

Data science for lazy people, Automated Machine Learning

PyConEs 2019

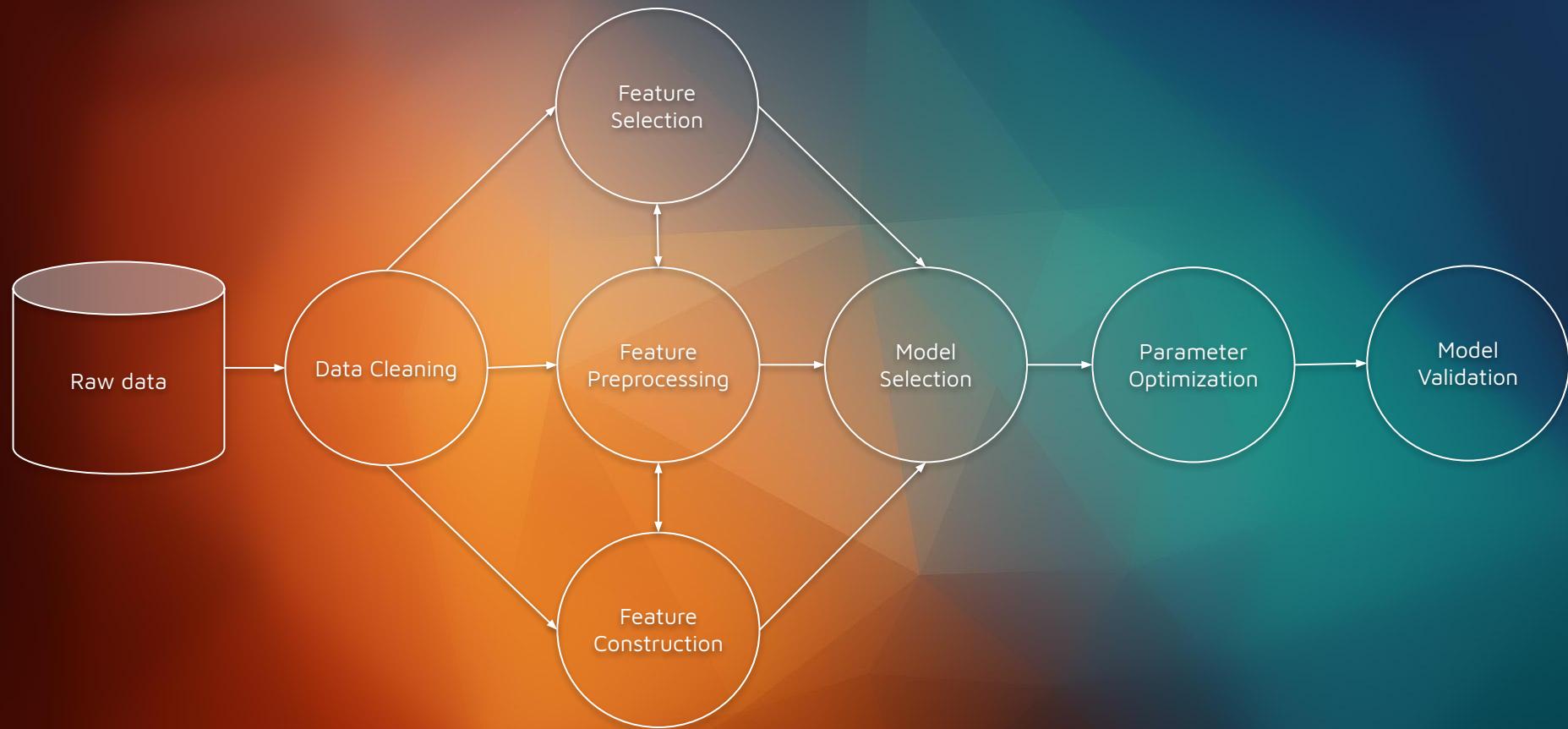
Diego Hueltes
Ravenpack

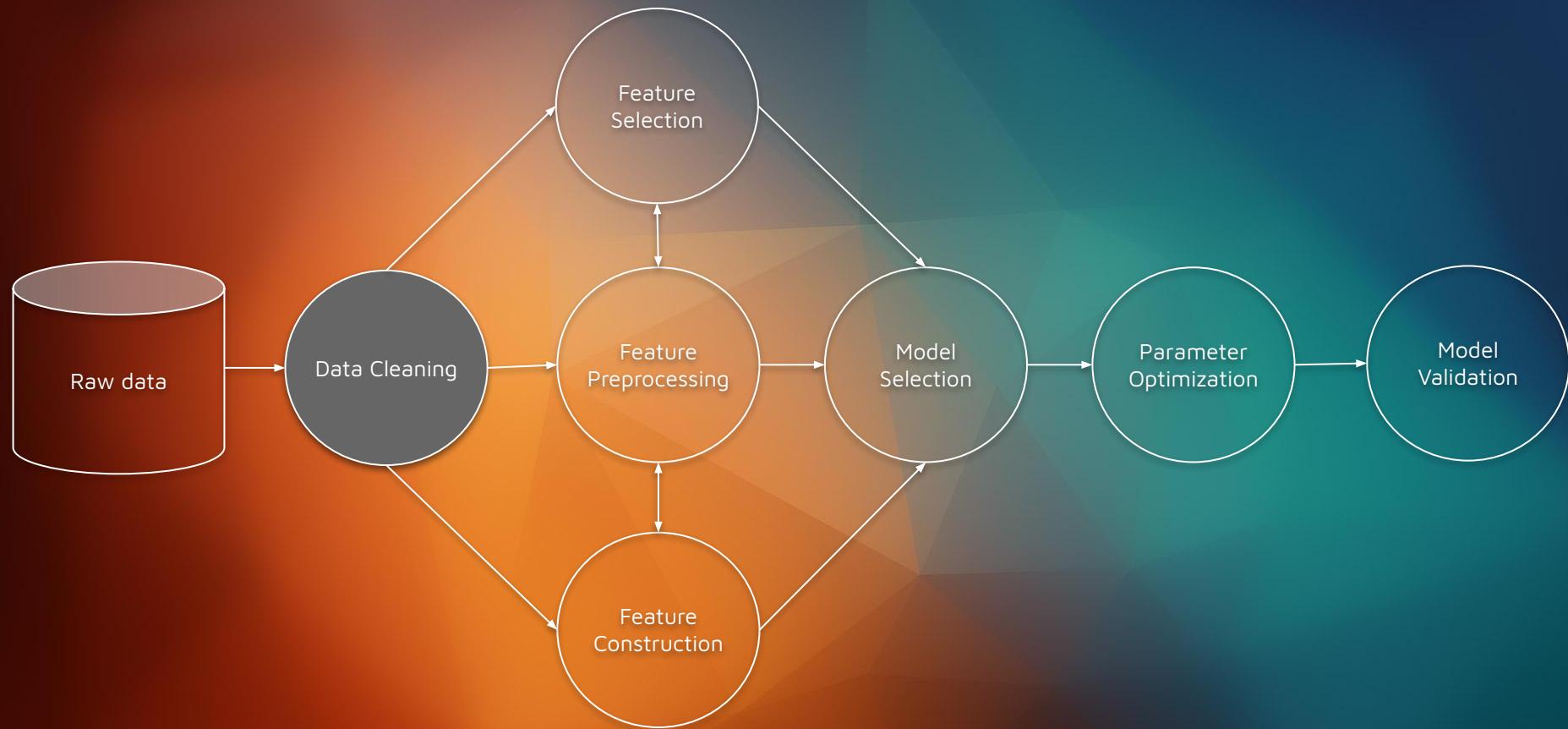


@diegohueltes



RavenPack is the leading big data analytics provider for financial services.





