IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

Banco de Portugal data centric-strategy for the adoption of a modern data architecture[1]

Caio Costa, Guilherme de Sousa and Hugo Matos,
Banco de Portugal

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

Banco de Portugal data centric – strategy for the adoption of a Modern Data Architecture

Caio Costa, Guilherme de Sousa, Hugo Matos[1]

Abstract

Banco de Portugal has been on a path towards becoming an increasingly data-driven central bank, in order to take the most advantage of information intelligence. Along with the establishment of an integrated data management program, the challenge of creating a vision and a strategy for the adoption of a Modern Data Architecture has arisen. It is intended to enable the coexistence of corporate analytical environments with new capabilities that enhance fast, agile and flexible access to structured and unstructured data in its most granular state, in order to allow business users to address new advanced analytics or machine learning use cases. This presentation will focus on the approach followed by Banco de Portugal, in a partnership between IT, Statistics and other mission areas, to identify a set of differentiating capabilities, relevant trends and new paradigms in Big Data/Data Science domains, the proof of concepts implemented to validate the business value, as well as the roadmap established for an iterative and governed adoption of a Modern Data Architecture.

Keywords: data strategy, advanced analytics, machine learning framework

1. Introduction

Banco de Portugal is on a path towards becoming an increasingly data driven central bank where the focus is more and more on "data centricity" rather than "application contricity".

One of the key elements of the Bank's four years strategic plan 2017-2020 was the creation of an integrated data management programme, aiming at creating the Bank's data warehouse together with two pillars: the data catalog and the master data system. This programme was a major transformational initiative, jointy coordinated by the IT and the Statistics Departments, in order to strongly contribute to the a better use of the available data in the Bank by means of rationalisation of the processes associated with data collection and processing and to promote its effective sharing throughout the whole organisation (Moreno, 2021).

While running this programme, new challenges emerged, such as: i) business areas were shifting more and more towards the collection of microdata (securities, loans, payments, ...), with higher frequency and more granularity; ii) projects involving big data were becoming more frequent; iii) new data science approaches were necessary to analyze and explore the new data landscape; and iv) security and privacy becoming a major concern.

It became then evident for the IT Department that we needed a specific IT programme to address all these challenges – the creation of our modern data architecture iniciative. It kicked-off late in 2019

[1] Caio Costa (ccfcosta@bportugal.pt), Guilherme de Sousa (garanha@bportugal.pt) and Hugo Azevedo Matos (hmatos@bportugal.pt), Information Systems and Technology Department, Banco de Portugal. The views expressed are those of the authors and not those of the Banco de Portugal. We thank to Sara Cândido (Banco de Portugal) and Filipa Lima (Banco de Portugal) for their valuable comments.

and since then we have been developing our data strategy and our artificial intelligence strategy as part of it.

The modern data architecture governance model foresees a straight collaboration between the business units and the IT department. Within the IT department, we have 3 units that are more directly involved in the programme: i) the BI and Information Management Unit; ii) the Systems and Applications Engineering Unit; and iii) the INOV# Inovation Lab.

The paper is organised as follows: after the introduction section, we describe the different dimensions of our modern data architecture in section 2. In section 3 we present the various use cases already in place. We conclude with section 4 with future developments and final remarks.

2. Modern Data Architecture

In our Modern Data Architecture strategy, we identified three clusters of analytical capabilities:

1) Corporate BI – analytical capabilities typically centrally guaranteed by IT, with the aim of providing certified information organized by domains.
2) Self-Service BI - analytical capabilities that guarantee greater autonomy to users to analyze and interact with analytical information, allowing the creation of interactive and dynamic outputs.
3) Advanced Analytics - analytical capabilities typically related to the automatic identification of patterns using self-learning algorithms and exploited autonomously by advanced users.

A key feature of our programme is to go step by step and to gradually promote the business autonomy and flexibility to interact with data, in environments and platforms governed by the IT department. Given that corporate BI and the self-service BI tools were already relatively well covered for the end user needs (through the application portfolio and powerBI dashboards), we focused on advanced analytical capabilities that we were missing and were demanded by the most savvy end users to boost their productivity, which will be the focus of this paper.

In the advanced analytics space, there are lots of new concepts that originates mislead interpretations so the first step was to build a common foundation of concepts.

