MONASH University

FIT5196 DATA WRANGLING

Week 7

Data Quality & Anomalies

By Jackie Rong , additional materials: Sailaja Rajanala

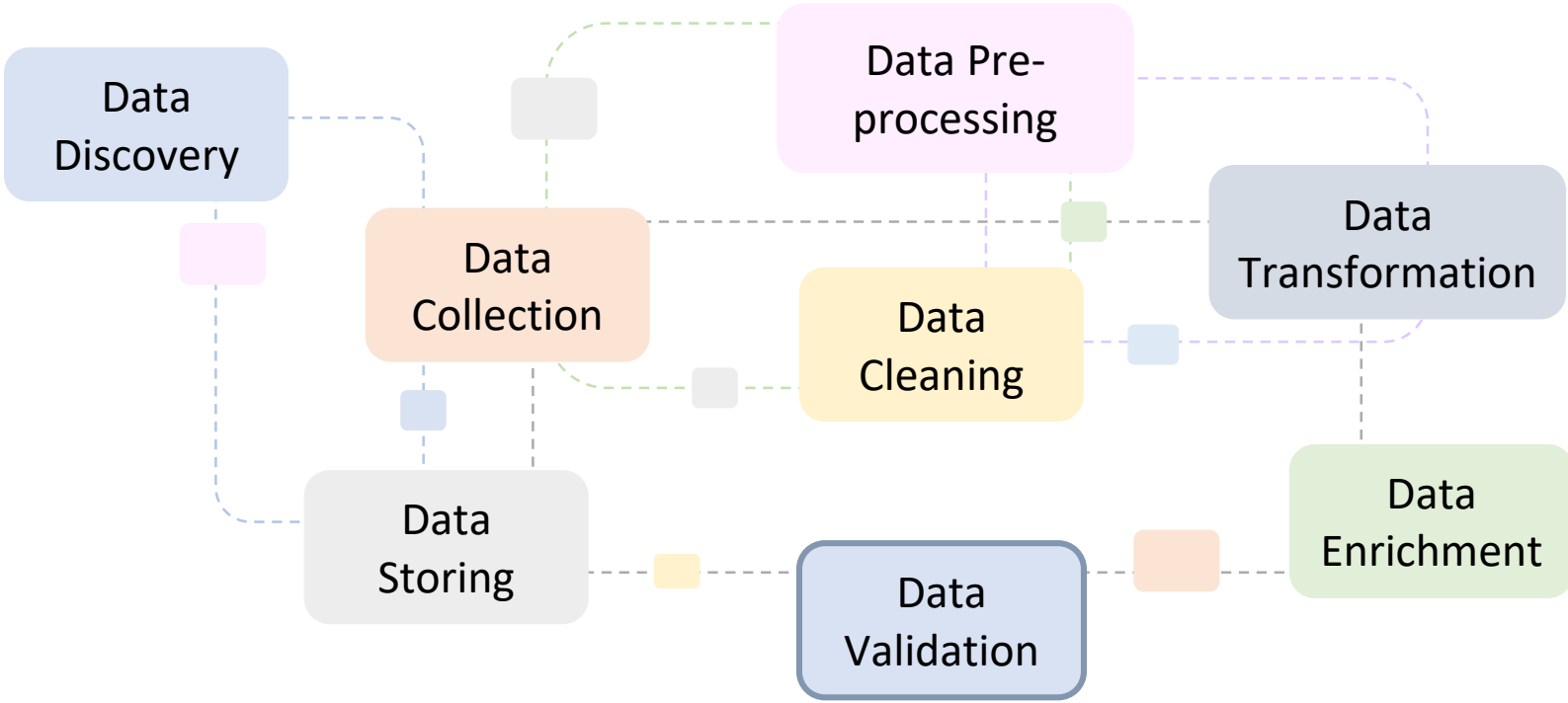
Faculty of Information Technology

Monash University

Data Wrangling Tasks (Recap)

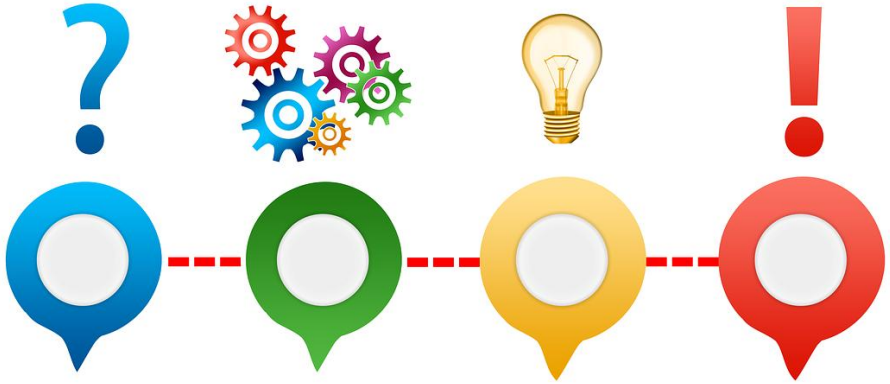
Data structuring is a critical aspect of managing and organizing data in a way that it can be efficiently accessed and manipulated.

The **main goal** is to enable data to be processed in an efficient manner, enhancing both speed and accessibility while minimizing resource usage.



Data Quality

- Definition of Data Quality
- Impact of Poor Data Quality
- Data Quality Dimensions and Measures
- Data Quality Challenges
- Data Anomalies and Data Quality Issues
- Data Quality Management Frameworks



MONASH University

Data Quality

- **Data quality** refers to the condition or state of data based on factors that influence its accuracy, completeness, reliability, relevance, and timeliness.
- High-quality data is essential for businesses, governments, and organizations to make informed decisions, improve operational efficiency, and gain competitive advantage.

https://www.capellasolutions.com/blog/10-data-management-challenges-every-it-director-must-conquer

Importance of Data Quality

- **Enhanced decision-making**
  - High-quality data ensures that decisions are based on accurate and factual information, reducing the risk of costly mistakes.
  - Stakeholders can make decisions with greater confidence when they trust the data's accuracy and completeness.

- **Regulatory compliance and risk management**
  - Maintaining high data quality helps organizations comply with these regulations, avoiding legal penalties.
  - Accurate and reliable data helps in identifying potential risks and vulnerabilities, allowing for proactive measures to mitigate them.

Importance of Data Quality

- **Operational efficiency**
  - Clean, accurate data streamlines operational processes, reducing time spent on corrections and verifications.
  - With reliable data, organizations can optimize resource allocation, ensuring that efforts and investments are directed where they are most needed.

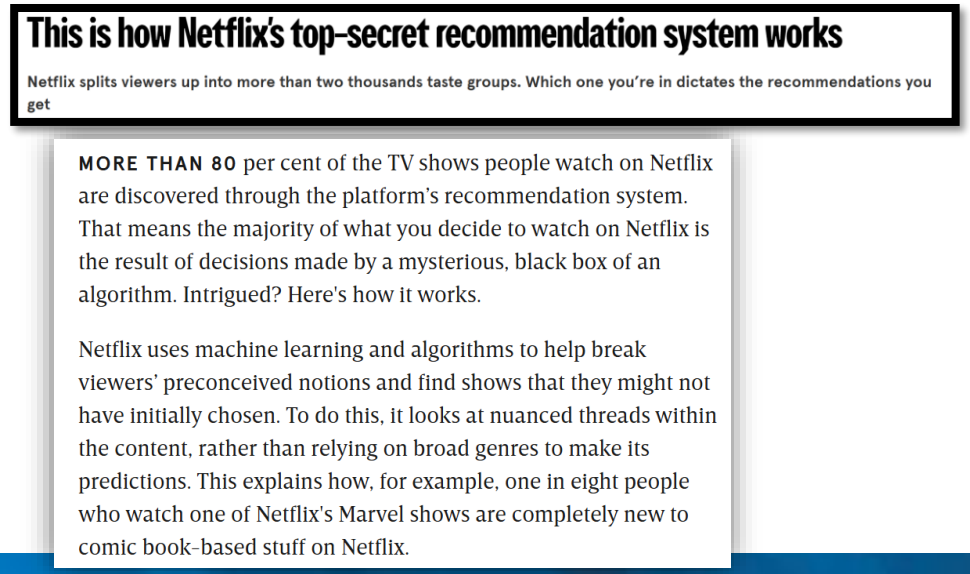
- **Customer satisfaction**
  - High-quality data enables personalized customer experiences by accurately understanding customer preferences and behaviours.
  - By analysing accurate data, organizations can identify areas for service improvement, enhancing overall customer satisfaction and loyalty.


Economics

Rapid-Fire Fulfillment

by Kasra Ferdows, Michael A. Lewis and Jose A.D. Machuca

Zara manufactures and distributes products in small batches. Instead of outside partners, the company manages all design, warehousing, distribution, and logistics functions itself. The result is a superresponsive supply chain exquisitely tailored to Zara's business model. Zara can design, produce, and deliver a new garment to its 600–plus stores worldwide in a mere 15 days.


This is how Netflix's top-secret recommendation system works

Netflix splits viewers up into more than two thousands taste groups. Which one you're in dictates the recommendations you get

MORE THAN 80 per cent of the TV shows people watch on Netflix are discovered through the platform's recommendation system. That means the majority of what you decide to watch on Netflix is the result of decisions made by a mysterious, black box of an algorithm. Intrigued? Here's how it works.

