

# Introducción a Data Versioning in MLOps

**By: Jose Alberto Arango Sánchez.**

**01.** ¿Qué es DVC?

**02.** Características

**03.** ¿Cómo funciona?

**04.** Instalación

**05.** Demo

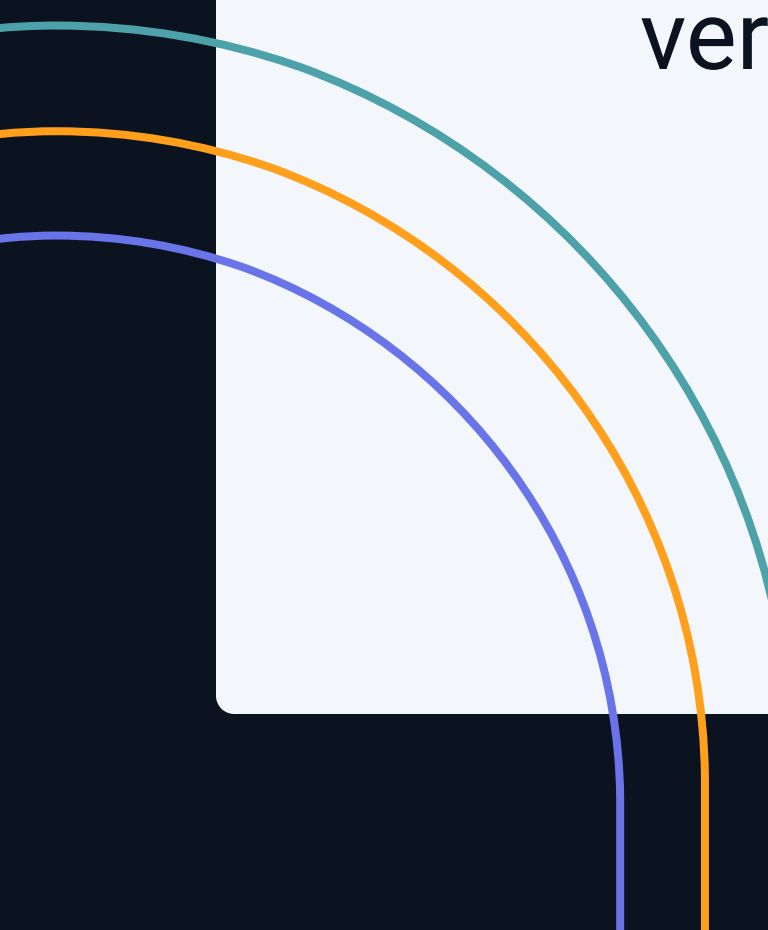
# Tabla de contenido





# ¿Qué es DVC?

Data Version Control (DVC) es un sistema de control de versiones de datos open source.



# Características de



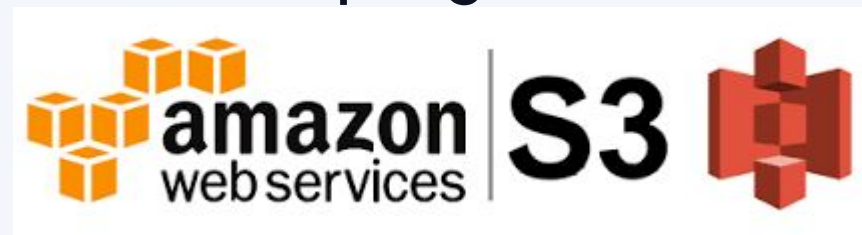
## Versionado

- Permite manejar modelos grandes y archivos de datos que no se pueden manejar con Git.
- Compatible con Git



## Almacenamiento

- Agnóstico al almacenamiento y lenguaje de programación.