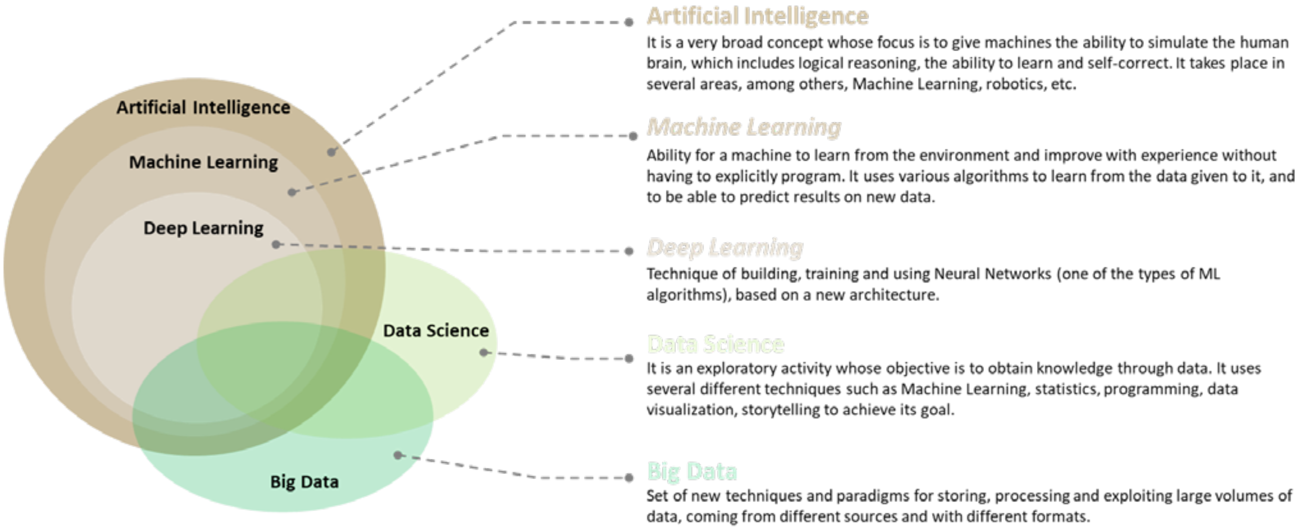


Figure 1 - Advanced Analytics concepts foundations.

Focusing on the transition to a modern data architecture and taking into account our business use cases, we started by defining two parallel paths:

1) Path "End users empowerment" - focused on providing new analytical environments with differentiating capabilities to end users take advantage of modern data exploration techniques.
2) Path 2 "IT Departament capacitation" – focused on empowering the IT Department with the skills, methodologies, frameworks and tools needed to deliver turnkey data science projects.

2.1 End users empowerment - New analytical environments

Based on a straight collaboration between the business units and the IT department, two main needs were identified:

- The first one, aimed to experienced data analysts, was related with having possibility to use SQL language to anaylize data in an environment with integrated access to corporate data and with write/persist permissions;
- The second one, aimed to data scientists, was related with providing environments suitable for the exploration of advanced techniques such as machine learning algorithms.

To answer to this needs, Banco de Portugal provided new analytical enviromnents, that enable autonomous exploration by end users, guided by governance and good practices.

2.1.1 SQL Data Labs

SQL data labs to allow users to interact with structured data available on corporate data stores using an SQL interface and persist the analysis results.

User can load adhoc data to join with corporate data and build analysis.

This data labs ensures that the execution of analysis doesn't interfere with scheduled batch corporate processes.

To support this data labs, we leveraged Microsoft SQL servers that we already had in-place and that were most of time idle, due to be used only for secondary replicas.

2.1.2 Data Science Labs

The Data Science Labs - Pitágoras are a new analytical environment for advanced data exploration with modern data science tools, runned on top of a platform with scalable and remote processing capacity, dedicated storage, code versioning (with Github) and a governance model.

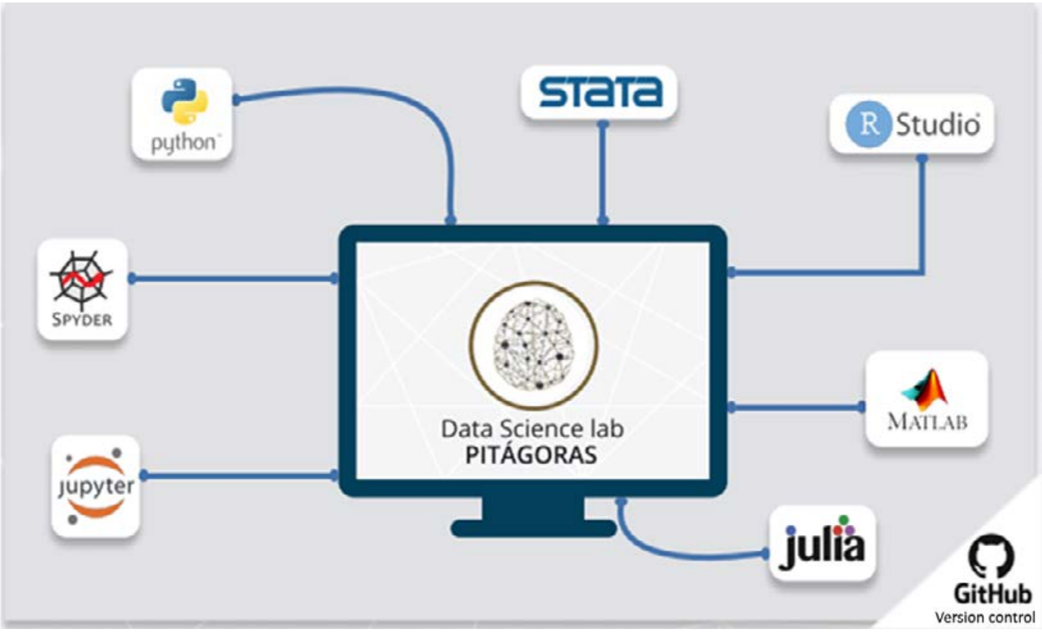


Figure 2 - Data Science Labs flavours.

Tools available for end users work with are Rstudio, Python, Spyder, Julia, Jupiter notebooks, Stata and Mathlab.

In terms of technology, Pitágoras is an in grid computing with containerized and customizable runtimes with singularity, which is a container runtime very similar to docker but designed for this type of use case, and based in CEPH, a Software defined storage that delivers both file and object storage with AWS S3 compatible interface capable of scalling to hundreds of petabytes.

With Pitágoras users are able to execute both batch and interactive jobs - such as submitting a background program, or running their development environment.

Regarding its current size we have 14 servers, totalling 7TB of RAM; 448 CPU Cores; 6 GPUs and 100TBs of storage. It can be scalled to as many servers as needed with ease since all the infrastructure is deployd automatically with Ansible.

We have also deployd a GitHub Enterprise infrastructure for our users source control.

2.2 IT Department capacitation

In order for the IT Department to be able to deliver turnkey data science projects, it was necessary to define project management methodologies suited to the specificities of this type of project, as well as define development and production cycles adapted to ML model, who promote automation, standardization and quality, leveraging good practices.

