

ML MODELS AND DATASET VERSIONING

Kurian Benoy



\$ WHOAMI

Open source contributor

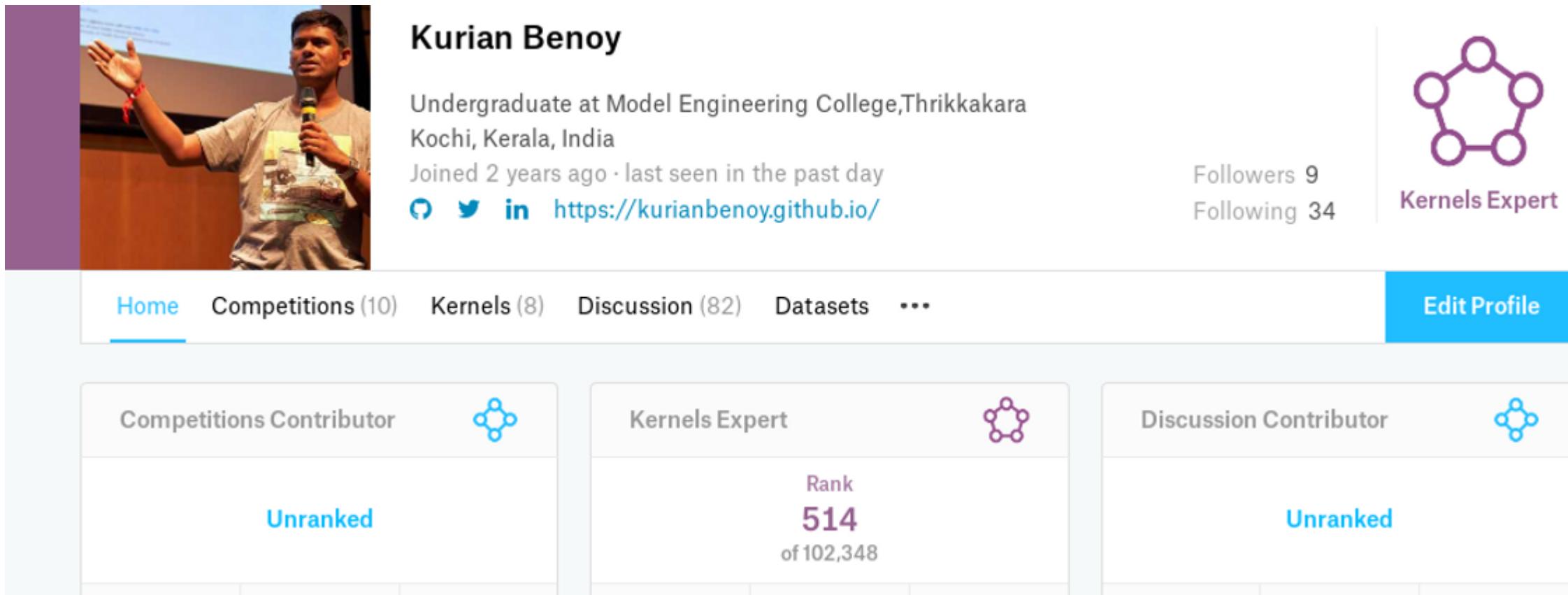
FOSSASIA OpenTechNights Winner

\$ WHOAMI

Open source contributor

FOSSASIA OpenTechNights Winner

Kaggle Expert in Kernels



A screenshot of a Kaggle profile page for Kurian Benoy. The profile features a photo of Kurian speaking at a podium. His name, "Kurian Benoy", is displayed in bold black text. Below his name is a bio: "Undergraduate at Model Engineering College,Thrikkakara Kochi, Kerala, India". It also shows he joined 2 years ago and was last seen in the past day. Social media links for GitHub, Twitter, and LinkedIn are provided, along with a link to his GitHub page: <https://kurianbenoy.github.io/>. On the right side, there's a purple icon of a neural network labeled "Kernels Expert". The profile stats show 9 followers and 34 following. Navigation links include Home (underlined), Competitions (10), Kernels (8), Discussion (82), Datasets, and an ellipsis. A blue "Edit Profile" button is located in the top right corner. Below the navigation, three cards show his expertise: "Competitions Contributor" (Unranked), "Kernels Expert" (Rank 514 of 102,348), and "Discussion Contributor" (Unranked).

