

| 00 |



> Open source tools  
for machine learning models  
and data sets versioning

Dmitry Petrov [iterative.AI](https://iterative.ai)

|HELLO|



**Dmitry Petrov**

PhD in Computer Science

Twitter: [@FullStackML](#)

**DVC**

Creator of

[DVC.org](#) project



Co-Founder & CEO > Iterative.AI > San Francisco, USA



ex-Data Scientist > Microsoft (BingAds) > Seattle, USA



ex-Head of Lab > St. Petersburg Electrotechnical University > Russia

## | AGENDA |

- > Do we need new tools for ML?
- > MLFlow
- > Git-LFS
- > DVC
- > Conclusion



| 01 |

> Do we need new tools for ML?

> MLFlow

> Git-LFS

> DVC

> Conclusion





**Andrew Ng**  @AndrewYNg · Jan 3



1/The rise of Software Engineering required inventing processes like version control, code review, agile, to help teams work effectively. The rise of AI & Machine Learning Engineering is now requiring new processes, like how we split train/dev/test, model zoos, etc.

 50

 1.1K

 3.5K



# | PROBLEM 1: ML IS SLOW |

## MY MODEL'S TRAINING

~~"MY CODE'S COMPILING."~~



@ xkcd

Solution: custom ML PIPELINES

## | PROBLEM 2: ML IS METRICS DRIVEN |

>> EXPERIMENT = CODE + OUTPUTS

Outputs include **metrics and graphs** *AUC, etc.*

**Solution:** metrics tracking

	A	B	C	D	E	F
1	<b>Date</b>	<b>Alpha</b>	<b>L1_ratio</b>	<b>mea</b>	<b>r2</b>	<b>mse</b>
2	2019-04-03 1:00 PM	1	1	0.649	0.04	0.862
3	2019-04-03 4:00 PM	1	0.5	0.648	0.046	0.859
4	2019-04-04 9:00 AM	1	0.2	0.628	0.125	0.823
5	2019-04-04 11:00 AM	1	0	0.619	0.176	0.799

# | PROBLEM 3: MESS WITH DATA ARTIFACTS |

>> EXPERIMENT = CODE + OUTPUTS + DATASET

Source code, Datasets, ML models

Solution: connect data to code

	A	B	C	D	E	F	G	H
1	<b>Date</b>	<b>Dataset</b>	<b>Model</b>	<b>Alpha</b>	<b>L1_ratio</b>	<b>mea</b>	<b>r2</b>	<b>mse</b>
2	2019-04-03 1:00 PM	data_v2	model_v7.p	1	1	0.649	0.04	0.862
3	2019-04-03 4:00 PM	data_v2	model_v7_l1-05.p	1	0.5	0.648	0.046	0.859
4	2019-04-04 9:00 AM	data_v2_upd_May	model_v7_l1-02_d3.p	1	0.2	0.628	0.125	0.823
5	2019-04-04 11:00 AM	data_v2_upd_May	model_v7_l1-zero_d3.p	1	0	0.619	0.176	0.799



## SUMMARY OF DIFFERENCES

Software engineering	Data science \ ML
Source code version control	Code versioning Versioning of datasets, ML models, ML pipelines and connect data to code
Code review	Metrics tracking and visualization
Agile methodology	-\_(\ツ)\_/-

| 02 |

> Do we need new tools for ML?

> **MLFlow**

> Git-LFS

> DVC

> Conclusion



## | MLFLOW INTRO |

Platform for the machine learning lifecycle

- > Tracking
- > Project
- > Models

```
$ pip install mlflow
```

## | MLFLOW TRACKING |

```
from mlflow import log_metric, log_param, log_artifact
log_param("lr", 0.03)
log_metric("loss", curr_loss)
log_artifact("loss")
```

```
$ mlflow ui
```

# | MLFLOW TRACKING UI |

Date	User	Source	Version	Parameters		Metrics		
				alpha	l1_ratio	mae	r2	rmse
2018-06-04 23:00:10	mlflow	train.py	05e956	1	1	0.649	0.04	0.862
2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.5	0.648	0.046	0.859
2018-06-04 23:00:10	mlflow	train.py	05e956	1	0.2	0.628	0.125	0.823
2018-06-04 23:00:09	mlflow	train.py	05e956	1	0	0.619	0.176	0.799

## MLFLOW SUMMARY

Feature	Result	Comment
Versioning ML models	+	
Versioning datasets	-	
Versioning ML pipelines	-	
Connecting data and code	-/+	Manual only
Tracking metrics	+	
Visualize metrics	+	

| 03 |

- > Do we need new tools for ML?
- > MLFlow
- > **Git-LFS** Git Large File Storage
- > DVC
- > Conclusion



## | GIT-LFS INTRO |

### > Install

```
$ brew install git-lfs  
$ git lfs install
```

### > Specify data-files type in a Git repository

```
$ git lfs track '*.p'  
$ git add .gitattributes
```



## | GIT-LFS ADD DATA FILES |

```
$ python mytrain.py # your code generates mymodel.p  
$ git add mytrain.py mymodel.p  
$ git commit -m 'Decay was added'  
$ git push
```

```
Uploading LFS objects: 100% (1/1),  
56 MB | 3.2 MB/s, done
```

