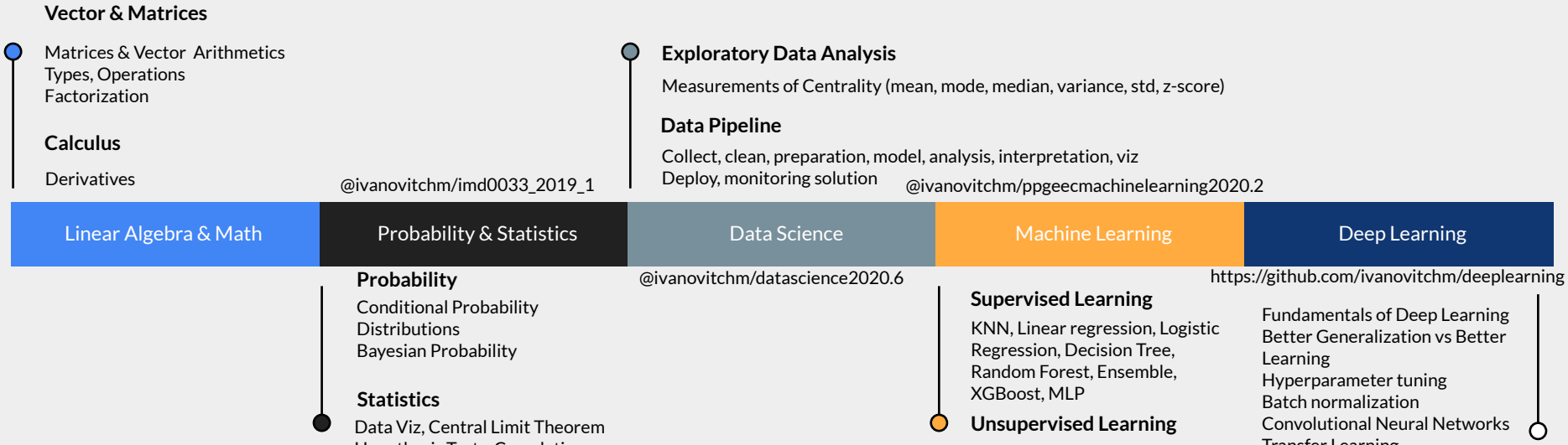
DCA0305
Machine Learning Based Systems Design

