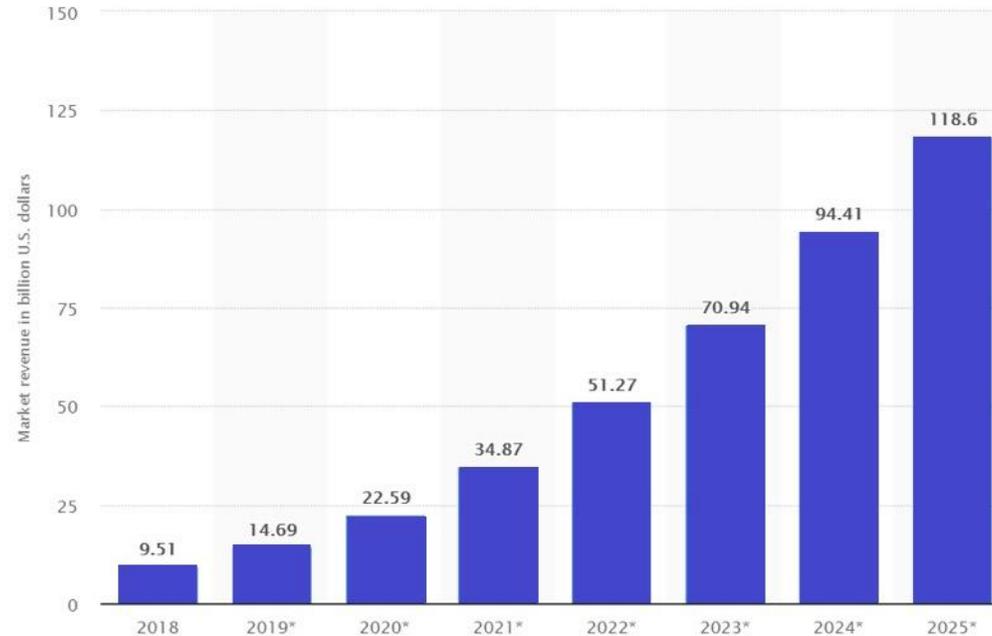


모바일 사용자와 인공지능을 통한 AI 학습데이터 수집 및 레이블링

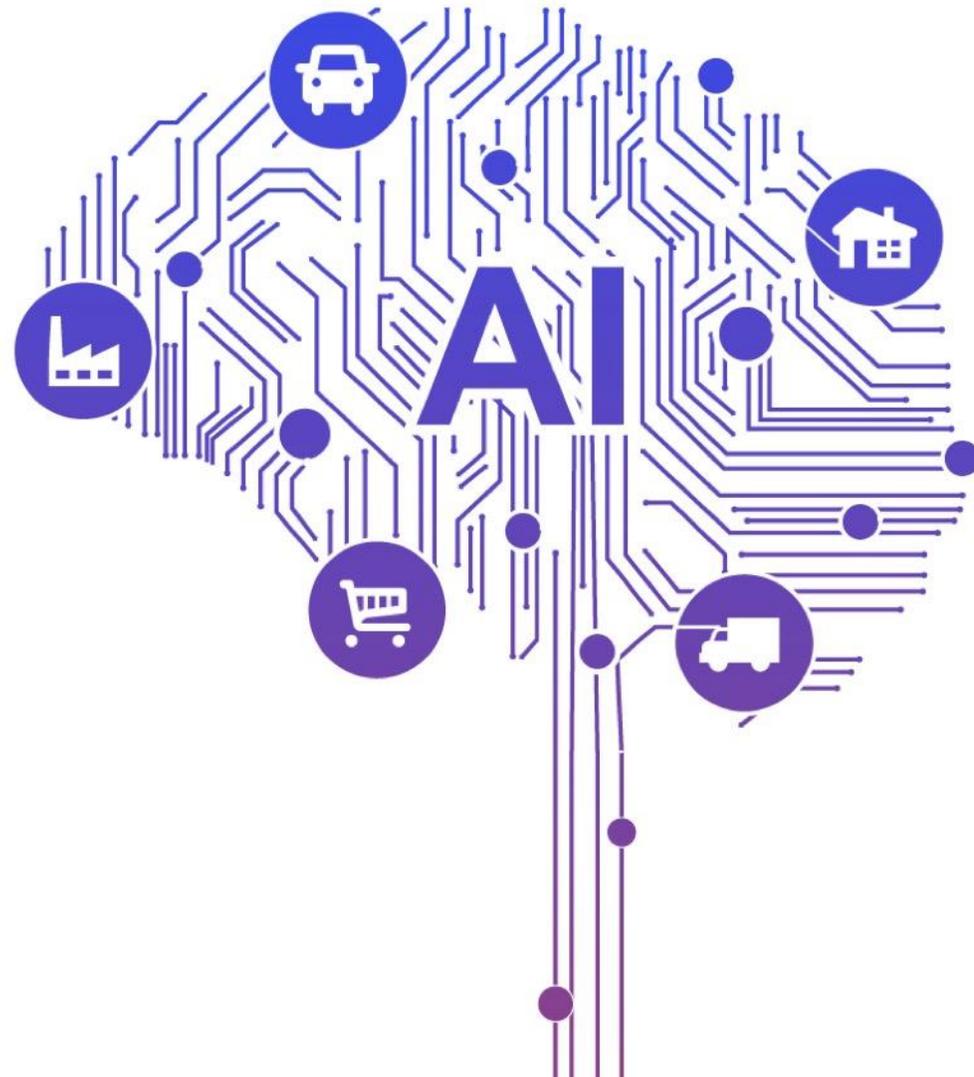


인공지능(AI)은 딥러닝(Deep learning)의 발전과 함께
이미지 인식, 텍스트 분석, 음성 Assistant, 자율주행 등과 같은
다양한 영역에서 **급속한 발전**을 보이고 있음



© Statista 2019

실제로 글로벌 마켓 리서치 업체 Statista에 따르면,
2025년까지 **AI전체 시장규모는 약 142조원**이 될 것으로 예상하고 있음





음성



텍스트



영상



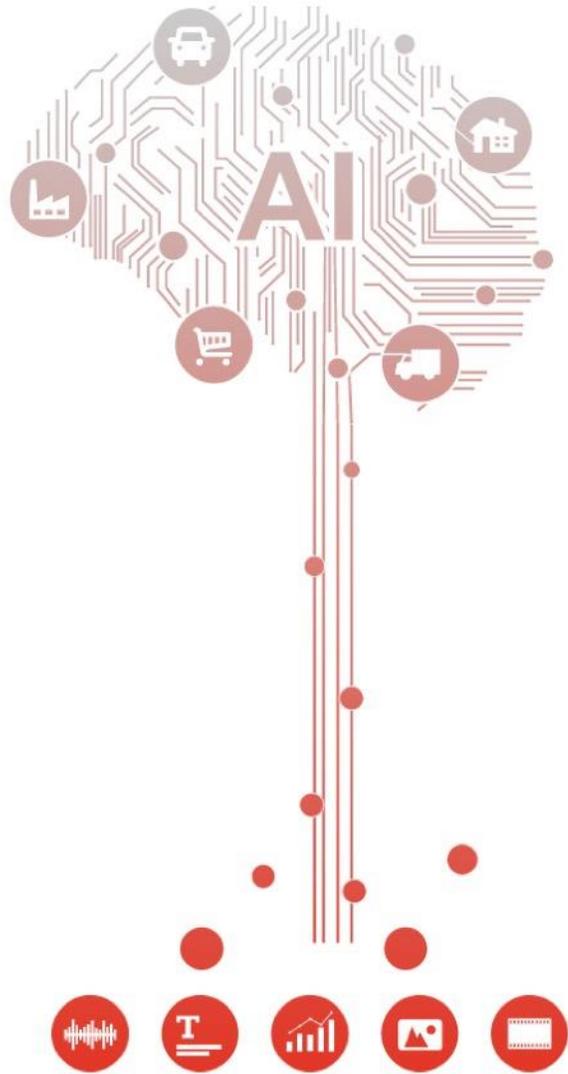
이미지



기타

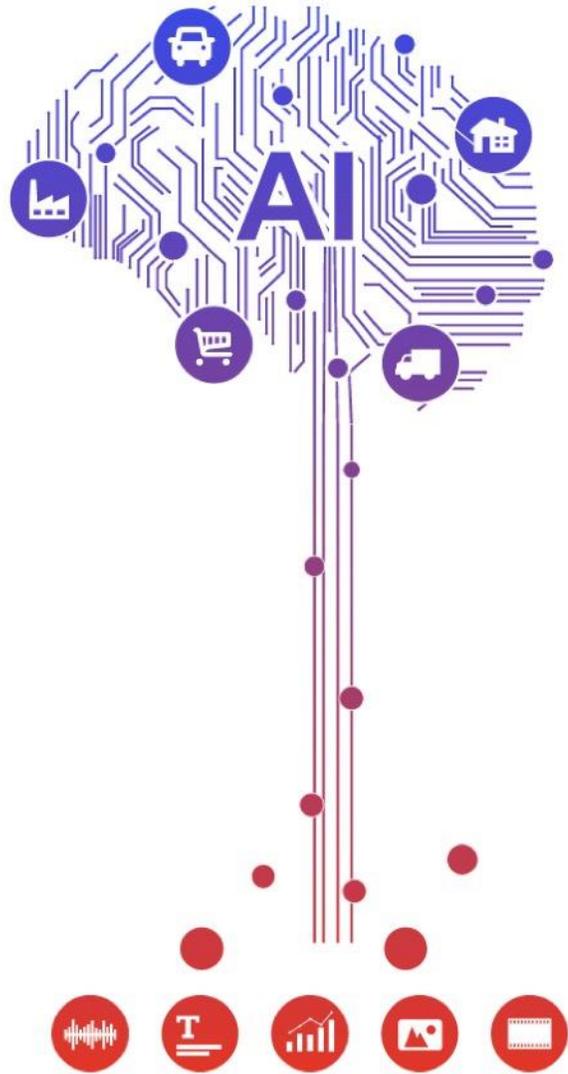
1

“높은 품질의 학습데이터”



