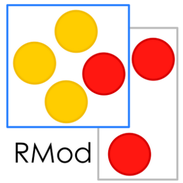
Machine Learning in Pharo

Sebastian JORDAN [1]
sebastian.jordan@inria.fr

Oleksandr ZAITSEV [2,1]
oleksandr.zaitsev@arolla.fr

[1] Inria, Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL

[2] Arolla, Paris



March 2022

Part 1:
ML Intro

- What is ML
- Why use ML
- Real life applications
- Types of ML problems

Part 2:
pharo-ai Library

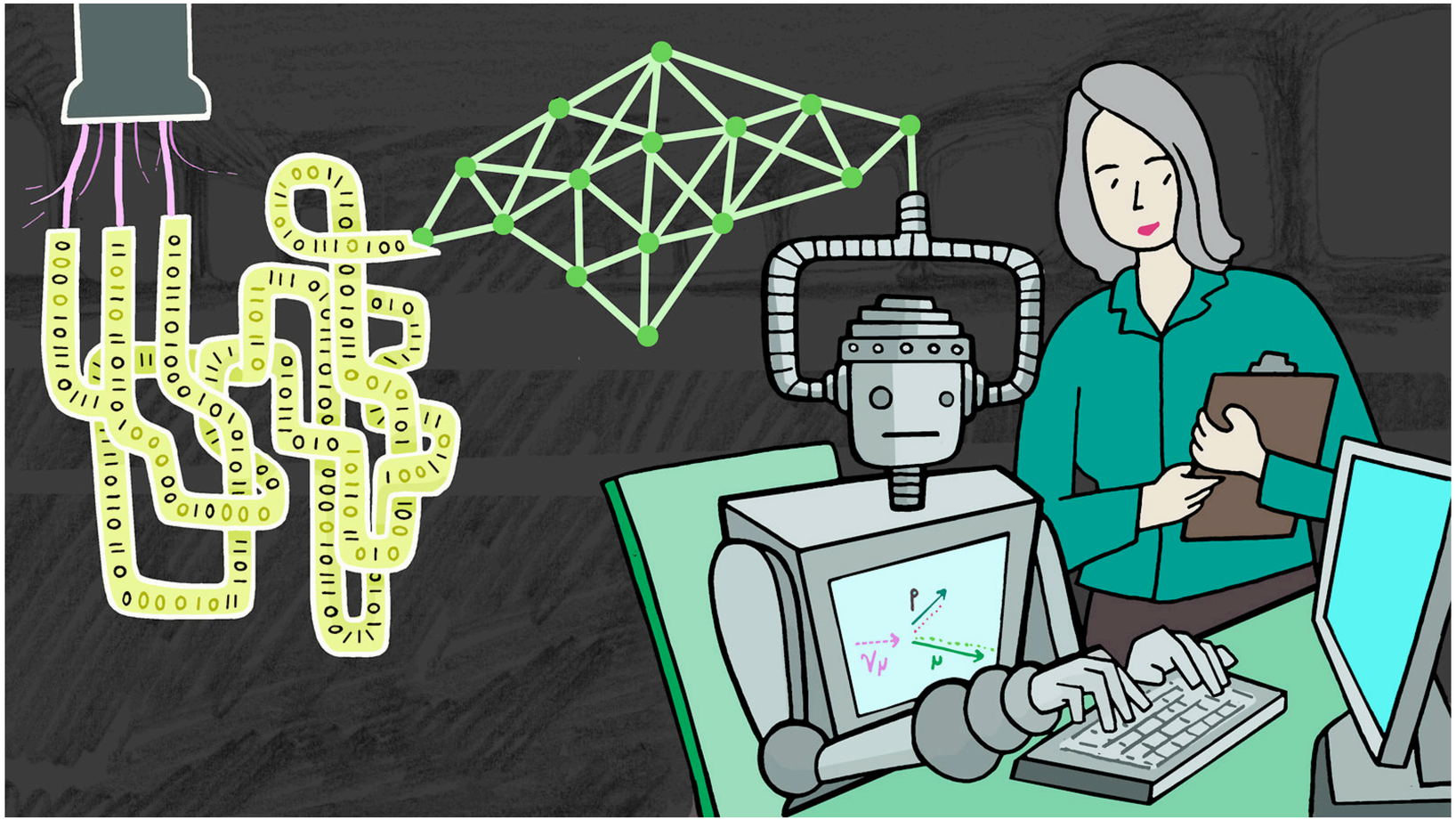
- Common ML algorithms
- What we have
- Future work
- Performance

