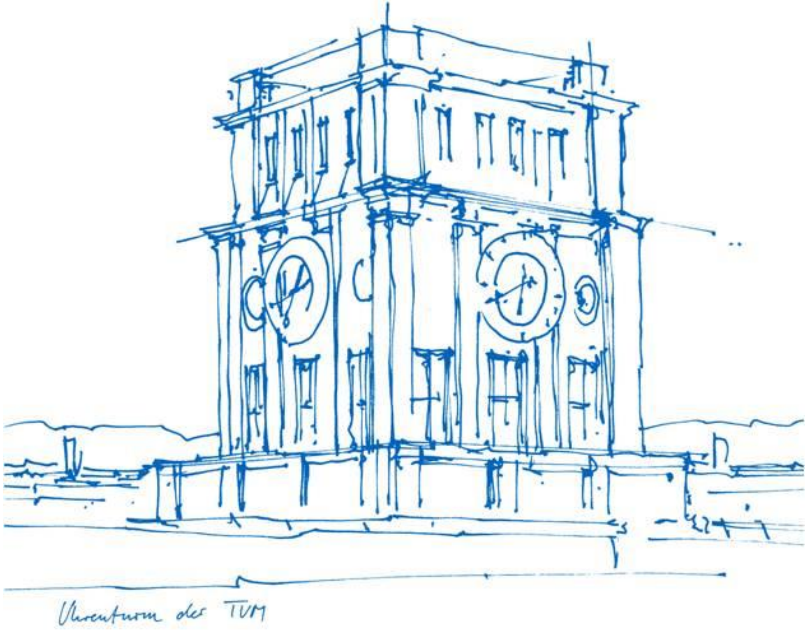
Information-optimal Abstaining for Reliable Classification of Building Functions

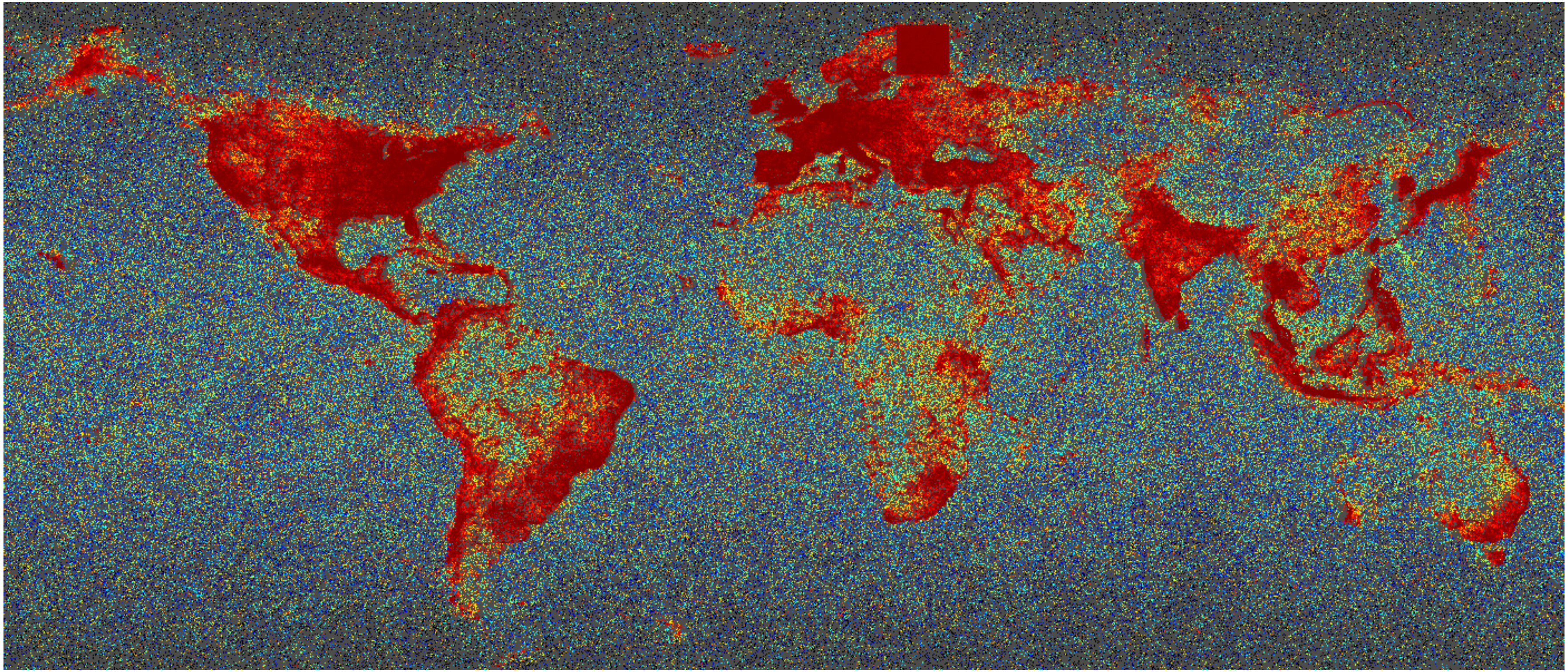
Gabriel Dax, **Martin Werner**

Technical University of Munich
Department of Aerospace and Geodesy
Professorship of Big Geospatial Data Management

