



Data Science & Machine Learning

# **1 - REPLICABILIDADE**

**Consultor:** Ricardo Manhães Savii



# **Ricardo Manhães Savii**

Machine learning  
Engineer  
@Thoughtworks

Sistemas Inteligentes desde  
2018.

---

Projetos de Processamento de  
Linguagem Natural,  
Classificação de Imagens e  
Aplicações em E-commerce.

---

MLOps (suporte para a criação  
e operacionalização de  
Modelos de Machine Learning).

PyData São Paulo.

---

# O que veremos neste módulo:

**01**

Roadmap de um  
projeto em  
ciência de dados

**02**

Experiment  
Tracking

**03**

Experiment  
Tracking e  
automação

**04**

Ferramentas  
Open Source

**05**

Noções de  
CI/CD

**06**

Montando seu  
primeiro ML CI

**07**

CI avançado com  
Git Ops

# **Alinhamento de Expectativas**

**Após o curso você será capaz de:**



- 1) Entender onde um cientista de dados se encaixa na entrega de uma projeto
- 2) Ferramentas que podem automatizar coisas maçantes no dia-a-dia.



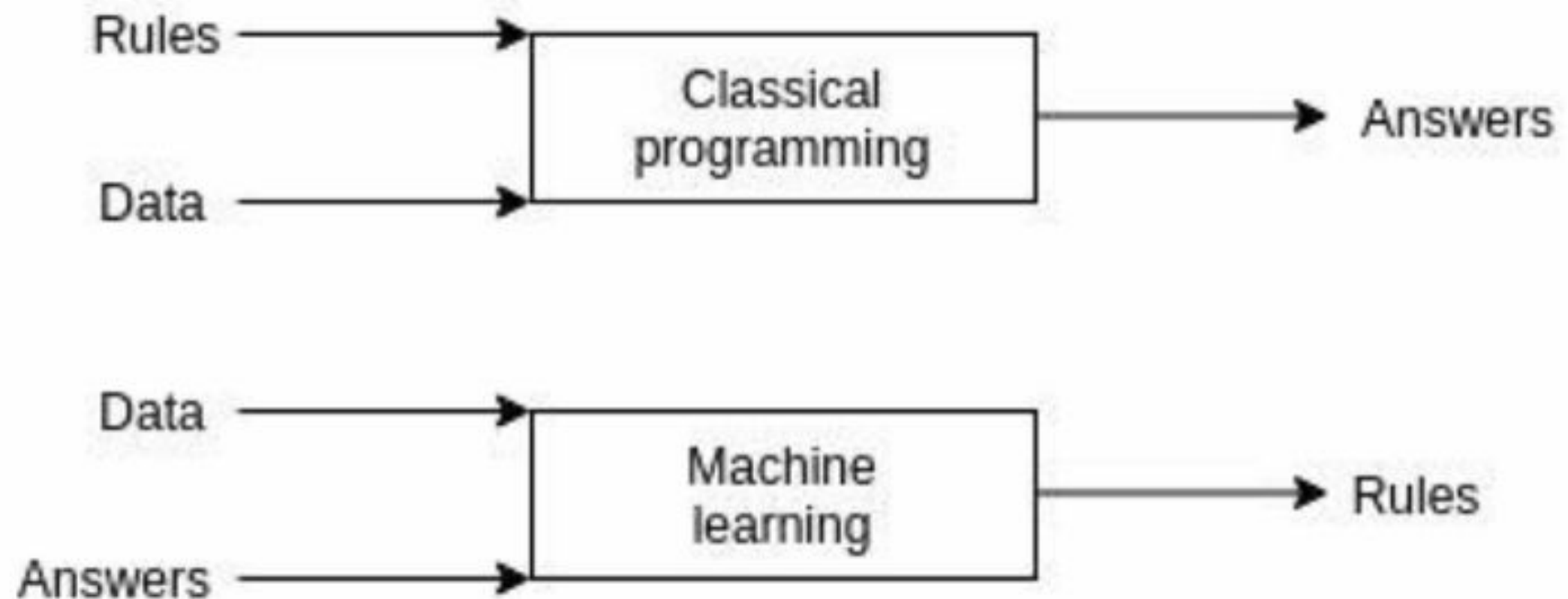
- 3) Percepções práticas em relação ao: Projeto e à entrega de valor
- 4) Habilidades práticas: Te aproximam da possibilidade de produtizar aplicações de ciência de dados.



## **2 - Visão de Experimentação em um Projeto de Software**

**Consultor:** Ricardo Manhães Savii

# Projeto de Software

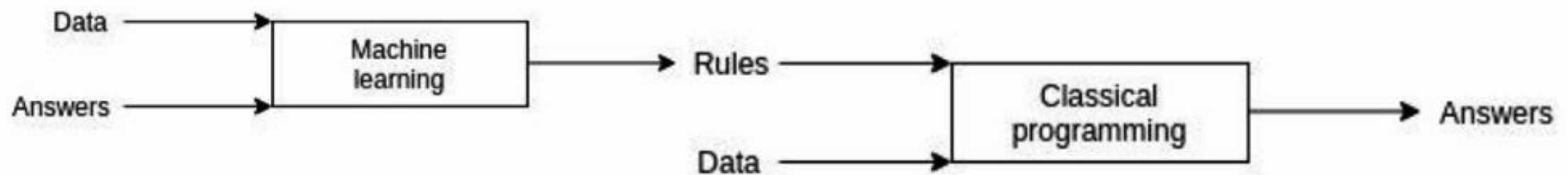


source: Deep Learning with Python by François Chollet

Leitura:

<https://medium.com/@ricardosavii/trying-to-turn-machine-learning-into-value-de9f28cde056>

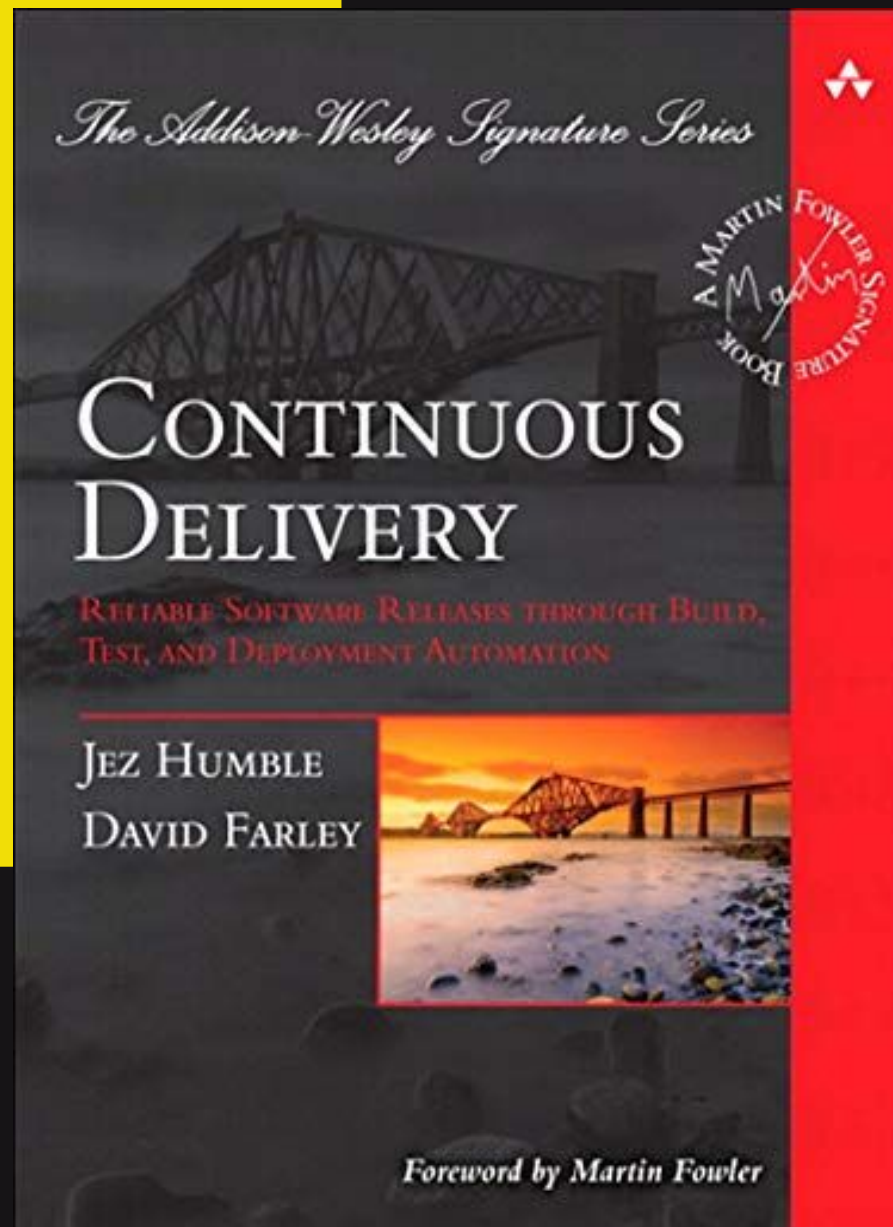
# Projeto de Software



Leitura:

<https://medium.com/@ricardosavii/trying-to-turn-machine-learning-into-value-de9f28cde056>

# Como entrega?



**Leitura:**

<https://www.amazon.com.br/Continuous-Delivery-Deployment-Automation-Addison-Wesley-ebook/dp/B003YMNVCO>