Netflix uses machine learning and algorithms to help break viewers' preconceived notions and find shows that they might not have initially chosen. To do this, it looks at nuanced threads within the content, rather than relying on broad genres to make its predictions. This explains how, for example, one in eight people who watch one of Netflix's Marvel shows are completely new to comic book-based stuff on Netflix.

Importance of Data Quality

- **Financial health**
  - Reducing errors and inefficiencies leads to significant cost savings, as less time is spent correcting mistakes or dealing with data-related issues.
  - Accurate data can uncover new opportunities for revenue generation, whether through improved customer targeting, product development, or market expansion.

- **Reputation and trust**
  - Consistently high-quality data builds trust among stakeholders, including customers, investors, and partners, by demonstrating reliability and commitment to excellence.
  - Organizations known for managing their data well are often seen as more reliable and trustworthy, positively impacting their brand reputation.

Hundreds of post office operators were wrongly convicted based on data produced by the faulty Horizon IT system the Post Office imposed on them in the late 1990s. Many were accused of theft, fraud and false accounting after the Horizon software made it appear as if money was going missing from their branches. link

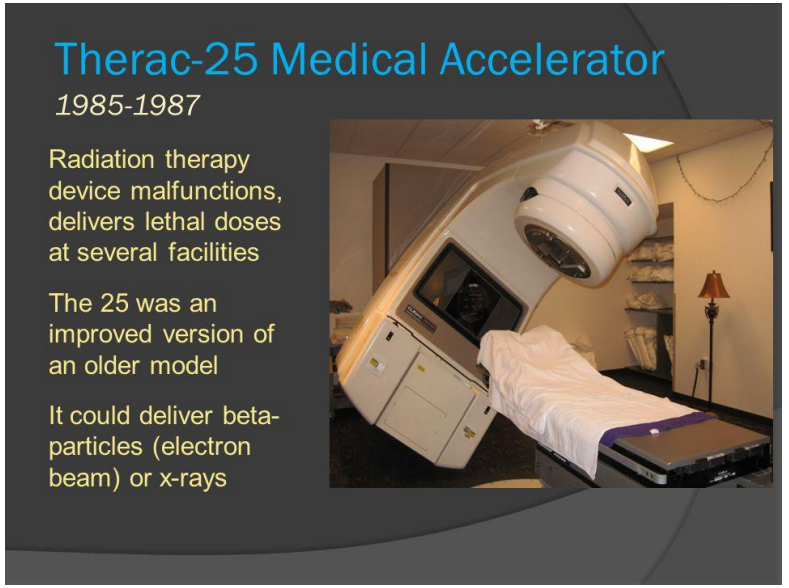
Importance of Data Quality

- **Innovation and growth**
  - High-quality data is a cornerstone for analytics and business intelligence, providing the insights necessary for innovation and strategic growth.
  - Organizations that leverage high-quality data effectively can gain a competitive edge by identifying trends, optimizing operations, and creating more value for their customers faster than their competitors.

Data Quality Failure: Test Case

The Therac-25 incident is a clear example of how data quality issues can have life-threatening consequences. Here's how data quality was a problem in this case:

- **Software Errors:** The Therac-25 software contained errors that led to incorrect calculations and radiation overdoses. These errors could be considered data quality issues within the software itself, as the software was processing and manipulating critical data (treatment settings) incorrectly.
- **Lack of Data Validation:** The software lacked proper checks to ensure the accuracy and validity of the data being entered (treatment settings). This could have involved features like dose range limitations or checks for conflicting settings.
- **Unrealistic Risk Assessments:** The manufacturer's initial risk assessment underestimated the likelihood of software errors and overdoses, essentially assigning very low probabilities to the possibility of software malfunction. This assessment was based on faulty data or assumptions.
- **In summary, the Therac-25 case highlights how flaws in software design, data handling within the software, and the lack of proper data validation all contributed to the radiation overdoses.** These issues ultimately stemmed from a lack of focus on data quality within the safety-critical system.



The **Therac-25** is a computer-controlled radiation therapy machine produced by Atomic Energy of Canada Limited (AECL) in 1982.

Therac-25.jpg (960×720) (bp.blogspot.com)

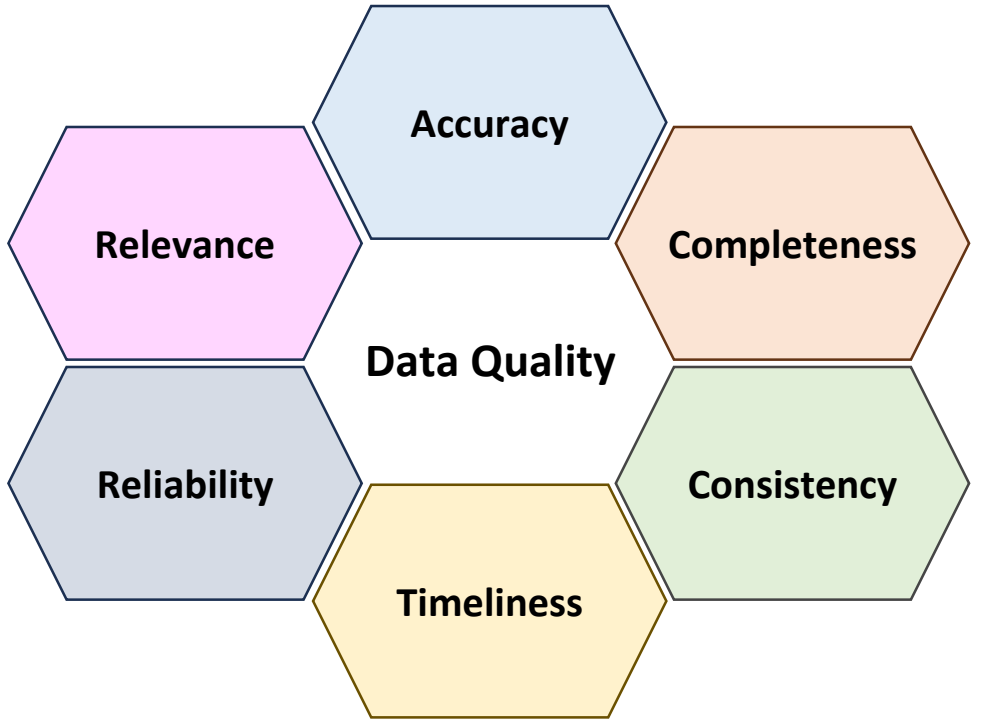
Nancy G. Leveson, Clark S. Turner, 1993. An Investigation of the Therac-25 Accidents (Abstract). Online Ethics Center. DOI: https://onlineethics.org/cases/therac-25/investigation-therac-25-accidents-abstract

Impacts of Poor Data Quality

- Inaccurate decision making
- Reduced efficiency and productivity
- Increased costs
- Damaged reputation
- Compliance and legal risks
- Customer dissatisfaction
- Misguided strategic initiatives
- Loss of competitive edge
- Data breaches and security edge
- Data breaches and security issues
- Analytical and forecasting errors

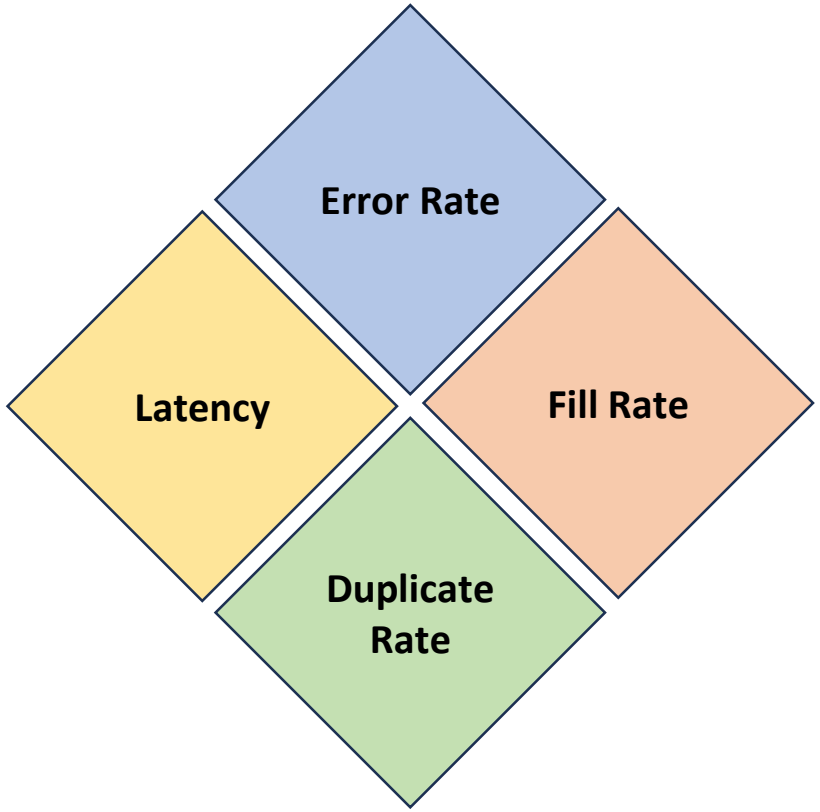
Data Quality Dimensions

- **Data quality dimensions** are the qualitative aspects or characteristics of data that contribute to its overall quality.
- They represent the broad categories or criteria used to assess the quality of data.
- These dimensions provide a framework for understanding what aspects of data need to be measured and managed to ensure its quality.



