

BIG DATA ANALYTICS

WEEK-05 | Data Preprocessing

**Yonsei University
Jungwon Seo**

데이터 전처리

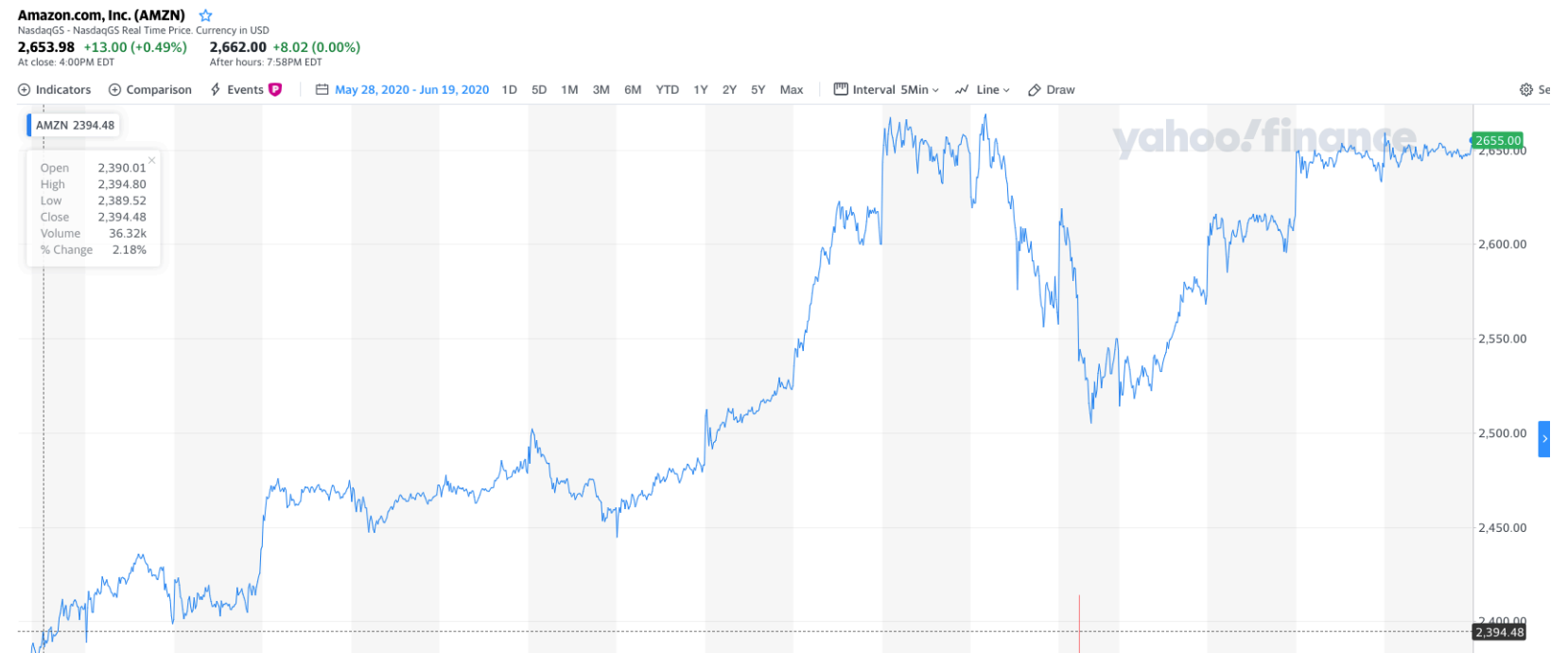
- Aggregation (집계)
- Sampling
- Dimensionality Reduction (차원축소)
- Feature subset selection
- Feature creation
- Discretization and Binarization (이진화)
- Attribute Transformation

Aggregation

- 둘 이상의 속성 (또는 객체)을 단일 속성 (또는 객체)으로 결합
- 목적
 - 데이터 축소
 - 속성 또는 객체 수 감소
 - 규모의 변화
 - 지역, 주, 국가 등으로 집계 된 도시
 - 연/월/일 데이터
 - 보다 안정적인 데이터
 - 집계 된 데이터는 변동성이 적은 경향을 보임

