Team Samba

Michael Fromm, Christian Lemke, Antoine Maurino, Mihaela Hanea, Sebastian Wagner
Supervisor: Evgeniy Faerman

Final Presentation

Agenda

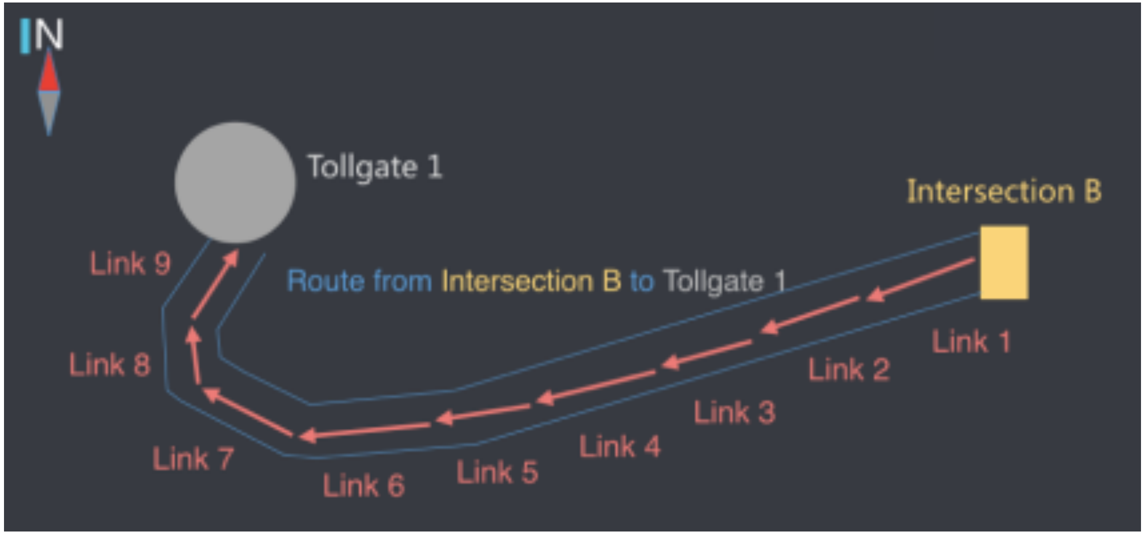
1. KDD CUP 2017
2. Data Preprocessing
3. Big Data Science Tool
4. Summary
5. Demo

1. KDD CUP 2017
2. Data Preprocessing
3. Big Data Science Tool
4. Summary
5. Demo

Overview

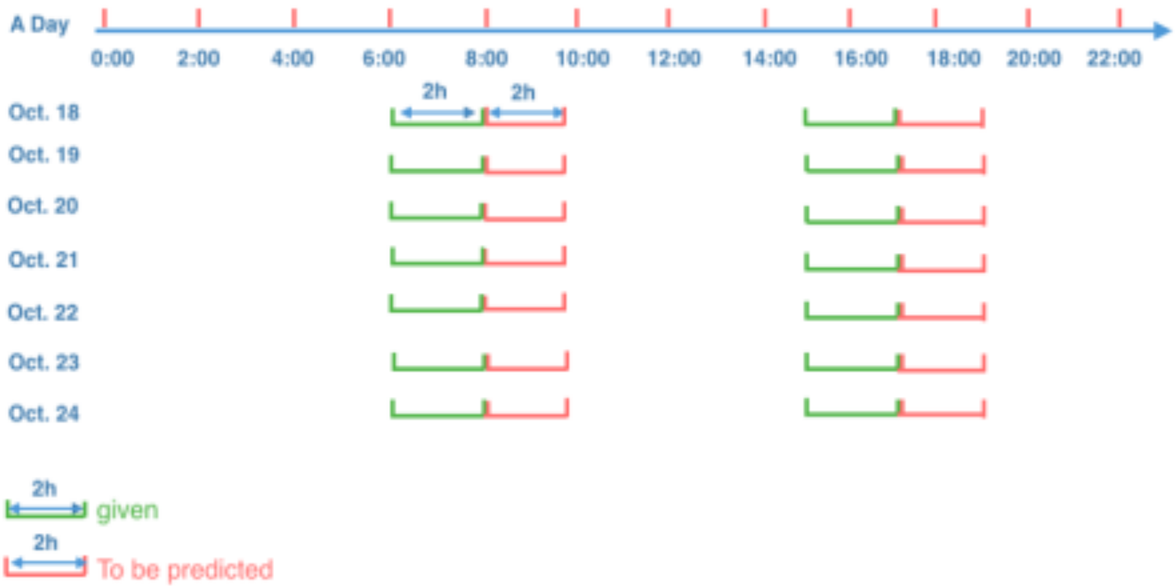
- **Topic:** Highway tollgates traffic flow prediction
- **Task:** Estimate the average travel time from intersections to tollgates in time windows

