Statistics Finland

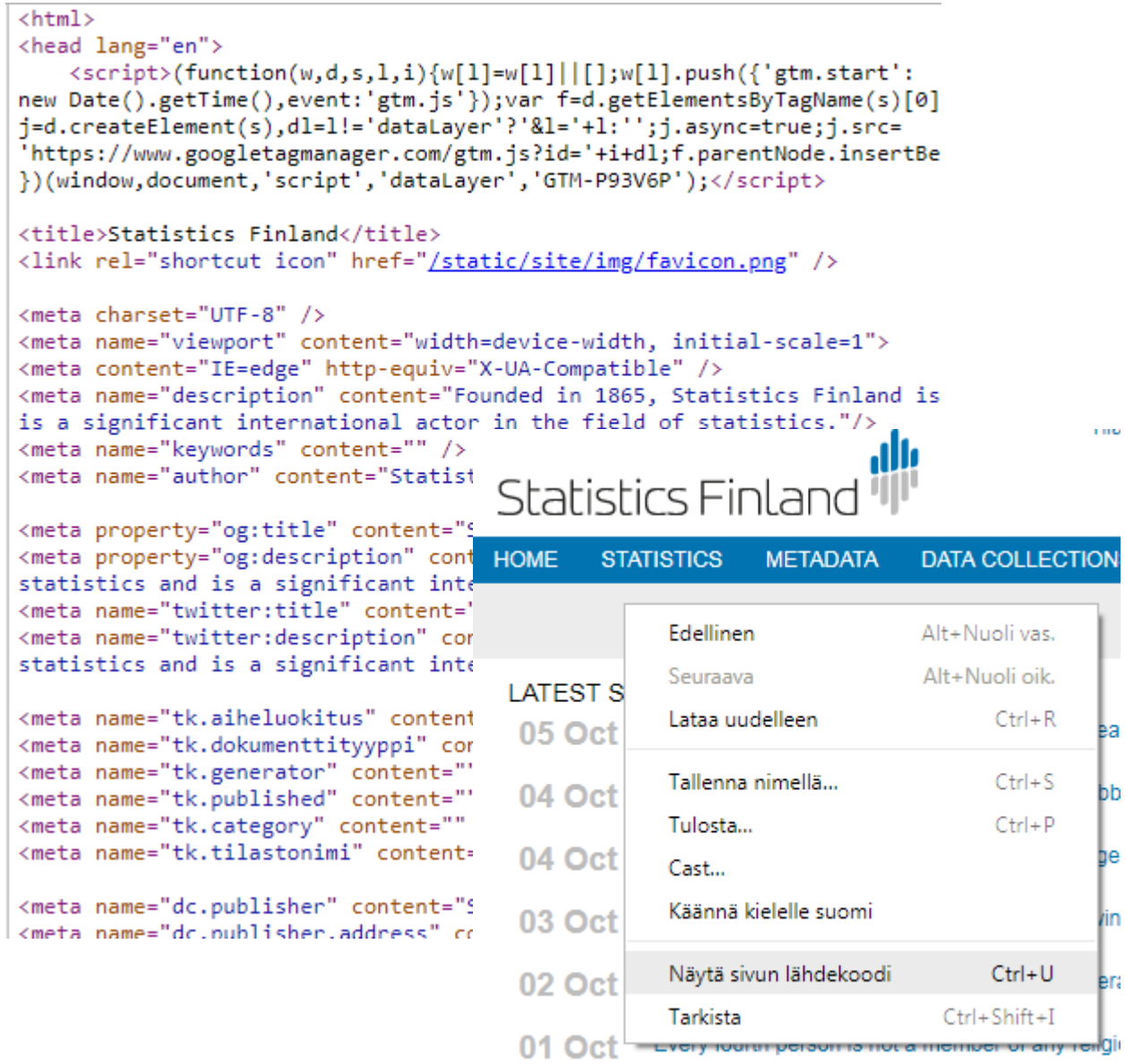
Scraping rental holiday homes in Finland

Matti Kokkonen / Katja Löytynoja

EFGS 2018 October 17th

What is web scraping?

