
The Hitchhiker’s Guide to
the Machine Learning
Engineering Galaxy
by Alyona Galyeva

Agenda

1. Introduction

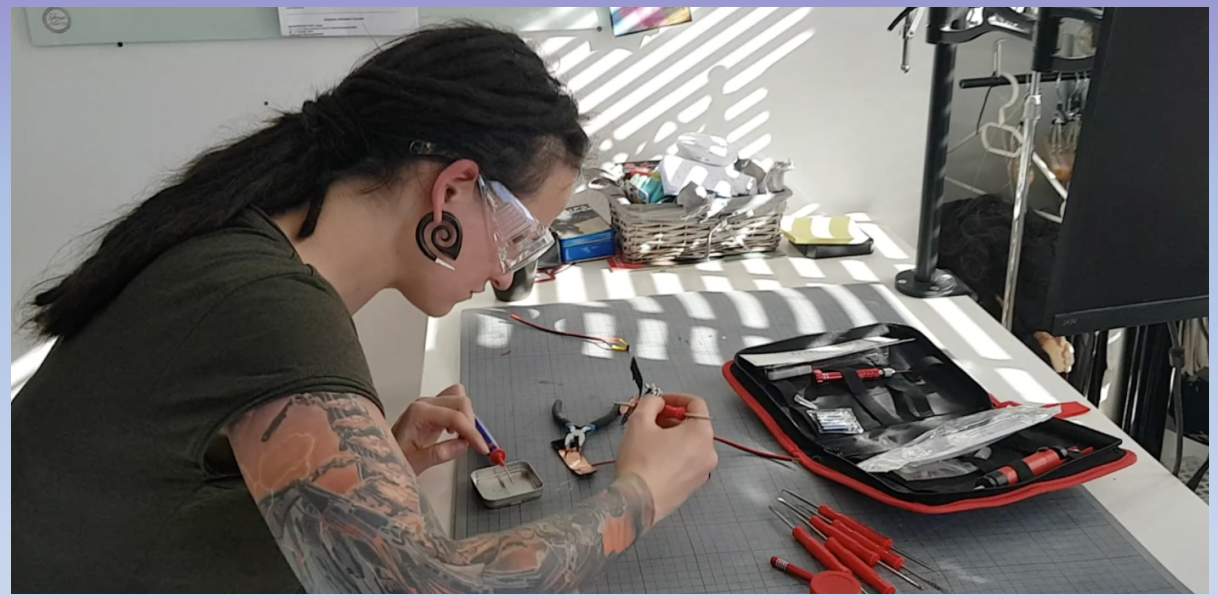
2. Traditional vs AI-powered software

3. MLOps

4. Hands-on with ML serving pipelines:
   - Batch serving
   - Online serving near real-time
   - Real-time serving with embedded model

Introduction


Jury
MAARTEN SUKEL
GAURAV AGARWAL
ALYONA GALYEVA
MICHEL VAN RIJTHOVEN
JUDIT SZAFFENAUER
WaiDATATHON
FOR SUSTAINABLE FUTURE
October 12, 2020
WaiDATATHON
FOR SUSTAINABLE FUTURE
CONGRATULATIONS TO EVERY PARTICIPANT!
YOU DID GREAT!!!
WOMEN IN AI



Traditional
vs
AI-powered software

Machine Learning Terminology

Model - algorithm that learns a solution to a problem from sample data

Data (training, validation, test):

- structured [numerical, categorical]
- unstructured [text, images, videos, audio]

Feature - attribute that is useful or meaningful to a problem

Prediction - sending unseen data to a model and making use of its output.

