

# Open Source for (Data) Science

Andreas Müller

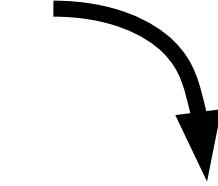
Associate Research Scientist  
Columbia University  
Scikit-learn Technical Committee



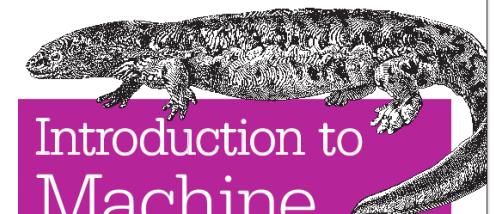
Alfred P. Sloan  
FOUNDATION



COLUMBIA UNIVERSITY  
IN THE CITY OF NEW YORK



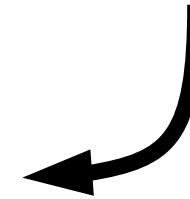
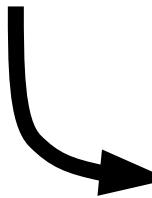
O'REILLY®



Introduction to  
Machine  
Learning  
with Python

A GUIDE FOR DATA SCIENTISTS

Andreas C. Müller & Sarah Guido



# Why Software Matters

“An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship.

The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures.”

- Buckheit and Donoho (1995)

