

DIY (Debug It Yourself)

Helping People Assess the Correctness of
Natural Language to SQL Systems

Arpit Narechania

arpitnarechania@gatech.edu

 @arpitnarechania



Adam Fourney

adamfo@microsoft.com

 @adamfourney



Bongshin Lee

bongshin@microsoft.com

 @bongshin



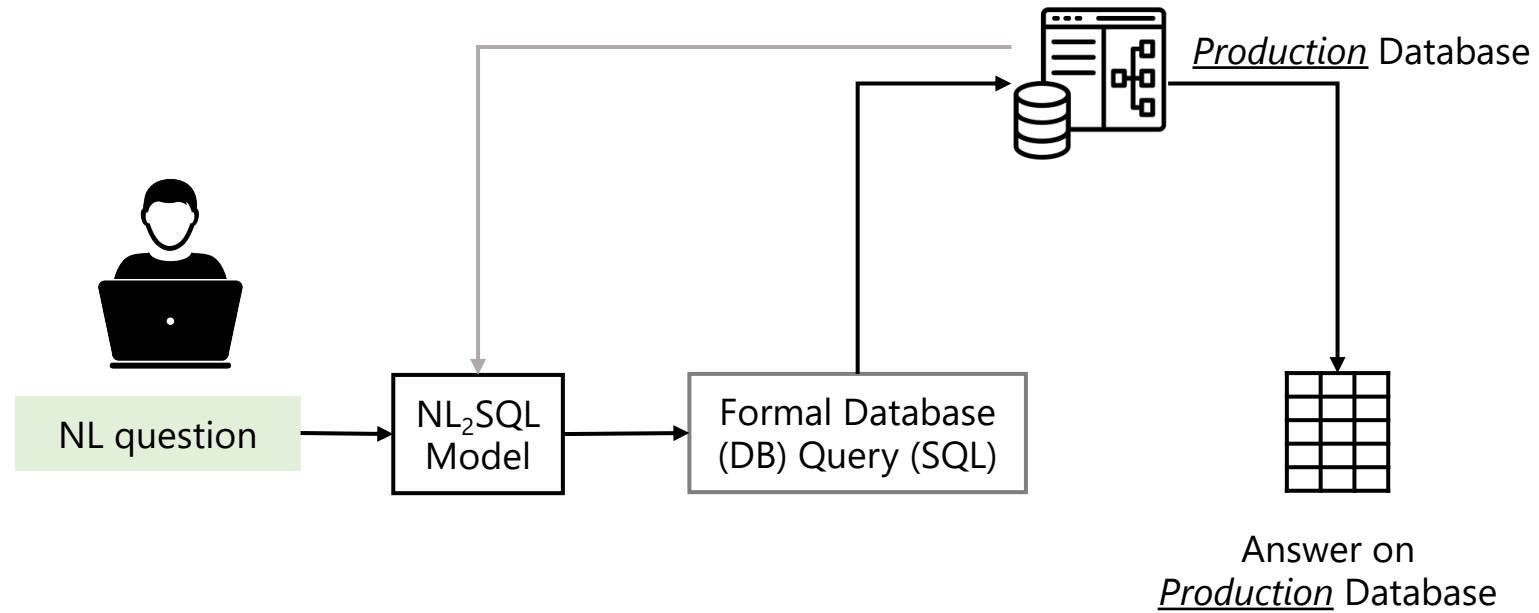
Gonzalo Ramos

goramos@microsoft.com

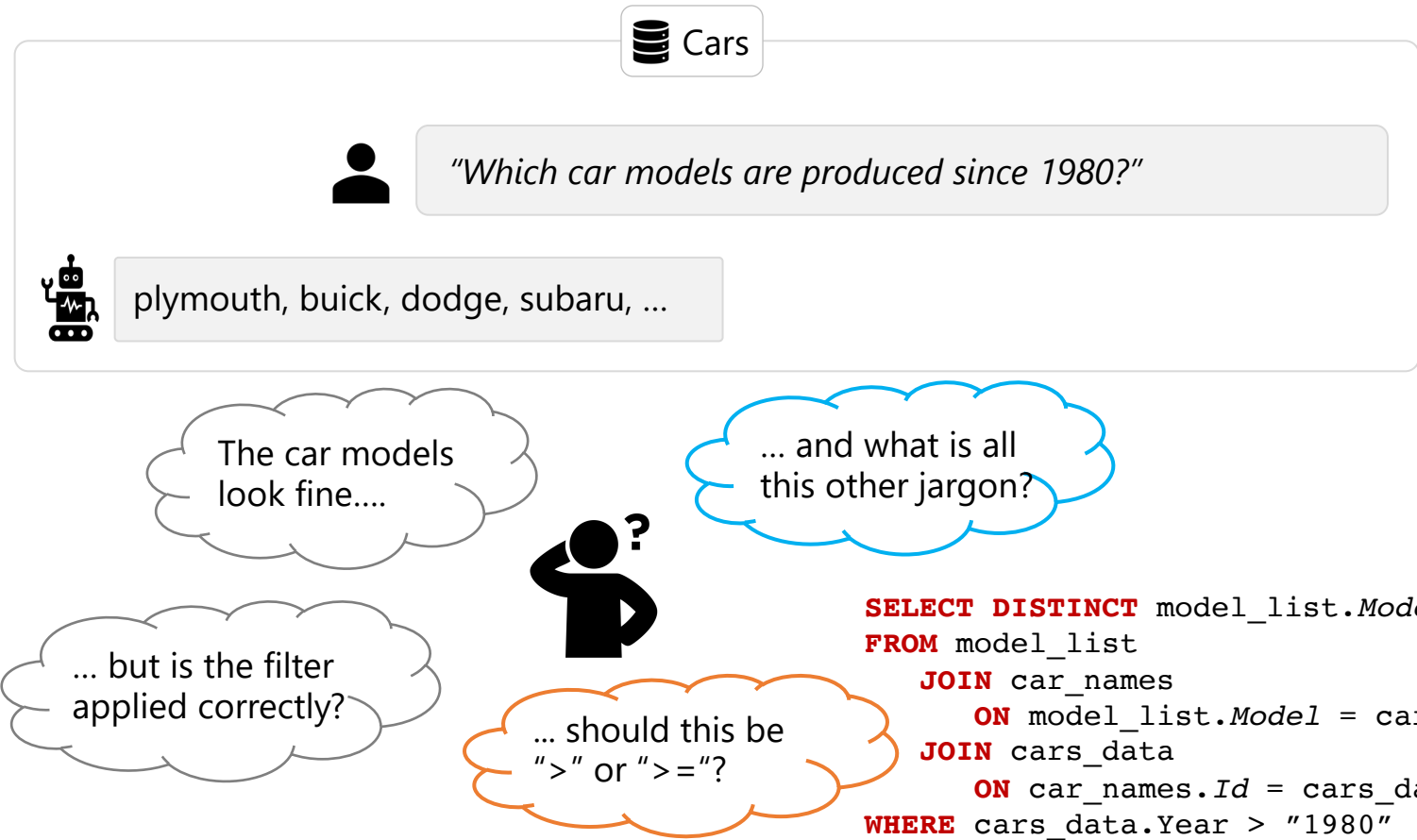
 @GonzaloWorks



NL₂SQL system



An NL₂SQL scenario

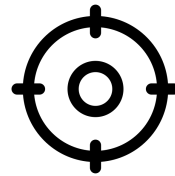


Research Questions

How can we improve Human-AI NL₂SQL scenarios by helping users to



Inspect system responses for correctness



Isolate errors, if any



Fix errors, if possible

Demo

Who can benefit from DIY?



Domain experts with limited DB experience.



Information workers who can't write complex DB queries.



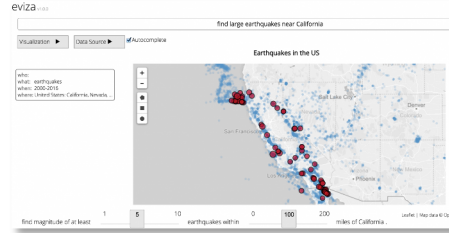
Computer scientists who took a DB class but are now rusty.

Recent Work

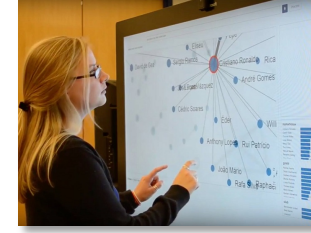
NLIs in Visualization



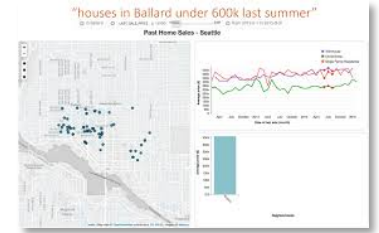
DataTone, *UIST 2015*



Eviza, *UIST 2016*



Orko, *TVCG 2018*

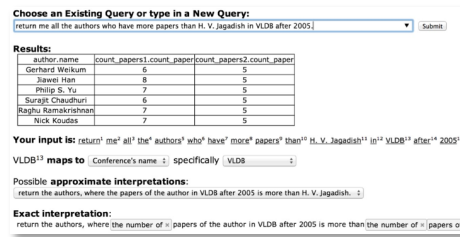


Evizeon, *TVCG 2018*

NLIs in Databases



QUEST, *VLDB 2013*



NaLIR, *SIGMOD 2014*

Natural Language Question : What are the countries that joined the NATO before 2004?	
