# WEEK 06 데이터 사이언스

OPEN CYBER UNIVERSITY OF KOREA

# 학습목표

- 1. 데이터 사이언스의 분야에 대해 이해한다.
- 2. 빅데이터 시대의 새로운 태도에 대해 고민한다.
- 3. 데이터가 바꿀 미래에 대해 고민하고 상상해 본다.

# 학습목차

- 1. 내 데이터 셋이 당신의 사고 방식을 바꾸게 해 주세요.
- 2. 빅데이터는 더 좋은 데이터
- 3. 의료 분야의 빅데이터 혁명

### 오늘의 수업 내용

### ◆ 데이터 사이언스

- ❖ 한스 로슬링 "Let my dataset change your mindset(내 데이터셋이 당신의 사고방식을 바꾸게 해주세요.)"
  - 개발 관련 통계를 그래픽 디스플레이로 변화시켜 제시
  - 이분법적 분류에서 벗어난 새로운 태도 필요
- ❖ 케네스 쿠키어 "Big data is better data(빅 데이터가 더 좋은 데이터)"
  - 머신러닝과 빅데이터
  - 빅데이터의 문제점: 예측에 의한 처벌, 일자리의 감소
- ❖ 조엘 셀라니키오 "The big-data revolution in health care(의료 분야의 빅데이터 혁명)"
  - 국제 보건 분야 자료 수집 방식의 혁명적 변화
  - 예방 가능한 질병의 자료를 디지털 형식으로 수집하는 방식





- 1) 데이터 사이언스란 무엇인가?
- 2) 오늘의 연사 소개
- 3) TED 강의 시청 및 강의 내용 정리/요약

# ◆ 백과 사전에서 찾아본 데이터 사이언스의 정의 (1/2)

- 데이터 과학(data science)이란, 데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다.
- 데이터 과학은 데이터를 통해 실제 현상을 이해하고 분석하는데 통계학, 데이터 분석, 기계학습과 연관된 방법론을 통합하는 개념으로 정의되기도 한다.
- 데이터의 구체적인 내용이 아닌 서로 다른 성질의 내용이나 형식의 데이터에 공통으로 존재하는 성질, 또는 그것들을 다루기 위한 기술의 개발에 착안점을 둔다는 특징을 가진다. 사용되는 기술은 여러분야에 걸쳐있으며 수학, 통계학, 계산기과학, 정보공학, 패턴인식, 기계학습, 데이터마이닝, 데이터베이스 등과 관련이 있다. 데이터 과학을 연구하는 사람을 데이터 과학자라고 한다.



- ◆ 백과 사전에서 찾아본 데이터 사이언스의 정의 (2/2)
  - 튜링상을 수상한 짐그레이(Jim Gray) 박사는 데이터 과학은 과학의 네번째 패러다임으로 정의하고 과학(경험, 이론, 계산, 그리고 이제 데이터)에 관한 모든 것이 바뀌고 있는데 이유는 정보기술과 데이터 범람(data deluge) 때문이라고 주장했다.
  - 데이터 과학은 생물학, 의학, 공학, 사회학, 인문과학 등의 여러 분야에 응용되고 있다.



- ◆ 앨런 튜링(Alan Turing)
  - ❖ 영국의 수학자·논리학자. 계산기가 어디까지 논리적으로 작동할 수 있는가에 대하여 처음으로 지적인 실험을 시도한 학자로 유명하며, 컴퓨터공학 및 정보공학의 이론적 토대를 마련하였다.



[출처] 두산백과(<u>http://www.doopedia.co.kr</u>) [사 진 출처] https://m.news1.kr/articles/?2098750#\_enliple



# ◆ 경영과학(Management Science) (1/2)

- ❖ 오퍼레이션 리서치(operations research:OR)를 중심으로 계량경제학적 분석이나 통계적 분석과 같은 수학적 기법에 의해서 경영현상을 해명하려는 넓은 의미의 경영학의 한 분야.
- ❖ 경영과학은 경영결정문제에 대한 과학적인 해결법을 개발함과 동시에, 의사결정의 기술로서 수학적 수법을 적용하여 문제에 대한 최적해(最適解)를 도출해 내는 결정 룰(decision rule)을 개발하려는 데 그 특징이 있다. 그리고 결정 룰을 컴퓨터프로그램화함으로써, 최적의 의사결정의 자동화가 이루어진다.
- ❖ 경영과학의 접근절차를 요약하면, 결정 모델의 구축에서 문제의 존재가 인식되면, 우선 현재의 시스템을 분석하고, 그 다음 현재의 시스템의 모델을 구축하며, 다시 새로운 모델을 개발하여, 실제로 테스트해 가는 순서를 거친다.

