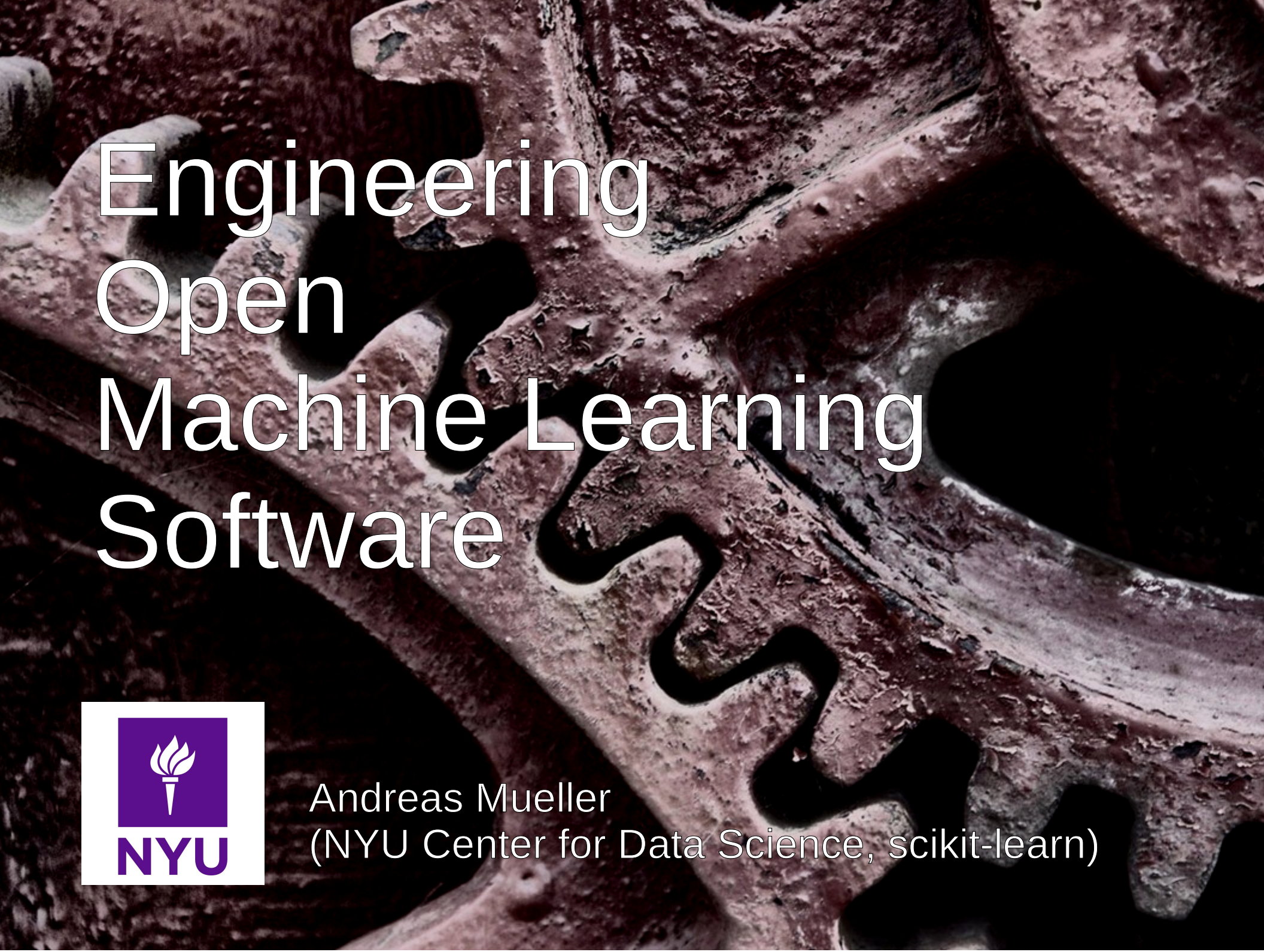
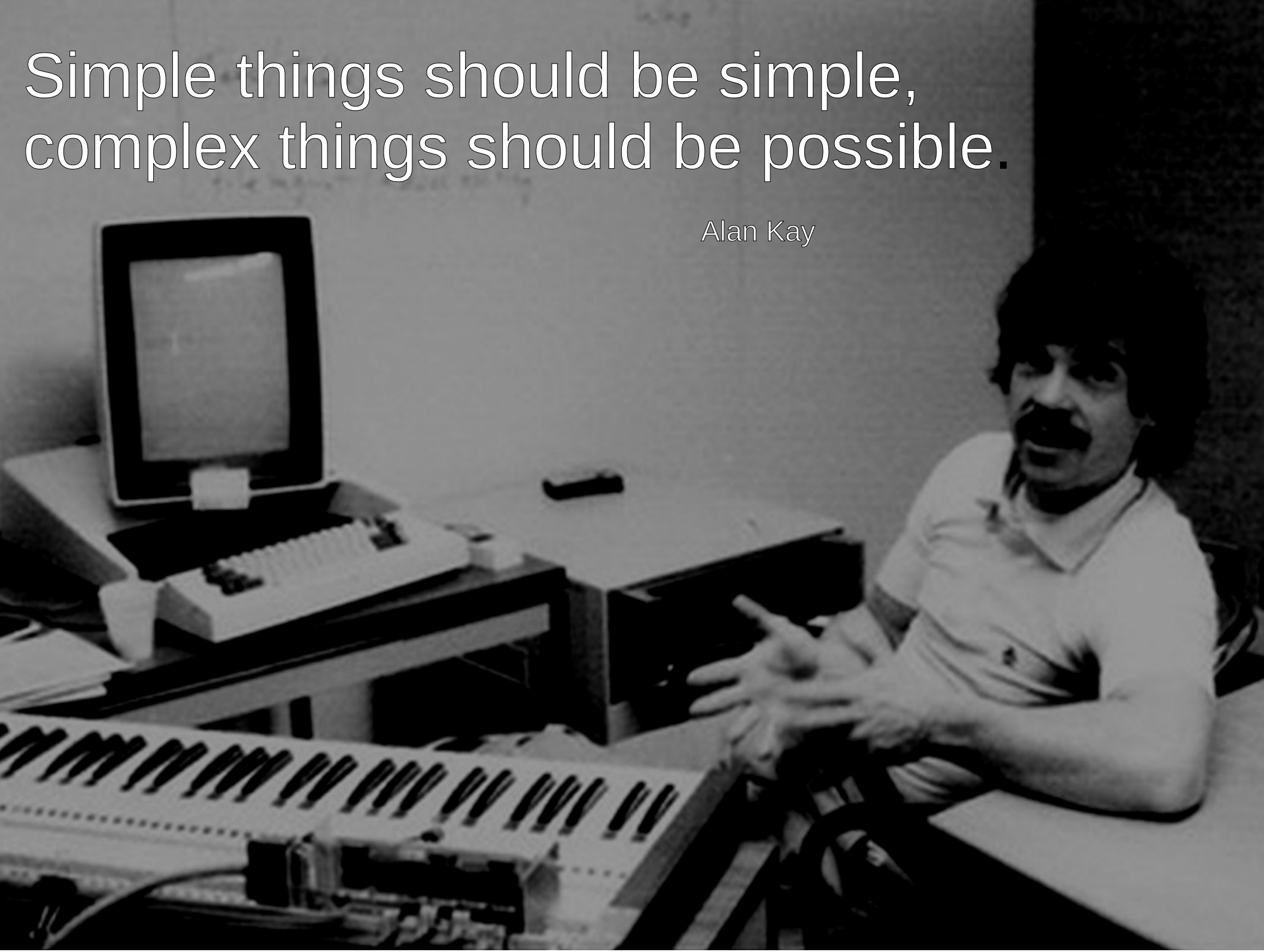


