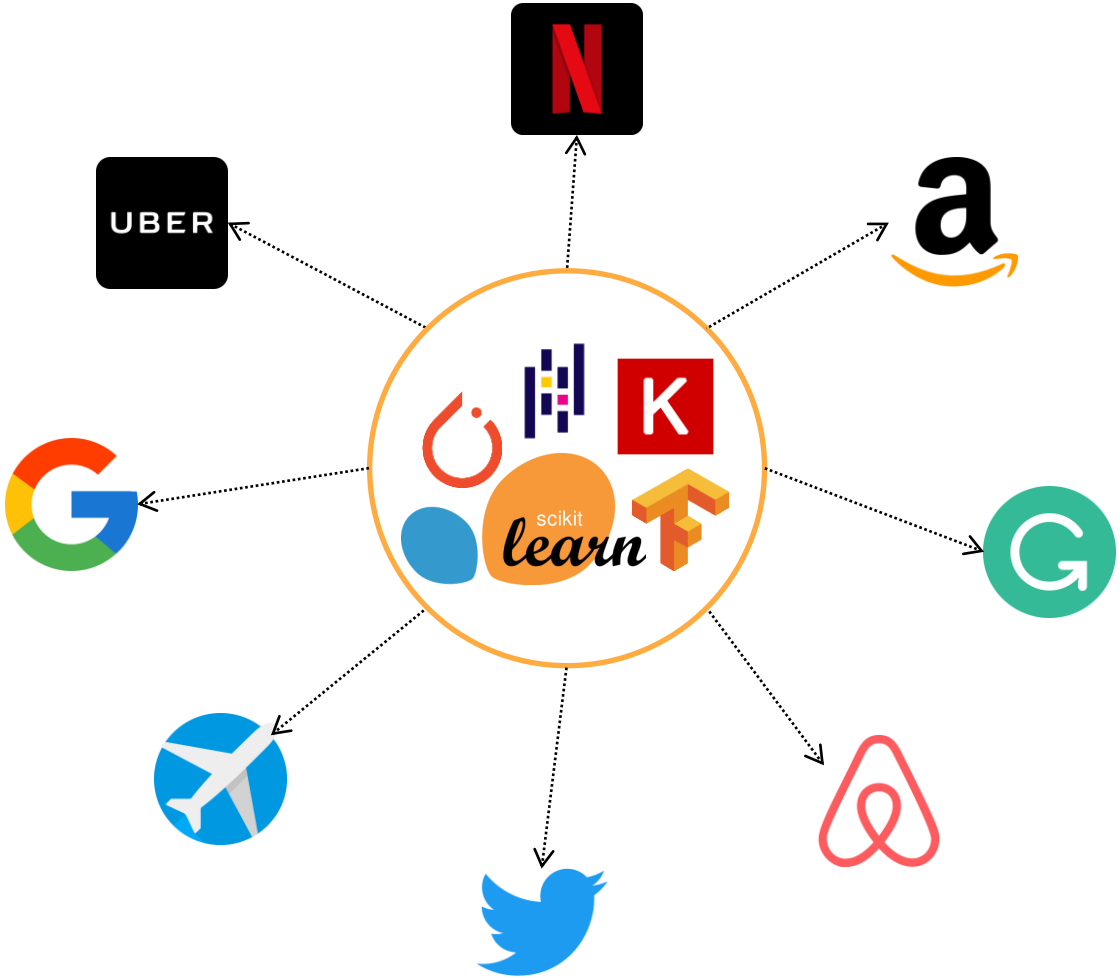
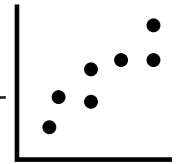
Unboxing Default Argument Breaking Changes in Scikit Learn

João Eduardo Montandon, Luciana Lourdes Silva, Cristiano Politowski, Ghizlane El Boussaidi, Marco Tulio Valente

23rd IEEE International Working Conference on Source Code Analysis and Manipulation


UBER
scikit
learn

```python
from sklearn.svm import SVC

X_train, X_test, y_train, y_test = ...

clf = SVC() # Just this?

clf.fit(X_train, y_train)
y_pred = clf.predict(X_test)
```



