문제해결과 기업성과 달성을 위한 정보시스템 활용

빅 데이터와 데이터 기반 의사결정 프레임워크

SNU Business School

비즈니스 애널리틱스(Business Analytics)

비즈니스 애널리틱스(Business Analytics)란?

· 비즈니스 애널리틱스: 조직 경영에 있어 인사이트를 도출하고 사업을 계획하기 위해 체계적이고 반복적으로 데이터를 탐색하고 조사하는 과정을 종합적으로 일컫는 용어
[source: https://en.wikipedia.org/wiki/Business_analytics, http://searchbusinessanalytics.techtarget.com/definition/business-analytics-BA]

- 데이터와 통계적 방법론의 효과적인 활용에 기반하여 조직 내에서 데이터 기반 의사결정(Data-driven decision making)을 가능케 한다

- 데이터 기반 의사결정(Data-driven decision making): 조직의 중요한 의사결정을 내릴 때 객관적으로 입증 가능한 데이터에 기반한 의사결정 방식 [source: http://whatis.techtarget.com/definition/data-driven-decision-management-DDDM]

비즈니스 애널리틱스의 범위와 종류

· 기술 분석(Descriptive Analytics)

- 과거와 현재에 일어나는 현상을 이해하기 위해 데이터를 활용

· 예측 분석(Predictive Analytics)

- 통계 및 기계학습(Machine learning) 등을 활용해 미래의 현상을 예측

· 규범 분석(Presecriptive Analytics)

- 최적화와 시뮬레이션 기법 등을 활용하여 의사결정을 지원

비즈니스 애널리틱스의 범위와 종류

· 예시: 재고 가격할인(Retail Markdown) 의사결정[Evans 2013]

- 상황: 대부분의 백화점(Department store)들이 계절 재고를 가격 할인을 통해 최소화하고 싶어함

- Question: 언제(when), 그리고 얼마나(how much) 가격 할인을 해야 하는가?

- 기술 분석: 비슷한 제품의 과거 데이터를 조사해봄(가격, 매출, 광고 효과 등)

- 예측 분석: 가격 데이터를 바탕으로 매출을 예측해봄

- 규범 분석: 매출 이익을 극대화하기 위한 가격과 광고의 적정 수준을 찾아냄

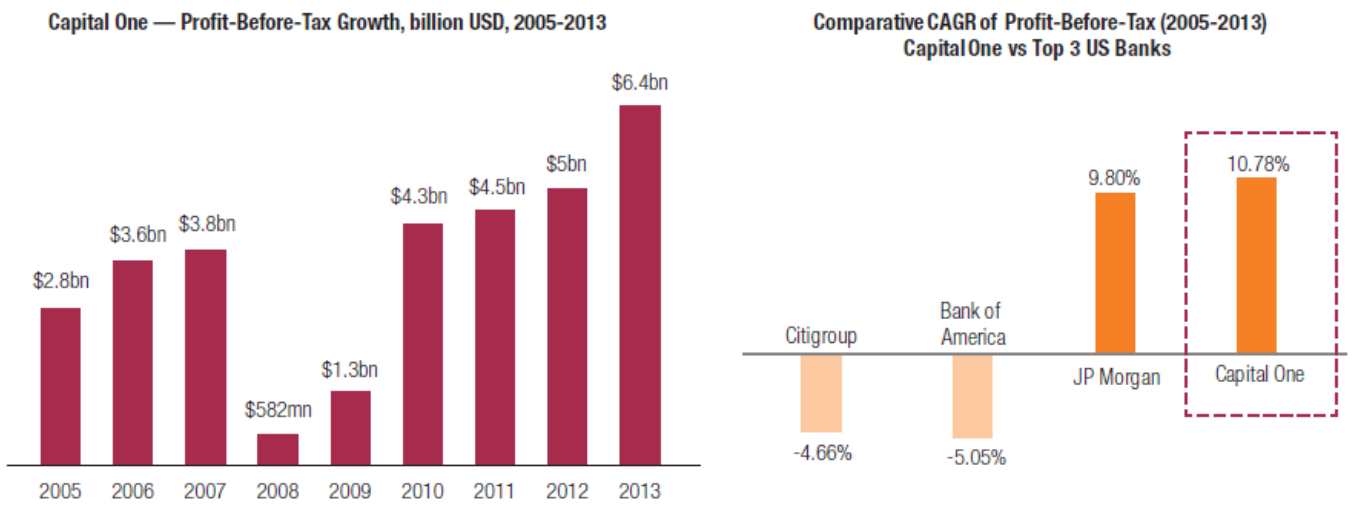
비즈니스 애널리틱스의 적용 사례

· 캐피털 원(Capital One)의 신용 분석 시스템[Buvat and KVJ 2014 / https://www.forbes.com/sites/greatspeculations/2015/09/11/how-does-capital-one-differentiate-itself-in-the-card-industry/#2e0146663cda]

- 캐피털 원의 주력 산업인 카드 대출은 상품에 따라 이자율, 보상 등에 큰 편차가 없는 산업으로 지속적인 경쟁 우위(competitive advantage)를 유지하기가 상당히 어려움

- 그럼에도 불구하고, 캐피털 원은 지속적인 성장세를 유지하며 Bank of America, JP Morgan 등 쟁쟁한 경쟁자들에 뒤처지지 않으며 우수한 실적을 보임

Figure 1: Profit-Before-Tax and Comparative CAGR Analysis of Capital One and Peers



Source: Capital One, Citigroup, JP Morgan Annual Reports

비즈니스 애널리틱스의 적용 사례

· 캐피털 원(Capital One)의 신용 분석 시스템[Buvat and KVJ 2014]

- 캐피털 원은 비즈니스 애널리틱스의 전사적 활용의 가치를 빠르게 이해하고, 창립 초창기인 90년대부터 데이터의 수집, 분석 및 활용을 통해 소비자의 구매 패턴(spending patterns)을 이해하고자 노력함

- 누구에게나 공개된 신용 관련 및 인구통계학적 정보를 바탕으로, 예측 모델을 수립해 고객들에게 맞춤 신용 서비스(personalized offer)를 제공함

  ■ 최신 컴퓨터로 100 milliseconds의 시간 내에 70%가 넘는 정확도로 50개 옵션 중 소비자들이 원하는 제품을 예측하는 데 성공함(“All these steps happened in under 100 milliseconds”)

  ■ 그 결과 2000년부터 2003년까지의 짧은 기간 내에 시장점유율을 3.3%에서 7.2%로 급격히 올릴 수 있었음

- 매출 예측 뿐만 아니라 고객 유지와 비용 절감을 위해서도 비즈니스 애널리틱스를 폭넓게 활용해 금융 산업에 신바람을 불러일으킴

  ■ 그 결과 짧은 기간 내에 고객 유지율을 87%까지 올리고, 비용을 83%나 절감할 수 있었음

비즈니스 애널리틱스의 적용 사례

· 캐피털 원(Capital One)의 신용 분석 시스템[Buvat and KVJ 2014]

- 뿐만 아니라, 캐피털 원은 비즈니스 애널리틱스를 활용해 혁신적인 신제품을 개발해 소비자 가치를 증진시키는 데에도 성공함

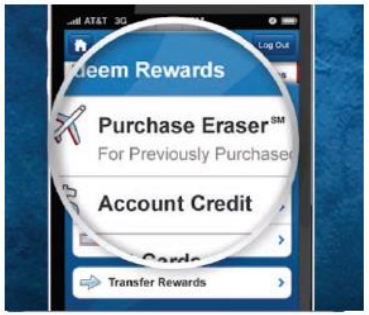
■ "잔액 전송 신용카드(Balance-transfer credit card)"의 개념을 만들어 다른 은행에서 고금리로 대출한 금액을 캐피털 원으로 전송해 더 낮은 금리로 상환할 수 있도록 함

■ "구매 삭제(Purchase eraser)" 서비스를 통해 카드 사용을 통해 축적한 보상 포인트(reward mile)를 활용해 비행, 크루즈, 호텔, 렌터카 등의 여행 비용을 지불할 수 있도록 함

■ 모바일 앱을 통해 타겟 소비자에게 필요할 만한 디지털 쿠폰(Digital coupons)을 제공하기도 함

Figure 3: Capital One's Mobile App Offerings

Capital One's Digital Service Offerings



Purchase Eraser App

Mobile Deals Program

비즈니스 애널리틱스의 적용 사례

· 카이사르 엔터테인먼트(Caesar's Entertainment)의 소비자 세분화 전략[Evans 2013; Welch and Westerman 2013]

- 카이사르 엔터테인먼트는 세계에서 4번재로 큰 도박 기업으로, 미국 전역에 걸쳐 50여개의 카지노 호텔과 7개의 골프장을 운영하고 있음

- 전통적으로 카지노는 "큰 손(whales)"들에게만 역량을 집중해 대부분의 카지노 이용자들은 상대적으로 차별화 되지 않은 경험만 할 수 있었음

■ 큰 손들에게 객실, 전용기 등 다양한 혜택을 제공하면서 카지노에 오래 머물게 해 매출을 늘림

비즈니스 애널리틱스의 적용 사례

· 카이사르 엔터테인먼트(Caesar's Entertainment)의 소비자 세분화 전략[Evans 2013; Welch and Westerman 2013]

- 하버드 경영대학원 교수인 게리 러브만(Garry Loveman)이 CEO로 취임하면서 큰 손들 뿐만 아니라 다른 부류의 고객들에게도 관심을 가지게 됨

  ■ 다양한 고객들에게 개인화된 서비스를 가능케 한 것은 데이터 기반 접근법(data-driven approach)

  ■ "*You can't just fail to service the tens of millions of people that constitute the middle of the market.*" – Garry Loveman

- 단순히 최신 기술을 맹목적으로 도입하는 것이 아니라 고객에게 초점을 맞추고 고객에게 보다 높은 가치를 제공하려는 기업 문화(customer-focused culture) 덕에 차별화된 고객 경험을 제공할 수 있었음

  ■ "The company's "aggressively collegial" culture seeks to complement technical novelty with on-the-ground operational experience"

