ï»¿.. _roadmap:

Roadmap

Purpose of this document

This document list general directions that core contributors are interested to see developed in scikit-learn. The fact that an item is listed here is in no way a promise that it will happen, as resources are limited. Rather, it is an indication that help is welcomed on this topic.

Statement of purpose: Scikit-learn in 2018

Eleven years after the inception of Scikit-learn, much has changed in the world of machine learning. Key changes include:

- Computational tools: The exploitation of GPUs, distributed programming frameworks like Scala/Spark, etc.
- High-level Python libraries for experimentation, processing and data management: Jupyter notebook, Cython, Pandas, Dask, Numba...
- Changes in the focus of machine learning research: artificial intelligence applications (where input structure is key) with deep learning, representation learning, reinforcement learning, domain transfer, etc.

A more subtle change over the last decade is that, due to changing interests in ML, PhD students in machine learning are more likely to contribute to PyTorch, Dask, etc. than to Scikit-learn, so our contributor pool is very different to a decade ago.

Scikit-learn remains very popular in practice for trying out canonical machine learning techniques, particularly for applications in experimental science and in data science. A lot of what we provide is now very mature. But it can be costly to maintain, and we cannot therefore include arbitrary new implementations. Yet Scikit-learn is also essential in defining an API framework for the development of interoperable machine learning components external to the core library.

Thus our main goals in this era are to:

- continue maintaining a high-quality, well-documented collection of canonical tools for data processing and machine learning within the current scope (i.e. rectangular data largely invariant to column and row order; predicting targets with simple structure)
- improve the ease for users to develop and publish external components
- improve interoperability with modern data science tools (e.g. Pandas, Dask) and infrastructures (e.g. distributed processing)

Many of the more fine-grained goals can be found under the API tag on the issue tracker.

Architectural / general goals

The list is numbered not as an indication of the order of priority, but to make referring to specific points easier. Please add new entries only at the bottom. Note that the crossed out entries are already done, and we try to keep the document up to date as we work on these issues.

1. Improved handling of Pandas DataFrames
    - document current handling
    - column reordering issue :issue:`7242`

      System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 72); backlink

      Unknown interpreted text role "issue".

    - avoiding unnecessary conversion to ndarray ~~:issue:`12147`~~

      System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 73); backlink

      Unknown interpreted text role "issue".

    - returning DataFrames from transformers :issue:`5523`

      System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 74); backlink

      Unknown interpreted text role "issue".

- getting DataFrames from dataset loaders ~~:issue:`10733`~~, ~~:issue:`13902`~~

  System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 75); *backlink*

  Unknown interpreted text role "issue".

  System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 75); *backlink*

  Unknown interpreted text role "issue".

- Sparse currently not considered ~~:issue:`12800`~~

  System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 77); *backlink*

  Unknown interpreted text role "issue".

2. Improved handling of categorical features

   - Tree-based models should be able to handle both continuous and categorical features :issue:`12866` and ~~:issue:`15550`~~.

     System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 81); *backlink*

     Unknown interpreted text role "issue".

     System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 81); *backlink*

     Unknown interpreted text role "issue".

   - ~~In dataset loaders :issue:`13902`~~

     System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 83); *backlink*

     Unknown interpreted text role "issue".

   - As generic transformers to be used with ColumnTransforms (e.g. ordinal encoding supervised by correlation with target variable) :issue:`5853`, :issue:`11805`

     System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 84); *backlink*

     Unknown interpreted text role "issue".

     System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 84); *backlink*

     Unknown interpreted text role "issue".

   - Handling mixtures of categorical and continuous variables

3. Improved handling of missing data

   - Making sure meta-estimators are lenient towards missing data, ~~:issue:`15319`~~

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 91); *backlink*

Unknown interpreted text role "issue".

- Non-trivial imputers ~~:issue:`11977`, :issue:`12852`~~

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 93); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 93); *backlink*

Unknown interpreted text role "issue".

- Learners directly handling missing data ~~:issue:`13911`~~

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 94); *backlink*

Unknown interpreted text role "issue".

