Computing the Relative Value of Spatio-Temporal Data in Data Marketplaces

Santiago Andrés Azcoitia

IMDEA Networks Institute, Universidad Carlos III de Madrid

Marius Paraschiv, Nikolaos Laoutaris, IMDEA Networks Institute

Developing the Science of Networks

Data is now a key production factor and marketplaces have appeared to help bring it to market and satisfy the ever-growing demand [1] …

[1] A Survey of Data Marketplaces and Their Business Models", Santiago Andrés Azcoitia, Nikolaos Laoutaris, SIGMOD Record

… including B2B general-purpose data marketplaces (DMs) trading ANY kind of data, …

[1] A Survey of Data Marketplaces and Their Business Models", Santiago Andrés Azcoitia, Nikolaos Laoutaris, SIGMOD Record

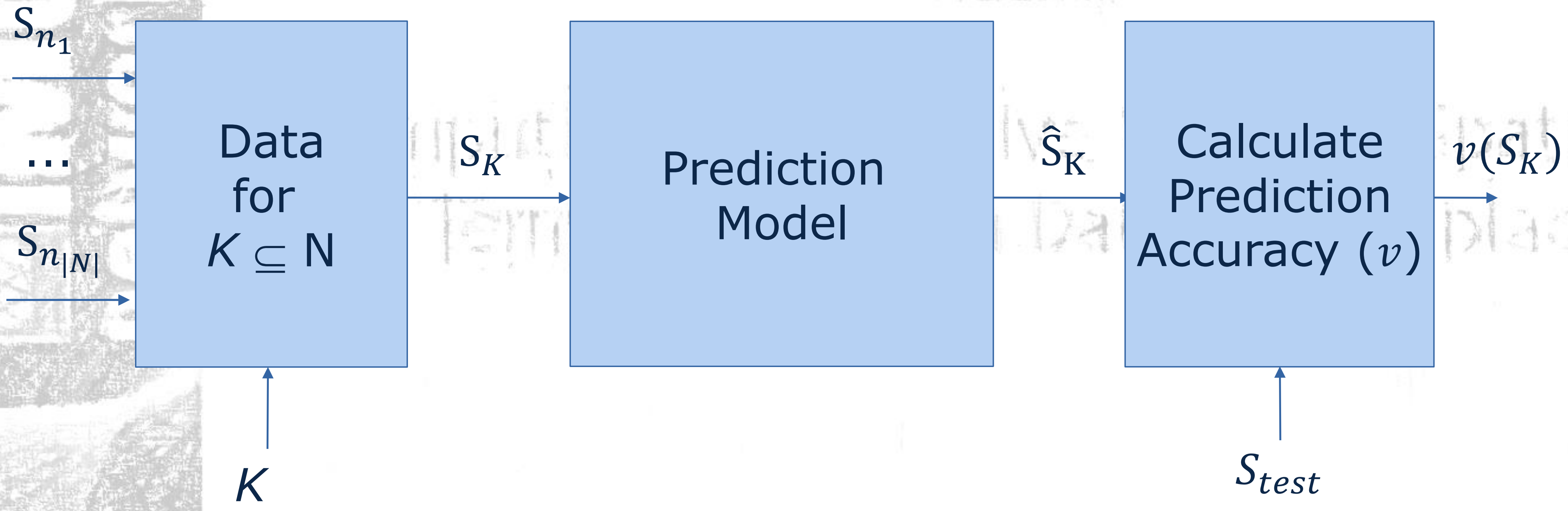
… domain-specific DMs, some of which are trading spatio-temporal data offered by data providers, …

[1] A Survey of Data Marketplaces and Their Business Models", Santiago Andrés Azcoitia, Nikolaos Laoutaris, SIGMOD Record

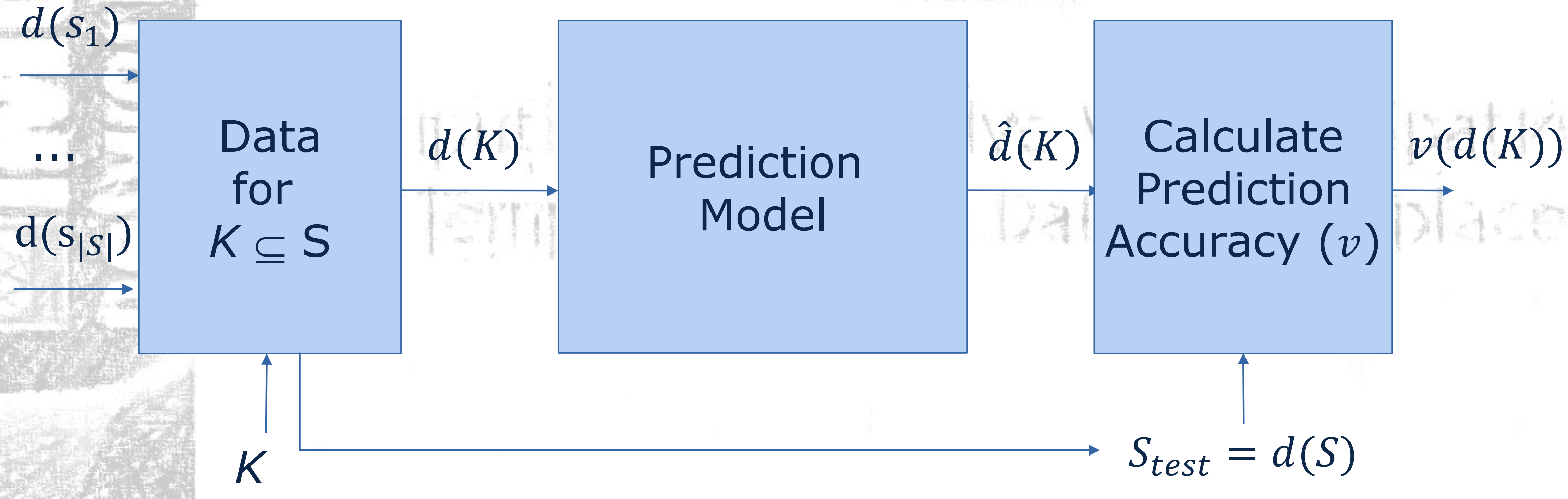
… and Personal Information Management Systems (aka PIMS) empowering people to take control and share their personal data, like their location.

