DI501 Introduction to Data Informatics

Lecture 4 – Data Preprocessing – Part I

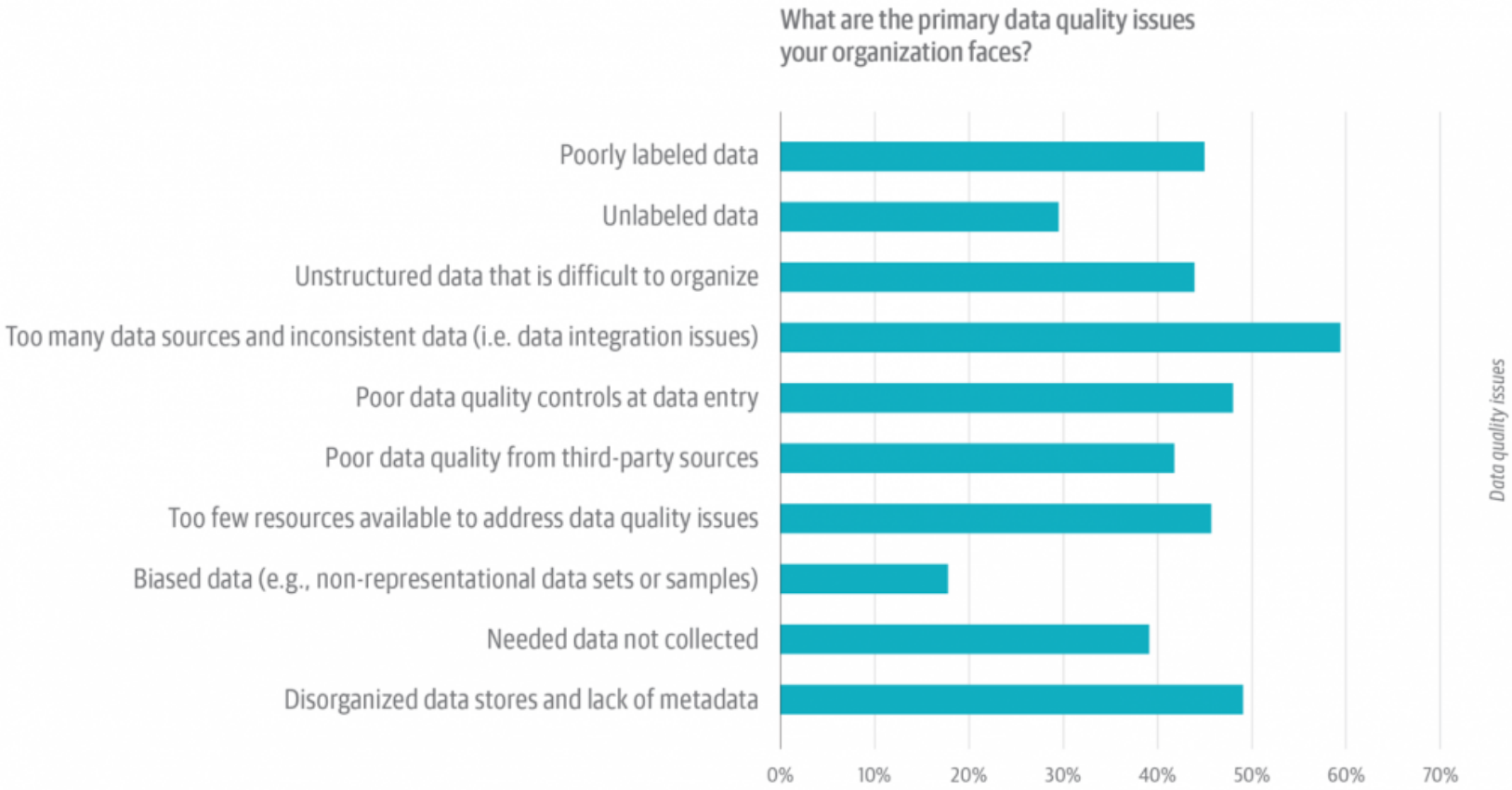
Introduction to Data Quality

Motivation

- Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.
- Low data quality results in significant problems in data analysis.
  - Incorrect conclusions can be drawn.

Motivation

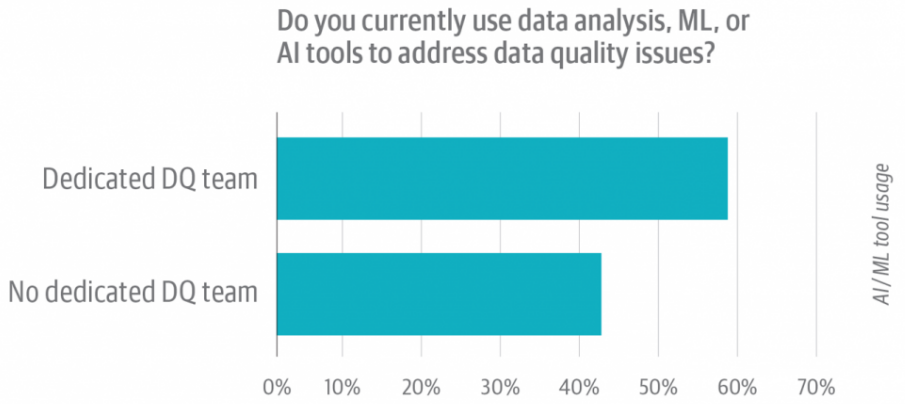
Data Quality Issues and Effects



Ref: https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/

Motivation

Data Quality Issues and Effects



Ref: https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/

Data Quality

- Data quality may be best defined as data that are fit for use by data consumers in their tasks (i.e., fitness for use).
  - The data must be accessible to the consumer.
    - E.g., the consumer knows how to retrieve data.
  - The consumer must be able to interpret data.
    - E.g., the data are not represented in a foreign language.
  - The data must be relevant to the consumer.
    - E.g., data are relevant and timely for use by the data consumer in the decision-making process.
  - The consumer must find the data accurate.
    - E.g., the data are correct, objective and come from reputable sources.

Data Quality Solutions

- **Data wrangling (aka data munging)**, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.
  - The goal of data wrangling is to assure quality and useful data. Data analysts typically spend the majority of their time in the process of data wrangling compared to the actual analysis of the data.

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

Data Quality

- Data profiling is an often-visual assessment that uses a toolbox of business rules and analytical algorithms to discover, understand and potentially expose inconsistencies in your data.
- This knowledge is then used to improve data quality as an important part of monitoring and improving the health of these newer, bigger data sets.
  - **Structure discovery,** also known as structure analysis, validates that the data that you have is consistent and formatted correctly.
  - **Content discovery** is the process of looking more closely into the individual elements of the database to check data quality.
  - **Relationship discovery** involves discovering what data is in use and trying to gain a better understanding of the connections between the data sets.

Ref: https://www.sas.com/en_us/insights/articles/data-management/what-is-data-profiling-and-how-does-it-make-big-data-easier.html

Data Quality

Data profiling methods:

- **Column profiling** provides statistical measurements associated with the frequency distribution of data values (and patterns) within a single column (or data attribute).
- The frequency distribution of column values exposes some insightful characteristics:
  - Range analysis
  - Sparseness
  - Cardinality
  - Uniqueness
  - Value distribution
  - Value absence

Ref: Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Data Quality

Data profiling methods:

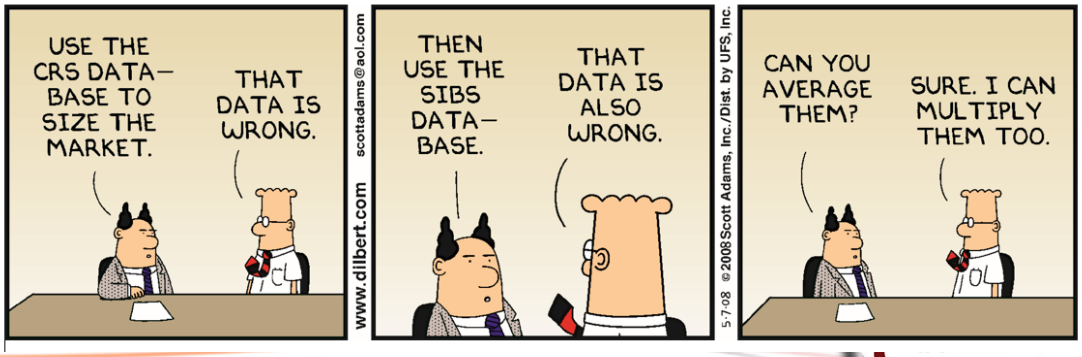
- **Column profiling:**
- The frequency analysis also provides summarization/aggregate values that can be used to characterize the data set, including:
  - Minimum value,
  - Maximum value,
  - Mean
  - Median
  - Standard deviation

Ref: Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Data Quality

Data profiling methods:

- **Column profiling:**
- Column profiling also looks at descriptive characteristics, feeding these types of analyses:
  - Type determination
  - Abstract type recognition
  - Overloading
  - Format evaluation



Ref: Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Data Quality

Data profiling methods:

- **Cross-column profiling** is made up of two processes: key analysis and dependency analysis.
  - Key analysis examines collections of attribute values by scouting for a possible primary key.
    - It is usually valid for older database systems
  - Dependency analysis is a more complex process that determines whether there are relationships or structures embedded in a data set.
    - Try to identify functional dependencies
- Computationally expensive.

Ref: https://www.sas.com/en_us/insights/articles/data-management/what-is-data-profiling-and-how-does-it-make-big-data-easier.html

Data Quality

Data profiling methods:

- **Cross-table profiling** uses foreign key analysis, which is the identification of orphaned records and determination of semantic and syntactic differences, to examine the relationships of column sets in different tables.
  - **Foreign key analysis**, which seeks to map key relationships between tables;
  - **Identification of orphaned records,** indicative of a foreign key relationship that is violated because a child entry exists where a parent record does not;
  - **Determination of semantic and syntactic differences**, such as when differently named columns holding the same values, or same-named columns hold different values.

Ref: Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Data Quality

Data profiling methods:

- **Data rule validation** verifies that data instances and data sets conform with predefined rules.
  - Helps to discover any anomalies.
  - Most data profiling tools allow for the definition of data rules, typically as assertions constructed from expressions.

Ref: Loshin, D. (2010). *The practitioner's guide to data quality improvement*. Elsevier.

Data Quality

Data profiling methods:

Cross-Column or Record Rule Examples

Dimension	Description	Example
Accuracy	A data element's value is accurate relative to a system of record when the value is dependent on other data element values for system of record lookup	Verify that the *last_name* field matches the system of record associated with the *customer_identifier* field
Accuracy	A data element's value is taken from a subset of a defined value domain based on other data attributes' values	Validate that the *purchaser_code* is valid for staff members based on *cost_center*
Consistency	One data element's value is consistent with other data elements' values	The *end_date* must be later than the *start_date*
Completeness	When other data element values observe a defined condition, a data element's value is not null	If *security_product_type* is "option" then the *underlier* field must not be null
Reasonableness	When other data element values observe a defined condition, a data element's value must conform to reasonable expectations	*Purchase_total* must be less than *credit_limit*
Currency	When other data element values observe a defined condition, a data element's value has been refreshed within the specified time period	If *last_payment_date* is after the *last_payment_due*, then refresh the *finance_charge*
Transformation	A data element's value is computed as a function of one or more other data attribute values	*Line_item_total* is calculated as *quantity* multiplied by *unit_price*

Ref: Loshin, D. (2010). *The practitioner's guide to data quality i*

Data Quality

Data profiling methods:

Examples of Table or Cross-Table Rules

Dimension	Description	Example
Accuracy	A data element's value is accurate when compared to a system of record when the value is dependent on other data element values for system of record lookup (including other tables)	The *telephone_number* for this office is equal to the *telephone_number* in the directory for this *office_identifier*
Consistency	One data element's value is consistent with other data elements' values (including other tables)	The *household_income* value is within 10% plus or minus the *median_income* value for homes within this *zip_code*
Completeness	When other data element values observe a defined condition, a data element's value is not null (including other tables)	Look up the customer's profile, and if *customer_status* is "preferred" then *discount* may not be null
Reasonableness	When other data element values observe a defined condition, a data element's value must conform to reasonable expectations (including other tables)	Today's *closing_price* should not be 2% more or less than the running average of closing prices for the past 30 days
Reasonableness	The value of a data attribute in one data instance must be reasonable in relation to other data instances in the same set	Flag any values of the *duration* attribute that are more than 2 times the standard deviation of all the *duration attribute values*
Currency	When other data element values observe a defined condition, a data element's value has been refreshed within the specified time period (including other tables)	Look up the product code in the supplier catalog, and if the price has been updated within the past 24 hours then *product_price* must be updated
Identifiability	A set of attribute values can be used to uniquely identify any entity within the data set	*Last_name, first_name, and SSN* can be used to uniquely identify any employee
Transformation	A data element's value is computed as a function of one or more other data attribute values (including other tables)	*Risk_score* can be computed based on values taken from multiple table lookups for a specific client application

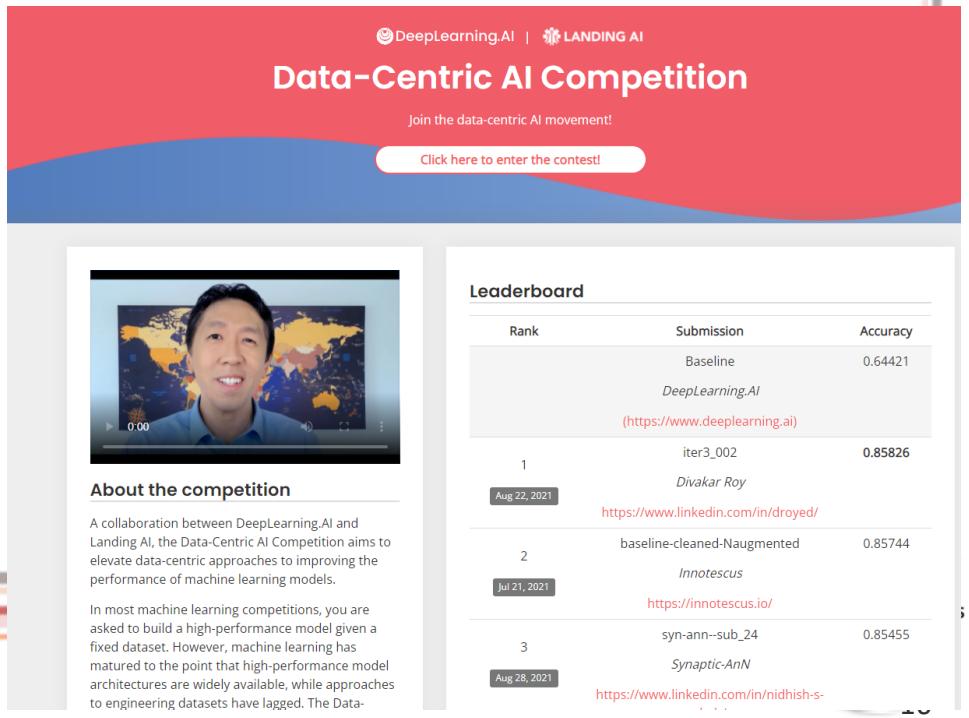
Ref: Loshin, D. (2010). *The practitioner's guide to data quality improv*

Data Quality

Important Terms

- There is a growing interest in Data-centric AI:
  - "Data is food for AI," says Andrew Ng, and he launched a campaign to shift the focus of AI practitioners from model/algorithm development to the quality of the data they use to train the models.

Ng observes that 80% of the AI developer's time is spent on data preparation.
This has been a widely shared estimate since the rise of "big data" in the late 2000s and the concomitant rise of "data scientists," known for their prowess in "data wrangling."



