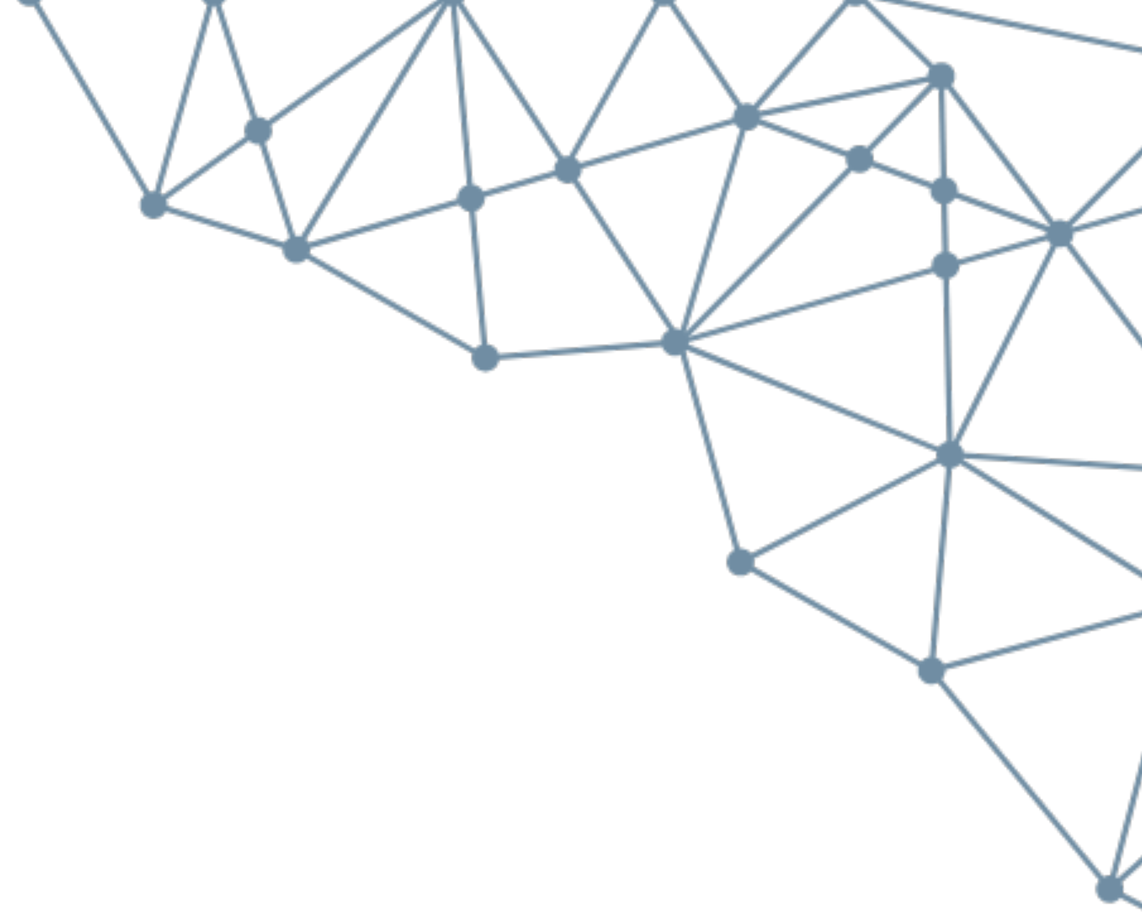
ABB - Session 2

Software 2.0, Data Engineering, & Machine Learning

Shaw Talebi

Today's Session



1. **Housekeeping**
    1.1. Announcements
    1.2. Homework 1
2. **Software 2.0**
    2.1. Machine Learning
    2.2. Data Engineering
3. **Example Code**
    3.1. ETL of Survey Data
    3.2. Training an ML Model

A Few Adjustments

1) Pre- and post-lecture Q&A (30 min)

2) Code walkthrough format

3) How to upload HW   

Homework 1

Shoutouts 🎉

File Organizer
Vladimir Belony

Resume Matcher
Saijai Osika

Report Image Extractor
Deborah Shutt

Automated Birthday Email Sender
Mathew Olajide

Personalized Mortgage Rate Emailer
Kalyan Mutyala

Football SuperLeague Leaderboard
Peirluigi Chiusolo

Bulk File Copier
Ronnie Rampersad

Radio Station Watcher
Sangeeta Bahri

Arxiv Paper Retriever
Fahad Ebrahim

News Feed Aggregator
Adam Rosenkoetter

Chess Grandmaster Ranking
Ludovic H

Python Project Summarizer
Bryce

iTunes Library Analysis
Rod Morrison

LI Internship Scraper and Emailer
Ewa Gros

Software 1.0

Rules are explicitly programmed into computer

You can do a lot with Software 1.0

But writing robust logic is hard…

… if possible.

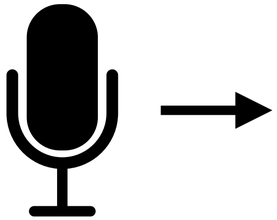
Software 1.0

Rules are explicitly programmed into computer

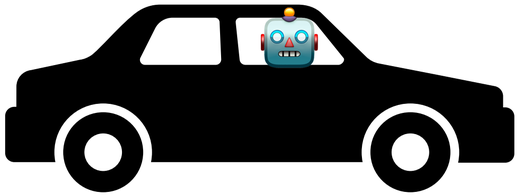
But writing robust logic is hard…

… if possible.

What is this?



Speech to text

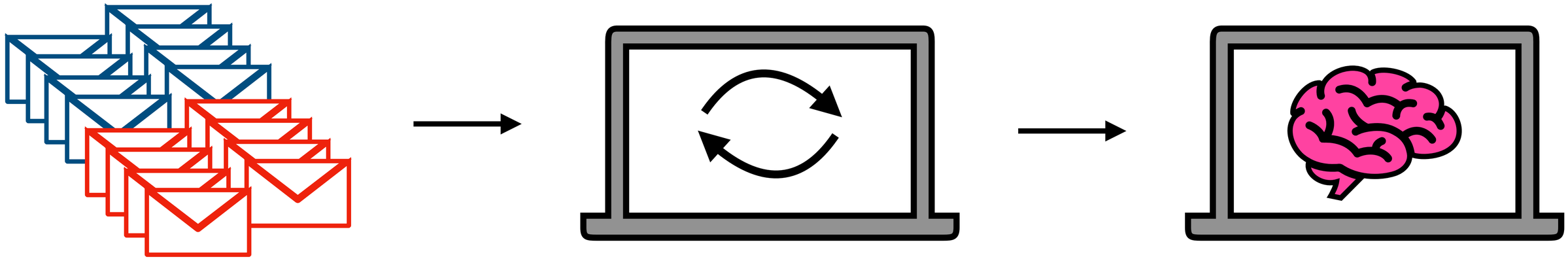


Self-driving

Software 2.0

Software 2.0

Programming computers by example (i.e. with data)



