5 plots in 5 minutes:
Using GitHub API to measure software adoption in CMS

Jim Pivarski

Princeton University – IRIS-HEP

August 11, 2020

Goal: find out how physicists are using software for data analysis

Workflow from detector $\rightarrow$ reconstruction $\rightarrow$ AODs is understood.

Workflow from AODs $\rightarrow$ analysis $\rightarrow$ proposed papers is not.

- Surveys are great, except when they're not: response rate can be correlated with the questions.
- Distributed collaborations: how do we know the survey questions have reached everyone we want to ask?
- Can we get a more direct method?

Direct method: look at their code!

GitHub API lets us query users and repositories (URL $\rightarrow$ JSON).

Can we identify "physicist" users?

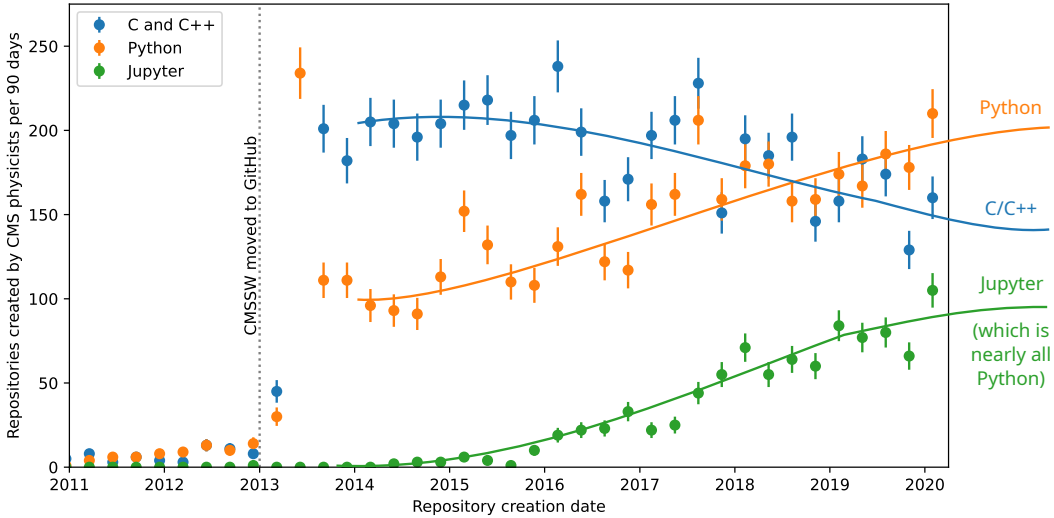
- CMSSW has been on GitHub since 2013.
- Assumption: most users who fork CMSSW are CMS physicists.
- Then examine their **non-fork** repositories.

Why GitHub/CMS? Until recently, all (free) GitHub repos were public, making them searchable by the API.

Large dataset: **3100 users** with **19 400 non-fork repos** spanning **7 years**.

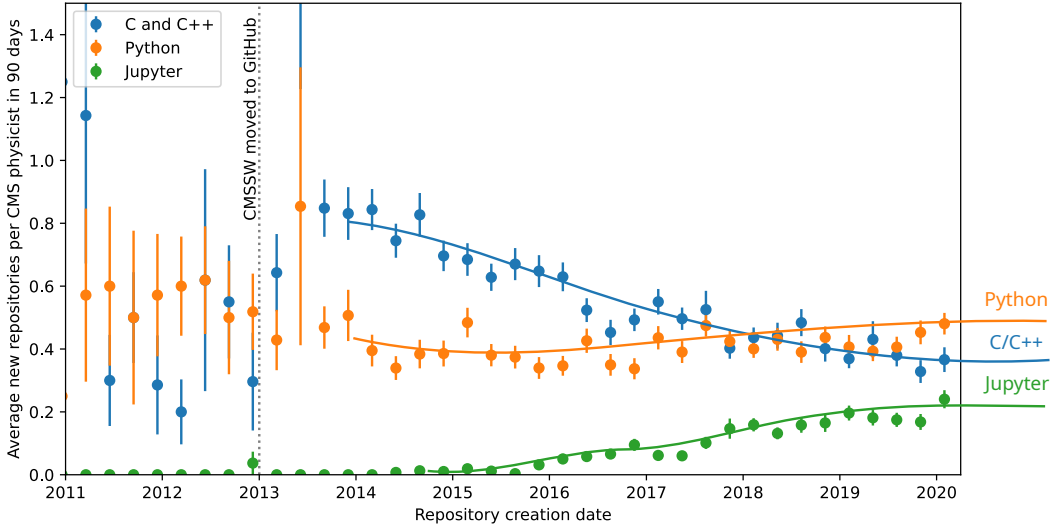
Plot #1: language choice

Using GitHub's algorithm for determining a repo's programming language.



