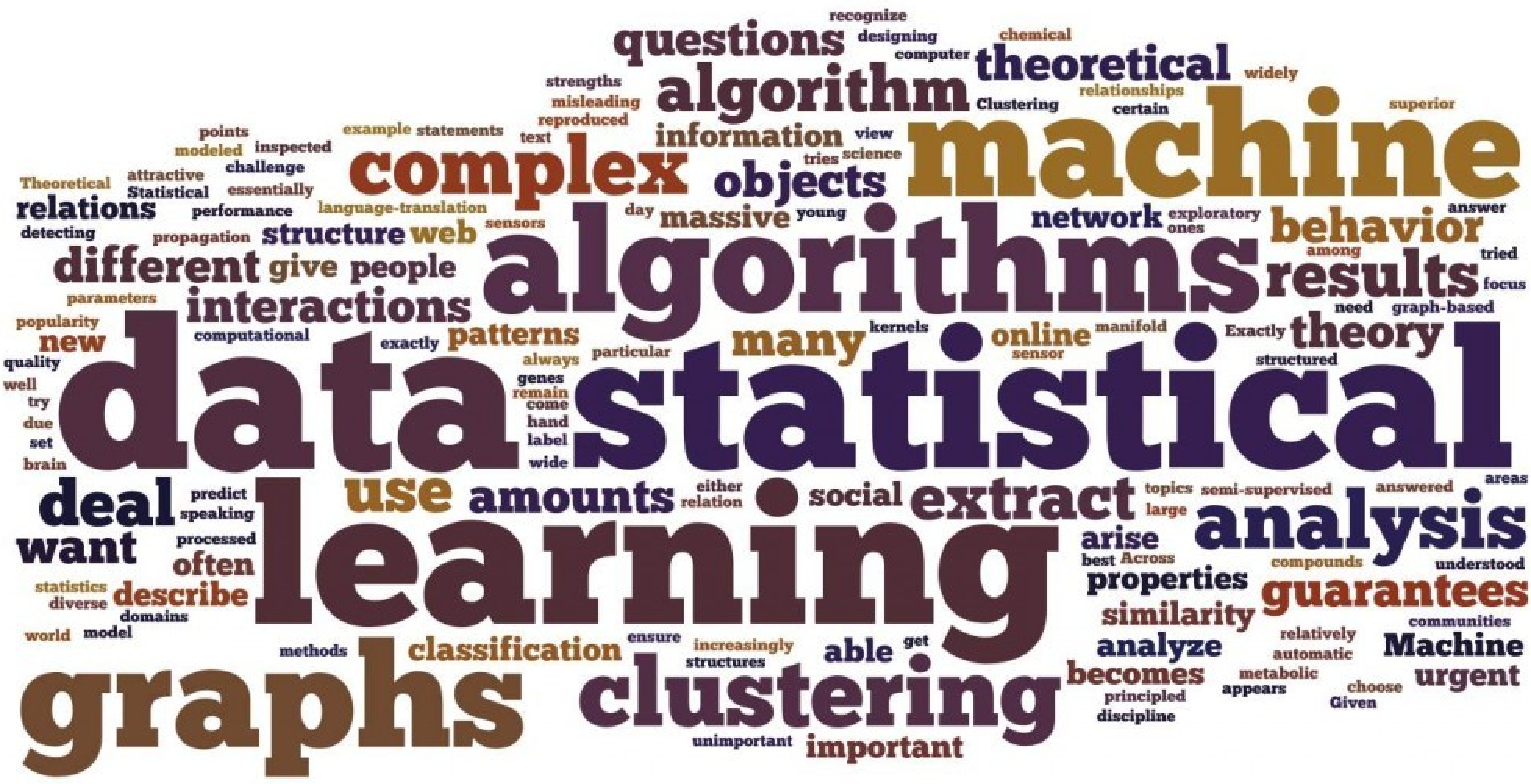
Learning from the Past: with Scikit-Learn

ANOOP THOMAS MATHEW
Profoundis Labs Pvt. Ltd.


recognize
designing
chemical
questions
computer
theoretical
widely
strengths
algorithm
relationships
misleading
Clustering
certain
superior
points
example
statements
reproduced
text
information
view
modeled
inspected
tries
science
machine
Theoretical
attractive
challenge
complex
Statistical
essentially
objects
relations
performance
language-translation
day
massive
young
network
exploratory
answer
detecting
propagation
structure
web
sensors
ones
behavior
different
give
people
algorithms
among
tried
results
focus
parameters
interactions
need
graph-based
popularity
kernels
manifold
Exactly
theory
new
computational
exactly
patterns
many
online
quality
always
particular
sensor
structured
well
genes
remain
try
come
data
statistical
due
hand
set
label
brain
wide
deal
predict
either
topics
semi-supervised
answered
areas
use
amounts
relation
social
extract
speaking
large
analysis
processed
arise
want
best
Across
often
compounds
understood
statistics
properties
learning
describe
guarantees
diverse
domains
similarity
communities
world
model
ensure
relatively
methods
increasingly
get
analyze
Machine
classification
able
structures
automatic
becomes
metabolic
urgent
graphs
clustering
principled
appears
choose
Given
discipline
unimportant
important

