kaggle DAYS

Challenges in Automated Machine Learning

Gilberto Titericz Jr

November / 2019

Spain

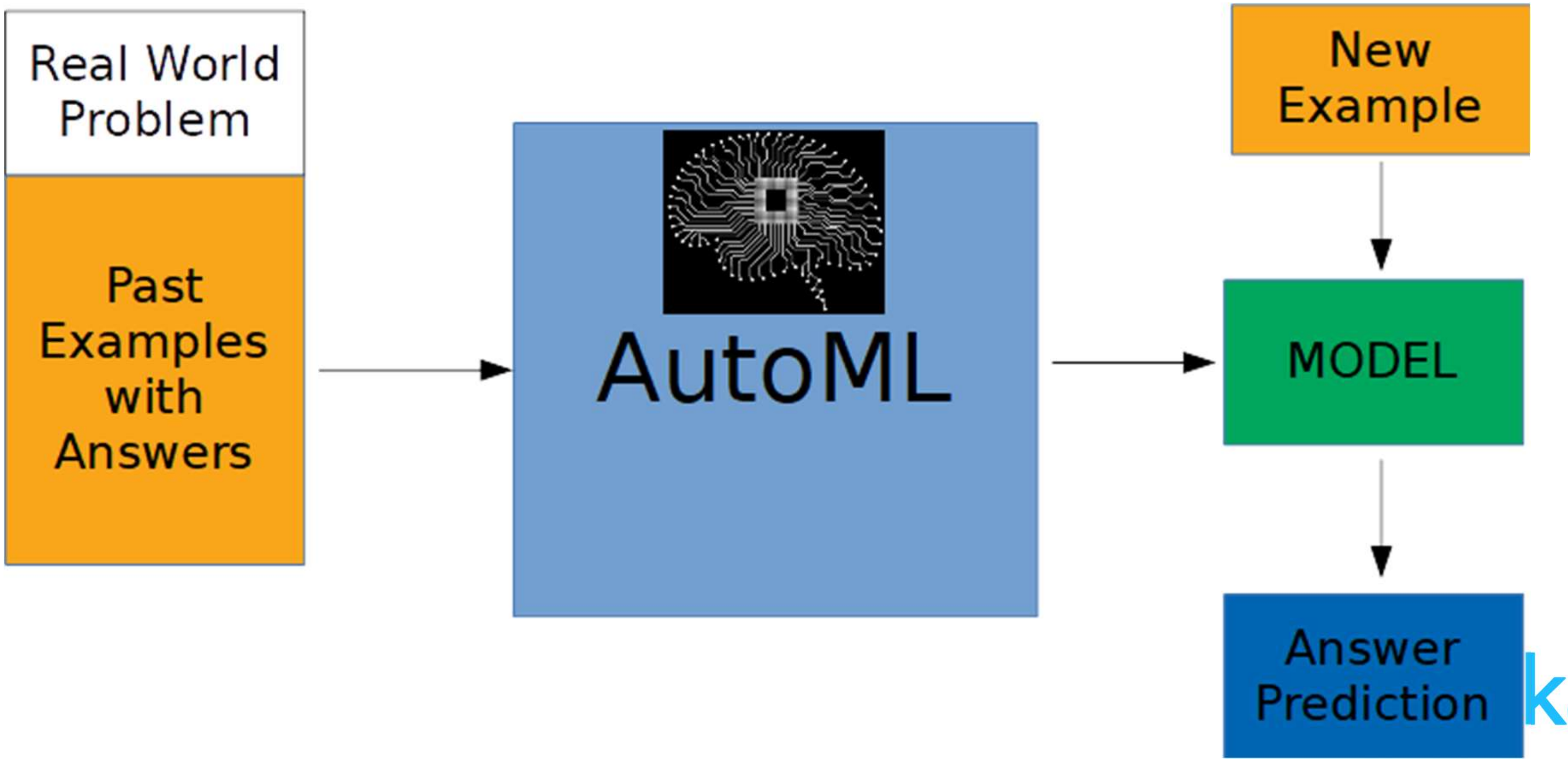
Automated Machine Learning
What is autoML?

- Wiki: “Is the process of automating end-to-end the process of applying machine learning to real-world problems”.

- It an intuitive tool that enables anyone to do Machine Learning and a productive tool for Data Scientists.

- AutoML is an A.I. that build A.I.’s.

Automated Machine Learning (AutoML)



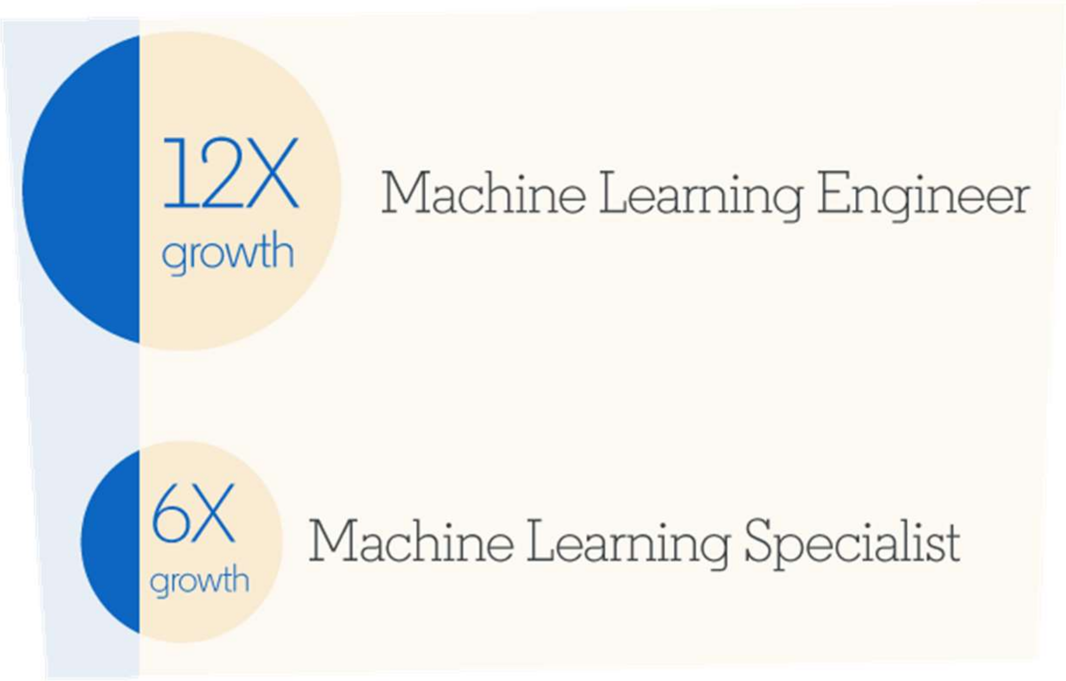
kaggle DAYS

Why AutoML ?

According Forbes,
the world generates
2.500.000.000.000.000.000
(2.5 Zetta) bytes of data every day.

Why AutoML?

