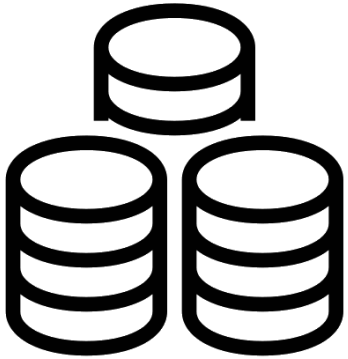
Big Data



Big Data

Big Data에 대한 이해

1. Big Data 개요
2. Big Data 정의
3. Big Data 특징과 의미
4. Big Data 분석 기법
5. Big Data 활용 사례
6. Data Mining 이해
7. Data Science Process
8. 데이터 과학자가 갖춰야 할 능력

1. Big Data 개요

❖ 빅 데이터(big data)란 기존 데이터베이스 관리 도구의 능력을 넘어서는 대량(수십 테라바이트)의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터 집합조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술이다.

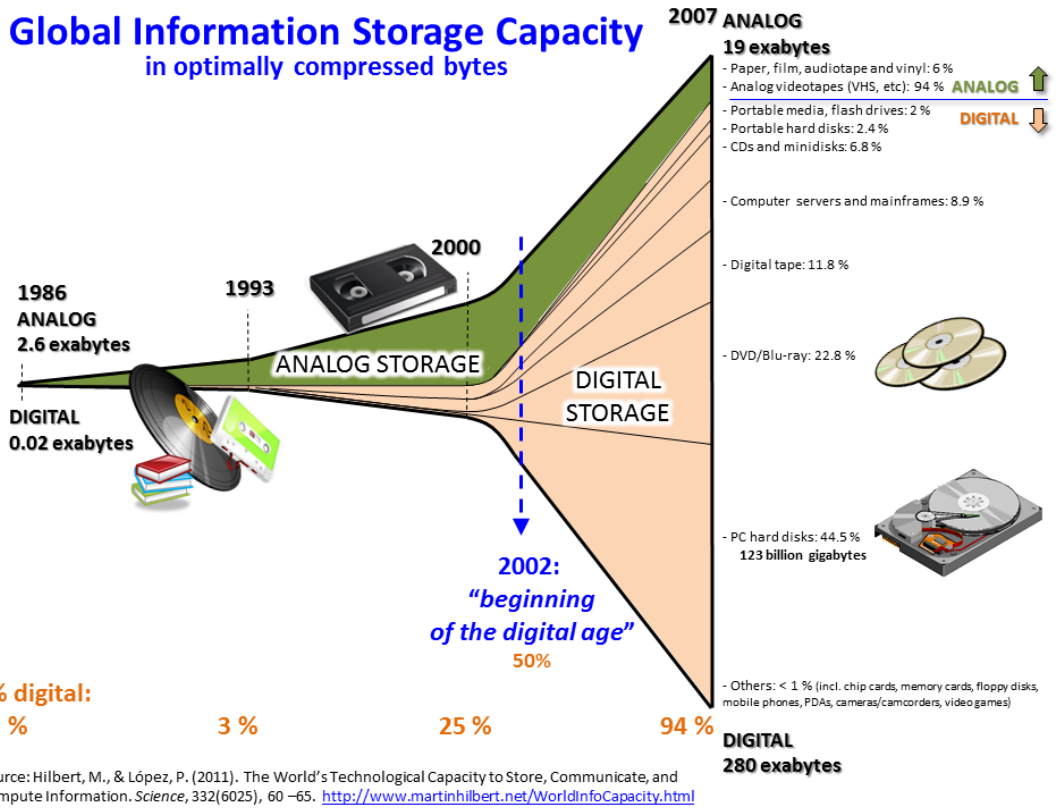
❖ 다양한 종류의 대규모 데이터에 대한 생성, 수집, 분석, 표현을 그 특징으로 하는 빅 데이터 기술의 발전은 다변화된 현대 사회를 더욱 정확하게 예측하여 효율적으로 작동케 하고 개인화된 현대 사회 구성원 마다 맞춤형 정보를 제공, 관리, 분석 가능케 하며 과거에는 불가능했던 기술을 실현시키기도 한다.

❖ 빅 데이터는 정치, 사회, 경제, 문학, 과학 기술 등 전 영역에 걸쳐서 사회와 인류에게 가치 있는 정보를 제공할 수 있는 가능성을 제시하며 그 중요성이 부각되고 있다.

❖ 하지만 빅데이터의 문제점은 사생활 침해와 보안 측면에 있다.

❖ 정의

빅 데이터는 통상적으로 사용되는 데이터 수집, 관리 및 처리 소프트웨어의 수용 한계를 넘어서는 크기의 데이터를 말한다. 빅 데이터의 크기는 단일 데이터 집합의 크기가 수십 테라바이트에서 수 페타바이트에 이르며, 그 크기가 끊임 없이 변화하는 것이 특징이다.