Machine Learning Roles

Data Scientist - EDA, feature engineering, models

Data Engineer - infra and pipelines around data

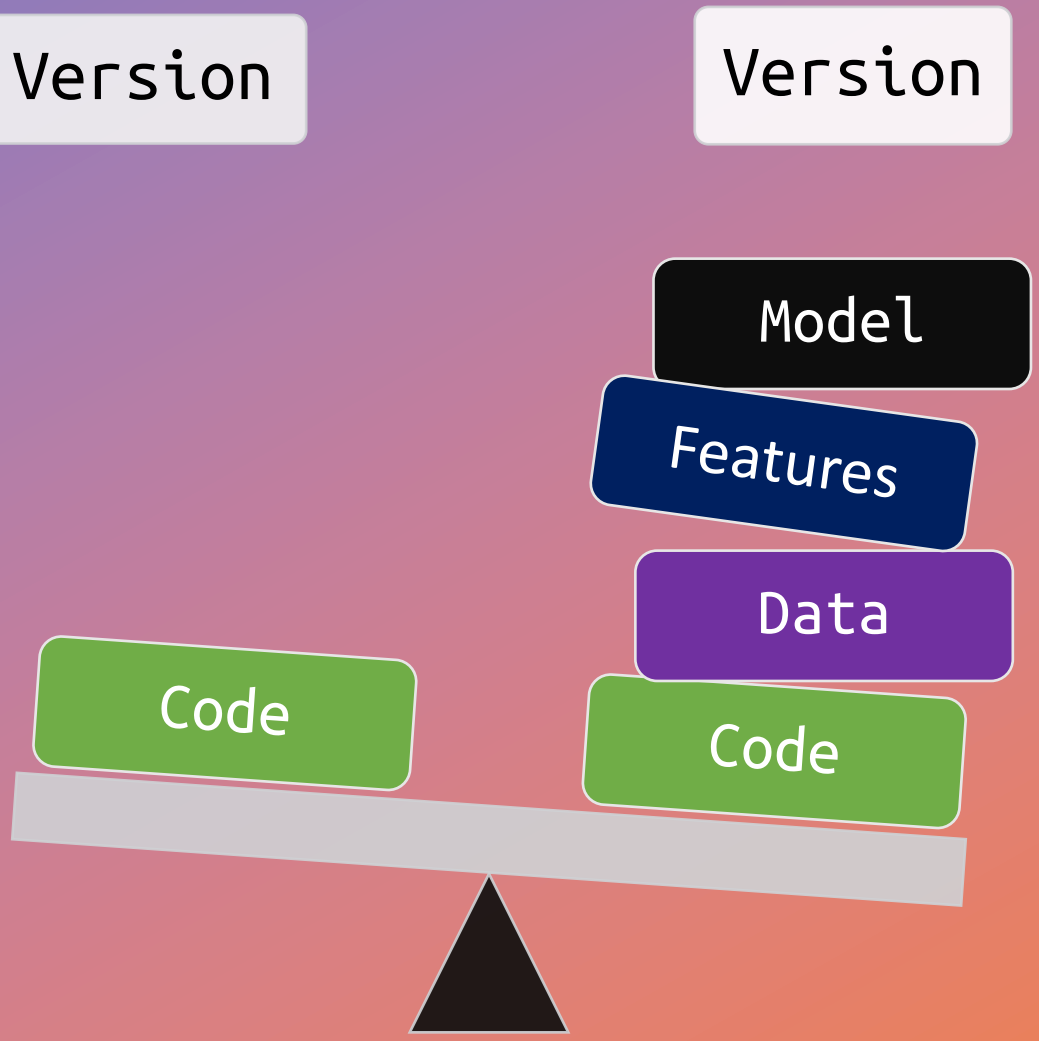
Machine Learning Engineer - infra and pipelines around models

