

# Course Introduction

Lesson 1

---

DVC tools for Data Scientists &  
Analysts

**2021**



## Lesson Outline

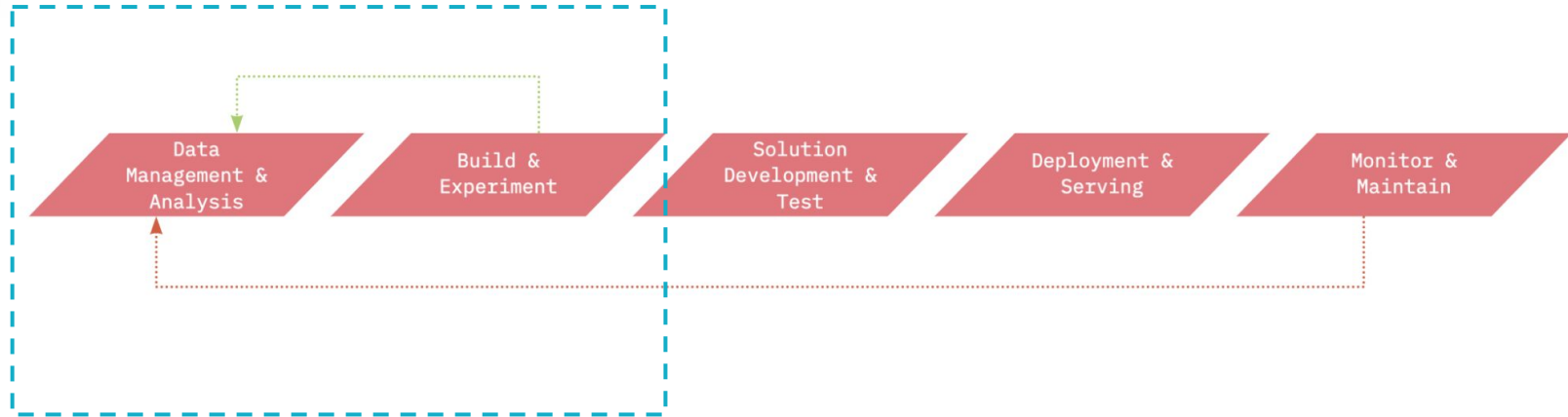
- ◇ Motivation
- ◇ What is DVC?
- ◇ What is DVC Studio?
- ◇ Course objectives
- ◇ Course structure



