Pearson

Get your Machine Learning workflow under control with DVC

Image by: Tang Yau Hoong

Mikołaj Bogucki
Advanced Computing & Data Science Lab

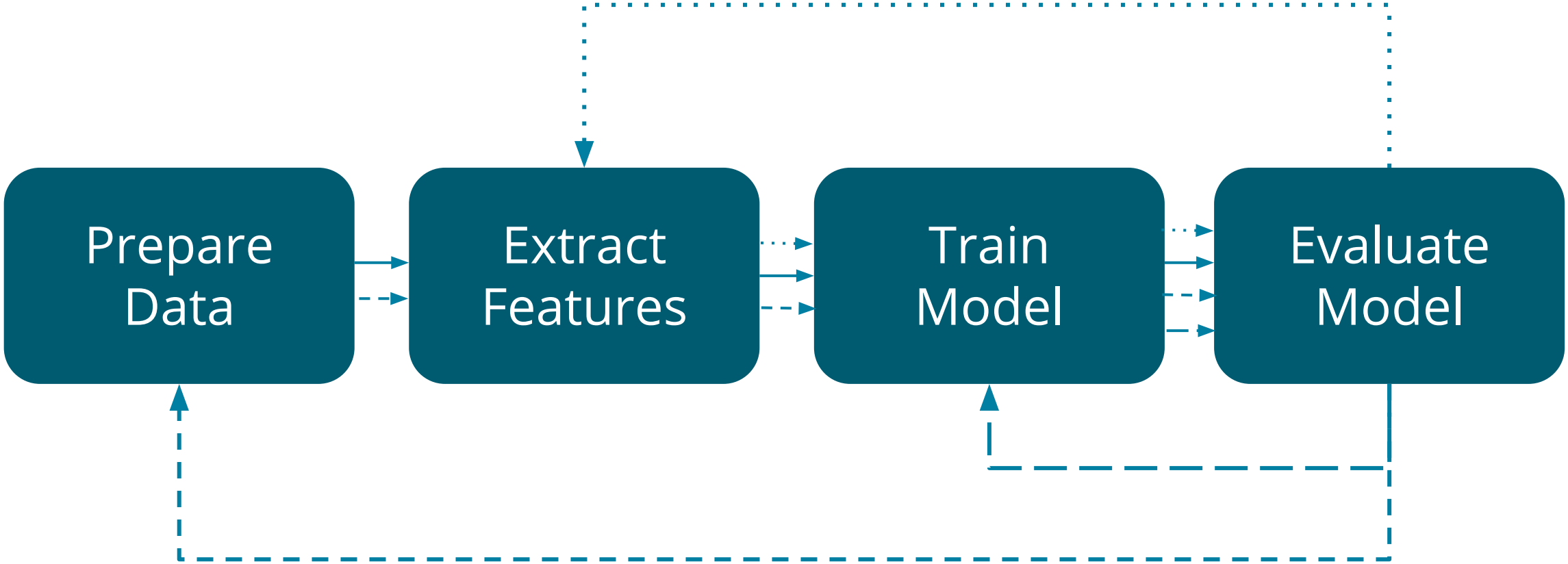
WhyR, 4 July 2018

Exploratory Data Science team

Pearson

Intro

Iterative Machine Learning Process



Source: https://dvc.org

Challenges

Challenges

- Discrepancy between versions of source code and versions of data files
- Effect of hyperparameter changes
- Recovering old ML experiments
- Reproducibility
- Sharing large data files with other colleagues

Our goal

Given relatively big data, we want to seamlessly experiment with new ML approaches.

DVC comes into play!

