Agenda

09:00 Welcome and Introductions
09:15 Data Science Lifecycle
10:30 Lab 1: Understanding and preparing data with S3, Glue and Athena
11:15 Lunch
12:00 Model training, testing and deploying with Sagemaker
12:45 Lab 2: Train, test and deploy your first model with Sagemaker
13:45 Break
14:00 Continuous Delivery of ML models
14:30 Lab 3: Continuous Delivery of ML models to Amazon SageMaker
15:15 Bring your own model
15:45 Wrap Up

Data Science Lifecycle

Machine Learning Immersion Day

Module 1

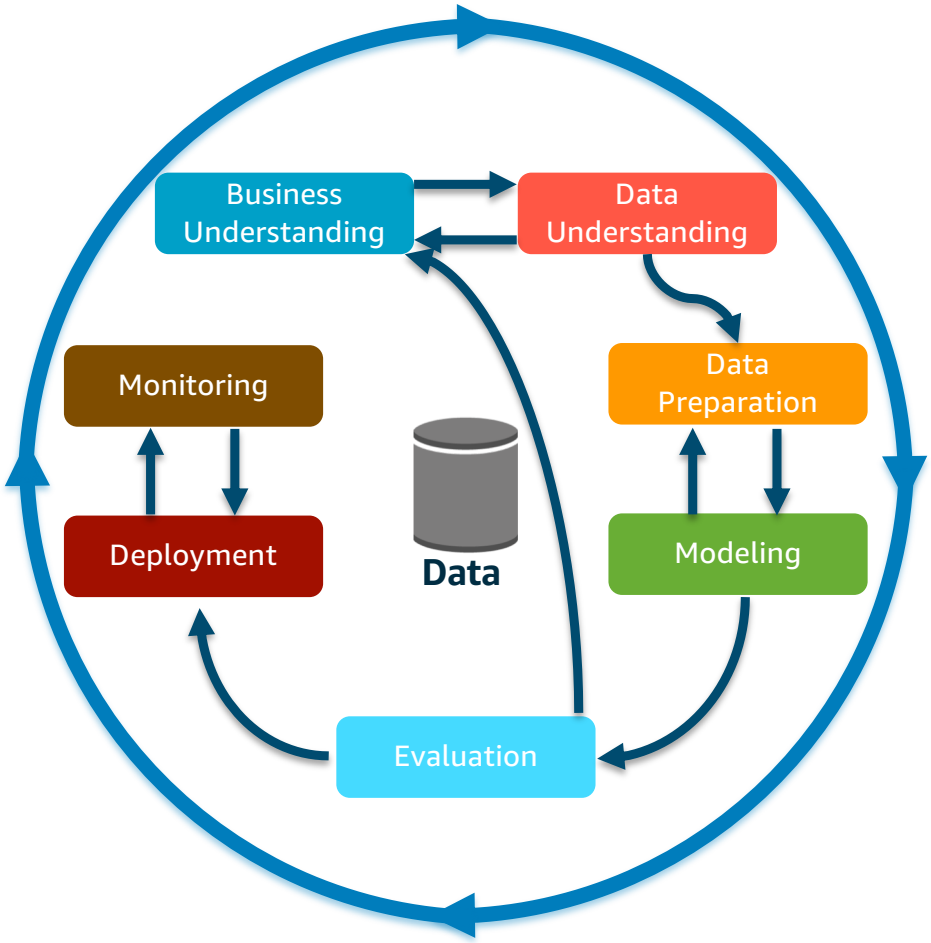
Machine Learning at AWS

Application Services	Use machine learning APIs without building and training your own model. Rekognition, Polly, Comprehend, Personalize, Forecast etc.
ML Platforms	SageMaker makes it easy for developers and data scientists to build, train and deploy models.
Frameworks	For the experts comfortable building and training their own models. AWS Deep Learning AMIs.

aws

CRISP-DM

- Cross Industry Standard Process for Data Mining
- Current de facto process for doing Data Science
- Highlights the cyclical and iterative natures of Data Science