- An amputation sample generator to make parts of a dataset go missing :issue:`6284`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 95); *backlink*

Unknown interpreted text role "issue".

4. More didactic documentation
   - More and more options have been added to scikit-learn. As a result, the documentation is crowded which makes it hard for beginners to get the big picture. Some work could be done in prioritizing the information.
5. Passing around information that is not (X, y): Sample properties
   - We need to be able to pass sample weights to scorers in cross validation.
   - We should have standard/generalised ways of passing sample-wise properties around in meta-estimators. :issue:`4497` :issue:`7646`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 107); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 107); *backlink*

Unknown interpreted text role "issue".

6. Passing around information that is not (X, y): Feature properties
   - Feature names or descriptions should ideally be available to fit for, e.g. . :issue:`6425` :issue:`6424`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 112); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 112); *backlink*

Unknown interpreted text role "issue".

- Per-feature handling (e.g. "is this a nominal / ordinal / English language text?") should also not need to be provided to estimator constructors, ideally, but should be available as metadata alongside X. :issue:`8480`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 114); *backlink*

Unknown interpreted text role "issue".

7. Passing around information that is not (X, y): Target information
   - We have problems getting the full set of classes to all components when the data is split/sampled. :issue:`6231` :issue:`8100`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 120); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 120); *backlink*

Unknown interpreted text role "issue".

   - We have no way to handle a mixture of categorical and continuous targets.
8. Make it easier for external users to write Scikit-learn-compatible components
   - More flexible estimator checks that do not select by estimator name ~~:issue:`6599`~~ :issue:`6715`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 127); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 127); *backlink*

Unknown interpreted text role "issue".

   - Example of how to develop an estimator or a meta-estimator, ~~:issue:`14582`~~

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 129); *backlink*

Unknown interpreted text role "issue".

   - More self-sufficient running of scikit-learn-contrib or a similar resource
9. Support resampling and sample reduction
   - Allow subsampling of majority classes (in a pipeline?) :issue:`3855`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 135);

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 136); *backlink*

Unknown interpreted text role "issue".

- Implement random forests with resampling :issue:`13227`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 136); *backlink*

Unknown interpreted text role "issue".

10. Better interfaces for interactive development
    - ~~__repr__ and HTML visualisations of estimators :issue:`6323` and :pr:`14180`~~.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 140); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 140); *backlink*

Unknown interpreted text role "pr".

    - Include plotting tools, not just as examples. :issue:`9173`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 142); *backlink*

Unknown interpreted text role "issue".

11. Improved tools for model diagnostics and basic inference
    - ~~alternative feature importances implementations, :issue:`13146`~~

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 146); *backlink*

Unknown interpreted text role "issue".

    - better ways to handle validation sets when fitting
    - better ways to find thresholds / create decision rules :issue:`8614`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 148); *backlink*

Unknown interpreted text role "issue".

12. Better tools for selecting hyperparameters with transductive estimators
    - Grid search and cross validation are not applicable to most clustering tasks. Stability-based selection is more relevant.
13. Better support for manual and automatic pipeline building
    - Easier way to construct complex pipelines and valid search spaces :issue:`7608` :issue:`5082` :issue:`8243`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 157); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 157); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 157); *backlink*

Unknown interpreted text role "issue".

- provide search ranges for common estimators??
- cf. searchgrid

14. Improved tracking of fitting
    - Verbose is not very friendly and should use a standard logging library :issue:`6929`, :issue:`78`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 164); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 164); *backlink*

Unknown interpreted text role "issue".

    - Callbacks or a similar system would facilitate logging and early stopping
15. Distributed parallelism
    - Accept data which complies with `__array_function__`
16. A way forward for more out of core
    - Dask enables easy out-of-core computation. While the Dask model probably cannot be adaptable to all machine-learning algorithms, most machine learning is on smaller data than ETL, hence we can maybe adapt to very large scale while supporting only a fraction of the patterns.
17. Support for working with pre-trained models
    - Estimator "freezing". In particular, right now it's impossible to clone a *CalibratedClassifierCV* with prefit. :issue:`8370`. :issue:`6451`

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 181); *backlink*

Unknown interpreted text role "issue".

