From the Ground Up:
Building a Data Science Team

Rui Filipe Pedro - 14/06/2022

Content

01

Getting into Data Science

Getting into Data Science

There is "more than one way to skin a cat"

Your own path depends on your objectives

There are many paths:

- Coding bootcamp + ML/DS online 101 course -> Data Scientist
- Computer Science w/ DS classes -> Data Scientist
- Non Computer Science degree + DS bootcamp -> Data Scientist
- And so on...

More knowledge == more options

02

Building the Team

Building the Team

Methodology

Roles

Work Plan

Infrastructure

Methodology

Methodology

Objectives:

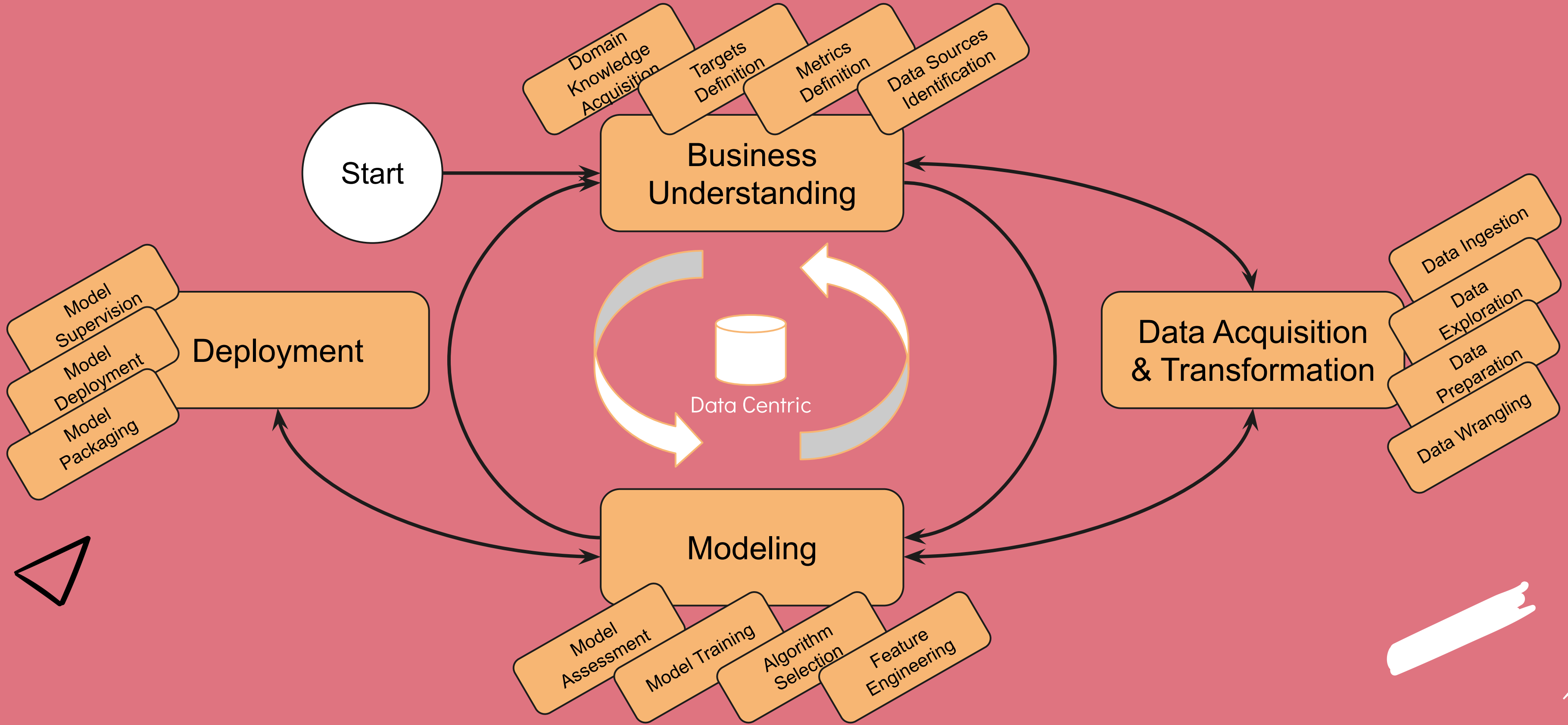
- Define a set of principles and processes for the implementation of Data Science projects
- Address technical and non-technical aspects of Data Science
- Set a standard for projects
- Provide recommendations for the Data Science tool-kit

Methodology

Several references:

- **KDD**: Knowledge Discovery in Databases
  - https://www.geeksforgeeks.org/kdd-process-in-data-mining/
  - https://en.wikipedia.org/wiki/Data_mining#Process
- **CRISP-DM**: Cross Industry Standard Process for Data Mining
  - https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining
- **TDSP**: Team Data Science Process
  - https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

