Fetching and Preparing Data

Amber Israelsen
AUTHOR | DEVELOPER | TRAINER

www.amberisraelsen.com

Course Overview

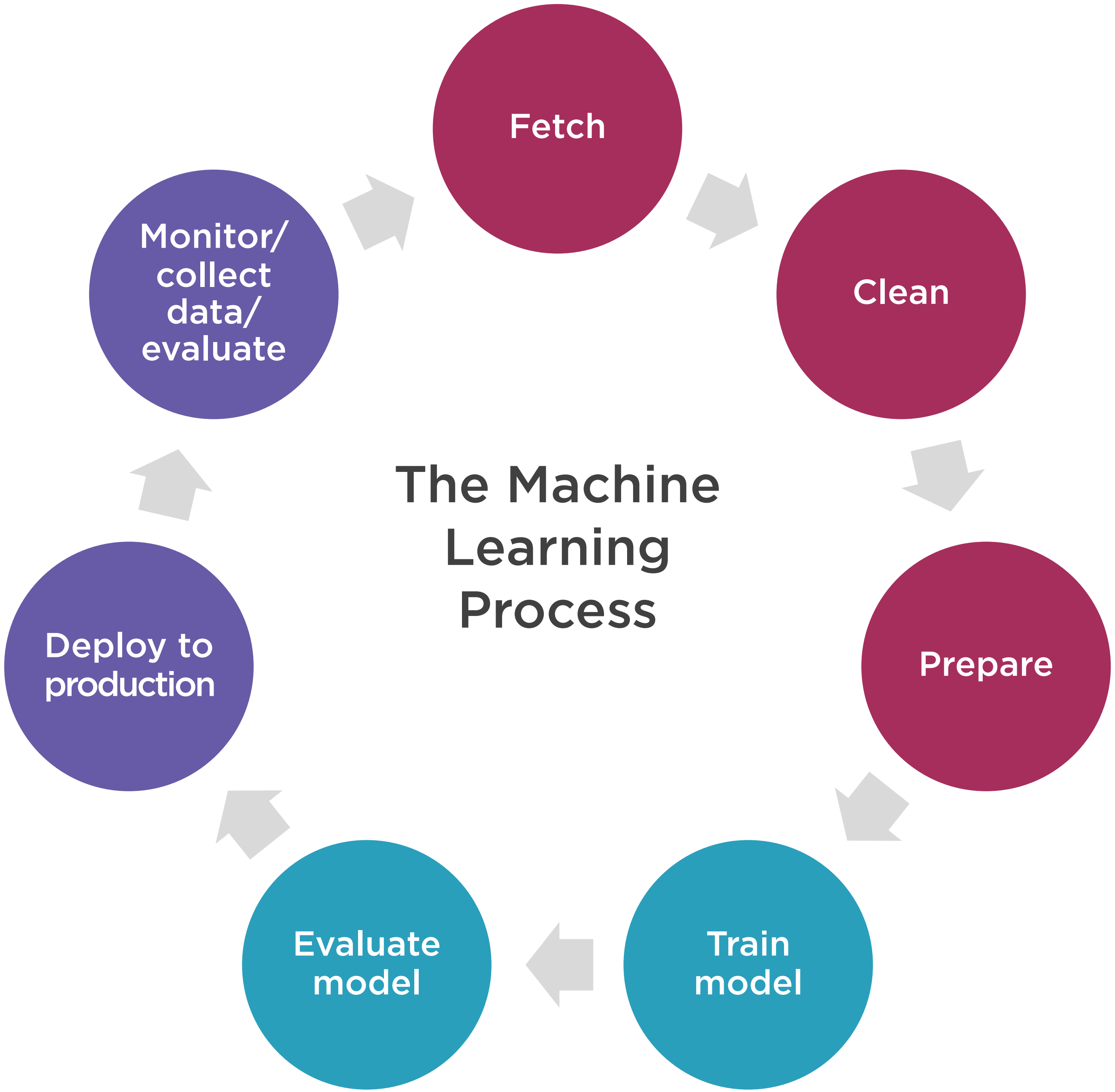
- Course Introduction
- Identifying Opportunities for Machine Learning
- Defining Machine Learning Problems
- Fetching and Preparing Data
- Training and Evaluating the Model
- Deploying and Monitoring the Model
- The AWS Machine Learning Stack
- Next Steps