Ground Truth SQL Query : SELECT country WHERE date of join < 2004	
Initial SQL Query : SELECT count (country) WHERE date of join < 2004	

Turn Side	NL Expression using Rule-Based Templates	Response
System:	Could you please confirm if you are asking total number of the results?	validate_agg (1, 2)
User:	No, that's not true.	negate()
System:	Could you please select one of the following options : (i) average number of results , (ii) all the results?	offer_agg (average, no_agg)
User:	It should be all the results .	select(2)
System:	Could you please confirm if date of join is 2004 ?	validate_where_change(6, 1)
User:	That's wrong.	negate()
System:	Could you please select one of the following options : (i) date of join is before 2004 , (ii) date of invitation is before 2004?	offer_where (date of join < ; date of invitation < ;)
User:	Date of join is before 2004.	select(1)

DialSQL, *ACL 2018*

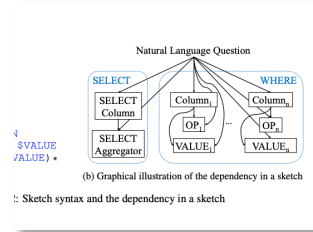
Question: "Greece held its last Olympics in what year?"

MAX(Year)	Country	City
1896	Greece	Athens
...	...	...
2004	Greece	Athens
2008	China	Beijing
2012	UK	London
2016	Brazil	Rio de Janeiro

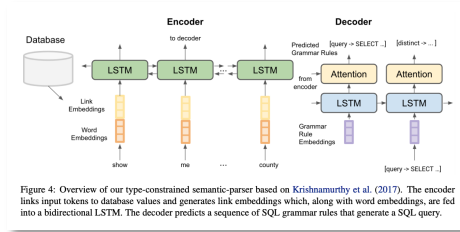
u : maximum value in column Year where Country is Greece.

Berant et al., *ICDE 2018*

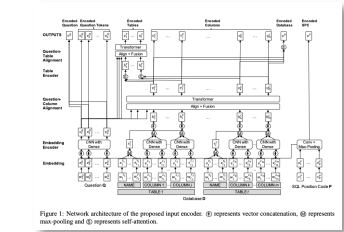
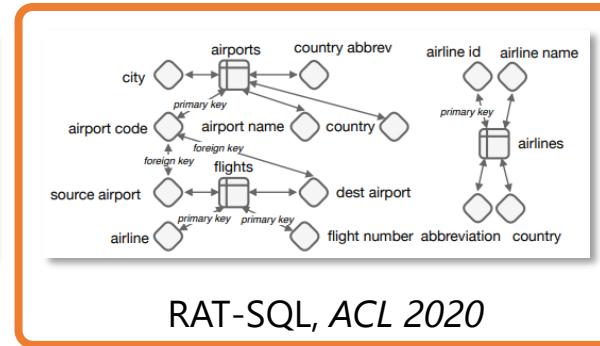
NL₂SQL DL models



