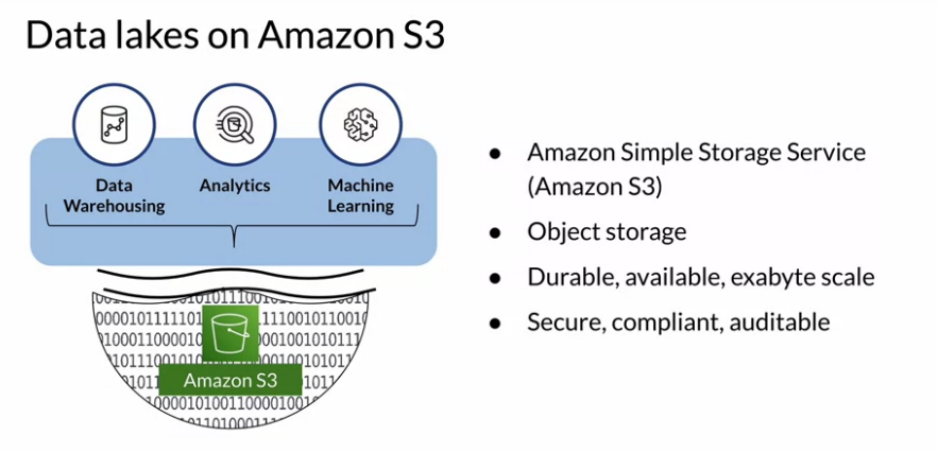
Analyze Datasets and Train ML Models using AutoML

Week 1 - Explore the Use Case and Analyze the Dataset

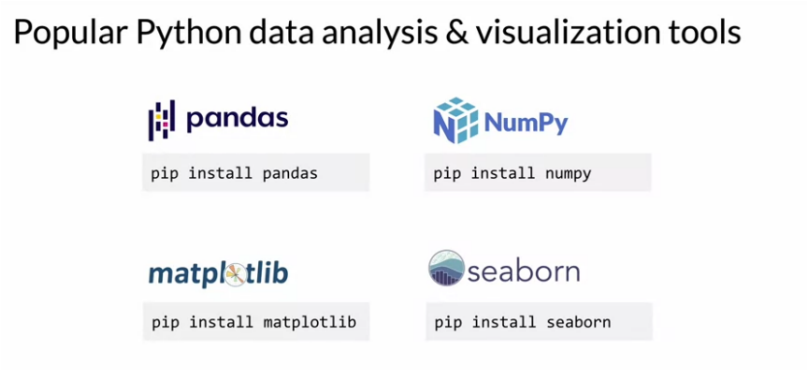
This week they talked about data ingestion, exploration and visualization.

Data Ingestion & Exploration



The data can be stored in S3, then with AWS Glue it will be made available for AWS Athena.

