DATUMO

The ultimate all-in-one data platform

www.datumo.com

All-in-one data platform

SAMSUNG ELECTRONICS
SK telecom
NAVER
LG U+
LG CNS

kt
SAMSUNG 삼성생명
SAMSUNG SDS
kakao enterprise
CJ 올리브네트웍스

ETRI 한국전자통신연구원
Upstage
Smilegate Entertainment
NSR 국가보안기술연구소
롯데정보통신

VOyAGERx
경동나비엔
서울아산병원 Asan Medical Center
KAIST
POSTECH

Est.
2018.11

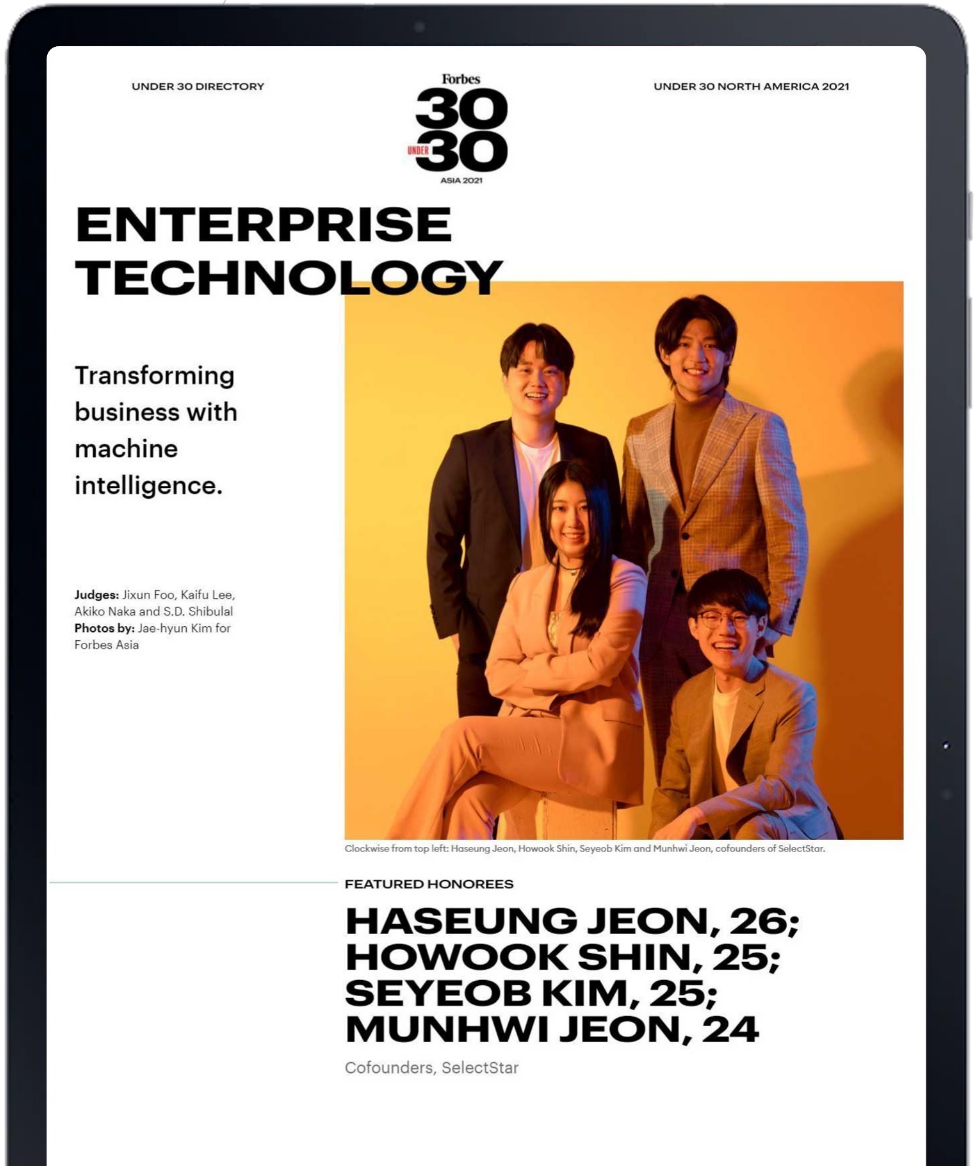
Clients
227+

Investment received
11M USD

Crowd-workers
240K+

Core staffs
100

All–in–one data platform



UNDER 30 DIRECTORY

Forbes 30 UNDER 30 ASIA 2021

UNDER 30 NORTH AMERICA 2021

ENTERPRISE TECHNOLOGY

Transforming business with machine intelligence.

Judges: Jixun Foo, Kaifu Lee, Akiko Naka and S.D. Shibulal
Photos by: Jae-hyun Kim for Forbes Asia

Clockwise from top left: Haseung Jeon, Howook Shin, Seyeob Kim and Munhwi Jeon, cofounders of SelectStar.

FEATURED HONOREES

HASEUNG JEON, 26;
HOWOOK SHIN, 25;
SEYEOB KIM, 25;
MUNHWI JEON, 24

Cofounders, SelectStar

Nominated as 2021 Forbes 30 Under 30 Asia, Enterprise Technology

From Model-centric to Data-centric AI

2021.03

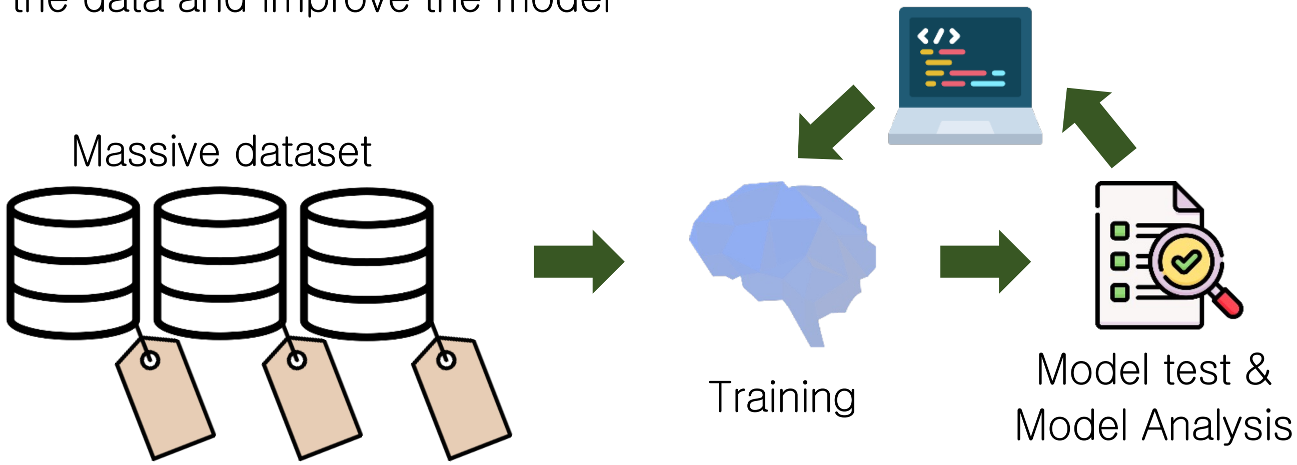
Andrew Ng (World-renowned AI scholar)

It is changing from "how to make AI" to "how to make data".

From Model–centric to Data–centric AI

Model–centric AI

Fix the data and improve the model



Model Fix

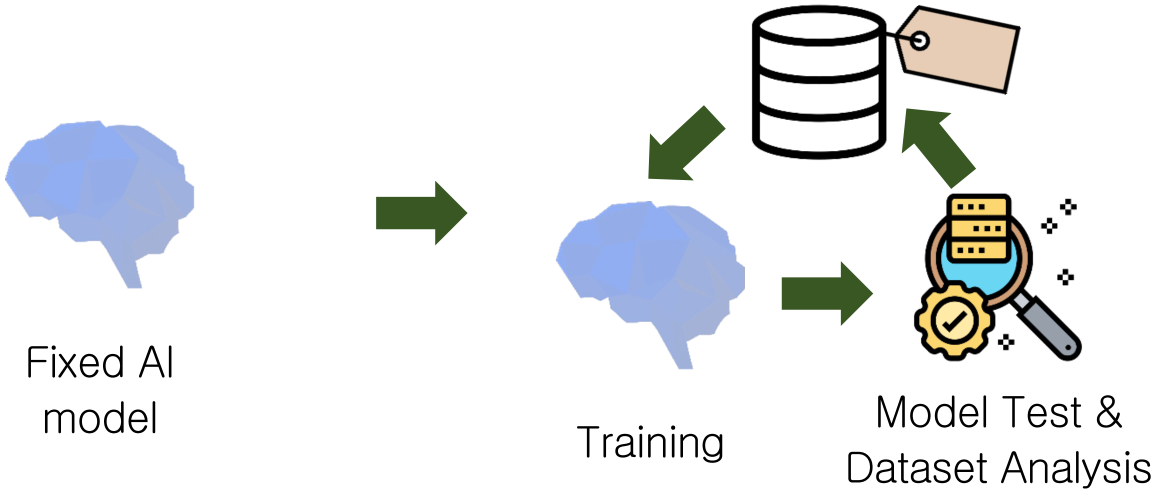
Massive dataset

Training

Model test &
Model Analysis

Data–centric AI

Fix the model and improve the dataset



Fix or add data

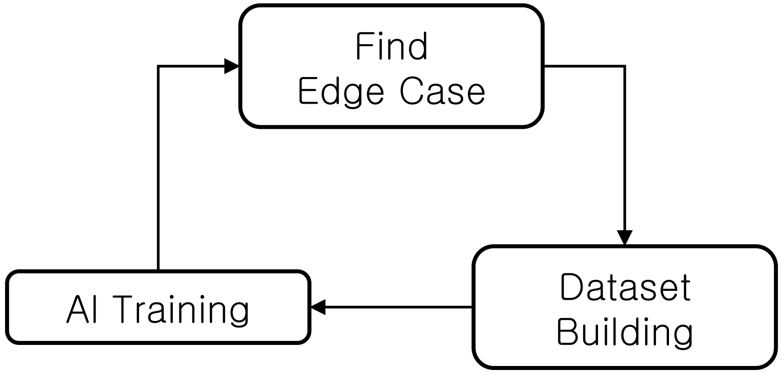
Fixed AI
model

Training

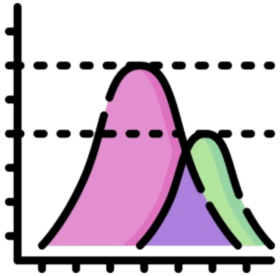
Model Test &
Dataset Analysis

Why Data-centric AI?

If AI performance issues stem from data, then the solution must come from the data itself.



Repeating edge case handling reduces AI performance gap between development and service.

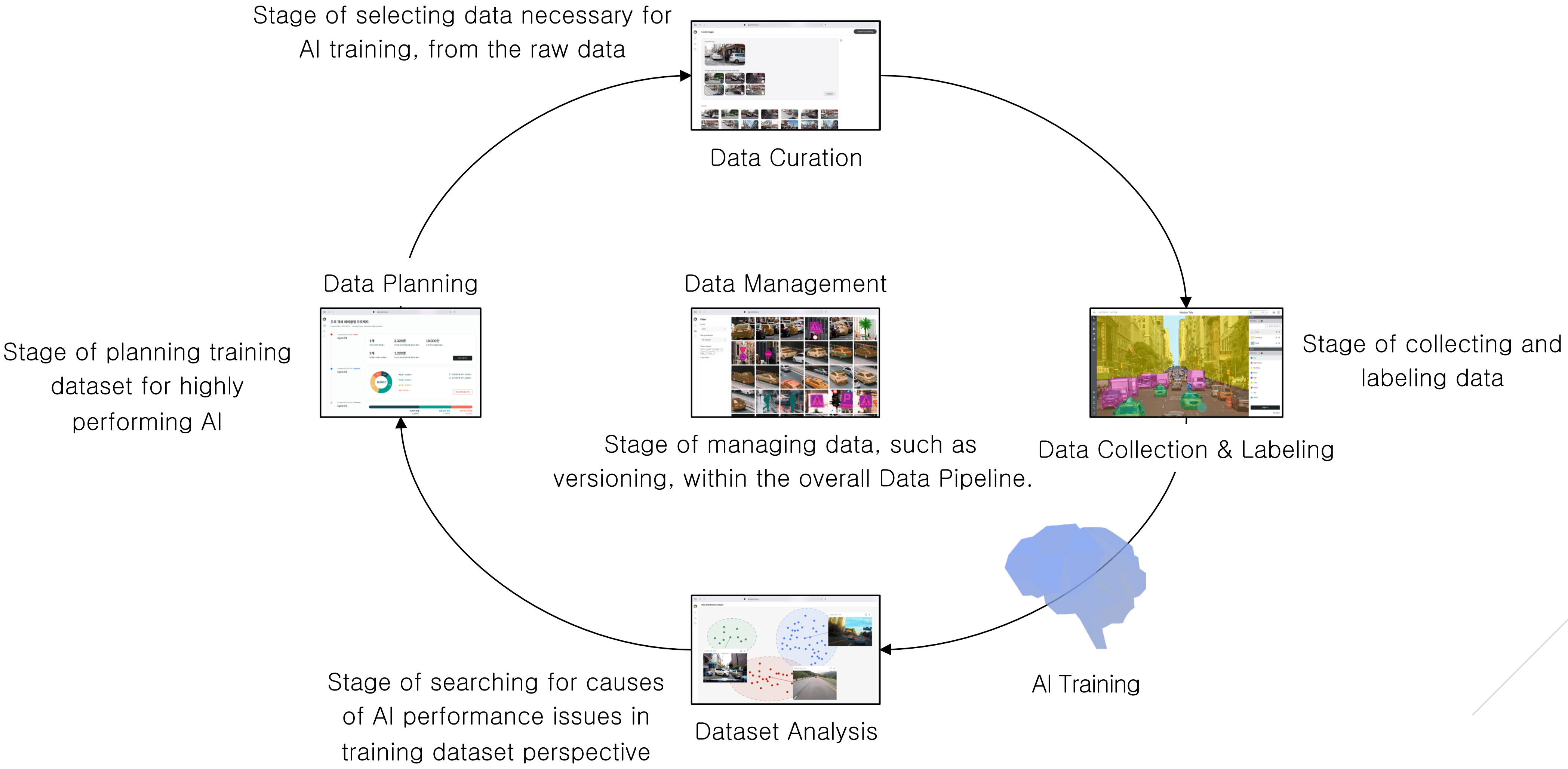


Data shifts cause AI performance degradation in service operation.

