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

http://scikit-learn.org

In a Nutshell, scikit learn...

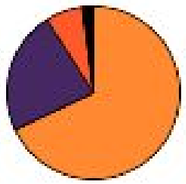
... has had 17,955 commits made by 458 contributors
representing 438,986 lines of code

... 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

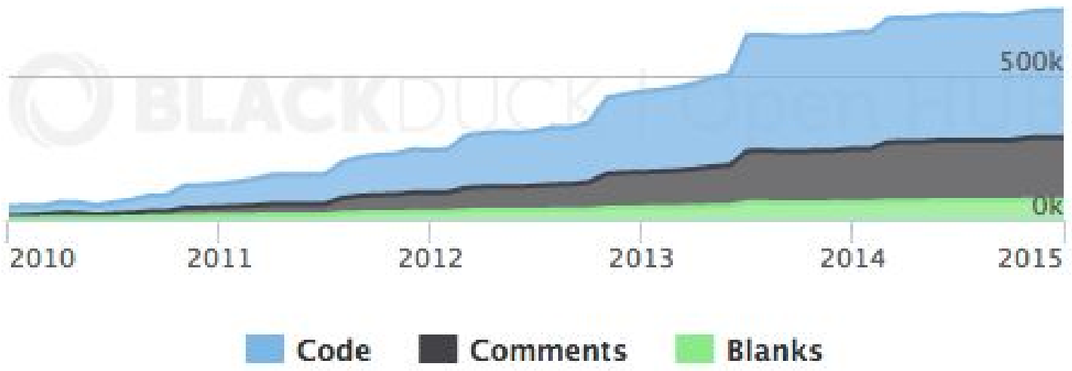
... took an estimated 119 years of effort (COCOMO model)
starting with its first commit in January, 2010
ending with its most recent commit 3 days ago

