

# ML MODELS AND DATASET VERSIONING

Kurian Benoy



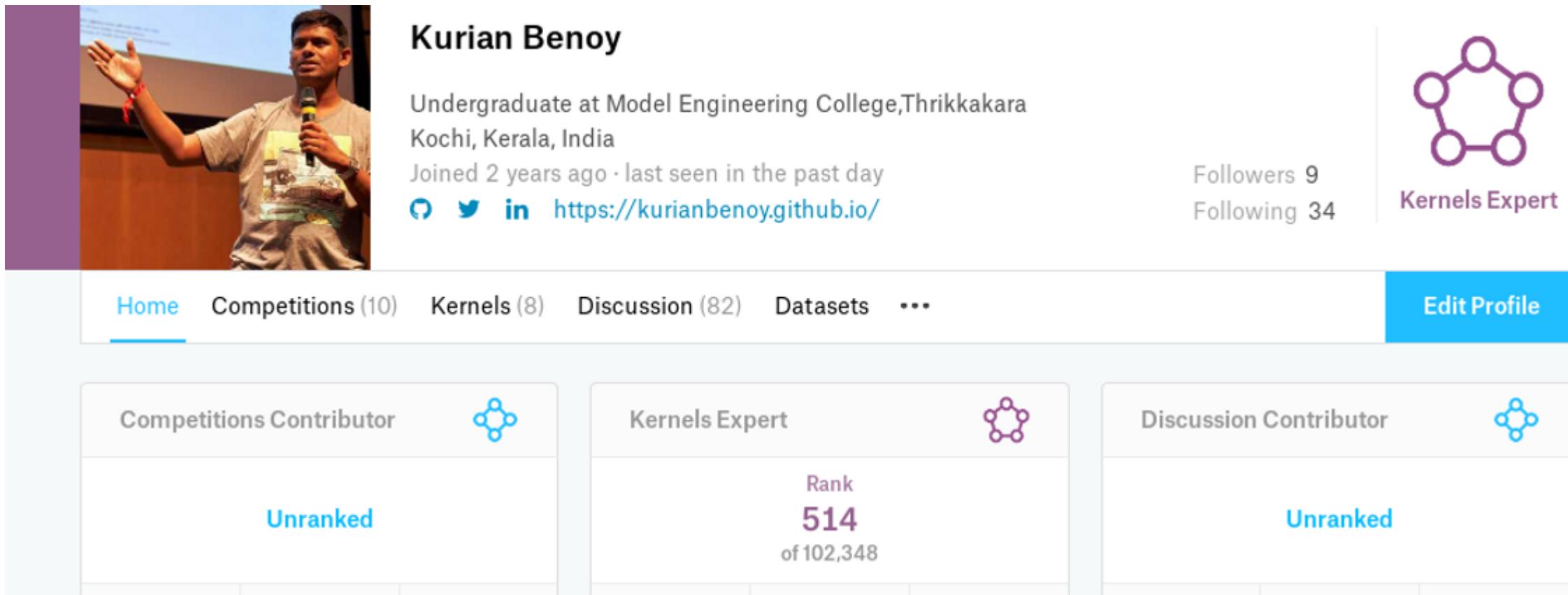
# \$ WHOAMI

---

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 GitHub URL: <https://kurianbenoy.github.io/>. On the right side, there's a purple "Kernels Expert" badge with a neural network icon. Below the bio, the follower count is 9 and the following count is 34. At the bottom of the profile, there are tabs for Home (selected), Competitions (10), Kernels (8), Discussion (82), Datasets, and an ellipsis (...). A blue "Edit Profile" button is located on the far right. Below the tabs, three cards show his expertise levels: "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

