# ML MODELS AND DATASET VERSIONING

Kurian Benoy

PYCON INDIA Chennai, 2019

# $ WHOAMI

Open source contributor

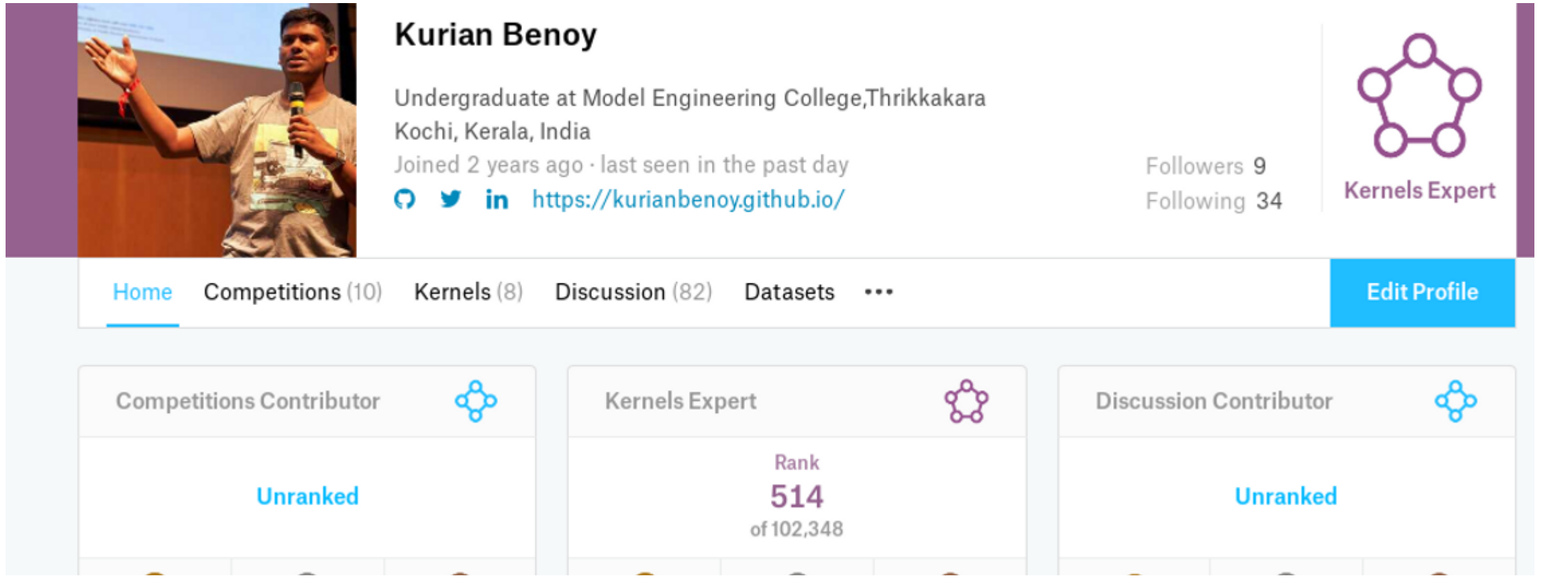
FOSSASIA OpenTechNights Winner

# $ WHOAMI

Open source contributor

FOSSASIA OpenTechNights Winner

Kaggle Expert in Kernels



# $ WHOAMI

Open source contributor

FOSSASIA OpenTechNights Winner

Kaggle Expert

Final Year BTech student @MEC

# OUTLINE

- Start up Adventures
- Challenges
- Model and Dataset versioning
- How I discovered DVC?
- Use case: Versioning dogs and Cats
- Conclusion

# Startup Adventures

# CHALLENGE 1: ML IS SLOW

# CHALLENGE 2: WORKING WITH ML PROJECTS

- Most software products take a few seconds to execute.

```
$ git clone project-repo
$ pip install -r requirements.txt
```

