daria

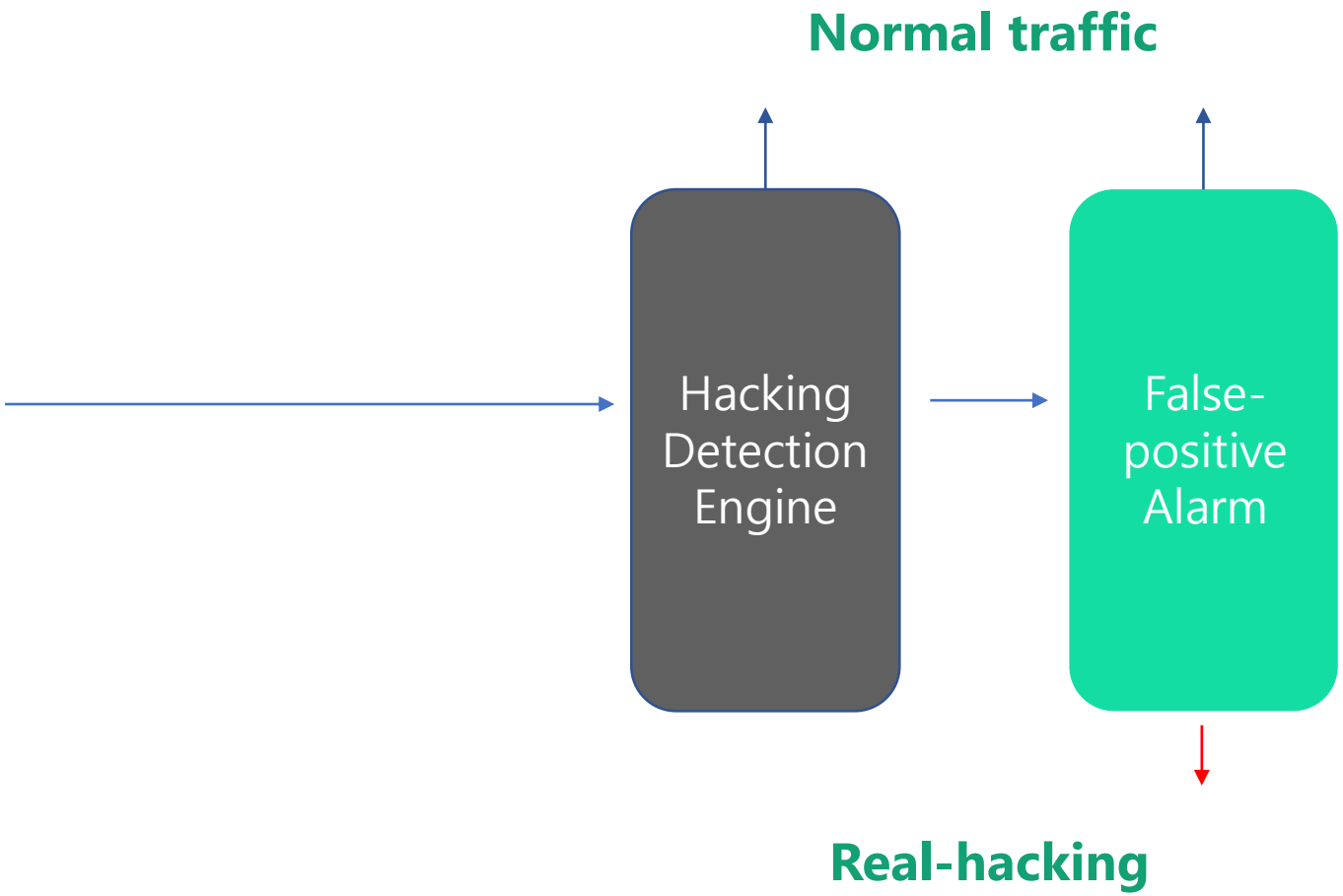
# AutoML 툴과 개발 과정의 어려움

최진영

XBRAIN

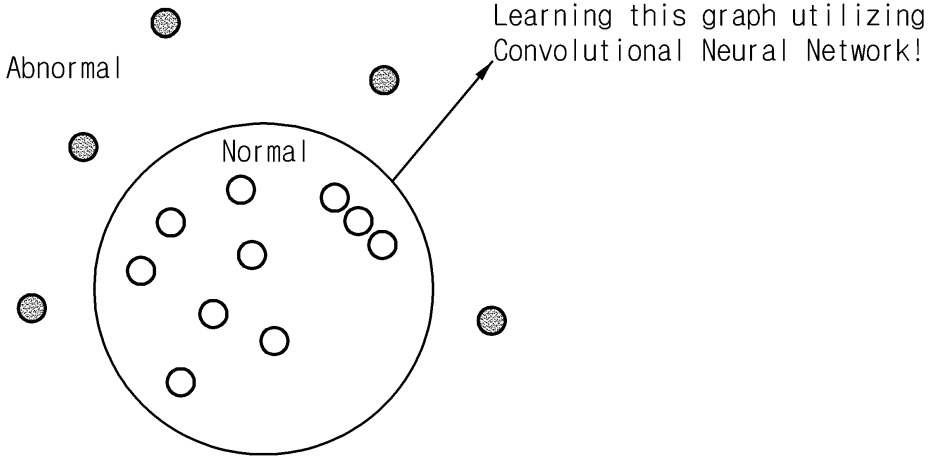
# Motivation

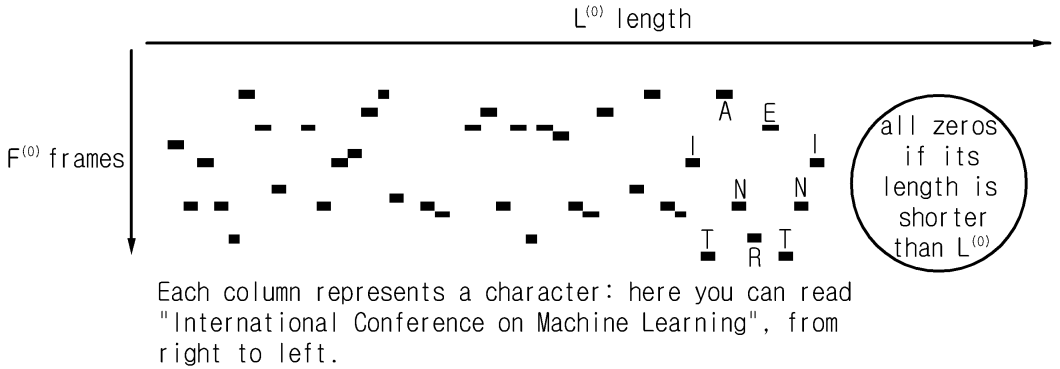
머신러닝을 활용한 웹 해킹 시도 오탐 탐지 (2014-2015)