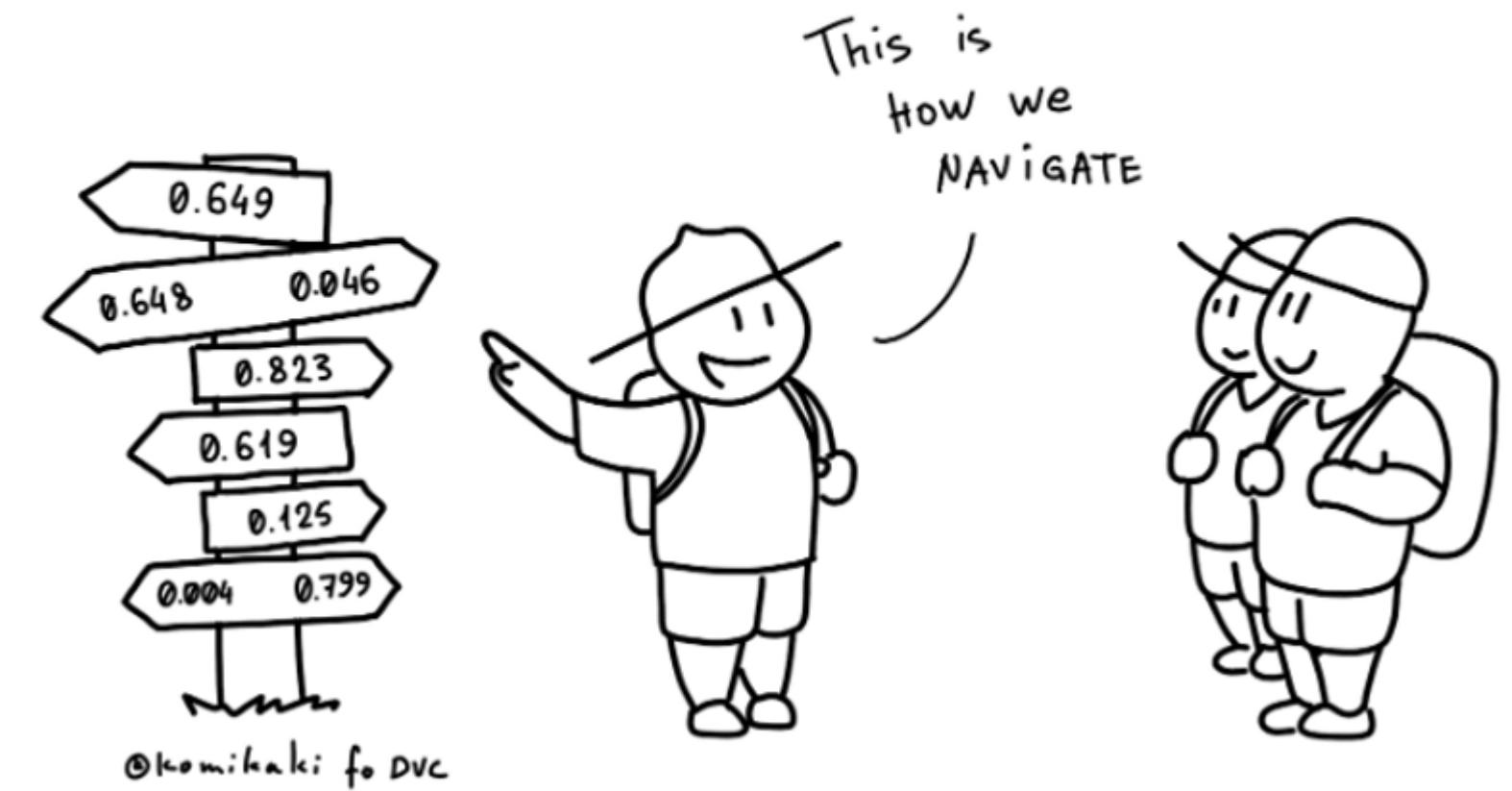
Business Needs

Bug Fixes

Configuration

ML IS METRICS DRIVEN

# CHALLENGE 3: METRIC DRIVEN

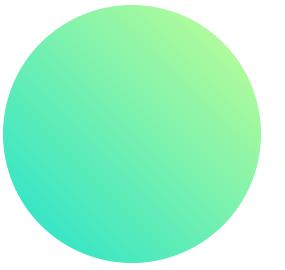


