Data Quality

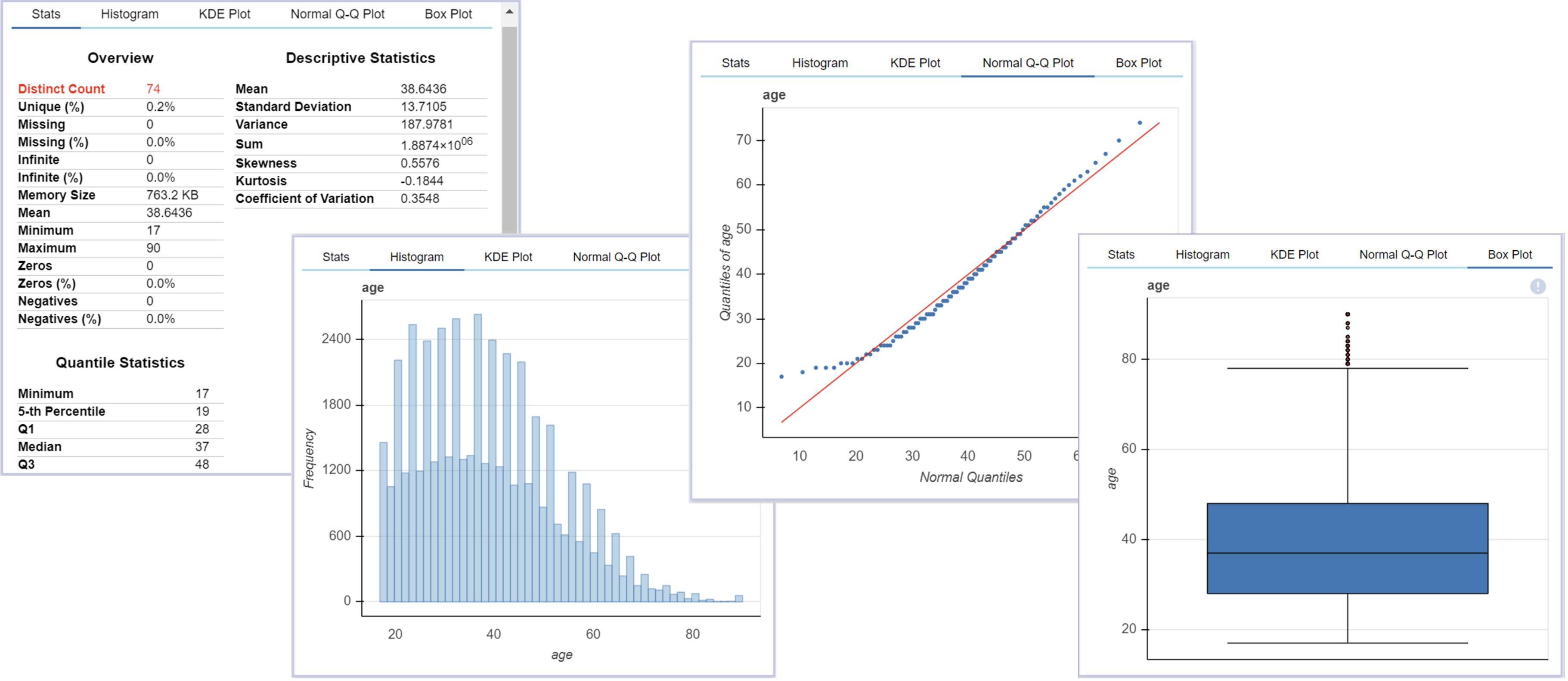
First Step to Verify Table

CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Data Quality

Previously Are Statistical; DataPrep For Example



Overview

Distinct Count	74
Unique (%)	0.2%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Memory Size	763.2 KB
Mean	38.6436
Minimum	17
Maximum	90
Zeros	0
Zeros (%)	0.0%
Negatives	0
Negatives (%)	0.0%

Descriptive Statistics

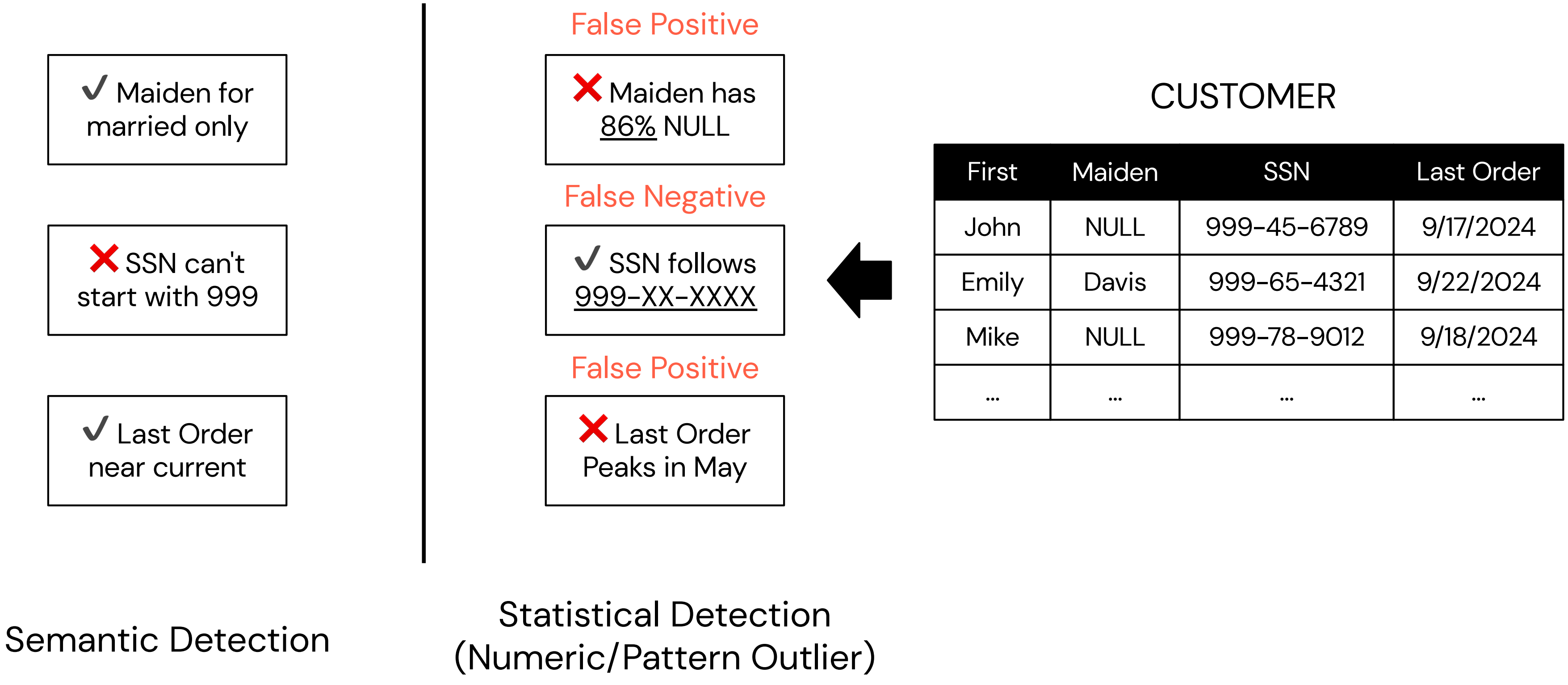
Mean	38.6436
Standard Deviation	13.7105
Variance	187.9781
Sum	1.8874×10^{06}
Skewness	0.5576
Kurtosis	-0.1844
Coefficient of Variation	0.3548

