의약품 제조와 AI

김 화 종

강의 목표

SNU 제약바이오 인력양성센터
경기시흥글로벌바이오캠퍼스

- 데이터 기반 의약품 품질관리의 중요성
- 공정, 설비, 원료, 제품의 품질관리 이해
- AI를 이용한 제조 품질관리
- 파이썬, 데이터 분석, 머신러닝 이해 및 적용

목차 1

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

목차 2

8. 머신러닝 모델
   - 주요 머신러닝 알고리즘과 특징 이해, • 손실 함수, 옵티마이저, 성능 지표 이해
   - 하이퍼파라미터 튜닝과 모델 최적화, • (실습) 다양한 머신러닝 모델 성능 비교
9. 딥러닝 모델
   - MLP, CNN, RNN, Transformer, Graph, • 이미지 분석, 전이학습, • (실습) 신경망 구현 및 응용
10. 시계열 예측
   - 생산량 예측, 리드타임 예측, • (실습) 시계열 예측 모델
11. 예지 보수(Predictive Maintenance)
   - 설비 및 부품 장애 예측, • (실습) 예측 모델
12. 패턴 인식과 이상 탐지
   - 시계열 패턴인식, 공정 이상 탐지, • (실습) 이상 탐지 모델
13. 객체 검출
   - 제조 환경의 객체 검출 (object detection), • (실습) 결함 예측 (defect detection)
14. 연속 공정과 AI
   - 연속 공정 데이터 분석과 최적 제어, • 데이터 무결성, QbD

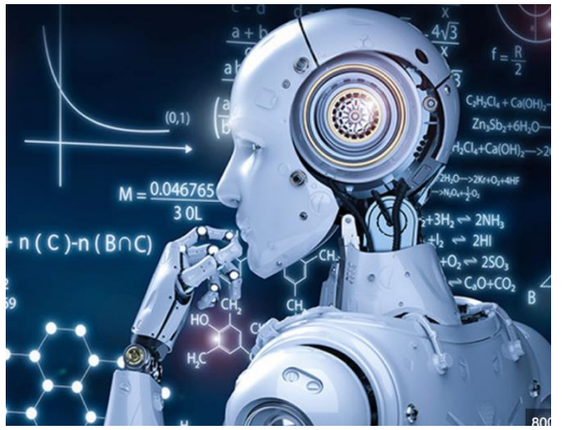
의약품 품질관리와 AI

AI Burden

SNU 제약바이오인력양성센터 경기시흥글로벌바이오캠퍼스

Bio Burden → Document Burden → Digital Burden → AI Burden



AI 도입 목적

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- 시간과 비용 절감 및 품질 개선
  - 제품기획, 시험, 설비구축, 수요예측, 생산, 운영, 마케팅, SCM, 피드백 분석 등 전 과정에 AI 적용중
  - 개발 시간과 비용 (원료, 설비, 운영, 에너지, 인력 등) 감소
  - 제품의 품질 향상
- 거의 모든 산업에 적용
  - **신약개발 → 임상시험 → 의약품생산 → 시판후관리** 에 적용

Terminology

- GMP, QbD (quality by design), continuouse manufacturing
- digital transformation (DT), industry 4.0, biopharma 4.0
- internet of thing (IOT), big data
- cloud, edge computing
- AI model, **machine Learning**, deep learning, LLM
- data integrity, data quality
- data security, data governance
- predictive maintenance
- smart factory, digital twin
- process design, optimization, control

FDA Discussion Paper

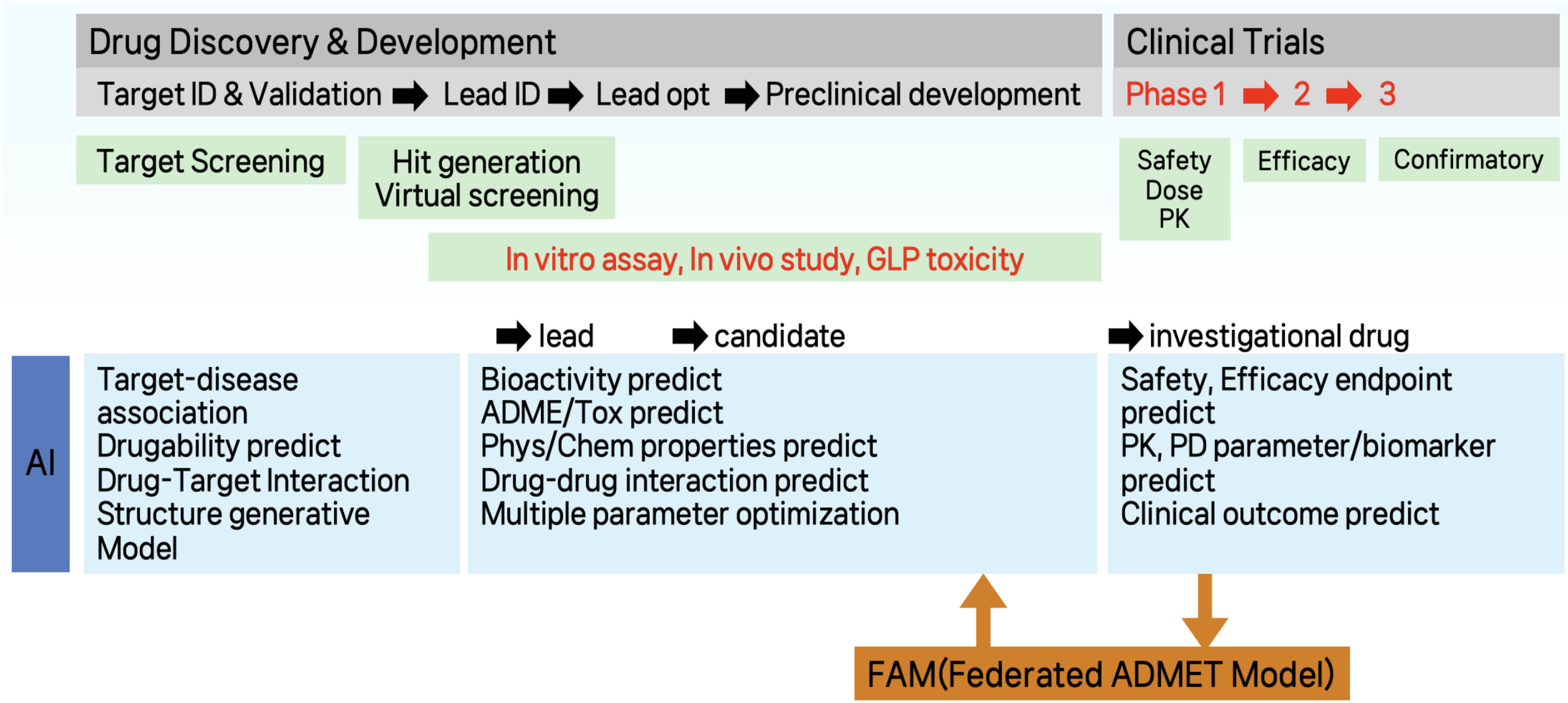
- https://www.fda.gov/media/165743/download

주요 내용

- 필요성
  - To produce quality drugs without excessive regulatory oversight
  - A need for more flexibility in manufacturing
  - AI for measurement, modeling, and control in pharmaceutical manufacturing
- 대상
  - New Drug Application (NDA), Abbreviated New Drug Application (ANDA), or Biologics License Application (BLA).

