

인공지능 데이터 품질 및 이슈

한국정보화진흥원 AI데이터추진단

김민준



CONTENTS

01 인공지능 데이터 개요

02 데이터 품질 관리 현황 및 이슈

03 품질 관리 개선방안 및 표준화

01

인공지능 데이터 개요

AI

AI의 급부상

인공지능이 연구수준을 넘어 범용기술(General Purpose Technology)로 발전함에 따라 기존 산업의 차별적 부가가치 창출을 견인하고, 획기적 변화를 촉발

인공지능 세계시장 규모(Staista, '18) : 73.5억\$('18)→898억\$('25), 2030년까지 글로벌 GDP 13조 달러 기여(맥킨지, '18)

인공지능 기술은 한국에서도 이미 제조업(자율주행차, 지능형 로봇, 스마트 공장 등) 서비스업(의료, 교육, 금융, 게임, 체육 등)과 융합되어 상용화 시작

의료 루닛, 뷰노, 딥노이드 등

교육 루이드, 노리, 바로풀기 등

법률 로엔비, 로메니저 등

체육 APK어플킹, KT 등

게임 엔씨소프트, 넷마블게임즈 등

AI스타트업에 대한 투자 증대, M&A 확대 등 인공지능 생태계가 구축되기 시작하면서 데이터, 플랫폼 등 AI자원 확보의 중요성 확대

국내 인공지능 분야 VC투자액은 '17년 216억원에서 '18년 1,223억원 약 5.6배 증가했으며, 해외 AI 스타트업 M&A도 '17년 115건인 것으로 조사(CBInsights, '18)

데이터 왜 중요한가?

데이터 역량 = 새로운 글로벌 국가 경쟁력

데이터가 경쟁력

- 대규모 데이터를 확보하는 글로벌 ICT기업 시장 주도



AI도 핵심은 데이터

- 머신러닝 기반 AI는 방대한 양질의 학습용 데이터가 성능을 좌우
- AI 경쟁력도 대량의 도메인별 데이터 확보 여부에 따라 결정

융합 및 혁신서비스의 밑거름

- 개인정보 활용 및 데이터 결합을 통한 혁신적인 서비스 및 제품 생산 가능
- 특히 방대한 데이터 축적으로 파급효과가 큰 의료·금융 분야와 타 산업분야 융합을 통해 혁신 성장의 견인차 역할

의료



개인맞춤형 건강 도우미,
AI닥터(닥터앤서) 등 활용

금융



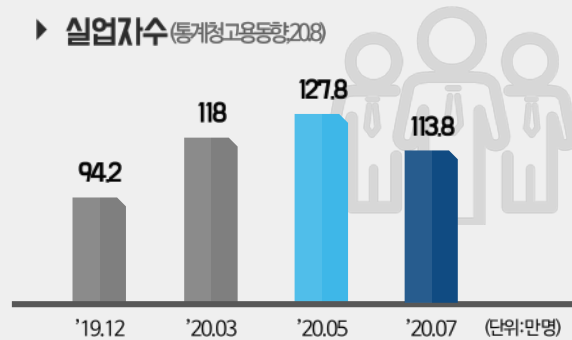
비금융데이터 신용평가,
보험금자동청구 등 활용

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

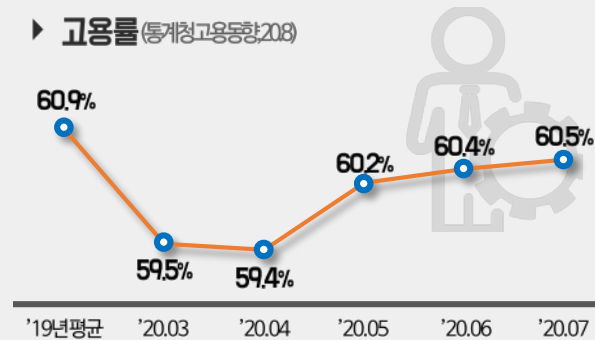
현 경제상황

□ 소폭 회복세이나, 불확실성은 여전

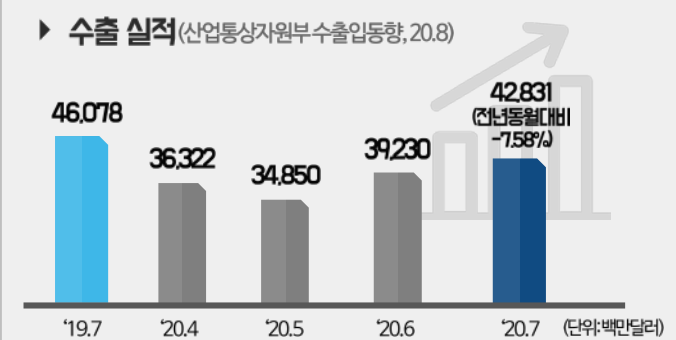
▶ 실업자수 (통계청고용동향, 20.8)