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

2. Big Data 정의

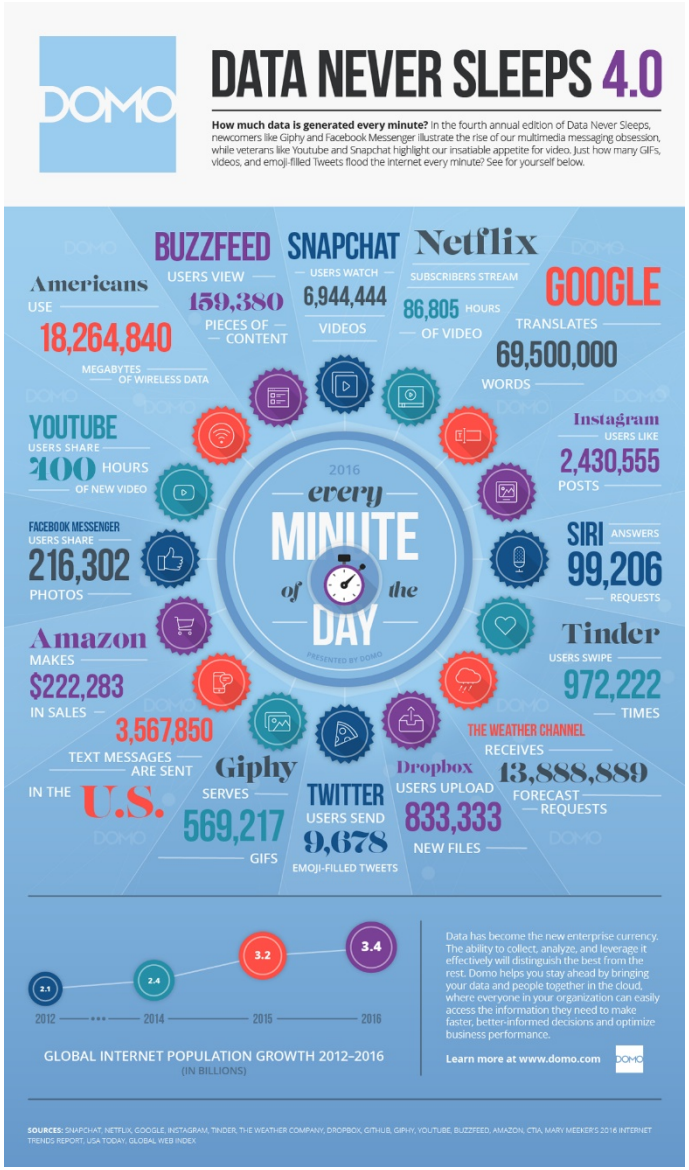
- 위키백과의 편집 현황의 시각화 자료(IBM 작성).

  수 테라바이트의 용량을 지닌 위키백과의 텍스트 및 이미지 자료는 빅 데이터의 고전적 사례에 속한다.

2. Big Data 정의

- 1분 동안 인터넷에서 생성되는 데이터의 양

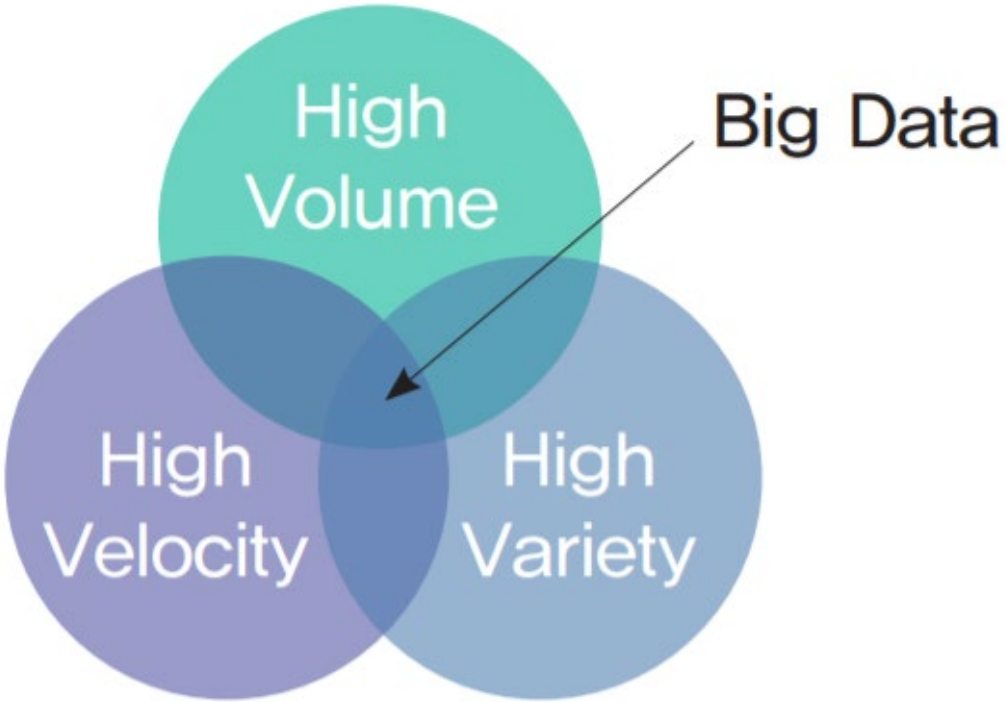
  이미지 출처: https://www.domo.com/blog/data-never-sleeps-4-0/



3. Big Data 특징 과 의미

❖ 특징

미국 유명 IT컨설팅업체인 Gartner는 빅 데이터의 공통적 특징으로 3V, 즉 대용량의 데이터 규모(High-Volume), 빠른 입출력 속도(High-Velocity), 다양성(High-Variety)을 말한다.



3. Big Data 특징 과 의미

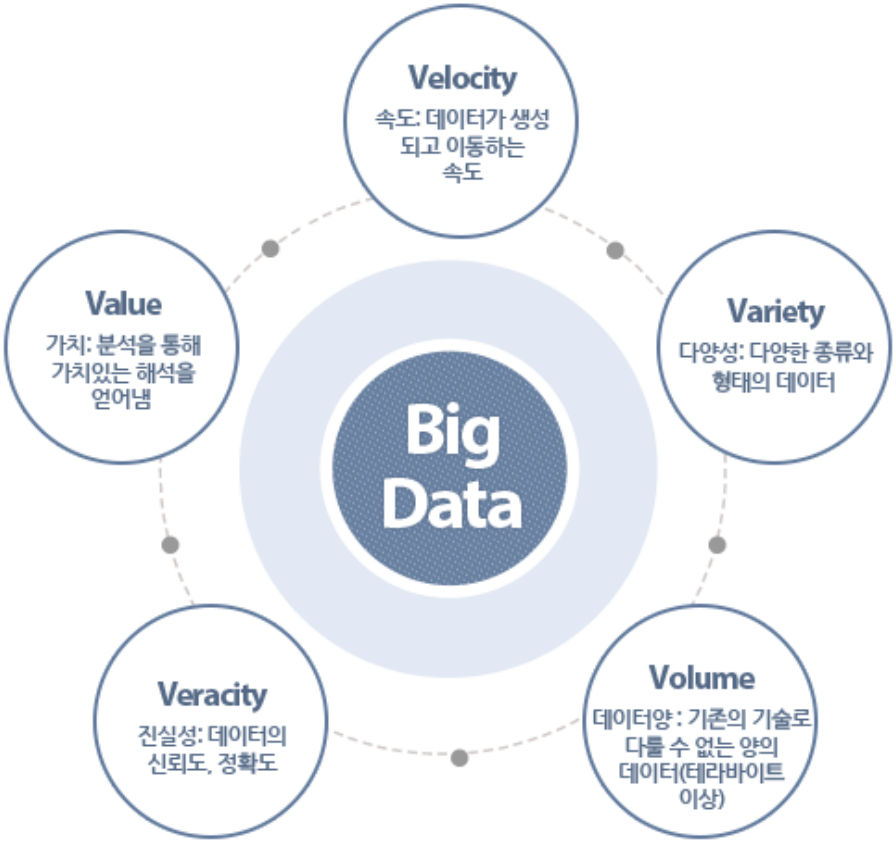
- 빠른 입출력 속도(High-Velocity)는 대용량의 데이터를 빠르게 처리하고 분석할 수 있는 속성이다. 융복합 환경에서 디지털 데이터는 매우 빠른 속도로 생산되므로 이를 실시간으로 저장, 유통, 수집, 분석처리가 가능한 성능을 의미한다.