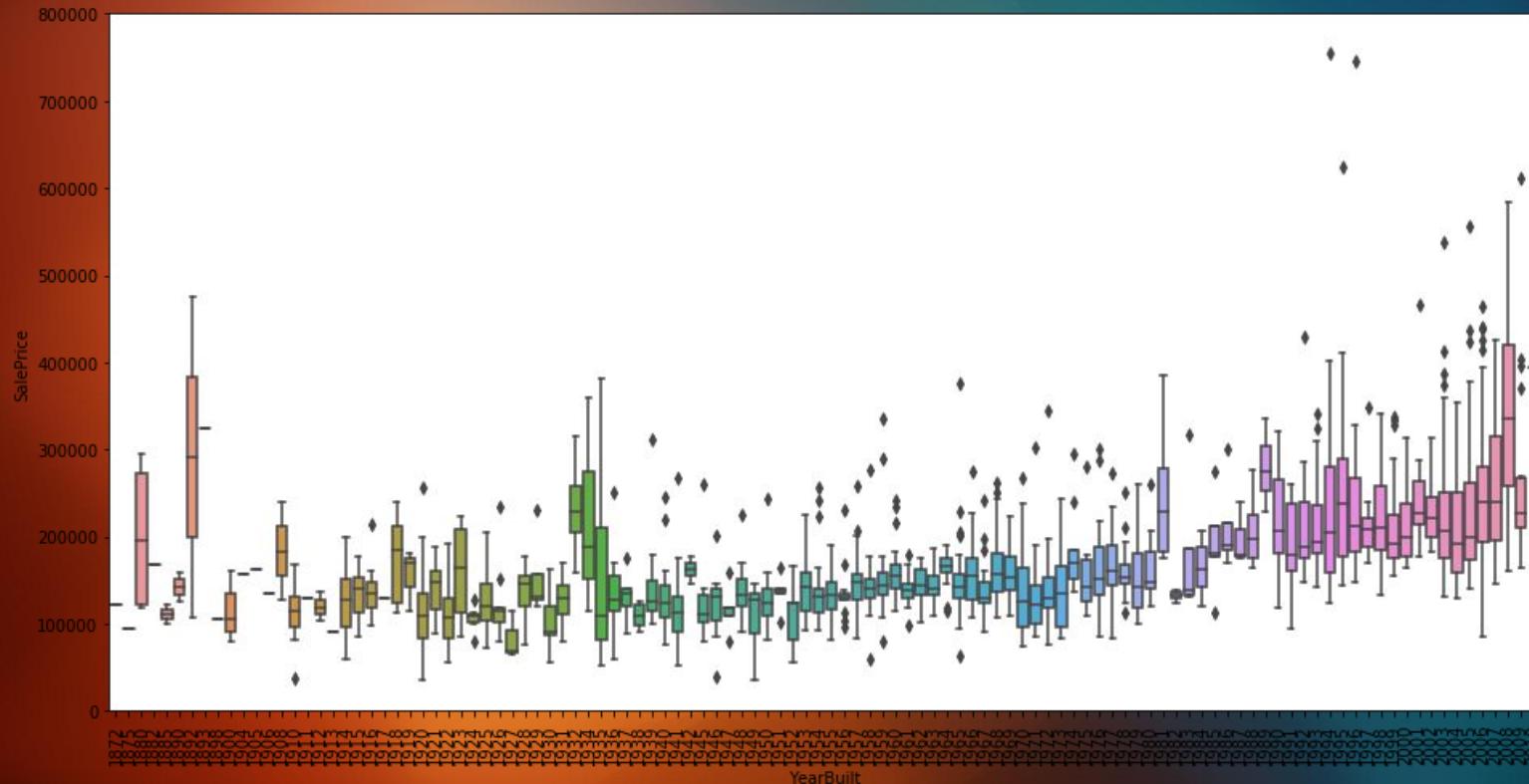
Data cleaning

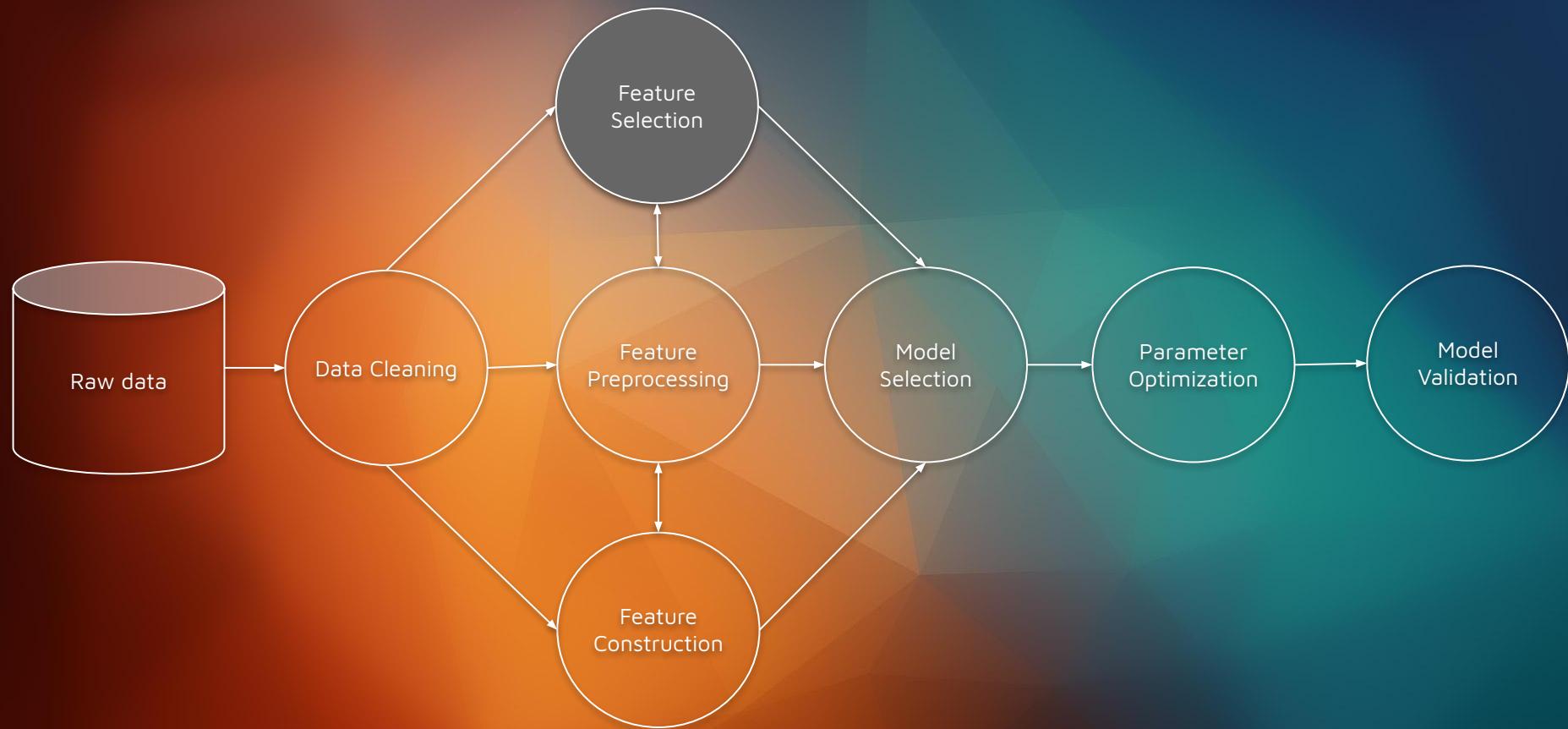


Data cleaning

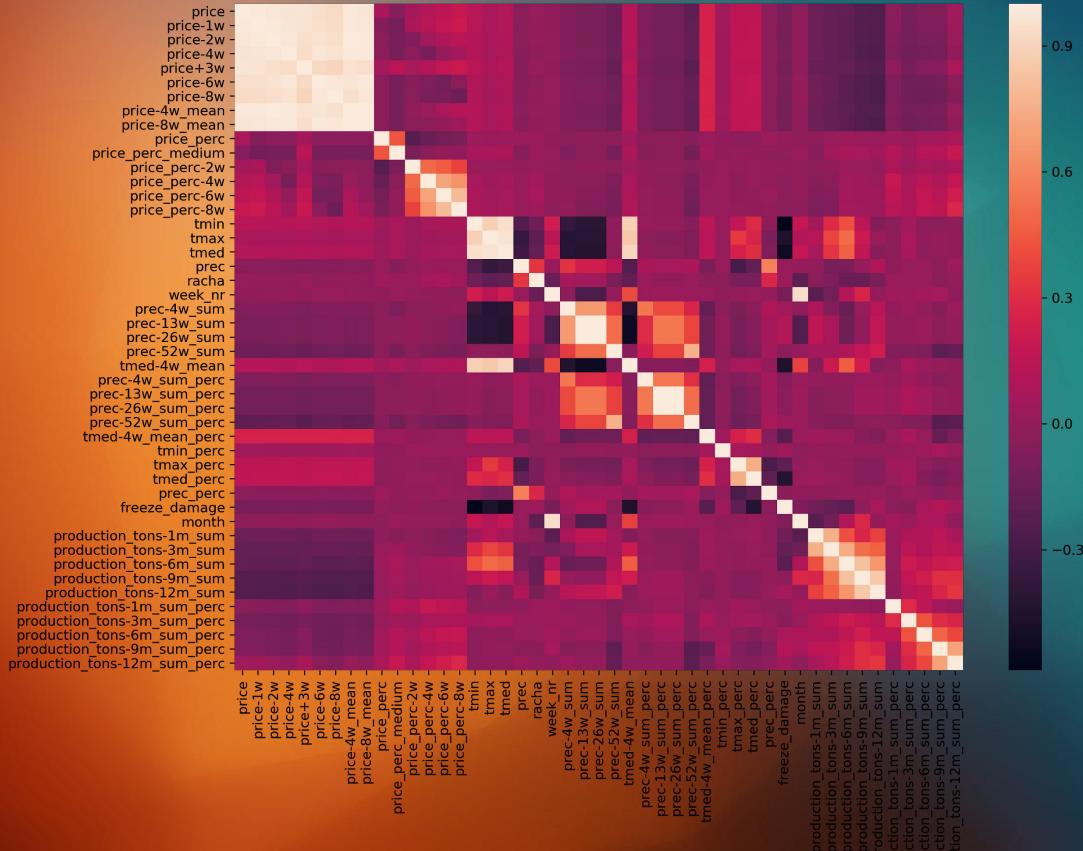
4.9	3.1	1.5	NA	Iris-setosa
5.4	3.7	1.5	NA	Iris-setosa
4.8	3.4	1.6	NA	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
5.7	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa
5.7	4.4	1.5	0.4	Iris-setosa
5.4	3.9	1.3	0.4	Iris-setosa
5.1	3.5	1.4	0.3	Iris-setosa
5.7	3.8	1.7	0.3	Iris-setossa
5.1	3.8	1.5	0.3	Iris-setosa

