

데이터 전처리 Data Preprocessing

02

데이터 전처리 개요

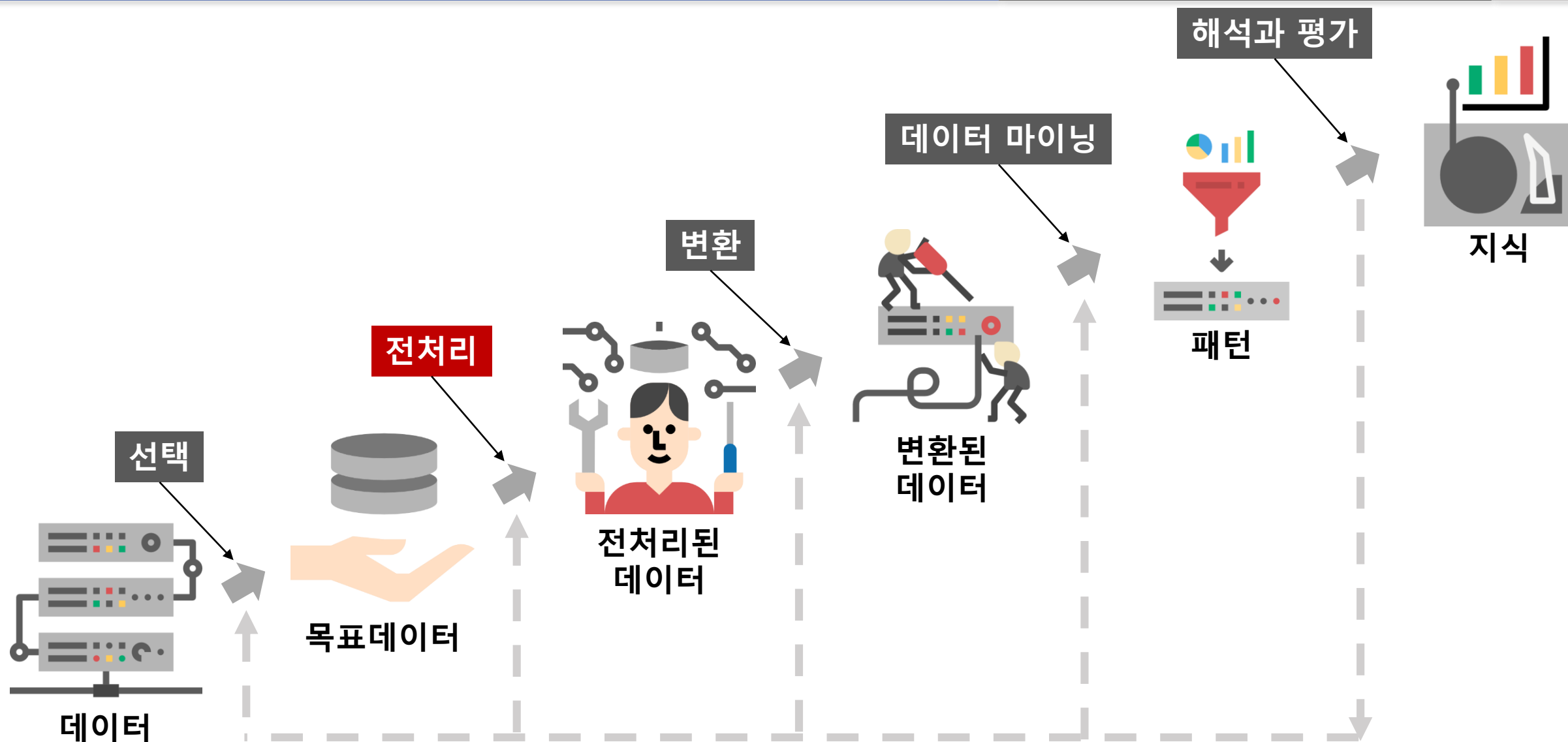
목차

1. 데이터 전처리
2. 데이터 품질
3. 데이터 전처리 단계

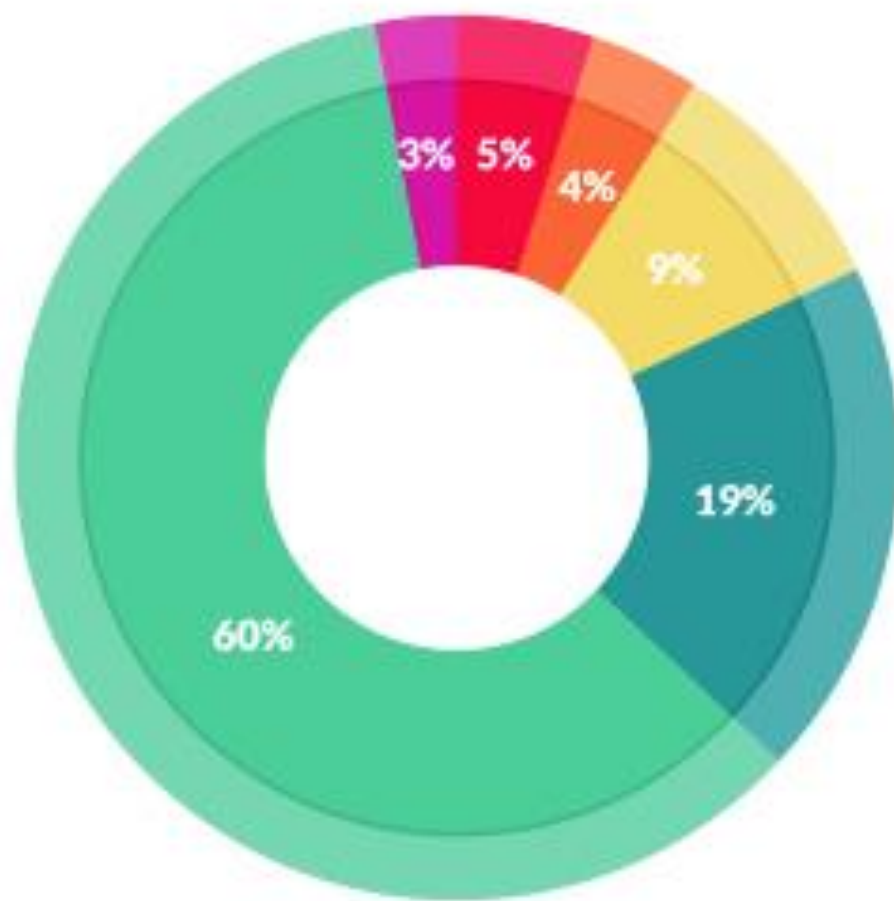


1. 데이터 전처리

데이터 분석 단계



데이터 과학자들이 가장 많은 시간을 소요하는 일

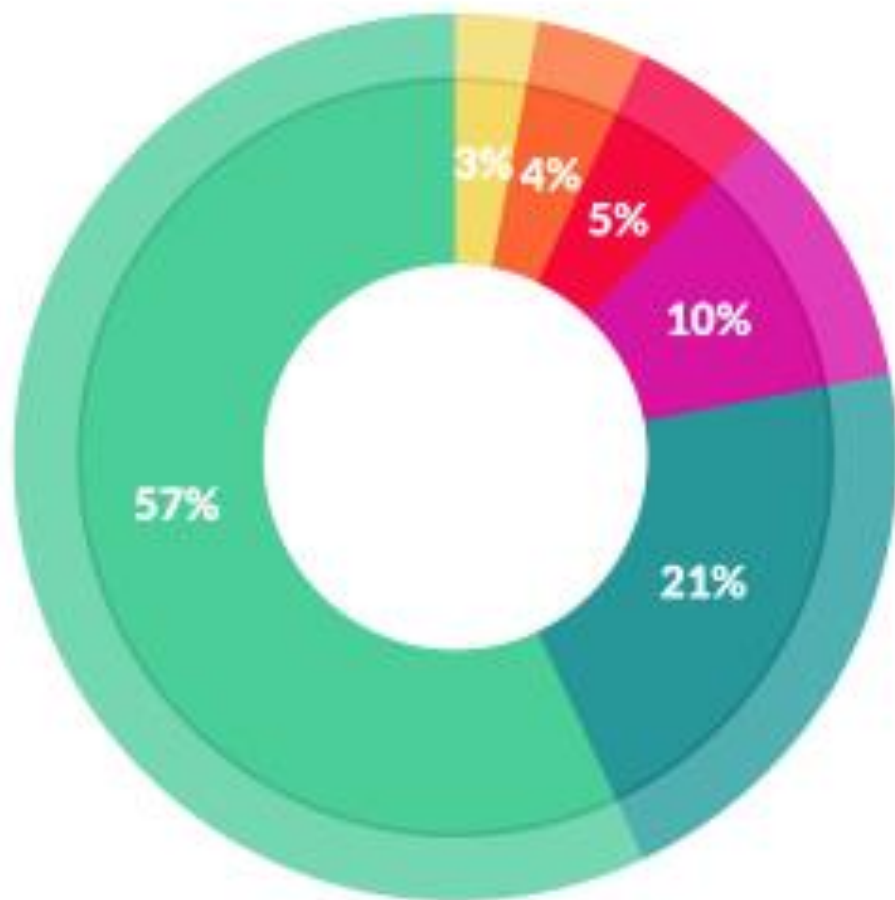


What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

CrowdFlower 2016 Data Science Report

데이터 과학에서 가장 즐겁지 않은 부분