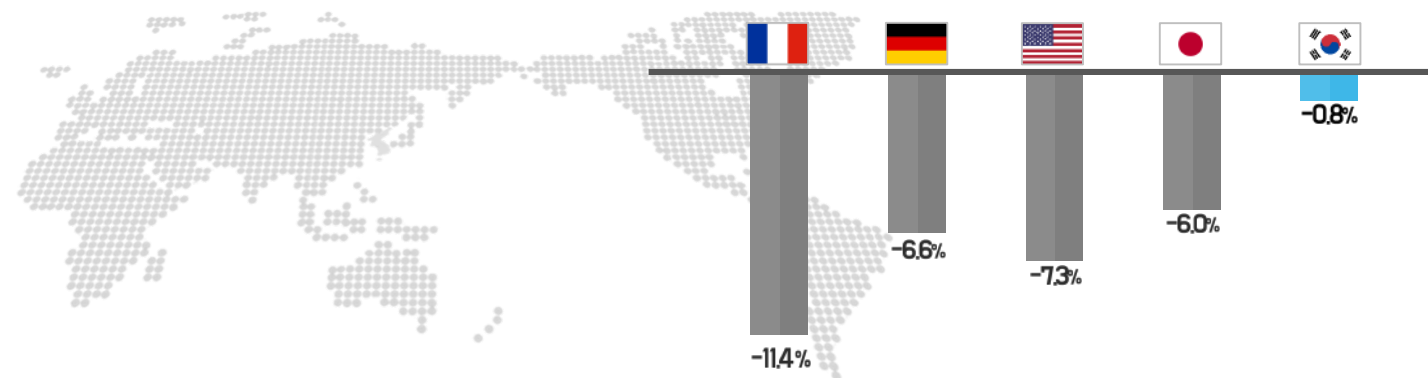
▶ 고용률 (통계청고용동향, 20.8)



▶ 수출 실적 (산업통상자원부수출입동향, 20.8)



□ OECD는 우리나라 경제성장을 1위(-0.8%) 전망('20.8)



위기를 기회로 : 한국형 뉴딜

실물경제 회복을 위한 **정부의 직접적인 투자 확대가 불가피**

회복탄력성
확보

팬데믹 하에서의
최소한의 삶 보장

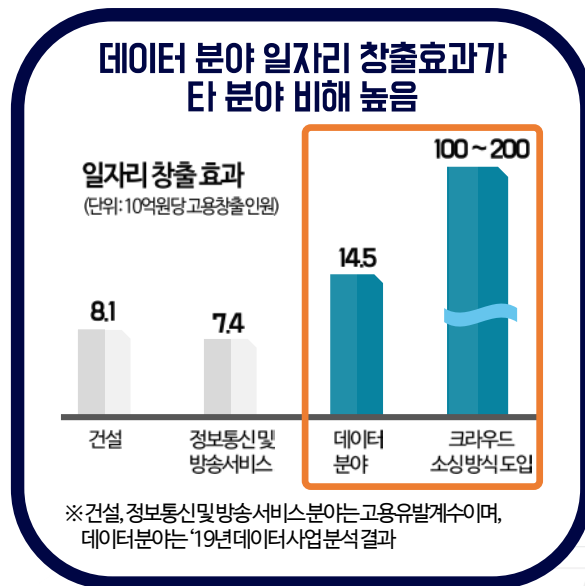
경기부양

경기침체와 마이너스 성장,
일상의 봉쇄, 고용쇼크 등
국내외 경제위기 극복

격차해소

소득격차, 교육격차,
디지털 격차 등
사회 제반 영역 격차해소

디지털 뉴딜 - 데이터 댐 기획 배경



**AI · 데이터 중심의
디지털 뉴딜**



디지털 뉴딜 주요내용

① D.N.A. 생태계 강화

(~'25) 31.9조원 투자, 56.7만개 일자리 창출

- 국민생활과 밀접한 분야데이터 구축·개방·활용
- 1·2·3차 산업으로 5G·AI 융합 확산
- 5G·AI 기반 지능형 정부
- K-사이버 방역체계 구축

② 교육 인프라 디지털 전환

(~'25) 0.8조원 투자, 0.9만개 일자리 창출

- 모든 초중고에 디지털 기반 교육 인프라
- 전국 대학·직업훈련기관 온라인 교육 강화

③ 비대면 산업 육성

(~'25) 2.1조원 투자, 13.4만개 일자리 창출

- 스마트 의료 및 돌봄 인프라 구축
- 중소기업 원격근무 확산
- 소상공인 온라인 비즈니스 지원

④ SOC 디지털화

(~'25) 10조원 투자, 19.3만개 일자리 창출

- 4대 분야 핵심 인프라 디지털 관리체계 구축
- 도시·산단의 공간 디지털 혁신
- 스마트 물류체계 구축

NIA 디지털 뉴딜 주요사업

디지털 뉴딜

D.N.A. 생태계 강화

- 양자암호통신 인프라 구축
- 5G 기반 정부 업무망 고도화
- 5G 공공선도 서비스 구축
- 클라우드 컴퓨팅산업 육성
공공 클라우드 플래그십 프로젝트
공적마스크 판매정보 서비스 공급지원
- 행정공공기관 정보시스템 클라우드 전면 전환
- AI 학습용 데이터 구축
- 빅데이터 플랫폼 및 네트워크 구축
- 공공데이터 품질관리·구축 신규지원 등
공공데이터 개방 및 품질개선 가속화
공공빅데이터 청년인턴십

교육인프라디지털전환

- 모든 학교 무선환경 구축

SOC 디지털화

- IoT·AI 기반 新데이터댐 구축

안전망 강화

사람 투자

- 공공와이파이 품질 고도화
- 무선인터넷 인프라 확대 구축
- 농어촌 등 통신망 고도화
- 전국민 디지털 역량강화 교육

'20년 추경 **8,318억원**

'21년 예정 **1조 2,260억원**

'22년 예정 **1조 4,858억원**

AI 학습용 데이터 구축 사업

AI 제품·서비스 및 기술 개발에 활용가치가 높은
대규모 **AI 학습용 데이터 구축 및 개방, 응용 개발**



AI 학습용 데이터 구축 사업

사업 기본방향

AI 기술 트렌드와 지역(지자체) 시장 수요 반영

- 국내 특유 상황을 반영하고, 산업적·사회적 수요가 높으나, 구축이 어려워 정부주도 구축이 필요한 지정과제 데이터
- 사회현안 및 데이터3법 등 변화하는 시장수요 반영을 위한 자유과제 데이터 지역 특유의 특성과 수요를 반영한 지역분야 데이터