Module Overview

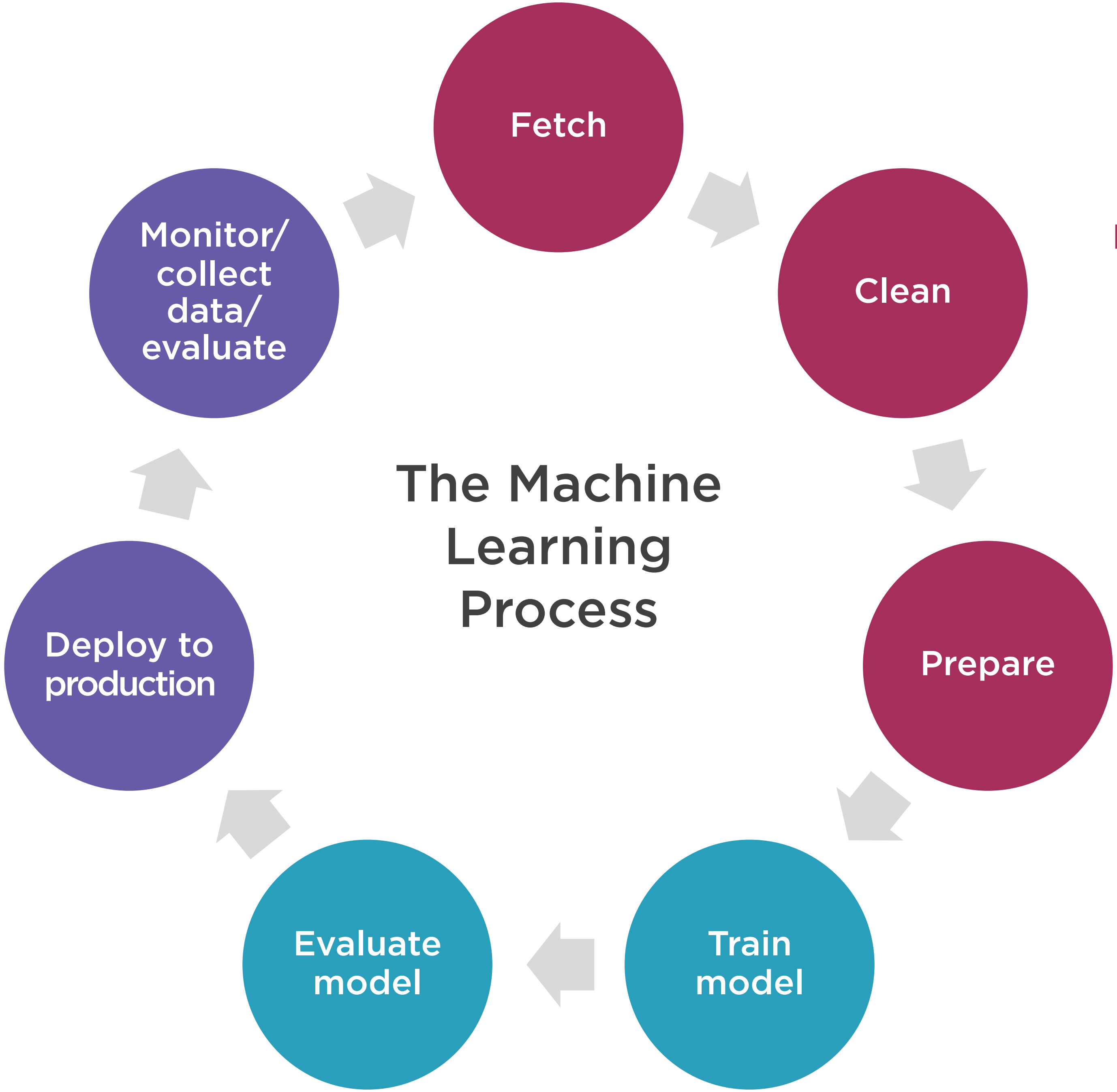
The machine learning process

- Overview
- Fetching data
  - AWS services
- Cleaning data
- Preparing data
  - Data visualizations
  - Feature engineering

Demo in SageMaker Studio


Fetch
Clean
Prepare
Train model
Evaluate model
Deploy to production
Monitor/ collect data/ evaluate
The Machine Learning Process

