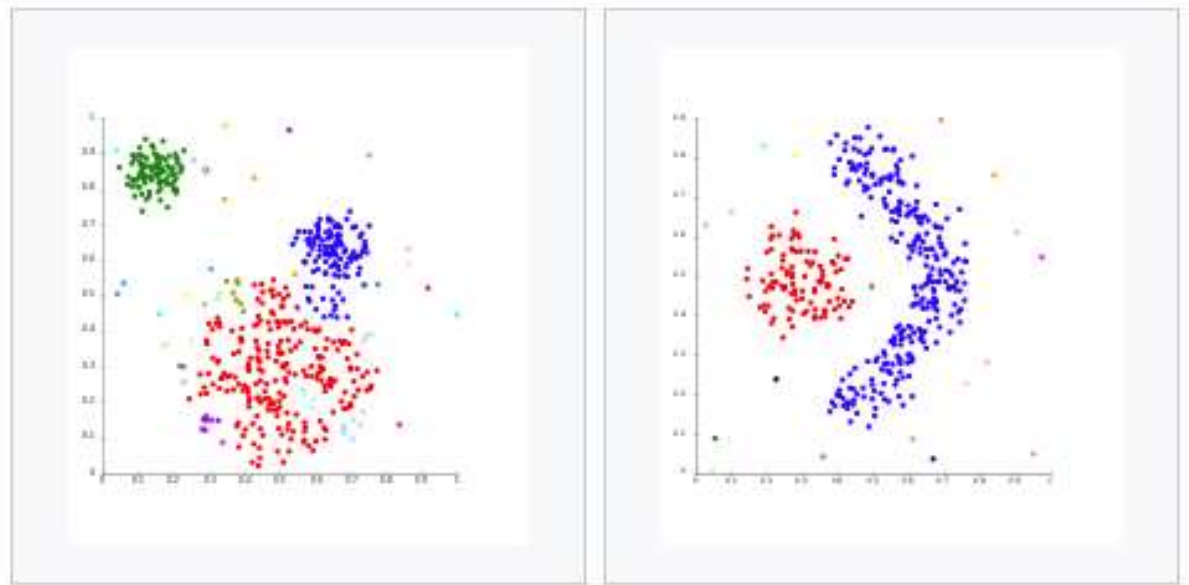


Unsupervised learning

- Clustering

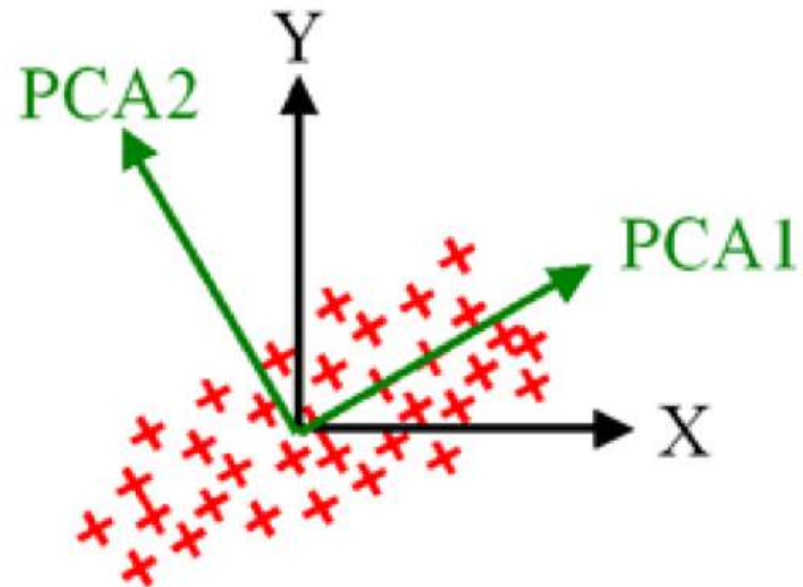
- ✓ 데이터를 요약하는 소수의 프로토 타입을 찾는 것
- ✓ 대용량 사진 데이터에서 풍경 사진, 강아지 사진, 바다 사진 등을 묶기
- ✓ 사용자의 활동 패턴을 몇 가지 유사한 행동으로 묶기
- ✓ 특정암을 유사한 유전자 발현 값이나 약물 반응성에 따라 하위 그룹으로 묶기
- ✓ ...