- 다양성(High-Variety)은 다양한 종류의 데이터를 의미하며 정형화의 종류에 따라 정형, 반정형, 비정형 데이터로 분류할 수 있다.

- 빅 데이터의 특징은 3V로 요약하는 것이 일반적이다. 즉 데이터의 양(Volume), 데이터 생성 속도(Velocity), 형태의 다양성(Variety)을 의미한다.

- 최근에는 가치(Value)나 복잡성(Complexity)을 덧붙이기도 한다.

3. Big Data 특징 과 의미



❖ Big Data 새로운 V

- 정확성(Veracity)

  빅 데이터 시대에는 방대한 데이터의 양을 분석하여 일정한 패턴을 추출할 수 있다. 하지만 정보의 양이 많아지는 만큼 데이터의 신뢰성이 떨어지기 쉽다. 따라서 빅 데이터를 분석하는데 있어 기업이나 기관에 수집한 데이터가 정확한 것인지, 분석할 만한 가치가 있는지 등을 살펴야 하는 필요성이 대두되었고, 이러한 측면에서 새로운 속성인 정확성(Veracity)이 제시되고 있다.

- 가변성(Variability)

  최근 소셜미디어의 확산으로 자신의 의견을 웹사이트를 통해 자유롭게 게시하는 것이 쉬워졌지만 실제로 자신의 의도와는 달리 자신의 생각을 글로 표현하게 되면 맥락에 따라 자신의 의도가 다른 사람에게 오해를 불러일으킬 수 있다. 이처럼 데이터가 맥락에 따라 의미가 달라진다고 하여 빅 데이터의 새로운 속성으로 가변성(Variability)이 제시되고 있다.

- 시각화(Visualization)

  빅 데이터는 정형 및 비정형 데이터를 수집하여 복잡한 분석을 실행한 후 용도에 맞게 정보를 가공하는 과정을 거친다. 이때 중요한 것은 정보의 사용 대상자에 이해 정도이다. 그렇지 않다면 정보의 가공을 위해 소모된 시각적, 경제적 비용이 무용지물이 될 수 있기 때문이다.

❖ 데이터와 그것의 사용 방법에 있어서 빅 데이터와 경영정보학이 구분된다.

- 경영정보학은 대상을 측정하고 경향을 예측하는 등의 일을 하기 위해 고밀도의 데이터로 구성된 기술적 통계를 활용한다.

- 빅 데이터는 큰 데이터 집합으로부터 일정한 법칙을 추론하여 결과 및 행동을 예측하기 위해 통계적 추론과 비선형 시스템 식별(nonlinear system identification)의 일부 개념을 활용한다.

4. Big Data 분석 기법

❖ 빅 데이터의 분석/활용을 위한 분석 기법은 분석 기술, 표현 기술로 나뉜다.

❖ 분석 기술

빅 데이터를 다루는 처리 프로세스로서 병렬 처리의 핵심은 분할 점령(Divide and Conquer)이다. 즉 데이터를 독립된 형태로 나누고 이를 병렬적으로 처리하는 것을 말한다.

- 아파치 하둡(Apache Hadoop) : 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 프리웨어 자바 소프트웨어 프레임워크
- 텍스트 마이닝 : 비/반정형 텍스트 데이터에서 자연 언어 처리 기술에 기반을 두어 유용한 정보를 추출, 가공
- 오피니언 마이닝 : 소셜미디어 등의 정형/비정형 텍스트의 긍정, 부정, 중립의 선호도를 판별
- 소셜 네트워크 분석 : 소셜 네트워크의 연결 구조 및 강도 등을 바탕으로 사용자의 명성 및 영향력을 측정
- 군집 분석 : 비슷한 특성을 가진 개체를 합쳐가면서 최종적으로 유사 특성의 군집을 발굴

➢ 대규모의 정형/비정형 데이터를 처리하는 데 있어 가장 기본적인 분석 인프라로 하둡이 있으며 데이터를 유연하고 더욱 빠르게 처리하기 위해 NoSQL 기술이 활용되기도 한다.

❖ 표현 기술

빅 데이터 분석 기술을 토해 분석된 데이트의 의미와 가치를 시각적으로 표현하기 위한 개술로 대표적인 것으로 R(프로그래밍 언어)

❖ 활용 사례

✓ 2008년 미국 대통령 선거

2008년 미국 대통령 선거에서 버락 오바마 미국 대통령 후보는 다양한 형태의 유권자 데이터베이스를 확보하여 이를 분석, 활용한 '유권자 맞춤형 선거 전략'을 전개했다. 당시 오바마 캠프는 인종, 종교, 나이, 가구형태, 소비수준과 같은 기본 인적 사항으로 유권자를 분류하는 것을 넘어서서 과거 투표 여부, 구독하는 잡지, 마시는 음료 등 유권자 성향까지 전화나 개별 방문을 또는 소셜 미디어를 통해 유권자 정보를 수집하였다. 수집된 데이터는 오바마 캠프 본부로 전송되어 유권자 데이터베이스를 온라인으로 통합 관리하는 '보트빌더(VoteBuilder.com)'시스템의 도움으로 유권자 성향 분석, 미결정 유권자 선별 , 유권자에 대한 예측을 해 나갔다. 이를 바탕으로 '유권자 지도'를 작성한 뒤 '유권자 맞춤형 선거 전략' 을 전개하는 등 오바마 캠프는 비용 대비 효과적인 선거를 치를 수 있었다.

