빅데이터 분석 전문가

01차시

빅데이터의 개념

분석방법론

환경 셋팅

R 프로그래밍 기본

: 기본 문법, 데이터 타입, 함수

조성진 강사

If you torture the data long enough

it will confess to anything.

- RONALD COASE -

강의 소개

- 빅데이터의 이해
- 얻고자 하는 데이터를 수집 가공한다.
- 데이터를 분석하여 정보를 얻는다.
- 데이터 분석 기획, 시각화를 수행한다.

❑약속 드리는 내용

- ❍ (확실한) 이해
- ❍데이터 기획,
- ❍원하는 데이터 수집하기
- ❍데이터 내 맘대로 가공하여 정보화하기

if 당연히 예습, 복습 필수일 경우

❑ 약속 드릴 수 없는 것

- ❍ 스킬 전달 형 강의

  ➔ 아는 것이 힘이 아닙니다. 아는 것을 사용하는 것이 힘 입니다.

- ❍ 생각 없이 따라만 치는 강의

- ❍ 과도한 이론수업 (구글에 좋은 자료 많습니다.)

  ➔ 필요할 땐 당연히 합니다. 오늘처럼요~ 직접 타이핑하세요!!

- ❍ 한번의 웃음이 없는 재미없는 강의

  ➔ 공부도 재미가 있어야 합니다.

  스스로 친해질 기회를 최대한 많이 부여하세요.

Ice Breaking

- 오늘의 옆 동료를 어디서 만나게 될 지 모릅니다.
- 적어도 2달 간 같이 성장해야 할 소중한 동료이자 라이벌입니다.
- 옆 자리분과는 매일 인사하세요!!!

Ice Breaking

- **자기 소개하기**
- **제한 시간: 1분 이내**
- **예시)**
  - 이름, 프로그래밍 수준
  - 수업에 참여하게 된 목표(개발, 기획, 통계)
  - 수업에서 얻고자 하는 것

본격적인 수업 시작 전에...

❑자가진단

- ❍데이터와 정보의 차이를 알고 있다.
- ❍피벗 테이블을 알고 있다.
- ❍정형데이터와 비정형데이터를 알고 있다.
- ❍그래프를 알고 있다.
- ❍GB(Giga) -> TB(Tera) -> PB(Peta) -> EB(Exa) -> ZB(Zeta) -> YB(Yotta)
- ❍크롤링 혹은 스크래핑을 알고 있다.

❑기존 데이터베이스 관리 도구의 능력을 넘어서는 대량의 정형 또는 심지어 데이터베이스 형태가 아닌 비정형의 데이터의 복합 조차 포함한 데이터로부터 가치를 추출하고 결과를 분석하는 기술

❑데이터로부터 의사 결정권자에게 도움을 줄 수 있는 정보를 추출하는 기법

❑ R, 하둡, 스파크, 카프카, 파이썬, 머신러닝, 딥러닝…

빅데이터 필수 용어

DIKW

- ❑지혜: 지식을 활용하는 것
- ❑지식: 정보의 이용에 대한 노하우
- ❑정보: 가공된 데이터
- ❑데이터: raw 데이터, 관찰된 객관적 사실

빅데이터 필수 용어

❑빅데이터

- ❍큰 데이터! 감당할 수 없을 정도의 거대한 데이터의 집합
- ❍일반적인 DB가 저장, 관리, 분석할 수 있는 범위를 초과하는 규모의 데이터 (McKinsey, 2011)
- ❍다양한 종류의 대규모로부터 저렴한 비용으로 가치를 추출하고 초고속 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아키텍처

➔ 새로운 개념이 아닌 이미 있던 데이터를 이제는 처리할 수 있는 기술이 나온 것

❑조쌤's 빅데이터 정리

❍과거에는 저장할 수 없을 정도로 큰 데이터(Big Data)를 저장하고

❍다양한 형태의 데이터를 빠르게 가공 / 분석하여

❍데이터를 사용 할 수 있게 되었다.