Unsupervised learning

- Subspace estimation

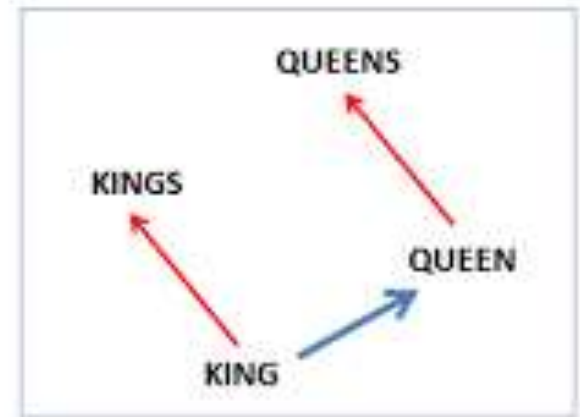
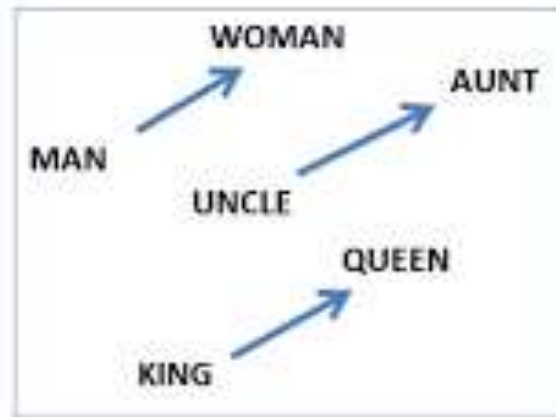
- ✓ 데이터와 관련된 특성을 정확하게 표현하는 소수의 변수를 찾는 것
- ✓ 가령 공의 궤적은 공의 속도, 직경, 그리고 질량 3가지로 설명이 가능함
- ✓ 옷을 입을 때 인체를 정확하게 표현하는 소수의 변수들 (허리, 어깨 사이즈...)
- ✓ Principal component analysis, auto-encoder
- ✓ ...



Unsupervised learning

- Representation learning

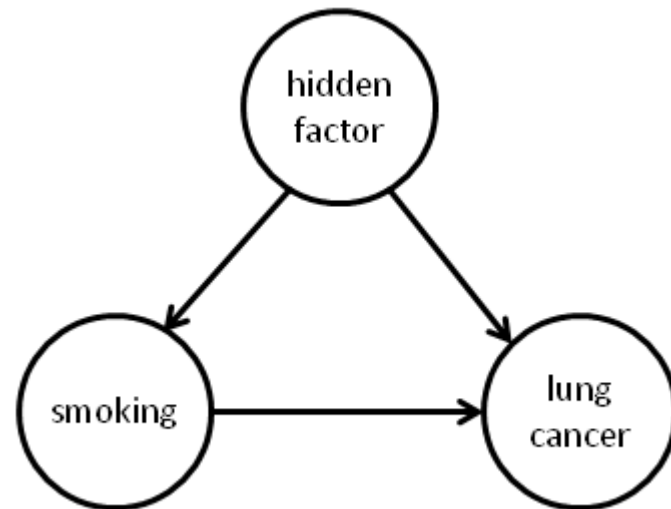
- ✓ 유클리드 공간 (R^n) 위의 특정 오브젝트를 특정 심볼과 잘 매칭시키는 것
- ✓ 주로 엔티티와 그것의 관계를 설명하는 데 활용
- ✓ 로마 - 이탈리아 + 프랑스 = 파리



(Mikolov et al., NAACL HLT, 2013)