비즈니스 애널리틱스의 적용 사례

· 카이사르 엔터테인먼트(Caesar's Entertainment)의 소비자 세분화 전략[Evans 2013; Welch and Westerman 2013]

- 고객의 위치 데이터(spatio-temporal data)를 활용해 실시간 맞춤 서비스를 제공하는 모바일 서비스 'Texpress'

■ Texpress를 통해 원격 체크인이 가능해 고객들의 대기시간을 최소화함

■ 고객의 현재 위치에 최적화된 특별 서비스를 SMS로 전송

• 예를 들어, 현재 프랑스 파리(Paris)에 있으면 에펠탑 무료 이용권을 전송

Figure 1: Caesars' guests can use its Texpress service to check in with a mobile phone and avoid lines when they arrive

비즈니스 애널리틱스의 적용 사례

· 카이사르 엔터테인먼트(Caesar's Entertainment)의 소비자 세분화 전략[Evans 2013; Welch and Westerman 2013]

- 웹사이트를 모바일 기기에 맞게 최적화해 언제 어디서든지 쉽게 카이사르가 소비자들에게 제공하는 서비스를 편하게 이용할 수 있도록 해 소비자들의 데이터를 보다 폭넓게 축적하고 이를 바탕으로 다시 개인화된 서비스를 제공

  ■ 보상 포인트(reward point) 관리

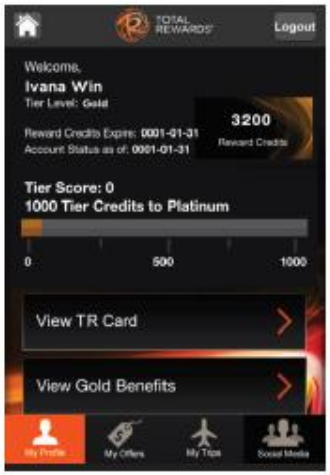
  ■ 카지노 위치 검색(casino locator)

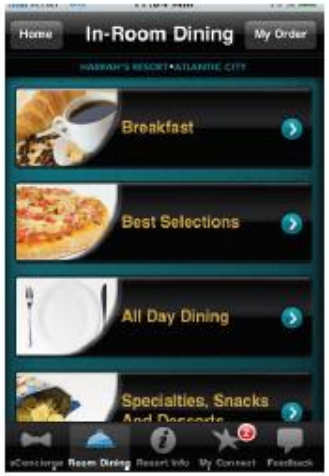
  ■ 객실 예약 및 룸 서비스 신청

- "Social, local, and mobile are huge areas of innovation and places to win."

Mobile Internet

Text Programs

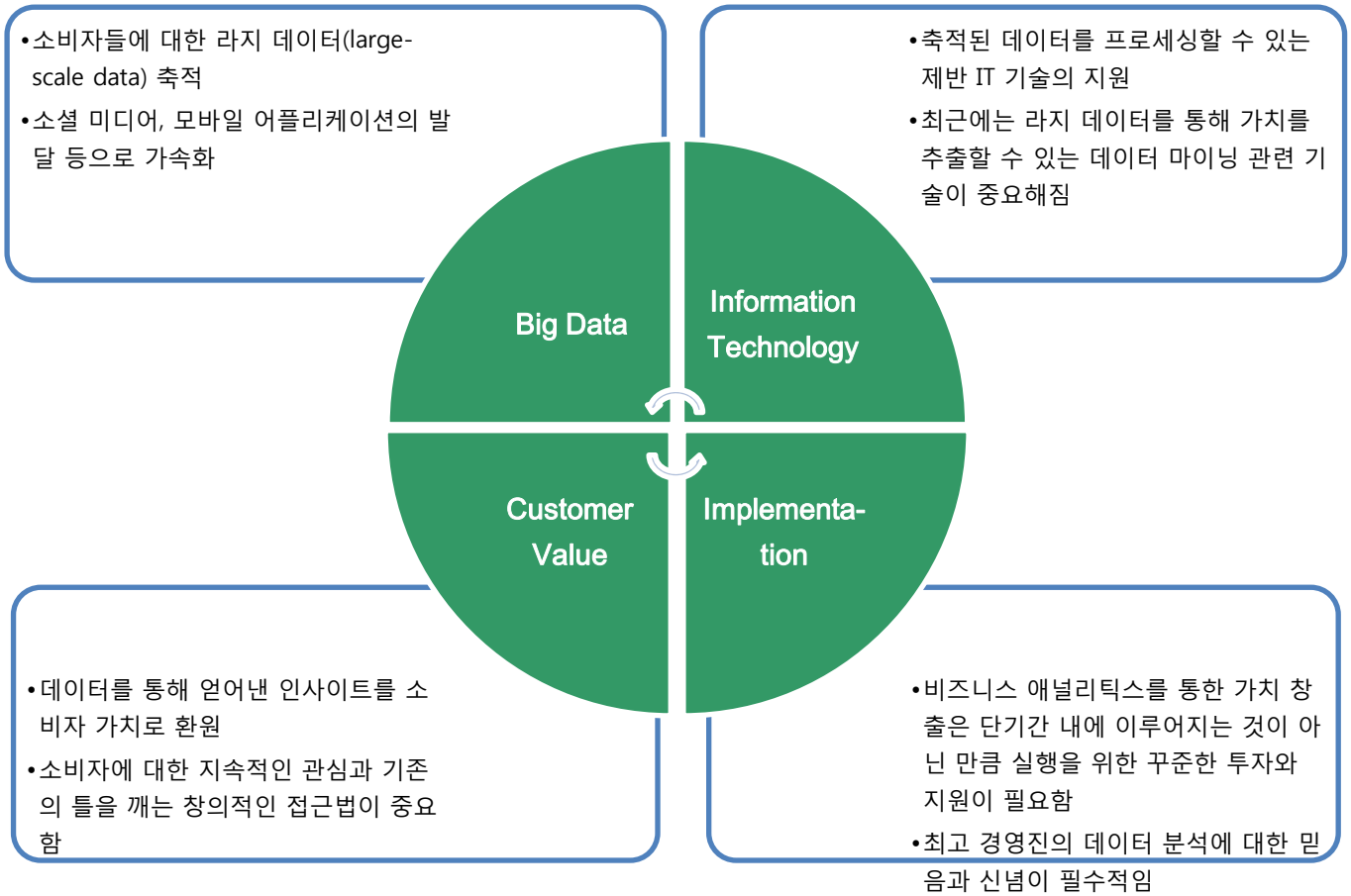




비즈니스 애널리틱스의 적용 사례

· 비즈니스 애널리틱스의 성공 요건

- 다양한 사례들을 통해 추출해낸 비즈니스 애널리틱스의 성공 요건은 다음과 같음



데이터 기반 의사결정(Data-driven Decision Making)

데이터 기반 의사결정이란?

· 데이터 기반 의사결정(Data-driven decision making): 조직의 중요한 의사결정을 내릴 때 객관적으로 입증 가능한 데이터에 기반한 의사결정 방식
[source: http://whatis.techtarget.com/definition/data-driven-decision-management-DDDM]

- MIT Center for Digital Business에 따르면, 데이터에 기반한 의사결정을 내리는 기업들이 그렇지 않은 기업들에 비해 4% 높은 생산 효율과 6% 높은 수익을 자랑한다고 함

- 미국 제조업 내에서 데이터 기반 의사결정은 2010년과 2015년 사이 5년 동안 3배 가량 증가한 것으로 밝혀짐
[Brynolffson and McElheran 2016]

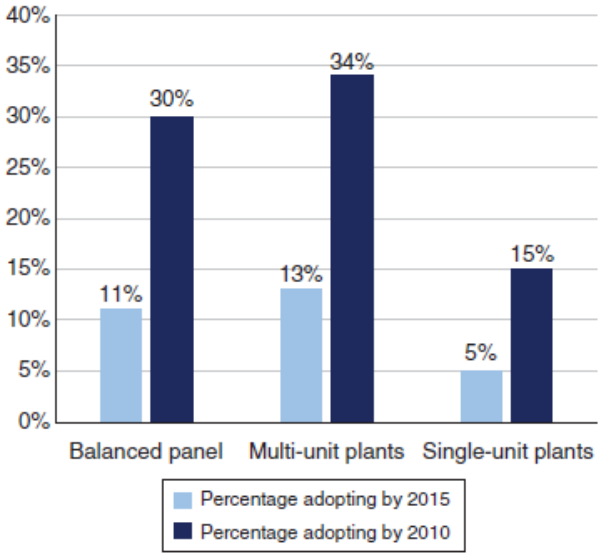


FIGURE 1. ADOPTION OF DATA-DRIVEN DECISION-MAKING IN US MANUFACTURING

데이터, 데이터베이스, 그리고 정보

· 데이터(Data)

- 경영 환경 속에서 수집된 사실(fact)과 수치(figure)
- 텍스트나 음성과 같은 질적 데이터(qualitative data)와 숫자로 표현되는 양적 데이터(quantitative data)를 모두 포함
- 비즈니스에서 데이터의 활용 사례

● 회계감사(Accounting audits)

● 수익성 분석(Financial profitability analysis)

● 경제 동향 분석(Economic trends)

● 마케팅 리서치(Marketing research)

● 생산 효율성(Operations management performance)

● 인적자원관리 측정(Human resource measurements)

데이터, 데이터베이스, 그리고 정보

· 데이터의 분류

- 범주형(Categorical or nominal) 데이터: 몇 가지 범주(category)중 하나로 분류되는 데이터

● 특정인의 혈액형: A, B, AB 혹은 O
● 전자메일 분류: 스팸 혹은 일반

- 순서형(Ordinal) 데이터: 범주형과 비슷하나 순서가 있는 데이터

● 신용등급

- 구간형(Interval) 데이터: 순서형과 비슷하나 구간 사이의 간격이 일정한 데이터

● 온도(섭씨 혹은 화씨)

- 비율형(Ratio) 데이터: 구간형과 비슷하나 절대적인 영점(zero point)가 존재하는 데이터

● 시간
● 절대온도(Degrees Kelvin)

