Aalto University
School of Science
and Technology

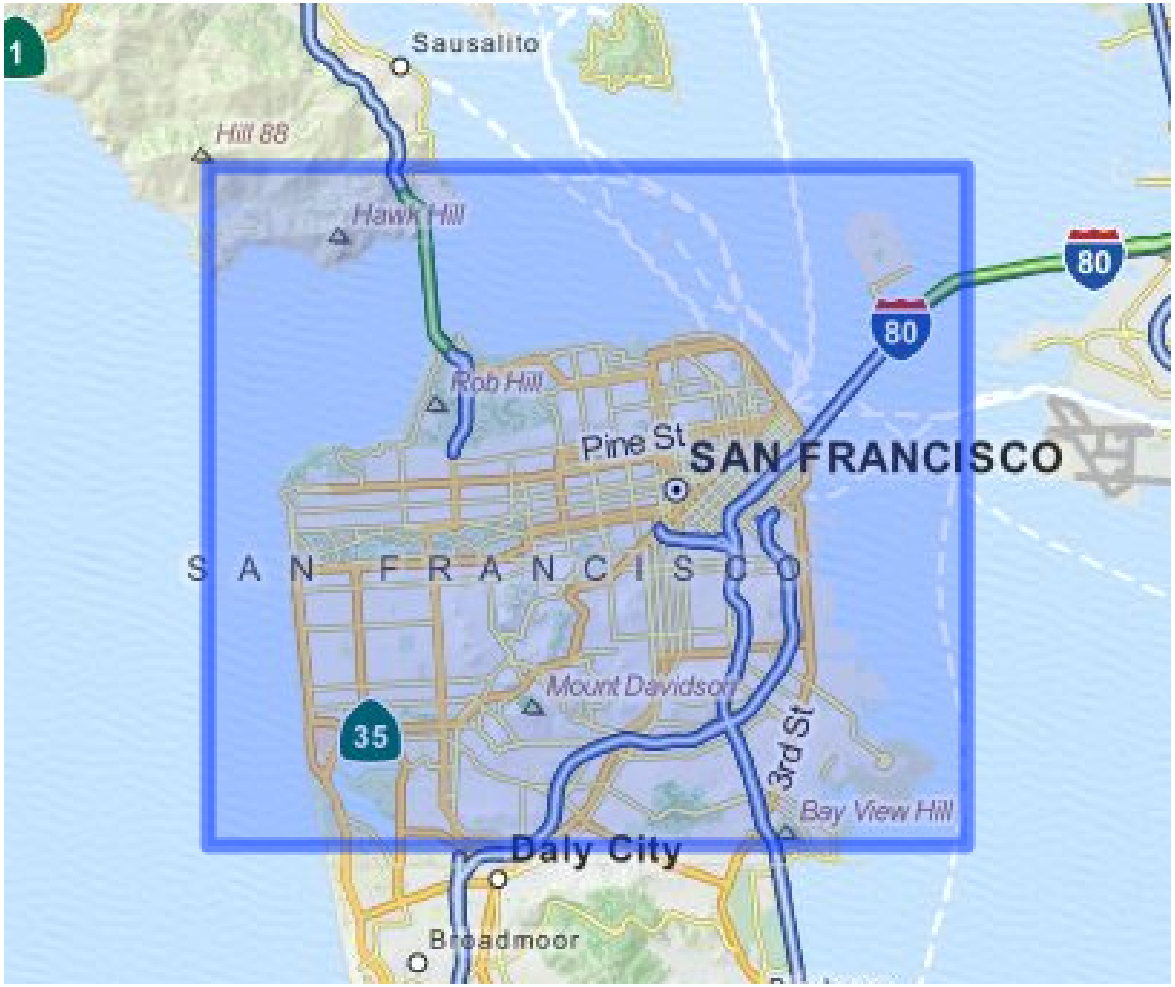
Exploring San Francisco's geography using geo-tagged photos