Data Quality

Data-centric AI Movement

- Data-Centric AI (DCAI) represents the recent transition from focusing on modeling to the underlying data used to train and evaluate models.
- Problems:
  - data collection/generation, data labeling, data preprocess/ augmentation, data quality evaluation, data debt, and data governance.
- Related term:
  - Data excellence: https://eval.how/dew2020/index.html

https://nips.cc/Conferences/2021/Schedule?showEvent=21860

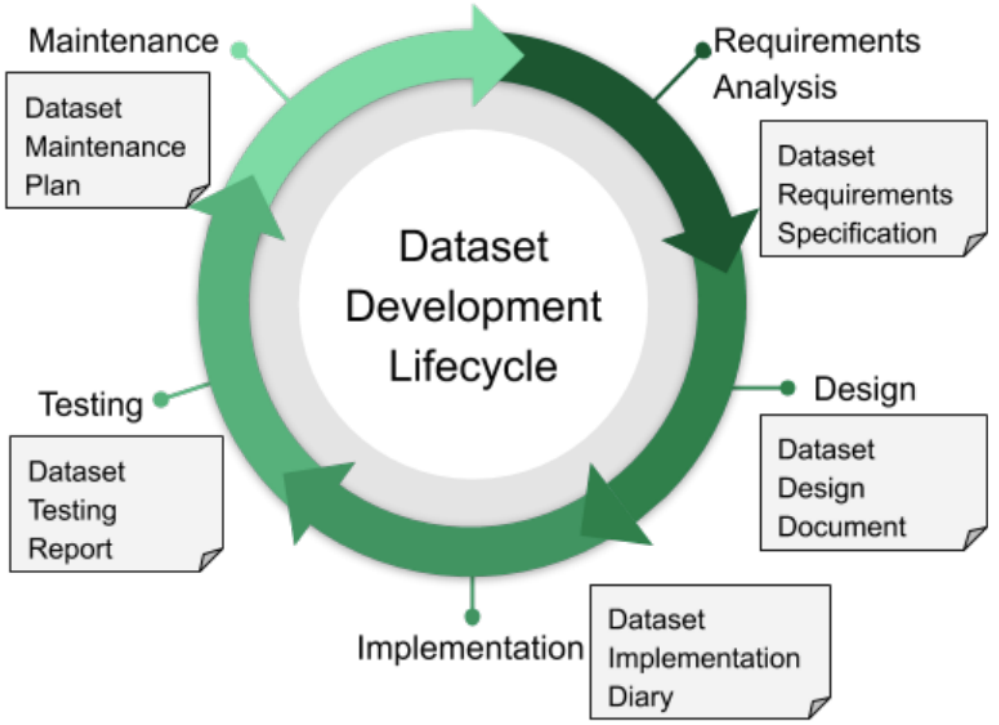
Problems with Machine Learning (ML) Datasets

- Bias-laundering: underrepresentation of minoritized groups and stereotype aligned correlations
- Reflecting historical patterns of social injustices
- Insufficient documentation and transparency regarding processes of dataset construction
- Problematic consent practices
- The lack of accountability to datafied and surveilled populations as well as groups impacted by data-driven decisions

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Dataset Development Cycle

- Documentation: a model of documentation practices throughout the dataset development lifecycle, drawing on software lifecycle practices.



Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Dataset Development Cycle

Documentation

Dataset Account	Questions Answered by the Document	Key Roles
Dataset Requirements Specification	Is data needed? By whom? What are its intended uses? What properties should it have? Do the uses necessitate constraints on collection and/or annotation?	Requirements owner; stakeholder; reviewer
Dataset Design Document	Is a new dataset needed (do existing datasets not meet requirements)? How will requirements be operationalized? What tradeoffs & assumptions are being made?	Design owner; domain expert; reviewer
Dataset Implementation Diary	How were the design decisions implemented? Why were they done this way? What unforeseen circumstances arose when implementing the design?	Implementation owner; data creator/labeler
Dataset Testing Report	Should I use the dataset? Does the dataset meets its requirements? Is the dataset safe to use? Is the dataset likely to have previously unforeseen consequences?	Data scientist; adversarial tester
Dataset Maintenance Plan	How will data staleness be detected and fixed? How will errors be fixed? How will affordances for manual interventions be provided?	Maintainer; funder; bug filer; data contester

Table 1: Critical document types for accountable dataset development. Each one is directly analogous to documentation types produced by the Software Development Lifecycle.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Dataset Development Cycle

Requirements

- Analogous to software requirements—covering both quantitative and qualitative factors.
- The needs of multiple stakeholders are collected and aggregated,
- Conflicts resolved through accountable mechanisms (including the keeping of accounts regarding the conflicts)

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Name of Dataset: Requirements Specification

Owner: *Name*; Created: *Date*; Last updated: *Date*

Vision

Brief summary of the envisioned data(set), its domains and scope.

Motivation

Problem and context that motivate why the data is needed.

Intended uses

Specific uses of the data that are intended.

Non-intended uses

What is the data not intended for? What should the data not be used for, and why?

Glossary of terms

If relevant, brief summary of acronyms and domain specific concepts for the general reader.

Related documents

List any related documents.

Data mocks

Include 2-3 typical examples of what the data instances should "look" like.

Stakeholders consulted

Whose needs were consulted and synthesised when creating this document? How were conflicting needs resolved?

Creation requirements

Where should the data come from? Include sources and collection methods

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Instance requirements

What requirements are there for data instances? Include any acceptable tradeoffs. Include numbers and types of instances, features, and labels.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Distributional requirements

What requirements are there for the distributions of your data? Include any acceptable tradeoffs. Include sampling requirements. If your data represents a set of people, describe who should be represented and in what numbers.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Data processing requirements

How should the data be annotated and filtered? Who should do the annotating? How should data be validated? Include any acceptable tradeoffs.

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Performance requirements

What can people who use this dataset for its intended uses expect?

- *Name of the requirement. Description.*
- *Name of the requirement. Description.*

Maintenance requirements

Should the data be regularly updated? If so, how often? For how long should the data be retained? Include any acceptable tradeoffs.

Sharing requirements

Should the data be made available to other teams within Google and/or open-sourced? If so, what constraints on data licensing, access, usage, and distribution are needed? Include any acceptable tradeoffs.

Caveats and risks

What would be the consequences of using data meeting the requirements described above?

Data ethics

Document your considerations of the ethical implications of the data and its collection.

Sign-off grid

Name	Role	Date
[redacted]		

Changelog

Editor	Comments	Date
[redacted]		

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Dataset Development Cycle

Design

- The primary account of this stage is the Dataset Design Document.
  - This document's primary roles are to lay out the plan of how requirements will be achieved, and to justify the design decisions that are made.
  - Issues to consider standardization, lossiness, sampling

Name of Dataset: Design Document

Owner: *Name*; Created: *Date*; Last updated: *Date*

Overview

High-level overview of the dataset.

Dataset Name: Name of dataset.

Primary Data Type(s): Primary data types; Eg: images, video, text.

Data Content: Eg. bounding boxes, image labels

Funding: How was the dataset funded?

Objective

What are the key objectives of the dataset? Is there a requirements specification?

Version

Current version; Differences to previous versions.

Background

Describe any relevant background of the dataset

Sources

System details; Where will the data come from? Selection and sampling criteria.

Annotations

Features and labels; Who are the annotators? How will they be trained?

Ratings: Rating tasks; Rating types; Rating procedures;

Data Quality

How is quality measured? How are metrics validated?

Characteristics

Characteristics of the dataset.

Expected Characteristics: Eg. How many instances, features, ratings.

Correlations: Acceptable correlations; Unacceptable correlations;

Acceptable and Unacceptable Conjunctional Datasets: Datasets that can and can not be used in conjunction with this dataset?

Population: Population represented.

Privacy Handling

How is privacy handled?

Maintenance

Who will maintain the dataset? Is there a maintenance plan? What are the recovery strategies if issues arise?

Sharing

Will the dataset be shared? How will access be controlled? How will the dataset be licensed?

Caveats

Describe known caveats

Data Ethics

Ethical considerations; Mitigation.

Work estimates

How much time will it take to collect the data; Costs are involved.

Related Datasets

Which existing datasests are related to this one? Why are they unsuitable?

Dataset Discovery Process: How did you search for other datasets?

Survey: High level overview of related datasets.

	Your Dataset	Other Dataset 1	Other Dataset 2
Documentation and DOI	Yours	Theirs	Theirs
Motivation and Intended use	Yours	Theirs	Theirs
Location	Yours	Theirs	Theirs
Size, Sampling and Filtering	Yours	Theirs	Theirs
Annotation and Labels	Yours	Theirs	Theirs
(Expected) Performance	How will your dataset improve on existing ones?	How does it not meet your requirements?	How does it not meet your requirements?
Examples	Example instance	Example instance	Example instance

Hutch
from s

y for machine learning datasets: Practices
arency (pp. 560-575).

Dataset Development Cycle

Testing

- A Dataset Testing Report is the account that specifies the evaluations that were done, as well as their results.
- Requirements testing (also called “acceptance testing”) checks a dataset against the stated requirements for its use.
- Requirements testing directly reframes the question of “how to evaluate whether the right data was collected with sufficient quantity”

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Dataset Development Cycle

Testing

- Adversarial testing of a dataset (also known as “test-to-fail” testing) aims to uncover unforeseen harms which may come from its use.
- Such risks vary widely in form, from harms to specific individuals such as privacy leaks, to harms to subgroups such as unforeseen correlations or stereotypes, to public relations risks via embarrassing data (including errors or omissions), to the possibility of malicious third parties using the data for nefarious ends.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Name of Dataset: Testing Report

Owner: *Name*; Created: *Date*; Last updated: *Date*

Summary

What is being tested?

Link to requirements specification.

Link to design document.

Meta-Testing

Is the data still needed? Are the data requirements still relevant and up-to-date?

Requirements Testing

Requirement tested	Results	Artifact
Requirement from requirements specification	Score or Results	Justification of the results or a link to artifact
Requirement from requirements specification	Score or Results	Justification of the results or a link to artifact
...	...	...

Untested Requirements

Untested Requirement	Reason for not testing
Requirement from requirements specification	Reason for not testing
Requirement from requirements specification	Reason for not testing
...	...

Adversarial Testing

Adversarial test	Results	Artifact
Describe tests	Score or Results	Justification of the results or a link to artifact
Describe tests	Score or Results	Justification of the results or a link to artifact
...	...	...

Other Testing

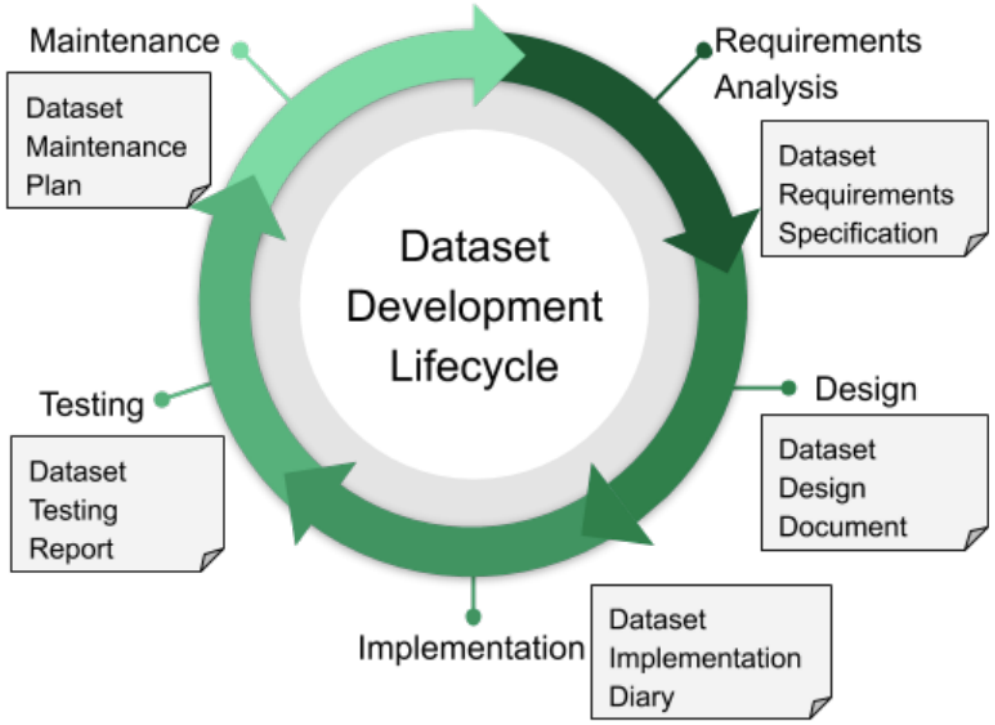
Test	Results	Artifact
Describe tests	Score or Results	Justification of the results or a link to artifact
Describe tests	Score or Results	Justification of the results or a link to artifact
...	...	...

Hutchinson
from softw

achine learning datasets: Practices
(pp. 560-575).

Dataset Development Cycle

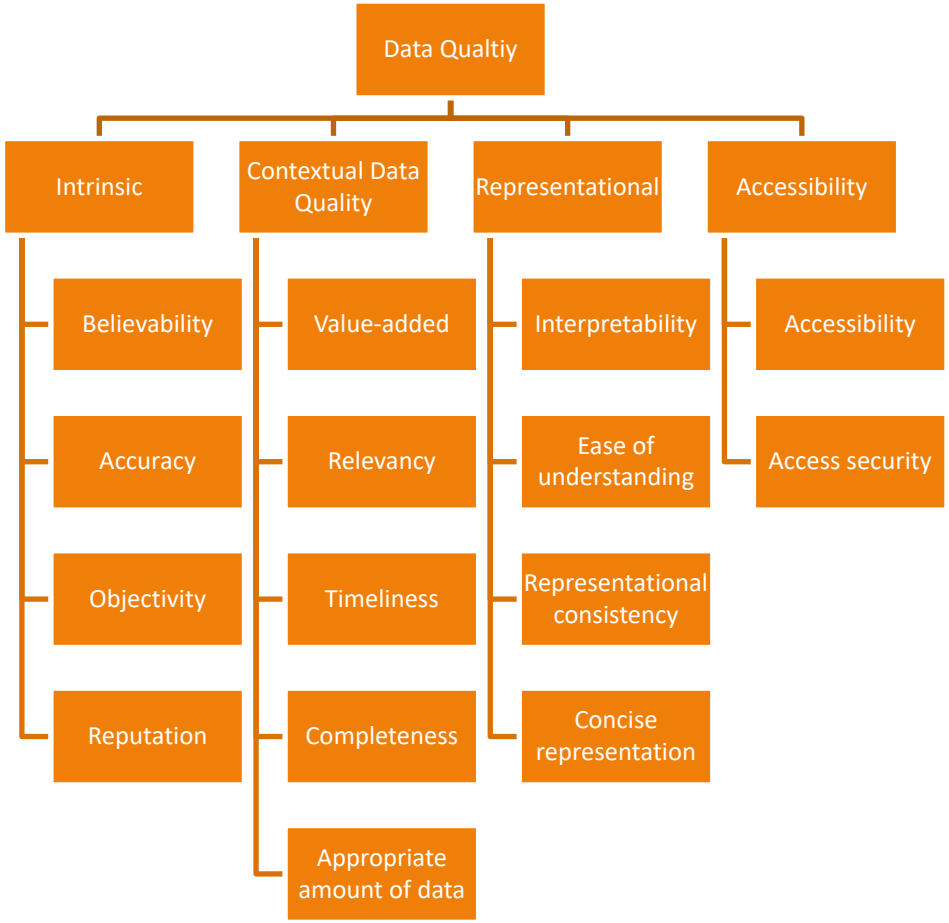
- Oversight: diverse oversight processes, including audits and reviews, which leverage these practices
- Maintenance: robust maintenance mechanisms, including those for addressing technical debt, correcting errors, postmortems and adapting to changing contexts



Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021, March). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 560-575).

Data Quality:

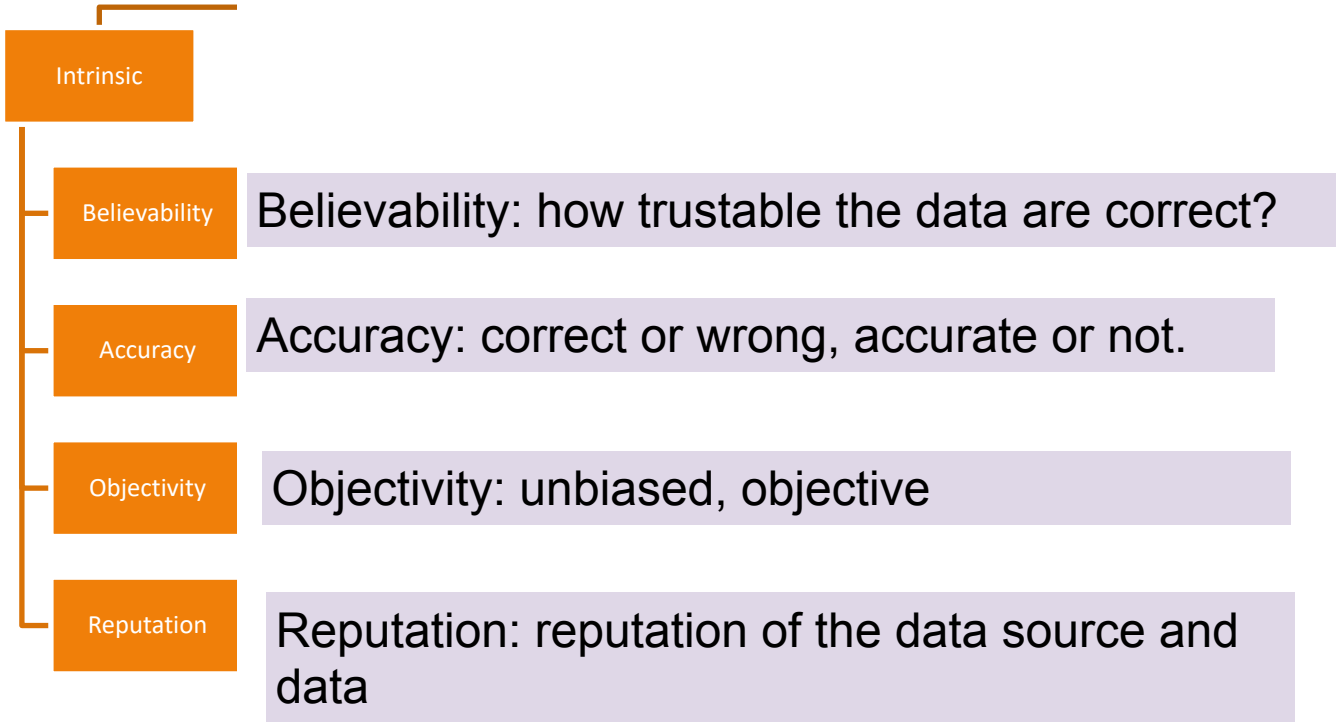
A Conceptual Framework



Accuracy, B. (1996). What Data Quality means to Data Consumers. *Journal of Management Systems*.

