

# Bil 470 / YAP 470

Introduction to Machine Learning (Yapay Öğrenme)

Batuhan Bardak

**Lecture 1:** Course outline and basic concepts of ML

**Date:** 13.09.2022

# Plan for today

- Course outline and materials
- Basic concepts and terminology of Machine Learning

# Logistics

- **Instructor:** Batuhan BARDAK
  - batuhanbardak@etu.edu.tr
- **Teaching Assistant:** ..
  - ....
- **Lectures:**
  - Monday: 08:30 - 10:30, Room: B07
  - Tuesday: 10:30 - 12:30, Room: 157

# What this class is

- **Fundamentals of ML:** supervised learning (e.g., linear regression, logistic regression, svm, boosting, deep learning), unsupervised learning (e.g., k-means, hierarchical clustering, PCA), bias/variance tradeoff, overfitting, advice for applying machine learning
- **More Recent topics of ML:** AutoML, Explainable AI, MLOps

# Prerequisites

- Basic algorithms and data structures
- Basic probability and statistics
- Basic linear algebra
- Good programming skills (especially in Python)

# Grading

- Midterm: %15
- Final: %25
- Homeworks: %20 (with Python)
- Paper Presentation: %10
- Course Project: %30 (done in groups of 2)

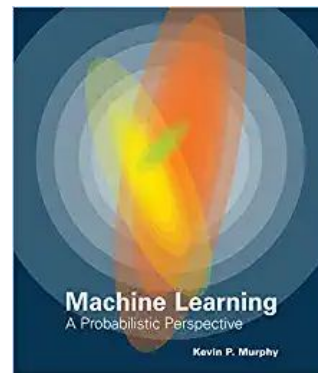
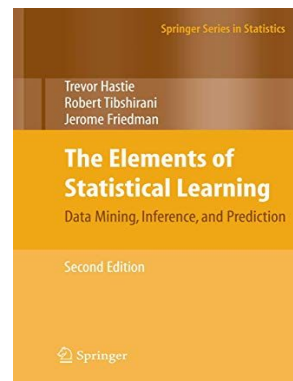
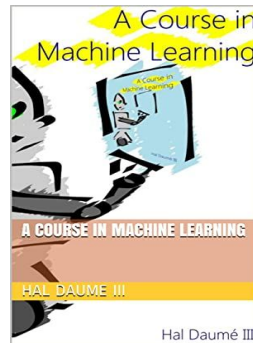
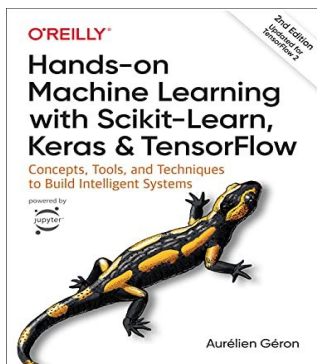
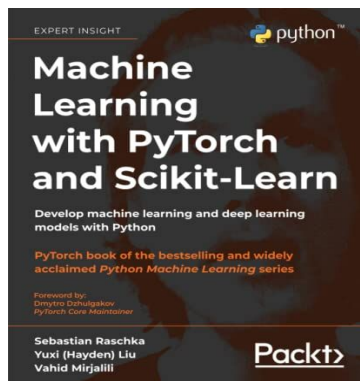
**Note:** Extra 5 points for submitting project final report as a paper to the IEEE conference.

# Course Project

- Some example domains:
  - Healthcare
  - Finance
  - Cybersecurity
  - Social media data
  - And many more..
- Grading
  - Proposal (% 5)
  - Literature Review (% 5)
  - Novelty (%5)
  - Github commits (%10)
  - Progress Report (%10)
  - Deployment (%15)
  - Project Presentation - in class & video (%20)
  - Final Report (%30)

# Reference Books

- Machine Learning with PyTorch and Scikit-Learn, Raschka, 2022
- Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow, Geron, 2nd Edition, 2017
- A Course in Machine Learning, Hal Daumé III, 2017
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, 2016
- Machine Learning: A Probabilistic Perspective, Murphy, MIT Press, 2012





# Communication

- Piazza
  - [piazza.com/etu.edu.tr/fall2022/yap470](https://piazza.com/etu.edu.tr/fall2022/yap470)
- Github
  - <https://github.com/bbardakk/TOBB-ETU-YAP470-2022>
- E-mail
  - [batuhanbardak@etu.edu.tr](mailto:batuhanbardak@etu.edu.tr)

# Course outline

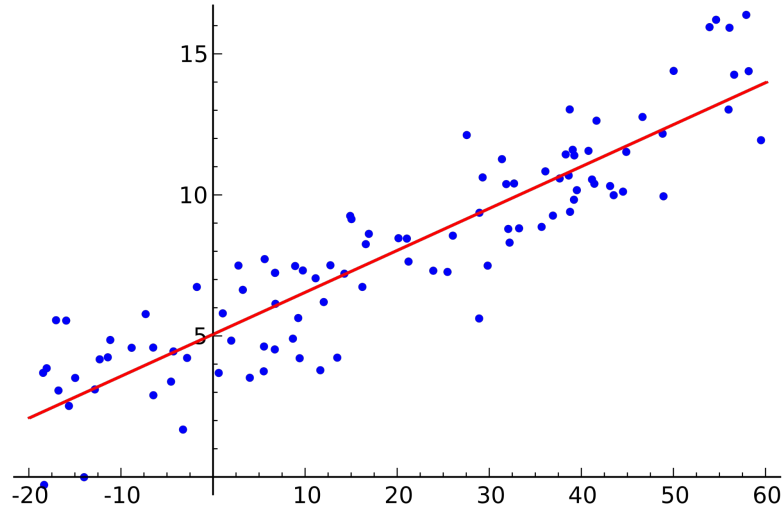
- **Week 1:** Introduction to ML and basic concepts
- **Week 2:** Machine Learning methodologies
- **Week 3:** Supervised Learning - Regression
- **Week 4:** Supervised Learning (Classification-1)
  - Logistic Regression, Neural Networks
- **Week 5:** Supervised Learning (Classification-2)
  - Naive Bayes, SVM
- **Week 6:** Supervised Learning (Classification-3)
  - Decision Tree, Ensemble Learning

# Course Outline

- **Week 7:** Feature selection and dimension reduction
- **Week 8:** Advice for applying machine learning & AutoML
- **Week 9:** Unsupervised Learning
  - Clustering (K-means, hierarchical clustering)
- **Week 10:** Deep Learning, Active Learning
- **Week 11:** MLOps
- **Week 12:** Ethics & Fairness in AI, Explainable AI

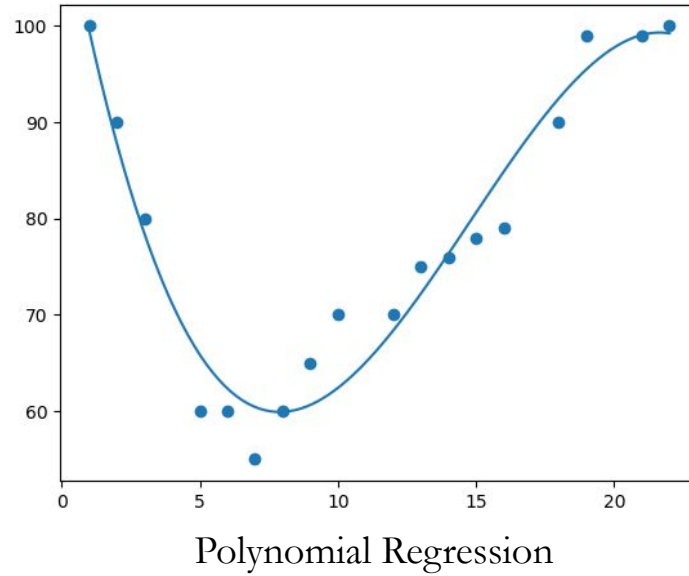
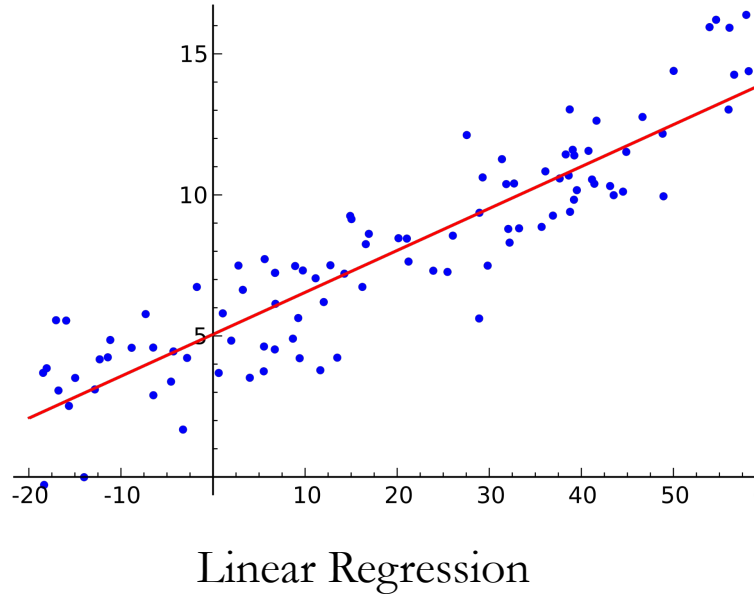
# Summary of Weeks

# Supervised Learning - Regression



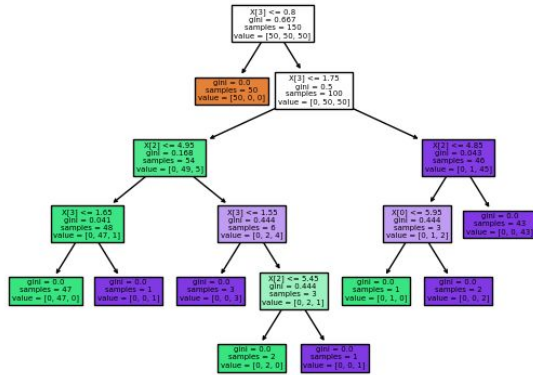
Linear Regression

# Supervised Learning - Regression



# Supervised Learning - Classification

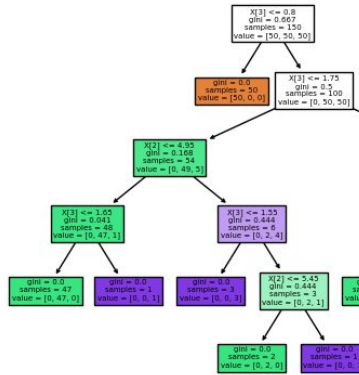
Decision tree trained on all the iris features