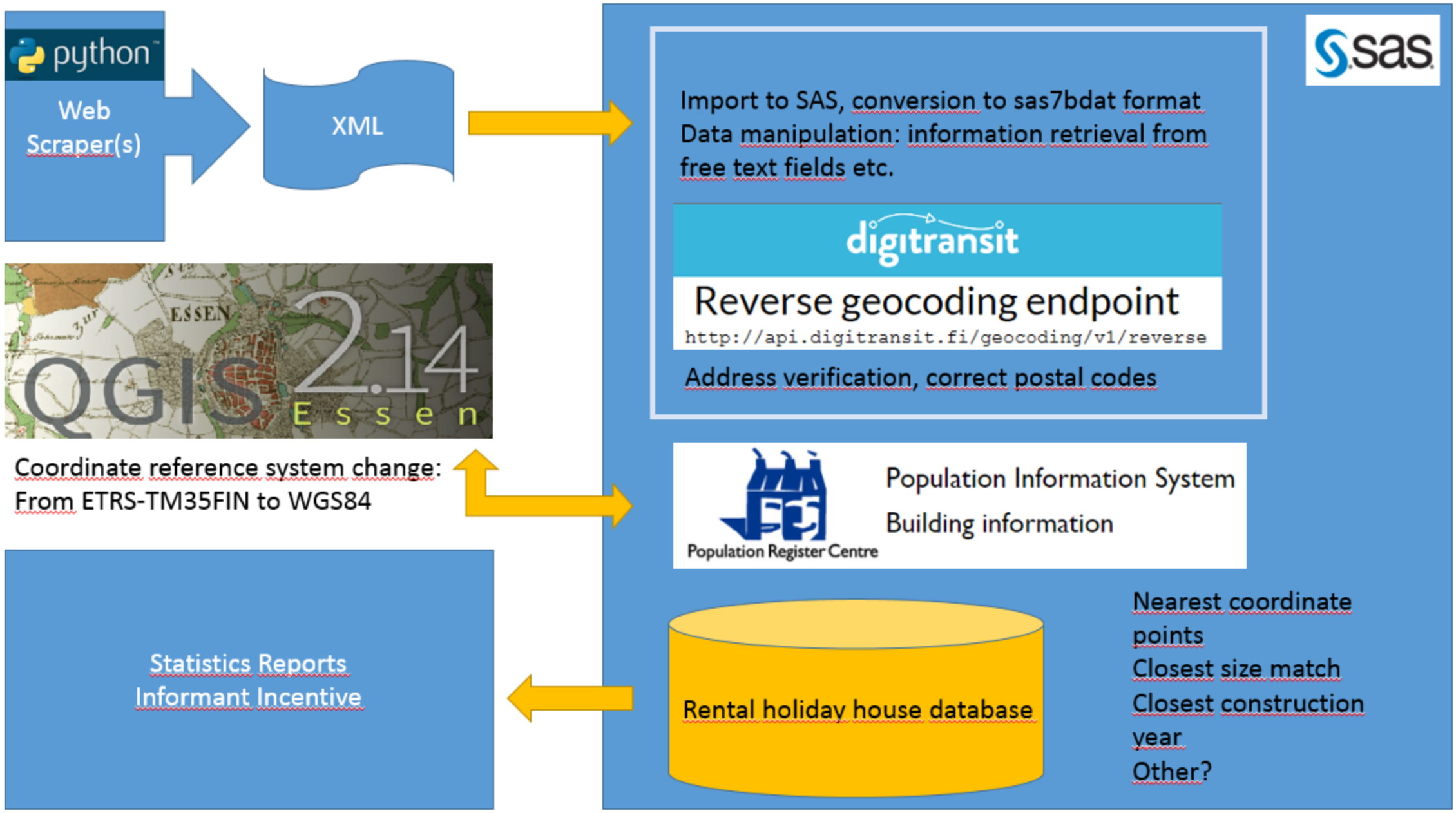
- Automated retrieval of elements from web site's source code
- Both visible and invisible for regular users
- Packages available for multiple programs, for example Python and R
- Regular html easy to scrape, JavaScript requires web browser automation



- Fast and relatively easy access to up-to-date data
- Minimizes perceived response burden
- Load on the server, immaterial rights -> scraping ethics
  - Identify yourself and provide contact information (user agent string)
  - Respect robots.txt and terms of use
  - Scrape during quiet periods
  - Do not reuse data commercially
  - Return value
  - Ask a permission
  - Alternatively use API or negotiate a direct access to server
- Statistical legislation and GDPR

Why and what?

