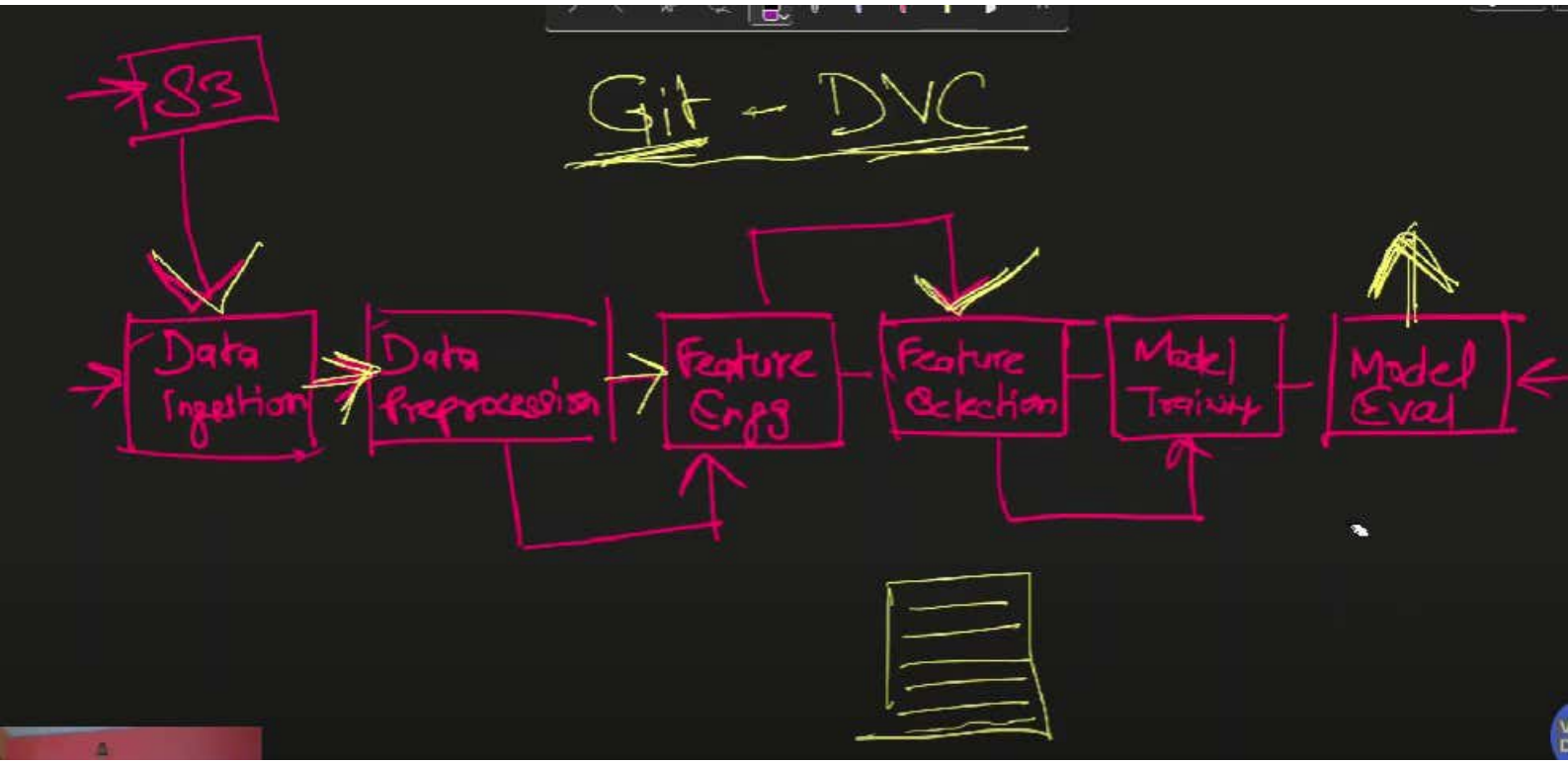
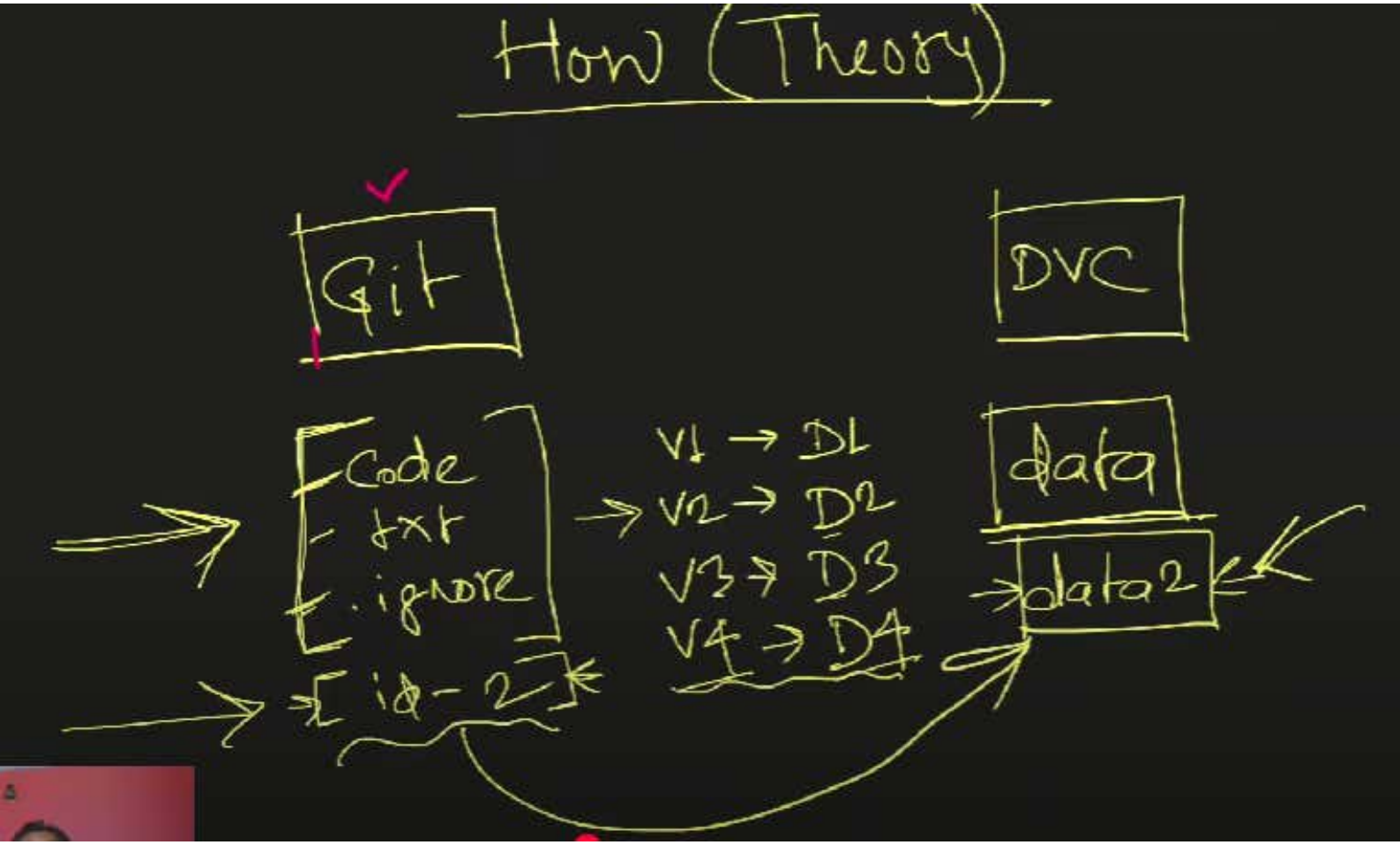