Data



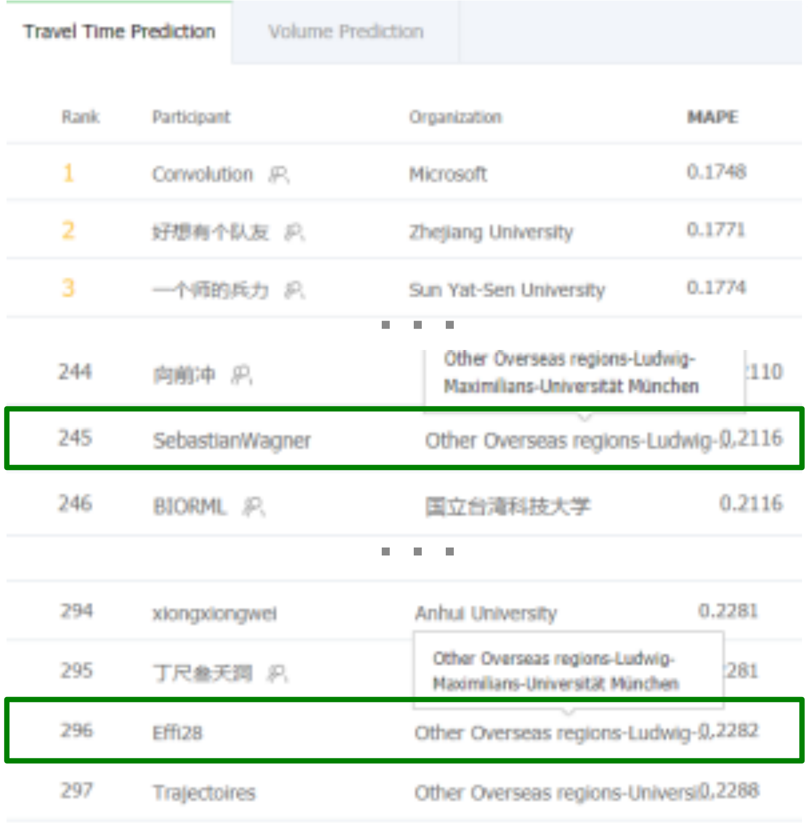
- 110000 data points
- 3 months time range
- 48 MB data size

Task



$$MAPE = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{T}\sum_{t=1}^{T}\left|\frac{d_{rt} - p_{rt}}{d_{rt}}\right|\right)$$

Our results



Rank	Participant	Organization	MAPE
1	Convolution	Microsoft	0.1748
2	好想有个队友	Zhejiang University	0.1771
3	一个师的兵力	Sun Yat-Sen University	0.1774
		■ ■ ■	
244	向前冲	Other Overseas regions-Ludwig-Maximilians-Universität München	110
245	SebastianWagner	Other Overseas regions-Ludwig-	0.2116
246	BIORML	国立台湾科技大学	0.2116
		■ ■ ■	
294	xiongxiongwei	Anhui University	0.2281
295	丁尺叁天洞	Other Overseas regions-Ludwig-Maximilians-Universität München	281
296	Effi28	Other Overseas regions-Ludwig-	0.2282
297	Trajectoires	Other Overseas regions-Universi	0.2288

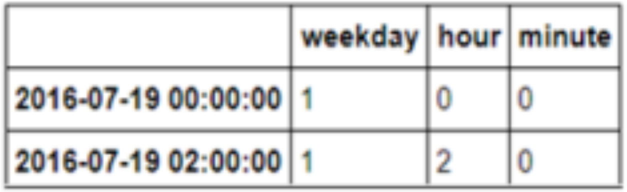
1. KDD CUP 2017
2. **Data Preprocessing**
3. Big Data Science Tool
4. Summary
5. Demo

Data Wrangling

- Specification of the input and output
- Data aggregation
- Handling missing values
- Manual feature selection

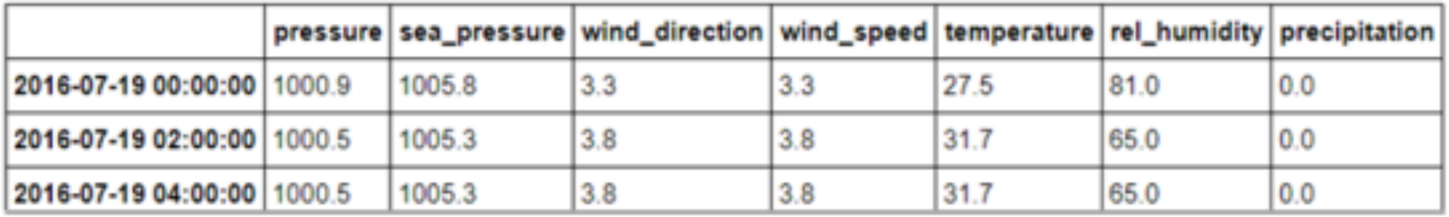
Input Features

X = Time Information (3 Features)