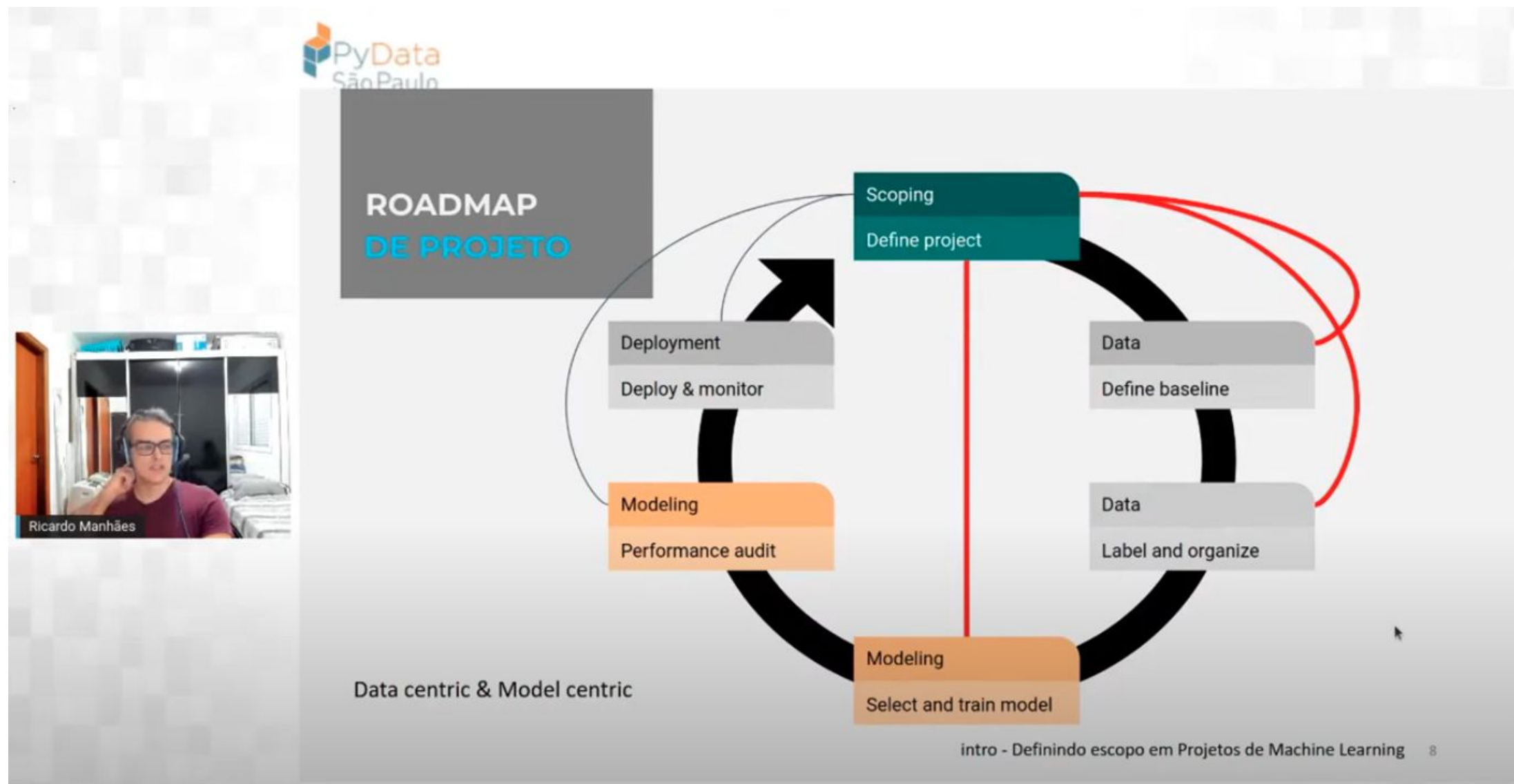
# **Software precisa de um bom código**



**Leitura:**

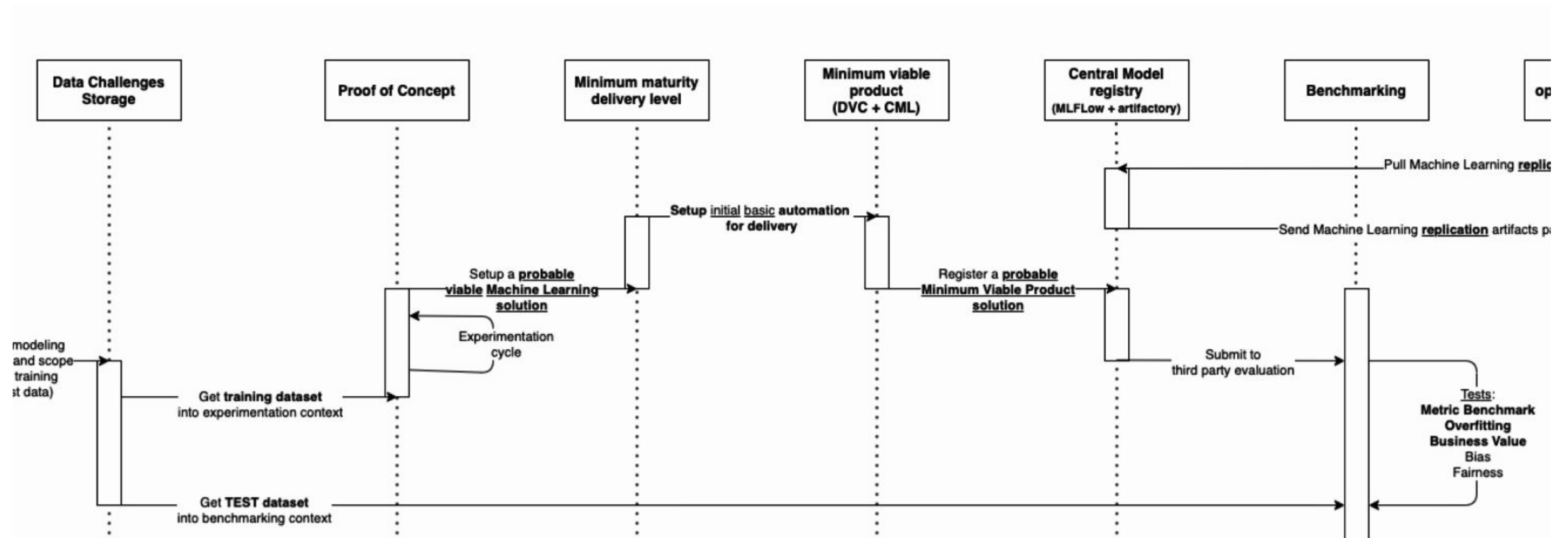
[https://www.amazon.com.br/Fluent-Python-Concise-Effective-Programming-ebook/dp/B0131L3PW4/ref=sr\\_1\\_1](https://www.amazon.com.br/Fluent-Python-Concise-Effective-Programming-ebook/dp/B0131L3PW4/ref=sr_1_1)

# E organização



<https://youtu.be/YLAKGcQ0ttw>

# Porque?



Leitura:

<https://medium.com/@ricardosavii/trying-to-turn-machine-learning-into-value-de9f28cde056>

# Projeto de Equipe

*“AI models, however, are ultimately a software artifact like any other that needs to be integrated within an application. The trouble with MLOps as it is most often pursued today is data scientists are constructing AI models in almost complete isolation from the rest of the organization.”*

— — *"The Road to AI Hell Starts with Good MLOps Intentions"* by

*Dmitry Petrov*

Leitura: <https://thenewstack.io/the-road-to-ai-hell-starts-with-good-mlops-intentions/>

# Resumo

Machine Learning é software, com diferenças, mas é um **software**.

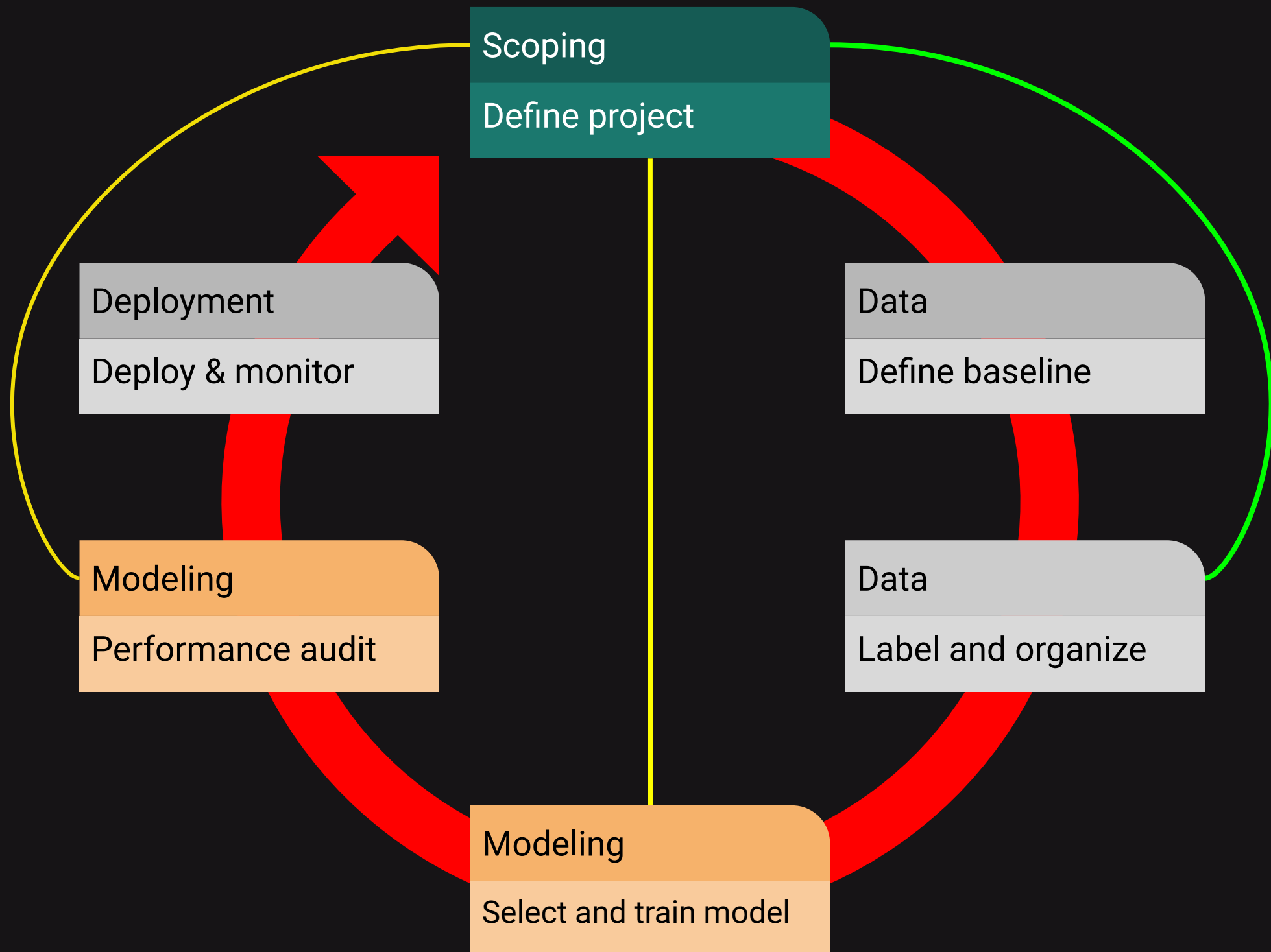
O trabalho de um cientista de dados é muitas vezes pensado como livre e criativo. Mas exige organização, protocolo e uso correto de ferramentas se quiser garantir resultados.



## **3- Experiment Cycle Tracking**

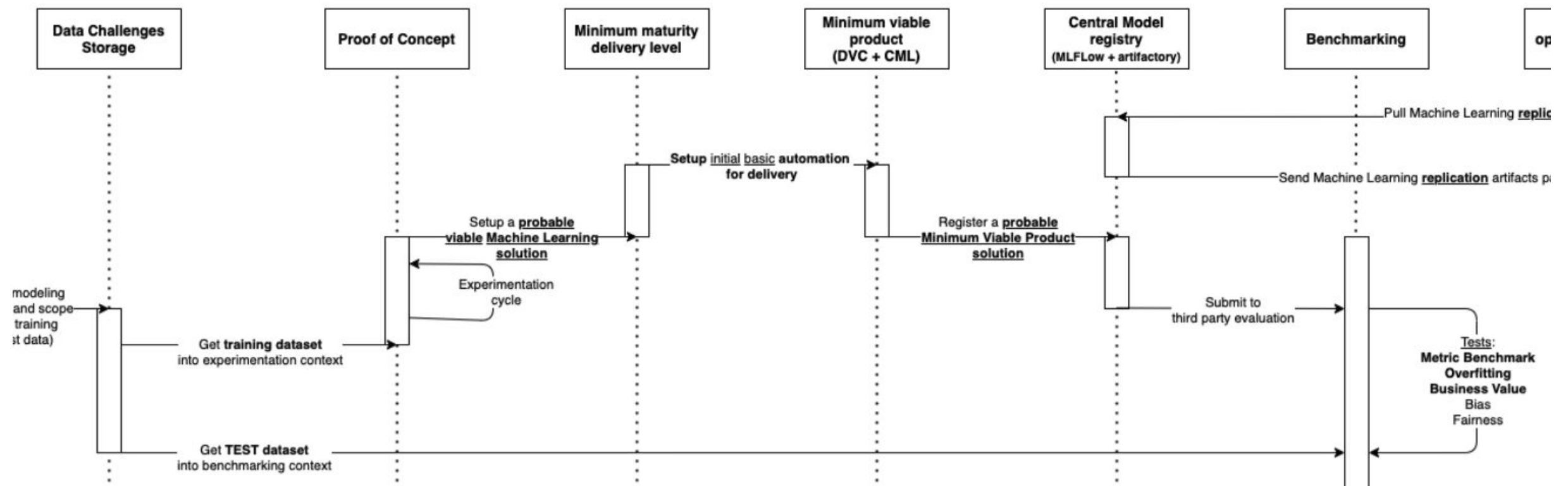
**Consultor:** Ricardo Manhães Savii

# Ciclo de Experimentação





# Esse ciclo é só uma parte





# Projeto de EQUIPE

*“AI models, however, are ultimately a software artifact like any other that needs to be integrated within an application. The trouble with MLOps as it is most often pursued today is data scientists are constructing AI models in almost complete isolation from the rest of the organization.”*

— — *"The Road to AI Hell Starts with Good MLOps Intentions"* by

*Dmitry Petrov*

Leitura:

<https://thenewstack.io/the-road-to-ai-hell-starts-with-good-mlops-intentions/>

# Automatize coisas!

“A well-engineered pipeline gets data scientists iterating much faster, which can be a big competitive edge” From *Engineering Practices in Data Science* By Chris Clark.

Leitura: <http://blog.untrod.com/2012/10/engineering-practices-in-data-science.html>

# Terminologia

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Reproducibility (Different team, different experimental setup )\*
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Replicability (Different team, same experimental setup )\*
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Leitura: <https://www.acm.org/publications/policies/artifact-review-badging>

# Replicabilidade automação

## What data scientists need to know about DevOps

A philosophical and practical guide to using continuous integration (via GitHub Actions) to build an automatic model training system.



