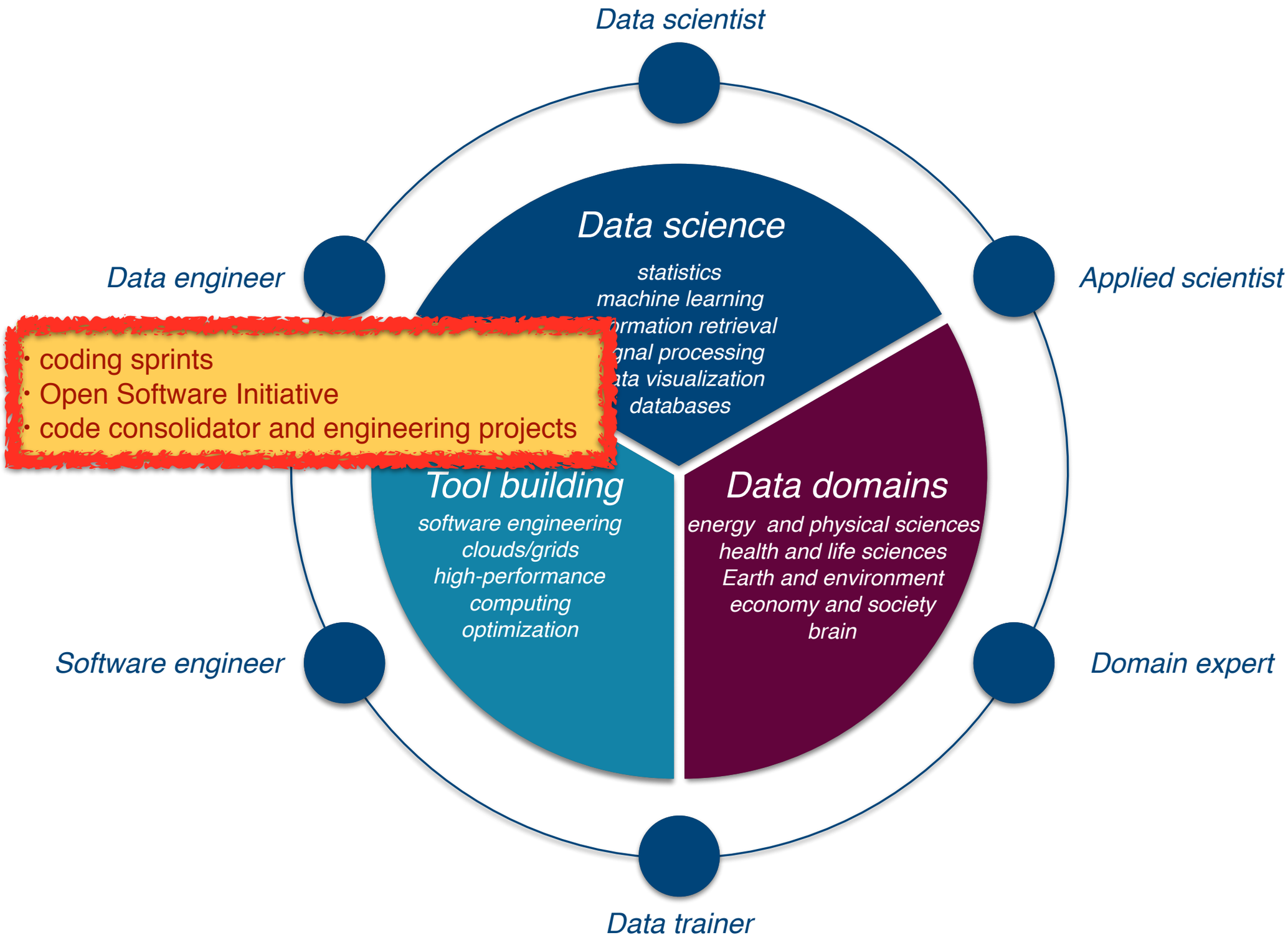
Paris-Saclay
Center for Data Science

Open Software Initiative (OSI)

Alexandre Gramfort
alexandre.gramfort@telecom-paristech.fr

Assistant Professor
LTCI, CNRS, Télécom ParisTech, Université Paris-Saclay

TOOLS IN THE DATA SCIENCE LANDSCAPE



scikit-learn: machine learning in Python — scikit-learn 0.14 documentation

scikit-learn: machine learning i...

