

Smart Workflows

How to develop your application

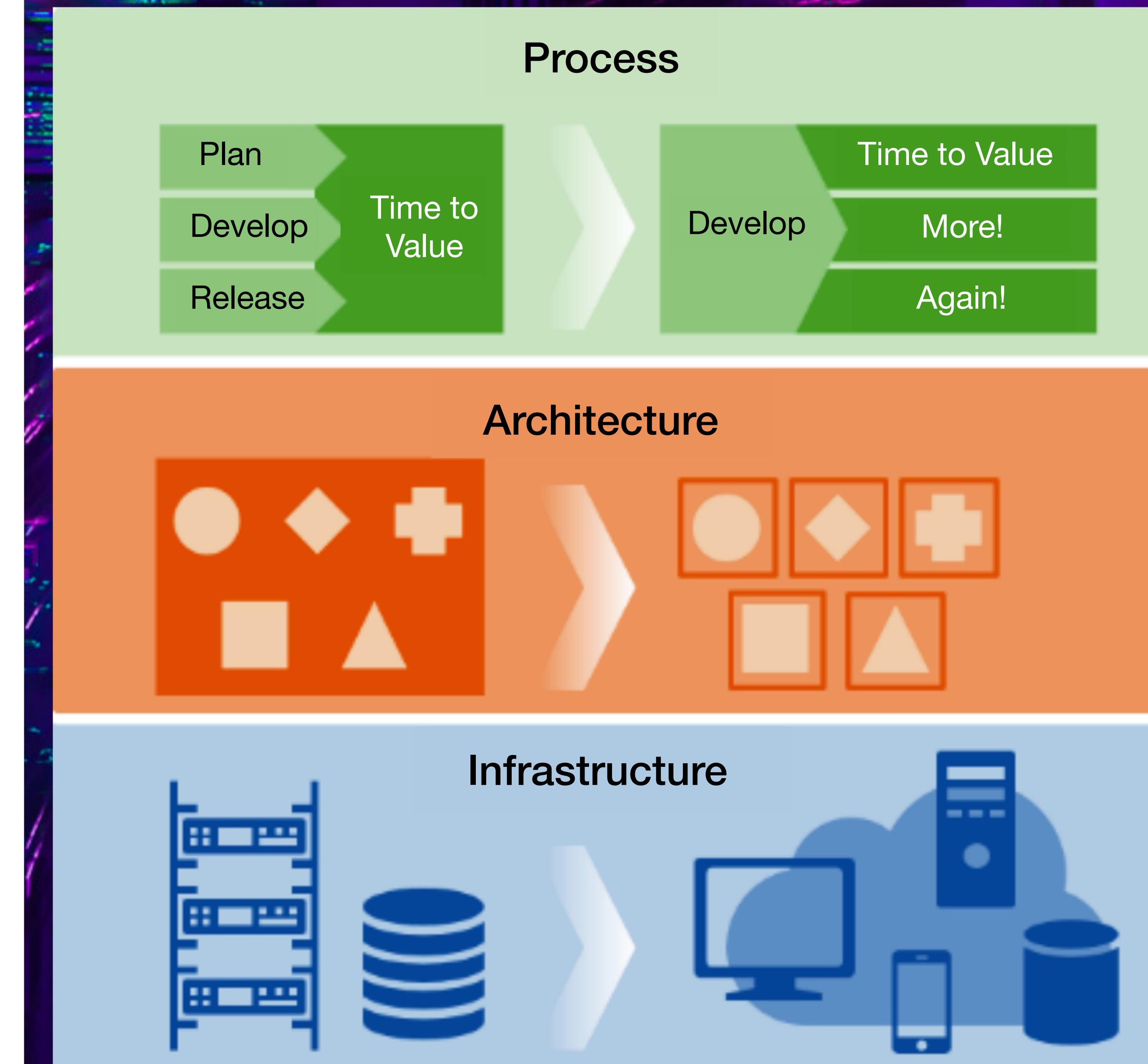


A paradigm shift

In the last decades the approach to computing has changed, not only because of new technologies, but also in its philosophy.

We often hear about:

- DevOps approach and Continuous Integration (CI)
- Microservices
- Internet of Things (IoT) and the Digital Continuum



Infrastructure

Internet of Things (IoT) and the Digital Continuum

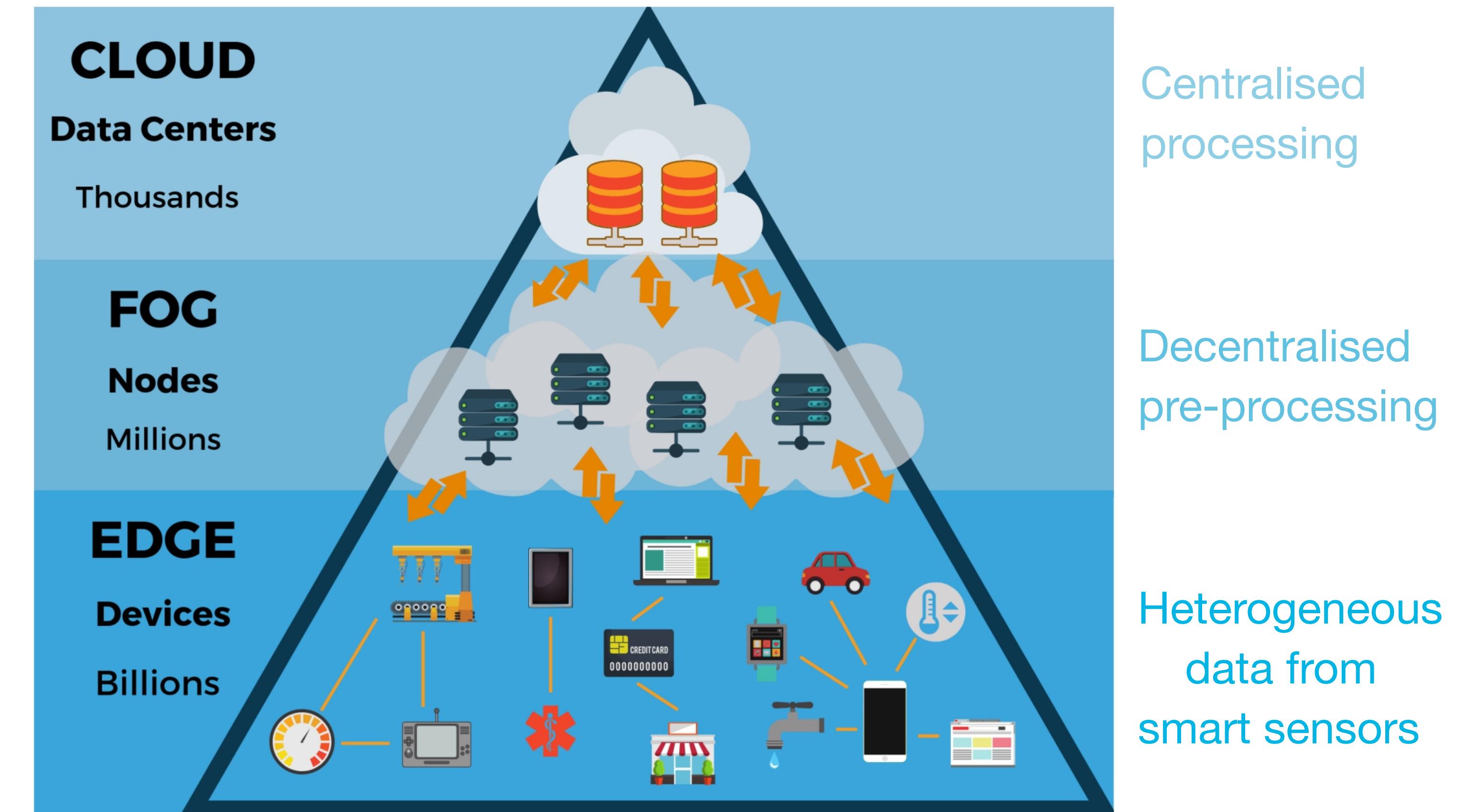
We used to have servers and storage disks...

