Copyright Notice

DeepLearning.AI

aws

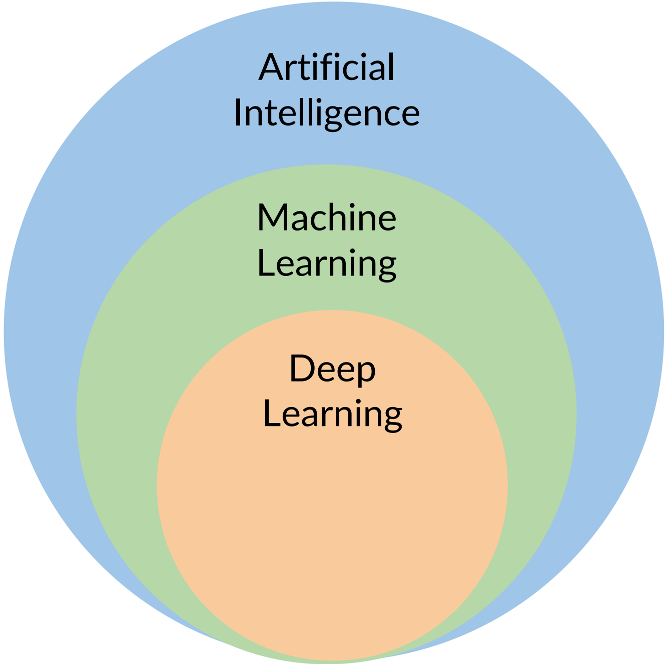
Practical Data Science

Explore the Use Case and Analyze the Dataset

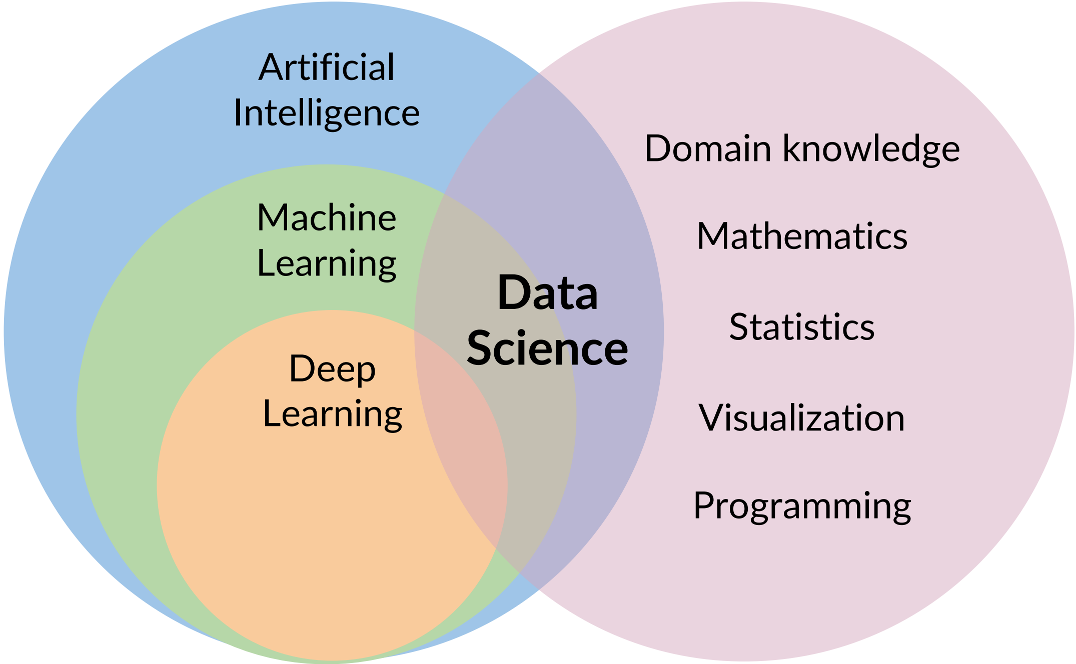
Practical Data Science in the Cloud

Introduction

AI, ML, DL, data science…?