scikit-learn.org/stable/

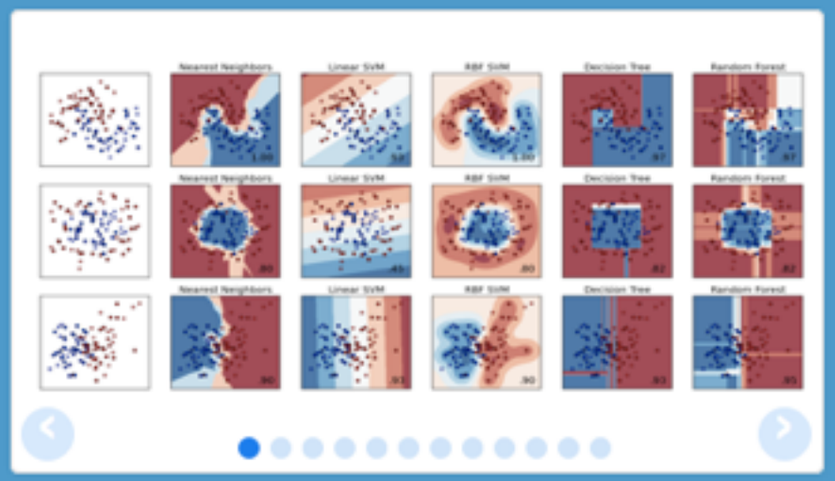
Google

scikit learn

Home | Installation | Documentation | Examples

Google™ Custom Search | Search

Fork me on GitHub



scikit-learn

Machine Learning in Python

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

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ... — *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ... — *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ... — *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *Isomap*, *non-negative matrix factorization*. — *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*. — *Examples*

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*. — *Examples*

scikit-learn: machine learning in Python — scikit-learn 0.14 documentation

scikit-learn: machine learning i...

scikit-learn.org/stable/

Google

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search

Fork me on GitHub

Classification

Identifying to which s
observation belong t

Applications: Spam
recognition.
Algorithms: *SVM*, *n*
forest, ...

Dimensionality

Reducing the numbe
consider.

Applications: Visua
efficiency
Algorithms: *PCA*, *Is*
matrix factorization.

Started in 2010 at:

université PARIS-SACLAY

In a Nutshell, scikit learn...

... has had 20,181 commits made by 650 contributors
representing 179,355 lines of code

... is mostly written in Python
with a well-commented source code

... has a well established, mature codebase
maintained by a very large development team
with stable Y-O-Y commits

... took an estimated 47 years of effort (COCOMO model)
starting with its first commit in January, 2010
ending with its most recent commit 3 days ago

source: https://www.openhub.net/p/scikit-learn

scikit-learn: machine learning in Python — scikit-learn 0.14 documentation

scikit-learn: machine learning i...

scikit-learn.org/stable/

Google

scikit learn

Home Installation Documentation Examples

Google™ Custom Search Search

Fork me on GitHub

Classification

Identifying to which s
observation belong t

Applications: Spam
recognition.
Algorithms: *SVM*, *n*
forest, ...

Dimensionality

Reducing the numbe
consider.

Applications: Visua
efficiency
Algorithms: *PCA*, *Is*
matrix factorization.

Started in 2010 at:

université PARIS-SACLAY

In a Nutshell, scikit learn...

... has had 20,181 co
representing 17

... is mostly wri
with a well-

... has a well
maintained
with stable

... took an estim
starting with its
ending with its m

Funding:

Inria INVENTEURS DU MONDE NUMÉRIQUE

TELECOM ParisTech

Google

NYU CENTER FOR DATA SCIENCE

Paris-Saclay Center for Data Science

criteo

source: https://www. scikit-learn

PeerIndex

fnrs
LA LIBERTÉ DE CHERCHER

Inria
INVENTORS FOR THE DIGITAL WORLD

TELECOM ParisTech

DataRobot

lovely

ŷhat

how about we...

okc

Spotify

INFONEA
Comma Soft AG

DATA PUBLICA

dataiku
Write your own data story

airbnb

AWeber
COMMUNICATIONS

Bestof MEDIA GROUP

RANGE SPAN

EVERNOTE

http://scikit-learn.org/stable/testimonials/testimonials.html

- Installed on 1% of debian systems
- 1200 job offers on Stack-Overflow
- Users: 60% academics & 40% industry

http://scikit-learn.org/stable/testimonials/testimonials.html

BEYOND SCIKIT-LEARN @ UPSA (… IN MY WORLD)

MNE
MEG + EEG ANALYSIS & VISUALIZATION

neo

NiBabel
Access a cacophony of neuro-imaging file formats

and certainly many more....

