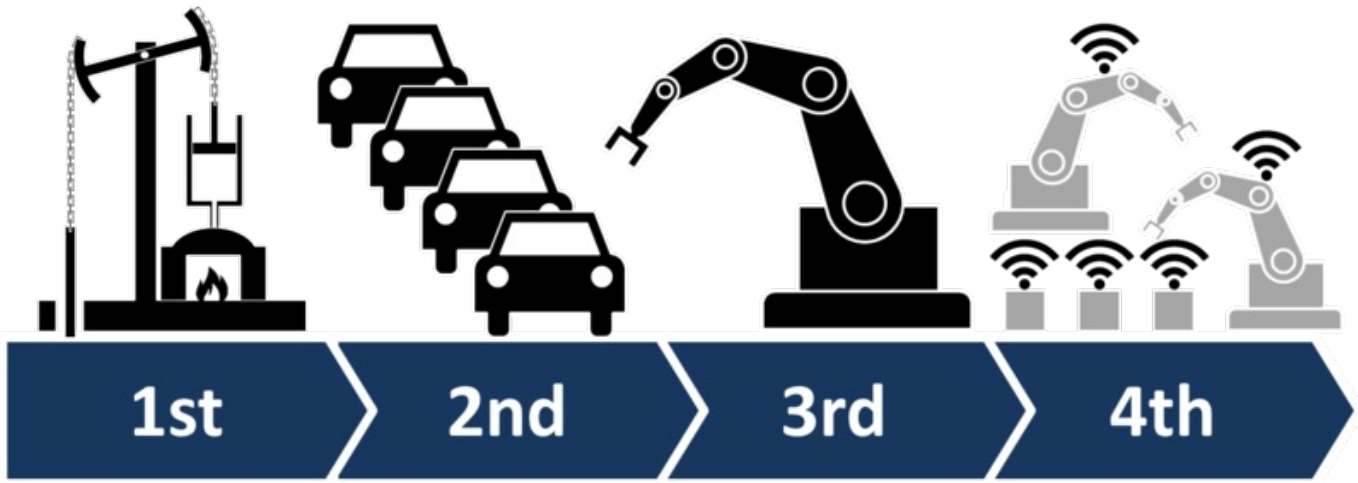
데이터 사이언스와 데이터 분석

데이터 사이언스

목차 데이터 사이언스와 데이터 분석

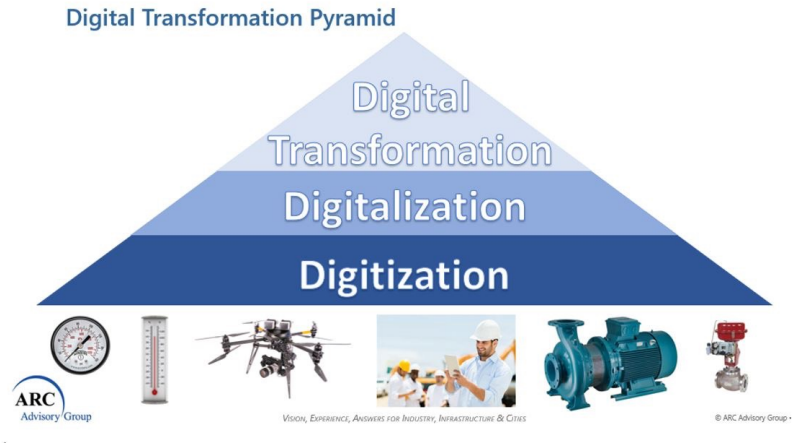
1. 디지털 변환(DX)과 데이터 사이언스

- DX(Digital Transformation)



https://www.forbes.com/

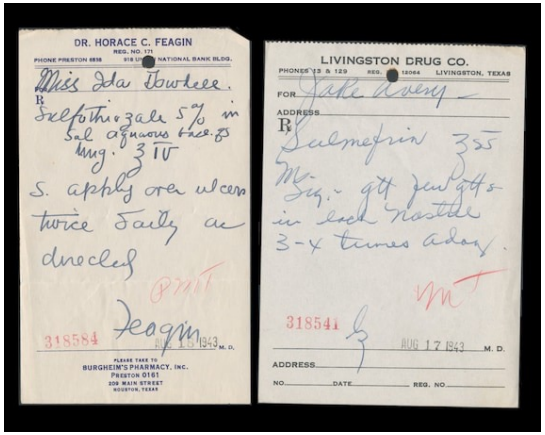
- DX란?
  - 디지털 변환(DX): 디지털 기술을 활용하여 비즈니스 모델, 프로세스, 조직 구조 등을 혁신하는 과정
  - 목적
    - 높은 효율성, 새로운 가치 창출, 고객 경험 개선
  - 기대 효과
    - 단순한 기술 도입을 넘어 전체적인 비즈니스 변화를 의미
    - 경쟁력을 강화하고 지속 가능한 성장을 추구
    - 산업 전반에 걸쳐 중요한 트렌드로 자리잡고 있음
  - 단, 이 모든 것들은 하루 아침에 이루어지는 것이 아님



https://www.arcweb.com/blog/what-digitization-digitalization-digital-transformation


