한국보건복지인력개발원
KOREA HUMAN RESOURCE DEVELOPMENT INSTITUTE
FOR HEALTH & WELFARE

국가인적자원개발컨소시엄
CHAMP Consortium for HRD Ability Magnified Program

의약품 빅데이터 분석 과정

- 1 차시 -

빅데이터 기본 이해

1차시. 빅데이터 기본 이해

• 학습목표

1. 빅데이터의 정의와 개념을 설명하고 빅데이터의 가치를 이해할 수 있습니다.
2. 빅데이터의 발전을 위한 국내외 정책과 동향을 파악할 수 있습니다.

• 학습하기

1. 빅데이터의 개념 및 가치 이해

빅데이터는 의약품 사업분야에서 하나의 추세로 자리잡고 있으며, 빅데이터를 통해 기업의 비즈니스 운영 방식을 전방위적으로 혁신하고 있습니다.

→ 비용을 최적화하거나 새로운 수익원을 창출

→ 경쟁력을 높일 훌륭한 기회를 제공

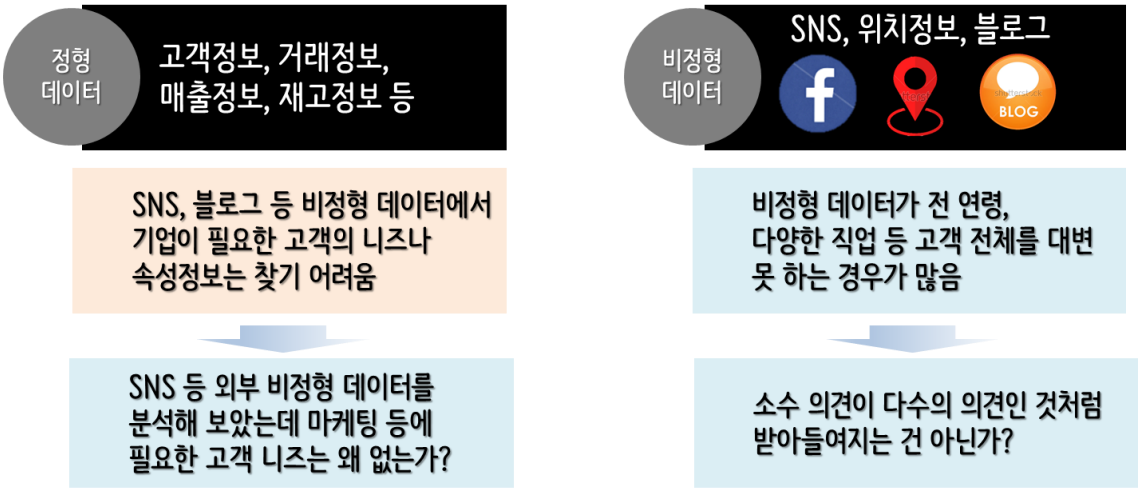
의약품 빅데이터는 의약품데이터연구에 있어서, 신약개발과 임상시험, 개인맞춤형의료, 약품감시 및 약품영향분석 그리고 비용절감과 효능연구에 사용하고 있습니다.

<의약품 빅데이터의 주요 활용가치>

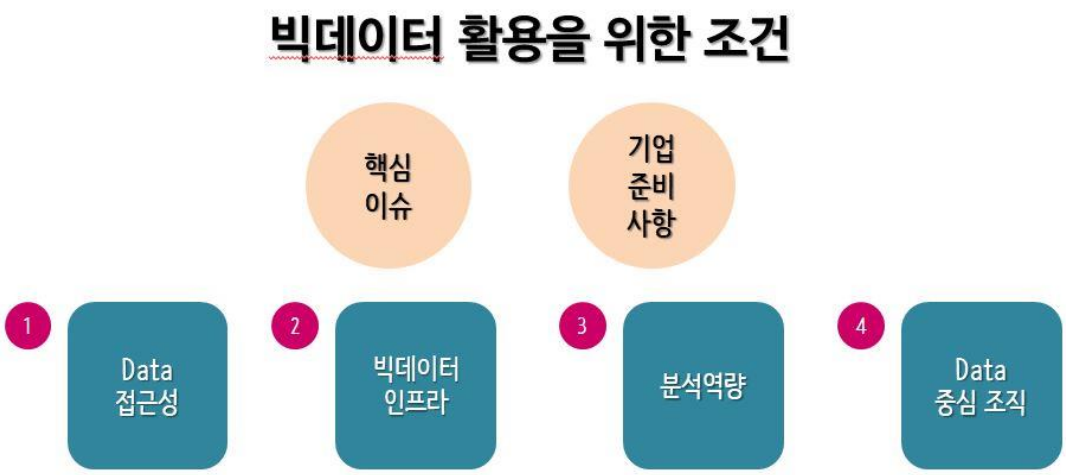
빅데이터의 개념은 일명 3V로 일컬어지는 volume(양), variety(다양성), velocity(속도)로 정의되었습니다. 이후 veracity(정확성)과 value(가치)가 추가되어 5V를 가지는 데이터 특성으로 이해되고 있습니다.

