Learning Scikit-Learn

Qiyang Hu
UCLA Office of Advanced Research Computing
July 28th, 2023

Outline

- Learning Scikit-learn (the basics)
  - High-level overview of scikit-learn libraries
    - According to a typical machine learning workflow
    - A lot of colab snippets as examples
  - Practical usage of scikit-learn
    - Deep learning and scikit-learn
    - Scikit-learn extension libraries

Today

- High-performance machine learning using scikit-learn
  - Overview of performance issues in machine learning
  - Making the computation faster
  - Processing large dataset

What can/can't be expected in the series?

✓ CAN	✗ CAN'T
• Review on Machine learning workflows	• Introduction to various Machine Learning models
• A ***BIG*** picture on scikit-learn's features, functions & components	• Discussions on the details of specific scikit-learn function interfaces
• Providing handy examples as demos (mainly for studying *after* the class)	• Line-by-line explanation on every demo code
• High-level introduction on high performance machine learning	• Lectures on detailed mechanism and implementations of HPML.

Why learning Scikit-Learn in the LLM era?

Scikit-Learn can help us

- Understand ML Models/Concepts
- Provide chains of thoughts
- Nail down the problem quickly

Beyond the Zero/Few Shots

Better & Efficient Prompts

Current LLMs cannot do everything yet.

- Needs to fix the hallucination problem.
- Needs to distill the domain knowledge
- Needs to finetune the pre-trained model (will mention a case example today.)

Learning Scikit-Learn in 5 minutes

- A Python machine learning framework
  - Library built on numpy, scipy, matplotlib
    - Started in 2007, publicly released in 2010
    - Is currently maintained by volunteers
- Installation/Loading
  - `conda install scikit-learn-intelex`
  - On H2: `module load anaconda3` `conda activate sklearn`
  - Using Google Colab
- Designed for easy-to-use productions
  - Simplicity
  - Qualitative code
    - Performance
    - Elegant APIs
  - Excellent docs: https://scikit-learn.org

```
import sklearn

# 2 samples, 3 features
X = [[ 1,  2,  3],
     [11, 12, 13]]

# classes of each sample
y = [0, 1]
```

Data

Modeling

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier(random_state=0)

clf.fit(X, y)
```

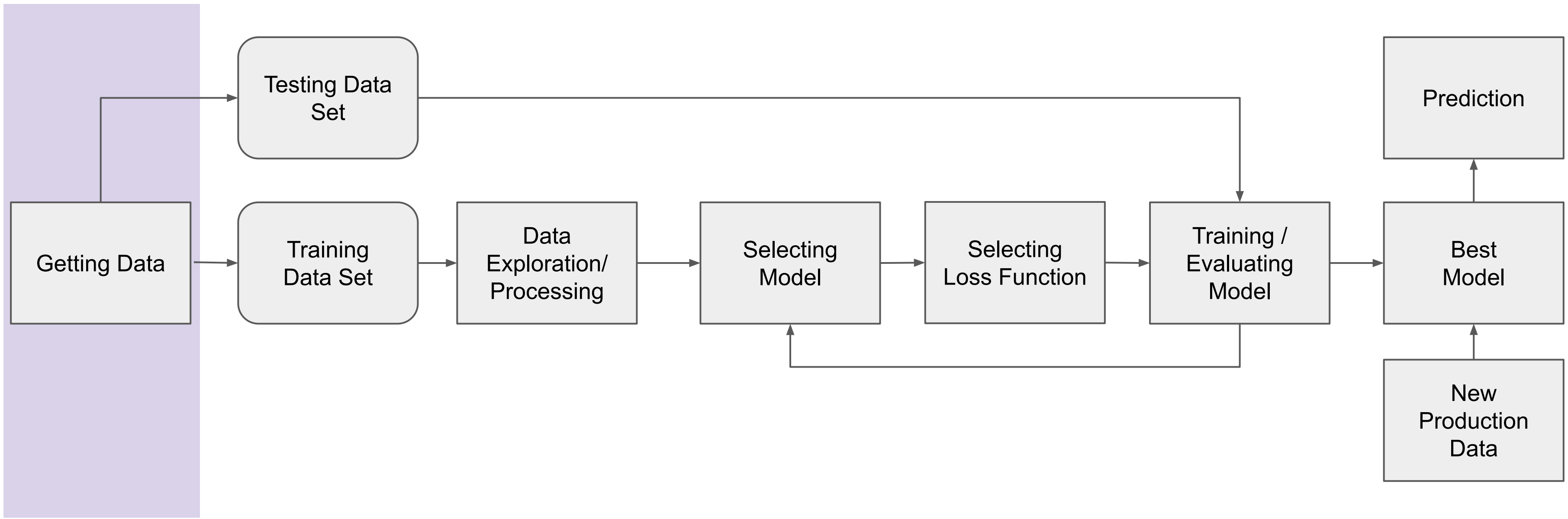
```
# predict classes of the training data
clf.predict(X)

