**PROJECT: General Search (Free Text Healthcare Provider Search)**

**Background / Problem Statement**

* Healthcare Provider Search was manual, rigid, and keyword-based—leading to poor patient experience.
* Existing search systems lacked semantic understanding, especially for unstructured, free-text patient queries.
* Absence of MLOps pipelines led to challenges in:
  + Model versioning and governance.
  + Consistent data pipelines across environments.
  + Continuous deployment and monitoring for model drift.

**UST Solution (Built Using AWS SageMaker MLOps AI Workbench)**

* Designed and deployed RoBERTa-based Transformer models for **free-text understanding** of patient queries.
* **Semantic NLP models** enabled contextual, intent-driven provider recommendations.
* AWS SageMaker MLOps AI Workbench powered the entire lifecycle using:
  + **Data Pipeline**: S3 + SageMaker Ground Truth + Clarify (for bias checks), DataOps, Data Versioning.
  + **CI/CD Driven Training Pipeline**: SageMaker Pipelines + GitHub + EventBridge.
  + **CI/CD Driven Inference Pipeline**: Load-balanced model hosting using VPC + IAM-secured endpoints.
  + **Model Monitoring Pipeline**: Real-time drift tracking with EventBridge triggers, auto-alerts, CloudWatch metrics.

**Business Impact**

* **60% improvement** in query relevance and accuracy for free-text searches.
* Enhanced patient satisfaction due to faster, more personalized provider matching.
* Seamless operationalization of models with faster experimentation-to-deployment cycle (~40% reduction in model release time).

**Cost Impact**

* 35% reduction in manual intervention costs through automated SageMaker pipelines.
* Scalable MLOps architecture reduced infrastructure sprawl and minimized model re-training overheads.
* Unified deployment + monitoring reduced maintenance cost by 25%.

**AWS SageMaker MLOps Benefits**

* Standardized CI/CD templates ensured **repeatable and governed deployment**.
* Version-controlled data/model lineage helped with regulatory compliance (HIPAA).

Event-driven retraining and pipeline orchestration led to **resilient production ML workflows**.