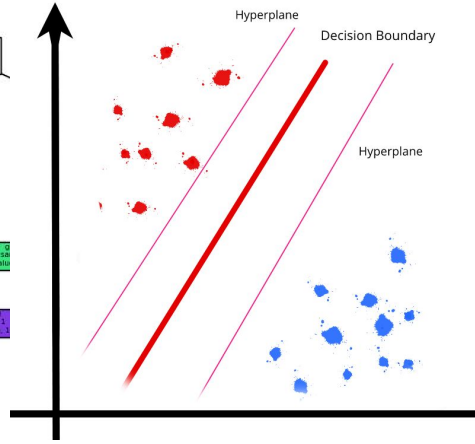
Decision Tree

# Supervised Learning - Classification

Decision tree trained on all the iris features



Decision Tree

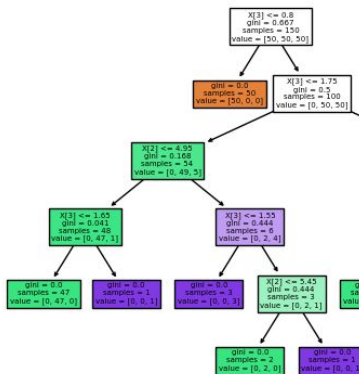


SVM

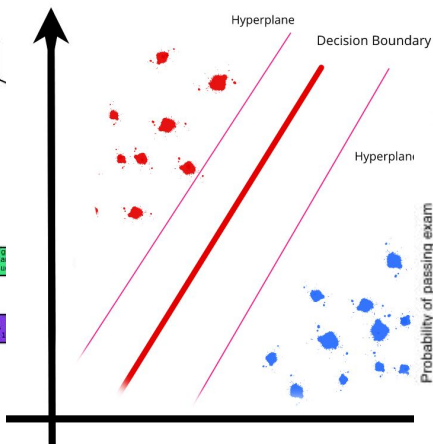


# Supervised Learning - Classification

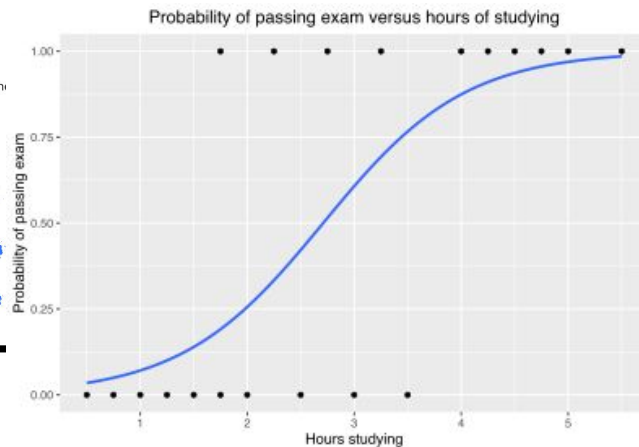
Decision tree trained on all the iris features



Decision Tree



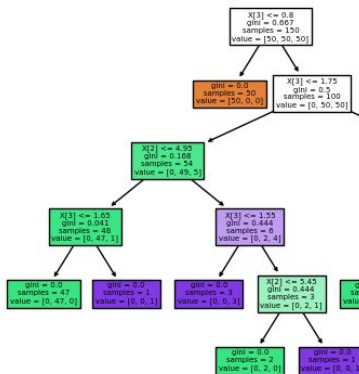
SVM



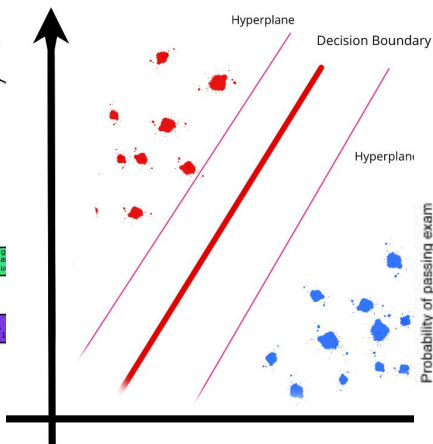
Logistic Regression

# Supervised Learning - Classification

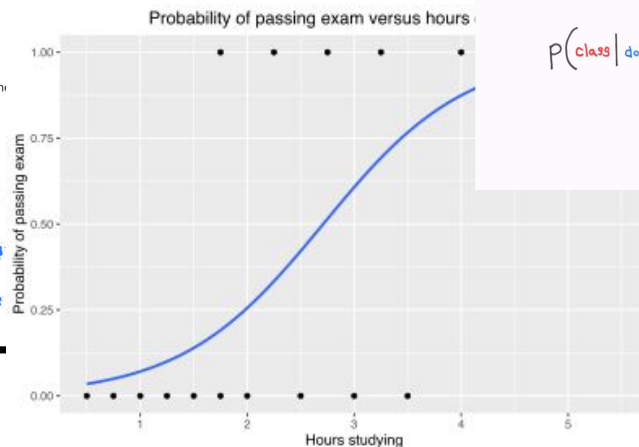
Decision tree trained on all the iris features



## Decision Tree



SVM



# Logistic Regression

GAUSSIAN  
**NAIVE BAYES**  
 CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

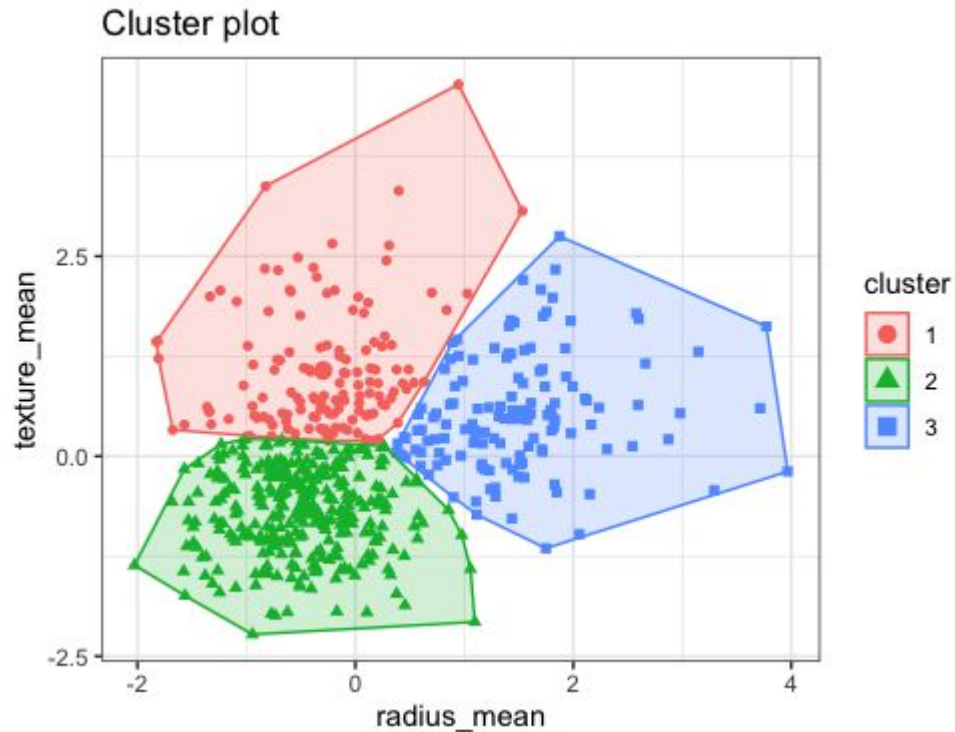
$$P(\text{class} | \text{data}) = \frac{P(\text{data} | \text{class}) \times P(\text{class})}{P(\text{data})}$$