빅데이터 특성	함정	기업의 고민
VOLUME (크고)	Big Data라고 부를 수 있는 Volume의 기준에 대한 국제적 합의가 없음	✓ 기업 내에 있는 데이터는 정말 빅 데이터인가? ✓ 사이즈는 얼마나 커야 하는가?
VARIETY (다양하고)	외부, 비정형 데이터를 찾고 있으나 어떤 데이터가 필요한지 어디에도 기준이 없음	✓ 외부에 있는 무수히 많은 데이터 중 우리 기업에 필요한 데이터는 무엇이고 어디서 구할 수 있는가?
VELOCITY (빠르고)	현재 기업의 IT 시스템 구조로는 정합성이 확보된 실시간 데이터 제공은 어려움 Ex) 콜센터 상담 메모도 각 지역 콜센터에서 데이터 센터로 야간에 이관돼야 분석 가능	✓ 실시간 제공을 위해서는 대규모적인 IT 투자가 이루어져야 하는데 가치가 있는 것인가?

데이터의 형태



이러한 빅데이터를 가치 있게 활용하기 위해서는 여러 가지 조건과 준비가 필요한데, 이를 빅데이터 활용을 위한 조건과 핵심이슈 그리고 준비사항 측면에서 정리해 본다면, 다음의 내용을 참고할 수 있습니다.

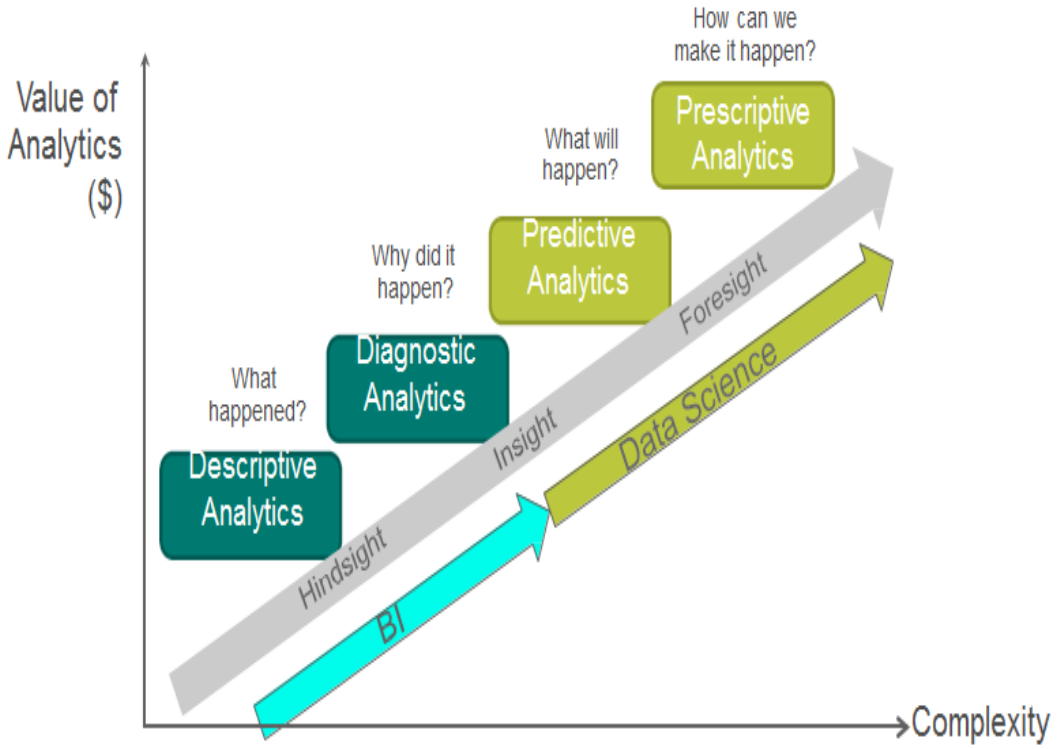


빅데이터 활용을 위한 조건	핵심 이슈	기업 준비사항
Data 접근성	- 외부의 제 3자 데이터 활용 가능성 - 내·외부 데이터의 체계적 결합 및 전사적 이용 가능성	- 프라이버시, 보안, 지적재산권, 법적 책임관련 사전 준비 - 외부 DB의 내부 활용방안
빅데이터 인프라	- 클라우드 기반 통합 분석 시스템 - 전사적 데이터 통합 활용 체계	- 분산된 데이터의 클라우드 기반 통합 데이터 공유 프로세스 정립
분석역량	- 대용량 데이터 분석기술(하둡 등) - 실시간 기반분석, 시각화 S/W 등	- 내부 DB와 결합분석을 통한 Warning System 구축 (VoC: Voice of Customer 등) - 실시간 의사결정 지원 방안
Data 중심 조직	- 전문적 분석조직 및 전문 인력 양성 - 데이터 기반 의사결정 조직 구조	- 빅 데이터 분석 전문 조직 검토 - Insight를 끌어낼 수 있는 전문가 채용

즉, 데이터 접근성 측면에서 지적재산권과 개인정보문제 등 법적책임관련에 대한 교육이 필요하고, 빅데이터 인프라 측면에서 흩어져있는 데이터들의 통합 활용체계구축 능력이 필요합니다. 실제로 현존하는 데이터의 대부분은 구조화 되어있지 않은 상태로 여러 기관 등에 쌓여있기 때문에, 이런 데이터를 쓸모 있는 상태로 자산화 하기 위해서는 운영 인프라에 대한 역량이 필요합니다. 데이터가 대용량 빅데이터가 됨에 따라, 스몰데이터의 분석기술과는 다른 분석역량이 요구됩니다.

