Équipe Scikit-learn

Prix de l'innovation Inria — Académie des sciences — Dassault Systèmes

scikit-learn : le logiciel open-source de machine learning devenu avec l'aide d'Inria depuis 2010 une référence mondiale dans l'industrie, l'enseignement et la recherche.

Historique du projet :

- 2006 : Prototypage du projet par D. Cournapeau financé par le Google Summer of Code
- 2010 : Prise en main du projet par l'équipe Parietal au sein d'Inria
- 2013 : Recrutement d'Olivier Grisel par Inria comme ingénieur senior sur scikit-learn
- 2018 : Lancement du consortium scikit-learn avec le soutien de 7 entreprises

1 Curriculum vitae de l'équipe

Scikit-learn est développé mondialement par une grande équipe de volontaires. Il a été initié et est dirigé en France par une équipe à Inria.

Gaël Varoquaux — CR Inria

Ancien élève de l'École Normale Supérieur Paris, Docteur en physique quantique (2008), Gaël travaille sur l'utilisation de l'apprentissage statistique pour comprendre le fonctionnement cérébral depuis 2008 au sein de l'équipe Inria Parietal. Il a dirigé la naissance du project scikit-learn au sein de cette équipe en 2010 et contribué au projet et à sa gestion depuis. Il dirige le consortium scikit-learn, entre la fondation Inria et des industriels. Il mène aussi une recherche scientifique en application de l'apprentissage statistique active, avec un facteur h de 38 (source).

Olivier Grisel — Ingénieur Inria

Ingénieur diplomé de l'École Nationale Supérieure de Techniques Avancées ParisTech (2003) et du MSc in Advanced Computing de l'Imperial College de Londres (2003), Olivier contribue au développement du projet scikit-learn depuis 2010 à titre personnel dans un premier temps puis en tant que mainteneur au sein de l'équipe Inria Parietal depuis 2013. Depuis 2018, Olivier officie en tant que directeur technique du consortium scikit-learn de la fondation Inria.

Alexandre Gramfort — DR Inria

Ancien élève de l'Ecole Polytechnique, diplômé du master MVA de l'ENS Paris-Saclay, Docteur en Informatique (2009), Enseignant Chercheur à Télécom ParisTech (2012-2017), Alexandre travaille au développement d'algorithmes d'optimisation, de traitement du signal et d'apprentissage statistique pour l'étude des données en neurosciences. Alors post-doc au sein de Parietal, il a contribué au design original et à l'optimisation des algorithmes de scikit-learn depuis sa création en 2010. Depuis son retour chez Inria en 2017, porteur d'une ERC Starting grant, Alexandre continue ses recherches au plus haut niveau avec un facteur h de 32 (source).

Betrand Thirion — DR Inria

Bertrand Thirion est le responsable de l'équipe Parietal, au sein du centre de recherche Inria Saclay Île de France. Il travaille sur les statistiques et le machine learning pour l'imagerie cérébrale. Il contribue aussi bien des algorithmes que du logiciel, avec une attention particulière aux applications à la neuroimagerie fonctionnelle. Il travaille au sein du centre d'imagerie cérébrale Neurospin. Bertrand dirige aussi l'institut de convergence DATAIA, qui unit IA et science des données à Saclay, le plus gros campus scientifique Français. Il a un facteur h de 41 (source).

Loïc Estève — Ingénieur de recherche Inria

Diplômé d'un magistère de Physique de l'École Normale Supérieure de Paris (2005) et d'un doctorat en Physique des Particules (2008). Après une incursion de quelques années dans le monde de la finance, Loïc a fait le choix de revenir dans le milieu académique pour travailler sur le développement de logiciels open-source au sein de l'équipe Inria Parietal. Il a contribué à scikit-learn depuis 2014 et est ingénieur de recherche au sein du Service de Développement et d'Expérimentation d'Inria Paris depuis 2017.

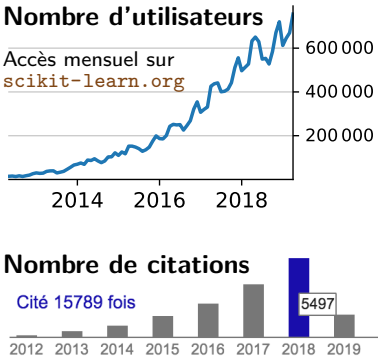
2 Principaux résultats

À l'échelle mondiale, scikit-learn est le 1er logiciel open source de machine learning piloté par une communauté de la recherche. Il rivalise en popularité avec les outils développés par les GAFA.

La vision scikit-learn : scikit-learn a été développé par l'équipe Inria Parietal depuis 2010 avec la vision de donner accès à l'apprentissage statistique au plus grand nombre, en commençant par les chercheurs en neurosciences. En fournissant un outil efficace, simple à prendre en main et très bien documenté avec des centaines d'exemples, les développeurs de scikit-learn ont contribué à démocratiser l'apprentissage statistique à la base de la révolution actuelle de l'intelligence artificielle. Avec un impact bien plus vaste que les neurosciences, les chercheurs et ingénieurs Inria à l'origine du succès de scikit-learn ont permis l'utilisation de l'apprentissage statistique dans toutes les sciences expérimentales de la chimie, la biologie ou la physique, ainsi que pour beaucoup d'applications industrielles.

Scikit-learn : une référence de l'apprentissage statistique Scikit-learn rassemble plus de 180 modèles d'apprentissage statistique différents. Il touche à beaucoup d'aspects de cette discipline des mathématiques appliquées et fournit l'ensemble des outils algorithmiques de références, tels qu'on peut les trouver dans un livre sur le sujet. Sa documentation –`http://scikit-learn.org`– forme à elle-même une introduction à l'apprentissage statistique. Elle est réputée très pédagogique et ferait plus d'un millier de pages si elle était sur papier.