- Demand for Data Scientists in 2018

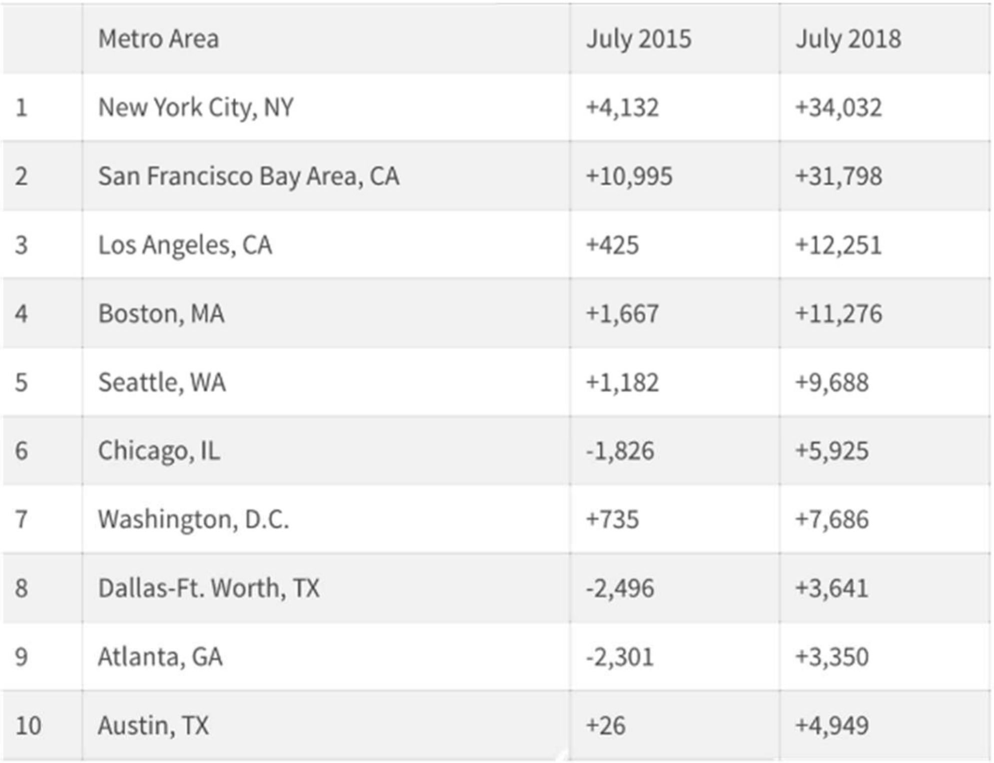


12X growth Machine Learning Engineer

6X growth Machine Learning Specialist

Source: https://economicgraph.linkedin.com/research/linkedin-2018-emerging-jobs-report

Why AutoML?



	Metro Area	July 2015	July 2018
1	New York City, NY	+4,132	+34,032
2	San Francisco Bay Area, CA	+10,995	+31,798
3	Los Angeles, CA	+425	+12,251
4	Boston, MA	+1,667	+11,276
5	Seattle, WA	+1,182	+9,688
6	Chicago, IL	-1,826	+5,925
7	Washington, D.C.	+735	+7,686
8	Dallas-Ft. Worth, TX	-2,496	+3,641
9	Atlanta, GA	-2,301	+3,350
10	Austin, TX	+26	+4,949

Source:
https://learning.linkedin.com/blog/tech-tips/why-it-s-really-good-to-be-a-data-scientist-right-now

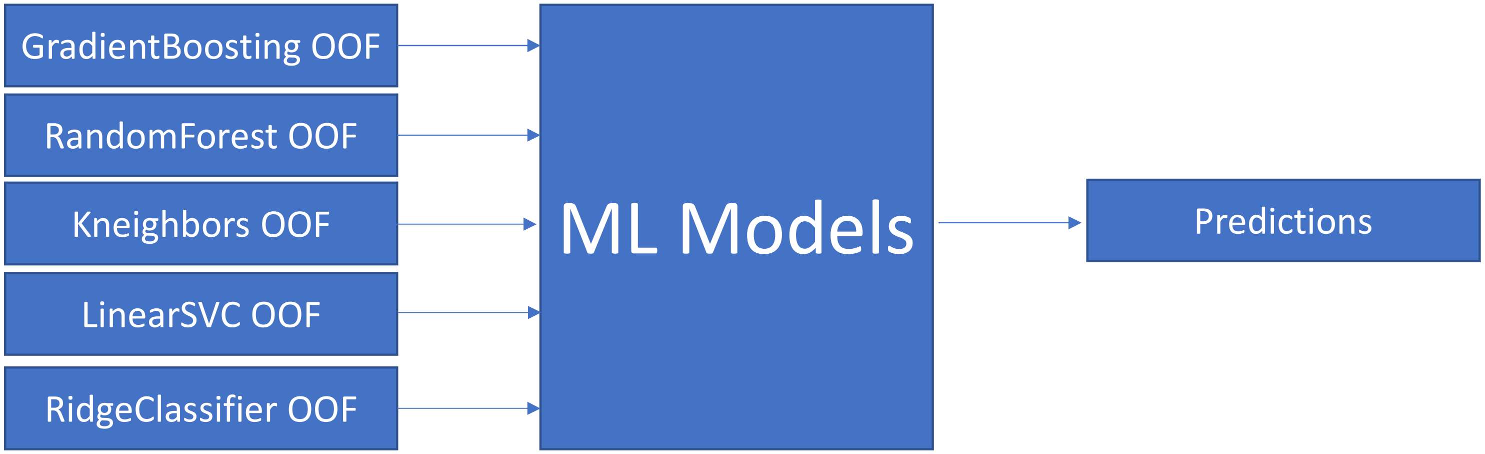
Why AutoML?

- 5-Fold CV, AUC Performance: (50 binary classification datasets)

Algorithm	Defaul HP	Tuned HP
GradientBoosting	0.826	0.891 (+6.57%)
RandomForest	0.810	0.861 (+5.1%)
KNeighbors	0.780	0.827 (+4.6%)
LinearSVC	0.772	0.811 (+4.0%)
RidgeClassifier	0.765	0.790 (+2.5%)
	0.790	0.836 (+4.6%)

Why AutoML?

- Ensembling:


GradientBoosting OOF

RandomForest OOF

Kneighbors OOF

LinearSVC OOF

RidgeClassifier OOF

ML Models

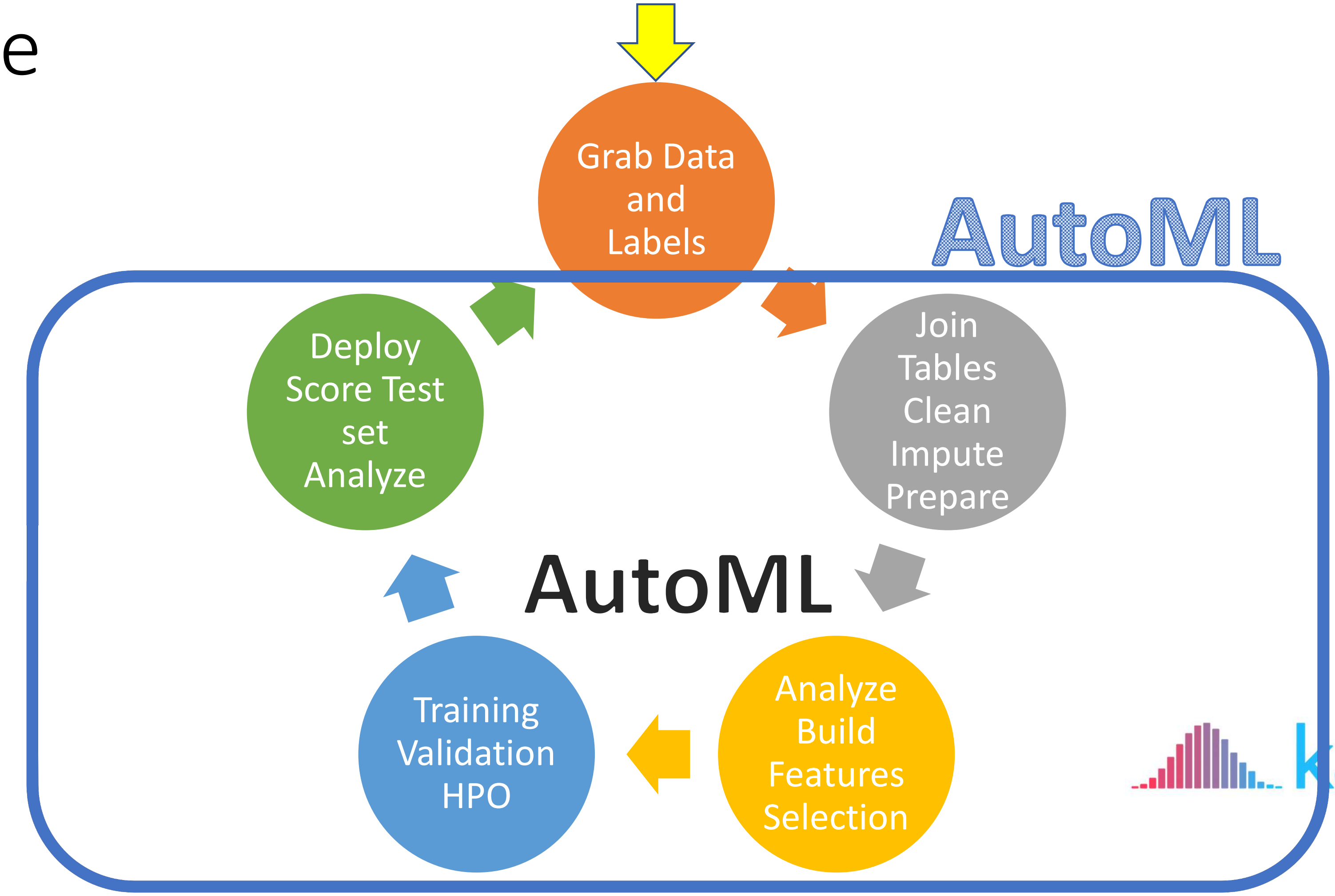
Predictions

Stacking Ensemble of HP Tuned models
AUC: 0.902 (+7.6% over best untuned model)

Why AutoML?

- Model Interpretability:
  - Why the model predicted it?
  - What are the most important features?
  - How features interact each other?

AutoML: Automation of a Machine Learning Cycle


Grab Data and Labels
AutoML
Join Tables Clean Impute Prepare
Analyze Build Features Selection
Training Validation HPO
Deploy Score Test set Analyze
AutoML

Gathering Data

- Define a problem.

- How to gather the data?

- How to store the data?

1 - Data Preparation/Wrangling

- Load the Data.
- Join Tables.
- Clean/Drop values or errors.
- Remove outliers (outlier detection).
- Data augmentation (add artificial examples).

2 - Analyze Data

- Feature Discovery.
- Feature Distribution.
- Correlations.
- Clusterings.
- Build Features (Feature Engineering).

3 - Train a Model

- Choose a range of algorithms.
- Choose a validation strategy.
- Hyperparameter tuning.
- Neural Architecture Searching.
- Feature Selection.
- Ensembling.

4 - Test a Model

- Analyze predictions.
- Check for inconsistences.
- Calculate relevant metrics.
- Compare models predictions.
- Prepare for model interpretability.

5 - Deploy a Model

- Deploy models to production.
- Check consistence.
- Maintain.
- Data health.