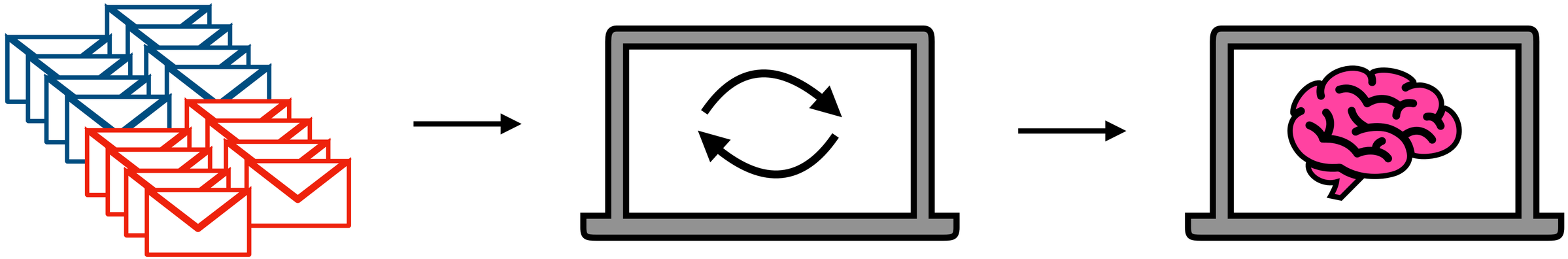
Gather spam/not spam examples

Pass to ML algorithm

ML Model

Machine Learning

Programming computers by example (i.e. with data)



Gather spam/not spam examples

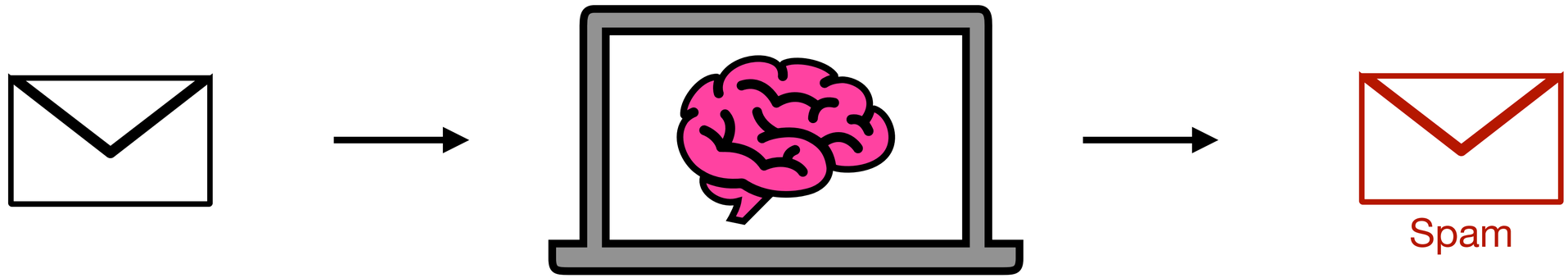
Pass to ML algorithm

ML Model

[1]

Machine Learning

Programming computers by example (i.e. with data)



