빅데이터는 나의 친구_나도 빅데이터 전문가 – 교육자료

1차시. 프롤로그

빅데이터
- 기존 데이터베이스 관리 도구의 능력을 넘어서는 수십 테라바이트의 정형 데이터, 또는 비정형 데이터를 포함한 데이터라고 할 수 있다.
- 그 규모가 방대하며, 생성 주기도 짧고, 수치 데이터뿐 아니라 문자와 영상 데이터를 포함하는 대규모 데이터이다.
- 기존 데이터베이스 관리 도구의 데이터 수집, 저장, 관리, 분석하는 역량을 넘어서는 규모를 말한다.

전 산업분야에서 빅데이터를 활용하는 사례
- 제품 개발에 활용된다.
- 예측적 유지 보수에도 활용된다.
- 사기를 방지할 수도 있다.

2차시. 빅데이터로 변화되는 우리의 일상

코로나19와 빅데이터에 대한 설명
- 우리나라의 방역시스템이 성공적일 수 있었던 이유 중 하나는 빅데이터 활용 역량이다.
- 빅데이터 분석을 통해 코로나 확진자 이동경로 파악과 접촉자 선별, 감염경로 파악, 감염확산 예측 등이 가능했다.
- 빅데이터 분석을 기초로 정부는 코로나19 극복에 대한 유용한 정보와 지침을 국민들에게 제공했다.

빅데이터를 활용하여 우리 삶의 질을 향상시킨 사례
- 심야버스 노선 최적화
- 빅데이터 기반 코로나 방역
- 농산물 가격 예측시스템

빅데이터가 우리 사회에 끼치는 긍정적인 영향
- 빅데이터를 통해 공공부문의 개인 맞춤형 서비스 제공 가능하다.
- 빅데이터를 선점한 기업들의 가치가 증가할 것이다.
- 시민을 위한 공공부문 서비스는 국가에 의해 획일적으로 조정되어 왔으나, 빅데이터로 인해 맞춤형 서비스를 제공할 수 있다.

- 2013년 미국의 국가안보국이 미국민들의 수백 만 건의 통화 기록과 인터넷 데이터 등의 개인정보를 무차별적으로 수집하고 감시하고 있다는 내용이 폭로된 바 있으며, 우리나라에서는 수사기관의 정당한 법 집행이란 이유로 카카오톡이 감청을 허용해준 사건이 있었다.

코로나 19에 대한 건강보험심사평가원의 의료 빅데이터 개방 사례
- 연구자들이 백신을 개발하기 위해 노력을 하였지만 '임상 데이터' 부족으로 결론에는 도달하지 못하고 있었다고 한다.
- 심평원이 보건의료 빅데이터를 개방하여 '코로나19 국제협력 연구'를 시작하면서 코로나19 연구에 속도가 나기 시작하였다고 한다.
- 단일보험체제인 우리나라 건강보험 제도적 특성으로 인해 보건의료 빅데이터가 구축되어 있었다.

3차시. 빅데이터가 뭐길래 난리야?

빅데이터는 다양한 정의
- 가트너 : "빅데이터는 큰 용량, 빠른 속도, 그리고 높은 다양성을 갖는 정보 자산이다"
- 맥킨지 : "빅데이터란 전형적인 데이터베이스 소프트웨어 툴의 능력으로 수집, 저장, 관리, 분석할 수 없는 규모의 데이터 집합을 의미한다."
- 삼성경제연구소 : "기존의 관리 분석 체계로는 감당할 수 없을 정도의 거대한 데이터의 집합을 지칭한다."

기존 데이터와 빅데이터와의 차이점
- 기존 데이터는 관계형 데이터 처리자가 수십 년 동안 저장 및 처리해 왔다.
- 과거 전 세계 데이터의 대부분을 차지하고 있었다.
- 기존 데이터는 조작하기 쉽고 기존의 데이터 처리 소프트웨어로 관리할 수 있었다.

