Jupyter Notebooks on GitHub: Characteristics and Code Clones

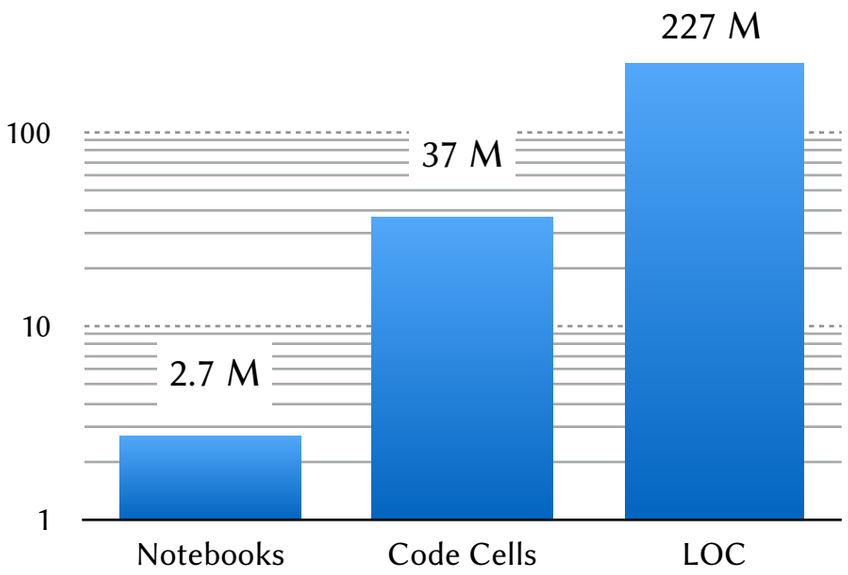
Malin Källén
Uppsala University

Ulf Sigvardsson
Independent

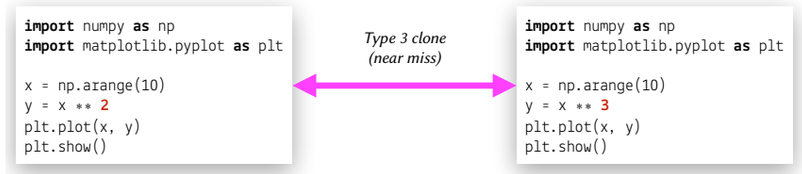
Tobias Wrigstad
Uppsala University

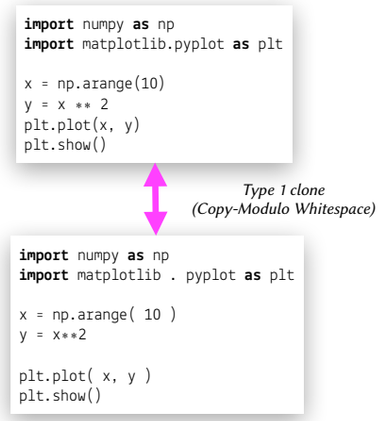
Our goal: to understand how ”data science programs” are written by looking at Jupyter Notebooks

Corpus of Jupyter Notebooks mined from GitHub



Code Clones





Copying code is very common in software development

In the small: copy-and-patch of a single function within a program

In the large: copy entire file tree across projects

Copying and software engineering

Makes it harder to fix bugs multiplied by copy

More code to understand and maintain

Copying and software engineering science

Makes it harder to study code

To draw conclusions from a code corpus — we need to understand the amount of cloning in the corpus which may otherwise skew our results

We analysed clones both on the level of files and the code cells in the notebooks

~1 M notebooks in our corpus are copies of other notebooks in the corpus

Notebook templates from courses and similar are a frequent source of clones

Only ~9.2 M unique code cells in our corpus

Sources of clones are often found within the same repository