Now, we still have servers and storage disks,
but we also talk about:

- **IoT**, where you generate, store and process data on an amount of objects:

computers, smartphones, sensors,
the washing machine...

- **Digital Continuum**: a design space that creates continuity between elements that exist separately, with the goal to transform data into information as soon as possible.



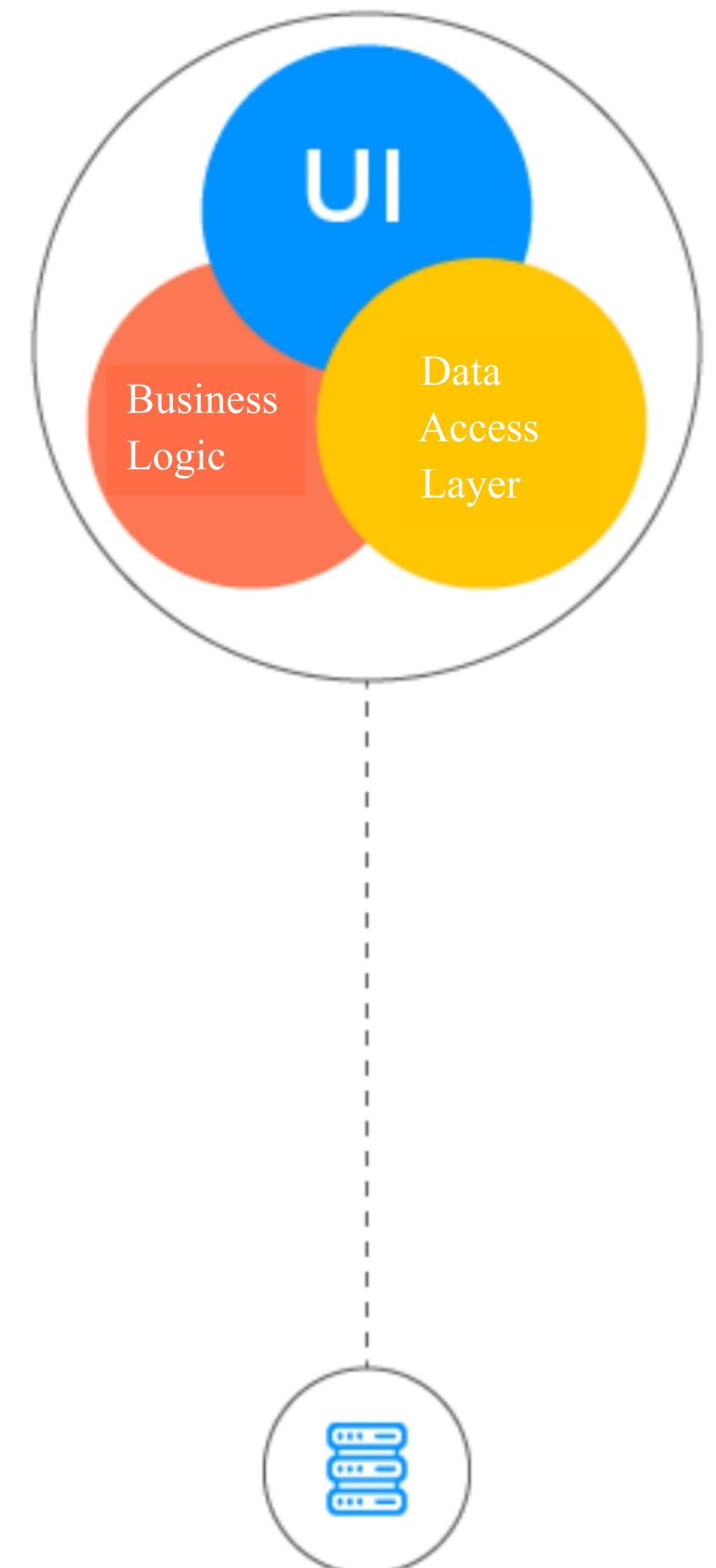
Architecture

Microservices

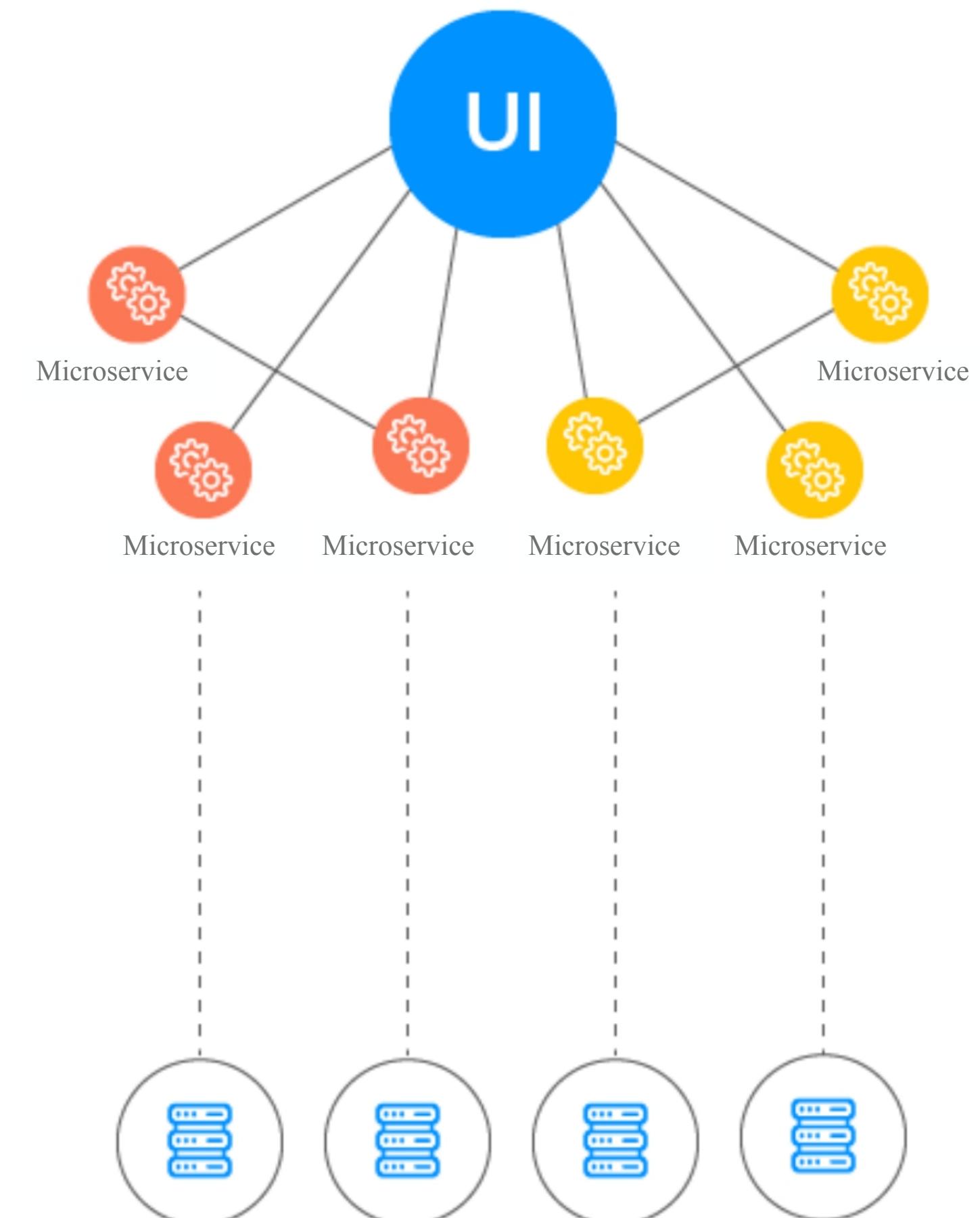
The way people structure a complex stack of computing applications has changed.

- **Monolithic architecture:** several “programs” installed on a server. Difficult to understand dependencies, do updates, re-use work done.
