THE DATA SCIENCE LAB
- Introduction -

COM 490 – Module 1a

Week 1

Week 1 - Agenda

- Introduction to the class
- Set up your lab environment

Meet the team

Sofiane Sarni
SDSC
Module 4

Pamela Delgado
SDSC
Module 3

Eric Bouillet
SDSC
Module 1
Module 2

Dongqing Wang
Doctoral Assistant

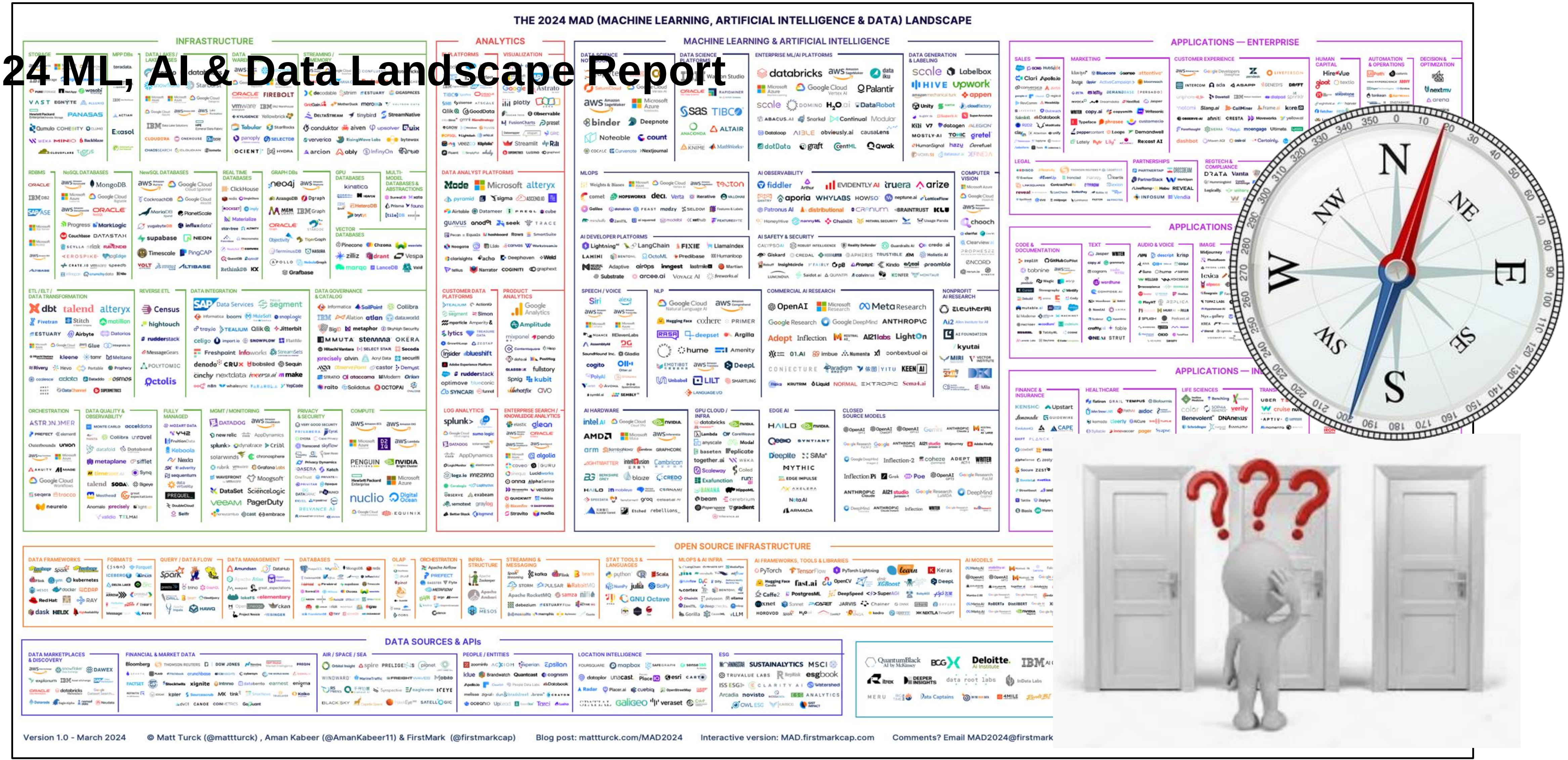
Daichi Kuroda
Doctoral Assistant

Yulun Jiang
Doctoral Assistant

Ziyi Zhang
Doctoral Assistant

What this lab is about?

2024 ML, AI & Data Landscape Report


THE 2024 MAD (MACHINE LEARNING, ARTIFICIAL INTELLIGENCE & DATA) LANDSCAPE
INFRASTRUCTURE
ANALYTICS
MACHINE LEARNING & ARTIFICIAL INTELLIGENCE
APPLICATIONS — ENTERPRISE
APPLICATIONS
APPLICATIONS — IN
OPEN SOURCE INFRASTRUCTURE
DATA SOURCES & APIs
Version 1.0 - March 2024
© Matt Turck (@mattturck) , Aman Kabeer (@AmanKabeer11) & FirstMark (@firstmarkcap)
Blog post: mattturck.com/MAD2024
Interactive version: MAD.firstmarkcap.com
Comments? Email MAD2024@firstmark
N
NE
E
SE
S
SW
W
NW

https://mattturck.com/landscape/mad2024.pdf

Lab Overview

- A journey through a real-world data science project
- Very hand-on and pragmatic