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

데이터 전처리

- 데이터를 분석 및 처리에 적합한 형태로 만드는 과정을 총칭하는 개념
- 데이터 전처리는 데이터 분석 및 처리 과정에서 중요한 단계
- 데이터 분석, 데이터 마이닝, 머신 러닝 프로젝트에 적용
- 일반적으로 데이터는 비어있는 부분이 많거나 정합성이 맞지 않는 경우가 많음
- 아무리 좋은 도구나 분석 기법도 품질이 낮은 데이터로는 좋은 결과를 얻을 수 없음

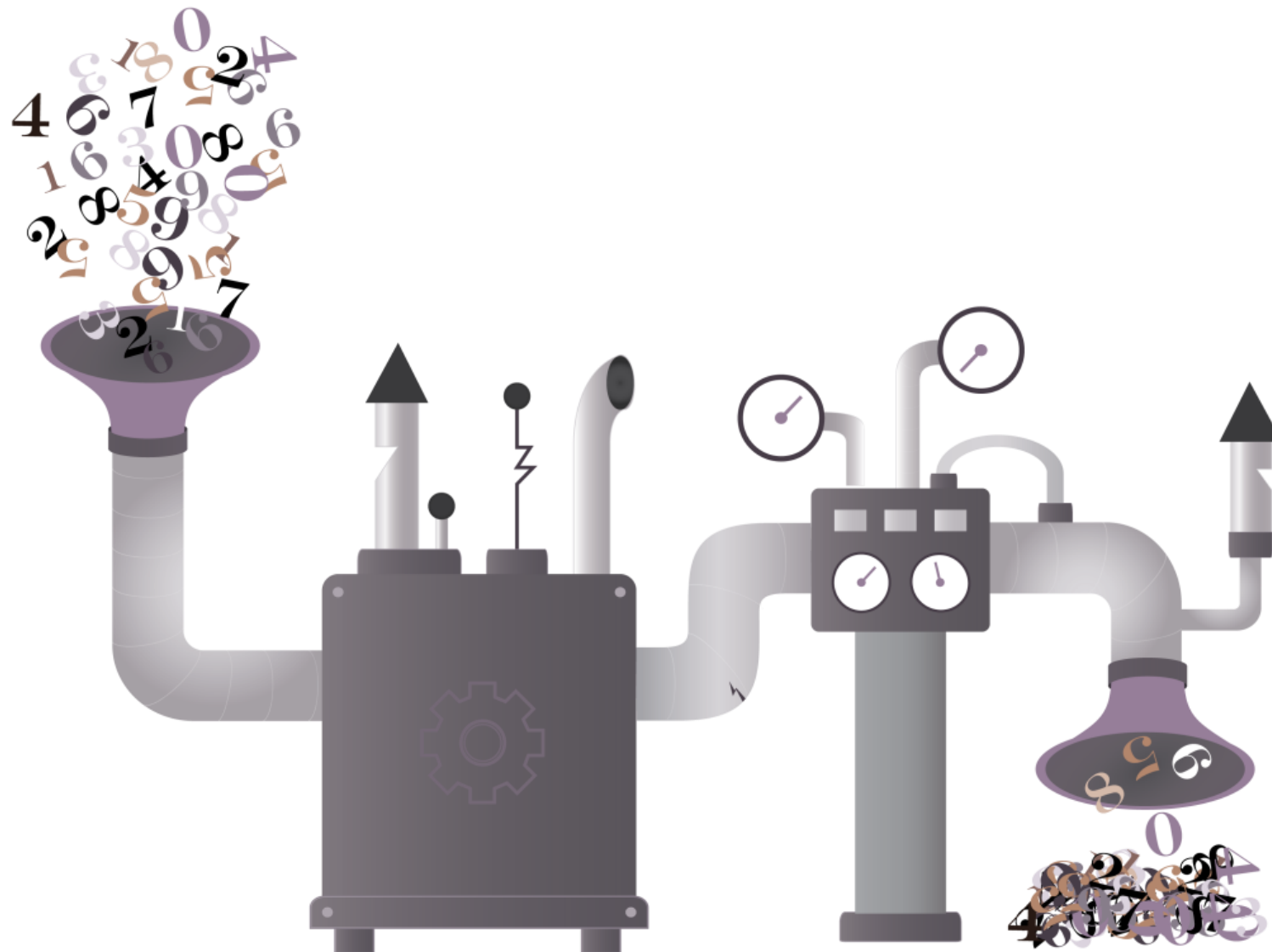
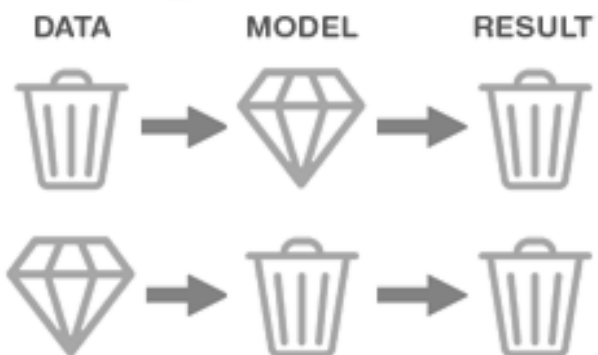
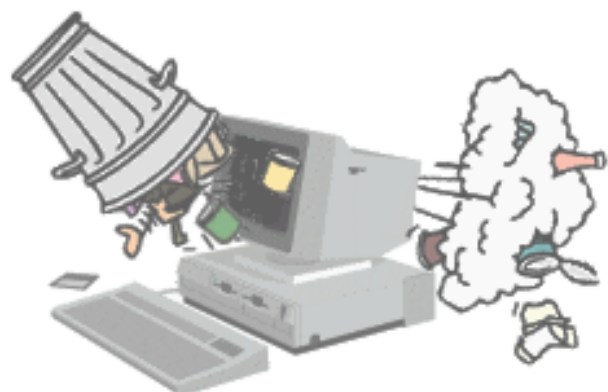