Elle O'Brien  • July 16, 2020 • 12 min read • [No comments](#)



Leitura: <https://dvc.org/blog/devops-for-data-scientists>



# Resumo

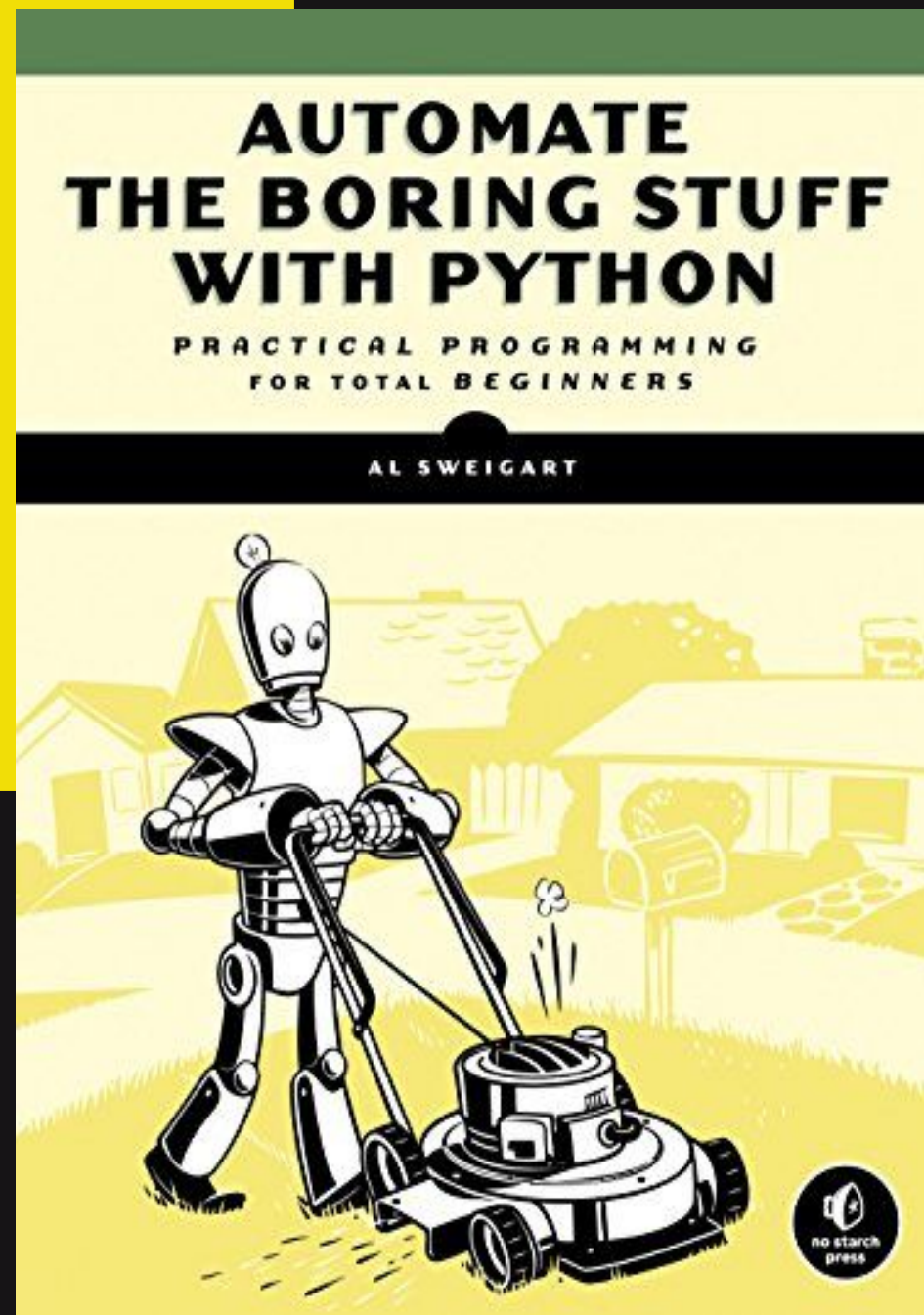
Replicabilidade e Reprodutibilidade são melhores quando há o mínimo de passos manuais.  
Pessoas erram, computadores são feitos para repetir operações incessantemente!



## **4 - Automatize coisas maçantes**

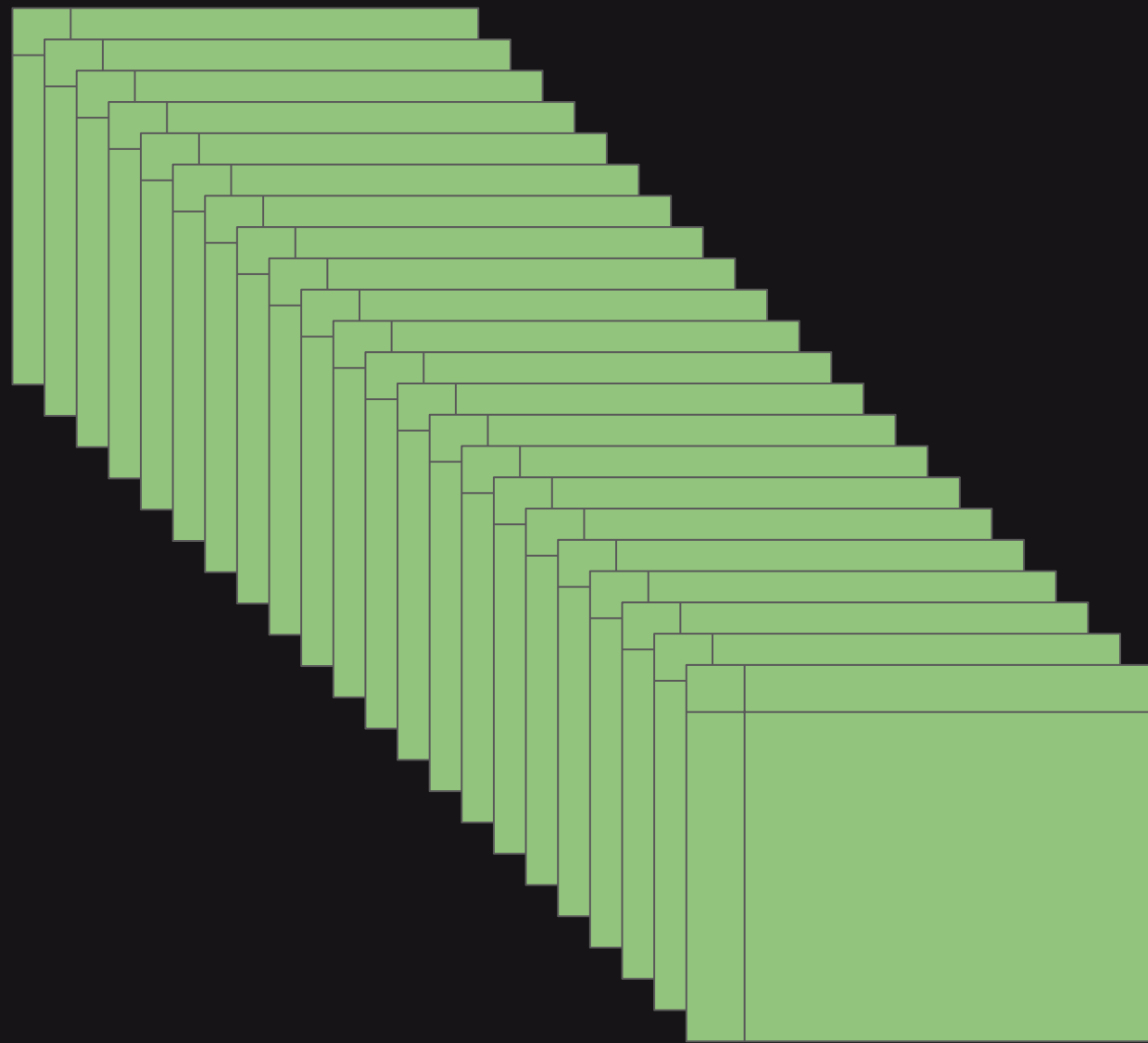
**Consultor:** Ricardo Manhães Savii

# Boring stuff



<https://www.amazon.com.br/Automate-Boring-Stuff-Python-Programming-ebook/dp/B00WJ049VU>

# O que é maçante?






# O que é maçante?



# Ciência de dados



The screenshot shows a Jupyter Notebook interface. At the top, the title bar reads "jupyter covid\_19\_dashboard" followed by "Last Checkpoint: Last Friday at 11:45 PM (unsaved changes)". The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running, and viewing. The notebook contains five code cells:

```
In [13]: # importing libraries

from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets

In [14]: # loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')

In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```



# Ciência de dados

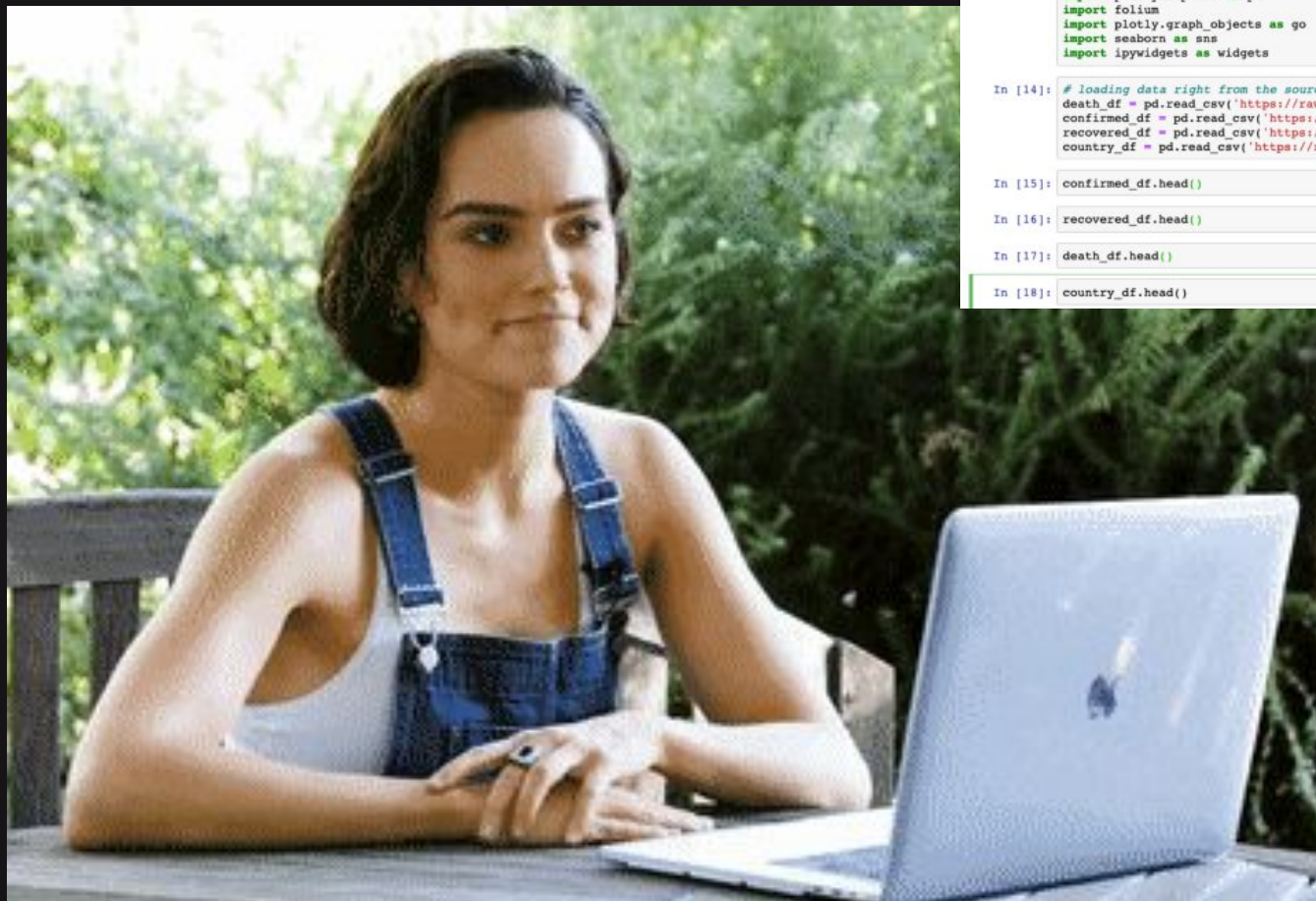


```
jupyter covid_19_dashboard Last Checkpoint: Last Friday at 11:45 PM (unsaved changes) ✓  
File Edit View Insert Cell Kernel Widgets Help  
+ -> Run Code  
In [13]: # importing libraries  
from __future__ import print_function  
from ipywidgets import interact, interactive, fixed, interact_manual  
from IPython.core.display import display, HTML  
  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import plotly.express as px  
import folium  
import plotly.graph_objects as go  
import seaborn as sns  
import ipywidgets as widgets  
  
In [14]: # loading data right from the source:  
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/death_df.csv')  
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df.csv')  
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df.csv')  
country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')  
  
