오픈소스 Deep Learning Framework 제작

이 태영 Tae Young Lee

Software Engineer

2014	2015	2016	2017
Change the world in IOT (Falinux) Game based on the IOT (KGC)	Python Network Programming Neural Network의 변천사를 통해 바라본 R에서 Deep Neural Net활용	Machine Learning In SPAM	[SK Planet] TensorFlow 기본 AutoML & AutoDraw [SK T academy] 딥러닝을 위한 TensorFlow Sequence Model and the RNN API Open Stack으로 바라보는 클라우드 플랫폼

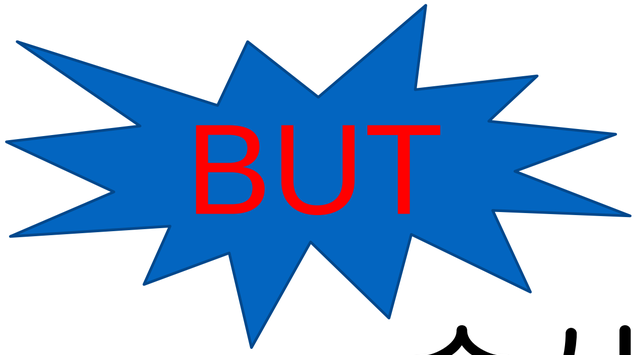
기초공사


골조공사


외장공사


내장공사


완성



순서 상 가장 중요한 것


설계

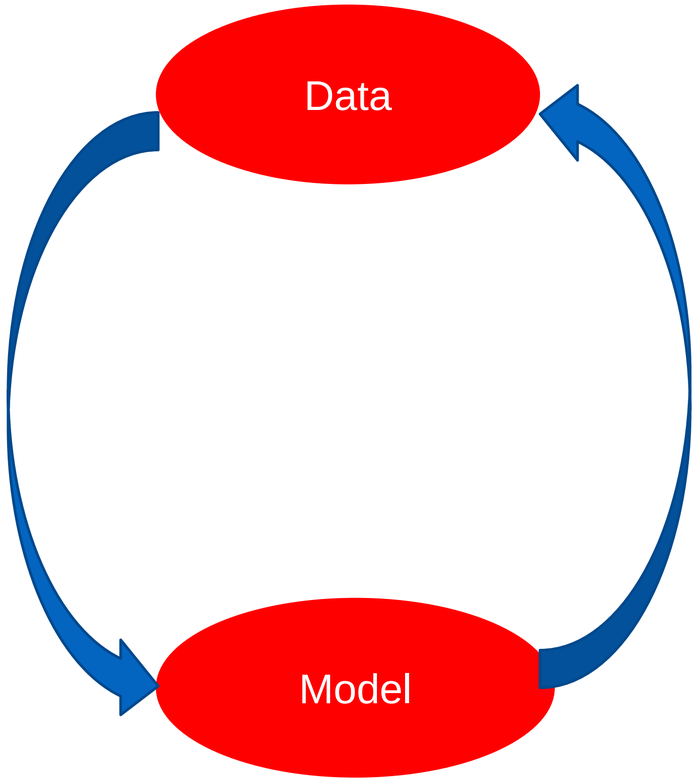
잊어서는 안 되는 것

공사설계 스케줄


건 축 주
사 업 결 정
대 지 매 입
토목설계업체 선정
규 모 결 정
건축설계 업체결정
설 계 계 약
지질조사시행 (20~30공구)
기획설계(S,D) (건축설계평면, 입면, 답면 STUDY)
FEEDBACK
결 정
NO
YES
PROCESS
기 본 설 계
건축 허가 접수
FEEDBACK
결 정
NO
YES
건축 허가 득
실 시 설 계 (C,D)
물량산출 및 내역작업 발주
도 면 납 품
시공회사 선정
일반적으로 4~5개 업체 선정 후 최저,최상 견적을 제외한 후 설계 업체 + 나머지 업체(3)의 평균 오차 범위 -5% 안의 업체 + 회사의 도급순위(평균 1,2군 업체) + 현장대리인(현장 소장) 의 경험을 복합고려하여 결정
공 사 착 공
공 사
준 공
사 용

Deep Learning 관점

설계하기 위한 Process 정립 (Deep Learning Process)



- ✓ Data
- ✓ Train
- ✓ Hyper Parameter Tuning
- ✓ Optimization
- ✓ Model
- ✓ Prediction

Framework 만들 준비

집 짓기 재료준비

1> 재료 알아보기

2> 재료 활용하기

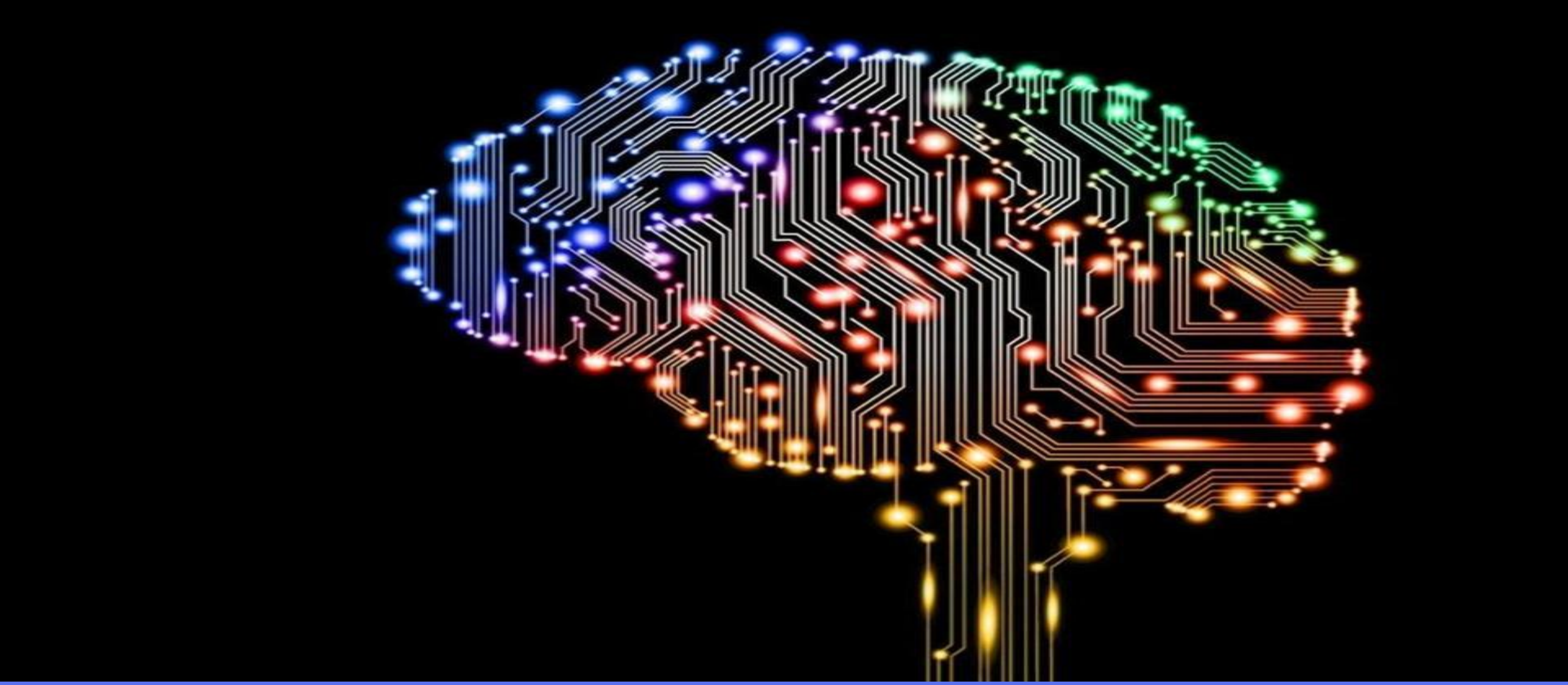
Deep Learning개념 알아보기

3> 구조 설계하기

4> 재료 조합하기

Framework개념 알아보기

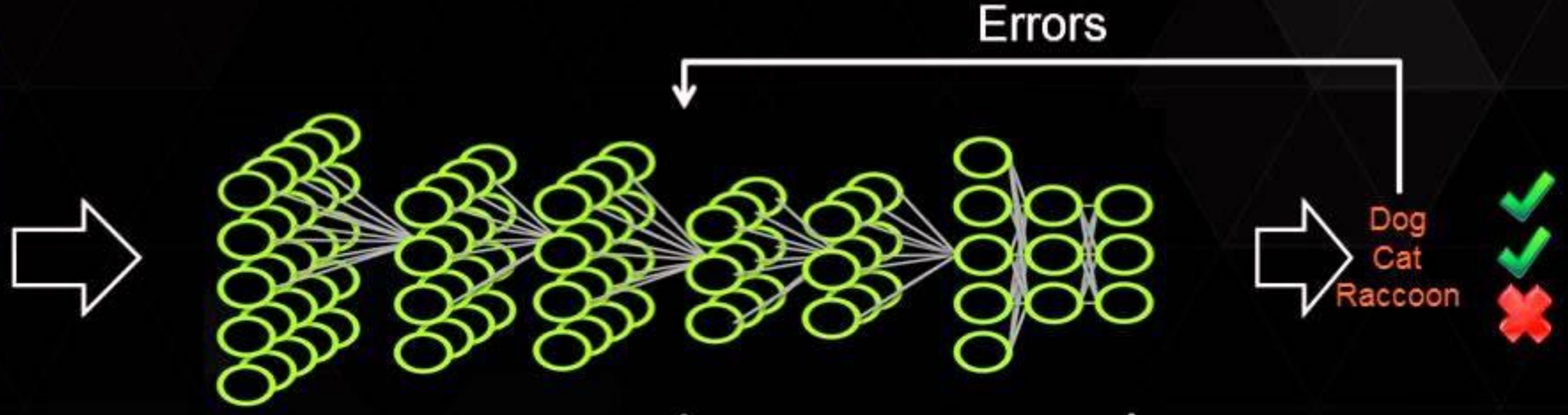
5> 활용해 보기



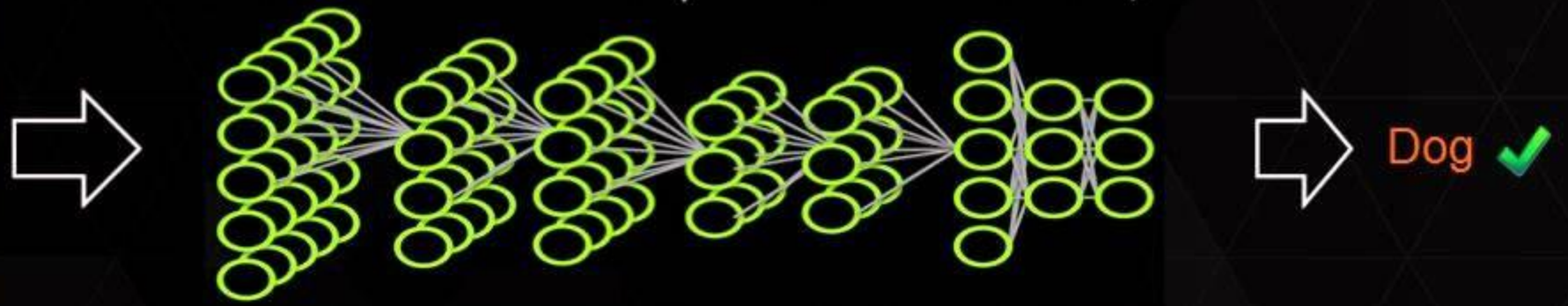
Deep Learning

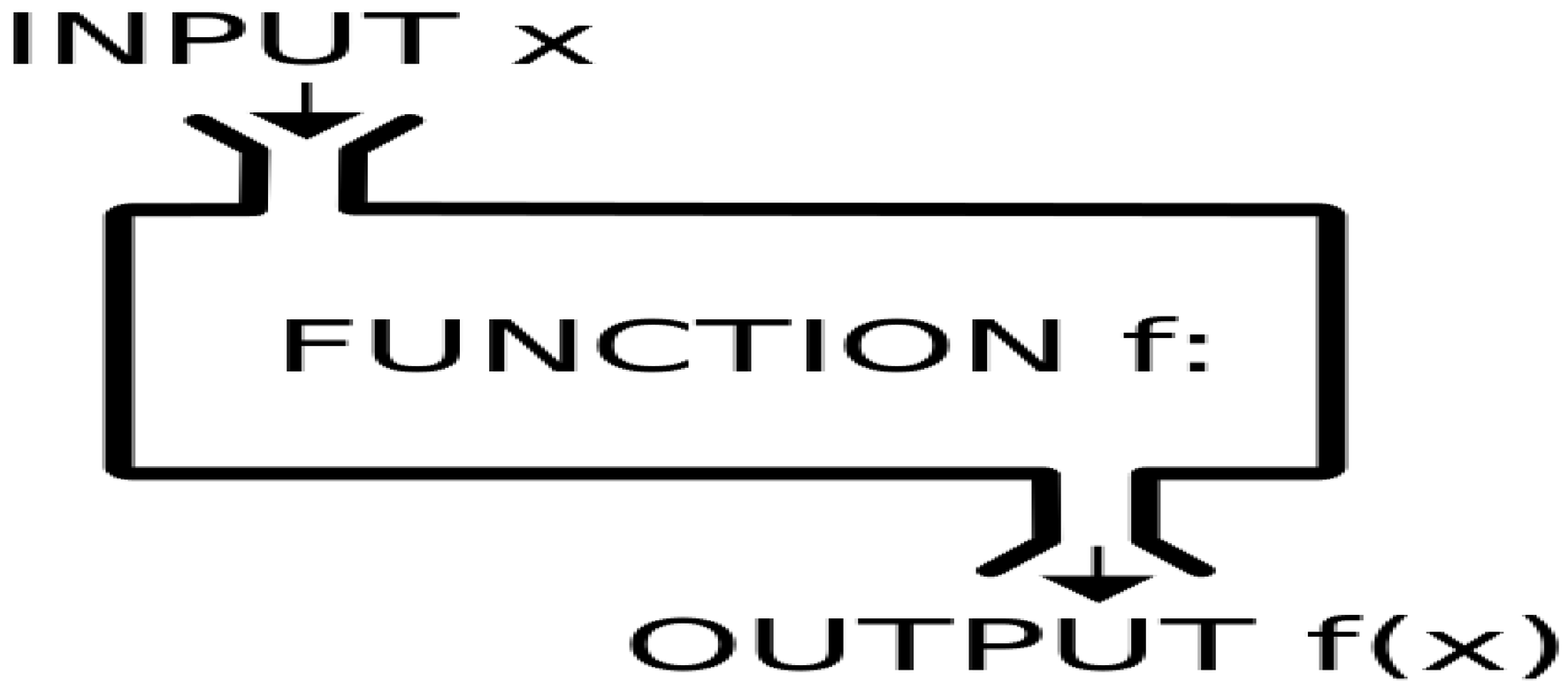
DEEP LEARNING APPROACH

Train:



Deploy:




INPUT x
FUNCTION f:
OUTPUT f(x)

Train:

Errors

MODEL

Dog
Cat
Raccoon

Deploy:

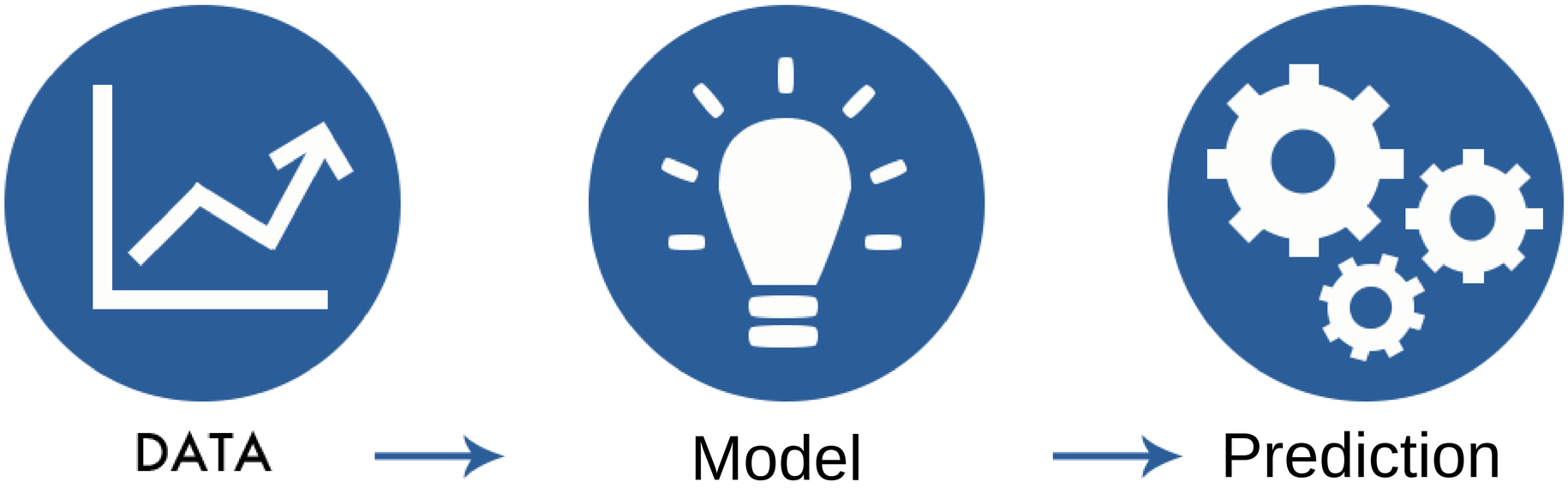
MODEL

Dog

x

$f: X \rightarrow Y$

y

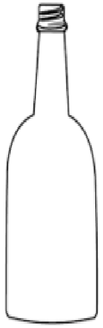

DATA
Model
Prediction

Data 관점

윤곽선 검출
Edge Detection

teapot

apple



bottle

윤곽선 템플릿들

생각해 봅시다.

