Improving Population Health

Group – Apache Doctors : Yanchen Dong, Vanessa Atwood, Liang Gong, Saphir Qi

March 2024

Business Problem & Objectives

Goal: Develop a product solution for Patient Health Monitoring and Recommendations System

The Health Insurance Group (HIG) has the Objective to develop an **AI solution** capable of **monitoring patient health records in real-time**, to provide proactive **recommendations** to patients on **lifestyle changes, doctor and ultimately Emergency Room visits**.

To fulfill this goal, our data Science team will **construct** a longitudinal patient record (**LPR**), by **aggregating data on prior medical visits, lab results, prescription data, and IoT devices** such as consumer wearables, smart-watches, and home diagnostic equipment.

Recommendation on Data Infrastructure: Data Ingestion and Storage

Utilize open-source technologies such as Apache Kafka for **real-time data ingestion** from various sources including EMRs, pharmacies, labs, and IoT devices.

Store the ingested data in a scalable and cost-effective manner using open-source distributed storage solutions like Apache Hadoop Distributed File System (HDFS)

Implement data governance and access controls to ensure compliance with HIPAA standards and patient consent requirements.

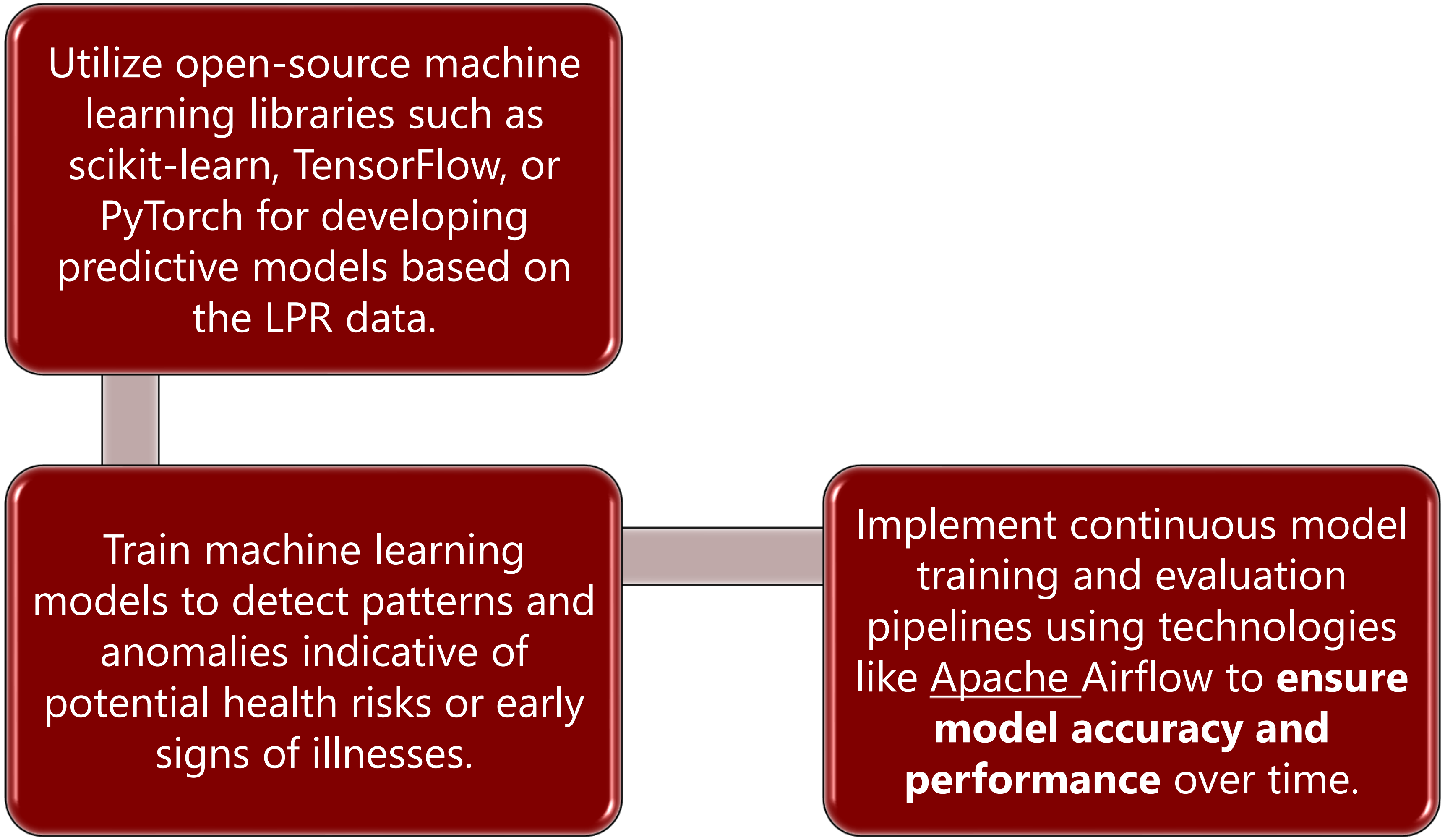
Recommendation on Data Infrastructure: Data Processing and Analysis