## CHALLENGE 4: NOT ABLE TO USE GIT



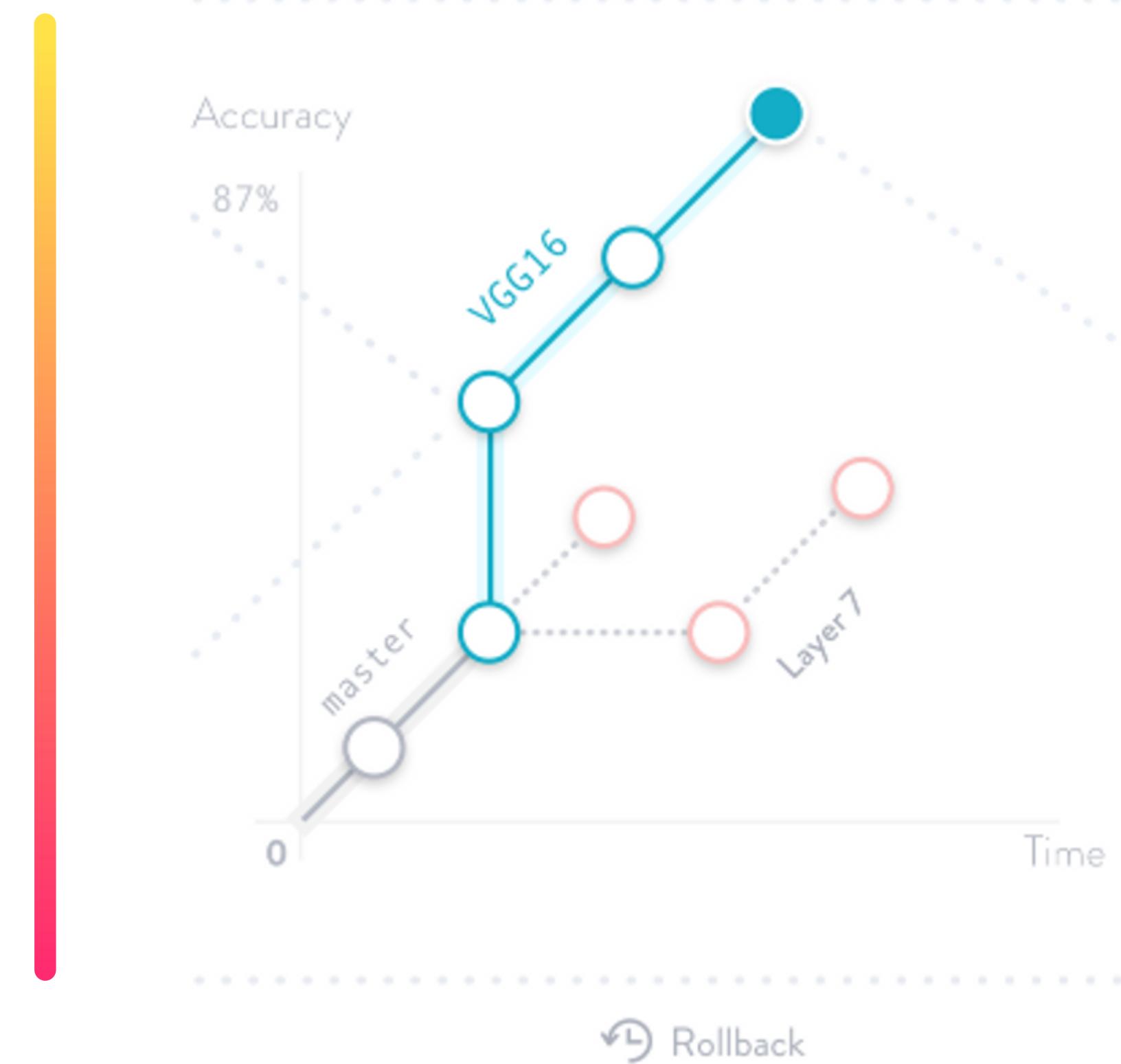
- git not suitable for projects > 1GB
- git clone becomes slow

