Data cleaning

INTRODUCTION TO KNIME

Emilio Silvestri
Data Scientist, KNIME

Data cleaning

Filter columns

#	Name	Department	Salary	Favorite food
1	Jenny	Sales	?	Pizza
2	Alex	Finance	22000	Sushi
3	Taylor	R&D	28000	Tacos
4	Alex	Finance	22000	Sushi
5	Sam	R&D	23000	Salad

Filter columns

Filter out

#	Name	Department	Salary	Favorite food
1	Jenny	Sales	?	Pizza
2	Alex	Finance	22000	Sushi
3	Taylor	R&D	28000	Tacos
4	Alex	Finance	22000	Sushi
5	Sam	R&D	23000	Salad

Filter rows

Employees from R&D

#	Name	Department	Salary	Favorite food
1	Jenny	Sales	?	Pizza
	Alex	Finance	22000	Sushi
3	Taylor	R&D	28000	Tacos
4	Alex	Finance	22000	Sushi
5	Sam	R&D	23000	Salad

Filter duplicates

#	Name	Department	Salary	Favorite foo
1	Jenny	Sales	?	Pizza
2	**Alex**	**Finance**	**22000**	**Sushi**
3	Taylor	R&D	28000	Tacos
4	**Alex**	**Finance**	**22000**	**Sushi**
5	Sam	R&D	23000	Salad

