Uncovering the hidden universe of rental units in Surrey

UBC Data Science for Social Good 2018

By: Jocelyn Lee, Andy Fink, Hyeongcheol Park, Zhe Jiang

Overview

- Introduction
- Data Sources and Collection
- Data Processing
- Classification Model Results
- Discussion and Future Work

The Hidden Housing Market

- Surrey is growing at a rapid rate
- Rental unit information for Surrey is incomplete
- Social consequences:
  - School overpopulation
  - Inadequate public transportation availability
  - Lack of available street parking
  - Unsafe secondary suite rentals
- Goal: provide the City of Surrey with up to date information on the **type**, **distribution** and **amount** of secondary suites

Data Sources

Open Sources:

Statistics Canada

SURREY'S OPEN DATA

CMHC SCHL

Non-Open Sources:

CITY OF SURREY the future lives here.

craigslist

VRBO Part of the HomeAway Family

airbnb

kijiji

QUANTITATIVE RHETORIC

Data Collection

- Different web crawlers built for different websites:
  - Most postings from Craigslist: **3,000~4,000 raw data monthly**
  - Other sources (mainly Kijiji and VRBO) comprise ~300 data monthly
  - Short-term rental very few: VRBO and Airbnb

- Crawler deployed on UBC server and collects data every day

- Current research was mainly based on data collected over the past 3 months

Data Cleaning & Processing

- Excluded non-Surrey region: latitude-longitude(GIS), title, location, url
- Standardization

description	housing_type	lat	long	location	price	source	title
	/ 2br - 550ft			(Ladner)	$1,600	Craigslist	2 bedroom su
	/ 4br - 2430f	49.029901	-123.07195	(delta)	$2,600	Craigslist	Delta Tsaww

- Deduplication
  - Set Theory (Deterministic Record Linkage)
  - Fuzzy Matching (Probabilistic RL)

lat	long	location	price	source	title
49.188634	-122.8632	(SURREY CE	$480	Craigslist	üí• AUG 1-FURNISHED PRIVATE ROOM - SINGLE OCCUPANCY-FEMALE STUDENT
49.188634	-122.8632	(SURREY CE	$480	Craigslist	AUG 1* BIG FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL
49.188634	-122.8632	(SURREY CE	$480	Craigslist	AUG 1* CLEAN FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL

- Missing value imputation for supervised-learning

lat	long	location	price	source	title
49.108673	-122.84547			Craigslist	1 bedroom in 2 bedroom basement only for girls

Manually Labelled Data and Proportions

Categories of Rental	% of Listings
Non-market Rental	0
Purpose-built	0.8
Entire Condo	**13.9**
Entire House or Townhouse	**25.0**
Basement Secondary Suite	**22.1**
Non-basement Secondary Suite	6.8
Laneway or Coach House	1.4
Unspecified Secondary Suite	4.5
Individual Rooms in a Condo or House	**19.8**
Non-housing Postings	5.7

Classification Example

" I am a student Punjabi girl. I need someone international Punjabi student to **share my one bedroom basement**. Internet included no laundry. Available immediately."

I am a student Punjabi girl. I need someone international Punjabi student to share my one bedroom basement.Internet included no laundry.Available immediately.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
s	city	country	date	description	lat	location	long	price	province	rooms	source	sqft	title	url	from	ID	Category	description_
	NA	NA	2018-05-24T		49.134266	(newton,sur	-122.83862	$1,400	BC	2br	Craigslist	NA	2 bedroom b	https://vanc	June_and_Ju	1204	3.1	
	NA	NA	2018-05-29T		49.173119	(Fraser/140)	-122.83444	$750	BC	1br	Craigslist	NA	750$/ 2Roon	https://vanc	June_and_Ju	1687	4	
	NA	NA	2018-06-13T		49.188386	(Surry Guild	-122.82174	$1,100	BC	2br	Craigslist	600ft	House with 2	https://vanc	June_and_Ju	2596	3.2	
	NA	NA	2018-06-07T	NA	NA	20041 55A A	NA	$1,230.00	BC	NA	Kijiji	NA	Renovated 1	https://www	June_and_Ju	4115	2.2	NA
	NA	NA	2018-06-01T		49.206217	(10237 133	-122.85394	$1,450	BC	1br	Craigslist	742ft	1 Bedroom P	https://vanc	June_and_Ju	914	3.2	
	NA	NA	2018-06-12T	NA	NA	196 68 Ave,	NA	$900.00	BC	NA	Kijiji	NA	1 bedr baser	https://www	June_and_Ju	4068	3.1	NA
	NA	NA	NA	NA	NA	241 King Edv	NA	NA	BC	NA	Kijiji	NA	Appartment	https://www	June_and_Ju	4277	NA	NA
	NA	**NA**	**2018-06**		**NA**	**NA**	**NA**	**$325**	**BC**	**NA**	**Craigsli**	**NA**	**One be**	**https://**	**June_a**	**2991**	**4**	