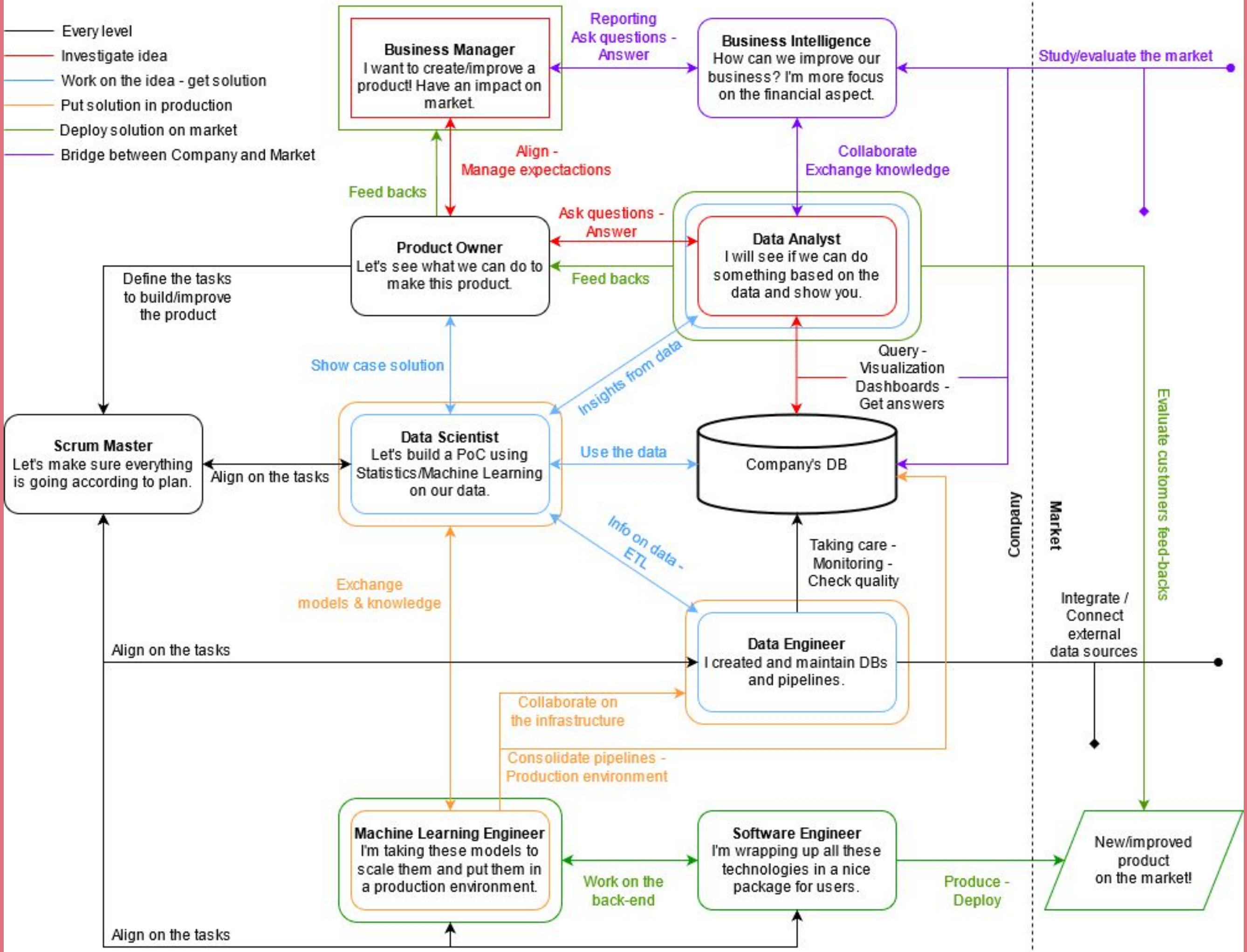
Methodology - Deriving our own



Roles

Roles

- Solution architect
- Project manager
- Data engineer
- Data scientist
- Application developer
- Project lead
- Machine Learning Engineer
- DevOps Engineer
- Data Mastermind
- MLOps Engineer
- Data Architect
- Data Analyst
- Business Analyst
- Data Science Manager
- Cognitive Champion
- Database Admin
- Statistician
- Data Journalist / Storyteller
- Software Engineer



Source: https://www.kdnuggets.com/2021/12/build-solid-data-team.html

Roles in practice

Product Team

- Software Dev
- Data Scientist
- Data Engineer
- Solution Architect
- QA
- Manager

Research Team

- Data Scientist
- Data Engineer
- ML Engineer
- ML Research Scientist

Data Science Only Team

- Project Lead
- Data Scientist
- Data Engineer
- Data Journalist
- MLOps Engineer

Work Plan

Work Plan

1. Planning a Schedule
2. Design Thinking

Work Plan – Planning

Design Thinking Workshop

Stream 1: Exploratory Data Analysis (EDA)

Stream 2: Datasets Definitions & Preparations

Stream 3: Models Discovery, Selection and Training

Stream 4: AI Business Story

Sprint 1: Explore

Sprint 2: Build

Sprint 3: Run & Deliver

Week 1

Week 3

Week 5

Work Plan - Design Thinking

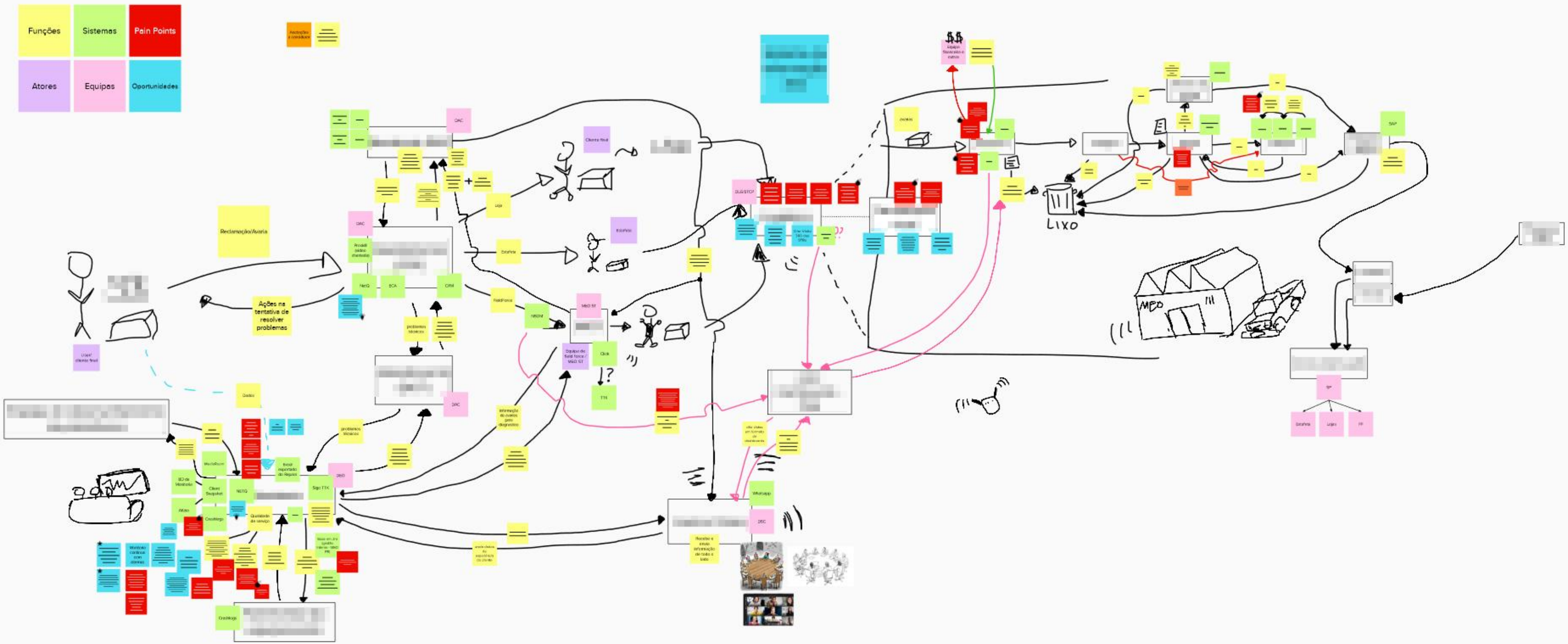
Way to kickstart and define all necessary information about Business Understanding