Engineering Open Machine Learning Software

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
Google Scholar
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.

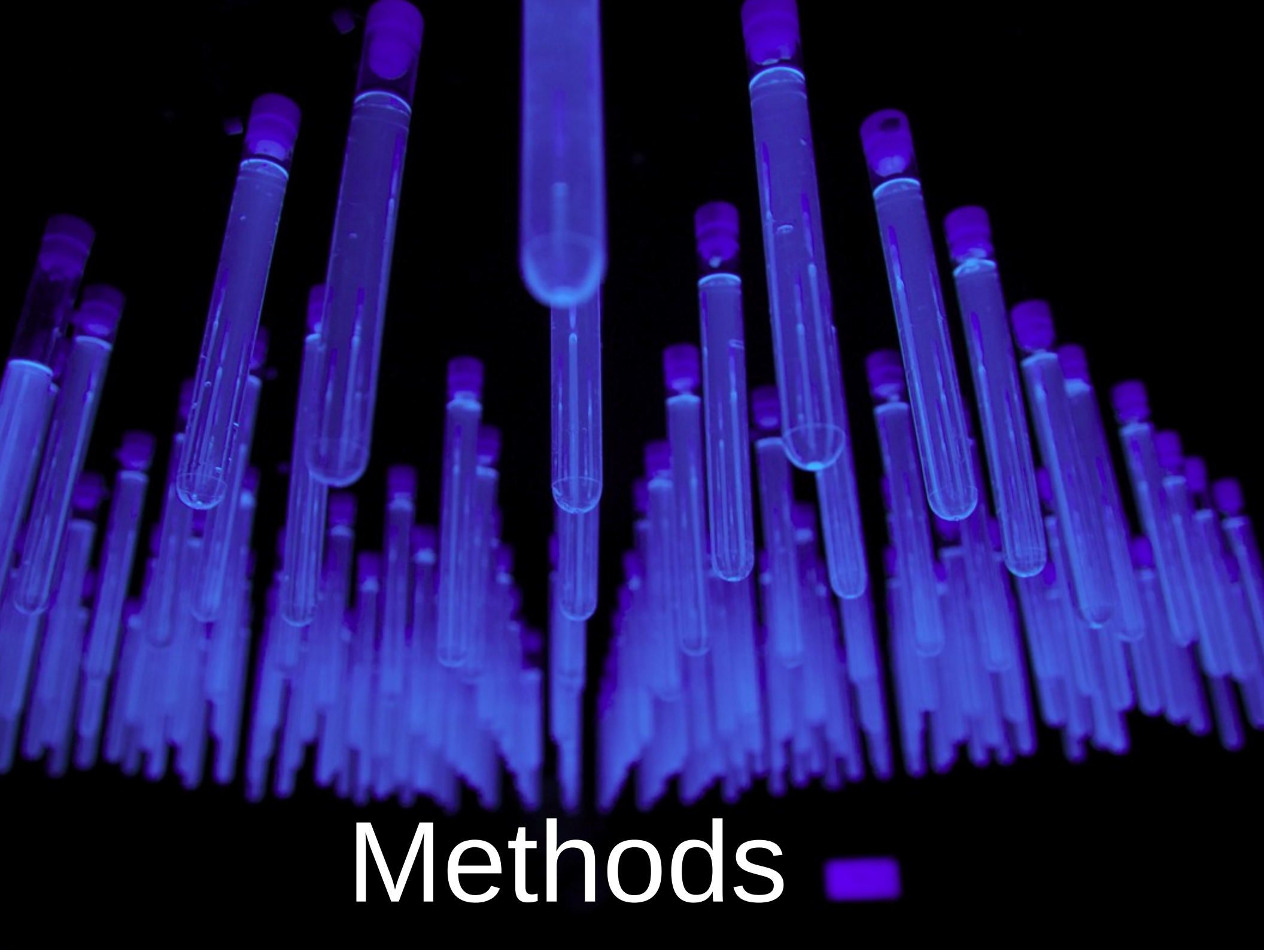
- lovely

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

- OkCupid

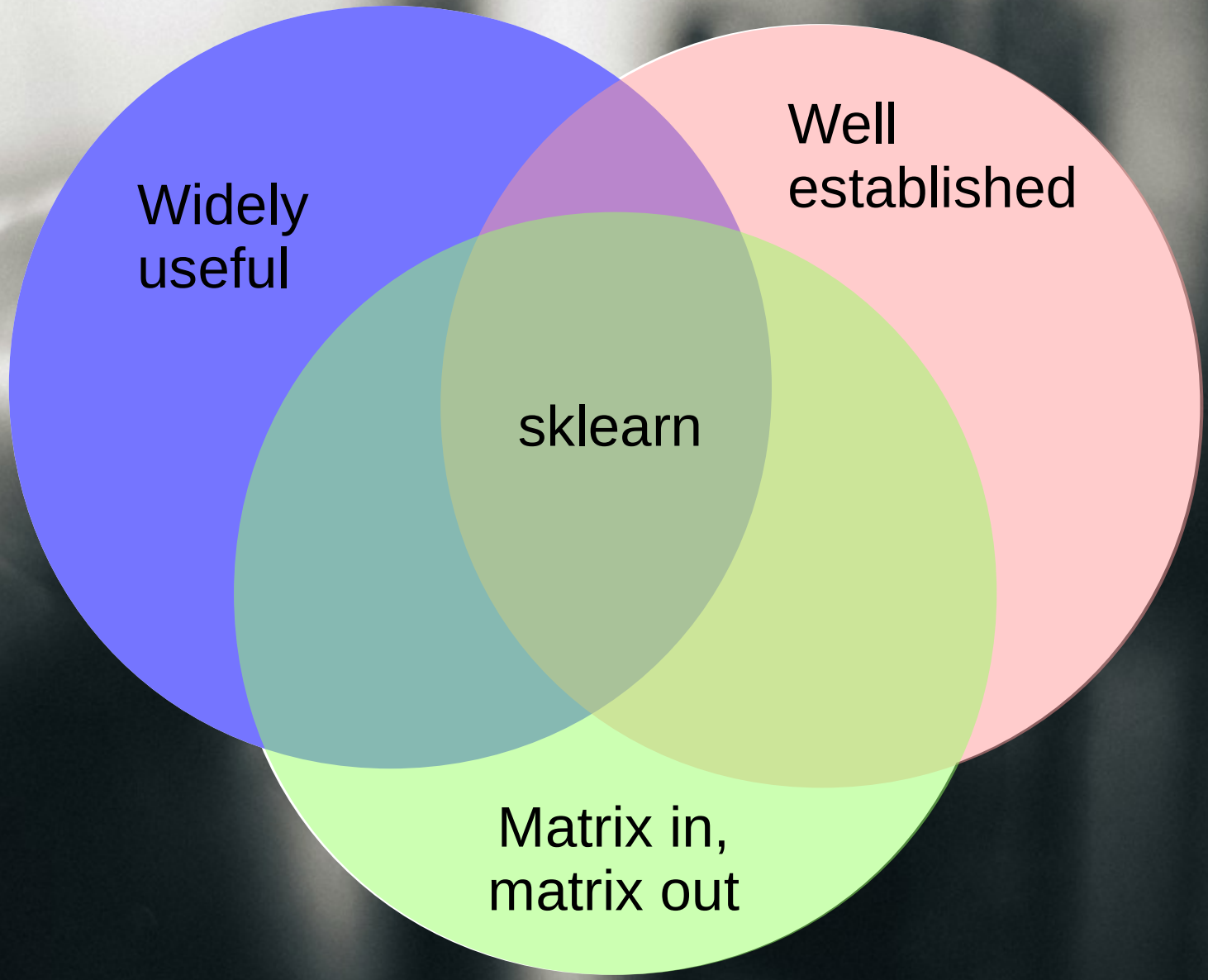
scikit-learn makes doing advanced analysis in Python accessible to anyone.