Problems with Such Classification

- It consumes too much to do manual labeling:
- So we built automatic **classifiers**.

- With the 1000-entry labeled dataset we had:
  - Some of the 10 classes had too few categories;
  - 1000 entries were not supportive enough to train a model to classify 10 categories;

- Shall we **condense** the current categories into fewer?

3 Category Classification

- Solution: Collapse into 3 categories:
  - 1 - Entire House or Condo 39.7%
  - 2 - Secondary Suites 34.8%
  - 3 - Individual Rooms 19.8%

(Non-housing ads excluded)

Final Classification Results

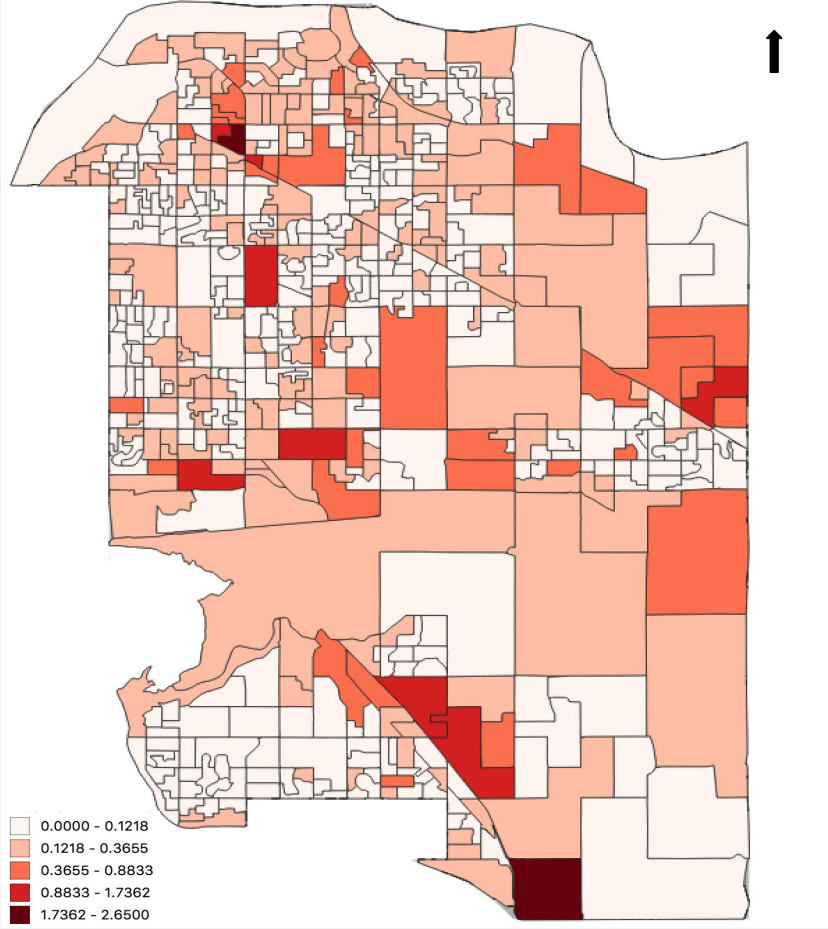
- From the Random Forest Classifier

Category	% Predicted	% Labelled
1 - Entire House or Condo	39.2	41.8
2 - Secondary Suites	37.6	37.0
3 - Individual Rooms	23.2	21.2

