

Engineering Open Machine Learning Software



Andreas Mueller
(NYU Center for Data Science, scikit-learn)



Mission:
Commoditize and Democratize
Machine Learning

Simple things should be simple,
complex things should be possible.

Alan Kay





Achievements



[Unwatch](#) ▾

1,072

[Unstar](#)

9,796

[Fork](#)

5,750

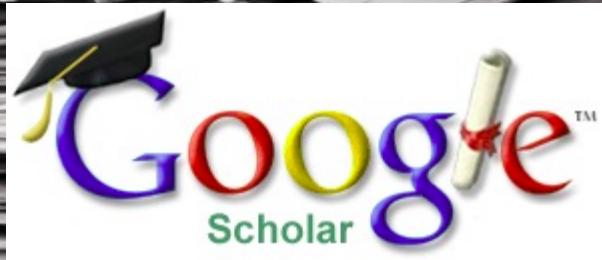


Downloads (All Versions):

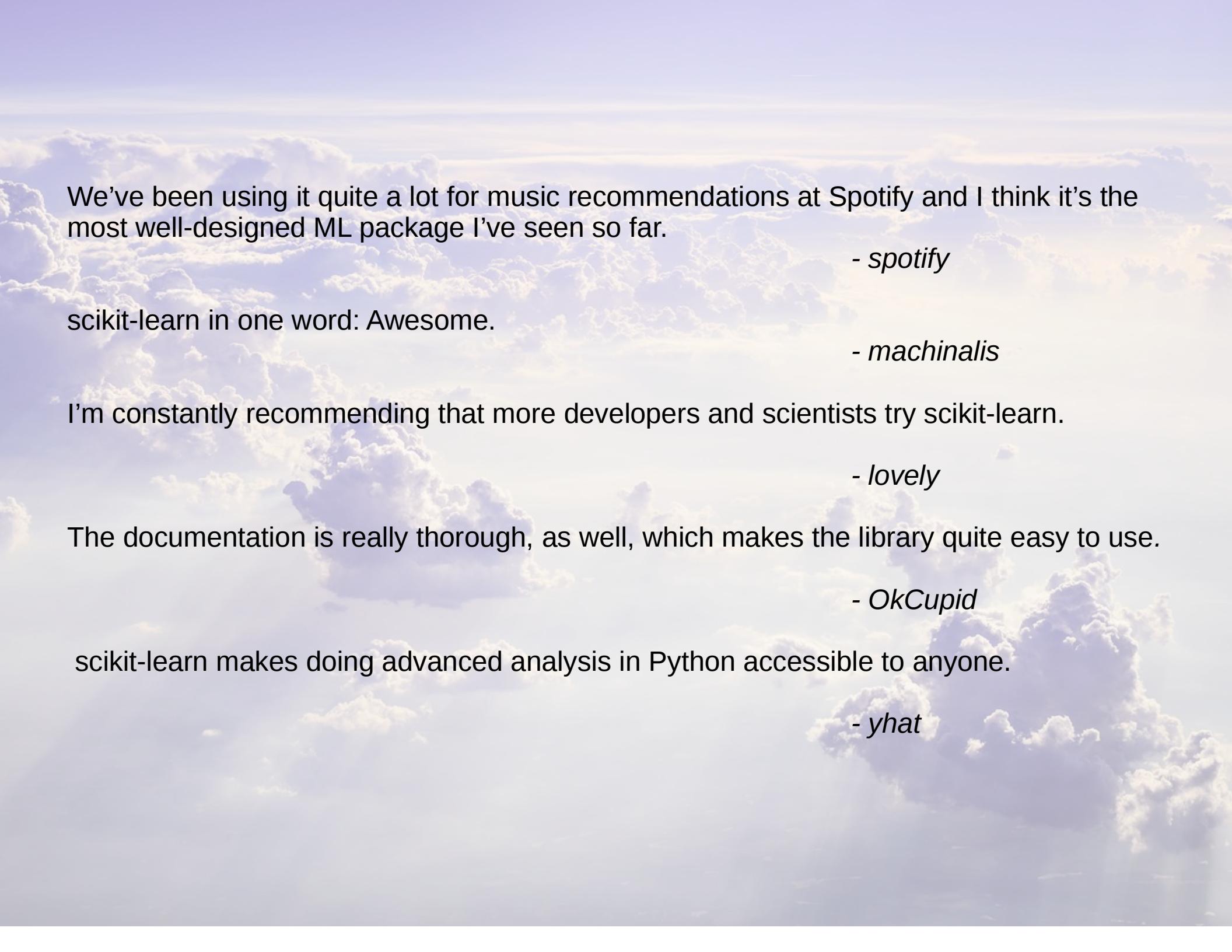
13745 downloads in the last day

95762 downloads in the last week

399200 downloads in the last month



Cited by 2499



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.

- *spotify*

scikit-learn in one word: Awesome.

- *machinalis*

I'm constantly recommending that more developers and scientists try scikit-learn.