SQLNet, *arXiv 2017*

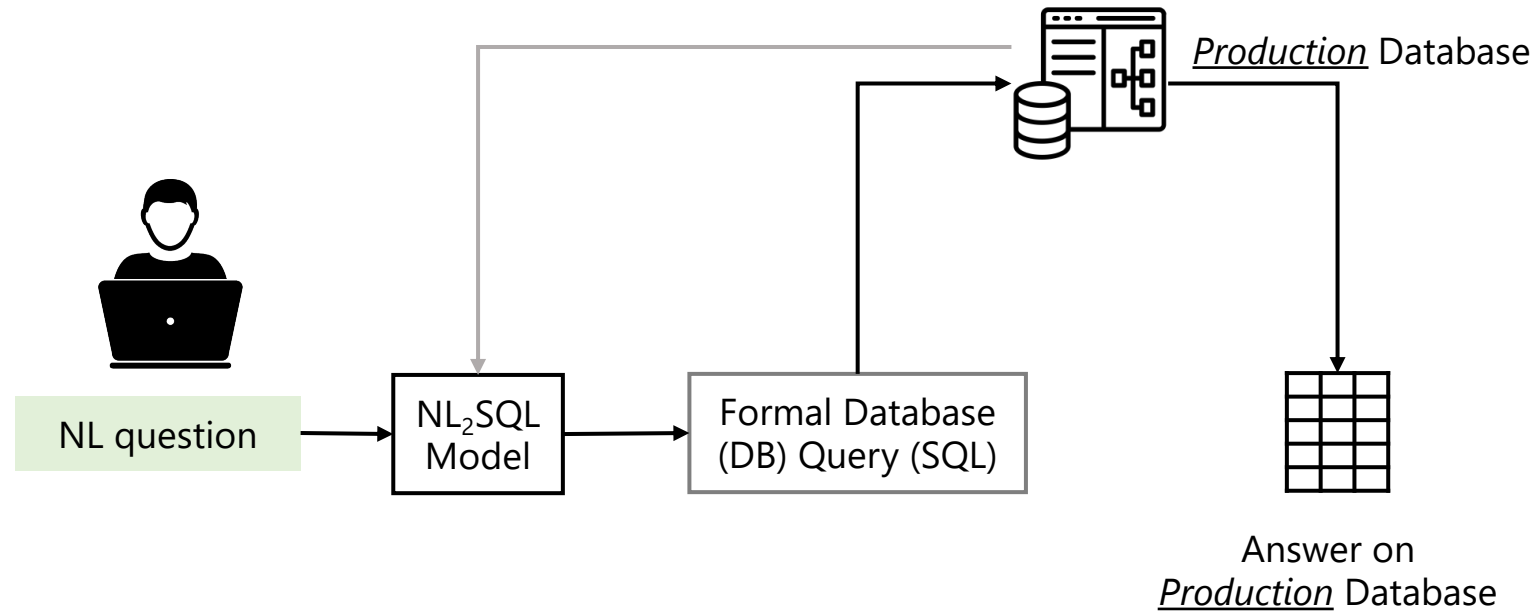


Lin et al., *arXiv 2019*

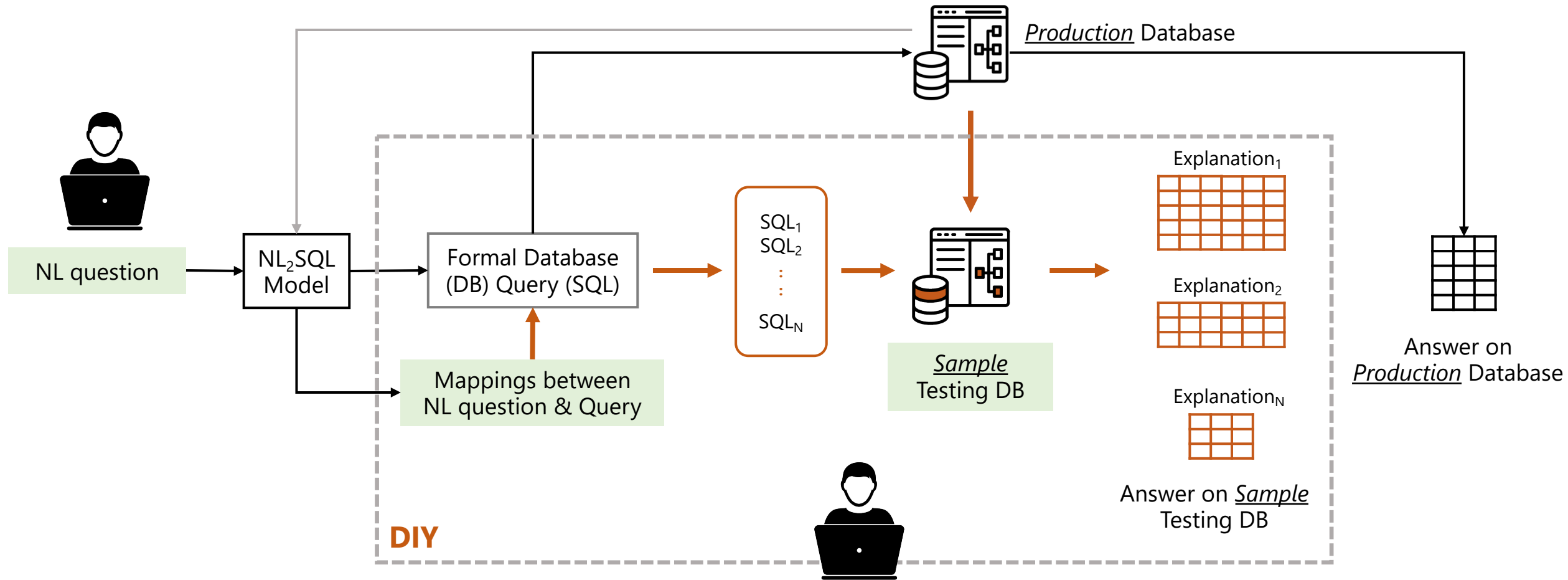


RYANSQL, *arXiv 2020*

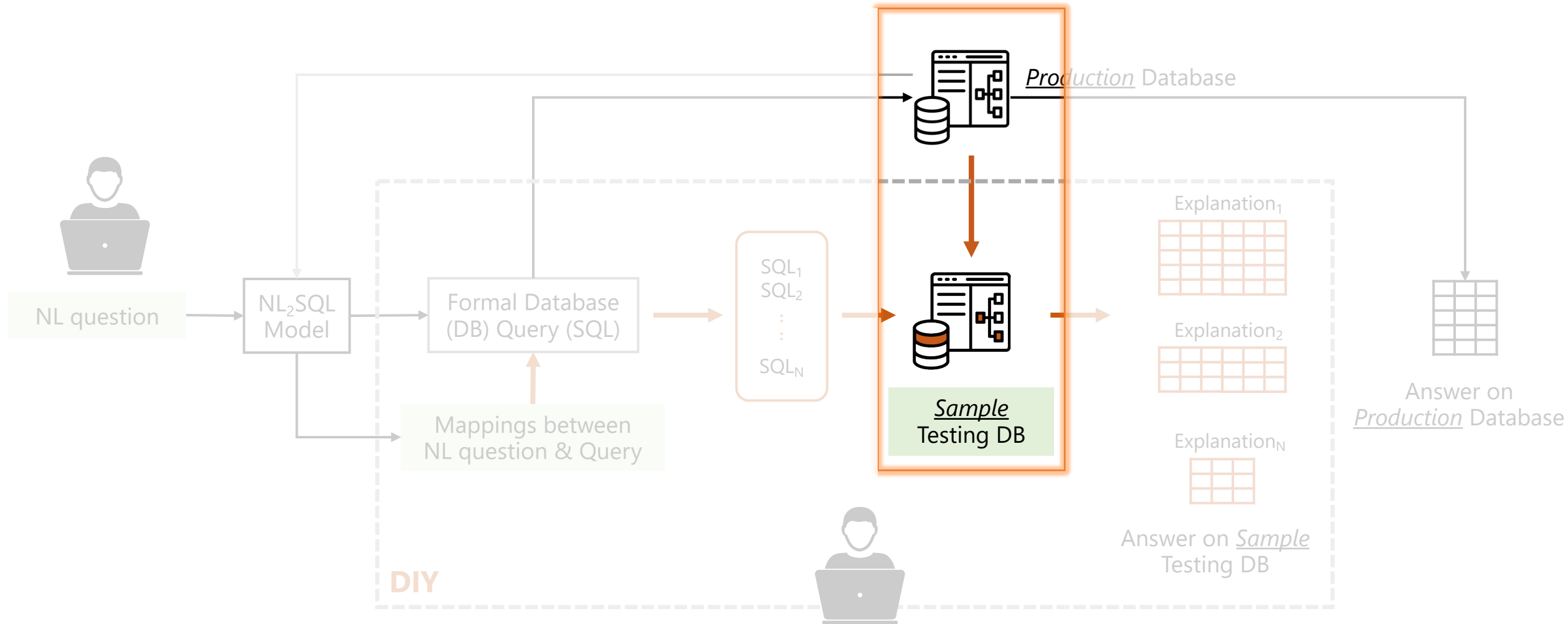
NL₂SQL systems: General Pipeline



DIY: Debug It Yourself



DIY: Sample Testing DB



Sample DB: Create Schema

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

```

SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
    
```

Sample DB: Create Schema

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

```

SELECT DISTINCT
    model_list.Model
FROM
    model_list
JOIN car_names
    ON model_list.Model
       = car_names.Model
JOIN cars_data
    ON car_names.Id
       = cars_data.Id
WHERE
    cars_data.Year > "1980"
    
```

Sample DB: Create Schema

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

```

SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
    
```

Sample DB: Create Schema

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
```



Create DB Schema (DDL)

```
CREATE TABLE model_list {
  Model varchar(255)
};
CREATE TABLE car_names {
  Id int,
  Model varchar(255)
};
CREATE TABLE cars_data {
  Id int,
  Year int
};
```



constraints are also created.



Sample Testing DB

model_list	
Model	

car_names	
Id	Model

cars_data	
Id	Year

Sample DB: Insert Data

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



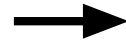
Primary Key



Foreign Key to the table on the left

SQL Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
```



SQL Create DB Schema (DDL)

SQL Insert Values into DB (DML)



Sample Testing DB

model_list	
Model	

car_names	
Id	Model

cars_data	
Id	Year

Sample DB: Insert Data

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

Smart Constraints

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

JOINS should be successful

Some cars should have **Year > 1980**; some should have **Year <= 1980**

Sample DB: Insert Data

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

Smart Constraints

JOINS should be successful