Data Quality:

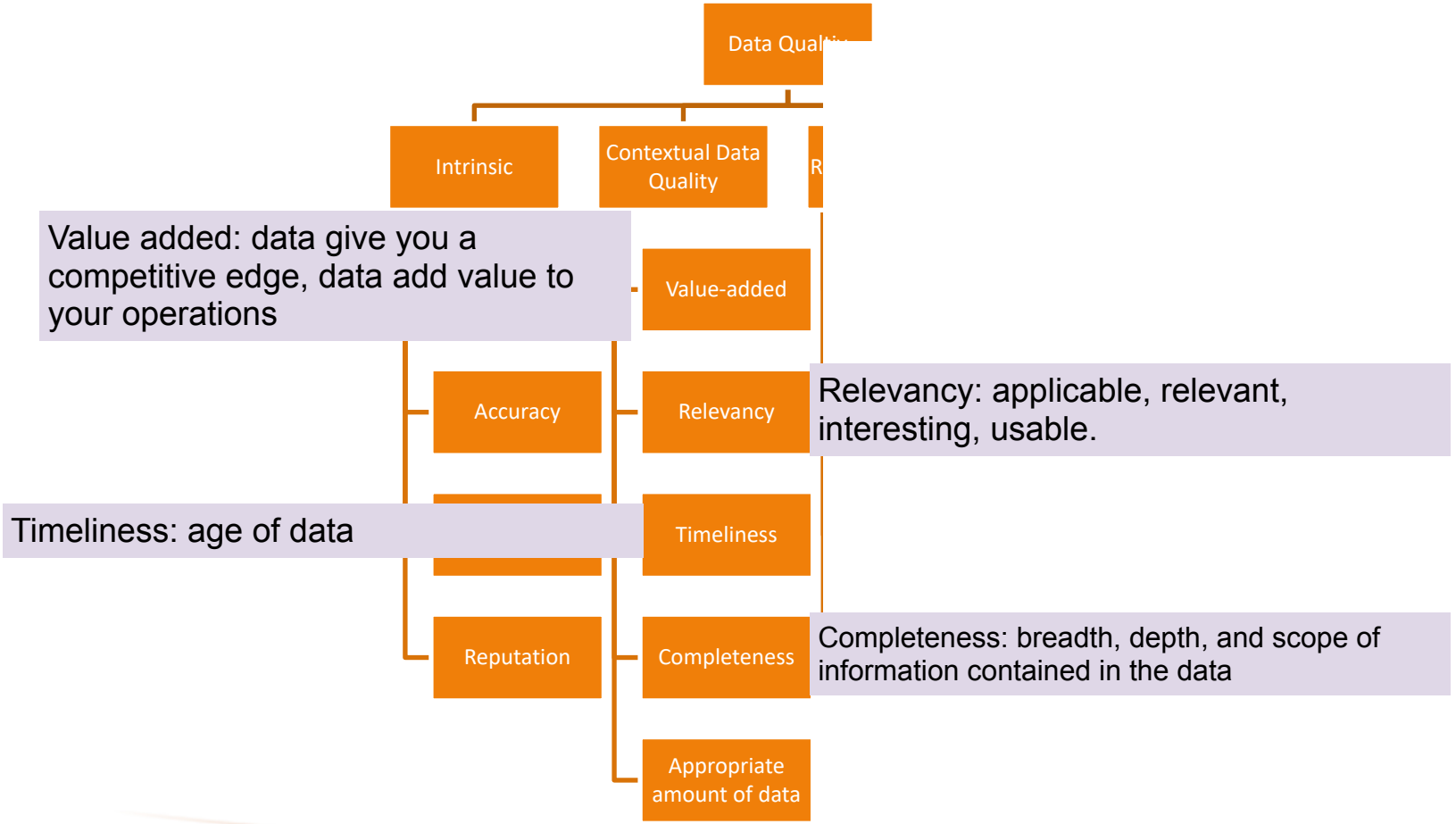
A Conceptual Framework



Accuracy, B. (1996). What Data Quality means to Data Consumers. *Journal of Management Systems*.

Data Quality:

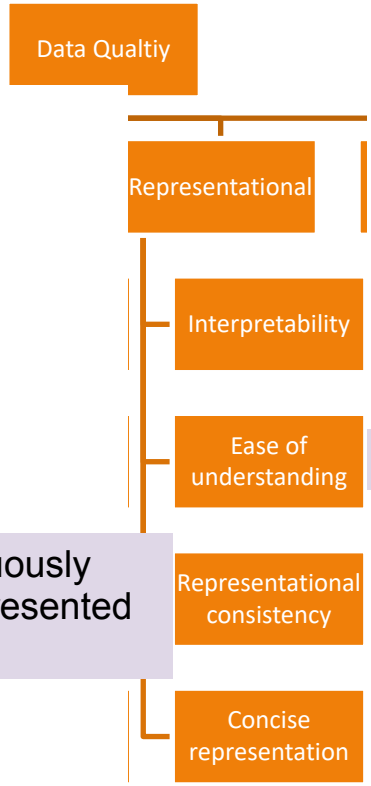
A Conceptual Framework



Accuracy, B. (1996). What Data Quality means to Data Consumers. *Journal of Management Systems*.

Data Quality:

A Conceptual Framework



Easily understood, clear and readable

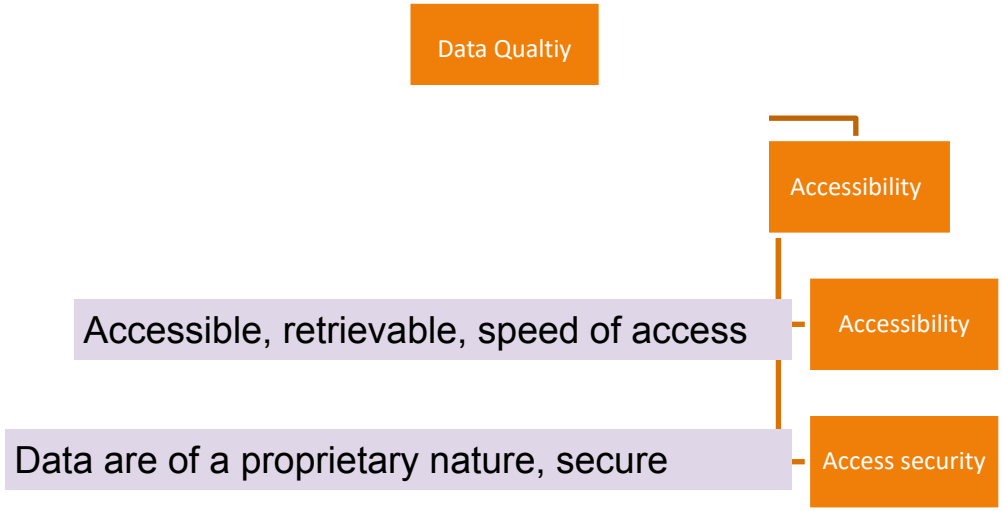
Representational consistency: Data are continuously presented in the same format, consistently represented and formatted, data are compatible with

Concise, well-organized, aesthetically pleasing, form of presentation

Accuracy, B. (1996). What Data Quality means to Data Consumers. *Journal of Management Systems*.

Data Quality:

A Conceptual Framework



Accuracy, B. (1996). What Data Quality means to Data Consumers. *Journal of Management Systems*.

Data Quality:

Tools for Measures

- **Deequ**, used internally at Amazon, verifies the quality of many large production datasets.
- It is a library built on top of Apache Spark.
- Dataset producers can add and edit data quality constraints.
- The system computes data quality metrics on a regular basis (with every new version of a dataset), verifies constraints defined by dataset producers, and publishes datasets to consumers in case of success.
  - In error cases, dataset publication can be stopped, and producers are notified to take action.
- Data quality issues do not propagate to consumer data pipelines, reducing their blast radius.

Ref: https://aws.amazon.com/blogs/big-data/test-data-quality-at-scale-with-deequ/

Data Quality:

Tools for Measures

- **Great Expectations** is a Python-based open-source library for validating, documenting, and profiling your data. It helps you to maintain data quality and improve communication about data between teams.
- An Expectation is a declarative statement that a computer can evaluate, and that is semantically meaningful to humans, like expect_column_values_to_be_unique or expect_column_mean_to_be_between.

Ref: https://greatexpectations.io/
https://medium.com/99p-labs/data-validation-measuring-completeness-consistency-and-accuracy-using-great-expectations-with-c0ad2924e425

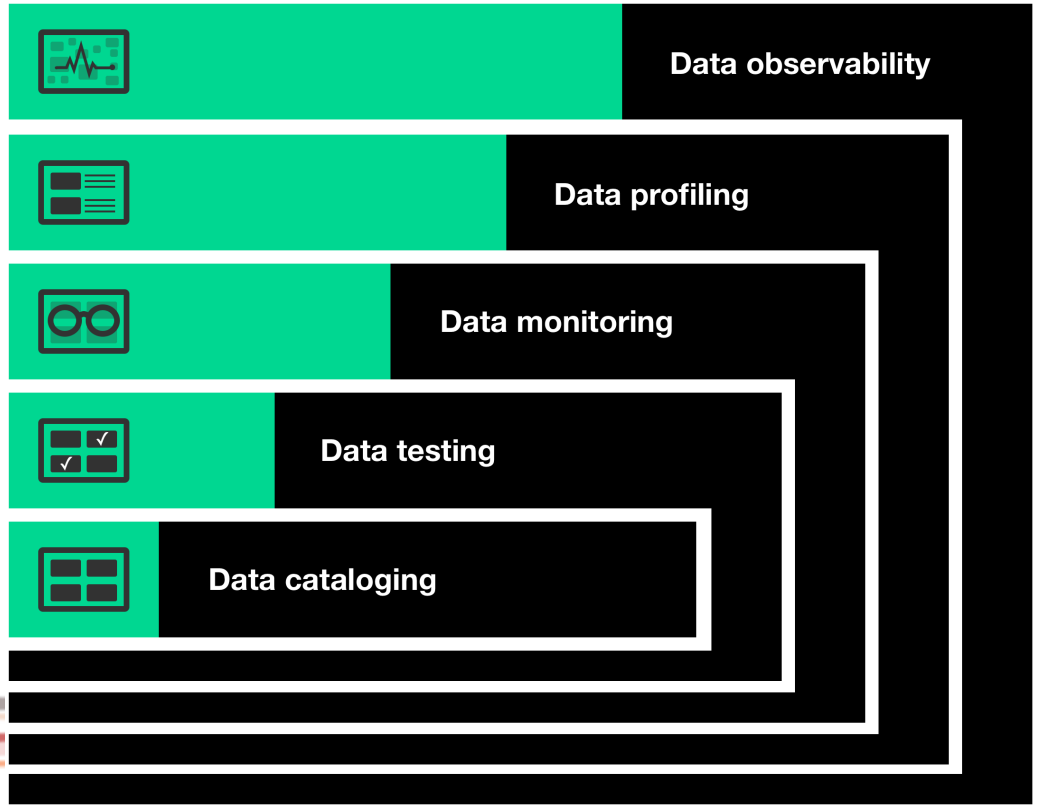
Data Quality:

Tools for Measures

- **SODA SQL and SODA Cloud:** a commercial solution targeting for ensuring data quality.

Soda SQL and **Soda Cloud** squarely address the challenges in testing, monitoring, profiling, and gaining observability into your data.

Use built-in metrics to define tests in Soda SQL or Soda Cloud that test data against quality thresholds and surface issues that occur throughout your data pipeline.



Ref: https://docs.soda.io/soda/data-monitoring.html

Data Quality:

Tools for Measures

- **Pytest** is used for writing unit tests against Python applications.
- It is also used for writing assertions to assert on the results of the queries, as a part of a data analytics pipeline.

```python
def test_column_difference():
    # test common cases
    test_df = pd.DataFrame([(1, 2), (3, 4)], columns=["A", "B"])
    test_df["A_minus_B"] = column_difference(test_df, col1="A", col2="B")
    assert all(test_df["A_minus_B"] == pd.Series([-1, -1]))
    test_df = pd.DataFrame([(5, 3), (10, 14), (0, -8)], columns=["A", "B"])
    test_df["A_minus_B"] = column_difference(test_df, col1="A", col2="B")
    assert all(test_df["A_minus_B"] == pd.Series([2, -4, 8]))
    # Include a third column
    test_df = pd.DataFrame([(1, 2, 100), (3, 4, 200)], columns=["A", "B", "C"])
    test_df["A_minus_B"] = column_difference(test_df, col1="A", col2="B")
    assert all(test_df["A_minus_B"] == pd.Series([-1, -1]))
    # Tets column of zeros
    test_df = pd.DataFrame([(1, 0), (3, 0)], columns=["A", "B"])
    test_df["A_minus_B"] = column_difference(test_df, col1="A", col2="B")
    assert all(test_df["A_minus_B"] == pd.Series([1, 3]))
    # Test empty dataframe
    test_df = pd.DataFrame(columns=["A", "B"])
    test_df["A_minus_B"] = column_difference(test_df, col1="A", col2="B")
    assert all(test_df["A_minus_B"] == pd.Series([]))
```

Ref: https://www.statestitle.com/resource/pytest-for-data-scientists/

Data Quality:

Tools for Measures

- **ETL tools** have built in solutions.
- ETL tools allow users to specify transforms through a graphical user interface (GUI).
  - ETL was used to blend data from multiple sources and load it into various targets.
    - It's often used to build a data warehouse.
    - Target may also be a big data platform such as Hadoop.
  - ETL can combine and surface transaction data from data stores so that it’s ready for business people to view in a format they can understand.
  - ETL is also used to migrate data from legacy systems to modern systems with different data formats.
  - It’s often used to consolidate data from business mergers, and to collect and join data from external suppliers or partners.

Ref: https://www.statestitle.com/resource/pytest-for-data-scientists/

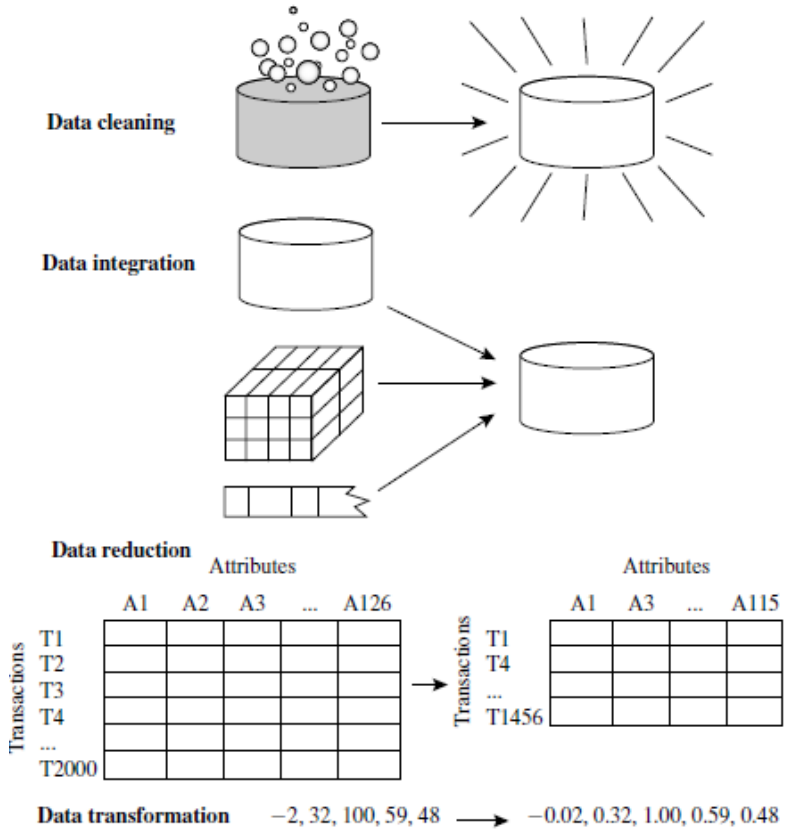
DI501 Introduction to Data Informatics

Lecture 4 – Data Preprocessing – Part 2

Data Preprocessing Tasks

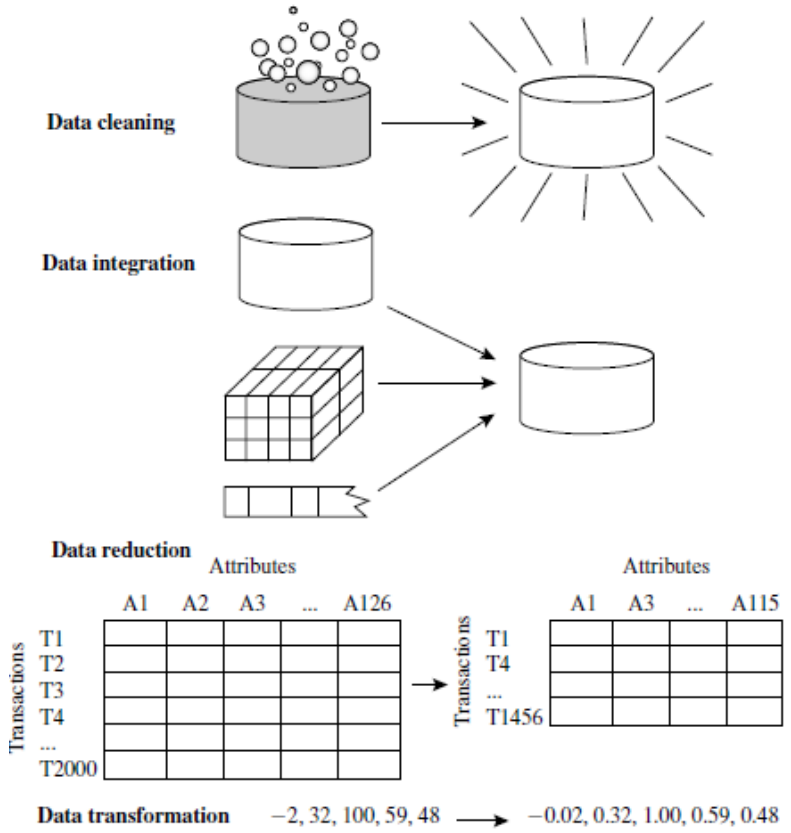
Data Preprocessing Tasks

- Data cleaning can be applied to remove noise and correct inconsistencies in data.
- Data integration merges data from multiple sources into a coherent data store such as a data warehouse.
- Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.