Objective is too teach students to replicate the scikit-learn model or at least contribute to it

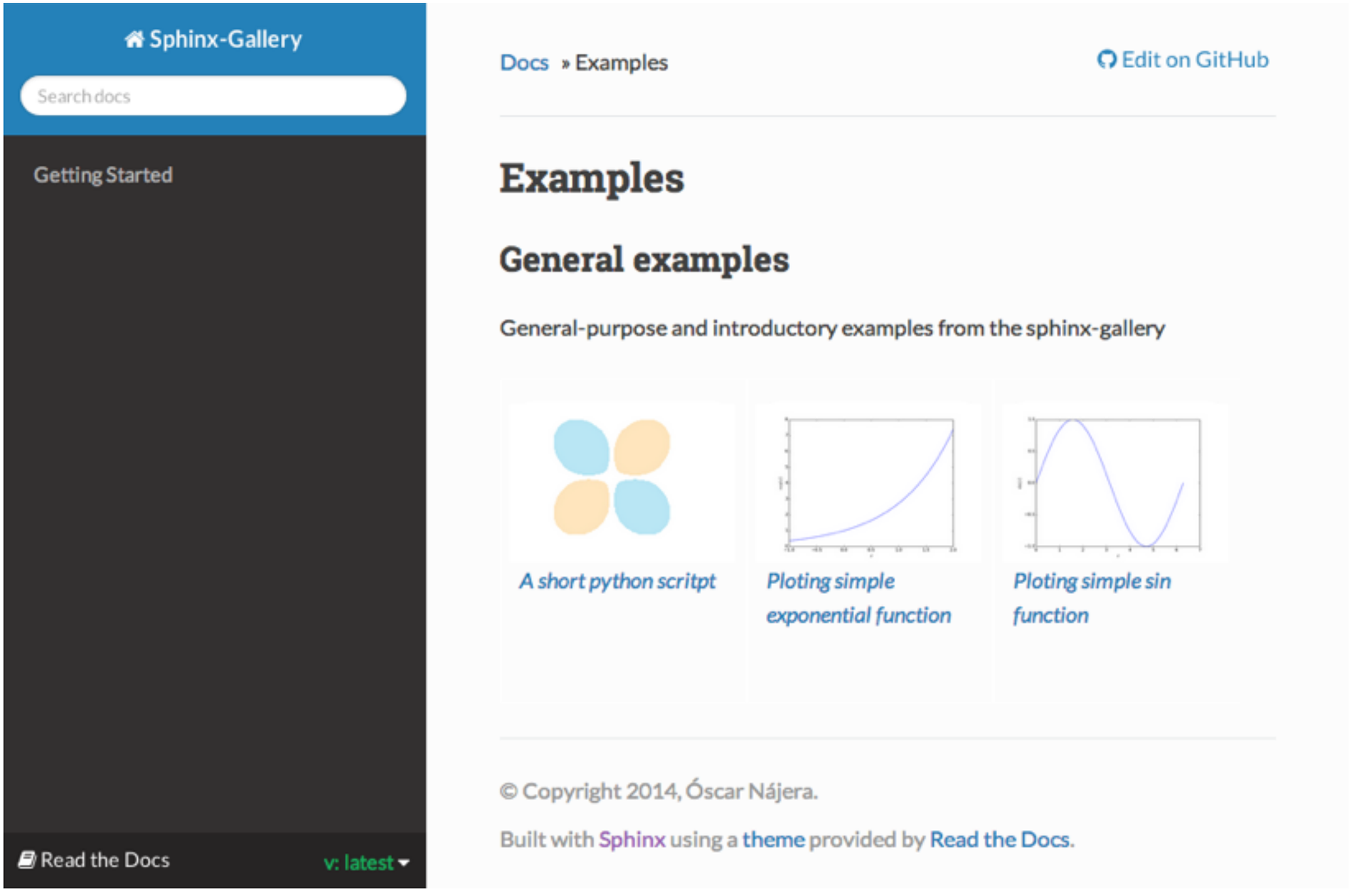
Open Software Initiative @ CDS 1

- 6 projects funded (5 missions doctorales + 1 code consolidator 3 months)
  - N. Goix (LTCI) : Scikit-Learn "Outlier/Novelty detection algorithm (IForest, etc.)"
  - M. Cherti (UpSud) : Scikit-Learn "Multivariate Additive Regression Splines (MARS)"
  - R. Brault (IBISC) : OperaLib "Large scale learning with operator valued kernels"
  - O. Najera (UpSud) : NiLearn "Sphinx-Gallery: Facilitate code documentation with generated example gallery"
  - S. Mishra (UpSud) : OpenMEEG/MNE "Integration of OpenMEEG into the MNE software via Python bindings"
  - Lorenzo De Santis (UpSud) : MNE "Port from C to Python dipole fitting solver for source localization"
  - Xin Su (LTCI) : START^2 "SyntheTic Aperture Radar Time series Toolbox"