빅데이터 필수 용어

❑3V

❍과거에는 저장할 수 없을 정도로 큰 데이터(Big Data)를 저장하고

❍다양한 형태의 데이터를 빠르게 가공 / 분석하여

❍데이터를 사용 할 수 있게 되었다.

❑3V

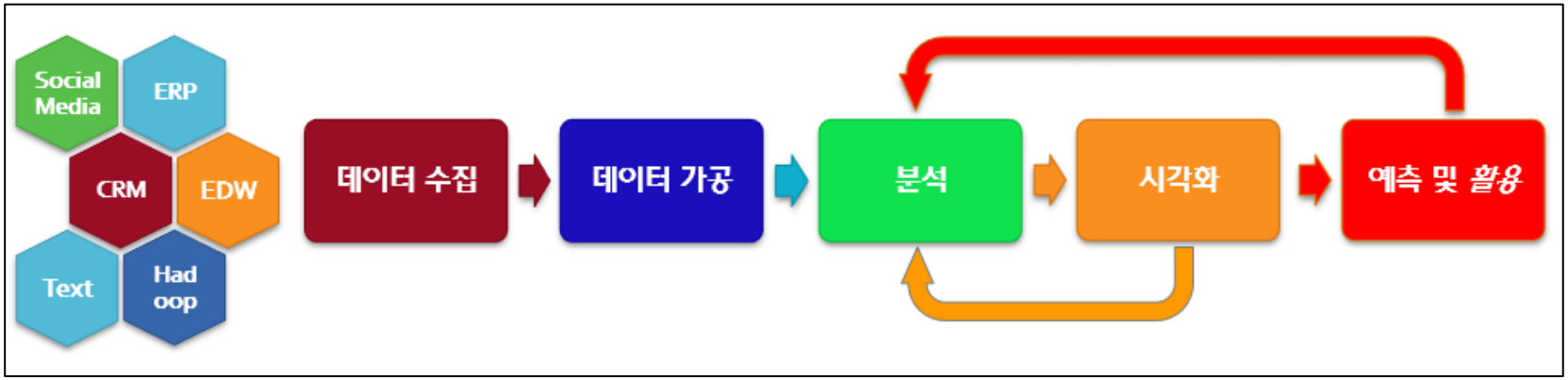
❍규모(Volume): 스마트폰, 유튜브, 교통 신호, 센서, 자율 주행차 등에서 생산 되는 엄청난 규모의 데이터