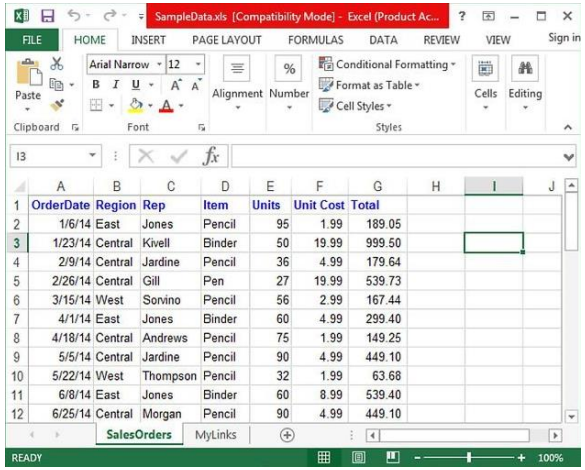
데이터, 데이터베이스, 그리고 정보

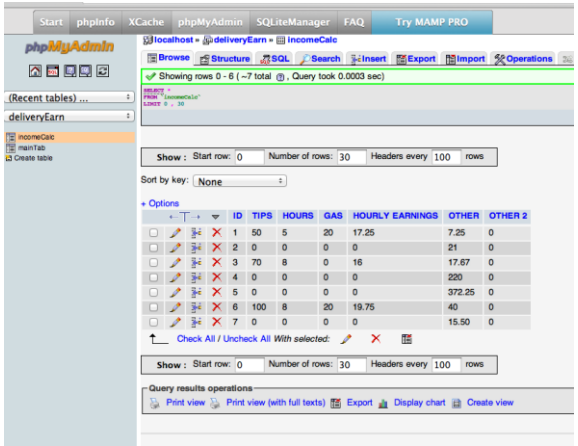
· 데이터베이스

- 데이터를 담고 있는 체계화된 컴퓨터 파일들의 총체

- 일반적으로는 행(row)과 열(column)로 이루어져 있는 관계형 데이터베이스를 의미함

- 하나의 데이터베이스 내에 실제 데이터를 담고 있는 여러 개의 표(table)들이 존재함





데이터, 데이터베이스, 그리고 정보

· 정보[source: http://www.diffen.com/difference/Data_vs_Information]

- 데이터를 가공하고, 규칙화하고, 분석해 조직의 의사결정에 도움을 줄 수 있도록 변환된 형태

- 데이터와 정보

● 한 반에서 각 학생의 시험 점수는 데이터이다

● 모든 학생의 시험 점수를 통계를 내 평균을 계산해 학생들의 수준에 대한 판단(의사결정)을 내리기 위해 사용한다면 이는 정보이다

- 데이터 자체만으로는 의미가 없고 데이터를 통해 정보를 얻어내고, 이를 통해 인사이트를 추출하는 것이 비즈니스 애널리틱스의 핵심이다!

데이터 기반 의사결정의 사례

· 구글의 데이터기반 인사평가(HR decisions using data)
[source: http://programs.online.utica.edu/articles/data-driven-decisions]

- 구글에서는 매니저(manager)의 역량이 팀의 성과에 얼마나 영향을 미치는지 알기 위해 인적자원 분석 부서(People Analytics Department)를 설립함

- 성과 리뷰와 종업원 설문을 통해 조사한 결과, 더 좋은 매니저를 가진 팀원들이 그렇지 않은 팀원들에 비해 더 좋은 성과를 낼 뿐 아니라 만족도와 근속년수도 높은 수치를 보였다고 함

- 나아가, 좋은 매니저의 질적 역량을 찾아내기 위해 “위대한 매니저 상(the Great Managers Award)”을 수여해 종업원들에게 우수한 매니저의 특질을 물어봄

- 그 결과, 우수한 매니저의 여덟 가지 특질을 찾아낼 수 있었으며, 이를 바탕으로 매니저들을 더 효과적으로 평가하고 동기부여할 수 있었다고 함

빅 데이터(Big Data)

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Volume (양): IT가 발달함에 따라 만들어지고 활용될 수 있는 데이터의 양이 기하급수적으로 늘어남
- Velocity (속도): 데이터의 입출력 및 분석 속도가 과거와는 다르게 굉장히 빨라짐
- Variety (다양성): 정형화된 데이터(수치형 데이터)뿐만 아니라 다양한 종류의 비정형 데이터가 분석의 대상으로 등장함
- Veracity (진실성): 데이터의 양(volume)이 증가하지만 데이터의 편향(bias), 노이즈(noise) 등이 증가하면서 진실된 데이터와 거짓된 데이터를 가려낼 수 있는 과정이 중요해짐

빅 데이터란?

· 빅 데이터 4V: VIDEO

- 수업 때 학생들에게 아래 영상을 보여주시면 이해가 빠를 것 같습니다

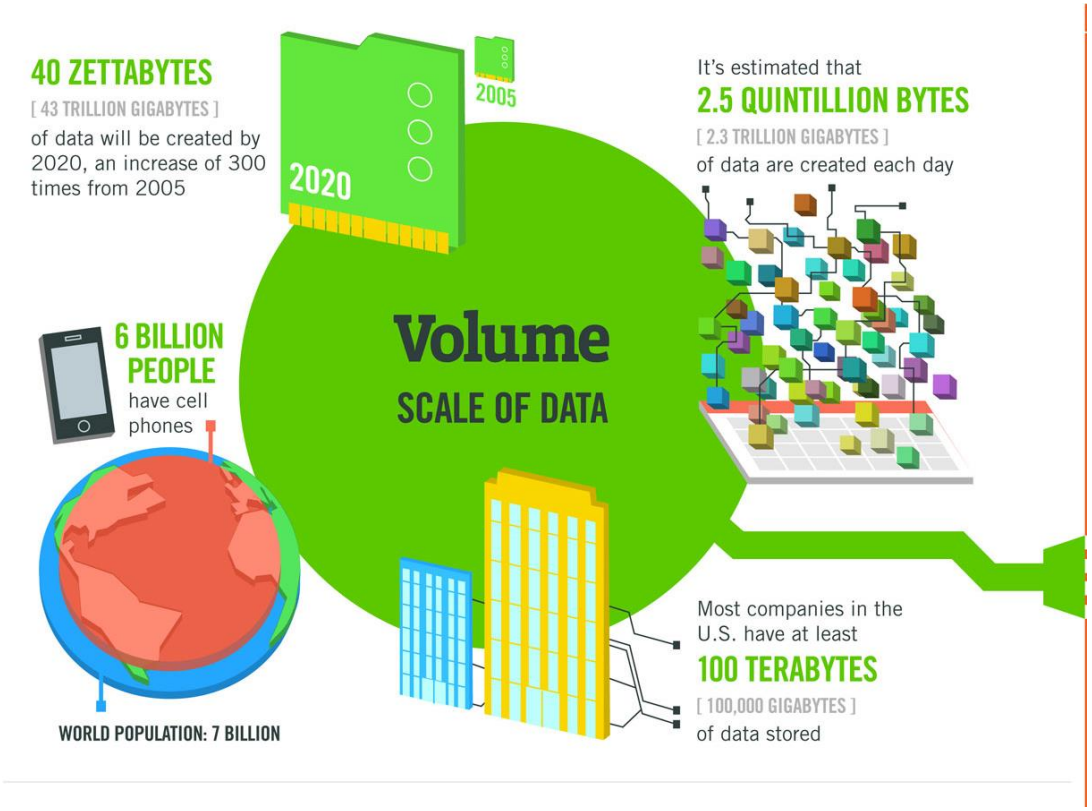
- VIDEO: The 4Vs of BIG DATA from IBM - Video Infographics

- link: https://www.youtube.com/watch?v=Hv397JnNWYc

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Volume (양)



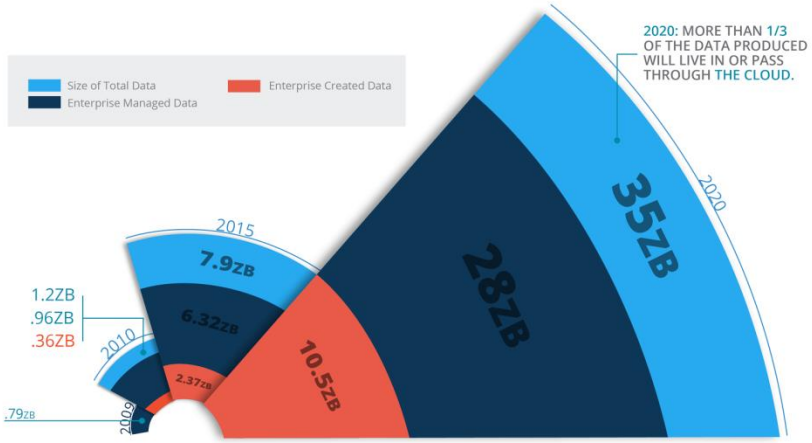
빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Volume (양): 웹(the Web), 소셜 미디어(social media), 사물인터넷(Internet of Things) 등이 발달하면서 발생하고 활용되는 데이터의 양이 기하급수적으로 늘어나게 됨

● 일례로, 페이스북의 데이터 웨어하우스는 300페타바이트(PB)가 넘는 데이터를 저장이 가능하며 하루에 들어오는 데이터의 양은 약 600테라바이트(TB)에 이른다[source: https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/]

● 2020년에는 전 세계적으로 35제타바이트(ZB)에 이르는 데이터가 생산되고 활용될 것으로 예상된다[source: https://www.impactradius.com/blog/7-vs-big-data/]



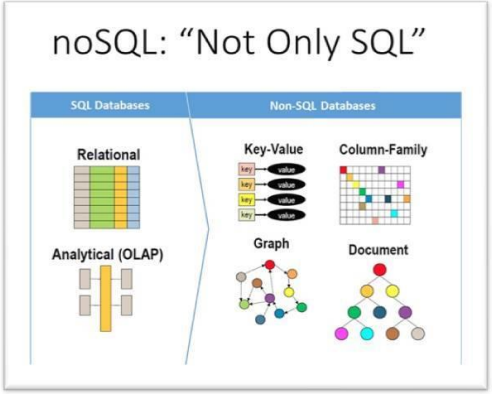
빅 데이터란?