# Python

1990s: The Scripting Era

*"Python as Alternative to Bash"*

2000s: The SciPy Era

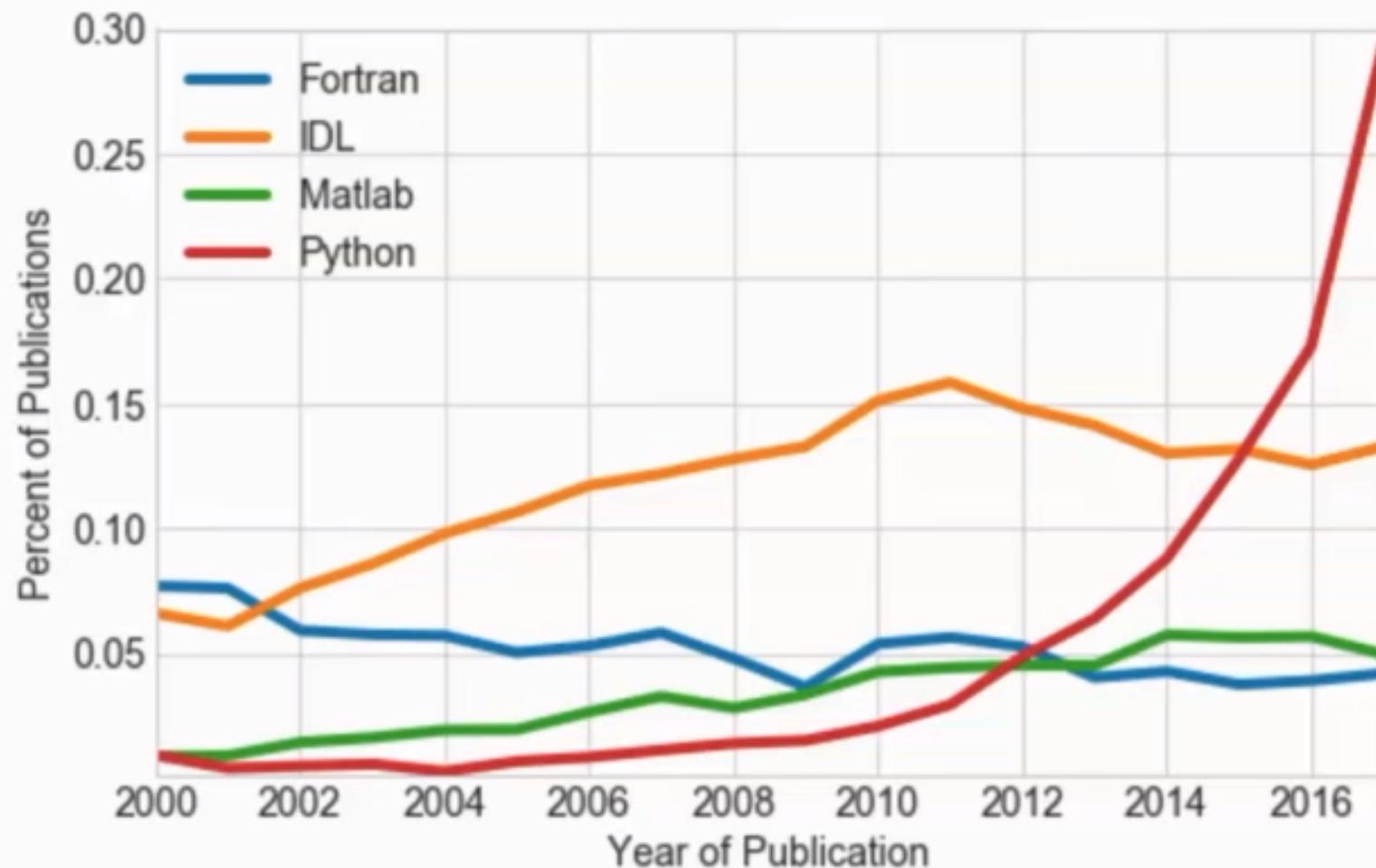
*"Python as Alternative to MatLab"*

2010s: The PyData Era

*"Python as Alternative to R"*

2020s: ???

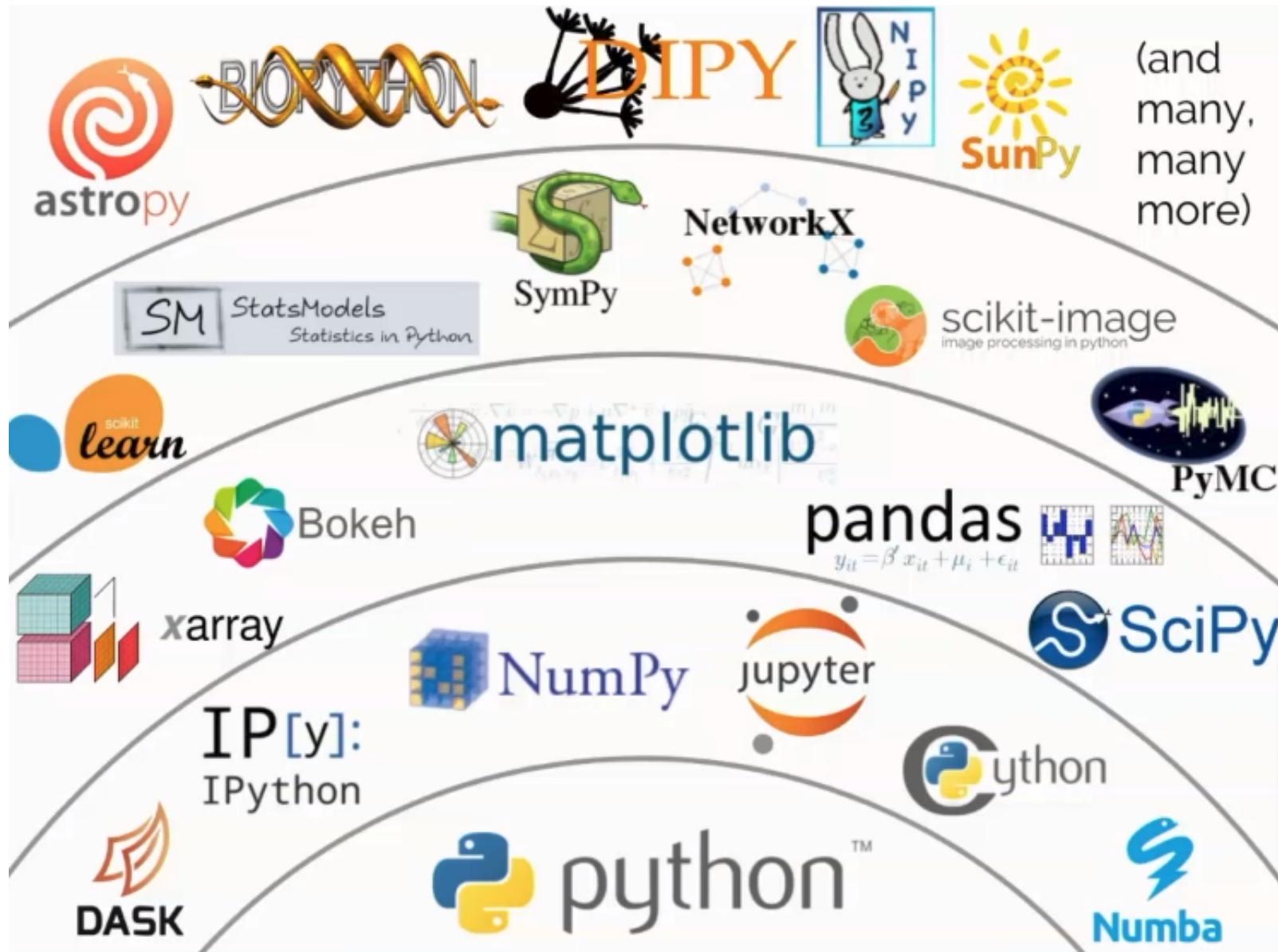
# Mentions of Software in Astronomy Publications:



Compiled from NASA ADS ([code](#)).

Thanks to Juan Nunez-Iglesias,  
Thomas P. Robitaille, and Chris Beaumont.

# The Scientific Python Ecosystem



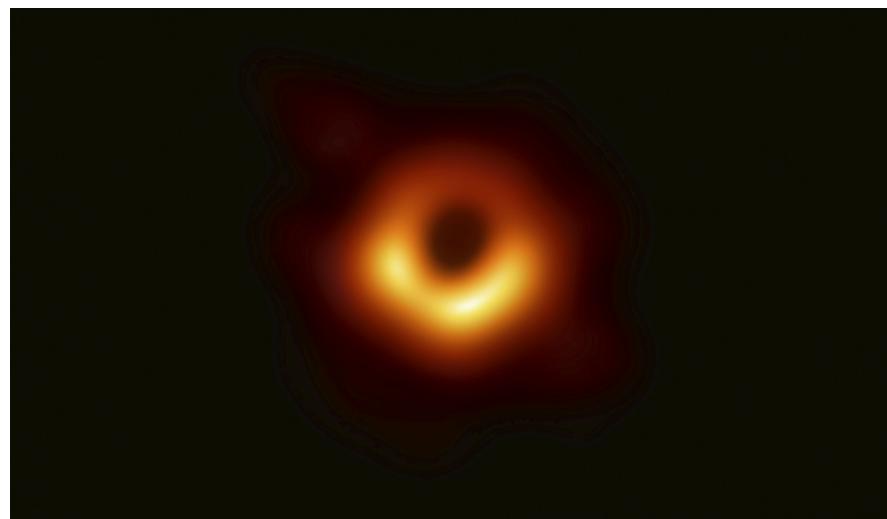
# Impact

“The scientific Python ecosystem is critical infrastructure for the research done at LIGO.”

David Shoemaker, LIGO Scientific Collaboration

These [SciPy/PyData] are fundamental to training biologists to have the computational skills they need to be effective in today's research realm.

Bonnie Hurwitz, Hurwitz lab, iPlant



# Transforming Science Through Cyberinfrastructure

*NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century*

## Executive Summary

Twenty-first century science and engineering (S&E) research is being transformed by the increasing availability and scales of computation and data. The national cyberinfrastructure (CI) ecosystem has thus become a key catalyst for discovery and innovation and now plays a critical role in ensuring US leadership in S&E, economic competitiveness and national security, consistent with NSF's mission. The vision and blueprint presented in this document have been developed by the NSF Office of Advanced Cyberinfrastructure (OAC) on behalf of NSF based on a synthesis of multiple community inputs through advisory bodies, requests for information (RFIs), workshops and conferences, and national initiatives.

