

Expedia Kaggle Competition

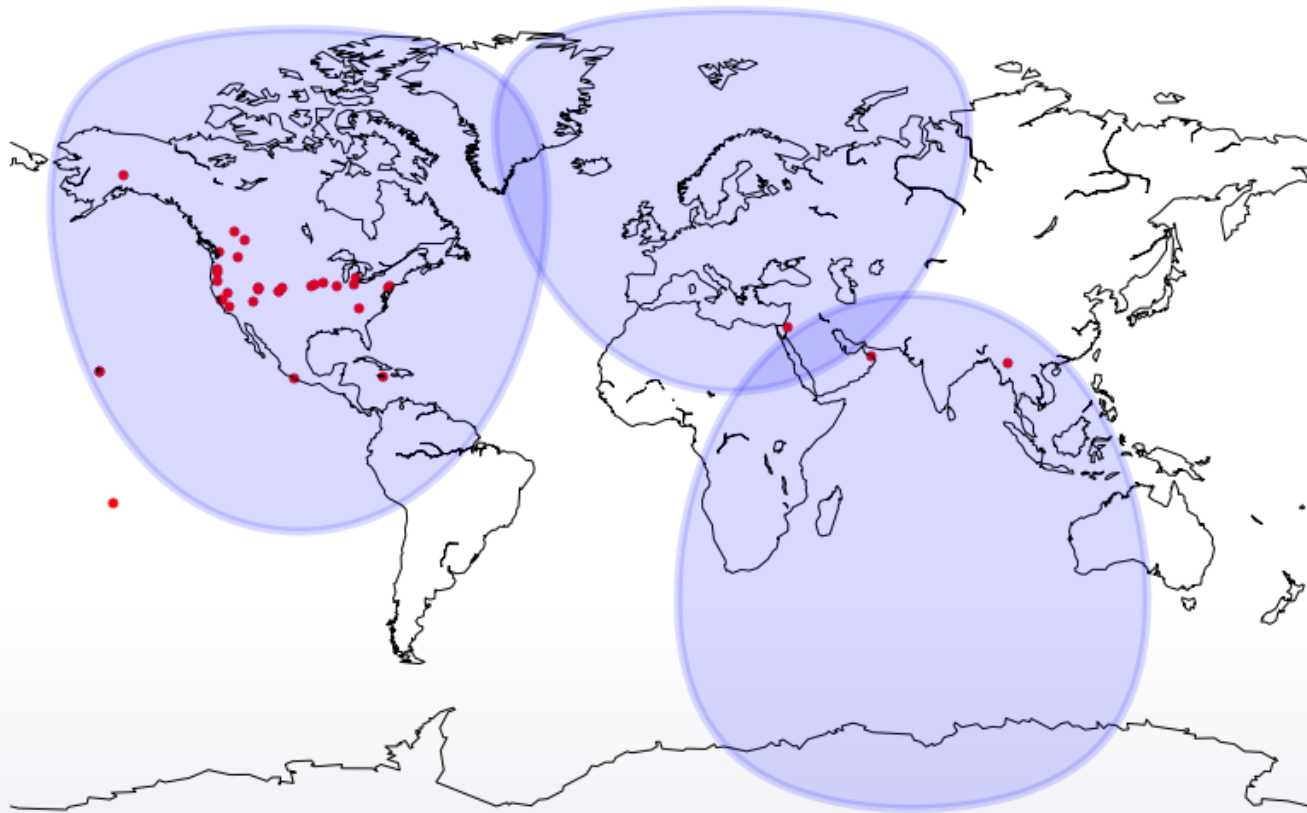


Data leakage

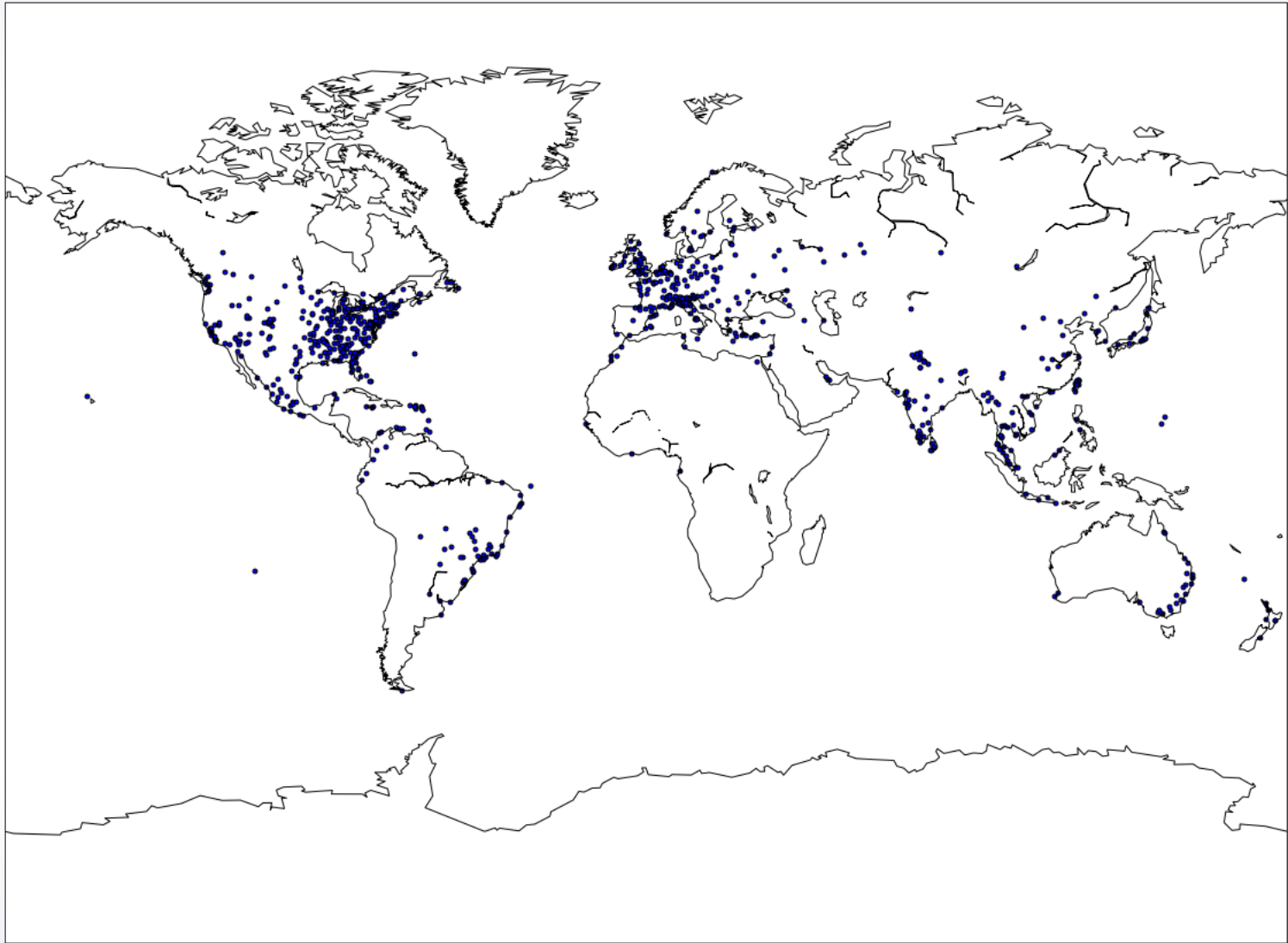
- destination_distance - user_city pair is a leak to true hotel location. A lot of matches between train and test.
- How to improve on that?
- Features based on counts on corteges of such nature
- Try to find the true coordinates