DVC

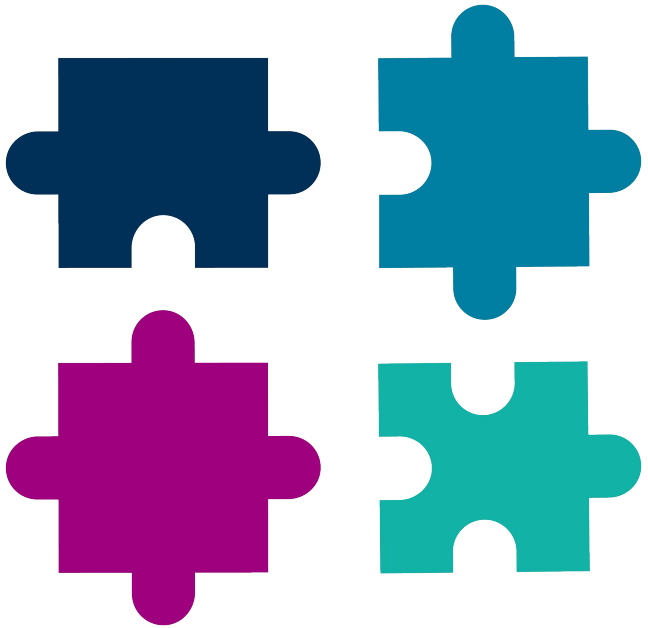
- DVC stands for **Data Version Control**.
- It's an **experiment management software** which works on top of Git repositories.
- DVC is a command-line tool.
- It supports all common OS: Mac OS, Linux and Windows.

DVC core features

- Building DAG pipelines which makes ML project reproducible
- Storing version of data files for each experiment
- Language agnostic
- Open-sourced
- Supporting cloud storage

DVC with Git

1. The master branch should store
   a stable version of our model.
2. Each experiment should be performed
   in a separate branch.
3. If experiment is successful, merge with master.



