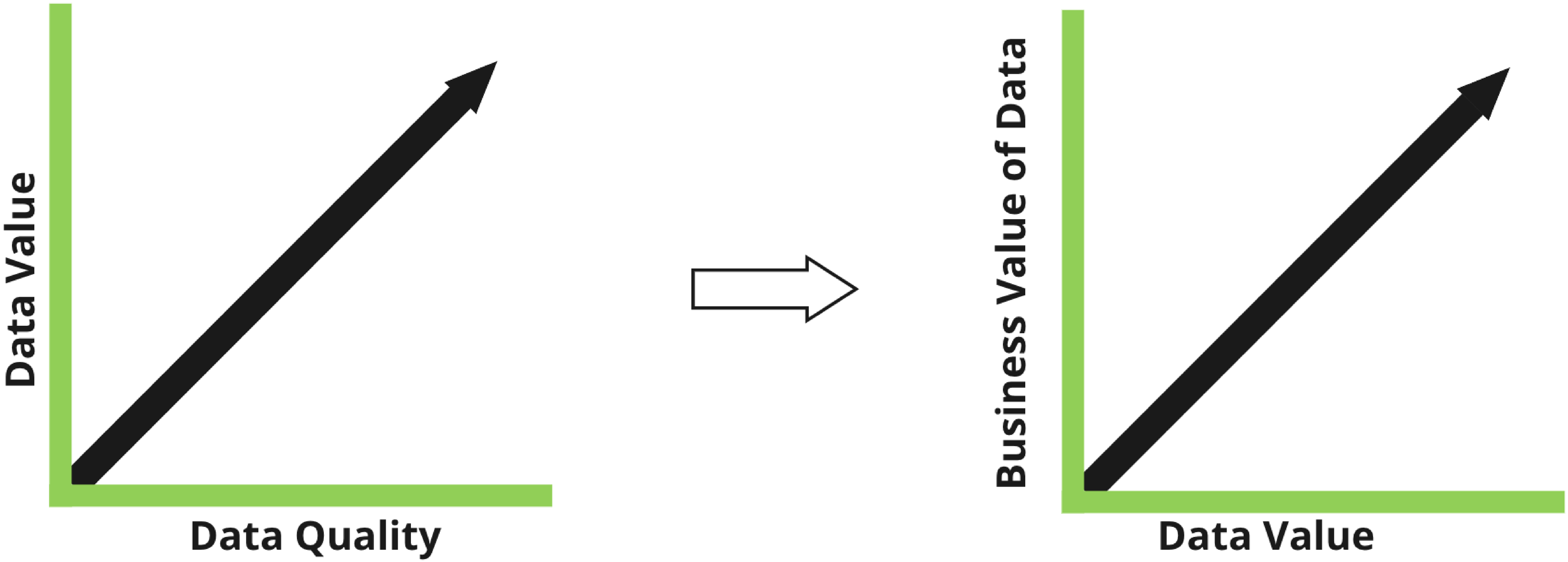
Importance of data quality

INTRODUCTION TO DATA QUALITY

Chrissy Bloom

Head of Enterprise Data Strategy & Governance

Data quality generates business value



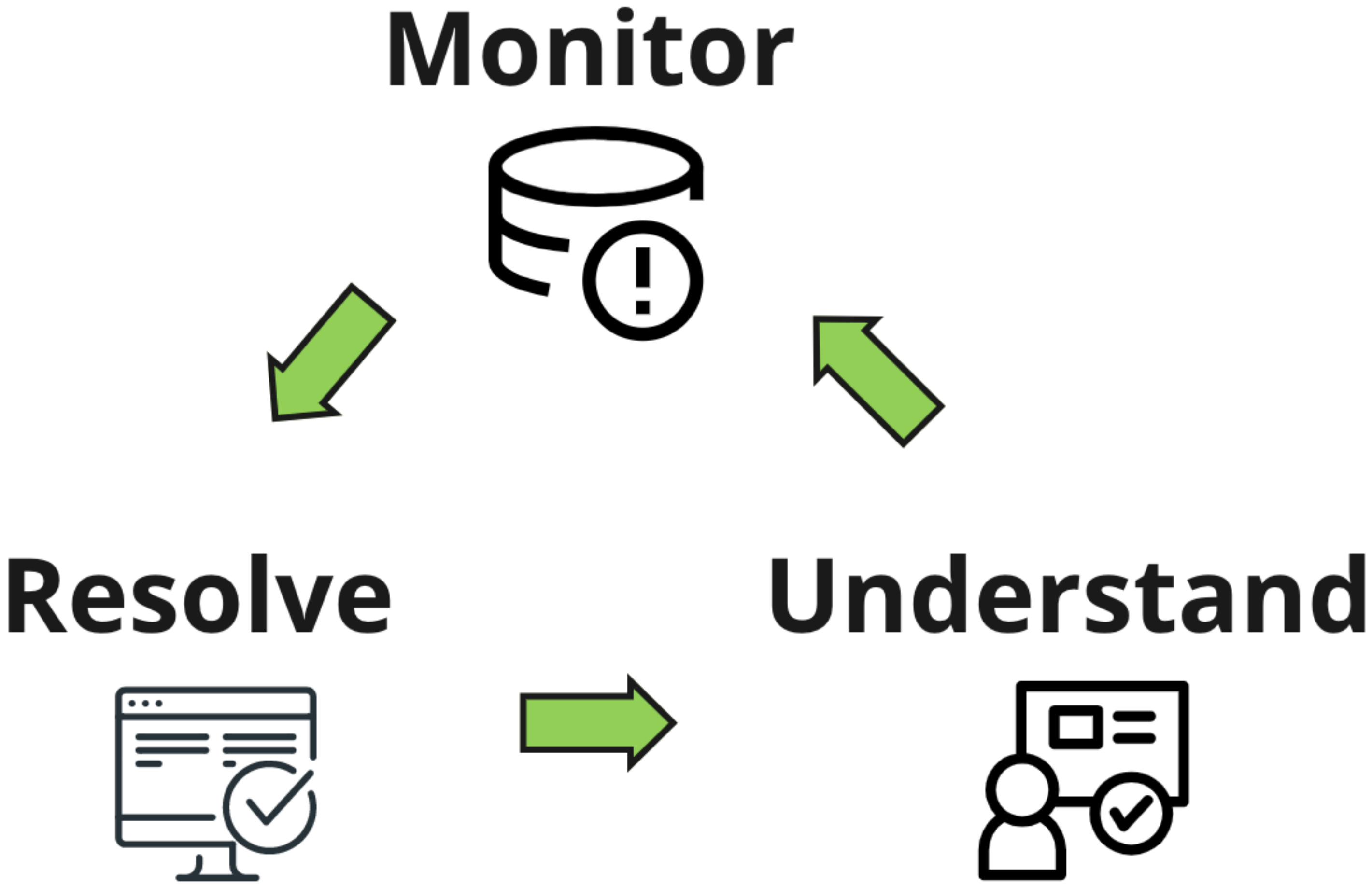
{1}

Activities for good data quality

Data used for decisions and processes:

- Is monitored for data quality
- Has timely issue resolution
- Is produced and consumed by people who understand data quality

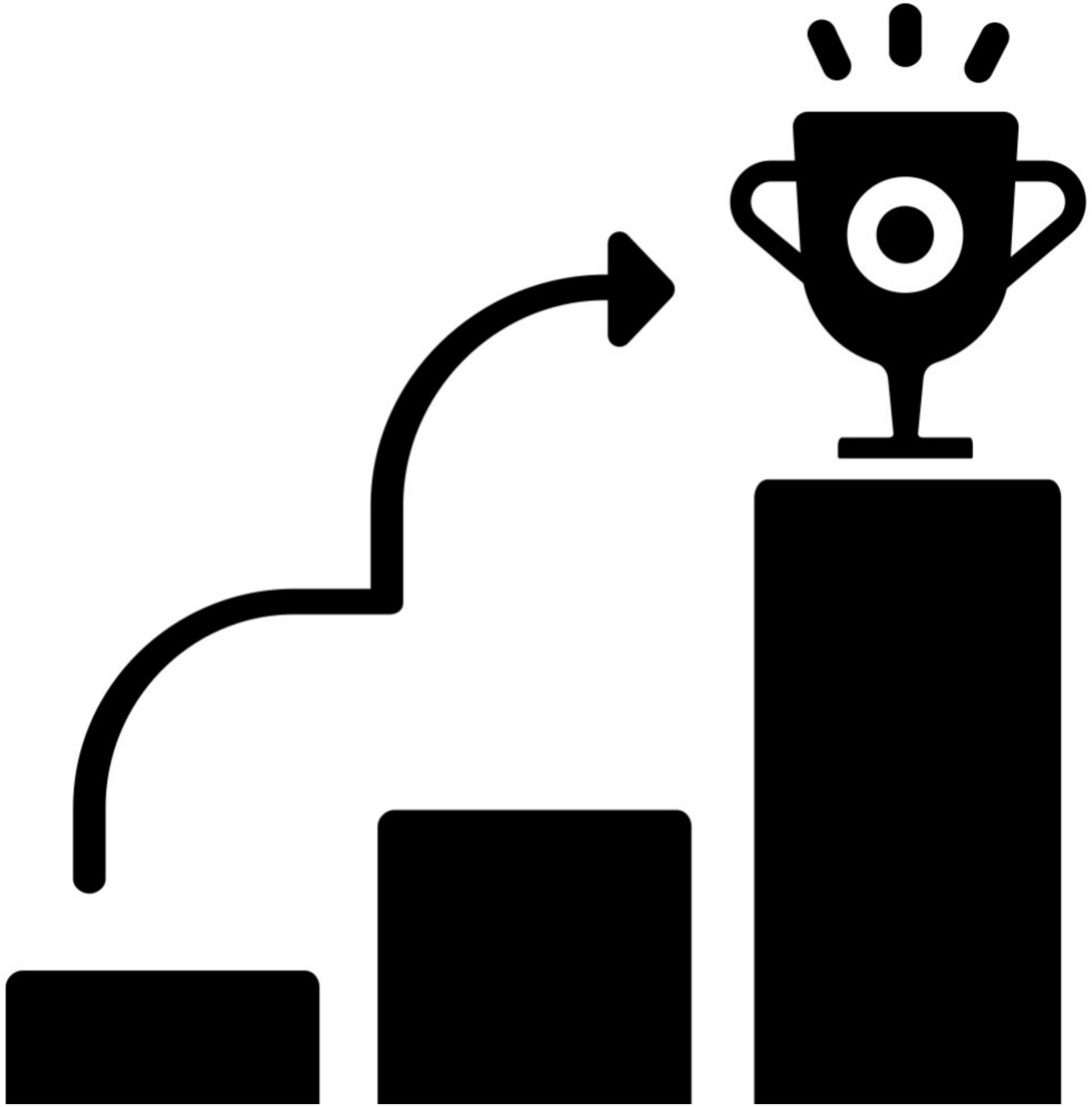
The data quality process is continuous



Value on the offense

Business Value on the Offense:

- Increased competitive edge
  - better decisions
- Increased customer satisfaction
  - better customer interactions



Value on the defense

Business Value on the Defense:

- Increased data risk mitigation
  - faster issue identification
- Increased process efficiencies
  - more streamlined processes

Checking for data quality

How do you know you can trust data to make business decisions?