Aristides Gionis Michael Mathioudakis
Géraud Le Falher

December 10, 2013

Outline

Dataset



San Francisco: 780K photos and 140K tags since January 2008

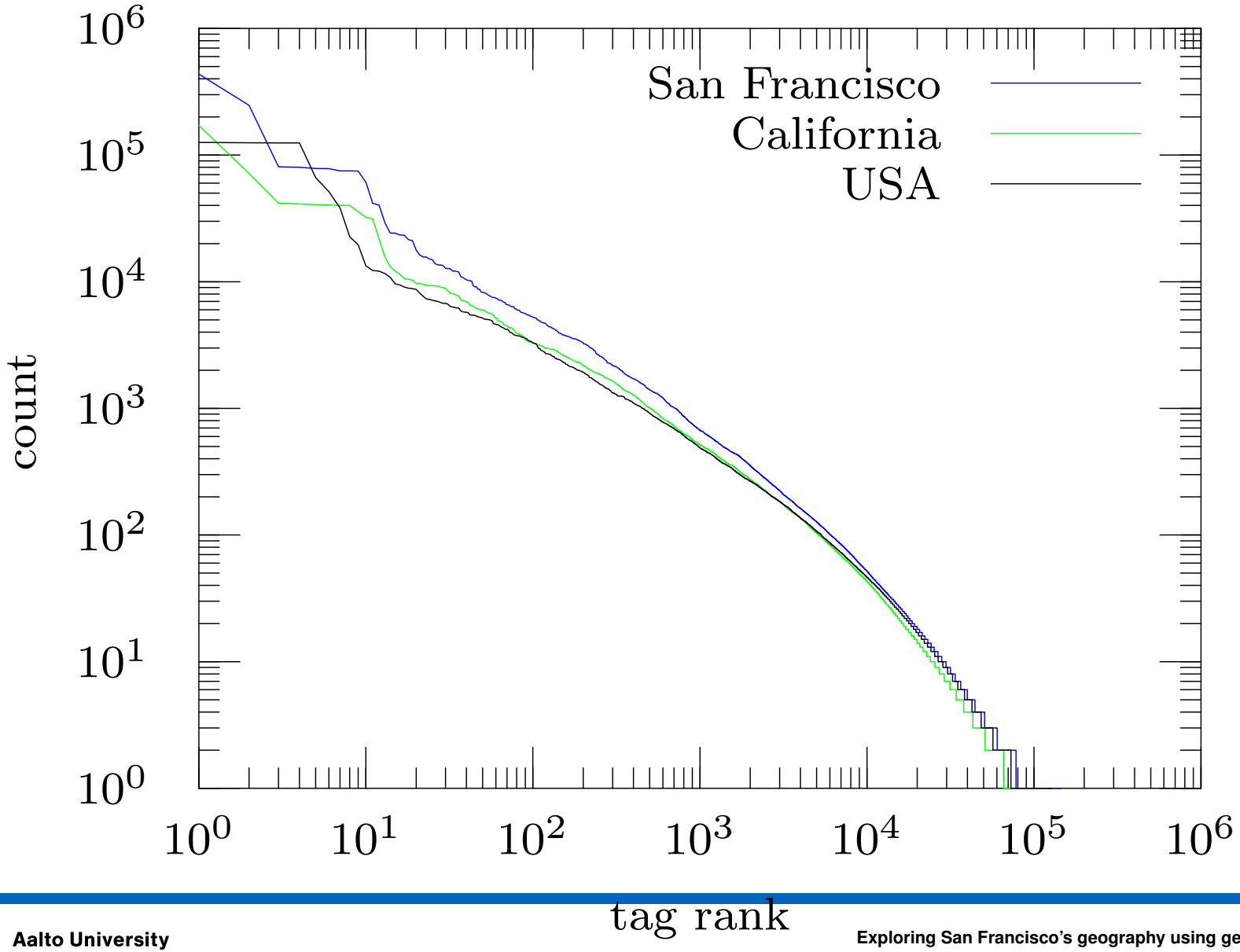
Photo metadata