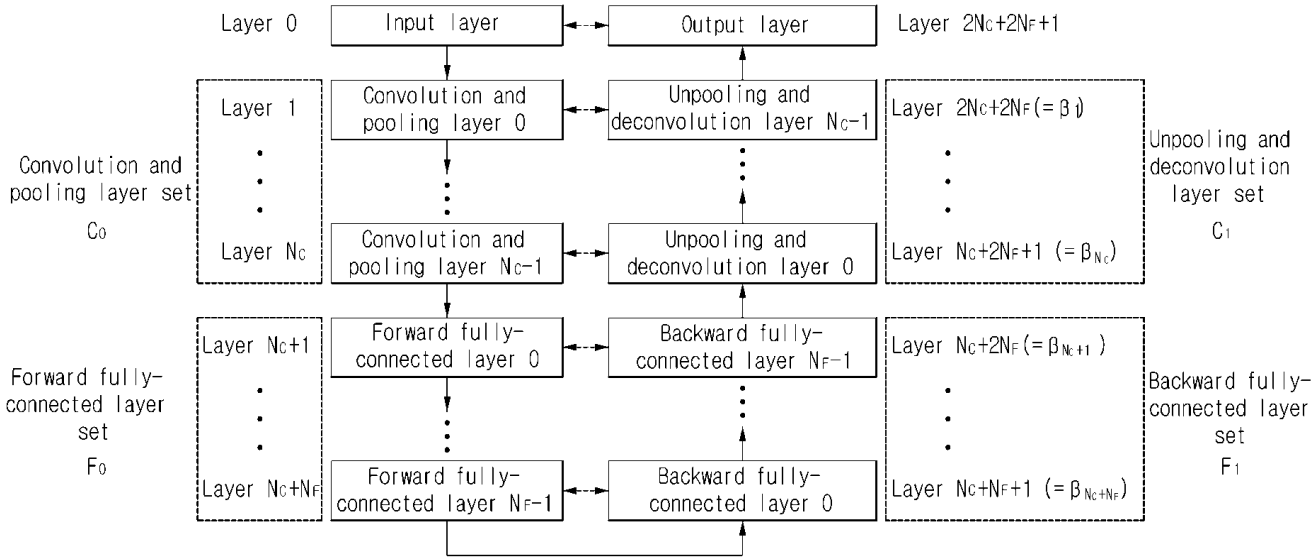


# Motivation

## Convolutional Neural Network 를 활용한 웹 해킹 이상 탐지 (2015)



# Motivation

## 반복적인 모델링 과정, 조금 더 편리하게 할 수는 없을까?

**Process of Weeks and Months**
통상 몇 주에서 수개월 걸리는 작업

**Data Source**
데이터

**Import Data**
데이터 들여오기

**Preprocessing & Preparation**
데이터 전처리 및 준비

- Form feature set
- Engineer features
- Select features/ reduce dimensionality for training
- Transform features for training
- Input missing values
- Handle outliers
- Check data types
- Split training set and test set

**Algorithm Selection / Hyperparameter**
알고리즘 선택 및 하이퍼패러미터 튜닝

- Identify learning algorithms from libraries