# **MODEL VERSIONING**



# TRACKING EXPERIMENTS

TRACKING METRICS

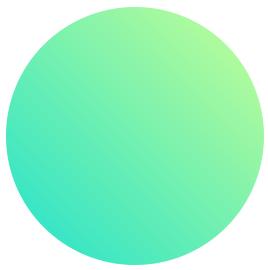


# why Model Versioning?

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

Models are outputs

# DATASET VERSIONING

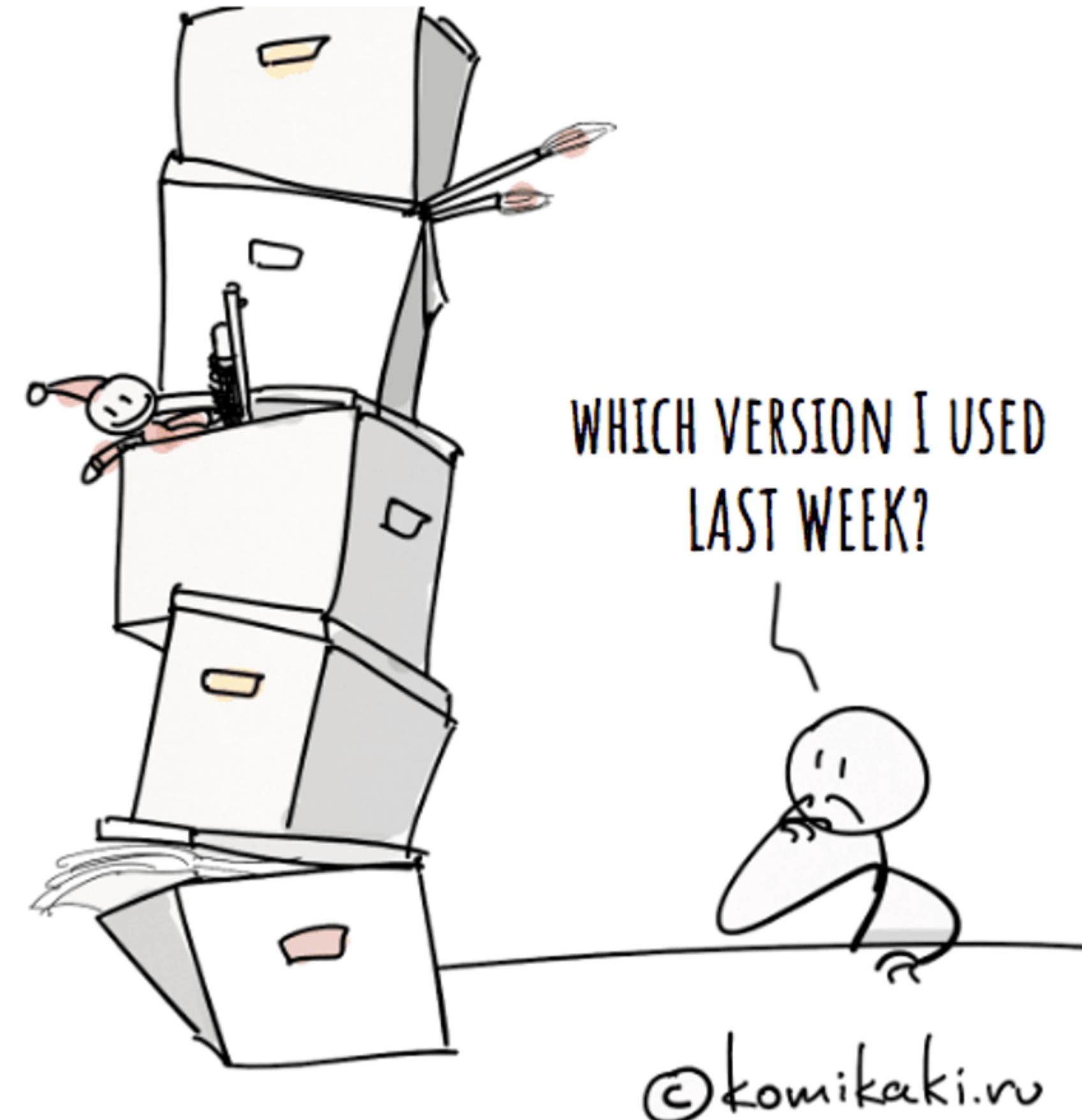






4 TB/day

# DATASETS MANAGEMENT

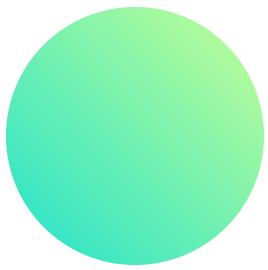


# Why Dataset management?

- > Moving Datasets around
  - > Datasets evolve, so versioning required
- »» EXPERIMENTS = CODE + DATA + OUTPUTS