Some cars should have **Year > 1980**; some should have **Year <= 1980**

Sample DB: Insert Data

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
```



SQL Create DB Schema (DDL)

SQL Insert Values into DB (DML)

```
INSERT INTO model_list
(Model) VALUES (...)

INSERT INTO car_names (Id,
Model) VALUES (...)

INSERT INTO cars_data (Id,
Year) VALUES (...)
```



Sample Testing DB

model_list	
Model	
buick	
chevrolet	
honda	
peugeot	
plymouth	

car_names	
Id	Model
306	cadillac
307	peugeot
345	honda
346	plymouth
347	buick

cars_data	
Id	Year
306	1979
307	1979
345	1980
346	1981
347	1982

Sample DB: Insert Data

continents	
Cont-Id	Continent
1	america
2	europa
3	asia
4	africa
5	australia
...	...

countries		
Count-ryId	Country-Name	Conti-nent
1	usa	1
2	germany	2
3	france	2
4	japan	3
5	italy	2
...	...	...

car_makers		
Id	Maker	Country
1	amc	1
2	volkswagen	2
3	bmw	2
4	gm	1
5	ford	1
6	honda	4
7	fiat	5
...	...	...

model_list	
Model	Maker
amc	1
audi	2
buick	4
chevrolet	4
honda	6
peugeot	11
plymouth	15
...	...

car_names		
Id	Make	Model
1	chevelle	chevrolet
2	skylark	buick
306	eldorado	cadillac
307	peugot504	peugeot
345	accord	honda
346	reliant	plymouth
347	skylark	buick
...	...	...

cars_data					
Id	Horse-power	Weight	Edispl	Accel-erate	Year
1	115	3090	133	17.50	1970
2	150	3436	318	11	1970
306	95	2833	198	15.5	1979
307	85	3070	200	16.70	1979
345	77	3530	183	20.10	1980
346	88	2279	97	19	1981
347	70	3050	2245	16.90	1982
...	...	...	...	...	...



Production Database



Primary Key



Foreign Key to the table on the left

SQL Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
  ON model_list.Model
    = car_names.Model
JOIN cars_data
  ON car_names.Id
    = cars_data.Id
WHERE
  cars_data.Year > "1980"
```



SQL Create DB Schema (DDL)



SQL Insert Values into DB (DML)



Sample Testing DB

model_list	
Model	
buick	
chevrolet	
honda	
peugeot	
plymouth	

car_names	
Id	Model
306	cadillac
307	peugeot
345	honda
346	plymouth
347	buick

cars_data	
Id	Year
306	1979
307	1979
345	1980
346	1981
347	1982

This Sample DB is can now effectively explain the generated SQL to the user.