From Model-centric to Data-centric AI

Data-centric AI is essential for
both creating a custom LLM and utilizing fine-tuning.

AI Data Flywheel



Data Curation

Stage of selecting data necessary for AI training, from the raw data

Data Collection & Labeling

Stage of collecting and labeling data

AI Training

Dataset Analysis

Stage of searching for causes of AI performance issues in training dataset perspective

Data Planning

Stage of planning training dataset for highly performing AI

Data Management

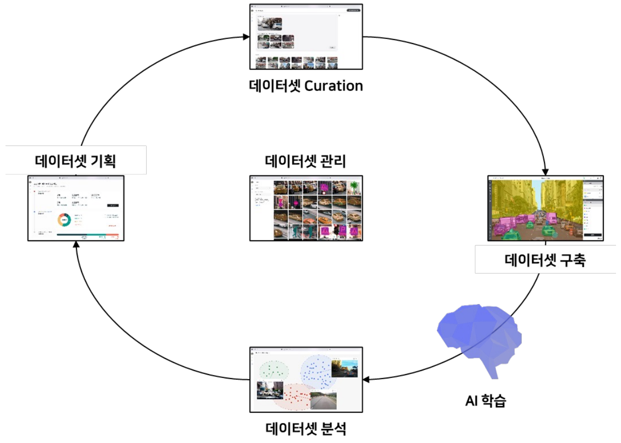
Stage of managing data, such as versioning, within the overall Data Pipeline.

From Model-centric to Data-centric AI

As LLM and other AI API offerings continue to expand...

Introducing
AI

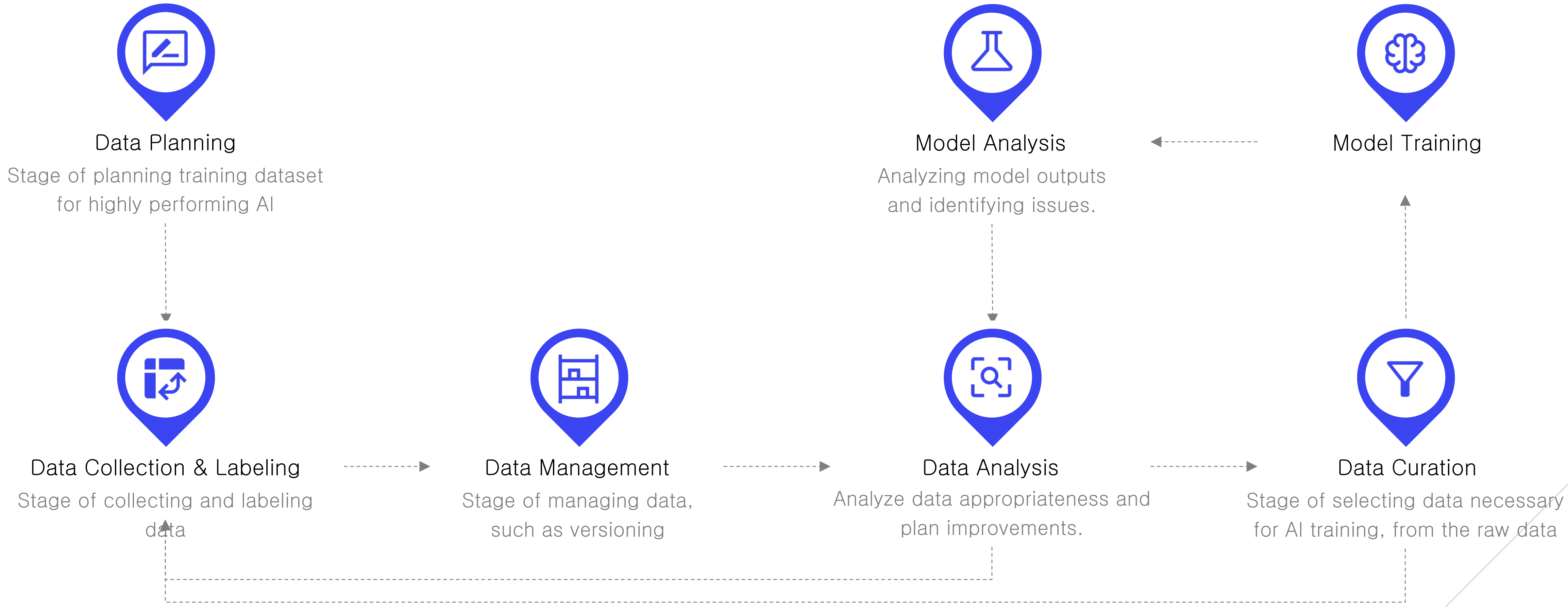
=



Introducing
Data Flywheel

DATUMO

AI Data Flywheel



Data Planning

Stage of planning training dataset for highly performing AI

Data Collection & Labeling

Stage of collecting and labeling data

Data Management

Stage of managing data, such as versioning

Data Analysis

Analyze data appropriateness and plan improvements.

Data Curation

Stage of selecting data necessary for AI training, from the raw data

Model Training

Model Analysis

Analyzing model outputs and identifying issues.

What is “Quality Data”?

Accuracy

Accurate data to fulfill the purpose

Balance

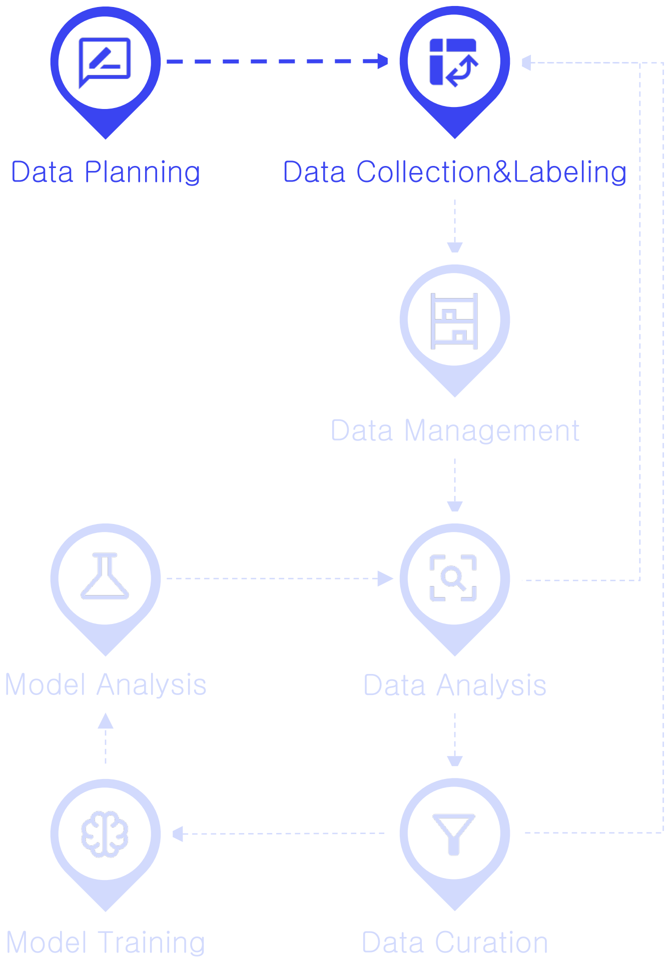
Unbiased dataset

Consistency

Consistent labeling results

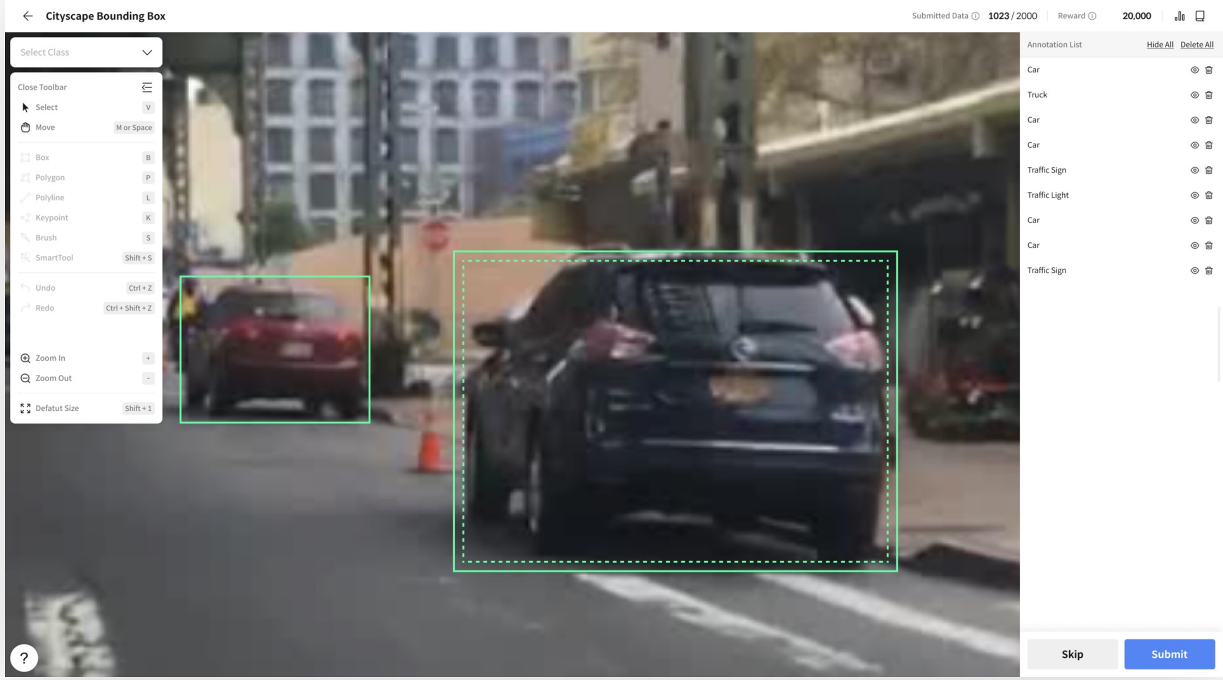
Coverage

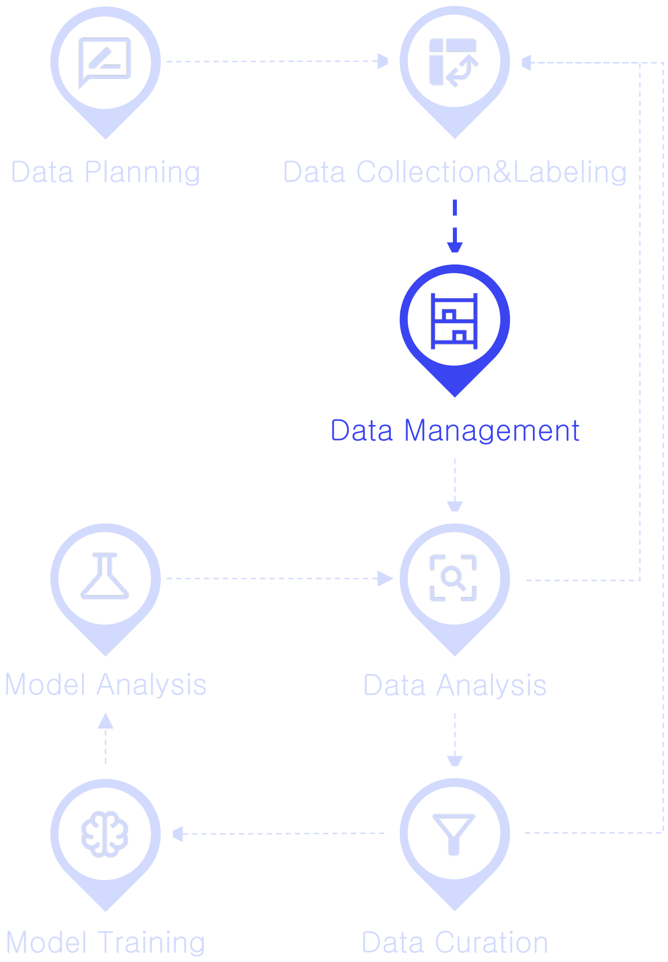
Enough case studies based on a variety of data



