ONLINE , AROUND-THE-WORLD OPEN SOURCE DATABASE CONFERENCE

PERCONA

LIVEONLINE

MAY 12 - 13th

2021

> whoami

Jorge Torres

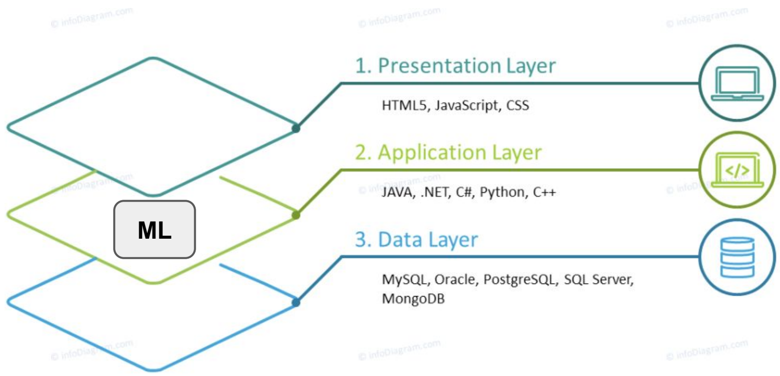
CEO@MindsDB

@torrmal

Machine Learning
Where does it belong best?

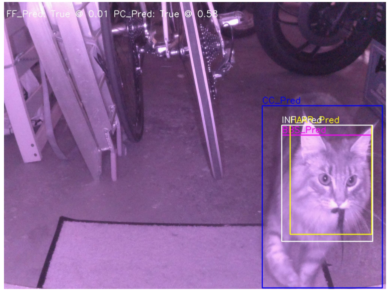
Application layer ML

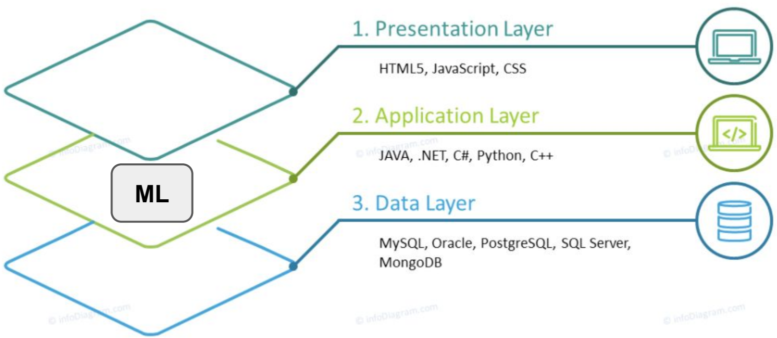
When applying ML, does it have to live in the application stack?



Application layer ML

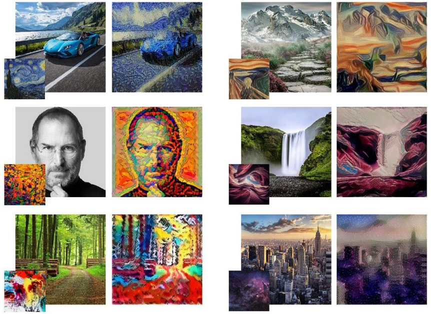
When applying ML, does it have to live in the application stack?

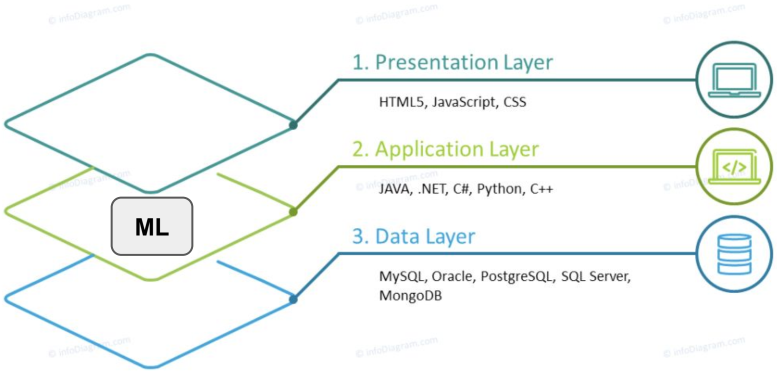




Application layer ML

When applying ML, does it have to live in the application stack?





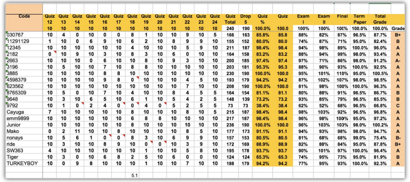
Machine Learning
In the data layer

mindsdb

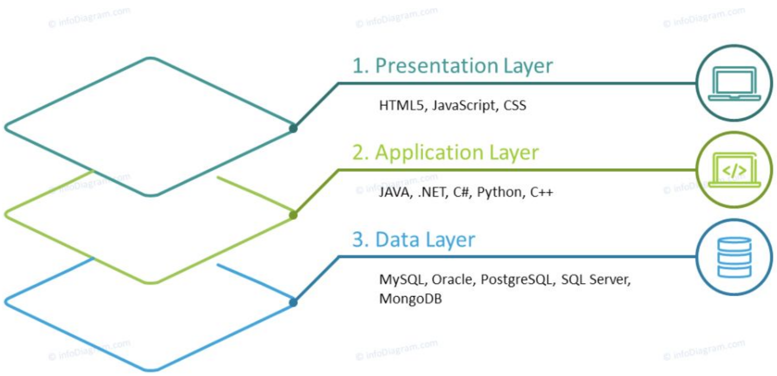
PERCONA
LIVEONLINE

Data layer ML

When applying ML, can it live in the data layer?