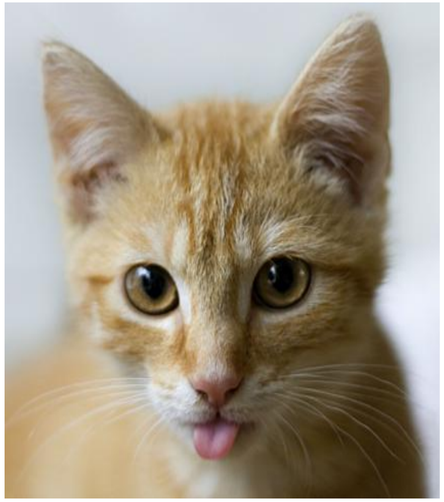
반대로 돌아간 **teapot**은**?**

무늬가 있는 **teapot**은**?**

특이하게 생긴 **teapot**은**?**

배경과 섞여 있는 특이한 **teapot**은**?**

사람은**?**



고양이는**?**

데이터 수집

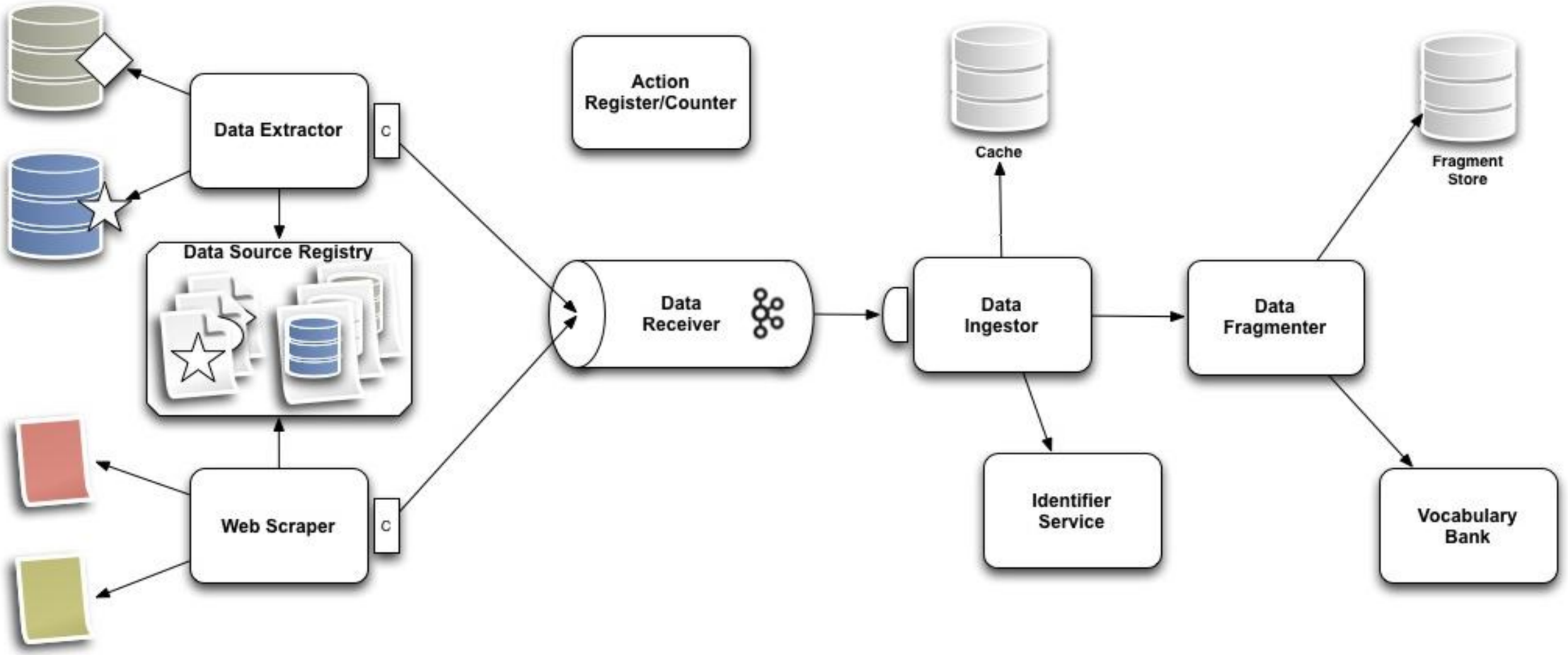
Streaming Data

Data Selection

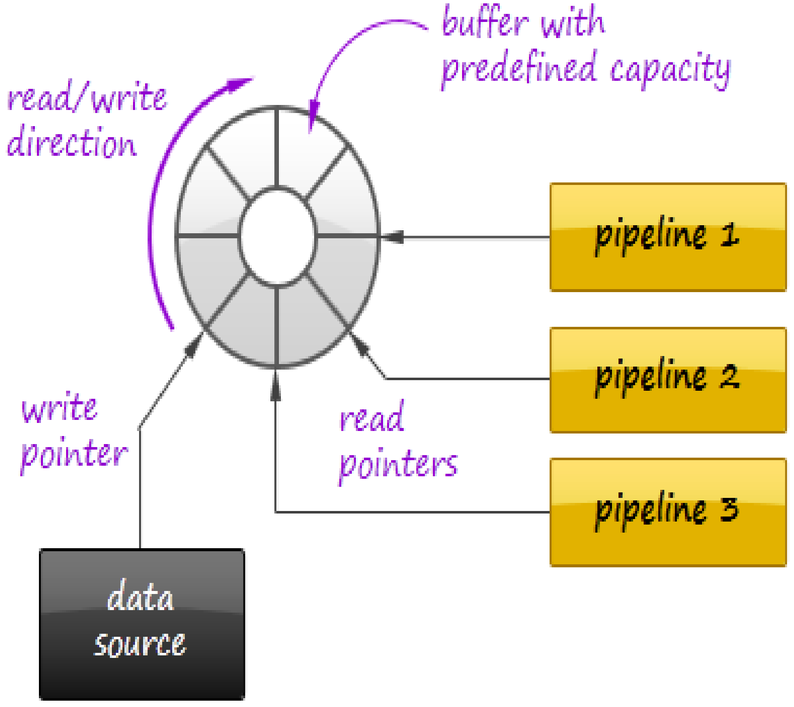
Data Cleaning

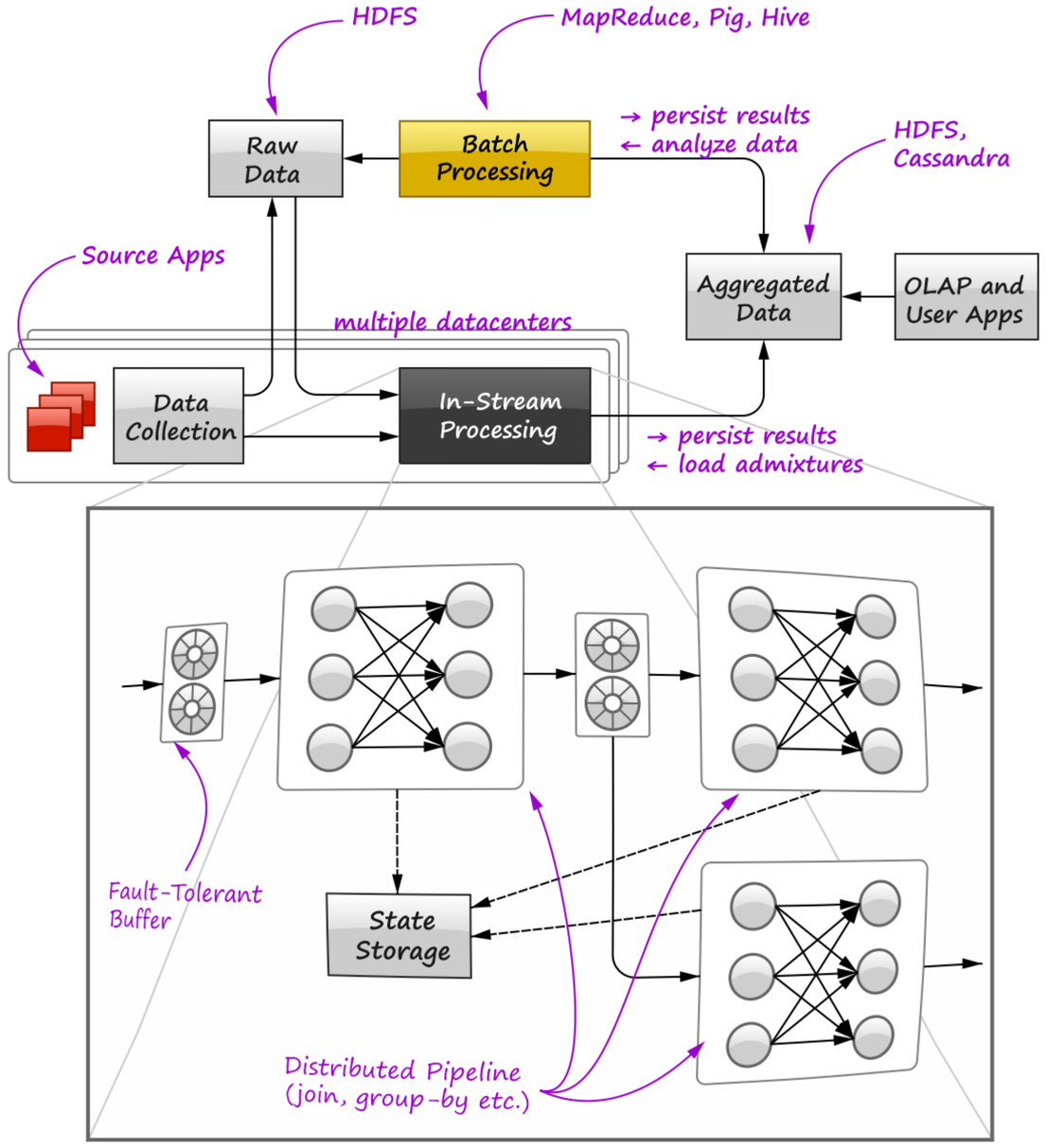
Data Augmentation

Data Pre-Processing


Data Extractor
C
Data Source Registry
Web Scraper
C
Action
Register/Counter
Data
Receiver
Cache
Data
Ingestor
Identifier
Service
Data
Fragmenter
Fragment
Store
Vocabulary
Bank

Streaming Data


buffer with
predefined capacity
read/write
direction
pipeline 1
pipeline 2
pipeline 3
write
pointer
read
pointers
data
source


HDFS
MapReduce, Pig, Hive
Raw
Data
Batch
Processing
→ persist results
← analyze data
HDFS,
Cassandra
Source Apps
Aggregated
Data
OLAP and
User Apps
multiple datacenters
Data
Collection
In-Stream
Processing
→ persist results
← load admixtures
Fault-Tolerant
Buffer
State
Storage
Distributed Pipeline
(join, group-by etc.)