A new vision. NSF envisions ***an agile, integrated, robust, trustworthy and sustainable CI ecosystem that drives new thinking and transformative discoveries in all areas of S&E research and education.*** This vision embodies the following overarching principles:

# Transforming Science Through Cyberinfrastructure

*NSF's Blueprint for a National Cyberinfrastructure Ecosystem for  
Science and Engineering in the 21st Century*

## - Python

### Executive Summary

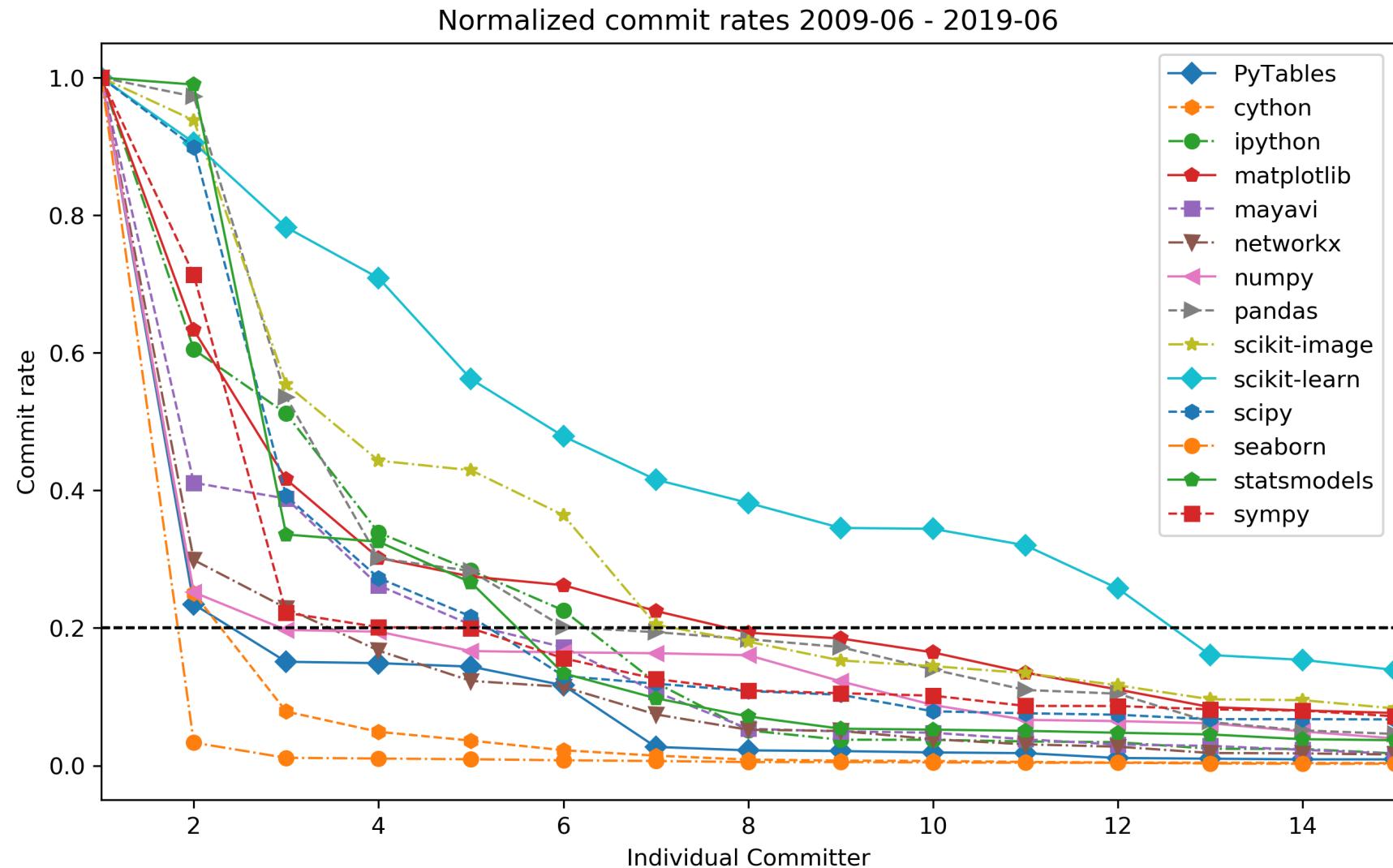
Twenty-first century science and engineering (S&E) research is being transformed by the increasing availability and accessibility of computation and data. The national cyberinfrastructure (CI) ecosystem has thus become a key catalyst for discovery and innovation and now plays a critical role in ensuring US leadership in S&E, economic competitiveness and national security, consistent with NSF's mission. The vision and blueprint presented in this document have been developed by the NSF Office of Advanced Cyberinfrastructure (OAC) on behalf of NSF based on a synthesis of multiple community inputs through advisory bodies, requests for information (RFIs), workshops and conferences, individual initiatives,