- 전통적인 데이터와 빅데이터 비교 요소로는 크기, 구성방법, 분석에 사용되는 방법, 파생되는 소스, 관리에 필요한 아키텍처 등이 있습니다.
- 기존 데이터는 기가바이트와 테라바이트 단위 측정, 빅데이터는 일반적으로 페타바이트, 제타바이트 또는 엑사바이트 단위로 측정한다.
- 빅데이터의 경우 초 단위로 생성되므로 데이터가 수집되는 동안 에도 분석할 수 있다.
- 정형데이터는 일반적으로 관계형 데이터베이스라고 하며, SQL을 사용하여 작성, 읽기 및 조작 가능하다.
- 구글 스프레드시트 또는 마이크로소프트 엑셀과 같은 스프레드시트 프로그램에 표시할 수 있는 모든 데이터를 말하는 것은 비정형 데이터가 아닌 정형데이터이다.
- 가트너가 제시한 빅데이터의 가장 대표적인 특징으로 3V는 데이터의 크기와 데이터의 속도, 데이터의 다양성이다.

4차시. 빅데이터는 어떻게 등장했지?

- 문자 형태의 숫자 체계는 기원전 4000년경에 이란에서 처음 사용되었을 것으로 추정되며, 장사를 위한 수단으로 사용된 이 새로운 방법은 점토판을 사용해서 각 상품마다 다른 숫자를 표기하는 것이었다.

- "무선통신이 완벽하게 적용되면 지구 전체가 거대한 두뇌로 전환될 것이다." 등으로 예언한 사람은 미국인 전기공학자이자 과학자, 발명가인 니콜라 테슬라이다.

- 빅데이터의 등장배경에는 페이스북, 트위터 등 소셜 네트워크 서비스 이용이 확산하고 소통방식이 변화하면서 소셜 데이터 혁명 발생이 작용되었다.

- 음성과 문자, 멀티미디어 등 다양한 비정형 데이터가 디지털화를 통하여 가공 및 유통이 자유롭게 이루어지면서 모바일과 소셜 네트워크 서비스 등 다양한 매체를 바탕으로 폭발적으로 증가하게 되었으며, 이러한 비정형 데이터를 통하여 고객의 숨겨진 욕구와 기호를 빠르게 파악하고 새로운 수익을 창출할 수 있을 것이라는 기대하고 있다.

5차시. 국가 데이터 정책_대한민국 데이터 119 프로젝트

- 데이터 3법 개정의 주요 내용으로 모호한 '개인정보' 판단 기준의 명확화를 들 수 있다.

- 데이터 3법 개정으로 민간의 데이터 활용 기반을 마련하였지만, 민간은 아직도 필요한 데이터 확보가 어렵거나 데이터 활용에 관한 제도가 미흡하다는 애로를 호소하고 있다.

- 데이터 분야 선제적 전략투자로 세계 최고의 경쟁력을 보유하고 있는 나라는 미국이다.

국가 데이터 정책의 기본방향
- 국민을 배려하는 사람과 서비스 중심
- 개별 정책보다는 체계적인 거버넌스 중심 설정
- 11대 실천과제와 9대 서비스 도출

국가 데이터 정책의 주요 내용
- 국민의 데이터 주권 강화를 위해 마이데이터 이용을 활성화
- 기업이 데이터를 활용하여 산업경쟁력을 강화하도록 지원하고,
- 장기적으로 범 국가 차원의 데이터 관리전략을 수립

6차시. 빅데이터가 꿈꾸는 미래_스마트시티를 중심으로 _ 1부

- 스마트주차시스템은 앱을 활용하여 유휴 주차장 정보 파악 및 예약을 통해 주차면 확보하는 시스템이다.