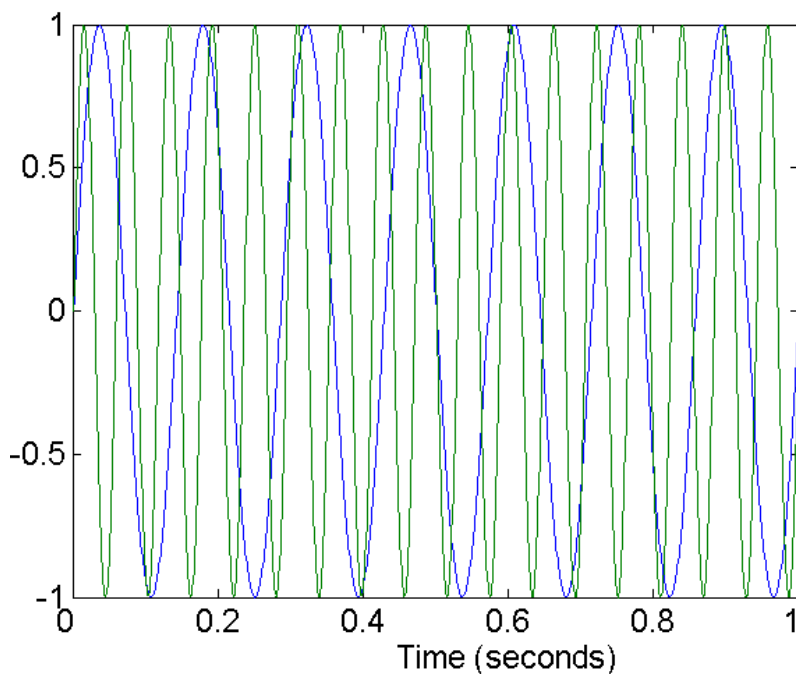
2. 데이터 품질

- 완벽한 데이터를 얻는다는 것은 실제에서는 불가능한 일
- 데이터 품질을 저해하는 주요 요인으로서는 크게 측정 오류와 수집 과정에서 발생하는 오류로 나눌 수 있음
- 측정 오류: 사람의 실수로 잘못된 단위로 기록을 하거나 측정 장비 자체의 한계 등 측정 과정에서 발생하는 오류
- 수집 과정 오류: 데이터의 손실, 중복 등의 문제로 발생하는 오류

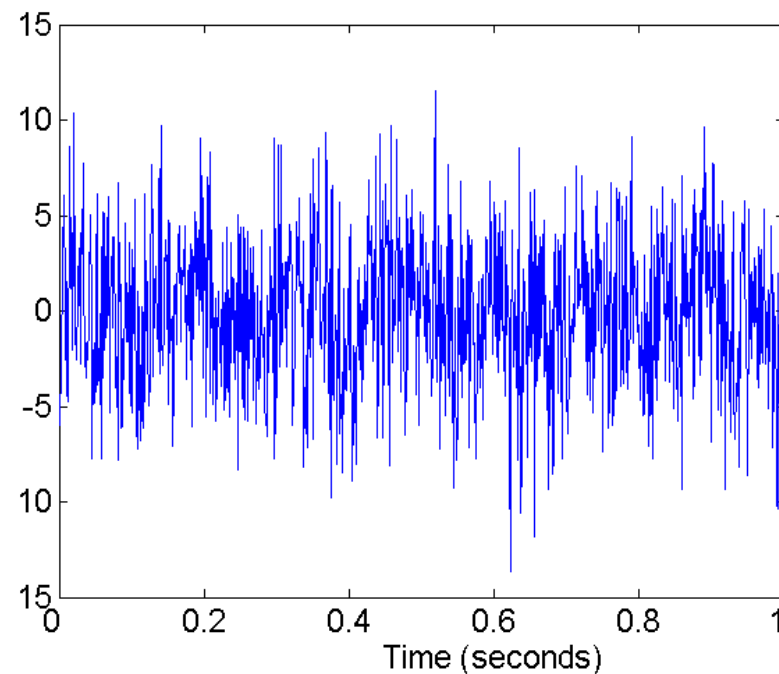
■ GIGO Garbage In Garbage Out



- 측정 과정에서 무작위로 발생하여 측정값의 에러를 발생시키는 것
- 실제 데이터는 매끈한 곡선 형태의 시계열 데이터였지만 측정 과정에서 잡음이 포함됨으로 인해 실제 값과 다른 데이터를 얻게 되어 실제 데이터의 형태를 잃어버릴 수도 있음



Two Sine Waves



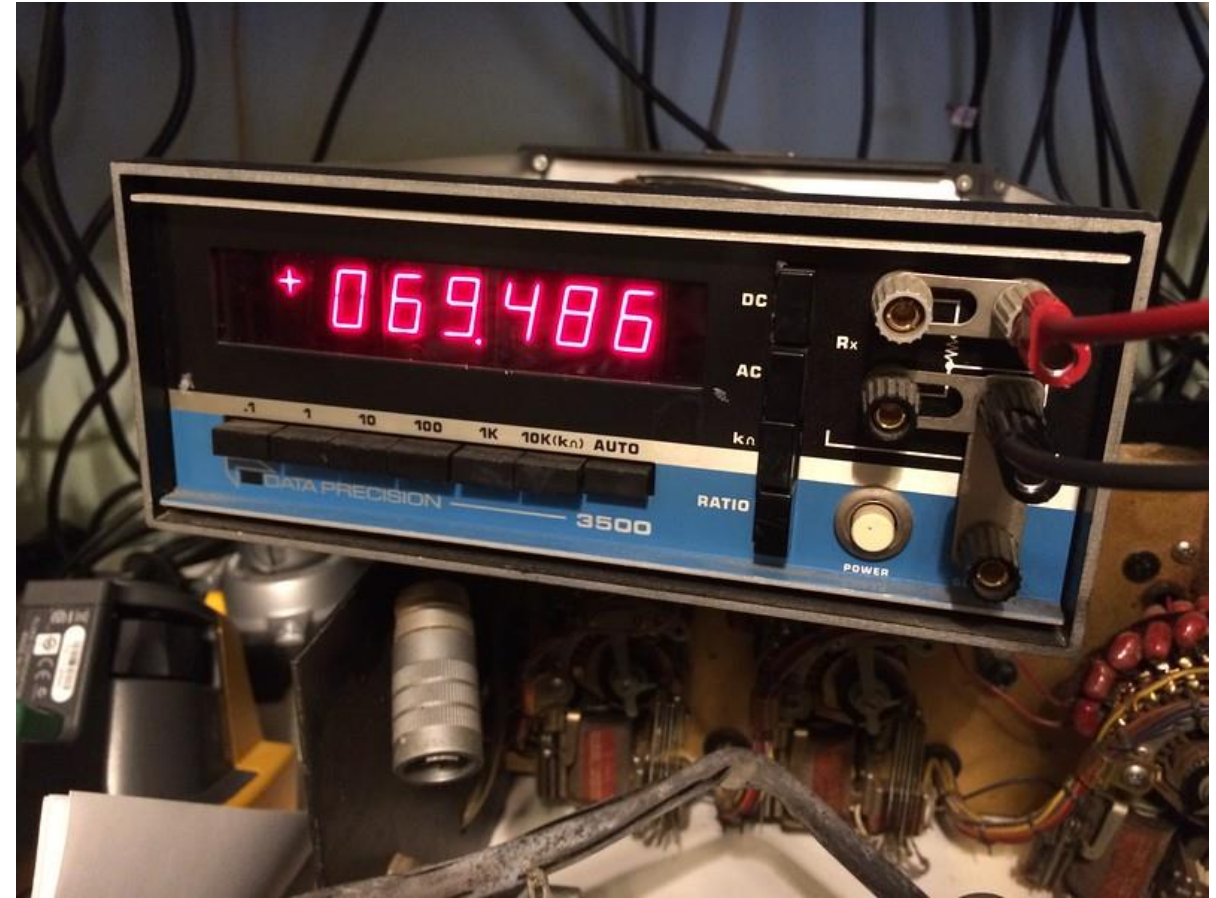
Two Sine Waves + Noise

- 어떠한 요인으로 인해 반복적으로 발생하는 왜곡이나 에러를 의미
- 일례로 카메라를 이용한 영상 데이터 획득에 있어 카메라 렌즈에 얼룩이 묻어 있다면 이에 해당하는 부분에서는 이 얼룩으로 인한 왜곡이 지속적으로 발생