5. Big Data 활용 사례

✓ 대한민국 제19대 총선

중앙선거관리위원회는 대한민국 제19대 총선부터 소셜 네트워크 등 인터넷 상의 선거 운동을 상시 허용하였다. 이에 소셜 미디어 상에서 선거 관련 데이터는 증폭되었으며, 2010년 대한민국 제5회 지방 선거 및 2011년 대한민국 재 보궐선거에서 소셜 네트워크 서비스의 중요성을 확인한 정당들 또한 SNS 역량 지수를 공천 심사에 반영하는 등 소셜 네트워크 활용에 주목했다. 이 가운데 여론 조사 기관들은 기존 여론조사 방식으로 예측한 2010년 제5회 지방 선거 및 2011년 재 보궐선거의 여론조사 결과와 실제 투표 결과와의 큰 차이를 보완하고자 빅 데이터 기술을 활용한 SNS 여론 분석을 시행했다. 그러나 SNS 이용자의 대다수가 수도권 20~30대에 쏠려 있기에, 빅 데이터를 이용한 대한민국 제19대 총선에 대한 SNS 분석은 수도권으로 한정되어 일치하는 한계를 드러내기도 하였다.

✓ 아마존 닷컴의 추천 상품 표시 / 구글 및 페이스북의 맞춤형 광고

아마존 닷컴은 모든 고객들의 구매 내역을 데이터베이스에 기록하고, 이 기록을 분석해 소비자의 소비 취향과 관심사를 파악한다. 이런 빅 데이터의 활용을 통해 아마존은 고객별로 '추천 상품'을 표시한다. 고객 한사람 한사람의 취미나 독서 경향을 찾아 그와 일치한다고 생각되는 상품을 메일, 홈 페이지상에서 중점적으로 고객 한사람 한사람에게 자동적으로 제시하는 것이다. 아마존 닷컴의 추천 상품 표시와 같은 방식으로 구글 및 페이스북도 이용자의 검색 조건, 나아가 사진과 동영상 같은 비정형 데이터 사용을 즉각 처리하여 이용자에게 맞춤형 광고를 제공하는 등 빅데이터의 활용을 증대 시키고 있다.

✓ 2014년 FIFA 월드컵 독일 우승과 '빅데이터'

브라질에서 개최된 2014년 FIFA 월드컵에서 독일은 준결승에서 개최국인 브라질을 7:1로 꺾고, 결승에서 아르헨티나와 연장전까지 가는 접전 끝에 1:0으로 승리를 거두었다. 무패행진으로 우승을 차지한 독일 국가대표팀의 우승의 배경에는 '빅데이터'가 있었다.

독일 국가대표팀은 SAP와 협업하여 훈련과 실전 경기에 'SAP 매치 인사이트'를 도입했다.
SAP 매치 인사이트란 선수들에게 부착된 센서를 통해 운동량, 순간속도, 심박수, 슈팅동작 등 방대한 비정형 데이터를 수집, 분석한 결과를 감독과 코치의 태블릿PC로 전송하여 그들이 데이터를 기반으로 전술을 짜도록 도와주는 솔루션이다. 기존에 감독의 경험이나 주관적 판단으로 결정되는 전략과는 달리, SAP 매치 인사이트를 통해 이루어지는 분석은 선수들에 대한 분석 뿐만 아니라 상대팀 전력, 강점, 약점 등 종합적인 분석을 통해 좀 더 과학적인 전략을 수립할 수 있다. 정보 수집에 쓰이는 센서 1개가 1분에 만들어내는 데이터는 총 12000여개로 독일 국가대표팀은 선수당 4개(골키퍼는 양 손목을 포함해 6개)의 센서를 부착했고, 90분 경기동안 한 선수당 약 432만개, 팀 전체로 약 4968만개의 데이터를 수집했다고 한다.

5. Big Data 활용 사례

✓ 통계학

데이터 마이닝이란 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석의 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술로, 수집되는 '빅 데이터'를 보완하고 마케팅, 시청률조사, 경영 등으로부터 체계화돼 분류, 예측, 연관분석 등의 데이터 마이닝을 거쳐 통계학적으로 결과를 도출해 내고 있다.

대한민국에서는 2000년부터 정보통신부의 산하단체로 사단법인 한국 BI 데이터마이닝 학회가 설립되어 데이터 마이닝에 관한 학술과 기술을 발전, 보급, 응용하고 있다. 또한 국내 · 외 통계분야에서 서서히 빅 데이터 활용에 대한 관심과 필요성이 커지고 있는 가운데 국가통계 업무를 계획하고 방대한 통계자료를 처리하는 국가기관인 통계청이 빅 데이터를 연구하고 활용방안을 모색하기 위한 '빅 데이터 연구회'를 발족하였다. 하지만 업계에 따르면, 미국과 영국, 일본 등 선진국들은 이미 빅 데이터를 다각적으로 분석해 조직의 전략방향을 제시하는 데이터과학자 양성에 사활을 걸고 있다. 그러나 한국은 정부와 일부 기업이 데이터과학자 양성을 위한 프로그램을 진행 중에 있어 아직 걸음마 단계인 것으로 알려져 있다.

5. Big Data 활용 사례

✓ 기업 경영

대규모의 다양한 데이터를 활용한 '빅데이터 경영'이 주목받으면서 데이터 품질을 높이고 방대한 데이터의 처리를 돕는 데이터 통합(Data Integration)의 중요성이 부각되고 있다.
데이터 통합(DI)은 데이터의 추출, 변환, 적재를 위한 ETL 솔루션이 핵심인데 ETL 솔루션을 활용하면 일일이 수많은 데이터를 기업 데이터 포맷으로 코딩하지 않아도 되고 데이터 품질을 제고할 수 있기 때문에 DI는 빅데이터 환경에 꼭 필요한 데이터 솔루션으로 평가받고 있는 단계까지 진입 되었다.
한편 비즈니스 인텔리전스(Business Intelligence, BI)보다 진일보한 빅데이터 분석 방법이 비즈니스 애널리틱스(Business analytics, BA)인데 고급분석 범주에 있는 BA는 기본적으로 BI를 포함하면서도 미래 예측 기능과 통계분석, 확률 분석 등을 포함해 최적의 데이터 기반 의사결정을 가능케 하는 것으로 평가받고 있기도 하다.

