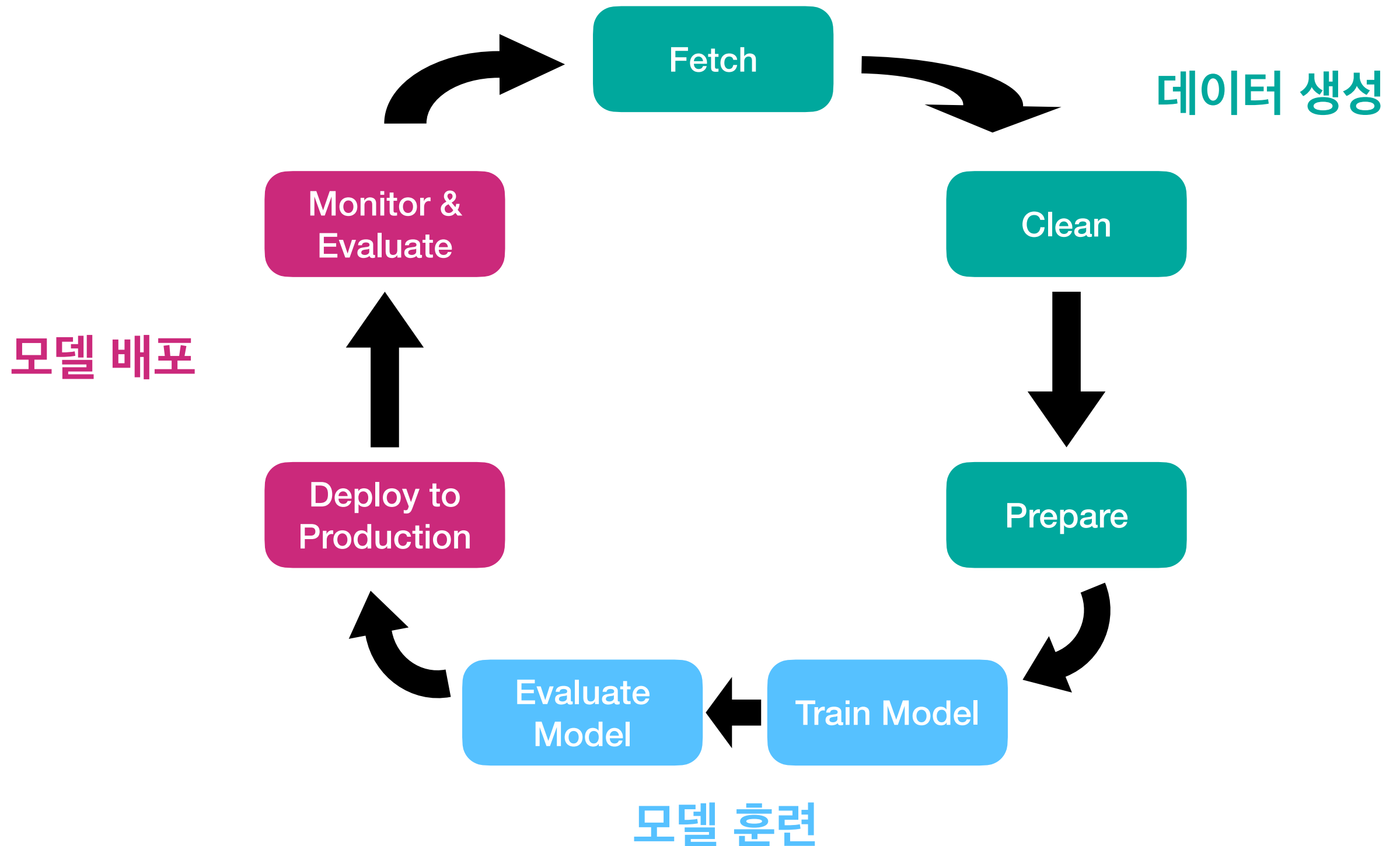


BIG DATA ANALYTICS

WEEK-04 | Data

Yonsei University
Jungwon Seo

머신러닝 Cycle



What is Data?

