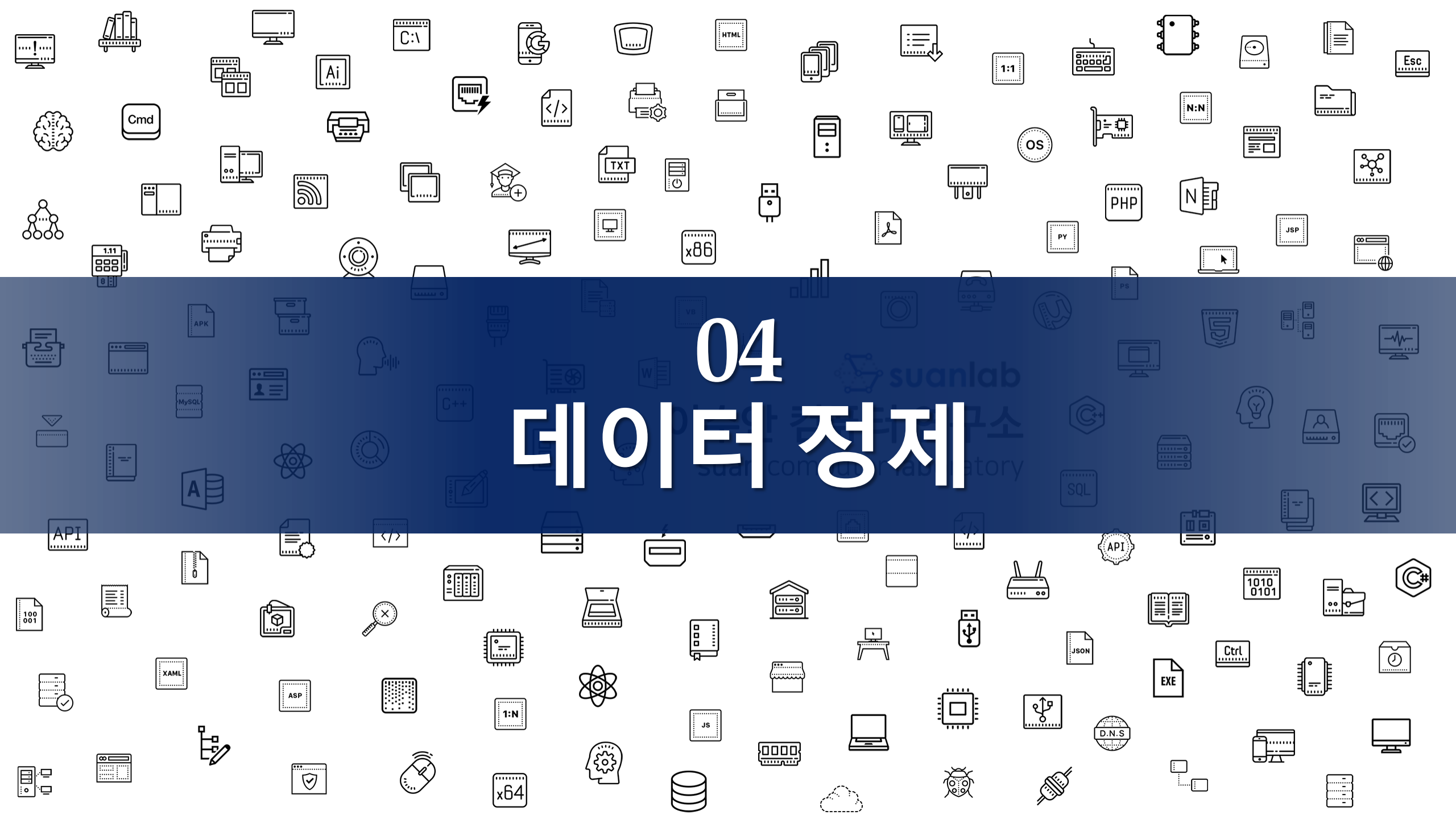
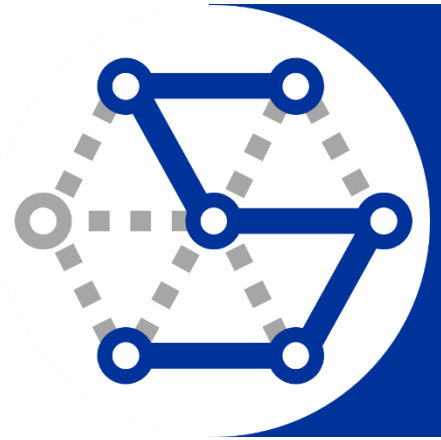
데이터 전처리
Data Preprocessing


