Data Community

Data preparation - practical foundations

Maciej Kępa

Roadmap

Data Community

01

AI/ML architecture - how it all fits together

Understand the full ML stack from data sources to production, including roles, cloud reference architectures, layers, tools, and a simple ETL + model demo.

02

Data Preparation - practical foundations

Learn practical techniques for building ML data pipelines, cleaning data, validating datasets, and preparing a dataset in PySpark

03

Feature Engineering – the art of extracting value from data

Discover how to create effective features, handle different types of data, apply encoding and normalization, monitor feature drift, and track features with Spark + MLflow.

04

ML Algorithms - the classical approach

Get familiar with core algorithms like regression, trees, and gradient boosting, understand classification, regression, clustering, their pros and cons, and try MLlib demos in Databricks.

05

Model Training in Practice

Learn how to split data, perform hyperparameter tuning, scale training, and train models with code and metrics using MLflow.

06

Deep Learning - leveling up

Understand neural network basics, choose between TensorFlow and PyTorch, leverage GPU/TPU scaling, and build a simple CNN with visualization.

07

ML Pipelines - automation and CI/CD

Build repeatable ML workflows with orchestration tools, CI/CD pipelines, model registry/versioning, and a Databricks pipeline demo.

08

MLOps - manage your ML solution

Explore ML lifecycle management with tracking, model monitoring, data drift handling, canary deployments, A/B testing, and monitoring demos.