❍형태(Variety): 엑셀, 워드, 통계치, 영상, 사진, 음성인식, 인스타 등에서 발생하는 다양한 형태

❍속도(Velocity): 데이터 저장, 분석, 로드의 속도가 매우 빨라짐

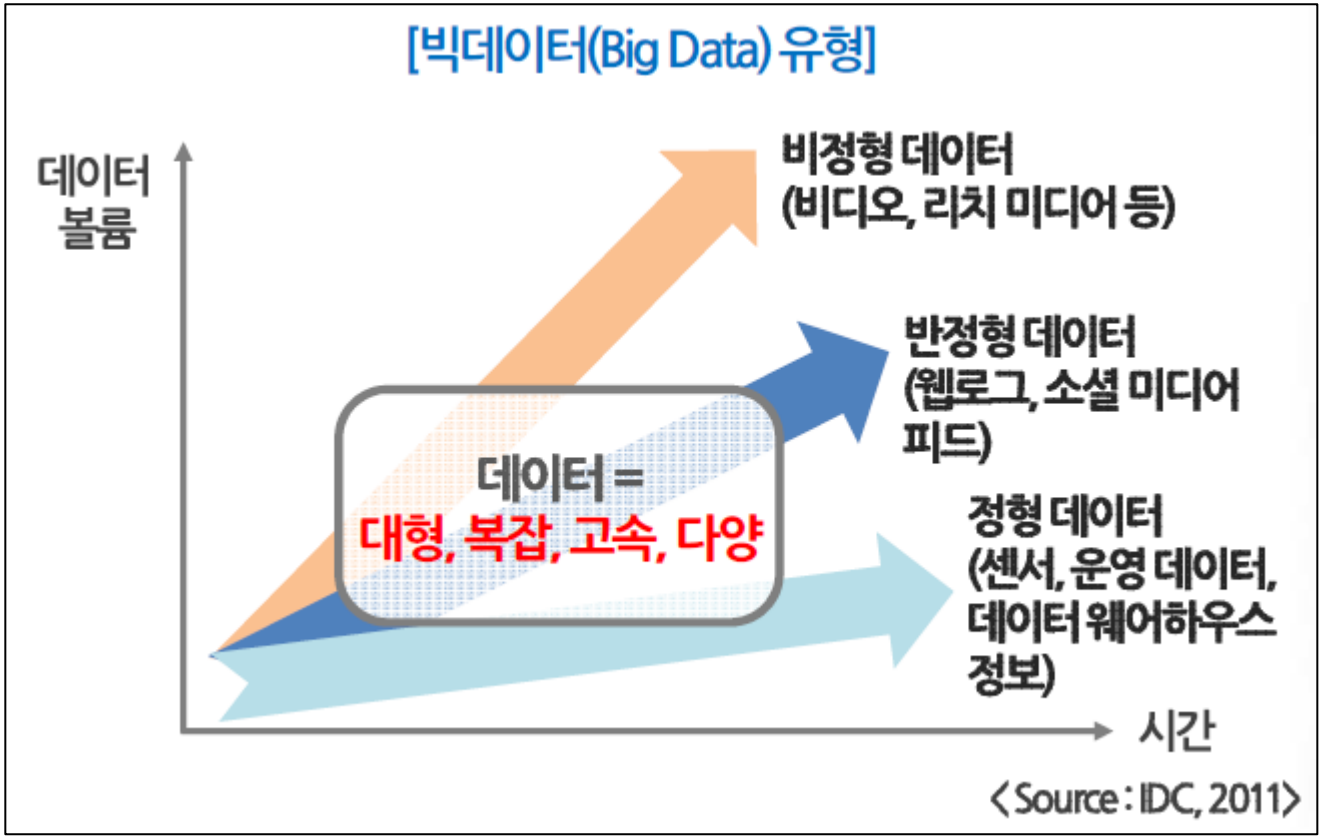
- HW 발전과 SW 기술의 발전

❑빅데이터 분석 프로세스



데이터로부터 가치를 추출하고 결과를 분석하는 기술

❑정형, 반정형, 비정형 데이터



❑정형, 반정형, 비정형 데이터

❑통계분석(Statistical learning)

❍데이터를 요약하고 기술하는 기술 통계와 샘플 데이터에서 전체 집단의 의미를 추정하는 통계적 추론이 주를 이룬다.

❑머신러닝(Machine learning)

❍인공지능(AI: Artificial intelligence)의 한 분야로 기계가 자동으로 데이터에서 중요한 패턴과 규칙을 학습하고 의사결정, 예측 등을 수행하는 기술

