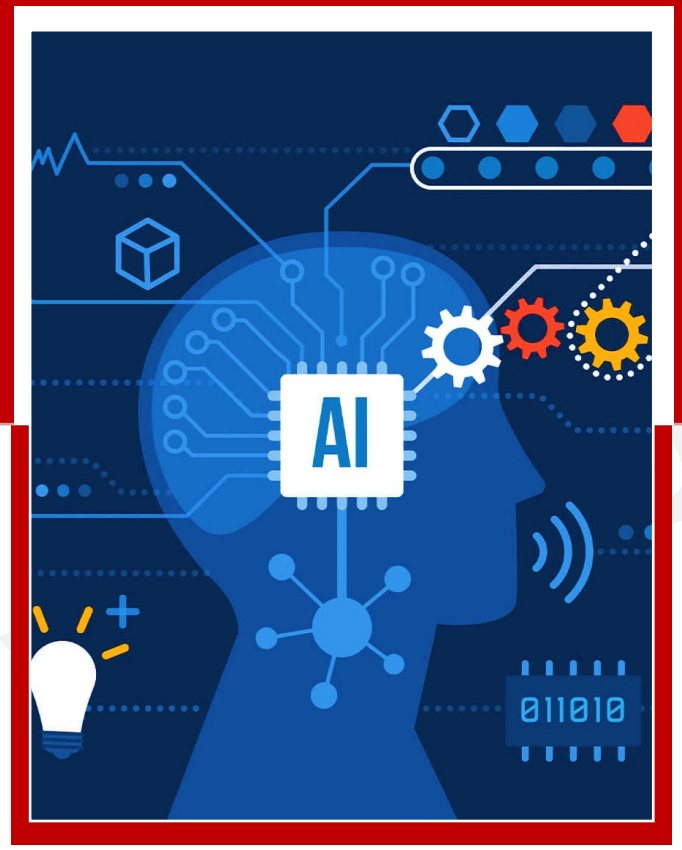
viettel

Data Science
Machine Learning Engineering

Trung tâm phân tích dữ liệu - Khối CNTT
Tổng công ty viễn thông Viettel

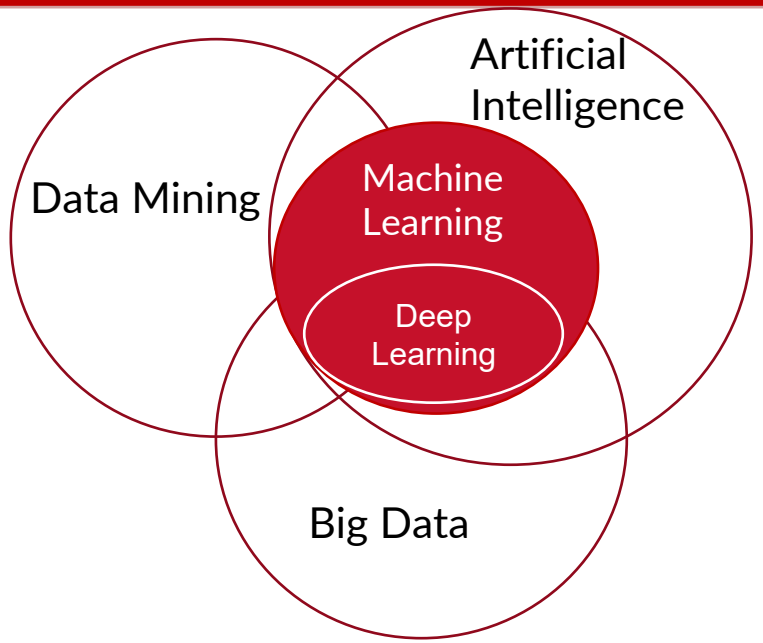
DATA SCIENTISTS



ENGINEER

What's Data Science?

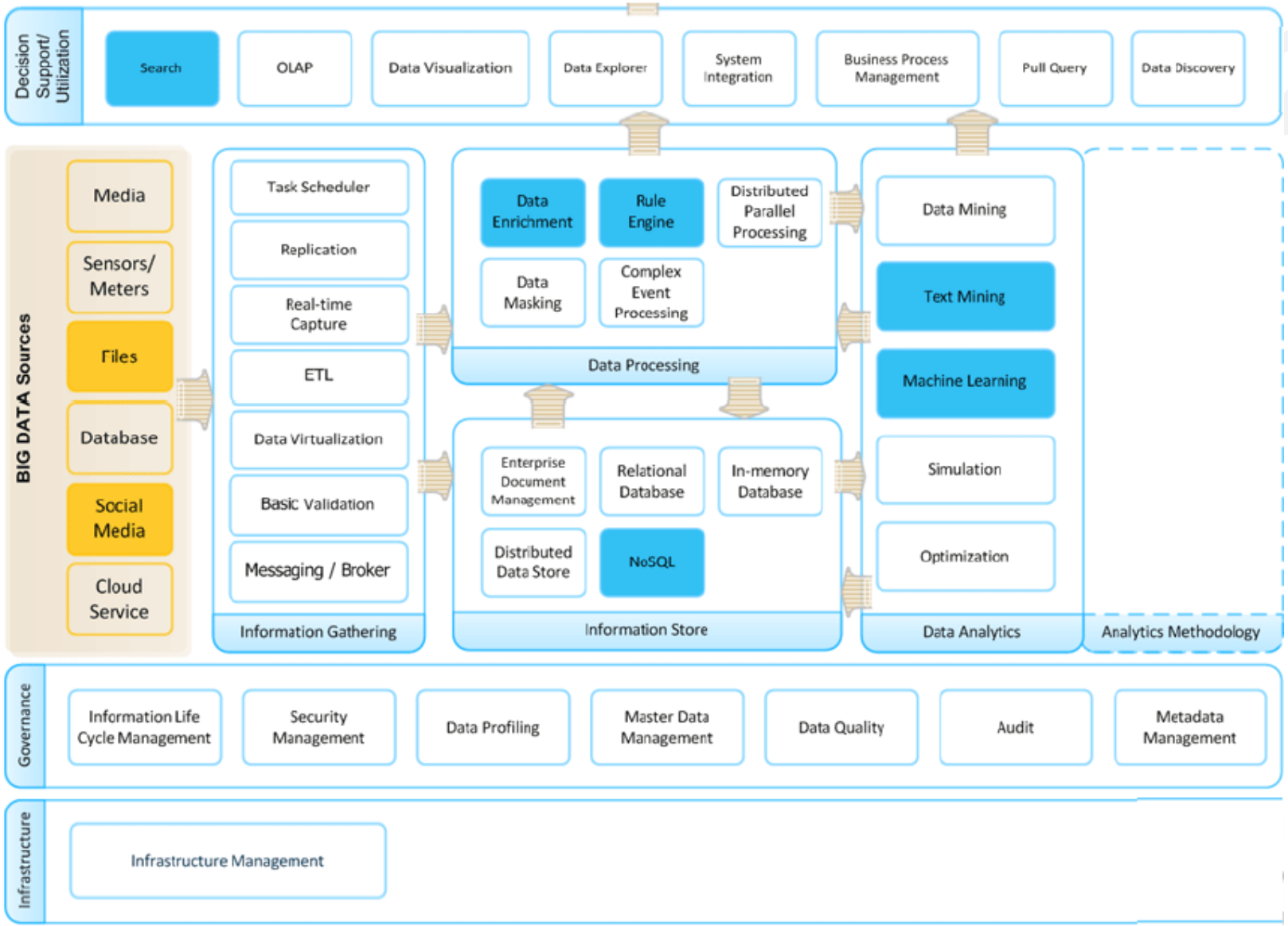
"A data scientist is some one who is better at statistic than any software engineer and better at software engineering than any statistician"
-Josh Wills