Data Quality Measures

- **Data quality measures** are the quantitative metrics or indicators used to evaluate the quality of data against the various dimensions.
- Measures are specific, measurable criteria used to assess how well the data meets quality standards.
- Measures are the practical tools used to perform the assessment.



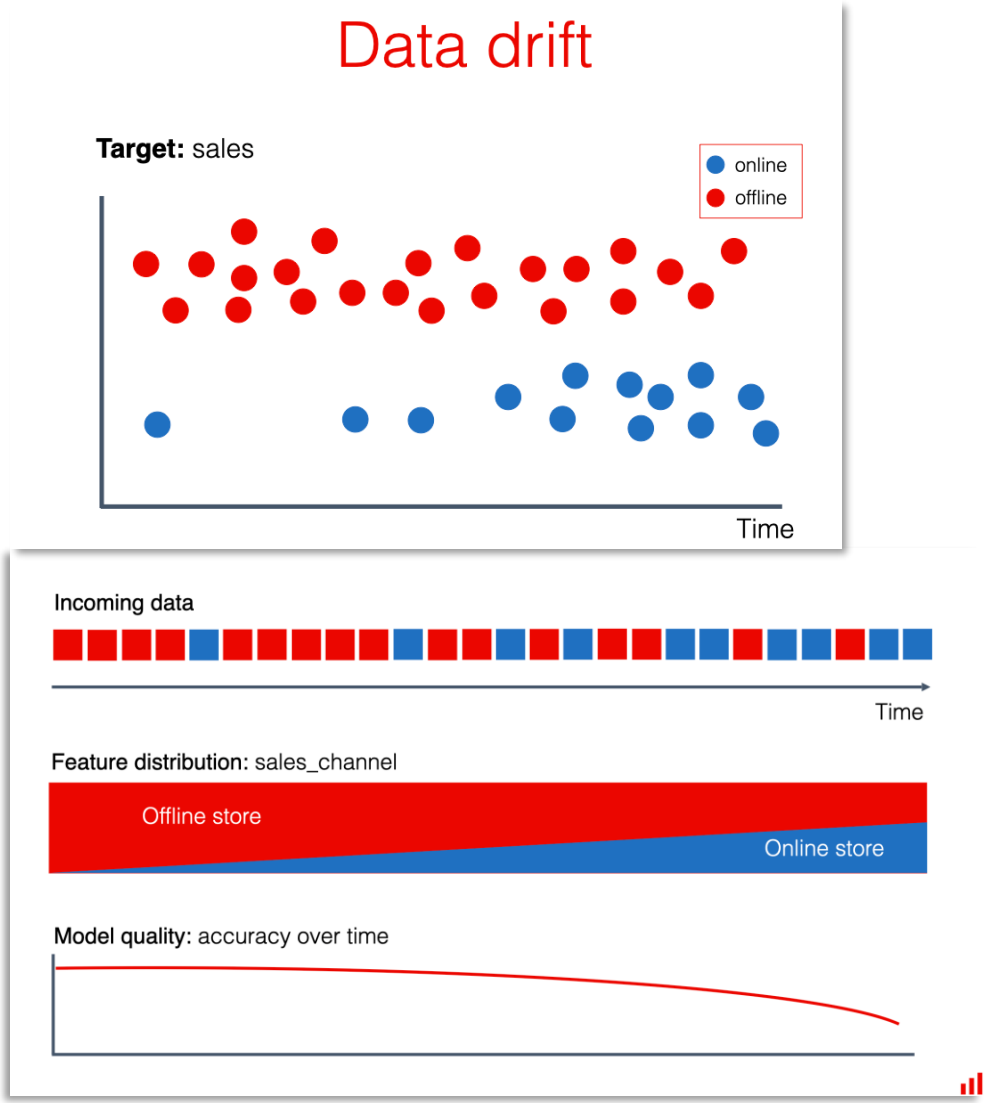
Data Quality Challenges

- **Data quality challenges** can arise from technical issues, organizational dynamics, data complexity, and the ever-evolving landscape of data sources and types.
- Addressing these challenges requires concerted effort across various levels of an organization.
  - **Volume and variety of data**
    - The sheer volume of data generated by modern businesses, combined with the variety of data types and sources, can make managing and maintaining quality a daunting task.
    - Integrating and ensuring consistency across diverse data sets is particularly challenging.
  - **Data silos**
    - Data stored in isolated silos within an organization can lead to inconsistencies, redundancies, and difficulties in achieving a unified view of data.
    - Breaking down these silos to ensure seamless data flow and integrity is a significant challenge.

Data Quality Challenges

- **Evolving data**
  - Data is not static; it changes and evolves over time.
  - Maintaining data quality in the face of changing business processes, regulatory requirements, and market conditions requires flexible and adaptive data management strategies.
- **Human error**
  - Data entry, interpretation, and management are prone to human error.
  - Even small mistakes can propagate through systems, leading to significant data quality issues.

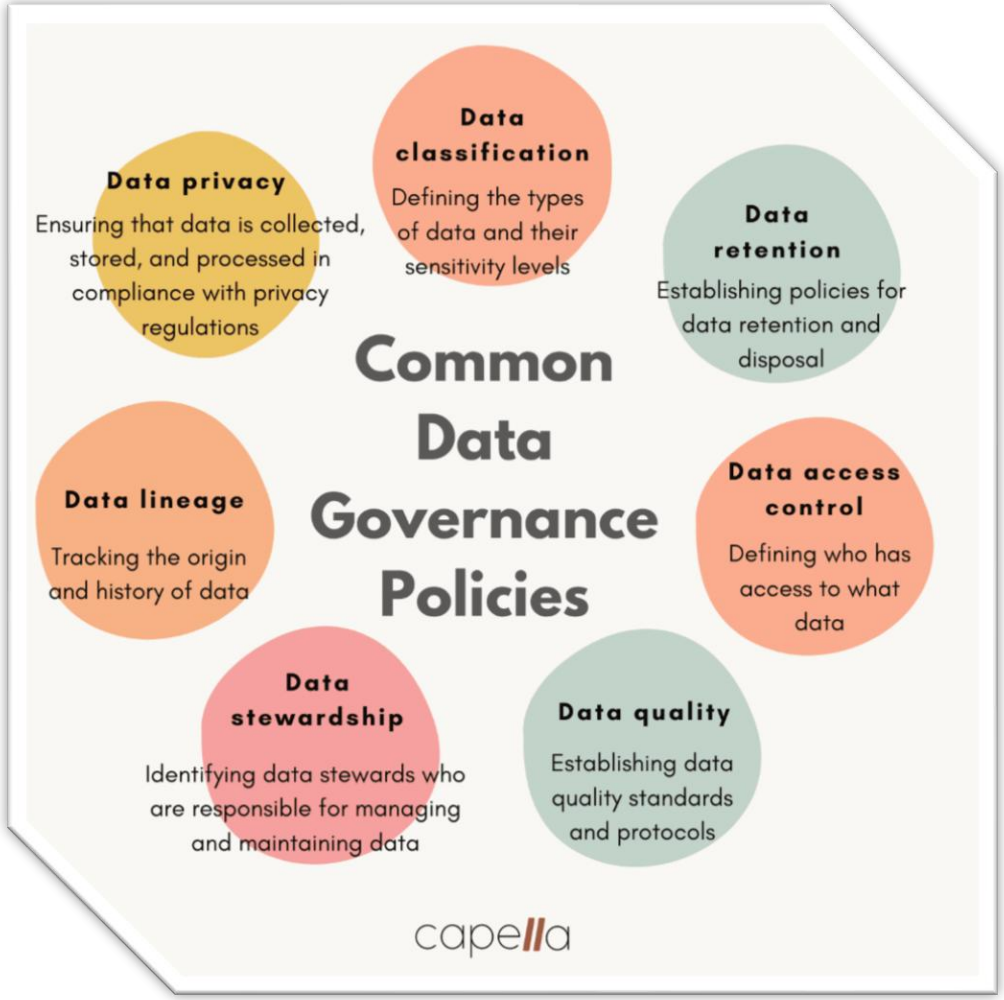
Source: Evidently AI



Data Quality Challenges

- **Lack of comprehensive data governance**
  - Without a robust data governance framework, it's challenging to establish standards, roles, policies, and procedures necessary for maintaining data quality.
  - Data governance provides the structure needed to address data quality systematically.