ivanovitch.silva@ufrn.br
@ivanovitchm

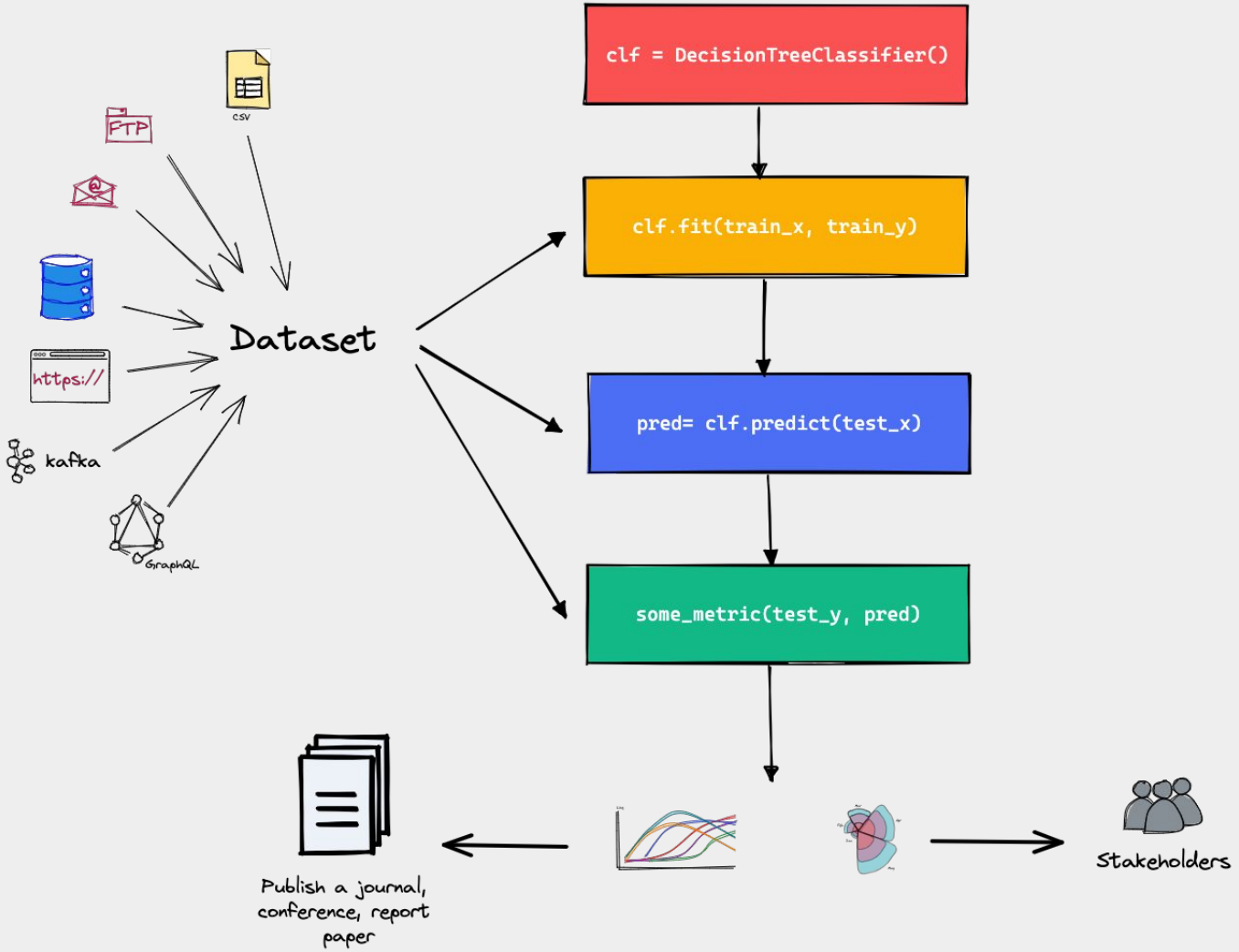
UFRN UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