In [15]: confirmed_df.head()  
  
In [16]: recovered_df.head()  
  
In [17]: death_df.head()  
  
In [18]: country_df.head()
```





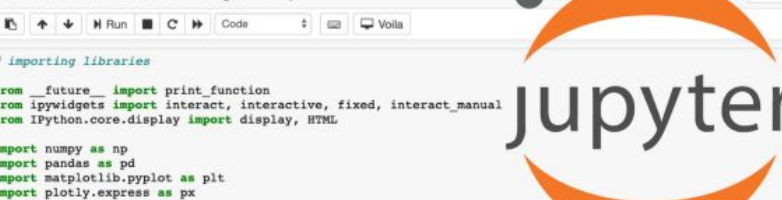
# Ciência de dados



```
jupyter covid_19_dashboard Last Checkpoint: Last Friday at 11:45 PM (unsaved changes) ✓  
File Edit View Insert Cell Kernel Widgets Help  
+ -> Run Code Voila  
In [13]: # importing libraries  
from __future__ import print_function  
from ipywidgets import interact, interactive, fixed, interact_manual  
from IPython.core.display import display, HTML  
  
import numpy as np  
import pandas as pd  
import matplotlib.pyplot as plt  
import plotly.express as px  
import folium  
import plotly.graph_objects as go  
import seaborn as sns  
import ipywidgets as widgets  
  
In [14]: # loading data right from the source:  
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/death_df.csv')  
confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df.csv')  
recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df.csv')  
country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')  
  
In [15]: confirmed_df.head()  
  
In [16]: recovered_df.head()  
  
In [17]: death_df.head()  
  
In [18]: country_df.head()
```



# Ciência de dados



The screenshot shows a JupyterLab environment. At the top, there's a header with the 'jupyter' logo, 'dashboards', and a status message: 'Last Checkpoint: Last Friday at 11:45 PM (unsaved changes)'. Below this is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, Help. To the right of the menu bar are icons for 'Trusted', a pencil icon, and 'Python 3'. The left sidebar contains a file explorer with a list of notebooks: 'In [13]:', 'In [14]:', 'In [15]:', 'In [16]:', 'In [17]:', and 'In [18:]'. The main workspace displays the code for the selected notebook, 'In [13]:'. The code is as follows:

```
# importing libraries

from __future__ import print_function
from ipywidgets import interact, interactive, fixed, interact_manual
from IPython.core.display import display, HTML

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import folium
import plotly.graph_objects as go
import seaborn as sns
import ipywidgets as widgets
```

The code continues with the next cell, 'In [14]:', which loads data from GitHub:

```
# loading data right from the source:
death_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/confirmed_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/recovered_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_data/country_df = pd.read_csv('https://raw.githubusercontent.com/CSSEGISandData/COVID-19/web-data/data/cases_country.csv')
```

The subsequent cells show the head of each dataframe:

```
In [15]: confirmed_df.head()

In [16]: recovered_df.head()

In [17]: death_df.head()

In [18]: country_df.head()
```

The Jupyter logo is overlaid on the right side of the image.



# Automatize coisas!

“A well-engineered pipeline gets data scientists iterating much faster, which can be a big competitive edge” From *Engineering Practices in Data Science* By Chris Clark.

Leitura:

<http://blog.untrod.com/2012/10/engineering-practices-in-data-science.html>

# Resumo

As vezes **nos acostumamos** tanto com coisas maçantes e **perdemos muito tempo** com coisas que **não trazem valor**.

Avalie bem o tempo que você gasta com o que realmente traz valor e o que não.

**Espero** que este módulo te traga alguns insights!



# **5 - Conhecendo os Pokémons da área**

**Consultor:** Ricardo Manhães Savii



# Pokémon or MLOps



<https://valohai.com/mlops-or-pokemon/>

# Reinventar a roda?

## 56 Groundbreaking Python Open-source Projects – Get started with Python

Python is booming and so is its Github page. This year was great for Python and we saw some very powerful python open-source projects to contribute to. Today, we're listing down some of the top python open-source projects; try contributing to at least one of these, it will help improve your Python skills.

## 56 Python Open-source Projects

The screenshot shows a Google Scholar search for 'scikit-learn'. The search bar at the top contains 'scikit-learn' and a search icon. Below the search bar, it indicates 'Articles' and 'About 65,500 results (0.04 sec)'. On the left side, there are filters for 'Any time' (with options: Since 2021, Since 2020, Since 2017, Custom range...), 'Sort by relevance' (with 'Sort by date' selected), 'Any type' (with 'include patents' unchecked and 'include citations' checked), 'Review articles', and 'Create alert'. The main results area displays three entries:

- [PDF] Scikit-learn: Machine learning in Python** by F Pedregosa, G Varoquaux, A Gramfort... - the Journal of machine ..., 2011 - jmlr.org. Description: Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level ... Cited by 47321.
- [HTML] Machine learning for neuroimaging with scikit-learn** by A Abraham, F Pedregosa, M Eickenberg... - Frontiers in ..., 2014 - frontiersin.org. Description: Statistical machine learning methods are increasingly used for neuroimaging data analysis. Their main virtue is their ability to model high-dimensional datasets, eg multivariate analysis of activation images or resting-state time series. Supervised learning is typically used in ... Cited by 874.