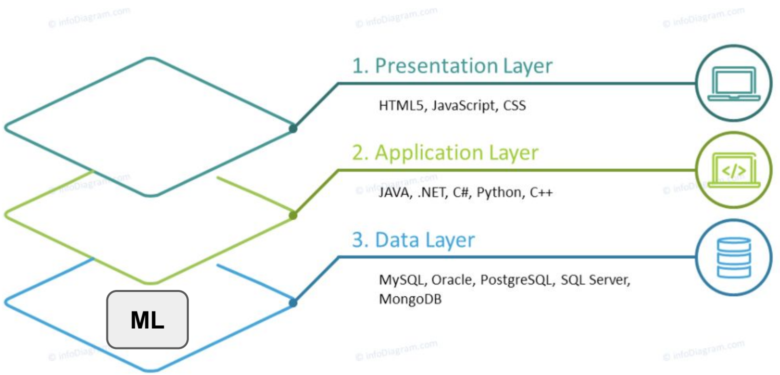


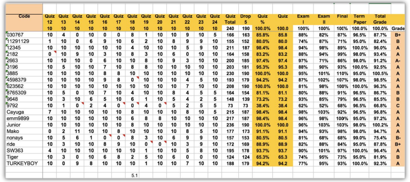
Code	Quiz 12	Quiz 13	Quiz 14	Quiz 15	Quiz 16	Quiz 17	Quiz 18	Quiz 19	Quiz 20	Quiz 21	Quiz 22	Quiz 23	Quiz 24	Quiz Total	Drop 5	Quiz %	Quiz	Exam I	Exam II	Final	Term Paper	Total Grade	
	10	10	10	10	10	10	10	10	10	10	10	10	10	240	190	100.0%	100.0	100%	100%	100%	100.0%	100.0%	Grade
030767	10	4	0	10	0	0	8	1	10	10	9	10	5	166	163	85.8%	85.8	88%	82%	87%	96.5%	87.7%	B+
11291129	1	10	0	6	9	10	6	2	10	8	9	9	10	155	152	80.0%	80.0	74%	87%	71%	95.0%	82.4%	B
12345	10	10	10	10	10	10	4	10	10	10	5	9	10	211	187	98.4%	98.4	94%	98%	89%	100.0%	96.0%	A
2182	0	10	9	10	3	10	8	3	10	6	0	10	10	164	158	83.2%	83.2	88%	94%	99%	98.0%	93.4%	A
2663	10	10	10	0	6	10	10	8	10	9	3	10	10	200	185	97.4%	97.4	97%	71%	86%	98.0%	91.2%	A-
3196	10	5	10	10	7	10	8	8	10	10	10	10	10	203	181	95.3%	95.3	88%	90%	93%	100.0%	92.5%	A
3885	10	10	10	10	8	8	10	10	10	10	10	10	10	230	190	100.0%	100.0	95%	101%	110%	95.0%	100.5%	A
4598379	10	10	10	10	9	8	0	10	10	10	4	5	10	193	179	94.2%	94.2	87%	102%	107%	98.0%	98.5%	A
623562	10	10	10	10	10	10	10	10	10	10	7	10	10	208	190	100.0%	100.0	81%	98%	100%	100.0%	96.3%	A
8765309	10	5	0	10	7	10	4	10	10	8	4	5	5	164	154	81.1%	81.1	80%	88%	91%	96.5%	86.7%	B
9648	10	3	10	6	5	10	6	1	10	5	4	2	5	148	139	73.2%	73.2	93%	85%	79%	96.5%	85.5%	B
9792	10	1	0	2	4	0	0	4	0	5	2	5	5	73	73	38.4%	38.4	62%	52%	68%	96.5%	66.8%	C
Cayuga	7	10	10	10	10	10	8	10	10	10	8	10	5	215	187	98.4%	98.4	98%	103%	92%	96.5%	96.8%	A
emm9899	10	10	10	10	10	10	6	8	8	10	10	10	10	217	187	98.4%	98.4	96%	98%	100%	95.0%	97.2%	A
Junior	10	10	10	10	10	10	8	10	10	10	10	10	10	236	190	100.0%	100.0	96%	103%	103%	98.0%	100.2%	A
Mako	0	2	11	10	10	8	10	10	10	10	8	5	10	177	173	91.1%	91.1	94%	93%	98%	98.0%	94.7%	A
nonays	10	5	6	1	0	0	8	3	10	6	9	9	10	157	153	80.5%	80.5	81%	68%	68%	99.0%	75.4%	B-
ride	10	3	10	10	8	9	10	0	0	10	3	9	10	172	169	88.9%	88.9	82%	88%	84%	95.0%	87.8%	B+
SW363	4	10	10	10	10	10	10	1	10	10	5	8	10	195	178	93.7%	93.7	90%	101%	97%	100.0%	96.4%	A
Tiger	10	3	0	10	6	8	2	5	10	6	0	0	10	124	124	65.3%	65.3	74%	95%	73%	95.0%	81.9%	B
TURKEYBOY	10	0	9	8	10	10	10	1	10	10	7	10	10	188	179	94.2%	94.2	77%	95%	93%	100.0%	92.3%	A
							5.1																



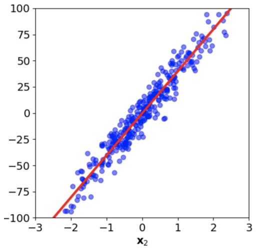
Data layer ML

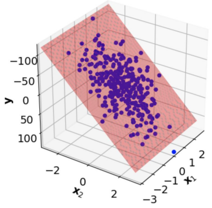
When applying ML, can it live in the data layer?