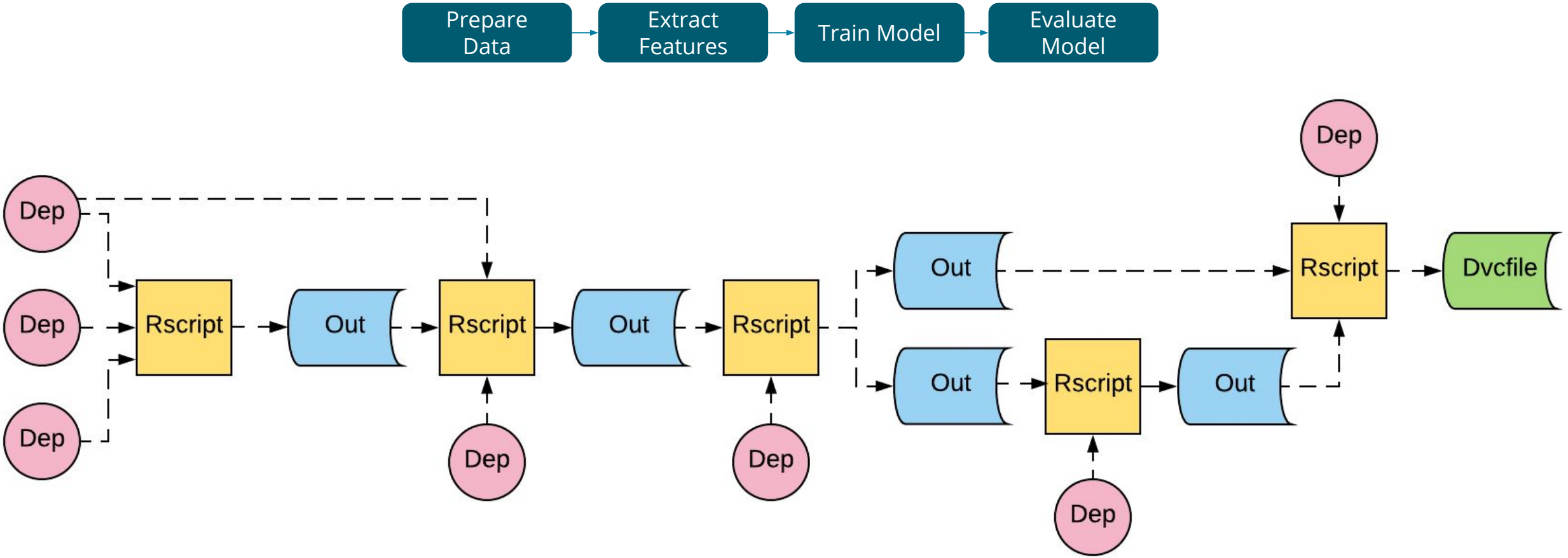
Walkthrough

How to start

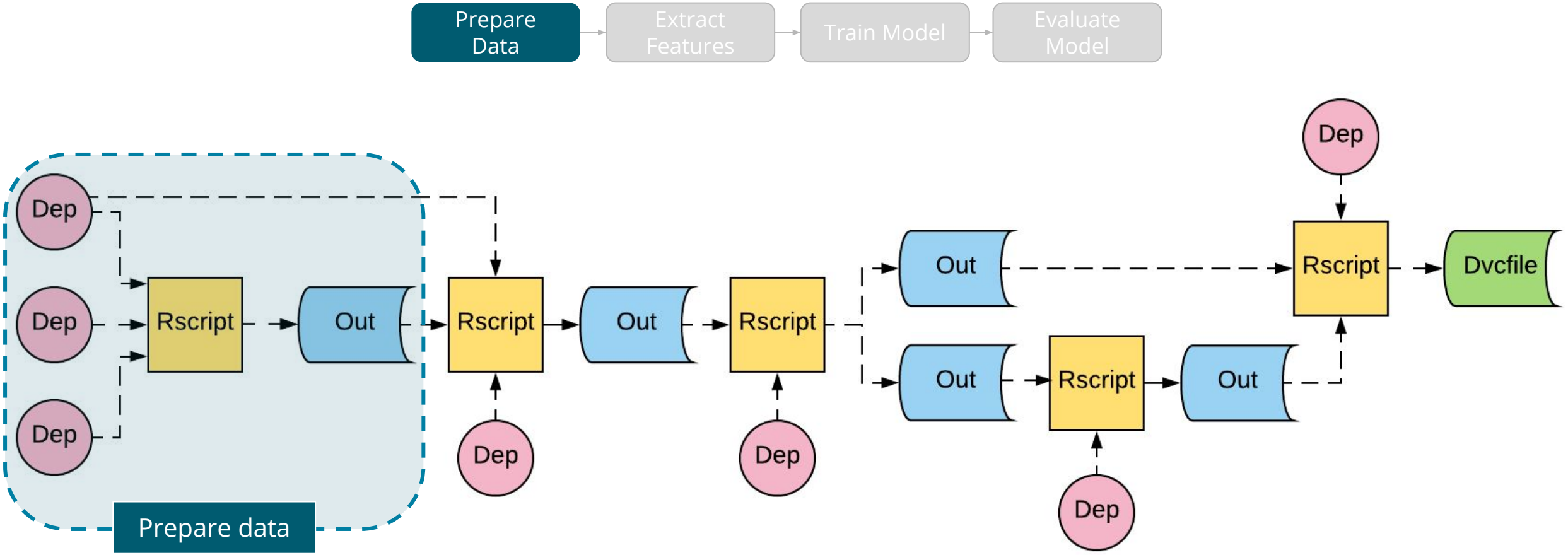
```
$ pip install dvc
$ git init
$ dvc init
$ ls .dvc
-> config
-> cache/
-> .gitignore
```

`cache/` directory will contain your data files. It is defined in `.gitignore`.