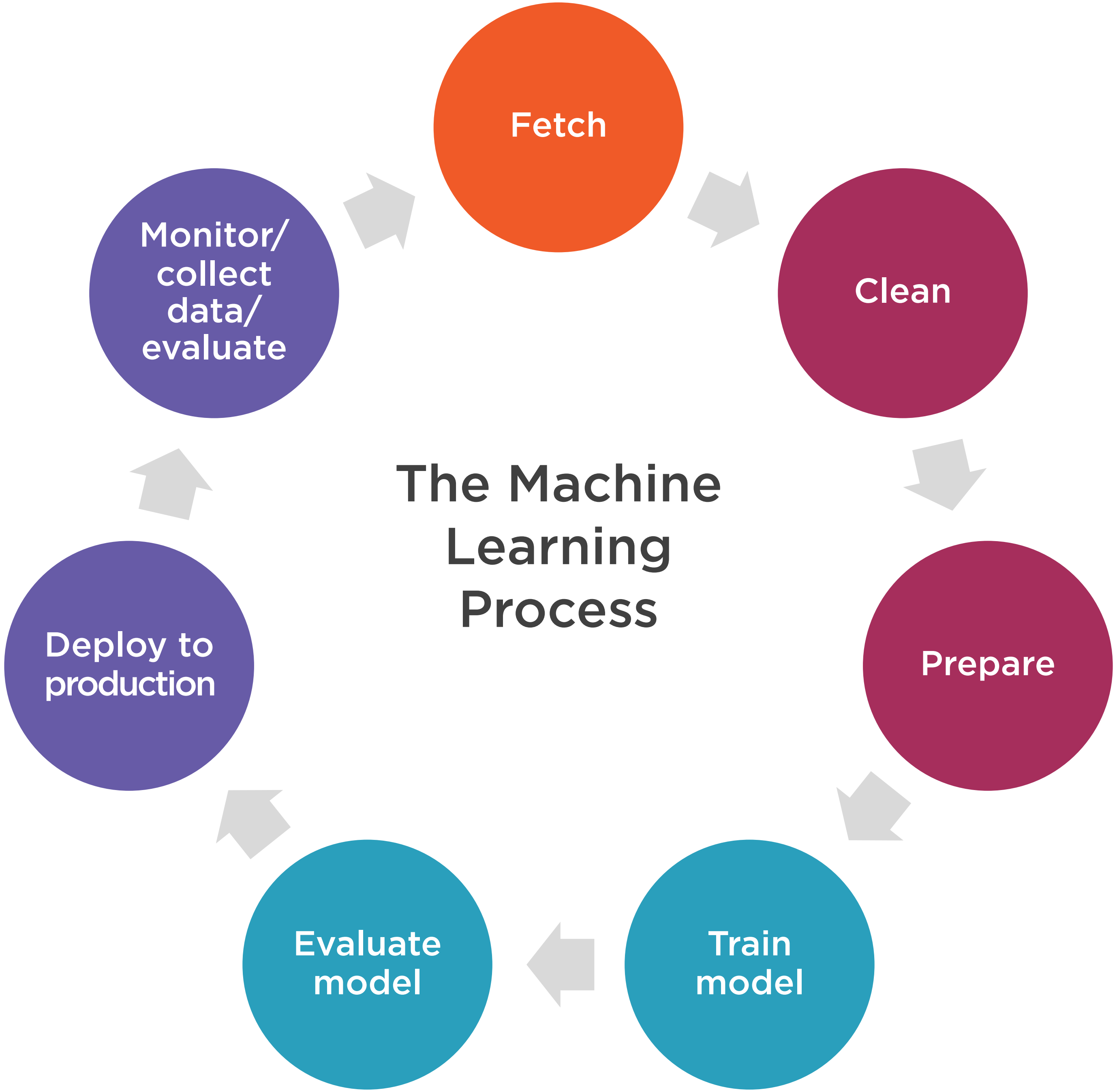
Fetching Data



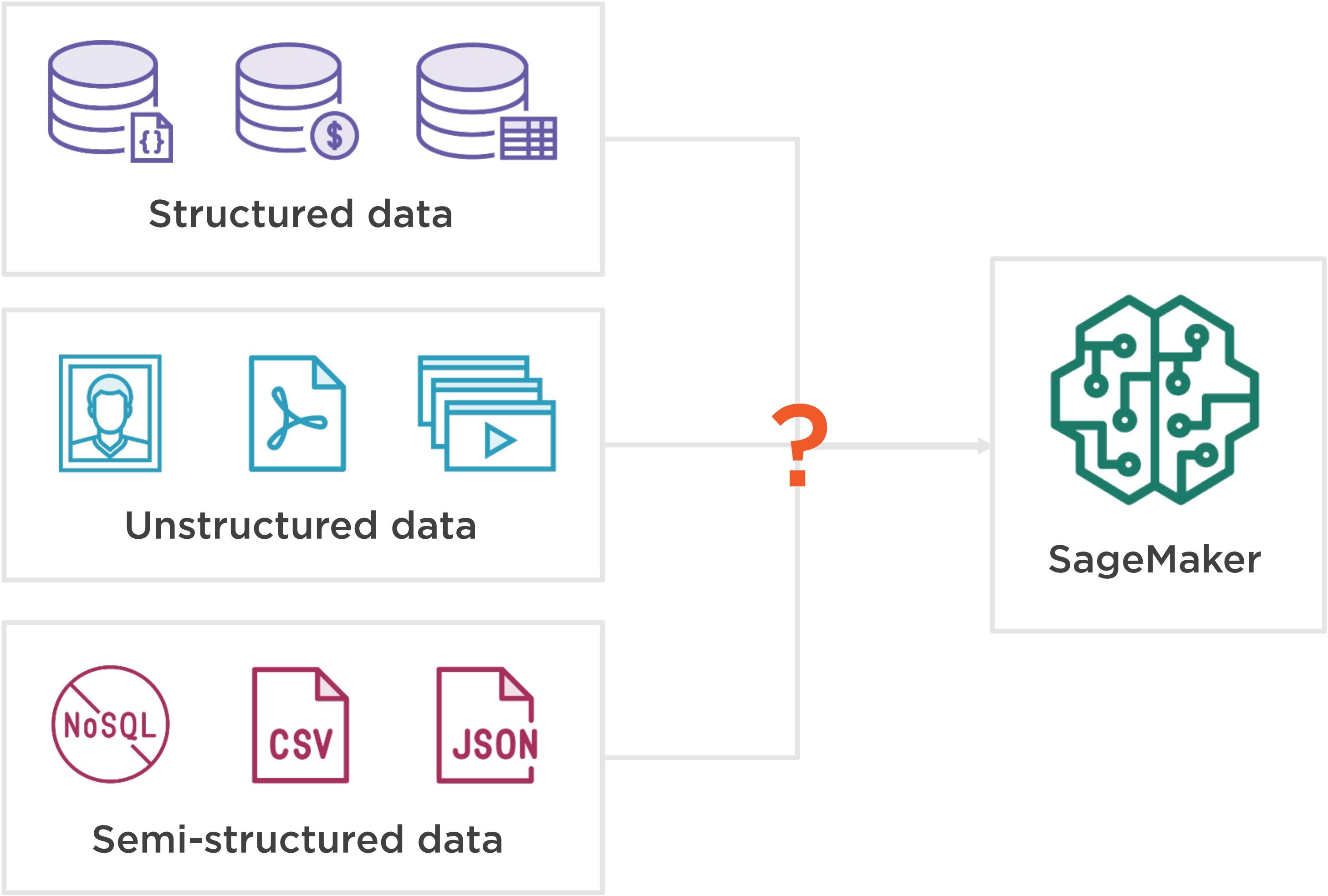
Most time-consuming part of the process

GOAL

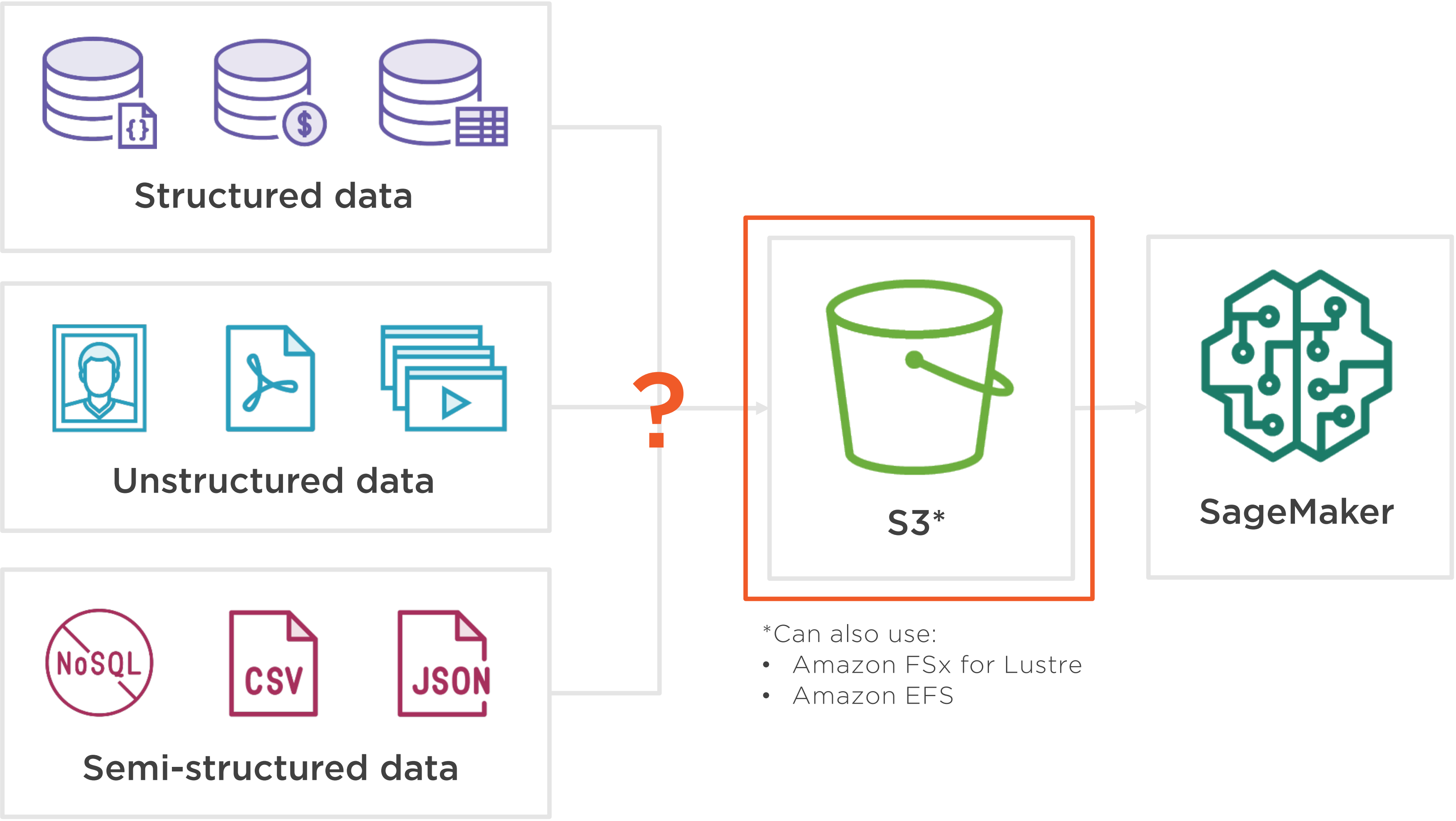
Fetch data from one or more data sources and get it into SageMaker