Kurian Benoy

Undergraduate at Model Engineering College,Thrikkakara Kochi, Kerala, India

Joined 2 years ago · last seen in the past day

GitHub Twitter LinkedIn <https://kurianbenoy.github.io/>

Followers 9 Following 34

Kernels Expert

Home Competitions (10) Kernels (8) Discussion (82) Datasets ... Edit Profile

Competitions Contributor

Unranked

Kernels Expert

Rank 514 of 102,348

Discussion Contributor

Unranked

\$ 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

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



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

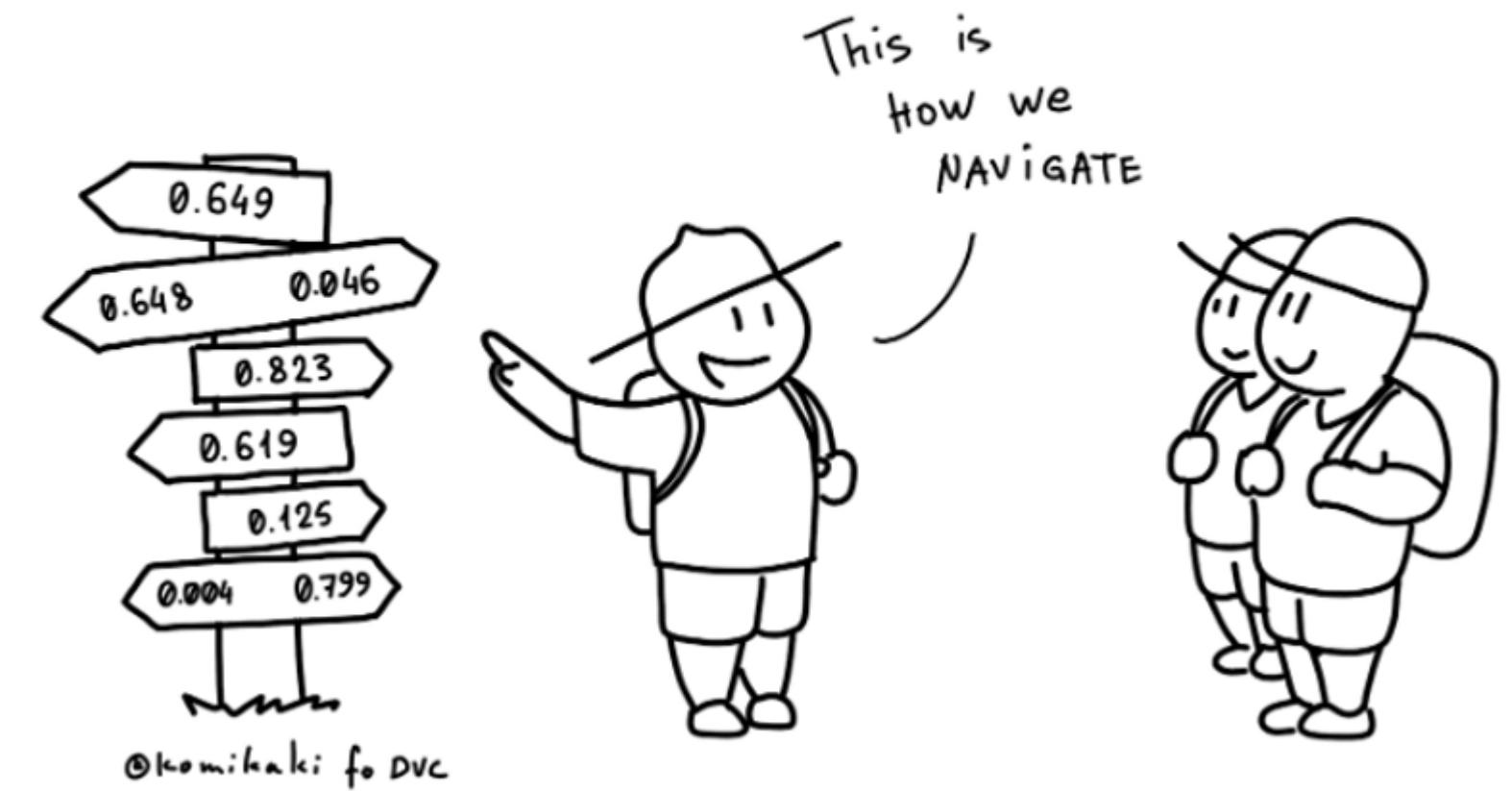
Business Needs

Bug Fixes

Configuration

ML IS METRICS DRIVEN

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



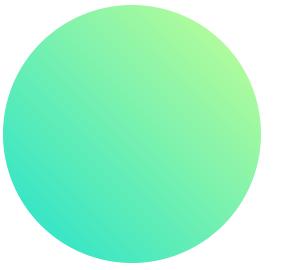
3.5K



SUMMARY OF DIFFERENCES

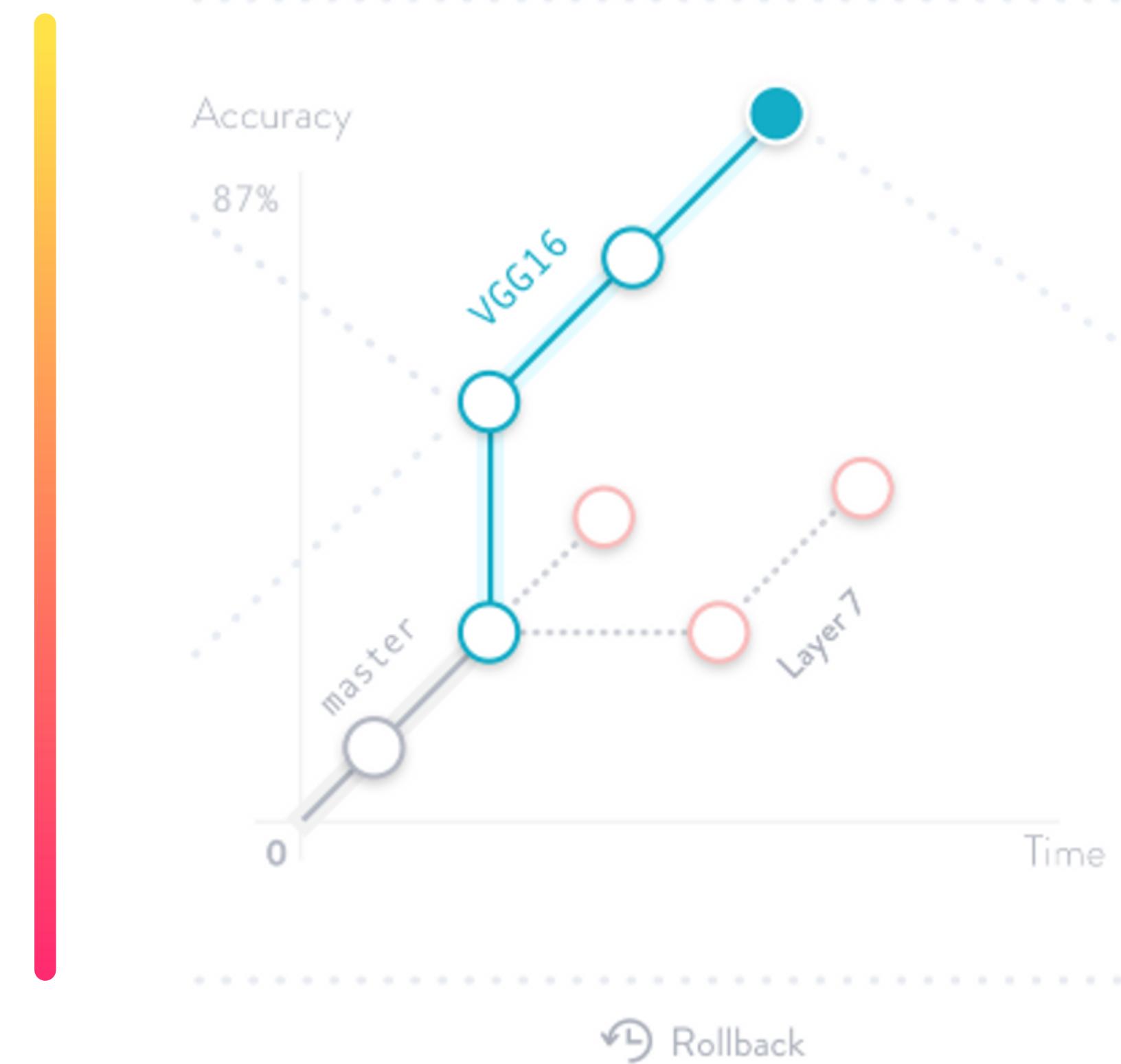
Software engineering	Data science \ ML
Source code version control	Code versioning Versioning of datasets, ML models, ML pipelines and connect data to code
Code review	Metrics tracking and visualization
Agile methodology	-_(ツ)_/-