Quantile Statistics

Minimum	17
5-th Percentile	19
Q1	28
Median	37
Q3	48

Data Quality

Previously Are Statistical; Need Semantics



✓ Maiden for married only

✗ SSN can't start with 999

✓ Last Order near current

Semantic Detection

False Positive

✗ Maiden has 86% NULL

False Negative

✓ SSN follows 999-XX-XXXX

False Positive

✗ Last Order Peaks in May

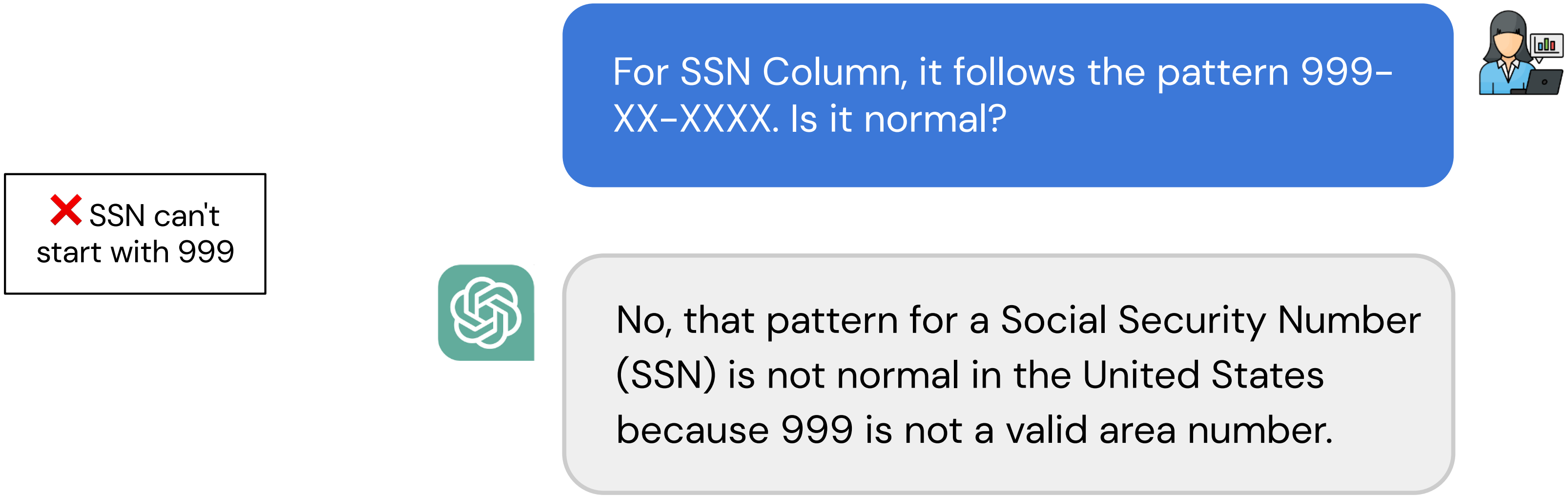
Statistical Detection (Numeric/Pattern Outlier)

CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	9/17/2024
Emily	Davis	999-65-4321	9/22/2024
Mike	NULL	999-78-9012	9/18/2024
...	...	...	...

Data Quality

Semantics Achieved by LLMs

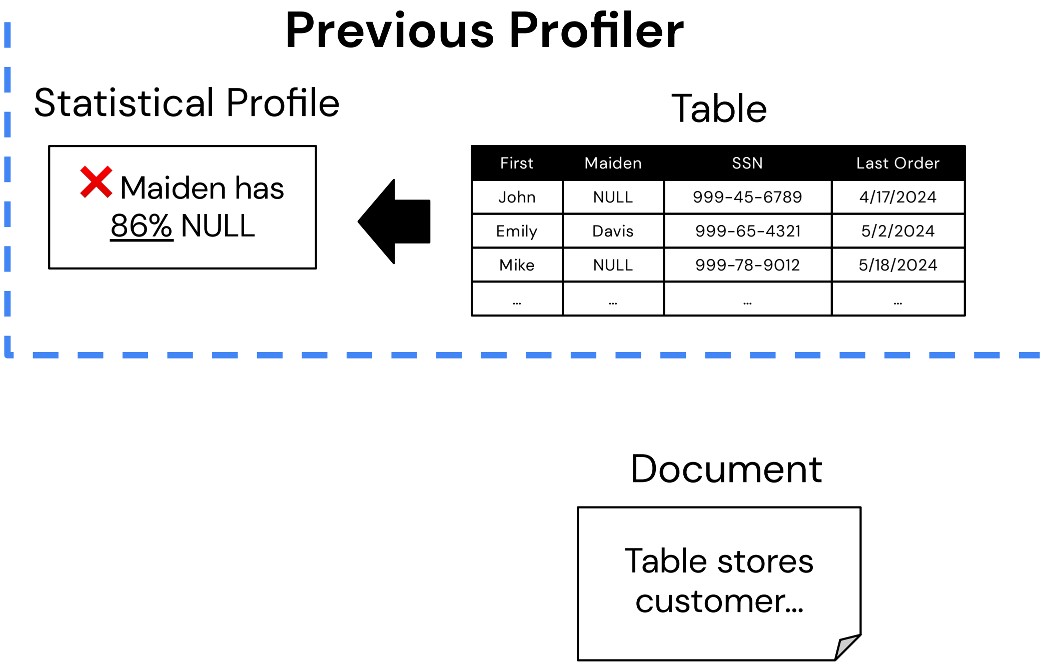


For SSN Column, it follows the pattern 999-XX-XXXX. Is it normal?

