Data/code management

NCML lab meeting

2024-05-15 Seung-Goo KIM

MAX PLANCK INSTITUTE
FOR EMPIRICAL AESTHETICS

Raw data backup

3-2-1 backup strategy steps

3 — data copies (1 primary and 2 backups)

2 — types of storage

1 — off-site storage

ILLUSTRATION: MAGLARA/ADOBE STOCK

©2019 TECHTARGET. ALL RIGHTS RESERVED TechTarget
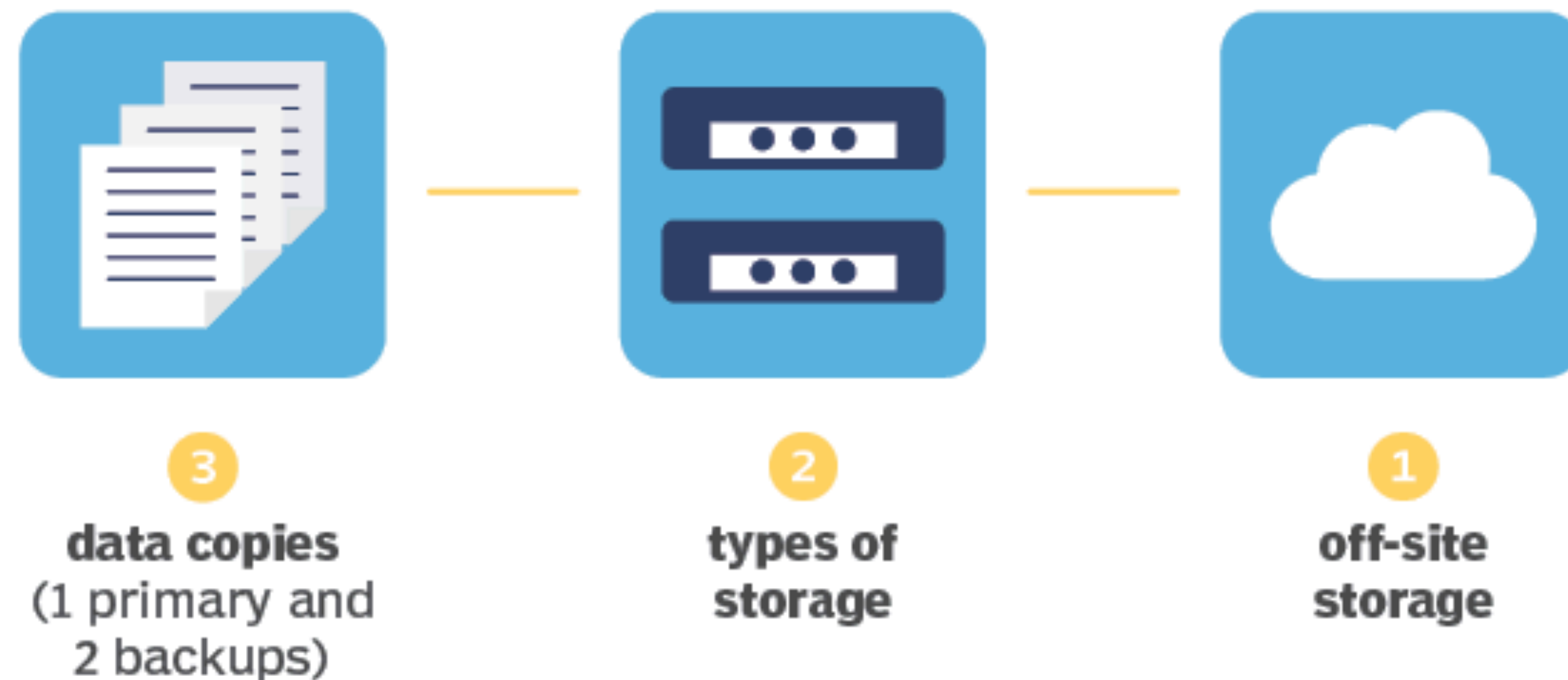
- 1. Original: laptop HDD

- 2. Copy in the same HDD

- 3. Copy in a server (still in the same building)

Raw data backup

3-2-1 backup strategy steps

3 data copies (1 primary and 2 backups)

2 types of storage

1 off-site storage

ILLUSTRATION: MAGLARA/ADOBE STOCK

©2019 TECHTARGET. ALL RIGHTS RESERVED

- 1. Original: laptop HDD

- 2. Copy in the same HDD

- 3. Cloud copy (in a different city)

"Scratch" (in progress) data backup

• At least just two?

3-2-1 backup strategy steps

3 data copies (1 primary and 2 backups)

2 types of storage

1 off-site storage

ILLUSTRATION: MAGLARA/ADOBE STOCK

©2019 TECHTARGET. ALL RIGHTS RESERVED TechTarget

MPIEA project folder and HPC folder
Institute's servers

- They are backed up everyday (so I was told). So, most of the time, you can recover what you had yesterday.

- But having something in just one local copy is definitely scary...😨

  - Whatever something that took long (1+ day) to compute, please copy it to Keeper or MPIEA servers.

# Code backup & version control

It's a good idea to always upload it on cloud

- **Manual**: `git` using Github or Gitlab

  - The server may not be secured (github uses it for AI training).

  - But so many people are using github.

- **Automatic**: a cloud folder using MPDL Keeper (powered by Seafile)

  - The server is secured.

  - But the client may be not.

Project structure

https://github.com/seunggookim/projtemp/

- `.gitignore`: `data`

  - github file size limit = 50 MB

  - git is text files, and is slow for binary files.

- What is git?

  - https://youtu.be/92sycL8ij-U?si=MUkW9F3KcoPPI_tb

- Why we need commits and branches?