Data Visualization

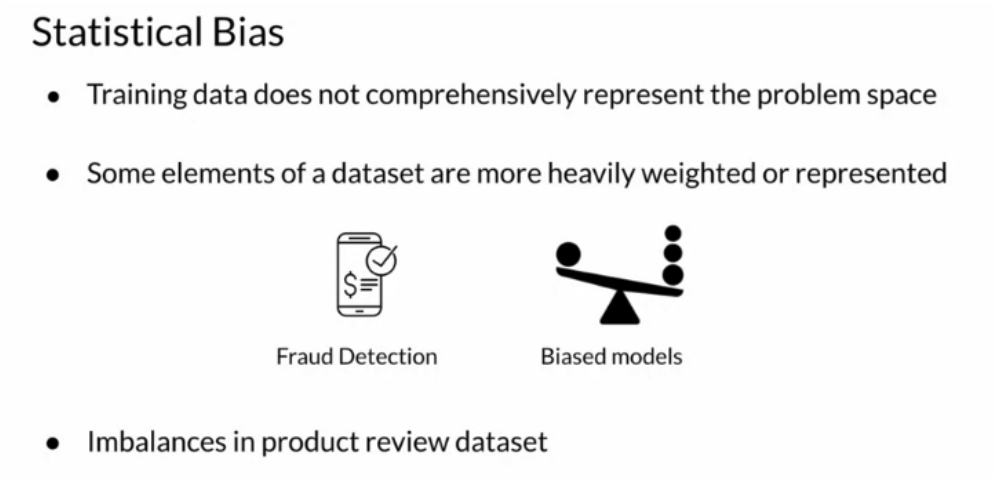


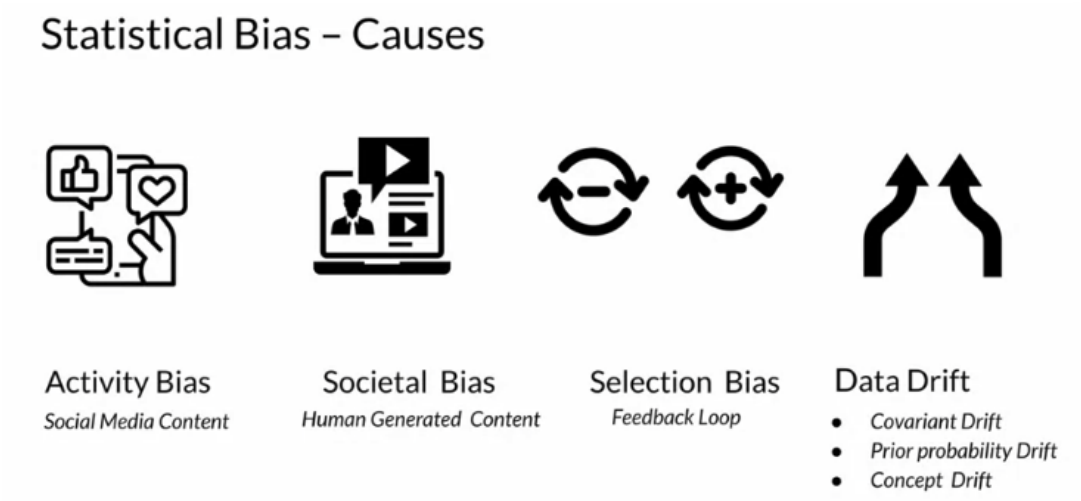
There are many ways and libraries to help with the visualization of your data.
Example:

- For visualization: histograms, series plots, and so on.
- For libraries: pandas, numpy, matploblib and seaborn.

Week 2 - Data Bias and Feature Importance

This week they talked about Statistical Bias, Bias Detection and Feature Importance.





They elaborate two metrics for imbalance:

Class Imbalance (CI): measures the imbalance in the number of members between different facet values. E.g. "Does a product_category have disproportionately more reviews than others?"

Difference in Proportions of Labels (DPL): measures the imbalance of positive outcomes between different facet values. E.g. "Does a product_category have disproportionately higher ratings than others?"

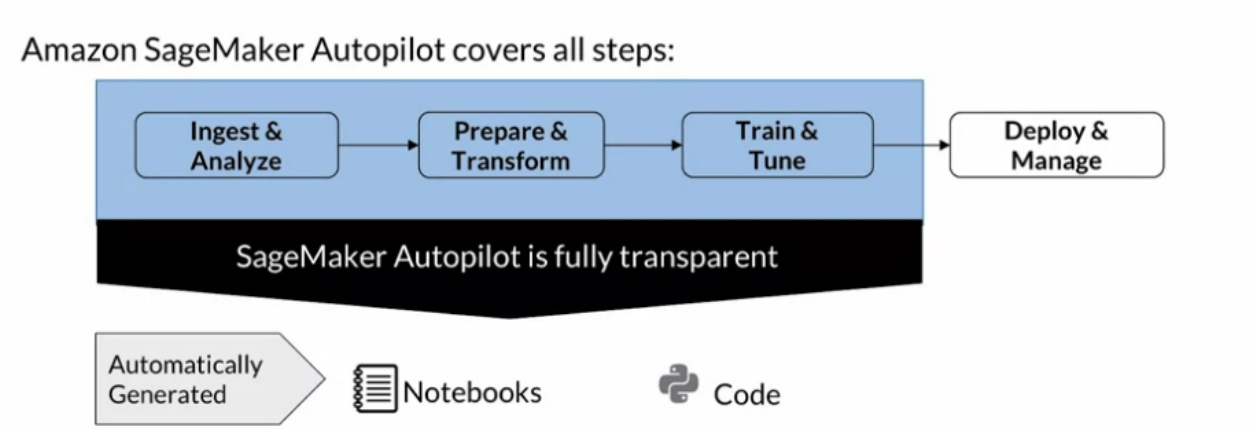
For **detecting statistical bias** they present two approaches/tools: **Amazon SageMaker Data Wrangler** and **Amazon SageMaker Clarify**.

