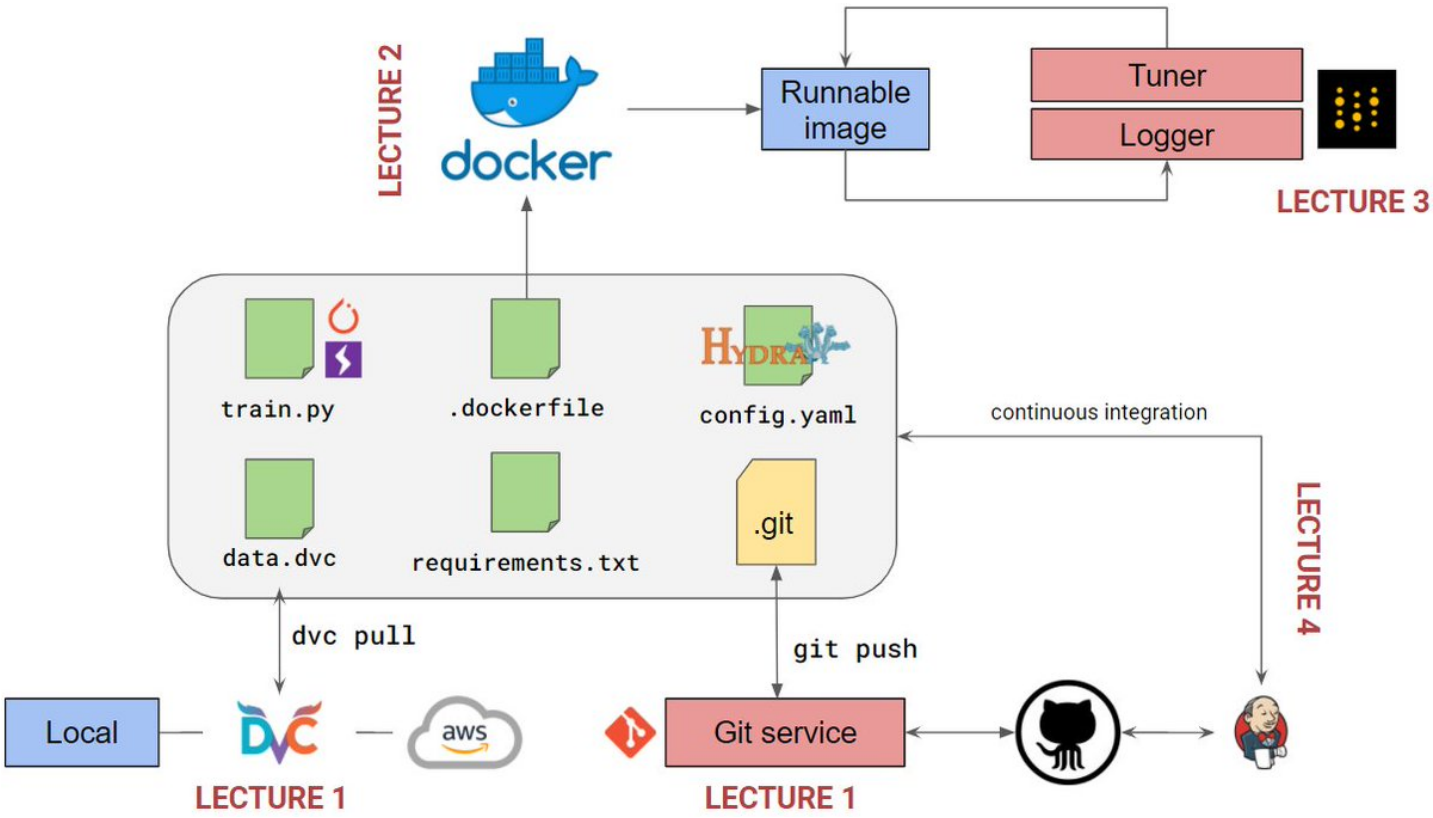
Twitter Thread by Simone Scardapane

Simone Scardapane
@s_scardapane

Reproducible deep learning
Lectures 3 and 4 are out!

With code versioning out of the way, it is time to look at data versioning (@DVCorg) and environment isolation (@Docker).

