



DeepLearning.AI

Data Transformation, Modeling and Serving

**Data Modeling and
Transformation
for Machine Learning**

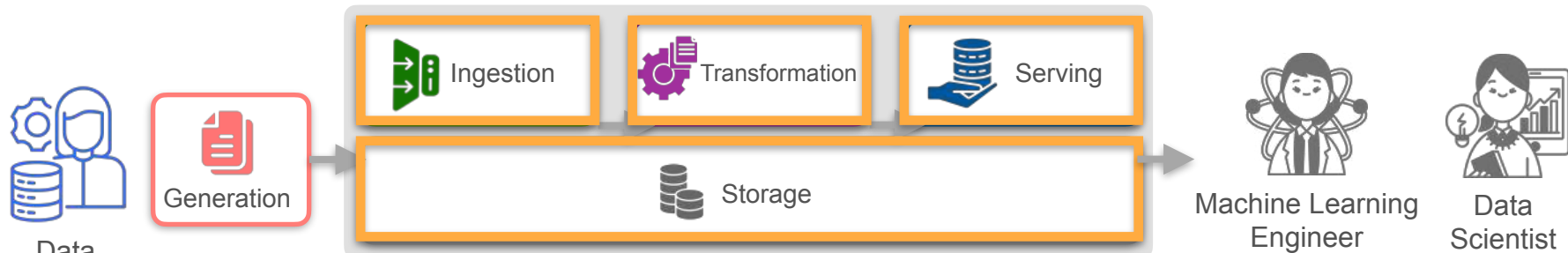


DeepLearning.AI

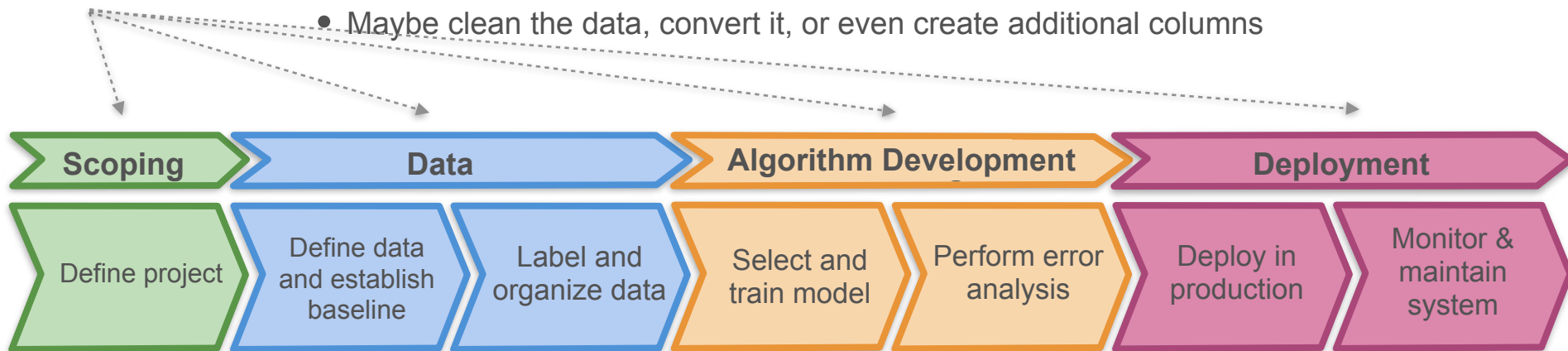
Data Modeling and Transformation for Machine Learning

Week 2 Overview

Data Engineering for Machine Learning

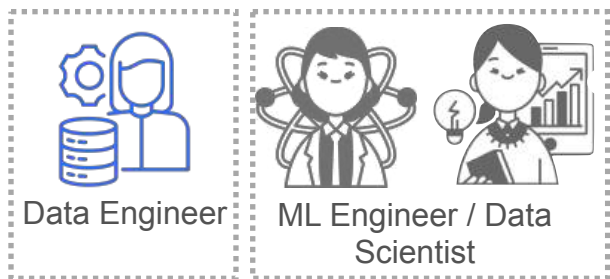


- Shape the data into a format suitable for the ML algorithm
- Maybe clean the data, convert it, or even create additional columns



Machine Learning Project Lifecycle Framework

Data Engineer, Data Scientist and Machine Learning Engineer



Separate teams



Serve raw data

Process the raw data



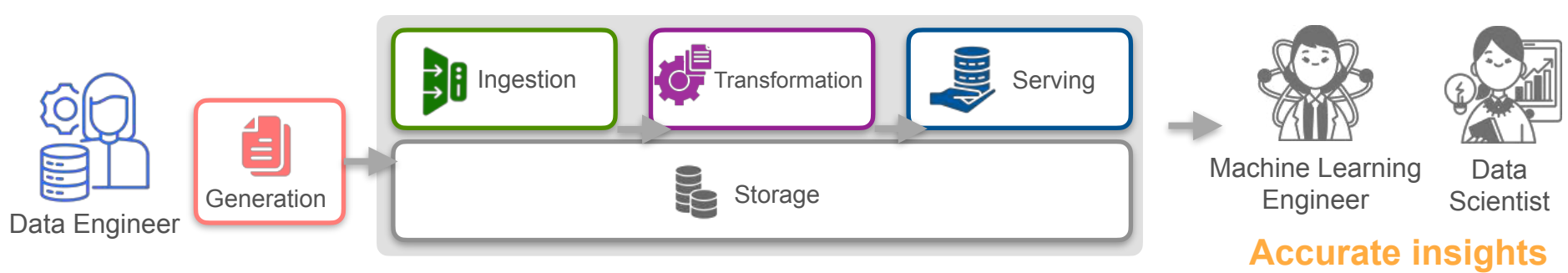
Process the raw data

Use the processed data for model training



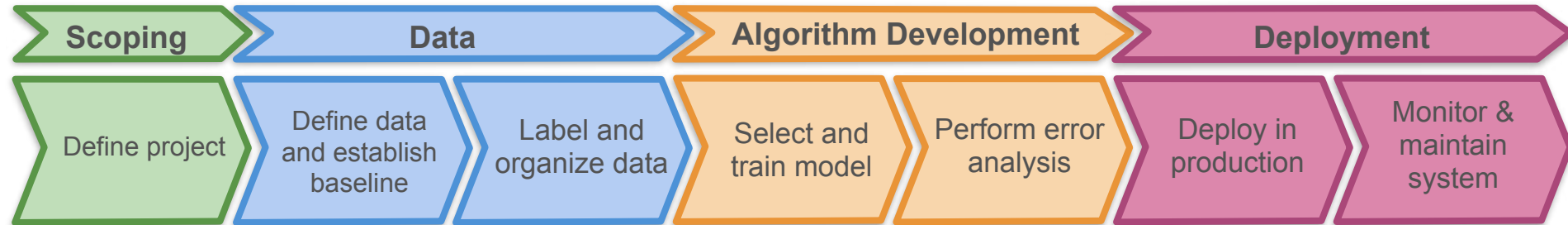
No mature ML team

Handle some extremely ML-specific tasks



Help organization adopt a **data-centric** approach to ML:

- Enhance the ML system by collecting high-quality data
- “Garbage in, garbage out”




Machine Learning Project Lifecycle Framework


This Week's Plan

- ML terminology and the the Machine Learning Project Lifecycle


Tabular

-  How to structure tabular data for classical ML algorithms

Image

-  How to prepare image data for classical and advanced ML algorithms

Text

-  How to preprocess text data and transform text into vectors
(Manually processing text to meet specific cost and system requirements)



DeepLearning.AI

Modeling and Processing Tabular Data for Machine Learning

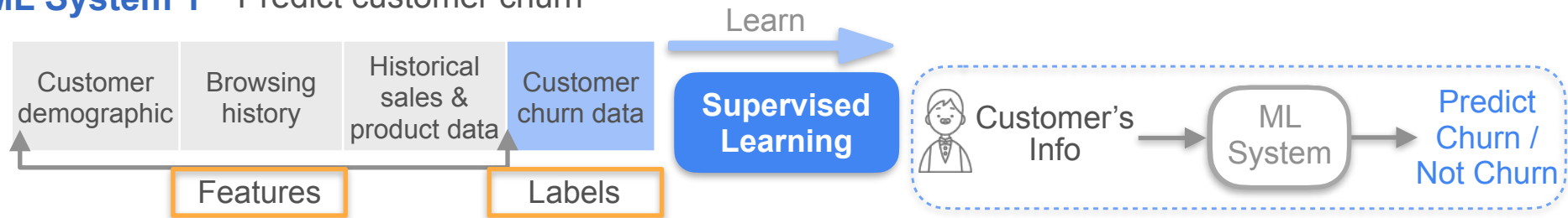
Machine Learning Overview

ML System 1

ML System 2

ML System 3

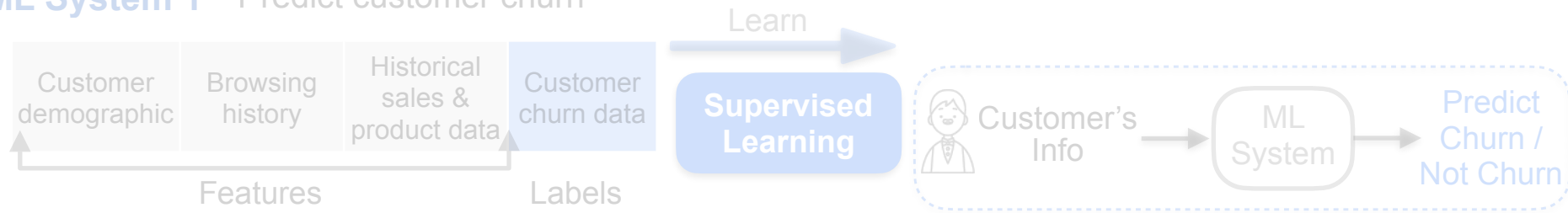
ML System 1 Predict customer churn



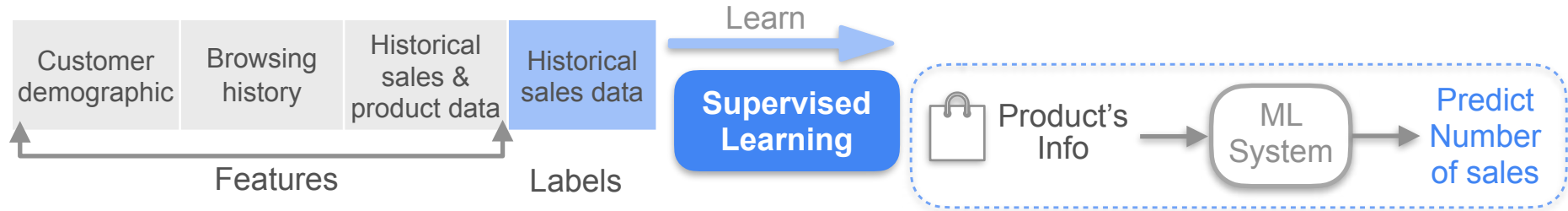
ML System 2

ML System 3

ML System 1 Predict customer churn



ML System 2 Predict the sales for the next new year holiday

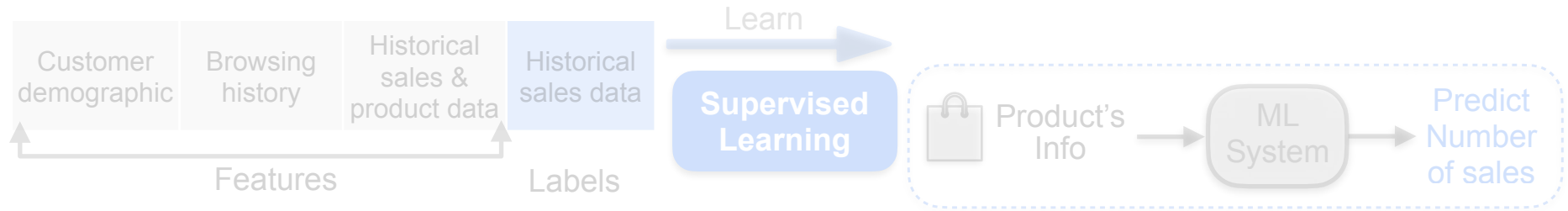


ML System 3

ML System 1 Predict customer churn



ML System 2 Predict the sales for the next new year holiday



ML System 3

ML System 1 Predict customer churn



ML System 2 Predict the sales for the next new year holiday



ML System 3

ML System 1 Predict customer churn