… from collection, processing, storage, use, security, and management of data



Data Quality Challenges

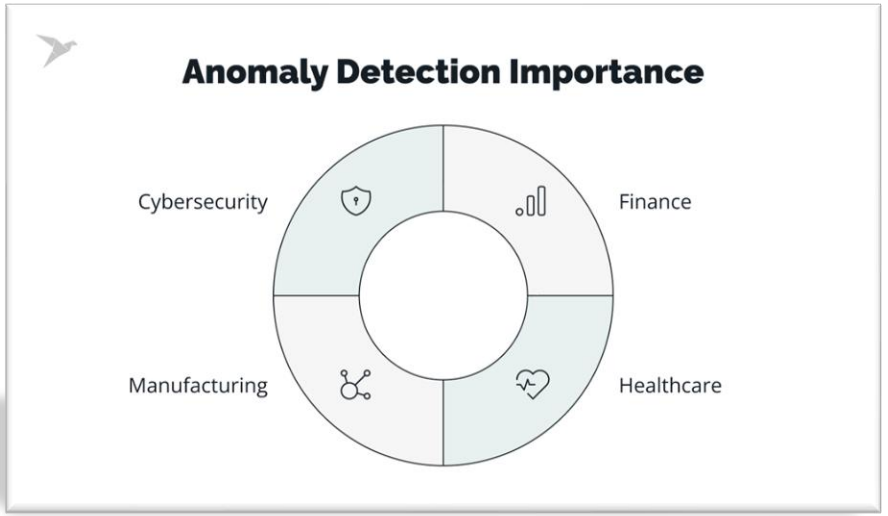
- **Complexity of data integration**
  - Integrating data from various sources, each with its formats, structures, and quality standards, can introduce errors and inconsistencies, complicating data quality efforts.
- **Inadequate data quality tools**
  - The lack of effective tools for data quality management can hinder an organization's ability to detect, correct, and prevent data quality issues.
  - Investment in the right tools and technologies is crucial for maintaining high data quality.
- **Poor data quality awareness**
  - A lack of awareness or understanding of the importance of data quality across the organization can result in inadequate prioritization of data quality initiatives.
  - Cultivating a data-quality culture is essential for overcoming this challenge.

Data Quality Challenges

- **Regulatory compliance**
  - Keeping up with and adhering to an ever-changing landscape of regulatory requirements related to data can be challenging.
  - Non-compliance can lead to data quality issues and legal penalties.
- **Resource constraints**
  - Allocating the necessary resources, including time, budget, and skilled personnel, to data quality initiatives can be difficult, especially for organizations with tight budgets or competing priorities.

Data Anomalies

- **Data anomalies** refer to irregularities or deviations in data that can indicate errors, inconsistencies, or unusual occurrences that deviate from expected patterns.
- Identifying and addressing data anomalies is crucial for maintaining data quality, ensuring accurate analysis, and supporting effective decision-making.
- There are primarily three types of data anomalies
  - Point Anomalies
  - Contextual Anomalies
  - Collective Anomalies



https://www.techmagic.co/blog/ai-anomaly-detection

Fraud detection
- Identifying fraudulent transactions
- Mitigating risks and financial losses

Healthcare
- Monitoring patient data for early signs of diseases
- Early detecting anomalies of health issues

Industrial systems
- Predicting equipment failures
- Optimizing operational efficiency

Predictive maintenance
- Anticipate maintenance needs
- Optimize asset utilization and extend its life
- Reduce maintenance costs

Data Anomalies

- **Point Anomalies**
  - A point anomaly occurs when a single data point significantly deviates from the rest of the data set.
  - This type of anomaly is the simplest form and is often detected through threshold-based methods or statistical analysis.
  - Such anomalies might indicate data entry errors, fraud, or other significant events.

In-class participation:

Task 1:

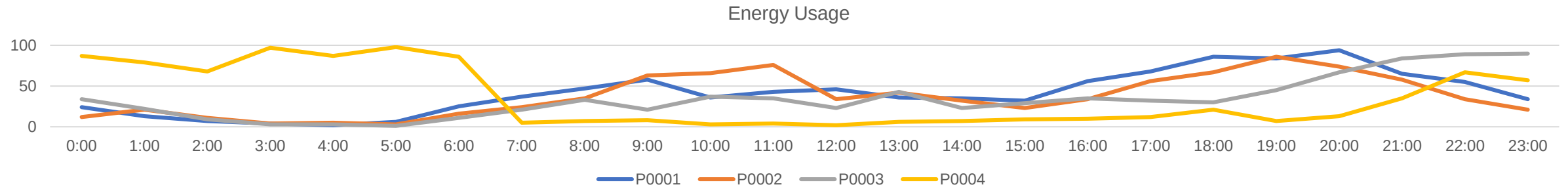
Find out the possible data anomalies in this table.

Staff_ID	First_Name	Last_Name	Level	Work_Hour
S001	John	Smith	D	6
S002	Kate	Joyce	C	8
S003	Mary	Wen	D	6
S004	Jenny	Wood	D	6
S005	Jon	Dolly	E	4
S006	Amy	Yeewood	A	10
S007	Addy	Zhang	B	9
S008	Allen	Fan	B	9
S009	James	Vu	A	10
S010	Anddy	Lee	D	500
S011	Jane	Jones	C	8
S012	Mike	Giacometti	C	8
S013	Anna	Nord	E	4
S014	Sunny	Johnson	E	4
S015	Ross	Hart	A	10

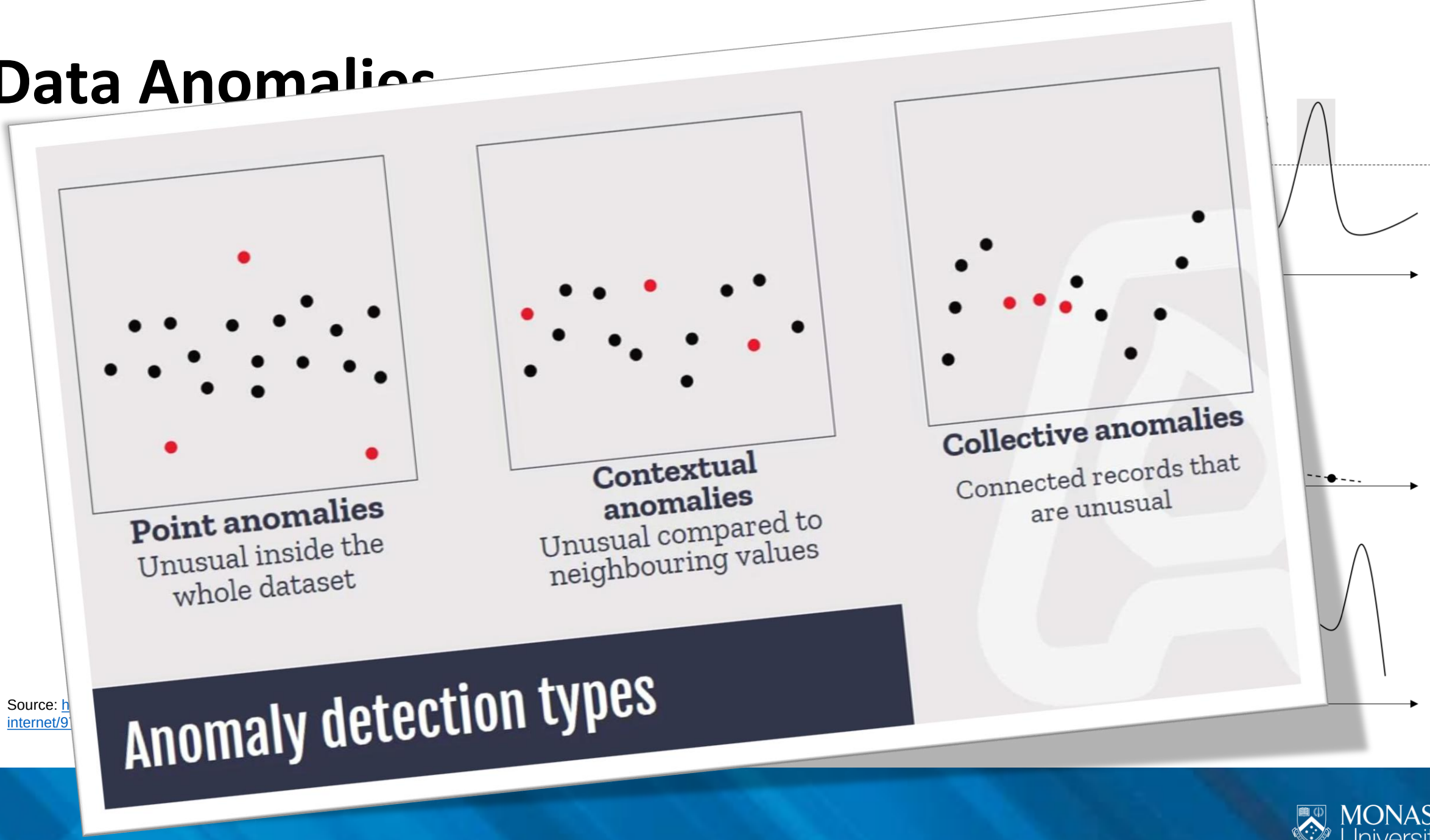
Data Anomalies

- **Contextual Anomalies**
  - Contextual anomalies, also known as conditional anomalies, occur when a data point is anomalous within a specific context or condition but not otherwise.
  - These anomalies can only be identified within a specific context, such as time or space.
  - Identifying contextual anomalies requires understanding the context and conditions under which data is expected to behave in a certain way.