04
데이터 정제

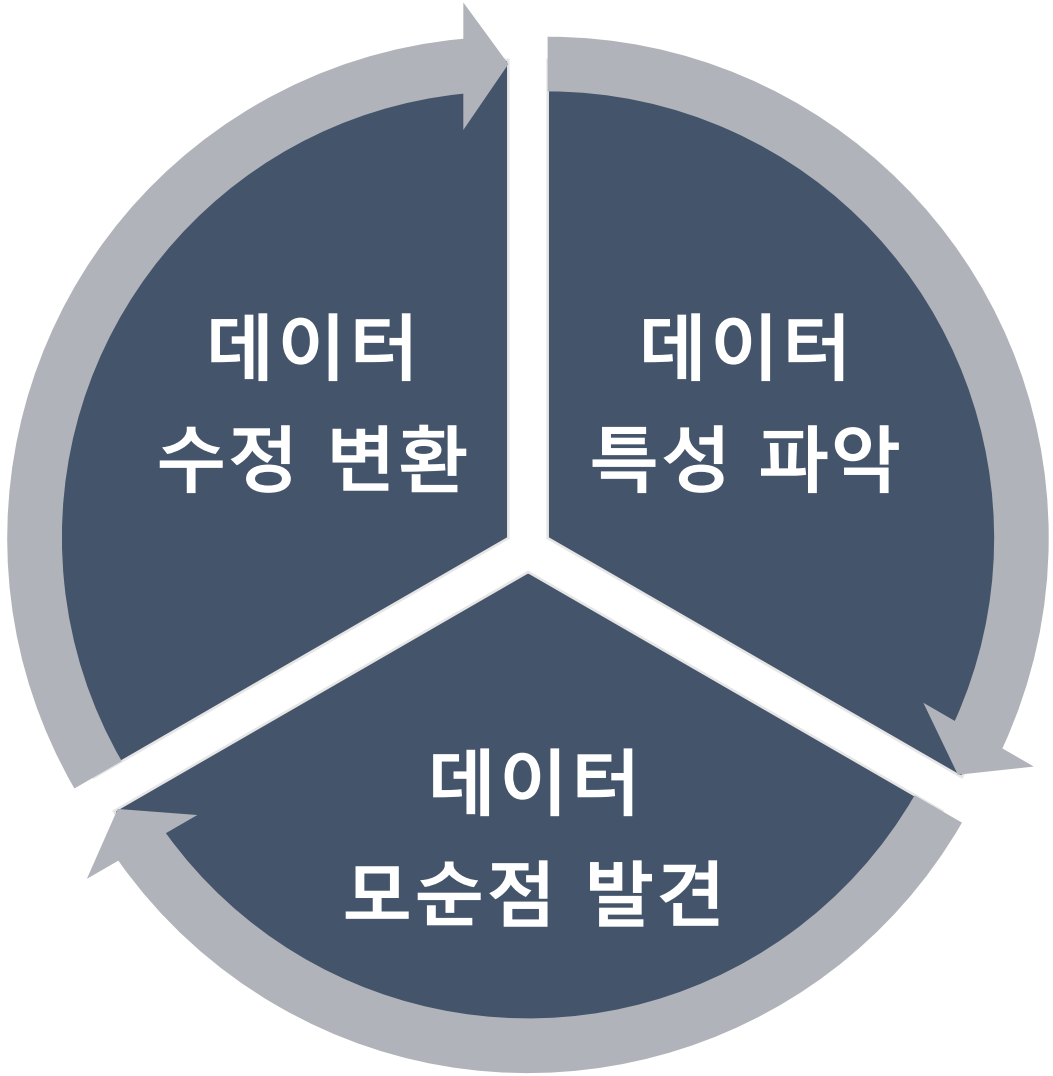
목차



1. 데이터 정제 절차

데이터 정제 절차

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.