No, that pattern for a Social Security Number (SSN) is not normal in the United States because 999 is not a valid area number.

✗ SSN can't start with 999

How semantics is applied



Previous Profiler

Statistical Profile

❌ Maiden has 86% NULL

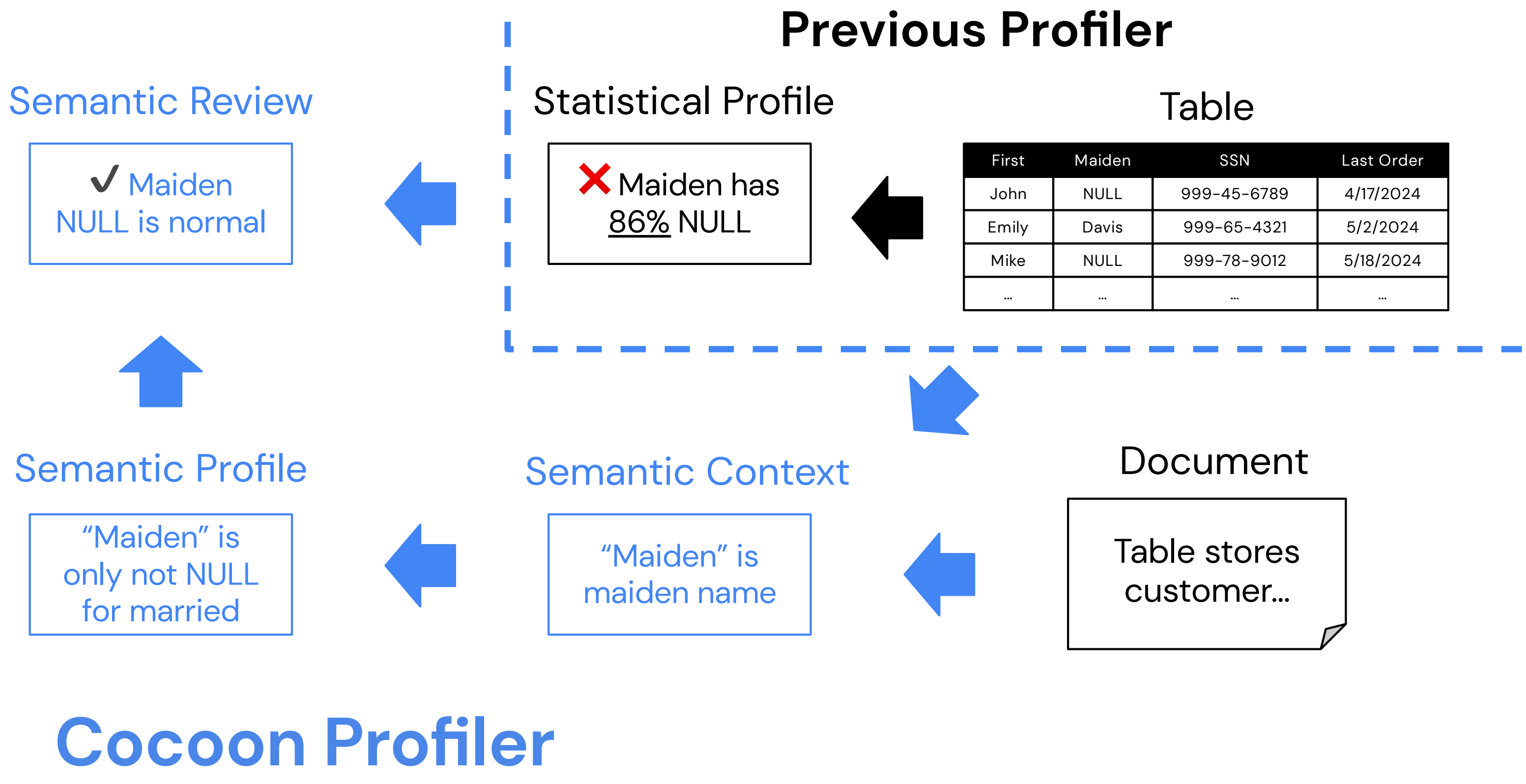
Table

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Document

Table stores customer...