- 데이터란 데이터 객체의 모음 및 속성 (attribute) 들
- 속성은 객체의 특성
 - 예 : 사람의 눈 색깔, 온도 등
 - variable, field, feature라고도 불림
- 속성의 모음은 객체를 설명
 - 객체 동의어: 레코드, 포인트, 사례, 샘플, 엔터티 또는 인스턴스

속성 값

- 속성 값은 속성에 지정된 숫자 또는 기호
- 속성과 속성 값의 구별
 - 동일한 속성은 다른 속성 값에 매핑 가능
 - 예 : 높이는 피트 또는 미터로 측정 가능
 - 다른 속성을 동일한 값 세트에 매핑 가능
 - 예 : ID 및 연령의 속성 값은 정수
 - ID에는 제한이 없지만 연령은 최대 값과 최소값 존재

속성 종류

- Nominal: 명목자료
 - 비교 X, 차이 O
 - 예: 우편번호, 주민번호 뒷자리, 피부색, 성별
- Ordinal: 서열자료
 - 순서가 있는 명목자료
 - 예: 점수 (10점 만점에 몇점), 학년, 키 (크다/보통/작다)
- Interval: 구간자료
 - True Zero가 존재하지 않음
 - 0도가 가장 낮은 온도인가? 1월 1일이 가장 이른 날짜인가?
 - 예: 달력 날짜, 온도 (섭씨/화씨)
- Ratio: 비율자료
 - True Zero가 존재
 - 절대 온도 0도가 가장 낮은 온도, 0M는 길이가 없음, 0초, 0번 등
 - 예: 절대온도, 길이, 시간, 수

이산 속성 및 연속 속성

- 이산 속성

- 한정적이거나 셀 수 있을 정도로 무한대의 값 집합만 있음
- 예 : 우편 번호, 횟수, 문서 모음에서 단어의 수
- 보통 정수 변수로 표시
- 참고 : 이진 속성은 특수한 경우의 이산 속성

- 연속속성

- 속성 값으로 실수를 가짐
- 예 : 온도, 높이 또는 무게.
- 하지만 유한한 갯수의 숫자로 측정되거나 표현 될 수밖에 없음
- 일반적으로 소수점을 포함한 수로 표현됨

데이터 세트 유형

- Record : 독립적으로 존재 가능함
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph: 다른 데이터와의 관계가 중요
 - World Wide Web
 - Molecular Structures
- Ordered: 위치/시간/순서가 데이터의 의미를 부여함
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

정형 데이터의 중요한 특성

- Dimensionality: 차원
 - Curse of Dimensionality : 차원의 저주
- Sparsity: 희소성
 - 유효한 부분만 카운팅
- Resolution: 해상도
 - 배율에 따라 패턴이 달라짐

Record Data

- 레코드 묶음으로 구성되는 데이터
- 각 레코드는 고정된 속성(attribute) 세트로 구성

온도	조망	습도	바람	테니스
보통	맑음	80	No	Yes
더움	맑음	75	Yes	No
더움	흐림	77	No	Yes
시원함	비	70	No	Yes
시원함	흐림	72	Yes	Yes
보통	맑음	77	No	No
시원함	맑음	70	No	Yes
보통	비	69	No	Yes
보통	맑음	65	Yes	Yes
보통	흐림	77	Yes	Yes
더움	흐림	74	No	Yes
보통	비	77	Yes	No
시원함	비	73	Yes	No

Data Matrix

- 데이터들이 숫자 속성으로 이루어져 있다면, 다차원 공간에서 점으로 표현 가능
- 여기서 각 차원은 고유 한 특성을 나타냅니다.
- $m \times n$ 행렬로 표시 (m : 데이터의 수, n : 속성의 수)

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- 각 문서는 단어 벡터로 표현
- 각 단어는 벡터의 구성 요소 (속성)
- 각 구성 요소의 값은 해당 단어가 문서에서 발생하는 횟수