Profoundis

Agenda

- Basics of Machine Learning
- Introduction some common techniques
- Let you know scikit-learn exists
- Some inspiration on using machine learning in daily life scenarios and live projects.

Profoundis

How to draw a snake?

and ...

Profoundis

How to draw a snake?

This is WEIRD!

and ...

Profoundis

Introduction

A lot of Data!

What to do???

Introduction

What is

Machine Learning

(Data Mining)?

(in plain english)

Machine Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E"

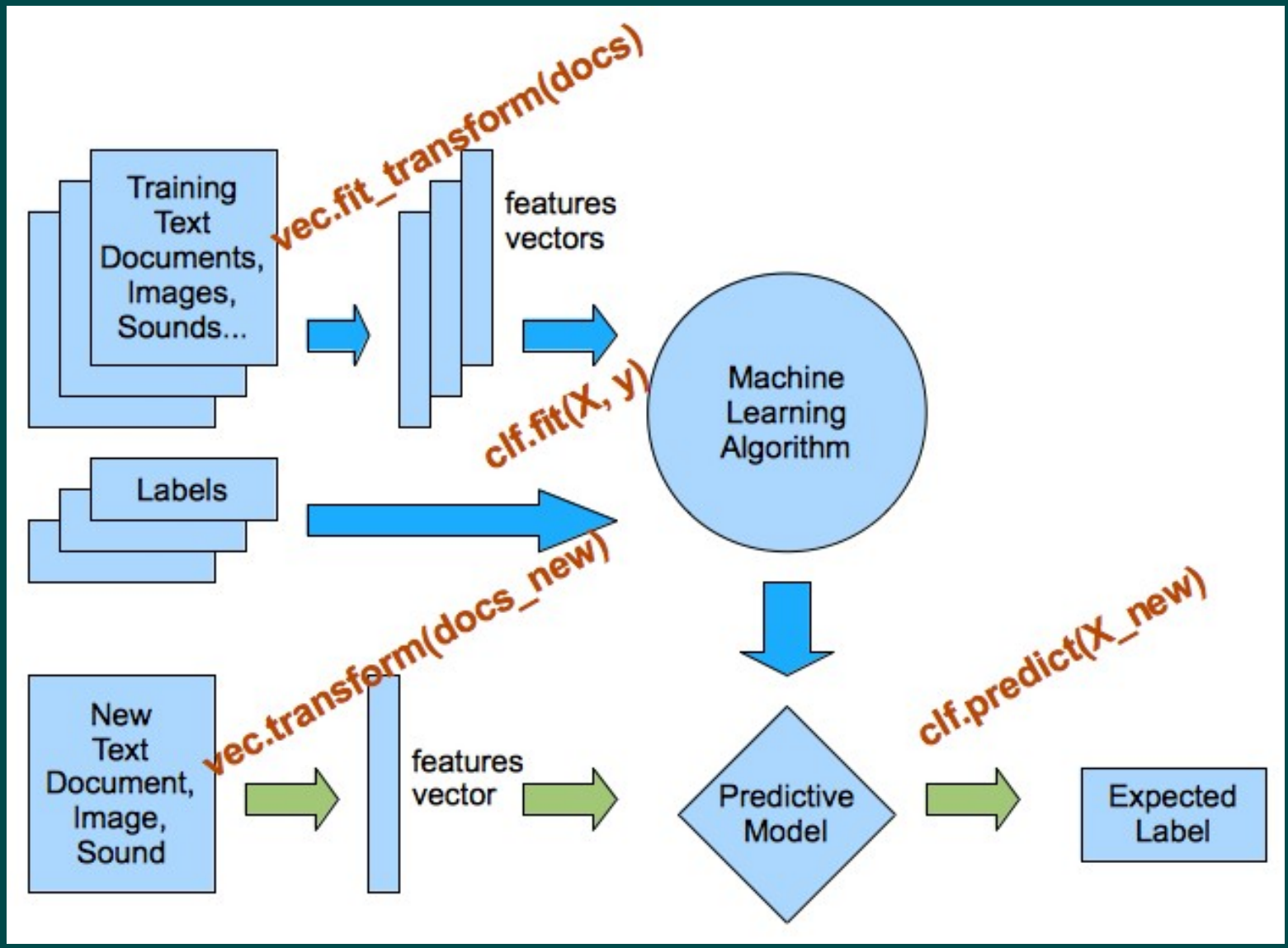
Tom M. Mitchell

Machine Learning

- Supervised Learning - model.fit(X, y)
- Unsupervised Learning - model.fit(X)

