USTH MASTER SPACE

Earth Observation - Astrophysics - Satellite Technologies

USTH

VIETNAM FRANCE UNIVERSITY

Lecture 5: Basic Machine Learning using Scikit-learn.

Hoàng Đức Thường

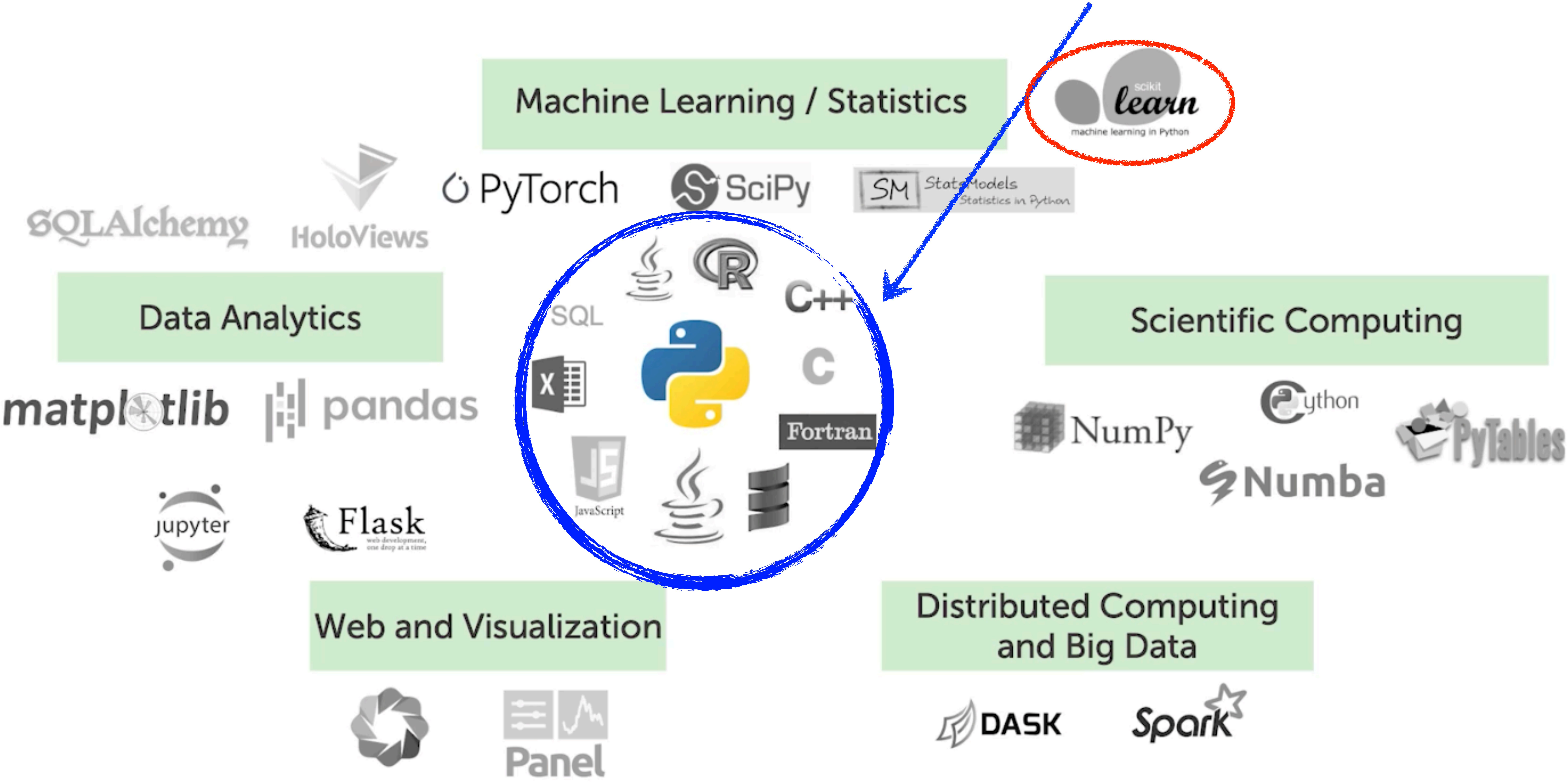
Department of Space and Applications (DSA), USTH

Hanoi, 2021

Lecture 5: Basic Machine Learning using Scikit-learn.