1

“높은 품질의 학습데이터”



2

“좋은 성능의 인공지능”

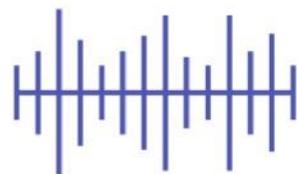
학습데이터 생산과정



1. 다양하고 많은 양의 데이터를 수집



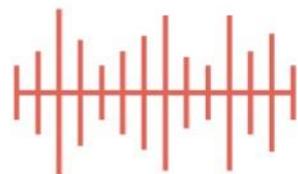
(서울, 25, 남)

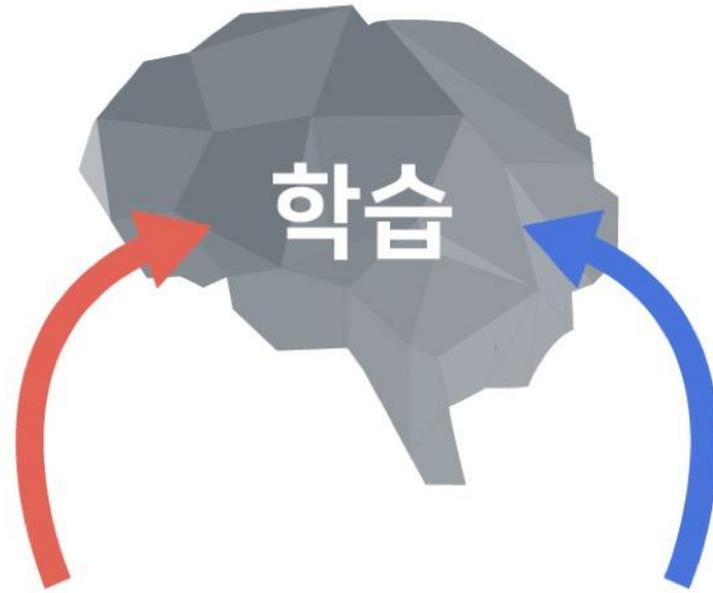


음성인식
시스피커



(경상도, 8, 여자)





다양함

지역, 연령, 성별

많은 양

1만 시간 이상

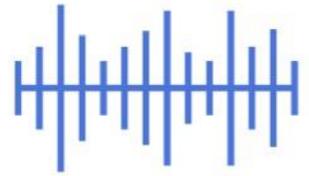


**범용성 + 높은 성능을
갖춘 AI**

학습데이터 생산과정



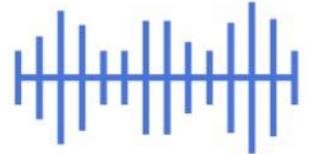
2. AI 학습을 위해 수집된 데이터 모두 가공



“노래 들어줘”



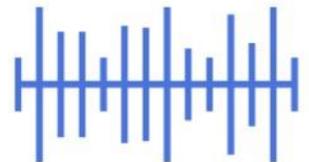
“음악 재생해줘”



“뮤직 플레이해줘”



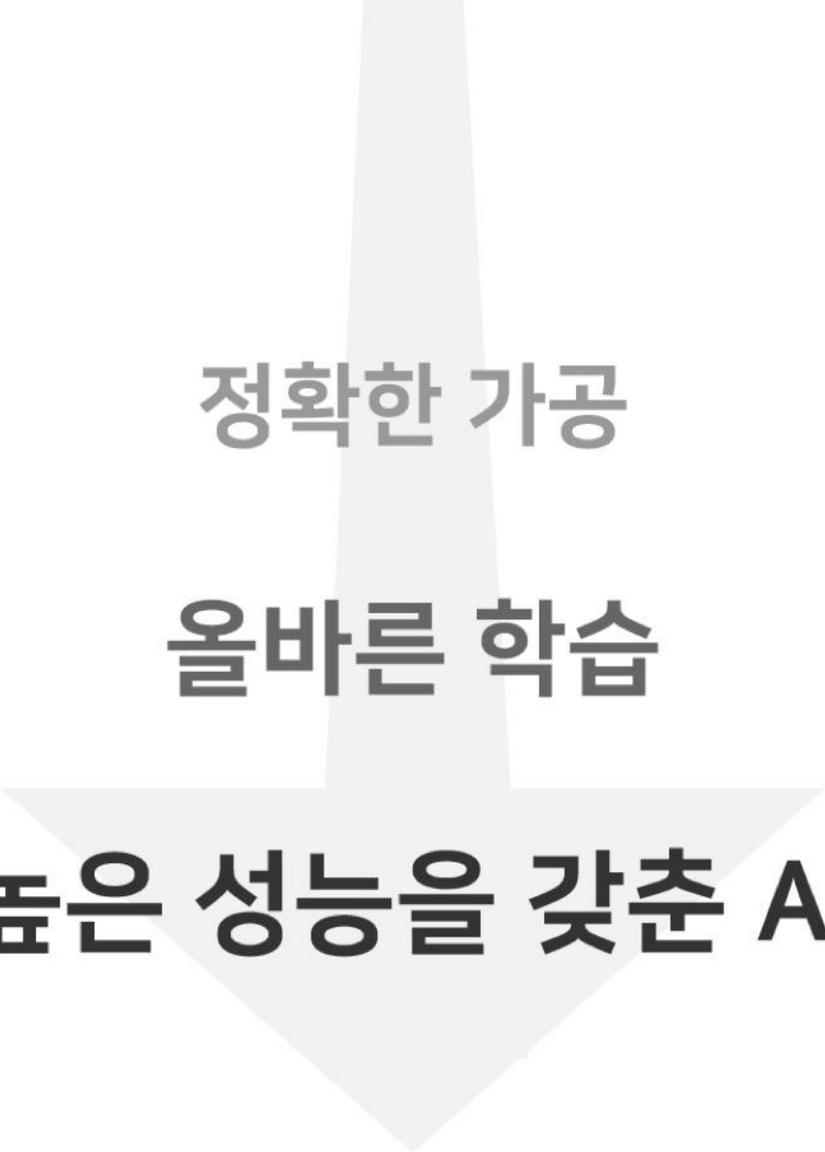
“노래 재생해주세요”



“음악 좀 들어봐야”



“헤이 디제이 드랍더비트”