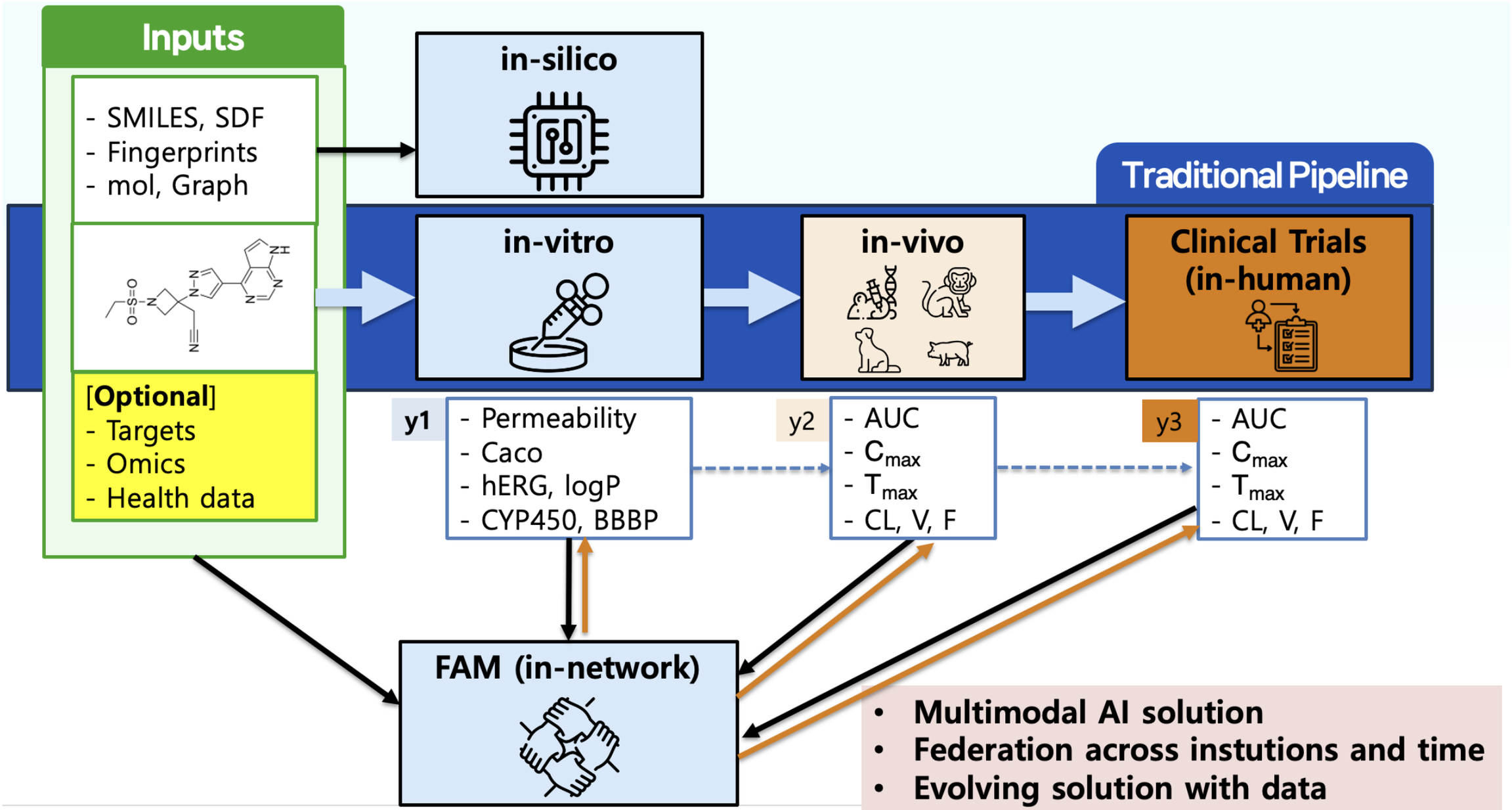
AI in Drug Discovery

- In-silico AI models are widely used for drug discovery and development



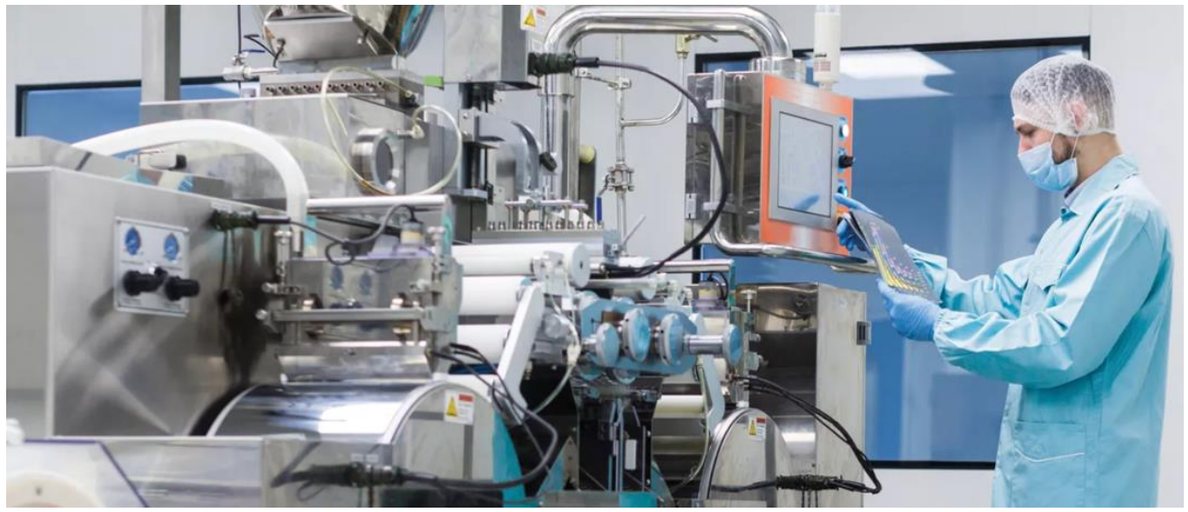
K-MELLODDY

- Federated Learning based AI model for drug discovery

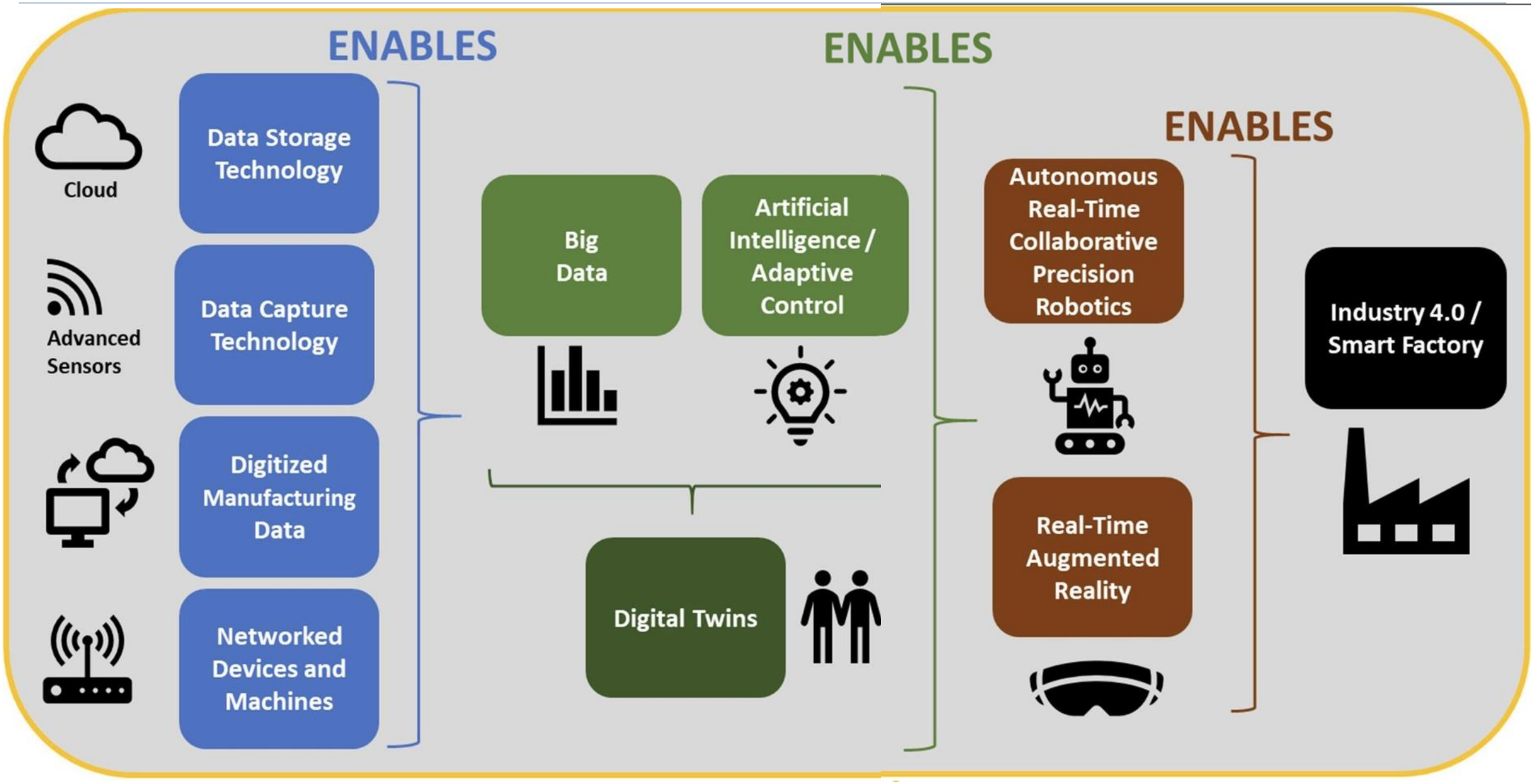


AI in Bio Manufacturing

- Predicting the impact of raw material variability
- Monitoring and control of bioreactors
- Process analytical technology
- Deviation management and change control
- Optimizing process design for scale up
- Predictive maintenance of equipments and utilities
- Predict in-process product quality and improve yields
- Visual inspection



Industry 4.0



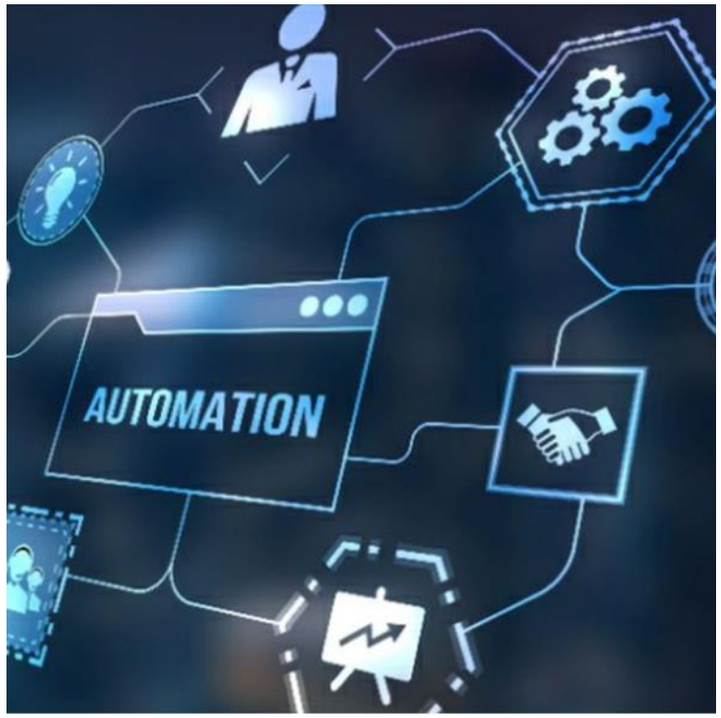
Industry 4.0 for pharmaceutical manufacturing: Preparing for the smart factories of the future, International Journal of Pharmaceutics Volume 602, 1 June 2021, 120554

BioPharma 4.0

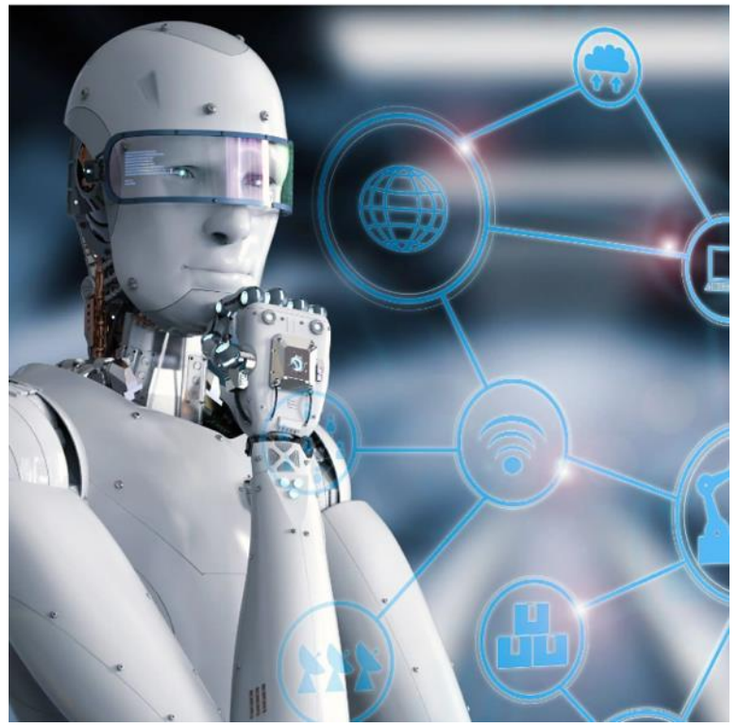
SNU 제약바이오인력양성센터 경기시흥글로벌바이오캠퍼스

Digitization → Automation → AI

- 문서의 **디지털화**
- 오류 최소화
- GMP 대응



- 관리의 **자동화**
- 분석의 자동화
- 실시간 분석



- Deviation **예측**
- 이상치, 품질 예측
- 시설,장치 예지 보수

AI 도입 이슈

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- AI 도입의 어려움
  - 성공했다는 결과만 기사화되고 어떻게 구축했는지 알려주지 않는다
  - 구축 기술을 이해하기 힘들다
  - AI 모델을 만들고 평가하기 위한 **데이터 확보**가 어렵다
- 기업마다 여건이 다르다
  - 다른 성공 사례를 그대로 도입할 수 없다
  - 시설, 방법, 목적, 구축기술 등이 다르다
  - 바이오 의약품 제조 공정은 특히 복잡하다
- 규제 산업에서의 AI 도입
  - AI 모델에 대한 명확하고 **객관적인 평가 기준**이 없다
  - AI의 장점(시간, 비용, 품질)을 위해 **AI 도입은 필수적 트렌드**

AI 도입의 어려움

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- 관성의 법칙
  - 기존 사업 영역과 사업 방식을 바꾸기 어렵다
  - 변화 선택에 따른 실패/책임에 대한 두려움이 크다
- 내부 역량의 한계
  - 기존 업무를 병행하며 새로운 기술을 배우고 적용하기 어렵다
  - AI 관련 지식 부족
- 외부 기관/전문가와의 협업 능력
  - 스스로 변신하기 어려운 부분은 AI 전문가와의 지속적 협업 필요
  - 관련 기업과 협력하는 능력이 필수
- 데이터의 가치를 놓친다
  - 이미 보유한 데이터의 이해, 이 데이터로부터 어떤 가치를 얻을지를 파악하지 못한다

AI in manufacturing 분야

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- Process **Design** and Scale-up
  - 최적의 프로세스 설계 파라미터 도출
- Process Monitoring and Fault **Detection**
  - 장비 장애, 부품, 생산품의 품질 예측
- Advanced Process **Control** (APC)
  - 센서 데이터 기반으로 최적의 제어
- Trend **Monitoring**
  - 고객 불만, 개선 피드백 분석

Advanced Process Control

- 기존의 방식으로는 동시에 여러 입력 변수의 변화를 예측하기 어려웠다.
- APC 는 센서 데이터를 AI 모델로 처리하여 원하는 결과를 얻기 위해 동적인 제어가 가능
- 향후 physics informed (또는 chemistry informed) AI 모델을 적용하여 성능을 더 개선될 것

Smart Monitoring and Maintenance

- 재고 관리, 예지 보수, 장애 예측 등이 가능하다
- 이미지 분석을 이용한 패키징, 라벨, vial 불량 예측
- 사람이 같이 개입하는 augmented Intelligence로 분류 성능을 높인다

AI is a Digital Reflex

- For QC Application

전통적 방식

- 문서, 규격 기반
- 경험, 노하우 중심
- **Rule centric**

보완

Digital Reflex

- 데이터 기반
- 이상징후 빠른 대응
- **AI centric**

AI Risk

- AI Error
  - 입력 데이터 오류에 의한 AI 오류
- AI misuse
  - 전문성 없이 단순히 AI 결과를 잘못 채택
- AI bias
  - 데이터의 편향성, 사회적 편향성이 AI 모델에도 반영될 수 있다
- Lack of transparency
  - AI의 구축, 활용, 설명에 대한 불명확성 (traceability와 explainability)으로 이해와 신뢰가 떨어지고 오류를 파악하기 어렵다
- Privacy and security
- Gaps in accountability : 여러 관계자의 책임 분산
- Obstacles to implementation

AI Limitation

- AI가 규제 과학을 대체할 수 없다
  - AI는 현미경, x-ray같은 훌륭한 **도구**이다
- AI는 논리적으로 동작하지 않는다
  - 반사운동 같이 입력 신호에 반응하는 것
  - 학습 데이터가 AI의 성능을 좌우한다
- AI는 설명력이 없다
  - 단계별로, 우리가 납득할 수 있는 답을 얻을 뿐

AI 대응

- AI가 잘 하는 영역
  - 예측, 생성, 추천, 최적화 작업
  - 속도도 빠르고, 지치지도 않고, 복제도 쉽다
- AI의 한계 영역
  - 감성적 능력 emotive capability
  - 창의성 creativity
  - 규제의 객관성