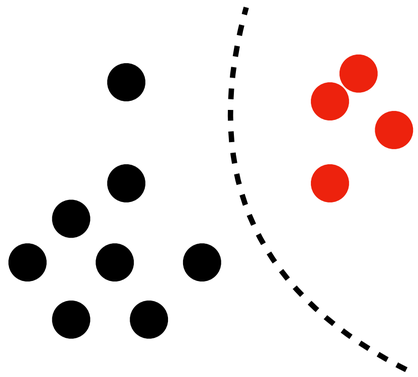
New Email

ML Model

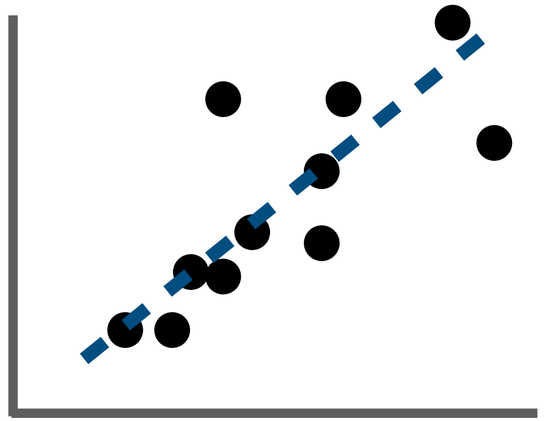
Prediction

3 Flavors of ML

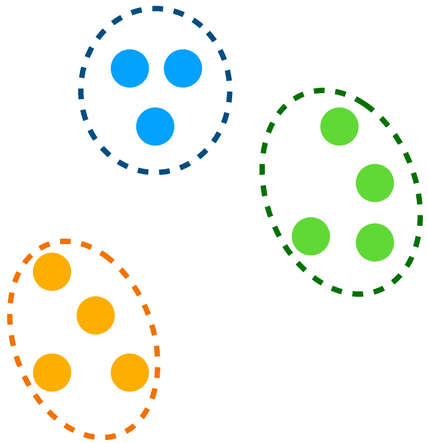
1) Classification



2) Regression

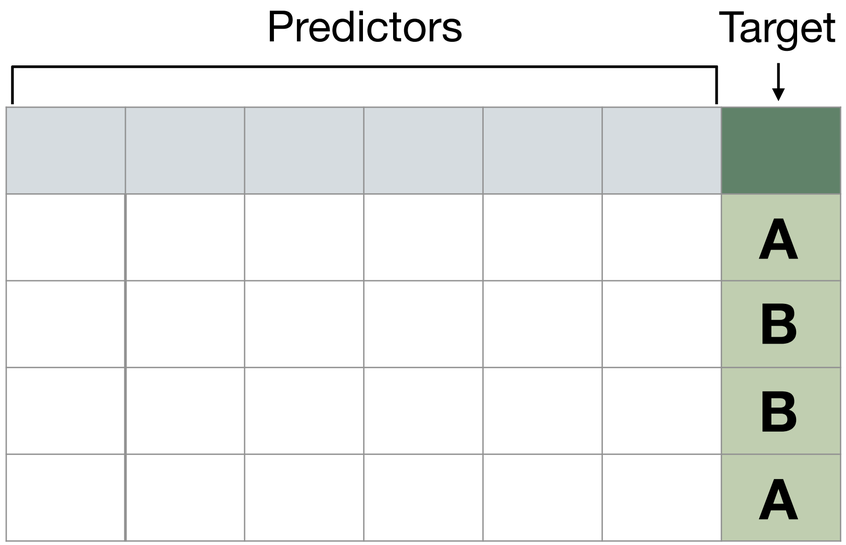


3) Clustering

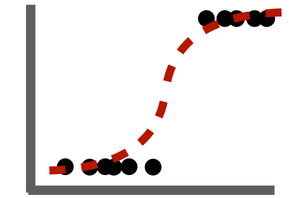


Flavor 1: Classification

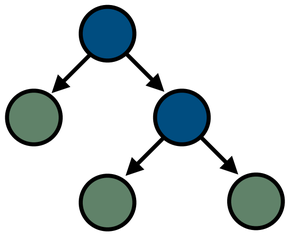
Labeling data with known categories



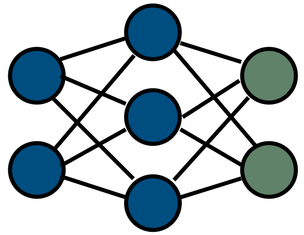
Training Data



Logistic Regression



Decision Tree Classifier



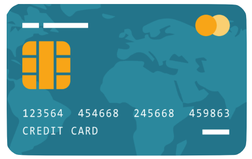
Neural Network

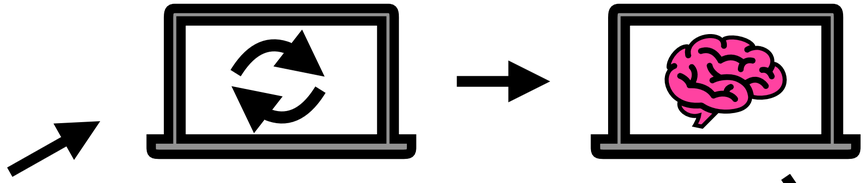
Techniques

[2] [3]