	weekday	hour	minute
2016-07-19 00:00:00	1	0	0
2016-07-19 02:00:00	1	2	0

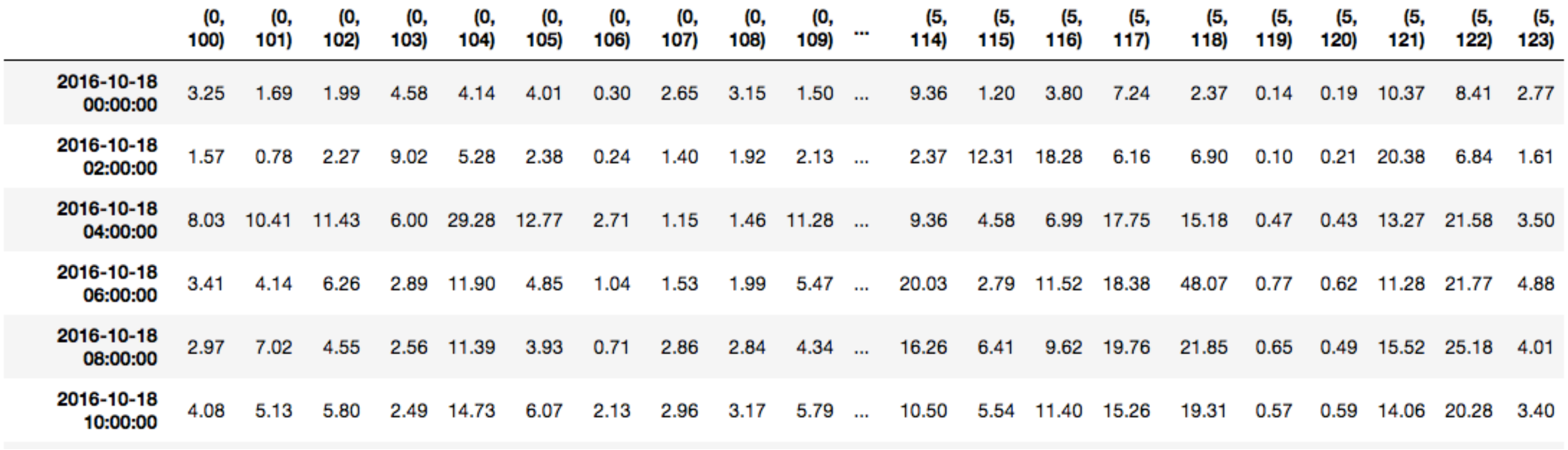
X = Weather (7 Features)



	pressure	sea_pressure	wind_direction	wind_speed	temperature	rel_humidity	precipitation
2016-07-19 00:00:00	1000.9	1005.8	3.3	3.3	27.5	81.0	0.0
2016-07-19 02:00:00	1000.5	1005.3	3.8	3.8	31.7	65.0	0.0
2016-07-19 04:00:00	1000.5	1005.3	3.8	3.8	31.7	65.0	0.0

Input Features

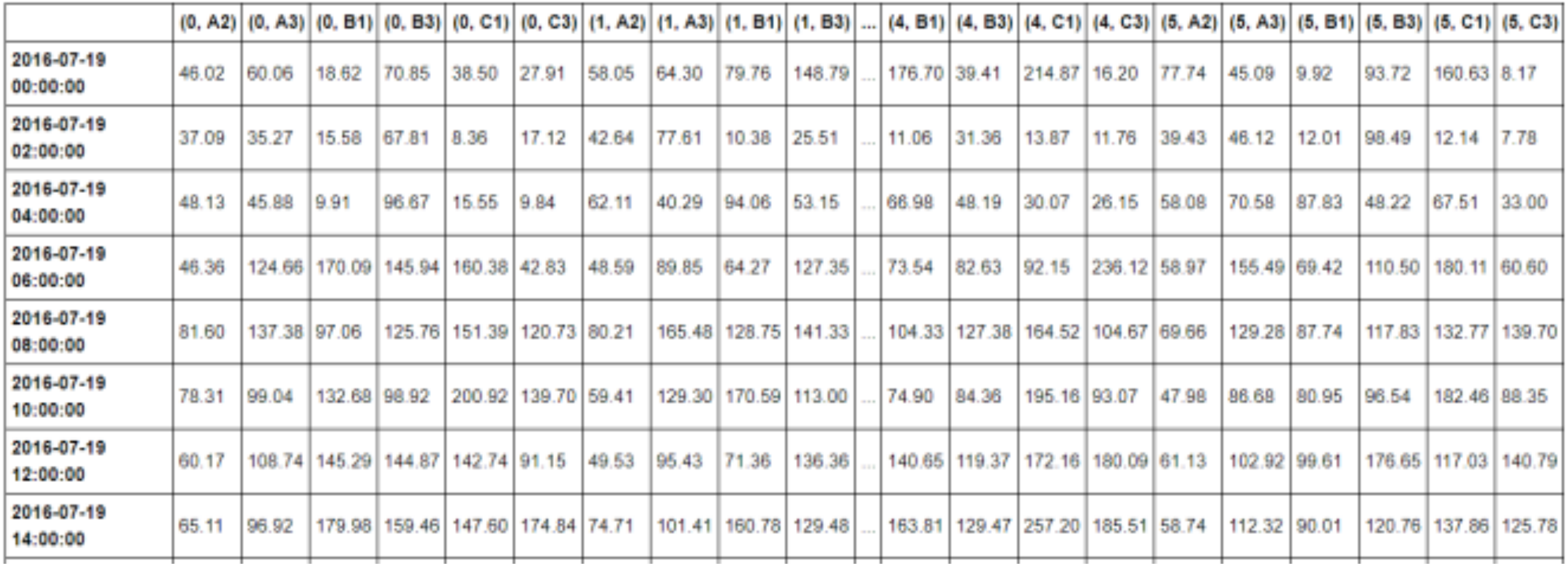
X = Current Situation (6 ·24 = 144 Features)