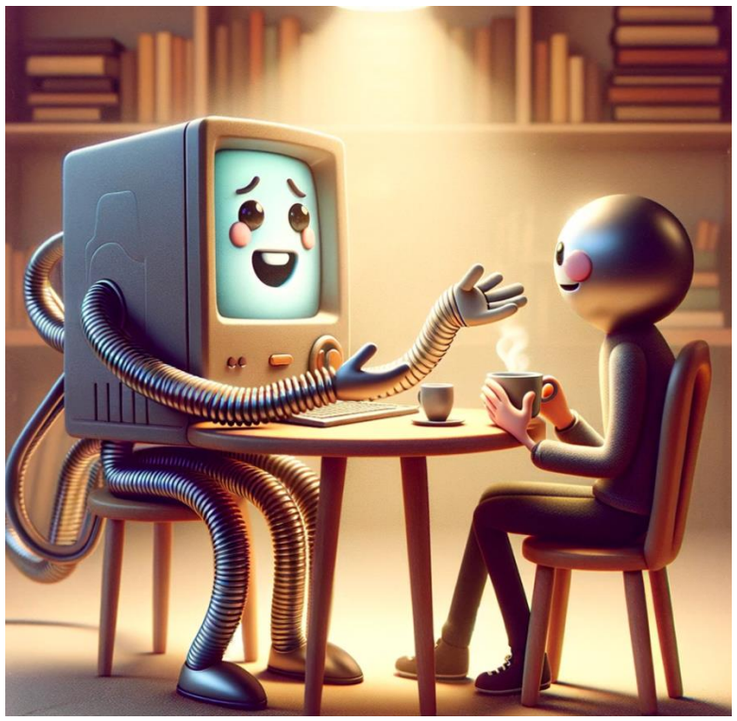
AI 도입과 규제

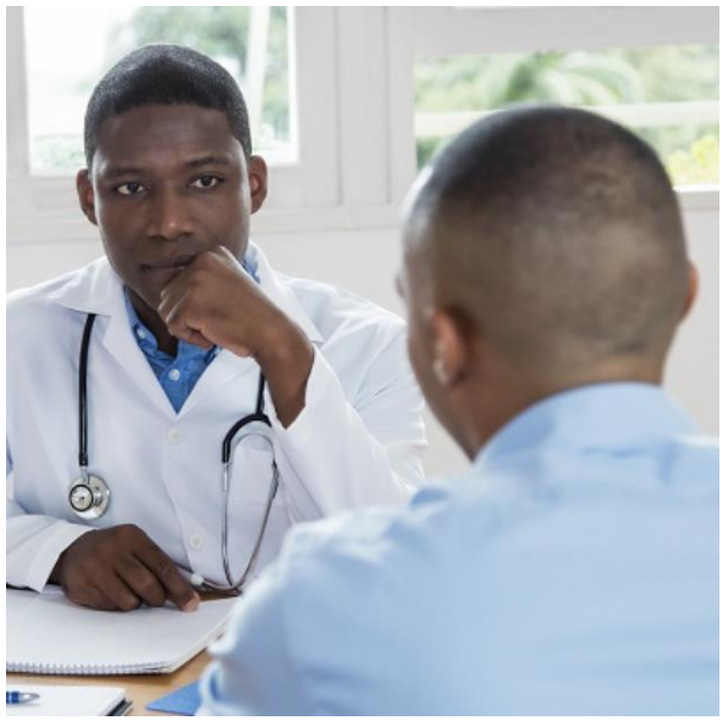
- DX/AI는 모든 산업 영역의 모든 단계에서 도입될 것
  - 신약개발, 기획, 마케팅, 생산, 피드백 등
- 모델 개발과 검증 데이터 확보가 관건
  - **목적 지향 데이터 수집**
  - **다수 기관의 데이터 기반 협력 방안 필요**
- 구축, 평가, 규제를 위한 새로운 **지수** 발굴 필요
  - 규제는 합격 여부를 판정하는 것

AI 원리 이해

AI의 정의

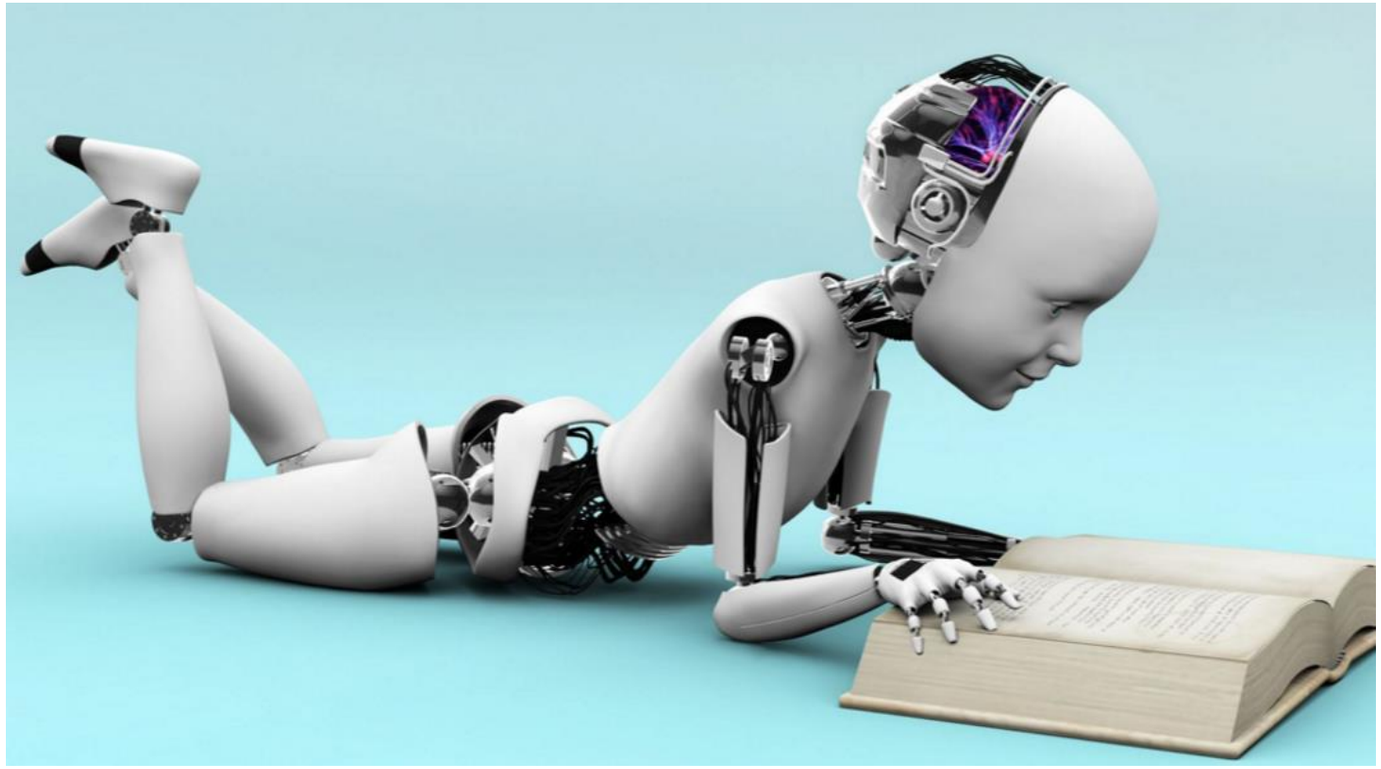
- AI (인공 지능)
  - "컴퓨터가 마치 **지능이 있는 것처럼** 똑똑하게 동작하는 것"
  - 거의 모든 산업에서 이미 활용 중
- AI 구현 방법
  1. 문법을 배워서 말을 할 줄 아는 능력을 구현?
  2. 전문가 지식이나 룰(공식)을 코딩한 "전문가 시스템(expert system)"?
  3. 데이터를 보고 스스로 학습하는 "머신러닝" 방식!





머신러닝 (machine learning)

- 현재 실용적으로 널리 사용되는 AI 구현 기술
  - 데이터를 보고 **스스로** 학습한다
- 사람의 지식이 개입되지 않는, **end-to-end 방식**으로 동작
  - 논리적인 해석이나 중간 값 확인을 필요로 하지 않는다
  - 사람은 **학습하는 기계**를 발명한 것



AI의 정의

- AI란
  - "컴퓨터가 마치 **지능이 있는 것처럼** 똑똑하게 동작하는 것"을 말한다
  - 내비게이터, 음악, 영화, 상품 추천, 예지정비 등
  - AI가 특정한 기술을 사용해야만 하는 것이 아니다. 예를 들어 신경망 (딥러닝) 을 사용해야만 하는 것은 아니다
- AI를 구현하는 기술
  - 사람처럼 "생각하는 능력"을 구현하려고 시도했고 언어학, 기호학 기반의 접근을 했으나 성공하지 못했다
  - 전문가의 지식을 코딩하는 "전문가 시스템"도 성공하지 못했다
- 머신러닝 기반 AI
  - 현재 동작하는 AI는, 데이터를 보고 스스로 학습하여 성능이 점차 개선되는 **"머신러닝" 방식의 AI이다**

DX/AI 도입 목적

- 현안 문제 해결
  - 가격/수요/판매/재고/물류/비용 예측, 경영 최적화, 안전, 장애관리, 예지보수, 자산관리, 빠른 제품시험, 제품설계, 지식관리, RPA 등
- 고객 요구 개선
  - 콜센터개선, 상담/거래 챗봇, 클레임 신속 분류/대응, 부품 준비, 마케팅 채널 분석, 타겟 마케팅 등
- 벤치마킹
  - 경쟁사 분석, 대체 상품 분석, 차별성 분석 등
- 혁신 제품/서비스 개발
  - 고객 분석, 시장 분석, 트렌드 분석, 신약개발, 의약품 제조 프로세스

AI 기본 기능

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- **예측 모델 (predictive model)**
  - 입력 데이터(X)를 보고 목적값(y)을 예측
  - 회귀 예측 (regression)
    - 수요예측, 재고예측, 판매예측, 잔여수명예측, 성능예측,약효예측 등 **수치 예측**
  - 분류 예측 (classification)
    - 사진판독, 양불판정, 기기 이상진단, 질병 예측, 단백질 결합 예측 등 **카테고리 예측**
- **생성 모델 (generative model)**
  - 어떤 조건에 맞는 텍스트, 이미지, 음악, 프로그램 코드, 약물후보 등을 생성하는 모델
  - chatGPT, BioNeMo

AI 활용 기능

- **최적화 (optimization)**
  - 최적의 제어변수 선택
  - 최적 운영환경 선택
  - 최적 설계 (기구, 약 제형 설계 등)
- **추천 (recommendation)**
  - 상품추천, 영화 추천, 음악 추천, 약 처방, 네비게이터,자율운전 등
- **군집화 (clustering)**
  - 특성이 유사한 샘플들을 그루핑하는 작업
  - 모든 연구는 군집화에서 시작된다

→ AI (머신러닝)는 위 다섯가지 중 하나의 작업을 수행한다

AI(ML) 적용 범위

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

1. 기획 (서비스/제품)
2. 수요 조사 (이용자 분석)
3. 설계 (매뉴얼 관리)
4. 시험 (시간 단축)
5. 최적 제어 (비용과 효율 개선)
6. 장애 예측 (장비, 부품, 제품 품질 관리(QC))
7. 수요 예측, 생산 계획, 재고 예측, 물류 최적화 (SCM)
8. 마케팅, 서비스/제품 추천 (CX)
9. 피드백 분석 (QA)

AI(ML) 이용자 니즈

- 외부 고객의 니즈 (서비스/제품에 대한)
  - **비용과 시간 절감**
  - **안전성** (UI 개선)
    - 오동작(실수)을 커버하는 인터페이스
  - **편리성** (UX 개선)
    - 설명이 필요 없는 인터페이스
  - **즐거움** (CX 개선)
    - Wants → Likes → Stickness
- 내부 고객의 니즈 (ML 솔류션에 대한)
  - 활용시의 **실제적인 효과** (정확도 개선, 비용과 시간 절감)
  - 도입 및 운영의 **편리성** (데이터 입력의 편리성 등)
  - **투자비용** (개발비, 장비, 교육비 등)

DX 추진 특징

- DX 도입 진입장벽이 낮아지고 있다
  - 네트워크 비용, 클라우드를 이용한 고속 연산 비용 감소
  - 공개 소프트웨어 사용으로 코딩의 진입 장벽이 낮아졌다
  - 기업간 경쟁이 치열하다
  - 그러나 **어디서, 어떻게 시작**해야 할지가 어렵다
- DX 추진 방법은 기업마다 다르다
  - 기업마다 비즈니스 모델, 보유한 데이터가 다르다
  - 의사결정자, 관리자, 실무자의 **이해도**가 서로 다르다

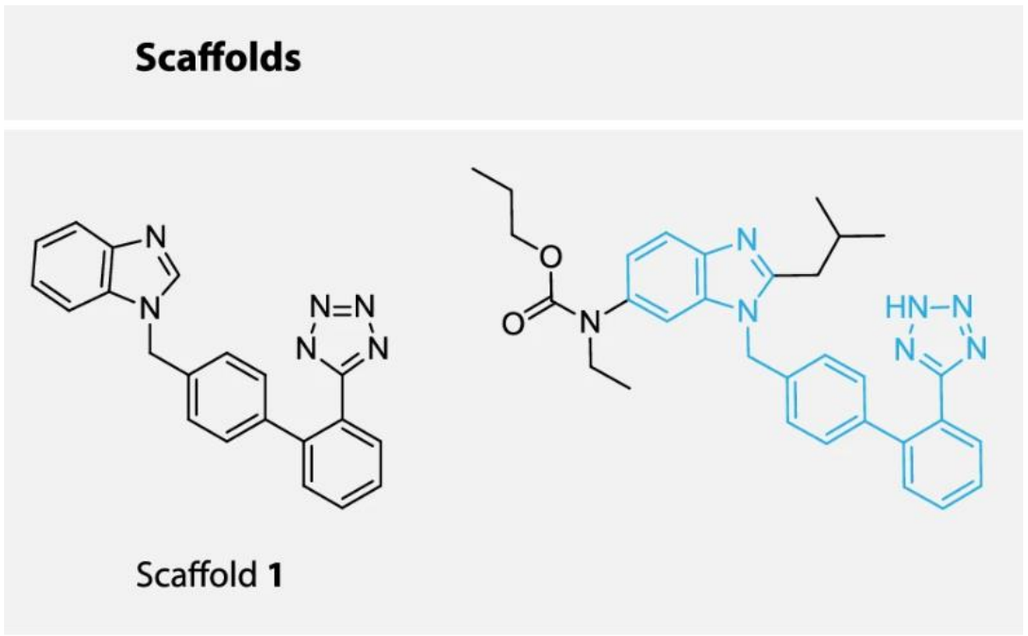
AI 활용 방법

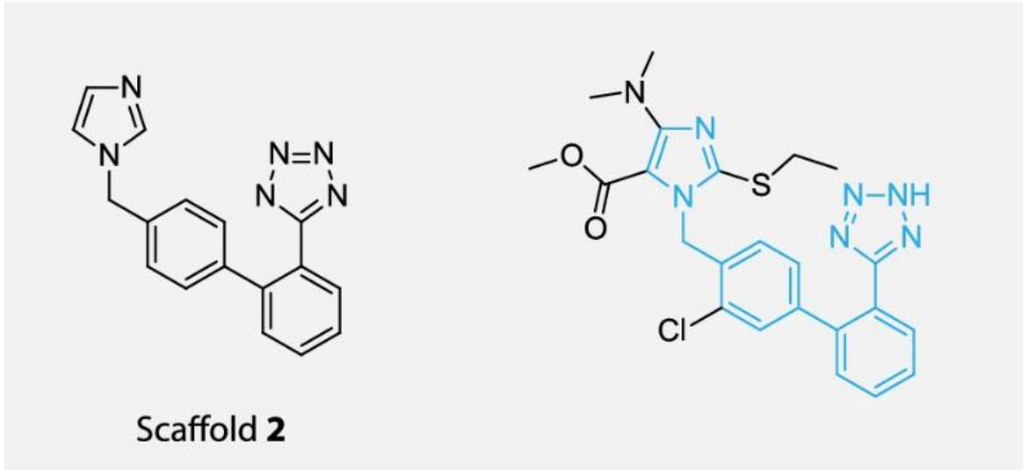
SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- AI 모델 직접 구현
  - 파이썬 프로그래밍
  - sklearn, tensorflow, keras, pytorch 등 패키지 활용
- AI 툴 이용
  - chatGPT (openAI)
  - gemini (Google)
  - Llama (Meta)
  - DALL-E (openAI)
  - Midjourney