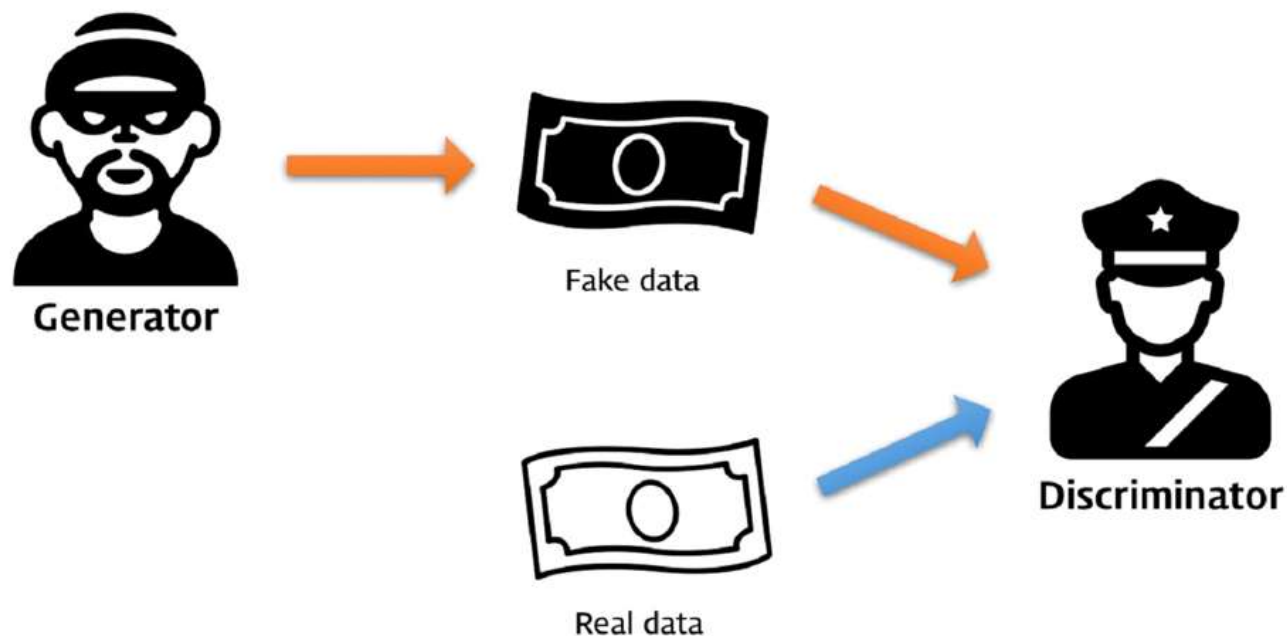
Unsupervised learning

- Causality and probabilistic graphical models
 - ✓ 우리가 관찰한 많은 데이터의 근본 원인에 대한 설명을 찾는 것
 - ✓ 집 값, 범죄, 교육, 임금과 같은 인구통계학 데이터가 어떻게 연관 있는 지
 - ✓ 유전자 그룹내에서 어떤 유전자의 발현이 어떤 질병의 발병에 영향을 주는 지