AI 학습용 데이터 활용성 강화

- AI 응용서비스 개발 역량 보유 기업·기관의 컨소시엄 참여
- AI 학습용 데이터 구축용 저작도구(어노테이션 툴 등)의 개발 및 소스코드 공개
- 데이터 활용 매뉴얼, 활용 예시 공개

클라우드 소싱 기반 데이터 구축과 일자리 창출

- 언제 어디서든 누구나 데이터 수집, 가공 등에 참여할 수 있는 클라우드 소싱 방식을 적극 도입하여 일자리 창출 효과 극대화

AI 허브('18~)

AI 기술 및 제품 서비스 개발에 필수적인 AI데이터, AI알고리즘, 고성능 컴퓨팅 자원을
국내 중소벤처, 스타트업이 쉽게 활용할 수 있도록 제공하는 **원스톱 플랫폼**



<https://www.aihub.or.kr>

AI 데이터 구축·개방

- 시급성·유망성이
높은 분야 발굴
→ AI 학습용 데이터
구축·개방
- 수어 영상, 대용량
동영상 콘텐츠 등
20개 분야 수행기관
선정('20. 5.)

AI 알고리즘

- 정부 AI R&D
결과물(AI모델)
→ AI 서비스용 open
API로 제공
- AI 허브 내 AI API
사용 신청을 통해
상시 제공 운영 중

고성능 컴퓨팅 자원

- GPU 기반 고성능
컴퓨팅 자원
→ 중소·벤처,
스타트업에 제공
- 자원 신청은 매년 초
접수를 받아 선정결과
별도 통보

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

[참고] AI 허브 데이터 개방 내역('17~'19년)

- '17~'18년 구축 데이터 11종



헬스케어

복내장, 당뇨망막증 등 주요 안과질환의
안저 촬영 이미지와 전문의 소견을 결합한 데이터



한국어 음성

2~3명이 서로 자연스럽게 연속하여 발화하는
한국어 음성과 발화 내용을 전사한 데이터



한국어 대화

음식점, 소매점, 학원 등 소상공인 업종에서 자주 발생하는 상황과
질문·답변·의도 등을 구조화한 대화 데이터



멀티 모달

영상 내에 사람의 감정, 행동, 상황, 대화 내용 등의 멀티모달정보를
결합시킨 동영상 데이터



기계 독해

다양한 질문과 정답, 정답을 포함하고 있는
지문의 쌍으로 구성된 지능형 QA용 텍스트 데이터

- '19년 구축 데이터 10종



한국어-영어 번역 말뭉치

한국어 인공지능 번역 기술개발 및 성능강화를 위한
한국어-영어 병렬 말뭉치 데이터 150만 문장



한국형 사물 이미지

한국형 객체, 장소, 상황 인식기술 개발 및 성능
강화를 위한 사물/거리 등 이미지 데이터 360만장



한국어 글자체 이미지

한글 광학문자인식(OCR) 성능개선을 위한 한글
글자체(손글씨 및 인쇄체) 이미지 500만장



인도 보행 영상

보행지원기술 개발을 위한 인도 보행 및 인도 위
객체(사람, 자전거, 펜스 등) 라벨링 데이터 200시간



멀티모달(융합) 영상

한국어 인공지능 번역 기술개발 및 성능강화를
위한 한국어-영어 병렬 말뭉치 데이터 150만 문장



사람 동작 영상

사람의 동작자세-생동 인식기술 개발을 위해
다양한 조건에서 사람 동작 영상 데이터 20만 클립



한국인 안전 이미지

여러 각도, 조도 등의 환경 하에서 안전인식기술
성능강화를 위한 한국인 얼굴 데이터 1,800만장



위해물품 엑스레이 이미지

위험물·도구 자동인식 기술개발 및 성능개선 위한
법적 도구, 반입금지물품 등의 X-ray 이미지 44만장



질병진단 이미지

국내에서 유행률, 중증도 등이 높은 주요 질환 관련
진단 이미지(초음파 영상 등) 및 결과 데이터 3만장



이상행동 CCTV 영상

주야간 적용가능한 이상행동 지능형 탐지기술개발 및
성능 강화를 위한 CCTV 영상 데이터 700시간

[참고] AI 허브 데이터 개방 내역('20년, 1차)



문서 요약 텍스트 AI 데이터

도서, SNS, 논문 등 분야별 원본 텍스트 데이터 35만건 이상과 원문의 핵심내용을 적절히 요약한 텍스트 데이터



전문분야 한-영 말뭉치 AI데이터

전문분야를 선정하고 분야별로 적정하게 배분한 한국어-영어 병렬 말뭉치 150만 건 이상



대용량 동영상 콘텐츠 AI데이터

다양한 분야별 원천비디오 데이터에 객체, 객체위치, 상황 등이 결합된 학습용 영상 데이터 500시간 이상



랜드마크 이미지 AI데이터

전국 50개 도시의 랜드마크(국내 공공기관, 주요 건물, 관광명소, 편의시설 등) 이미지 총 500만장 이상



딥페이크 변조영상 AI데이터

성별, 나이 등을 균등하게 고려하여 제작된 원본 영상을 기반으로 합성(변조)된 동영상 데이터 20만개 이상



자율주행드론 비행영상 AI데이터

다양한 비행 환경(시간, 장소 등)에서 획득된 영상과 장애물 등의 객체가 레이블링 된 영상 총 300시간 이상



수어(手語)영상 AI데이터

일상생활 및 특정상황에 사용되는 단어(4,000단어 이상) 및 문장(2,000문장 이상)을 표현하는 동영상 20만개



한국인 대화음성 AI데이터

성별, 지역 등 분야별 원본 음성 데이터 4,000시간 이상과 음성을 문자로 변환한 텍스트 데이터 400만 문장 포함



시각정보기반 질의응답 AI데이터

다양한 객체, 상황 등을 표현하는 이미지 115만장 및 질문·답변 데이터 690만개



사람 인체·자세 3D AI데이터

70여개 이상 사람의 행동 및 동작으로 3D 모델 50만건, 2D 이미지 200만장 이상

[참고] AI 허브 데이터 개방 내역('20년, 1차)



