Module 2

Setting up Machine Learning Environment

Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

Class Modules

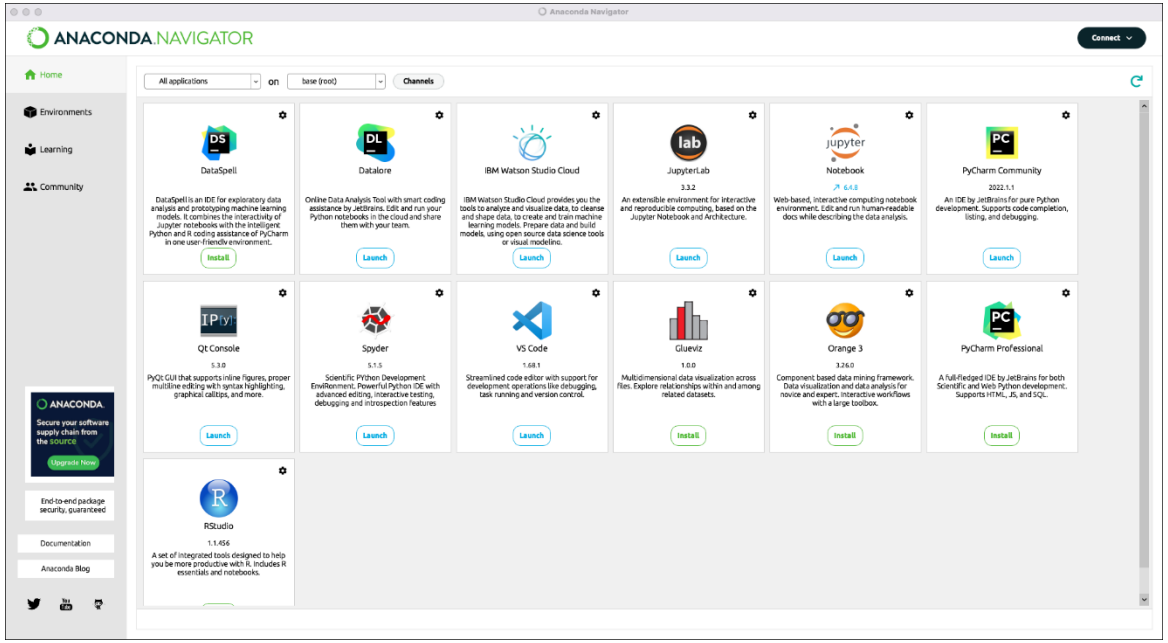
- Module 1- Introduction to Deep Learning
- Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)

Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry

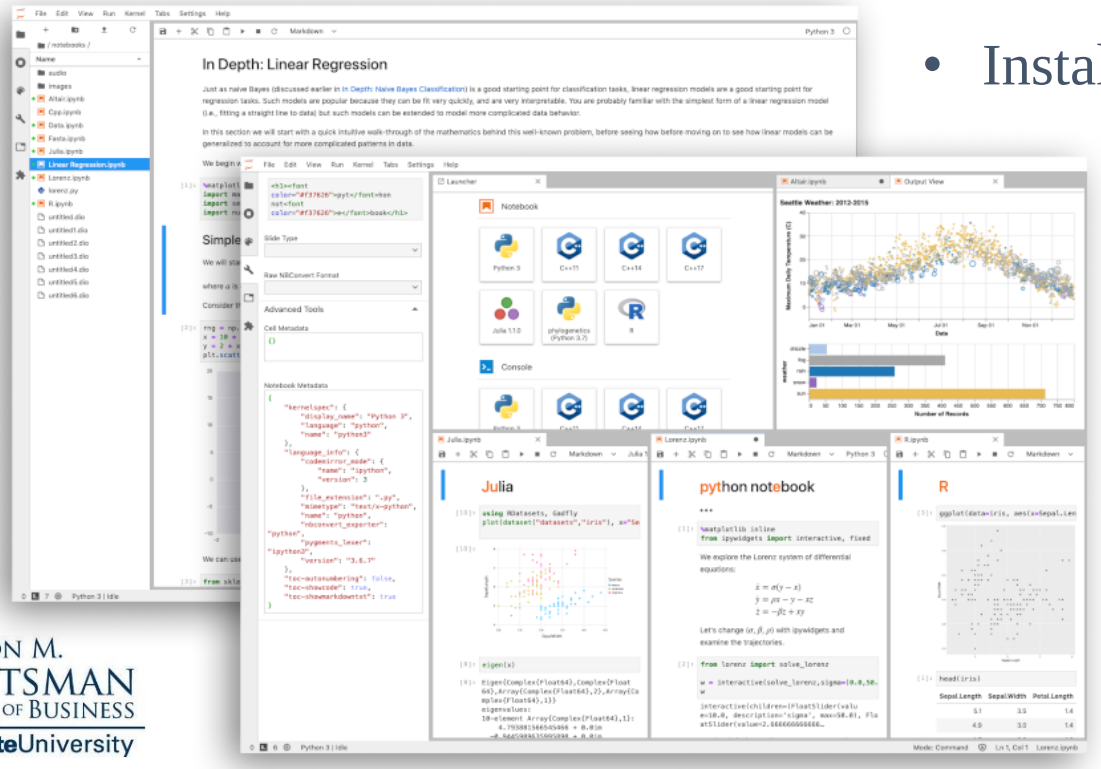
Install through

- Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management with conda environments.
- Anaconda offers the easiest way to perform data science and machine learning on a single machine.
- Install Anaconda @ https://www.anaconda.com/products/distribution



JupyterLab

- JupyterLab is the latest web-based interactive development environment for notebooks, code, and data
- Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are Julia, Python and R



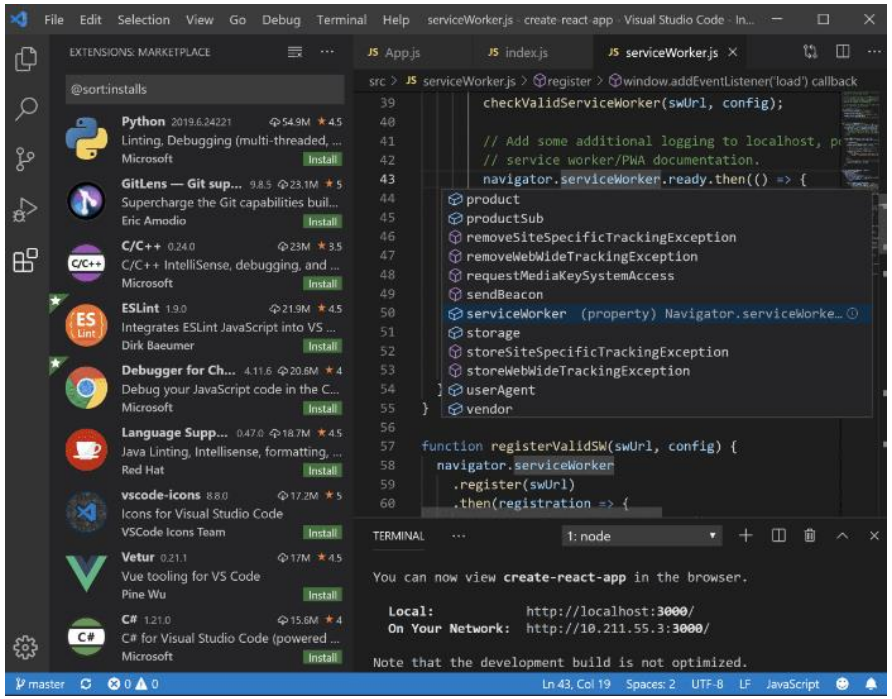
- Install JupyterLab @ http://jupyter.org/install

Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry

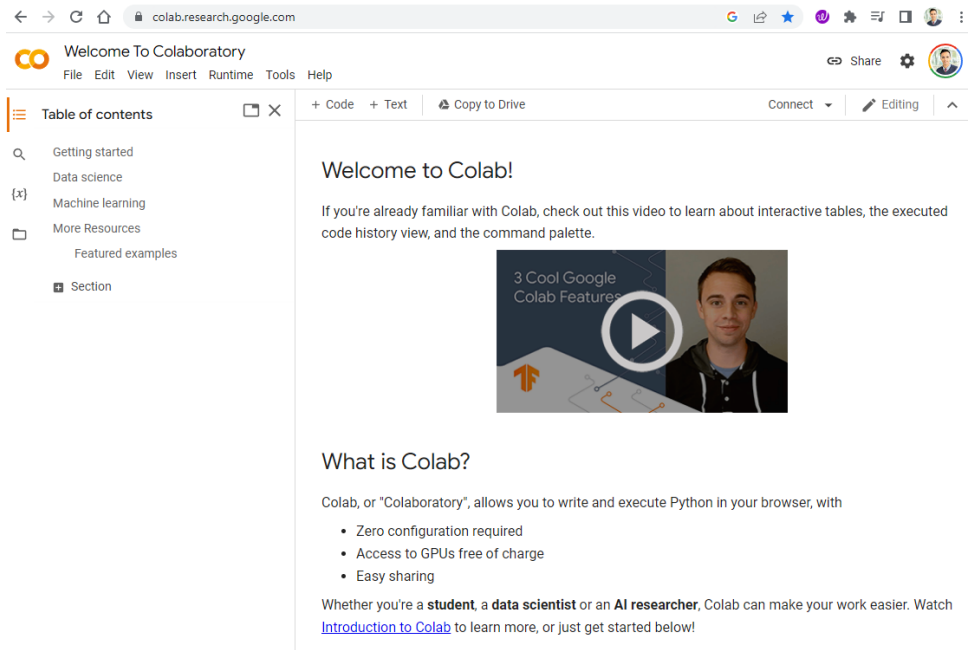
VS Code

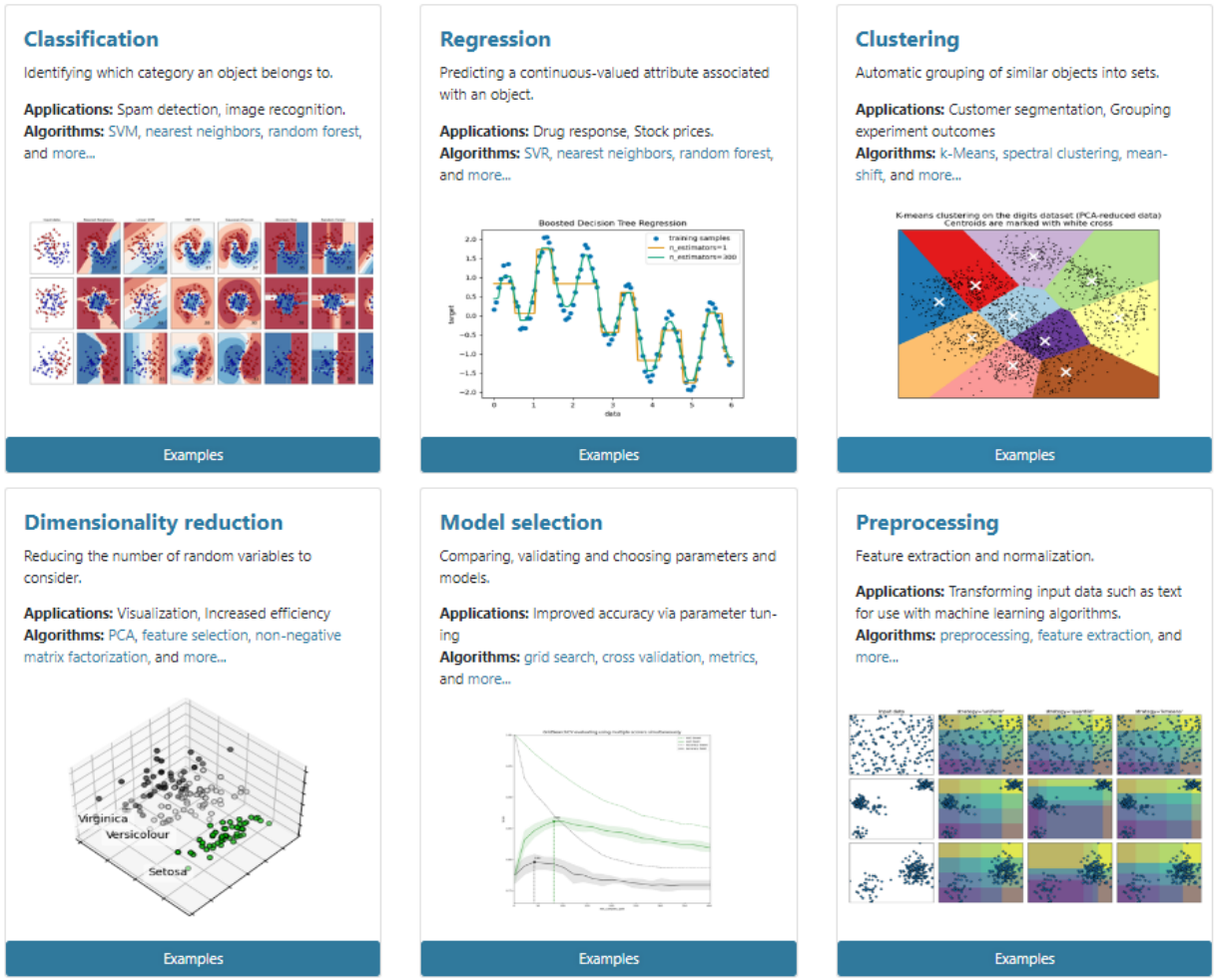
- VS Code is one of the most popular source code editors
- Features include support for debugging, syntax highlighting, intelligent code completion, code refactoring, and embedded Git.
- Install VS code @ https://code.visualstudio.com/



