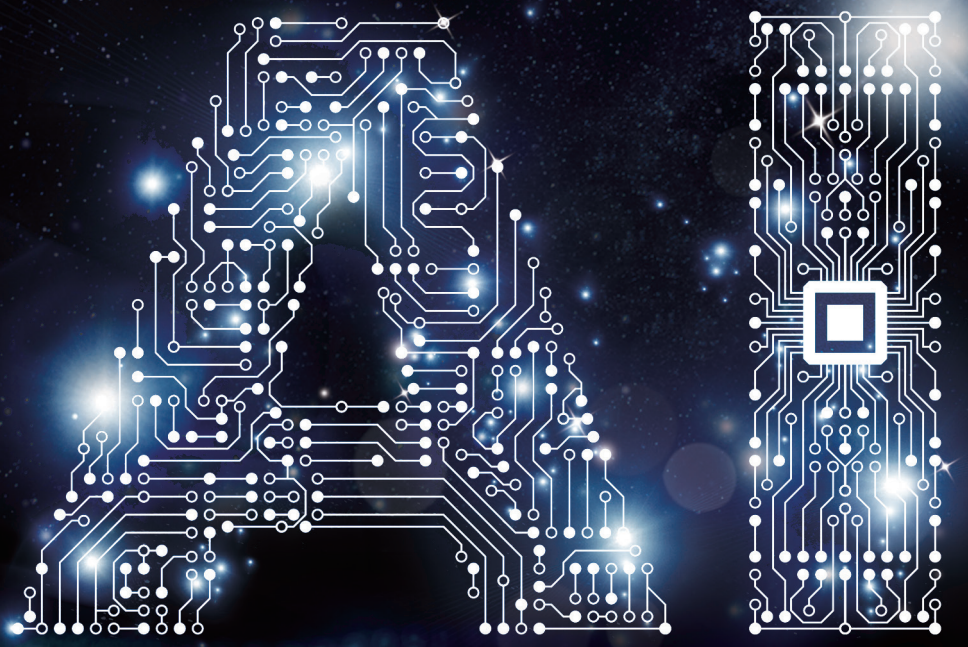
AI INSIGHT REPORT

VOL.01
2019. 9

인공지능 현장에서 듣는 AI데이터 중장기 구축 방향

AI전문가 심층 인터뷰



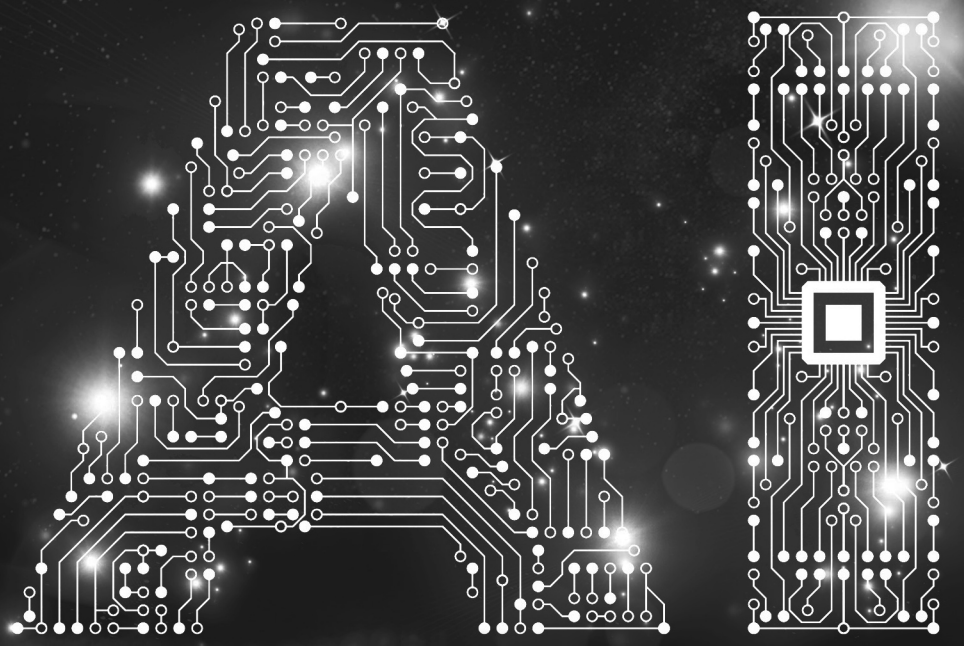
NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

AI INSIGHT REPORT

VOL.01
2019. 9

인공지능 현장에서 듣는 AI데이터 중장기 구축 방향

AI전문가 심층 인터뷰



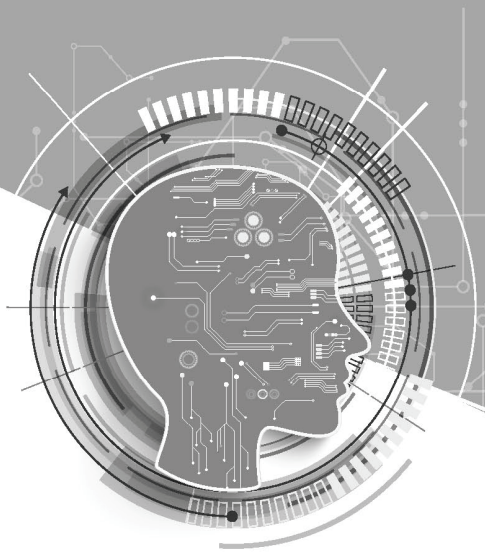
NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY

AI INSIGHT REPORT

인공지능 현장에서 듣는
AI데이터 중장기 구축 방향

- AI전문가 심층 인터뷰 -

제 1 호 (2019.9.)