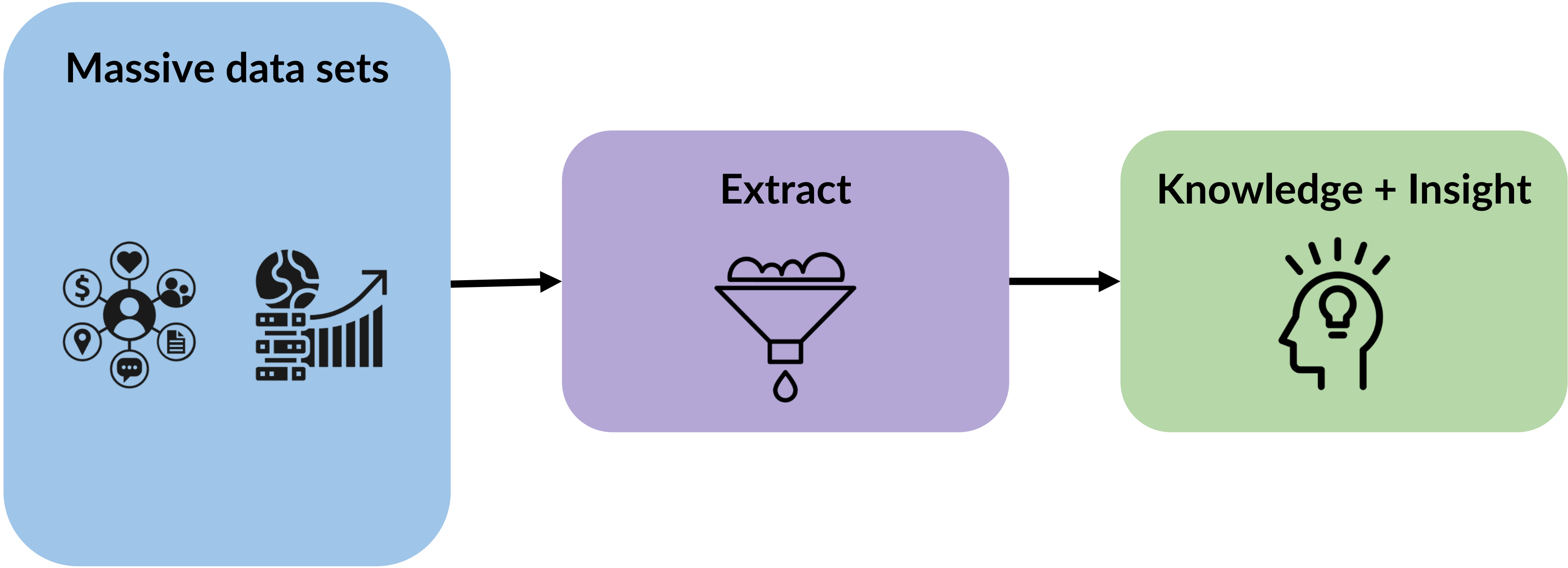


AI, ML, DL, data science...?