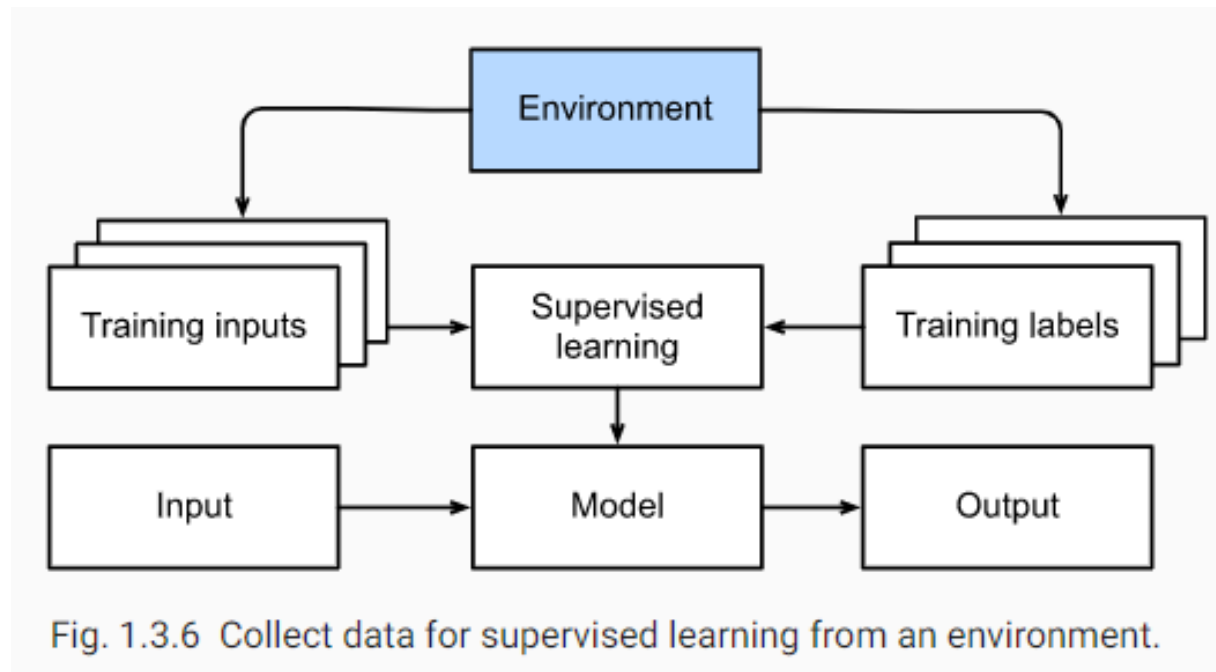
Unsupervised learning

- Generative adversarial networks (GANs)
 - ✓ 이미지와 오디오 같이 복잡한 데이터를 합성하는 방식 제공
 - ✓ 실제와 가짜 데이터가 같은 지 틀린 지 체크하는 방식의 메커니즘으로 작동



Interacting with an Environment

- 우리가 관심 가지지 않았던 것들
 - ✓ 데이터가 실제로 어디서 오는 지
 - ✓ ML 모델이 무언가 생성한 결과가 실제로 어떤 일을 일으키는 지
 - ✓ 우리는 보통 데이터를 미리 수집하고 환경과 격리된 상태로 학습을 함



Interacting with an Environment

- Offline learning
 - ✓ 외적으로 주의를 분산시키는 것에서 격리되어 특정 태스크에 대해서만 고민
 - ✓ 대신 문제 정의를 하는 것에 있어서 다소 제한적이게 됨
 - ✓ 단순한 예측 문제가 아니라 지능형 에이전트를 꿈꾼다면?
 - ✓ 어떤 행동을 선택할 지
 - ✓ 그 행동이 어떻게 환경에 영향을 주는 지
 - ✓ 영향을 준 환경의 관찰 값이 어떻게 또 영향을 받는 지

Interacting with an Environment

- Intelligent agent
 - ✓ 환경까지 고려한다면 문제정의가 완전히 새롭게 됨
 - ✓ 이전에 했던 것을 기억하는 지
 - ✓ 우리를 도우려는 지 / 견제하려는 지
 - ✓ 아니면 많은 경우에 신경을 쓰지 않는 지
 - ✓ 변화하는 지
 - ✓ 미래 데이터는 과거 데이터와 유사한 가?
 - ✓ 패턴이 시간에 따라 자연스럽게 변하는 지
 - ✓ 아니면 특정 응답에 따라 변하는 지

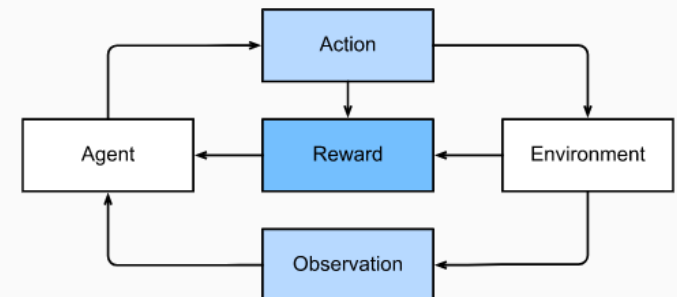


Fig. 1.3.7 The interaction between reinforcement learning and an environment.