질병진단 이미지 시데이터

4,000명에서 추출된 유방암 병리 이미지 10만건,
부비동(DICOM) 데이터 8,000건



도로주행영상 시데이터

실도로 주행데이터 Use-Case 기반 자율주행 원천데이터
수집, 총 학습용 데이터 60만 5천 프레임 이상



도로환경 파노라마 이미지 시데이터

도심지 파노라마(3,400km, 338,910장, 94시간 규모),
평면영상 2,711,280장, 753시간 규모)



치매진단 뇌파영상 시데이터

치매관련 의료 영상데이터와 진단정보를 결합한
시데이터 28만건



피트니스 자세 이미지 시데이터

200,000개의 비디오로 구성하며, 각 운동 종목당 최소
2,400개의 비디오 포함



감성 대화 말뭉치 시데이터

우울증 등 심리 장애로 인한 사회문제 해결을 위한 감성대화
코퍼스 데이터 270,000문장, 발화음성 10,000건



K-Fashion 이미지 시데이터

저작권 문제가 해결된 패션 이미지의 패션 요소 정보를
어노테이션한 이미지 120만장 이상



위성영상 객체판독 시데이터

광학 및 레이다 영상을 이용한 5종의 시데이터 120만건 구축



한국인 재식별 이미지 시데이터

대한민국의 실내/외 구축된 공공 CCTV 환경을 고려한
한국인(1,000명) 재식별 데이터셋 400만장



사람 인체·자세 3D 시데이터

치아 및 치주질환 진단과 치료계획 수립을 위한 파노라마
5,000장(5,000명), CBCT 200,000장(2,000명)

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

[참고] AI 허브 데이터 개방 내역('20년, 2차)

10개 영역, 72개 과제 (주제지정45개, 지역6개, 자유21개). **총 150종 구축하여 AI 허브에 개방**



자연어(97개)

음성인식 고도화 등을 위한 음성 및 말뭉치데이터 등 32종



미디어(47개)

영상내 음성·글자 인식, 야외소음 환경, 스포츠(축구, 농구, 골프) 사람동작데이터 등 11종



헬스케어(97개)

암, 치매, 피부질환 진단을 위한 의료 영상데이터, 진료 및 건강데이터 등 20종



안전(6개)

안전감시 및 안전활동 자동화를 위한 영상데이터, 노후시설물 이미지데이터 등 16종



자율주행(5개)

자율주행기술 발전을 위한 차선·횡단보도, 신호등·도로표지판데이터 등 14종



기타(5개)

소상공인, 제조 분야의 업무 자동화 및 비용 절감을 위한 영상, 이미지데이터 등 8종



농축수산(4개)

농·수산물 지능형 로봇 개발과 작물 질병·해충 진단을 위한 영상 및 이미지데이터 등 12종



지역분야(6개)

지역(지자체) 수요를 반영한 자유주제 AI 학습용 데이터 선정 및 구축 추진



국토환경(3개)

환경 오염 모니터링과 기후 변화예측을 위한 이미지데이터 및 위급상황데이터 등 10종



자유(21개)

최신 시장 수요 등을 반영한 자유주제 AI 학습용 데이터 선정 및 추진

02

데이터 품질 관리 현황 및 이슈

AI

품질 관리 현황

AI 학습용 데이터 구축 및 품질 관리 공통기준 마련('19년)

- 데이터 구축 단계별 절차 체크리스트 마련

요구사항	체크리스트	
데이터 품질 요구사항	원시데이터구성 원시데이터소스정보	원시데이터속성 데이터라벨링품질관리 AI활용품질관리
데이터구축요구사항	데이터획득 데이터정제	데이터라벨링 데이터검수

품질 이슈

품질기준 미흡

- 품질관리 절차 준수여부, 데이터의 라벨링 오류 위주
- 데이터의 다양성, AI모델 유효성 등 실질적·종합적인 검증에는 한계

사후 유지보수 체계 부재

- 데이터 구축단계 이후 사후 활용단계에서의 품질 유지보수·업데이트 체계 부재
- 해외 개방 데이터셋의 경우, 활용단계에서 품질오류 개선, 라벨링 방법 추가, 데이터 추가 등 지속적으로 유지보수·업데이트

도메인 전문성 한계

- 의료·자율주행차·농림수산 등 분야별 전문적 검증에는 한계
- 데이터를 활용할 활용기업이 참여하는 품질검증 체계는 부재

수행기관 책임감 담보 미흡

- 양질의 데이터 구축을 위해서는 수행기관의 품질역량의 제고와 책임감이 핵심
- 구축한 데이터의 품질결과에 따른 인센티브 및 제재방안 부재

품질 이슈

AI 학습용 데이터 구축 관련 품질관리의 구조적 특성(한계점)

대량의데이터(수만~수백만건)를 수집-정제-가공-품질관리 전과정을
대부분 수작업 처리, 구축단계에서 **완벽한 품질확보가 어려움**

해외사례

'04년 구축 시작된 MS COCO, '16년 구축 시작된 Google Open Images의 현재 재현율
(라벨링해야 하는 사물 중 라벨링된 비율)이 각각 83%, 43%에 불과