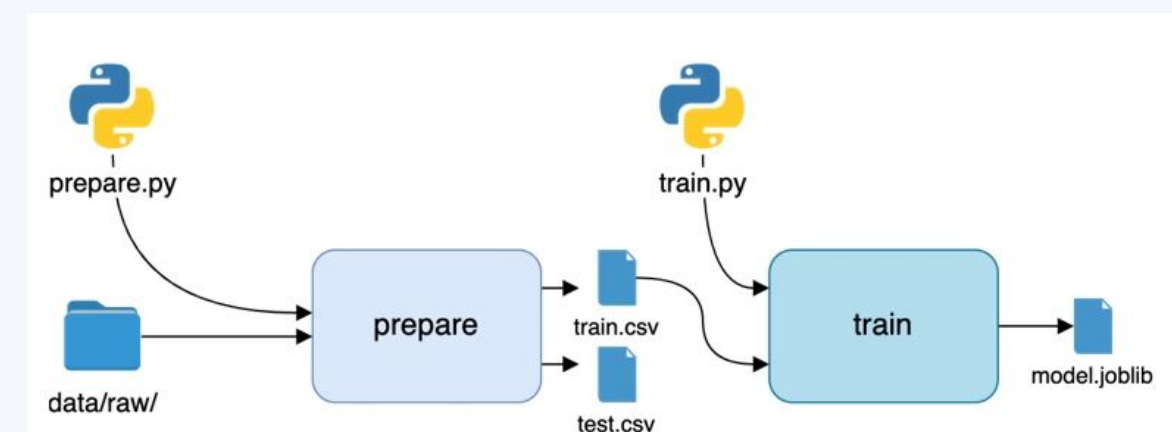
DVC Storage


































Google Cloud Storage

## Replicabilidad

- Ayuda a administrar la complejidad de los pipelines de ML para que pueda entrenar el mismo modelo repetidamente.



# Benchmark

	Open Source	Data Format Agnostic	Cloud/Storage Agnostic* (Supports most common cloud and storage types)	Simple to Use	Easy Support for BIG Data
	 (Apache 2.0)				
 <b>DELTA LAKE</b>	 (Apache 2.0)				
 <b>Git Large File Storage</b>	 (MIT)				
 <b>Pachyderm</b>	  (Non-standard license)				
 <b>DOLT</b>	 (Apache 2.0)				

# Principales comandos



```
$ dvc init # initialize the repo
$ dvc add . # add the files that have been changed
$ dvc commit -m "making some changes" # commit the updates with a message
$ dvc remote add newremote s3://bucket/path # point the repo to an S3
bucket for storage
$ dvc push # push the changes to the DVC repo hosted in the default S3
bucket
$ dvc pull # pull the latest changes from the DVC repo hosted in the
default S3 bucket
```