We don't calculate this in naive bayes classifiers

Chris Albon

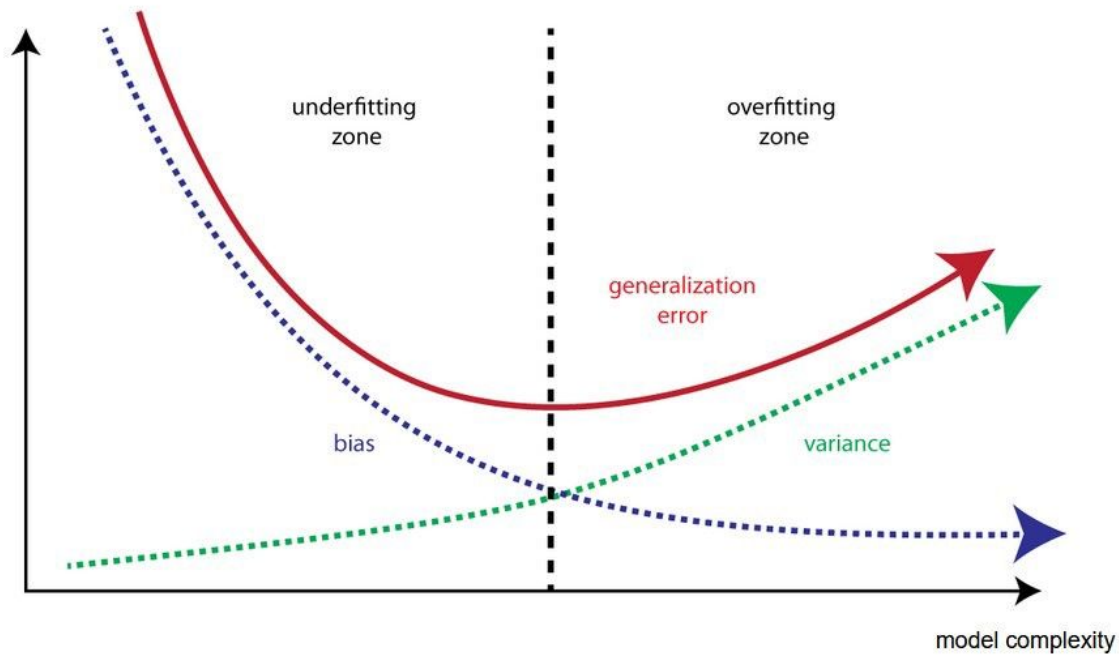
# Naive Bayes

# Unsupervised Learning



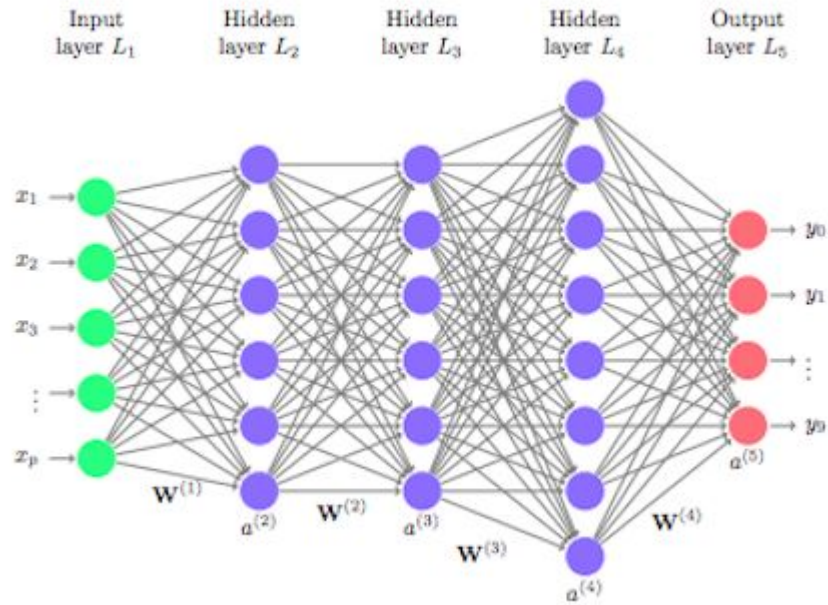
# Advices for applying machine learning?

the bias vs. variance trade-off

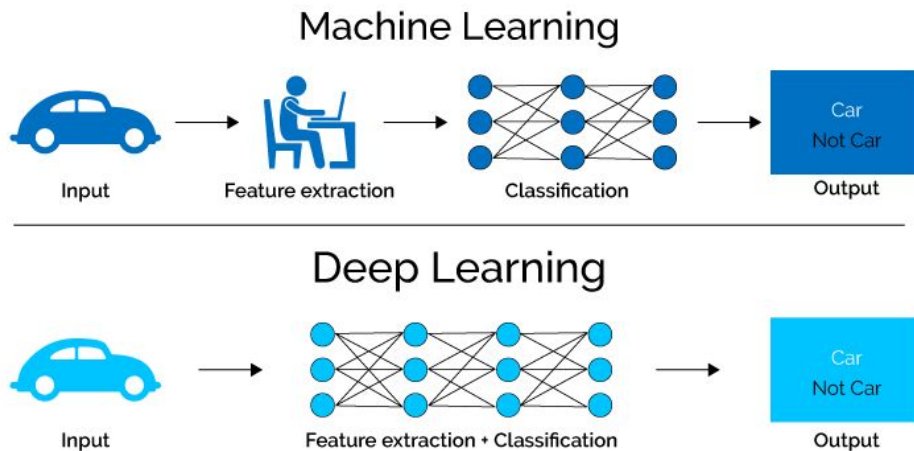
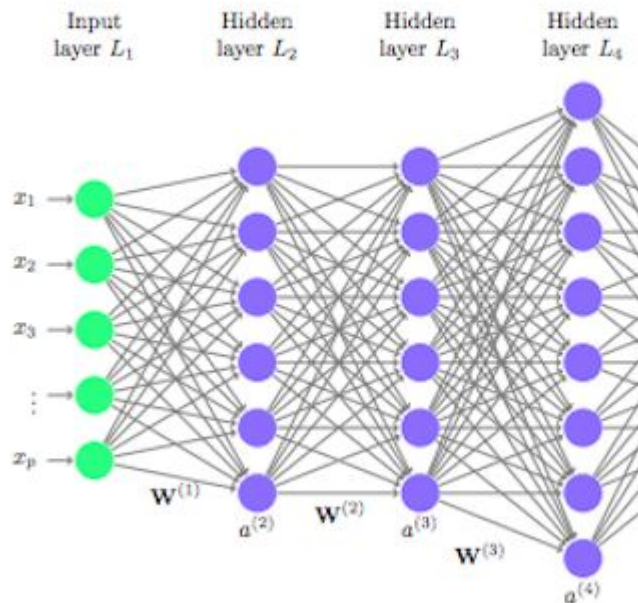


What is next ?

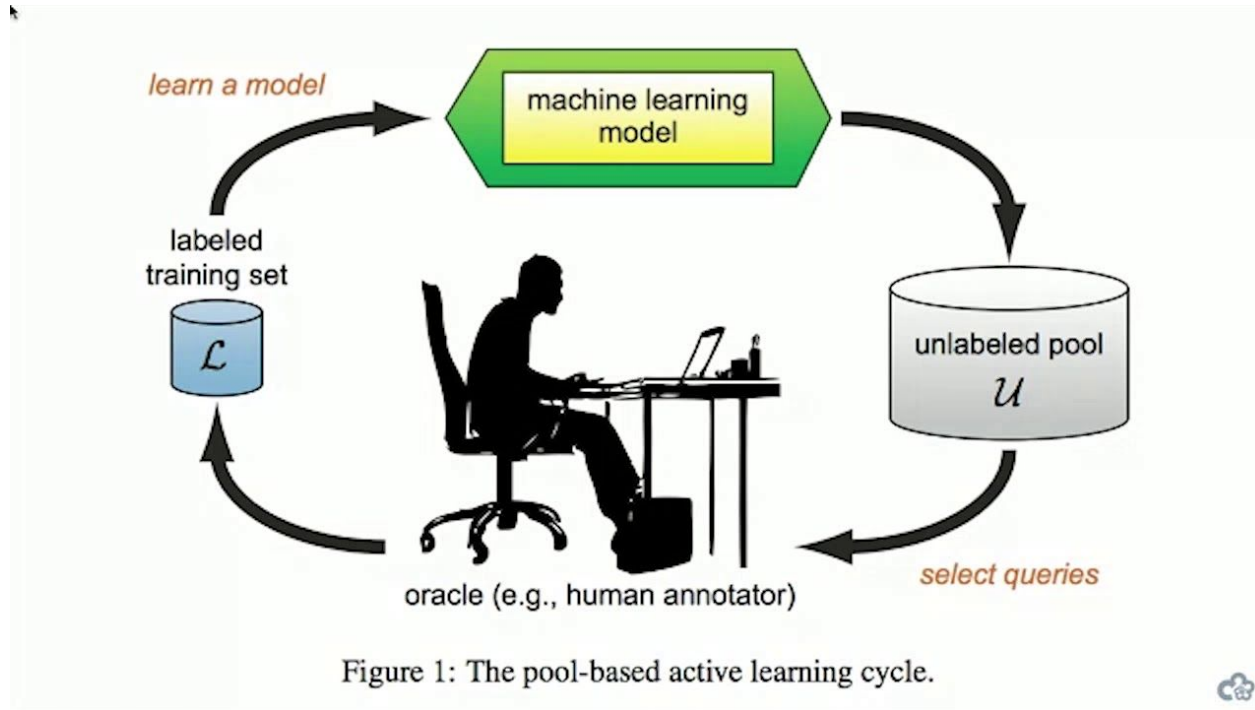
# Deep Learning



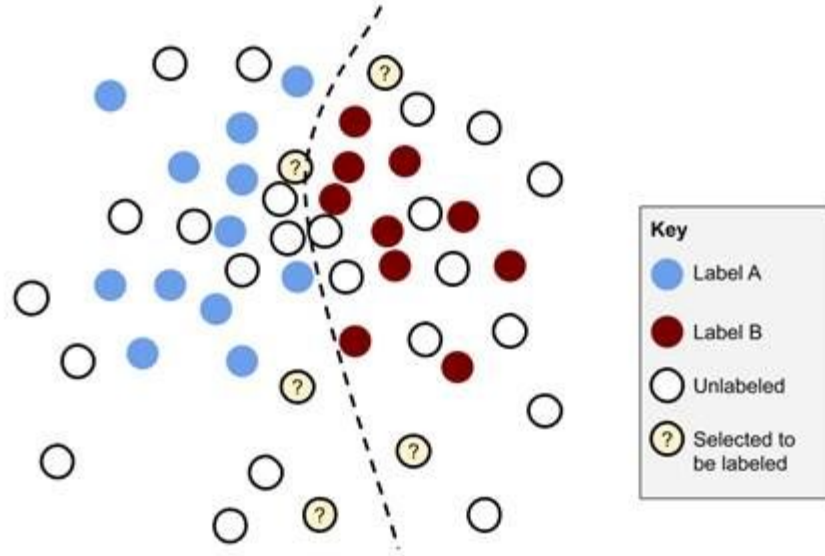
# Deep Learning



# Active Learning




# Active Learning






# MLOps

The logo for MLFlow is a dark blue rectangle. At the top, there is a glowing, wavy line of blue dots and lines, resembling a neural network or data flow. Below this graphic, the text "An open source platform for the machine learning lifecycle" is written in white, sans-serif font, centered within the rectangle.

An open source platform for the  
machine learning lifecycle

**MLFlow**

# MLOps

A dark blue banner with a glowing blue network pattern of dots and lines in the background.