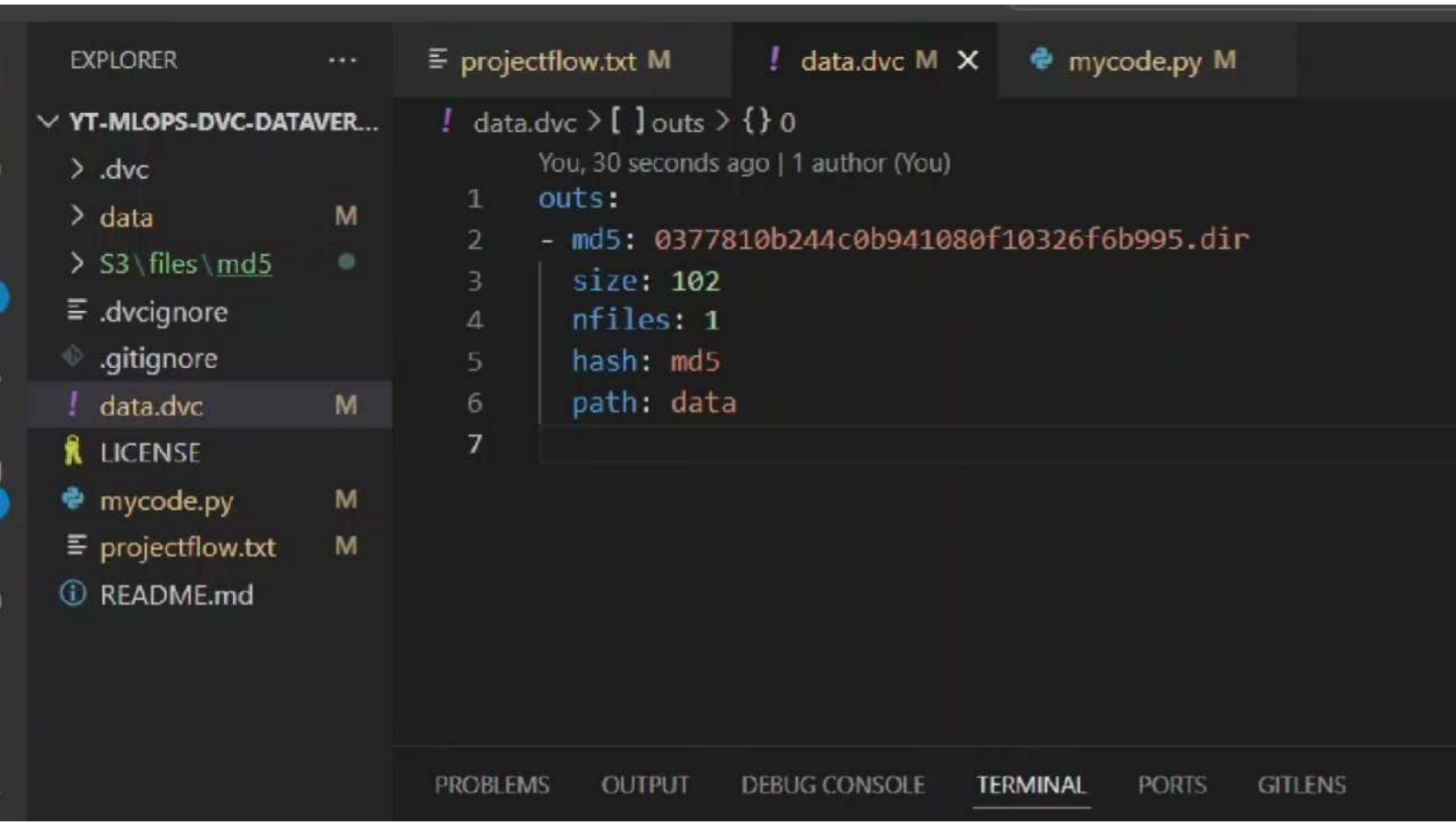