Aggregation Example

5분 단위



1시간 단위



Data Sampling

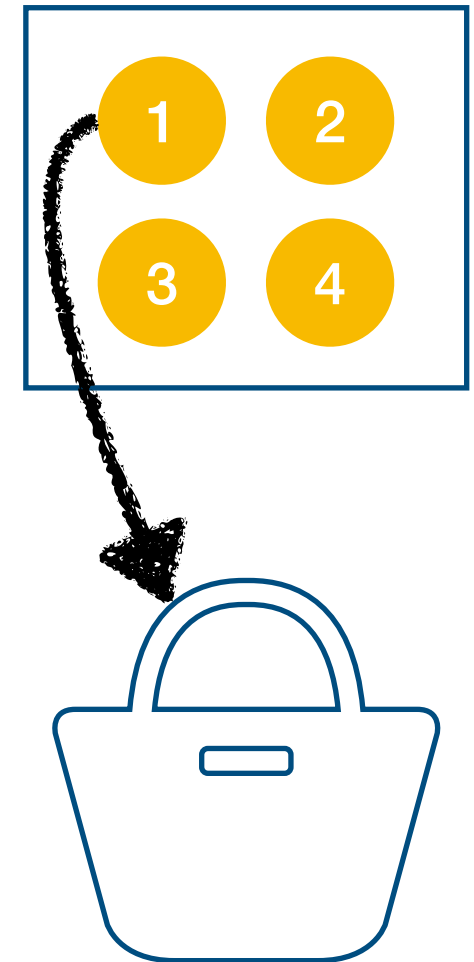
- 샘플링은 데이터 선택에 사용되는 주요 기술
 - 데이터의 예비 조사와 최종 데이터 분석에 종종 사용
- 전체 관심 데이터 세트를 얻는 것이 너무 비싸거나 시간이 많이 걸리기 때문에 통계학자는 표본을 추출
- 관심있는 전체 데이터 세트를 처리하는 것이 너무 비싸거나 시간이 많이 걸리므로 샘플링은 데이터 마이닝에 사용

How to Sample?

- 효과적인 샘플링의 핵심 원리
- 샘플을 사용하면 샘플이 대표적 일 경우 전체 데이터 세트를 사용하는 것과 비슷한 효과
- 표본이 원래 데이터 세트와 대략 동일한 속성 (관심있는)을 갖는 경우 대표 표본

Types of Sampling

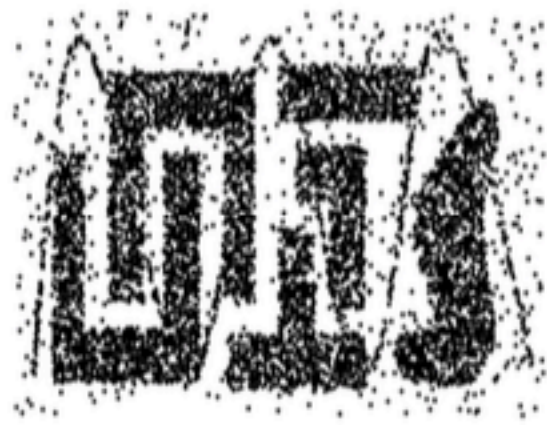
- 간단한 랜덤 샘플링
 - 특정 항목을 선택할 확률은 동일
- 교체 (replacement) 없이 샘플링
 - 각 항목을 선택하면 모집단에서 제거
- 교체 샘플링
 - 표본에 대해 선택된 개체는 모집단에서 제거되지 않음
 - 교체를 통한 샘플링에서 동일한 객체를 두 번 이상 선택가능
- 계층화 된 샘플링
 - 데이터를 여러 파티션으로 분할 후 각 파티션에서 임의의 샘플을 추출



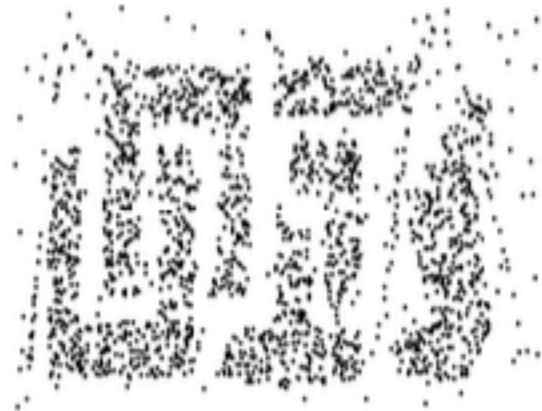
교체(replacement)의 여부에 따른 장단점은 무엇인가?

샘플 사이즈를 결정하는 합리적인 전략은?