Métriques d'utilisations Comme scikit-learn est un logiciel libre, il est difficile d'avoir des chiffres exacts de son nombre d'utilisateurs. Cependant les statistiques du site web revèlent plus de 42 millions de visites en 2018 et 700 000 utilisateurs mensuels actifs (figure sur la droite). GitHub qui héberge le code source du projet reporte prêt de 17,000 forks et 35,000 stars (source). scikit-learn représente 39 années de travail pour un seul homme (source). C'est le troisième logiciel libre de machine learning le plus populaire, derrière deux logiciels développés par Google (source). Un sondage réalisé il y a quelques années a compté 63% d'utilisateurs dans l'industrie, et 34% dans le milieu académique. Le papier académique de référence [2] est cité 16 000 fois sur Google scholar depuis 2012 avec 5500 citations rien qu'en 2018 (figure sur la droite).



Examples d'utilisations industrielles De nombreux acteurs privés utilisent scikit-learn dans leurs produits et leurs équipes de data scientists. Beaucoup ne communiquent pas dessus, mais nous avons cependant réuni des témoignages. Nous en listons quelques uns ci-dessous, pour illustrer la diversité et l'importance des cas d'applications, allant de la banque à la recommandation de musique.

J.P.Morgan : "Scikit-learn is an indispensable part of the Python machine learning toolkit at JPMorgan. It is very widely used across all parts of the bank for classification, predictive analytics, and very many other machine learning tasks. Its straightforward API, its breadth of algorithms, and the quality of its documentation combine to make scikit-learn simultaneously very approachable and very powerful." *Stephen Simmons, VP, Athena Research, JPMorgan*

Spotify : "Scikit-learn provides a toolbox with solid implementations of a bunch of state-of-the-art models and makes it easy to plug them into existing applications. We've been using it quite a lot for music recommendations at Spotify and I think it's the most well-designed ML package I've seen so far." *Erik Bernhardsson, Engineering Manager Music Discovery & Machine Learning, Spotify*

booking.com : "At Booking.com, we use machine learning algorithms for many different applications, such as recommending hotels and destinations to our customers, detecting fraudulent reservations, or scheduling our customer service agents. Scikit-learn is one of the tools we use when implementing standard algorithms for prediction tasks. Its API and documentations are excellent and make it easy to use. The scikit-learn developers do a great job of incorporating state of the art implementations and new algorithms into the package. Thus, scikit-learn provides convenient

access to a wide spectrum of algorithms, and allows us to readily find the right tool for the right job." *Melanie Mueller, Data Scientist*

Best of media group : "Scikit-learn is our #1 toolkit for all things machine learning at Bestofmedia. We use it for a variety of tasks (e.g. spam fighting, ad click prediction, various ranking models) thanks to the varied, state-of-the-art algorithm implementations packaged into it. In the lab it accelerates prototyping of complex pipelines. In production I can say it has proven to be robust and efficient enough to be deployed for business critical components.' *Eustache Diemert, Lead Scientist Bestofmedia Group*

Data Robot : "DataRobot is building next generation predictive analytics software to make data scientists more productive, and scikit-learn is an integral part of our system. The variety of machine learning techniques in combination with the solid implementations that scikit-learn offers makes it a one-stop-shopping library for machine learning in Python. Moreover, its consistent API, well-tested code and permissive licensing allow us to use it in a production environment. Scikit-learn has literally saved us years of work we would have had to do ourselves to bring our product to market." *Jeremy Achin, CEO & Co-founder DataRobot Inc.*

Dataiku : "Our software, Data Science Studio (DSS), enables users to create data services that combine ETL with Machine Learning. Our Machine Learning module integrates many scikit-learn algorithms. The scikit-learn library is a perfect integration with DSS because it offers algorithms for virtually all business cases. Our goal is to offer a transparent and flexible tool that makes it easier to optimize time consuming aspects of building a data service, preparing data, and training machine learning algorithms on all types of data." *Florian Douetteau, CEO, Dataiku*

Plus de témoignages sur `https://scikit-learn.org/stable/testimonials/testimonials.html`

Consortium scikit-learn Le consortium scikit-learn hébergé par la fondation Inria est né en septembre 2018 avec le soutien de 7 entreprises : Microsoft, BCG, AXA, BNP Paribas-Cardif, Intel, NVIDIA, et Dataiku. Ce partenariat/mécénat démontre l'impact industriel de scikit-learn et il permettra le financement à long terme du logiciel. À l'heure actuelle, le montant du mécénat est de 400 k€ annuel. Les priorités du consortium sont établies après consultation des mécènes et de la communauté de développeurs. Elles se focalisent sur l'amélioration de la bibliothèque et de sa documentation.

L'écosystème et la communauté scikit-learn scikit-learn est un logiciel basé sur l'écosystème scientifique open source Python et principalement les librairies Numpy, Scipy et Cython. Le succès rapide de scikit-learn a donc été rendu possible par des centaines de contributeurs open source. Le logiciel scikit-learn a lui même été écrit par prêt de 1300 personnes (source) avec des contributions majeures en France à Inria, Télécom ParisTech et à l'international avec Andreas Mueller à Columbia University et plus récemment Joël Nothman à l'université de Sydney.

3 Publications

[1] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, et al. API design for machine learning software : experiences from the scikit-learn project. *arXiv preprint arXiv :1309.0238*, 2013. **426 citations**.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn : Machine learning in Python. *Journal of machine learning research*, 12(Oct) :2825–2830, 2011. **15 789 citations**.