'스마트시티'에 대한 정의
- "첨단 정보통신 기술(ICT)과 빅데이터 등 신기술을 접목해 각종 도시 문제를 해결하고 삶의 질을 개선할 수 있는 도시 모델"
- "경제, 이동성, 환경, 인간, 생활, 행정 등 다양한 주요 분야에서 우수하고 지속 가능한 경제 발전과 높은 삶의 질을 창출하는 발전된 도시"
- "첨단 정보통신 기술을 이용해 도시의 모든 인프라를 네트워크화한 미래형 첨단 도시"

- 구 도심지를 재개발하는 사업은 스마트시티와 관련이 없다.

- 건물에너지관리시스템은 빌딩 내의 에너지 관리 설비의 다양한 정보들을 실시간으로 수집하고 분석함으로서 건물 에너지를 효율적으로 사용할 수 있는 시스템이다.

- 스마트 쓰레기 관리는, 쓰레기통에 센서를 설치하여 쓰레기통이 비어 있을 경우에는 그대로 지나가고, 가득 채워져 있을 경우에만 쓰레기를 처리하여 청소업무의 효율성을 높이는 시스템이다.

- 지능형 방범서비스는 특히, 어린이 보호구역, 우범 지역, 골목길, 공원 등에 설치하여 그 효과를 극대화할 수 있다.

- 스마트시티 인덱스에서 제시된 글로벌 스마트시티 구축 동향에서

- 지구 전체 면적에서 도시가 차지하는 비중은 2%에 불과하지만, 지구 온실가스의 70%는 도시에서 만들어지고 있다고 한다.

7차시. 빅데이터가 꿈꾸는 미래_스마트시티를 중심으로 _ 2부

- 스마트 교통은 운전자가 실시간으로 기상여건을 확인할 수 있게 함으로써 안전운전을 보조해주는 서비스이다.

스마트시티가 가져올 일상의 변화에서의 에너지 자립화
- 공공건물 및 유휴공간에 태양광 발전을 비롯한 재생에너지 인프라를 확충하여 도시 내 필요한 전력을 스스로 생산한다.
- 시민들이 스마트 그리드를 이용하여 전력을 거래할 수 있다.
- 프랑스, 스웨덴, 네덜란드, 중국 등의 국가는 도로태양광을 도입하여 시범사업을 추진하고 있다.

부의 제3차 스마트도시 종합계획에서 밝힌 스마트시티 구축 필요성
-인구와 경제분야에서 저출산 · 고령화 심화, 저성장 · 공유경제 등 산업구조 변화를 들 수 있다.
- 기후 및 환경분야에서 기후변화 및 환경오염으로 지속가능한 도시모델로 될 수 있다.
- 기술과 산업분야에서 4차산업혁명으로 초연결 · 지능사회 출현 및 신산업이 대두되고 있다.

스마트 시티 국가시범단지
- 세종시는 공공데이터에 기반한 스마트도시로, 7대 서비스 구현에 최적화된 공간 계획을 마련하고 있다.
- 부산 에코델타시티는 로봇 및 물 관리 관련 신산업 육성을 중점적으로 추진하고 있다.
- 세종시는 AI를 활용한 모빌리티, 헬스케어, 교육, 에너지·환경, 거버넌스, 문화·쇼핑, 일자리에 대한 최적화된 공간 계획을 마련하고 있다.

8차시. 정부는 빅데이터로 뭘 할까?_공공부문

공공부문 빅데이터 활용으로 기상청과 농촌진흥청의 기상데이터와 농산물 생산성 예측 모델
- 농수산물의 수매를 예측하기란 매우 어려우며, 그 이유는 기후와 시장 상황 변동으로 인하여 매년 달라지기 때문이다.
- 정부가 수급조절을 위해 노력하고 있으나 역부족으로, 농산물 가격을 예측할 수 있다면 수급조절이 가능하다.
- 기상청과 농촌진흥청은 농작물에 영향을 미치는 기상요인을 분석하여 '농작물 생산성 예측 모형'을 개발하였다.