[출처] 두산백과(http://www.doopedia.co.kr)



- ◆ 경영과학(Management Science) (2/2)
  - ❖ 칸토로비치: 러시아의 수학자·경제학자. 1939년 발표한《생산조직과 생산계획의 수학적 방법》에서 선형계획(線型計畫)을 경제계획의 수단으로 응용하였다.
  - ❖ 단치그: 미국의 수학자로, 선형 계획법의 정립에 큰 기여를 하여, 흔히 "선형 계획법의 아버지"로 부른다. 댄치그가 고안한 단체법은 선형 계획법을 푸는 방법으로 널리 쓰이고 있다.





[출처] 두산백과(<a href="http://www.doopedia.co.kr">http://www.doopedia.co.kr</a>), 위키백과(<a href="http://ko.Wikipedia.org">http://ko.Wikipedia.org</a>)

[사 진 출처] <a href="http://whatsnewonnetflix.com">http://whatsnewonnetflix.com</a>



- ◆ 강연자 소개
  - ❖ Hans Rosling(한스 로슬링)
    - 스웨덴의 의사, 학자, 연설가, 국제보건학 교수, 갭마인더 재단 회장
    - 2017년 사망, '데이터를 생명으로 가져올 수 있는 사람'

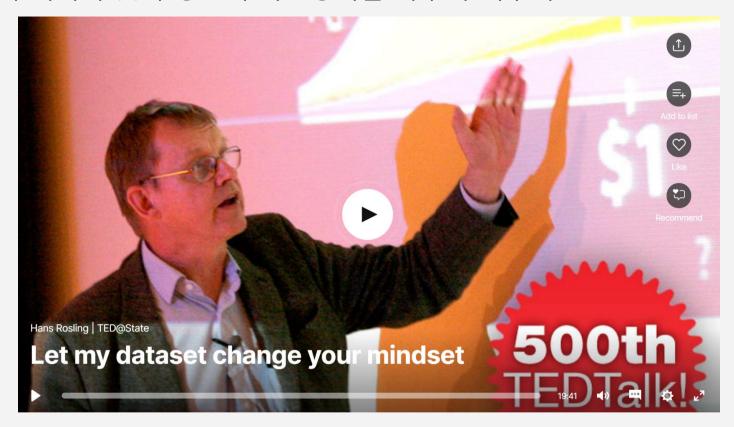


[출처] ko.wikipedia.org



# 3) TED 강의 시청 및 강의 내용 정리/요약

- ◆ TED Talks 시청
  - \* Hans Rosling, Let my dataset change your mindset
  - ❖ <내 데이터 셋이 당신의 사고방식을 바꾸게 해주세요.>



