

Alexandre Gramfort

Dept. TSI, Telecom ParisTech, Institut Mines-Télécom

alexandre.gramfort@telecom-paristech.fr

Projet Résilience
Jan, 2015



- Développer les **outils de machine learning** pour analyser les logs des machines sous SlapOS
- et “Apprendre à predire les pannes”
- **Contraintes:**
 - Données volumineuses: Algorithmes avec complexité (quasi-)linéaire en mémoire et temps de calcul
 - Nécessité de traiter les données sous forme de flot (pas tout en mémoire)
 - Algorithmes supervisés et non-supervisés pour traiter des données non-étiquetées
 - Facilité de déploiement des modèles sur le cloud

The screenshot shows the official scikit-learn website. At the top, there's a navigation bar with links for Home, Installation, Documentation, Examples, Google Custom Search, and a search bar. A "Fork me on GitHub" button is also visible. On the left, there's a grid of nine small plots illustrating various machine learning algorithms like SVM, Random Forest, and AdaBoost. The main title "scikit-learn" is prominently displayed in large white letters, followed by the subtitle "Machine Learning in Python". Below the title is a bulleted list of features:

- 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 belongs 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

<http://scikit-learn.org>

In a Nutshell, scikit learn...

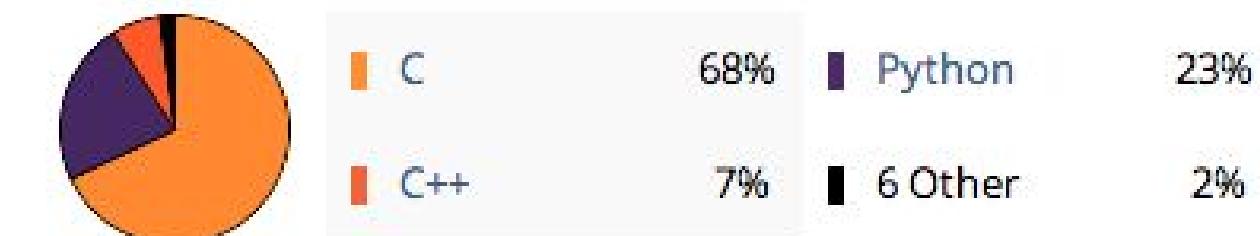
... has had 17,955 commits made by 458 contributors representing 438,986 lines of code

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

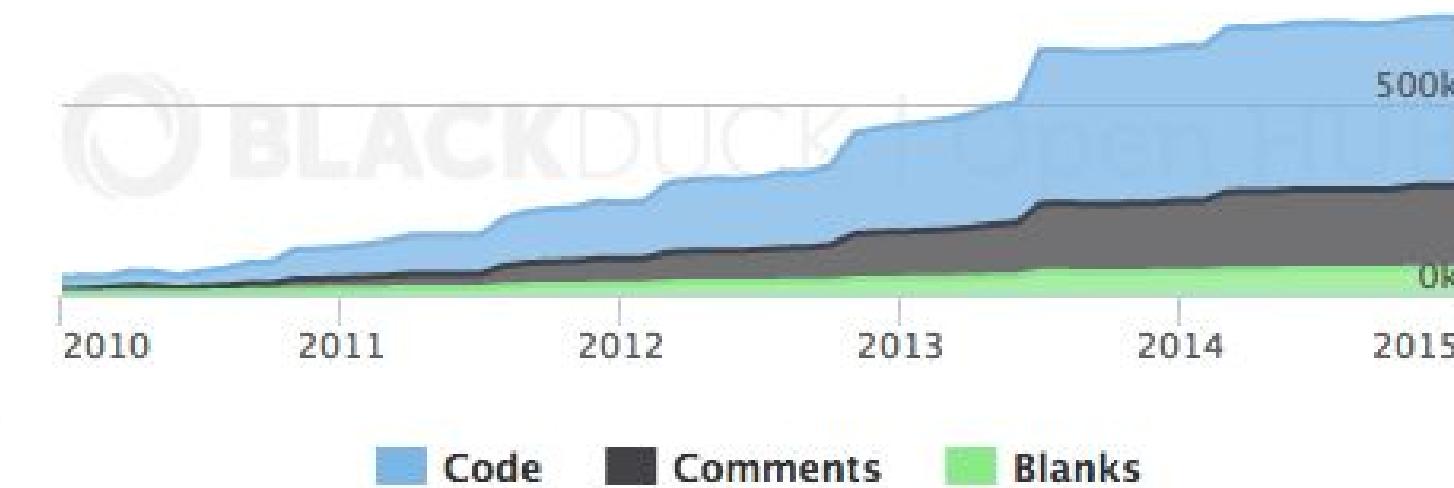
... has a codebase with a long source history
maintained by a very large development team
with decreasing Y-O-Y commits