	(0, 100)	(0, 101)	(0, 102)	(0, 103)	(0, 104)	(0, 105)	(0, 106)	(0, 107)	(0, 108)	(0, 109)	...	(5, 114)	(5, 115)	(5, 116)	(5, 117)	(5, 118)	(5, 119)	(5, 120)	(5, 121)	(5, 122)	(5, 123)
2016-10-18 00:00:00	3.25	1.69	1.99	4.58	4.14	4.01	0.30	2.65	3.15	1.50	...	9.36	1.20	3.80	7.24	2.37	0.14	0.19	10.37	8.41	2.77
2016-10-18 02:00:00	1.57	0.78	2.27	9.02	5.28	2.38	0.24	1.40	1.92	2.13	...	2.37	12.31	18.28	6.16	6.90	0.10	0.21	20.38	6.84	1.61
2016-10-18 04:00:00	8.03	10.41	11.43	6.00	29.28	12.77	2.71	1.15	1.46	11.28	...	9.36	4.58	6.99	17.75	15.18	0.47	0.43	13.27	21.58	3.50
2016-10-18 06:00:00	3.41	4.14	6.26	2.89	11.90	4.85	1.04	1.53	1.99	5.47	...	20.03	2.79	11.52	18.38	48.07	0.77	0.62	11.28	21.77	4.88
2016-10-18 08:00:00	2.97	7.02	4.55	2.56	11.39	3.93	0.71	2.86	2.84	4.34	...	16.26	6.41	9.62	19.76	21.85	0.65	0.49	15.52	25.18	4.01
2016-10-18 10:00:00	4.08	5.13	5.80	2.49	14.73	6.07	2.13	2.96	3.17	5.79	...	10.50	5.54	11.40	15.26	19.31	0.57	0.59	14.06	20.28	3.40

Output Features

Y = Average Travel Time (6·6 = 36 Features)