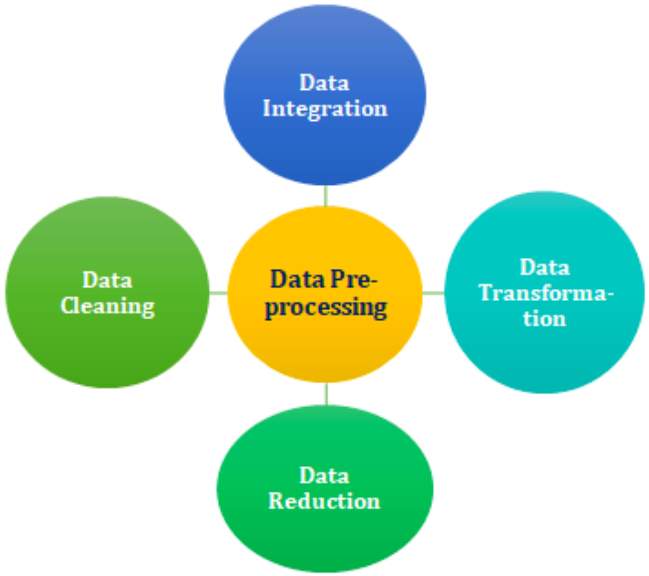
Data Preprocessing Tasks

- Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0.
  - This can improve the accuracy and efficiency of mining algorithms involving distance measurements.
- **These techniques are not mutually exclusive; they may work together.**



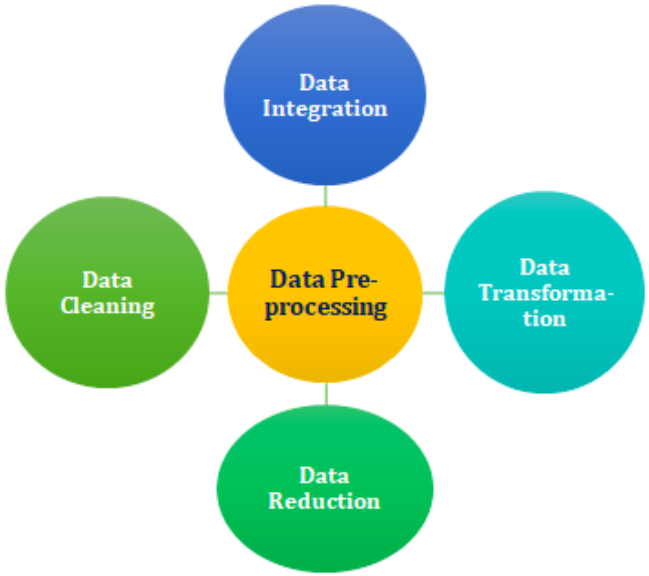
Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, data cubes, or files
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization
  - Concept hierarchy generation



Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, data cubes, or files
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization
  - Concept hierarchy generation



Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
  - Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set.
- The semantic heterogeneity and structure of data pose great challenges in data integration.
  - How can we match schema and objects from different sources? This is the essence of the entity identification problem.
  - Are any attributes correlated? Duplication causes problems.
  - How can we detect and resolve data value conflicts?

Data Integration

- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Data Integration

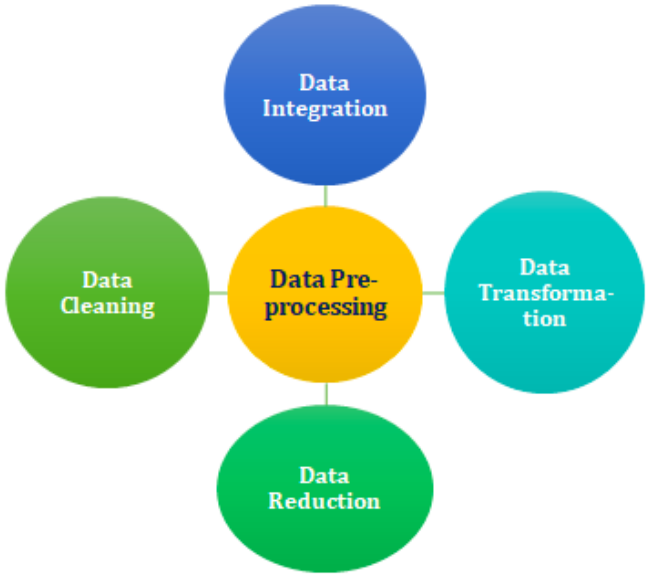
- Redundant data occur often when data in multiple databases are integrated.
  - Object identification: The same attribute or object may have different names in different databases.
  - Derivable data: One attribute may be a “derived” attribute in another table, e.g., annual revenue.
- Redundant attributes may be detected by correlation analysis and covariance analysis (feature selection-variable ranking).
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality.

Data Integration

- Feature reduction:
  - Eliminating the features which are highly correlated with others.
  - Redesigning features.
- Feature selection:
  - Selecting the most appropriate subset of attributes.
- Attribute generation:
  - Transforming a number of features into a new attribute.

Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, data cubes, or files
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization
  - Concept hierarchy generation



Data Cleaning

- Data in the real world is dirty: Lots of potentially incorrect data, e.g., instrument faults, human or computer errors, transmission errors
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., Occupation=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., Salary=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names,
    - e.g., Age=“42”, Birthday=“03/07/2010”
    - e.g., Was rating “1, 2, 3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records
  - Intentional (e.g., *disguised missing data*)
    - e.g., Jan. 1 as everyone’s birthday?

Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Data Quality:

Problem Detection

- Detection of inconsistencies
  - Data manager and administrator who are experienced in the domain work on the data.
- If your familiarity with the data of the domain is limited:
  - Look for any explicit errors
  - Look for any provided domain dependent business rules
  - Look for any outliers
  - Look for any disguised missing data
  - Look for any missing data

Data Cleaning for Explicit Errors

- Data which can be analyzed and corrected using business rules and domain knowledge (metadata):
  - For instance,
    - what are the data type and domain of each attribute?
    - what are the acceptable values for each attribute?
  - Examples:
    - faulty instrument: A rule can identify certain values within a range and discard the others
    - human or computer error: use domain knowledge to identify them
      - e.g., Salary = "−10" (an error)

Data Cleaning for Explicit Errors

- Data inconsistencies emerged due to the discrepancies between records
  - This may indicate a normalization problem in the database
  - Use business rules to detect them
    - Age = “42”, Birthday = “03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
- Look for the inconsistent use of codes and any inconsistent data representations
  - e.g., “2010/12/25” and “25/12/2010” for date.
- Field overloading is another error source that typically results when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes
  - e.g., an unused bit of an attribute that has a value range that uses only, say, 31 out of 32 bits.

Data Cleaning for Explicit Errors

- The data should also be examined regarding unique rules, consecutive rules, and null rules.
  - A unique rule says that each value of the given attribute must be different from all other values for that attribute.
  - A consecutive rule says that there can be no missing values between the lowest and highest values for the attribute, and that all values must also be unique (e.g., as in check numbers).
  - A null rule specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values should be handled.

Data Cleaning for Explicit Errors

- There are a number of different commercial tools that can aid in the discrepancy detection step.
  - Data scrubbing tools use simple domain knowledge (e.g., knowledge of postal addresses and spell-checking) to detect errors and make corrections in the data. These tools rely on parsing and fuzzy matching techniques when cleaning data from multiple sources.
  - Data auditing tools find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions. (e.g., correlation and clustering to find outliers)

Rebecca	by Daphne du Maurier (Mass Market Paperback)	$6.29	****
Sonnet 19.	Craig W.J., ed. 1914. The Oxford Shakespeare		
The Big Four	Agatha Christie, Mass market paperback	5.39	10%

Data Cleaning for Explicit Errors

- Most errors, however, will require data transformations.
- Once we find discrepancies, we typically need to define and apply (a series of) transformations to correct them.
- Commercial tools can assist in the data transformation step.
- Data migration tools allow simple transformations to be specified such as to replace the string “gender” by “sex.”

Noisy Data

- Noise is a random error or variance in a measured variable.
- Incorrect attribute values may be due to:
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

How to Handle Noisy Data

- Given a numeric attribute such as, say, price, how can we “smooth” out the data to remove the noise?
  - Binning
  - Regression
  - Outlier analysis
  - Concept hierarchies are a form of data discretization that can also be used for data smoothing.

How to Handle Noisy Data?

- Binning:
  - First sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

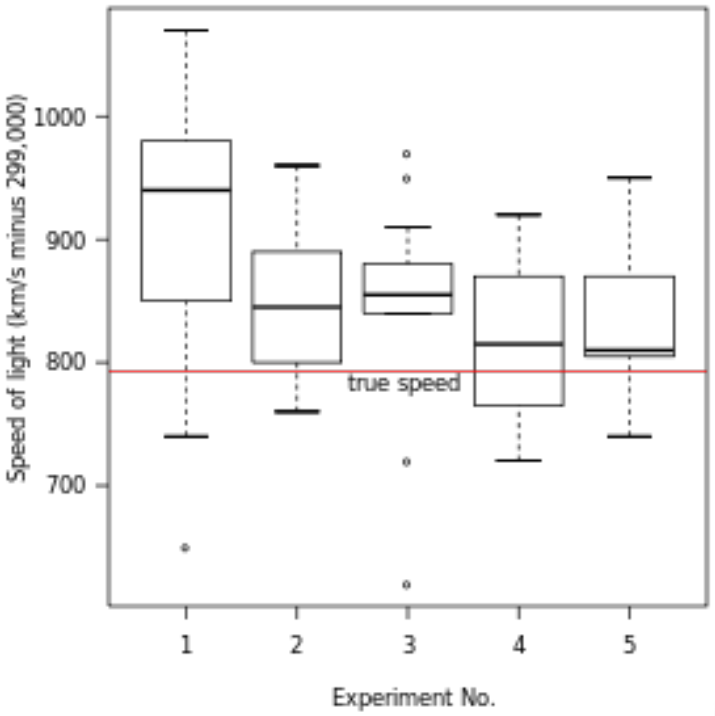
Partition into (equal-frequency) bins:
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

How to Handle Noisy Data?

- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)
- Box plots
  - Discard outliers?
    - Outliers are different from the noise data.
      - Noise may distort the normal objects and blur the distinction between normal objects and outliers.



How to Handle Noisy Data?

- How to identify outliers?
  - An outlier is an unusual score, relative to the data set. It will influence the mean and the standard deviation.
  - A data set might have no outliers, one outlier, or several outliers.
  - One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.

How to Handle Noisy Data?

- How to identify outliers?
  - An outlier is an unusual score, relative to the data set. It will influence the mean and the standard deviation.
  - A data set might have no outliers, one outlier, or several outliers.
  - One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.

RECALL: The first quartile, denoted Q_1, is the value in the data set that holds 25% of the values ***below*** it.
The third quartile, denoted Q_3, is the value in the data set that holds 25% of the values ***above*** it.
Interquartile Range = Q_3-Q_1

How to Handle Noisy Data?

- Filters
  - Moving average, median filters
  - Used in digital signal processing and time series analysis.
    - Well-known examples are Henderson moving average, Spencer's 15-point moving average (convolution of four filters)

EQUATION 15-1
Equation of the moving average filter. In this equation, $x[\,]$ is the input signal, $y[\,]$ is the output signal, and M is the number of points used in the moving average. This equation only uses points on *one side* of the output sample being calculated.

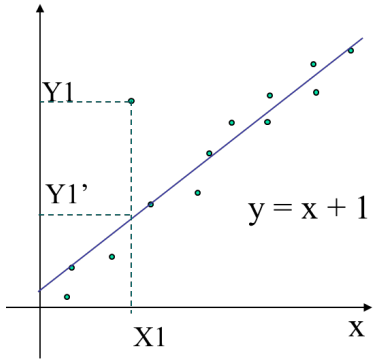
$$y[i] = \frac{1}{M} \sum_{j=0}^{M-1} x[i+j]$$

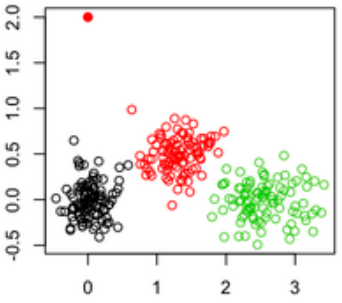
Where $x[\,]$ is the input signal, $y[\,]$ is the output signal, and M is the number of points in the average. For example, in a 5 point moving average filter, point 80 in the output signal is given by:

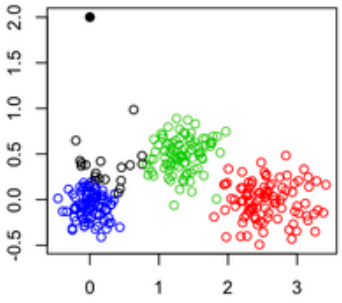
$$y[80] = \frac{x[80] + x[81] + x[82] + x[83] + x[84]}{5}$$

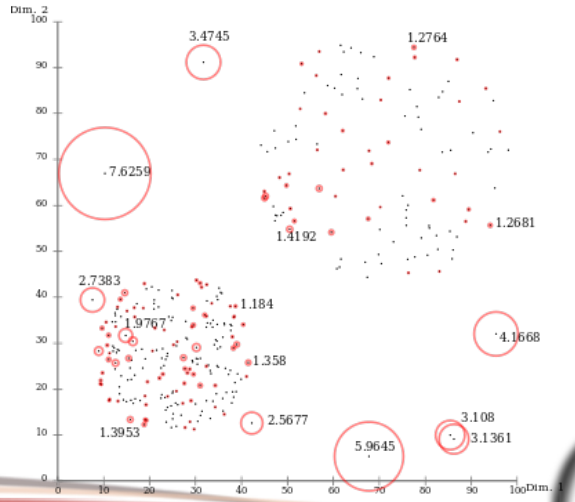
How to Handle Noisy Data?

- Regression
  - smooth by fitting the data into regression functions.
- Clustering
  - detect and remove outliers by clustering.







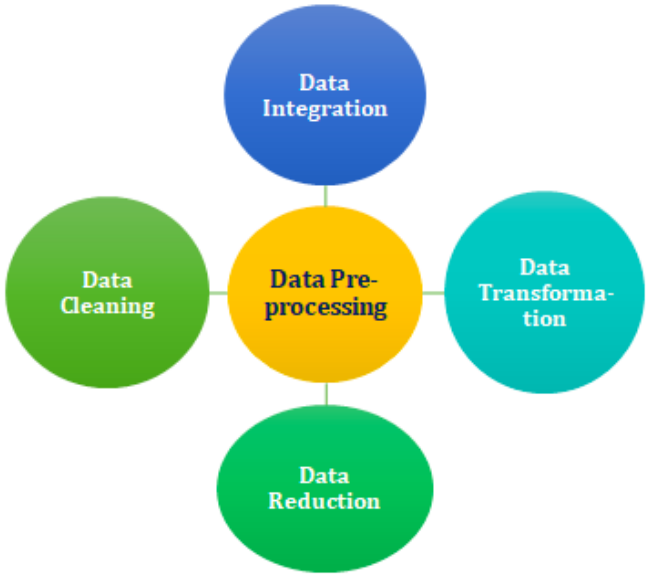


DI501 Introduction to Data Informatics

Lecture 4 – Data Preprocessing – Part III

Major Tasks in Data Preprocessing

- Data integration
  - Integration of multiple databases, data cubes, or files
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data reduction
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- Data transformation and data discretization
  - Normalization
  - Concept hierarchy generation



Missing Data

- Data science methods are usually applied on rectangular data sets.
  - Rows represent units (AKA cases, observations, or subjects depending on context)
  - Columns represent variables (AKA attributes, features, characteristics) measured for each unit.
  - Entries are real numbers or categories (ordered or unordered).
- Missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value.