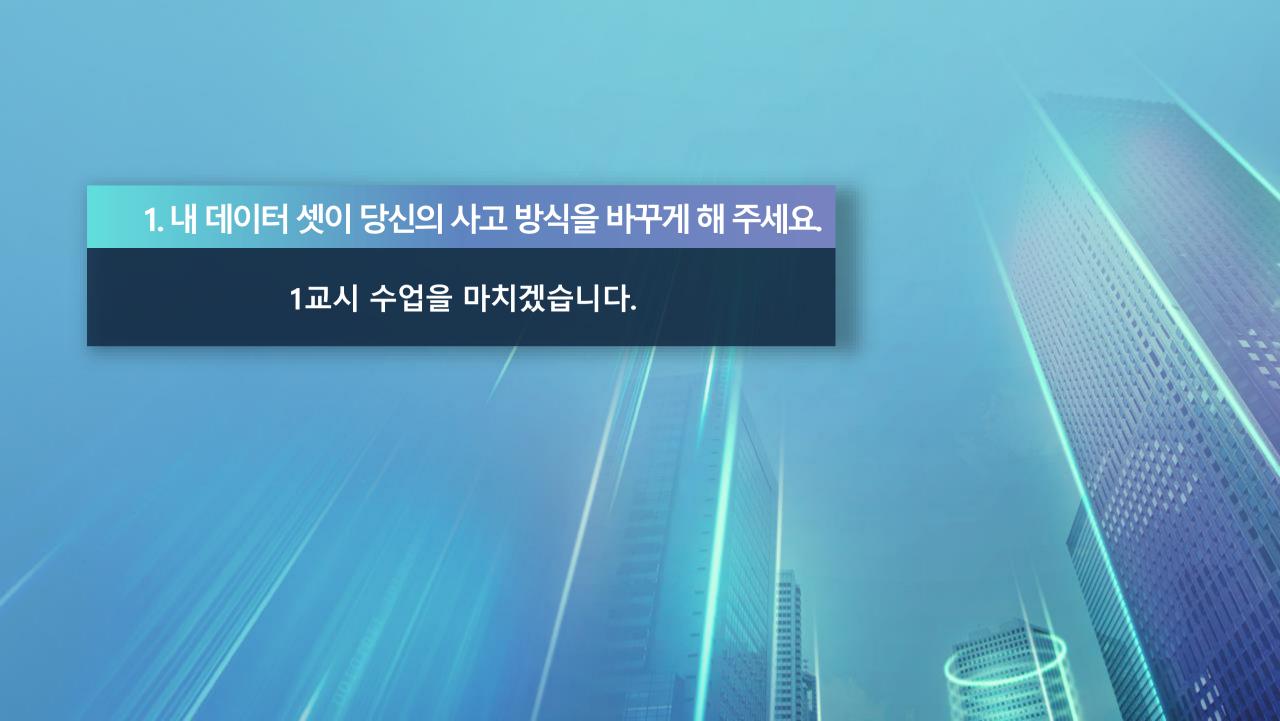


# 3) TED 강의 시청 및 강의 내용 정리/요약

#### ◆ 강연 요약

- ❖ Hans Rosling은 개발 관련 통계를 그래픽 디스플레이로 변환해 보여주는 Gapminder World를 만들었고, 개발 도상국에 대한 관점을 바꾸었다.
- ❖ '개발'의 측면에서 세상을 볼 때, 많은 사람들의 사고방식은 실제 자료와 일치하지 않는다.
- ❖ 또한, 서구 세계와 개발 도상국을 '우리', '그들'과 같이 나누어 정의한다.
- ❖ 그러나 이분법적인 분류는 의미가 없다.
- ❖ 개도국은 선진국의 발전 과정을 그대로 따라가고 있지 않고, 같은 국가 안에서도 큰 차이가 생길 수 있다.
- ❖ 우리에게는 새로운 태도가 필요하다.







- 1) 빅데이터란 무엇인가?
- 2) 오늘의 연사 소개
- 3) TED 강의 시청 및 강의 내용 정리/요약

### 1) 빅데이터란 무엇인가?

# ◆ 빅데이터(Big data) (1/3)

❖ 빅데이터는 포괄적인 개념으로, 단일화된 개념 정의보다는 다양한 관점에서의 정의가 존재한다. 미국의 IT 리서치 회사인 가트너(Gartner, Inc)는 정보 자산의 측면에서 "향상된 시사점(insight)과 더 나은 의사결정을 위하여 사용되는 비용 효율이 높고, 혁신적이며, 대용량·고속 및 다양성의 특성을 가진 정보 자산"이라 정의한 바 있으며, 미국의 경영컨설팅 회사인 맥킨지앤컴퍼니(McKinsey & Company)는 데이터 규모의 측면에서 "일반적인 데이터베이스 관리도구가 저장·관리·분석할 수 있는 범위를 초과하는 규모의 데이터"로 정의한 바 있다. 또한 미국의 IT 마켓리서치 회사인 IDC(International Data Corporation)는 업무수행의 측면에서 "다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고 데이터의 초고속 수집·발굴·분석을 지원하도록 고안된 차세대 기술 및 아키텍처"라고 정의한 바 있다.