- **Microservices:** Service Oriented Architecture (SOA) that structures an application as a collection of loosely coupled services.
 - Makes the application easier to understand, develop and test.
 - Parallelises development.
 - Orchestration layer needed to coordinate operations.

Monolithic Architecture



Microservice Architecture



Process

Time to Value

Time to value: from the original idea to have a product available for other people to use or to produce a scientific result.

Before: I have an idea, I write code, I release.

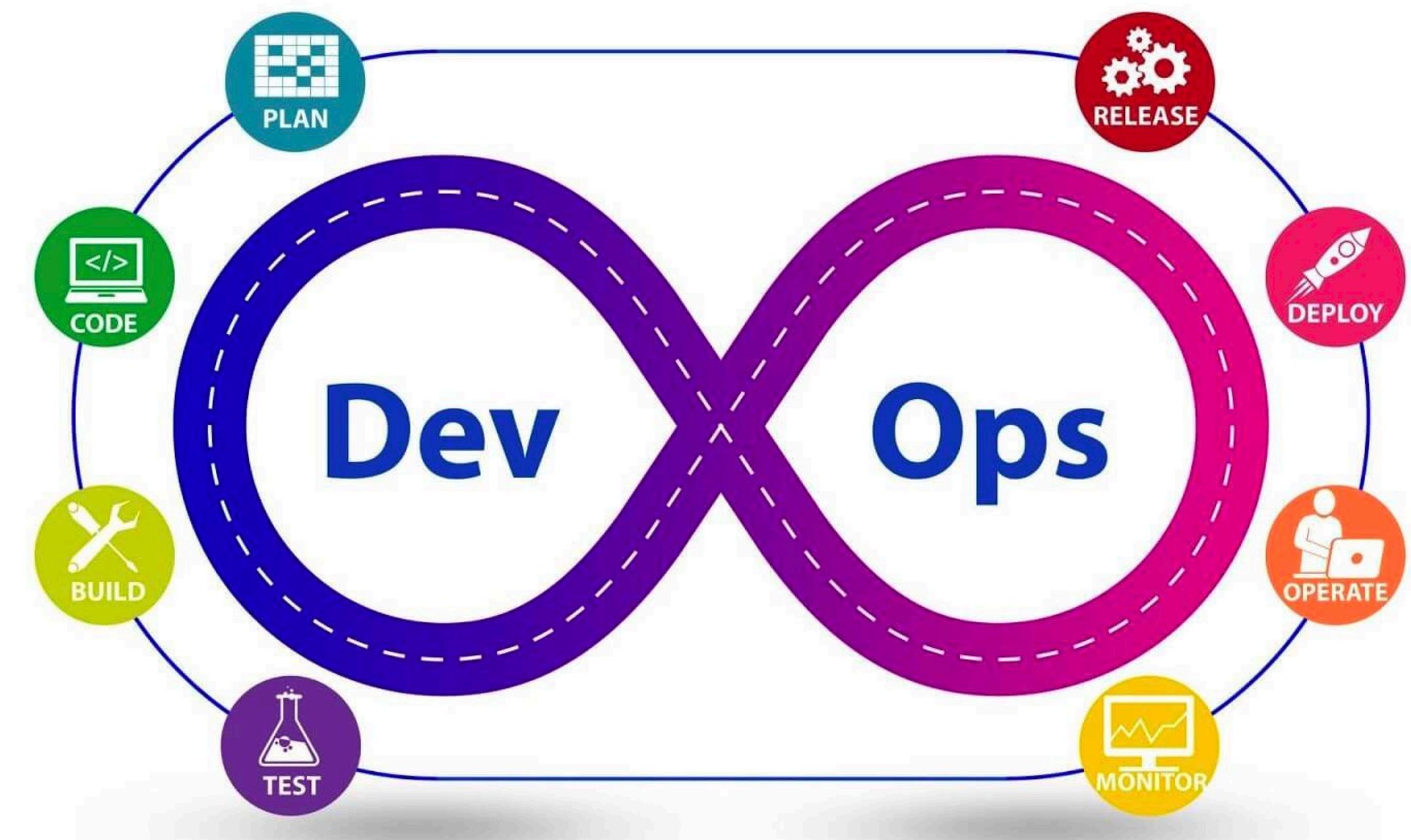
Now: keep time to value as short as possible. Make a change available whenever it is ready and tested, with the possibility to rollback.