Digitization
Digitalization
Digital Transformation

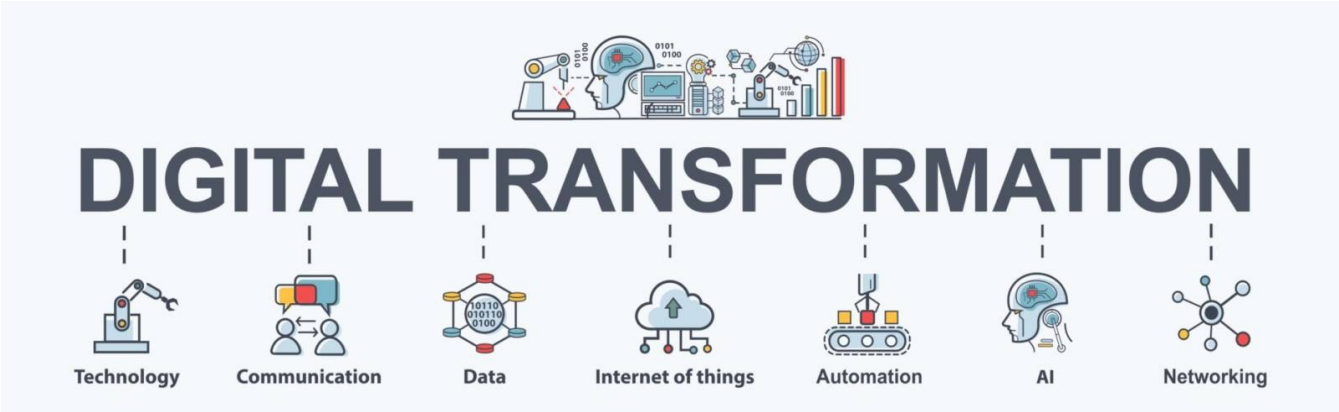
- Digitization
  - 아날로그 정보를 디지털 형식으로 변환하는 과정
    - 예) 종이 문서를 스캔하여 PDF 파일로 변환
  - 디지털 변환의 첫 단계
  - 목적: 정보 접근성을 높이고, 저장 및 검색을 용이하게 하는 행위
  - 장점: 물리적 공간 절약, 정보 손실 방지
  - 모든 조직이 처음 디지털화 단계를 거치며, 이를 통해 데이터를 분석 가능하게 만듦



https://www.slideshare.net/slideshow/electronic-health-records-34408135/34408135

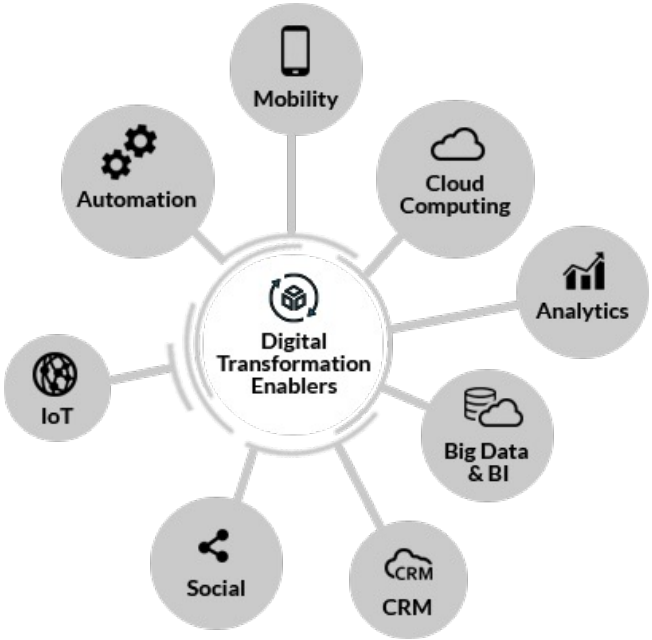
- Digitalization
  - 기존 비즈니스 프로세스를 디지털 기술을 활용하여 개선하는 과정
  - Digitization이 데이터를 디지털로 변환하는 것이라면, Digitalization은 그 데이터를 활용해 업무 효율을 높이는 것
    - 예) 종이 기반 회계 시스템을 디지털 회계 소프트웨어로 전환하는 것
    - BI tools, Office programs
  - 업무 속도를 높이고, 오류를 줄이며, 데이터 기반 의사결정을 가능하게 함
  - 디지털 변환의 중간 단계로, 디지털 도구의 사용을 극대화
  - On-going process

- Digital Transformation
  - 조직 전반의 근본적인 변화를 의미
    - 우리 사회가 나아갈 방향
  - 단순한 디지털 도구 사용을 넘어, 비즈니스 모델, 문화, 고객 경험 등 모든 측면에서 혁신을 추구
    - 예) 온라인 쇼핑몰이 고객 데이터를 분석하여 맞춤형 마케팅 전략을 개발
  - Digital Transformation은 데이터와 기술을 중심으로 이루어지며, 지속적인 변화와 적응이 필요
  - 조직은 이를 통해 새로운 가치를 창출하고, 경쟁력을 확보할 수 있음



https://smartway2.com/

- DX와 데이터 사이언스
  - 데이터 사이언스는 DX의 핵심 요소로, 데이터 분석을 통해 의사결정을 지원
    - 데이터 사이언스는 대규모 데이터에서 패턴과 인사이트를 도출하여
    - 비즈니스 혁신에 기여
  - 두 분야는 상호 보완적으로 작용하여, 조직의 디지털화와 경쟁력 강화를 도움
    - 데이터 사이언스는 DX의 성공을 위한 필수 요소