# predict classes of new data
clf.predict([[4, 5, 6], [14, 15, 16]])
```

Predicting

Outline

- Learning Scikit-learn
  - High-level overview of scikit-learn libraries
    - According to a typical machine learning workflow
    - A lot of colab snippets as examples
  - Practical usage of scikit-learn
    - Deep learning and scikit-learn
    - Scikit-learn extension libraries
- High-performance machine learning using scikit-learn
  - Overview of performance issues in machine learning
  - Making the computation faster
  - Processing large dataset

Simplified workflow for a machine learning project



Data input and data loader

- Data format can be input directly as:
    - Dense data: numpy.ndarray
    - Sparse data: scipy.sparse.matrix
- Data can be loaded from standard datasets:
    - Toy datasets
    - Real world datasets

Loaders and Fetchers returns a dictionary-like **bunch** object

```
>>> b = Bunch(a=1, b=2)
>>> b['b']
2
>>> b.b
2
>>> b.a = 3
>>> b['a']
3
>>> b.c = 6
>>> b['c']
6
```

sklearn.datasets.**load**_

- boston
- iris
- diabetes
- digits
- linnerud
- wine
- breast_cancer
- sample_image[*s*]

sklearn.datasets.**fetch**_

- olivetti_faces
- 20newsgroups[_*vectorized*]
- lfw_[people/pairs]
- covtype
- rcv1
- kddcup99
- california_housing
- openml

Data Generator

sklearn.datasets.**make**_

For classification and clustering:

- blob
- classification
- gaussian_quantiles
- hastie_10_2
- circles
- moons
- multilabel_classification
- biclusters
- checkerboard

