

要約：

Jupyter Notebookのベストプラクティスはこれ！！

A Large-scale Study about Quality and Reproducibility of Jupyter Notebooks

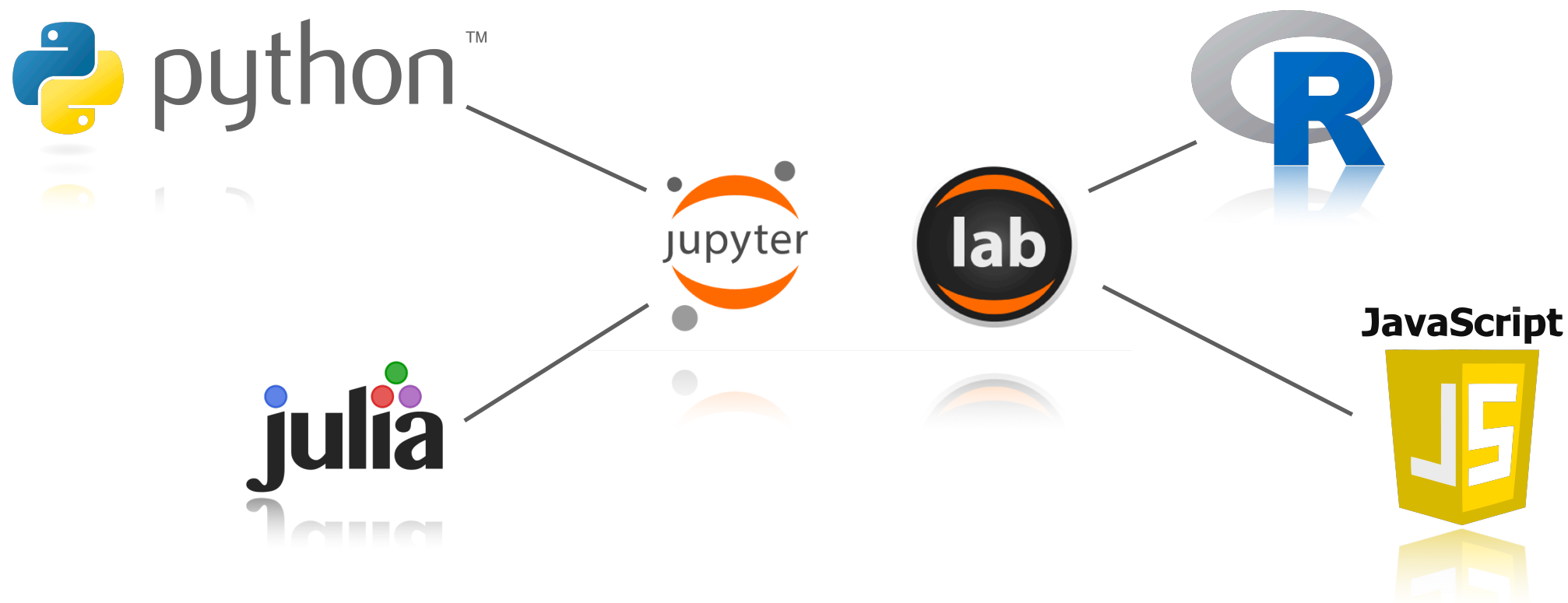
Joaõ Felipe Pimentel, Leonardo Murta, Vanessa Braganholo,
and Juliana Freire

MSR '19 Proceedings of the 16th International Conference on Mining Software Repositories
Pages 507-517

発表者 木村 祐太

Jupyter Notebook

- 2013年にリリース
- IPythonから始まり，現在では複数の言語をサポート
- インタラクティブなシステムで，広く使われている



Jupyter Notebookの問題点

- 再現性に乏しく， 予期せぬ挙動を起こしやすい

再現性の調査

GitHubリポジトリに含まれるノートブック(1,450,071件) 

有効なPythonノートブックは863,878件 

エラーなしで実行できたものは24.11% 

Pythonのバージョンや
実行順序が記載されている

「結果が再現できたものは4.03%」

先行研究

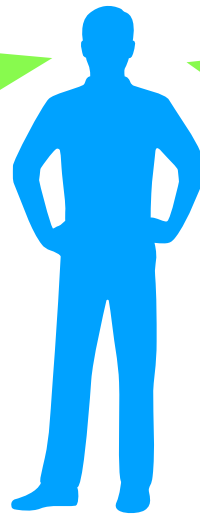
- Jupyter Notebookに対して否定的な研究が多数
- しかし、**再現性に影響を与える特性は調べられていない**