MODEL VERSIONING



TRACKING EXPERIMENTS

TRACKING METRICS

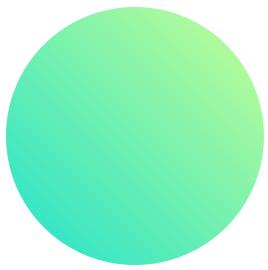


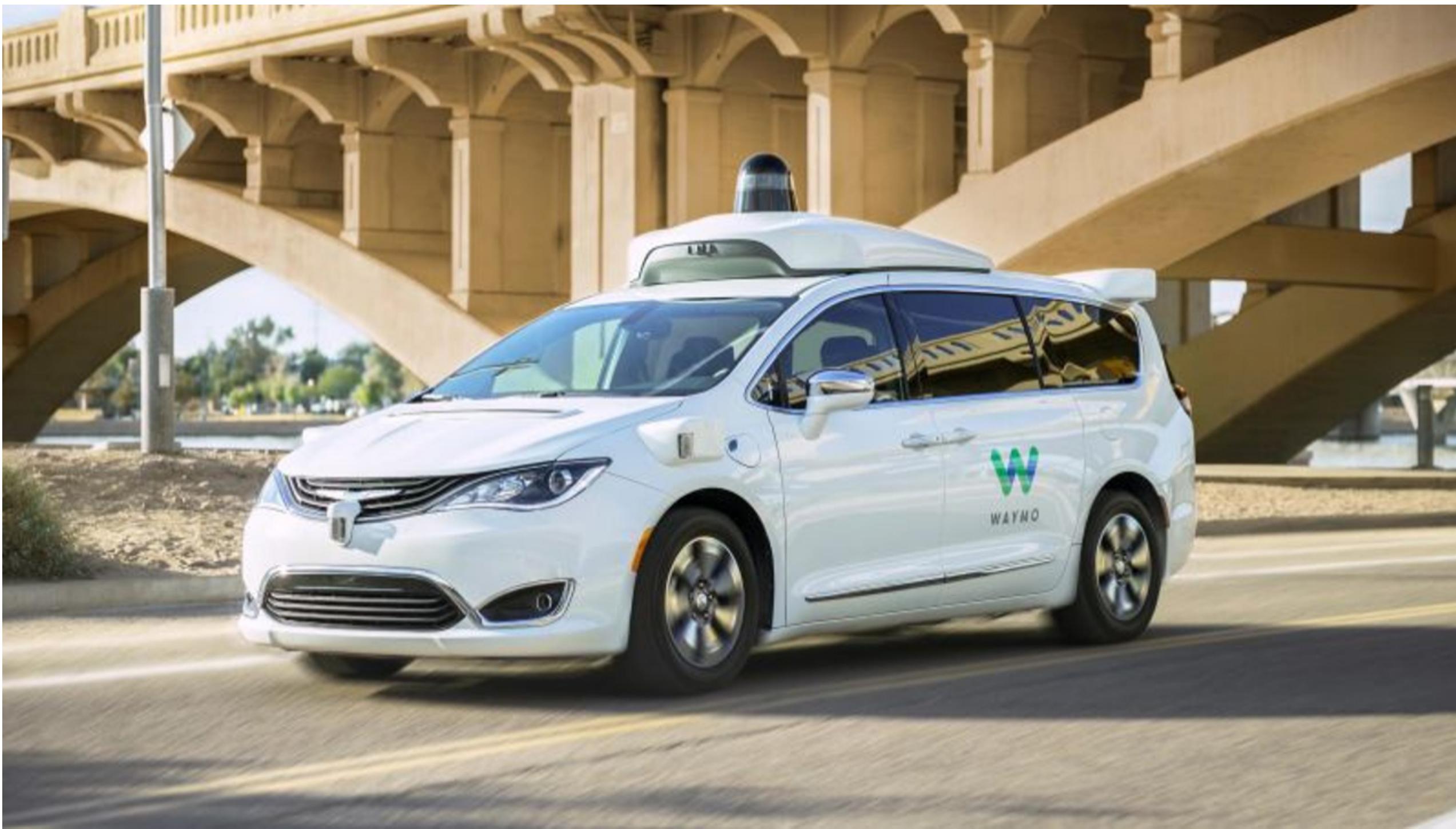
why Model Versioning?