Research Scientist - new algorithms for models

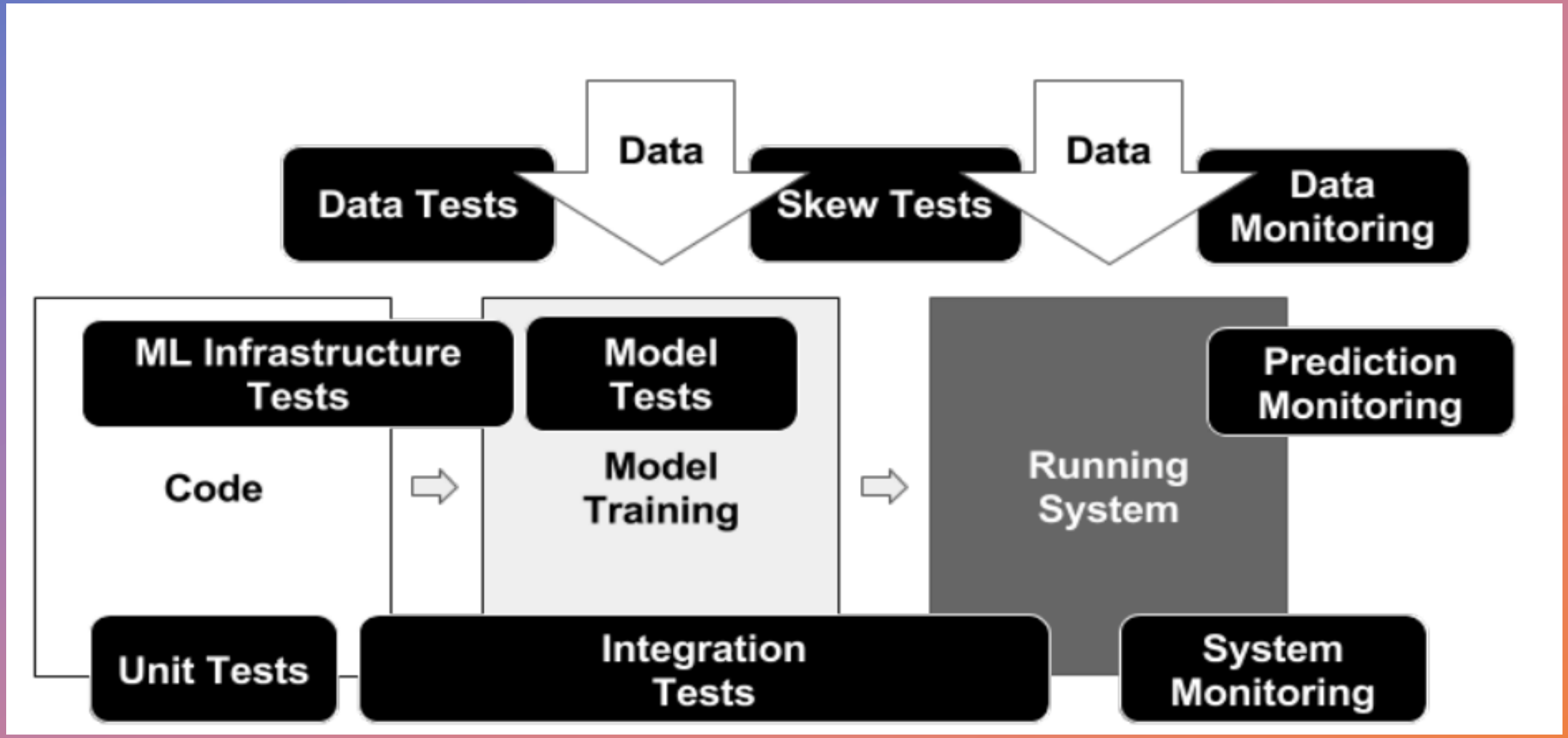
Data analyst - insights from structured data

Developer - production systems enabling end users to access models

Traditional vs AI-software



ML Testing and Monitoring



Source: https://research.google/pubs/pub46555/

Common ML Challenges

Data quality - “garbage in, garbage out” (accuracy, completeness, consistency, timeliness)

