Data Engineering Project

Module 6
Machine Learning Pipeline

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

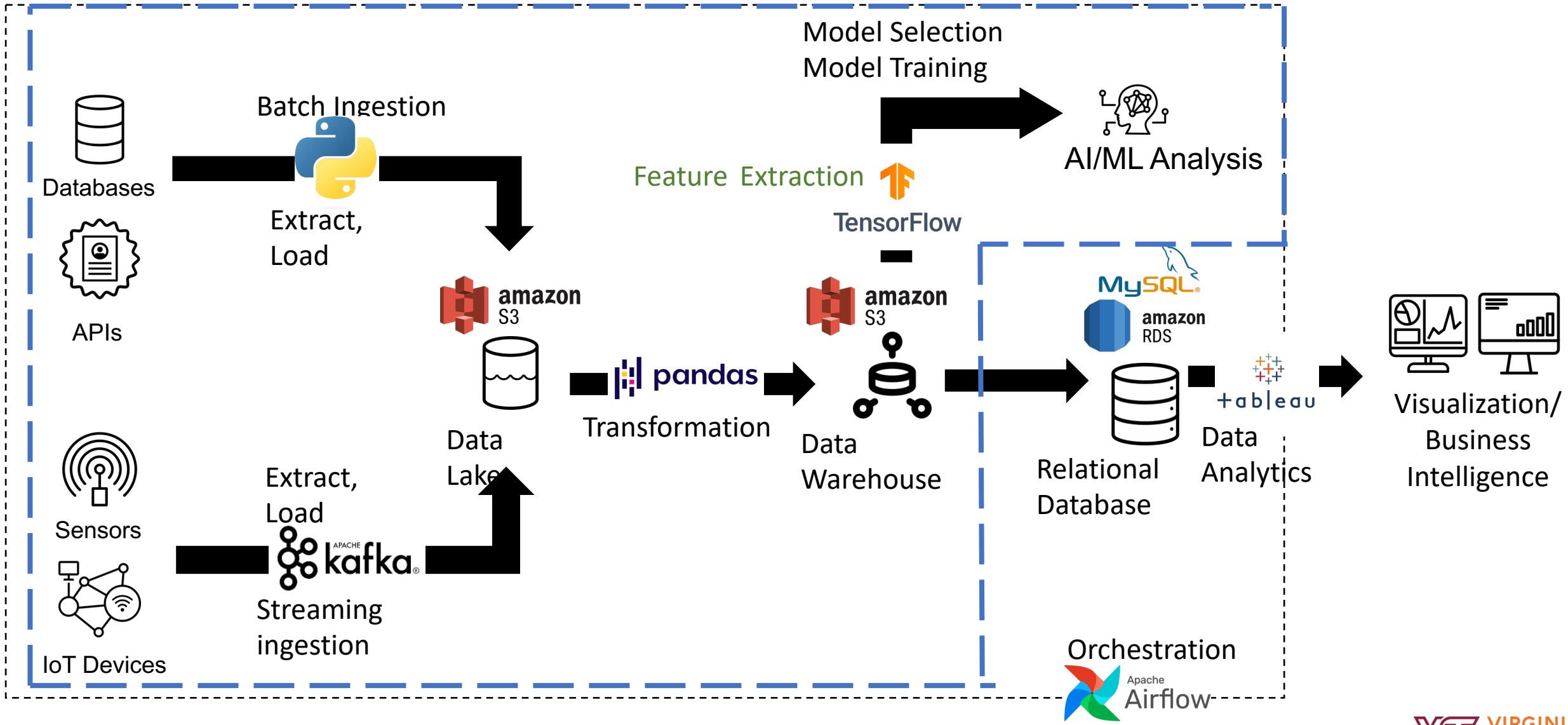
Objectives

- Machine Learning Pipeline
  - Feature Extraction
- Building and Training the ML Model
  - From Tensors to TensorFlow