- › To keep track of experiments
 - › Choose the best ideas
- ›› EXPERIMENTS = CODE + OUTPUTS

Models are outputs

DATASET VERSIONING

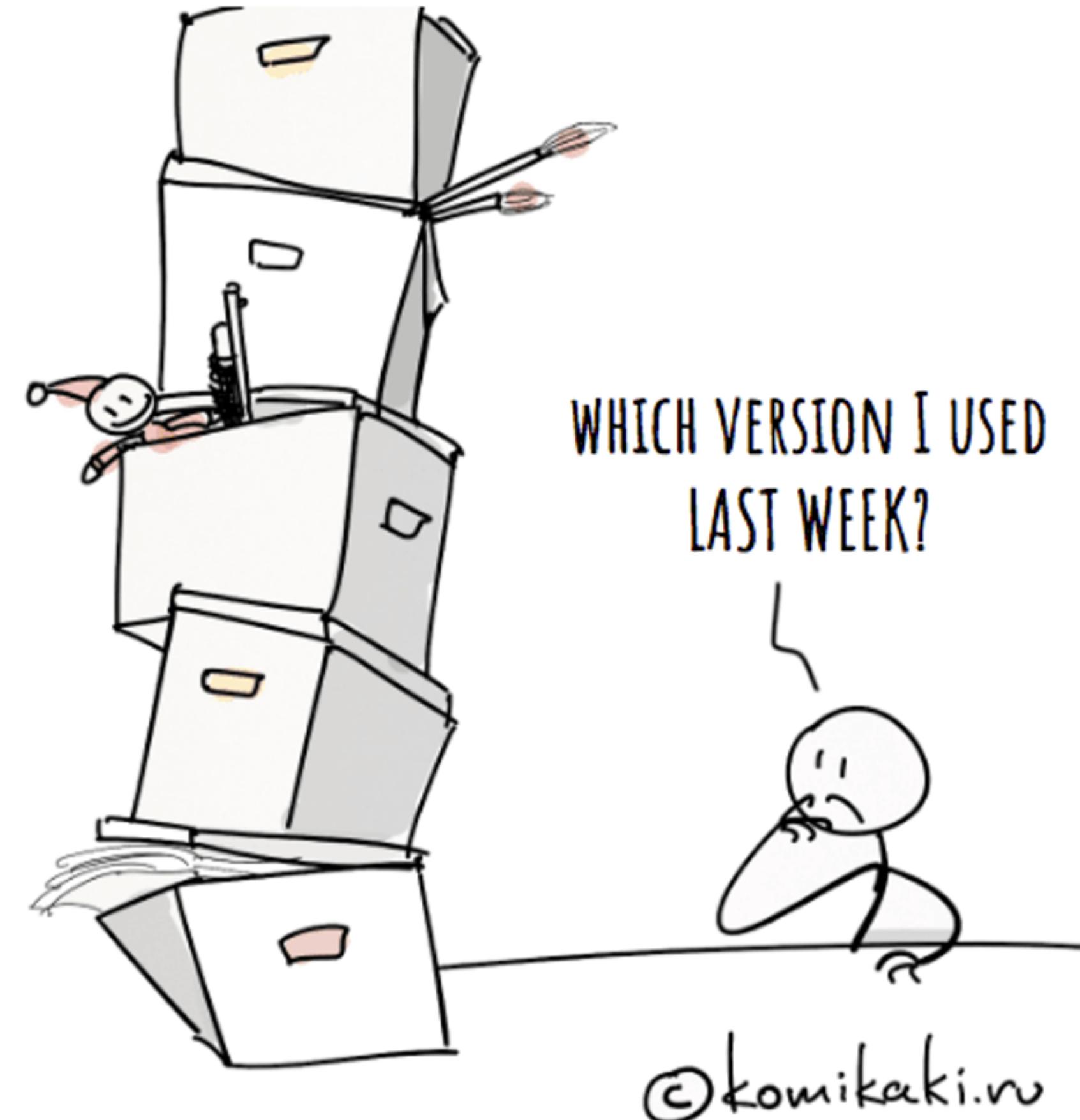






4 TB/day

DATASETS MANAGEMENT

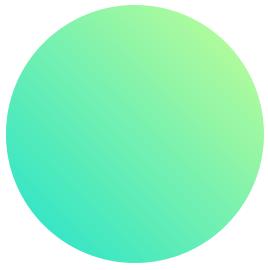


Why Dataset management?