Adapted from existing Product Design Thinking to Data projects

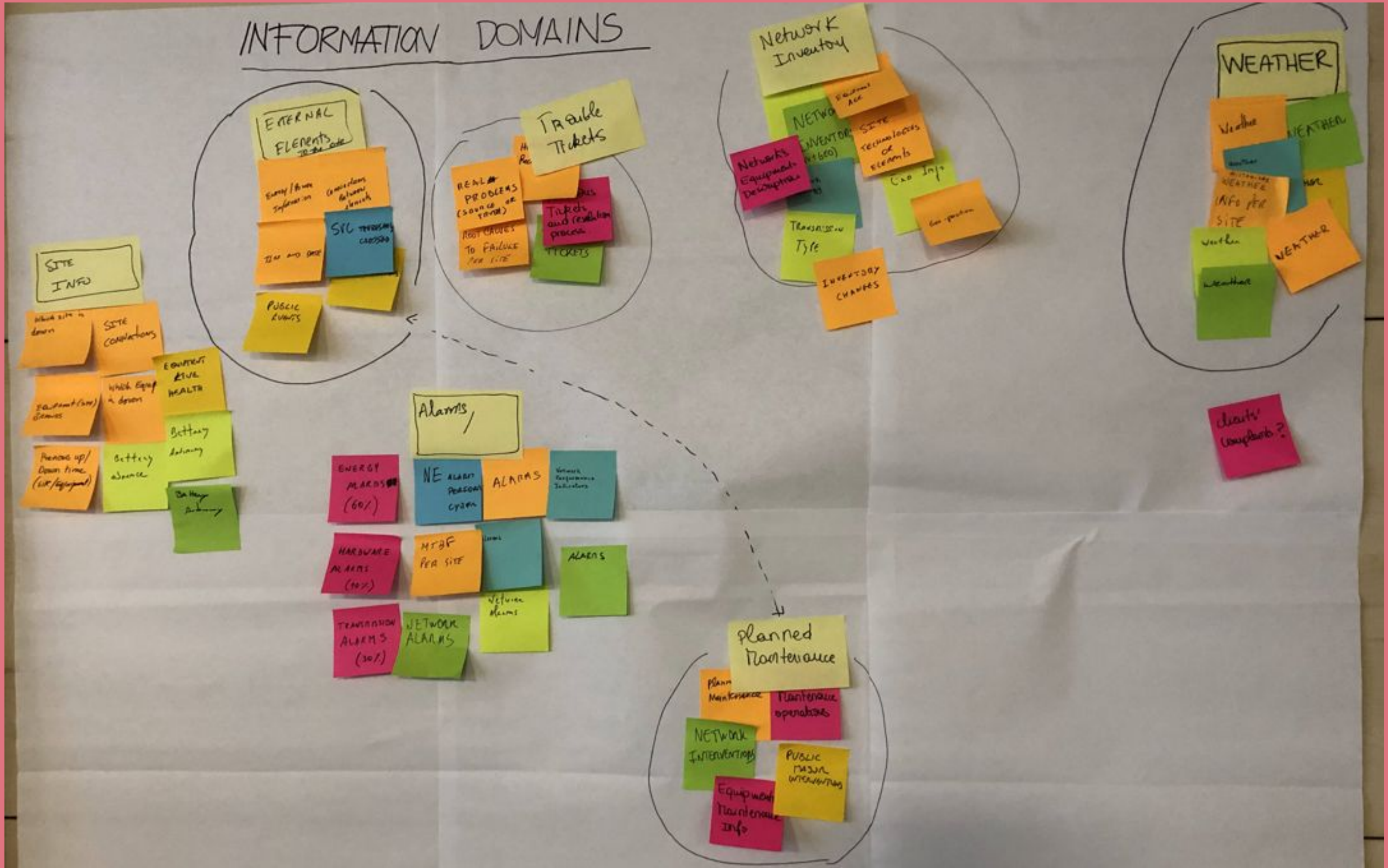
Define:

- The problem to solve
- Stakeholders
- Data sources (discovery and evaluation)
- Vision and problem statements
- Success criteria: business and technical metrics
- Execution: planning, timeline, level of involvement and communication