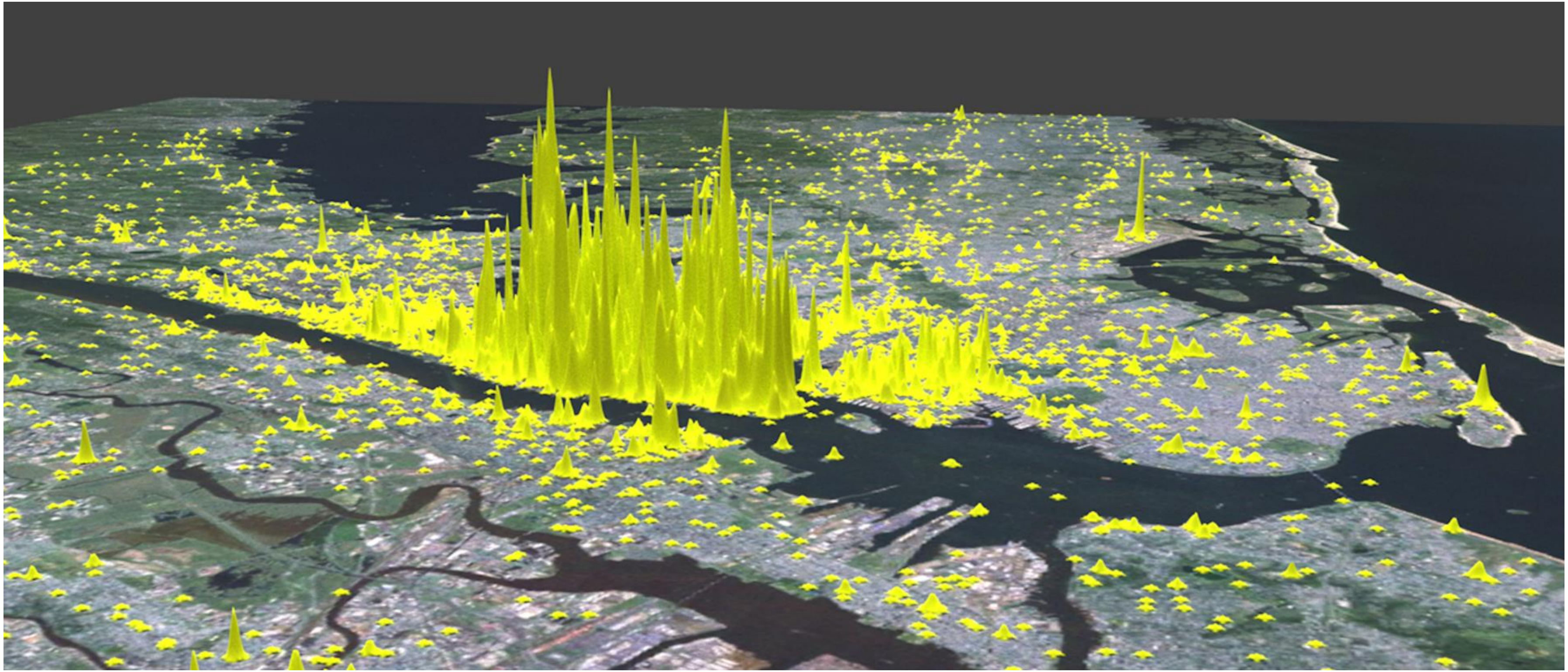
martin.werner@tum.de



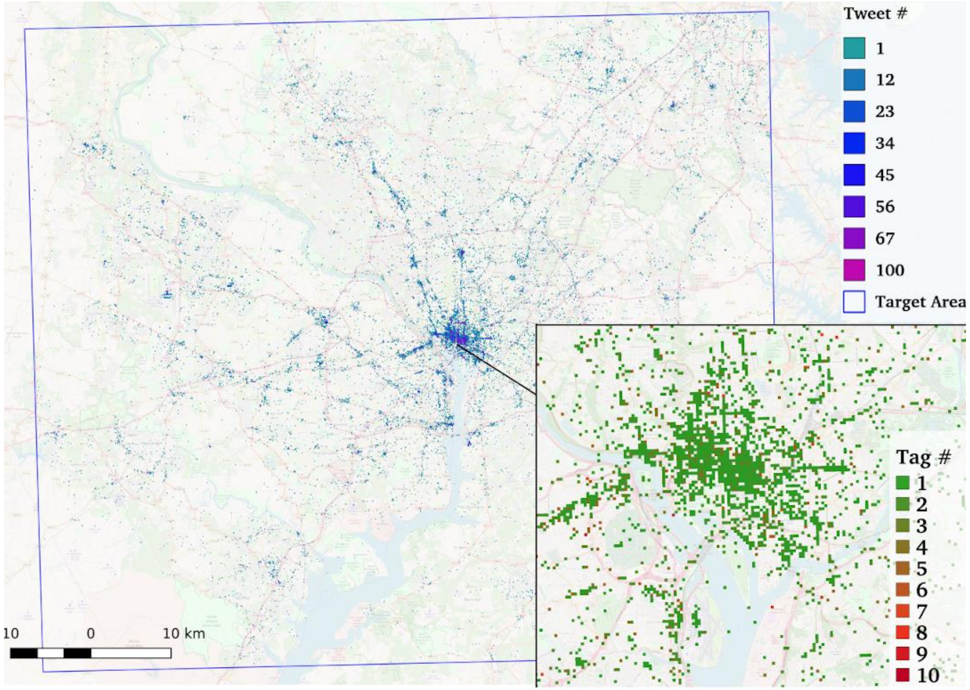
Social Media is a