Cleaning Data

Mihaela Danci
Data Analyst

linkedin.com/in/mihaela-danci/

Ensuring high-quality data is an essential prerequisite for accurate data analysis.

Data Quality Characteristics

- Completeness
- Consistency
- Accuracy
- Uniqueness
- Standardization
- Relevance

Data Quality Characteristics

Completeness

Consistency

Accuracy

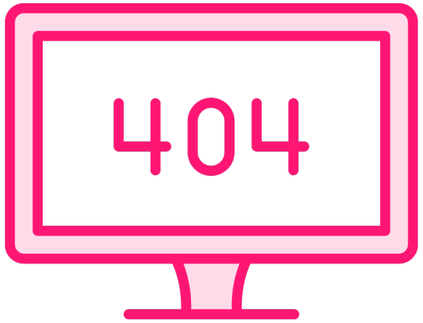
Uniqueness

Standardization

Relevance

Causes of Poor Data Quality

Human error



System error

Poor Quality Dataset

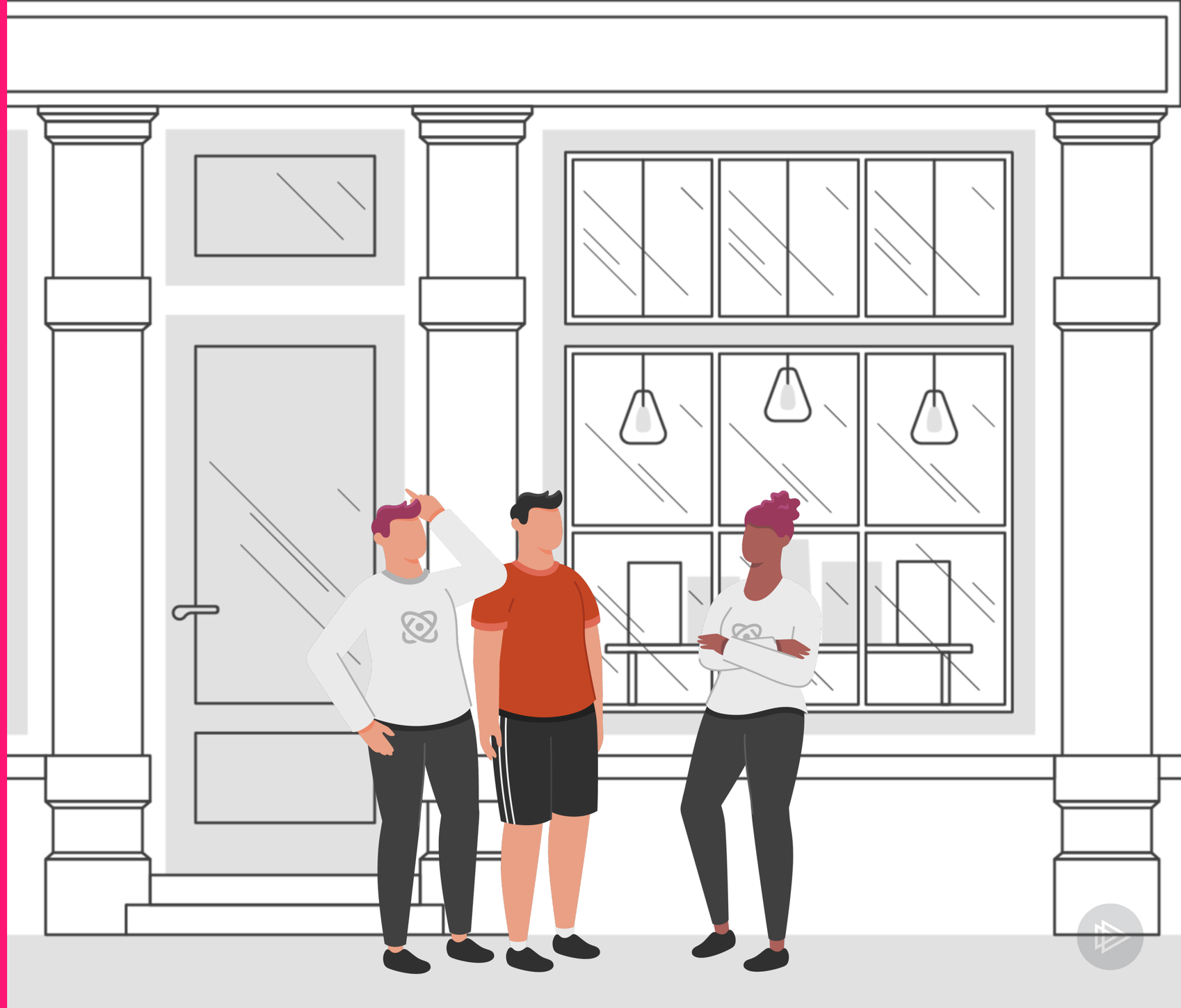
Duplicates

Inconsistent data

Incorrect data