Process

DevOps approach

- DevOps (Development and Operations) is a philosophy and practice focused on agility, collaboration, and automation within IT and development team processes.
- The goal is to **bridge the gap between IT operations and development** to improve communication and collaboration, create more seamless processes, and align strategy and objectives for faster and more efficient delivery.
- **DevOps philosophy principles:**
 - Automation
 - Iteration
 - Self-service
 - Continuous improvement
 - Collaboration
 - Continuous testing

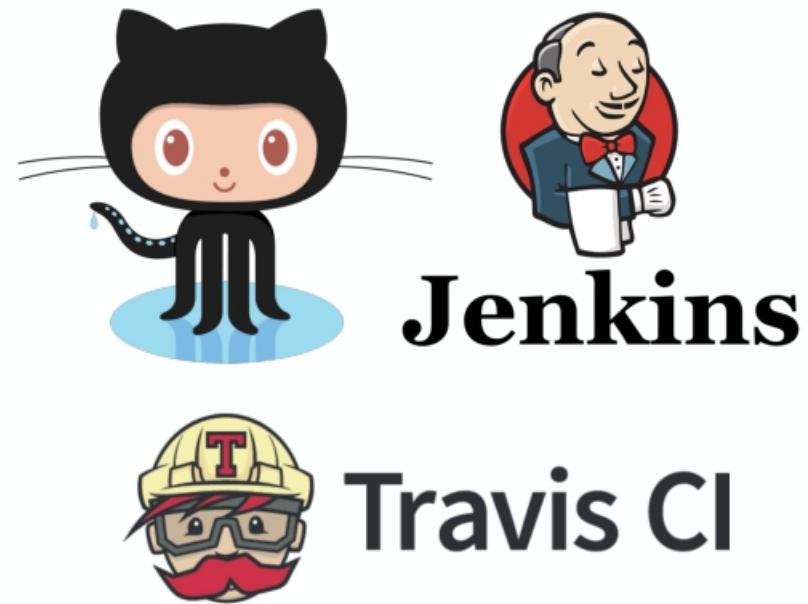


Process

Continuous Integration, Delivery and Deployment

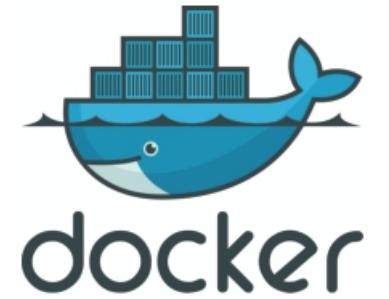
Continuous Integration (CI):

- a software development practice in which developers regularly merge their code changes into a shared repository where those updates are automatically tested
- ensures that the most up-to-date and validated code is always readily available to developers.



Continuous Delivery:

- code changes are automatically built, tested, and packaged for release into production.



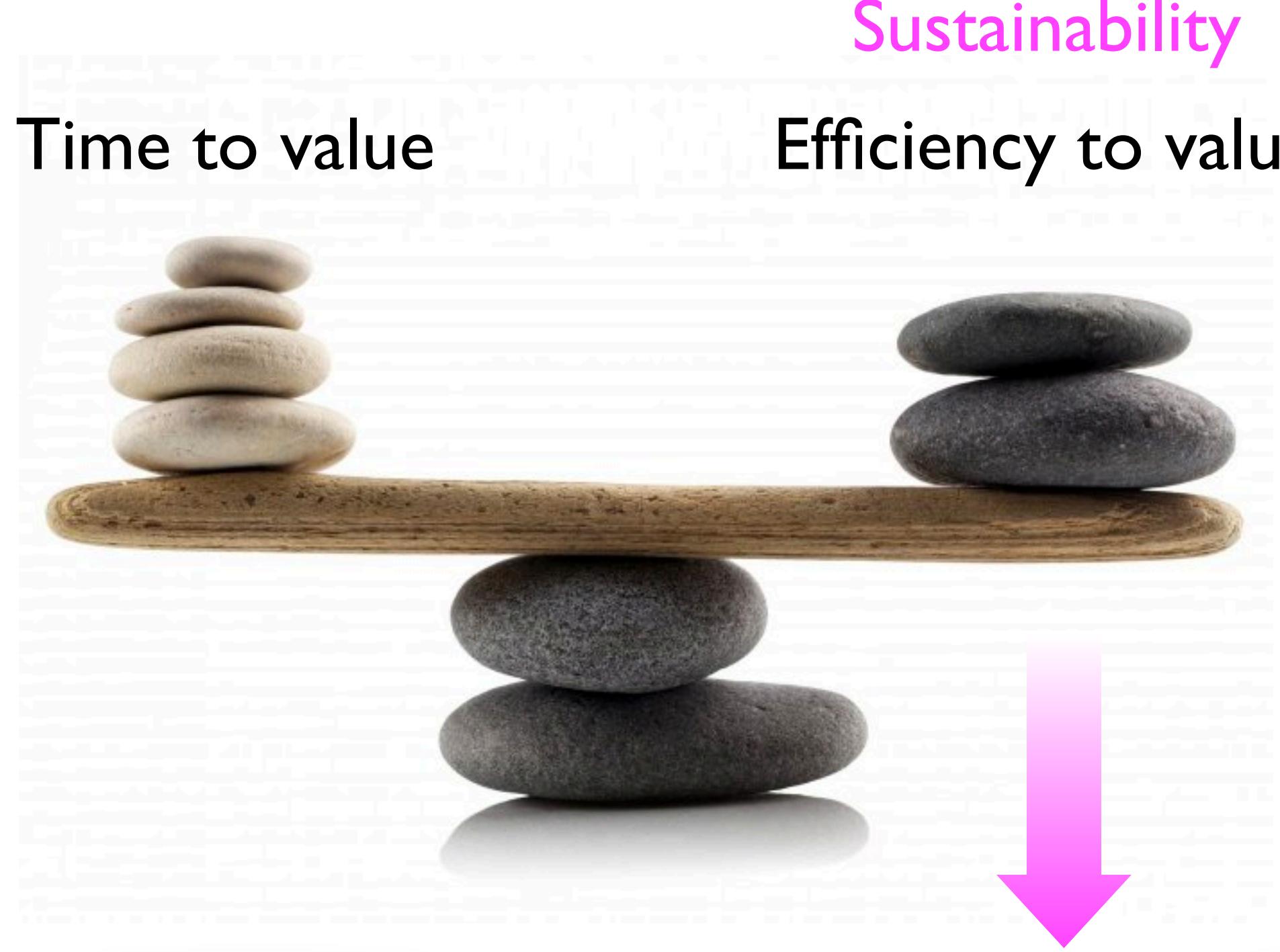
Continuous Deployment:

- every validated change is automatically released to users.



