A BRIEF INTRODUCTION TO BIG DATA AND MACHINE LEARNING

Dr. Lablanche Pierre-Yves
African Institute for Mathematical Sciences

HOW **BIG** IS BIG DATA?

"… It's not about the size …" - *unknown author*

VOLUME **VELOCITY**

VARIETY **VERACITY**

The 4 V's

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015

4.4 MILLION IT JOBS

will be created globally to support big data, with 1.9 million in the United States

Volume

SCALE OF DATA

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005

It's estimated that

2.5 QUINTILLION BYTES

[2.3 TRILLION GIGABYTES]

of data are created each day

6 BILLION PEOPLE

have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least

100 TERABYTES

[100,000 GIGABYTES]

of data stored

Variety

DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES

[161 BILLION GIGABYTES]

30 BILLION PIECES OF CONTENT

are shared on Facebook every month

By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

4 BILLION+ HOURS OF VIDEO

are watched on YouTube each month

400 MILLION TWEETS

are sent per day by about 200 million monthly active users

Velocity

ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

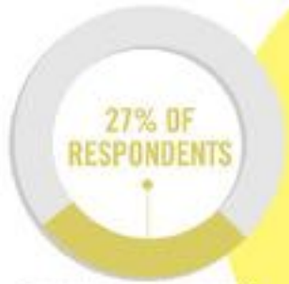
– almost 2.5 connections per person on earth

Veracity

UNCERTAINTY OF DATA

1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate

Poor data quality costs the US economy around

$3.1 TRILLION A YEAR

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

IBM

BUT WHAT IS "*DATA*"?

"Data" is information … can be pretty much everything.

(Note *data* is the plural of *datum*)

For this course a ***data set*** is a collection of ***samples*** with their ***features***

DATA SET - SAMPLE - FEATURE

A table is an easy way to visualise a data set :

DATA SET	Feature 1	Feature 2	…	Feature M
Sample 1	value [1,1]	value [1,2]	…	value [1,M]
Sample 2	value [2,1]	value [2,2]	…	value [2,M]
…	…	…	…	…
Sample N	value [N,1]	value [N,2]	…	value [N,M]

IMPORTANT : a "value" is not necessarily a number

DATA SET - SAMPLE - FEATURE

Example : The forever famous "iris data set"

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	name
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa

THE BIG DATA CHALLENGE

- What can we do with these Big Data ?
- What can we learn from such Big Data ?
- How to deal with Big Data ?

MACHINE LEARNING

Machine Learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Such algorithms operate by building a model based on inputs and using that to make predictions decisions, rather than following only explicitly programmed instructions.

http://en.wikipedia.org/wiki/Machine_Learning

WHAT MACHINE LEARNING CAN DO :