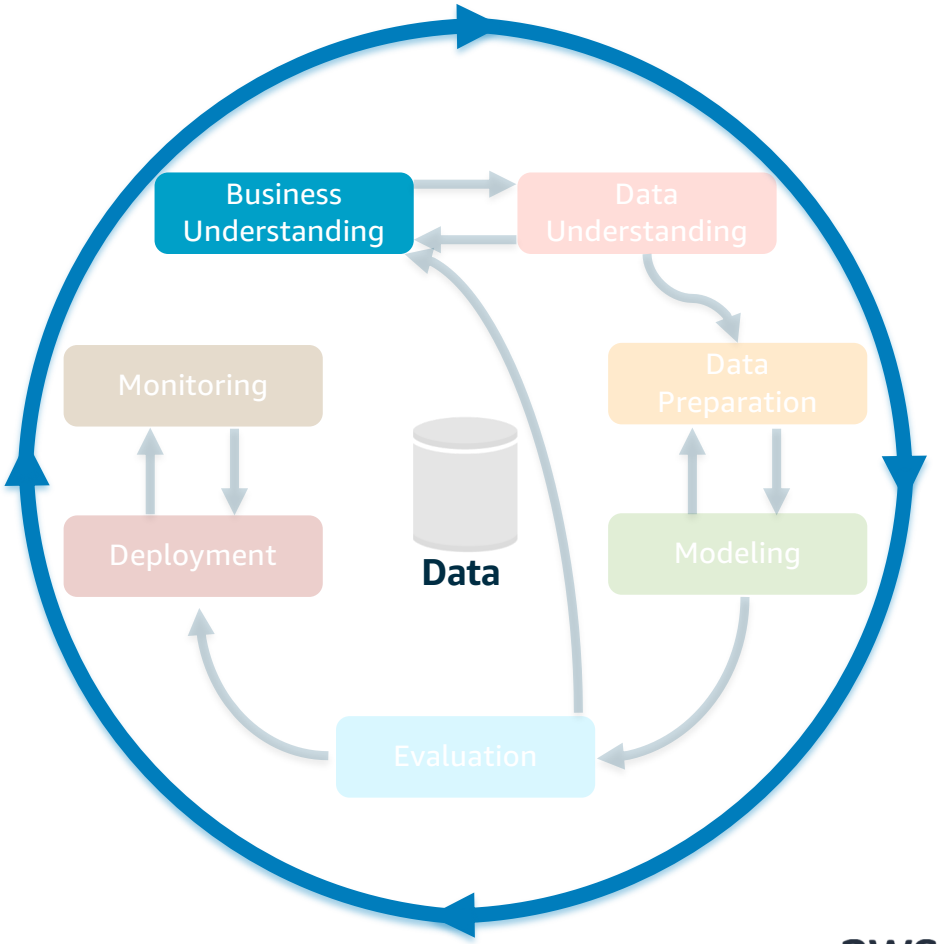


aws

Phase 1: Business Understanding

Don’t dive into the data immediately. First take some time to understand

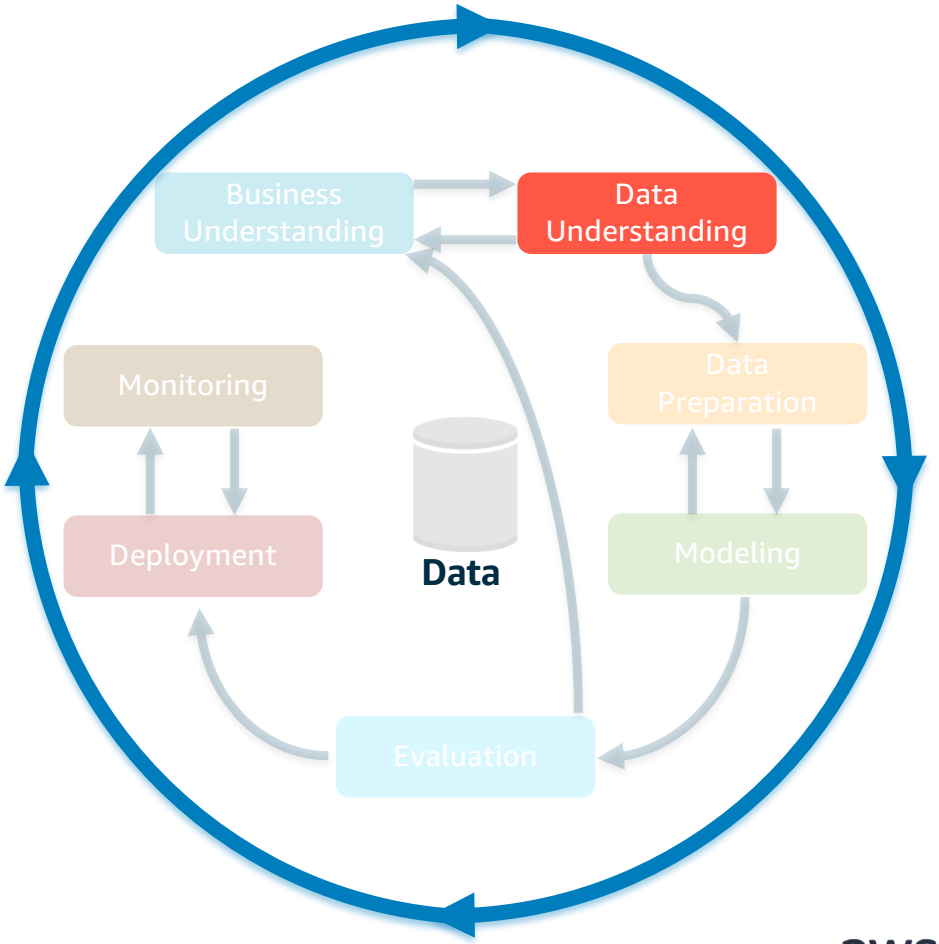
- Business objectives
- Surrounding context
- ML problem category
- Prelim. Project plan



Phase 2: Data Understanding

Exploring the data provides us with necessary information such as

- Data quality and cleanliness
- Interesting patterns in the data
- Likely paths forward once we start modeling