How semantics is applied



Previous Profiler

Statistical Profile

✗ Maiden has 86% NULL

Table

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
…	…	…	…

Document

Table stores customer…

Semantic Context

“Maiden” is maiden name

Semantic Profile

“Maiden” is only not NULL for married

Semantic Review

✓ Maiden NULL is normal

Cocoon Profiler

Cocoon: Semantic Table Profiling Using LLMs

Zachary Huang, Eugene Wu
HILDA @ SIGMOD 2024

COLUMBIA UNIVERSITY

Table profiling

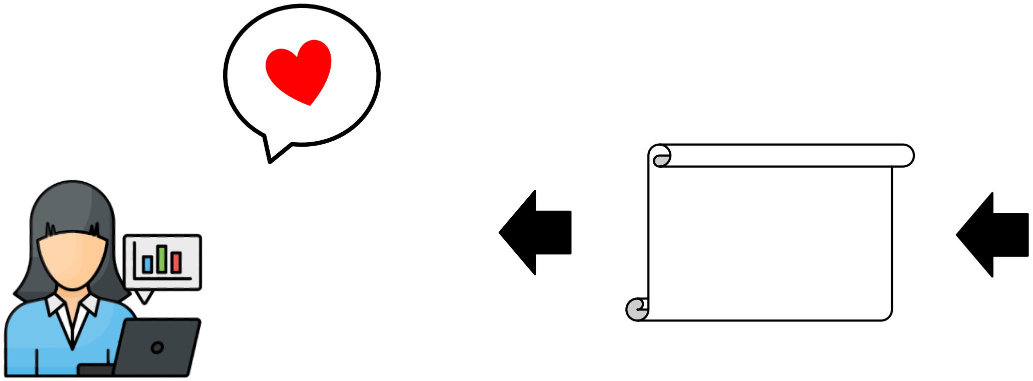
First step to understand table

CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Table profiling

First step to understand table



Interpretable representation

CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Table profiling

First step to understand table

❌ Maiden has 86% NULL

✓ SSN follows 999-XX-XXXX

❌ Last Order Peaks in May

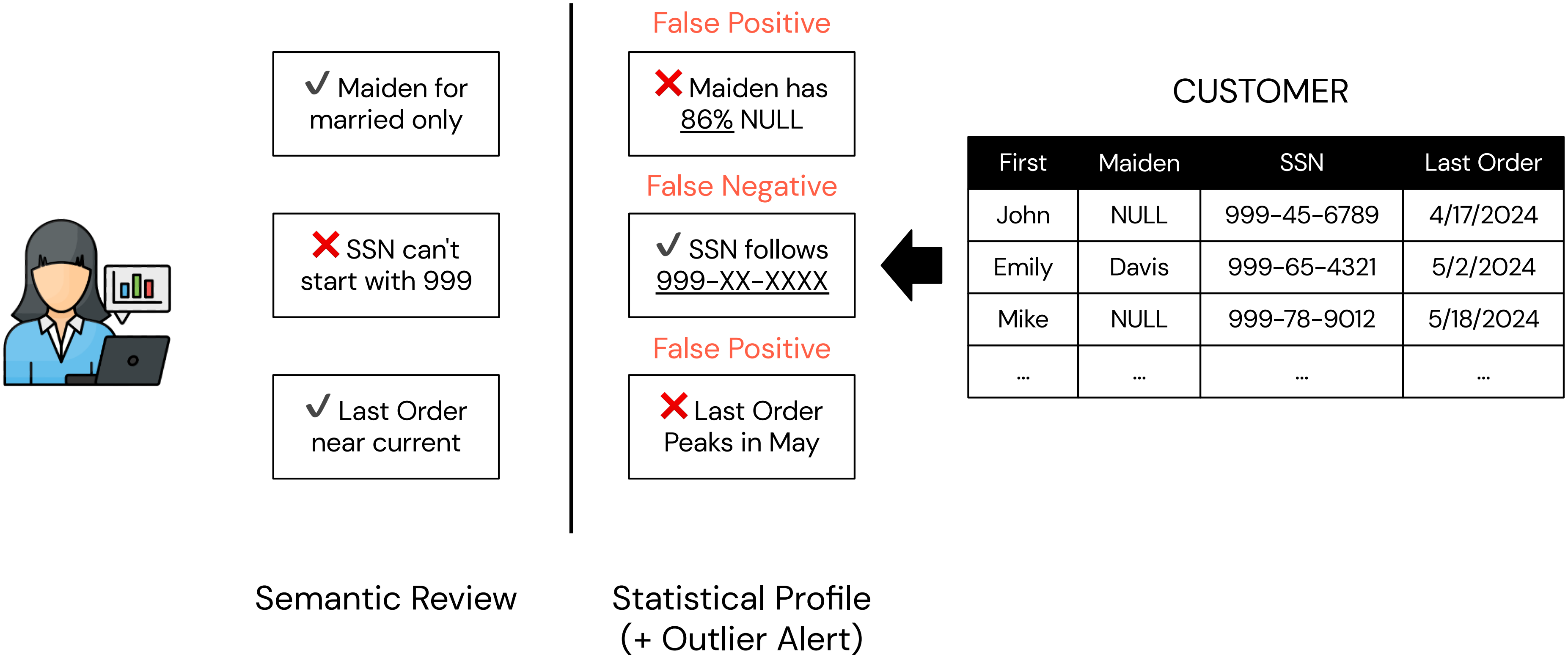
CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Statistical Profile
(+ Outlier Alert)

Table profiling

First step to understand table



✓ Maiden for married only

✗ SSN can't start with 999

✓ Last Order near current

Semantic Review

False Positive

✗ Maiden has 86% NULL

False Negative

✓ SSN follows 999-XX-XXXX

False Positive

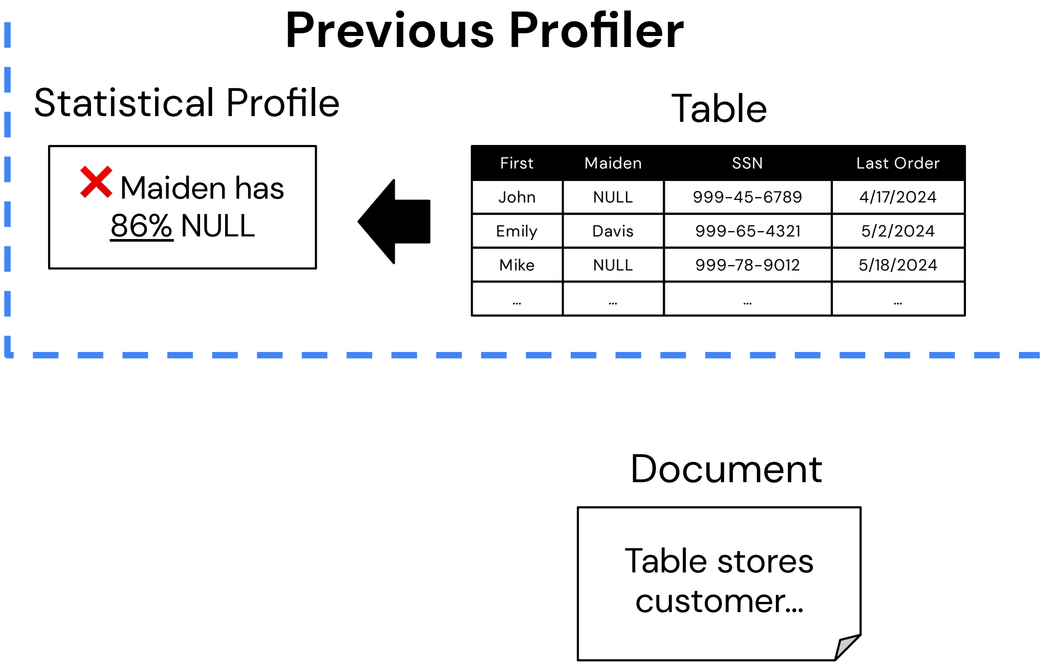
✗ Last Order Peaks in May

Statistical Profile (+ Outlier Alert)