- Prediction Accuracy: **91%** with an out of bag error of 11%

Spatial Distribution of Online Postings

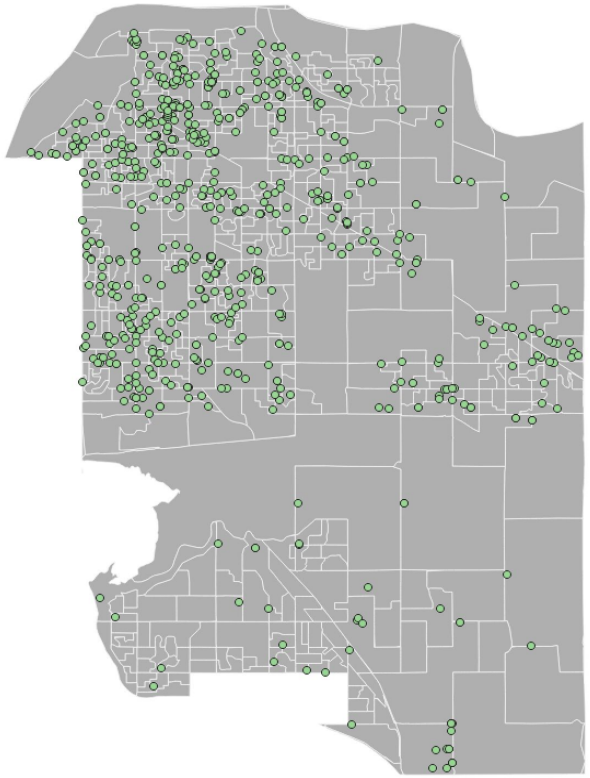
- Maps created using QGIS 3.2.3
- Counts measured using Dissemination Areas
- Highest posting densities in Douglas and City Center, high density in Cloverdale