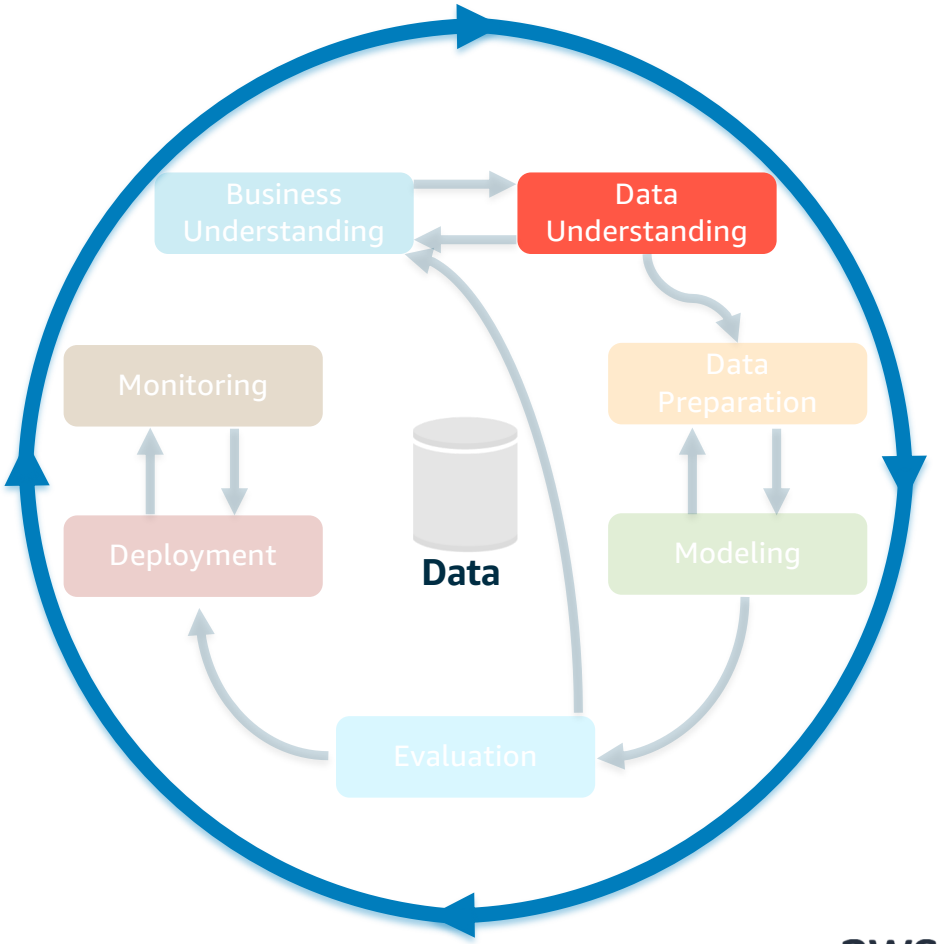


aws

Phase 2: Data Understanding

Tools

- Warehouse + SQL
- Python / R
- Local disk



Phase 2: Data Understanding

~~Tools~~

- ~~Warehouse + SQL~~
- ~~Python / R~~
- ~~Local disk~~

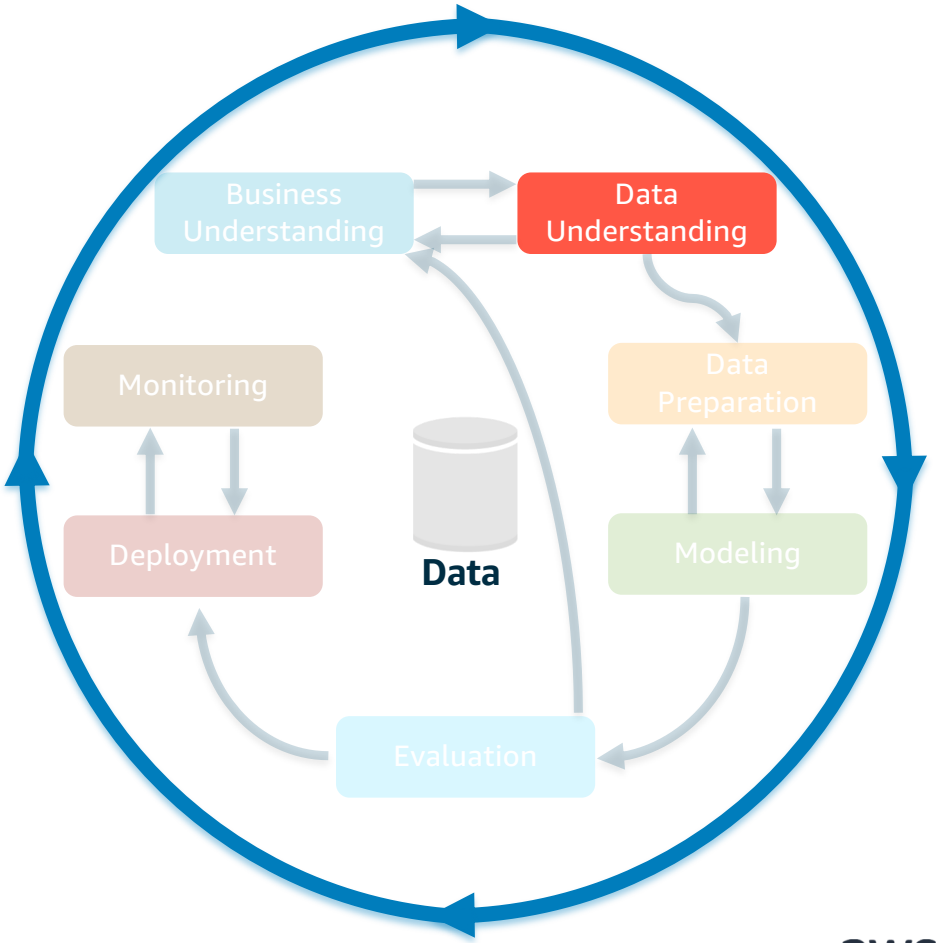
New Tools

SageMaker

Amazon Athena

Amazon QuickSight