AutoML Challenges

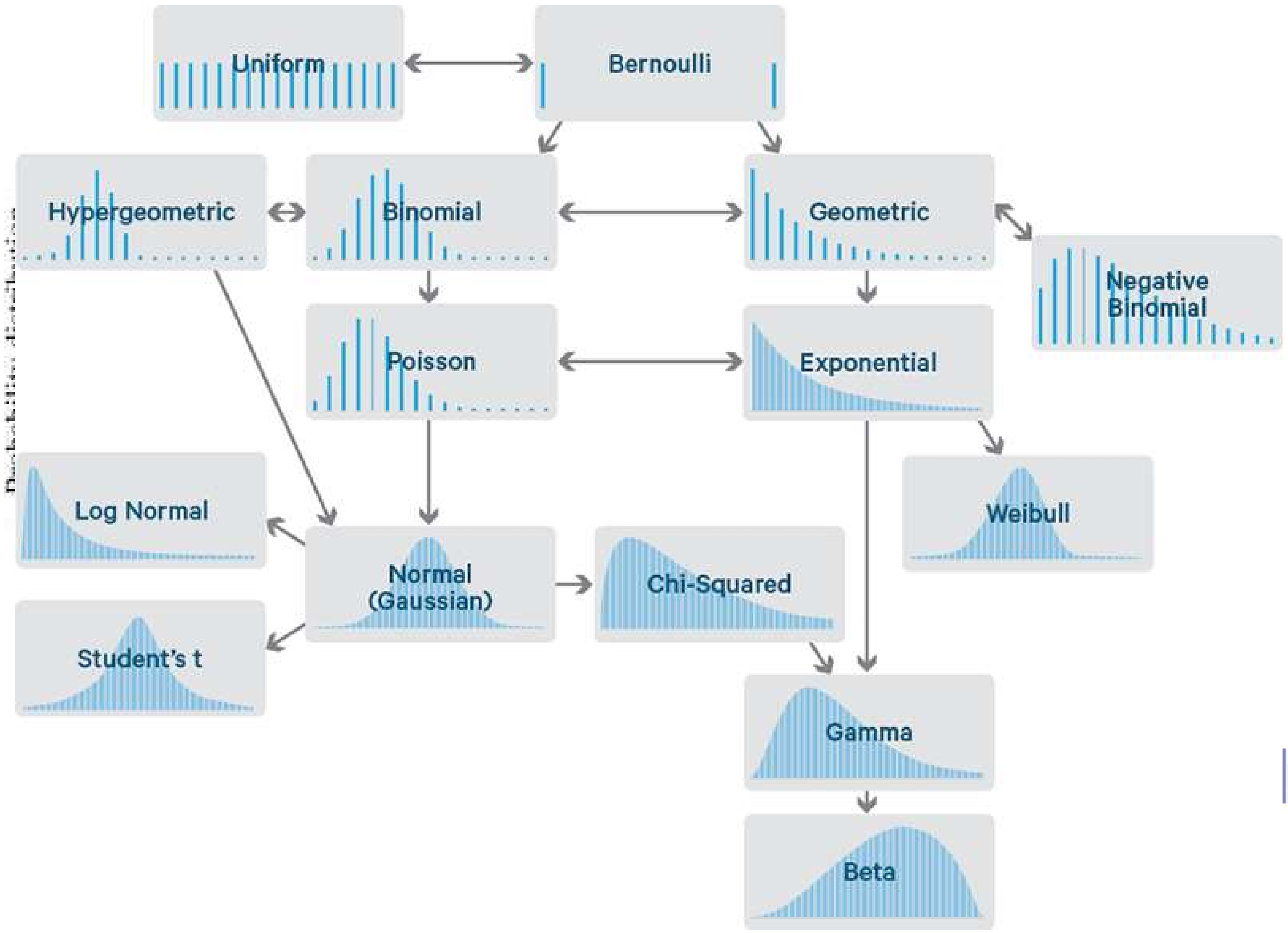
Automatic Feature Discovery

This is a test	80455	1234567890	1,80
I like data	80995	9876543211	5-6
I like data	45665	1928376450	1.55
Ask a DS about data	12336	1122334455	2.05

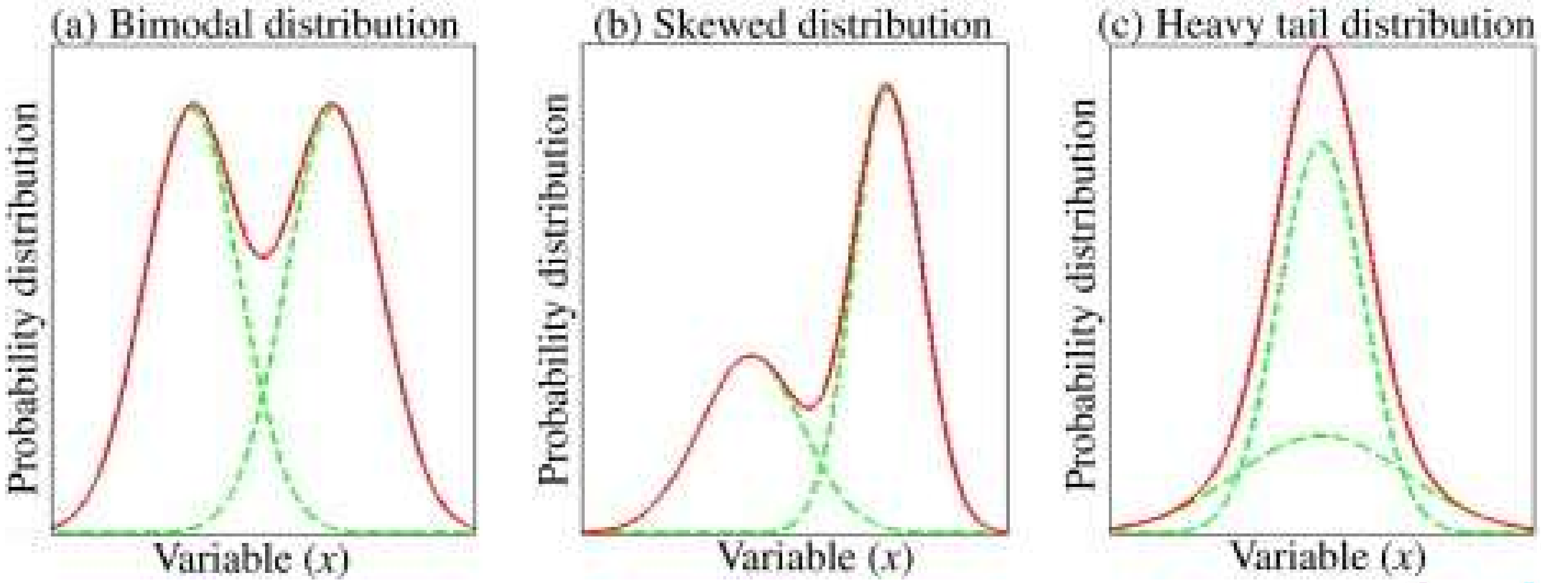
Automatic Feature Discovery

Categorical Feature	Zip Codes	TimeStamp	Human Height
This is a test	80455	1234567890	1,80 m
I like data	80995	9876543211	5 ft-6 in
I like data	45665	1928376450	1.55 m
Ask a DS about data	12336	1122334455	2.05 m

Automatic Feature Discovery



Automatic Feature Discovery



Automatic Feature Engineering

Tabular Data:

1	2	3	4	5	6
Cat 1	Cat 2	num	**AVG(num)byCat1**	**Cat1+Cat2**	**Cat1 Frequency**
RED	DARK	1	3.5	REDDARK	2
GREEN	DARK	10	6	GREENDARK	2
GREEN	DARK	2	6	GREENDARK	2
RED	LIGHT	6	3.5	REDLIGHT	2
BLUE	LIGHT	3	3	BLUELIGHT	1

Automatic Feature Engineering

NLP :

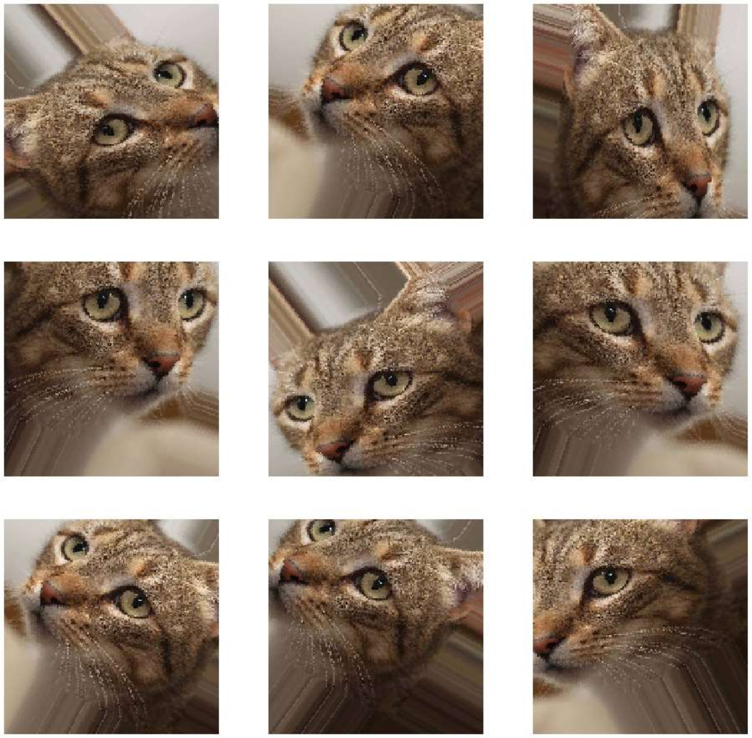
- Detect Language.
- Find Language specific relations between the words.
- Cleaning the text.
- Fix Misspellings.
- Traditional approaches like: ngrams, tf-idf and bag of words.
- Extracted Embeddings from Pretrained Models.

Automatic Feature Engineering

Image Classification:

- Extracted Embeddings from pretrained models:
  - Raw, PCA, ICA, Isomap, AutoEncoder, etc…

- Pure CNN model:
  - Preprocessings: filter, edge detection, equalization.
  - Augmentations: horizontal/vertical flip, rotation, deformation, distortion, equalization, add noise.



