iterative

# Course Introduction

Lesson 1

DVC tools for Data Scientists & Analysts

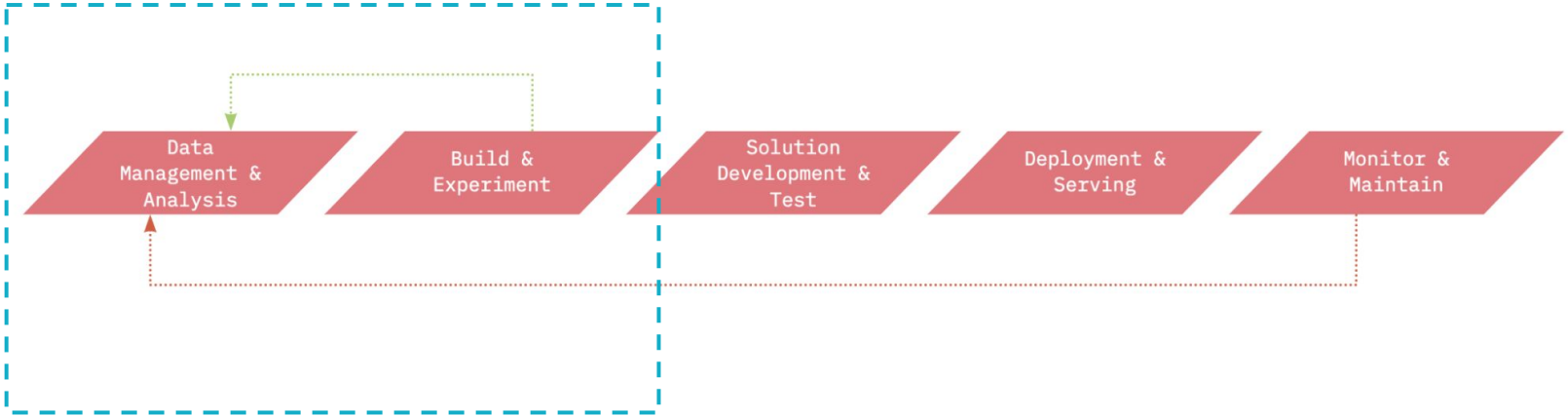
2021

# Lesson Outline

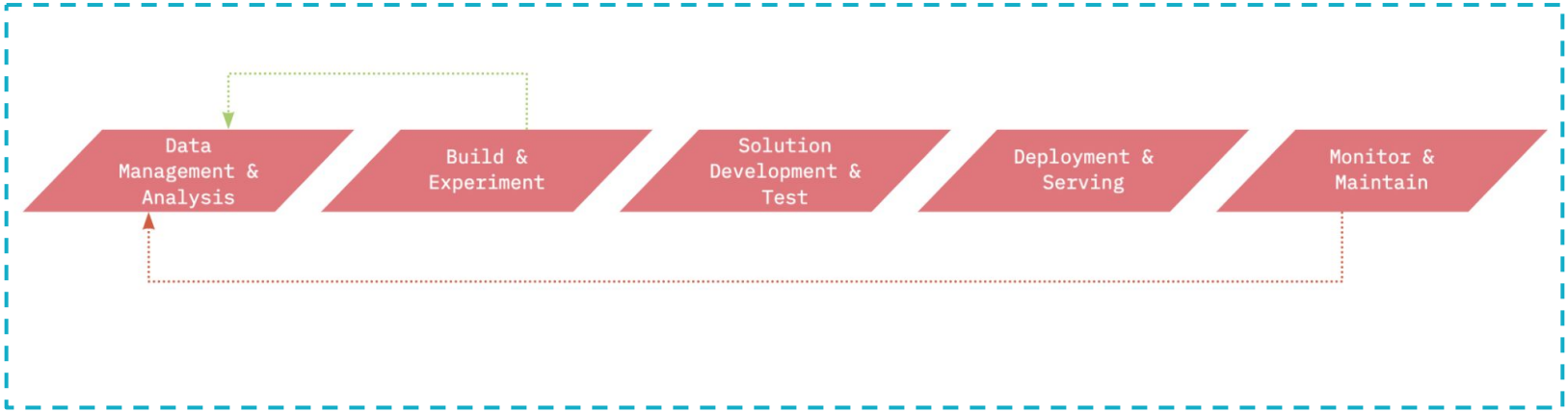
- Motivation
- What is DVC?
- What is DVC Studio?
- Course objectives
- Course structure