Languages



C	68%	Python	23%
C++	7%	6 Other	2%

Lines of Code



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

DATA PUBLICA

lovely

DataRobot

EVERNOTE

Best of MEDIA GROUP

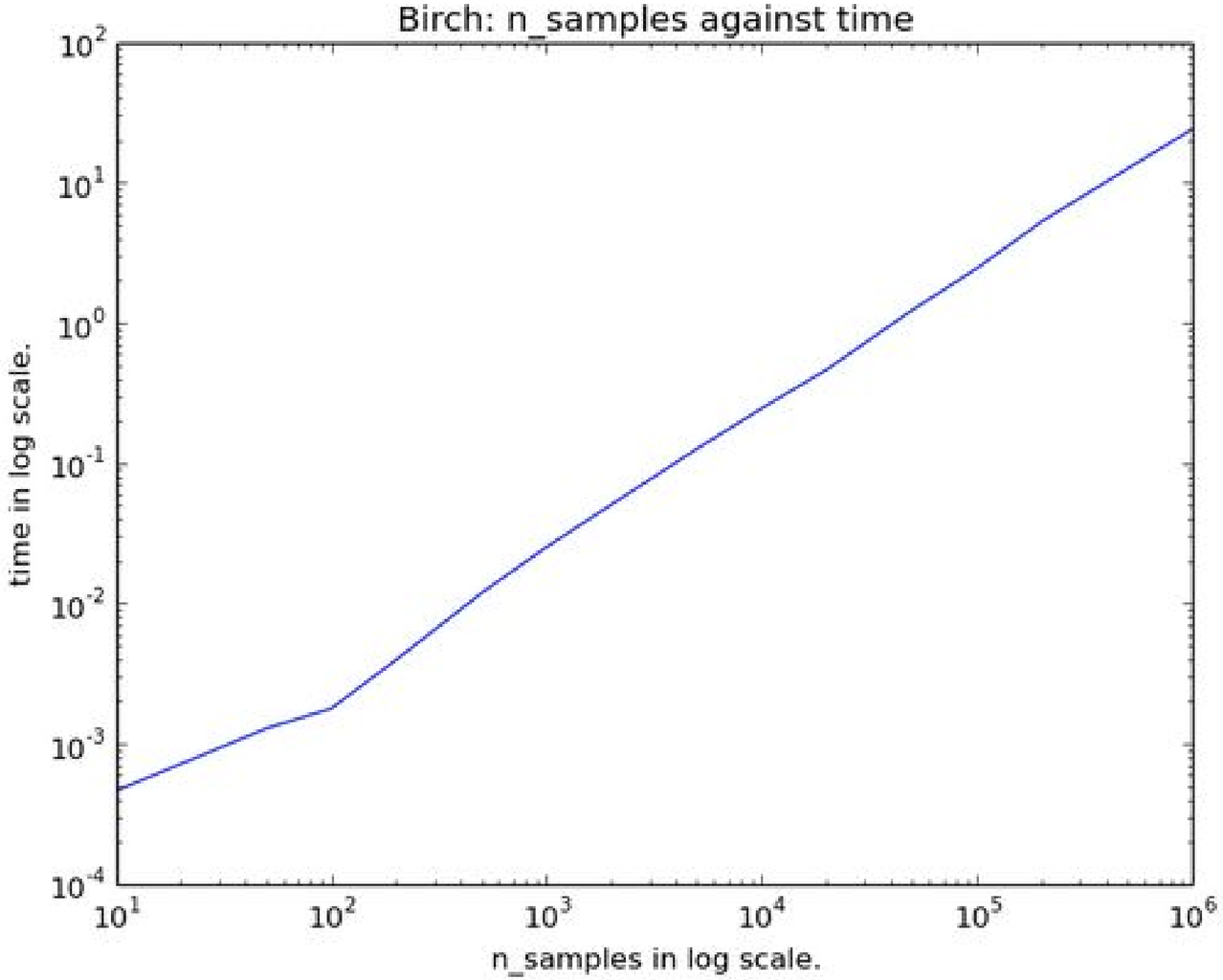
ŷhat

source: 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]

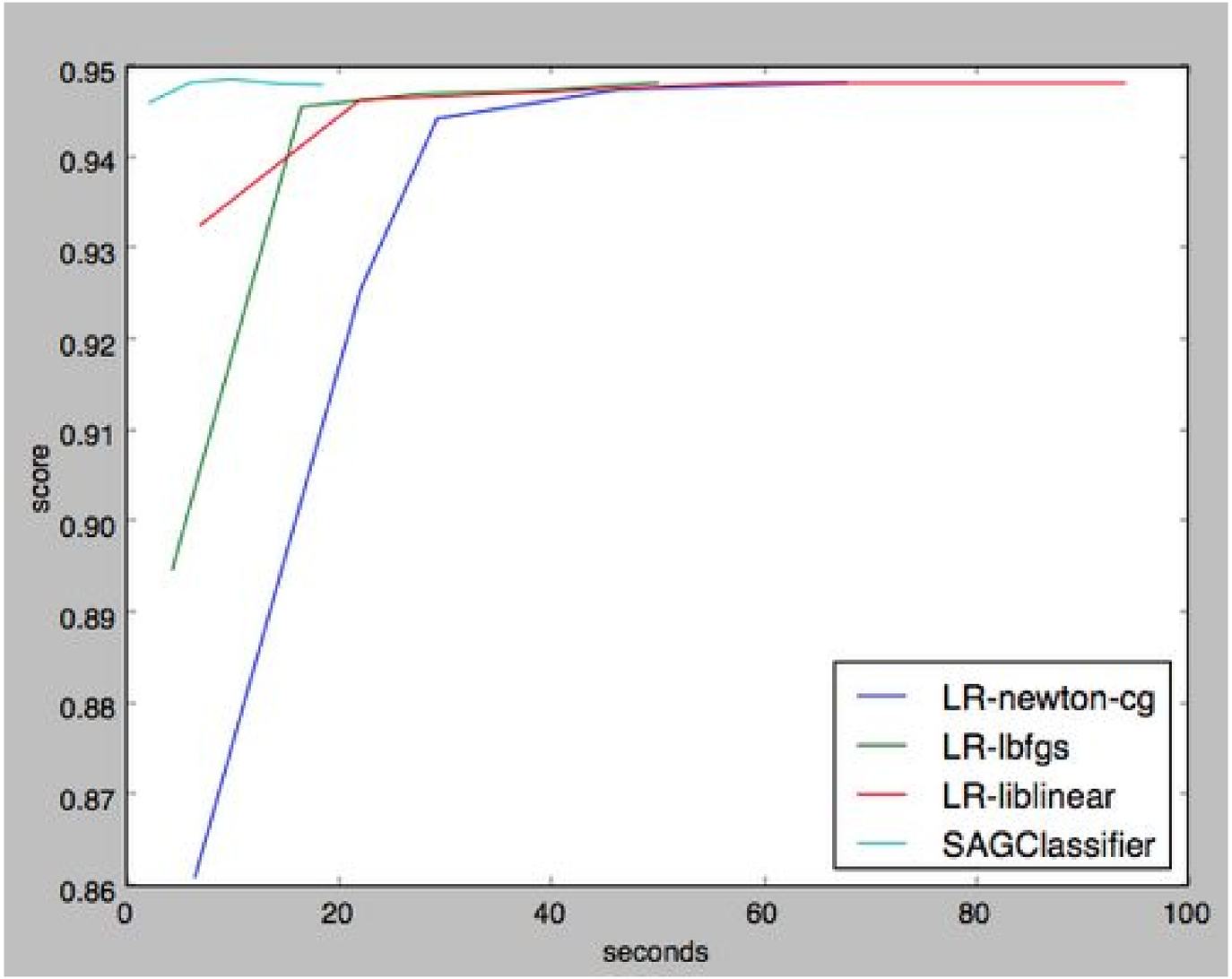
[1] Large-Scale Machine Learning with Stochastic Gradient Descent, Léon Bottou, http://leon.bottou.org/publications/pdf/compstat-2010.pdf
[2] Minimizing Finite Sums with the Stochastic Average Gradient, Mark Schmidt, Nicolas Le Roux, Francis Bach, http://arxiv.org/abs/1309.2388
[3] Tian Zhang, Raghu Ramakrishnan, Maron Livny BIRCH: An efficient data clustering method for large