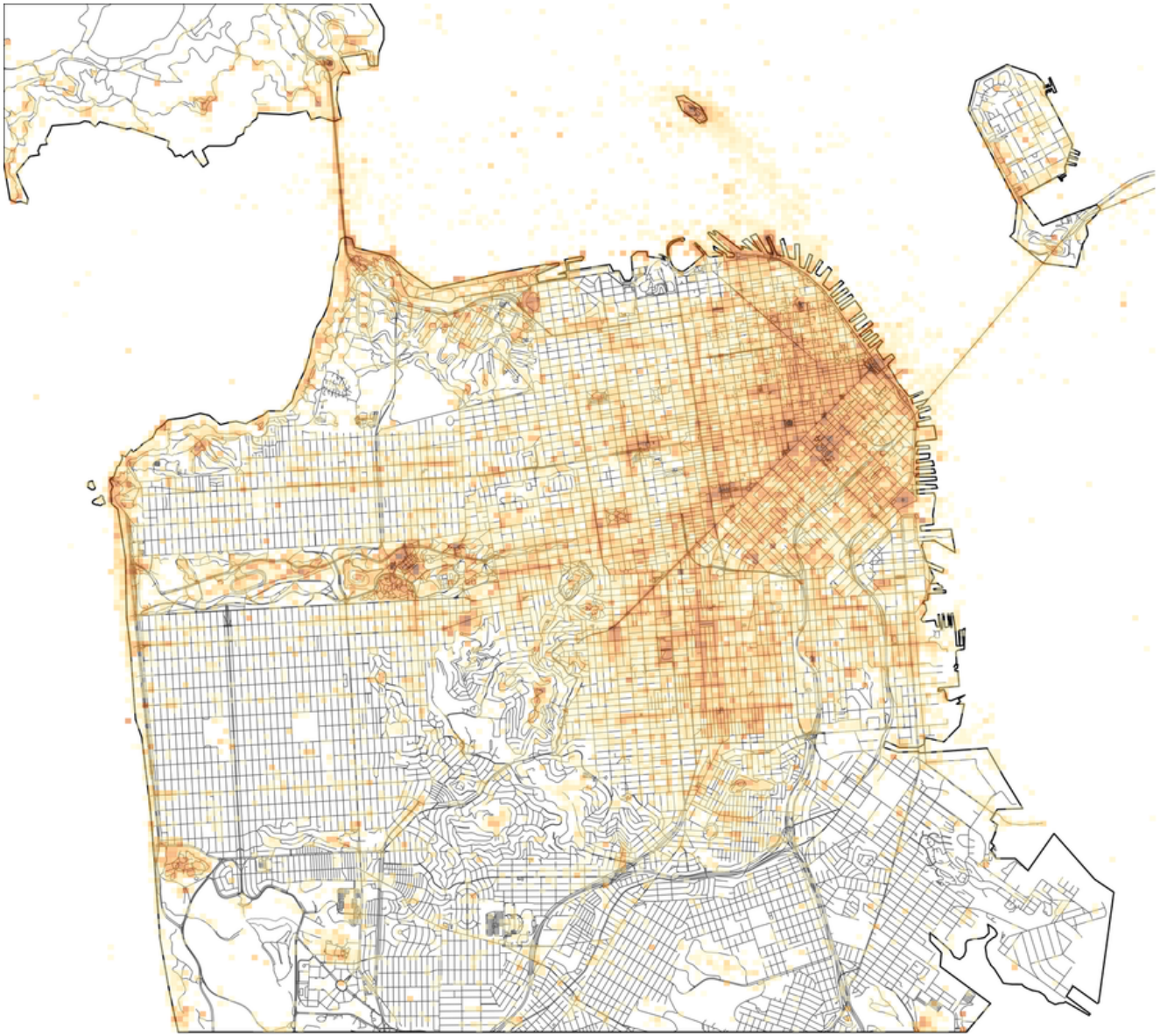
```
"loc"      : [-122.392501, 37.77515],
"taken"    : "2008-03-24 14:55:40",
"user_id"  : "37417902@N00",
"tags"     : ["sanfrancisco", "california",
              "bridge", "chinabasin"]
```