Profoundis

Supervised Learning



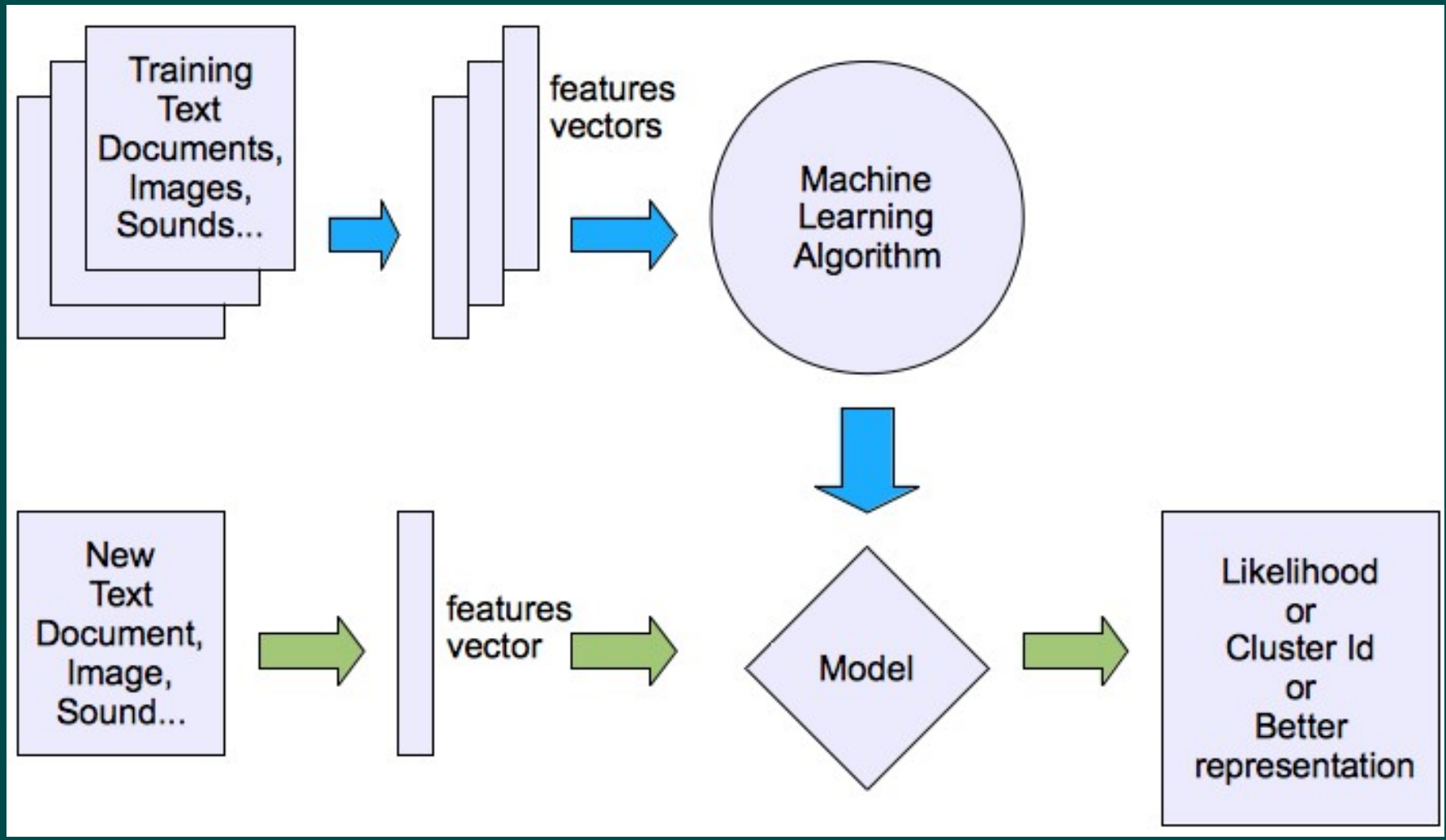
Profoundis

For example ...

```
from sklearn.linear_model import Ridge as RidgeRegression
from sklearn import datasets
from matplotlib import pyplot as plt

boston = datasets.load_boston()
X = boston.data
y = boston.target
clf = RidgeRegression()
clf.fit(X, y)
clf.predict(X)
```

Unsupervised Learning



Profoundis

For example ...

```
from sklearn.cluster import KMeans
from numpy.random import RandomState
rng = RandomState(42)
k_means = KMeans(3, random_state=rng)
k_means.fit(X)
```

What can Scikit-learn do?