Roots

- 근간이 되는 이론 및 연구들
 - ✓ 베르누이 / 가우시안 분포
 - ✓ 통계적 추정, 이상치
 - ✓ Fisher가 발표한 연구들 (Linear Discriminant Analysis, Iris 데이터 ...)
 - ✓ 새넌의 정보 이론 연구들
 - ✓ 튜링 테스트
 - ✓ 뇌과학과 심리학
 - ✓ 생명과학

Roots

- 이런 관심에도 불구하고...
 - ✓ 2005년까지 AI winter, 왜?
 - ✓ 신경망 학습은 너무 비싼 연산
 - ✓ 메모리에 비해서 계산 성능이 따라오지 못함
 - ✓ MNIST (6만)이 크다고 느껴질 정도로 데이터가 상대적으로 적었음
 - ✓ 상대적으로 적은 데이터와 계산 성능으로 잘 동작하는 모델들이 존재
 - ✓ SVM
 - ✓ Decision Tree
 - ✓ ...

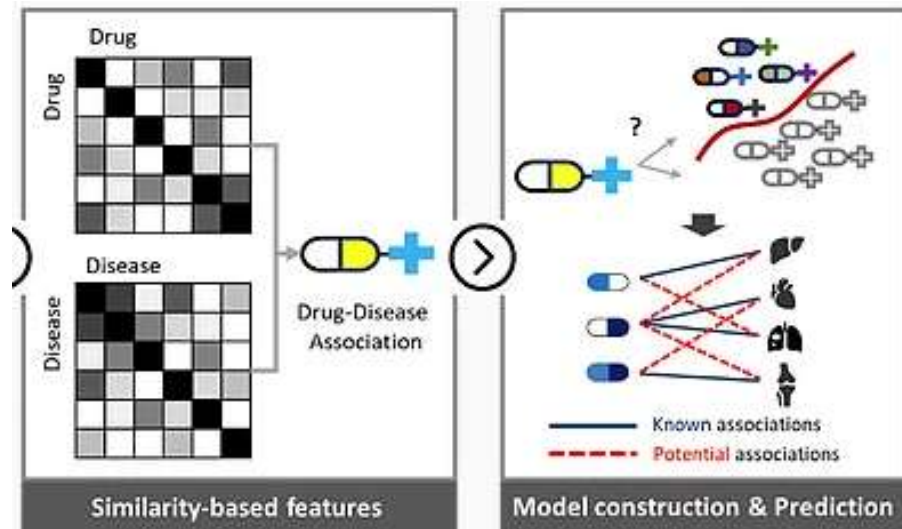
The Road to Deep Learning

- 딥러닝이 대세가 되게 된 요인들
 - ✓ 수억 명의 온라인 유저 데이터
 - ✓ 싸고 고품질의 센서 데이터
 - ✓ 저렴한 데이터 저장 및 계산 성능 (GPU 활용)

Table 1.5.1 Dataset vs. computer memory and computational power

Decade	Dataset	Memory	Floating Point Calculations per Second
1970	100 (Iris)	1 KB	100 KF (Intel 8080)
1980	1 K (House prices in Boston)	100 KB	1 MF (Intel 80186)
1990	10 K (optical character recognition)	10 MB	10 MF (Intel 80486)
2000	10 M (web pages)	100 MB	1 GF (Intel Core)
2010	10 G (advertising)	1 GB	1 TF (Nvidia C2050)
2020	1 T (social network)	100 GB	1 PF (Nvidia DGX-2)

Success stories



Summary

- ✓ 머신러닝 -> 특정 태스크의 성능을 경험 (데이터)을 활용해 향상 시키는 것
- ✓ 데이터를 잘 표현할 수 있는 적절한 방법을 찾는 것
- ✓ 데이터의 볼륨, 저렴한 센서, 계산 성능 (GPU)의 향상 -> 딥러닝 트리거
- ✓ 전체 시스템의 최적화가 좋은 퍼포먼스의 열쇠

Thanks