An open source platform  
machine learning life

MLFlow



**Continuous Machine  
Learning (CML) is CI/CD  
for Machine Learning  
Projects**

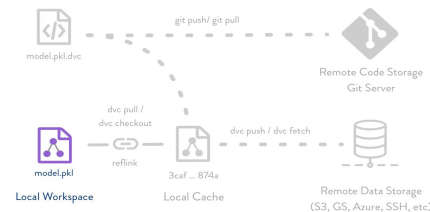
# MLOps

An open source platform  
machine learning life

**MLFlow**

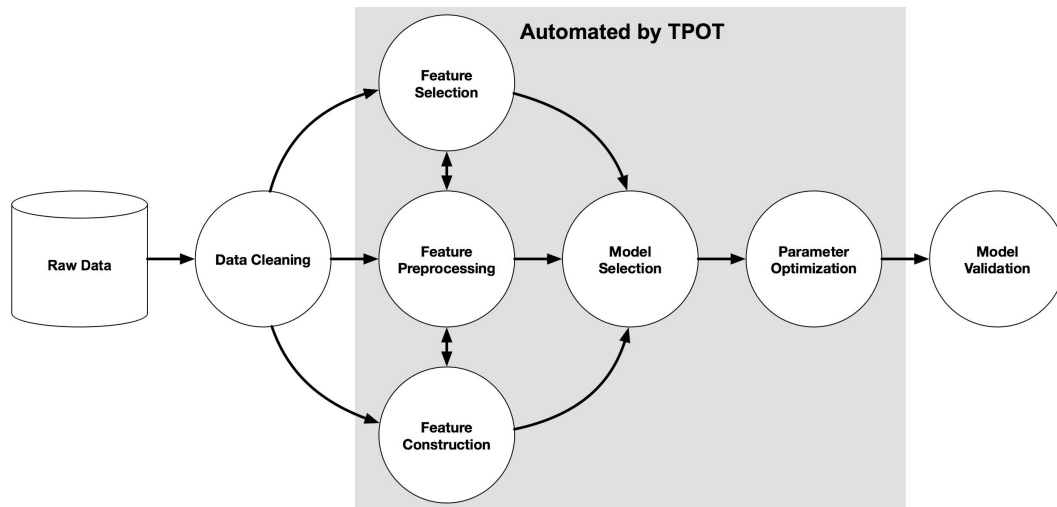


**Continuous Machine  
Learning (CML)  
for Machine Learning  
Projects**

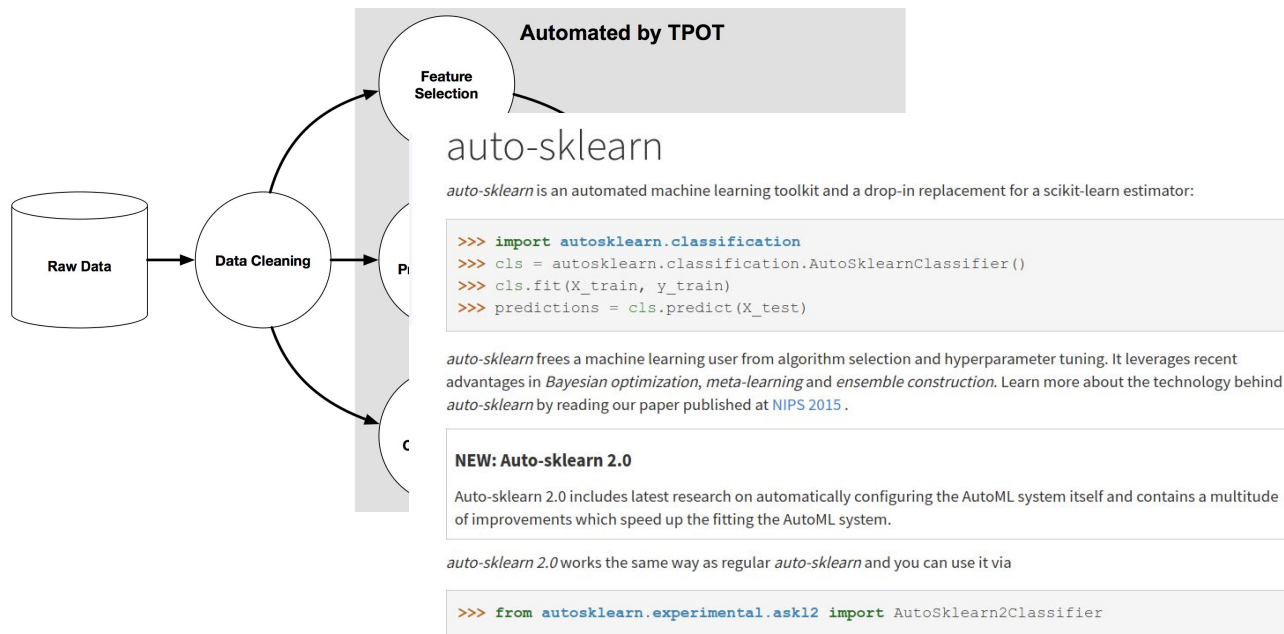


**Data Version  
Control (DVC)**

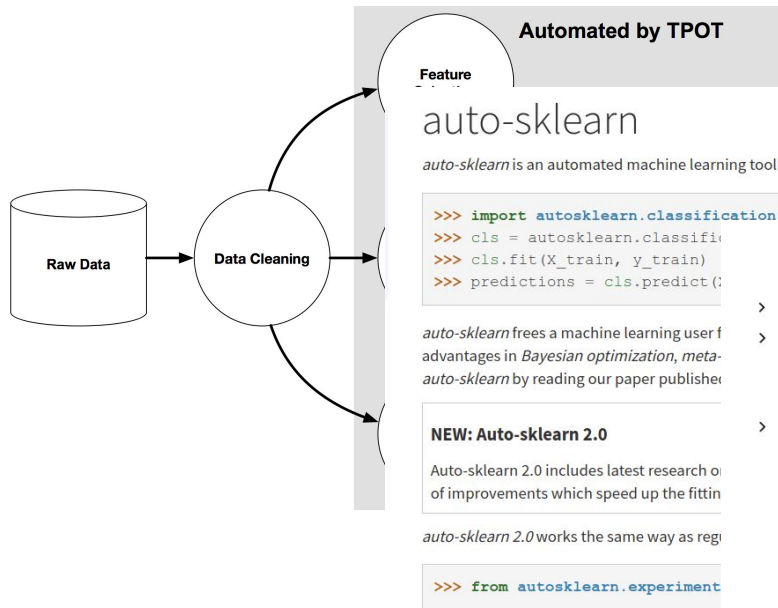
# AutoML



# AutoML



# AutoML



AutoKeras: An AutoML system based on Keras. It is developed by [DATA Lab](#) at Texas A&M University. The goal of AutoKeras is to make machine learning accessible for everyone.

## Example

Here is a short example of using the package.

```
import autokeras as ak

clf = ak.ImageClassifier()
clf.fit(x_train, y_train)
results = clf.predict(x_test)
```

For detailed tutorial please check [here](#)

# Ethics & Fairness in AI

## BIAS in the AI Model

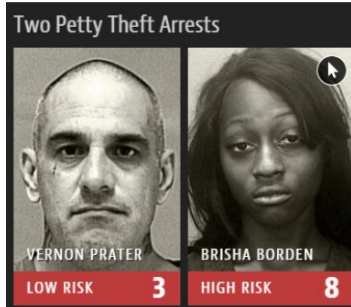


Fig 1: Both the defendants have same criminal history and committed the same crime but have different risk score.

Prediction Fails Differently for Black Defendants		
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Fig 2: White defendant's risk scores are skewed toward lower-risk category, however, for not for black defendants.

# Ethics & Fairness in AI

## BIAS in the AI Model

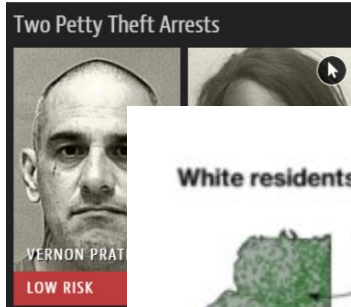


Fig 1: Both the defer committed the same



Fig 2: White defender

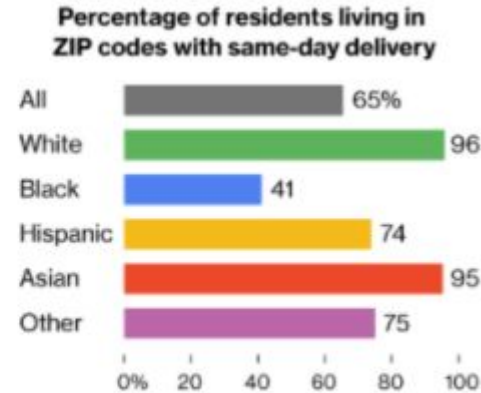
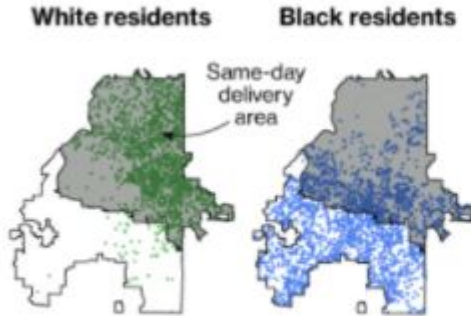
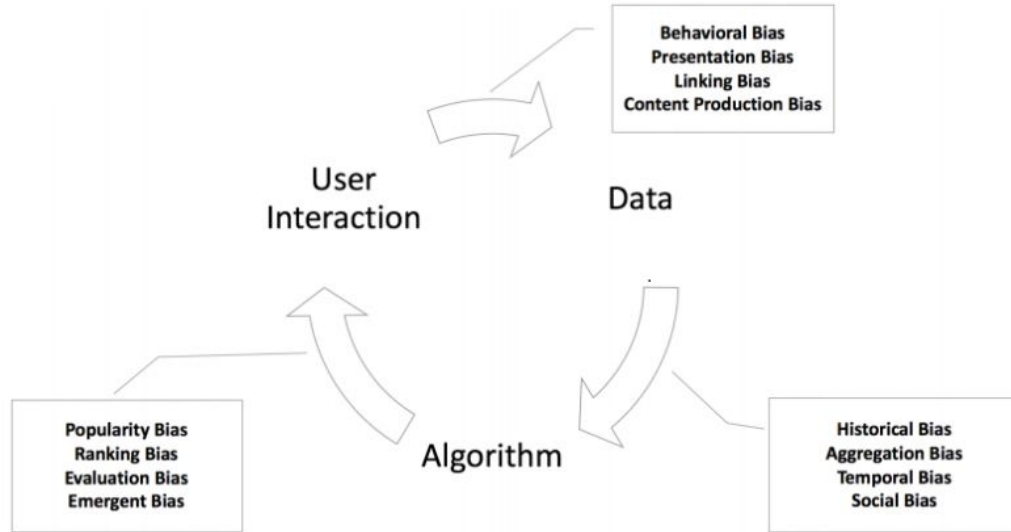