Data Collection & Labeling

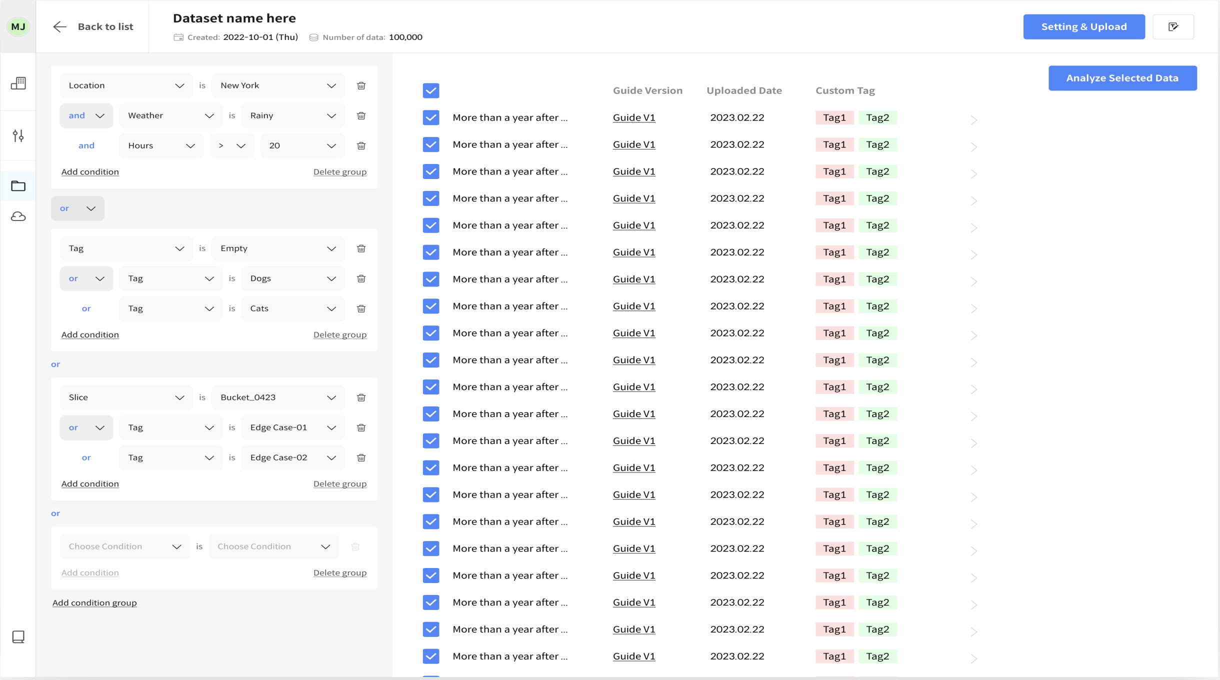
Plan collection and labeling carefully to ensure accuracy, balance, consistency and coverage.





Data Management

Versioning is important to keep up with the ever-changing guidelines, sources, and basic metadata of data.



← Back to list

Dataset name here

Created: 2022-10-01 (Thu) Number of data: 100,000

Location is New York
and Weather is Rainy
and Hours > 20
Add condition Delete group

or

Tag is Empty
or Tag is Dogs
or Tag is Cats
Add condition Delete group

or

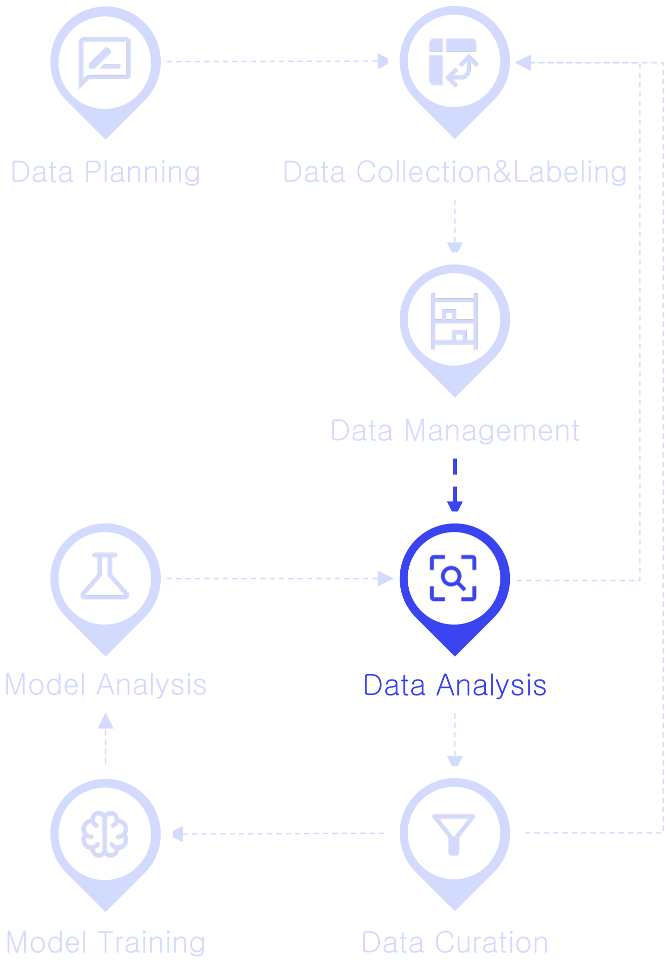
Slice is Bucket_0423
or Tag is Edge Case-01
or Tag is Edge Case-02
Add condition Delete group

or

Choose Condition is Choose Condition
Add condition Delete group

Add condition group

☑		Guide Version	Uploaded Date	Custom Tag
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2
☑	More than a year after ...	Guide V1	2023.02.22	Tag1 Tag2



Data Planning

Data Collection&Labeling

Data Management

Model Analysis

Data Analysis

Model Training

Data Curation

Data Analysis

Analyze data appropriateness. E.g. **coverage**

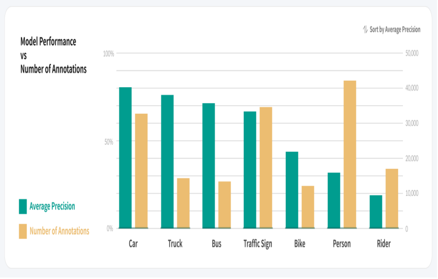
Location of image capture

Time of image capture

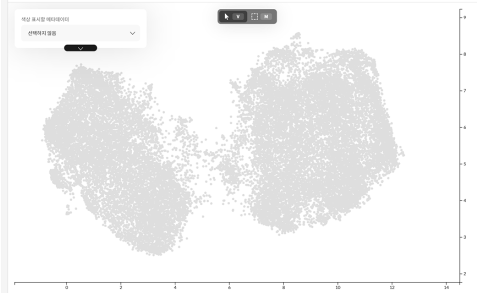
Age groups of provided vocal data

Type of voice recording device

Metadata



Statistics of labels per class

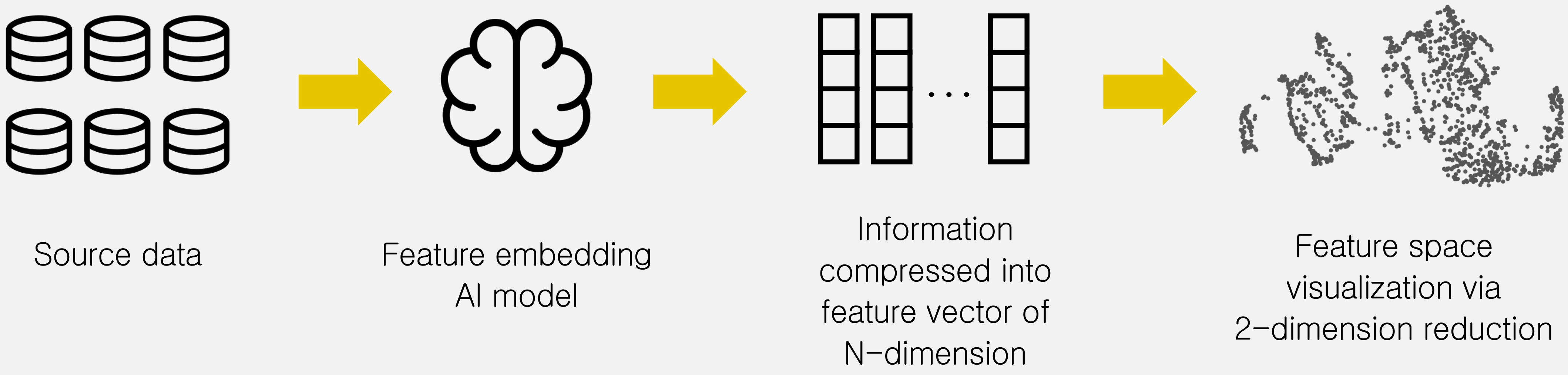


색상 표시할 메타데이터

선택하지 않음

Feature Space

What is feature space?



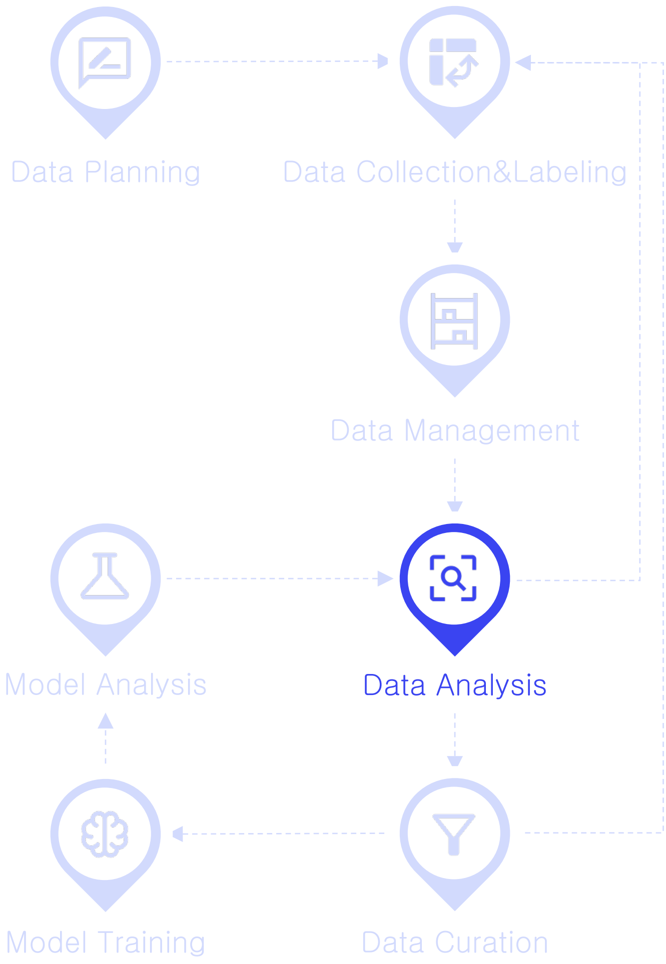
Source data

Feature embedding AI model

Information compressed into feature vector of N–dimension

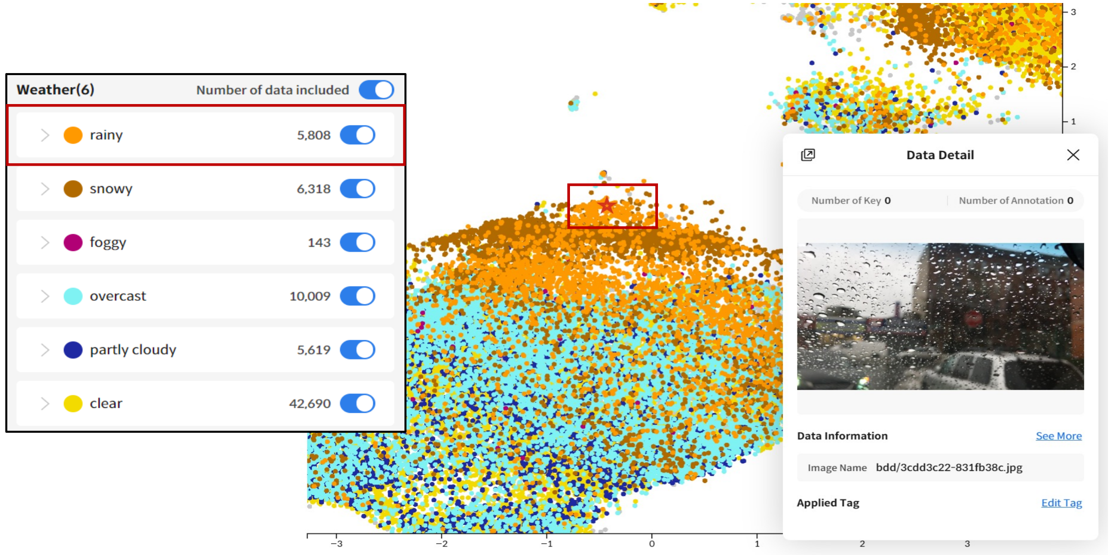
Feature space visualization via 2–dimension reduction

When feature is extracted from data, the data with similar characteristics
in the perspective of the trained AI model tend to stay close in feature space



Data Analysis

AI scenario analysis according to feature space distribution based on metadata



Weather(6)	Number of data included
● rainy	5,808
● snowy	6,318
● foggy	143
● overcast	10,009
● partly cloudy	5,619
● clear	42,690