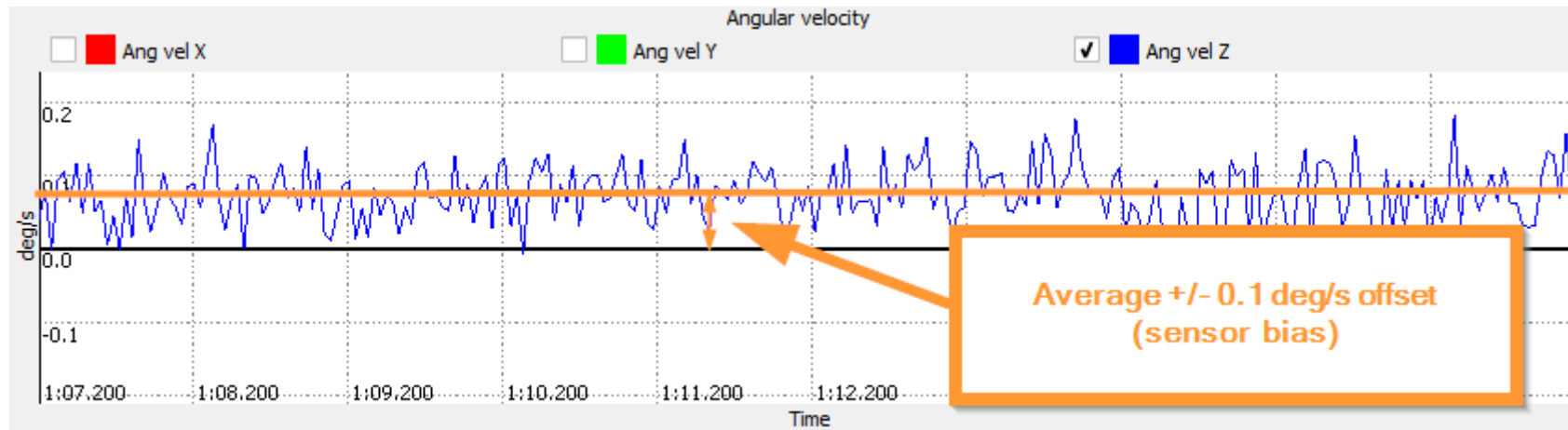
<https://www.cis.rit.edu/htbooks/mri/chap-11/chap-11.htm>

- 동일한 대상을 반복적으로 측정하였을 때의 각 결과의 친밀성을 나타내는 것
- 측정 결과의 표준편차^{standard deviation}로 나타낼 수도 있음
- 예를 들어 동일한 1g을 측정하는데 있어 각각의 측정 결과가 {1.015, 0.990, 1.013, 1.001, 0.986}인 경우 이들의 표준편차는 0.013이므로 이 때의 정밀도는 0.013이라 말할 수 있음



<https://pmanning.smugmug.com/Electronics/Data-Precision-3500-DMM/>

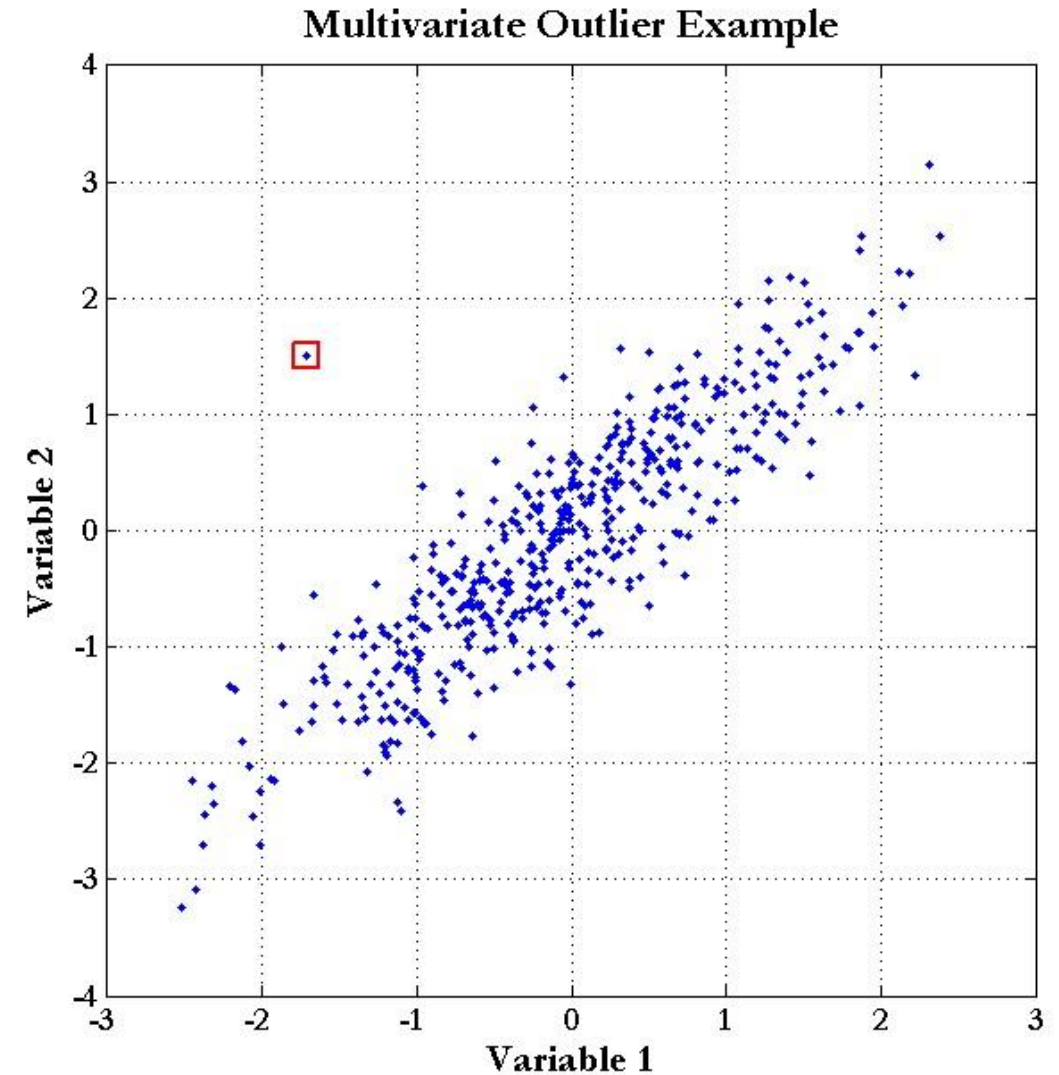
- 측정 장비에 포함된 시스템적인 변동으로 앞서 영점 조절 되지 않은 체중계가 좋은 예
- 정밀도에서 언급된 예제의 경우 1g에 대한 측정 평균은 1.001이며 이 측정 장비에는 0.001만큼의 바이어스가 포함되어 있음을 알 수 있음



<https://base.xsens.com/hc/en-us/articles/209611089-Understanding-Sensor-Bias-offset->