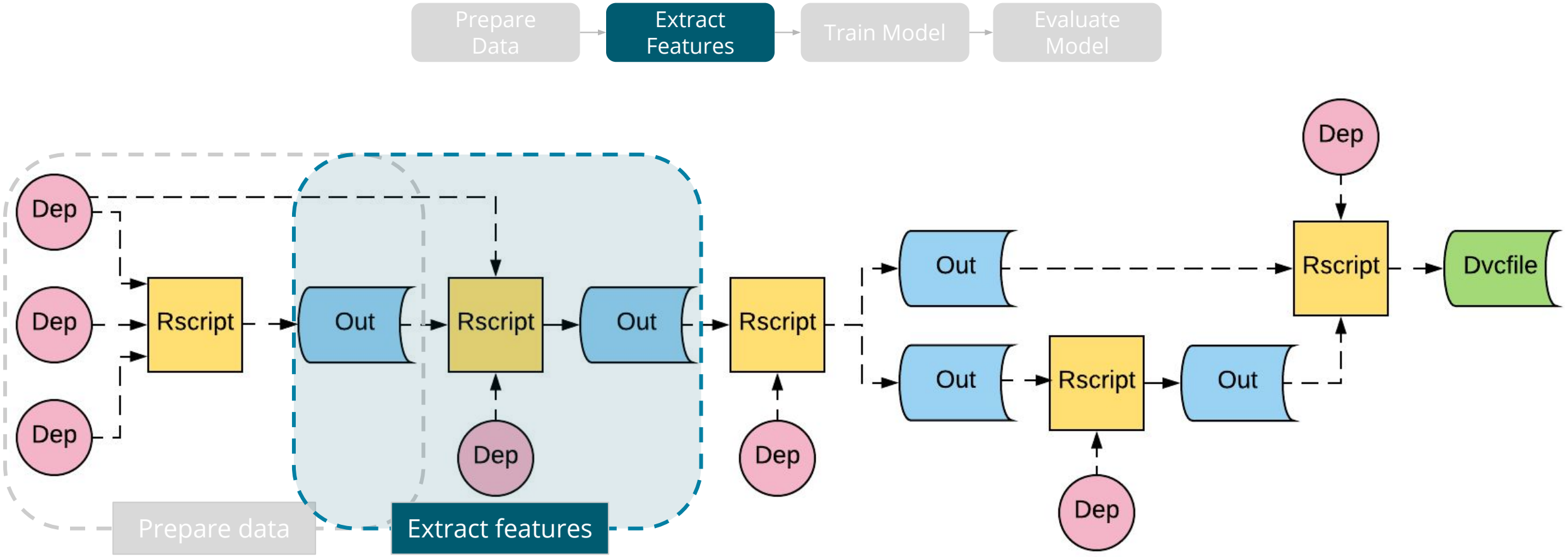
DVC pipeline



DVC pipeline



DVC pipeline



How to create a single task

`dvc run`

- Allows us to build lightweight DAG pipelines.
- It's important to define **dependencies** (-d), **outputs** (-o)

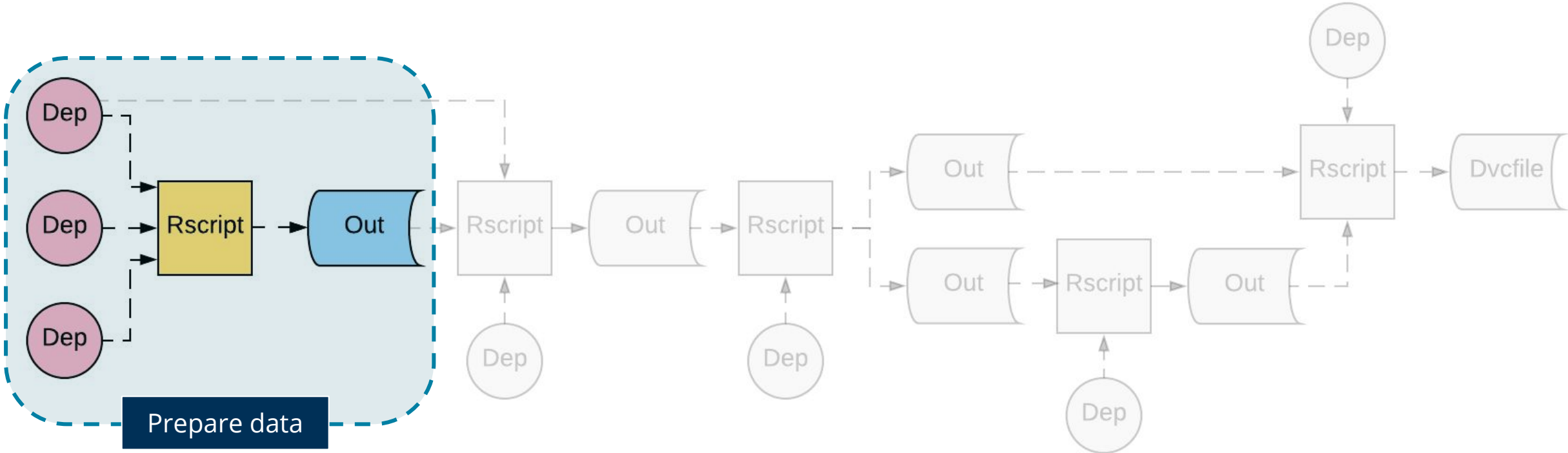
  and a **script** to execute.

```
$ dvc run -d data/titanic.csv -d code/conf.R -d code/data_preparation.R -o
data/clean_titanic.csv Rscript code/data_preparation.R
```