	Energy Usage																						
Property	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
P0001	24	13	7	4	2	6	25	37	47	58	36	43	46	36	35	32	56	68	86	84	94	65	55	34
P0002	12	21	11	4	5	3	16	24	35	63	66	76	34	42	32	23	34	56	67	86	74	58	34	21
P0003	34	22	9	3	3	1	11	21	33	21	37	35	23	43	23	29	35	32	30	45	67	84	89	90
P0004	56	43	21	35	37	32	43	26	11	21	35	14	22	17	16	9	23	97	63	59	66	46	78	89



Data Anomalies



Source: h
internet/9

Data Anomalies

- Detecting anomalies can be challenging, especially in large or complex datasets.
- Techniques for anomaly detection include statistical methods, machine learning models, and domain-specific rules or thresholds.
- Once detected, it's important to investigate anomalies to determine their cause, which could range from simple data entry errors to indications of systemic issues or fraudulent activity.

Data Quality Issues

- **Data quality issues** refer to problems in the data that can negatively impact its usefulness, reliability, and accuracy for decision-making, analysis, and operational processes.
- These issues often arise from various sources throughout the data lifecycle, from collection and storage to processing and analysis.
- Addressing these issues is critical for organizations to ensure that their data assets provide value and support their objectives.

Data Quality Issues

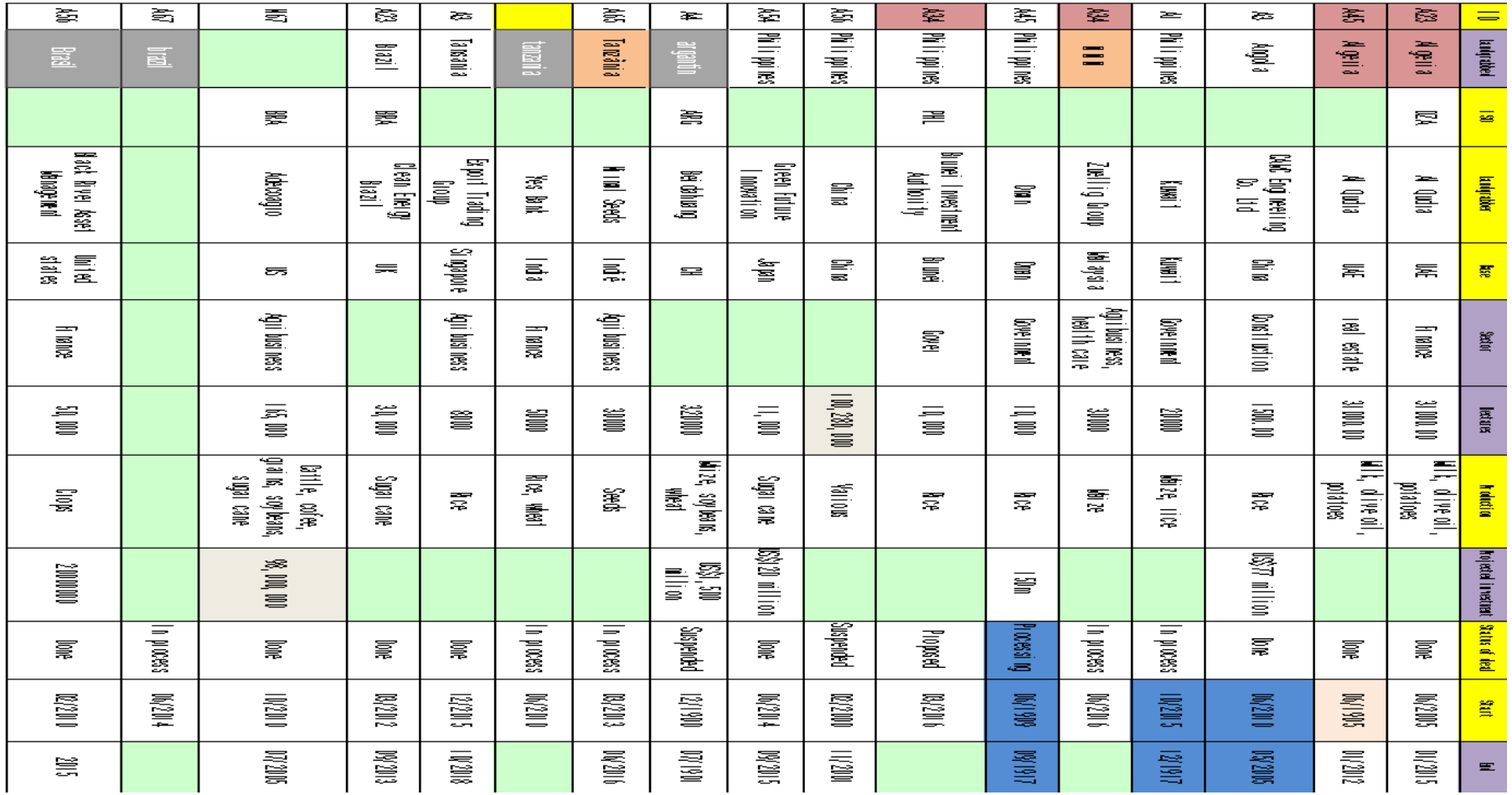
- Looking for errors
  - Incomplete data
  - Inaccurate data
  - Inconsistent data
  - Data Duplication

ID	Landgrabbed	ISO	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Start	End
A23	Algeria	DZA	Al Qudra	UAE	Finance	31000.00	Milk, olive oil, potatoes		Done	06/2005	01/2015
A45	Algeria		Al Qudra	UAE	real estate	31000.00	Milk, olive oil, potatoes		Done	06/1905	01/2012
A3	Angola		CAMC Engineering Co. Ltd	China	Construction	1500.00	Rice	US$77 million	Done	06/2010	05/2015
A1	Philippines		Kuwait	Kuwait	Government	20000	Maize, rice		In process	10/2015	12/1917
A34	[redacted]		Zuellig Group	Malaysia	Agribusiness, health care	30000	Maize		In process	06/2016	08/2020
A45	Philippines		Oman	Oman	Government	10,000	Rice	150m	Processing	06/1909	09/1917
A34	Philippines	PHL	Brunei Investment Authority	Brunei	Gover	10,000	Rice		Proposed	03/2016	
A56	Philippines		China	China		100,280,000	Various		Suspended	02/2000	11/2001
A54	Philippines		Green Future Innovation	Japan		11,000	Sugar cane	US$120 million	Done	06/2014	09/2015
A4	argantin	ARG	Beidahuang	CH		320000	Maize, soybeans, wheat	US$1,500 million	Suspended	12/1900	07/1901
A65	Tanzania		Nirmal Seeds	India	Agribusiness	30000	Seeds		In process	03/2013	06/2016
	Tanzania		Yes Bank	India	Finance	50000	Rice, wheat		In process	06/2010	06/2017
A3	Tansania		Export Trading Group	Singapore	Agribusiness	8000	Rice		Done	12/2015	10/2018
A23	Brazil	BRA	Clean Energy Brazil	UK		30,000	Sugar cane		Done	03/2012	09/2013
M67		BRA	Adecoagro	US	Agribusiness	165,000	Cattle, cofee, grains, soybeans, sugar cane	98,000,000	Done	10/2010	07/2005
A67	brazil	BRA	Archer Daniels Midland	US	Agribusiness	12,000	Oil palm		In process	06/2014	01/2015
A56	Brasil		Black River Asset Management	United states	Finance	50,000	Crops	20000000	Done	02/2010	2015

ID	Landgrabbed	ISO	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Start	End
A23	Algeria	DZA	Al Qudra	UAE	Finance	31000.00	Milk, olive oil, potatoes		Done	06/2005	01/2015
A45	Algeria		Al Qudra	UAE	real estate	31000.00	Milk, olive oil, potatoes		Done	06/1905	01/2012
A3	Angola		CAMC Engineering Co. Ltd	China	Construction	1500.00	Rice	US$77 million	Done	06/2010	05/2005
A1	Philippines		Kuwait	Kuwait	Government	20000	Maize, rice		In process	10/2015	12/1917
A34	[redacted]		Zuellig Group	Malaysia	Agribusiness, health care	30000	Maize		In process	06/2016	08/2020
A45	Philippines		Oman	Oman	Government	10,000	Rice	150m	Processing	06/1909	09/1917
A34	Philippines	PHL	Brunei Investment Authority	Brunei	Gover	10,000	Rice		Proposed	03/2016	
A56	Philippines		China	China		100,280,000	Various		Suspended	02/2000	11/2001
A54	Philippines		Green Future Innovation	Japan		11,000	Sugar cane	US$120 million	Done	06/2014	09/2015
A4	arganti n	ARG	Beidahuang	CH		320000	Maize, soybeans, wheat	US$1,500 million	Suspended	12/1900	07/1901
A65	Tanzánia		Nirmal Seeds	Indiê	Agribusiness	30000	Seeds		In process	03/2013	06/2016
	Tanzania		Yes Bank	India	Finance	50000	Rice, wheat		In process	06/2010	06/2017
A3	Tansania		Export Trading Group	Singapore	Agribusiness	8000	Rice		Done	12/2015	10/2018
A23	Brazil	BRA	Clean Energy Brazil	UK		30,000	Sugar cane		Done	03/2012	09/2013
N67		BRA	Adecoagro	US	Agribusiness	165,000	Cattle, cofee, grains, soybeans, sugar cane	98,000,000	Done	10/2010	07/2005
A67	brazil	BRA	Archer Daniels Midland	US	Agribusiness	12,000	Oil palm		In process	06/2014	01/2015
A56	Brasil		Black River Asset Management	United states	Finance	50,000	Crops	20000000	Done	02/2010	2015