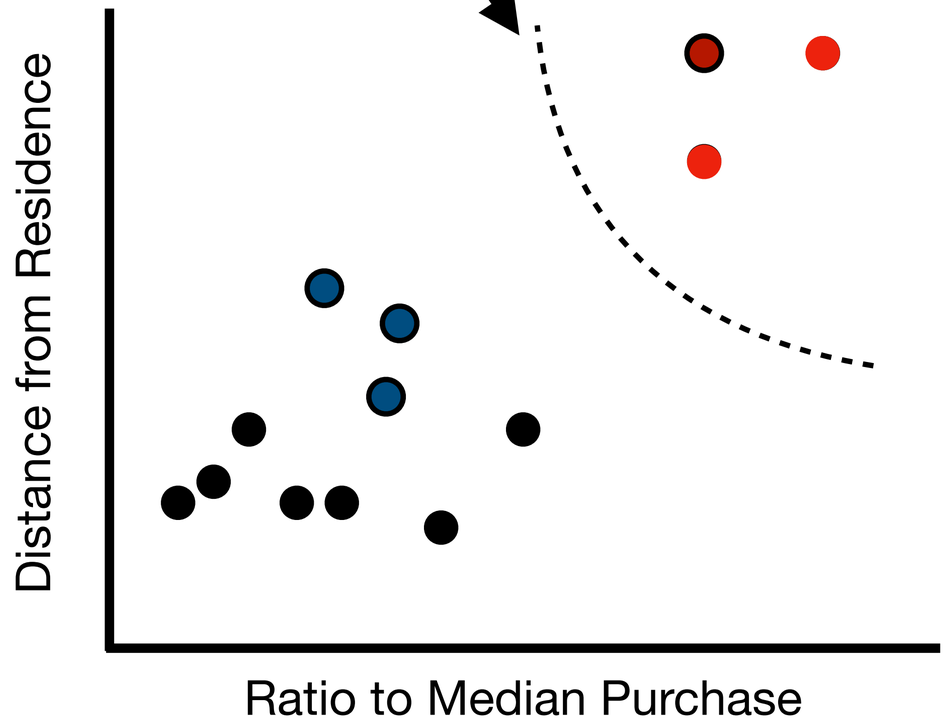
Flavor 1: Classification

Example: Fraud Detection



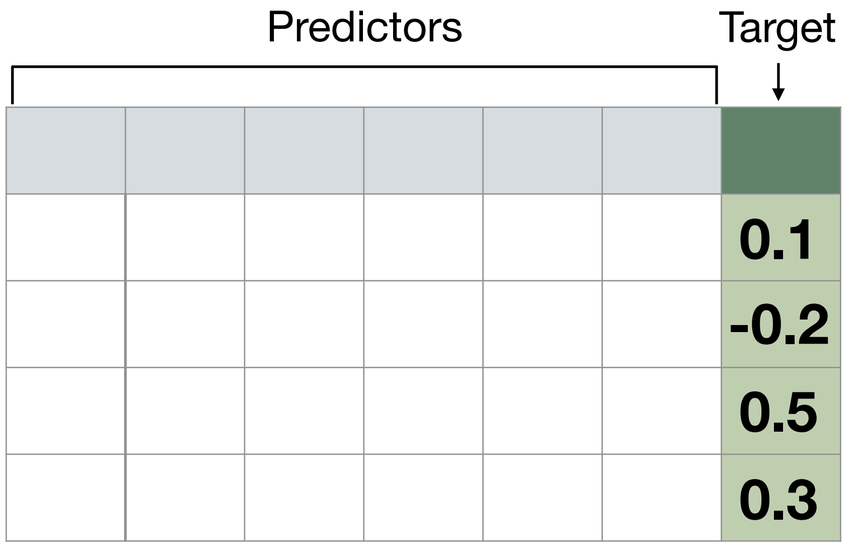


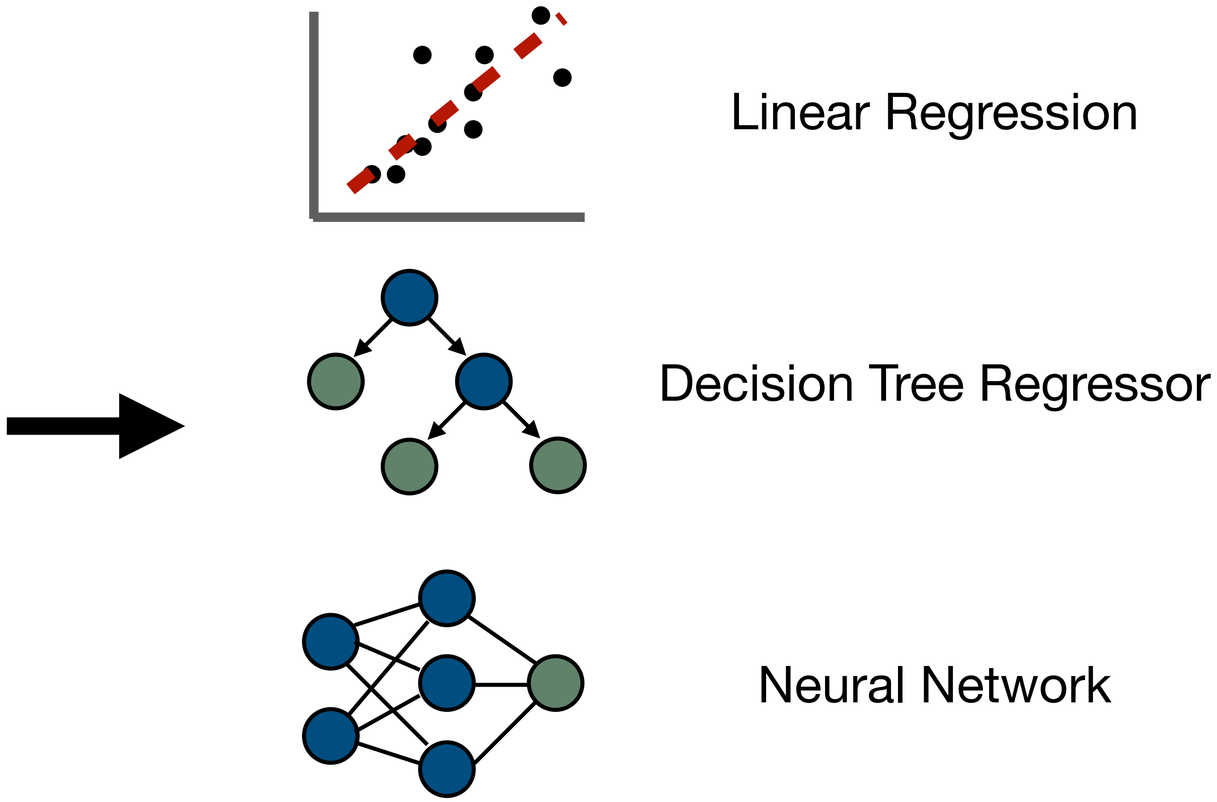
Ratio to Median Purchase	Distance from Residence	Fraud Flag
1.5	10	0
0.8	5	0
1.0	2	0
2.2	55	1
1.3	1	0
1.9	42	1
0.75	3	0
1.1	2	0



Flavor 2: Regression

Predicting a continuous value





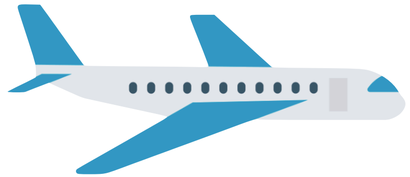
Training Data

Techniques

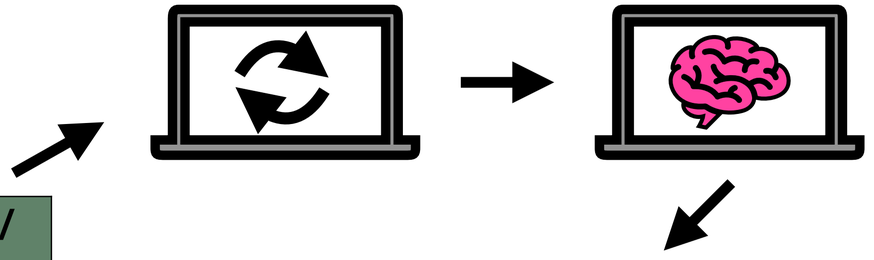
[3] [4]

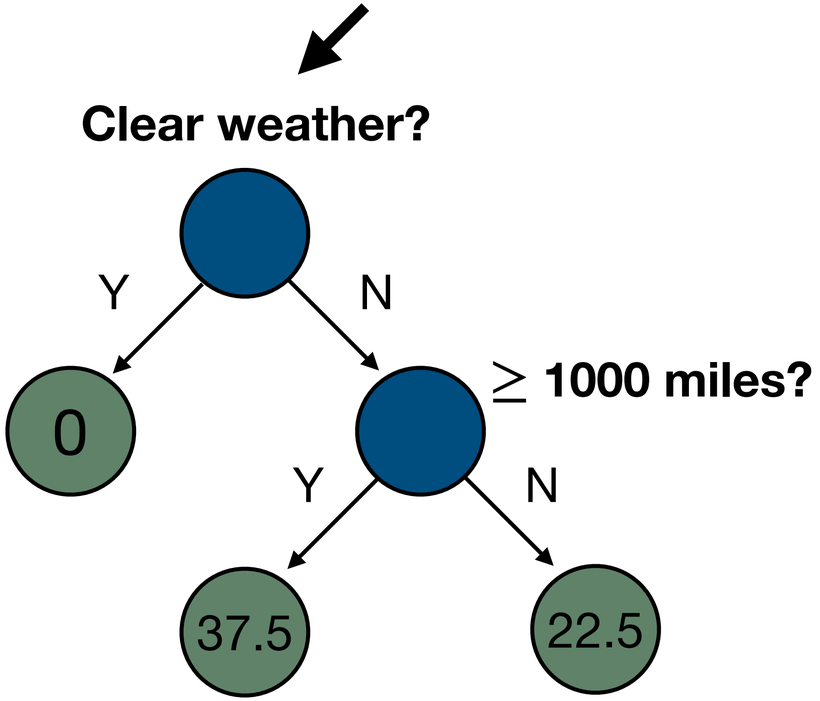
Flavor 2: Regression

Example: Estimating Arrival Times



Distance (miles)	Weather Conditions	Minutes Delayed/ Early
500	Clear	5
750	Rain	20
600	Clear	-5
800	Fog	25
400	Clear	5
1200	Snow	30
950	Clear	-10
1100	Thunderstorms	45