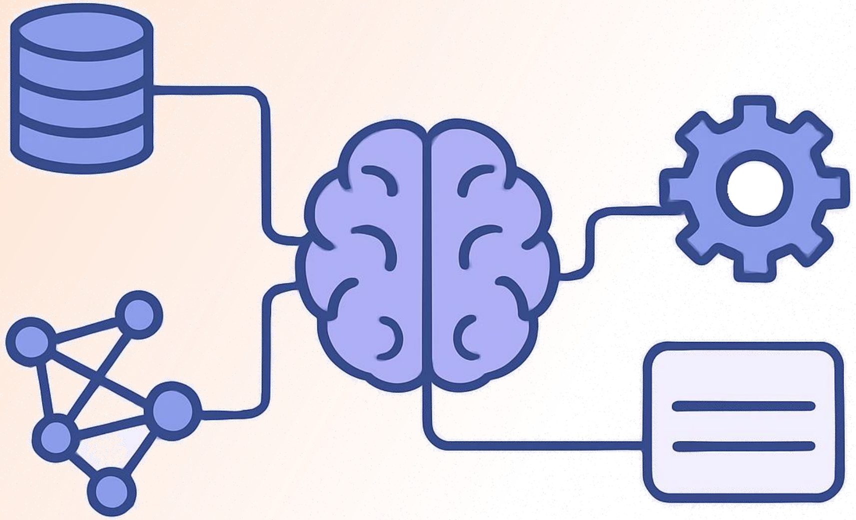
09

Generative AI and LLMs - the new wave of technology

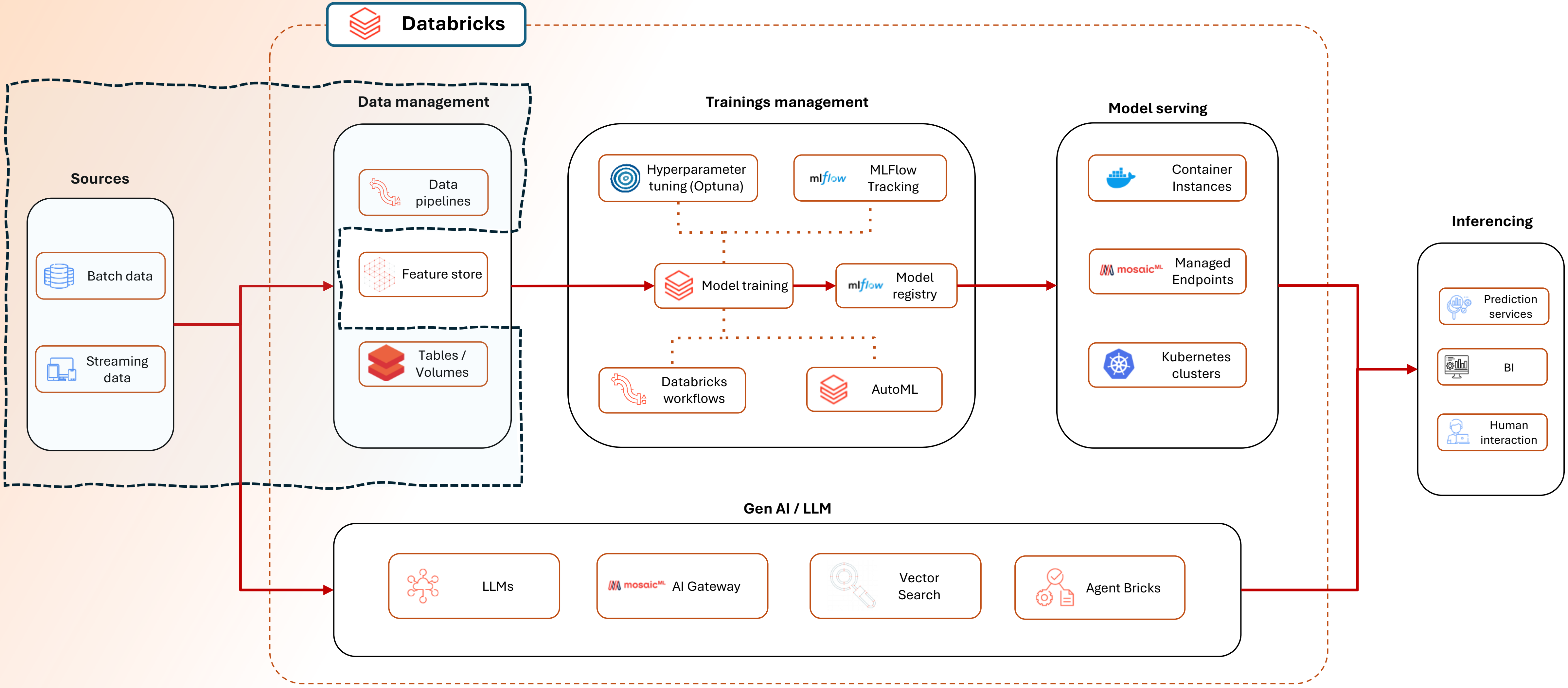
Learn how LLMs differ from classical ML, understand transformers, RAG, AI agent orchestration, and try a simple RAG demo in Azure OpenAI.

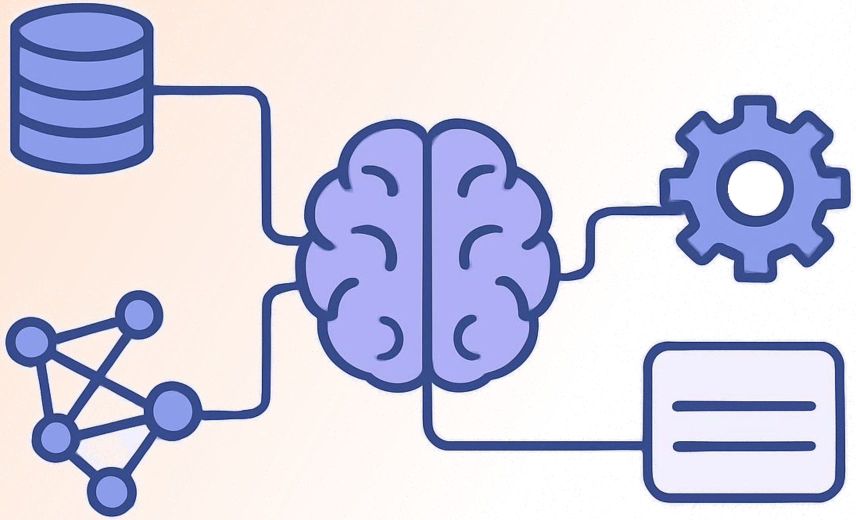
Agenda

1. Introduction
2. From raw to model-ready data
3. Workshop: data cleaning
4. Workshop: data validation
5. Workshop: data pipeline
6. Best practices