Sample Size



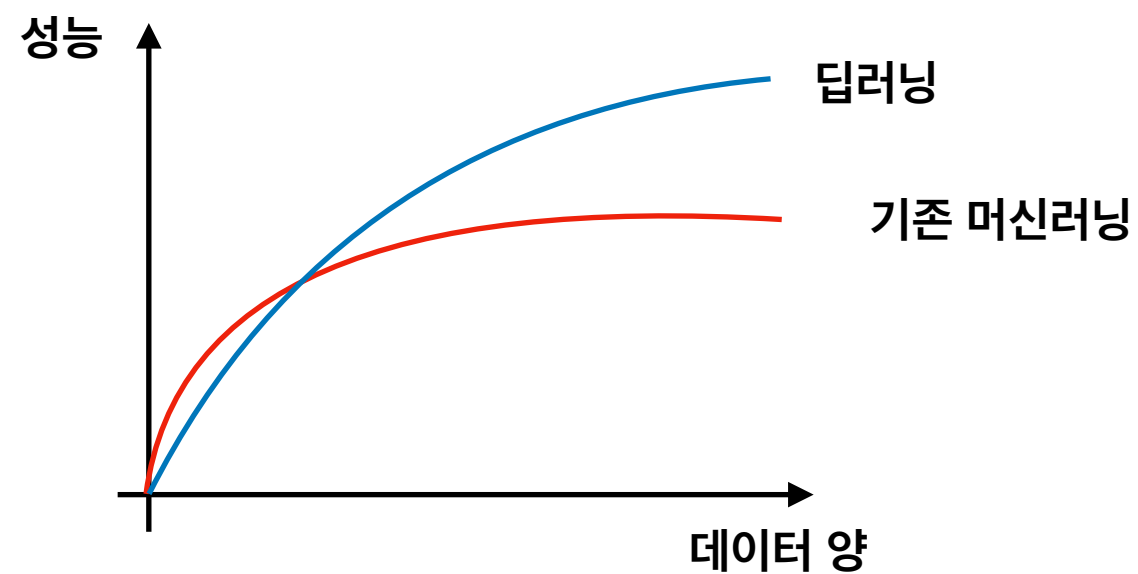
8000 points



2000 Points

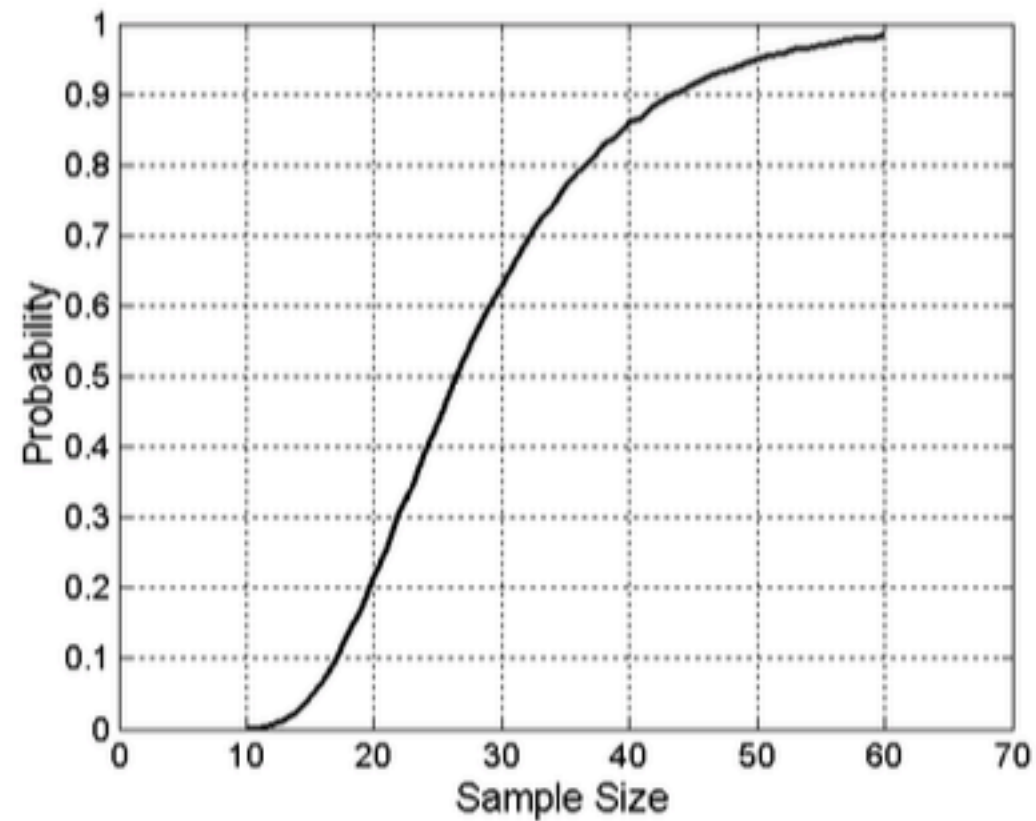


500 Points



Sample Size

- 10 개 그룹 각각에서 하나 이상의 물체를 얻는 데 필요한 표본 크기는?