- **4 Modules**
  - Module 1 – Review of Data Science with Python
  - Module 2 – Big data wrangling and query
  - Module 3 – Big data processing & Machine Learning with Apache Spark
  - Module 4 – Real time data acquisition and processing

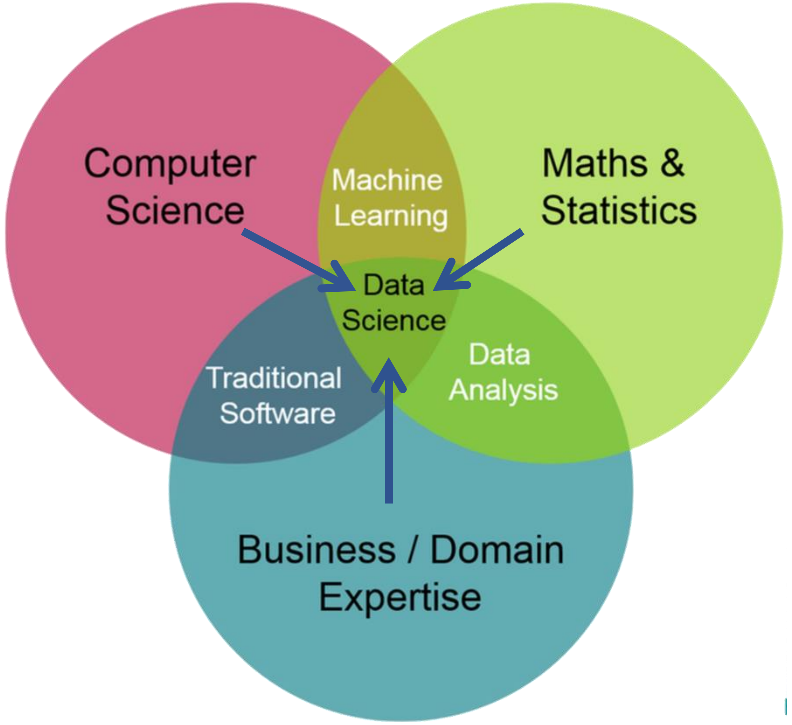
Agenda 2025 - Module 1a

19.02	Introduction to Data Science with Python
26.02	(Bigger) Data Science with Python
05.03	Introduction to Big Data Technologies
12.03	Big Data Wrangling with Hadoop
19.03	Advanced Big Data Queries
26.03	Introduction to Spark
02.04	Spark Data Frames
09.04	Advanced Spark
16.04	Introduction to Stream Processing
30.04	Stream Processing with Kafka
07.05	Advanced Stream Processing
14.06	Final Project Q&A
22.05	Final Project Videos Due before midnight
28.05	Oral Sessions

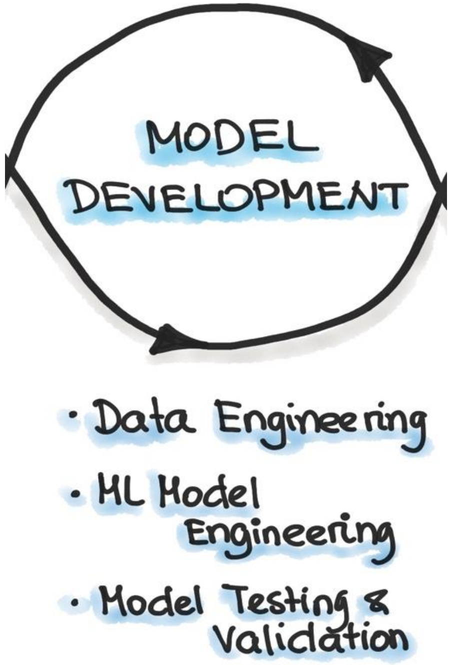
Lab Overview

- 50% (Big) Data/Feature Engineering
- 30% (Big) Data Science
- 20% Build foundations for ML-Ops