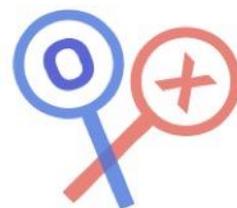


정확한 가공

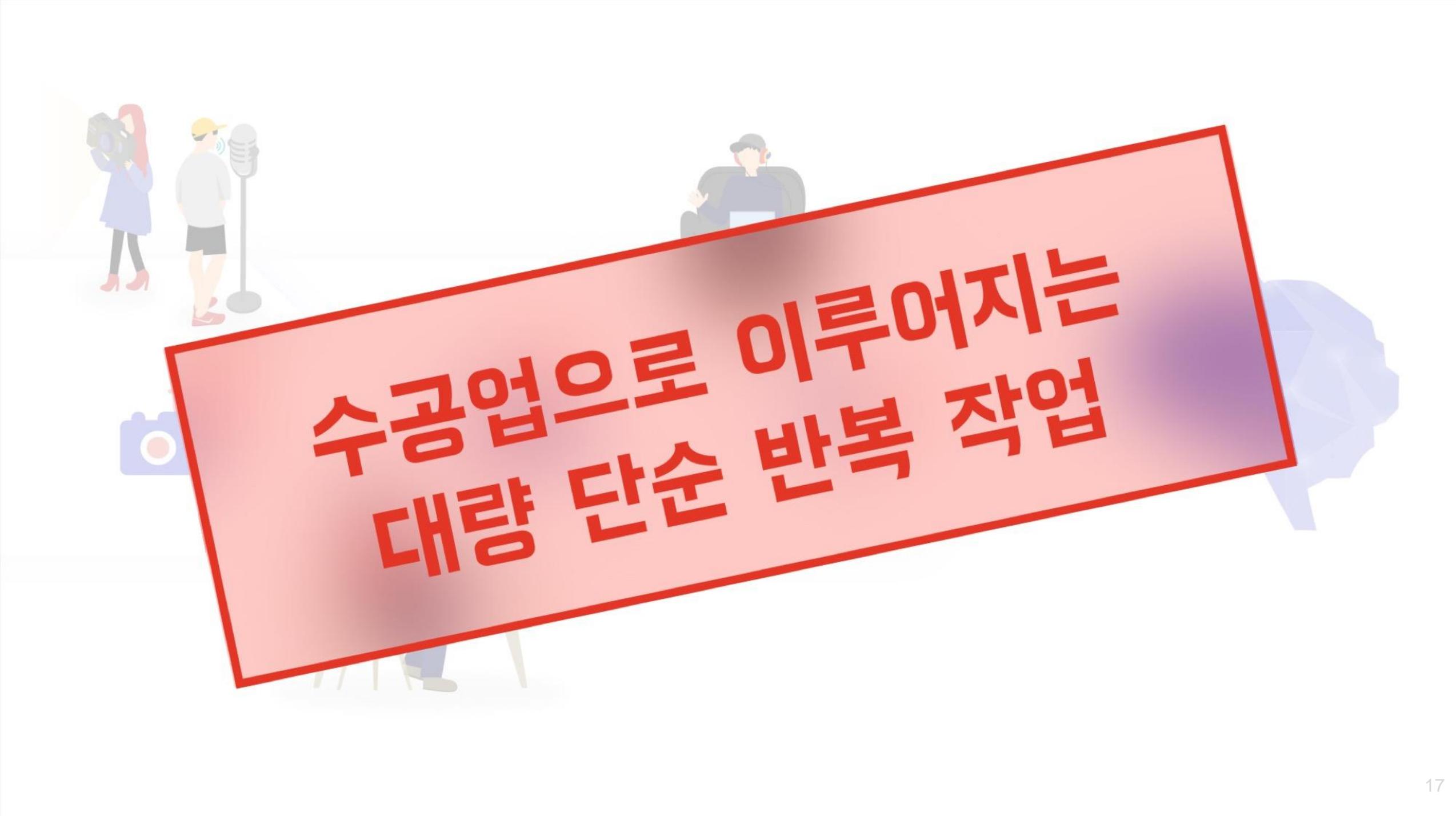
올바른 학습

높은 성능을 갖춘 AI

학습데이터 생산과정



3. 가공된 데이터 모두 품질 검사



수공업으로 이루어지는
대량 단순 반복 작업



금전

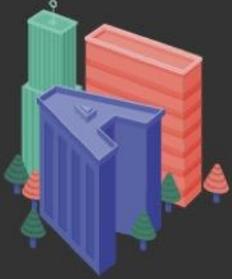


인력



시간



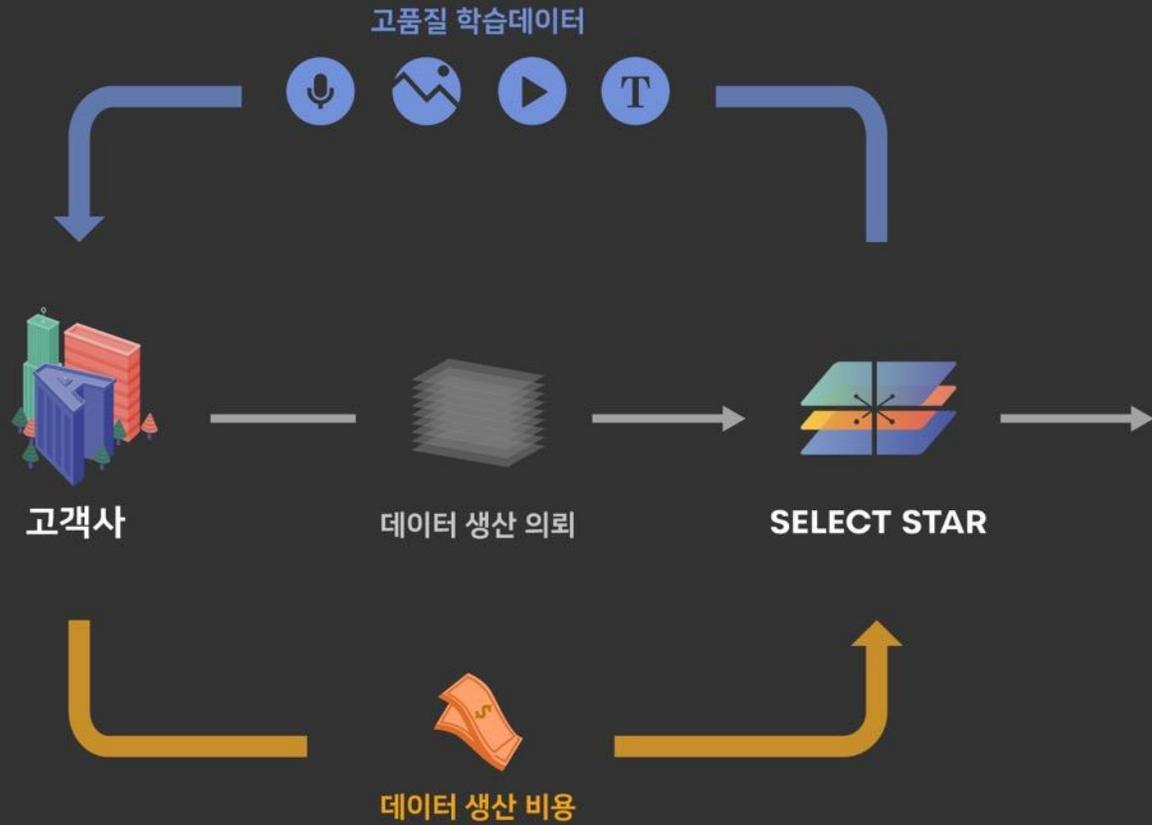


고객사

수많은 모바일 사용자를 통한 데이터 수집 및 가공 서비스



모바일 사용자

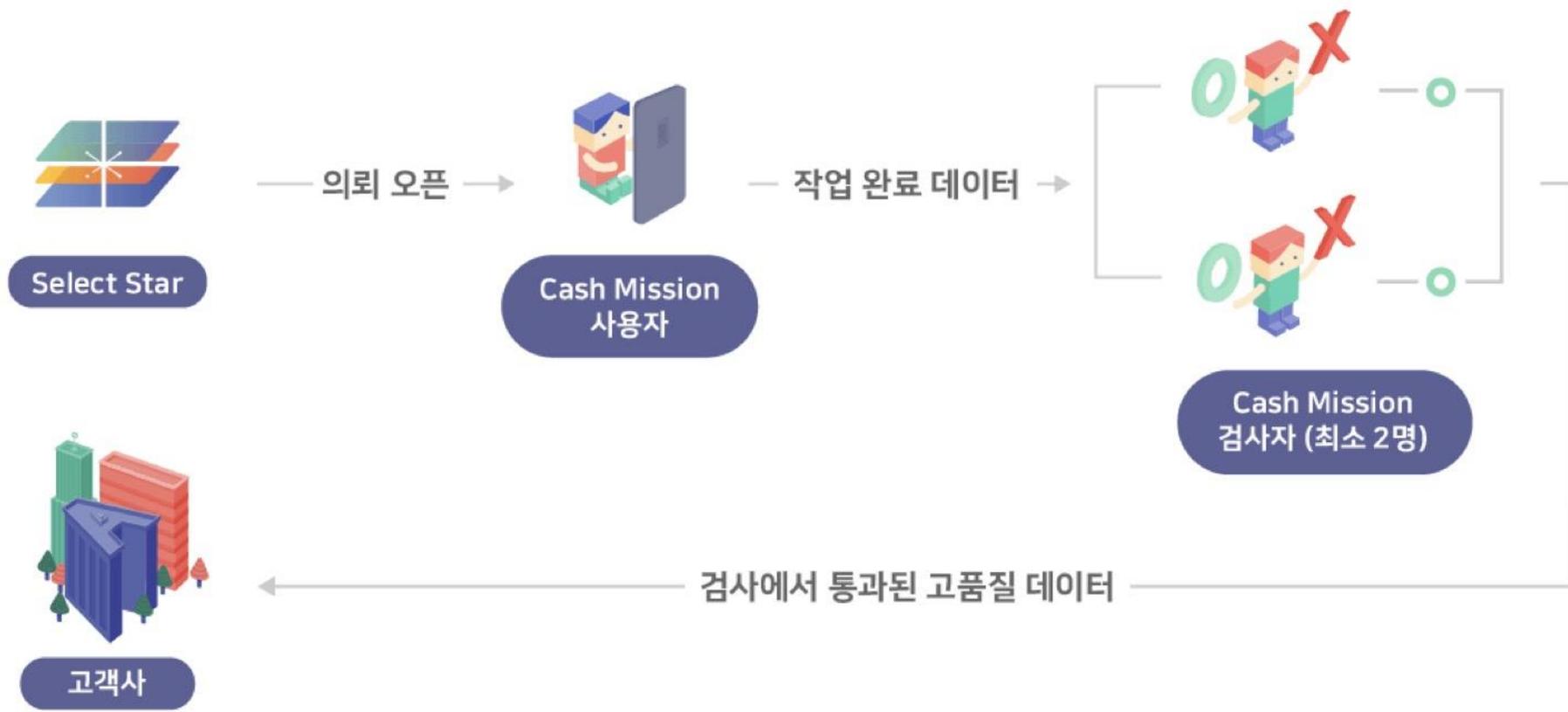


데이터를 원하는 기업/연구자
양질의 데이터 | 빠르고 정확하게

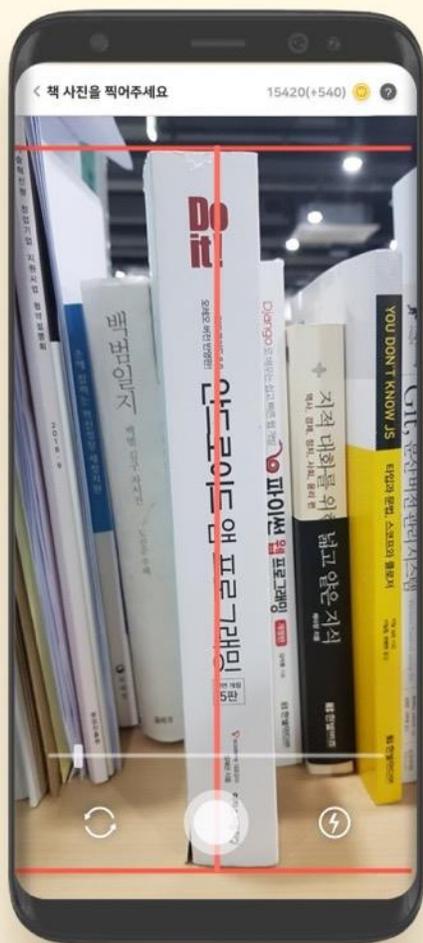
일반 유저
매력 있는 보상 | 언제 어디서든 투잡 가능



이중 검수 시스템으로 데이터 품질을 보장합니다