데이터 특성 파악

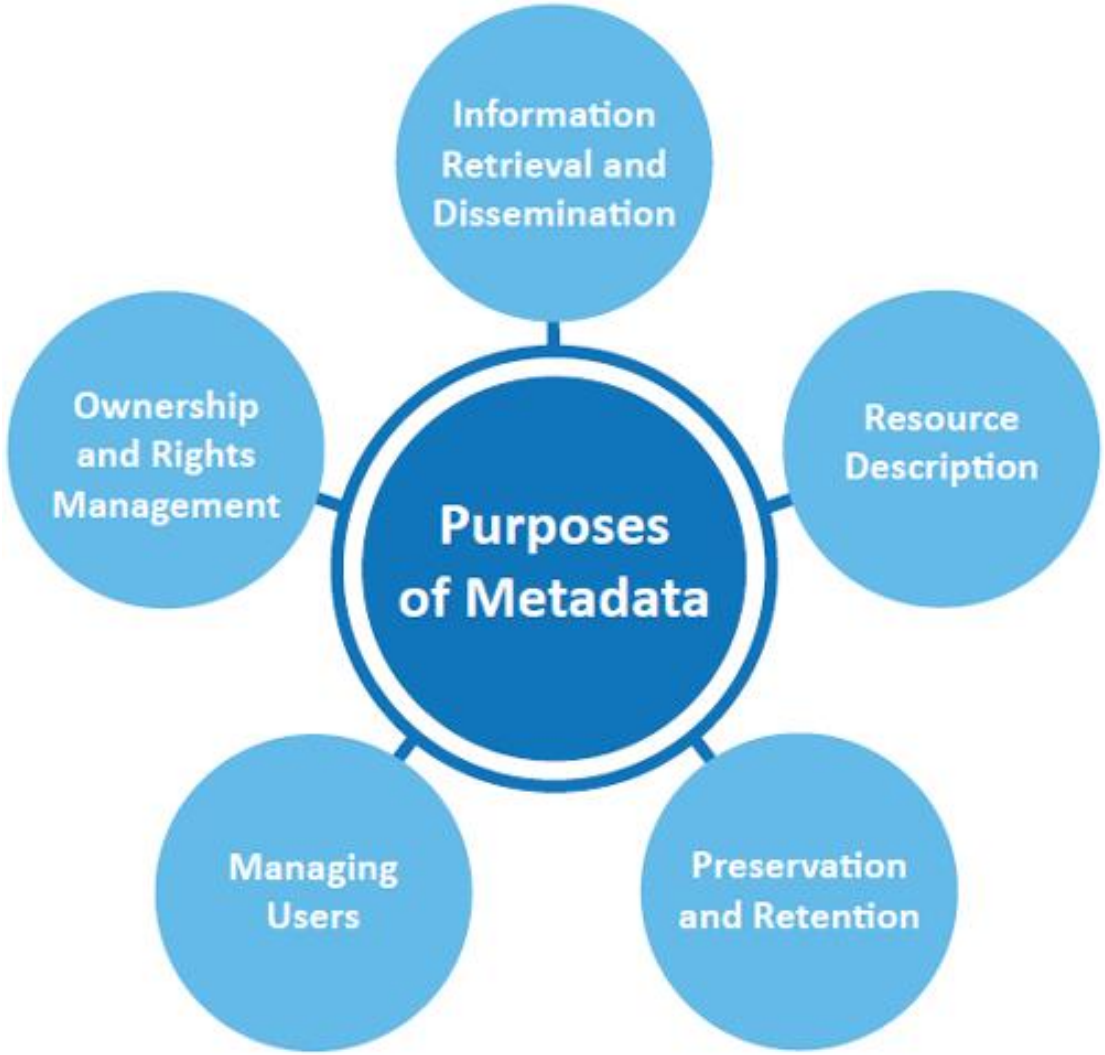
- 속성의 데이터 타입과 도메인(속성 값의 범위)
- 속성 값의 분포 특성(대칭, 비대칭 등)
  - 대칭/비대칭 분포
  - 실제 값의 주요 분포 범위
  - 값의 표준편차
- 속성 간의 의존성
  - 속성 A의 값이 같은 데이터의 속성 B 값이 반드시 같다면, 속성 A와 속성 B 간의 함수적 종속성 존재 (A→B)