# 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

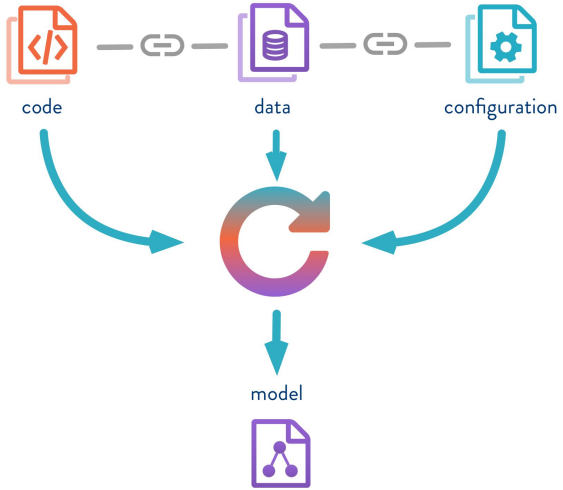
# Good practices for ML projects

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

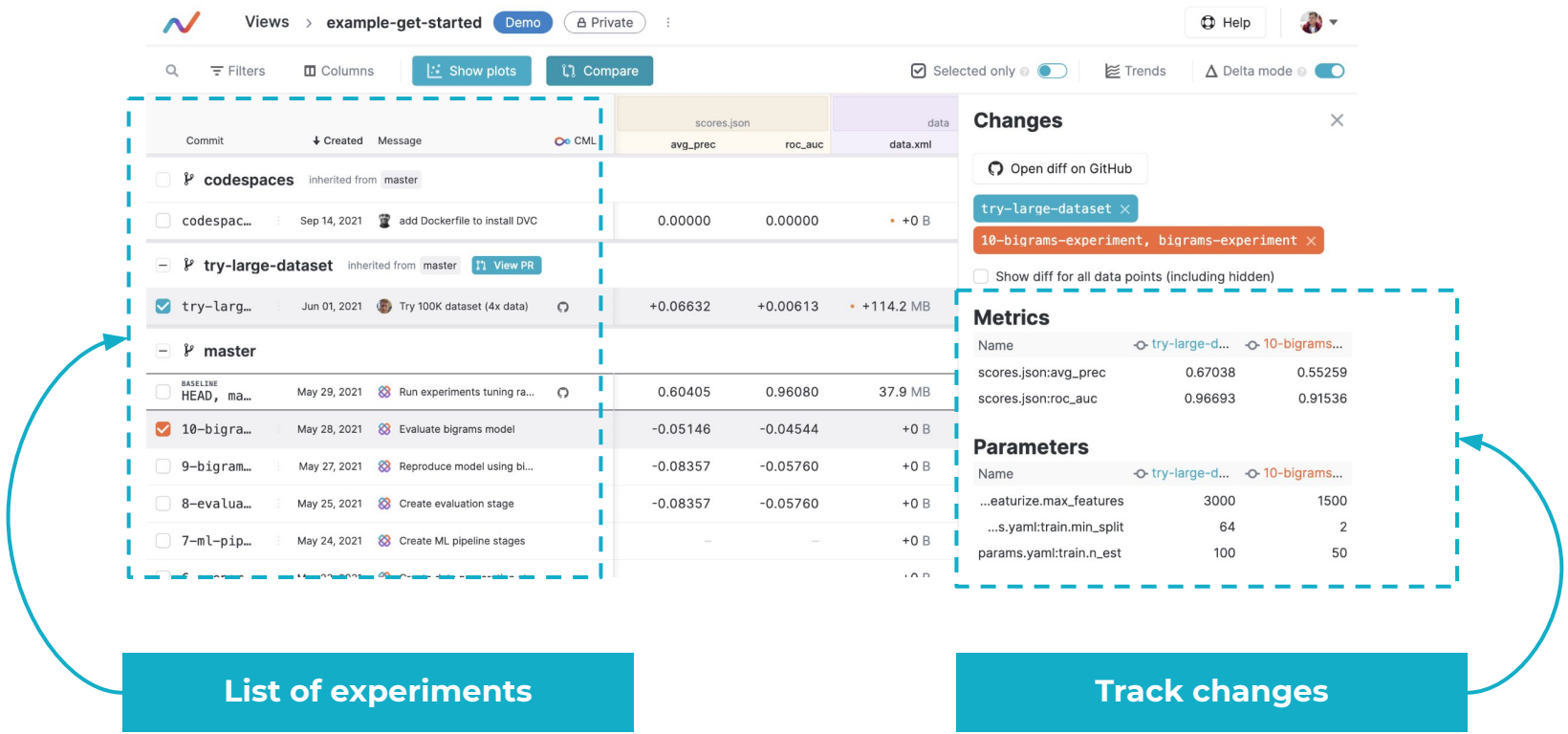


# DVC Team

## Welcome video

# What is DVC Studio?

# **Studio:** UI for ML experiments and metrics tracking