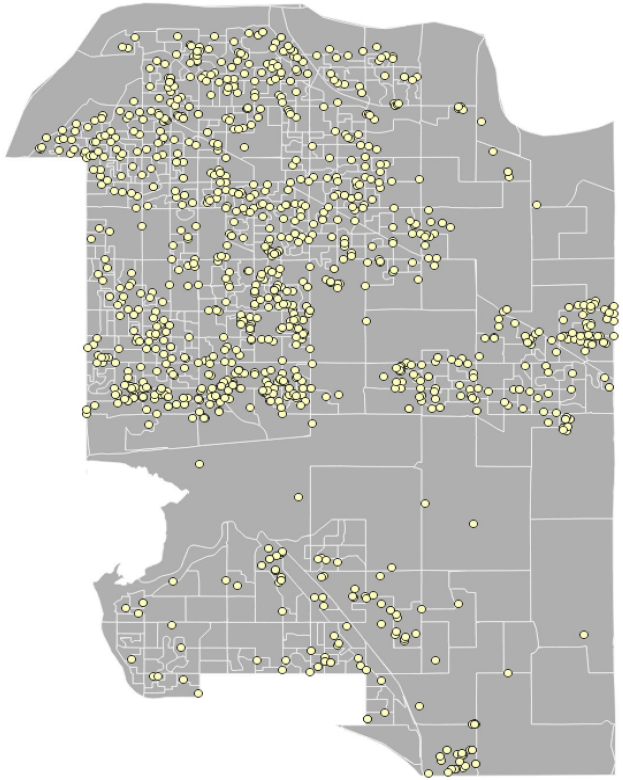
% of online posts per DA

Spatial Distribution of Online Postings

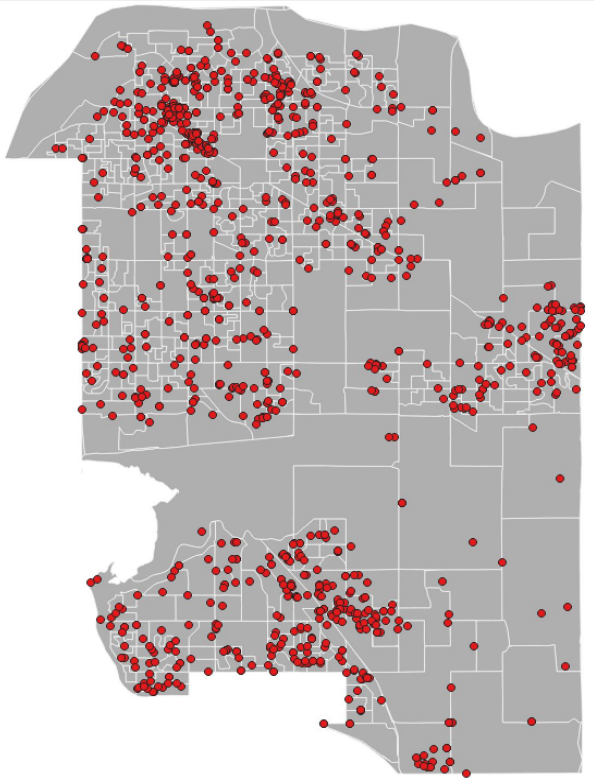
Private Room



Secondary Suite

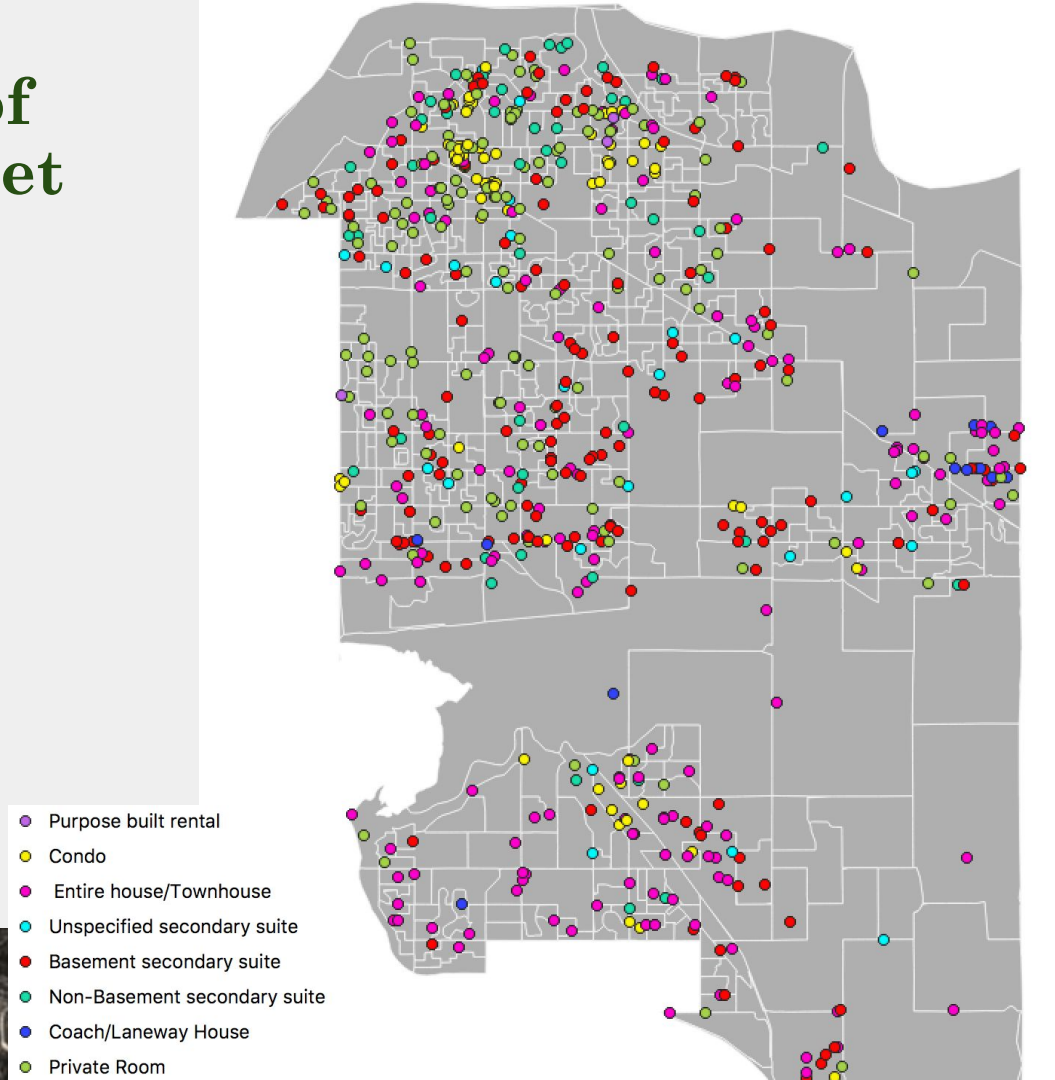


Entire Property