Use Apache Spark for **distributed data processing and analytics**, enabling efficient processing of large-scale datasets and complex machine learning algorithms.

Leverage Apache Hive for **SQL-based querying and analysis of structured data** stored in the distributed storage system.

Implement data quality checks and data cleansing techniques to ensure accuracy and reliability of the LPR data.

Recommendation on Data Infrastructure: Machine Learning Model Development

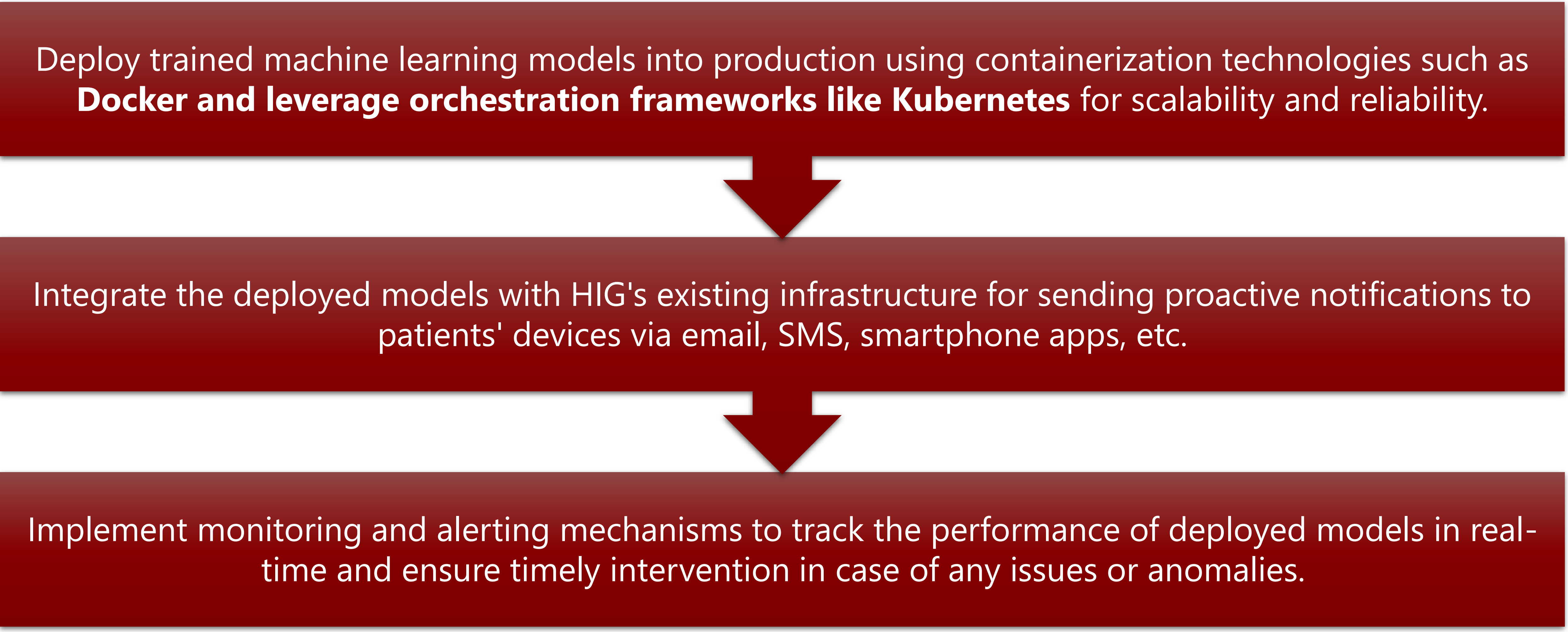


Utilize open-source machine learning libraries such as scikit-learn, TensorFlow, or PyTorch for developing predictive models based on the LPR data.