Data cleaning

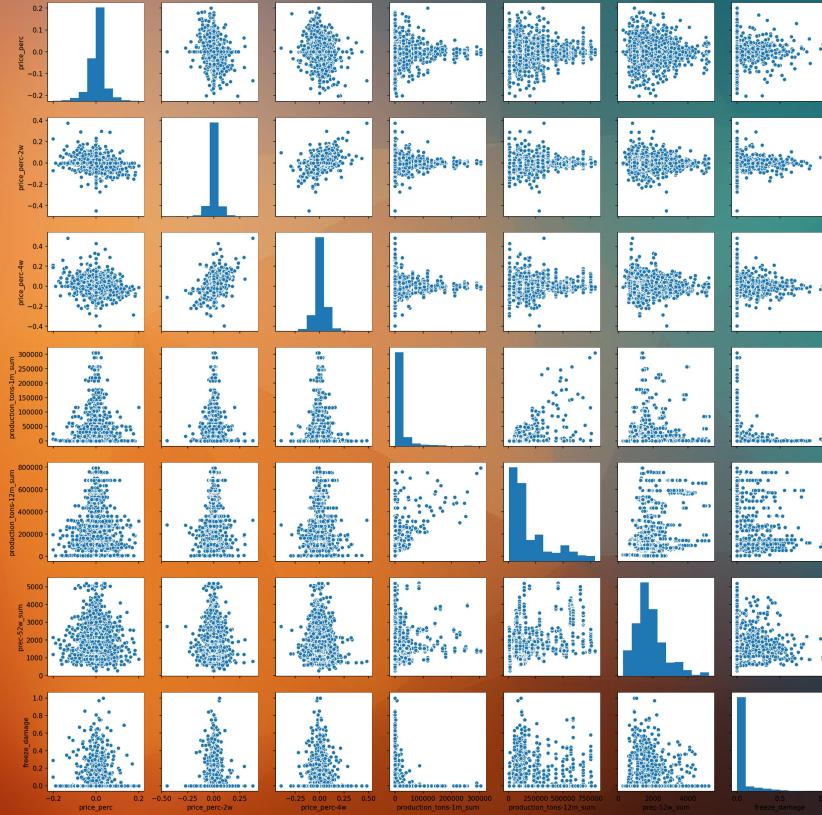


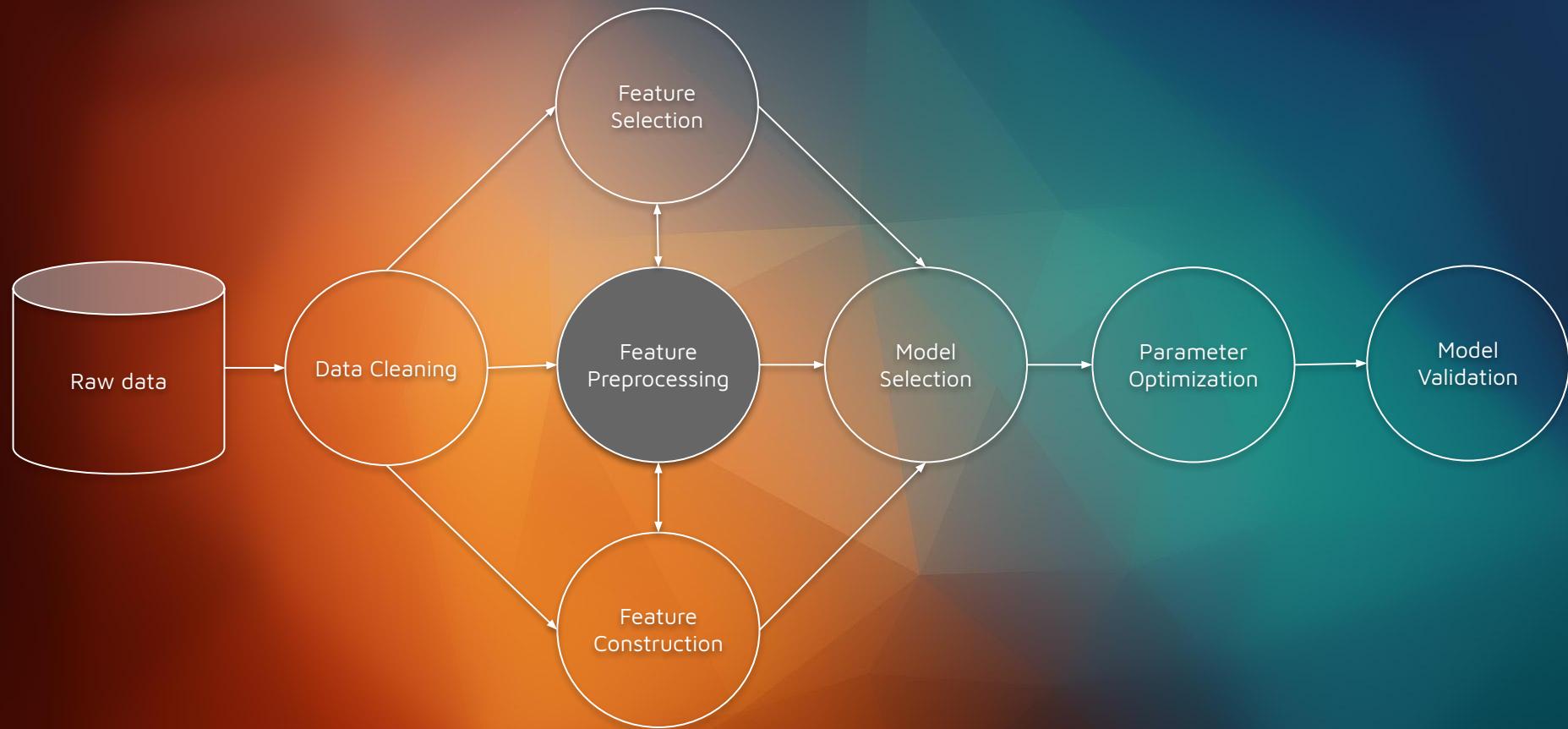


Feature selection

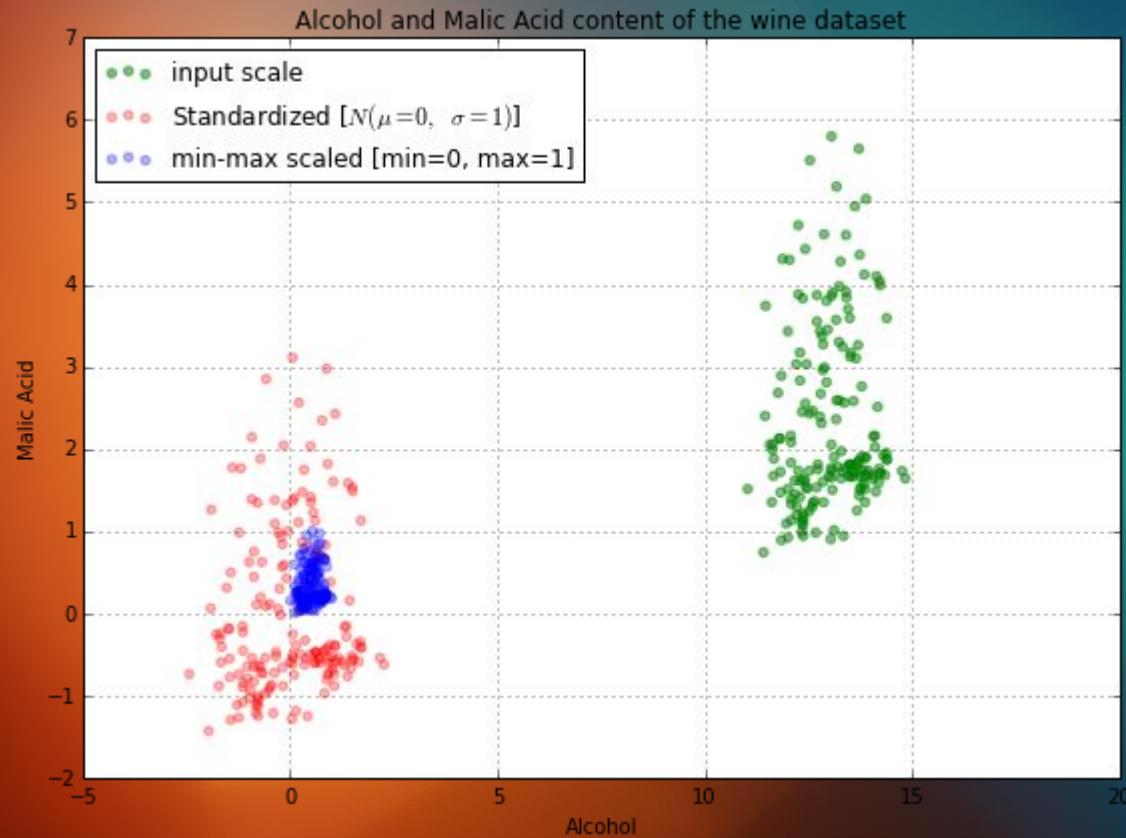


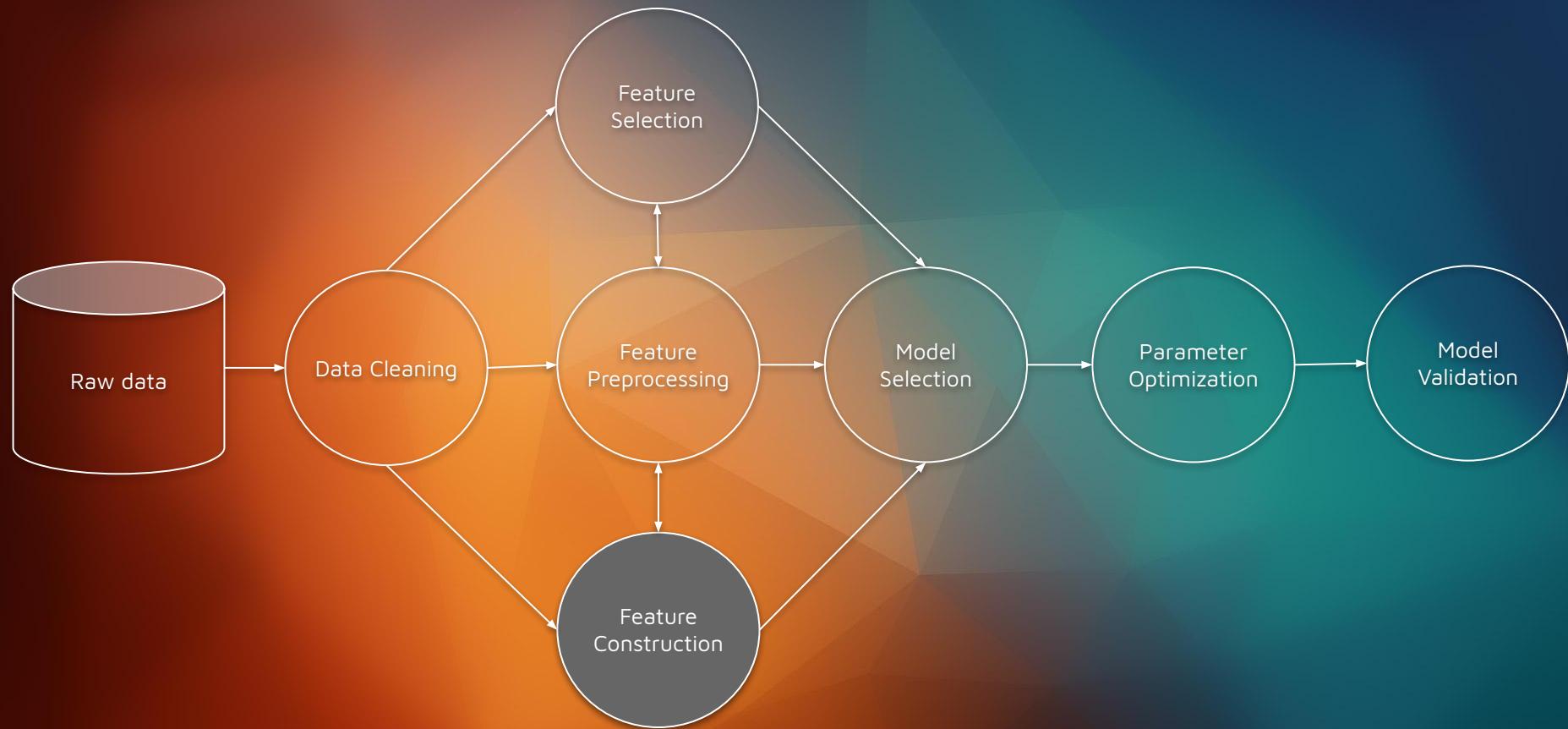
Feature selection





