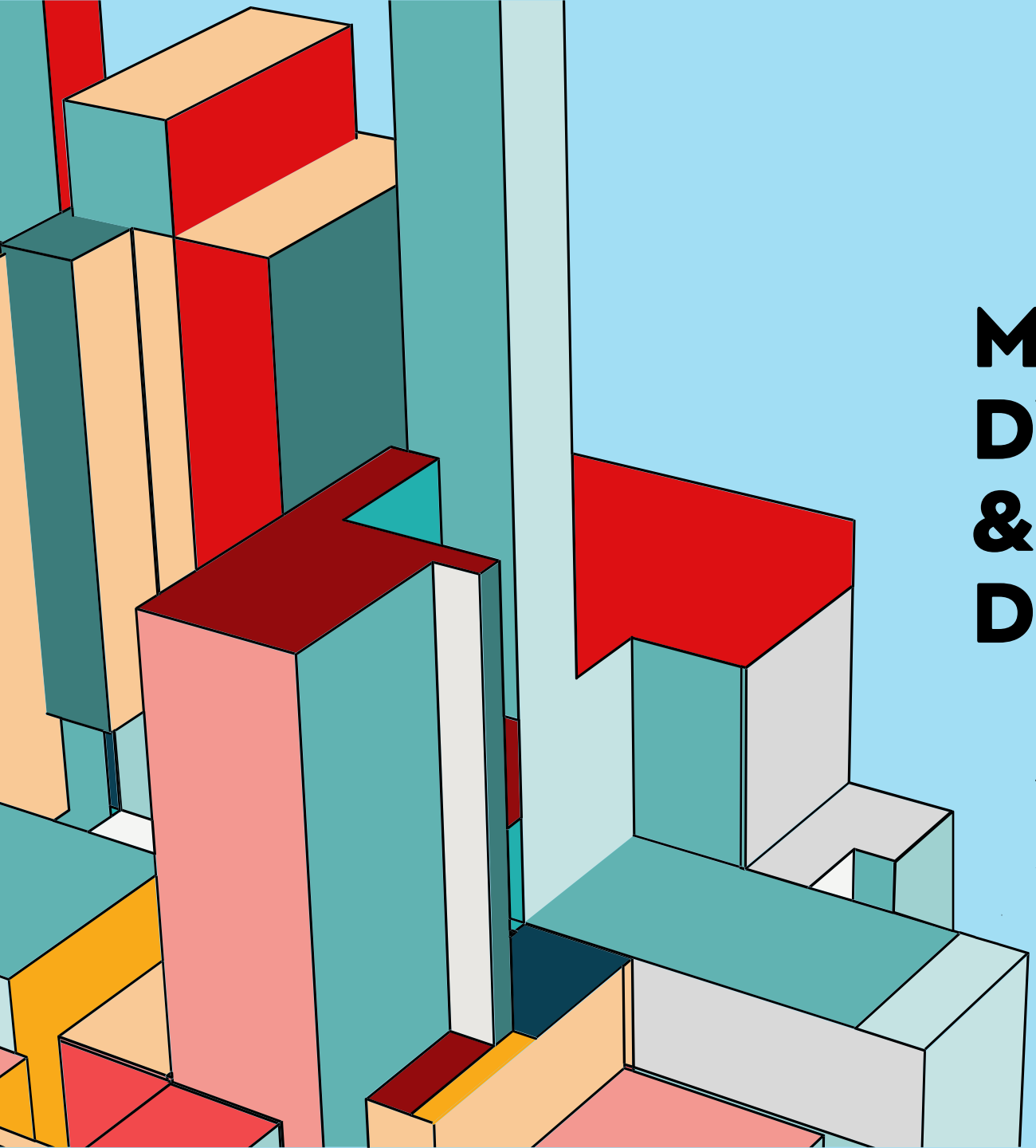

MlOPs Assignment 1: DVC Tools Comparison & DP Models

Anusha Bhat

July 3, 2025

Tools Comparison: Lakefs And Gitlfs

- LakeFS is powerful but slightly complex
  - Great for teams, pipelines, and production ML.
- GitLFS is **easy to use**
  - Best for **individual projects** and lightweight workflows.
- Both tools **allowed for seamless pushing of csv versions** for integration in model building.

Comparison	LakeFs	GitLFS
Ease of Installation	• Challenging • Requires external storage setup (e.g. S3), access keys, and bucket configuration. • Need to use LakeFS CTL for CLI or Boto3 library in python to push and pull.	• Straightforward • Requires installation of GitLFS and repository setup on GitHub • Can use standard git commands.
Ease of Data Versioning	• Easy • Git-like branches & commits for data in object storage	• Easy • Tracks CSVs easily using Git and LFS pointers
Version Switching	Can easily branch and rollback like Git to switch.	Manually have to check previous commits if versions aren't kept separately.
UI	• Can view the raw CSV files within the repository directly on LakeFS cloud website.	• Can view the repository on GitHub but CSV files are pointers to GitLFS.
Model Training Integration	Can version data alongside code for model training in pipeline if using .	• Limited capacity • Have to commit changes via terminal, separate from the model training pipeline.
Cloud Dependency	• Needs external object storage linked to the repository	• Linked to a Git remote such as a GitHub repository. • No need to configure a storage bucket.

Model Comparison: DP Vs. Non-DP

- The DP model **performed similarly in accuracy** to the non-DP model
  - Minimal increase in error
  - Minimal drop in model explanation power
  - **Small trade-off in RMSE and R^2**
- The DP model epsilon signals a **moderate data privacy level,** with room for improvement.
  - Allows **for gain of a decent privacy guarantee** without significantly decreasing model accuracy
  - Suitable for applications that need some privacy, but model performance is critical

Model	RMSE	R^2	Epsilon	Delta
Non-DP	177.95	0.59	-	-
DP	179.06	0.585	0.784	4.165e 05