Part 3:
Hands-On Tutorials

- Clustering simple example
- Clustering credit card users with K-means
- Predicting house pricing with linear regression

Part 1:
Machine Learning Introduction

What is Machine Learning



Source: symmetrymagazine.org

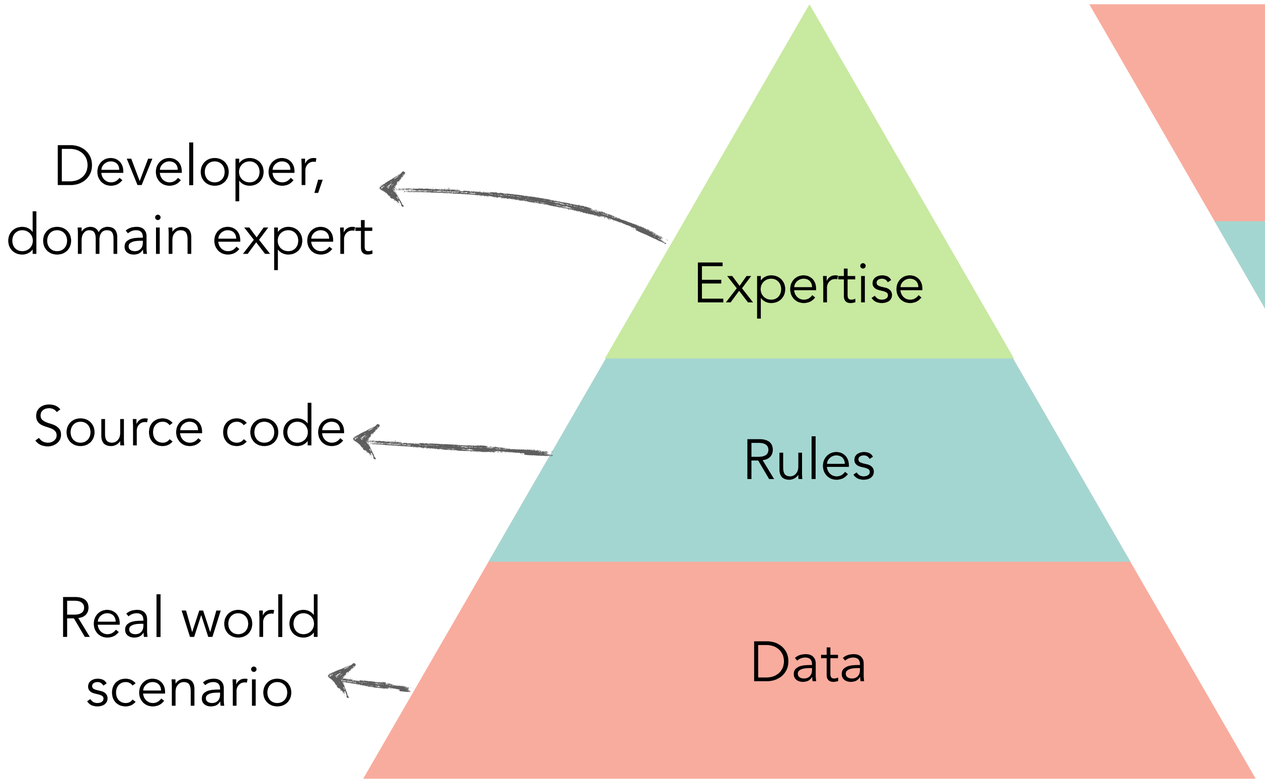
« Is the field of study that gives computers the ability to learn without being explicitly programmed. »