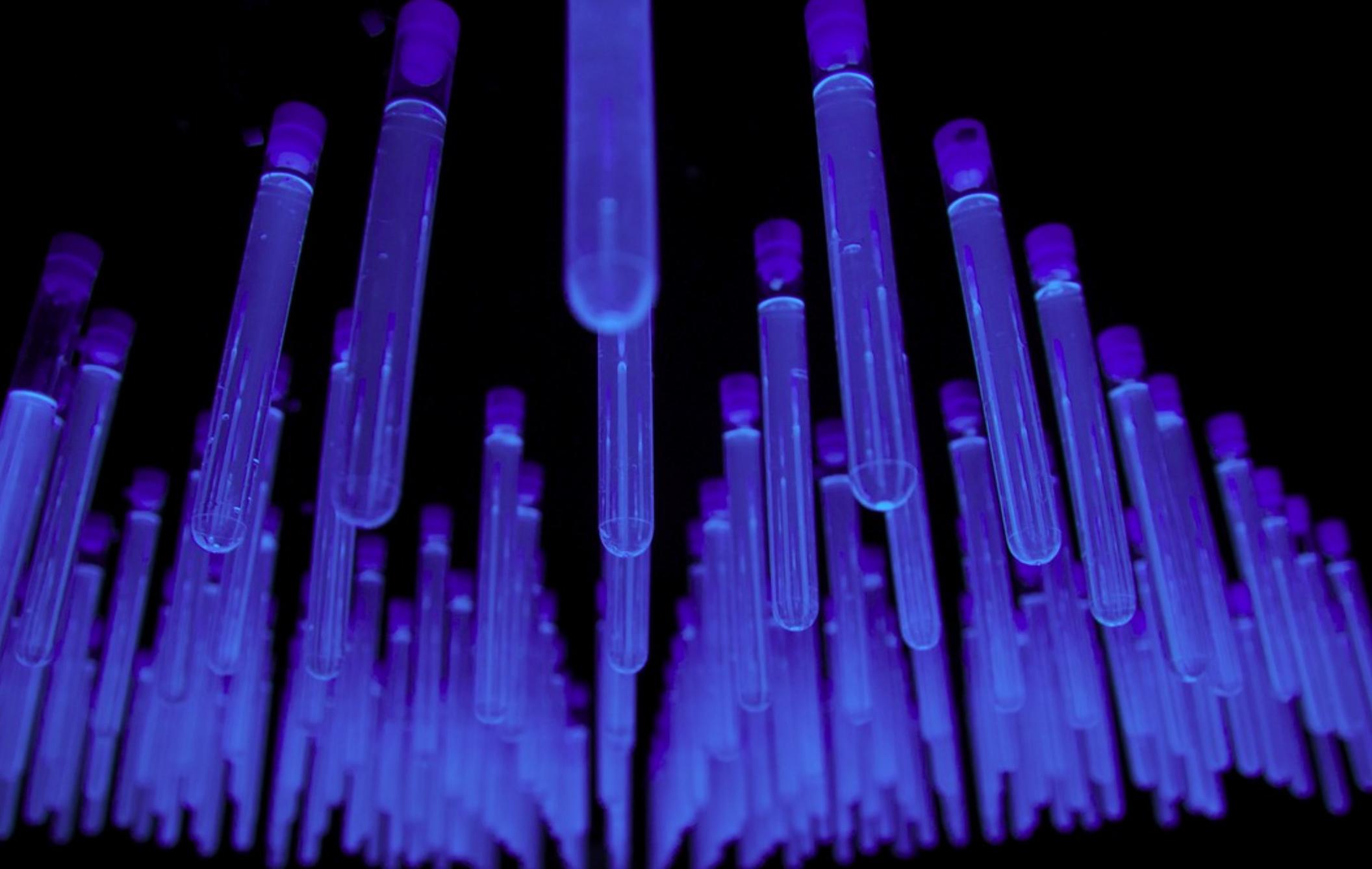
- *lovely*

The documentation is really thorough, as well, which makes the library quite easy to use.