빅데이터 분석의 가치는, 데이터 기반 Evidence를 통해 미래를 예측하고 그 미래가 우리사회에 이익이 되는 방향으로 펼쳐질 수 있도록, 지금 어떠한 의사결정을 할 것인가에 도움이 될 때 빛을 발하게 됩니다.

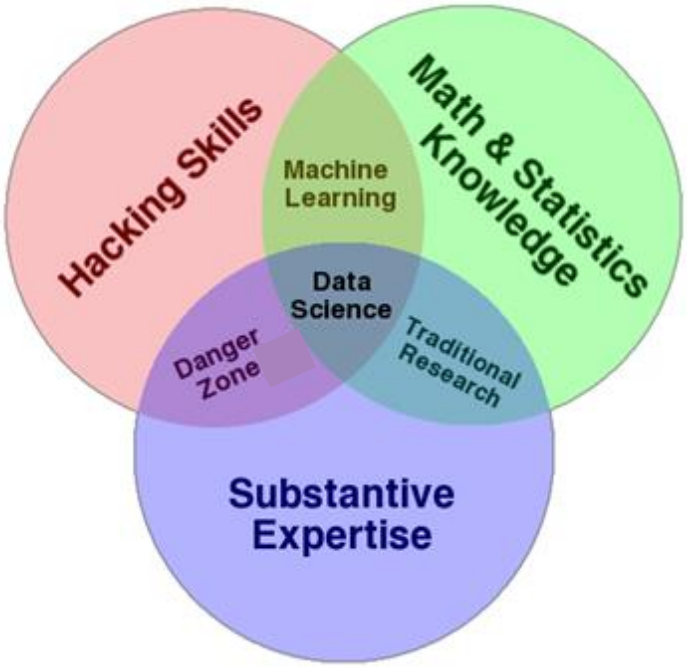
<빅데이터 분석의 가치>



※ 출처:한국보건복지인력개발원 '4차산업혁명과 창의적사고' 강의교재

의약품 빅데이터를 제대로 이해하고 활용하기 위해서는 통계학, 경제학, 정보 기술, 수학 등 다학제적(multidisciplinary) 이해가 필요하고, 이 외에도 통합적 사고와 통찰력을 갖추어야 합니다.

Alluvium의 CEO이자 설립자 이며, Machine Learning co-author로 널리 알려진 미국의 데이터사이언티스트 'Drew Conway의 [data science venn diagram]'을 통해서도 다학제적 균형 있는 역량의 필요성을 확인할 수 있습니다.



※ 출처 : Data science venn diagram/ Drew Conway 2016

위의 데이터 과학의 밴다이어그램에서 필요한 역량에서 알 수 있듯이, 여러분이 앞으로 빅데이터를 활용하기 위해서는, 다음의 3가지 스킬에 대한 내용을 두루 갖추어야 합니다.

첫 번째, Substantive Expertise, 의약품 분야에서의 전문지식입니다.
본인이 실무를 하고 있는 분야에서의 전문성으로, 의약품 분야에 대한 도메인 날리지(Domain knowledge)와 전문분야의 배경지식, 실무에서 습득된 전문적인 데이터에 대한 내용의 이해를 수반합니다. 이는 의약품 관련업무에 재직중인 제약업계, 연구계, 학계 종사자등 본인의 실무분야에서의 전문성을 의미하여, 이미 습득하고 있거나, 습득중인 영역입니다.

두 번째, Hacking Skills입니다. 여기서의 해킹 스킬은 우리가 흔히 생각하는 은행해킹이나 개인정보 보완 해킹과 같은 나쁜 스킬이 아닌, 컴퓨터 활용기술입니다. 즉, 컴퓨터를 이용하여 의약품관련 원하는 정보를 찾아내고, 명령어를 실행하고 데이터를 활용하고 자료문서를 작성 하는 등의 데이터 기술을 의미합니다. 의약업계 종사자들이 직무를 하고 있는 현업과정에서 이미 습득하고 입사했거나 그때그때 습득 중인 기술로서 큰 문제가 되는 경우는 드뭅니다.