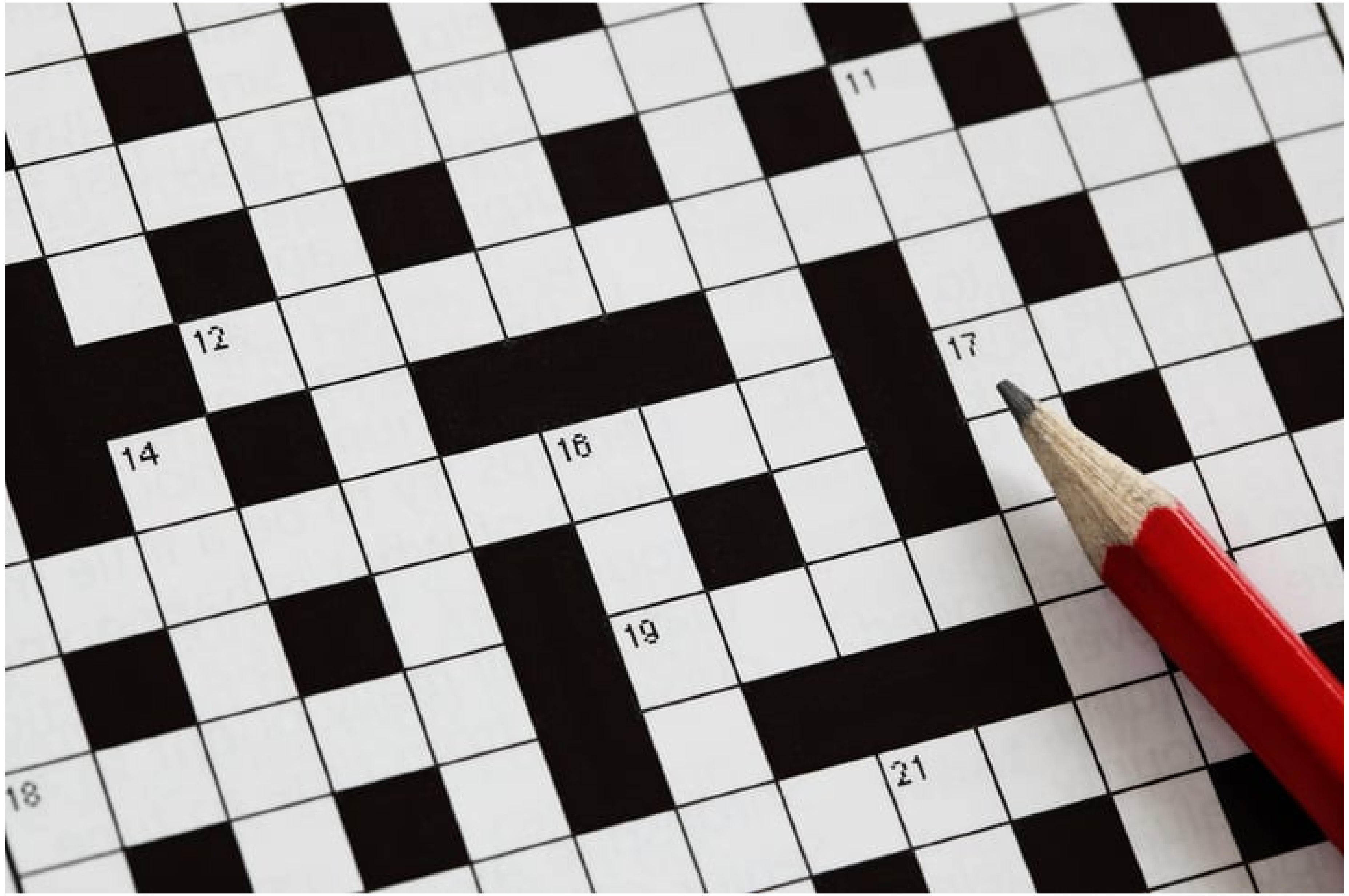
Duplicate row

Handle missing values

#	Name	Department	Salary	Favorite food
1	Jenny	Sales	?	Pizza
2	Alex	Finance	2200	shi
3	Taylor	R&D	28000	Tacos
4	Alex	Finance	22000	Sushi
5	Sam	R&D	23000	Salad

Missing value

Handle missing values



Data types

#	Name	Salary	Birthday	Remote
1	Jenny	25000	12/12/1990	TRUE
2	Alex	22000	04/03/1998	FALSE
3	Taylor	28000	11/02/1987	FALSE
4	Alex	22000	04/03/1998	FALSE
5	Sam	23000	02/07/1989	TRUE

Data types

String

#	Name	Salary	Birthday	Remote
1	Jenny	25000	12/12/1990	TRUE
2	Alex	22000	04/03/1998	FALSE
3	Taylor	28000	11/02/1987	FALSE
4	Alex	22000	04/03/1998	FALSE
5	Sam	23000	02/07/1989	TRUE

Data types

String

Number

#	Name	Salary	Birthday	Remote
1	Jenny	25000	12/12/1990	TRUE
2	Alex	22000	04/03/1998	FALSE
3	Taylor	28000	11/02/1987	FALSE
4	Alex	22000	04/03/1998	FALSE
5	Sam	23000	02/07/1989	TRUE

Data types

	String	Number	Date&Time	
#	**Name**	**Salary**	**Birthday**	**Remote**
1	Jenny	25000	12/12/1990	TRUE
2	Alex	22000	04/03/1998	FALSE
3	Taylor	28000	11/02/1987	FALSE
4	Alex	22000	04/03/1998	FALSE
5	Sam	23000	02/07/1989	TRUE