2.2.1 CRISP-DM

CRISP-DM

To manage our data science projects, we decided to base on CRISP-DM methodology.

The CRISP-DM methodology (Cross Industry Standard Process for Data Mining) is an open standard process model that describes the data science life cycle. It aims to standardize and describe the phases in a data science project, in order to help plan, organize and implement, us-ing common approaches and vocabulary.

It consists of 6 phases, namely:

1. Business Understanding - initial phase that focus on understanding the value-add from a business perspective and translate it to a data science problem definition. Includes definition of business objectives and what success means, and creation of the preliminary plan to achieve those success criteria.

2. Data Understanding - focus on identify, collect and explore initial datasets to know what can be expected and achieved from the data.

3. Data Preparation - prepare the data sets to be used by algorithms, which includes select, clean and create data using data preparation processes.

4. Modelling - phase responsible to build the model that answers to the business needs, as-sessing distinct model techniques against each other.

5. Evaluation - verify if the results meet the success criteria and define next steps. Outcome could be proceed to deploy, iterate again on previous phases to understand why the objec-tives were not achieved and try different approaches, or even to close the project. Also in-cludes reviewing and documenting the processes.

6. Deployment - last phase who is responsible to plan and implement the deployment as well as the monitoring and maintenance of the model after going to production. Depending on the business objectives, the deployment can be integrate the model with an existing software or build a dashboard/output to support decision-making.

The sequence of the phases is not rigid and moving back and forth between phases is common.

This methodology was created many years ago, and despite there are some new project man-agement frameworks used for delivering data science projects (generic project management frameworks like Scrum, Kanban, and also more data science oriented like TDSP, SEMMA), it stills maintains to be one of the most used ones.

2.2.2 Machine learning framework

In order to enforce automation, the use of standards and good pratices, we build a custom machine learning framework, composed by technology and processes.

This framework define development and production cycles adapted to ML model, leveraging good practices in software development.

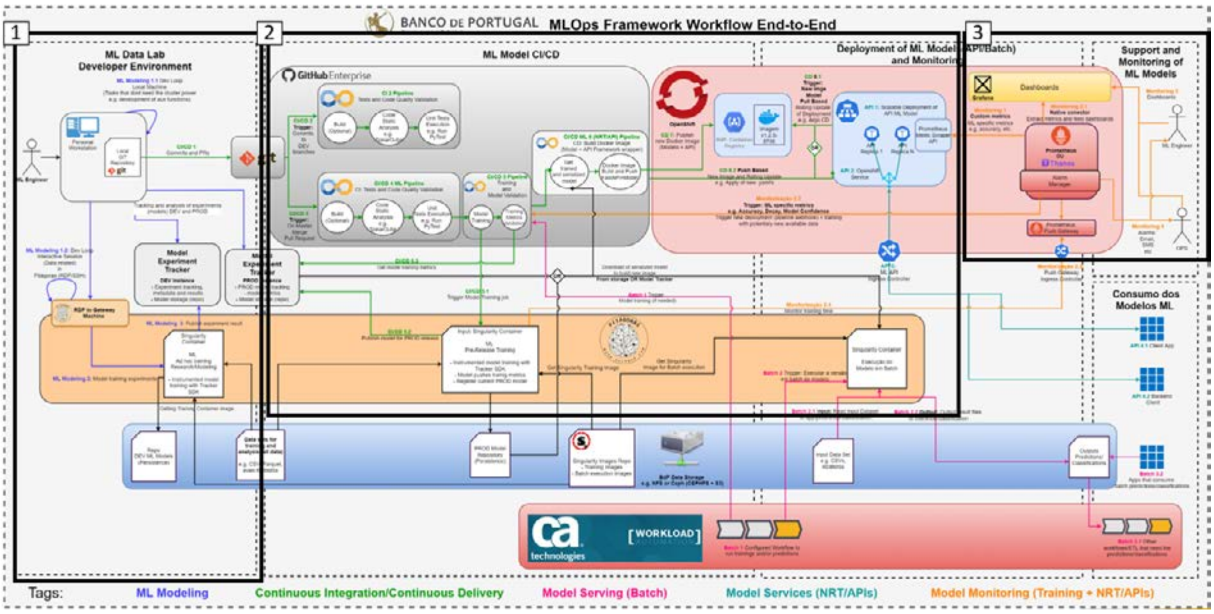


Figure 3 - BdP Machine Learning Framework

The framework is composed of 3 main areas:

1) Development: model development

2) CI/CD Training Pipeline: code quality validation, automatic tests, (re) training, deploying new versions (batch / api)

3) Monitoring & Alerting: collection and monitoring of model metrics, trigger actions based on rules (ex: retrain)

In terms of the development area, our goals where to provide some guidelines regarding good code practices, project structure and model development standardization. To achieve this, we built and example project in which we took all of these aspects into account to demonstrate what has been documented. So far, our top requirements are that models are developed with the kedro framework which brings at almost no cost a lot of good practices, and for the rest APIs to be developed with fast api.

Regarding the CI/CD pipeline, our objective was to have everything automated end to end with the good DevOps practices, plus everything that deploying ML requires: in other words, MLOps. The idea is that this pipeline is generic enough so that everytime a new ML project arrives, it can be reused with just very few changes – we are using github actions, so everything is "infrastructure as code". In the end, after the tests have passed, the model has been trained in Pitágoras and built into a docker image, it is deployed in our Openshift Cluster.