In this class I will briefly introduce Scikit-learn tool using for machine learning and data analysis.

Data Science & Anaconda



Machine Learning / Statistics

Data Analytics

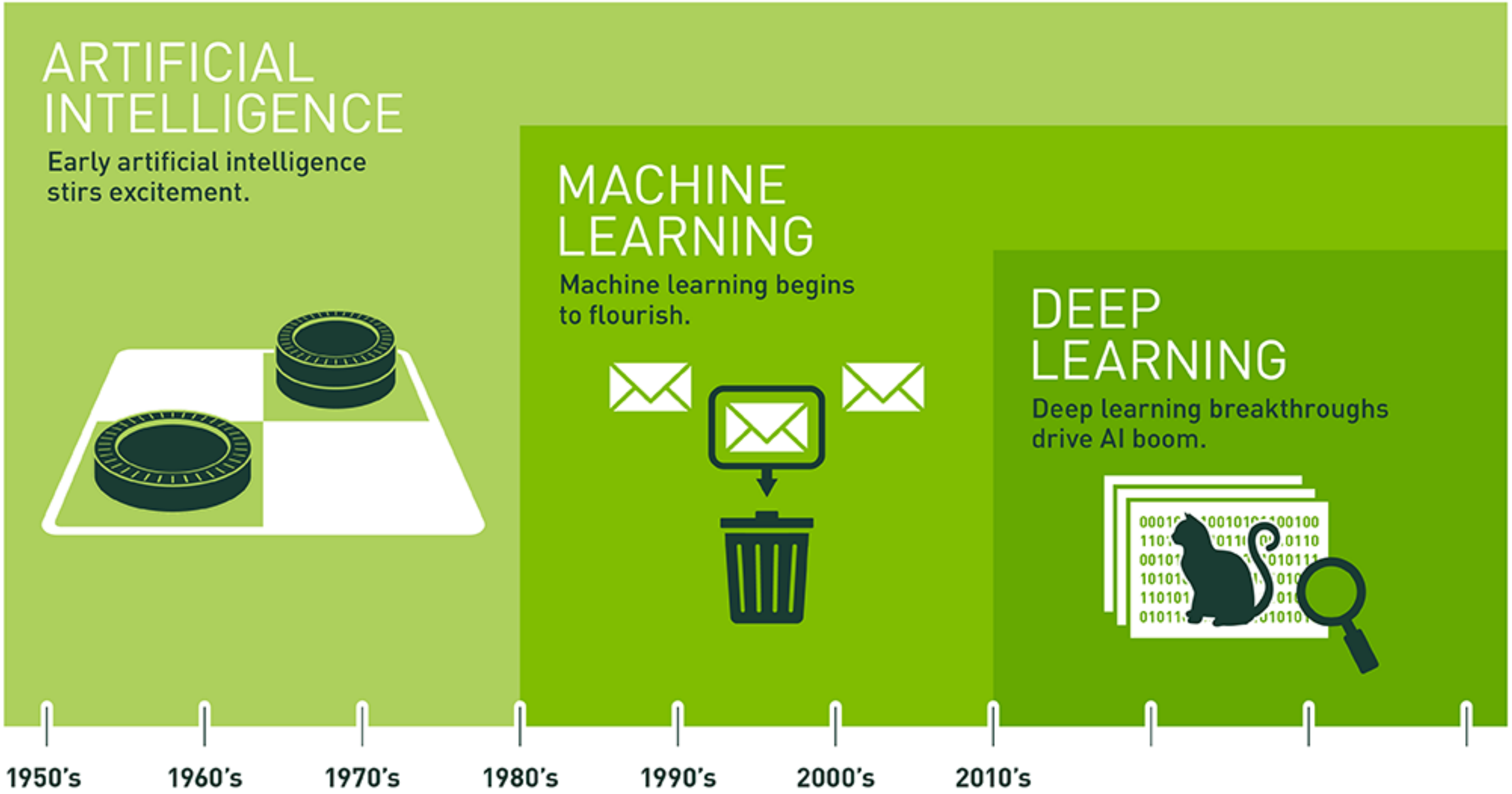
Scientific Computing

Web and Visualization

Distributed Computing and Big Data

AI, ML, DL

The relationship of AI, Machine Learning (ML), and Deep Learning (DL).



ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.

MACHINE LEARNING

Machine learning begins to flourish.