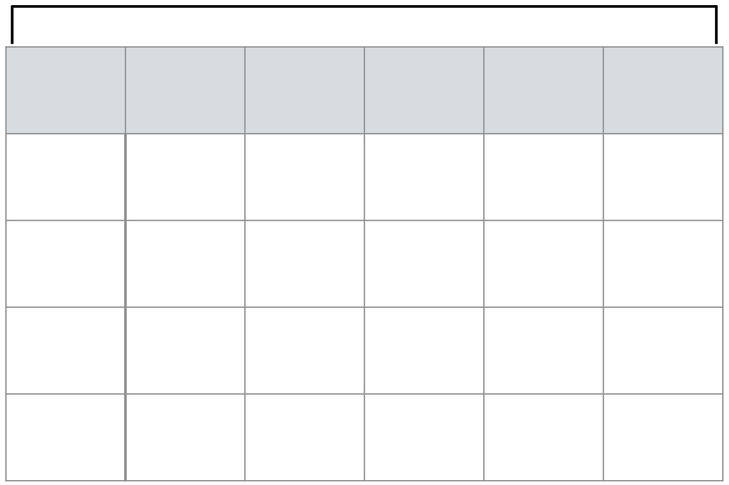




Flavor 3: Clustering

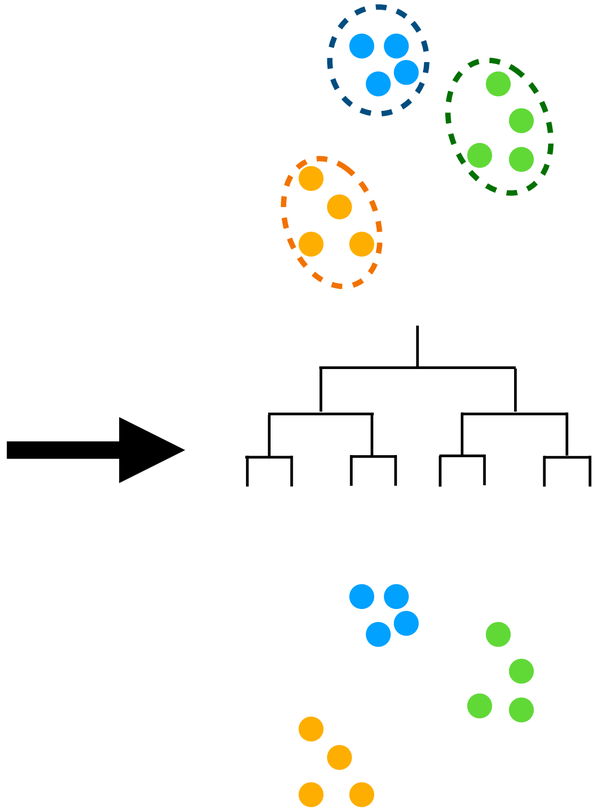
Grouping data based on similarity

Predictors



No target needed!

Training Data



Gaussian Mixture Model

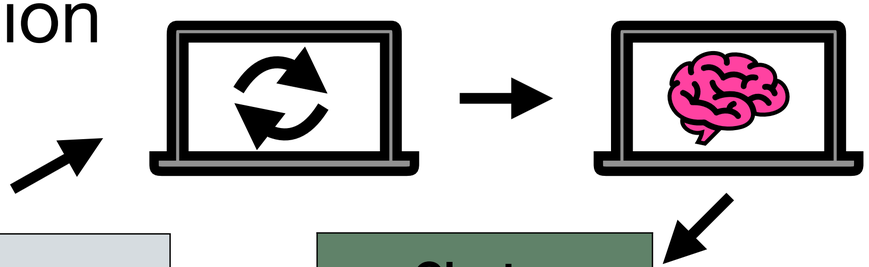
Hierarchical Clustering

K-Means

Techniques

Flavor 3: Clustering

Example: Customer Segmentation



Age	Sex	Country
25	Male	USA
30	Female	Canada
22	Female	UK
28	Male	Australia
35	Female	Germany
40	Male	France
27	Female	USA
33	Male	Canada
29	Female	UK
31	Male	Australia

Cluster
2
1
2
1
3
3
2
1
1
1

1 = Middle-aged, non-European/US

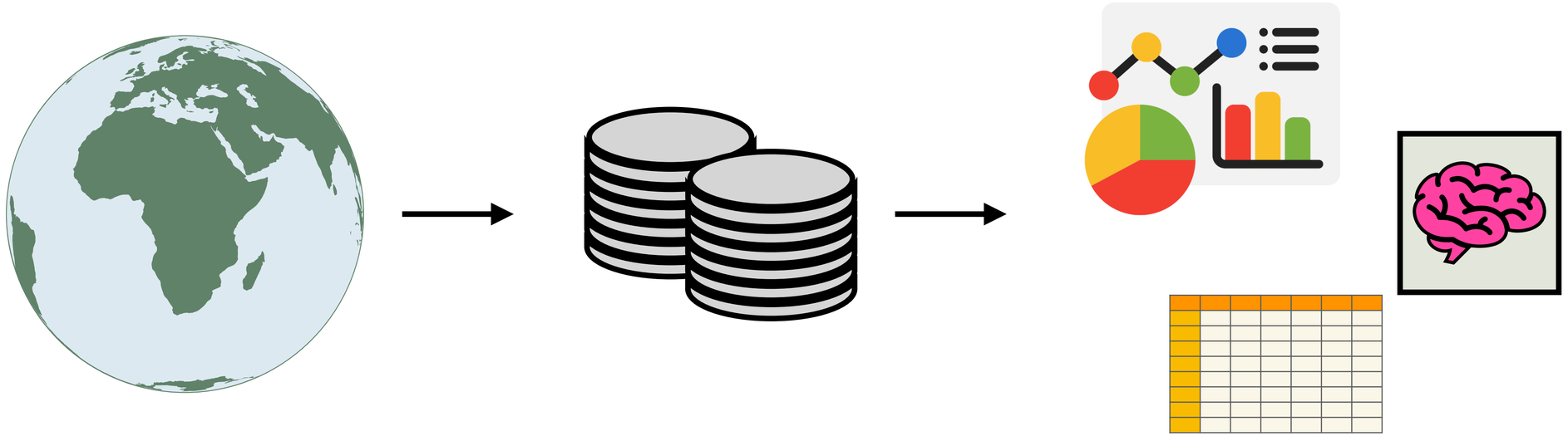
2 = Young, US/UK

3 = Middle-aged, European

Data Engineering

Data Engineering

Making data available for analytics and ML applications



Raw data

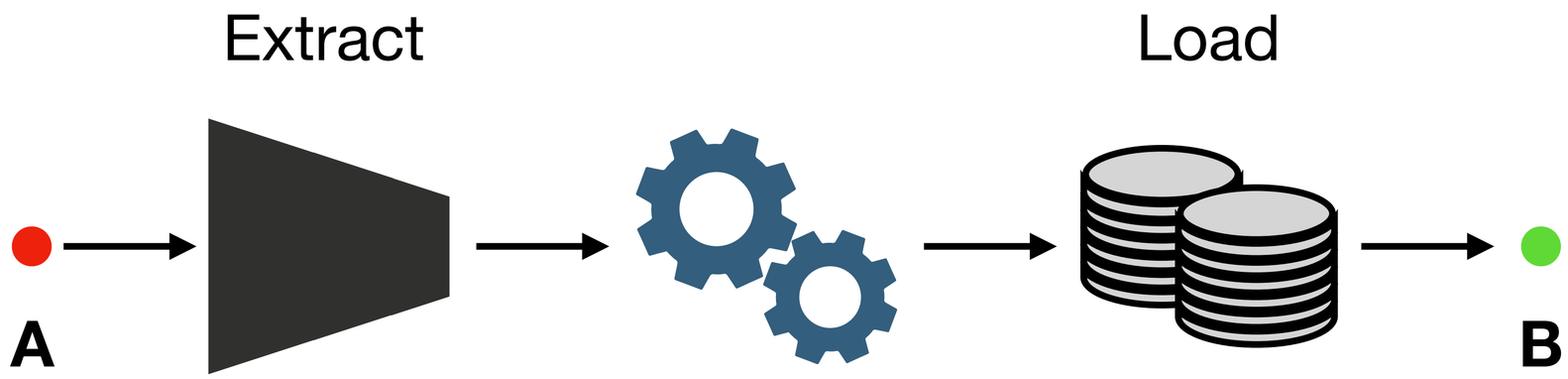
Use case-ready data

Use Cases

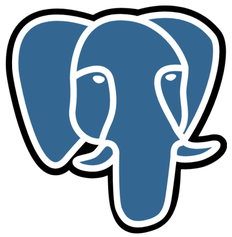
[6] [7]