DCA

A Long Pathway



A typical ML workflow



My first ML models in production

when you write very long code

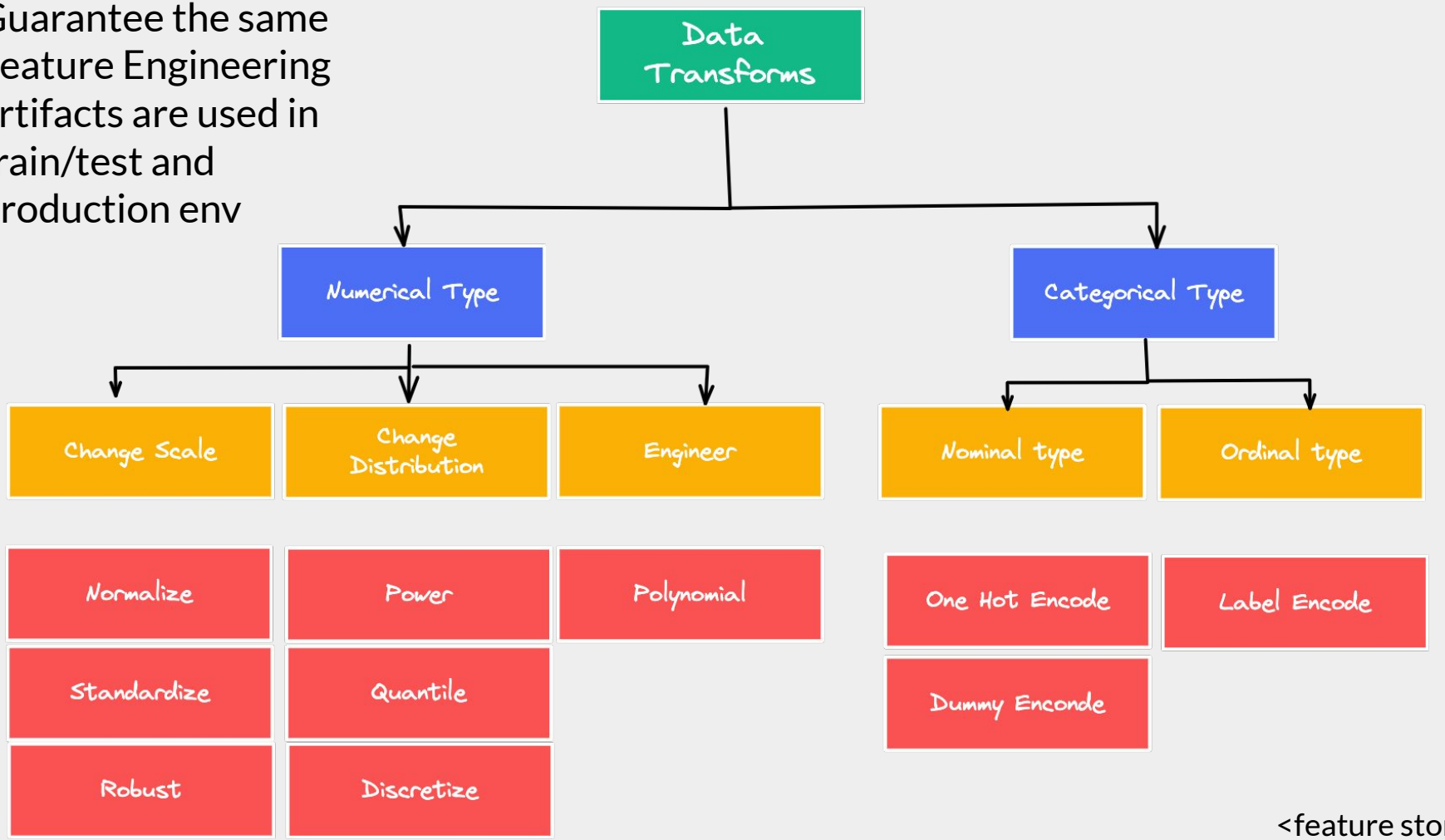
NAMING VARIABLES

Giving them meaningful names, according to their use.

Giving them the most compact names possible, for less storage usage.

Giving them random names like "ahshjdn" or "yeetus".

Guarantee the same Feature Engineering artifacts are used in train/test and production env



<feature store>

education sciences

MDPI

Article

Predictive Models for Imbalanced Data: A School Dropout Perspective

Thiago M. Barros [1,*,†], Plácido A. Souza Neto [1,†] and Ivanovitch Silva [2,†] and Luiz Affonso Guedes [2,†]

1 Federal Institute of Rio Grande do Norte (IFRN), 1559 Tirol Natal, Brazil; placido.neto@
2 Federal University of Rio Grande do Norte (UFRN), 59078-970 Natal, Brazil; ivan@imd.
affonso@dca.ufrn.br (L.A.G.)
* Correspondence: thiago.medeiros@ifrn.edu.br
† These authors contributed equally to this work.

Received: 24 July 2019; Accepted: 10 November 2019; Published: 15 November 2019

Abstract: Predicting school dropout rates is an important issue for the smooth
educational system. This problem is solved by classifying students into two classes
activities related statistical datasets. One of the classes must identify the students who have the tendency to persist. The other class must identify the students who have the tendency to dropout. This problem often encounters a phenomenon that masks out the obtained results. This study delves into this phenomenon and provides a reliable educational data mining technique that accurately predicts the dropout rates. In particular, the three data classifying techniques, namely, decision tree, neural networks and Balanced Bagging, are used. The performances of these classifies are tested with and without the use of a downsample, SMOTE and ADASYN data balancing. It is found that among other parameters geometric mean and UAR provides reliable results while predicting the dropout rates using Balanced Bagging classifying techniques.