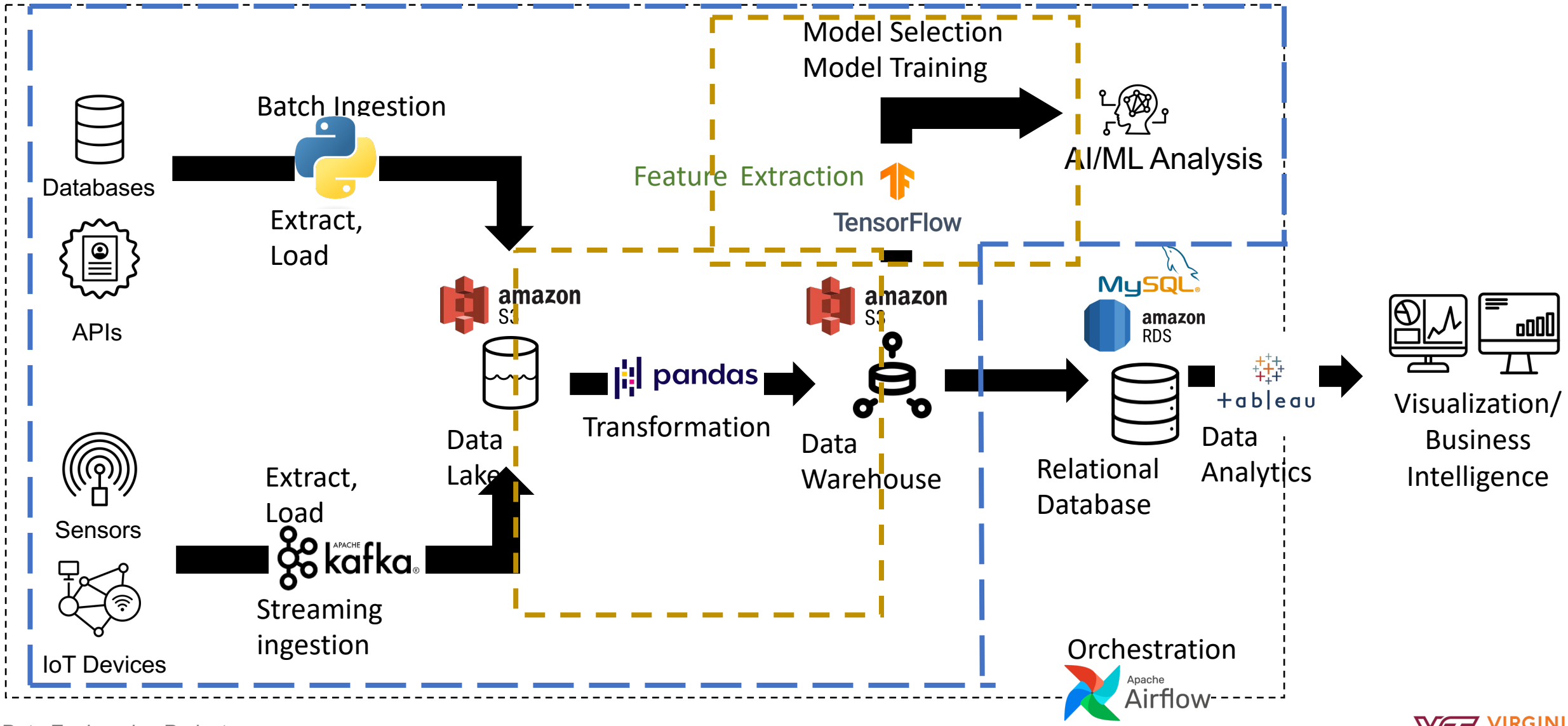
Machine Learning Pipeline

The end-to-end construct that orchestrates the flow of data into, and output from, a machine learning model
(or set of multiple models)

Custom Data Engineering Pipeline



Custom Data Engineering Pipeline



Machine Learning Pipeline Overview

Step 1
Acquire the dataset needed to perform our machine learning task and perform EDA/cleaning on it

Step 2

- Perform feature extraction
- Divide the dataset into a training set and a testing set
- Advice: have an 80/20 split of the data as training and testing

Step 3

- Choose a model specific to the machine learning task and build it
- Use the training dataset to train said model
- Verify your model works on the test dataset