Open Software Initiative @ CDS 1

- 6 projects funded (5 missions doctorales + 1 code consolidator 3 months)
  - N. Goix (LTCI) : Scikit-Learn "Outlier/Novelty detection algorithm (IForest, etc.)"
  - M. Cherti (UpSud) : Scikit-Learn "Multivariate Additive Regression Splines (MARS)"
  - R. Brault (IBISC) : OperaLib "Large scale learning with operator valued kernels"
  - O. Najera (UpSud) : NiLearn "Sphinx-Gallery: Facilitate code documentation with generated example gallery"
  - S. Mishra (UpSud) : OpenMEEG/MNE "Integration of OpenMEEG into the MNE software via Python bindings"
  - Lorenzo De Santis (UpSud) : MNE "Port from C to Python dipole fitting solver for source localization"
  - Xin Su (LTCI) : START^2 "SyntheTic Aperture Radar Time series Toolbox"

OSI : O. Najera “Sphinx-Gallery”



http://sphinx-gallery.readthedocs.io/en/latest/

https://github.com/sphinx-gallery/sphinx-gallery

Welcome to Sphinx-Gallery's documentation!

A Sphinx extension that builds an HTML gallery of examples from any set of Python scripts.

It is extracted from the scikit-learn project and aims to be an independent general purpose extension.

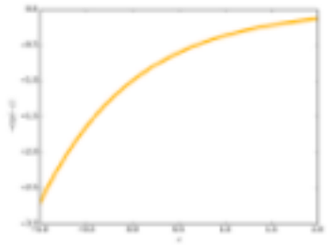
Why Sphinx-Gallery?

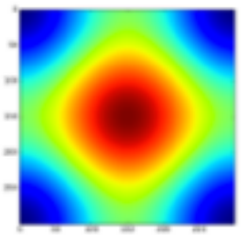
- Simple examples that run out of the box are the best way to learn a library
- Pleasing, organized, visual layouts
- Links, searching, backlinks throughout examples and documentation

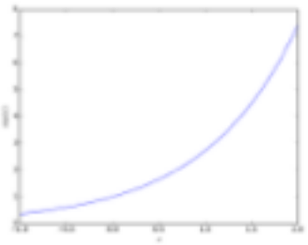
What does it look like?

Here is an example gallery generated from a few Python scripts.

Examples using `numpy.linspace`







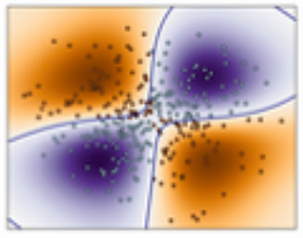
http://sphinx-gallery.readthedocs.io/en/latest/

OSI : O. Najera “Sphinx-Gallery”

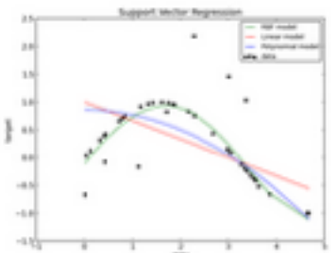
http://scikit-learn.org/stable/auto_examples/index.html

Support Vector Machines

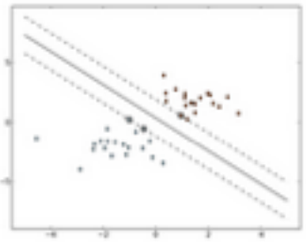
Examples concerning the `sklearn.svm` package.



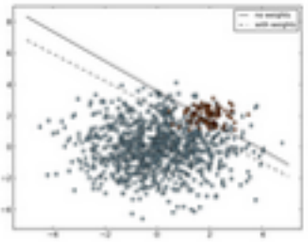
Non-linear SVM



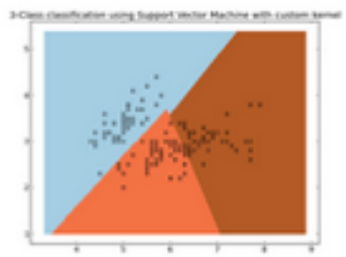
Support Vector Regression (SVR) using linear and non-linear kernels



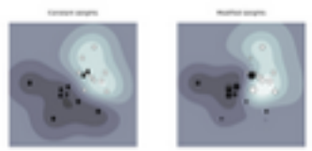
SVM: Maximum margin separating hyperplane