Continuous Monitoring and Feedback.





- Sustainability of the Cloud**
- Building materials
 - Location
 - Renewable energy
 - Efficient power and cooling
 - Server population (energy footprint, lifetime...)
 - Shared resources
 - ...

- Sustainability in the Cloud**
- Development language
 - Data design and usage
 - Algorithmic efficiency
 - Resource allocation
 - On-demand scaling
 - Monitoring of non-functional properties (performance, efficiency)
 - ...

- Co-design of infrastructure and application**
- Better hardware (heterogeneous computing)
 - Better software
 - Better tools (for design, analysis and modelling)



Git

A Version Control System:

- You make constant changes to the code, releasing new versions
- Keep the revisions straight, storing modifications in a central repository
- Make it easy to collaborate: download/upload new revisions
- Efficient storage of file changes
- File integrity checks

The screenshot shows the GitHub web interface. At the top is the GitHub logo (a cat icon) and the word "GitHub". Below the logo is a navigation bar with links: Search or jump to..., Pull requests, Issues, Marketplace, and Explore. The main area shows a user profile picture and the user's repositories. A pink circle highlights the "Pull requests" tab in the repository header. Another pink circle highlights the "Switch branches/tags" dropdown in the repository commit list. The repository details include the owner (Sara Vallero), repository name (VirtualServers), and a brief description: "Configuration for Nginx Virtual Servers on the Virgo Kubernetes cluster at CNAF." The repository has 57 forks and 0 stars. The commit list shows several commits, with one commit highlighted in pink and labeled "My commits". The commit details show the commit hash (6e6a96c), date (on 10 Jun), and number of commits (2 commits). The repository also includes sections for README.md, VirtualServers, and Configuration for Nginx Virtual Servers on the Virgo Kubernetes cluster at CNAF.

The Hub (GitHub)

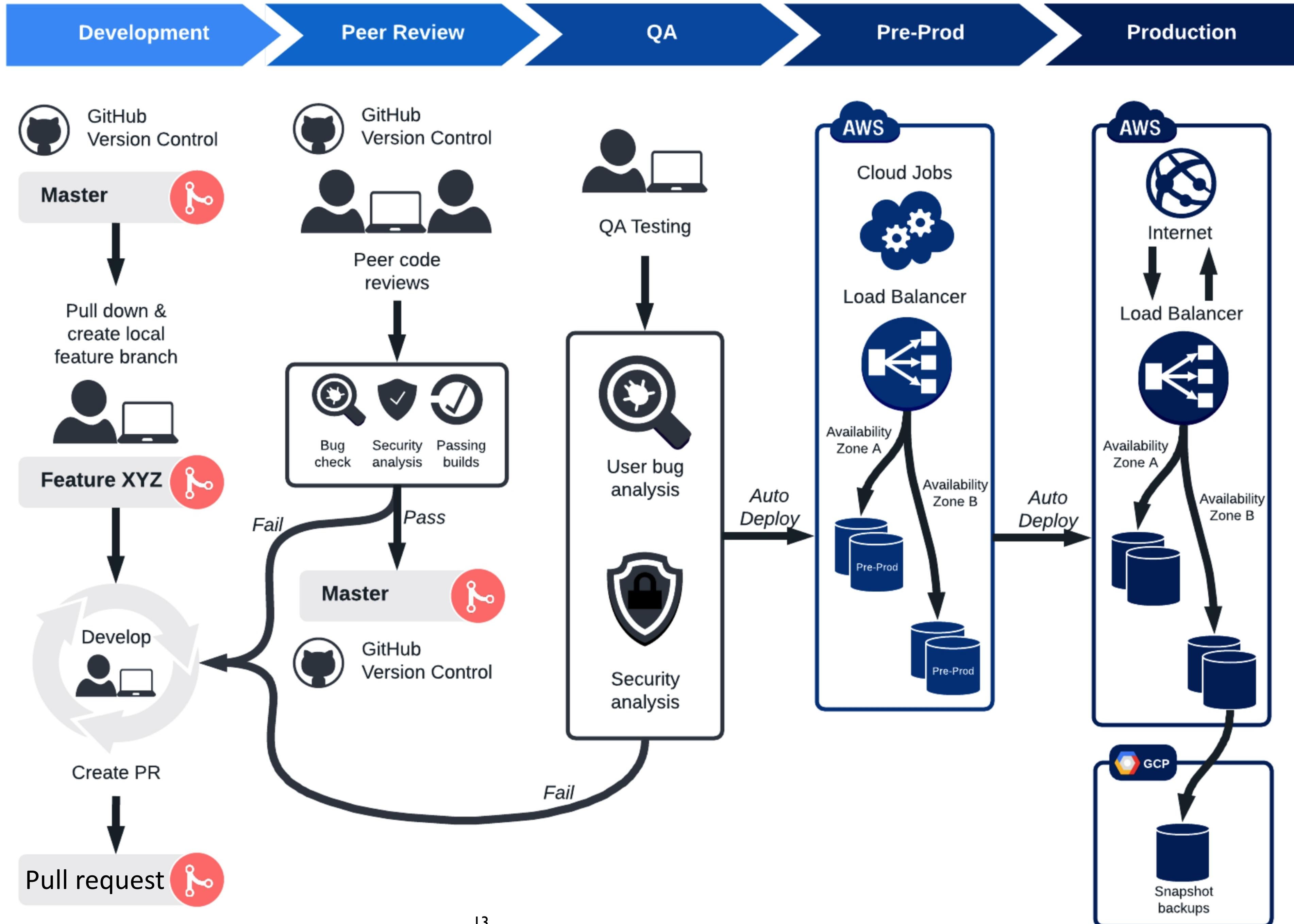
A Web Platform to:

- Store your projects
- Collaborate
- Host your project's documentation
- Implement Continuous Development
- Automatic builds
- WebHooks

DevOps workflow

Atomic commits:

implement a new feature, commit. If something goes wrong you can always go back. Each commit version is working with its own set of features.