ML System 2 Predict the sales for the next new year holiday



ML System 3 Segment customers into groups based on similar purchasing behaviors



Scoping

Define project

Data

Define data
and establish
baseline

Label and
organize data

Algorithm Development

Select and
train model

Perform error
analysis

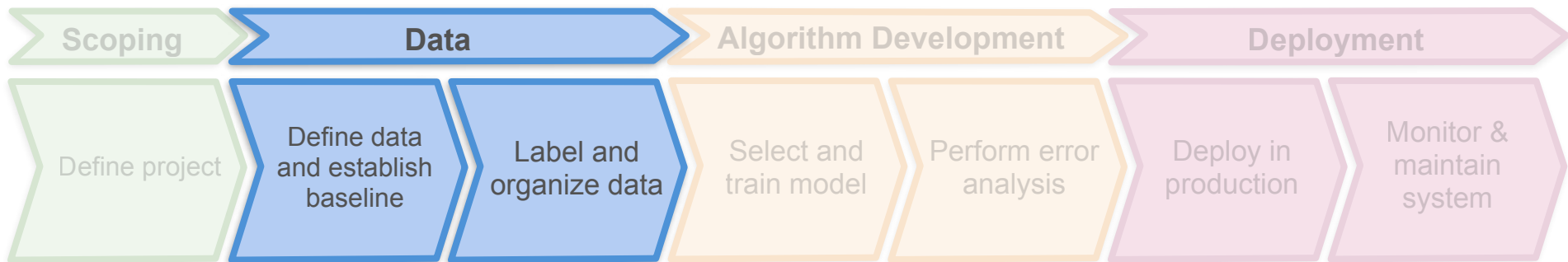
Deployment

Deploy in
production

Monitor &
maintain
system

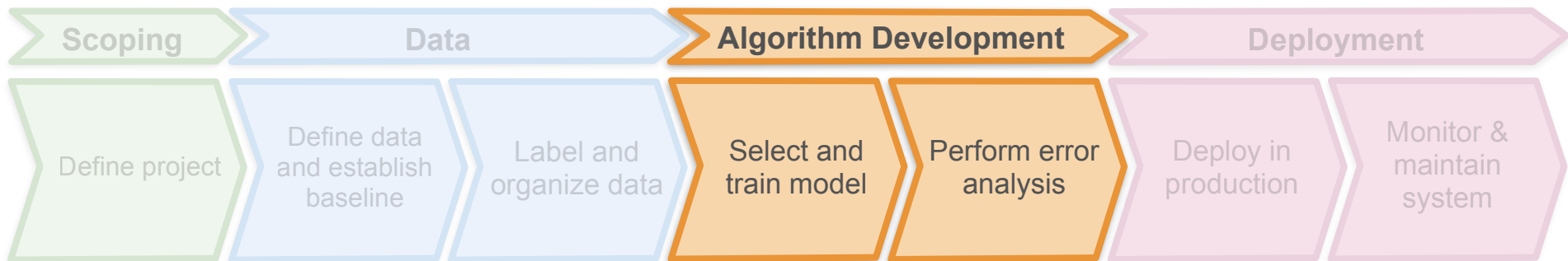


ML Engineer /
Data Scientist



Determine what features and labels you need to collect

Data Engineer



Split the data into:

- training set
- test set



Use the training set to train several ML algorithms

Classical ML algorithms:

- Linear regression
- Logistic regression
- Decision trees
- Random forest and boosted trees

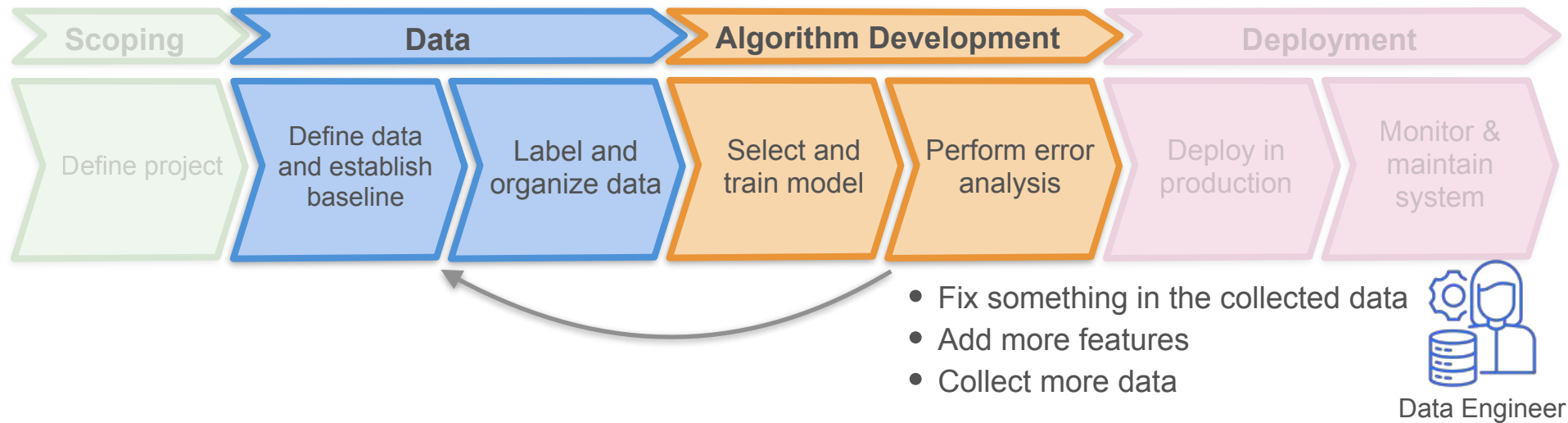
Complex ML algorithms:

- Deep neural networks
- Convolutional neural networks
- Recurrent neural networks
- Large language models

Select the best model through cross-validation



Evaluate the model performance using the test set



Split the data into:

- training set
- test set



Use the training set to train several ML algorithms

Classical ML algorithms:

- Linear regression
- Logistic regression
- Decision trees
- Random forest and boosted trees