Last but not least, monitoring and alerting. One of the guidelines of the ML Framework is that the models need to expose a metrics endpoint so that we can monitor it with a Prometheus, Grafana and AlertManager stack. This allows us to monitor the model's accuracy so that we can be allerted when it goes bellow the defined threshold and even trigger an automatic training.

3. Use cases

In this section we describe the use cases we have worked so far with various business units.

3.1 SQL Data labs

We start with quality control for individual loans. Experts from the Statistics Department wanted to find outliers in loans information (from our Central Credit Register – CCR). The information has an analytical model defined which is available to the end users. For the first phase of the use case, it was used a SQL data lab, where the experts from Statistics autonomously did the construction / prototyping /and creation of a MVP on Quality Control Processes and at a later satge, after certification, these rules were incorporated into the CCR system (in production environment by the IT Department).

In the second phase, the business wanted an algorithm to detect outliers, so they selected the isolation forest algorithm and created the process in python – using our Data Science Lab - Pitágoras platform. In this case whenever they need they run manually the process, because as we mention we are still defining the best way to incorporate ML code done by end users into corporate systems. (just for a glimpse the Isolation forest (deal with 17 million of new records per month | and in less than 1 hour they are able to train the algorithm).

Another use case worth mentioning is ad hoc analysis for illicit financial activity. Ingestion and analysis of data collected during inspection actions, is done in a completely autonomous way by business users, using skills that are familiar to them (SQL), in an governed environment that has high computing power and that enhances the sharing of information between stakeholders, a SQL data lab.

3.2 Data Science labs

Regading the machine learning use cases using the framework established and implemented by IT department in project mode we have the following 4 use cases, which were first experimented in our innovation lab and once approved for adoption followed our pipeline for IT project implementation.

1) Automatic validation of credit contracts

The goal is to validate automatically compliance with regulatory standards in the Credit Agreement Drafts, moving from a sampling approach (number of drafts contracts and number of validation rules) to a universal approach (potentially all drafts contracts and all rules). The tool automatically identifies, even if partially, the clauses that do not comply with legal and regulatory requirements, improving the performance of the Conduct Supervision department.

2) Supervisory Board written procedures

The goal is to accelerate the analysis and reply to Supervisory Board written procedures for which a RPA and rules engine were developed. The tool extracts and structures data from a sample of previous written procedure since 2014 and provides NPL capabilities to assist document content and analysis.

3) Automatic classification and response to information requests by banking clients

The goal is to automatically classify information requests and propose the corresponding automatic replies, depending on the classification and content, evaluating thus the feasibility of an automatic answer.

4) Non structured data analysis in acquisition processes

The goal is to have automatic analysis of the acquisition processes, data extraction from acquisition documents (ex.: contract, contract decision, public procurement portal) and audit controls' validation (ex.: mandatory clauses in contracts, amount consistency between documents).

4. Final remarks

For the current 2021-2025 strategic plan we will continue working across the data value chain to fully benefit from new capabilities (data lake and distributed processing engine - Spark) and to promote the use of good practices, such as data lake organization, workload segregation, data ingestion patterns, etc. We will work with the Payments Department towards a new data collection paradigm of individual payments data and with the Statistics Department in order to define a data quality control framework where a set of common consistency and validation rules can be applied irrespectively of the data domain.

We conclude with the key success factors throughout our data journey:

1) Data architecture

Defining the path from traditional to a modern analytics architecture, with business use cases.

Choosing the right technology for the most use cases.

Significant investments in IT infrastructures and software.

2) Data security

Defining very clear rules and principles in terms of information security policy.

Increased use of micro data raises the bar in terms of security issues.

Particularly relevant when considering cloud environments.

3) Data governance

Setting an enterprise-wide roles and responsibilities.

All departments should be involved in the decisions.

Cultural / Organizational change.

Adequate expectations to management is vital.

4) Data skills

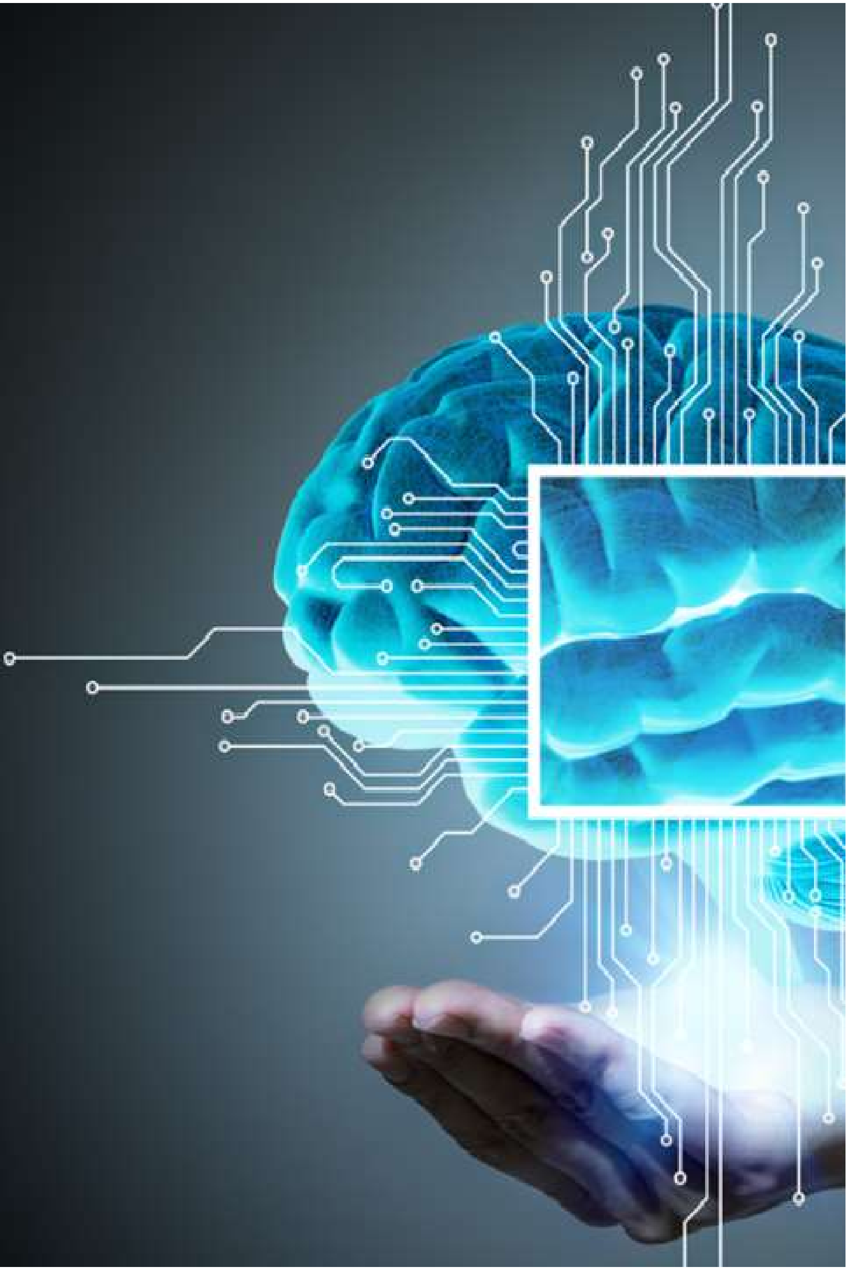
New roles are needed (e.g. big data / machine learning engineers, data scientists).