15 parameters!

```
class SVC(BaseSVC):
    """ ... """
    def __init__(
        self, *, C=1.0, kernel="rbf", degree=3, gamma="auto",
        coef0=0.0, shrinking=True, probability=False, tol=1e-3,
        cache_size=200, class_weight=None, verbose=False, max_iter=-1,
        decision_function_shape="ovr", break_ties=False, random_state=None
    ):
```

```
clf = SVC() # How?
```

```
class SVC(BaseSVC):
    """ ... """
    def __init__(
        self, *, C=1.0, kernel="rbf", degree=3, gamma="auto",
        coef0=0.0, shrinking=True, probability=False, tol=1e-3,
        cache_size=200, class_weight=None, verbose=False, max_iter=-1,
        decision_function_shape="ovr", break_ties=False, random_state=None
    ):
```

Default Argument Values!

```
clf = SVC() # SWEEET!
```

Changed in version 0.22: The default value of `gamma` changed from 'auto' to 'scale'.

```
class SVC(BaseSVC):
    """ ... """
    def __init__(
        self, *, C=1.0, kernel="rbf", degree=3, gamma="scale",
        coef0=0.0, shrinking=True, probability=False, tol=1e-3,
        cache_size=200, class_weight=None, verbose=False, max_iter=-1,
        decision_function_shape="ovr", break_ties=False, random_state=None
    ):
```

```
clf = SVC() # Wait... NOOOOO!
```

Changed in version 0.22: The default value of `gamma` changed from 'auto' to 'scale'.

```
class SVC(BaseSVC):
    """ ... """
    def __init__(
        self, *, C=1.0, kernel="rbf", degree=3, gamma="scale",
        coef0=0.0, shrinking=True, probability=False, tol=1e-3,
        cache_size=200, class_weight=None, verbose=False, max_iter=-1,
        decision_function_shape="ovr", break_ties=False, random_state=None
    ):
```

Default Argument Breaking Change (DABC)

```
clf = SVC() # Wait... NOOOOO!
```

Investigate the occurrence of DABCs in Scikit-Learn

Library

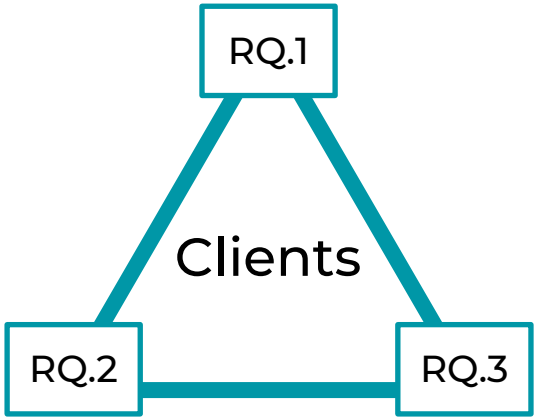
RQ.1: What are the Most Common DABCs?

RQ.2: In which version the DABCs were introduced?

RQ.3: In which modules the DABCs were introduced?

Clients

RQ.4: How clients are vulnerable to DABCs?



Library

On October 11th, 2022.