All information in a small thread. ■ /n

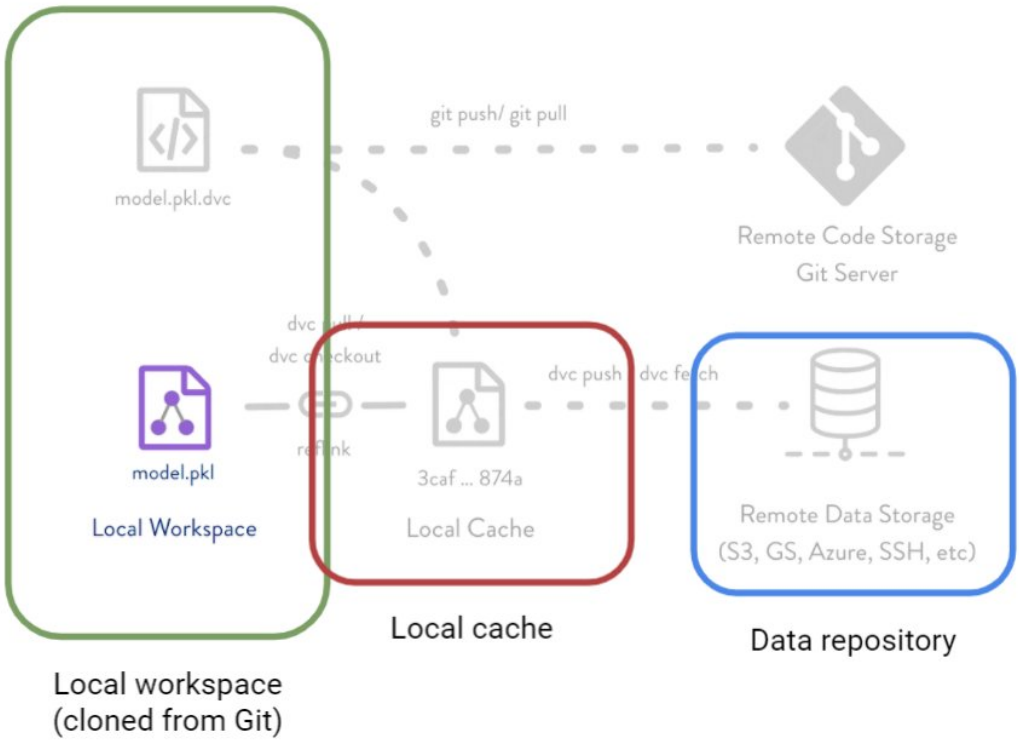


If you know Git, you (almost) know @DVCorg!

A fantastic tool to secure your data in a number of remotes, or to create "data repositories" from which to immediately get folders and artifacts.

My intro to DVC: https://t.co/2m3cXGAPN6

/n



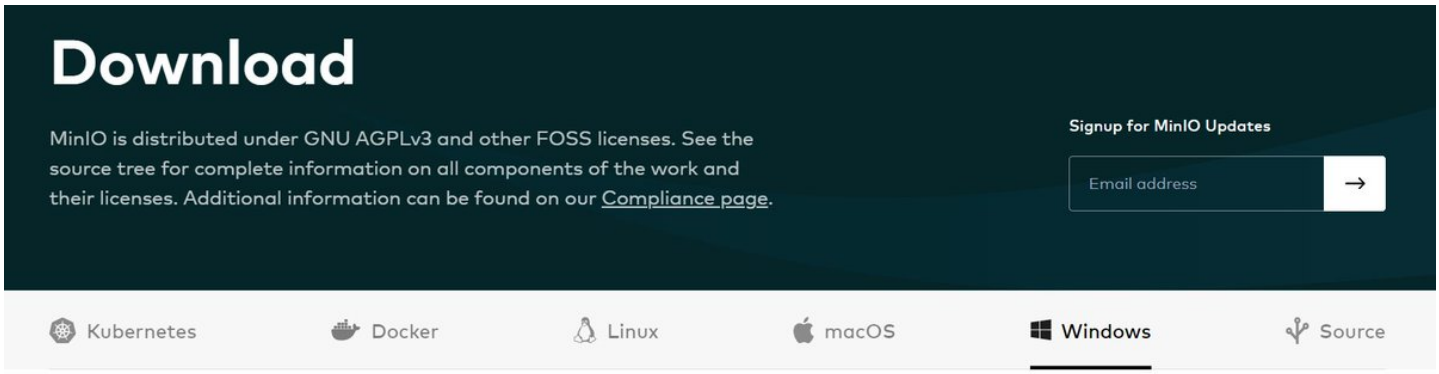
iterative/dvc: Data Version Control | Git for Data & Models

For the course, I created a simple exercise tasking you with initializing DVC on the repository, and syncing the data locally and remotely.

To simulate an S3-like interface, we use a small https://t.co/91bFj7KSPG server and boto3.