```
(n_samples=100, n_features=2, *,
centers=None, cluster_std=1.0,
center_box=- 10.0, 10.0,
shuffle=True, random_state=None,
return_centers=False)
```

For regression:

- regression
- friedman[*1/2/3*]
- sparse_uncorrelated

For manifold learning:

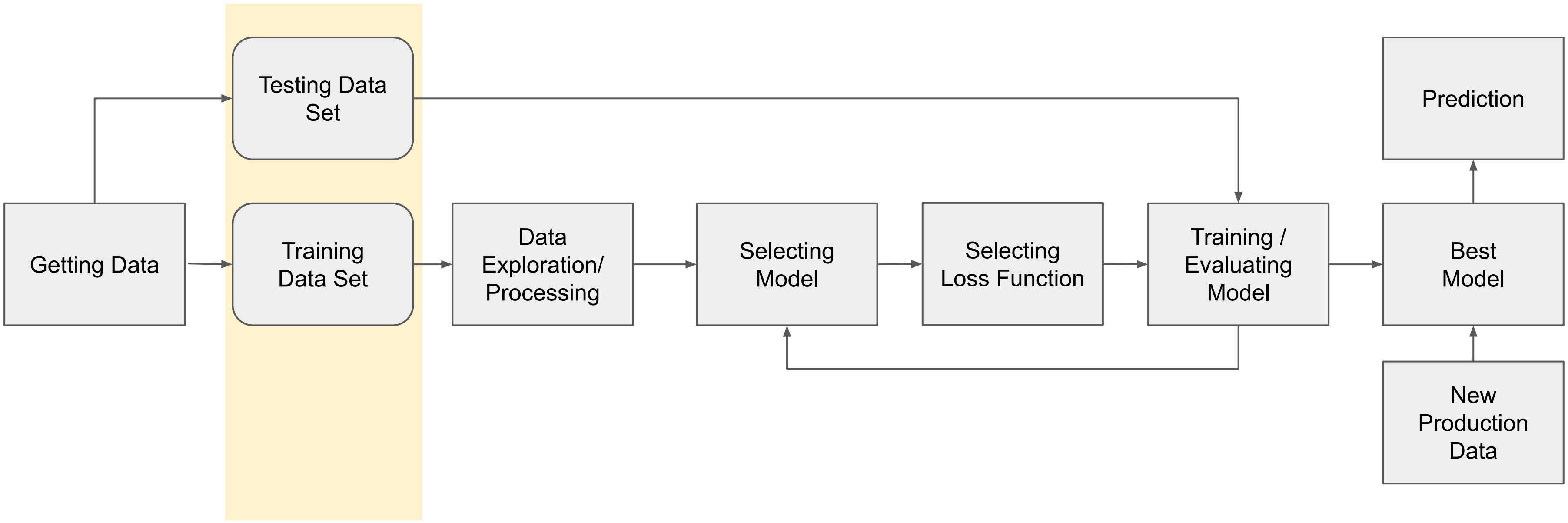
- s_curve
- swiss_roll

For decomposition:

- low_rank_matrix
- sparse_coded_signal
- spd_matrix
- sparse_spd_matrix

bit.ly/lskl_01

Workflow for a machine learning project

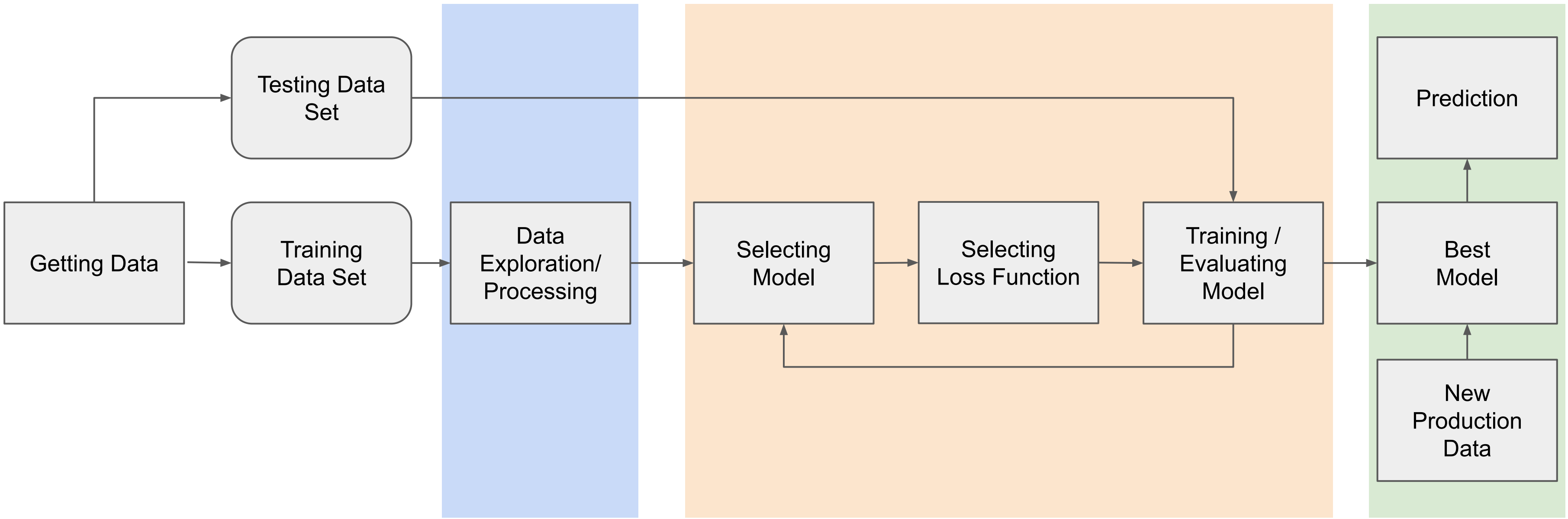


Split training and testing dataset

- Essential for an unbiased evaluation of prediction performance
  - Process related with model evaluation and selection
  - Internally, scikit-learn uses cross-validation iterators to split
- Multiple splitting methods
  - Stratified splitting
  - Group splitting
  - Time series splitting
  - Predefined splitting
- Sklearn's `train_test_split`
  - A wrapper of a single-call ShuffleSplit
  - Only allows for stratified splitting
  - As a base for the default cross-validations

```
>>> import numpy as np
>>> from sklearn.model_selection import train_test_split
>>> from sklearn import datasets
>>> from sklearn import svm

>>> X, y = datasets.load_iris(return_X_y=True)
>>> X.shape, y.shape
((150, 4), (150,))
>>> X_train, X_test, y_train, y_test = train_test_split(
...     X, y, test_size=0.4, random_state=0)

