

ML OPS Assignment #1

Data Versioning and Differential Privacy

Danil Meresench

Ian Vogt

Monica Zhang

Jason Rajan



Tool Comparison: DagsHub vs DVC

	Data Version Control (DVC)	DagsHub
Ease of installation	Download from DVC website and guided installation. Did not need to make a DVC account. Could start using directly using Terminal with <i>dvc init</i> . Supports multiple storage types such as S3, Azure, and Google Drive.	Only requires dvc and git installations + a DagsHub signup. Zero configuration storage setup for each project.
Ease of data versioning	Data versioning available through the Terminal with a couple of commands. Easy and straightforward, with both written and video documentation available for reference.	Can push locally versioned data to your repository through the command prompt or with the dagshub python package <code>upload_files</code> module. Can manage code, notebooks, data, models in one place.
Ease of switching between versions for the same model	Can quickly switch between versions through commands in the Terminal. A DVC Python API is also available to read and stream different version of data from remote storage.	Can pull versioned datasets stored in your DagsHub repository into Jupyter notebook directly without needing local storage.
Effect of DP on model accuracy/metrics	No noticeable impact on model accuracy/metrics	No noticeable impact on model accuracy/metrics