5. Big Data 활용 사례

✓ 구글 번역

구글에서 제공하는 자동 번역 서비스인 구글 번역은 빅 데이터를 활용한다. 지난 40년 간 컴퓨터 회사 IBM의 자동 번역 프로그램 개발은 컴퓨터가 명사, 형용사, 동사 등 단어와 어문의 문법적 구조를 인식하여 번역하는 방식으로 이뤄졌다. 이와 달리 2006년 구글은 수억 건의 문장과 번역문을 데이터베이스화 하여 번역시 유사한 문장과 어구를 기존에 축적된 데이터를 바탕으로 추론해 나가는 통계적 기법을 개발하였다. 캐나다 의회의 수백만 건의 문서를 활용하여 영어-불어 자동번역 시스템 개발을 시도한 IBM의 자동 번역 프로그램은 실패한 반면 구글은 수억 건의 자료를 활용하여 전 세계 58개 언어 간의 자동번역 프로그램 개발에 성공하였다. 이러한 사례로 미루어 볼 때, 데이터 양의 측면에서의 엄청난 차이가 두 기업의 자동 번역 프로그램의 번역의 질과 정확도를 결정했으며, 나아가 프로젝트의 성패를 좌우했다고 볼 수 있다.

6. Data Mining 이해

❖ 데이터 마이닝은 대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아 내는 것을 말하며, KDD(데이터베이스 속의 지식발견: knowledge-discovery in databases)라고도 일컫는다.

❖ 개요

통계학에서 패턴인식에 이르는 다양한 계량 기법을 사용한다. 데이터 마이닝 기법은 통계학 쪽에서 발전한 탐색적자료분석, 가설 검정, 다변량 분석, 시계열 분석, 일반선형모형 등의 방법론과 데이터베이스 쪽에서 발전한 OLAP(온라인 분석 처리: On-Line Analytic Processing), 인공지능 진영에서 발전한 SOM, 신경망, 전문가 시스템 등의 기술적인 방법론이 쓰인다.

❖ 응용 분야

신용평점 시스템(Credit Scoring System)의 신용평가모형 개발, 사기탐지시스템(Fraud Detection System), 장바구니 분석(Market Basket Analysis), 최적 포트폴리오 구축과 같이 다양한 산업 분야에서 광범위하게 사용되고 있다.

❖ 단점

자료에 의존하여 현상을 해석하고 개선하려고 하기 때문에 자료가 현실을 충분히 반영하지 못한 상태에서 정보를 추출한 모형을 개발할 경우 잘못된 모형을 구축하는 오류를 범할 수 있다.

❖ 적용 분야

- 분류(Classification) : 일정한 집단에 대한 특정 정의를 통해 분류 및 구분을 추론(예: 경쟁자에게로 이탈한 고객)
- 군집화(Clustering) : 구체적인 특성을 공유하는 군집을 찾는다. 군집화는 미리 정의된 특성에 대한 정보를 가지지 않는다는 점에서 분류와 다르다.(예:유사 행동 집단의 구분)
- 연관성(Association) : 동시에 발생한 사건간의 관계를 정의한다.(예:장바구니안에 동시에 들어 가는 상품들의 관계 규명)
- 연속성(Sequencing) : 특정 기간에 걸쳐 발행하는 관계를 규명한다. 기간의 특성을 제외하면 연관성 분석과 유사하다.(예:슈퍼마켓과 금융상품 사용에 대한 반복 방문)
- 예측(Forecasting) : 대용량 데이터 집합 내의 패턴을 기반으로 미래를 예측한다.(예: 수요예측)

❖ 데이터 마이닝 중요 사항

데이터 마이닝의 가장 중요한 사항은 데이터를 수집하고 가공하는 이유가 무엇인지 이를 통해서 원하는 결과를 얻기 위하여 어떤 기법을 사용해야 하는지에 대한 이해와 선택이다. 데이터 분석은 지하에 묻힌 광물을 찾아낸다는 뜻을 가진 mining란 용어로 부르게 된 것은 데이터에서 정보를 추출하는 과정이 탄광에서 석탄을 캐거나 대륙붕에서 원유를 채굴하는 작업처럼 숨겨진 가치를 찾아낸다는 특징을 가졌기 때문이다.

❖ 통계학과 데이터 마이닝의 유사점

데이터에서 정보를 찾아낸다는 관점에서 보면 데이터 마이닝과 통계학과 매우 비슷하다. 데이터를 탐색하고 분석하는 이론을 개발하는 학문 분야가 통계학이기 때문이다.

데이터 마이닝에서 주고 사용하고 있는 방법론인 로지스틱 회귀분석(logistic regression), 주성분 분석(principal component analysis), 판별 분석(discriminant analysis), 군집 분석(clustering analysis) 등은 통계학에서 사용되고 있는 분석 방법론이다.

❖ 통계학과 데이터 마이닝의 차이점

통계학은 비교적 적은 실험데이터를 대상으로 하는데 반해 데이터 마이닝은 비 계획적으로 축적된 대용량의 데이터를 대상으로 한다.

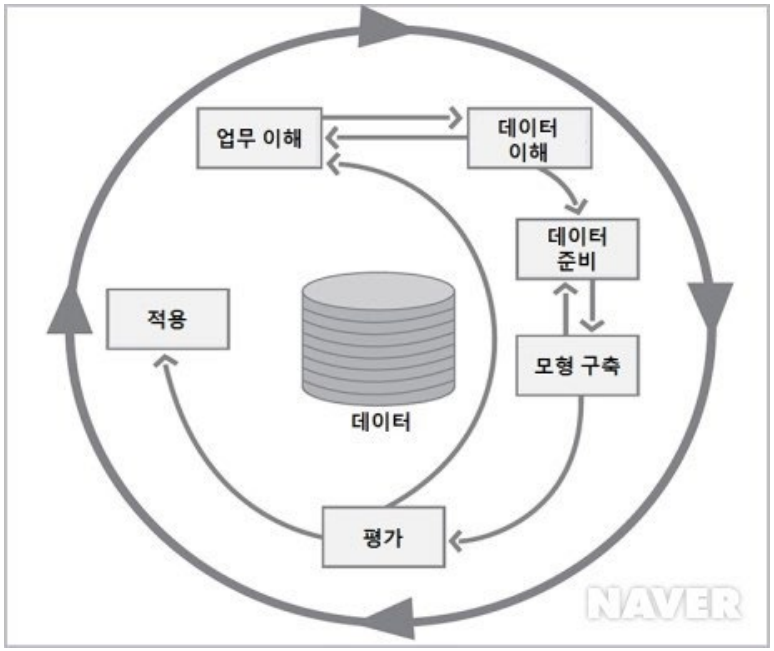
통계학은 추정(estimation)과 검정(testing)이라는 이론을 중시하는 특징을 가졌다면 데이터 마이닝은 이해하기 쉬운 예측모형의 도출에 주목한다.