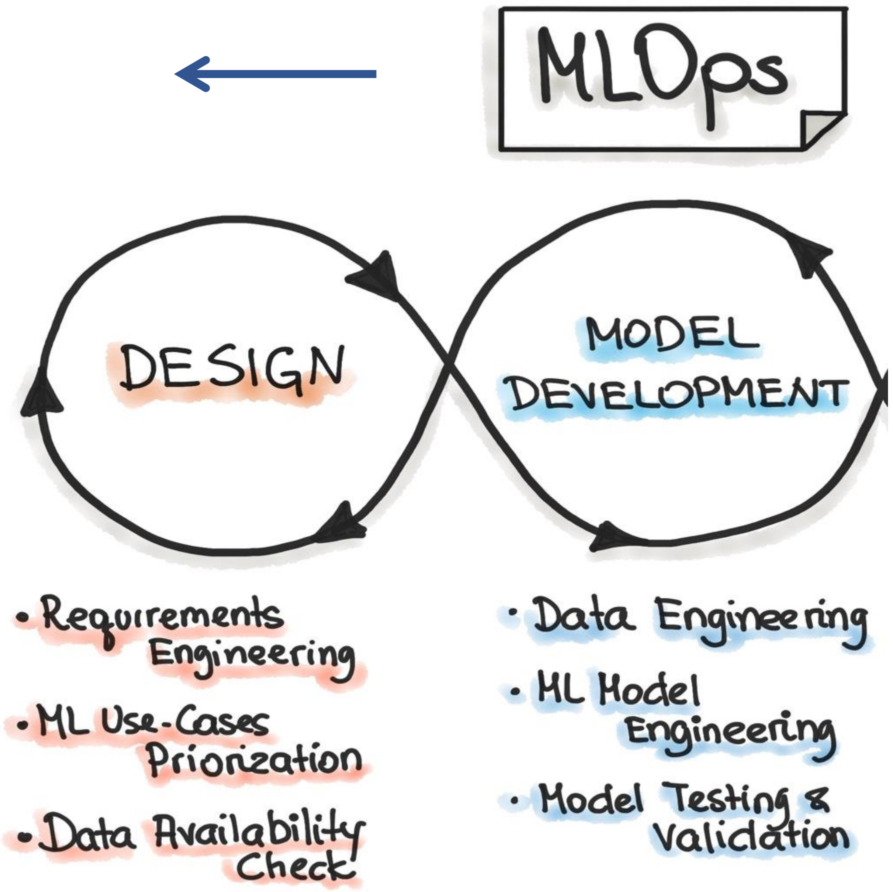
Drew Conway's Venn Diagram



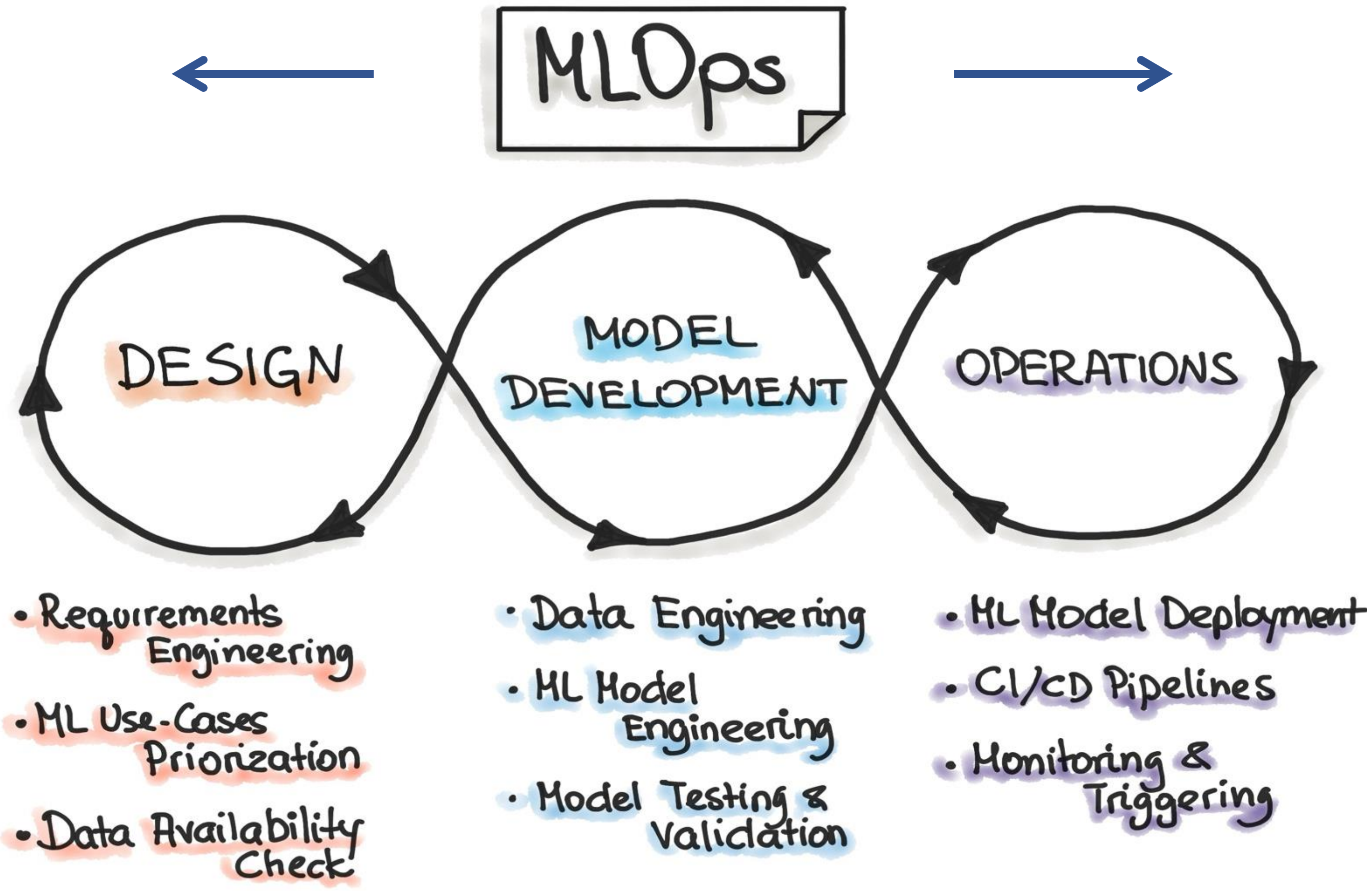
Lab Overview



Lab Overview


MLOps
DESIGN
MODEL DEVELOPMENT
• Requirements Engineering
• ML Use-Cases Priorization
• Data Availability Check
• Data Engineering
• ML Model Engineering
• Model Testing & Validation