CUSTOMER

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

How semantics is applied



Previous Profiler

Statistical Profile

✗ Maiden has 86% NULL

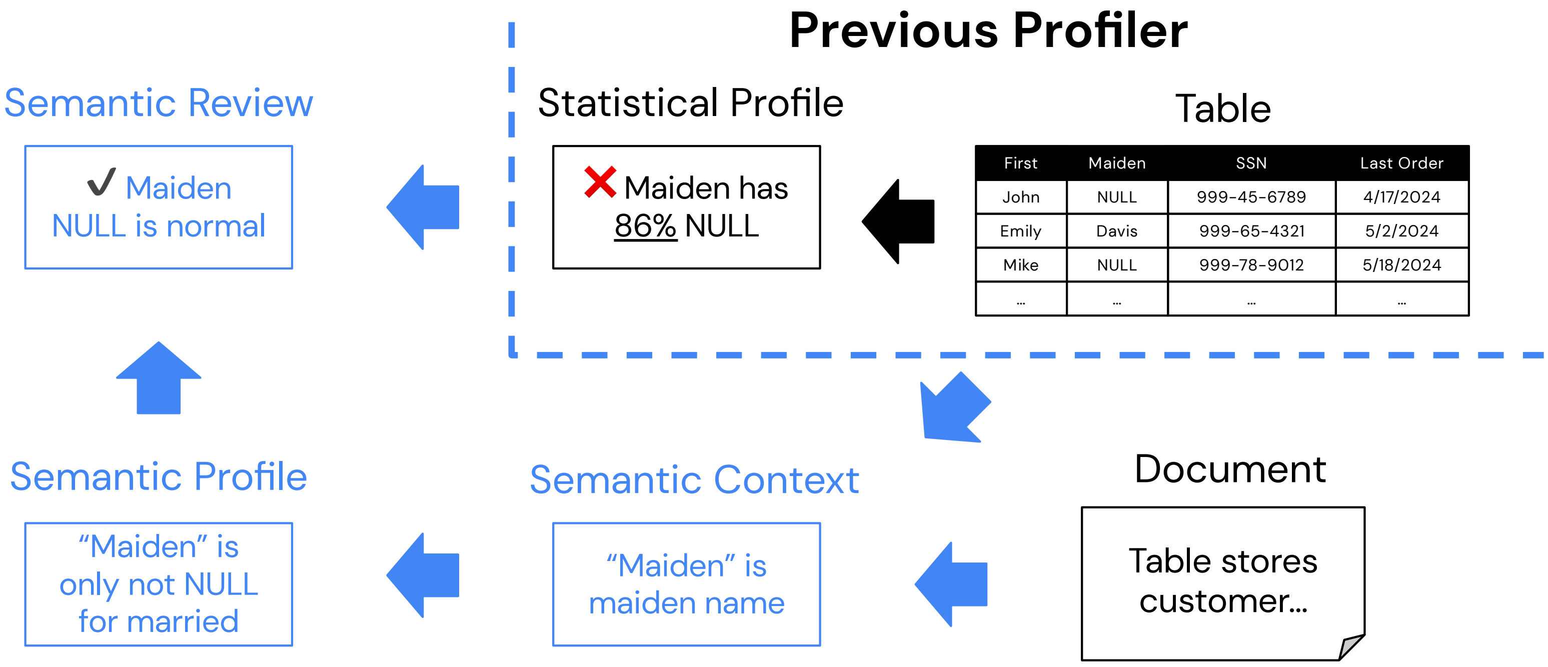
Table

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Document

Table stores customer...

How semantics is applied



Previous Profiler

Table

First	Maiden	SSN	Last Order
John	NULL	999-45-6789	4/17/2024
Emily	Davis	999-65-4321	5/2/2024
Mike	NULL	999-78-9012	5/18/2024
...	...	...	...

Statistical Profile

✗ Maiden has 86% NULL

Semantic Review

✓ Maiden NULL is normal

Document

Table stores customer...

Semantic Context

“Maiden” is maiden name

Semantic Profile

“Maiden” is only not NULL for married

Cocoon Profiler

LLMs for Semantic Table Profiling

Challenges

- **Many data quality issues to address**

  Duplication, Missing Values, Numeric Outliers, String Outliers…

- **Each issue needs Semantic Context, Profile and Review**

LLMs for Semantic Table Profiling

Challenges

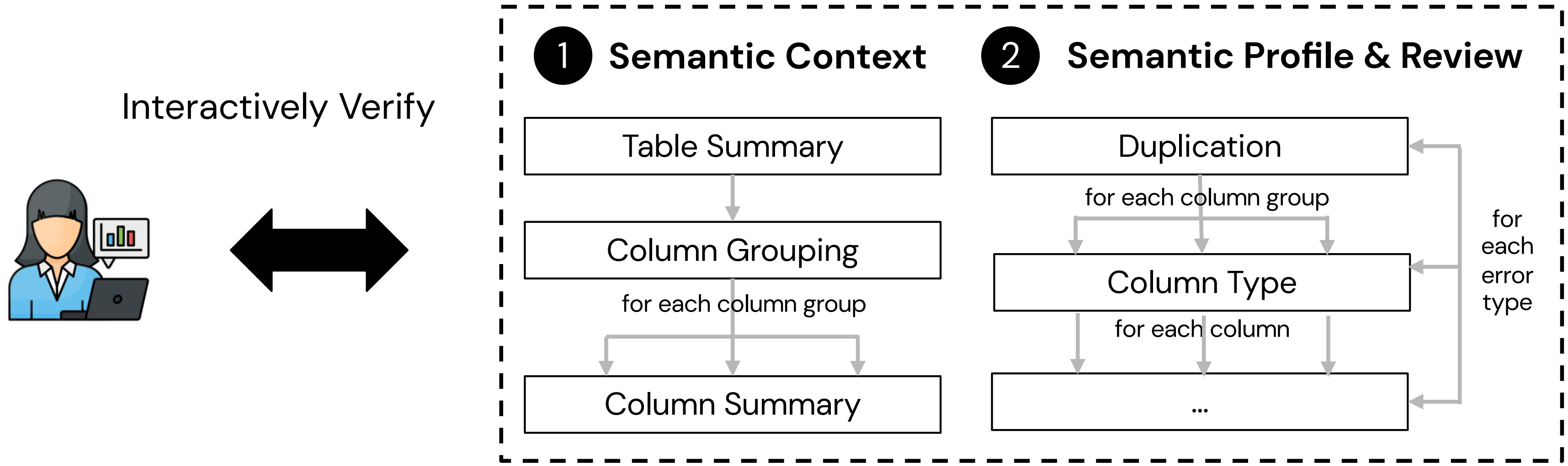
- **Many data quality issues to address**

  Duplication, Missing Values, Numeric Outliers, String Outliers…

- **Each issue needs Semantic Context, Profile and Review**

Task Decomposition, but how?