Feature preprocessing

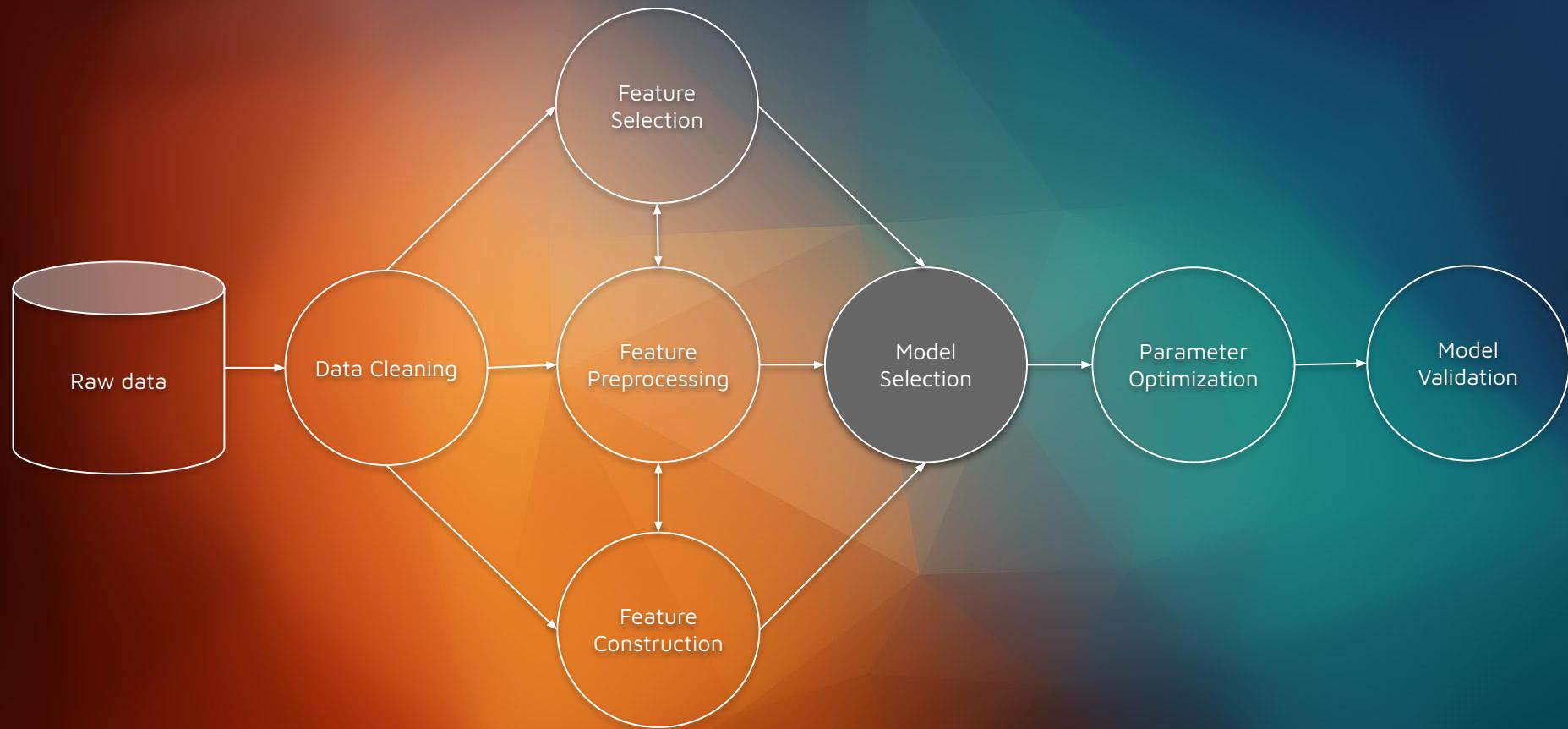




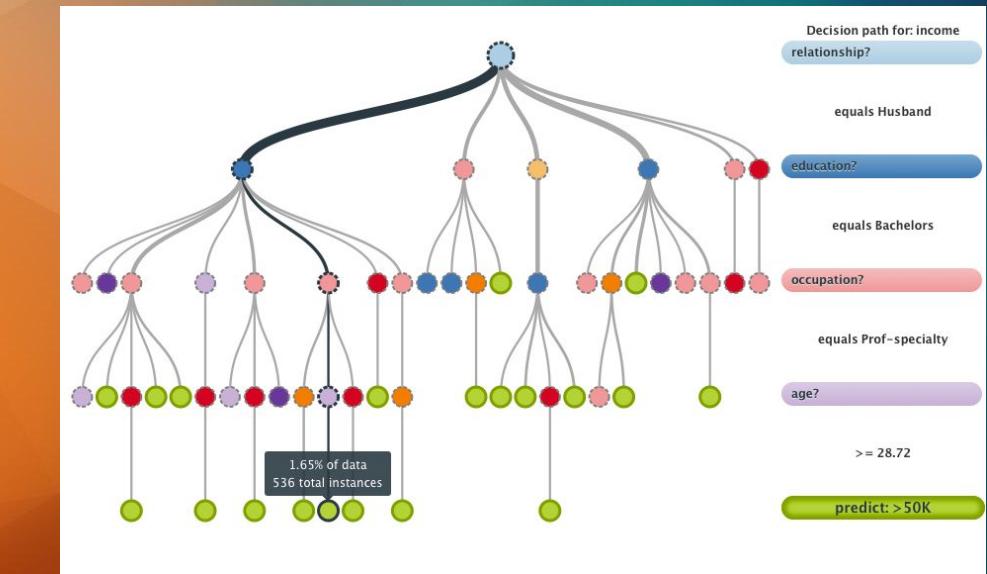
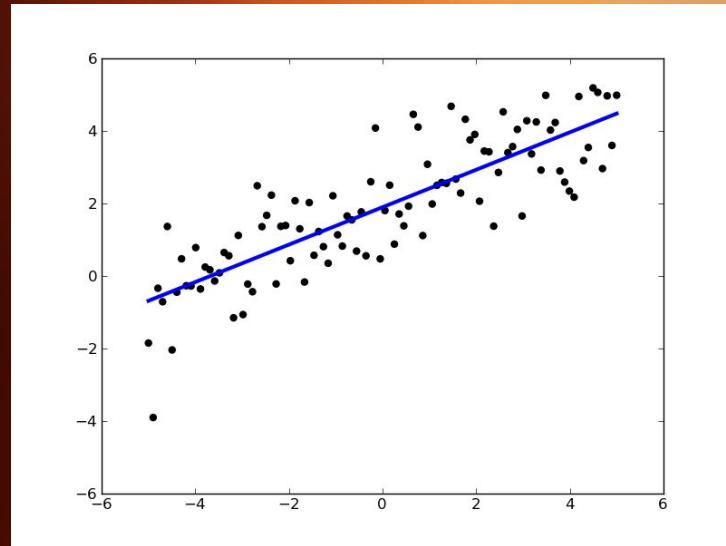
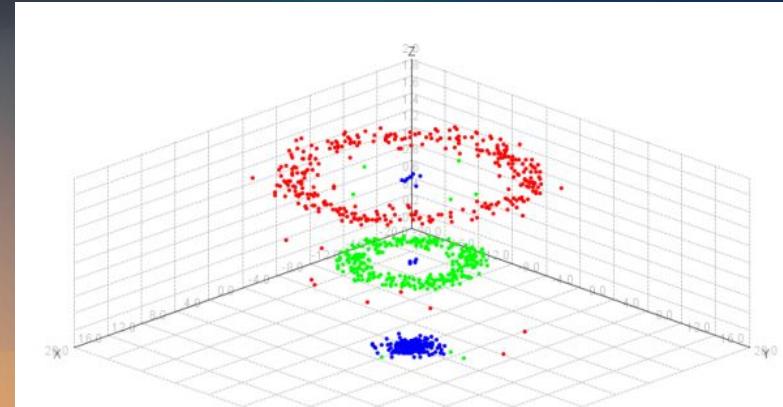
Feature construction