- Once we you run it, it creates `.dvc meta data` file which corresponds with `.dvc/cache` directory.

dvc run

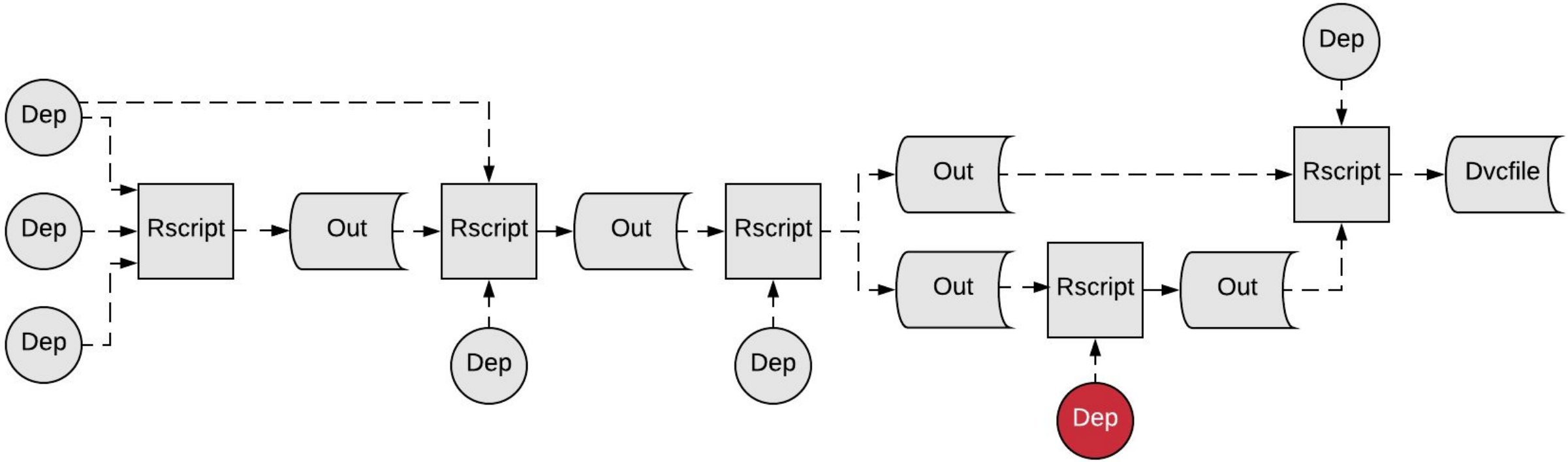
```
$ dvc run -d data/titanic.csv -d code/conf.R -d code/data_preparation.R
-o data/clean_titanic.csv Rscript code/data_preparation.R
```