Keywords: dropout rates; accuracy paradox; imbalanced learning; downsample; g-mean predict; mlp; decision tree; Balanced Bagging; UAR; SMOTE; ADASYN

Concept/Data Drift

g1 EDUCAÇÃO

Evasão escolar de crianças e adolescente aumenta 171% na pandemia, diz estudo

Levantamento da organização Todos Pela Educação mostra que 244 mil crianças de 6 a 14 anos estavam fora da escola no segundo trimestre de 2021.

Por g1 ... São Paulo
...lizado há 3 meses

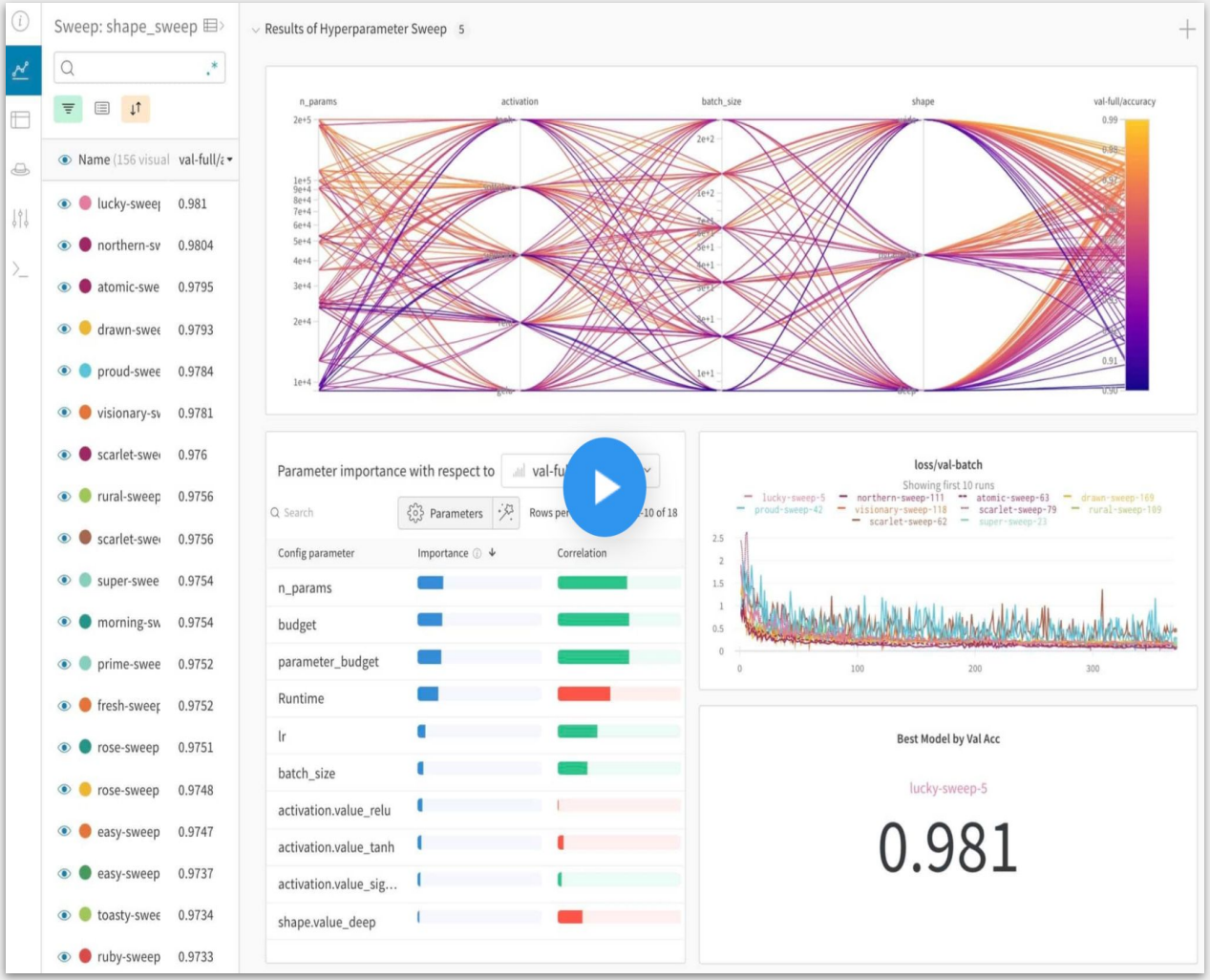
...is de 650 mil crianças ...am da escola durante a ...demia