데이터 평가

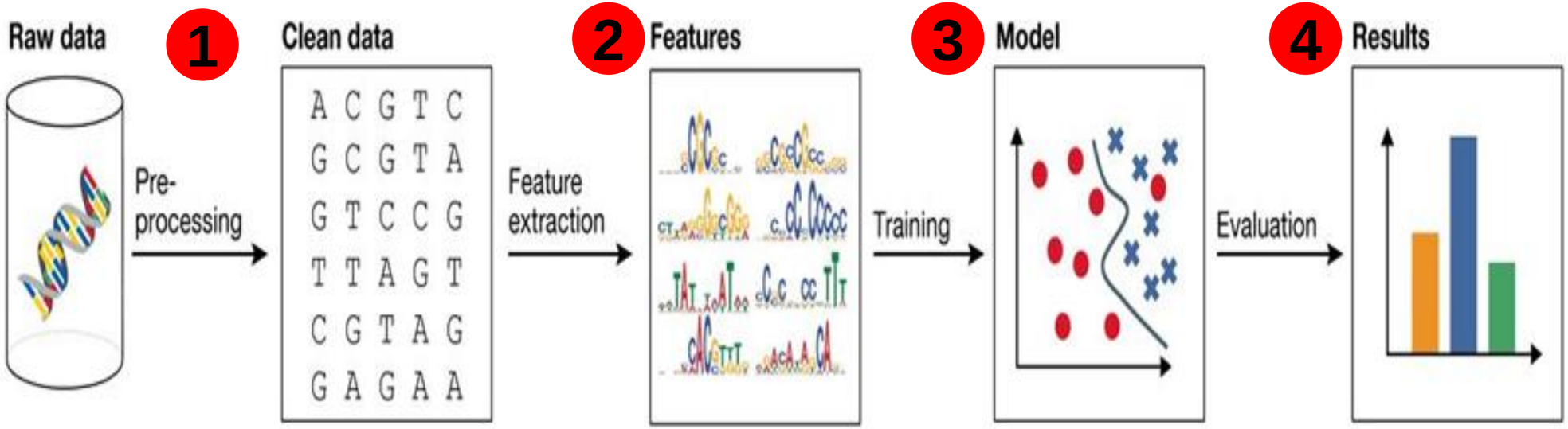
1 벽돌을 만들 재료에 불순물 제거 — Raw Data

2 벽돌 재료 배합 — Clean Data

3 벽돌 만들기 — Features

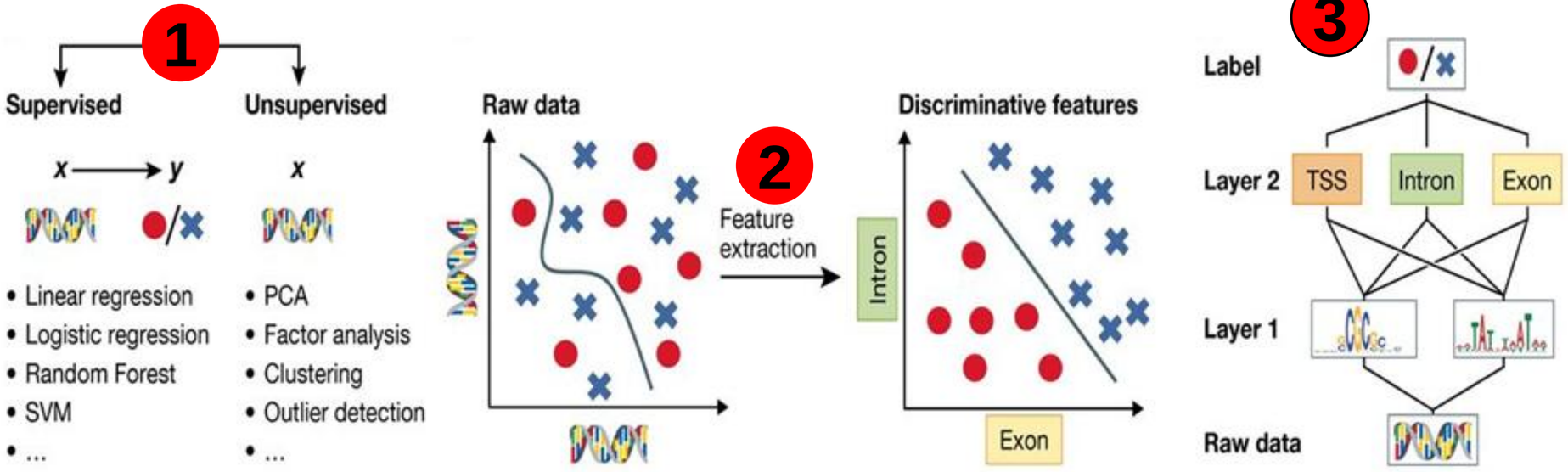
4 벽돌 내구성 테스트 — Model

Evaluation

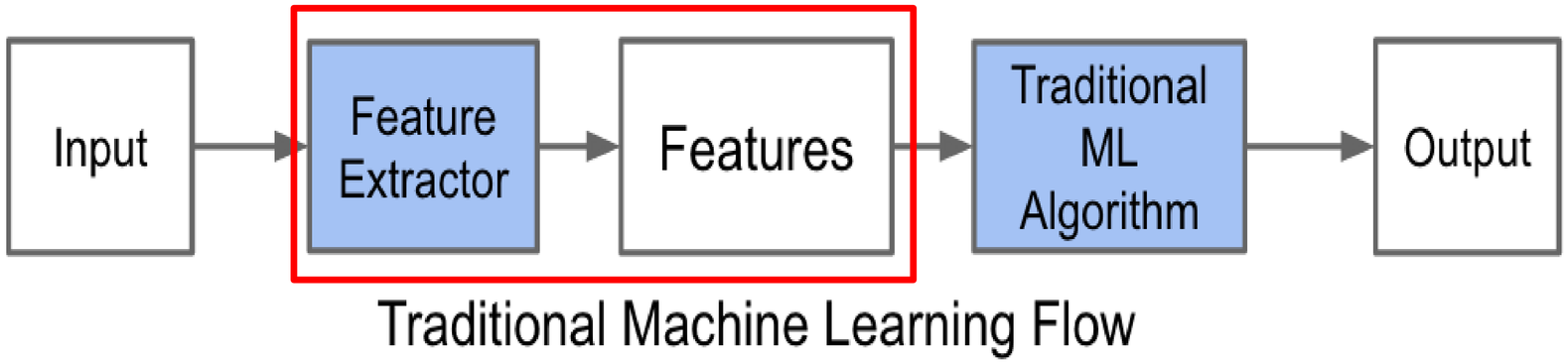


1. 재료의 특성 확인 (Label) — Supervised
2. 재료의 속성 확인 — Unsupervised
3. 구분 지어 주기 — Feature Extraction

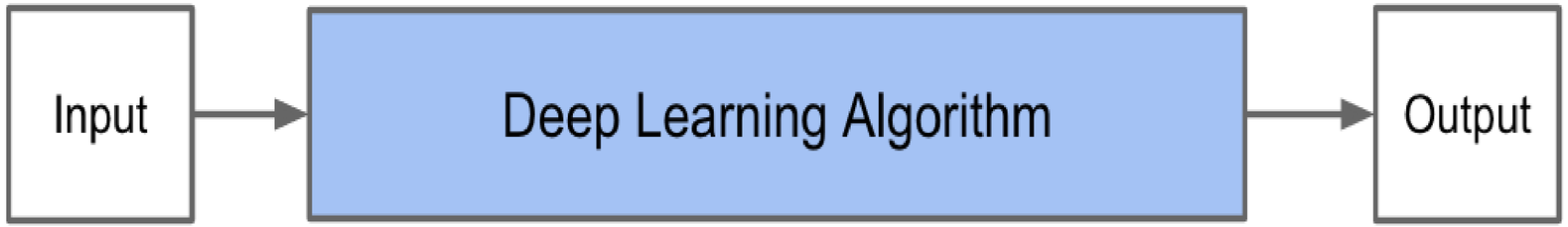
기본 재료 확보



재료 확보 Process 개선

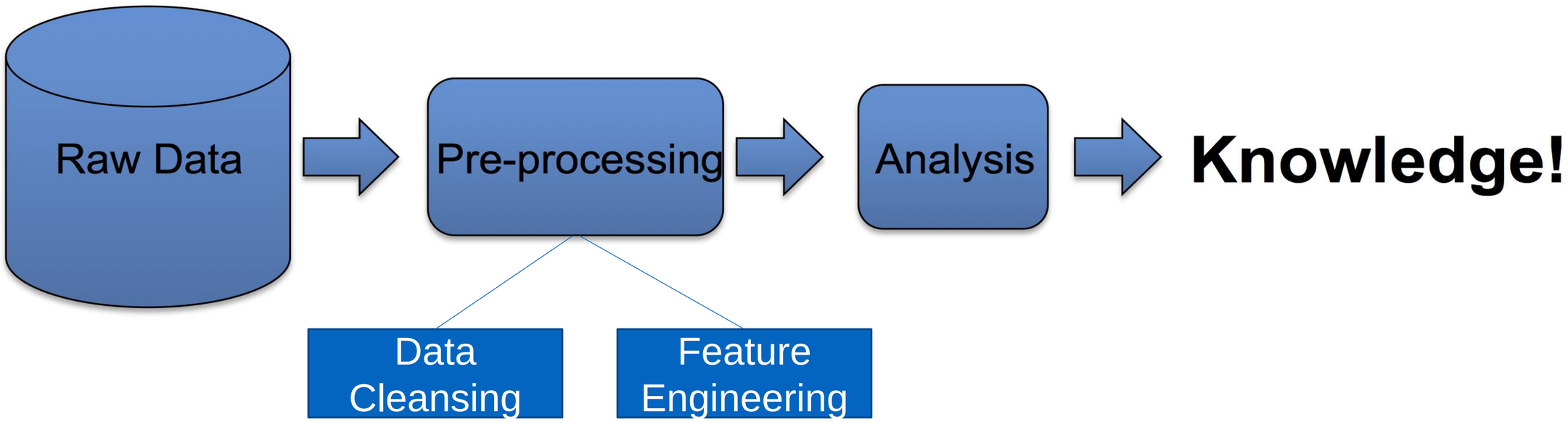


Traditional Machine Learning Flow



Deep Learning Flow

재료 확보 Process 세부 내역확인



Deep Learning이나 Machine Learning을 하기 위해선 Data를 변환 필요

Pre-Processing(Image)

Viewpoint variation(시점 변화). 객체의 단일 인스턴스는 카메라에 의해 시점이 달라질 수 있다.
Scale variation(크기 변화). 비주얼 클래스는 대부분 그것들의 크기의 변화를 나타낸다(이미지의 크기뿐만 아니라 실제 세계에서의 크기까지 포함함).
Deformation(변형). 많은 객체들은 고정된 형태가 없고, 극단적인 형태로 변형될 수 있다.
Occlusion(폐색). 객체들은 전체가 보이지 않을 수 있다. 때로는 물체의 매우 적은 부분(매우 적은 픽셀)만이 보인다.
Illumination conditions(조명 상태). 조명의 영향으로 픽셀 값이 변형된다.
Background clutter(배경 분규류). 객체가 주변 환경에 섞여(blend) 알아보기 힘들게 된다.
Intra-class variation(내부클래스의 다양성). 분류해야할 클래스는 범위가 큰 것들이 많다.
예를 들어 의자 의 경우, 매우 다양한 형태의 객체가 있다.

좋은 이미지 분류기는 각 클래스간의 감도를 유지하면서 동시에 이런 다양한 문제들에 대해 변함 없이 분류할 수 있는 성능을 유지해야 한다.

Pre-Processing (정형)

Data Cleansing (Null값 처리 방법)

-Delete(삭제)

. 가장 간단한 방법

. 누락된 데이터가 10%이상이 될 경우 고려가 필요함

. 누락된 데이터 자체가 의미가 있을 경우 고려가 필요함

예: 설문지에서 나이가 필수가 아닌 경우 누락된 나이는 개인정보를 민감하게 반응하는 사람을 대표할 수도 있음

-Filling(채우기)

.data를 채우는 방법

.범주형(Categorical)데이터는 Unknown 혹은 특정 값으로 변환

.연속형 데이터는 0 혹은 평균 등을 사용가능

Pre-Processing (정형)

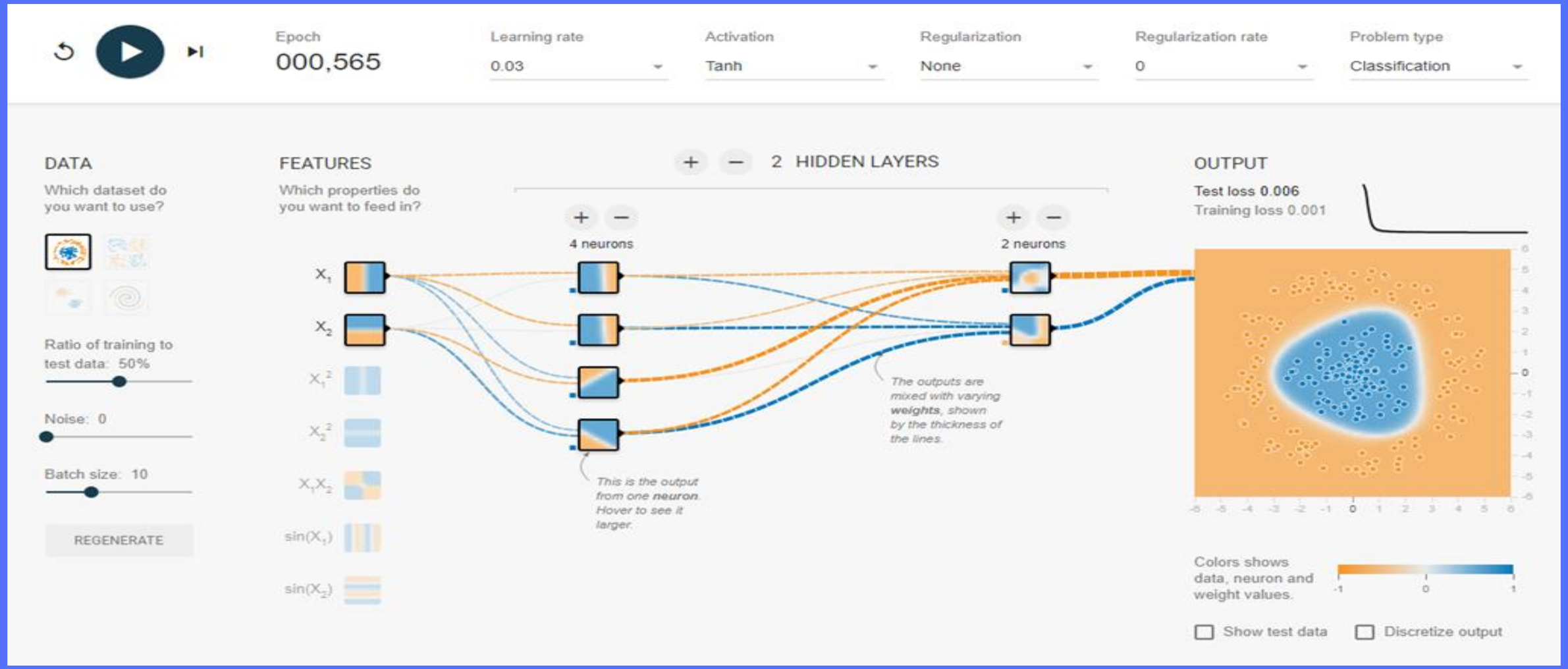
Feature Engineering

. Category Data 변환
. Continuous Data를 범주형으로 변환
. Feature를 더해서 새로운 Feature를 만듦
. Feature의 중요도를 찾아서 필요 없는 Feature 제거
. 연속형을 범주형 데이터로 변환

Feature Engineering을 하기 위한 필수 라이브러리

. PANDAS : 데이터를 보고 간단한 연산을 수행
. Scikit-Learn : 각종 머신러닝 모델
. Xgboost : Gradient Boosting 라이브러리
. Matplotlib : 데이터 시각화

Deep Learning 과정 Simulation(1)



Deep Learning 과정 Simulation(2)

wevi: word embedding visual inspector

Everything you need to know about this tool - Source code

Control Panel

Config:

```
{"hidden_size":5,"random_state":1,"learning_rate":0.2}
```

Training data (context|target):

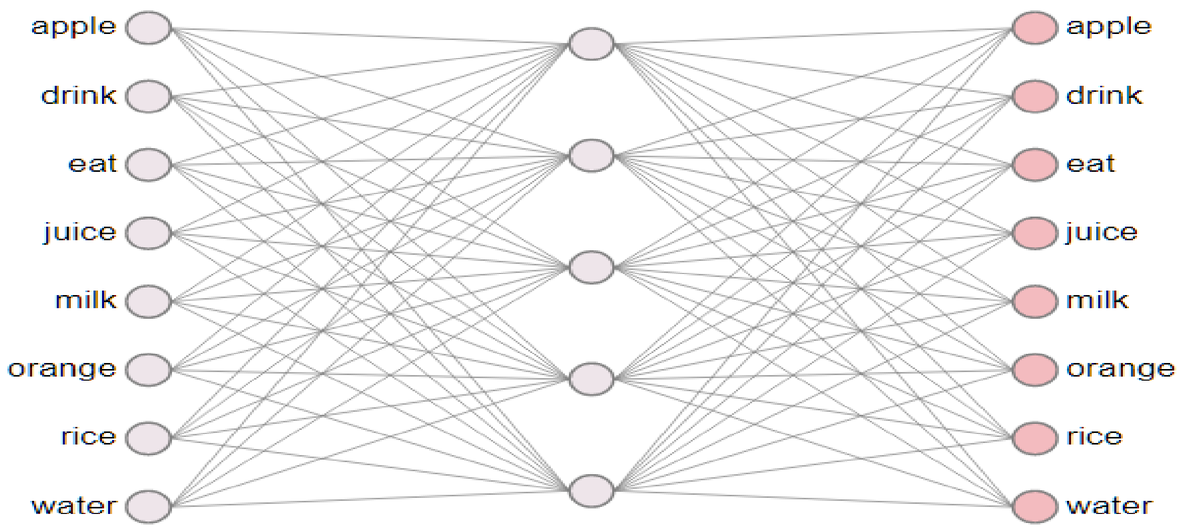
```
eat|apple,eat|orange,eat|rice,drink|juice,drink|milk,drink|water,orange|juice,apple|juice,rice|milk,milk|drink,water|drink,juice|drink
```