fascinating data source…



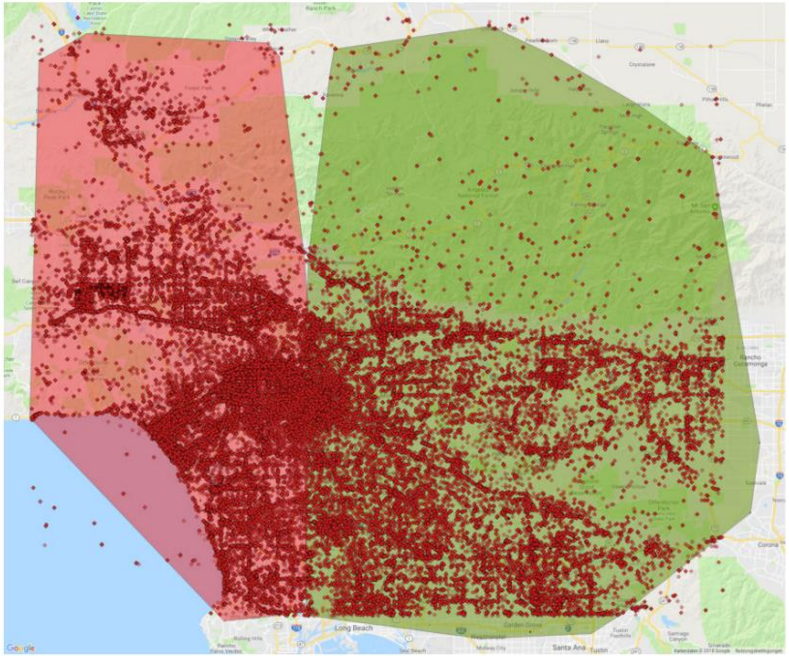
… but does it help?



… and if it helps, on which scale?