Google colab

- Colab is a free hosted Jupyter notebook-style environment that runs entirely in the cloud and requires no setup to use. It also provides access to machine learning libraries and computing resources including GPU.
- Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. https://colab.research.google.com/





- Scikit-learn is an open-sourced Python library and includes a variety of unsupervised and supervised learning techniques.
- It is based on technologies and libraries like Matplotlib, Pandas and NumPy and helps simplify the coding task.
- Install Scikit-learn @ https://scikit-learn.org/stable/install.html

Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

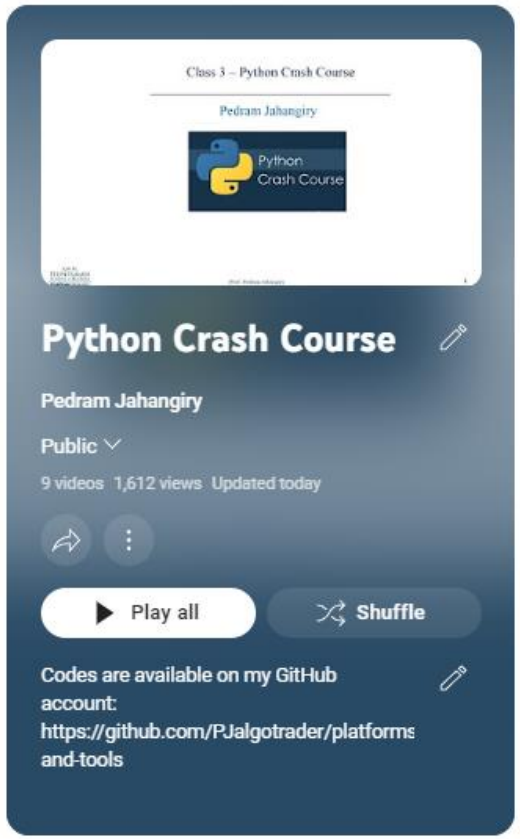
PyCaret

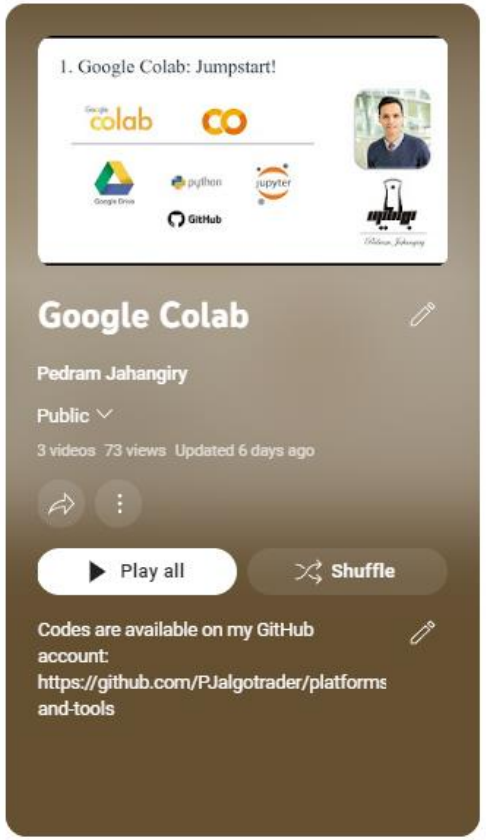
- PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows.
- PyCaret is essentially a Python wrapper around several machine learning libraries and frameworks
- Install PyCaret @ https://pycaret.gitbook.io/docs/get-started/installation

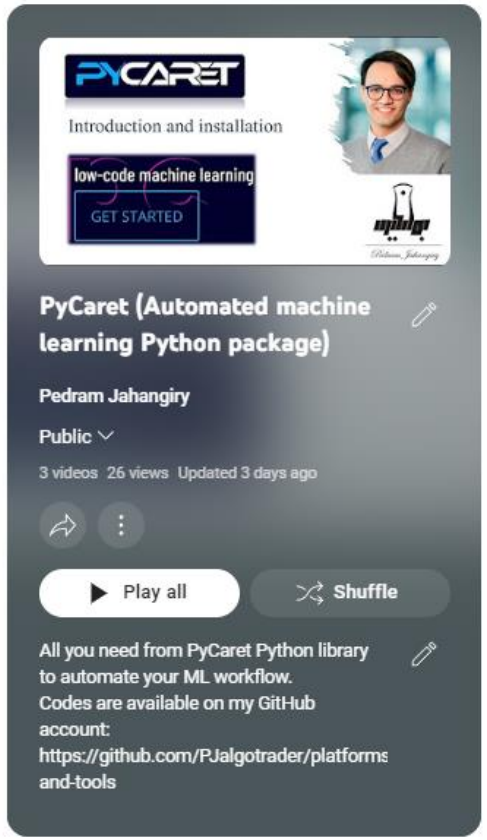
```
# load dataset
import pandas as pd
train = pd.read_csv('train.csv')
test  = pd.read_csv('test.csv')

# init setup
from pycaret.classification import *
s = setup(train, target= 'target')

# model training and selection
best = compare_models()

# analyze best model
evaluate_model(best)


# predict on new data
predictions = predict_model(best, data =test )

# save best pipeline
save_model(best, 'my_best_pipeline')
```

ID	Name
'lr'	Logistic Regression
'knn'	K Nearest Neighbour
'nb'	Naives Bayes
'dt'	Decision Tree Classifier
'svm'	SVM - Linear Kernel
'rbfsvm'	SVM - Radial Kernel
'gpc'	Gaussian Process Classifier
'mlp'	Multi Level Perceptron
'ridge'	Ridge Classifier
'rf'	Random Forest Classifier
'qda'	Quadratic Discriminant Analysis
'ada'	Ada Boost Classifier
'gbc'	Gradient Boosting Classifier
'lda'	Linear Discriminant Analysis
'et'	Extra Trees Classifier
'xgboost'	Extreme Gradient Boosting
'lightgbm'	Light Gradient Boosting
'catboost'	CatBoost Classifier

Available YouTube playlists







Platforms and Packages

Listed below are some Python packages and platforms that will be used in the deep learning and deep forecasting courses.