"활용단계에서 **지속적 유지보수** 필요"

[참고] OCR AI 학습용 데이터 오류 예시(1)



뉴 노멀 시대
선도를 위한
ICT 표준의
역할

[참고] OCR AI 학습용 데이터 오류 예시(2)



03

품질관리 개선방안 및 표준화

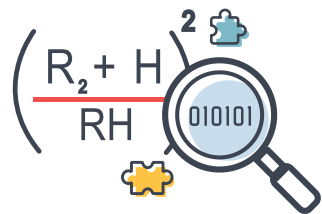
AI

'데이터 품질이 좋다'고 말할 수 있으려면

- 일반적으로 데이터는 사용자가 운영, 의사결정, 계획을 하는데 있어 의도된 활용에 적합한 경우에 **데이터 품질이 좋다**고 말함
- 또한, 데이터가 나타내고자 하는 **실제 세상의 구성을 정확하게 표현하고 있는 경우** 품질이 높다고 간주

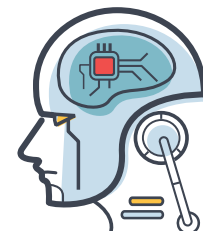
※ 출처: en.Wikipedia.org/wiki/Data_quality

AI 학습용 데이터가 품질이 좋다는 의미



- 01 실제 세상의 특성이 다양하고 편향없이 반영되어 있다
- 02 지도학습용 정답(Ground Truth)을 신뢰할 수 있다
- 03 머신러닝에 활용하기 적합한 형태이다

품질관리 개선방안



01
글로벌 수준의 품질평가 및 유지보수 체계를 벤치마킹

02
수행기관의 전문성·책임성을 대폭 강화

03
사전교육 및 컨설팅을 강화하여
수행기관의 품질관리 인식 및 역량 제고

품질관리 개선방안

(1) 정량적 품질기준 확립



- 해외에서 데이터 품질검증에 활용되는 적합성, 정확성, 유효성 등에 대한 종합적인 평가기준 및 검증방법 적용
- 사업협약 및 사업초기 단계에서 수행기관과 구체적으로 합의

- 수행기관과 **품질기준에 대한 구체적 합의 + 사전교육 및 컨설팅**을 통해 수행기관이 **스스로 품질관리를 강화**할 수 있도록 유도
- 특히, **사업착수 초기에 집중적인 품질검증**을 통해 **반복 재작업 최소화 유도**

[참고] 품질검증 항목 및 지표



AI 학습용 데이터 품질확보를 위해
핵심 품질지표 및 세부지표를 발굴하고 향후 지속 개선

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

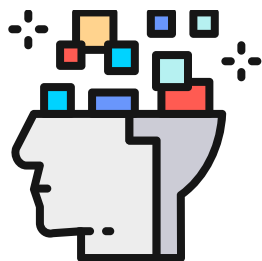


검증대상	검증대상	주요내용
품질관리 절차준수	절차준수율	공통기준상 품질관리 절차준수율 (품질관리 책임자 지정, 라벨링 작업자 교육 등)
원시데이터	기계학습 적합성	획득데이터가 기계학습으로 문제를 해결하는데 적합하게 구성되어 있는지 점검 *통계적 다양성, 품질(해상도, 음질 등) 등
라벨링 데이터	라벨링 정확성	데이터 레이블이 빠짐없이 정확하고 일관성 있게 작업되었는지 검사 -재현율, 정확도(정답 라벨링 비율) 등
	기계학습 유효성	AI 알고리즘을 훈련시켰을 때 목표로 했던 수준의 성능을 달성하는지 시험 *분류 성능, 탐지 성능, 기계번역 등

[참고] 품질검증 항목 및 지표

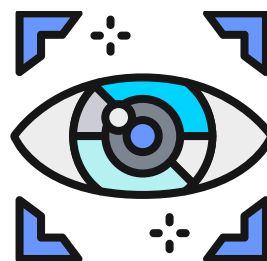
(1) 원시데이터 기계학습 정확성

다양성
(diversity)



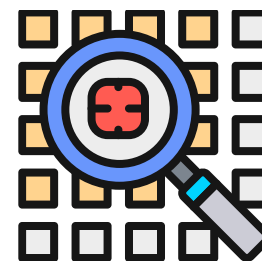
활용목적을 달성할 수 있도록
인공지능이 처리해야하는
실제 세상의 데이터와 유사한
특성과 변동성을 가진
데이터로 구성

신뢰성
(trustworthiness)



원시데이터는 반드시 신뢰할
수 있는 출처로부터 획득

공정성
(fairness)

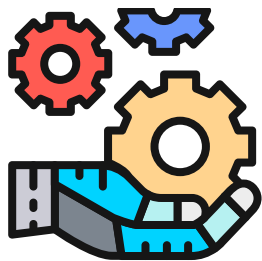


지역적 편견, 사회적 편견,
인종적 편견 등 의도가
없을지라도 데이터 속에
포함될 수 있는 편향성 제거

[참고] 품질검증 항목 및 지표

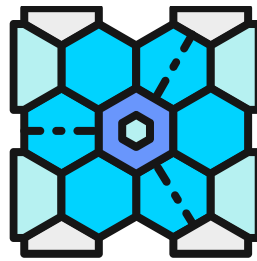
(1) 원시데이터 기계학습 정확성

충분성
(sufficiency)



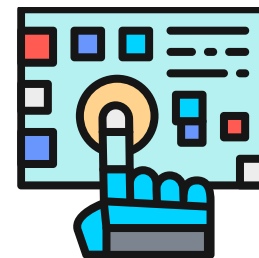
카테고리와 인스턴스는 학습
에 **유용한 수량**이어야 함
* MS COCO 데이터셋은 객체 당
5,000개 이상 인스턴스 수집