9차시. 기업은 빅데이터로 뭘 할까?_마케팅부문

구글의 독감 확산 예측모델인 구글 플루 트렌즈 자사의 마케팅 활용
- 구글은 감기나 독감 검색빈도가 높은 지역을 지도에 표시함으로써 독감의 확산에 대한 예측이 가능하다고 판단하였다.
- 구글은 2009년도부터 검색정보와 위치를 기반으로 감기 바이러스 확산 상황을 알려주는 플루 트렌드 서비스를 제공하였다.
- 구글의 입장에서는 각종 언론에 자신들의 존재를 알림으로서 혁신적인 IT기업의 이미지를 강화할 수 있었다.

10차시. 기업은 빅데이터로 뭘 할까?_제조부문

인더스트리 4.0
- 인더스트리 4.0의 개념은 독일 신제조업 전략이라고 할 수 있다.

- 인더스트리 4.0의 최초명칭은 "사이버물리생산체계"라는 딱딱한 이름이었고, 개념도 복잡하였다.
- 메르켈 총리에 의해 국민들이 이해하기 쉬운 단순하고 상징적 표현인 인더스트리 4.0으로 정해지게 되었다고 한다.

- 독일이 인더스트리 4.0을 준비하게 된 계기와 우리나라의 IT 및 자동차 산업과는 직접적인 관련이 없다.

스마트팩토리에서 각종 센서의 역할
- 제조설비 및 엔지니어링 기구에 센서와 RFID가 장착되어 있어 기계가 사용되고 있는 환경에 대한 모든 필수적인 정보들이 관리자에게 전송된다.
- 관리자는 제품에 센서를 설치함으로써 제품이 생산되는 전 과정을 한자리에 앉아서 모니터링 하면서 피드백한다.
- 기계들끼리 M2M 시스템을 통해 자동으로 작업 지시가 이루어짐으로써 생산의 효율성이 높아진다.

- 가상물리생산시스템(CPPS)의 일종으로, 현실에 존재하는 설비를 디지털로 똑같이 구현하는 것을 디지털 트윈이라고 한다.

- 수율 개선으로 생산성을 높이고 비용을 줄이는 것 역시 빅데이터 도입으로 얻을 수 있는 명확한 장점이라고 할 수 있다.

11차시. 기업은 빅데이터로 뭘 할까?_의료부문

정부의 데이터 3법 개정과 보건의료 빅데이터 활용과의 관계
- 보건의료 관련 각 기관에 분산된 빅데이터를 연계하고 통합한 후 비식별화하여 민간 연구자에게 제공하는 것이 가능해졌다.
- 다양한 분야에 의료 빅데이터가 적극 활용될 전망이다.
- 빅데이터 기반 제품과 서비스의 질 개선 및 신규 비즈니스 모델 발굴이 가능할 것으로 기대되고 있다.

우리나라의 보건의료 빅데이터의 특징에 대해 설명
- 전 국민 단일 건강보험 체제인 우리나라는 다른 나라에 비해 보건의료 빅데이터가 매우 잘 축적돼 있다.
- 많은 의료기관들이 전자의무기록(EMR)을 도입해 디지털 의료정보를 보유하고 있다.
- 데이터의 소유권과 의료정보의 개인정보 보호, 데이터의 표준화 부족 등 많은 이슈들이 산재되어 있는 것도 사실입니다.

우리나라의 보건의료 빅데이터 현황
- 건강보험공단이 보험료와 진료, 검진 등 보유한 데이터 건수는 약 3조4천억 건이다.

- 건강보험심사평가원은 진료과 투약 내역, 의약품 등 약 3조건을 축적해둔 상태이다.
- 국립암센터도 국내 암 발생 현황을 일괄 보유하고 있다.

- 개인정보보호법 제23조에 따르면 건강에 대한 정보를 가명처리하고 정보제공 주체인 개인의 '동의'를 받아야만 이를 활용할 수 있다고 명시되어 있으나, 유전정보 또는 다른 정보와의 결합 등으로 개인식별이 가능할 수 있으므로 충분히 안전하다고 볼 수 없다.