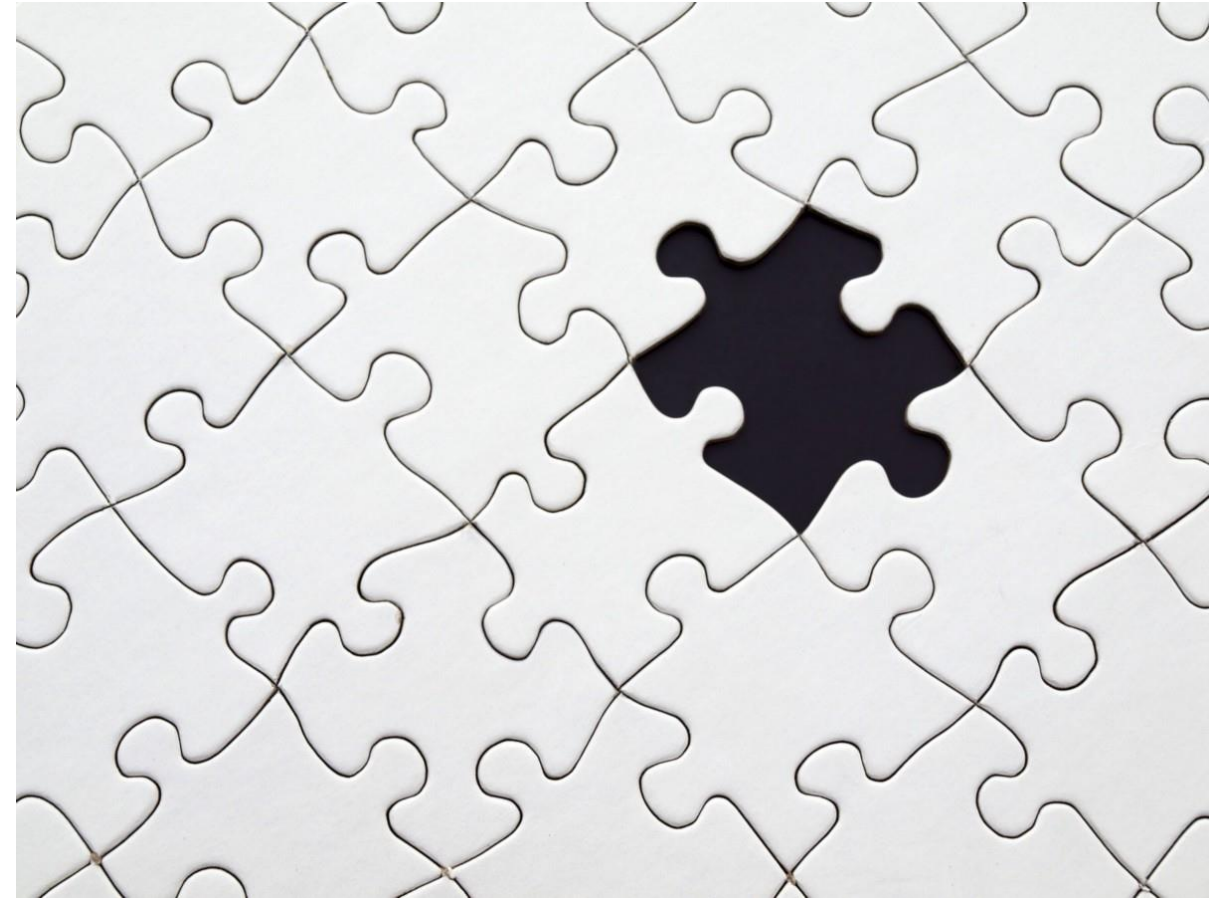
- 정확도는 정확도와 바이어스에 기인하는 것이지만 이를 이용하여 명시적으로 나타낼 수 있는 수식은 없음
- 다만 정확도는 유효 숫자^{Significant digit}의 사용에 있어 중요한 측면을 가지고 있음
- 이는 공학이나 과학에서 기본적으로 다루는 개념으로 수의 정확도에 영향을 주는 숫자를 의미
- 예를 들어, 측정에 있어 이는 측정 장비의 한계로 인해 정확하지 않은 자리의 수를 측정함에 따라 발생할 수 있는 문제로 자를 이용한 길이 측정을 가정
- 자의 최소 눈금이 1mm라면, 1mm 단위로 길이를 측정하게 될 것이며 이 경우 항상 $\pm 0.5\text{mm}$ 만큼의 오차를 가지게 됨
- 이 자를 이용하여 측정한 길이가 10.3mm 였다면 1mm 미만의 값인 0.3mm 라는 수치는 의미가 없음을 알 수 있음

- 대부분의 데이터와 다른 특성을 보이거나 특정 속성의 값이 다른 개체들과 달리 유별난 값을 가지는 데이터를 의미
- 이상치의 중요한 점은 잡음과는 다르다는 것
- 잡음이 임의로 발생하는 예측하기 어려운 요인임에 반해 이상치는 적법한 하나의 데이터로서 그 자체가 중요한 분석의 목적이 될 수도 있음
- 예를 들어 네트워크의 침입자 감시와 같은 응용에 있어서는 대다수의 일반 접속 중 예외적으로 발생하는 불법적인 접속 시도와 같은 이상치를 찾는 것이 주된 목표



<https://madhureshkumar.wordpress.com/2015/06/18/trend-and-outlier/>

- 데이터의 결측은 일반적인 경우는 아니지만 드물게 발생하는 문제
- 설문조사의 경우 몇몇 사람들은 자신의 나이나 몸무게와 같은 사적인 정보를 공개하는 것을 꺼리는 경우가 발생하며 이러한 값들은 조사에 있어 결측값으로 남게 됨



- 때에 따라서는 동일한 개체에 대한 측정 데이터가 다르게 나타나는 경우가 발생할 수 있는데 이러한 경우를 모순 또는 불일치값이라 표현
- 예를 들어, 고객의 주소와 우편번호를 저장해 놓은 데이터를 생각해보면, 주소가 동일한 지역임에도 불구하고 어떠한 이유로 우편번호가 상이한 경우가 발생할 수 있음
- 이런 경우에는 주소를 확인해서 우편번호를 정정하는 작업이 필요

Sales	Region
19,593	North
20,541	Wst
20,642	Eest
27,769	E
22,895	Sth

<https://edu.gcfglobal.org/en/excel-tips/a-trick-for-finding-inconsistent-data/1/>

allthembooks [High Replication] [Community S](#)

Text Search > «empty».shared_index

Query

Query string:

Learn more about [text search syntax](#).

Run Query

« Prev 20 40 42 Next 20 »

ID	bookAuthors	bookCategory	bookCondition	bookDate	bookDescription
4794849949646848	James Stewart	Engineering and Applied Sciences	New	2013-12-31 00:00:00	Success in your calculus course ...
4663887806857216	Mark W. ...	Bussiness	New	2013-12-31	Business model innovation is

<https://stackoverflow.com/questions/20861697/inconsistent-values-for-getnumberfound-in-search-api>

- 데이터의 중복은 언제든지 발생 가능
- 문제는 중복된 데이터 사이에 속성의 차이나 값의 불일치가 발생할 수 있다는 것
- 기본적으로 모든 속성 및 값이 동일하다면 하나의 데이터는 삭제할 수 있지만, 그렇지 않은 경우에는 두 개체를 합쳐서 하나의 개체를 만들거나, 응용에 적합한 속성을 가진 데이터를 선택하는 등의 추가적인 작업을 필요로 하게 됨

First Name	Last Name	City	Age
John	Barrow	London	32
Aaron	Gomez	Manchester	28
Ralph	Johnson	Newcastle	25
Monica	Robert	Plymouth	37
Craig	Miller	Leicester	30
Paul	Martin	Manchester	32
Claire	Rudolf	London	22
Eric	Plymouth	Birmingham	39
Monica	Robert	Plymouth	37