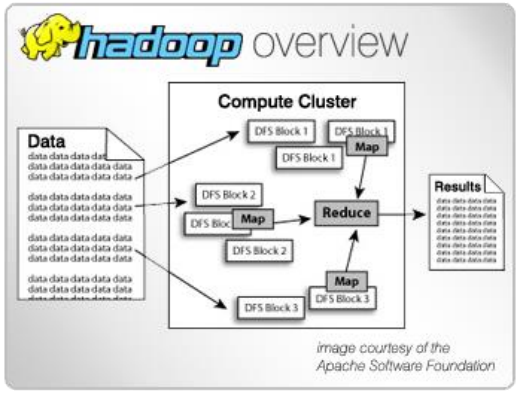
· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Volume (양): 기하급수적으로 늘어나는 데이터를 저장/처리하기 위하여 다양한 빅 데이터 관련 도구들이 개발됨

● NoSQL (Not only SQL): 전통적으로 널리 사용되었던 관계형 데이터베이스 모델(Relational Database Model)에 기반한 관계형 데이터베이스 관리 시스템(Relational Database Management Systems)의 단점인 분산 처리가 불가능하다는 것을 극복하기 위해 제안된 데이터베이스 시스템들을 총칭함. 데이터의 일관성을 포기하는 대신 여러 대의 컴퓨터로 다량의 데이터를 분산하여 저장하는 것이 가능해짐

● 아파치 하둡(Apache Hadoop): 대량의 자료를 처리할 수 있는 분산 처리 시스템을 지원하는 프리웨어 자바 소프트웨어 프레임워크. 야후, 페이스북 등이 하둡을 자사의 데이터와 관련된 여러 분야에 폭넓게 활용하고 있는 것으로 알려져 있음

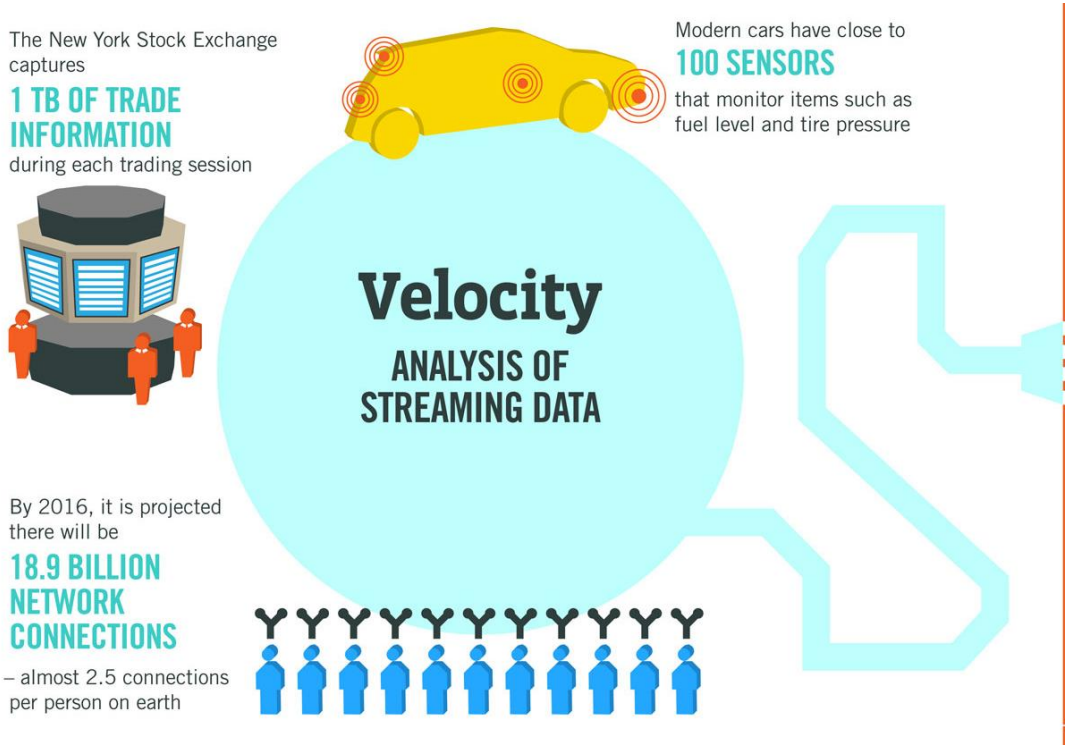




빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Velocity (속도)



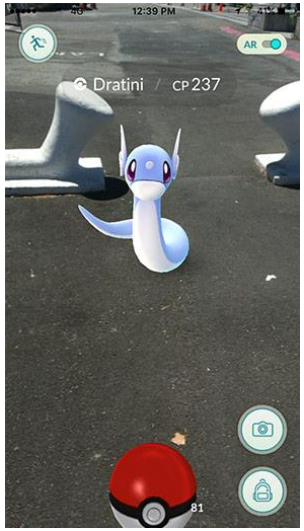
빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Velocity (속도): 대량의 데이터가 생성될 뿐만 아니라, 데이터의 생성 주기도 짧아지며 아주 짧은 시간 동안만 활용 가치가 있음

● 센서, 사물 인터넷 등이 널리 활용되면서 데이터를 실시간으로(real-time) 주고 받으며 처리하는 것이 중요해짐

● 자율주행자동차, 우버(Uber), 포켓몬 고(Pokemon Go) 등 시시각각으로 변하는 시공간 데이터(spatio temporal)를 활용하는 어플리케이션이 많아지면서 매 시간 엄청난 양의 데이터가 만들어지는 동시에 사라지고 있음



빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

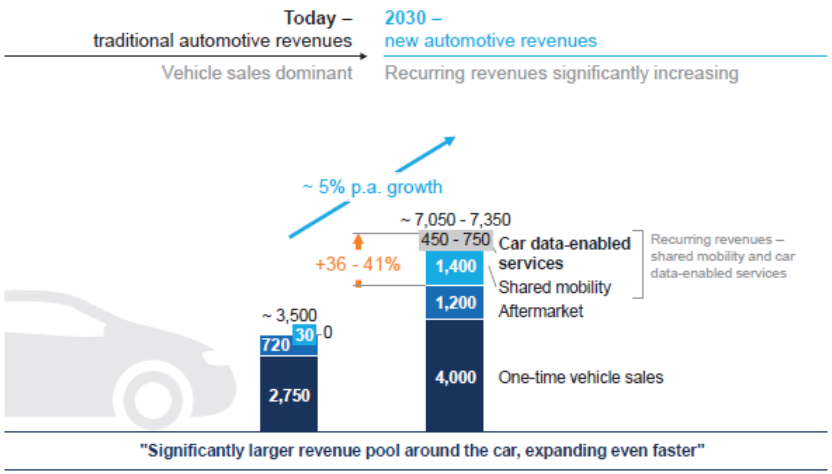
- Velocity (속도): 고속의 데이터의 가치를 붙잡기 위해 '연결(connection)'의 가치가 갈수록 높아지고 있음

● 센서와 사물인터넷 등으로 "연결된 자동차(connected car)'의 경우 클라우드로 매 시간 25기가바이트(GB)의 데이터를 전송할 것으로 예상됨[source: https://qz.com/344466/connected-cars-will-send-25-gigabytes-of-data-to-the-cloud-every-hour/]

● 이렇게 생성된 데이터는 주차(parking), 내비게이션(navigation), 타겟 광고(targeted advertisements), 보안, 카풀(carpool) 등에 활용될 수 있으며 2030년에는 4천5백억에서 7천5백억 달러에 이르는 시장을 형성할 것으로 예상됨 [Bertoncello et al 2016]

Exhibit 2 Car-generated data may become a USD 450 - 750 billion market by 2030

USD billions



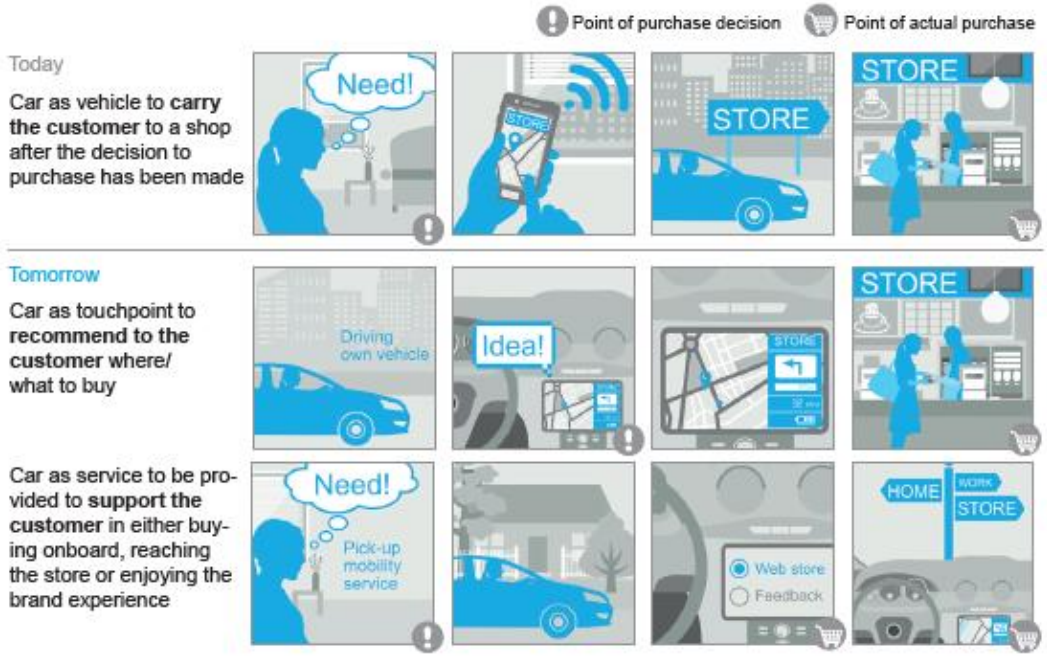
빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Velocity (속도): 고속의 데이터의 가치를 붙잡기 위해 '연결(connection)'의 가치가 갈수록 높아지고 있음