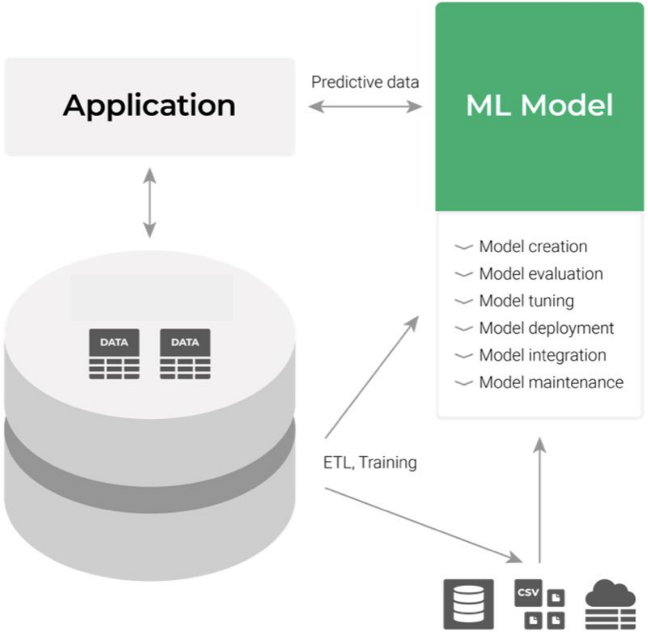


Code	Quiz 12	Quiz 13	Quiz 14	Quiz 15	Quiz 16	Quiz 17	Quiz 18	Quiz 19	Quiz 20	Quiz 21	Quiz 22	Quiz 23	Quiz 24	Quiz Total	Drop 5	Quiz %	Quiz	Exam I	Exam II	Final	Term Paper	Total Grade	
	10	10	10	10	10	10	10	10	10	10	10	10	10	240	190	100.0%	100.0	100%	100%	100%	100.0%	100.0%	Grade
030767	10	4	0	10	0	0	8	1	10	10	9	10	5	166	163	85.8%	85.8	88%	82%	87%	96.5%	87.7%	B+
11291129	1	10	0	6	9	10	6	2	10	8	9	9	10	155	152	80.0%	80.0	74%	87%	71%	95.0%	82.4%	B
12345	10	10	10	10	10	10	4	10	10	10	5	9	10	211	187	98.4%	98.4	94%	98%	89%	100.0%	96.0%	A
2182	0	10	9	10	3	10	8	3	10	6	0	10	10	164	158	83.2%	83.2	88%	94%	99%	98.0%	93.4%	A
2663	10	10	10	0	6	10	10	8	10	9	3	10	10	200	185	97.4%	97.4	97%	71%	86%	98.0%	91.2%	A-
3196	10	5	10	10	7	10	8	8	10	10	10	10	10	203	181	95.3%	95.3	88%	90%	93%	100.0%	92.5%	A
3885	10	10	10	10	8	8	10	10	10	10	10	10	10	230	190	100.0%	100.0	95%	101%	110%	95.0%	100.5%	A
4598379	10	10	10	10	9	8	0	10	10	10	4	5	10	193	179	94.2%	94.2	87%	102%	107%	98.0%	98.5%	A
623562	10	10	10	10	10	10	10	10	10	10	7	10	10	208	190	100.0%	100.0	81%	98%	100%	100.0%	96.3%	A
8765309	10	5	0	10	7	10	4	10	10	8	4	5	5	164	154	81.1%	81.1	80%	88%	91%	96.5%	86.7%	B
9648	10	3	10	6	5	10	6	1	10	5	4	2	5	148	139	73.2%	73.2	93%	85%	79%	96.5%	85.5%	B
9792	10	1	0	2	4	0	0	4	0	5	2	5	5	73	73	38.4%	38.4	62%	52%	68%	96.5%	66.8%	C
Cayuga	7	10	10	10	10	10	8	10	10	10	8	10	5	215	187	98.4%	98.4	98%	103%	92%	96.5%	96.8%	A
emm9899	10	10	10	10	10	10	6	8	8	10	10	10	10	217	187	98.4%	98.4	96%	98%	100%	95.0%	97.2%	A
Junior	10	10	10	10	10	10	8	10	10	10	10	10	10	236	190	100.0%	100.0	96%	103%	103%	98.0%	100.2%	A
Mako	0	2	11	10	10	8	10	10	10	10	8	5	10	177	173	91.1%	91.1	94%	93%	98%	98.0%	94.7%	A
nonays	10	5	6	1	0	0	8	3	10	6	9	9	10	157	153	80.5%	80.5	81%	68%	68%	99.0%	75.4%	B-
ride	10	3	10	10	8	9	10	0	0	10	3	9	10	172	169	88.9%	88.9	82%	88%	84%	95.0%	87.8%	B+
SW363	4	10	10	10	10	10	10	1	10	10	5	8	10	195	178	93.7%	93.7	90%	101%	97%	100.0%	96.4%	A
Tiger	10	3	0	10	6	8	2	5	10	6	0	0	10	124	124	65.3%	65.3	74%	95%	73%	95.0%	81.9%	B
TURKEYBOY	10	0	9	8	10	10	10	1	10	10	7	10	10	188	179	94.2%	94.2	77%	95%	93%	100.0%	92.3%	A
							5.1																

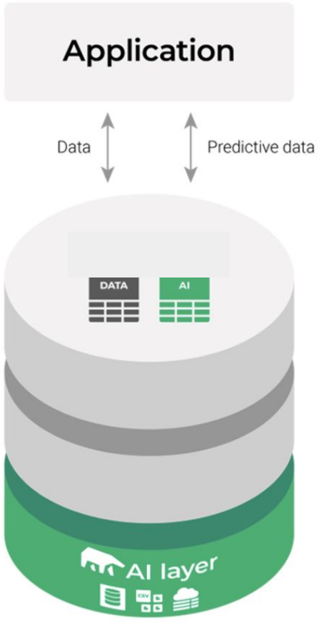




Application layer ML vs Data layer ML



vs



How does it work?

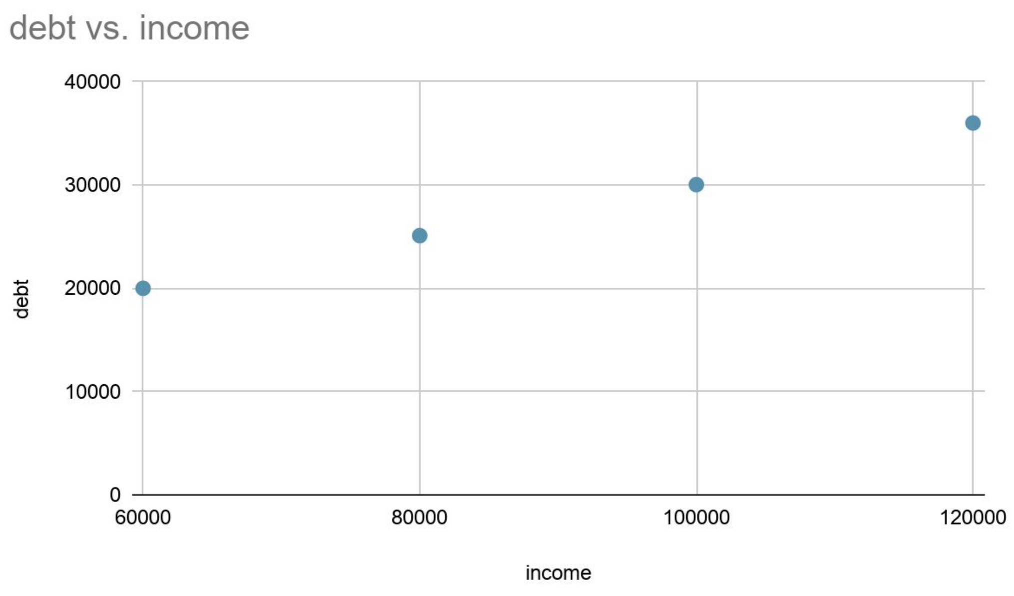
AI Tables

ML Predictive models as Database tables

Simple Example, from DB Tables to ML model

```
SELECT income, debt
FROM income_table
```