- Check that it is the latest data available
- Check for duplicate records
- Check that the dataset is complete
- Check for completeness of expected values
- Check for valid values

If data meets your data quality criteria, it is fit for use.

Data Quality Checklist

☒ Timeliness
☒ Uniqueness
☒ Completeness
☒ Validity

Let's practice!

INTRODUCTION TO DATA QUALITY

Data quality terms and concepts

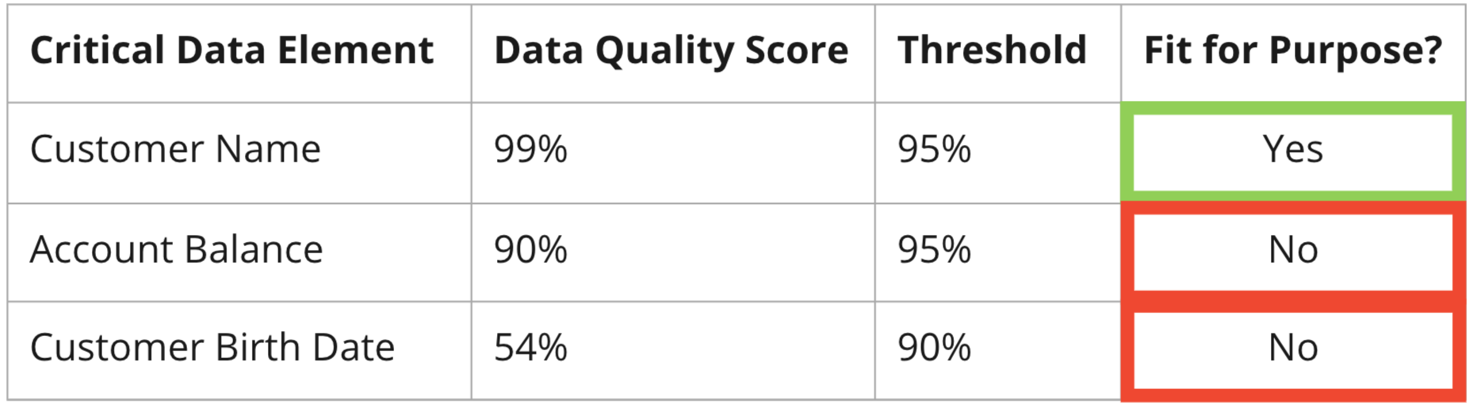
INTRODUCTION TO DATA QUALITY

Chrissy Bloom

Head of Enterprise Data Strategy & Governance

Defining data quality

- **Data Quality:** a measurement of the degree to which data is fit for purpose
- Good data quality = trust in data
  - Better business decisions
  - Better equipped business processes
- Often good data quality is assumed
- Data quality needs to be measured and monitored to ensure that data is fit for use.

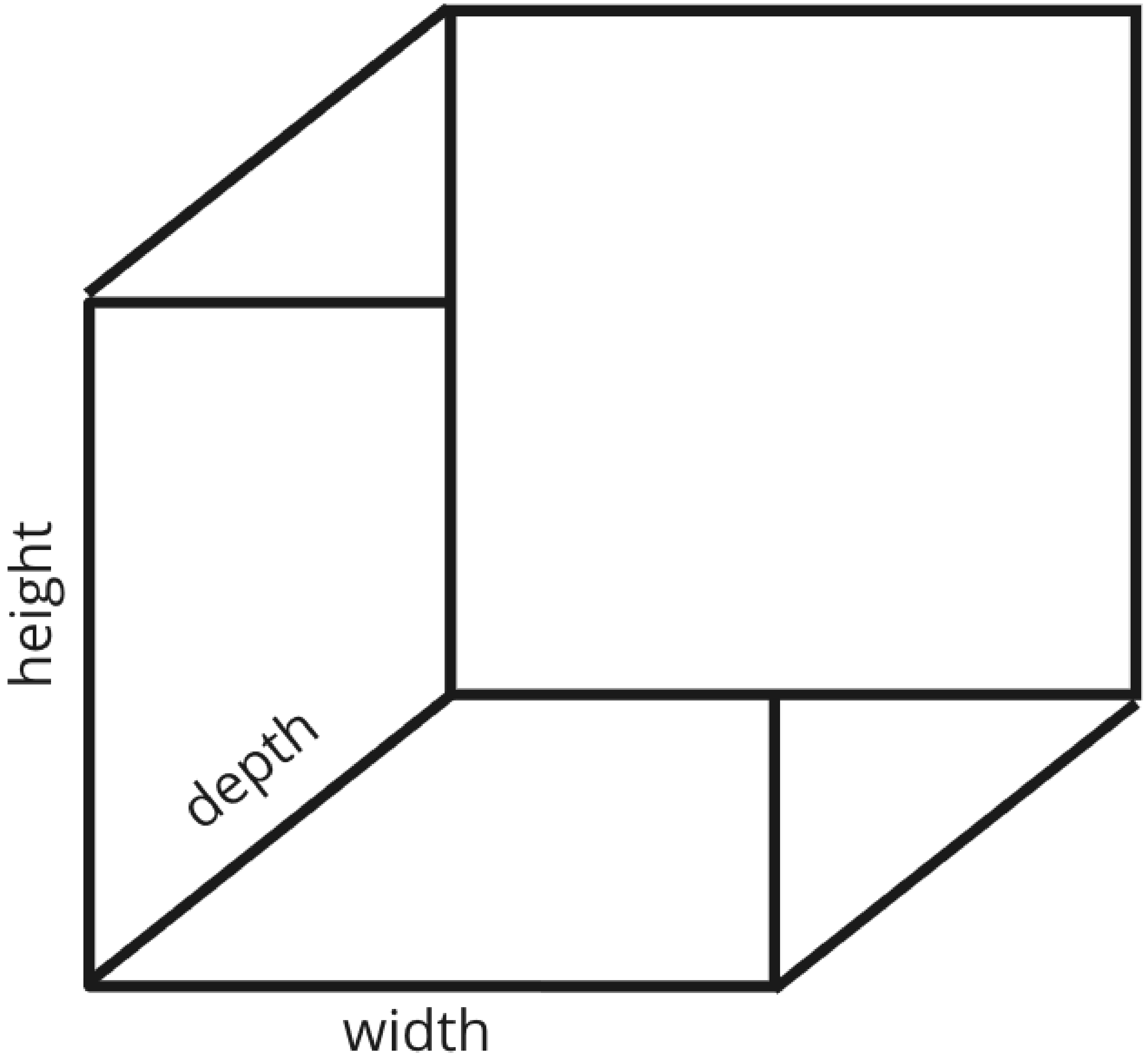


Critical Data Element	Data Quality Score	Threshold	Fit for Purpose?
Customer Name	99%	95%	Yes
Account Balance	90%	95%	No
Customer Birth Date	54%	90%	No

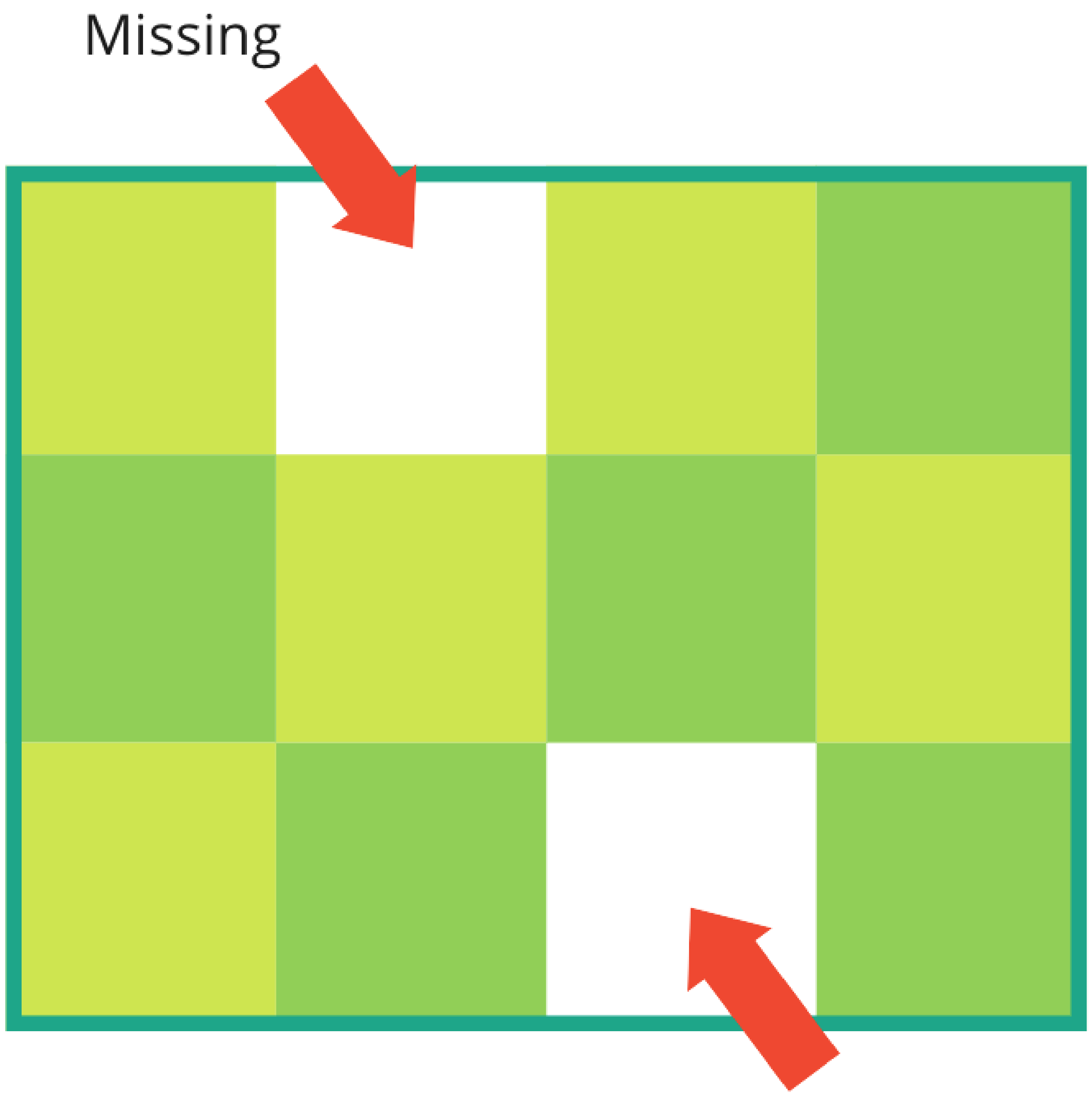
Defining data quality dimensions

Data Quality Dimension: a measurement of a specific attribute of a data's quality