... took an estimated 119 years of effort (COCOMO model)
starting with its first commit in January, 2010
ending with its most recent commit 3 days ago

Languages



Lines of Code



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



DATA PUBLICA

*inria*
INVENTORS FOR THE DIGITAL WORLD

Spotify®



PHIMECA

RANGE SPAN

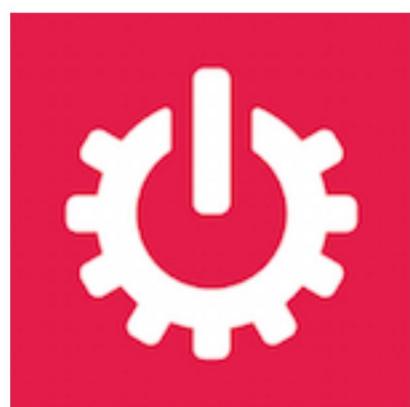
ŷhat



lovely



EVERNOTE

Best of
MEDIA
GROUPsource: <https://www.openhub.net/p/scikit-learn>

- Implémentation des méthodes de l'état de l'art en **classification supervisé**:
 - Averaged stochastic gradient descent (ASGD) [1]
 - Stochastic Average Gradient (SAG) [2]
- Implémentation d'un **algorithme non-supervisé de clustering online**:
 - BIRCH [3]

[1] Large-Scale Machine Learning with Stochastic Gradient Descent, Léon Bottou, <http://leon.bottou.org/publications/pdf/compstat-2010.pdf>

[2] Minimizing Finite Sums with the Stochastic Average Gradient, Mark Schmidt, Nicolas Le Roux, Francis Bach, <http://arxiv.org/abs/1309.2388>

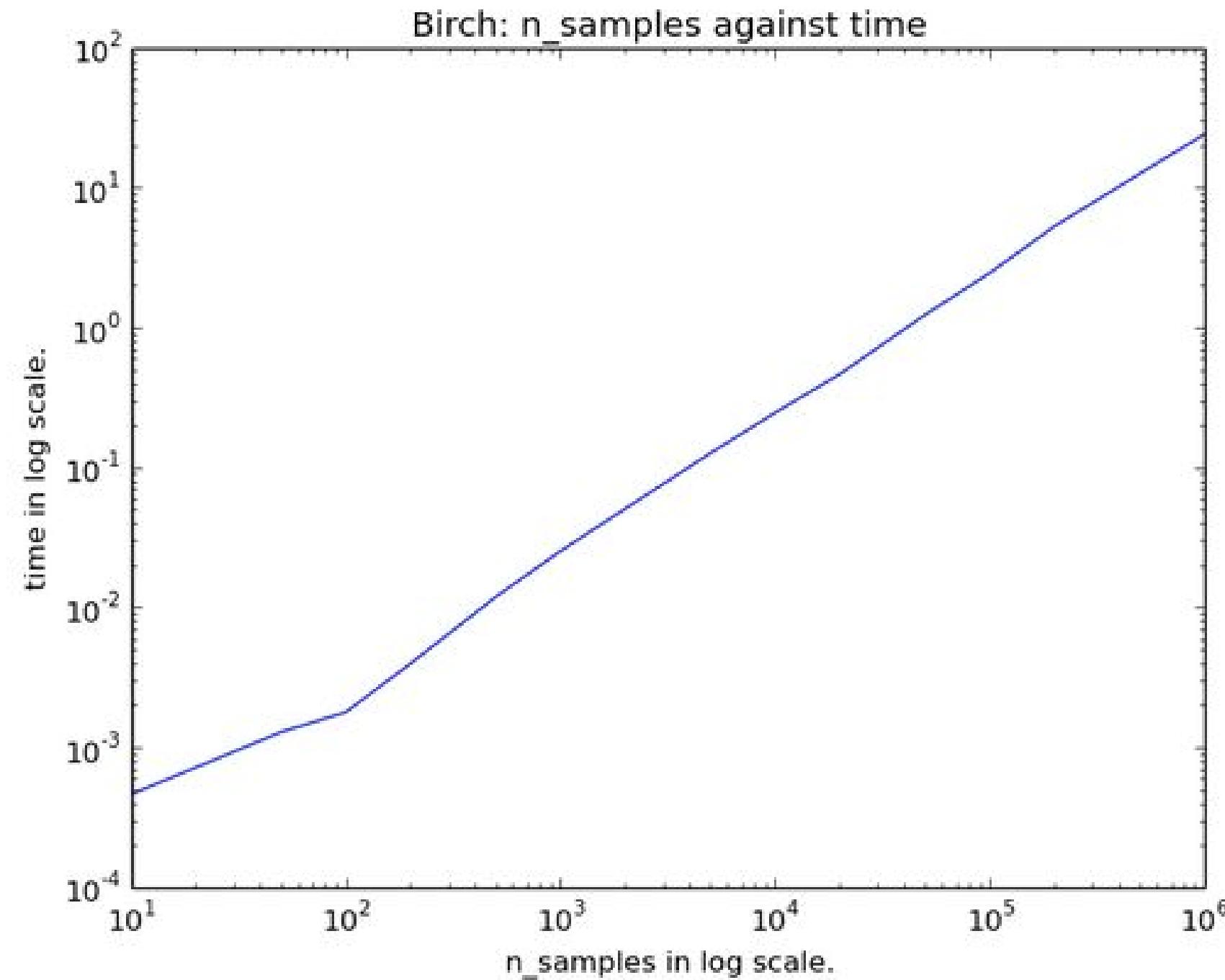
[3] Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large