**List of experiments**

**Track changes**

# DVC Studio Team

## Welcome video

# Course objectives

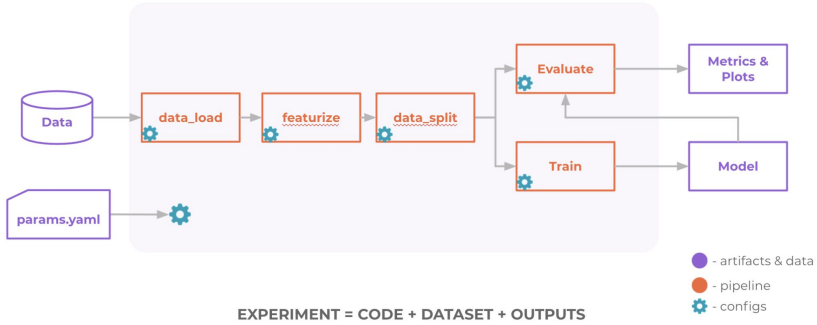
# Course objectives

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

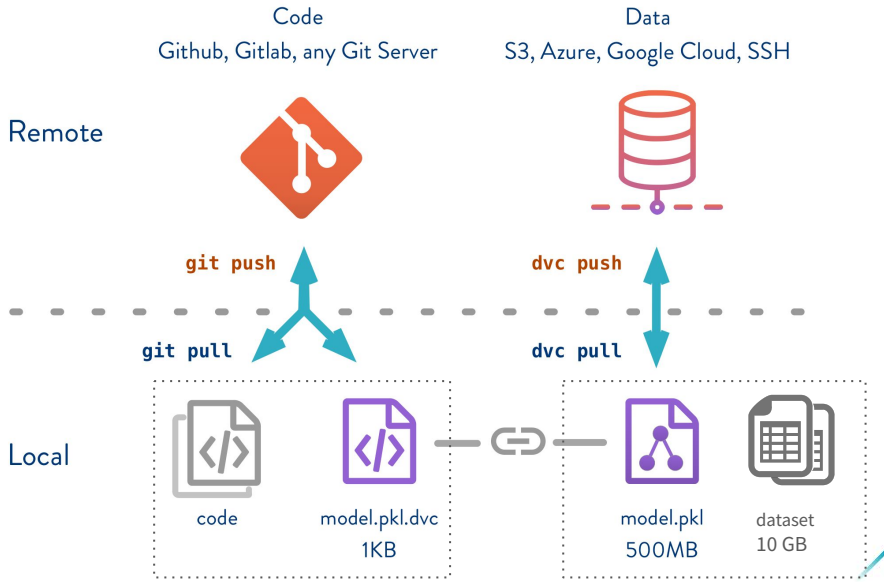



EXPERIMENT = CODE + DATASET + OUTPUTS

# What will you learn?

How to...