- *OkCupid*

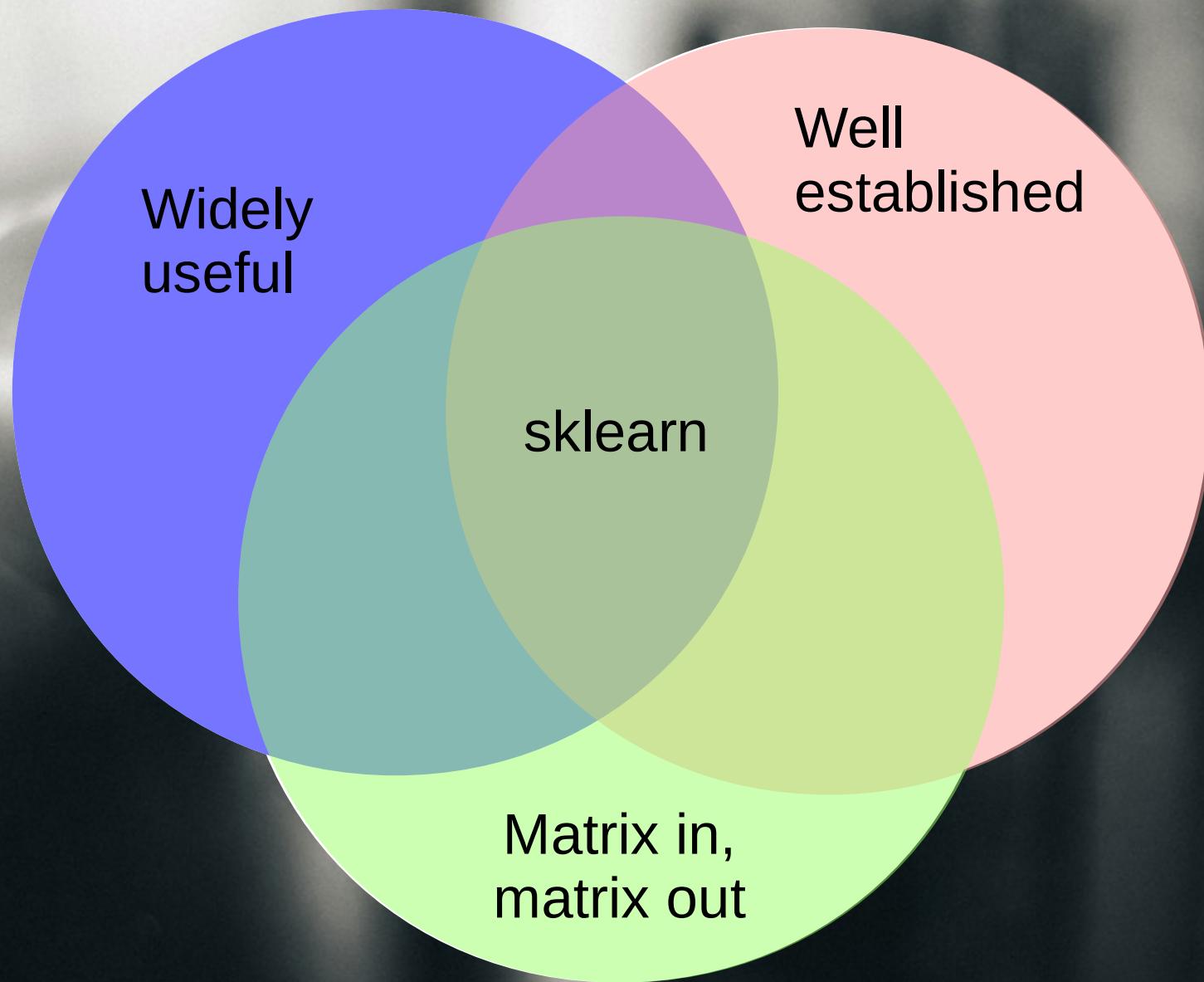
scikit-learn makes doing advanced analysis in Python accessible to anyone.