Fig 3: Source - Bloomberg



# Ethics & Fairness in AI

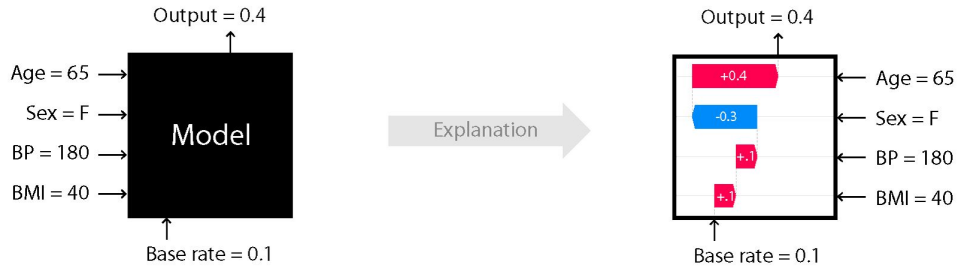


## BIAS in the AI Model

# Explainable AI



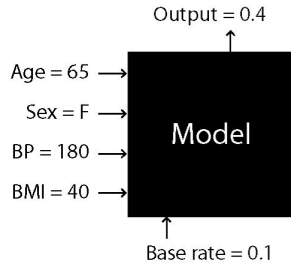
SHAP



# Explainable AI



SHAP



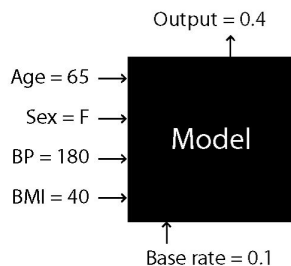
Explanation



# Explainable AI

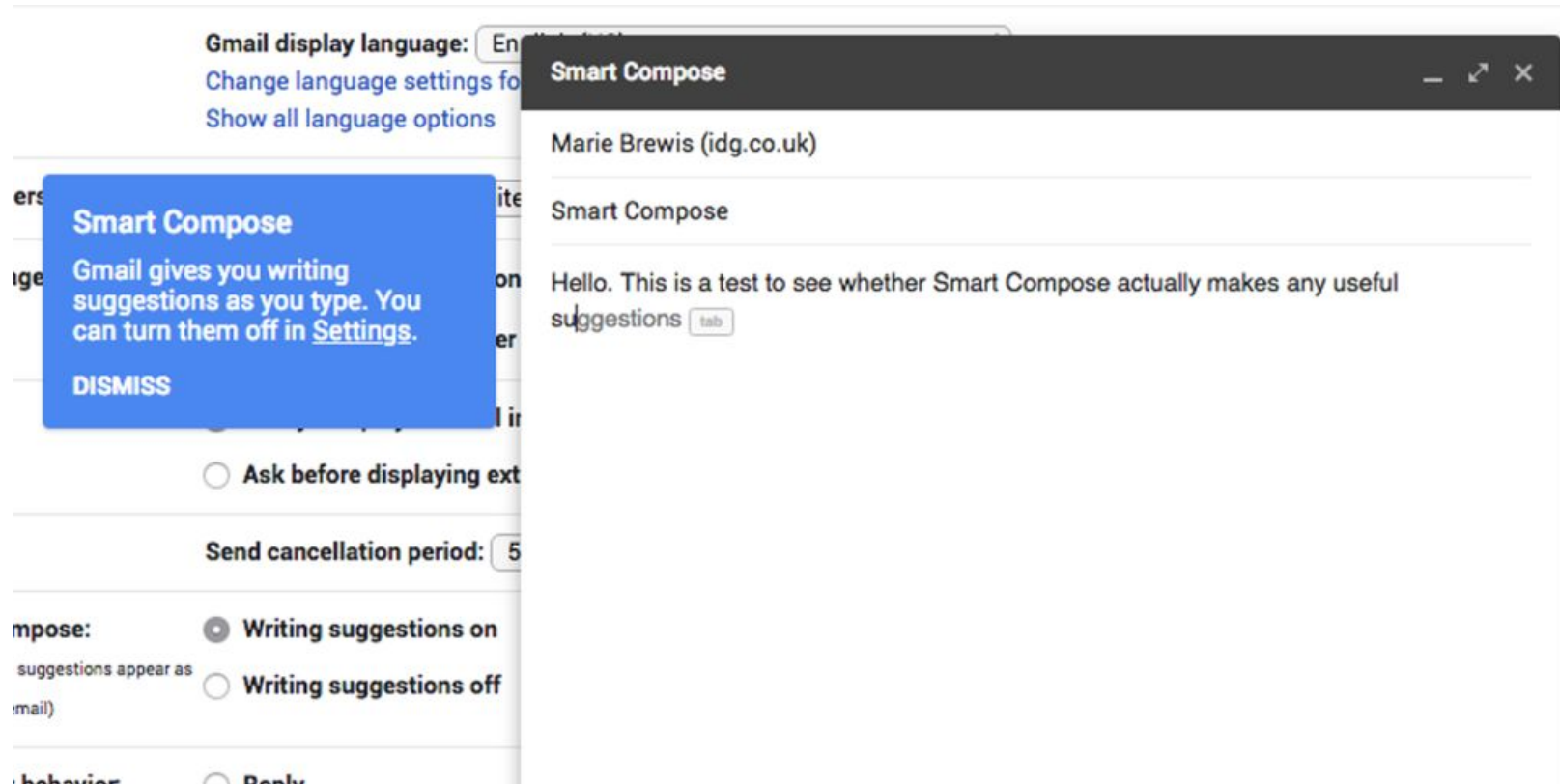


SHAP

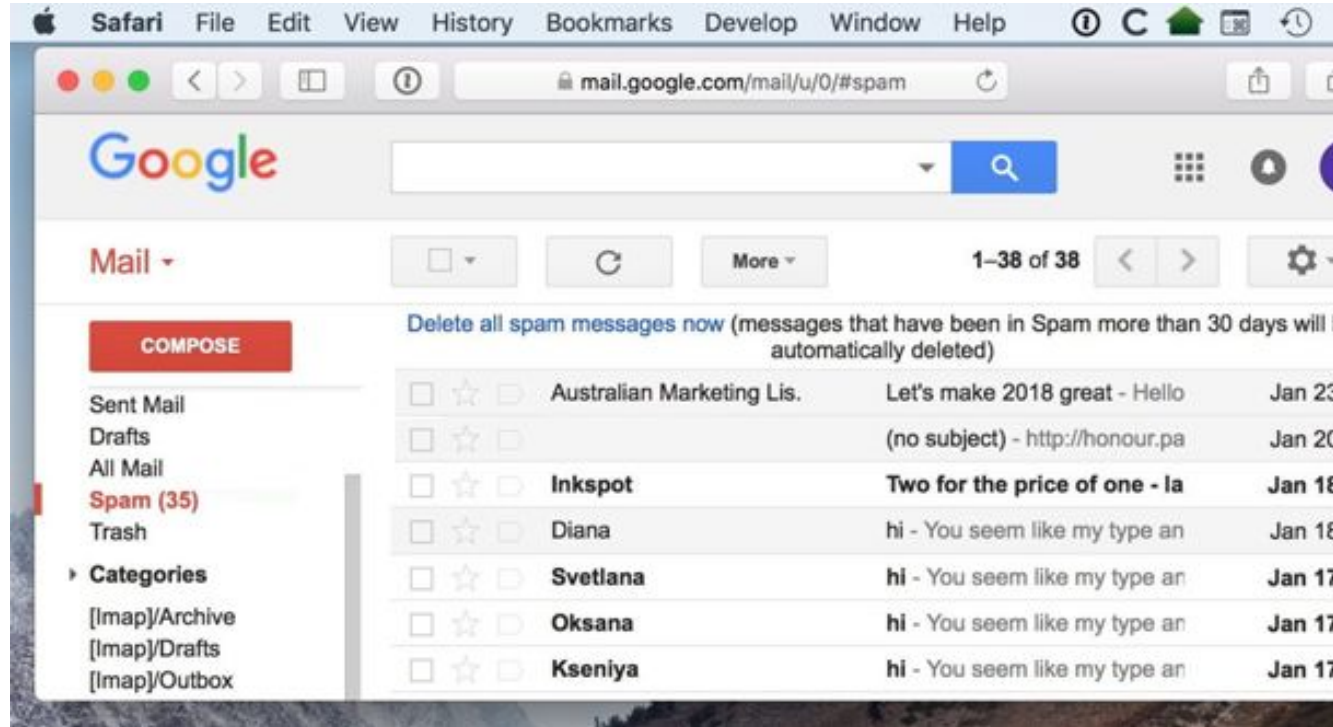


# Applications

# Text Generation



# Spam Detection



# Recommendation Systems

ISBN  
9781787125933

ISBN  
978-1787125933

2nd

Packt Publishing

September 20,  
2017

English

## Frequently bought together



+



+



Total price: **\$130.14**

[Add all three to Cart](#)

- ☒ **This item:** Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn... by Sebastian Raschka Paperback **\$41.99**
- ☒ Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems... by Aurélien Géron Paperback **\$56.16**
- ☒ The Hundred-Page Machine Learning Book by Andriy Burkov Paperback **\$31.99**

Add to List

Share [Embed](#)

Have one to sell?

