✋ Machine Learning for Tangible Effects:

Natural Language Processing for Uncovering the Illicit Massage Industry

PhD Thesis Defense: Rui Ouyang

Sept. 5, 2023

Computer Science
Harvard University

Update:
🎥 The talk is now on youtube !
📜 & Thesis is now on Arxiv

Related Works -- 2 other classifiers

- Classified Yelp businesses with random forest
- Classified individual Yelp reviews
- Demands Estimate
- Correlated socioeconomic factors with Rubmaps listings (per county and per census tract) - income, airport dist.
- Analyzed (explicit) Rubmaps reviews: correlated text features suggesting trafficking / exploitation

Video shortlink: tinyurl.com/nro-defense-video
These slides: tinyurl.com/nro-defense-slides

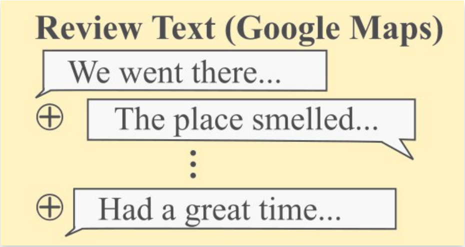
- My website: nrobot.dev
- My contact: nouyang@alum.mit.edu

[Edit from Sept. 15]

Outline of Talk

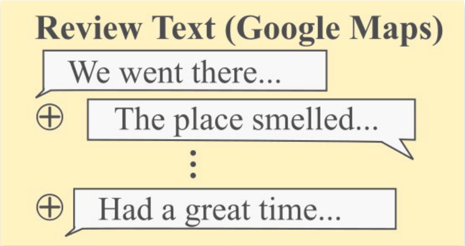
Outline of Talk

- **Part 1** (20 mins)
  The Google Places dataset: illegal activity in plain sight
- **Part 2** (20 mins)
  The Forum dataset: two case studies and a hackathon



Outline of Talk

- **Part 1** (20 mins)
  The Google Places dataset: illegal activity in plain sight
- **Part 2** (20 mins)
  The Forum dataset: two case studies and a hackathon
- **Part 3** (10 mins)
  My research journey:
  | Scotiabank | Digger Finger | Fiducial Force Sensor |
  Call-to-Action
  Acknowledgements



Content Warning

This work may contain sexist and racist language or topics

Discretion advised

(Should be SFW)

Feel free to leave any time ! There's cookies outside.

The Google Places Dataset: Illegal Activity in Plain Sight

Part I

The Google Places Dataset: Illegal Activity in Plain Sight

Part I

Sections

1. Introduction
2. Classifier
3. Vulnerability Insights with Named Entity Recognition
4. Fairness with Ablation & Business Hours

What makes a massage parlor illicit?

Establishments with registered business names that ostensibly provide massage, wellness, and/or spa services while in fact deriving some clientele and revenue through the provision of commercial sex acts.

> – V. Bouche and S. M. Crotty, “Estimating demand for illicit massage businesses in Houston, Texas,”Journal of Human Trafficking, vol. 4, no. 4, pp. 279–297, Oct. 2018.

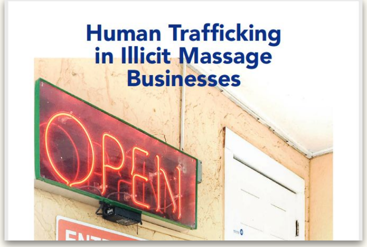
What makes a massage parlor illicit?

Establishments with registered business names that ostensibly provide massage, wellness, and/or spa services while in fact deriving some clientele and revenue through the provision of commercial sex acts.

– V. Bouche and S. M. Crotty, "Estimating demand for illicit massage businesses in Houston, Texas,"Journal of Human Trafficking, vol. 4, no. 4, pp. 279–297, Oct. 2018.

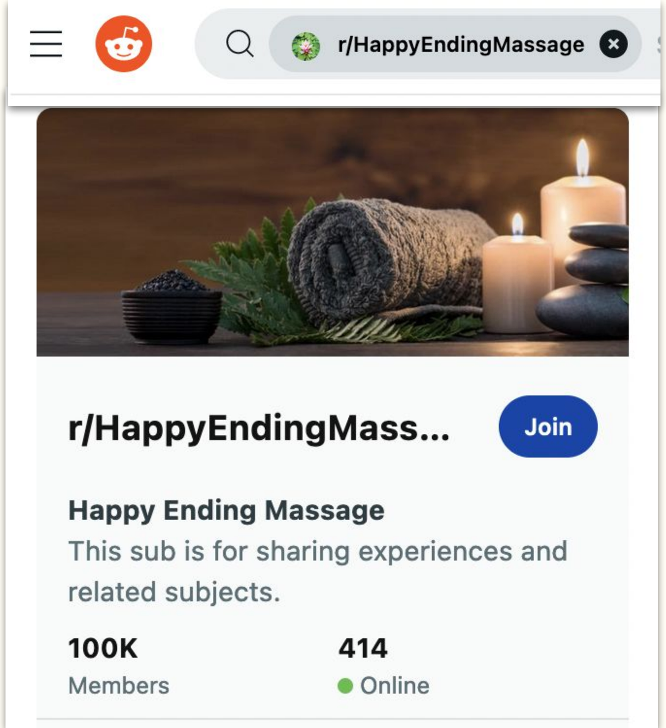
In 2018:

- 11,000+ IMBs in the United States
- Combined annual revenue of $2.5 billion
- 100+ locations in Manhattan alone that received visits in first half of 2023



Keyhan, Rochelle, et al. "Human Trafficking in Illicit Massage Businesses (Report)." (2017).