Missing data

Shape the data's appearance in the output



Build-in functions

Refinement of displayed data

Keeping the underlying data unchanged

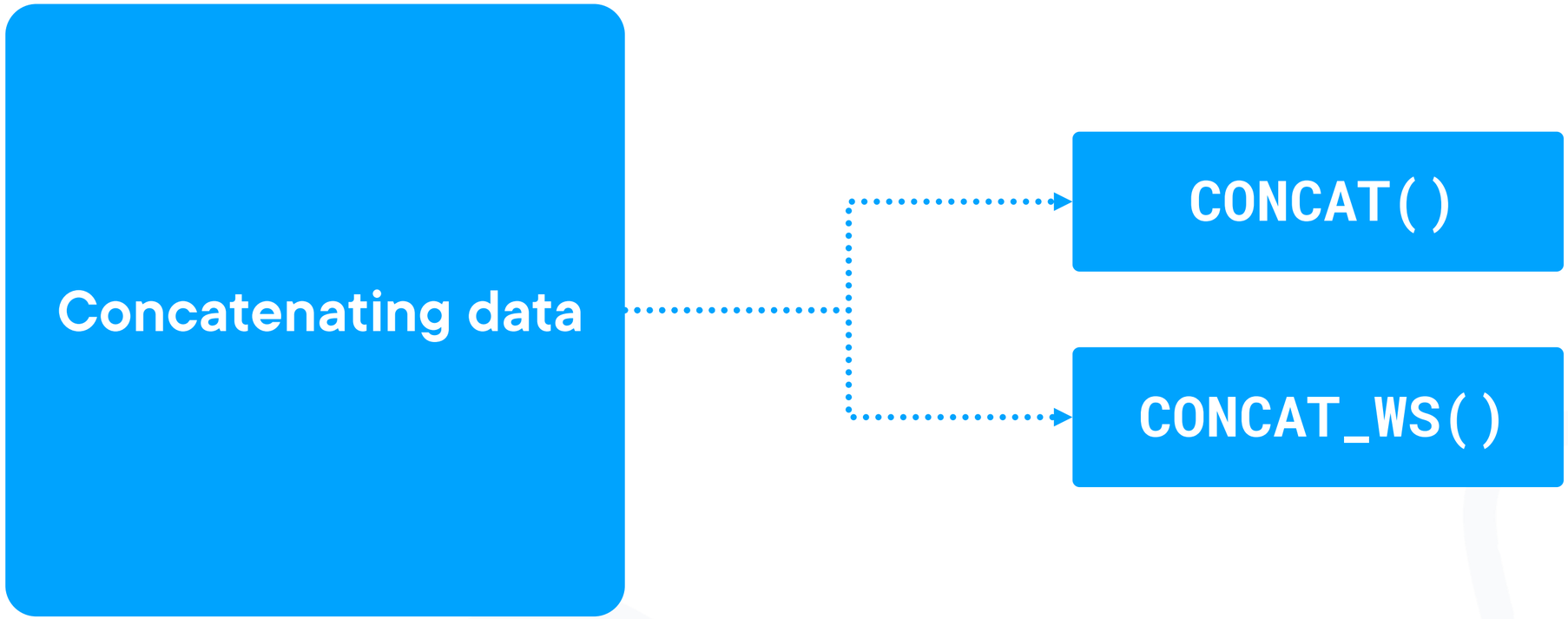
```
-- Removing white spaces
SELECT TRIM( '      abc      ') AS result;

-- Removing specific characters
SELECT TRIM( '&' FROM '&abc&') AS result;

SELECT LTRIM( '&abc' , '&') AS result;

SELECT RTRIM('abc&', '&') AS result;
```

```
-- Output
abc

abc

abc

abc
```

Text Manipulation Functions

Product	UPPER(Product)	LOWER(Product)
Bike	BIKE	bike
Jacket	JACKET	jacket
Helmet	HELMET	helmet