- Scikit-learn** by O Kramer - Machine learning for evolution strategies, 2016 - Springer. Description: ... scikit-learn is widely used in many commercial applications and is also part of many research projects and publications. The scikit-learn implementations are the basis of many methods introduced in this book. To improve efficiency, some algorithms are implemented ... Cited by 50.

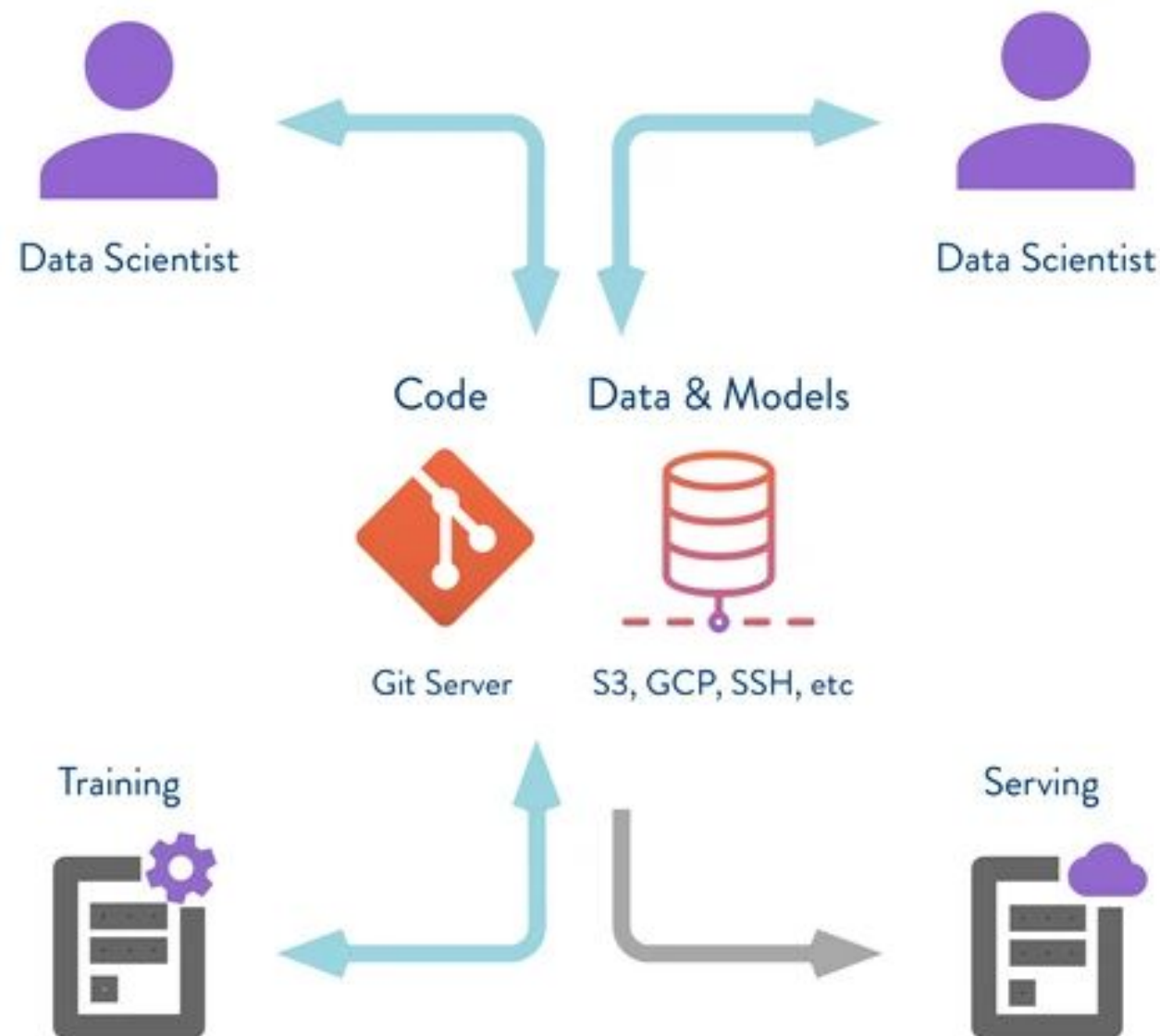
At the bottom, there is a link to a book: **[book] Mastering Machine Learning with scikit-learn** by G Hackling - 2017 - books.google.com.

<https://data-flair.training/blogs/python-open-source-projects/>  
[https://scholar.google.com/scholar?hl=en&as\\_sdt=0%2C5&q=scikit-learn&btnG=](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=scikit-learn&btnG=)

# Experiment tracking tools

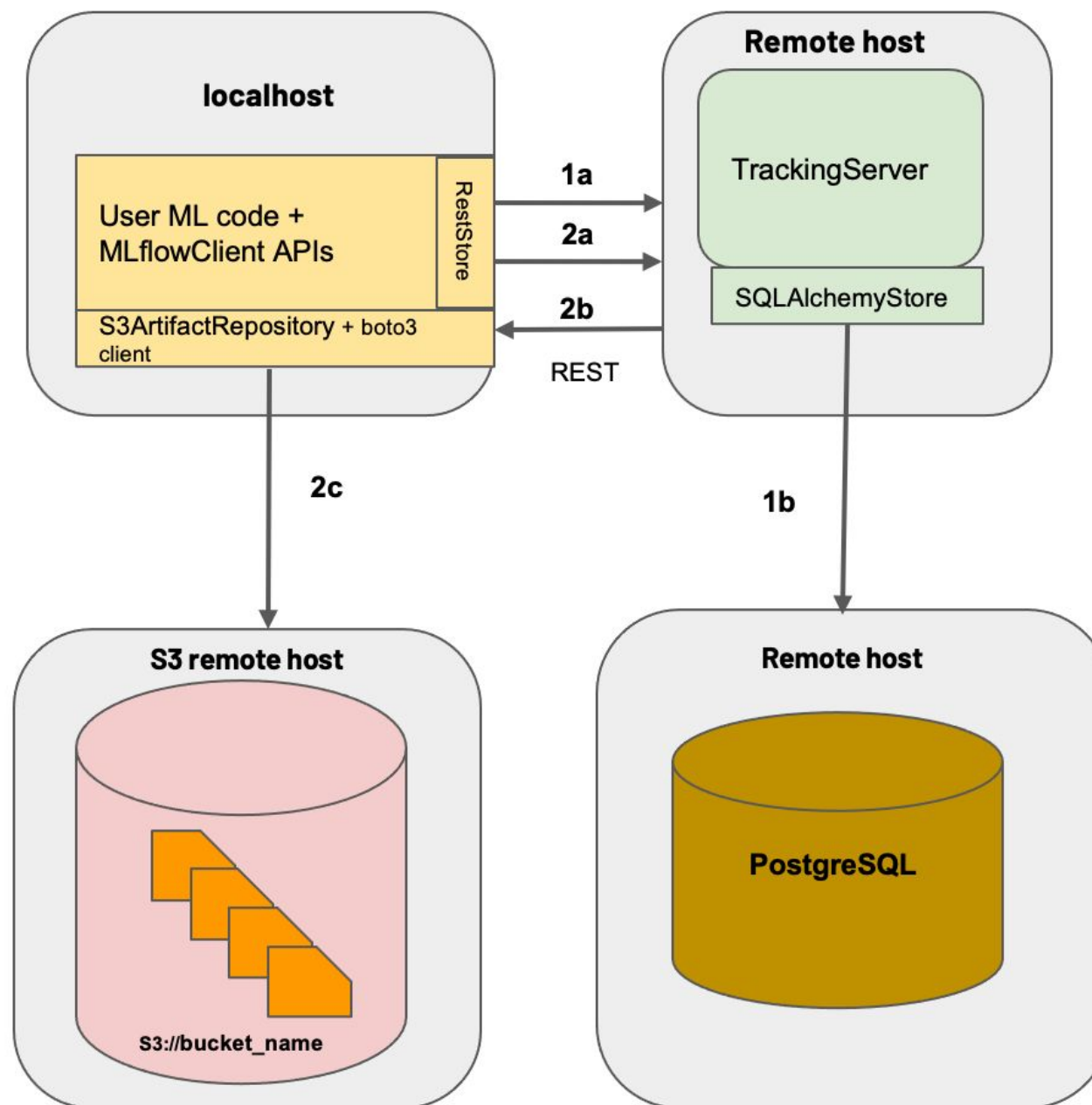
The logo for DVC (Data Version Control) features the letters 'D', 'V', and 'C' in a stylized, overlapping font. The 'D' is teal, the 'V' is purple, and the 'C' is orange.The logo for mlflow features the text 'mlflow' in a sans-serif font. The 'ml' is black, and the 'flow' is blue. The 'o' in 'flow' is stylized with a circular arrow inside it.

# DVC (data version control)



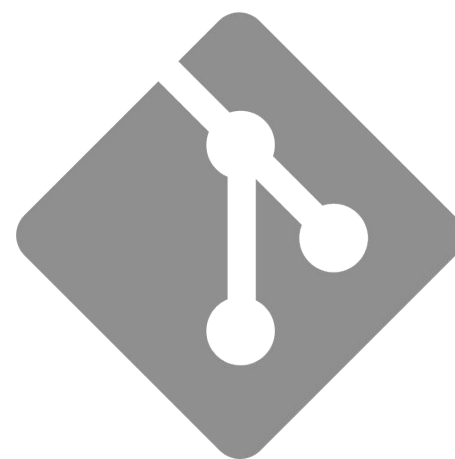


# MLflow



**Scenario 4:** `mlflow server --backend-store-uri postgresql://URI --default-artifact-root S3://bucket_name --host remote_host`

# Diferenças (DVC)



git

# Diferenças (MLflow)

MLflow is an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry. MLflow currently offers four components:

## MLflow Tracking

Record and query experiments: code, data, config, and results

[Read more](#)

## MLflow Projects

Package data science code in a format to reproduce runs on any platform

[Read more](#)

## MLflow Models

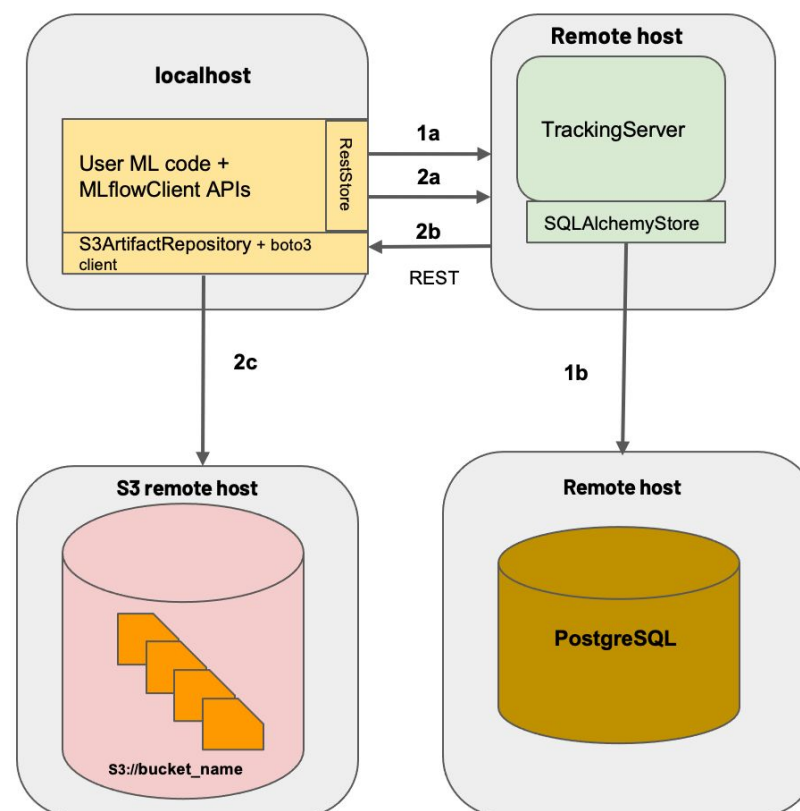
Deploy machine learning models in diverse serving environments

[Read more](#)

## Model Registry

Store, annotate, discover, and manage models in a central repository

[Read more](#)



# Resumo

Machine Learning é uma área em extrema expansão de conhecimento e aplicações, não é fácil acompanhar.

Se manter atualizado fará toda a diferença, pois muitos projetos interessantes surgem a todo momento.

**DVC\*** e **MLflow** são dois projetos que te ajudarão muito na questão de **controle** e **replicabilidade** de seus projetos.





## **6 - Replicação na Prática**

**Consultor:** Ricardo Manhães Savii

# Replicabilidade

## Terminology


A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology(VIM); see the [Appendix](#) for details.

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Reproducibility (Different team, different experimental setup )\*
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Replicability (Different team, same experimental setup )\*
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Fonte:

<https://www.acm.org/publications/policies/artifact-review-badging>


# Projeto divórcios

 Search or jump to... Pull requests Issues Marketplace Explore


ricoms / divorce-predictor Public

<> Code Issues Pull requests Actions Projects Wiki Security

main 2 branches 0 tags

 ricoms ran new experiment

.dvc	setup initial project
.github/workflows	correct push comm
divorce_predictor	set experiment for D
dvc_plots	setup new experime
ml	ran new experiment
scripts	setun initial project

**UCI** Machine Learning Repository  
Center for Machine Learning and Intelligent Systems

Check out the [beta version](#) of the new UCI Machine Learning Repository we are currently testing! [Contact us](#) if you h

## Divorce Predictors data set Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** Participants completed the Personal Information Form and Divorce Predictors Scale.

<b>Data Set Characteristics:</b>	Multivariate, Univariate	<b>Number of Instances:</b>	170	<b>Area:</b>	Life
<b>Attribute Characteristics:</b>	Integer	<b>Number of Attributes:</b>	54	<b>Date Donated</b>	2019-07-24
<b>Associated Tasks:</b>	Classification	<b>Missing Values?</b>	N/A	<b>Number of Web Hits:</b>	120430

**Source:**  
Dr. Mustafa Kemal Yöntem, Nevşehir Hacı Bektaş Veli University, Faculty of Education, Department of Educational Sciences, [mus](#)  
Dr. Kemal ADEM, Alanya University, Faculty of Education and Administrative Sciences, Department of Management Information

# DVC

[Features](#)[Doc](#)[Blog](#)[Community ▼](#)

## Open-source Version Control System for Machine Learning Projects



Download  
(macOS)



Watch video  
How it works



We're on [GitHub](#) ★ 8857

<https://dvc.org/>



# Vamos ao código

The screenshot shows a VS Code editor window titled "experiment.py — Untitled (Workspace)". The Explorer sidebar on the left shows a project structure with folders like "divorce-predictor" and "experiment", and files like "cml\_on\_pr.yaml" and "train\_main.yaml". The main editor area displays the code for "experiment.py".