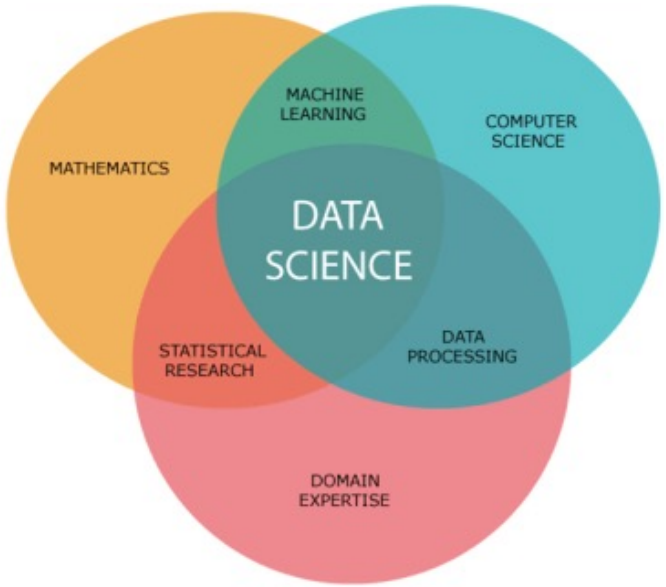
https://medium.com/@techhunt2195/big-data-facilitate-artificial-intelligence-and-drive-digital-transformation-4aef03a4a058

2. 데이터 사이언스

- 데이터 사이언스란?
  - 데이터를 통해 유의미한 인사이트를 도출하고, 문제를 해결하는 학문
  - 데이터 사이언티스트는 데이터를 분석하여 비즈니스 문제를 해결하고, 데이터 기반 의사결정을 지원

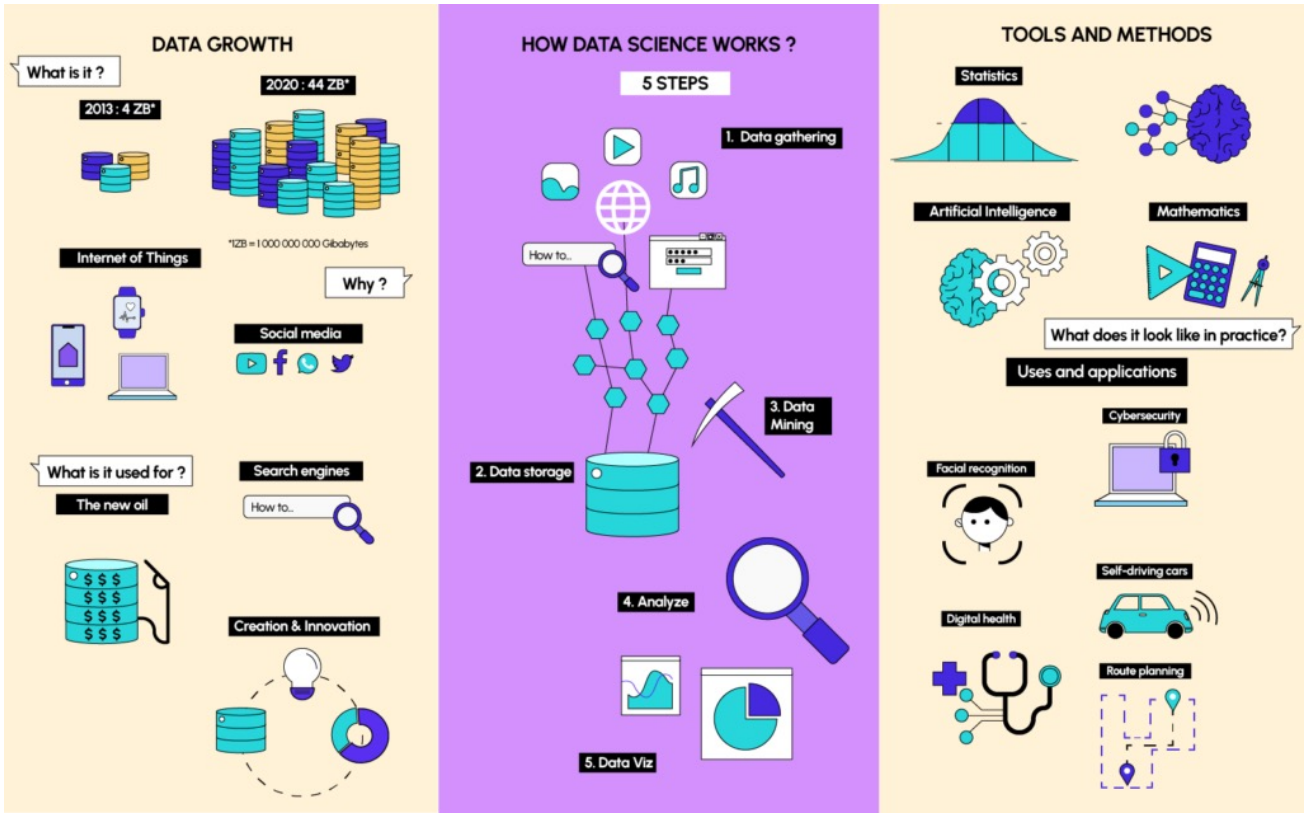
https://artoftesting.com/what-is-data-mining