Reinforcing and adapting the skills of employees to the new challenges.

Establishing a training programme is essential.

References

Moreno, M C (2021): "Data Governance: an orchestra of people, processes and technology", IFC Bulletin No 54, Issues in Data Governance.



Banco de Portugal Data Centric

Strategy for the adoption of a Modern Data Architecture

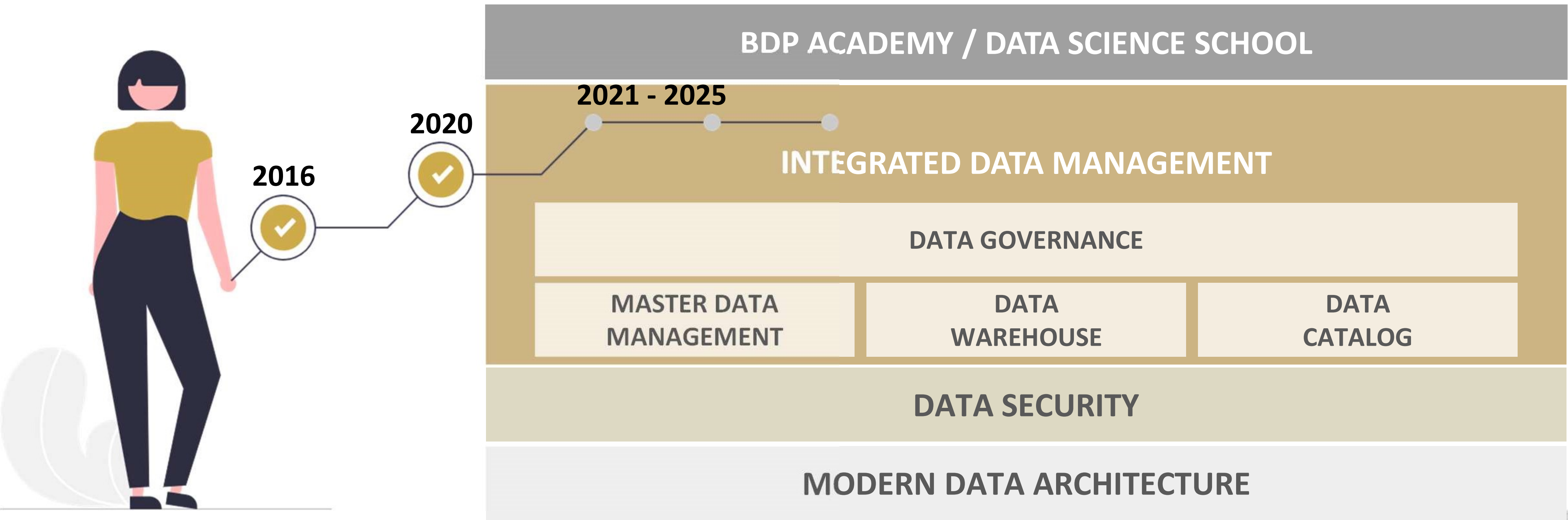
Caio Costa, Senior Data Engineer
(ccfcosta@bportugal.pt)

Guilherme de Sousa, Systems and Applications Engineer
(garanha@bportugal.pt)

BANCO DE PORTUGAL
EUROSISTEMA

February, 2022

Banco de Portugal on a path towards becoming an increasingly **data centric central bank**



In our **Data Strategy**, there are three main complementary **clusters of analytical capabilities**

Corporate BI

Analytical capabilities typically centrally guaranteed by IT, with the aim of providing certified information organized by domains.

Dashboards Reports Portal

Target audience:

Data User Data Analyst

Self-Service BI

Analytical capabilities that guarantee greater autonomy to users to analyze and interact with analytical information, allowing the creation of interactive and dynamic outputs.

Self-Service BI Dashboards

Target audience:

Data Analyst

Descriptive analytics

Advanced Analytics

Analytical capabilities typically related to the automatic identification of patterns using self-learning algorithms and exploited autonomously by advanced users.