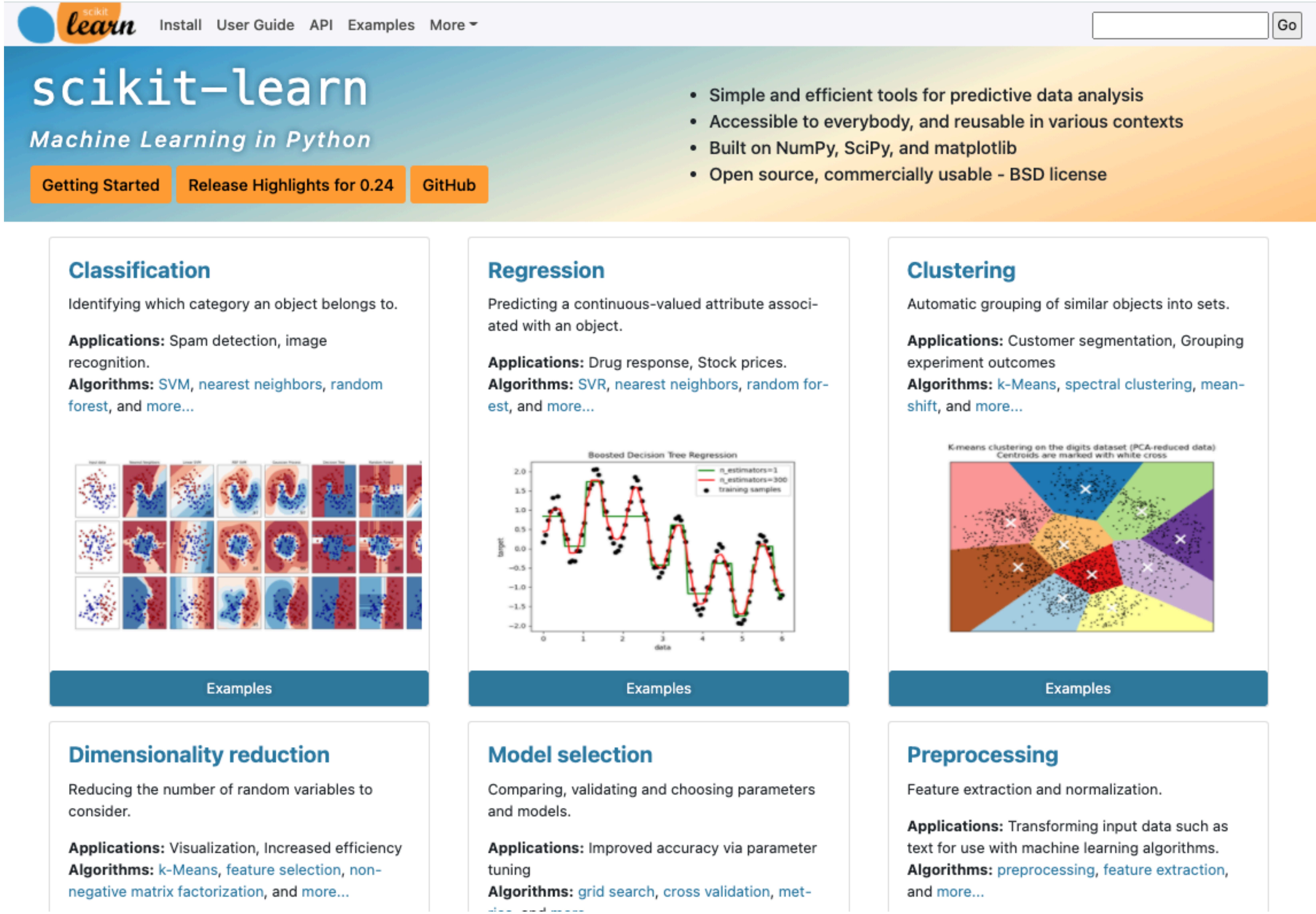
DEEP LEARNING

Deep learning breakthroughs drive AI boom.

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Source: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Scikit-learning



scikit-learn

Machine Learning in Python

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: SVM, nearest neighbors, random forest, and more...

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, spectral clustering, mean-shift, and more...

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning
Algorithms: grid search, cross validation, met-

Preprocessing

Feature extraction and normalization.

Applications: Transforming input data such as text for use with machine learning algorithms.
Algorithms: preprocessing, feature extraction, and more...

Source: https://scikit-learn.org/stable/index.html

Scikit-learning for Machine Learning

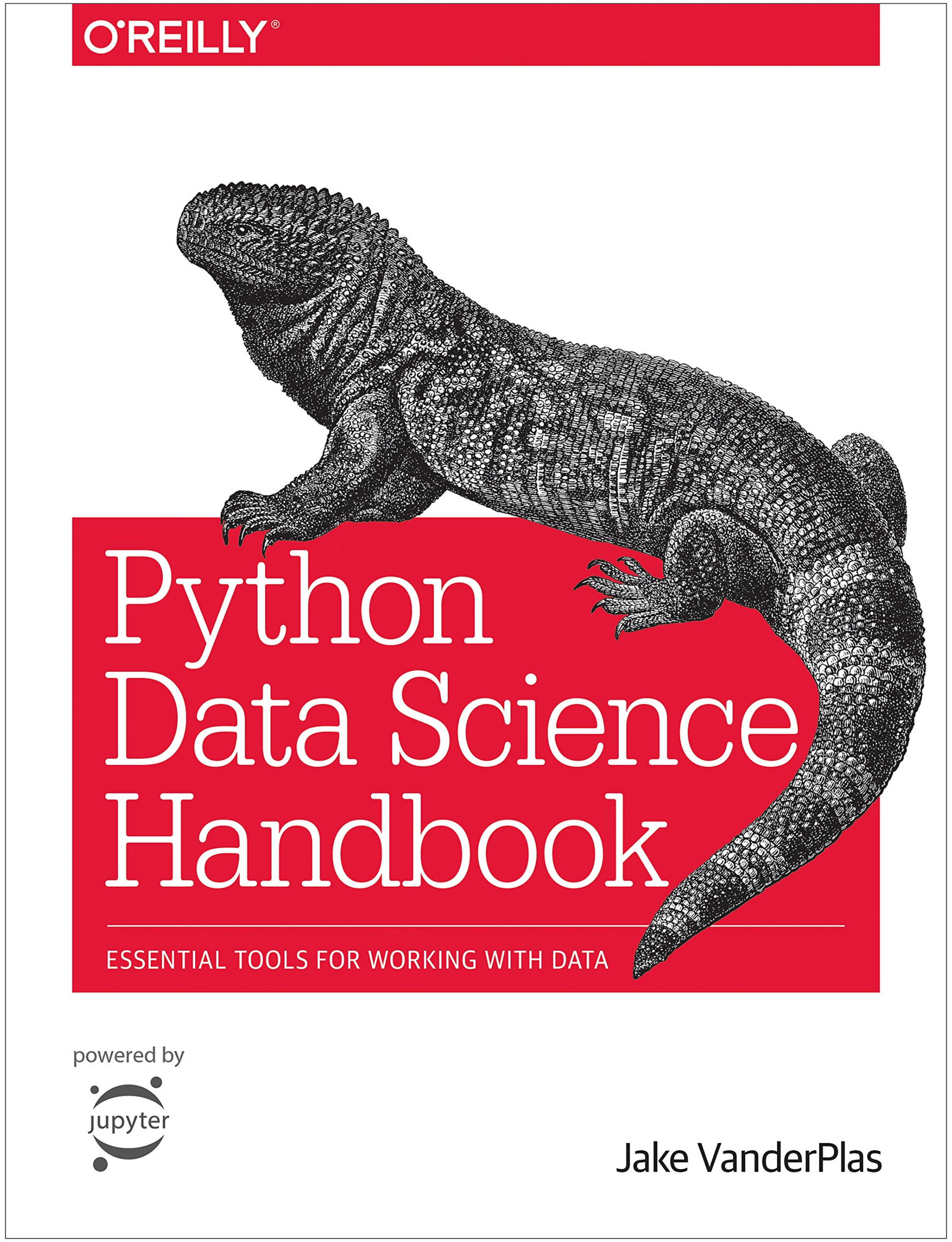
- Scikit-learn is a tool for data analysis. In general, a learning problem considers a set of N samples data and then predict unknown quantities.Learning a problem can be classified into:

1. **Supervised learning**: Data have additional attributes, and the problem either:
    - Classification: Identifying which category of an object belongs to. Example: Spam detection
    - **Regression**: Predicting an attribute associated with an object. Example: Parameter estimations, minimum chi-square, stock price prediction.

2. **Unsupervised learning**: Training data to discover groups of similar examples.
    - **Clustering**: Automatic group of similar onbjects into sets, example: Customer segmentation. Methods: K-means.
    - Model selection: Comparing, validating, choosing parameters and model, example PCA method.

Scikit-learning book

GitHub: https://github.com/jakevdp/PythonDataScienceHandbook



O'REILLY®

Python Data Science Handbook

ESSENTIAL TOOLS FOR WORKING WITH DATA

powered by jupyter

Jake VanderPlas

Scikit-learning example

The K-means algorithm divides a set of N samples of variable X into K number of clusters C, each cluster described by the mean μ_j (centroids).

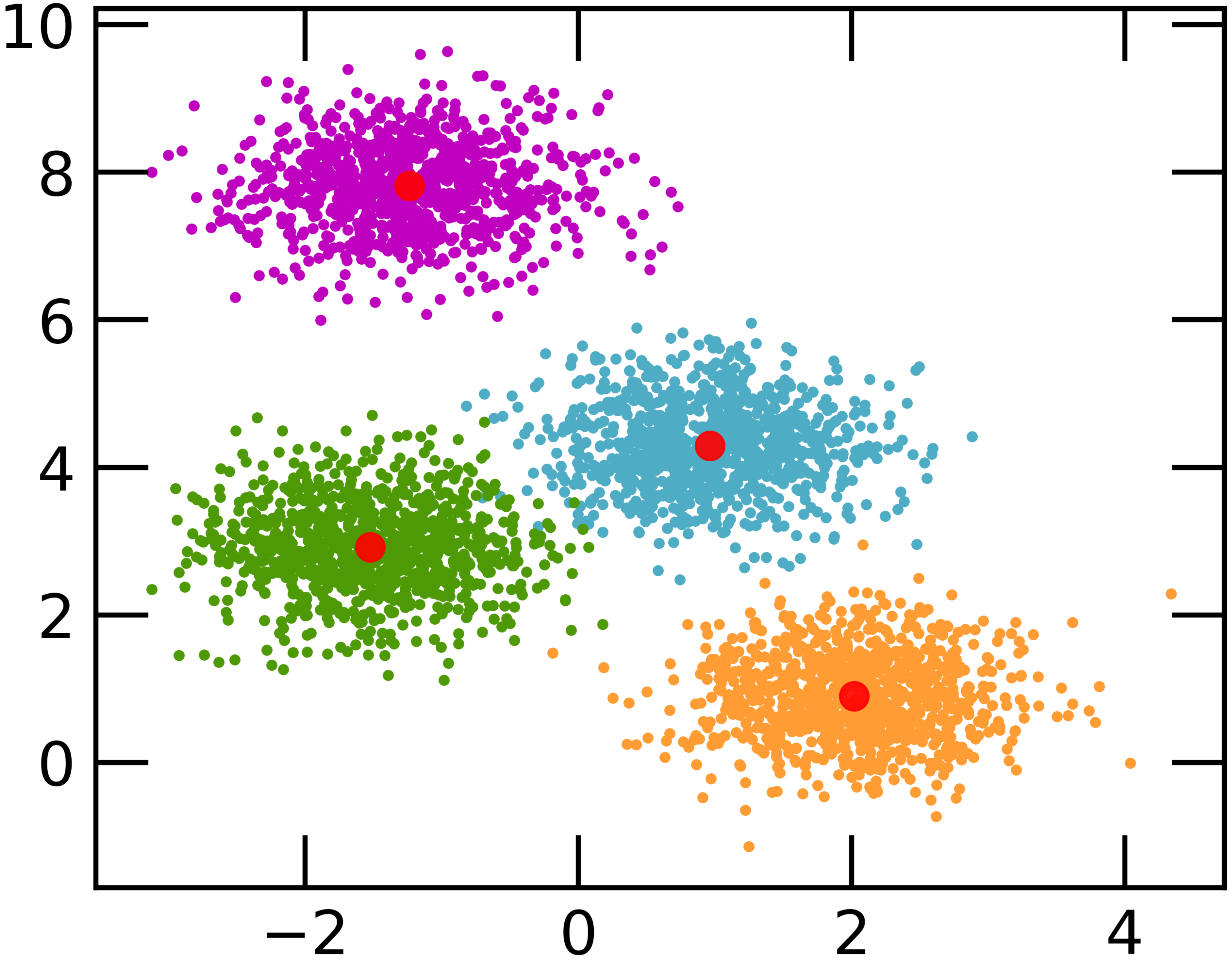
$$\sum_{i=0}^{n} \underbrace{min}_{\mu_j \in C}(||x_i - \mu_j||^2)$$

Notation: || || is norm of vector.

Approach:

1. Select random Kas first centers.
2. Distribute data points to cluster that have closest center.
3. If the data distribution is stabled. The method is done.
4. Update centers for each cluster by estimate the mean again of all the data points in the current clusters after distribution step 2.
5. Redo step 2.

- Example of clustering K-Means



Summary

- **Lecture 0**: Introduction to Jupyter notebook and Python.

In this class I review tools for data analysis and visualization: Python and Notebook.

- **Lecture 1**: Fundamental statistics quantities: Mean, median, standard deviation (variance), correlation.

In this class I review statistic quantities as mean, median, variance, and associated errors. The correlation between two random variables.

- **Lecture 2**: Probability distributions: Binomial, Uniform, Normal (or Gaussian), Poisson, Gamma, T-Student's, Chi-square, the central limit theorem.

In this class I review several special probability distributions and simple applications.

Summary

- **Lecture 3**: Hypothesis testing, model fitting and parameter estimation.

In this class we studied the use of the χ^2 statistic as a (1)hypothesis test of a model and (2) estimate the best fit parameter estimation.

- **Lecture 4**: Principal components analysis (PCA) and Bayesian methods.

In this class We focus on PCA dimensionality reduction to model variables and Bayesian statistic applications.

- **Lecture 5**: Basic Machine Learning using Scikit-learn tool.

In this class I briefly introduce Scikit-learn tool using for machine learning and data analysis.