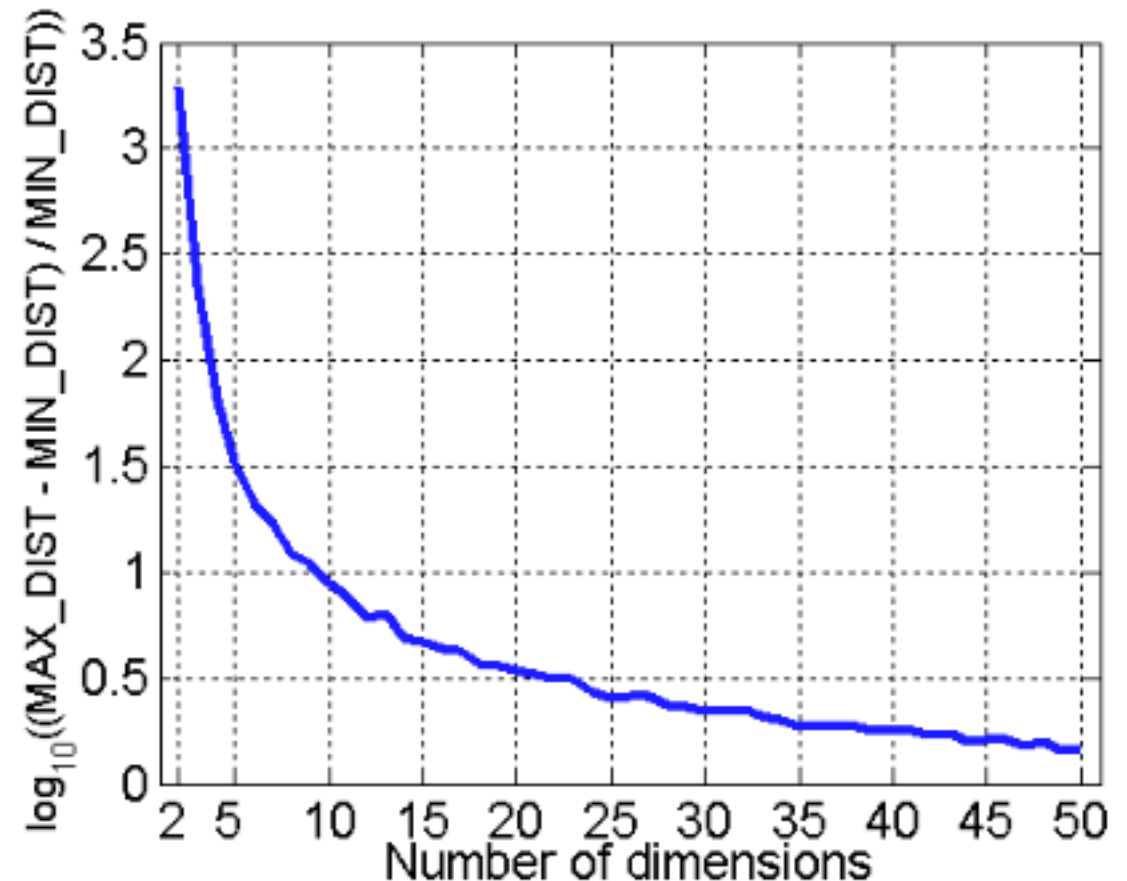


Reservoir Sampling

- 전체 데이터의 크기를 모르거나, 전체 데이터가 엄청 큰 경우 샘플링 방법은?
- 비축/저장소 샘플링
- r 샘플의 저장소를 유지
 - 처음 10 개 항목을 메모리에 보관
 - i 번째 항목이 도착하면 ($i > r$)
 - 확률 (r / i)로 새로운 아이템을 유지 (오래된 아이템을 버리고 무작위로 교체 할 것을 선택 $1/r$ 확률)
 - 확률 $1 - (r / i)$ 로 오래된 아이템을 유지