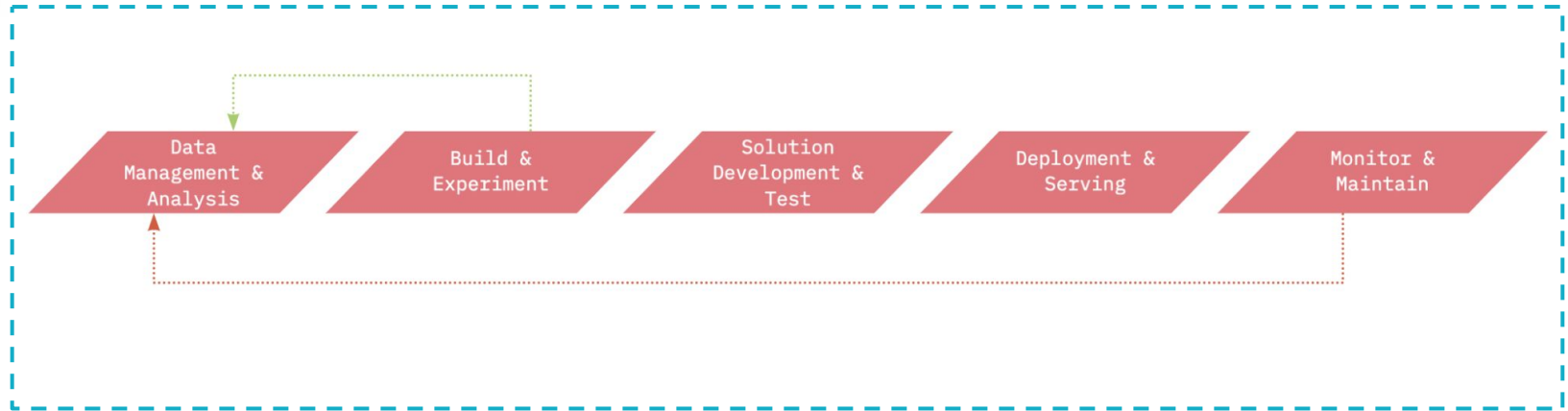


# Motivation

# Machine Learning Workflow



# Machine Learning Workflow



# Common DS/ML Issues



- ◇ Difficult sharing & collaborating
- ◇ Inefficiency & work duplication
- ◇ Slow updates
- ◇ Pipelines not reliable or reproducible
- ◇ Data quality issues
- ◇ Model metrics tracking

# Good practices for ML projects



## 1. Project structure & dev environment

- ◇ Organize a project repository
- ◇ Environment dependencies control

## 2. Coding (software development)

- ◇ Follow style-guides
- ◇ Code version control (Git)

## 3. Documentation & task tracking

- ◇ Document your code, experiments, and findings
- ◇ Task tracking

## 4. ML pipelines development & experiments

- ◇ Automated pipelines
- ◇ Control run params
- ◇ Models and artifacts version control
- ◇ Experiment results tracking
- ◇ Reproducible experiments



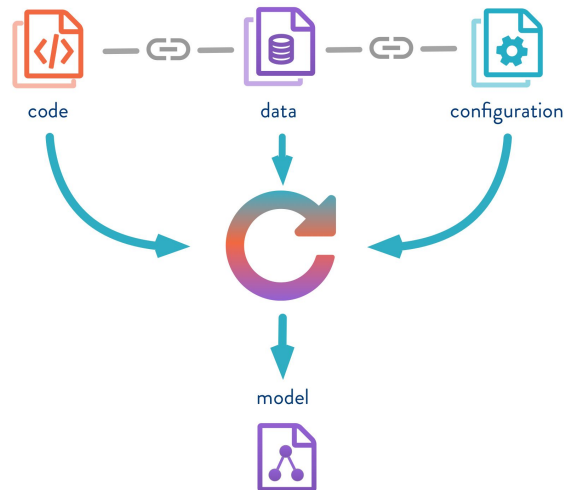
# What is DVC?



# What is DVC?



- ◆ Platform to manage machine learning experiments and pipelines
- ◆ Tool for data and model versioning
- ◆ Data access, sharing and collaboration tool
- ◆ Link between your code and data





**DVC Team**

**Welcome video**



# What is DVC Studio?

# Studio: UI for ML experiments and metrics tracking



Views > example-get-started Demo Private

Search Filters Columns Show plots Compare Selected only Trends Delta mode

Commit	Created	Message	avg_prec	roc_auc	data
codespaces		inherited from master			
codespac...	Sep 14, 2021	add Dockerfile to install DVC	0.00000	0.00000	+0 B
try-large-dataset		inherited from master			
try-larg...	Jun 01, 2021	Try 100K dataset (4x data)	+0.06632	+0.00613	+114.2 MB
master					
BASELINE HEAD, ma...	May 29, 2021	Run experiments tuning ra...	0.60405	0.96080	37.9 MB
10-bigra...	May 28, 2021	Evaluate bigrams model	-0.05146	-0.04544	+0 B
9-bigram...	May 27, 2021	Reproduce model using bl...	-0.08357	-0.05760	+0 B
8-evalua...	May 25, 2021	Create evaluation stage	-0.08357	-0.05760	+0 B
7-ml-pip...	May 24, 2021	Create ML pipeline stages	-	-	+0 B