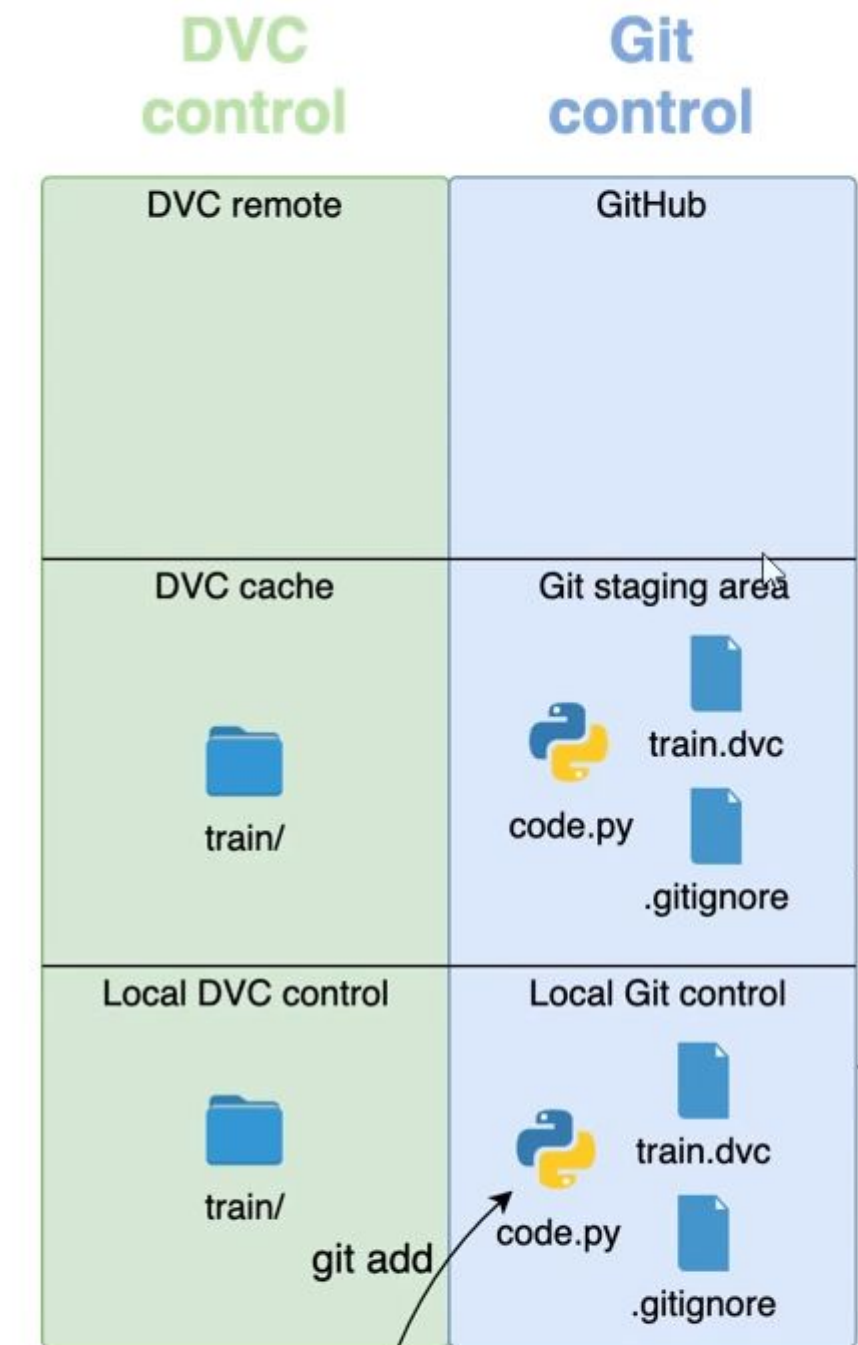
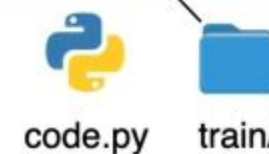