메타데이터

메타데이터Metadata

https://namu.wiki/w/메타데이터

- '데이터에 대한 데이터'
- 데이터에 관한 구조화된 데이터로 다른 데이터를 설명해 주는 데이터
- 메타데이터의 종류
  - 기술용 메타데이터descriptive metadata
    - 정보 자원의 검색을 목적으로 한 메타데이터
    - 발견, 식별, 선정, 병치, 평가, 링크, 가용성
    - 전통적 도서관 편목
  - 관리용 메타데이터administrative metadata
    - 자원 관리를 용이하게 하기 위한 메타데이터
    - 보존용 메타데이터의 요소
  - 구조용 메타데이터structural metadata
    - 복합적인 디지털 객체들을 묶어주기 위한 메타데이터
    - 오디오와 텍스트 결합



https://cdn-images-1.medium.com/max/1000/1*y3bST2DnCmDTQzD5YnamTw.png

데이터 모순점 발견

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 잘못 설계된 데이터 입력 폼이 존재
- 데이터 입력에서 사람의 실수로 발생
- 응답자가 자신의 정보가 누설되는 것을 원하지 않기에 발생하는 의도적인 오류
- 만료된 데이터 (바뀐 주소 등)
- 데이터 표현의 모순
- 일치하지 않는 코드의 사용
- 데이터를 기록하는 계측 장치의 오류나 시스템 오류
- 원래의 의도와 다른 목적으로 데이터를 부적절하게 사용
- 데이터 통합 과정에서 주어진 속성이 다른 데이터베이스에서 다른 이름을 사용

필드 오버로딩Field Overloading

- 데이터 분석가는 코드 사용의 불일치와 데이터 표현의 불일치를 주의해야 함
- 필드 오버로딩은 개발자가 기존에 정의된 속성에서 사용하지 않은 일부를 새로운 속성의 정의로 사용할 때 발생
- 필드 오버로딩에 대한 검토
  - 유일 규칙Unique Rule: 주어진 속성의 값이 같은 속성의 다른 값들과는 달라야 함
  - 일관 규칙Consecutive Rule: 최소값과 최대값 사이에 결측치가 없어야 함
  - 무 규칙Null Rule: 공백, 물음표, 특수문자 등과 같이 데이터가 없음을 나타내는 다른 문자의 사용 가능

데이터 수정 변환

- 모순점이 발견된 데이터에 대해서 수정 변환 필요
- 데이터 수정 변환 시 오류 발생 가능성도 높고, 많은 시간이 필요
- 어떤 수정 변환은 더 많은 모순이 생길 수 있음
- 어떤 모순은 다른 것들이 모두 수정된 후에 감지될 수 있음
  - 년도 속성에 20001 이라는 값은 다른 날짜 값들이 일관된 형식(YYYY)으로 변환될 때 발견
  - 성별 속성에 'emale'이라는 값은 다른 성별 값들이 일관된 형식(1:male, 2:female)으로 변환될 때 발견