viettel

Data Flow

Lộ trình để tham gia nghành khoa học dữ liệu



viettel

ML in research vs. in production

	Research	Production
Objectives	Model performance	Different stakeholders have different objectives
Computational priority	Fast training, high throughput	Fast inference, low latency
Data	Static	Constantly shifting
Fairness	Good to have	Important
Interpretability	Good to have	Important

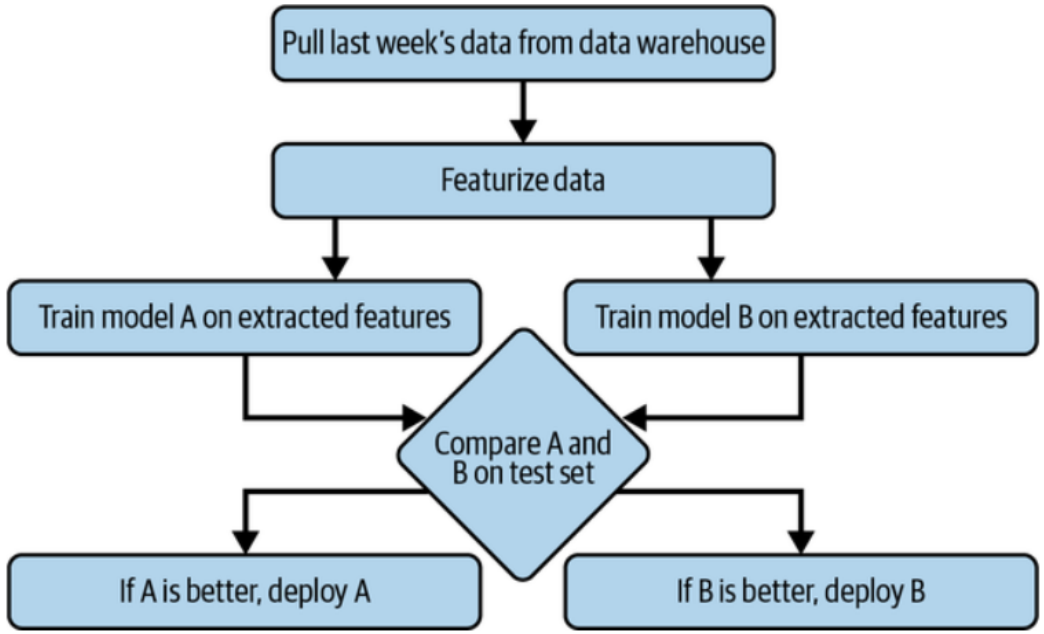
ML team
Highest accuracy

Product
Fastest inference

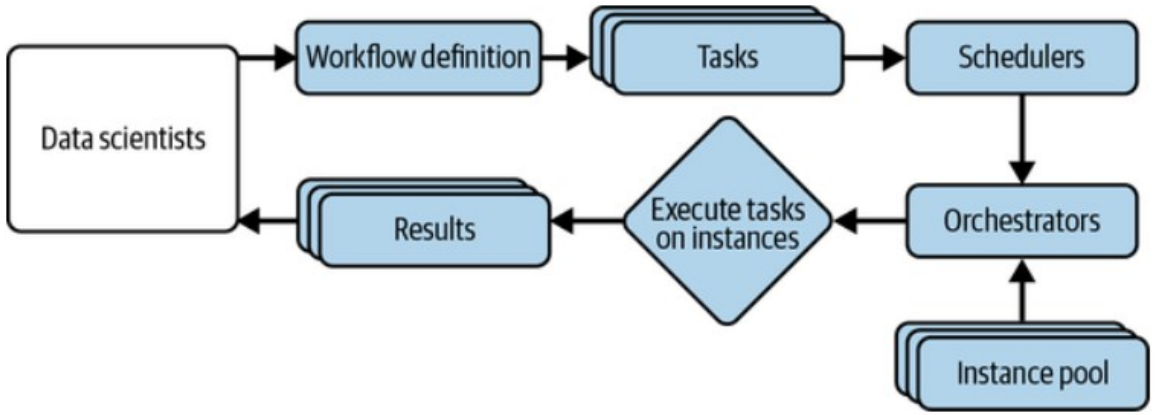
Sales
Sell more ads

Manager
Maximizes profit
= laying off ML teams

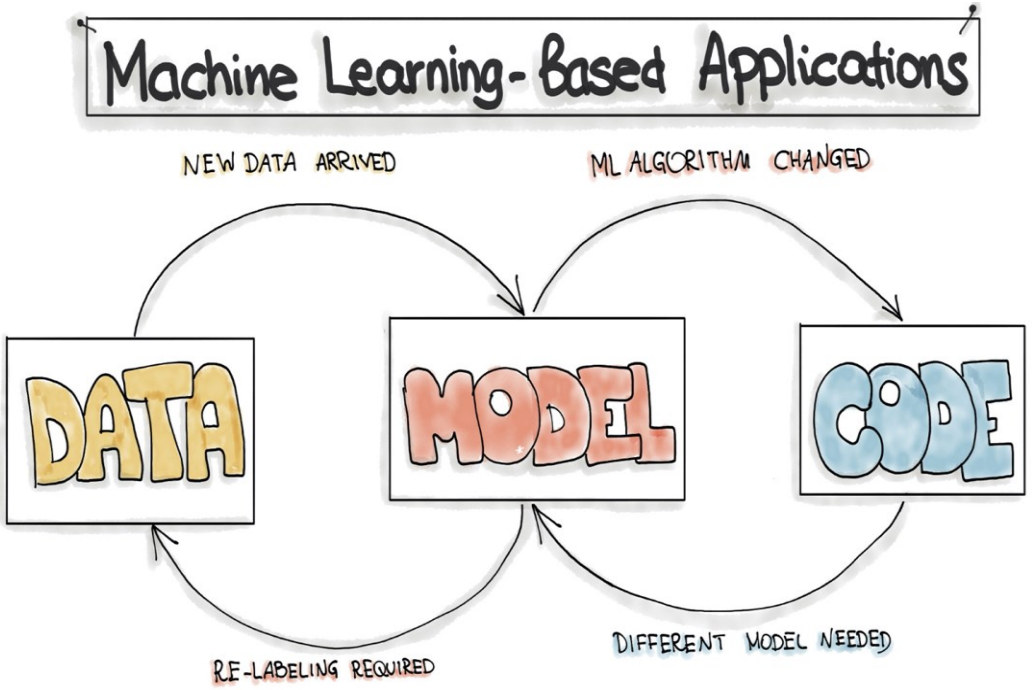
A DAG of a simple ML workflow

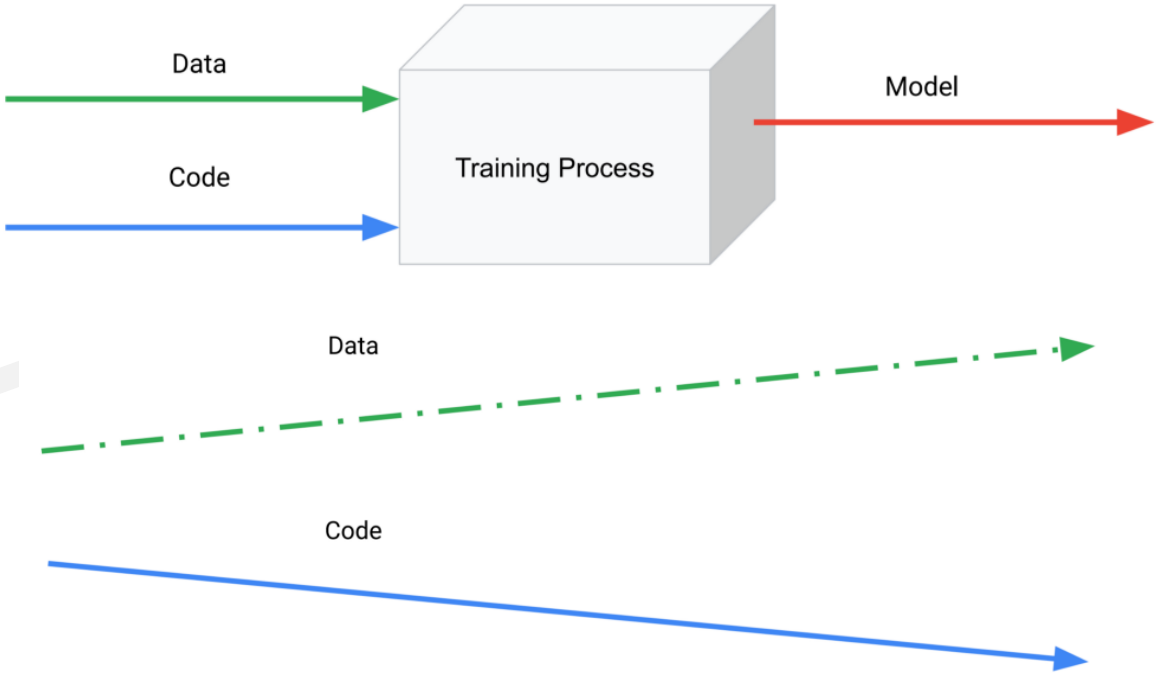


DS workflow



MLops





MLops

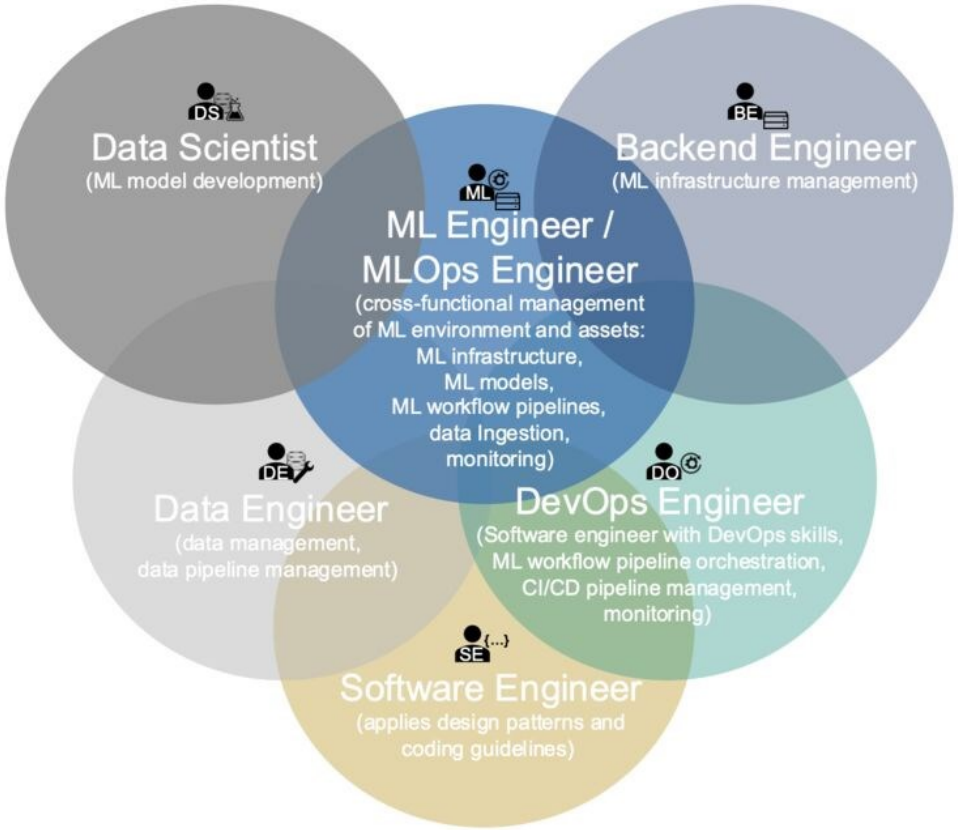
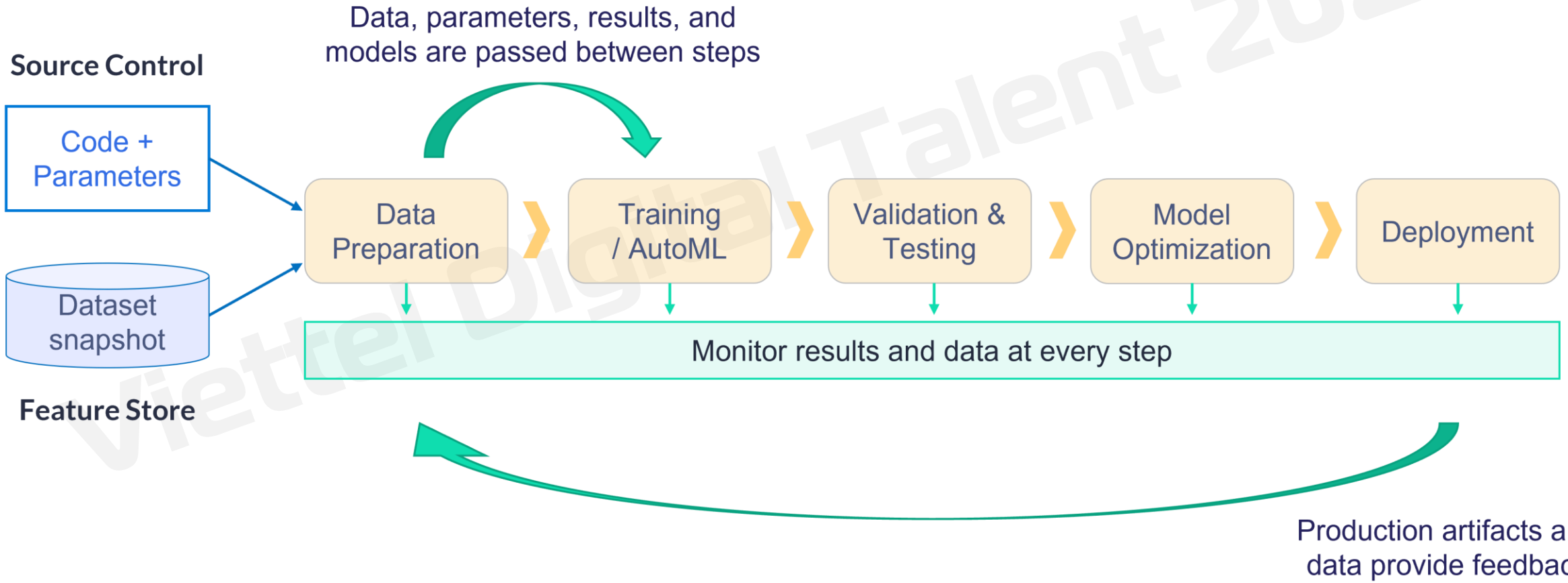