이미지/영상 데이터



촬영 수집



Classification

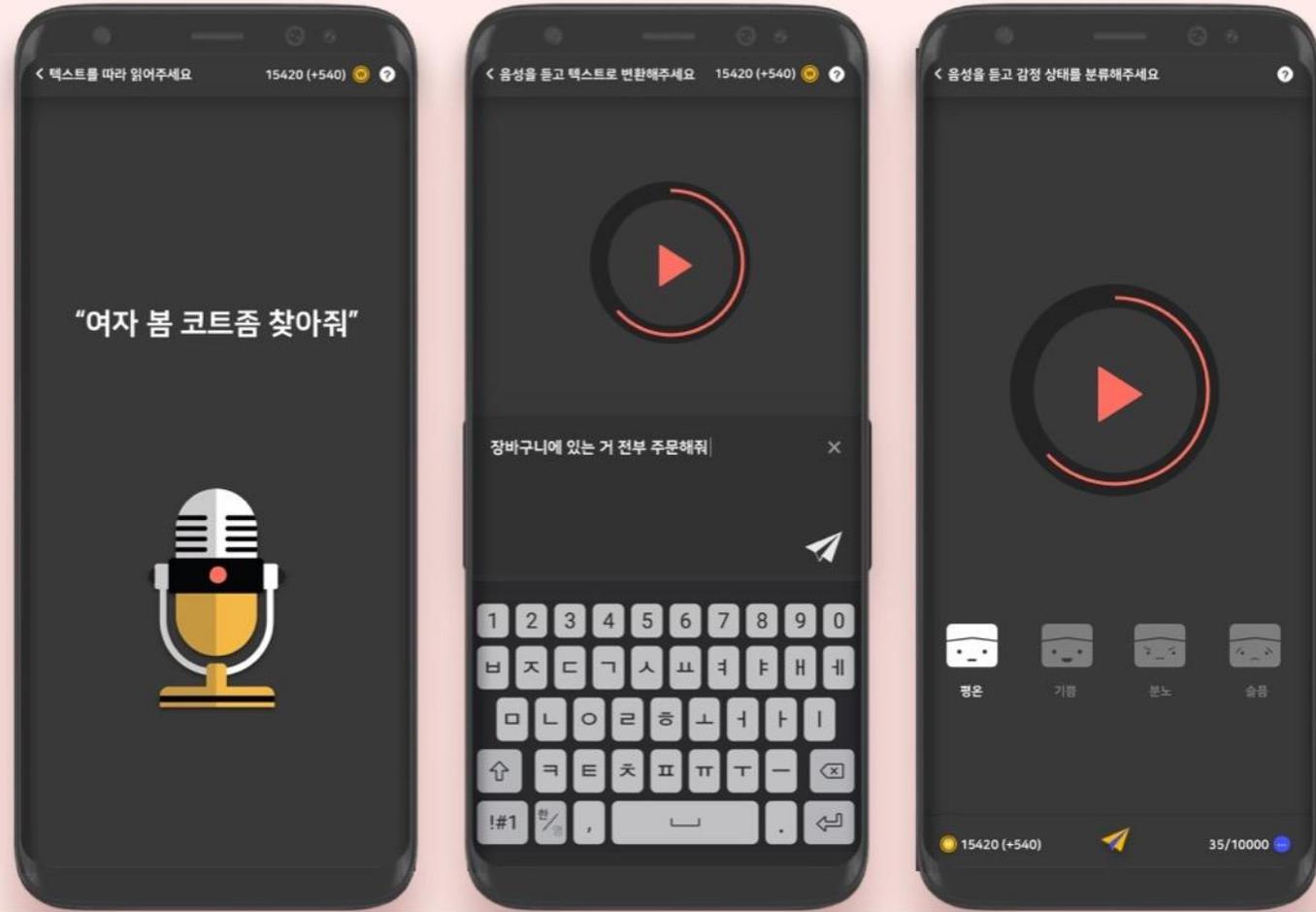


Bounding Box



Instance Segmentation

음성 데이터



녹음 수집

Transcription

Classification

텍스트 데이터



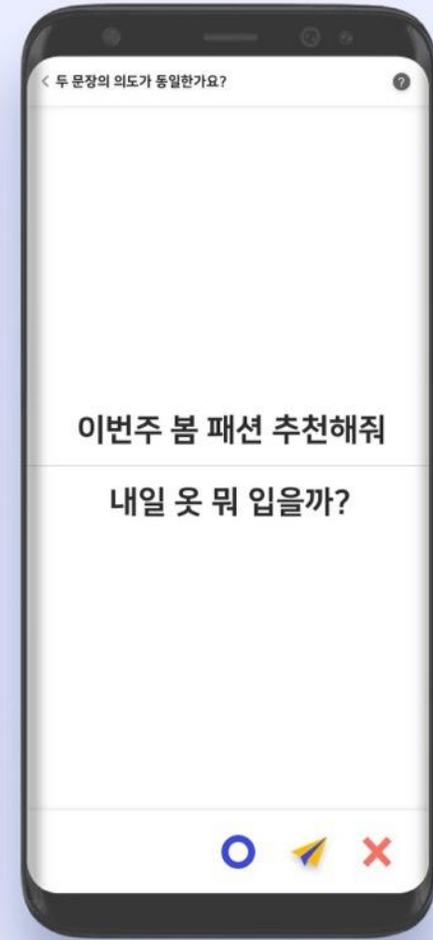
상황별 텍스트 수집



OCR Transcription



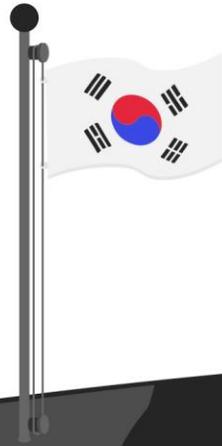
Classification



etc



빠르고 다양한 데이터 생산



국내 사용자
7천명

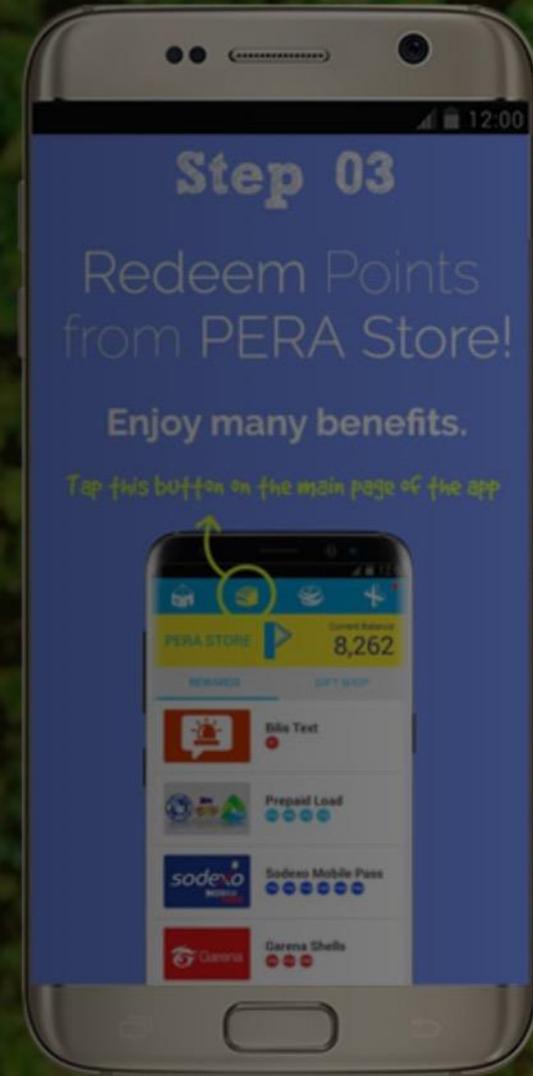


필리핀 사용자
1.7만명



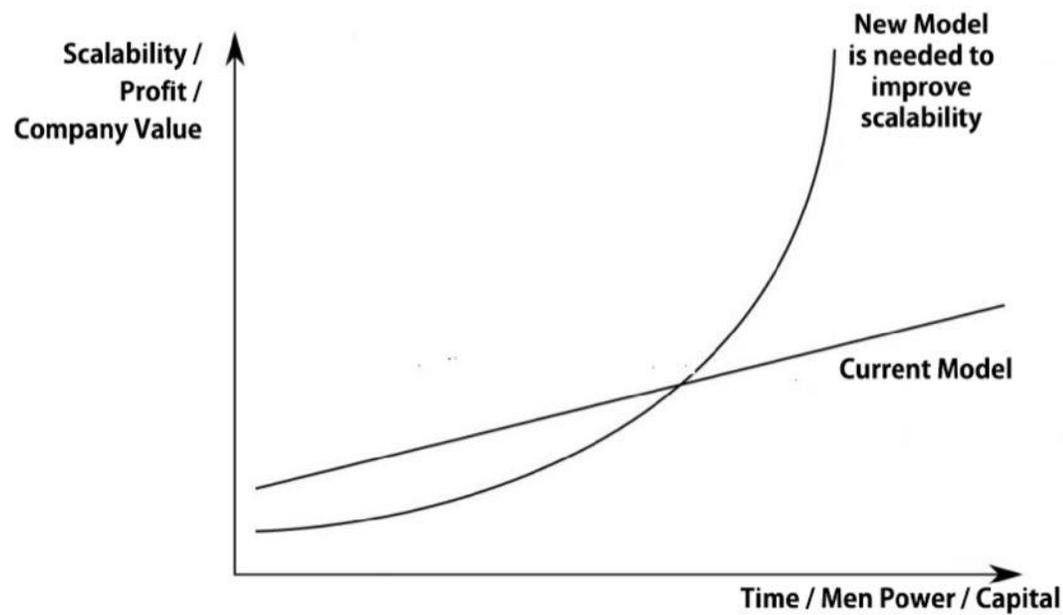
200만 필리핀 사용자를 보유한
Pera Swipe의 파트너십.