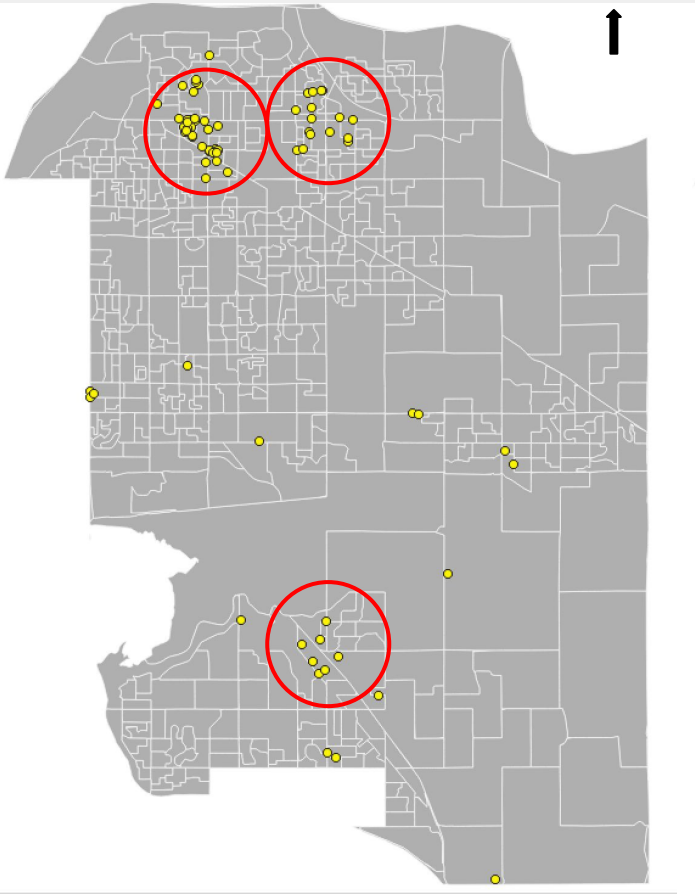
Spatial Distribution of Manually Classified Set

- Manually classified set
- Each dot represents an individual posting
- Noticeable clusters in City Centre, Cloverdale and South Surrey