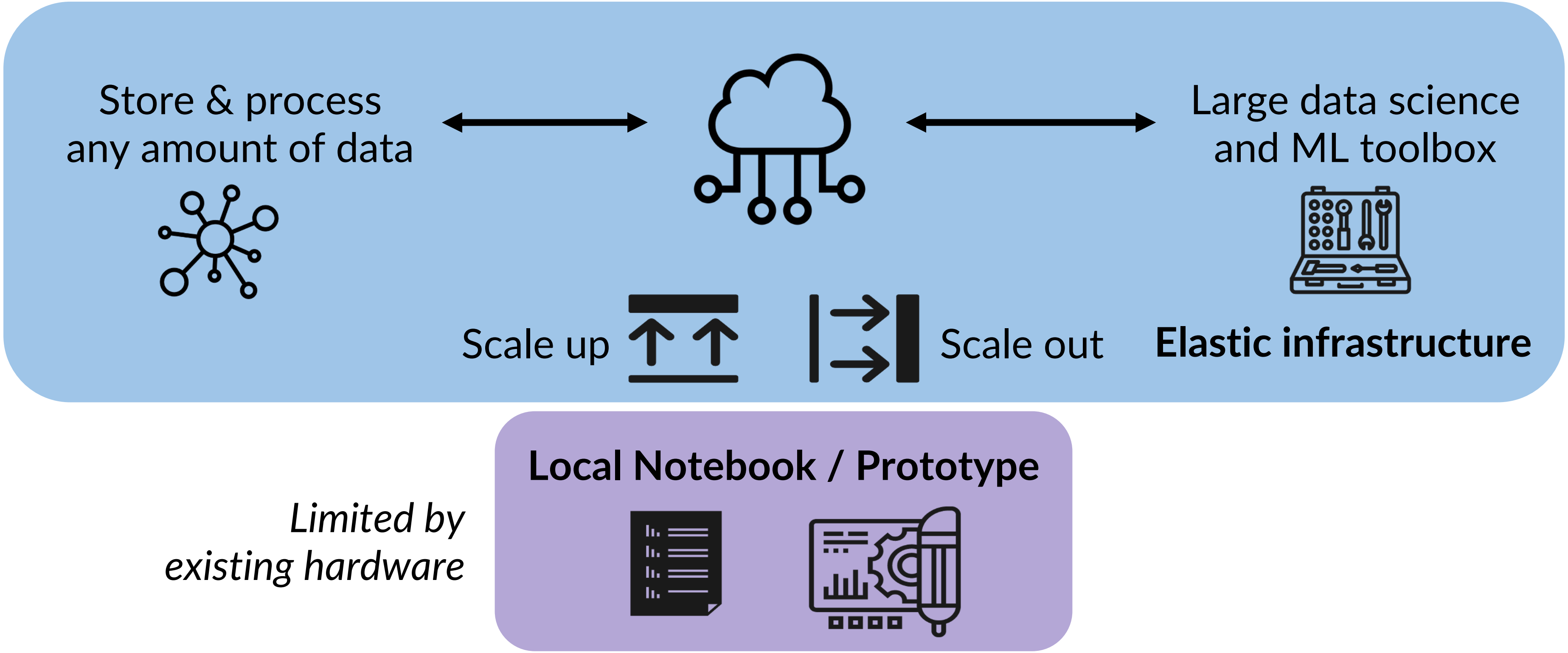
Practical Data Science?

Practical data science



... in the *Cloud*?

Practical data science in the cloud



Store & process any amount of data

Large data science and ML toolbox

Scale up

Scale out

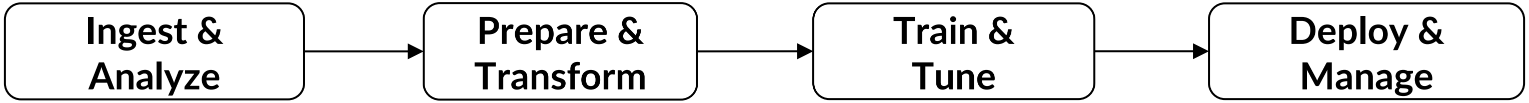
Elastic infrastructure

Local Notebook / Prototype

Limited by existing hardware

Data science and ML toolbox

Machine Learning Workflow



Ingest & Analyze	Prepare & Transform	Train & Tune	Deploy & Manage
Data exploration	Feature engineering	Automated ML	Model deployment
Bias detection	Feature store	Model train and tune	Automated pipelines
Amazon S3 & Amazon Athena	Amazon SageMaker Data Wrangler	Amazon SageMaker Autopilot	Amazon SageMaker Endpoints
AWS Glue	Amazon SageMaker Processing Jobs	Amazon SageMaker Training & Debugger	Amazon SageMaker Batch Transform
Amazon SageMaker Data Wrangler & Clarify	Amazon SageMaker Feature Store	Amazon SageMaker Hyperparameter Tuning	Amazon SageMaker Pipelines