Missing Data

- Missing-Data: Data that are missing for some variables (not all) for some cases (not all).
- Latent (or unobserved) Data: Data that are missing on a variable for all cases.
- Unit non-response: Data that are missing on all variables for some cases.
- Disguised Data: Missing data values are not explicitly represented as such, but are coded as valid data values.
  - Can be treated as missing data.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	1	6		2		4	A	T
4	2	2		1				F
5	5	1		4		3		T
2	1	2		1		4		F
3	2	3		7		5	B	
4	2	4		8			D	F
1	3	1		6			D	
6	4	2		4			C	F
2	5	3		1		3	D	T
3	1	4				2		F
8	3			3			C	T

Missing Data

- Conventional tools and techniques presume that all variables in a specified model are measured for all cases.
  - The default method applied is simply to delete cases (or variables) with any missing data on the variables (or cases) of interest, a method known as complete case analysis.
  - The most obvious drawback of the complete case analysis is that it often deletes a large fraction of the sample, leading to a severe loss of information and biased estimates and loss of statistical power.
  - We also have imputation methods that improve the statistical power if certain assumptions are met.

Why are the Values Missing?

- By Design
  - May be completely at random
    - 50% of units selected randomly for each interview
    - 50% randomly selected for follow-up
    - Effective when there are too many units or high costs
- Intentionally Missing—Researcher controlled
  - Boys not asked when first menstruation
  - Drop some units from analysis
  - Sometimes unintentionally imputed
  - Imputing doesn't necessarily hurt

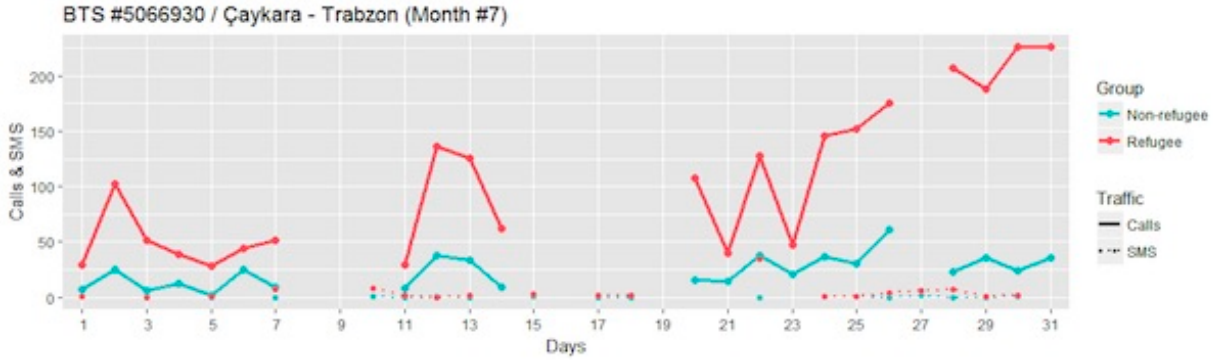
Why are the Values Missing?

- Refusals—We may know mechanism
  - Adjusted for gender, race, education
  - May be missing at random
  - Otherwise, bias is likely without auxiliary variables
    - E.g., income may be missing. We may correlate variables such as education to explain missingness.
- Missing because of "don't know" responses
  - Between agree and disagree?
  - Can we impute a better value? Should we?

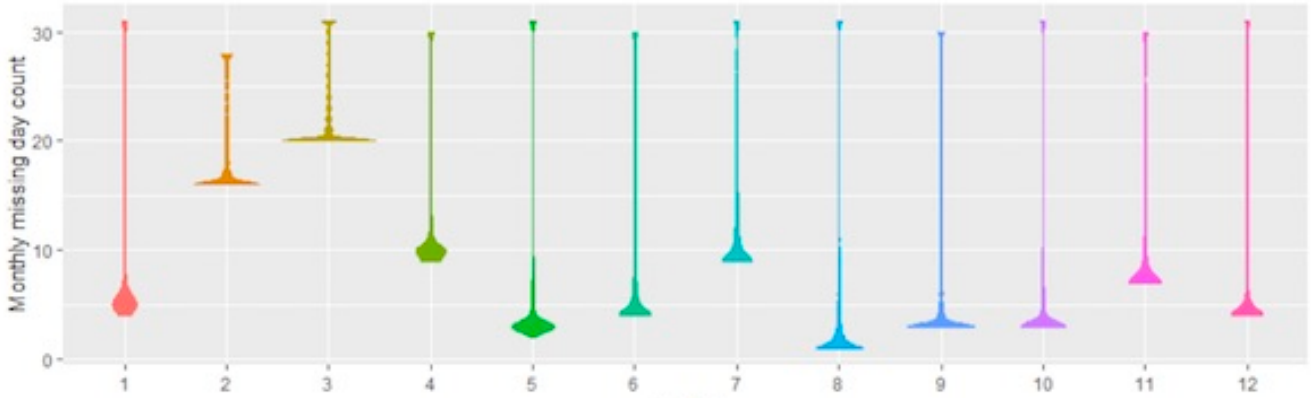
Why are the Values Missing?

- Missing by researcher error
  - May be missing completely at random
  - May reflect researcher bias
  - Missing observation worse than missing value
- Sampling problems
- Code reason value is missing
  - NLSY97, uses 5 types of missing values
  - Treat each differently

Why are the Values Missing?



Data is missing due to sampling errors in a well-known mobile call data records dataset.



Violin plots indicating the number of days in a month that has no record for a specific base transceiver receiver (BTS) in the telecom dataset, grouped by months.

Ref: http://d4r.metu.edu.tr/dataquality.html

Why are the Values Missing?

- Variables may contain missing values for several reasons. The RAND HRS (Health and Retirement Study) data set's missing data codes are as follows:

  MISSING CODES:

Code	Reason for missing
.	Reference person did not respond to this wave
.D	Don't know
.R	Refused
.X	Does not apply (specifics depend on variable)
.Q	Data not available because of HRS and AHEAD survey instrument differences in Wave 2 or 3
.U	Reference person is not married (for spouse variables)
.V	Spouse did not respond this wave (for spousal variables)
.S	Information not available due to skip patterns, typically because the interview is by proxy respondent
.M	Other missing

General Steps for Analysis with Missing Data

1. Identify patterns/reasons for missingness and recode correctly
2. Understand distribution of missing data
3. Decide on best method of analysis

Ref: Missing Data & How to deal: An overview of missing data by Melissa Humphries

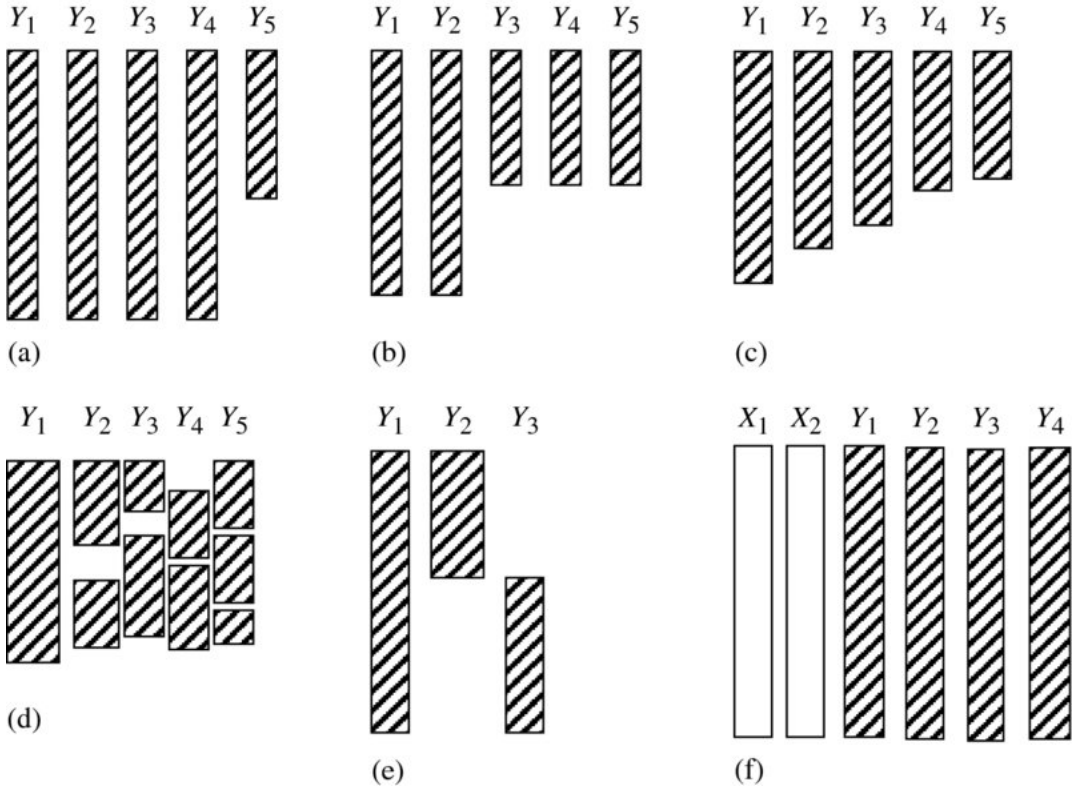
Step 1: Understand Your Data

- Attrition due to social/natural processes
  - Example: School graduation, dropout, death
- Skip pattern in survey
  - Example: Certain questions only asked to respondents who indicate they are married
- Intentionally missing as part of data collection process
- Random data collection issues
- Respondent refusal/Non-response
  - Identify skip patterns and/or sampling strategy from documentation

Missingness Pattern and Missingness Mechanism

- Before dealing with missing data, we have to understand why each value is missing.
- Accordingly, we can impute values or we can delete observations or variables where you do not intend to impute a value
- Missingness pattern describes which values are missing and which values are observed in the data matrix.
  - It describes the location of the missing values and not the reasons for missingness.
- Missingness mechanism (or mechanisms) concerns the relationship between missingness and the values of variables in the data matrix.

Missingness Patterns



a) Univariate nonresponse

b) Unit nonresponse pattern

c) Monotone

d) General

e) File matching

f) Factor analysis, with two factors and four measured variables.

Step 1: Understand Your Data

Patterns of Missing Values: Example

MISSING DATA PATTERNS

	1	2	3	4	5	6	7	8	9	10
HLTH	x	x	x	x						
CHILDS	x	x	x	x	x	x	x	x	x	x
HAP_GEN	x	x			x	x	x			
INCOME98	x		x		x	x		x	x	
AGE	x	x	x	x	x		x	x		x
EDUC	x	x			x	x	x			

- What is problem with
  - HLTH? (Health)
  - INCOME98?
  - EDUC?
  - HAP_GEN (General happiness)

X indicates full.

Empty indicates at least one tuple in the database does not have this attribute.

Step 1: Understand Your Data

Patterns of Missing Values

- Throw out 81 people in pattern 2?
  - We have data on five of the six variables
  - Income might not be a key predictor
- Why is health missing in patterns 5 to 10—Was this by design?

550 individuals have all six attributes.
81 have five attributes: only INCOME98 is missing in these tuples.
77 instances do not have HAP_GEN and EDUC but have the other four attributes.
…

PATTERN FREQUENCIES

Pattern	Freq
1	550
2	81
3	77
4	30
5	27
6	2
7	12
8	21
9	4
10	14

Sum:818

Step 1: Understand Your Data

Amount of Missing Values

PROPORTION OF DATA PRESENT

	HLTH	CHILDS	HAP_GEN	INC	AGE	EDUC
HLTH	.90					
CHILDS	.90	1.00				
HAP_GEN	.77	.82	.82			
INCOME98	.76	.83	.70	.83		
AGE	.90	.99	.81	.82	.99	
EDUC	.77	.82	.82	.70	.81	.822

	1	2	3	4	5	6	7	8	9	10
HLTH	x	x	x	x						

Pattern	Freq
1	550
2	81
3	77
4	30
5	27
6	2
7	12
8	21
9	4
10	14

Sum:818

Let's construct the proportion tables:
Example for HLTH:
Sum of frequencies of pattern 1-4: 738
So HLTH presence proportion is 738/818=.9022

Step 2: Missing Data Mechanism

- Consider the probability of *missingness*
  - Are *certain groups* more likely to have missing values?
    - Example: Respondents in service occupations less likely to report income
  - Are *certain responses* more likely to be missing?
    - Example: Respondents with high income less likely to report income
- Certain analysis methods assume a certain probability distribution.

Step 2: Missing Data Mechanisms

- Missing Completely at Random (MCAR)
  - The probability that a unit is missing a variable is completely independent of the value of the variable and of the values of other variables.
- Missing at Random (MAR)
  - The probability that a variable is missing for a unit depends on the value of an other variable(s) for that unit but not on the value of the missing variable itself.
- Missing not at Random (MNAR)
  - The probability that a variable is missing for a unit depends on the value of the variable itself.

Step 2: Missing Data Mechanisms

- Missing Completely at Random (MCAR)
  - The probability that a unit is missing a variable is completely independent of the value of the variable and of the values of other variables.
- Missing at Random (MAR)
  - The probability that a variable is missing for a unit depends on the value of an other variable(s) for that unit but not on the value of the missing variable itself.
- Missing not at Random (MNAR)
  - The probability that a variable is missing for a unit depends on the value of the variable itself.

Step 2: Missing Data Mechanism

Missing Completely at Random (MCAR)

- Suppose we have *p* variables. Data can be characterized as MCAR if the missingness, or the probability of missing data, on an outcome variable $X_j(j=1,...,j,...p)$ is unrelated
  a) to the value of X_j itself or
  b) to values on any of the remaining *(p-1)* variables in the data.
- Note that MCAR allows for the possibility that missingness on one variable may be related to missingness on another.
  - For example, two or more variables may always be missing together or observed together. This typically occurs when data are integrated from multiple sources.

Step 2: Missing Data Mechanism

Missing Completely at Random (MCAR)

- If data are MCAR, complete data sample is a random subsample of original target sample.
- MCAR is the strongest assumption that is commonly made and also the best situation to be in.
- For most data sets, the MCAR assumption is unlikely to be precisely satisfied.
  - One situation in which the assumption is likely to be satisfied is when data are missing by design.
    - For example, a researcher may decide that a brain scan is just too costly to administer to everyone in her study. Instead, she does the scan for only a 25% random subsample. For the remaining 75%, the brain-scan data are MCAR.

Step 2: Missing Data Mechanism

Testing MCAR Assumption

- Assumption - Part I: the probability that any variable is missing cannot depend on the value of any other variable in the model of interest.
  - For example,
    - Apply t-test for differences in means of X_i (i=1,...,j-1,j+1,...p) between those having missing X_j and those having observed X_j,
    - We can also employ chi-squared test to check frequencies
    - Generate classification models that uses X_i (i=1,...,j-1,j+1,...p) to predict missingness on X_j and check the classification performance.
- Assumption - Part II: the probability that X_j is missing cannot depend on the value of X_j itself.
  - It is not possible to test this because we don't know the missing values.

Raw Data

ID	Var1	Var2	Var3
1	9	7	.
2	.	3	5
3	7	4	.
4	9	4	6
5	6	2	7
6	.	.	5

Missingness Ind.

ID	D1	D2	D3
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	0	0
6	1	1	0

Step 2: Missing Data Mechanism

Missing Completely at Random (MCAR)

- Is the missingness random?
  - D1, D2, D3 should be uncorrelated with anything else observed!
    - D1 is not correlated with Var2 and Var3
    - D2 is not correlated with Var1 and Var3
    - D3 is not correlated with Var1 and Var2
- We can, for example, fit logistic regression models and check the prediction performance

Step 2: Missing Data Mechanism

Missing Completely at Random (MCAR)

- Is the missingness random?
  - D1, D2, D3 should be uncorrelated with anything else observed!

Example:

- Suppose V1=gender and V3=pregnancy.
- Male individuals are not required to answer pregnancy related questions.
- Hence, if gender=male then D3=1 as they are missing.
- The correlation between gender (V1) and D3 will be high.
- For MCAR, we **check for no or very small correlation between V and D values**.

Step 2: Missing Data Mechanisms

- Missing Completely at Random (MCAR)
  - The probability that a unit is missing a variable is completely independent of the value of the variable and of the values of other variables.
- Missing at Random (MAR)
  - The probability that a variable is missing for a unit depends on the value of an other variable(s) for that unit but not on the value of the missing variable itself.
- Missing not at Random (MNAR)
  - The probability that a variable is missing for a unit depends on the value of the variable itself.

Step 2: Missing Data Mechanism

Missing at Random (MAR)

- Data are missing at random (MAR) if the probability of missing data on X_j is unrelated to the value of X_j after controlling statistically for other variables in the analysis.
- That is, the probability that any variable is missing depends on any other variable in the model of interest, but not on the potentially missing values themselves.
  - e.g., probability of missingness depends on whether the subject is assigned to the control group or the treatment group, but does not depend on the value of the response variable.
- MAR still allows for the possibility that missingness on one variable may be related to missingness on another.