>>> X_train.shape, y_train.shape
((90, 4), (90,))
>>> X_test.shape, y_test.shape
((60, 4), (60,))
```

Workflow for a machine learning project



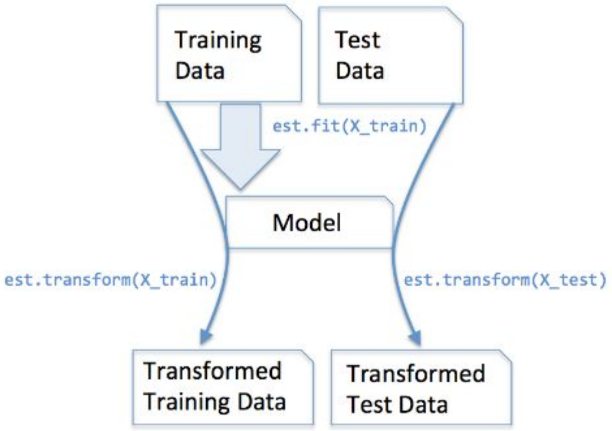
Core objects (estimators)

Preprocessing, Feature extraction

`fit()` calculates the parameters μ and σ and saves them as internal objects.

Transformer

```
transformer.transform(data)
transformer.fit_transform(data)
```



Estimator

```
estimator.fit(data)
estimator.fit(data, targets)
estimator.partial_fit(data,targets)
```

Predictor

```
predictor.predict(data)
```

For Classification Algorithms

- `predictor.predict_prob(data)`
- `predictor.decision_function(data)`

`fit()` fits a model based on some training data and is capable of inferring some properties on new data.

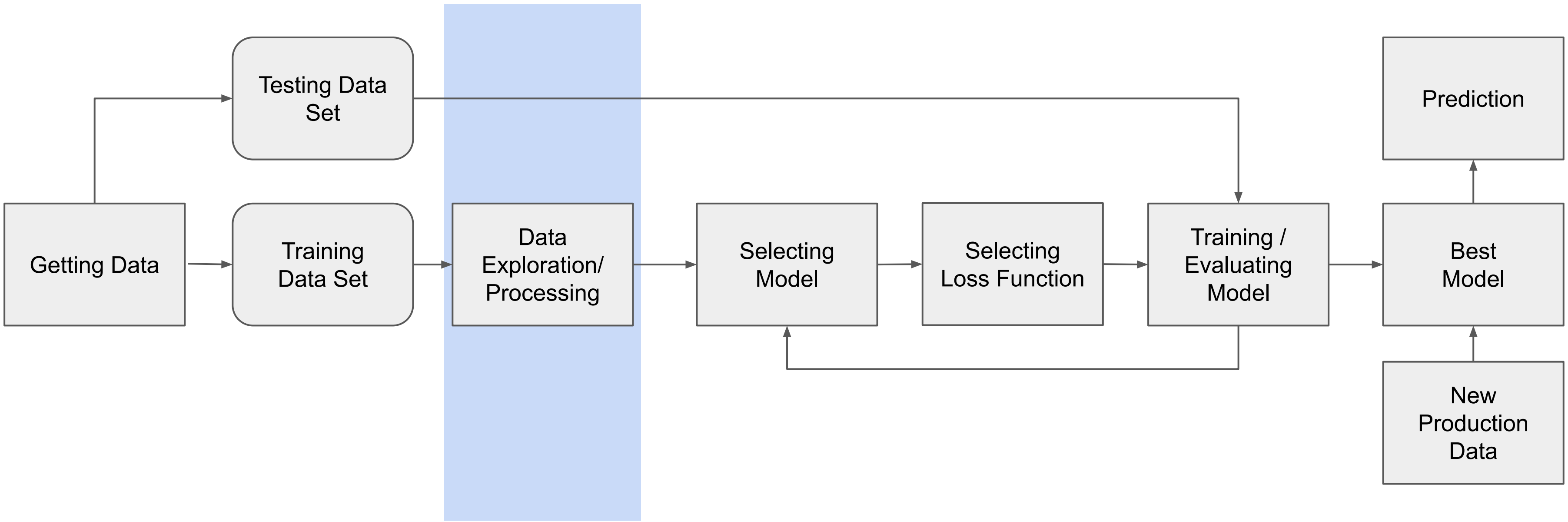
Model

```
model.score(data)
```

Modeling, Learning

Qiyang Hu

Workflow for a machine learning project



Qiyang Hu

Preprocessing:

```
from sklearn.preprocessing import
StandardScaler
sc = StandardScaler()
sc.fit_tranform(X_train)
sc.transform(X_test)
```

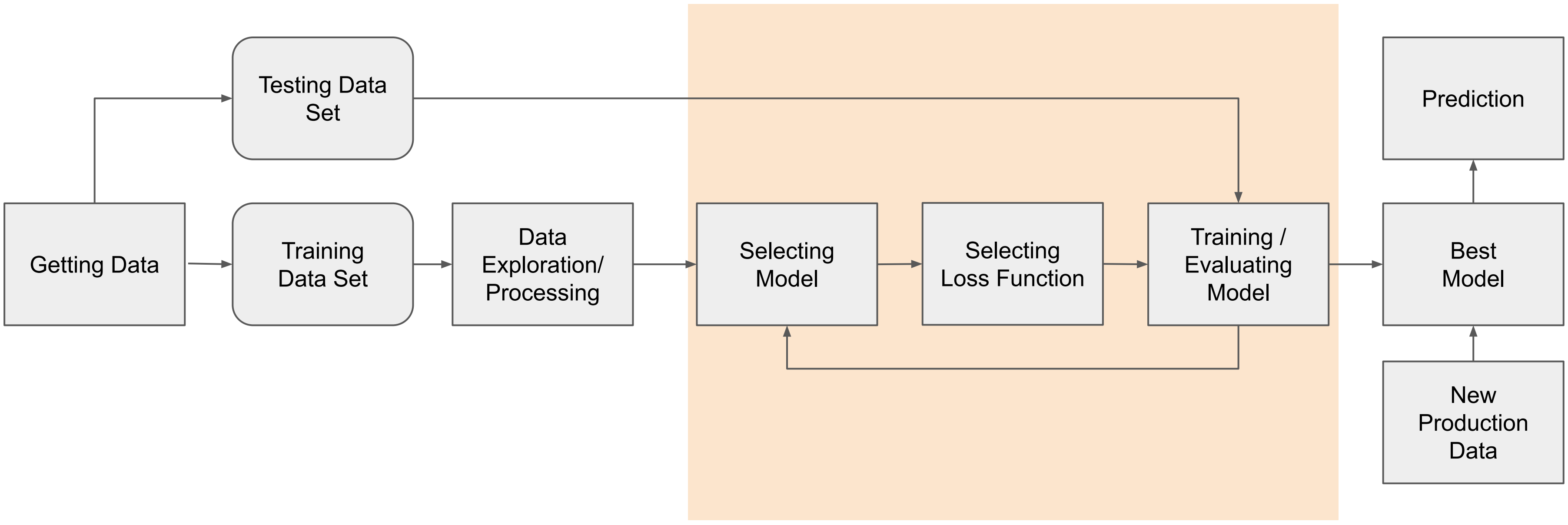
sklearn.**preprocessing**.

- Standardization, or mean removal and variance scaling
  - StandardScaler / RobustScaler
  - MinMaxScaler / MaxAbsScaler
  - KernelCenterer
- Non-linear transformation
  - QuantileTransformer
  - PowerTransformer
- Normalization
  - normalize
  - Normalizer
- Encoding categorical features