Pela primeira vez desde 2005, país registrou queda de matrículas na educação infantil

PERSPECTIVA 2022

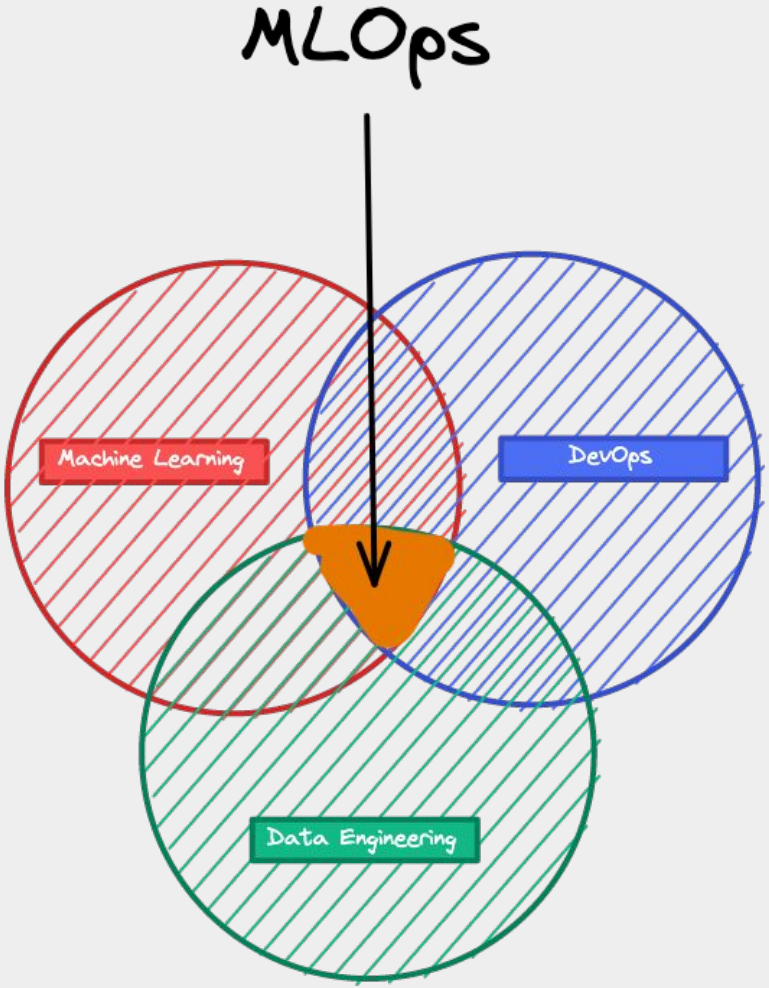
Evasão escolar bate recordes durante a pandemia

Colégios abertos, mas sem alunos. Com cerca de 240 mil estudantes fora das salas de aula a desistência é o maior desafio de 2022. Especialistas afirmam que esse é o pior cenário em 20 anos