Data Pipeline

Getting data from point A to point B

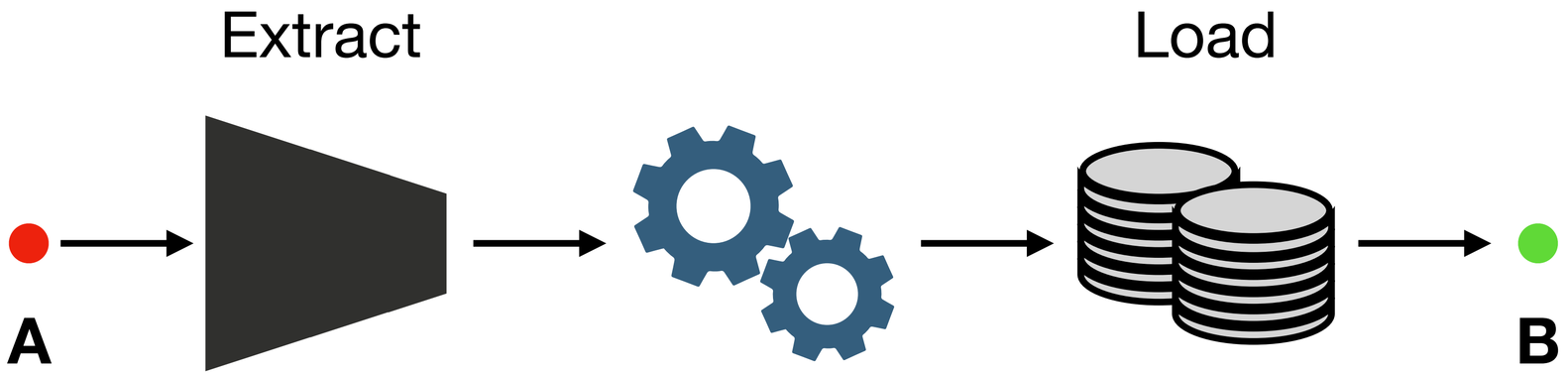


Parsing sales reports and
storing in structured format



Data Pipeline

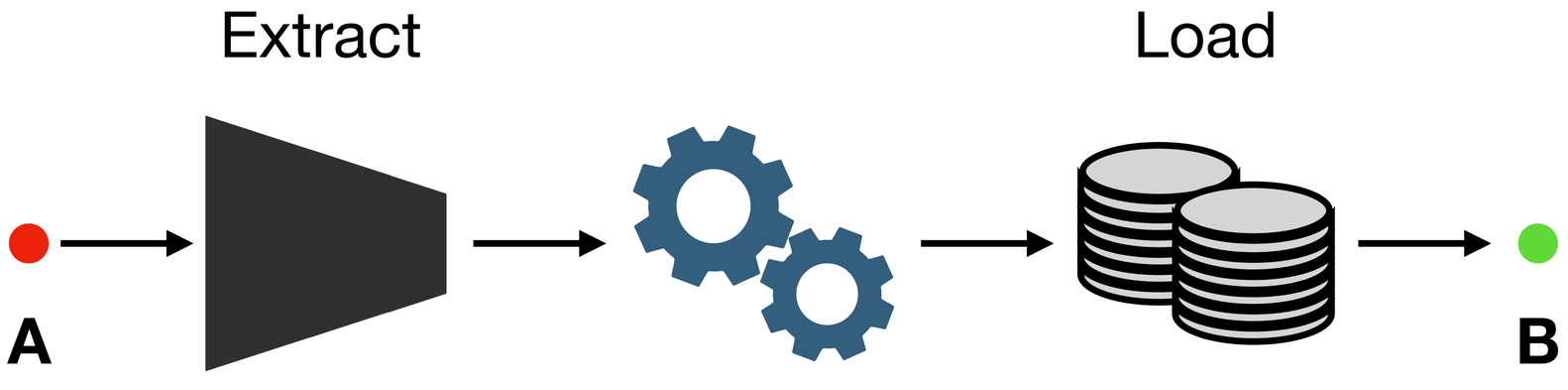
Getting data from point A to point B



Parsing sales emails and
labelling with status

Data Pipeline

Getting data from point A to point B



Grabbing transcriptions and
storing them in vector database

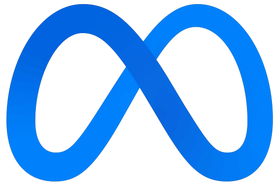
E: Extract

Acquiring data from its source

APIs

→ **Lead Data**

→ **Sales Data**



→ **Social Data**

Custom Extracts

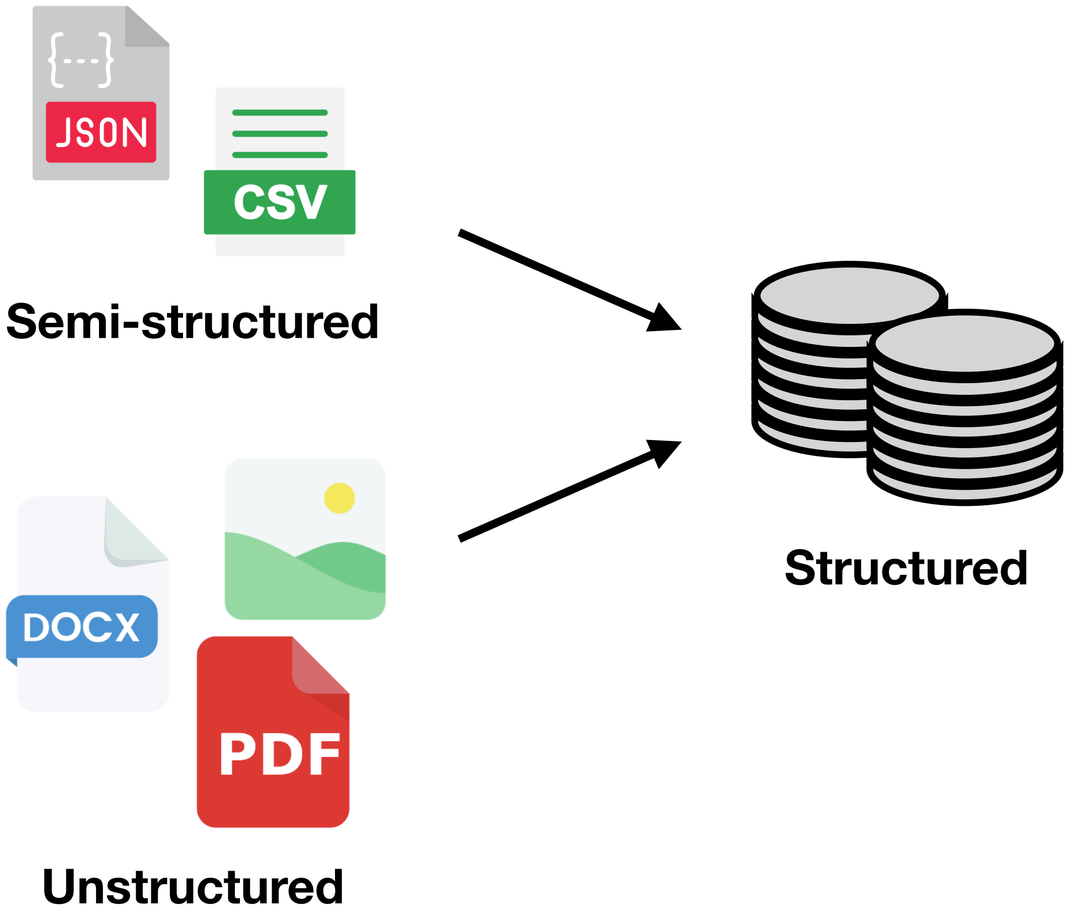
Scraping Public Webpages

Docs from File System

Sensor Data

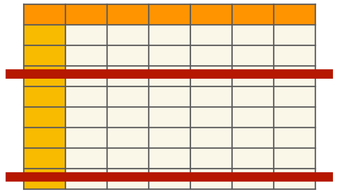
T: Transform

Translating data into a useful form

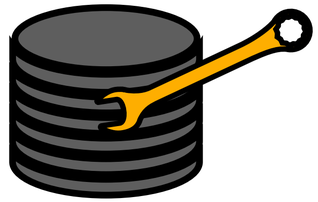


Common Tasks

- Managing data types and ranges



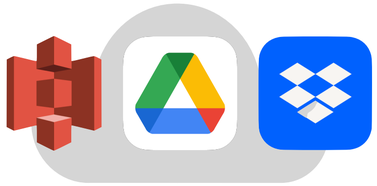
- Deduplication
- Imputing missing values
- Handling special characters and values
- Feature engineering