<https://www.opentechguides.com/how-to/article/excel-2016/127/remove-duplicate-data.html>



3. 데이터 전처리 기법

데이터 전처리 단계



데이터 수집 Data Collection

분석이나 학습에 필요한 데이터를 부분 혹은 전체를 수집하는 작업



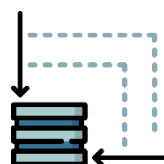
데이터 정제 Data Cleansing

비어있는 데이터나 잡음, 모순된 데이터 등을 정합성이 맞도록 교정하는 작업



데이터 통합 Data Integration

여러 개의 데이터베이스, 데이터집합 또는 파일을 통합하는 작업



데이터 축소 Data Reduction

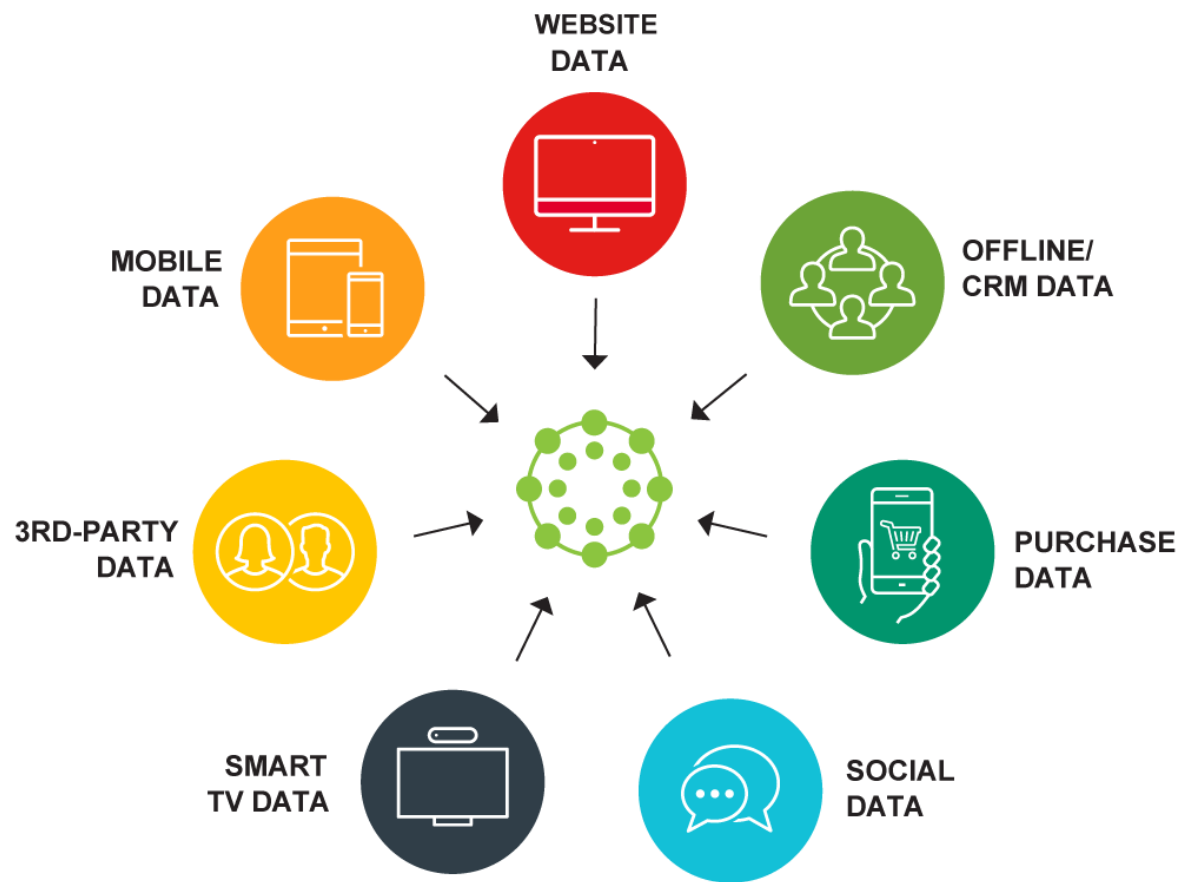
샘플링, 차원축소, 특징 선택 및 추출을 통해 데이터 크기를 줄이는 작업



데이터 변환 Data Transformation

데이터를 정규화, 이산화 또는 집계를 통해 변환하는 작업

데이터 수집 Data Collection



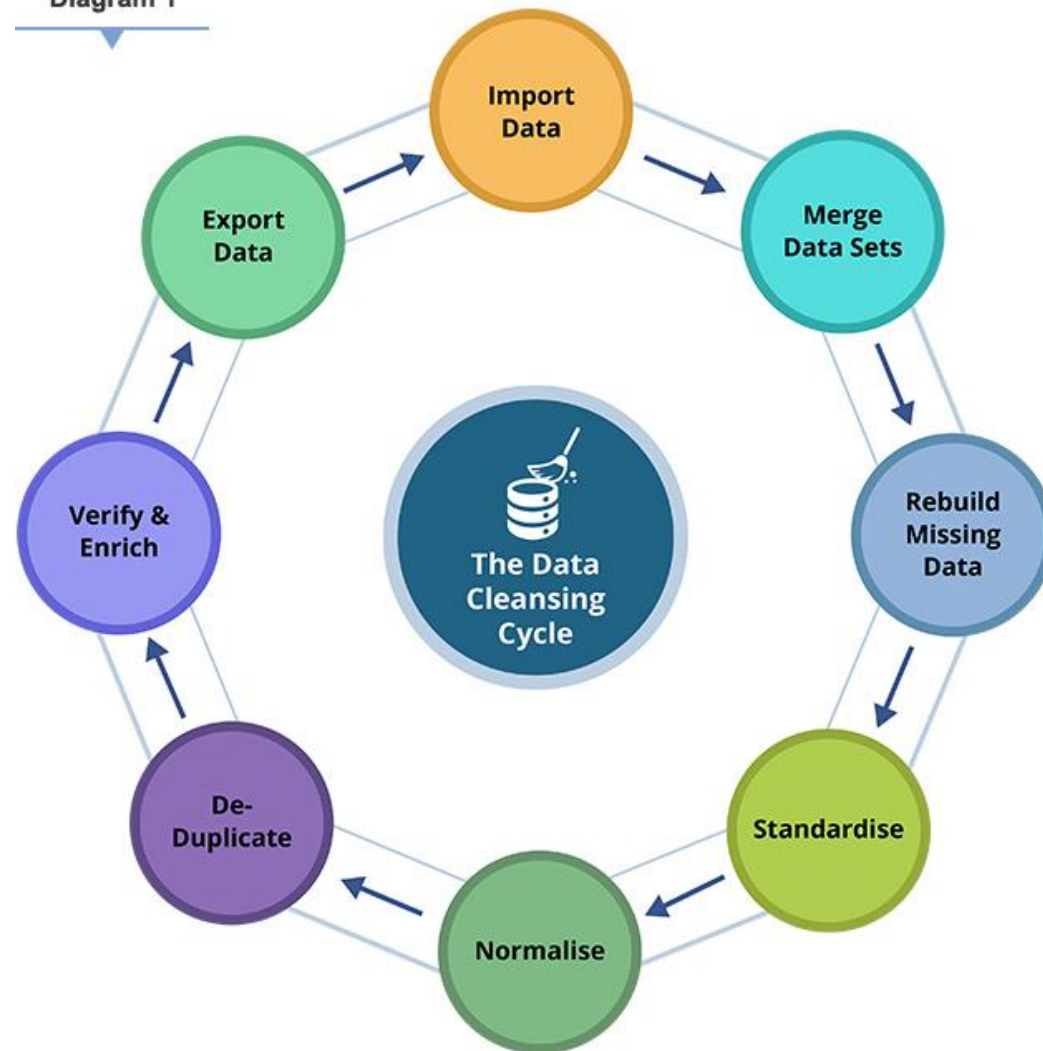
<http://xcademy.in/data-management-platform/data-collection/>