Data Labs Machine Learning Data Lake

Target audience:

Data Scientist

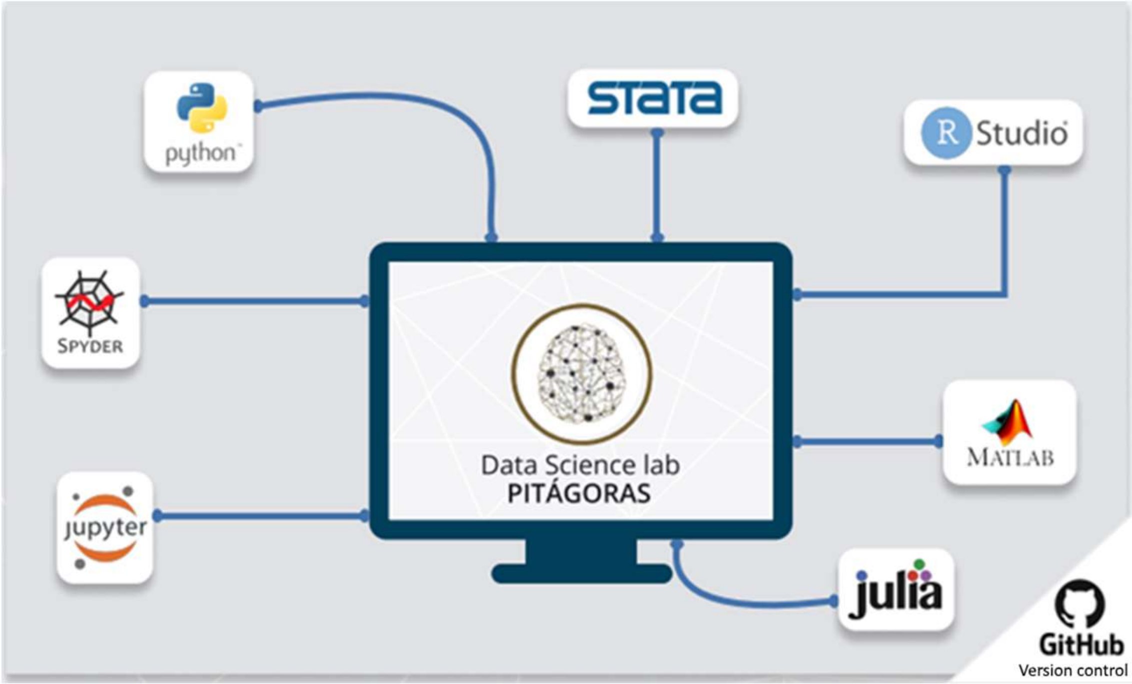
Predictive analytics

In the **Advanced Analytics cluster,** with a target audience of advanced users, we provided **new analytical environments**

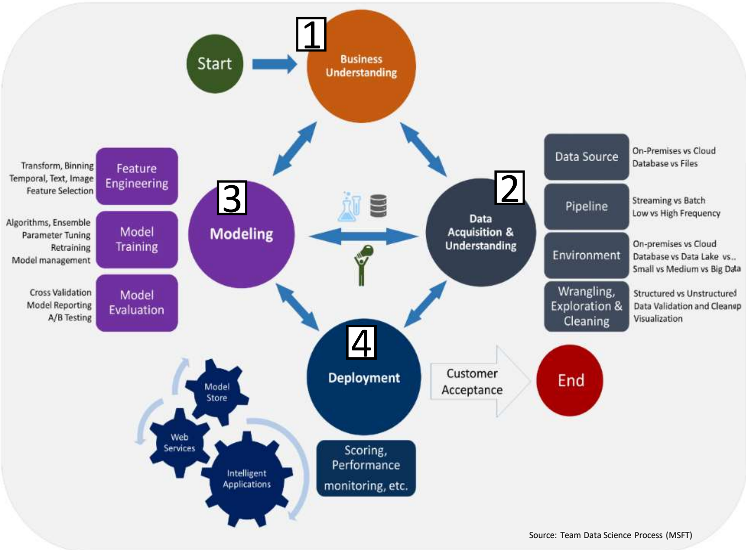
SQL DATA LAB	Interact with structured data using an **SQL interface** and persist the analysis results, with **high computing capacity** and without interfering with the execution of the scheduled batch corporate processes. Users have **write permissions on the environment**.
DATA SCIENCE LAB (PITÁGORAS)	Autonomy to use new data exploration techniques such as Machine Learning, Natural Language Processing through multiple tools / languages. With **scalable and remote processing** capacity, **dedicated storage** and **code versioning**.

The **Data Science Lab** is a custom built open-source platform aimed for **data scientists**

- Grid computing solution
- Containerized and customizable runtimes and IDE's (singularity containers)
- Dedicated storage
- Batch executions
- Interactive executions (IDE's, Notebooks, etc.)
- + GitHub Enterprise Server



The IT Department provides **turnkey data science projects** using standard methodology and MLOps pipelines