마지막 세 번째, Math & Statistics Knowledge입니다. 밴다이어그램에서 눈 여겨 봐야 하는 대목은 해당분야의 전문성(Substantive Expertise)과 컴퓨터 활용능력(Hacking Skills)만 있고 수학적 통계적지식(Math & Statistics Knowledge)이 없는 경우를 '위험영역(Danger Zone)'으로 규정했다는 지점입니다. 즉, 의약품 빅데이터를 다룰 수 있는 전문인재가 되기 위해서는, 수학적 통계적인 교육이 반드시 필요하다는 의미입니다. 수학적 논리와 체계적인 사고력으로 대변되는 기술과 통계적인 지식이 없다면, 아무리 의약품 전문분야지식이 출중하고 데이터 활용 기술이 뛰어날지라도, 그 데이터를 분석한 결과를 잘못 이해하고 해석하며 엉뚱하게 적용하여, 결국 잘못된 의사결정(Wrong decision making)을 초래할 수 있다는 의미입니다. 완벽하게 추출되고 구조화 된 의약품분석결과를 받아본다 할지라도, 예를 들어 선형 회귀 분석을 실행하고 계수 R로 나타내온 분석결과를 받아본다 해도, 수학적 통계적 지식이 없다면, 이들 분석결과가 의미하는 바를 이해하지 못합니다.

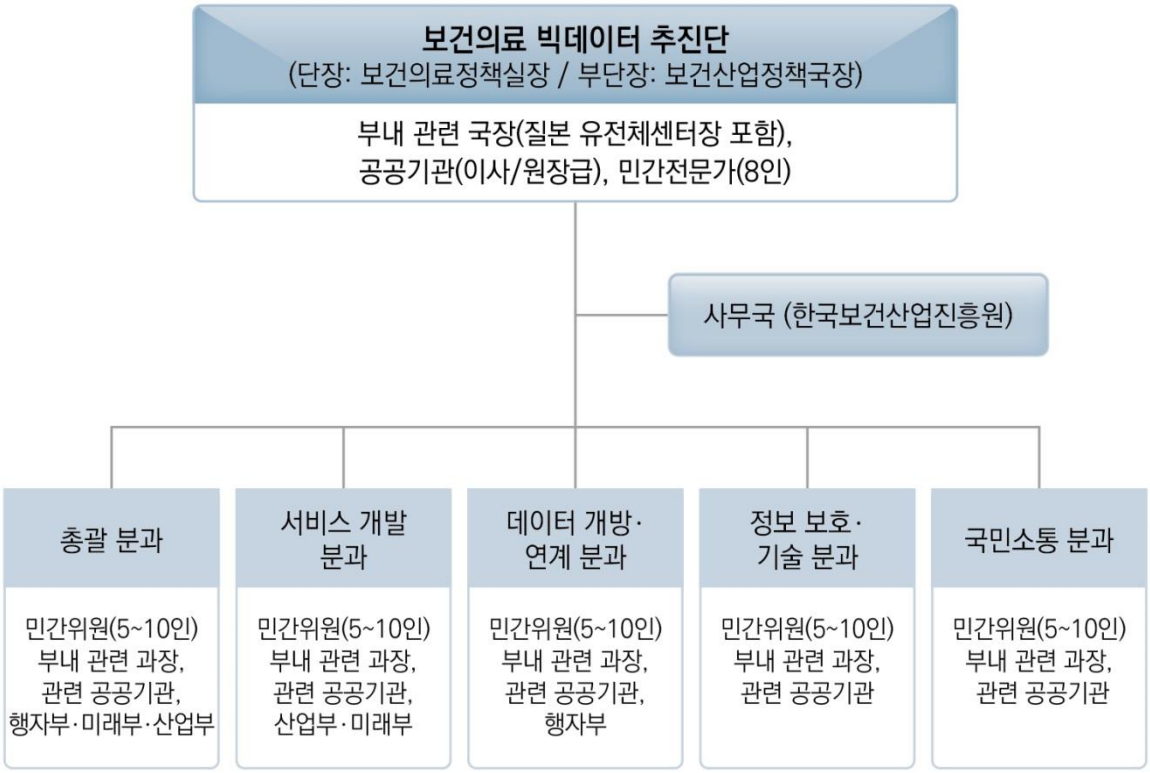
그러므로 의약품 빅데이터 과정에서는 수학적 통계적 분석과정이 필수적이며, 본 사이버 과정을 통해 여러분이 습득하게 될 통계과정은 바로 그런 측면에서 의미가 있습니다.

2. 빅데이터 국내외 정책 및 트렌드

1. 국내 정책 및 트렌드

의약품 빅데이터에 대한 국내정책을 파악하기 위해서는 보건복지부의 빅데이터 활용확대 추진전략을 참고할 필요가 있습니다.

2017년 3월, 보건복지부는 4차 산업혁명 대응과 국민건강 증진을 위해 빅데이터 추진단을 발족하고, 보건의료 빅데이터 추진전략 수립과 국가차원의 로드맵을 마련하겠다고 밝혔습니다.. 추진단은 보건의료 빅데이터 활용 체계 마련을 위한 비전·목표·추진전략 등을 수립하고, 관련 법·제도 개선대책, 전문인력 양성방안, 전담 거버넌스 마련 등을 다루며, 건강증진·질병예방, 보건의료 가치향상, 미래 보건의료 설계 등을 주요 방향성으로 하여, 민·관 데이터 수요를 발굴하고 서비스 모델을 개발한다는 목표를 가지고 있습니다.. 이렇듯 빅데이터는 국민 건강증진과 보호를 위한 핵심기술로 평가받고 있습니다. 특히, 정부는 데이터 연계시스템구축, 기관 간 분석자료 공유·활용 네트워크, 보건의료 빅데이터 활성화에 집중할 예정으로, 보건의료빅데이터 가치창출의 기조를 여실히 나타내고 있습니다.