[출 처] 두산백과(http://www.doopedia.co.kr)



### 1) 빅데이터란 무엇인가?

# ◆ 빅데이터(Big data) (2/3)

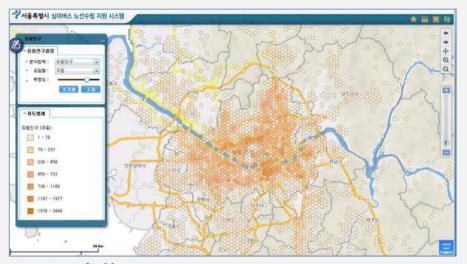
- ❖ 즉, 빅데이터는 스마트 기기의 대중화와 더불어 생활환경의 디지털화가 가속화되고 그로 인하여 발생하는 정보량이 기하급수적으로 증가함에 따라 데이터를 수집·축적·분석하여 새로운 가치를 창출하는 것이라 볼수 있으며, 이러한 빅데이터의 특징은 데이터의 종류와 양이 방대하고 생성 속도가 빠르며, 활용 가치가 높고 가변적이라는 점이다.
- ❖ 그러나 빅데이터 그 자체는 데이터의 단순한 집적물일 뿐이기 때문에, 이를 분석하는 기술과 전문 인력이 필수적이다.
- ❖ 빅데이터를 활용한 사례로는 소비 패턴 분석을 통한 상품 추천 서비스, 검색어 분석을 통한 독감예보 서비스, 과거 데이터에 기반한 범죄예방 시스템 구축, 유전자 데이터 공유를 통한 질병치료체계 구축, 지능형 교통안내 시스템, 유권자 데이터베이스에 기반한 유권자 맞춤형 선거전략 등이 있다.

[출 처] 두산백과(http://www.doopedia.co.kr)

### 1) 빅데이터란 무엇인가?

### ◆ 빅데이터를 활용한 버스 노선

- ❖ 서울시 심야버스 노선 수립은 공공부문 빅데이터 사업 중 가장 성공적인 사례로 평가 받음
- ❖ 폐기되거나 의미 없이 쌓였던 데이터를 활용해 실제 생활의 질 향상
- ❖ 서울시의 교통 데이터, KT의 통화량 통계 데이터를 합해 유동인구 밀집도 분석과 시각화 작업을 진행하여 노선별, 요일별 소비자의 패턴을 분석해 심야 버스 노선을 최적화함



[출 처] https://m.etnews.com/201312240348



# 2. 빅데이터는 더 좋은 데이터

# 2) 오늘의 연사 소개

### ◆ 강연자 소개

- ❖ Kenneth Cukier(케네스 쿠키어)
  - 미국 언론인이자 기술 및 사회 분야 작가
  - 홍콩 월스트리트 저널 아시아 판 기술 편집자
  - 공동 집필한 Big Data 관련 서적으로 유명(뉴욕 타임즈 베스트 셀러)
  - 전 세계 21개 언어로 번역되는 등 폭발적 인기를 누림



[출처] commons.wikipmedia.org



# ◆ Kenneth Cukier(케네스 쿠키어) (1/2)

- Kenneth Cukier(1968년 출생)는 미국의 언론인이자 기술 및 사회 분야의 작가이다. 그는 1999 년 프랑스 정부의 감시 능력을 설명하기 위해 "Frenchelon"이라는 용어를 만들었고, 2001 년 홍콩 월스트리트 저널 아시아 에디션의 기술 편집자로 활동하며 The New York Times, Financial Times, Foreign Affairs 외의 여러 간행지에도 기고를 하였다. 또한 2008 년 그는 International Bridges to Justice의 이사회에 임명되었으며, 2015 년 그는 열악한 어린이들에게 클래식 악기를 제공하는 Open String Foundation의 이사회에 합류했다. 그는 2016 년에는 영국 국제 문제 연구소인 Chatham House의 관리 위원으로 선출되었고 2017 년에는 University of Oxford's Said Business School에서 associate fellow로 지명되어 인공 지능과 비즈니스에 관해 강의를 하였다.
- 특히, 그는 The Economist에서의 활동과 2013년 Viktor Mayer-Schönberger와 공동 집필한 'Big Data: A Revolution that Will Transform How We Work, Live and Think'이라는 책으로 유명하다.



# ◆ Kenneth Cukier(케네스 쿠키어) (2/2)

- 저서인 Big Data는 New York Times 베스트셀러로 선정되어 전 세계 21개 언어로 번역되었으며, Financial Times와 McKinsey Business Book of the Year Award의 최종 후보였다. 이후 둘은 Big Data의 후속 연구인 Learning With Big Data: The Future of Education를 발표하기도 하였다.
- 이외에도 Cukier는 개인 정보 보호가 소규모 데이터 시대의 핵심과제로 빅 데이터 시대에는 도전이 자유 의지, 도덕적 선택, 인간 의지, 인간 기관을 보호하는 것이라 하였다. 또한 현재의 빅 데이터 혁명을 구텐베르크 활자(Gutenberg Press) 및 인쇄 혁명(Printing Revolution)과 유사한 변화의 주요 원인이라 언급하며, 이는 큰 이점이 있지만 '기본적인 자유를 유지하고' 빅 데이터가 Orwell's Big Brother의 다른 버전이 되는 것을 막기 위해 제한이 필요하다고 경고했다. 그는 빅데이터와 데이터 사이언스에 대한 그의 활동으로 오하이오 주(Ohio State) 비텐 베르크 대학교(Wittenberg University in Springfield)에서 Humane Letters 명예박사 학위를 수여 받았다.