[Sell on Amazon](#)

**amazon book clubs**  
early access

[Add to book club](#)

Not in a club? [Learn more](#)

## Products related to this item

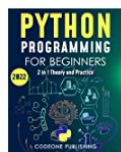
Page 1 of 35

Sponsored



Python Machine Learning: Machine Learning and Deep Learning with Python, ...  
Sebastian Raschka

Newly updated for TensorFlow 2.0, this widely acclaimed book is a reference you'll keep coming back to as you build your machine learning systems  
★★★★★ 384  
Paperback  
**\$49.99**



Python Programming for Beginners: The #1 Python Programming Crash Course for...  
Codeone Publishing  
★★★★★ 356

**#1 Best Seller**  
**\$19.45**



Python Programming for Beginners: The Ultimate Crash Course to Learn Python in 7 Days...  
Andrew Park  
★★★★★ 184

**#1 Best Seller**  
**\$13.95**



Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep L...  
Sebastian Raschka

This book is a comprehensive guide to machine and deep learning using PyTorch's simple to code framework  
★★★★★ 131  
Paperback  
**\$46.79**



Clean Code in Python: Develop maintainable and efficient code, 2nd Edition  
Mariano Anaya  
Paperback

Discover how to apply industry-approved coding practices to design clean, sustainable, and readable real-world Python code  
★★★★★ 89  
**\$44.64**



Hands-On Data Science for Marketing: Improve your marketing strategies with...  
Yoon Hyup Hwang  
★★★★★ 63  
Paperback

**\$44.99**

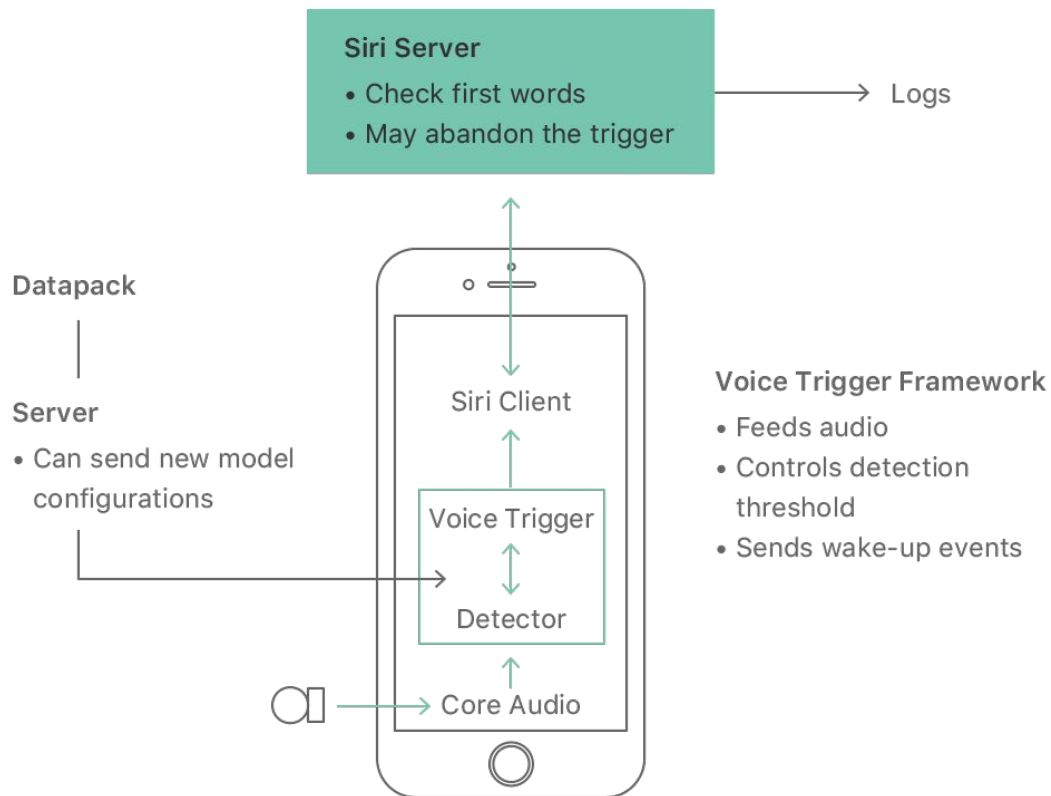


Python Automation Cookbook: 75 Python automation ideas for web scraping, data...  
Jaime Buelta

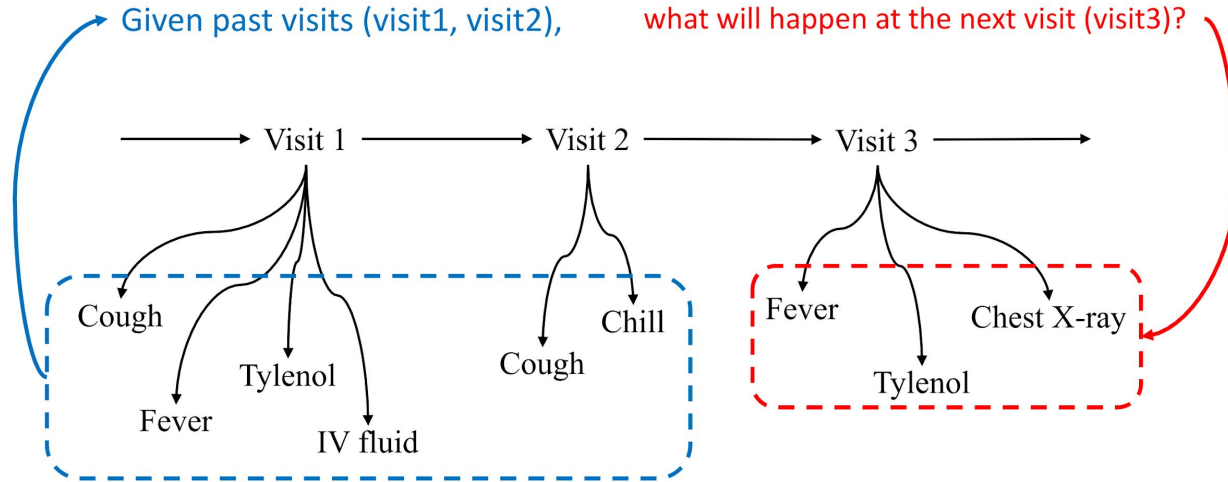
Get a firm grip on core processes including browser automation, web scraping, Word, Excel, and GUI automation with Python 3.8 and higher  
★★★★★ 78  
Paperback  
**\$29.99**



# Speech Recognition



# AI in Healthcare



# AI in Healthcare

## Unstructured Imaging Data: Data

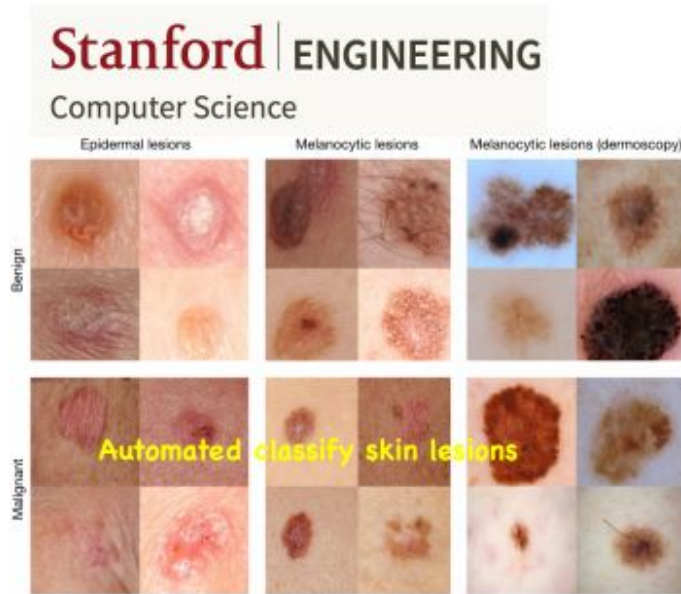
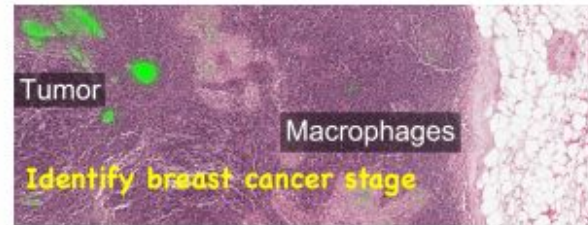
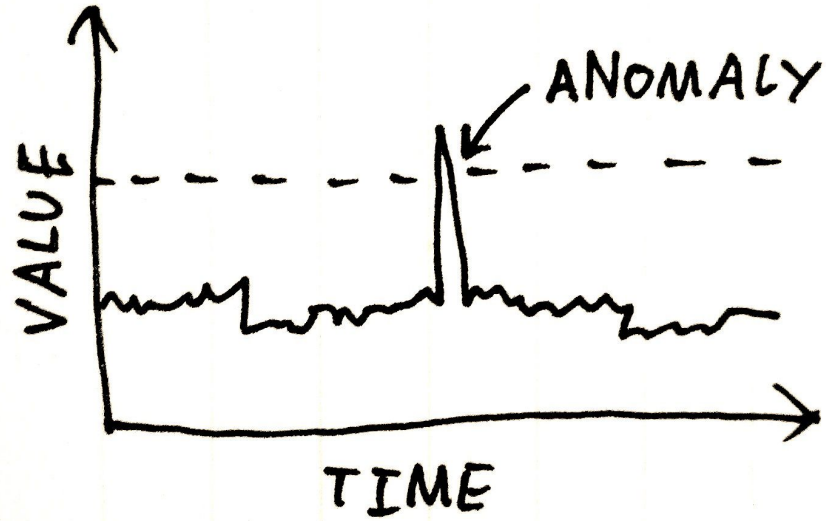


Figure 1. Examples of retinal fundus photographs that are taken to screen for DR. The image on the left is of a healthy retina (A), whereas the image on the right is a retina with referable diabetic retinopathy (B) due to a number of hemorrhages (red spots) present.



# AI in Cybersecurity



# AI in Cybersecurity



(RHSA-2021:0736) Critical: java-1.8.0-ibm security update  
2021-03-04 22:36:10

👍🚫🔒🌐 cvss 7.5  
⚙️ 5.1

ID RHSA-2021:0736

Type redhat

Reporter RedHat

Modified 2021-03-04 22:38:34

## Description

IBM Java SE version 8 includes the IBM Java Runtime Environment and the IBM Java Software Development Kit.