### Changes

Open diff on GitHub

try-large-dataset x

10-bigrams-experiment, bigrams-experiment x

Show diff for all data points (including hidden)

### Metrics

Name	try-large-d...	10-bigrams...
scores.json:avg_prec	0.67038	0.55259
scores.json:roc_auc	0.96693	0.91536

### Parameters

Name	try-large-d...	10-bigrams...
...eaturize.max_features	3000	1500
...s.yaml:train.min_split	64	2
params.yaml:train_n_est	100	50

List of experiments

Track changes



# DVC Studio Team

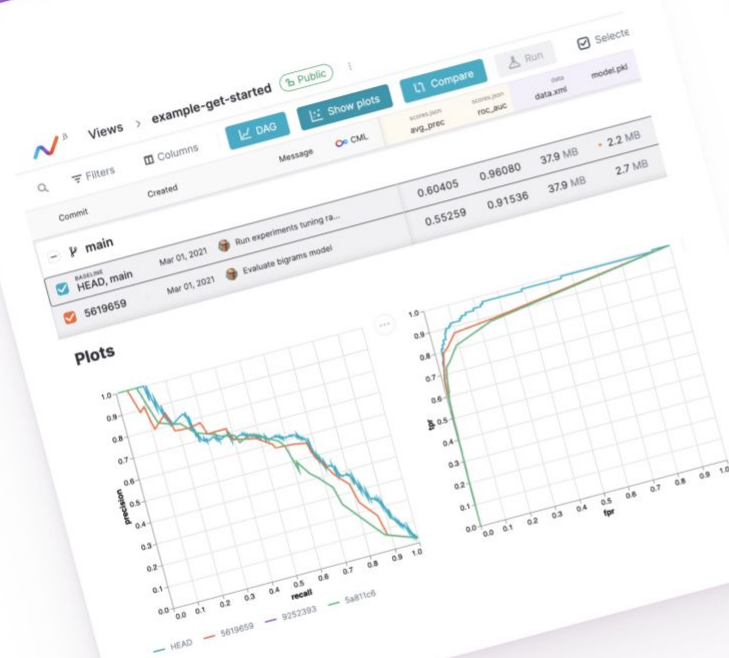
Welcome video



# Course objectives

# Course objectives

1. Improve ML experimenting & development processes
2. Bring good engineering practices into ML
3. Improve team collaboration
4. Learn & integrate tools for ML projects

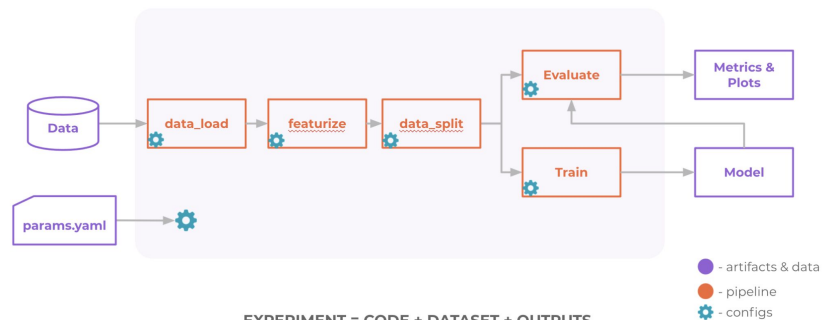


# What will you learn?

How to...



