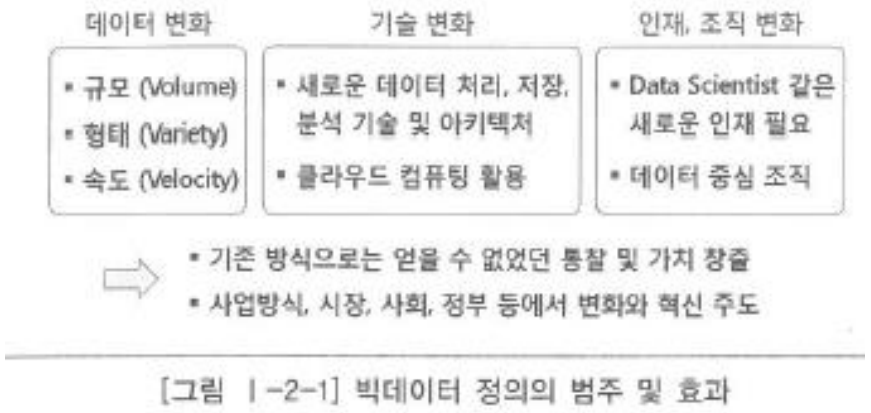
데이터분석 전문가 가이드

과목 1. 데이터 이해
제 2장 데이터의 가치와 미래

출처 : 데이터분석 전문가 가이드, 한국데이터베이스진흥원

제 1절 빅데이터의 이해

1. 정의
   - “ 빅데이터는 일반적인 데이터베이스 소프트웨어로 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터다.” (McKinsey ,2011) => 데이터 규모에 중점
   - “빅데이터는 다양한 종류의 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집 ·발굴 ·분석을 지원하도록 고안된 차세대 기술 및 아키텍처다.” (IDC, 2011) => 분석 비용 및 기술에 중점
   - 데이터와 데이터 처리, 저장 및 분석 기술에 의미 있는 정보 도출에 필요한 인재나 조직까지도 빅데이터라는 개념에 포함 제안 (노무라 연구소) => 포괄적 범위
   - “빅데이터란 대용량 데이터를 활용해 작은 용량에서는 얻을 수 없었던 새로운 통찰이나 가치를 추출해 내는 일이다. 나아가 이를 활용해 시장, 기업 및 시민과 정부의 관계 등 많은 분야에 변화를 가져오는 일이다.” (메이어-쉰베르거와쿠키어, 2013) => 사회 ·정치 ·경제 ·문화적 변화를 포함한 정의
   - 빅데이터의 정의를 종합하면
     - 첫째, 3V로 요약되는 데이터 자체의 특성 변화에 초점을 맞춘 좁은 범위의 정의
     - 둘째, 데이터 자체뿐 아니라 처리, 분석 기술적 변화까지 포함하는 중간 범위의 정의
     - 셋째, 인재, 조직 변화까지 포함해 빅데이터를 넓은 관점으로 정의