Step 2: Missing Data Mechanism

Testing MAR Assumption

- We can check whether the probability of missingness on a given variable X_j is related to any of the remaining variables X_i (i=1,...,j-1,j+1,...p) in the data set.
  - If missingness on X_j is related to other variables, then these other variables must be included in any analysis of variable X_j to correct for biases in parameter estimates that would otherwise occur.
- Does the probability of missingness depend on the values that are missing?
  - Unfortunately MAR assumption is not completely testable as the missing values are not known.
  - Alternatively we can put many variables in X, especially those that are highly correlated with X_j. Hence, we can reduce or eliminate the residual dependence of missingness on X_j itself.

Step 2: Missing Data Mechanism

Missing at Random (MAR)

- The term MAR is confusing because data are not really missing at random as the missingness seems to depend on some of the variables in the data set.
- The missingness data is a random pattern after you control for
  - Variables in your analysis
  - Auxiliary variables
  - Probability of missingness NOT dependent on unobserved variables
- Correlate variables (Var1, Var2, Var3) with D1, D2, D3
- Consider auxiliary variables--race, gender, age, education

Raw Data

ID	Var1	Var2	Var3
1	9	7	.
2	.	3	5
3	7	4	.
4	9	4	6
5	6	2	7
6	.	.	5

Missingness Ind.

ID	D1	D2	D3
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	0	0
6	1	1	0

Step 2: Missing Data Mechanism

Missing at Random (MAR)

- The term MAR is confusing
  missing at random as the m
  some of the variables in th
- The missingness data is a r
  for
  - Variables in your analysis
  - Auxiliary variables
  - Probability of missingn
    variables
- Correlate variables (Var1, V
- Consider auxiliary variables

Auxiliary variables are the variables in your data set which you do not necessarily use for modelling. But they can be very helpful in detecting the reasons of missingness. For example, in your analysis, the 'religion' variable is not required but it may be helpful for understanding why certain meals are not consumed by some people and missing in your data set.

Step 2: Missing Data Mechanism

Missing at Random (MAR)

- The term MAR is confusing because data are not really missing at random as the missingness seems to depend on some of the variables in the data set.

Previous Example:

- Male individuals are not required to answer pregnancy related questions.
- Assume pregnancy question=V3, gender=V1 variables.
- If gender=male, the D3 will be 1 as they are missing.
- There is a correlation between V and D, so this is «missing at random».
- Because it is related to observed variables but not unobserved variables.

Step 2: Missing Data Mechanism

MAR and MCAR

- MCAR is a special case of MAR. That is, if the data are MCAR, they are also MAR.
- MAR is considerably weaker assumption than MCAR.
- The difference between MAR and MCAR is whether or not other variables in the data set are associated with whether a unit has missing data on a particular variable.
  - For example, are older people more likely to refuse to respond to the income variable?

Step 2: Missing Data Mechanism

Ignorable

- Missing data mechanism is ignorable If
  - data are MAR, and
  - the parameters governing the missing-data mechanism are distinct from the parameters in the model to be estimated.
- Second condition is unlikely to be violated in real-world cases.
  - Hence, MAR and ignorability are used interchangeably in practice.
- Any general purpose method for handling missing data must assume that the missing data mechanism is ignorable.
  - If missing data mechanism is ignorable, there is no need to model the missing data process to get valid, optimal estimates of parameters.

Step 2: Missing Data Mechanisms

- Missing Completely at Random (MCAR)
  - The probability that a unit is missing a variable is completely independent of the value of the variable and of the values of other variables.
- Missing at Random (MAR)
  - The probability that a variable is missing for a unit depends on the value of an other variable(s) for that unit but not on the value of the missing variable itself.
- Missing not at Random (MNAR)
  - The probability that a variable is missing for a unit depends on the value of the variable itself.

Step 2: Missing Data Mechanism

MNAR

- If the MAR assumption is violated, the data are said to be MNAR.
- Missingness is related to the unknown (missing) value.
  - For example, the probability that a particular question is not answered is dependent on the answer itself.
- In that case, the missing-data mechanism is non-ignorable, and valid estimation requires that the missing-data mechanism be modeled as part of the estimation process.

Step 2: Missing Data Mechanism

MNAR

- Every MNAR situation is different, the model for the missing-data mechanism must be carefully tailored to each situation.
- Furthermore, there is no information in the data that would help you choose an appropriate model, and no statistic that will tell you how well a chosen model fits the data.
- It is probably a good idea to enlist the help and advice of someone who has real expertise in this area.
- It is also recommended that you try different models for the missing-data mechanism to get an idea of how sensitive the results are to model choice.

Step 2: Missing Data Mechanism

Summary of Methods

ID	Var1	Var2	Var3
1	9	7	.
2	.	3	5
3	7	4	.
4	9	4	6
5	6	2	7
6	.	.	5

ID	D1	D2	D3
1	0	0	1
2	1	0	0
3	0	0	1
4	0	0	0
5	0	0	0
6	1	1	0

Exploring Missing Data Mechanisms

- Can’t be 100% sure about probability of missing (since we don’t actually know the missing values)
- Little’s MCAR Test is the most common test for MCAR
- Many missing data methods assume MCAR or MAR but our data often are MNAR

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

Imputation methods are generally classified into two categories:

- A deterministic imputation method: determines one and only one possible value for imputing each missing case.
  - Once the imputation scheme is set up, the imputation result is unique.

- A random imputation method: draws imputation values randomly either from the observed data or from the predicted distribution.
  - Multiple sets of imputations can be created to capture the uncertainty between imputations via any random imputation method.
  - Generally, a random imputation method adds more variability to the statistics computed from an imputed data set than a deterministic imputation method.

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Simple deterministic imputation method:
  - Deductive imputation: deduces missing values from available information, such as similar items in previous surveys, related items in current surveys, etc.
    - To apply this method, the user needs to find some deterministic relationship between the missing item and items from other resources
  - Overall or cell mean imputation (also called adjusted mean imputation or substitution method)

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Simple deterministic imputation method:
  - Deterministic hot deck imputation:
    - Hot deck imputation does not employ any explicit statistical model.
    - Its major disadvantage is that it cannot recover typical values for objects with certain characteristics if no such subject responds to a survey.
    - It employs many methods:
      - Sequential nearest neighbor hot deck imputation (Traditional hot deck)
      - Multivariate matching
      - Distance function matching

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Simple random imputation methods:
  - Overall or cell mean imputation with random disturbance
  - Random hot deck method
  - Overall random imputation
  - Approximate Bayesian Bootstrap (ABB)
  - Within-class random imputation

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Model based deterministic imputation methods: works well if the data constructor knows the reasons of non response hence satisfying the assumptions.
  - However assumptions are often nonverifiable.
  - Example methods:
    - Ratio imputation
    - Predicted regression imputation
    - EM algorithm

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Model based random imputation methods:
  - Some example methods:
    - Draw imputations from predicted distributions
    - Random regression imputation
    - Ratio with random disturbance imputation
- Imputation methods related to Bayesian theories:
  - Some example methods:
    - Adjusted data augmentation
    - Sequential imputation method

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Types of Imputations

- Recent Techniques:
  - Imputation Using Deep Learning (Datawig)
  - Imputation Using Multivariate Imputation by Chained Equation (MICE)

Ref: Hu, M. X., & Salvucci, S. (2001). A Study of Imputation Algorithms. Working Paper Series.

Step 3: Deal with Missing Data

Missing Data Solutions: Summary

- Ignore the tuple.
- Fill in the missing value manually.
- Use a global constant to fill in the missing value.
- Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value.
- Use the attribute mean or median for all samples belonging to the same class as the given tuple.
- Use the most probable value to fill in the missing value.
  - we can impute probabilistically
  - e.g., multiple imputation – unbiased estimates, values are imputed with uncertainty

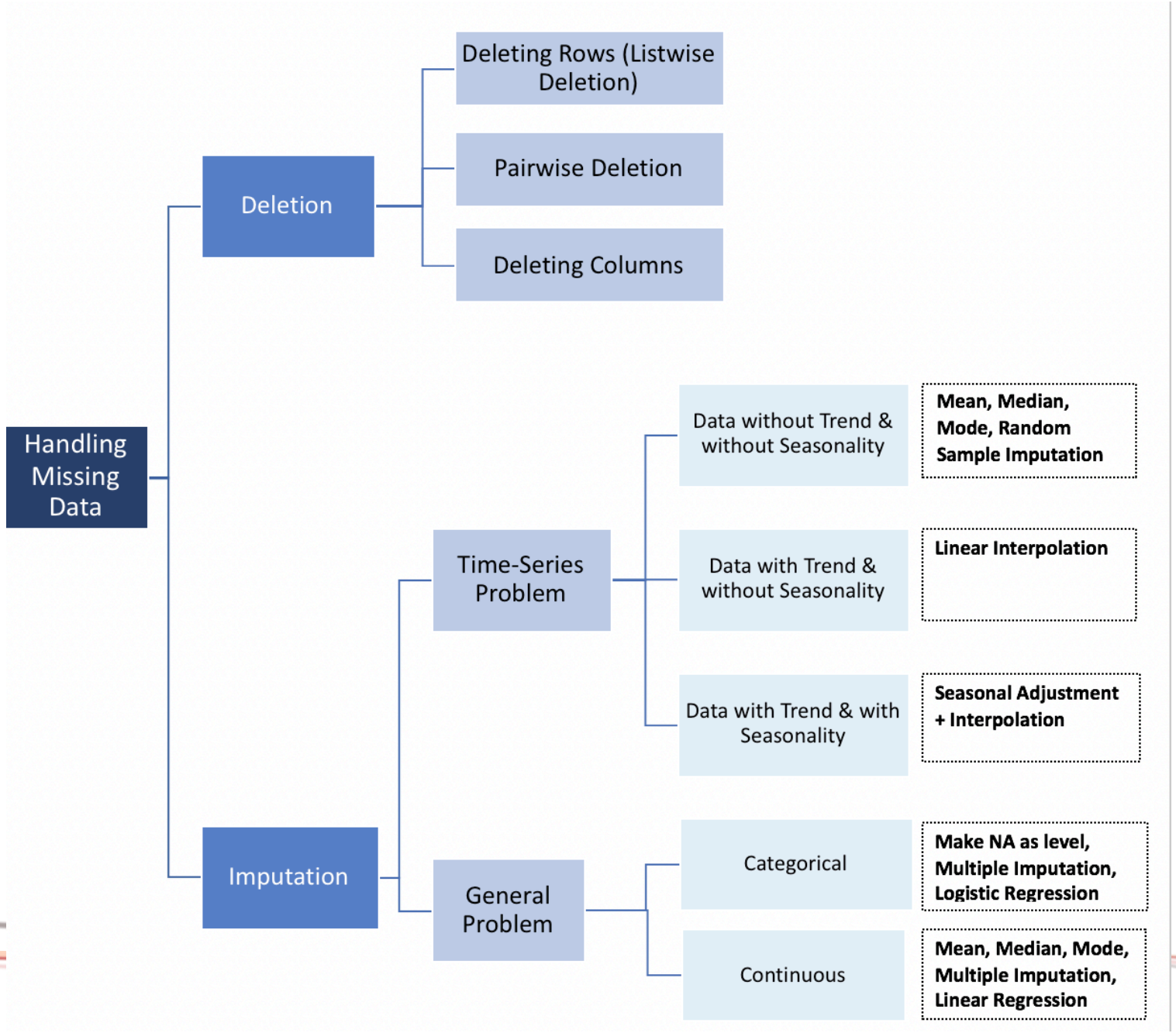
Step 3: Deal with Missing Data

Which Method to Use?

- There is general agreement that a good method should:
  - Minimize bias: missing data can introduce bias into parameter estimates but a good method makes that bias as small as possible.
  - Maximize the use of available information: Avoid discarding any data, and use the available data to produce good parameter estimates.
  - Yield good estimates of uncertainty: accurate estimates of standard errors, confidence intervals and p-values.
- In addition, it would be nice if the method does not necessitate making unnecessarily restrictive assumptions about the missing-data mechanism.

Step 3: Deal with Missing Data

- Use what you know about
  - Why data is missing
  - Distribution of missing data
- Decide on the best analysis strategy to yield the least biased estimates
  - Deletion Methods
    - Listwise deletion, pairwise deletion, variable deletion
  - Single Imputation Methods
    - Mean/mode substitution, dummy variable method, single regression
  - Model-Based Methods
    - Maximum Likelihood, Multiple imputation


Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

Deletion Methods

- Listwise deletion
  - AKA «**complete case analysis**»
- Pairwise deletion

Listwise Deletion (Complete Case Analysis)

- Only analyze cases with available data on each variable
- Advantages:
  - Simplicity
  - Comparability across analyses
- Disadvantages:
  - Reduces statistical power (because lowers n)
  - Doesn't use all information
  - Estimates may be biased if data not MCAR*

Gender	8th grade math test score	12th grade math score
~~F~~	~~45~~	~~.~~
~~M~~	~~.~~	~~99~~
F	55	86
F	85	88
F	80	75
~~.~~	~~81~~	~~82~~
F	75	80
~~M~~	~~95~~	~~.~~
M	86	90
F	70	75
~~F~~	~~85~~	~~.~~

*NOTE: List-wise deletion often produces *unbiased regression slope estimates* as long as missingness is not a function of outcome variable.

Pairwise Deletion (Available Case Analysis)

- Analysis with all cases in which the variables of interest are present.
- Advantage:
  - Keeps as many cases as possible for each analysis
  - Uses all information possible with each analysis
- Disadvantage:
  - Can't compare analyses because sample different each time

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

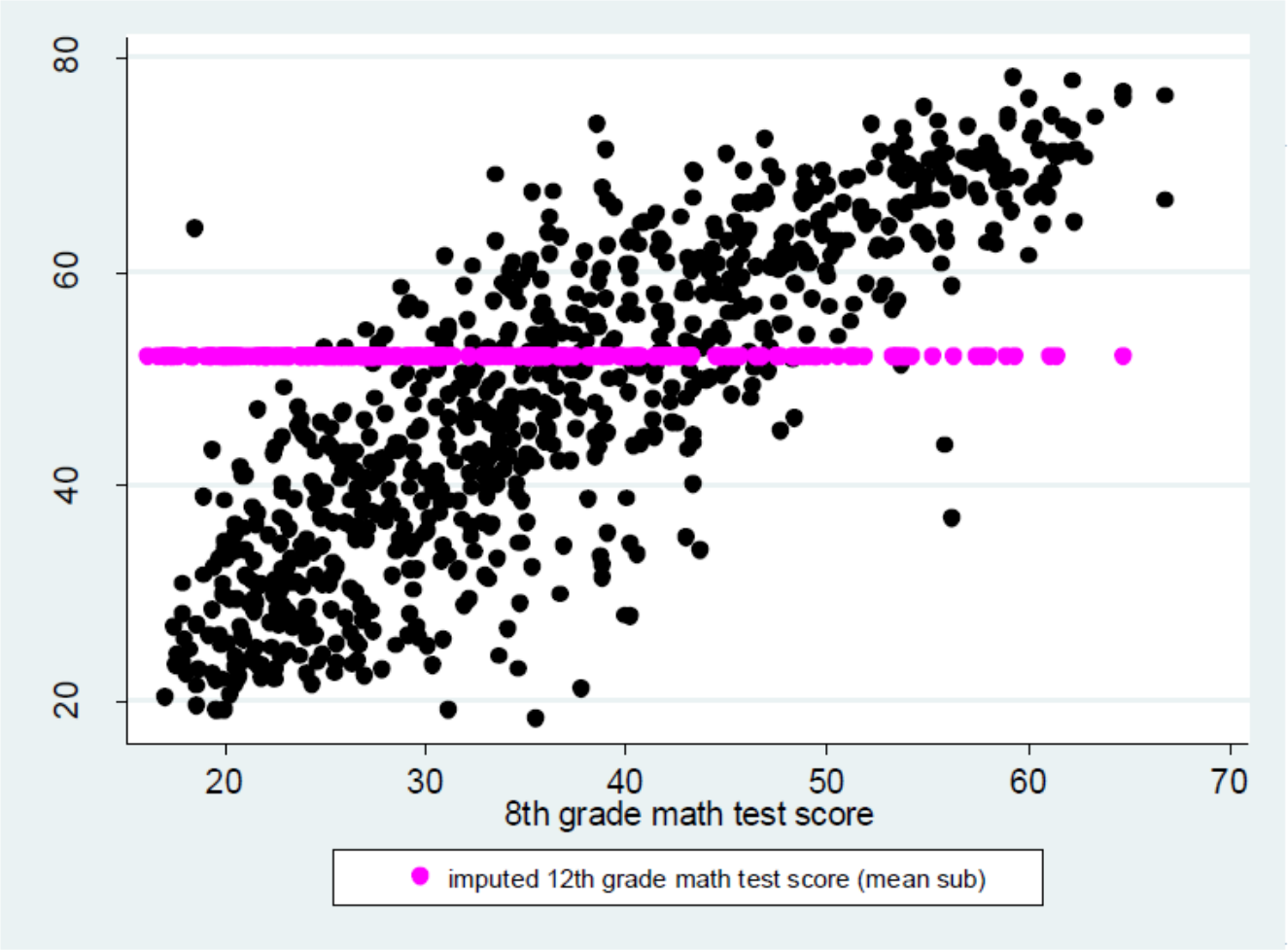
Single Imputation Methods

- A single data set with non-missing values for all observations on all variables is produced from the original data set that contains missing values.
- Single imputation does not reflect the uncertainty about the prediction of the missing values.
- Methods:
  - Mean/Mode substitution
  - Dummy variable control
  - Conditional mean substitution
  - Other variations (e.g., “last observation carried forward” for time series or sequential data)

Mean/Median/Mode Substitution

- Replace missing value with sample mean, median or mode
- Run analyses as if all complete cases
- Advantages:
  - Simplicity
  - Can be used by complete case analysis methods
- Disadvantages:
  - Reduces variability
  - Weakens covariance and correlation estimates in the data (because ignores relationship between variables)

Mean/Median/Mode Substitution



Y axis shows the 12th grade math test scores.
When imputed with mean/mode, it affects the distribution of the dataset negatively.