Google colab

General Python libraries

NumPy

pandas

matplotlib

seaborn

Machine Learning libraries

scikit learn

dmlc XGBoost

LightGBM

PyCaret

Deep Learning libraries

Keras

TensorFlow

HUGGING FACE

Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

Setting up Deep Learning Environment

Google colab

	Personal Workstation	Could Platforms	Google Colaboratory
Pros	• Full control over hardware and software • Work offline • Fixed cost	• Powerful computing resources • Scalability • Ease of use • Cost-effective: Pay-as-you-go • Collaboration	• Powerful computing resources (GPU, TPU) • Ease of use • Collaboration • No need to set up a local environment
Cons	• Scalability • Maintenance (both hardware and software)	• Expensive for large-scale experiments • Dependency on the provider • Limited control • Internet connection • Security	• Time limit • Hardware limitation • Data storage • Limited control • Internet connection • Security

Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry

The modern machine learning landscape

- From 2016 to 2020, the entire machine learning and data science industry has been dominated by these two approaches:
  1. Deep learning
  2. Gradient boosted trees

- Most practitioners of deep learning use Keras, often in combination with its parent framework TensorFlow.
- This means you'll need to be familiar with Scikit-learn, XGBoost, and Keras

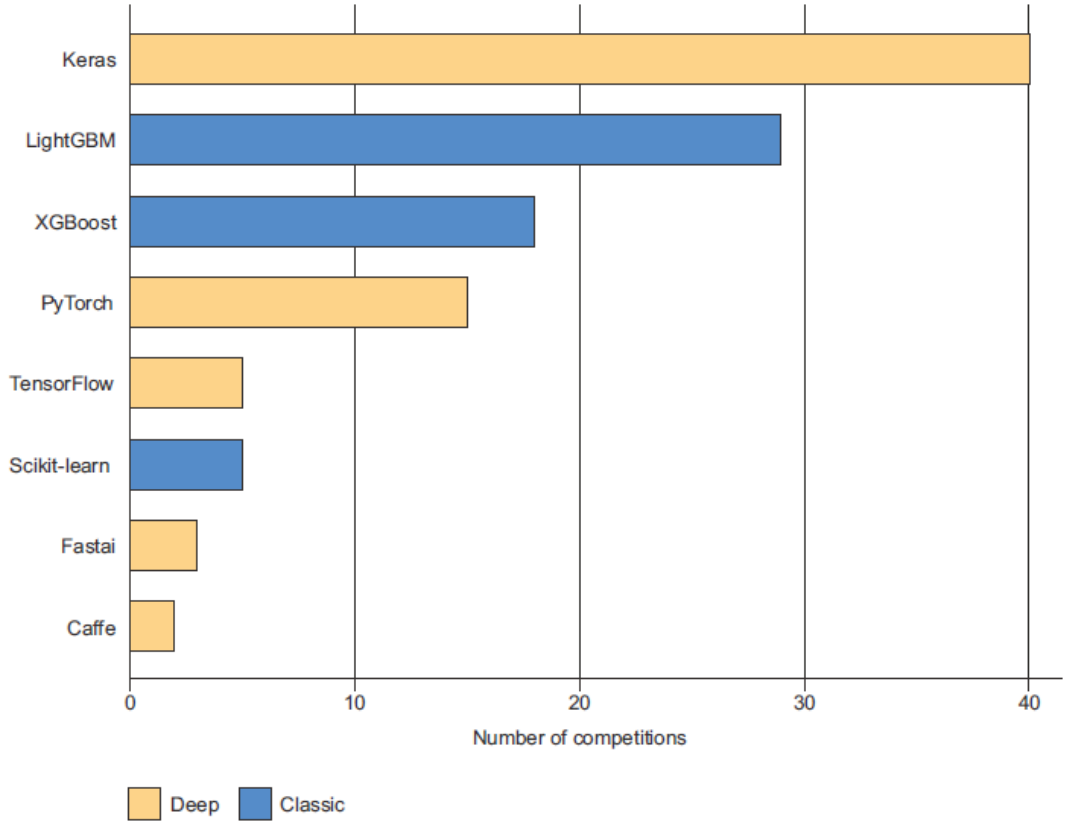


Figure 1.12 Machine learning tools used by top teams on Kaggle

The modern machine learning landscape

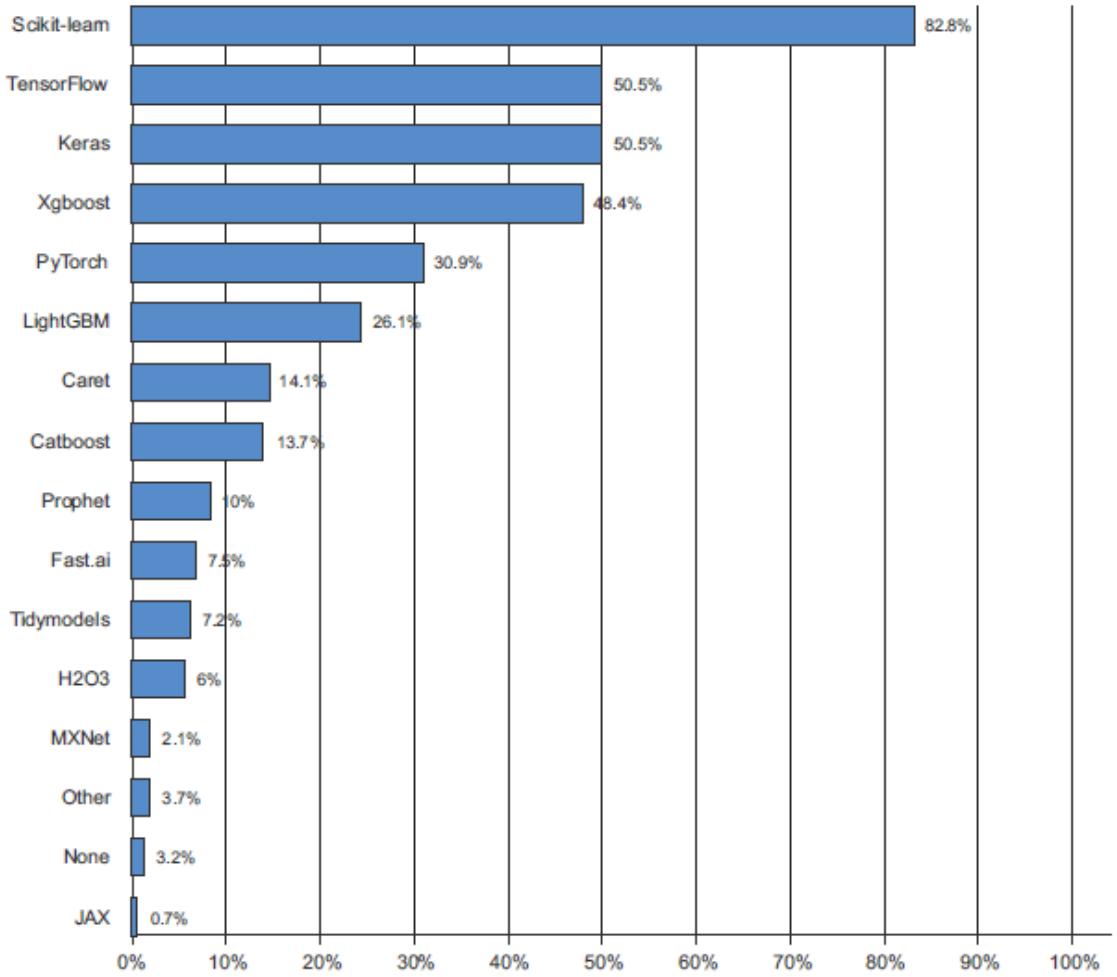


Figure 1.13 Tool usage across the machine learning and data science industry (Source: www.kaggle.com/kaggle-survey-2020)

- Kaggle also runs a yearly survey among machine learning and data science professionals worldwide.
- This survey is one of our most reliable sources about the state of the industry!!!
- This figure shows the percentage of usage of different machine learning software frameworks.

Pedram Jahangiry

Class Modules

- ✓ Module 1- Introduction to Deep Learning
- ✓ Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)

Jon M. Huntsman School of Business
UtahStateUniversity

Pedram Jahangiry