[1] A Survey of Data Marketplaces and Their Business Models", Santiago Andrés Azcoitia, Nikolaos Laoutaris, SIGMOD Record

DMs can combine data from different sources, whose relative value is useful for i) for buyers to select suitable data, and ii) for DMs to retribute data providers



DMs can combine data from different sources, whose relative value is useful for i) for buyers to select suitable data, and ii) for DMs to retribute data providers



We have computed the value of data from companies and individuals in different settings and prediction tasks

CHI

Demand prediction in Chicago (Jan-Sep 2019)

11 MM rides from 6,469 cars grouped in 16 companies

NYC

Demand prediction in NYC (Apr-May 2019)

65 MM rides from 33 companies in 261 districts

Travel time prediction in Porto (Jul’13 – Jun’14)

1,71 MM ride trajectories from 448 taxis

We resorted to the Shapley value and to other simpler methods to calculate the relative value of data from companies and individuals

#	Metric description	Complexity
1	Shapley value, average marginal contribution of S_{ni} to every combination of the rest of datasets: $$\phi(n_i) = \sum_{K \subseteq N \setminus \{n_i\}} \frac{\|K\|!(\|N\| - \|K\| - 1)!}{\|N\|!} [v(S_K \cup S_{n_i}) - v(S_K)],$$	$O(2^{\|N\|})$ Approx. $O(\|N\|^2)$
2	Leave-one-out – marginal contribution of a source to the rest of the sources in a transaction $$LOO(n_i) = v(S_N) - v(S_{N-\{n_i\}})$$	O(N)
3	Equitable	-
4	Proportional to the volume of data $\|S_{ni}\|$	-
5	Any other context-specific heuristics?	-

We have computed the value of data from companies and individuals in different settings and prediction tasks related to spatio-temporal data

CHI	Demand prediction in Chicago (Jan-Sep 2019) 11 MM rides from 6,469 cars grouped in 16 companies
NYC	Demand prediction in NYC (Apr-May 2019) 65 MM rides from 33 companies in 261 districts
	Travel time prediction in Porto (Jul'13 – Jun'14) 1,71 MM ride trajectories from 448 taxis

institute imdea networks

Data Transparency Group

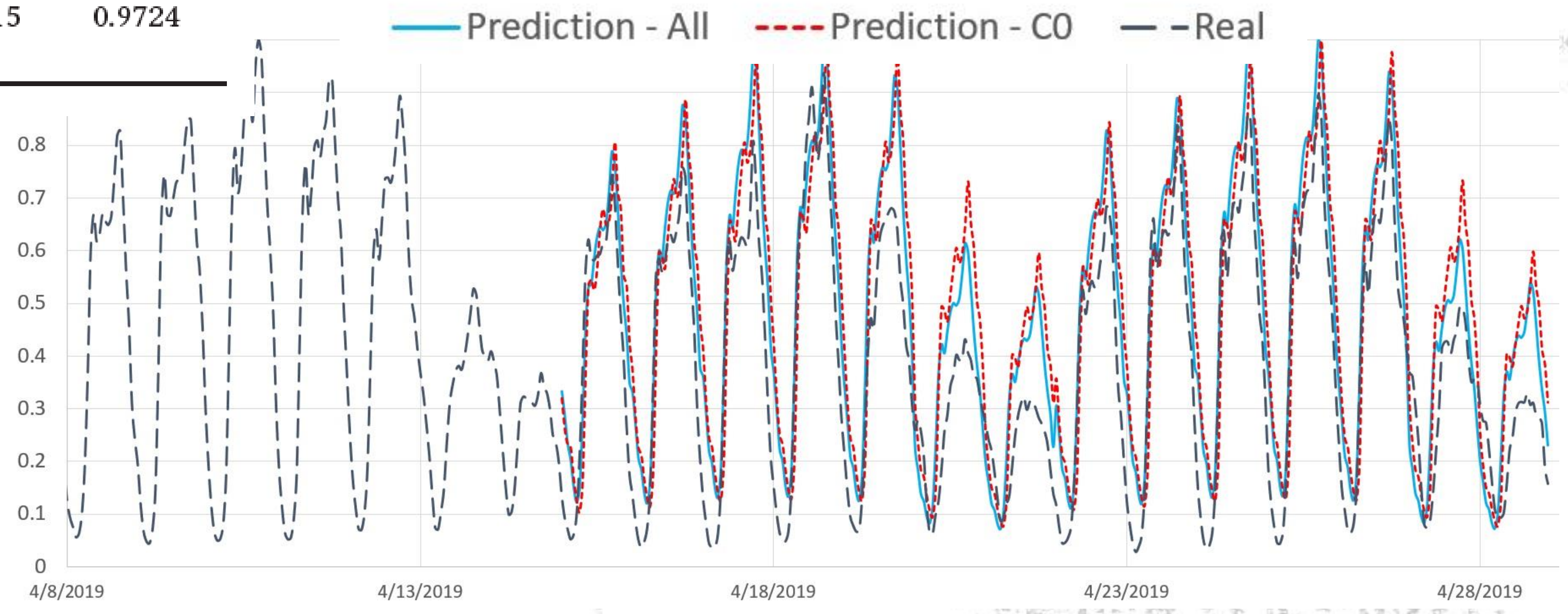
acm SIGSPATIAL

At city level, all companies are able to reconstruct demand with a high accuracy…