즉 데이터 마이닝은 기업활동 과정에서 자연스럽게 축적된 대량의 데이터를 분석해 기업 경영에 필요한 가치 있는 정보를 추출하기 위하여 사용한다.

이러한 이유로 데이터 마이닝을 “규모, 속도 그리고 단순성의 통계학(statistics at scale, speed and simplicity)”이라 부른다.

❖ 데이터 마이닝 분석과정

데이터 마이닝은 기업 경영 활동 과정에서 발생하는 데이터를 분석하기 위한 목적으로 개발되었기 때문에 다양한 산업 분야에 공통적으로 적용되는 표준화 처리 과정이 제시되었다. 데이터 마이닝 표준 처리 과정(CRISP-DM, Cross Industry Standard Process for Data Mining)은 비즈니스 이해(Business Understanding), 데이터 이해(Data Understanding), 데이터 준비(Data Preparation), 모형(Modeling), 평가(Evaluation), 적용(Deployment)의 6단계로 구성되어 있다.



7. Data Science Process

❖ 데이터 과학이란 과연 무엇일까? 다음 질문들에 대한 해답을 찾는 과정을 생각해 보자

- 주택 가격을 예측하는 방법은?
- 초등생 자녀의 수학 능력과 상관 관계가 높은 변수는 무엇일까?
- 훌륭한 직원을 뽑는 인터뷰 방법은 무엇일까?
- 웹사이트를 개선하는 방법은?
- TV 광고가 제품 판매에 얼마만큼의 영향을 줄까?
- 온라인 광고에서 클릭 여부를 예측하는 방법은?
- 비싼 와인이 더 맛있을까?
- 대학 진학을 할 때 전공이 중요할까, 학교가 중요할까?

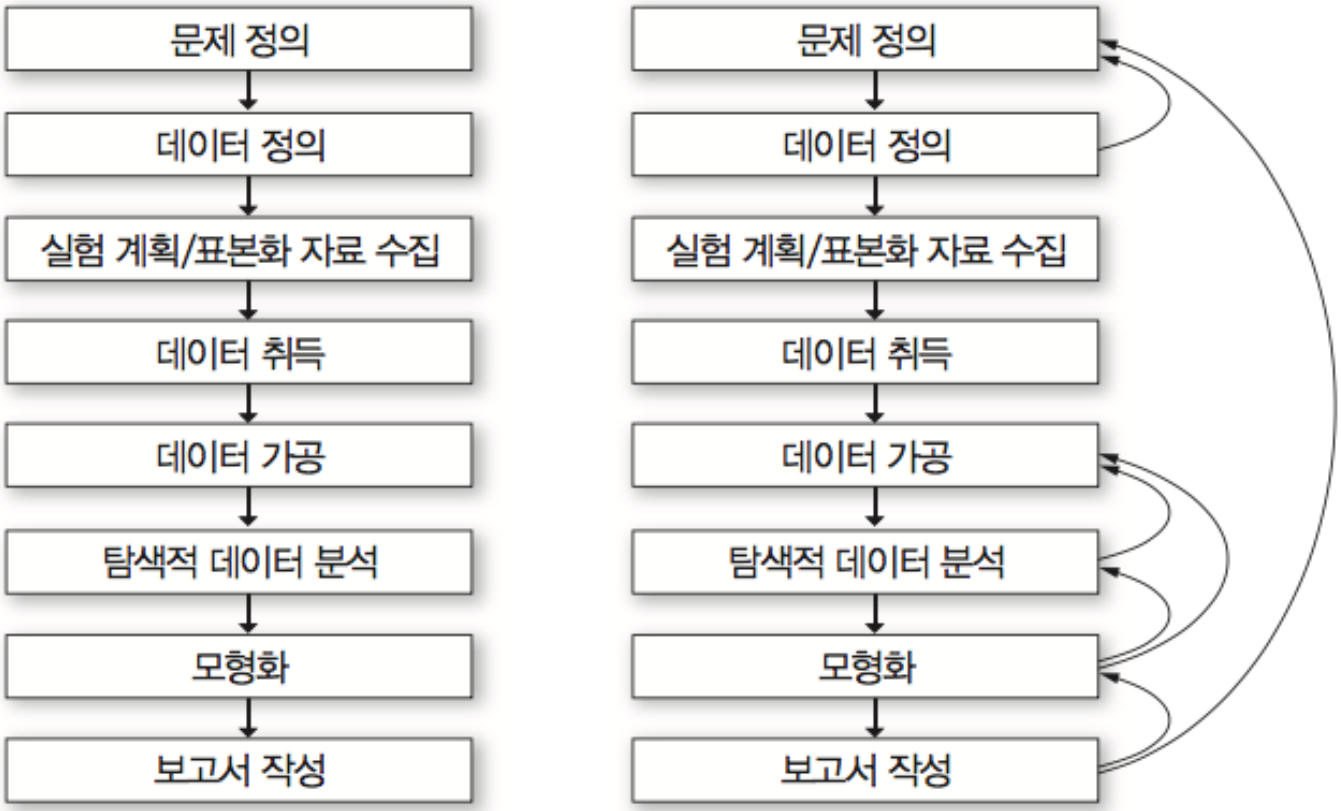
❖ 위와 같은 질문에 대한 답은 여러 방법으로 찾을 수 있다. 하지만 기존에 가지고 있던 데이터나 실험으로 얻은 데이터를 기반으로 답을 찾는다면 그것을 데이터 과학이라고 부를 수 있다. 즉 데이터 과학은 데이터를 사용하여 위와 같은 질문에 합리적인 답을 내릴 수 있게 해주는 활동을 말한다.