— Arthur Samuel 1959

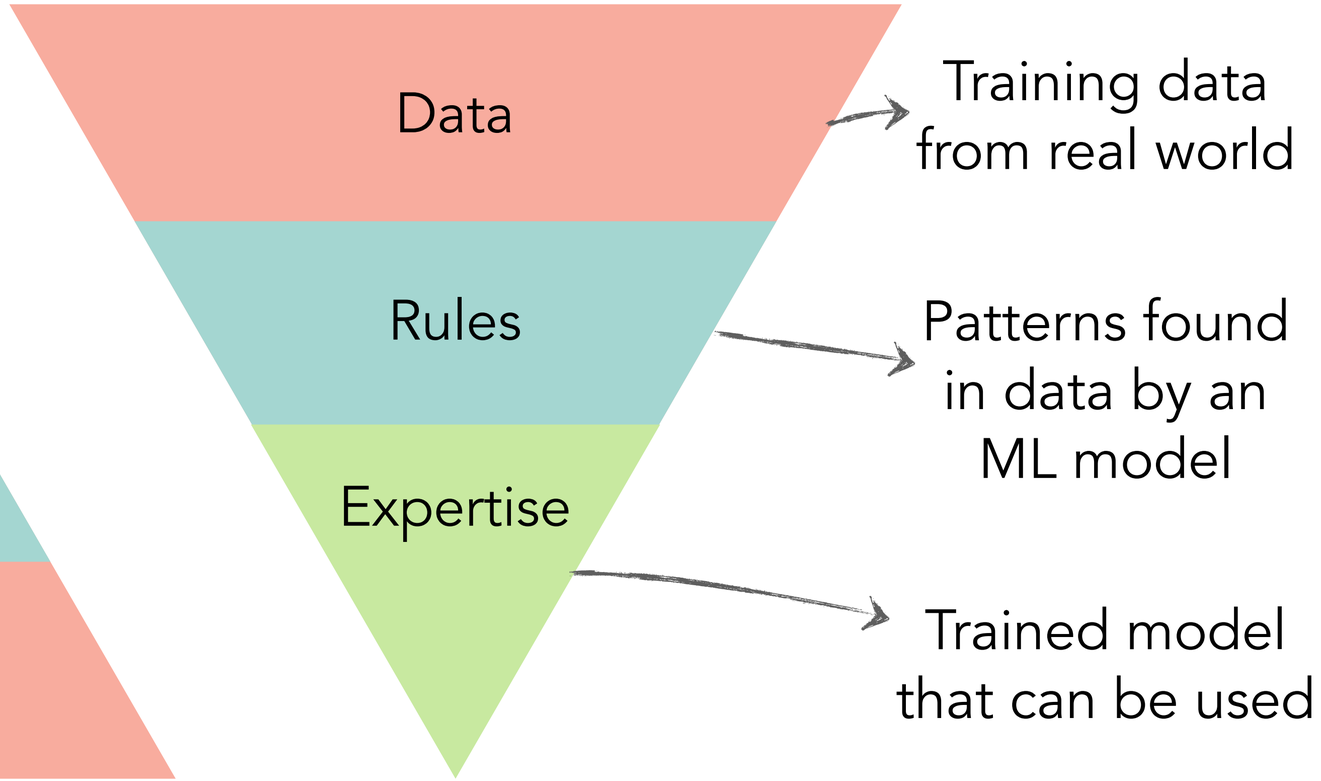
*« A computer program is said to learn from experience **E** with respect to some task **T** and some performance measure **P**, if its performance on **T**, as measured by **P**, improves with experience **E**. »*