```
divorce-predictor > divorce_predictor > experiment > experiment.py > Experiment > setup
1  import json
2  from pathlib import Path
3
4  import numpy as np
5  import pandas as pd
6  from joblib import dump
7  from sklearn.dummy import DummyClassifier
8  from sklearn.model_selection import cross_validate
9
10 from divorce_predictor.base import BaseExperiment
11
12
13 class Experiment(BaseExperiment):
14     def __init__(self, X: np.ndarray, y: np.ndarray):
15         self.X = X
16         self.y = y
17
18     def setup(self):
19         self.scoring = ["roc_auc", "accuracy"]
```

A tooltip for the `cross_validate` function is visible, showing its signature and a brief description: "(function) cross\_validate: (estimator, X, y=None, \*, groups=None, scoring=None, cv=None, n\_jobs=None, verbose=0, fit\_params=None, pre\_dispatch='all', error\_score=np.nan) -> Any".

The terminal window at the bottom shows the output of a `git push` command, indicating that the code has been pushed to the `main` branch of the `divorce-predictor` repository.

```
Check JSON.....(no files to check)Skipped
Check python ast.....(no files to check)Skipped
Check for added large files.....Passed
Detect Private Key.....Passed
black.....(no files to check)Skipped
flake8.....(no files to check)Skipped
[main e966bc4] correct push command to main
1 file changed, 1 insertion(+), 1 deletion(-)
→ divorce-predictor git:(main) git push
```

# Resumo

Nesta aula eu demonstrei **repetibilidade**.

Para demonstrar **replicabilidade** precisarei de  
você.

No exercício prático espero que consiga instalar o  
projeto e reproduzi-lo. Boa sorte!

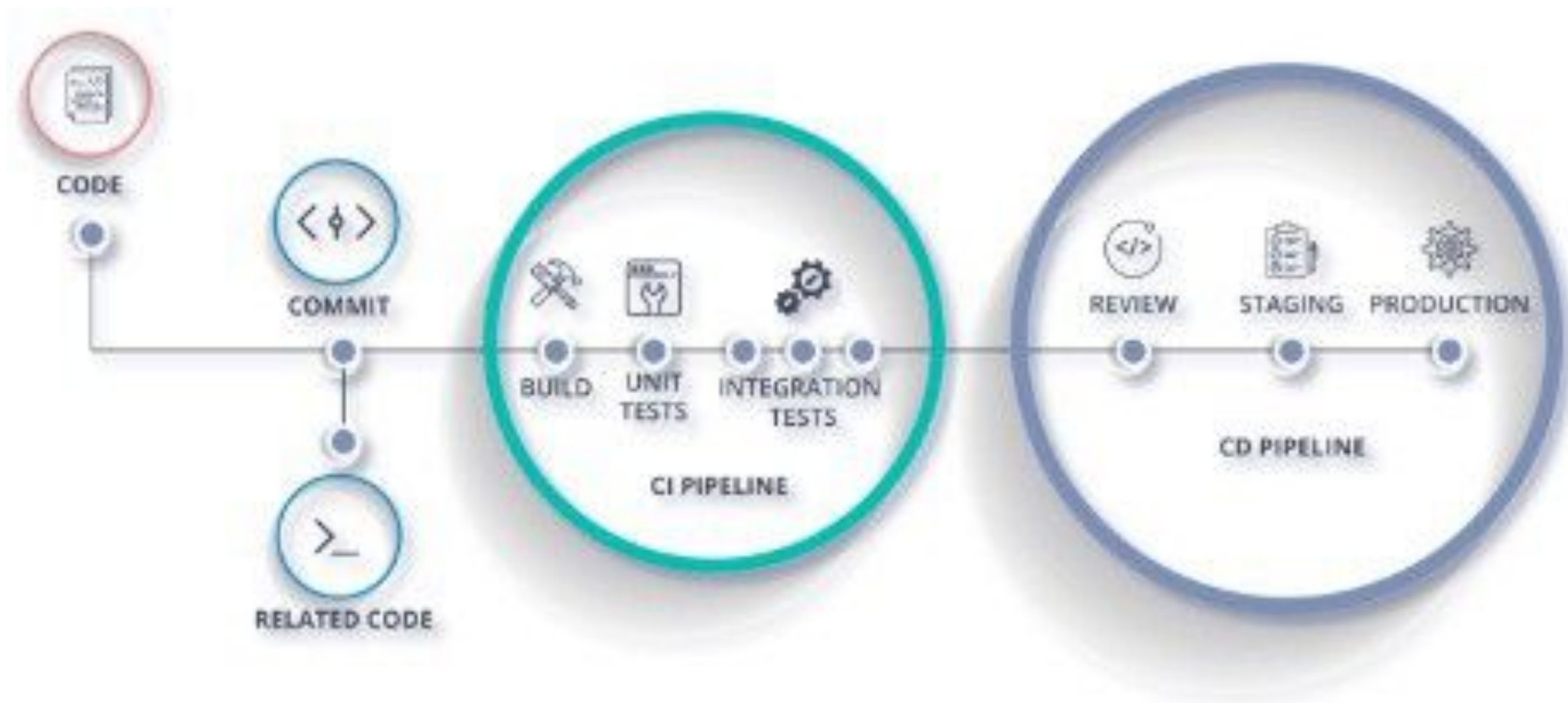




# **10 - Noções de CI/CD**

**Consultor:** Ricardo Manhães Savii

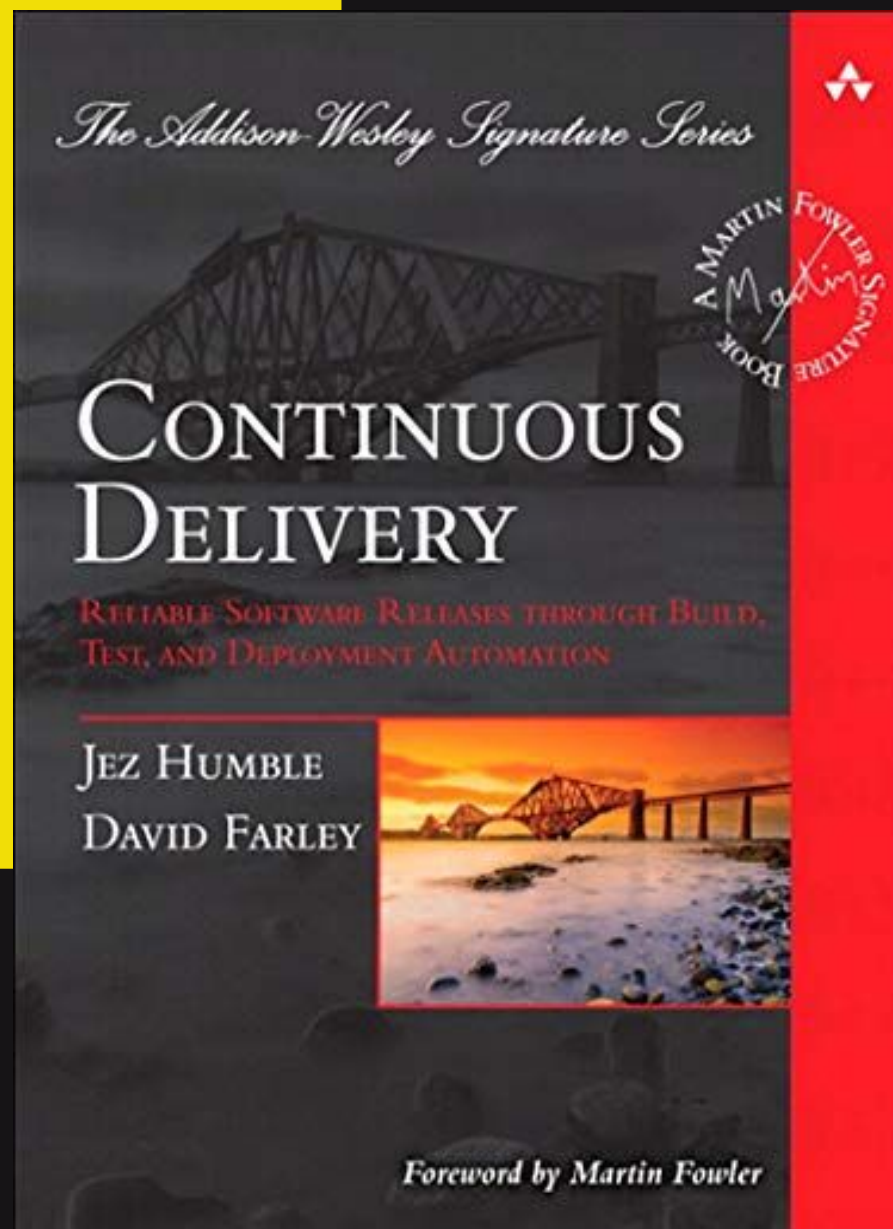
# CI/CD



Fonte:

<https://medium.com/mlearning-ai/trying-to-turn-machine-learning-into-value-de9f28cde056>

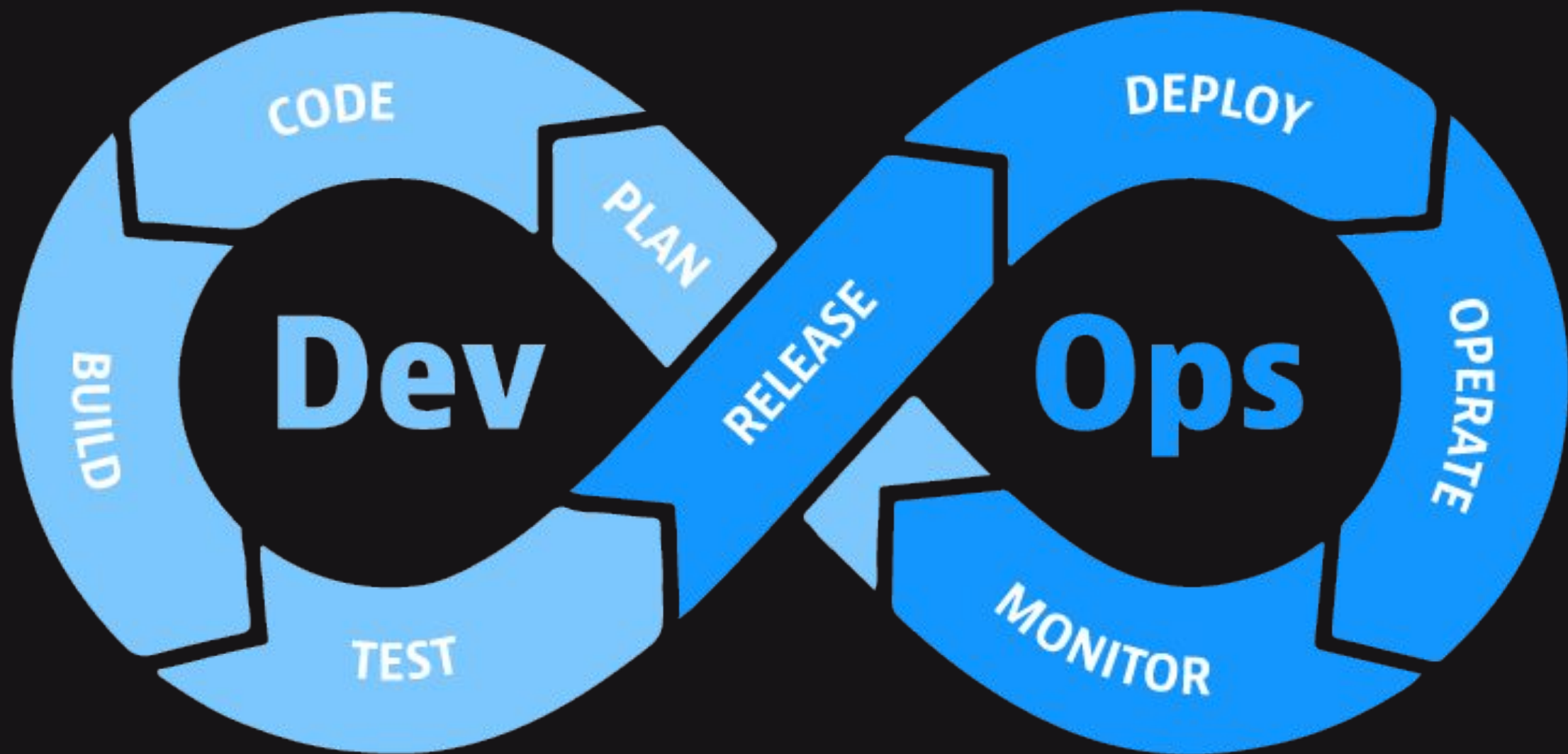
# Continuous Delivery



**Leitura:**

<https://www.amazon.com.br/Continuous-Delivery-Deployment-Automation-Addison-Wesley-ebook/dp/B003YMNVCO>

# DevOps



<https://www.dynatrace.com/news/blog/what-is-devops/>

# Plataformas de CI/CD



<https://github.com/features/actions>  
<https://www.jenkins.io/>



# Vamos ao github

The screenshot shows the GitHub interface for the repository 'ricoms / divorce-predictor'. The repository is public and has a 'main' branch with 2 branches and 0 tags. The commit history shows a recent commit by 'ricoms' titled 'ran new experiment' with the hash 'fb1f451'. The file list includes folders like '.dvc', '.github/workflows', 'divorce\_predictor', 'dvc\_plots', 'ml', 'scripts', and 'tests', as well as files like '.dvcignore', '.editorconfig', '.flake8', and '.gitignore'. Each file has a description of its purpose.

Search or jump to... / Pull requests Issues Marketplace Explore

ricoms / divorce-predictor Public

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

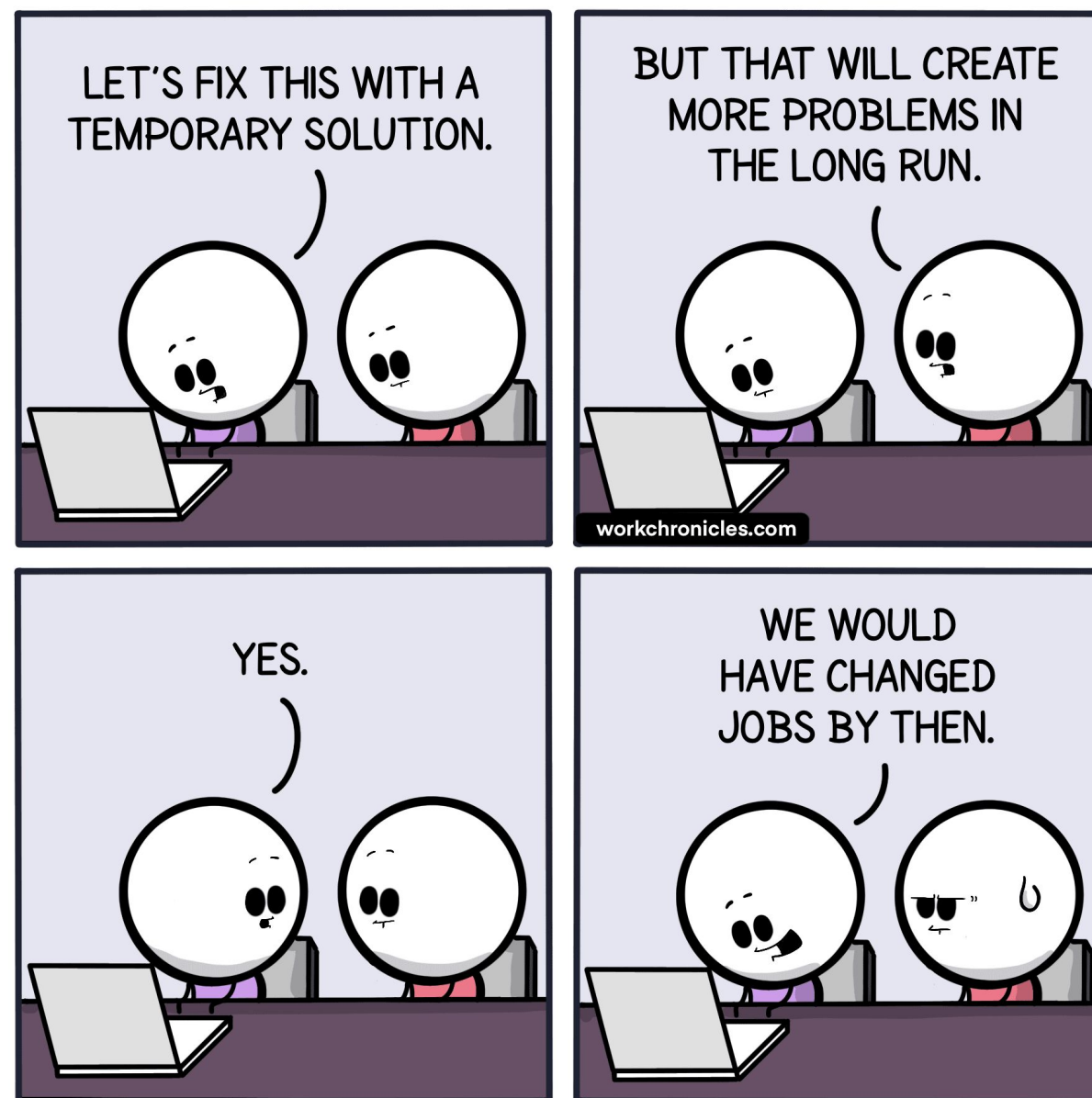
main 2 branches 0 tags Go to file

ricoms ran new experiment fb1f451 1

.dvc	setup initial project setup and initial get_data for dvc repro