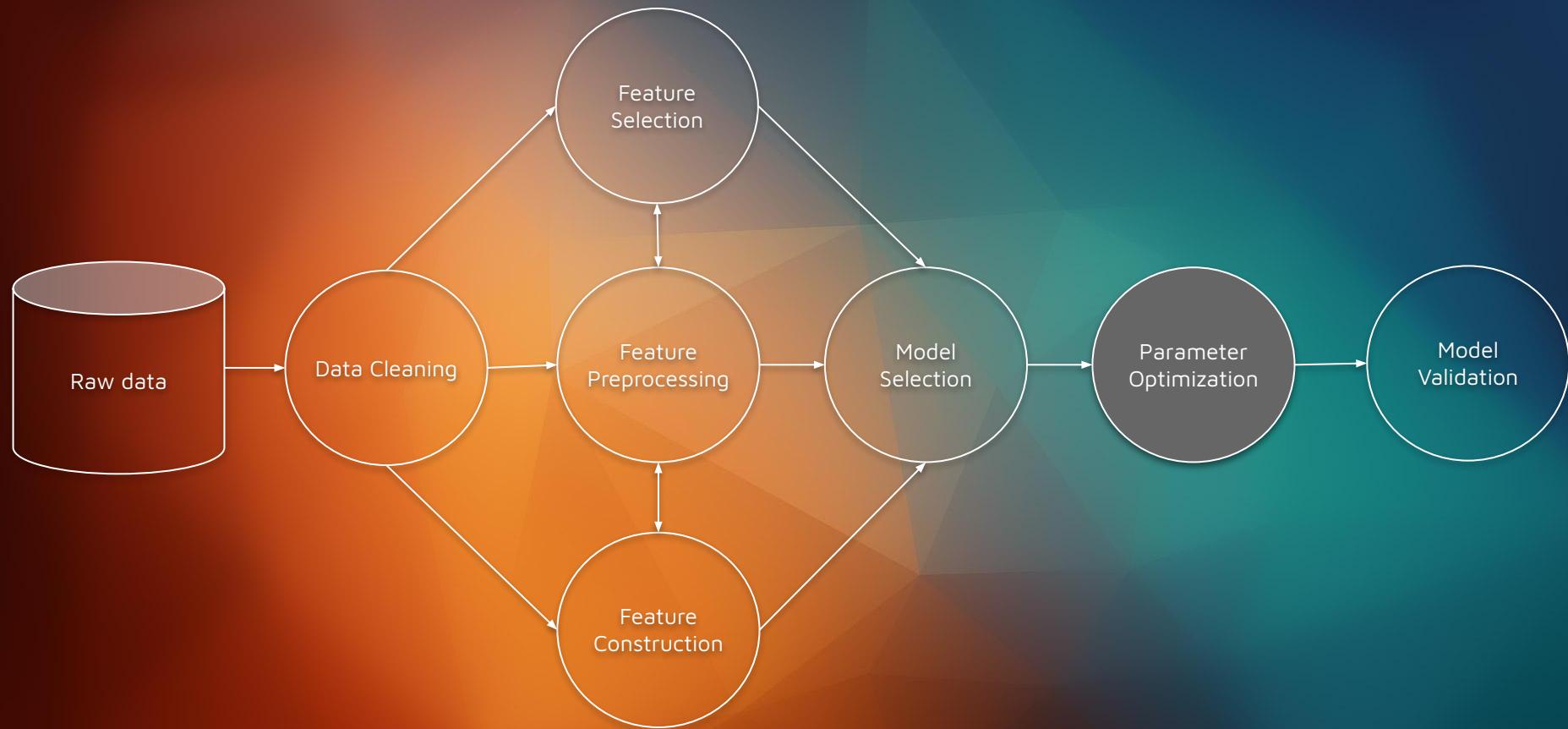


MakeAGIF.com



Model selection

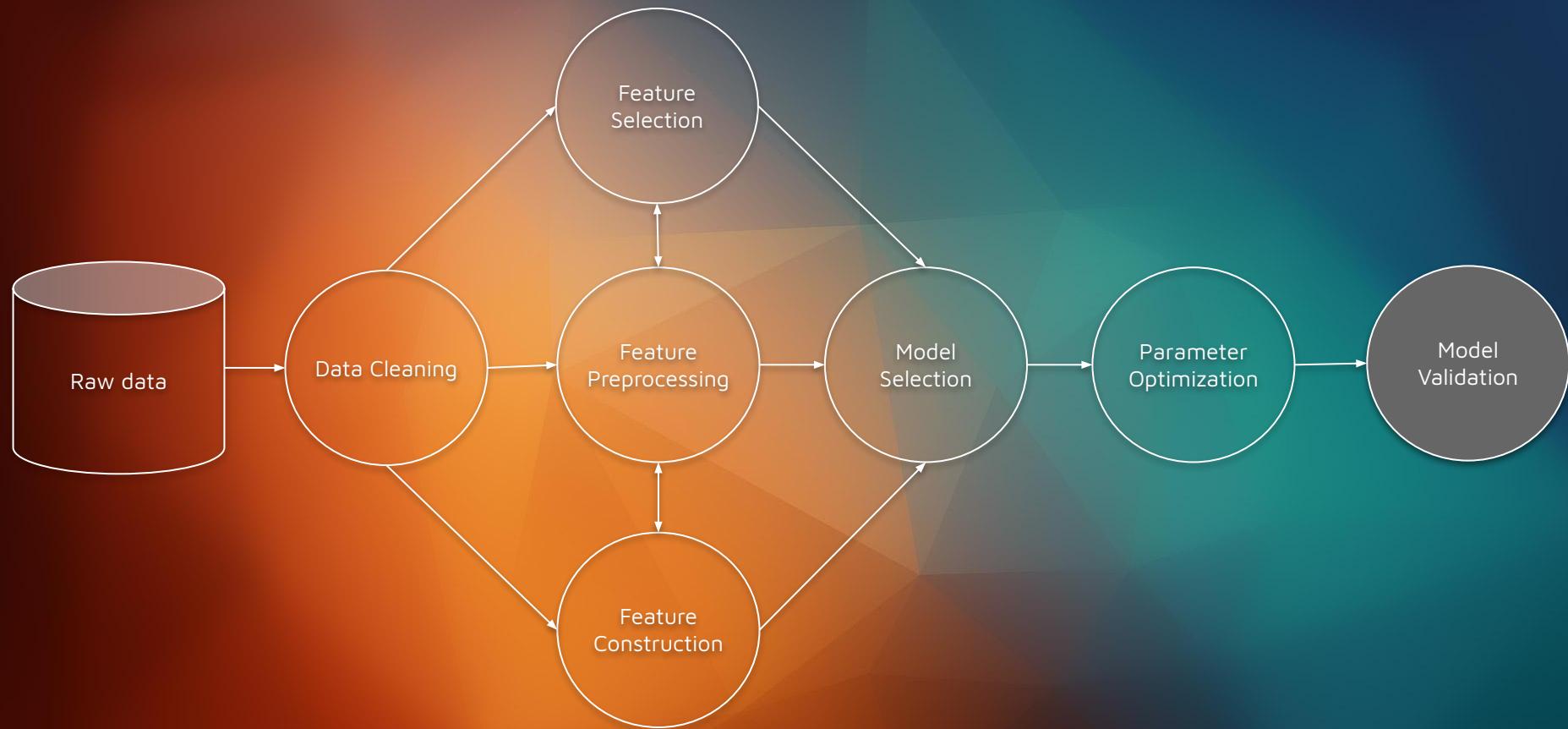




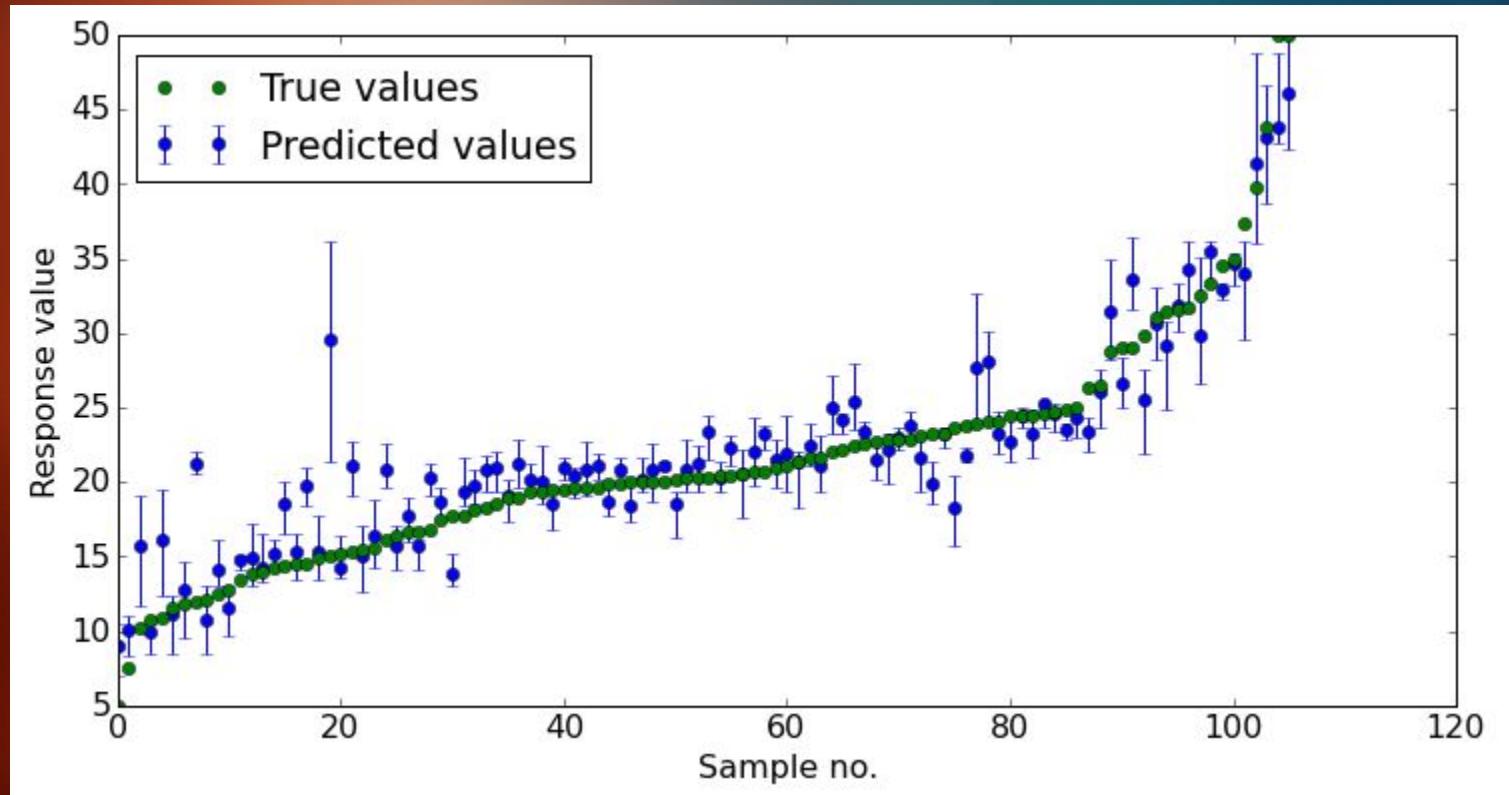
Parameter optimization

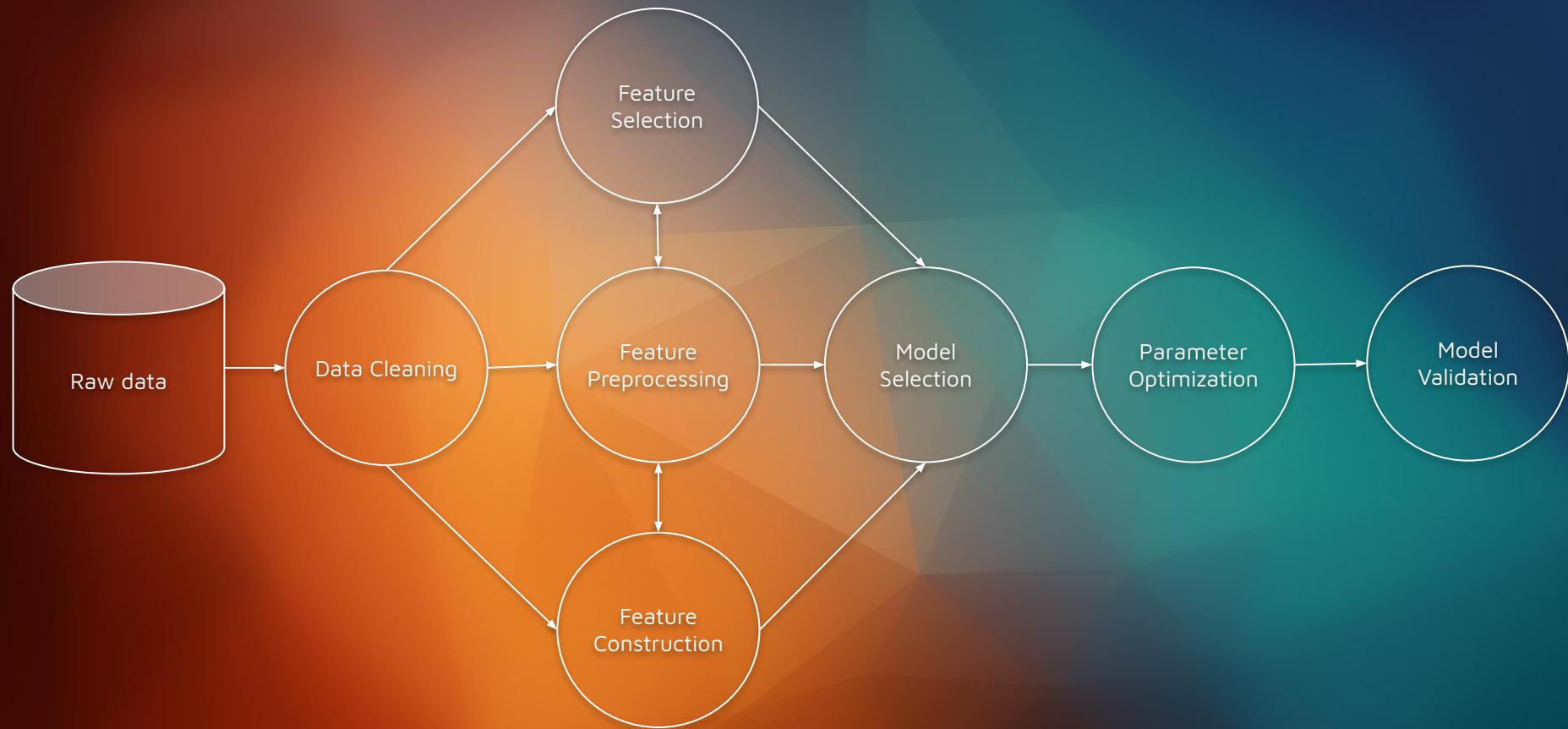
```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                      max_depth=2, max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
                      oob_score=False, random_state=0, verbose=0, warm_start=False)
```