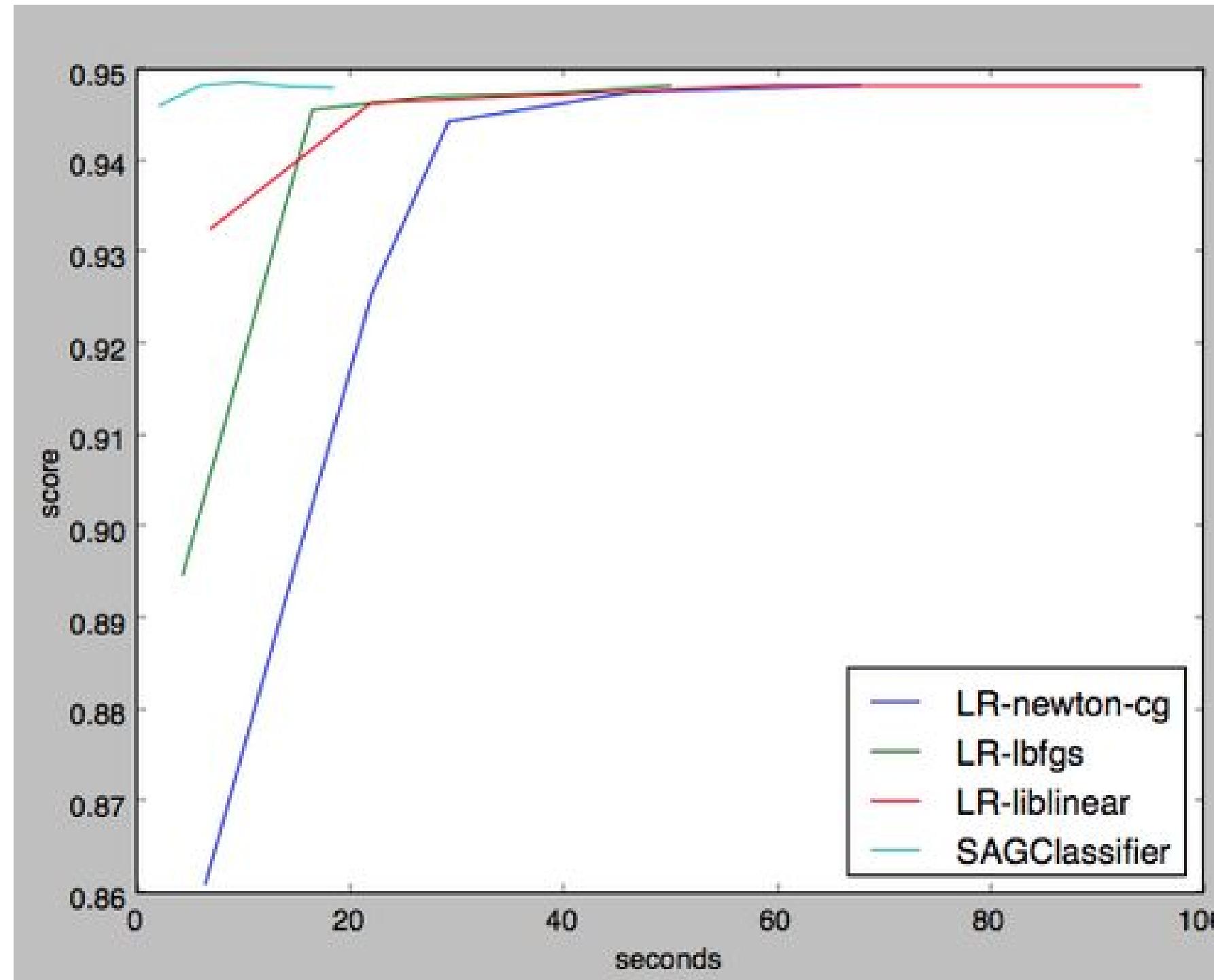
The screenshot shows the scikit-learn documentation page for the `Birch` clustering algorithm. The top navigation bar includes links for Home, Installation, Documentation (with a dropdown menu), Examples, Google Custom Search, and a Search bar. A red banner on the right says "Fork me on GitHub". The main content area has a light blue header with the title `sklearn.cluster.Birch`. Below it is a code block showing the class definition:

```
class sklearn.cluster.Birch(threshold=0.5, branching_factor=50, n_clusters=3, compute_labels=True, copy=True)
```