※ 출처 : 보건의료빅데이터 추진단 (2017. 보건복지부)

정부는 2016년 12월, 관계부처 합동의 [제4차 산업혁명에 대응한 지능정보 사회 중장기 종합대책]을 발표했습니다.. 종합대책에서는 급속도로 발전하는 지능정보기술이 ICT 산업 뿐 아니라 모든 미래 산업에 근본적인 영향을 미쳐, 국가 경쟁력을 판가름한다는 인식하에, 우수한 지능정보기술을 선점하여 향후 고부가가치 지능정보기술 생태계 기반을 구축하는 것이 중요하다는 점을 강조하였습니다..

<지능정보기술의 개념>

※ 출처 : 제4차 산업혁명에 대응한 지능정보사회 중장기 종합대책(관계부처 합동 2016.12월)

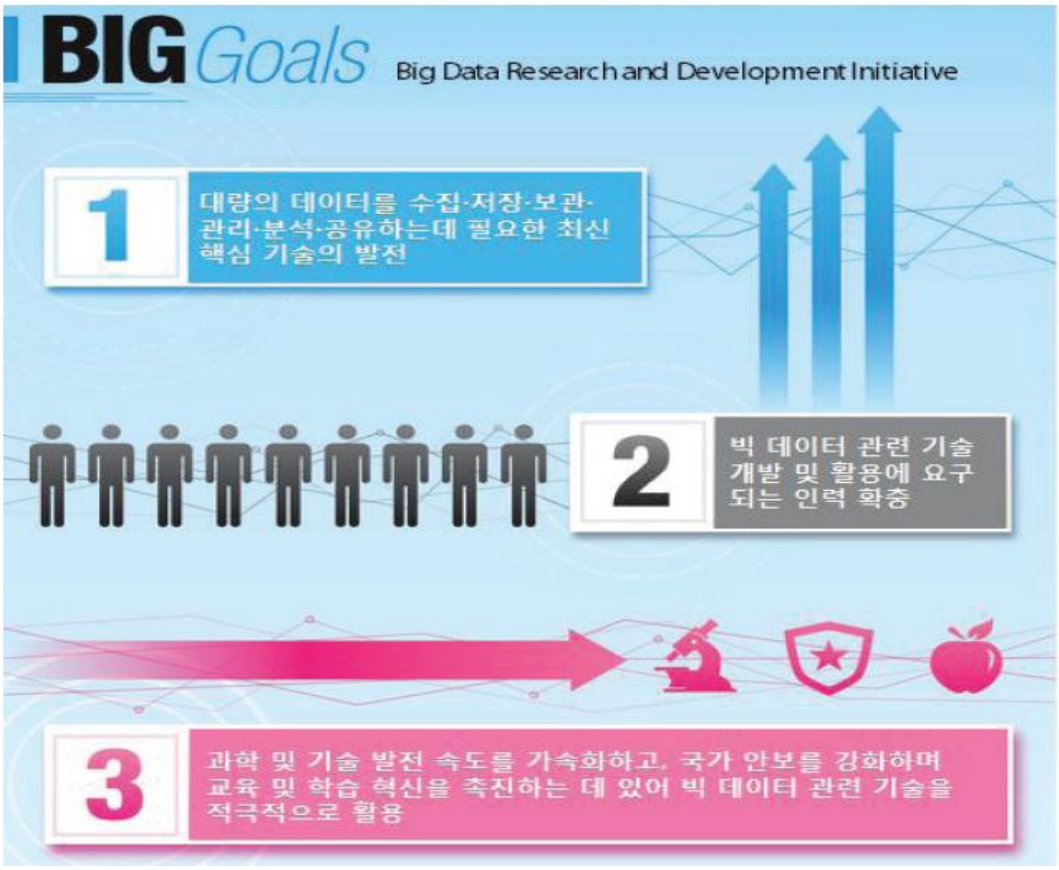
현재의 지능정보기술에 대해서는, 국내와 선진국간 격차는 분명히 존재한다고 평가하고, 지능정보기술 생태계를 선점하지 못할 경우, 글로벌 ICT 기업들이 지능정보 생태계를 독점하고 국내 기업들은 이에 종속되어 혁신적이고 주도적인 기업활동에 한계가 있게 된다는 우려를 담았습니다. 지능정보화 촉진을 위한 정책 목표로는, 공공서비스 및 민간산업 전반에 지능정보기술 도입을 조기 확산하여 생산성 향상 및 국가경쟁력을 확보한다는 내용을 담았습니다. 특히, 의료, 제조, 금융 등 기존산업이, 데이터와 지능정보기술에 기반한 맞춤형 제조·서비스 산업으로 변모하여 고부가가치 창출을 하고, 국민에게 안전하고 편리한 고품질의 지능화된 공공서비스를 제공하도록 한다는 것입니다.

2. DID(Digging into Data Challenge)

미국의 의약품 빅데이터 정책은 미국정부의 '빅데이터 R&D 이니셔티브'를 통해 알 수 있습니다.

