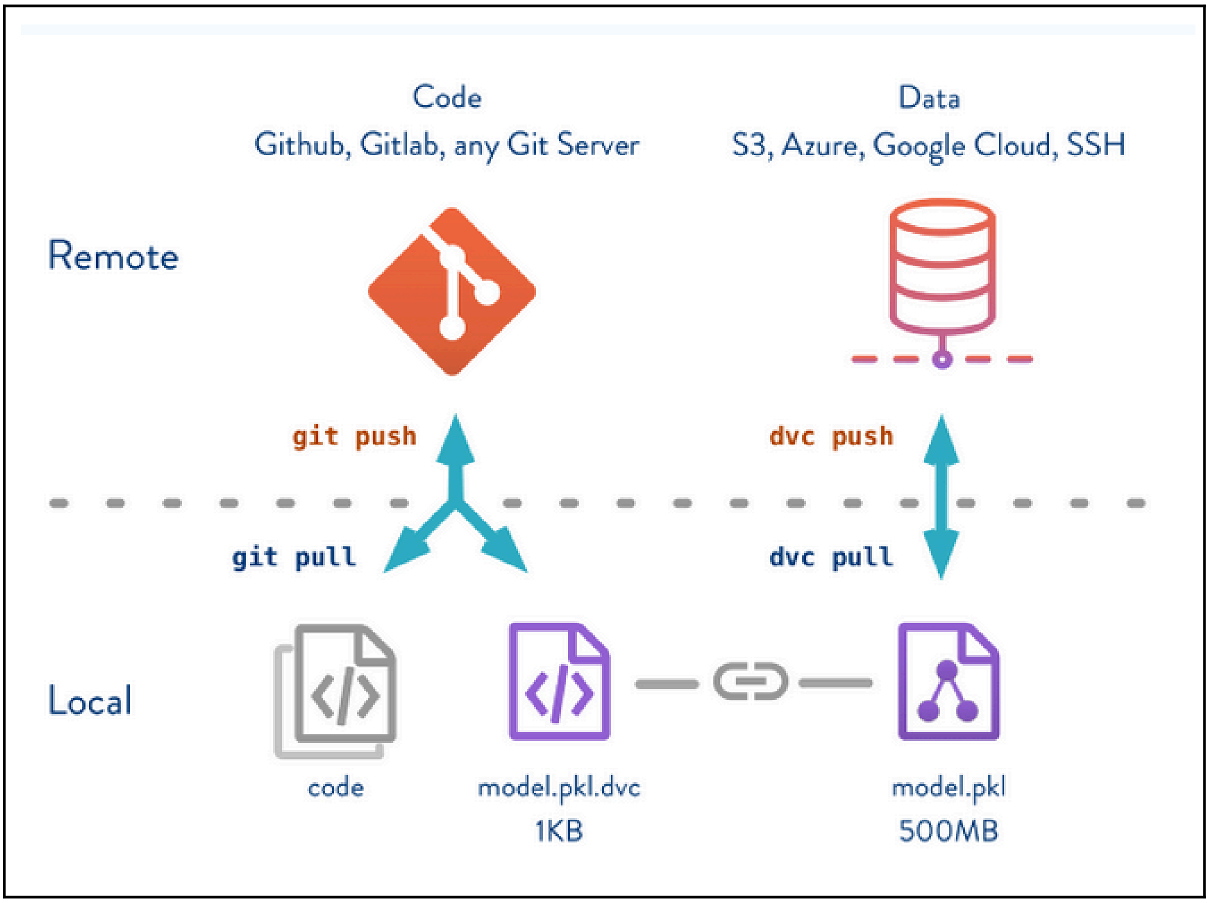
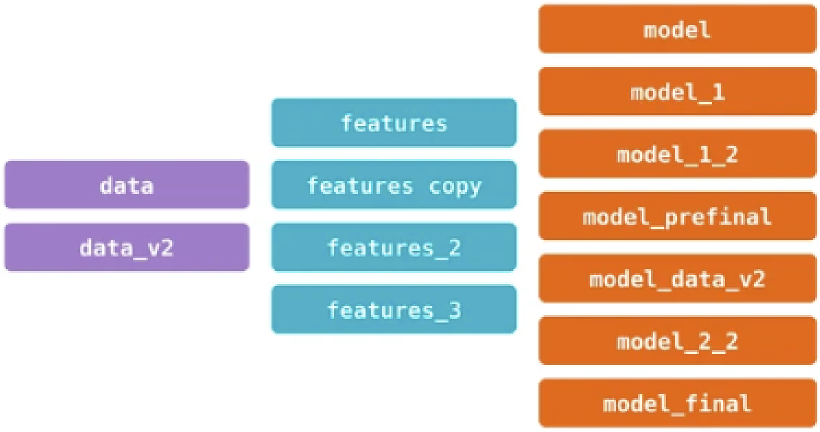
Data Version Control (DVC)

Even with all the success we've seen in machine learning, especially with deep learning and its applications in business, data scientists still lack best practices for organizing their projects and collaborating effectively. This is a critical challenge: while ML algorithms and methods are no longer tribal knowledge, they are still difficult to develop, reuse, and manage.

DVC (Data Version Control) is a prominent open source tool that works alongside git. GIt manages the code and small metadata files (.dvc files), while DVC handles the large data and model files, storing them in separate local or cloud-based storage (AWS S3, Google Cloud Storage).



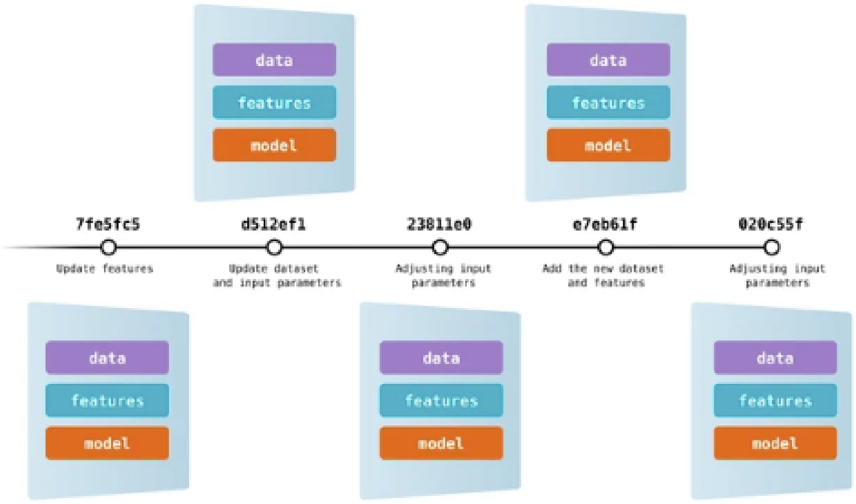
It manages the data the same way the git manages the code. It uses a Git-like model to bring the best practices of software engineering to data, AI/ML and data science teams.



Exponential complexity of data science projects

DVC is useful, if you want to store and process data files or datasets to produce other data or machine learning models and you want to:

1. Track and save data and machine learning models the same way you capture code.
2. Create and switch between versions or data and ML models easily.
3. Understand how datasets and ML artifacts were built in the first place.
4. Adopt engineering tools and best practices in data science projects.



DVC matches the right versions of data, code, and models for you 💘.