Lab Overview


MLOps
DESIGN
MODEL DEVELOPMENT
OPERATIONS
• Requirements Engineering
• ML Use-Cases Priorization
• Data Availability Check
• Data Engineering
• ML Model Engineering
• Model Testing & Validation
• ML Model Deployment
• CI/CD Pipelines
• Monitoring & Triggering

Class Format

- **Labs on Wednesday – 13h10 to 16h00**
  - Theory and general introduction to exercises
  - Exercise sessions of 30min to 40min each, and 10min recap between sessions
  - Classes are recorded (Zoom*), and videos are made available after the class

- **Office hours**
  - Interactive communication via Ed forum(*)
  - Outside class hours on demand - time to be adapted according to students’ schedule

*Details on Moodle

Communication

- **Moodle**
  - https://moodle.epfl.ch/course/view.php?id=15635
  - Class materials (slides), form groups, oral schedule, and other useful links

- **Ed (*)**
  - For real-time intra/inter group communication, and to reach us outside class hours
  - Channels:

General	For our general announcements or to forward EPFL guidelines
Labs	Discussions related to the lectures and labs
Assignments	Channel for each assignment (A1, ...), and one for the final
Social	Looking for a team, or a team-mate ?

  - Etiquette:
    - **DO** Answer questions in a comment under a thread
    - **DO** Help each other with technical issues etc.,
    - **DO NOT** provide solutions to assignment

*Details on Moodle

Lab Assessment

- 40% Final project
  - Collaborative project, in groups of ~5 students
  - Due before final week of semester
    - 6-7min video presentation
    - Code
  - Mini oral presentation (group) during the final week

- 60% Continuous assessment
  - One take-home assignment per module 1 to 3
  - To complete in groups, within 3 weeks each
  - Assignments are related to the final project

Lab Assessment – Important Dates

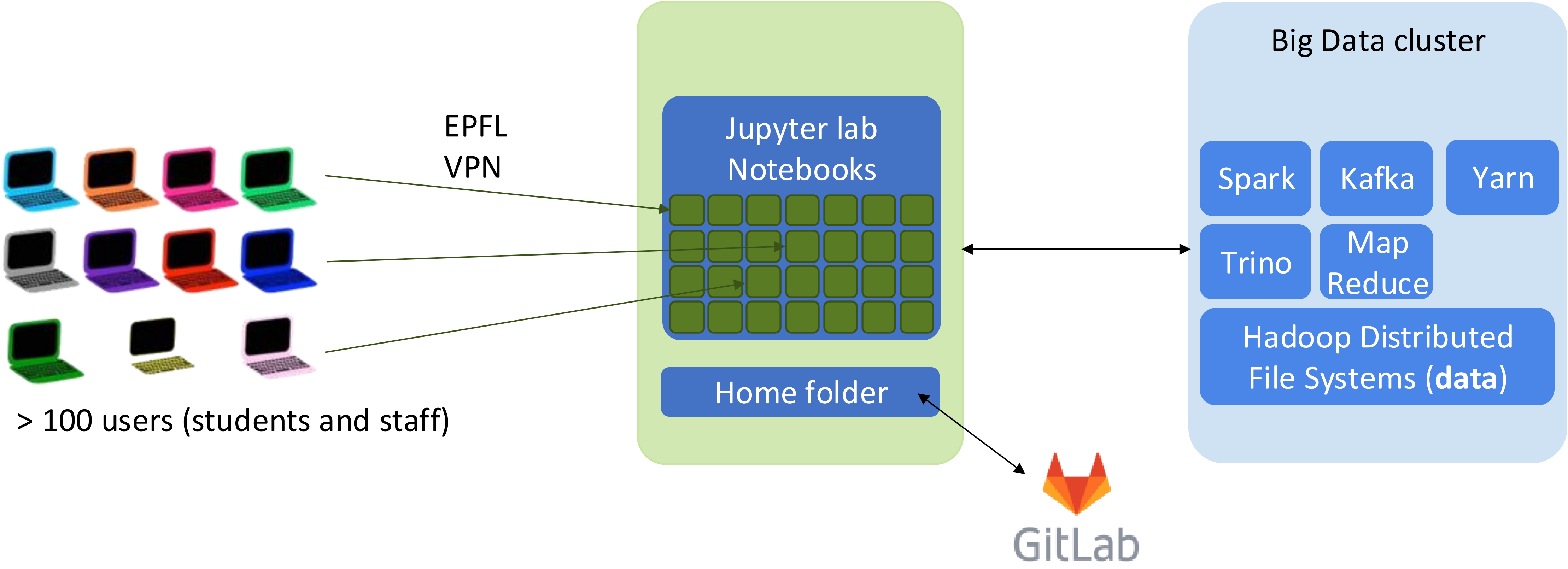
Tuesday	18.03	– Assessment 1 is due by midnight
Tuesday	08.04	– Assessment 2 is due by midnight
Tuesday	06.05	– Assessment 3 is due by midnight
Thursday	22.05	– Short video final assessment is due by midnight
Wednesday	28.05	– Oral sessions (10mins per group)
Friday	30.05	– Final assessment is due by midnight

Programming Languages

- **Programming Languages**
  - Mainly Python
    - Numpy, pandas, scikit-learn, matplotlib, PySpark, ...
  - Also SQL(-like)
  - And a pinch of Linux Shell command lines

- **Developer tools**
  - Git (gitlab)
  - Hadoop big data stack command lines (hdfs, yarn, ...)
  - Jupyter notebooks

Programming Environment



1 **BYOL**: Students work remotely using their laptops. Nothing to install – only web browser is needed.