2. 결측값 처리

결측값Missing values

- 값이 존재하지 않고 비어있는 상태
- NANot Available 또는 NULL 값
  - NA: 결측값
  - NULL: 값이 없다
- 분석 대상의 속성 값이 상당 부분 비어있게되면, 분석 대상 데이터가 충분하지 않은 상태이므로 제대로 된 분석을 수행하기 어려움

결측값 구분

https://en.wikipedia.org/wiki/Missing_data

- MCAR Missing Completely At Random
  - 결측값이 관측된 데이터와 관측되지 않은 데이터와 독립적이며 완전 무작위로 발생
  - 데이터 분석 시 편향되지 않아서 결측값이 문제가 되지 않는 경우
  - 데이터가 MCAR인 경우는 거의 없음
- MAR Missing At Random or MCAR Missing Conditionally At Random
  - 결측값이 조건이 다른 변수에 따라 조건부로 무작위 발생되는 경우
  - 변수의 조건에 따른 결측값이 설명할 수 있는 경우
  - 데이터 분석 시 편향이 발생할 수도 있음
- MNAR Missing Not At Random
  - MCAR 또는 MAR이 아닌 데이터
  - 무시할 수 없는 무응답 데이터 (누락된 이유가 존재)
  - 결측값이 무작위가 아니라서 주도면밀한 추가 조사가 필요한 경우

결측값 처리 방법

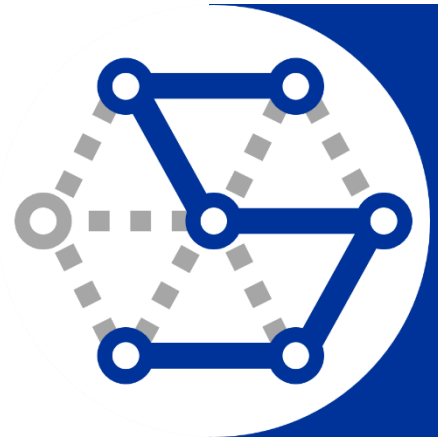
- 결측값 데이터 개체 또는 속성의 제거
  - 결측값이 발생한 데이터 개체를 분석 과정에서 제거하거나 해당 속성을 제거
  - 데이터가 충분히 많이 있다면 고려할만한 방법
  - 데이터 내에 결측치를 가진 데이터나 속성이 많은 경우 대부분의 정보가 제거될 수 있음
  - 실제로는 많이 사용하지 않는 방법
- 수동으로 결측값 입력
  - 결측값이 발생한 데이터를 다시 조사 및 수집하여 입력
  - 매우 고비용의 소모적인 방법
  - 결측값이 많은 경우 비현실적인 방법
- 전역상수global constant를 사용한 결측값 입력
  - 단순하고 명확한 방법
  - 예를 들어, 결측값을 0으로 입력
  - 전역상수 값이 분석 결과를 왜곡할 수 있음

결측값 처리 방법

- 결측값의 무시
  - 알고리즘이나 응용에 따라서는 결측치가 발생한 속성을 무시하고 분석을 수행할 수도 있음
  - 예를 들어, 개체들 사이의 유사성 계산에 있어 많은 수의 속성이 있는 경우 이 중 하나의 속성이 없다면 이를 제외하고 유사성을 계산할 수 있도록 알고리즘을 조정하는 것
  - 하나의 속성 값이 없더라도 유사성을 계산하는데 미치는 영향이 크지 않다면 이러한 방법도 적용 가능
  - 데이터 간 결측값을 가진 속성들이 산재해 있다면 너무 많은 데이터가 제외될 수 있음
  - 속성이 몇 개 없어 하나의 속성이라도 무시하기 힘든 경우라면 이러한 방법의 적용은 좋지 않음
- 결측값의 추정
  - 일반적으로 많이 사용되는 방법
  - 결측값이 발생한 데이터와 유사한 데이터를 사용하여 결측값을 추정하는 방법
  - 결측값을 추정하는 방법에 따라 다양한 형태가 존재