❖ 데이터 과학이란 '컴퓨터 도구를 효율적으로 이용하고, 적절한 통계학 방법을 사용하여 실제적인 문제에 답을 내리는 활동'이라고 정의할 수 있다.

7. Data Science Process

❖ 데이터 과학 프로세스

1. 문제 정의(problem definition) : 현실의 구체적인 문제를 명확하게 표현하고 통계적, 수리적 언어로 '번역'하는 작업
2. 데이터 정의(data definition): 변수(variable), 지표(metric) 등을 정의
3. 실험 계획(design of experiment)/표본화(sampling) : 데이터를 직접 수집해야 하는 경우는 보통 두 가지 목적 중 하나다. 첫째는 어떤 처리의 효과를 알아내기 위한 통제 실험, 둘째는 모집단을 대표하는 표본을 얻기 위한 표본화다. 소스 데이터(source data)가 이미 존재하는 경우에는 불필요
4. 데이터 취득(data acquisition) : 다양한 형태의, 다양한 시스템에 저장된 원데이터를 분석 시스템으로 가져오는 활동
5. 데이터 가공(data processing, data wrangling) : 데이트를 분석하기 적당한 표 형태로 가공하는 작업, 데이터 변환
6. 탐색적 분석과 데이터 시각화(exploratory data analysis, data visualization) : 시각화와 간단한 통계량을 통하여 데이터의 패턴을 발견하고 이상치를 점검하는 분석
7. 모형화(modeling) : 모수 추정, 가설검정등의 활동과 모형분석, 예측분석 등을 포괄
8. 분석 결과 정리(reporting) : 분석 결과를 현실적인 언어로 이해하기 쉽도록 번역해내는 작업



데이터 분석 과정에 대한 이상적 관점(왼쪽)과 현실적 관점(오른쪽)

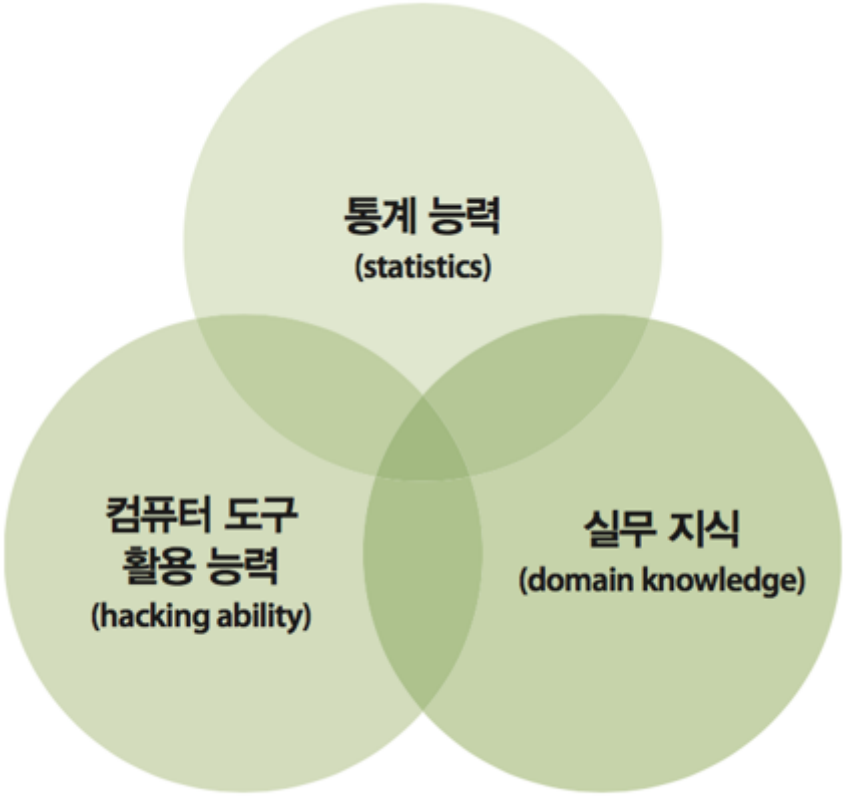
❖ 선형적이면서 직선적인 과정은 현실적이지 않다. 실제 데이터 분석에서는 다음과 같은 경우가 생기게 된다.

- 데이터 수집에서 문제가 생기면 필요한 데이터나 문제를 수정해야 할 수도 있다.
- 탐색적 데이터 분석에서 수집된 데이터의 문제가 발견되기도 한다. 새로 데이터를 수집하거나 문제 가설등을 바꿔야 할 수도 있다.
- 모형화 작업에서 의미 있는 결과를 도출하지 못할 수도 있다. 그러한 경우에도 그 결과 역시 정리하고 알려야 한다.
- 모형화의 결과가 유의미하지 않은 이유를 알아내기 위해 탐색적 데이터 분석을 다시 시행해야 할 수도 있다.
- 일반적으로 모형화 단계에서는 여러 다양한 모형을 시도하게 된다. 모형 결과 자체도 탐색적 데이터 분석을 해야 할 경우가 많다.
- 분석 결과 정리와 공유 후에 여러 피드백을 받게 된다. 이것은 새로운 문제 정의, 데이터 정의 등의 단계로 선순환적으로 이어지게 된다.

❖ 데이터 분석 프로세스를 도식적으로, 선형적으로 이해하는 것은 피해야 할 것이다. 대신에 데이터가 말해주는 내용을 좇아서 능동적으로 적응해나가는, 점진적이면서 순환적(iterative)과정으로 이해하는 것이 더 현실에 가깝다.