- Find right algorithms for given training data
- Select algorithms and modeling techniques
- Tune hyperparameters

**Model Assessment / Validation**
예측모델의 측정 및 검증

- Select optimization metric
- Validate partitioning
- Check for overfitting
- Analyze ROC curves, confusion matrix, etc. for classification models
- Analyze RMSE, MAE, coefficient of determination for regression models
- Score and compare models

**Black Box Model**

**Deployment**
예측모델 배포

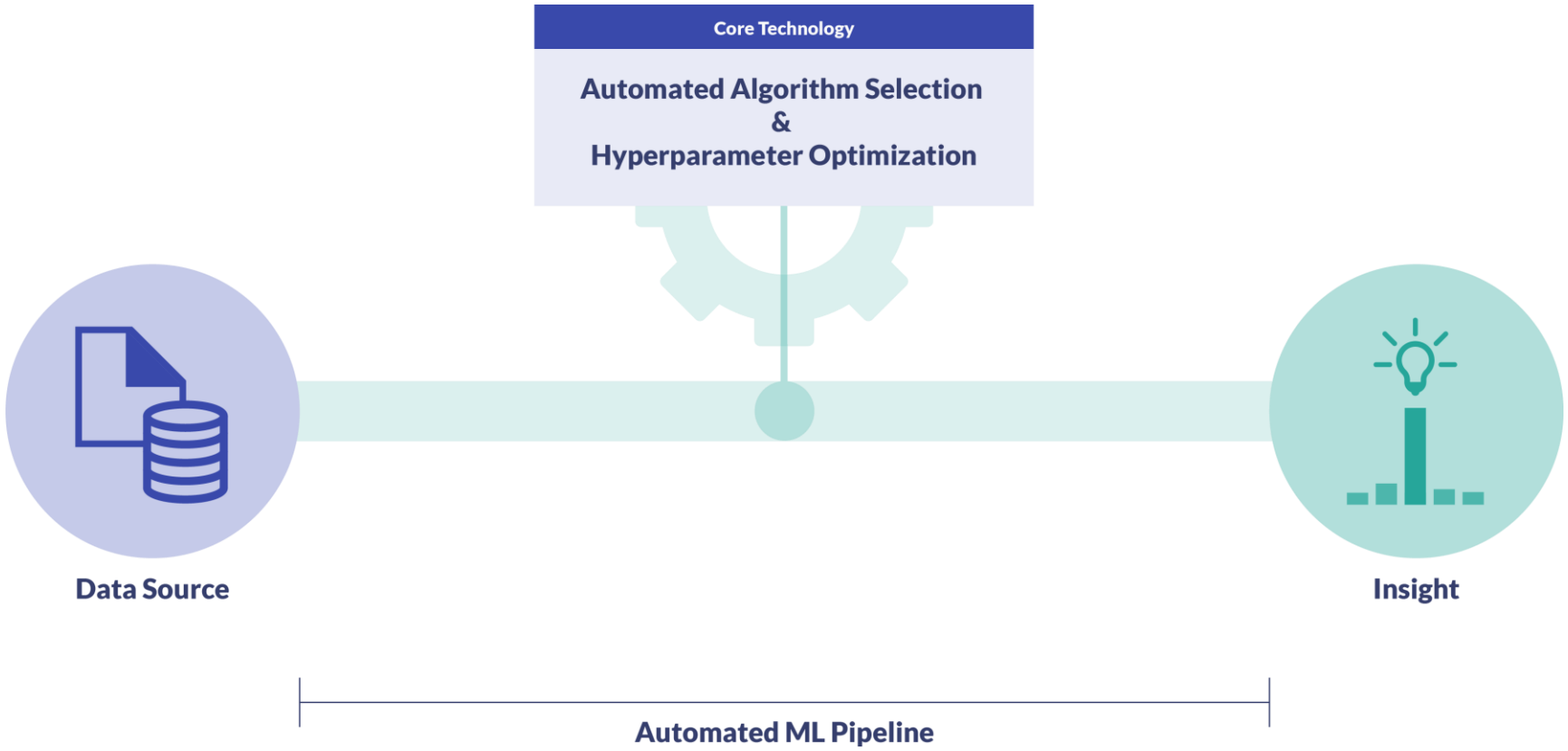
- Implement model in application
- Prepare code in multiple languages based on predictions
- Test deployment API

