Techniques for data cleansing

Readings for today

- Müller, H., & Freytag, J. C. (2003). Problems, Methods, and Challenges in Data Cleansing. Berlin: HUB-IB-164.

Topics

1. Data cleansing

2. Types of anomalies

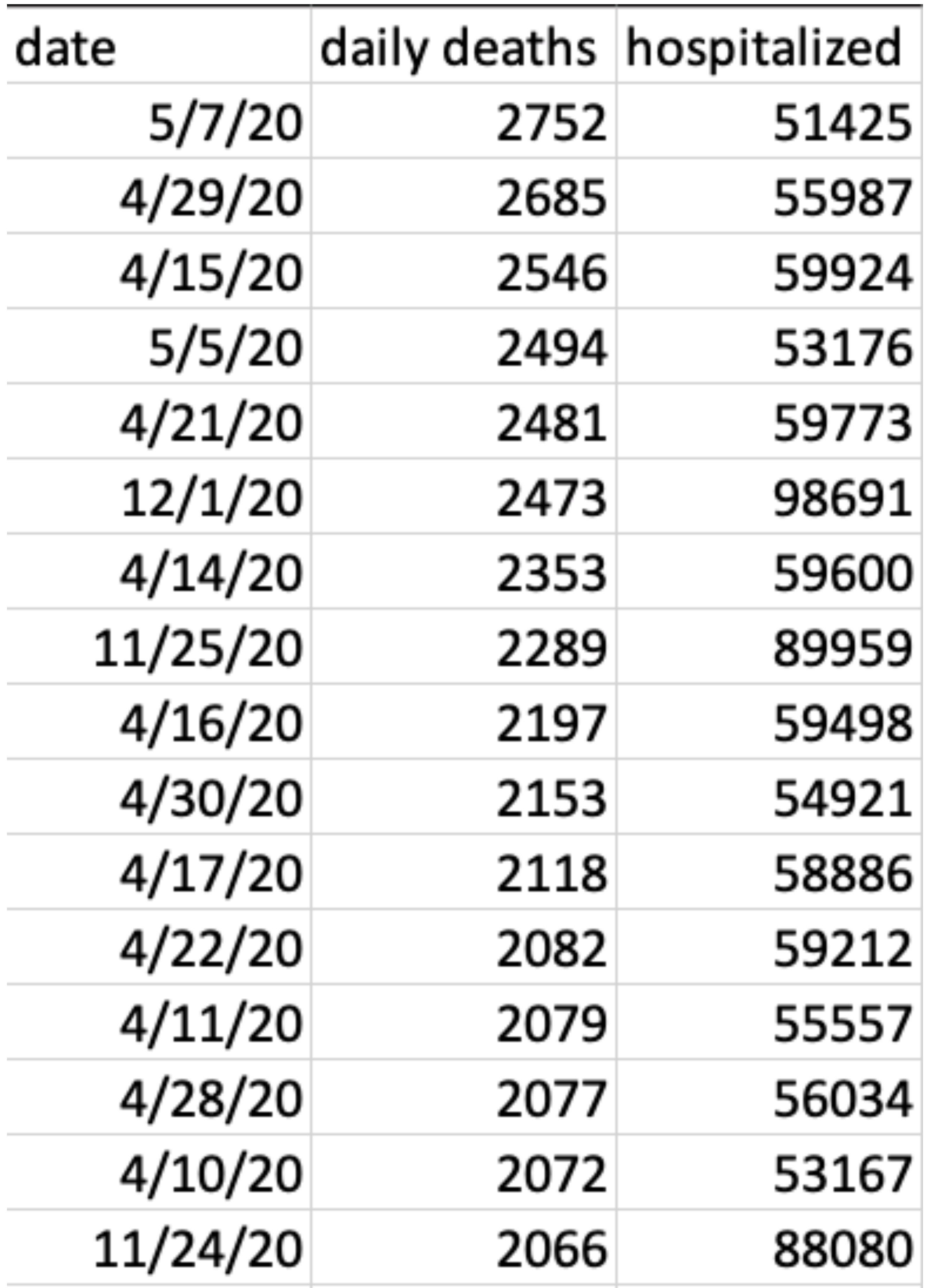
3. Data quality criterion

Data cleansing

Real data is messy

Once you have a tidy data table, there are lots of ways that errors can be present.