Sample DB: Human-in-the-loop

The machine-generated data in the **sample DB** **may not** always be appropriate, e.g.,

- “Show me car manufacturers on Mars.” does not have a positive constraint, **yet!**

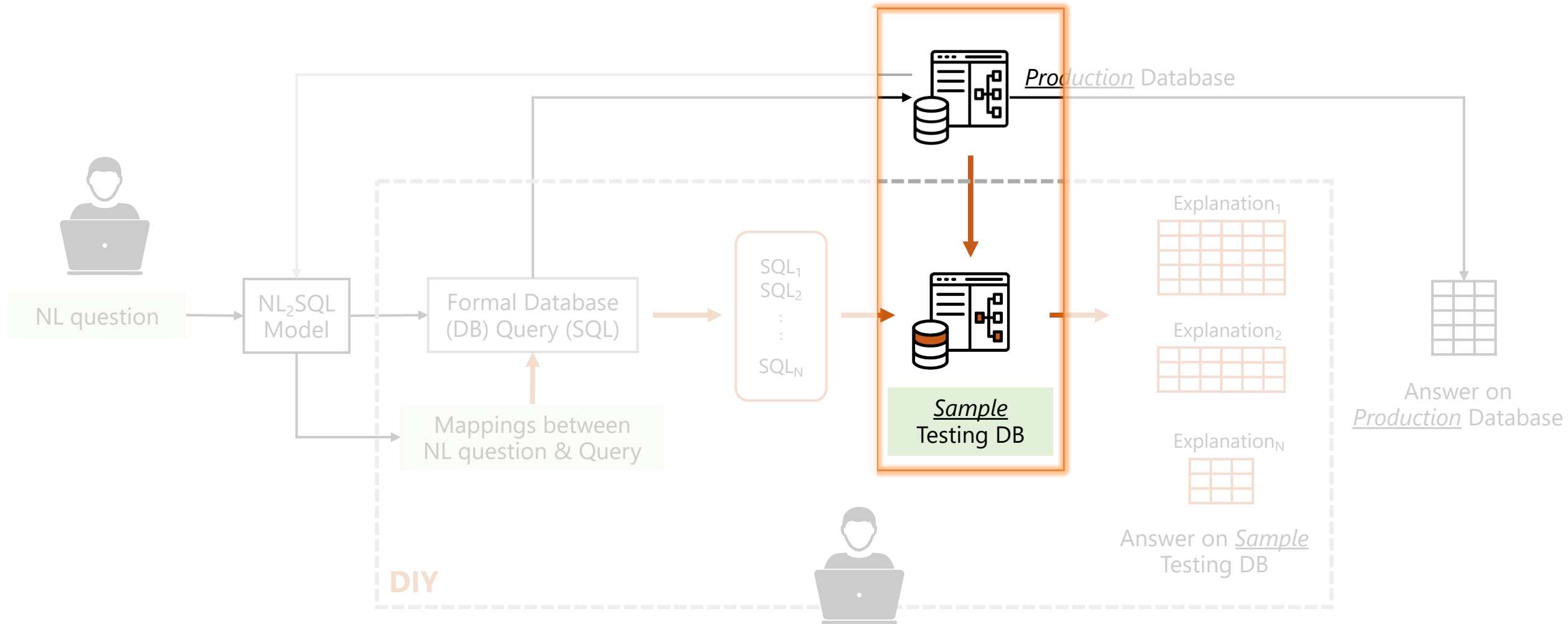
DIY leverages a **human-in-the-loop approach** enabling users to modify sample database to:

- create a more appropriate data sample, or
- debug using different what-if scenarios

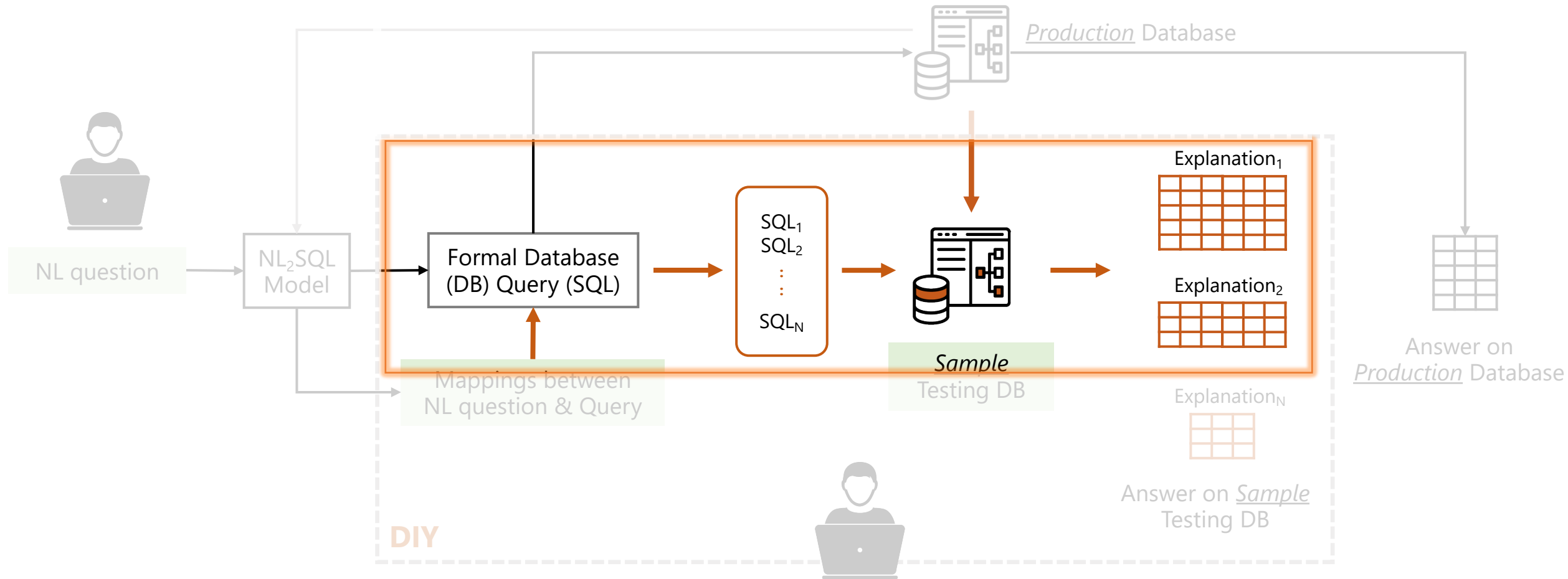
cars_data		
Accelerate x	Year ▲ x	+ x
12	1970	🗑️
11	1972	🗑️
12	1972	🗑️
10.5	1972	🗑️
11.5	1973	🗑️
📊 x	📊 x	+ 🔄