Step 4
Save the trained model and push it to the storage space

Step 5
Use it to make data-informed decisions

The Machine Learning Process

- Training an ML model: ML algorithm & training dataset to learn from
- ML model: the model artifact that is created by the training process
  - Types of models: e.g., Supervised ML, Unsupervised ML

Example: use a financial dataset to make a prediction model based on the "close" price of different companies

Model: LSTM will be trained on the data in the training set

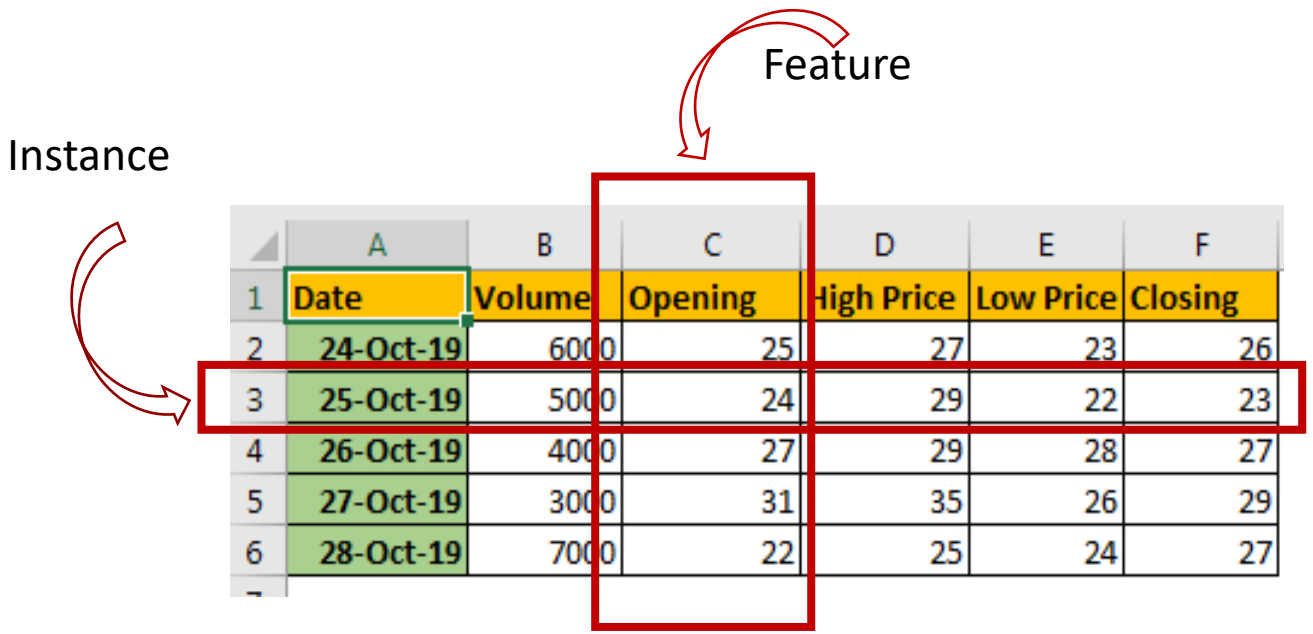
- Determine the features to be used in the training process is critical

Features (towards Feature Extraction)

Datasets: instances + features

Instances are described through features also known as attributes

E.g., in a stock market dataset for a particular company, the *open, close, high, and low price of a stock on a particular day* is called an instance, while "opening" is a feature



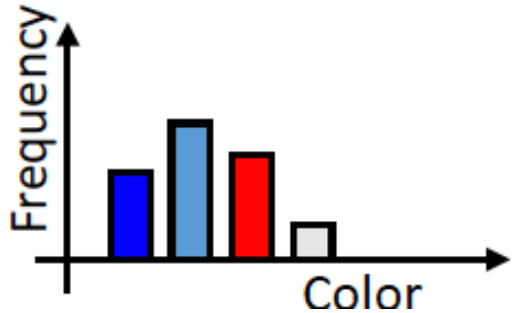
	A	B	C	D	E	F
1	Date	Volume	Opening	High Price	Low Price	Closing
2	24-Oct-19	6000	25	27	23	26
3	25-Oct-19	5000	24	29	22	23
4	26-Oct-19	4000	27	29	28	27
5	27-Oct-19	3000	31	35	26	29
6	28-Oct-19	7000	22	25	24	27