目的

良い慣習と悪い慣習を
分析して、再現性に関する
知見を提供しよう

否定的な意見が多いけど
Jupyterって
本当に使えないの？

Jupyterの開発のために
良い慣習を提案



データセット

- ・ 対象：GitHubの265,143リポジトリに含まれる.ipynbファイル
- ・ 期間：2013年1月1日 ～ 2018年4月16日

※データセットの詳細・内訳は予備資料参照

実験の前に

- **Prospective Data**

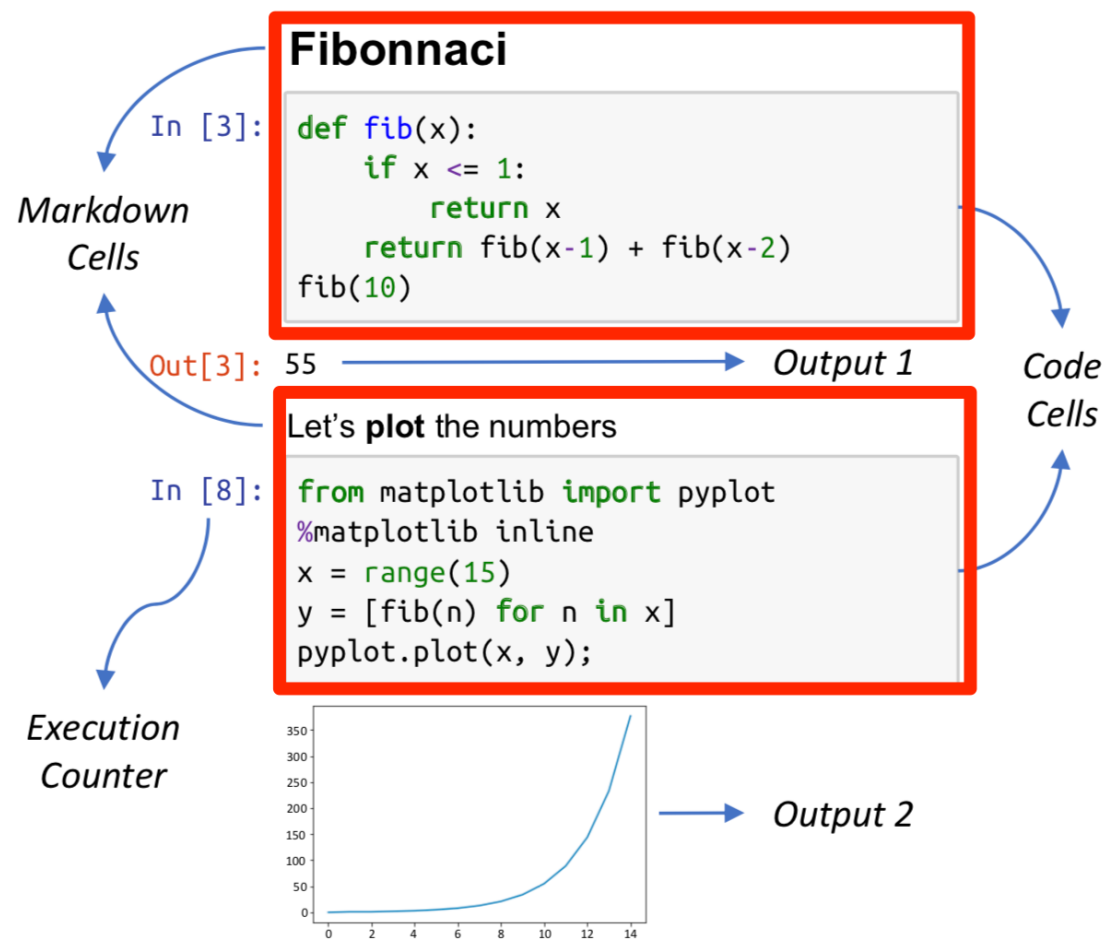
- ・ エディタに記述されているコードやファイル名

- **Retrospective Data**

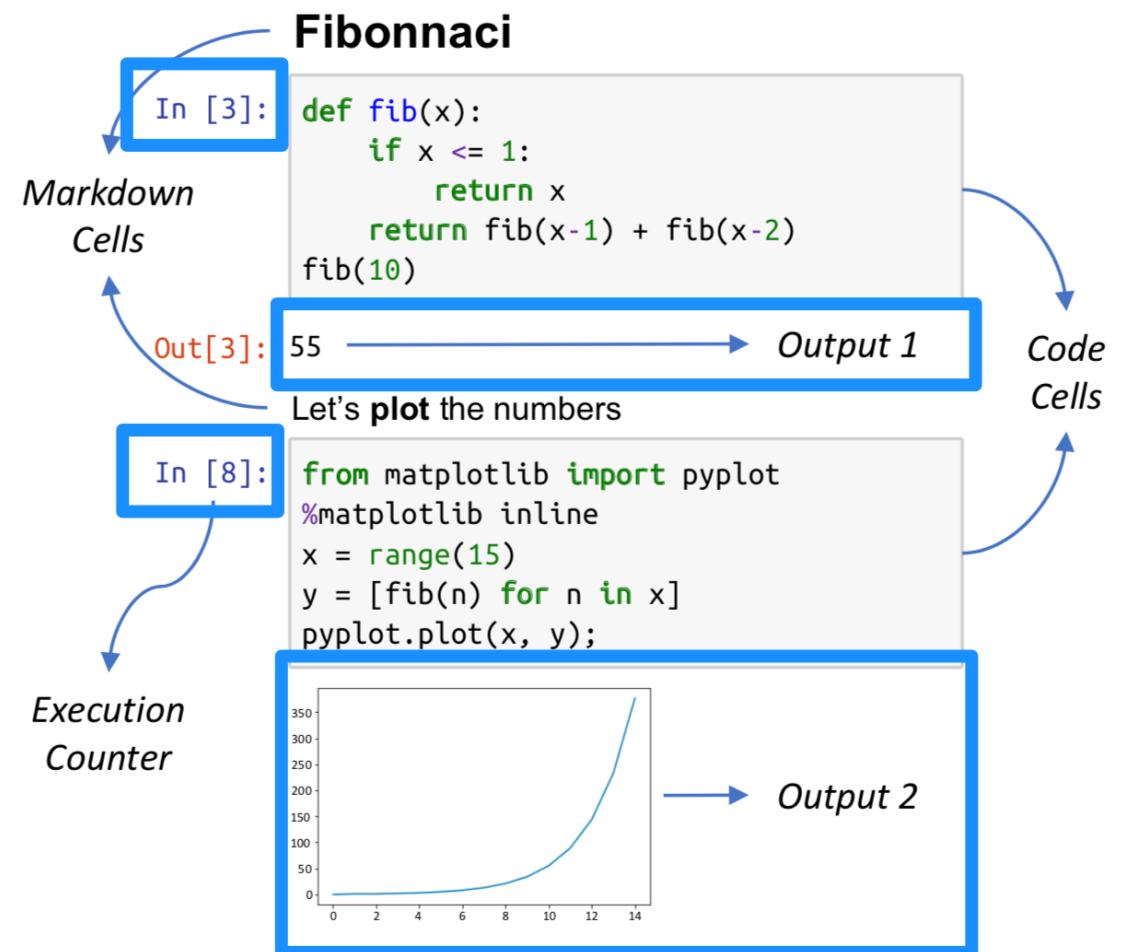