Presets: Fruit and juice

Update and Restart | Update Learning Rate

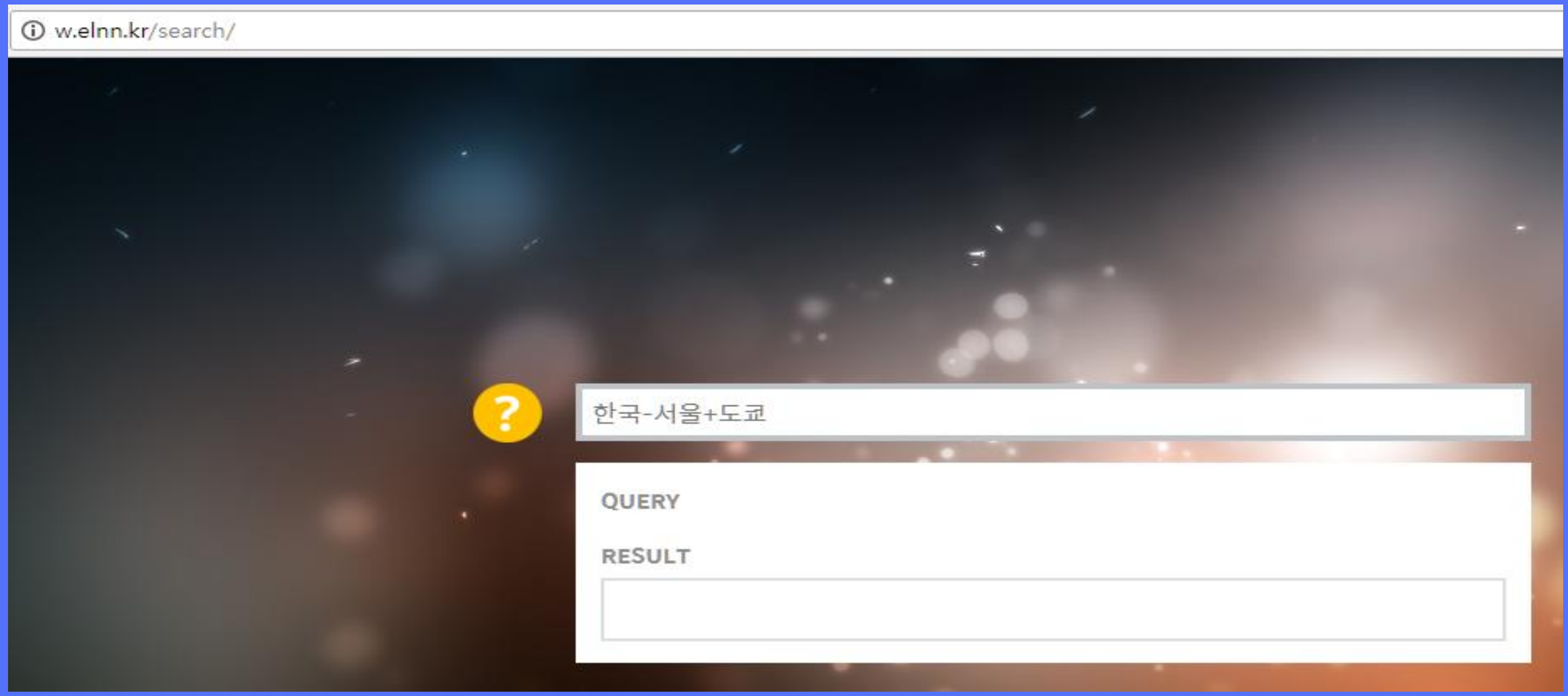
Next | 20 | 100 | 500 | PCA

Neurons

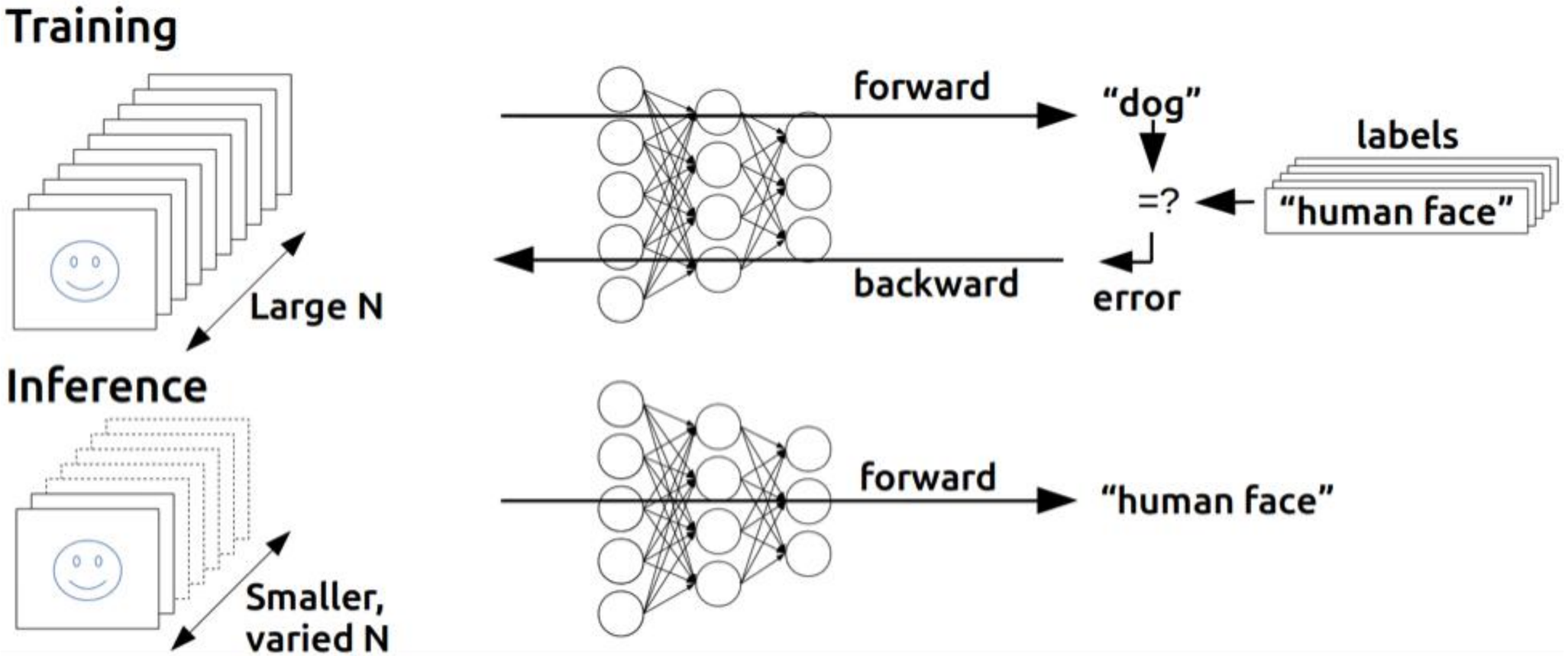


왕|왕국,여왕|왕국,왕|왕궁,여왕|왕궁,왕|왕가,여왕|왕가,왕|조지,여왕|메리,
남자|쌀,여자|쌀,남자|농부,여자|농부,남자|집,여자|집,남자|조지,여자|메리

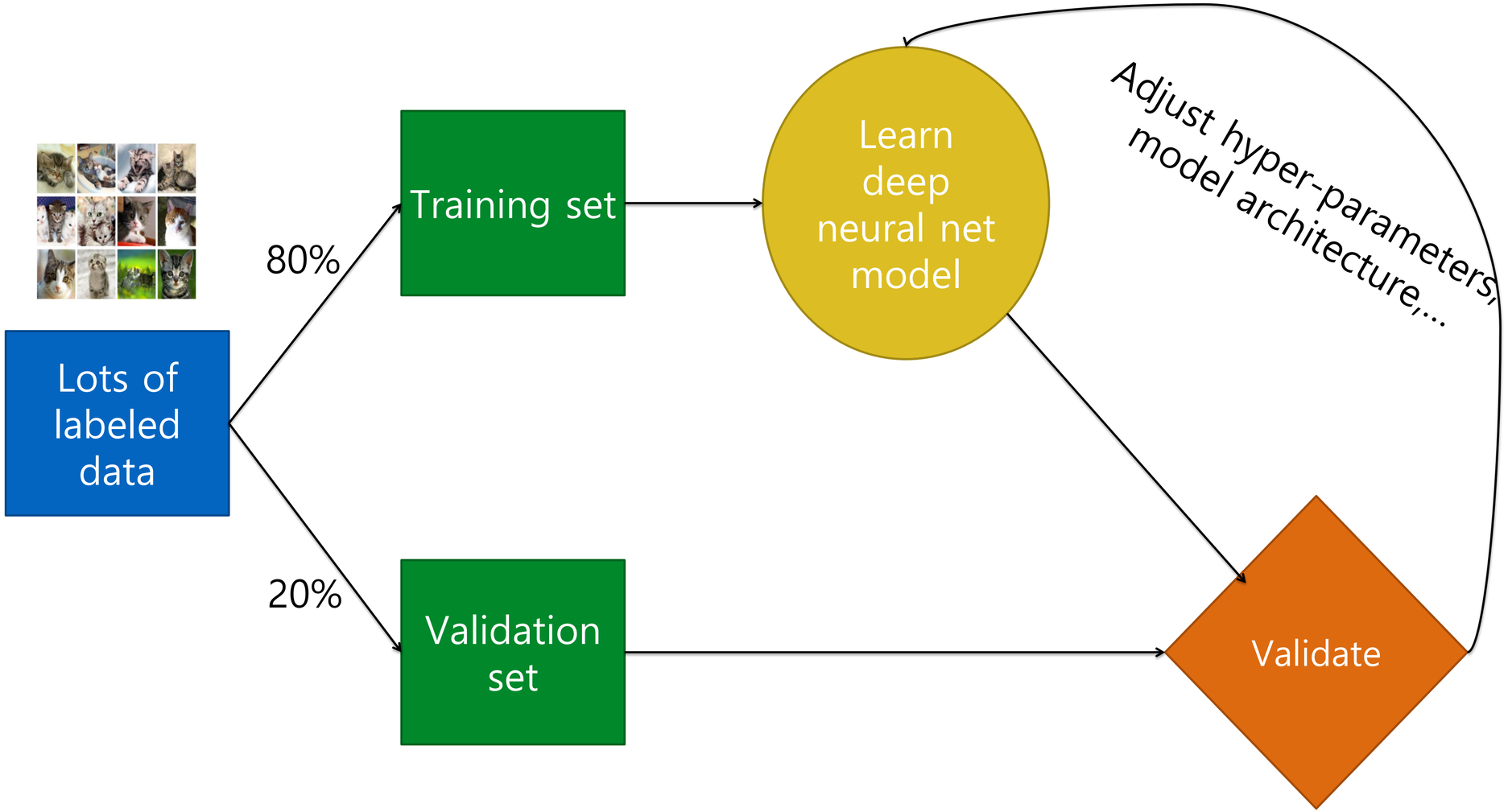
Deep Learning 과정 Simulation(3)



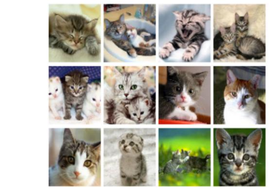
Model Training



Deep Learning Workflow



Transfer learning with deep features



$$H=\begin{bmatrix}20 & 77 & 21 & 31 & 4 & 5 & 1 & 13\\ 35 & 141 & 26 & 46 & 9 & 0 & 4 & 28\\ 76 & 301 & 49 & 91 & 23 & 2 & 7 & 43\\ 70 & 280 & 53 & 84 & 21 & 5 & 10 & 54\\ 71 & 258 & 44 & 76 & 19 & 4 & 4 & 32\\ 21 & 86 & 12 & 25 & 5 & 0 & 4 & 22\\ 92 & 350 & 53 & 102 & 19 & 1 & 5 & 40\\ 2 & 7 & 0 & 2 & 0 & 0 & 0 & 0\\ 56 & 221 & 31 & 62 & 6 & 2 & 3 & 23\\ 29 & 126 & 17 & 35 & 7 & 2 & 1 & 18\\ 55 & 192 & 39 & 52 & 15 & 2 & 5 & 32\\ 39 & 141 & 20 & 38 & 8 & 1 & 3 & 27\end{bmatrix}$$