Data Quality Issues

- Looking for errors
  - Incomplete data
  - Inaccurate data
  - Inconsistent data
  - Data Duplication



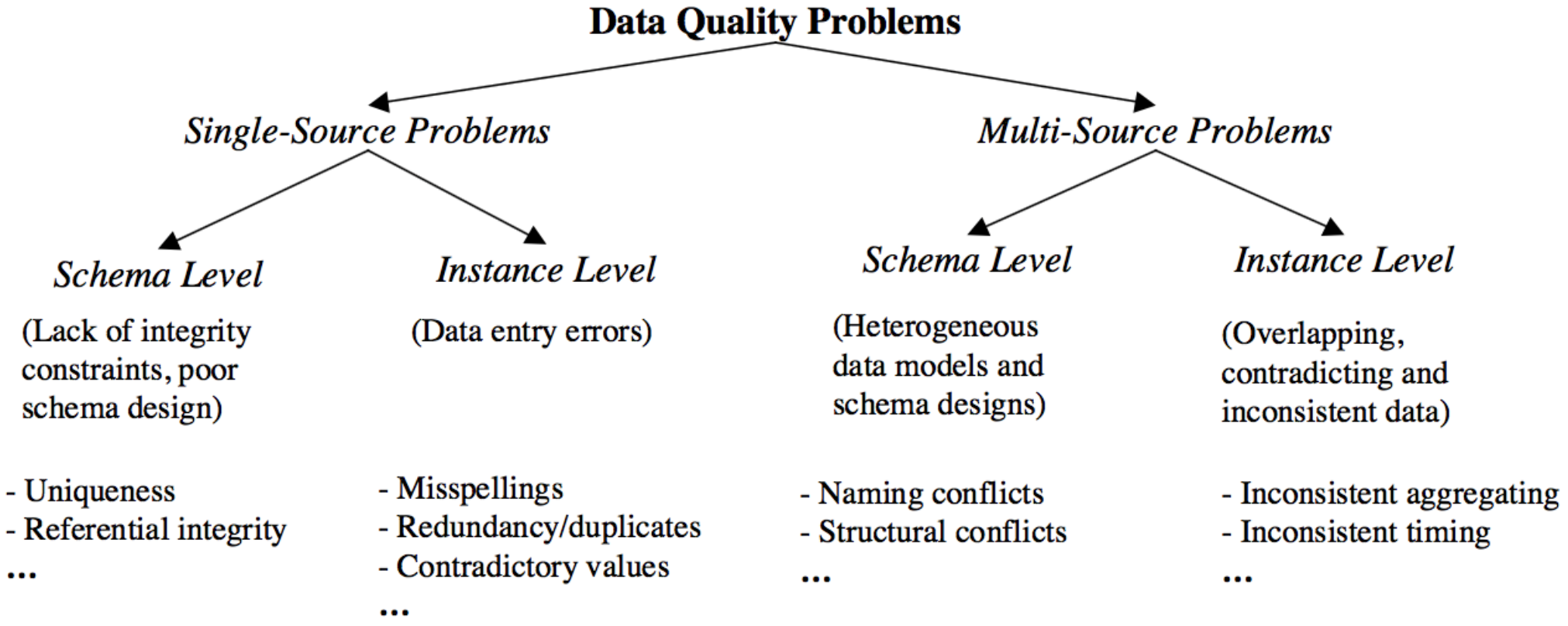
ID	Landgrabbed	ISO	Landgrabber	Base	Sector	Hectares	Production	Projected investment	Status of deal	Start	End
A23	Algeria	DZA	Al Qudra	UAE	Finance	31 000.00	Milk, olive oil, potatoes		Done	06/2005	01/2015
A45	Algeria		Al Qudra	UAE	real estate	31 000.00	Milk, olive oil, potatoes		Done	06/1905	01/2012
A3	Angola		CAMC Engineering Co. Ltd	China	Construction	1500.00	Rice	US$77 million	Done	06/2010	05/2005
A1	Philippines		Kuwait	Kuwait	Government	20000	Maize, rice		In process	10/2015	12/1917
A34	[illegible]		Zuellig Group	Malaysia	Agribusiness, health care	30000	Maize		In process	06/2016	
A45	Philippines		Oman	Oman	Government	10,000	Rice	150m	Processing	06/1909	09/1917
A34	Philippines	PHL	Brunei Investment Authority	Brunei	Gover	10,000	Rice		Proposed	03/2016	
A56	Philippines		China	China		100,280,000	Various		Suspended	02/2000	11/2001
A54	Philippines		Green Future Innovation	Japan		11,000	Sugar cane	US$120 million	Done	06/2014	09/2015
A4	argantin	ARG	Beidahuang	CH		320000	Maize, soybeans, wheat	US$1,500 million	Suspended	12/1900	07/1901
A65	Tanzânia		Mirnal Seeds	Indiê	Agribusiness	30000	Seeds		In process	03/2013	06/2016
	tanzania		Yes Bank	India	Finance	50000	Rice, wheat		In process	06/2010	
A3	Tansania		Export Trading Group	Singapore	Agribusiness	8000	Rice		Done	12/2015	10/2018
A23	Brazil	BRA	Clean Energy Brazil	UK		30,000	Sugar cane		Done	03/2012	09/2013
M67		BRA	Adecoagro	US	Agribusiness	165,000	Cattle, cofee, grains, soybeans, sugar cane	98,000,000	Done	10/2010	07/2005
A67	brazil								In process	06/2014	
A56	Brasil		Black River Asset Management	United states	Finance	50,000	Crops	20000000	Done	02/2010	2015

Data Quality Issues

- Poor data standardization
- Lack of data timeliness
- Data relevance issues
- Poor data security and privacy
- Complex data structures
- Data accessibility issues

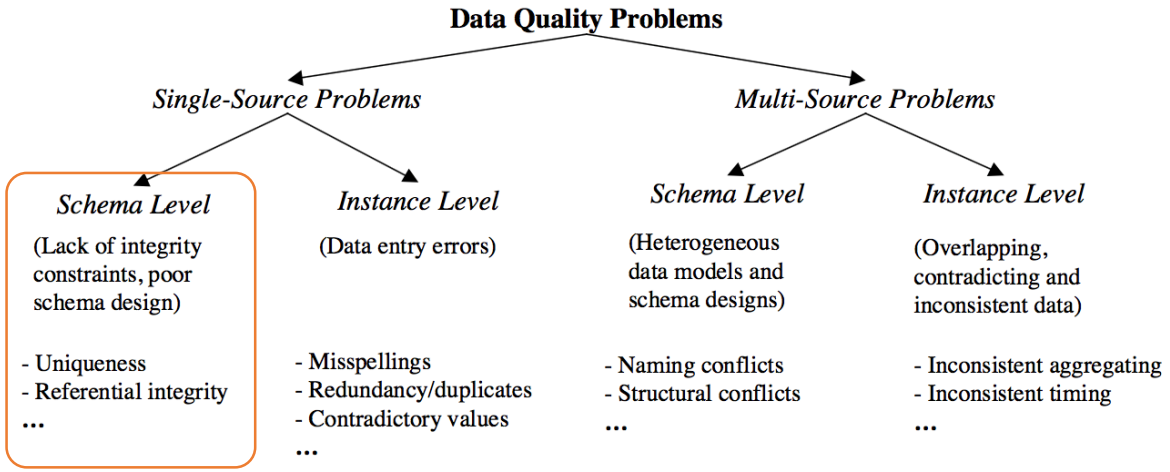
Data Quality Issues

- Data quality problems can be categorized based on data sources.



From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Single-Source Problems

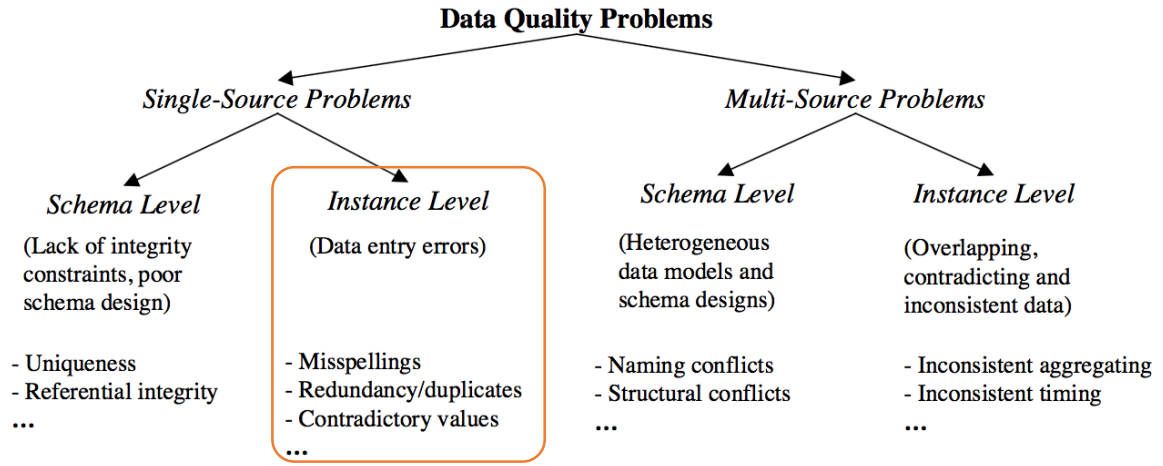


Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Illegal values	bdate=30.13.70	values outside of domain range
Record	Violated attribute dependencies	age=22, bdate=12.02.70	age = (current date – birth date) should hold
Record type	Uniqueness violation	emp_1=(name="John Smith", SSN="123456") emp_2=(name="Peter Miller", SSN="123456")	uniqueness for SSN (social security number) violated
Source	Referential integrity violation	emp=(name="John Smith", deptno=127)	referenced department (127) not defined