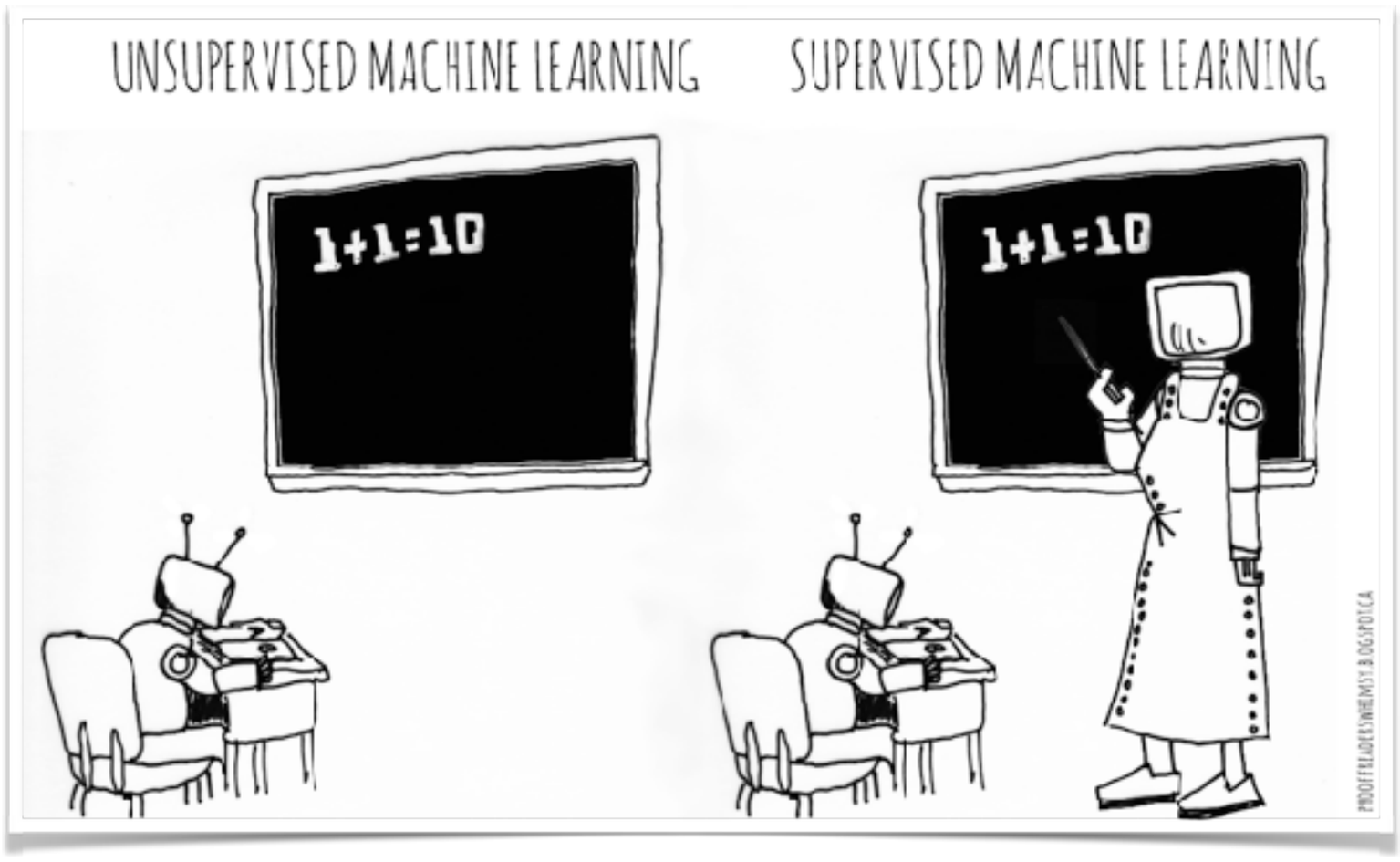
- Classification / Regression
- Pattern Recognition
- Outliers Detection
- Clustering
- Dimensionality Reduction
- Knowledge Extraction
- etc.

Machine Learning is not a blackbox that will solve your problem it is a **toolbox**

HOW MACHINE LEARN ?