Some labeled data

Extract features with neural net trained on different task

80%

20%

Training set

Validation set

Learn simple model

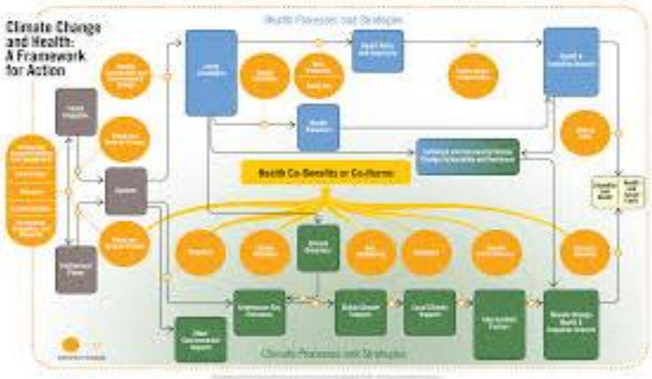
Validate

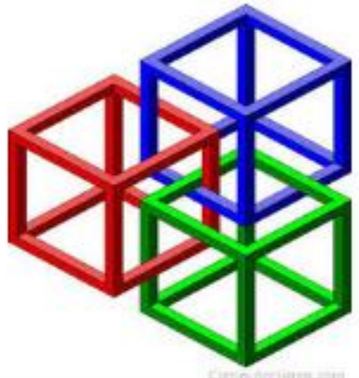
Deploy in production

Google

Framework

전체 이미지 동영상 도서 뉴스 더보기 설정 도구

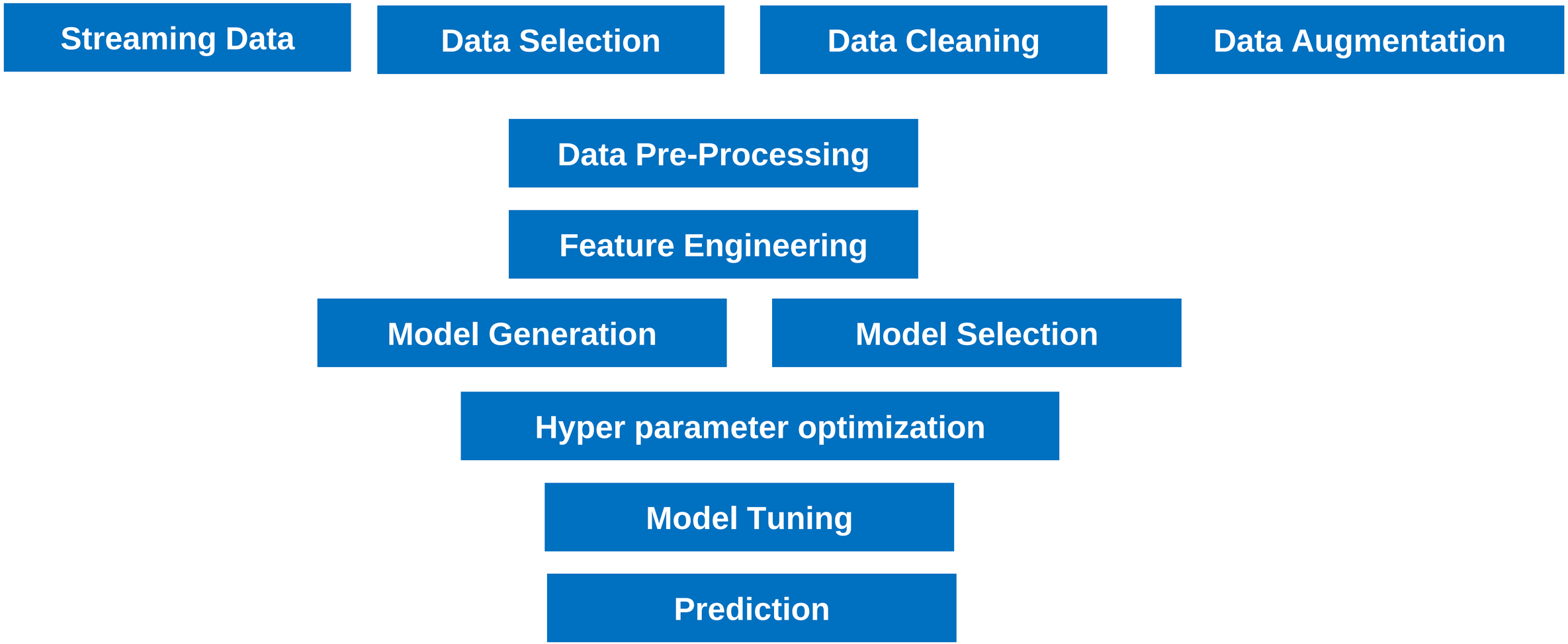




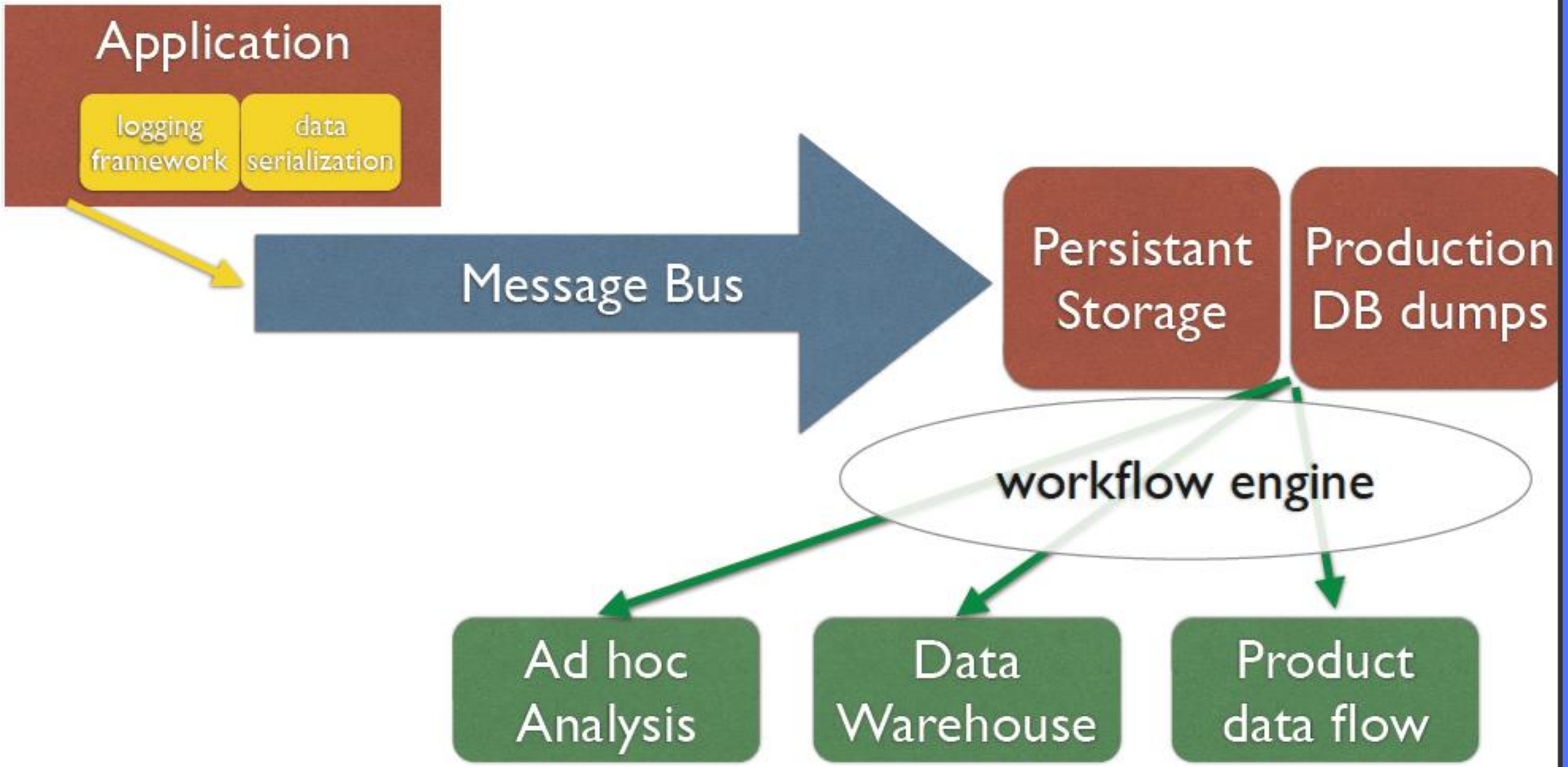
Framework

Data

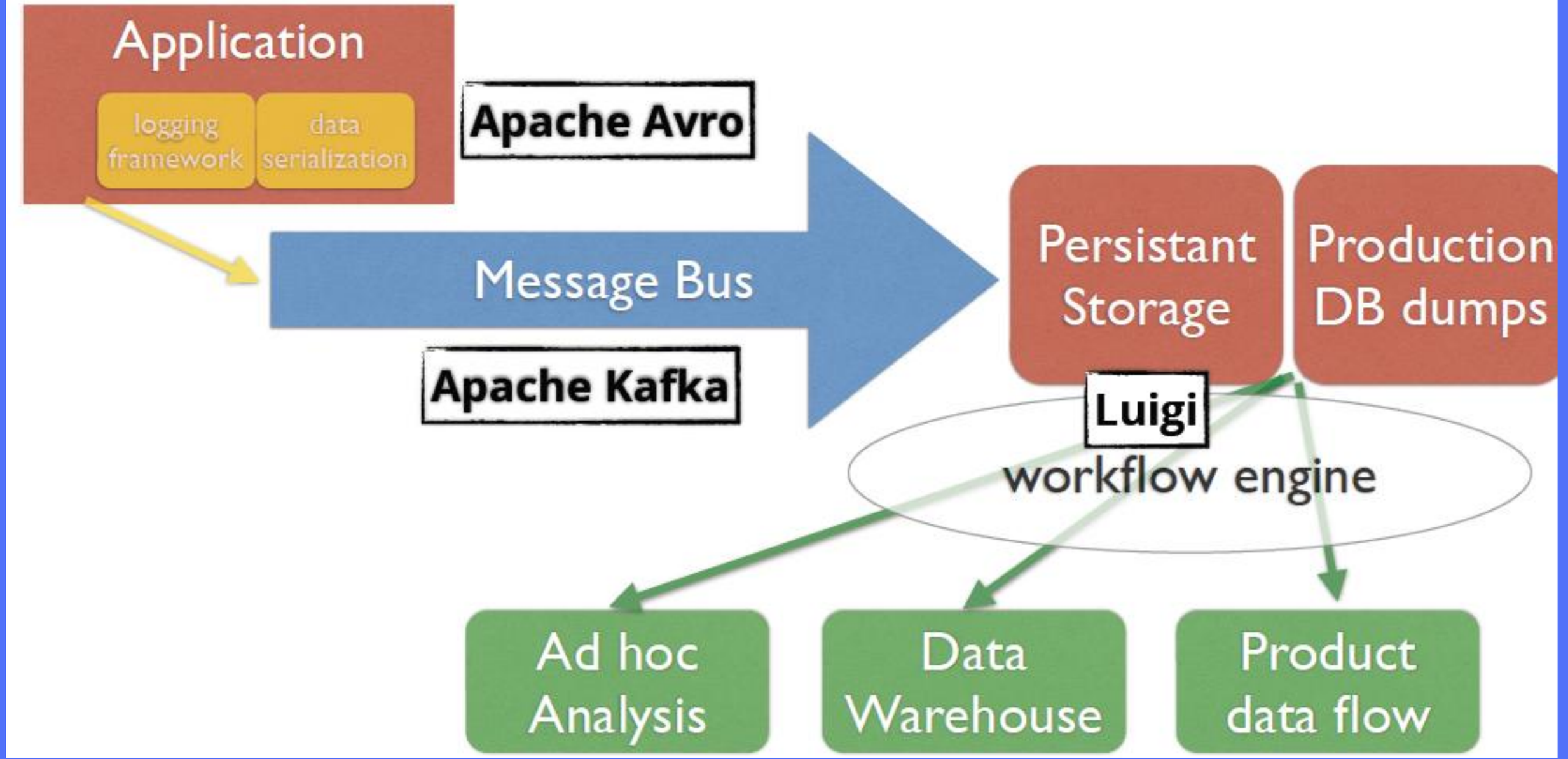
The Perspective of Data Flow



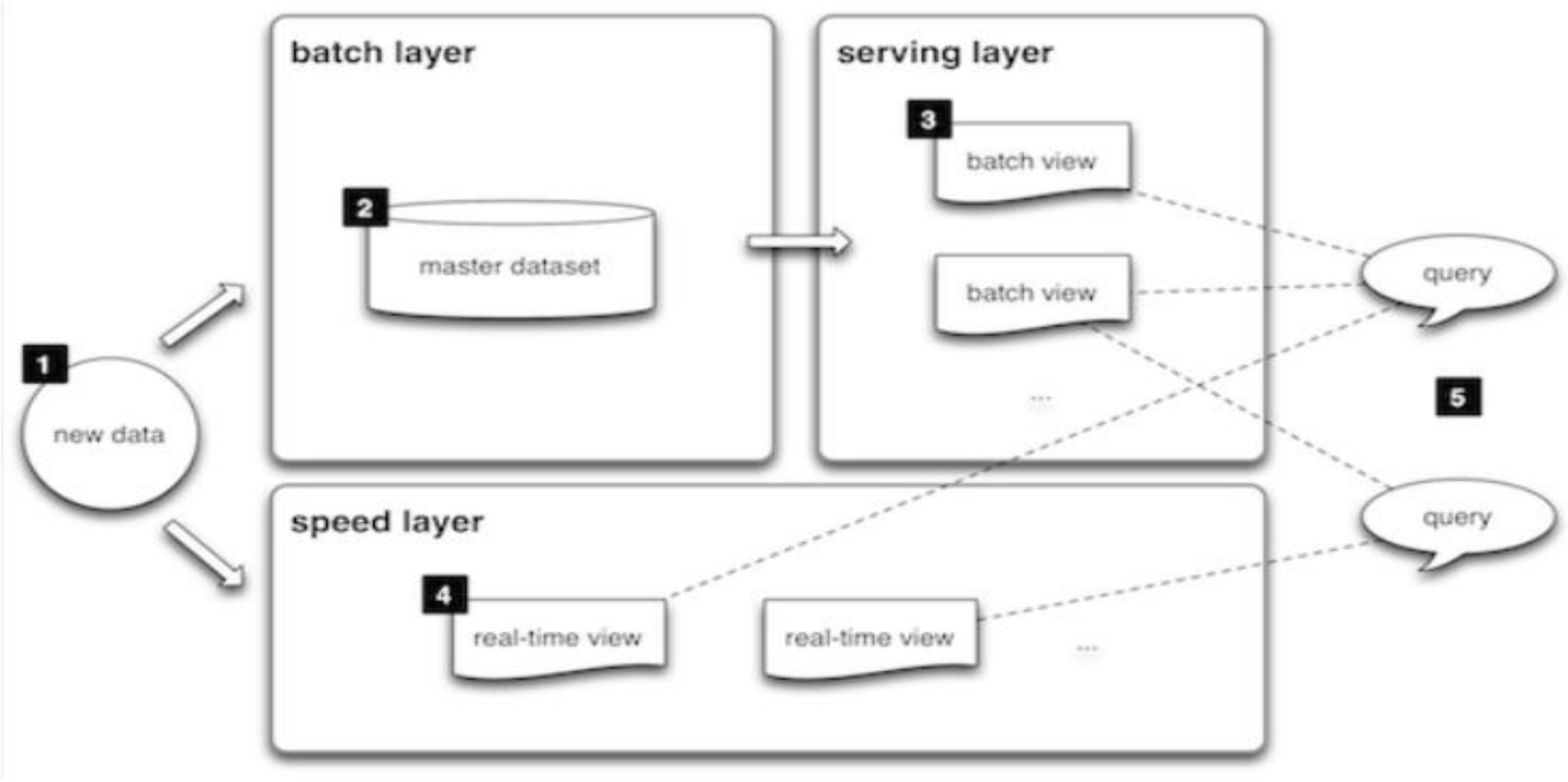
SYSTEM OVERVIEW



SYSTEM OVERVIEW (OPINIONATED)

