Python for Data Science

Introduction to Machine Learning

Dr. Ilkay Altintas

- Explain what machine learning is
- List three applications of machine learning encountered in everyday life

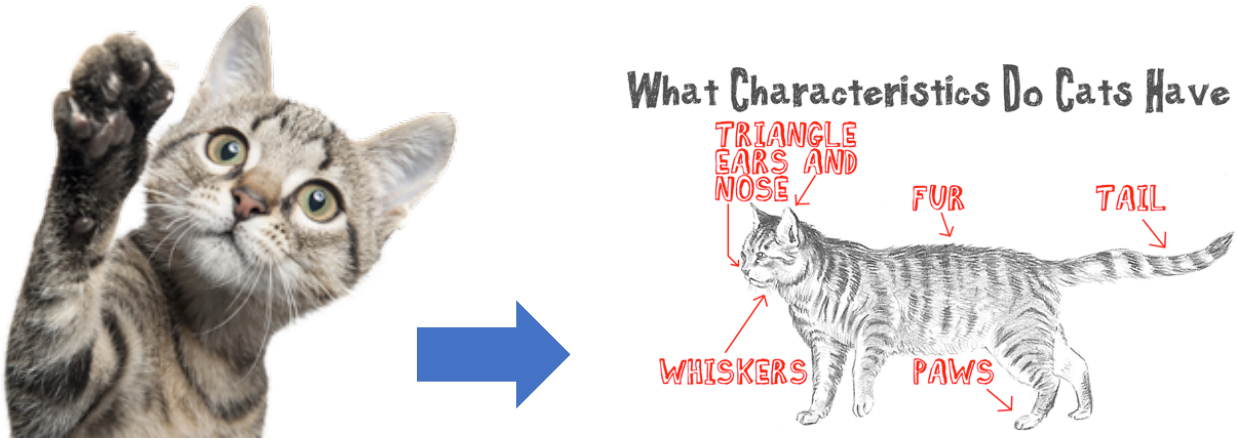
Machine Learning is…

… learning from data

Machine Learning is…

… learning from data

… on its own



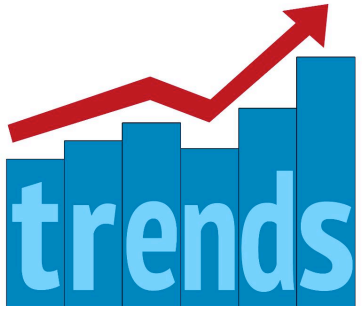
Machine Learning is…

… learning from data

… on its own

… discovering hidden patterns

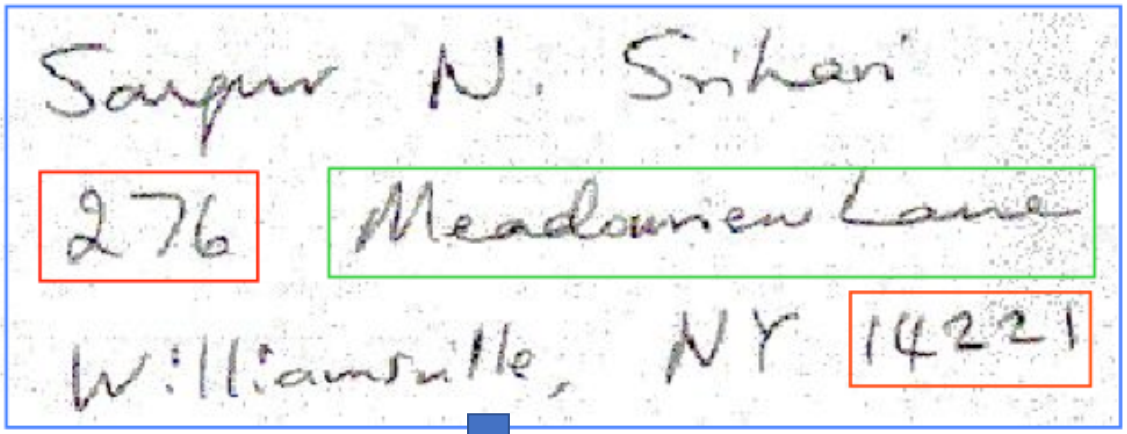
… data-driven decisions



Applications of Machine Learning

Credit Card Fraud Detection

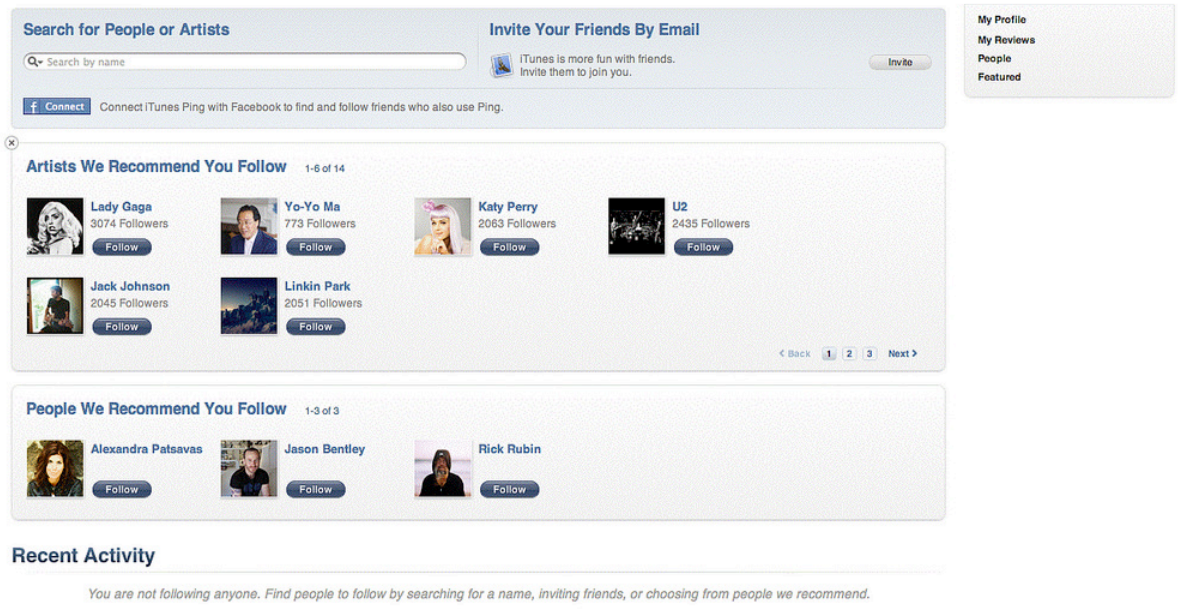
Handwritten Digit Recognition



ZIP Code: 14221

Primary number: 276

Recommendations on Websites



Machine Learning and Data Science

- Data mining
- Predictive analytics
- Data science
- Big Data

Categories of Machine Learning

- Describe the main categories of machine learning techniques
- Summarize how supervised learning differs from unsupervised learning

Classification

Goal: Predict category

Sunny

Windy

Rainy

Cloudy

Image source:
http://www.davidson.k12.nc.us/parents___students/inclement_weather

Regression

Goal: Predict numeric value

Cluster Analysis

Goal: Organize similar items into groups.