Automatic Feature Engineering

Tabular + NLP + Images

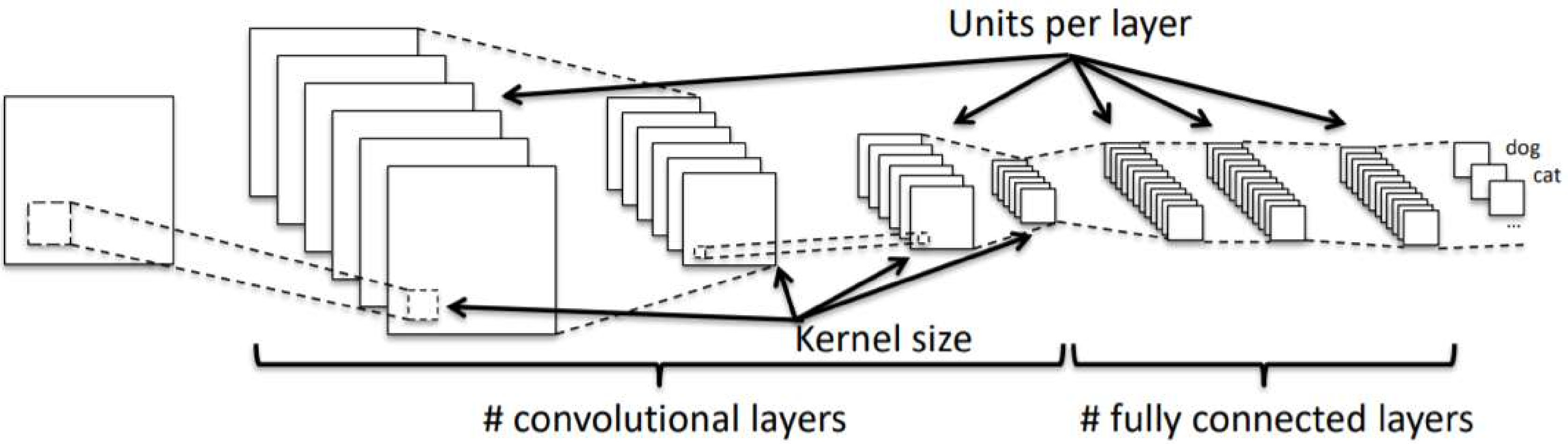
Automatic Model Fit

- Set a Validation Strategy (Time based, Random/Stratified, Grouped).
- Set a relevant Metric.
- Train a model: Linear, Decision Trees, Deep Learning, etc…
  - Preprocess data for each model type.
- Optimize Hyperparameter and Neural Architecture Search.

Hyperparameter Optimization

- LightGBM (GBDT) Parameters:
    - Around 60 hyperparameters to set.

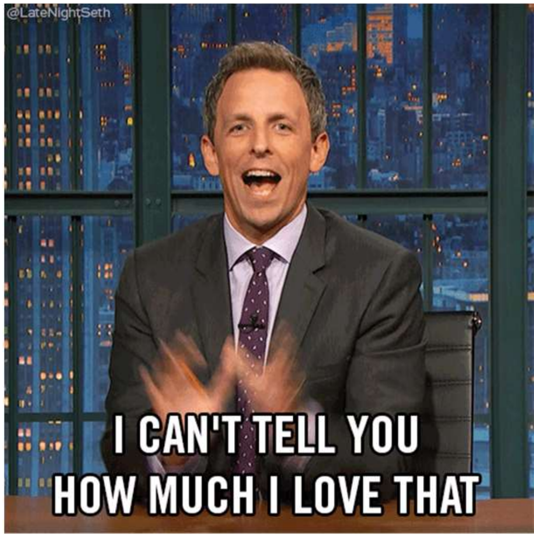
- Neural Nets:
    - Some hyperparametrs and a large number of architectures to try.



Hyperparameter Optimization

```
Class lightgbm.LGBMClassifier(
        boosting_type=‘gbdt’,
        num_leaves=31,
        max_depth=-1,
        learning_rate=0.1,
        n_estimators=100,
        subsample_for_bin=200000,
        objective=‘binary’,
        class_weight=None,
        min_split_gain=0.0,
        min_child_weight=0.001,
        min_child_samples=20,
        subsample=1.0,
        subsample_freq=0,
        colsample_bytree=1.0,
        reg_alpha=0.0,
        reg_lambda=0.0,
)
```

Hyperparameter Optimization



- Algorithms:
  - Grid Search
  - Random Search
  - Gradient Based Optimization
  - Bayesian Optimization
  - Bayesian Optimization + Hyperband
  - Guess (sklearn-autoML)

kaggle DAYS

Hyperparameter Optimization

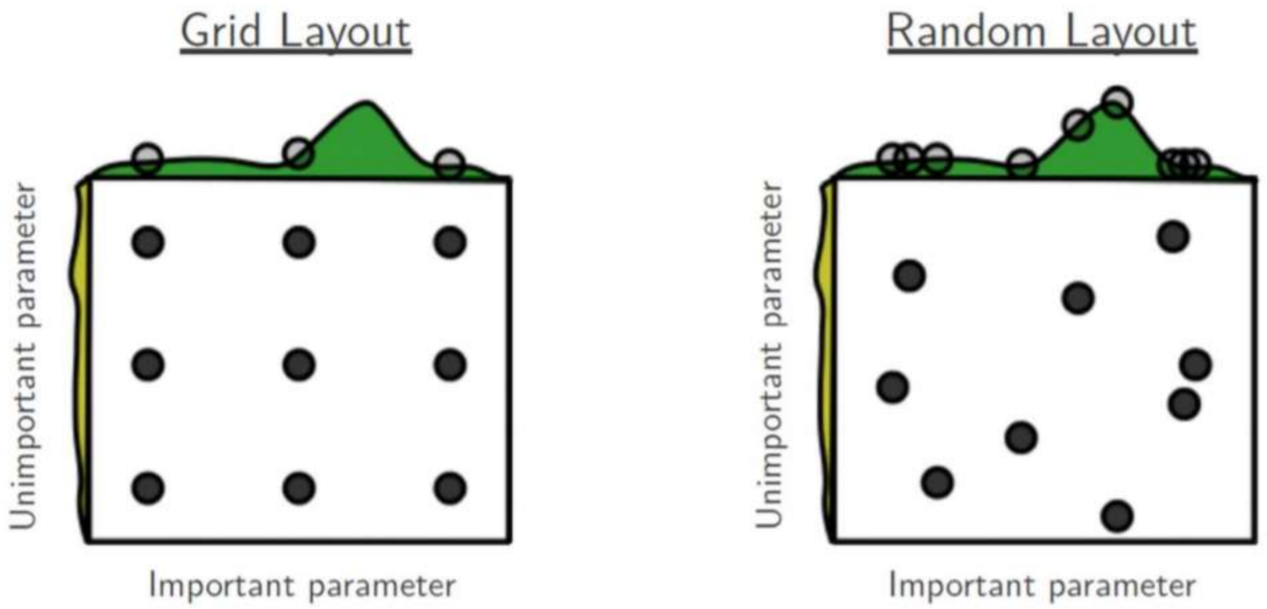


Image source: Bergstra & Bengio, JMLR 2012

Hyperparameter Optimization

- Grid Search:
  - Parameters:
    1. Param1 = { 0.001, 0.01, 0.1, 1, 10.0 }
    2. Param2 = { 1, 2, 4, 8, 16 }
    3. Param3 = { 1, 10, 100, 1000, 10000 }

Iterations: 5 x 5 x 5 = 125 (can be parallelized)

Hyperparameter Optimization

- Random Search:
  - Parameters:
    1. Param1 = random value(0.001, 10.0)
    2. Param2 = random value(1, 16)
    3. Param3 = random value(1, 10000)

Iterations: N (can be parallelized)

Hyperparameter Optimization

- Bayesian Optimization:
  - Starts from M initial HP combinations.
  - Parameters:

    Maps the performance of the M initial parameters in a Gaussian Space and use the priors to calculate the next HP point to query.