Not that niche



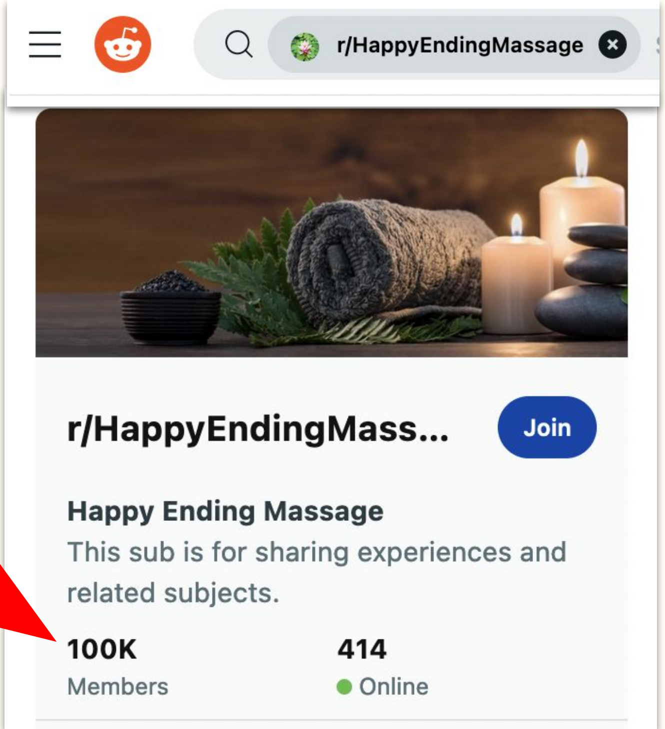
r/HappyEndingMass...

Happy Ending Massage

This sub is for sharing experiences and related subjects.

100K	**414**
Members	● Online

Not that niche



r/HappyEndingMass...

Happy Ending Massage

This sub is for sharing experiences and related subjects.

100K	414
Members	● Online

How is it linked to human trafficking?

IMI Employees

- Generally immigrant women, often undocumented, often in debt (travel loans)
- English barriers
- Often supporting family

Combination of sex and labor trafficking

How is it linked to human trafficking?

IMI Employees

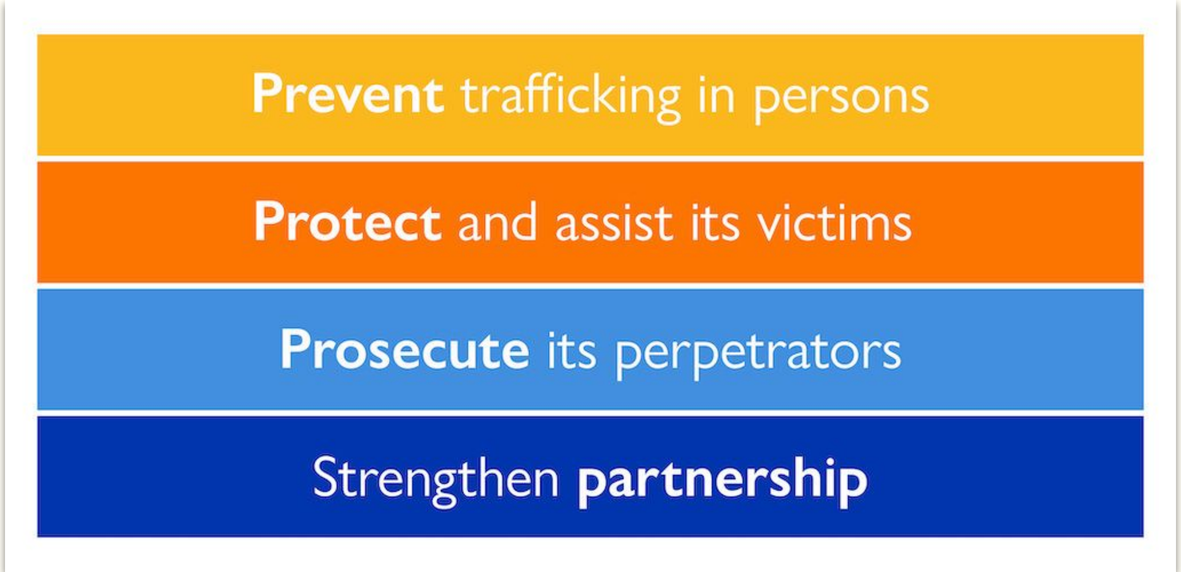
- Generally immigrant women, often undocumented, often in debt (travel loans)
- English barriers
- Often supporting family

Combination of sex and labor trafficking

> "**Trafficking in persons**" shall mean the **recruitment**, transportation, transfer, harbouring or receipt of persons,
>
> by means of the threat or use of force or other forms of coercion, of abduction, **of fraud, of deception**, of the abuse of power or **of a position of vulnerability** or of the giving or receiving of payments or benefits
>
> to achieve the consent of **a person having control over another person, for the purpose of exploitation.**