- Use data quality dimensions to quantify how fit for purpose data is.
  - **Completeness**
  - **Validity**
  - **Uniqueness**
  - Consistency
  - Timeliness
  - Accuracy



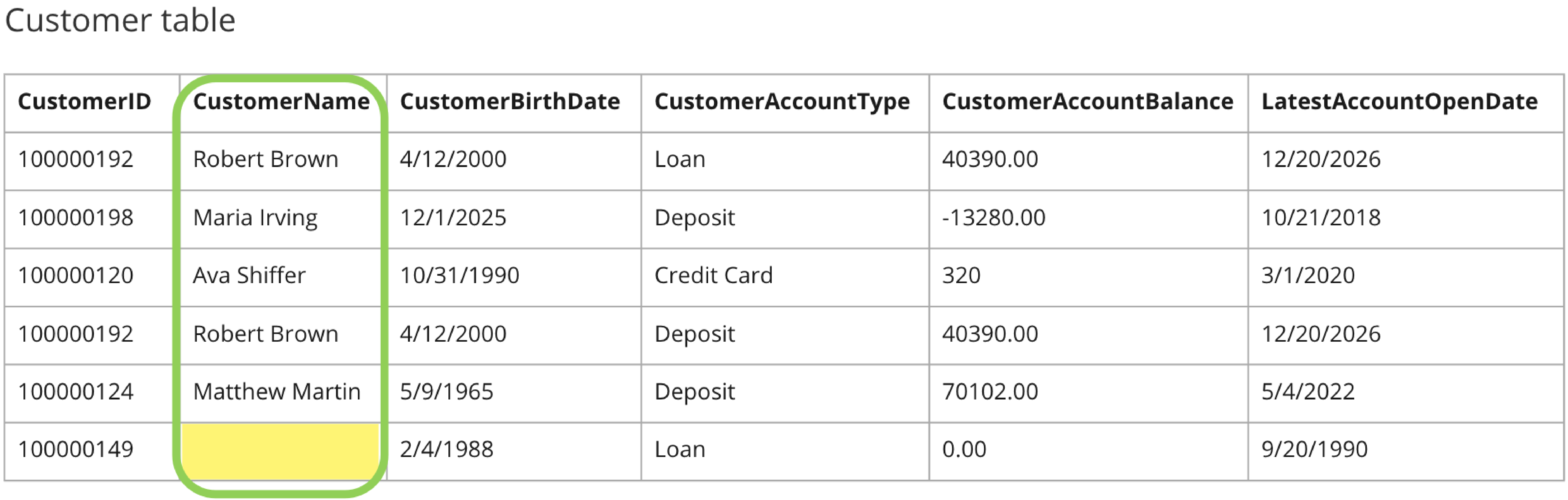
Completeness as a data quality dimension



Completeness:

- Dataset level: measures the degree to which all expected records in a dataset are present.
- Data element level: measures the degree to which all records have data populated when expected.
- Business issues due to incomplete data:
  - Numbers may be skewed
  - Customers may be affected

Completeness example



Customer table

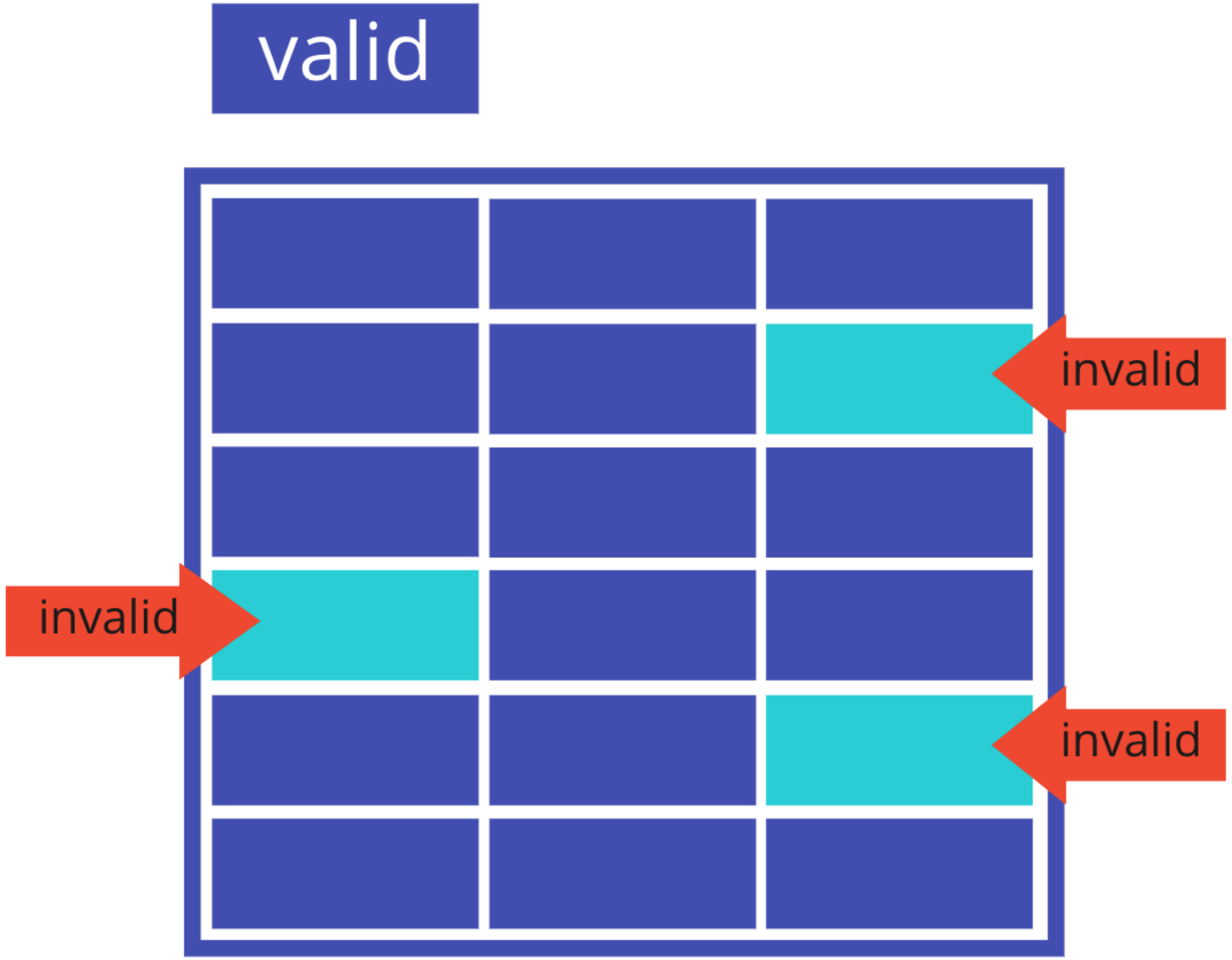
CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

All records must have a value populated in the CustomerName field.

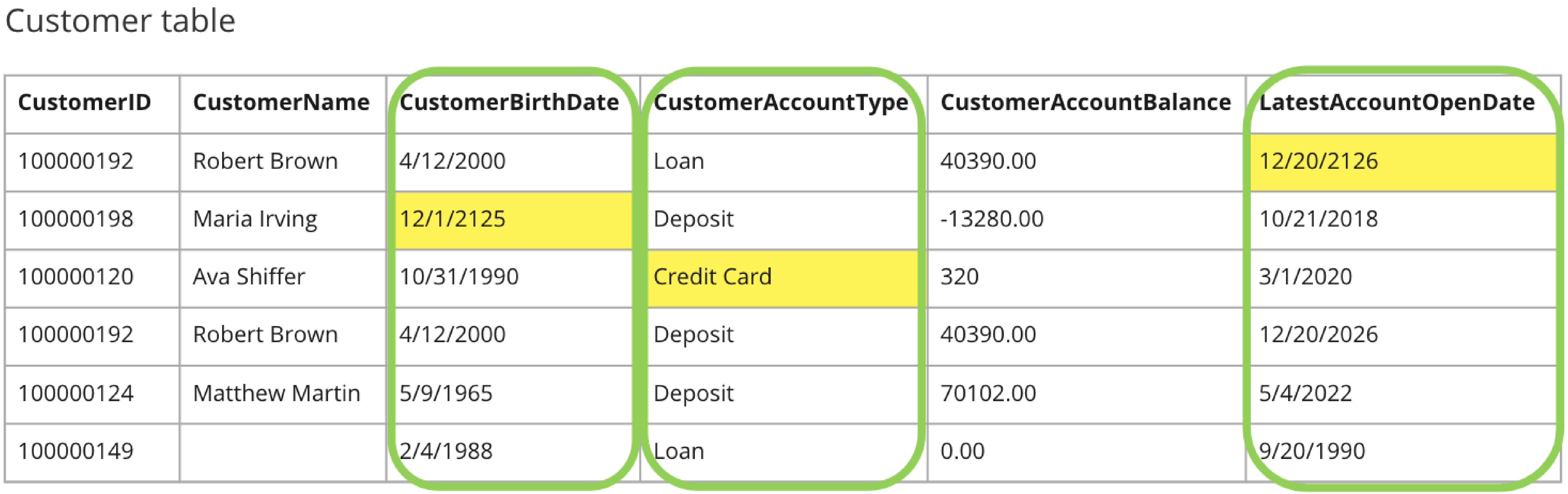
Validity as a data quality dimension

Validity: measures the degree to which the values in a data element are valid

- Requires business context
- Define list or criteria for valid values
- Numeric measurement of validity = count of valid/total count