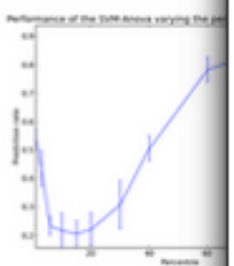
SVM: Separating hyperplane for



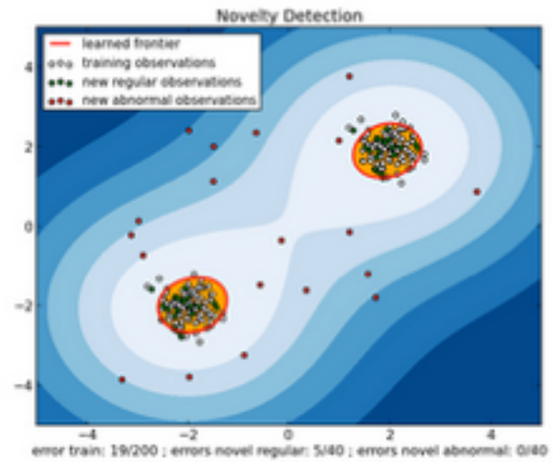
SVM with custom kernel



SVM: Weighted samples



SVM-Anova: SV univariate featu selection

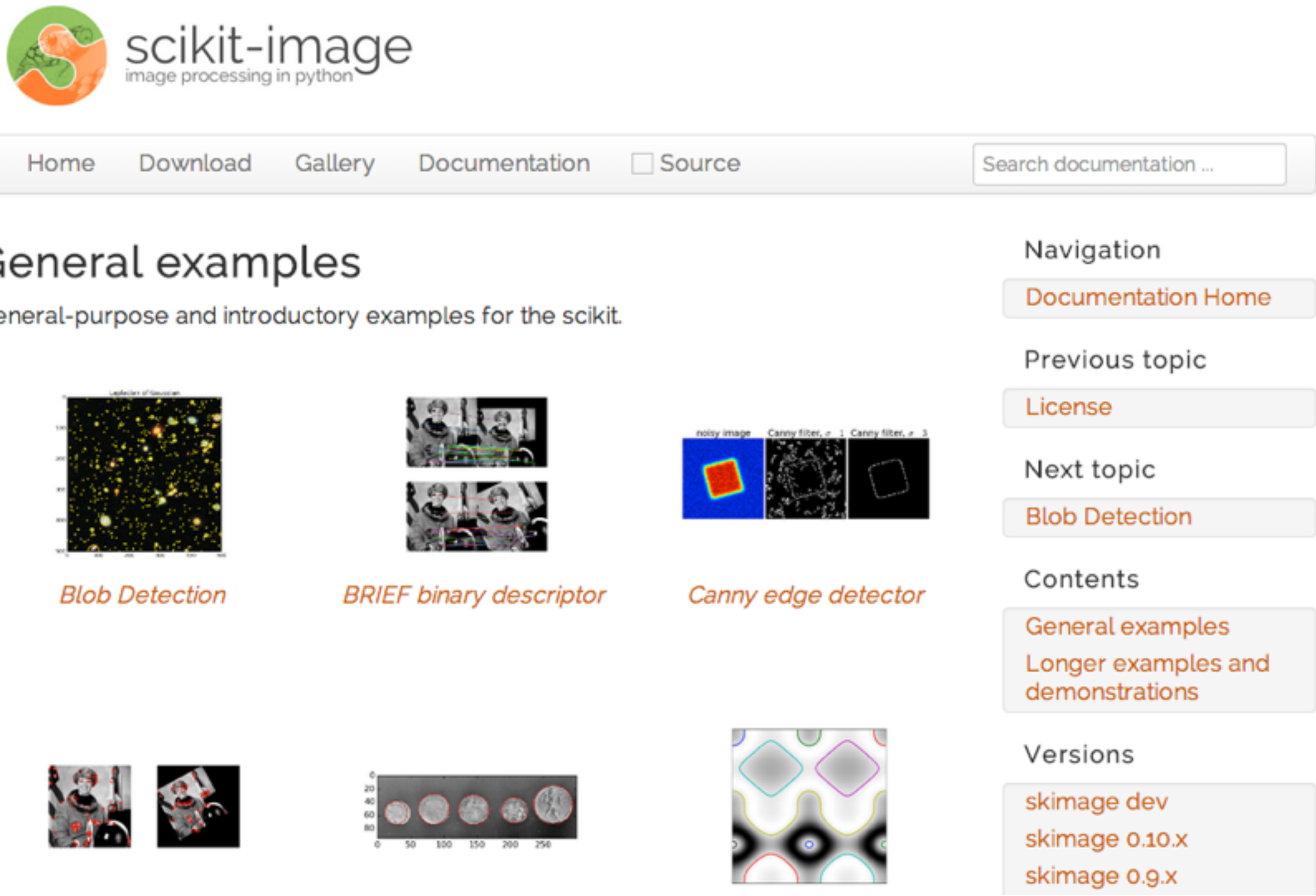


One-class SVM with non-linear kernel (RBF)

An example using a one-class SVM for novelty detection.