● 차량 데이터의 활용 예상 – 차량 내부 광고(in-car advertising)

Exhibit 24 **Car data could enable in-car advertising, allowing for decision making or purchasing online while driving** ILLUSTRATIVE

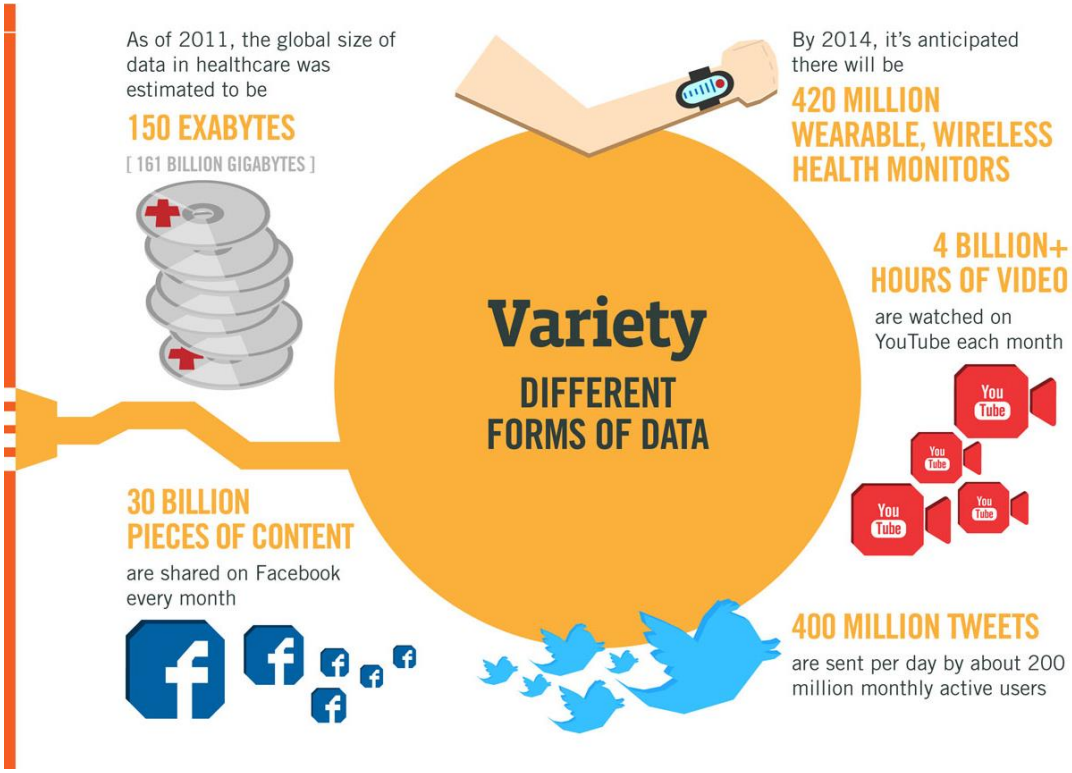


SOURCE: McKinsey

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Variety (다양성)



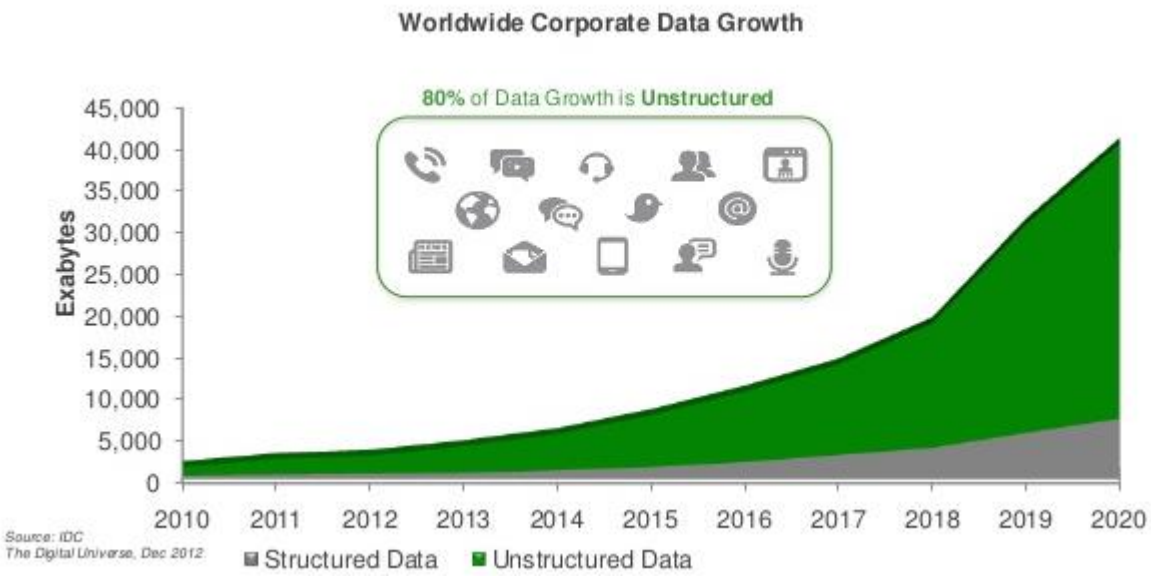
빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Variety (다양성): 과거와 달리 테이블에 저장할 수 있는 정형 데이터(structured data) 뿐만 아니라 텍스트, 이미지, 영상 등 비정형 데이터(unstructured data)의 가치가 증대됨. 소셜 미디어의 발달과 함께 비정형 데이터의 양이 폭발적으로 증가하고 있음

● 유투브(Youtube)에는 매 분 300시간 분량의 영상이 업로드되고 있음[source: https://fortunelords.com/youtube-statistics/]

● 인스타그램(Instagram)에는 매일 9,500만 개의 사진과 비디오가 업로드되고 있음[source: https://blog.hootsuite.com/instagram-statistics/]

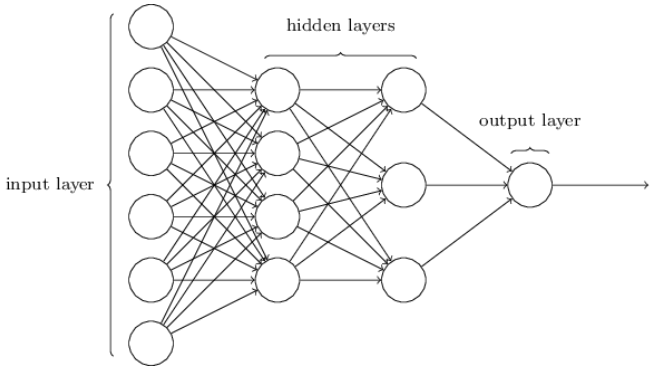


빅 데이터란?

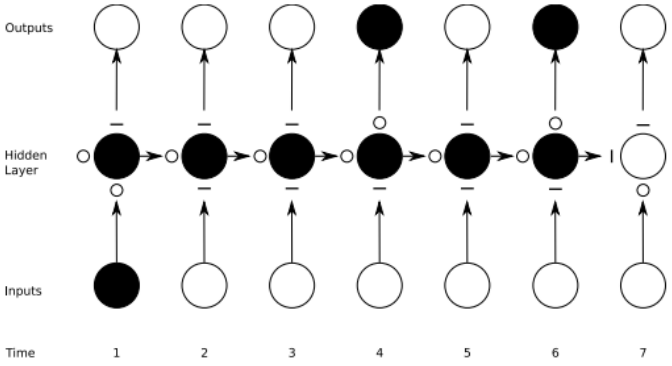
· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Variety (다양성): 딥 러닝(deep learning) 등 분석 기술이 급진적으로 발달하면서 비정형 데이터 분석의 가치도 동반 상승하고 있음[source: http://machinelearningmastery.com/inspirational-applications-deep-learning/]

● 순환형 신경망(Recurrent Neural Network): 일부 학자들이 "궁극의 인공 신경망 구조"라고 주장하듯 굉장히 강력한 모델로 글, 유전자, 손글씨, 주가 등 배열(sequence)의 형태를 갖는 데이터에서 패턴을 인식함[source: https://deeplearning4j.org/kr/lstm]



일반적인 신경망 구조



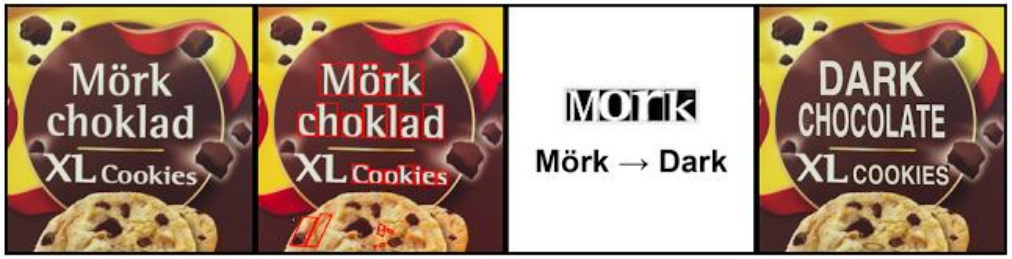
순환형 신경망 구조

빅 데이터란?