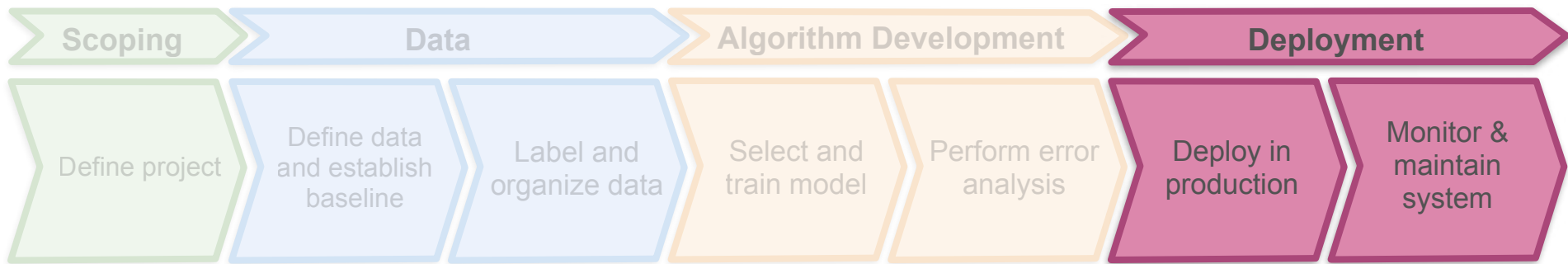
Complex ML algorithms:

- Deep neural networks
- Convolutional neural networks
- Recurrent neural networks
- Large language models

Select the best model through cross-validation



Evaluate the model performance using the test set



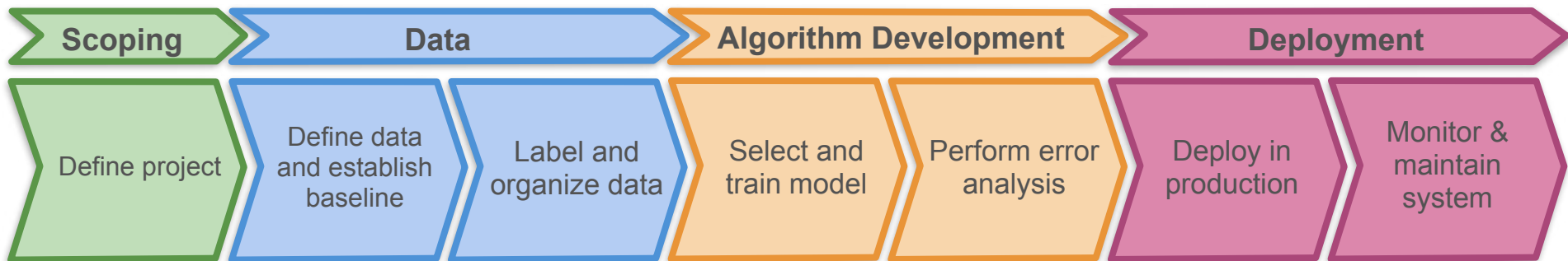
ML Engineer /
Data Scientist

- Check to make sure the system's performance is good and reliable
- Write software to put system into production
- Monitor the system, track data, and maintain system



Data Engineer

- Prepare and serve the data that is needed for the deployed model
- Serve an updated set of data to re-train and update the model



Data Engineer

Set up the pipeline to serve data that supports these phases



DeepLearning.AI

Modeling and Processing Tabular Data for Machine Learning

Modeling Data for Traditional Machine Learning Algorithms

Numerical Tabular Form

Customer	No. of items purchased	Date of last purchase	Customer income	Minutes on Platform	Account type	Churned
	14	7/5/2024	\$50,000	15	Family	No
	9	3/4/2024	\$40,000	13	Platinum	Yes
	Null	8/12/2024	Null	Null	Null	Yes
	2	8/24/2024	Null	35	Basic	No

Numerical Tabular Form

What most classical ML algorithms expect as training data

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Purchases per minute	Churned
0.93	0.90	0.5	0.24	1	0.93	0
0.57	0.29	0.4	0.20	2	0.69	1 Churn
0.07	0.03	0.35	0.64	0	0.06	0 Not Churn
...

Features **Labels**

- No missing values or duplicate rows
- Each column consists of numerical values that are within a similar range

Feature Engineering

Feature Engineering

Any change or processing done to a raw column, and any creation of new features

- Handling missing values
- Feature scaling
- Converting categorical columns into numerical ones
- Creating new columns by combining or modifying existing ones

Handling Missing Values

Understand why the values are missing and then determine the most appropriate way

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Churned
14	28	\$50,000	15	Family	No
9	9	\$40,000	13	Platinum	Yes
Null	12	Null	Null	Null	Yes
2	1	Null	35	Basic	No
...

- Delete the entire column or row (if there's no risk of losing valuable data)
- Impute the missing values with summary statistics
 - Replace missing values with the column mean or median
 - Replace missing values with values from a similar record

Handling Missing Values

Understand why the values are missing and then determine the most appropriate way

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Churned
14	28	\$50,000	15	Family	No
9	9	\$40,000	13	Platinum	Yes
2	1	\$35,000	35	Basic	No
...

- Delete the entire column or row (if there's no risk of losing valuable data)
- Impute the missing values with summary statistics
 - Replace missing values with the column mean or median
 - Replace missing values with values from a similar record

Scaling Numerical Features

Scale features so that the values of each feature end up within a similar range

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Churned
14	28	\$50,000	15	Family	No
9	9	\$40,000	13	Platinum	Yes
2	1	\$35,000	35	Basic	No
...

- Training an ML algorithm is based on solving an optimization problem:
 - If values vary drastically → take longer for the optimization algorithm to converge
- Certain ML algorithms are based on distance metrics:
 - Their accuracies can be affected by different ranges of values

Scaling Numerical Features

Scale features so that the values of each feature end up within a similar range

No. of items purchased	Days since last purchase	Customer income
14	28	0.5
9	9	0.4
2	1	\$35,000
...

$$\frac{\$50,000 - \$0}{\$100,000 - \$0} = 0.5$$

$$\frac{\$40,000 - \$0}{\$100,000 - \$0} = 0.4$$

Standardization

value - column mean

column standard deviation

Resulting value has mean of 0 and variance of 1

Min: \$0
Max: \$100,000

Min-Max Scaling

value - column min

column max - column min

Resulting value is between 0 and 1

Converting Categorical Columns into Numerical Ones

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Churned
14	28	0.5	15	Family	No
9	9	0.4	13	Platinum	Yes
2	1	0.35	35	Basic	No
...