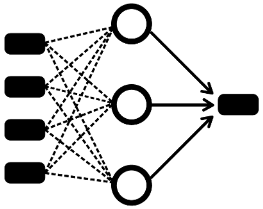
Machine Learning Workflow

Ingest & Analyze	Prepare & Transform	Train & Tune	Deploy & Manage
Data exploration	Feature engineering	Automated ML	Model deployment
Bias detection	Feature store	Model train and tune	Automated pipelines
Amazon S3 & Amazon Athena	Amazon SageMaker Data Wrangler	Amazon SageMaker Autopilot	Amazon SageMaker Endpoints
AWS Glue	Amazon SageMaker Processing Jobs	Amazon SageMaker Training & Debugger	Amazon SageMaker Batch Transform
Amazon SageMaker Data Wrangler & Clarify	Amazon SageMaker Feature Store	Amazon SageMaker Hyperparameter Tuning	Amazon SageMaker Pipelines

Use Case and Dataset

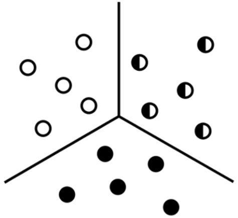
Introduction

Popular ML tasks and learning paradigms



Classification & Regression

Supervised



Clustering

Unsupervised

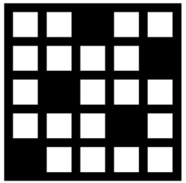
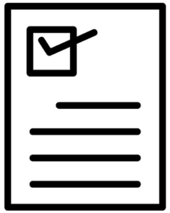