Reproducibility - inherent element of randomness

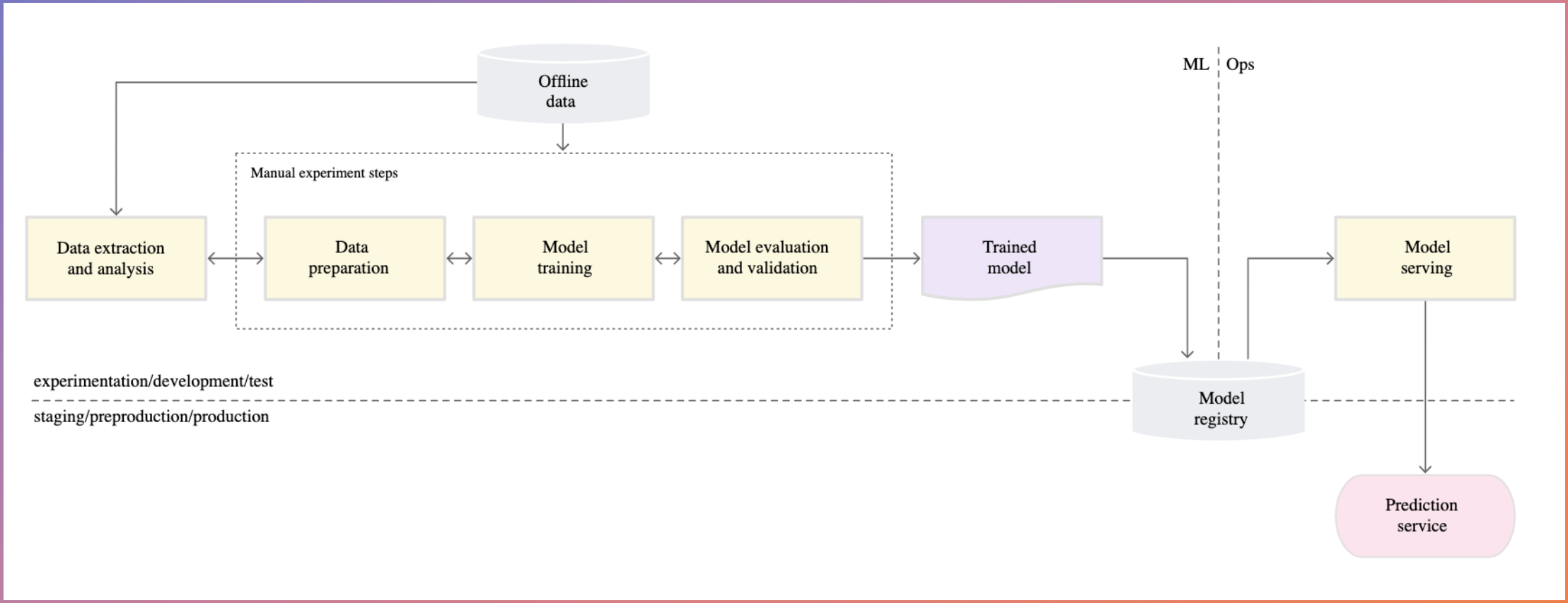
Data Drift - data change over time

Scale - infrastructure needs per each step

Multiple Objectives - balanced optimizations

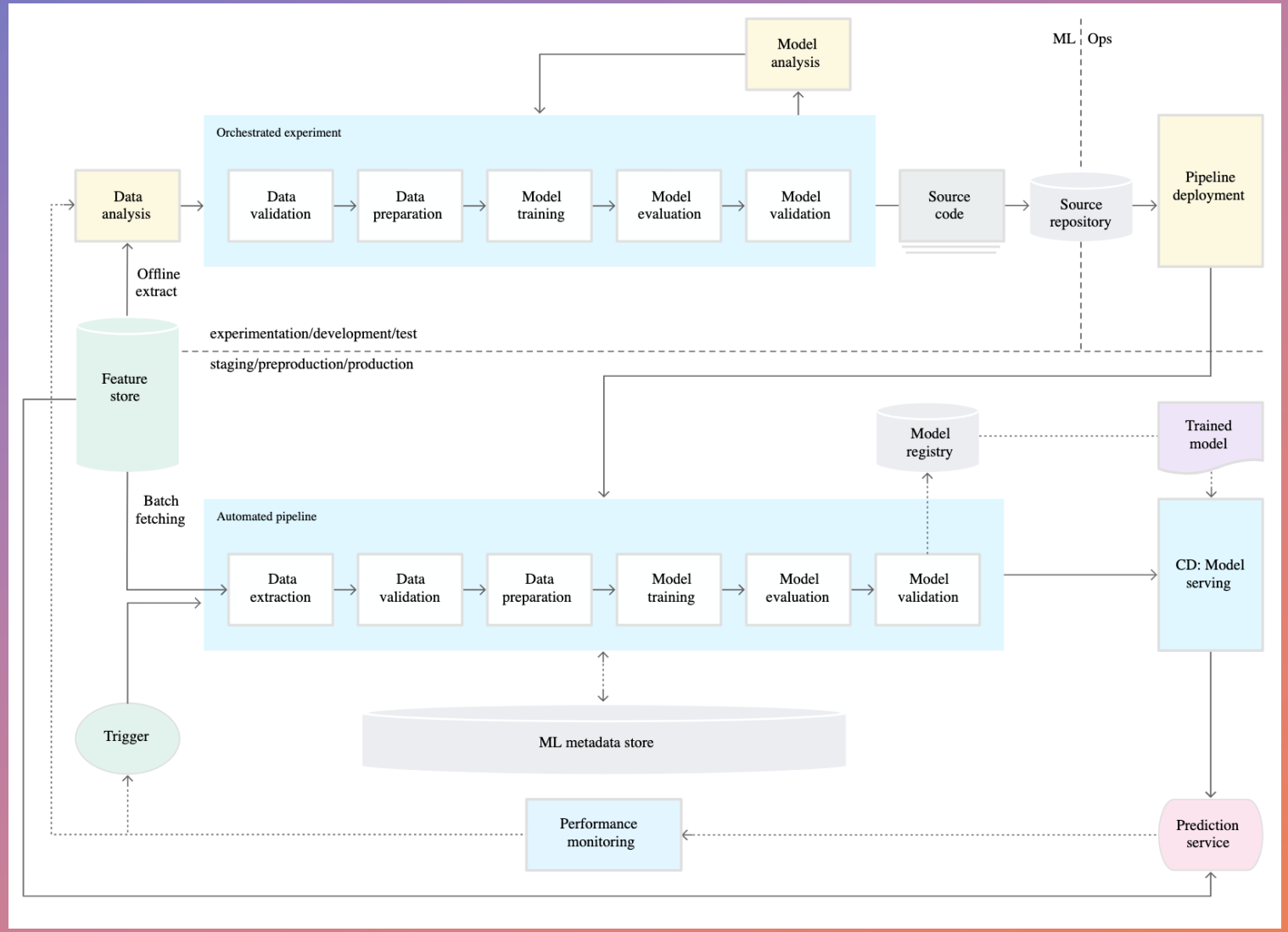
MLOps

MLOps level 0: Manual process



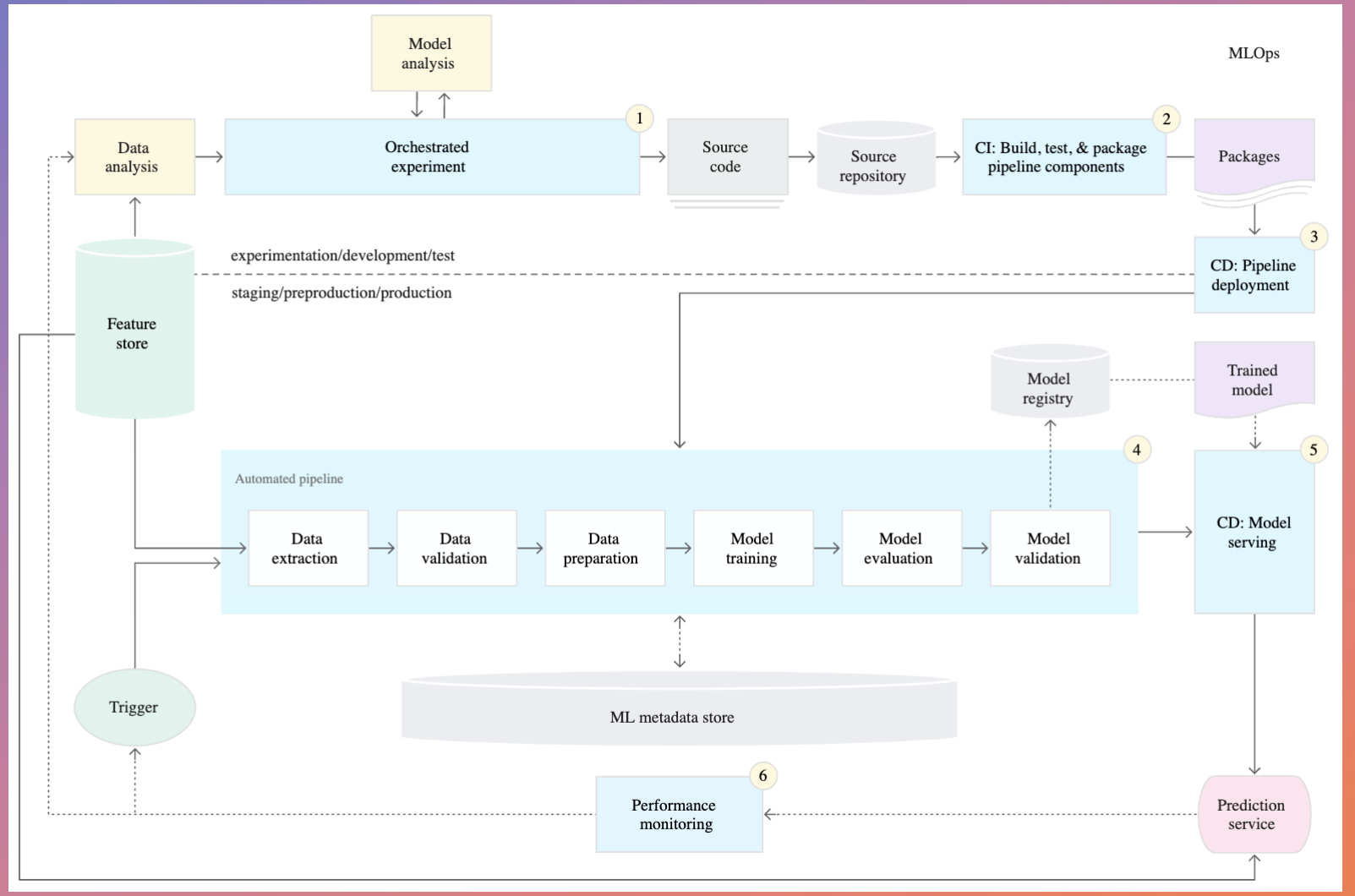
Source: https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

MLOps level 1: ML pipeline automation



Source: https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

MLOps level 2: CI/CD pipeline automation