Contents

NIA 한국정보화진흥원

AI Insight Report

- **"AI INSIGHT REPORT"**는 급변하는 인공지능산업의 기술, 서비스, 정책에 대한 시의성 있는 정보 제공과 심도 있는 분석을 위해 한국정보화진흥원에서 기획·발간하는 보고서입니다.
- 한국정보화진흥원의 승인 없이 본 보고서의 무단전제와 복제를 금하며, 인용 시에는 반드시 "한국정보화진흥원, 「AI INSIGHT REPORT」" 임을 밝혀주시기 바랍니다.
- 본 보고서의 내용은 한국정보화진흥원(NIA)의 공식 견해와 다를 수 있습니다.

▶ 작 성 한국정보화진흥원 지능데이터본부
AI데이터팀 홍효진 수석(hhyoj@nia.or.kr)
㈜마인즈앤컴퍼니 이혜정 이사(haejung.lee@mindslab.ai)

▶ 기 획 오성탁 본부장, 윤억수 팀장

▶ 발행인 문용식

▶ 보고서 온라인 서비스 www.nia.or.kr, www.aihub.or.kr
https://ko-kr.facebook.com/kict.bigdata

발 간 사

사람부터 사물에 이르기까지 실시간으로 쏟아지고 있는 방대한 데이터는 숨 고를 틈도 없이 우리를 4차 산업혁명의 한가운데로 몰아가며, 인류의 새로운 자원으로 주목받고 있습니다.

구슬이 서 말이라도 꿰어야 보배가 되듯이, 의미 없는 정보 조각에 불과했던 데이터가 '인공지능'을 만나며 전 세계적으로 무한한 가능성과 새로운 가치를 만들어내고 있습니다. 알파고 이후 인공지능에 대한 대중적 관심이 환기되면서, 인공지능이 의료, 금융, 교통, 제조 등 여러 산업의 생산성을 획기적으로 향상시킨 사례는 이미 곳곳에서 읽을 수 있습니다. 이를 반영하듯, 인공지능 기술 기업의 시장 가치는 나날이 높아지고 있으며, 2019년 시가총액 기준 글로벌 10대 기업 중 7개 기업이 인공지능 기술을 직접 개발·투자 중입니다.

인공지능은 앞으로 경제, 사회 전반에 우리가 이전까지 경험하지 못했던 수준 이상의 상당한 파급력을 미칠 것으로 보입니다. 이로 인한 변화가 우리 사회의 긍정적 발전의 동력이 되기 위해서는 공공과 민간의 지속적인 정보 공유, 대화, 협력이 절실합니다. 특히, 인공지능의 원료가 되는 데이터를 누구나 자유롭게 쓸 수 있고, 새로운 기술을 마음껏 시도할 수 있는 환경을 만드는 것이 중요합니다.

이에 한국정보화진흥원은 인공지능이 우리 경제와 사회에 초래할 변화상을 풀어내고, 능동적으로 대응할 수 있는 통찰을 제공하기 위해 'AI INSIGHT REPORT'를 발간합니다. 모쪼록 이 보고서가 인공지능 분야에 직·간접적으로 몸담고 계신 모든 분에게 지능화 시대에 유연하게 대처하는데 유용한 가늠자 역할을 할 수 있기를 바랍니다.

2019. 9.

한국정보화진흥원 원장 문 용 식

요 약 summary

◇ AI전문가 심층 인터뷰 개요

○ **(목적)** AI데이터 중장기 구축 방향 수립을 위해 산·학·연 AI 전문가의 현장 수요 파악 및 현실감 있는 Insight 획득

* 정부는 인공지능 제품·서비스 개발 및 성능 향상에 필수적인 인공지능 학습용 데이터를 구축하고, AI허브(www.aihub.or.kr)에 공개

○ **(논의)** AI데이터의 중요성, 중장기 구축 방향, 공공의 역할

○ **(방식)** 총 4개월간('19.5~8월) 개별 방문, 간담회 등 추진

○ **(전문가)** 총 20명(AI스타트업, AI 활용 전문기업, 학계 및 연구계)

◇ 인터뷰 주요 내용

1 방대한 양질의 데이터가 인공지능 시대의 핵심 자산

○ 기본적인 AI알고리즘 대부분이 오픈소스로 공개되어 있으나, AI데이터는 충분히 우위를 점할 수 있는 전략적 요소

* 실제로 중국은 방대한 데이터를 동력으로 강국으로 부상

2 기술 격차 해소를 위한 최적의 방법은 AI데이터 확대

○ AI데이터 구축은 데이터 수집, 가공, 모형 적용에 이르는 데이터 가치 사슬 전반에 대한 경험을 제공하기 때문에, 인공지능 역량 축적의 중요한 시발점 역할

3 기술 트렌드를 반영한 범용성 높은 글로벌 선도 데이터 구축

○ Transfer Learning 등 최신 기술 트렌드를 반영하여, 인공지능 기술 선도에 기여할 수 있는 데이터 타입별(이미지, 영상, 텍스트, 음성 등) 대규모 범용 AI데이터 구축