균일성
(uniformity)



분류/탐지/인식/이해/예측
카테고리 별 인스턴스 수량의
균일성 및 **비율고려**

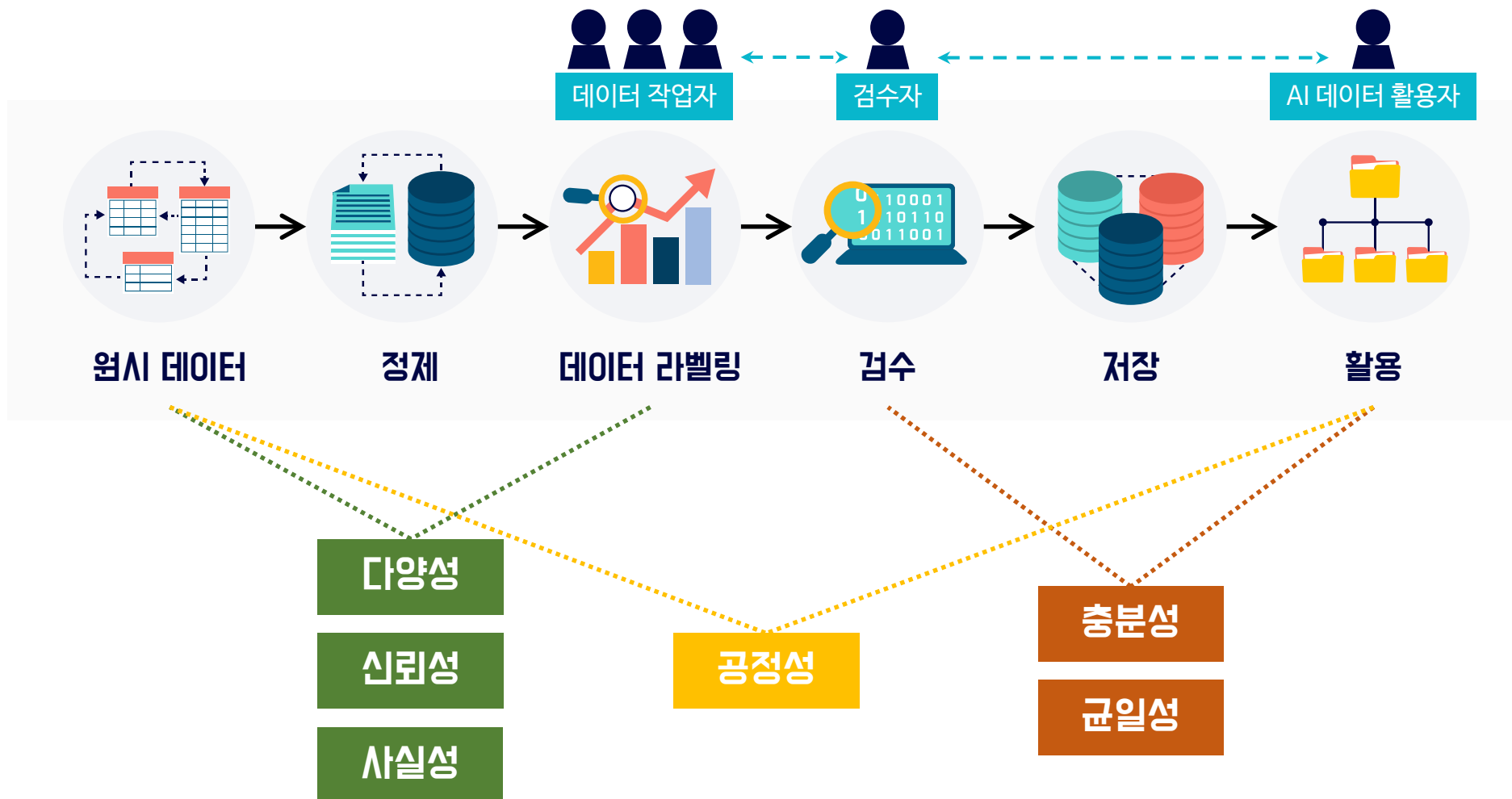
사실성
(reality)



원시데이터를 인위적인
환경과 조건하에 획득해야
하는 경우 반드시 **실제 환경과
상황 특성을 반영**

[참고] 품질검증 항목 및 지표

(1) 원시데이터 기계학습 정확성



뉴 노멀 시대
선도를 위한
ICT 표준의
역할

[참고] 품질검증 항목 및 지표

(2) 라벨링 데이터 정확성



어노테이션의 기능목적 적합성

- 인공지능으로 해결하려는 문제의 목적에 적합한 어노테이션 방법을 적용해야 함

* CCTV 사람 행동탐지를 위해서는 키포인트 표시 필요



객체 분류의 명확성

- 객체 분류 시 표시해야 하는 것과 아닌 것의 구분, 카테고리 간 중의성 제거 등 분류체계가 명확해야 함

* 도로상 데이터의 객체분류 체계 - 보행자, 차, 도로표지, 안전표지, 신호기, 기타



어노테이션 특성 정보의 포괄성

- 현재 최신의 인공지능 기술에서 요구하는 특성 정보를 포함하고 있어야 함

* 간판글씨 속성 - 폴리곤 좌표, 문자방향, 진행방향, 문자아트, 특수문자, 가려짐

[참고] 품질검증 항목 및 지표

(3) 라벨링 데이터 기계학습 유효성



학습 유효성

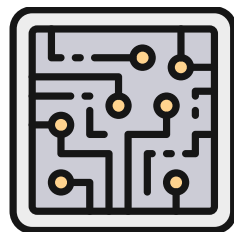
- 기계학습 유효성 달성 목표를 수립하고 관리하여야 한다.
- 잘 알려진 최신의 기계학습 알고리즘 또는 자체 알고리즘을 활용한다.
- 유효성 측정지표 및 목표치를 설정해야 한다.