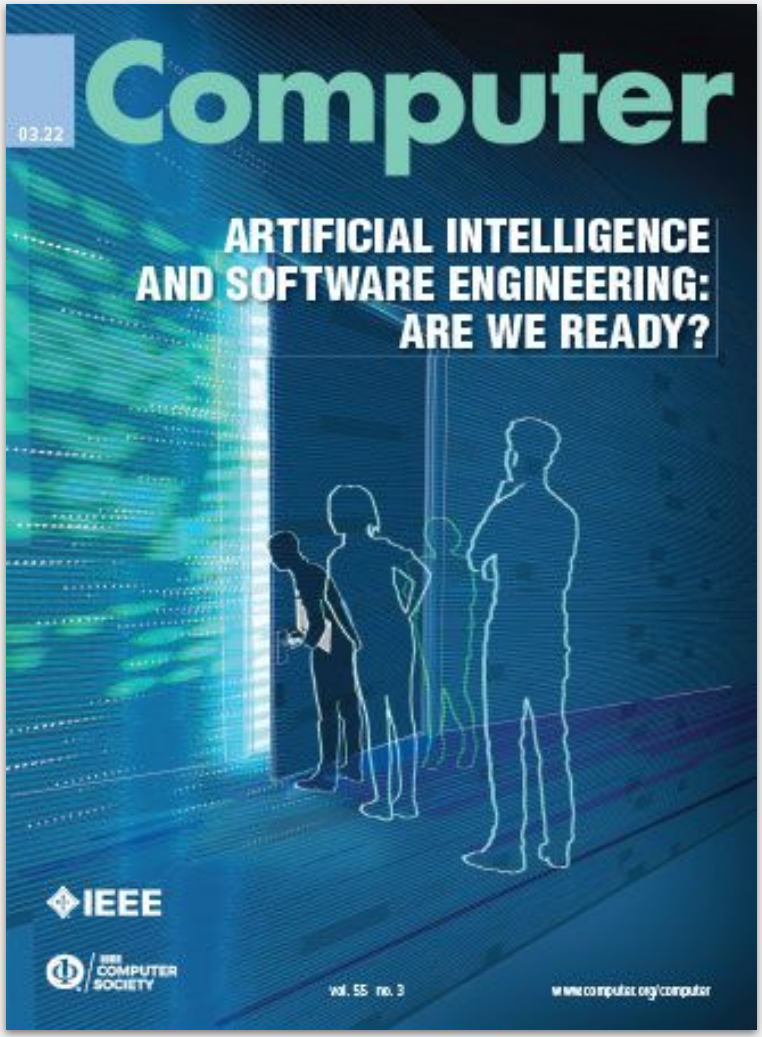
What settings were used in the last experiment?

experiment tracking
dataset versioning
model management



Syllabus

Introduction to Data Science
Clean Code Principles
Fundamentals of ML
Building Reproducible Model Workflow
Deploy a Scalable ML Pipeline in Production
ML Model Scoring and Monitoring