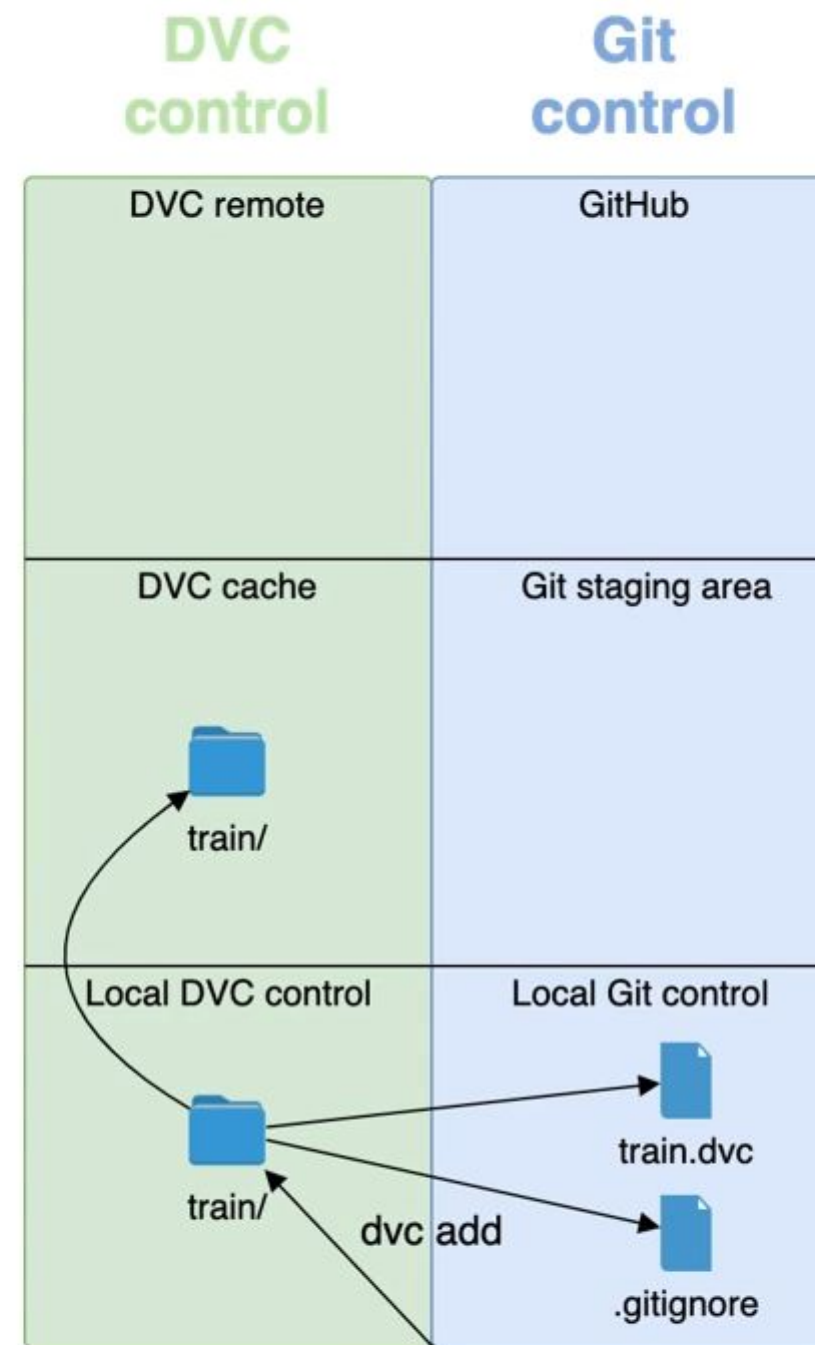
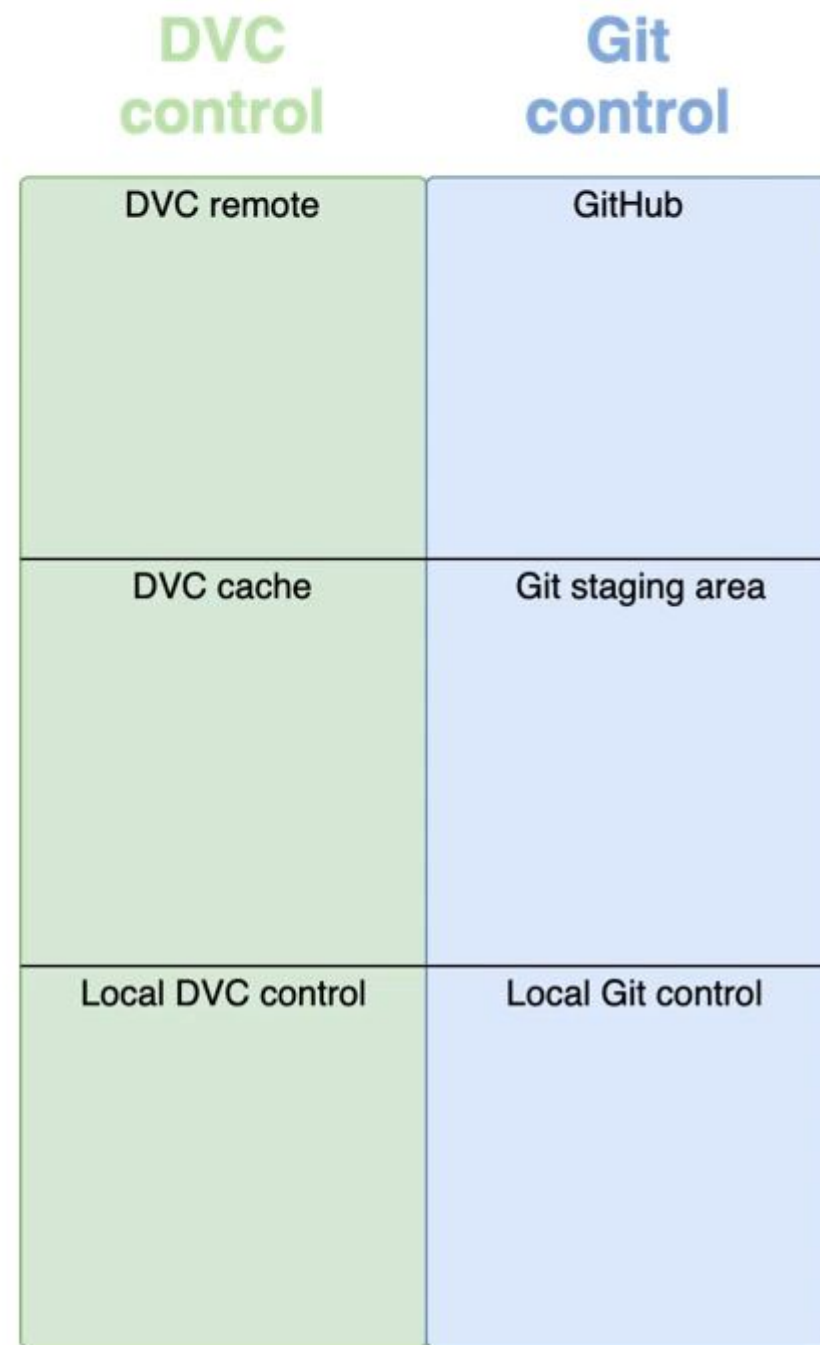