Model validation





lazy

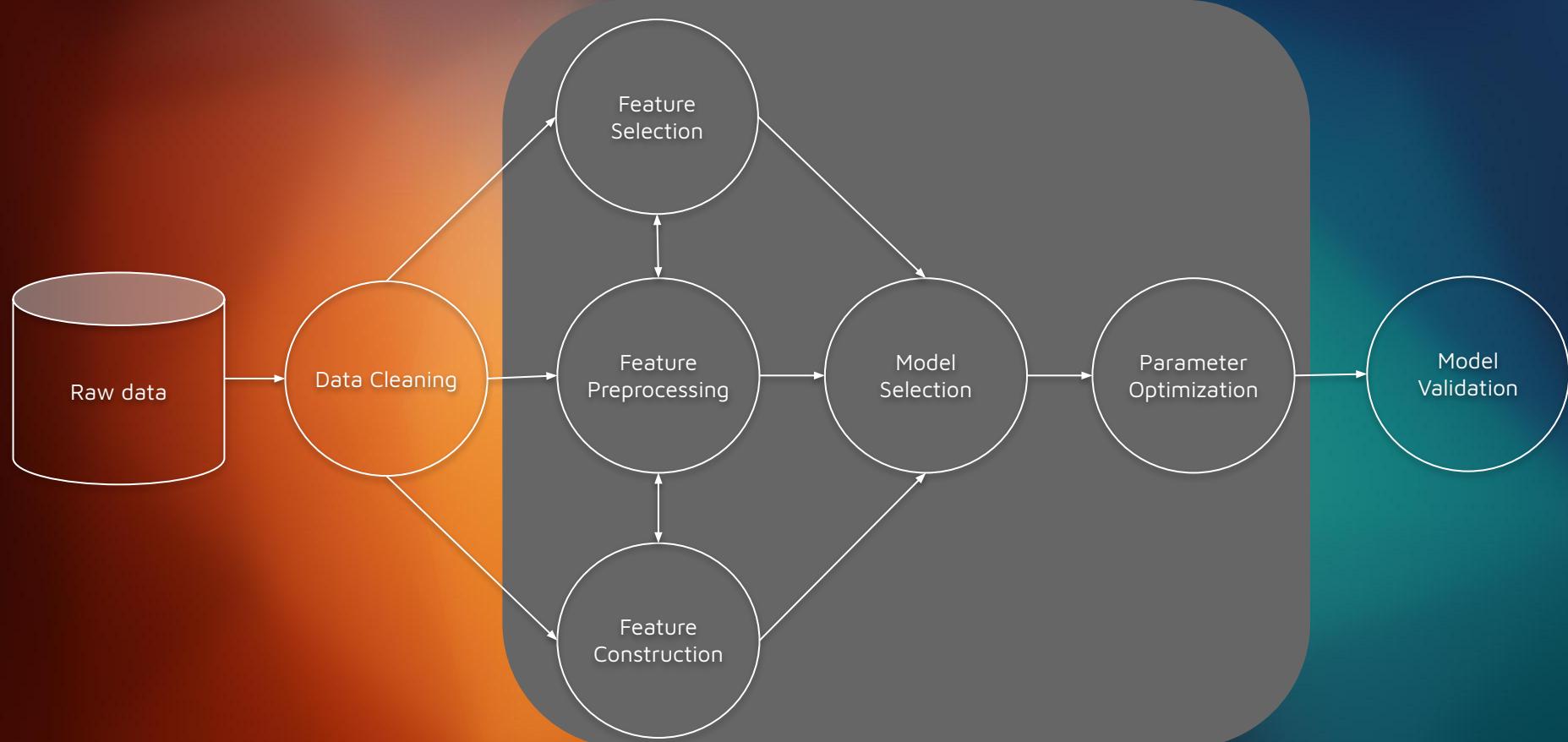
Oxford dictionary

Unwilling to work or use energy

lazy

Diego dictionary
~~Oxford dictionary~~

Unwilling to work or use energy
in repetitive tasks



naive_bayes.GaussianNB
tree.DecisionTreeClassifier
feature_selection.SelectFwe
ensemble.ExtraTreesRegressor
neighbors.KNeighborsRegressor xgboost.XGBRegressor
ensemble.RandomForestRegressor
kernel_approximation.RBFSampler preprocessing.MaxAbsScaler
preprocessing.MinMaxScaler linear_model.LogisticRegression
preprocessing.PolynomialFeatures naive_bayes.MultinomialNB
feature_selection.SelectFromModel preprocessing.RobustScaler
linear_model.ElasticNetCV
neighbors.KNeighborsClassifier
cluster.FeatureAgglomeration
feature_selection.VarianceThreshold preprocessing.Normalizer
ensemble.GradientBoostingClassifier decomposition.PCA
ensemble.feature_selection.SelectPercentile naive_bayes.BernoulliNB
kernel_approximation.Nystroem svm.LinearSVR
ensemble.GradientBoostingRegressor ensemble.ExtraTreesClassifier
ensemble.feature_selection.SelectPercentile naive_bayes.BernoulliNB
ensemble.GradientBoostingRegressor ensemble.AdaBoostRegressor
kernel_approximation.Nystroem svm.LinearSVC linear_model.LassoLarsCV
xgboost.XGBClassifier preprocessing.StandardScaler
tpot.builtins.ZeroCount tree.DecisionTreeRegressor
decomposition.FastICA ensemble.RandomForestClassifier
ensemble.GradientBoostingRegressor preprocessing.Binarizer

Automated Machine Learning

TPOT

Auto-Sklearn

Auto-Weka

Machine-JS

DataRobot

BigML

auto_ml

Google Cloud AutoML

AutoKeras

H2O AutoML

...



TPOT is a Python tool that automatically creates and optimizes **machine learning pipelines** using **genetic programming**.

<https://github.com/EpistasisLab/tpot>



Source: <http://www.genetic-programming.org/gpbook4toc.html>

TPOT

```
from tpot import TPOTClassifier, TPOTRegressor

tpot = TPOTClassifier()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)

tpot = TPOTRegressor()
tpot.fit(X_train, y_train)
predictions = tpot.predict(X_test)
```

TPOT - configuration

```
TPOTClassifier(config_dict = {  
    'sklearn.ensemble.RandomForestClassifier' : {  
        'n_estimators' : [100],  
        'criterion' : ["gini", "entropy"],  
        'max_features' : np.arange(0.05, 1.01, 0.05),  
        'min_samples_split' : range(2, 21),  
        'min_samples_leaf' : range(1, 21),  
        'bootstrap' : [True, False]  
    },  
    'sklearn.feature_selection.RFE' : {  
        'step' : np.arange(0.05, 1.01, 0.05),  
        'estimator' : {  
            'sklearn.ensemble.ExtraTreesClassifier' : {  
                'n_estimators' : [100],  
                'criterion' : ['gini', 'entropy'],  
                'max_features' : np.arange(0.05, 1.01, 0.05)  
            }  
        }  
    }  
})
```

auto-sklearn

auto-sklearn frees a machine learning user from algorithm selection and hyperparameter tuning. It leverages recent advantages in **Bayesian optimization, meta-learning** and **ensemble construction**

<https://github.com/automl/auto-sklearn>

auto-sklearn

```
import autosklearn.classification  
import autosklearn.regression  
  
automl = autosklearn.classification.AutoSklearnClassifier()  
automl.fit(X_train, y_train)  
predictions = automl.predict(X_test)  
  
automl = autosklearn.regression.AutoSklearnRegressor()  
automl.fit(X_train, y_train)  
predictions = automl.predict(X_test)
```

auto-sklearn custom config

include_estimators

exclude_estimators

include_preprocessors

exclude_preprocessors

Olive oil full dataset...