- ・ 実行時の出力や実行順序番号

実験の前に

Prospective Data



Retrospective Data



Research Questions

<Prospective Data>

1. Jupyter Notebookの機能はどんな感じで使われてる？
2. ファイルの名付けはどんな感じ？
3. モジュールや関数, クラスの利用はどんな感じ？
4. Jupyter Notebookではどのくらいテストされてるの？

RQ1

機能の使われ方は？

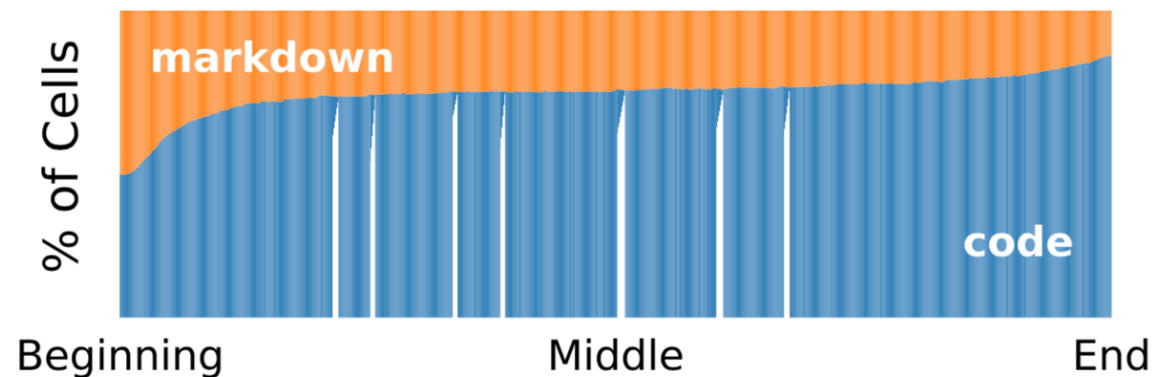


Fig. 5. Distribution of cell types in the notebook.