- *yhat*



Methods

Scoping



Simplicity

```
est = Est()  
est.fit(X_train, y_train)  
est.score(X_test, y_test)
```

```
grid = GridSearchCV(svm, param_grid)  
grid.fit(X_train, y_train)  
grid.score(X_test, y_test)
```

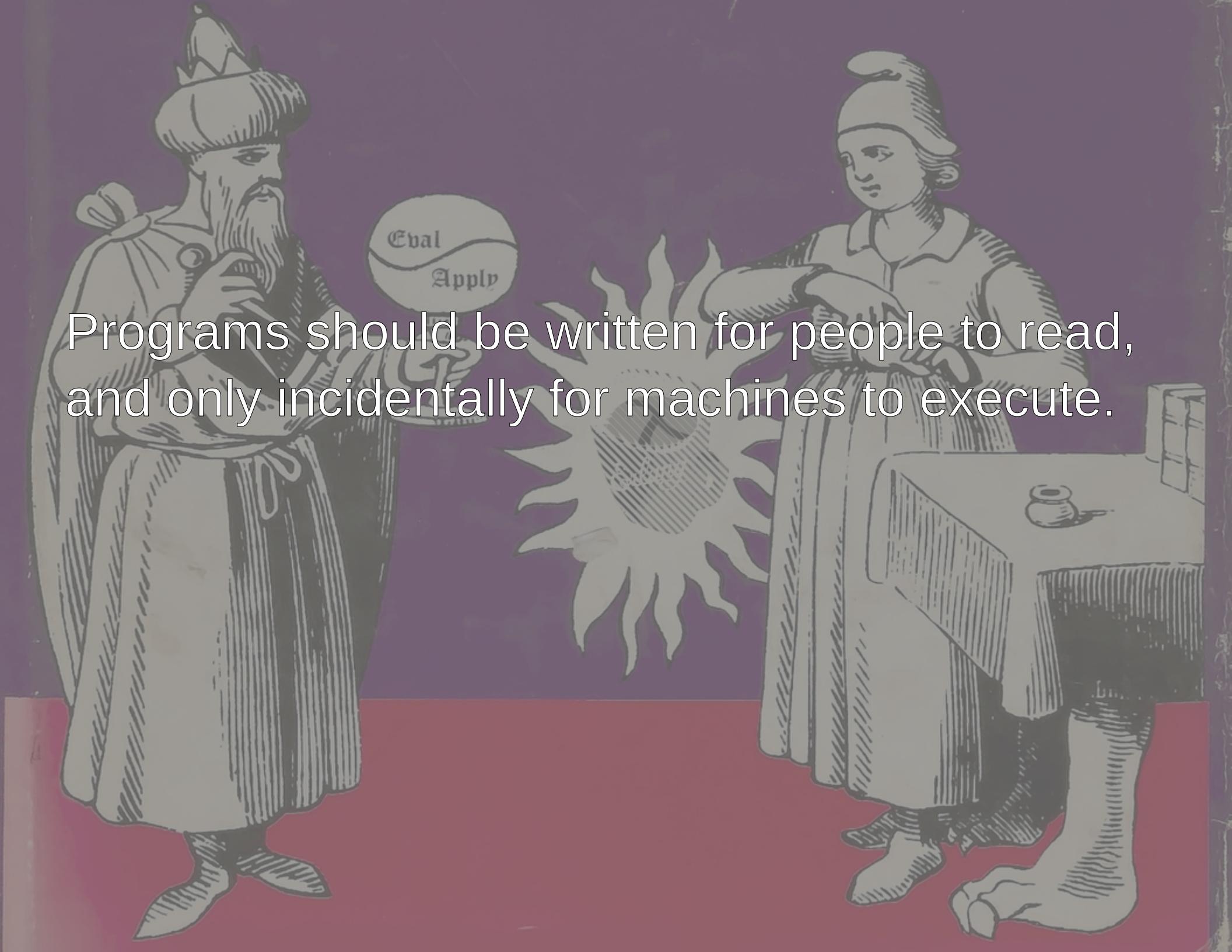
Consistency

Flat Class Hierarchy, Few Types

- Numpy arrays / sparse matrices
- Estimators
- [Cross-validation objects]
- [Scorers]