1. Build automated pipelines and reproducible experiments



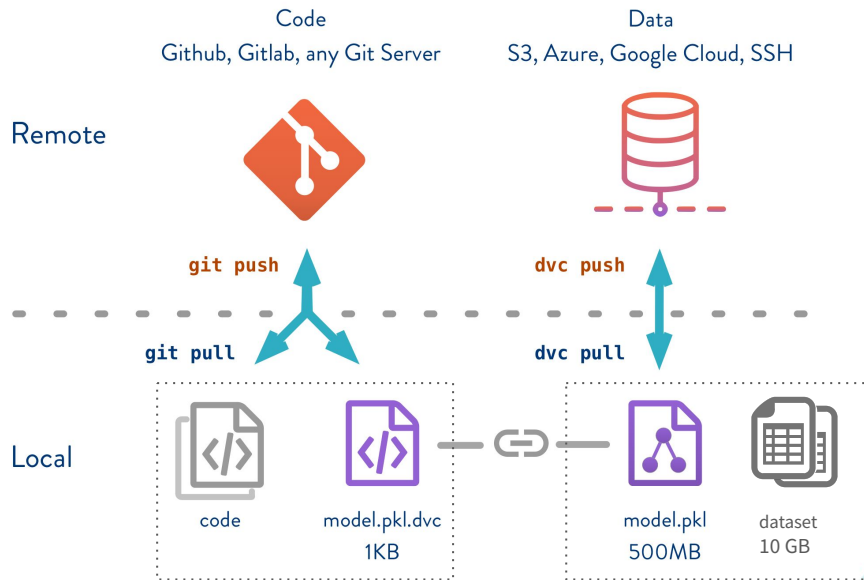


# What will you learn?

How to...



1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning

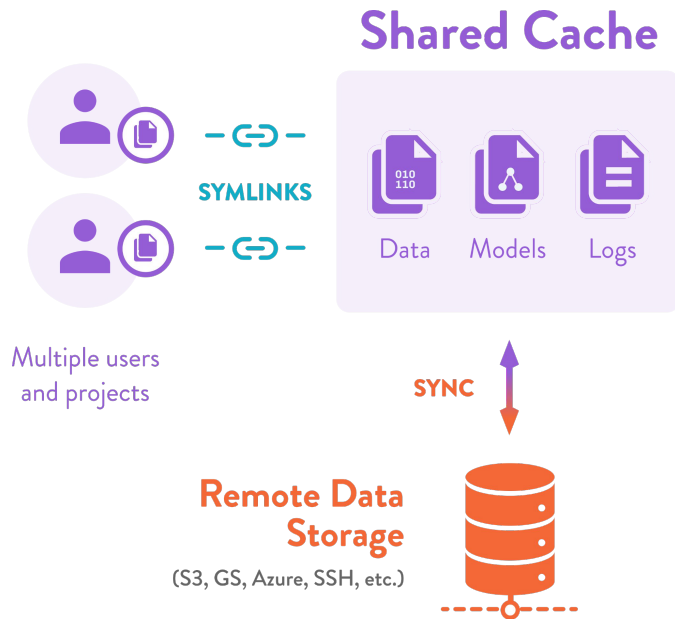


# What will you learn?

How to...



1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning
3. Organize your project code and team collaboration