- ノートブック内で約1/4が
Markdownのセル



使われ方は、
ヘッダーや段落区切りなど
有効に利用されているとは
言えない…

RQ2

名付けはどんな感じ？

デフォルト名やコピーファイルは約3%
大体の人がきちんと名付けしている

ただし、約27%のファイル名は
OSが違えば使えない名前



RQ3

関数などの使い方は？

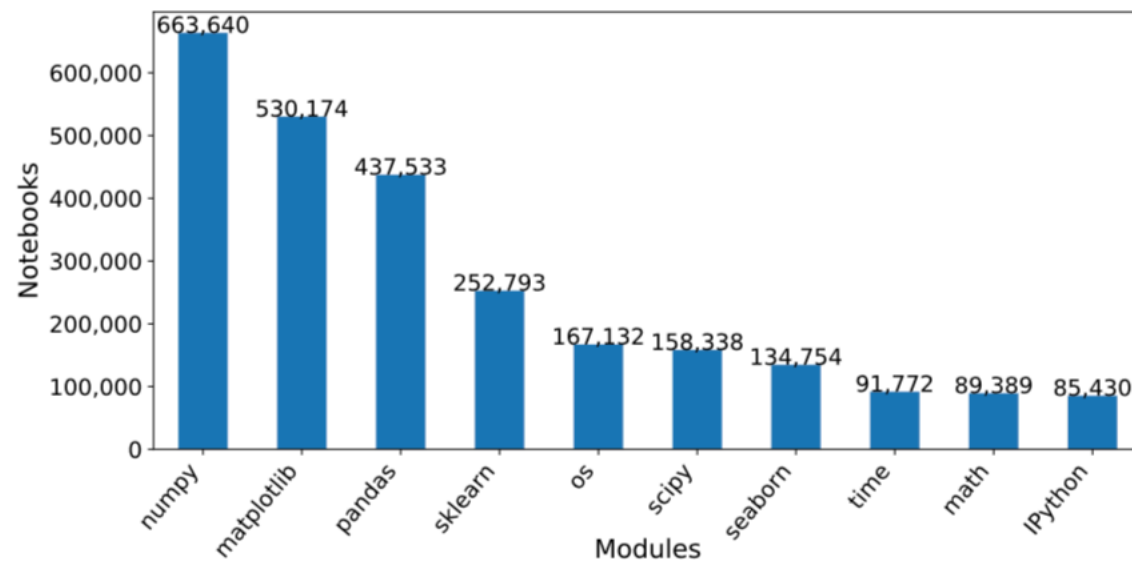


Fig. 7. Top 15 most imported modules.

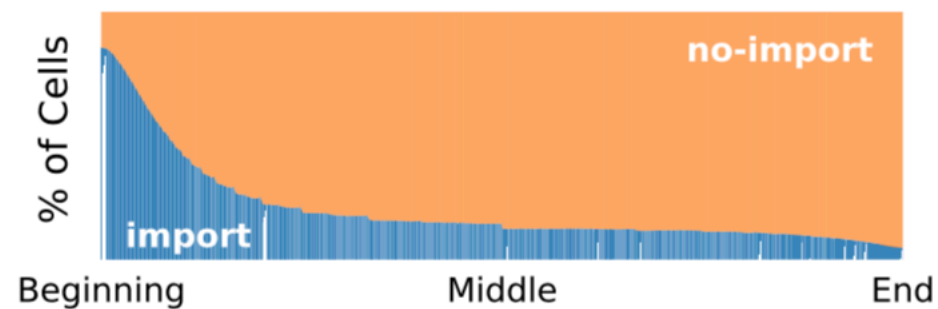


Fig. 8. Distribution of imports in notebooks.