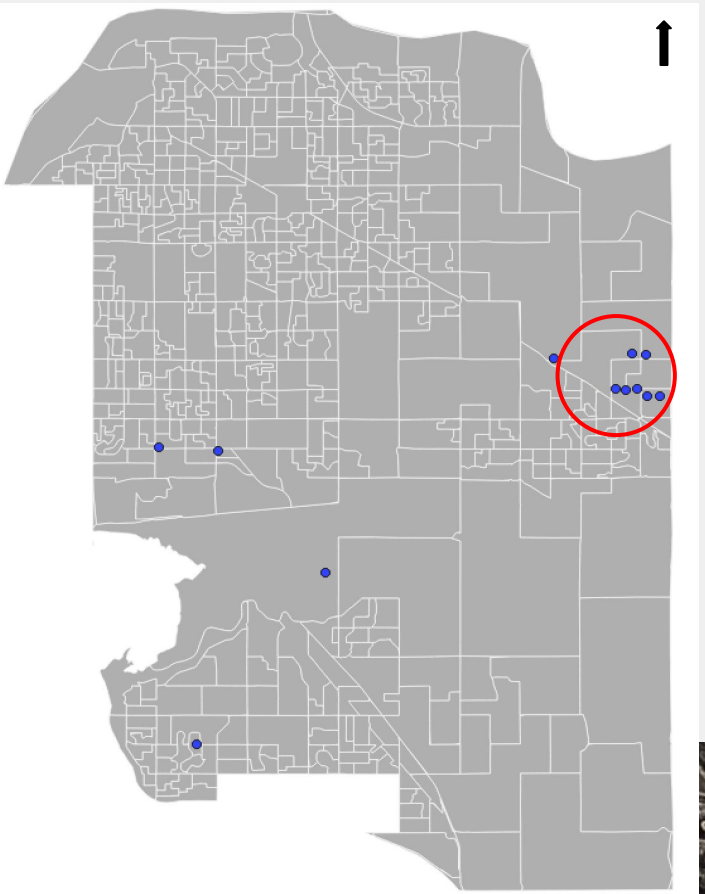


Cluster Examples

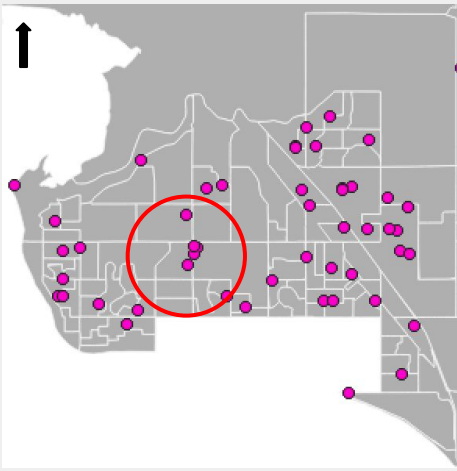
Condos



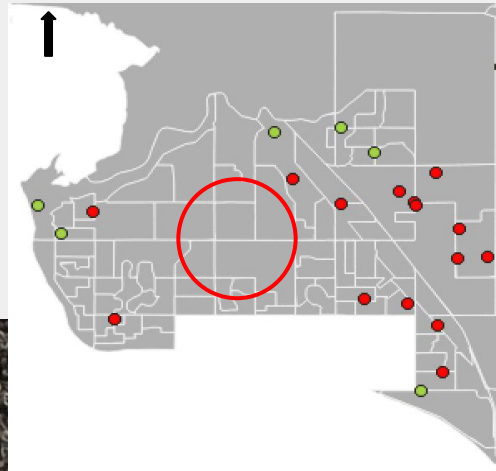
Coach/Laneway Houses



Entire Houses



Basement/Private Rooms



Discussion

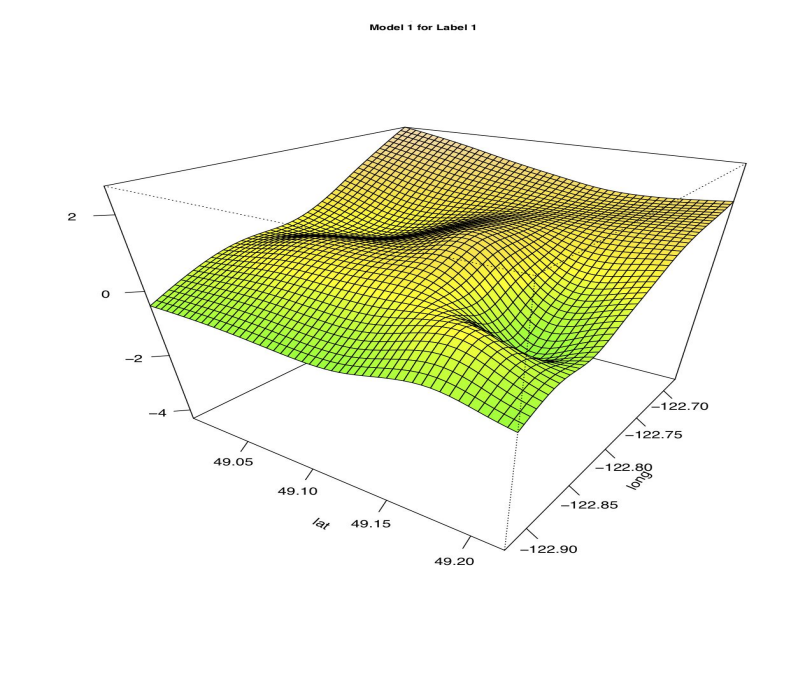
- Current dataset for supervised learning is small:
  - Distribution of categories might be different in real situation;
  - Classifier model possibly overfitting;
- Data was collected over only 3 months;
- Two other models were not ensembled, could have been used to increase accuracy.

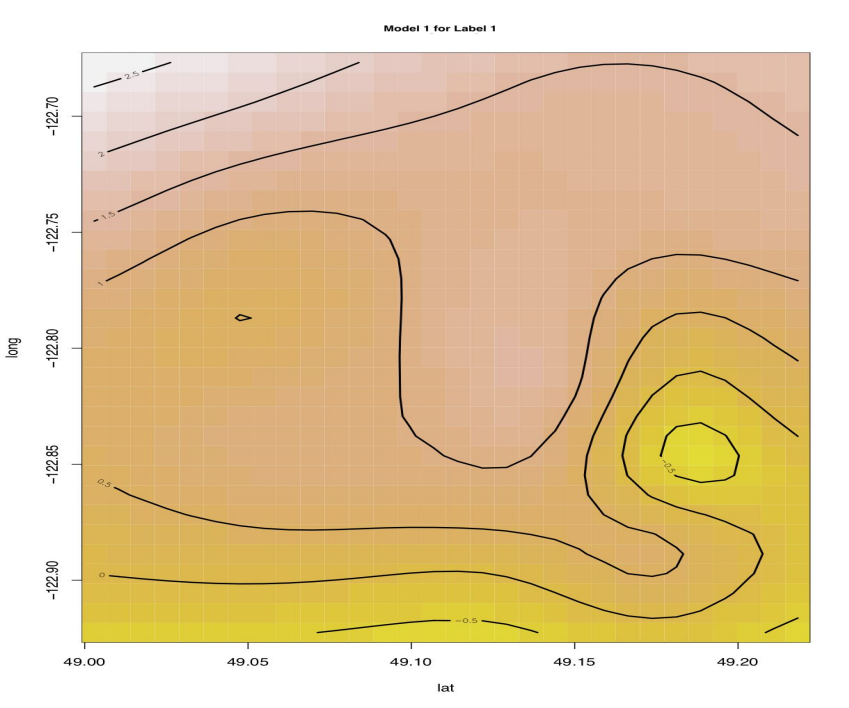
Future Work

- Validation and analysis over a time-series;
- Pipeline development: a set of user-friendly automatic tools;
- More robust classifiers with Natural Language Interpretation:
  - Better data imputation: from addresses, descriptions
  - More features generated from titles/descriptions
  - Ensembled methods

Thanks for watching!
Questions?

Final Classification Results





Other Classification Results

- From the Naive Bayes Model (without normalization)

Category	% Predicted	% Labelled
1 - Entire House or Condo	28.07	39.7
2 - Secondary Suites	46.04	34.8
3 - Individual Rooms	20.89	19.8

- Prediction Accuracy: 75%

Other Classification Results

- From the Generalized Additive Model with Majority Voting

Category	% Predicted	% Labelled
1 - Entire House or Condo	46.0	41.8
2 - Secondary Suites	37.0	37.0
3 - Individual Rooms	16.9	21.2

- Prediction Accuracy: 82.53%