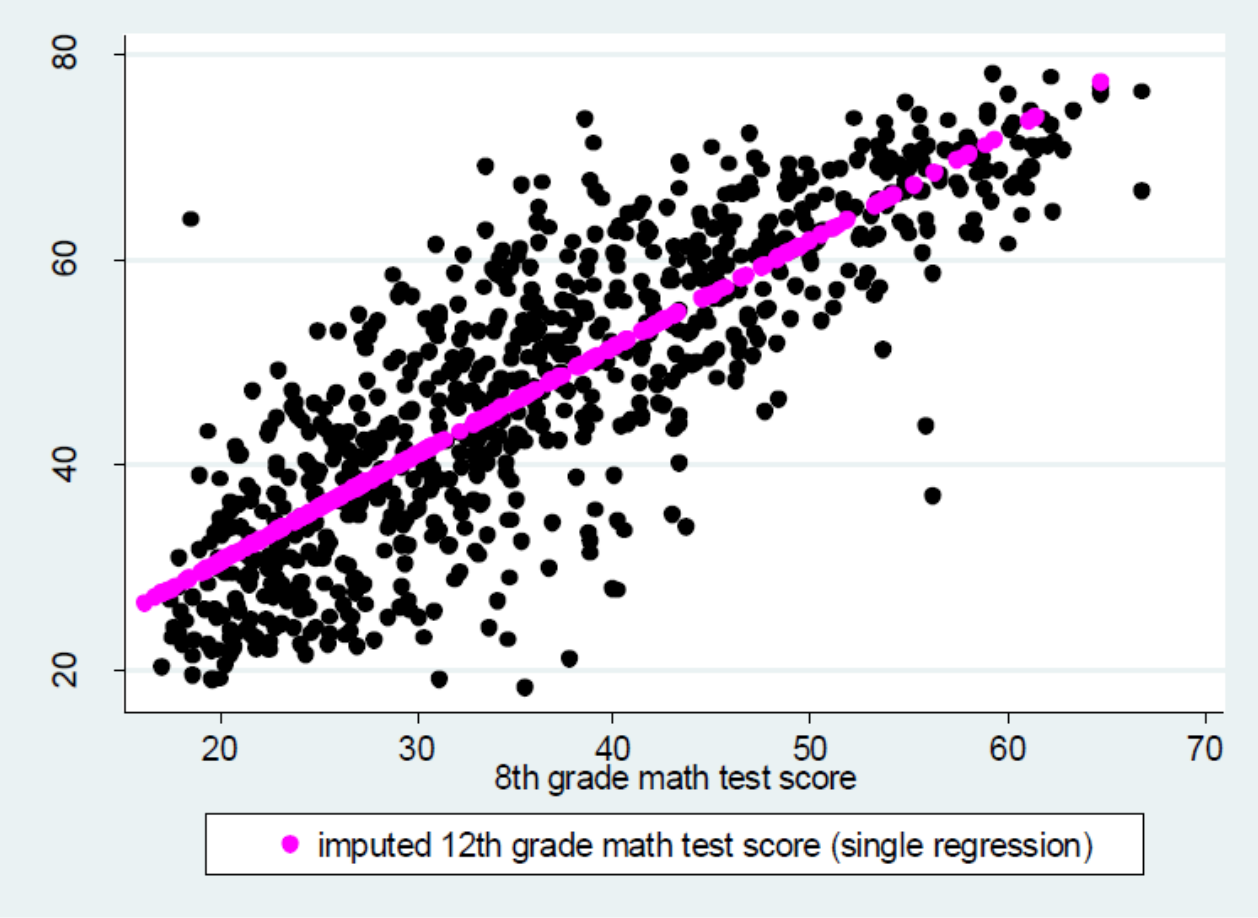
Dummy Variable Adjustment

- Create an indicator for missing value (1=value is missing for observation; 0=value is observed for observation)
- Impute missing values to a constant (such as the mean)
- Include missing indicator in regression
- Advantage: Uses all available information about missing observation
- Disadvantage: Results in biased estimates
- Results are not biased if value is missing because of a legitimate skip.

Regression Imputation

- Replaces missing values with predicted score from a regression equation.
- Advantage:
  - Uses information from observed data
- Disadvantages:
  - Overestimates model fit and correlation estimates
  - Weakens variance

Regression Imputation



Model Based Methods

- Maximum likelihood
- Multiple imputation

Model Based Methods:

Maximum Likelihood Estimation

- ML estimate: value that is most likely to have resulted in the observed data
- Identifies the set of parameter values that produces the highest log-likelihood.
- Conceptually, process the same with or without missing data
  - Advantages:
    - Uses full information (both complete cases and incomplete cases) to calculate log likelihood
    - Unbiased parameter estimates with MCAR/MAR data
  - Disadvantages
    - SEs biased downward—can be adjusted by using observed information matrix
- Inference based methods such as Bayesian formula or decision tree is used to find the most probable value for the missing data.

Multiple Imputation

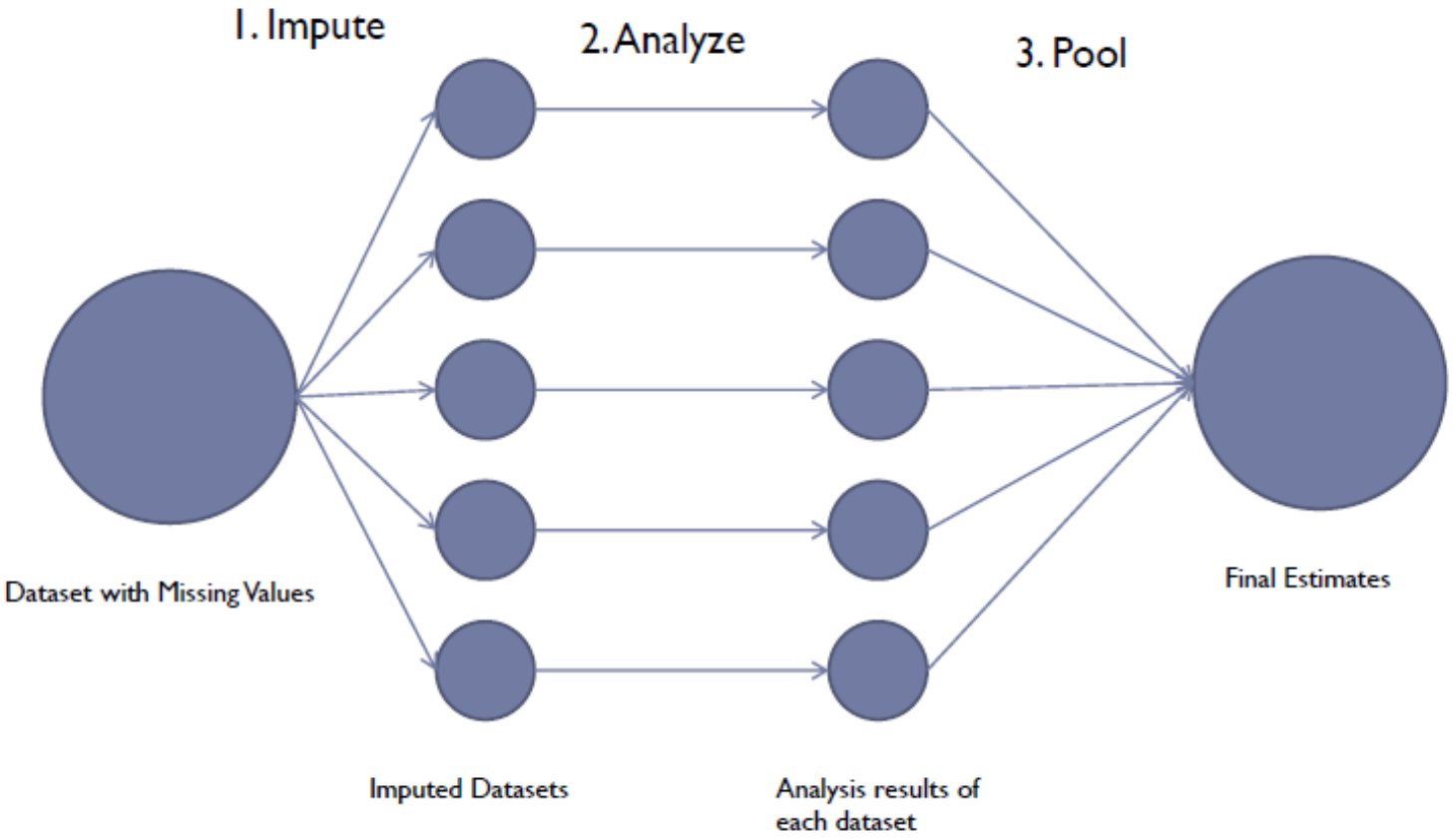
- The missing values are imputed multiple times, resulting in several complete data sets.
- The imputation should be random, so that the imputed data sets are different from each other.
- In addition, the imputation should be conducted such that the variation in the imputed values for a given missing value reflects the uncertainty in our ability to predict that value.

Multiple Imputation

1. Impute: Data is “filled in” with imputed values using specified regression model
   - This step is repeated m times, resulting in a separate dataset each time.
2. Analyze: Analyses performed within each dataset
3. Pool: Results pooled into one estimate

- Advantages:
  - Variability more accurate with multiple imputations for each missing value
  - Considers variability due to sampling AND variability due to imputation
- Disadvantages:
  - Cumbersome coding
  - Room for error when specifying models

Multiple Imputation Process



Multiple Imputation:

Example

- Multiple imputation (say, five) impute values for each missing value, each of which is predicted from a slightly different model and each of which also reflects sampling variability.
- We use each set of imputed values to form (along with the observed data) a completed dataset (ds_1,ds_2, … ds_n).
- Within each completed dataset a standard analysis can be run.
- Then inferences can be combined across datasets.

Multiple Imputation:

Example

- Suppose we want to make inferences about a regression coefficient, β.
- We obtain estimates $\hat{\beta}_m$ in each of M datasets as well as standard errors $s_1, s_2, \ldots, s_M$.
- To obtain an overall point estimate, we then simply average over the estimates from the separate imputed datasets.

  That is, $\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_m.$
- A final variance estimate V_β reflects the variance within and between

$$V_\beta = W + \left(1 + \frac{1}{m}\right) B,$$

where $W = \frac{1}{m} \sum_{m=1}^{M} s_m^2$, and $B = \frac{1}{m-1} \sum_{m=1}^{M} (\hat{\beta}_m - \hat{\beta})^2$.

Multiple Imputation Process

Toy Example

- Lets have a toy dataset `my_ds`

```
import statsmodels.imputation.mice as mice
import statsmodels.api as sm

imp = mice.MICEData(my_ds)
model = mice.MICE('Y ~ X1 + X2', sm.OLS, imp)
result = model.fit(0,3)
for result in model.results_list:
  print(result.summary())
print(results.summary())
```

Three datasets (on the right) are generated.

	Y	X1	X2
0	0.56	-0.26	-0.58
1	0.08	0.58	-0.25
2	-1.31	0.47	NaN
3	-0.51	NaN	-0.25
4	1.01	-0.27	1.33
5	0.22	NaN	-0.44
6	-0.54	0.26	NaN
7	0.25	-0.26	-0.70
8	-0.38	-1.23	0.01
9	-0.83	NaN	NaN

	Y	X1	X2
0	0.56	-0.26	-0.58
1	0.08	0.58	-0.25
2	-1.31	0.47	-0.58
3	-0.51	-0.27	-0.25
4	1.01	-0.27	1.33
5	0.22	0.47	-0.44
6	-0.54	0.26	0.01
7	0.25	-0.26	-0.70
8	-0.38	-1.23	0.01
9	-0.83	0.47	-0.25

	Y	X1	X2
0	0.56	-0.26	-0.58
1	0.08	0.58	-0.25
2	-1.31	0.47	0.01
3	-0.51	-0.26	-0.25
4	1.01	-0.27	1.33
5	0.22	-0.27	-0.44
6	-0.54	0.26	1.33
7	0.25	-0.26	-0.70
8	-0.38	-1.23	0.01
9	-0.83	0.47	1.33

	Y	X1	X2
0	0.56	-0.26	-0.58
1	0.08	0.58	-0.25
2	-1.31	0.47	-0.25
3	-0.51	-0.26	-0.25
4	1.01	-0.27	1.33
5	0.22	-0.26	-0.44
6	-0.54	0.26	-0.25
7	0.25	-0.26	-0.70
8	-0.38	-1.23	0.01
9	-0.83	-0.26	1.33

Multiple Imputation Process

Toy Example

- If we run the regression in each dataset, we'll get slightly different results because of those differences in the imputed values (output edited for readability):

```
R-squared:                       0.209
R-squared:                      -0.017
F-statistic:                     0.9243
Prob (F-statistic):              0.440
                coef    std err          t      P>|t|
-------------------------------------------------------
Intercept     -0.0654      0.233     -0.281      0.787
X1            -0.1892      0.427     -0.443      0.671
X2             0.4729      0.415      1.139      0.292
=======================================================
R-squared:                       0.118
Adj. R-squared:                 -0.134
F-statistic:                     0.4691
Prob (F-statistic):              0.644
                coef    std err          t      P>|t|
-------------------------------------------------------
Intercept     -0.1696      0.245     -0.691      0.512
X1            -0.4180      0.473     -0.884      0.406
X2            -0.0424      0.311     -0.136      0.895
=======================================================
R-squared:                       0.042
Adj. R-squared:                 -0.232
F-statistic:                     0.1535
Prob (F-statistic):              0.861
                coef    std err          t      P>|t|
-------------------------------------------------------
Intercept     -0.1862      0.256     -0.729      0.490
X1            -0.2770      0.512     -0.541      0.605
X2             0.0108      0.356      0.030      0.977
```

```
MICE RESULTS
=========================================
            Coef.  Std.Err.    t    P>|t|
-----------------------------------------
Intercept -0.1404    0.2561 -0.5482 0.5836
X1        -0.2948    0.4904 -0.6010 0.5478
X2         0.1471    0.4890  0.3009 0.7635
=========================================
```

$$\hat{\beta}_I = \frac{(-0.0654) + (-0.1696) + (-0.1862)}{3} = -0.1404$$

$$\hat{\beta}_{X1} = \frac{(-0.1892) + (-0.4180) + (-0.2770)}{3} = -0.2948$$

$$\hat{\beta}_{X2} = \frac{(0.4729) + (-0.0424) + (0.0108)}{3} = 0.1471$$

Recommendations

- Perform initial descriptive analyses of data to identify the nature and extent of missing data.
- When item nonresponse on scales is relatively minor, consider using single imputation or substitution.
- If the amount of missing data in the entire set is very small, consider using single imputation.
- Experts on missing data often claim that missingness on any individual variable can be 50% or higher and optimal methods will handle the analyses quite well.
  - 1-2% missing is rather low.
  - 10-15% or higher moderate.
  - level of 25% or higher relatively high.
- If the amount of missing data is moderate or large and the variables related to missingness cannot be included in all analyses, use multiple imputation.

Disguised Missing Data

- The problem of disguised missing data arises when missing data values are not explicitly represented as such, but are coded with values that can be misinterpreted as valid data.
  - Example: Jan. 1 as everyone's birthday?
- Causes:
  - the use of form-based electronic data entry systems with rigid edit checks, included to prevent data entry errors.
  - deliberate fraud.
  - the lack of a standard missing data representation.

Disguised Missing Data

Spatial Database Example

- Typical problems:
  - GPS devices may malfunction due to a problem in the projection system and consequently inaccurate coordinate information may be obtained.
  - Longitude information may be entered in lieu of latitudes and vice versa.
    - This situation arises when these coordinates are very close in a particular town.
  - Users may enter the coordinate information inaccurately in the database.
  - The coordinate information may be obtained before GPS devices get ready (GPS devices should receive information from at least three satellites for an accurate measure.)