4 산업계가 요구하는 도메인별 데이터 구축 병행

○ 상용화 가능성은 높으나, 현실적으로 데이터가 부족하여 상용화가 더딘 전략 도메인을 선정하여 AI데이터 구축

* 프로젝트 성공 가능성 타진 시 주로 해외 유수의 범용 데이터를 사용하고, AI기업은 특화된 응용 제품·서비스 개발에 집중하는 경우가 대다수

5 국가 간 데이터 바이어스(bias)를 제거하고, 상용화 수준의 서비스 개발을 위해 한국형 AI데이터 구축

○ 한국어 음성·말뭉치, 한국인 안면 이미지, 한글 OCR 등 한국형 AI데이터에 대한 높은 수요를 반영한 데이터 확대

6 공공성과 사회적 수요가 높으며, 법제도 제약으로 인해 민간이 추진하기 어려운 AI데이터 구축

○ 기후, 환경, 교통, 복지, 의료영상, CCTV영상, 법률·판례·특허 등 사회문제 해결형 AI데이터 구축·공개

7 인공지능산업의 생태계 조성까지 고려한 AI데이터 구축

○ '데이터 개방'이 생태계의 끝이 아님을 인지하고, AI데이터 생애주기 全 단계에 걸친 종합·체계적인 후속 조치 마련

* 데이터 활용·서비스 개발 가이드 제공, 타 기관 및 사업과의 제휴, 글로벌 오픈소스 플랫폼 및 경연대회와의 연계, 공개된 데이터의 지속적인 사용과 업데이트를 위한 장치 필요

8 인공지능 개발자의 사용 유도를 통한 AI데이터 활용도 제고

○ AI허브에 공개된 데이터를 활용, 논문 발표 및 인용 건수를 증가시키고 새로운 비즈니스가 창출되는 선순환 구조 정착을 위해 R&D 연계 및 챌린지(수시 경진대회) 적극 지원

9 풍부한 한국형 AI데이터 구축을 위한 데이터 규제 해결 절실

○ 인공지능 학습용으로 제대로 가공할 수 있도록 쓸만한 원천 데이터 공개를 위한 적극적인 데이터 규제개선 필요

Ⅰ AI전문가 심층 인터뷰 개요

□ 추진 목적

○ 인공지능에 대한 현장수요에 대응하고, “데이터·AI경제”의 조기 활성화를 위해 AI데이터 중장기 구축 방향 수립

* 정부는 인공지능 제품·서비스 개발 및 성능 향상에 필수적인 인공지능 학습용 데이터를 구축하고, AI허브(www.aihub.or.kr)에 공개

○ 산업계(스타트업, 벤처, 대기업 등), 학계 및 연구계 등 다양한 도메인에서 활동 중인 AI전문가의 현실감 높은 Insight 획득

□ 주요 논의 사항

○ AI데이터의 중요성, 중장기 구축 방향, 공공 부문의 역할 등

□ 추진 방식

○ 기간 : ’19. 5월 ~ 8월(4개월 간)

○ 조사 방식 : Ideation 회의, 개별 방문 인터뷰, 간담회 등

○ 조사 대상 : AI전문가 총 20명

* 전문가 명단은 보고서 후면에 기재

< AI전문가 구성 >

소속	AI스타트업	AI 활용 전문기업	학계 및 연구계	합계
참여 인원	8명	8명	4명	20명

[참고1] AI데이터 구축 사업(17년~)

◇ 정부는 인공지능 제품·서비스 개발 및 성능 향상에 필수적인 인공지능 학습용 데이터를 구축하고, AI허브(www.aihub.or.kr)에 공개

- ’17년 본 사업을 통해 법률, 특허, 일반상식, 한국형 이미지 등 AI데이터 4종을 구축하여, ’18년 1월부터 AI허브에서 제공 중
- ’19년 1월에는 관광, 농업, 헬스케어 등 AI데이터 7종을 개방하였고, 6월에 한국어 음성 등 4종을 추가 개방
- ’19년 말에는 AI데이터의 종류와 수량을 대폭 확대하여, 10종* 2,500여만건의 AI데이터를 공개할 계획

  * 한·영 번역 말뭉치, 한국형 사물 이미지, 한글 글자체 이미지, 인도 보행 영상, 멀티모달 영상, 사람 동작 영상, 한국인 안면 이미지, 위험물 X-ray 이미지, 질병 진단 이미지, 이상행동 CCTV 영상

  - (1차 공개) AI허브에 10종 AI데이터 중 일부 공개(’19년 7월 말)

< AI데이터 구축·현황 및 계획 >

출처 : 과학기술정보통신부 보도자료(2019.6.14.), ‘한국 사람의 말과 감성을 이해하는 똑똑한 AI 시대를 앞당긴다.’