Work Plan - Design Thinking

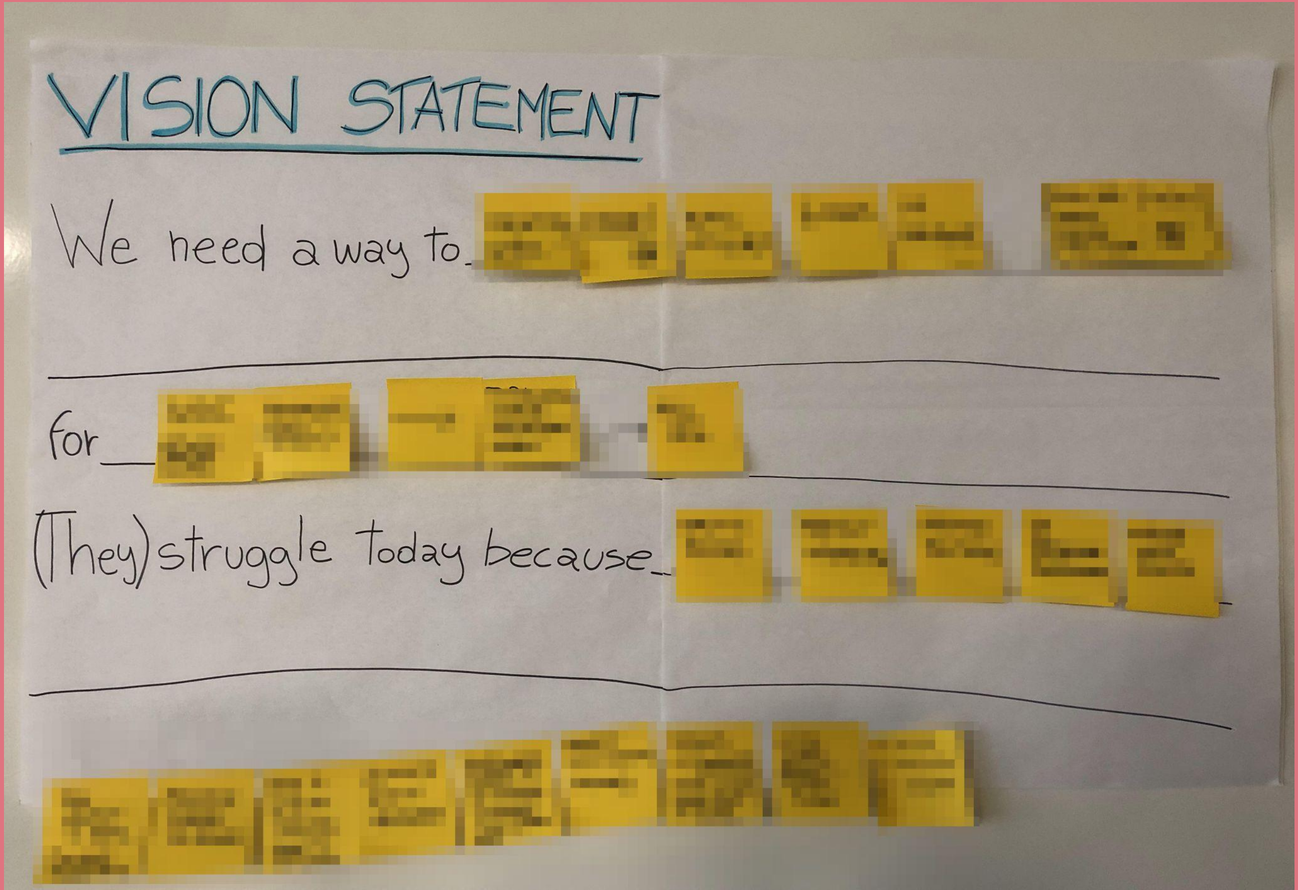


Work Plan - Design Thinking

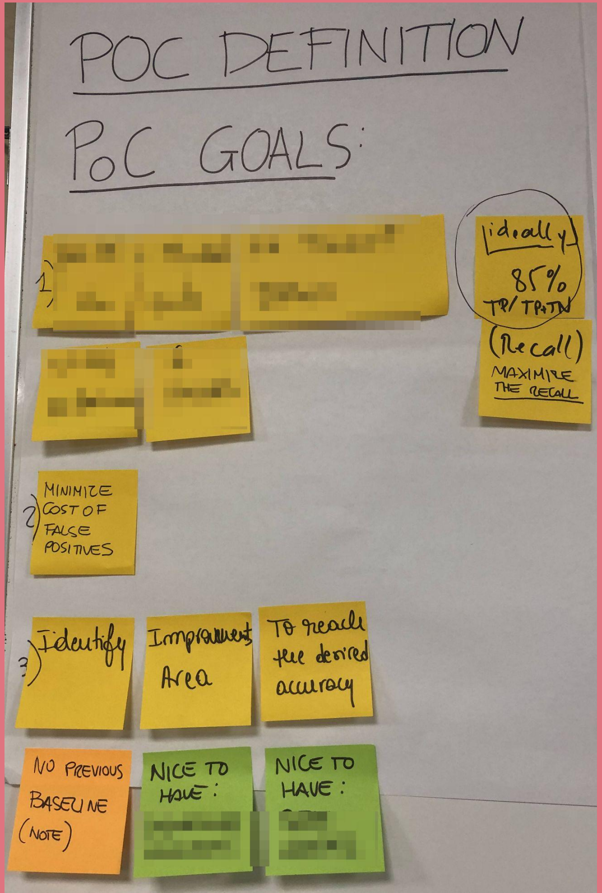


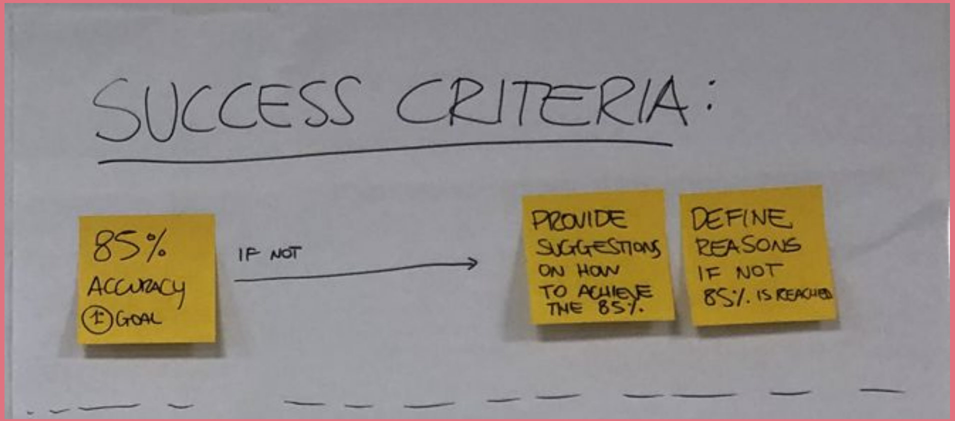
Work Plan - Design Thinking

Work Plan – Design Thinking



Work Plan – Design Thinking





Infrastructure

Infrastructure

Creating the tool-kit for data science

Any combination of closed-source, open-source, cloud, on-prem technology

Objectives:

- Create a standard way of working
- Ensure all relevant artifacts are versioned
- Streamline the lifecycle processes
- Align with the industry - embrace MLOps, containers, etc

Infrastructure

1 Project Structure

2 Versioning

3 Streamlining

4 Deploying

Infrastructure - Project Structure

Few options:

- Dzone article: https://dzone.com/articles/data-science-project-folder-structure
- TDSP template: https://github.com/Azure/Azure-TDSP-ProjectTemplate
- Cookiecutter-DS: https://drivendata.github.io/cookiecutter-data-science/

Infrastructure - Project Structure

Option #3: CookieCutter for Data Science

```
├── LICENSE
├── Makefile           <- Makefile with commands like `make data` or `make train`
├── README.md          <- The top-level README for developers using this project.
├── data
│   ├── external       <- Data from third party sources.
│   ├── interim        <- Intermediate data that has been transformed.
│   ├── processed      <- The final, canonical data sets for modeling.
│   └── raw            <- The original, immutable data dump.
│
├── docs               <- A default Sphinx project; see sphinx-doc.org for details
│
├── models             <- Trained and serialized models, model predictions, or model summaries
│
├── notebooks          <- Jupyter notebooks. Naming convention is a number (for ordering),
│                         the creator's initials, and a short `-` delimited description, e.g.
│                         `1.0-jqp-initial-data-exploration`.
│
├── references         <- Data dictionaries, manuals, and all other explanatory materials.
│
├── reports            <- Generated analysis as HTML, PDF, LaTeX, etc.
│   └── figures        <- Generated graphics and figures to be used in reporting
│
```

```
├── requirements.txt   <- The requirements file for reproducing the analysis environment, e.g.
│                         generated with `pip freeze > requirements.txt`
│
├── setup.py           <- Make this project pip installable with `pip install -e`
├── src                <- Source code for use in this project.
│   ├── __init__.py    <- Makes src a Python module
│   │
│   ├── data           <- Scripts to download or generate data
│   │   └── make_dataset.py
│   │
│   ├── features       <- Scripts to turn raw data into features for modeling
│   │   └── build_features.py
│   │
│   ├── models         <- Scripts to train models and then use trained models to make
│   │   │                 predictions
│   │   ├── predict_model.py
│   │   └── train_model.py
│   │
│   └── visualization  <- Scripts to create exploratory and results oriented visualizations
│       └── visualize.py
│
└── tox.ini            <- tox file with settings for running tox; see tox.readthedocs.io
```

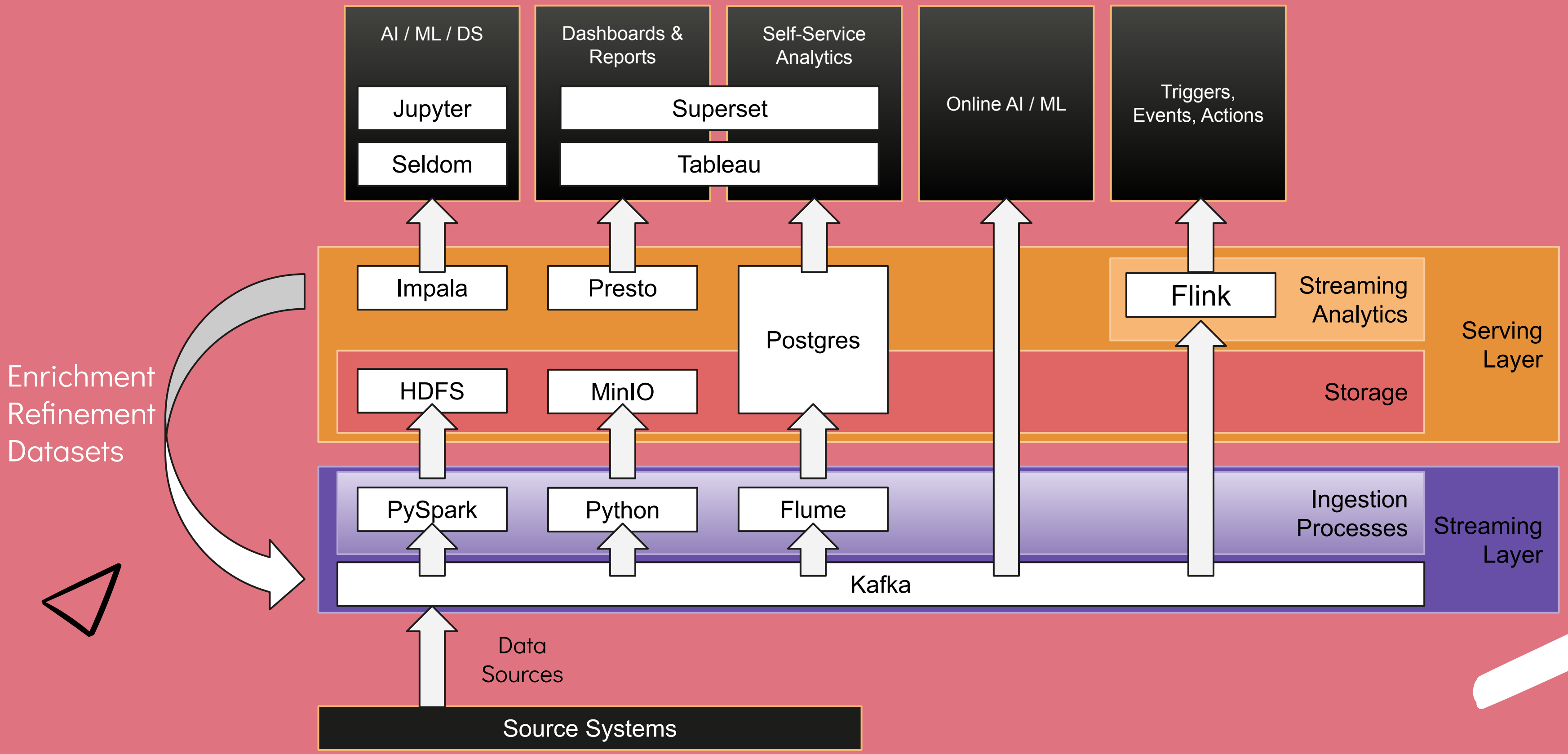
Source: https://drivendata.github.io/cookiecutter-data-science/