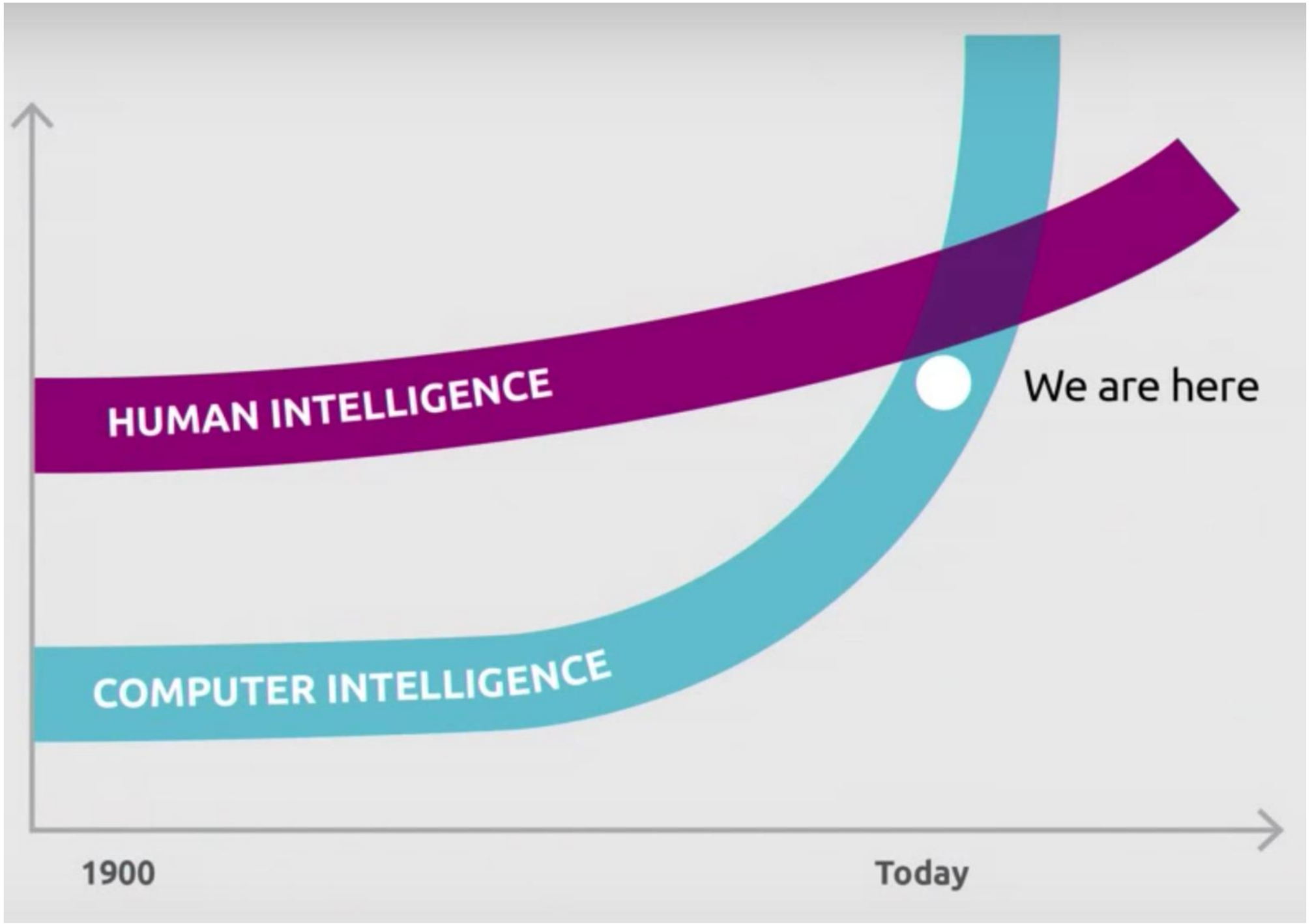
AI Impact

- 보고, 듣고, 쓰고, 말하는 지적 능력이 인간 수준을 넘었다
- 그림 그리기, 프로그래밍, **신약개발** 등으로 확대





At Crossing Point



Paradigm Shift

Logic-driven

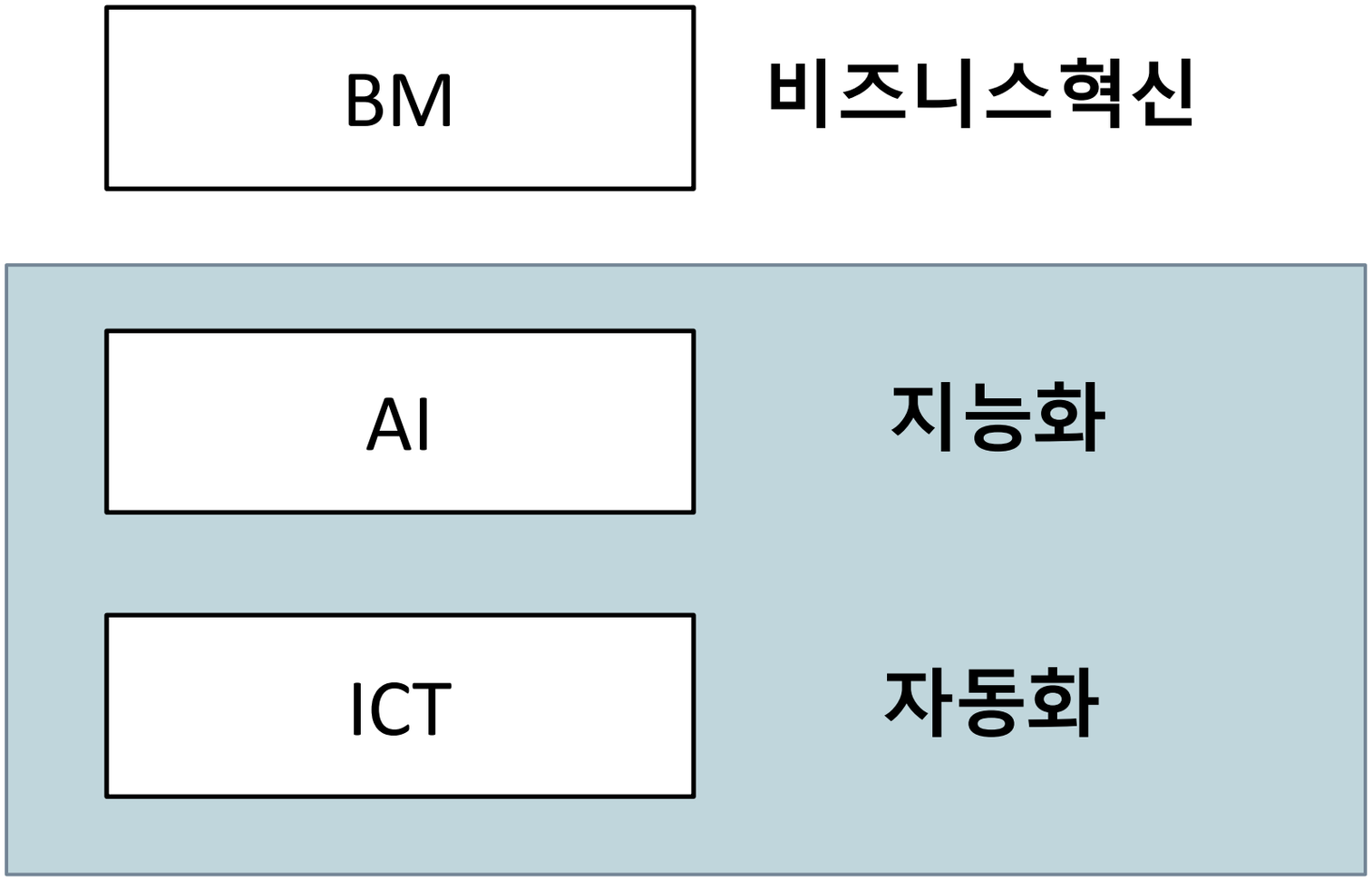
- **Natural Science**
  - Physics, Chemistry, Biology, Mathmatics
- **Experience-centric**
  - Knowledge-based

2012~2022

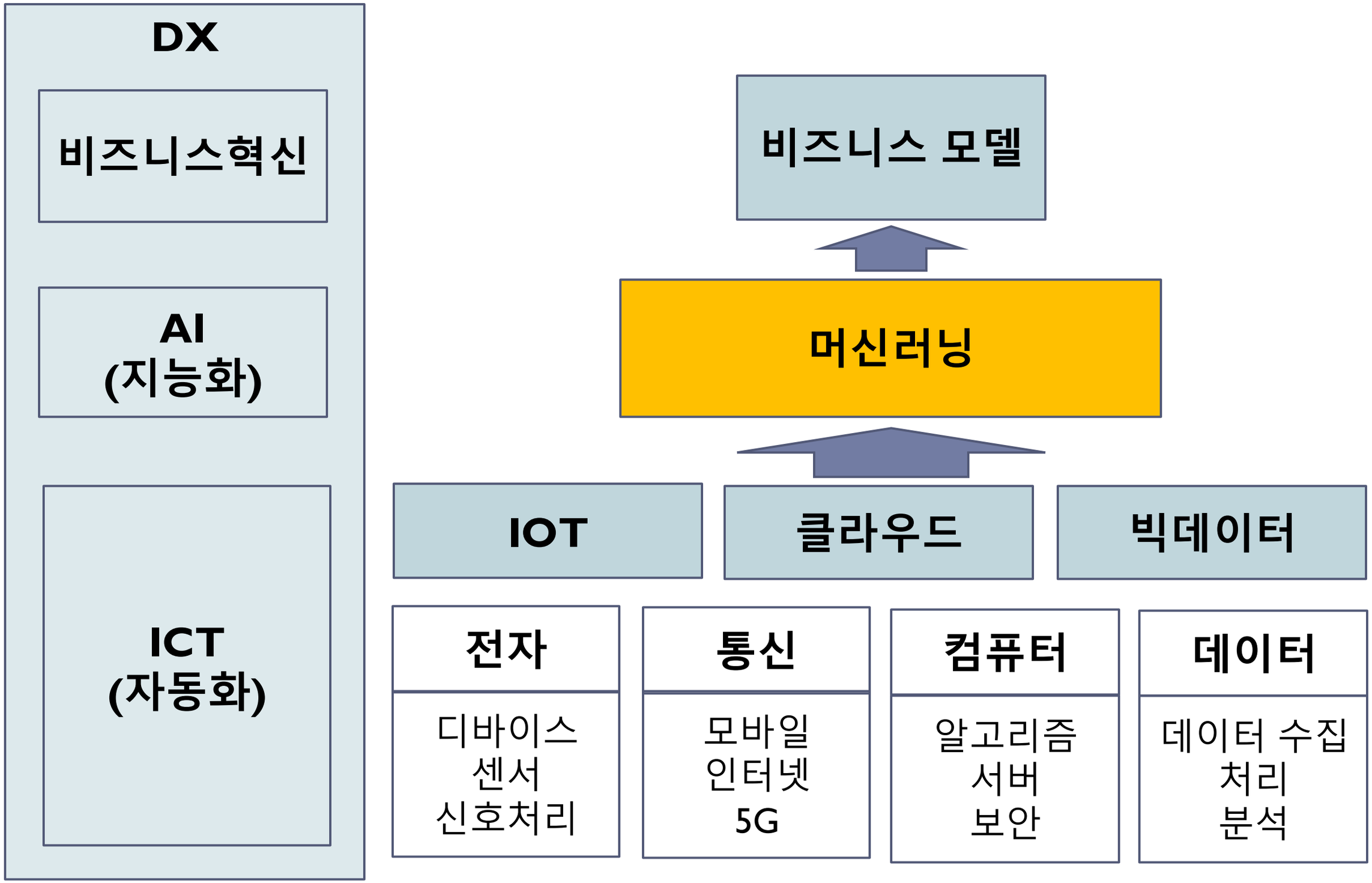
Data-driven

- **Data Science**
  - End-to-end model
- **AI-centric**
  - Machine Learning

DX의 범위



DX의 구성


DX
비즈니스혁신
AI
(지능화)
ICT
(자동화)
비즈니스 모델
머신러닝
IOT
클라우드
빅데이터
전자
디바이스
센서
신호처리
통신
모바일
인터넷
5G
컴퓨터
알고리즘
서버
보안
데이터
데이터 수집
처리
분석