– 지도학습, 비지도학습

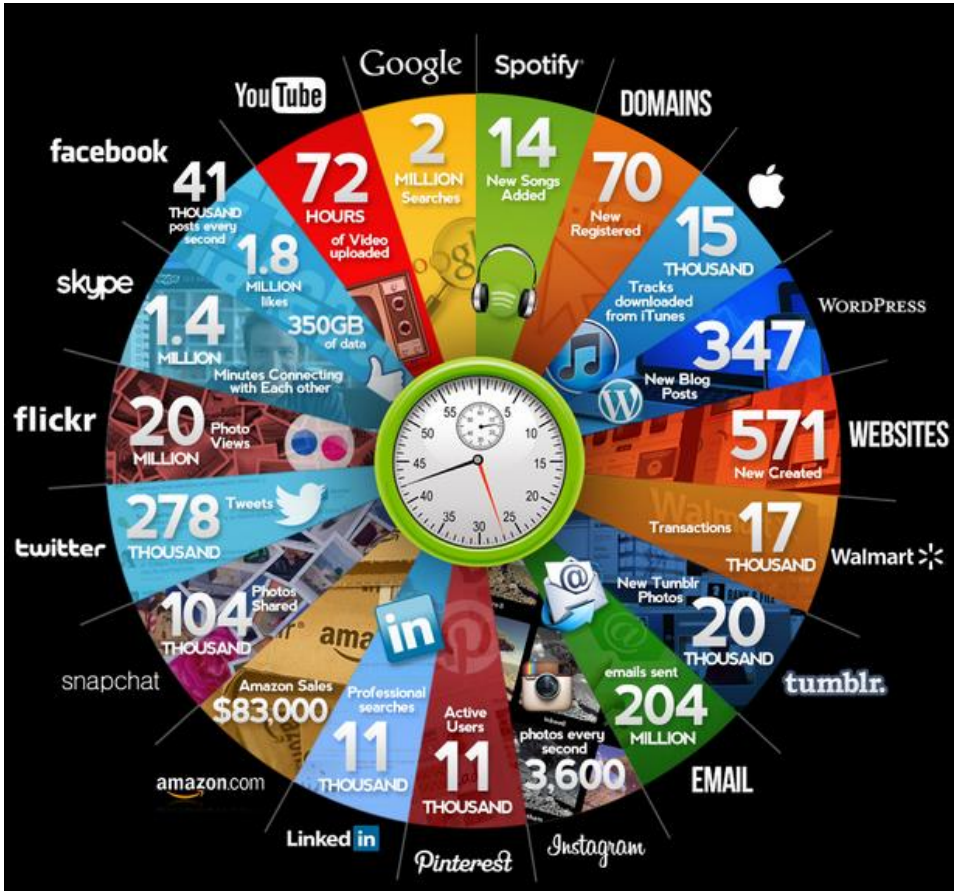
❑딥러닝(Deep learning)

- ❍머신러닝의 한 분야로 다중 신경망에서 잘 작동하는 학습 방법론
- ❍CNN(Convolutional Neural Network)
  - 합성곱 신경망, 이미지 인식 문제에 뛰어난 성과를 보이는 알고리즘
- ❍RNN(Recurrent Neural Network)
  - 순환 신경망, 시간 순서의 특징을 반영할 수 있는 모델로 이전 시점의 학습 정볼르 다음 시점의 학습에 반영할 수 있는 구조가 특징

❑추천시스템(Recommendation system)

- ❍정보 필터링 기술의 일종으로 특정 사용자가 관심을 가질만한 정보 (영화, 음악, 뉴스, 이미지, 웹 페이지)등을 추천하는 것
- ❍큐레이션이라고도 불림
- ❍전자상거래의 핵심으로 아마존, 넷플릭스 등이 가장 좋은 사례