[참고2] AI데이터 공개 플랫폼, AI허브(AI Hub)

◇ **'AI허브'는 국내 중소·벤처·스타트업이 인공지능 학습용 데이터, 알고리즘, 고성능 컴퓨팅 자원을 쉽게 활용할 수 있도록 원스톱으로 제공하는 플랫폼**

- **(이용대상)** 국내 기업 및 대학, 연구기관, 개인 등 누구나 계정을 발급받아 포털에서 제공한 서비스를 활용 가능
- **(준수사항)** 제공한 학습용 데이터를 활용한 제품·서비스 및 R&D 과정에서 산출된 논문, 제품 등 결과물에 데이터 출처 명시
  * 다운로드를 받은 데이터는 제3자에게 재배포 불가

< AI허브 포털(www.aihub.or.kr) >

Ⅱ 인공지능산업과 AI데이터

□ 방대한 양질의 데이터가 인공지능 시대의 자산이자, 경쟁력

○ 기본적인 AI알고리즘 대부분이 오픈소스로 공개되어 있으나, AI데이터는 우리가 충분히 우위를 점할 수 있는 전략적 요소

* 인공지능의 핵심요소는 데이터, 알고리즘, 컴퓨팅 파워이며, 그 중 인공지능 학습을 위한 AI데이터는 도메인별 모델 성능 향상에 필수

> *"AI데이터 구축은 고속도로를 까는 일과 같이 국가의 경쟁력이에요. 알리바바의 마윈은 데이터를 new oil이라 했고, 엔비디아의 젠슨 황은 new source code라고 했지요"*

○ 실제로 중국은 방대한 데이터를 동력으로 인공지능 강국으로 부상

* 인공지능 특허출원 건수 : 중국 19개 vs 미국 12개('16~'18년)[1)]

** 세계 100대 AI스타트업 중 유니콘 기업 수 : 중국 5개 vs 미국 5개[2)]

> *"중국은 데이터 활용에 제한이 거의 없습니다. 중국은 족집게 과외를 하듯, 수능 전 문제지만 많이 풀어보듯, 데이터 품질 수준을 99.9% 까지 높이고 이를 기반으로 상용화 기술 성능을 확보했습니다."*

○ 특히, AI응용 서비스 개발과 모델 성능 향상에 집중하는 중소·벤처·AI스타트업에게 데이터 확보는 생존을 위한 필수 자원

> *"모두가 구글과 페이스북이 될 필요는 없습니다. AI기업들이 해야 할 일은 서비스 출시를 위한 기반 데이터를 확보하는 것입니다. 더 나아가 서비스 정교화를 위한 데이터를 지속적으로 수집할 수 있는 '데이터 순환 루프(loop)를 구축하는 것, 그것이 AI기업의 핵심 전략이 되어야 합니다."*
>
> *(Andrea Ng)* [3)]

1) 니혼게이자이신문, '주요 10개국 특허 기관에 출원된 AI 관련 특허 통계' 자료, 19.3
2) CB Insights, 'AI 100:The Artificial Intelligence Startups Redefining Industries', 19.2
3) TechTarget(2017), 'How to win in the AI era? For now, it's all about the data'

[참고3] 국내 스타트업의 AI데이터 활용 사례

◇ AI스타트업 '㈜스켈터랩스'는 AI허브에서 제공하는 한국어 대화 AI데이터를 활용하여, AI 대화엔진 성능 테스트 진행

- ㈜스켈터랩스는 AI허브에 공개된 '총 10만건의 한국어 대화 데이터'를 활용해 자사의 'AI기반 대화엔진 성능 테스트'를 진행
- 평가 결과, F1 스코어가 72%로 글로벌 기업(A사: 67%, B사: 66%) 대비 월등히 높은 성능을 기록

  * F1 스코어: 정밀도, 재현율 고려한 AI 기술 분류 평가 수치

> ***"인공지능 기술 개발 경쟁력은 데이터 확보 여부에 달려있습니다. 이번 한국정보화진흥원이 공개한 250만여 인공지능 학습용 데이터는 스타트업의 기술 개발 과정에 자산이 되었습니다"***
>
> ***- 조성진 ㈜스켈터랩스 CTO***

◇ 수원시, 이상행동 CCTV 영상 데이터 학습('19년 AI데이터 구축 사업으로 추진 중)으로 범죄 잡는 AI CCTV 도입

- AI전문기업 ㈜마인즈랩은 수원시, 수원남부경찰서와 함께 '19년 말 AI허브에 공개될 '이상행동 CCTV 영상 데이터'를 활용하여 지능형 CCTV 모델을 구축하고, 실제 CCTV 관제 시스템에 적용할 계획
- CCTV 카메라에 포착된 사람의 수상한 행동을 AI를 통해 분석하고, 위험이 감지되면 경찰서, 소방서 등에 즉시 알려 신속하게 대응

출처 : IT동아(2019.8.22.), '스켈터랩스, 인공지능 기반 대화엔진 정확도와 재현율 테스트 공개'
한국경제(2019.06.12.), '마인즈랩-수원시, 범죄 잡는 AI CCTV 도입한다'

□ 인공지능기술 격차 해소를 위한 최적의 방법은 AI데이터 확대

○ 도메인별 AI모델 성능을 개선하고, 상용기술 수준을 개선하기 위해서는 누구나 양질의 풍부한 데이터를 자유롭게 활용하여, 새로운 기술을 마음껏 시도할 수 있는 환경 조성이 선결 조건

- 아무리 높은 기술력을 갖춘 AI스타트업이라도 오랜 시간과 높은 비용이 수반되는 데이터 생산과 가공 비용은 큰 부담으로 작용

  * AI 적용 전체 과정 중 데이터 관련 업무가 평균 1/3 이상의 시간 소요(Kaggle, '18)

> ***"예를 들어 음성 데이터의 경우 돈이 너무 많이 들기 때문에 정말로 스타트업이 만들 수가 없어요. 수만 시간짜리 데이터가 필요하거든요."***