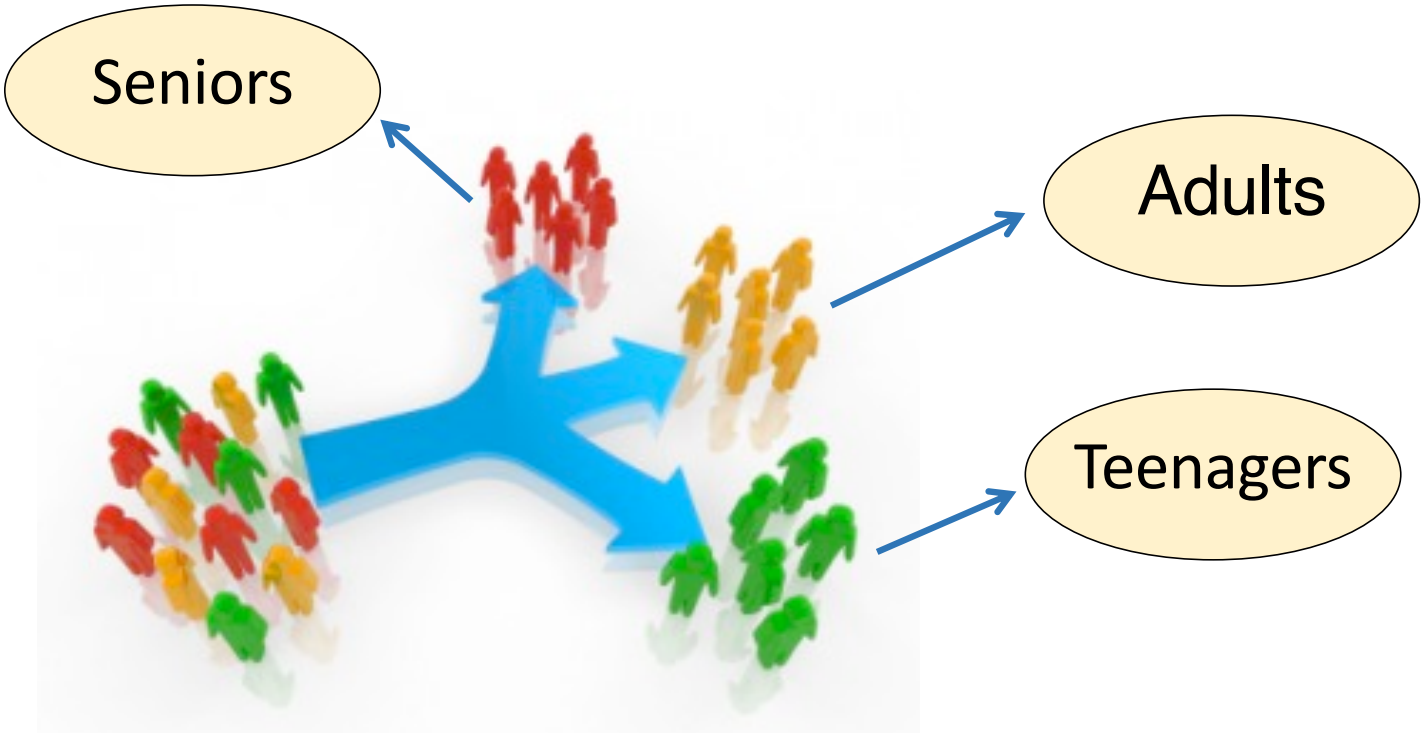


Image source: http://www.monetate.com/blog/the-intrinsic-value-of-customer-segmentation

Association Analysis

Goal: Find rules to capture associations between items.