Source: https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

ML / AI Infrastructure

DATA PREPARATION

Data Exploration & Processing
Paxata, TRIFACTA, alteryx, databricks, Superb AI, iguazio, Qubole, HIVE, SAS

Data Version Control
Pachyderm, data iku, DVC

Feature Engineering and Storage
FEATURE LABS, Featuretools, tecton, Google Feast

Data Labeling
scale, Labelbox, figure eight, Amazon SageMaker, Google AI, Azure Machine Learning

Data Quality Checks
great_expectations

MODEL BUILDING

Hosted Notebooks Management
Amazon SageMaker, Azure Machine Learning, databricks, CLOUDERA, CO, DOMINO, Deepnote

Model Management, Version Tracking and Storage
databricks, Amazon SageMaker, Google AI, Azure Machine Learning, iguazio, ALGORITHMIA, DOMINO, data iku, SAS

Experiment Tracking
Weights & Biases, comet, TensorBoard, mlflow

Model Optimization Hyper Parameter
SIGOPT, Weights & Biases, Amazon SageMaker, anyscale, comet

Auto ML
H2O.ai, DataRobot, Amazon SageMaker, Google AI, Azure Machine Learning, data iku

Model Training
Amazon SageMaker, Azure Machine Learning, Google AI, Kubeflow, anyscale, SAS, iguazio, PerceptiLabs