– United Nations Palermo Protocols, which was adopted in 2000 and now ratified by 178 parties

The 4Ps Framework



Prevent trafficking in persons

Protect and assist its victims

Prosecute its perpetrators

Strengthen **partnership**

The United Nations Global Plan of Action to Combat Trafficking in Persons

How we can help: The 4Ps Framework

- **Prevention**
  - **Monitoring**
  - Laws
  - **Reducing demand**
  - **Public awareness**

How we can help: The 4Ps Framework

- **Prevention**
    - **Monitoring**
    - Laws
    - **Reducing demand**
    - **Public awareness**

- **Protection**
    - Rescue **(victim-identification)**
    - Rehabilitation **(access to help and long-term opportunity)**
    - Re-integration (voluntary repatriation)

How we can help: The 4Ps Framework

- **Prevention**
    - **Monitoring**
    - Laws
    - **Reducing demand**
    - **Public awareness**

- **Protection**
    - Rescue **(victim-identification)**
    - Rehabilitation **(access to help and long-term opportunity)**
    - Re-integration (voluntary repatriation)

Who: Lawmaker

Why: Effect of policy

How we can help: The 4Ps Framework

- **Prevention**
  - **Monitoring**
  - Laws
  - **Reducing demand**
  - **Public awareness**

Who: Lawmaker
Why: Effect of policy

- **Protection**
  - Rescue **(victim-identification)**
  - Rehabilitation **(access to help and long-term opportunity)**
  - Re-integration (voluntary repatriation)

Who: Non-profit
Why: Job training location

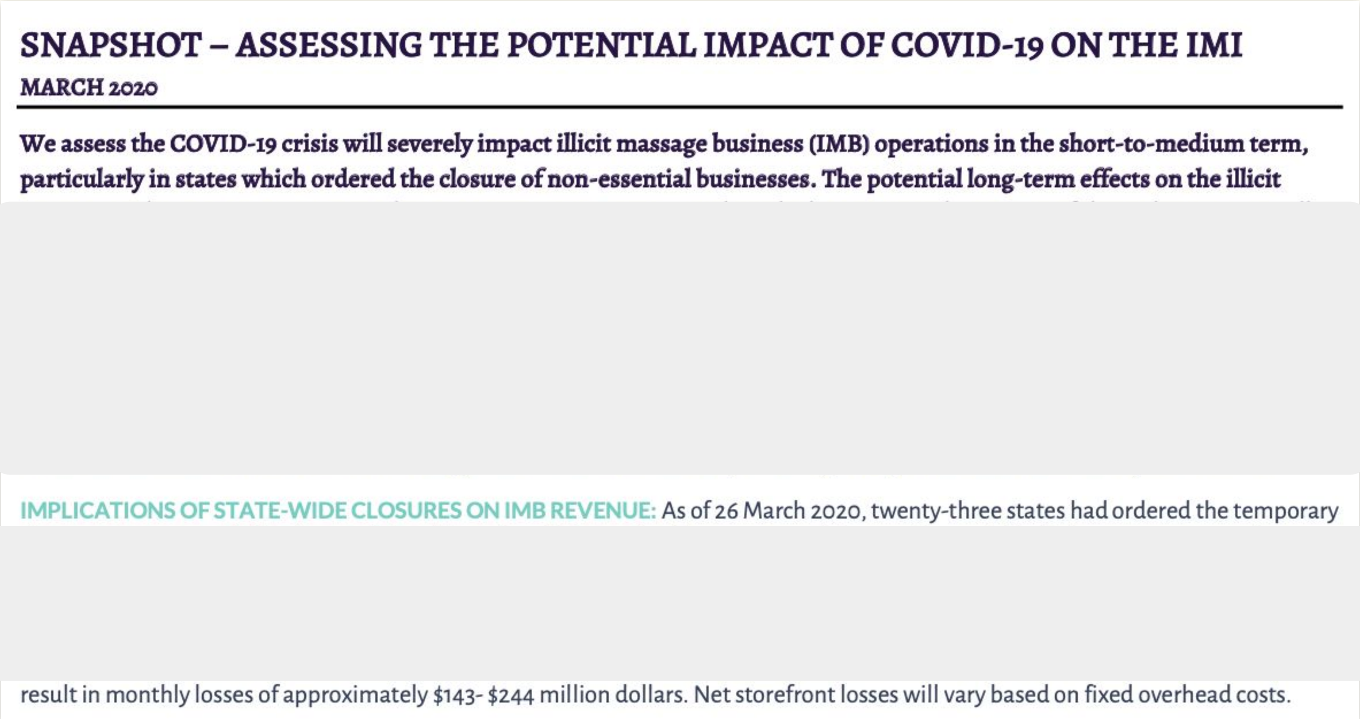
The Google Places Dataset: Illegal Activity in Plain Sight

Part I

Sections

1. Introduction
2. Classifier
3. Vulnerability Insights with Named Entity Recognition
4. Fairness with Ablation & Business Hours

The Network



SNAPSHOT – ASSESSING THE POTENTIAL IMPACT OF COVID-19 ON THE IMI

MARCH 2020

We assess the COVID-19 crisis will severely impact illicit massage business (IMB) operations in the short-to-medium term, particularly in states which ordered the closure of non-essential businesses. The potential long-term effects on the illicit