Fetch
Clean
Prepare
Train model
Evaluate model
Deploy to production
Monitor/ collect data/ evaluate
The Machine Learning Process

Getting Data into SageMaker

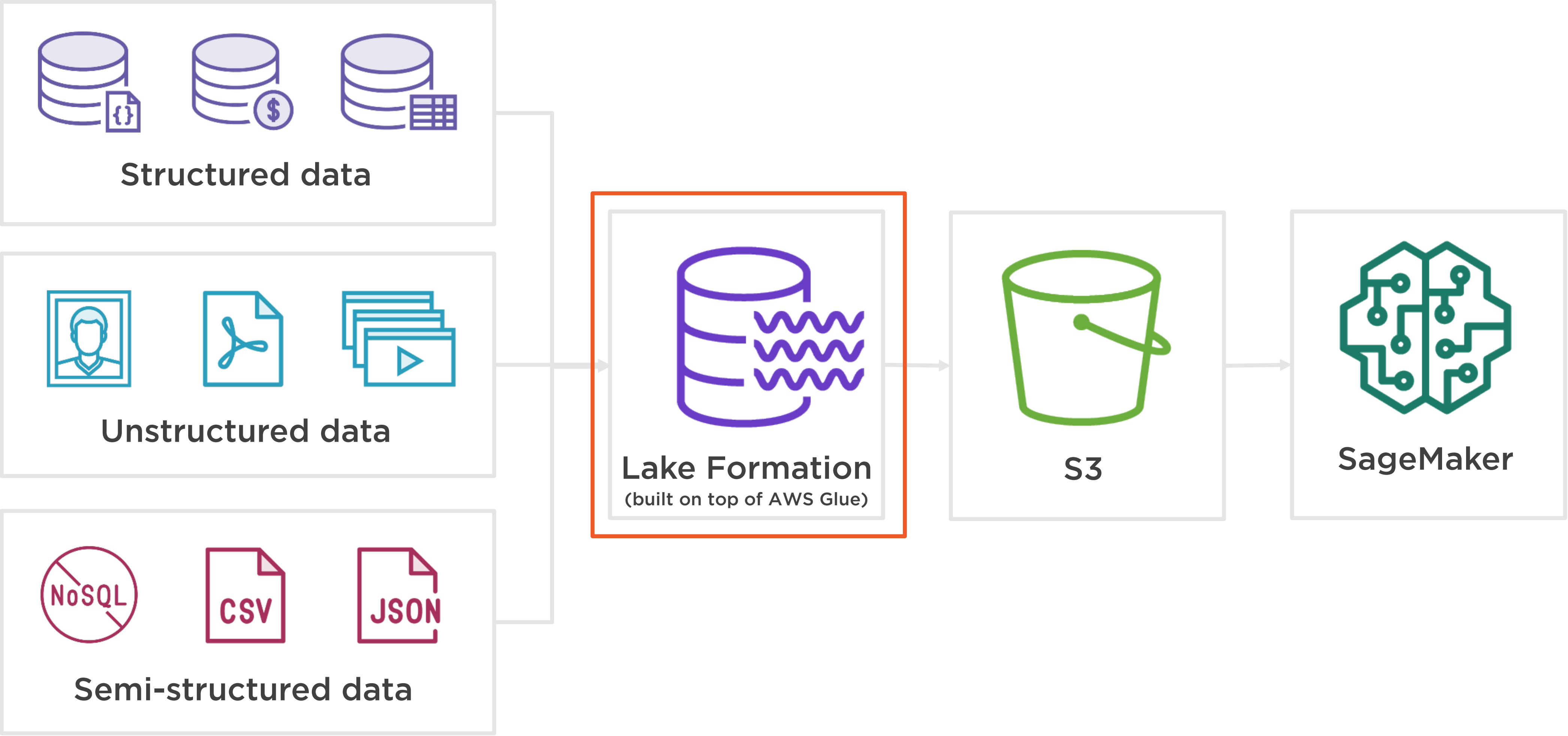


Getting Data into SageMaker

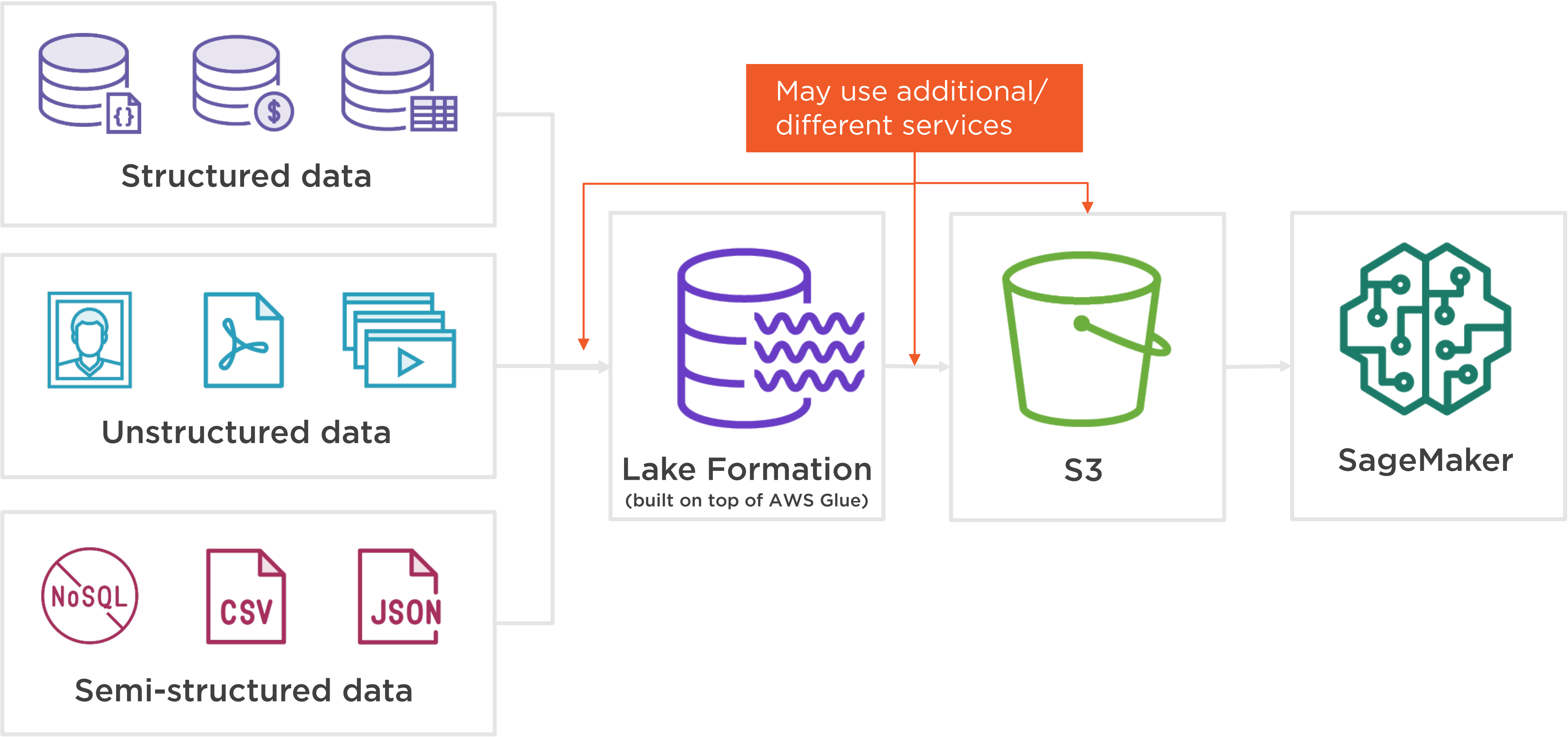


*Can also use:
- Amazon FSx for Lustre
- Amazon EFS

Getting Data into SageMaker



Getting Data into SageMaker



Two Types of Data Ingestion

Batch Processing

Periodically collect and send data

Can be activated on certain conditions or on a set schedule

Use when there is not a need for real-time processing

Generally cheaper and easier

Stream Processing

AWS Services Used for Batch Processing

SERVICE	HIGHLIGHTS
AWS Glue	• Fully managed ETL service • Runs on serverless Apache Spark environment • Uses crawlers to infer schemas and stores them in the Data Catalog
Data Pipeline	• Managed orchestration for data-driven workflows • Moves data between AWS compute and storage resources, or on-premises to AWS • Can store data in DynamoDB, RDS, Redshift, and S3
Database Migration Service (DMS)	• Migrate data between databases, either in AWS or on-premises • Supports homogeneous or heterogeneous migrations • Manages the infrastructure for you

Two Types of Data Ingestion

Batch Processing

Periodically collect and send data

Can be activated on certain conditions or on a set schedule

Use when there is not a need for real-time processing

Generally cheaper and easier

Stream Processing

Real-time processing