  - OrdinalEncoder/LabelEncoder
  - OneHotEncoder
- Discretization
  - KBinsDiscretizer
  - Binarizer
- Custom transformers
  - FunctionTransformer
  - PolynomialFeatures

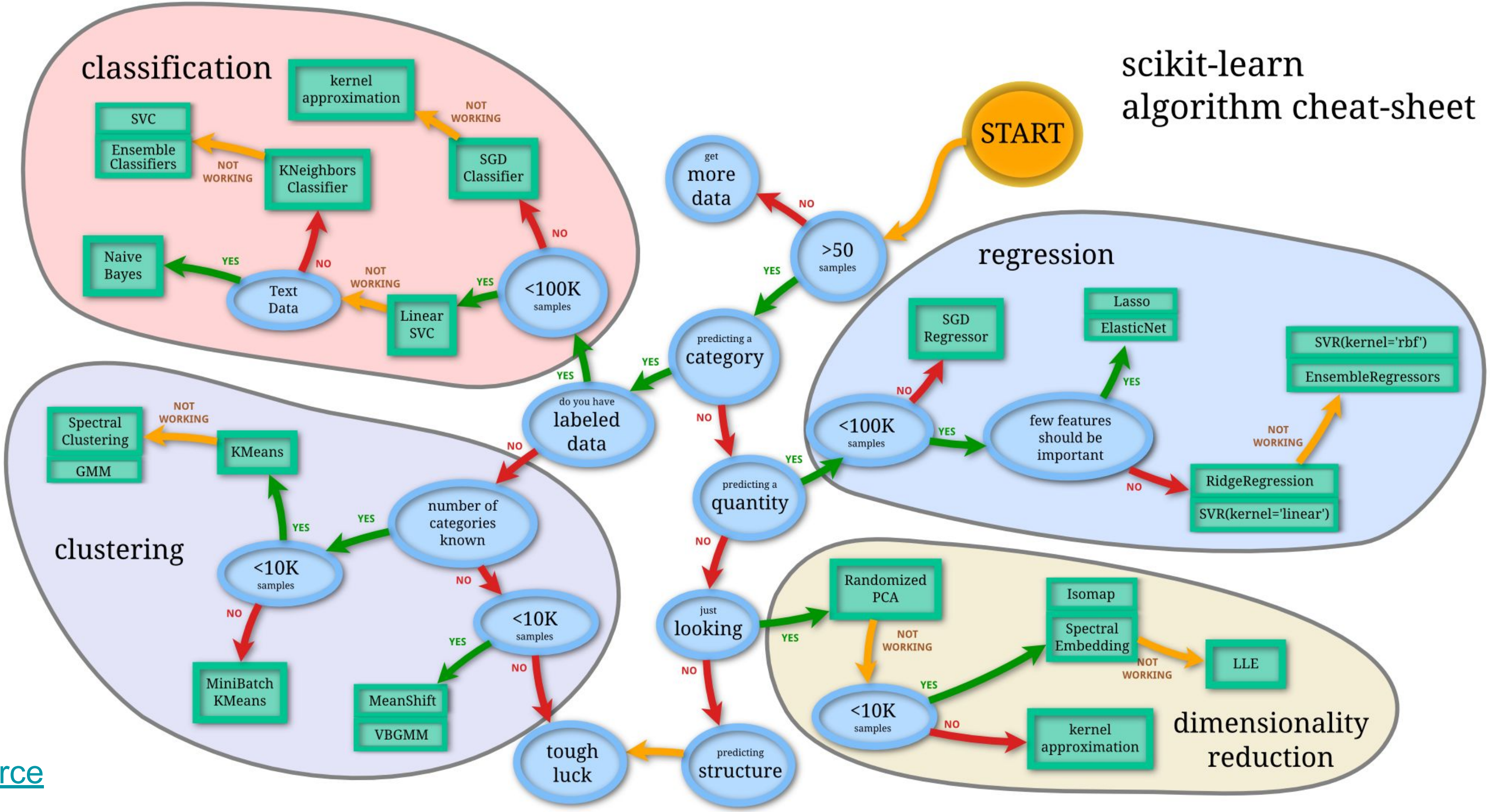
sklearn.**imput**.

- Imputation of missing values
  - SimpleImputer
  - IterativeImputer
  - KNNImputer
  - MissingIndicator

Workflow for a machine learning project



Choosing the right estimator (algorithm)



[Source](#)

Pseudo-code template for modeling and learning