The more sophisticated science becomes, the harder it is to communicate results. Papers today are longer than ever and full of jargon and symbols. They depend on chains of computer programs that generate data, and clean up data, and plot data, and run statistical models on data. These programs tend to be both so sloppily written and so central to the results that it's contributed to a replication crisis, or put another way, a failure of the paper to perform its most basic task: to **report what you've actually discovered, clearly enough that someone else can discover it for themselves.**

- James Somers



Reproducibility



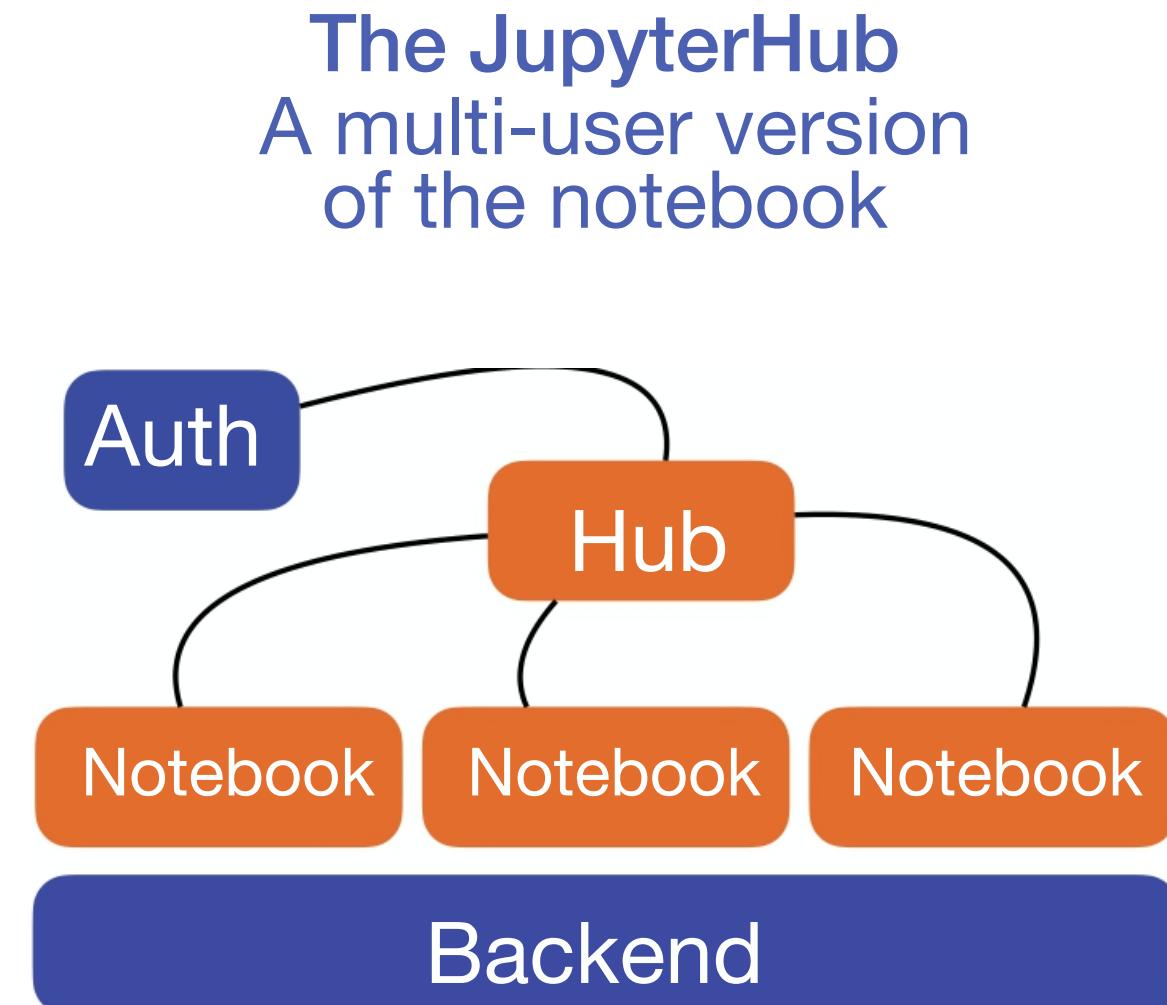
JupyterHub

The **Jupyter Notebook** is an open-source web application that allows you to interactively create and share documents that contain:

- visualisations
- narrative text
- live code
- equations

Uses include:

- data cleaning and transformation
- numerical simulation
- statistical modeling
- data visualisation
- machine learning



Multiple language support (kernels)

The screenshot shows the JupyterHub interface with multiple tabs open. On the left, a file browser displays a directory with several Jupyter notebooks and a terminal window. The main area contains sections for "Hands-on Day 1", "You'll learn", "Input data", "Visualization", and "Dataset description". In the bottom right, there is an "Exercise 2" section with a histogram titled "Distribution of m_bb". The histogram has two overlapping distributions: a blue one labeled "signal" and an orange one labeled "background". The x-axis is labeled "m_bb" and ranges from 0.0 to 3.0. The y-axis is labeled "counts" and ranges from 0 to 6000.

File Edit View Run Kernel Spark Tabs Settings Help

IPython / Filter files by name

Name Last Modified

- MLCourse-2122 2 days ago
- inputForML_day... 10 minutes ago
- sparkmonitor_k... 6 months ago
- sparkmonitor_s... 6 months ago
- Test.ipynb 7 months ago
- Untitled.ipynb 6 months ago
- Untitled1.ipynb 6 months ago

Terminal 1 inputForML.ipynb custom_magics.py

Hands-on Day 1

- course slides

You'll learn

- familiarize with jupyter notebooks, numpy and pandas

Input data

- efficient data format: convert CSV to Parquet
- create input vector with features for MLLib. Format of the input depends on chosen ML library

Visualization

- explore dataset, plot features
- correlation matrix

Dataset description

Exercise 2

- plot a few input variables and try to understand which ones are more promising to distinguish signal from background