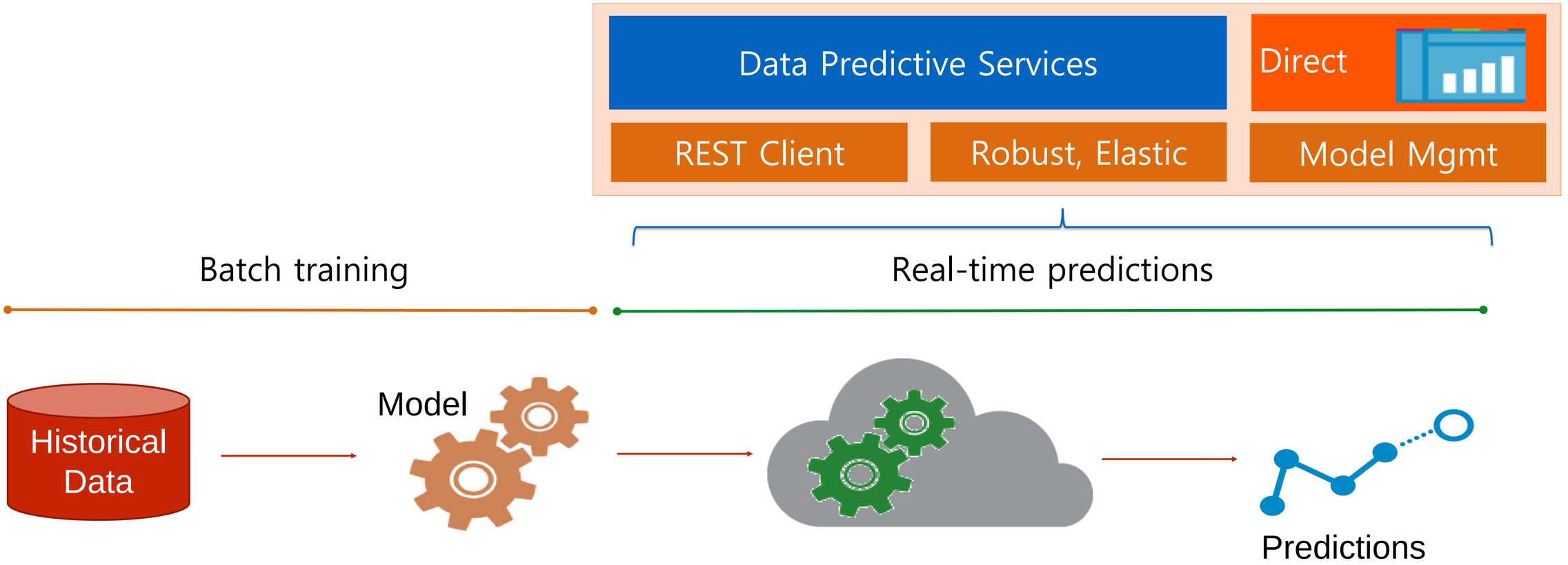


LAMBDA ARCHITECTURE

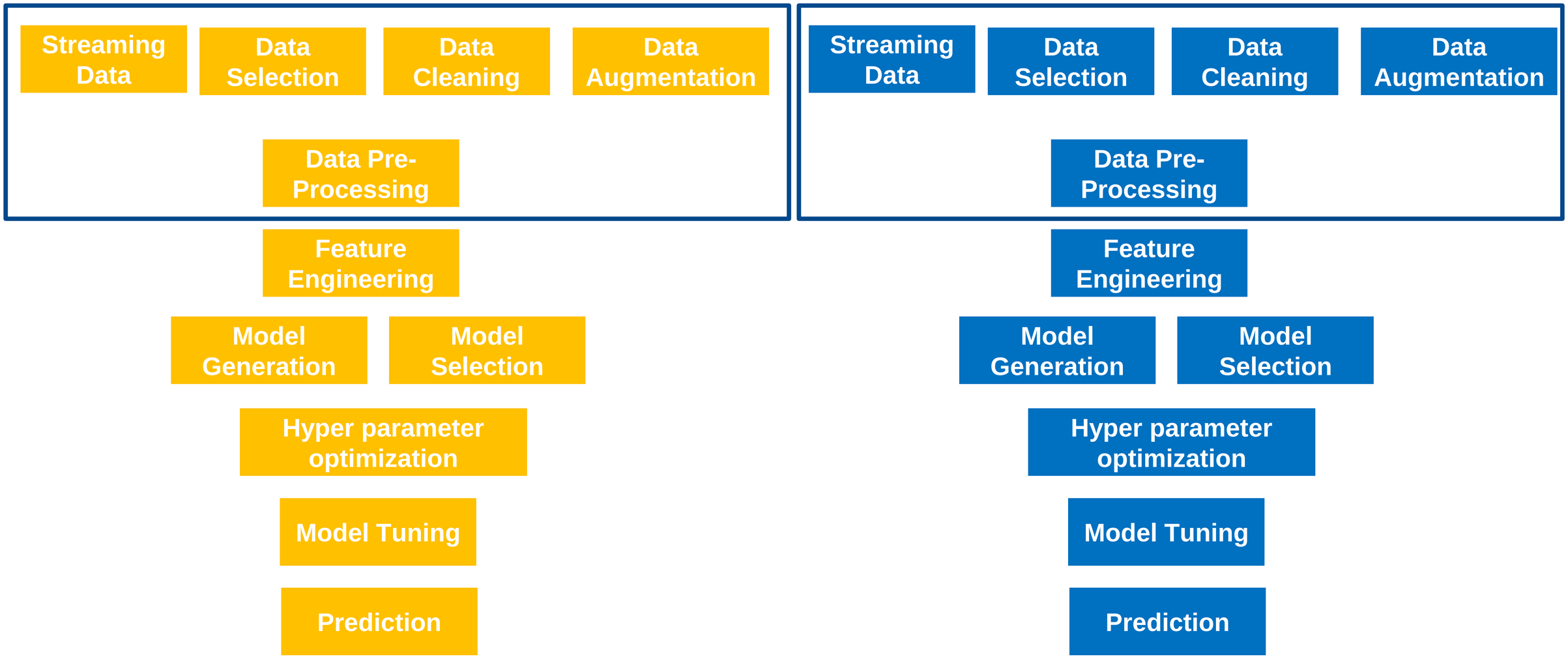


http://lambda-architecture.net/

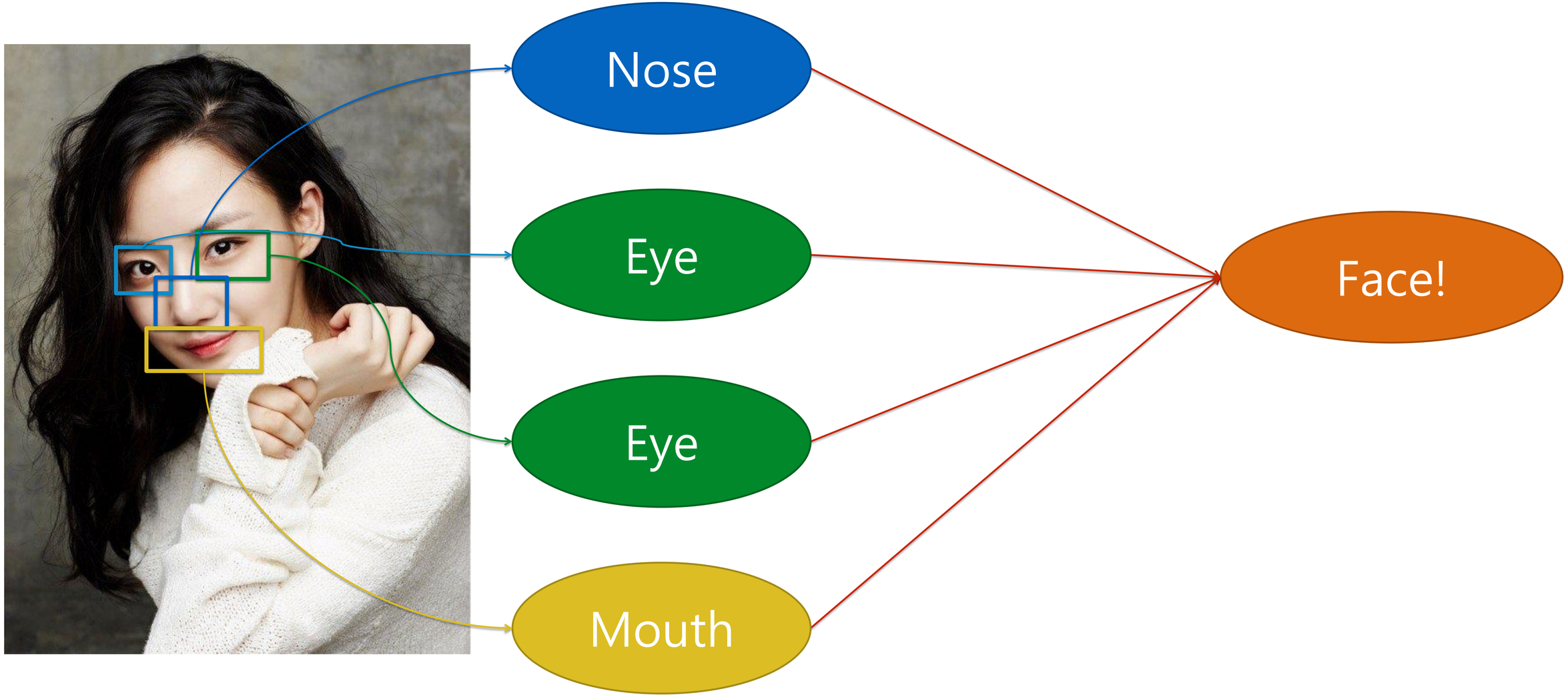
The Perspective of System



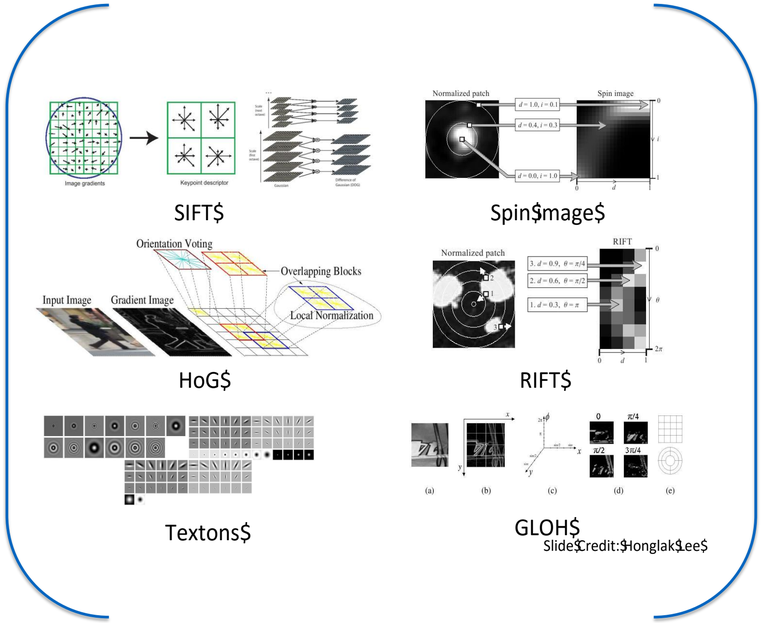
The Perspective of Data Scope



The Perspective of Data Scope



Extract features

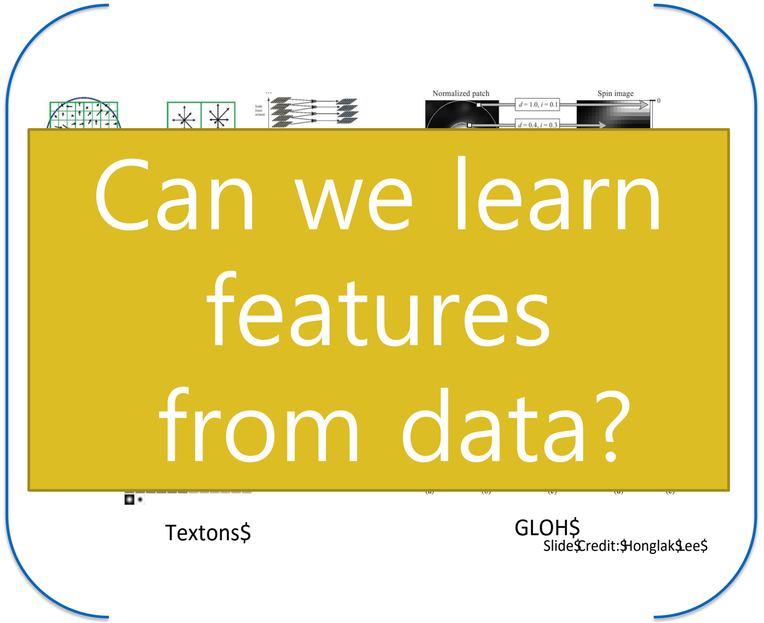


Slide$Credit:$HonglakLee

Use simple classifier e.g., logistic regression, SVMs

Face

Extract features

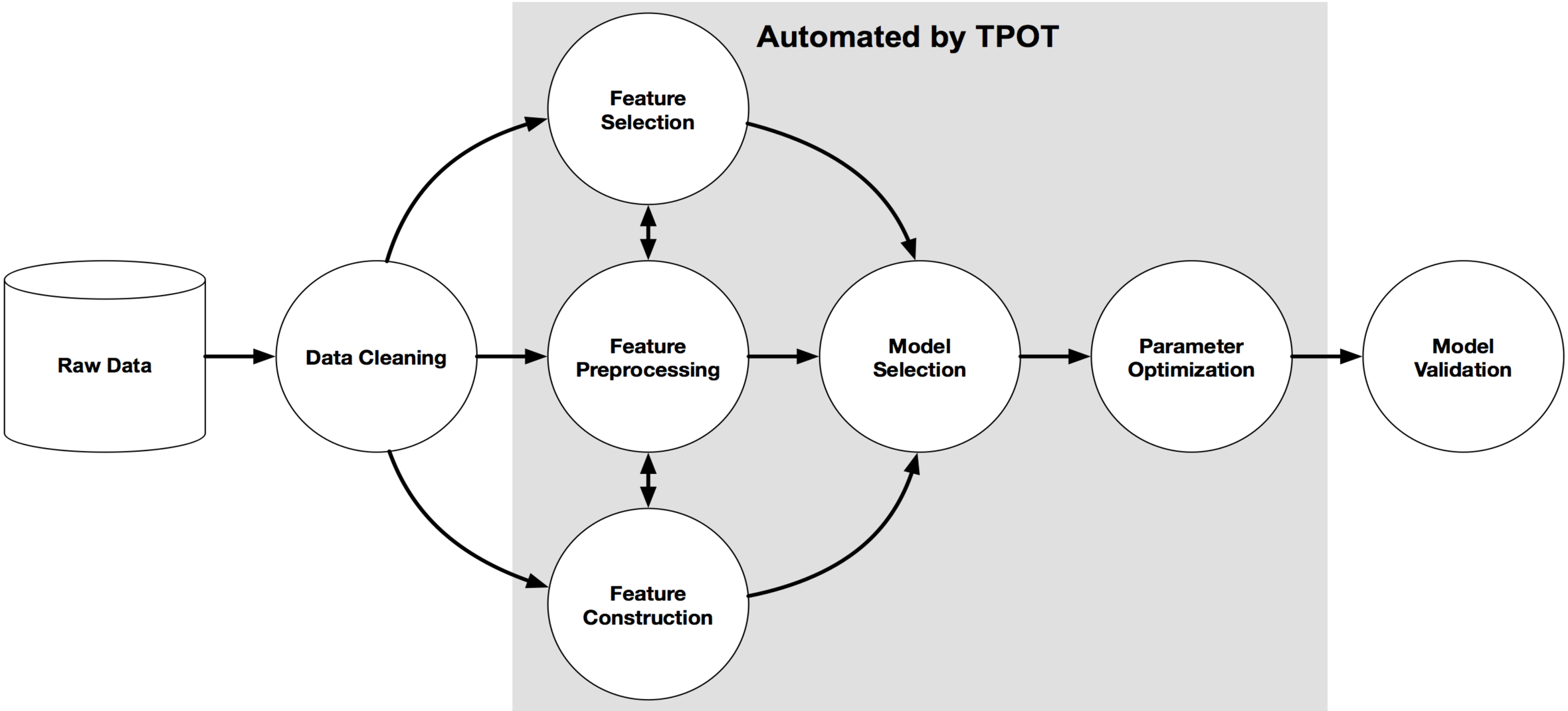


Use simple classifier e.g., logistic regression, SVMs

Face

Optimization

설계 자동화를 위한 고려사항(Automation Machine Learning)



AutoML

Hyper Paremter를 어떻게 탐색하는가에 대한 것이 아닌 끝까지 Train 을 진행하지 않고 사전에 결과를 예측하여 전체적인 탐색 속도를 향상시키는 것에 있다.

실제 적용 시에는 Hyper Parameter 의 Range 를 지정하고 탐색 Matirx 를 구성하고 실제로 Train 을 반복적으로 실행하는 환경은 구성을 해야 함.

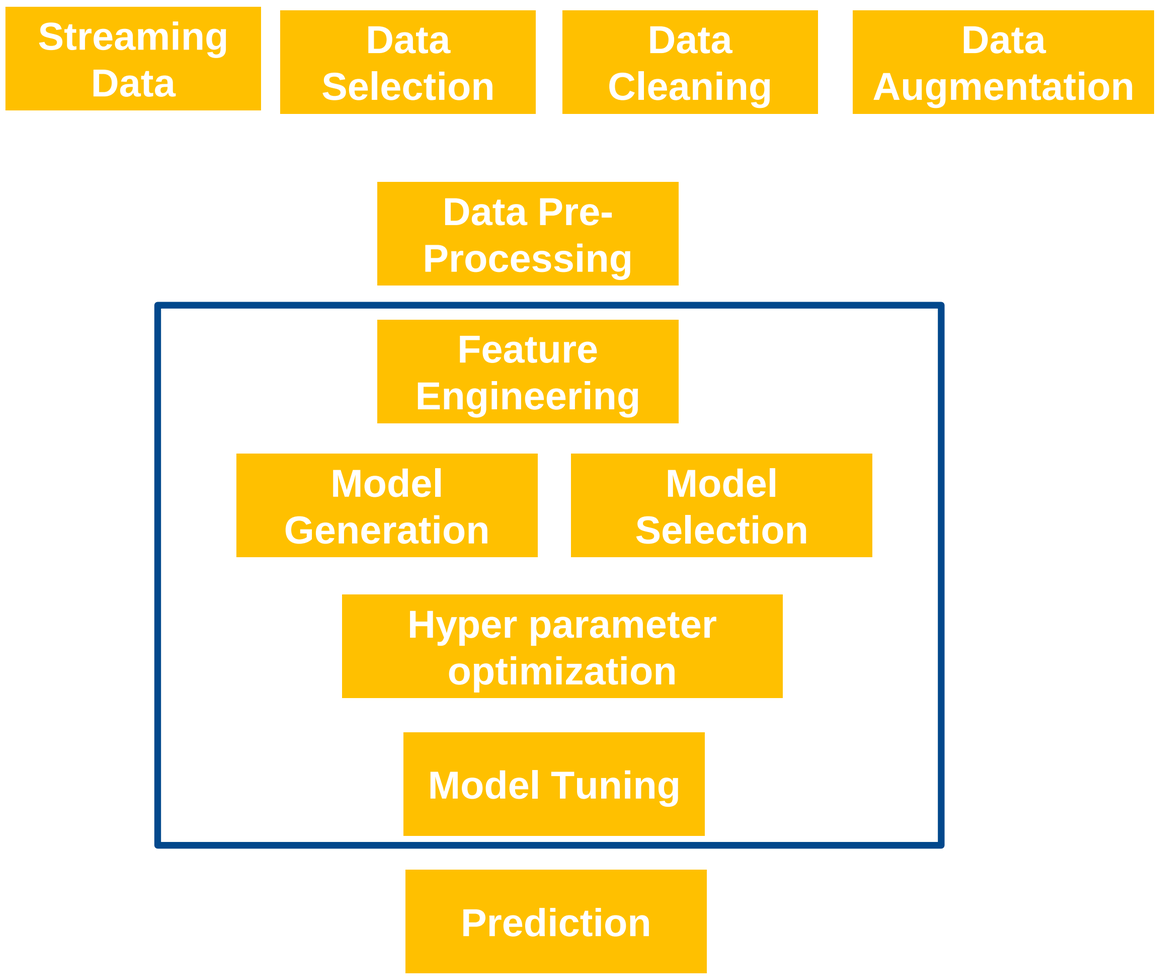
그 후에 특정 Iter 가지 진행한 Accuracy Learning Curve 정보를 가지고 위에 설명한 데이터를 예측하여 나온 Accuracy 가 지금까지 우리가 가지고 있는 최대 Accuracy 와 비교하여 그 이하라면 중단하는 형태로 구현을 하면된다.