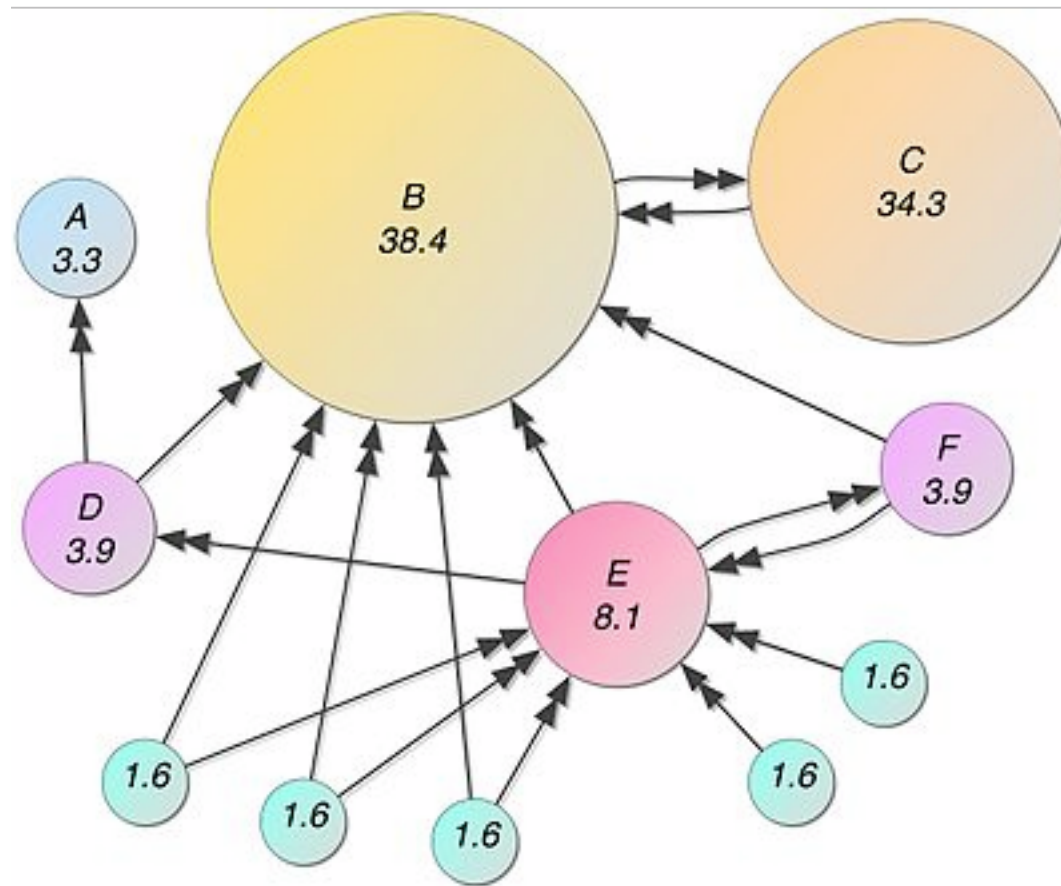
단어	문서 1	문서 2	문서 3	단어빈도
텍스트	1	0	0	1
마이닝	1	0	0	1
비정형	1	1	0	2
데이터	1	2	0	3
다루다	1	0	0	1
정형	0	1	0	1
복잡하다	0	1	0	1
어렵다	0	1	0	1
오늘	0	0	1	1
단어주머니	0	0	1	1
배우다	0	0	1	1

Transaction Data

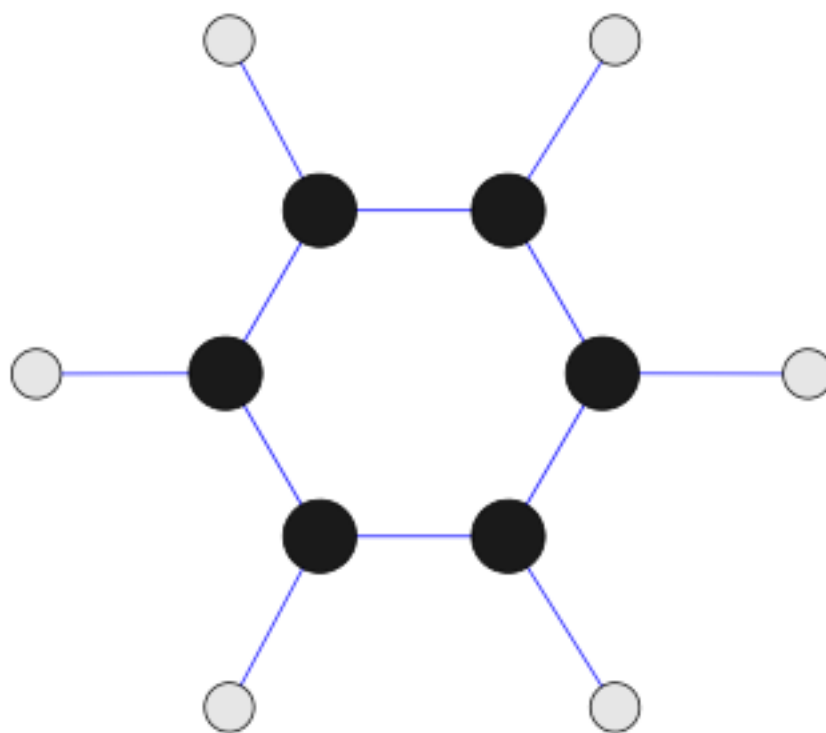
- 각 레코드 (트랜잭션)에는 일련의 항목이 포함되는 특수한 유형의 레코드 데이터.
- 예를들어, 마트에서 한번에 결제하는 데이터셋안에는, 구매한 여러 물품들이 포함

