Scikit-learn:
Machine learning in Python land
Gaël Varoquaux
1. The vision: enabling machine learning
2. The tool: a Python library
3. The project: a community
The vision: enabling machine learning
1 The vision: enabling machine learning

Machine-learning is now everywhere
The vision: enabling machine learning

Machine-learning is now everywhere

scikit-learn = an enabler
The vision: enabling machine learning

Machine-learning is now everywhere

scikit-learn = an enabler

Machine learning for everybody and for everything

Machine learning without learning the machinery
Machine learning in a nutshell

Machine learning is about making prediction from data
Expert systems
- Building decision rules

The 80s
- Eatable?
- Mobile?
- Tall?

The 90s
- Learn these from observations
- Statistical learning

2000s
- Model the noise in the observations
- Big data

Artificial Intelligence
- Machine learning on hard problems
Expert systems
- Building decision rules

Machine learning
- Learn these from observations

The 80s
The 90s

Big data today
Artificial Intelligence tomorrow
Machine learning on hard problems
1 Machine learning: a historical perspective

**Expert systems**  
- Building decision rules  

**Machine learning**  
- Learn these from observations  

**Statistical learning**  
- Model the noise in the observations  

The 80s

The 90s

2000s

G Varoquaux
Machine learning: a historical perspective

Expert systems
- Building decision rules

Machine learning
- Learn these from observations

Statistical learning
- Model the noise in the observations

Big data
- Many observations

The 80s
The 90s
2000s
today
1 Machine learning: a historical perspective

**Expert systems**
- Building decision rules

**Machine learning**
- Learn these from observations

**Statistical learning**
- Model the noise in the observations

**Big data**
- Many observations

**Artificial Intelligence**
- Machine learning on hard problems
Machine learning in a nutshell: an example

Face recognition

Andrew  Bill  Charles  Dave
Face recognition

Andrew  Bill  Charles  Dave

G Varoquaux
Machine learning in a nutshell

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

“Nearest neighbor” method
Machine learning in a nutshell

A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) image, find the image that is most similar.

"Nearest neighbor" method

How many errors on already-known images?

... 0: no errors
A simple method:

1. Store all the known (noisy) images and the names that go with them.
2. From a new (noisy) images, find the image that is most similar.

“Nearest neighbor” method

How many errors on already-known images?

... 0: no errors

Test data ≠ Train data
A single descriptor:
one dimension
A single descriptor: one dimension

Which model to prefer?
**Machine learning in a nutshell: regression**

A single descriptor: one dimension

---

**Problem of “over-fitting”**

- Minimizing error is not always the best strategy (learning noise)

- Test data \( \neq \) train data
Machine learning in a nutshell: regression

A single descriptor:
one dimension

Prefer simple models

= concept of “regularization”

Balance the number of parameters to learn with the amount of data

G Varoquaux
A single descriptor:
one dimension

Bias

variance tradeoff

Machine learning in a nutshell: regression

Balance the number of parameters to learn with the amount of data

G Varoquaux
Machine learning in a nutshell: regression

A single descriptor:
one dimension

Two descriptors:
2 dimensions

More parameters
Machine learning in a nutshell: regression

A single descriptor:
one dimension

⇒ need more data

“curse of dimensionality”

Two descriptors:
2 dimensions

More parameters
Machine learning in a nutshell: classification

Example:
recognizing hand-written digits
Example:
recognizing hand-written digits

 Represent with 2 numerical features
Machine learning in a nutshell: classification
1. Machine learning in a nutshell: unsupervised

Stock market structure

- ConocoPhillips
- American express
- Raytheon
- Boeing
- Apple
- Pepsi
- Navistar
- GlaxoSmithKline
- Microsoft
- Kimberly-Clark
- Ryder
- SAP
- Goldman Sachs
- Colgate-Palmolive
- Wal-Mart
- General Electrics
- Sony
- Pfizer
- Amazon
- Marriott
- Novartis
- Coca Cola
- 3M
- Cablevision
- Toyota
- CVS
- Bank of America
- Valero Energy
- ConocoPhillips
- Exxon
- Total
- Caterpillar
- Yahoo
- Texas instruments
- IBM
- HP
- Apple
- Amazon
- SAP
- GlaxoSmithKline
- Sanofi-Aventis
- Novartis
- Unilever
- Comcast
- Ford
- Mc Donalds
- Kraft Foods
- Xerox
- Navistar
- DuPont de Nemours
- Home Depot
- Ryder
- Marriott
- Pfizer
- Procter Gamble
- Yahoo
- Texas instruments
- IBM
- HP
- Amazon
- Apple
- Samsung
- Lenovo
- Google
- Facebook
- Instagram
- Twitter
- Netflix
- Airbnb
- Uber
- Airbnb
Machine learning in a nutshell: unsupervised

Stock market structure

Unlabeled data more common than labeled data

G Varoquaux
Machine learning

Mathematics and algorithms for fitting predictive models

Regression

Classification

Notions of overfit and test error
Machine learning is everywhere

- Image recognition
- Marketing (click-through rate)
- Movie / music recommendation
- Medical data
- Logistic chains (eg supermarkets)
- Language translation
- Detecting industrial failures
We built a machine learning library

In Python, in 2010
Real statisticians use R

- And real astronomers use IRAF
- Real economists use Gauss
- Real coders use C assembler
- Real experiments are controlled in Labview
- Real Bayesians use BUGS stan
- Real text processing is done in Perl
- Real Deep learner is best done with torch (Lua)
- And medical doctors only trust SPSS

G Varoquaux
Our stack

Python, what else?

- General purpose
- Interactive language
- Easy to read / write

Fantastic scientific ecosystem
scikit-learn goals

Machine learning for all
No specific application domain
No requirements in machine learning

High-quality Pythonic software library
Interfaces designed for users

Community-driven development
BSD licensed, very diverse contributors

http://scikit-learn.org
Between research and applications

Machine learning research

- Conceptual complexity is not an issue
- New and bleeding edge is better
- Simple problems are old science

In the field

- Tried and tested (aka boring) is good
- Little sophistication from the user
- API is more important than maths

Solving simple problems matters
Solving them really well matters a lot