City-wide accuracy by company

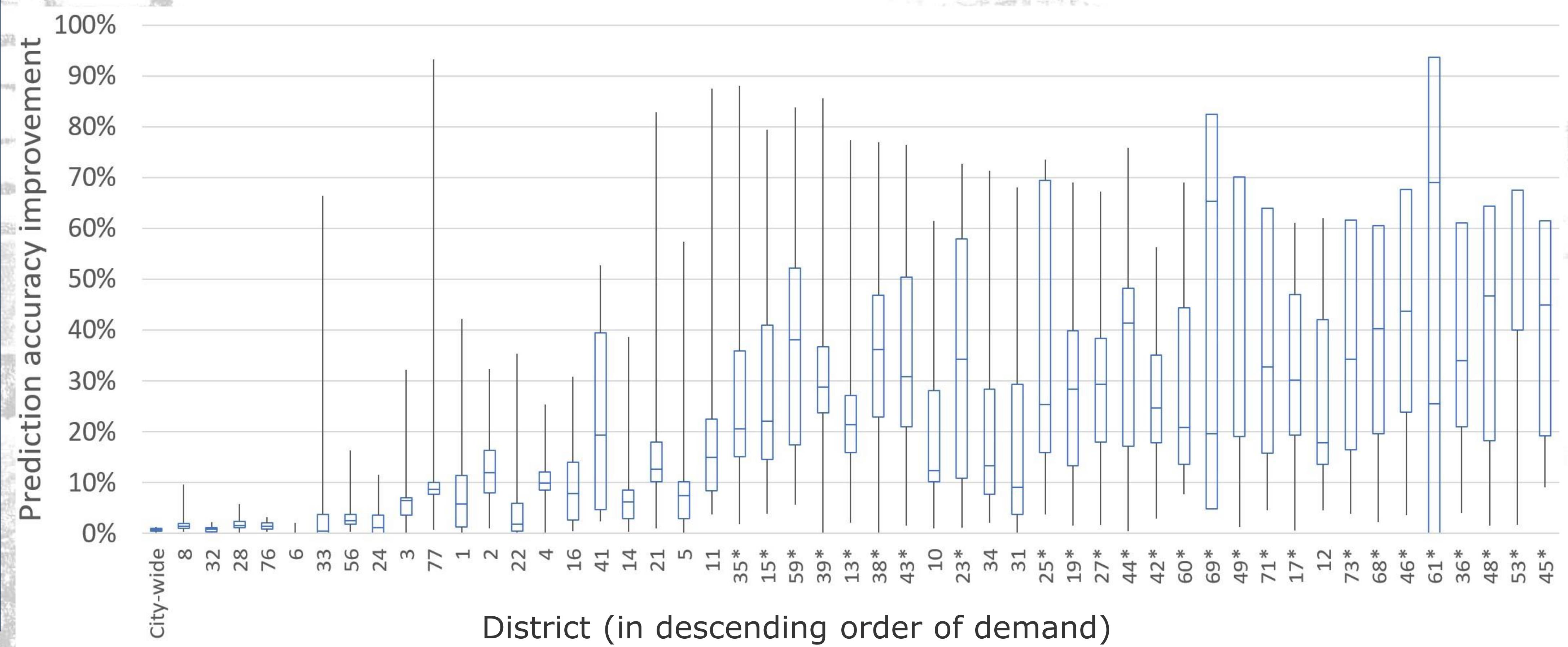
Co	Accuracy	Co	Accuracy	Co	Accuracy
All	0.9833	C5	0.9736	C11	0.9659
C0	0.9686	C6	0.9800	C12	0.9845
C1	0.9835	C7	0.9804	C13	0.9725
C2	0.9794	C8	0.9797	C14	0.9767
C3	0.9737	C9	0.9861	C15	0.9724
C4	0.9801	C10	0.9829		

Predicted vs real normalized plot



… but they must combine their data in smaller districts. The smaller the district, the more value companies get by cooperating and sharing data.

Box-plot (over companies) of potencial prediction accuracy by combining datasets



imdea networks institute

Data Transparency Group

acm SIGSPATIAL

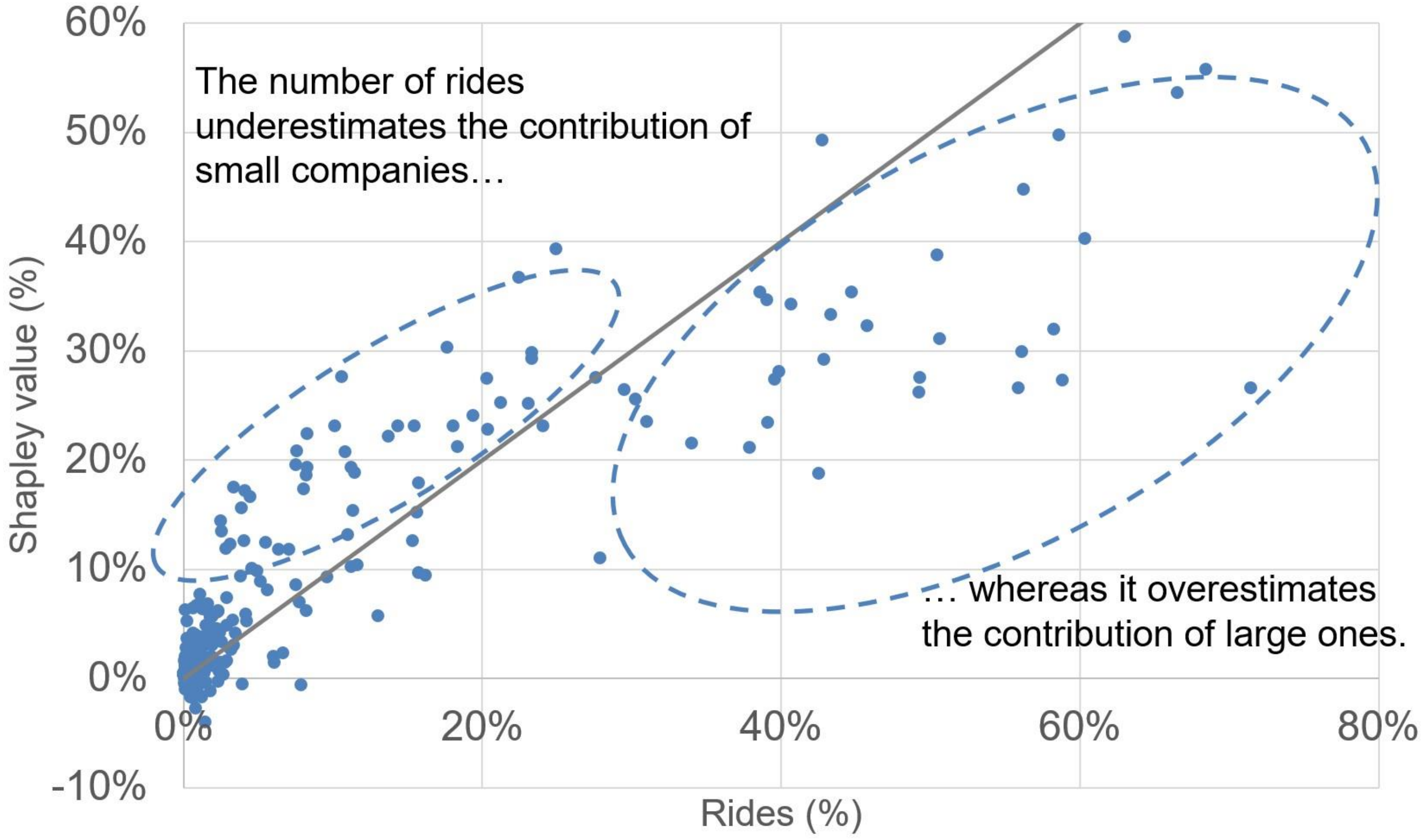
Their Shapley values strongly differ between and across districts. LOO values are not useful nor correlated to Shapley values ($R^2 = 0.38$), with negatives.