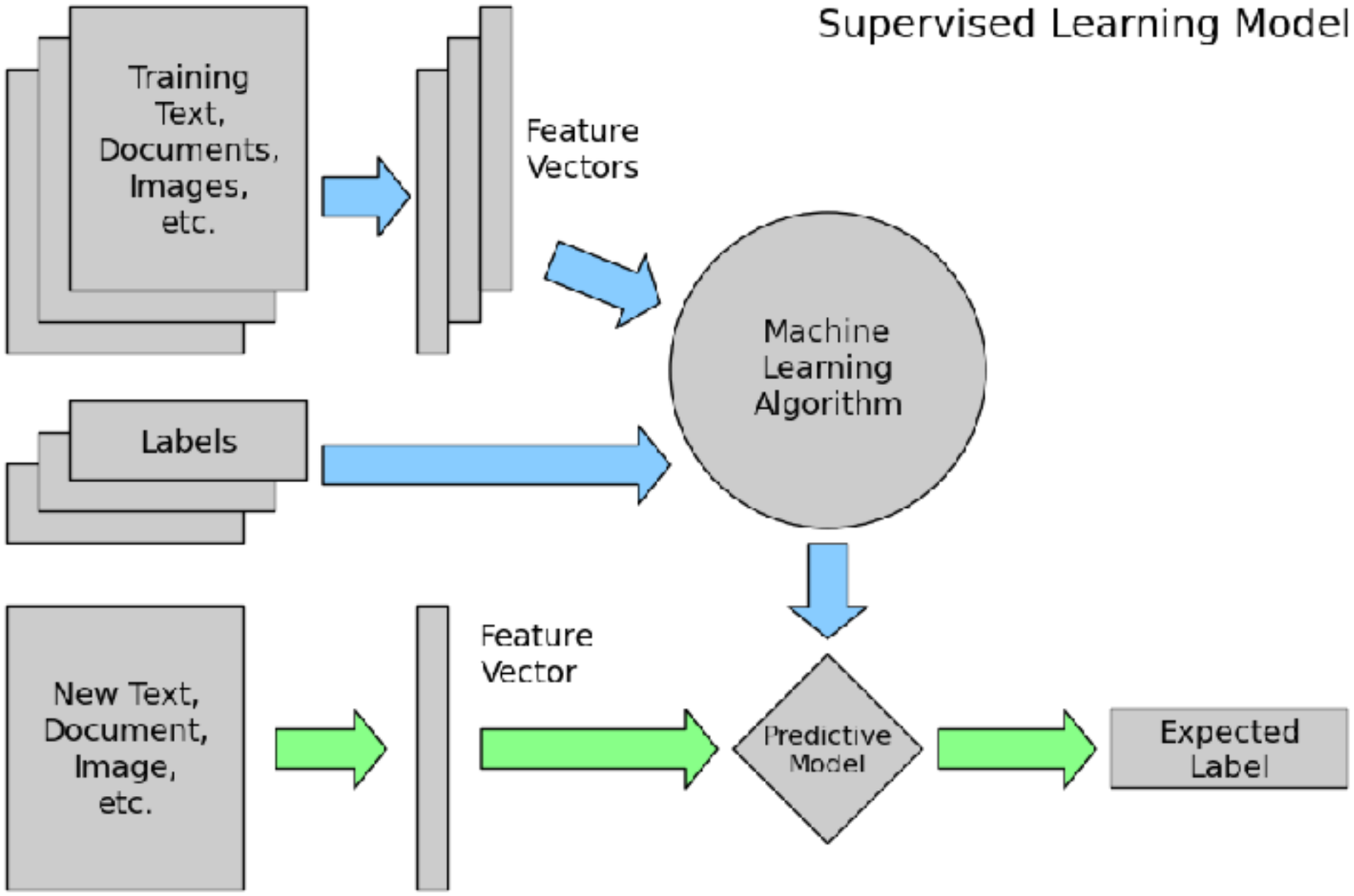
Different types of learning algorithms :