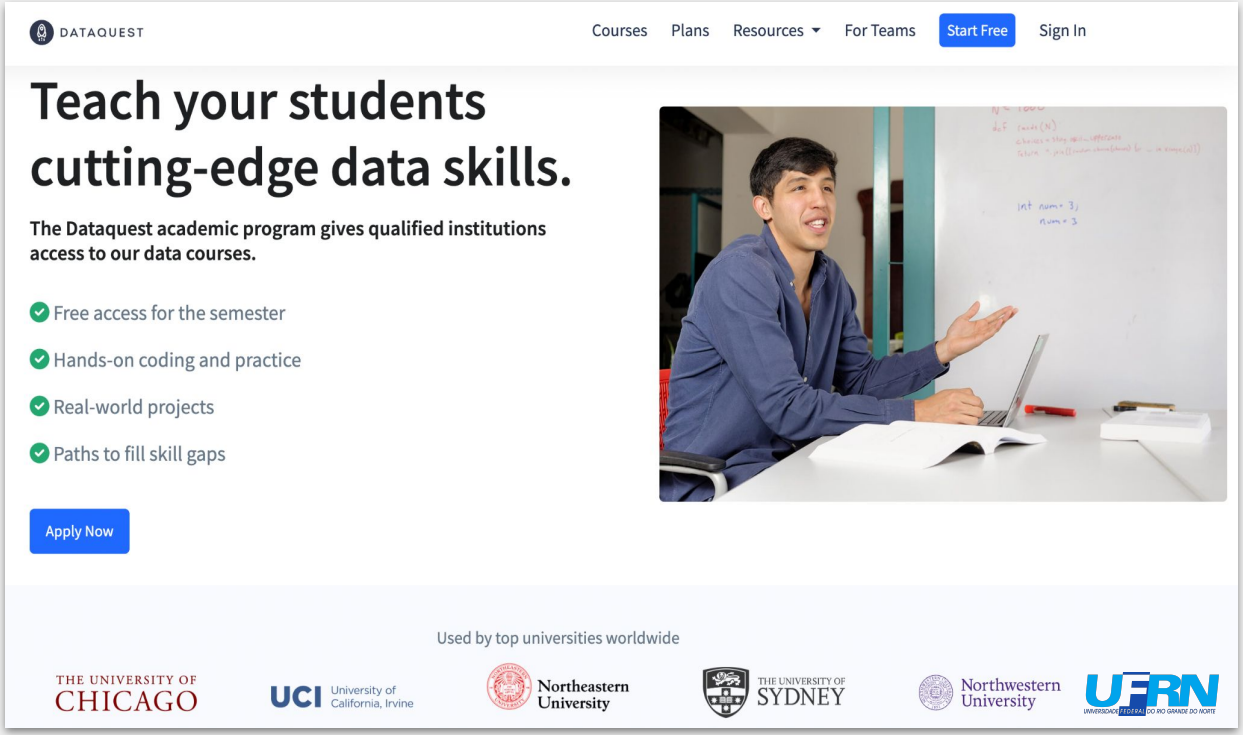


S. Warnett and U. Zdun, "*Architectural Design Decisions for the Machine Learning Workflow*" in **Computer**, vol. 55, no. 03, pp. 40-51, 2022. Doi: 10.1109/MC.2021.3134800

H. Washizaki, et al.,"*Software-Engineering Design Patterns for Machine Learning Applications*" in **Computer**, vol. 55, no. 03, pp. 30-39, 2022. doi: 10.1109/MC.2021.3137227

A. Mashkoor, T. Menzies, A. Egyed and R. Ramler, "*Artificial Intelligence and Software Engineering: Are We Ready?*" in **Computer**, vol. 55, no. 03, pp. 24-28, 2022. doi: 10.1109/MC.2022.3144805

R. Sangwan, Y. Badr, S. Srinivasan and P. Mukherjee, "*On the Testability of Artificial Intelligence and Machine Learning Systems*" in **Computer**, vol. 55, no. 03, pp. 101-105, 2022. doi: 10.1109/MC.2021.3132710