Pera Swipe의 잠금화면 →
Select Star 플랫폼으로 접속.



Human-in-the-Loop 방식의 한계

모바일 크라우드소싱으로 한번에 많은 사용자가 참여하여
전체 작업 시간은 줄었지만 개개인의 레이블링 작업 시간은 여전히 수작업으로,
투입되는 인원/시간 대비 아웃풋이 비례함.

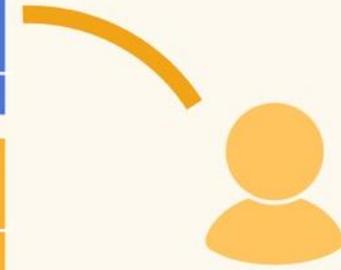


**생산성 및 작업 효율을 획기적으로 높일
기술 개발 필요성에 직면!**

반자동 레이블링 시스템으로 쉽고 빠른 데이터 가공



반자동 AI
가공 시스템



사용자
수정작업



초기 데이터 레이블링



자율주행용



얼굴인식용



일반용

모델 선택 및 Transfer Learning



반자동 AI
가공 시스템



사용자
수정작업

재학습 및
성능 향상



초기 데이터 레이블링



자율주행용



얼굴인식용

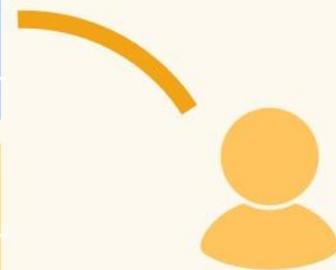


일반용

모델 선택 및 Transfer Learning



반자동 AI
가공 시스템



사용자
수정작업

Active Learning

재학습 및
성능 향상



초기 데이터 레이블링



자율주행용



얼굴인식용



일반용

모델 선택 및 Transfer Learning



반자동 AI
가공 시스템

작업 난이도 예측
통한 작업자 배치



사용자
수정작업

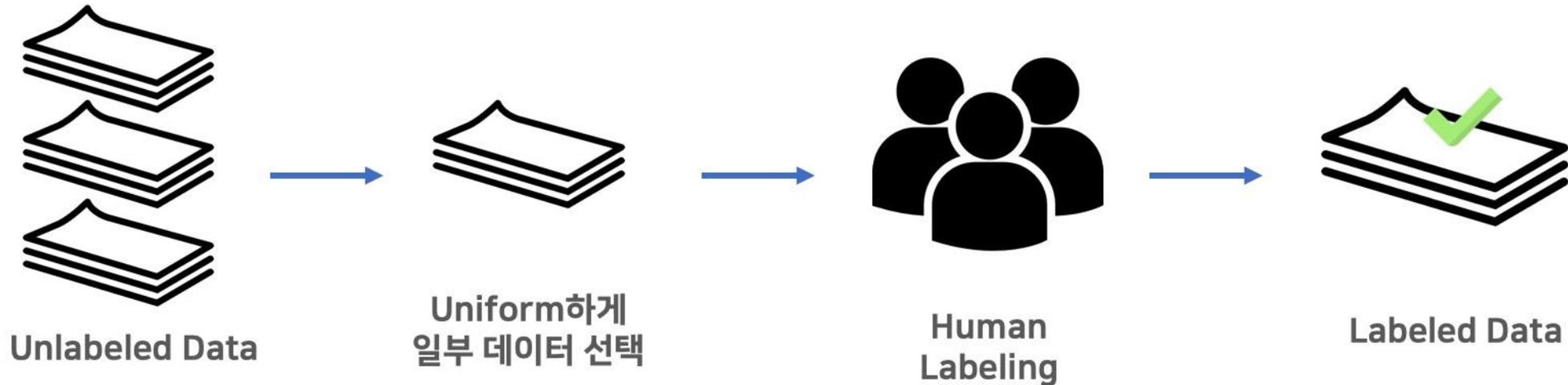
재학습 및
성능 향상

Active Learning