In [11]: `#plot invariant mass distribution`
`plotSignalvsBg(df, 'm_bb')`

Distribution of m_{bb}

counts

0.0 0.5 1.0 1.5 2.0 2.5 3.0

0 1000 2000 3000 4000 5000 6000

signal background

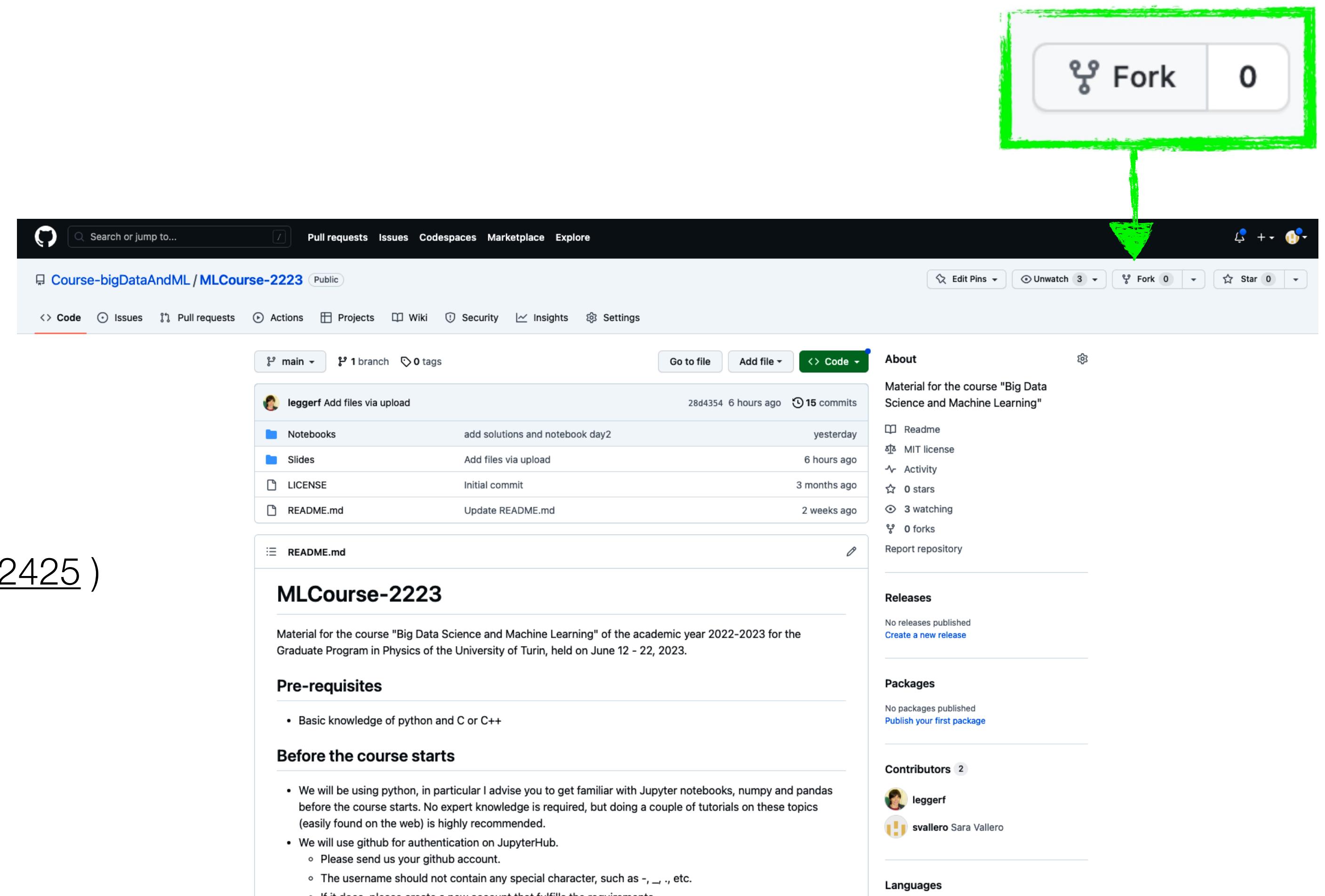


A very simple (and useful) example of DevOps workflow

During the hands-on sessions, you will have a persistent storage to save your Notebooks, but you can also save them on GitHub!

STEP 1: Fork

- login to GitHub (<https://github.com/>)
- go to the course repository
(<https://github.com/Course-bigDataAndML/MLCourse-2425>)
- fork the repository in your private space