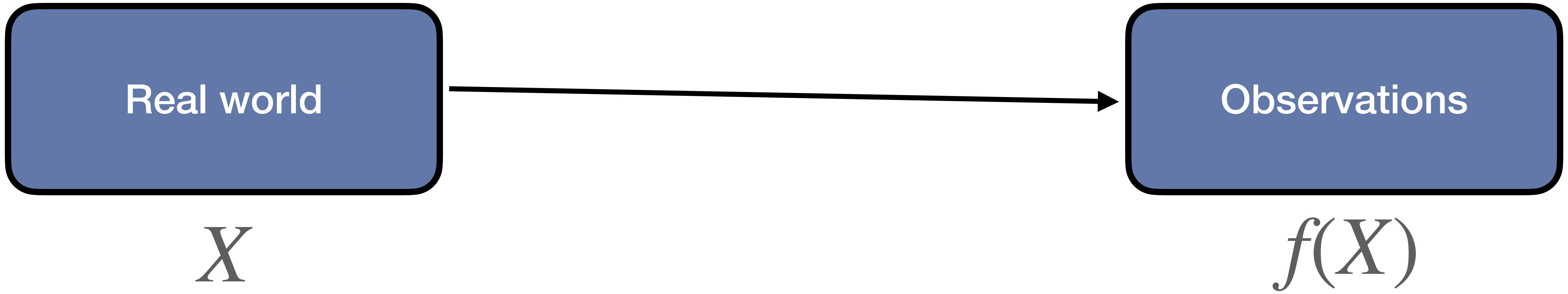
- Some are obvious.
- Some are hidden



date	daily deaths	hospitalized
5/7/20	2752	51425
4/29/20	2685	55987
4/15/20	2546	59924
5/5/20	2494	53176
4/21/20	2481	59773
12/1/20	2473	98691
4/14/20	2353	59600
11/25/20	2289	89959
4/16/20	2197	59498
4/30/20	2153	54921
4/17/20	2118	58886
4/22/20	2082	59212
4/11/20	2079	55557
4/28/20	2077	56034
4/10/20	2072	53167
11/24/20	2066	88080

What is data cleansing?

Data cleansing: The identification & accounting for anomalies in your data.



It is assumed that your observations have a veridical mapping to entities in the real world.

Definition of terms

Anomaly: Property of data that renders it an incorrect representation of the world.

Data: Symbolic representation of information.

Tuple: List of discrete values from a finite set.

Feature vector: Collection of observations.

(Müller & Freytag 2003)

Types of anomalies

Types of anomalies

Syntactical: Errors in labels or formats.

Semantic: Errors in the fundamental value of observations themselves.

Coverage: Gaps in the collection process.

(Müller & Freytag 2003)

Types of anomalies

Syntactical:

- Lexical errors: Discrepancy between structure of data & format.

Subject ID	Trial	Accuracy	RT
S001	1	correct	incorrect
S001	2	correct	599
S001	3	incorrect	240
S002	1	incorrect	692
S002	2	correct	476
S002	3	correct	301

(Müller & Freytag 2003)

Types of anomalies

Syntactical:

- **Lexical errors:** Discrepancy between structure of data & format.
- **Domain format errors:** Value for attribute does not match domain.

Name	Trial	Cond	RT
Smith, John	1	A	380
Doe, Jane	2	B	599
Smith, Karen	3	A	240
Pain Max	1	A	692

Missing comma in name format

(Müller & Freytag 2003)

Types of anomalies

Syntactical:

- Lexical errors: Discrepancy between structure of data & format.
- Domain format errors: Value for attribute does not match domain.
- Irregularities: Non-uniform use of values, units, or observations.