Flipped Classroom

Active Learning

Complementary Material

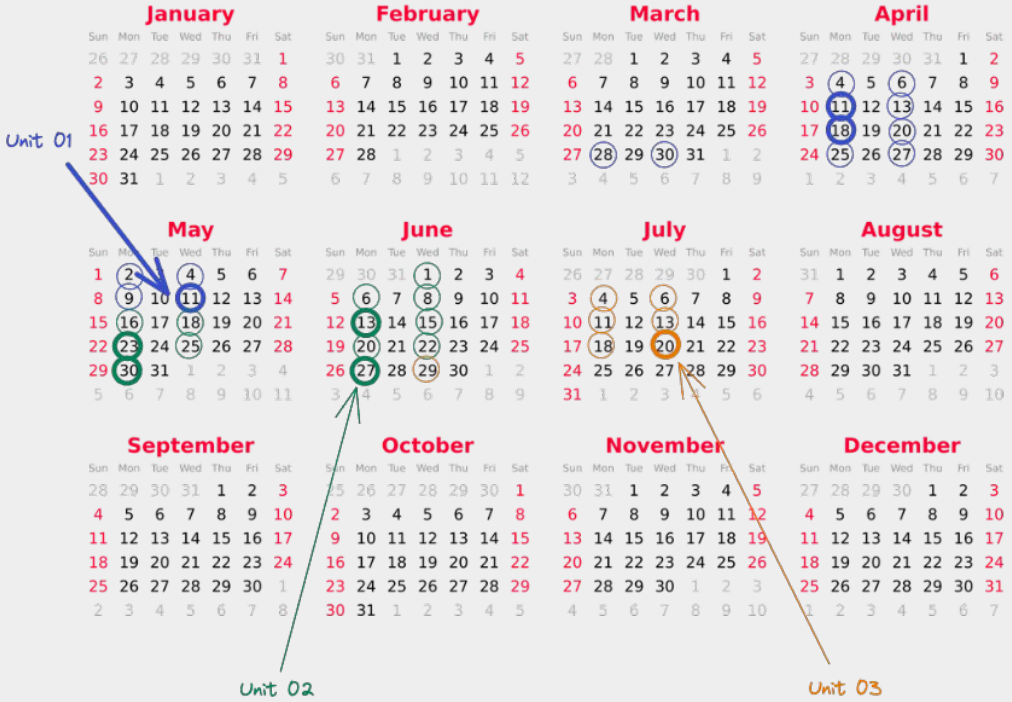
Career Paths

Data Scientist With Python
Data Analyst With R
Data Analyst in Python
Data Engineering
Business Analyst

Dataquest Method

Learn with real-world data
Complete exercises and get feedback
Build your portfolio with projects

2022
Calendar



Unit 01

28/03 - Planning
30/03 - Planning
04/04 - Course Outline
06/04 - DQ git and version control
11/04 - UT1 (10% grade)
- DQ Elements of CLI
13/04 - DQ Text processing CLI
18/04 - UT2 (20% grade)
- Clean Code principles
- DQ Function: advanced
20/04 - Clean Code principles
25/04 - Produciton ready code
27/04 - Production ready code
02/05 - Project
04/05 - Project
09/05 - Project
11/05 - Project
- UT3 (70% grade)

Unit 02

16/05 - Building a data pipeline
- DQ Context managers
- DQ Introduction to decorators
18/05 - Building a data pipeline
- DQ Decorators: advanced
23/05 - U2T1 (20% grade)
- Building a reproducible model workflow (BRW) PART I
- Intro MLOps, Tools, Env, ML Pipes
- MLFlow
25/05 - BRW PART I
30/05 - U2T2 (20% grade)
- BRW PART II
- Fundamentals of ML
01/06 - BRW PART II
06/06 - BRW PART III
- ETL, Data Checks
- Data Segragation
08/06 - BRW PART III
13/06 - U2T3 (20% grade)
- BRW PART IV
- Train, Validaton
- Experiment Tracking
15/06 - BRW PART IV
20/06 - BRW PART V
- Final pipeline, release and deploy
22/06 - BRW PART V
27/06 - U2T4 (40% grade)

Unit 03

27/06 - Deploy a scalable ML Pipeline in production
- DVC
29/06 - DVC
04/07 - FastAPI, CI/CD, Heroku
06/07 - FastAPI, CI/CD, Heroku
11/07 - Final Project
13/07 - Final Project
18/07 - Final Project
20/07 - Final Project
- U3T1 (100% grade)

Tools and Open Sources Solutions Adopted in this Course

Work in Progress
WiP

Second time the course is offered

The subject is new, we don't have all the answers

We are all learning too!!!

We appreciate you:

a) enthusiams for trying out new things

b) patience bearing with things that don't quite work

c) feedback to improve the course

Inspired by Chip Huyen CS329S - ML Systems Design

ivanovitchm / mlops Public

Pin | Unwatch 11 | Fork 7 | Star 51

Code | Issues | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

main | 1 branch | 0 tags | Go to file | Add file | Code

ivanovitchm update readme — fb0b39f 2 days ago — 98 commits

Images	start a new semester	2 days ago
.gitignore	Added week 09 example 01	2 months ago
LICENSE	Initial commit	5 months ago
README.md	update readme	2 days ago

README.md

About

Repository for DCA0305, an undergraduate course about Machine Learning Workflows and Pipelines

Readme
GPL-3.0 License
51 stars
11 watching
7 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

https://github.com/ivanovitchm/mlops