Validity example

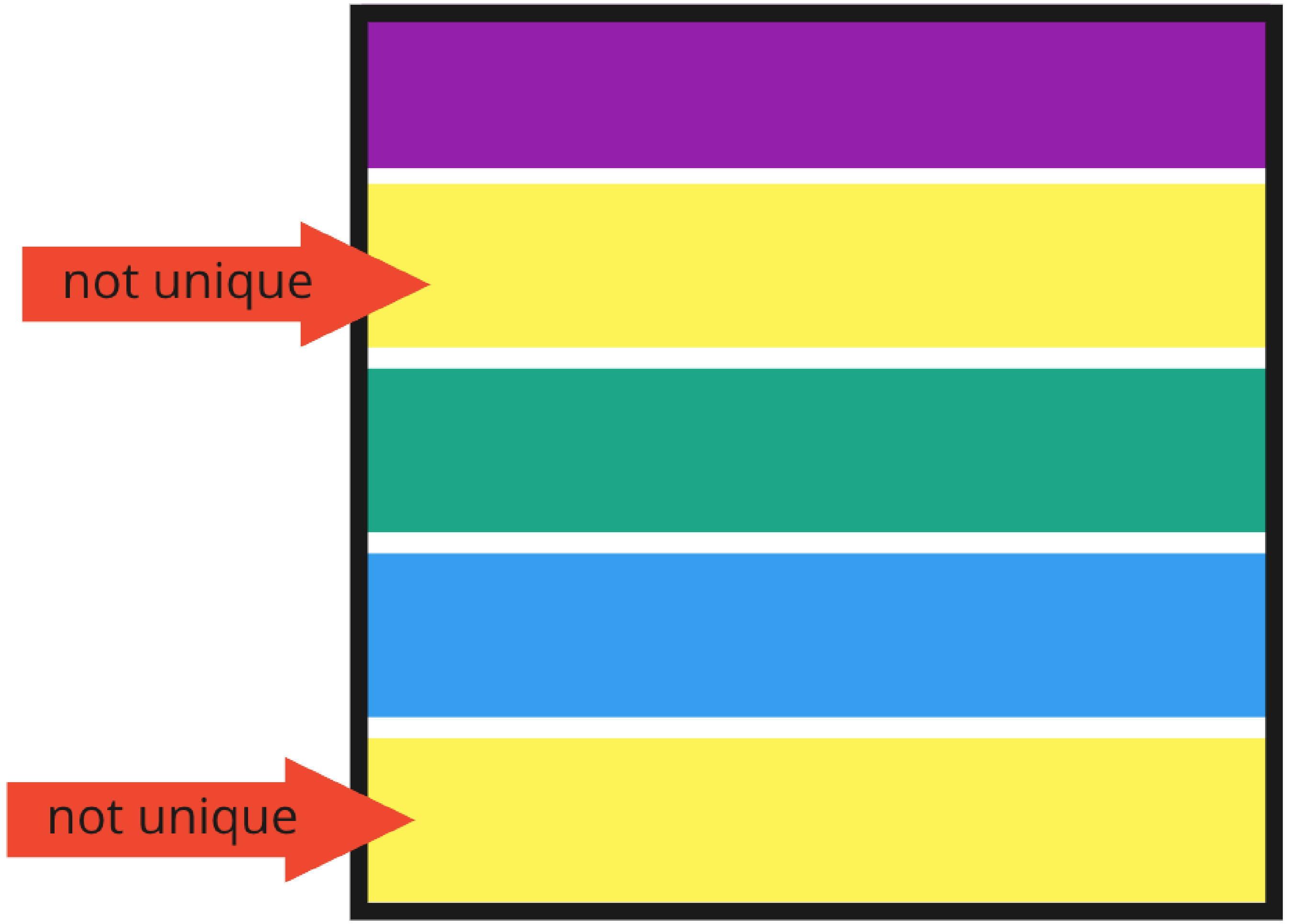


Customer table

CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2126
100000198	Maria Irving	12/1/2125	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Deposit	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

- CustomerBirthDate value must be a date in the past.
- CustomerAccountType value must be either Loan or Deposit.
- LatestAccountOpenDate value must be a date in the past.

Uniqueness as a data quality dimension

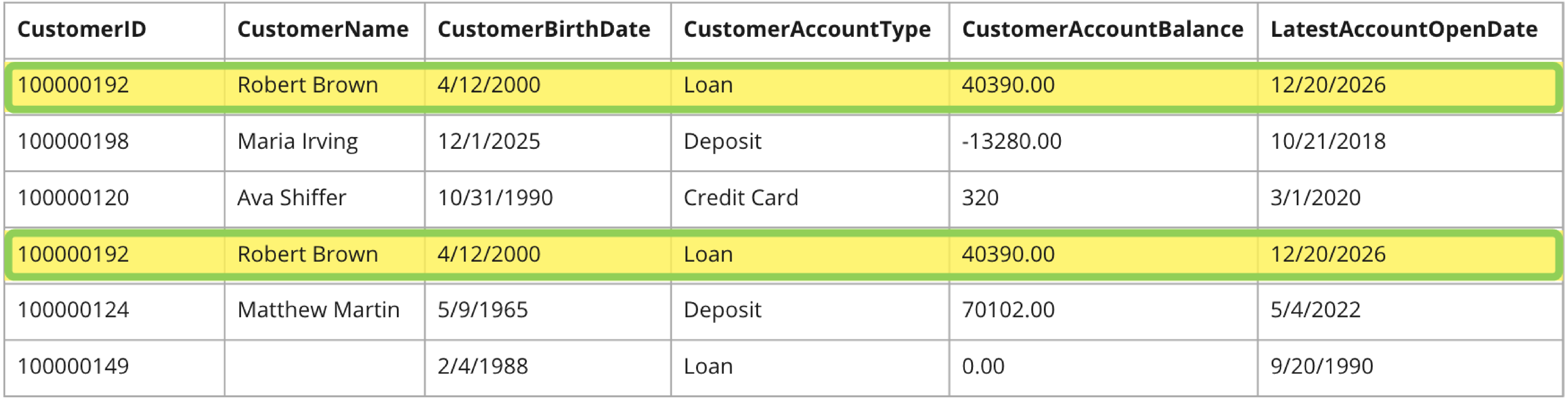


Uniqueness: measures the degree to which the records in a dataset are not duplicated

- Requires business context to define criteria for determining unique records
- May need to look for duplicates in one or multiple columns to identify errors

Uniqueness example

Customer table



CustomerID	CustomerName	CustomerBirthDate	CustomerAccountType	CustomerAccountBalance	LatestAccountOpenDate
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000198	Maria Irving	12/1/2025	Deposit	-13280.00	10/21/2018
100000120	Ava Shiffer	10/31/1990	Credit Card	320	3/1/2020
100000192	Robert Brown	4/12/2000	Loan	40390.00	12/20/2026
100000124	Matthew Martin	5/9/1965	Deposit	70102.00	5/4/2022
100000149		2/4/1988	Loan	0.00	9/20/1990

All records must have a unique CustomerID and CustomerName.

Let's practice!

INTRODUCTION TO DATA QUALITY

Bonus data quality dimensions

INTRODUCTION TO DATA QUALITY

Chrissy Bloom

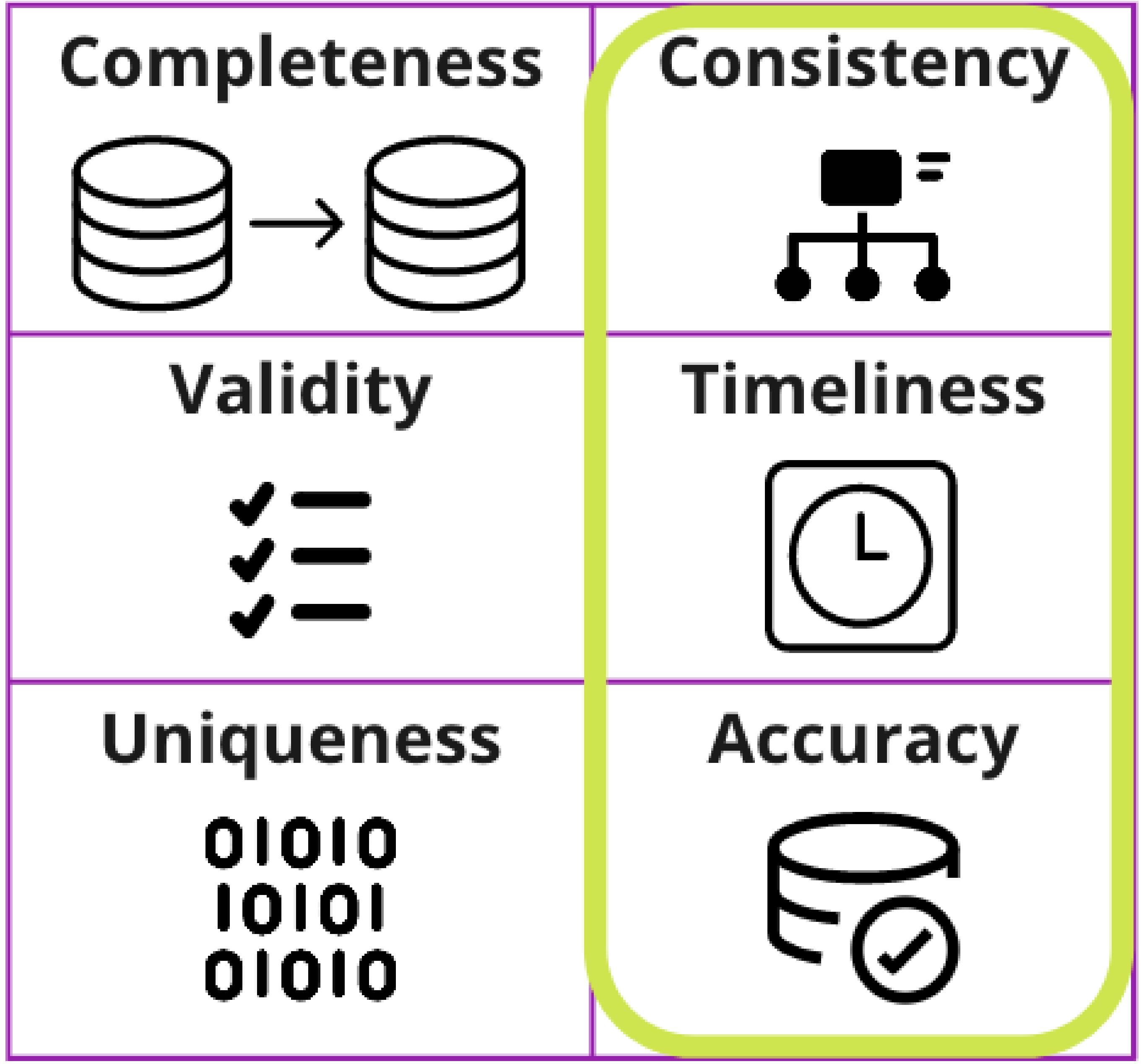
Head of Enterprise Data Strategy & Governance

datacamp

Recalling what a dimension is

Data Quality Dimension: a measurement of a specific attribute of a data's quality

- Use data quality dimensions to quantify how fit for purpose data is.