income	debt
60000	20000
80000	25100
100000	30040
120000	36010



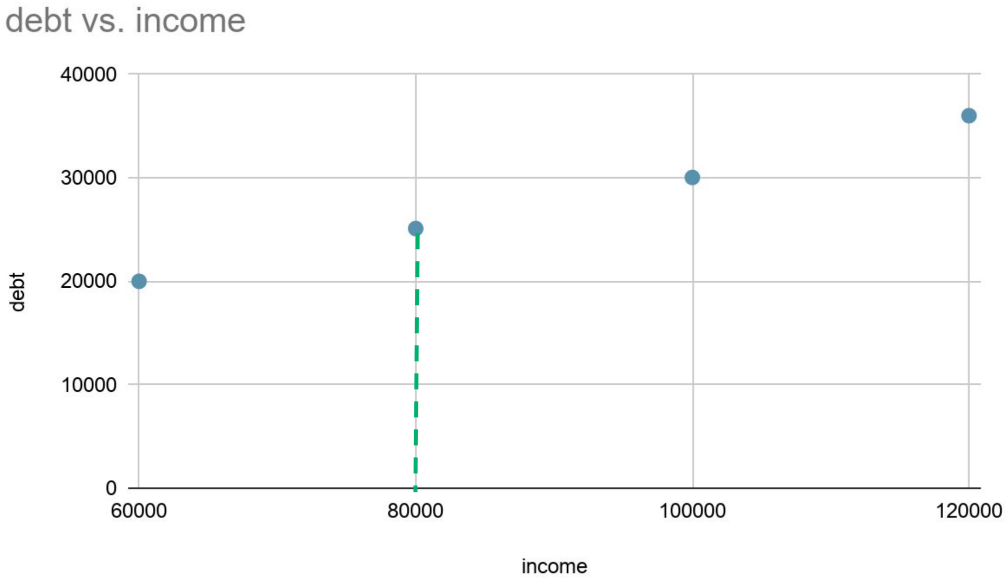
DB Tables and queries

`income_table`

income	debt
60000	20000
80000	25100
100000	30040
120000	36010

```
SELECT income, debt FROM income_table
WHERE income = 80000
```

income	debt
80000	**25100**



You could query the database for information in this table, and if your search criteria has a match, you get results.

mindsdb

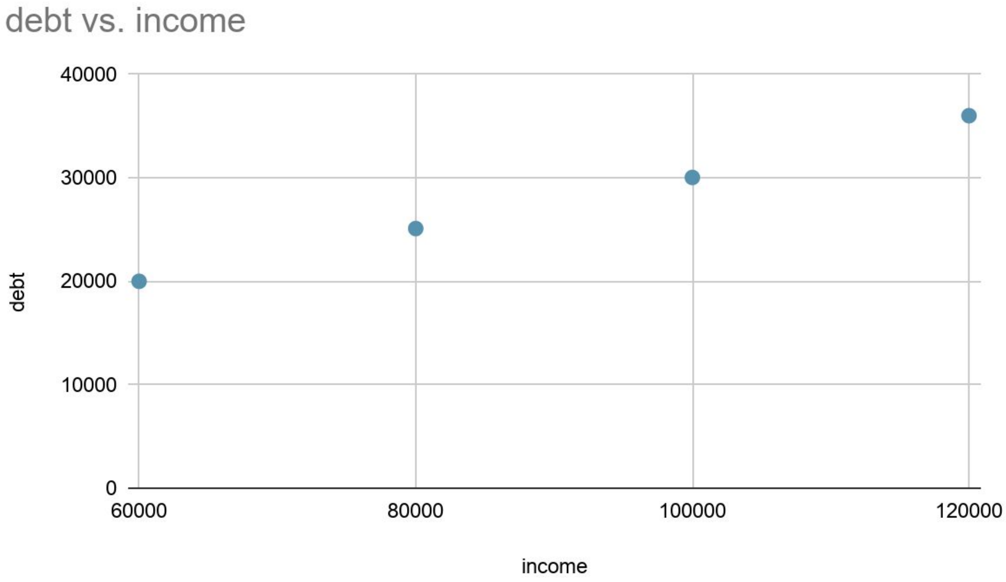
DB queries and mismatches

`income_table`

income	debt
60000	20000
80000	25100
100000	30040
120000	36010

```
SELECT income, debt FROM income_table
WHERE income = 90000
```

income	debt
90000	NULL

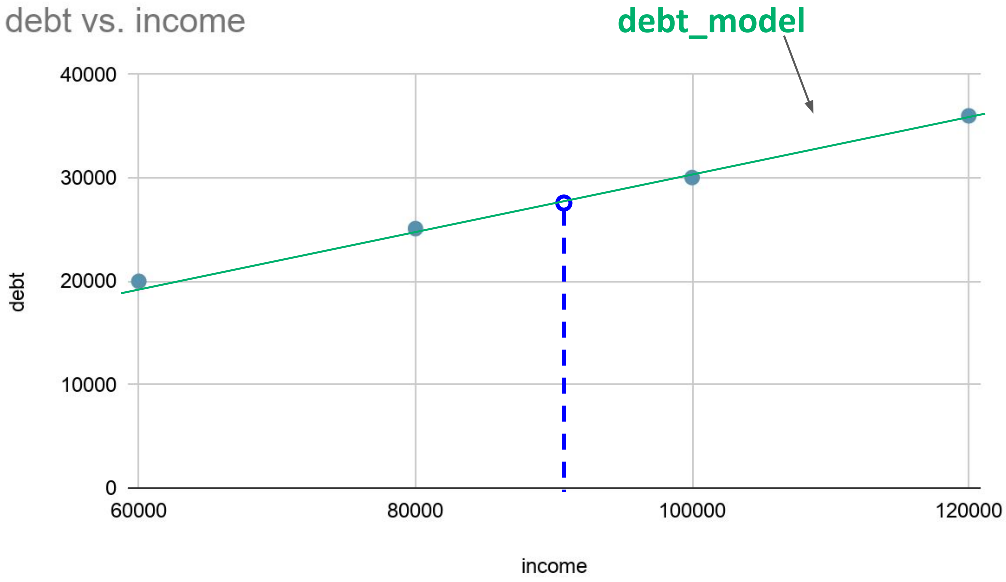


However, if there is no match for your search criteria you get empty results.
Even if your search criteria is close to some data.

Machine Learning as DB Tables

```
SELECT
  income, debt, predicted_debt
FROM
  debt_model
WHERE
  income = 90120
```

income	debt
90120	**28010**



There are hard problems in data layer ML!