# ¿Cómo funciona?



`dvc add data/raw/train`

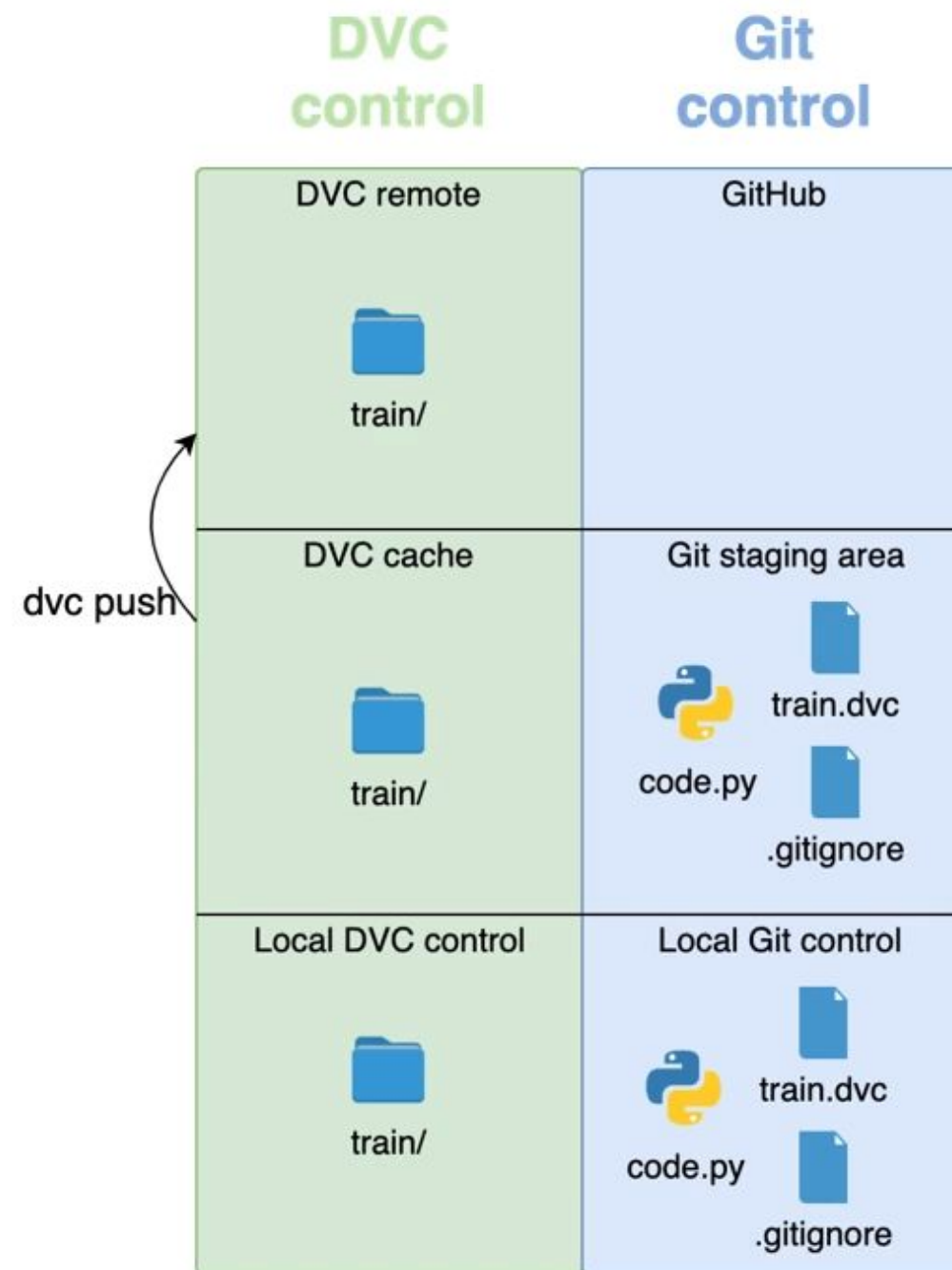