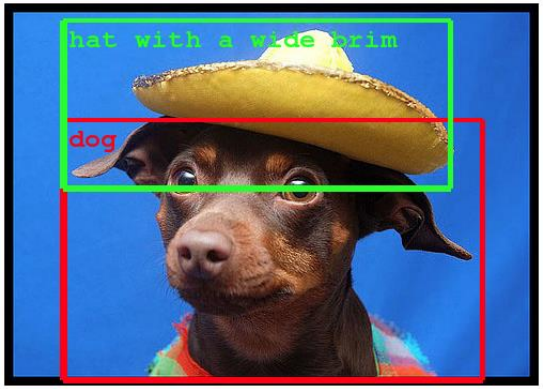
· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Variety (다양성): 딥 러닝(deep learning) 등 분석 기술이 급진적으로 발달하면서 비정형 데이터 분석의 가치도 동반 상승하고 있음[source: http://machinelearningmastery.com/inspirational-applications-deep-learning/]

● 자동 변역(Automatic Machine Translation): 딥 러닝을 활용해 텍스트와 이미지를 자동으로 번역함



● 사물 탐지(Object detection): 딥 러닝을 활용해 이미지 내에 나타나는 사물을 탐지함



빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Variety (다양성): 딥 러닝(deep learning) 등 분석 기술이 급진적으로 발달하면서 비정형 데이터 분석의 가치도 동반 상승하고 있음[source: http://machinelearningmastery.com/inspirational-applications-deep-learning/]

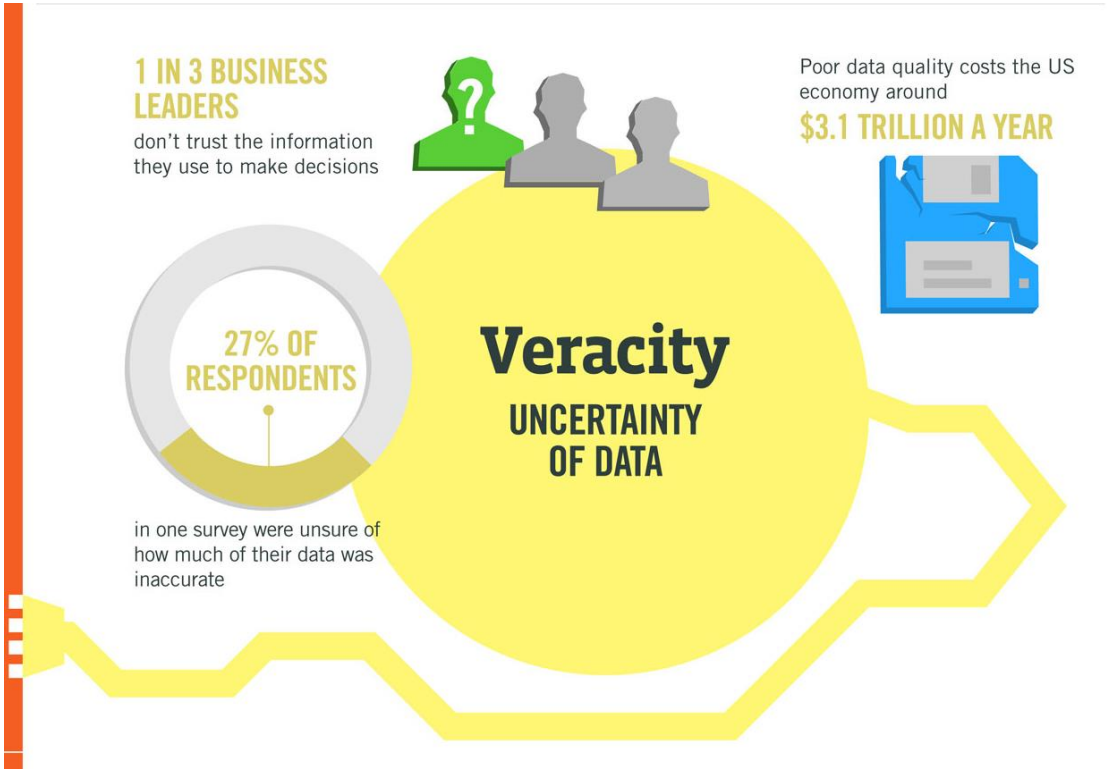
● 자동 텍스트 생성(Automatic text generation): 딥 러닝을 활용해 자동으로 텍스트를 생성함. 아래는 셰익스피어의 작품을 학습해 기계가 만들어낸 희극의 일부임

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.

Clown:
Come, sir, I will make did behold your worship.