자동화와 지능화

- 자동화
  - 프로그래머가 코딩한 **로직대로** 동작한다
  - 정해진 알고리즘(플로우 차트)대로 업무를 빠르고 정확하게 수행
  - 개발자(사람)의 지식과 경험이 성능을 좌우
- 지능화
  - 데이터를 보면서 모델(소프트웨어)의 성능이 점차 개선된다
  - 학습에 사용하는 데이터의 양과 **다양성**에 따라서 성능이 향상
  - 개발자 역할은 모델을 잘 만들고 좋은 데이터를 공급하는 것

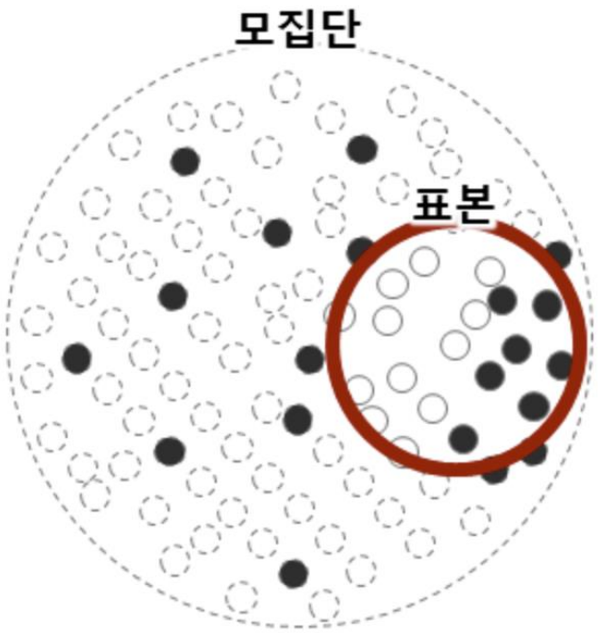
DX 관련 기술 1

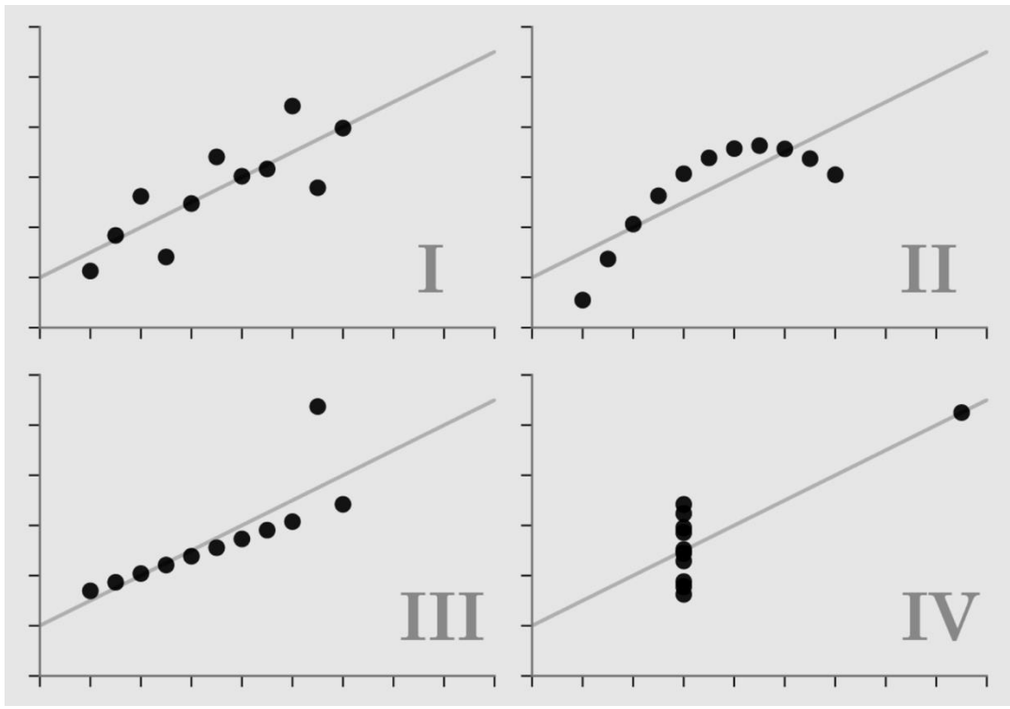
SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- 인공지능(AI)
  - 지능이 있는 것처럼 **똑똑하게** 일을 처리하는 소프트웨어
- 머신러닝(Machine Learning)
  - 컴퓨터가 데이터를 많이 볼수록 성능을 점차 개선하는 방법
- 빅데이터(Big Data) 분석
  - 대량의, 그리고 다양한 데이터 분석으로 일반적인(통계적인) 분석으로는 찾지 못하던 **새로운 insight**를 얻는 것
- 클라우드(Cloud)
  - 데이터 및 컴퓨팅 파워를 한 곳에 집중하여 비용 효율적으로 빅데이터 분석과 AI 구현을 가능하게 하는 기술
- IOT (Internet of Thing, 사물인터넷)
  - 사물에 센서, 통신 기능을 추가하여 빅데이터 수집을 용이하게 하는 기술

DX 관련 기술 2

- 데이터마이닝
  - 기 구축된 데이터베이스에서 **새로운** 지식을 얻는 것
- 비즈니스 인텔리전스
  - 데이터 분석을 통해 새로운 **비즈니스** 전략을 얻는 것
- 통계분석
  - **샘플 데이터**로부터 전체 데이터의 속성을 파악하는 것
  - 가설, 검정, 추정, 오차범위, 신뢰구간 등을 사용하며 수학적(논리적) 증명을 필요로 한다

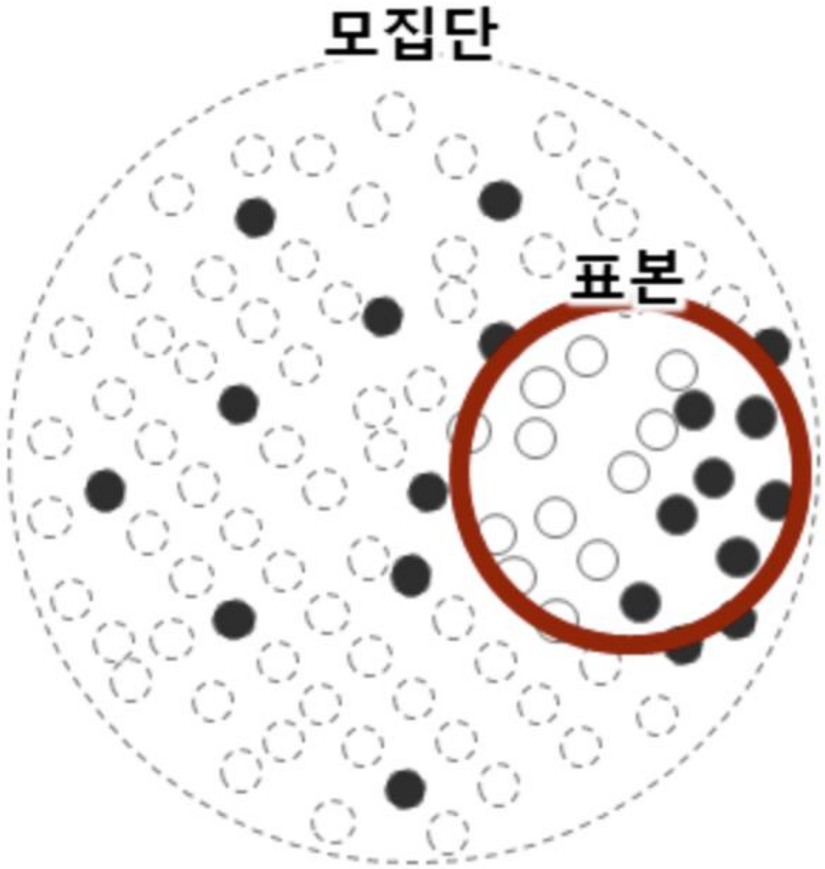




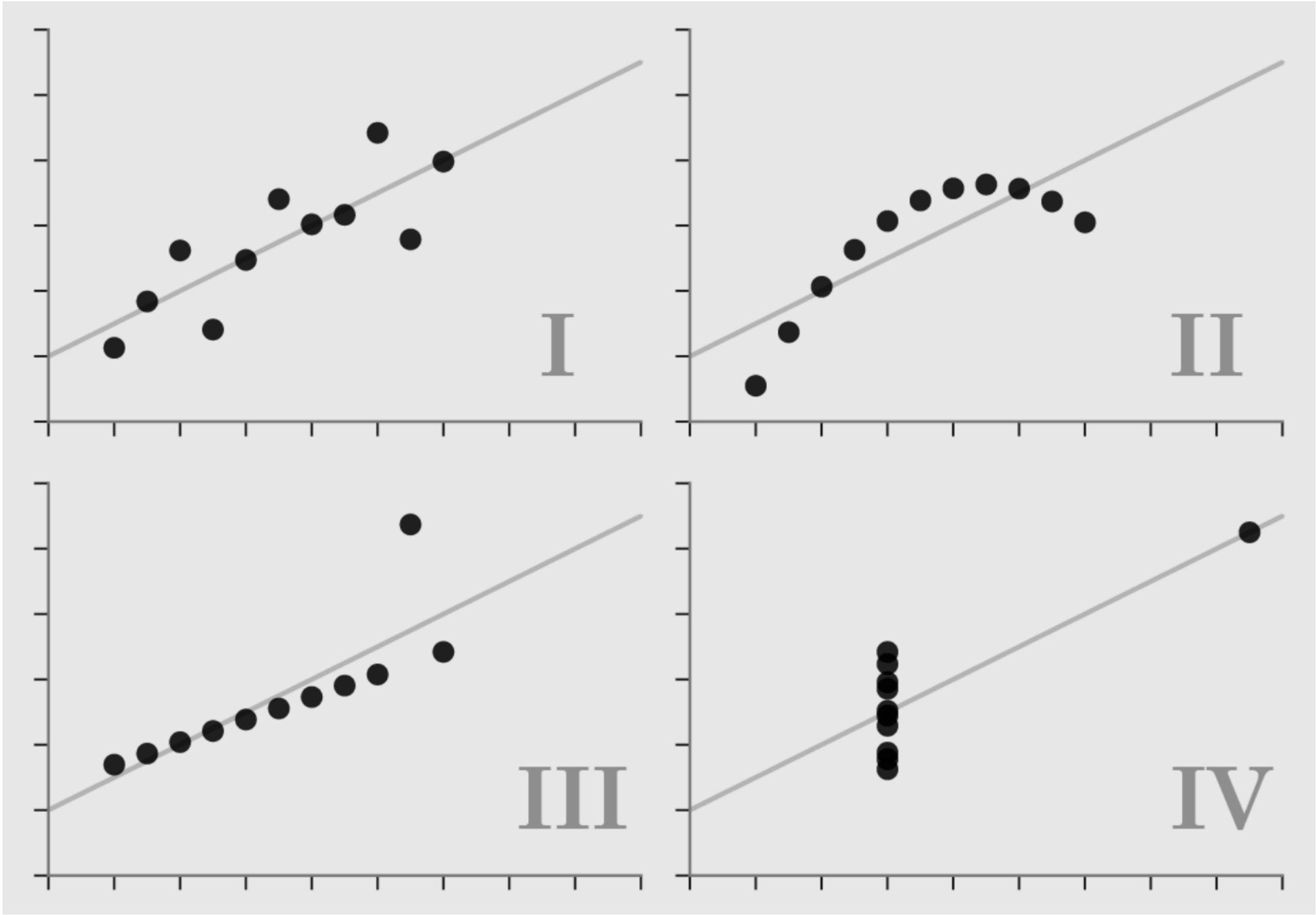
빅데이터 분석

통계 분석

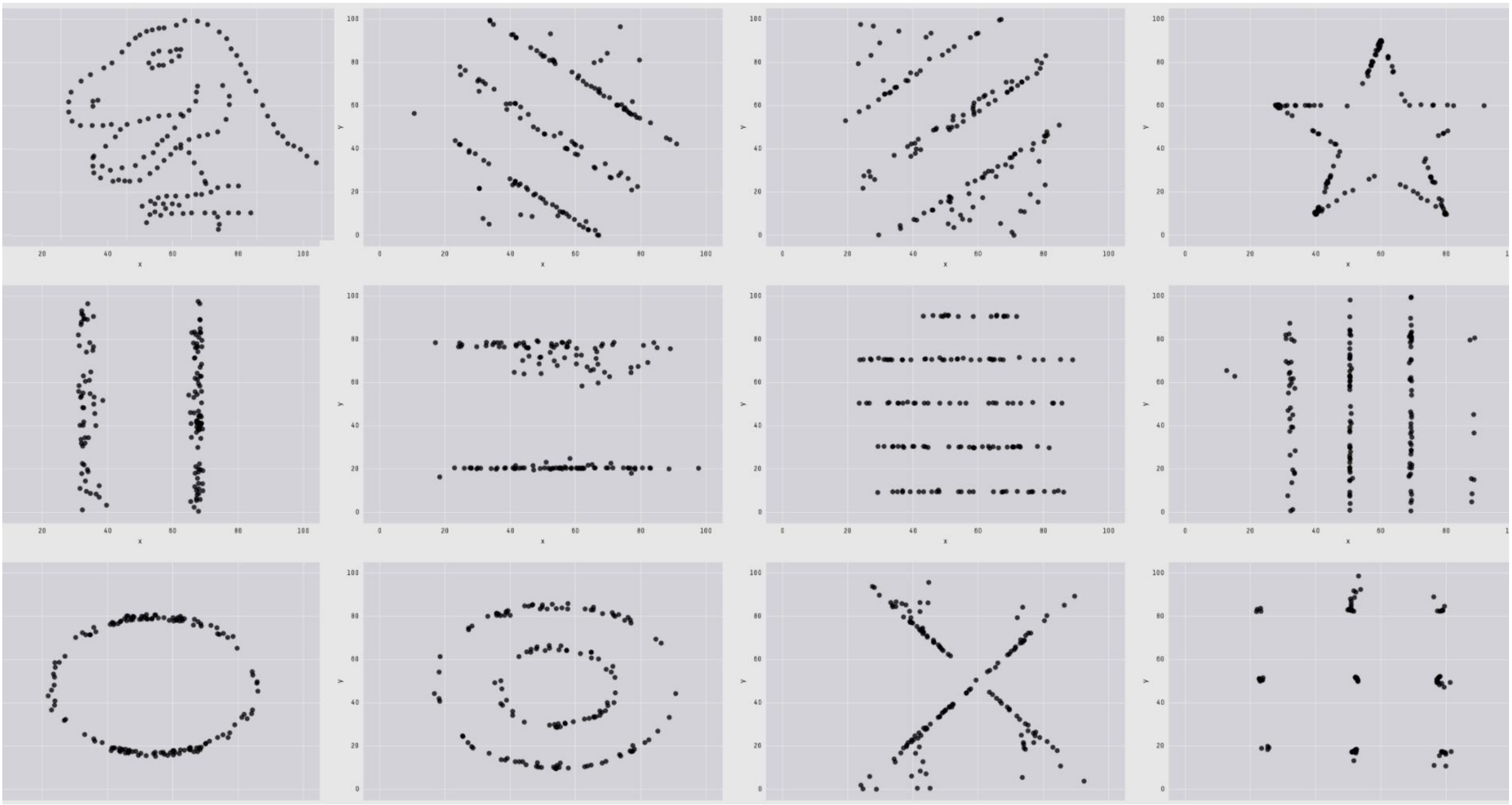
- 표본(sample)으로부터 모집단(population)의 특성을 설명하는 것
- 수학적 설명을 위해서 오차범위, 신뢰구간, 가설, 검증 등을 필요로 한다