`git add --all`



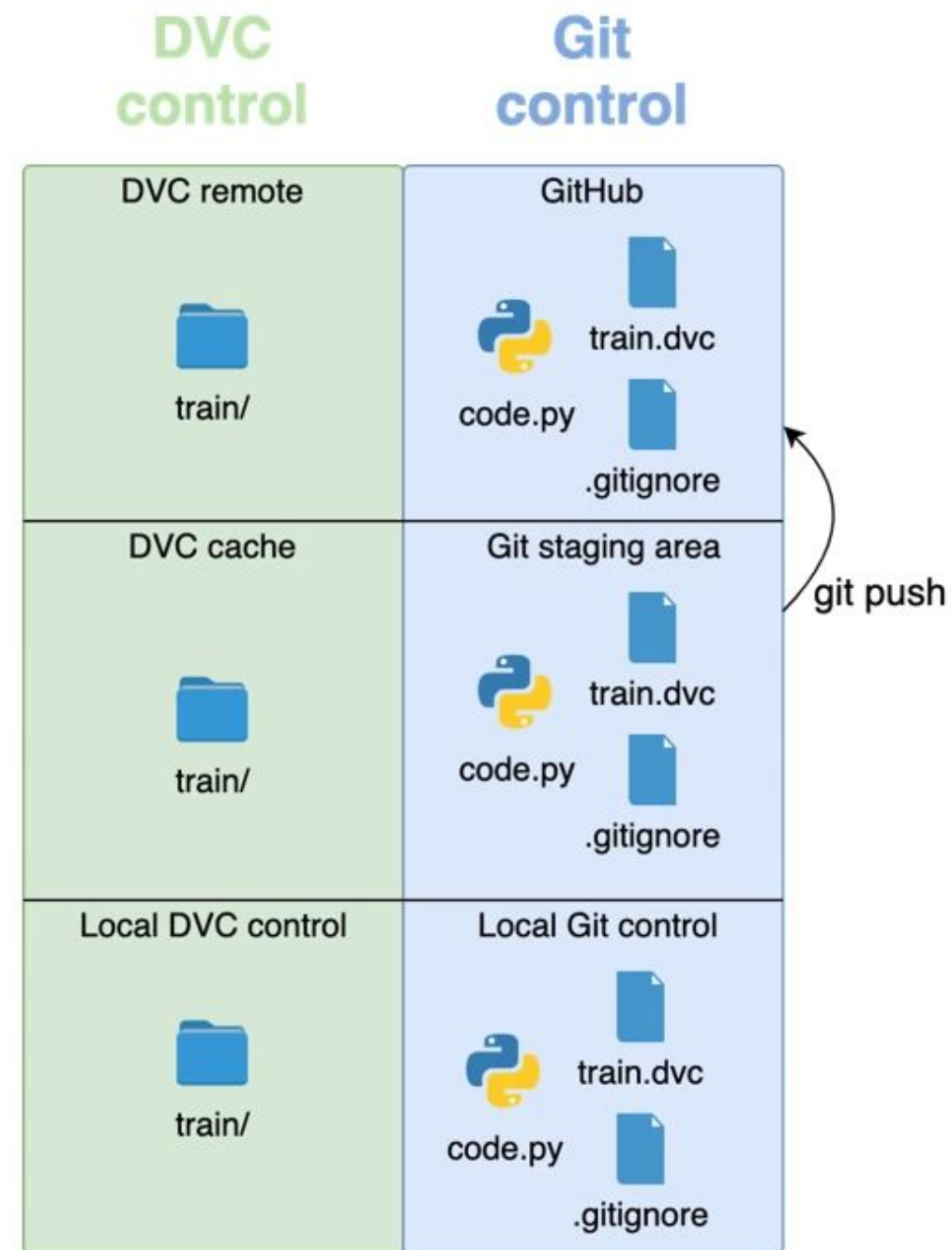
# ¿Cómo funciona?



