ML MODELS AND DATASET VERSIONING

Kurian Benoy

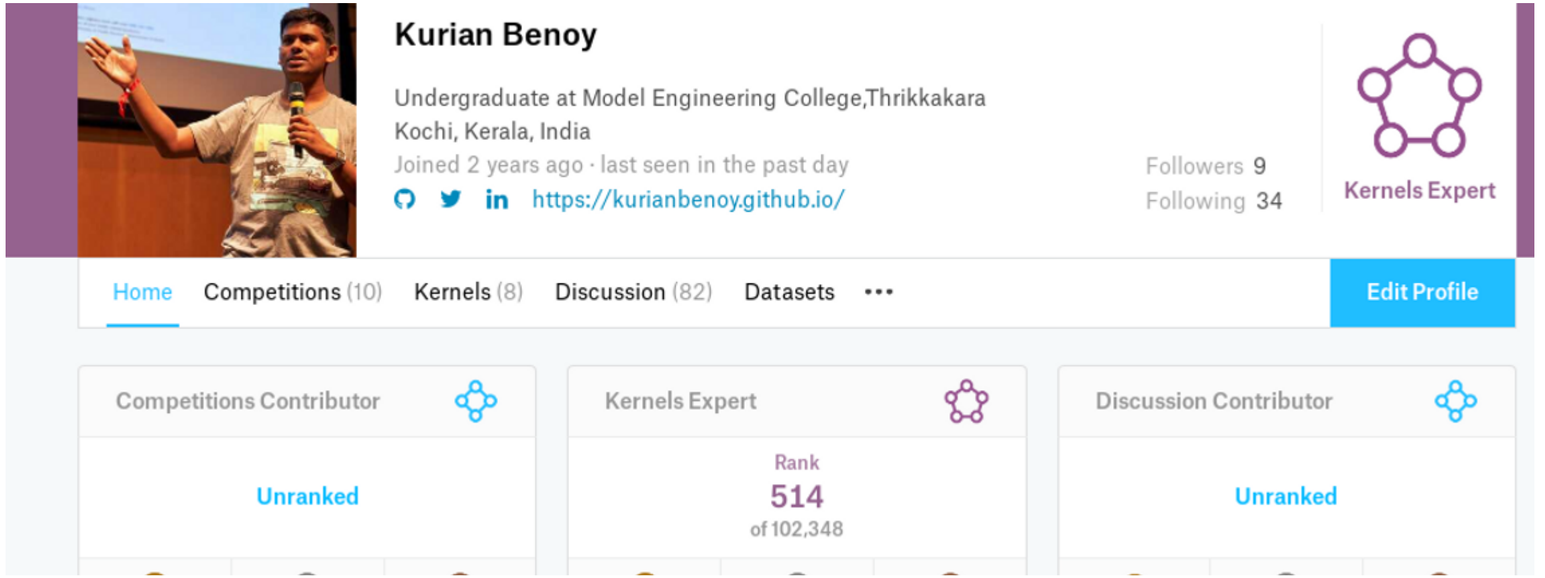
PYCON INDIA Chennai, 2019

$ 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

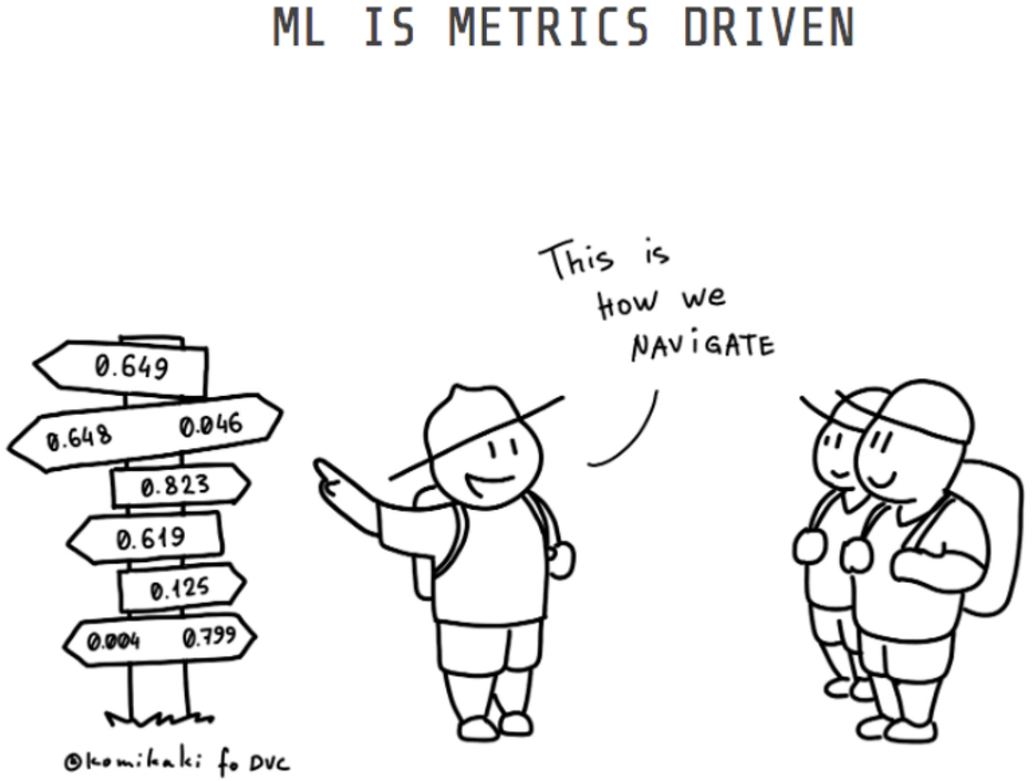
- Algorithms
- More Training
- Experiments