Data is loaded and manipulated as it's recognized (through constant monitoring)

Use when real-time data is required (e.g., stock prices)

More expensive

AWS Services Used for Stream Processing

Amazon Kinesis

- **Kinesis Video Streams**
- **Kinesis Data Streams**
- **Kinesis Data Firehose**
- **Kinesis Data Analytics**

The Kinesis Family of Services

SERVICE	HIGHLIGHTS
Video Streams	• Ingests, processes/streams and stores streaming video and audio data • Automatically provisions and scales infrastructure
Data Streams	• Ingests, processes and stores streaming data, breaking it into “shards” • Data has to be processed (e.g., Lambda, Data Analytics) before storing (which is optional) • Data retention of 24 hours by default (can be extended to 7 days)
Data Firehose	• Ingests, processes and stores streaming data, without “shards” • Can stream directly to storage (processing is optional) • If data delivery to S3 fails, the retries are automatic, but data is discarded after 24 hours
Data Analytics	• Analyzes streaming data • Automatically provisions and scales infrastructure • Enables SQL querying and custom Java applications

Running SQL Queries on Your Data

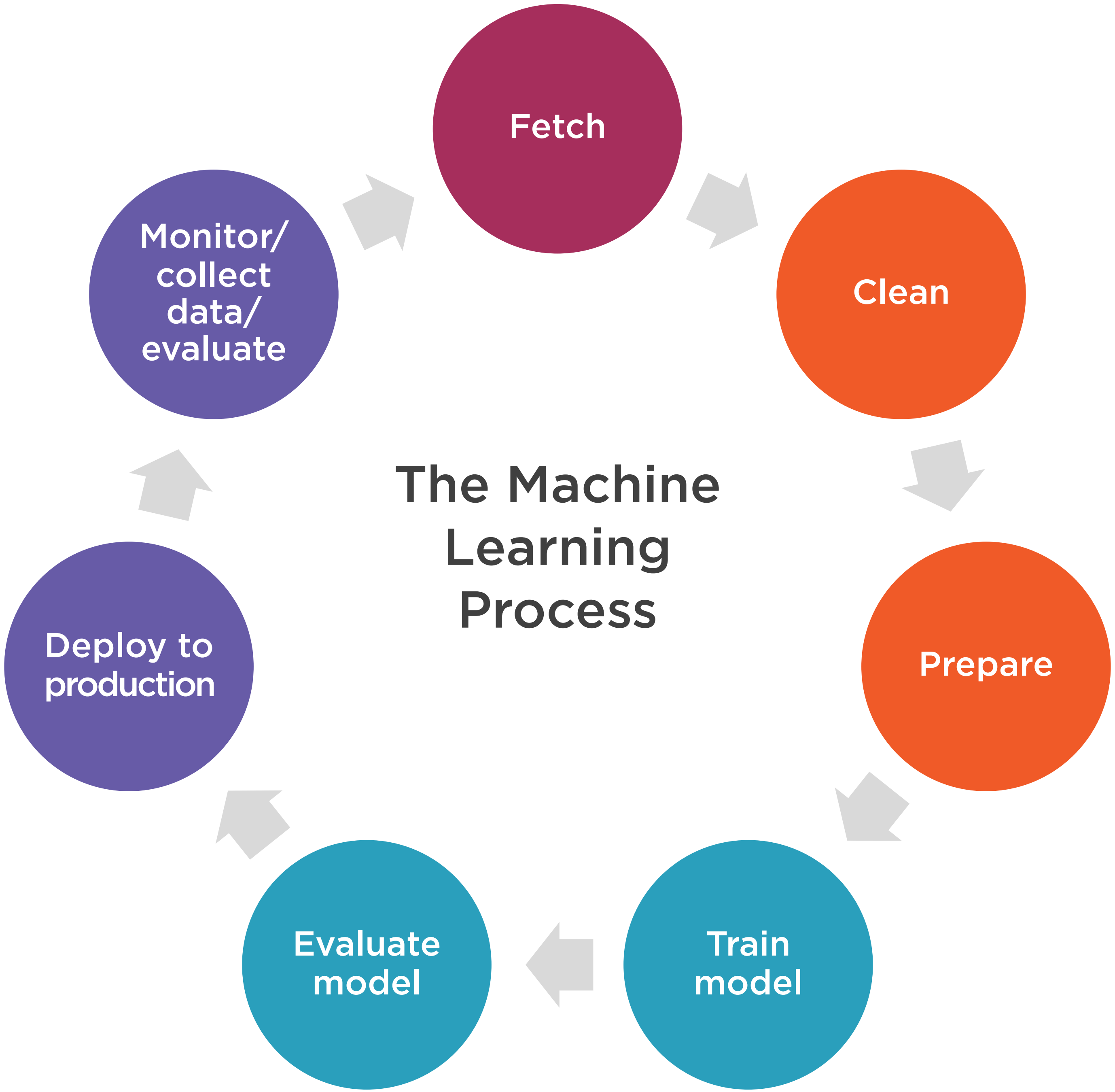
Amazon Athena

Redshift Spectrum

Dealing with Massive Amounts of Data

Amazon Elastic MapReduce (EMR)

Cleaning and Preparing Data


The Machine Learning Process
Fetch
Clean
Prepare
Train model
Evaluate model
Deploy to production
Monitor/ collect data/ evaluate

Data is Messy!