Categories of Machine Learning Techniques

- Classification
- Cluster Analysis
- Regression
- Association Analysis

Supervised vs. Unsupervised

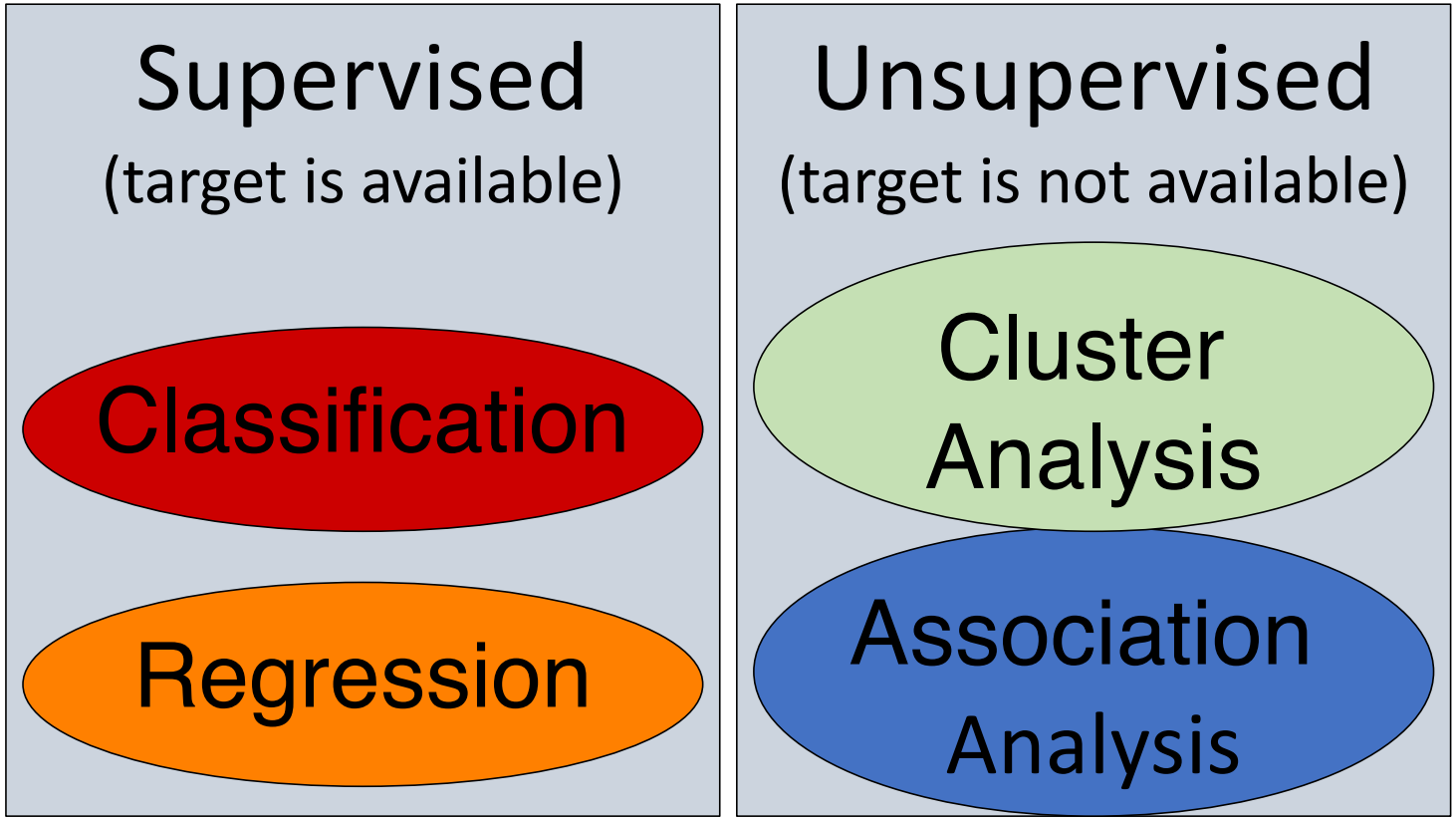
- Supervised Approaches
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

Supervised vs. Unsupervised

- Supervised Approaches
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

- Unsupervised Approaches
  - Target is unknown or unavailable
  - ‘unlabeled’ data
  - Cluster analysis & association analysis are unsupervised.

Categories of Machine Learning Techniques



scikit-learn

Machine Learning in Python

- Identify key strengths of scikit-learn
- Explain why it is a leading library for Machine Learning
- Navigate your way to find the right tool in scikit-learn
- Search for tutorials that provide problem specific examples using library functions