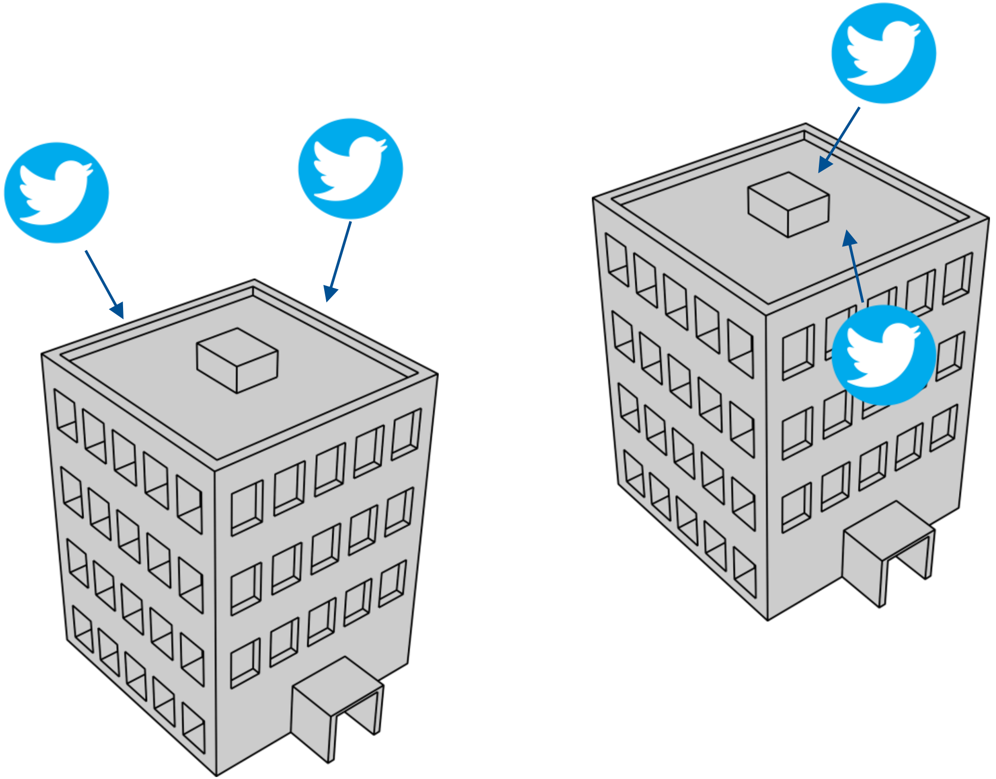


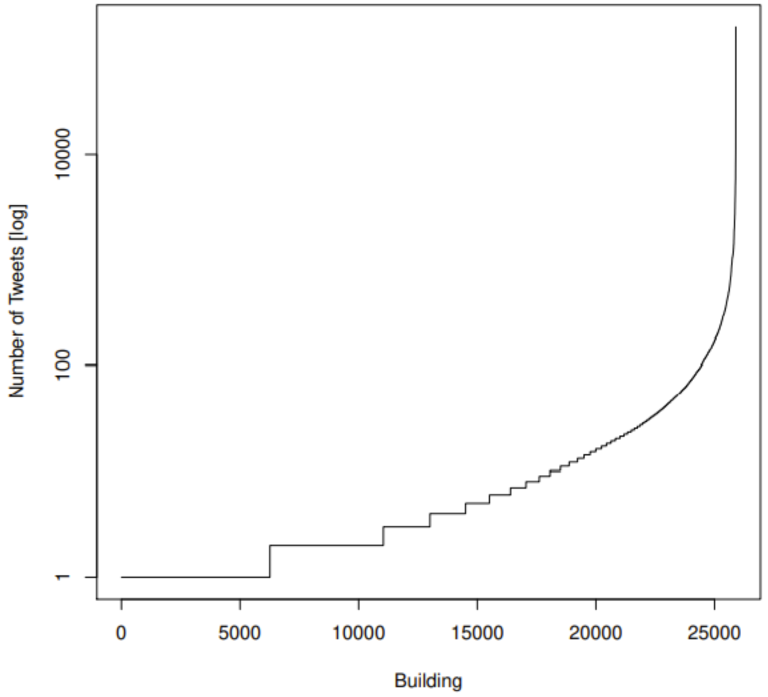
Leichter, A., Wittich, D., Rottensteiner, F., Werner, M., & Sester, M. (2018). IMPROVED CLASSIFICATION OF SATELLITE IMAGERY USING SPATIAL FEATURE MAPS EXTRACTED FROM SOCIAL MEDIA. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLII-4*, 335–342. https://doi.org/10.5194/isprs-archives-XLII-4-335-2018



(b) Tweets collected for the area of Los Angeles.

… for each building ? Of course not. But!





Observations

- Most tweets are not georeferenced
  ➔ we need to be very careful with removing data

- Most tweets are not relevant for the location of their origin
  ➔ we need to find the needle in the haystack

- Some tweets are very determining
  ➔ „I am at @applestorebln“

- Choose a simple, relevant, and valuable machine learning task with ground truth
  ➔Classify „residential“ (mainly for living) or „commercial“ (e.g., shops, malls) buildings

Main Question: How do we decide, when to use a tweet in the classification system?

Information-Optimal Abstaining (I)

Step 1: Introduce Decision Thresholds:

$$y_l = \arg\max\left(\frac{\phi_i(x_l)}{\tau_i}\right), 0 < \tau_i \le 1$$

Step 2: Introduce an additoinal class for instances that we don't classify

$$y_l = \begin{cases} \arg\max\left(\frac{\phi_i(x_l)}{\tau_i}\right) & \text{if } \max\left(\frac{\phi_i(x_l)}{\tau_i}\right) \ge 1 \\ m+1 & \text{else} \end{cases}$$

Step 3: Find τ in a parameter-free way.