The background of the image is a collection of various old, rusty tools, including wrenches, pliers, and screwdrivers, scattered on a green surface. The tools are in different states of disrepair, with some showing significant rust and others appearing more worn.

- Maintainability



Programs should be written for people to read,
and only incidentally for machines to execute.



delete

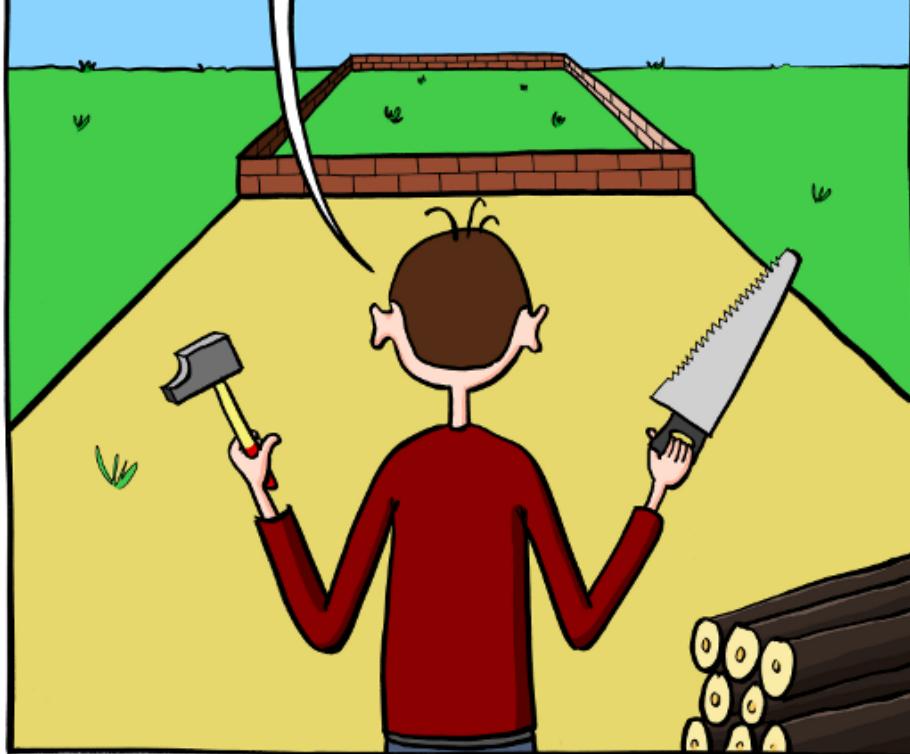
avoid code;
avoid code rot!