Completeness

Consistency

Validity

Timeliness

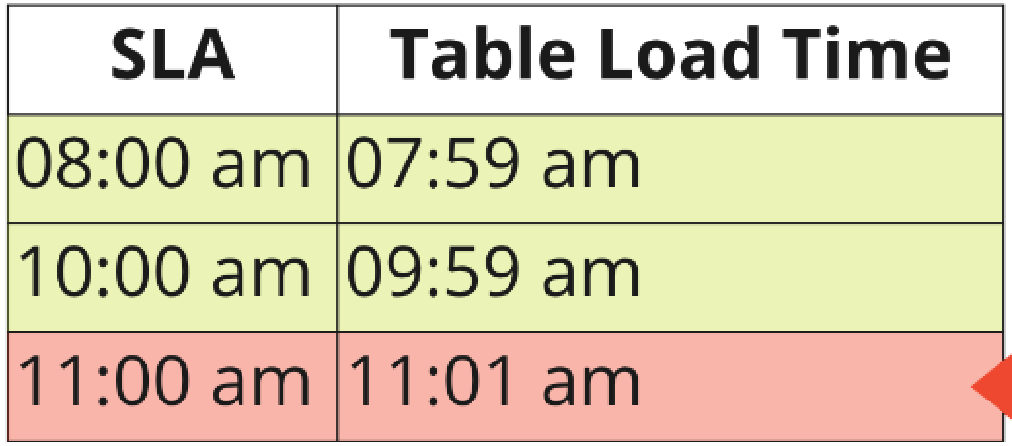
Uniqueness

Accuracy

Timeliness as a dimension

Timeliness: measures the degree to which a dataset is available when expected

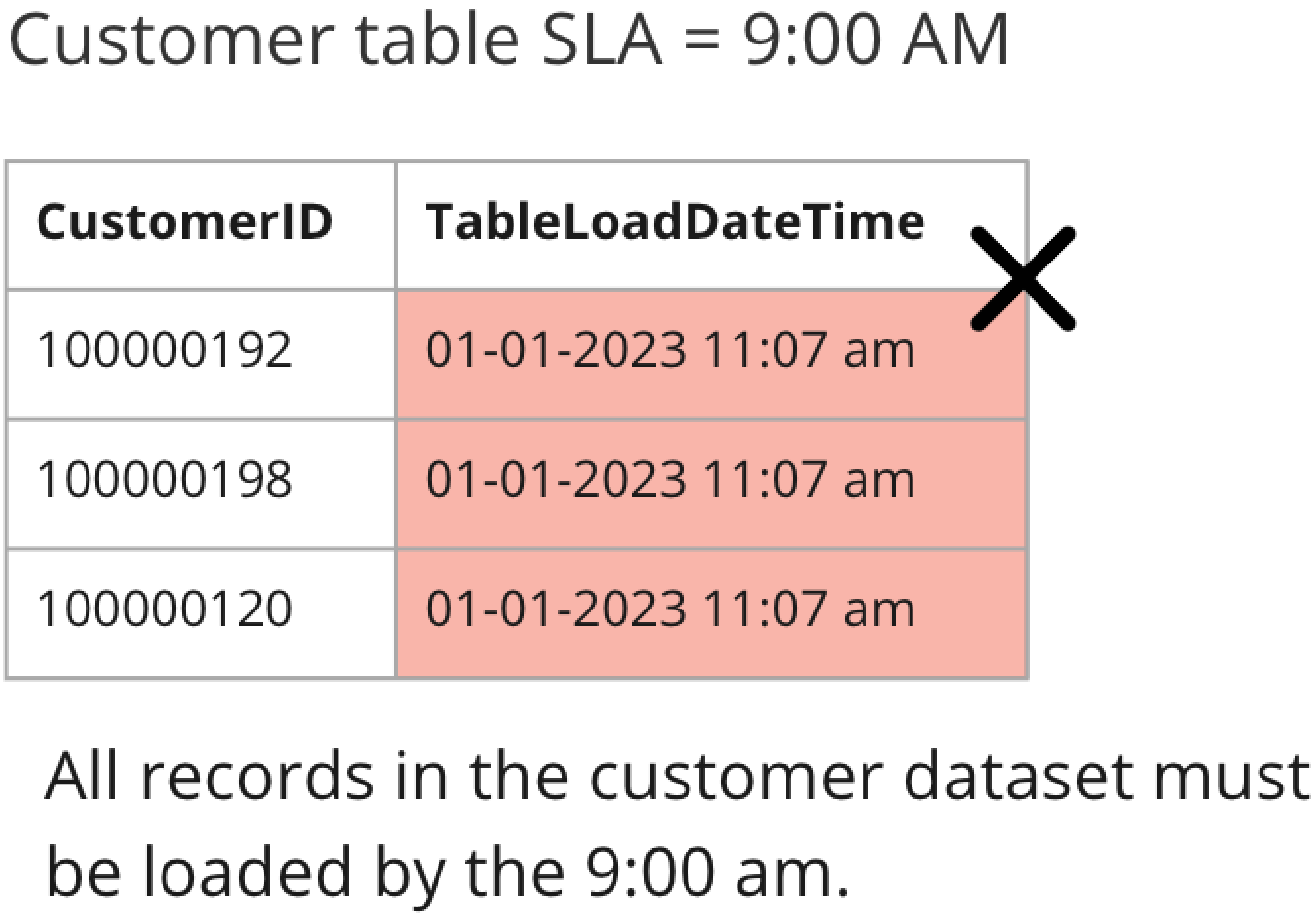
- depends on service level agreements set up between technical and business resources



SLA	Table Load Time
08:00 am	07:59 am
10:00 am	09:59 am
11:00 am	11:01 am

Missed the SLA

Timeliness example



Customer table SLA = 9:00 AM

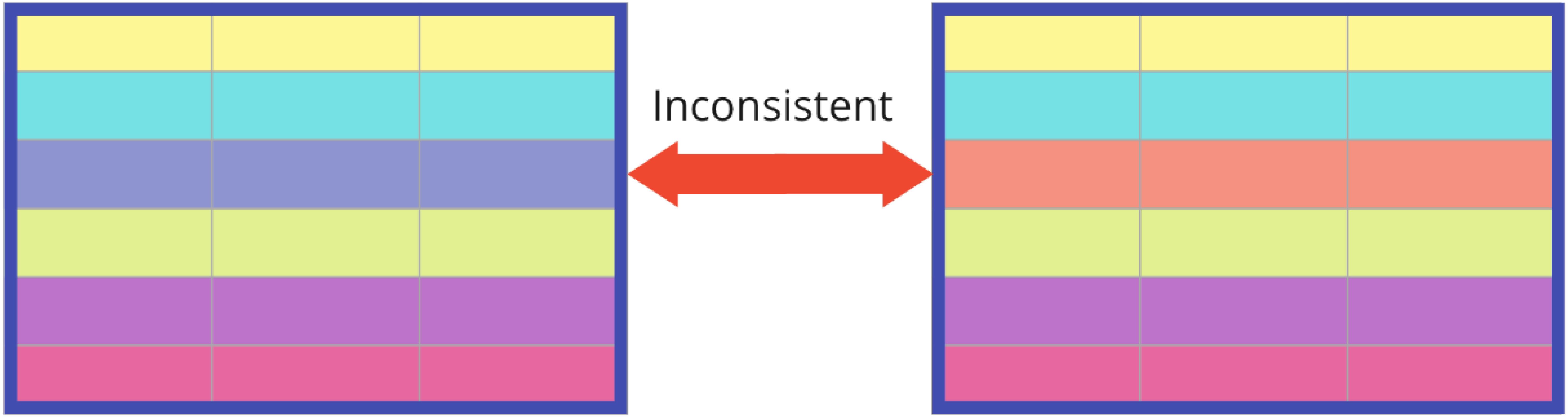
CustomerID	TableLoadDateTime
100000192	01-01-2023 11:07 am
100000198	01-01-2023 11:07 am
100000120	01-01-2023 11:07 am

✗

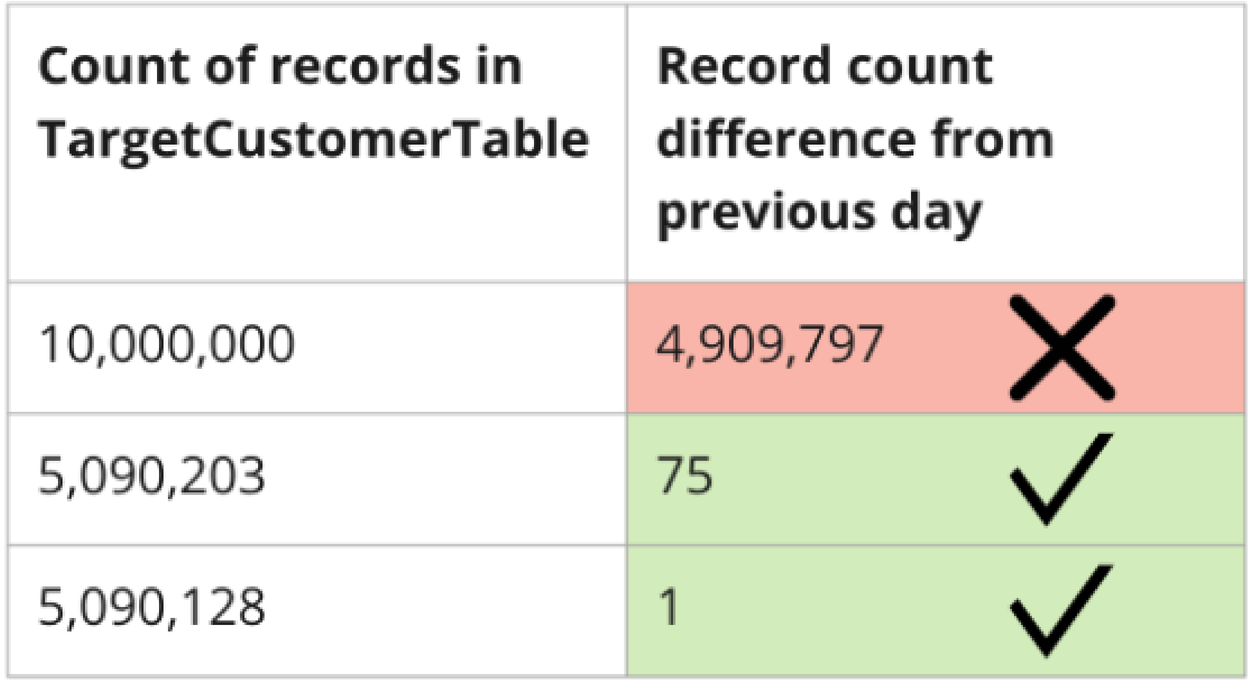
All records in the customer dataset must be loaded by the 9:00 am.

Consistency as a dimension

Consistency: measures the degree to which data is the same across all instances of the data.