- 데이터 사이언스란?
  - 데이터 수집, 저장, 처리, 분석, 시각화 및 해석을 포함한 다양한 기술과 방법론 사용
  - 통계학, 컴퓨터 과학, 수학 등 여러 분야의 지식을 통합하여 데이터에서 가치 추출



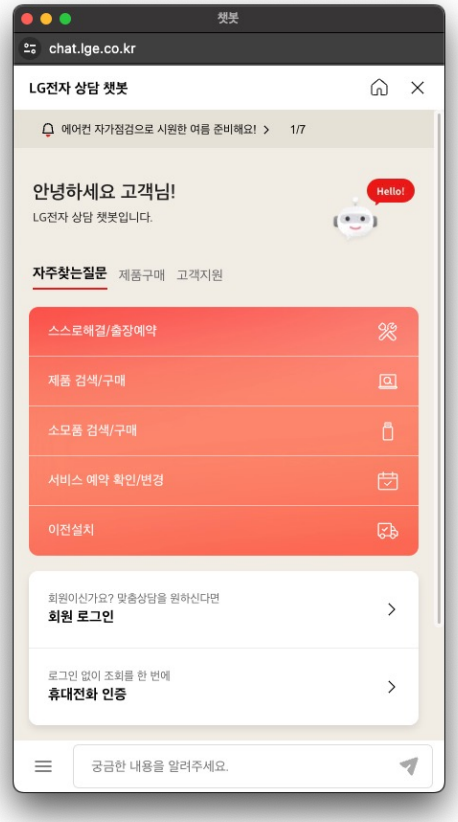
https://medium.com/analytics-vidhya/introduction-to-data-science-28deb32878e7

- 데이터 사이언스



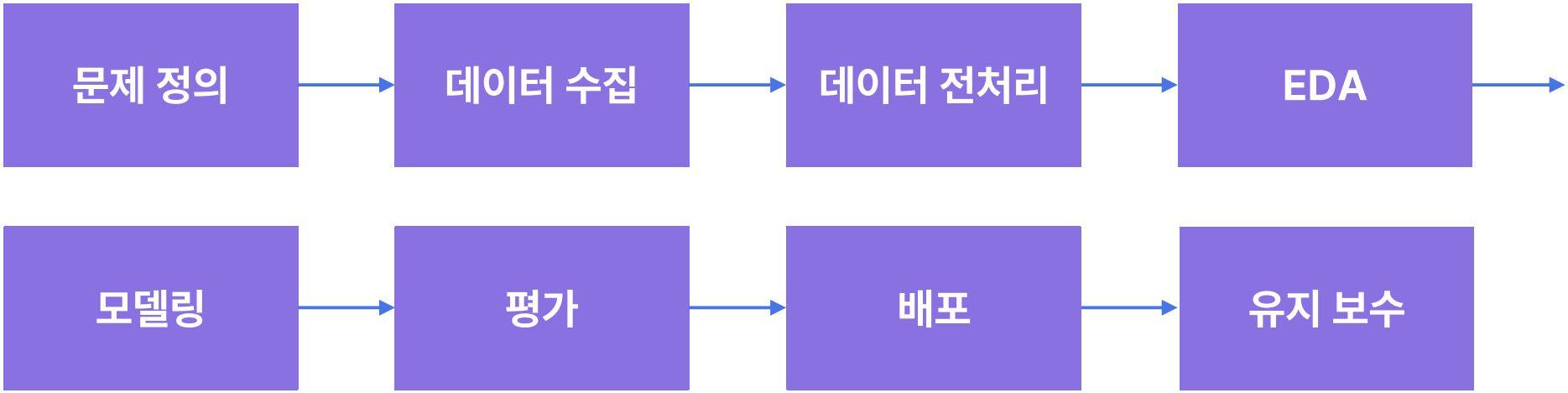
https://datascientest.com/en/data-science-definition

- 데이터 사이언스의 필요성
  - 데이터의 폭발적인 증가와 복잡성 증대 → 데이터 사이언스의 중요성 증대
  - 데이터 기반 의사결정
    - 정확도와 신뢰성을 높이며, 리스크를 최소화
  - 새로운 비즈니스 기회와 혁신적인 솔루션을 제공
  - 고객 맞춤형 서비스 제공 가능
  - 기업의 경쟁력과 효율성