Image Processing

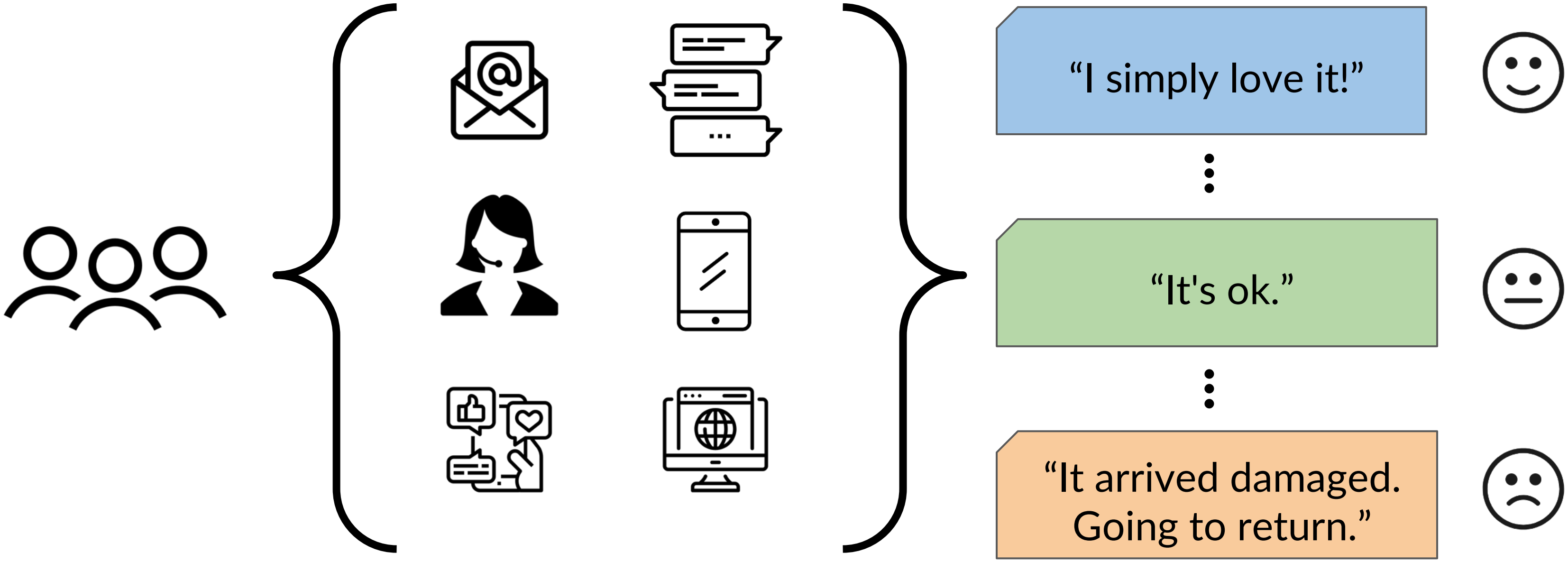
Computer Vision



Text Analysis

NLP / NLU

Multi-class classification for sentiment analysis of product reviews

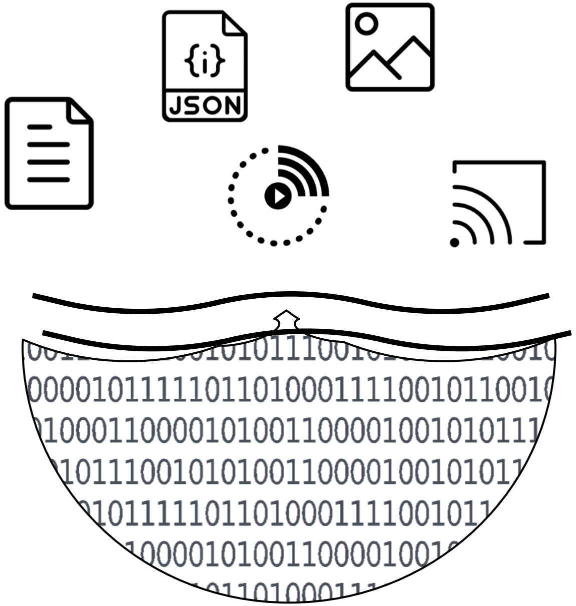


Working with product reviews data

Input feature for model training	Label for model training
Review Text	**Sentiment**
I simply love it!	1 (positive)
It's ok.	0 (neutral)
It arrived damaged, going to return	-1 (negative)

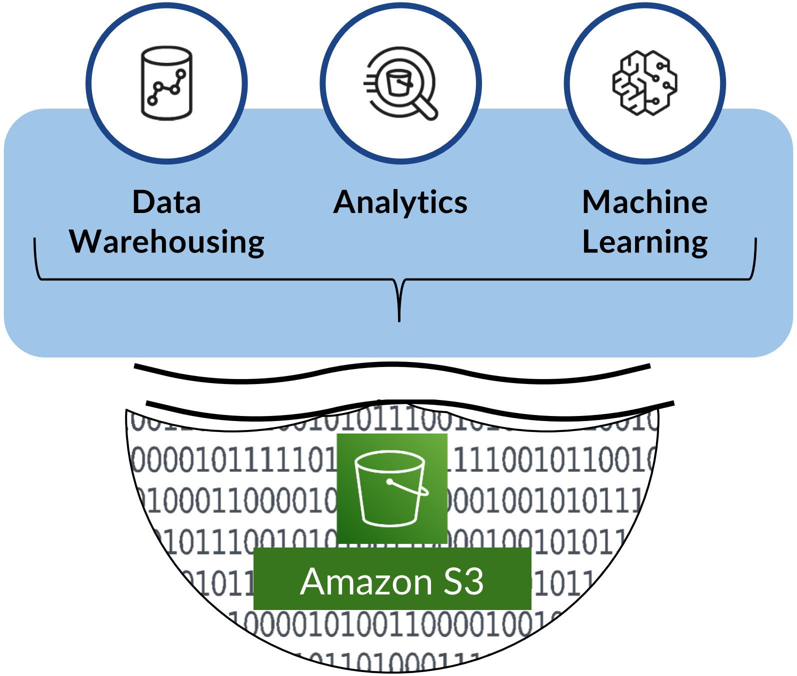
Data Ingestion & Exploration

Ingest data into data lakes



- Centralized and secure repository
- Store, discover and share data at any scale
  - structured relational data
  - semi-structured data
  - unstructured data
  - streaming data
- Governance