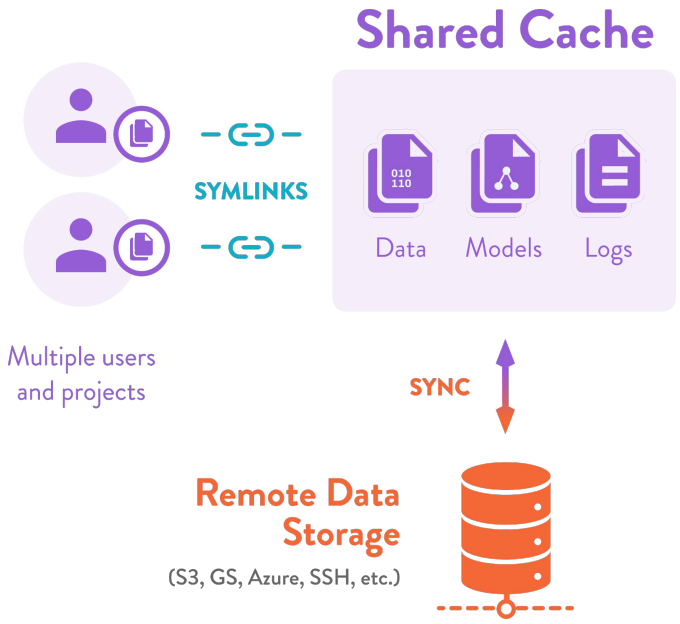
1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning



# What will you learn?

How to...

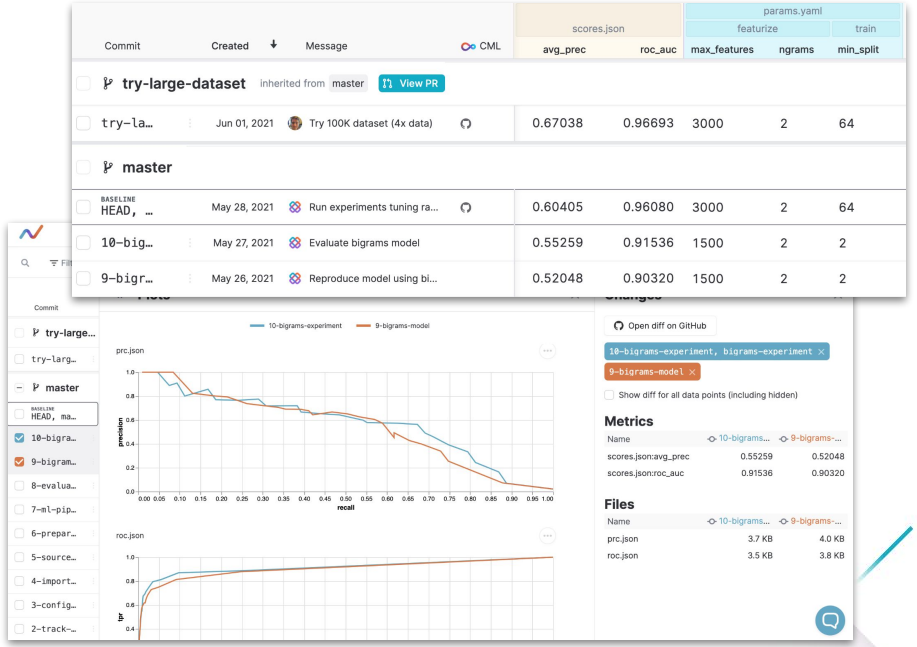
1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning
3. Organize your project code and team collaboration



# What will you learn?

How to...

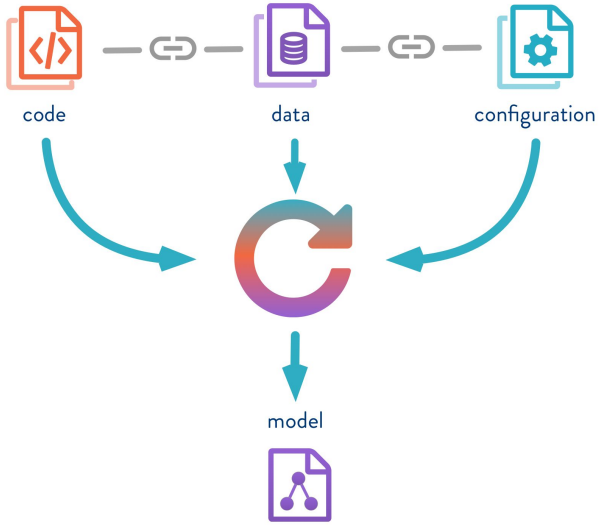
1. Build automated pipelines and reproducible experiments
2. Manage data and model versioning
3. Organize your project code and team collaboration
4. Visualize metrics & collaborate on ML experiments



