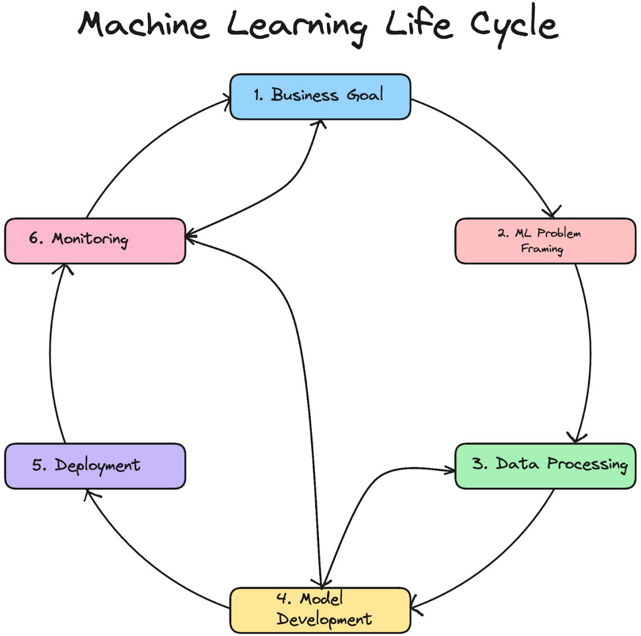
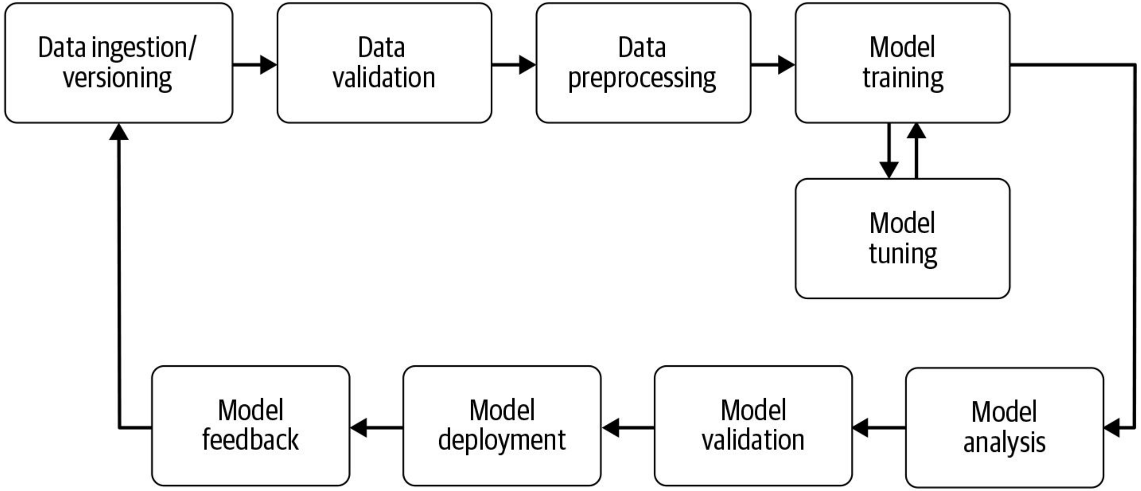
Machine Learning Model Deployment

Introduction to ML Pipeline

What is Machine Learning Pipeline?