TID	물건
1	식빵, 콜라, 우유
2	맥주, 빵
3	맥주, 콜라, 기저귀, 우유
4	맥주, 빵, 기저귀, 우유
5	콜라, 기저귀, 우유

Graph Data

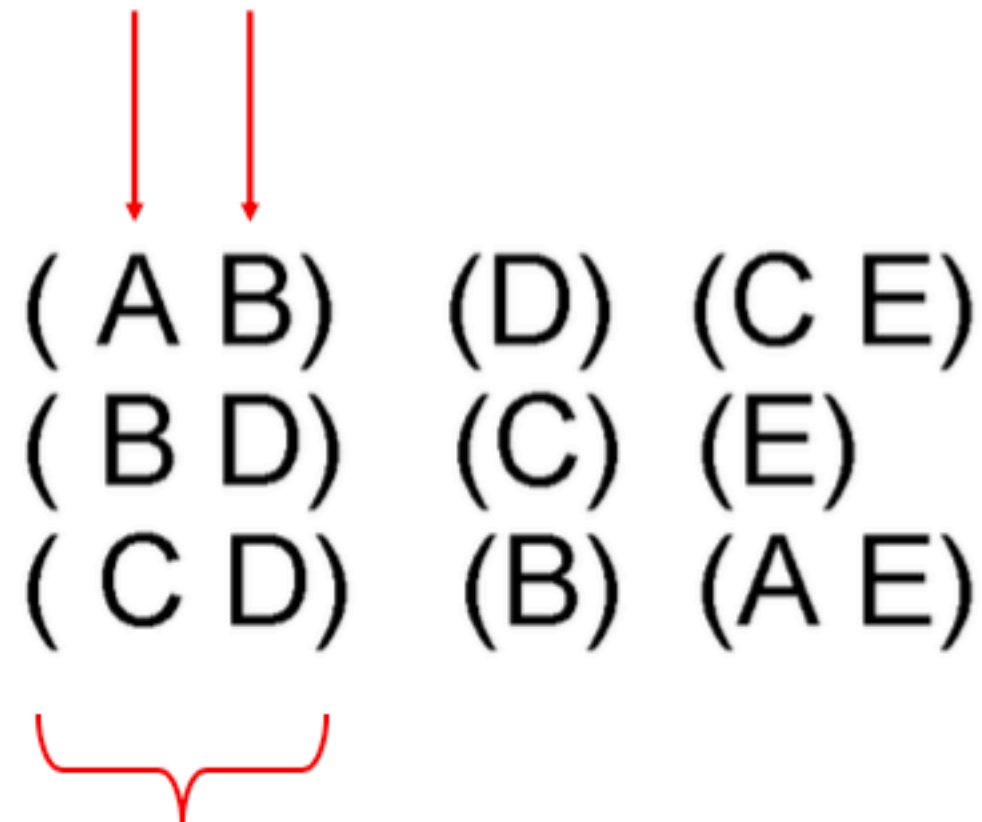


Chemical Structures



Sequential data

Items/Events



(A B)	(D)	(C E)
(B D)	(C)	(E)
(C D)	(B)	(A E)