- yhat

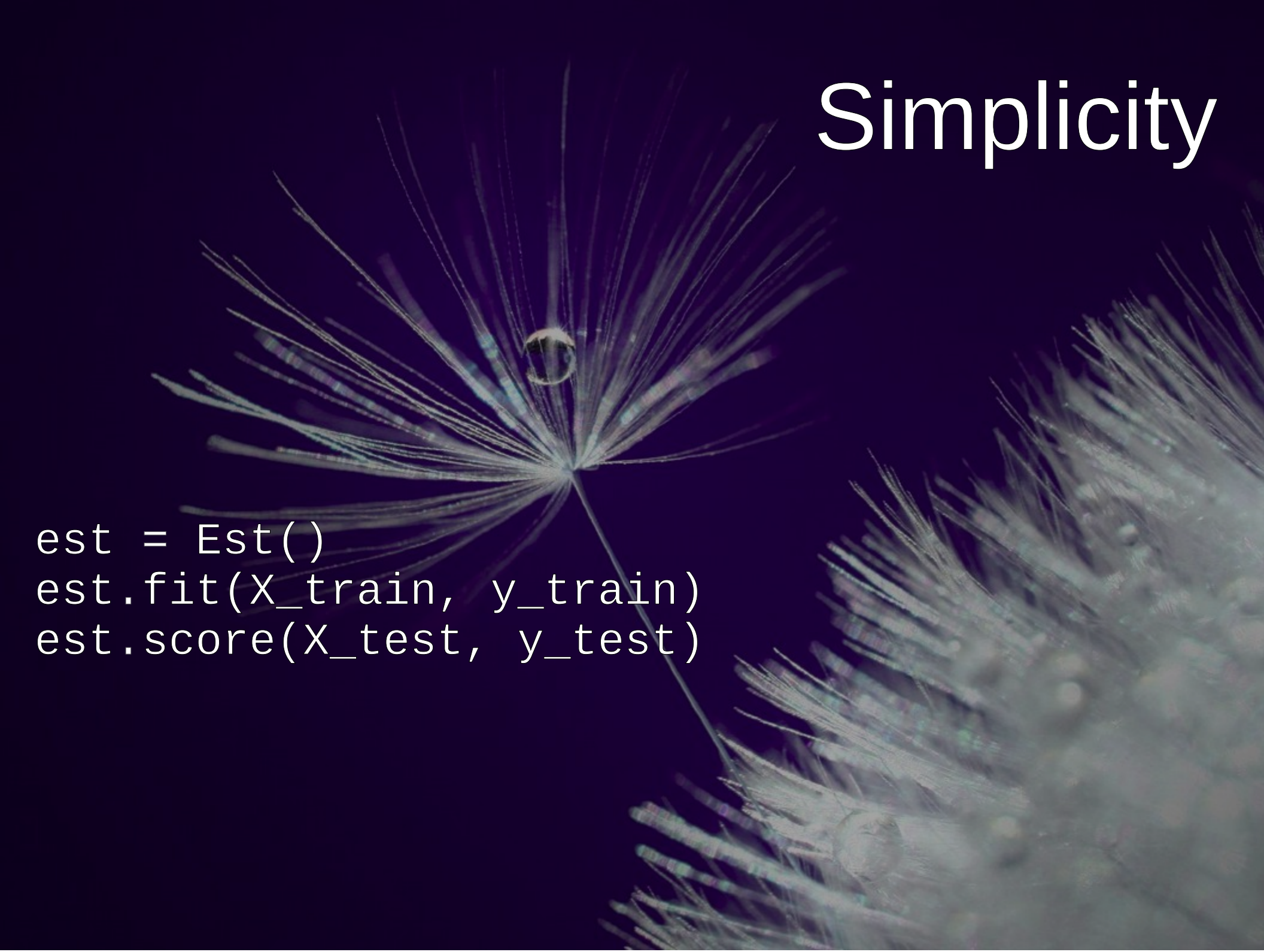


Methods

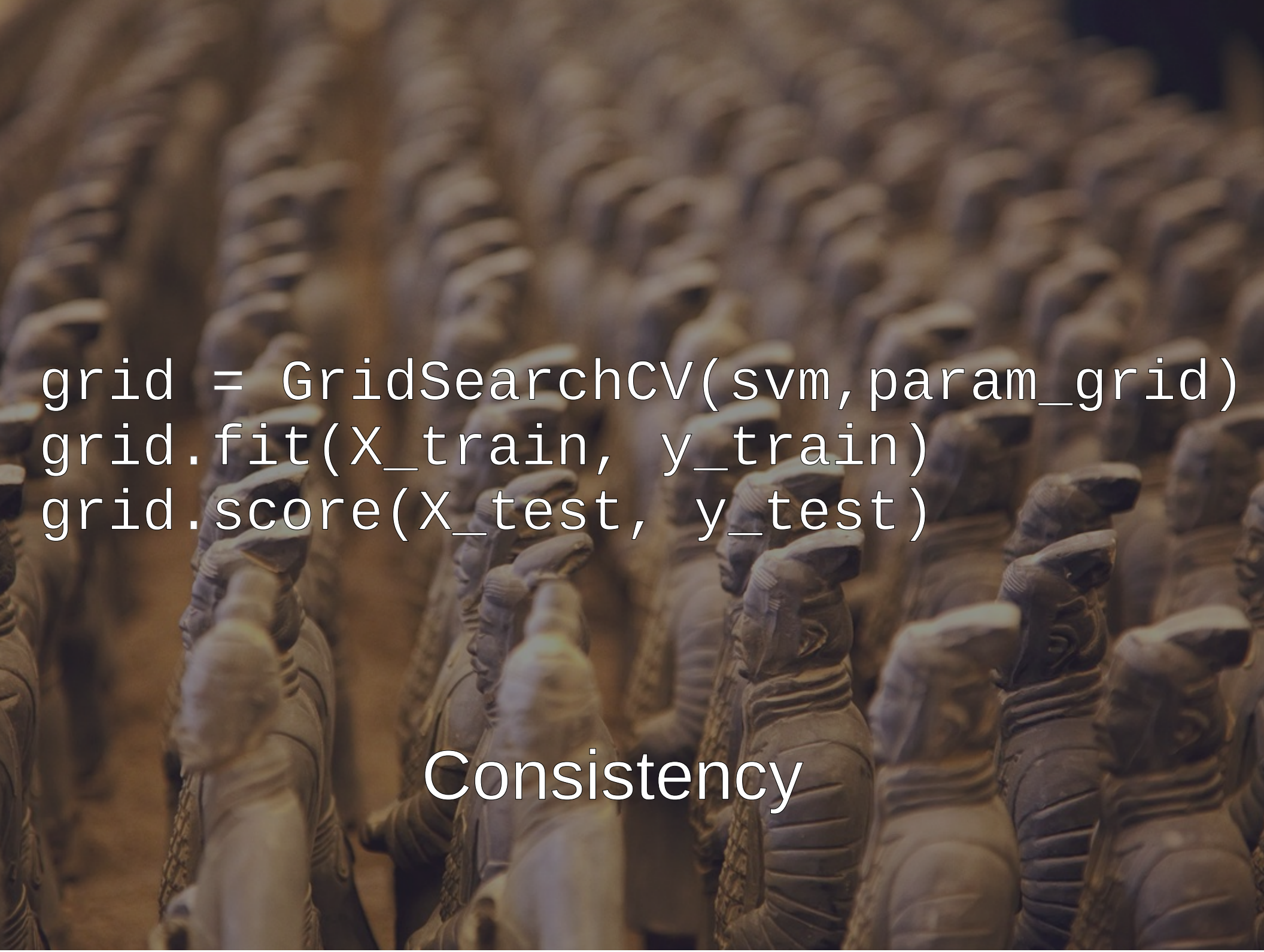
Scoping


Widely useful
Well established
sklearn