Multivariate time series forecasting with MindsDB

> whoami

Patricio Cerda-Mardini

Machine Learning Research Engineer

@paxcema

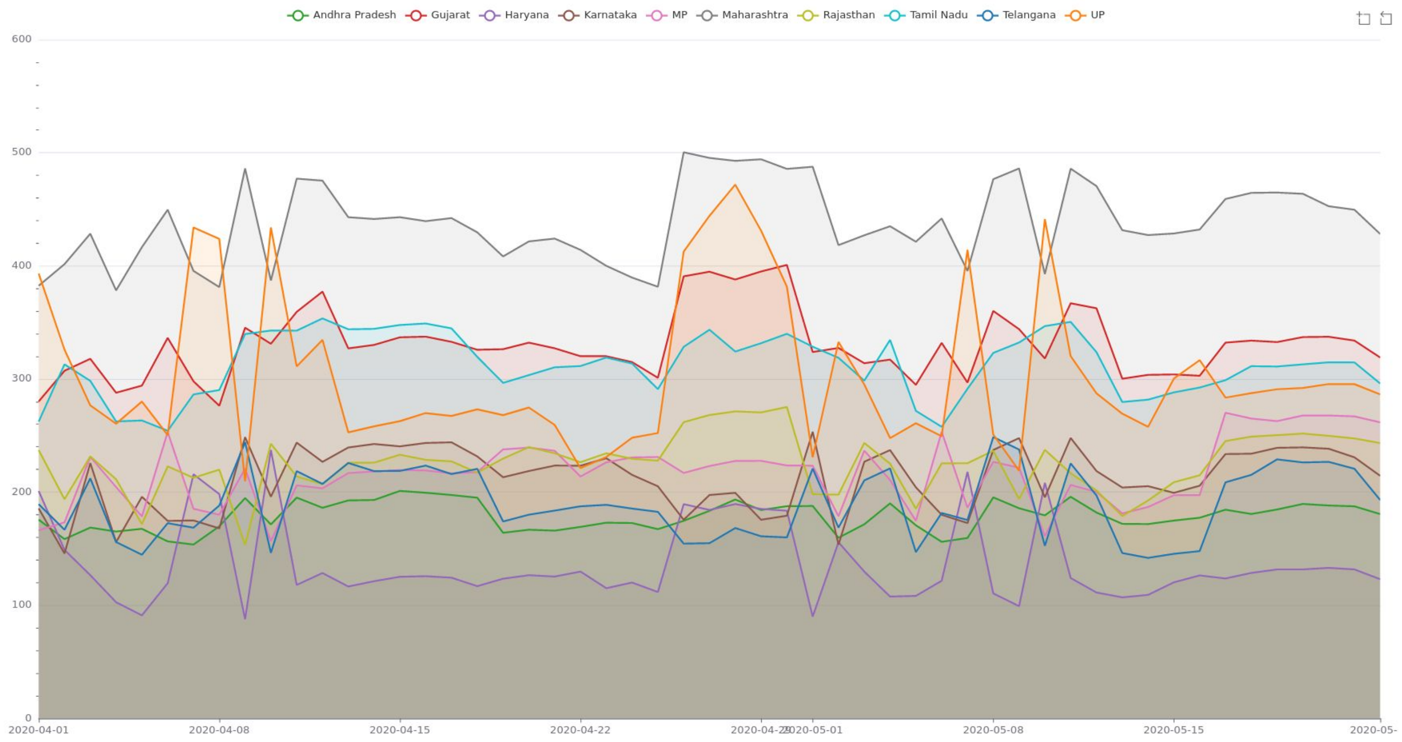
Example

Let’s consider electrical power consumption forecasting

	States	Regions	latitude	longitude	Dates	Usage
0	Punjab	NR	31.519974	75.980003	02/01/2019 00:00:00	119.9
1	Haryana	NR	28.450006	77.019991	02/01/2019 00:00:00	130.3
2	Rajasthan	NR	26.449999	74.639981	02/01/2019 00:00:00	234.1
3	Delhi	NR	28.669993	77.230004	02/01/2019 00:00:00	85.8
4	UP	NR	27.599981	78.050006	02/01/2019 00:00:00	313.9
5	Uttarakhand	NR	30.320409	78.050006	02/01/2019 00:00:00	40.7

Source: kaggle.com/twinkle0705/state-wise-power-consumption-in-india

Example

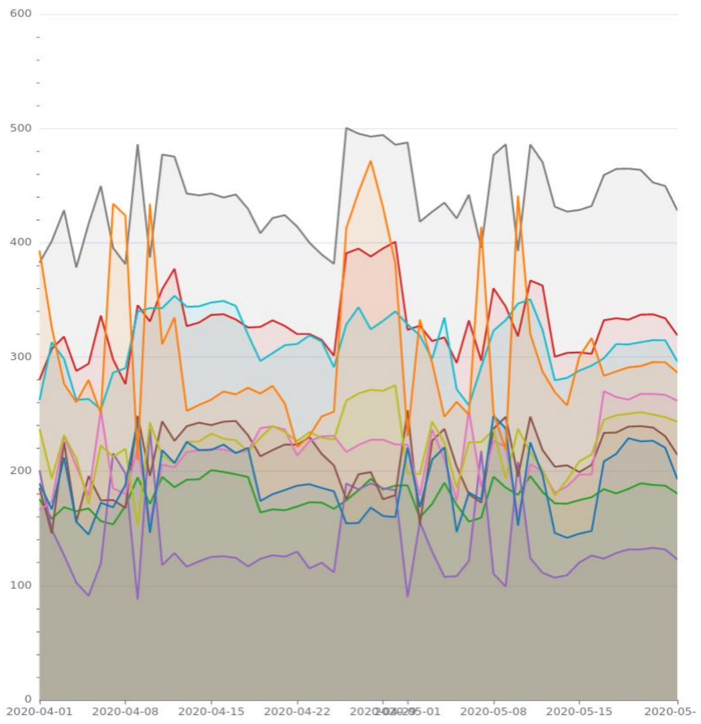


Andhra Pradesh
Gujarat
Haryana
Karnataka
MP
Maharashtra
Rajasthan
Tamil Nadu
Telangana
UP
600
500
400
300
200
100
0
2020-04-01
2020-04-08
2020-04-15
2020-04-22
2020-04-29
2020-05-01
2020-05-08
2020-05-15
2020-05-