Spherical geometry

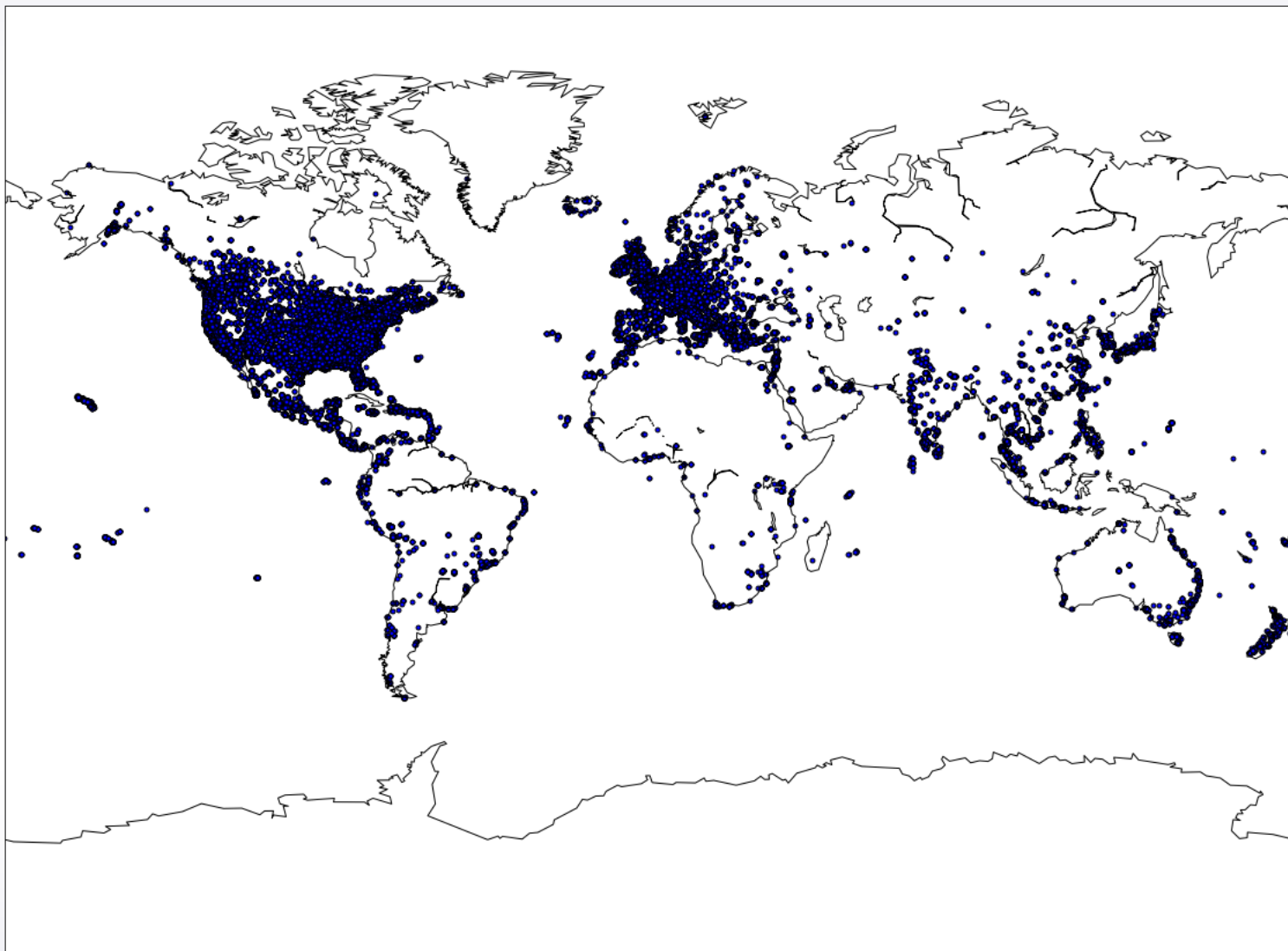
$$d = 2r \arcsin \left(\sqrt{\text{hav}(\varphi_2 - \varphi_1) + \cos(\varphi_1) \cos(\varphi_2) \text{hav}(\lambda_2 - \lambda_1)} \right)$$
$$= 2r \arcsin \left(\sqrt{\sin^2 \left(\frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$