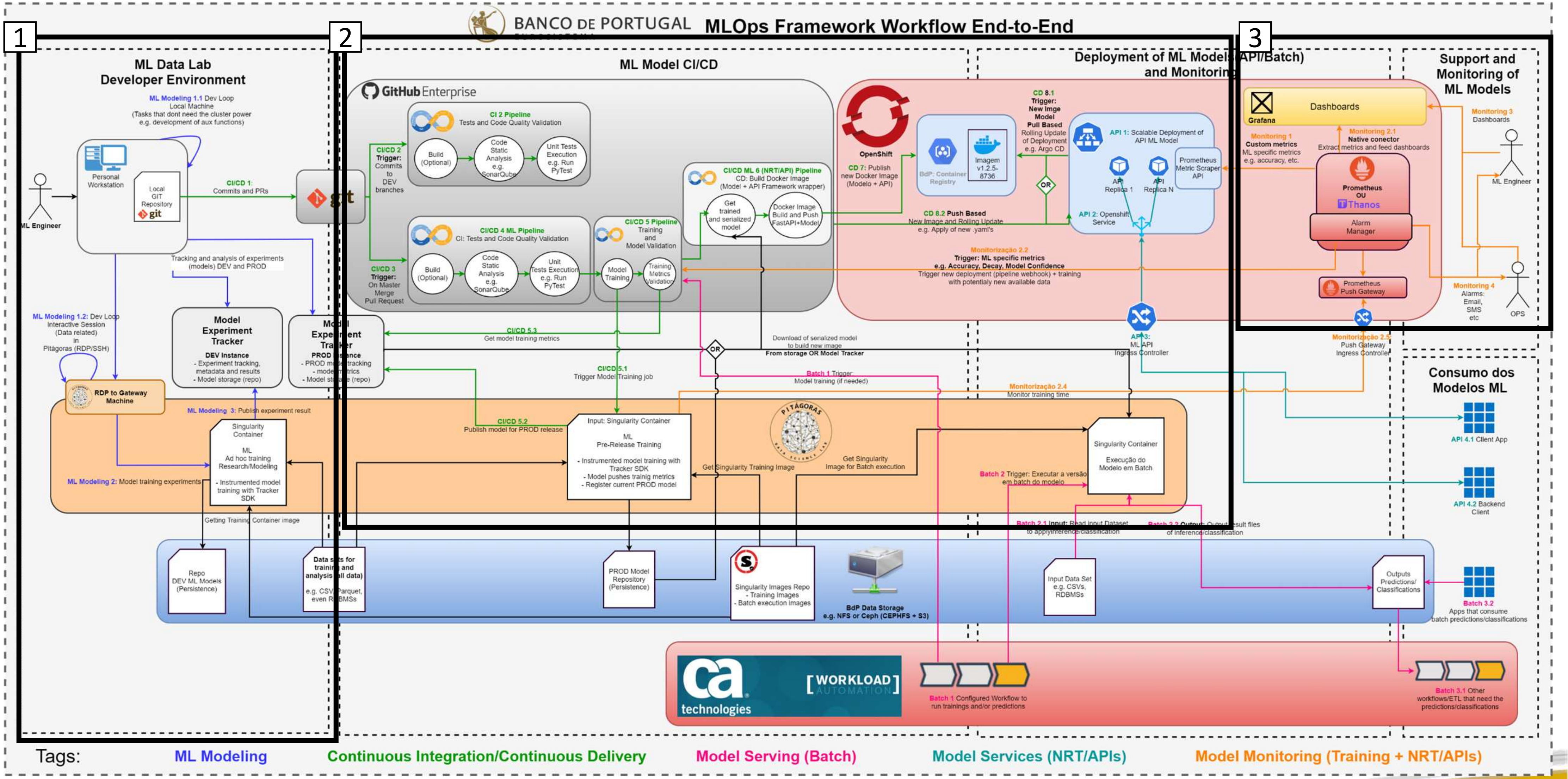
Skills:
(not exhaustive)

- Subject-Matter Expert
- Data Scientist
- Machine Learning Engineer

- Data Science processes typically run through this life cycle based on **methodology (CRISP-DM)**, with well-defined steps that follow a logical sequence (identified by numbers)
- For each stage, **objectives** are defined, **techniques** to achieve them and **outputs**.
- In order to ensure that this process is **efficient**, the practice of **MLOps** is in place to **increase automation and improve the quality of the ML model** development and production cycle.

MACHINE LEARNING FRAMEWORK

The objective of this framework is to **increase automation**, promote **standardization** and **improve** the **quality** of the development and production cycle of ML models, leveraging **good practices** in software development.



BANCO DE PORTUGAL
MLOps Framework Workflow End-to-End
1
2
3
ML Data Lab
Developer Environment
ML Model CI/CD
Deployment of ML Models (API/Batch) and Monitoring
Support and Monitoring of ML Models
Consumo dos Modelos ML
ML Engineer
OPS
Personal Workstation
Local GIT Repository
CI/CD 1: Commits and PRs
GitHub Enterprise
CI 2 Pipeline Tests and Code Quality Validation
CI/CD 4 ML Pipeline CI: Tests and Code Quality Validation
CI/CD 5 Pipeline Training and Model Validation
CI/CD ML 6 (NRT/API) Pipeline CD: Build Docker Image (Model + API Framework wrapper)
Model Experiment Tracker
RDP to Gateway Machine
OpenShift
Prometheus
Alarm Manager
Grafana
Dashboards
Prometheus Push Gateway
BdP Data Storage e.g. NFS or Ceph (CEPHFS + S3)
ca technologies
[WORKLOAD] AUTOMATION
Tags:
ML Modeling
Continuous Integration/Continuous Delivery
Model Serving (Batch)
Model Services (NRT/APIs)
Model Monitoring (Training + NRT/APIs)

Banco de Portugal has deployed use cases that take advantage of the new analytical capabilities

Self-service

Illicit financial activity **(adhoc analysis)**	• Ingestion and analysis of data collected during **inspection actions**, in a completely autonomous way by business users, using skills that are familiar to them (SQL), in an governed environment that has high computing power and that enhances the sharing of information between stakeholders - **SQL Data Lab**.
Central Credit Register **(quality control processes)**	• Implementation of quality control checks autonomously by business users **using both traditional (SQL) and modern (machine learning techniques) approaches** – expert users with skills is SQL & Python. • After certified by IT, the quality control checks were incorporated into corporate system processes. • **SQL Data Lab** – used for apply quality control rules using SQL queries and to prepare data to apply ML techniques; • **Data Science Lab** – creation and execution of the ML (isolation forest) algorithm for outliers identification.