Week 3 - Train a model with Amazon SageMaker Autopilot

This week they talked about AutoML (the concept), what it comprehends and also about AutoPilot (the Automated Machine Learning tool from AWS).

AutoML will cover:

1. Ingest & Analyze: something like EDA and bias detection.
2. Prepare & Transform: data preparation.
3. Train & Tune: train different algorithms, evaluate them and compare.
4. Deploy & Manage: deploy the select model to an endpoint API.

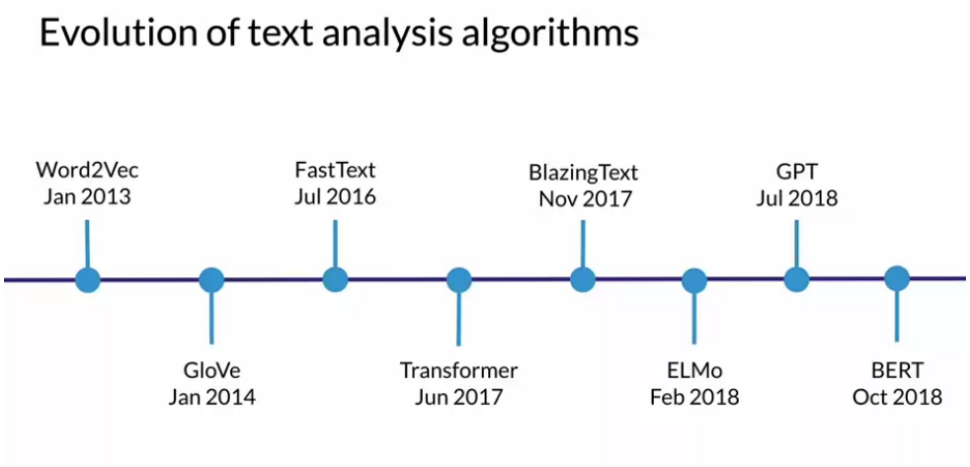


Week 4 - Built-in algorithms

This week they talked about built-in algorithms, how they worked and different problem spaces:

- Classification & Regression (supervised)
- Clustering (unsupervised)
- Image Processing (computer vision)
- Text Analysis (NLP)

The focus was on Text Analysis algorithms with some brief explanation of each concept over time.



They also talked about Amazon SageMaker BlazingText and their parameters.

Amazon SageMaker BlazingText hyper-parameters for text classification

Parameter Name	Recommended Ranges or Values	Description
epochs	[5-15]	Number of complete passes through the dataset
learning_rate	[0.005-0.01]	Step size for the numerical optimizer
min_count	[0-100]	Discard words that appear less than this number
vector_dim	[32-300]	Number of dimensions in vector space
word_ngrams	[1-3]	Number of words n-gram features to use
early_stopping	True or False	Stop training if validation accuracy stops improving
patience	[5-15]	Number of epochs before early stopping

They also talked about how to deploy a model and provided snippets.

```
text_classifier = estimator.deploy(
    initial_instance_count=1,
    instance_type='ml.m4.xlarge', ...)
```

Increase instance_count > 1 to easily scale out