How to use shell commands in Databricks

Files, Shell, Bash, Scripting

One of the **original ways to use computers** to process data was developed for the Unix Environment and followed its philosophy. Shortly summarized as:

- Write programs to do one thing and do it well.
- Write programs to work together.
- Write programs to handle text streams because that is a universal interface.

Most of the commands developed within the Unix ecosystem, more than 30 years ago, are still relevant today

Why start this way?

- Ease of execution of commands (no need to copy and paste every time)
- Powerful programming constructs

MORE INFO:

[BDA 2023 Lab1 - Bash commands intro.pdf](#)

Files, Shell, Bash, Scripting

The screenshot shows the Databricks Clusters interface. On the left sidebar, a red circle with the number '1' highlights the 'Apps' icon. At the top, a red circle with the number '2' highlights the 'Apps' tab in the navigation bar. Below the 'New cluster' button, a red circle with the number '3' highlights the 'Launch Web Terminal' button. The 'Web Terminal' section is visible, along with the 'RStudio Server' section.

If the Web Terminal is not enabled, follow these steps in the next page.

Files, Shell, Bash, Scripting

databricks malmeida@novaims.unl.pt

Compute

All-purpose compute Job compute

Filter compute you have access to Created by

Create compute

State	Name	Runtime	Active mem...	Active cores	Active DBU / h	Source	Creator	Notebooks
✓	test1	12.2	15 GB	2 cores	1	UI	malmeida@novaims....	-

Advanced

- > Third-party iFraming prevention Enabled ☒
- > Download button for notebook results Enabled ☒
- > Upload data using the UI Enabled ☒
- > Notebook Exporting Enabled ☒
- > Notebook Table Clipboard Features Enabled ☒
- > Web Terminal Not enabled ☐
- > DBFS File Browser Not enabled ☐

User Settings

Admin Settings

Delete Account

Log out

Notebooks

1. Go to the Settings
2. Click on Admin Settings.
3. Click the Workspace Settings tab.
4. In the Advanced section, click the Web Terminal toggle.
5. Refresh the page.

Shell commands

- Shell is an environment in which we can run our commands, programs, and shell scripts.
- There are different flavors of a shell, just as there are different flavors of operating systems.
- Each flavor of shell has its own set of recognized commands and functions.
- In this course we will use **Bash** ("Bourne Again Shell") shell as the main shell interpreter.

Tip: If the shell is running on your local filesystem:

- You can press the up arrow to cycle through previous commands
- When using windows, you can right-click to paste (instead of ctrl-v).

Shell commands

Practice and more information:

BDA 2024 Lab 1 – Bash commands intro

Python review (will not be covered in this class)

BDA-Lab1-p1 - Python review.html

BDA-Lab1-p2- Python review.html

End