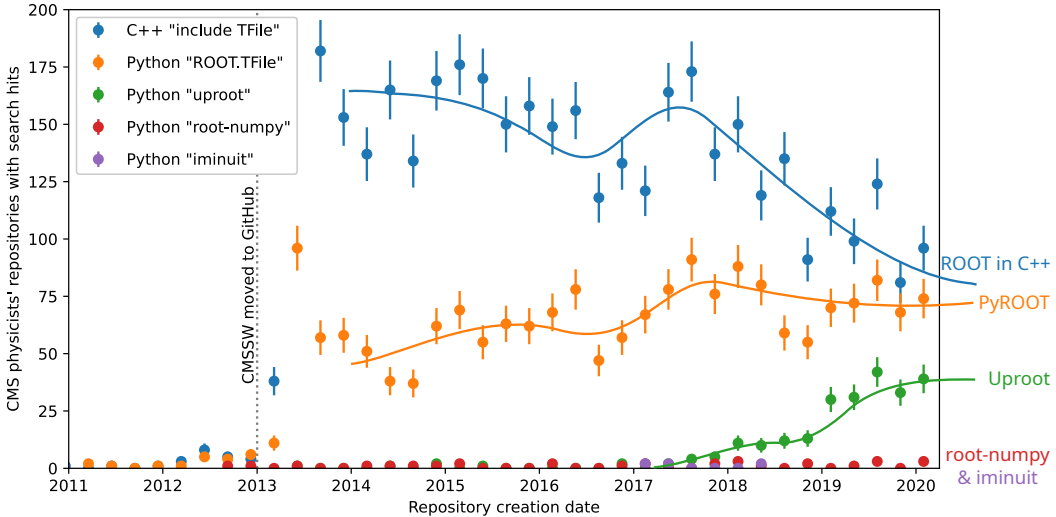
Plot #2: language choice by user

Same thing, but with average number of repos per user instead of total repos.



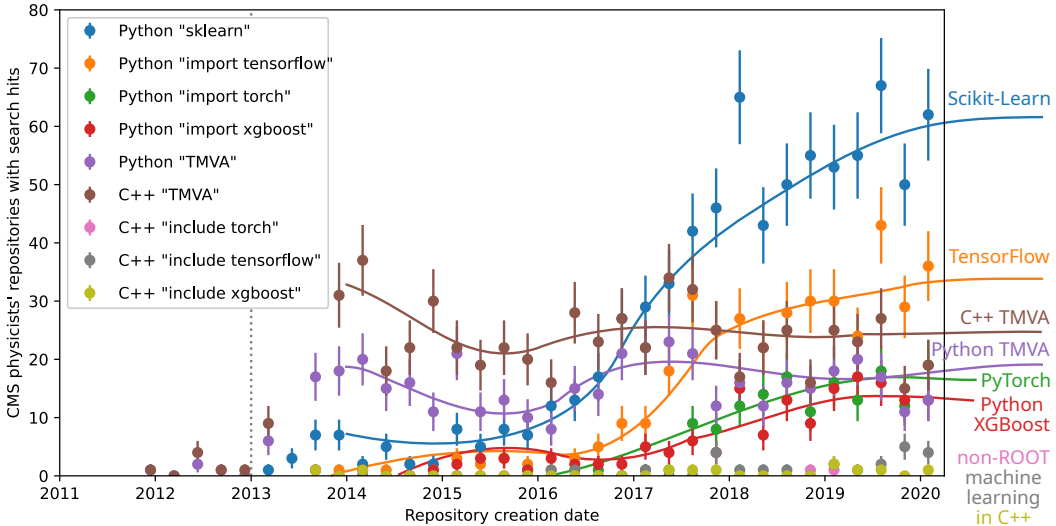
Plot #3: search for package imports

Number of repos that match a search string (within C/C++ or Python/Jupyter files).



Plot #4: what machine learning packages do they use?

Same technique. Dominance of Scikit-Learn (over TensorFlow and Torch) is surprising.



Plot #5: Did machine learning drive Python adoption?

Not really. Basic analysis tools (NumPy, Matplotlib, Pandas) outweigh Pythonic ML.