Figure 3. Roles and their intersections contributing to the MLOps paradigm

Machine Learning workflow

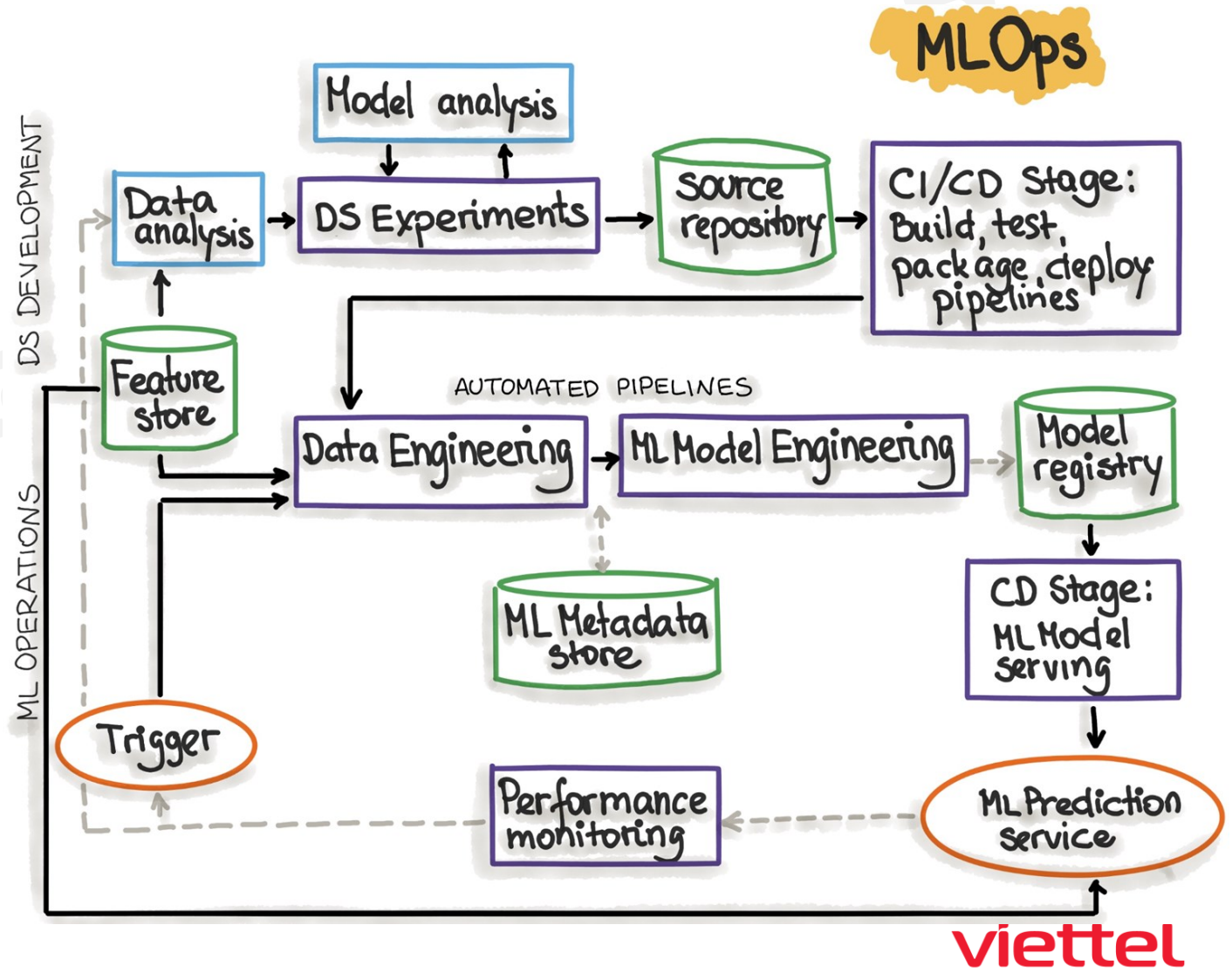
- Manual executions
- Disconnection b/w DS & Ops engineers
- Infrequent release iterations
- No CI/CD
- No monitoring



viettel

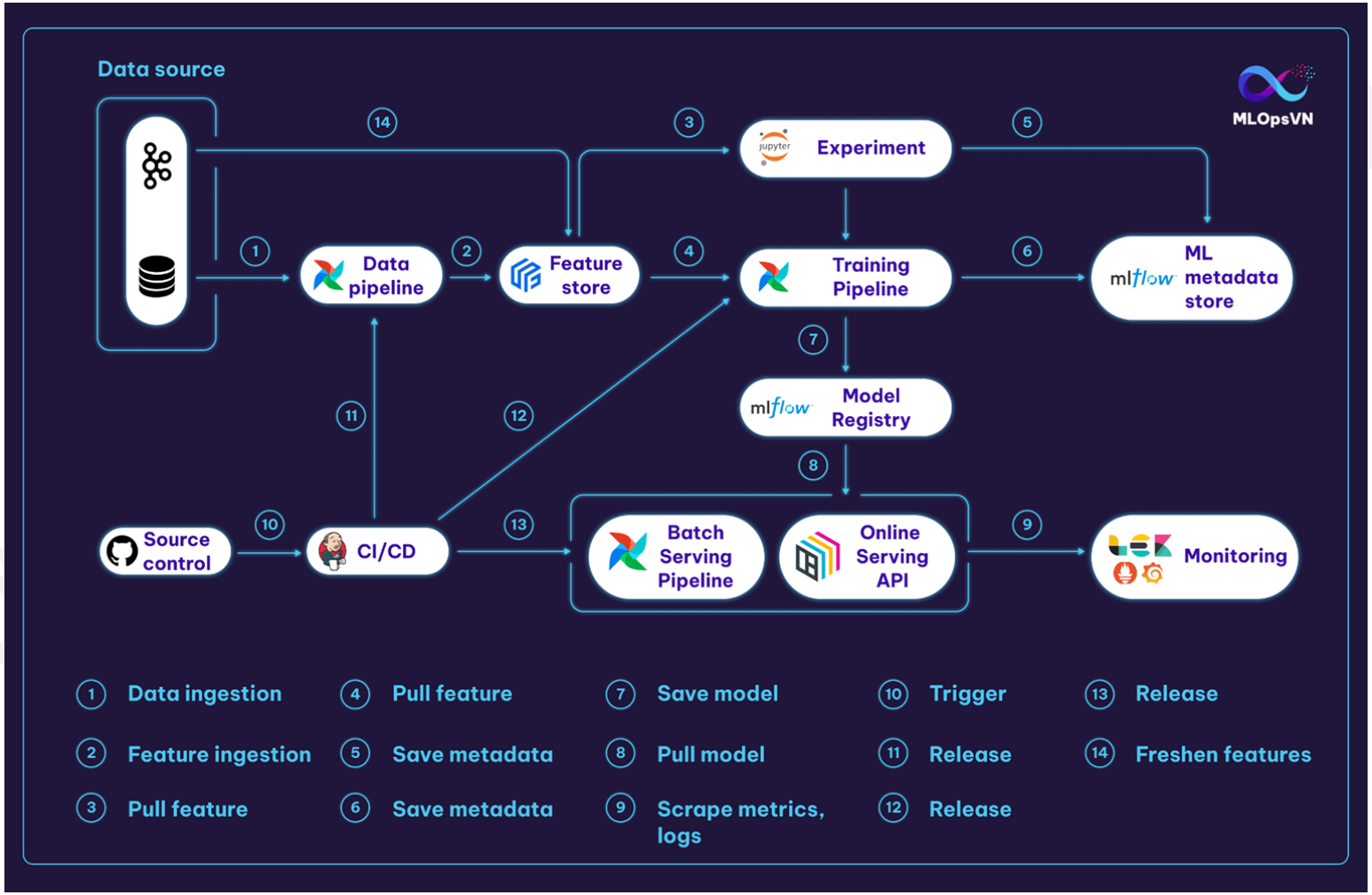
Components of MLops

- Source control
- CI/CD tool
- Feature Store
- Data/ML Pipelines
- Model Registry
- ML Metadata Store
- Performance Monitoring