Shapley value, LOO and nº rides (Rd%) for three small districts

Co	15 SV	15 LOO	15 Rd(%)	17 SV	17 LOO	17 Rd(%)	19 SV	19 LOO	19 Rd(%)
1	11.2	0.5	2.5	14.0	0.6	8.3	2.0	0.0	3.4
2	1.8	-0.1	0.8	0.0	-0.1	0.5	1.5	0.0	0.5
3	1.0	0.0	0.3	0.2	0.0	0.5	0.3	0.0	0.0
4	0.4	-0.1	0.2	0.2	0.0	0.0	0.4	0.0	0.1
5	2.3	-0.1	0.9	0.4	0.0	0.5	0.7	0.0	0.8
6	16.4	-1.2	37.9	28.0	8.7	56.2	24.1	3.3	38.6
7	1.1	-0.3	0.4	0.2	0.0	0.4	0.2	0.2	0.5
8	1.1	-0.1	0.8	0.3	0.4	1.4	1.5	0.2	0.5
9	-0.3	0.0	0.2	0.0	0.0	0.2	-0.6	-0.1	0.3
10	2.3	0.4	1.4	0.2	-0.2	0.8	0.9	0.1	0.7
11	0.6	0.0	0.3	0.2	0.1	0.5	1.4	0.0	0.5
12	4.4	-0.1	1.9	0.4	0.1	0.9	2.4	0.1	1.9
13	17.9	0.8	18.1	0.3	-0.2	1.3	4.3	0.0	1.3
14	16.7	-0.9	34.0	17.2	0.0	27.6	26.4	1.9	50.4
15	0.4	0.0	0.1	0.8	0.0	0.3	0.4	0.0	0.1
16	0.2	-0.1	0.2	0.0	0.0	0.8	2.4	0.1	0.5

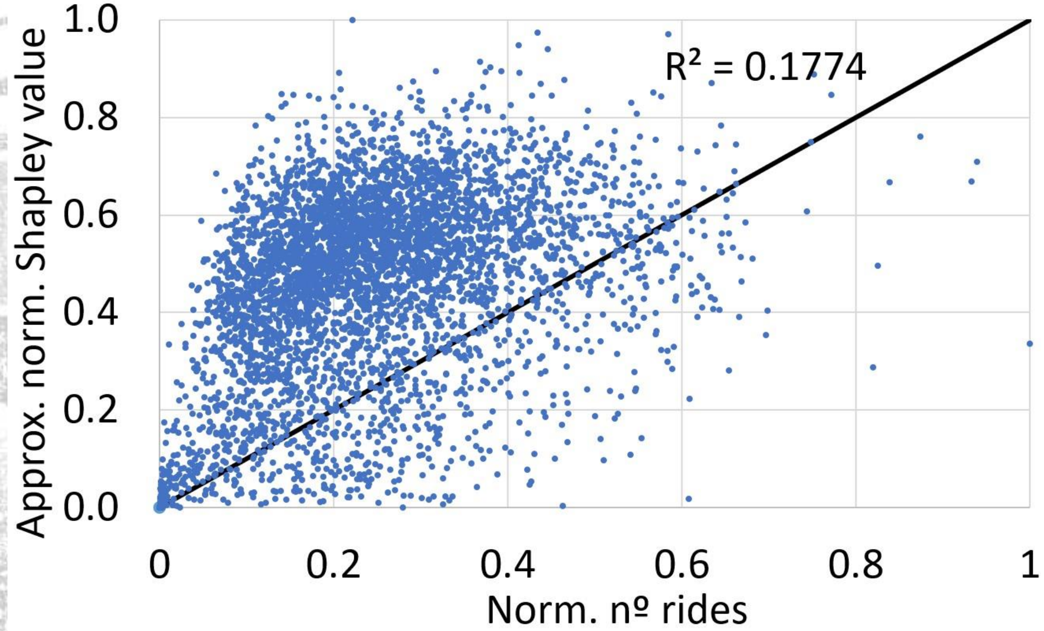
The number of rides reported by a company does not necessarily reflect the average value its data provides to the prediction task…

Shapley value vs. nº rides by company in small districts of Chicago

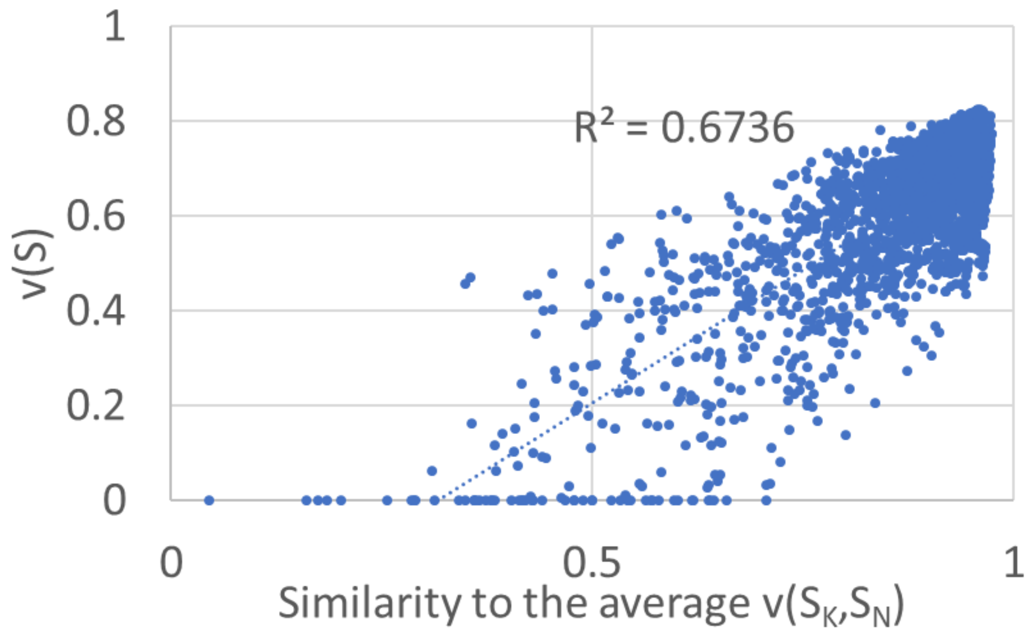


… nor does it reflect the value of individuals at city level, that instead seems to to be more correlated to the similarity of the input to the average (R^2 = 0.6736)

Shapley value vs. nº rides by driver at city level



Shapley value vs. averageness at city level



institute imdea networks
Data Transparency Group
acm SIGSPATIAL

We have computed the value of data from companies and individuals in different settings and prediction tasks related to spatio-temporal data

CHI

Demand prediction in Chicago (Jan-Sep 2019)