- **Supervised**
- **Unsupervised**
- Semi-supervised
- Reinforcement



SUPERVISED LEARNING

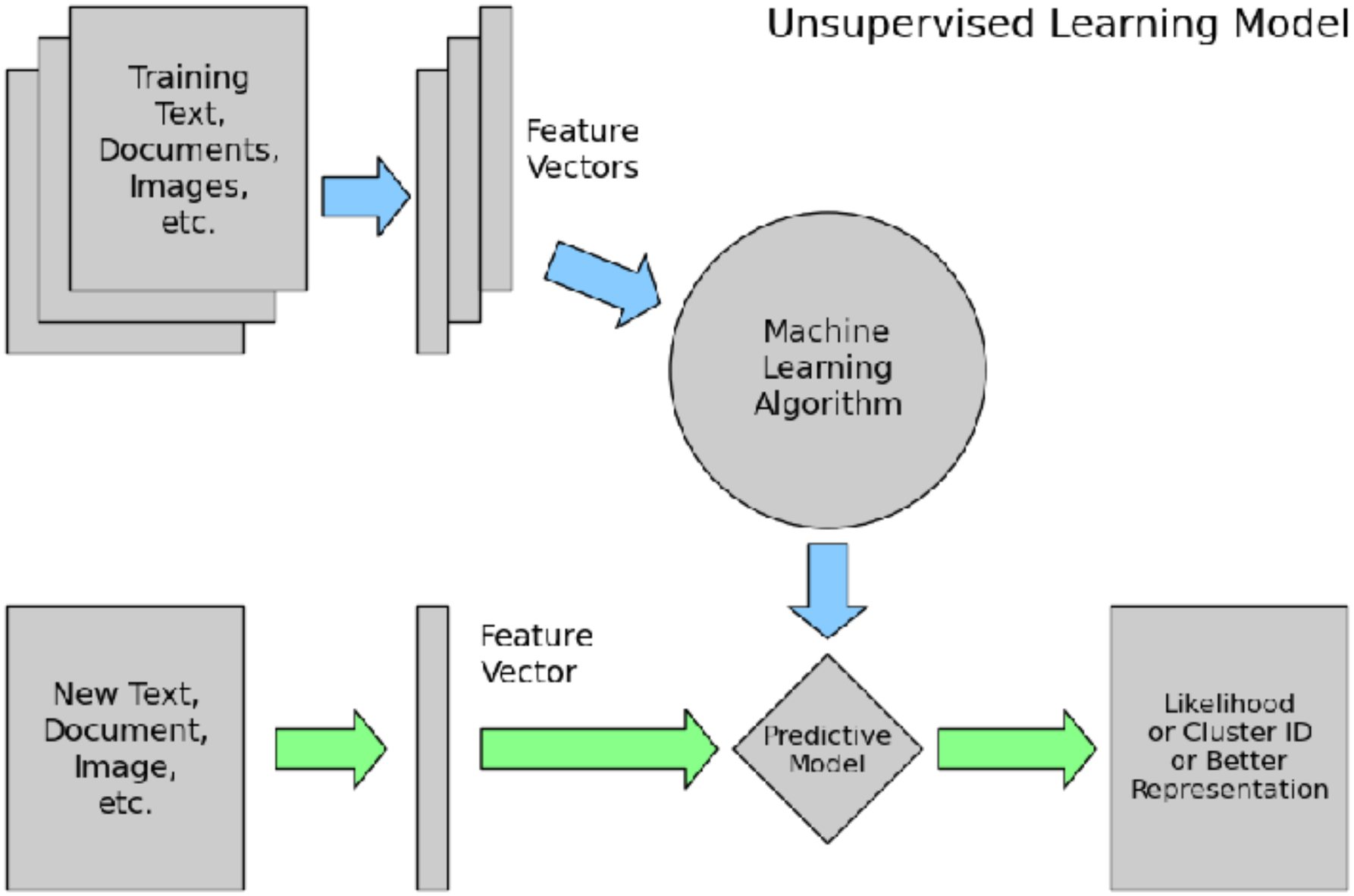
The machine learns from known *labelled* data