Source code, Datasets

# HOW I DISCOVERED DVC





# DATA VERSION CONTROL(DVC)

---

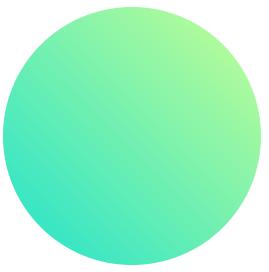


- › Experiment and Dataset tracking
- › Open-source(3500+ stars)
- › Built to adopt the best practises of ML
- › Works well with git
- › Language and framework agnostic

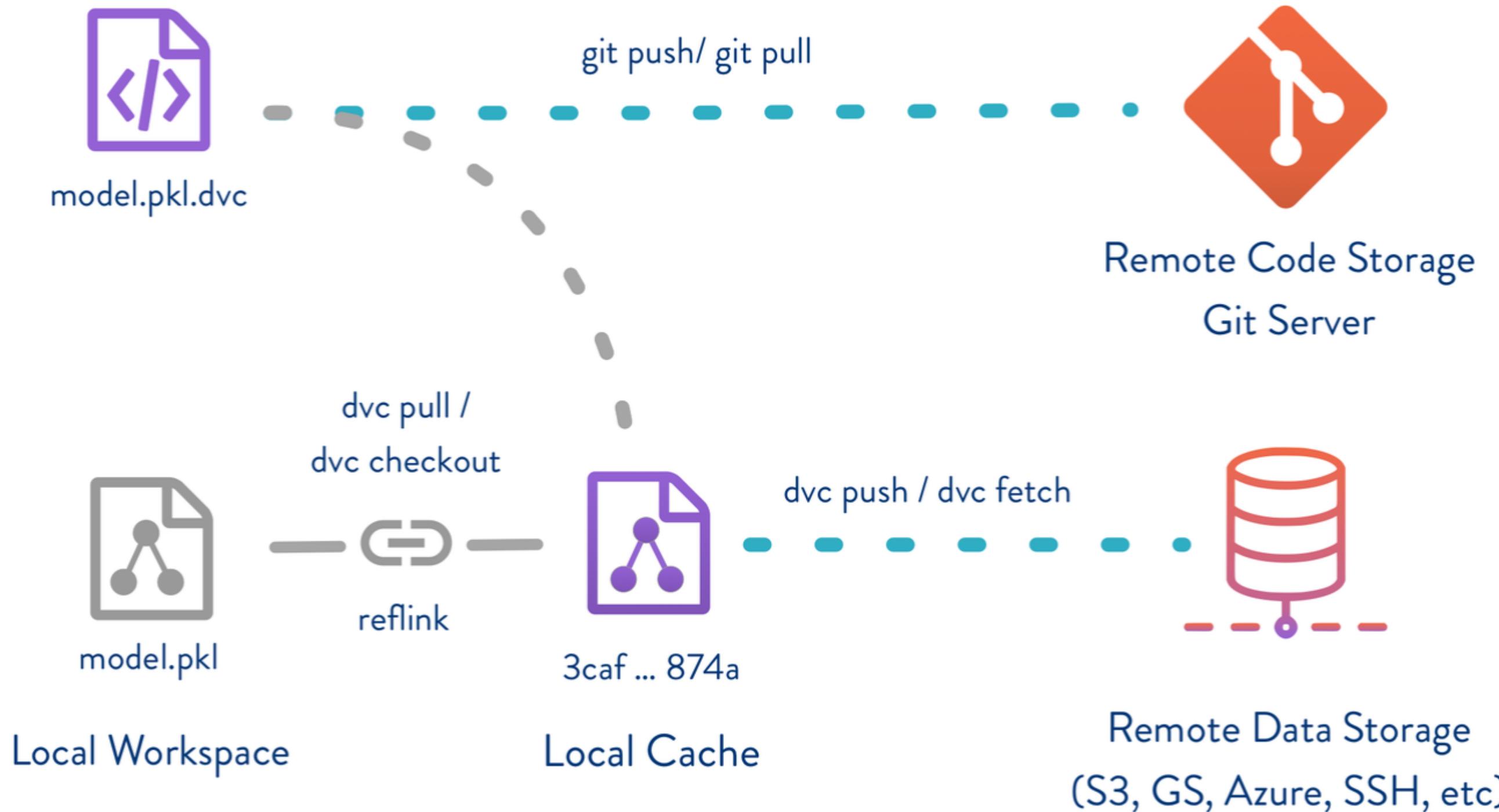
# VERSIONING CATS & DOGS



# DEMO TIME



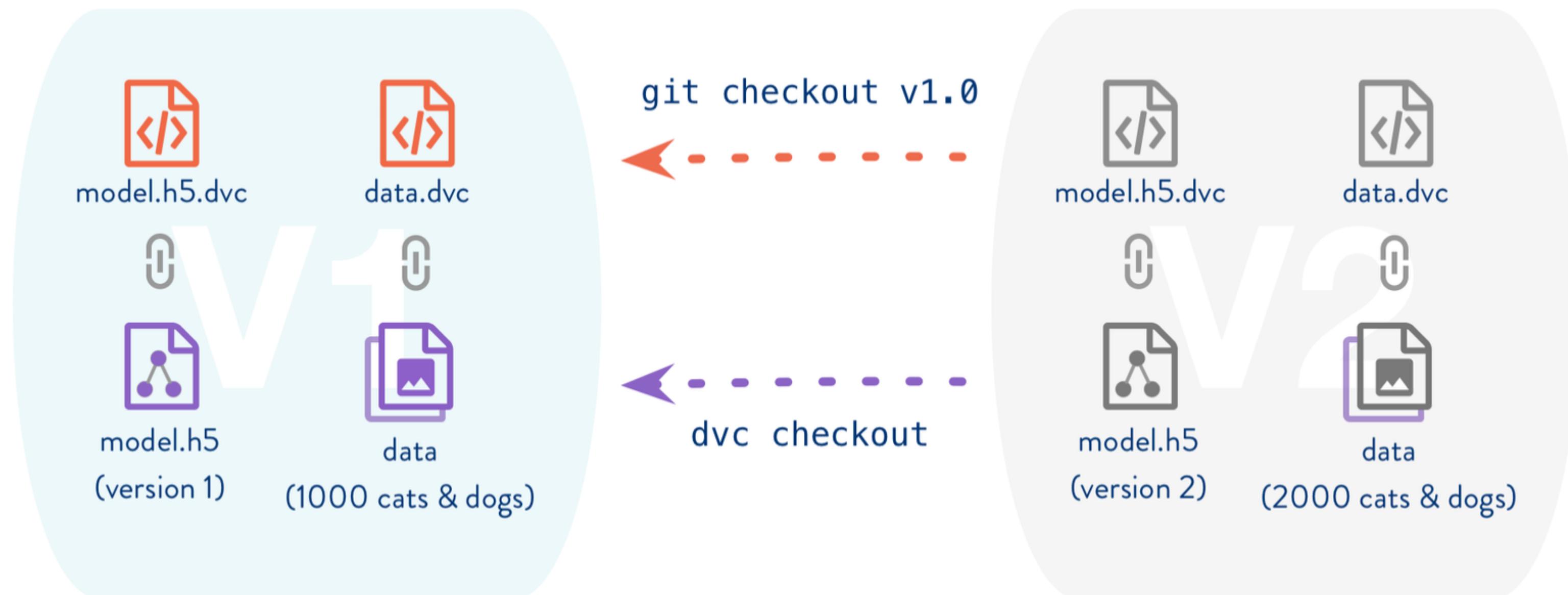
# DVC WORKFLOW



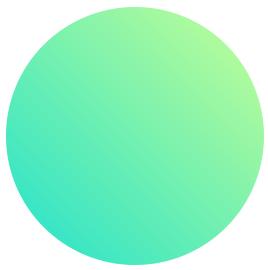
# Tracking data

- 1 Tracking 1000 cats and dogs**
- 2 Add 1000 more labelled images of cats & dogs**