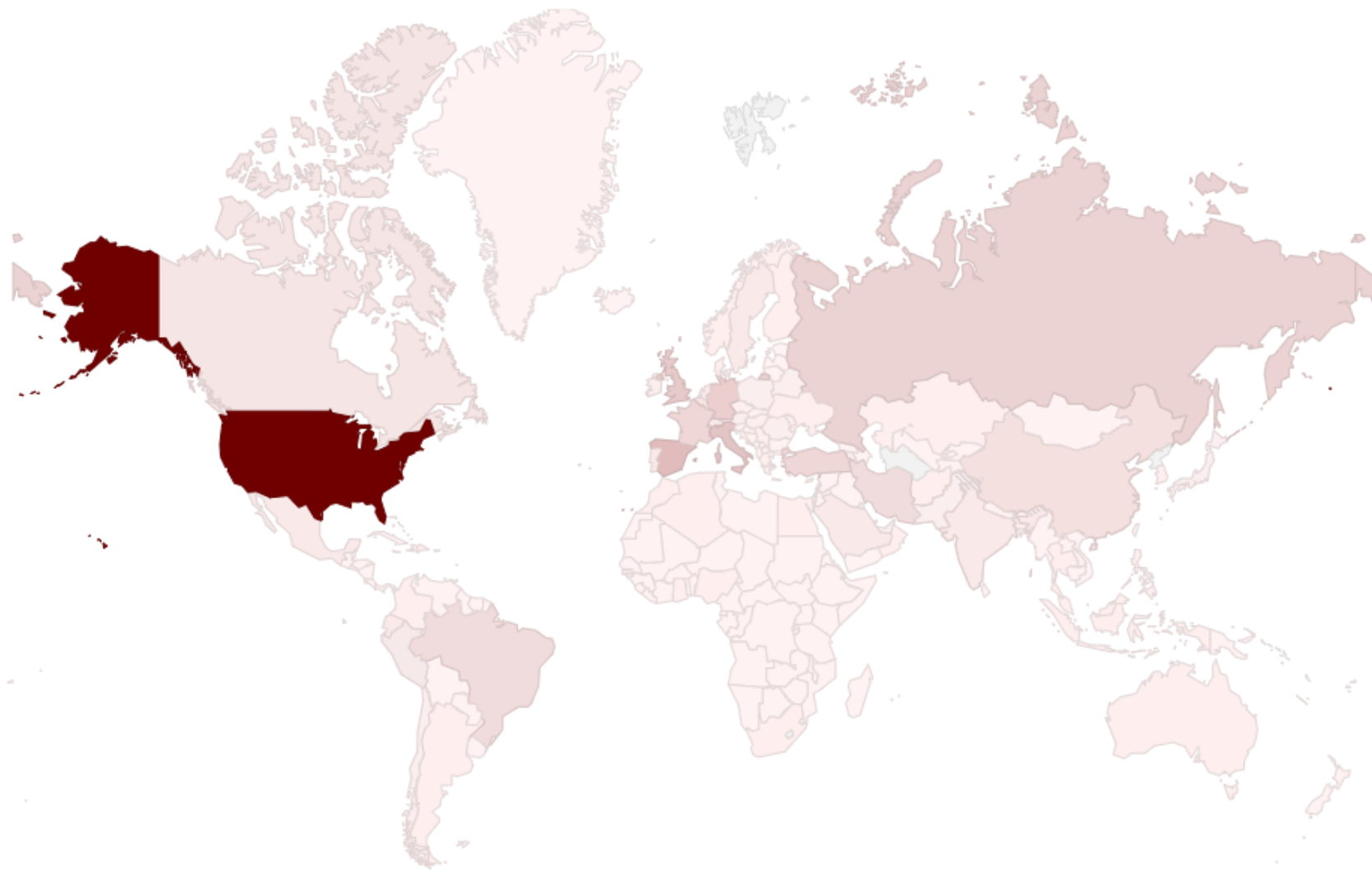
**An element of
the sequence**

Gene Sequence

**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Spatio-temporal Data

국가별 현황



① 용어 설명

* 발생률: 100만명당 발생률 (=확진자/인구수*1,000,000)

오늘

어제

사망률, 완치율, 발생률*



Data Quality

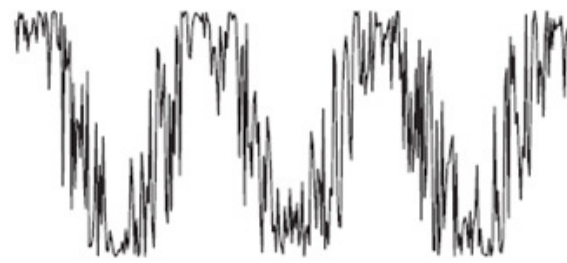
- 어떤 종류의 데이터 품질 문제가 있나?
- 데이터 문제를 어떻게 감지 할 수 있나?
- 이 문제들에 대해 무엇을 할 수 있나?
- 데이터 품질 문제의 예 :
 - Noise 와 outliers
 - Missing value
 - Duplicate data

Noise

- 잘못된 관측 또는 무작위적 오류
- 예) 옛날 전화기 잡음, 티비 Noise



(a) Time series.



(b) Time series with noise.