viettel

Components of MLops



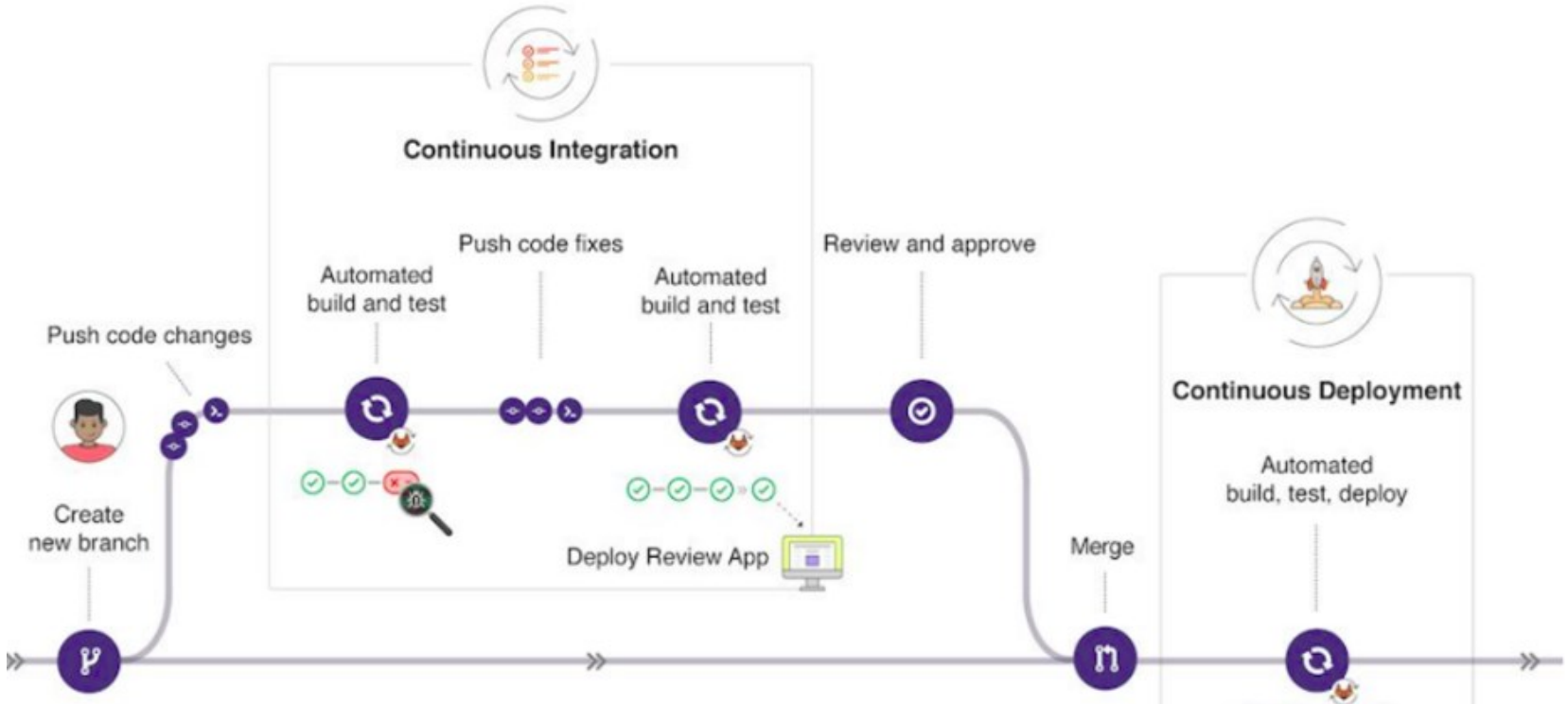
CI/CD automation

- Git workflow
- Code version control

- Data version control

Continuous...	Chi tiết	Tool
Integration (CI)	Tự động hóa quá trình kiểm thử	Gitlab CI hoặc Jenkins
	Tự động hóa quá trình đóng gói code và môi trường	
Delivery (CD)	Tự động hóa quá trình deploy serving API	
	Tự động hóa quá trình deploy các pipeline	
Training (CT)	Tự động hóa quá trình run các data pipeline chuẩn bị feature cho model	Airflow hoặc Kubeflow Pipelines
	Tự động hóa quá trình run training pipeline để train model	
Monitoring (CM)	Tự động hóa quá trình theo dõi và cảnh báo hiệu năng model và tài nguyên hệ thống	Prometheus và Grafana