11 MM rides from 6,469 cars grouped in 16 companies

NYC

Demand prediction in NYC (Apr-May 2019)

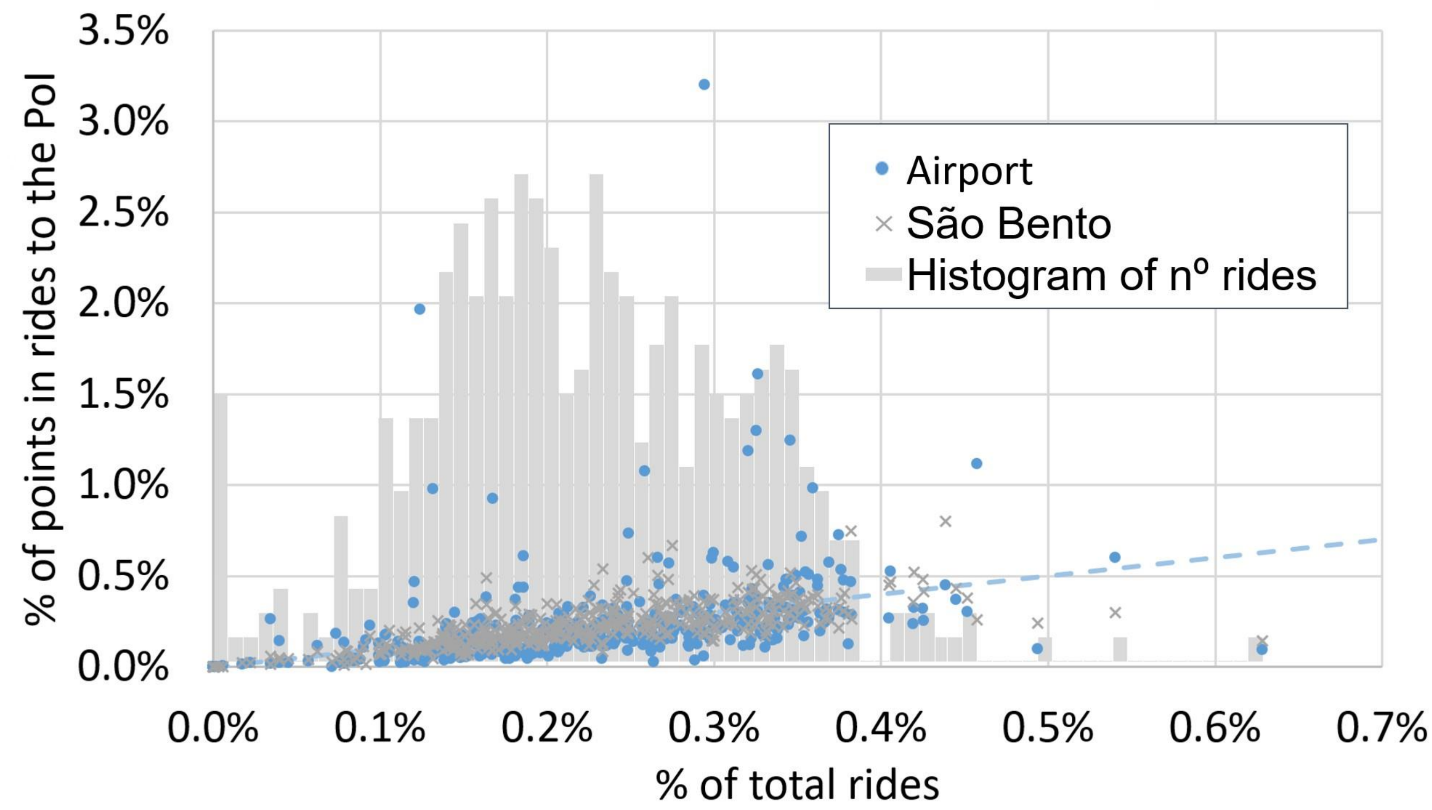
65 MM rides from 33 companies in 261 districts

Travel time prediction in Porto (Jul’13 – Jun’14)

1,71 MM ride trajectories from 448 taxis

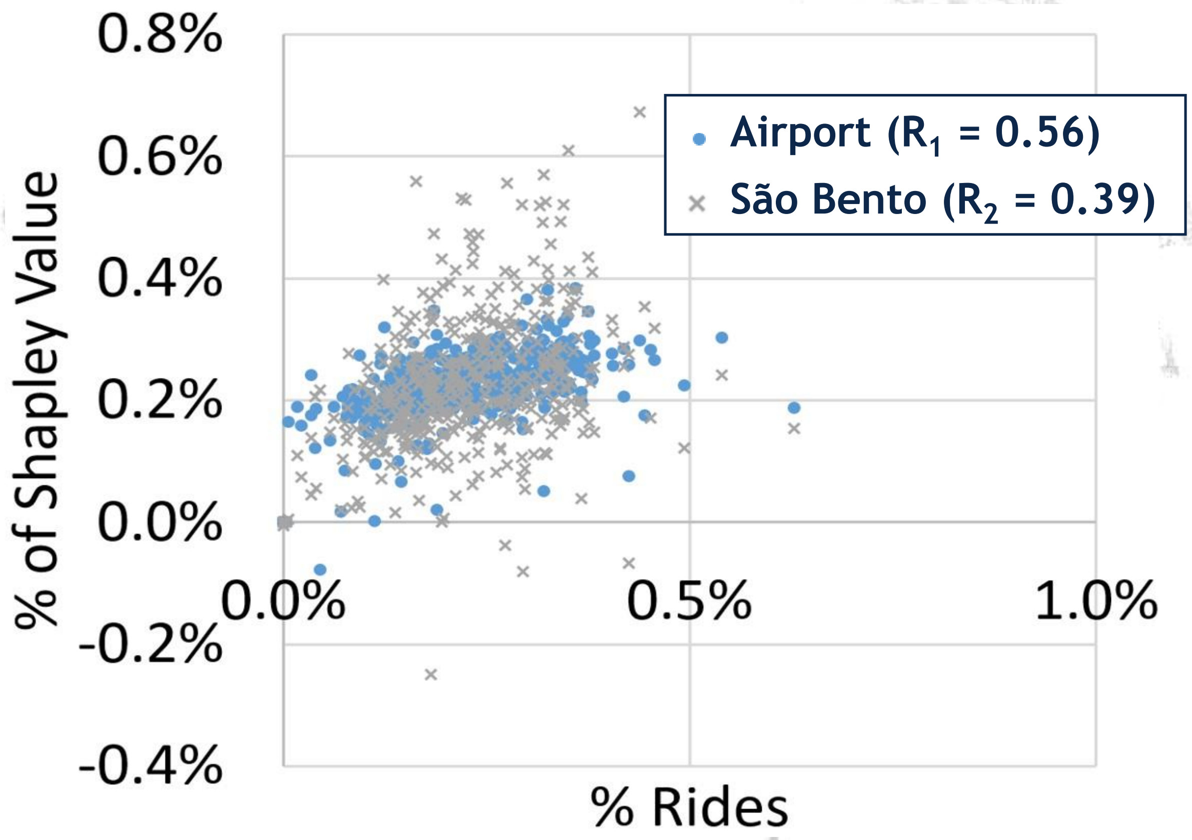
In a completely different setting, we predicted travel time to Porto’s airport and to São Bento station, based on data of individual taxis