VIOLA:
I'll drink it.
```

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Veracity (진실성)



빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Veracity (진실성): 데이터의 양이 늘어나고 형식이 다양해지면서 데이터의 불확실성(uncertainty)에 대한 우려의 목소리도 높아지고 있음

● 비정형 데이터의 경우 노이즈(noise)가 많이 포함되어 있을 수 있음 – 예를 들어, 텍스트 메시지나 온라인 채팅에서 쓰이는 언어들은 문법이나 규칙에 맞지 않는 은어나 약어 등이 많이 포함되어 있을 수 있다

● 다양한 출처의 데이터를 통합하여 활용하다 보니 데이터의 형식이나 단위가 맞지 않을 수 있고 중복된 데이터가 존재할 가능성이 있음

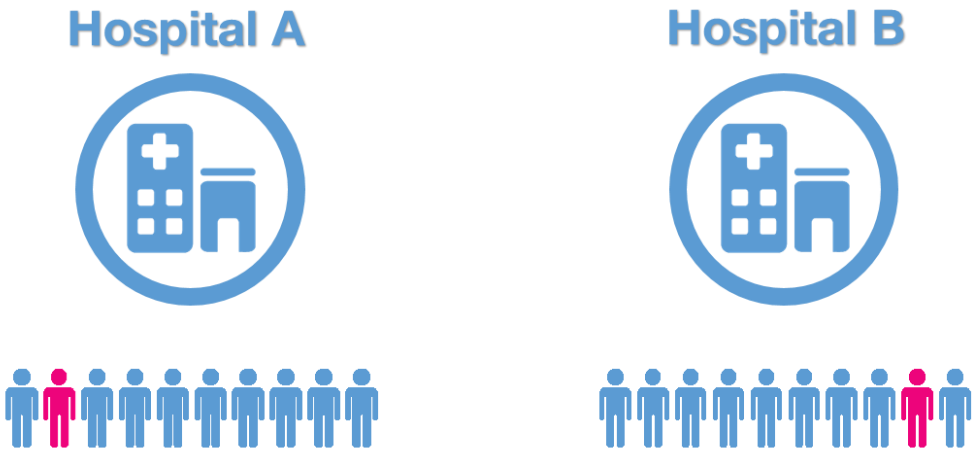
● 잘못 입력된 값, 결측치(missing value) 등

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Veracity (진실성): 데이터의 질(quality)를 높이기 위한 다양한 시도가 행해지고 있음

● 엔터티 레졸루션(Entity resolution): 서로 다른 데이터베이스에서 같은 레코드(record)를 찾아 내 데이터 내 중복(redundancy)를 없애는 작업. 데이터 전처리, 테러 방지(counter-terrorism), 소셜 네트워크 등의 분야에서 폭넓게 활용됨



Patient Name: Vivian Christensen
Visit ID: 837720
Date of Birth: 3/20/1953
SSN: 000-86-6628
MRN: 9968427
Dx: 414.00

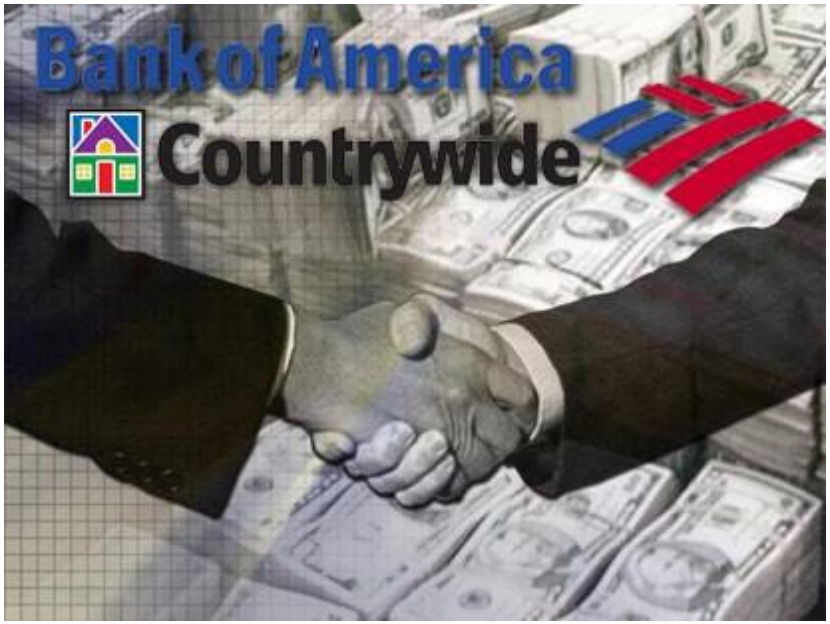
Patient Name: Viv Christensen
Visit ID: 483005
Date of Birth: 3/20/1953
SSN: 000-68-6628
MRN: 0099523461
Dx: 493.01

빅 데이터란?

· 빅 데이터 4V[source: Laney 2001, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, https://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/]

- Veracity (진실성): 데이터의 질(quality)를 높이기 위한 다양한 시도가 행해지고 있음

● 엔터티 레졸루션의 활용 사례: Bank of America와 Countrywide Mortgage가 M&A를 시도할 때 핵심적인 과제가 Countrywide Mortage의 고객 데이터베이스에서 전자 메일 주소를 가져오는 것으로 추측됨. 이 때 두 기업의 대용량 데이터베이스에서 중복되는 엔트리를 찾아내기 위해 엔티티 레졸루션 기법이 활용될 수 있었음[Herzrog et al 2010]

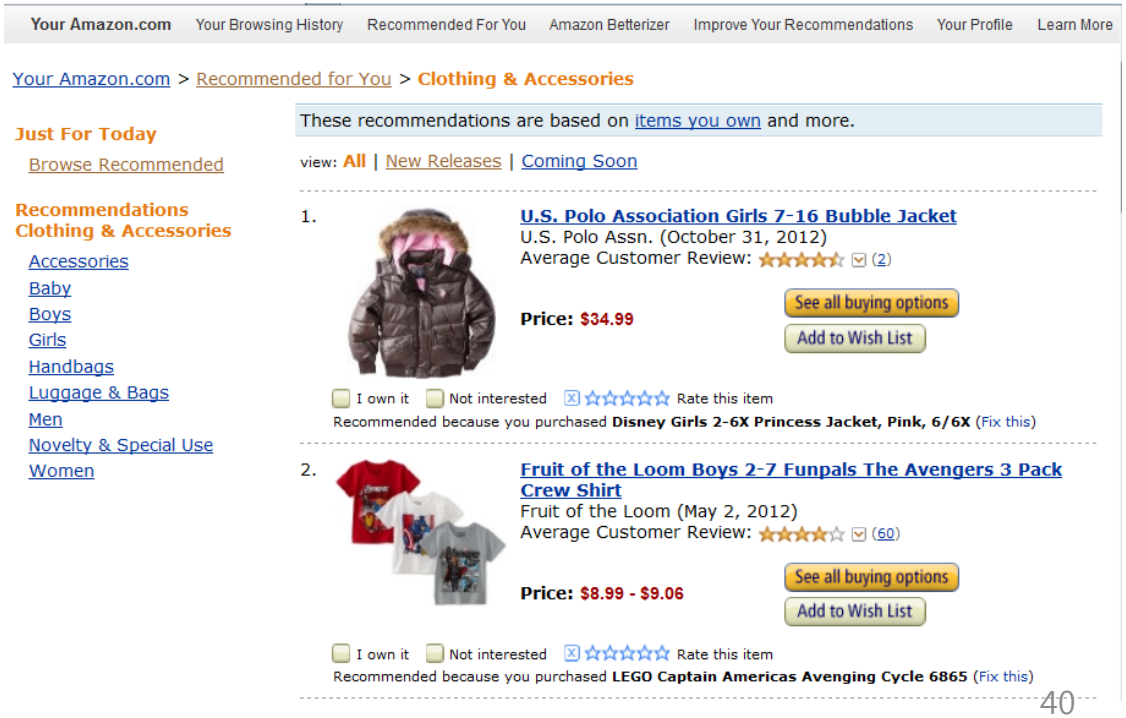


빅 데이터와 우리의 미래

· 아마존(Amazon)의 추천 시스템 – 딥러닝 기반 추천시스템

"아마존은 고객에게 상품을 팔아서가 아니라 고객의 구매의사결정을 도와줘 돈을 번다:"

- 제프 베조스(Jeff Bezos), CEO of Amazon



빅 데이터와 우리의 미래

· 아마존의 추천 시스템: 영상

- **수업 때 학생들에게 아래 영상을 보여주시면 이해가 빠를 것 같습니다**

- VIDEO: How Big Data Is Used In Amazon Recommendation Systems To Change Our Lives

- link: https://www.youtube.com/watch?v=S4RL6prqtGQ

빅 데이터와 우리의 미래

· 아마존은 추천시스템의 기반이 되는 딥러닝 프레임워크 'DSSTNE'를 대중에게 공개
[source: https://www.wired.com/2016/05/amazons-giving-away-ai-behind-product-recommendations/]

- 아마존의 추천시스템은 단순히 "딥러닝 기술"을 활용하는 것이 능사가 아니라 그에 걸맞는 데이터와 플랫폼이 기반이 되어야 한다는 것을 시사함

CADE METZ BUSINESS 04.09.15 08:09 PM

NOW ANYONE CAN TAP THE AI BEHIND AMAZON'S RECOMMENDATIONS

GETTY IMAGES

Amazon helped show the world how machines can learn. As far back as the late '90s, the company's online retail site would track every book, CD, and movie you purchased. As time went on, it would develop a pretty good sense of what you liked, serving up product recommendations its code predicted would catch your eye.

It wasn't rocket science. It was an algorithm. But it worked. And in the years since, the field of so-called

KLINT FINLEY BUSINESS 05.16.16 6:03 PM

AMAZON'S GIVING AWAY THE AI BEHIND ITS PRODUCT RECOMMENDATIONS

Amazon has become the latest tech giant that's giving away some of its most sophisticated technology. Today the company unveiled DSSTNE (pronounced "destiny"), an open source artificial intelligence framework that the company developed to power its product recommendation system. Now any company, researcher, or curious tinkerer can use it for their own AI applications.

It's the latest in series of projects recently open sourced by large tech companies all focused on a branch of AI called deep learning. Google, Facebook, and Microsoft have mainly used these systems for tasks like image and speech recognition. But given Amazon's core business, it's not surprising that the online retailer's version is devoted to selling merchandise.