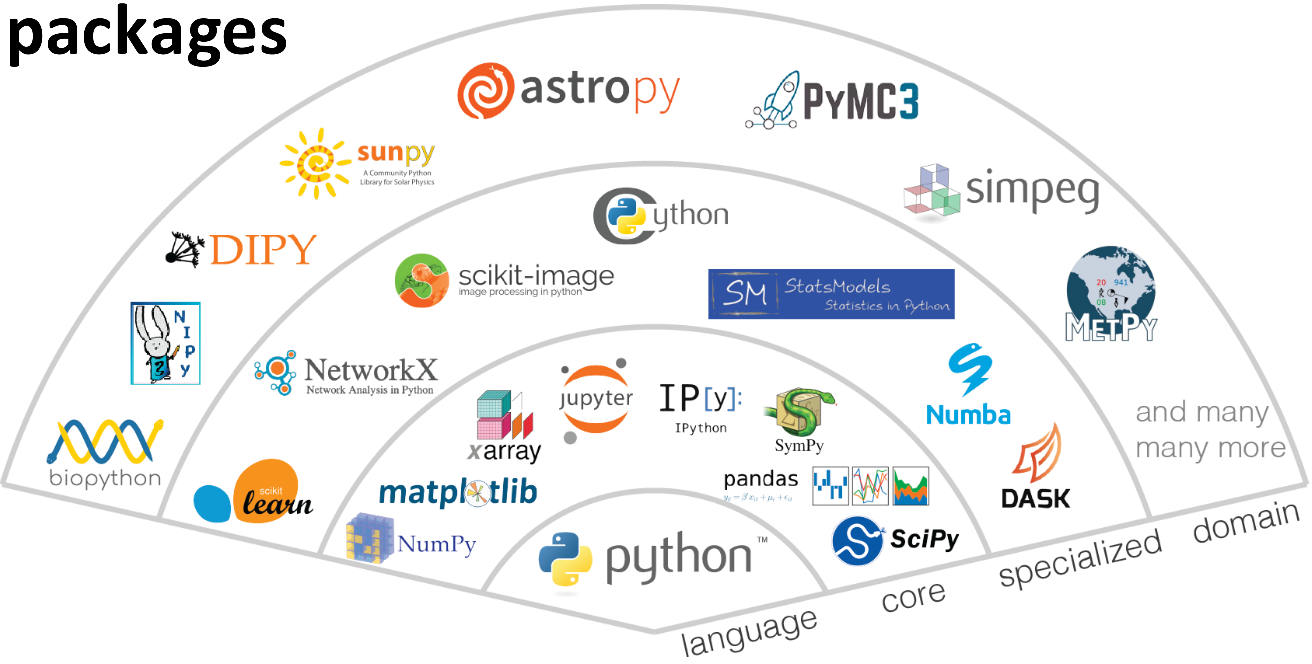
2 Students work in teams, write and share code and environment in jupyter notebooks and gitlab

3 All data stored, and compute intensive processing executed on the distributed Big Data cluster.

Gentle Introduction to Data Science With Python

Python Data Science Ecosystem

- **Python**
  - Core programming language used in the class
- **Python Math & Data Science packages**
  - Numpy
  - Pandas
  - Scikit-Learn
  - …



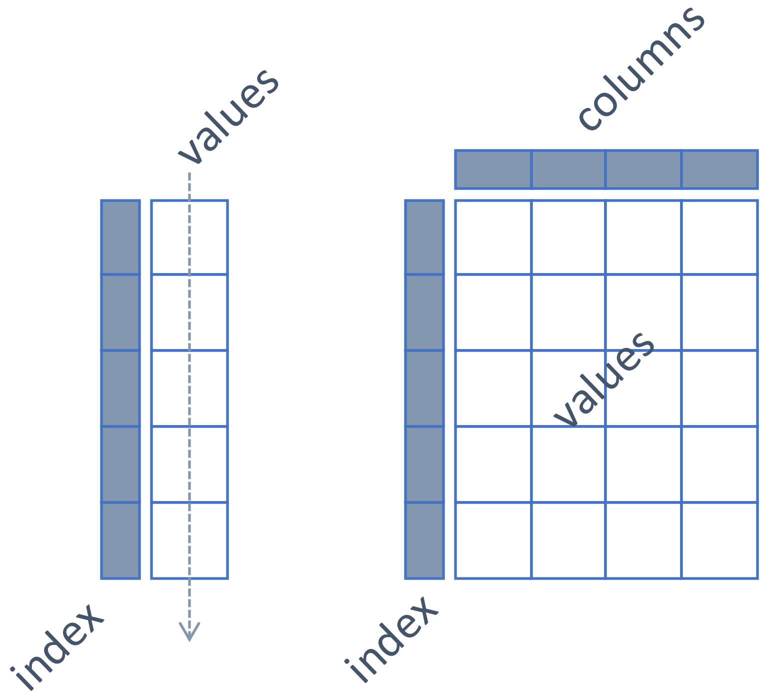
and many more …

Python Data Science Ecosystem

- **Numpy**
  - Core library for scientific computing in Python
  - Provides a high-performance multidimensional array object, <N>-D
  - Large collection of high-level mathematical functions to operate on arrays objects
  - Optimized for size and performance
- **SciPy**
  - Built on NumPy
  - Mathematical library for Scientific and Technical Computing
    - Integration, linear optimization, spatial, stats, FFT, ...