8. 데이터 과학자가 갖춰야 할 능력

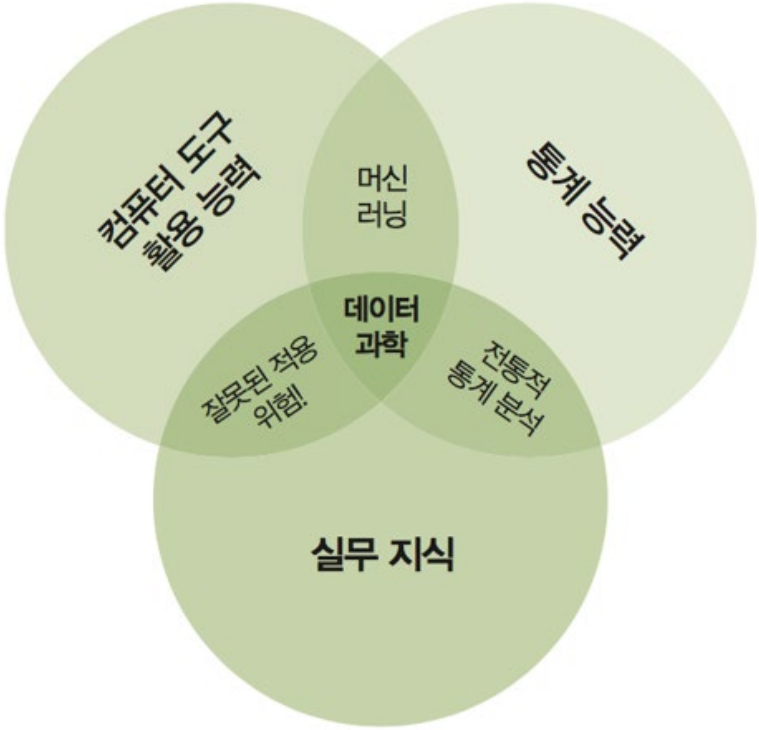
❖ 데이터 과학자가 되기 위해서는 통계 능력, 컴퓨터 도구 활용 능력, 실무 지식이 필요하다.



데이터 과학자가 갖춰야 할 능력 벤다이어그램 1

8. 데이터 과학자가 갖춰야 할 능력

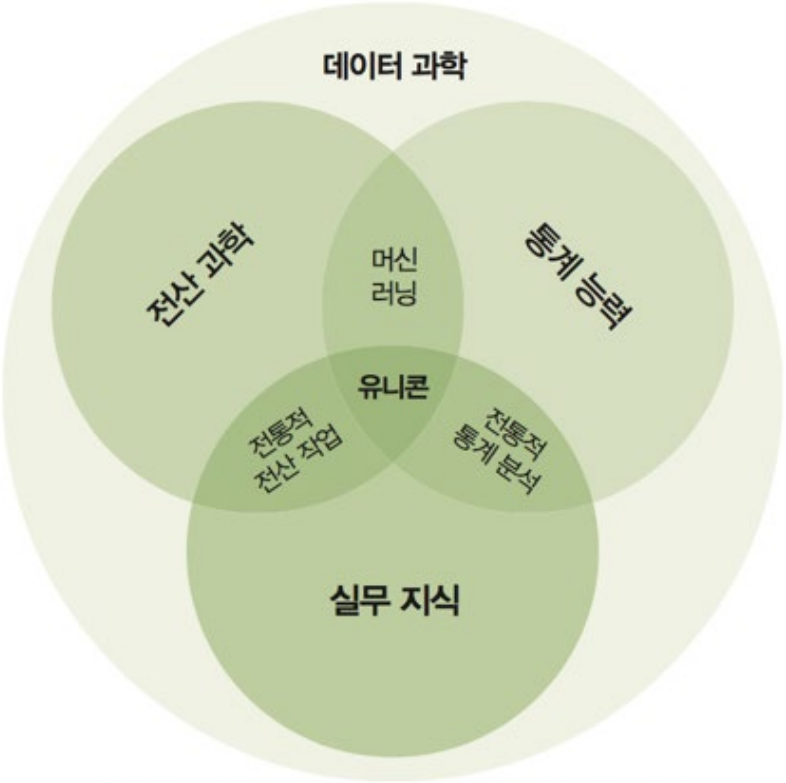
❖ 머신 러닝, 예측 분석을 잘하려면 실무 영역을 아주 잘 알아야 한다. 실무 지식이 부족한데 해킹과 통계만 잘하는 사람의 작업 결과는 “공허한 분석, 의미 없는 시스템”라 부를 수 있다.(https://goo.gl/gKkJP)



데이터 과학자가 갖춰야 할 능력 벤다이어그램 2

8. 데이터 과학자가 갖춰야 할 능력

❖ 세 능력을 다 갖춘 사람을 실존하지 않는 존재인 “유니콘”으로 묘사한 것이 재미있다. 이전 다이어그램과 마찬가지로 “머신 러닝”은 옳은 정의가 아니다.(https://goo.gl/YRkr7z)



데이터 과학자가 갖춰야 할 능력 벤다이어그램 3

8. 데이터 과학자가 갖춰야 할 능력

❖ 데이터 과학은 '컴퓨터 도구를 효율적으로 이용하고 적절한 통계학 방법을 사용하여 실제적인 문제에 답을 내리는 활동'으로 데이터 과학자가 갖춰야 할 능력은 ① 실제적인 문제를 통계적으로 표현하고, ② 컴퓨터 도구를 사용하여 시각화와 데이터 가공과 모형화를 한 후에, ③ 그를 이용하여 실제적인 언어로 의미 있는 결과를 만들어내는 능력의 조합이다.

❖ 또한 데이터 과학자에게 필요한 능력은 다른 사람들과 협업할 수 있는 태도이다.

- 데이터 과학은 그 성격상 절대 독불장군이 성공할 수 없는 분야다. 자체로 서기보다는 적용 분야의 전문가, 제품 기획가, 개발자, 데이터 엔지니어 등과 공생하는 도우미이다.

❖ 또 다른 능력은 문서나 말로 협업자들과 대화할 수 있는 소통 능력이다. 듣고, 말하고 읽고, 쓰기를 잘하는 것이 중요하다. 더불어 인문학적 지식, 사회 전반에 관한 관심과 폭넓은 독서가 도움이 된다. 현실의 구체적 문제를 수리적으로 번역하고 통계/수리적 분석 결과를 또한 사람들이 이해할 수 있도록 알기 쉽게 표현하는 것이 데이터 과학의 중요한 기술이다.

참고

- https://ko.wikipedia.org/wiki/빅_데이터
- https://m.blog.naver.com 데이터마이닝 포스트
- 따라하며 배우는 데이터 과학(권재명 지음) - Jpub