.github/workflows	correct push command to main
divorce_predictor	set experiment for DummyClassifier in main
dvc_plots	setup new experiment
ml	ran new experiment
scripts	setup initial project setup and initial get_data for dvc repro
tests	setup entire code experiment and save an initial experiment ru
.dvcignore	setup initial project setup and initial get_data for dvc repro
.editorconfig	setup initial project setup and initial get_data for dvc repro
.flake8	setup initial project setup and initial get_data for dvc repro
.gitignore	test using docker within CI, include .lock file to accelerate insta



# Design de Sistemas



Comics about work. Made with love & lots of coffee.  
Join [r/workchronicles](https://www.reddit.com/r/workchronicles). Or find on Webtoons, IG, Twitter, FB

Work Chronicles  
[workchronicles.com](https://workchronicles.com)

<https://martinfowler.com/bliki/DesignStaminaHypothesis.html>

# Resumo

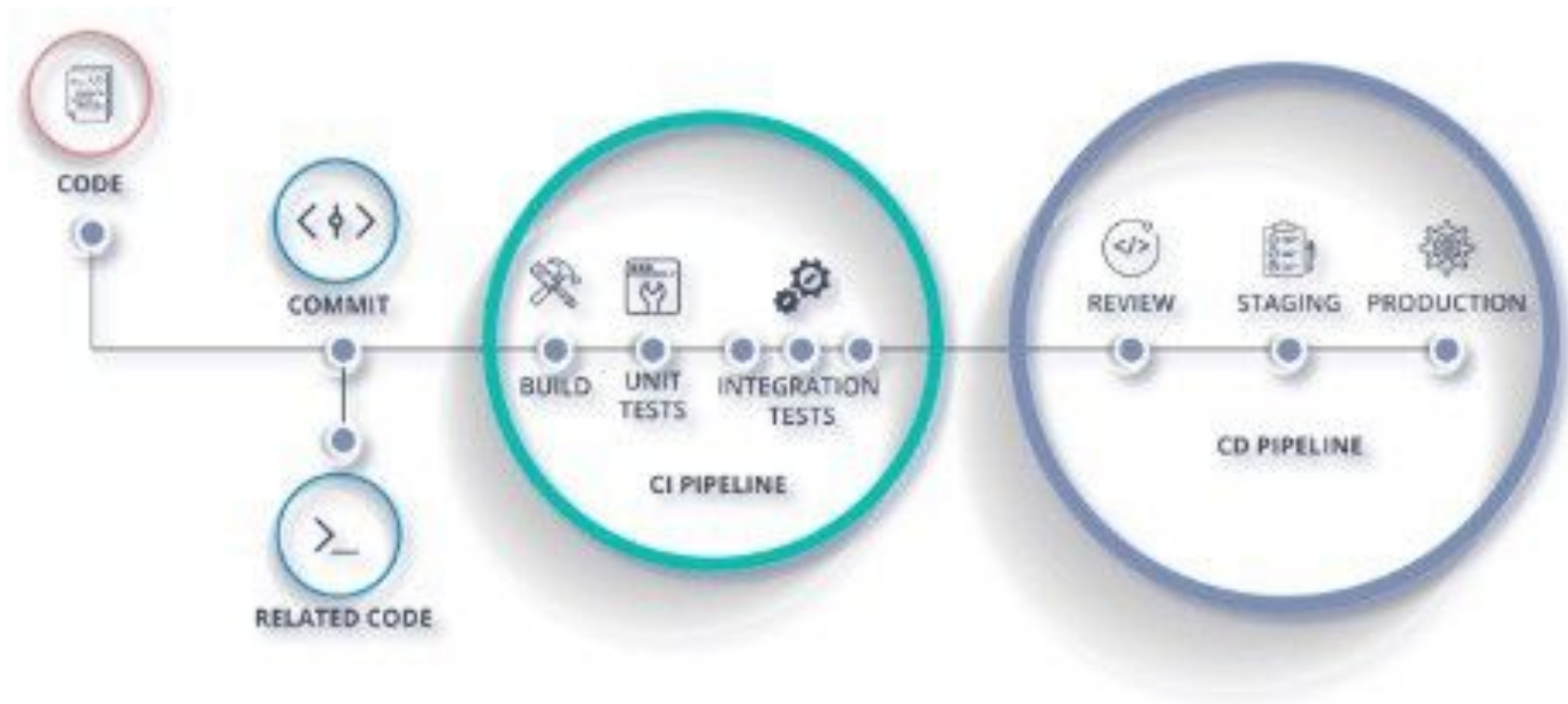
CI/CD, Continuous Delivery, DevOps, MLOps são uma sequência de metodologias e funções que procuram melhores formas de **entregar e garantir funcionalidade** de sistemas computacionais.



# **11 - Montando meu primeiro CI**

**Consultor:** Ricardo Manhães Savii

# CI/CD



Fonte:

<https://medium.com/mlearning-ai/trying-to-turn-machine-learning-into-value-de9f28cde056>

# GitHub Actions

GitHub Actions

Free, Pro, & Team ▾

English ▾

🔍 Search topics,

## GitHub Actions

Automate, customize, and execute your software development workflows right in your repository with GitHub Actions. You can discover, create, and share actions to perform any job you'd like, including CI/CD, and combine actions in a completely customized workflow.

Quickstart

Overview



Guides [View all →](#)

### Learn GitHub Actions

Whether you are new to GitHub Actions or interested in learning all they have to offer, this guide will help you us...

Popular

### Workflow syntax for GitHub Actions

[Learn GitHub Actions](#)

What's new [View all →](#)

### Input types for manual workflows

November 10

[Conditional execution of steps in actions](#)

<https://docs.github.com/en/actions>



# Vamos ao código

The screenshot shows the GitHub interface for the repository 'ricoms / divorce-predictor'. The repository is public and has a 'main' branch with 2 branches and 0 tags. The commit history shows a recent commit by 'ricoms' titled 'ran new experiment' with the hash 'fb1f451'. The file list includes folders like '.dvc', '.github/workflows', 'divorce\_predictor', 'dvc\_plots', 'ml', 'scripts', and 'tests', as well as files like '.dvcignore', '.editorconfig', '.flake8', and '.gitignore'. Each file has a description of its purpose.

Search or jump to... / Pull requests Issues Marketplace Explore

ricoms / divorce-predictor Public

<> Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 2 branches 0 tags Go to file

ricoms ran new experiment fb1f451 1

.dvc	setup initial project setup and initial get_data for dvc repro
.github/workflows	correct push command to main
divorce_predictor	set experiment for DummyClassifier in main
dvc_plots	setup new experiment
ml	ran new experiment