Dimensionality Reduction

- 차원 축소
- 차원의 저주
 - Curse of Dimensionality
- 차원이 증가하면 데이터가 차지하는 공간에서 데이터가 점점 희박 (sparse) 해짐
- 군집 및 이상치 탐지에서 중요한 점간의 밀도 및 거리 정의는 의미가 점점 떨어짐



- 임의로 생성된 500개의 점
- 임의의 쌍이 두점간의 최대/최소 거리의 차이를 계산

Dimensionality Reduction

- 목적

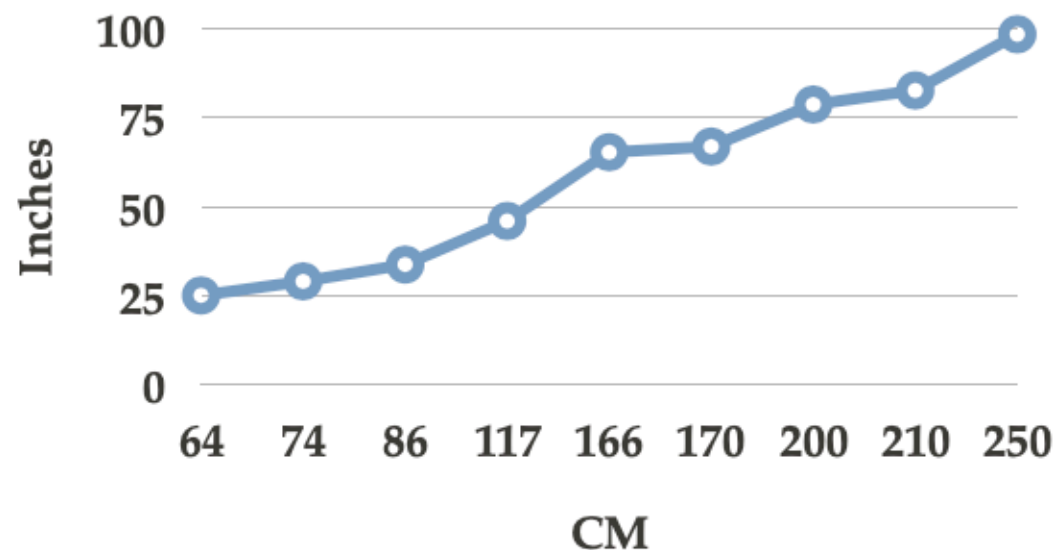
- 차원의 저주 피하기
- 데이터 마이닝 알고리즘에 필요한 시간과 메모리 양을 줄임
- 데이터를 보다 쉽게 시각화 가능
- 관련없는 기능을 제거하거나 노이즈를 줄이는 데 도움

- 기법

- Principle Component Analysis (PCA) : 주성분분석
- Singular Value Decomposition (SVD) : 특이값 분해
- Others: supervised and non-linear techniques