Turnkey projects

Information Requests Classification and Response (Conduct)	**Automatic classification of requests for information and generation of response text** inov#) • Evaluate the feasibility of automatically classifying requests for information sent to Banking Conduct Supervision. • Depending on the classification and content, evaluate the feasibility of proposing an answer automatically.
Credit Agreement Draft Validation	**Validate automatically compliance with regulatory standards in the Credit Agreement Drafts** • In a step-by-step approach from a sampling approach (number of drafts contracts and number of validation rules) to a universal approach (potentially all drafts contracts and all rules). • Automatically identify, even if partial, clauses that do not comply with legal and regulatory requirements, improving the performance of the Conduct Supervision department.

For a success implementation of a Data Strategy, these were the **Key Success Factors**

DATA ARCHITECTURE

Defining the path from traditional to a modern analytics architecture, with business use cases.

Choosing the right technology for the most use cases.

Significant investments in IT infrastructures and software.

DATA SECURITY

Defining very clear rules and principles in terms of information security policy.

Increased use of micro data raises the bar in terms of security issues.

Particularly relevant when considering cloud environments.

DATA GOVERNANCE

Setting an enterprise-wide roles and responsibilities.

All departments should be involved in the decisions.

Cultural / Organizational change.

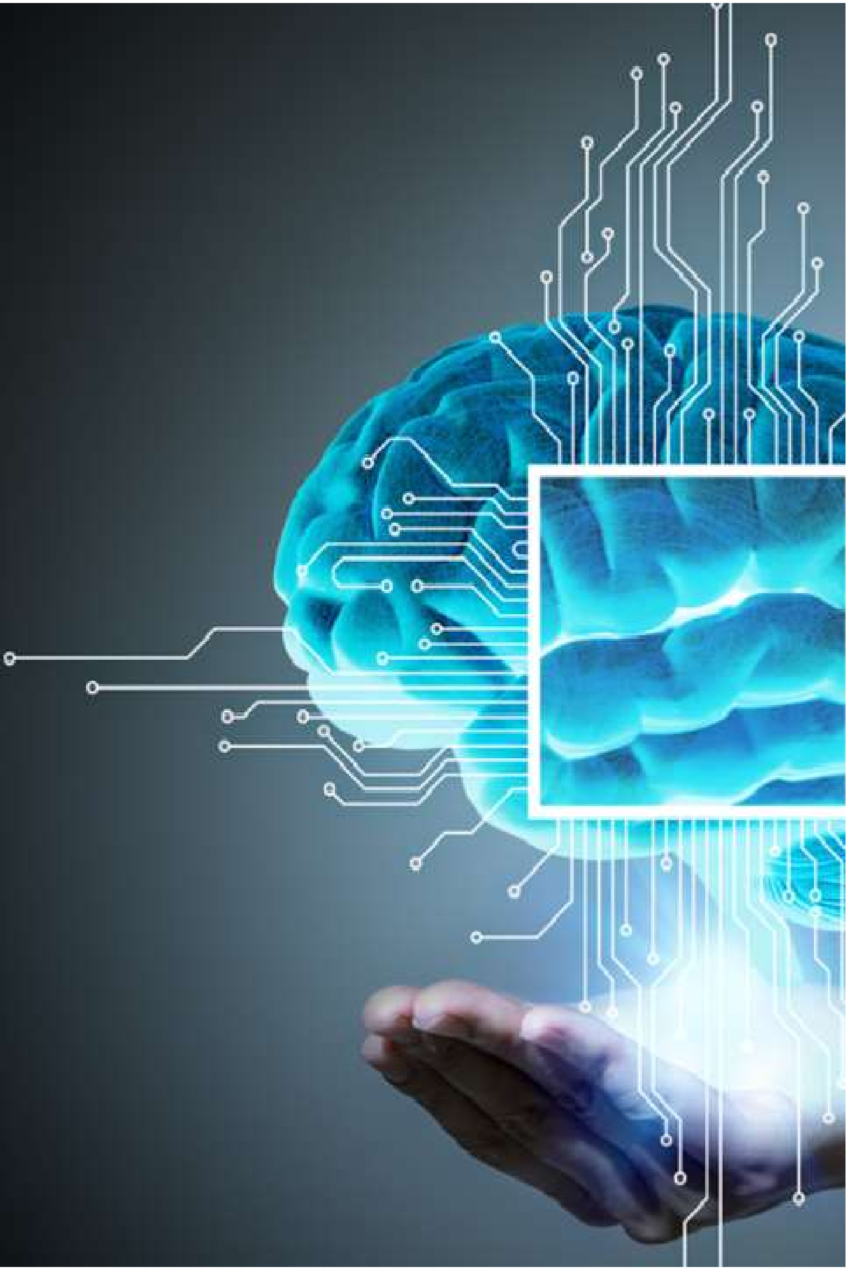
Adequate expectations to management is vital.

DATA SKILLS

New roles are needed (e.g. big data / machine learning engineers, data scientists).

Reinforcing and adapting the skills of employees to the new challenges.

Establishing a training program is essential.



Banco de Portugal Data Centric

Strategy for the adoption of a Modern Data Architecture

Caio Costa, Senior Data Engineer
(ccfcosta@bportugal.pt)

Guilherme de Sousa, Systems and Applications Engineer
(garanha@bportugal.pt)

BANCO DE PORTUGAL
EUROSISTEMA

February, 2022