[그림 Ⅰ-2-1] 빅데이터 정의의 범주 및 효과

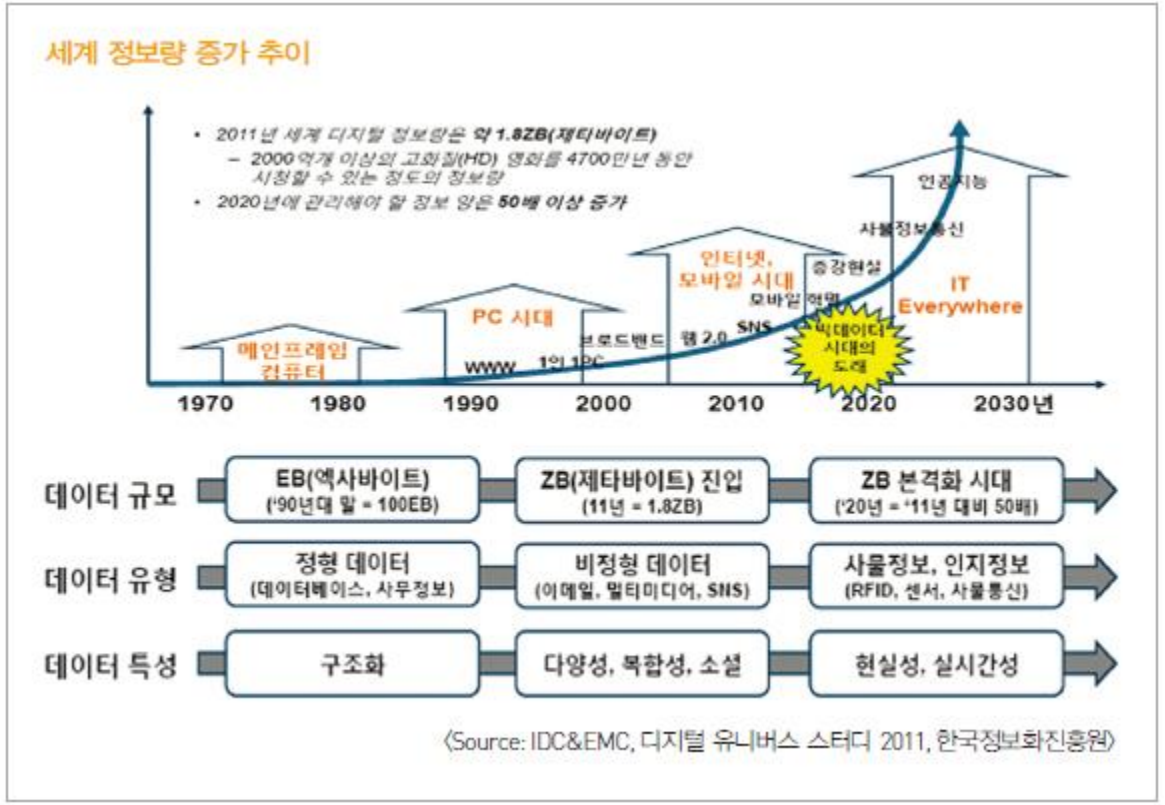
제 1절 빅데이터의 이해

2. 출현 배경

① 산업계 – 고객데이터 축적

② 학계 – 거대 데이터 활용 과학 확산

③ 관련 기술 발전(디지털화, 저장기술, 인터넷보급, 모바일혁명, 클라우드 컴퓨팅)



제 1절 빅데이터의 이해

3. 빅데이터 기능

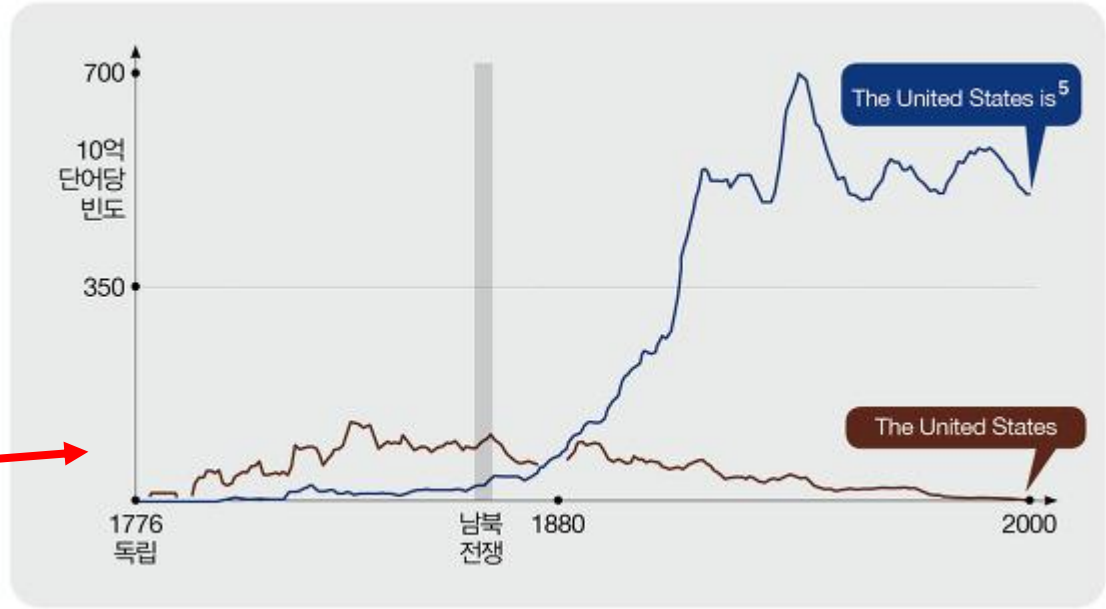
① 산업혁명의 석탄, 철
산업혁명에서 했던 역할을 차세대 산업혁명에서 해낼것으로 기대

② 21세기의 원유
빅데이터도 원유처럼 비즈니스, 공공기관 대국민 서비스, 경제 성장에 필요한 정보를 제공

③ 렌즈
현미경이 생물학 발전에 미쳤던 영향만큼이나 데이터 산업 전반에 영향(예, 구글의 Ngram Viewer)

④ 플랫폼
다양한 사업자들이 공동으로 사용하는 플랫폼을 빅데이터 형태로 제공

=> 차세대 산업혁신에 꼭 필요한 요소



제 1절 빅데이터의 이해

4. 빅데이터가 만들어 내는 본질적인 변화

① 사전처리 -> 사후처리
미리 계획된 필요한 정보만 수집하고 필요하지 않는 정보를 버리는 관료주의 시대의 데이터 사전 처리에서 가능한 한 많은 데이터를 모으고 사후에 숨은 정보를 찾아내는 방식

② 표본조사 -> 전수조사
전수조사에 비용 감소와 데이터 기술과 통계도구의 발전하였고, 전수조사는 샘플링이 주지 못하는 패턴이나 정보를 제공

③ 질 -> 양
구글의 자동번역 시스템 구축은 데이터의 질보다 양이 중요함을 보여줌.