Cocoon System Design



Interactively Verify

1 Semantic Context

Table Summary

Column Grouping

for each column group

Column Summary

2 Semantic Profile & Review

Duplication

for each column group

Column Type

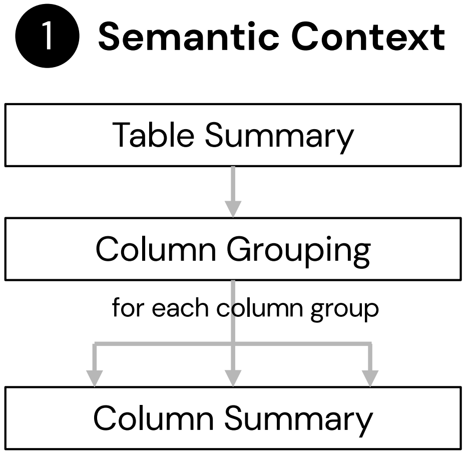
for each column

...

for each error type

Cocoon System Design

Semantic Context



1 Semantic Context

Table Summary

Column Grouping

for each column group

Column Summary

Table Summary: Given table samples (first 5 rows) and document, describe the table in NL

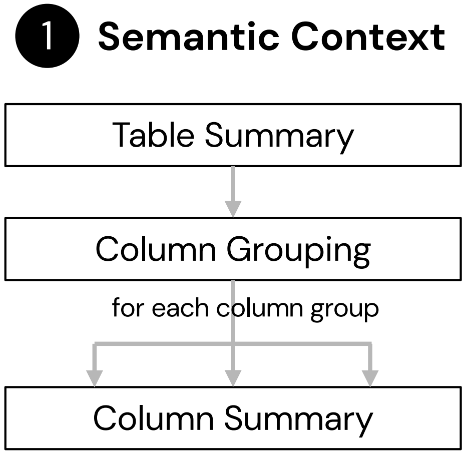
Column Grouping: Given the Table Summary, cluster columns into groups. This is to

1. Find columns together express single concept (E.g., longitude, latitude)
2. Help human understand

Column Summary: Given samples for each Column Group and Table Summary, describe the group in NL

Cocoon System Design

Semantic Context



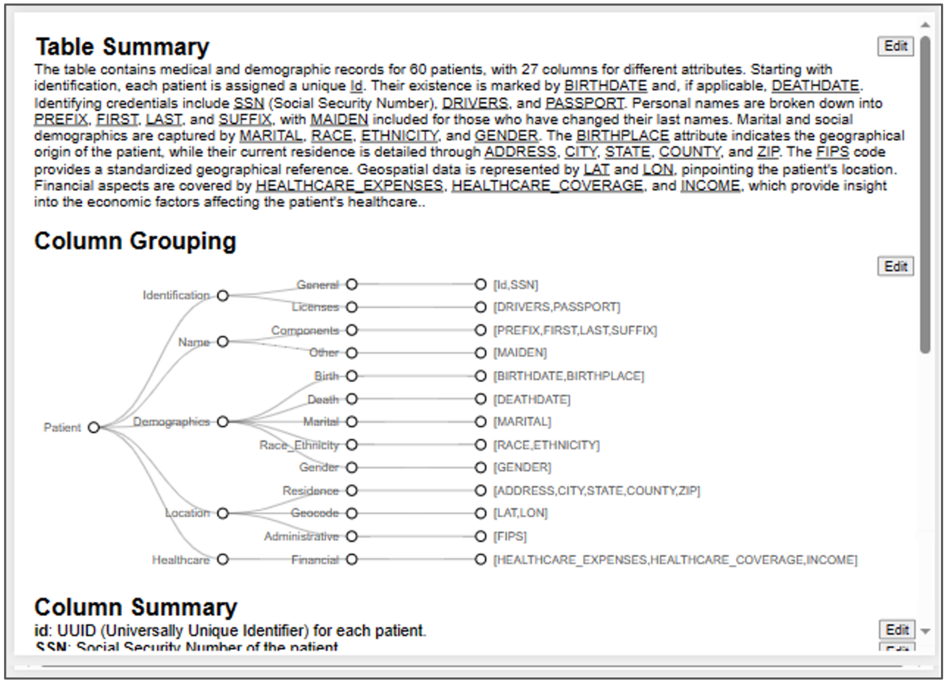


Table Summary

The table contains medical and demographic records for 60 patients, with 27 columns for different attributes. Starting with identification, each patient is assigned a unique Id. Their existence is marked by BIRTHDATE and, if applicable, DEATHDATE. Identifying credentials include SSN (Social Security Number), DRIVERS, and PASSPORT. Personal names are broken down into PREFIX, FIRST, LAST, and SUFFIX, with MAIDEN included for those who have changed their last names. Marital and social demographics are captured by MARITAL, RACE, ETHNICITY, and GENDER. The BIRTHPLACE attribute indicates the geographical origin of the patient, while their current residence is detailed through ADDRESS, CITY, STATE, COUNTY, and ZIP. The FIPS code provides a standardized geographical reference. Geospatial data is represented by LAT and LON, pinpointing the patient's location. Financial aspects are covered by HEALTHCARE_EXPENSES, HEALTHCARE_COVERAGE, and INCOME, which provide insight into the economic factors affecting the patient's healthcare..

Column Grouping

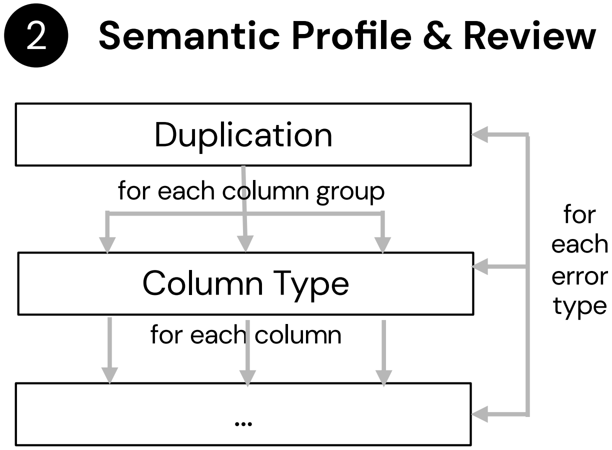
Column Summary

id: UUID (Universally Unique Identifier) for each patient.
SSN: Social Security Number of the patient

Interface for human feedback

Cocoon System Design

Semantic Profile & Review



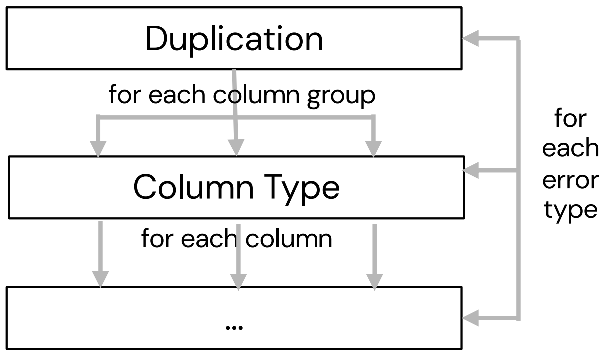
We consider errors:

- Duplication
- Column Type
- Uniqueness
- Disguised Missing Value
- Missing Value
- Numeric Outliers
- String Outliers
- Missing Record

Cocoon System Design

Semantic Profile & Review

2 Semantic Profile & Review



Duplication

for each column group

Column Type

for each column

...

for each error type

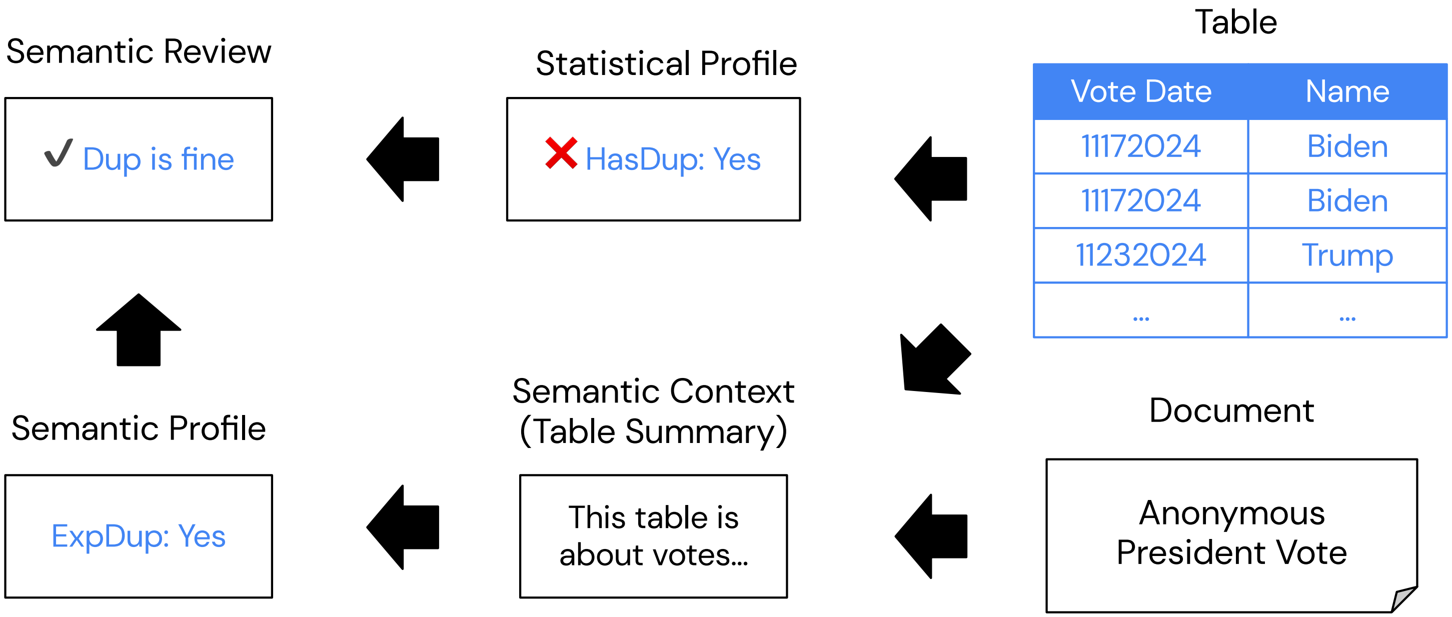
We consider errors:

- Duplication
- Column Type
- Uniqueness
- Disguised Missing Value
- Missing Value
- Numeric Outliers
- String Outliers
- Missing Record

Design: What are the Semantic Profile & Review

Cocoon System Design

Semantic Profile & Review for Duplication



Semantic Review

✓ Dup is fine

Statistical Profile

✗ HasDup: Yes

Table

Vote Date	Name
11172024	Biden
11172024	Biden
11232024	Trump
...	...

Semantic Profile

ExpDup: Yes

Semantic Context (Table Summary)

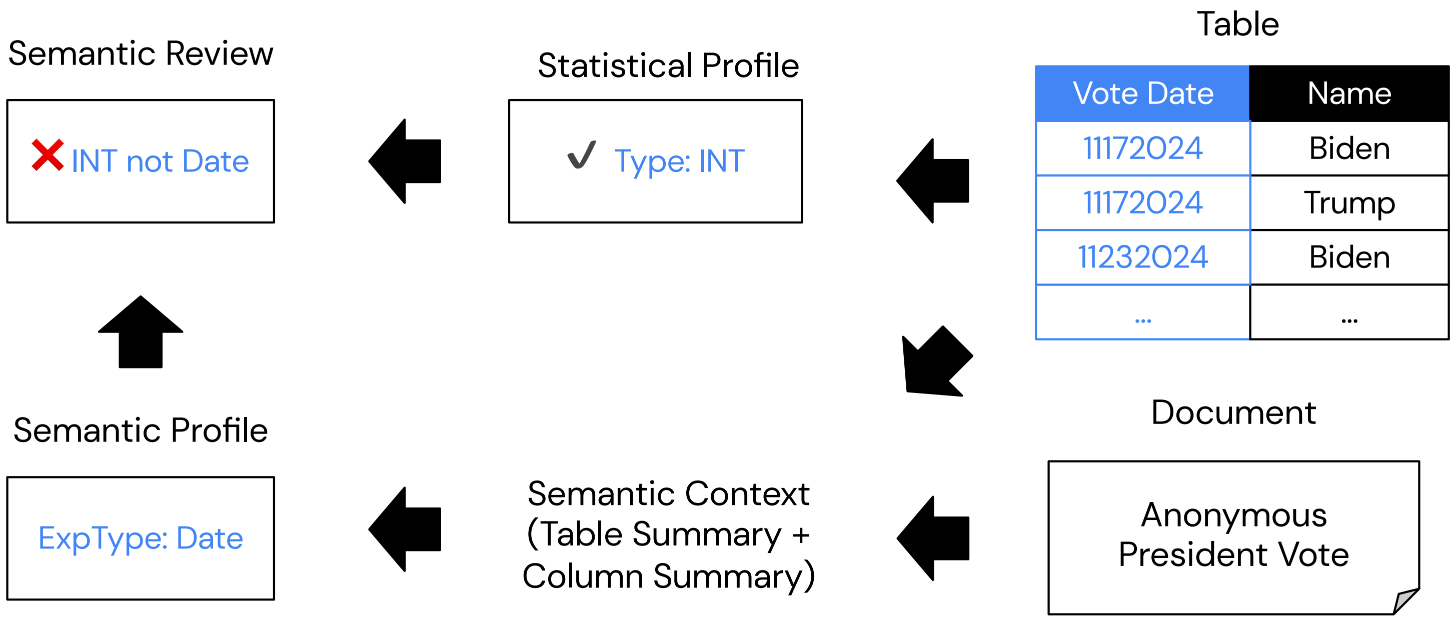
This table is about votes...