반자동 레이블링 시스템 : Active Learning

학습에 가장 도움이 되는 데이터부터
레이블링하면서 학습하는 방법론

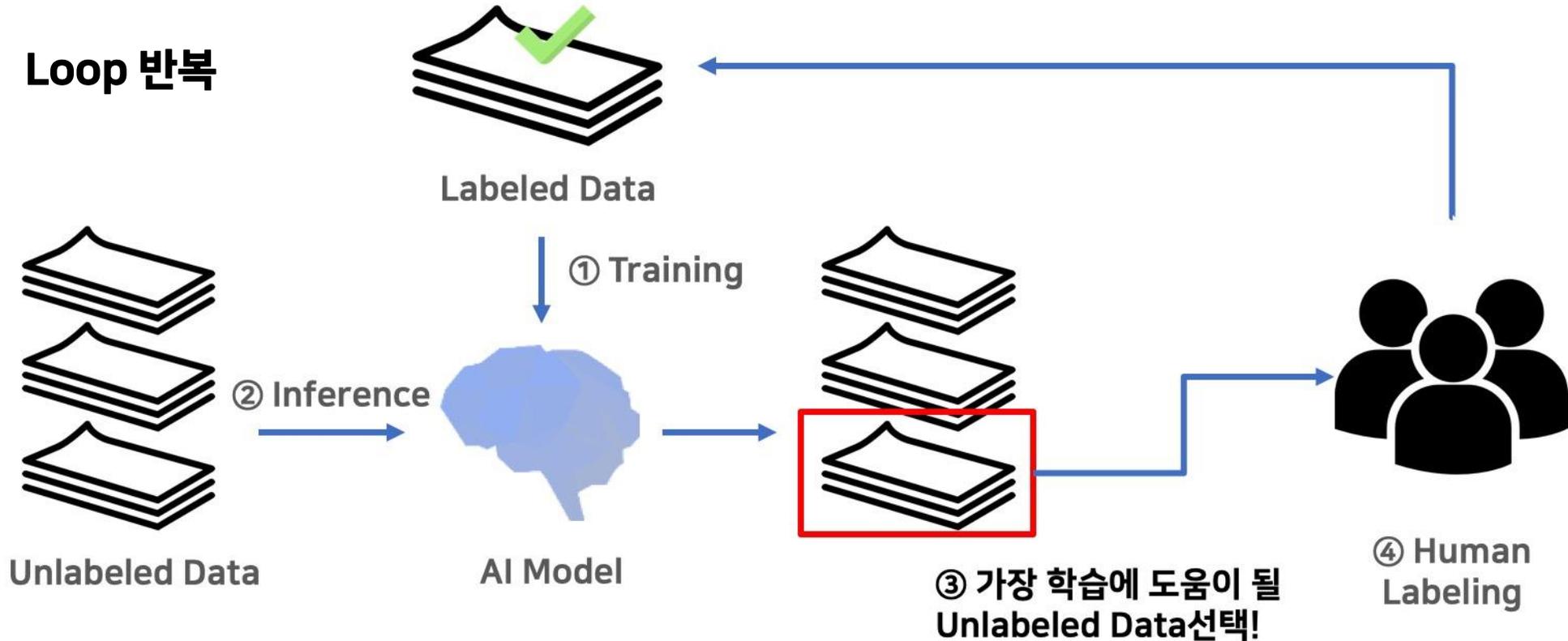
초기 세팅



반자동 레이블링 시스템 : Active Learning

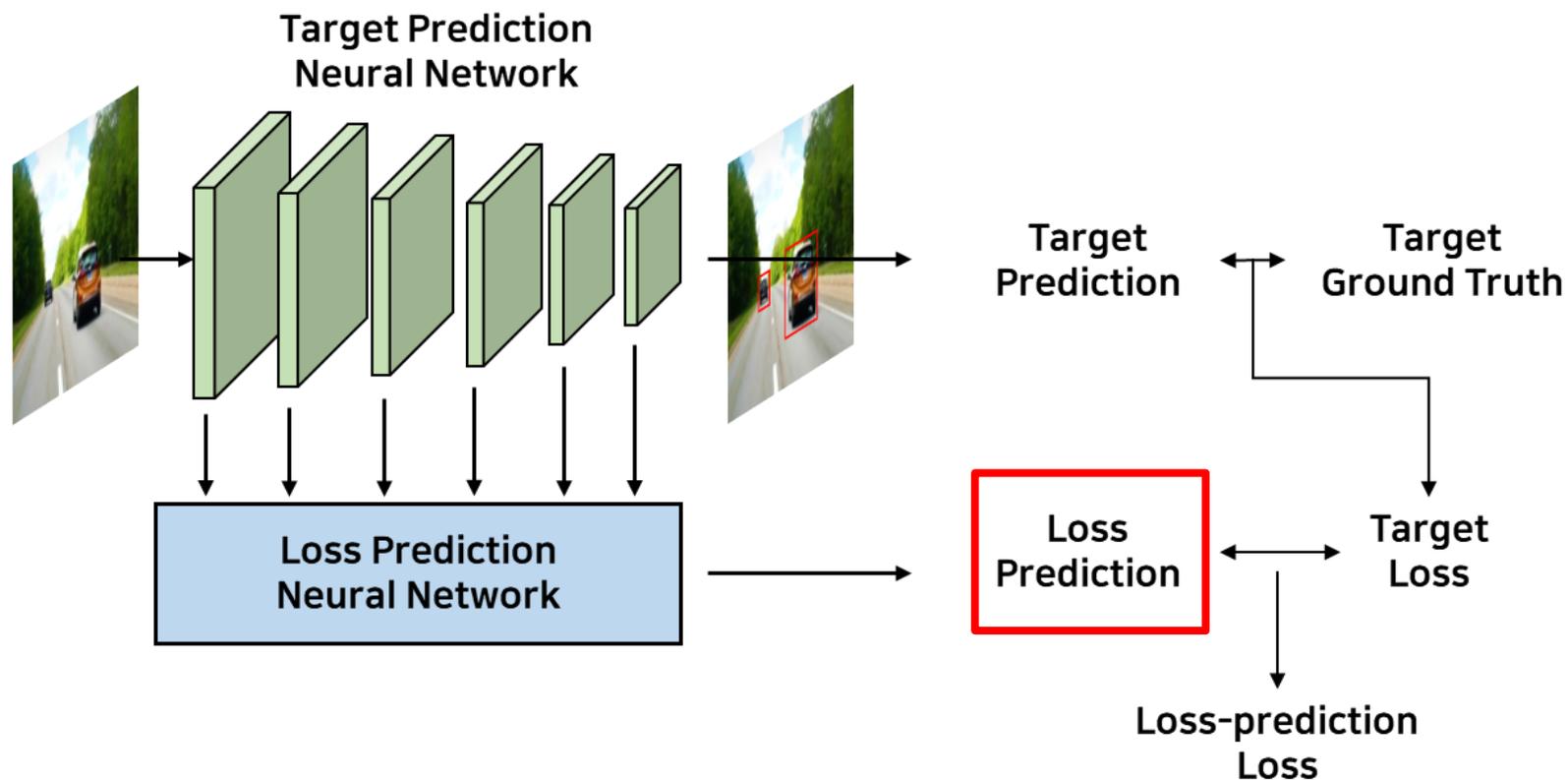
학습에 가장 도움이 되는 데이터부터
레이블링하면서 학습하는 방법론

Loop 반복



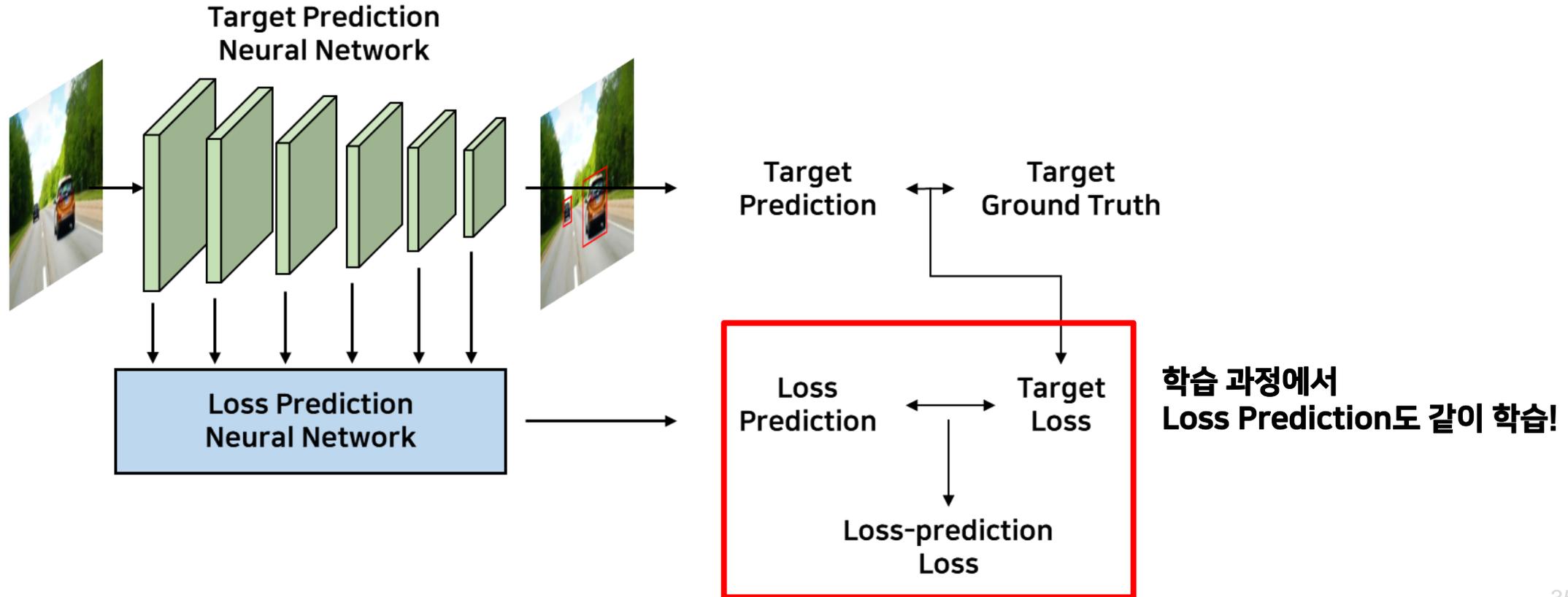
Learning Loss for Active Learning (Yoo, D., & Kweon, I. S., CVPR 2019)

학습에 가장 도움이 될 레이블링 대상을
Unlabeled Data의 Training Loss를 예측하여 선택하자!



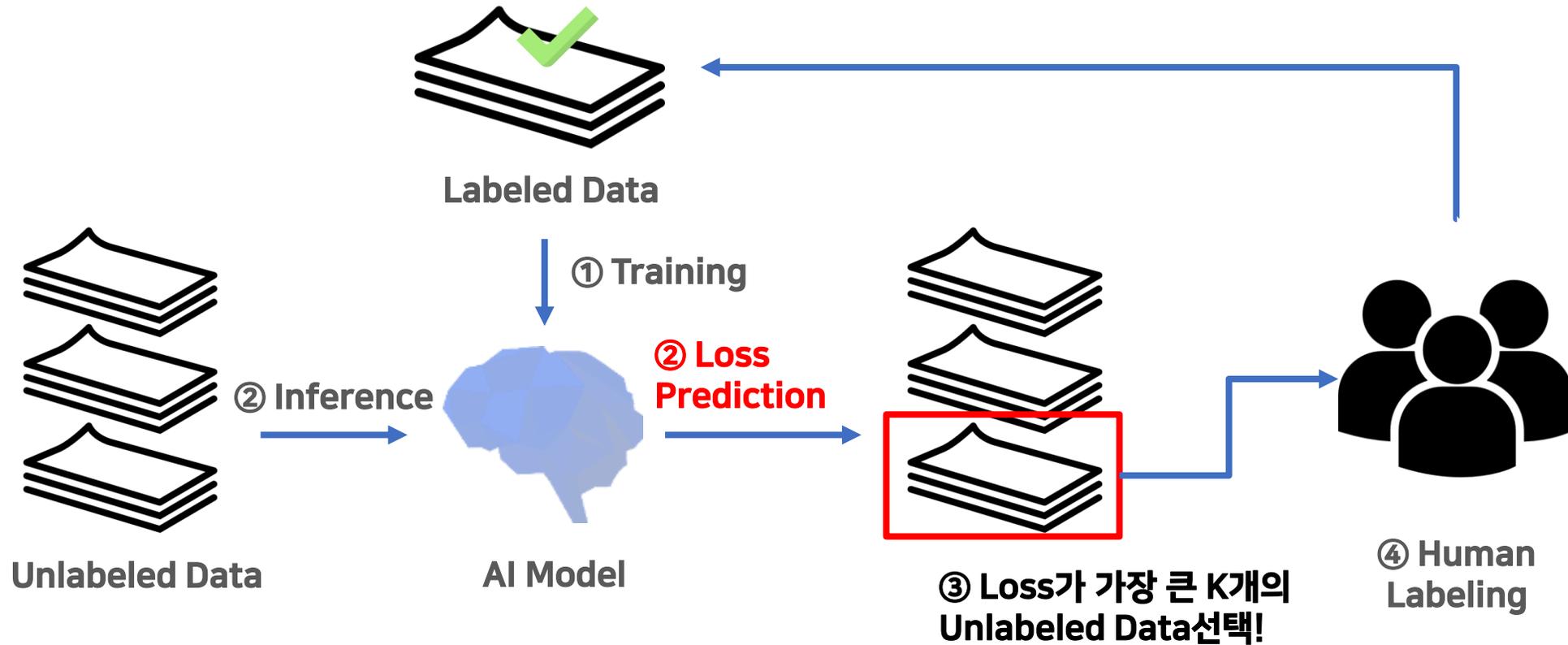
Learning Loss for Active Learning (Yoo, D., & Kweon, I. S., CVPR 2019)

학습에 가장 도움이 될 레이블링 대상을
Unlabeled Data의 Training Loss를 예측하여 선택하자!



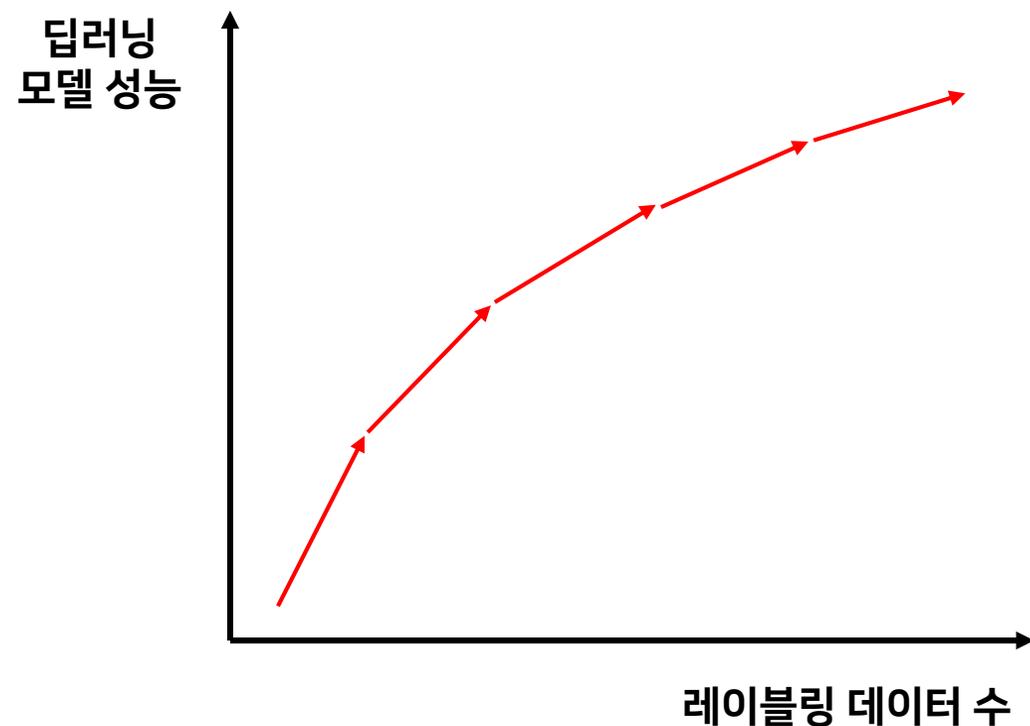
Learning Loss for Active Learning (Yoo, D., & Kweon, I. S., CVPR 2019)

학습에 가장 도움이 될 레이블링 대상을
Unlabeled Data의 Training Loss를 예측하여 선택하자!