*질의응답 F1-score(90점), 근거 - SQuAD2의 리더보드는 2019년 11월 말 이후 최저 성능이 F1-score 90.037임

- 데이터셋은 train/validation/test로 분할하되 분할된 각각의 데이터셋이 전체 데이터셋과 유사한 분포를 갖도록 해야 한다.

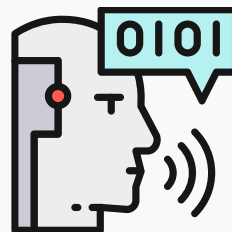
[참고] 품질검증 항목 및 지표

[핵심 품질지표 별로 검증 주기 및 방법을 달리하여 품질 검증 방법 효율화]



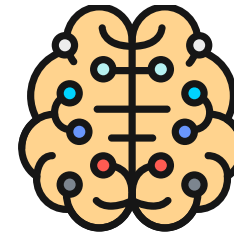
적합성

전체의 1% 데이터 구축시점부터
전문가가 데이터 샘플과 규격간
비교검토를 주기적으로 실시



정확성

전체의 5% 이상 데이터가 누적된
시점부터 데이터 샘플의 수작업
검사(의미적 정확성) 및 라벨링
데이터 구조에 대한 전수검사
(구문적 정확성)를 지속적으로
실시



유효성

전체의 70% 이상 데이터를
사용하여 훈련/검증/테스트셋으로
분할하고 알고리즘 훈련 완료 후
성능 측정

원시데이터 적합성, 라벨링 정확성 검증 제공('20. 10)
유효성 검증 실시('20. 12)

품질관리 개선방안

(2) 도메인 전문성 강화



구축한 데이터 평가시 분야별 산·학·연 전문가, 활용기업 등이
참여하는 ‘품질평가자문단’을 운영

- 부처 전문기관(스마트팜-농진청 등), 대형과제 책임자(의료-Dr. Answer 등)의 참여·검증을 통해 부처 정책 및 R&D·상용화 계획과의 연계성 제고

- 분야별 **품질기준 합의** 및 **데이터 활용성 정기검증**에 참여

품질관리 개선방안

(3) 사후 유지보수 체계 마련



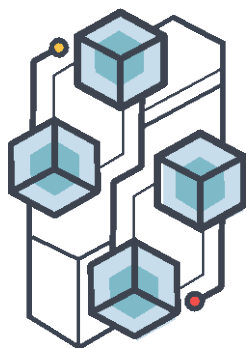
무상 유지보수 의무기간(1년간)을 설정, 해커톤 대회 개최 등을 통해
품질개선, 데이터 추가 등 유지보수 추진

- 정부출연금의 2%를 하자 보증금(현금/보증보험증권) 납부, 1년 내에 하자보수 완료 시 수령(미흡 시 환수 조치 → 부정당업체 지정 및 입찰참여제한)
- 무상 유지보수 의무기간 이후에는 SW 분야(개발비의 10~15%) 등을 벤치마킹 하여 별도의 유지보수 예산 반영 추진

- 기 구축된 데이터('17~'19)에 대한 올해 전수조사를 거쳐 필요시 **내년에 유지보수** 추진

품질관리 개선방안

(4) 수행기관 책임감 제고



수행기관에 2인 교차검증 의무화, AI 허브에 데이터별 구축기업
(과제책임자 포함)과 품질평가 결과* 공개

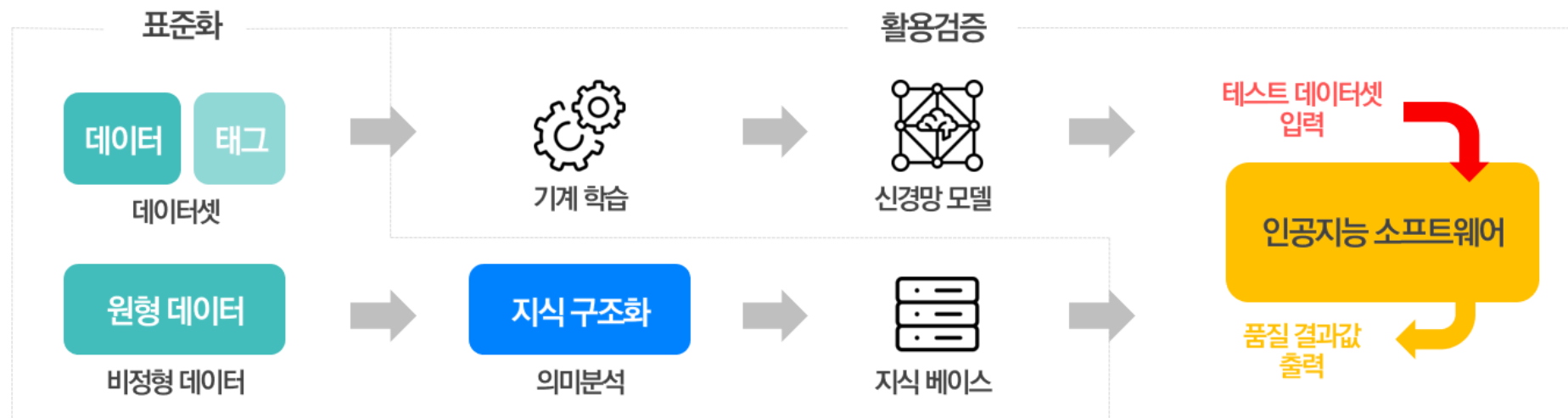
* 예) 자율차 데이터 라벨링 재현율 98% : 신뢰수준 97%, 오차율 $\pm 3\%$ 수준에서 검증