This update upgrades IBM Java SE 8 to version 8 SR6-FP25.

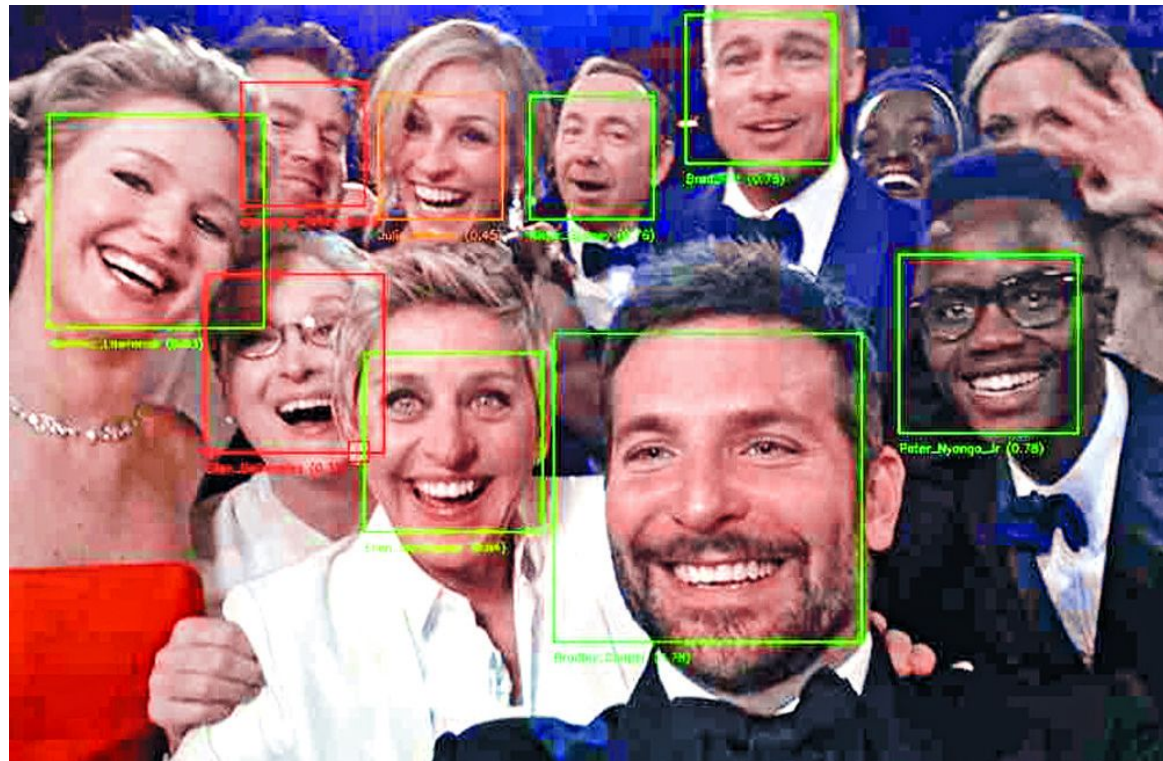
## Security Fix(es):

- IBM JDK: Stack-based buffer overflow when converting from UTF-8 characters to platform encoding (CVE-2020-27221)
- OpenJDK: Unexpected exceptions raised by DOMKeyInfoFactory and DOMXMLSignatureFactory (Security, 8231415) (CVE-2020-2773)
- OpenJDK: Credentials sent over unencrypted LDAP connection (JNDI, 8237990) (CVE-2020-14781)

# Natural Language Processing - Named Entity Recognition

Investigation after his disparaging texts about President Trump PERSON were uncovered, was fired. Credit T.J. Kirkpatrick PERSON for The New York Times By Adam Goldman ORG and Michael S. Schmidt Aug PERSON . 13 CARDINAL , 2018 WASHINGTON CARDINAL — Peter Strzok PERSON , the F.B.I. GPE senior counterintelligence agent who disparaged President Trump PERSON in inflammatory text messages and helped oversee the Hillary Clinton PERSON email and Russia GPE investigations, has been fired for violating bureau policies, Mr. Strzok PERSON said Monday DATE . Mr. Trump and his allies seized on the texts — exchanged during the 2016 DATE campaign with a former F.B.I. GPE Lisa Page — in PERSON assailing the Russia GPE investigation as an illegitimate “witch hunt.” Mr. Strzok PERSON , who rose over 20 years ago DATE at the F.B.I. GPE to become one of its most experienced counterintelligence agents, was a key figure in the early months DATE of the inquiry. Along with writing the texts, Mr. Strzok PERSON was accused of sending a highly sensitive search warrant to his personal email account. T

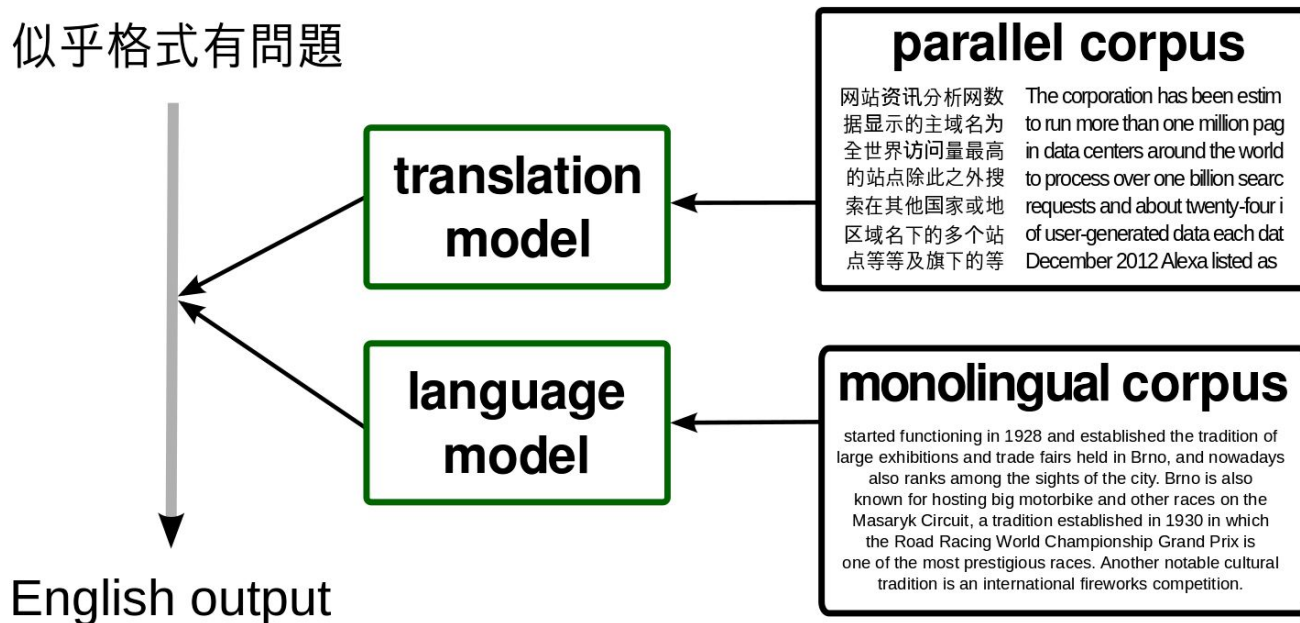
# Face Recognition





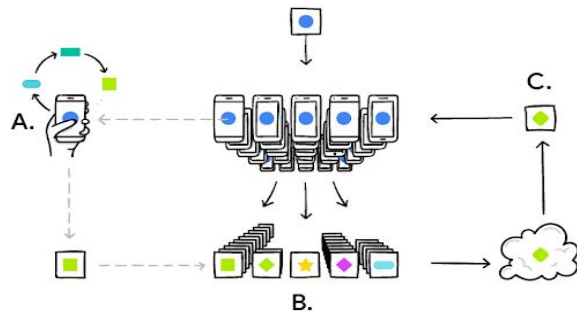
# Machine Translation

似乎格式有問題





# Privacy Preserving ML



## TensorFlow Privacy

This repository contains the source code for TensorFlow Privacy, a Python library that includes implementations of TensorFlow optimizers for training machine learning models with differential privacy. The library comes with tutorials and analysis tools for computing the privacy guarantees provided.

The TensorFlow Privacy library is under continual development, always welcoming contributions. In particular, we always welcome help towards resolving the issues currently open.

### Latest Updates

2020-12-21: A new vectorized version of the TF 2 optimizer is available, which can deliver much faster performance. We recommend trying it first, and to fall back to using the original non-vectorized version only if this fails. We are thankful to the authors of this paper for spurring this change.