- 데이터 수집이 데이터 처리 분석 및 모델 생성의 첫 과정
- 목적과 목표가 되는 정보를 수집하고 측정하기 위해 정의가 필요
- 문제의 정의와 문제해결을 위한 데이터 분석 기획과 시나리오가 중요
- 문제를 식별하고 탐색함으로써 정보 수집 시기 및 방법을 결정
- 데이터 종류에 따라서 내부 또는 외부, 질적 또는 양적 데이터 수집

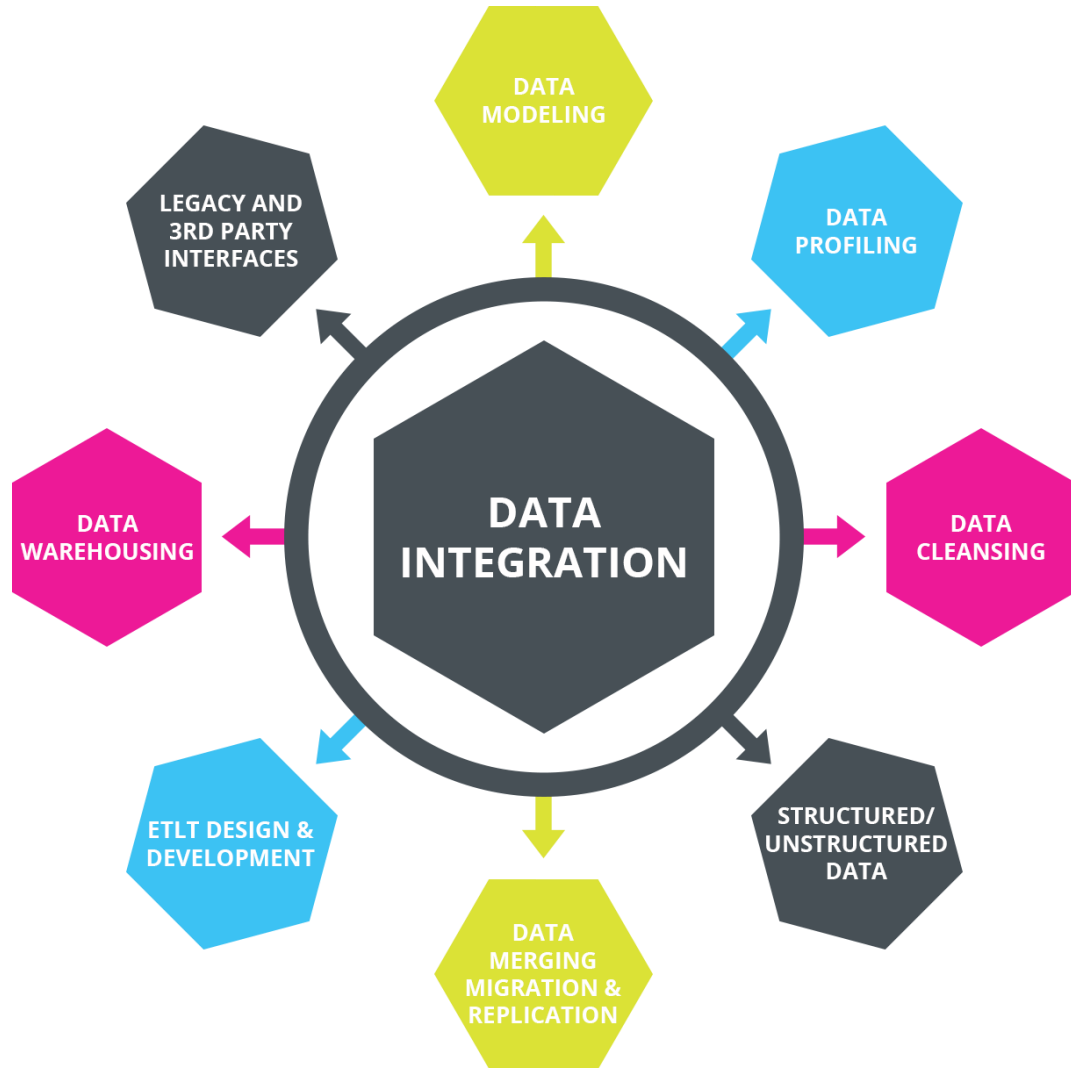
데이터 정제 Data Cleansing

- 데이터를 활용할 수 있도록 만드는 과정
- 데이터의 누락값, 불일치, 오류의 수정
- 컴퓨터가 읽을 수 없는 요소의 제거
- 숫자나 날짜 등의 형식에 대해 일관성 유지
- 적합한 파일 포맷으로 변환

Diagram 1



<https://www.dataentryoutsourced.com/blog/cxos-guide-to-marketing-and-sales-data-cleansing-and-enrichment/>



- 서로 다른 출처의 여러 데이터를 결합
- 서로 다른 데이터 세트가 호환이 가능하도록 통합
- 같은 객체, 같은 단위나 좌표로 데이터를 통합
- 링크드 데이터의 핵심 목표 중 하나는 데이터 통합을 완전히 또는 거의 완전히 자동화하는 것