# 3) TED 강의 시청 및 강의 내용 정리/요약

- ◆ TED Talks 시청
  - \* Kenneth Cukier, Big data is better data
  - ❖ <빅데이터는 더 좋은 데이터>





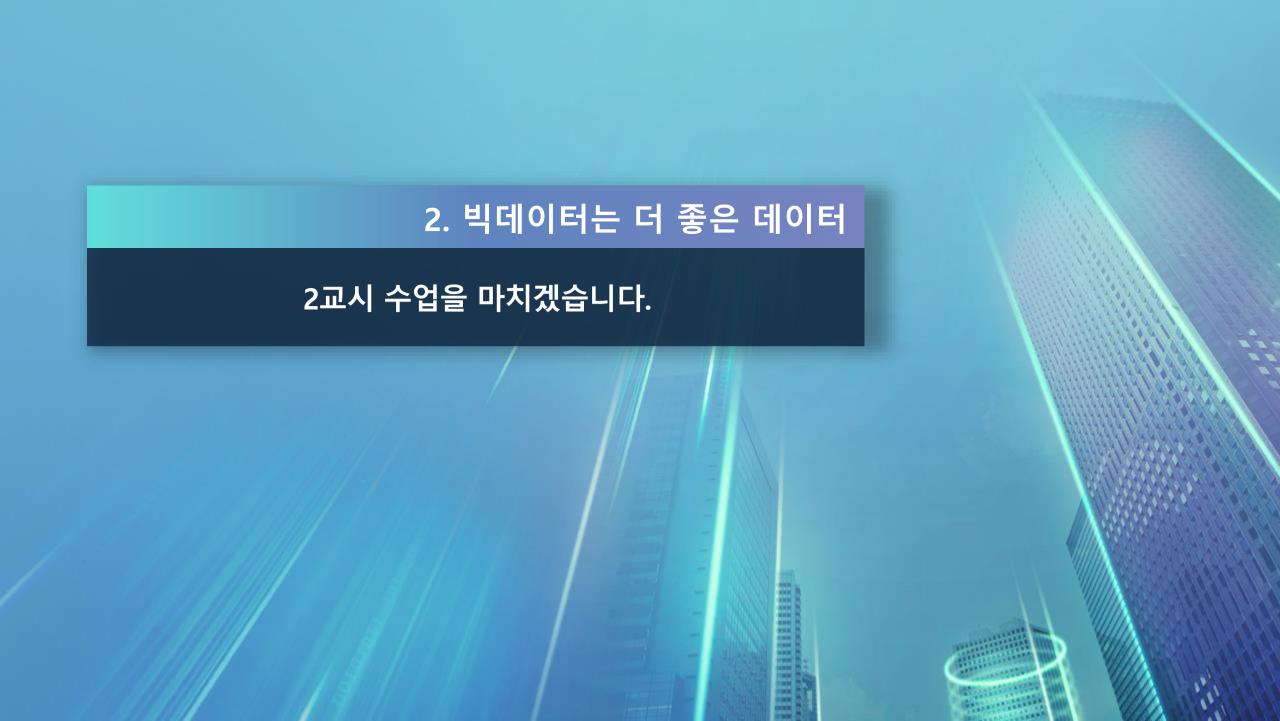


# 3) TED 강의 시청 및 강의 내용 정리/요약

### ◆ 강연 요약

- 지금은 데이터가 넘쳐나는 사회이다.
- 이전의 어느 때보다 많은 양의 자료를 쉽게 가공하고, 공유, 저장할 수 있게 되었다.
- 빅데이터는 많은 것을 보고, 새로운 것을 배울 수 있게 한다.
- 빅데이터 활용의 인상적인 한 분야는 기계학습이다.
- 사람이 기계에게 문제 해결 방식을 직접 설명하지 않고, 기계가 데이터들의 공통점을 직접 파악해 문제를 처리하게 만드는 것이다.
- 반면, 빅데이터는 문제점도 야기할 수 있다.
- 먼저, 우리가 취할 행동을 예상하는 알고리즘이 발명된다면, 예측에 의해 처벌을 받을 수도 있다.
- 또한, 일자리의 감소도 문제이다.