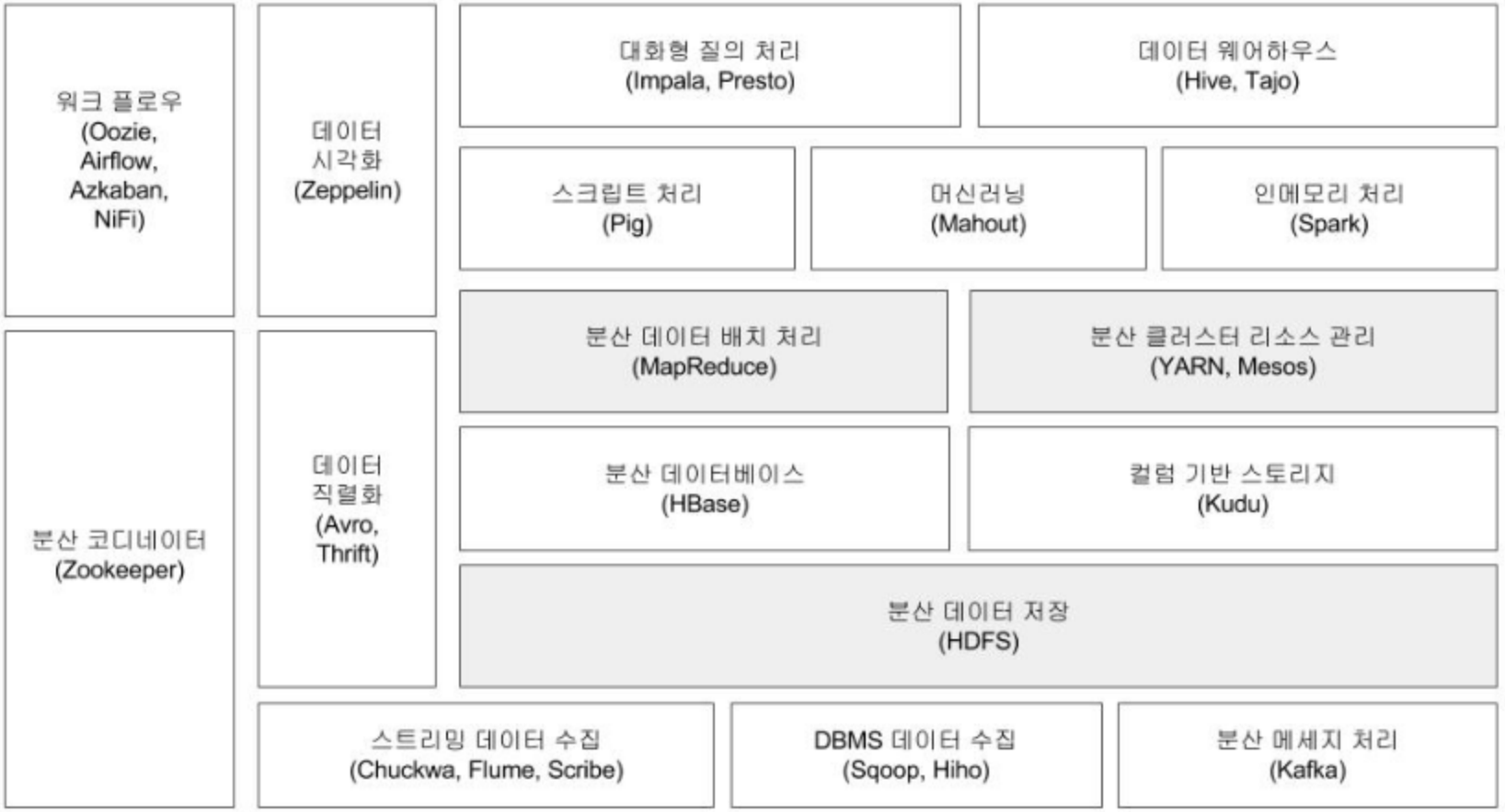
빅데이터 분석 방법론



인터넷에서 60초 동안에 일어나는 일

어떻게 데이터를
빠르게 쌓을 수 있을까요?