Source : http://www.astroml.org/sklearn_tutorial/general_concepts.html

UNSUPERVISED LEARNING

The machine learns from unknown data



Source : http://www.astroml.org/sklearn_tutorial/general_concepts.html

OTHER LEARNING APPROACH

- Semi-supervised learning:
  Make use of both labelled and unlabelled data

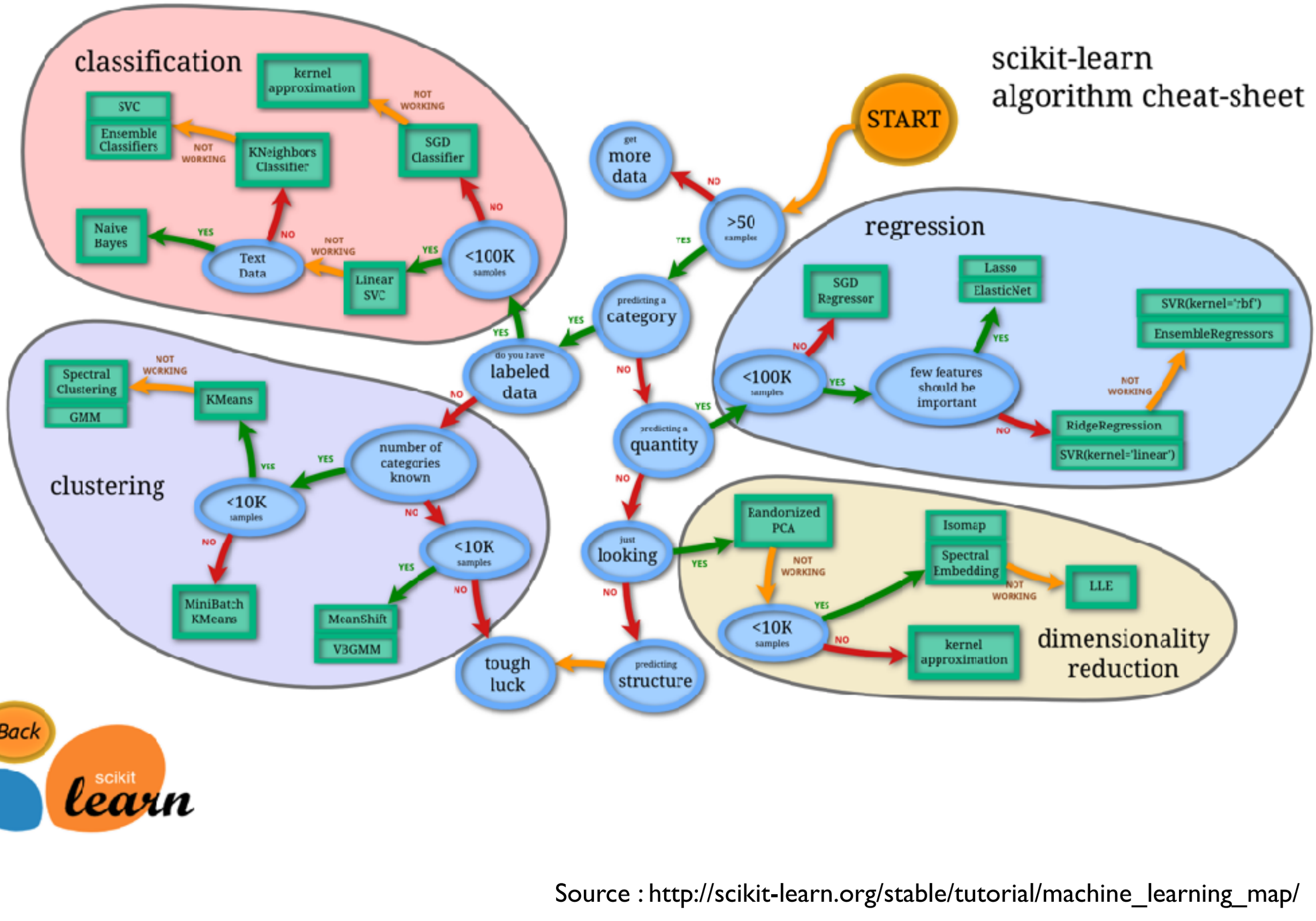
- Reinforcement learning :
  The machine learns to react to an environment

PYTHON SCIKIT LIBRARY

Standard Python Library for machine learning :

```
> from sklearn import my_algo
> algo = my_algo(parameters)
> algo.fit(training_data)
> algo.predict(new_value)
```

scikit-learn algorithm cheat-sheet



Source : http://scikit-learn.org/stable/tutorial/machine_learning_map/