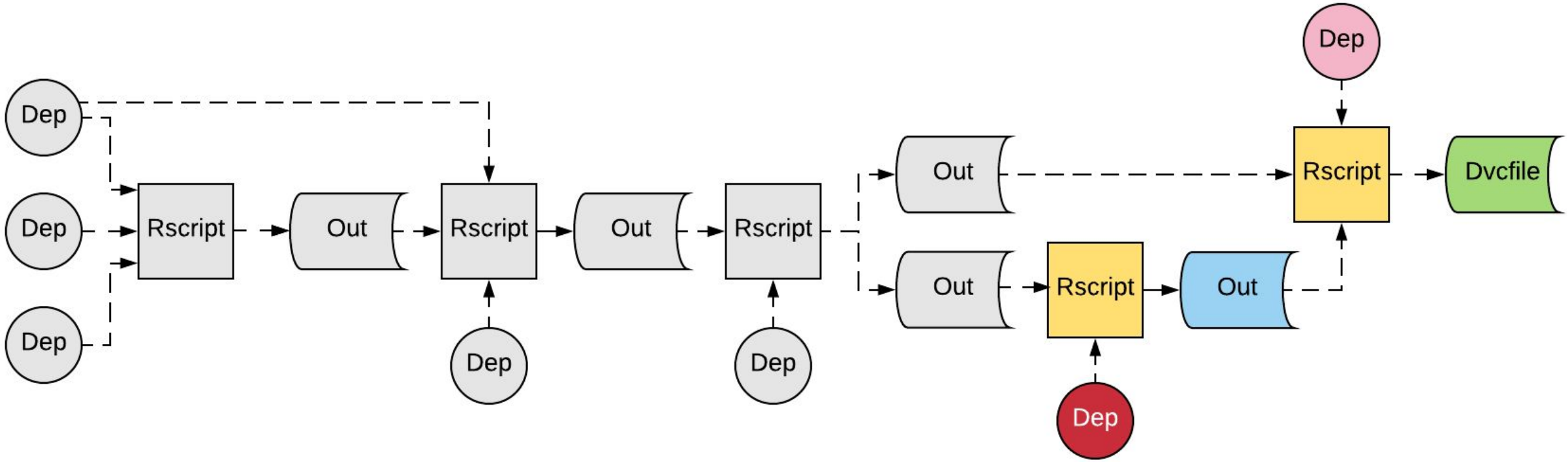
`dvc repro`

- A command which allows to reproduce our DVC pipeline.
- It automatically **detects** only these tasks which **have to be reproduced**.