빅데이터 에코 시스템



빅데이터 분석 방법론

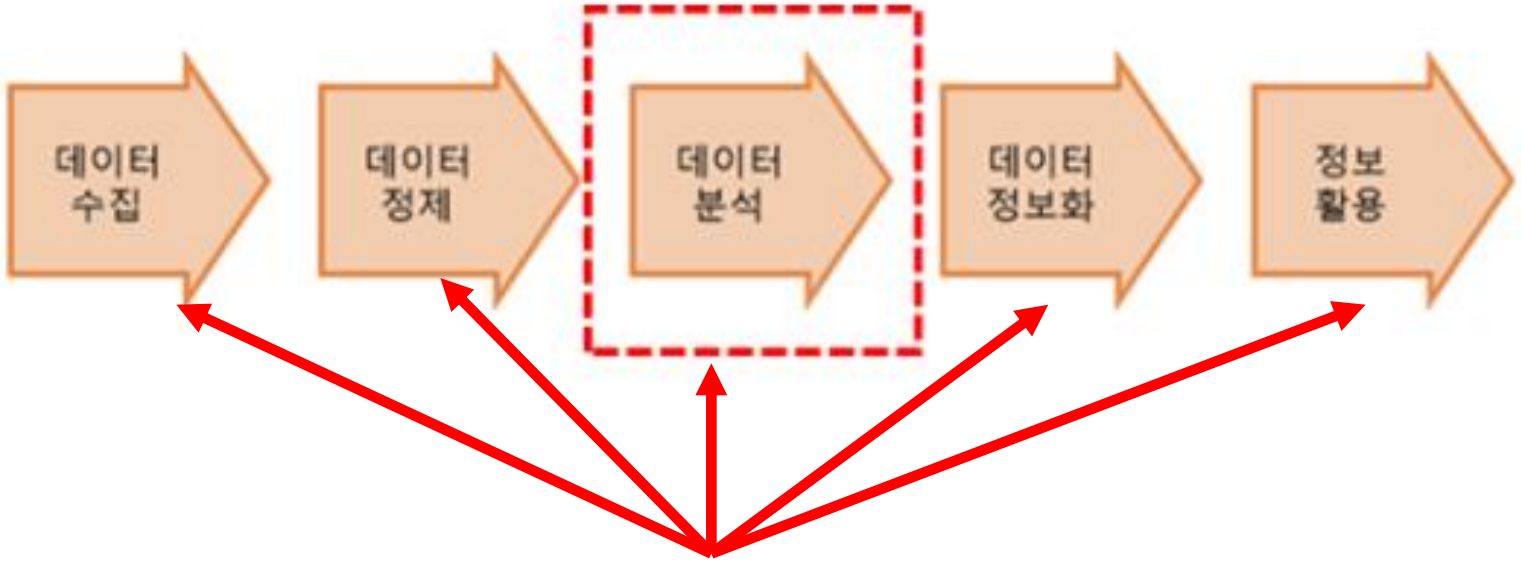
데이터 사이언스
(DATA SCIENCE)

그래서 우리 강의는 뭘 하는거다?

빅데이터 분석 방법론

R을 활용한 데이터 분석

빅데이터 분석 방법론



R을 활용한 데이터 분석

❑R 프로그램 설치

- ❍공식홈(https://www.r-project.org/)
- ❍통계 및 그래프 작업을 위한 인터프리터 프로그래밍 언어

❑R Studio 설치

- ❍공식홈(https://www.rstudio.com/)
- ❍R IDE(통합 개발 환경)

❑Java 설치

- ❍공식홈(https://java.com/ko/download/)
- ❍R 패키지 사용을 위한 JRE 설치

❑R 프로그래밍

- ❍통계 계산과 그래픽을 위한 프로그래밍 언어이자 SW 환경
- ❍1993년 오클랜드 대학교에서 개발되었고, 현재는 R 코어팀이 개발 중
- ❍윈도우, 맥OS, 리눅스와 유닉스에서 사용이 가능
- ❍그래프 기능이 강점
- ❍수 많은 패키지를 가지고 있고, CRAN이라는 생태계가 존재
  - npm, maven, 안드로이드 마켓?과 비슷

❑R 프로그래밍 장점

❍무료 - OSS

- ❍SPSS, MATLAB은 상당히 고가의 가격
- ❍오픈소스지만 상용에 전혀 뒤지지 않는 다양한 기능
- ❍그래픽 관련 패키지가 뛰어남
- ❍수 많은 패키지
  - 이미 모든 통계 기법이 패키지로 구현되어 있음

❑R 프로그래밍 단점

❍정보보호 없음

❍메모리 관련하여 매우 큰 데이터 집합을 이용할 때 문제가 발생할 수 있다.

- 컴퓨터 기술의 발달로 어느 정도 해결됨

❍프로그램 자체에서 한글 지원 안 함

왜?

요즘 배우는 언어인데

정보보호도 안 되고, 메모리,

한글 지원도 안 돼요?

R은 생각보다 오래 된 언어

R 1993년 출생

Java 1995년 출생

하지만 이런 단점보다는
장점이 더 매력적

모든 프로그래밍의 기본?

Hello! World

❑R 기본 사용법 배우기

- ❍패키지 다루기

❑기본 문법

- ❍기본 연산
- ❍변수
- ❍데이터 타입
- ❍함수