+

Code

- Business Needs
- Bug Fixes
- Configuration

CHALLENGE 3: METRIC DRIVEN



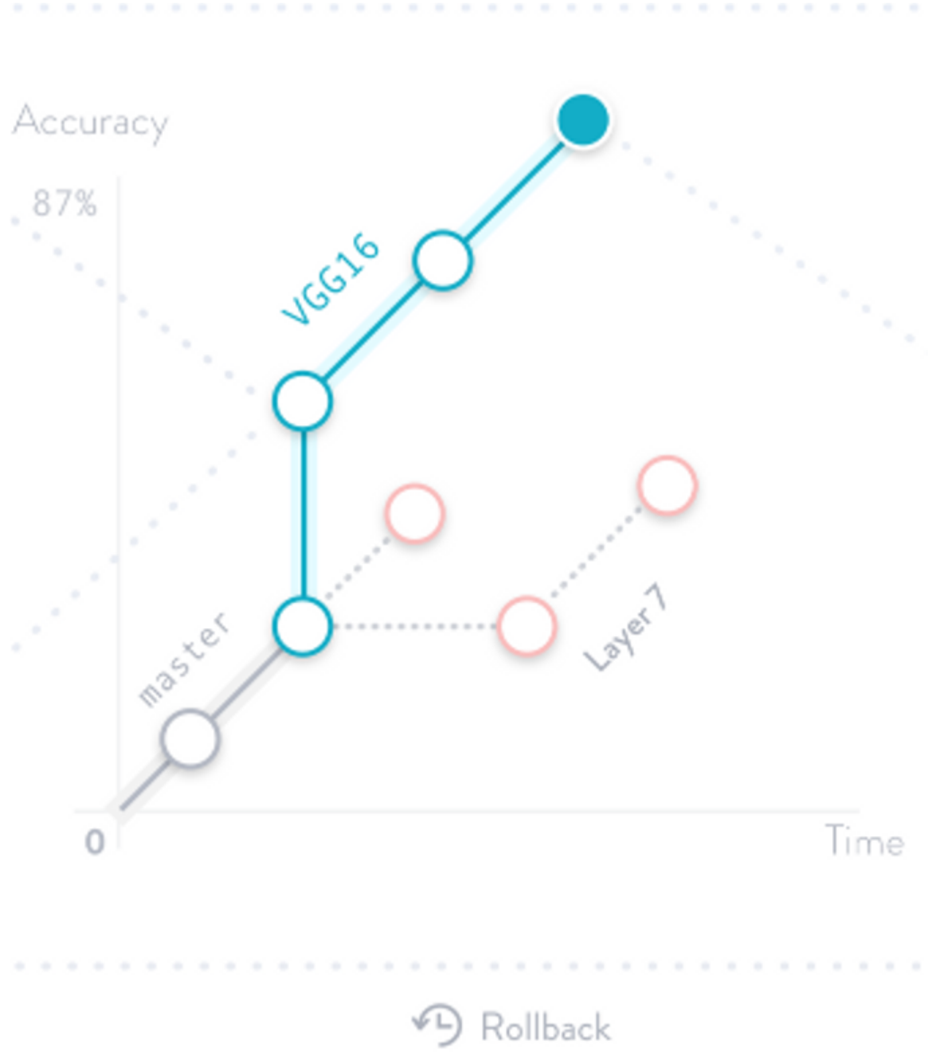
CHALLENGE 4:
NOT ABLE TO USE GIT

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

MODEL VERSIONING

TRACKING EXPERIMENTS

TRACKING METRICS



Why Model Versioning?