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,0.0,16924.63,3.0,20.8439957728900005,811831296.0,6337482752.0,3825119.0,2014-10-26,(

The probabilty that this user will cause a problem is 0.77

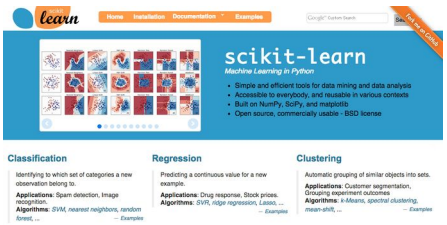
Example Log Messages:

```
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



http://scikit-learn.org

In a Nutshell, scikit learn...

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

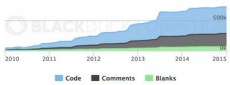
-- is mostly written in C
with a very well-commented source code

-- 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



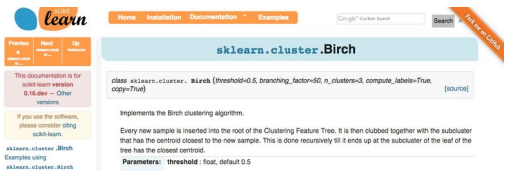
Code Comments Blanks

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

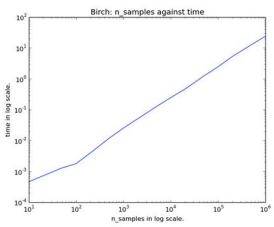
Source: https://www.openhub.net/p/scikit-learn

- 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

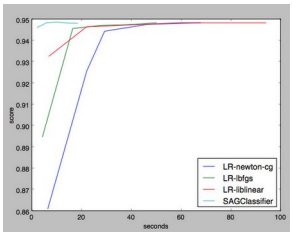


http://scikit-learn.org/dev/modules/generated/sklearn.cluster.Birch.html



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,0.0,16924.63,3.0,20.843995772890000 5,811831296.0,6337482752.0,3825119.0,2014-10-26,(

The probabilty that this user will cause a problem is 0.77

Example Log Messages:

```
slapuser11,17557.0,None,0.0,16924.63,1.0,10.843995772890000 5,811831296.0,6337482752.0,3825119.0,2015-02-02,04:00:02,0
slapuser10,17557.0,None,0.0,16924.63,3.0,20.843995772890000 5,811831296.0,6337482752.0,3825119.0,2014-10-26,00:00:02,0
slapuser5,8479.0,None,0.0,3839.94,24.0,4.7951274804400005,806901632.0,9005555712.0,1000870.0,2014-07-21,00:15:01,0
```