Inconsistent column names

Inconsistent naming/ abbreviations, blank values, nulls, and dashes

Inconsistent decimal formats, NaN, outlier/ typo of “450”

Inconsistent formatting, two different currencies, missing values

First_Name	LastName	State	Country	ZIP	Age	Gender	Salary
Becky	Johnson	Utah	-	84103-0437	44.0	F	€85,000
Wes	Byers	Washington	United States	98735	53	Male	$147,000
Tony	Herrera	Texas	USA	75002	32.0		90000
I-Chin	Chang	CA		9006	37	Female	$120000
Damian	Wilson		U.S.		27	Male	N/A
Kalene	Brown	AZ	null	85937	NaN	Female	75000
E.J.	Smith	Florida		32008	450		Unknown

Punctuation in names

Inconsistent formatting, blank values, one ZIP is only 4 digits

Inconsistent naming, blank values

Handle Missing Data

Remove rows or columns that have the missing data

Fill in the missing values

- The column mean or median
- Zero
- Null
- Imputation (your best guess)

Handle Outliers

Could be mistakes

Make it harder to get accurate predictions

Generally want to remove outliers

Handle Format

- Spacing
- Casing
- Punctuation
- Decimal points
- Special characters
- Currencies
- Abbreviations

Data Visualization and Analysis

Data Visualization and Analysis

Better understand your data and feature relationships

Descriptive Statistics

Rows

Columns

Mean

Median

Standard deviation

Count and most/least frequent values

Gaining Insights with Visualizations

Is there correlation between features?

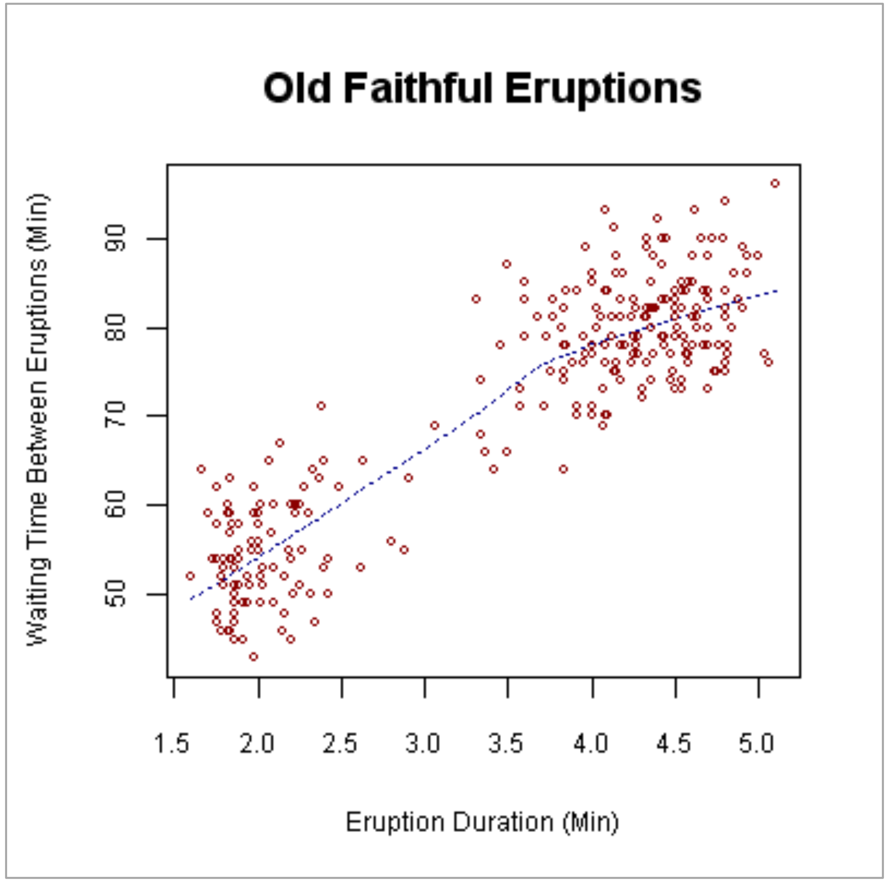
What are the mean, min, max values?

Are there any interesting patterns?

Are there any outliers?

Are there any features we need to add?

Scatter Plot



Public Domain
https://commons.wikimedia.org/w/index.php?curid=646999

Used to show relationship between two variables

Positive correlation: line slopes from lower left to upper right

Negative correlation: line slopes from upper left to lower right

In this example:

- Positive correlation between wait time and duration
- Short-wait-short-duration
- Long-wait-long-duration

Correlation Matrix

	Poverty	Breast Cancer	Stroke	Obesity	High Blood Pressure
Poverty	1.0	0.04	0.12	0.01	-0.0
Breast Cancer	0.04	1.0	0.4	0.33	0.27
Stroke	0.12	0.4	1.0	0.2	0.12
Obesity	0.01	0.33	0.2	1.0	0.6
High Blood Pressure	-0.0	0.27	0.12	0.6	1.0

Used to quantify relationships between variables

Correlation of 1: variables are perfectly correlated (both move in the same direction)

Correlation of -1: the two variables are perfectly negatively correlated (move in opposite directions)

Correlation of 0: there is no linear relationship

Correlation Matrix

	Poverty	Breast Cancer	Stroke	Obesity	High Blood Pressure
Poverty	1.0				
Breast Cancer	0.04	1.0			
Stroke	0.12	0.4	1.0		
Obesity	0.01	0.33	0.2	1.0	
High Blood Pressure	-0.0	0.27	0.12	0.6	1.0

Used to **quantify relationships** between variables

Correlation of 1: variables are perfectly correlated (both move in the same direction)

Correlation of -1: the two variables are perfectly negatively correlated (move in opposite directions)

Correlation of 0: there is no linear relationship

Correlation Matrix

	Poverty	Breast Cancer	Stroke	Obesity	High Blood Pressure
Poverty	1.0	0.04	0.12	0.01	-0.0
Breast Cancer	0.04	1.0	0.4	0.33	0.27
Stroke	0.12	0.4	1.0	0.2	0.12
Obesity	0.01	0.33	0.2	1.0	0.6
High Blood Pressure	-0.0	0.27	0.12	0.6	1.0