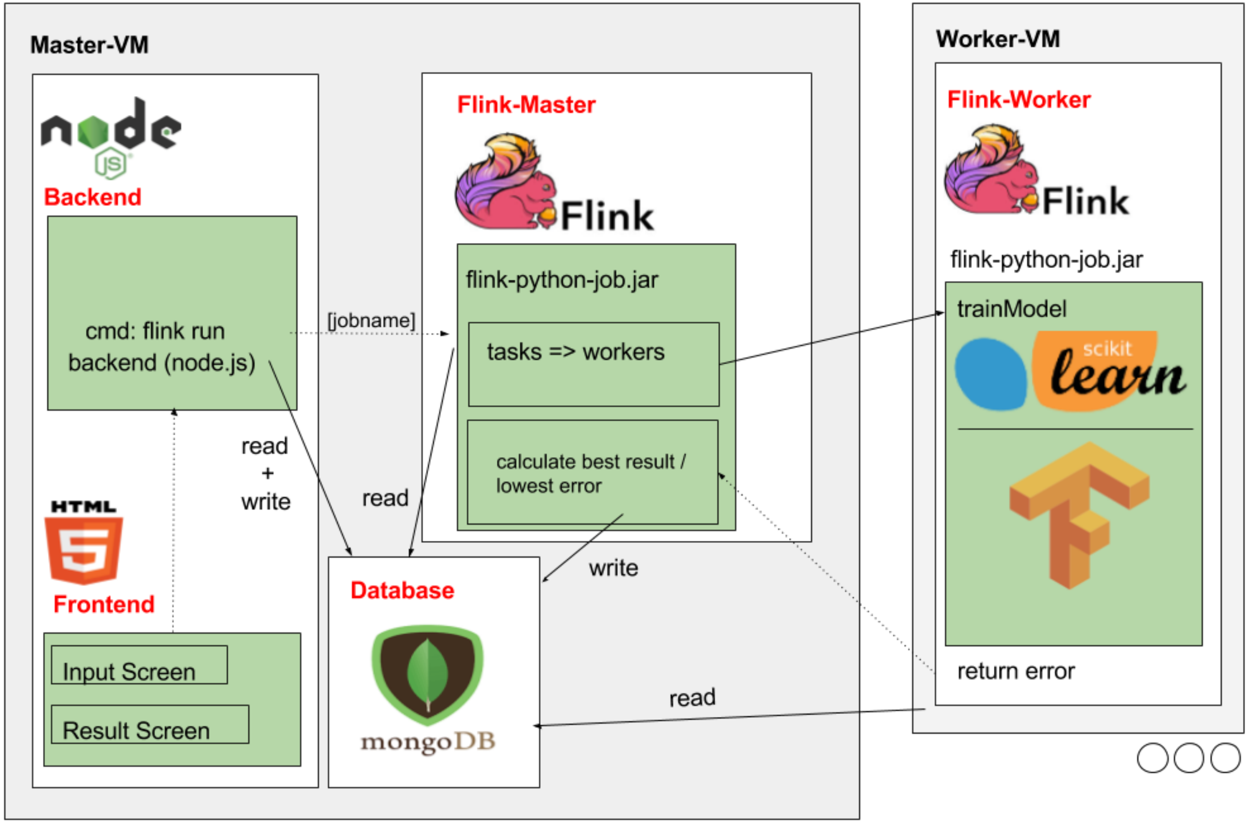
	(0, A2)	(0, A3)	(0, B1)	(0, B3)	(0, C1)	(0, C3)	(1, A2)	(1, A3)	(1, B1)	(1, B3)	...	(4, B1)	(4, B3)	(4, C1)	(4, C3)	(5, A2)	(5, A3)	(5, B1)	(5, B3)	(5, C1)	(5, C3)
2016-07-19 00:00:00	46.02	60.06	18.62	70.85	38.50	27.91	58.05	64.30	79.76	148.79	...	176.70	39.41	214.87	16.20	77.74	45.09	9.92	93.72	160.63	8.17
2016-07-19 02:00:00	37.09	35.27	15.58	67.81	8.36	17.12	42.64	77.61	10.38	25.51	...	11.06	31.36	13.87	11.76	39.43	46.12	12.01	98.49	12.14	7.78
2016-07-19 04:00:00	48.13	45.88	9.91	96.67	15.55	9.84	62.11	40.29	94.06	53.15	...	66.98	48.19	30.07	26.15	58.08	70.58	87.83	48.22	67.51	33.00
2016-07-19 06:00:00	46.36	124.66	170.09	145.94	160.38	42.83	48.59	89.85	64.27	127.35	...	73.54	82.63	92.15	236.12	58.97	155.49	69.42	110.50	180.11	60.60
2016-07-19 08:00:00	81.60	137.38	97.06	125.76	151.39	120.73	80.21	165.48	128.75	141.33	...	104.33	127.38	164.52	104.67	69.66	129.28	87.74	117.83	132.77	139.70
2016-07-19 10:00:00	78.31	99.04	132.68	98.92	200.92	139.70	59.41	129.30	170.59	113.00	...	74.90	84.36	195.16	93.07	47.98	86.68	80.95	96.54	182.46	88.35
2016-07-19 12:00:00	60.17	108.74	145.29	144.87	142.74	91.15	49.53	95.43	71.36	136.36	...	140.65	119.37	172.16	180.09	61.13	102.92	99.61	176.65	117.03	140.79
2016-07-19 14:00:00	65.11	96.92	179.98	159.46	147.60	174.84	74.71	101.41	160.78	129.48	...	163.81	129.47	257.20	185.51	58.74	112.32	90.01	120.76	137.86	125.78

1. KDD CUP 2017
2. Data Preprocessing
3. **Big Data Science Tool**
4. Summary
5. Demo

Prediction task

- Algorithms
  - Linear regression (Scikit-learn)
  - Support vector machine (Scikit-learn)
  - Feed-forward neural network (TensorFlow)
- Distribution of model learning (Apache Flink)
- Select best model for learning

Architecture



1. KDD CUP 2017
2. Data Preprocessing
3. Big Data Science Tool
4. **Summary**
5. Demo

Challenges

- Follow the motto:

  "Do not separate responsibilities! Everyone is responsible for everything."

- Rotation of Scrum master
- Security issues
- Dynamic rescaling not supported by Flink 1.3

Learnings

- Python 3
- Sklearn
- Numpy + Pandas
- Linear Regression
- SVR
- TensorFlow (lowlevel)

- Soft skills
- IT-Security
- Flink, Clusters
- MongoDB
- Scrum
- Web Dev

Expected outcome

- ✓ Selection of models for traffic flow prediction problem
- ✓ Documentation of models and explanation of hyperparameters
- ✓ Model selection framework in Flink
- ✓ GUI for model selection framework for arbitrary dataset
- ✓ Best model for traffic flow prediction problems