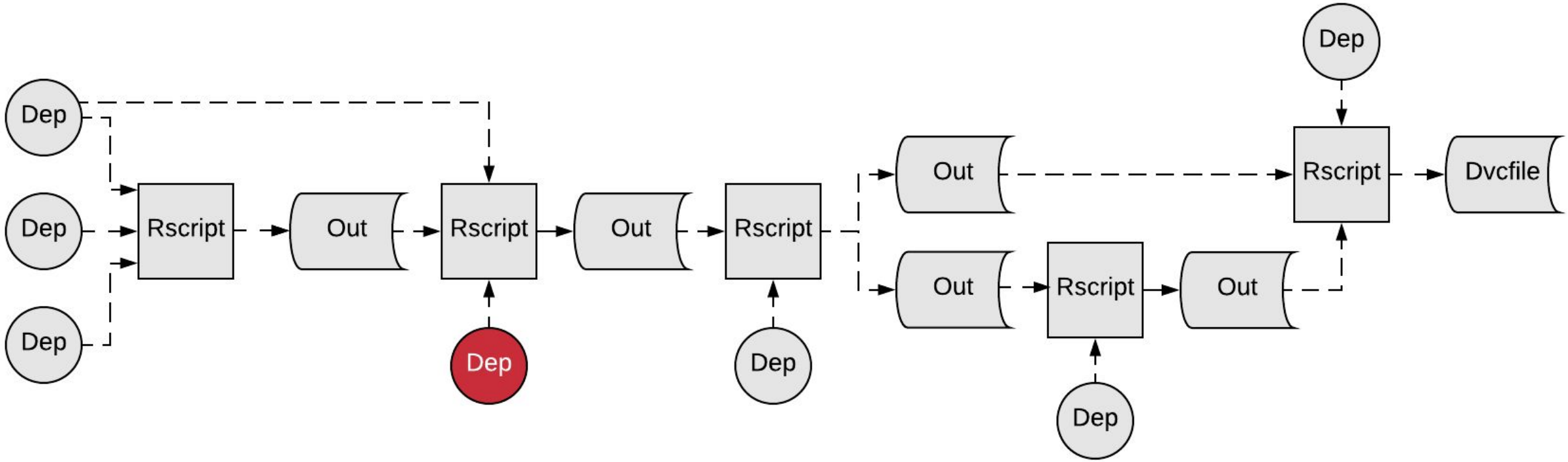
dvc repro


Dep
Dep
Dep
Rscript
Out
Rscript
Dep
Out
Rscript
Dep
Out
Out
Rscript
Dep
Rscript
Out
Dvcfile

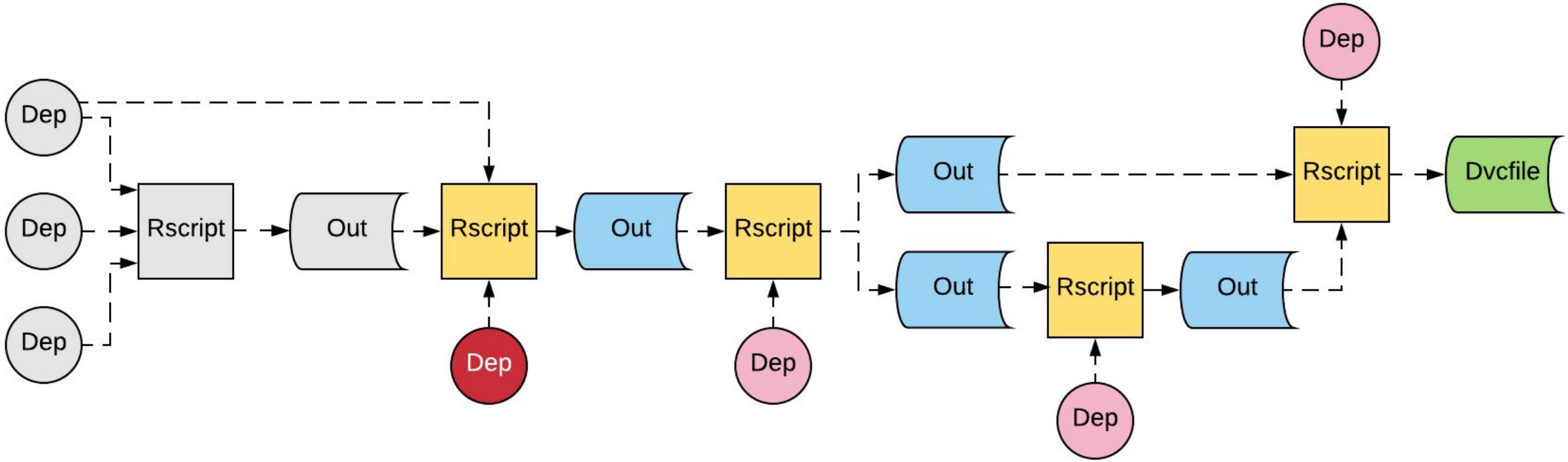
dvc repro


Dep
Dep
Dep
Rscript
Out
Rscript
Dep
Out
Rscript
Dep
Out
Out
Rscript
Dep
Out
Rscript
Dvcfile