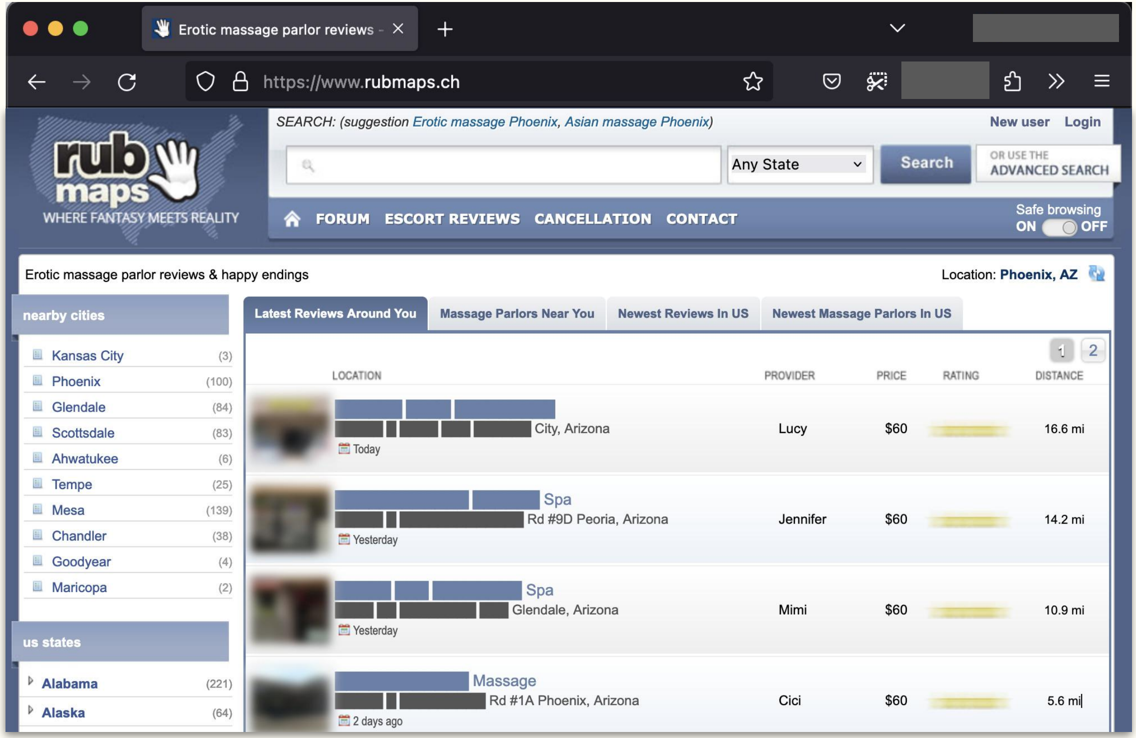
[redacted]

IMPLICATIONS OF STATE-WIDE CLOSURES ON IMB REVENUE: As of 26 March 2020, twenty-three states had ordered the temporary

[redacted]

result in monthly losses of approximately $143- $244 million dollars. Net storefront losses will vary based on fixed overhead costs.

Rubmaps



Erotic massage parlor reviews -

https://www.rubmaps.ch

rub maps
WHERE FANTASY MEETS REALITY

SEARCH: (suggestion Erotic massage Phoenix, Asian massage Phoenix)

New user Login

Any State

Search

OR USE THE ADVANCED SEARCH

FORUM ESCORT REVIEWS CANCELLATION CONTACT

Safe browsing ON OFF

Erotic massage parlor reviews & happy endings

Location: **Phoenix, AZ**

nearby cities

Kansas City	(3)
Phoenix	(100)
Glendale	(84)
Scottsdale	(83)
Ahwatukee	(6)
Tempe	(25)
Mesa	(139)
Chandler	(38)
Goodyear	(4)
Maricopa	(2)

us states

Alabama	(221)
Alaska	(64)

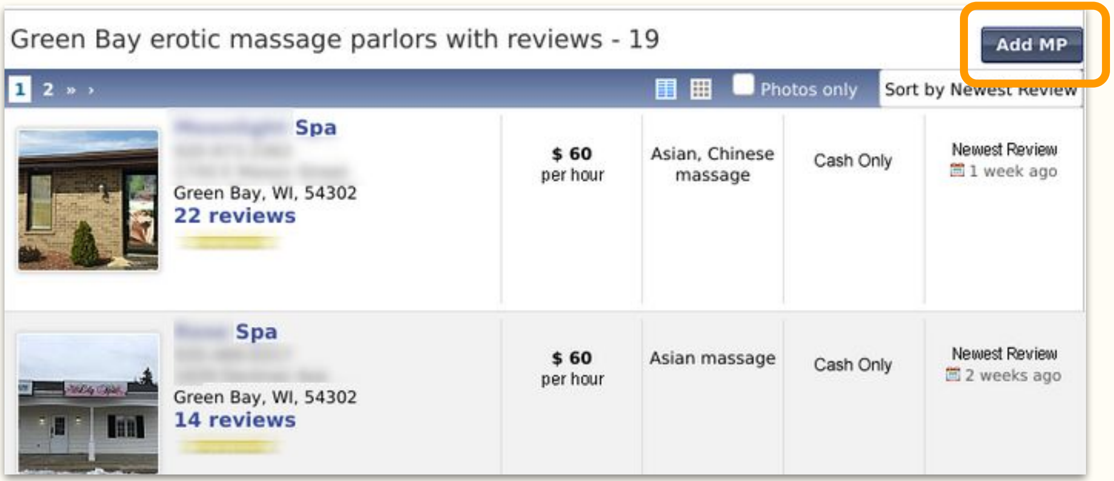
Latest Reviews Around You | Massage Parlors Near You | Newest Reviews In US | Newest Massage Parlors In US