Feature Extraction - Definition

Feature Extraction: reducing the number of features (attributes) of a dataset by creating new features from the existing ones

New reduced set of features → to summarize most of the information contained in the original set of features

- Images
  - Feature: pixels
  - Feature: the distribution of colors in the RGB space over the pixels of an image



Feature extraction for Text Data

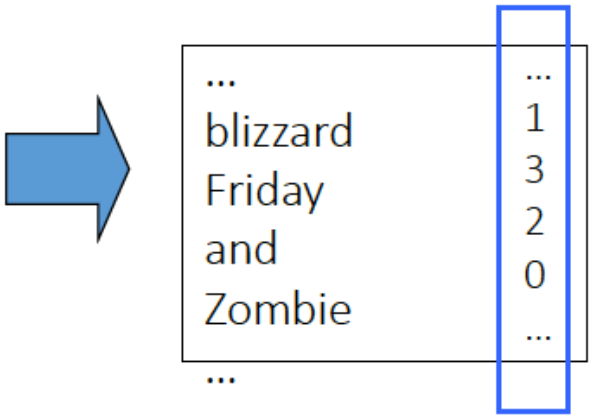
Text can be represented as a set of terms (Bag-Of-Words model)

Terms can be:

- Unigrams (“cluster“, “analysis“..)
- Bigrams (“cluster analysis”, ...)
- *n*-grams

Typical feature extraction from text: transform the document into a vector of term frequencies

The region is preparing for blizzard conditions Friday, with the potential for more than two feet of snow in the Fairfax City area. Conditions are expected to deteriorate Friday afternoon, with the biggest snowfall, wind gusts and life-threatening conditions Friday night and Saturday.



Building and Training the ML Model

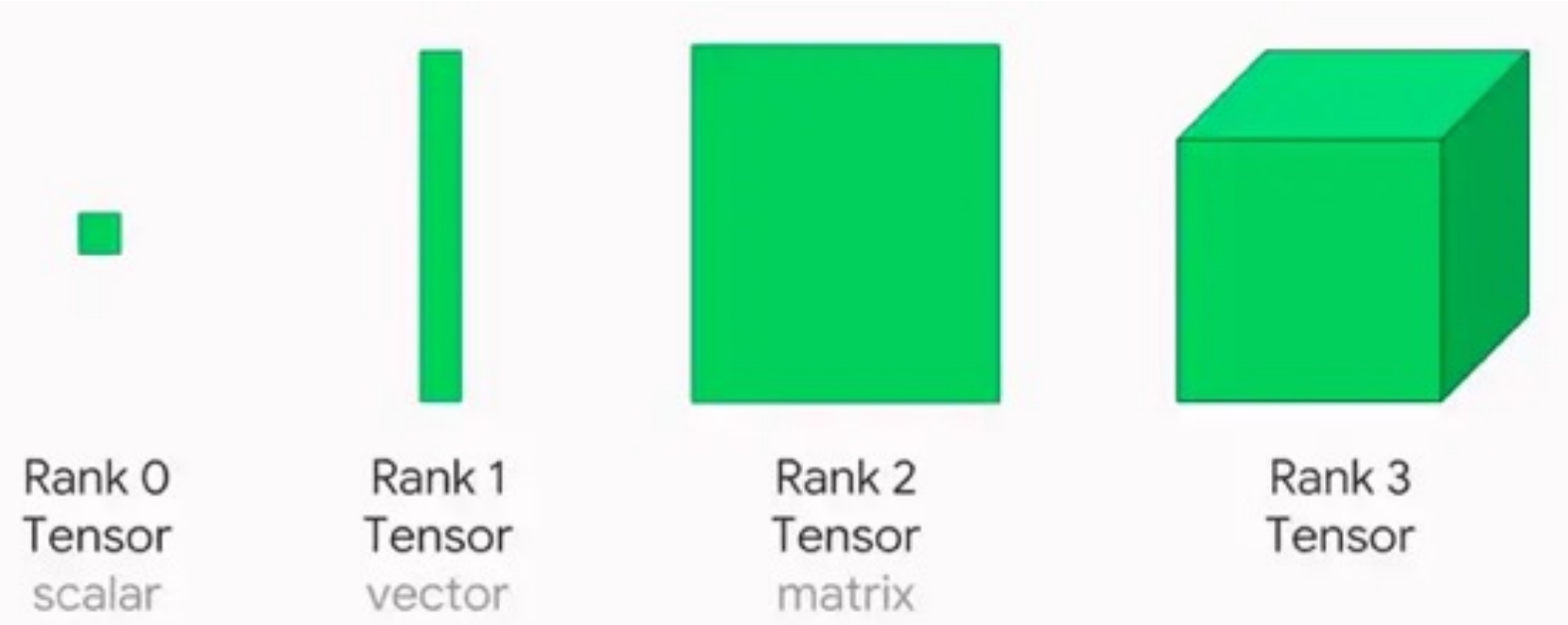
Use of TensorFlow Python library to build and train the ML model