Challenges



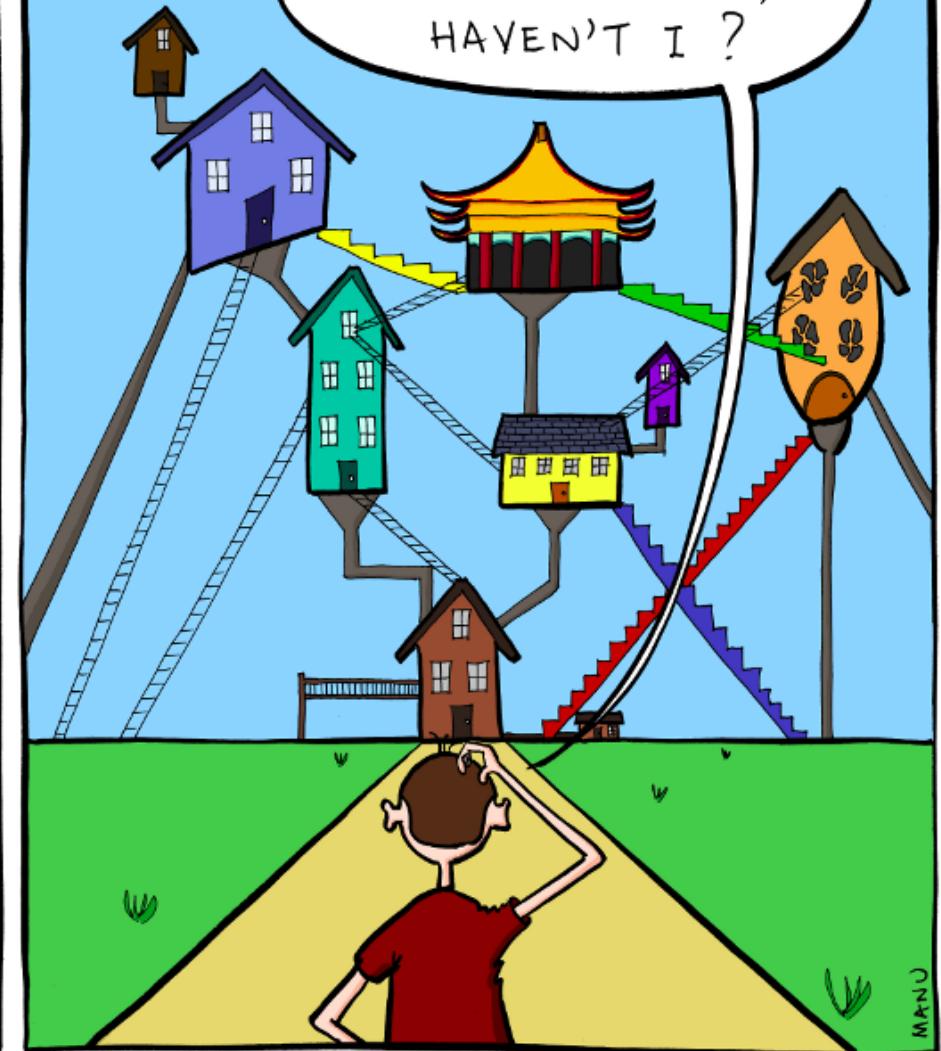
THE LIFE OF A SOFTWARE ENGINEER.

CLEAN SLATE. SOLID
FOUNDATIONS. THIS TIME
I WILL BUILD THINGS THE
RIGHT WAY.



MUCH LATER...

OH MY. I'VE
DONE IT AGAIN,
HAVEN'T I ?