Learning Loss for Active Learning (Yoo, D., & Kweon, I. S., CVPR 2019)

학습에 가장 도움이 될 레이블링 대상을
Unlabeled Data의 Training Loss를 예측하여 선택하자!



레이블링이 진행될수록
모델 성능과 작업 효율이 향상!



초기 데이터 레이블링



자율주행용



얼굴인식용



일반용

모델 선택 및 Transfer Learning



반자동 AI
가공 시스템

작업 난이도 예측
통한 작업자 배치



사용자
수정작업

재학습 및
성능 향상

Active Learning

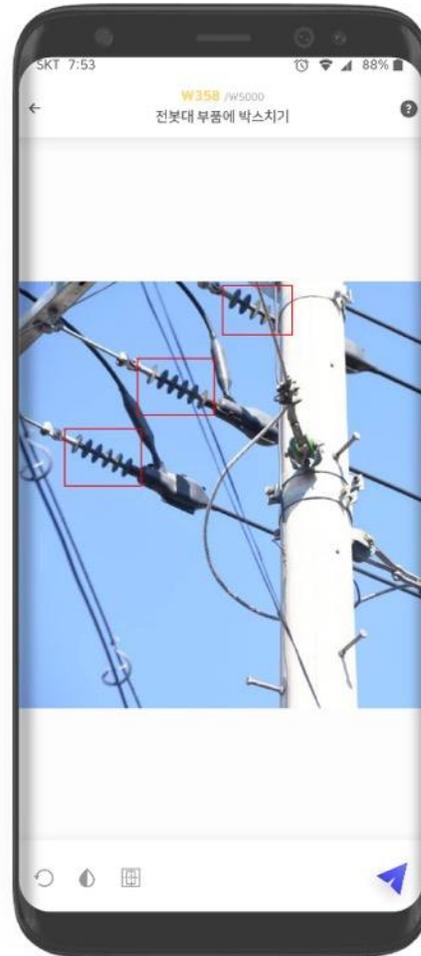
반자동 레이블링 시스템



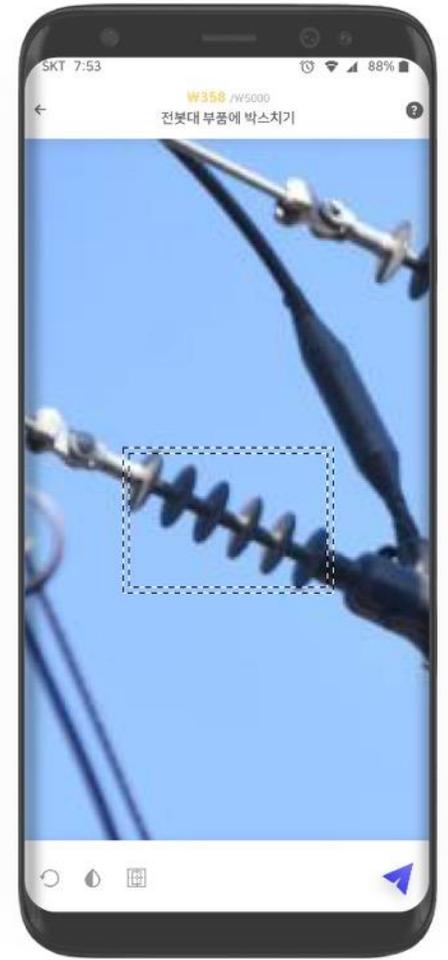
작업자 레벨에 따라
참여 가능 난이도 조절



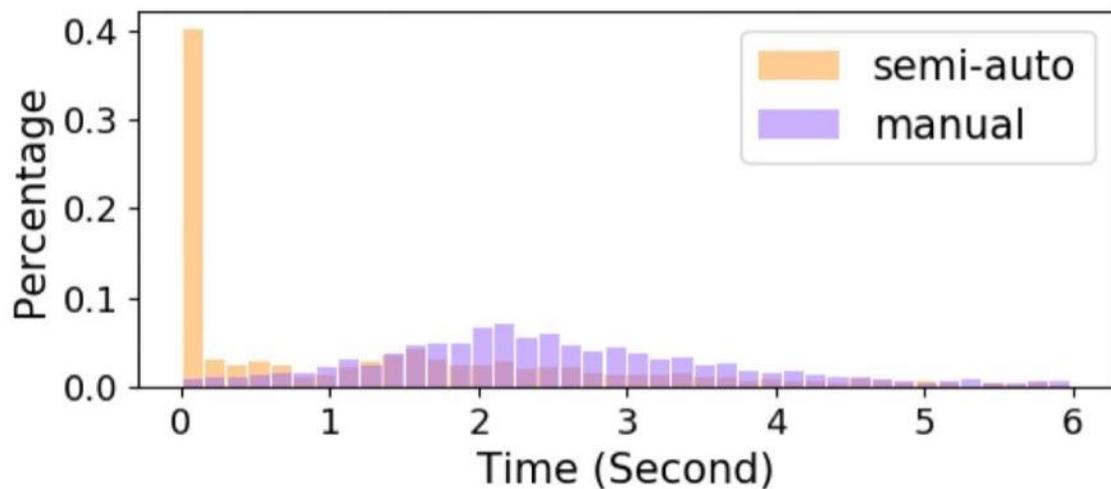
작업 난이도에 따라
작업 보상 차등



모델 예측된
BBox 선택 후
수정 제출

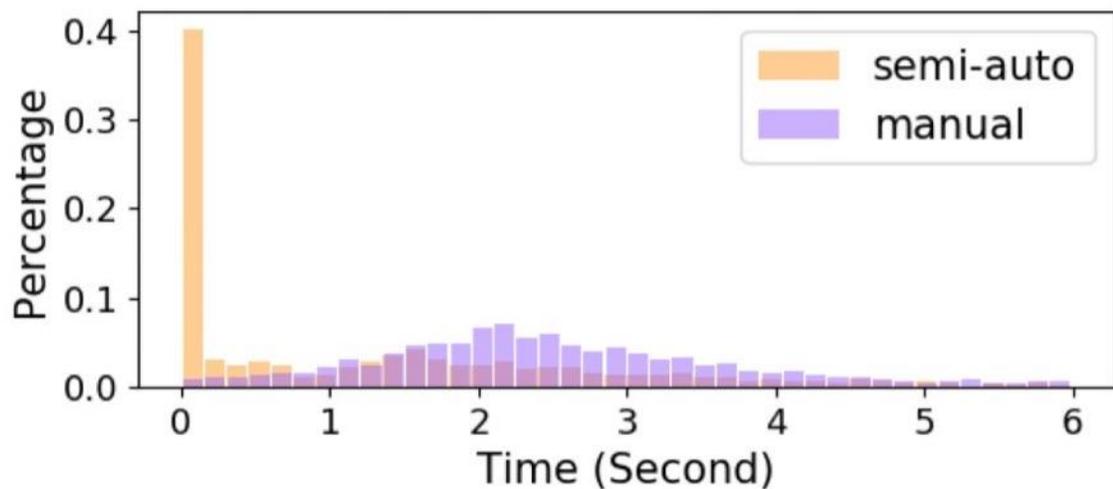


UC Berkeley에서 Deep Drive Dataset 구축 시
Fast-RCNN 모델 기반으로 반자동 레이블링 시도.



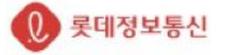
수동 레이블링 대비
60% 작업 시간 절감

UC Berkeley에서 Deep Drive Dataset 구축 시
Fast-RCNN 모델 기반으로 반자동 레이블링 시도.



수동 레이블링 대비
60% 작업 시간 절감
개발 1차 목표!