**Predictions**
예측

**Iterative Work**
반복적인 작업

# AutoML

## 사람의 개입 없이 머신러닝(비지도학습) 을 적용하는 기술




*KDNuggets*

# AutoML

**주어진 데이터와 알고리즘에 대해 Loss 함수를 최소화하는 하이퍼파라미터를 찾는 일.**

Loss function

$$A^*_{\lambda^*} = \text{argmin}_{A^{(j)} \in \mathcal{A}, \lambda \in \Lambda^{(j)}} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}\left(A^{(j)}_{\lambda}, \mathcal{D}^{(i)}_{\text{train}}, \mathcal{D}^{(i)}_{\text{validation}}\right)$$

Algorithm & Hyperparameter

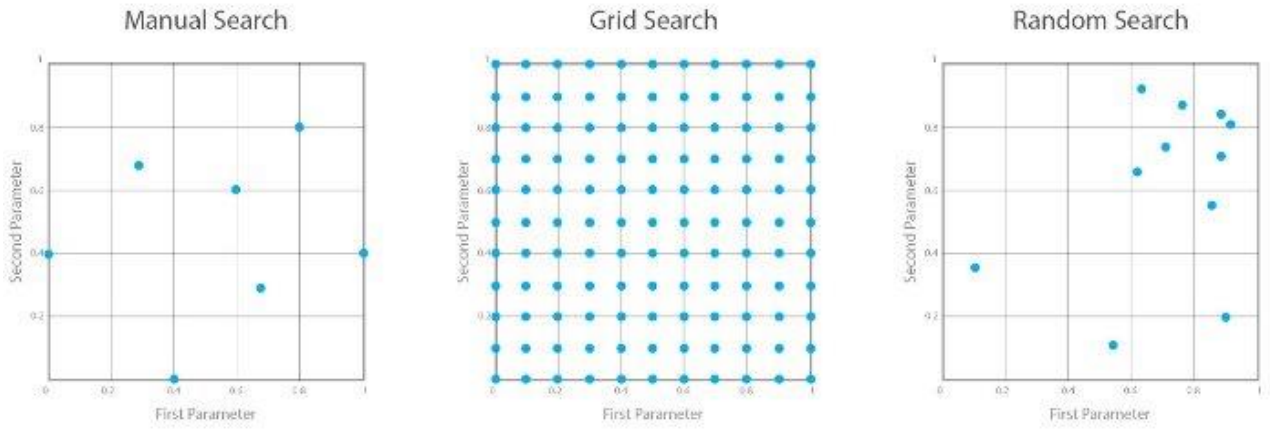
Data

*Thornton et al, 2013.*

# AutoML – Hyperparameter Optimization

## 모델 선택과 관련된 다양한 방법론

- Manual search
- Random search
- Grid search
- **Bayesian optimization**
- Evolutionary algorithm