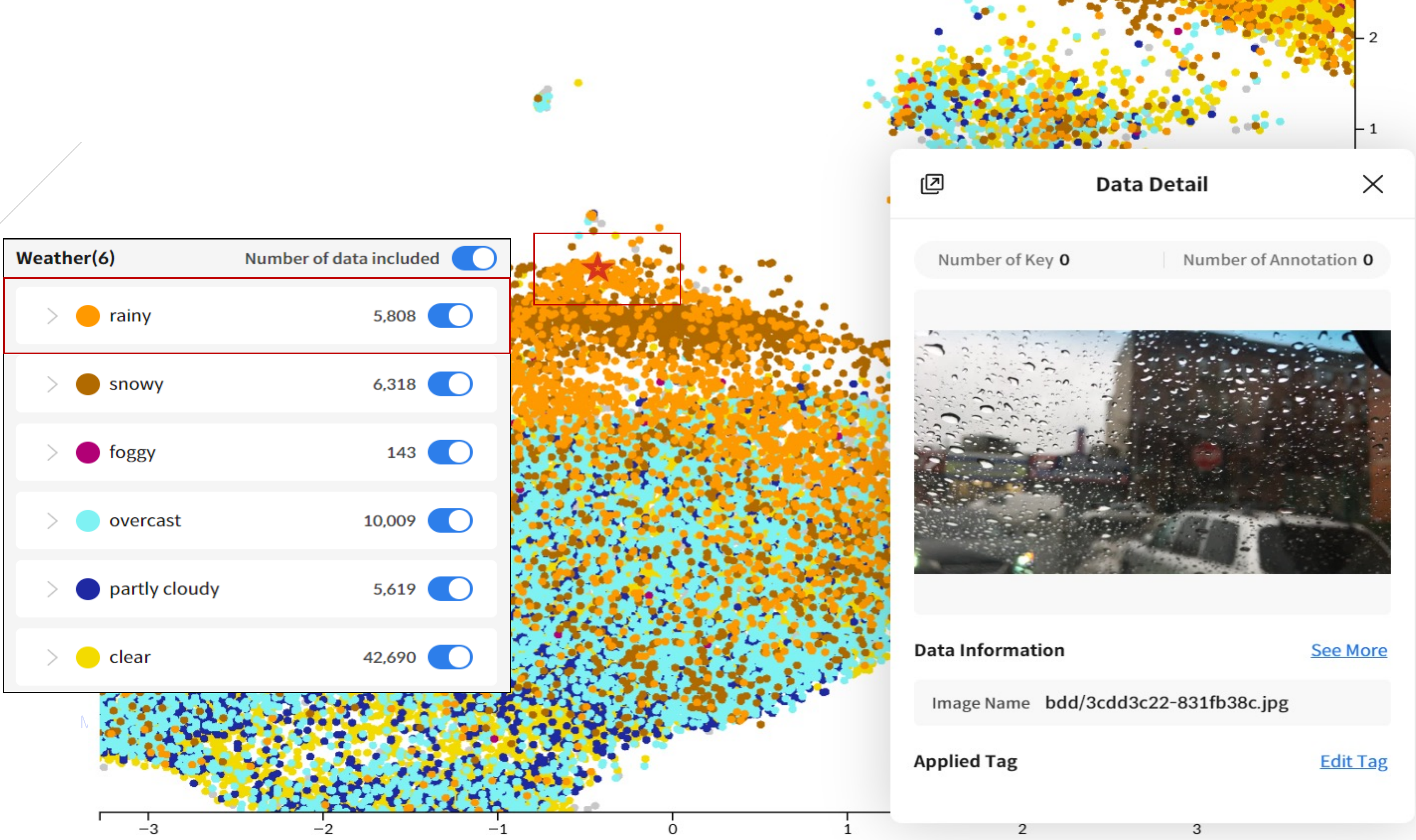
Data Detail

Number of Key 0 | Number of Annotation 0

Data Information

Image Name bdd/3cdd3c22-831fb38c.jpg

Applied Tag



Weather(6) Number of data included

●	rainy	5,808
●	snowy	6,318
●	foggy	143
●	overcast	10,009
●	partly cloudy	5,619
●	clear	42,690

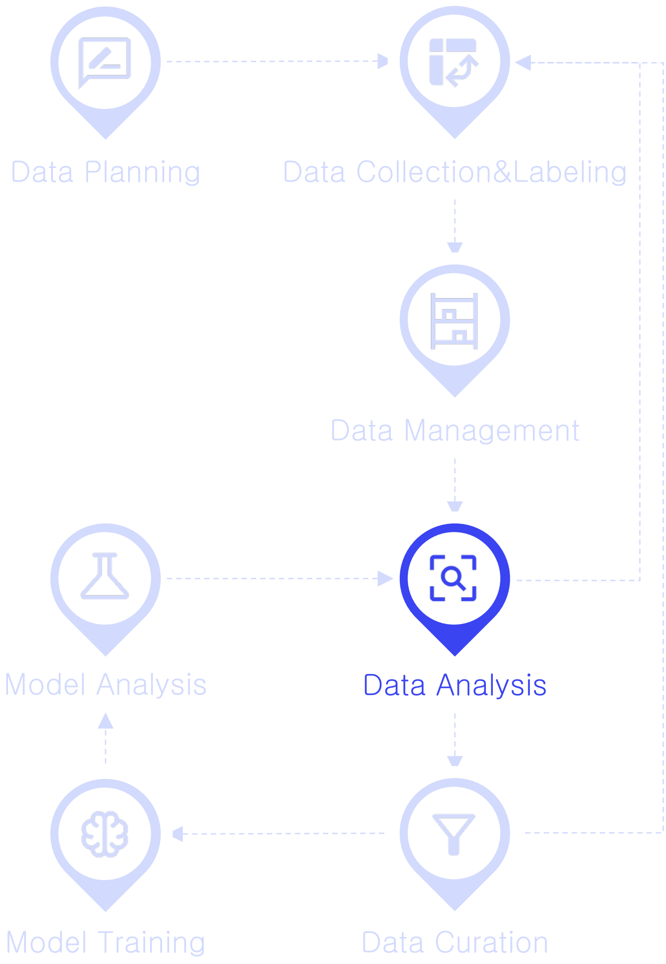
Data Detail

Number of Key **0** | Number of Annotation **0**

Data Information See More

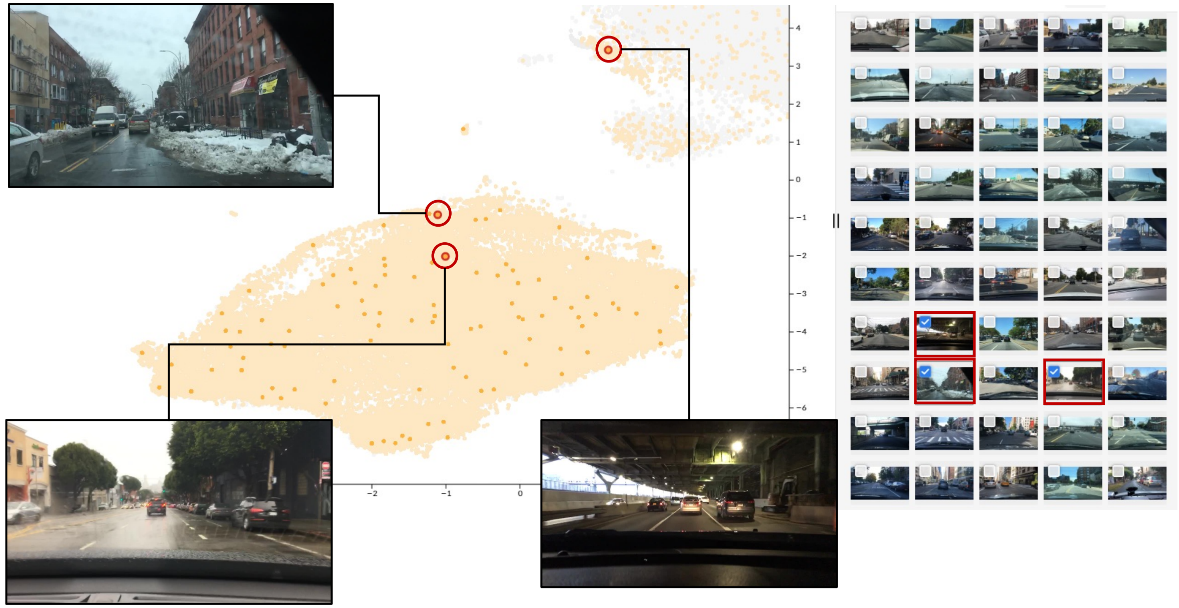
Image Name bdd/3cdd3c22-831fb38c.jpg

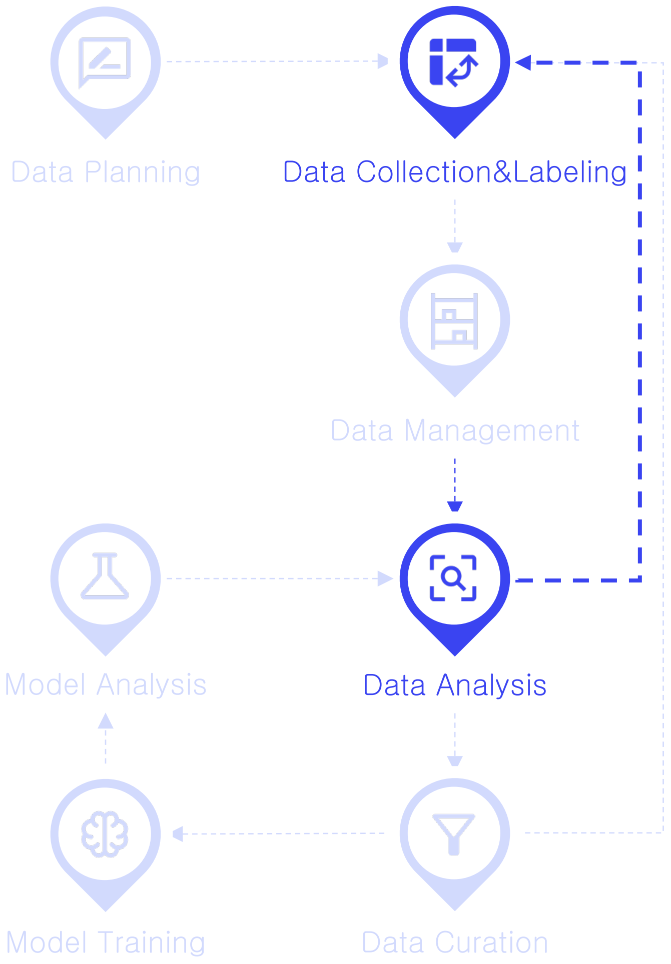
Applied Tag Edit Tag



Data Analysis

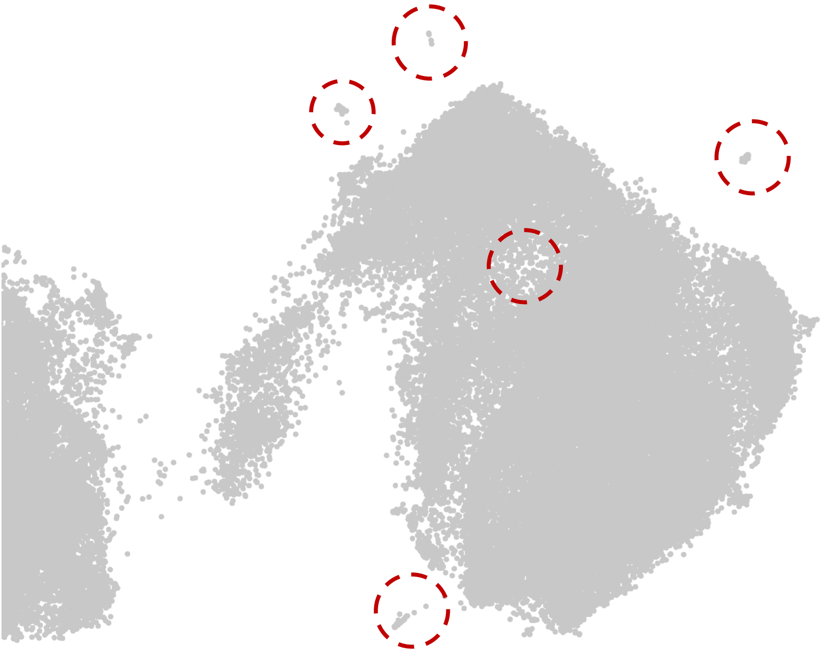
Scenario analysis by reviewing main images chosen via mathematical curation to increase coverage

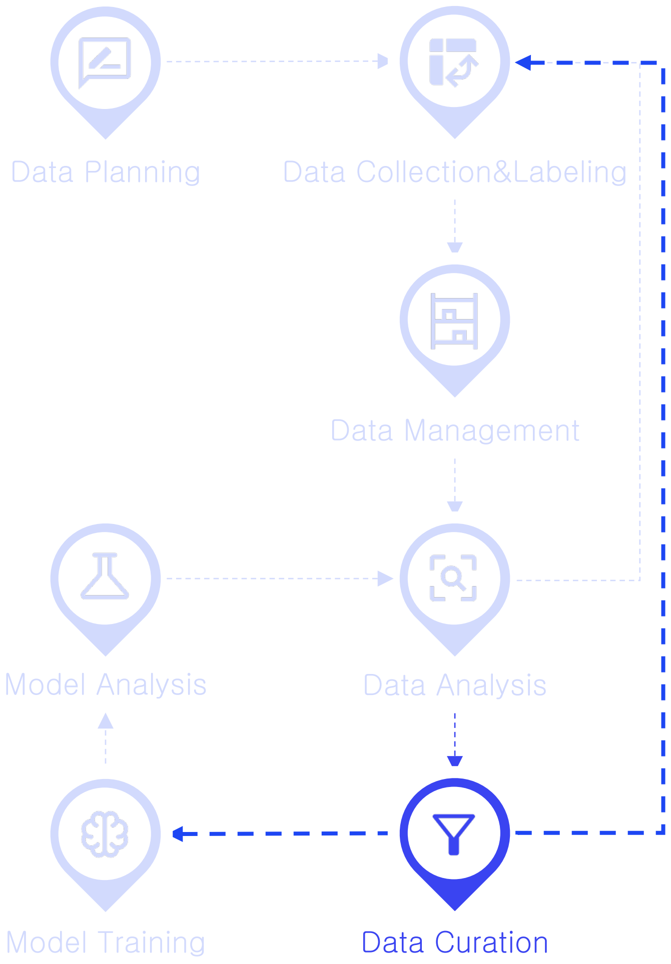




Data Analysis

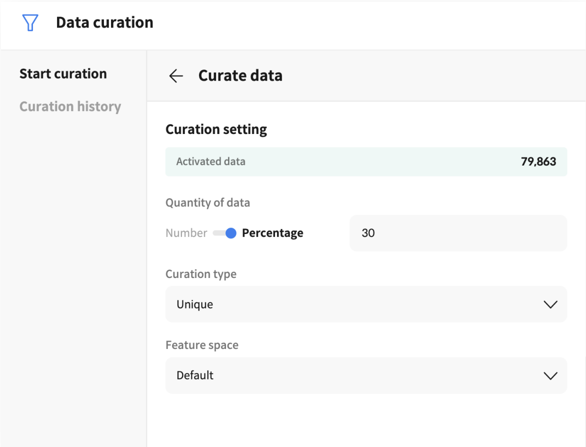
Collect and process additional data that is lacking in the feature space