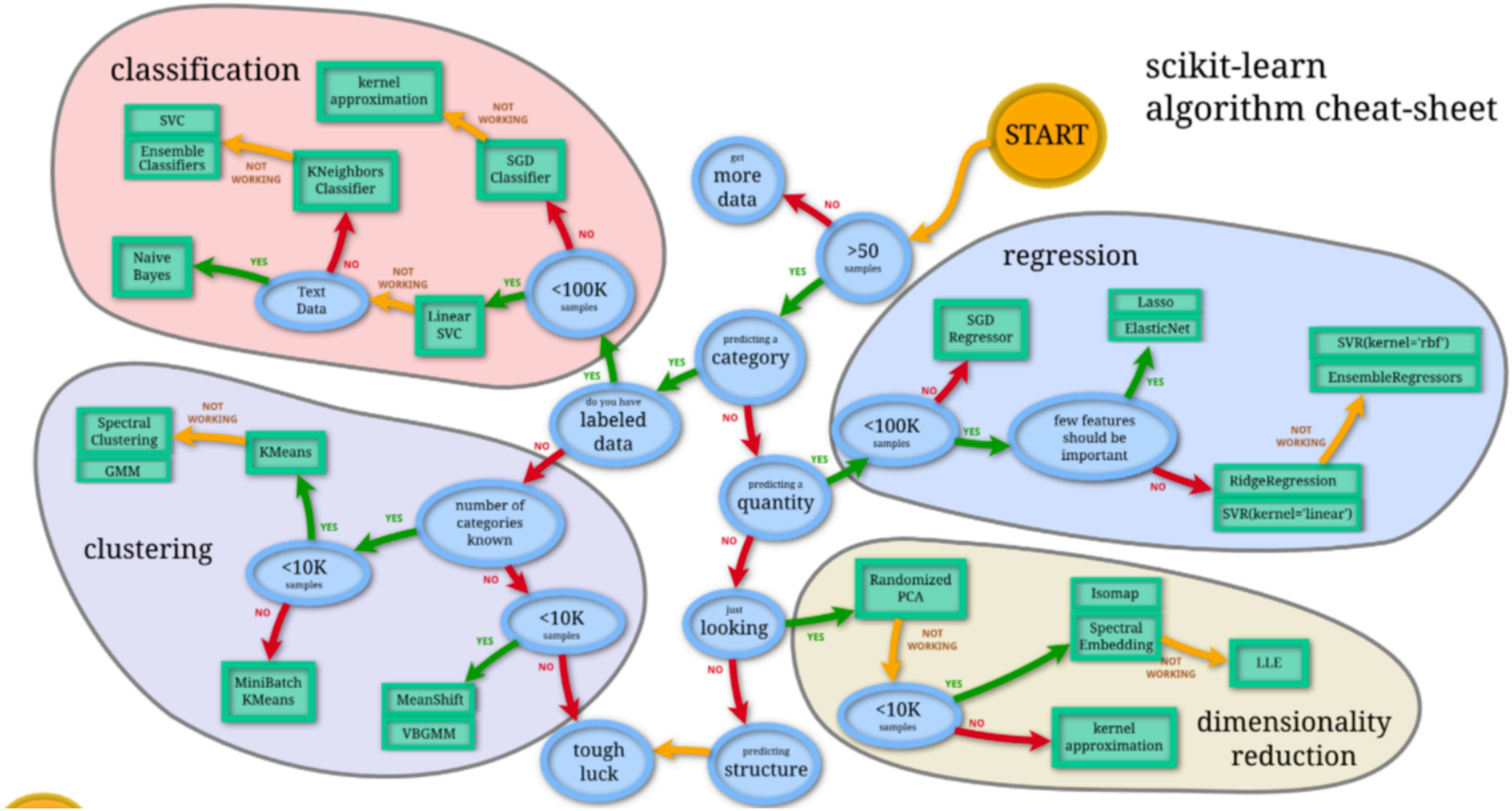
scikit-learn

- Open source library for Machine Learning in Python
- Built on top of NumPy, SciPy, matplotlib
- Active community for development
- Improved continuously by developers

Preprocessing Tools

- Utility Functions for
  - Transforming raw feature vectors to suitable format
- Provides API for
  - Scaling of features: remove mean and keep unit variance
  - Normalization to have unit norm
  - Binarization to turn data into 0 or 1 format
  - Handling of missing values
  - Generating higher order features
  - Build custom transformations

Python for Data Science



Provides organized tutorials with specifics.

Quick Start

A very short introduction into machine learning problems and how to solve them using scikit-learn. Introduced basic concepts and conventions.

User Guide

The main documentation. This contains an in-depth description of all algorithms and how to apply them.