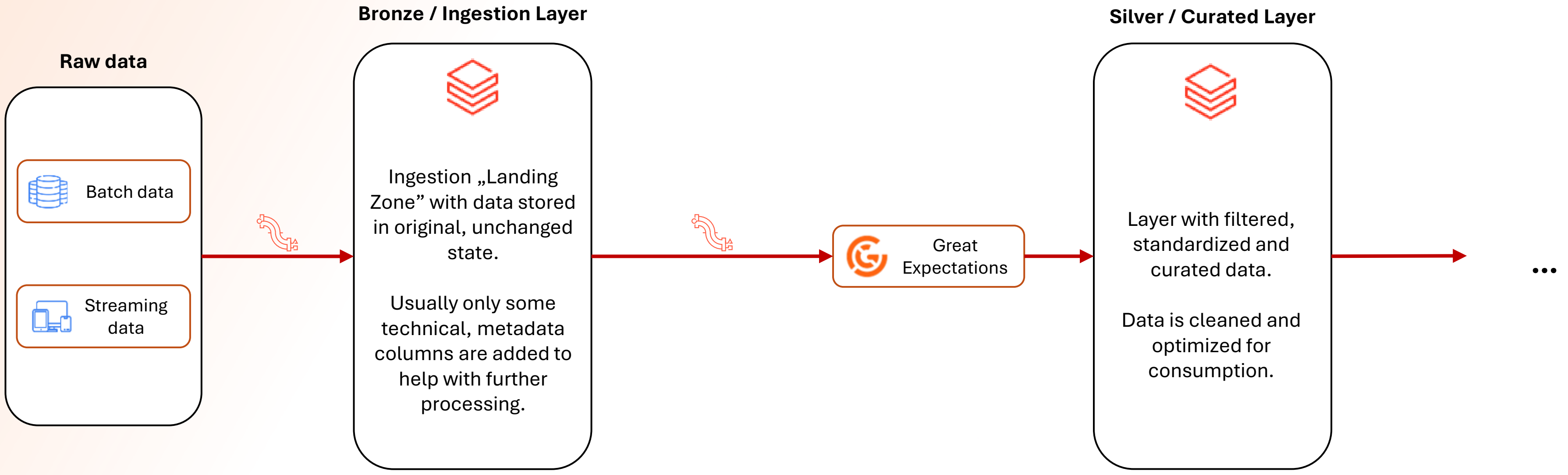
Introduction


Databricks
Sources
Batch data
Streaming data
Data management
Data pipelines
Feature store
Tables / Volumes
Trainings management
Hyperparameter tuning (Optuna)
mlflow MLFlow Tracking
Model training
mlflow Model registry
Databricks workflows
AutoML
Model serving
Container Instances
mosaic ML Managed Endpoints
Kubernetes clusters
Inferencing
Prediction services
BI
Human interaction
Gen AI / LLM
LLMs
mosaic ML AI Gateway
Vector Search
Agent Bricks