## - R

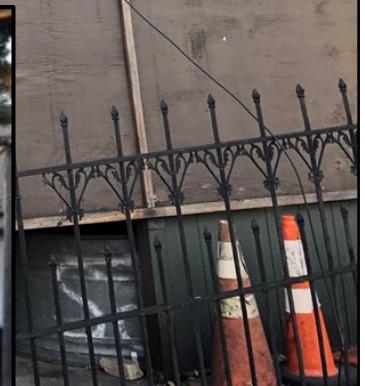
## - Open Source

A new vision. NSF envisions ***an agile, integrated, robust, trustworthy and sustainable CI ecosystem that drives new thinking and transformative discoveries in all areas of S&E research and education.*** This vision embodies the following overarching principles:

# Evolution and Support



# Scholz Factor



1111



# Machine Learning with Python and Scikit-learn

## Authors

The following people are currently core contributors to scikit-learn's development and maintenance:



Joris Van den  
Bossche



Loïc Estève



Thomas J Fan



Alexandre  
Gramfort



Olivier Grisel



Yaroslav  
Halchenko



Nicolas Hug



Adrin Jalali



Guillaume  
Lemaître



Jan Hendrik  
Metzen



Andreas Mueller



Vlad Niculae



Joel Nothman



Hanmin Qin



Bertrand Thirion  
Tom Dupré la  
Tour



Nelle Varoquaux



Gael Varoquaux



Roman Yurchak

## Emeritus Core Developers

Alexander Fabisch  
Alexandre Passos  
Angel Soler Gollonet  
Arnaud Joly  
Chris Gorgolewski  
David Cournapeau  
David Warde-Farley  
Eduard Duchesnay  
Fabian Pedragosa  
Gilles Louppe  
Jacob Schreiber  
Jake Vanderplas  
Jaques Grobler  
Jarrod Millman  
Kyle Kastner  
Lars Buitinck  
Manoj Kumar  
Mathieu Blondel  
Matthieu Brucher  
Noel Dawe  
Paolo Losi  
Peter Prettenhofer  
Raghav Rajagopalan  
Robert Layton  
Ron Weiss  
Satrajit Ghosh  
Shiqiao Du  
Thouis (Ray) Jones  
Vincent Dubourg  
Vincent Michel  
Virgile Fritsch  
Wei Li

1,323 contributors

## Authors

The following people are currently core contributors to scikit-learn's development and maintenance:



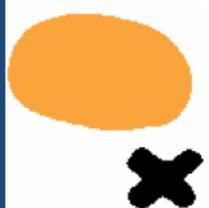
Joris Van den  
Bossche



Loïc Estève



Thomas J Fan



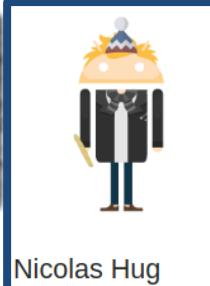
Alexandre  
Gramfort



Olivier Grisel



Yaroslav  
Halchenko



Nicolas Hug



Adrin Jalali



Guillaume  
Lemaître



Jan Hendrik  
Metzen



Andreas Mueller



Vlad Niculae



Joel Nothman



Hanmin Qin



Bertrand Thirion



Tom Dupré la  
Tour



Nelle Varoquaux



Gael Varoquaux



Roman Yurchak

## Emeritus Core Developers

Alexander Fabisch  
Alexandre Passos  
Angel Soler Gollonet  
Arnaud Joly  
Chris Gorgolewski  
David Cournapeau  
David Warde-Farley  
Eduard Duchesnay  
Fabian Pedragosa  
Gilles Louppe  
Jacob Schreiber  
Jake Vanderplas  
Jaques Grobler  
Jarrod Millman  
Kyle Kastner  
Lars Buitinck  
Manoj Kumar  
Mathieu Blondel  
Matthieu Brucher  
Noel Dawe  
Paolo Losi  
Peter Prettenhofer  
Raghav Rajagopalan  
Robert Layton  
Ron Weiss  
Satrajit Ghosh  
Shiqiao Du  
Thouis (Ray) Jones  
Vincent Dubourg  
Vincent Michel  
Virgile Fritsch  
Wei Li

 **1,323** contributors

Our Mission:

Commoditize and Democratize Machine Learning

## Scikit-learn: Machine learning in Python

17173

2011

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, ...

Journal of machine learning research 12 (Oct), 2825-2830

About 100 citations / day

About 4 citations since this meeting started!

scikit-learn / scikit-learn ✓

Used by ▾

58,454



François Chollet ✅

@fchollet

Following

Replying to @ylecun

The only thing Keras took from Torch7 is the name of the Sequential class. Checking out early Keras versions demonstrates they have virtually nothing in common. The promo line "in the spirit of Torch" was only meant as "it's simple/minimalist". Sklearn was a far bigger influence.

1616

# Activities

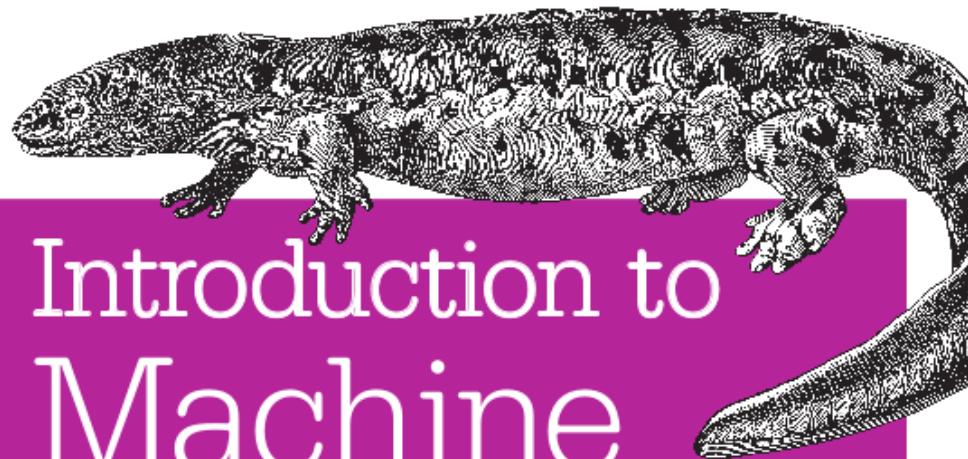
- Extreme Gradient Boosting (Nicolas Hug)
- Preprocessing, usability
- Maintenance, bug fixes, user support

# Future Work

- Model understanding and inspection
- Visualization
- Automatic Machine Learning (AutoML) & dabl

# Changing Scientific Software Funding?

O'REILLY®



# Introduction to Machine Learning with Python

A GUIDE FOR DATA SCIENTISTS

Andreas C. Müller & Sarah Guido



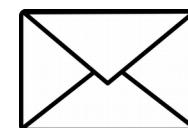
[amueller.github.io](http://amueller.github.io)



@amuellerml



@amueller



andreas.mueller  
@columbia.com