Python Data Science Ecosystem

- **Pandas**
  - 1D or 2D structures
  - Built on top of NumPy
    - NumPy stores your data in arrays
    - Pandas takes the arrays, …
      … and gives you labelled index to it
    - Basically dictionary based NumPy *ndarray*
  - Powerful & flexible data munging library
  - Recommended reading: [pandas documentation]



1-D
(Series)

2-D
(DataFrame)

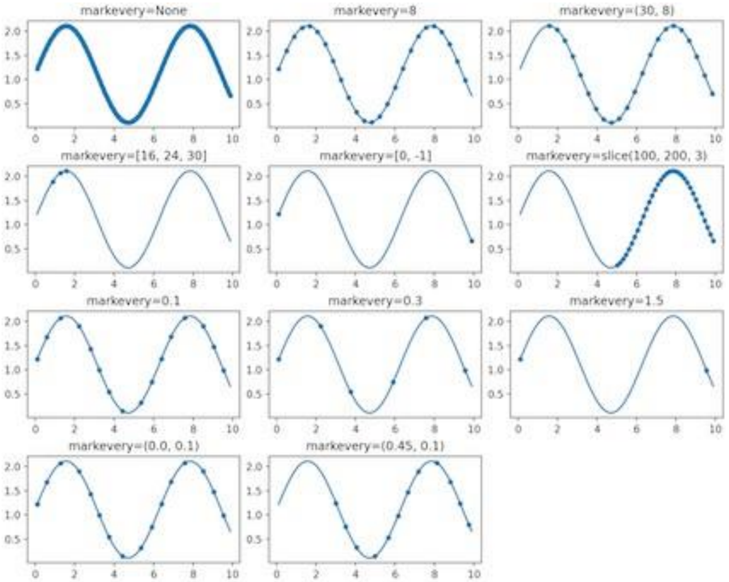
Python Data Science Ecosystem

- **Scikit-learn - Machine Learning in Python**
  - Model algorithms (Classification, Regression, Clustering, NN, ...)
  - Performance metrics
  - Model hyper-paremeter tunings
  - Model Training, Validation
  - Feature selection
  - Data Processing, Pipelines
  - ...
- **PyTorch, TensorFlow**
  - AI, Deep Learning
  - GPU-based optimization
  - ...

Python Data Science Ecosystem

- **Matplotlib**
  - The library for creating visualizations in Python
  - Pandas’ default visualization engine

    `pandas.DataFrame.plot()`
  - Powerful, but low level programming interface
  - Best for quick and basic data exploration
- **Alternatives**
  - Plotly
  - Seaborn, folium, bokeh, osmnx, vispy, pygal, cufflinks, ...



Today's check list – key objectives

- **You have access to EPFL network (VPN)**
  - Otherwise: → https://vpn.epfl.ch
- **You have registered for the class on IS-Academia**
  - Otherwise: → http://is-academia.epfl.ch
- **You have access to our Moodle page and have bookmarked it**
  - https://moodle.epfl.ch/course/view.php?id=15635
  - Contact us to add you to the list
- **You have access to our programming environment (JupyterHub)**
  - You can login to your assigned jupyter notebook with your usual EPFL (gaspar) username and password
- **You have access to the exercises of module 1a**
  - You can login and access https://dslabgit.datascience.ch/course/2025/module-1a
- **You master the ABCs of building and validating a predictive model with Scikit-learn**

Start your engines

Bootstrapping into Jupyter notebooks

Jupyter Hub – Login

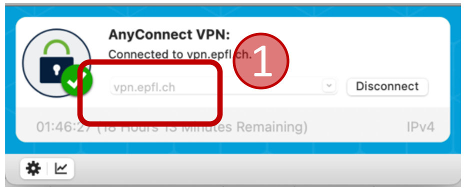
1. Must be on EPFL network (VPN)
2. Sign in https://groups.epfl.ch/ and in “My groups” search for **com-490** to find your assigned Jupyter hub server

   You should see **ic-spark-com-490-**... If not, come to us

3. Based on the above, in a browser (Firefox, Safari, Chrome), sign in with your EPFL (gaspar) username and password

ic-spark-com-490-1	iccluster082.iccluster.epfl.ch
ic-spark-com-490-2	iccluster083.iccluster.epfl.ch
ic-spark-com-490-3	iccluster084.iccluster.epfl.ch
ic-spark-com-490-4	iccluster086.iccluster.epfl.ch
ic-spark-com-490-5	iccluster087.iccluster.epfl.ch

4. Start My Server



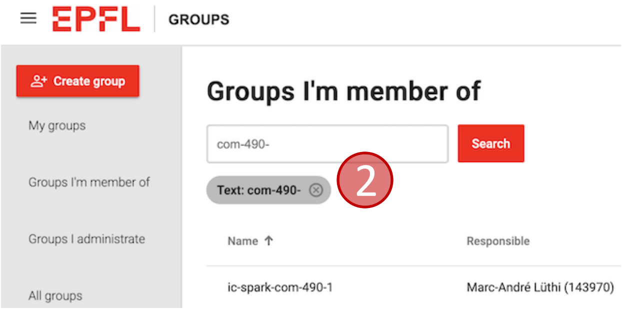
1

AnyConnect VPN:
Connected to vpn.epfl.ch.