Subject ID	Trial	Cond	RT
S001	1	A	380
S001	2	B	599
S001	3	A	240
S002	1	A	692
S002	2	B	0.476
S002	3	A	301

Units change from milliseconds to seconds

(Müller & Freytag 2003)

Types of anomalies

Semantic:

- Integrity constraint violations: Value does not match constraints of attribute.

Subject ID	Age	DOB
S001	41	4/14/1979
S002	24	12/25/1996
S003	-1	1/1/1984
S004	28	1/24/1991

Age cannot be negative

(Müller & Freytag 2003)

Types of anomalies

Semantic:

- Integrity constraint violations: Value does not match constraints of attribute.
- Contradictions: Values violate a dependency.

Subject ID	Age	DOB
S001	41	4/14/1979
S002	24	12/25/1996
S003	37	1/1/1984
S004	68	1/24/1991

Age does not match date of birth

(Müller & Freytag 2003)

Types of anomalies

<u>Semantic</u>:

- **Integrity constraint violations:** Value does not match constraints of attribute.
- **Contradictions:** Values violate a dependency.
- **Duplicates:** 2 or more data points represent the same thing.

Subject ID	Age	DOB
S001	41	4/14/1979
S002	24	12/25/1996
S003	37	1/1/1984
S004	37	1/1/1984

Likely 2 of the same entry

(Müller & Freytag 2003)

Types of anomalies

Semantic:

- Integrity constraint violations: Value does not match constraints of attribute.
- Contradictions: Values violate a dependency.
- Duplicates: 2 or more data points represent the same thing.
- Invalid tuples: General case for all other semantic errors.

Subject ID	Age	DOB
S001	41	4/14/1979
S002	24	12/25/1996
S003	37	1/1/1984
S004	128	1/24/1891

Outlier or extreme value

(Müller & Freytag 2003)

Types of anomalies

Coverage:

- Missing value: Omission of an observation or data point

Subject ID	Age	DOB
S001	41	4/14/1979
S002	24	12/25/1996
S003		1/1/1984
S004	128	1/24/1891

Missing observation

(Müller & Freytag 2003)

Types of anomalies

Coverage:

- **Missing value:** Omission of an observation or data point
- **Missing tuples:** Omission of a full variable or feature.

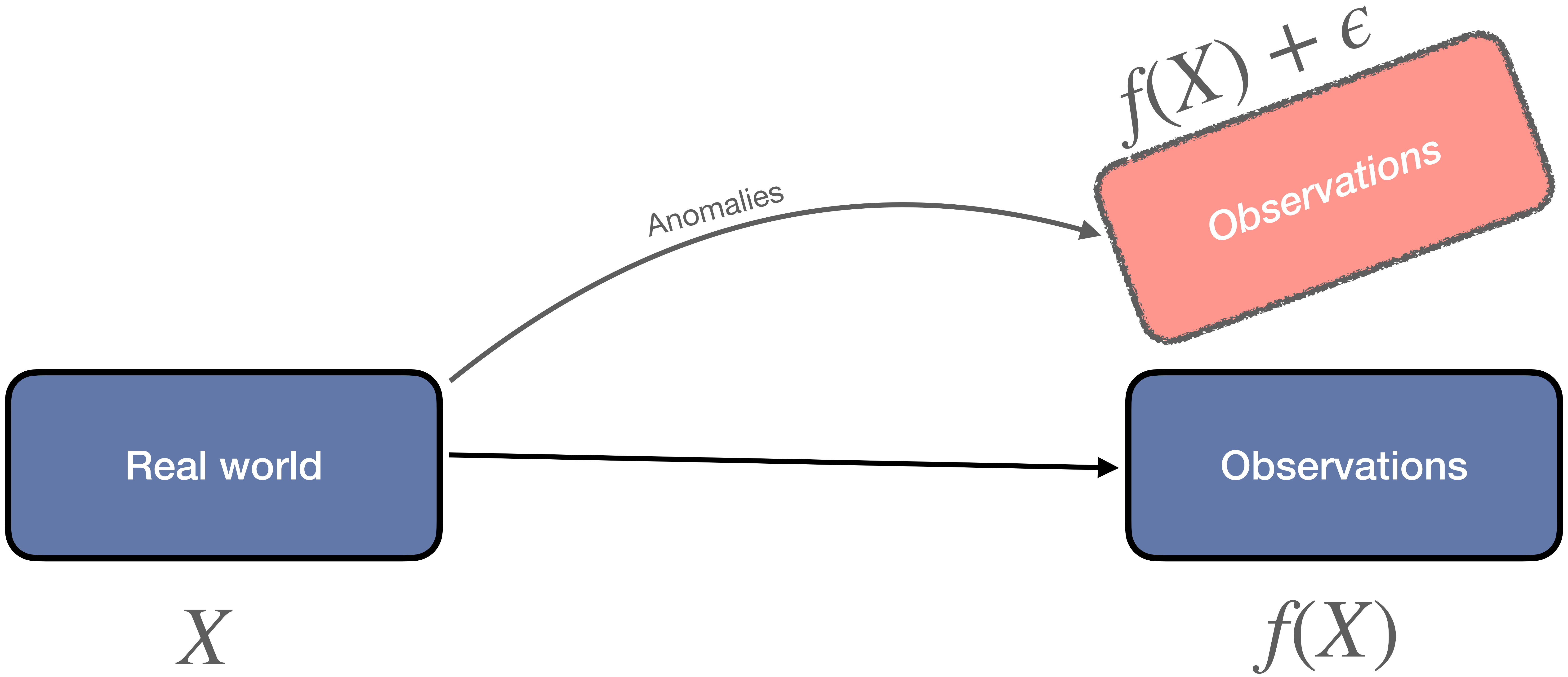
Subject ID	**Age**	**DOB**
S001	41	
S002	24	
S003	37	
S004	128	