scripts	setup initial project setup and initial get_data for dvc repro
tests	setup entire code experiment and save an initial experiment ru
.dvcignore	setup initial project setup and initial get_data for dvc repro
.editorconfig	setup initial project setup and initial get_data for dvc repro
.flake8	setup initial project setup and initial get_data for dvc repro
.gitignore	test using docker within CI, include .lock file to accelerate insta



# Resumo

Montar um CI (Continuous Integration) não é uma habilidade necessária para um Cientista de Dados.

Mas ter noções deste processo te ajudará a mostrar seu valor em empresas e equipes em que estiver.



# **13 - Resumo do Módulo**

**Consultor:** Ricardo Manhães Savii

# Replicabilidade

## Terminology

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology(VIM); see the [Appendix](#) for details.

- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Reproducibility (Different team, different experimental setup )\*
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Replicability (Different team, same experimental setup )\*
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

# Replicabilidade

## Terminology

A variety of research communities have embraced the goal of reproducibility in experimental science. Unfortunately, the terminology in use has not been uniform. Because of this we find it necessary to define our terms. The following are inspired by the International Vocabulary for Metrology(VIM); see the [Appendix](#) for details.

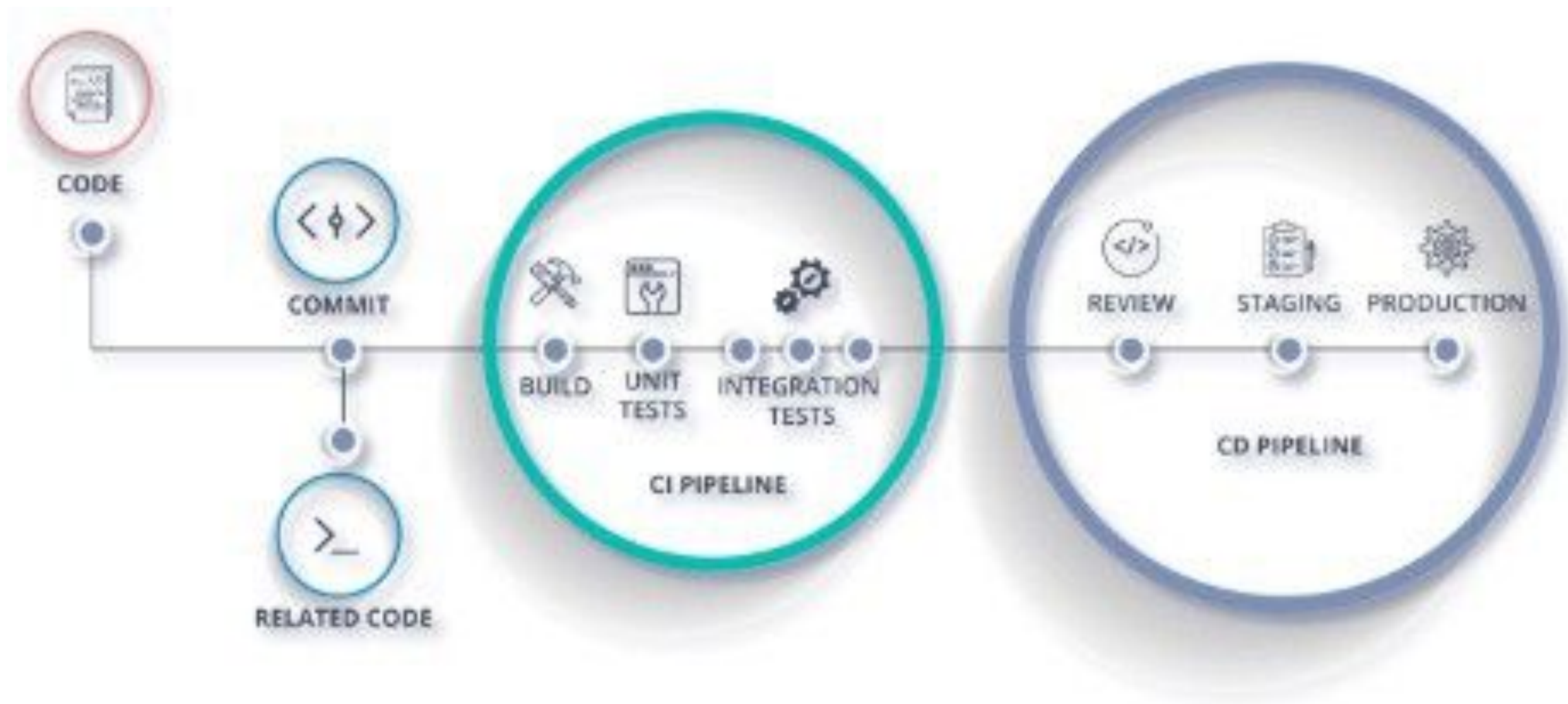
- Repeatability (Same team, same experimental setup)
  - The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation.
- Reproducibility (Different team, different experimental setup )\*
  - The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts.
- Replicability (Different team, same experimental setup )\*
  - The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently.

Fonte:

<https://www.acm.org/publications/policies/artifact-review-badging>



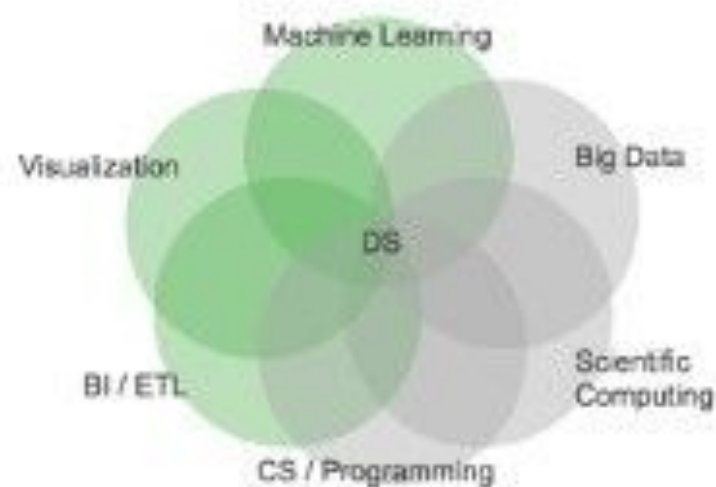
# CI/CD



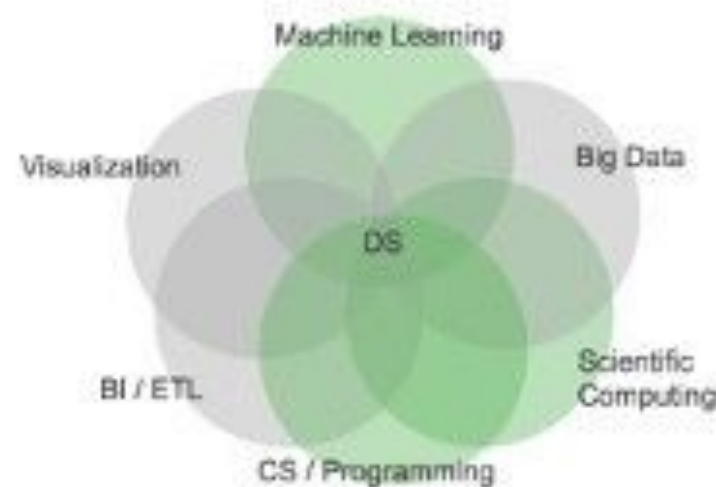
Fonte:

<https://medium.com/mlearning-ai/trying-to-turn-machine-learning-into-value-de9f28cde056>

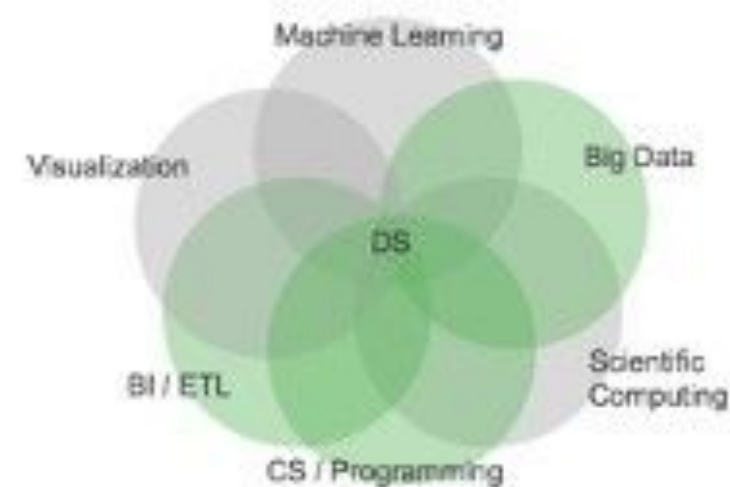
# Equipes e papéis



Statistician / Analyst



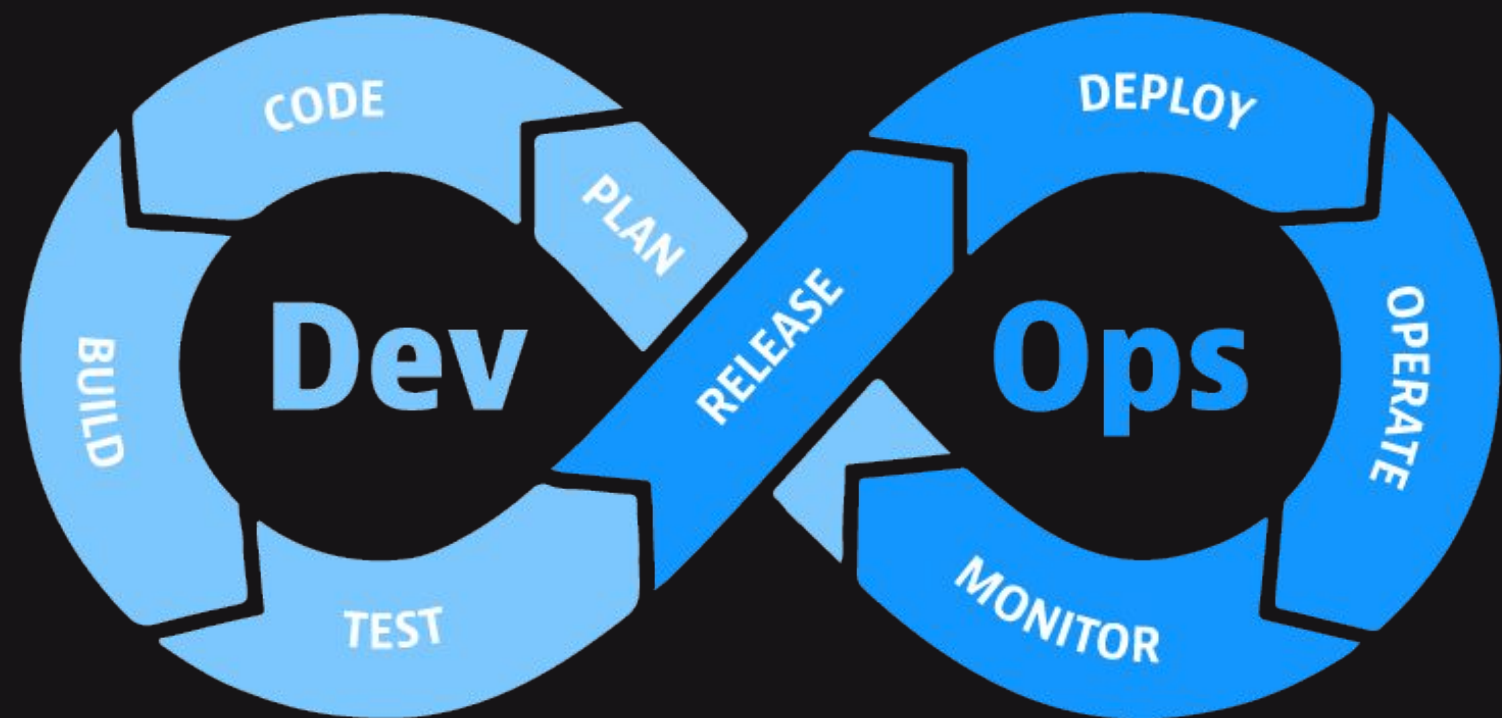
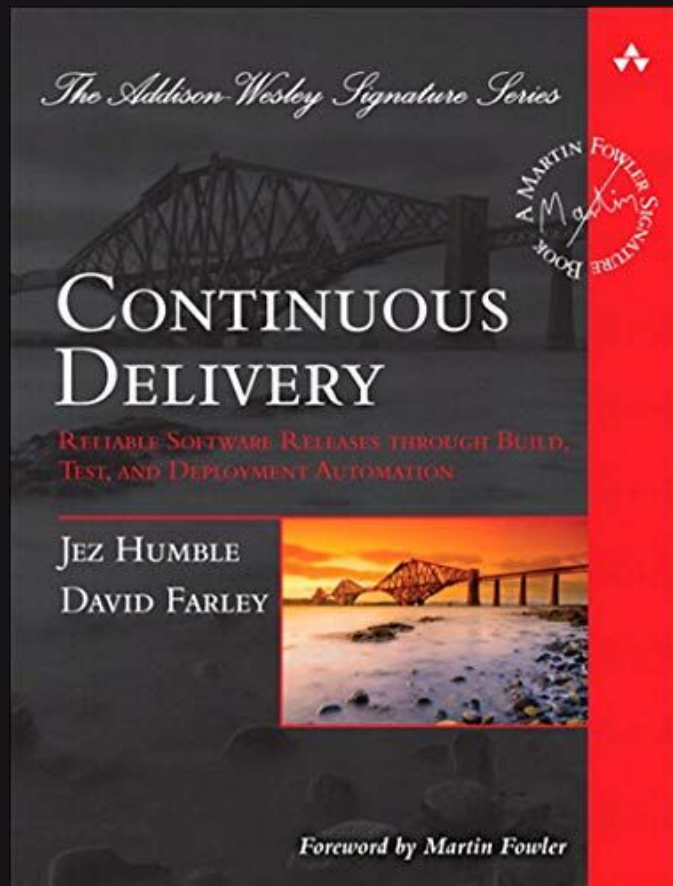
Research / Computational  
Scientist



Developer / Engineer

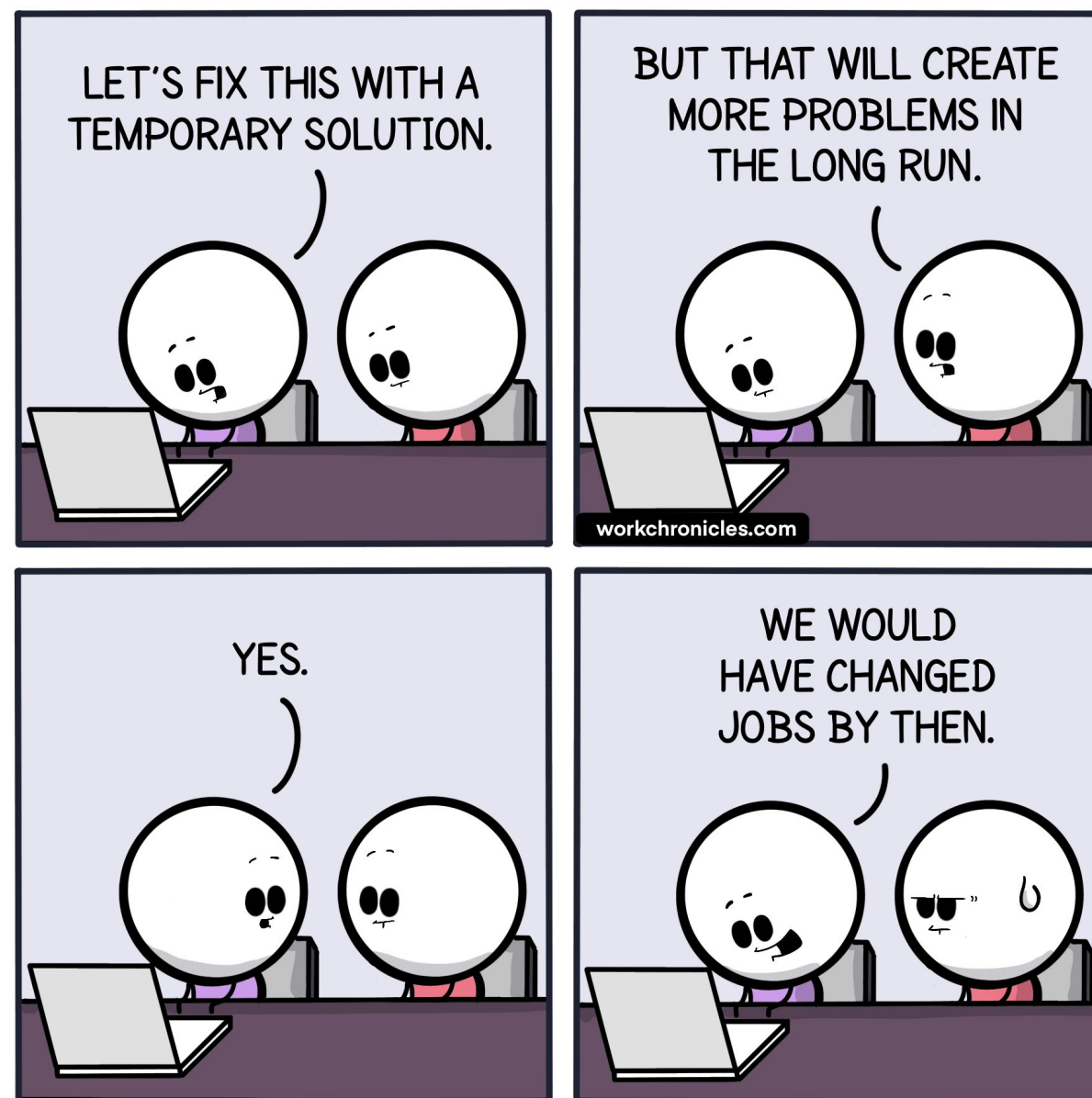


# Continuous Delivery



<https://www.amazon.com.br/Continuous-Delivery-Deployment-Automation-Addison-Wesley-ebook/dp/B003YMNVCO>

# Design de Sistemas



Comics about work. Made with love & lots of coffee.  
Join [r/workchronicles](https://www.reddit.com/r/workchronicles). Or find on Webtoons, IG, Twitter, FB

Work Chronicles  
[workchronicles.com](https://workchronicles.com)

<https://martinfowler.com/bliki/DesignStaminaHypothesis.html>

# Resumo

CI/CD, Continuous Delivery, DevOps, MLOps são uma sequência de metodologias e funções que procuram melhores formas de **entregar e garantir funcionalidade** de sistemas computacionais.



**FIM**