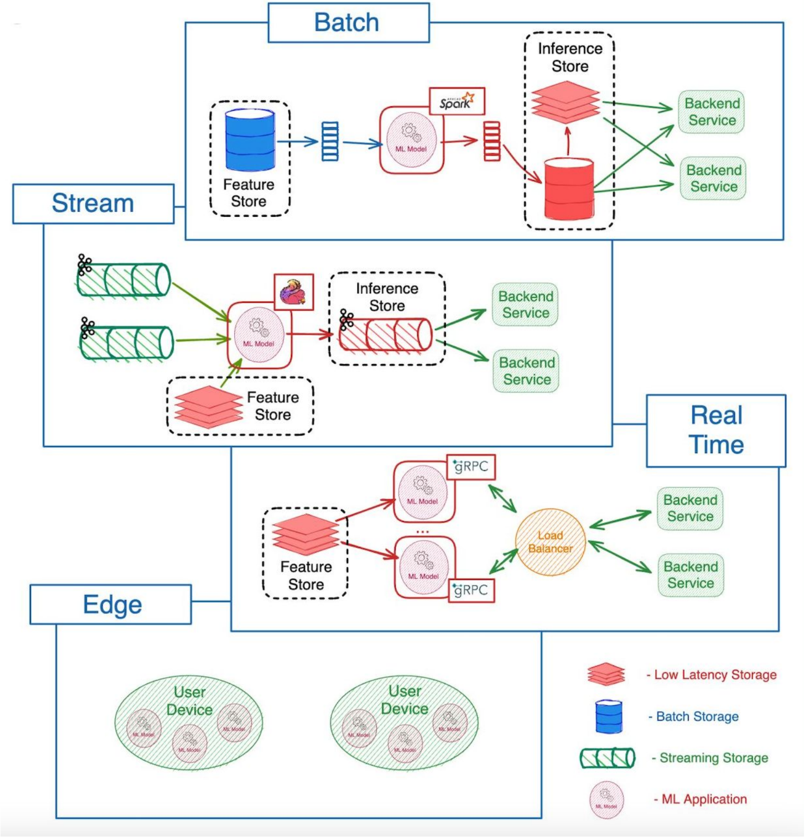
Type of ML Deployment

➔ **Batch:** In batch deployment, ML models process large volumes of data at scheduled intervals, ideal for tasks like end-of-day reporting or monthly analytics.

➔ **Stream:** Stream deployment enables ML models to process and analyze data in real-time as it flows