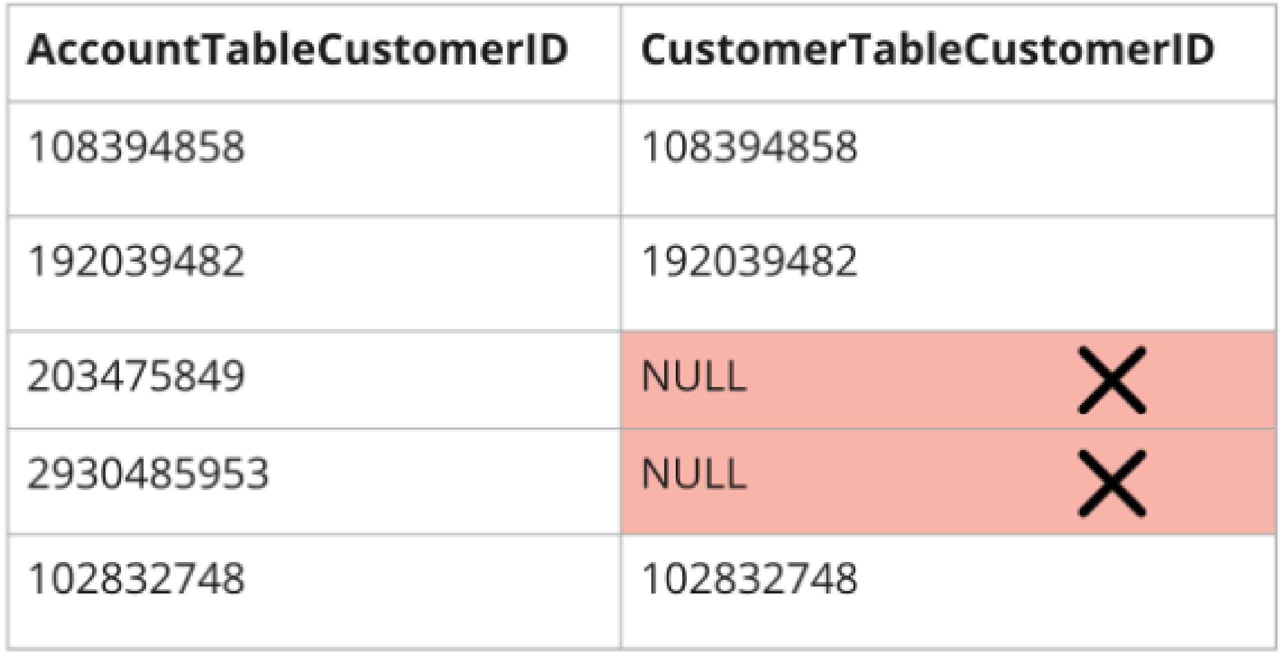


Consistency examples



Count of records in TargetCustomerTable	Record count difference from previous day
10,000,000	4,909,797 ✗
5,090,203	75 ✓
5,090,128	1 ✓

The count of records loaded today must be within +/- 5% of the count of records loaded yesterday.



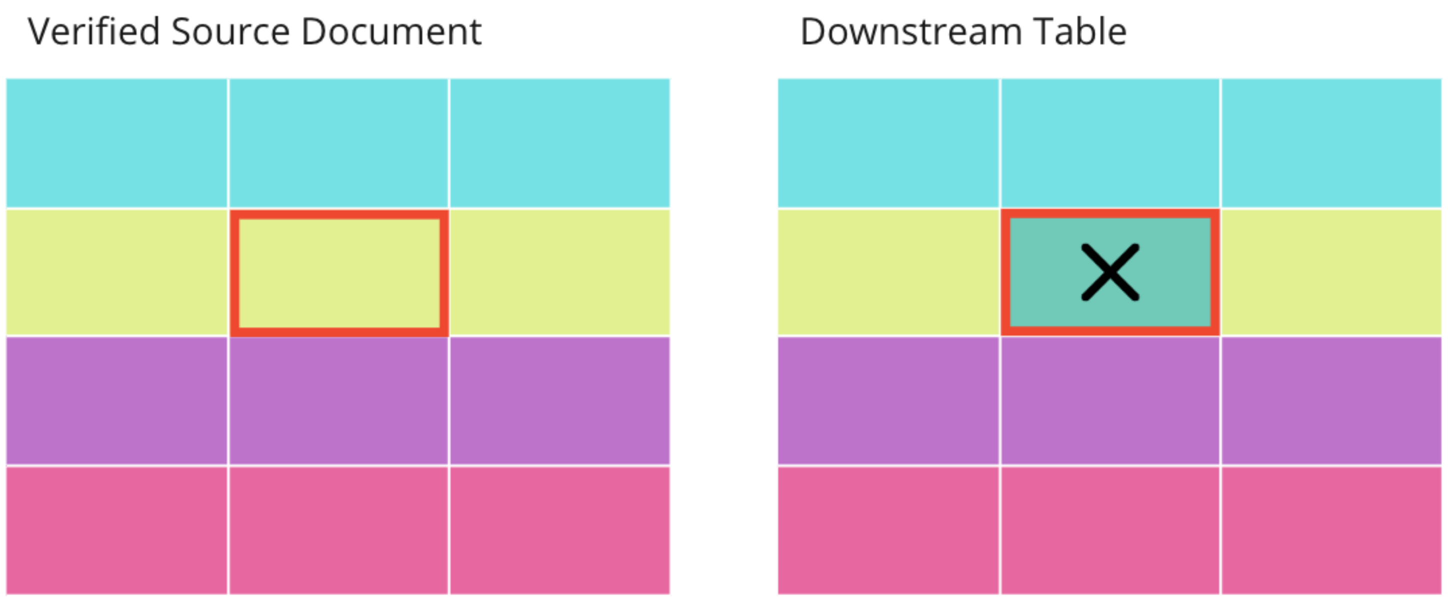
AccountTableCustomerID	CustomerTableCustomerID
108394858	108394858
192039482	192039482
203475849	NULL ✗
2930485953	NULL ✗
102832748	102832748

All Customer ID values in the AccountTable must also be present in the CustomerTable.

Accuracy as a dimension

Accuracy: measures the degree to which data is correct and represents the truth.

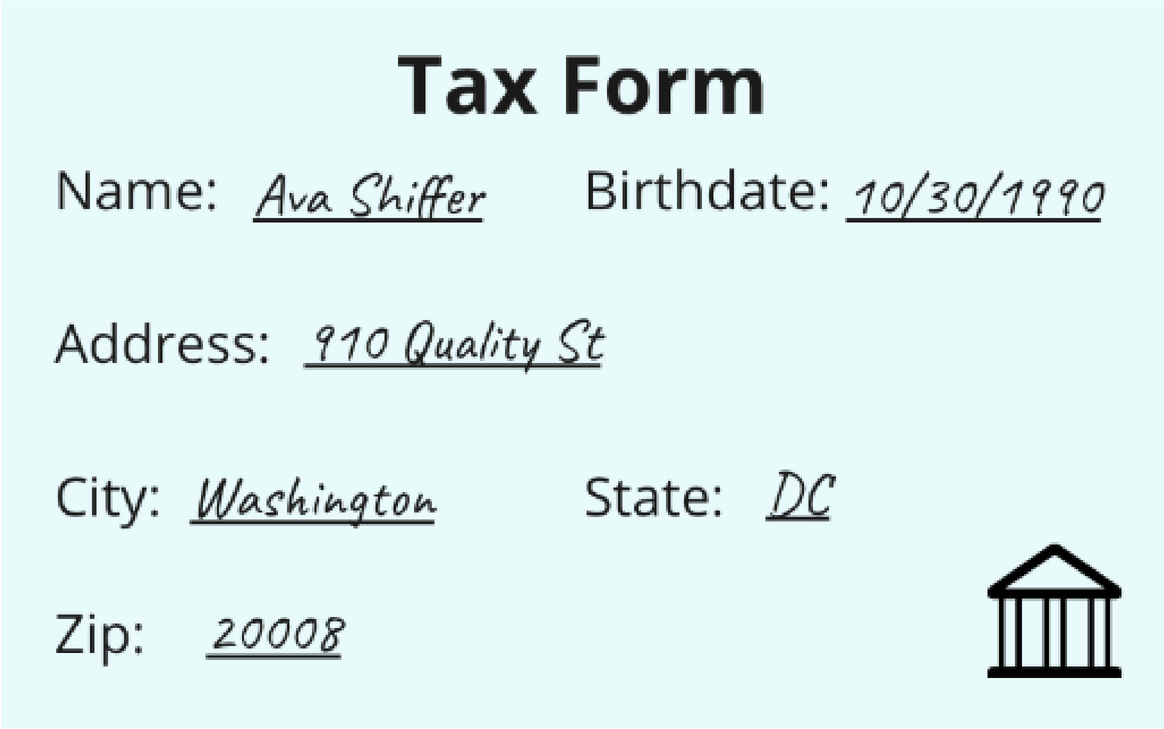
- challenging to measure because it relies on the source of truth being available and accurate



Verified Source Document

Downstream Table

Accuracy example



Tax Form

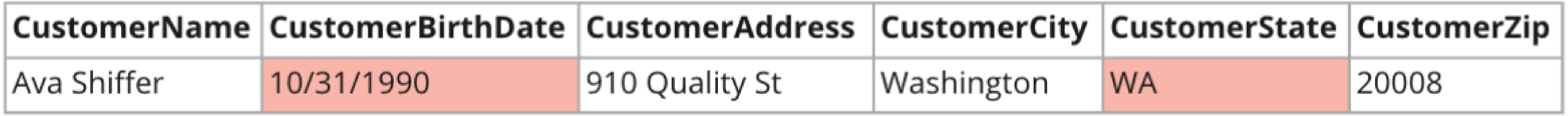
Name: Ava Shiffer Birthdate: 10/30/1990

Address: 910 Quality St

City: Washington State: DC

Zip: 20008

All records in the Customer Table must have accurate Customer Name, Customer Birthdate, and Customer Address fields when compared to the Tax Form.



CustomerName	CustomerBirthDate	CustomerAddress	CustomerCity	CustomerState	CustomerZip
Ava Shiffer	10/31/1990	910 Quality St	Washington	WA	20008

Let's practice!