Data types

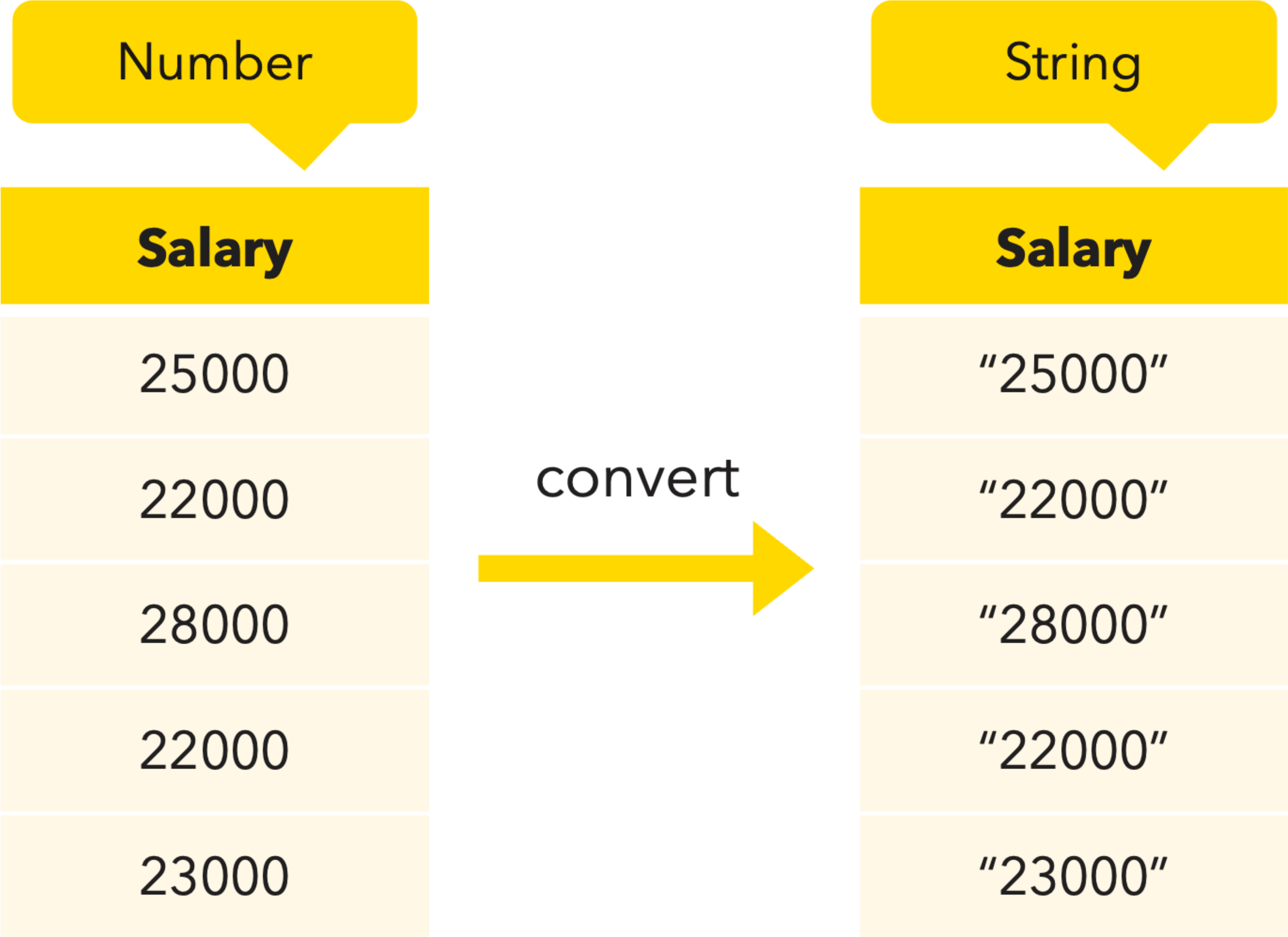
	String	Number	Date&Time	Boolean
#	**Name**	**Salary**	**Birthday**	**Remote**
1	Jenny	25000	12/12/1990	TRUE
2	Alex	22000	04/03/1998	FALSE
3	Taylor	28000	11/02/1987	FALSE
4	Alex	22000	04/03/1998	FALSE
5	Sam	23000	02/07/1989	TRUE