○ 현재 빠르게 확장·고도화 중인 국내 인공지능산업은 상용화에 실질적인 도움이 되는 수준의 현장 데이터에 대한 수요 급증

  * 국내 인공지능산업은 원천기술개발 중심의 초기 단계에서 他산업과 융합되어 상용화가 시작되는 과도기인 것으로 유추[4)]

○ 특히, AI데이터 구축은 데이터 수집, 가공, 모형 적용까지 데이터 가치 사슬 전반에 대한 경험을 제공하기 때문에, 국가 사회적으로 인공지능 역량 축적의 중요한 시발점 역할 수행

> ***"외국 데이터에만 의존하면 아무도 데이터를 구축해본 경험이 없게 됩니다. 데이터를 구축하는 것은 안 해본 사람은 못해요."***
>
> ***"데이터를 구축하는 과정은 결과로는 데이터가 남고, 과정으로는 구축 경험이 남습니다. 데이터 구축 사업은 데이터 구축하는 인력을 키워내는 사업이라고 생각합니다."***

4) 한국정보화진흥원 내부자료(AI스타트업, AI전문기관 대상 설문조사 결과, '19년 5월)

Ⅲ AI데이터 중장기 구축 방향

□ 기술 트렌드를 반영한 범용성 높은 글로벌 선도 데이터 구축

○ Transfer Learning* 등 최신 기술 트렌드를 반영하여, 인공지능기술 선도에 기여 할 수 있는 주요 데이터 타입별(이미지, 영상, 텍스트, 음성, 등) 대규모 범용 AI데이터 구축

* 하나의 특정한 과업(task)만 처리할 수 있는 알고리즘이 아닌, 범용지능으로 자기지도학습(self-supervised learning)을 통해 새로운 과업에 적합한 의미 있는 가공(fine tunning)을 하는 기술로, 현재 전 세계적으로 각광

- 실제로 AI서비스 개발 과정에서 범용 데이터로 사전학습을 하고, 미세조정(fine tunning)을 거쳐야 우수한 성능 확보 가능

> ***"범용 데이터는 거의 모든 문제 해결을 위해 도움이 되는 데에 중요합니다. 이미지넷의 경우 클래스 분류만 생각하면 그걸 어디에 쓰나 싶지만, 사전학습을 시키는 것과 시키지 않는 것에는 넘을 수 없는 수준의 차이가 있습니다. 범용 데이터를 기반으로 성능을 최대한 높여야 합니다."***

○ 특히, 유튜브 등 동영상 데이터 사용이 기하급수적으로 증가하고 있어, 국가 차원의 대용량 동영상 AI데이터 구축 시급

> ***"기술 개발 및 서비스 상용화의 전망으로 보았을 때, 앞으로는 이미지 데이터보다 영상 데이터의 수요가 많아질 것입니다. 따라서 최대한 많은 동영상 데이터를 확보하는 것이 중요할 것입니다."***

○ '주제'의 범용성이 아닌, '사용'의 범용성 추구를 위해 카테고리(도메인, 데이터 종류)별 공공성이 높은 AI데이터 구축 필요

* 기업이 스스로 데이터를 오픈하는 것은 불가능하므로, 정부가 직접 레퍼런스 데이터를 구축·개방할 필요(예. 한국어 손글씨체)

□ 산업계가 요구하는 도메인 기반 데이터 구축 병행

○ '시장 수요가 많은 데이터'가 곧 '범용 데이터'가 되므로, 구체적인 도메인에서 범용성을 갖춘 대용량 AI데이터 구축

○ AI응용서비스 개발과 조기 성과 도출이 가능하도록 산업계의 다양한 인공지능 프로젝트에 기반한 AI데이터 구축 고려

- 상용화가 활발하게 진행 중인 의료, 교육, 법률, 자율주행 등 세부 도메인별 예상되는 AI응용 서비스를 고려한 AI데이터 구축

  * 프로젝트 성공 가능성 타진 시 주로 해외 유수의 범용 데이터를 사용하고, AI기업은 특화된 응용 제품·서비스 개발에 초점을 두는 경우가 다수

○ 상용화 가능성이 높음에도 불구하고 현실적으로 데이터가 부족하여 상용화가 더딘 전략 도메인을 선정하여, AI데이터를 구축

- AI스타트업 중 규모의 경제를 실현할 만큼 성장한 사례는 아직 없으므로, 데이터 부족 문제 해결을 통해 성공사례 발굴 필요

> ***"스타트업이 많긴 한데 아직 지배적인 플레이어가 없습니다. 고만고만한 플레이어들 밖에 없습니다. 물론 그 안에서 스마트한 팀들도 있지만, 아직 국내에서 스케일업을 해서 규모를 갖춘 경우들이 많지는 않습니다. 이런저런 규제로 인해 데이터 제너레이션이 안 되다보니 그 다음 스텝으로 나가기가 어렵습니다."***

○ 다만, 최대한 많은 기업에게 혜택이 돌아갈 수 있는 규격화된 데이터 도출을 위해 면밀하고 체계적인 수요조사 필요

> ***"시장 수요는 범위가 좁아질수록 특정 기업에 편중될 위험이 있습니다. 따라서 시장 수요는 충분히 고려하지만, 많은 기업들이 사용할 수 있는 파급력이 큰 데이터를 구축하는 것이 중요하다고 생각합니다."***