*KDNuggets*

# AutoML – Bayesian Optimization

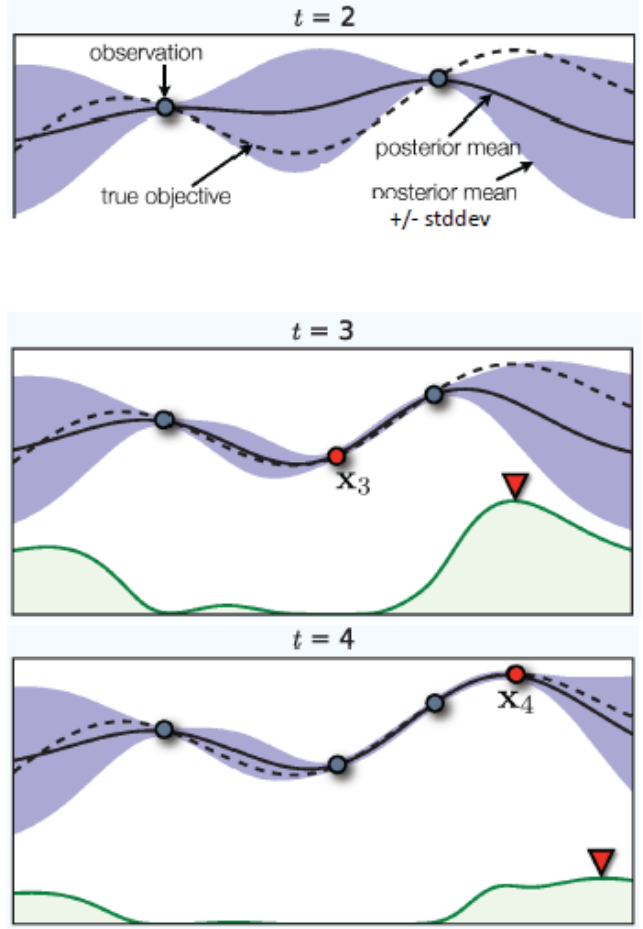
## 베이지안 최적화 - 최적 하이퍼파라미터 조합을 모델기반의 최적화 기법 활용하여 탐색

- 탐색 대상 함수와 해당 하이퍼파라미터를 쌍(pair)을 대상으로 **Surrogate Model 을 만들고 평가를 통해 순차적으로** 업데이트해 가면서 **최적의 하이퍼파라미터 조합을 탐색**한다.
  1. Build a surrogate probability model of the objective function
  2. Find the hyperparameters that perform best on the surrogate
  3. Apply these hyperparameters to the true objective function
  4. Update the surrogate model incorporating the new results
  5. Repeat steps 2–4 until max iterations or time is reached

- 왜 Bayesian 이라고 부르는가?

  It is called **Bayesian** because it uses the famous "**Bayes' theorem**", which states (simplifying somewhat) that the posterior probability of a model (or theory, or hypothesis) ***M*** given evidence (or data, or observations) ***E*** is proportional to the likelihood of ***E*** given ***M*** multiplied by the prior probability of ***M***.

$$P(M|E) \propto P(E|M)P(M)$$




*Eric Brochu et al, 2010*

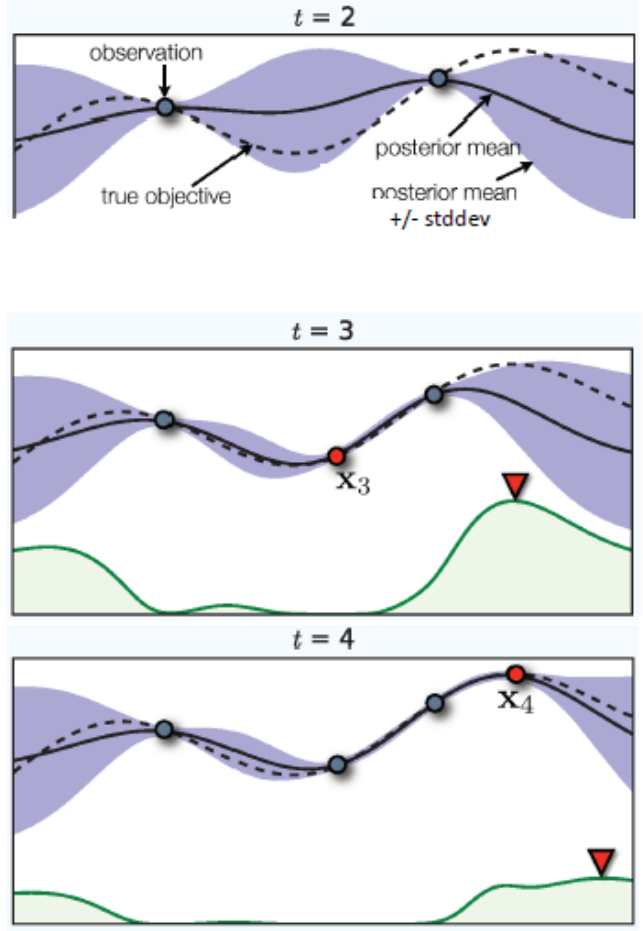
# AutoML - SMBO

## Sequential Model-Based global Optimization (SMBO)

SMBO$(f, M_0, T, S)$

1 $\mathcal{H} \leftarrow \emptyset$,
2 For $t \leftarrow 1$ **to** $T$,
3 $x^* \leftarrow \text{argmin}_x \; S(x, M_{t-1})$,
4 Evaluate $f(x^*)$, ▷ *Expensive step*
5 $\mathcal{H} \leftarrow \mathcal{H} \cup (x^*, f(x^*))$,
6 Fit a new model $M_t$ to $\mathcal{H}$.
7 **return** $\mathcal{H}$