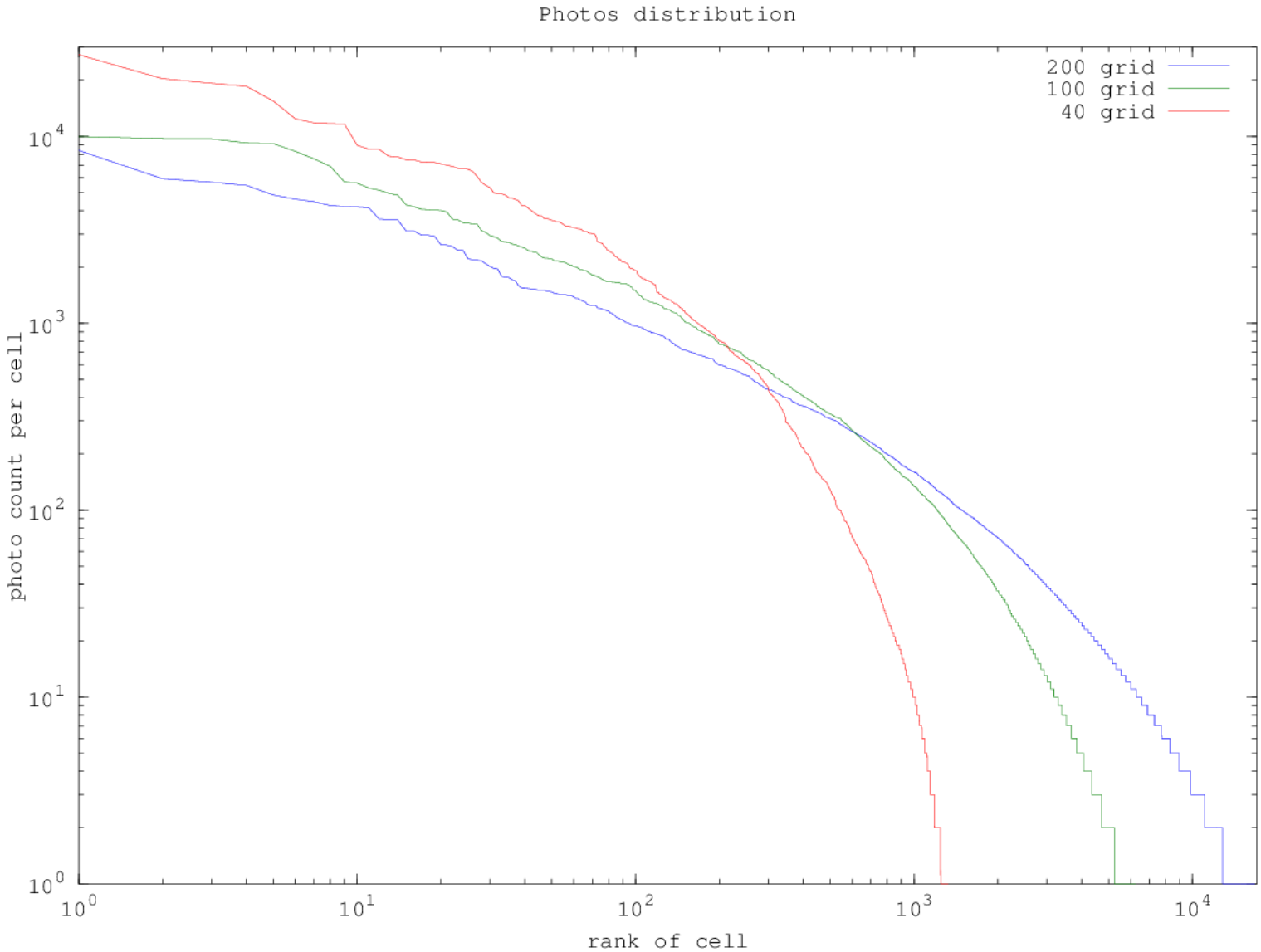
Problem

- The goal is to find association rules
  - tag: in which location it appears
  - locations: which tags describe it
  - city: which locations are interesting
- Applications
  - Flickr website
  - Tourism