1 2

LOCATION	PROVIDER	PRICE	RATING	DISTANCE
[redacted] [redacted] City, Arizona Today	Lucy	$60	[redacted]	16.6 mi
[redacted] Spa [redacted] Rd #9D Peoria, Arizona Yesterday	Jennifer	$60	[redacted]	14.2 mi
[redacted] Spa [redacted] Glendale, Arizona Yesterday	Mimi	$60	[redacted]	10.9 mi
[redacted] Massage [redacted] Rd #1A Phoenix, Arizona 2 days ago	Cici	$60	[redacted]	5.6 mi

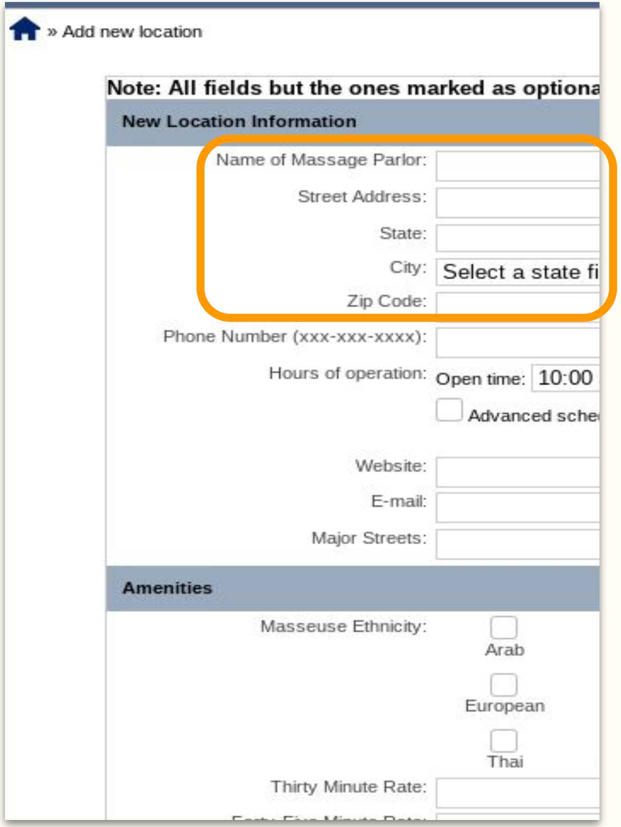
Rubmaps (video)

User-added listings



Green Bay erotic massage parlors with reviews - 19

[redacted] Spa [redacted] [redacted] Green Bay, WI, 54302 22 reviews [redacted]	$ 60 per hour	Asian, Chinese massage	Cash Only	Newest Review 1 week ago
[redacted] Spa [redacted] [redacted] Green Bay, WI, 54302 14 reviews [redacted]	$ 60 per hour	Asian massage	Cash Only	Newest Review 2 weeks ago



» Add new location

Note: All fields but the ones marked as optiona

New Location Information

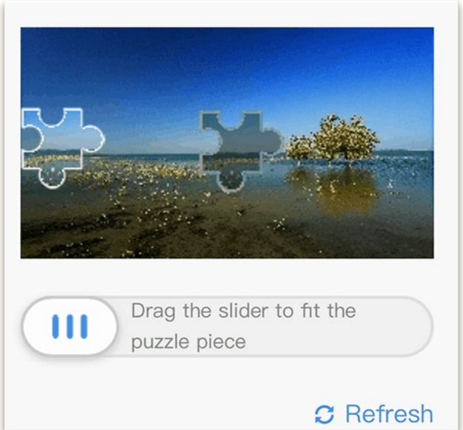
Name of Massage Parlor:
Street Address:
State:
City: Select a state fi
Zip Code:
Phone Number (xxx-xxx-xxxx):
Hours of operation: Open time: 10:00
☐ Advanced sche
Website:
E-mail:
Major Streets:

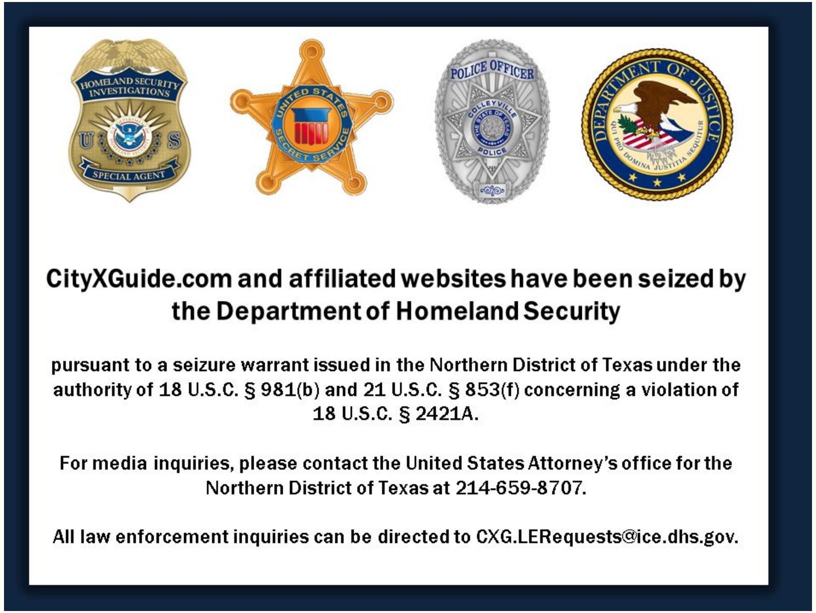
Amenities

Masseuse Ethnicity:
☐ Arab
☐ European
☐ Thai
Thirty Minute Rate:

Monitoring over time

- Scrape frequency / CAPTCHA arms race
- Law enforcement action





CityXGuide.com and affiliated websites have been seized by the Department of Homeland Security

pursuant to a seizure warrant issued in the Northern District of Texas under the authority of 18 U.S.C. § 981(b) and 21 U.S.C. § 853(f) concerning a violation of 18 U.S.C. § 2421A.

For media inquiries, please contact the United States Attorney's office for the Northern District of Texas at 214-659-8707.

All law enforcement inquiries can be directed to CXG.LERequests@ice.dhs.gov.

2020, $15 million forfeiture

Key Insight

Can we use larger, more stable, well-known website as complementary (or replacement) data source?

- **Rubmaps:** High precision
    - United States only
    - Does not have text (requires subscription)
    - Scraping-based

Key Insight

Can we use larger, more stable, well-known website as complementary (or replacement) data source?

- **Rubmaps:** High precision
  - United States only
  - Does not have text (requires subscription)
  - Scraping-based

- **Google Maps:**
  - World-wide
  - Up to 5 reviews per business
  - API calls

Key Insight

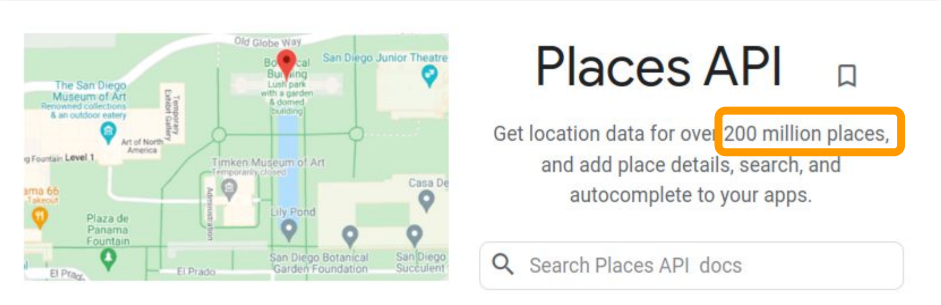
Can we use larger, more stable, well-known website as complementary (or replacement) data source?

- **Rubmaps:** High precision
  - United States only
  - Does not have text (requires subscription)
  - Scraping-based
- **Google Maps:**
  - World-wide
  - Up to 5 reviews per business
  - API calls

Rubmaps: ~11,000 locations

vs.

Google Maps:

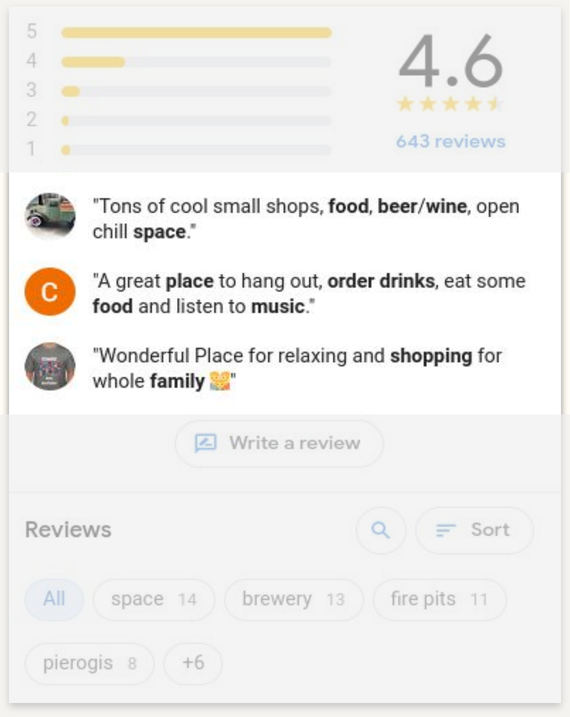


Places API

Get location data for over 200 million places, and add place details, search, and autocomplete to your apps.