Used to **quantify relationships** between variables

Correlation of 1: variables are perfectly correlated (both move in the same direction)

Correlation of -1: the two variables are perfectly negatively correlated (move in opposite directions)

Correlation of 0: there is no linear relationship

Correlation Matrix

	Poverty	Breast Cancer	Stroke	Obesity	High Blood Pressure
Poverty	1.0	0.04	0.12	0.01	-0.0
Breast Cancer	0.04	1.0	0.4	0.33	0.27
Stroke	0.12	0.4	1.0	0.2	0.12
Obesity	0.01	0.33	0.2	1.0	0.6
High Blood Pressure	-0.0	0.27	0.12	0.6	1.0

Used to **quantify relationships** between variables

Correlation of 1: variables are perfectly correlated (both move in the same direction)

Correlation of -1: the two variables are perfectly negatively correlated (move in opposite directions)

Correlation of 0: there is no linear relationship

Correlation Matrix

	Poverty	Breast Cancer	Stroke	Obesity	High Blood Pressure
Poverty	1.0	0.04	0.12	0.01	-0.0
Breast Cancer	0.04	1.0	0.4	0.33	0.27
Stroke	0.12	0.4	1.0	0.2	0.12
Obesity	0.01	0.33	0.2	1.0	0.6
High Blood Pressure	-0.0	0.27	0.12	0.6	1.0

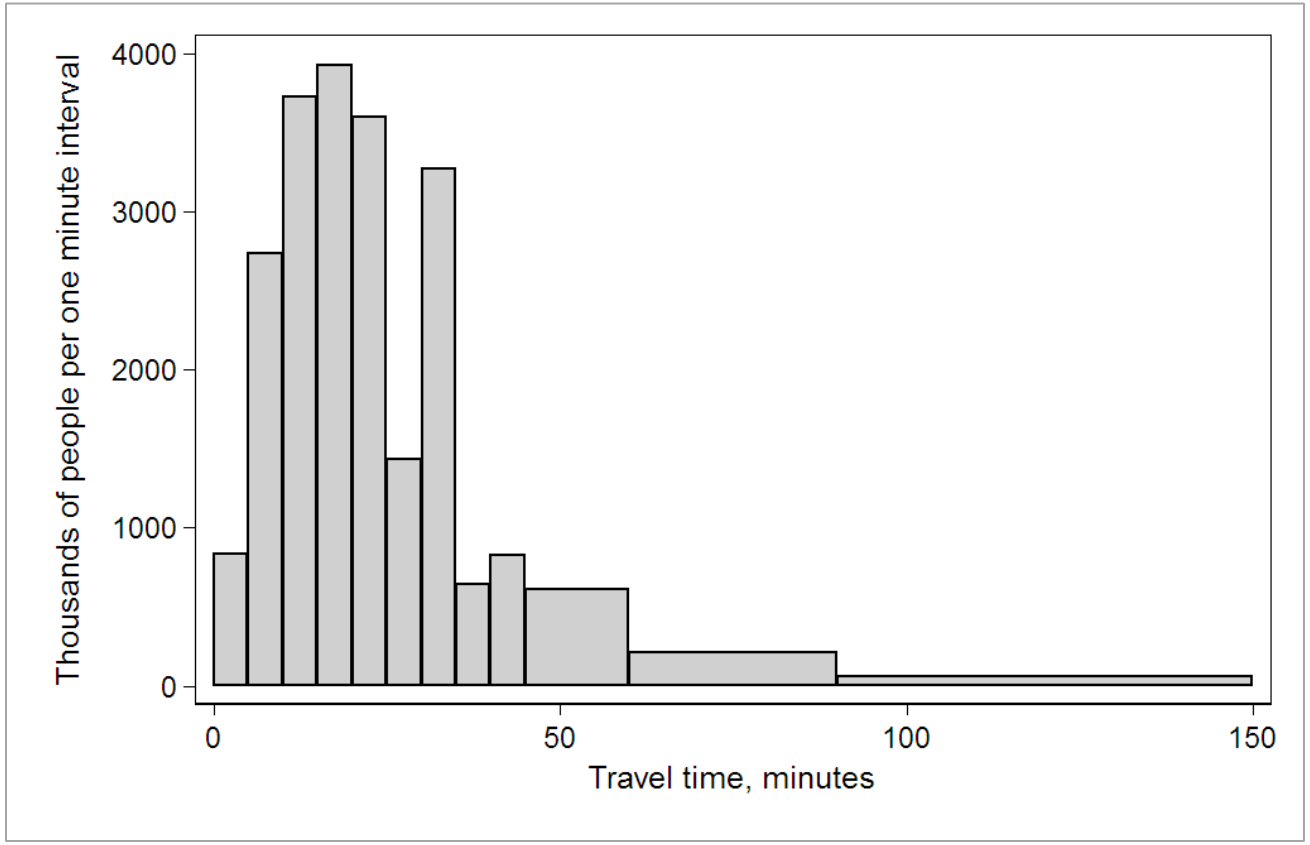
Used to **quantify relationships** between variables

Correlation of 1: variables are perfectly correlated (both move in the same direction)

Correlation of -1: the two variables are perfectly negatively correlated (move in opposite directions)