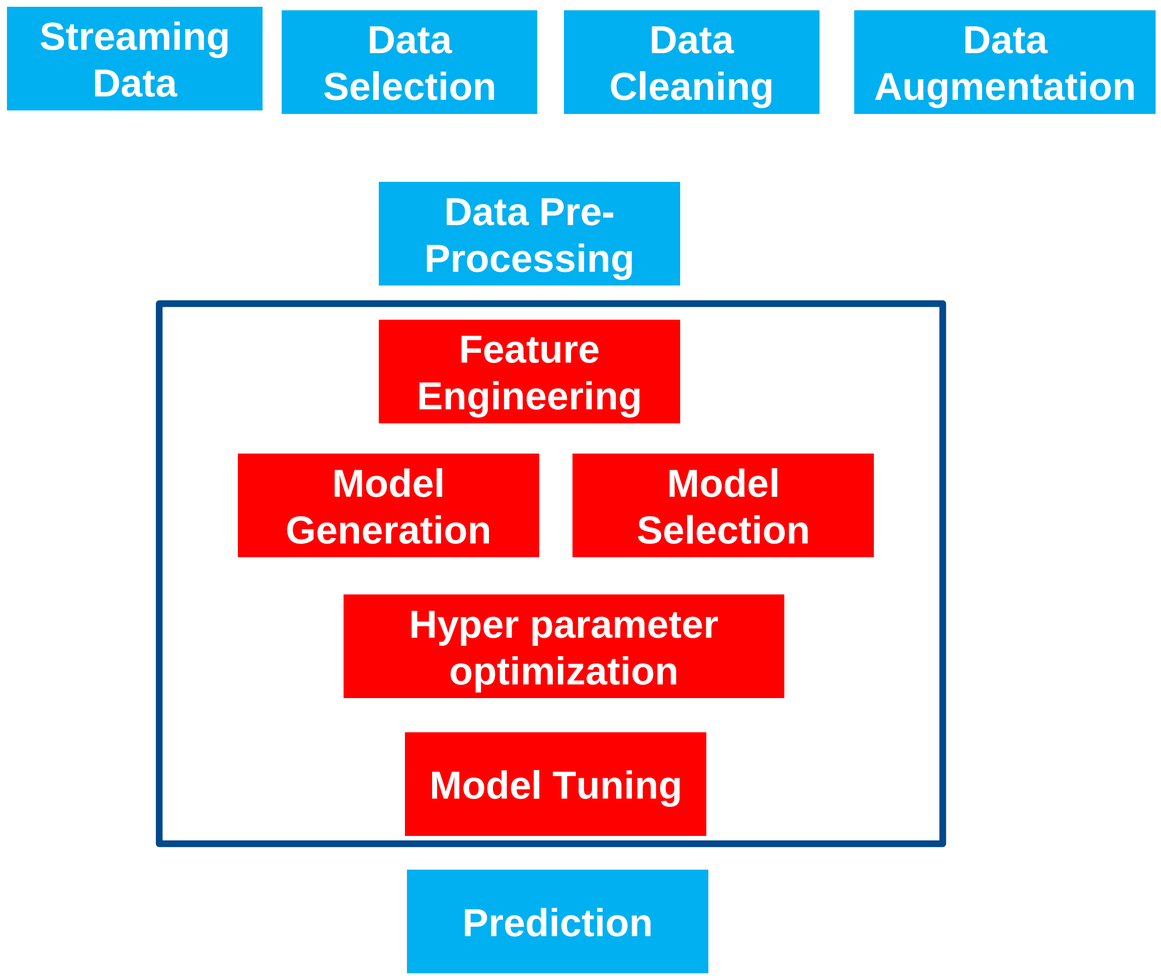
실제 간단한 DNN 에 적용하여 테스트해본 결과 위의 논문과 같이 모델을 구성하여 예측하는 과정이 CPU 기준으로 10분정도 소요되어 아주 작은 데이터에 대해서 테스트 시에는 크게 효용이 없을 것으로 보이며 Real BigData 에 대해서만 의미가 있다.

또, GPU 서버와 CPU 서버를 분리하여 예측 작업을 분리하여 처리하여 GPU 서버에서의 작업은 계속해서 진행하면 Over Head 를 줄 일 수 있을 것 같다.

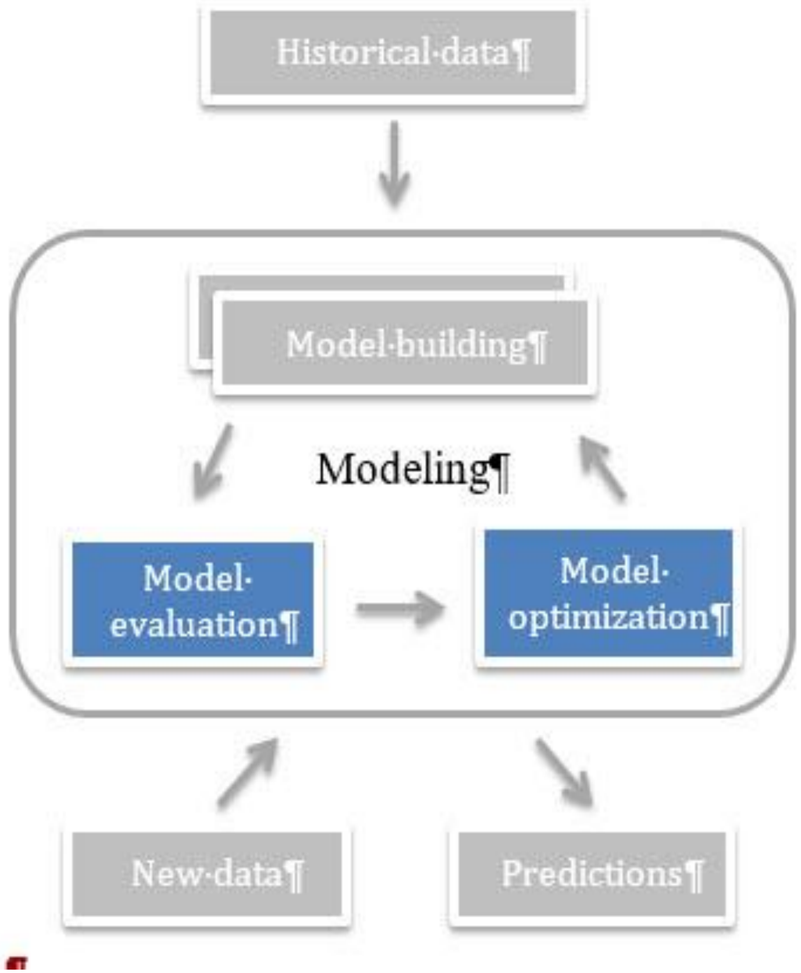
실제 예측 모델에서 예측한 Accuracy 와 실제로 훈련을 계속 했을때의 정확도가 얼마나 오차가 있는지는 Sample 데이터로 수행을 했을 때는 큰 차이가 없었다.