OSI : O. Najera “Sphinx-Gallery”

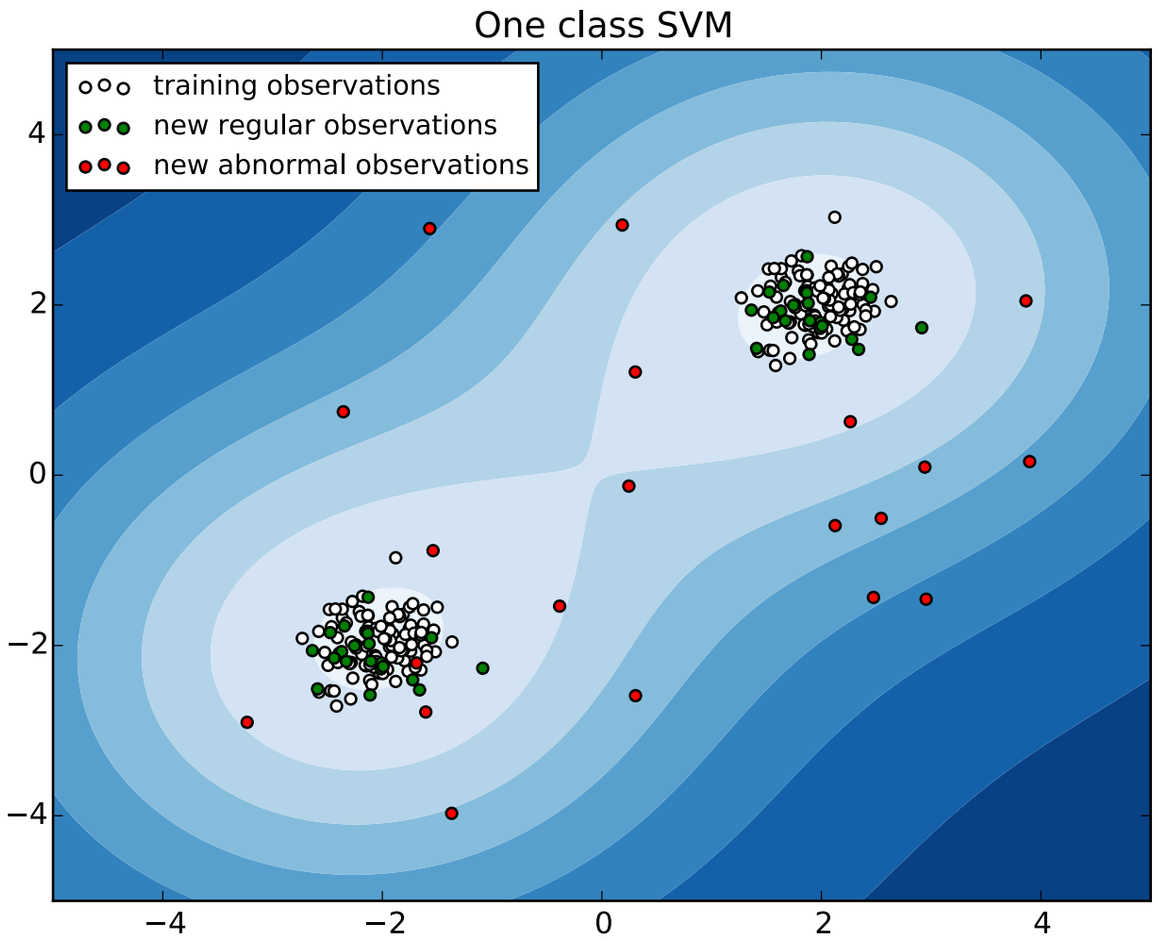
http://scikit-image.org/docs/dev/auto_examples/