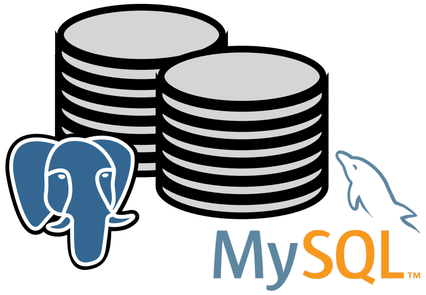
L: Load

Making data available for ML training or inference

Project Directory
MB-scale, few sources, 1 use

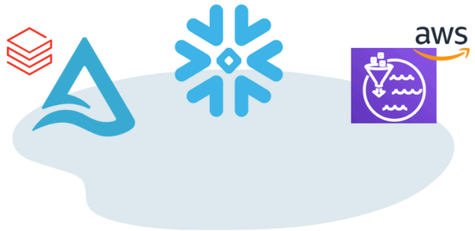


Simple Storage
GB-scale, few sources, few uses



Database
GB-scale, many sources, many uses

Data Warehouse
TB-scale, many sources, many uses

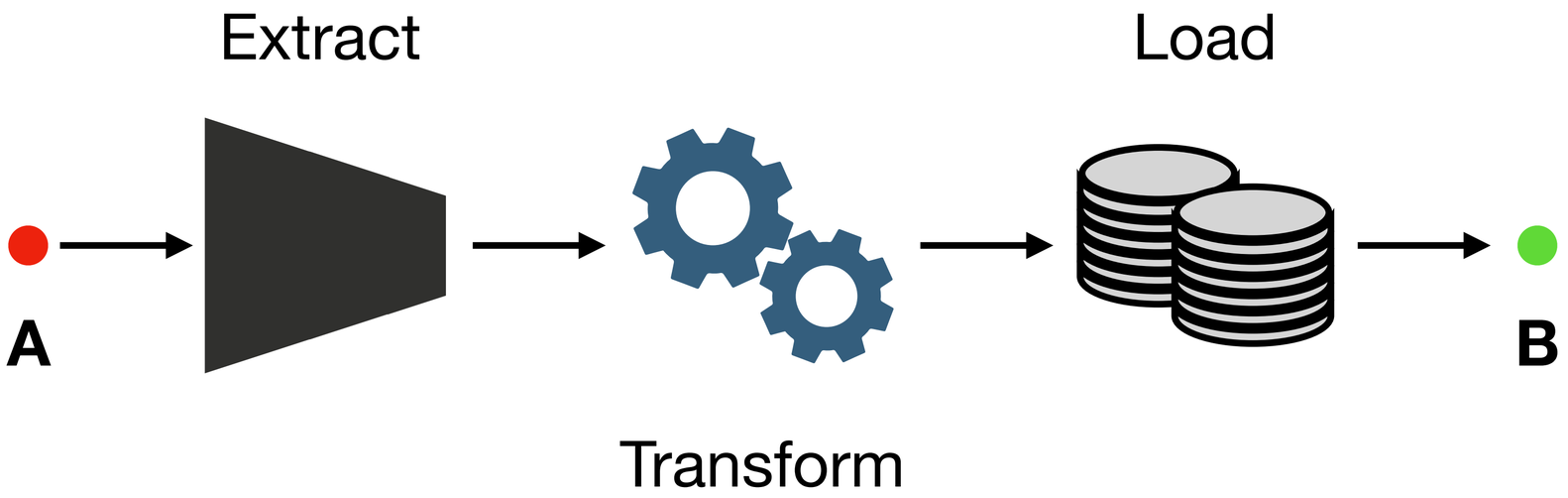


Data Lake
PB-scale, many sources, endless uses

Examples

Example 1

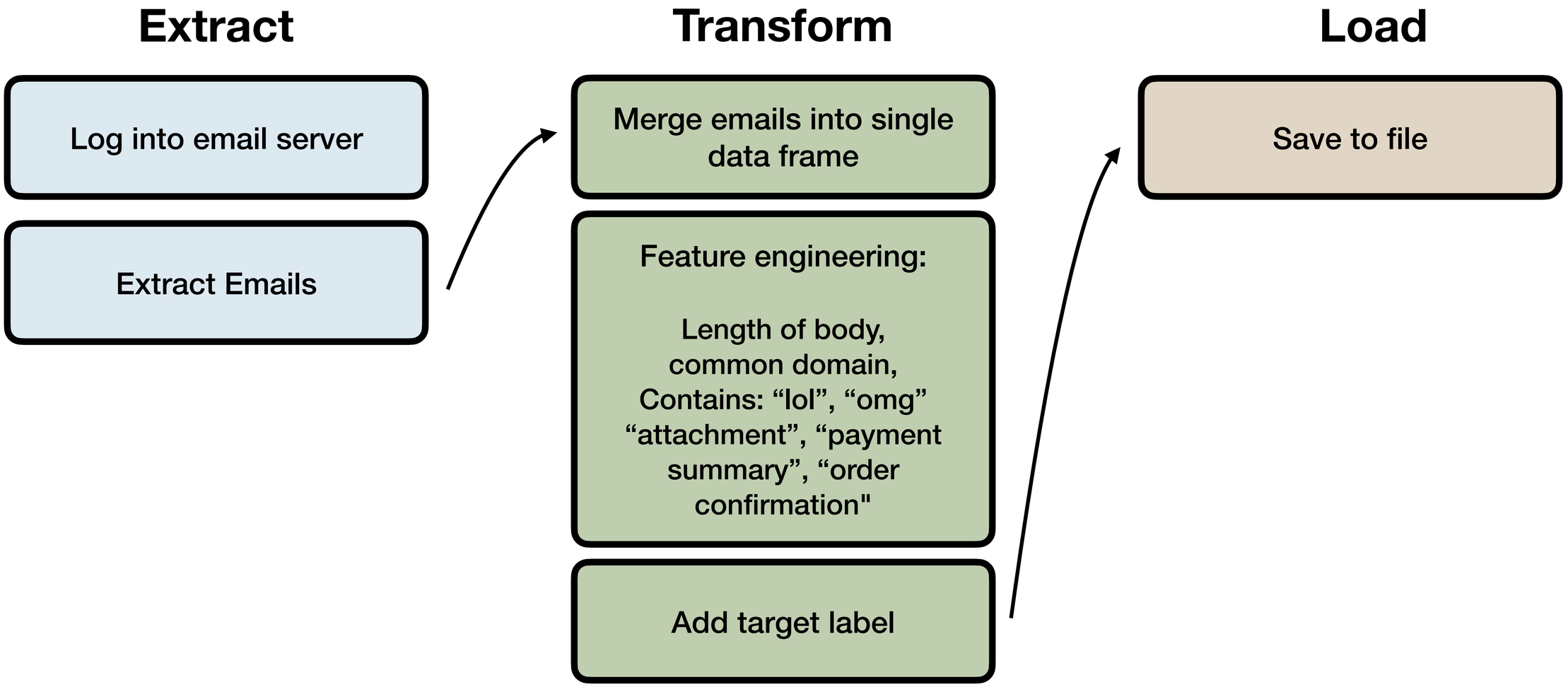
ETL of Gmail Data (Overview)



Feature engineering and labeling personal/not personal emails

Example 1

ETL of Gmail Data (Flowchart)

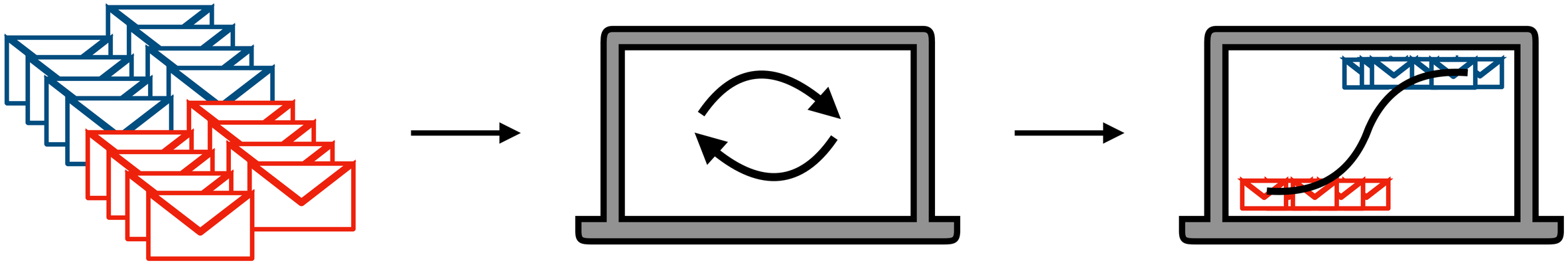


Example 1

ETL of Gmail Data (Example)

Example 2

Training an Email Classifier (Overview)