S3
Git - DVC
Data Ingestion
Data Preprocessing
Feature Engg
Feature Selection
Model Training
Model Eval


How (Theory)
Git
DVC
Code
.txt
.ignore
id-2
V1 → D1
V2 → D2
V3 → D3
V4 → D4
data
data2

1. Create git repo and clone it in local.
2. Create code.py and add code to it. (it will save a csv file to a new "data" folder)
3. Do a git add-commit-push before initializing dvc.
4. Now we do "dvc init" (creates .dvcignore, .dvc)
5. Now do "mkdir S3" (creates a new S3 directory)
6. Now we do "dvc remote add -d myremote S3"
7. Next "dvc add data/"
   Now it will ask to do: ("git rm -r --cached 'data'" and "git commit -m "stop tracking data"")
   Because initially we were tracking data/ folder from git so now we remove it for DVC to handle.
8. Again we do "dvc add data/" (creates data.dvc) then "git add .gitignore data.dvc"
9. Now - "dvc commit" and then "dvc push"
9. Do a git add-commit-push to mark this stage as first version of data.
10. Now make changes to code.py to append a new row in data, check changes via "dvc status"
11. Again - - "dvc commit" and then "dvc push"
12. Then git add-commit-push (we're saving V2 of our data at this point)
13. Check dvc/git status, everything should be upto date.
14. Now repeat step 10-12 for v3 of data.

Extension - DVC has to be installed
dvc init (creates .dvcignore, .dvc)
<locate dvc to global if any error thrown>
---
mkdir %temp%/dvcstore or mkdir temp
dvc remote add -d myremote %temp%/dvcstore or dvc remote add -d myremote temp
---
dvc add data/ (creates data.dvc)
git add .gitignore data.dvc
dvc commit - dvc push
dvc status - git status
dvc push
git log --oneline
git checkout <hash> - git checkout master
dvc pull


EXPLORER
projectflow.txt M
data.dvc M
mycode.py M
YT-MLOPS-DVC-DATAVER...
.dvc
data M
S3\files\md5
.dvcignore
.gitignore
data.dvc M
LICENSE
mycode.py M
projectflow.txt M
README.md
data.dvc > [ ] outs > {} 0
You, 30 seconds ago | 1 author (You)
outs:
- md5: 0377810b244c0b941080f10326f6b995.dir
size: 102
nfiles: 1
hash: md5
path: data
PROBLEMS
OUTPUT
DEBUG CONSOLE
TERMINAL
PORTS
GITLENS