- 1) 의료 분야와 데이터 사이언스
- 2) 오늘의 연사 소개
- 3) TED 강의 시청 및 강의 내용 정리/ 요약

### 1) 의료 분야와 데이터 사이언스

### ◆ 구글의 독감 예측 시스템

- ❖ 구글의 독감 관련 검색어 활용
- ❖ 어느 지역에 독감이 많이 발생하는지 미국 지도와 함께 공시
- ❖ 우리나라의 국민건강보험공단에서도 감기, 눈병, 식중독, 천식, 피부염의 알람 서비스 제공(기간별 진료 데이터 및 SNS 상의 키워드 빈도수 기반 질병 예측 시스템)

### ◆ 삼성 서울 병원의 의료 데이터 활용

- ❖ 유전적, 의학적 현상, 평상시 생활 등의 개인 의료 데이터 기반으로 개인별 맞춤 치료 서비스 제공
- ❖ 의료 데이터를 분석 후 개인에 맞는 건강관리 및 치료 방법 제안
- ❖ 더 나아가 질병 또한 예측하는 서비스
- ❖ 치료 중심이 아닌 건강 관리 및 예방에 초점



### 1) 의료 분야와 데이터 사이언스

### ◆ 동아 ST 제약회사의 빅데이터 기술 활용

- ❖ 아주대 병원과 함께 복합체 개발 과정에서 빅데이터 기술 활용
- ❖ 처방전 및 약물의 부작용 사례를 활용해 약 개발 시 비용과 리스크 감소
- ❖ 예를 들어, 일반적으로 관절염 약과 소화제를 같이 처방하는 경우가 많은데, 복합제를 개발하는 등 빅데이터 기술을 활용

### ◆ IBM 왓슨(Watson)의 의료 분야 활용

- ❖ IBM의 인공지능 슈퍼컴퓨터 왓슨을 활용하여 의료 데이터를 분석 후 환자를 진단하여 의사에게 치료법을 추천함
- ❖ 수백만 건의 진단서, 환자 기록, 의료 서적 등 방대한 데이터를 바탕으로 빠른 분석을 통해 확률이 가장 높은 병명, 성공 가능성이 높은 치료법을 동시에 의사에게 조언함



### ◆ 강연자 소개

- ❖ Joel Selanikio(조엘 셀라니키오)
  - 미국 조지타운 대학병원의 소아과 주치의 겸 소아과 교수
  - 처음에는 대학 졸업 후, 시스템 분석가로 일하다 의학 학위를 받고 질병관리본부에서도 근무함
  - 공공보건에 대한 데이터를 수집하기 위한 오픈소스 소프트웨어를 만들기 위해 DataDyne.org를 설립



[출 처] commons.wikipmedia.org



# ◆ Joel Selanikio(조엘 셀라니키오) (1/2)

- Joel Selanikio는 미국 조지타운 대학병원의 소아과 주치의 겸 소아과 조교수로 1986년 필라델피아 해버포드 대학(Haverford College)을 졸업하고 체이스 맨해튼(Chase Manhattan)에서 시스템 분석가로 일했다. 브라운 대학(Brown University)에서 의학 학위를 취득했고 조지아 주 애틀랜타에서 레지던트를 했다. 이후 Selanikio는 질병관리본부(CDC; Center for Disease Control and Prevention)에서 근무했으나, 공공 보건에 관한 데이터를 수집하기 위한 오픈 소스 소프트웨어를 만들기 위하여 직장을 떠나 DataDyne.org 회사를 시작하였다.
- 그는 2004년 인도양 지진과 쓰나미 이후 셀라니키오는 인도네시아 아체에서 국제구조위원회와 함께 일했으며, 2014~2015년 시에라리온에서 에볼라가 유행하는 동안 루나르에서 국제의료단과 함께 일했다.