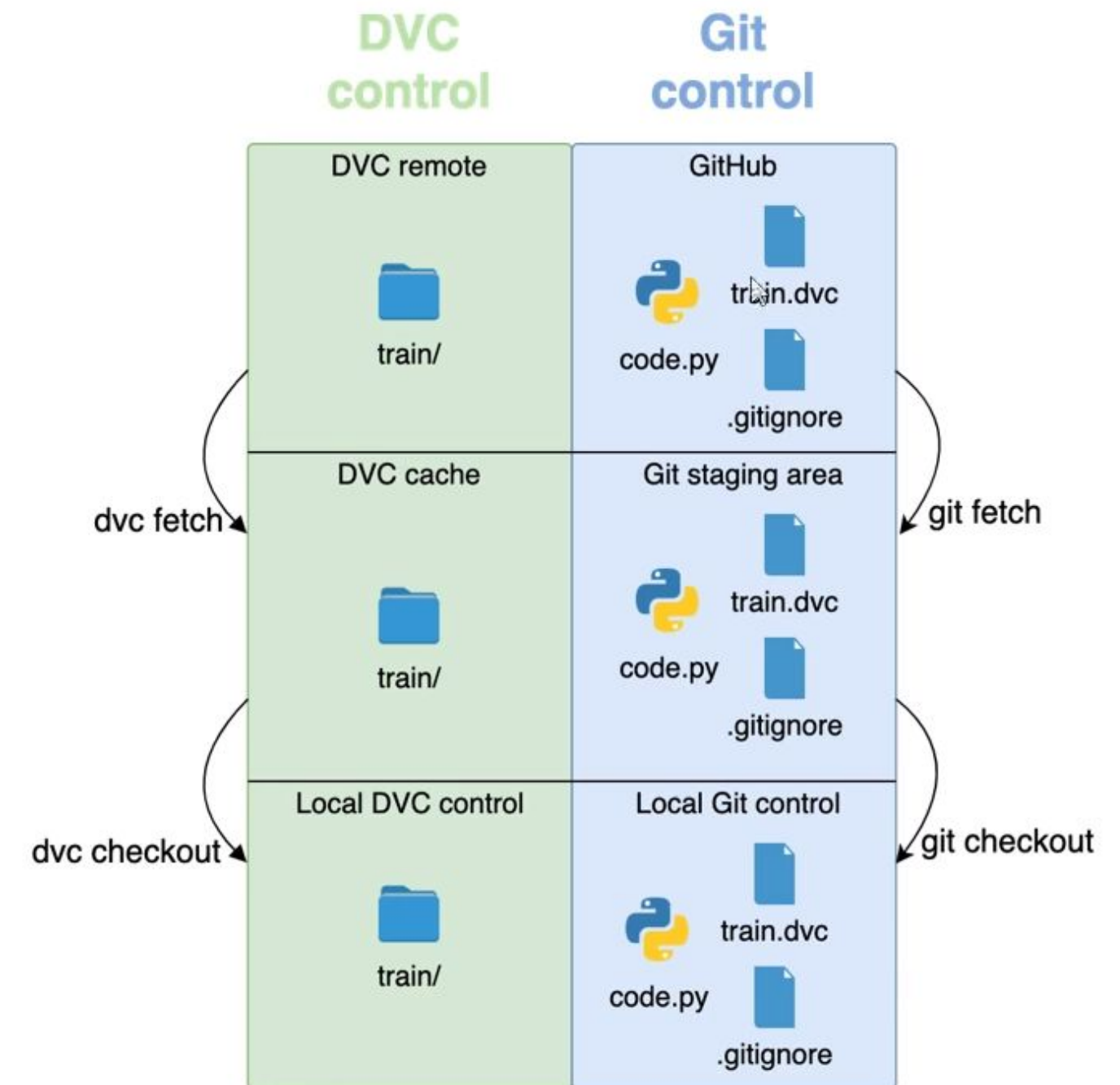
## dvc push



## git push




## dvc pull






# Instalación

 Note that Python 3.8+ is needed to get the latest version of DVC.

```
$ pip install dvc
```

 Requires [Miniconda](#) or [Anaconda Distribution](#).

```
$ conda install -c conda-forge mamba # installs much faster than  
$ mamba install -c conda-forge dvc
```



---



# Demo



# Muchas gracias!

# ¿Alguna duda?

Science for everyone !!!



# Recursos de interés

- Effective MLOps: Model Development
- CI/CD for Machine Learning (GitOps).

<https://www.wandb.courses/pages/w-b-courses>

- A curated list of MLOps projects

<https://mlops.toys/#projects>