Missing variable

(Müller & Freytag 2003)

Data quality criterion

Remember the goal



How do anomalies distort the mapping between world and observations?

Data quality criterion

I. Accuracy: Exact, uniform, & complete representations of the world.

- Integrity: Data set contains representations of all desired aspects of the world.

  - Completeness: All the unique variables & observations are present.

  - Validity: No contradictions or invalid tuples.

(Semantic)

(Müller & Freytag 2003)

Data quality criterion

I. Accuracy: Exact, uniform, & complete representations of the world.

- Integrity: Data set contains representations of all desired aspects of the world. (Semantic)

- Consistency: Data set is uniform & free of contradictions. (Syntactic)

  - Schema conformity: No lexical or domain formatting errors.

  - Uniformity: All observations of the same variable have the same format.

(Müller & Freytag 2003)

Data quality criterion

I. Accuracy: Exact, uniform, & complete representations of the world.

- Integrity: Data set contains representations of all desired aspects of the world. (Semantic)

- Consistency: Data set is uniform & free of contradictions. (Syntactic)

- Density: For all n observations & p variables, you have exactly $n \times p$ values. (Coverage)

(Müller & Freytag 2003)

Data quality criterion

I. Accuracy: Exact, uniform, & complete representations of the world.

- Integrity: Data set contains representations of all desired aspects of the world. (Semantic)
- Consistency: Data set is uniform & free of contradictions. (Syntactic)
- Density: For all n observations & p variables, you have exactly $n \times p$ values. (Coverage)

II. Uniqueness: There are no duplicate entries or variables.

(Müller & Freytag 2003)