결측값 추정 방법

- 속성의 평균값을 사용하여 결측값 추정
  - 속성의 평균값을 결측값에 채워넣는 방법
  - 분석 결과를 왜곡시킬 위험성 존재
- 같은 클래스에 속하는 속성의 평균값 사용
  - 주어진 데이터와 같은 클래스(분류)에 속하는 튜플 들의 속성 평균값 사용
  - 동일 유형에 속하는 데이터의 평균값을 사용하므로 왜곡 가능성 줄임
- 가장 가능성이 높은 값으로 결측값 추정
  - 회귀분석, 베이지안Bayesian 기법, 의사결정트리 기법 등의 통계 또는 마이닝 기법을 활용하여 결측값 예측
  - 분석에 의해 가능성이 높은 값을 찾아내는 방법
  - 가장 효과적이고 높은 정확도의 결측값 예측 가능
  - 결측값을 채우기 위한 분석 가설을 세우는 등의 복잡성 존재



3. 잡음 제거

잡음Noise data

- 측정된 변수(속성)에서의 오류나 오차 값
- 오류나 오차 값에 의해 경향성 훼손 발생
- 잡음에 대한 훼손을 줄이기 위해 데이터 평활화 기법smoothing technique 존재
- 데이터 평활화 기법
  - 구간화Binning
  - 회귀Regression
  - 군집화Clustering