Interactive Tabular Widget

DIY: Sample Testing DB



DIY: Explaining SQL



Explaining SQL: FROM, JOIN

SQL Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
JOIN car_names
    ON model_list.Model
        = car_names.Model
JOIN cars_data
    ON car_names.Id
        = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

Explaining SQL: FROM, JOIN

DIY

SQL

SQL Generated Query

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

```
SELECT
  *
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
```

DIY

Explainer Steps

- 1
- 2
- 3
- 4

For each record in **model_list**, choose each corresponding record in **car_names** where **model_list.Model** equals **car_names.Model**.

For each record in **car_names**, choose each corresponding record in **cars_data** where **car_names.Id** equals **cars_data.Id**.

DIY



Sample Testing DB

model_list		car_names		cars_data	
Model		Id	Model	Id	Year
buick		306	cadillac	306	1979
chevrolet		307	peugeot	307	1979
honda		345	honda	345	1980
peugeot		346	plymouth	346	1981
plymouth		347	buick	347	1982

model_list		car_names		cars_data	
Model		Id	Model	Id	Year
peugeot		307	peugeot	307	1979
honda		345	honda	345	1980
plymouth		346	plymouth	346	1981
buick		347	buick	347	1981

Explaining SQL: FROM, JOIN

DIY Explainer Steps

- 1
- 2
- 3
- 4

Natural Language Explanations

For each record in **model_list**, choose each corresponding record in **car_names** where **model_list.Model** equals **car_names.Model**.

For each record in **car_names**, choose each corresponding record in **cars_data** where **car_names.Id** equals **cars_data.Id**.

Tabular Visualization

model_list		car_names		cars_data	
Model	Id	Model	Id	Year	
peugeot	307	peugeot	307	1979	
honda	345	honda	345	1980	
plymouth	346	plymouth	346	1981	
buick	347	buick	347	1981	

Explaining SQL: WHERE

Natural Language Explanations

Tabular Visualization

DIY Explainer Steps

- 1
- 2
- 3
- 4

For each record in **model_list**, choose each corresponding record in **car_names** where **model_list.Model** equals **car_names.Model**.

For each record in **car_names**, choose each corresponding record in **cars_data** where **car_names.Id** equals **cars_data.Id**.

model_list	car_names		cars_data	
Model	Id	Model	Id	Year
peugeot	307	peugeot	307	1979
honda	345	honda	345	1980
plymouth	346	plymouth	346	1981
buick	347	buick	347	1981

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

- 1
- 2
- 3
- 4

Keep those records whose **cars_data.Year** is greater than 1980.

model_list	car_names		cars_data	
Model	Id	Model	Id	Year
peugeot	307	peugeot	307	1979
honda	345	honda	345	1980
plymouth	346	plymouth	346	1981
buick	347	buick	347	1981

Explaining SQL: SELECT

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

DIY Explainer Steps



Large Explanations

For each record in **model_list**, choose each corresponding record in **car_names** where **model_list.Model** equals **car_names.Model**.

For each record in **car_names**, choose each corresponding record in **cars_data** where **car_names.Id** equals **cars_data.Id**.

model_list		car_names		cars_data	
Model	Id	Model	Id	Year	
peugeot	307	peugeot	307	1979	
honda	345	honda	345	1980	
plymouth	346	plymouth	346	1981	
buick	347	buick	347	1981	

Keep those records whose **cars_data.Year** is greater than 1980.

model_list		car_names		cars_data	
Model	Id	Model	Id	Year	
peugeot	307	peugeot	307	1979	
honda	345	honda	345	1980	
plymouth	346	plymouth	346	1981	
buick	347	buick	347	1981	

Choose the **model_list.Model**.

model_list
Model
plymouth
buick

Table Visualization

Explaining SQL: DISTINCT

```
SELECT DISTINCT
  model_list.Model
FROM
  model_list
  JOIN car_names
    ON model_list.Model
      = car_names.Model
  JOIN cars_data
    ON car_names.Id
      = cars_data.Id
WHERE
  cars_data.Year > "1980"
```

DIY Explainer Steps

1

2

3

4

For each record in **model_list**, choose each corresponding record in **car_names** where **model_list.Model** equals **car_names.Model**.

For each record in **car_names**, choose each corresponding record in **cars_data** where **car_names.Id** equals **cars_data.Id**.

model_list		car_names		cars_data	
Model	Id	Model	Id	Year	
peugeot	307	peugeot	307	1979	
honda	345	honda	345	1980	
plymouth	346	plymouth	346	1981	
buick	347	buick	347	1981	

1

2

3

4

Keep those records whose **cars_data.Year** is greater than 1980.

model_list		car_names		cars_data	
Model	Id	Model	Id	Year	
peugeot	307	peugeot	307	1979	
honda	345	honda	345	1980	
plymouth	346	plymouth	346	1981	
buick	347	buick	347	1981	

1

2

3

4

Choose the **model_list.Model**.