From raw to model-ready data

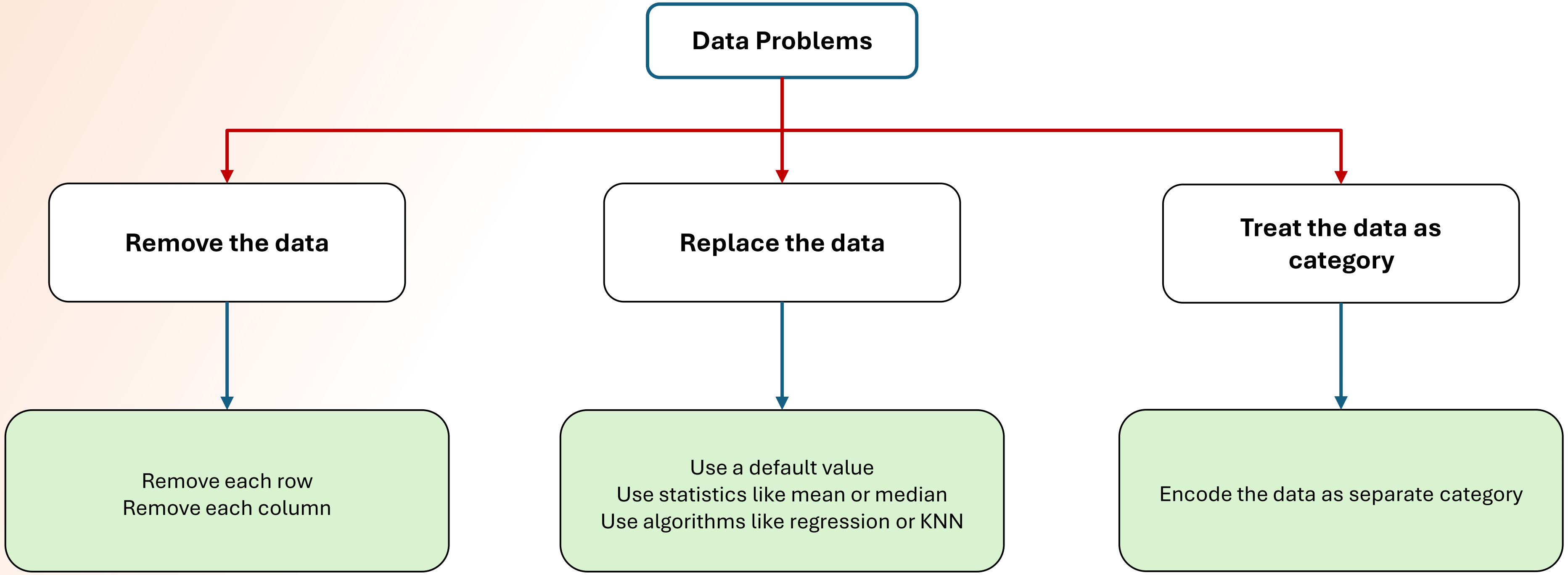
Data pipeline



Most common data problems

1	**Missing values**	Data points are absent due to collection issues or system gaps
2	**Inconsistent data types**	Numeric fields stored as strings or mixed formats breaking transformations
3	**Duplicates**	Repeated records inflate counts and distort model training
4	**Outliers or impossible values**	Extreme values that skew distributions and can mislead model behavior
5	**Inconsistent categorical labels**	Variations like “USA”, “Usa”, “United States” cause incorrect grouping

Handling data quality



Rules for handling data quality:

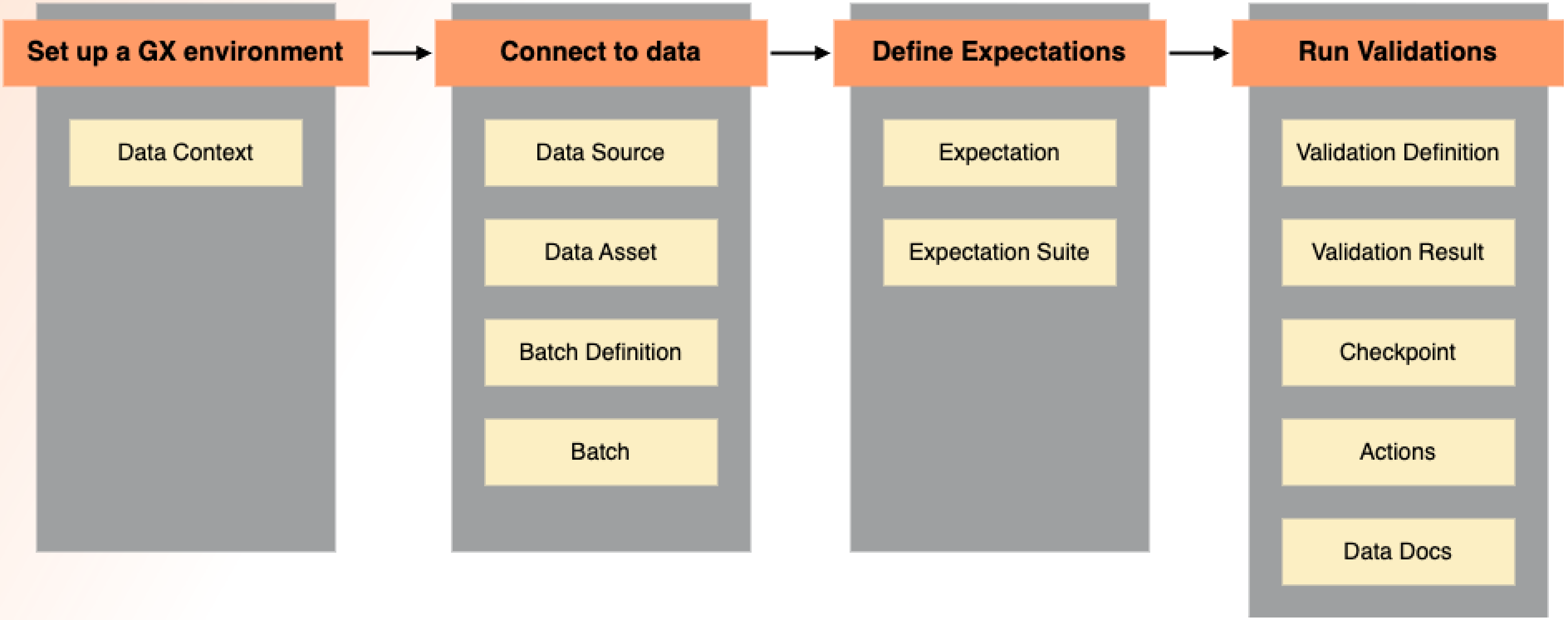
There is no „best” way to do it

Understand the reasons

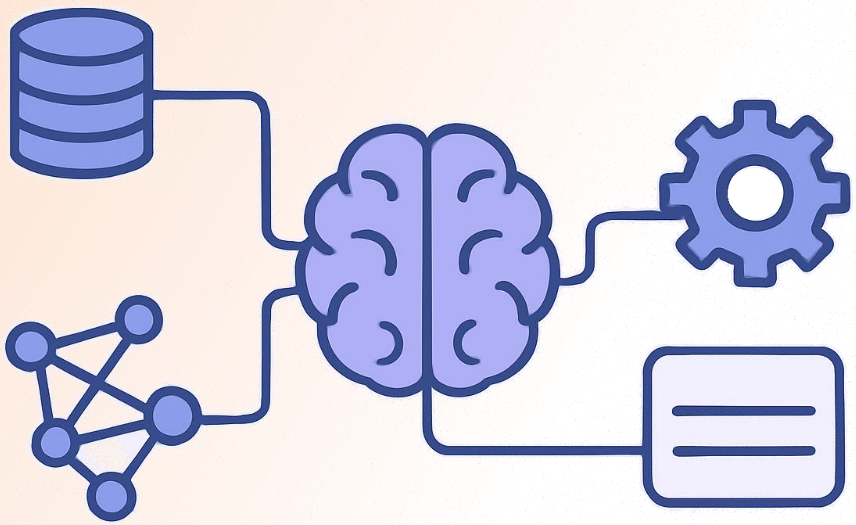
Consider and validate the impact and potential bias

Assess the amount and pattern of missing data

Great Expectations



DEMO



Summary

Takeaways

- **There is no universal cleaning recipe:** every dataset has its own problems, so methods must be chosen based on context, not habit.
- **You can't clean what you don't understand**: meaningful preprocessing starts with understanding semantics, business logic, and how data is generated.
- **Validate early, validate always:** automated checks catch data problems before they hit training or production.
- **Consistency beats cleverness:** stable schemas, clear types, and predictable formats are more beneficial than the most complicated feature engineering.

Thank you!

Contact:

https://www.linkedin.com/in/maciej-kepa

https://github.com/maciejkepa/ai-ml-in-practice