- データ解析に用いられる有名モジュール
- モジュールのインポートは大体はじめの方にされる

RQ3

関数などの使い方は？

- 関数はよく使われるがクラスはあまり使われない
- ループや条件が使われているノートブックの約7割が関数を定義している

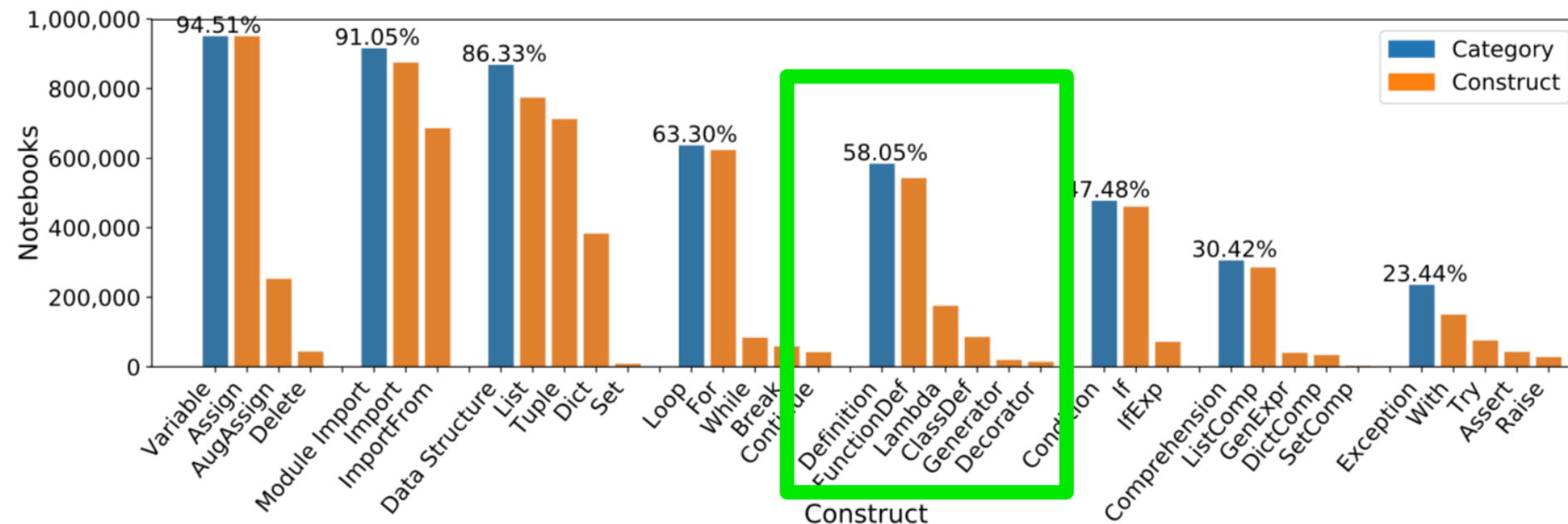


Fig. 9. Distribution of Python constructs in notebooks. This figure groups constructs into categories. The constructs of a category appear on the right of the category bar. A category corresponds to the union of its constructs.

RQ4

テストしてますか？

- テストモジュールをインポートしている
- testなどのようなワードが含まれているモジュール

➔ 1.54%しか存在しない

データ解析などによく使われる
➔ 要求されるテストの数が膨大



Research Questions

<Retrospective Data>

- 5. ユーザは出力をノートブックに保存する？
- 6. 実行どんな感じで行ってる？
- 7. 再現性はどんな感じ？

RQ5

出力結果を残してる？

TABLE I
OUTPUT FORMATS IN CELLS AND NOTEBOOKS

Format	% of cells with output	% of executed notebooks
Text	95.79%	81.98%
Image	31.41%	51.00%
HTML/JS	22.90%	36.86%
Error	3.22%	14.90%
Formatted	1.78%	1.93%
Extension	1.04%	0.00%
PDF	0.12%	0.12%

- 実行されたノートブックは全体の85%
- さらに出力が残っているものは85%
- 大体はテキストが出力されている

RQ6

どんな感じで実行されてる？

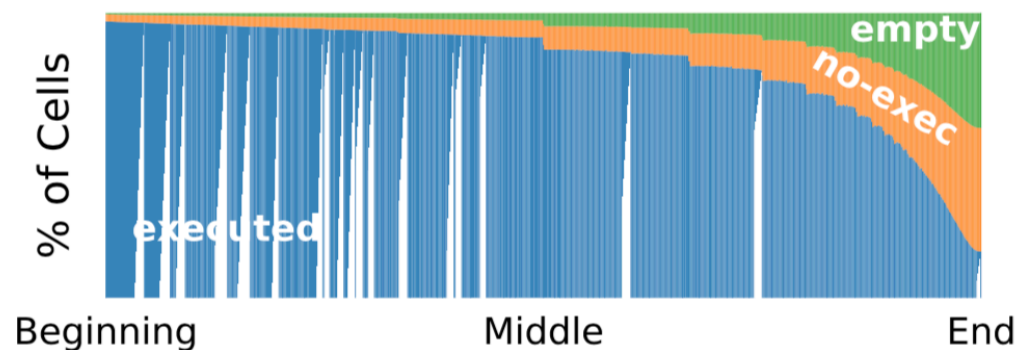


Fig. 10. Distribution of code cells in notebooks.

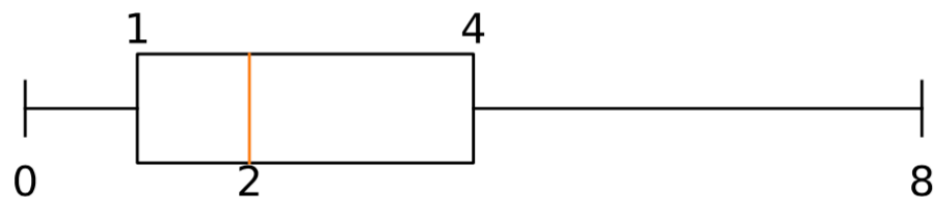


Fig. 11. Distribution of skips. Max. outlier: 220.