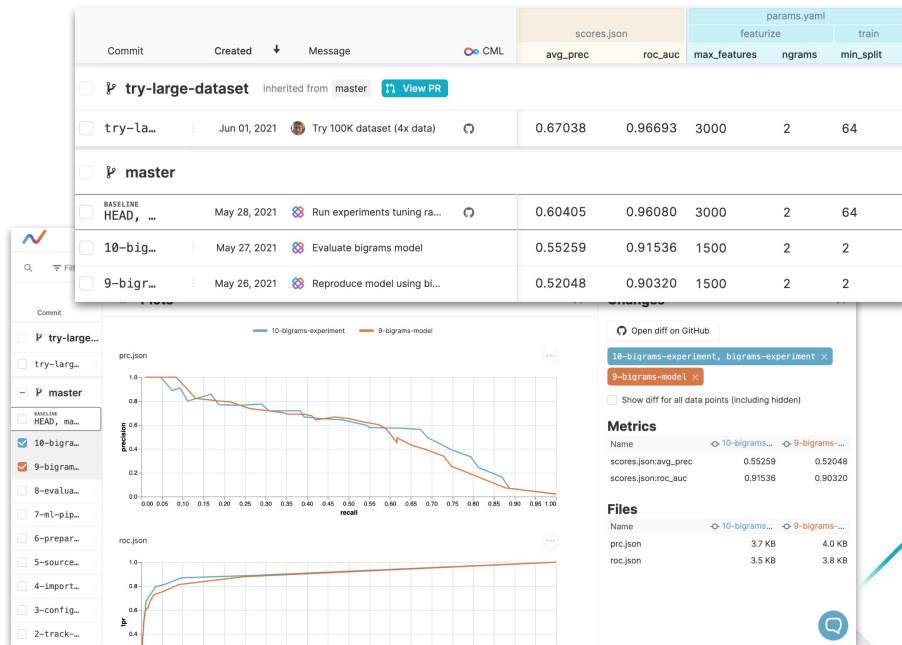


# What will you learn?

How to...



1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning
3. Organize your project code and team collaboration
4. Visualize metrics & collaborate on ML experiments

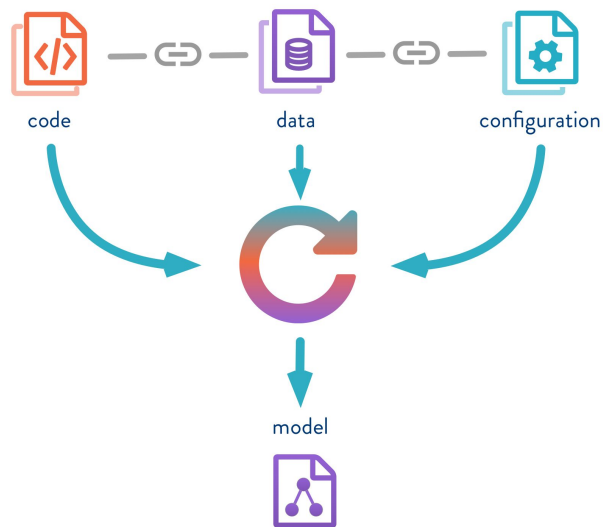


# What will you learn?

How to...



1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning
3. Organize your project code and team collaboration
4. Visualize metrics & collaborate on ML experiments
5. Integrate DVC and DVC Studio into your own project





# Course structure



# Course lessons

**Lesson 1.** Course Introduction

**Lesson 2.** Practices and Tools for Efficient Collaboration in ML Projects

**Lesson 3.** Pipeline Automation and Configuration Management

**Lesson 4.** Versioning Data and Models

**Lesson 5.** Visualizing Metrics & Comparing Experiments with DVC and Studio

**Lesson 6.** Experiment Management and Collaboration

**Lesson 7.** Tools for Deep Learning

**Lesson 8.** Review of Advanced Topics and Use Cases

# Course content and tools



## Format

- ◇ Video lectures with slides
- ◇ Code examples and demos
- ◇ Discussions in Discord

## Tools

- ◇ Jupyter Notebooks
- ◇ Python
- ◇ Git
- ◇ DVC
- ◇ DVC Studio

# Important Prerequisites



## Skills

- ◇ Basic knowledge of Python
- ◇ Basic CLI
- ◇ Basic Git

## System

- ◇ Software: Python, Git, Docker, DVC
- ◇ ~ 1 GB disk space
- ◇ min 4 GB RAM is recommended



# Checklist before take-off



1. Python installed
2. Python packages: pip, virtualenv
3. Git installed
4. Registered at the class Discord channel
5. **Say Hello** to the class and share your expectations of this course