Code: https://t.co/KDSX80aqJs

/n

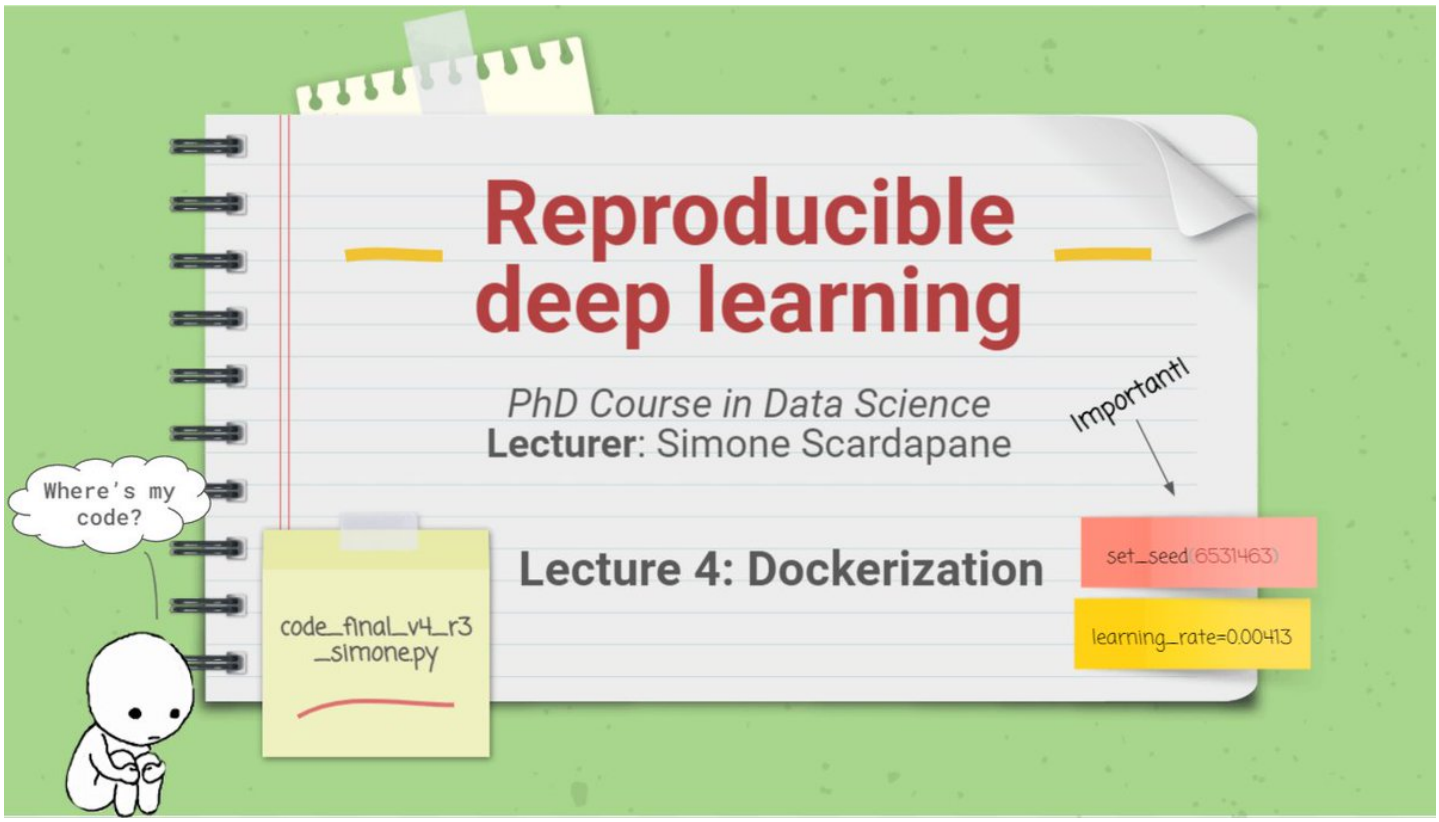


Next up, it is time to "dockerize" your environment!

Docker has become an almost de-facto standard, and knowing it is practically indispensable today.

A very quick introduction, glossing over a number of details: https://t.co/XSrUZNhd3g

/n



In the corresponding exercise, you will learn about creating a working environment in Docker, packaging the entire training loop, and pushing/pulling an image from the Hub.

Code is here: https://t.co/YAAN3YB4tV

/n

That's it for today!

Next week, we conclude with an overview of @weights_biases, continuous integration, Git hooks, and unit tests.

The full code (divided into branches) is here: https://t.co/zBcwOeJhRy

Credits to the fantastic documentation from both @DVCorg @Docker ■