Other Versions

- scikit-learn 0.18 (stable)
- scikit-learn 0.19 (development)
- scikit-learn 0.17
- scikit-learn 0.16
- scikit-learn 0.15

Tutorials

Useful tutorials for developing a feel for some of scikit-learn's applications in the machine learning field.

API

The exact API of all functions and classes, as given by the docstrings. The API documents expected types and allowed features for all functions, and all parameters available for the algorithms.

Additional Resources

Talks given, slide-sets and other information relevant to scikit-learn.

Development

Information on how to contribute. This also contains useful information for advanced users, for example how to build their own estimators.

Flow Chart

A graphical overview of basic areas of machine learning, and guidance which kind of algorithms to use in a given situation.

FAQ

Frequently asked questions about the project and contributing.

Related packages

Other machine learning packages for Python and related projects. Also algorithms that are slightly out of scope or not well established enough for scikit-learn.

http://scikit-learn.org/stable/documentation.html

Clustering

http://scikit-learn.org/stable/modules/clustering.html#clustering

- sklearn.cluster gives algorithms for grouping of unlabeled data

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

Dimensionality Reduction

- Enables you to reduce features while preserving variance
- scikit-learn has capabilities for:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition
  - Factor Analysis
  - Independent Component Analysis
  - Matrix Factorization
  - Latent Dirichlet Allocation

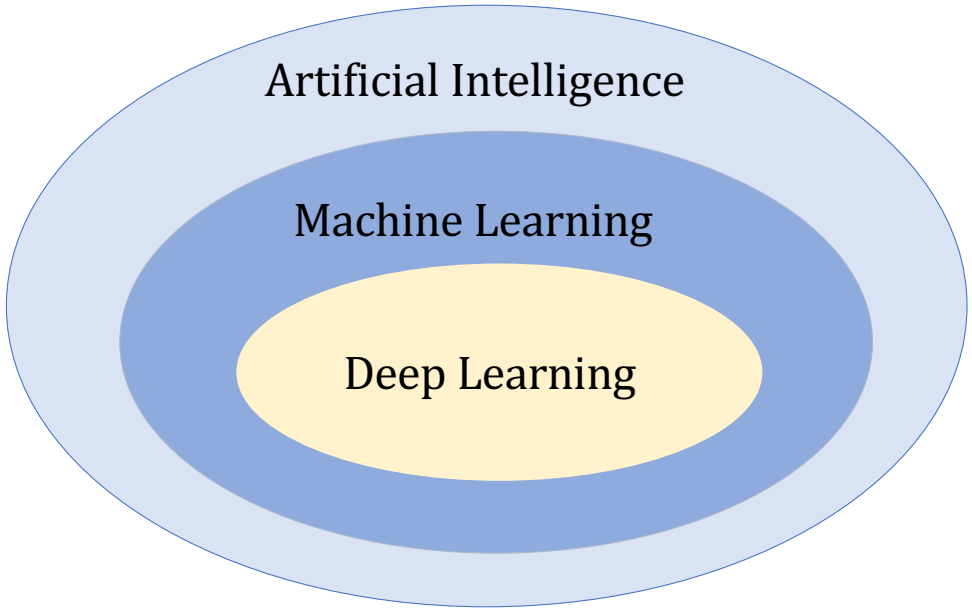
Model Selection

- Provides methods for Cross Validation
- Library functions for tuning hyper parameters
- Model Evaluation mechanisms to measure model performance
- Plotting methods for visualizing scores to evaluate models

Summary of scikit-learn

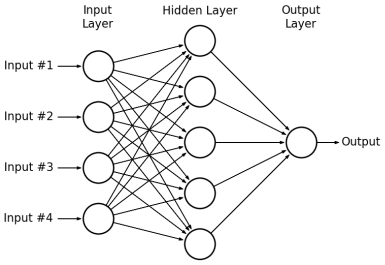
- Extensive set of tools for full pipeline in Machine Learning
- Dependable due to community support
- Provides easy to use API for training, and making predictions
- Collection of the best, most popular, algorithms in one place