Matrix in, matrix out

Simplicity

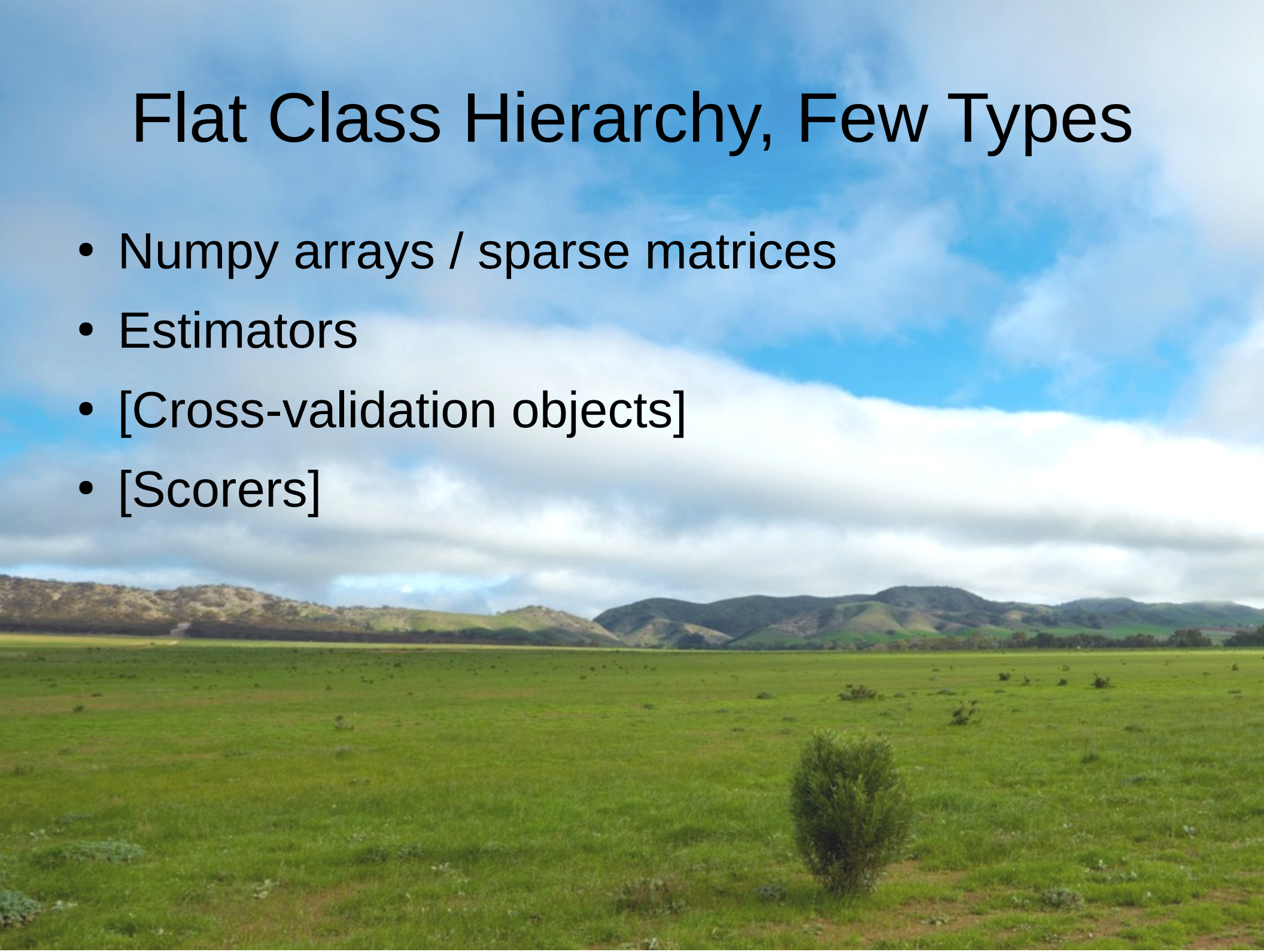


```
est = Est()
est.fit(X_train, y_train)
est.score(X_test, y_test)
```