Disguised Missing Data

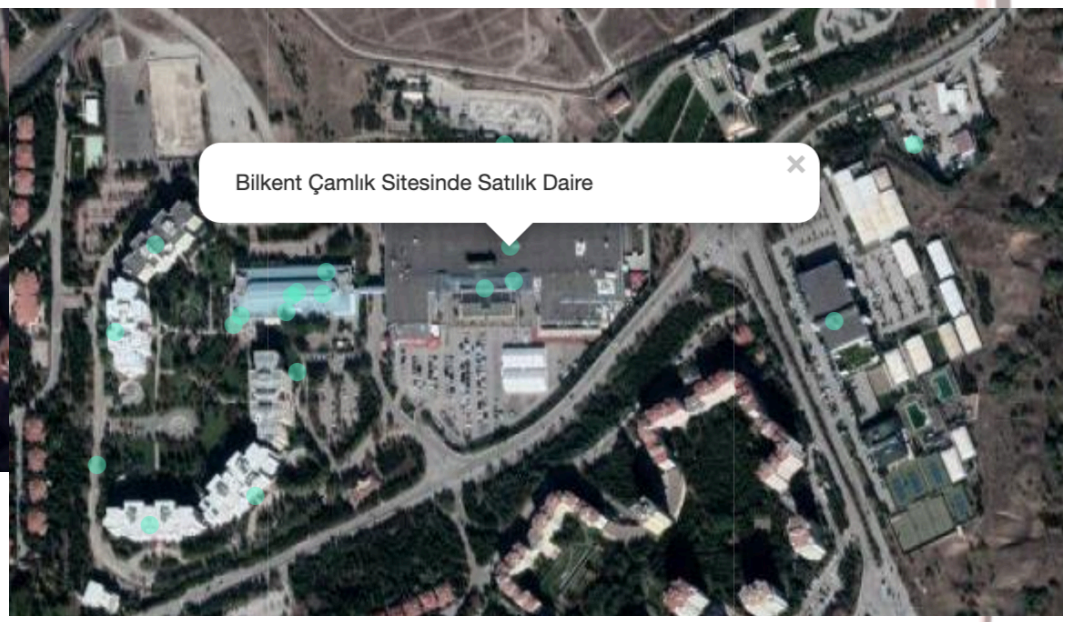
Spatial Database Example

- Typical problems:
  - The borders of a town or city may be changed or may have been defined incorrectly in the system's database.
  - Users may enter the same coordinates to the database systematically for any given points. The reasons are numerous:
  - Users may not know the exact location;
  - Users may enter the same values to save time and it is easy;
  - Users may use dated reports to enter values.

Disguised Missing Data

Spatial Database Example

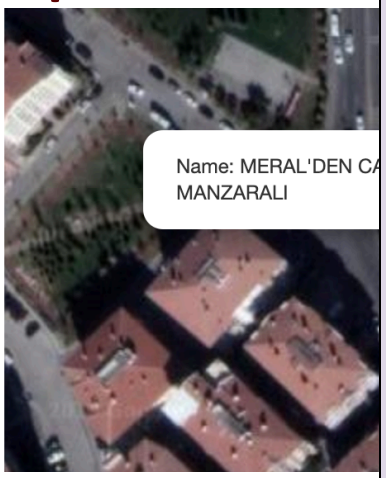
Several ad listings were posted in the same place with very similar coordinates. These ads were posted in the center of a real estate agency. An example for a disguised missing data.



An ad listing was posted with an inaccurate coordinate specifically in the Bilkent Center Shopping Mall (the provided textual address does not match with the coordinates). The error is not repetitive in nature. It is an erroneous data.

Disguised Missing Data

Spatial Database Example

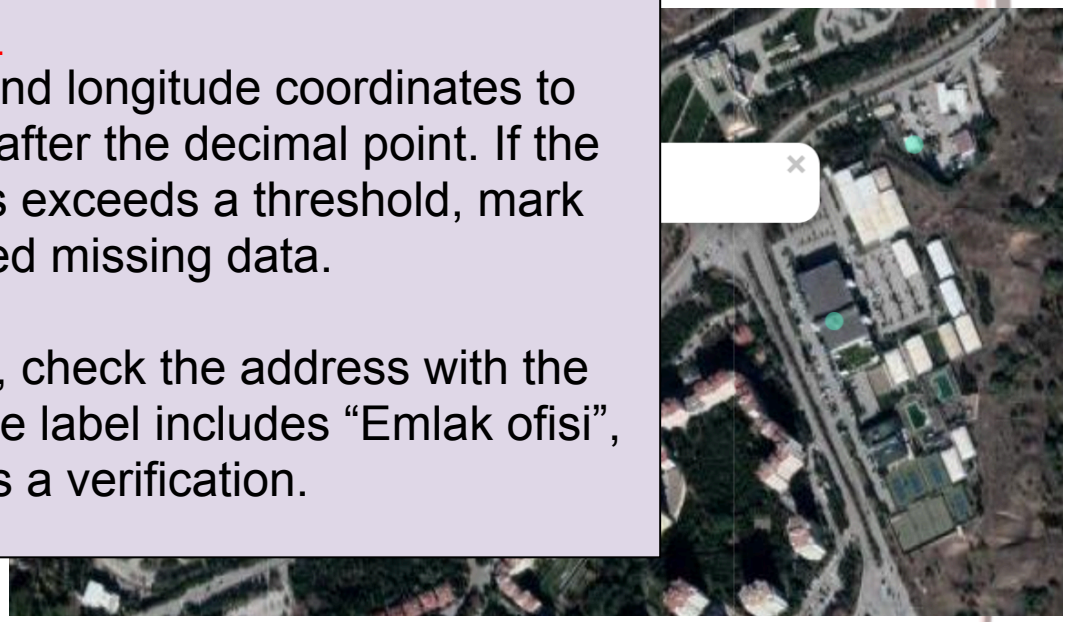


Several ad listings were posted in the same place with very similar coordinates. These ads were posted in the center of a real estate agency. An example for a disguised missing data.

Simplest Option:

Round latitude and longitude coordinates to three/four digits after the decimal point. If the number of points exceeds a threshold, mark them as disguised missing data.

For the left case, check the address with the Google API. If the label includes “Emlak ofisi”, it can be used as a verification.



An ad listing was posted with an inaccurate coordinate specifically in the Bilkent Center Shopping Mall (the provided textual address does not match with the coordinates). The error is not repetitive in nature. It is an erroneous data.

Disguised Missing Data

Example: Pima Indians diabetes database

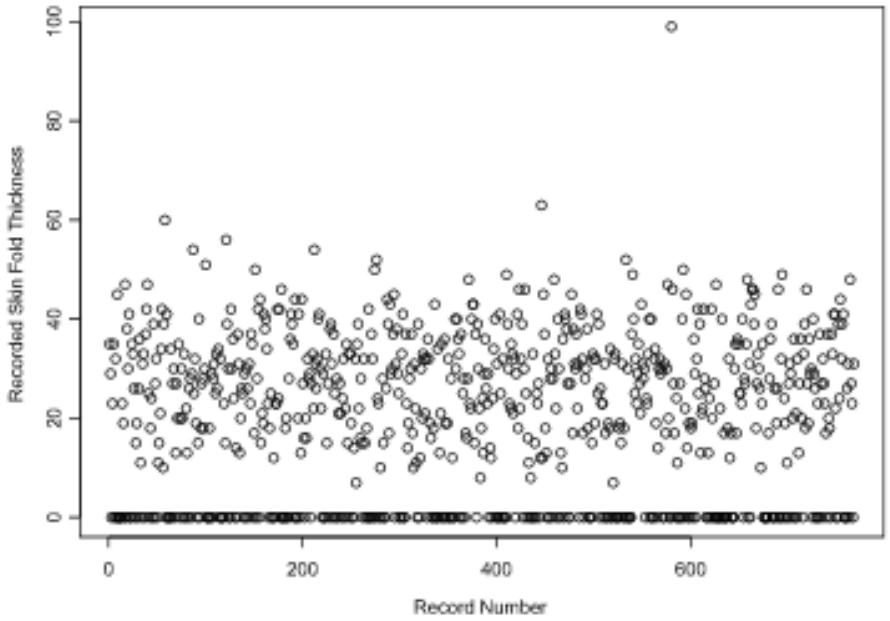


Figure 7: Plot of recorded skin fold thickness (TSF) values from the Pima Indians diabetes database. Although the zero values are visually suspicious in this plot, they cannot be detected using automated outlier detection algorithms.

Ref: http://kdd.org/exploration_files/12-Pearson.pdf

Disguised Missing Data Detection

- Outliers
- Inliers: data values that lies in the interior of the statistical distribution (of the nominal data values) but which are nevertheless in error.
- Easy detection:
  - Q-Q plots

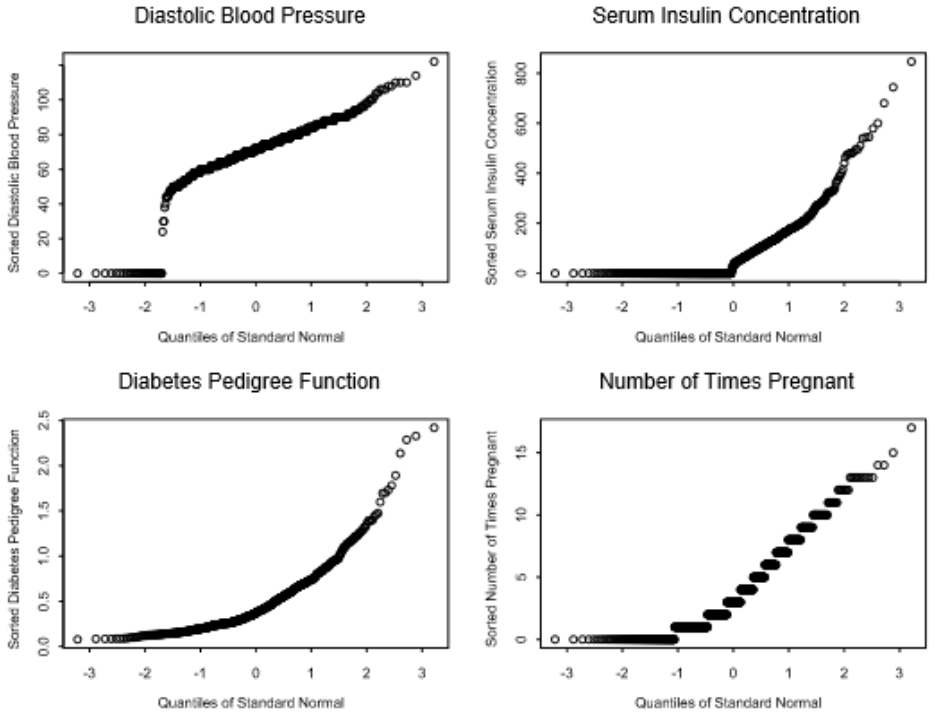


Figure 9: Normal Q-Q plots constructed from four of the Pima Indians diabetes variables: DIA (upper left), INS (upper right), DPF (lower left), and NPG (lower right).

Ref: http://kdd.org/exploration_files/12-Pearson.pdf

Disguised Missing Data Detection

Automatic detection is possible:

Check Hua, Ming, and Jian Pei. "Cleaning disguised missing data: a heuristic approach." Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007.

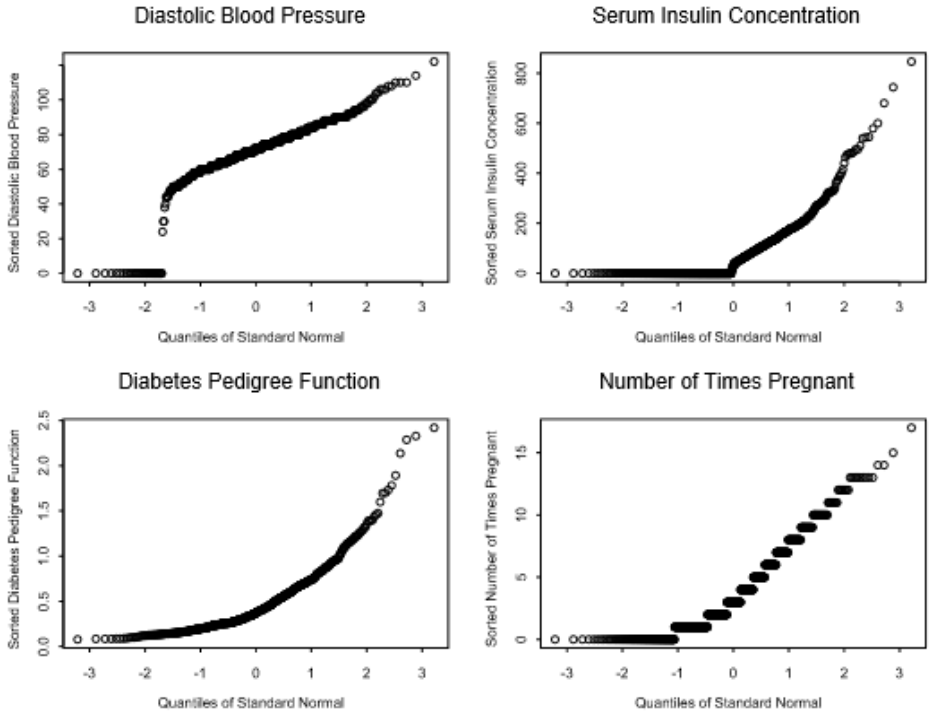


Figure 9: Normal Q-Q plots constructed from four of the Pima Indians diabetes variables: DIA (upper left), INS (upper right), DPF (lower left), and NPG (lower right).

Ref: http://kdd.org/exploration_files/12-Pearson.pdf

Disguised Missing Data

Spatial Database Example

- Road safety accidents are stored in a relational database, which is used by all traffic inspectorates in Turkey. When an accident takes place, the local officer should enter the details of the accident, i.e. time and date of the accident, town, city, the road name, road type, accident type, number of dead and wounded people, latitude and longitude of the accident. All of which are mandatory.
- Disguised missing data can appear in two different ways:
  - The coordinates may not be on the road and are too far away from the valid range (outliers).
  - Coordinates may appear as valid values i.e. on the road and within the valid range (town boundaries). This type of data is called as inliers.

Disguised Missing Data

Spatial Database Example

The same coordinates are entered systematically to the system. For example, latitude 36 and longitude 37 were always entered instead of indicating the data entry as missing or unknown in city of Osmaniye in Turkey in year 2006 and 2007.

Year	Month	Day	Day of the Week	Province	Local Officer	Name of the Road	# Dead People	# Wounded People	X Coordinate	Y Coordinate
2006	5	8	Monday	OSMANİYE-MERKEZ	OSMANİYE-OSMANİYE	52-07	0	0	36,000	37,000
2006	1	10	Tuesday	OSMANİYE-MERKEZ	OSMANİYE-OSMANİYE	52-07	0	2	36,000	37,000
2006	5	3	Wednesday	DÜZİÇİ	OSMANİYE-OSMANİYE	52-08	0	1	36,000	37,000
2006	6	3	Saturday	BAHÇE	OSMANİYE-OSMANİYE	52-10	0	0	36,000	37,000
2006	6	27	Tuesday	BAHÇE	OSMANİYE-OSMANİYE	52-10	0	0	36,000	37,000
2006	10	6	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	0	36,000	37,000
2006	2	17	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	0	36,000	37,000
2006	7	28	Friday	TOPRAKKALE	OSMANİYE-OSMANİYE	52-06	0	1	36,000	37,000

Later case occurs due to GPS, manual entry or system problems. Users may enter spurious coordinate values with minor changes in the decimals. For example, the points where Y coordinates vary between 41.050 and 41.102 and X coordinates vary between 29.000 and 29.005 are detected as disguised missing

Year	Month	Day	Day of the Week	Province	Local Officer	Name of the Road	# Dead People	# Wounded People	X Coordinate	Y Coordinate
2008	1	7	Wednesday	KAĞITHANE	İSTANBUL-Trafik De	01-01	0	0	29,000	41,0607
2008	4	27	Saturday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,0638
2008	12	10	Saturday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	7	29,004	41,0669
2008	6	14	Tuesday	KAĞITHANE	İSTANBUL-Trafik De	01-01	0	0	29,004	41,1006
2008	3	5	Friday	BEŞİKTAŞ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,1015
2008	10	10	Friday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,000	41,1102
2008	5	5	Thursday	ŞİŞLİ	İSTANBUL-Trafik De	01-01	0	0	29,003	41,1394
2008	1	29	Tuesday	ŞİŞLİ	İSTANBUL-Bölge Tra	02-02	0	0	29,003	41,1013
2008	2	5	Friday	ŞİŞLİ	İSTANBUL-Bölge Tra	02-05	0	0	29,005	41,1001

Disguised Missing Data Detection

- Hua and Pei's method based on Embedded Unbiased Sample (EUS) Heuristic can be used.
  - This approach is based on the heuristic that only a small number of values are frequently used as disguised missing values (one or two in an attribute) in real world data and these values are randomly distributed in the database.
  - If the GPS data is provided accurately to the system, we expect to see strong correlations between the coordinates and the town, coordinates and the name of the traffic inspectorate, coordinates and the road names, coordinates and the city. If a point is used as a disguise value such as [29;41.05] in city of İstanbul, then it will appear with different attributes values in the database i.e. the same coordinates with different towns, cities, and road names.
  - The aim is to measure the distribution similarity between attribute couples of the dataset and the projected subset.

Disguised Missing Data

- Once identified, treat them as missing data!