Convert Data Types

Number

Salary
25000
22000
28000
22000
23000

Convert Data Types



Number

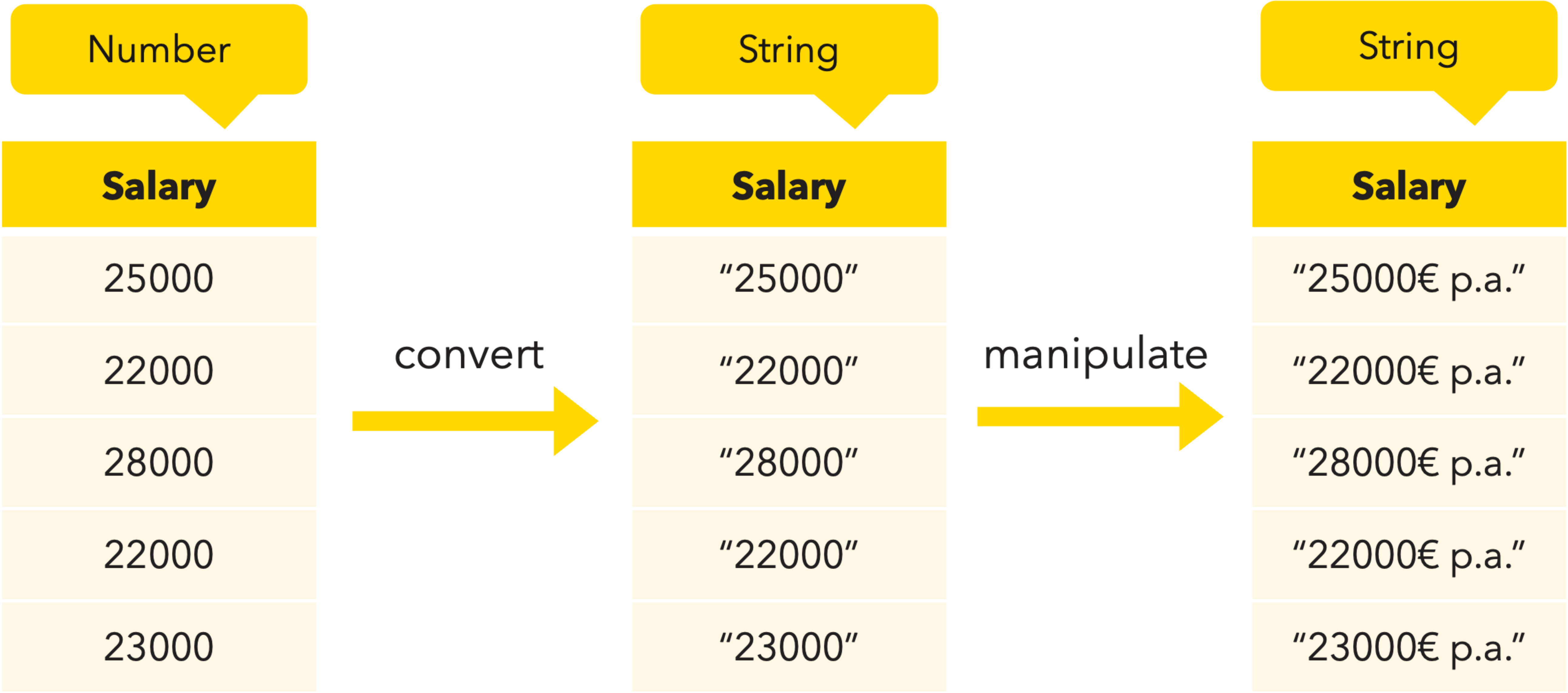
Salary
25000
22000
28000
22000
23000

convert

String

Salary
“25000”
“22000”
“28000”
“22000”
“23000”

Convert Data Types



Number

Salary
25000
22000
28000
22000
23000

convert

String

Salary
“25000”
“22000”
“28000”
“22000”
“23000”

manipulate

String

Salary
“25000€ p.a.”
“22000€ p.a.”
“28000€ p.a.”
“22000€ p.a.”
“23000€ p.a.”

Clean strings

Email
jenny.brówñ @EXAMPLE.COM
alex.léé @EXAMPLE.COM
taylor.tóñs @EXAMPLE.COM
alex.léé @EXAMPLE.COM
sóphia.kim @EXAMPLE.COM

Clean strings

Email
jenny.brówñ @EXAMPLE.COM
alex.léé @EXAMPLE.COM
taylor.tóñs @EXAMPLE.COM
alex.léé @EXAMPLE.COM
sóphia.kim @EXAMPLE.COM

clean

Email
jenny.brown@example.com
alex.lee@example.com
taylor.tons@example.com
alex.lee@example.com
sophia.kim@example.com

Let's practice!

INTRODUCTION TO KNIME

Cleaning data in KNIME Analytics Platform

INTRODUCTION TO KNIME

Full Name

Instructor

Let's practice!

INTRODUCTION TO KNIME