## Data

- Schema
- Sampling over Time
- Volume

+

## Model

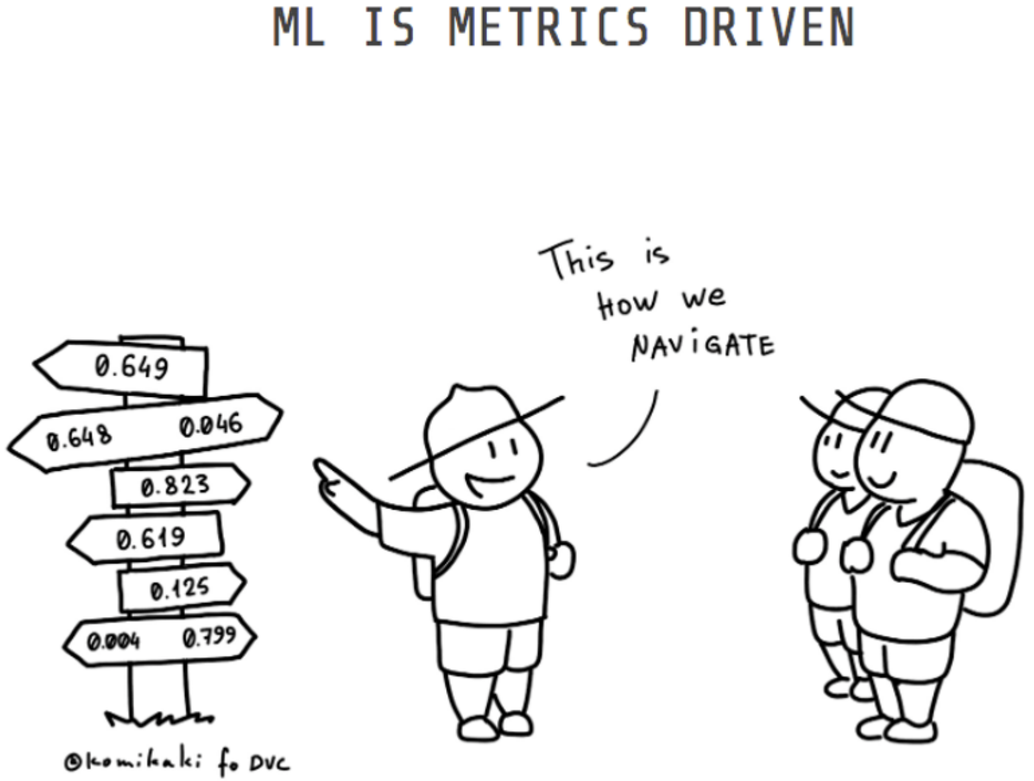
- Algorithms
- More Training
- Experiments