Infrastructure - Versioning

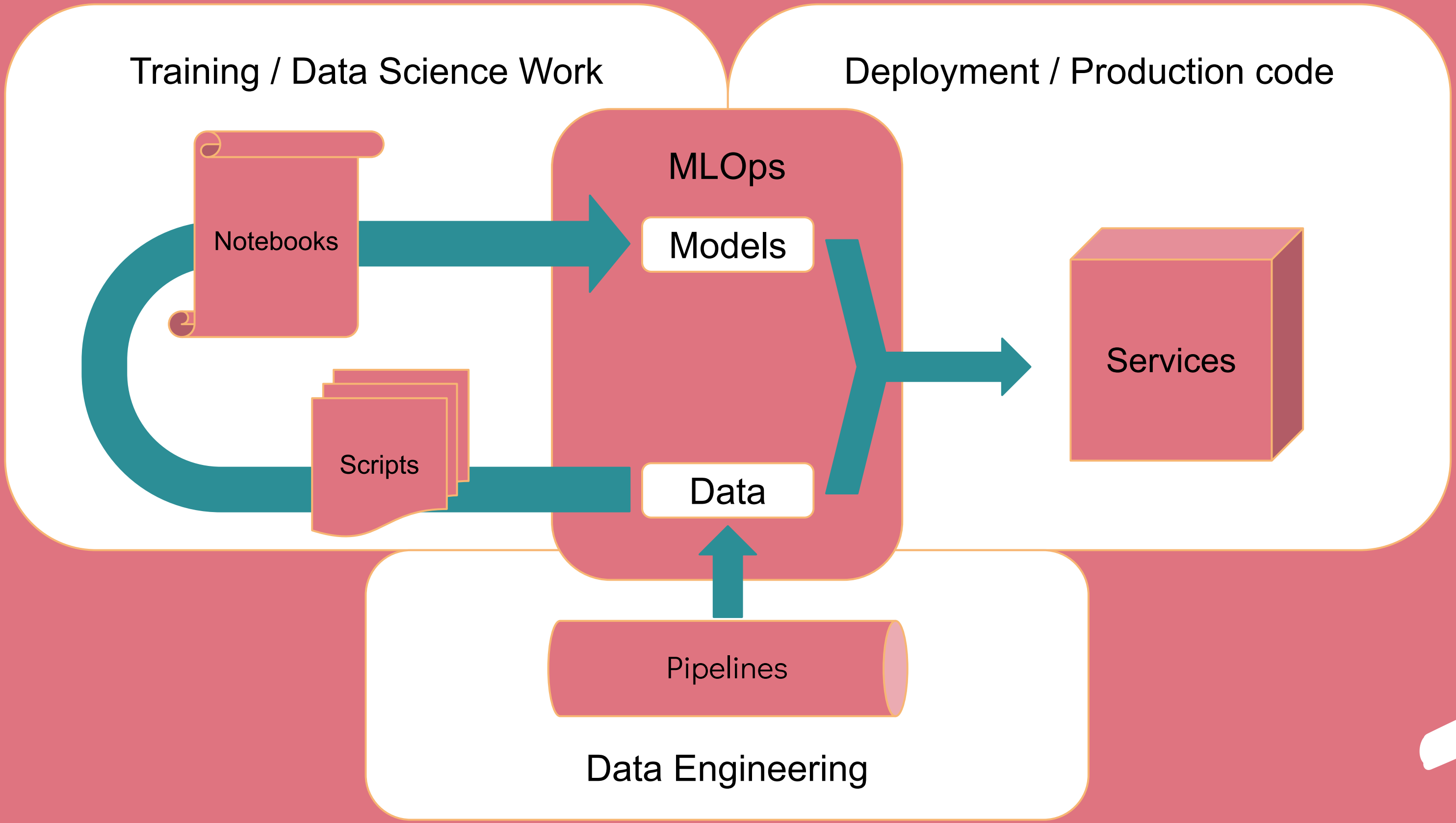
There are no excuses!

- Code:
  - git
  - svn
- Data:
  - DVC
  - Weights & Biases
- Models:
  - MLFlow
  - Neptune
  - Weights & Biases
  - Comet ML