Deep Learning

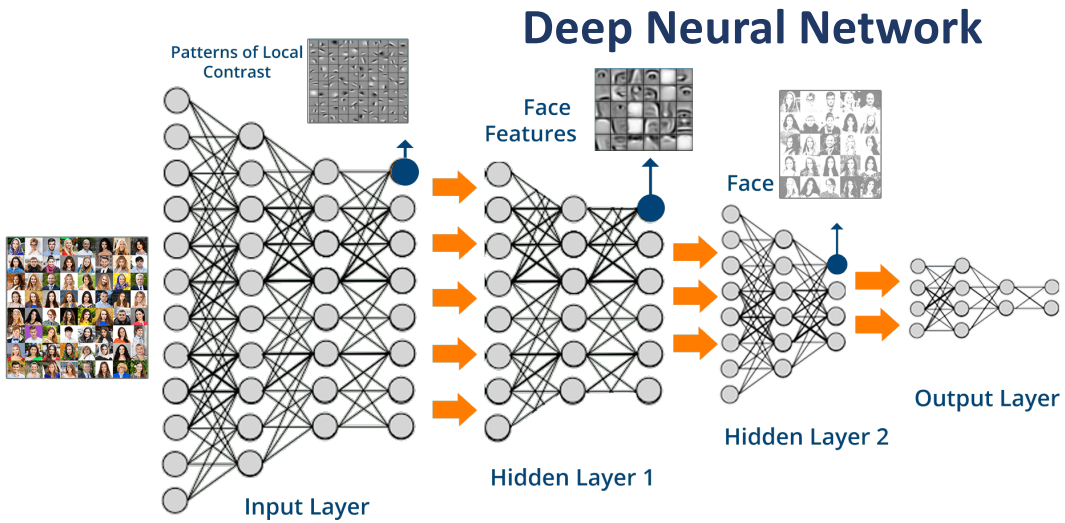


Deep Learning is a subfield of Machine Learning

Deep Learning



http://www.astroml.org/book_figures/appendix/fig_neural_network.html



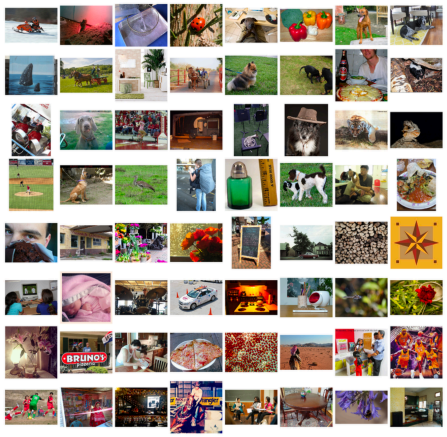
https://cdn.edureka.co/blog/wp-content/uploads/2017/05/Deep-Neural-Network-What-is-Deep-Learning-Edureka.png

- 'Deep' refers to the many layers in model
  - Allows for learning at different levels of abstraction
  - Leads to automatic feature learning & excellent performance

Applications of Deep Learning

- Image classification
- Speech recognition
- Handwriting recognition
- Self-driving cars
- Drug design
- Precision medicine
- Disease detection
- Targeted ads
- Stock market analysis

ImageNet



- Database
  - Developed for computer vision research
  - > 14,000,000 images hand-annotated
  - > 22,000 categories
- ILSVRC History
  - Started in 2010
  - Image classification task: 1,000 object categories
  - Image classification error rate
    - 2011: ~25% (conventional image processing techniques)
    - 2012: 15.3% (AlexNet)
    - 2015: 3.57% (ResNet; better than human performance)
    - 2016: 2.99% (16.7% error reduction)
    - 2017: 2.25% (23.3% error reduction)

Python Deep Learning Libraries

- TensorFlow
  - https://www.tensorflow.org/
  - ML framework developed by Google
- Keras
  - https://keras.io/
  - High-level NN API. Runs on TensorFlow, CNTK, or Theano
- PyTorch
  - https://pytorch.org/
  - ML framework developed by Facebook
- Caffe & Caffe2
  - https://caffe2.ai/docs/caffe-migration.html
  - Caffe2 now merged into PyTorch
- Apache MXNet
  - https://mxnet.apache.org/
  - DL framework used by AWS