Feature Creep

Two Language Problem



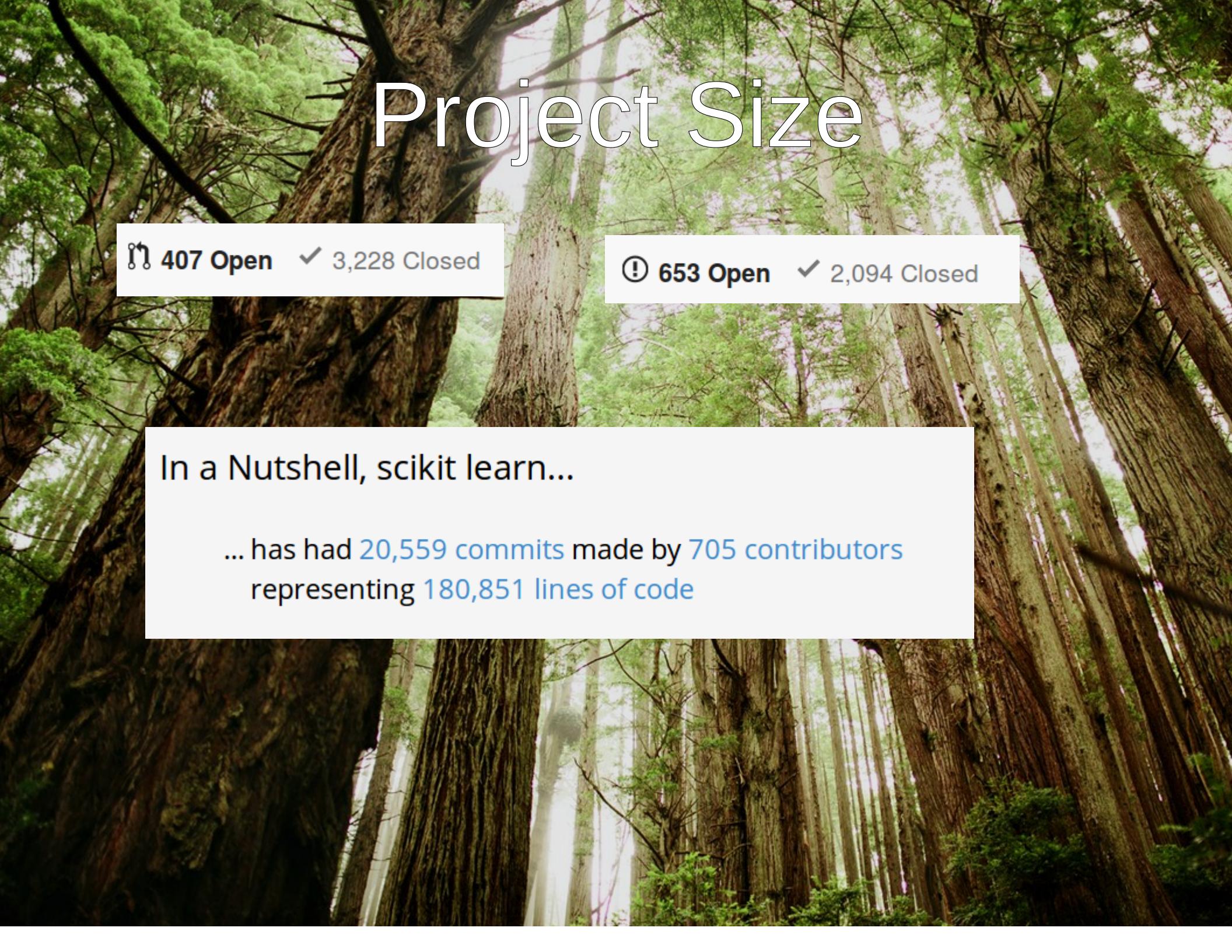
Project Size

! 407 Open ✓ 3,228 Closed

! 653 Open ✓ 2,094 Closed

In a Nutshell, scikit learn...

... has had [20,559 commits](#) made by [705 contributors](#)
representing [180,851 lines of code](#)



The background image shows a vertical wooden planks of a barn. A white-painted window frame is positioned on the right side. The window has six panes: three in the top row and three in the bottom row. The top two panes are dark, while the bottom one is lighter. The overall texture is weathered and reddish-brown.

Open
Questions

Data Structures

Categorical Variables



Better Defaults
Benchmarking ?