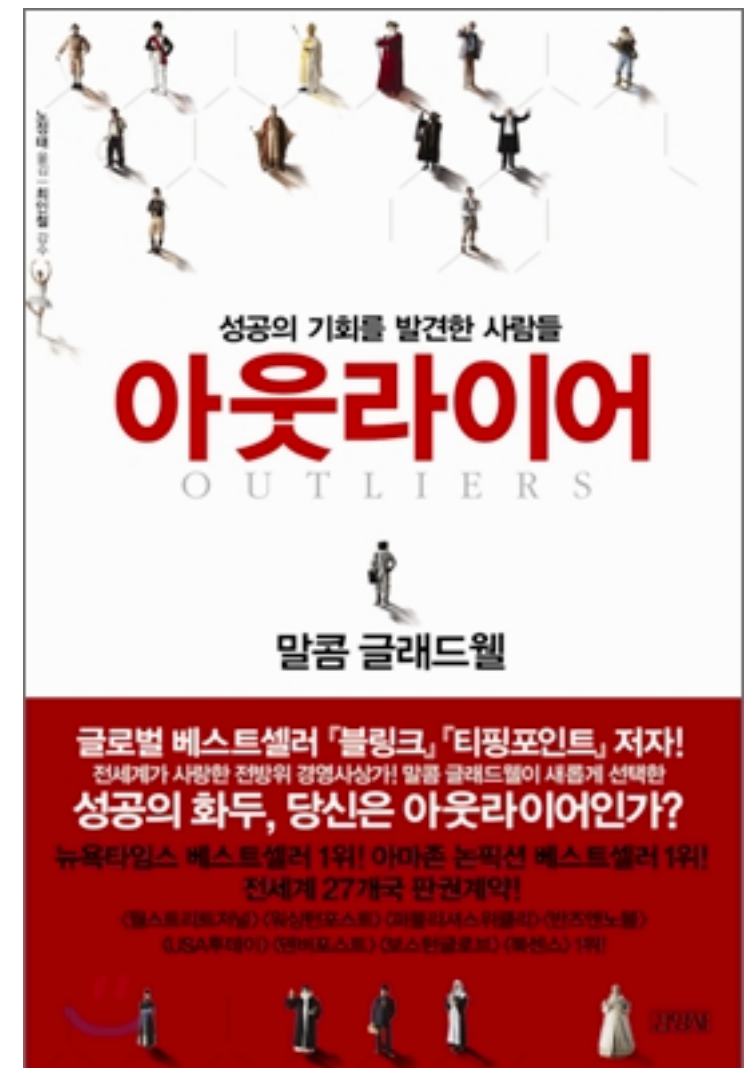
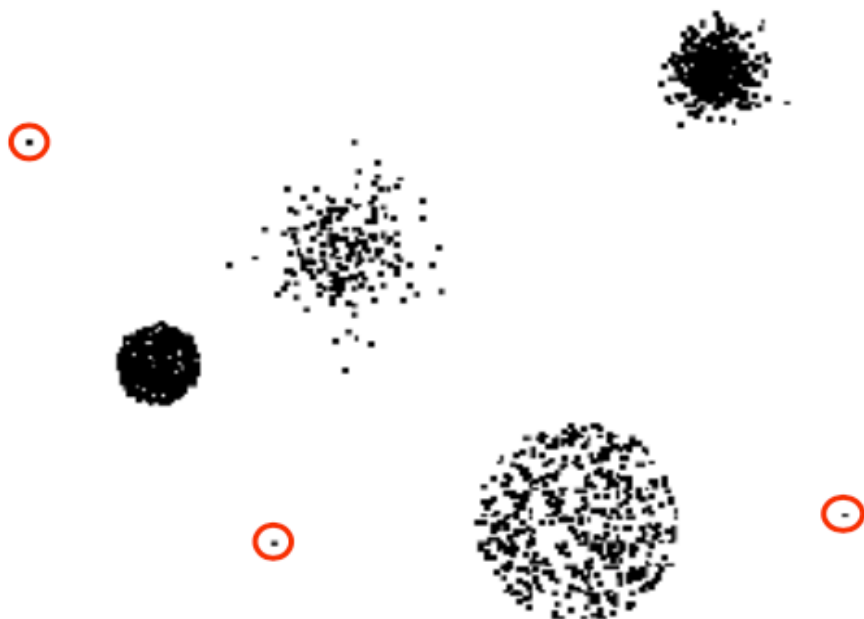


(그림 출처: Pang-Ning Tan et al, Introduction to Data Mining, Addison-Wesely, 2005)

https://commons.wikimedia.org/wiki/File:TV_noise.jpg

Outliers

- 특이 치는 데이터 집합의 다른 데이터 개체와 상당히 다른 특성을 가진 데이터 개체
- 예) 이메일 해외접속 알림



Noise와 Outlier의 차이는?

Missing Values

- 결측값이 생기는 이유
 - 정보가 항상 수집되지는 않음
 - 예 : 사람들은 나이와 체중을 숨기려고 함)
 - 모든 경우에 속성을 적용 할 수있는 것은 아님
 - 예 : 연간 소득은 어린이에게는 적용되지 않음)
- 결측값 처리
 - 데이터 객체 제거
 - 결측값 추정
 - 분석 중 결측값 무시
 - 가능한 모든 값으로 대체 (확률에 따라 가중치가 부여됨)

Duplicate Data

- 데이터 세트는 서로 중복되거나 거의 중복되는 데이터 객체를 포함 할 가능성이 있음.
- 출처가 다른 데이터를 병합 할 때 발생하는 주요 문제
 - 예 :이메일 주소가 여러개인 같은 사람
- 데이터 청소
 - 중복 데이터 문제를 처리하는 프로세스