One Hot Encoding

Account type	Basic	Family	Platinum
Family	0	1	0
Platinum	0	0	1
Basic	1	0	0
...

Ordinal Encoding

Account type	Account type
Family	2
Platinum	3
Basic	1
...	...

middle
most
expensive
cheapest

Embeddings

More on
this later



DeepLearning.AI

Modeling and Processing Tabular Data for Machine Learning

Processing Tabular Data for Classical Machine Learning Algorithms Using Scikit-Learn (Part 1)

scikit-learn

Machine Learning in Python

[Getting Started](#)
[Release Highlights for 1.5](#)

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts

Two Processing Methods:

- Standardization for the numerical columns
- One-hot encoding for the categorical columns

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [logistic regression](#), and [more...](#)

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, stock prices.
Algorithms: [Gradient boosting](#), [nearest neighbors](#), [random forest](#), [ridge](#), and [more...](#)

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, grouping experiment outcomes.
Algorithms: [k-Means](#), [HDBSCAN](#), [hierarchical clustering](#), and [more...](#)

Download Dataset

CustomerID	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
1	22	25	14	4	27	Basic	Monthly	598	9	1
2	41	28	28	7	13	Standard	Monthly	584	20	0
3	47	27	10	2	29	Annual	Annual	757	21	0
...

<https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset>

Preparing Data for Training a Machine Learning Model



ML Engineer Team

Dataset

CustomerID	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
1	22	25	14	4	27	Basic	Monthly	598	9	1
2	41	28	28	7	13	Standard	Monthly	584	20	0
3	47	27	10	2	29	Annual	Annual	757	21	0
...

80%

20%

Training Dataset

Customer_id

Standardized numerical columns

One-hot encoded categorical columns

Testing Dataset

Customer_id

Standardized numerical columns

One-hot encoded categorical columns



Preparing Data for Training a Machine Learning Model

1. Split the data into training and test sets

2. Process the training data

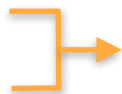
- a. Numerical columns → standardize
- b. Categorical columns → one hot encoding
- c. Combine processed columns with the Customer ID into a Pandas data frame
- d. Convert Pandas data frame into a parquet file



Makes it easier

3. Process the test data

- a. Numerical columns → standardize
- b. Categorical columns → one hot encoding
- c. Combine processed columns with the Customer ID
- d. Convert Pandas data frame into a parquet file



Use the same computed statistics used on the training set



DeepLearning.AI

Modeling and Processing Tabular Data for Machine Learning

Processing Tabular Data for Classical Machine Learning Algorithms Using Scikit-Learn (Part 2)

Preparing Data for Training a Machine Learning Model

1. Split the data into training and test sets

2. Process the training data

- a. Numerical columns → standardize
- b. Categorical columns → one hot encoding
- c. Combine processed columns with the Customer ID into a Pandas data frame
- d. Convert Pandas data frame into a parquet file

3. Process the test data

- a. Numerical columns → standardize
- b. Categorical columns → one hot encoding
- c. Combine processed columns with the Customer ID
- d. Convert Pandas data frame into a parquet file



DeepLearning.AI

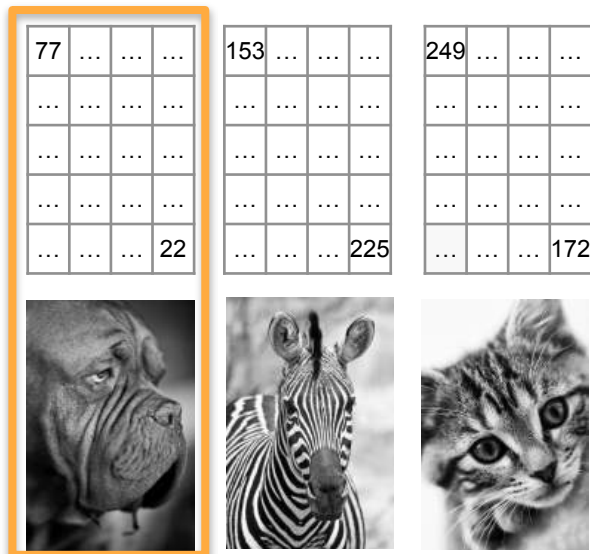
Modeling and Processing Unstructured Data for Machine Learning

Modeling Image Data for Machine Learning Algorithms

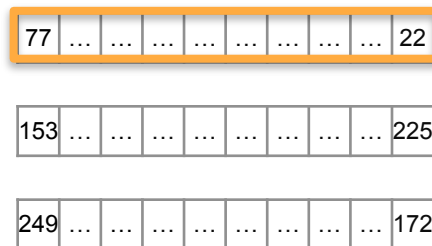
Training an ML Algorithm on Image Data

Traditional ML algorithms

No. of items purchased	Days since last purchase	Customer income	Minutes on Platform	Account type	Purchases per minute	Churned
0.93	0.90	0.5	0.24	1	0.93	0
0.57	0.29	0.4	0.20	2	0.69	1
0.07	0.03	0.35	0.64	0	0.06	0
...



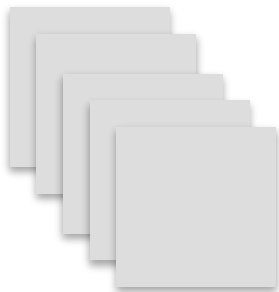
→
Train a traditional ML algorithm



- Lose spatial information that can be extracted from the relative location of pixels
- Can create a high-dimensional vector of features
 - e.g. 1000 pixels by 1000 pixels → vector of size 1 million
- Affect the performance of the ML algorithm

Training an ML Algorithm on Image Data

Convolutional Neural Network (CNN)



Each layer tries to identify more image features to help with the ML task

- First layer: Generic features
- Later layers: Complex patterns and textures



ML Engineer Team

- Start with pre-trained CNN algorithms
- Fine tune these models for the specific task

Preparing Image Data for the Training an ML Algorithm



Resizing



Scaling the pixels



Flipping



Rotating



Cropping



Adjusting brightness



**Data
Augmentation**

Technique used to create new versions of existing images
(Increases the size & variety of training data)