- 実行されていないセルや空のセルは後ろの方に集まりやすい

- 実行順序は整理されていない

スキップの例

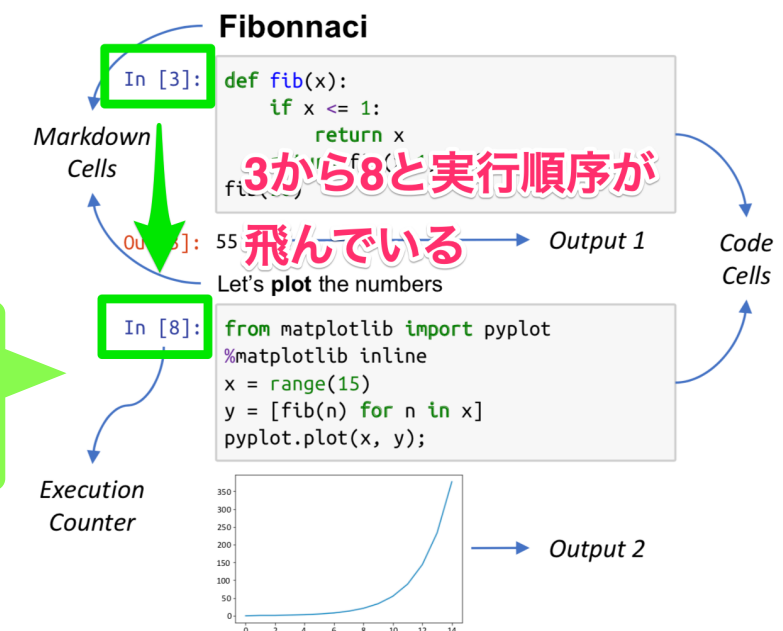


Fig. 1. A notebook example with markdown, code, and output.

RQ7

再現性どう？

- モジュールの依存関係を表記しているリポジトリに含まれる Python ノートブックは **118,483 / 863,878 (13.72%)**
 - requirements.txt, setup.py, Pipfile のいずれかを含むもの
- インストールに失敗した数：75,059 / 118,483
 - requirements.txt：61.17% → 記述の形式が固まっている
 - setup.py：67.55% → コードによる記述のためエラーを含みやすい
 - Pipfile：65.20% → 登場してから年数が経っていない

RQ7

再現性どう？

表：インストール時のエラー

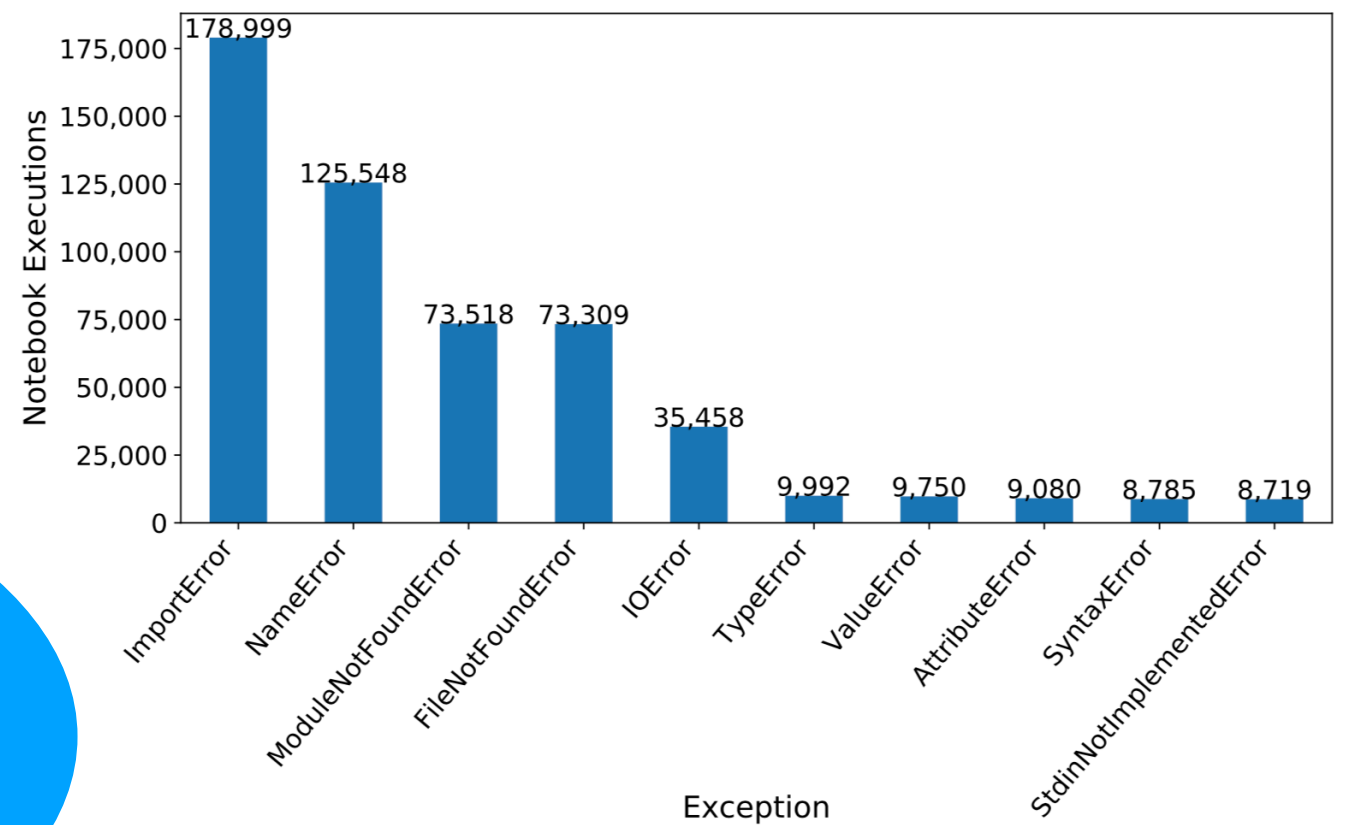
利用不能なファイル	35.04%
構文エラーや依存関係の競合を含むファイル	24.77%
Pythonのバージョンに関係するファイル	25.67%
外部のライブラリを必要とするファイル	8.73%