*J. Bergstra et al, 2013*




*Eric Brochu et al, 2010*

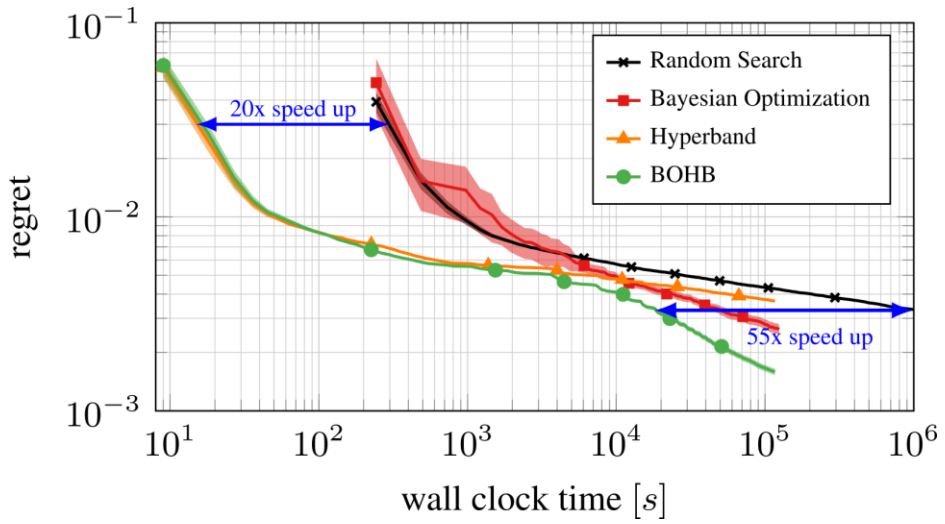
**결과에 영향을 끼치는 주요 요소**

- Surrogate function (or response surface)
  - 어떠한 모형으로 Evaluation points 를 fitting 할 것인지?
  - Gaussian Process, Random Forest Regression, TPE, etc.
- Acquisition function
  - Next evaluation point 를 정하는 기준
  - Probability of Improvement, Expected Improvement, GP-UCB, etc.

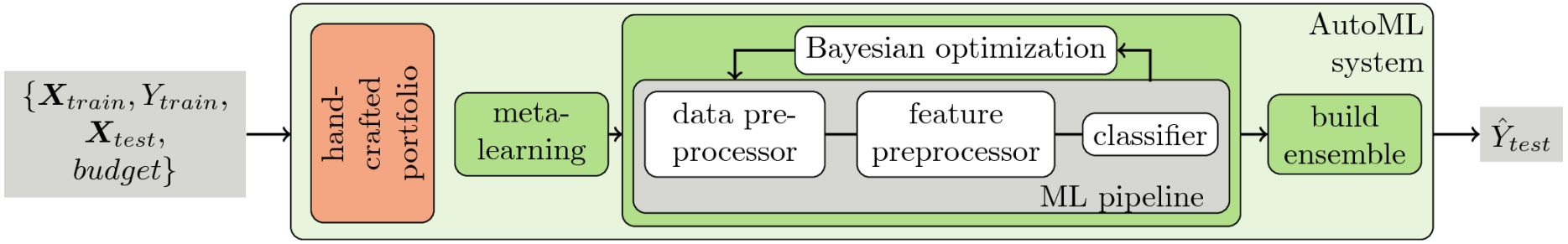
# Libraries

## 다양한 방법론들과 오픈소스 라이브러리들

- auto-sklearn
- Bayesian Optimization
  - Hyperopt (TPE)
  - Spearmint (GP)
  - SMAC (Random forest regression, Hutter et al, 2011)
  - Etc..
- Hyperband (Li et al, 2016)
- BOHB (Falkner et al, 2018)
  - Hyperband + Bayesian Optimization