model_list
Model
plymouth
buick

1

2

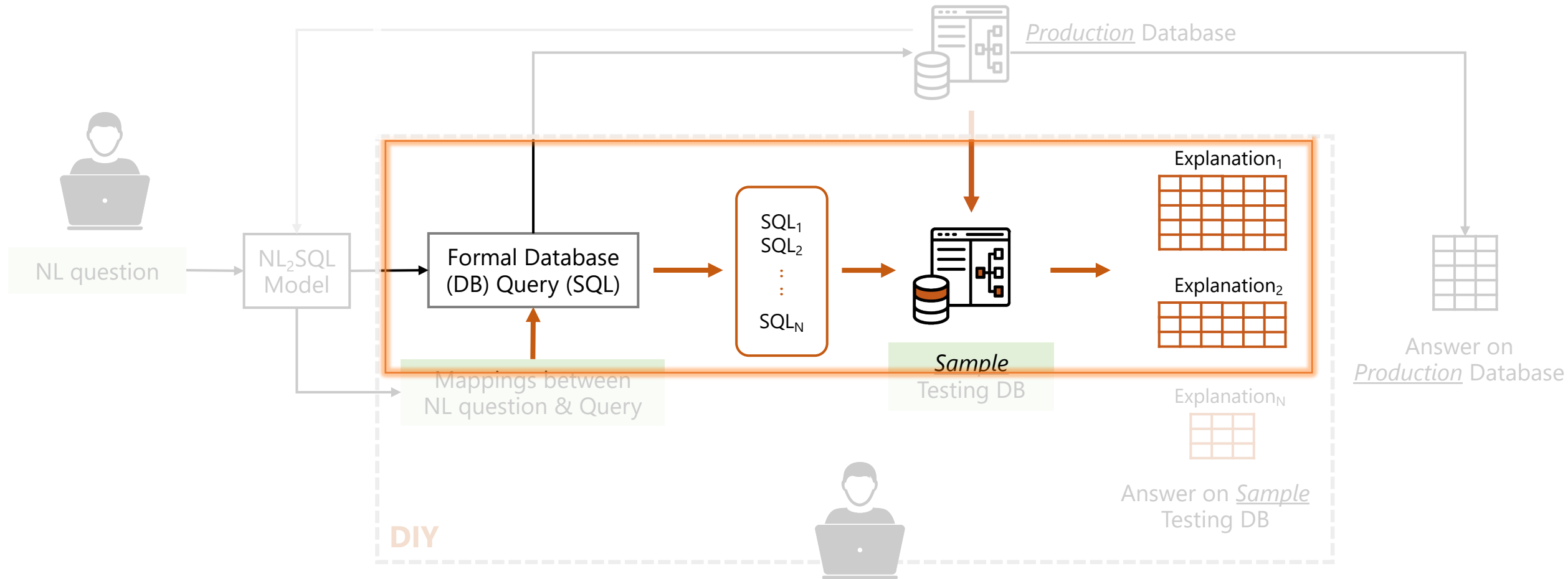
3

4

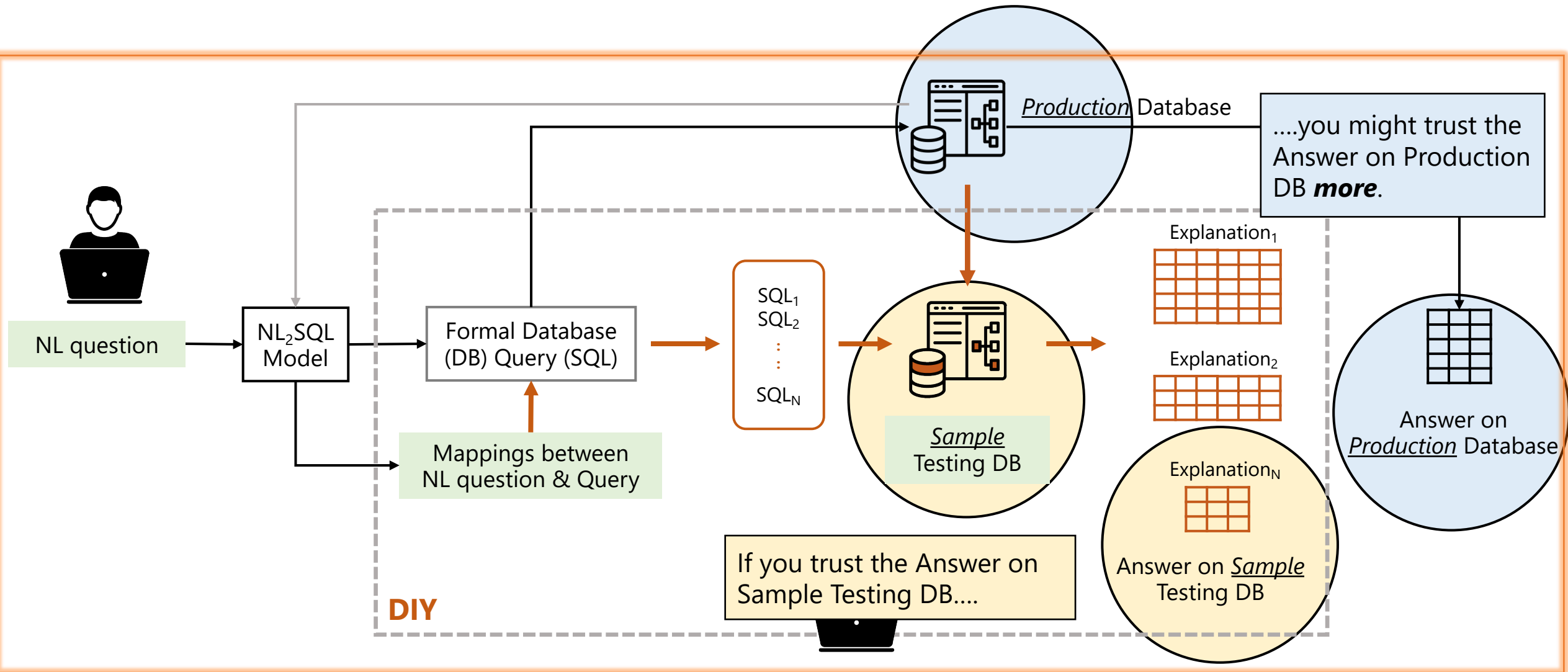
Keep unique records.

model_list
Model
plymouth
buick

DIY: Explaining SQL



DIY: Debug It Yourself



Exploratory User Study

- Understand how DIY helps users to **inspect for, isolate, and fix errors** in NL2SQL scenarios.
- **12** users: UX Designers, Design Researchers, Data Scientists, Program Managers, Cloud Architects, etc.
- **~90** min sessions on Teams; remote due to Covid-19
- **8** tasks categorized by
 - Complexity (3 easy, 2 medium, 3 hard questions)[†]
 - Correctness (4 questions with errors)

[†] Yu, Tao, et al. "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task." arXiv preprint, 2018.