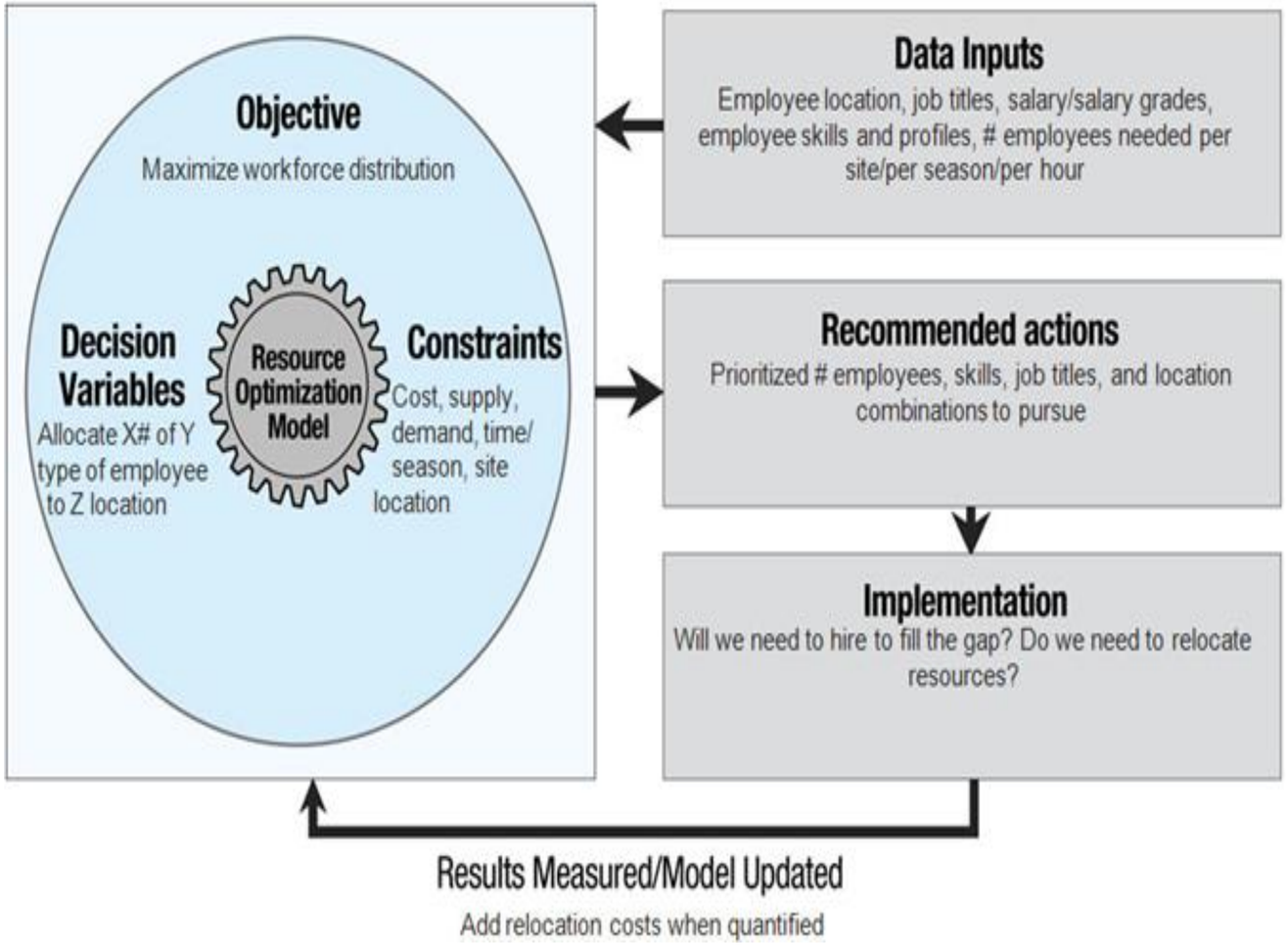
Traditional vs AutoML



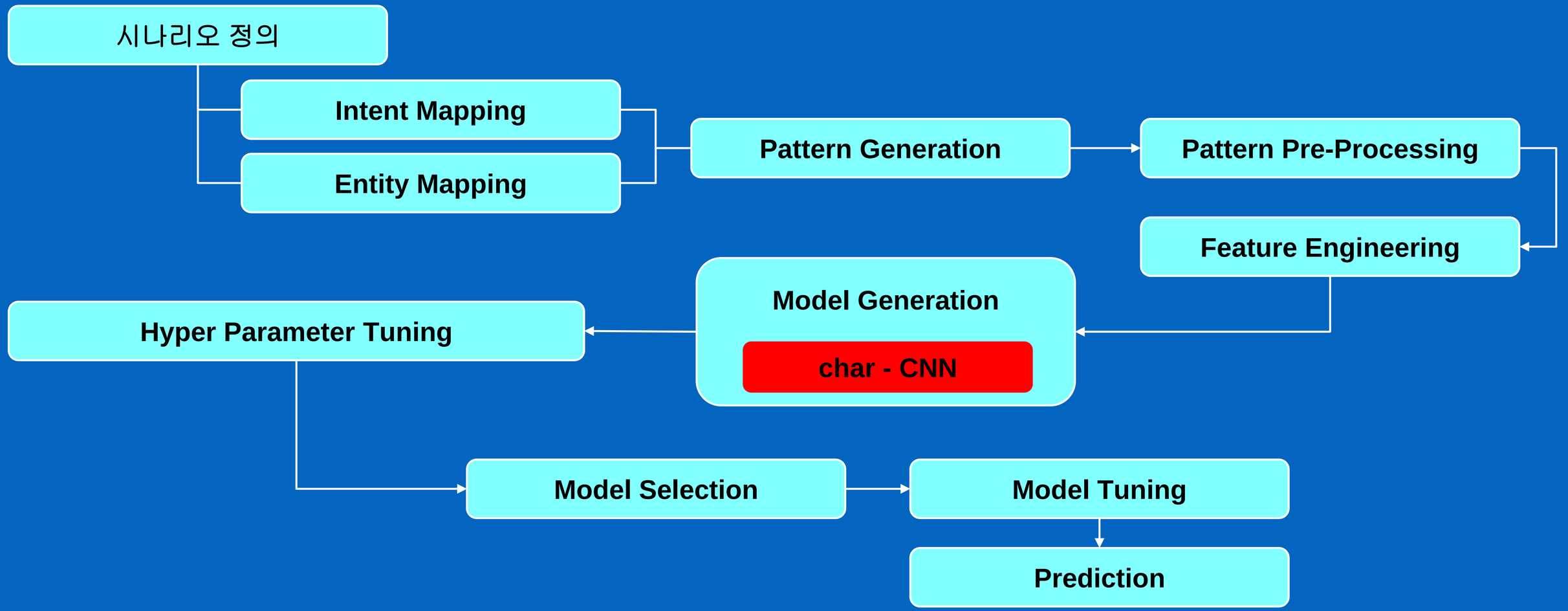


Model Optimization

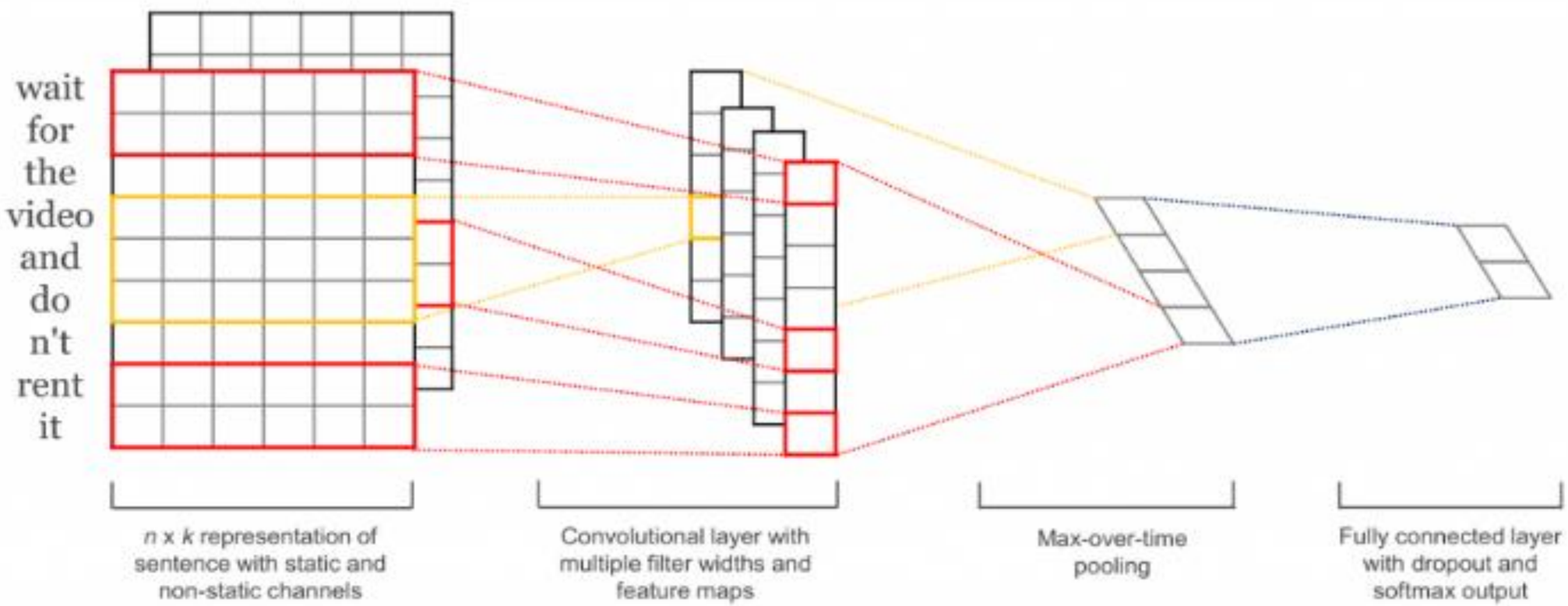




The Perspective of The Automation

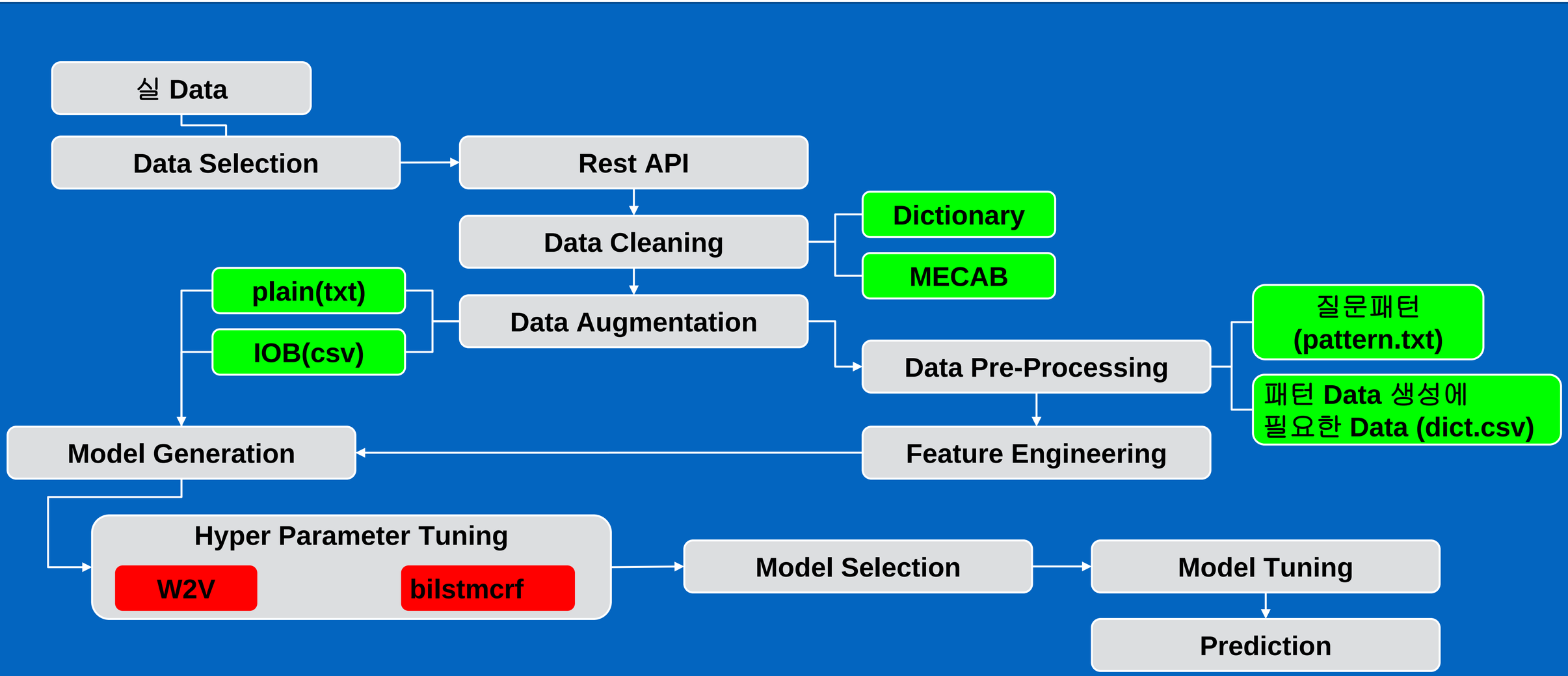


Char-CNN (Intent를 알아내는 법)



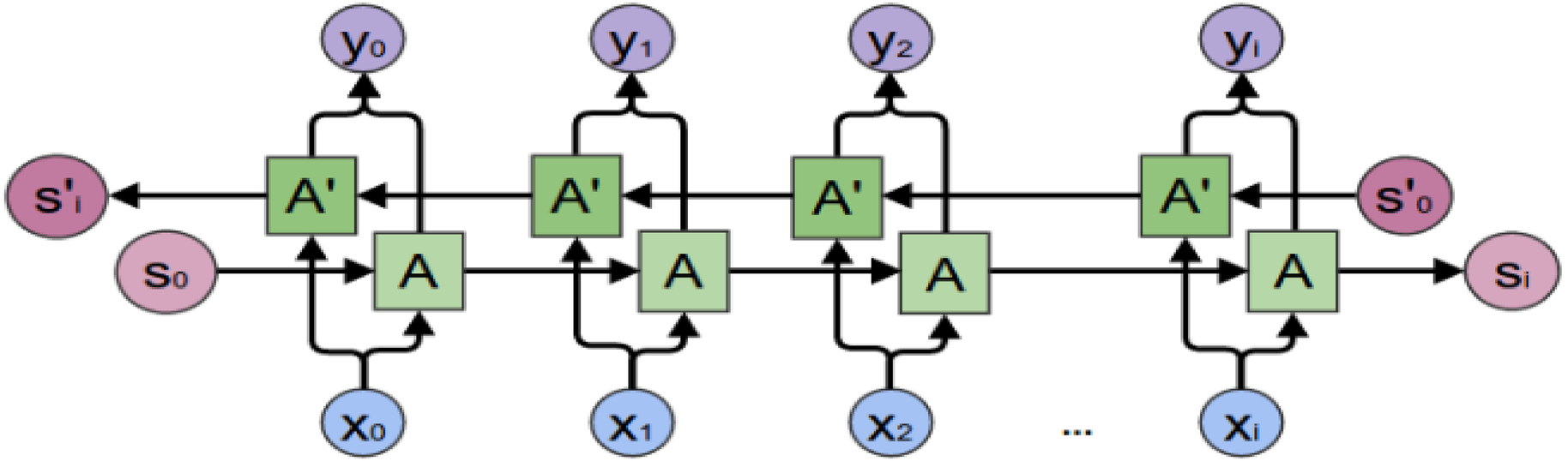
벡터변환 특징 추출 추상화 값 추출

The Perspective of The Automation

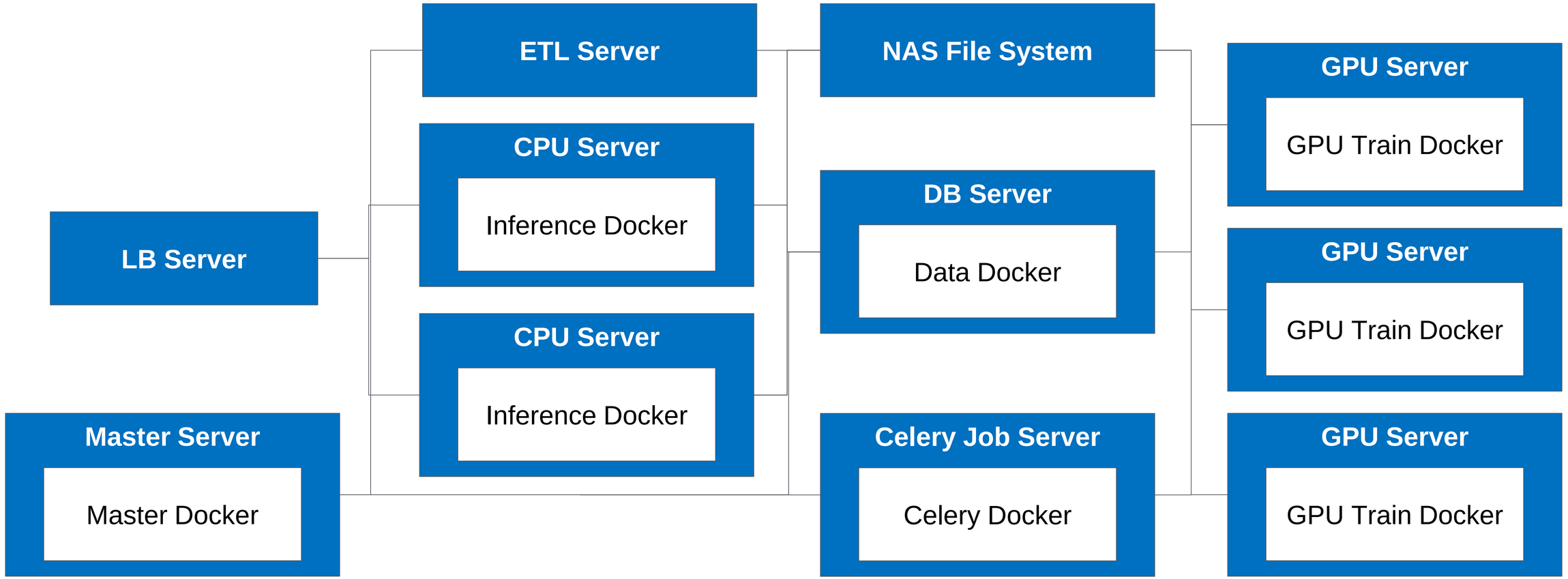


Named Entity Recognition 알아내기

Bidirectional LSTM(양방향 Layer)
- RNN 기반의 모델
- 특정 위치에 있는 단어의 태깅 활용
- IOB Label을 활용한 Supervised Learning Model 생성
  (I-이어지는 어휘, O-어휘 아님, 공백(OUT), B – 시작되는 어휘



Architecture Overview



RPA(Robotic Process Automation)

Cognitive Continuum

Machine Learning

Predictive Analytics

Business Intelligence

dynamic

static

3. Active Learning Points Prioritized

4. Machine Self Adjusted

5. Machine extends support capability

Routine Production Services

Engagement Capability

System Orchestration

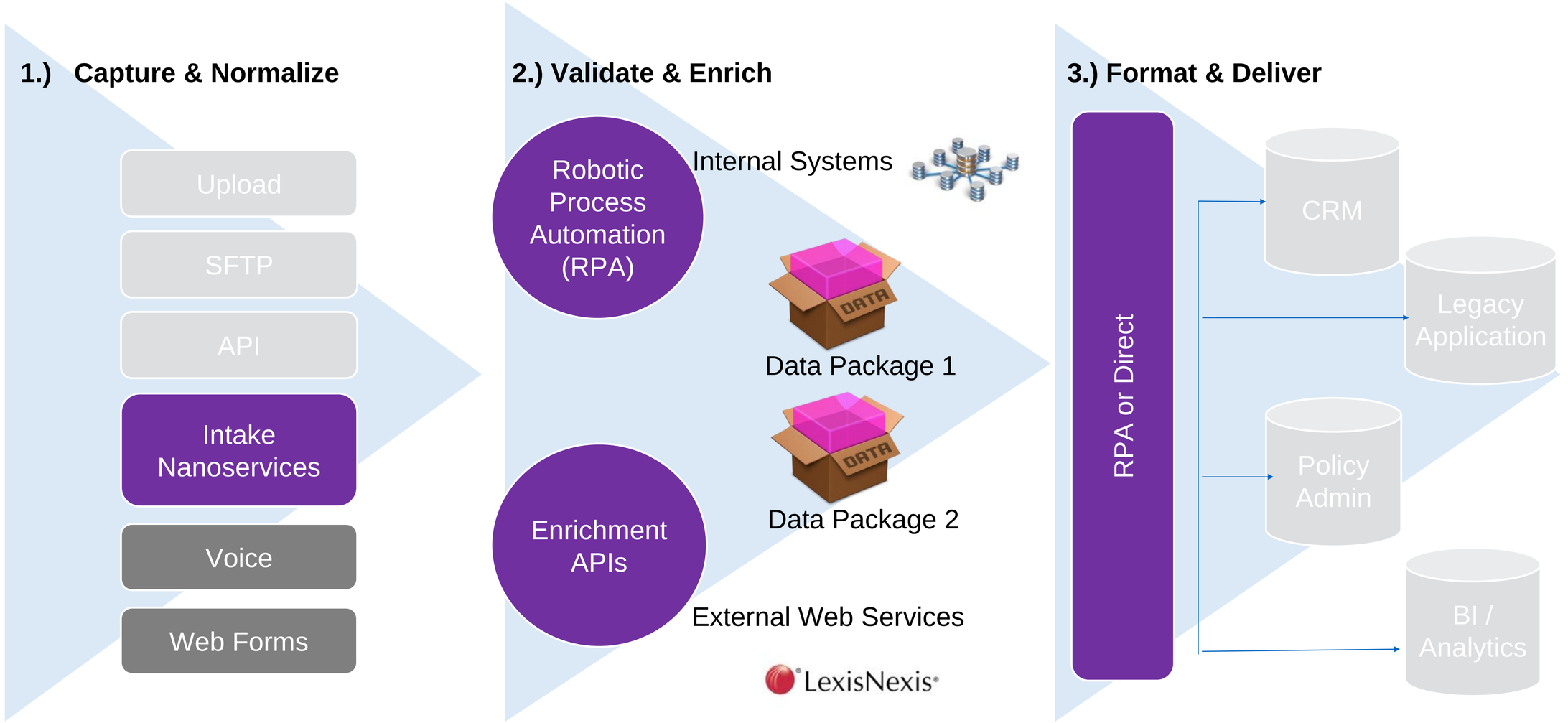
Natural Language Processing

Robotics Process Automation

1. Customer Hits Portal Link For Human Assistance

2. Human addresses problem – text captured and converted for machine ingestion

Data as a Service (DaaS)



Self-Serve Monitoring & Compliance Dashboards

Processors

Human Agent

Leo

UiPath
Robotic Process Automation

Workflow

Input
- eMail
- Excel
- Manual

Data Processing
- Validation
- Business Rules

Output
- Status
- Output values

Reporting
- Notifications
- Dashboards
- Reports

Engines

Unstructured Input Interpretation

salesforce

SAP

hp

SmartBuy
Experience a simpler way to shop.

Credentials Vault

Mission Control Center


Thank
you