Train machine learning models to detect patterns and anomalies indicative of potential health risks or early signs of illnesses.

Implement continuous model training and evaluation pipelines using technologies like Apache Airflow to **ensure model accuracy and performance** over time.

Recommendation on Data Infrastructure: Model Deployment and Integration



Deploy trained machine learning models into production using containerization technologies such as **Docker and leverage orchestration frameworks like Kubernetes** for scalability and reliability.

Integrate the deployed models with HIG's existing infrastructure for sending proactive notifications to patients' devices via email, SMS, smartphone apps, etc.

Implement monitoring and alerting mechanisms to track the performance of deployed models in real-time and ensure timely intervention in case of any issues or anomalies.

Recommendation on Data Infrastructure: Data Visualization and Reporting

Implement open-source tools such as Apache Superset or Redash to **query, visualize, and create custom charts and dashboards**. This helps HIG team in understanding trends, patterns, and making informed decisions.

Recommended Key Features

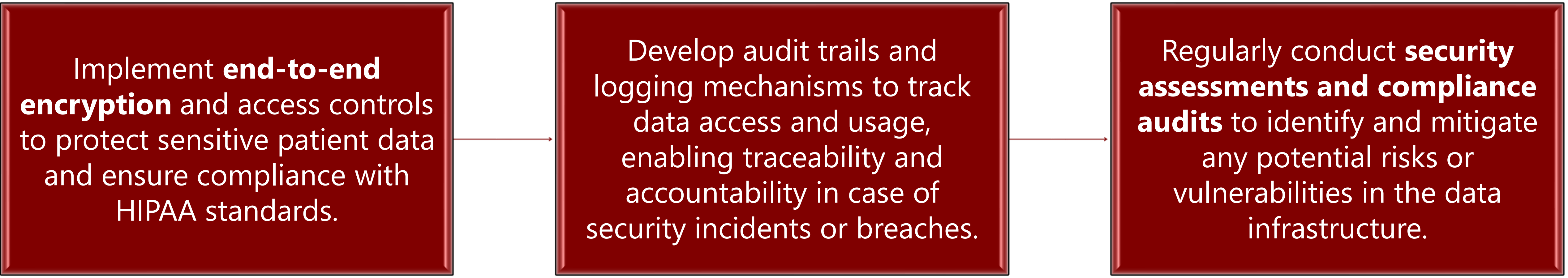
Customizable Dashboards: Allow users to customize dashboards based on their role (e.g., healthcare providers, HIG analysts, patients) and their specific needs or interests.

Interactive Reports:

Enable **dynamic interaction with the data**, such as drilling down into specific metrics, filtering, or slicing data to view different dimensions.

Alerts and Notifications: Integrate alerting mechanisms that can notify stakeholders about **critical health trends** or **anomalies** detected by the AI models.

Recommendation on Data Infrastructure: Compliance and Security



Implement **end-to-end encryption** and access controls to protect sensitive patient data and ensure compliance with HIPAA standards.

Develop audit trails and logging mechanisms to track data access and usage, enabling traceability and accountability in case of security incidents or breaches.

Regularly conduct **security assessments and compliance audits** to identify and mitigate any potential risks or vulnerabilities in the data infrastructure.

Recommendation on Data Infrastructure: Proposed Infrastructure is "Big Data"

The proposed infrastructure can be referred to as "Big Data" as it involves the processing and analysis of large volumes of **heterogeneous data** sources including electronic medical records, IoT device data, and other healthcare-related data.

The use of distributed storage and processing technologies like Apache Hadoop and Spark enables scalability and performance optimization for handling big data workloads effectively.

Additionally, the incorporation of machine learning algorithms for predictive analytics adds another dimension to the complexity and scale of data processing involved, further justifying the characterization of the infrastructure as "Big Data".