"We are releasing DSSTNE as open source software so that the promise of deep learning can extend beyond speech and language understanding and object recognition to other areas such as search and recommendations," the Q&A section of Amazon's DSSTNE GitHub page reads. "We hope that researchers around the

빅 데이터와 우리의 미래

· 이제는 '추천'을 넘어서 '예측' 배송까지[source: https://techcrunch.com/2014/01/18/amazon-pre-ships/]

- 데이터를 기반으로 한 예측배송 시스템을 특허 등록함

Amazon Patents "Anticipatory" Shipping — To Start Sending Stuff Before You've Bought It

Posted Jan 18, 2014 by Natasha Lomas (@riptari)

Next Story

Amazon's plans for autonomous flying delivery drones are so last year. The ecommerce juggernaut is purportedly working on something far more dystopian: pre-shipment.

Amazon has filed a patent for a shipping system designed to cut delivery times by predicting what buyers are going to buy before they buy it — and shipping products in their general direction, or even right to their door, before the sales click even (or ever) falls.

Which really is one more step towards cutting out human agency entirely from the ecommerce roundabout. Why not have machines autonomously buy stuff from other

Crunchbase

Amazon

FOUNDED
1994

OVERVIEW
Amazon is an e-commerce retailer formed originally to provide consumers with products in two segments. It offers users with merchandise and content purchased for resale from vendors and those offered by third-party sellers. Operating in North American and international markets, Amazon provides its services through websites such as amazon.com and amazon.ca. It also enables authors, musicians, filmmakers, ...

LOCATION
Seattle, WA

CATEGORIES
Crowdsourcing, E-Commerce, Internet, Delivery, Retail, Software

FOUNDERS
Jeff Bezos

WEBSITE
http://amazon.com

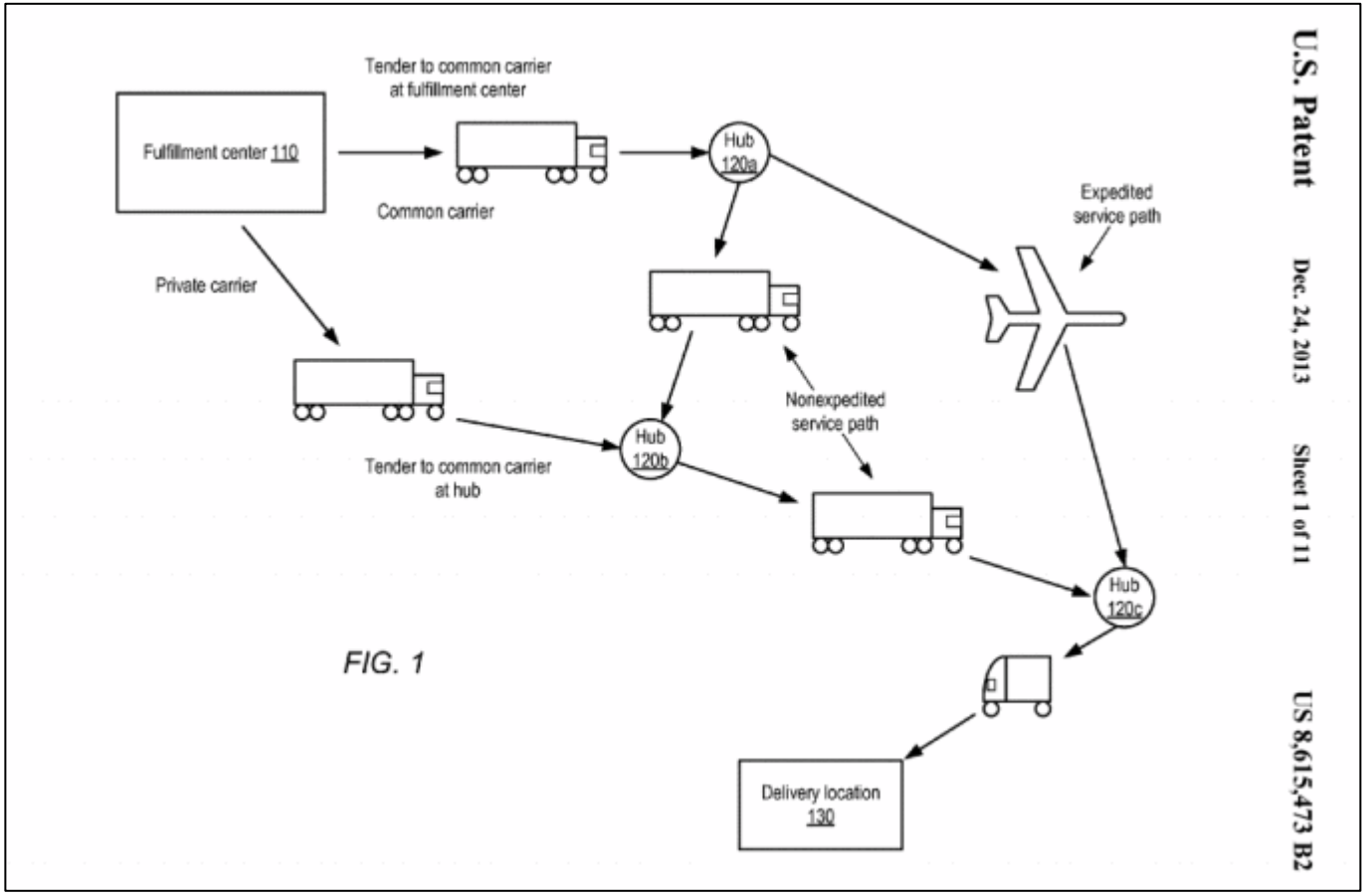
Full profile for Amazon

NEWSLETTER SUBSCRIPTIONS

The Daily Crunch
Get the top tech stories of the day delivered to your inbox

빅 데이터와 우리의 미래

· 이제는 '추천'을 넘어서 '예측' 배송까지[source: https://techcrunch.com/2014/01/18/amazon-pre-ships/]



빅 데이터와 우리의 미래

· 유투브의 자막 제공 서비스

- 자연어 학습 데이터를 바탕으로 자막을 일일이 사람을 만드는 것이 아니라 기계가 자동으로 생성함

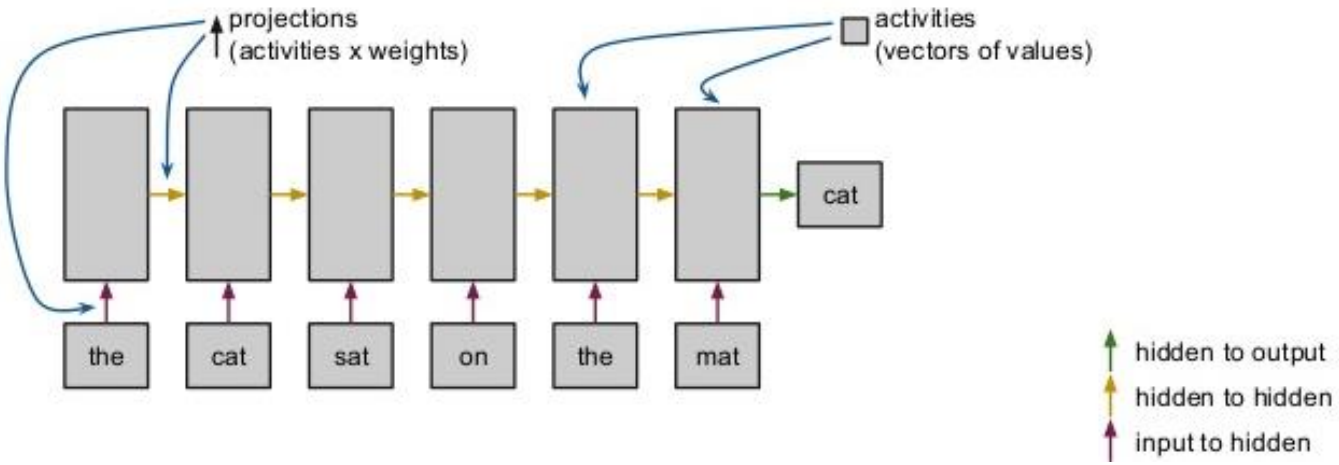
- 여기서도 텍스트 분석을 위해 순환신경망(RNN)이 활용됨

빅 데이터와 우리의 미래

· 유투브의 자막 제공 서비스

- 처음에는 오류가 많았으나 데이터가 축적되고 딥러닝 기술이 정교해지면서 정확도가 거의 완벽한 수준에 가까워 졌으며, 이제는 영어 외에도 다른 언어로도 지원됨

- 과거에 전통적인 선형 기계학습 모델(SVM, softmax 등) 만으로 해결할 수 없는 문제를 새로운 기술(딥러닝)이 발달하면서 복합적인 시스템을 만들어 해결한 대표적인 사례라고 할 수 있음



빅 데이터와 우리의 미래

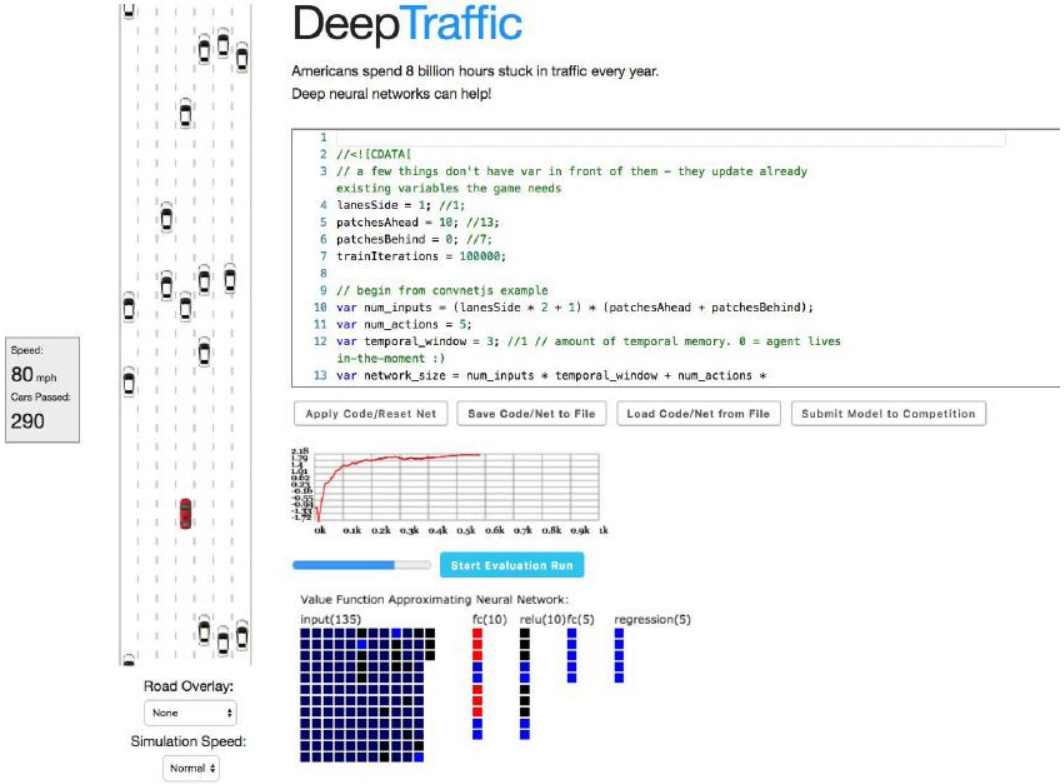
· 자율주행자동차

- 자동차의 자율 주행 관련 기술이 발달되어 상용화를 눈앞에 두고 있음

빅 데이터와 우리의 미래

· 자율주행자동차

- 자율주행자동차 구동의 핵심도 역시 다량의 데이터를 바탕으로 한 ‘딥러닝 기술’

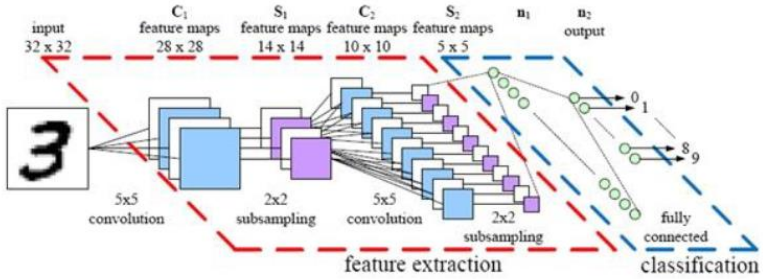


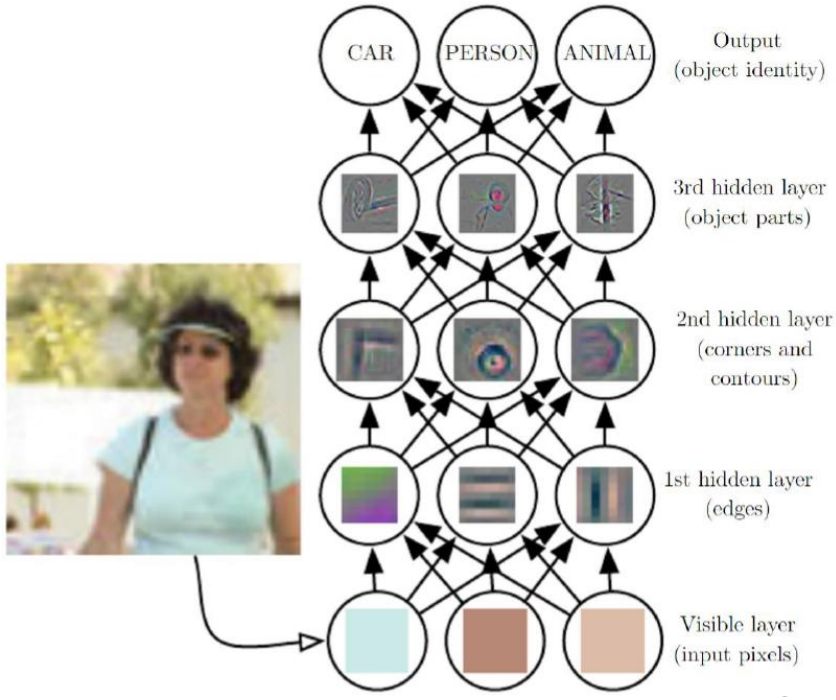
빅 데이터와 우리의 미래

· 자율주행자동차

- 자율주행자동차 구동의 핵심도 역시 다량의 데이터를 바탕으로 한 '딥러닝 기술'

- 이미지 데이터를 연산을 통해 컴퓨터가 인지할 수 있도록 한 기술이 기반이 됨





빅 데이터와 우리의 미래

· 자율주행자동차

- 테슬라 자율주행자동차: https://www.youtube.com/watch?v=EgDUjCqog2U

위 영상을 보여주시면 학생들 이해에 도움이 될 것 같습니다

References

· Bertoncello et al (2016). Monetizing car data. Mckinsey&Company

· Brynolfsson and McElheran (2016). The rapid adoption of data-driven decision-making. American Economic Review. 106(5). pp133-139

· Buvat and KVJ (2014). Doing business the digital way: How Capital One fundamentally disrupted the financial services industry. Capgemini Consulting

· Evans (2013). Business Analytics: Methods, Models, and Decisions. Pearson Education

· Herzog, T. H., Scheuren, F., & Winkler, W. E. (2010). Record linkage. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 535-543.

· Laney (2001). 3D Data management: Controlling data volume, velocity, and variety. Meta Group

· Welch and Westerman (2013). Caesars Entertainment: Digitally personalizing the customer experience. Capgemini Consulting