Iterations: M initial points + N iterations (serial)

Hyperparameter Optimization

- Bayesian Optimization

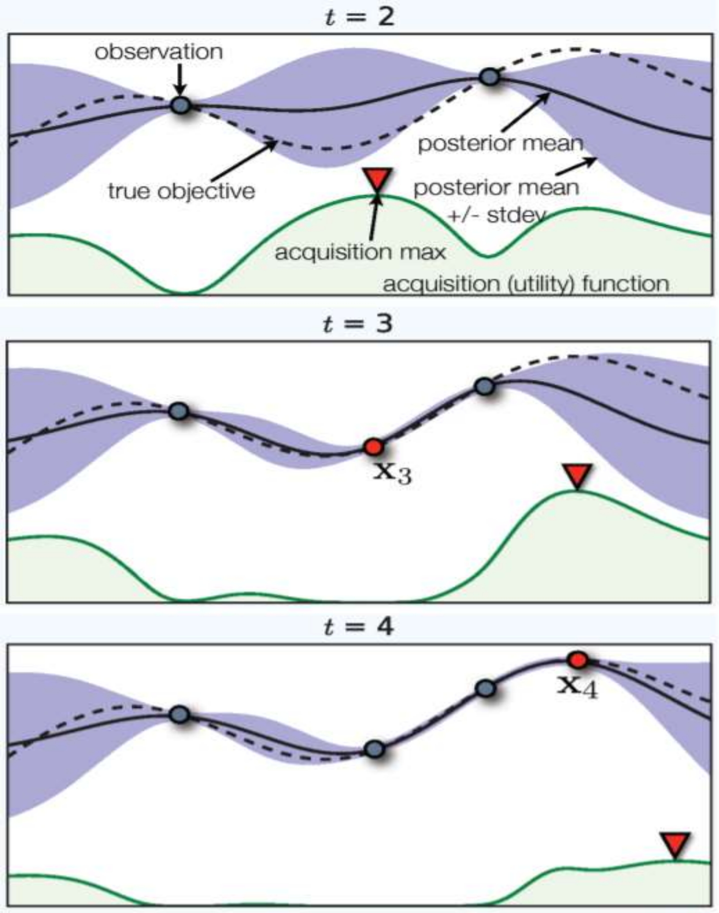


Image source: Brochu et al, 2010

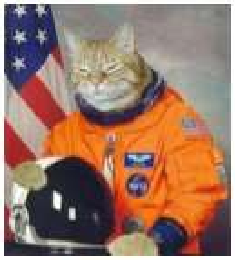
Ensembling

- Combine N fitted models in order to improve overall accuracy.
- Model Selection Algorithm.
- Training Strategy.

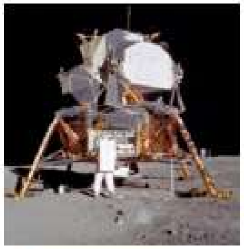
Model Interpretability

- Understand a model is very important (trust).

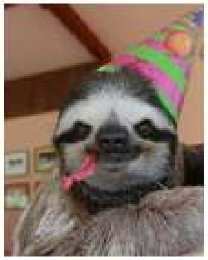
- Add mechanisms to enable model interpretation.