- New statistics where target group identified through web scraping
- Largest domestic booking agencies and market platforms
- Started out with 11 scrapers, 5 still in use
- Variables scraped
  - Address
  - Coordinates
  - Construction year
  - Area (m2)
  - Number of beds
  - Rental price (usually per week or weekend) if available
  - Advertiser (used to identify doubles)
  - URL to advertisement (an unique id)

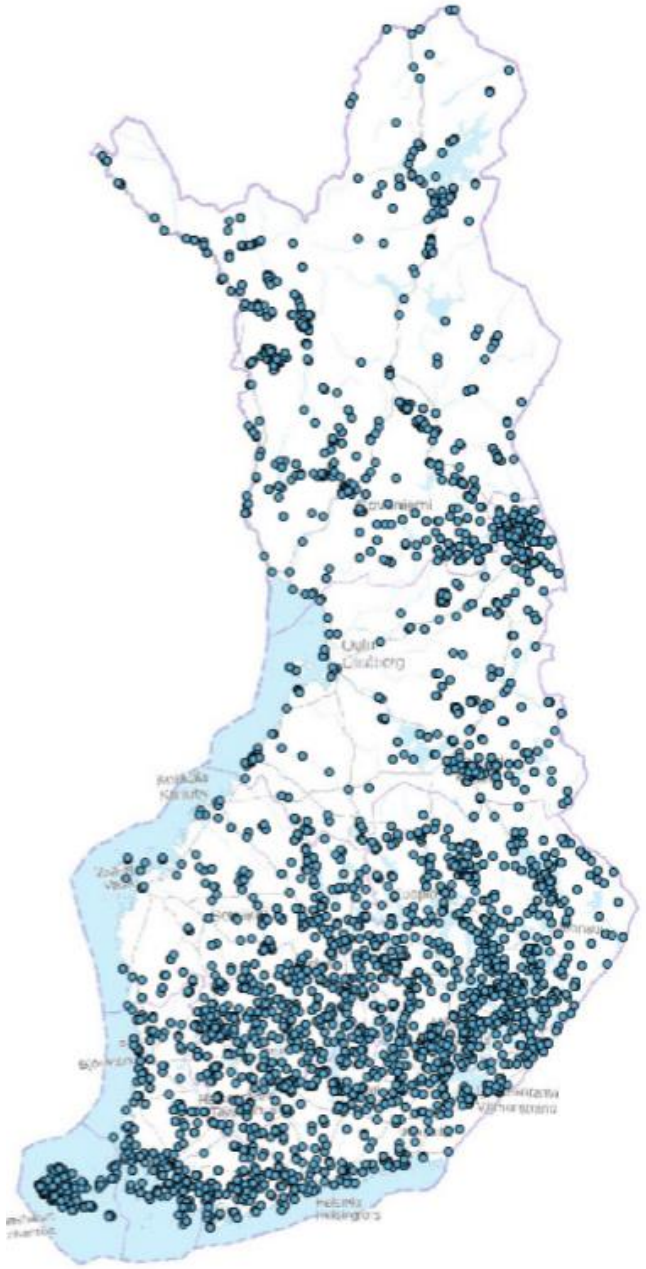

python
Web Scraper(s)
XML
Import to SAS, conversion to sas7bdat format
Data manipulation: information retrieval from free text fields etc.
sas
digitransit
Reverse geocoding endpoint
http://api.digitransit.fi/geocoding/v1/reverse
Address verification, correct postal codes
QGIS 2.14 Essen
Coordinate reference system change:
From ETRS-TM35FIN to WGS84
Population Register Centre
Population Information System
Building information
Statistics Reports
Informant Incentive
Rental holiday house database
Nearest coordinate points
Closest size match
Closest construction year
Other?