+

## Code

- Business Needs
- Bug Fixes
- Configuration

# CHALLENGE 3: METRIC DRIVEN



# CHALLENGE 4:
# NOT ABLE TO USE GIT


- git not suitable for projects > 1GB
- git clone becomes slow

**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

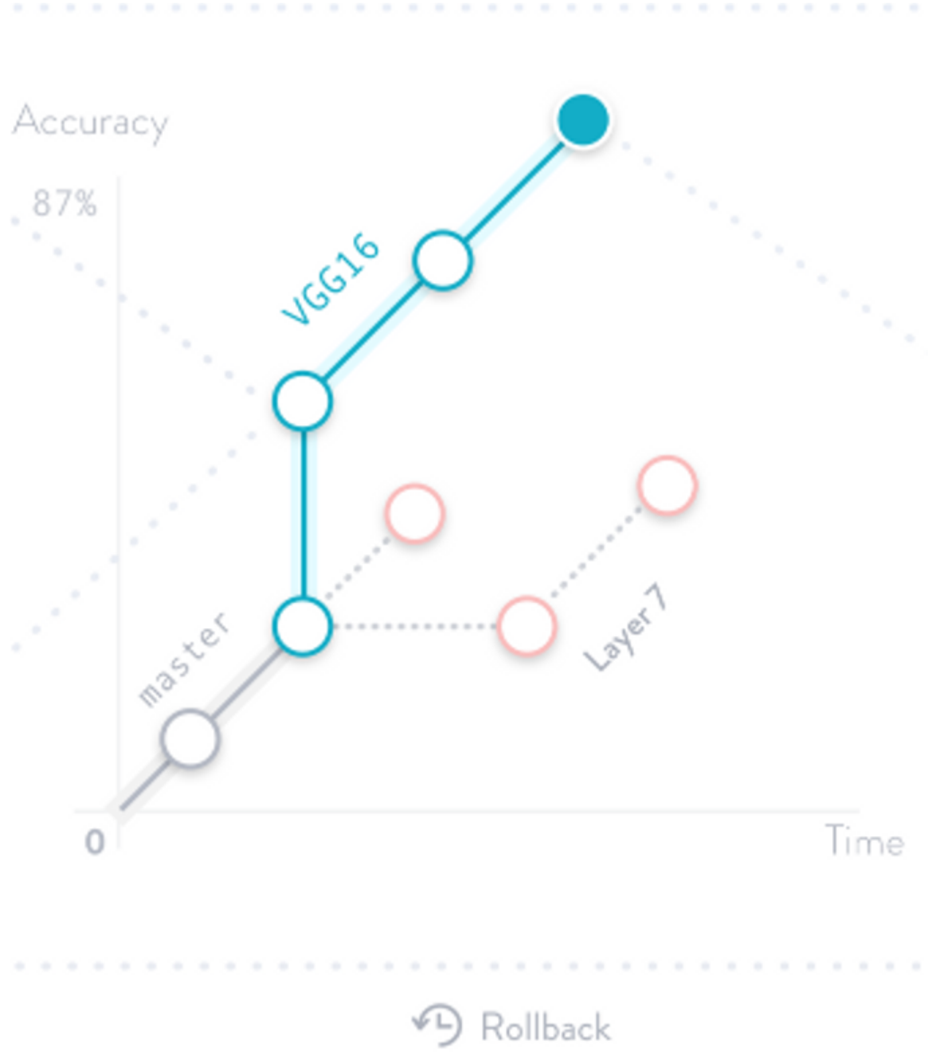
## SUMMARY OF DIFFERENCES

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

# MODEL VERSIONING

# TRACKING EXPERIMENTS

## TRACKING METRICS



# Why Model Versioning?

> To keep track of experiments