in, suitable for applications like fraud detection or live social media analysis.

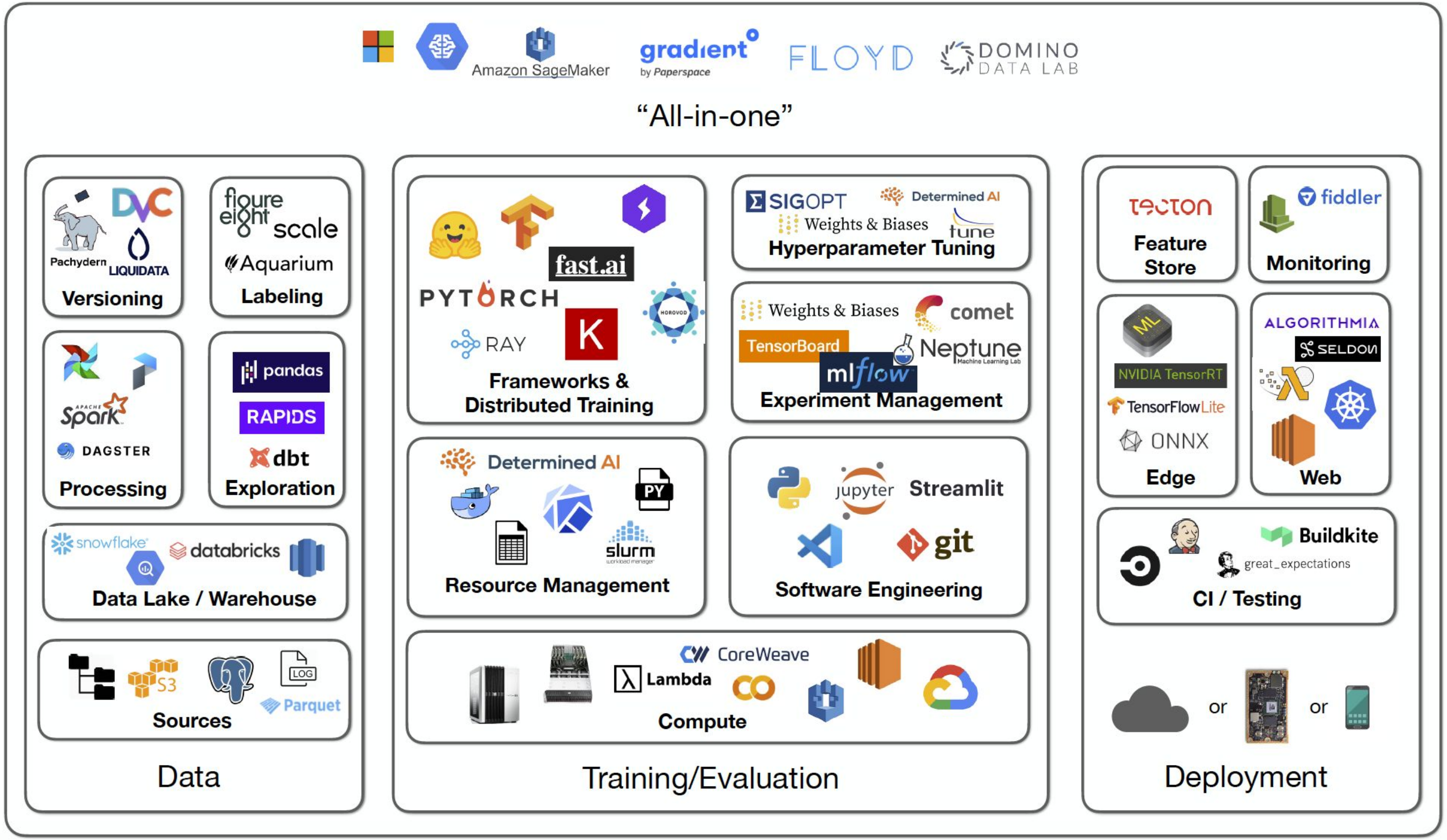
➔ **Realtime:** Realtime deployment allows ML models to provide instant predictions or decisions in response to incoming data, essential for use cases like recommendation systems or autonomous driving.