Search Places API docs

Feasibility

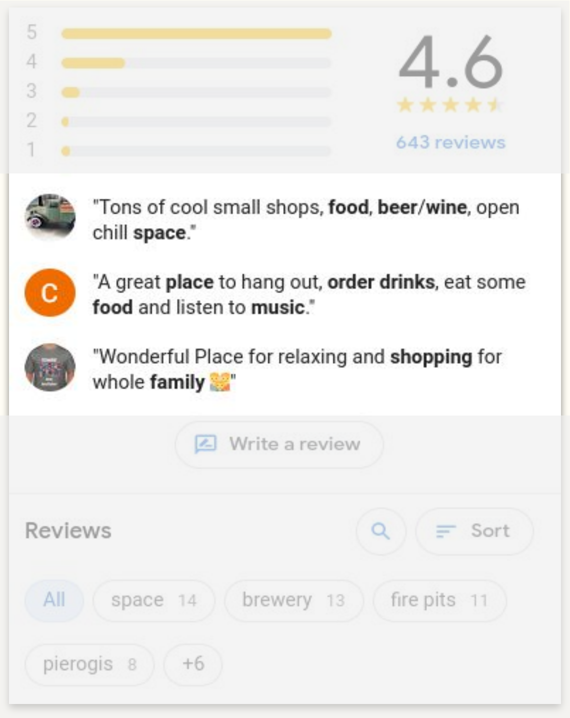


"Tons of cool small shops, **food**, **beer**/**wine**, open chill **space**."

"A great **place** to hang out, **order drinks**, eat some **food** and listen to **music**."

"Wonderful Place for relaxing and **shopping** for whole **family** 🥰"

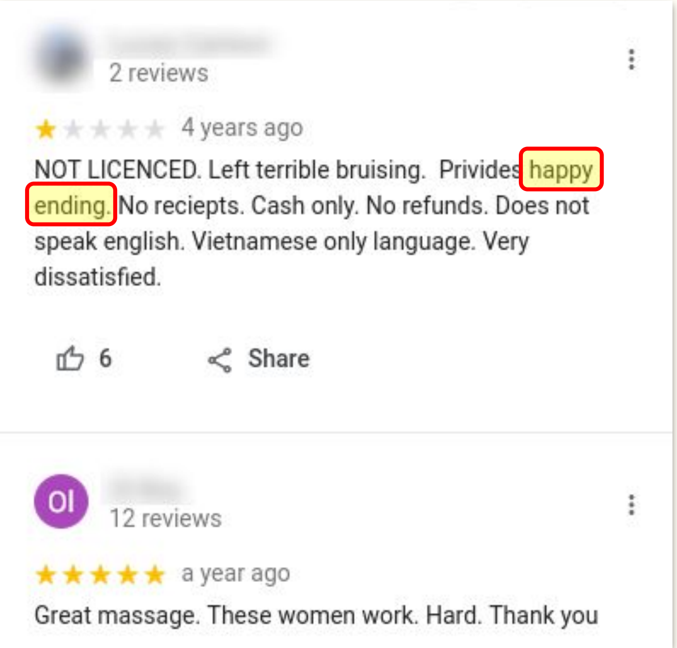
Feasibility



"Tons of cool small shops, **food**, **beer/wine**, open chill **space**."

"A great **place** to hang out, **order drinks**, eat some **food** and listen to **music**."

"Wonderful Place for relaxing and **shopping** for whole **family** 🥰"



NOT LICENCED. Left terrible bruising. Privides happy ending. No reciepts. Cash only. No refunds. Does not speak english. Vietnamese only language. Very dissatisfied.

Great massage. These women work. Hard. Thank you

- Mapped advertisements, showed different locations (globally) in supply and demand

  Ramchandani, P., Bastani, H., & Wyatt, E. (2021). **Unmasking Human Trafficking Risk in Commercial Sex Supply Chains with Machine Learning**. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3866259

Same phone number used in both recruitment and sales

Related Works -- 2 other classifiers

- Classified Yelp businesses with random forest
  - Maria Diaz and Anand Panangadan. "**Natural language-based integration of online review datasets for identification of sex trafficking businesses.**" IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI) (2020)
- Classified individual Yelp reviews
  - Ruoting Li, Margaret Tobey, Maria Mayorga, Sherrie Caltagirone, and Osman Ozaltn. "**Detecting human trafficking: Automated classification of online customer reviews of massage businesses.**" SSRN Electronic Journal (2021)
- Demands Estimate
  - Bouche, V., & Crotty, S. M. Estimating demand for illicit massage businesses in Houston, Texas. Journal of Human Trafficking (2018)
- Correlated socioeconomic factors with Rubmaps listings (per county and per census tract) - income, airport dist.
  - Anna White, Seth Guika2ema, and Bridgette Carr. "**Why are you Here? modeling illicit massage business location characteristics with machine learning**." Journal of Human Trafficking (2021)
- Analyzed (explicit) Rubmaps reviews: correlated text features suggesting trafficking / exploitation
  - Vries, Ieke de and Jason Radford. "**Identifying online risk markers of hard-to-observe crimes through semi-inductive triangulation: The case of human trafficking in the United States**." The British Journal of Criminology (2021)

Prior Classifier Limitations

- Limited to cities in 12 states
- Static releases by Yelp

Issue:

- Easy to move across cities, states under law enforcement pressure

Ground Truth

Label definition:

1 = a "flagged" a.ka. **illicit** business

0 = not concerned, a.k.a a **legal** business

Ground Truth

- No list of all massage parlors in U.S., let alone all illicit massage parlors
- Instead:
  Use Places API to find co-occurring massage parlors as a negative class
- Co-occuring = Same cities as Rubmaps
- If a business is **not** listed in Rubmaps, consider it a **legal** business

Label definition:

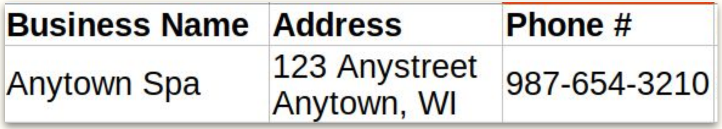
1 = a "flagged" a.ka. **illicit** business
0 = not concerned, a.k.a a **legal** business

How to get Google Place IDs from Rubmaps?