> Choose the best ideas

>> EXPERIMENTS = CODE + OUTPUTS

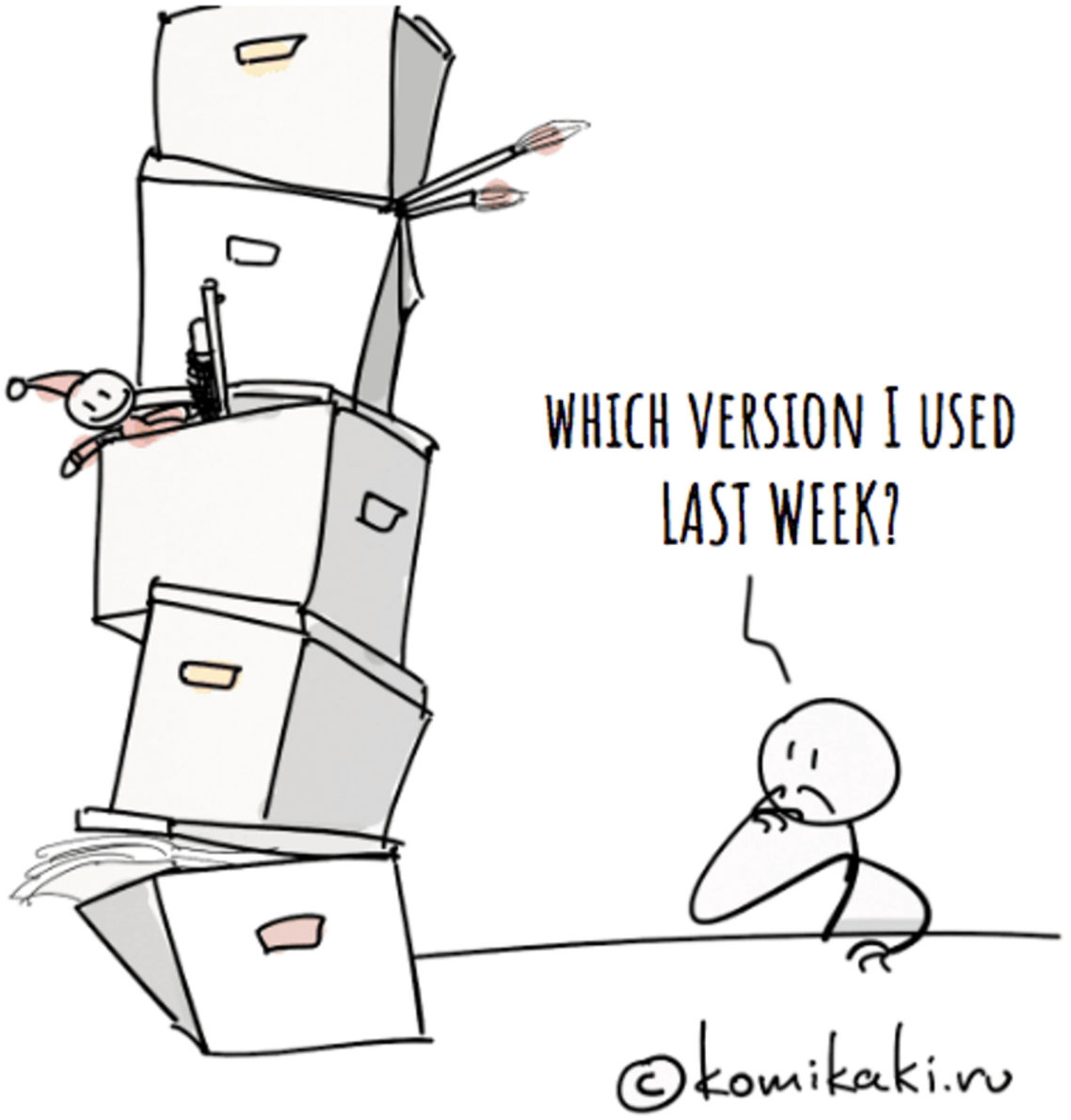
Models are outputs

# DATASET VERSIONING


WAYMO


4 TB/day

# DATASETS MANAGEMENT



# Why Dataset management?

> Moving Datasets around

> Datasets evolve, so versioning required

>> EXPERIMENTS = CODE + DATA + OUTPUTS

Source code, Datasets, ML Models

# HOW I DISCOVERED DVC

# DATA VERSION CONTROL(DVC)

" For the Datascientist, by the Datascientist"