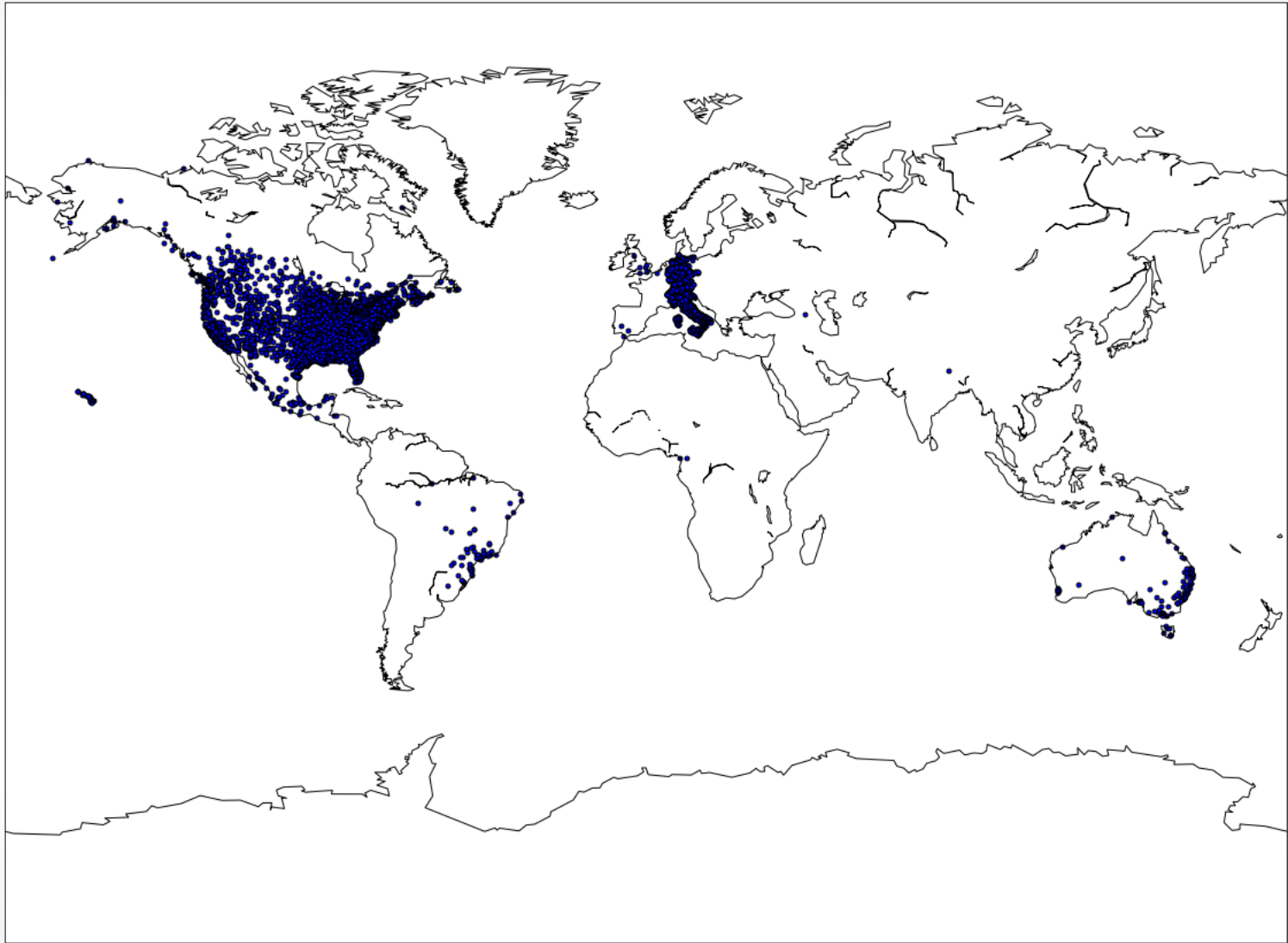
Hotel cities. Old version



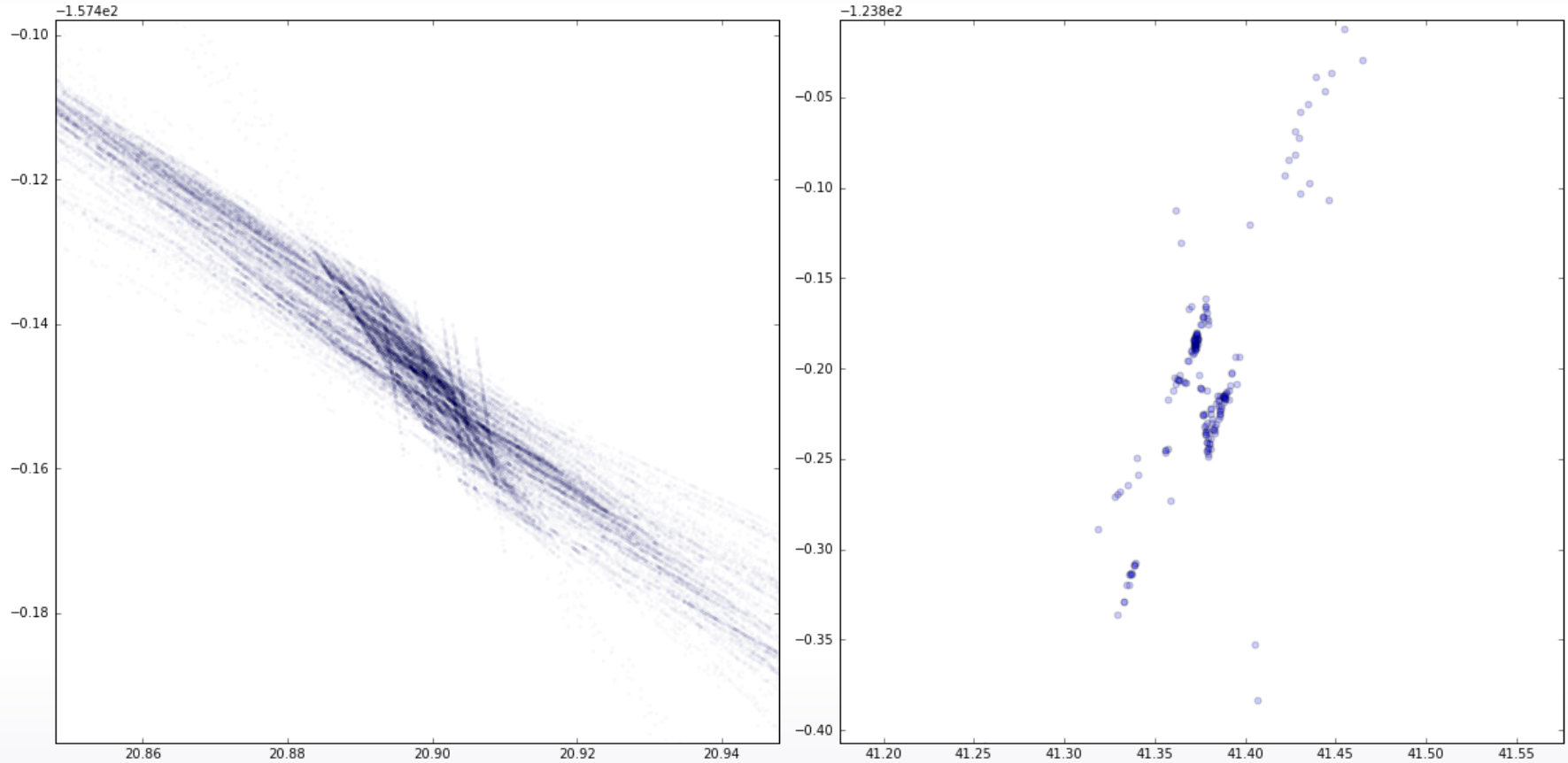
Hotels cities. New version



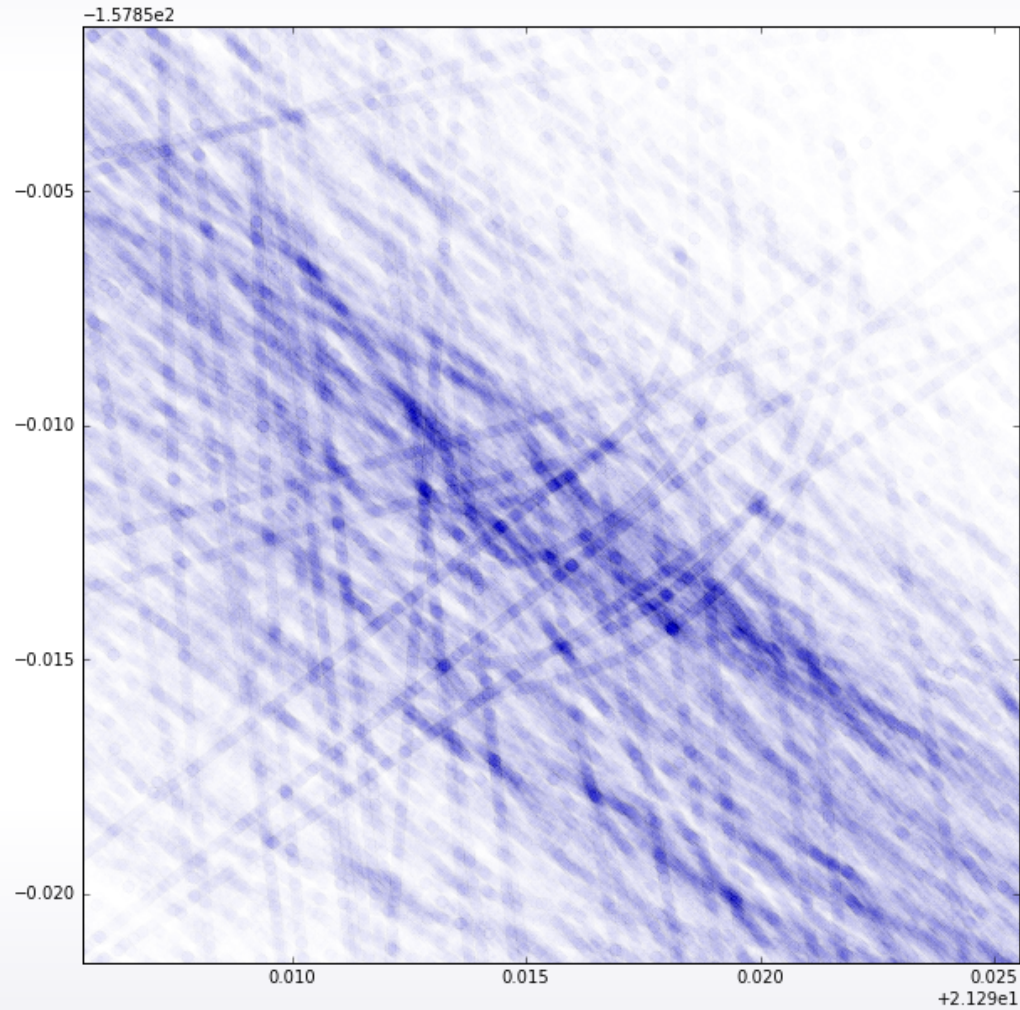