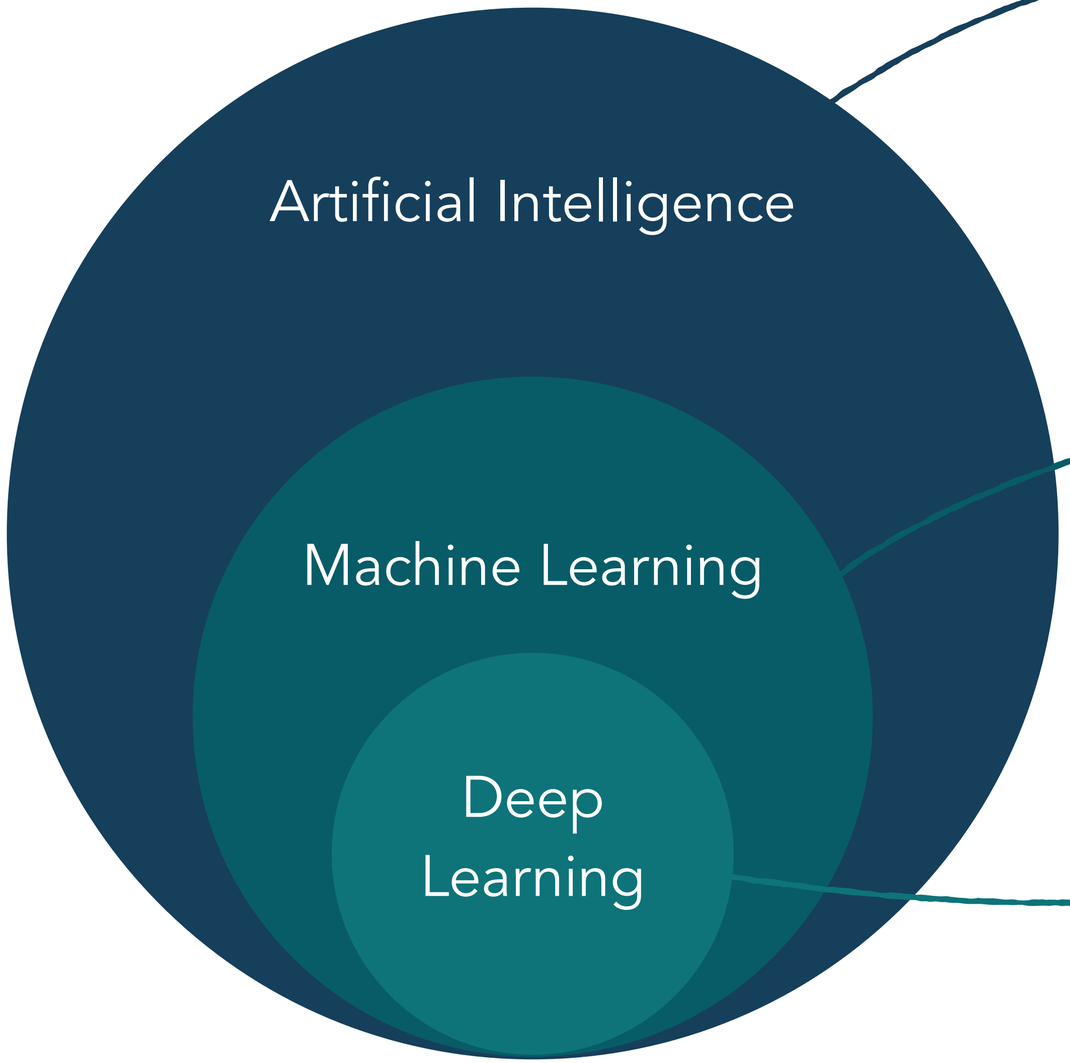
— Tom Mitchell 1997

Top-down



Bottom-up





Techniques which enable machines to mimic human behaviour

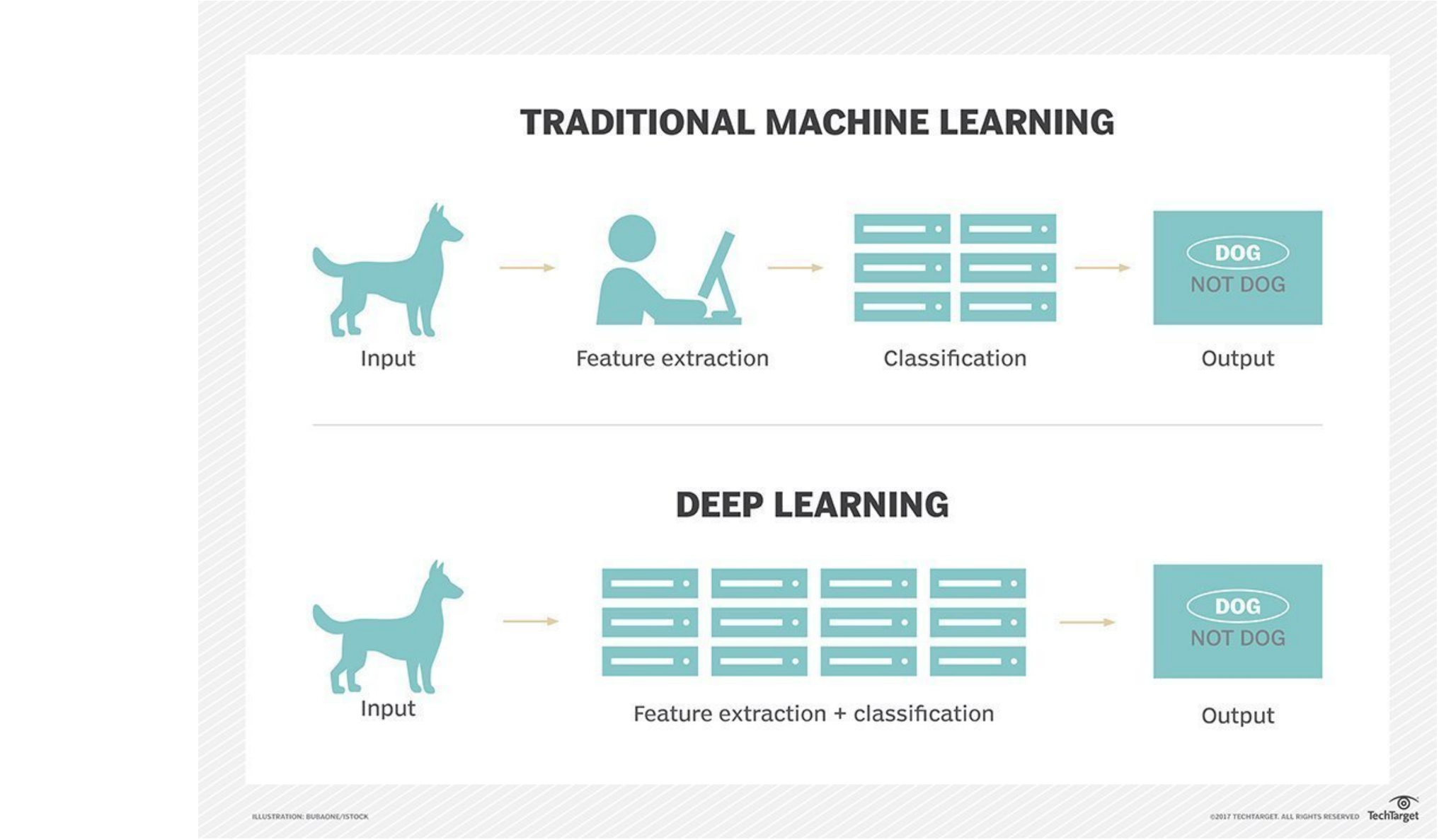
(A, knowledge based, …)*

Subset of AI techniques which use statistical methods to enable machines to improve with experience

(linear regression, k-means, naive bayes, …)

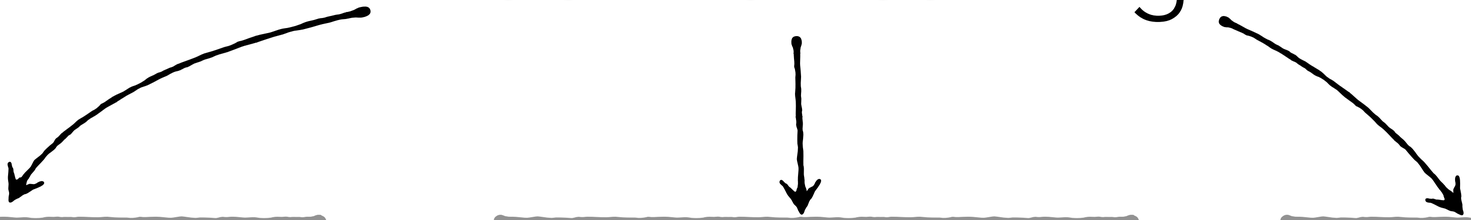
Subset of ML which uses deep neural networks

(convolutional networks, GAN, Seq2seq, …)



Source: TechTarget

Machine Learning



Supervised

- Labeled data
- First learn from examples, then apply to new data

(classification, regression,…)

Unsupervised

- No labeled data
- Extract patterns from the data

(clustering, anomaly detection,…)

Reinforcement

- Reward based learning
- Learn how to act in a certain environment
- Maximize reward

(game AI, self-driving cars, trading, …)

When to use machine learning (and when not to…)

Machine Learning is great for …

1. Problems for which existing solutions require **a lot of fine-tuning** or **long list of rules** *(e.g., face detection)*
2. **Complex problems** for which using a traditional approach yields no good solution *(e.g., playing chess)*
3. **Fluctuating environments**: machine learning can adapt to new data *(e.g., financial market)*
4. **Getting insights** about complex problems and large amounts of data *(e.g., unsupervised learning)*

5 Key Limitations of Machine Learning

1. **Ethics**: we trust data and algorithms more than personal insights
2. **Data**: require good amount of training data (often labeled data)
3. **Interpretability**: many machine learning algorithms produce results that can not be easily explained
4. **Nondeterminism**: based on randomness, contain noise, not well suited for tasks that require precision
5. **Reproducibility**: hard to reproduce and test

Some Examples of Applications

- Analysing images to classify them
- Detecting tumors in brain scans
- Automatically classifying news articles
- Flagging offensive comments
- Summarising long documents
- Chatbots and personal assistants
- Forecastings
- Voice comprehension
- Detecting credit card fraud
- Segmenting clients based on purchases
- Personalised recommendations
- Game AI