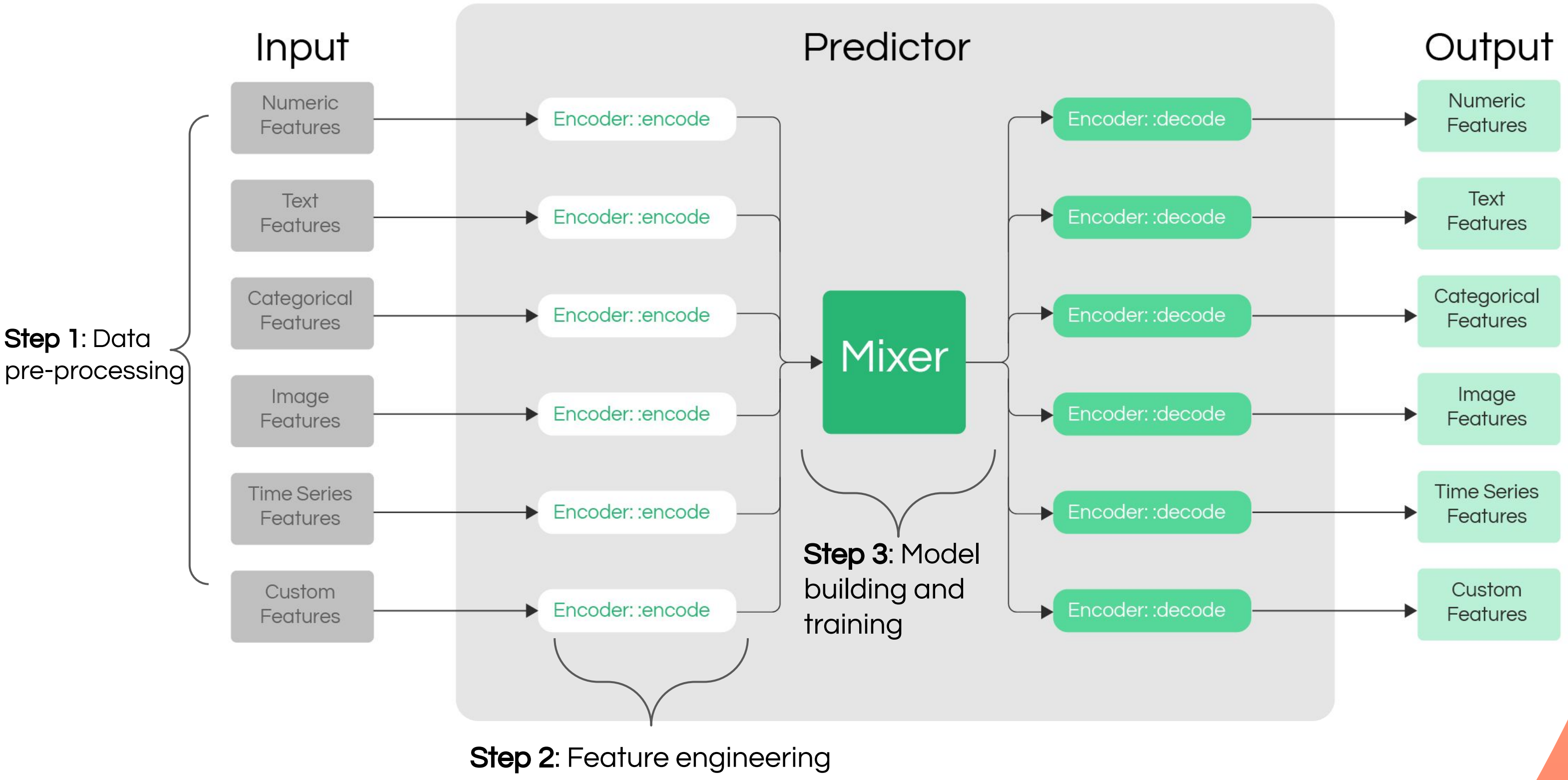
Time series forecasting can be challenging

1. Data pipeline instancing
2. High group cardinality
3. Efficient use of resources

MindsDB automates this process

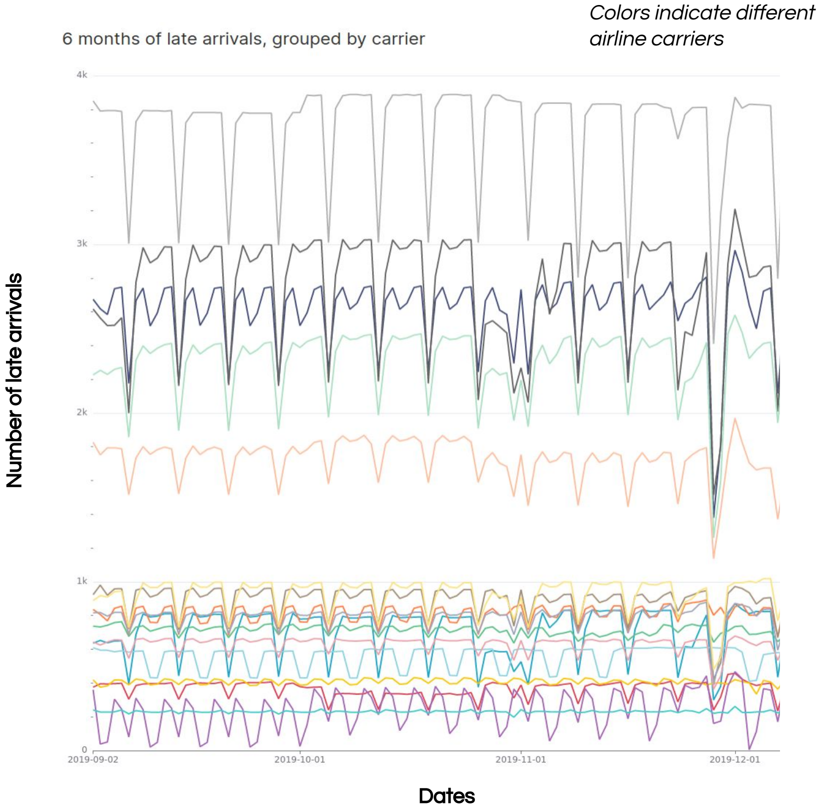


Flexible encoder-mixer philosophy

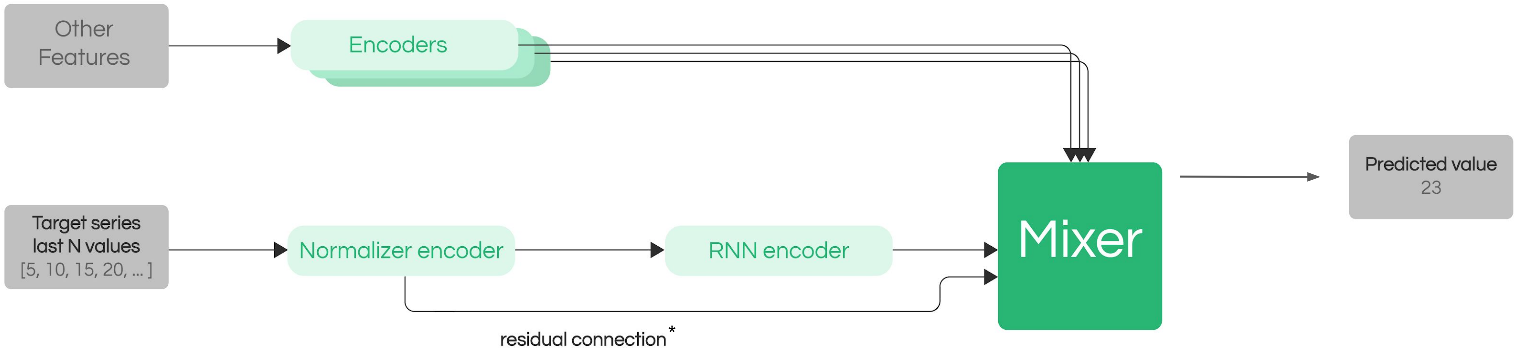


Dynamic Normalization

1. Store statistics for each series (one per group combination)
2. Normalize input using these stats
3. Mixer learns to predict normalized values



Mixers



- Neural network mixer for time series has two streams
  a. learned autoregressive process yields base prediction
  b. secondary stream handles for fine-tuning

- Gradient booster mixer uses LightGBM. MindsDB supports Optuna for stepwise hyperparameter search.