동일한 평균, 표준편차, 상관계수

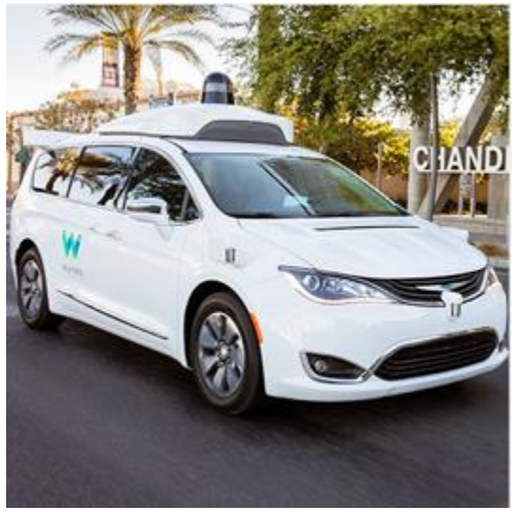


동일한 평균, 표준편차, 상관계수



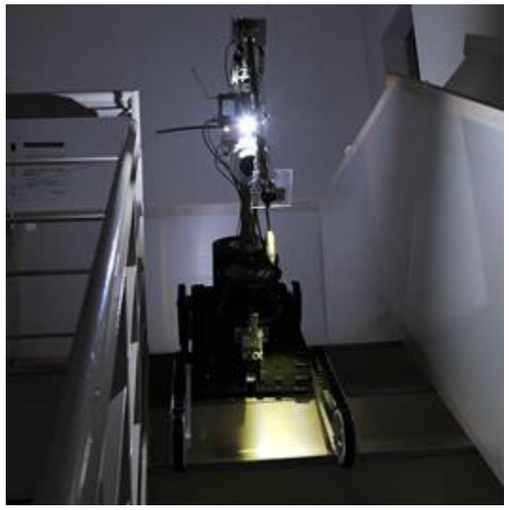
Robots





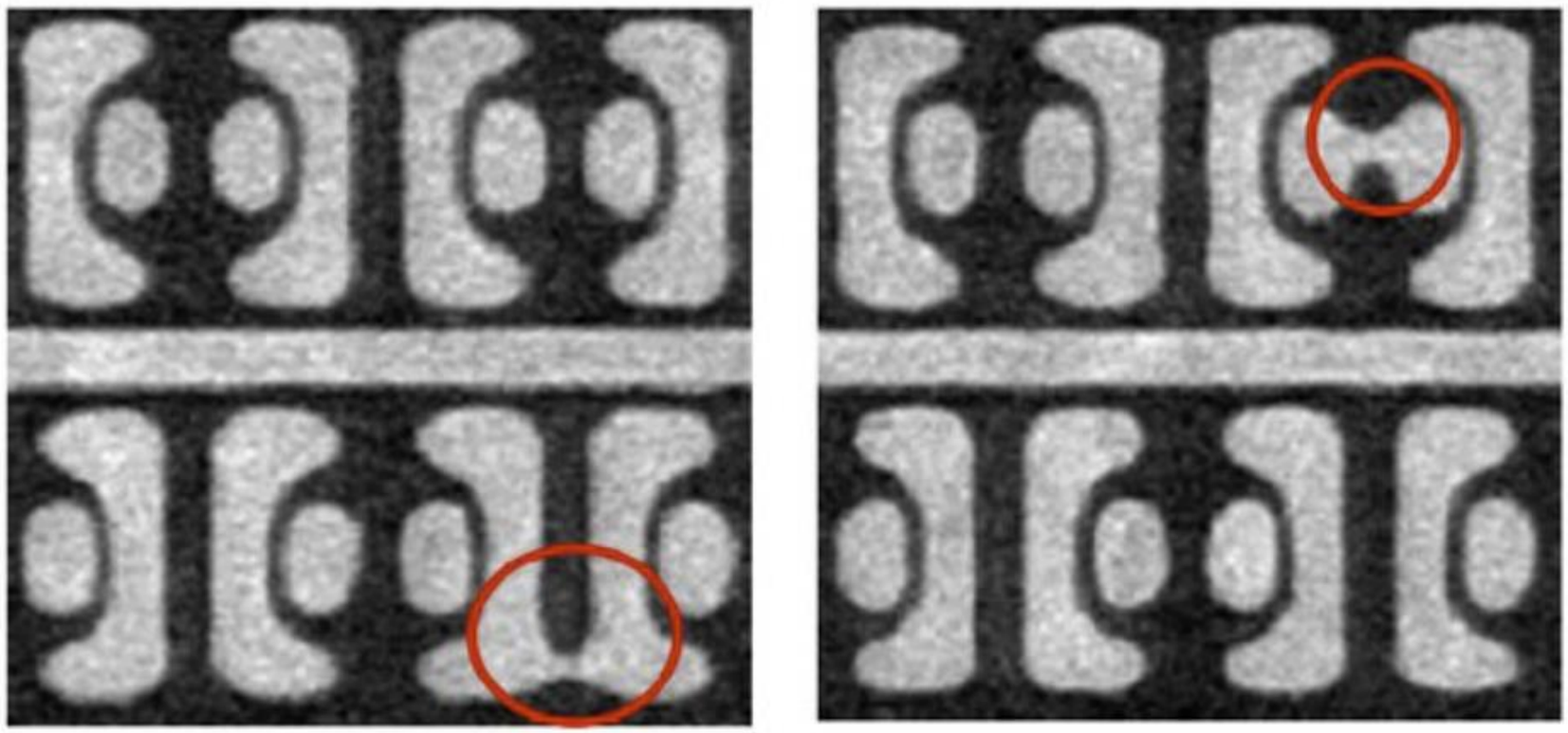




defect detection

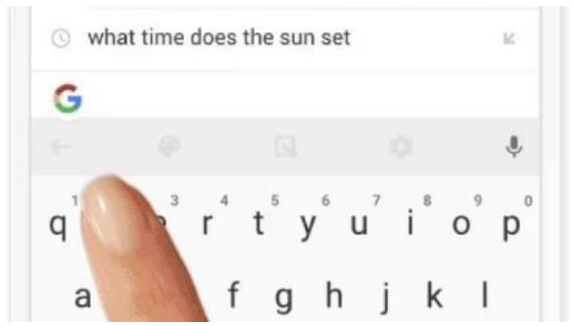
- 사람이 찾기 어려운 작은 결함을 찾는다
- 이미지 기반 작업 공정 실시간



Data, Data, Data

Federated Learning

Gboard



MACHINE LEARNING LEDGER ORCHESTRATION FOR DRUG DISCOVERY

JUNE 2019 - MAY 2022

MELLODDY

powered by aws

PHARMA PARTNERS

AMGEN · astellas · AstraZeneca · BAYER · Boehringer Ingelheim · gsk · Janssen · MERCK · NOVARTIS · SERVIER

PUBLIC PARTNERS

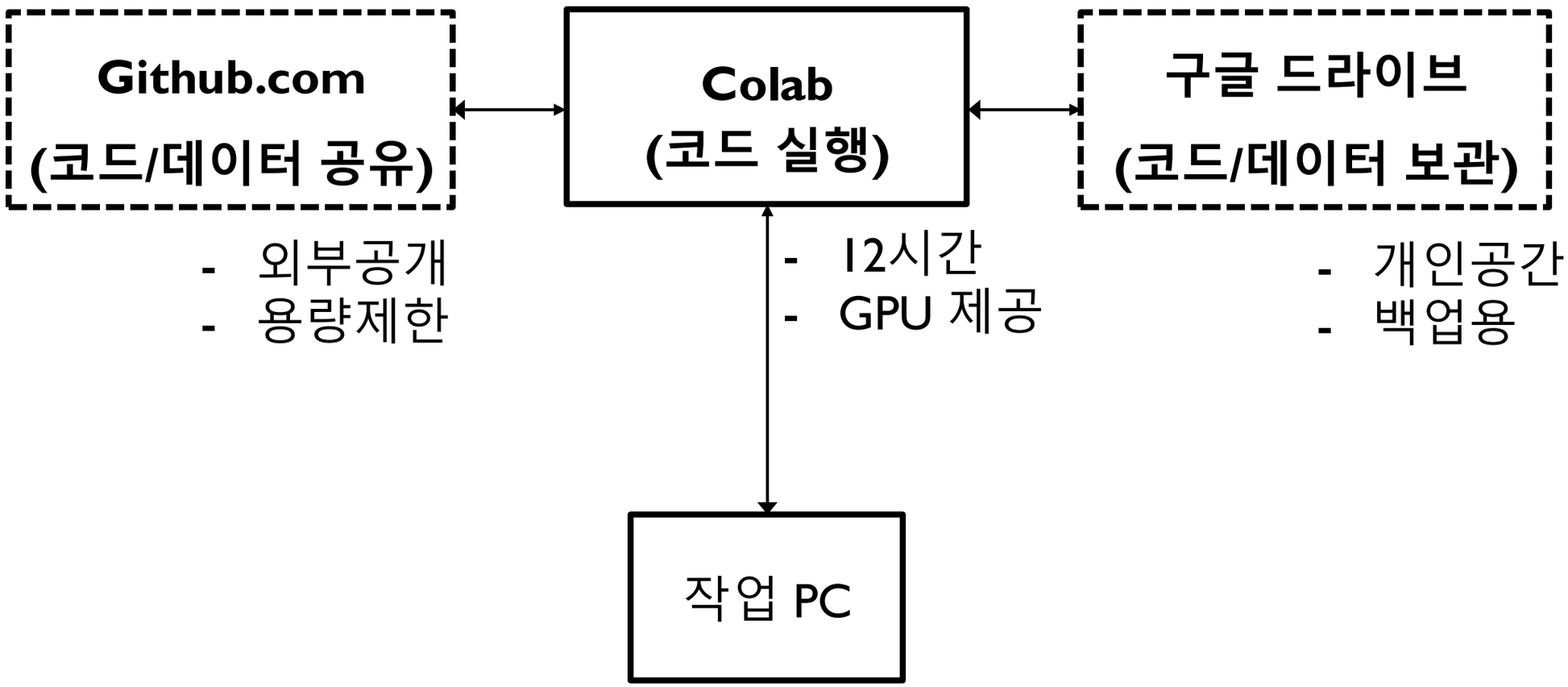
MŰEGYETEM 1782 · IKTOS · KU LEUVEN · loodse · NVIDIA · OWKIN · SUBSTRA FOUNDATION

파이썬 프로그래밍

파이썬 실행 환경

- 쥬피터 노트북 (Jupyter notebook)
  - PC에서는 Anaconda 설치
  - 구글 코랩 사용 colab
    - Chrome으로 접속
    - 구글 계정 가입(gmail) – 구글 드라이브 사용
  - github.com 가입
- 쥬피터 노트북 특징
  - 모든 작업을 셀 단위로 수행
  - 코드 셀 – 프로그램 영역
  - 마크다운 셀 – 도큐먼트 영역

코랩 실행 환경


Github.com
(코드/데이터 공유)
- 외부공개
- 용량제한
Colab
(코드 실행)
- 12시간
- GPU 제공
구글 드라이브
(코드/데이터 보관)
- 개인공간
- 백업용
작업 PC

쥬피터 팁

- 각 셀은 프로그램 코드 또는 마크업 문서 중 하나로 사용
  - 셀을 실행하려면 Shift + Enter 를 입력한다 (셀 실행후 커서가 다음 셀로 이동)
  - 셀을 Ctrl + Enter로 실행하면 셀 실행 후에 현재의 셀에 남아 있는다
- colab에서 셀에 대한 명령어 (ctrl + m + ...)