- open-source library for numerical computation and large-scale machine learning
- based on Neural Networks
- developed by Google Brain Team to conduct machine learning research
- “TensorFlow is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms”

Tensors

A tensor is an N-dimensional array of data

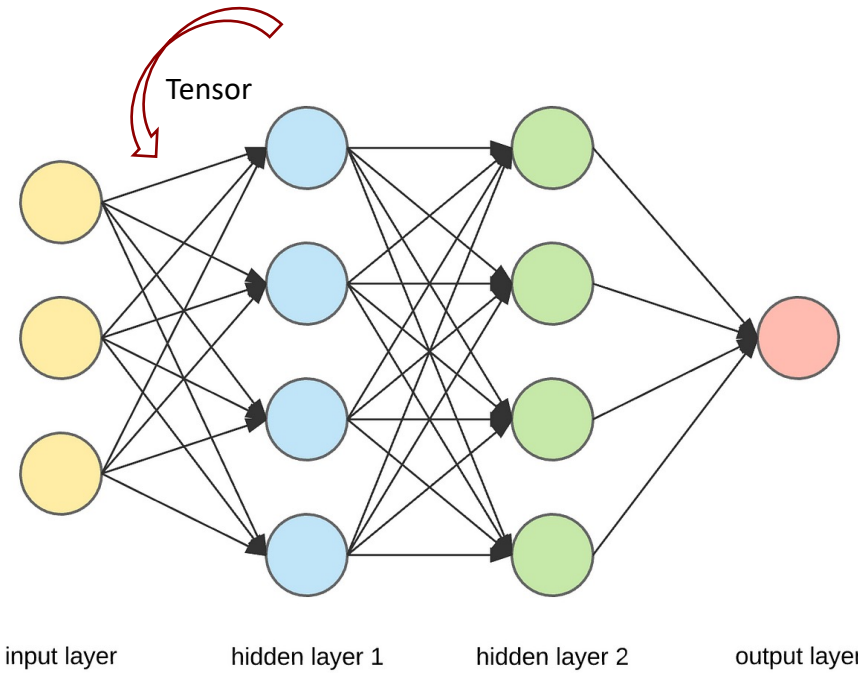
A standard in scientific computing simulations, machine learning settings including deep learning



How are Tensors used in ML?