□ 국가 간 데이터 바이어스(bias)를 제거하고, 상용화 수준의 서비스 개발을 위해 한국형 AI데이터 구축

○ 해외 유수의 범용 데이터*가 많이 공개되어 있으나, 이는 주로 알고리즘 학습이나 실험 수준의 연습용 데이터가 대부분

* 안면, 랜드마크, 음식, 음성대화, 텍스트 등 대부분이 서구권(서양인, 영어) 데이터로, 우리나라 상황에 적용에는 부적합

○ 우리나라 실정에 맞는 상용화 수준의 서비스 개발을 위해서는 한국에 특화된 실무용 AI데이터 필요

> *"데이터 국가 바이어스가 큽니다. 국가적 차원에서 한국인 안면 이미지, 랜드마크 이미지, 음식 이미지 같은 것을 구축하면 스타트업에서 많이 사용할 것입니다."*

○ 한국어 음성, 한국어 말뭉치, 한국인 안면 이미지, 한글 OCR 등 한국형 AI데이터에 대한 수요는 높으나, 데이터양이 턱없이 부족

* AI데이터 구축 사업('17~) : 한국어 음성·대화, 한국어 말뭉치, 한국인 안면 이미지, 한글 OCR, 한국형 사물 이미지 등 한국형 AI데이터 구축

○ 활용 가능한 해외 유수의 범용 데이터가 이미 많으므로, 정공법으로 품질에 집착하는 고전적인 방식보다는 데이터의 양을 증가시키는 것이 기업들에게 현실적으로 큰 도움

> *"같은 예산을 사용한다고 했을 때 100시간짜리 5-star 데이터보다 1만 시간 짜리 3-star 데이터가 더 좋다고 생각합니다."*
>
> *"한국어 음성인식 데이터의 경우, 작년에 NIA에서 공개한 1천 시간은 턱없이 부족해요. 적어도 음성 10만 시간은 돼야 해요."*

< 한국형 AI데이터 현황 >

데이터	주요 내용
한국어 음성·대화	▶ 데이터 구축 비용이 상당히 높아 개별 기업이 감당하기에는 어려운 수준 ▶ 해외 데이터* 역시 데이터의 양이 불충분 예) LibriSpeech(영어 음성) : 1천시간, VoxCeleb2(영상 추출 음성): 화자 7천명, 2천시간 ▶ AI 허브(www.aihub.or.kr) 공개 데이터 - 한국어 음성 1천 시간, 한국어 대화 20시간
한국어 말뭉치	▶ 저작권 이슈로 인해 현실적으로 사용 가능한 수준의 한국어 말뭉치 데이터 부족 ▶ AI 허브(www.aihub.or.kr) 공개 데이터 - 한국어-영어 병렬 말뭉치 160만 문장 구축·공개 예정('19년 말)
한국인 안면 이미지	▶ 초상권 이슈로 인해 데이터 구축 비용이 높아, 개별 스타트업이 대량의 데이터 확보가 어려운 영역 ▶ 해외 데이터* 역시 데이터의 양이 불충분 예) Celeb B : 1만명의 안면 이미지, 총 20만장 ▶ AI 허브(www.aihub.or.kr) 공개 데이터 - 현재 한국인 200명의 안면 이미지 데이터가 공개 중이며, '19년 말 600명의 안면 이미지 데이터 공개 예정
한글 OCR	▶ 글자 영역의 바운딩 박스와 텍스트 내용이 함께 태깅된 데이터는 국내외 모두 부족한 상황 ▶ 특히 한글 손글씨 데이터는 全無하여, 개별 기업이 이벤트를 해서 모으고 있는 상황 ▶ AI 허브(www.aihub.or.kr) 공개 데이터 - '19년 말 한글 인쇄체 300만자(字), 필기체 200만(字), Text in the wild 10만장(張) 공개 예정
한국형 사물 이미지	▶ 해외 범용 데이터는 있으나, 한국의 특수성을 담은 데이터는 부족 ▶ AI 허브(www.aihub.or.kr) 공개 데이터 - '19년 말 한국의 유적 건조물, 상품, 주요 도시의 랜드마크 이미지 등 360만장 공개 예정

□ 공공성과 사회적 수요가 높으며, 법제도 제약으로 인해 민간이 추진하기 어려운 AI데이터는 공공 부문에서 구축 · 공개

○ 기후, 환경, 교통, 복지 등 정부와 공공기관이 보유한 공공 데이터를 활용해 사회문제 해결형 AI데이터 구축·공개 필요

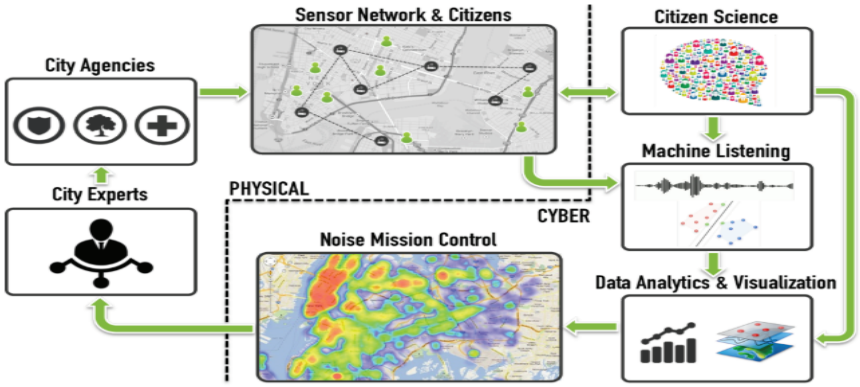
- 공공 데이터를 활용한 AI데이터는 각 영역에서 발생하는 사회문제 해결뿐 아니라, 산업계에서도 활용성이 높을 것으로 기대
  * (예) 기후 데이터는 에너지산업, 환경 데이터는 건설산업, 교통 데이터는 자동차 산업, 복지 데이터는 보건의료산업 등에서 활용 가능
- 이미 존재하는 공공 데이터를 활용함으로써 데이터 수집·생산의 어려움을 제거하고, 데이터 구축 비용 절감 가능