Exploration of data



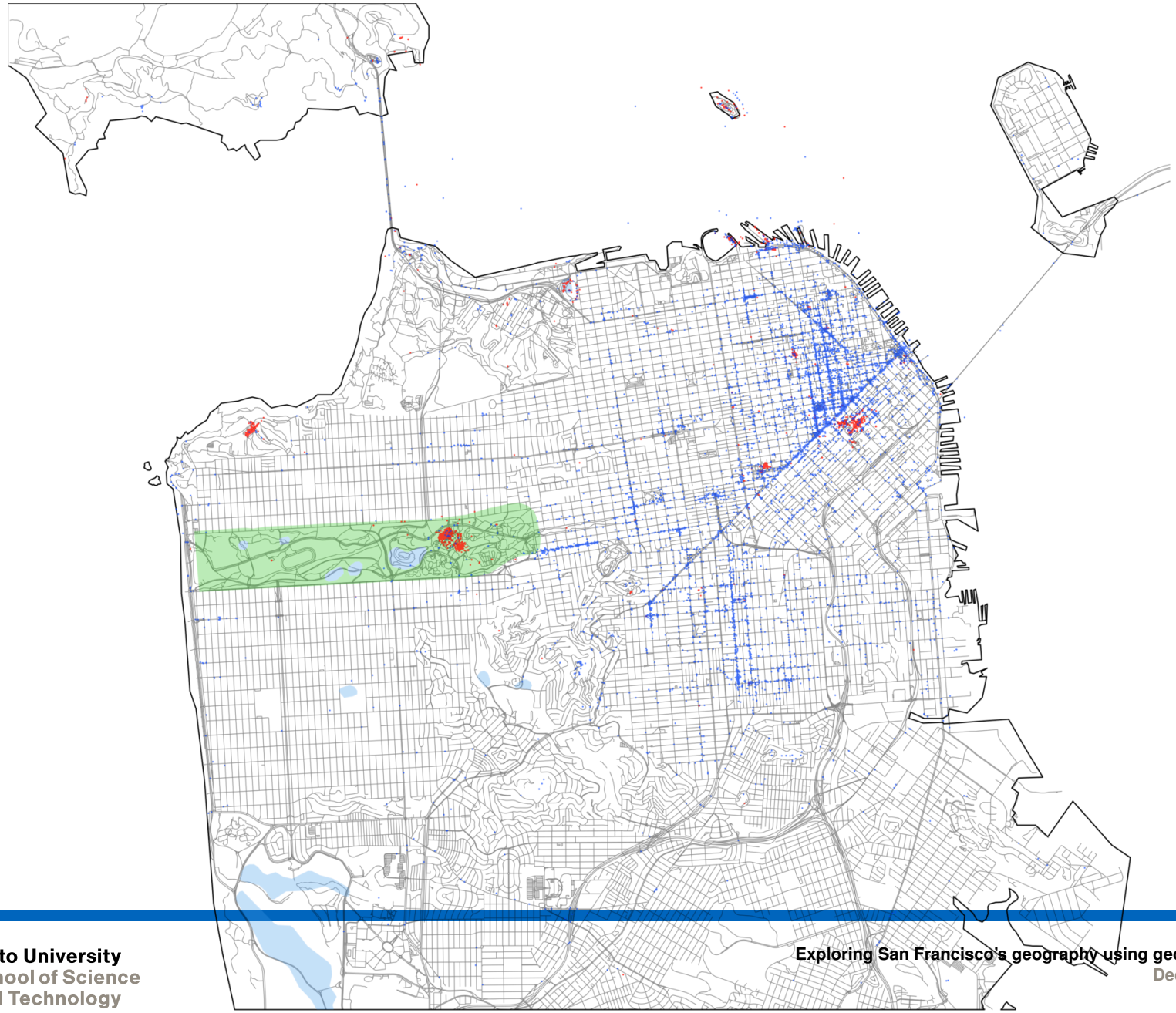



Photos distribution
200 grid
100 grid
40 grid
photo count per cell
rank of cell
10^0
10^1
10^2
10^3
10^4

Preliminary cleaning

- Avoid spurious photos by removing tags of the same user set almost in the same place at almost the same time.
- Keep only tags used enough times by enough users over enough time.