```
                   ⎧ linear_model ⎫
                   ⎪ svm          ⎪
                   ⎪ tree         ⎪
                   ⎪ naive_bayes  ⎪
from sklearn.      ⎨ multioutput  ⎬     import  SpecModel
                   ⎪ ensemble     ⎪
                   ⎪ cluster      ⎪
                   ⎪ decomposition⎪
                   ⎩ ...          ⎭

model = SpecModel( hyperparameter )

model.fit( X, y )

y_pred = model.predict( X_new )

s = model.score( X_new )
```

```
penalty='l2', tol=0.0001, C=0.1,
fit_intercept=True,
solver='liblinear', max_iter=100,
multi_class='ovr', n_jobs=1,  ...
```

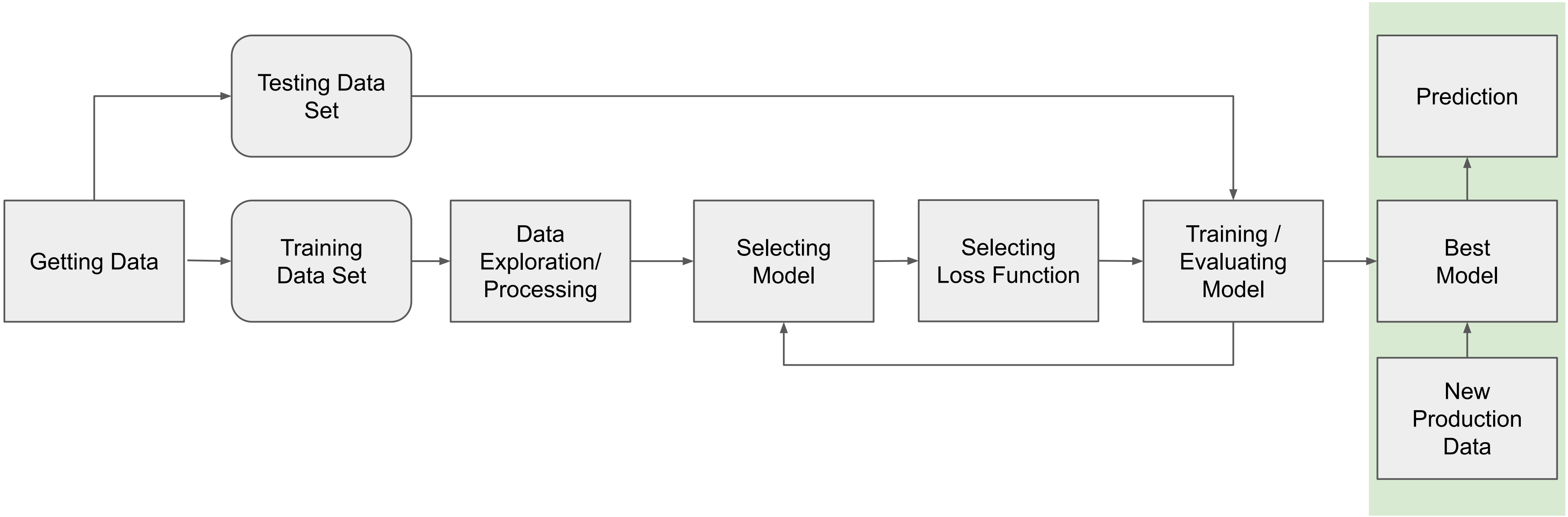
```
LogisticRegression
LogisticRegressionCV
PassiveAggressiveClassifier
Perceptron
RidgeClassifier
RidgeClassifierCV
SGDClassifier
LinearRegression
Ridge
RidgeCV
SGDRegressor
ElasticNet
ElasticNetCV
Lars
LarsCV
Lasso
LassoCV
LassoLars
LassoLarsCV
LassoLarsIC
OrthogonalMatchingPursuit
OrthogonalMatchingPursuitCV
ARDRegression
BayesianRidge
PoissonRegressor
GammaRegressor
HuberRegressor
RANSACRegressor
...
```

Meta-estimator: as an assembly of base estimators

- Internally accept base models as parameters

- Meta Estimators
  - ensemble
    - Averaging method
    - Boosting method
      - AdaBoostClassifier
      - AdaBoostRegressor
      - BaggingClassifier
      - BaggingRegressor
      - ExtraTreesClassifier
      - ExtraTreesRegressor
      - GradientBoostingClassifier
      - GradientBoostingRegressor
      - IsolationForest
      - RandomForestClassifier
      - RandomForestRegressor
      - RandomTreesEmbedding
      - StackingClassifier
      - StackingRegressor
      - VotingClassifier
      - VotingRegressor
      - HistGradientBoostingRegressor
      - HistGradientBoostingClassifier
  - multiclass
    - OneVsRestClassifier
    - OneVsOneClassifier
    - OutputCodeClassifier
  - multioutput
    - ClassifierChain
    - MultiOutputRegressor
    - MultiOutputClassifier
    - RegressorChain
  - model_selection
    - Splitter Classes/Functions
    - Hyper-param optimizers
    - Model Validation
      - StratifiedKFold
      - train_test_split
      - GridSearchCV
      - RandomizedSearchCV
      - cross_validate
      - cross_val_score
      - learning_curve
      - ...
  - pipeline
    - FeatureUnion
    - Pipeline
    - make_pipeline
    - make_union

Workflow for a machine learning project



Model persistence (saving/restoring a trained model)

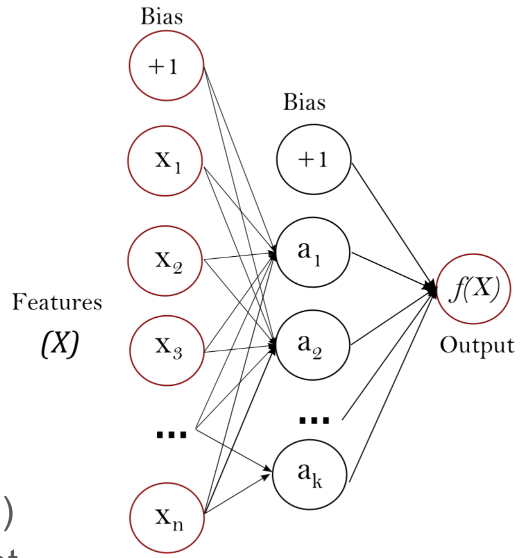
- Python’s built-in serialization:
  - Using `pickle` or `joblib`: dump and load
  - Custom transformers in Pipeline cannot be serialized by pickle or joblib
    - Consider using Neuralxle’s module to save custom pipeline in step wise
  - Pickled model better to be deployed using containers to avoid portability issues
  - A more secure format: skops
- Other exporting formats
  - Open Neural Network Exchange (ONNX)
    - sklearn-onnx
  - Predictive Model Markup Language (PMML)
    - sklearn2pmml

Outline

- Learning Scikit-learn
  - High-level overview of scikit-learn libraries
    - According to a typical machine learning workflow
    - A lot of colab snippets as examples
  - Practical usage of scikit-learn
    - Deep learning and scikit-learn
    - Scikit-learn extension libraries
- High-performance machine learning using scikit-learn