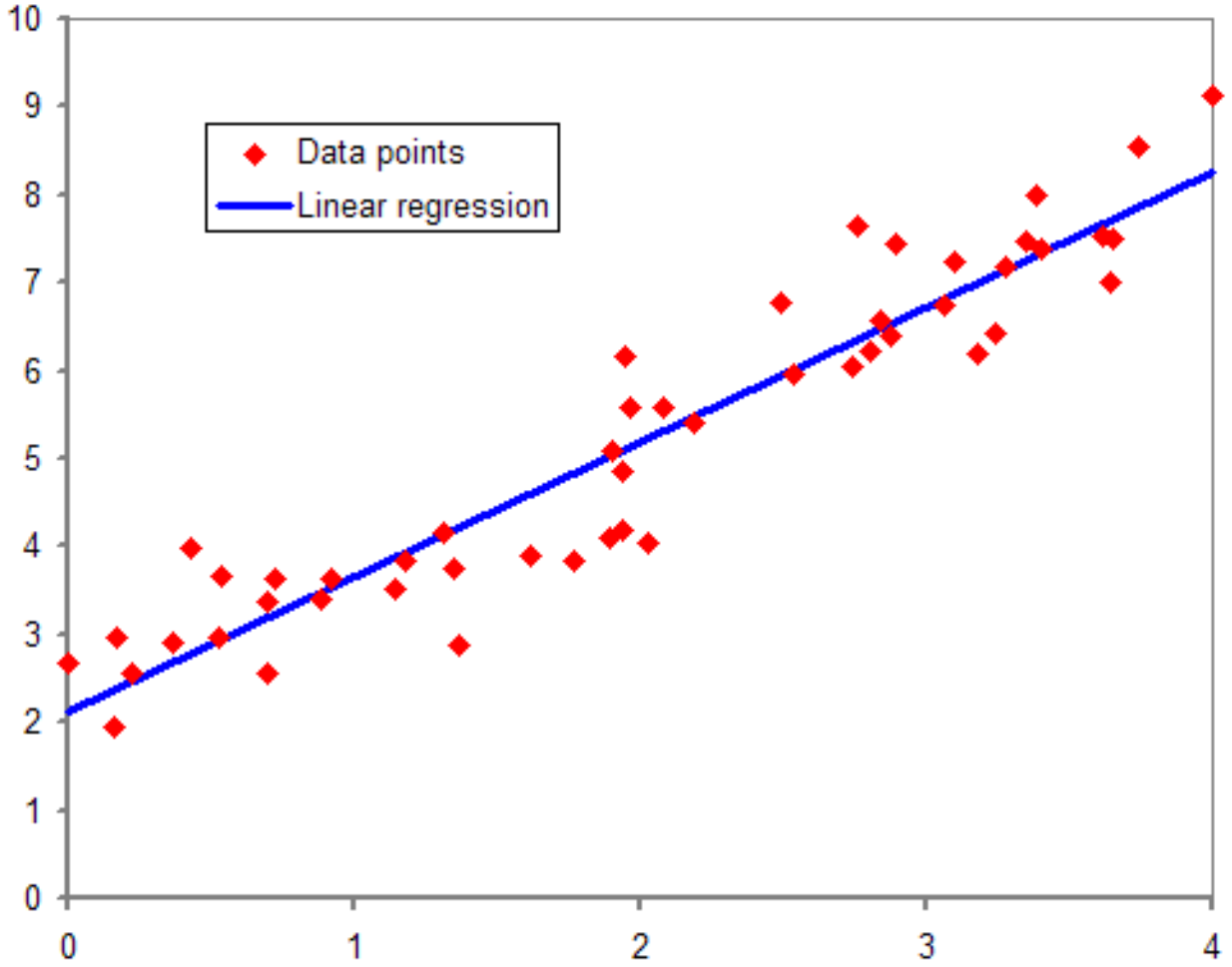
구간화Binning

- 정렬된 데이터 값들을 몇 개의 빈(혹은 버킷)으로 분할하여 평활화하는 방법
- 이웃neighborhood(주변 값)들을 참조하여 정렬된 데이터를 매끄럽게 함
  - 평균값 평활화smoothing by bin means : 빈bin에 있는 값들이 그 빈의 평균값mean으로 대체
  - 중앙값 평활화smoothing by bin medians : 빈bin에 있는 값들이 그 빈의 중앙값median으로 대체
  - 경계값 평활화smoothing by bin boundaries : 빈bin의 최소값과 최대값이 그 빈의 경계값이 되며, 두 경계값 중 가까운 쪽 값으로 대체

회귀Regression

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 회귀함수를 이용한 데이터 평활화 기법
- 선형 회귀 분석linear regression analysis은 하나의 속성이 다른 하나의 속성을 예측하는 데 이용할 수 있도록 선형 관계를 찾음
- 다중 회귀 분석multiple regression analysis은 두 개 이상의 속성을 가지고 다른 속성을 예측하는 데 이용할 수 있도록 단순 선형 회귀 분석의 확장
- 선형 회귀 분석 또는 다중 회귀 분석을 이용하여 평활화