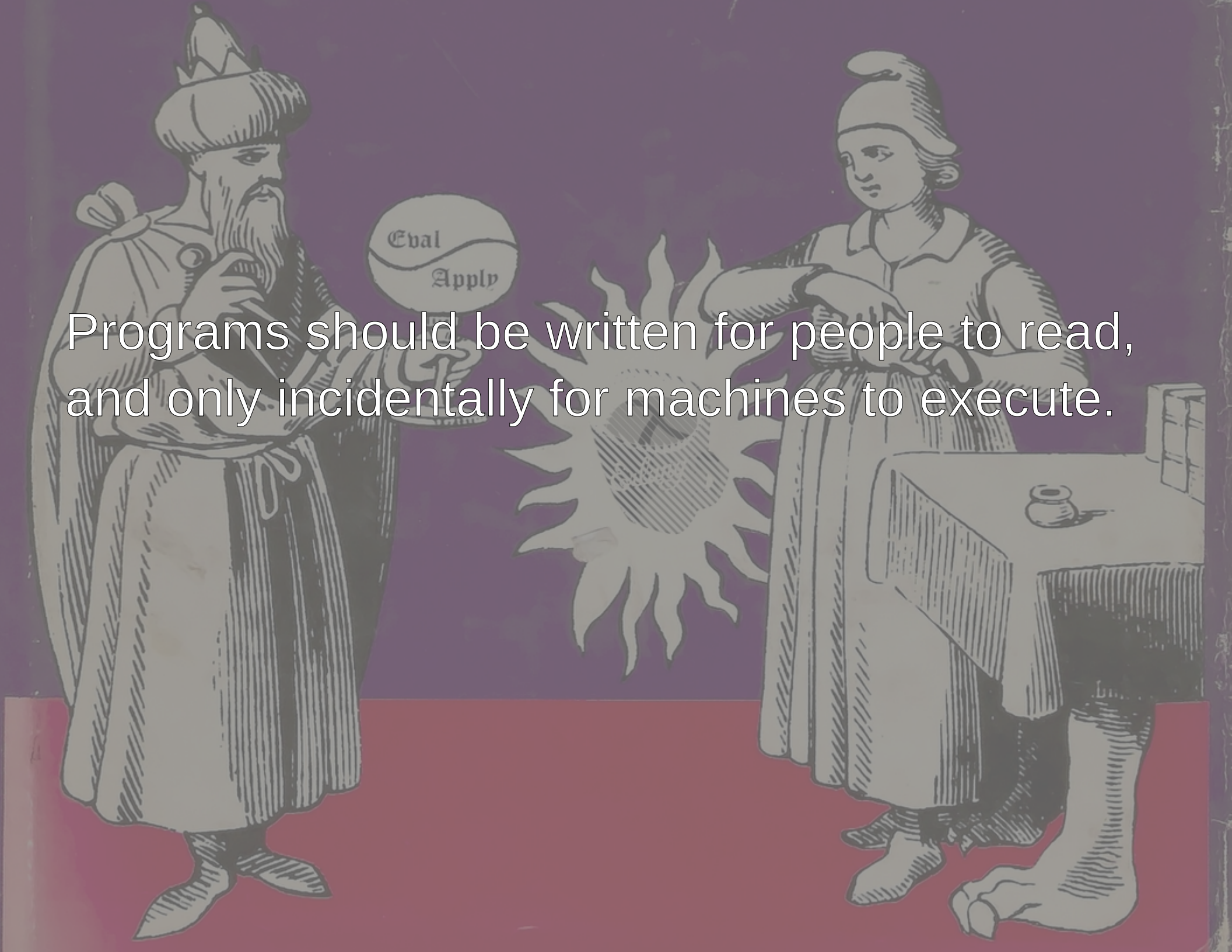

grid = GridSearchCV(svm,param_grid)
grid.fit(X_train, y_train)
grid.score(X_test, y_test)
Consistency