➔ **Edge:** Edge deployment involves running ML models on local devices close to the data source, reducing latency and bandwidth usage, which is crucial for IoT applications and smart devices.



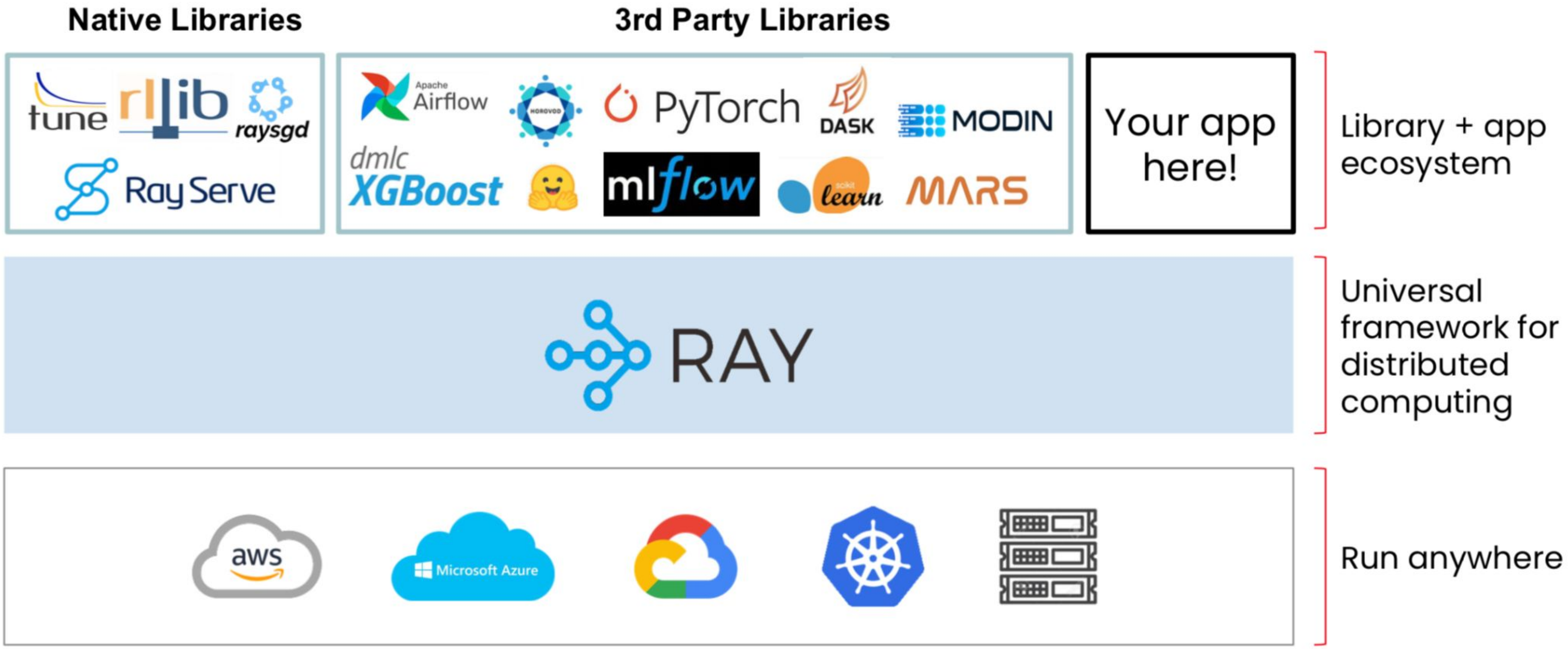
Infrastructure and Integration

- **Hardware and Software:** Setting up the right environment for model deployment.
- **Integration:** Seamlessly integrating the model with existing systems and applications.



Benefits of Deploying ML Models

Focus on new models, not maintaining existing models || **Prevention** of bugs || **Creation** of records for debugging and reproducing results || **Standardization ||** **Allows models** to handle real-time data and large user bases.



Native Libraries

3rd Party Libraries

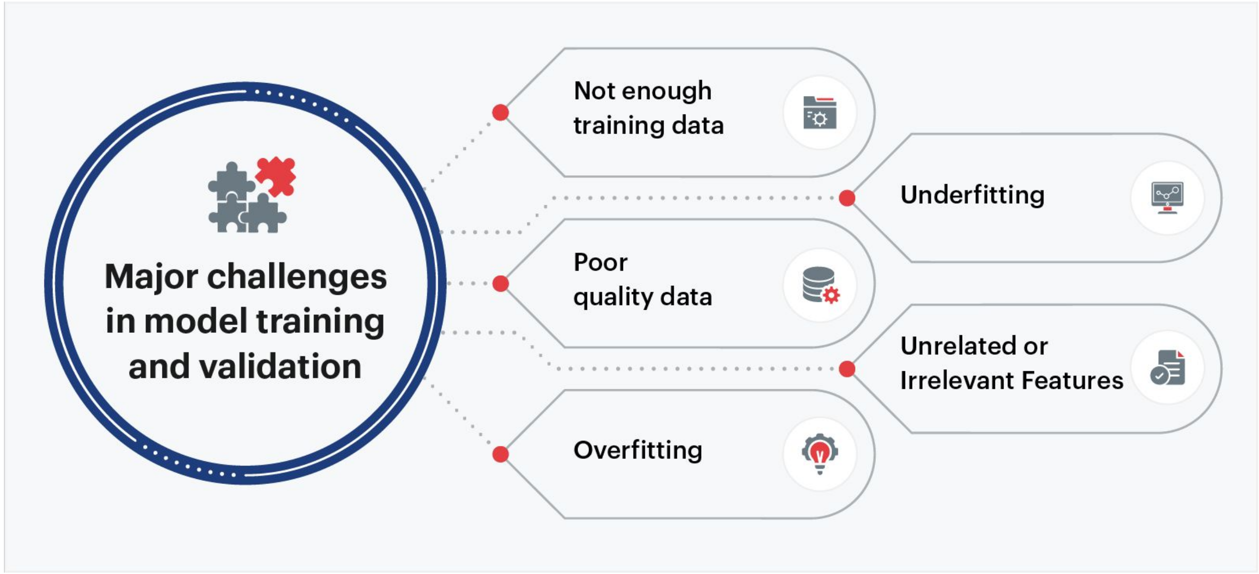
Your app here!