- > 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"**



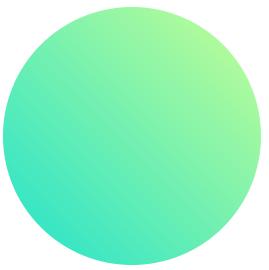
IDVC SUMMARY

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

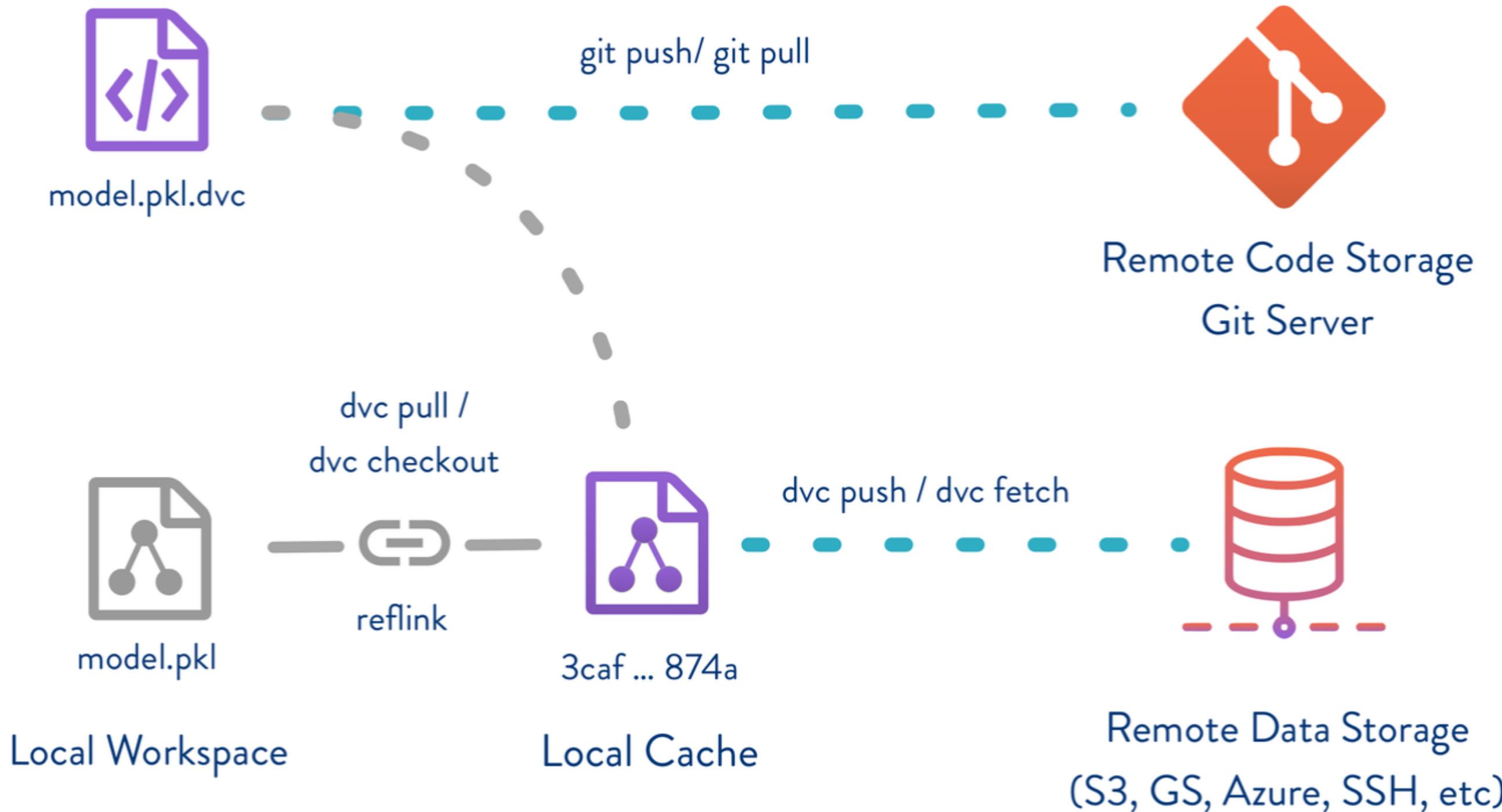
VERSIONING CATS & DOGS



DEMO TIME



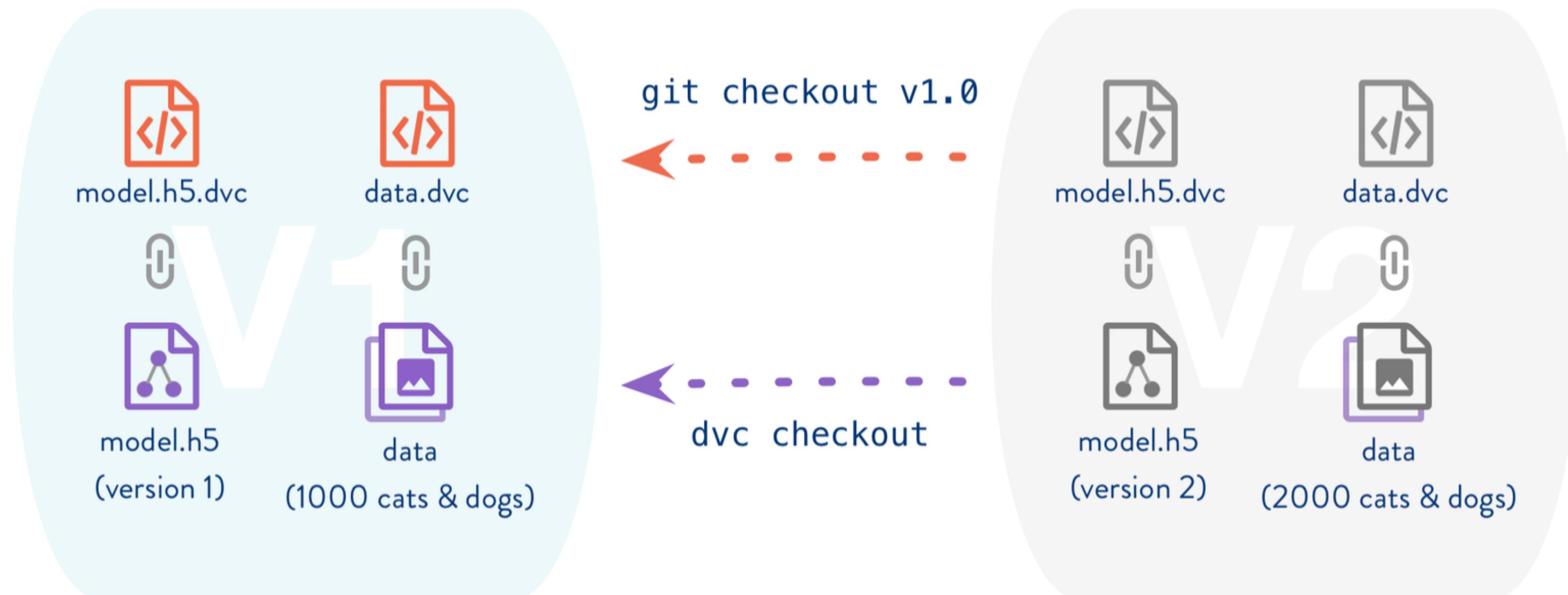
DVC WORKFLOW



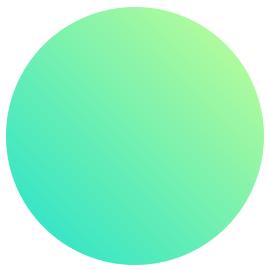
Tracking data

- 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

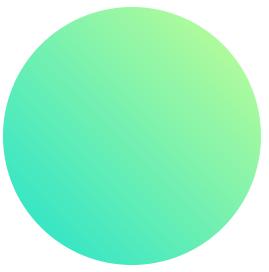
- Think about your processes
- Try to version control for your big projects
- Create the best practises from now on

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