Statistics

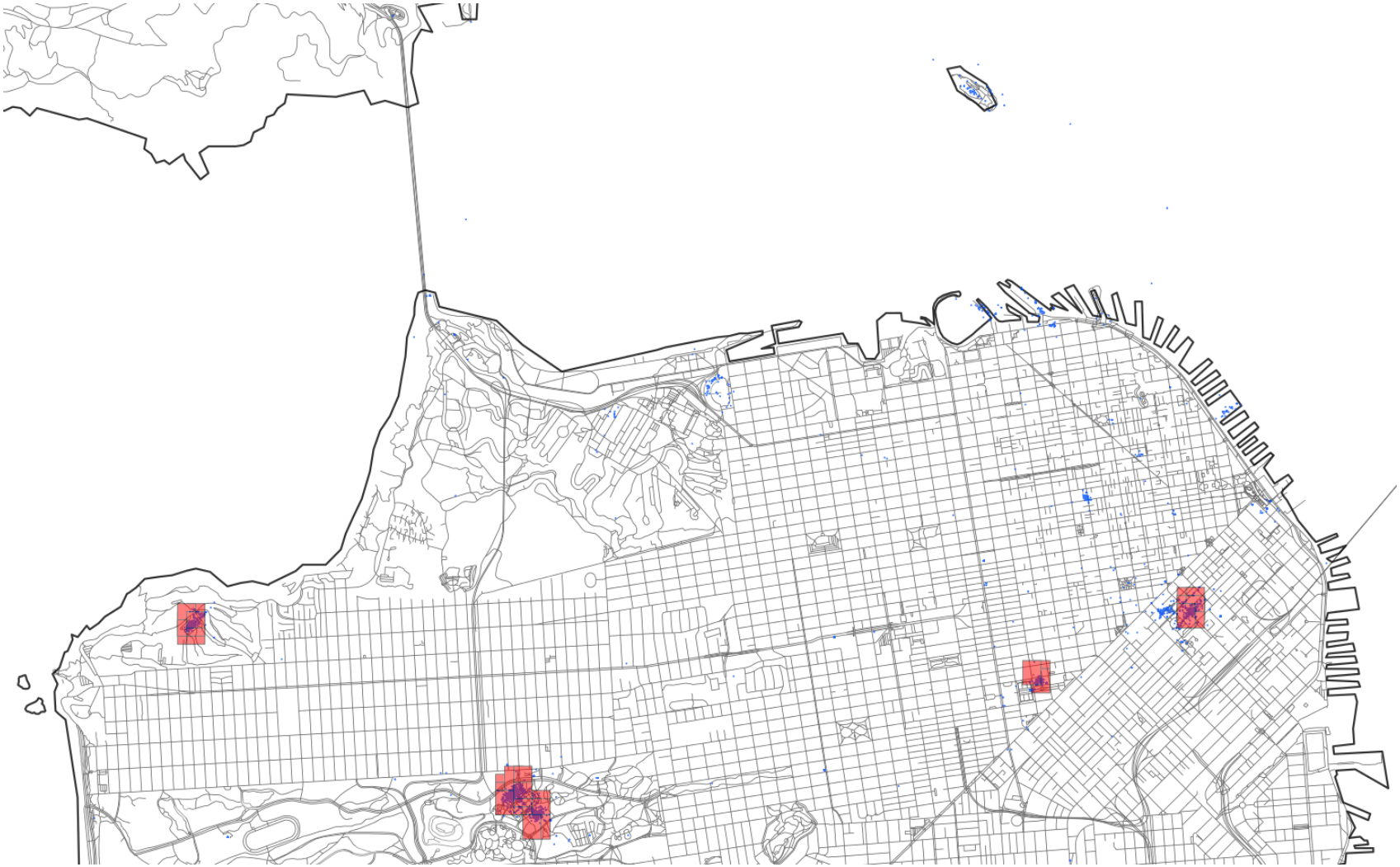


- First, Entropy and Kullback Leibler divergence with background photos:
  $H(museum) = 3.565$ $H(street) = 6.875$

Lowest entropy	Highest divergence
californiaacademyofsciences	zoo
conservatoryofflowers	treasureisland
sfmoma	conservatoryofflowers
deyoung	oceanbeach
deyoungmuseum	ucsfschoolofdentistry
ucsfschoolofdentistry	ucsf
attpark	japantown
cityhall	sfmoma