Data Curation

Curate dataset with coverage in mind, label, and train AI model


Data curation

Start curation
Curation history

Curate data

Curation setting

Activated data 79,863

Quantity of data

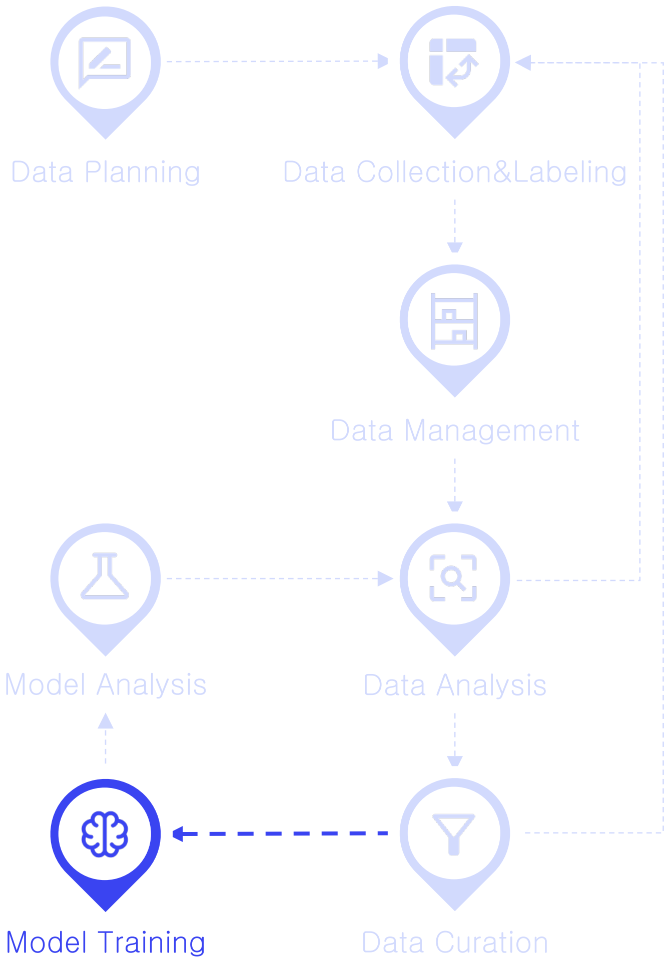
Number ● Percentage 30

Curation type

Unique

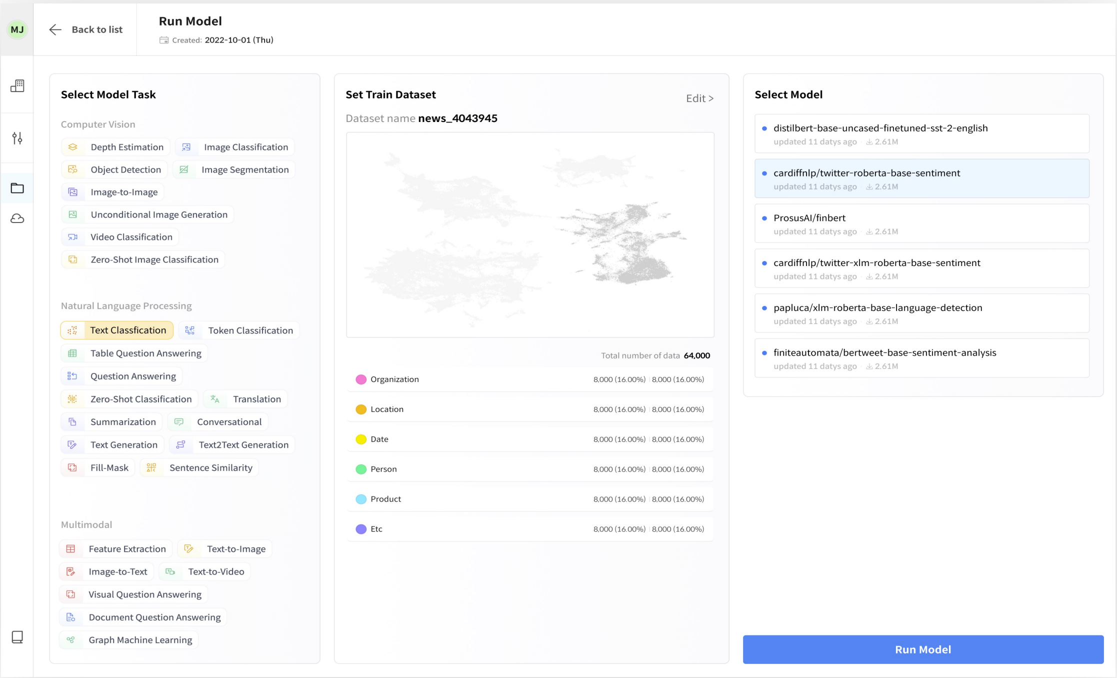
Feature space

Default



Model Training

Select and train the model you want from the list of models with an open API



Run Model

Created: 2022-10-01 (Thu)

Select Model Task

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classfication
- Token Classification
- Table Question Answering
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text2Text Generation
- Fill-Mask
- Sentence Similarity

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Set Train Dataset

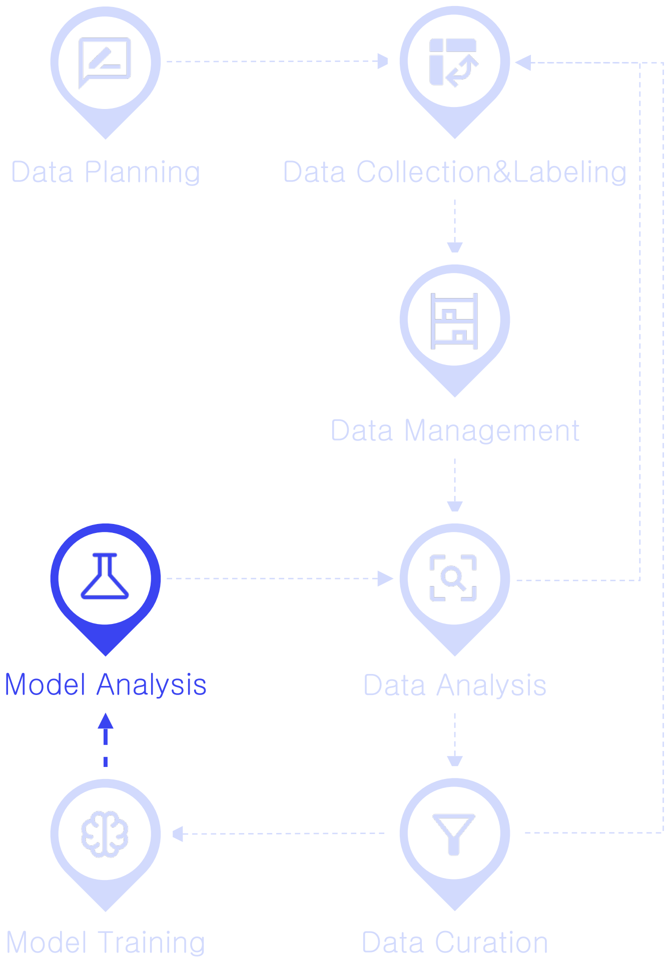
Dataset name **news_4043945**

Total number of data **64,000**

Organization	8,000 (16.00%)	8,000 (16.00%)
Location	8,000 (16.00%)	8,000 (16.00%)
Date	8,000 (16.00%)	8,000 (16.00%)
Person	8,000 (16.00%)	8,000 (16.00%)
Product	8,000 (16.00%)	8,000 (16.00%)
Etc	8,000 (16.00%)	8,000 (16.00%)

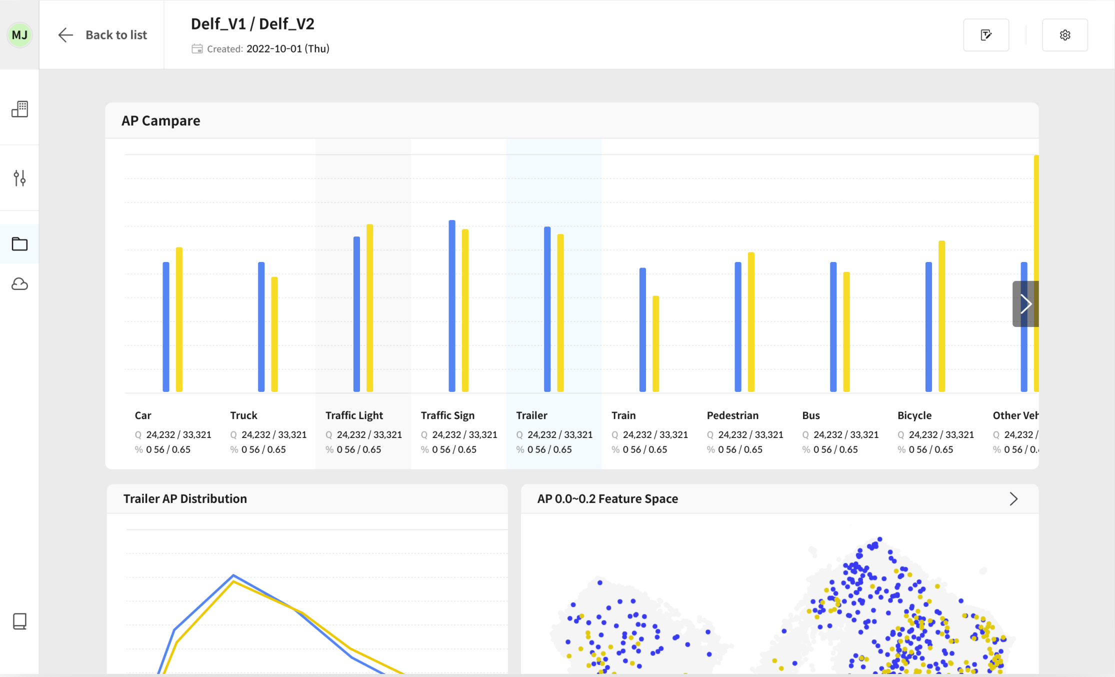
Select Model

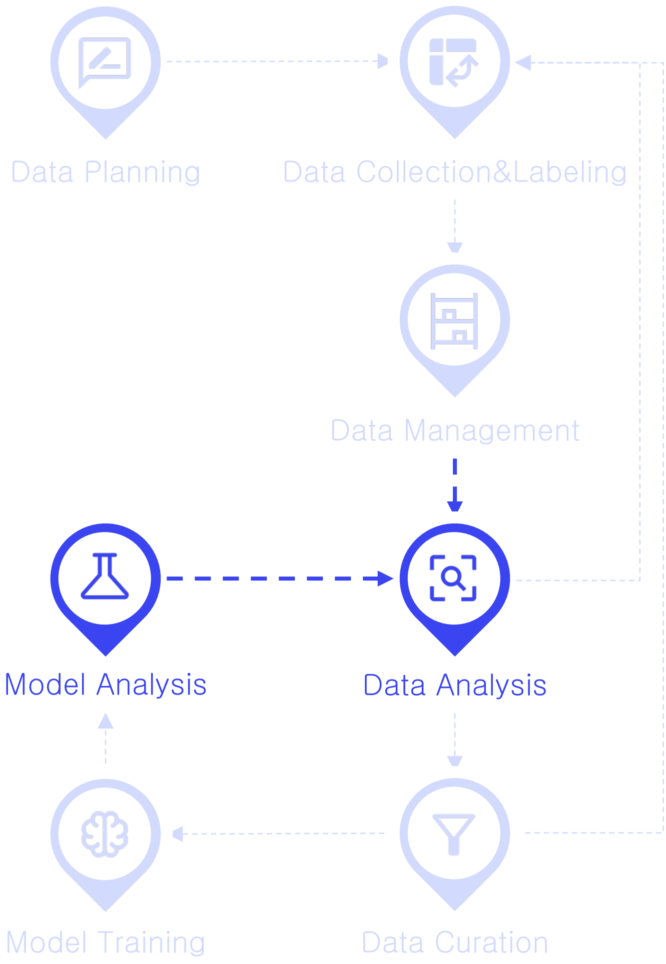
- distilbert-base-uncased-finetuned-sst-2-english
  updated 11 datys ago · 2.61M
- cardiffnlp/twitter-roberta-base-sentiment
  updated 11 datys ago · 2.61M
- ProsusAI/finbert
  updated 11 datys ago · 2.61M
- cardiffnlp/twitter-xlm-roberta-base-sentiment
  updated 11 datys ago · 2.61M
- papluca/xlm-roberta-base-language-detection
  updated 11 datys ago · 2.61M
- finiteautomata/bertweet-base-sentiment-analysis
  updated 11 datys ago · 2.61M