Correlation of 0: there is no linear relationship

Histogram



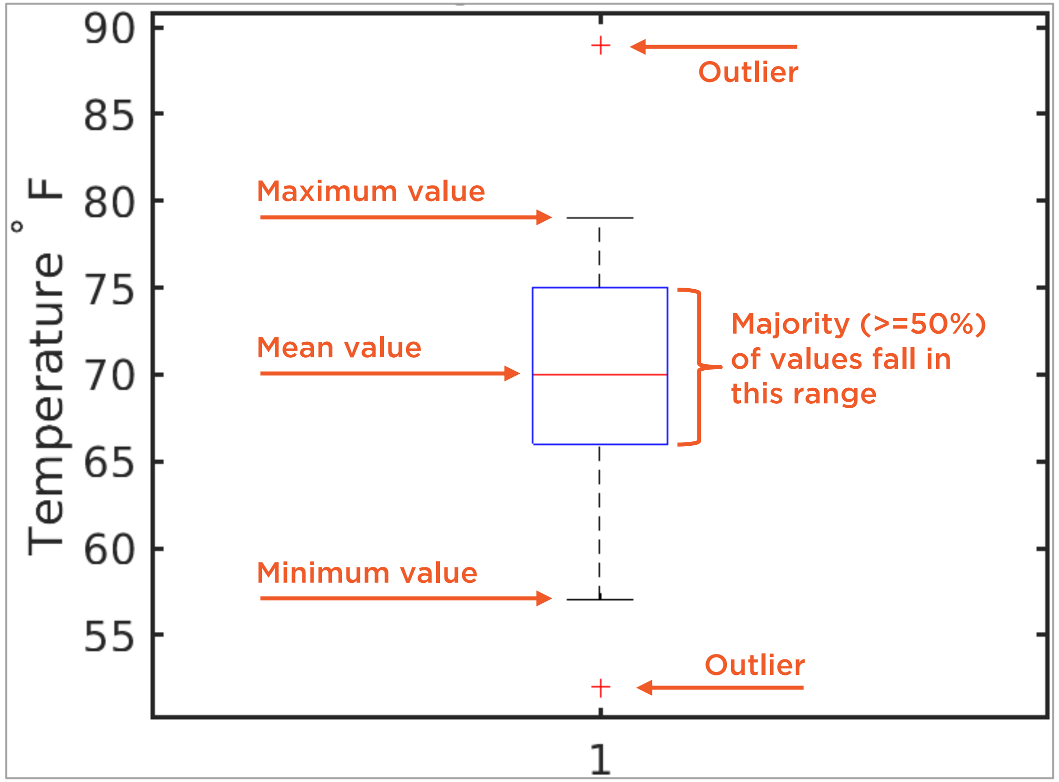
Used to show distribution of data

Values are grouped into "bins"

In this example:

- The majority of people are commuting less than 30 minutes

Box Plots



By Ever.chae - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=84823719

Used to show **distribution** of data

In this example:

- The majority of temperatures are between 67 and 75
- Max temp is 79
- Min temp is 57
- Mean temp is 70
- There are some outlier temps

Feature Engineering

Feature Engineering

The process of transforming raw data into features that better represent the underlying problem

Goal: Increase the model's predictive power

Auto Loan Approvals

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	Yes	11326	36498	718	Yes
SUV	2019	Medium	Yes	8984	32099	785	Yes
Sedan	2016	Small	Yes	58446	9650	690	Yes
Truck	2020	Large	No	316	64800	620	No
Coupe	2019	Medium	Yes	7290	31000	750	Yes

DIMENSIONALITY REDUCTION

Are there any features we can drop?

Handling Scale

Example

- Measurements
  - Inches
  - Kilometers
  - Yards
- Age and income

Ways to handle

- Normalization
  - Rescale data so that values are between 0 and 1
- Standardization
  - Rescale distribution of data so that mean is 0 is standard deviation is 1

Auto Loan Approvals

Target

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	Yes	11326	36498	718	Yes
SUV	2019	Medium	Yes	8984	32099	785	Yes
Sedan	2016	Small	Yes	58446	9650	690	Yes
Truck	2020	Large	No	316	64800	620	No
Coupe	2019	Medium	Yes	7290	31000	750	Yes

Auto Loan Approvals

Target

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	Yes	11326	36498	718	Yes
SUV	2019	Medium	Yes	8984	32099	785	Yes
Sedan	2016	Small	Yes	58446	9650	690	Yes
Truck	2020	Large	No	316	64800	620	No
Coupe	2019	Medium	Yes	7290	31000	750	Yes

Binary categorical variables

Auto Loan Approvals

Target

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	~~Yes~~ 1	11326	36498	718	~~Yes~~ 1
SUV	2019	Medium	~~No~~ 0	8984	32099	785	~~Yes~~ 1
Sedan	2016	Small	~~Yes~~ 1	58446	9650	690	~~Yes~~ 1
Truck	2020	Large	~~No~~ 0	316	64800	620	~~No~~ 0
Coupe	2019	Medium	~~Yes~~ 1	7290	31000	750	~~Yes~~ 1

Binary categorical variables

Categorical Data

Describes Categories or Groups

NOMINAL

Order does Not matter

- {Red, Yellow, Blue}
- {Yes, No}

ORDINAL

Order does matter

- {Small, Medium, Large}
- {Hot, Hotter, Hottest}

Auto Loan Approvals

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	Medium	1	11326	36498	718	1
SUV	2019	Medium	0	8984	32099	785	1
Sedan	2016	Small	1	58446	9650	690	1
Truck	2020	Large	0	316	64800	620	0
Coupe	2019	Medium	1	7290	31000	750	1

Ordinal values
(order matters)

Auto Loan Approvals

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	~~Medium~~ 10	1	11326	36498	718	1
SUV	2019	~~Medium~~ 10	0	8984	32099	785	1
Sedan	2016	~~Small~~ 5	1	58446	9650	690	1
Truck	2020	~~Large~~ 15	0	316	64800	620	0
Coupe	2019	~~Medium~~ 10	1	7290	31000	750	1

Ordinal values
(order matters)

One-to-one mapping
Small = 5
Medium = 10
Large= 15

Auto Loan Approvals

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	10	1	11326	36498	718	1
SUV	2019	10	0	8984	32099	785	1
Sedan	2016	5	1	58446	9650	690	1
Truck	2020	15	0	316	64800	620	0
Coupe	2019	10	1	7290	31000	750	1

Numerical values

Auto Loan Approvals

Type	Year	Size	Used	Miles	Price	Credit Score	Loan Approved
Truck	2018	10	1	11326	36498	718	1
SUV	2019	10	0	8984	32099	785	1
Sedan	2016	5	1	58446	9650	690	1
Truck	2020	15	0	316	64800	620	0
Coupe	2019	10	1	7290	31000	750	1

Nominal values
(order doesn't matter)

Numerical encoding
not recommended

THE SOLUTION: one-hot encoding

One-hot Encoding

	Type
1	Truck
2	SUV
3	Sedan
4	Truck
5	Coupe

	Type_Truck	Type_SUV	Type_Sedan	Type_Coupe
1				
2				
3				
4				
5				

One-hot Encoding

	Type
1	Truck
2	SUV
3	Sedan
4	Truck
5	Coupe

	Type_Truck	Type_SUV	Type_Sedan	Type_Coupe
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	0	1

Tools for Preparing and Visualizing Your Data

SageMaker and Jupyter Notebooks

AWS Glue

Amazon QuickSight

Getting Some Human Help

SageMaker Ground Truth

Mechanical Turk (Human Workforce)

Demo

Fetching and preparing data in SageMaker Studio

Key Points to Remember

Fetching and transforming data

- Get data from various sources into S3
- AWS services
  - Lake Formation
  - AWS Glue
  - Data Pipeline
  - Database Migration Service (DMS)
  - Kinesis
  - EMR
  - Athena
  - Redshift Spectrum
  - QuickSight
  - Ground Truth

Cleaning data

- Sanitize data to handle missing values, outliers and formatting

Common data visualizations

- Scatter plot
- Correlation matrix
- Histogram
- Box plots

Feature engineering

- Handle scaling issues
- Categorical data describes categories or groups
  - Nominal: order does not matter
  - Ordinal: order does matter
- One-hot encoding

Up Next:
Training and Evaluating the Model