Nº rides reported by taxi driver



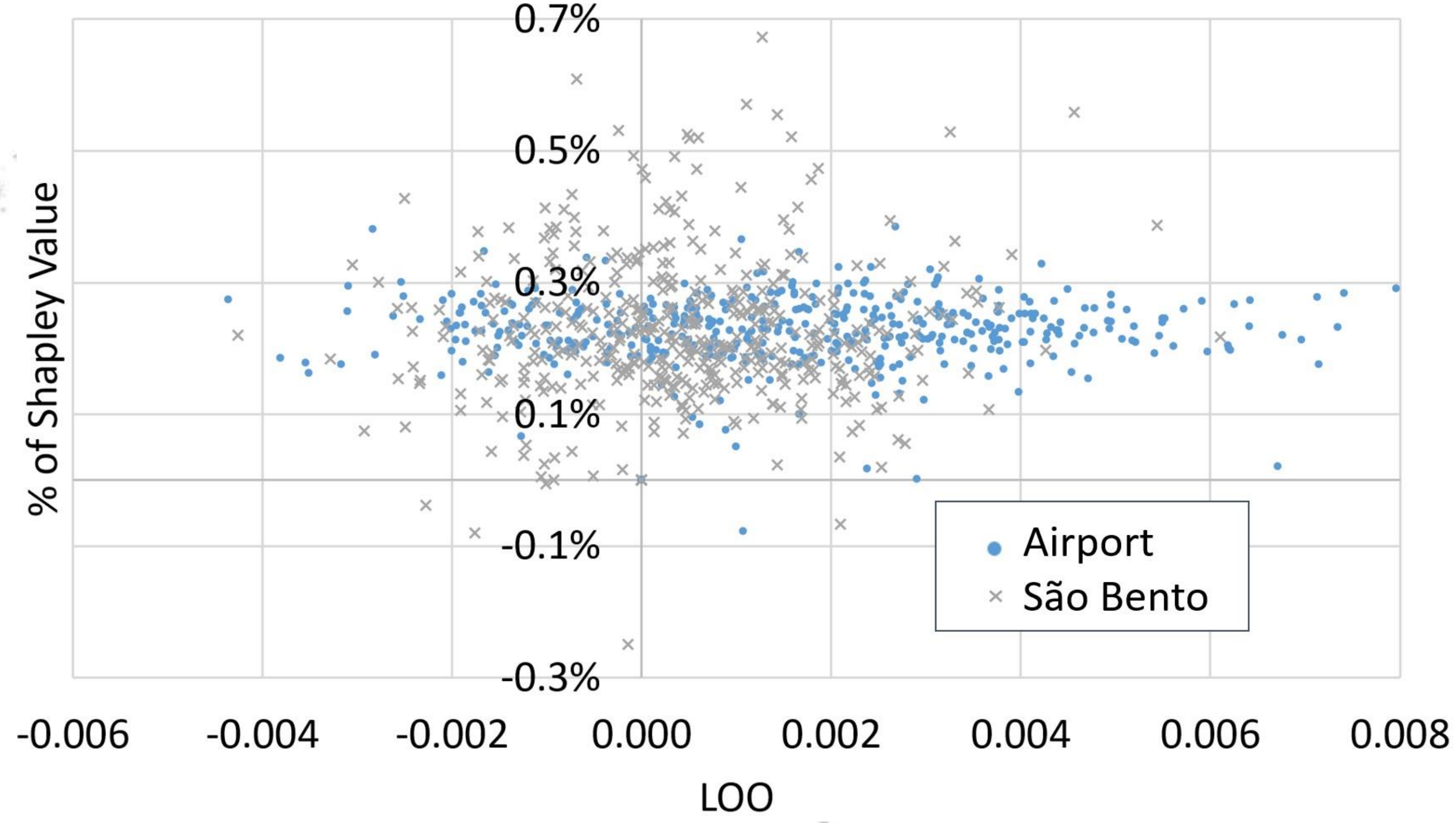
Similar to the case of Chicago, the Shapley value is different for each driver, and it is not significantly correlated with the nº rides they report or with LOO…

Shapley value vs. % rides reported by each taxi



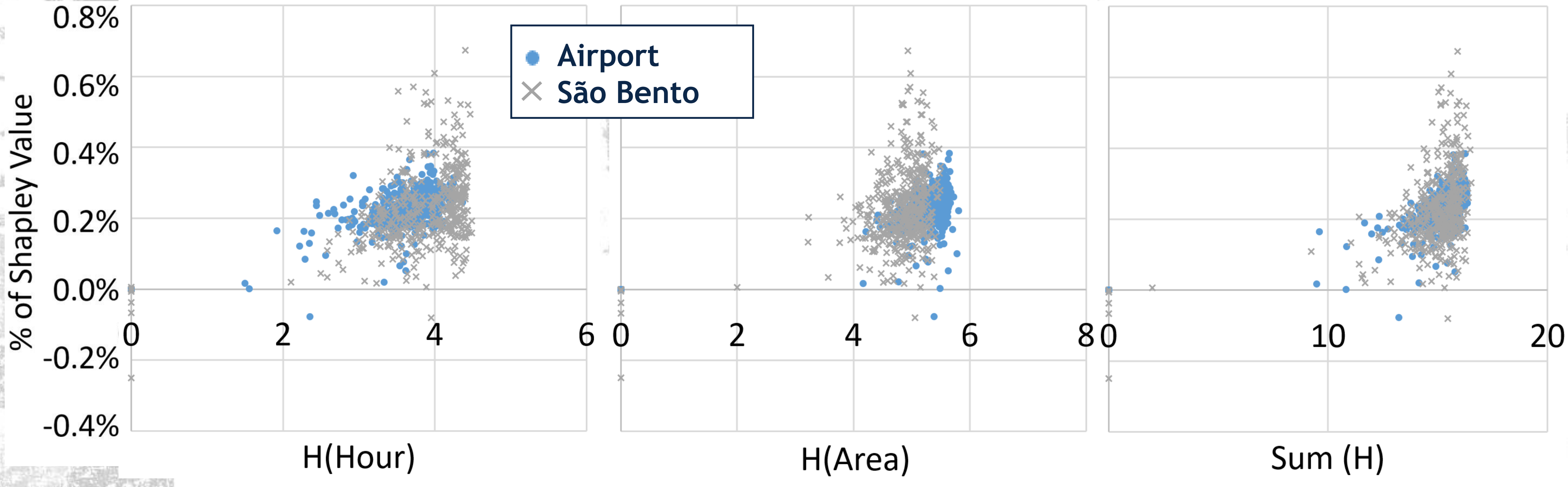
… nor with LOO-values.