With a "[source]" link to the right. The text explains that it implements the Birch clustering algorithm, which inserts new samples into the root of a Clustering Feature Tree and clubs them with subclusters until a leaf node's centroid is closest. There are also sections for parameters, examples, and related modules.



BIRCH has an almost linear complexity
BIRCH scales to massive datasets like the logs of SlapOS



The SAG solvers is about 10 times faster to reach optimal performance compared to alternative solvers present before in scikit-learn.

SlapOS Log Prediction

```
slapuser10,17557.0,None,0.0,16924.63,3.0,20.8439957728900005,811831296.0,6337482752.0,3825119.0,2014-10-26,0
```

The probability that this user will cause a problem is 0.77



Example Log Messages:

```
slapuser11,17557.0,None,0.0,16924.63,1.0,10.8439957728900005,811831296.0,6337482752.0,3825119.0,2015-02-02,04:00:02,0
slapuser10,17557.0,None,0.0,16924.63,3.0,20.8439957728900005,811831296.0,6337482752.0,3825119.0,2014-10-26,00:00:02,0
slapuser5,8479.0,None,0.0,9839.94,24.0,4.7961274004400005,806981632.0,9069555712.0,1000070.0,2014-07-21,00:35:01,0
```

Alexandre Gramfort

Dept. TSI, Telecom ParisTech, Institut Mines-Télécom

alexandre.gramfort@telecom-paristech.fr

Projet Résilience
jan, 2015



- Développer les **outils de machine learning** pour analyser les logs des machines sous SlapOS

- et “Apprendre à predire les pannes”

- **Contraintes:**

- Données volumineuses: Algorithmes avec complexité (quasi-)linéaire en mémoire et temps de calcul
- Nécessité de traiter les données sous forme de flot (pas tout en mémoire)
- Algorithmes supervisés et non-supervisés pour traiter des données non-étiquetées
- Facilité de déploiement des modèles sur le cloud

A. Gramfort

Projet Réalisation

The screenshot shows the official scikit-learn website at <http://scikit-learn.org>. The header includes the logo, navigation links for Home, Installation, Documentation, Examples, and a search bar. A banner on the right says "Scikit-learn 0.19.1". The main content area features a grid of small images representing various machine learning applications. Below this, three sections are listed: Classification, Regression, and Clustering, each with a brief description, applications, and algorithms. At the bottom, there's a footer with links to GitHub, Stack Overflow, and a "Get Involved" button.

scikit-learn
Machine Learning in Python

Classification

Identifying to which set of categories a new observation belongs.
Applications: Spam detection, Image recognition.
Algorithms: SVM, nearest neighbors, random forest, ...