Table 1. Examples for single-source problems at schema level (violated integrity constraints)

From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

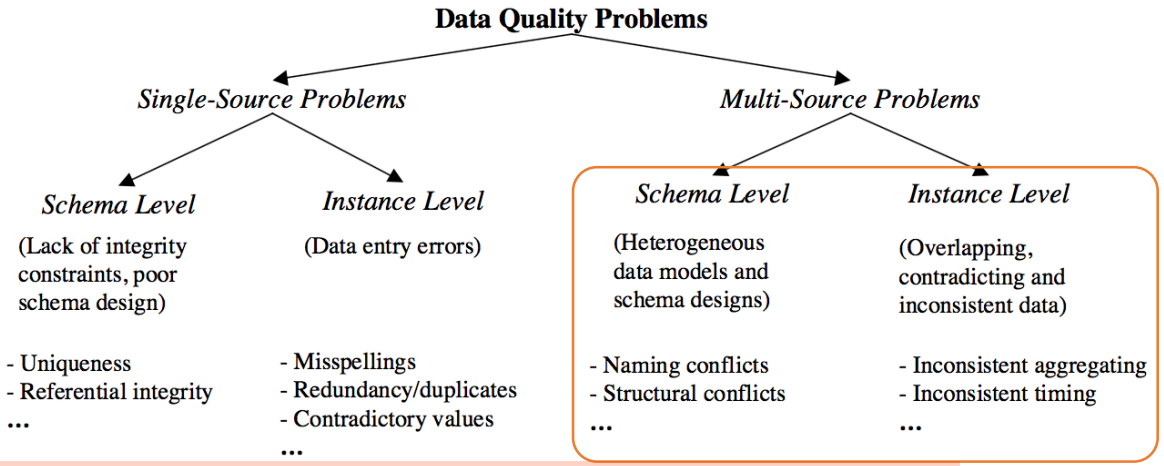
Single-Source Problems



Scope/Problem		Dirty Data	Reasons/Remarks
Attribute	Missing values	phone=9999-999999	unavailable values during data entry (dummy values or null)
	Misspellings	city="Liipzig"	usually typos, phonetic errors
	Cryptic values, Abbreviations	experience="B"; occupation="DB Prog."	
	Embedded values	name="J. Smith 12.02.70 New York"	multiple values entered in one attribute (e.g. in a free-form field)
	Misfielded values	city="Germany"	
Record	Violated attribute dependencies	city="Redmond", zip=77777	city and zip code should correspond
Record type	Word transpositions	$name_1$= "J. Smith", $name_2$="Miller P."	usually in a free-form field
	Duplicated records	emp_1=(name="John Smith",...); emp_2=(name="J. Smith",...)	same employee represented twice due to some data entry errors
	Contradicting records	emp_1=(name="John Smith", bdate=12.02.70); emp_2=(name="John Smith", bdate=12.12.70)	the same real world entity is described by different values
Source	Wrong references	emp=(name="John Smith", deptno=17)	referenced department (17) is defined but wrong

From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Multi-Source Problems



Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

In-class participation:

Task 2:

Let's try to find out the problems in these examples.

Schema level, there are **name conflicts** (synonyms Customer/Client, Cid/Cno, Sex/Gender) and **structural conflicts** (different representations for names and addresses).

Instance level, we note that there are **different gender representations** ("0"/"1" vs. "F"/"M") and presumably a **duplicate record** (Kristen Smith).

The latter observation also reveals that while **Cid/Cno are both source-specific identifiers,** their contents are not comparable between the sources; different numbers (11/493) may refer to the same person while different persons can have the same number (24).

Requires both schema integration and data cleaning; the third table shows a possible solution. Note that the schema conflicts should be resolved first to allow data cleaning, in particular detection of duplicates based on a uniform representation of names and addresses, and matching of the Gender/Sex values.

From "Data Cleaning: Problems and Current Approaches" by Rahm and Do

Data Quality Issues

- Data quality problems can be categorized based on data type.
  - **Syntactical Anomalies**: format and values
    - Lexical errors
      - data format discrepancies in terms of database
      - e.g., spelling errors, typos in terms of linguistics.
    - Domain format errors
      - inconsistent value format of an attribute
      - e.g., Buntine, Wray Lindsay v.s. Wray L. Buntine
    - Irregularities
      - the non-uniform use of values, units and abbreviations?
      - e.g., salary in difference currencies.

Data Quality Issues

- Data quality problems can be categorized based on data type.
  - Syntactical Anomalies: format and values
  - **Semantic Anomalies**: comprehensiveness and non-redundancy
    - Integrity constraint violations
    - Contradictions
      - violation of dependencies between attributes
      - e.g., AGE and DOB.
    - Duplicates: observations representing the same entity.
    - Invalid observations → logical inconsistencies like 1000 year old person.

Data Quality Issues

- Data quality problems can be categorized based on data type.
  - Syntactical Anomalies: format and values
  - Semantic Anomalies: comprehensiveness and non-redundancy
  - **Coverage Anomalies**: **missing values**
    - Missing values: due to omissions while collecting the data
    - Missing observations

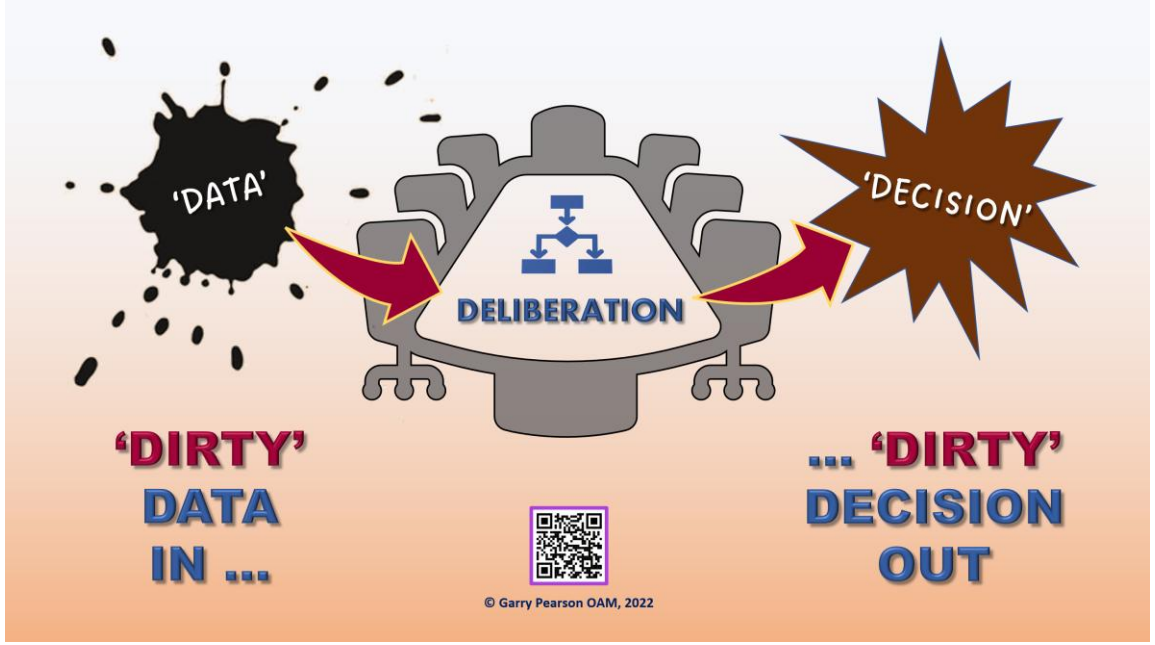
They're **gaps or distortions in how well your data covers the population, space, time, or category combinations you're supposed to analyze.** Different from row-level missing values, these are **missing records** (or lopsided records) for whole slices of the universe.

Common types

- **Under-coverage:** parts of the population/time/space never (or rarely) recorded.
  Ex: Store #42 has no Sundays; Q2 has zero rows.
- **Over-coverage:** segments are overrepresented (duplicates, bursty collectors).
  Ex: Device sends the same reading 10×.
- **Sparse regions:** data exists but far below expected density.
- **Boundary gaps/drift:** start/end dates shifted.
- **Group-attribute coverage holes:** a field is present overall but missing within a subgroup.
  Ex: "income" absent for ages >65.