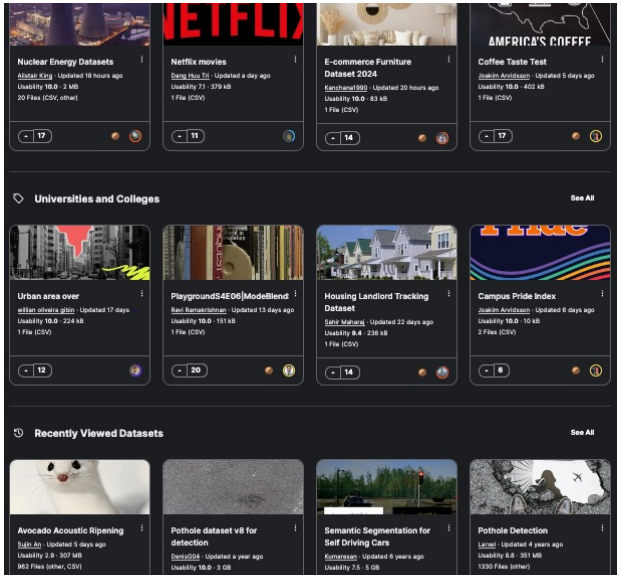
- 데이터 사이언스를 통해 얻을 수 있는 주요 기대 효과:
  - 효율성 향상: 프로세스 최적화 및 비용 절감
  - 의사결정 지원: 데이터 기반의 정확한 의사결정 가능
  - 고객 이해: 고객 행동 분석을 통해 맞춤형 서비스 제공
  - 새로운 비즈니스 모델 창출: 데이터에서 도출된 인사이트를 통해 새로운 사업 기회 발굴
  - 리스크 관리: 데이터 분석을 통해 리스크를 사전에 식별하고 관리

3. 데이터 사이언스의 워크플로우와 도구

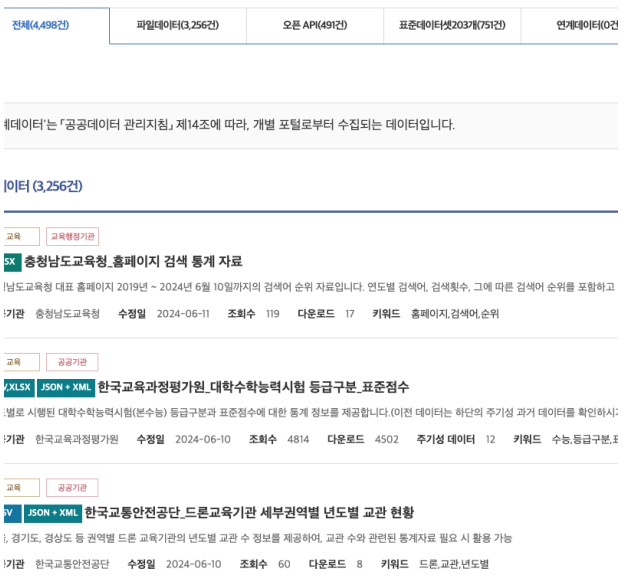

문제 정의
데이터 수집
데이터 전처리
EDA
모델링
평가
배포
유지 보수

- 문제 정의
  - 해결하고자 하는 비즈니스 문제를 명확히 정의
  - 현상의 문제점, 문제의 원인을 파악
  - 해결 방안에 대한 가설수립
  - 전체 분석 목적을 진단
  - 현업 전문가의 의견이 가장 중요한 단계

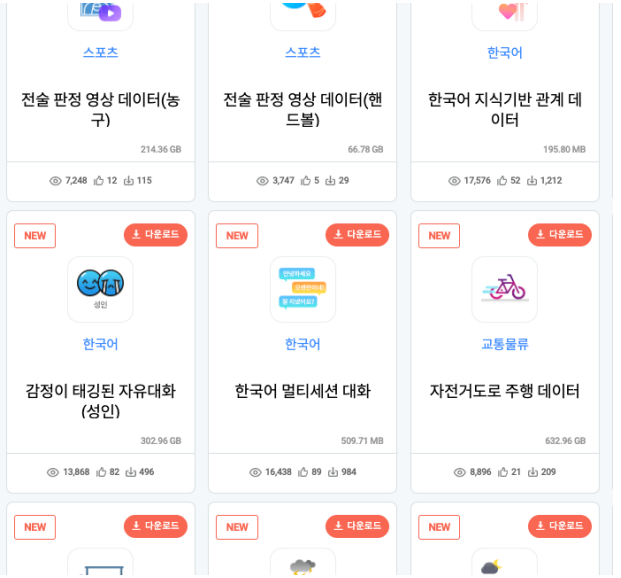
- 데이터 수집
  - 문제를 해결하기 위해 필요한 데이터를 정의
  - 다양한 출처에서 데이터를 수집
    - 미보유/기보유 데이터의 명확한 구분
    - 데이터 수집에 가용할 수 있는 시간/인력/예산 계획
  - 문제정의 과정에서 논의된 기간/범위의 데이터 수집
  - 기보유 데이터의 경우 선별 과정이 필요
    - 데이터 유형(이미지, 비디오, 텍스트 등)
    - 라벨링 여부
    - 가용할 수 있는 데이터/스토리지/리소스 파악