Data lakes on Amazon S3



- Amazon Simple Storage Service (Amazon S3)
- Object storage
- Durable, available, exabyte scale
- Secure, compliant, auditable

AWS Data Wrangler

- Open source Python library
- Connects pandas DataFrames and AWS data services
- Load/unload data from
  - data lakes
  - data warehouses
  - databases

```
!pip install awswrangler

import awswrangler as wr
import pandas as pd

# Retrieving the data directly from Amazon
S3
df = wr.s3.read_csv(
         path='s3://bucket/prefix/')
```

Register data with AWS Glue Data Catalog

AWS Glue
Data Catalog

Name	reviews
Database	dsoaws_deep_learning
Classification	csv
Location	s3://<bucket>/<prefix>

- Creates reference to data ("S3-to-table" mapping)
- Just metadata / schema stored in tables
- No data is moved
- *AWS Glue Crawlers* can be set up to automatically
  - infer data schema
  - update data catalog

Register data with AWS Glue Data Catalog

AWS Glue
Data Catalog

Name	reviews
Database	dsoaws_deep_learning
Classification	csv
Location	s3://<bucket>/<prefix>

```
import awswrangler as wr

# Create a database in the
# AWS Glue Data Catalog
wr.catalog.create_database(
         name=...)

# Create CSV table (metadata only) in the
# AWS Glue Data Catalog
wr.catalog.create_csv_table(
         table=...,
         column_types=...,
     ...)
```

Query data with Amazon Athena

Amazon Athena

- Query data in S3
- Using SQL
- No infrastructure to set up
- Schema lookup in AWS Glue Data Catalog
- No data to load

Python

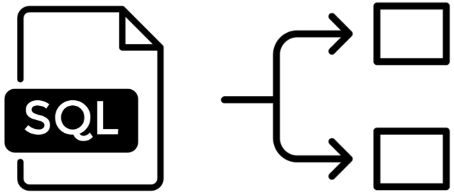
```python
import awswrangler as wr

# Create Amazon Athena S3 bucket
wr.athena.create_athena_bucket()

# Execute SQL query on Amazon Athena
df = wr.athena.read_sql_query(
    sql=...,
    database=...)
```

SQL

SQL

```sql
'SELECT product_category FROM reviews'
```

Query data with Amazon Athena



- Complex analytical queries
- Gigabytes > Terabytes > Petabytes
- Scales automatically
- Runs queries in parallel
- Based on Presto
- No infrastructure setup /
  no data movement required

Data Visualization

Popular Python data analysis & visualization tools

pandas

```
pip install pandas
```

NumPy

```
pip install numpy
```

matplotlib

```
pip install matplotlib
```

seaborn

```
pip install seaborn
```

How many reviews are in each *sentiment class*?