Library + app ecosystem

Universal framework for distributed computing

Run anywhere

Challenges in ML Deployment

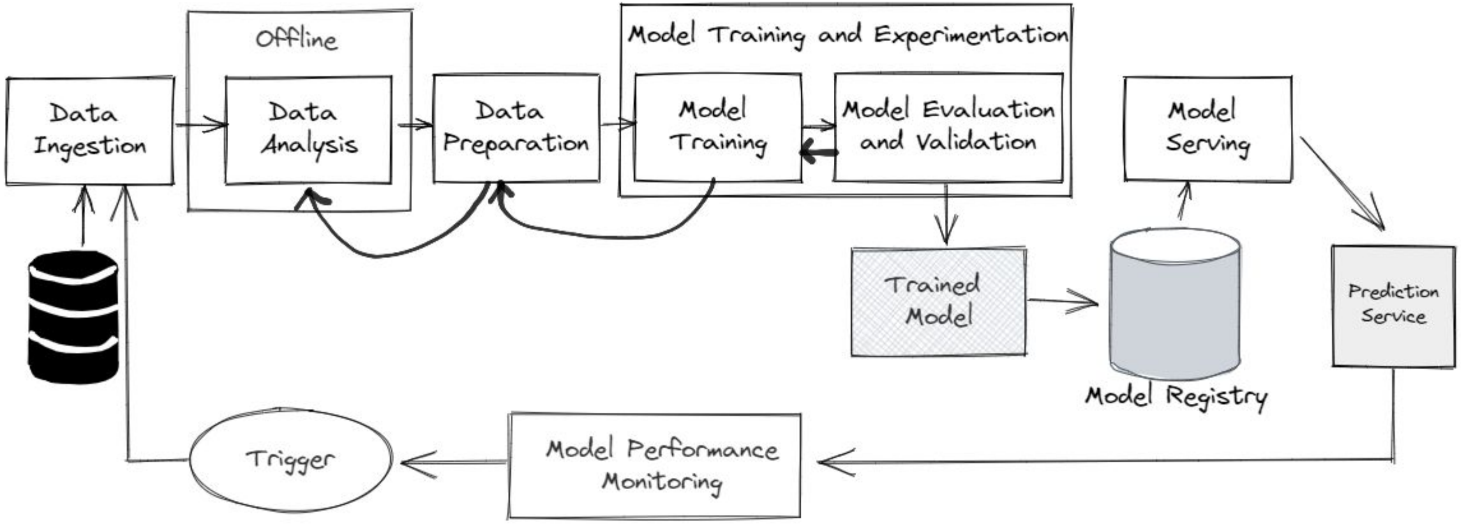

Major challenges in model training and validation
Not enough training data
Underfitting
Poor quality data
Unrelated or Irrelevant Features
Overfitting

As per research, only 13% of ML models ever make it to production. This is a huge gap, considering the possibilities that AI model deployment can bring to the organization.

- Data Management: Making sure the model gets the right kind of data.
- Model Scalability and Performance: Ensuring that their model can effectively scale as it keeps adding more complex information.
- Integration with Existing Systems: Fitting the model into current computers and software.
- Monitoring and Maintenance: Watching and fixing the model over time.
- Security and Privacy: Protecting data and keeping it private.
- Resource Management: Using computer resources like memory and power wisely.
- Versioning and Model Management: Keeping track of different versions of the model.
- Regulatory Compliance: Making sure the model follows the laws, rules, and regulations.
- User Acceptance and Trust: Getting people to trust and accept the model.
- Explainability and Transparency: Being able to explain how the model works.
- Cost Management: Managing how much it costs to use the model.

Data and Model Management

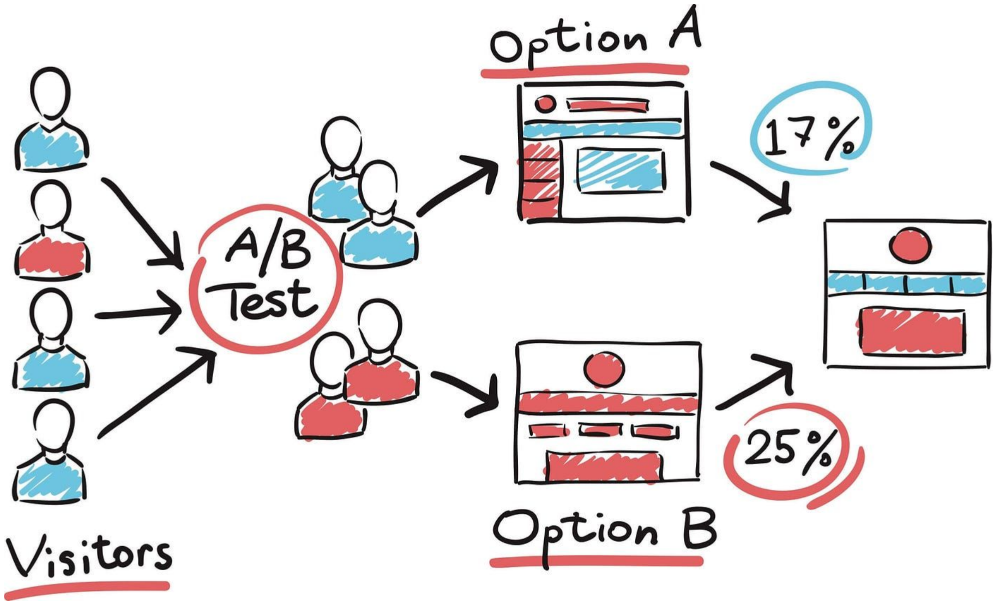
Data Pipelines: Building and maintaining data pipelines for continuous data flow.
Model Versioning: Tracking and managing different versions of models.