AWS Glue



aws

Phase 3-4: Data Preparation & Modeling

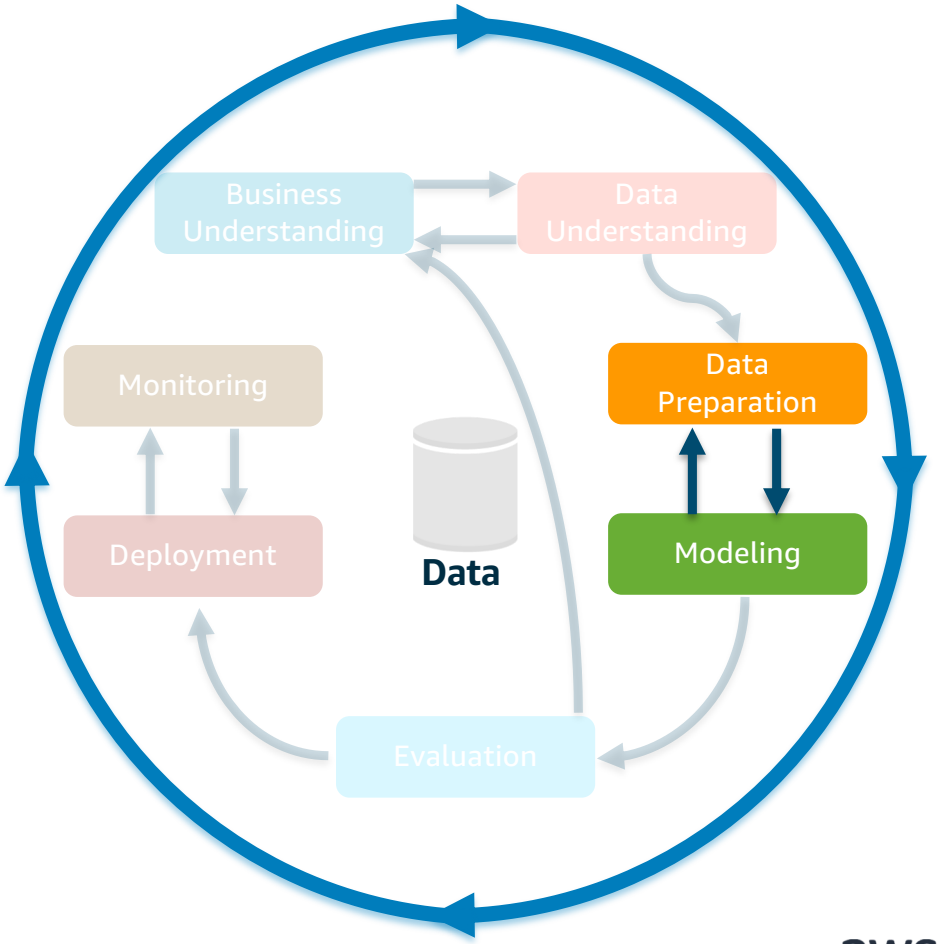
Data Preparation

- Normalization
- Feature selection - relevance
- Feature extraction - derived

Modeling

- Training a model & receiving an output

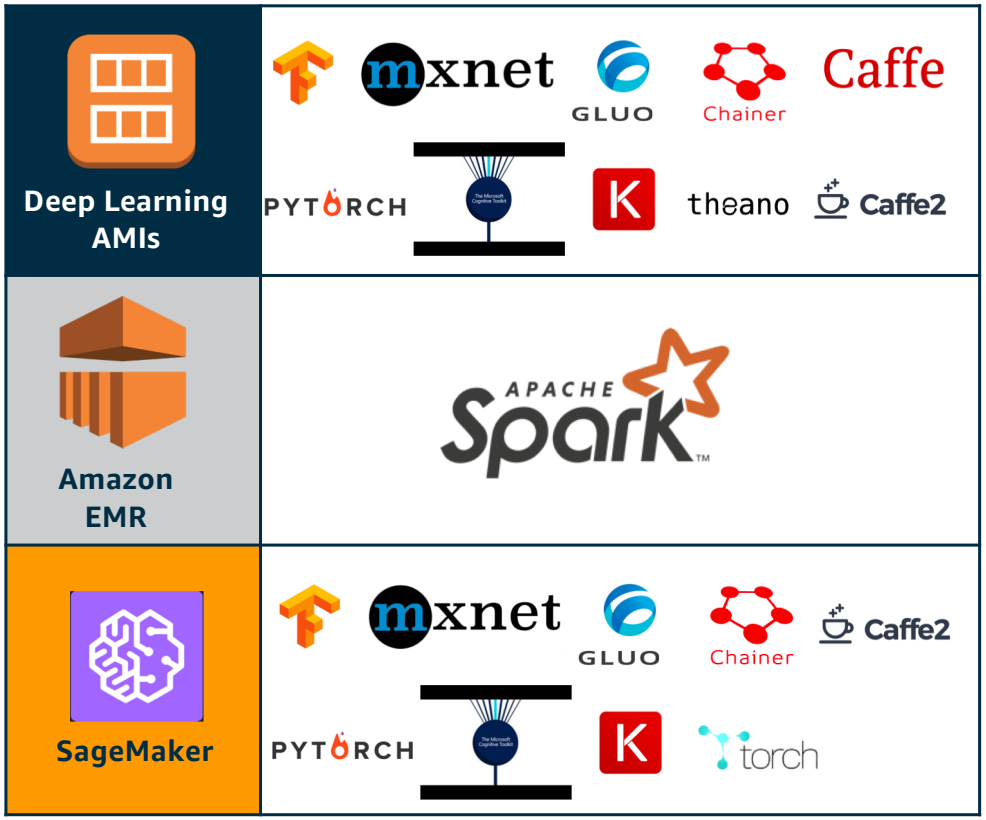
Notice that the two phases are interconnected.

