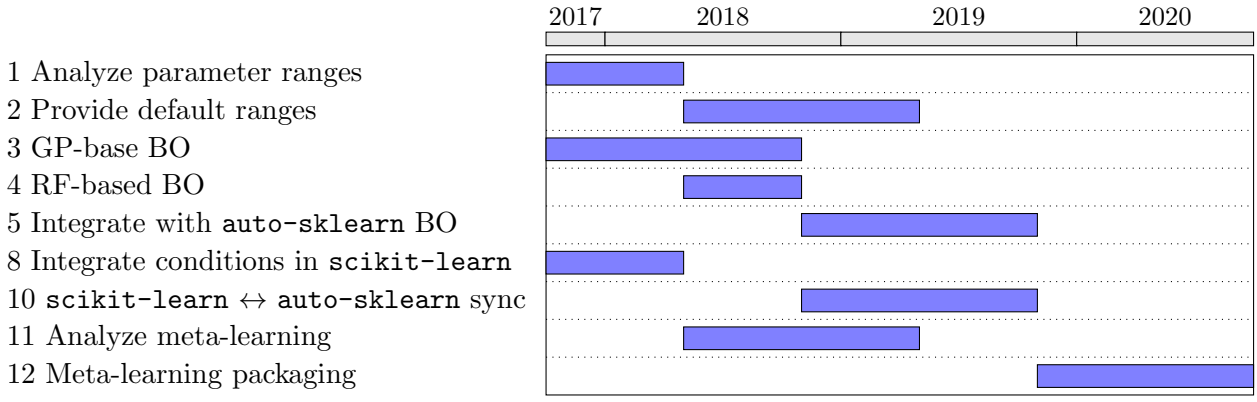

Figure 1: Project timeline

1 Change of proposed work

Due to the reduced budget, and given the current state of scikit-learn development, we will remove the tasks 6 (Searching pipeline steps), 7 (Transformation conditions) and 9 (Feature encodings in `scikit-learn`) from the project. A revised project timeline can be found in Figure 1. Task 6 is already implemented in the current version of `scikit-learn`, while tasks 7 and 9 are in the process of being implemented, partially as part of a different funded project.

2 Overlap with Other Funded Work

Another project of mine relating to `scikit-learn` has recently been funded by the Alfred P. Sloan Foundation under the title "Extensions and Maintenance of Scikit-learn", which focusses on integration of `scikit-learn` with the pandas library, better support for missing and categorical data, and better tools for understanding and visualizing models. That work does not include any aspects of automation, benchmarking or meta-learning, and removing task 9 from this proposal eliminates any overlap.

3 Educational Aspects

There are two separate aspects of this proposal that relate to education. The first is outreach to contribute to open source projects, through coding sprints and collaboration with the Women in Machine Learning and Data Science group. The other aspect is the training of students in the use of machine learning software, in particular the software developed as part of this project. I am teaching an annual course on Applied Machine Learning[1] at the Columbia Data Science Institute. This elective is part of the graduate curriculum of the Data Science program, but is open to students from other programs. Notably, this years course had participants from programs in Astronomy, Statistics, Economics, Urban Development, Applied Mathematics and Finance. This class is focused on teaching data analysis skills and software tools that can be applied to real-world problems. Tools developed as part of this project will be included in the course as they are completed. This will also provide feedback for further refinement of the software. The course material is publicly available and licensed under the CC-0 license (equivalent to public domain under US legislation).

[1] https://amueller.github.io/applied_ml_spring_2017/lectures.html

4 Licensing of Software Products

All software produced as part of this project will be licensed using the BSD three-clause licensed reproduced below:

```
Copyright (c) 2007-2017 The scikit-learn developers.
All rights reserved.

Redistribution and use in source and binary forms, with or without
modification, are permitted provided that the following conditions are met:

  a. Redistributions of source code must retain the above copyright notice,
     this list of conditions and the following disclaimer.
  b. Redistributions in binary form must reproduce the above copyright
     notice, this list of conditions and the following disclaimer in the
     documentation and/or other materials provided with the distribution.
  c. Neither the name of the Scikit-learn Developers  nor the names of
     its contributors may be used to endorse or promote products
     derived from this software without specific prior written
     permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS"
AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE
IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE
ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR
ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL
DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER
CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT
LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY
OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH
DAMAGE.
```

The BSD three-clause license and other variations of the BSD license are recognized licenses by the Open Source Initiative[2]. The BSD three-clause license is the license currently used by the `scikit-learn` project, which will facilitate integration with the existing package. This is also the license used in other projects within the scientific Python ecosystem, such as NumPy, SciPy, Pandas and others, and is a de-facto standard in the community. The three-clause BSD license is a permissive open source license (as opposed to a copyleft license like the GPL) that allows unlimited use and modification of the code, both in a scientific and commercial contexts. The open nature of the license encourages scientists and companies to contribute code without fear of losing intellectual property rights (such as patents) because of their contributions, while encouraging usage of the software without the possibly severe consequences of a copyleft license.

[2] https://opensource.org/licenses

5 Quantitative Usage Metrics

We will use two main metrics to quantify the adoption of our software: Usage in open source code published on GitHub, and contributors. Thanks to the Google BigQuery interface for open source repositories on GitHub[3] and the Github code search feature it is possible to quantitatively analyze a large amount of scientific software and experimentation code. This even allows for fine-grained analysis of which features of the software are used. For code that is contributed to the `scikit-learn` package, we will use this code analysis exclusively, as contributors to parts of the package are hard to track[4]. For code that lives in a separate package, we will use contributors and code analysis statistics.

Year 1 At least 10 open source projects or research projects using the provided features

Year 2 At least 20 open source projects or research projects using the provided features, at least 2 external contributors to the project.

Year 3 At least 50 open source projects or research projects using the provided features, at least 5 external contributors to the project.

We do not include citations into our metrics, as citations for software are unfortunately rare, and a paper to describe the meta-learning project could only be published at the end of the third year. Counting citations to `scikit-learn` related publications is unlikely to reflect the specific outcomes of this project.

6 User and Community Engagement

There are two kinds of communities that we need to engage for this project: Those that already use machine learning, but could save time and potentially achieve better results by using automation, and those that do not use machine learning yet, because of the current obstacles in doing so. Reaching the first group is fairly easy, as the `scikit-learn` project is already established in the scientific community as one of the most popular machine learning tools. Any improvements to the `scikit-learn` package are therefore automatically available to these users. Researchers that are using `scikit-learn` are also likely to engage with the `scikit-learn` community, and therefore making them aware of related tools outside of the `scikit-learn` main package is possible via mailing lists, references on the website, and social media. In addition, we identified "early adopters" in several disciplines which whom we will be in direct contact during the project.

Reaching communities in which machine learning is not already established is somewhat harder, but is helped the central role of the Data Science Institute within Columbia, and the role of Columbia as a host of the Northeast Big Data Innovation Hub (NEBDIH). Within Columbia, the Data Science Institute is a contact point for machine learning and data processing needs for all institutes in the university. This is implemented explicitly in the Collaboratory[5], which offers a help-desk for data science needs. This program enables us to build bridges to communities that are currently not using machine learning, and less engaged in the scientific software community. Beyond Columbia, the NEBDIH is tightly integrated into a wide array of research communities that rely on data processing for their research. These are the ideal communities to reach out to, for advertising simplified machine learning tools.

[3] https://cloud.google.com/bigquery/public-data/github

[4] The number of editors of a file is not a good proxy for the number of contributors.

[5] http://collaboratory.columbia.edu