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 181); *backlink*

Unknown interpreted text role "issue".

18. Backwards-compatible de/serialization of some estimators
    - Currently serialization (with pickle) breaks across versions. While we may not be able to get around other limitations of pickle re security etc, it would be great to offer cross-version safety from version 1.0. Note: Gael and Olivier think that this can cause heavy maintenance burden and we should manage the trade-offs. A possible alternative is presented in the following point.
19. Documentation and tooling for model lifecycle management
    - Document good practices for model deployments and lifecycle: before deploying a model: snapshot the code versions

(numpy, scipy, scikit-learn, custom code repo), the training script and an alias on how to retrieve historical training data + snapshot a copy of a small validation set + snapshot of the predictions (predicted probabilities for classifiers) on that validation set.
    - Document and tools to make it easy to manage upgrade of scikit-learn versions:
        - Try to load the old pickle, if it works, use the validation set prediction snapshot to detect that the serialized model still behave the same;
        - If joblib.load / pickle.load not work, use the versioned control training script + historical training set to retrain the model and use the validation set prediction snapshot to assert that it is possible to recover the previous predictive performance: if this is not the case there is probably a bug in scikit-learn that needs to be reported.

20. Everything in Scikit-learn should probably conform to our API contract. We are still in the process of making decisions on some of these related issues.
    - *Pipeline <pipeline.Pipeline>* and *FeatureUnion* modify their input parameters in fit. Fixing this requires making sure we have a good grasp of their use cases to make sure all current functionality is maintained. :issue:`8157` :issue:`7382`

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 217); *backlink*
>
> Unknown interpreted text role "issue".

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 217); *backlink*
>
> Unknown interpreted text role "issue".

21. (Optional) Improve scikit-learn common tests suite to make sure that (at least for frequently used) models have stable predictions across-versions (to be discussed);
    - Extend documentation to mention how to deploy models in Python-free environments for instance ONNX. and use the above best practices to assess predictive consistency between scikit-learn and ONNX prediction functions on validation set.
    - Document good practices to detect temporal distribution drift for deployed model and good practices for re-training on fresh data without causing catastrophic predictive performance regressions.

Subpackage-specific goals

:mod:`sklearn.ensemble`

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 238); *backlink*
>
> Unknown interpreted text role "mod".

- ~~a stacking implementation, :issue:`11047`~~

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 240); *backlink*
>
> Unknown interpreted text role "issue".

:mod:`sklearn.cluster`

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 242); *backlink*
>
> Unknown interpreted text role "mod".

- kmeans variants for non-Euclidean distances, if we can show these have benefits beyond hierarchical clustering.

:mod:`sklearn.model_selection`

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst, line 247); *backlink*

Unknown interpreted text role "mod".

- ~~multi-metric scoring is slow :issue:`9326`~~

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst`, **line 249**); ***backlink***

  Unknown interpreted text role "issue".

- perhaps we want to be able to get back more than multiple metrics
- the handling of random states in CV splitters is a poor design and contradicts the validation of similar parameters in estimators, SLEP011
- exploit warm-starting and path algorithms so the benefits of *EstimatorCV* objects can be accessed via *GridSearchCV* and used in Pipelines. :issue:`1626`

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst`, **line 254**); ***backlink***

  Unknown interpreted text role "issue".

- Cross-validation should be able to be replaced by OOB estimates whenever a cross-validation iterator is used.
- Redundant computations in pipelines should be avoided (related to point above) cf daskml

:mod:`sklearn.neighbors`

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst`, **line 263**); ***backlink***

Unknown interpreted text role "mod".

- ~~Ability to substitute a custom/approximate/precomputed nearest neighbors implementation for ours in all/most contexts that nearest neighbors are used for learning. :issue:`10463`~~

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst`, **line 265**); ***backlink***

  Unknown interpreted text role "issue".

:mod:`sklearn.pipeline`

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\scikit-learn-main\doc\[scikit-learn-main][doc]roadmap.rst`, **line 269**); ***backlink***

Unknown interpreted text role "mod".

- Performance issues with *Pipeline.memory*
- see "Everything in Scikit-learn should conform to our API contract" above