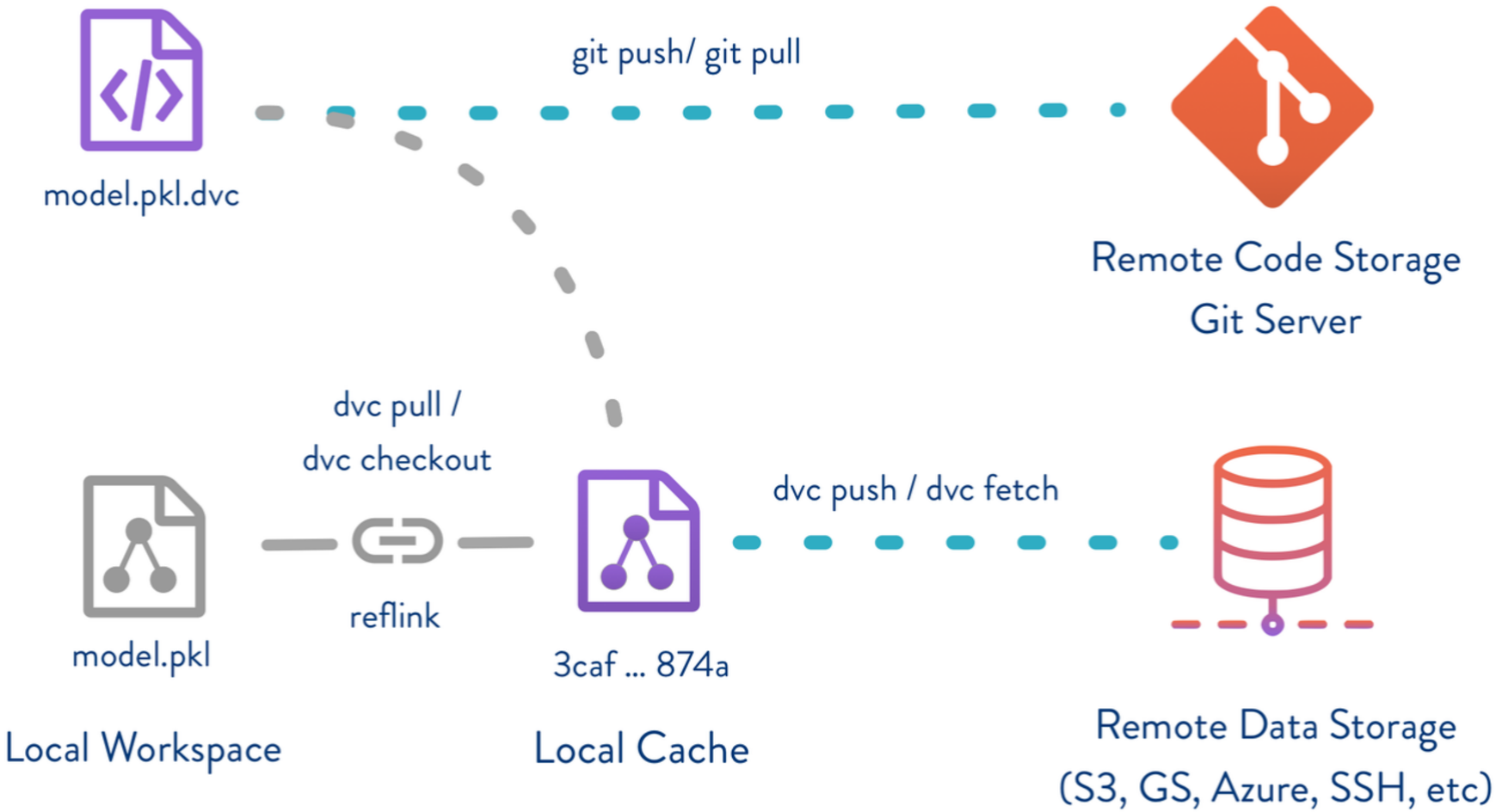
## DVC SUMMARY

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

# VERSIONING CATS & DOGS

# DEMO TIME

# DVC WORKFLOW

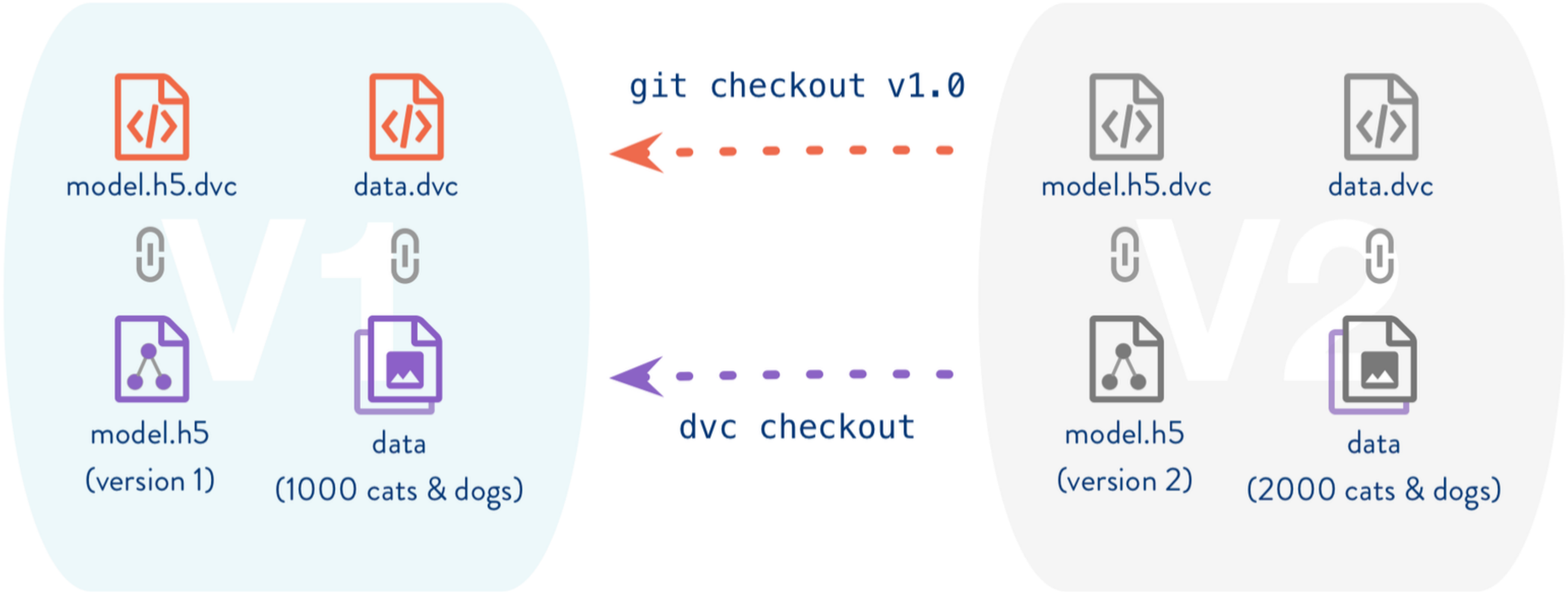


# Tracking data

1. Tracking 1000 cats and dogs

2. Add 1000 more labelled images of cats & dogs

# SWITCHING VERSIONS



# CONCLUSION

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

Nick Elprin,
CEO, DominoLabs.

- Think about your processes(ML projects)

- Think about your processes
- Try to version control for your projects

- Think about your processes
- Try to version control for your big projects
- Create the best practises from now on

# THANK YOU

- Twitter: kurianbenoy2
- Email : kurian.bkk@gmail.com

Speaker Deck: bit.ly/mlversion19

# APPENDIX

# Other Tools for versioning

ML Flow - Tracking Models, Metrics

Git-LFS - Tracking Large files

Jovian - JupyterNB based tracking

Neptune.Ml

Hangar Py - Versioning Tensor Data