Data Engineer

TensorFlow



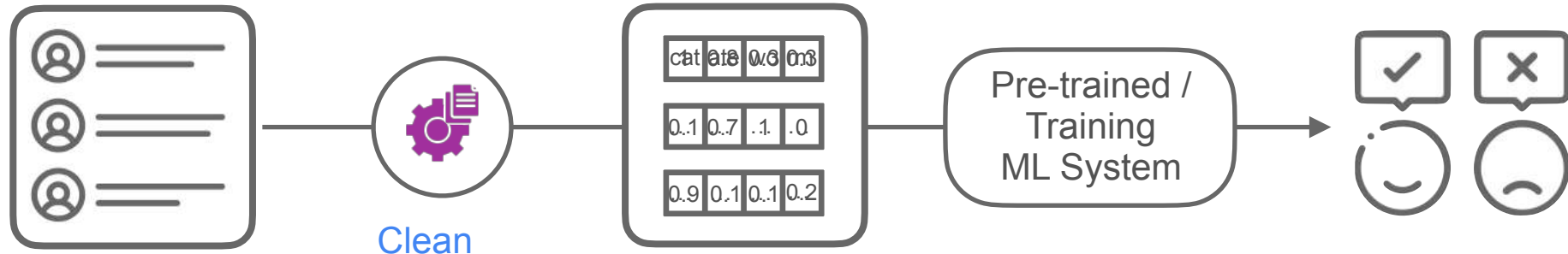
DeepLearning.AI

Modeling and Processing Unstructured Data for Machine Learning

Preprocessing Textual Data for Analysis and Text Classification

Pre-processing Texts for ML

Sentiment Analysis



Customer Reviews



This is a wonderful
price for the amount
#@% you get



Depends on the type of ML algorithm:

- **Classical ML algorithm** requires numerical data
- **LLMs** can work with tokens or words

Pre-processing Texts for ML



Textual data might contain
typos, inconsistencies, &
repetitions

May contain words or
characters not relevant
to the NLP task

Training LLMs is
expensive and time
consuming



Data Engineer



Clean and
high-quality
data

Remove any
irrelevant
words or
characters



ML Engineer Team

Train a classical or
advanced ML model

Other use cases

Processing Text

Cleaning

Removing punctuations, extra spaces, characters that add no meaning

Normalization

Tokenization

Removal of
Stop Words

Lemmatization

Reviews

This is a wonderful price for the
amount #@% you get

Great product! Big amt

I bought this for my son as his hair is
thinning. I don't know yet how well is
helping. He said the smell is great.

Cleaned Reviews

This is a wonderful price for the
amount you get

Great product Big amt

I bought this for my son as his hair is
thinning I don't know yet how well is
helping He said the smell is great

Processing Text

Cleaning

Converting texts to consistent format:

- Transforming to lower-case
- Converting numbers or symbols to characters
- Expanding contractions

kg → kilograms

lbs → pounds

DE / D.E → data engineering

Normalization

Tokenization

Removal of
Stop Words

Lemmatization

Cleaned Reviews

This is a wonderful price for the amount you get

Great product Big **amt** → amount

I bought this for my son as his hair is thinning I **don't** know yet how well is helping He said the smell is great

do not

Normalized Reviews

this is a wonderful price for the amount you get

great product big amount

i bought this for my son as his hair is thinning i do not know yet how well is helping he said the smell is great

Processing Text

Cleaning

Normalization

Tokenization

Removal of
Stop Words

Lemmatization

Splitting each review into individual tokens
(words, subwords, short sentences)

Normalized Reviews

this is a wonderful price for the amount
you get

great product big amount

i bought this for my son as his hair is
thinning i do not know yet how well is
helping he said the smell is great

Tokenized Reviews

[this, is, a, wonderful, price, for, the,
amount, you, get]

[great, product, big, amount]

[i, bought, this, for, my, son, as, his,
hair, is, thinning, i, do, not, know, yet,
how, well, is, helping, he, said, the,
smell, is, great]

Processing Text

Cleaning

- Removing frequently used words such as “is”, “are”, “the”, “for”, “a”
- Define your own list of stop words
- Or use built-in set of NLP libraries

spaCy



NLTK



Gensim



TextBlob

Normalization

Tokenization

Removal of
Stop Words

Lemmatization

Tokenized Reviews

[this, is, a, wonderful, price, for, the, amount, you, get]

[great, product, big, amount]

[i, bought, this, for, my, son, as, his, hair, is, thinning, i, do, not, know, yet, how, well, is, helping, he, said, the, smell, is, great]

Stop Words Removed

[this, wonderful, price, amount, you, get]

[great, product, big, amount]

[i, bought, this, my, son, his, hair, thinning, i, do, not, know, yet, how, well, helping, he, said, smell, great]

Stop words: {is, a, for, the, as, are}

Processing Text

Cleaning

Replacing each word with its base form or lemma (using NLP libraries)

getting / got → get

Normalization

Tokenization

Removal of
Stop Words

Lemmatization

Stop Words Removed

[this, wonderful, price, amount, you, get]

[great, product, big, amount]

[i, bought, this, my, son, his, hair, thinning, i, do, not, know, yet, how, well, helping, he, said, smell, great]

Tokenized and Lemmatized Reviews

[this, wonderful, price, amount, you, get]

[great, product, big, amount]

[i, buy, this, my, son, his, hair, thin, i, do, not, know, yet, how, well, help, he, say, smell, great]



DeepLearning.AI

Modeling and Processing Unstructured Data for Machine Learning

Text Vectorization and Embedding

Traditional Vectorization

Bag of Words

Term-Frequency Inverse-Document-Frequency (TF-IDF)

The corpus

Reviews	
[this, wonderful, price, amount, you, get]	→ A document
[great, product, big, amount]	
[I, buy, this, my, son, his, hair, thin, I, do, not, know, yet, how, well, help, he, say, smell, great]	

The vocabulary

[this, wonderful, price, amount, you, get, great, product, big, I, buy, my, son, his, hair, thin, do, not, know, yet, how, well, help, he, say, smell]



this	wonderful	price	amount	you	get	great	product	big	I	buy	my	son	his	hair	thin	do	not	not	know	yet	how	well	help	he	say	smell

Example

“purchase”
“buy” → High frequency,
little meaning

“break”
“exceptional” → Low frequency,
more significant

Bag of Words

Term-Frequency Inverse-Document-Frequency (TF-IDF)

Each entry: number of occurrences

- Only takes into account the word frequency in each document
- Some frequently appearing words might carry little meaning



The
corpus



Reviews									
[this, wonderful, price, amount, you, get]					→ A document				
[great, product, big, amount]									
[I, buy, this, my, son, his, hair, thin, I, do, not, know, yet, how, well, help, he, say, smell, great]									

this	wonderful	price	amount	you	get	great	product	big	I	buy	my	son	his	hair	thin	do	not	not	know	yet	how	well	help	he	say	smell
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	1	0	0	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Bag of Words

Term-Frequency Inverse-Document-Frequency (TF-IDF)

Account for the weight and rarity of each word

TF: the number of times the term occurred in a document divided by the length of that document

IDF: how common or rare that word is in the entire corpus.