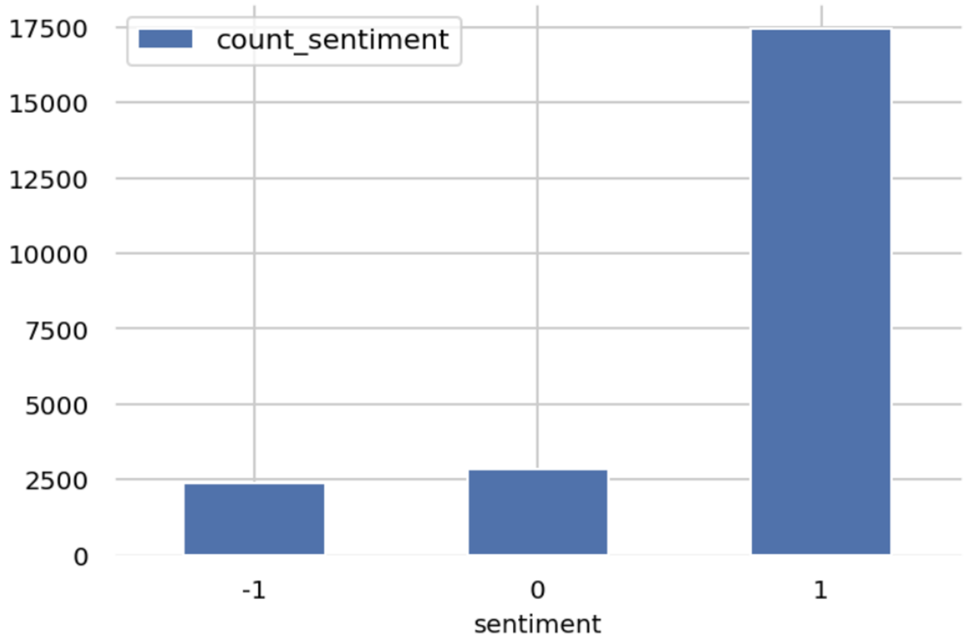
SQL Query

```sql
SELECT sentiment, COUNT(*) AS count_sentiment
FROM dsoaws_deep_learning.reviews
GROUP BY sentiment
ORDER BY sentiment DESC, count_sentiment
```

Python visualization code

```python
import matplotlib.pyplot as plt
chart = df.plot.bar(
         x="sentiment",
     y="count_sentiment")

plt.xlabel("sentiment")
plt.show(chart)
```

How many reviews are in each *sentiment class*?



What is the distribution of review lengths? *(number of words)*

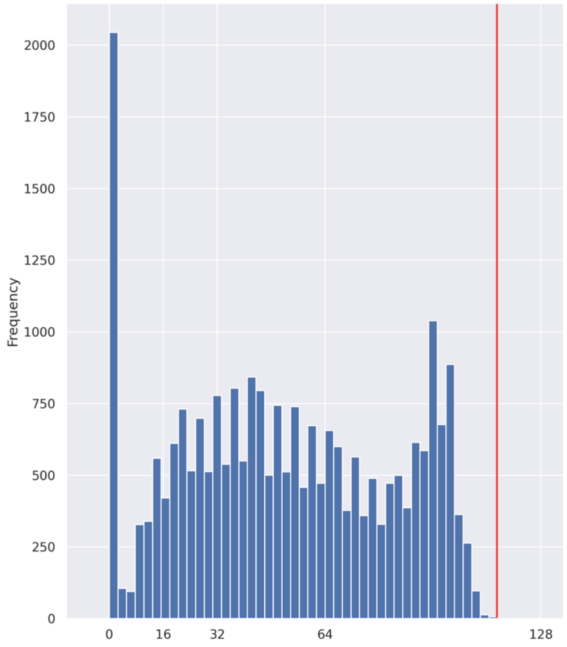
SQL Query

```
SELECT CARDINALITY(SPLIT(review_body, ' ')) as num_words
FROM dsoaws_deep_learning.reviews
```

Python visualization code

```
summary = df["num_words"].describe(
     percentiles=[0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 1.00])

df["num_words"].plot.hist(
     xticks=[0, 16, 32, 64, 128, 256], bins=100,
     range=[0, 256]).axvline(x=summary["100%"], c="red")
```

What is the distribution of review lengths? *(number of words)*



```
mean        52.51
std         31.38
min          1.00
10%         10.00
20%         22.00
30%         32.00
40%         41.00
50%         51.00
60%         61.00
70%         73.00
80%         88.00
90%         97.00
100%       115.00
```