Kaggle

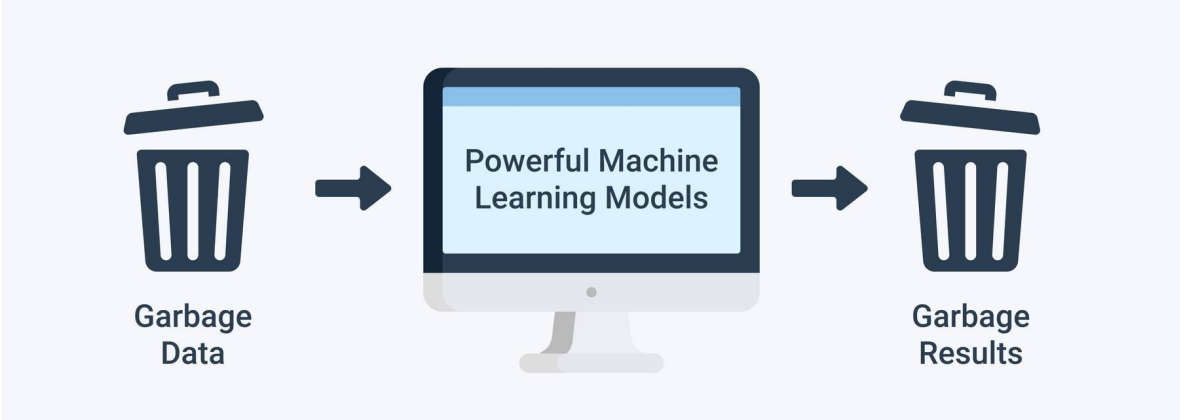


공공데이터 포털



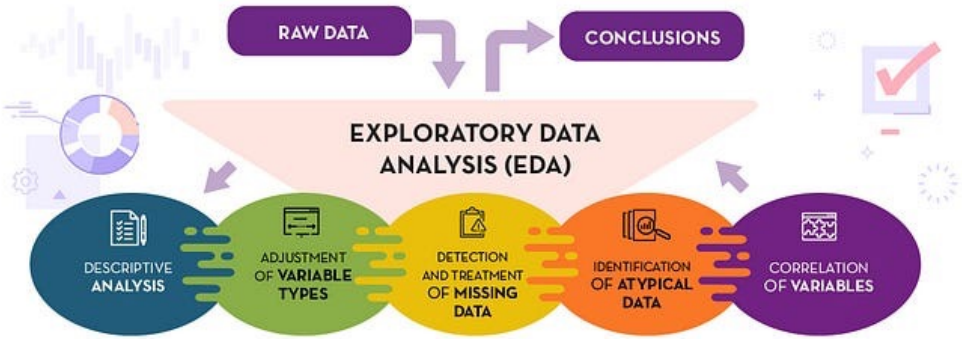
AI Hub

- 데이터 전처리
  - 데이터 분석이나 모델링을 수행하기 전에 데이터를 정리하고 구조화하는 과정
  - GIGO(Garbage In, Garbage Out)
  - 일반적으로 아래 과정들이 포함
    - 결측치 처리
    - 이상치 탐지
    - 정규화 등



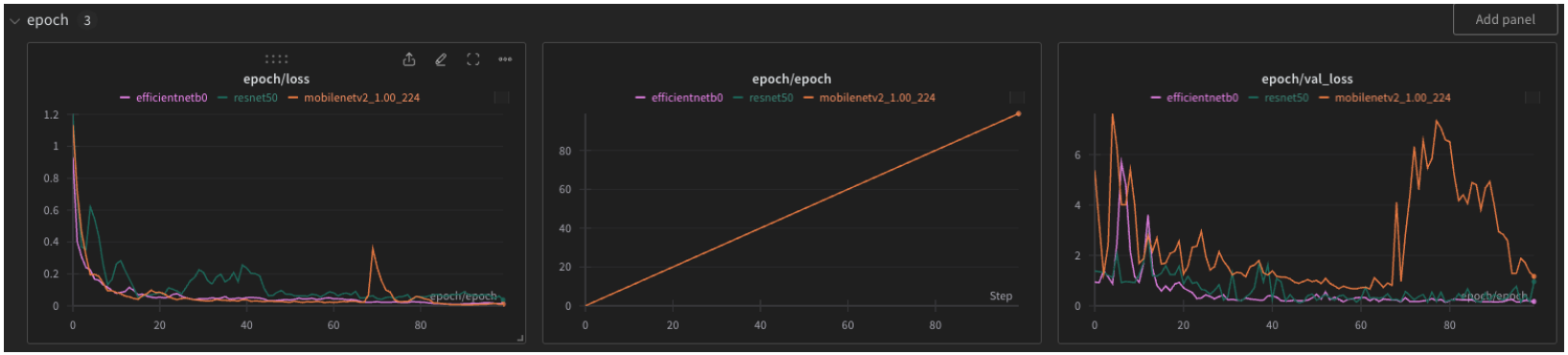
https://monkeylearn.com/blog/data-preprocessing/