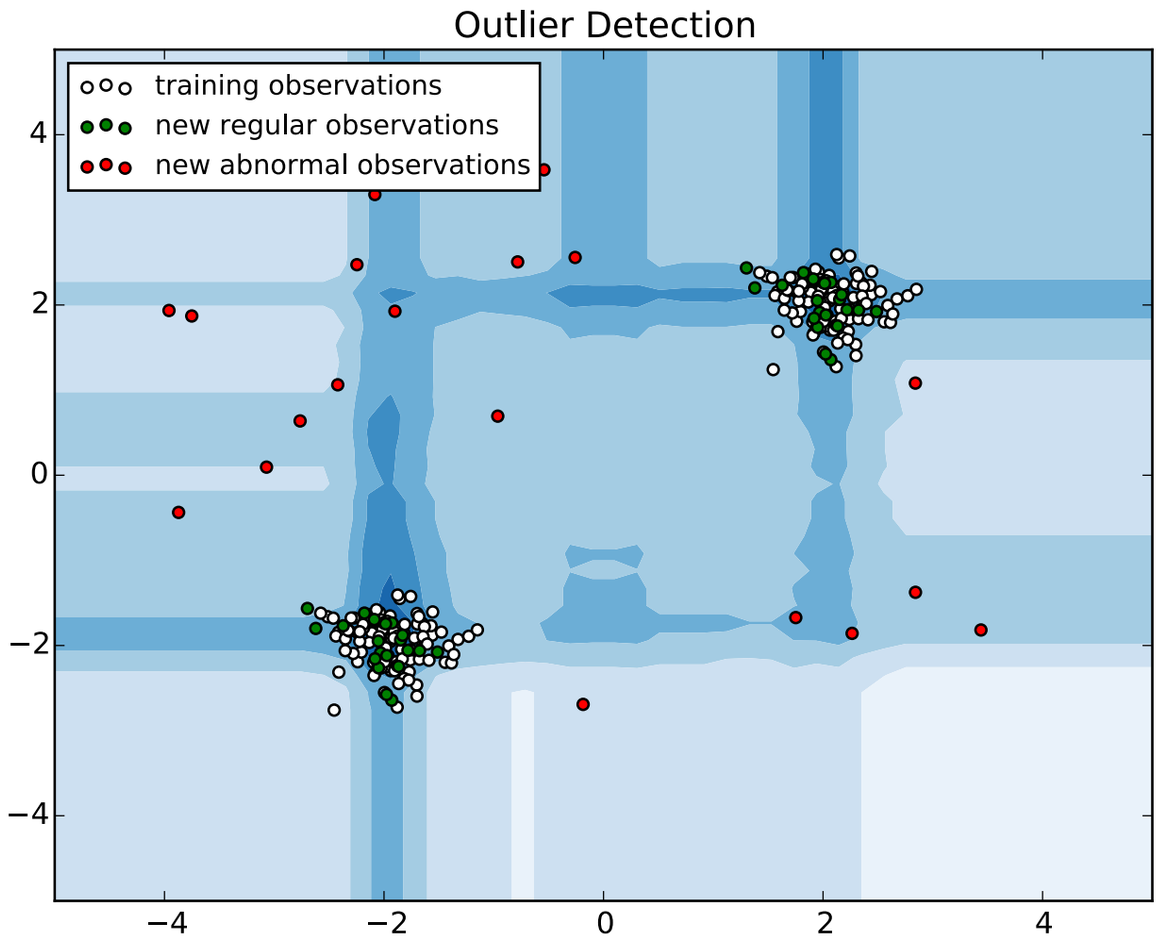


Open Software Initiative @ CDS 1

- 6 projects funded (5 missions doctorales + 1 code consolidator 3 months)
  - N. Goix (LTCI) : Scikit-Learn "Outlier/Novelty detection algorithm (IForest, etc.)"
  - M. Cherti (UpSud) : Scikit-Learn "Multivariate Additive Regression Splines (MARS)"
  - R. Brault (IBISC) : OperaLib "Large scale learning with operator valued kernels"
  - O. Najera (UpSud) : NiLearn "Sphinx-Gallery: Facilitate code documentation with generated example gallery"
  - S. Mishra (UpSud) : OpenMEEG/MNE "Integration of OpenMEEG into the MNE software via Python bindings"
  - Lorenzo De Santis (UpSud) : MNE "Port from C to Python dipole fitting solver for source localization"
  - Xin Su (LTCI) : START^2 "SyntheTic Aperture Radar Time series Toolbox"

OSI : N. Goix “Outlier detection in Scikit-Learn”