~~QUESTION~~
~~THE~~
~~ANSWERS~~

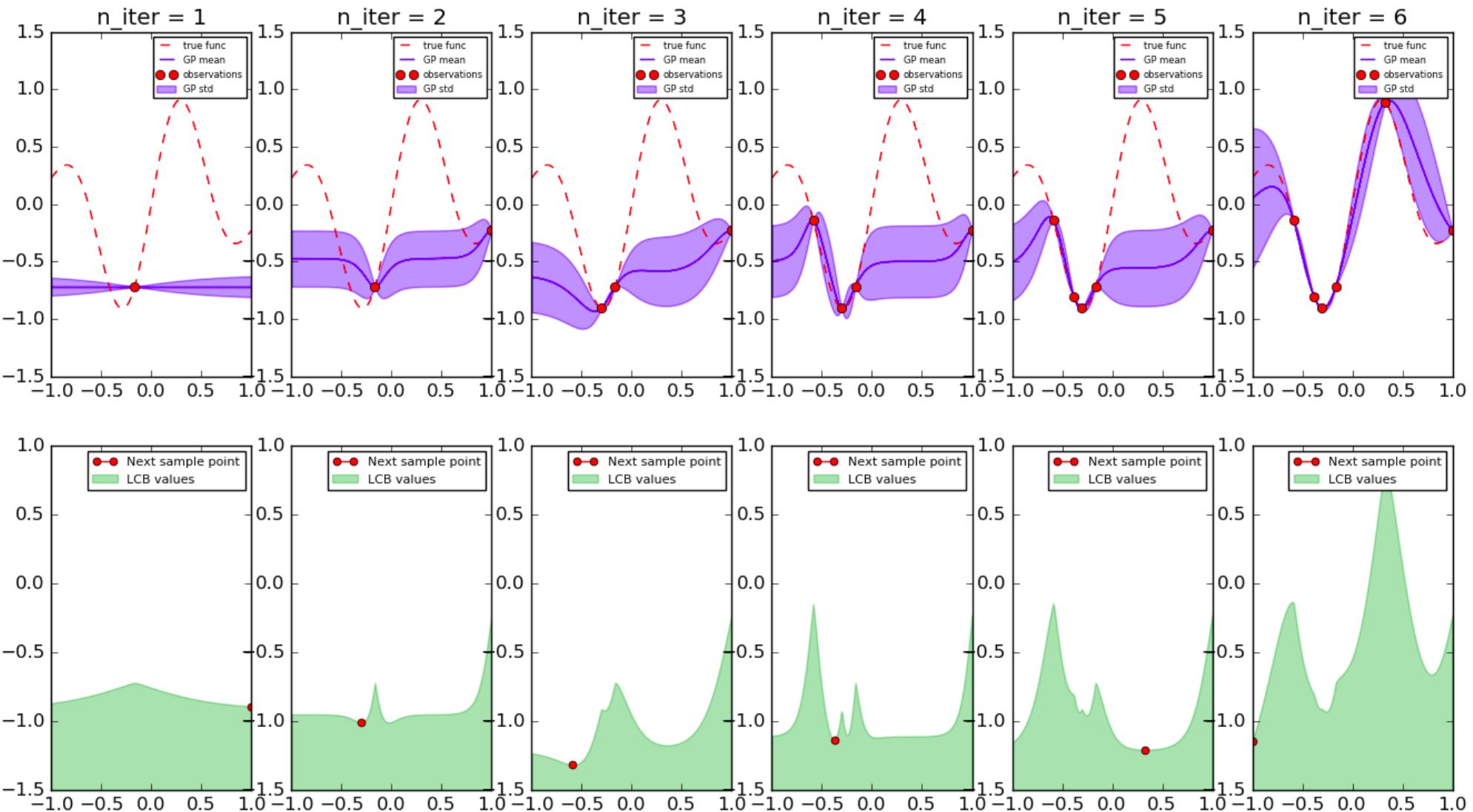
Correctness Testing

Outlook



Bayesian parameter optimization

Gaussian process based minimization

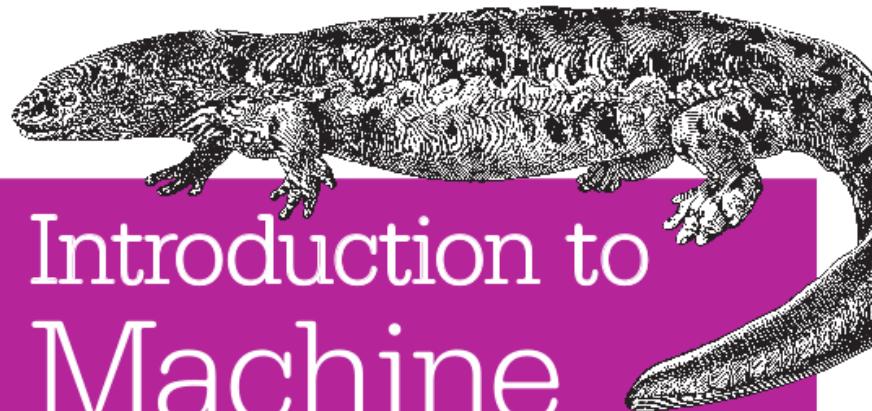


Better
Feature Name
Support



HELLO
my name is

O'REILLY®



Introduction to Machine Learning with Python

A GUIDE FOR DATA SCIENTISTS

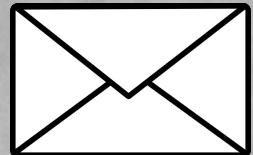
Andreas C. Müller & Sarah Guido



@amuellerml



@amueller



amueller@nyu.edu