## | GIT-LFS RETRIEVE DATA FILES |

```
$ git clone https://github.com/dmpetrov/my-lfs-repo
$ cd my-lfs-repo
$ du -sh mymodel.p # data file does not contain data yet
4.0K  mymodel.p
$ git pull

Downloading LFS objects: 75% (3/4),
44 MB | 4.5 MB/s
```

## | GIT-LFS PROS/CONS |

### > PROS

- > Simple, like Git

### > CONS

- > Limited by data size <2Gb, <500Mb even better

- > Not every Git server supports Git-LFS

- > No ML\Data Science specific

## | GIT-LFS SUMMARY |

Feature	Result	Comment
Versioning ML models	+	Limited by size
Versioning datasets	-/+	
Versioning ML pipelines	-	
Connecting data and code	+	
Tracking metrics	-	
Visualize metrics	-	

| 04 |

- > Do we need new tools for ML?
- > MLFlow
- > Git-LFS
- > **DVC** Data Version Control
- > Conclusion



## | DVC INTRO |

Website: <http://DVC.org>

> Install

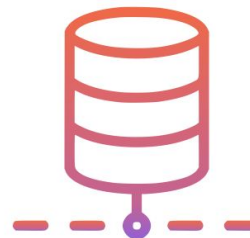
```
$ pip install dvc  
$ dvc init
```

> Git-like tool no infrastructure is required

Code  
Github, Gitlab, any Git Server

Data  
S3, Azure, Google Cloud, SSH

Remote



**git push**

**dvc push**

**git pull**

**dvc pull**

Local



code



model.pkl.dvc  
1KB



model.pkl  
500MB

## | DVC ADD DATA FILES |

> Push data to storage

```
$ dvc add data.xml  
$ dvc push
```

> Push meta information to Git server

```
$ git add .gitignore data.xml.dvc  
$ git commit -m "add source data to DVC"  
$ git push
```



## | DVC RETRIEVE DATA |

```
$ git clone https://github.com/dmpetrov/my-dvc-repo
$ cd my-dvc-repo
$ dvc pull
...

$ du -sh data.xml
7G data.xml
```

## | DVC PARTIAL DATA RETRIEVING |

```
$ git clone https://github.com/dmpetrov/my-dvc-repo
$ cd my-dvc-repo
$ dvc pull train.dvc
...

$ du -sh cnn_model.p
54M cnn_model.p
```

# | DVC CHECKOUT |

> Checkout data

```
$ git checkout vgg16_exp2  
$ dvc checkout
```



## | DVC SPEED |

> Copy 50G directory with millions of images ~10 min

What about DVC?

```
$ git checkout image_update_20190310
```

```
$ time dvc checkout
```

```
real    0m12.958s
```

```
user    0m11.567s
```

```
sys     0m1.725s
```

# | DVC PIPELINES |

```
$ dvc add data/data.xml
$ dvc run -d src/prepare.py -d data/data.xml -o data/prepared \
    python src/prepare.py data/data.xml
$ dvc run -d src/featurization.py -d data/prepared -o
data/features \
    python src/featurization.py data/prepared data/features
$ dvc run -d src/train.py -d data/features -o model.pkl \
    python src/train.py data/features model.pkl
```

# | DVC PIPELINES: SHOW |

```
$ dvc pipeline show --ascii train.dvc --commands
```

```
+-----+  
| python src/prepare.py data/data.xml |  
+-----+
```

```
*  
*  
*
```

```
+-----+  
| python src/featurization.py data/prepared data/features |  
+-----+
```

```
*  
*  
*
```

```
+-----+  
| python src/train.py data/features model.pkl |  
+-----+
```

## | DVC PIPELINES: REPRODUCIBILITY |

### > Reproduce your project

```
$ dvc repro
```

### > Reproduce

```
$ dvc repro train.dvc
```

### > Version DVC pipeline

```
$ git add train.dvc
```

```
$ git commit -m 'Reproduce with dataset update 2019-05-02'
```

## | DVC SUMMARY |

Feature	Result	Comment
Versioning ML models	+	
Versioning datasets	+	
Versioning ML pipelines	+	
Connecting data and code	+	
Tracking metrics	-/+	Final metrics only
Visualize metrics	-	



| 05 |

> Do we need new tools for ML?

> Git-LFS

> MLFlow

> DVC

> Conclusion



## SUMMARY

Feature	MLFLOW	Git-LFS	DVC
Versioning ML models	+	+	+
Versioning datasets	-	-/+	+
Versioning ML pipelines	-	-	+
Connecting data and code	-/+	+	+
Tracking metrics	+	-	-/+
Visualize metrics	+	-	-

# | THE WORLD IS CHANGING |

Data science as different from software  
as software was different from hardware

Nick Elprin, Domino Data Lab

Hardware

|

Waterfall

Software

|

Agile

DS/ML

|

-\\_(\ツ)\\_/-

# | HOW TO DESIGN OUR FUTURE |

| Think about processes

| Try new ML tools

| Share your feedback

# | THANK YOU |



## > Questions

Twitter [@FullStackML](#)

Email [dmitry@iterative.ai](mailto:dmitry@iterative.ai)

## > Actions

Visit [dvc.org](https://dvc.org)

Star [github.com/iterative/dvc](https://github.com/iterative/dvc)

| 06 |

> Appendix



**Andrew Ng**  @AndrewYNg · Jan 3



2/I'm also seeing many AI teams use new processes that haven't been formalized or named yet, ranging from how we write product requirement docs to how we version data and ML pipelines. This is an exciting time for developing these ideas!



17



147



998