Clustering
Classification
Regression

Profoundis

Terminology

- **Model** the collection of parameters you are trying to fit
- **Data** what you are using to fit the model
- **Target** the value you are trying to predict with your model
- **Features** attributes of your data that will be used in prediction
- **Methods** algorithms that will use your data to fit a model

Steps for Analysis

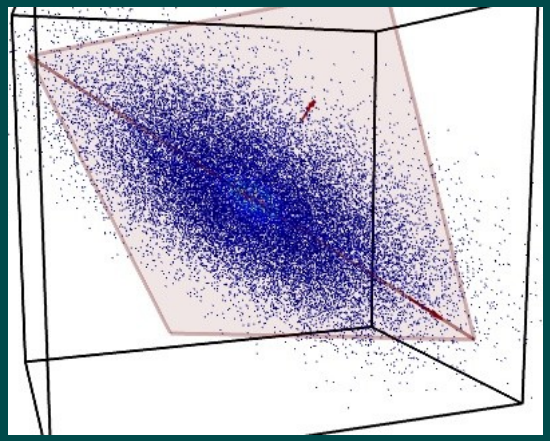
- Understand the task. See how to measure the performance.
- Choose the source of training experience.
- Decide what will be input and output.
- Choose a set of models to the output function.
- Choose a learning algorithm.

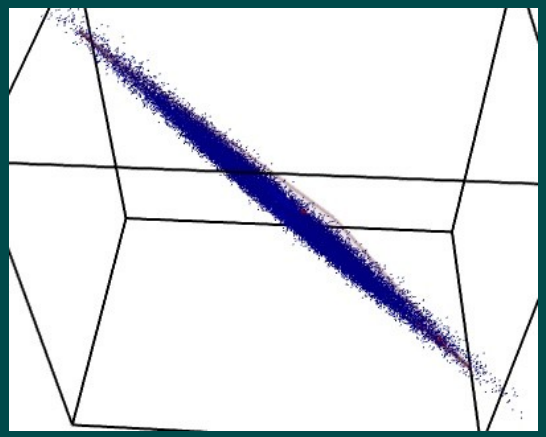
Steps for Analysis

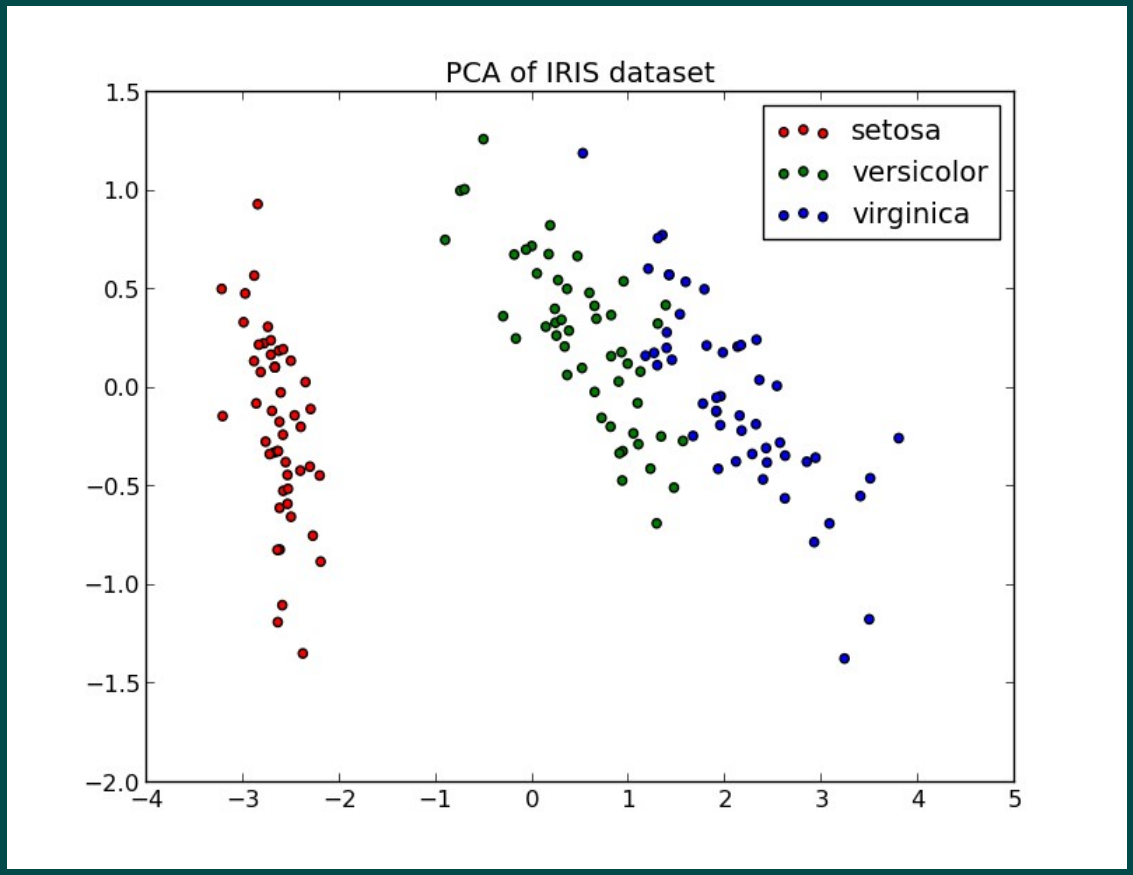
- Understand the task. See how to measure the performance. Find the right question to ask.
- Choose the source of training experience.
  - Keep training and testing dataset separate. Beware of *overfitting* !
- Decide what will be input and expected output.
- Choose a set of models to approximate the output function. (use dimensinality reduction)
- Choose a learning algorithm. Try different ones ;)