Similarity and Dissimilarity

- 유사성 (Similarity)
 - 두 데이터 객체의 유사성에 대한 수치 측정
 - 객체가 더 비슷할 때 더 높음
 - 보통 $[0,1]$ 범위
- 비유사성 (Dissimilarity)
 - 두 데이터 객체가 얼마나 다른지에 대한 수치 측정
 - 물체가 더 비슷할 때 더 낮아짐
 - 최솟값은 보통 0 (같은 객체)
 - 상한이 다름 (얼마나 다를지는 보장되지 않음)

Similarity/Dissimilarity for Simple Attributes

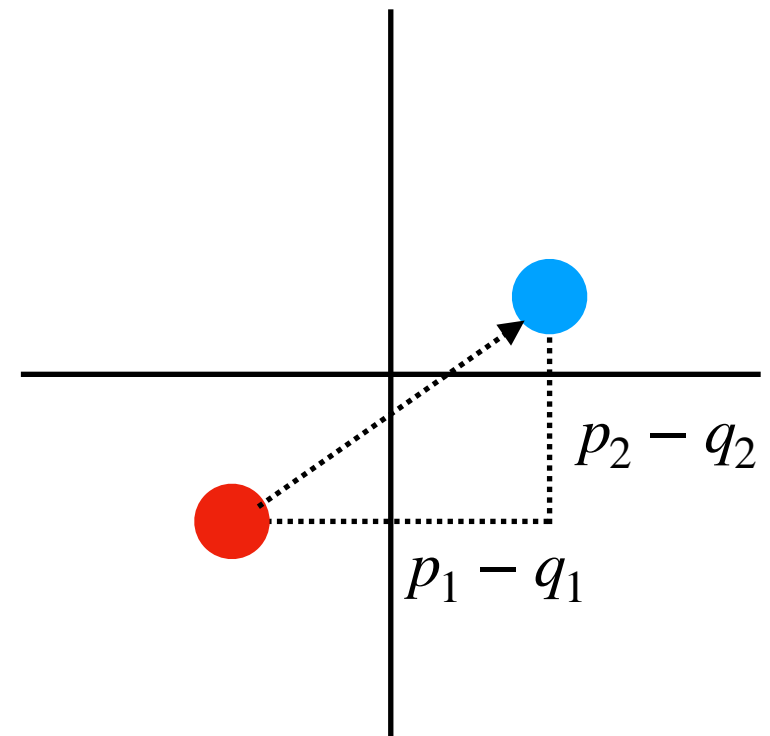
- p와 q는 두 데이터 객체의 속성 값

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

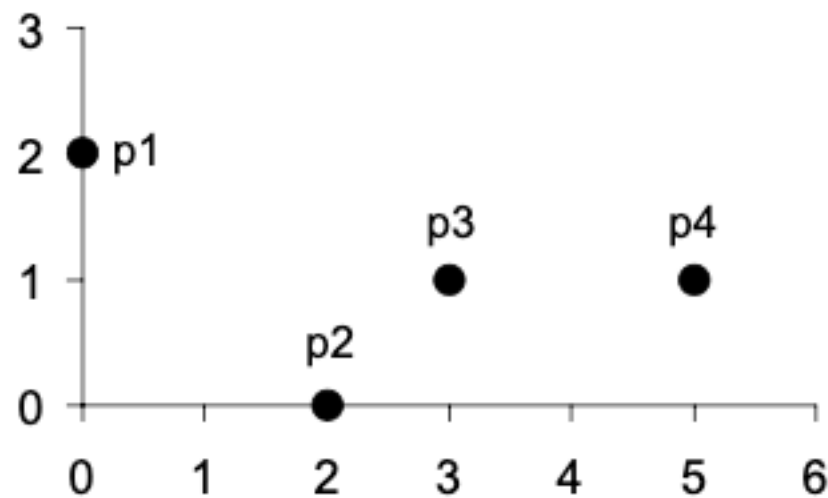
Euclidean Distance

- 유클리드 거리
- 여기서 n 은 차원 (속성)의 수이고 p_k 및 q_k 는 p, q 의 k 번째 속성 (feature)

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$



Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

$$\mathbf{dist} = \left(\sum_{k=1}^n | \mathbf{p}_k - \mathbf{q}_k |^r \right)^{\frac{1}{r}}$$

- Minkowski Distance는 유클리드 거리의 일반화 형태
- r 은 매개변수, 여기서 n 은 차원 (속성)의 수이고 p_k 및 q_k 는 p, q 의 k 번째 속성 (feature)

Minkowski Distance: Examples

- $r = 1$. 도시 블록 (맨해튼, 택시, L_1 표준) 거리
- $r = 2$. 유클리드 거리
- $r \rightarrow \infty$. "최고" (L_{\max} 표준, L_∞ 표준) 거리

Properties of a Distance Function

- 유클리드 거리와 같은 거리에는 다음과 같은 속성을 포함
 - $d(p,q) \geq 0$ for all p and q and $d(p,q)=0$ only if $p=q$.
 - $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
 - $d(p,r) \leq d(p,q) + d(q,r)$ for all points p, q , and r . (Triangle Inequality)
- 위의 조건을 만족하는 거리를 metric이라 부름

Properties of a Similarity Function

- 유사도의 속성
 - $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
 - $s(p, q) = s(q, p)$ for all p and q . (Symmetry)
 - where $s(p, q)$ is the similarity between points (data objects), p and q .