Flat Class Hierarchy, Few Types

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



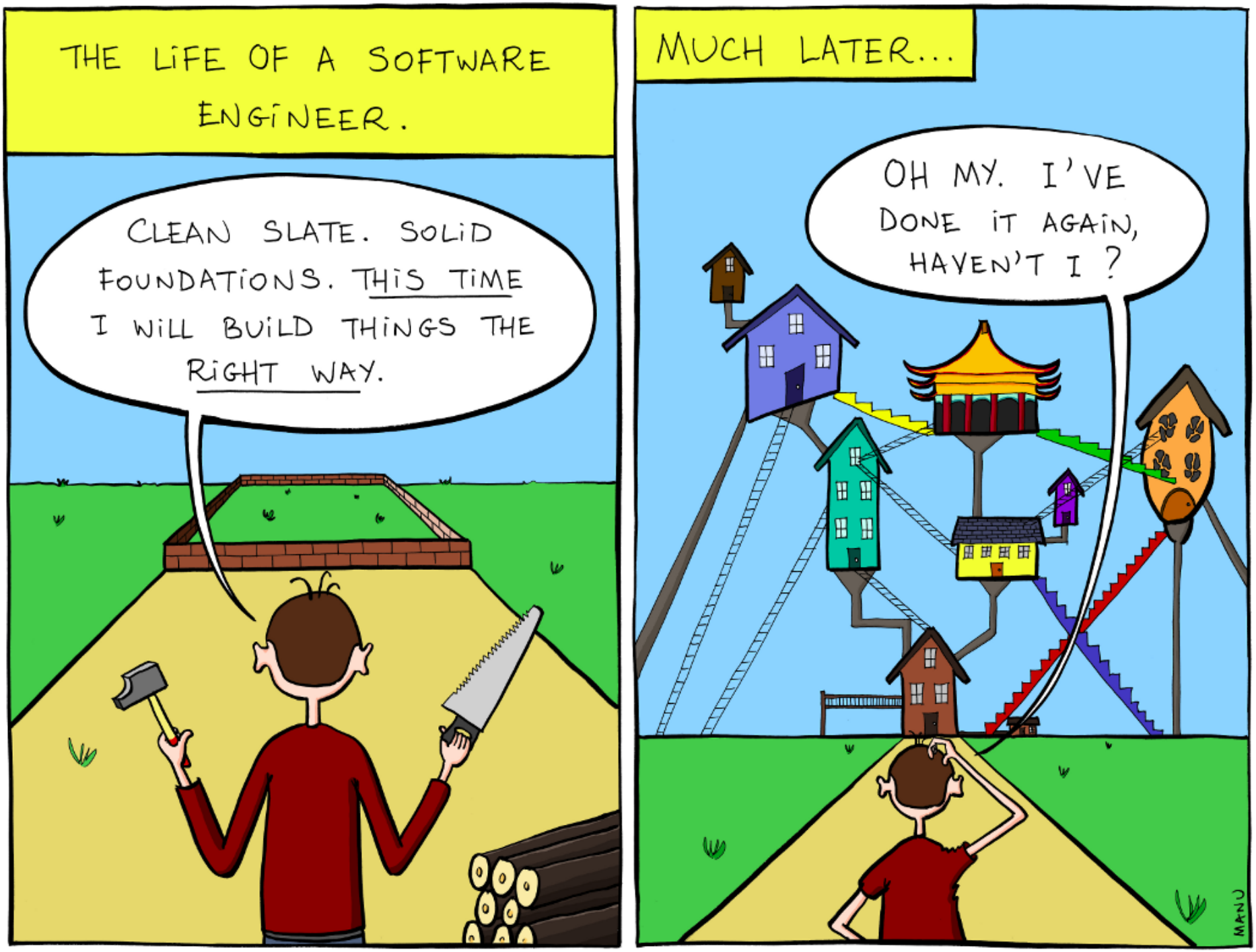
- Maintainability



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


F12
delete
avoid code;
avoid code rot!