Example - SQL usage

Training:

```
INSERT INTO predictors( name,
                        predict,
                        select_data_query,
                        Training_options )

VALUES ( 'PowerConsumption',
         'Usage',
         'SELECT * FROM training_data;',
         '{"timeseries_settings": {
             "order_by": ["Dates"],
             "group_by": ["States"],
             "window": 10
}}')
```

Example - SQL usage

Querying:

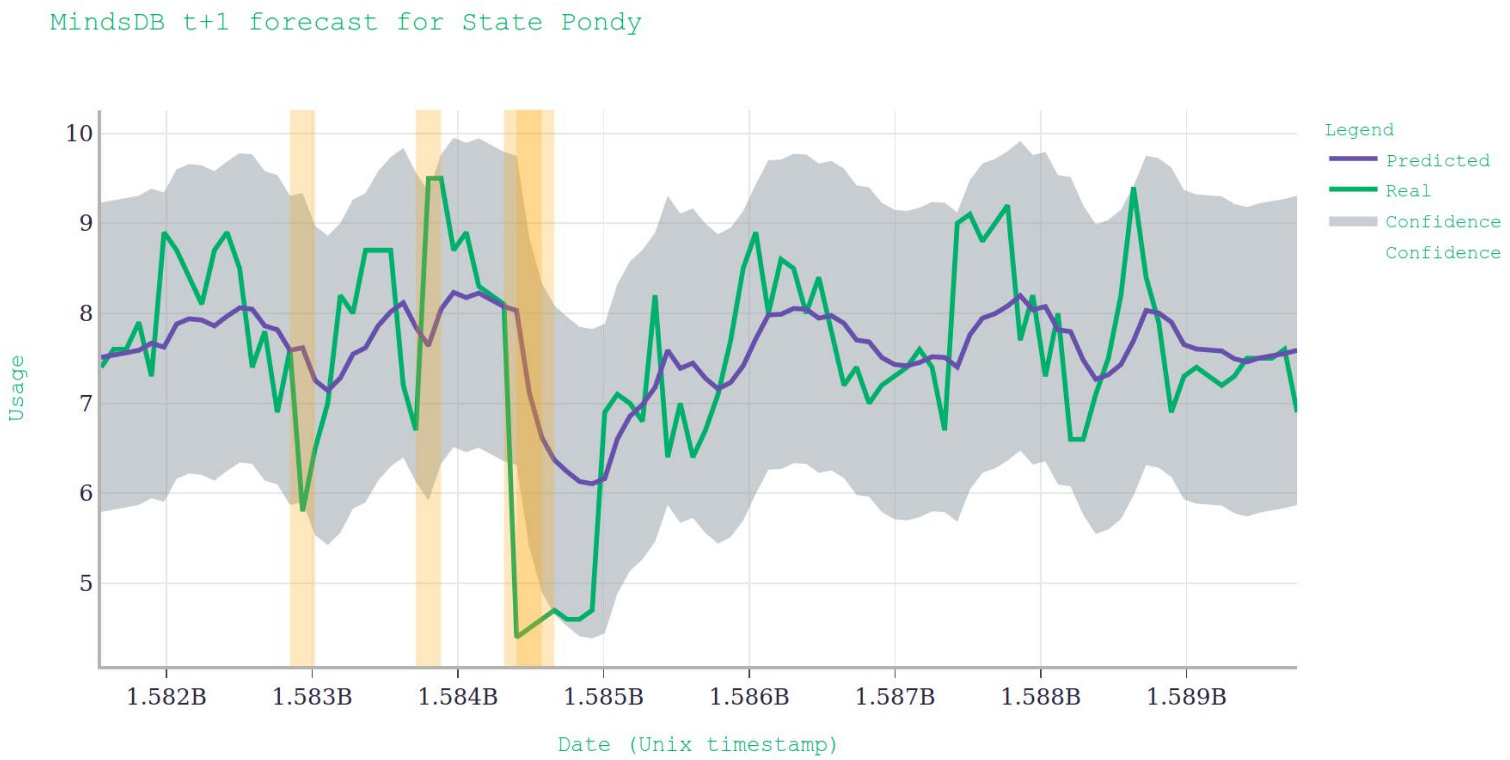
1) Conditional single prediction

```
SELECT Usage FROM mindsdb.PowerConsumption WHERE Dates =
"2021/03/14 12:34:56";
```

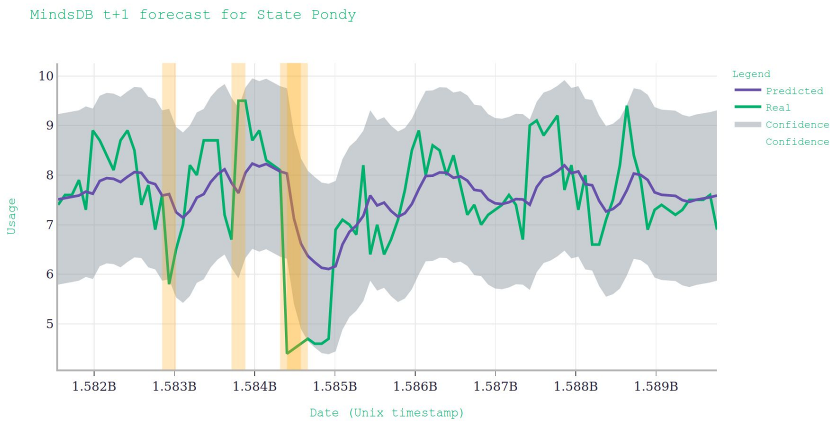
2) Batch prediction

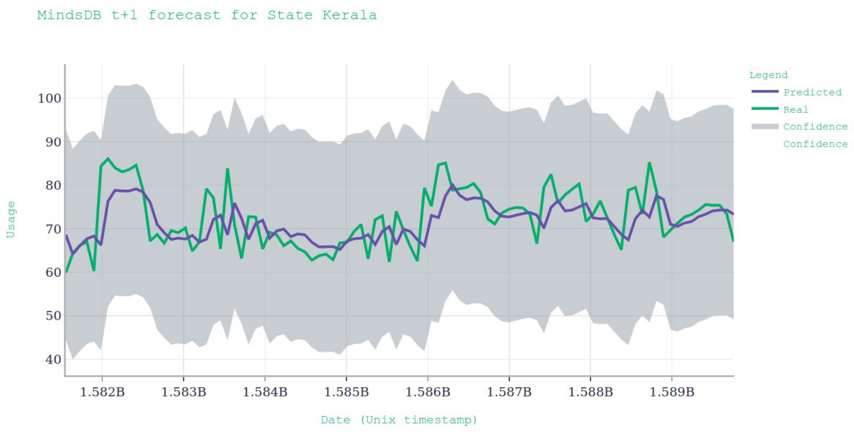
```
SELECT d.Dates, d.Usage as PrevUsage, p.Usage FROM
data.test
AS d LEFT JOIN mindsdb.PowerConsumption AS p ON 1=1;
```