vpn.epfl.ch

Disconnect

01:46:27 (18 Hours 13 Minutes Remaining) IPv4



EPFL GROUPS

Create group

My groups

Groups I'm member of

Groups I administrate

All groups

Groups I'm member of

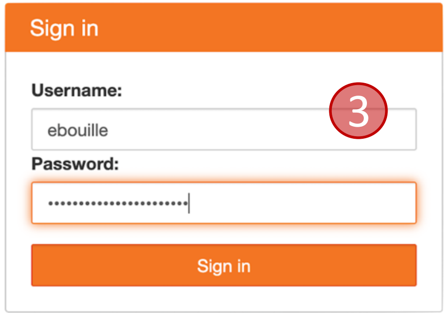
com-490-

Search

Text: com-490-

2

Name ↑	Responsible
ic-spark-com-490-1	Marc-André Lüthi (143970)



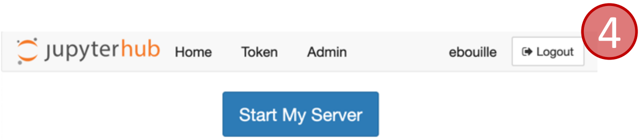
Sign in

3

Username: ebouille

Password: ••••••••••••••••••••••

Sign in



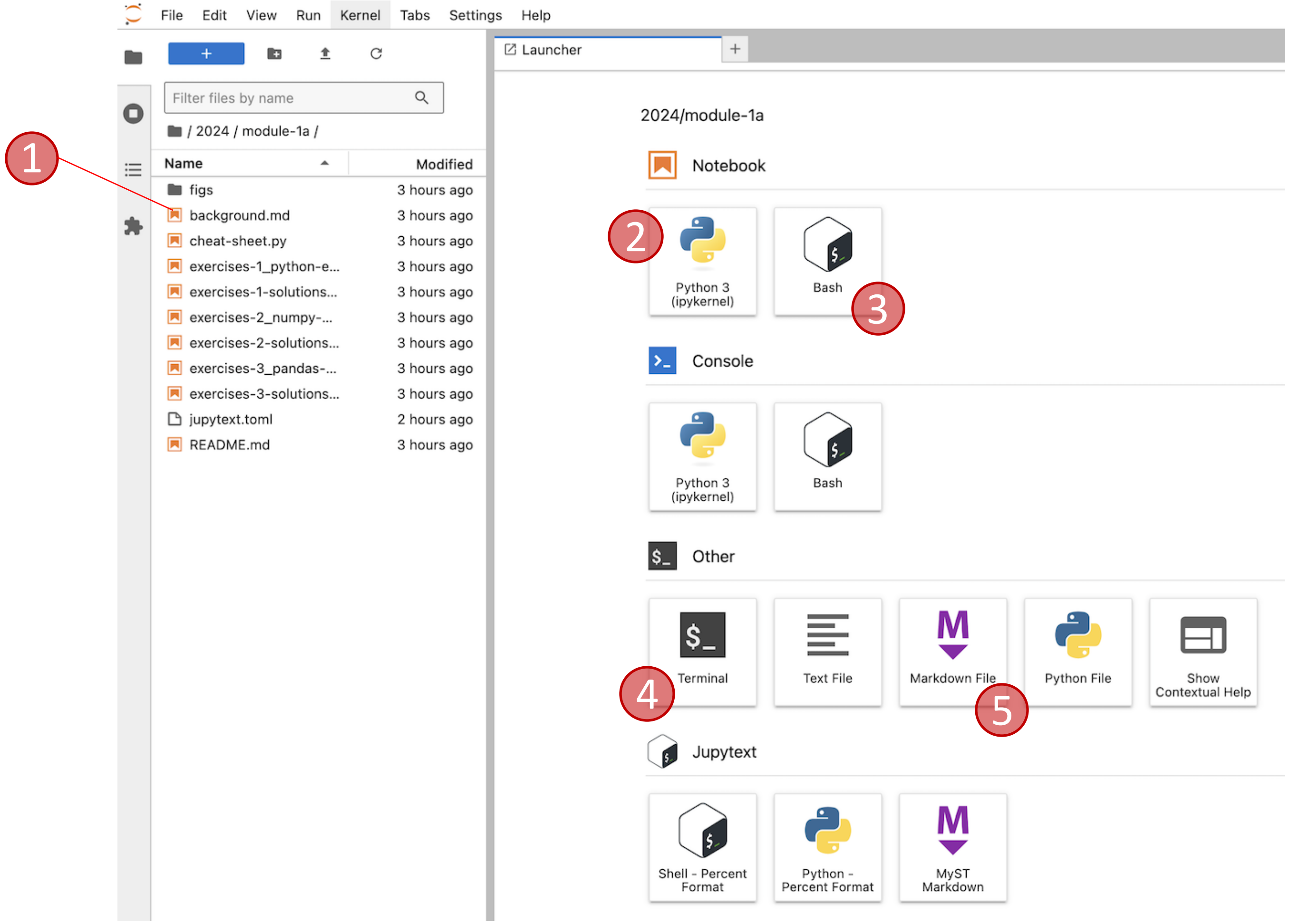
4

jupyterhub Home Token Admin ebouille Logout

Start My Server

Jupyter Lab – Interactive sessions

1. Folders and files of weekly lab
   E.g. module-1a, module-1b, ...
2. New python notebooks
3. New shell script (bash) notebooks
4. New terminal (bash/linux)
5. Markdown .md files (README, doc)