Anomaly to quality criterion

Criterion

Anomaly

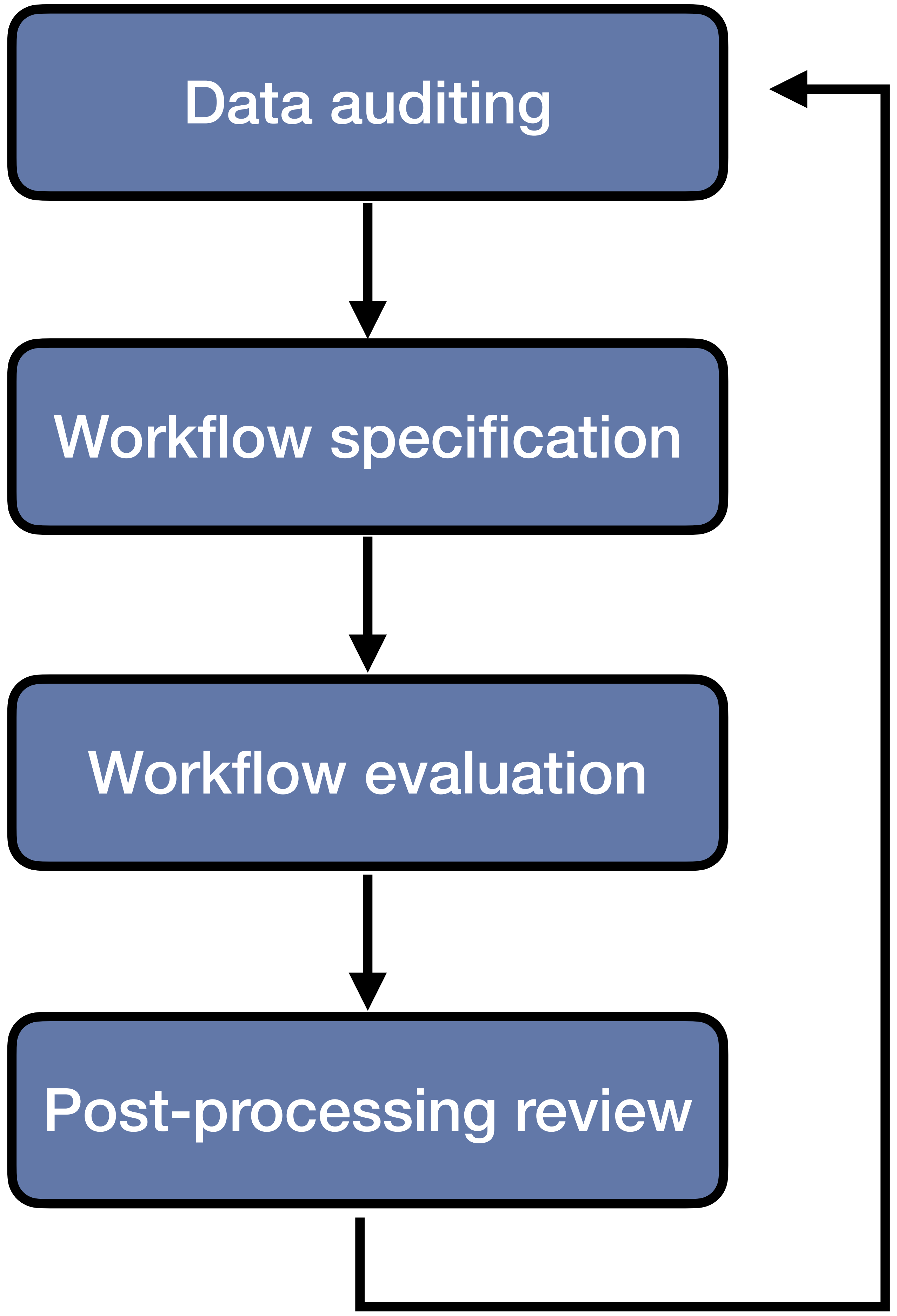
	Completeness	Validity	Schema Conformity	Uniformity	Density	Uniqueness
Lexical Errors		🟦	🟥	🟦	🟦	🟦
Domain format errors		🟦	🟥	🟦		🟦
Irregularities		🟦		🟥		🟦
Constraint Violations		🟥				
Missing Values					🟥	🟦
Missing Tuples	🟥					
Duplicates						🟥
Invalid Tuples		🟥				

🟥 Directly impacts

🟦 Indirectly impacts

(Müller & Freytag 2003)

Steps of data cleansing



- Defines a logic for cleansing pipelines.
- Leave original data untouched.
- Automate as many steps as possible

(Müller & Freytag 2003)

Take home message

Developing a formal process for identifying anomalies, correcting identified anomalies, and evaluating for quality makes your data a more veridical representation of the real word.