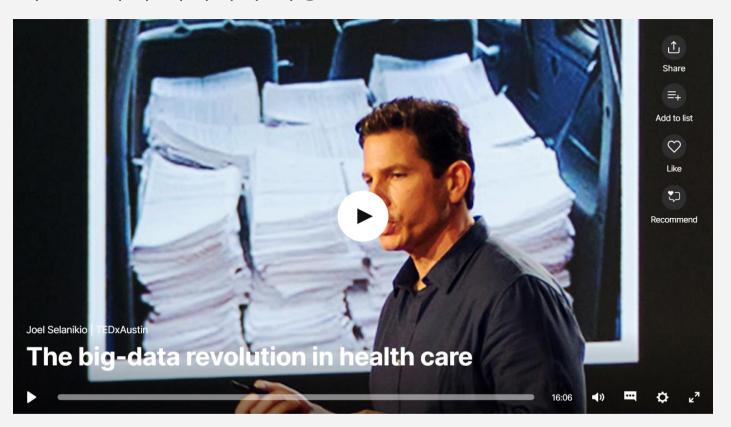
# ◆ Joel Selanikio(조엘 셀라니키오) (2/2)

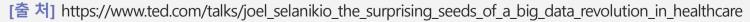
- 또한 2013년의 TED를 비롯하여 2009년 Lemelson-MIT Innovation Prize Lecture, 2010년 World Economic Forum "Tech for Society" Panel, Davos, 2011년 Royal Society of Medicine Lecture, 2013년 Ivey Global Health Conference 등에서 강연을 하였다.
- 이외에도 그는 2005년 Haverford College Award, 2008년 Tech Museum Award for Health, 2008년 Stockholm Challenge, 2009년 Fast Company Magazine Social Enterprise of the Year, 2009년 Wall Street Journal Award for Technological Innovation in Healthcare, 2009년 Lemelson-MIT Prize \$100,000 Award for Sustainable Innovation, 2012년 ComputerWorld 21st Century Achievement Award 등 다양한 상들을 수여 받았다.



# 3) TED 강의 시청 및 강의 내용 정리/요약

- ◆ TED Talks 시청
  - ❖ Joel Selanikio, The big-data revolution in health care
  - ❖ <의료 분야의 빅데이터 혁명>







# 3) TED 강의 시청 및 강의 내용 정리/요약

### ◆ 강연 요약

- 국제 보건 분야에서는 자료를 수집할 때 대부분 종이를 이용하였다.
- 종이를 이용하면 결과를 컴퓨터에 입력해야 해서, 데이터 분석을 시작하기까지 많은 시간이 소요된다.
- 때문에, 예방 가능한 질병인데도 조치를 받지 못해 아이들이 사망하는 일들이 일어났다.
- Joel Selanikio는 이 문제를 해결하기 위해 여러 프로그램에서 자료를 디지털 형식으로 수집하는 방식을 교육하였다.
- 또한, 만나서 교육할 수 있는 사람의 수에는 한계가 있기 때문에,
   소프트웨어 '매그파이'를 만들어 직접 기기에 자료를 입력할 수 있게 하였다.