Dirty Data

- Dirty data manifests itself in three different ways:
  - missing data
  - not missing but wrong data
  - not missing and not wrong but unusable



Source: https://polgovpro.blog/2022/06/22/dirty-data-in-dirty-decisions-out/

Dirty Data

- Dirty data manifests itself in three different ways:
  - **missing data**
    - Missing data where there is no Null-not-allowed constraint → *The column is optional. It's OK for it to be blank because the business doesn't require that value for every row.*
      - Examples: middle_name, apartment_number, secondary_phone, marketing_opt_in_date (only exists if the user opted in)
    - Missing data where Null-not-allowed constraint should be enforced → *The value is mandatory for correctness, integrity, safety, or compliance. A blank here is an error.*
      - Examples: order_id, transaction_timestamp, amount, email_for_login, patient_identifier, allergy_status (recorded as "NKA/Unknown/HasAllergies" — but not left blank).
  - not missing but wrong data
  - not missing and not wrong but unusable

Dirty Data

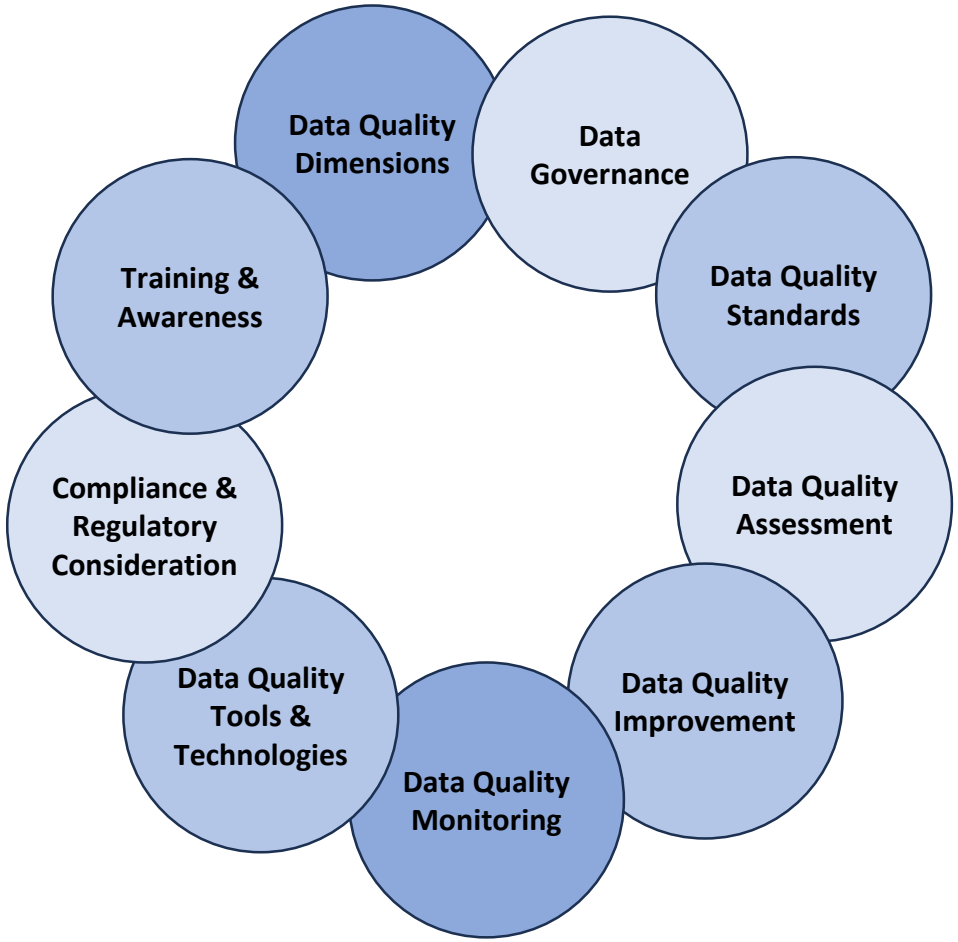
- Dirty data manifests itself in three different ways:
  - missing data
  - **not missing but wrong data**
    - Integrity constraints
      - violation of data type constraint, including value range
      - violation of non-null uniqueness constraint, i.e., duplicated data
      - violation of referential integrity
      - Wrong categorical data
      - Outdated temporal data
      - Inconsistent spatial data
    - Data Entry error involving a single table
      - Data entry error involving a single field: erroneous entry, misspelling, extraneous data
      - Data entry error involving multiple fields: entry into wrong fields, wrong derived-field data
  - not missing and not wrong but unusable

Dirty Data

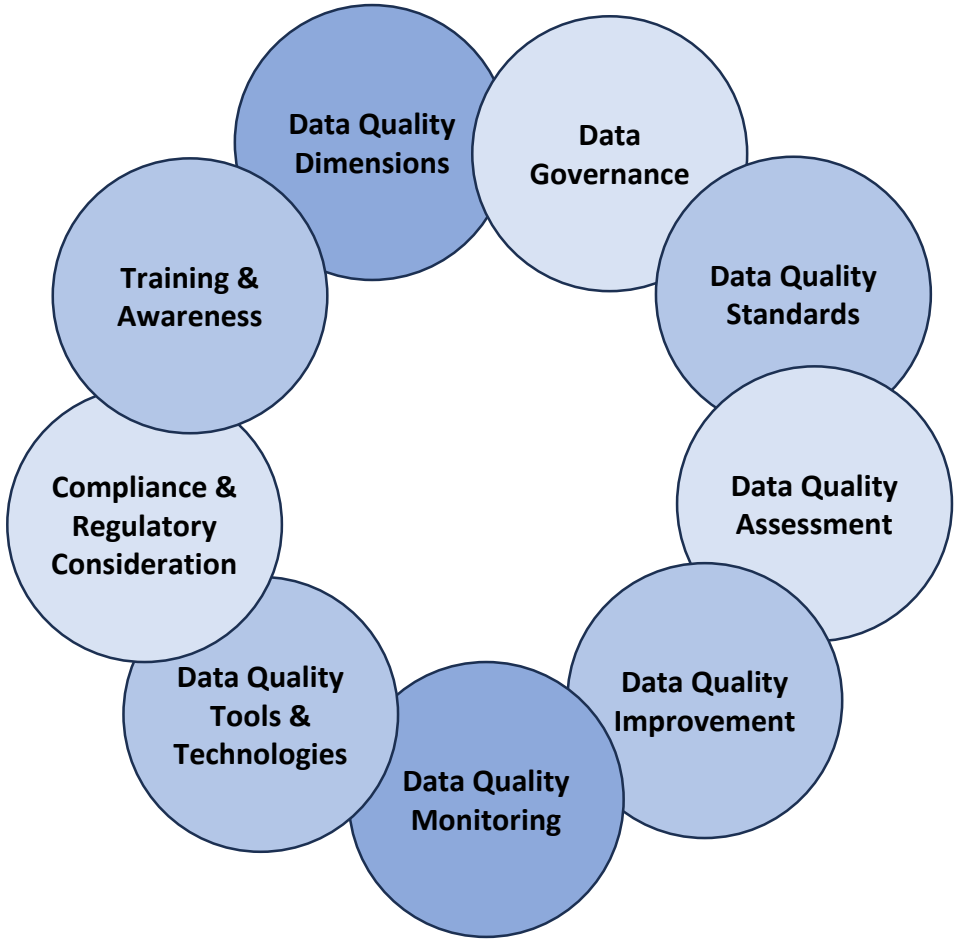
- Dirty data manifests itself in three different ways:
  - missing data
  - not missing but wrong data
  - **not missing and not wrong but unusable**
    - Different data for the same entity across multiple databases
      - Ambiguous data due to the use of abbreviation (Dr. for doctor or drive)
      - Incomplete context (e.g., Sydney of Australia or Canada)
      - The use of abbreviation (e.g., ste for suite, rd for road, st for street, etc)
      - Alias/nick name (e.g., Bill Clinton, President Clinton)
      - Encoding formats (e.g, ASCII, ...)
      - Representations (e.g., negative number, precision, fraction)
      - Measurement units (e.g., data, time, currency, weight, area, etc.)
      - Uses of special characters (e.g., space, dash, parenthesis in phone numbers) in concatenated data

Data Quality Management Frameworks

- **Data Quality Management Frameworks** are structured approaches to ensuring that an organization's data is accurate, complete, reliable, and suitable for its intended use.
- These frameworks provide the principles, policies, standards, processes, and metrics necessary to manage the quality of data effectively throughout its lifecycle.
- Implementing a robust Data Quality Management Framework is essential for organizations that rely on data for decision-making, compliance, and operational efficiency.



Data Quality Management Frameworks



- The implementation of a Data Quality Management Framework is an iterative process that requires engagement from stakeholders across the organization.
- It begins with a clear understanding of the organization's data quality needs and involves the development of a tailored framework that addresses those needs.
- Success relies on strong governance, clear communication, effective use of technology, and a commitment to continuous improvement.

The Role of Machine Learning in Data Quality

- **Machine Learning** (ML) play increasingly vital roles in enhancing data quality by automating and refining the processes involved in identifying, correcting, and preventing data quality issues.
- ML's capabilities enable organizations to handle vast volumes of data more efficiently, uncover hidden insights, and improve the overall integrity and value of their data assets.
  - **Automated Error Detection**
  - **Data Cleansing**
  - **Predictive Data Quality**
  - **Enhanced Data Matching and Merging**
  - **Natural Language Processing (NLP)**
  - **Data Governance and Metadata Management**
  - **Data Enrichment**
  - **Continuous Monitoring and Improvement**

Summary & To-do List

- Please download and read the materials provided on Moodle.
- Review the content learnt from Week 7.

- Assessments
  - Complete Group Assessment 1 (Due: 11:55 pm, Monday, 15 September 2025)

    → All the group members must click submit button.
    → In text citations in the report for all the references.

- Next week: Data Cleansing