○ 의료영상, CCTV영상, 저소득층을 위한 AI 교육 데이터, 법률·판례·특허 등 민간이 만들기 어려운 분야는 공공 부문에서 추진

< 미국 뉴욕주립대의 SONYC 데이터 >

▶ SONYC UST(Urban Sound Tagging) 데이터는 뉴욕시의 소음 상황에 대한 사운드 레벨 데이터

▶ 뉴욕시 전역에 설치된 50여개의 센서에서 수집된 데이터를 크라우드소싱 데이터 가공 플랫폼인 Zooniverse(www.zooniverse.org)에서 시민참여를 통해 데이터 어노테이션을 실시



출처 : 뉴욕시 소음 데이터 클라우드 포털(https://wp.nyu.edu/sonyc/)

Ⅳ 공공 부문의 역할 및 시사점

□ 인공지능산업의 생태계 조성까지 고려한 AI데이터 구축 필요

○ '데이터 개방'이 생태계의 끝이 아님을 인지하고, AI데이터 생애주기 全 단계에 걸친 종합·체계적인 후속 조치 마련

- 데이터 활용·서비스 개발 가이드 제공, 타 기관 및 사업과의 제휴, 글로벌 오픈소스 플랫폼 및 경연대회와의 연계 등 AI데이터 활용 전반에 대한 종합적인 고려 필요

> ***"AI 허브를 사람들이 많이 찾는 오픈소스 플랫폼과 연동하는 것도 하나의 방법입니다. 그들은 그 플랫폼 안에서 데이터를 사용하고 또 유지 보수하며 데이터의 양질을 높여가기 때문입니다."***

- 공개된 데이터의 지속적인 사용과 업데이트를 위한 장치 필요

  * (예) AI허브(www.aihub.or.kr)에 데이터 공개 시, 구축한 기업·기관의 실명제 도입으로 책임감, 권한, 자부심을 동시에 강화하는 방안 고려

> ***"사용자들이 많이 사용하는 데이터의 대부분은 버전이 계속 업그레이드 되어 도메인에서 지속적으로 사용되는 것입니다."***
>
> ***"구글은 데이터셋을 설계하고 구축한 인력의 이름을 명시합니다. 우리도 참여 인력의 이름을 넣어 그들의 크레딧을 보장해 주면 좋겠습니다."***

○ 인공지능을 활용한 다양한 분야의 수요기업, 개발자 등과 같은 AI데이터의 실질적인 사용자의 적극적인 참여 유도 필요

> ***"인공지능이 제대로 상용화 수준까지 올라오려면, 데이터 헤게모니가 있어야 합니다. 수익을 창출할 수 있는 기업들, 즉 데이터 사용자들이 과제 수행에 참여해야 현실성을 높일 수 있습니다."***

□ 인공지능 개발자의 사용 유도를 통한 AI데이터 활용도 제고

○ 진정한 의미의 AI데이터 이용 활성화란, 논문 작성, 학회, 워크샵, 챌린지 등에 AI데이터 활용사례가 축적되고, 이것이 자연스럽게 개발자 커뮤니티와 인공지능 업계에 확산되는 것

- 뉴스, 기사 등 일반적인 미디어를 활용한 직접적인 홍보보다 실제 개발자들이 모여 있는 딥러닝 관련 채널 활용이 효과적

> ***"홍보를 위해서는 데이터를 가지고 논문을 쓰고 발표를 하고, 소위 말하는 선수들이 계속 보게 되면 활용도가 높아질 수 밖에 없습니다."***
>
> ***"데이터 설계 및 구축부터 활용방안까지 많은 이야기들이 오고 가기 때문에 학회의 주제로 논의를 해보고 데이터를 실제 사용해도 좋을 것 같습니다."***

○ AI허브에 공개된 한국형 AI데이터를 기반으로 논문 발표 및 인용 건수가 증가하고, 새로운 비즈니스가 창출되는 선순환 구조 정착을 위해 R&D 연계 및 챌린지(수시 경진대회) 적극 지원

- AI허브에 데이터뿐 아니라, AI서비스 개발공정별 산출물*까지 공개됨으로써 AI응용개발 및 지식의 폭발적 성장에 기여

  * (예) 버전별 소스코드, 데이터, 테스트 결과 등

> ***"한국형 AI데이터 기반 연구를 통해 논문 인용 수를 향상시키면, 해당 도메인의 실제 연구자들에게 모티베이션이 되고, 기업들은 한국형 데이터가 잘만 나오면 다들 활용할 것입니다."***
>
> ***"구축된 학습데이터 활용을 전제로 하는 챌린지 등을 개최하면 데이터 활용 유도가 가능하며, 학습데이터가 일단 유의미한 것이라면 스타트업의 참여가 활발할 것으로 기대됩니다."***