The
corpus



Reviews	
[this, wonderful, price, amount, you, get]	→ A document
[great, product, big, amount]	
[I, buy, this, my, son, his, hair, thin, I, do, not, know, yet, how, well, help, he, say, smell, great]	

Bag of Words

Term-Frequency Inverse-Document-Frequency (TF-IDF)

Account for the weight and rarity of each word

TF: the number of times the term occurred in a document divided by the length of that document

IDF: how common or rare that word is in the entire corpus.



The corpus



Reviews									
[this, wonderful, price, amount, you, get] → A document									
[great, product, big, amount]									
[I, buy, this, my, son, his, hair, thin, I, do, not, know, yet, how, well, help, he, say, smell, great]									

this	wonderful	price	amount	you	get	great	product	big	I	buy	my	son	his	hair	thin	do	not	not	know	yet	how	well	help	he	say	smell
0.33	0.44	0.44	0.33	0.44	0.44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.43	0	0	0.43	0.56	0.56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.16	0	0	0	0	0	0.16	0	0	0.43	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22	0.22

Word Embedding

Word

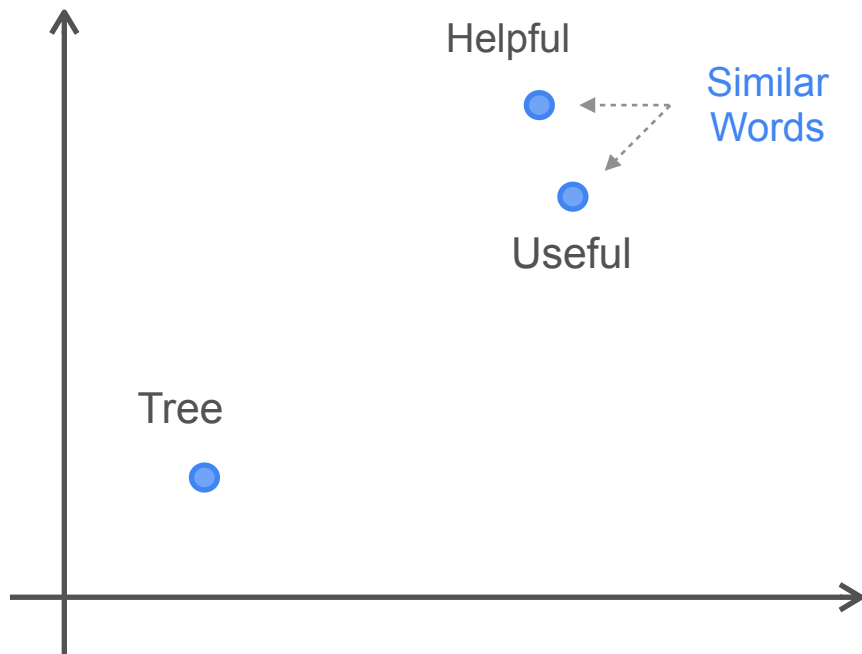
word2vec, GLOVE

Trained to learn the embeddings of words from their co-occurrences



Word Embedding

Vector that captures the semantics meaning of the word



Word Embedding

Reviews	
[this, wonderful, price, amount, you, get]	
[great, product, big, amount]	
[I, buy, this, my, son, his, hair, thin, I, do, not, know, yet, how, well, help, he, say, smell, awesome]	

▼ **Represent each review, not just each word, by one vector**

Word Embedding (great)

+

Word Embedding (product)

+

Word Embedding (big)

+

Word Embedding (amount)



Vector that represents the sentence

Does not account for the position of the words in the sentence

“A man ate a snake” \neq “A snake ate a man”

Sentence Embedding

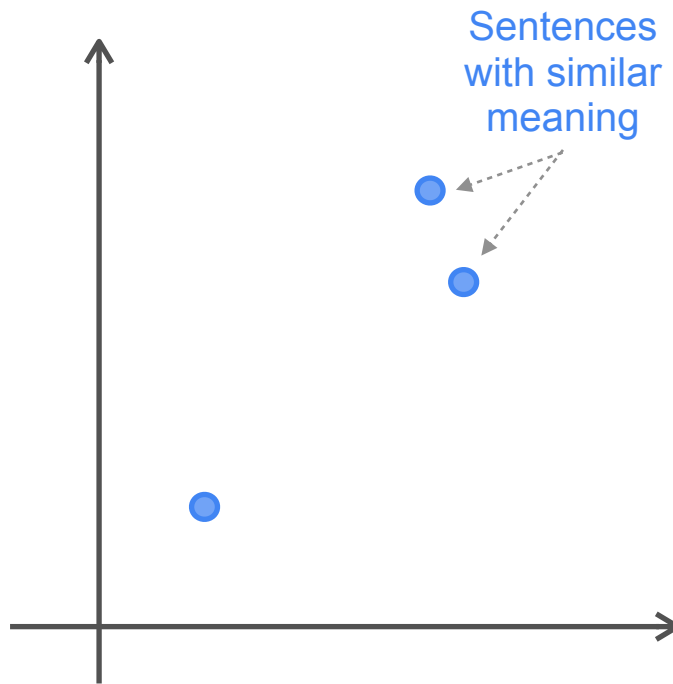
Sentence



Pre-trained NLP models

Sentence Embedding

- Vector that reflects the semantic meaning of the sentence
- Lower dimension than the vector generated by TF-IDF



Sentence Embedding

Sentence

Pre-trained NLP models

based on Large Language Models (LLM)