Part 2:
pharo-ai library

We introduce pharo-ai v0.8

a modular library for shallow machine learning in Pharo

 github.com/pharo-ai

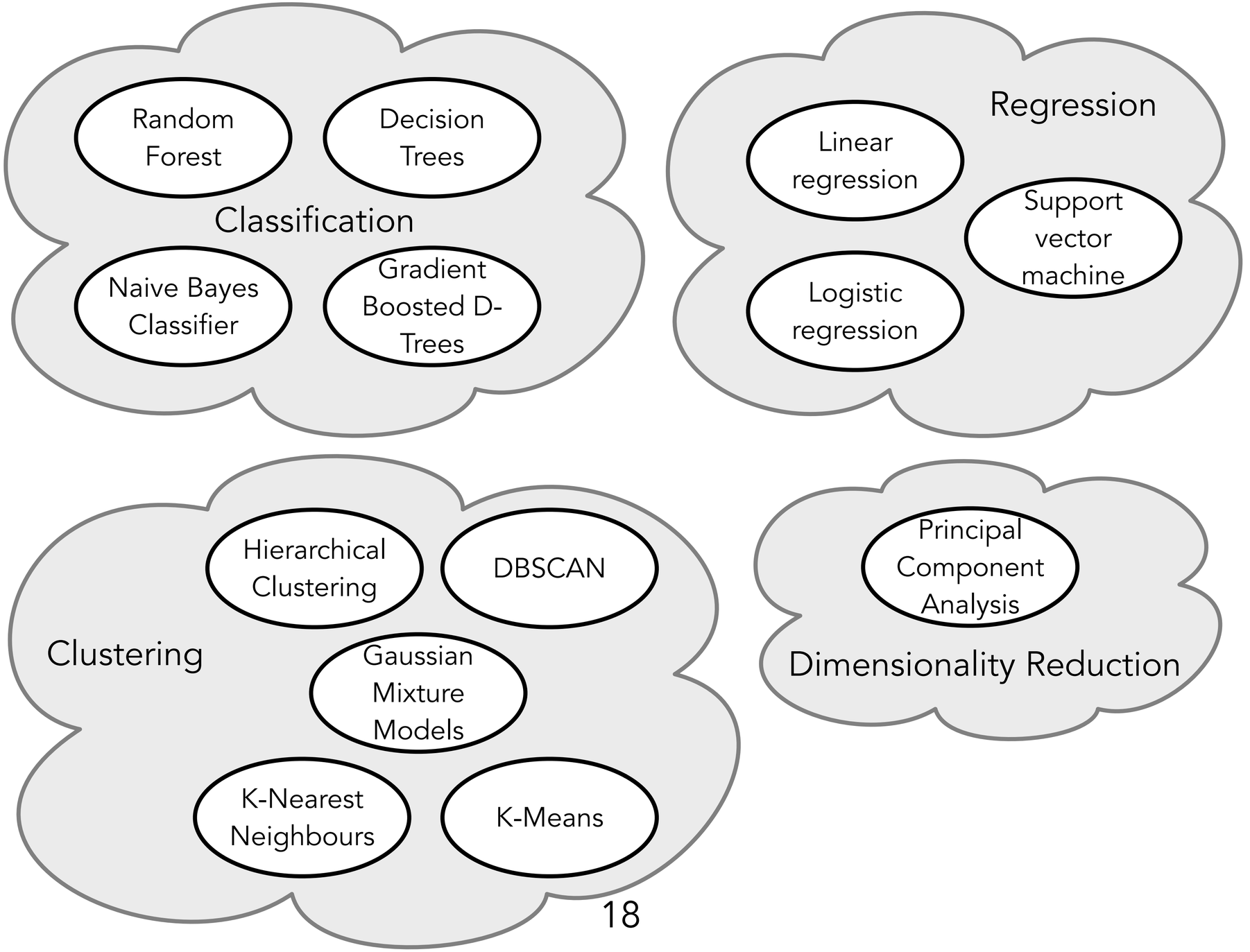
Why do we need a ML library in Pharo

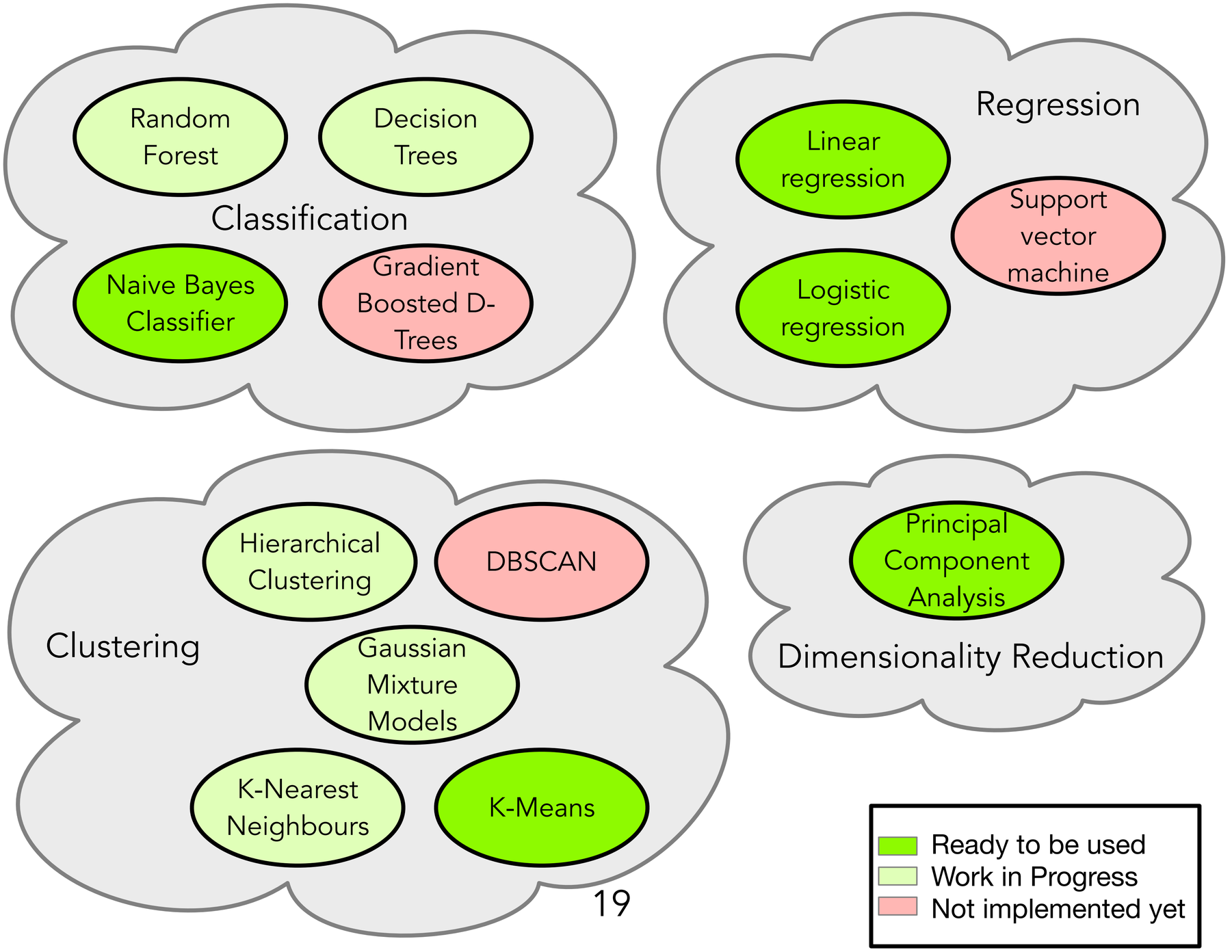
- We want to provide tools for the Pharo community people interested in doing ML and AI.

- We would like to contribute to the work that is currently being developed by different people (Univ. Chile, Object Profile-Chile, PolyMathOrg, Semantics-Bolivia, CIRAD-France).

How do we position ourselves

	Python	**R**	**Pharo**
Data Analysis & Manipulation	pandas	data.frame, dplyr	DataFrame
Algebra & Statistics	numpy, scipy	MASS, SparseM	PolyMath
Shallow Learning	scikit-learn	caret, ml3	pharo-ai
Deep Learning	TensorFlow, Keras	TensorFlow, Keras	TensorFlow, Keras
Visualisation	matplotlib	ggplot	Roassal


Random Forest
Decision Trees
Classification
Naive Bayes Classifier
Gradient Boosted D-Trees
Regression
Linear regression
Support vector machine
Logistic regression
Hierarchical Clustering
DBSCAN
Clustering
Gaussian Mixture Models
K-Nearest Neighbours
K-Means
Principal Component Analysis
Dimensionality Reduction


Random Forest
Decision Trees
Classification
Naive Bayes Classifier
Gradient Boosted D-Trees
Regression
Linear regression
Support vector machine
Logistic regression
Hierarchical Clustering
DBSCAN
Clustering
Gaussian Mixture Models
K-Nearest Neighbours
K-Means
Principal Component Analysis
Dimensionality Reduction
Ready to be used
Work in Progress
Not implemented yet