Motivation 1: 데이터 압축

- 데이터 마이닝 / 기계 학습 알고리즘을 더 빨리 훈련 가능
- 더 적은 스토리지 / 메모리 사용

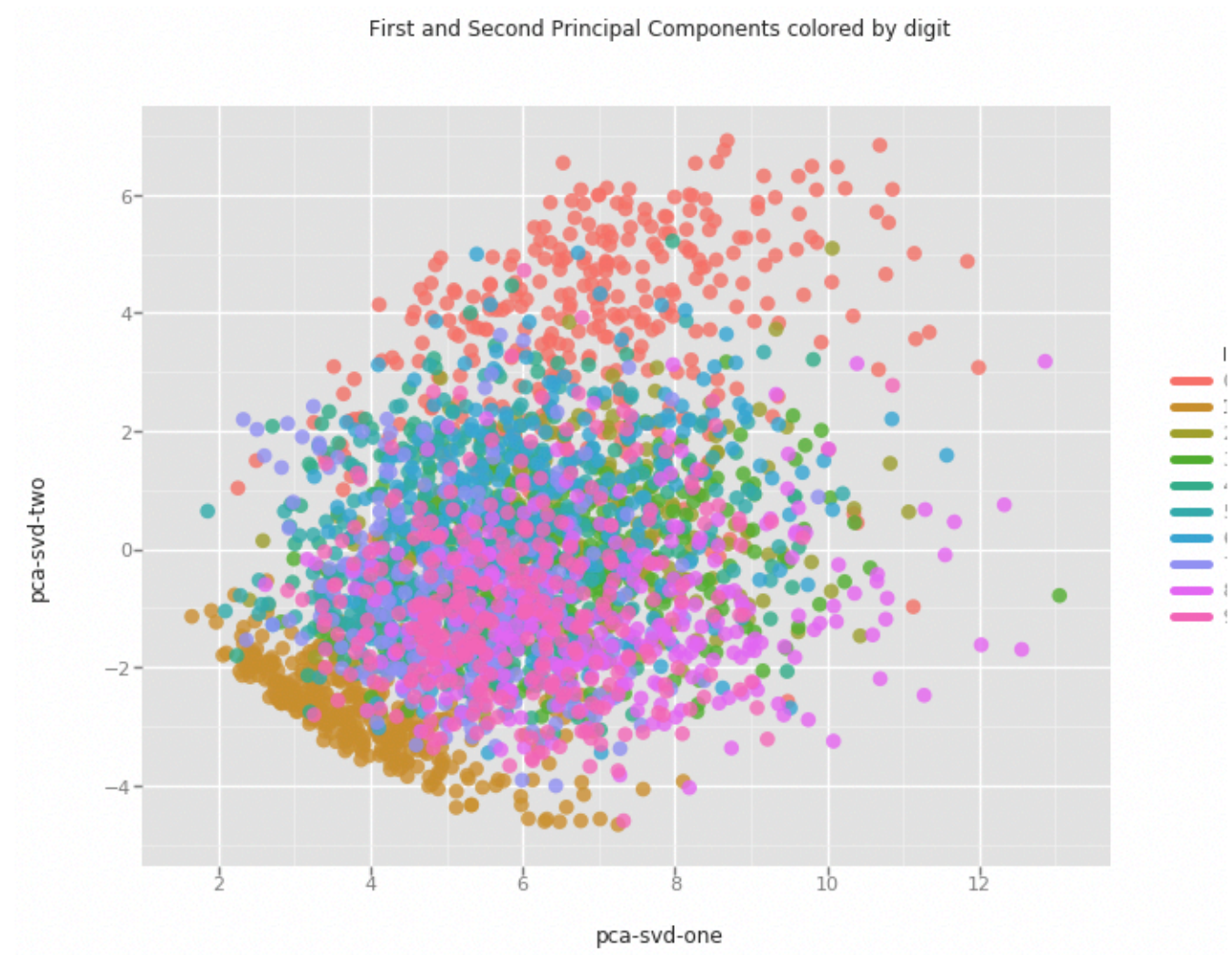
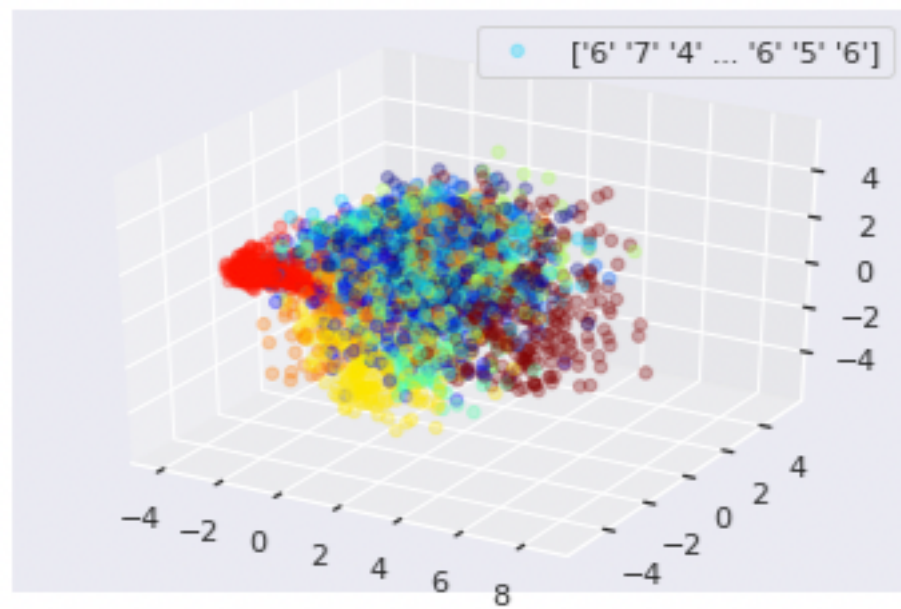