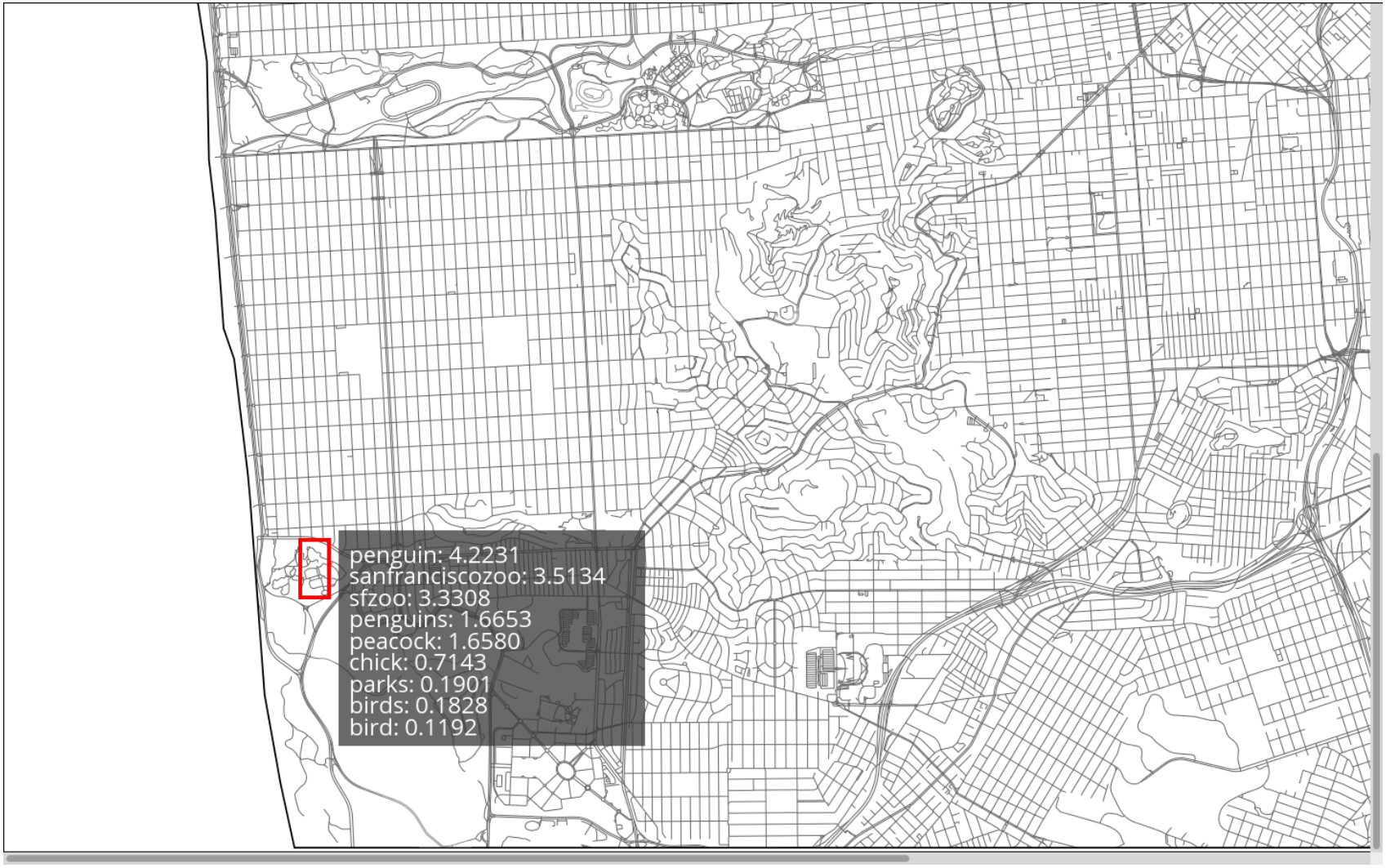
- But it provide only one value for each tag, i.e. it loses spatial information
- Enter Kulldorff Spatial Scan Statistic: for a region R, compute the KL divergence considering only R and its complementary: $t(R) \log \frac{t(R)}{b(R)} + (1 - t(R)) \log \frac{1-t(R)}{1-t(R)}$
- Computed for all supported tags over all reasonnably sized regions

Where are the museums?



What is this place?


augusterodin: 5.4883
legionofhonor: 5.3298
94121: 2.5414
museodearte: 1.9376
fineartsmuseum: 1.7584
muse: 1.2857
konst: 1.1895
sztuka: 1.1895
sining: 1.1895
musee: 1.0414
muzeum: 1.0406
kunst: 1.0221
museudearte: 1.0201
museu: 0.9582
museo: 0.7706
artes: 0.7333


penguin: 4.2231
sanfranciscozoo: 3.5134
sfzoo: 3.3308
penguins: 1.6653
peacock: 1.6580
chick: 0.7143
parks: 0.1901
birds: 0.1828
bird: 0.1192

Where to go in San Francisco?

Future work

- Evaluate result using ground truth: tourist guide, manual annnotation
- Use less parameters and more information from the data
- Extend to different scales: city, region, country, world
- Consider more type of entities: time, users


Thank you
Questions?