  - Overview of performance issues in machine learning
  - Making the computation faster
  - Processing large dataset

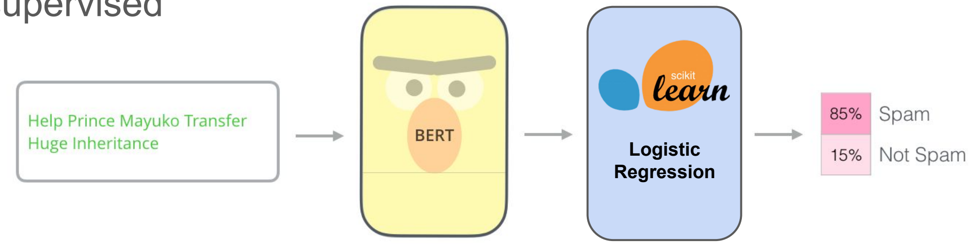
Deep Learning and Scikit-Learn

- Neural networks in scikit-learn
    - Multi-layer Perceptron: `MLPRegressor`, `MLPClassifier`
        - `from sklearn.neural_network import MLPClassifier`
        - `MLPClassifier(solver='lbfgs', alpha=1e-5,`
          `hidden_layer_sizes=(5, 2))`
    - Not versatile as Tensorflow and Pytorch, ***but***
        - Much simpler and straightforward (esp. for early-stopping)
        - Directly support sparse data as input, saving memory a lot
        - Natively support partial-fit (will discuss in next talk)!



bit.ly/3PuFTHw

- Works together with modern exotic deep learning models
    - Large NLP models: pre-trained, semi-supervised
    - Scikit-Learn for supervised fine-tuning
        - As a top classifier layer
        - Defining the workflow



Scikit-learn extension libraries

- Libraries adopting scikit-learn functionalities
  - Data formats: sklearn_pandas, sklear_xarray, ...
  - Auto-ML: auto-sklearn, Featuretools, Neuraxle, …
  - Model visualization: dtreeviz, eli5, …
  - Model selection: scikit-optimize, sklearn-deap, ...
  - Model export: onnxmltools, sklearn2pmml, …
  - Parallelization: sk-dist
  - Plotting: scikit-plot
- Libraries compatible with scikit-learn interfaces
  - Time-series models: tslearn, sktime, seglearn, …
  - Deep learning: keras, skorch, ...
  - Other regression/classification: xgboost, ML_Ensemble, gplearn, …
  - Decomposition and clustering: lda, hdbscan, ...

SciKits means a huge family of SciPy libraries

- scikit-learn
- scikit-opt
- scikit-image
- scikit-sparse
- scikit-statsmodels
- ...

Outline

- Learning Scikit-learn
  - High-level overview of scikit-learn libraries
    - According to a typical machine learning workflow
    - A lot of colab snippets as examples
  - Practical usage of scikit-learn
    - Deep learning and scikit-learn
    - Scikit-learn extension libraries
- High-performance machine learning using scikit-learn
  - Overview of performance issues in machine learning
  - Making the computation faster
  - Processing large dataset

See you Next Friday!

bit.ly/lskl_survey1

Backup Slides

Difference between SVC and SGD with hinge loss

- LinearSVC:
  - implemented in terms of liblinear
  - use the *full* data and solve a convex optimization problem with respect to these data points.
- SVC with kernel='linear' parameter
  - implemented in terms of libsvm
  - use the *full* data and solve a convex optimization problem with respect to these data points.
- Stochastic Gradient Descent with loss='hinge' parameter
  - treat the data in *batches* and performs a gradient descent aiming to minimize expected loss with respect to the sample distribution, assuming that the examples are iid samples of that distribution.
  - is typically used when the number of samples is very big or not ending. Observe that you can call the *partial_fit* function and feed it chunks of data.

Key Takeaways

- Learning Scikit-learn means:
  - Understanding the machine learning workflows
  - Learning its API conventions
  - Familiarizing its documentations
  - Inspecting the example codes
  - Considering the scikit-learn extensive libraries