> To keep track of experiments

> Choose the best ideas

>> EXPERIMENTS = CODE + OUTPUTS

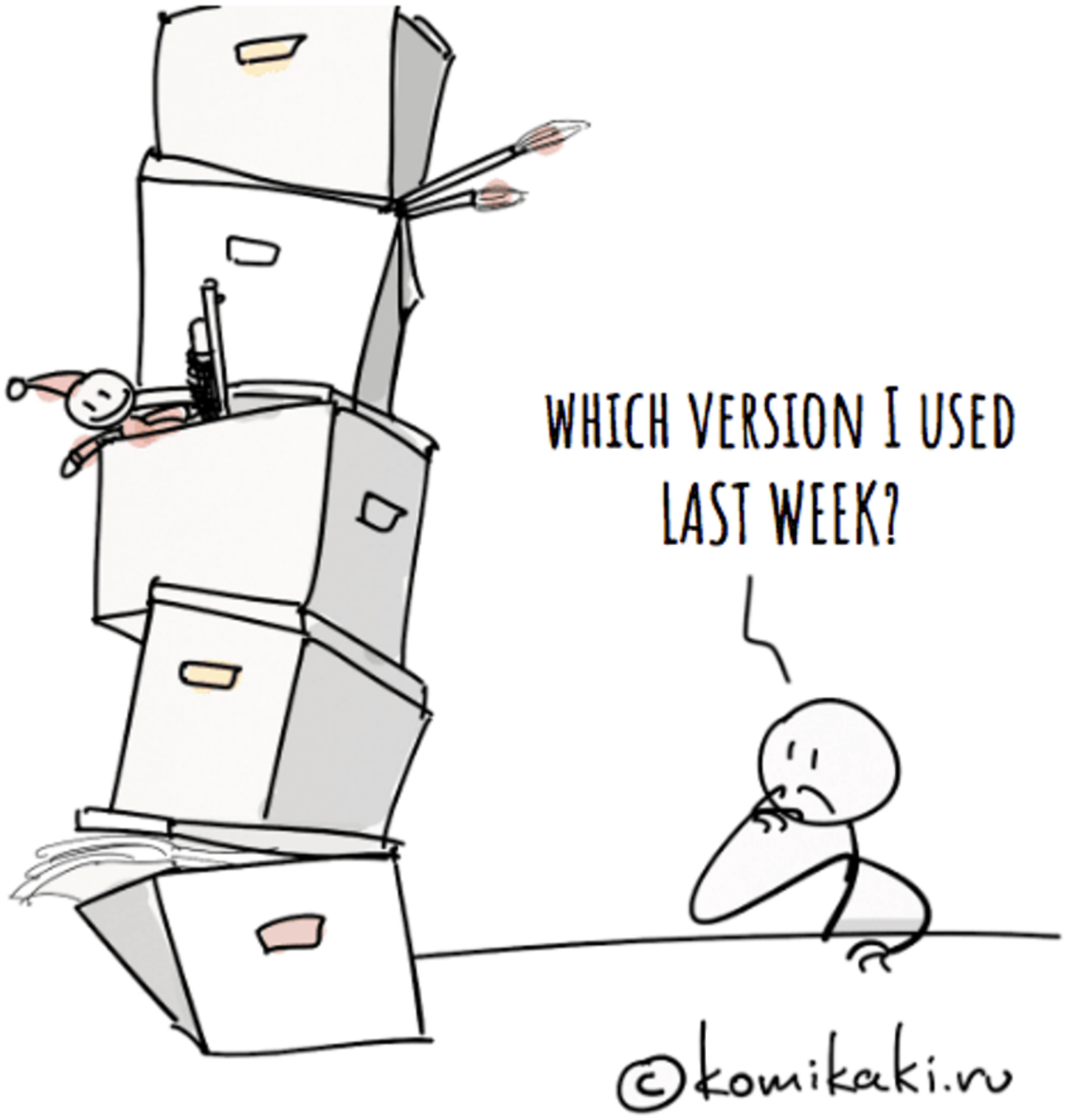
Models are outputs