Text Manipulation Functions



Text Manipulation Functions

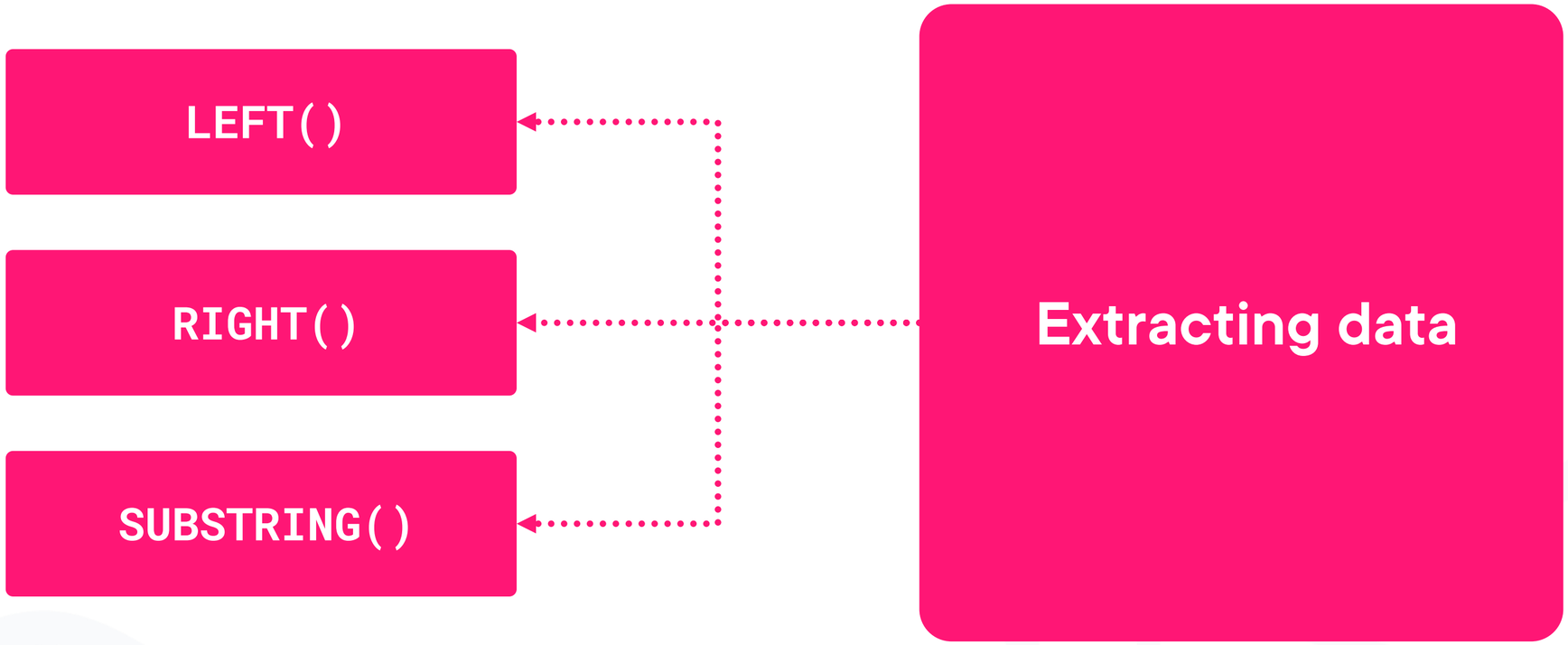
Concatenating data

Replacing data

REPLACE( )

Extracting data

Text Manipulation Functions



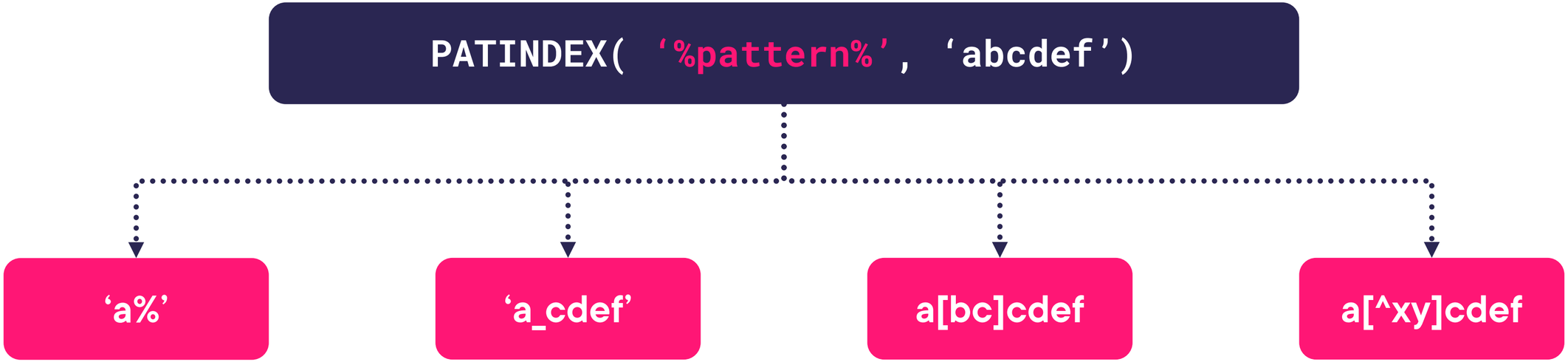
Searching for Patterns

```
PATINDEX( '%pattern%', expression)
```

```
PATINDEX( 'a%', 'abc')

1
```

```
PATINDEX( 'x%', 'abc')

0
```