Test in Google Colab, clean dataset.

TPOT: 55% Accuracy

auto-sklearn: 56% Accuracy

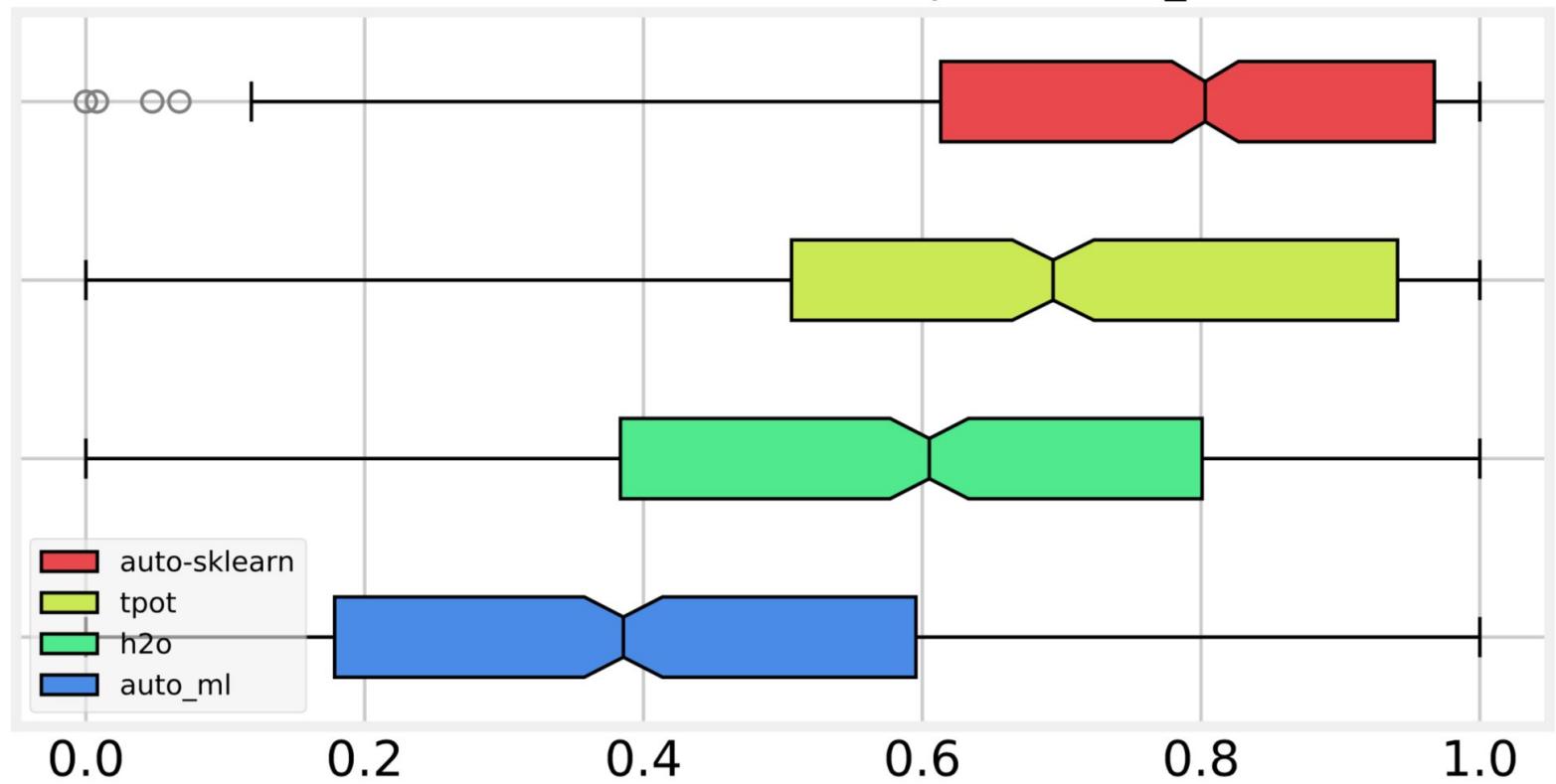
H2O automl: 51% Accuracy

Benchmarking Automatic Machine Learning Frameworks

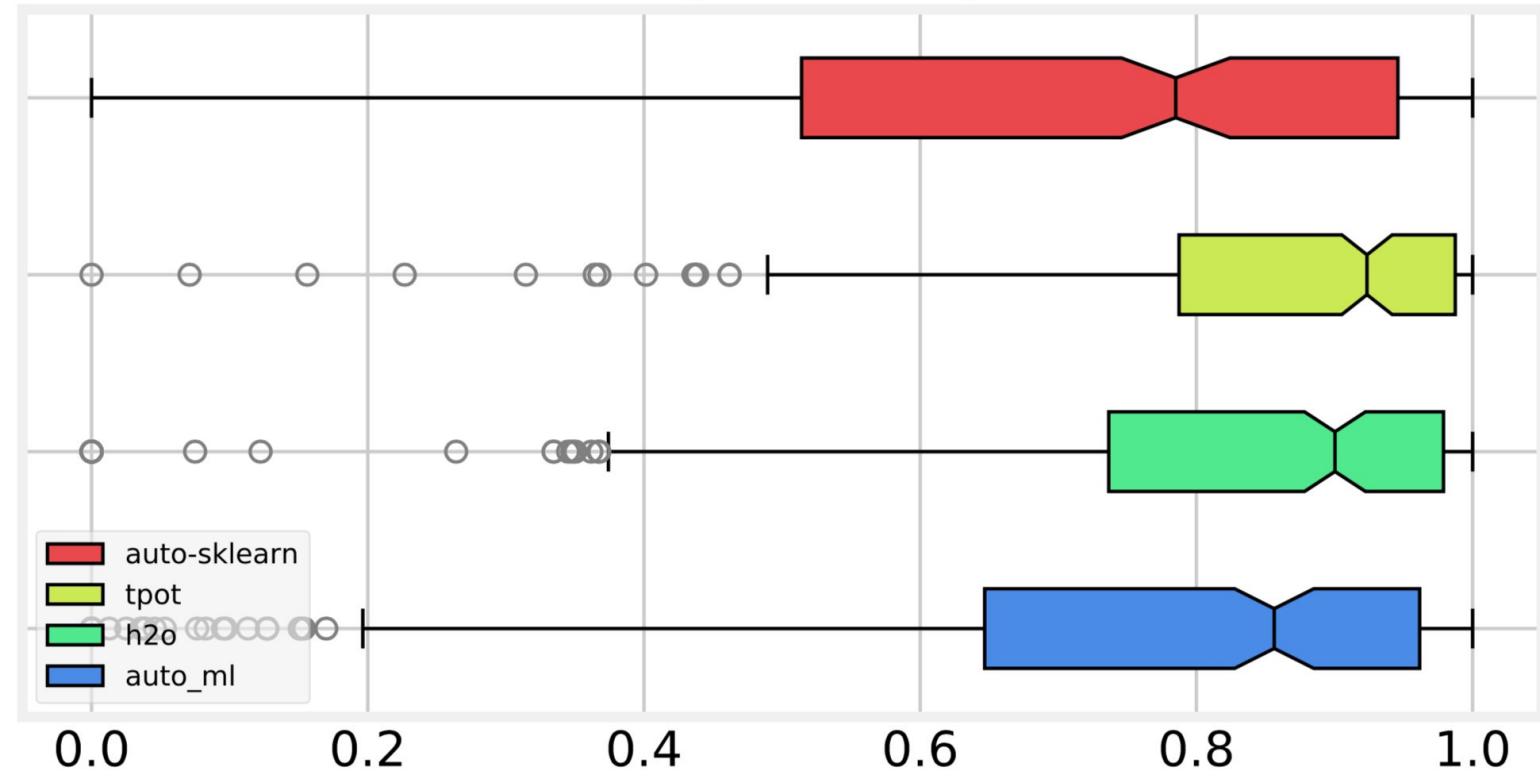
Adithya Balaji, Alexander Allen

<https://arxiv.org/abs/1808.06492>

Raw Per Model Classification Comparison (F1_SCORE)



Raw Per Model Regression Comparison (MSE)



Automated Machine Learning – A Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb

<https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8>

Automated Machine Learning

- Machine Learning for beginners
- Exploratory analysis
- Selective discovering
- New ideas for your model
- Model optimization

Progress isn't made by early risers. It's
made by lazy people trying to find
easier ways to do something.

– Robert A. Heinlein



 RavenPack



@diegohueltes



DiegoHueltes

<https://www.ravenpack.com/careers/>