Future work

- Adding more models

  e.g. ensemble learning, recurrent networks

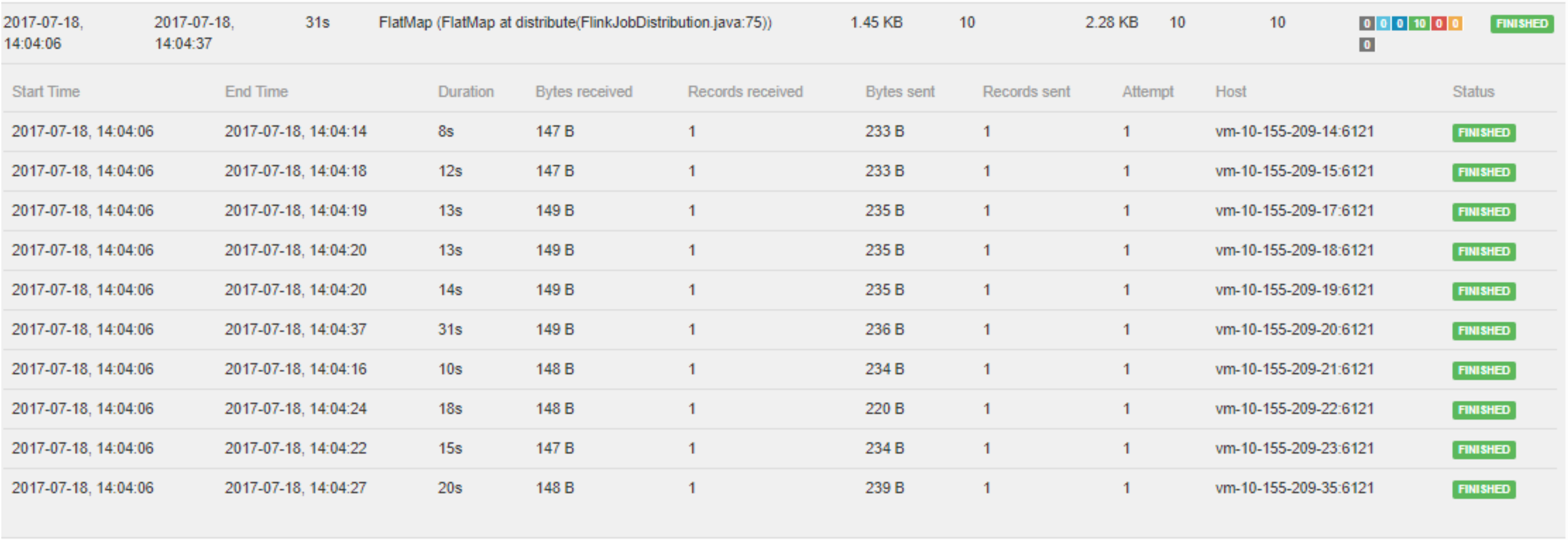
- Adding authentication
- Dashboards
- GPU computation for neural nets

1. KDD CUP 2017
2. Data Preprocessing
3. Big Data Science Tool
4. Summary
5. Demo

Live Demo

Thanks for your attention

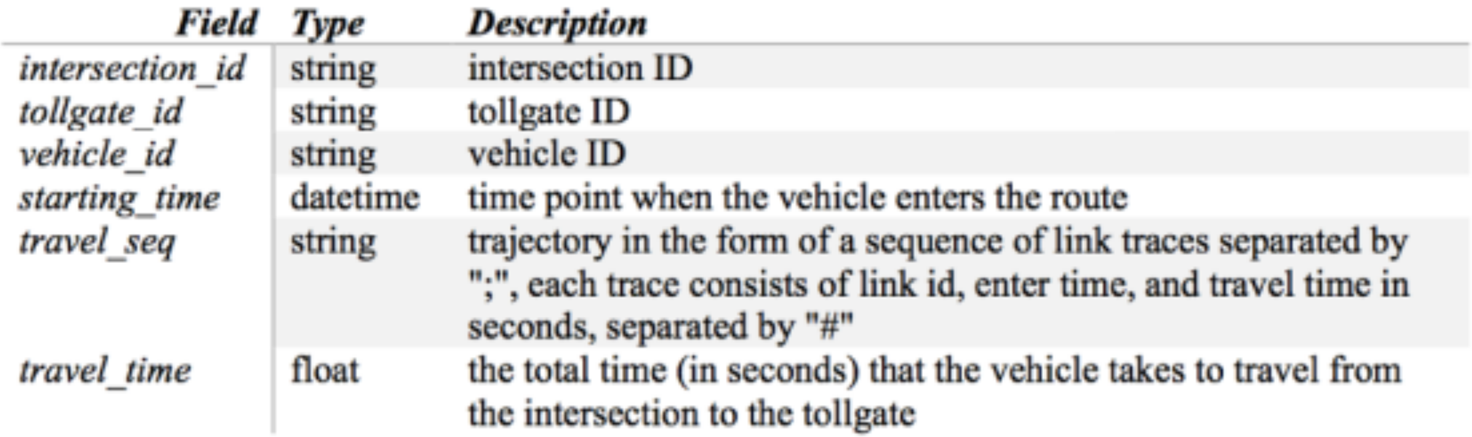
Scalability



2017-07-18, 14:04:06	2017-07-18, 14:04:37	31s	FlatMap (FlatMap at distribute(FlinkJobDistribution.java:75))	1.45 KB	10	2.28 KB	10	10	0 0 0 10 0 0 0	FINISHED