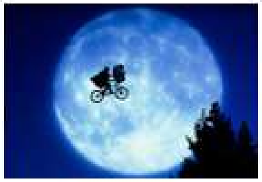
person



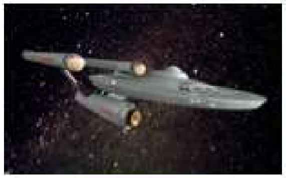
cement mixer



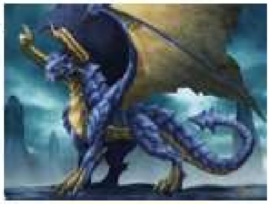
person



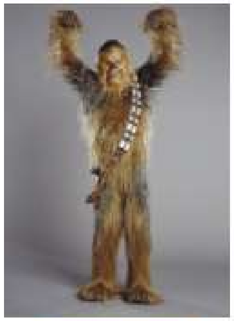
spotlight



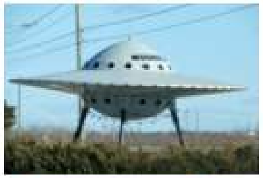
flying boat



American lobster

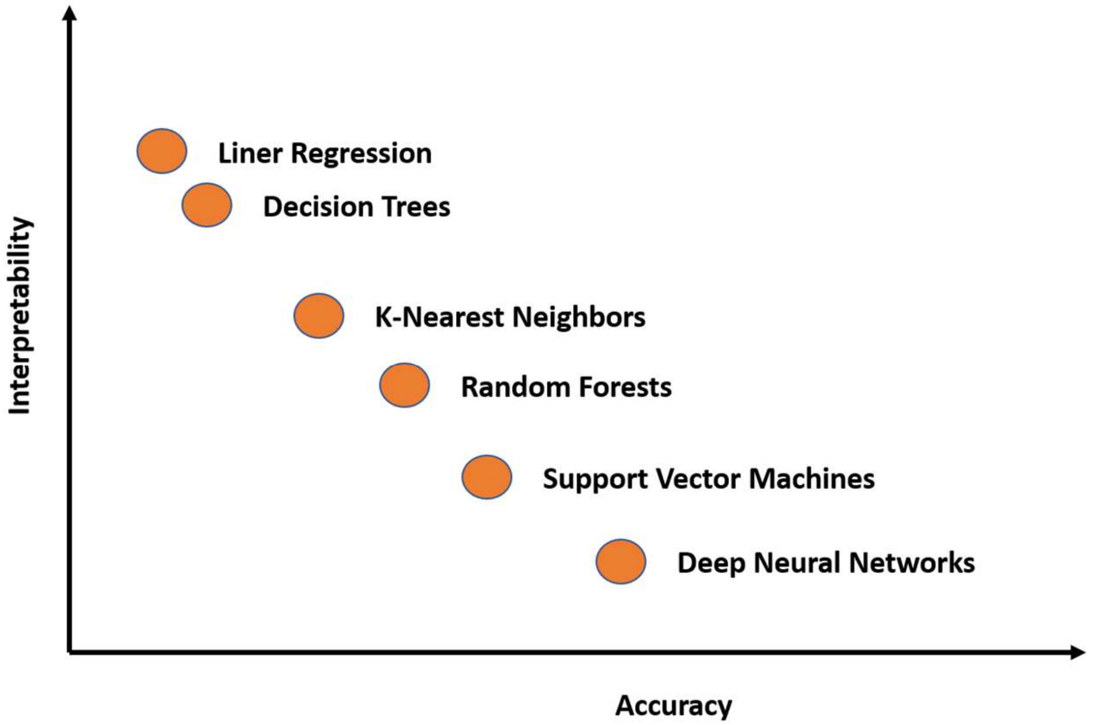


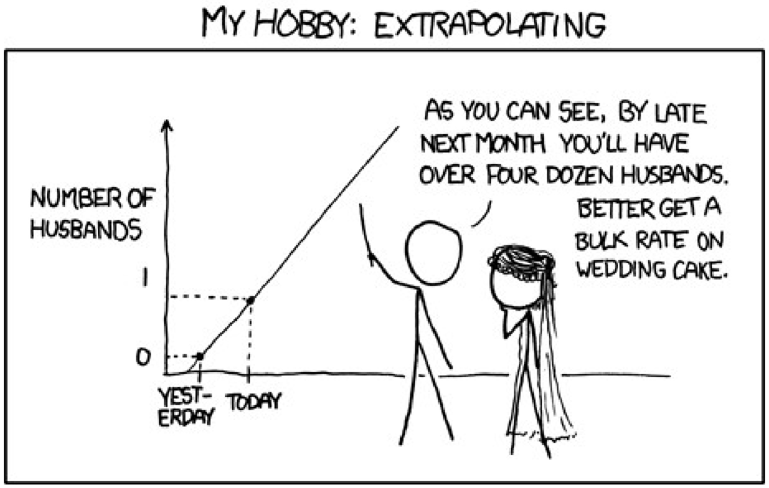
hunting dog



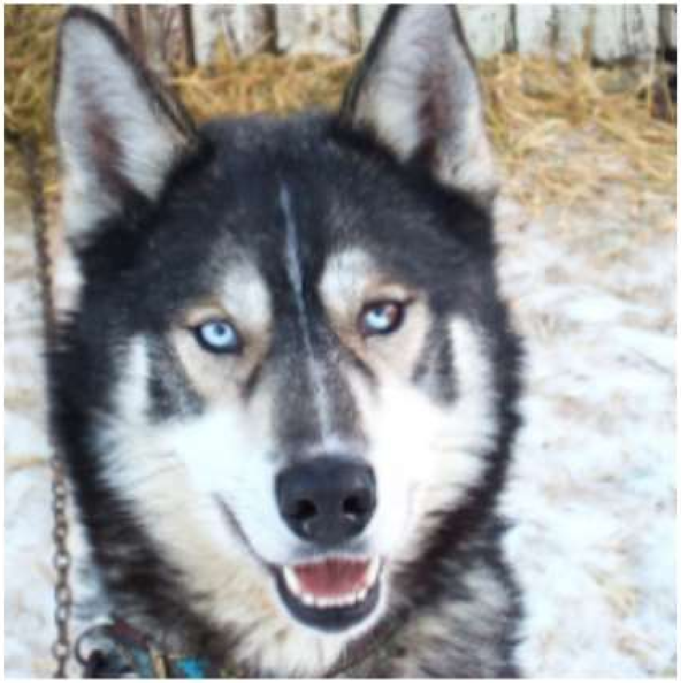
stealth bomber

Model Interpretability

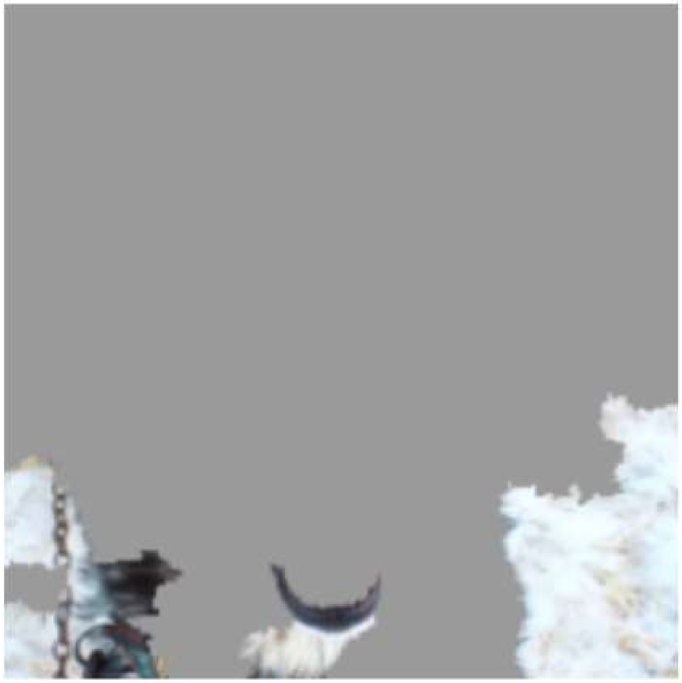



MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

Model Interpretability



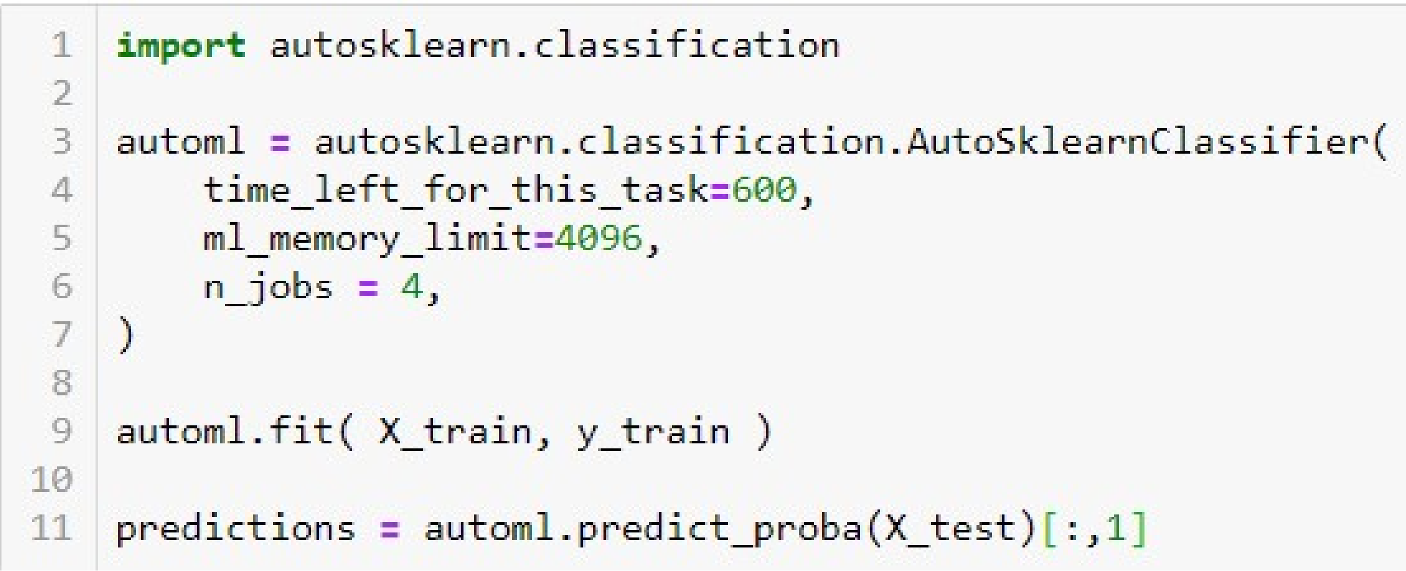
(a) Husky classified as wolf



(b) Explanation

LIME: https://arxiv.org/abs/1602.04938

AutoML

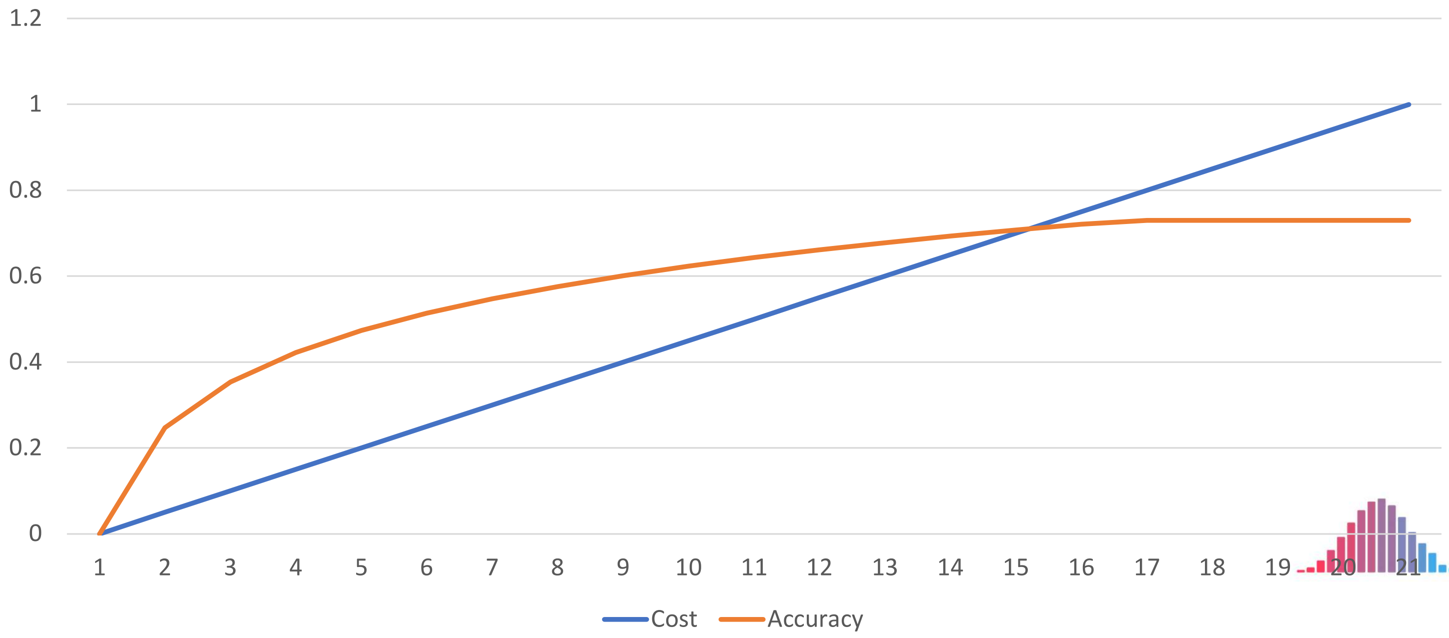


```python
import autosklearn.classification

automl = autosklearn.classification.AutoSklearnClassifier(
    time_left_for_this_task=600,
    ml_memory_limit=4096,
    n_jobs = 4,
)

automl.fit( X_train, y_train )