- 품질평가 결과*에 따라 차년도 사업 선정 시 가·감점을 반영

* 평가결과 등급제 (A, B, C, D 등) 등 구체적 방안 마련 예정

AI 학습용 데이터 표준체계 구축

대규모 고품질의 AI 학습용 데이터 구축·활용을 촉진하기 위한
인공지능 추론·학습용 데이터 표준화 및 활용 검증 방안 연구



학습의 목적을 고려한 AI 학습용 데이터의 품질의 **보편적인 기준**을 만들기 위한 연구 지속

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

AI 학습용 데이터 표준체계 구축

데이터 획득, 정제, 가공, 검수, 활용의 전주기별 품질관리 및 검증 방안 정립

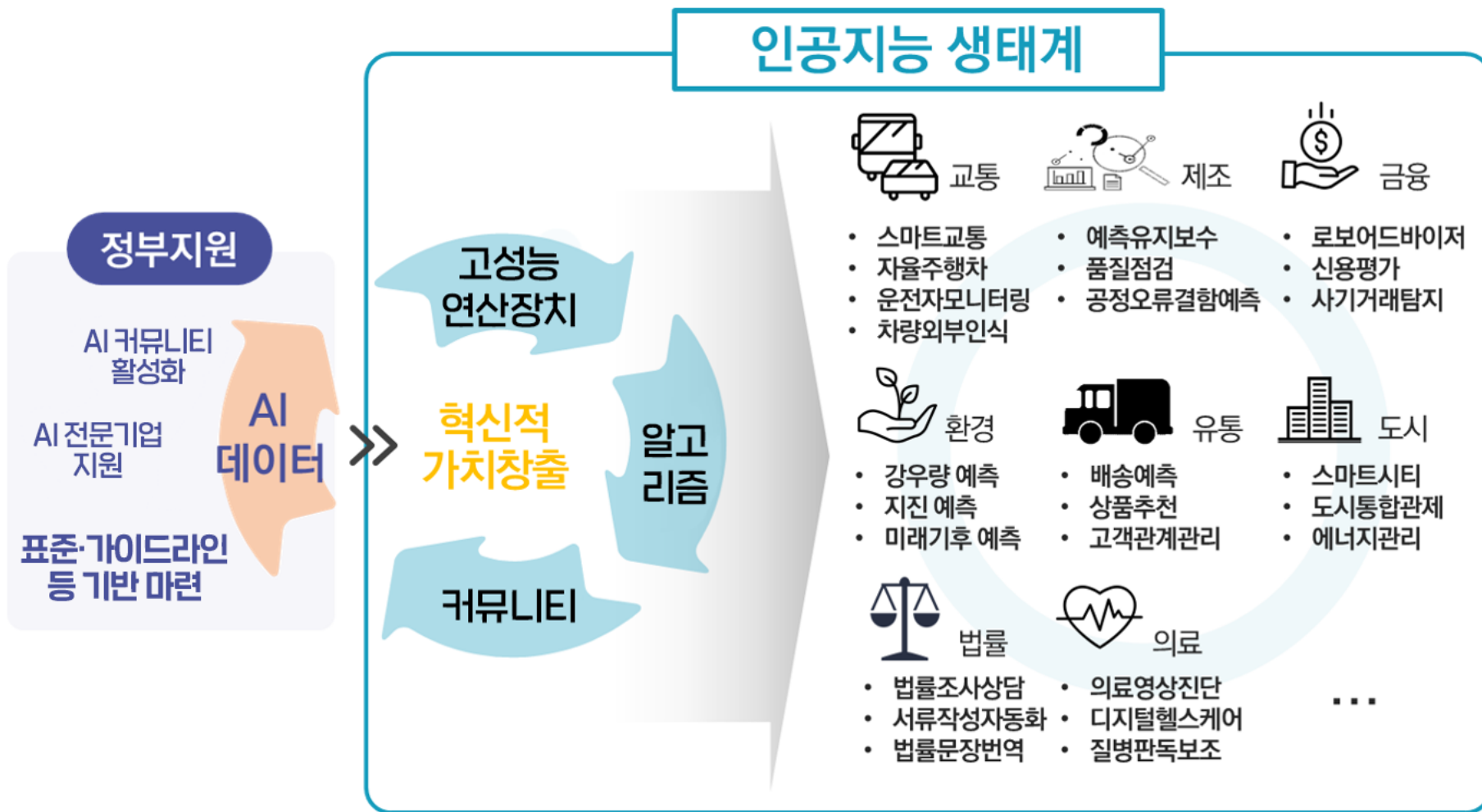
구분	내용				
구축절차	문제정의	데이터수집	정제·가공	데이터검수	데이터활용
단계별 산출물 (수행기업)	문제유형 정의서	구축공정 설계서, 어노테이션매뉴얼, 데이터셋 검수결과, 데이터셋 분석결과			데이터셋명세, 시범서비스(프로토타입)
가이드라인	데이터구축 요구사항	AI학습용데이터구축가이드라인			가이드북
품질검증	품질기준마련	단계별 산출물 점검 및 피드백			품질검증 보고서

AI 학습용 데이터 구축 및 품질 관리 공통기준 단체 표준(안) 제시('20)

향후 "국내 및 국제 표준" 제시를 위한 초석 마련

인공지능 데이터 구축 그리고 생태계 활성화

뉴 노멀 시대
선도를 위한
ICT 표준의
역할





감사합니다