Model Evaluation
TensorBoard, Streamlit

Model Explainability
fiddler, TensorBoard

PRODUCTION

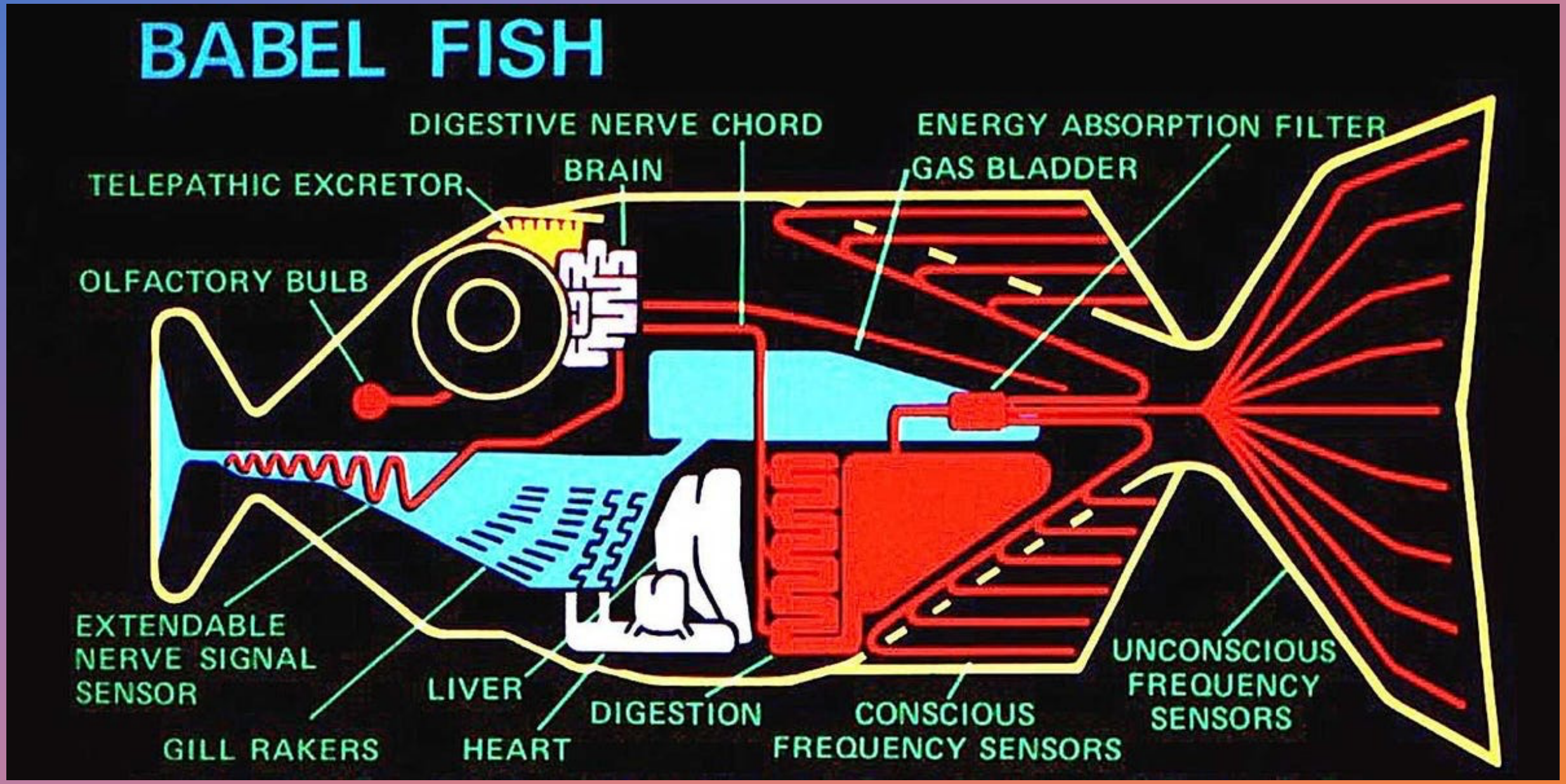
Model Observability
arize, DataRobot, Other startups