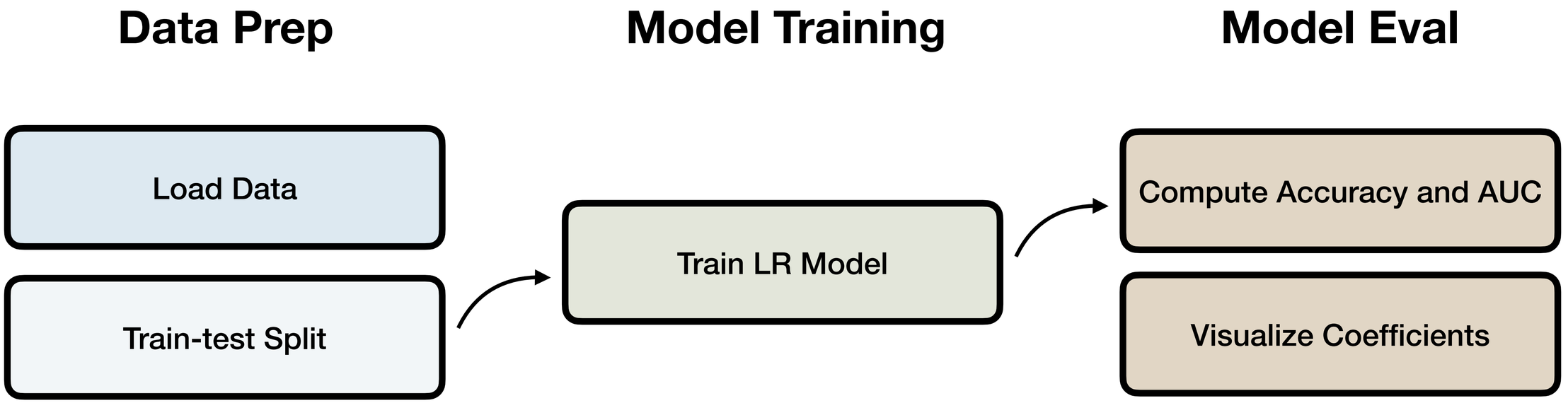
Dataset of personal and not personal emails

Logistic Regression Trainer

Logistic Regression Model

Example 2

Training an Email Classifier (Flowchart)



Example 2

Training an Email Classifier (Example)

Homework 2

Project 💻

Build a Simple ETL Pipeline

Bonus: train a ML model with it!

Pre-work ✍️

Session 3: Introduction to LLMs

Session 3: Prompt Engineering

Session 3: OpenAI API

References

[1] Machine learning: the power and promise of computers that learn by example

[2] sklearn Classifier Comparison

[3] An Introduction to Decision Trees | Gini Impurity & Python Code

[4] sklearn Supervised Learning

[5] sklearn Unsupervised Learning

[6] How Data Engineering Works

[7] How to Build Data Pipelines for ML Projects (w/ Python Code)