Shapley vs. LOO values



Interestingly, the diversity of data reported, measured as Shannon's entropy (H) of key spatio-temporal features, shows a stronger correlation in this case

Pearson correlation of Shapley values with data features



$R_1 = 0.69$, $R_2 = 0.57$ | $R_1 = 0.62$, $R_2 = 0.58$ | $R_1 = 0.69$, $R_2 = 0.60$

Bottomline…

- Shapley seems to be a “necessary evil” to capture the importance of data to a given ML task,…
- … which simpler heuristics based on volume and LOO fail to approximate
- BUT, we found context-specific heuristics measuring valuable inherent features of data, such as its averageness or its spatio-temporal diversity, which do better approximate Shapley values
- Not only are they *faster* to calculate, but they are *more explainable* to end users, as well
- We are working on:
  - Computing the value of data and identifying other such heuristics in new settings and tasks
  - Designing components for data marketplaces to “select” data and calculate payoffs based on such pre-calculated heuristics

Thank you!

Now it is Q&A time!

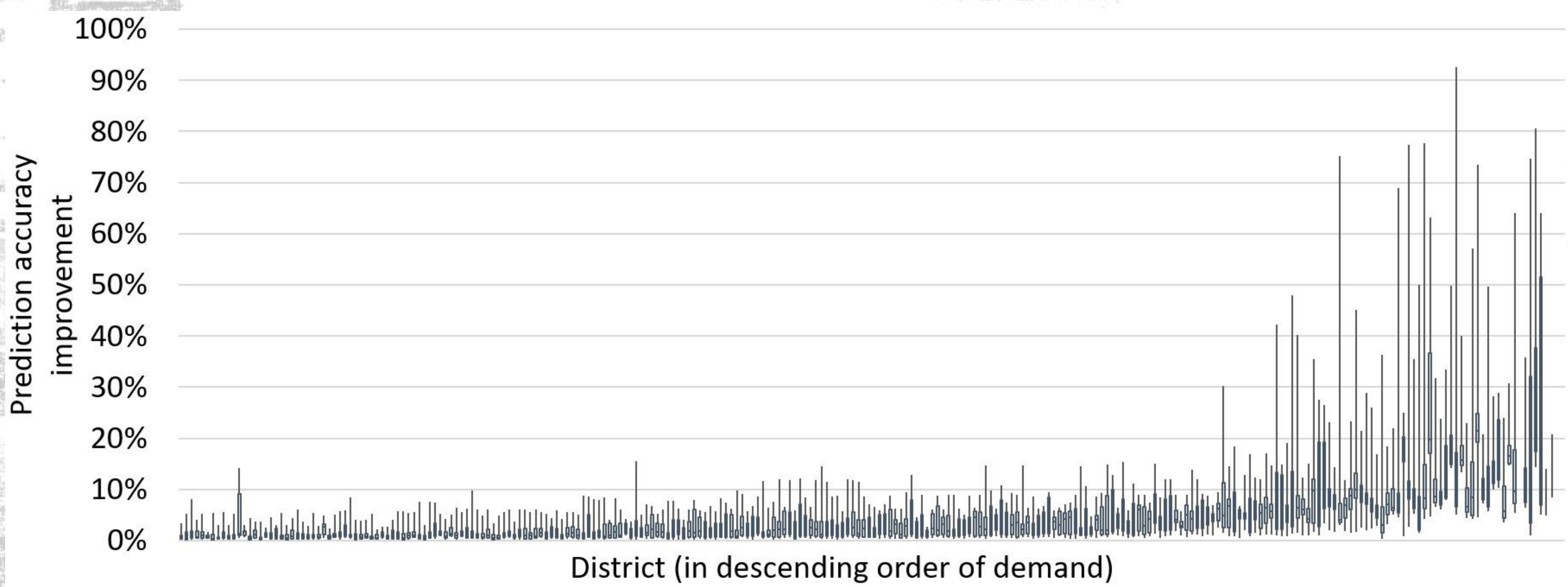
For more information please contact:

Santiago Andrés Azcoitia
santiago.azcoitia@imdea.org

Data Transparency Group

Results in NYC lead to the same conclusion, first most taxi companies are able to predict demand in 219/261 districts…

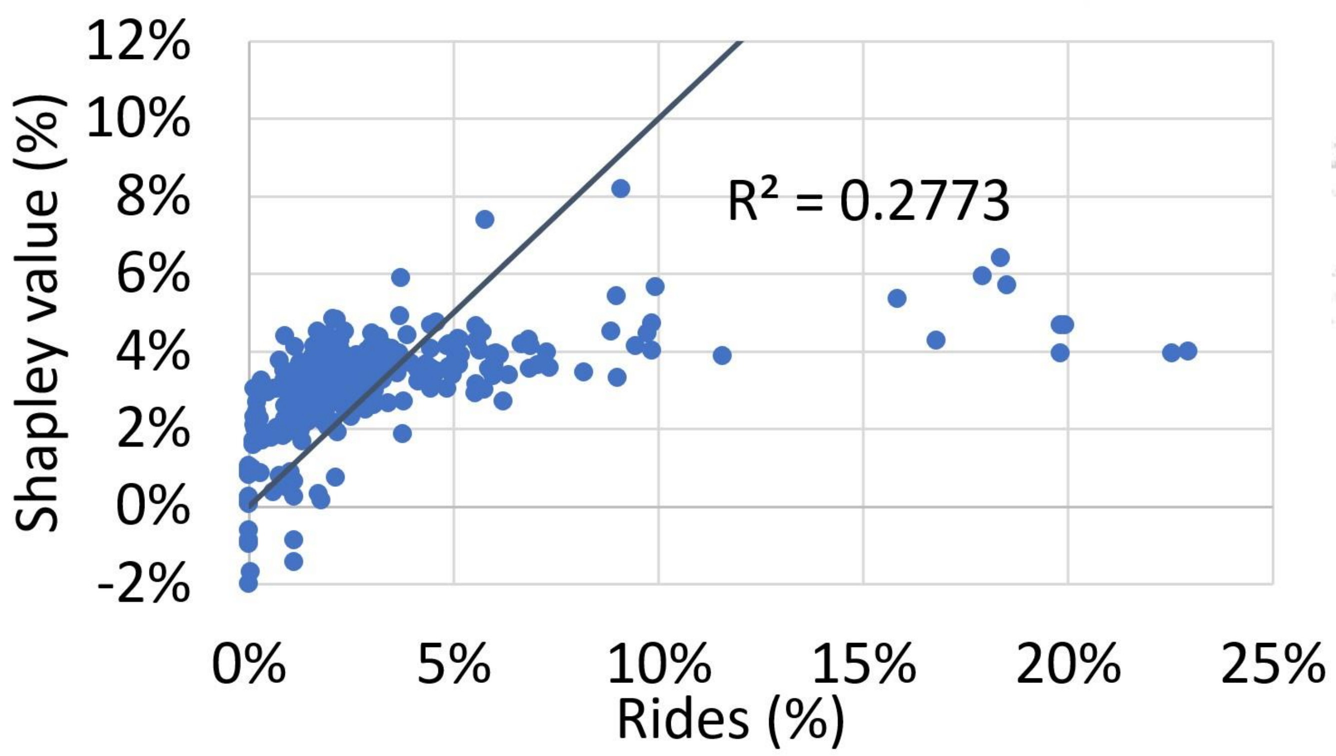
Box-plot (over companies) of potencial prediction accuracy by combining datasets