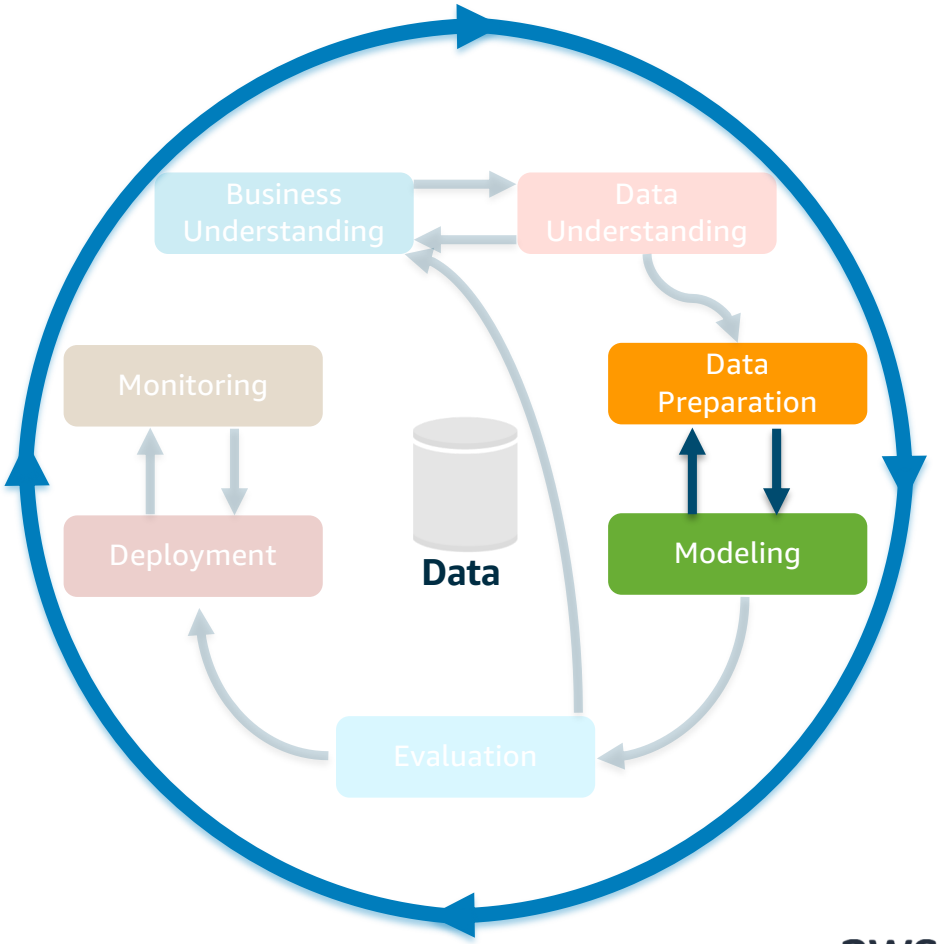


aws

Phase 3-4: Data Preparation & Modeling

Services and Tools



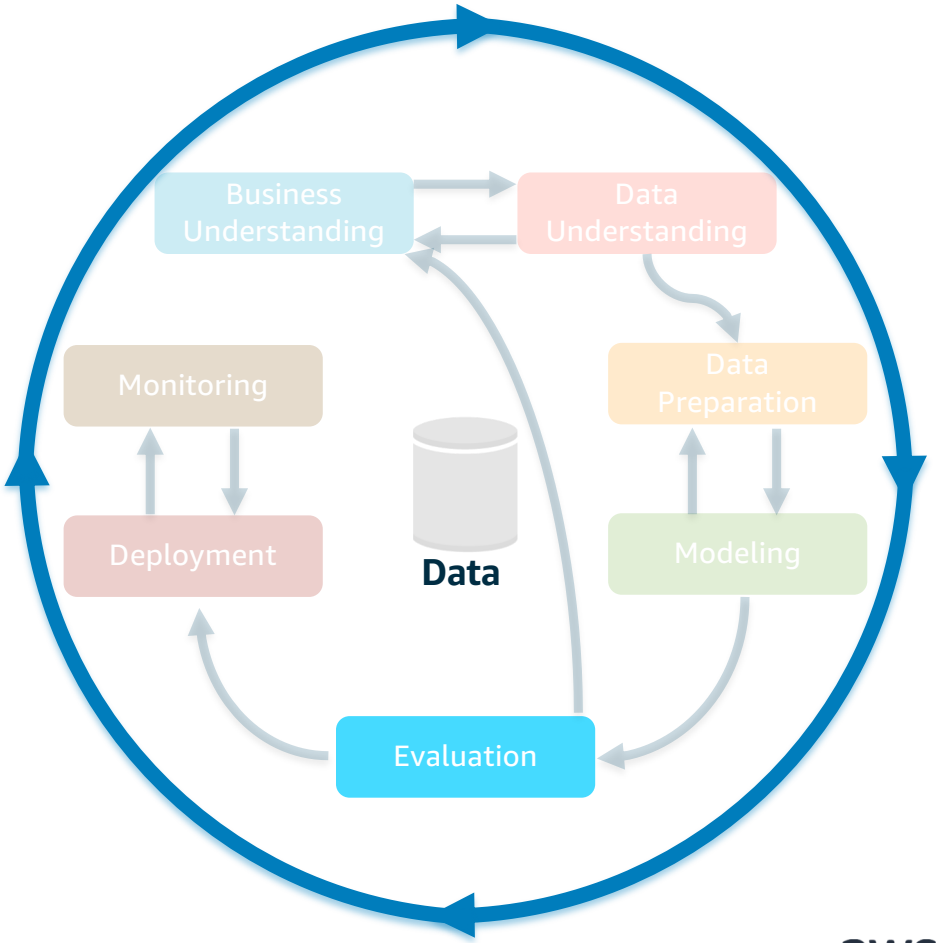


aws

Phase 5: Evaluation

Tie back into original business objectives:

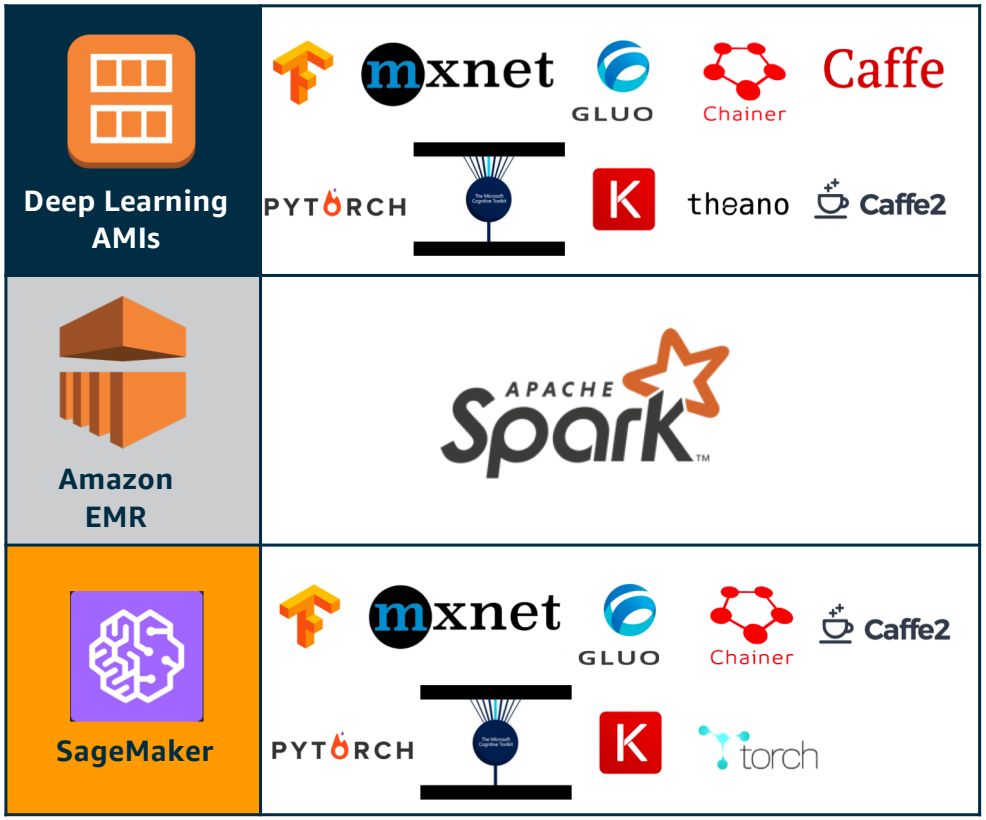
- False positive & false negative rates
- Performance acceptable?
- Actually deployable?
- Model fails =>
  - Are business objectives realistic?
  - Other data sources we can use?
  - What domain expertise can we leverage?