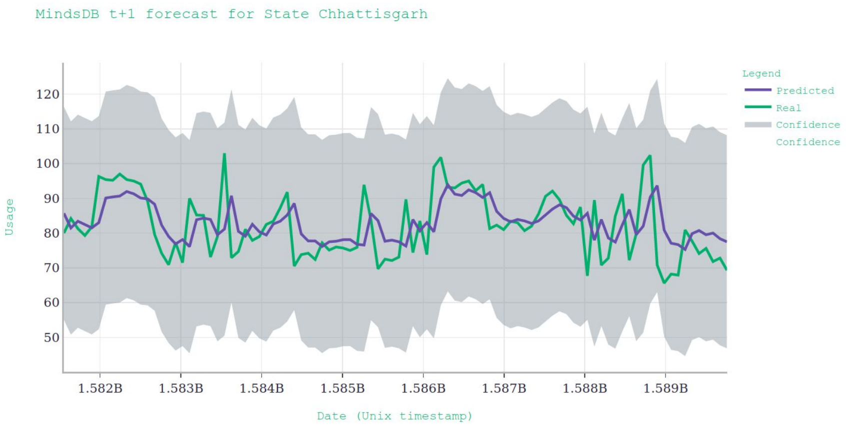
Single group forecasting

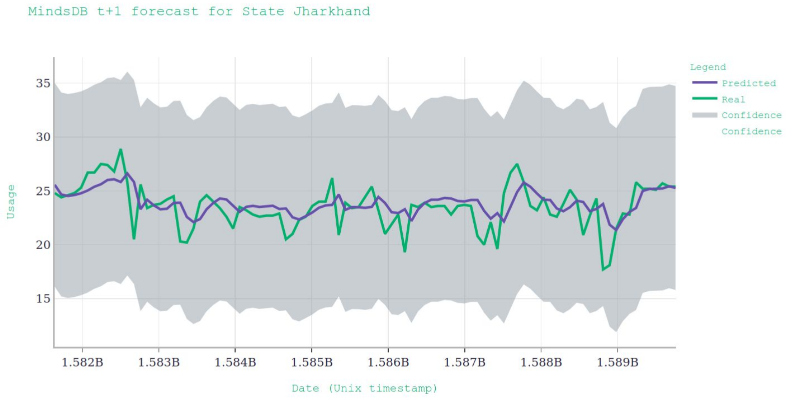


Multivariate forecasting

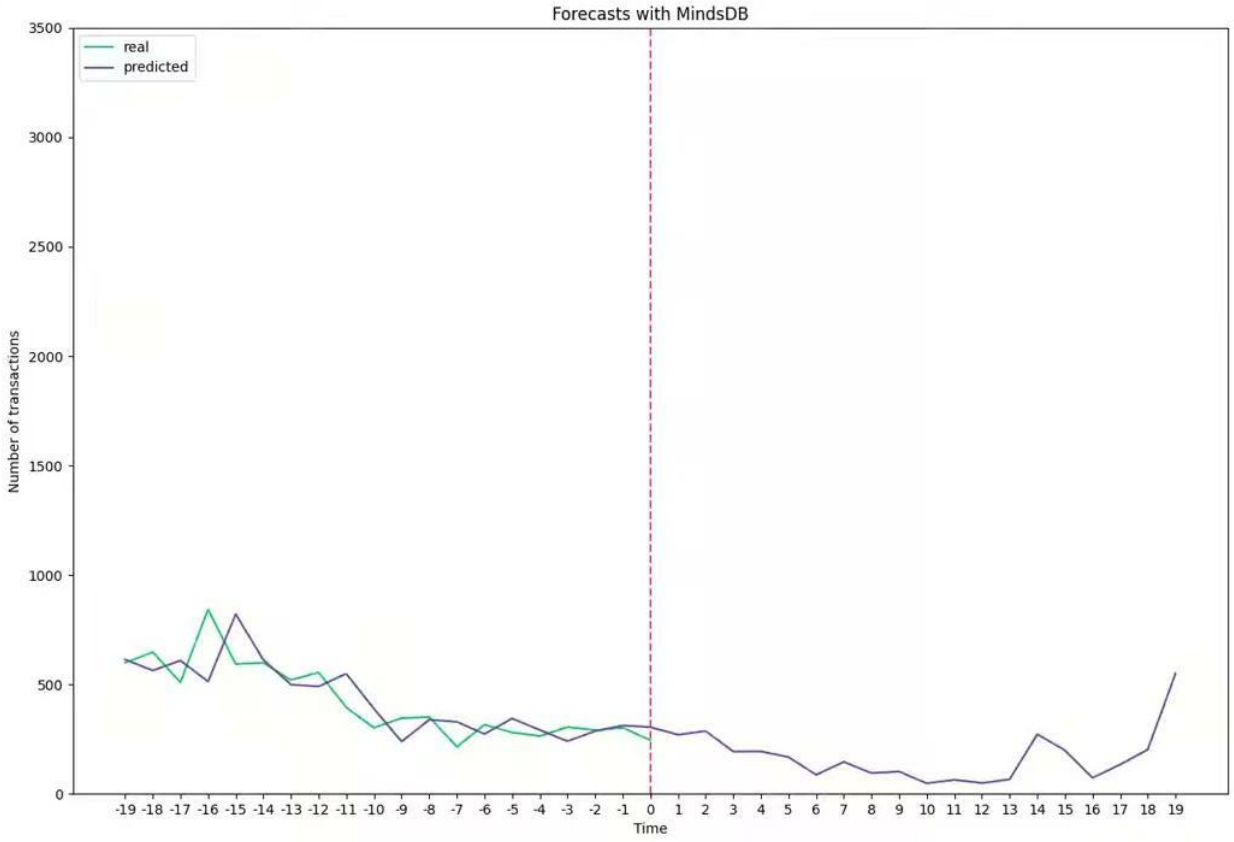








Visualized forecasting



Future Work

- Predicting data streams (e.g. Redis, Kafka)
- Improving forecasts for long horizons with multiple imputation
- Detecting gradual anomalies
- Modin integration

Questions?

You can find us at:

- @torrmal
- @paxcema

/mindsdb/mindsdb

THANK YOU !

ONLINE, AROUND-THE-WORLD OPEN SOURCE DATABASE CONFERENCE

PERCONA
LIVEONLINE
MAY 12 - 13th
2021