# 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

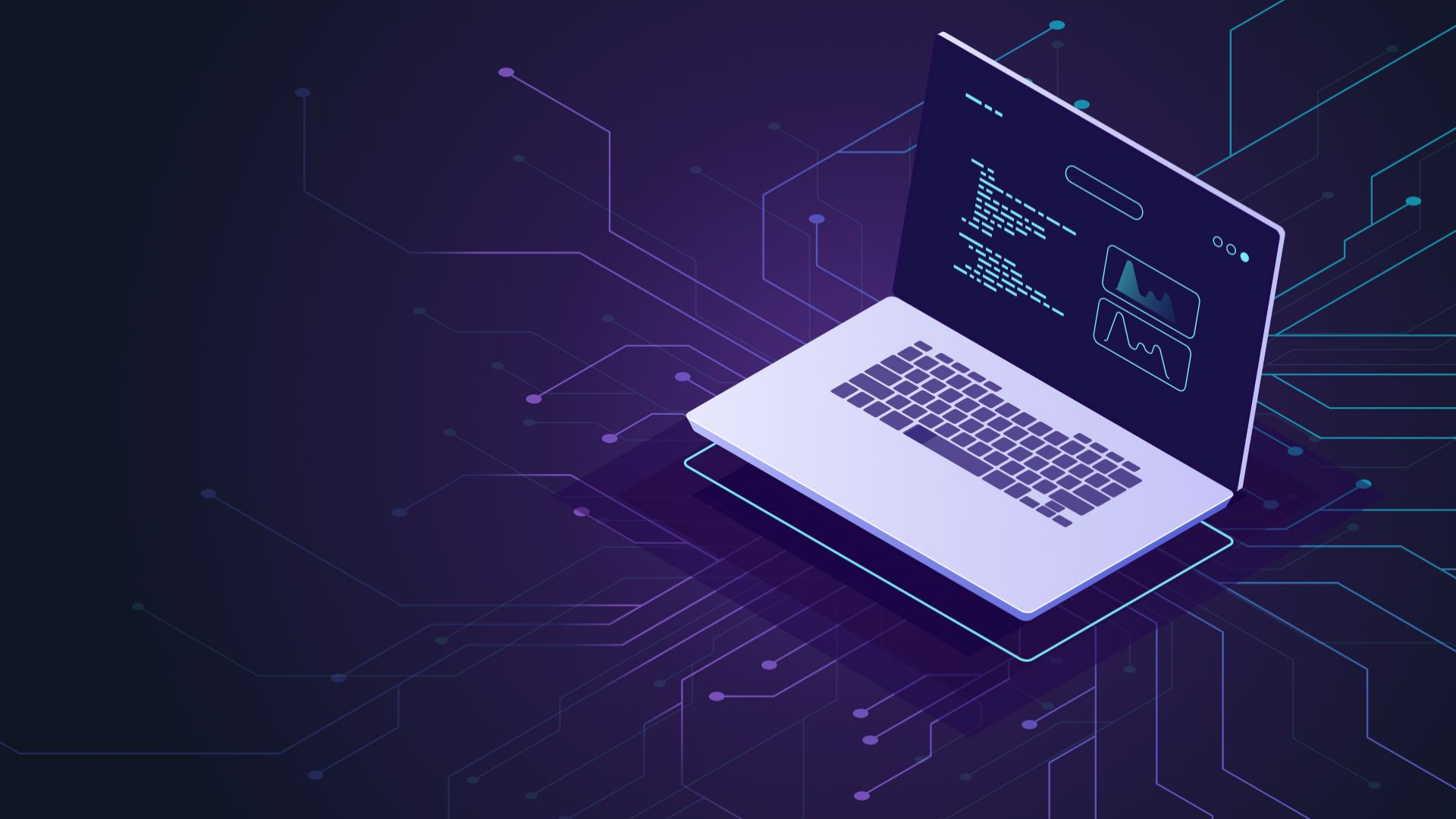
# Demo

## 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