Regression

Predicting a continuous value for a new example.
Applications: Drug response, Stock prices.
Algorithms: SVR, ridge regression, Lasso, ...

Clustering

Automatic grouping of similar objects into sets.
Applications: Customer segmentation, Grouping experiment outcomes.
Algorithms: k-Means, spectral clustering, mean-shift, ...

<http://scikit-learn.org>

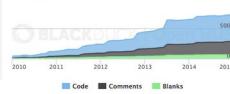
In a Nutshell, scikit learn...

- has had 17,955 commits made by 458 contributors representing 438,986 lines of code
- is mostly written in C with a very well-commented source code
- has a codebase with a long source history maintained by a very large development team with decreasing Y-O-Y commits
- took an estimated 119 years of effort (CCOMO model) starting with its first commit in January, 2010 ending with its most recent commit 3 days ago

Languages



Lines of Code



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



- Implémentation des méthodes de l'état de l'art en
classification supervisée:

- Averaged stochastic gradient descent (ASGD) [1]
- Stochastic Average Gradient (SAG) [2]

- Implémentation d'un **algorithme non-supervisé de clustering online:**

- BIRCH [3]

[1] Large-Scale Machine Learning with Stochastic Gradient Descent, Léon Bottou,
<http://leon.bottou.org/publications/pdf/compsstat-2010.pdf>

[2] Minimizing Finite Sums with the Stochastic Average Gradient, Mark Schmidt, Nicolas Le Roux, Francis Bach, <http://arxiv.org/abs/1309.2388>

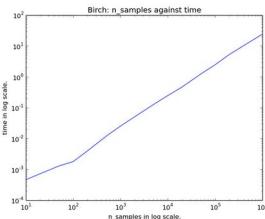
[3] Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large

A. Gramfort

Projet Réalisation

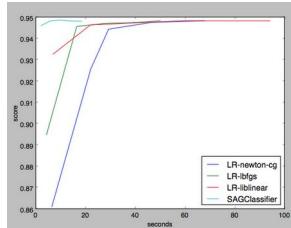
The screenshot shows a web browser displaying the scikit-learn documentation. The URL in the address bar is <http://scikit-learn.org/dev/modules/generated/sklearn.cluster.Birch.html>. The page title is "sklearn.cluster.Birch". The header includes links for Home, Installation, Documentation, Examples, and a search bar. The main content area contains the class definition for `sklearn.cluster.Birch`, its docstring, and parameters. A sidebar on the left provides navigation links for Periods, Next, Up, Previous, and Other versions, along with information about the current scikit-learn version (0.16.dev) and a link to the GitHub repository.

http://scikit-learn.org/dev/modules/generated/sklearn.cluster.Birch.html



BIRCH has an almost linear complexity
BIRCH scales to massive datasets like the logs of SlapOS

<http://scikit-learn.org/dev/modules/generated/sklearn.cluster.Birch.html>



The SAG solvers is about 10 times faster to reach optimal performance compared to alternative solvers present before in scikit-learn.

SlapOS Log Prediction

```
slapuser10,17557,0,None,0,0,16924,63,3,0,20,843995772890005,811831296,0,6337482752,0,3825119,0,2014-10-26,
```

The probability that this user will cause a problem is 0.77



Example Log Messages:

```
slapuser11,17557,A,None,A,16924,43,1,0,18,43995772890005,811831296,0,6337482752,0,3825119,A,2815-42-40,84,00-02,0  
slapuser10,17557,A,None,A,16924,43,3,0,20,43995772890005,811831296,0,6337482752,X,3825119,A,2814-39-40,84-02,0  
slapuser9,2079,A,None,A,903534,14,A,-7931174894489000,,000001531,A,3809976,A,2014-07-21,06:35:01,2
```