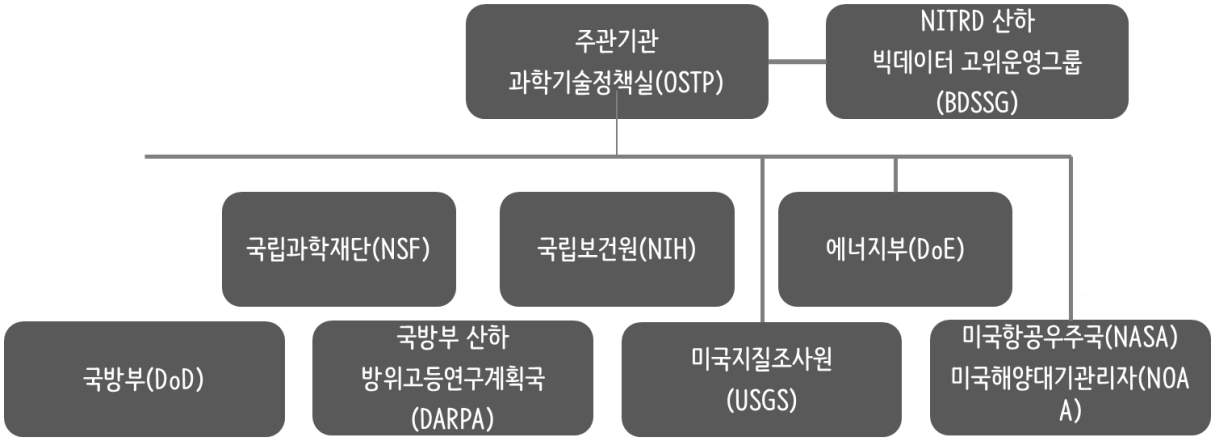
2016년 5월 23일, 미국 정부는 연방정부의 빅데이터 연구 개발(R&D) 전략 계획, "The Federal Big Data Research and Development Strategic Plan"을 공개했습니다. 미국정부는 빅데이터 기술을 개발하고, 빅데이터 응용프로그램 등을 활용하는 차세대 데이터 과학자를 양성하기 위해 2012년 빅데이터 연구개발 이니셔티브(Big Data Research and Development Initiative)를 시작한 바 있으며, 백악관에 최고 데이터 과학자를 채용하는 등의 노력을 기울여 왔습니다. 이번에 발표한 전략계획은 신흥 빅데이터의 기능을 강조하고 연방 빅데이터 연구개발 계획을 확장하기 위한 지침으로, 빅데이터 연구개발 이니셔티브의 중요한 이정표가 될 것이라고 미 정부는 밝히고 있습니다.

<'빅데이터 R&D 이니셔티브'의 세부목표>



※ 출처 : 미국 과학기술정책실(OSTP, 2012년 3월)/한국인터넷진흥원 인터넷&시큐리티이슈 (Aug.2012)

<'빅데이터 R&D 이니셔티브' 참여기관>



※ 출처 : 미국 과학기술정책실(OSTP, 2012년 3월)/한국인터넷진흥원 인터넷&시큐리티이슈 (Aug.2012)

특히 빅데이터 사용 장려를 위해서는 적용프로그램인 'DID(Digging into Data Challenge)'를 통해 디지털 서적이나 신문, 웹 서치 데이터, 음성 기록 등 다양한 종류의 대규모 데이터를 활용할 수 있는 연구방법을 개발하고 있습니다.

※ 출처 : https://diggingintodata.org/(2017.12)

미국정부의 빅데이터 관련 여러 정책 중 의약품 빅데이터 관련하여 우리가 주목할 지점은 미국 국립보건원 전략계획인 'NIH-Wide strategic plan'이며, 주요 내용은 아래와 같습니다.

1) 정밀의학의 적용을 통해 수천 명의 암 환자들의 생존율 증가 확인

2) 다양한 인플루엔자 바이러스 변종에 대해 광범위한 항체 결합 반응을 유도하는 후보 백신의 임상시험 진입

* 포괄적인 독감 백신(universal flu vaccine)을 향한 필수과정

3) 보건문제에 있어서 차별성을 경험하는 사람들의 건강 증진과 질병 예방을 위한 효과적이고 개인 맞춤형인 행동/사회적인 치료방안 개발

4) 약물 유전체학을 실제 임상에 적용하여 다수의 약물 사용에서 개선된 결과 도출

5) 2016년 남아프리카공화국에서 시작될 것으로 예상되는 새로운 HIV 백신 유효성 시험에서, HIV 바이러스 침입 후 최소 50% 이상의 보호 결과 확보