CI/CD workflow



CI/CD user workflow

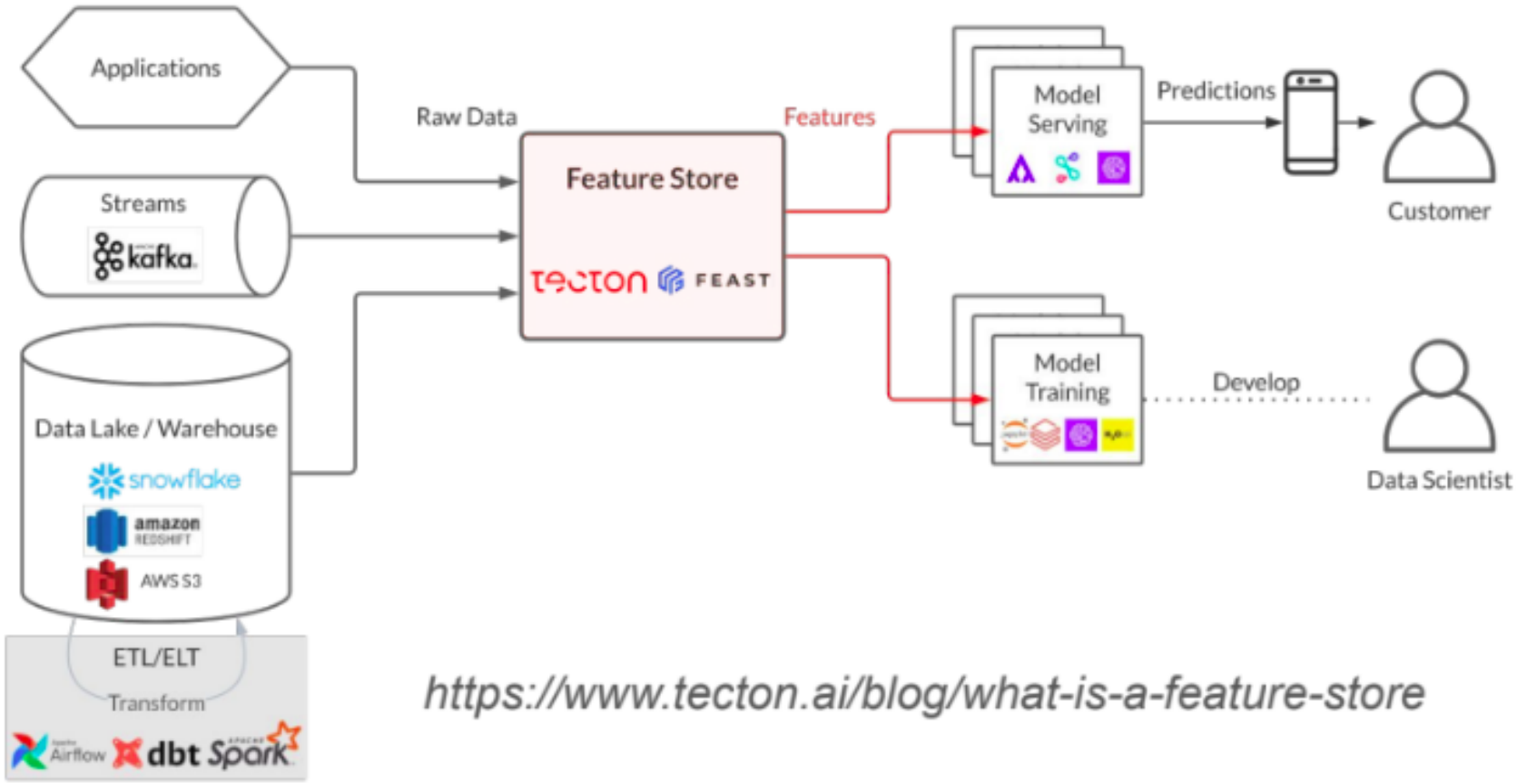
(https://docs.gitlab.com/ee/ci/introduction/)

viettel

Data pipeline

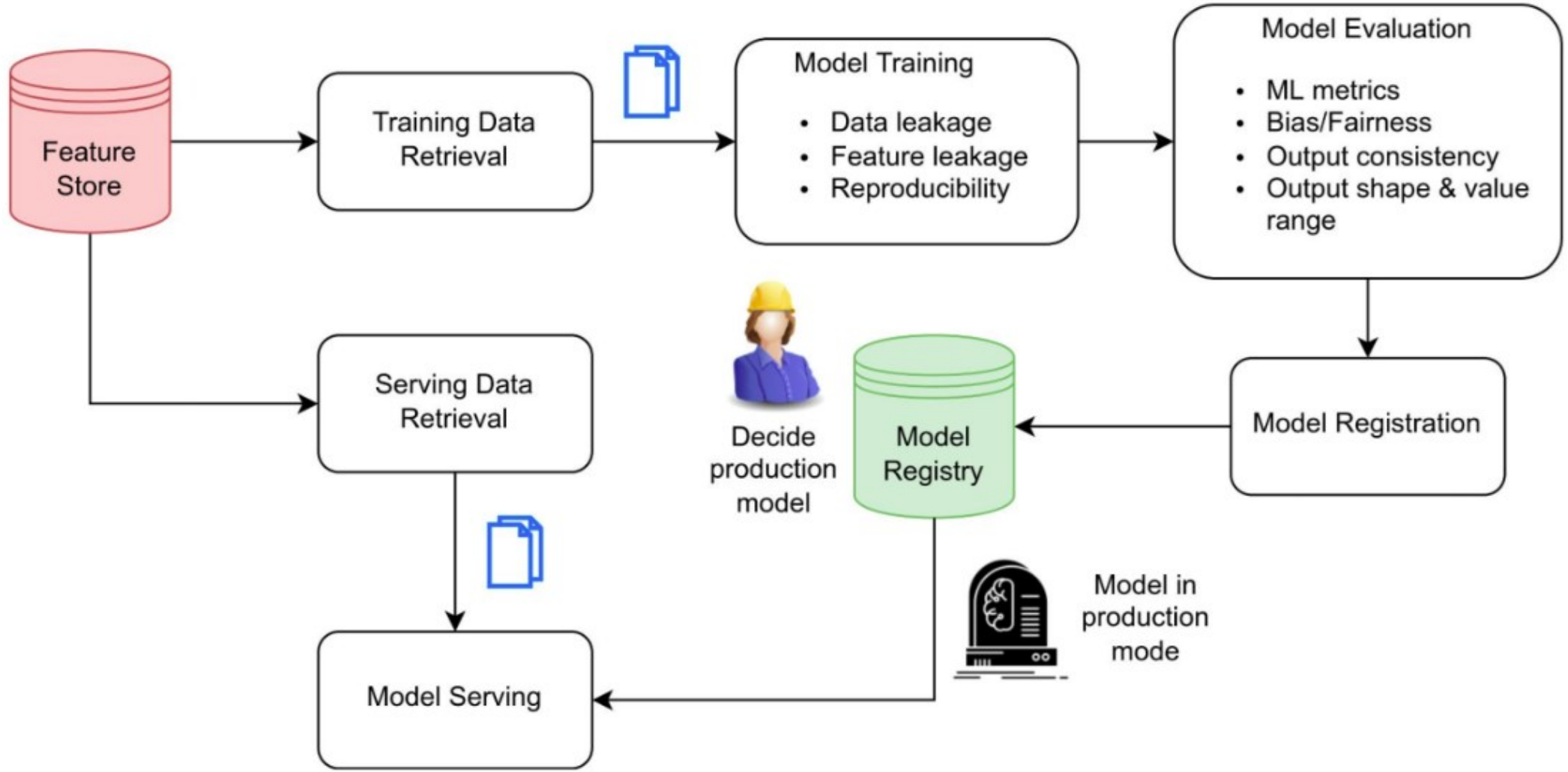
Tên công đoạn	Công việc cụ thể
Data ingestion	Data provenance: lưu trữ thông tin về các dữ liệu nguồn
	Metadata catalog: lưu trữ thông tin về dữ liệu bao gồm: kích thước, định dạng, người sở hữu, người có quyền truy cập, thời gian sửa đổi gần nhất, .v.v.
	Data formatting: chuyển dữ liệu sang format khác để dễ dàng xử lý
	Privacy Compliance: đảm bảo các dữ liệu nhạy cảm (PII), ví dụ thông tin họ tên khách hàng đi kèm CMND/CCCD, đã được ẩn đi
Data cleaning	Xử lý outlier/missing values
	Loại bỏ các features không liên quan hoặc các sample bị lặp lại
	Thay đổi thứ tự các cột
	Thực hiện các phép biến đổi
Data exploration & validation	Data profiling: hiển thị thông tin cơ bản về các feature như kiểu dữ liệu, tỉ lệ missing value, phân bố dữ liệu, các con số thống kê như min, max, mean, .v.v.
	Visualization: xây dựng các dashboard về phân bố hoặc độ skew của dữ liệu
	Validation: sử dụng các user-defined rule (ví dụ như tỉ lệ missing value < 80%), hoặc dựa vào thống kê để xác định độ lệch phân bố