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

HOW I DISCOVERED DVC

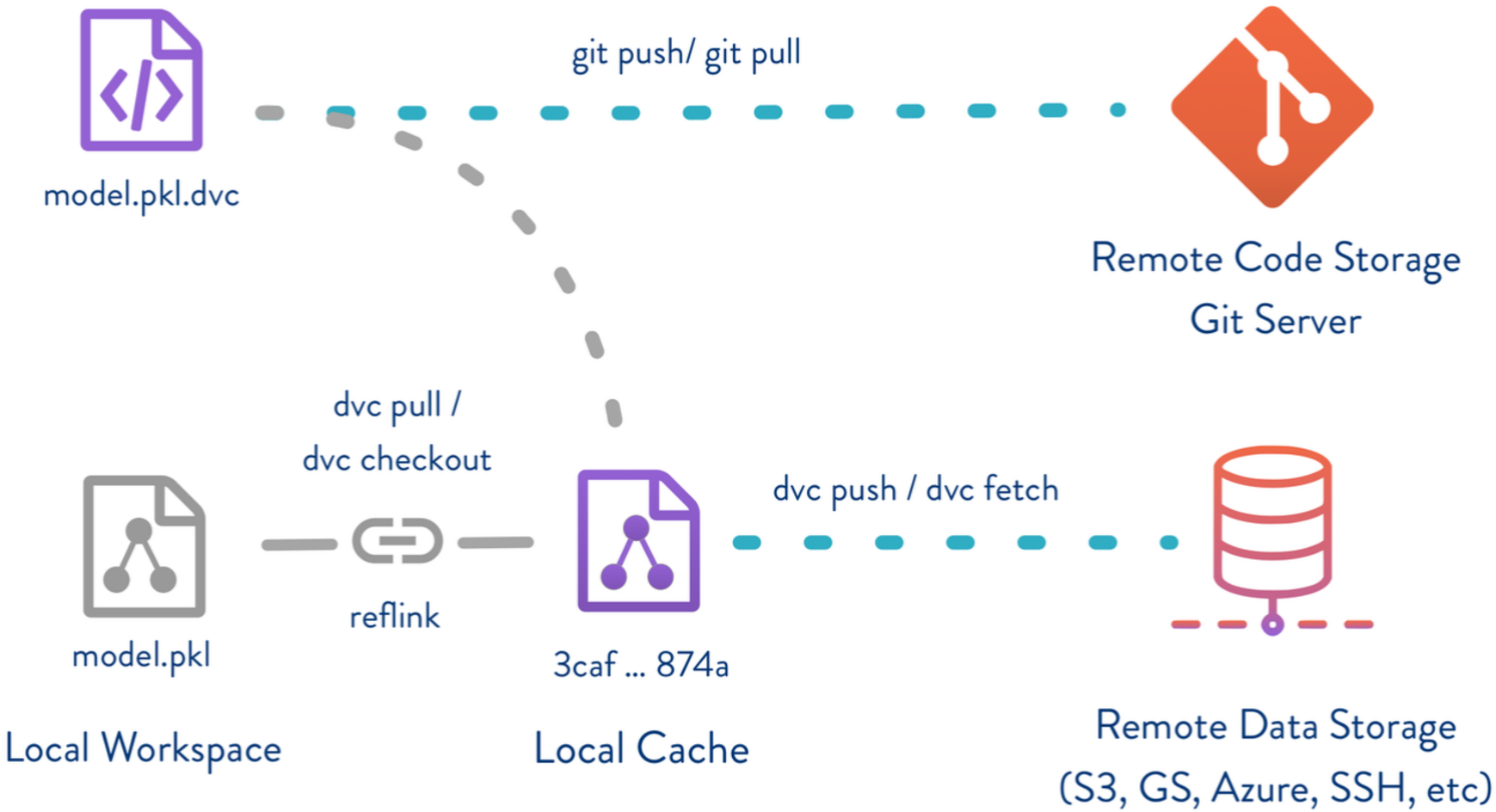
DATA VERSION CONTROL(DVC)