User cities. New version



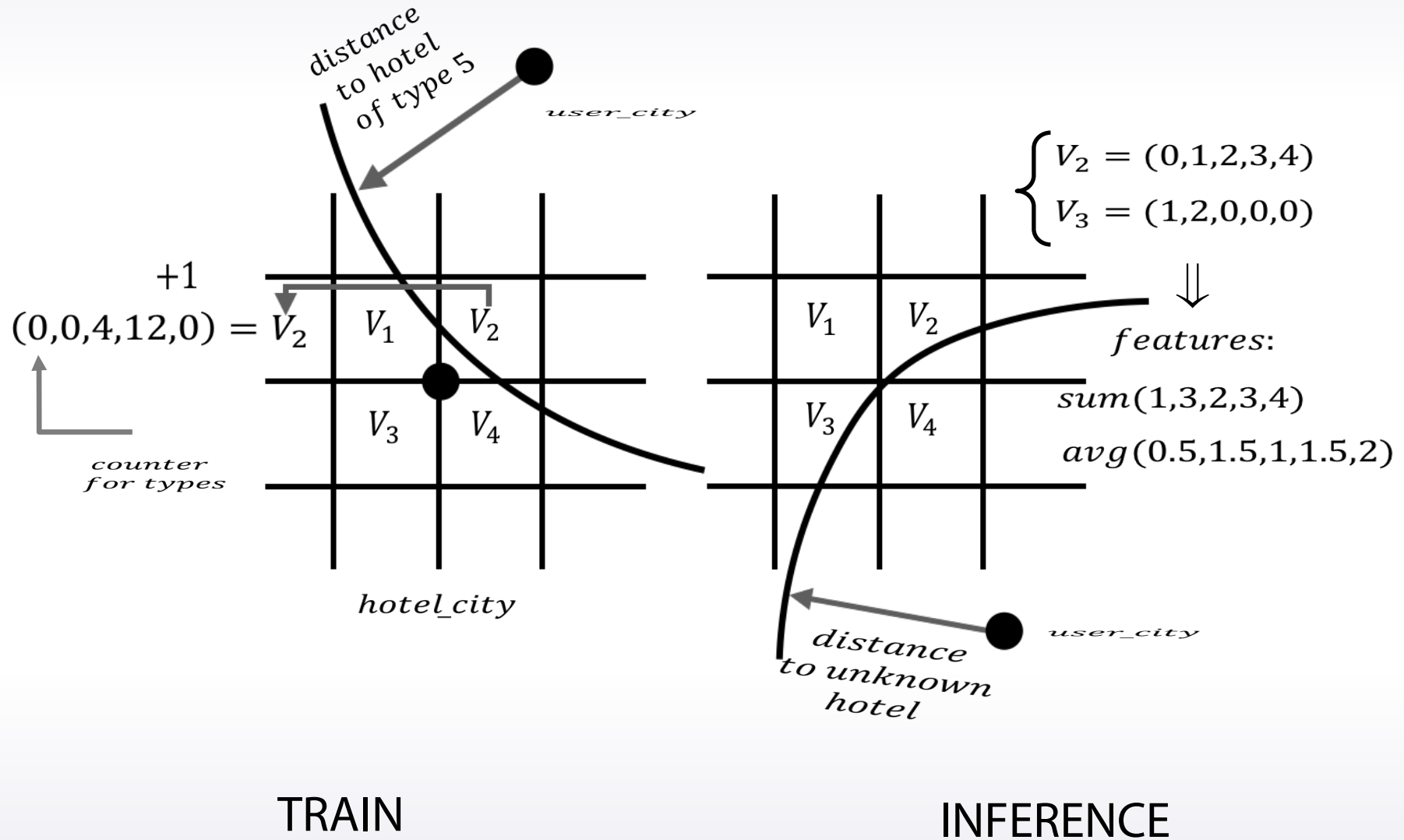
Trying to find the true coordinates of hotels (fail?)



Trying to find the true coordinates of hotels (fail?)



Counters in grid cells








Final model

- Out-of-fold feature generation. 2013<->2014
- Xgboost
- 16 hours of training

Results

- Public – 3rd
- Private - 4th

#	△pub	Team Name	Kernel	Team Members	Score [?]	Entries	Last
1	—	idle_speculation		 ★★★★★	0.60219	1	1y
2	—	beluga		 ★★★★★	0.53218	64	1y
3	▲1	Victor		 ★★★★★	0.53134	50	1y
4	▼1	Ala Mode		  ★★★★★ ★★★★★	0.52995	26	1y