Some Common Algorithms

Principal Component Analysis



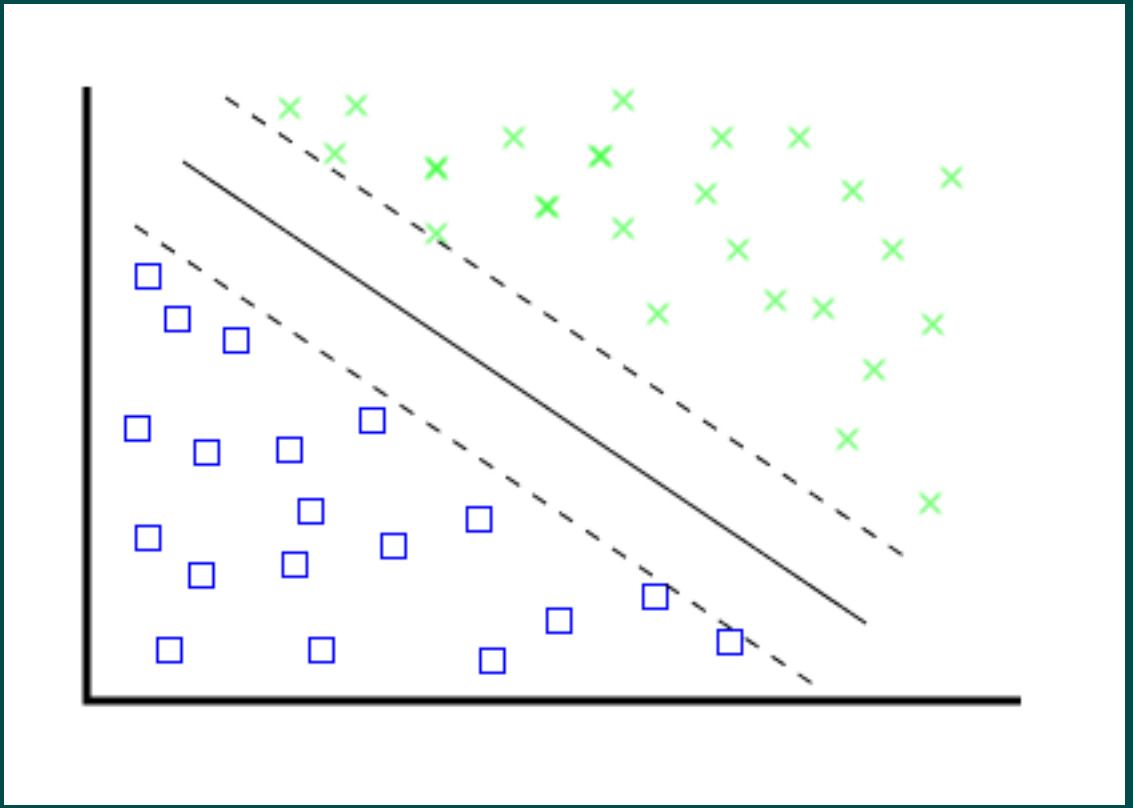




Profoundis

Some Common Algorithms

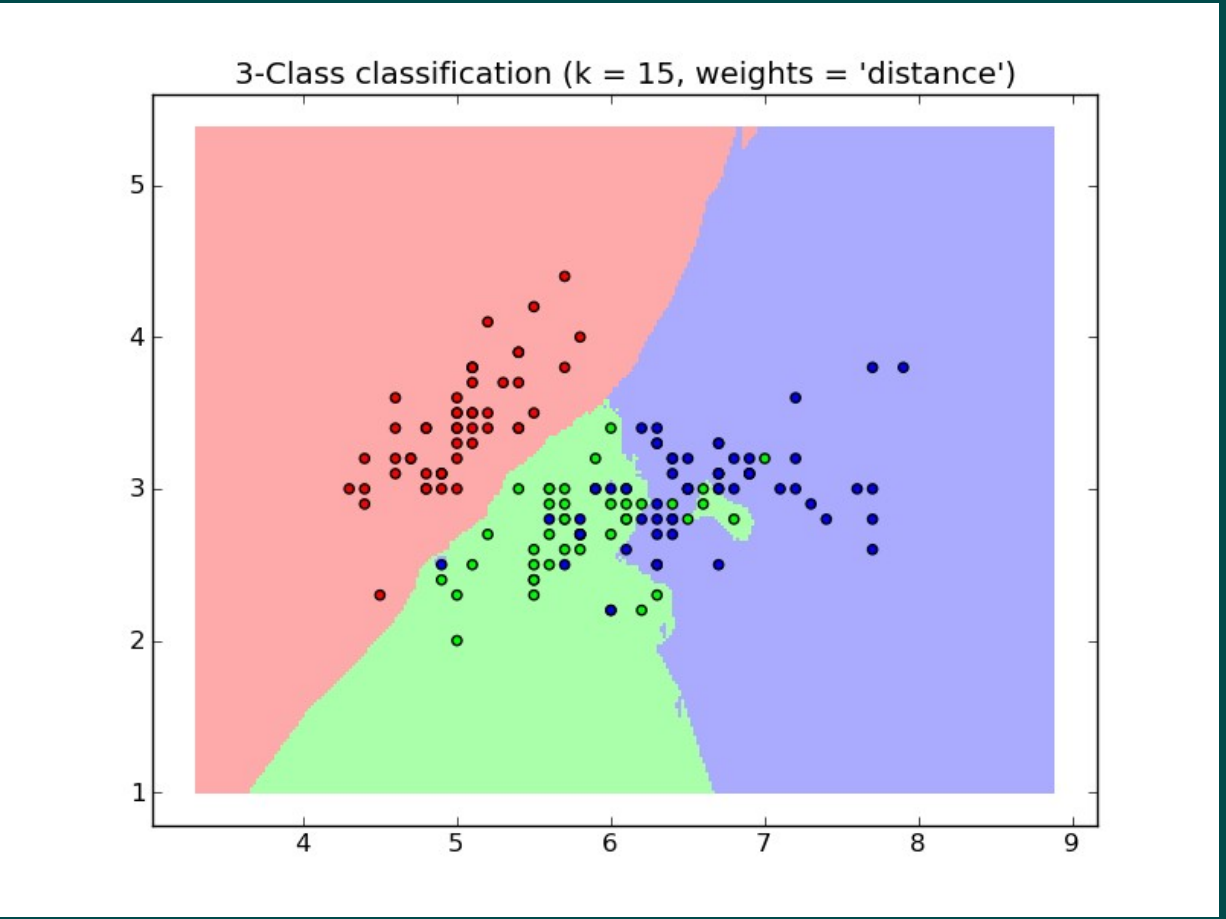
- Support Vector Machine



Profoundis

Some Common Algorithms

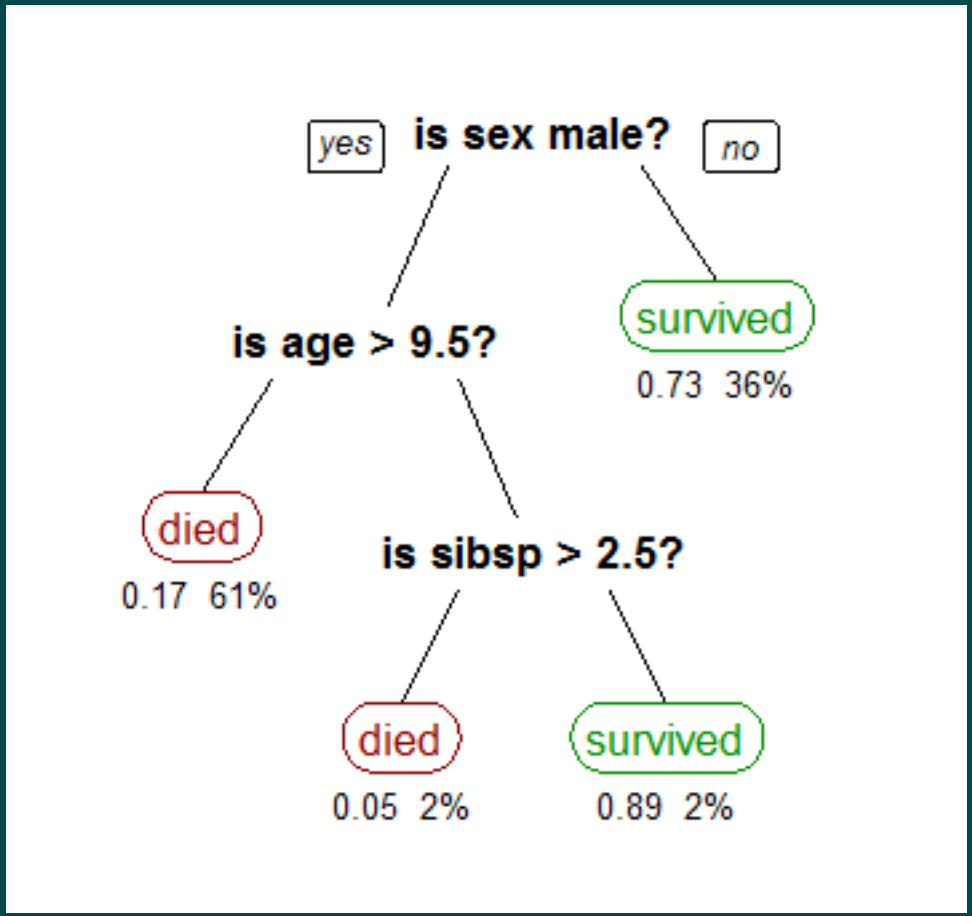
- Nearest Neighbour Classifier