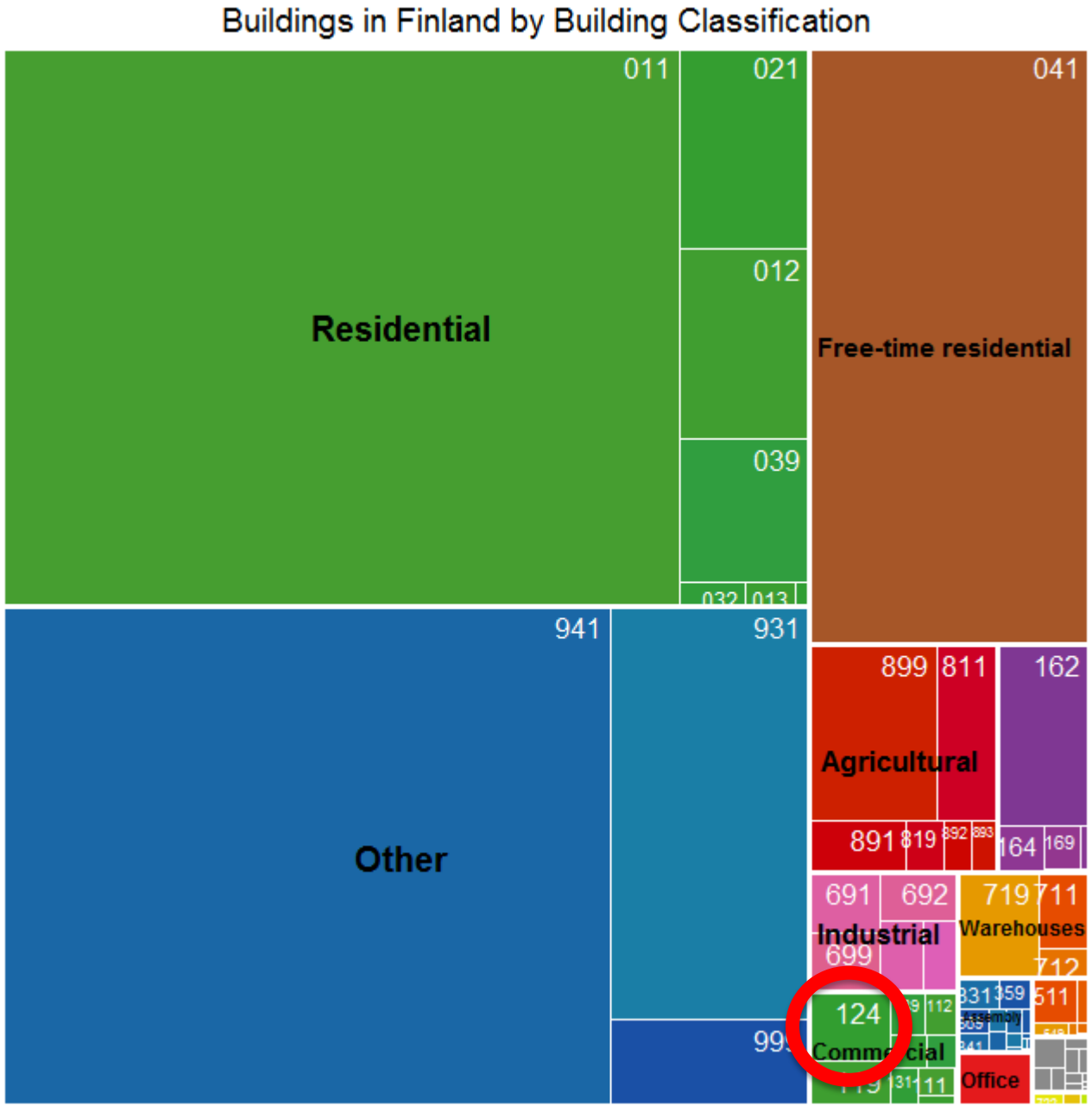
Raw data

Coordinates	Address		
	Not available	Available	Total
Not available	0.6 %	14.7 %	15.3 %
Available	50.5 %	34.2 %	84.7 %
Total	51.1 %	48.9 %	100.0 %

After reverse geocoding

Coordinates	Address		
	Not available	Available	Total
Not available	0.7 %	1.4 %	2.1 %
Available	0.1 %	97.8 %	97.9 %
Total	0.8 %	99.2 %	100.0 %





Target group?

Classification of Buildings 1994

011	One-dwelling houses
012	Two-dwelling houses
039	Other blocks of flats
041	Free-time residential buildings
124	Rental holiday cottages and flats
811	Cowsheds, pighouses, hen-houses, etc.
899	Other buildings in agriculture, forestry and fishing
931	Sauna buildings
941	Outbuildings
999	Buildings n.e.c.