- h : help
- a : 위에 셀 추가 (above)
- b : 아래에 셀 추가 (below)
- d : 셀 삭제 (delete)
- m : 문서 모드로 전환 (mark up)
- y : code 모드로 전환
- o : 출력 보이기/안보이기 (output)
- '-' : 셀 둘로 나누기
- 'shift' 두 개 이상의 셀 선택 후에 (ctrl + 클릭으로) : 셀 합치기

파이썬 기초 문법

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- 기본 변수
  - 정수, 소수, 문자열, 논리값을 표현한다
- 리스트
  - 임의의 데이터를 목록을 만들어 담을 수 있다. [ ] 로 표현된다
- 튜플
  - 상수화된 리스트이다. 값의 변경이 불가하다. ( ) 로 표현된다
- 사전(dictionary)
  - 모든 항목이 항상 "키(key)"와 "값(value)" 짝으로 구성. { }로 표현된다
- 논리 흐름
  - 조건의 만족을 논리적으로 판단한다 if, else, elif 사용
  - 반복을 위하여 for, while을 사용한다
- 함수
  - 사용자가 임의의 기능을 정의할 수 있다. def를 사용

기호 사용

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- [ ]
  - 리스트를 표현, x = [1,2,3]
  - 인덱싱을 표현, z = x[0]
- ( )
  - 튜플을 표현
  - 함수 호출시 인자를 입력, x = function(a,b,c)
- { }
  - 사전(dictionary)을 표현
  - 집합(set) 을 표시

판다스

- 데이터프레임(DataFrame)을 사용하기 위한 파이썬 패키지
  - 데이터프레임은 2차원 테이블 구조로 "액셀"과 같은 용도로 사용된다
  - (참고) 1차원의 데이터를 다루기 위해서는 리스트가 사용된다
- 데이터 프레임 생성 방법
  - 딕셔너리로부터 만드는 방법
  - 배열, 리스트, 튜플로부터 만드는 방법
  - csv 파일을 읽어 만드는 방법
  - 액셀 파일을 읽어 만드는 방법
- Series
  - 데이터프레임에서 한 컬럼을 취하면 시리즈가 된다
  - 시리즈는 리스트처럼 1차원 데이터인데, 인덱스가 붙어 있다

데이터 프레임 생성

```
## 딕셔너리에서 데이터프레임을 만드는 예

x = {'city': ['서울', '부산', '대구', '대전', '광주'],
     'population': [990, 350, 250, 154, 150],
     'temp': [17, 18, 18, 19, 20]}

df = pd.DataFrame(x)
```

	city	population	temp
0	서울	990	True
1	부산	350	True
2	대구	250	True
3	대전	154	True
4	광주	150	False

주요 함수

- 인덱스 설정
  - set_index() 사용
  - 인데스 원상복구: reset_index() 사용
- 데이터프레임 합치기 (디폴트로 행 방향으로 합친다-샘플 추가)
  - concat() 사용
- 원하는 행 찾기, loc, iloc()
  - loc(): 특정 인덱스에 해당하는 샘플을 얻는다
  - iloc(): 정수형 순서 번호로 샘플을 얻는다
- 특정 컬럼을 기준으로 순서 정렬하기
  - df.sort_values(['인구’])
- 데이터프레임 저장¶
  - to_csv(): csv 파일로 저장
  - to_excel(): 액셀 파일로 저장
  - 한글이 깨지는 경우 encoding='utf-8', 'cp949','euc-kr' 등의 옵션 사용

주요 함수

- 파일을 읽어 데이터프레임 만들기
  - read_csv(): csv 파일을 데이터프레임으로 읽기
  - read_excel(): 액셀, xlsx 파일을 데이터프레임으로 읽기merge
- 특정 열(컬럼)을 기준으로 두 데이터프레임 합치기
  - (참고) 같은 인덱스를 기준으로 합칠때는 join()을 사용한다
- groupby
  - 데이터프레임을 특정 조건에 맞는 그룹을 세분화하여 처리한다
  - 내부적으로 여러개의 데이터프레임으로 나누어진다

numpy

- **numpy**
  - 데이터프레임은 테이블 구조의 데이터를 편리하게 조작하기 위해서 사용한다
  - 숫자만으로 구성된 데이터를 대상으로 "연산"을 하기 위해서는 어레이(array)로 표현되어야 한다
    - 어레이를 ndarray(n-dimensional array)라고도 부른다
  - 어레이를 사용하기 위해 넘파이 라이브러리를 사용한다
- **리스트, 데이터프레임, 어레이의 차이**
  - 리스트는 데이터를 1차원 "목록"으로 만들어 조작하는 용도 (추가, 삭제, 변형 등)
  - 데이터프레임은 2차원 테이블 구조의 데이터를 만들고 조작하는 용도
  - 어레이는 연산을 위해 숫자만으로 구성된, 수학의 매트릭스와 같은 용도

주요 함수

SNU 제약바이오인력양성센터 경기시흥글로벌바이오캠퍼스

- arange
  - range 타입의 "범위" 데이터를 생성한 후 이를 어레이로 만들어준다
- reshape
  - 어레이의 구조(shape)를 바꾼다
- savetxt()
  - 어레이를 csv 파일로 저장한다
- loadtxt()
  - csv 파일을 어레이로 읽기 (데이터프레임으로 읽지 않음)
  - 대용량 데이터인 경우, 데이터프레임으로 읽는 것보다 속도가 빠르다

함수 정의

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- **함수 사용**
  - 반복 사용되는 작업은 함수로 만들면 편리하게 다시 사용할 수 있다
  - 함수를 정의할 때 def를 사용한다
  - 함수를 호출할 때 인자(arguments)를 넘겨줄 수 있다
  - 함수 실행 결과로 어떤 값을 받으려면 return 문을 사용한다
- **lambda, 익명 함수 정의**
  - def, return을 사용하지 않고 간단히 함수를 정의할 수 있다
- **map, 리스트에 함수 적용하기**
  - map의 첫번째 인자에는 함수를, 두번째 인자에는 데이터를 넣는다
  - map은 리스트 외에도 튜플, 배열에 대해서도 사용할 수 있다
  - lambda를 사용하여 함수 내용을 직접 정의할 수도 있다

```
list(map(lambda i: i*10, data)
```

- **apply, 시리즈나 데이터프레임에 함수 적용**
  - apply는 리스트, 튜플, 배열에는 사용할 수 없다

파일 다루기

SNU 제약바이오 인력양성센터 경기시흥글로벌바이오캠퍼스

- 파일 열기 open(), with로 열기
- csv, xlsx 파일 다루기
- zip 파일 다루기
- 폴더 내 파일 목록 다루기 (glob, listdir)
- 폴더/파일 생성과 삭제
- **read_csv 옵션**
  - nrows=1000 # 상위 1000줄만 읽기
  - skiprows = 3 # 처음 3행 건너뛰기
  - skipfooter = 1000 # 맨 뒤의 1000행은 읽지 않기
  - usecols= (0,2,4) # 해당 컬럼만 읽기

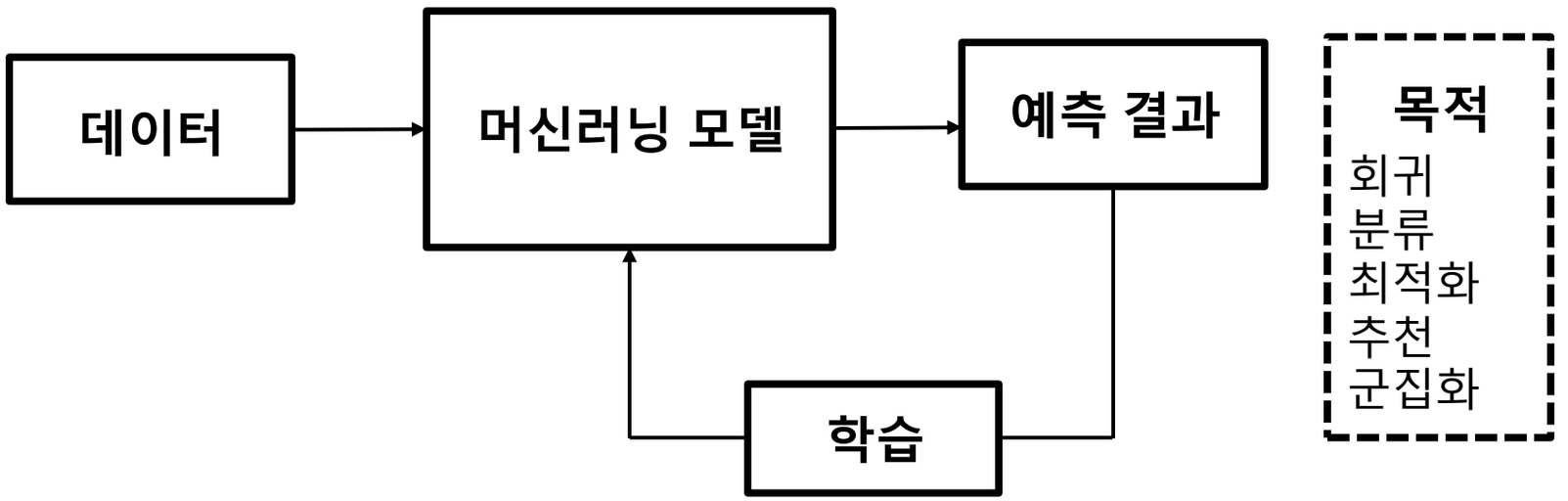
폴더 다루기

- 액셀 한글 인코딩
  - 문자의 인코딩은 기본적으로 utf-8 을 사용하나 한글의 경우 다른 인코딩으로 저장되는 경우가 있다
    - encoding = 'cpc949', 'euc-kr'을 선택해야 하는 경우도 있다
- 현재 폴더 위치 보기
  - getcwd()를 사용한다 (current working directory)
- 폴더의 파일 목록 얻기
  - glob() 사용: 파일 타입을 편리하게 지정할 수 있다
  - listdir() 사용: 경로를 지정할 수 있다
- os 패키지에서 제공하는 함수
  - 폴더 생성 mkdir(), 폴더 위치 이동 chdir()
  - 파일 이름 변경 rename(), 파일 삭제 remove(), 폴더 삭제 rmdir()

머신러닝 개요

머신러닝 모델

- 수치를 예측하는 회귀, 카테고리를 예측하는 분류, 최적의 추천 등을 수행하는 소프트웨어
- 데이터를 보고 학습하여 점차 성능이 개선된다



지도 및 비지도학습

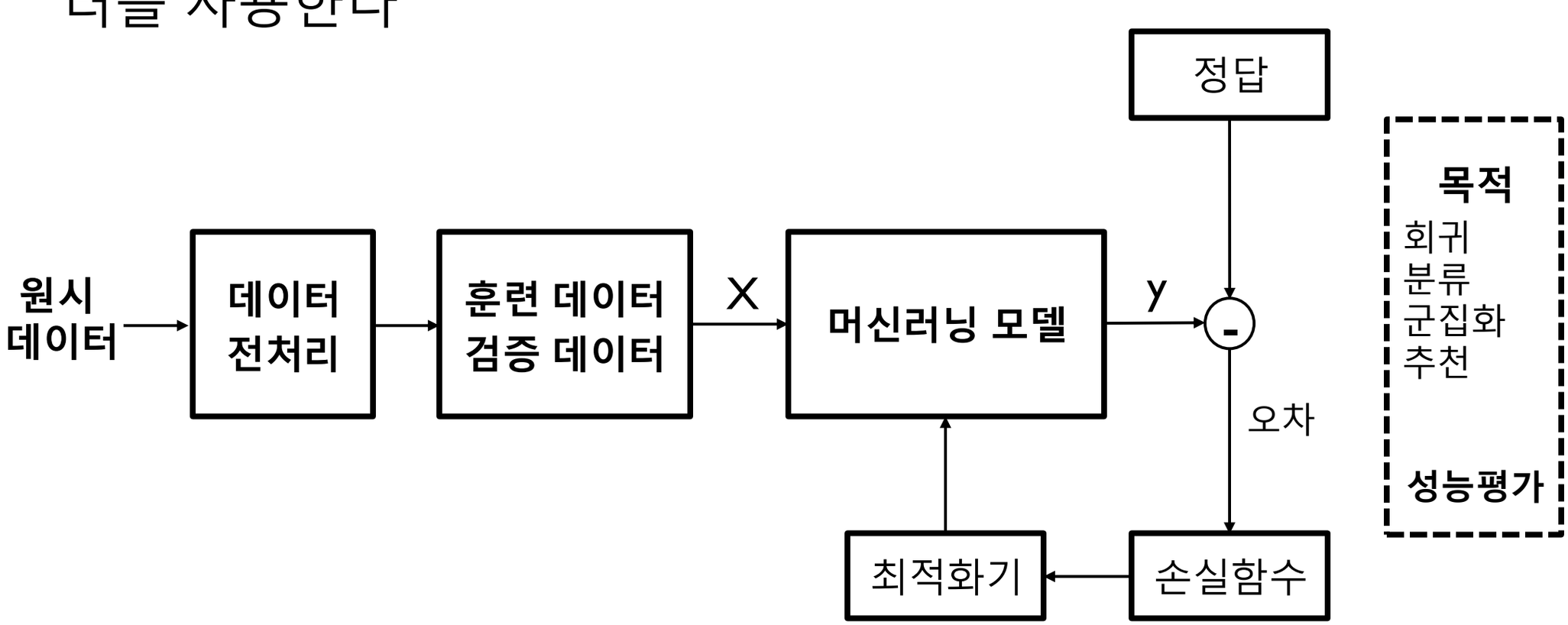
- 지도 학습(supervised learning)
  - 정답이 있고 이를 예측하는 학습
  - 정답을 목적변수 (target variable) 또는 레이블(label)이라고 부른다.
- 비지도 학습 (unsupervised learning)
  - 정답이 없이 데이터에 내포된 의미(insight)를 찾는 것
  - 탐색적 분석, 시각화, 연관분석, 클러스터링(군집화) 등
  - 주성분분석(PCA), 차원축소, t-SNE
  - 언어 모델 등에서 대량의 텍스트를 이용한 사전 학습 등
  - 표현학습 (representation learning, 임베딩 학습) 등
    - 생성 모델은 표현학습을 통해서 이미지, 텍스트 등을 표현하는 방법을 미리 사전학습해야 한다