Open-Source



Sentence Transformers

Closed-Source



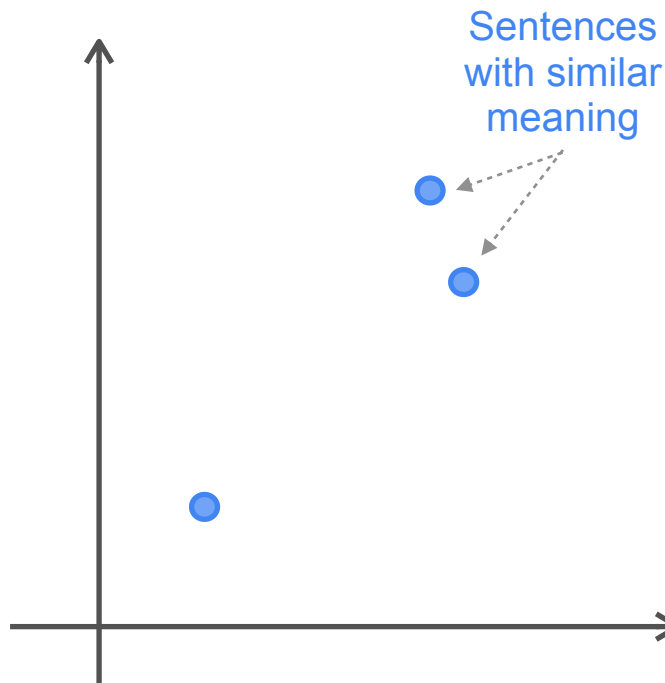
OpenAI

AI

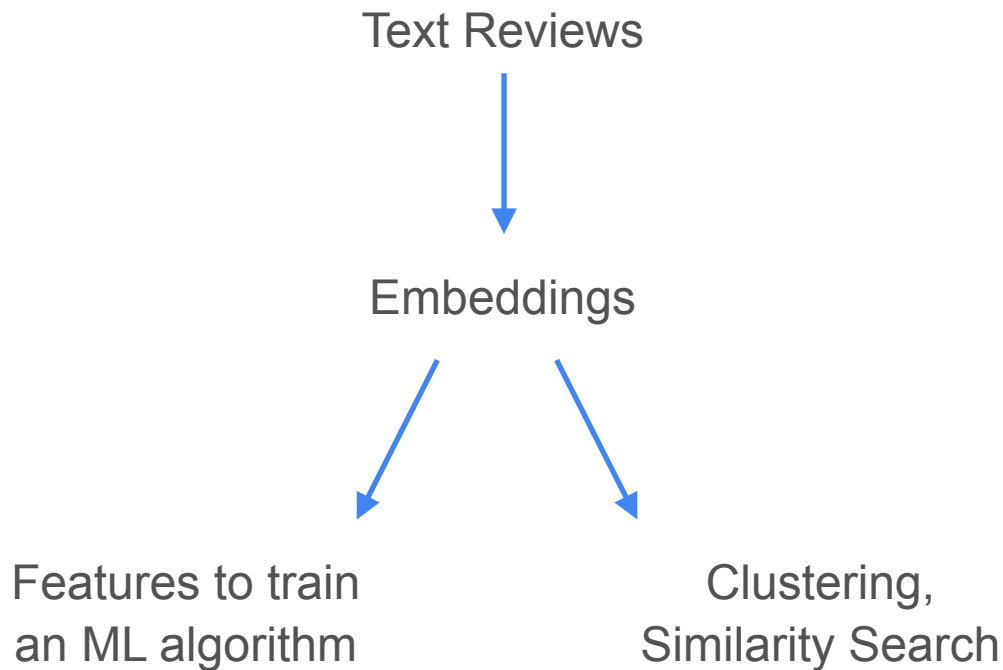
Gemini

Sentence Embedding

- Vector that reflects the semantic meaning of the sentence
- Lower dimension than the vector generated by TF-IDF



Sentence Embeddings



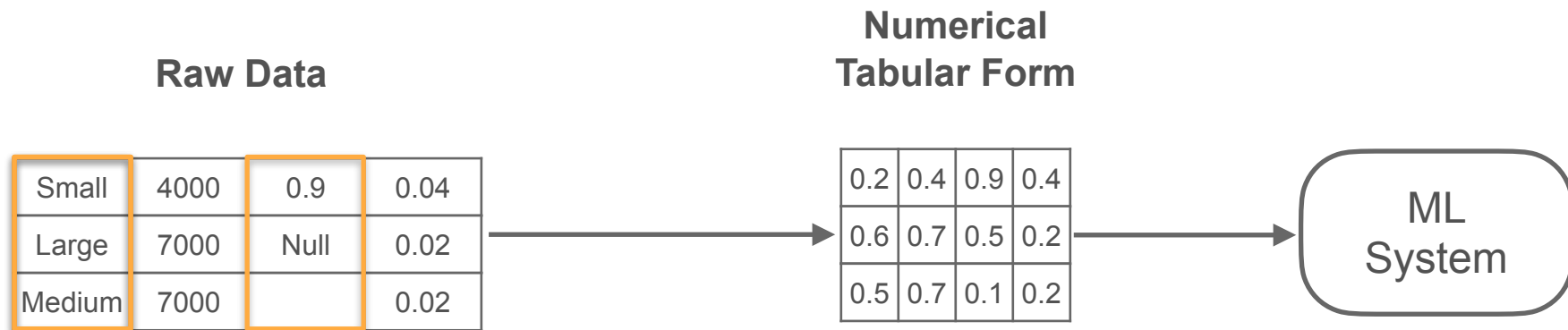


DeepLearning.AI

Data Modeling and Transformation for Machine Learning

Week 2 Summary

Tabular Data for Training an ML Algorithm



1. Impute data or delete empty records/columns
2. Convert categorical columns to numerical ones
 - One hot encoding, Ordinal encoding, etc.
3. Scale the numerical features

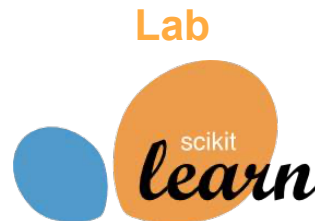


Image Data for Training an ML Algorithm



Classical ML algorithms

Unroll the images in a long sequence of pixels

Convolutional neural networks

- Pre-processing techniques:
 - Image reshaping
 - Image normalization
- Data augmentation:
 - Flipping
 - Rotation
 - Adding distortions

Textual Data for Training an ML Algorithm