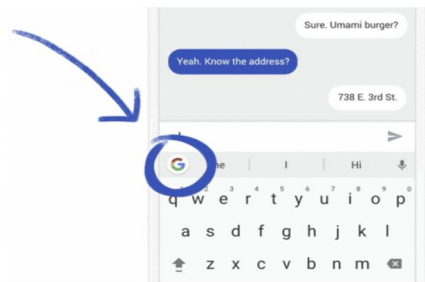
### Setting up TensorFlow Privacy

#### Dependencies

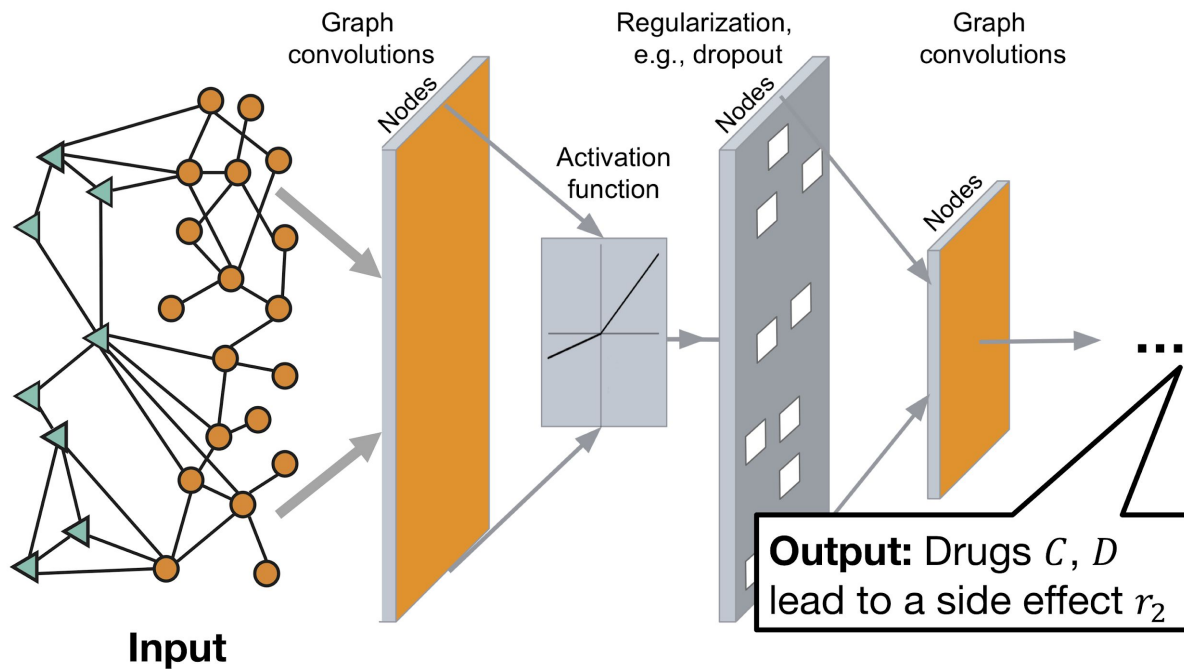
This library uses TensorFlow to define machine learning models. Therefore, installing TensorFlow ( $\geq 1.14$ ) is a prerequisite. You can find instructions [here](#). For better performance, it is also recommended to install TensorFlow with GPU support (detailed instructions on how to do this are available in the [TensorFlow installation documentation](#)).

In addition to TensorFlow and its dependencies, other prerequisites are:

We're currently testing Federated Learning in [Gboard on Android](#), the Google Keyboard. When Gboard shows a suggested query, your phone locally stores information about the current context and whether you clicked the suggestion. Federated Learning processes that history on-device to suggest improvements to the next iteration of Gboard's query suggestion model.



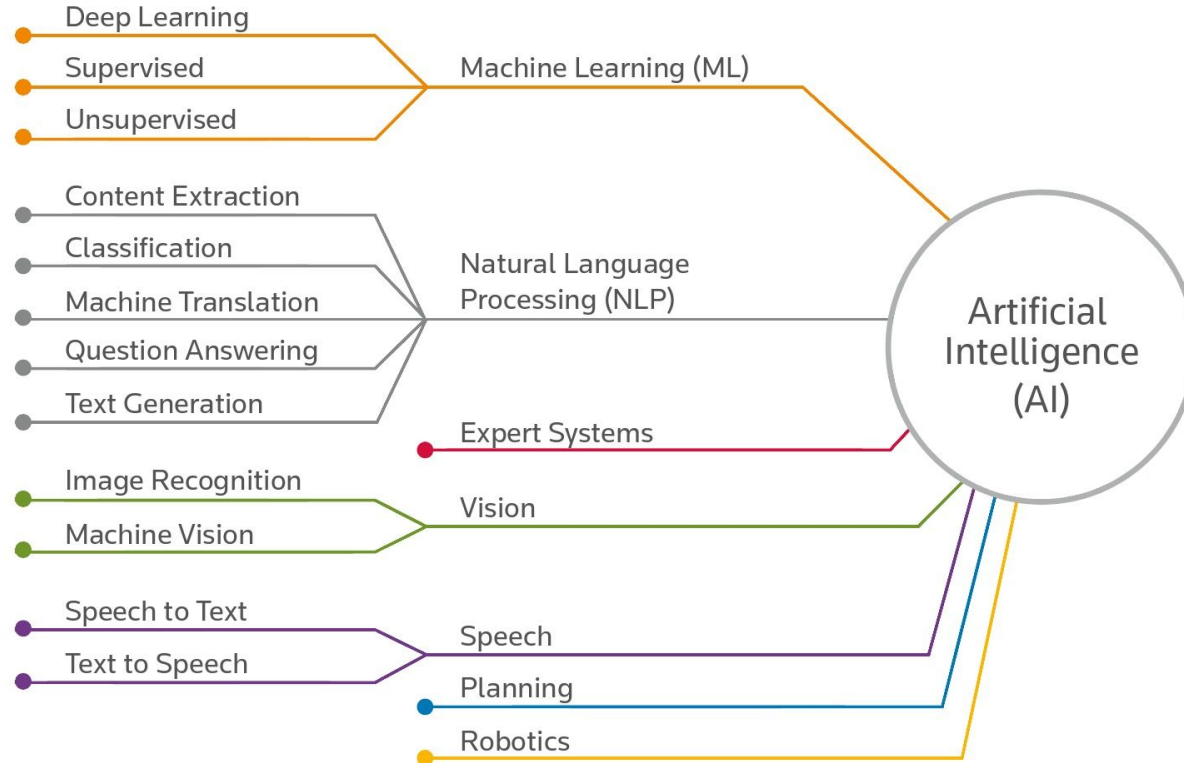
# Graph ML



# Many Other Applications

- Finance
- E-commerce
- Information Extraction
- Social Networks
- Web Search
- Computer Vision and robotics
- Computational Biology
- Fraud Detection
- Etc.

# What is Artificial Intelligence (AI) ?



# What is Machine Learning?

- Learning is any process by which a system improves performance from experience”
  - Herbert Simon
- Another definition is done by Tom Mitchell
  - Machine Learning is the study of the algorithms that
    - improve their performance  $P$
    - at some task  $T$
    - with experience  $E$
  - A well-defined learning task is given by  $\langle P, T, E \rangle$

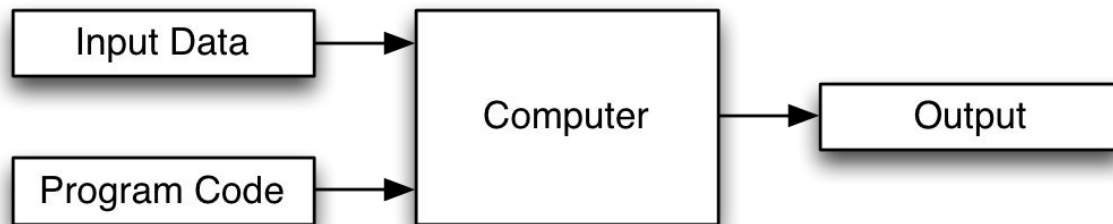
T: Recognizing hand-written words

P: Percentage of words correctly classified

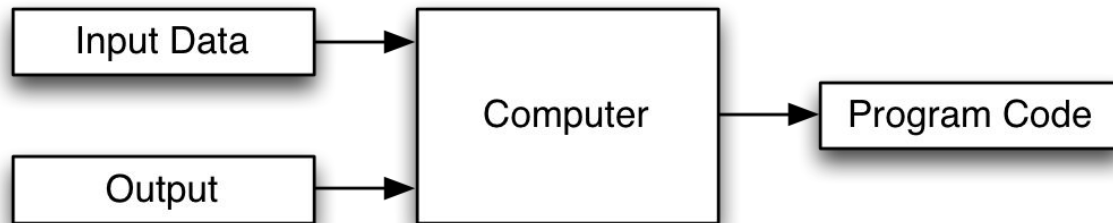
E: Database of human-labeled images of handwritten words

# Traditional Programming vs Machine Learning

## Traditional Software Development



## Machine Learning Programming



# Why Study Machine Learning?

- Algorithms
  - Many basic effective and efficient algorithms available.
- Data
  - Large amounts of on-line data available.
- Computing
  - Large amounts of computational resources available.

# Machine Learning in a Nutshell

- Every machine learning algorithm consists of the following basic steps:
  - Data collection
  - Representation
  - Modeling
  - Evaluation
  - Optimization



# Rules of Machine Learning

- <https://developers.google.com/machine-learning/guides/rules-of-ml>

## Before Machine Learning

**Rule #1: Don't be afraid to launch a product without machine learning.**

Machine learning is cool, but it requires data. Theoretically, you can take data from a different problem and then tweak the model for a new product, but this will likely underperform basic **heuristics**. If you think that machine learning will give you a 100% boost, then a heuristic will get you 50% of the way there.

For instance, if you are ranking apps in an app marketplace, you could use the install rate or number of installs as heuristics. If you are detecting spam, filter out publishers that have sent spam before. Don't be afraid to use human editing either. If you need to rank contacts, rank the most recently used highest (or even rank alphabetically). If machine learning is not absolutely required for your product, don't use it until you have data.

# What is Machine Learning?

**Samples**  
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

**Features**  
(attributes, measurements, dimensions)

**Class labels**  
(targets)

**Petal**

**Sepal**



# Machine Learning Terminology

- **Dataset:** A table with the data from which the machine learns. The dataset contains the features and the target to predict.
- **Instance:** The thing about which you want to make a prediction. For example, the instance might be a web page that you want to classify as either "about cats" or "not about cats".
- **Label:** An answer for a prediction task either the answer produced by a machine learning system, or the right answer supplied in training data. For example, the label for a web page might be "about cats".
- **Feature:** A property of an instance used in a prediction task. For example, a web page might have a feature "contains the word 'cat'".

# Machine Learning Terminology

- **Example:** An instance (with its features) and a label.
- **Model:** A statistical representation of a prediction task. You train a model on examples then use the model to make predictions.
- **Metric:** A number that you care about. May or may not be directly optimized.
- **Objective:** A metric that your algorithm is trying to optimize.
- **Pipeline:** The infrastructure surrounding a machine learning algorithm. Includes gathering the data from the front end, putting it into training data files, training one or more models, and exporting the models to production.
- **Prediction:** what the ML model “guesses” what the target value should be based on the given features.

# Next Class:

Continue with Machine Learning Concepts & Machine Learning by Examples