머신러닝 모델

- 머신러닝에서는 모델을 만드는 것이 핵심이다.
  - 스팸 메일을 찾아내는 모델
  - 수익 예측 모델
  - 도난 카드 사용 검출 모델
- 수식은 가장 명확한 모델이다
  - 그러나 현실의 복잡한 현상은 수식으로 모델링하기 어렵다
  - 데이터 기반의 모델이 필요하다 – **머신러닝 모델**
- 모델은 구조와 파라미터로 구성된다.
  - 모델 구조: 모델의 동작 방식 (선형모델, 트리모델, 신경망 모델 등)
  - 모델 파라미터: 모델이 잘 동작하도록 학습된 가중치(weight)
    - y =ax + b
    - 위의 입출력 관계는 **모델 구조**이고, a, b 는 **모델 파라미터**임

머신러닝 모델 구현

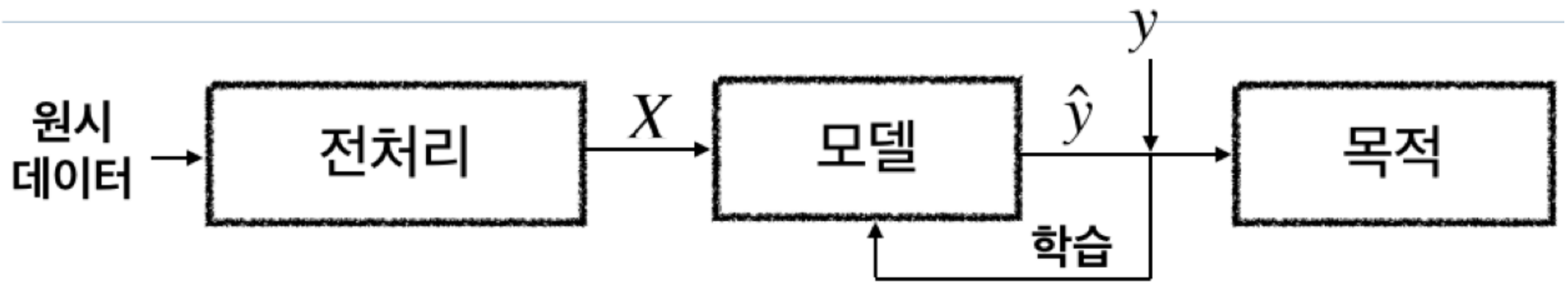
- 데이터 전처리
  - 모델의 성능을 높이기 위해서 결측치 처리, 이상치 처리, 스케일링, 카테고리 인코딩 등을 수행하는 것
- 훈련/검증 데이터
  - 모델을 만드는 데는 훈련 데이터를, 성능을 검증하는 데는 검증 데이터를 사용한다



훈련과 검증

- 모델이 데이터를 이용하여 학습하는 과정을 훈련 (training)이라고 한다.
- 모델 파라미터 값 (예를 들어 선형 회귀에서 가중치의 값)의 초기값은 랜덤한 값을 준다
- 모델을 훈련시킨 후에는 모델이 제대로 동작하는지 검증하기 위해서 검증데이터를 사용한다

머신러닝 프로세스



결측치 처리
오류값 처리
스케일링
데이터 변환
. 카테고리 변환
. 로그, 역수 변환
특성공학
. 차원축소
. PCA

선형모델
로지스틱회귀
SVM
결정트리
랜덤포레스트
그라디언트부스트
kNN, Bayes
CNN
RNN

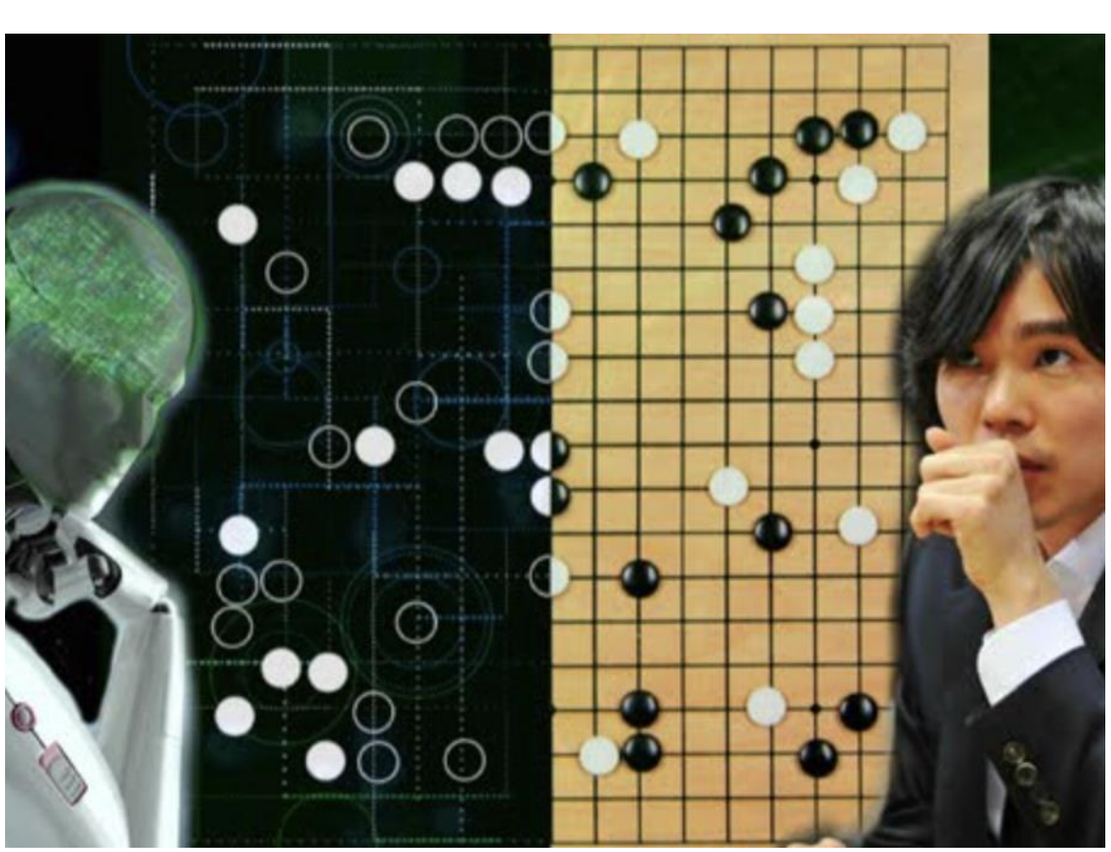
클러스터링
설명적 분석
. EDA, 시각화
회귀 예측
분류 예측
추천

머신러닝 모델 비교

머신러닝 유형		알고리즘	특징
지도학습	선형 계열	선형 모델, SVM 로지스틱회귀	곱셈과 덧셈으로 점수를 구하고 이를 이용하여 회귀와 분류 예측
	신경망	MLP, CNN, RNN, Transformer Graph Network	매트릭스 연산을 기반으로 점수를 계산하며 활성화 함수 도입
	트리 계열	결정 트리, 랜덤포레스트, 그라디언트부스팅	True/False 선택을 반복하여 회귀와 분류 예측 수행. 스케일링이 필요없다
	기타	kNN, 베이즈	특성 공간상의 거리를 기준, 또는 조건부 확률을 기준으로 예측
비지도학습	클러스터링	k-means, DBSCAN	특성 공간상 거리 또는 유사도를 기준으로 샘플을 그루핑
	데이터 변환	스케일링, 로그변환, 카테고리 인코딩	모델의 성능을 높이 위한 데이터 전처리
	차원 축소	PCA, t-SNE	계산량을 줄이고 모델 성능을 향상, 또는 시각화를 위한 차원 축소

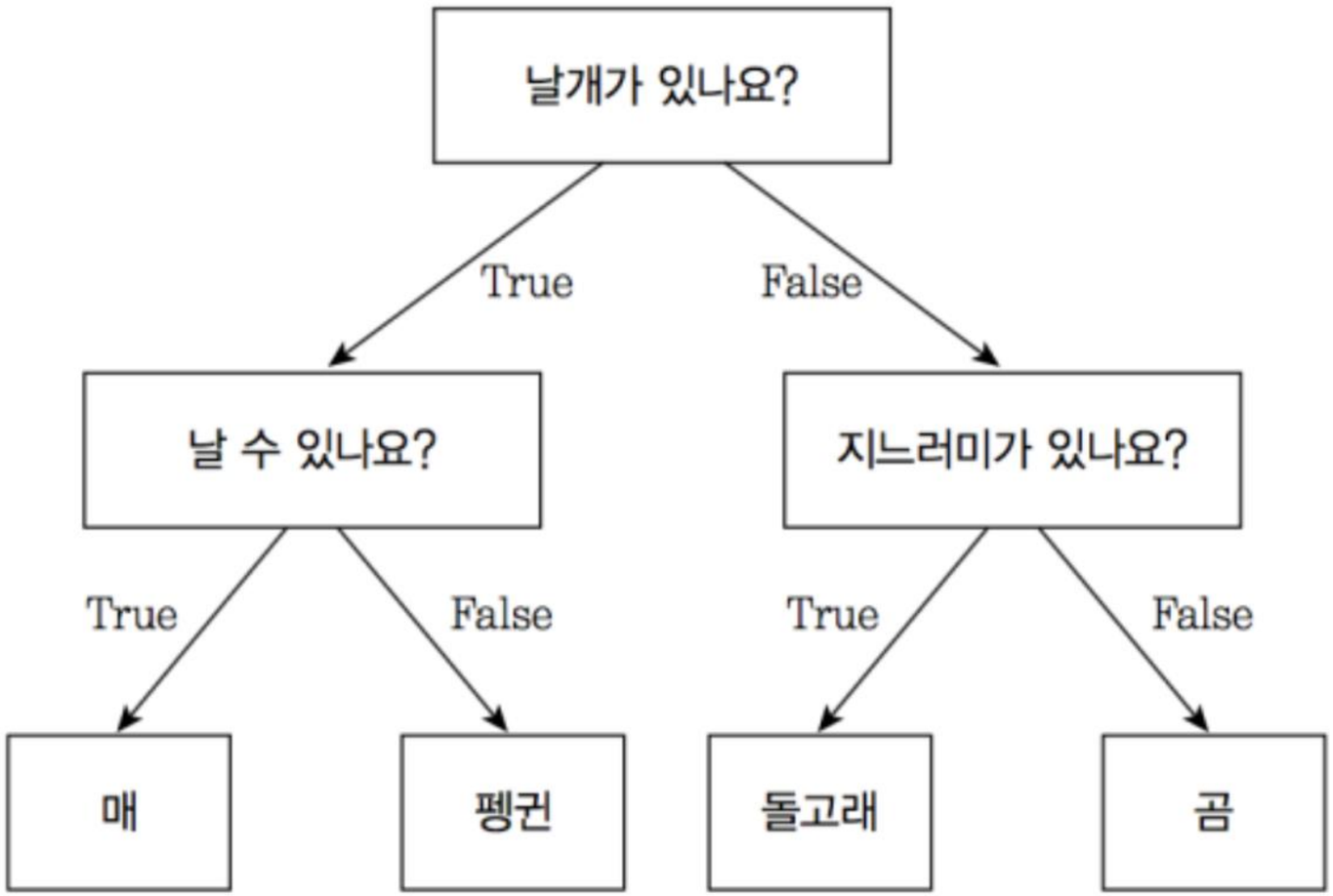
강화 학습

- Reinforcement Learning
- 샘플 입력마다 정답(label)이 있지 않지만 최종적으로 달성해야 할 목표를 알려주고 보상을 통해 학습 방법을 학습시킨다
- 게임과 같이 룰이 있고 시뮬레이션이 가능한 경우에만 동작



결정트리

- 분류 모델
  - 하위 그룹에 가능하면 같은 종류의 샘플이 모이도록 한다
- 회귀 모델
  - 하위 그룹의 샘플의 분산(variance)이 적도록 나눈다



머신러닝 모델 종류

머신러닝 유형		알고리즘	특징
지도학습	선형 계열	선형 모델, SVM 로지스틱회귀	곱셈과 덧셈으로 점수를 구하고 이를 이용하여 회귀와 분류 예측
	신경망	MLP, CNN, RNN, Transformer Graph Network	매트릭스 연산을 기반으로 점수를 계산하며 활성화 함수 도입
	트리 계열	결정 트리, 랜덤포레스트, 그라디언트부스팅	True/False 선택을 반복하여 회귀와 분류 예측 수행. 스케일링이 필요없다
	기타	kNN, 베이즈	특성 공간상의 거리를 기준, 또는 조건부 확률을 기준으로 예측
비지도학습	클러스터링	k-means, DBSCAN	특성 공간상 거리 또는 유사도를 기준으로 샘플을 그루핑
	데이터 변환	스케일링, 로그변환, 카테고리 인코딩	모델의 성능을 높이 위한 데이터 전처리
	차원 축소	PCA, t-SNE	계산량을 줄이고 모델 성능을 향상, 또는 시각화를 위한 차원 축소

제조 데이터처리