  - https://youtu.be/lJu5xwbGgRk?si=ccW8tuZbu8Gq5JuK
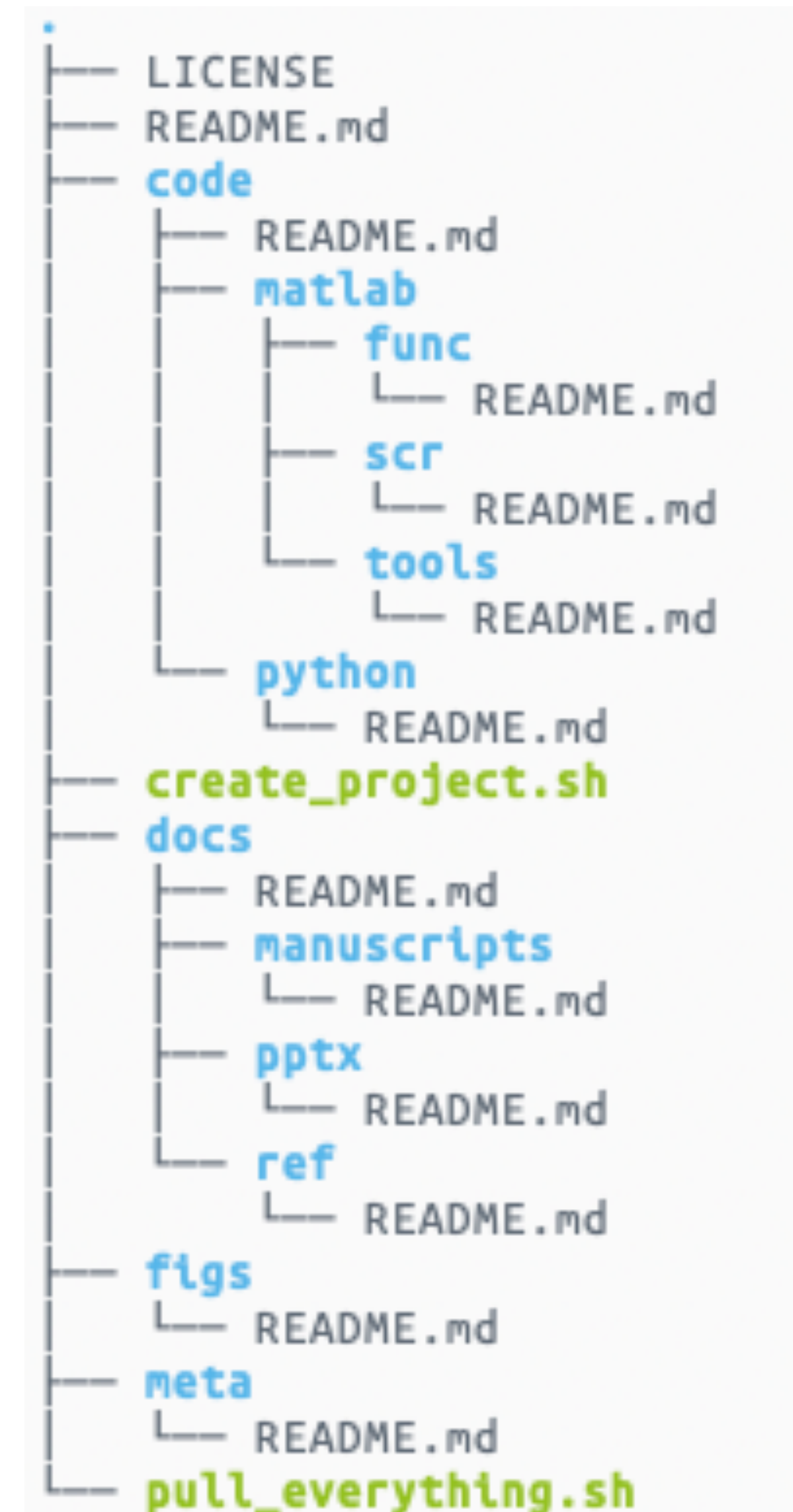
Data directory structure

Hierarchy works well

```
[seung-goo.kim@login02:~/usrfolder/musencemo/data
(main)$ tree -L 1
.
├── bhv
├── essentia
├── fs
├── madmom
├── ohbm2022
├── ohbm2024
├── raw
├── s6xuaica
├── s6xuaica_2023-05-22
├── s6xuaica_rest
├── sim
└── spmants
```

```
seung-goo.kim@login
[(main)$ tree raw -L
raw
├── ds003085-downlo
│   ├── annex-uuid
│   ├── CHANGES
│   ├── dataset_des
│   ├── doc
│   ├── stim
│   ├── sub-01
│   ├── sub-02
│   ├── sub-03
│   └── sub-04
```

```
spmants
├── run_aroma.sh
├── sub-01
├── sub-02
├── sub-03
├── sub-04
├── sub-05
├── sub-06
└── sub-08
```

```
s6xuaica
├── cca
│   └── avg_n37
├── commontimes.mat
├── enc
│   └── avg_n37
├── fmri
│   ├── happy
│   ├── mask.nii.gz
│   ├── mask.png
│   ├── mean.nii.gz
│   ├── mean.png
│   ├── prob.nii.gz
│   ├── prob.png
│   ├── sadln
│   └── sadsh
├── me_cca.log
├── me_cca.mat
├── me_import.log
├── me_import.mat
├── musencemo_cca.log
├── rate
│   ├── happy
│   ├── sadln
│   └── sadsh
└── stim
    ├── filtcoch-pca-95pcs-2D.png
    ├── filtcoch-pca-95pcs-3D.png
    ├── filtcoch-pca.mat
    ├── happy
    ├── openl3-l02-pca.mat
    ├── openl3-l02-pca.png
```