# SWITCHING VERSIONS



# CONCLUSION



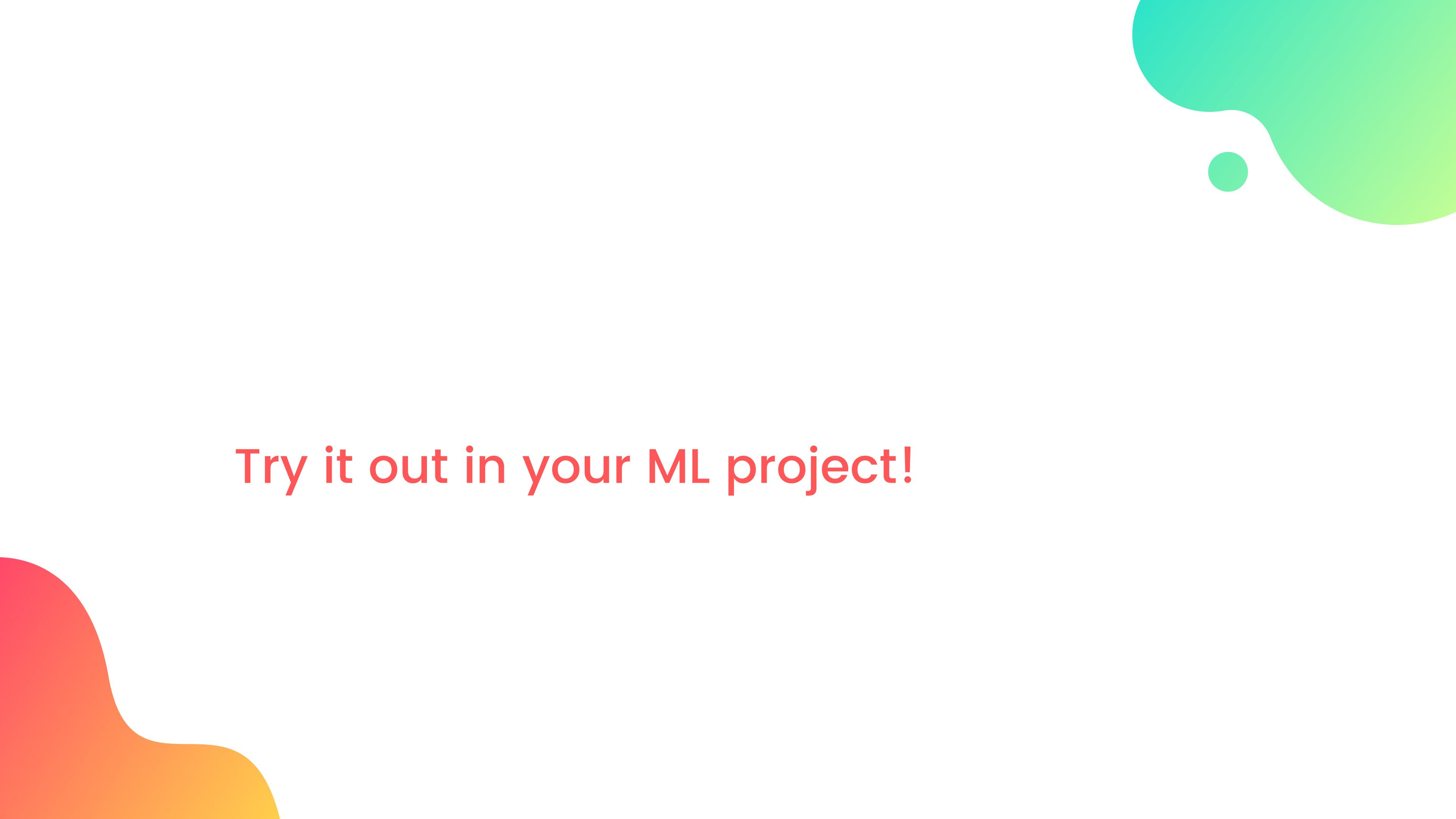


**"Data science as different from software  
as software was different from hardware."**

Nick Elprin,  
CEO, DominoLabs.

- Think about your processes(ML projects)

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



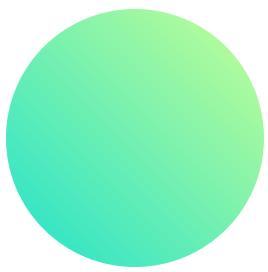
Try it out in your ML project!

# THANK YOU

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

Speaker Deck: [bit.ly/mlversion19](https://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