Document

Anonymous President Vote

Cocoon System Design

Semantic Profile & Review for Data Type



Semantic Review

❌ INT not Date

Statistical Profile

✓ Type: INT

Table

Vote Date	Name
11172024	Biden
11172024	Trump
11232024	Biden
...	...

Semantic Profile

ExpType: Date

Semantic Context (Table Summary + Column Summary)

Document

Anonymous President Vote

Cocoon System Design

Final Output

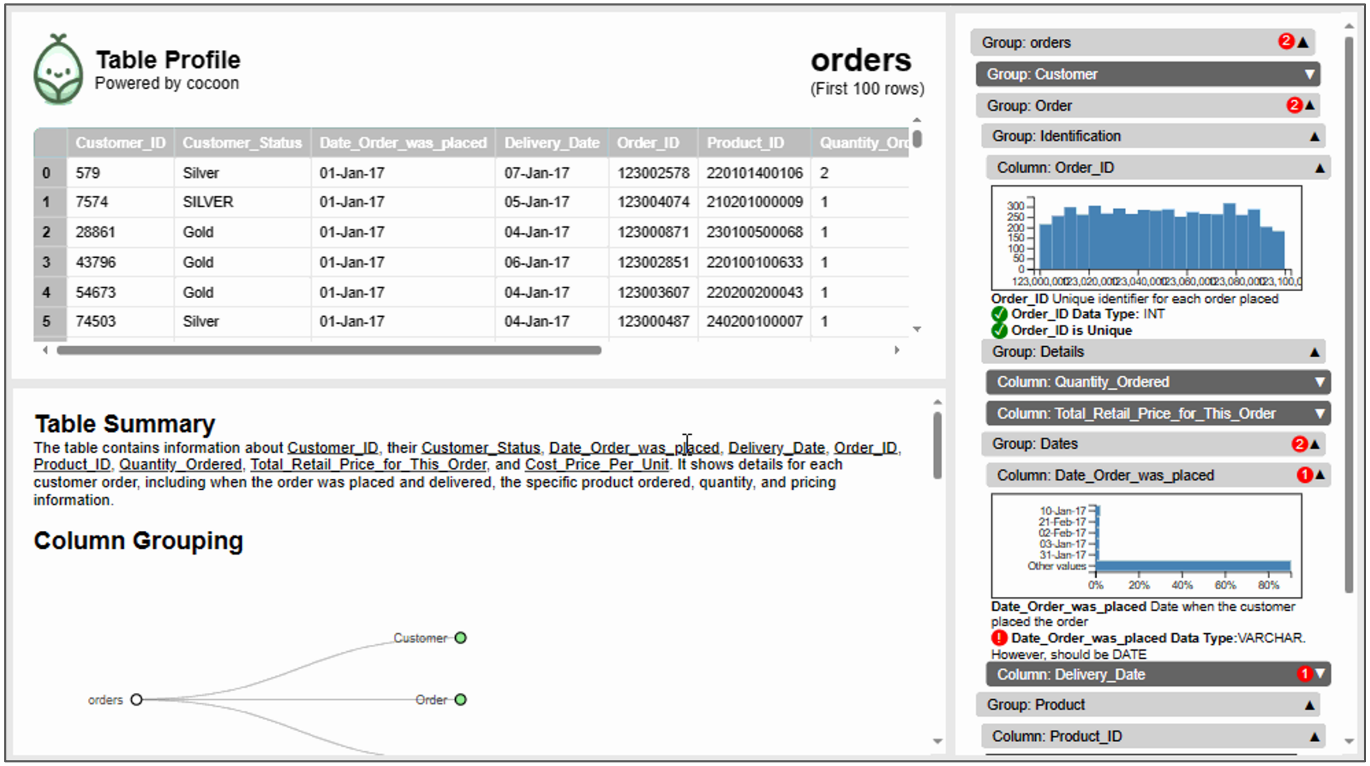


Table Profile
Powered by cocoon

orders
(First 100 rows)

	Customer_ID	Customer_Status	Date_Order_was_placed	Delivery_Date	Order_ID	Product_ID	Quantity_Ord
0	579	Silver	01-Jan-17	07-Jan-17	123002578	220101400106	2
1	7574	SILVER	01-Jan-17	05-Jan-17	123004074	210201000009	1
2	28861	Gold	01-Jan-17	04-Jan-17	123000871	230100500068	1
3	43796	Gold	01-Jan-17	06-Jan-17	123002851	220100100633	1
4	54673	Gold	01-Jan-17	04-Jan-17	123003607	220200200043	1
5	74503	Silver	01-Jan-17	04-Jan-17	123000487	240200100007	1

Table Summary

The table contains information about Customer_ID, their Customer_Status, Date_Order_was_placed, Delivery_Date, Order_ID, Product_ID, Quantity_Ordered, Total_Retail_Price_for_This_Order, and Cost_Price_Per_Unit. It shows details for each customer order, including when the order was placed and delivered, the specific product ordered, quantity, and pricing information.

Column Grouping

Group: orders 2 ▲

Group: Customer ▼

Group: Order 2 ▲

Group: Identification ▲

Column: Order_ID ▲

Order_ID Unique identifier for each order placed

✓ **Order_ID Data Type:** INT

✓ **Order_ID is Unique**

Group: Details ▲

Column: Quantity_Ordered ▼

Column: Total_Retail_Price_for_This_Order ▼

Group: Dates 2 ▲

Column: Date_Order_was_placed 1 ▲

Date_Order_was_placed Date when the customer placed the order

! **Date_Order_was_placed Data Type:**VARCHAR. However, should be DATE

Column: Delivery_Date 1 ▼

Group: Product ▲

Column: Product_ID ▲

Conclusion

- Statistical profiling has false positives/negatives
- LLMs help interactively improve profiling with semantics
- Task decomposition improves quality