- Represent input data and output data
  - and the hidden layers

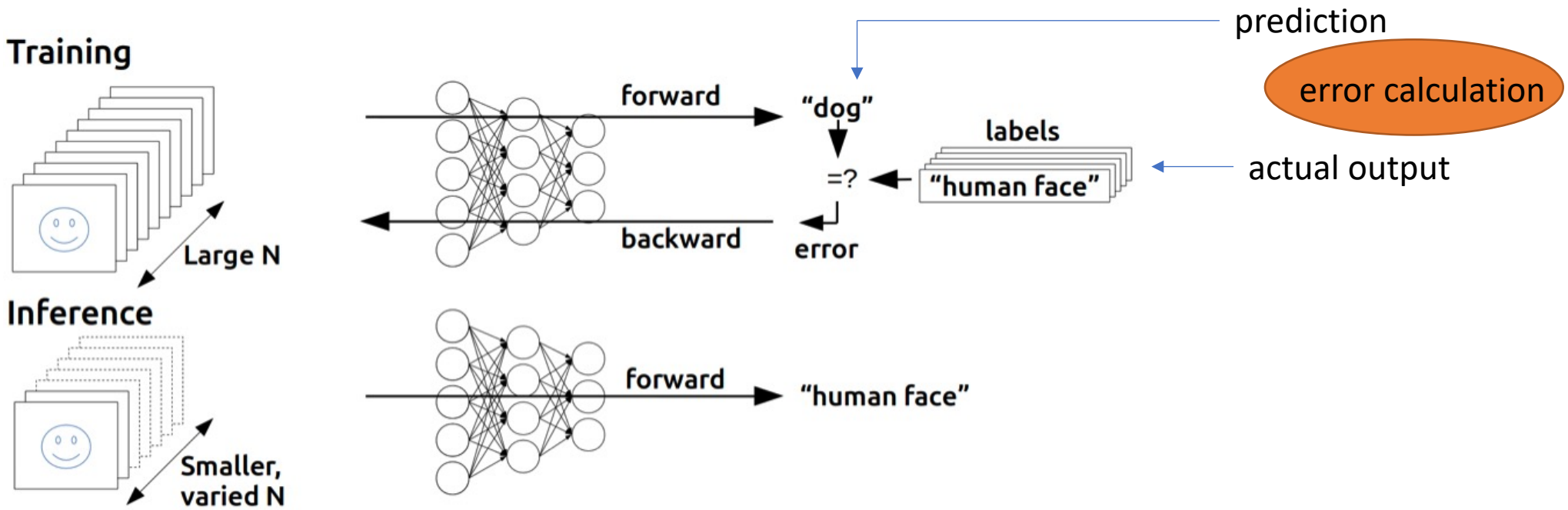
- Hidden layers can find **features** within the data and allow the following layers to operate on those features



Learning: Backpropagation

Backpropagation is a process involved in training a neural network
It takes the error rate of a forward propagation and
feeding this loss backward through the neural network layers to fine-tune the model

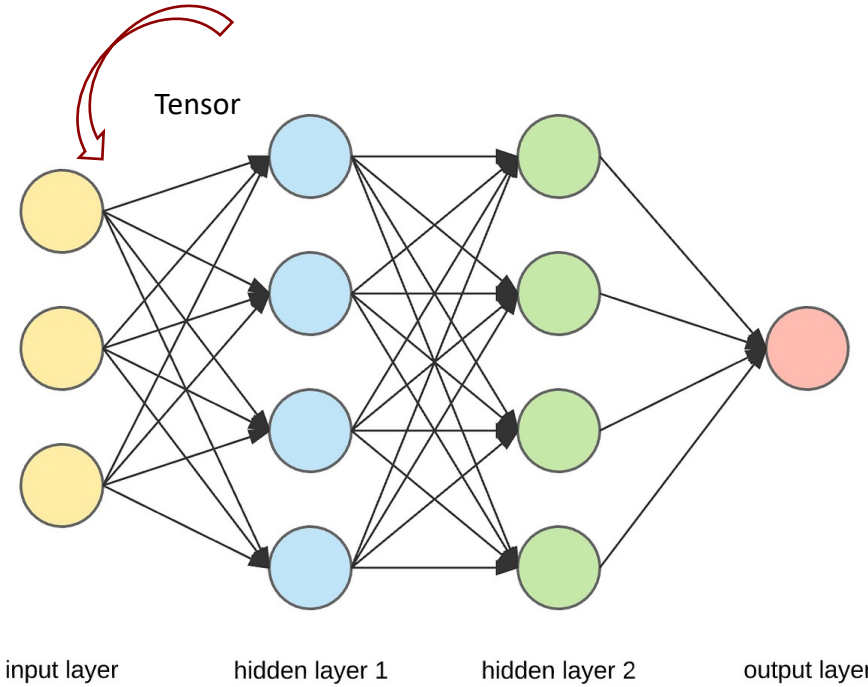
Backpropagation is the essence of neural net training.