# Demo

**Where to find more material,  
useful links, and Discord  
channel**

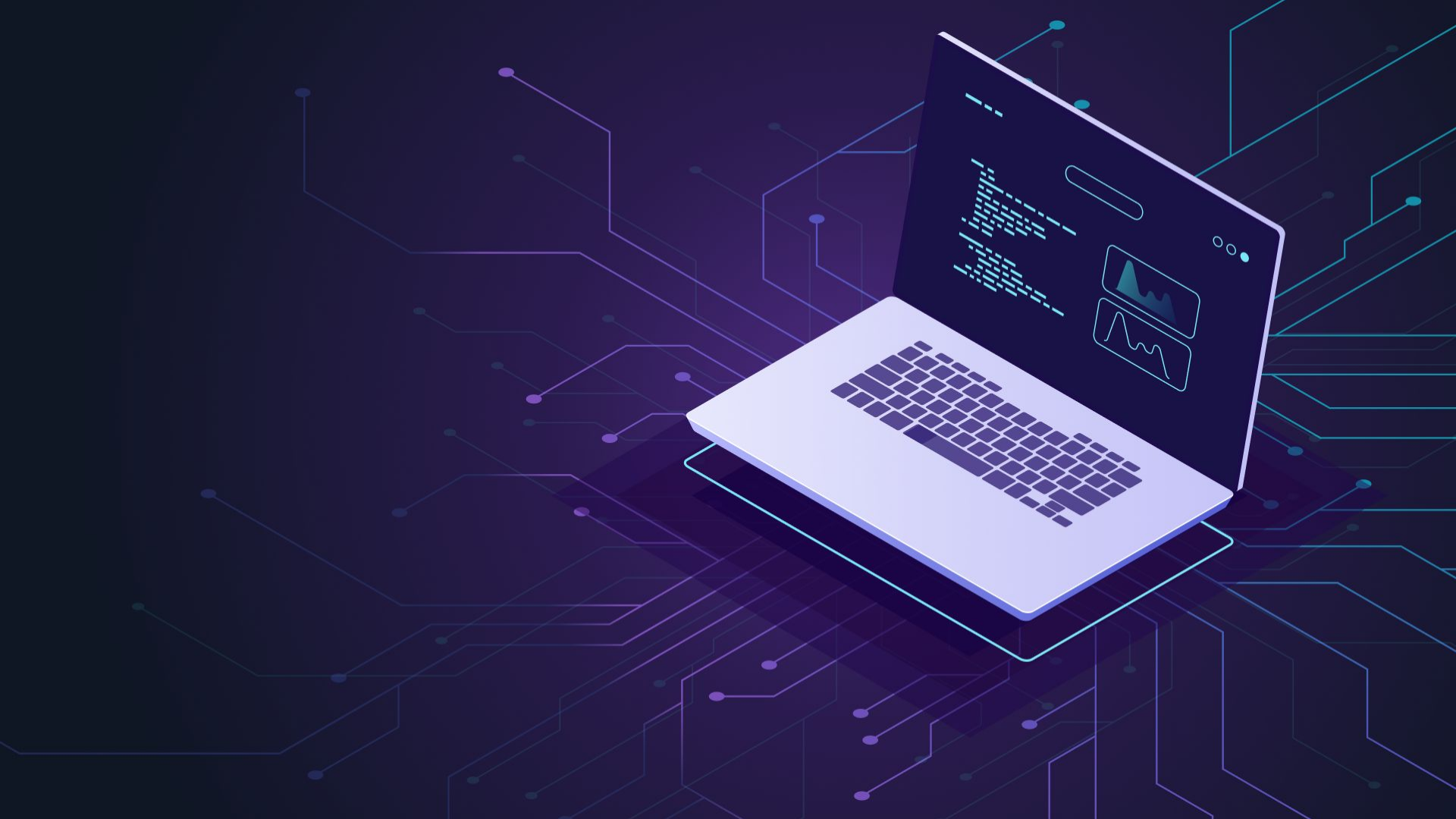


**What have we learned?**

# What have we learned?



1. Course objectives and structure
2. What is DVC
3. What is DVC Studio





# Links



Data Science blueprint

<https://data-science-blueprint.readthedocs.io/en/latest/presentation/schema.html>