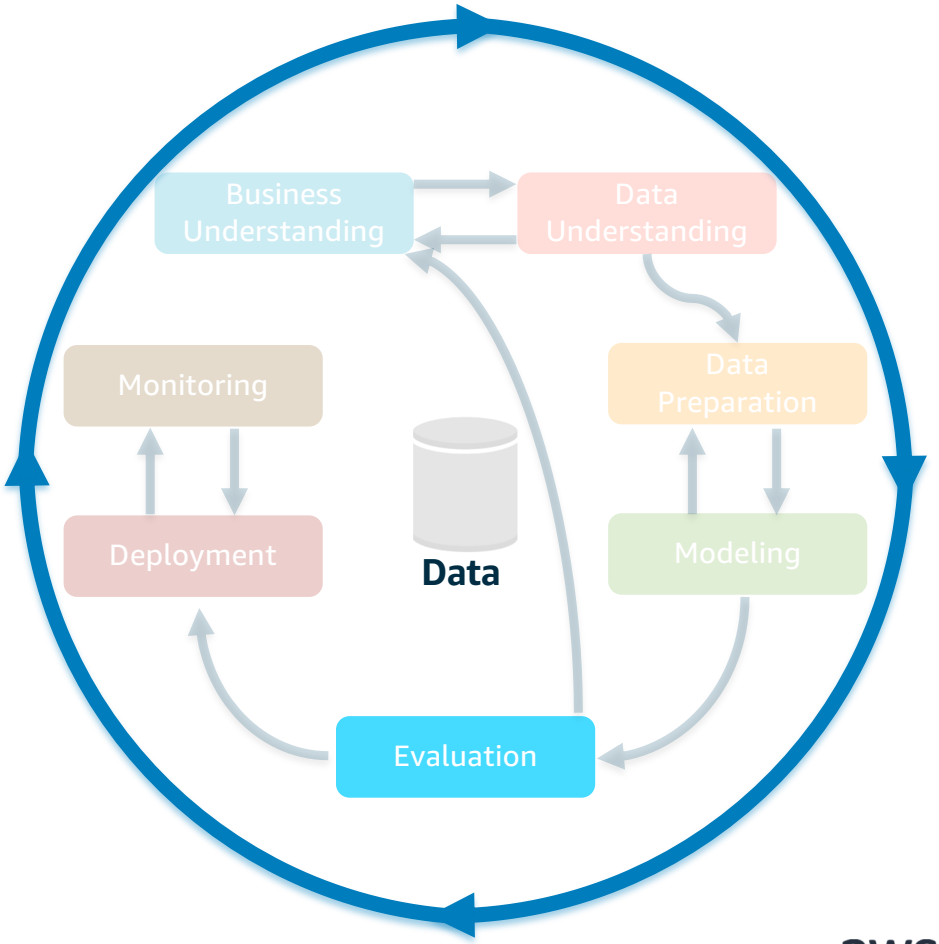


aws

Phase 5: Evaluation

Services and Tools





aws

Phase 6: Deployment

Services used

AWS Lambda

AWS Step Functions

AWS Batch

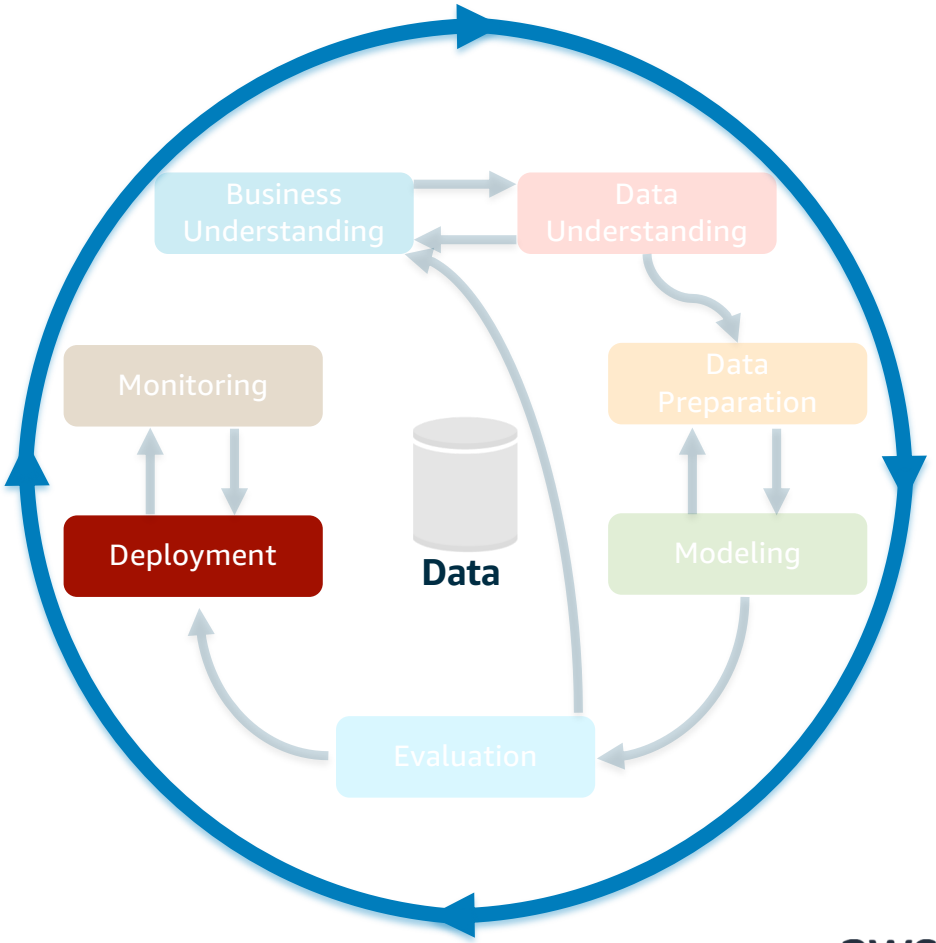
SageMaker

AWS CodeBuild

AWS CodeDeploy

AWS CodePipeline

AWS CloudFormation



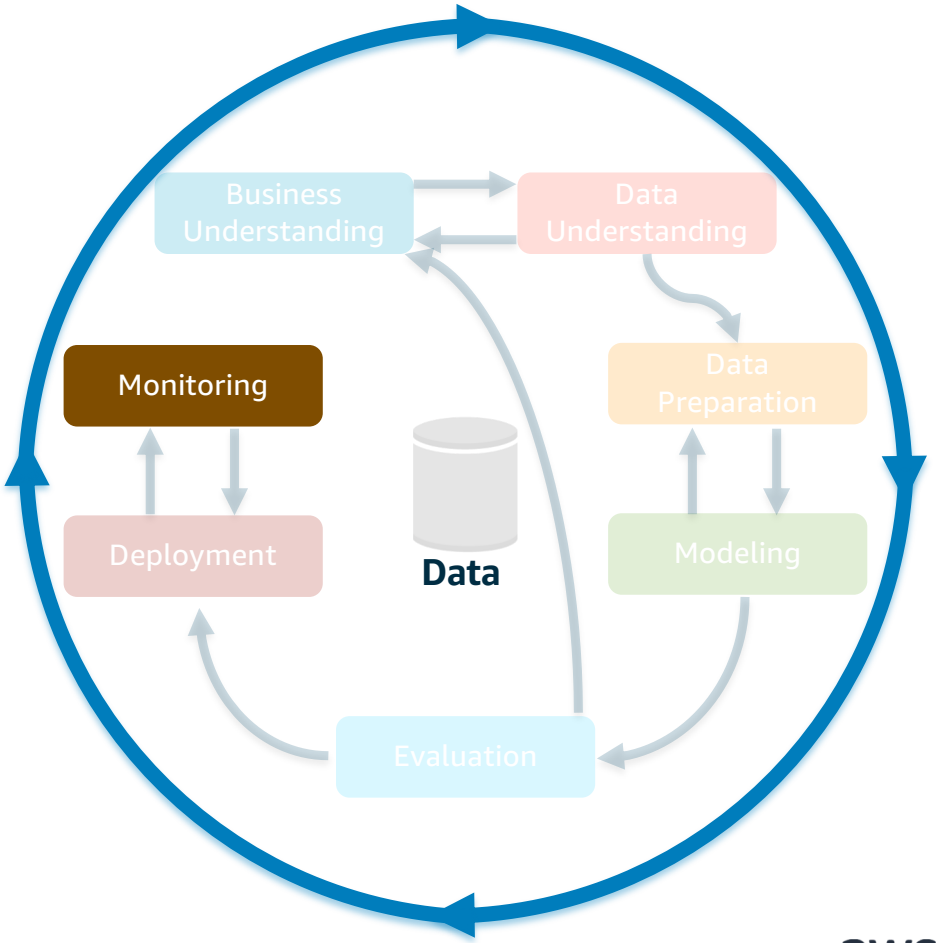
aws

Phase 7: Monitoring

Services used

Amazon CloudWatch

Amazon QuickSight



aws

Data Understanding & Preparation

aws

Phase 2-3: Data Understanding & Preparation - GLUE

What data do I have? What about ETL? GLUE

Data Catalog

Discover

Apache Hive Metastore compatible

Integrated with AWS services

Automatic crawling

Job Authoring

Develop

Auto-generates ETL code

Python and Apache

Spark

Edit, debug, and share

Job Execution

Deploy

Serverless execution

Flexible scheduling

Monitoring and alerting

aws

Lab 1: Understanding and preparing data with S3, Glue and Athena

Questions?