Profoundis

Some Common Algorithms

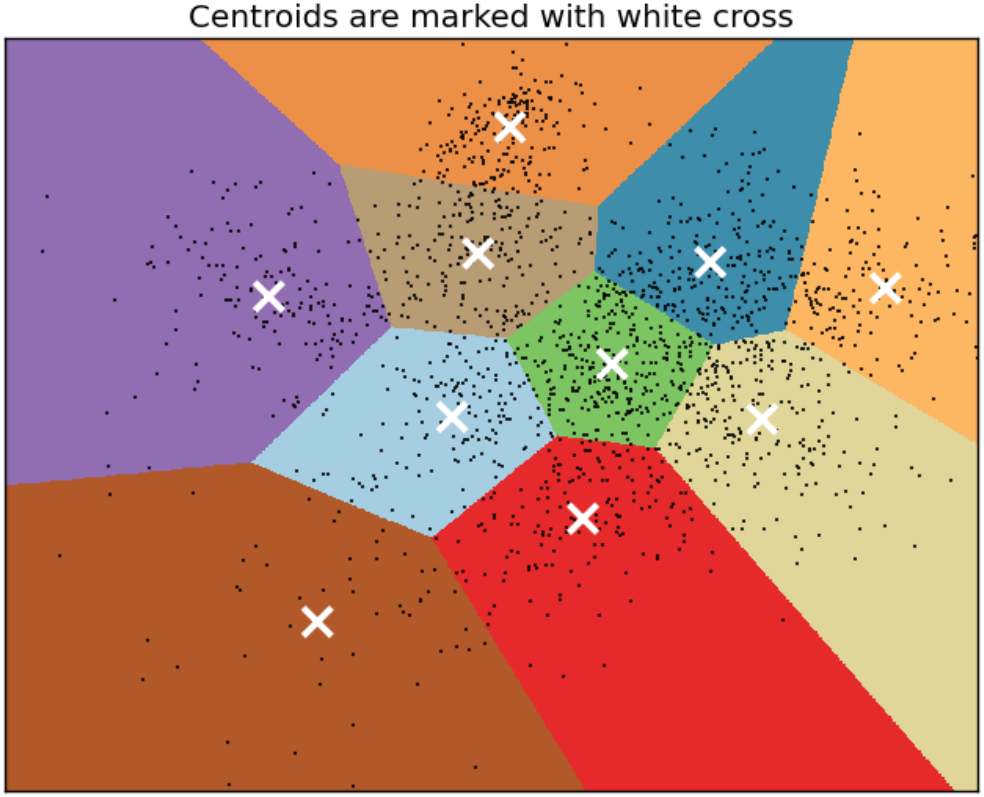
- Decision Tree Learning



Profoundis

Some Common Algorithms

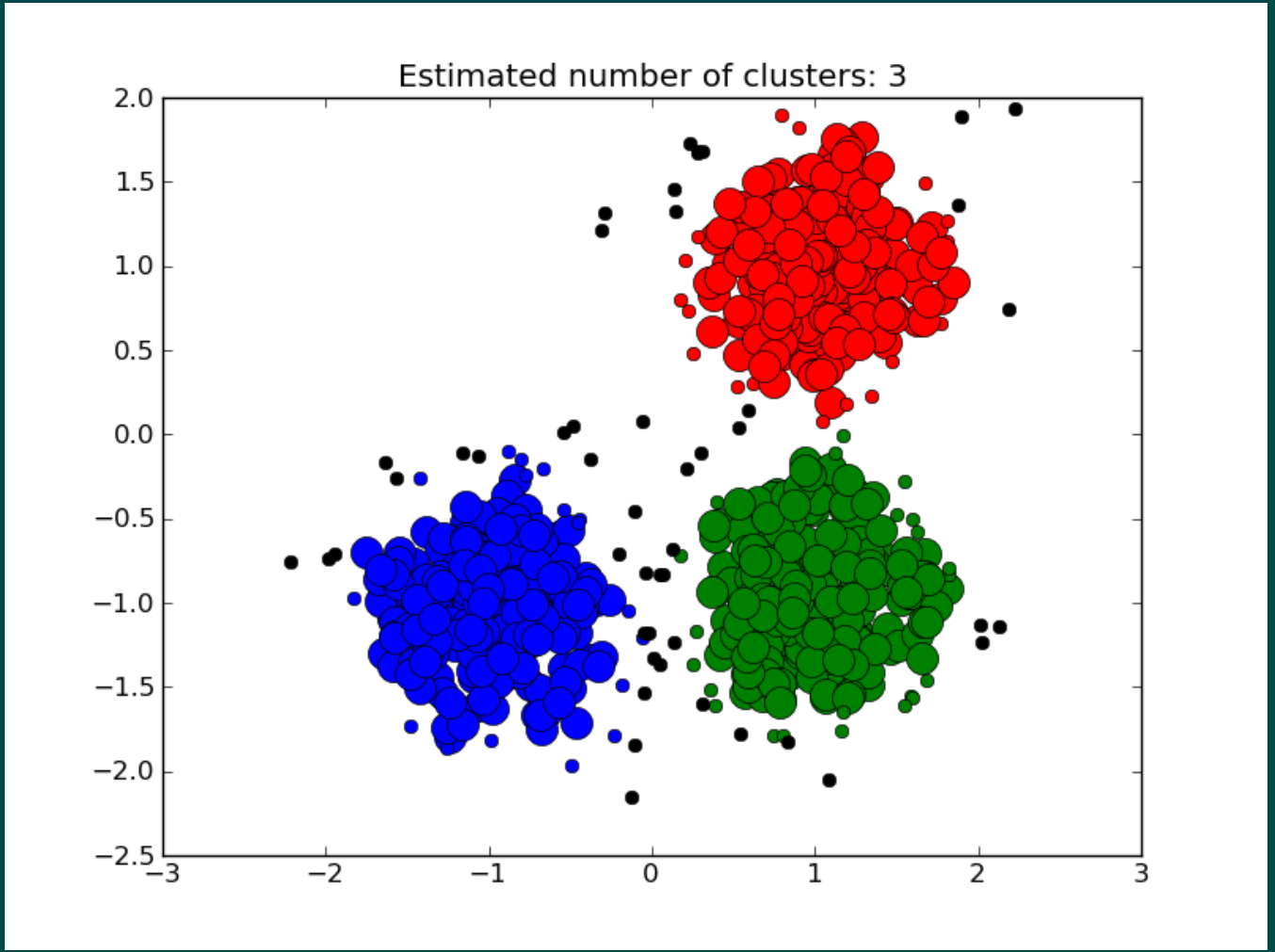
k-means clustering



Profoundis

Some Common Algorithms

DB SCAN Clustering



Profoundis

Some Example Usecases

- Log file analysis
- Outlier dectection
- Fraud Dectection
- Forcasting
- User patterns

Profoundis

A few comments

- nltk is a good(better) for text processing
- scikit-learn is for medium size problems
- for humongous projects, think of mahout
- matplotlib can be used for visualization
- visualize it in browser using d3.js
- have a look at pandas for numerical analysis

Conclusion

- This is just the tip of an iceberg.
- Scikit-learn is really cool to hack with.
- A lot of examples

(http://scikit-learn.org/stable/auto_examples/index.html)

Final words

pip install scikit-learn

Its all in the internet.

Happy Hacking!