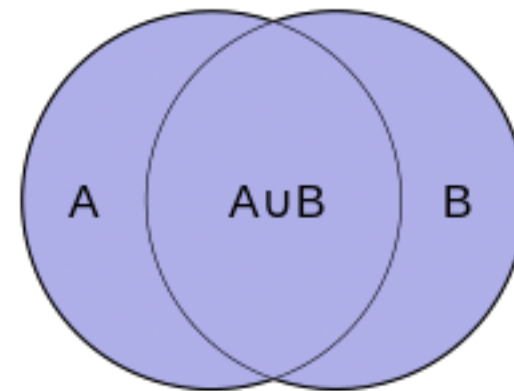
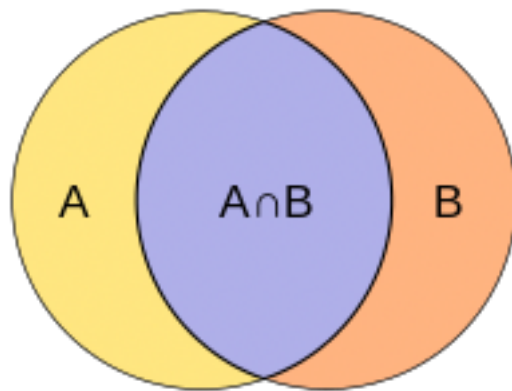
Simple Matching Coefficient (SMC)

- 일반적인 상황은 객체 p와 q에 이진(binary) 속성 만있는 것
- 다음 수량을 사용하여 유사점 계산
 - $M01$ = p가 0이고 q가 1 인 속성 수
 - $M10$ = p가 1이고 q가 0 인 속성 수
 - $M00$ = p가 0이고 q가 0 인 속성 수
 - $M11$ = p가 1이고 q가 1 인 속성 수
- Simple Matching and Jaccard Coefficients
 - $SMC = \text{일치 수} / \text{속성 수}$
 $= (M11 + M00) / (M01 + M10 + M11 + M00)$
 - $J = \text{11 개의 일치} / \text{0이 아닌 속성 값의 수}$
 $= (M11) / (M01 + M10 + M11)$

Jaccard Similarity/ Coefficient

- 범주 속성에 사용
- 집합에도 적용 가능

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Vectors (벡터)

- $n \times 1$ 행렬
 - 일반적으로 소문자로 표현
 - n 행
 - 1 열

$$y = \begin{bmatrix} 100 \\ 230 \\ \vdots \\ 530 \end{bmatrix}$$

Dot product

- 내적
- 주어진 두 벡터 a 와 b

$$\vec{a} = (a_1, a_2, a_3, \dots) \qquad \vec{b} = (b_1, b_2, b_3, \dots)$$

$$\vec{a} \cdot \vec{b} = \sum_{i=1}^n a_i b_i = a_1 b_1 + a_2 b_2 + \dots + a_n b_n$$

Cosine Similarity

- d1과 d2가 두 문서 벡터 인 경우

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

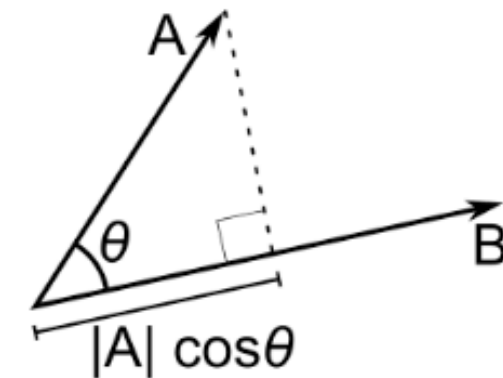
$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \cdot d_2 = 3 \cdot 1 + 2 \cdot 0 + 0 \cdot 0 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 2 = 5$$

$$\|d_1\| = (3 \cdot 3 + 2 \cdot 2 + 0 \cdot 0 + 5 \cdot 5 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 2 + 0 \cdot 0 + 0 \cdot 0)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1 \cdot 1 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 1 \cdot 1 + 0 \cdot 0 + 2 \cdot 2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$



E.O.D