Start Time	End Time	Duration	Bytes received	Records received	Bytes sent	Records sent	Attempt	Host	Status
2017-07-18, 14:04:06	2017-07-18, 14:04:14	8s	147 B	1	233 B	1	1	vm-10-155-209-14:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:18	12s	147 B	1	233 B	1	1	vm-10-155-209-15:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:19	13s	149 B	1	235 B	1	1	vm-10-155-209-17:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:20	13s	149 B	1	235 B	1	1	vm-10-155-209-18:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:20	14s	149 B	1	235 B	1	1	vm-10-155-209-19:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:37	31s	149 B	1	236 B	1	1	vm-10-155-209-20:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:16	10s	148 B	1	234 B	1	1	vm-10-155-209-21:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:24	18s	148 B	1	220 B	1	1	vm-10-155-209-22:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:22	15s	147 B	1	234 B	1	1	vm-10-155-209-23:6121	FINISHED
2017-07-18, 14:04:06	2017-07-18, 14:04:27	20s	148 B	1	239 B	1	1	vm-10-155-209-35:6121	FINISHED

KDD CUP 2017 - Data



Field	Type	Description
intersection_id	string	intersection ID
tollgate_id	string	tollgate ID
vehicle_id	string	vehicle ID
starting_time	datetime	time point when the vehicle enters the route
travel_seq	string	trajectory in the form of a sequence of link traces separated by ";", each trace consists of link id, enter time, and travel time in seconds, separated by "#"
travel_time	float	the total time (in seconds) that the vehicle takes to travel from the intersection to the tollgate

Data statistics:

- 110000 trajectories
- 3 months
- 48 MB

Results

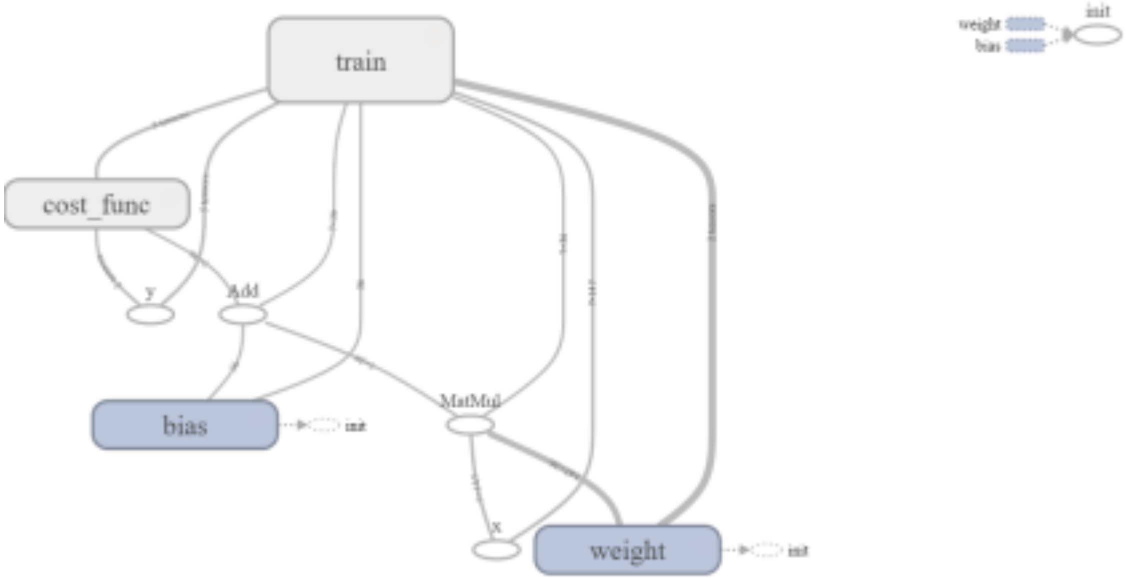
Sklearn:

- Linear Regression - TI - MAPE ?0.8?
- SVR - TI - MAPE 0.200

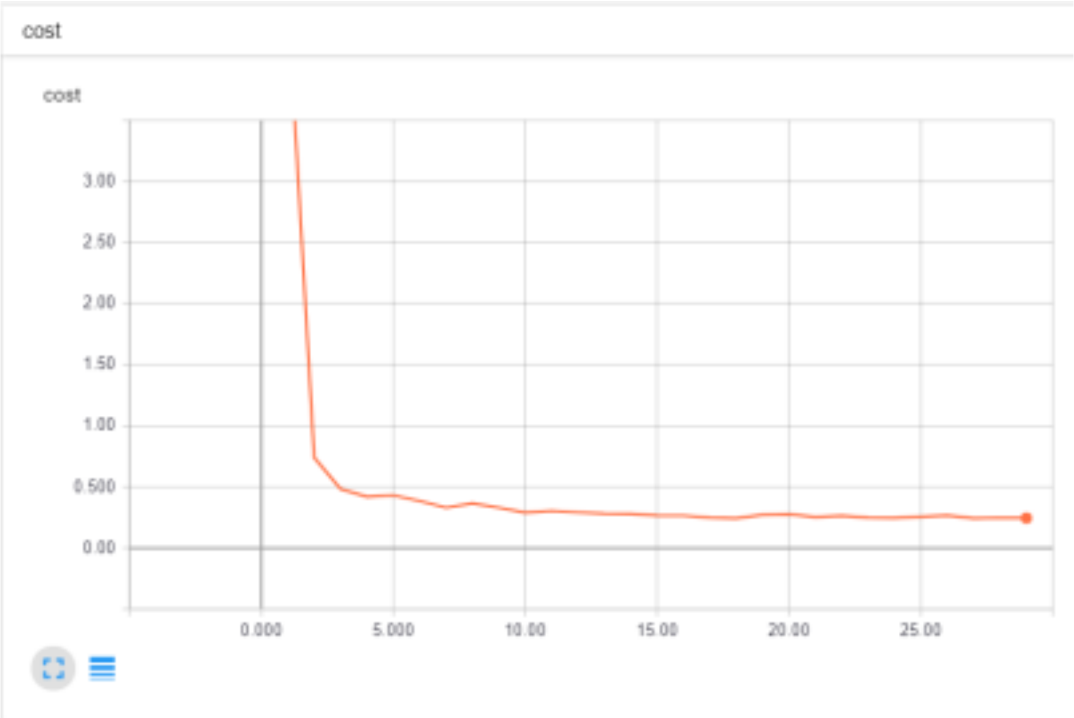
TensorFlow:

- Linear Regression - TI - MAPE 0.8
- NN - CS - MAPE 0.55
- DNN - MAPE ?

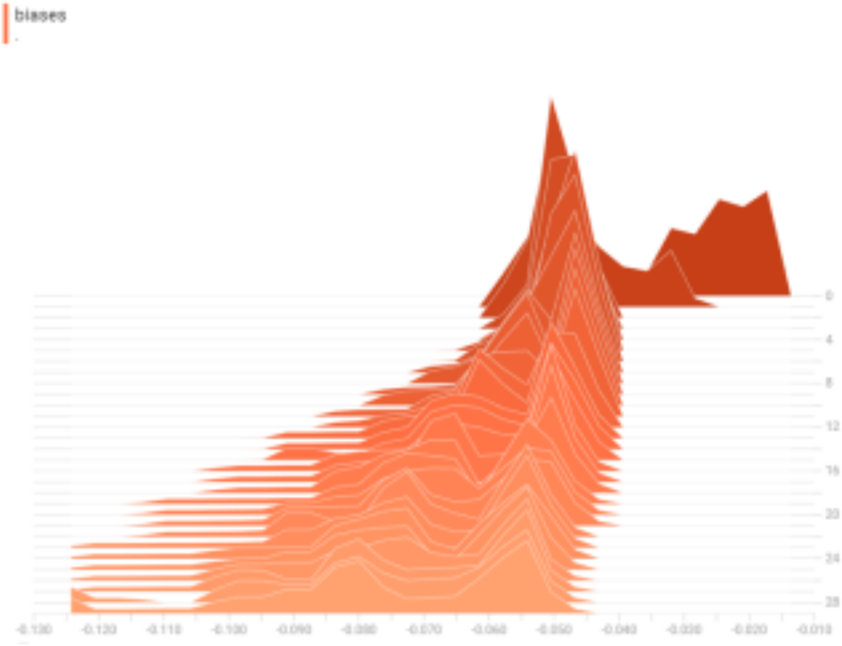
Neural network model

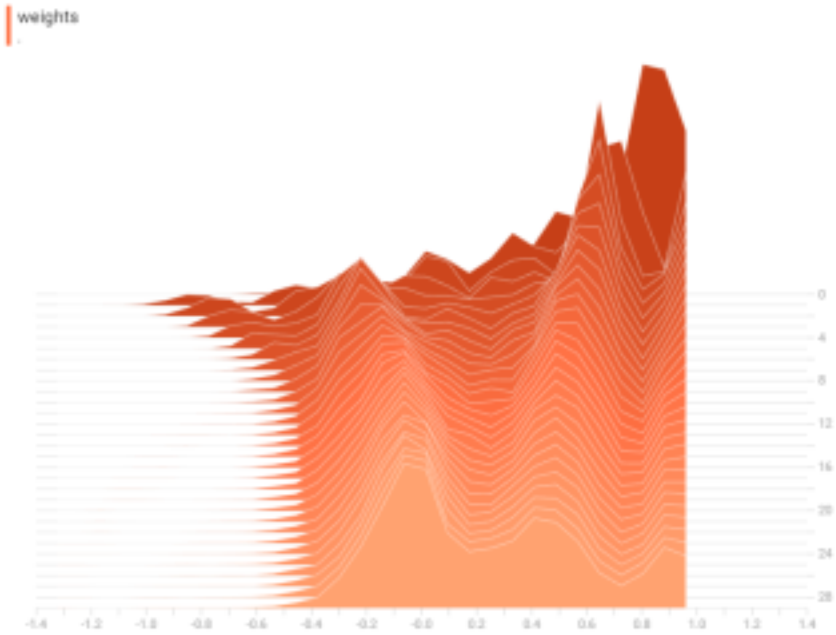


Training error



Learning process


biases


weights