- 보건의료 빅데이터 활용의 목적은 빅데이터 기반 제품과 서비스의 질 개선 및 신규 비즈니스 모델 개발이다.

12차시. 기업은 빅데이터로 뭘 할까?_통신, 에너지, 금융, 항공, 교육부문

빅데이터와 통신부문의 연관성
- 통신업계에서의 빅데이터 적용은 효과적인 마케팅과 사업전략 수립을 통해 이익을 극대화할 수 있다.
- 통신분야는 데이터의 송수신으로 인해 데이터의 양이 폭발적으로 늘어나고 있다.
- 통신 빅데이터에는 경제 활동 현황과 인구 이동 패턴 등 다양한 정보가 있다.

통신 빅데이터와 코로나 19 대응
- 도시의 유동인구 현황을 파악하고 의료 자원을 배분하는 등 코로나 확산 방지 정책에 활용하였다.
- 정부가 코로나19 확산을 막기 위해 이동을 제한하는 정책을 시행했을 때, 실제로 정책 효과가 있었는지 확인하는 데도 유용하게 활용하였다.
- 한국을 포함한 일부 국가에서는 통신 데이터를 전염병 확산을 연구하기 위한 모델링 연구에도 활용하였다.

통신 빅데이터를 활용한 서울시 생활 이동 데이터
- 코로나19 대응이나 통근 및 통학시간 개선 등 서울시의 각종 정책 수립에 활용될 예정이다.
- 출퇴근 시간 혼잡도가 높게 나타난 노선의 증차를 결정하거나 대중교통 인프라 수요가 높은 지역에 버스노선을 신설할 수도 있다.
- 청년 공공주택 부지로 2~30대 통근인구가 많은 지역을 선정하는 등 도시 공간 구조개선에도 활용될 전망이라고 한다.

에너지 분야에서의 통신 빅데이터 활용
- 전봇대의 전력데이터 정보를 바탕으로 폭염 때 특정 지역 소비자 전력 이용량을 예측해 단계별로 에너지 절감 문자메시지를 전송한다.
- 지진 발생 시 전봇대의 기울기나 진동 등 충격 감지 정보를 수집해 사고 대비·복구에 활용한다.
- 독거노인의 전력 사용량이 갑자기 줄면 사회복지사에게 자동으로 통보해주기도 한다.

기업은 은행에 개인정보 이용 및 활용동의서를 제출하기 때문에 은행은 기업 여신 심사에 빅데이터를 활용할 수 있으며, 빅데이터를 활용하여 기업 관련 중요 정보를 객관적으로 분석하고 부실 징후를 예측할 수 있다.

13차시. 나는 빅데이터로 뭘 할까?_창업

- 공공데이터란 공공기관이 만들어내는 모든 자료나 정보, 국민 모두의 소통과 협력을 끌어내는 공적인 정보를 말하며, 무료로 누구나 이용 가능하다.

- 공공데이터 신청이 반려된 경우 명확한 근거를 바탕으로 신청인에게 반려된 이유를 통보해 주며, 제공신청이 반려됐다면 분쟁조정 신청으로 한 번 더 공공데이터 요청이 가능하다.

14차시. 누구나 알 수 있는 빅데이터 분석기법

빅데이터 분석
- 빅데이터 분석은 대량의 데이터로부터 숨겨진 패턴과 알려지지 않은 정보를 찾아내기 위한 과정이다.
- 개인이나 기업 등에서 자료를 토대로 어떠한 의사 결정을 할 때에 중요한 정보로써 사용되다.
- 효과적인 데이터 분석을 위해서 일반적으로 빅데이터 분석 플랫폼을 구축하는 경우가 많다.

15차시. 데이터가 스스로 답한다_AI

- 과거 인공지능은 대량의 데이터를 몇 초 만에 분석할 수 없었으나, 현재는 실시간으로 언제나 접근이 가능한 데이터와 분석도구로 인하여 빠른 분석을 가능하다.