INTRODUCTION TO DATA QUALITY

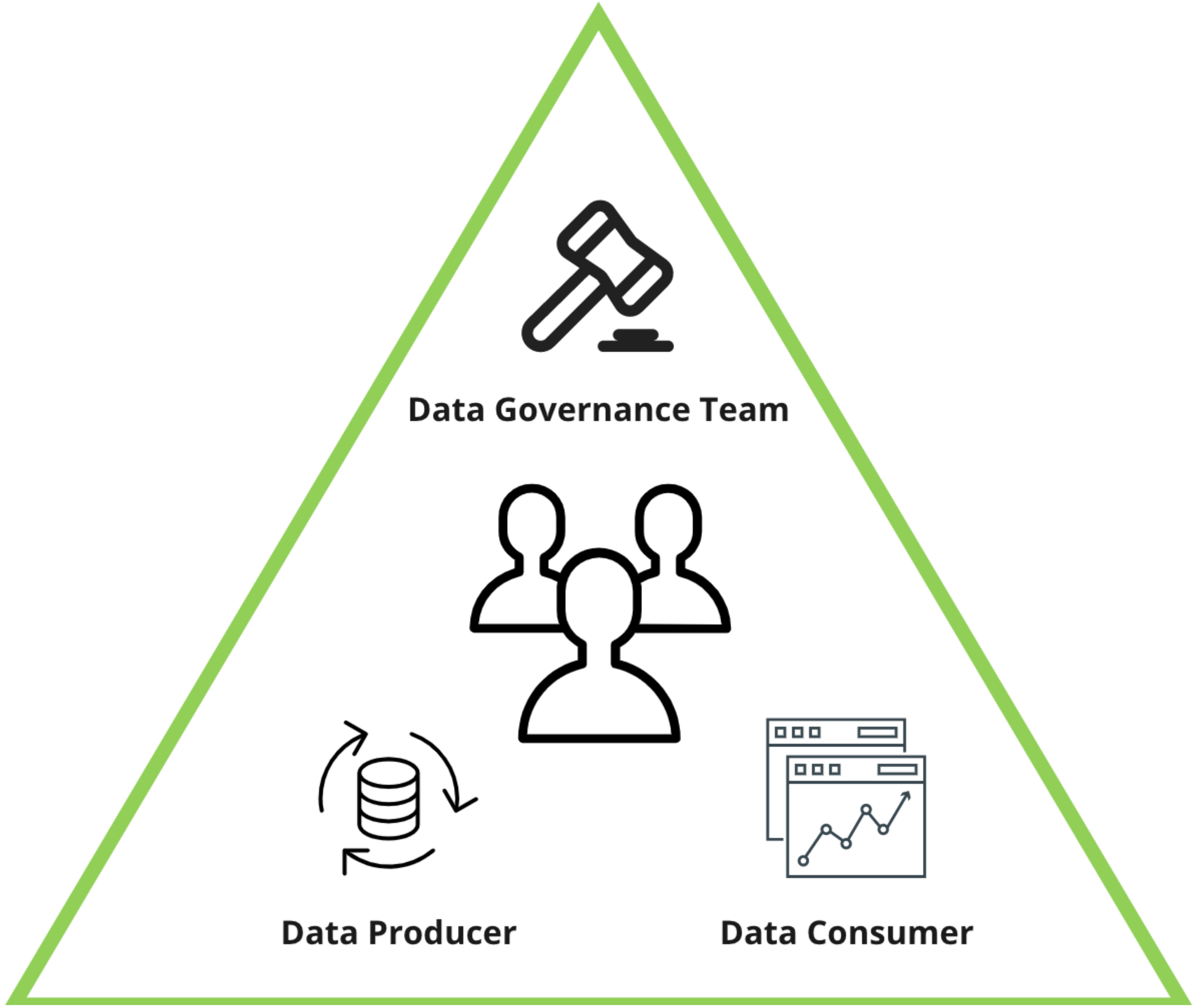
Data quality roles and responsibilities

INTRODUCTION TO DATA QUALITY

Chrissy Bloom

Head of Enterprise Data Strategy & Governance

Roles and responsibilities involved in data quality activities



Roles

- Individuals who serve a specific data quality function
- Often not a full time job
- Identified by assessing everyone who interacts with data

Responsibilities

- Functions and activities related to data quality that each role is responsible for

Data Producers

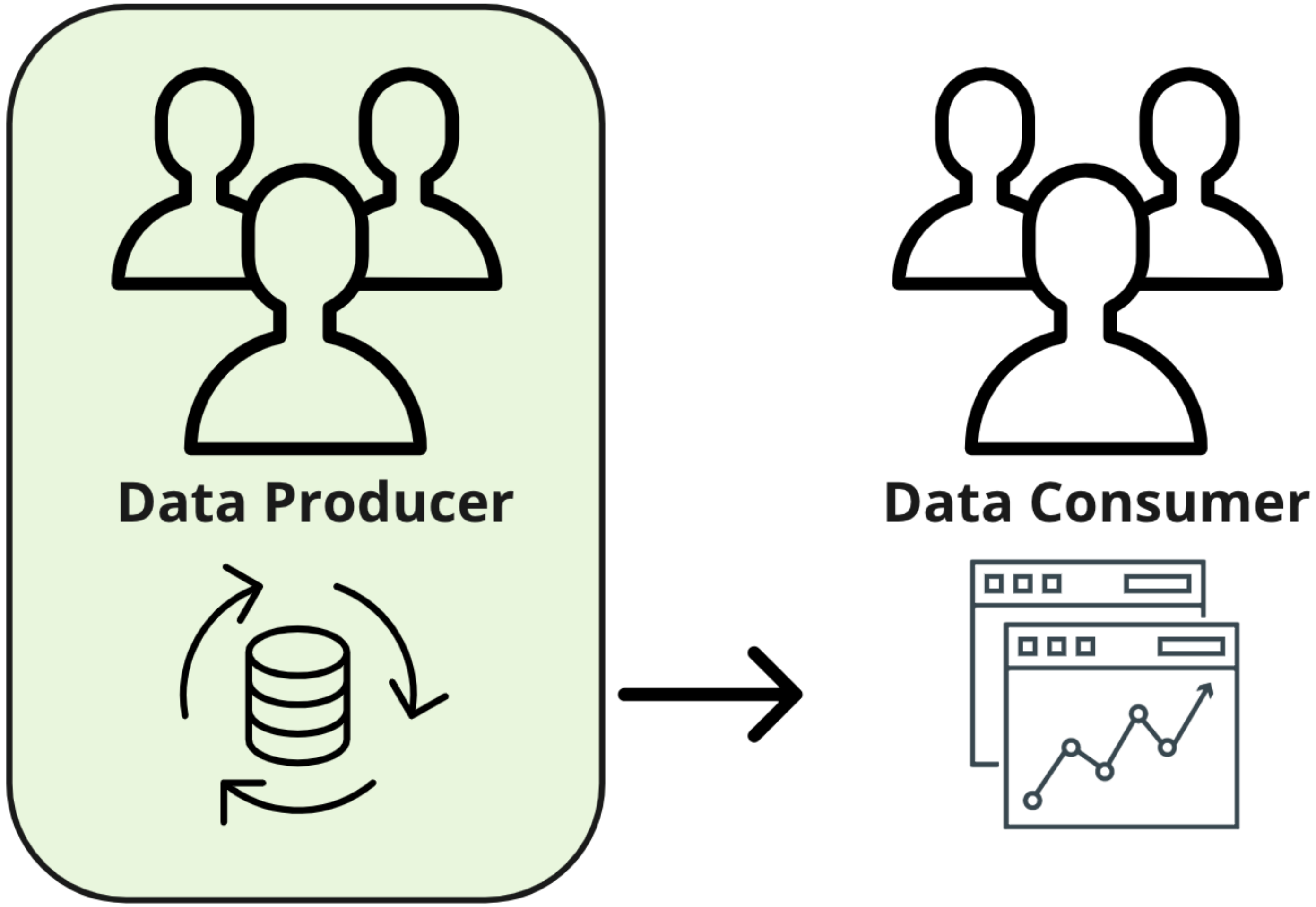
Data Quality Role: Data Producer

Description: Individuals who create, collect, process, transform, or store data

Data Quality Responsibilities:

- Implements data quality rules
- Ensures remediation of data quality issues
- Responsible for technical data quality rules

Examples: System and database owners, ETL developers, report writers and data scientists who create data



Data Consumers

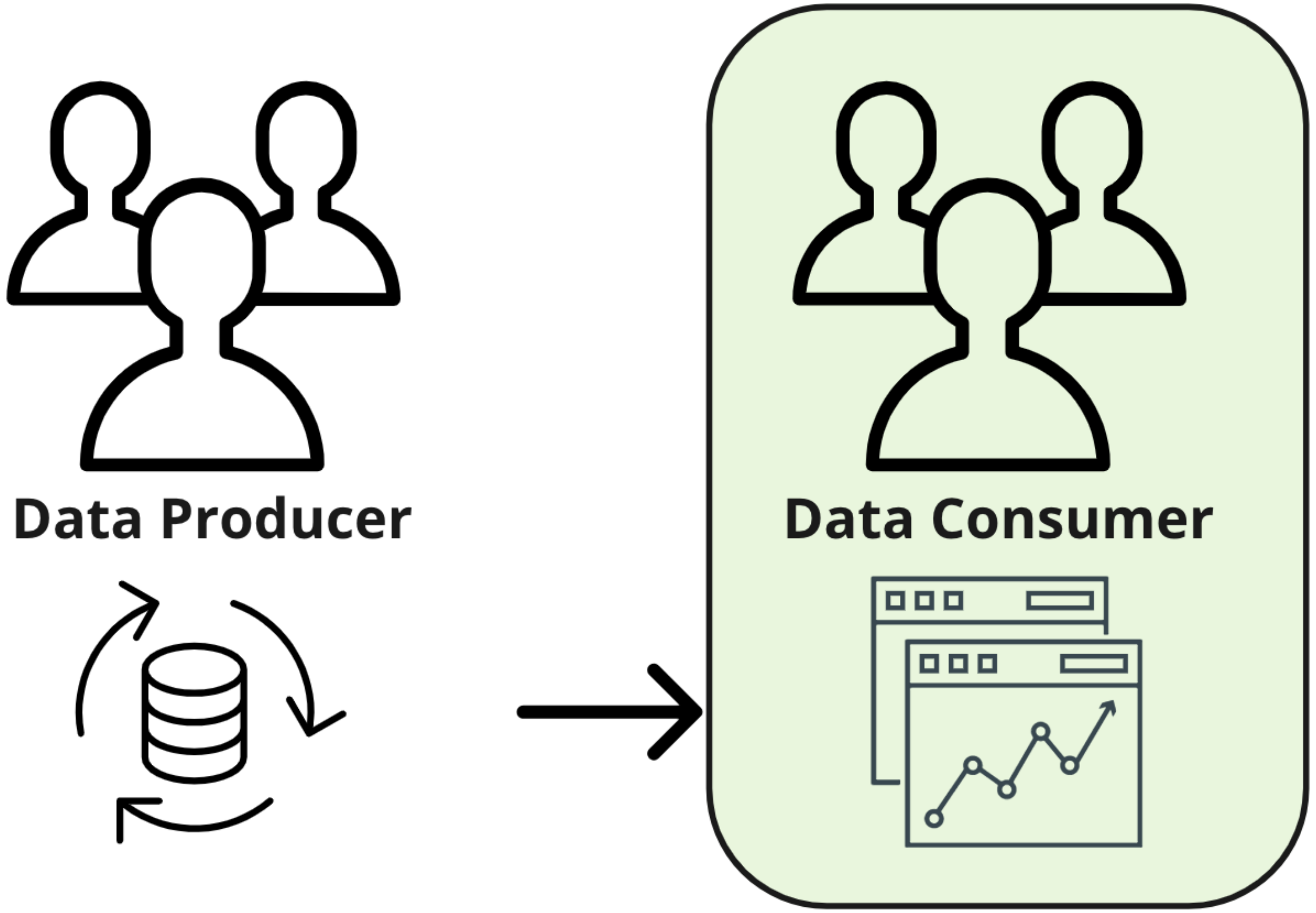
Data Quality Role: Data Consumer

Description: Individuals or applications who use the data produced by data producers

