Empowering Natural Language to Visualization Neural Translation using Synthesized Benchmarks
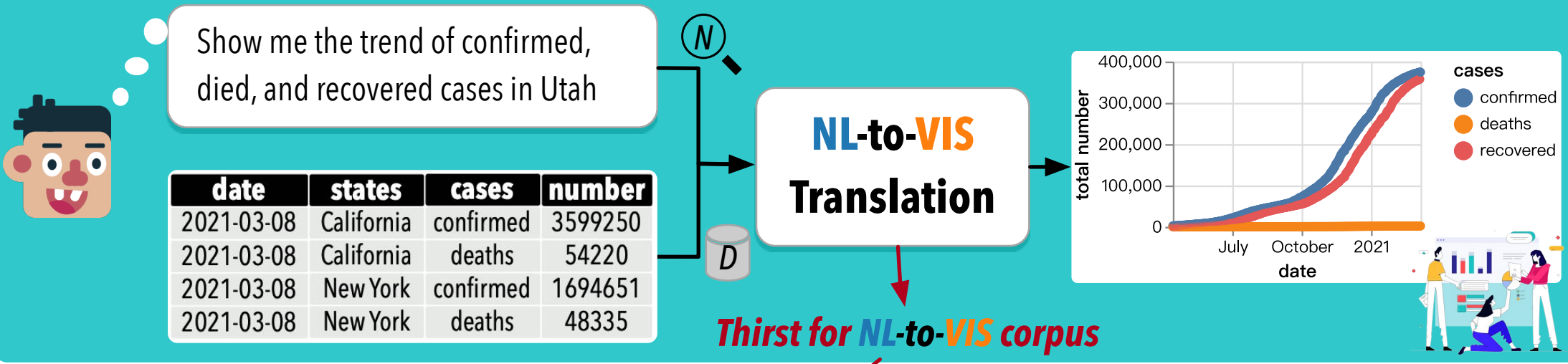
Yuyu Luo   Guoliang Li   Chengliang Chai
Tsinghua University, China

Jiawei Tang
American School of Doha, Qatar

Natural Language to Visualization Translation

Show me the trend of confirmed, died, and recovered cases in Utah

date	states	cases	number
2021-03-08	California	confirmed	3599250
2021-03-08	California	deaths	54220
2021-03-08	New York	confirmed	1694651
2021-03-08	New York	deaths	48335

NL-to-VIS Translation

Thirst for NL-to-VIS corpus

nvBench: NL-to-VIS Benchmark

Statistics of nvBench

#-Domains	#-Databases	#-Tables
105	153	780
Categorical (%)	Temporal (%)	Quantitive (%)
68.78 %	11.58 %	19.64 %
Avg (#-Rows)	Max (#-Rows)	Min (#-Rows)
1309.65	183,978	1

VIS Types	#-VIS	#-(NL, VIS)
Bar Chart	5523	19407
Pie Chart	520	1750
Line Chart	380	1562
Scatter Chart	226	1041
Stacked Bar	359	1172
Grouping Line	72	271
Grouping Scatter	127	547
All Types	7247	25750

Sample (NL, VIS) examples in nvBench

Database: products_for_hire

Table: payments

...	payment_type_code	amount_paid_in_full_yn	payment_date	aoumnt_due	amount_paid
...	Check	1	2018-03-09 16:28:00	369.52	206.27
...	Cash	1	2018-03-03 13:39:44	278.60	666.45
...	Credit Card	0	2018-03-22 15:00:23	840.06	135.70
...	Check	0	2018-03-22 02:28:11	678.29	668.40
...	...	...	...	...	...

NL Query

What are the payment date of the payment with amount paid higher than 300 or with payment type is 'Check, and count them by a line chart

VIS Query (Vega-Zero, a variant of Vega-Lite)

```
mark      line
data      payments
encoding  x payment_date y aggregate count payment_date
transform filter amount_paid > 300 or payment_type_code='check' group x
```

VIS Specification (Vega-Lite)
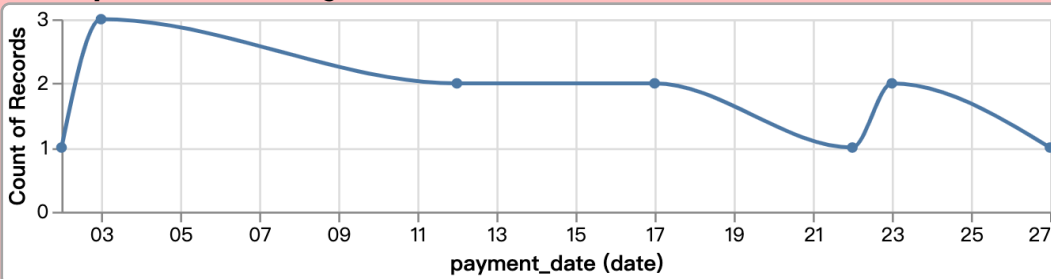
Properties

nvBench_id: *2617*
DB_id: *products_for_hire*
Chart: *Line*
Hardness: *Hard*

Synthesizing nvBench from NL-to-SQL Benchmarks

n_Q Show flight number, origin, destination of all flights in the alphabetical order of the departure cities.

Q **Select** *flno, origin, destn*
From *Flight*
Order By *origin*

NL2SQL-to-NL2VIS Synthesizer

n_{21} Give me a histogram to compare the number of flights to each destination city?

n_{22} How many flights in each destination city? Return a bar chart.

t_2 **mark** bar
data Flight
encoding x destn y aggregate count destn
transform group x

n_{11} I prefer a pie chart to understand how many flights from each origin city.

n_{12} Show me the proportion of the number of flights by each origin city.

t_1 **mark** pie
data Flight
encoding x origin y aggregate count origin
transform group x

nvBench: https://github.com/TsinghuaDatabaseGroup/nvBench