Model Analysis

Check the performance of the results of the model on the dashboard





Data Planning

Data Collection&Labeling

Data Management

Model Analysis

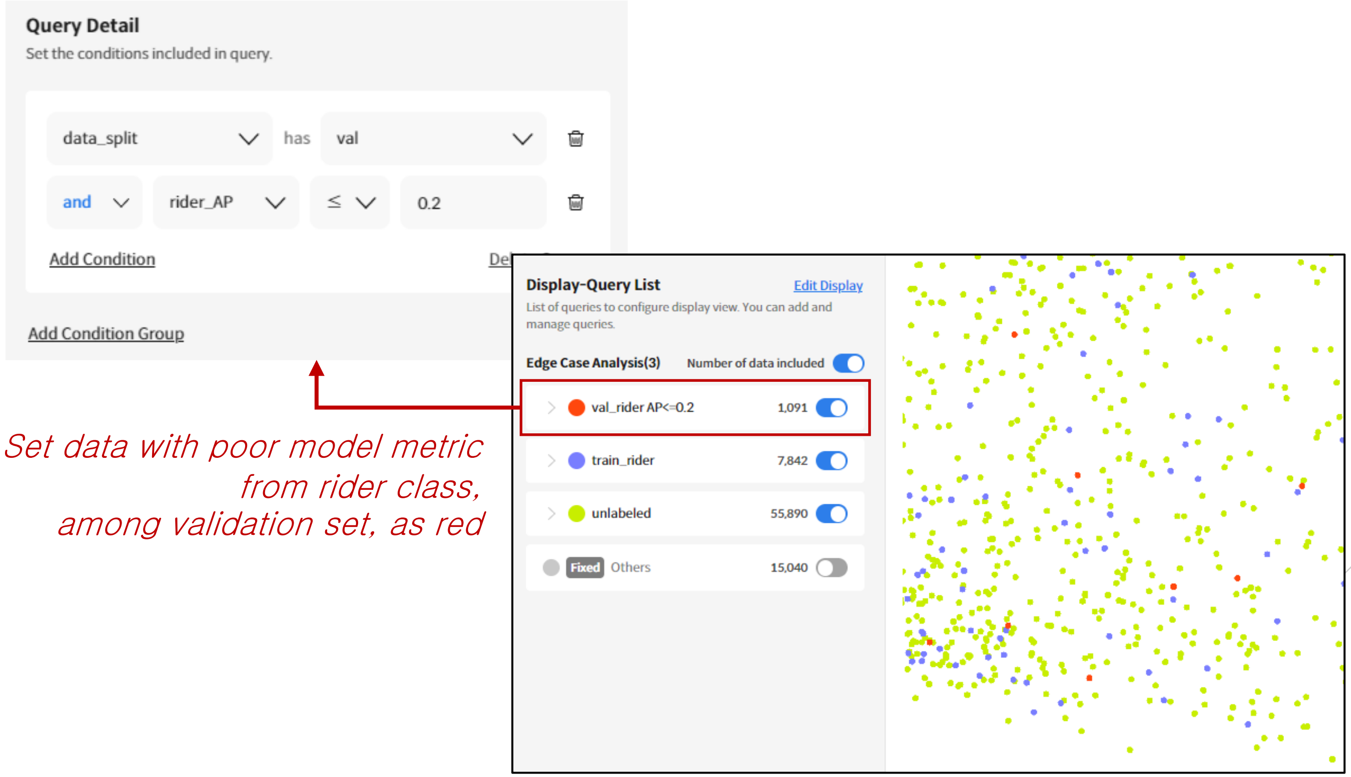
Data Analysis

Model Training

Data Curation

Model & Data Analysis

Combine multiple information such as model metric, data split, and others as queries and mark in color on feature space.



Query Detail

Set the conditions included in query.

data_split has val

and rider_AP ≤ 0.2

Add Condition

Add Condition Group

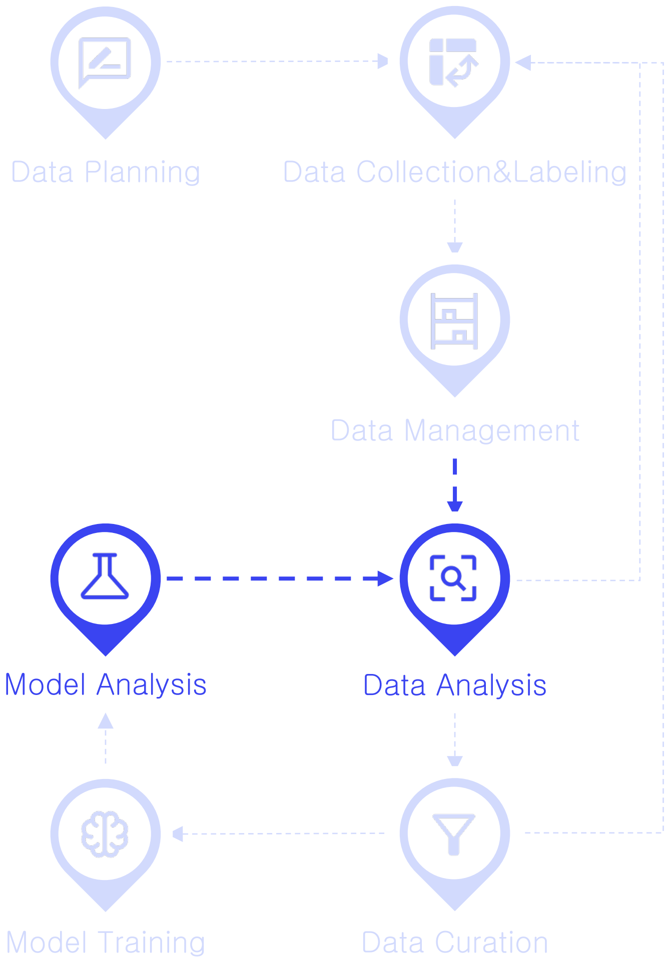
Display-Query List

Edit Display

List of queries to configure display view. You can add and manage queries.

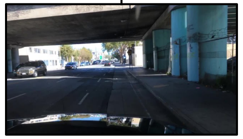
Edge Case Analysis(3)	Number of data included
● val_rider AP<=0.2	1,091
● train_rider	7,842
● unlabeled	55,890
● Fixed Others	15,040

Set data with poor model metric from rider class, among validation set, as red



Model & Data Analysis

Define edge cases by finding common features among data



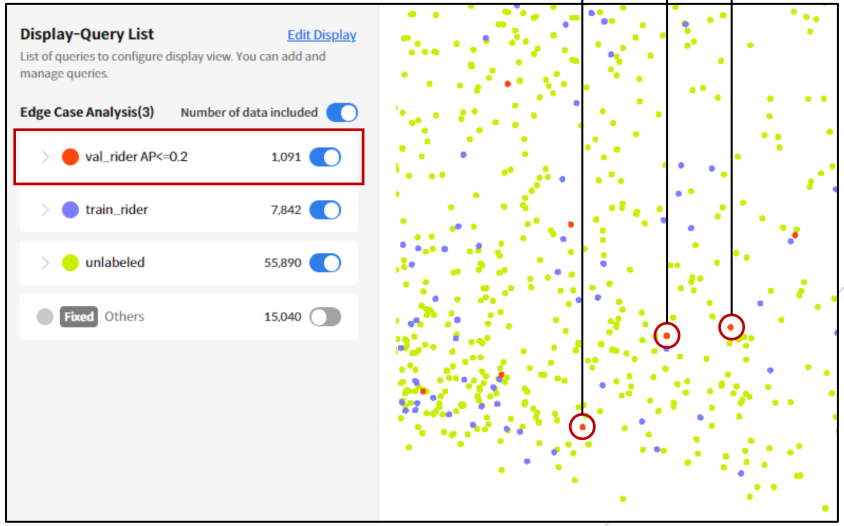
Define edge case as:
“Dark from the shadows under bridges“

↓

Notice lack of train set around edge cases

↓

Check additional data within the area that may be additionally labeled and trained

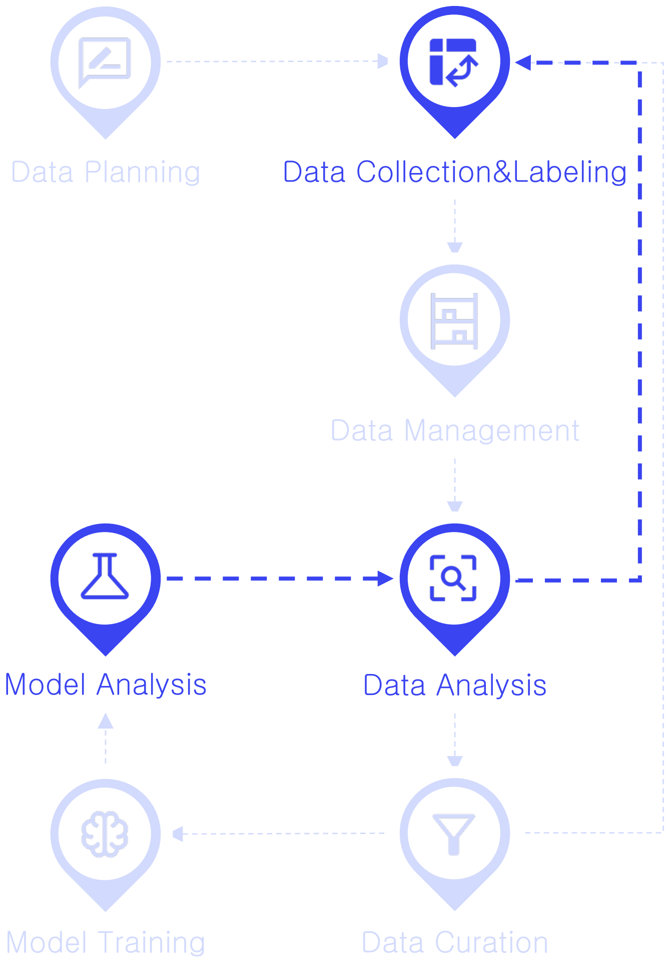


Display-Query List Edit Display

List of queries to configure display view. You can add and manage queries.

Edge Case Analysis(3) Number of data included

●	val_rider AP<=0.2	1,091
●	train_rider	7,842
●	unlabeled	55,890
○ Fixed	Others	15,040