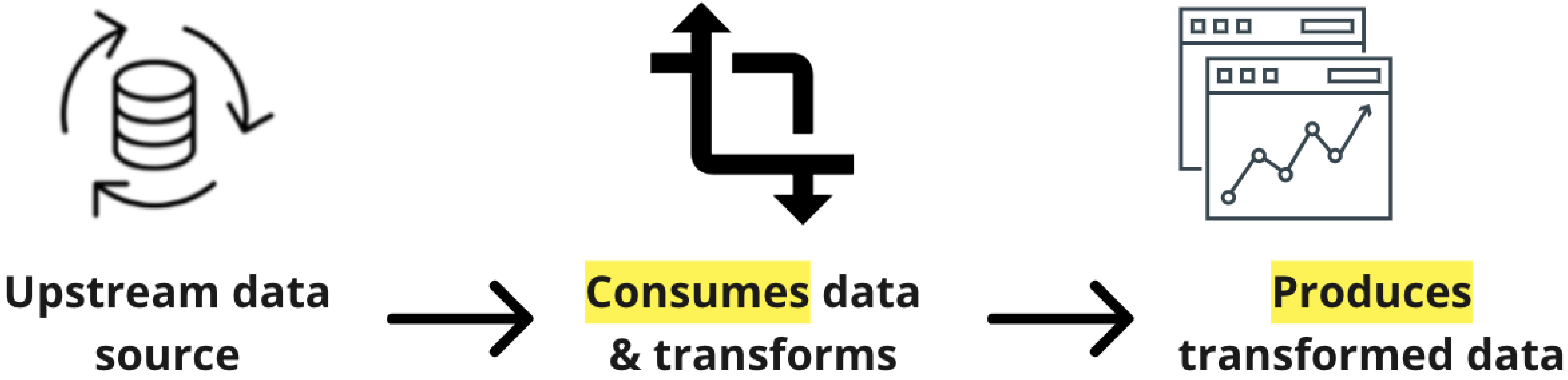
Data Quality Responsibilities:

- Advises on data quality rules to implement
- Accountable for understanding quality of data before using it
- Reports data quality issues

Examples: Report writers, data scientists, ETL developers, business managers, executives

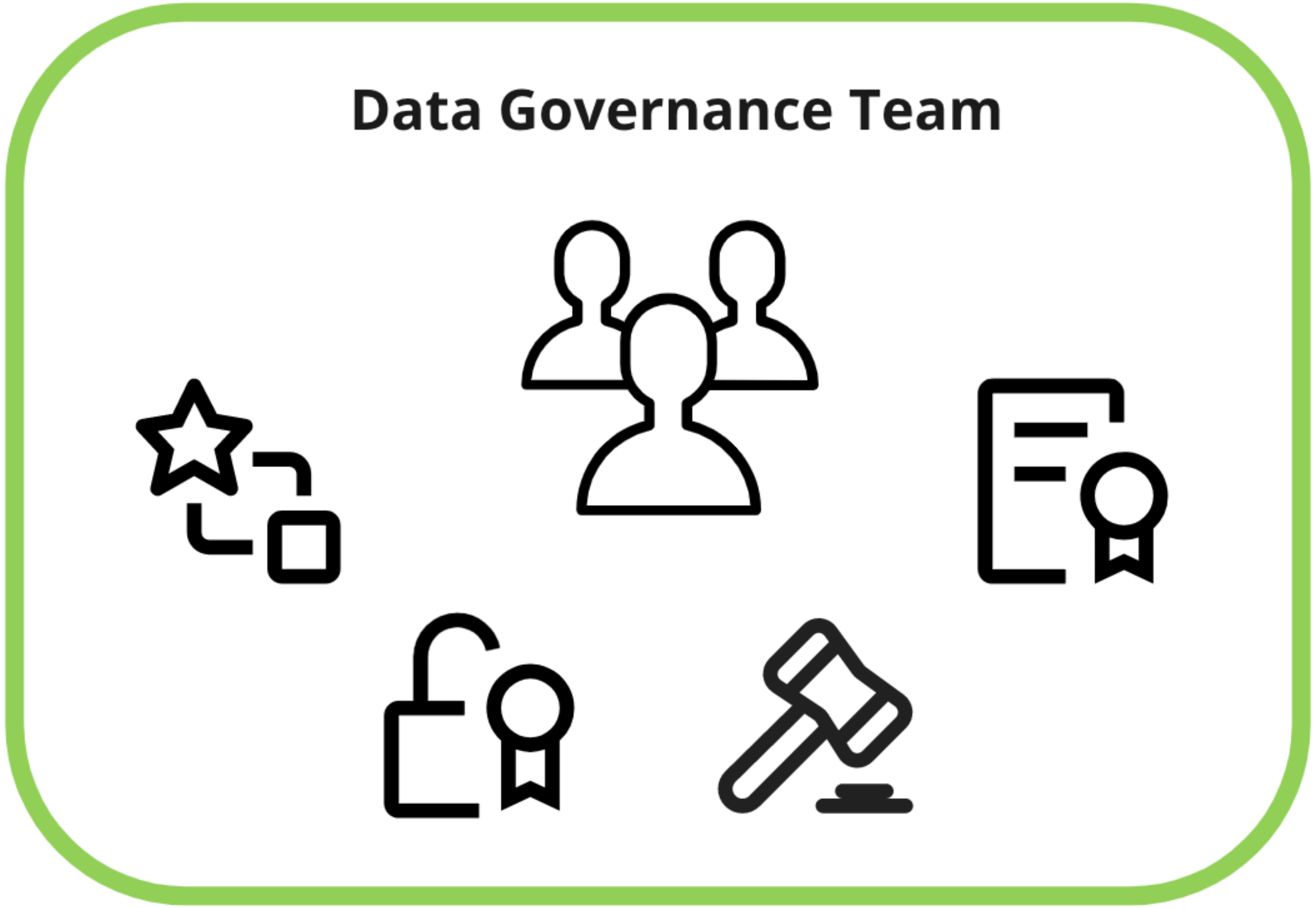


Both a producer and consumer



Upstream data source → **Consumes** data & transforms → **Produces** transformed data

Data Governance Team



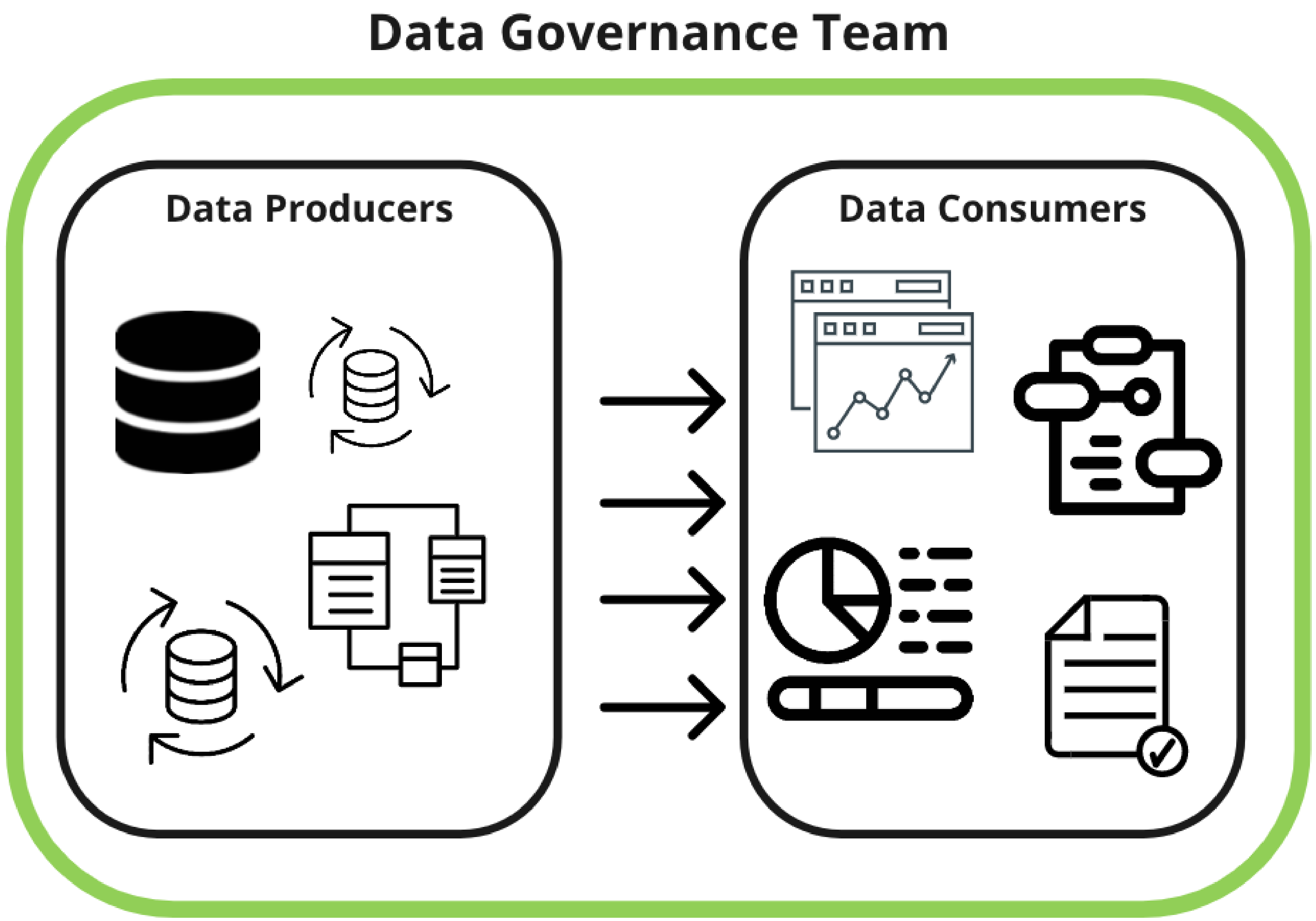
Data Quality Role: Data Governance Team

Description: Team responsible for overall data quality oversight and governance

Data Quality Responsibilities:

- Define and enforce data quality policies
- Define data quality roles and responsibilities
- Monitor data quality dashboards
- Ensure appropriate data quality tools, processes, and training are available

Data Quality Roles and Responsibilities Applied



Scenario: Report writer identifies a data quality issue or determines a data quality rule is needed

- **Data Consumer (Report writer):** Alerts data producer and recommends a data quality rule for implementation. Decides whether to use the data.
- **Data Producer (Source System Owner):** Implements data quality rule and remediates the issue
- **Data Governance Team:** Oversees the process. Makes a data quality dashboard

Let's practice!

INTRODUCTION TO DATA QUALITY