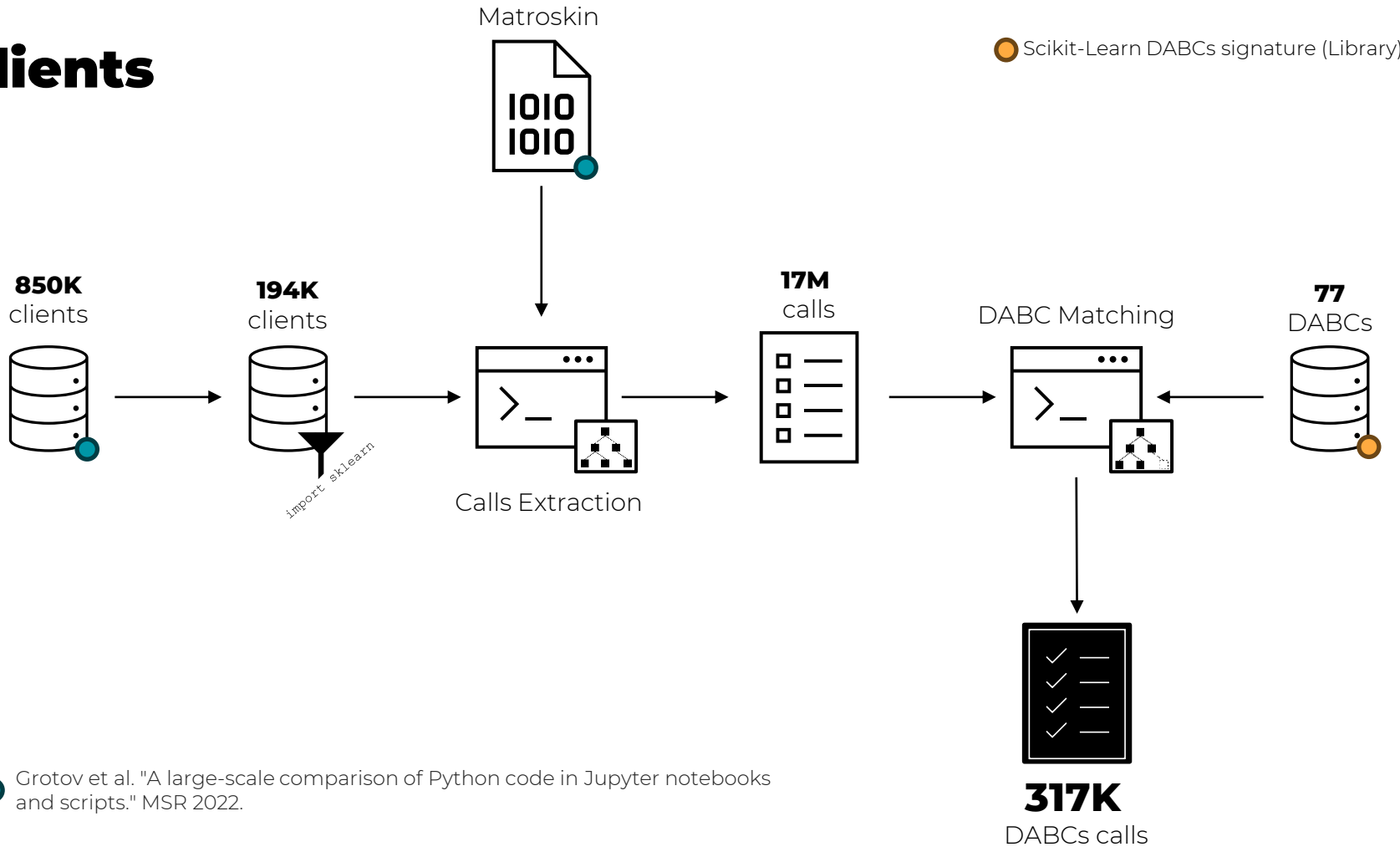
1. Cloned scikit-learn repository
2. Checked-out version 1.1.2 (latest public release)
3. Looked for `/\.\.versionchanged:: .+/` regex
4. Manually removed unrelated ones

```
class SVC(BaseSVC):
    """C-Support Vector Classification.
    ...
    Parameters
    ----------
    ...
    gamma : {'scale', 'auto'} or float, default='scale'
        Kernel coefficient for 'rbf', 'poly' and 'sigmoid'.

        - if ``gamma='scale'`` (default) is passed then it uses
          1 / (n_features * X.var()) as value of gamma,
        - if 'auto', uses 1 / n_features
        - if float, must be non-negative.

        .. versionchanged:: 0.22
           The default value of ``gamma`` changed from 'auto' to 'scale'.
    """
```

Library

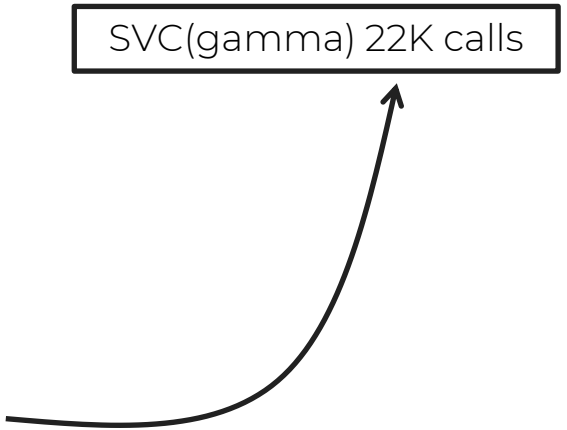
- **[RQ.1]** 77 DABCs in total.
- **[RQ.1]** 56 DABCs in class constructors.
- **[RQ.2]** DABCs in all majors between 0.19 and 1.1.
- **[RQ.3]** Model Training and Evaluation gather 79% of all DABCs

Clients



Grotov et al. "A large-scale comparison of Python code in Jupyter notebooks and scripts." MSR 2022.

Clients



- 10 DABCs concentrate 250K (79%)DABC calls.
- 67K (35%) clients are vulnerable.
- 312K (98%) calls deal with Model Training and Evaluation.

To Takeaway

- **[Researchers]** DABCs are a reality.
- **[Maintainers]** Default Arguments updates should follow strict version guidelines.
- **[Clients]** Avoid relying on default values.

jemaf@ufmg.br

Unboxing Default Argument Breaking Changes in Scikit Learn (SCAM 2023)