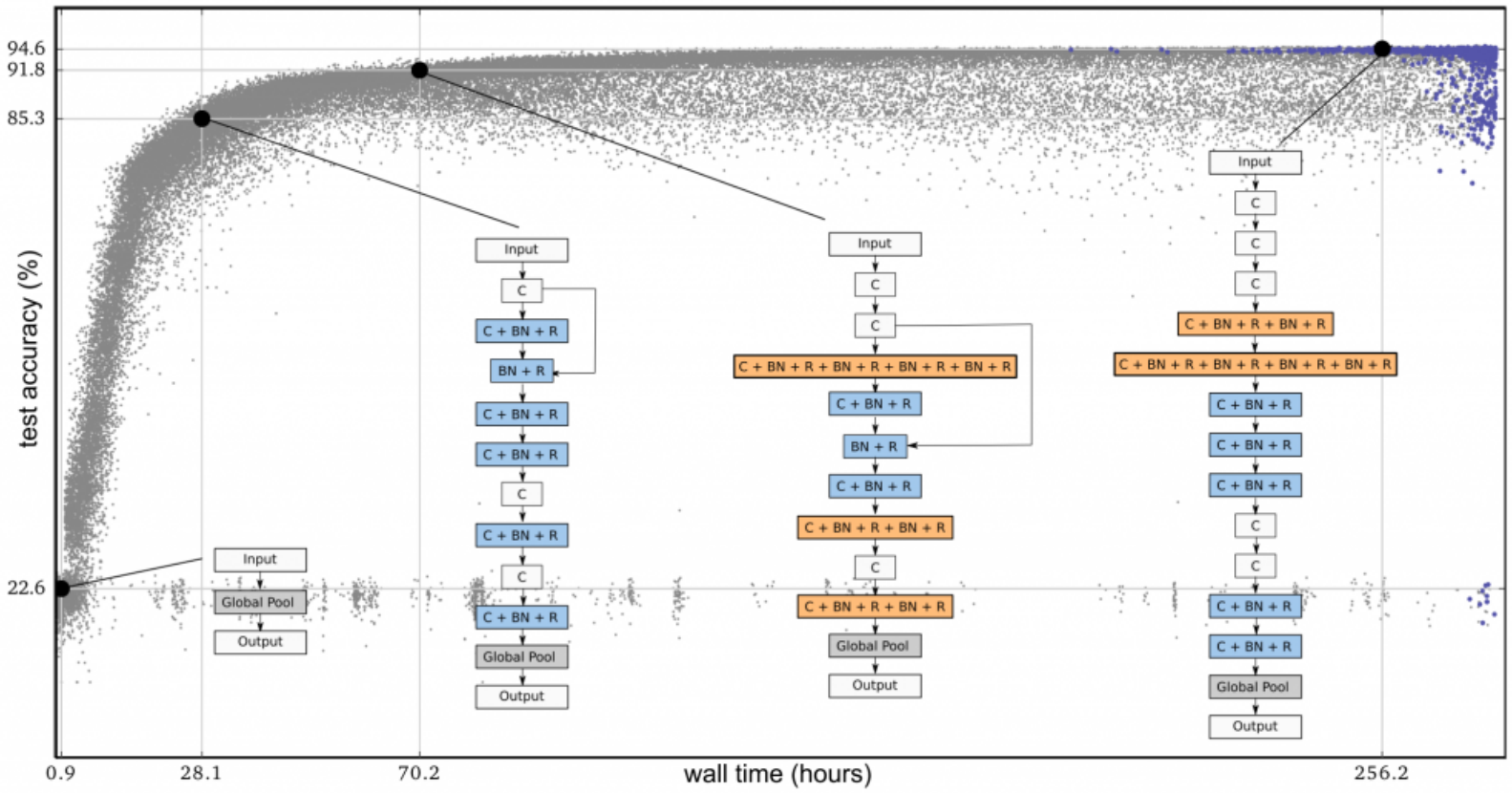

automl.org, BOHB




automl.org, auto-sklearn

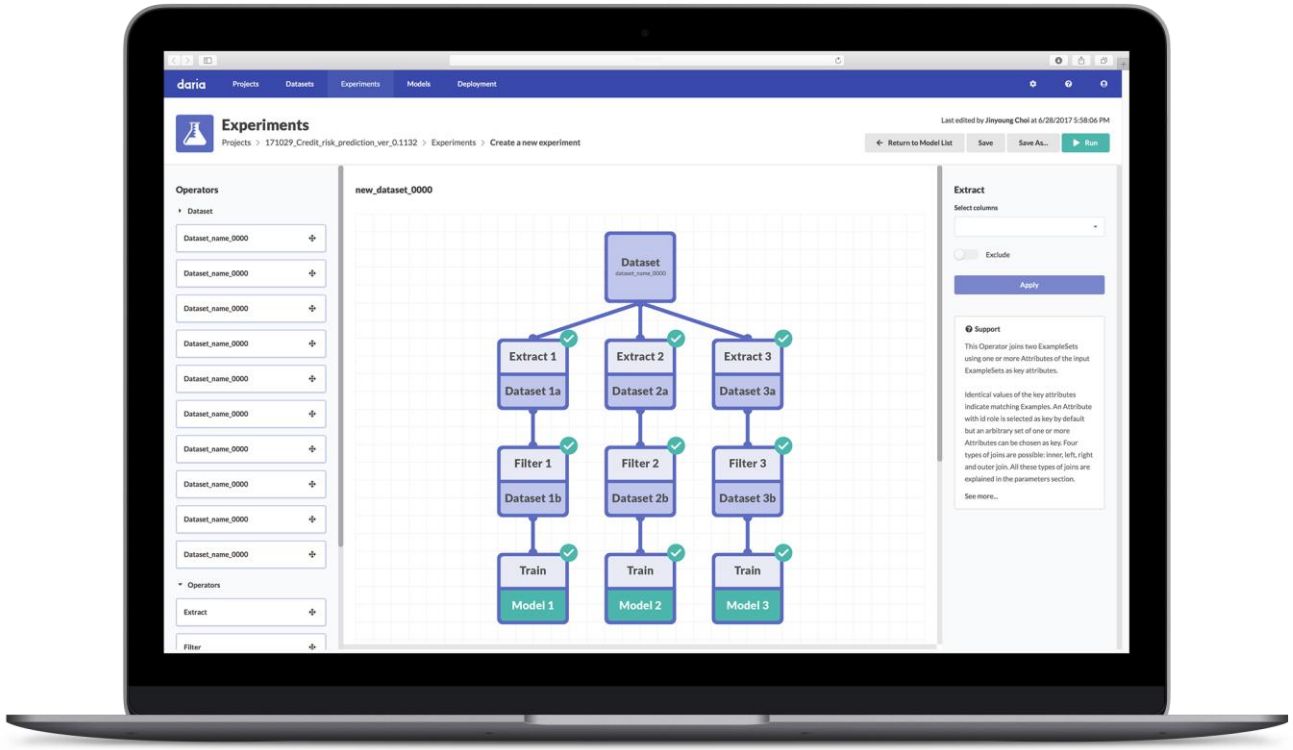
# AutoML – Deep Learning

## Neural Architecture Search (NAS)



# Daria, Your Partner in Machine Learning Greatness

## 조직의 예측분석 역량을 증진하기 위한 손쉽고 실용적인 도구




### Remove Barriers

기업들이 AI 시스템을 구축하기 위한
기술적/ 재무적 장벽을 걷어냅니다

### Streamline Workflows

자동화된 머신러닝을 통해서 긴 시간이
소요되는 반복적인 작업을 간소화합니다

### Easy Start

입문자를 위해 직관적인 GUI를 통해서
손쉽게 예측분석을 시도해볼 수 있습니다

# 개발 과정에서의 어려움

**초기 회사로서 과업을 수행하며 느꼈던, 그리고 느끼고 있는 어려운 점들.**

## 시장의 성숙도로 인한 제품의 사용성 검증의 어려움

- 제한된 자원으로 소프트웨어 개발 위해서는 **다양한 사용자 페르소나 (직군, 산업군 등)과의 빠른 제품 검증**이 중요