Name	Modified
figs	3 hours ago
background.md	3 hours ago
cheat-sheet.py	3 hours ago
exercises-1_python-e...	3 hours ago
exercises-1-solutions...	3 hours ago
exercises-2_numpy-...	3 hours ago
exercises-2-solutions...	3 hours ago
exercises-3_pandas-...	3 hours ago
exercises-3-solutions...	3 hours ago
jupytext.toml	2 hours ago
README.md	3 hours ago

Jupyter Lab – Exercises module 1a

1. Start a new terminal session
2. Open a terminal and in the terminal, type:

```
git clone git@dslabgit.datascience.ch:course/2025/module-1a.git
```

3. Press enter
4. You should have a new folder

```
./module-1a
```

5. If git clone does not work for you, download the file module-1a.zip from moodle in the same terminal

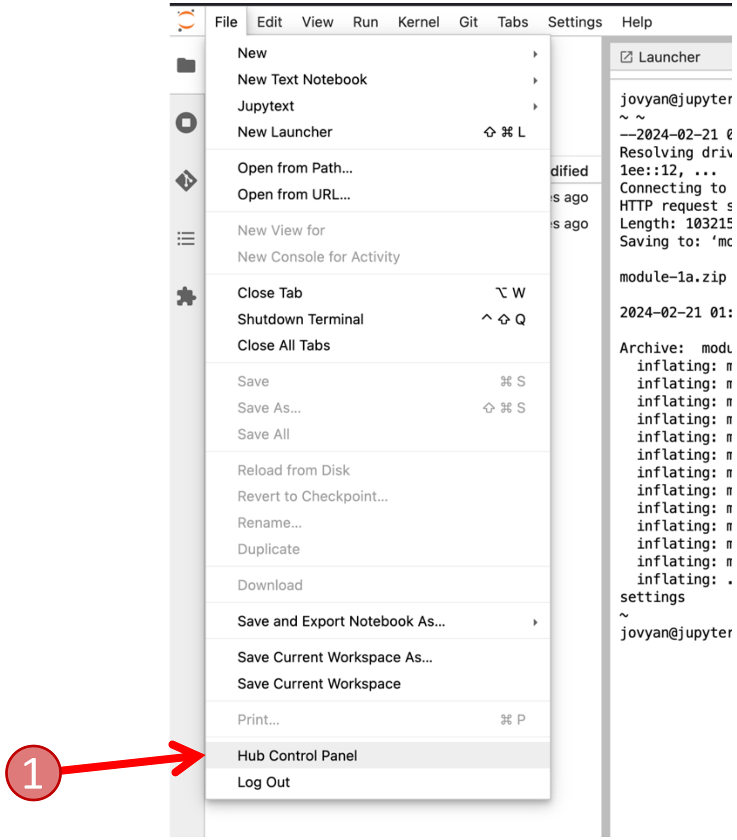
```
wget -O module-1a.zip https://drive.switch.ch/index.php/s/IWccU0aqEgbLCRk/download
unzip module-1a.zip
```

6. You should a new folder

```
./work/module-1a.zip
```

Jupyter Lab – Exercises module 1a

- If you need to restart your jupyter lab server

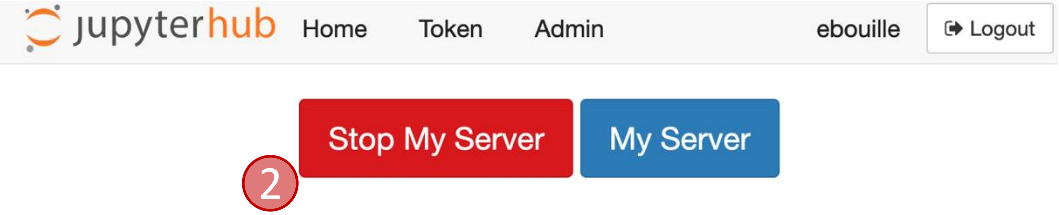


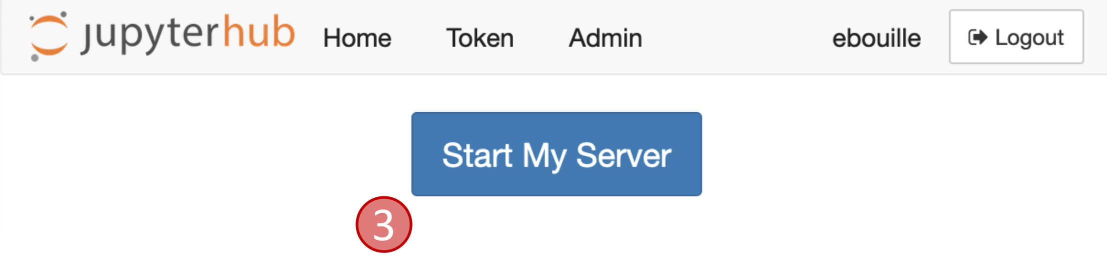
```
jovyan@jupyter
~ ~
--2024-02-21 0
Resolving driv
1ee::12, ...
Connecting to
HTTP request s
Length: 103215
Saving to: ‘mc

module-1a.zip

2024-02-21 01:

Archive:  modu
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: m
  inflating: .
settings
~
jovyan@jupyter
```


jupyterhub
Home
Token
Admin
ebouille
Logout
Stop My Server
My Server
2


jupyterhub
Home
Token
Admin
ebouille
Logout
Start My Server
3