Step 3.1: Remember Normalized Mutual Information: a measure for the degree of dependance of two random variables.

Difficult to compute? Yes, but…

$$\mathrm{NI}(T,Y) = \frac{I(T,Y)}{H(T)},$$

$$H(T) = -\sum_{i=1}^{m} P(T=i)\log_2 P(T=i)$$

$$I(T,Y) = \sum_{i=1}\sum_{j=1} P(T=i,Y=j)\cdot$$

$$\cdot\log_2 \frac{P(T=i,Y=j)}{P(T=i)P(Y=j)}$$

Information-Optimal Abstaining (I)

Step 3.2: Consider the extended confusion matrix…

T	**Y** 1	2	$\ldots$	m	$m+1$
1	c_{11}	c_{12}	$\cdots$	c_{1m}	$c_{1(m+1)}$
2	c_{21}	c_{22}	$\cdots$	c_{2m}	$c_{2(m+1)}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
m	c_{m1}	c_{m2}	$\cdots$	c_{mm}	$c_{m(m+1)}$

Step 3.3: … and approximate I(T,Y) from it and …

$$I(T,Y) \approx I(C) = -\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} c_{ij}\log_2\left(\frac{c_{ij}}{C_i\sum_{i=1}^{m}\frac{c_{ij}}{n}}\right)}{\sum_{i=1}^{m} C_i \log_2 \frac{C_i}{n}},$$

Step 3.4 … now run grid search and Powell algorithm to find a good candidate for τ

$$\tau^* = \arg\max \mathrm{NI}(t, y = \phi^\tau(x))$$

Results – Baselines

Classifier	Training		Test	
	Commercial	Residential	Commercial	Residential
Ridge	0.97	0.97	0.50	0.50
Perceptron	1.00	1.00	0.50	0.48
kNN	0.72	0.79	0.40	0.57
RF	1.00	1.00	0.49	0.53
X-Tree	1.00	1.00	0.50	0.52
SVC-L2	0.99	0.99	0.50	0.50
SVC-L1	0.91	0.91	0.50	0.51
ElasticNet	0.77	0.70	0.54	0.42
MN-NB	0.99	0.99	0.52	0.52
SVC-L1/2	0.94	0.94	0.50	0.50

Results – Abstaining

Classifier	Abstain-Rate	Commercial		Residential	
		Precision	Recall	Precision	Recall
MN-NB1	63%	0.54	0.19	0.57	0.21
MN-NB2	72%	0.53	0.14	0.58	0.17
MN-NB3	89%	0.55	0.04	0.62	0.09
SGD-L2	99%	0.17	0.00	0.83	0.01
SGD-L1	96%	0.56	0.01	0.76	0.04

High Abstaining rates lead to high precision with low recall.

Put simply:

For a few buildings, we can be pretty sure that they are residential…

Ensembles…

Classifier	Abstain-Rate	Training		Test	
		Commercial	Residential	Commercial	Residential
BIRP	54 %	0.60	0.78	0.82	0.26
HRF1	58 %	0.70	0.23	0.75	0.38
AVE	-	0.59	0.85	0.73	0.41
AVE-A	16 %	0.61	0.72	0.75	0.37
AVE-F1	-	0.59	0.86	0.74	0.52
AVE-F1-A	16 %	0.61	0.72	0.75	0.37

Model Parameters are **crucial**! Best Individual Residential Precision (BIRP) and Highest Residential F1 (HRF1) **are reached by the same model** (Multinomial Bayes), but with extremely different smoothing parameter (smallest and largest).

Simple model blending reveals even more that the abstaining mechanism abstains **differently** across classifiers, such that we successfully increase performance with a variant of bagging.

Conclusion / Remarks

Take-Home Message:

- Social Media Data is a difficult yet valuable data source
- There is interesting information to find
- Blind machine learning (end-to-end) does not work (50% accuracy and precision)

Reproducibilty:

- The source code of all of this work is available: https://github.com/mwernerds/agile21_abstaining
- For reproduction, we **cannot share the data** (ethical, privacy, and legal bounds on sharing social media observations), but we provide both the relevant tweet IDs and synthetic data proxies built from openly available literature

Thanks / Questions

Gabriel Dax, Martin Werner
Technical University of Munich
Department of Aerospace and Geodesy
Professorship of Big Geospatial Data Management

Website: www.bgd.lrg.tum.de/