Data redundancy

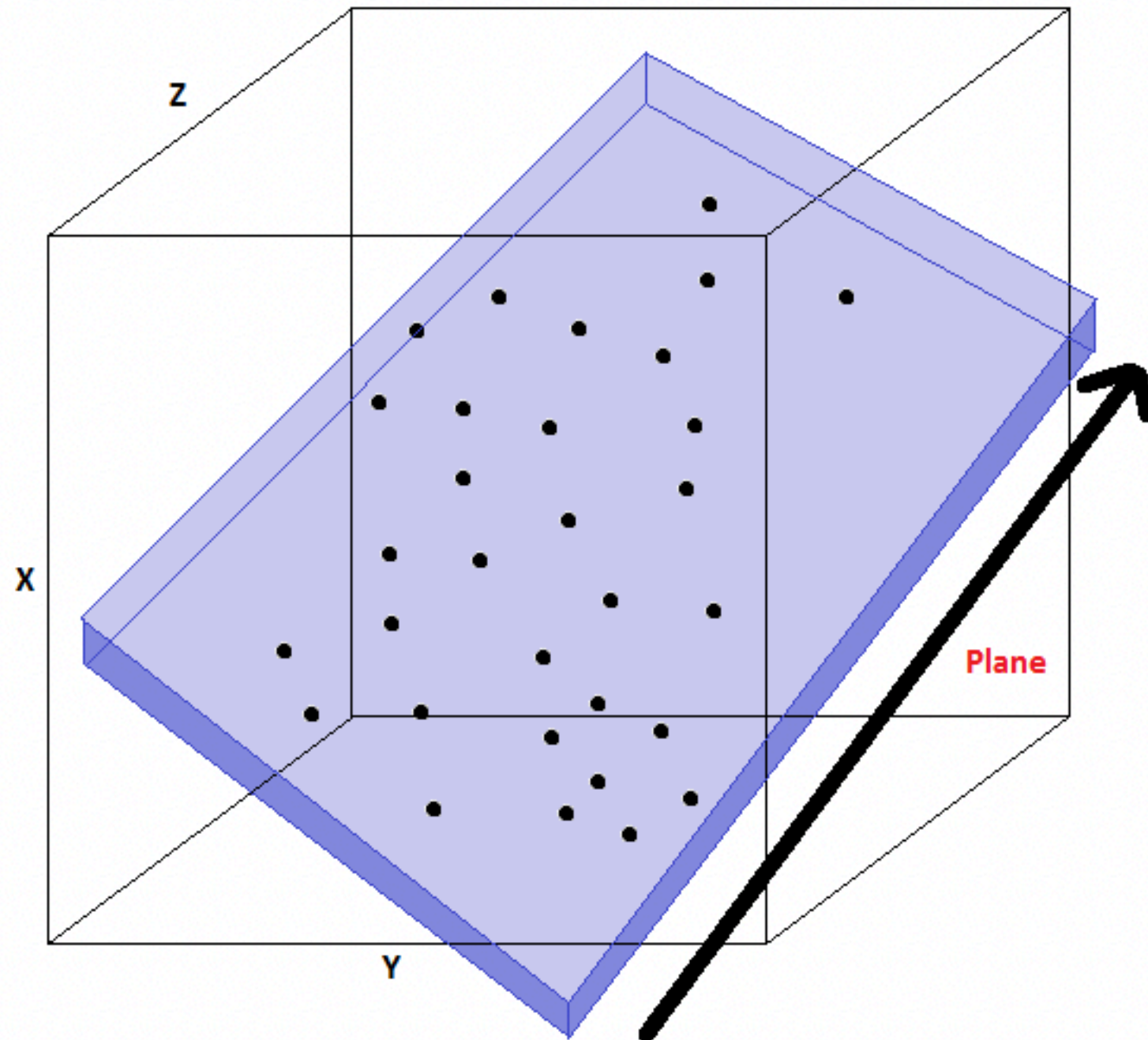
데이터 압축 예시 1



데이터 압축 예시 2



데이터 압축 예시 3

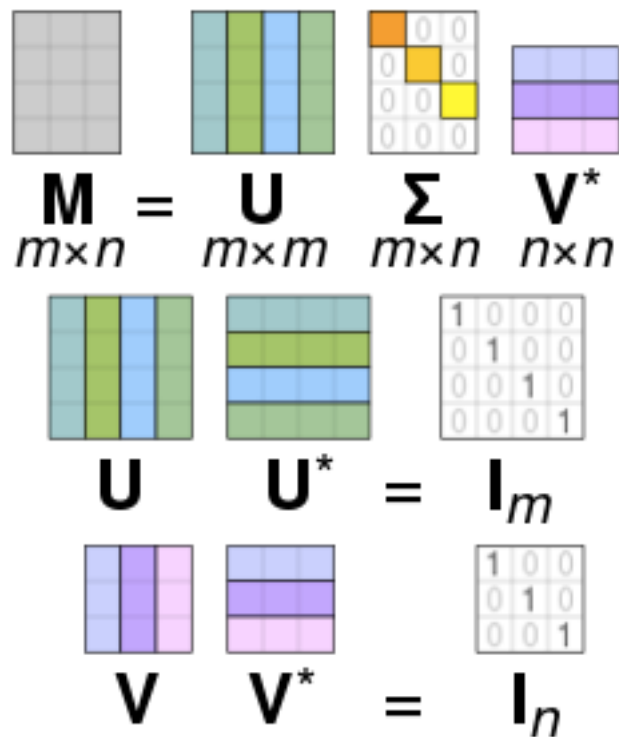


Motivation 2: 시각화

- 고차원 데이터는 시각화하기 어려움
- 인간이 시각적으로 인지할 수 있는 최대 차원은?
- 시각화의 중요성
 - 시각화를 통해 새로운 인사이트를 얻을 수 있고, 알고리즘 개선에 큰 도움
 - 차원 축소는 작업을 수행하는 데 도움이 됨
 - 데이터를 시각적으로 제 3자에게 보여 줄 수 있다면, 설명에 용이

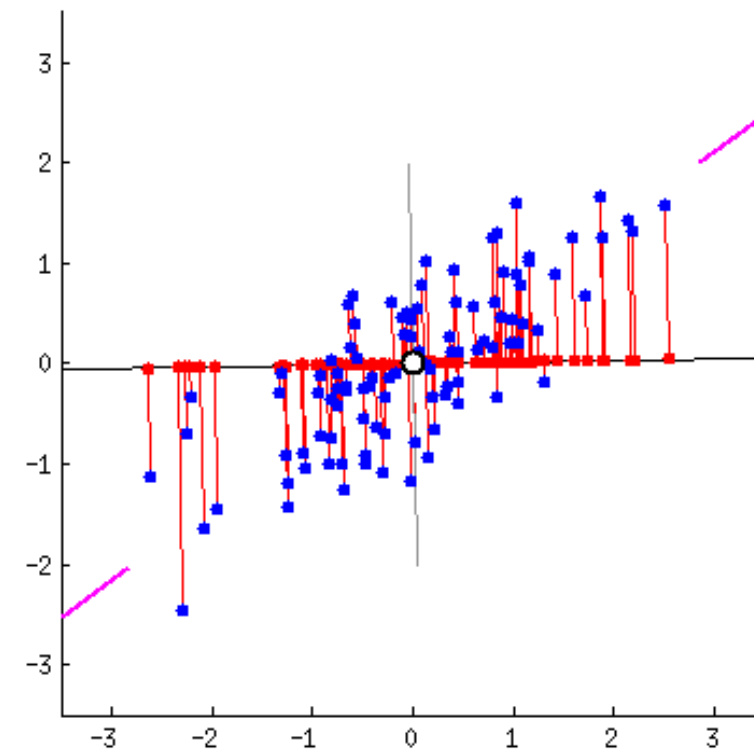
차원 축소 알고리즘

- 특이값 분해 (SVD)
- 주성분 분석 (PCA)



The diagram illustrates the SVD decomposition of matrix M into three matrices: U , Σ , and V^* . The dimensions of these matrices are given as $m \times n$, $m \times m$, and $n \times n$ respectively. The matrix Σ is shown as a diagonal matrix with non-zero elements on the diagonal. The matrices U and V are shown as orthogonal matrices, with U^* and V^* representing their conjugate transposes. The equations $U U^* = I_m$ and $V V^* = I_n$ are also shown, indicating that U and V are orthogonal matrices.

$$\begin{matrix} \text{Matrix} & \text{Dimensions} \\ M & m \times n \\ U & m \times m \\ \Sigma & m \times n \\ V^* & n \times n \end{matrix}$$
$$U U^* = I_m$$
$$V V^* = I_n$$



* Source: https://ko.wikipedia.org/wiki/%ED%8A%B9%EC%9D%B4%EA%B0%92_%EB%B6%84%ED%95%B4

** Source: <https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues>

E.O.D