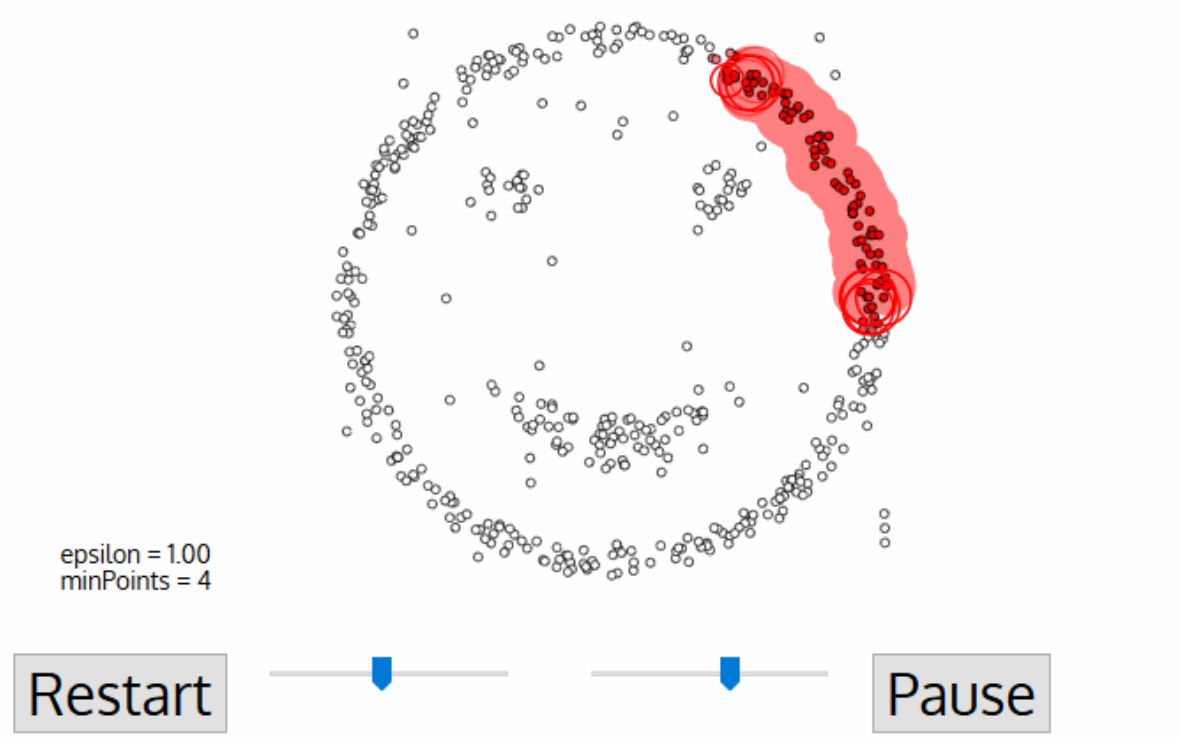


https://upload.wikimedia.org/wikipedia/commons/thumb/b/be/Normdist_regression.png/300px-Normdist_regression.png

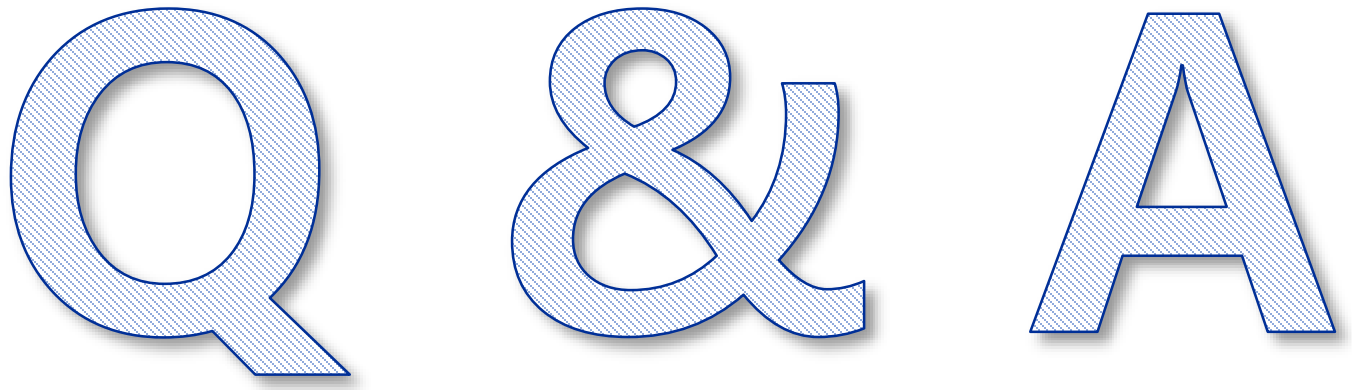
군집화Clustering

이현호, Python과 SQL을 활용한 실전 데이터 전처리, 카오스북, 2018.

- 군집화: 유사한 값들끼리 그룹화하는 과정
- 이상값outlier: 어떤 군집에도 속하지 않은 값
- 이상값은 경향성을 훼손하고, 오류 데이터일 가능성이 높음
- 이상값에 대해서 평활화 수행



https://static.wixstatic.com/media/5dad68_abfdb1fe2b0f4f64bfad9e03e1a70b73~mv2.gif


Q & A