Searching for Patterns



Overview

Explore data issues and effective solutions

- Manipulating text
- Converting data
- Managing NULLs

Demo

Text manipulation functions

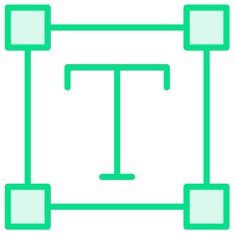
- Removing white spaces
- Removing unwanted characters
- Concatenating fields
- Extracting string parts
- Searching for patterns

Data Types

Data Types

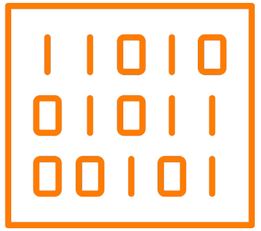
Exact numeric

Approximate numeric

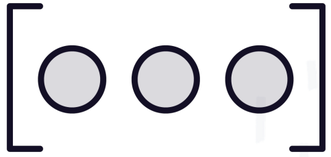


Character

Date/time



Binary



Other

Operators

‘Tom’	+	‘Smith’	=	‘TomSmith’
2	+	2	=	4
‘Tom’	+	2	=	‘Tom2’

Data type precedence

Operators

‘Tom’	+	‘Smith’	=	‘TomSmith’
2	+	2	=	4
‘Tom’	+	’2’	=	‘Tom2’

Data type precedence

Conversion Types

Implicit

Explicit

Explicit Conversions

CAST()

CONVERT()

PARSE()

They'll throw an error if the conversion is not possible.

Explicit Conversions

CAST()

CONVERT()

PARSE()

TRY_CAST()

TRY_CONVERT()

TRY_PARSE()