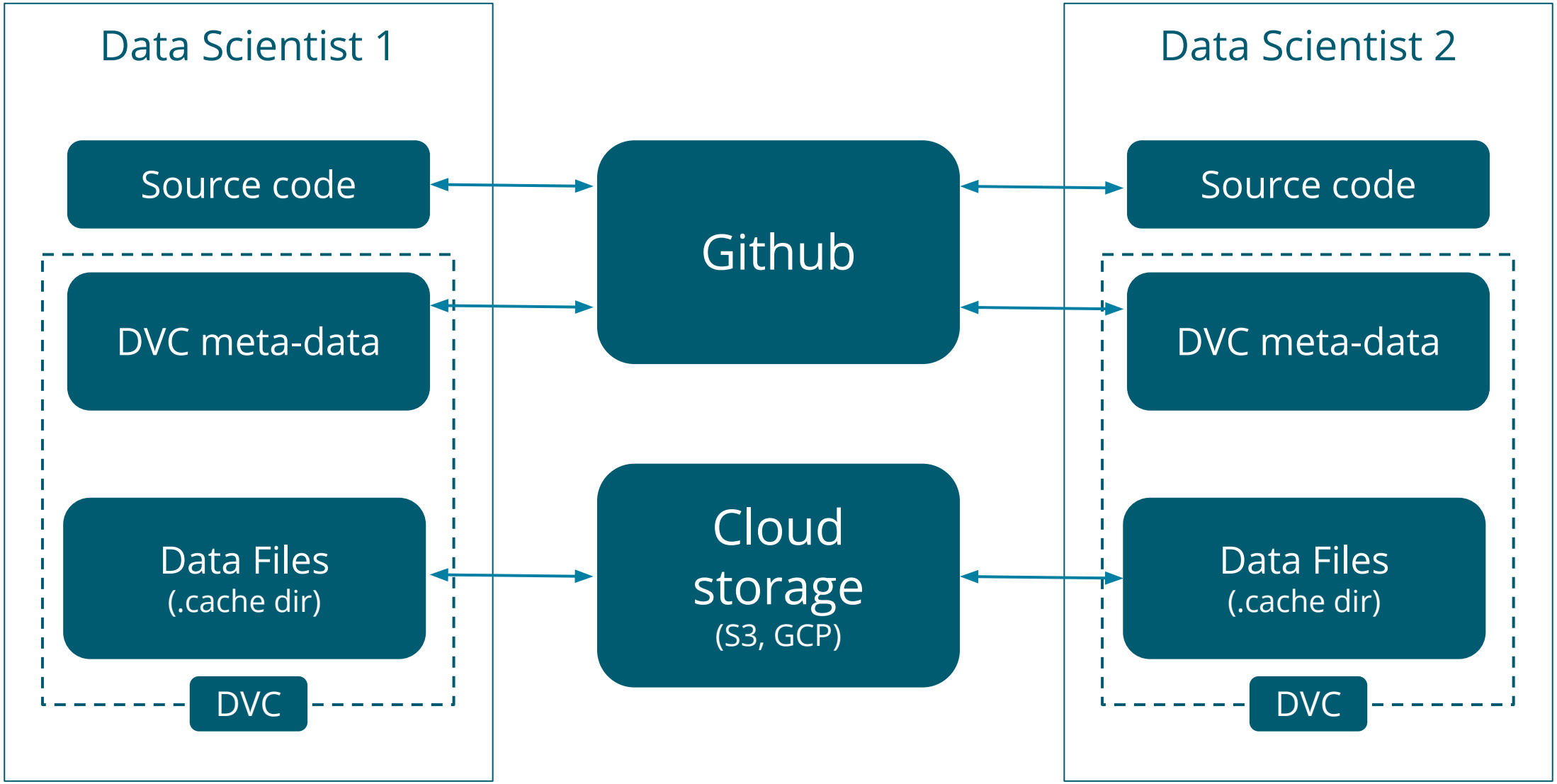
dvc repro


Dep
Dep
Dep
Rscript
Out
Rscript
Dep
Out
Rscript
Dep
Out
Out
Rscript
Dep
Rscript
Out
Dvcfile

dvc repro


Dep
Dep
Dep
Rscript
Out
Rscript
Dep
Out
Rscript
Dep
Out
Out
Rscript
Dep
Out
Rscript
Out
Dep
Rscript
Dvcfile

Team workflow



Source: https://dvc.org

Caveats

Keep in mind

- Rapid development of DVC
- DVC developers are very responsive - just create a github issue!

Useful links

- DVC homepage dvc.org
- Our team blog: ioki.pl/blog

ALWAYS LEARNING