- 정부는 인공지능과 빅데이터 활용에 있어 개인정보보호와 사생활 침해 등의 개인정보가 유출되어 범죄에 활용되지 않도록 법률 정비를 이행해나가야 한다.

16차시. 맛만 보는 빅데이터 시각화 기술

- 데이터 활용능력은 수집 및 정제의 데이터 가공 단계부터 분석 기법을 활용한 데이터 분석 및 분석 결과를 시각적으로 표현하는 능력을 말한다.

통계자료의 시각화 발전
- 샤를 미나르(Charles Minard)가 나폴레옹의 러시아 침공을 지도로 만들었는데, 지금도 대표적인 통계 그래프 중 하나로 인용되고 있다.

- 데이터 시각화에 불을 지핀 것은 다름 아닌 컴퓨터의 등장이었으며, 컴퓨터는 많은 양의 데이터를 매우 빠른 속도로 데이터의 시각화를 처리할 수 있게 되었다.
- 오늘날 데이터 시각화는 과학과 예술의 조합으로 빠르게 진화하고 있으며, 향후 몇 년간 기업 환경을 획기적으로 변화시킬 것입니다.

17차시. 필요한 데이터는 나에게 맡겨라_IoT

- 빅데이터 분석을 시각화하면 일반인들도 정보를 빠르게 이해할 수 있다.

- 데이터 시각화의 기능으로는 새로운 동향 파악하여 미래를 예측할 수 있는 기능이 있다.

- 인간은 획득하는 정보의 80% 이상이 시각을 통해 얻어진다고 하므로, 데이터 시각화는 데이터 분석에 대한 전문지식이 없어도 누구나 쉽게 데이터 인사이트를 찾을 수 있다.

- 모든 사물에서 쏟아져 나오는 데이터를 실시간으로 분석하고 관리할 수 있는 기술이 없다면 우리 일상은 어떤 변화도 일어나지 않을 것이며, IoT와 빅데이터는 별개의 기술적 트렌드로 보이지만 결국 하나의 큰 기류로 합쳐질 것으로 예상된다.

- IoT는 Internet of Things의 줄임말로 직역하면 사물들의 인터넷을 말합니다.

- 미국 애틀랜타의 맥키니는 건물 자동화 및 제어시스템 제공회사로, 빌딩 내에 수만 개의 센서를 설치하고, 이를 통해 데이터를 수집 및 분석하여 실시간으로 모니터링한다.

18차시. 데이터 저장의 새로운 패러다임_Data Lake

데이터 저장소인 데이터 웨어하우스
- 데이터 웨어하우스는 미리 결정된 목적을 위해 비즈니스 애플리케이션에서 수집 및 생성하는 데이터의 저장소이다.
- 데이터 저장 전 미리 정의된 스키마를 적용하여 이 저장소에 저장하기 전에 데이터를 정리하고 구성해야 한다.
- 데이터 웨어하우스에 저장된 데이터는 이미 처리되었기 때문에 높은 수준의 분석이 용이하다.

데이터 저장소인 데이터 레이크
- 원시 데이터를 기본 형식으로 저장하는 방대한 저장소입니다.
- 데이터 레이크의 장점중 하나는 다양한 구조의 데이터를 저장할 수 있다는 것입니다.
- 저장된 각 데이터 요소에는 고유한 식별자와 메타데이터가 태그되어 있으므로 필요할 때 더 쉽게 쿼리할 수 있으며, 미리 정의된 스키마가 없습니다.

19차시. 마지막으로 빅데이터 분석가가 되려면?
- 일반적으로 데이터 엔지니어, 데이터 분석가, 데이터 과학자로 빅데이터 분야 대표 직업 3개를 정의하여 빅데이터 분야의 직무를 구분하고 있다.
- 통계 분야에 대한 역량과 코딩 역량 및 머신 러닝과 딥 러닝의 생태계 및 알고리즘에 대한 이해 역량은 데이터 과학자의 직무역량을 말한다.