- Experiment and Dataset tracking
- Open-source(3500+ stars)
- Build to adopt the best practises of ML
- Works well with git
- Language and framework agnostic

VERSIONING CATS & DOGS

DEMO TIME

DVC WORKFLOW

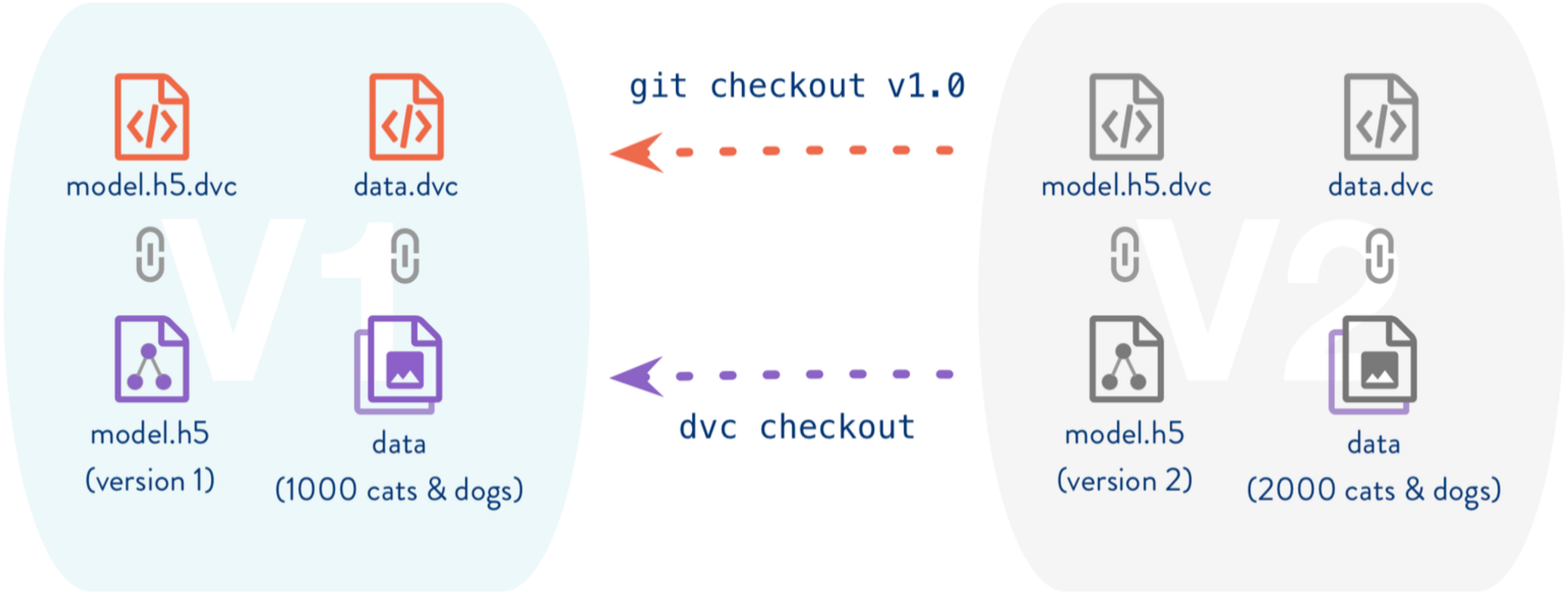


Tracking data

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

Try it out in your ML project!

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