predictions = automl.predict_proba(X_test)[:,1]
```

kaggle DAYS

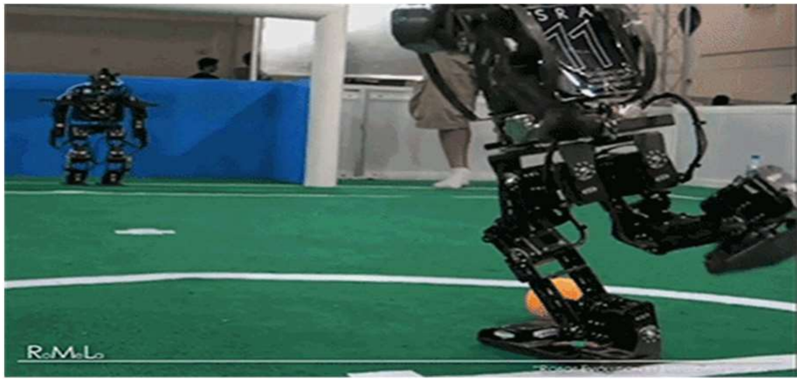
AutoML Challenges

- Cost and Model Performance vs Time
  - More time, better performance, higher costs.



AutoML Challenges

- Dataset Join/Merge
- Feature Discovery
- Feature Processing
- Feature Imputation
- Feature Engineering
- Feature Selection
- Model Selection
- Ensemble Models
- Hyper Parameter Optimization
- Big Data/Scalability (Large RAM or distributed systems).
- Cost
- Maintainability.
- Interpretability.
- Deploy



kaggle DAYS

AutoML Advantages

- Near zero complexity for the user.
- Reduce costs of hiring ML Experts.
- Reduces human bias a errors.
- Increases productivity (“time to reward”).
- Fast Insights about the data and performance.
- State-of-the-art in ML.
- Easy Scalable.
- Easy Deployable.

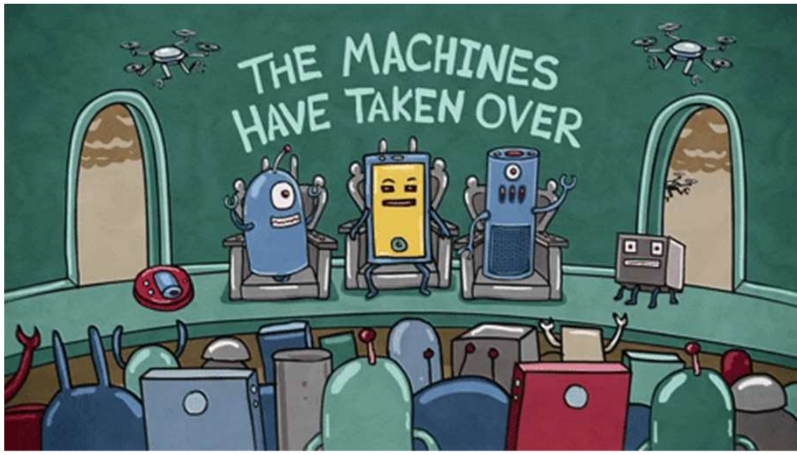
kaggle DAYS

Thank You

Find me @

www.linkedin.com/in/giba1

https://www.kaggle.com/titericz



kaggle DAYS