Amazon S3

A/B Testing

- **Objective Comparison:** A/B testing allows for an objective comparison of two model versions to determine which performs better based on specific metrics.
- **Real-World Application:** It is widely used to optimize user experiences, such as testing different recommendation systems or ad strategies to enhance engagement or conversion rates.
- **Statistical Significance:** The technique ensures that performance differences are statistically significant and not due to random chance by using control and treatment groups along with statistical tests.



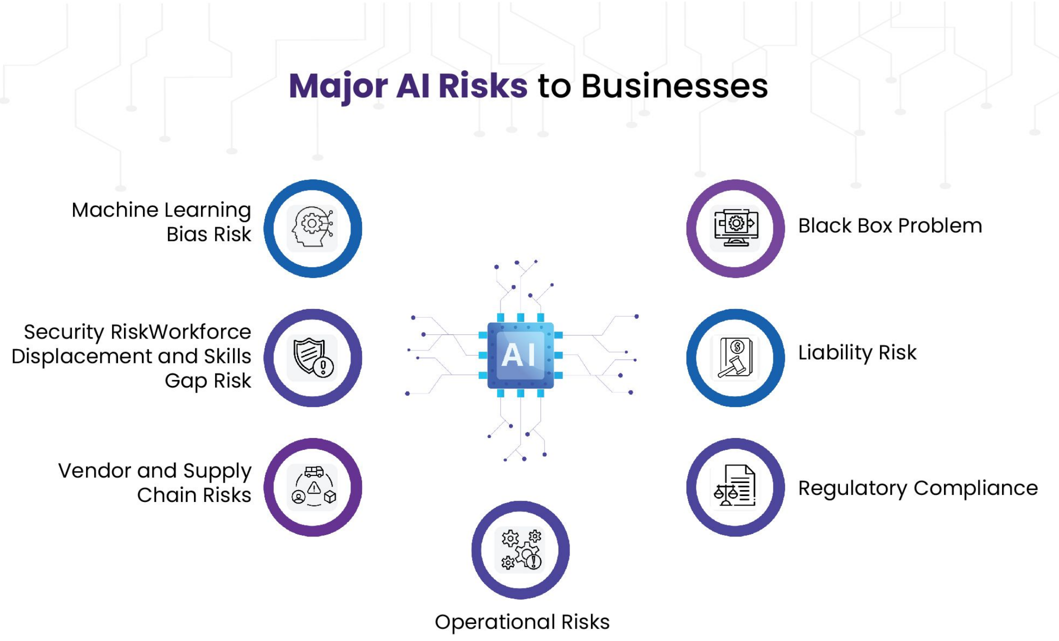
Security, Compliance and Bias

Security: Ensuring the security of machine learning models involves protecting sensitive data from unauthorized access and breaches through robust encryption, secure APIs, and access controls

Compliance: Adhering to industry regulations and standards, such as GDPR or HIPAA, is critical to ensure the legal and ethical use of data in machine learning deployments. This involves data anonymization, user consent, and regular compliance audits.

Bias Detection: Identifying and mitigating bias in ML models is crucial to prevent unfair and discriminatory outcomes. This involves using diverse training datasets, applying fairness-aware algorithms, and conducting bias impact assessments

Continuous Monitoring: Regular monitoring and updating of deployed models are essential to maintain security, compliance and fairness. This involves real-time performance tracking, automated alerts for anomalies, and periodic model retraining.



Major AI Risks to Businesses

Machine Learning Bias Risk

Security RiskWorkforce Displacement and Skills Gap Risk

Vendor and Supply Chain Risks

Operational Risks

Black Box Problem

Liability Risk

Regulatory Compliance