### 주차 정리

### ◆ 데이터 사이언스

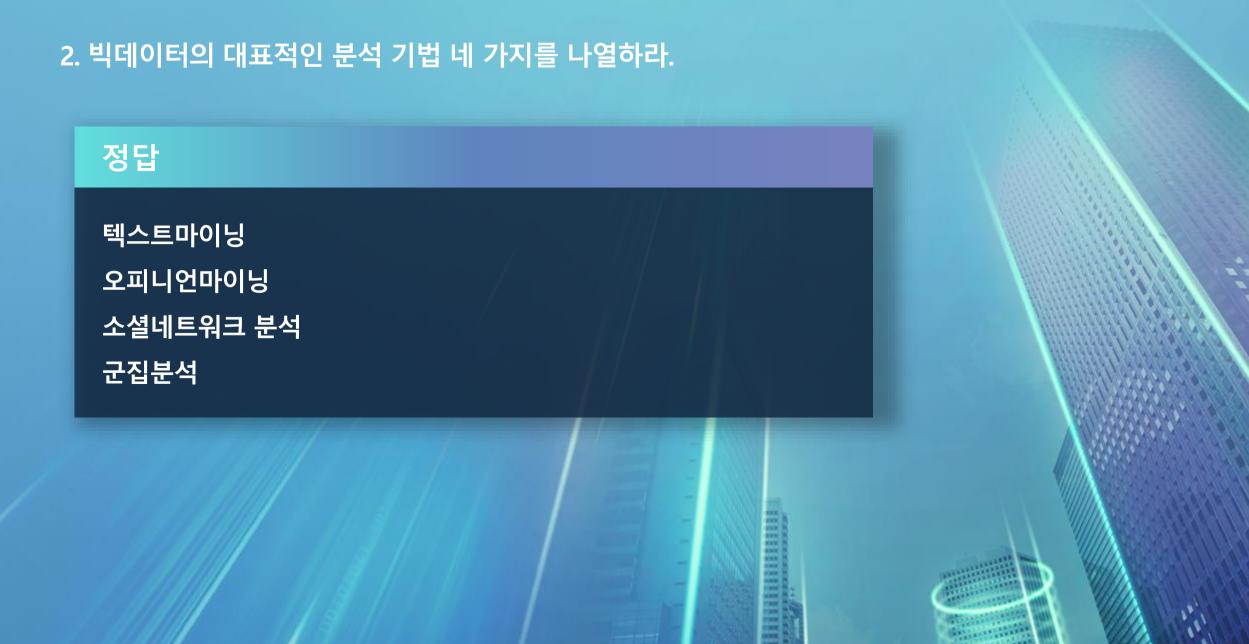
- ❖ 한스 로슬링 "Let my dataset change your mindset(내 데이터셋이 당신의 사고방식을 바꾸게 해주세요.)"
  - 개발 관련 통계를 그래픽 디스플레이로 변화시켜 제시
  - 이분법적 분류에서 벗어난 새로운 태도 필요
- ❖ 케네스 쿠키어 "Big data is better data(빅 데이터가 더 좋은 데이터)"
  - 머신러닝과 빅데이터
  - 빅데이터의 문제점: 예측에 의한 처벌, 일자리의 감소
- ❖ 조엘 셀라니키오 "The big-data revolution in health care(의료 분야의 빅데이터 혁명)"
  - 국제 보건 분야 자료 수집 방식의 혁명적 변화
  - 예방 가능한 질병의 자료를 디지털 형식으로 수집하는 방식



1. 데이터 사이언스의 정의에 대해 설명하라.

### 정답

데이터 과학(data science)이란, 데이터 마이닝(Data Mining)과 유사하게 정형, 비정형 형태를 포함한 다양한 데이터로부터 지식과 인사이트를 추출하는데 과학적 방법론, 프로세스, 알고리즘, 시스템을 동원하는 융합분야다.



3. 조엘 셀라니키오는 국제 보건 분야에서 자료를 수집 시, 대부분 종이를 이용하던 것이 데이터 분석에 가장 큰 장애물이라고 보았다.



# 정답:0

맞는 설명(T)이다.

이에 대한 해결책으로 소프트웨어 매그파이를 만들어 직접 기기에 데이터를 입력할 수 있도록 했다.

4. 케네스 쿠키어는 빅데이터 활용의 인상적인 분야를 기계학습으로 보고 인류의 일자리가 더 늘어날 것으로 예측했다.



정답 : X

틀린 설명이다.

케네스 쿠키어는 일자리의 감소로 문제가 생길 수 있다는 점을 지적했다. 또한 예측에 의해 처벌을 받을 수 있는 등, 빅데이터의 문제점도 함께 지적했다.

#### 차주 예고

#### ◆ 글로벌 경영

- ❖ 스캇 스와스트, The global business next door<글로벌 기업의 다음 문>
  - 성공적인 글로벌 회사는 세계적인 대기업이 아니다?
  - 수출 기업은 국내 기업보다 많은 성장 기회를 가지고 있다.
- ❖ 마르그레테 베스타게르, The new age of corporate monopolies <기업 독점의 새로운 시대>
  - 규제와 경쟁 조약은 같은 수준에서 경쟁할 수 있도록 한다.
  - 시장이 모두를 위한 시장이기 위해서 적절한 경쟁 규제가 필요하다.
- ❖ 아네트 하우저, The 3 agencies with the power to make or break economies<경제를 만들거나 무너뜨릴 수 있는 힘을 가진 3개 단체>
  - 아네트 하우저는 국가신용 비영리 평가 기관 INCRA 개발하였다.
  - 미국의 신용 평가 시장은 세 기관이 독점하고 있으며, 등급 산출 기준을 공개하고 있지 않으나, INCRA는 평가 내용, 지표, 방법을 모두 공개한다.