6) NIH가 지원한 임상시험을 통해 지금까지 임상적으로 유용하다고 여겨진, 적어도 6가지 이상의 치료법이 실제로는 가치가 없음을 증명

7) 새로운 구조생물학적 방법들로 약물 스크리닝과 최적화 과정을 개선

8) 12가지 희귀질환에 대한 FDA 승인 치료법에 직접적으로 기여

9) 모바일건강(mHealth) 기술이 건강 증진 및 질병 예방에 효과적이라는 증거 확보

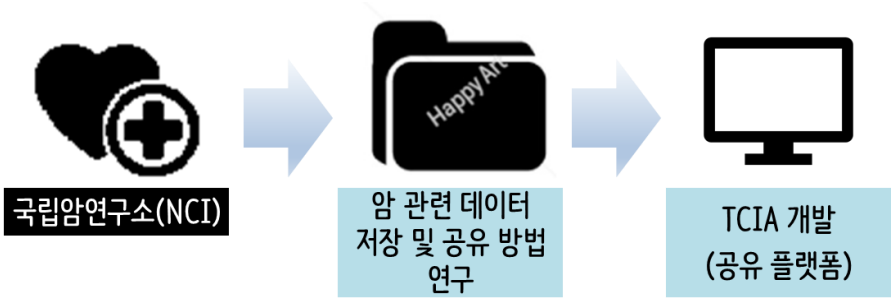
10) 실시간으로 혈중 알코올 농도를 측정할 수 있는 입는 바이오센서를 개발해 알코올 관련 부상과 질병 예방에 효과적이라는 사실 입증

11) 척추손상 환자들의 마비를 되돌리고 일부 정상기능을 회복하는 기술 개발

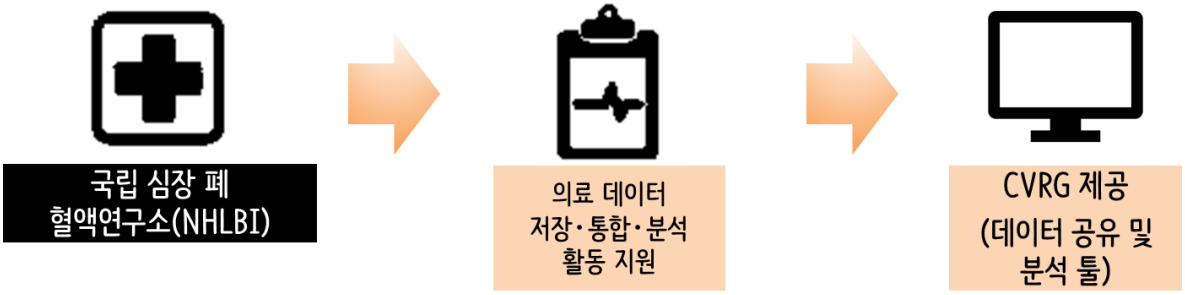
12) 호흡기 합포체 바이러스에 대한 백신의 유효성 테스트를 수행하여 아동기 폐렴의 중요한 원인 중 하나에 대한 해결책 마련

13) 인공췌장 연구로 저혈당 위험 없이 당뇨병을 보다 잘 관리할 수 있는 진전된 시도 마련

14) NIH가 의과학 연구를 어떻게 잘 지원하고 과학적인 방법을 적용할 것인지에 대한 베스트 모델 연구기관으로서의 역할을 공고히 함.



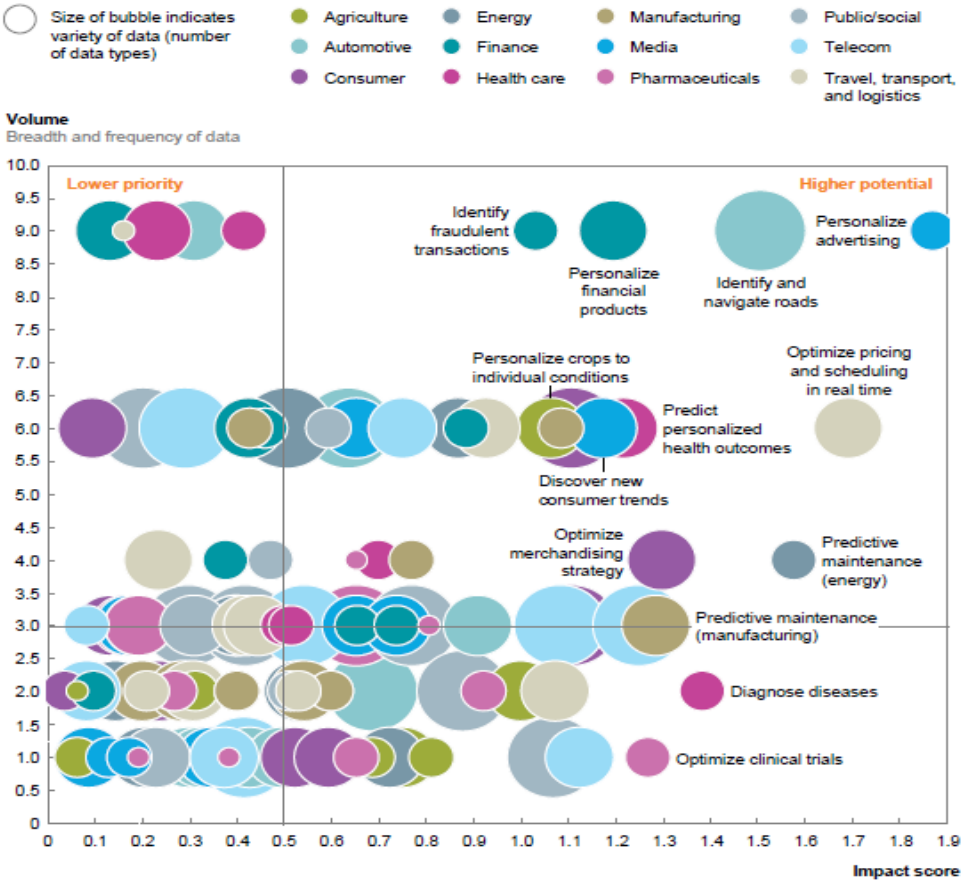
국립보건원 산하 국립암연구소(NCI)는 암 관련 데이터의 저장 및 공유 방법을 연구하고 있으며, 의료 이미지 및 영상 데이터 공유 플랫폼인 'TCIA'를 개발하였습니다. 이를 통해, 의사들의 암 치료와 연구지원 및 환자들의 암 발견 가능성 향상을 위해 노력하고 있으며, 유전자 분석 기술을 응용, 대규모의 암 세포 관련 데이터를 축적하기 위해 'TCGA'프로젝트 운영하고 있습니다.