Also in pharo-ai ecosystem

Graph algorithms

A-Priori

NLP

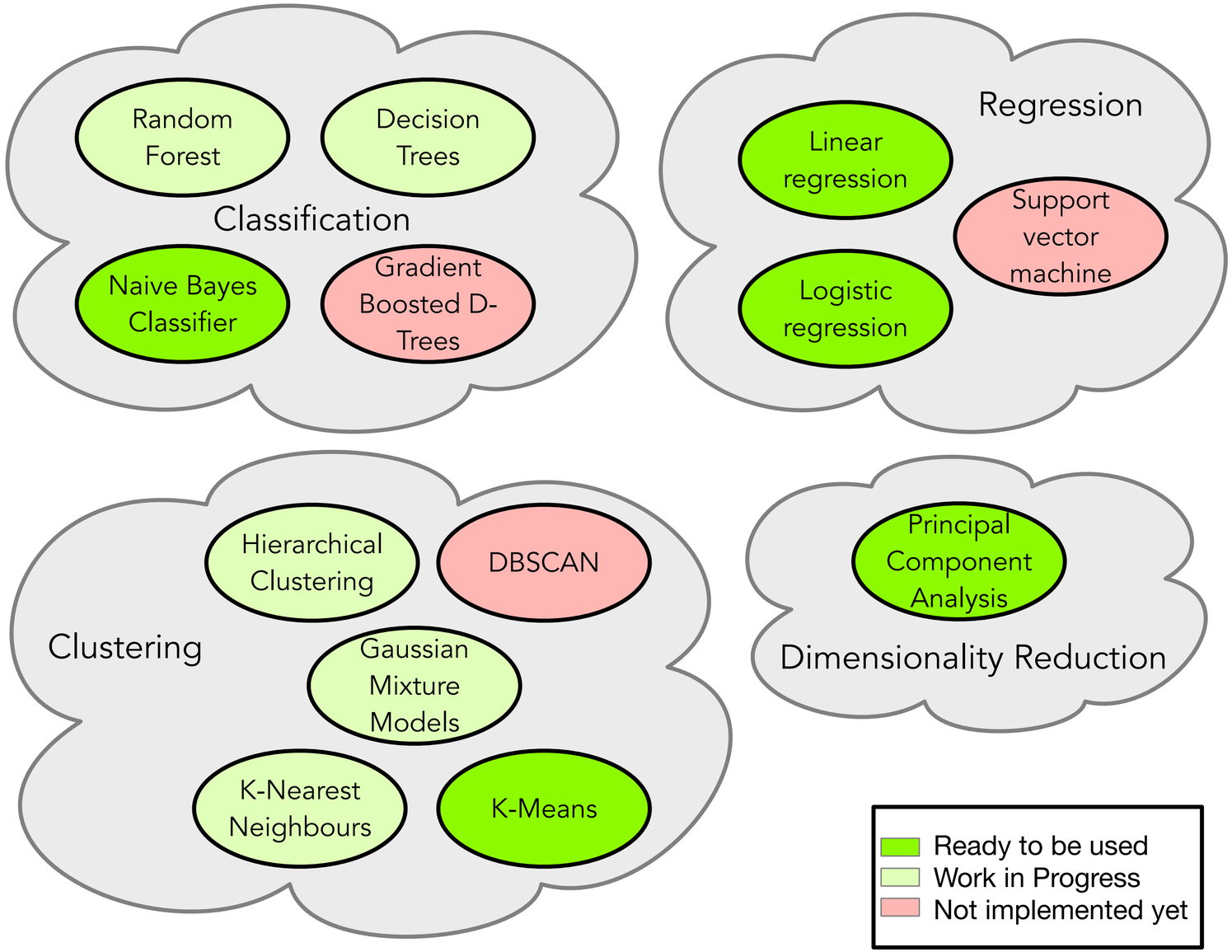
DataFrame

Metrics

Datasets

Normalisation

PolyMath



Roadmap

- Finish the Work In Progress algorithms

  work in progress -> ready to be used

- Implement the missing algorithms

  not implemented -> ready to be used

- Performance benchmarking against scikit-learn

  Pharo -> Pharo + LAPACK

- Standardise the API for all the algorithms

- Documentation

  Wiki, Book, Website

Visit Us !
Play, Use, and Contribute

Start here

Pharo-ai Wiki: https://github.com/pharo-ai/wiki

Visit Us !
Play, Use, and Contribute

Other ML projects in Pharo: https://github.com/pharo-ai/awesome-pharo-ml

Part 3:
Hands-On Tutorials

Hands-On

- Clustering simple example
- Clustering credit card users with K-means
- Predicting house price with linear regression