Data Planning

Data Collection&Labeling

Data Management

Model Analysis

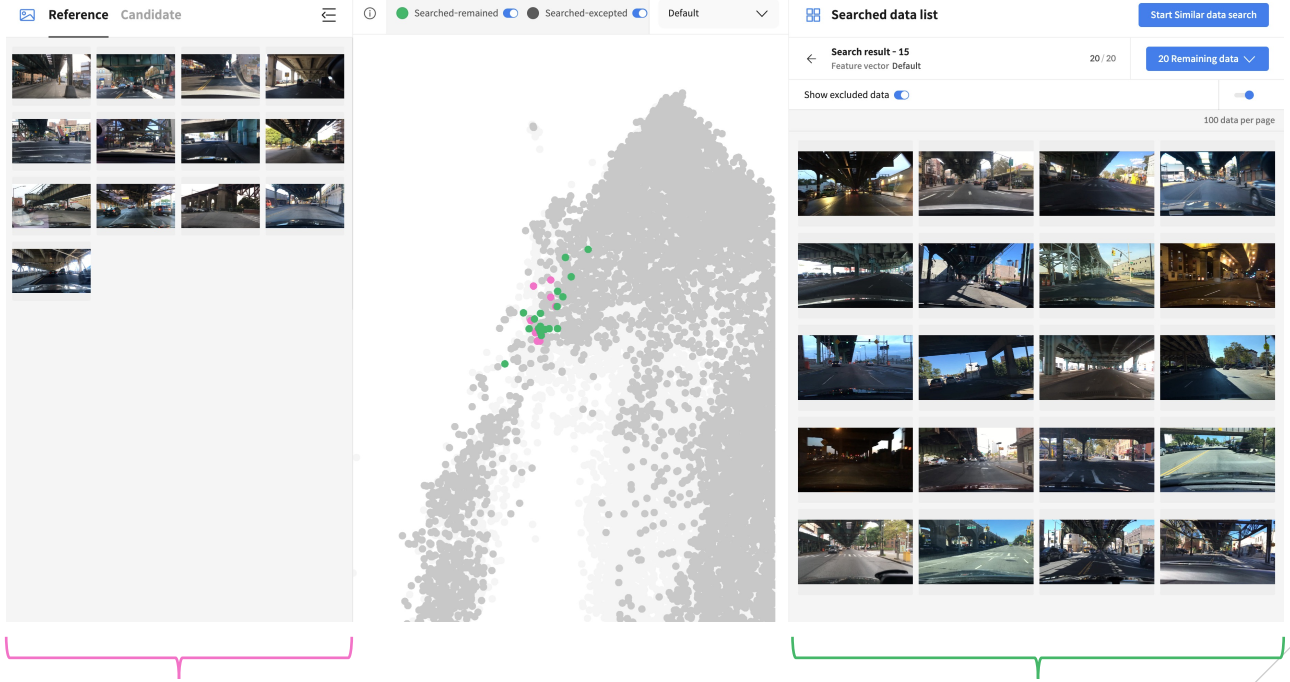
Data Analysis

Model Training

Data Curation

Model & Data Analysis

Curation of data similar to edge cases from unlabled data and labeling them

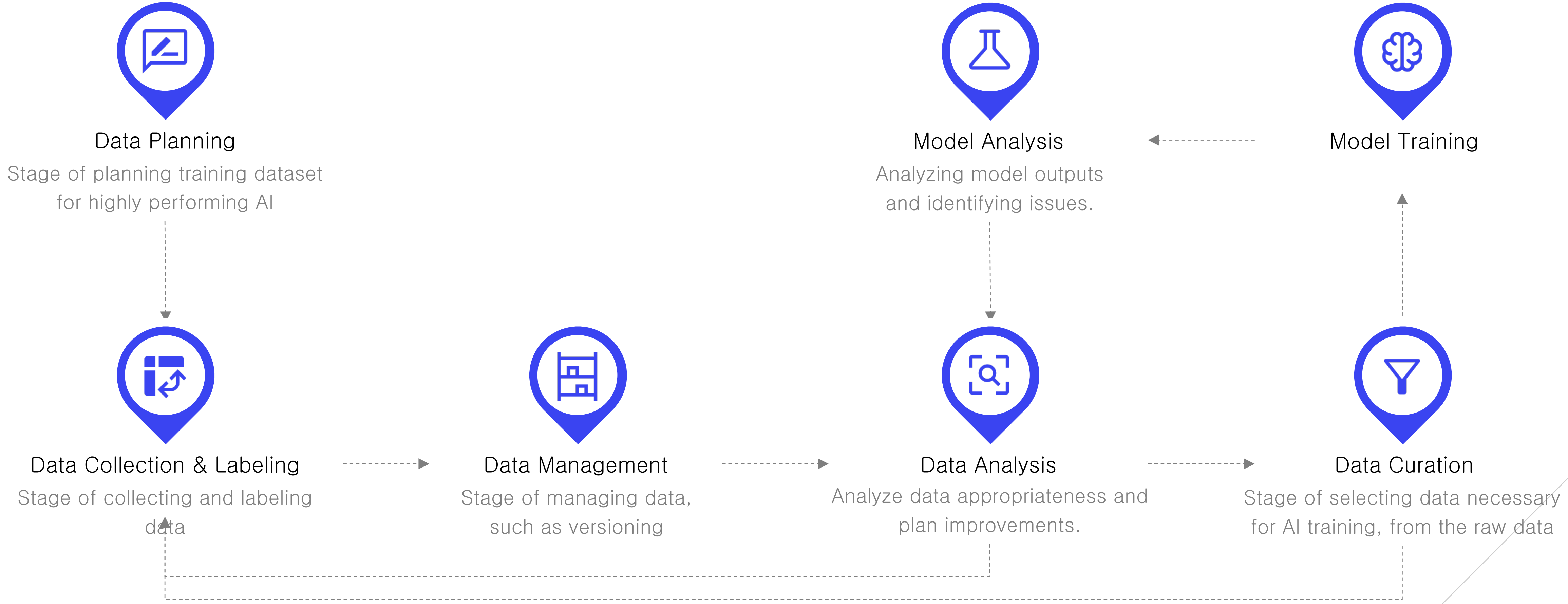


Select edge case image

Similar data search result

DATUMO

AI Data Flywheel



Data Planning
Stage of planning training dataset for highly performing AI

Data Collection & Labeling
Stage of collecting and labeling data

Data Management
Stage of managing data, such as versioning

Data Analysis
Analyze data appropriateness and plan improvements.

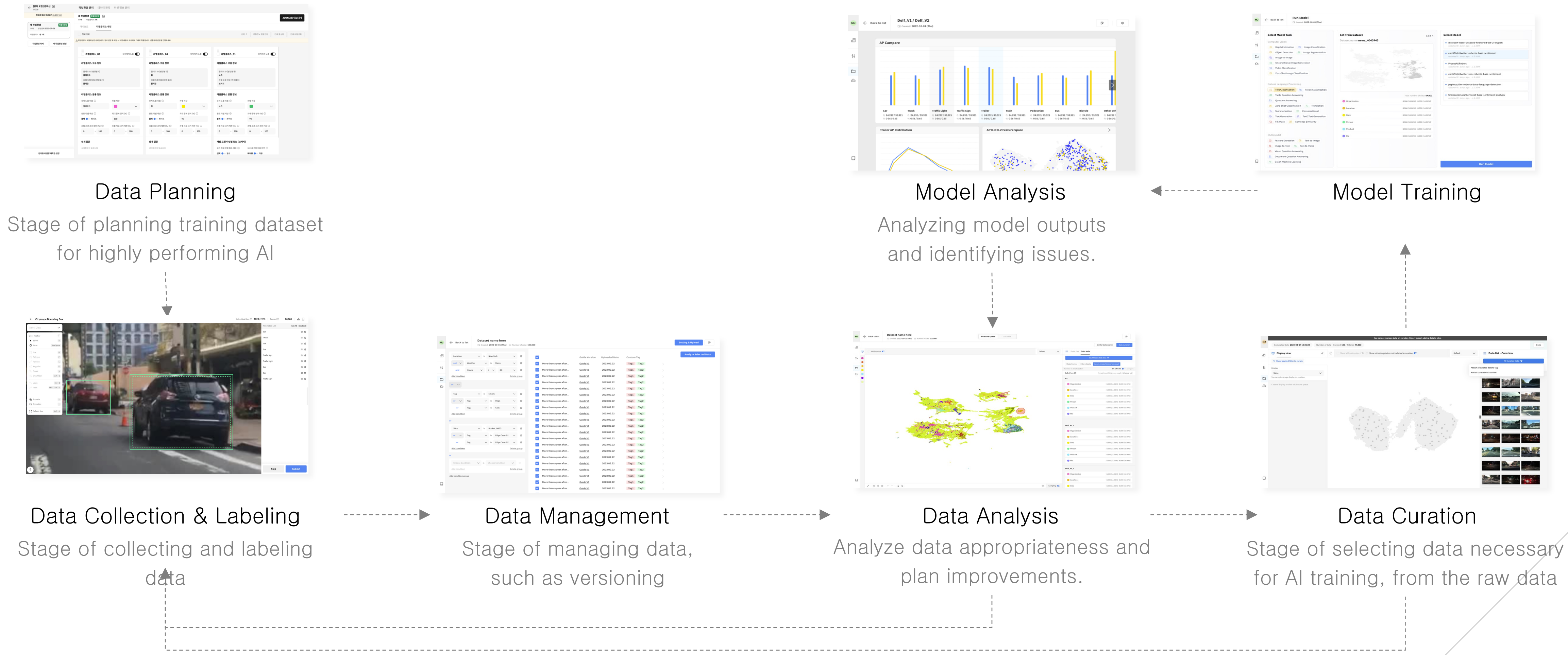
Data Curation
Stage of selecting data necessary for AI training, from the raw data

Model Training

Model Analysis
Analyzing model outputs and identifying issues.

DATUMO

AI Data Flywheel with DATUMO SCOPE



Data Planning

Stage of planning training dataset for highly performing AI

Data Collection & Labeling

Stage of collecting and labeling data

Data Management

Stage of managing data, such as versioning

Data Analysis

Analyze data appropriateness and plan improvements.

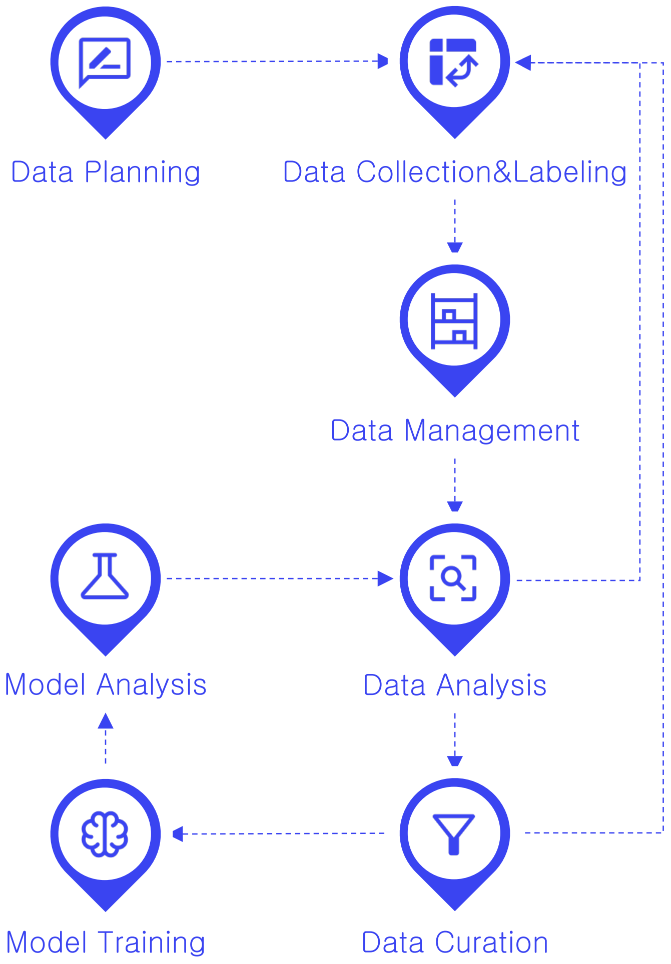
Data Curation

Stage of selecting data necessary for AI training, from the raw data

Model Training

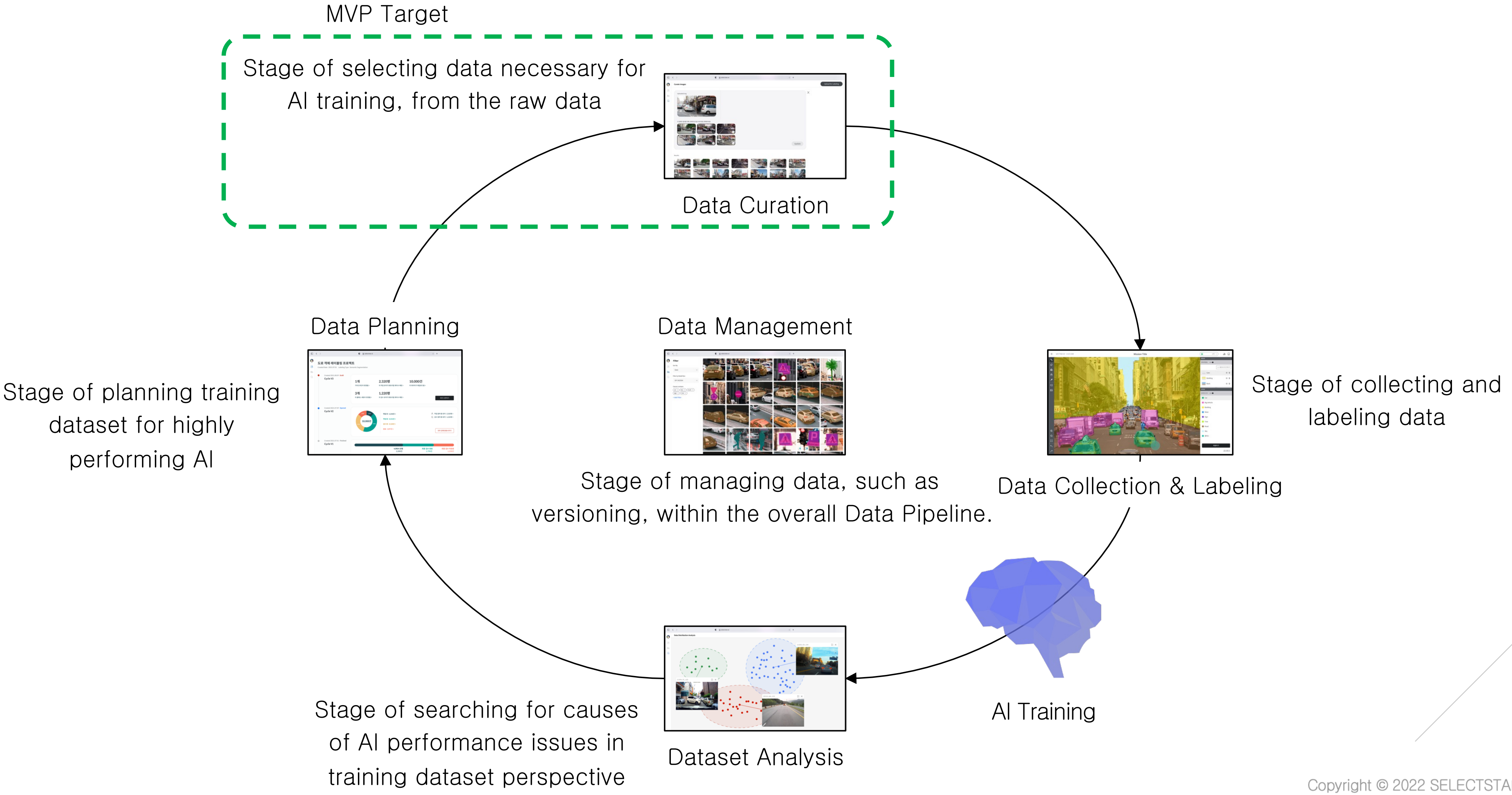
Model Analysis

Analyzing model outputs and identifying issues.