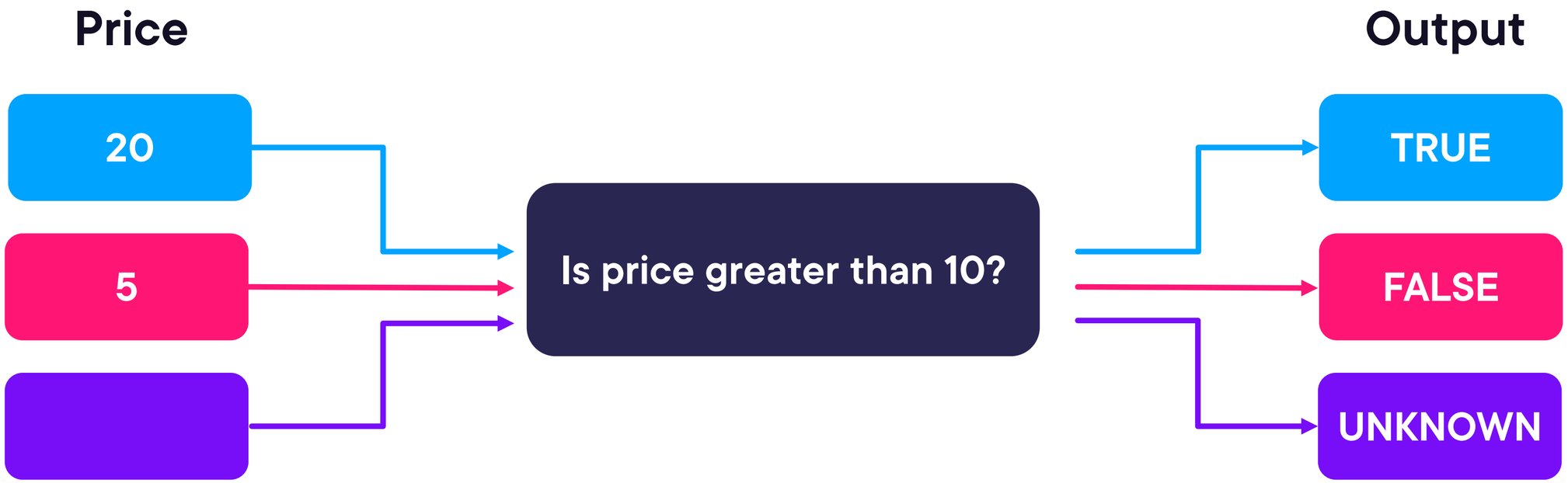
Working with NULLs

NULLs represent missing or unknown values.

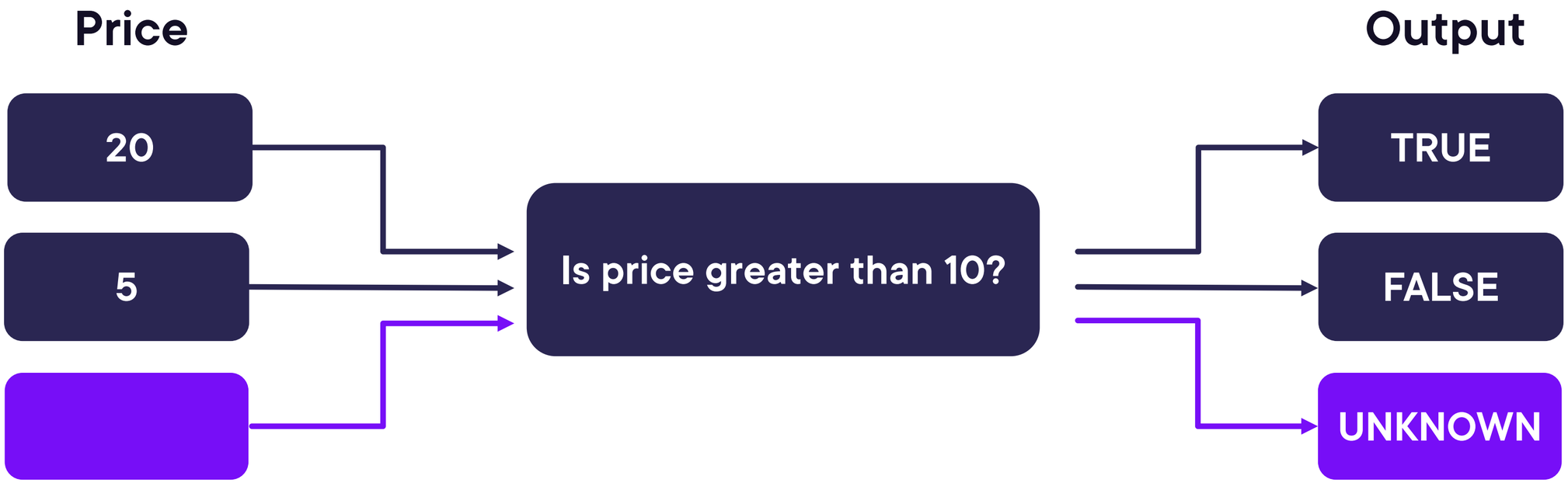
Step 1: Identifying Missing Values

Three-valued predicate logic: TRUE, FALSE, UNKNOWN

Identifying Missing Values



Identifying Missing Values



Identifying Missing Values

Identifying Missing Values

IS NULL

IS NOT NULL

Step 2 : Handling Missing Values

Handling Missing Values

Find the missing values

Exclude the missing values from analysis

Replace the missing values

Handling Missing Values

```
ISNULL( expression, replacement_value)
```

```
COALESCE( expression_1, expression_2,… expression_n)
```

Handling Missing Values

```
COALESCE( expression_1, expression_2,… expression_n)
```

```
CASE
WHEN (expression_1 IS NOT NULL) THEN expression_1
WHEN (expression_2 IS NOT NULL) THEN expression_2
…
ELSE expression_n
END
```

Replacing Missing Values

```
NULLIF( expression, expression)
```

Demo

Handling NULL values

- ISNULL()
- COALESCE()
- NULLIF()

Summary

Data cleaning

Text manipulation functions

- TRIM()
- CONCAT()
- UPPER()

Change data types

- CAST()
- CONVERT()

Handling NULLs

- COALESCE()

Up Next:

Using Subqueries and CTEs