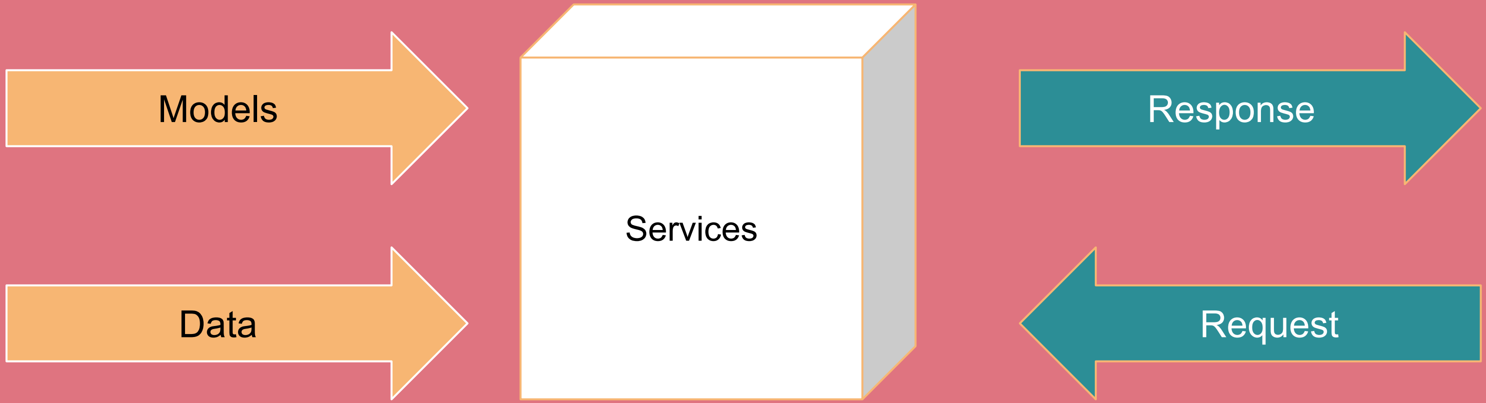
Infrastructure - Streamlining



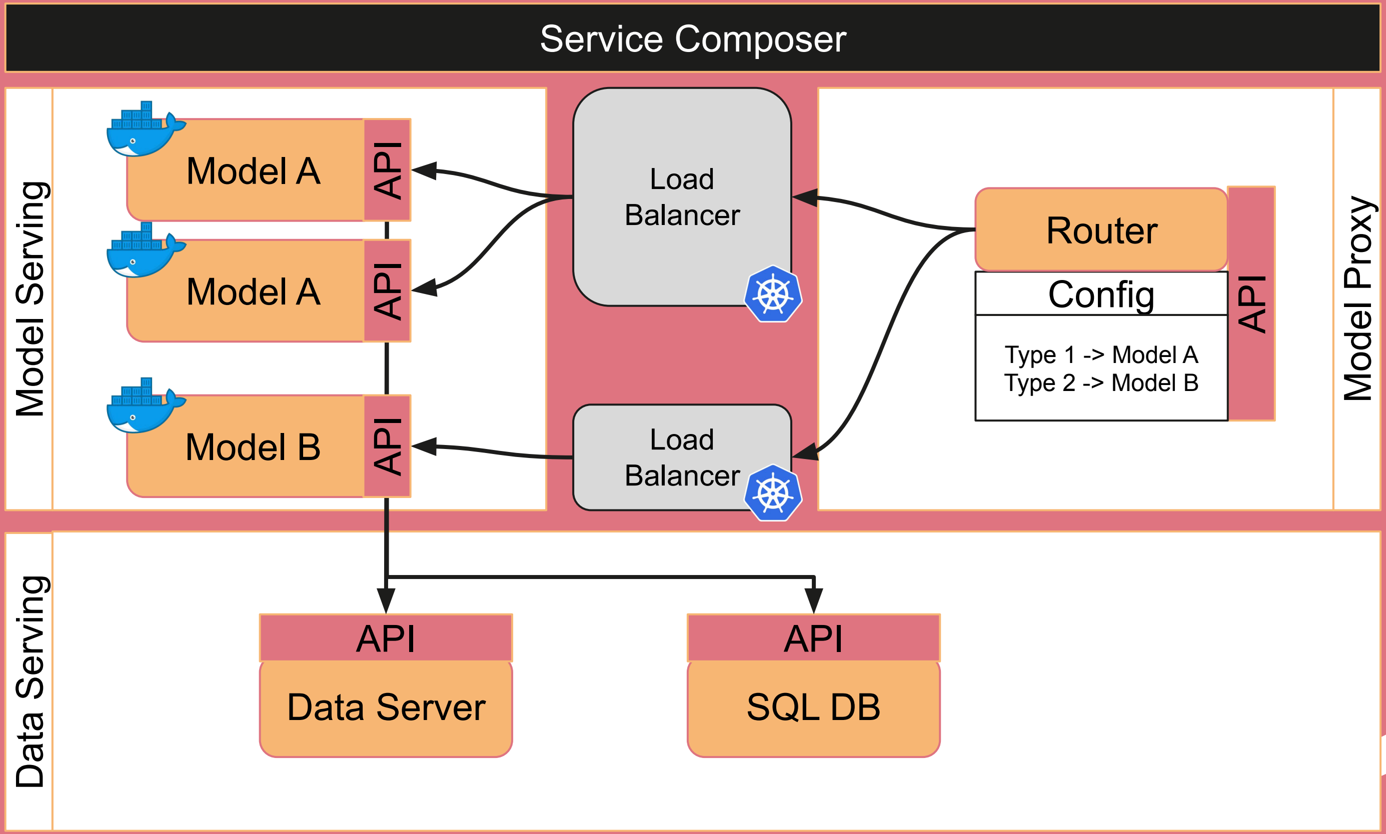
Infrastructure - Deploying



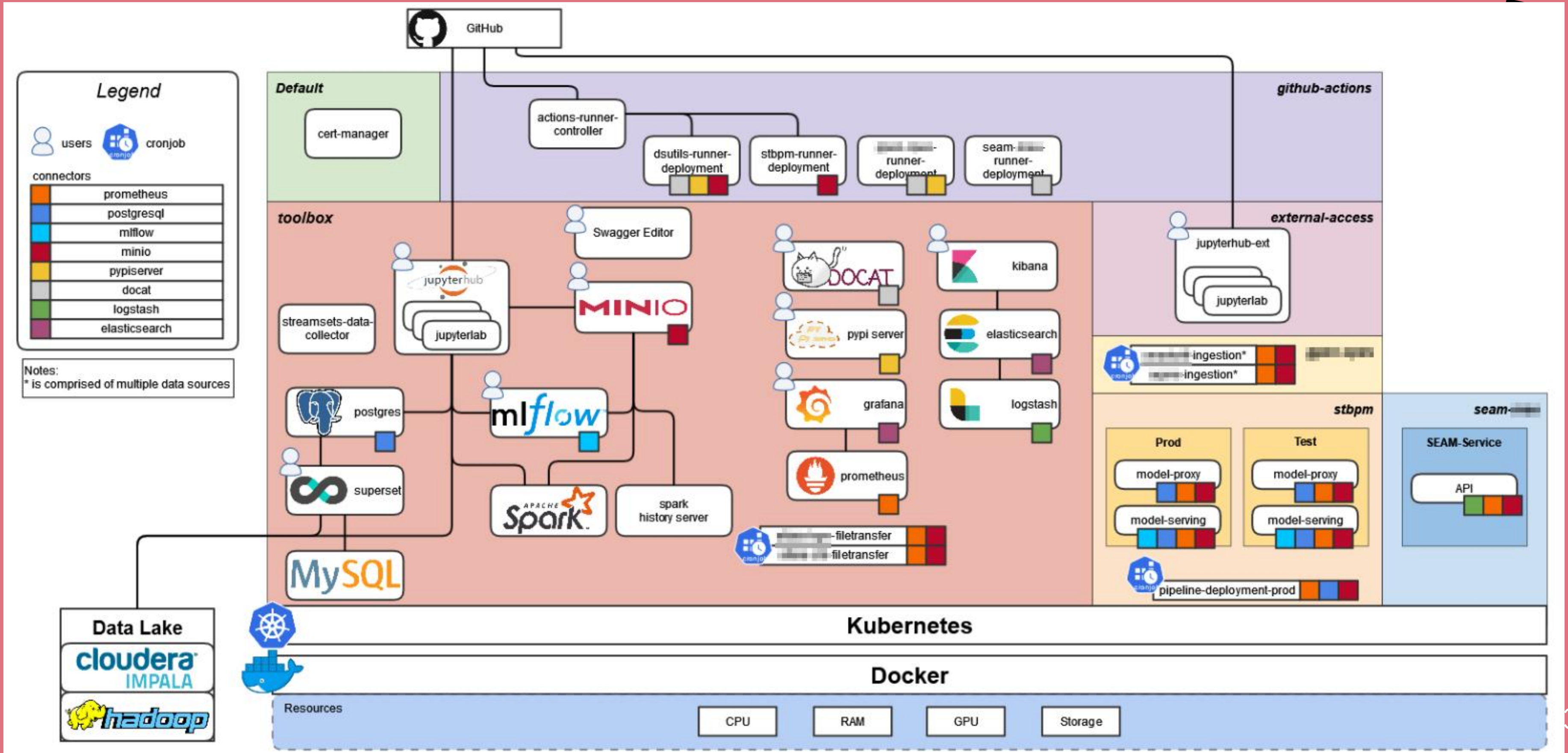
Infrastructure - Deploying



Infrastructure - Deploying



Infrastructure

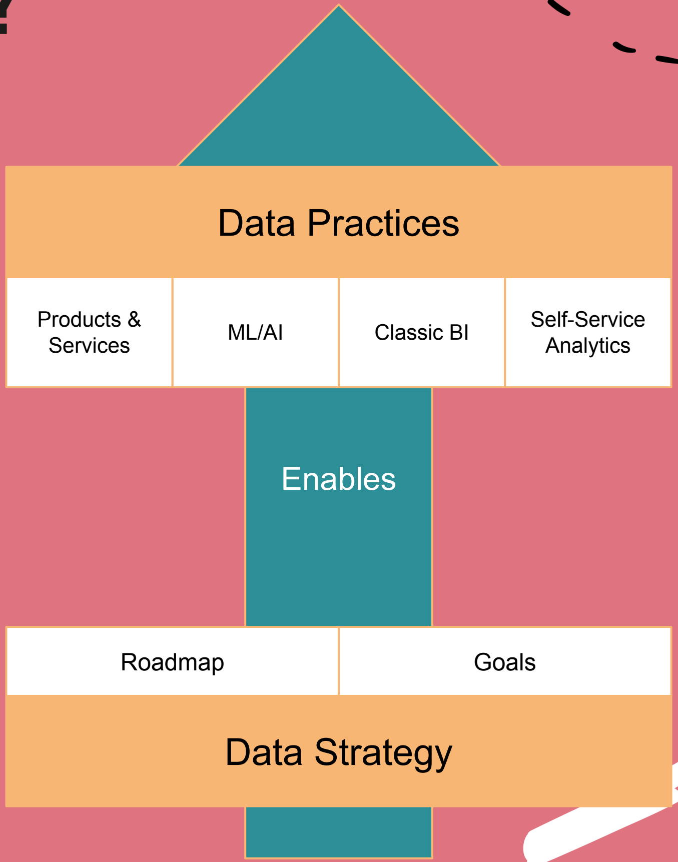


03

Data Strategy

What is it?

- A data strategy is the foundation of all data practices
- It is a framework to achieve a data driven culture
- It is a long-term plan that defines people, processes, and technology to put into place to solve data challenges and support business goals



Why do we need it?

- Data has become crucial to companies' success
- Most companies remain badly behind the curve
- More than 70% of employees have access to data that shouldn't have
- Rogue data propagates in silos
- Companies' data tech often isn't up to the demands put on it
- Slow and inefficient business processes
- Data privacy, integrity and quality that undercut the ability to analyze
- Lack of clarity of business needs and goals
- Inefficient data movement between different parts of the business and/or duplication of data by multiple BUs
- Inefficiency due to lack of roles, rewards and structure:
  - Lack of data standards and literacy
  - Lack of vision, sponsorship and leadership
- The problem: every project addresses data issues as a one-off, built from scratch activities.

Changing the mindset

Data as a
BYPRODUCT

Data as a
BUSINESS ASSET

Transform Data
into Information

Drive Business
Decisions and AI
products

The 5 Pillars

Vision	• Business Strategy and Value • Organizational Goals
Data Governance	• Organization Structure • Data and Information Management
Data Architecture	• Data Framework • Ecosystem Technology
People & Culture	• Data Literacy • Data-driven Mindset • Roles and Rewards
Roadmap	• Execution plan • Responsibility and leadership

04

Takeaways

Takeaways

- Define a clear vision and mission
- Data Science is not only for Data Scientists
- Tech is important but business goals define the value

Thank you!

Reach out!
Email: ruifilipe.rspedro@gmail.com
Linkedin: https://www.linkedin.com/in/rui-filipe-pedro-532774a3/