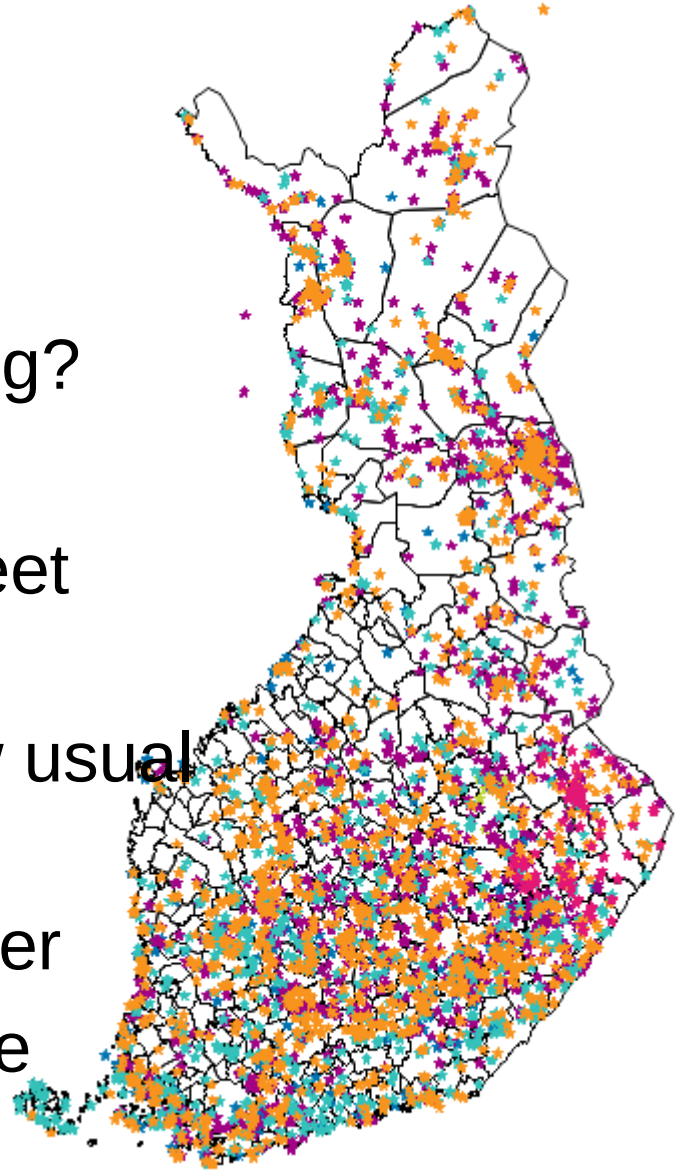
Statistics Finland

Käyttötarkoitus				
rakkayt_vtj	Frequency	Percent	Cumulative Frequency	Cumulative Percent
011	69	10.53	69	10.53
012	3	0.46	72	10.99
021	1	0.15	73	11.15
041	252	38.47	325	49.62
119	1	0.15	326	49.77
124	36	5.50	362	55.27
129	1	0.15	363	55.42
139	4	0.61	367	56.03
151	1	0.15	368	56.18
162	7	1.07	375	57.25
222	1	0.15	376	57.40
511	1	0.15	377	57.56
699	1	0.15	378	57.71
719	3	0.46	381	58.17
811	1	0.15	382	58.32
891	1	0.15	383	58.47
892	1	0.15	384	58.63
899	4	0.61	388	59.24
931	119	18.17	507	77.40
941	145	22.14	652	99.54
999	3	0.46	655	100.00

- A sample of the data was linked to Population Information System's Building Registry by
  - Nearest coordinate match
  - Closest size match
  - Closest construction year match
  - Postal code was used for subsetting Building Registry
  - The closest match for all of the above was selected
- Subsetting by postal code and street name could be quicker

Quality

- Coordinate precision and accuracy
  - WGS84 vs. building register ESTR-TM35FIN
  - How coordinates are created for the add? Geocoding?
  - From three to seven decimals
  - Coordinates may point to same spot even when street number is different -> address vs. coordinates
  - Address may point to advertiser, not cottage -> How usual is this? In which sites is this common?
  - Reverse geocoding may provide wrong street number
- Are all adds about cottages? There are row houses, tree tents, igloos, even bicycles
- Removing doubles and buildings in Accommodation Statistics



Next steps

- Qualities of reverse geocoding and linking to Building Registry
- Informant incentive
- Optimizing code
  - Reverse geocoding through API takes time
  - Currently linking to Building Registry is highly compute intensive
- Building a strong theoretical background as the data can be considered a non-probability sample