- 대부분 데이터 분석 및 모형 개발의 중요성에 대해서는 알고 있지만, **문제 정의에 많은 어려움을 겪는다.**
  - 일부 전통적인 산업군(금융, 통신, 게임 등) 을 제외하고는 예측 모형 개발을 주기적으로 수행하는 곳이 제한적
  - 과거에 비해 데이터 분석을 위한 인프라 환경이 개선되고는 있지만, 여전히 많은 영역에서는 어려워 함.
  - 다행히 최근에 개발 프레임워크의 다양화 및 교육 환경이 개선되며 풀이 늘어나는 추세.

  ** 이에 따라 여러 도메인에서 새로운 예측 모형 개발 사례들이 증가하고 있음.*

## 인력 채용의 어려움 – 데이터 분석가 < 소프트웨어 개발

- 데이터 분석 및 예측 모형 개발에 대한 인력 풀의 부족
- 최근 데이터 분석 및 과학에 대한 수요 증대로 인해 많은 인력들이 양성되고 있음.
- 종종 우리 회사에 지원하시는 분들 중에 훌륭한 “데이터 분석가 " 분들이 계시지만,
- 엑스브레인에서는 다리아와 같은 **시스템을 개발하는 것에 흥미를 느끼는 소프트웨어 엔지니어링 역량을 갖춘 분**들이 우선적으로 필요.

  **물론 “데이터 분석가 " 분들도 조만간 채용 계획이 있습니다. ☺*