RQ7

再現性どう？



インポートエラー
がすごく多い！



図：実行時のエラー

RQ7

再現性どう？

- ・ 最後まで実行できたもの：
- ・ 下のセルになる程異なる出力結果を持つ

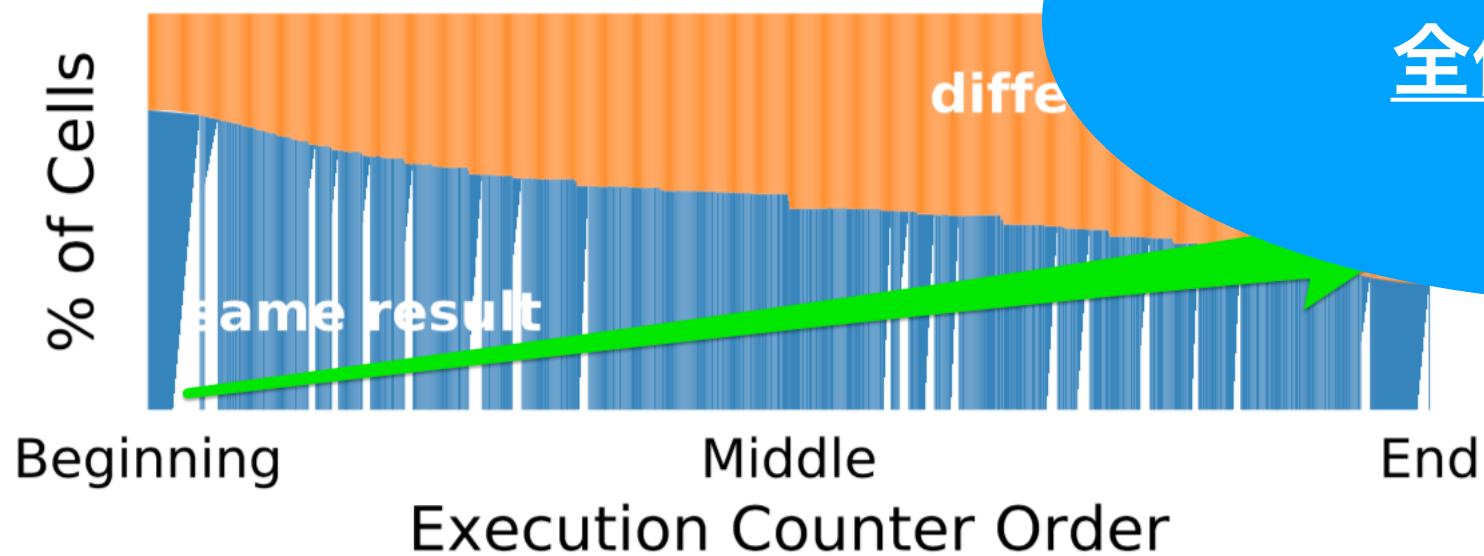


Fig. 13. Distribution of cell reproducibility.