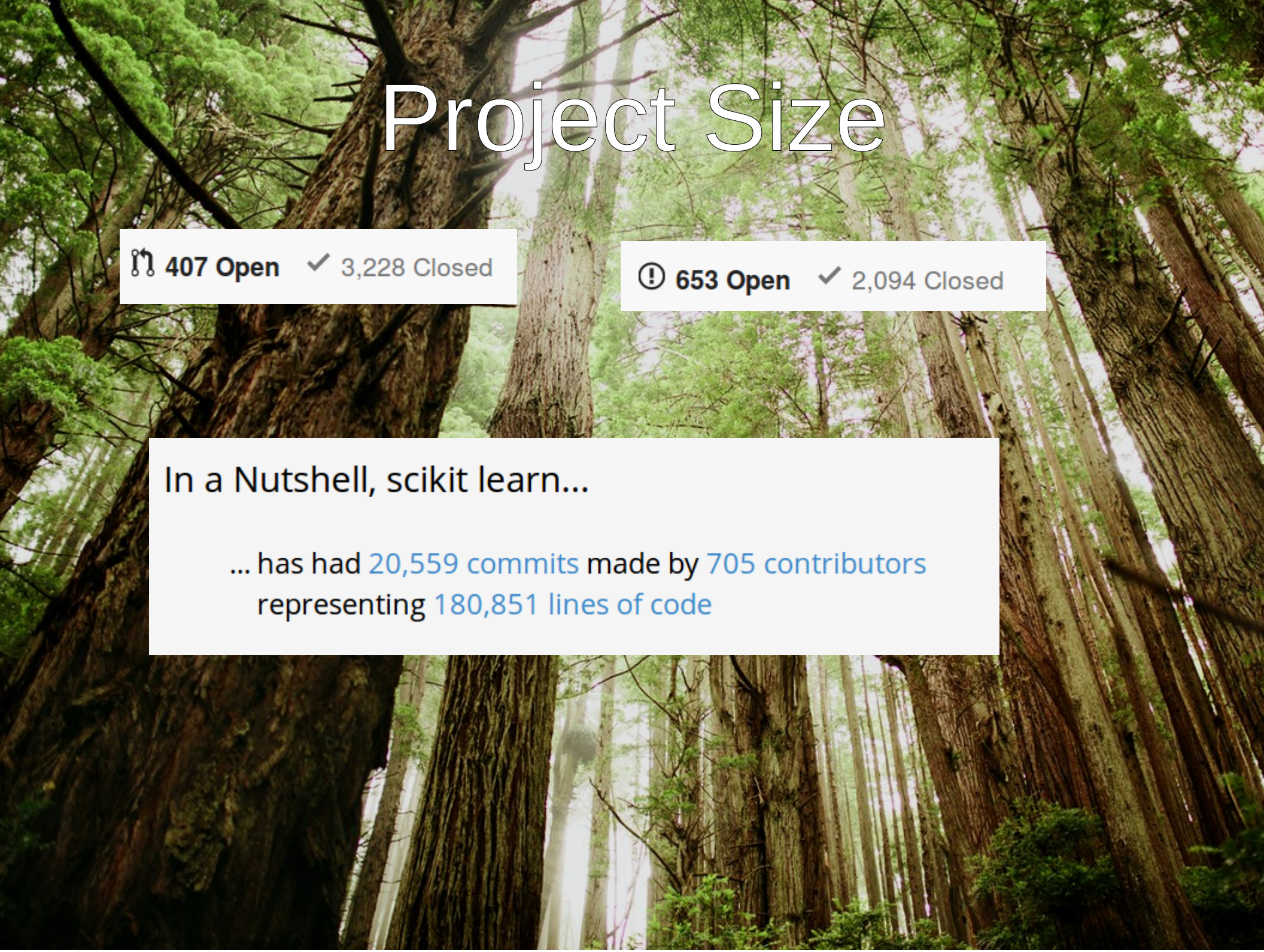
Challenges



Feature Creep

Two Language Problem

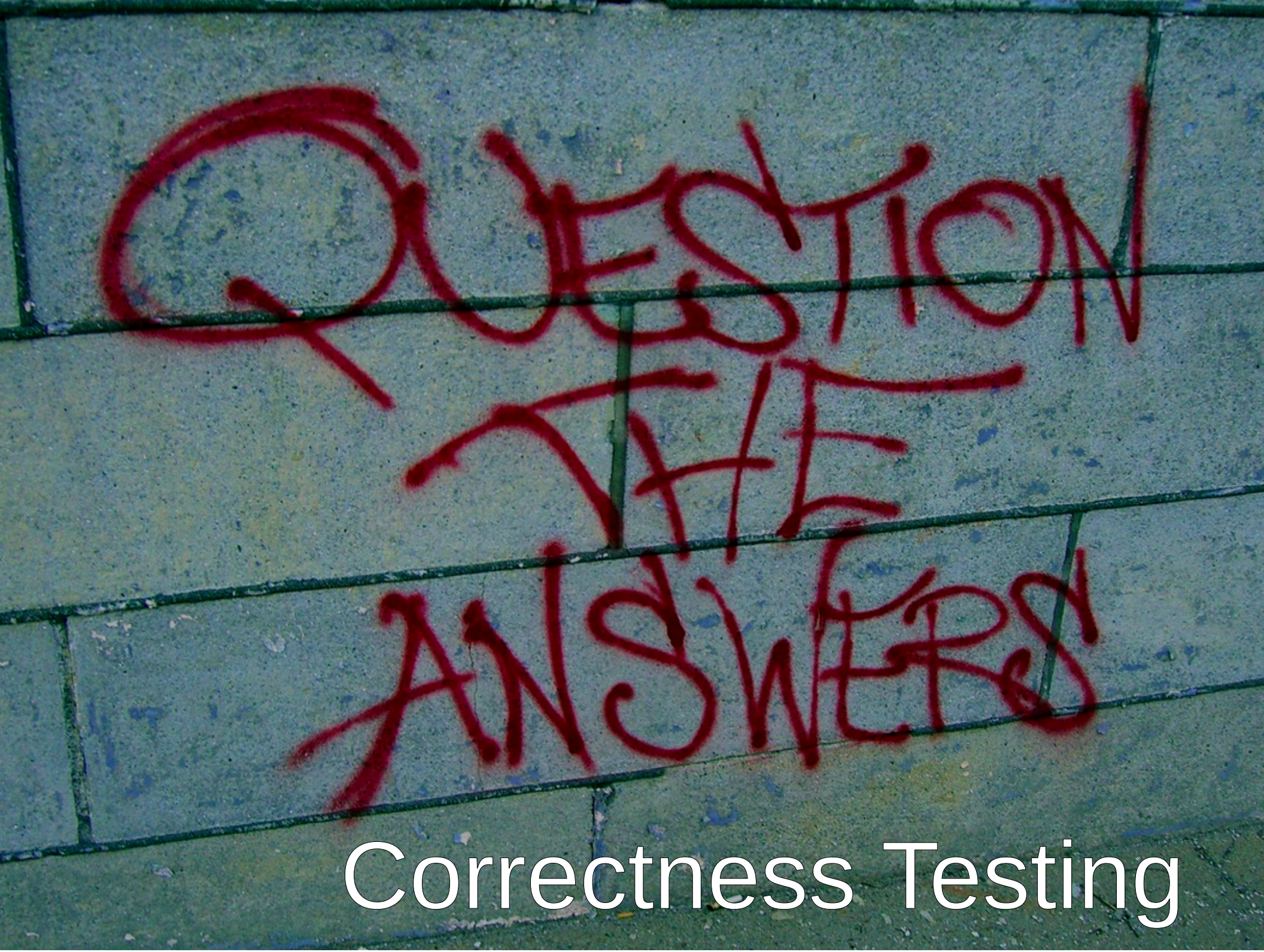
Project Size




Open
Questions

Data Structures
Categorical Variables

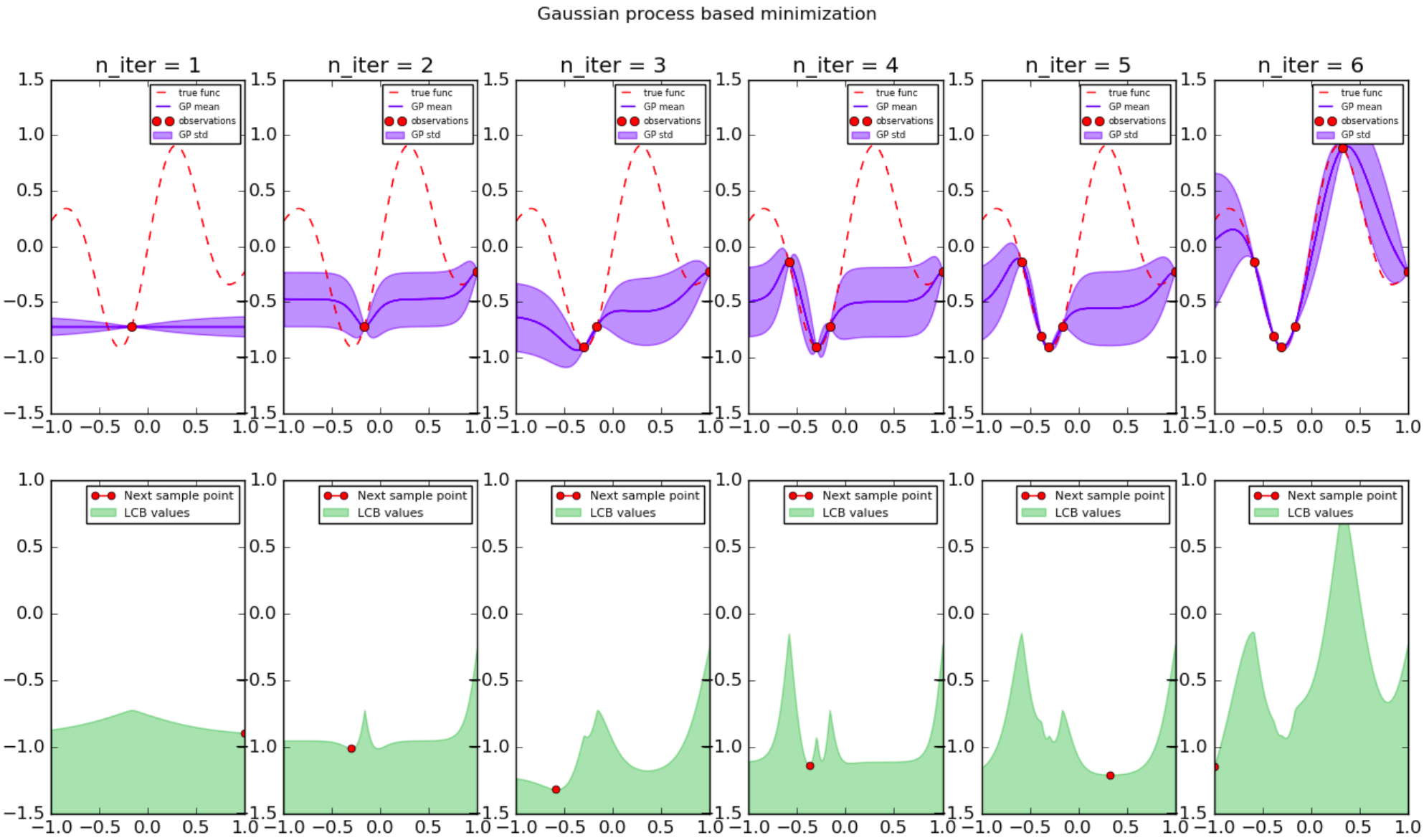

Better Defaults
Benchmarking ?



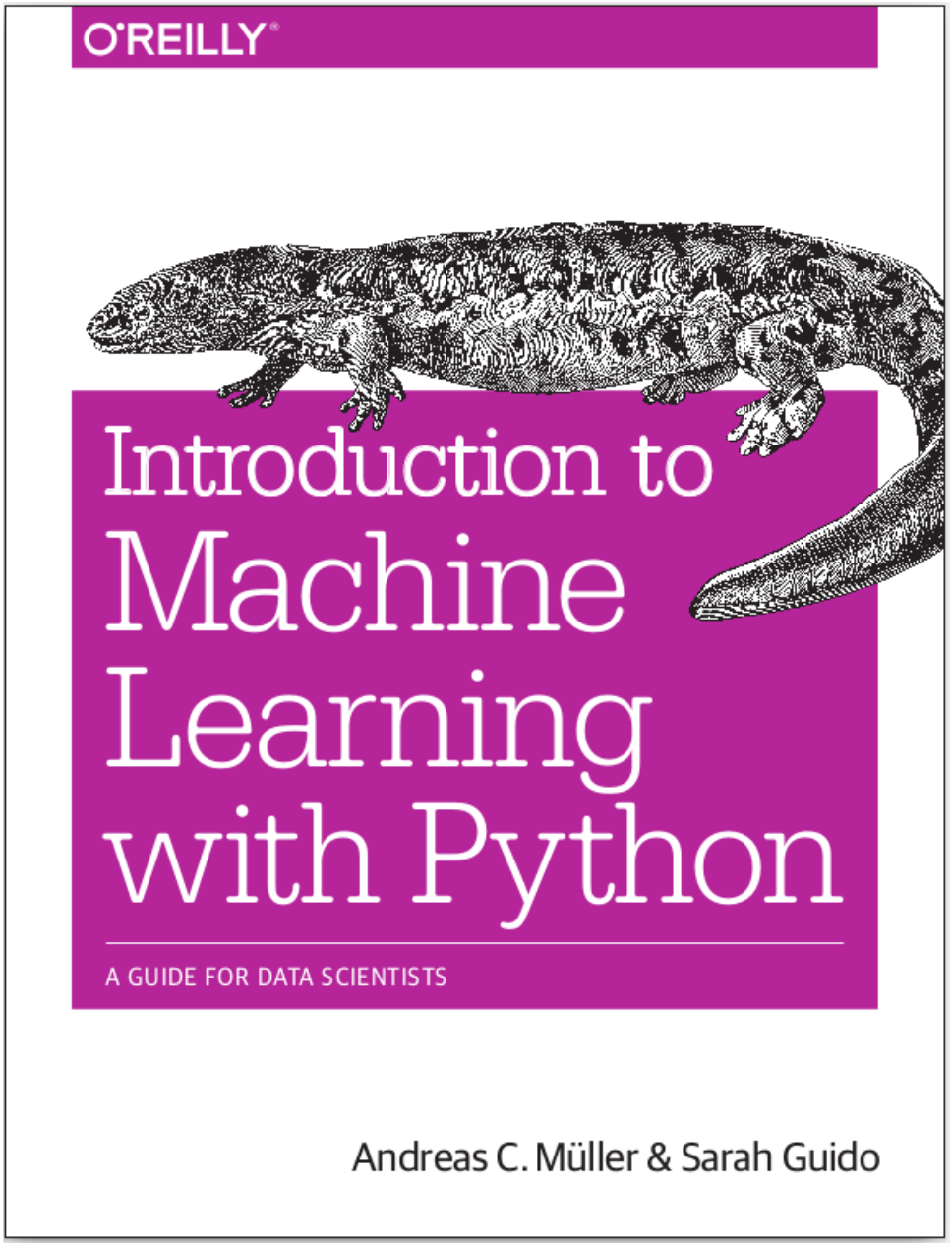
Correctness Testing

Outlook

Bayesian parameter optimization

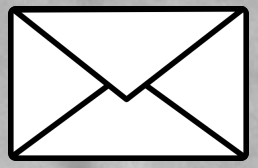


Better Feature Name Support


O'REILLY
Introduction to
Machine
Learning
with Python
A GUIDE FOR DATA SCIENTISTS
Andreas C. Müller & Sarah Guido

@amuellerml

@amueller



amueller@nyu.edu