Feature Store



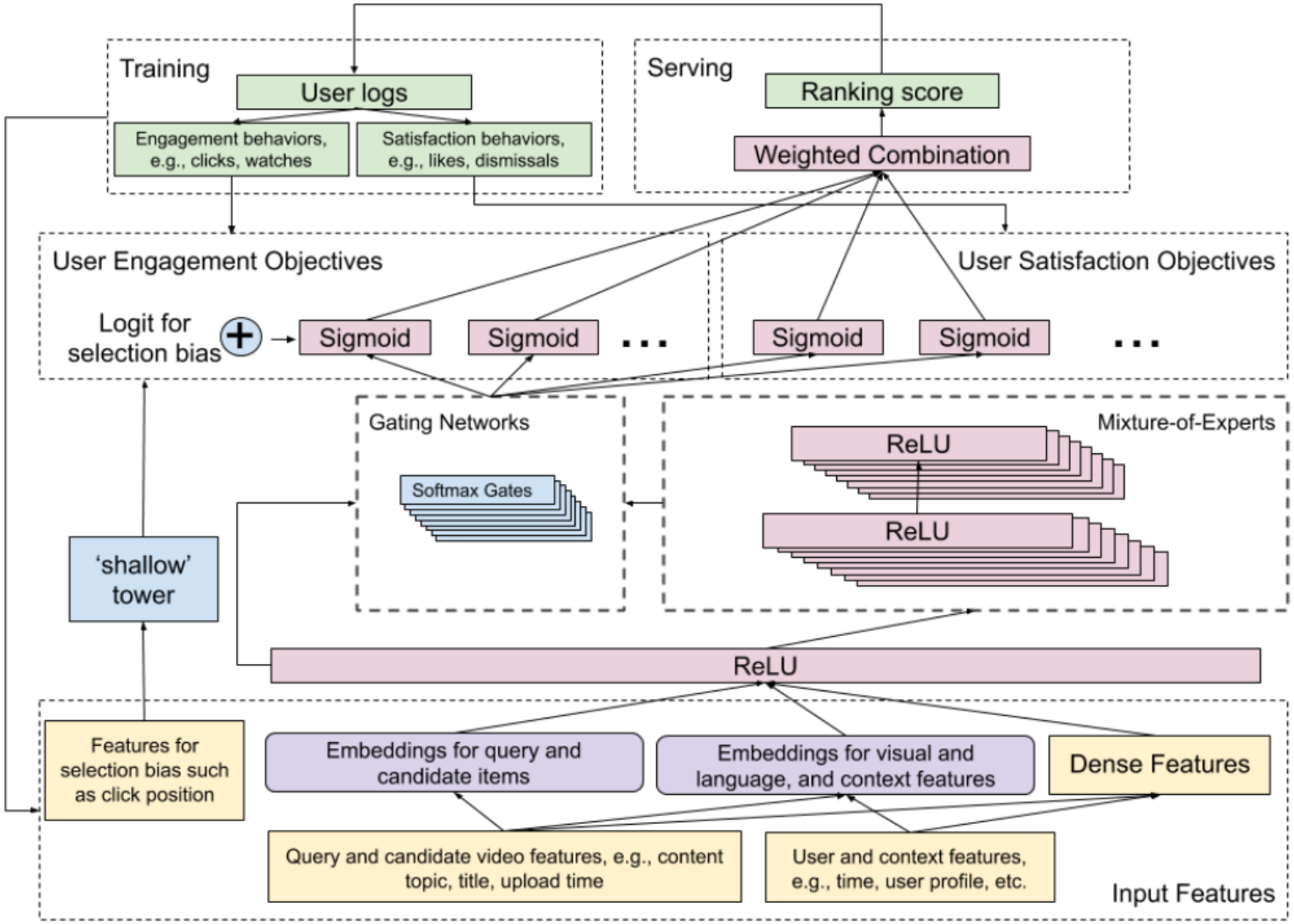
https://www.tecton.ai/blog/what-is-a-feature-store

ML pipeline


Feature Store
Training Data Retrieval
Model Training
• Data leakage
• Feature leakage
• Reproducibility
Model Evaluation
• ML metrics
• Bias/Fairness
• Output consistency
• Output shape & value range
Model Registration
Model Registry
Decide production model
Serving Data Retrieval
Model Serving
Model in production mode

Mlops tool stack

Sample of proposed ranking system



viettel

Thanks for your attention!