□ 풍부한 한국형 AI데이터 구축을 위한 데이터 규제 해결 절실

○ 인공지능은 곧 데이터 싸움이나, 법제도 제약으로 인해 활용 가능한 한국에 특화된 데이터가 상당히 부족하다는 것에 업계 전반이 공감

> *"정책적으로 규제에 막혀 있는 문제가 해결되어야 하는 것이 중요합니다. 데이터가 나오지 않으면 스타트업 입장에서는 아무것도 분석할 것이 없습니다. 이러한 규제 문제가 해결되지 않고 정부에서 수요를 예측해서 데이터를 만든다고 하면 make sense 하지 않습니다."*
>
> *"데이터의 저작권 및 개인정보 이슈를 해결 부탁드립니다. 데이터를 쓰려고 해도 항상 법무팀에서 막혀 챌린지, 기술 개발이 어렵습니다."*

○ 특히 헬스케어, 모빌리티 등은 국내 데이터의 중요성이 높고 사업화 잠재력도 크지만, 데이터 규제로 인해 해외와 달리 아직 디지타이징(digitizing) 수준이 낮은 분야

> *"모빌리티, 헬스케어 등의 분야는 데이터 포텐셜이 많은 영역입니다. 사람 몸, 도시는 디지타이징 할 부분이 너무너무 많습니다."*
>
> *"사람의 신체와 정신은 아직 데이터화 수준이 낮고 잠재력이 높은 반면, 국가 간 건강정보 공유가 원활하지 않을 뿐만 아니라 인종 등 한국인의 물리적 특성 반영이 중요하기 때문에 국내 데이터 구축이 꼭 필요합니다."*

○ 인공지능 학습용으로 제대로 가공할 수 있도록 쓸만한 원천 데이터 공개를 위한 적극적인 데이터 규제 개선 필요

* 인공지능 분야에서 상대적으로 침체되었던 일본은 최근 연구목적일 경우 비식별 정보를 제한 없이 사용(재전송 금지)할 수 있도록 저작권법을 개정

- 법제도 이슈에 저촉되지 않기 위해 오직 공개 목적으로만 가공된 데이터는 인공지능학습뿐 아니라, 통계분석 자료로도 활용 불가능

〈 심층 인터뷰 참여 전문가 명단 〉

구분	성명	직위	소속	전문 분야
AI스타트업 (8개社)	김민철	대표	㈜스위트케이	· AI기반 사람동작인식 및 기계독해 플랫폼
	김현수	대표	㈜슈퍼브에이아이	· AI 기반 데이터 가공 솔루션 및 AI데이터 구축·가공
	남세동	대표	㈜보이저엑스	· 딥러닝 기반 응용 솔루션 및 서비스
	박민우	대표	㈜크라우드웍스	· 크라우드소싱 오픈 플랫폼 개발 및 데이터 구축·가공
	신호욱	대표	㈜셀렉트스타	· 크라우드소싱 플랫폼 개발 및 데이터 가공
	오성식	이사	㈜아크릴	· 공감형 인공지능 플랫폼 및 서비스
	윤석원	대표	㈜테스트웍스	· AI데이터 구축·가공 및 SW Testing · ICT Social Venture
	이영수	이사	㈜마인즈앤컴퍼니	· AI기업 컨설팅 및 AI데이터 분석 모델링 · Kaggle Competitions Master

* 가나다순으로 정리

〈 심층 인터뷰 참여 전문가 명단(계속) 〉

구분	성명	직위	소속	전문 분야
AI 활용 전문기업 (8개社)	배윤정	대표	㈜메디플러스 솔루션	· ICT 기반 건강관리 서비스 및 알고리즘
	양상환	리더	네이버 D2SF	· 국내 기술 스타트업 엑셀러레이터
	윤덕호	이사	㈜코난 테크놀로지	· 자연어처리·자연어 이해 기반 AI솔루션
	인치원	실장	카카오M	· AI 기술 개발 및 전략 기획 · AI스타트업 투자
	정규환	이사	㈜뷰노	· AI기반 진단 보조 의료기기 개발
	채규열	책임	㈜LG전자 CTO 컨버전스센터	· 임베디드 카메라AI SW
	하정우	리더	네이버 Clova	· AI 기술 및 활용 서비스
	한상기	대표	테크프론티어	· AI 기반 사업전략수립 및 정책 컨설팅
학계 및 연구계 (4명)	구영현	교수	세종대 컴퓨터공학과	· 빅데이터·AI 기반 기술 및 플랫폼 연구
	김종엽	교수	건양대학교병원	· 의료 빅데이터 분석 및 AI기반 의료융합과학 · 이비인후과 전문의
	김학래	교수	중앙대 문헌정보학과	· AI 및 데이터 사이언스 · 공공데이터전략위원회 실무위원
	이명진	연구위원	환경정책평가연구원	· AI기반 환경평가시스템 개발 및 기후변화 연구

* 가나다순으로 정리

AI INSIGHT REPORT

- 제1호(2019.9), 「인공지능 현장에서 듣는 AI데이터 중장기 구축 방향」

AI INSIGHT REPORT VOL.01

인공지능 현장에서 듣는 AI데이터 중장기 구축 방향

NIA 한국정보화진흥원
NATIONAL INFORMATION SOCIETY AGENCY