<http://www.matricis.com/en/integration-solutions/data-integration/>

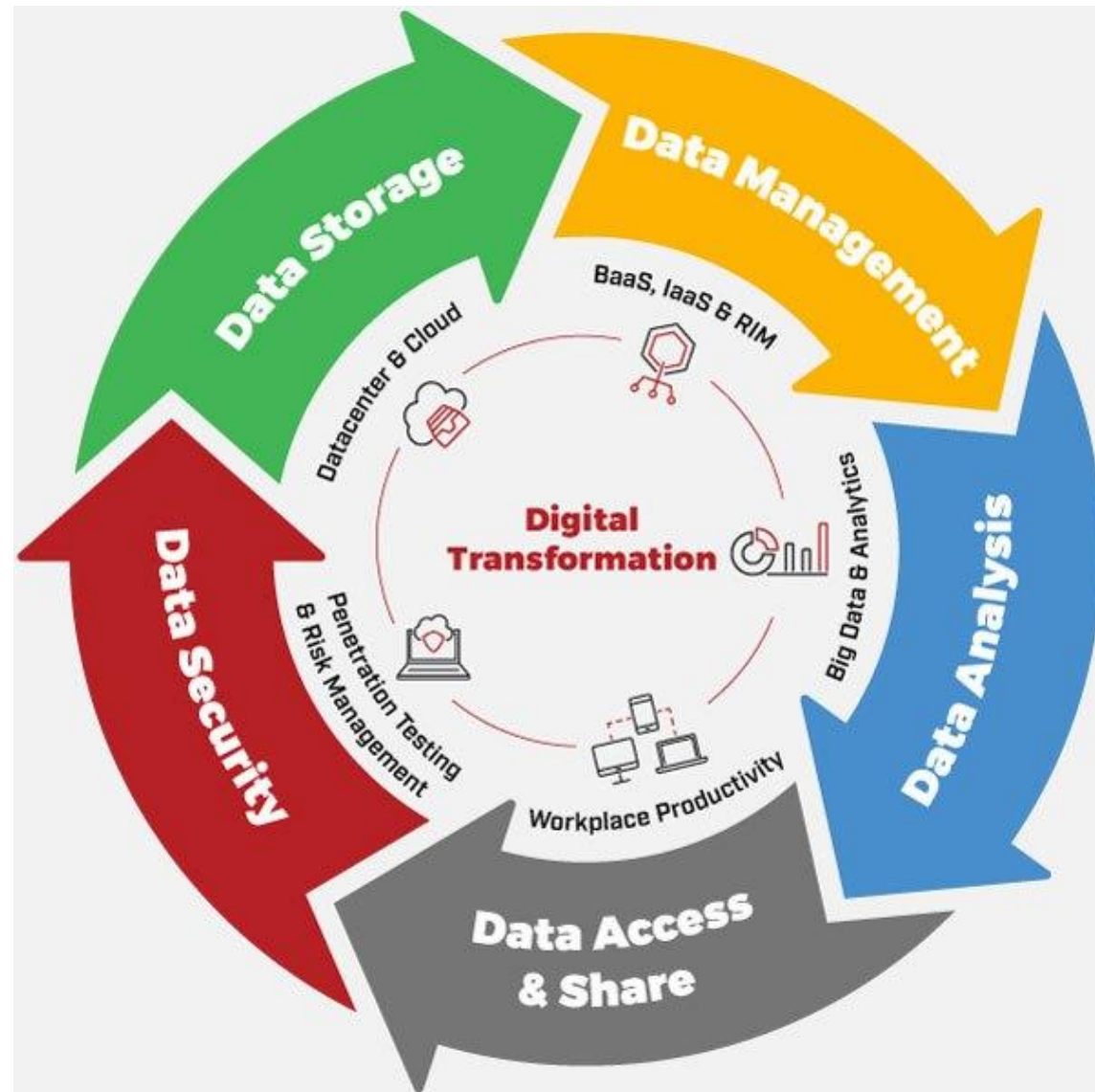
데이터 축소 Data Reduction

- 일반적으로 데이터는 매우 크기 때문에 대용량 데이터에 대한 복잡한 데이터 분석은 실행하기 어렵거나 불가능한 경우가 많음
- 데이터 축소는 원래 용량 기준보다 작은 양의 데이터 표현결과를 얻게 되더라도 원 데이터의 완결성을 유지하기 위해 사용
- 데이터를 축소하면 데이터 분석 시 좀 더 효과적이고 원래 데이터와 거의 동일한 분석 결과를 얻어낼 수 있는 장점



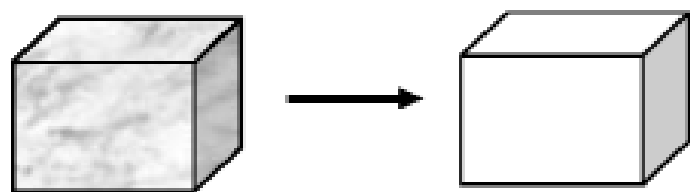
<https://www.cohesity.com/blog/cohesity-data-reduction-lock-stock-barrel/>

- 데이터를 한 형식이나 구조에서 다른 형식이나 구조로 변환
- 원본 데이터와 대상 데이터간에 필요한 데이터 변경 내용을 기반으로 데이터 변환이 간단하거나 복잡 할 수 있음
- 데이터 변환은 일반적으로 수동 및 자동 단계가 혼합되어 수행
- 데이터 변환에 사용되는 도구 및 기술은 변환되는 데이터의 형식, 구조, 복잡성 및 볼륨에 따라 크게 다를 수 있음

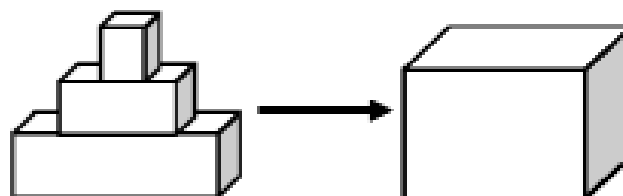


그림으로 보는 데이터 전처리 기법

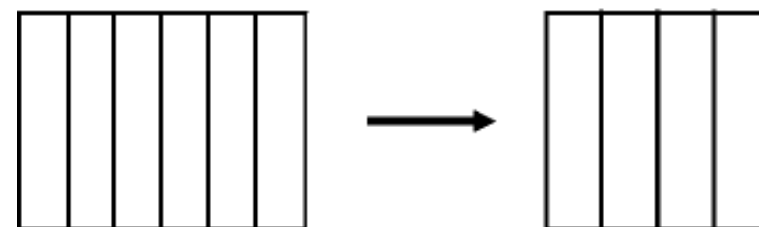
Data cleaning



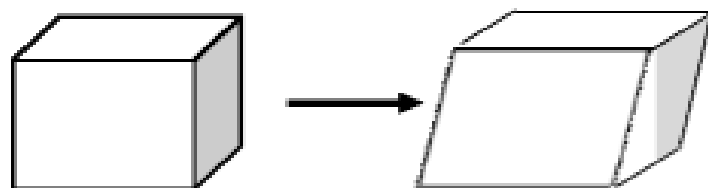
Data normalization



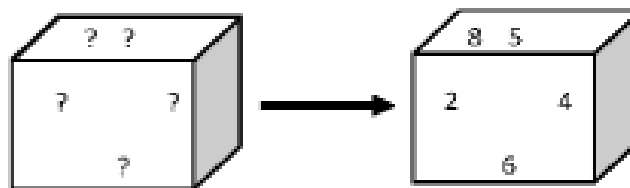
Feature selection



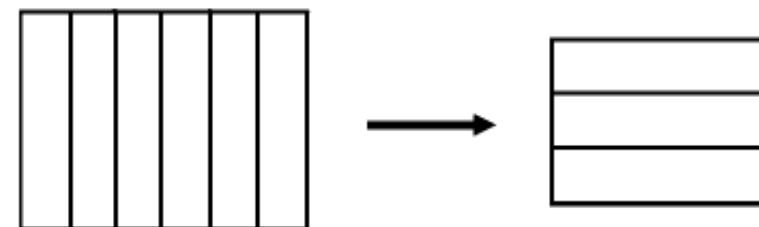
Data transformation



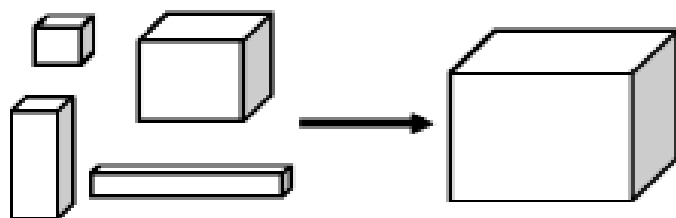
Missing values imputation



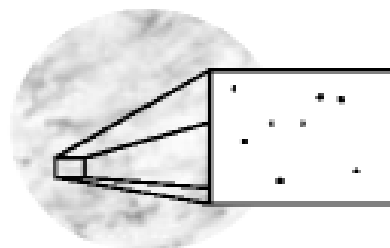
Instance selection



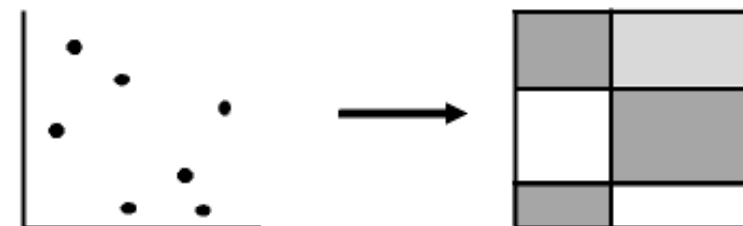
Data integration



Noise identification



Discretization



<https://bdataanalytics.biomedcentral.com/articles/10.1186/s41044-016-0014-0>

- 집계 Aggregation
- 샘플링 Sampling
- 차원 축소 Dimensionality Reduction
- 특징 선택 Feature subset selection
- 특징 생성 Feature creation
- 이산화와 이진화 Discretization and Binarization
- 속성 변환 Attribute Transformation
- ...



<https://medium.com/datadriveninvestor/data-cleaning-for-data-scientist-363fbbf87e5f>

Q & A