- EDA(Explorative Data Analysis)
  - 데이터의 특성을 파악 하고, 시각화를 통해 인사이트 도출
  - 통계적 기법, 머신러닝 모델, 경험적 근거를 활용하여 데이터가 갖는 의미를 분석
  - 시각화와 수치화를 통한 커뮤니케이션
  - 데이터 사이언티스트의 역량이 가장 크게 부각되는 구간

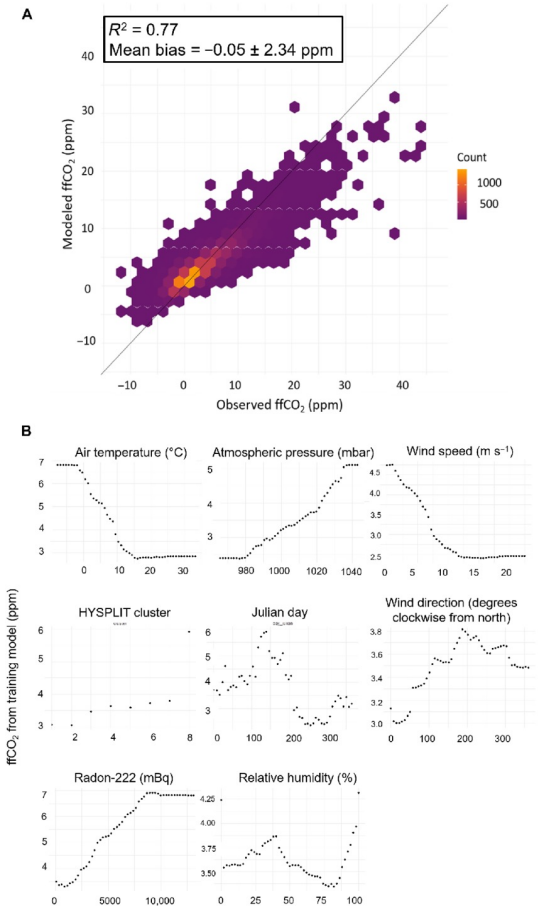


https://medium.com/@zuhailinasir_26370/data-science-process-ad8d1978377e

- 모델링
  - 문제 해결에 적합한 통계/머신러닝 모델을 선정하는 과정
  - 모델의 선정 기준
    - 풀려는 문제에 따라: 분류, 회귀, 차원 축소, 생성
    - 다루는 데이터에 따라: 테이블 데이터, 이미지, 텍스트, 비디오, 오디오
    - 학습할 데이터의 규모에 따라
    - 모델 가용 여부에 따라
    - 사전학습 여부에 따라

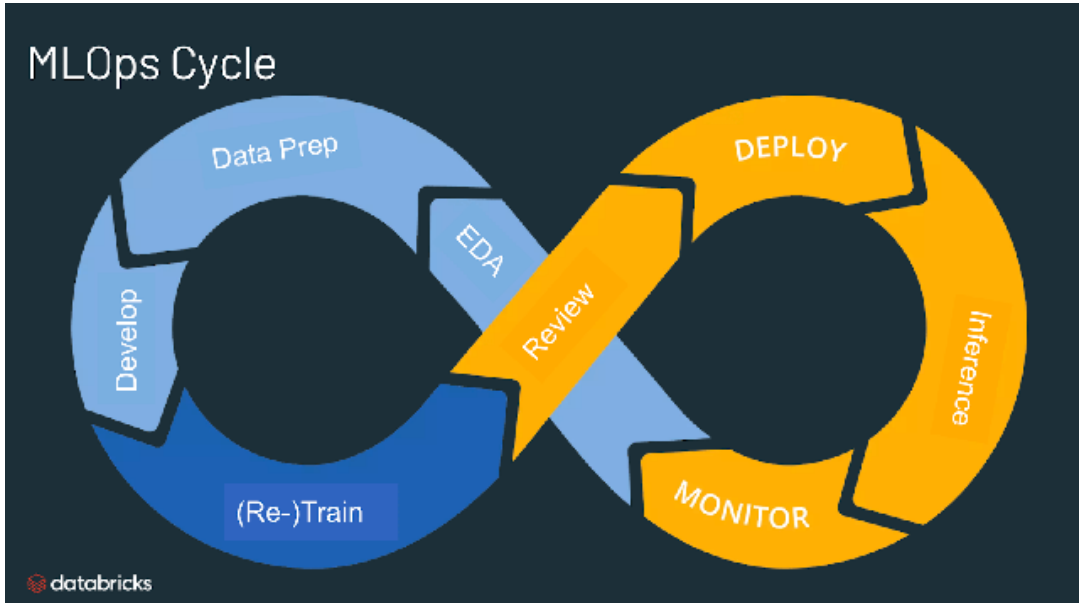


- 평가
  - 모델의 성능을 객관적으로 평가하고, 문제 해결 능력을 정량화하는 과정
    - 모델을 튜닝하여 성능을 높이거나
    - 데이터를 추가하거나
    - 전처리 과정을 개선
  - 설정한 가설이 타당한지 여부를 판단하는 과정
  - Academia에서 주의깊게 검토하는 과정
    - 평가지표가 현실 세계를 잘 반영하는지
    - 왜 해당 평가 기준을 선정했는지
    - 평가 지표를 어떻게 해석해야 하는지
    - 선택한 기준에 사각지대는 없는지