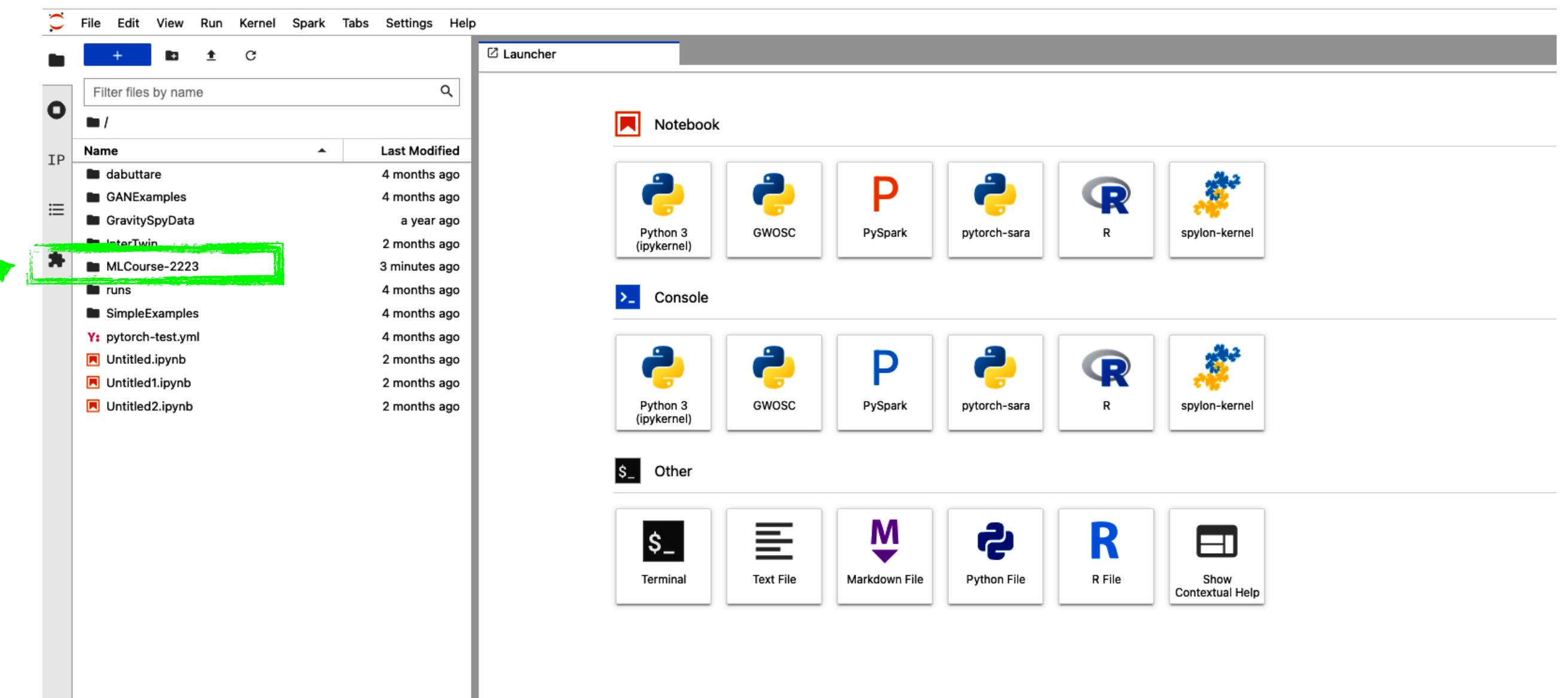
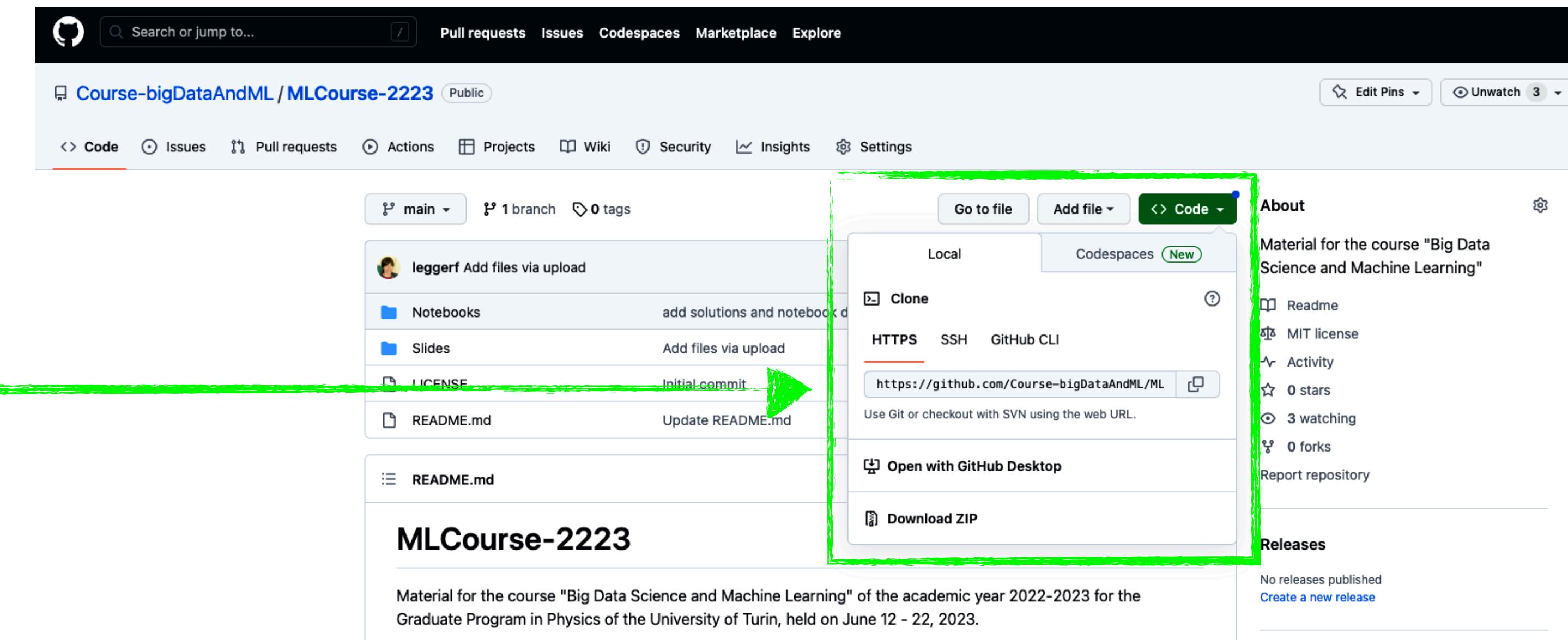
STEP 2: Clone

- You have been redirected to your private Fork of the repository
- Copy the Clone URL
- Clone your repository using the command line:

```
$ git clone https://github.com/  
Course-bigDataAndML/  
MLCourse-2425.git
```

(This has to be executed in a Jupyter Terminal as explained in one of the hands-on sessions)

- You should now have a directory called:
MLCourse-2425
- The directory is available in your JupyterHub space!



STEP 3: Status

- Browse the directory and start a Notebook.
- Make local modifications to the Notebooks (i.e. solve the exercises) and save them
- Now you can go back to the terminal and see what happened...

```
$ git status
```

On branch main

Your branch is up to date with 'origin/main'.

Changes not staged for commit:

(use "git add <file>..." to update what will be committed)

(use "git restore <file>..." to discard changes in working directory)

modified: inputForML.ipynb



This is the file you have changed and you want to commit.

Untracked files:

(use "git add <file>..." to include in what will be committed)

Notebooks/Day1/.ipynb_checkpoints/



These are the files that will not go in the repository unless you explicitly add them (and you don't want to do that).

no changes added to commit (use "git add" and/or "git commit -a")

STEP 4: Commit

- Add your changes
- Commit your changes
(remember the atomic commits)
- Check the status
- So far you made only local modifications

```
$ git add Notebooks/Day1/inputForML.ipynb
```

```
$ git status
```

On branch main

Your branch is up to date with 'origin/main'.

Changes to be committed:

(use "git restore --staged < file>..." to unstage)

modified: Notebooks/Day1/inputForML.ipynb

Untracked files:

(use "git add < file>..." to include in what will be committed)

Notebooks/Day1/.ipynb_checkpoints/

STEP 5: Push and Pull

- Send your changes to the central repository:

git push

- Get the latest version of the code, including changes made by others: **git pull**

```
$ git commit -m "All exercises solved"
```

[main 942ef43] All exercises solved

1 file changed, 7354 insertions(+), 22 deletions(-)

```
$ git status
```

On branch main

Your branch is ahead of 'origin/main' by 1 commit.

(use "git push" to publish your local commits)

...

To take home

- DevOps approach for fast time (efficiency) to value
- Version Control System (Git) to book-keep code changes and ease collaboration
- Make your workflow understandable and reproducible (Notebooks)

Wrap-up

No BigData/MachineLearning expert is such if they cannot compute a **real-size problem** (scale-out/scale-up)

We introduced several concepts and tools which are at the basis of **modern computing strategies**, now you know how to place them in the right context.

We hope this will help you in your daily work.