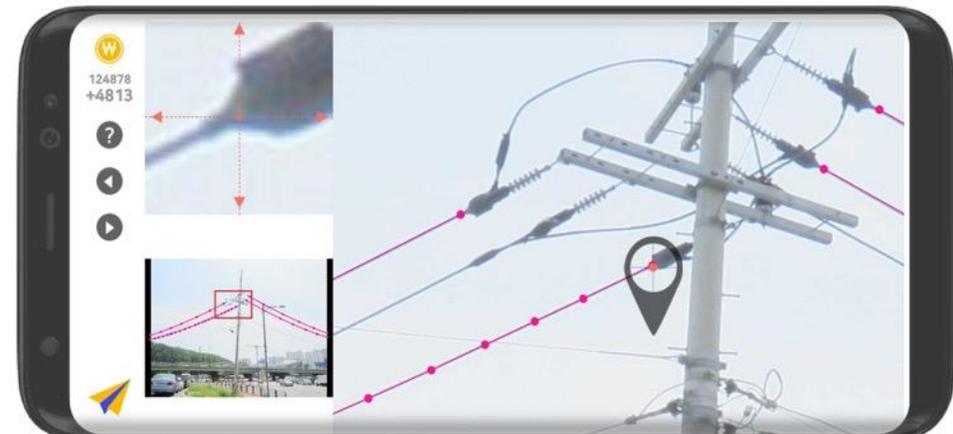
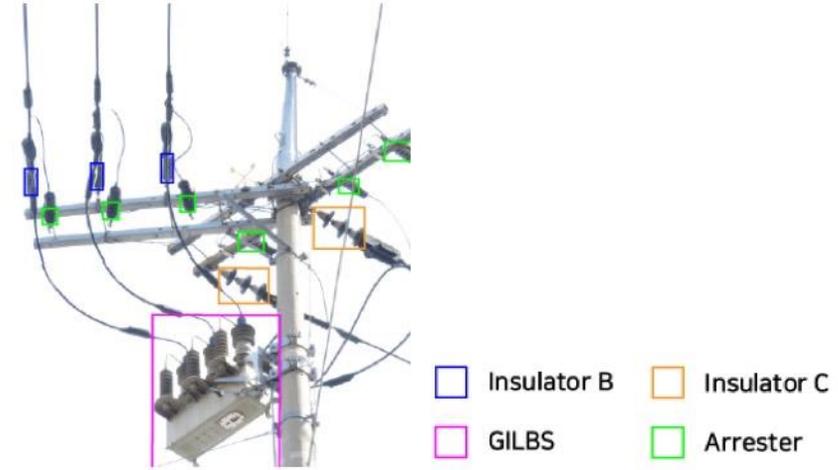
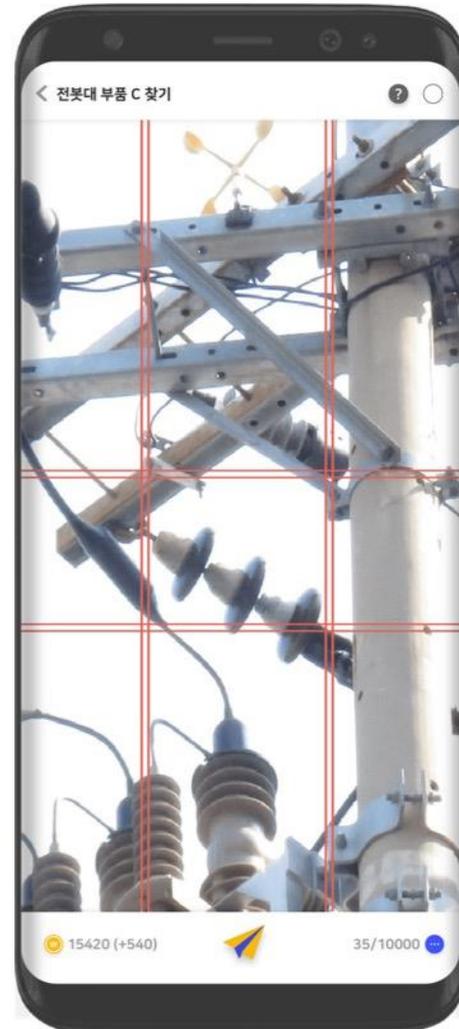
설립 후 9개월, 40개 고객사



설립 후 9개월,
40개 고객사

이미지/영상 데이터

한국전력 & KAIST AI 연구센터
전봇대 부품 Bounding Box & Line Segmentation



설립 후 9개월, 40개 고객사

이미지/영상 데이터

한국정보화진흥원
인공지능 학습용 데이터 구축 사업
인도(人道) 보행 영상 수집 (4개월 내 40만장 구축)



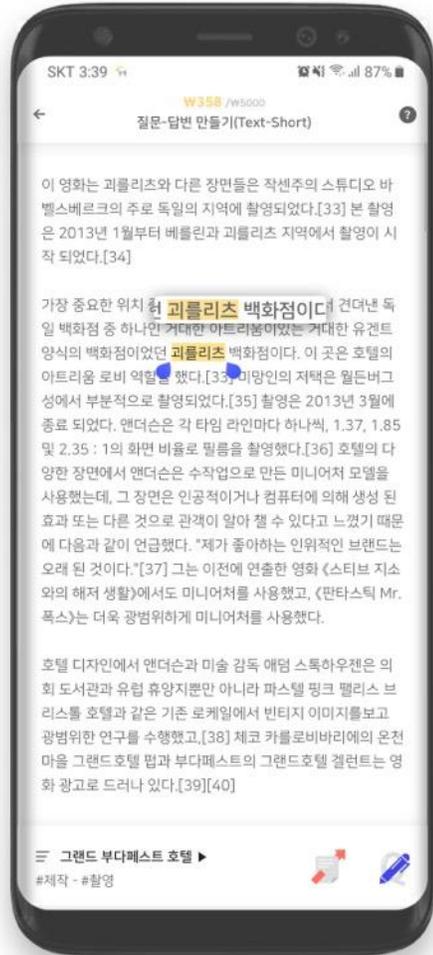
설립 후 9개월, 40개 고객사

텍스트 데이터

LG CNS

한국어 기계독해 질문 & 답변 생성

1400명이 참여하여 KorQuAD 2.0 구축

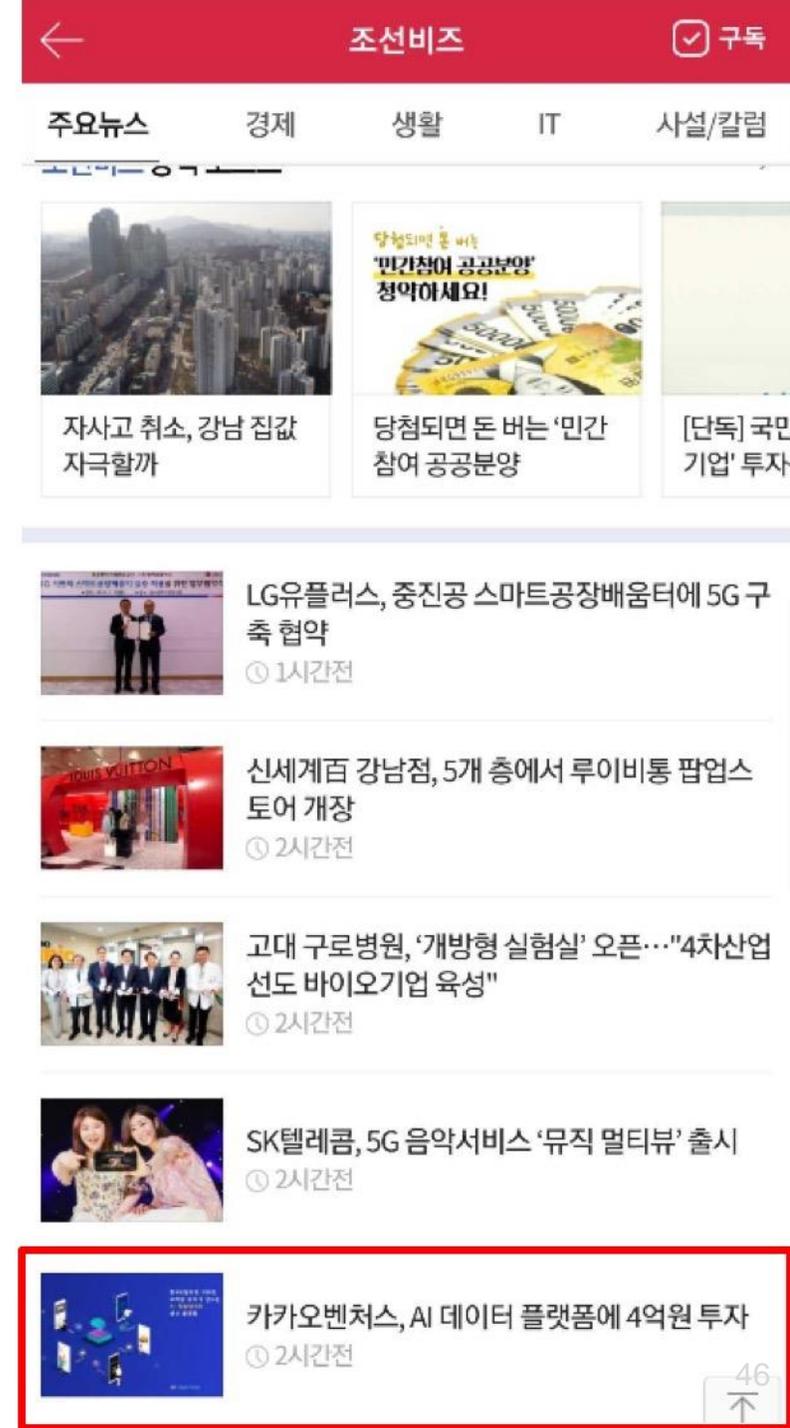


kakaoverventures

카카오벤처스 4억원 투자유치

셀렉트스타는 최근 빠른 사용자 경험 개선 속도와 많은 데이터 생산 실행력을 인정받아 카카오벤처스로부터 4억원의 시드 투자를 유치하였습니다.

셀렉트스타는 앞으로도 다양한 데이터의 대량생산을 위한 유저 풀 확장과 의뢰 수행능력 고도화를 위한 인공지능 어시스턴스 시스템 개발, 데이터 생산 파이프라인 안정화 연구를 통해 최선, 최고의 데이터를 생산하기 위해 끊임없이 노력하겠습니다.



THANK YOU.

채용 문의

신호욱 대표이사
hawk@selectstar.ai

서비스 문의

황민영 영업이사
michael@selectstar.ai