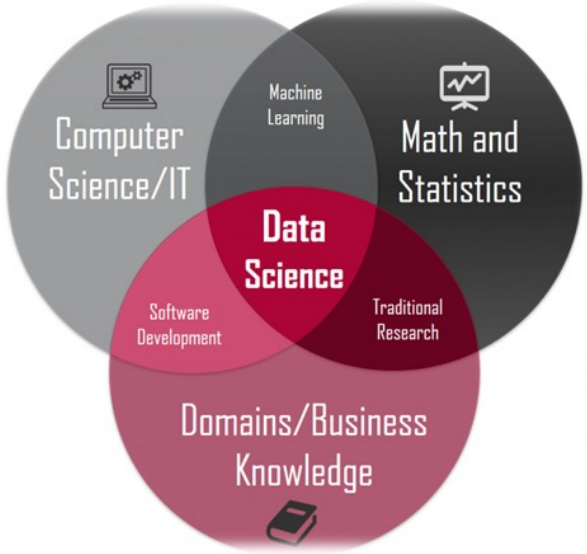
https://www.researchgate.net/figure/Evaluation-of-random-forest-machine-learning-model-A-Scatter-plot-of-hourly-observed_fig4_360129122

- 배포 및 유지보수
  - 모델을 실제 환경에 배포하여 활용
  - 배포 된 이후에도 모델 성능 모니터링 및 업데이트 필요

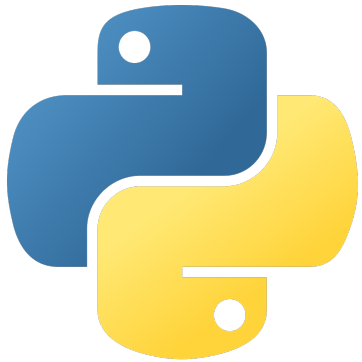


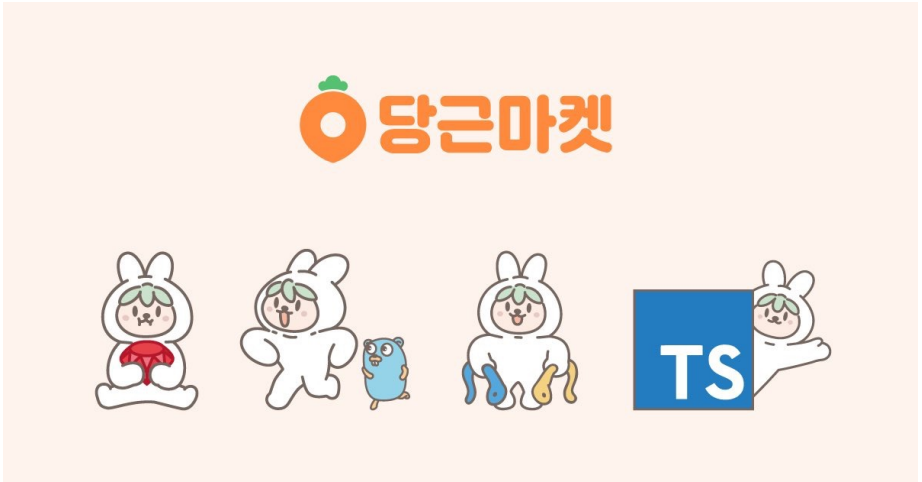
https://www.databricks.com/kr/glossary/mlops

- 데이터 사이언스에 필요한 지식
  - 프로그래밍 능력
    - 데이터를 처리하고 분석하기 위한 프로그래밍 기술 필요
    - Python, R 등
  - 도메인 지식
    - 특정 분야의 전문 지식이 데이터 분석의 정확성과 유용성을 높임
    - 데이터 사이언스는 도구 학문
  - 수학
    - 확률: 데이터 분포, 정보 이론, 모델 이해를 위한 확률론 개념 요구
    - 통계학: 데이터 분석 및 해석을 위한 기초적인 통계 개념 필요
    - 미적분, 선형 대수 등



- 프로그래밍 언어: 파이썬
  - 데이터 사이언스에서 가장 널리 사용되는 프로그래밍 언어
  - 특징
    - 쉬운 문법과 높은 가독성
    - 객체 지향 프로그래밍, 함수형 프로그래밍 등 다양한 프로그래밍 패러다임 지원
    - 커뮤니티, 강력한 라이브러리 지원
    - 데이터 처리, 분석, 시각화, 머신러닝 등 다양한 작업에 적합





https://www.snippod.com/l32f

- 파이썬의 데이터 사이언스 생태계
  - 모듈 확장성: 필요에 따라 다양한 모듈 및 패키지를 쉽게 추가 가능
  - 사용자 정의 모듈 성성 가능
  - 모듈 버전 업데이트 및 기여도 활성화

https://www.linkedin.com/pulse/python-libraries-data-analysis-priya-kumar-ac4xc