What did we find?

- **Benefits**
 - *"It is important to have this transparency...*
 - *to show people how the system is working...*
 - *and to let them control it. This [DIY] is a great example of that." – P08*
- **Pressure points**
 - Confusion between Production Database and Sample Testing Database.
 - Manipulating Data even if it is a testing sample does not seem right w.r.t. Ethics, Integrity.
- **Debugging Rationales**
 - To double-check the answer and/or strategy
 - There was not enough domain expertise to trust the answer
 - There were one or more errors detected
 - The capability existed.
- **Debugging Strategies**

Debugging Strategies

Edit a record

Consider the following sample data table.

cars_data		
Accelerate	Year	+
12	1970	🗑️
11.5	1970	🗑️
11	1970	🗑️
12	1970	🗑️
10.5	1970	🗑️
📄 ✂️	📄 ✂️	+ 🔗

Group records with the same **cars_data.Year** together.

cars_data	
Year	Accelerate
— 1970 (5 records)	(12, 11.5, 11, 12, 10.5)
	12
	11.50
	11
	12
	10.50

Choose the **average** of **cars_data.Accelerate**, **cars_data.Year**.

cars_data	
Avg(Accelerate)	Year
11.40	1970

Delete a record

Consider the following sample data table.

cars_data		
Accelerate	Year	+
12	1970	🗑️
11.5	1970	🗑️
11	1970	🗑️
12	1970	🗑️
10.5	1970	🗑️
📄 ✂️	📄 ✂️	+ 🔗

Group records with the same **cars_data.Year** together.

cars_data	
Year	Accelerate
— 1970 (5 records)	(12, 11.5, 11, 12, 10.5)
	12
	11.50
	11
	12
	10.50

Choose the **average** of **cars_data.Accelerate**, **cars_data.Year**.

cars_data	
Avg(Accelerate)	Year
11.40	1970

Randomize all column values

Consider the following sample data table.

cars_data		
Accelerate	Year	+
17.4	1970	🗑️
24.8	1970	🗑️
17	1970	🗑️
15.7	1970	🗑️
16.4	1970	🗑️
📄 ✂️	📄 ✂️	+ 🔗

Group records with the same **cars_data.Year** together.

cars_data	
Year	Accelerate
+ 1970 (5 records)	(17.4, 24.8, 17, 15

Choose the **average** of **cars_data.Accelerate**, **cars_data.Year**.

cars_data	
Avg(Accelerate)	Year
18.26	1970

Add a new record

Consider the following sample data table.

cars_data		
Accelerate	Year	+
12	1970	🗑️
11.5	1970	🗑️
11	1970	🗑️
12	1970	🗑️
10.5	1970	🗑️
📄 ✂️	📄 ✂️	+ 🔗

Group records with the same **cars_data.Year** together.

cars_data	
Year	Accelerate
— 1970 (5 records)	(12, 11.5, 11, 12, 10.5)
	12
	11.50
	11
	12
	10.50

Choose the **average** of **cars_data.Accelerate**, **cars_data.Year**.

cars_data	
Avg(Accelerate)	Year
11.40	1970

Debugging Strategies II

Modify Entity Mappings

Entities Detected in the Question

What is the **average** **acceleration** of cars each **year** ?

average of cars_data.Accel... cars_data.Year

Sample Data View

0 >>

Consider the following sample data table.

cars_data		
Accelerate	Year	+
12	1970	🗑
11.5	1970	🗑
11	1970	🗑
12	1970	🗑
10.5	1970	🗑
📄 ✂	📄 ✂	+ ↻

Steps

1

Group records with the same cars_data.Year together.

cars_data	
Year	Accelerate
— 1970 (5 records)	(12,11.5,11,12,10.5)
	12
	11.50
	11
	12
	10.50

Answer on the Sample Data

>> 2

Choose the **average** of cars_data.Accelerate, cars_data.Year.

cars_data	
Avg(Accelerate)	Year
11.40	1970

Opportunities for Future

Facilitate Data Exploration

- Manipulating mappings instead of paraphrasing to query new information (e.g., change "**greater than**" to "**less than**" to find values along the other dimension).

Handle Ambiguities between Conversational and Formal Language

- "**since** 1980" can be perceived as "**greater than** 1980" or "**greater than or equal to** 1980".

Learn and Teach SQL

- As a Natural Language plugin for online tools like **SQL Try it Editor**[‡].

[‡] https://www.w3schools.com/sql/trysql.asp?filename=trysql_op_in

DIY: Helping People Assess the Correctness of Natural Language to SQL Systems



Arpit Narechania
arpitnarechania@gatech.edu
 @arpitnarechania



Adam Fourney
adamfo@microsoft.com
 @adamfourney



Bongshin Lee
bongshin@microsoft.com
 @bongshin



Gonzalo Ramos
goramos@microsoft.com
 @GonzaloWorks

PUBLICATION
bit.ly/DIY-pub

TALK SLIDES
bit.ly/DIY-slides

Entities Detected in the Question

What is the **average** **acceleration** of cars each **year** ?

average of cars_data.Accelerate cars_data.Year

Sample Data View

0 »

Consider the following sample data table.

cars_data		
Accelerate	Year	+
17.4	1979	🗑️
24.8	1979	🗑️
17	1980	🗑️
15.7	1981	🗑️
16.4	1981	🗑️
📄 🔍		📄 🔍 + ↻

Steps

1

Group records with the same **cars_data.Year** together.

cars_data	
Year	Accelerate
— 1979 (2 records)	(17.4, 24.8)
	17.40
	24.80
+ 1980 (1 record)	(17)
+ 1981 (2 records)	(15.7, 16.4)

Answer on the Sample Data

» 2

Choose the **average** of **cars_data.Accelerate**, **cars_data.Year**.

cars_data	
Avg(Accelerate)	Year
21.10	1979
17	1980
16.05	1981