Build, manage, and analyze AI training data using DATUMO SCOPE, and deploy AI services without development.

Data Pipeline



MVP Target

Stage of selecting data necessary for AI training, from the raw data

Data Curation

Stage of collecting and labeling data

Data Collection & Labeling

AI Training

Dataset Analysis

Stage of searching for causes of AI performance issues in training dataset perspective

Data Planning

Stage of planning training dataset for highly performing AI

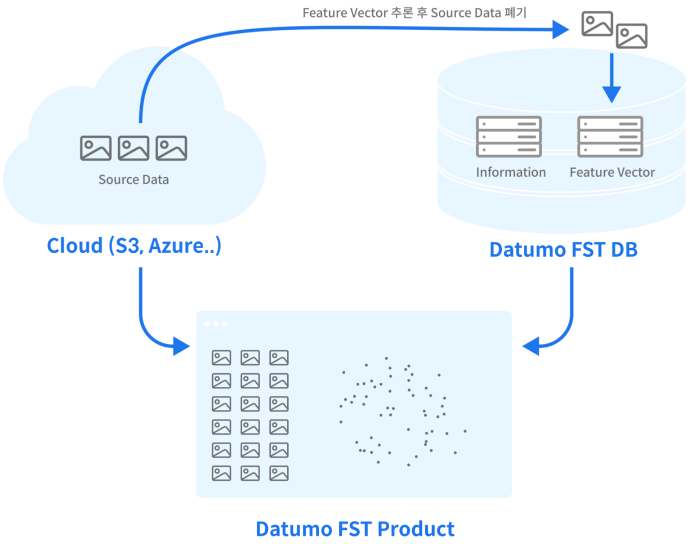
Data Management

Stage of managing data, such as versioning, within the overall Data Pipeline.

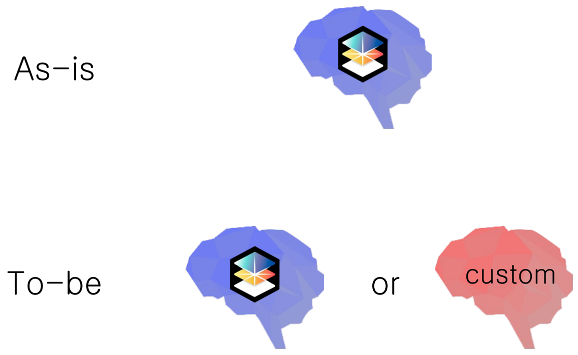
Datumo Feature Space Tool

Integration with user's AI lifecycle

No need to upload source data when utilizing Cloud Integration

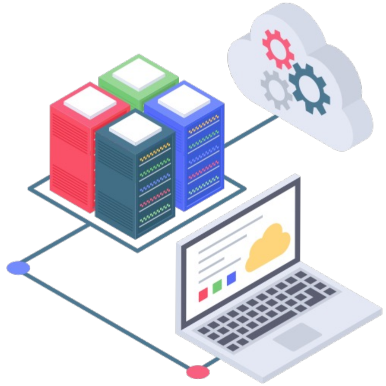


Upload feature vector extracted from user's AI model
('23.01 Release)



Feature Embedding Model

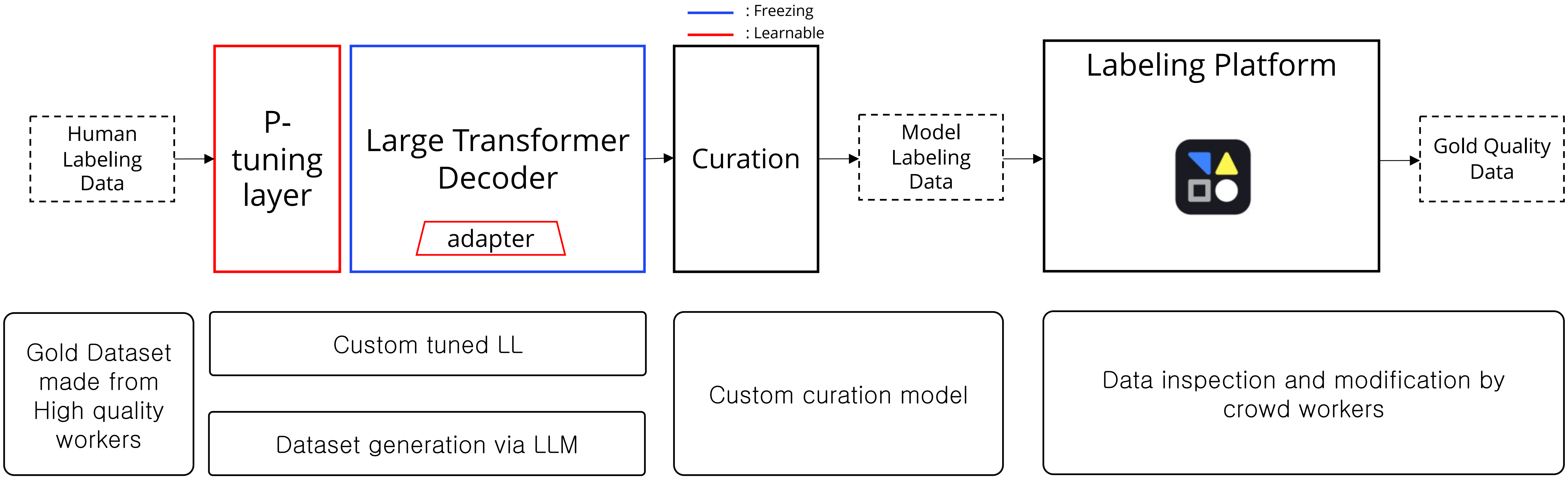
Enable the use of intranet environment by providing on-premise
('23.03 Release)



NLP Technology

Semi-automated text data generation.

high–quality data production pipeline utilizing our in–house tuned Large Language Model, curation algorithm, and labeling platform with AI–generated data.



— : Freezing

— : Learnable

Human Labeling Data

P-tuning layer

Large Transformer Decoder

adapter

Curation

Model Labeling Data

Labeling Platform

Gold Quality Data

Gold Dataset made from High quality workers

Custom tuned LL

Dataset generation via LLM

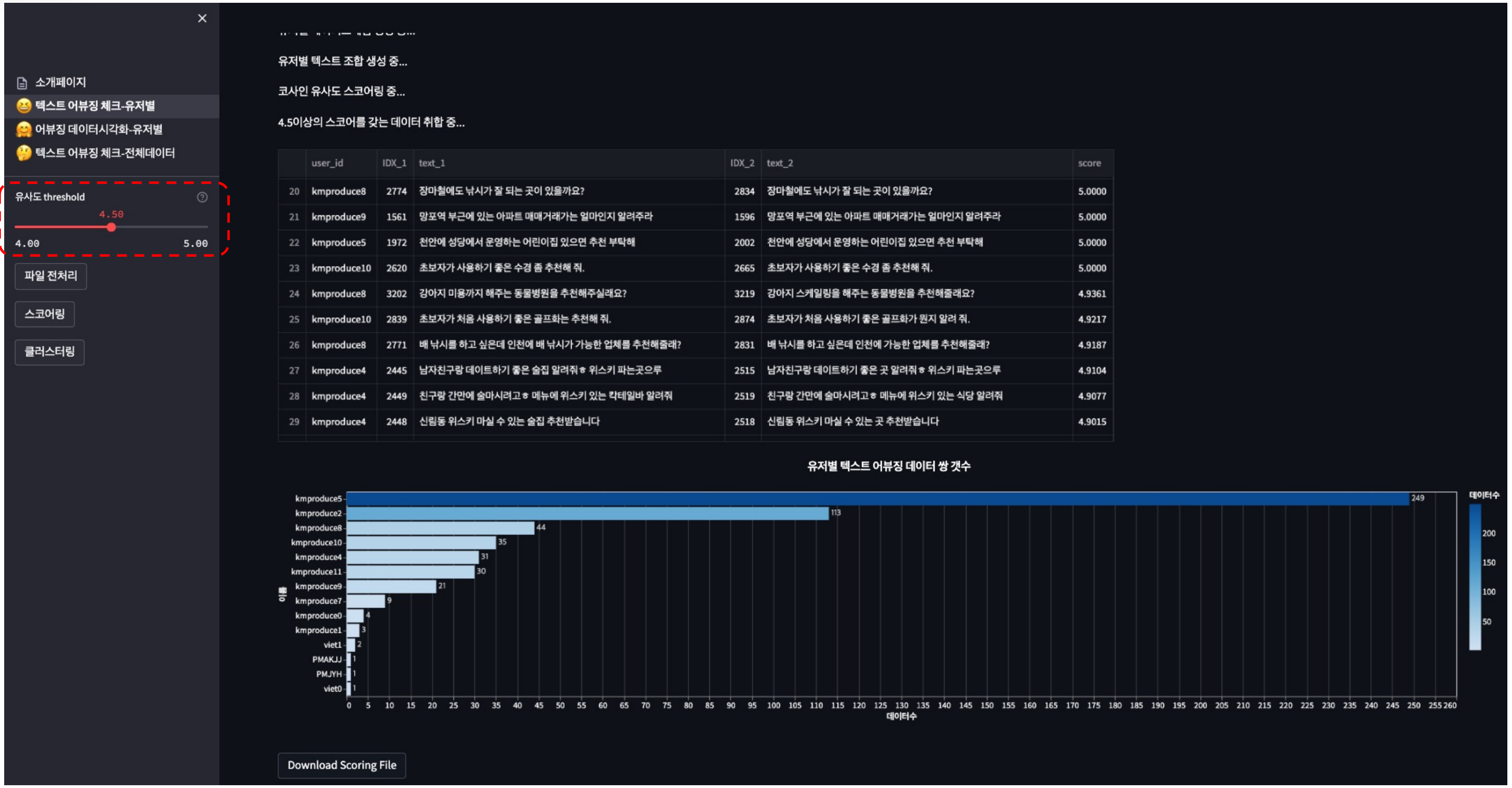
Custom curation model

Data inspection and modification by crowd workers

DATUMO

NLP Technology

Filtering of similar text submissions.



유사도 threshold 4.50 (4.00 – 5.00)

유저별 텍스트 조합 생성 중...

코사인 유사도 스코어링 중...

4.5이상의 스코어를 갖는 데이터 취합 중...

	user_id	IDX_1	text_1	IDX_2	text_2	score
20	kmproduce8	2774	장마철에도 낚시가 잘 되는 곳이 있을까요?	2834	장마철에도 낚시가 잘 되는 곳이 있을까요?	5.0000
21	kmproduce9	1561	망포역 부근에 있는 아파트 매매거래가는 얼마인지 알려주라	1596	망포역 부근에 있는 아파트 매매거래가는 얼마인지 알려주라	5.0000
22	kmproduce5	1972	천안에 성당에서 운영하는 어린이집 있으면 추천 부탁해	2002	천안에 성당에서 운영하는 어린이집 있으면 추천 부탁해	5.0000
23	kmproduce10	2620	초보자가 사용하기 좋은 수경 좀 추천해 줘.	2665	초보자가 사용하기 좋은 수경 좀 추천해 줘.	5.0000
24	kmproduce8	3202	강아지 미용까지 해주는 동물병원을 추천해주실래요?	3219	강아지 스케일링을 해주는 동물병원을 추천해줄래요?	4.9361
25	kmproduce10	2839	초보자가 처음 사용하기 좋은 골프화는 추천해 줘.	2874	초보자가 처음 사용하기 좋은 골프화가 뭔지 알려 줘.	4.9217
26	kmproduce8	2771	배 낚시를 하고 싶은데 인천에 배 낚시가 가능한 업체를 추천해줄래?	2831	배 낚시를 하고 싶은데 인천에 가능한 업체를 추천해줄래?	4.9187
27	kmproduce4	2445	남자친구랑 데이트하기 좋은 술집 알려줘ㅎ 위스키 파는곳으루	2515	남자친구랑 데이트하기 좋은 곳 알려줘ㅎ 위스키 파는곳으루	4.9104
28	kmproduce4	2449	친구랑 간만에 술마시려고ㅎ 메뉴에 위스키 있는 칵테일바 알려줘	2519	친구랑 간만에 술마시려고ㅎ 메뉴에 위스키 있는 식당 알려줘	4.9077
29	kmproduce4	2448	신림동 위스키 마실 수 있는 술집 추천받습니다	2518	신림동 위스키 마실 수 있는 곳 추천받습니다	4.9015

Different abuse criteria can be set depending on the data by using similarity thresholds

- Similarity scores between pairs of text submitted by users are checked, with a range of 0 to 5 indicating how similar they are.
- The number of similar text pairs can be visualized for each user.

Free Data Labeling Service for academic purposes

CVPR JUNE 19-24 2022
NEW ORLEANS • LOUISIANA

InstaOrder
(w. POSTECH)

NEURAL INFORMATION PROCESSING SYSTEMS

KLUE : Korean Language Understanding Evaluation
(w. Upstage, NAVER, NYU, KAIST)

EMNLP 2022

KOLD: Korean Offensive Language Dataset
(w. KAIST)

EMNLP 2022

Analyzing Norm Violations in Real-Time Live-Streaming Chat
(w. Softly.AI, USC)

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE

Split GCN

Free offerings

- Try Datumo Scope Now! For FREE
- Free Data Labeling Service for Academic purposes

For more benefits, please contact contact@datumo.com

We aim to innovate the AI training data market through technology

contact@datumo.com

DATUMO.COM