④ 인과관계 -> 상관관계
정교한 이론적 틀에 맞추 인과관계보다는 신속한 의사결정을 원하는 비즈니스에 인사이트를 줄수 있는 상관관계가 중요해짐

제 2절 빅데이터의 가치와 영향

1. 빅데이터의 가치

빅데이터의 가치 산정이 어려운 이유

① 데이터활용 방식 : 재사용, 재조합(mashup), 다목적용 개발

② 새로운 가치 창출

③ 분석 기술 발전

제 2절 빅데이터의 가치와 영향

2. 빅데이터의 영향

① 기업 : 혁신, 경쟁력, 생산성 향상

② 정부 : 환경 탐색, 상황 분석, 미래 대응

③ 개인 : 목적에 따라 활용

=> 생활 전반의 스마트화

제 3절 비즈니스 모델

1. 빅데이터 활용 사례
   - 구글의 페이지랭크(PageRank) 알고리즘
   - IBM의 왓슨
   - 정부에서도 대국민 서비스 개선을 빅데이터 활용
   - 정치인은 선거 승리를 위해서
   - 가수는 팬들의 음악 청취 기록 분석을 통해 실제 공연에서 부를 노래 순서를 짜는데 활용

제 3절 비즈니스 모델

2. 빅데이터 활용 기본 테크닉

1. 연관규칙학습(Association Rule Learning)
2. 유전 알고리즘(Genetic Algorithms)
3. 회귀분석(Regression Analysis)
4. 유형분석(Classification Tree Analysis)
5. 기계학습(Machine Learning)
6. 소셜네트워크 분석(Social Network Analysis)
7. 감정분석(Sentiment Analysis)

커피를 구매하는 사람이 탄산음료를 더 많이 사는가?

이 사용자는 어떤 특성을 가진 집단에 속하는가?

최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?

기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화중에서 어떤 것을 가장 보고 싶어할까?

구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?

새로운 환불 정책에 대한 고객의 평가는 어떤가?

특정인과 다른 사람이 몇 촌 정도의 관계인가?

제 3절 비즈니스 모델

2. 빅데이터 활용 기본 테크닉

1. 연관규칙학습(Association Rule Learning)
2. 유전 알고리즘(Genetic Algorithms)
3. 회귀분석(Regression Analysis)
4. 유형분석(Classification Tree Analysis)
5. 기계학습(Machine Learning)
6. 소셜네트워크 분석(Social Network Analysis)
7. 감정분석(Sentiment Analysis)

커피를 구매하는 사람이 탄산음료를 더 많이 사는가?

이 사용자는 어떤 특성을 가진 집단에 속하는가?

최대의 시청률을 얻으려면 어떤 프로그램을 어떤 시간대에 방송해야 하는가?

기존의 시청 기록을 바탕으로 시청자가 현재 보유한 영화중에서 어떤 것을 가장 보고 싶어할까?

구매자의 나이가 구매 차량의 타입에 어떤 영향을 미치는가?

새로운 환불 정책에 대한 고객의 평가는 어떤가?

특정인과 다른 사람이 몇 촌 정도의 관계인가?

제 4절 위기 요인과 통제 방안

1. 위기 요인

① 사생활침해

② 책임 원칙 훼손

영화‘마이너리 리포트’에 나오는 것처럼 범죄 예측 프로그램에 의해 범행을 저지르전에 체포될 수도 있음.

③ 데이터 오용

포드가 자동차를 만들려고 했을때 사람들의 의견을 물었다면 사람들은 더 빠른 말이 필요하다고 대답을 했을 것이라는 비유 => 빅데이터 기반 미래 예측은 높은 정확도를 갖지만 항상 맞을 수는 없음.

제 4절 위기 요인과 통제 방안

2. 통제 방안

① 동의에서 책임으로

② 결과 기반 책임 원칙 고수

③ 알고리즘 접근 허용

제 5절 미래의 빅데이터

① 데이터 : 모든것이 데이터화

② 기술 : 진화하는 알고리즘, 인공지능(AI)

③ 인력 : 데이터 사이언티스트, 알고리즈미스트(Algorithmist)

알고리즈미스트 : 빅데이터가 발생시키는 문제를 중간자 입장에서 중재하며 해결해 주는 새로운 직업