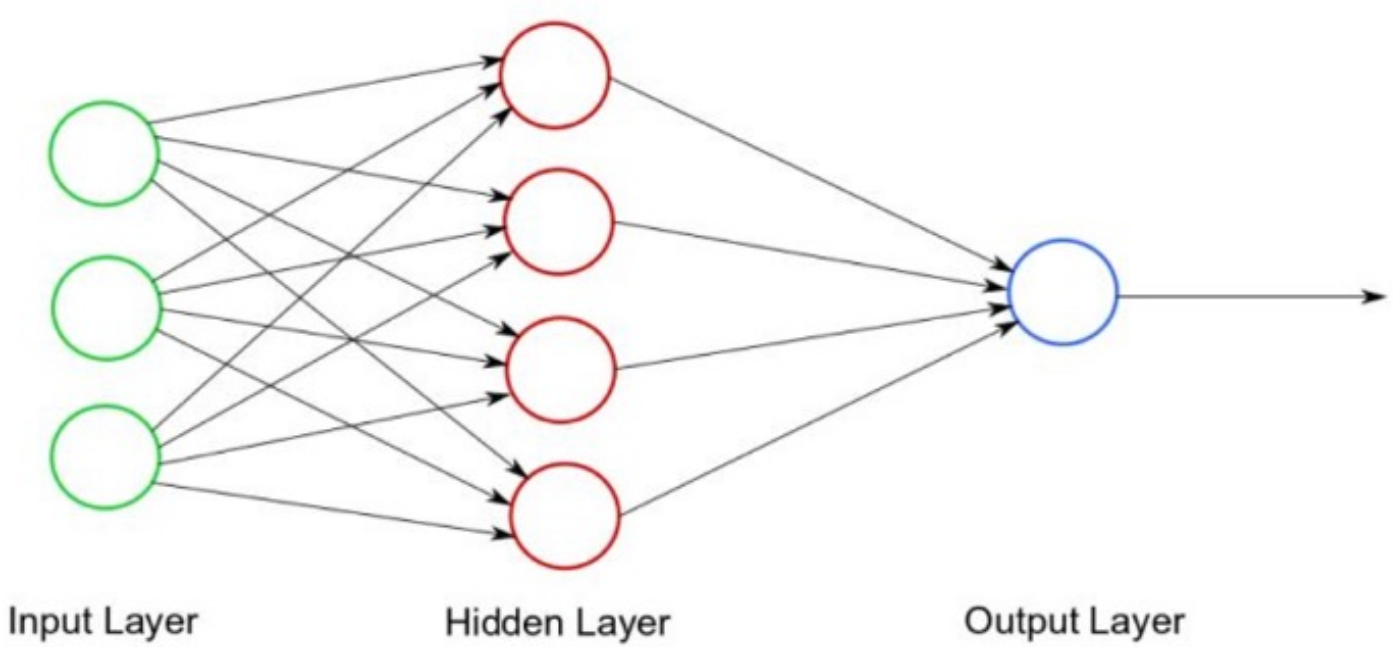
From: http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/

From Tensors to TensorFlow

- Key idea: express a numeric computation as a graph
- Graph nodes are operations with any number of inputs and outputs
- Graph edges are tensors which flow between nodes



Provenance matters



Logging:

- Data
  - Input
  - Output
  - Intermediate
- Features
- Structure

Summary

- Machine Learning Pipeline
  - Feature Extraction
- Building and Training the ML Model
  - From Tensors to TensorFlow

Data Engineering Project

Module 6
Machine Learning Pipeline

Nektaria Tryfona, PhD

Electrical and Computer Engineering
Virginia Tech

Machine Learning Models

Training an ML model: ML algorithm & training dataset to learn from

ML model: the model artifact that is created by the training process

Supervised (labeled datasets)	Used for
Support Vector Machines	Document classification, spam filtering
Logistic Regression	Image classification
Unsupervised (unlabeled datasets)	**Used for**
Clustering	Anomaly detection, cluster segmentations
Principal Component Analysis	Feature reduction and reducing dimensionality