ベストプラクティス

- ファイル名は短く， 詳細はノートブック内のMarkdownで
- 下のセルにしっかり注意を払いましょう
- 関数などに抽象化し， きちんとテストしましょう
- 必要なファイルをきちんと記述し， それら全てのバージョンがわかるようにしましょう
- 依存関係をテストするためにクリーンな環境を使用して、それらすべてが宣言されているかどうかを確認しましょう

ベストプラクティス

- インポートは初めにしましょう
- リポジトリ内のデータにアクセスするために、相対パスを使いましょう
- コミットする前に、最初から最後までもう一度実行して確認しましょう

まとめ・所感

- Jupyter Notebookの良いところ・悪いところを分析
- 7つのRQから8つのベストプラクティスに関する知見が得られた
- Jupyter Notebookは個人の検証用に用いているイメージ
- 個々人のコーディングする際の意識に依存している

予備資料

対象データの内訳

表：抽出されたノートブック数

全データ数(checkpointは除く)	1,450,071
有効なノートブック数	1,423,676
コピーされたノートブック数	264,510
利用されたノートブック数	1,159,166

対象データの内訳

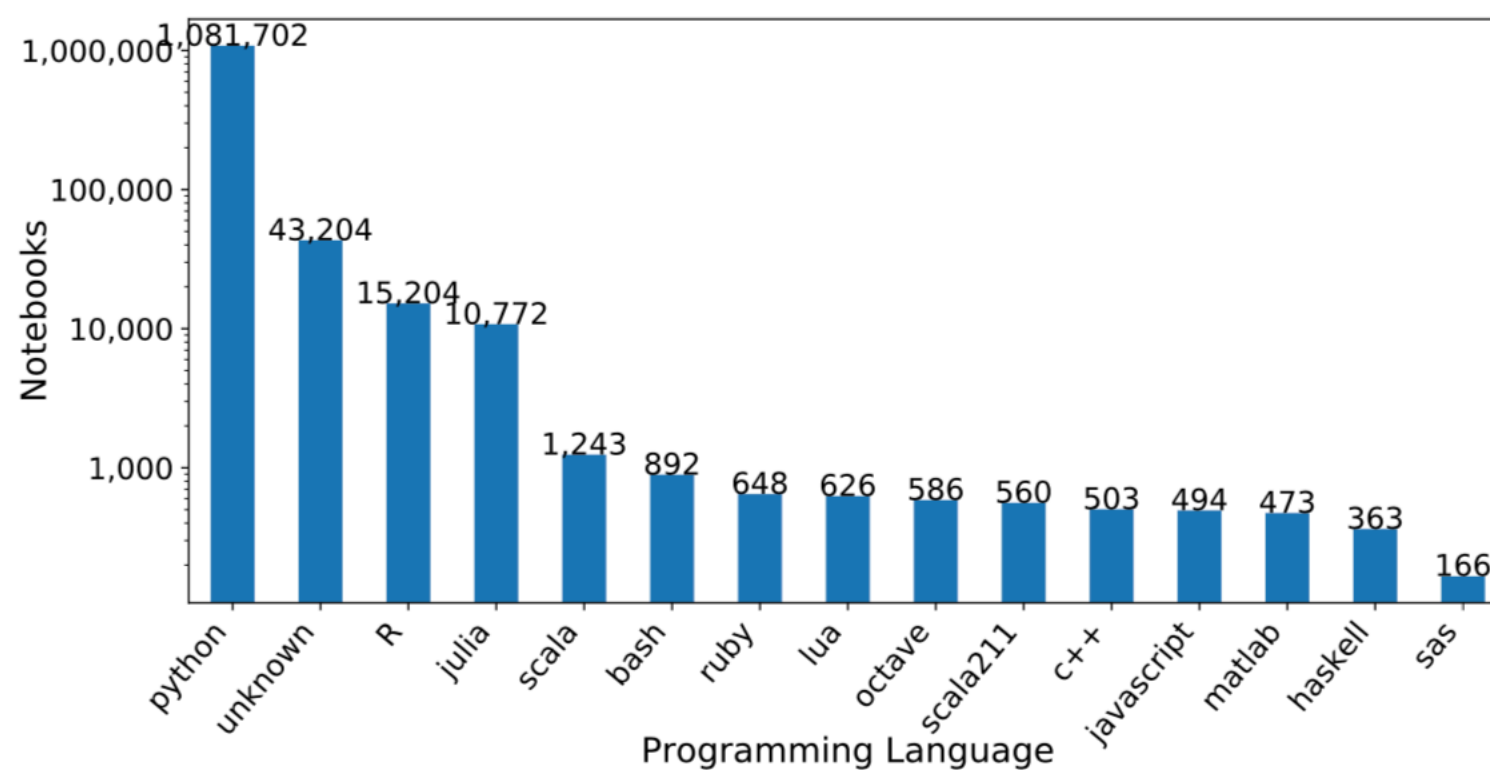


Fig. 3. Top 15 most declared programming languages. Notebooks axis in logarithmic scale.