Data Transparency Group

… and the number of rides reported by drivers in small districts is again weakly correlated with the Shapley value, which also holds for LOO

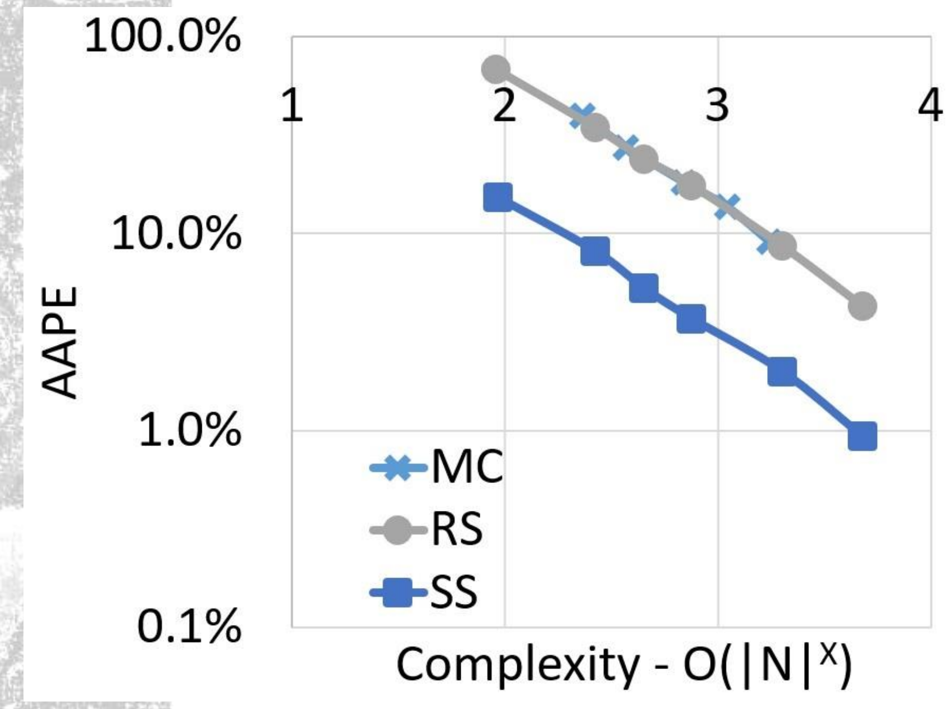
Shapley value vs. nº rides by company in small districts of NYC



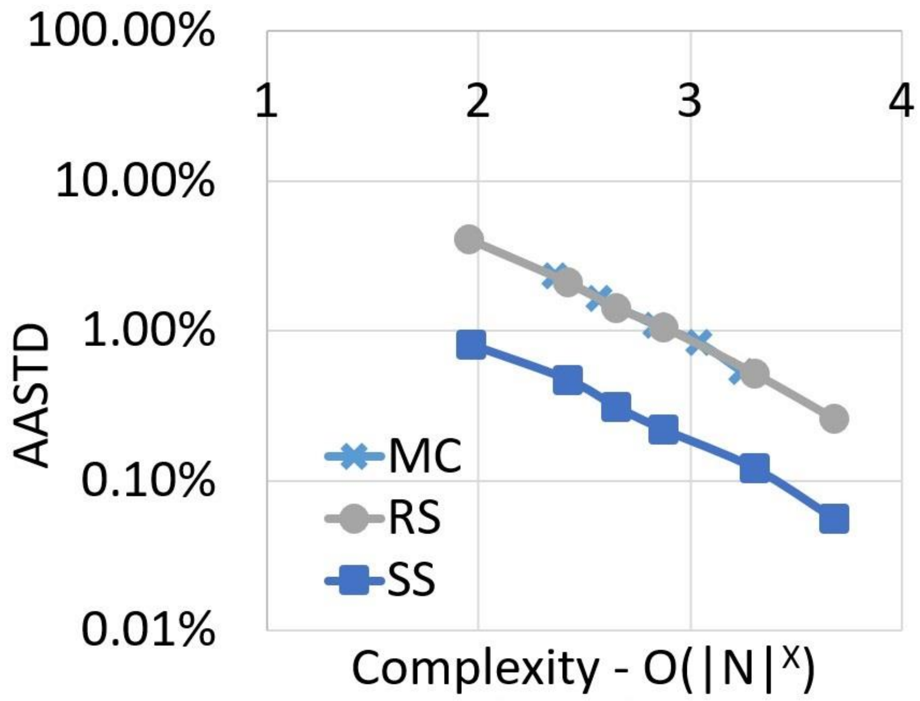
Data Transparency Group

We tested several approximations to the Shapley value and we found that structured sampling outperforms Monte Carlo and Random Sampling, …

Complexity vs Error



Complexity vs Robustness



imdea networks institute
Data Transparency Group
acm SIGSPATIAL

… and it is able to approximate payoff distributions based on Shapley values with an error of less than 10 in $O(N)$ to $O(N^2)$ computation time.

Complexity vs Error