국립보건원 산하 국립심장폐혈액연구소(NHLBI)는 의료 데이터의 저장·통합·분석 활동을 지원하고 있으며, 이를 위해 데이터 공유 및 분석 툴인 'CVRG'를 제공하고 있습니다. 국립보건원이 추진 중인 '1000 Genomes Project'를 통해 해독된 약 200테라바이트의 인체 유전자데이터를 공개하고, 클라우드 서비스인 아마존 웹 서비스를 통해 누구나 데이터에 접근할 수 있도록 할 예정입니다.

최근 빅데이터는 의약품 뿐 아니라 헬스케어 등 산업분야 전반에서 활용가치가 높아지고 있어, 각국 정부는 빅데이터 정책에 대한 다양한 변화도입을 검토하고 있습니다.

헬스케어부문 등 산업분야 전반에서 데이터의 활용



※ 출처 : McKinsey Global Institute analysis(2016.12)

3. 영국의 정책 및 트렌드

영국 정부는 데이터 시대를 대비하기 위해, 데이터 역량 강화 전략(A Strategy for UK Data Capability, 2013)을 발표하고, 오픈 데이터 로드맵(Open Data Roadmap, 2015)을 통해 더 많은 데이터 개방과 오픈데이터 재활용에 역점을 두고 있습니다.

의약품 빅데이터 관련하여, 영국의 NICE(National Institute for Health and Care Excellence)는 비용 효과성을 판단하기 위해 삶의 질에 초점을 맞춘 관점으로 모든 의료 개입과 치료를 평가하는 기관으로, 빅데이터 기반 비용효과성을 측정하고 있습니다.

개인 맞춤 의학에서의 빅데이터 활용 목적은 각 환자에 대한 올바른 치료법을 파악하는 것입니다. 인공 지능에 기반한 임상적 의사결정 지원 시스템은 수백만 건의 환자 기록, 게놈 시퀀스, 기타 건강 및 행위 데이터를 조합하여 특정 성질을 가진 특정 개체에 가장 효과적인 치료 과정을 식별할 수 있게 합니다.

의약품 등 치료비용에 이러한 통찰력을 적용하면 가장 비용 효과적인 의약품을 알 수 있

을 뿐 아니라, 특정 개인은 본인에게 효과가 없는 치료법을 피할 수도 있습니다. 이는 의약품, 수술 및 기타 개입의 효능을 극대화하는 동시에, 의료 폐기물과 유해한 부작용을 감소시킵니다. 실제 사례로, University College London의 연구원은 슈퍼컴퓨터 시뮬레이션을 사용해 50개의 약물 중에서 특정 유방암 돌연변이에 대한 최상의 치료법을 결정했습니다. 대규모 환자 데이터 통합은 세부적인 건강 및 질병 상태에 대한 환자 치료부문에서 새로운 통찰력을 지원할 것으로 기대됩니다.

지난 몇 년간 데이터기반 예측 분석은 제약 부문에서도 상당한 효과를 이끌어냈습니다. Merck가 후원한 과제에서, 딥러닝에 힘입은 알고리즘을 활용하여, 약물 개발에 쓰일 수 있는 잠재적인 분자를 확인한 바 있습니다. 최근 휴스턴 감리교 팀은 병원의 환자 차트를 인간보다 30배 빠른 속도로 분석하여, 유방암 위험을 예측하는 텍스트 분석 알고리즘을 개발하기도 하였습니다.

<딥러닝 알고리즘을 활용한 예측 분석 사례>

- 헬스케어 : 스캔, 생검, 오디오 및 기타 데이터에서 질병 진단
- 통신 : 개인 고객 관련 생애주기 및 변동 위험 예측
- 제약 : R&D 비용과 출시 속도 단축을 위한 실험 결과 예측
- 재무 : 신속하고 편견이 적은 의사결정을 통한 정확한 고객 신용 위험 평가
- 공공 부문 : 복잡한 상호 작용 및 잠재적인 결과를 고려한 공공 정책 결정 (예 : 주택 정책) 최적화