Model Compliance & Audit
fiddler, SAS, Other startups

Model Deployment and Serving
Amazon SageMaker, Kubeflow, ALGORITHMIA, DataRobot, Google AI, Azure Machine Learning, PerceptiLabs, iguazio, SAS

Model Validation
arize, fiddler, SAS, Other startups

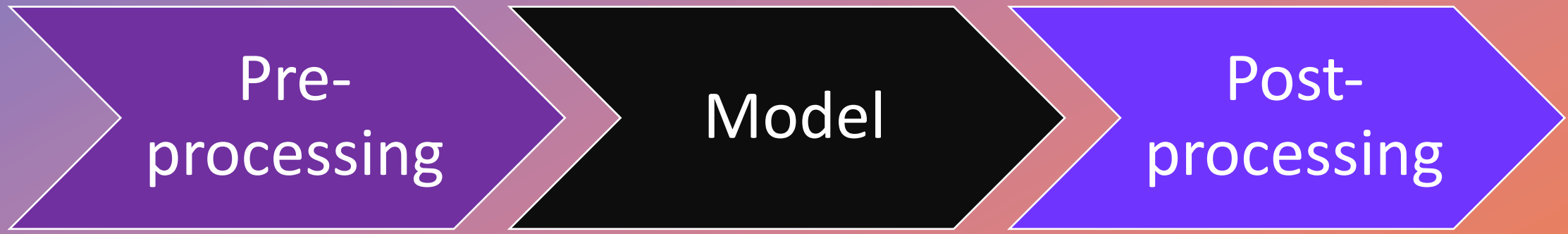
Platform Specific Model Builds
OctoML


BABEL FISH
DIGESTIVE NERVE CHORD
ENERGY ABSORPTION FILTER
BRAIN
GAS BLADDER
TELEPATHIC EXCRETOR
OLFACTORY BULB
EXTENDABLE NERVE SIGNAL SENSOR
LIVER
DIGESTION
UNCONSCIOUS FREQUENCY SENSORS
GILL RAKERS
HEART
CONSCIOUS FREQUENCY SENSORS


DON'T
PANIC

Hands-on with ML serving pipelines

ML serving pipeline



In order to answer the question "How to deploy a model?" we need to understand how end users are going to interact with our model:

- interactive or non-interactive
- single record or batch
- synchronous or asynchronous
- real-time or non-real-time

Today we will explore 3 flavours of model deployments:

- batch serving
- online serving near real-time
- real-time serving with embedded model

Hands-on with ML pipelines

Batch serving

Batch inference is about using data distributed processing infrastructure to carry out inference asynchronously on a large number of instances at once.

What to optimize: throughput, not latency-sensitive

End user: usually no direct interactions with a model. User interacts with the predictions stored in a data storage as a result of the batch jobs.

Validation: offline

Let's try: https://github.com/EzheZhezhe/ML-Batch-Serving

Hands-on with ML pipelines

Online serving near real-time

Online inference is about responding with a prediction to the request of the end user with a low latency.

What to optimize: latency

End user: usually interacts with a model directly available through an API

Validation: offline and online via A/B testing

Let's try: https://github.com/EzheZhezhe/ML-Online-Near-real-time-Serving

Hands-on with ML pipelines

Real-time serving with embedded model

Real-time serving with embedded model is about distributed event-at-a-time processing with millisecond latency and high throughput.

What to optimize: latency and throughput

End user: usually no direct interactions with a model.

Validation: offline and online via A/B testing

Let's try: https://github.com/EzheZhezhe/ML-Real-time-serving-with-Embedded-Model

Happy deployment

All materials could be found here

https://github.com/EzheZhezhe/The-Hitchhiker-Guide-to-the-Machine-Learning-Engineering-Galaxy

https://github.com/EzheZhezhe/ML-Batch-Serving

https://github.com/EzheZhezhe/ML-Online-Near-real-time-Serving

https://github.com/EzheZhezhe/ML-Real-time-serving-with-Embedded-Model