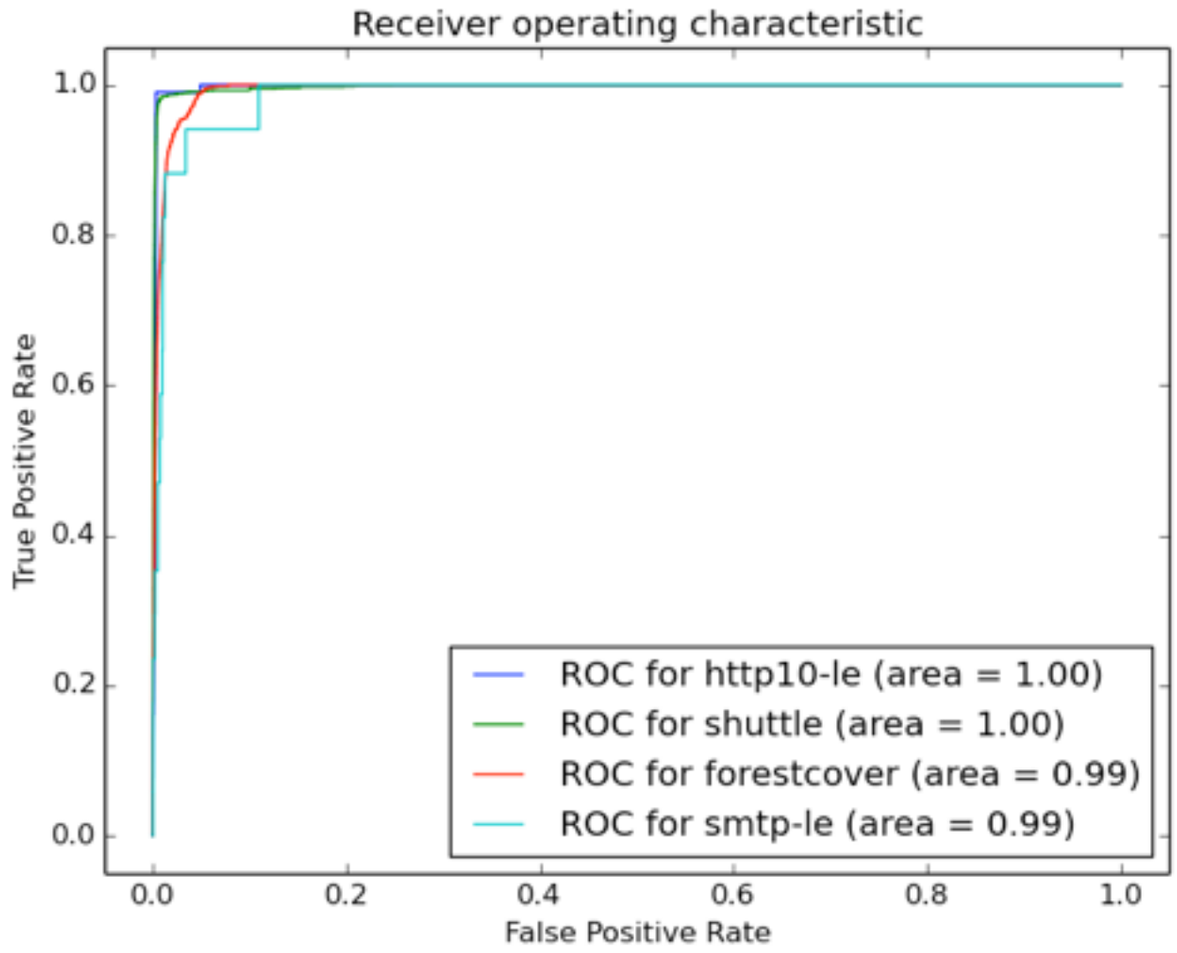




http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

F.T. Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on, pages 413--422, Dec 2008.

OSI : N. Goix "Outlier detection in Scikit-Learn"



Result on
KDD 99 dataset

http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html

F.T. Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on, pages 413--422, Dec 2008.

How to apply?

- Process:
  - Software project proposes eventually a topic. List from 2015:
    - https://github.com/paris-saclay-cds/open-software-initiative/wiki/Projects
  - Student applies by writing a project proposal of 2 pages with a roadmap
  - Proposal is refined by iterating between mentor and student

Coding sprints

MNE (http://martinos.org/mne) and NiLearn (http://nilearn.github.io/)

1 week sprint in June 2014 at Telecom ParisTech
15 attendees (from Saclay, UK, Finland, Austria, Germany, USA)

CONCLUSIONS

- OSI:
  - offers great training for students (team development, collaborative work)
  - offers visibility for the students !
  - teaches students how to write code that others can read so good code !
  - makes students discover a new scientific field