- Business can change names, address misspellings, etc.
- Use "Find Place" request: Takes text input, returns Place ID(s)
  Put in business name and address
- Final results use Google Places IDs provided by collaborator

Example:

Rubmaps Data



Business Name	Address	Phone #
Anytown Spa	123 Anystreet Anytown, WI	987-654-3210

Google Data - Match?

Name: Any Town Spa

Address: 132 Any St. #3, Anytown, WI

Place ID: =3cs34lk8geh

Listed on Rubmaps: ~4,700 businesses

- Turnover:
  Keep only locations reviewed since since
  Jan. 1st, 2019 (up until April 1st, 2021)

- Total:
  4,719 businesses

- Cities: ~1,700
  Geocode city name to GPS point

nearby_search() for each of ~1,700 cities

- Cities from illicit class

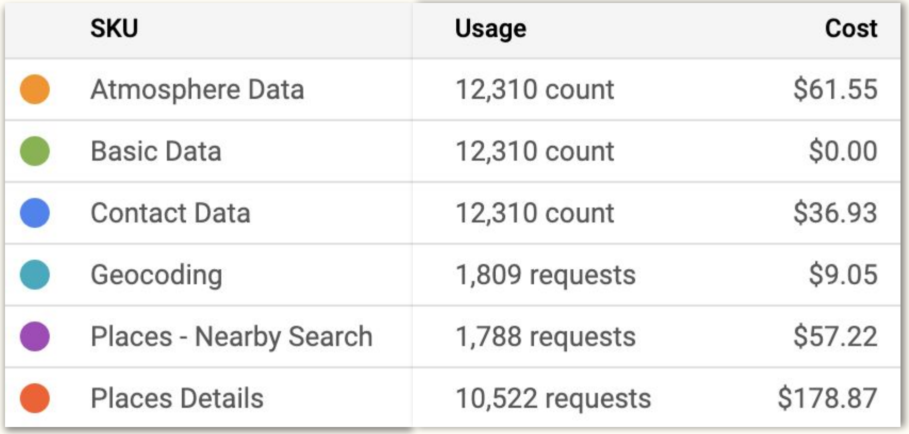
-
```
places.nearby_search(
    GPS lat, lon
    keyword=massage
    type=SPA )
```

- Returns 0–20 businesses (ordered by distance)

- Total:
  17,247 places

Subset (API cost) = ~7,000 businesses

- Total:
  17,247 places
- Overlap:
  1,541 places listed in Rubmaps
  (~9 % prevalence)
- Subset to half:
  7,431 places



	SKU	Usage	Cost
●	Atmosphere Data	12,310 count	$61.55
●	Basic Data	12,310 count	$0.00
●	Contact Data	12,310 count	$36.93
●	Geocoding	1,809 requests	$9.05
●	Places - Nearby Search	1,788 requests	$57.22
●	Places Details	10,522 requests	$178.87

Total: $344

Dataset Summary

Place IDs

- 4,719 positive
- 7,431 negative

12,150 total

`place_details()`

- Up to 5 reviews per place
- 55,385 reviews total

Dataset Summary

Place IDs

- 4,719 positive
- 7,431 negative

12,150 total

`place_details()`

- Up to 5 reviews per place
- 55,385 reviews total

Baseline algorithm:
Always guess most frequent class

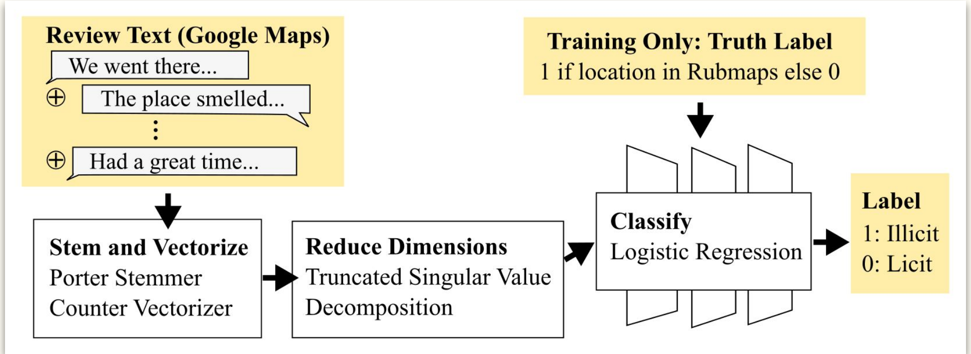
Null accuracy: 61.2%

Methods

- Download review text
  & Clean (pre-process)
- Turn text into numbers **(bag-of-words)**
- Classify into illicit (label 1) or legal (label 0)
- Evaluate with **5-fold cross-validation**

Methods

- Download review text
  & Clean (pre-process)
- Turn text into numbers **(bag-of-words)**
- Classify into illicit (label 1) or legal (label 0)
- Evaluate with **5-fold cross-validation**



Review Text (Google Maps)

We went there...

⊕ The place smelled...

⋮

⊕ Had a great time...

Training Only: Truth Label

1 if location in Rubmaps else 0

Stem and Vectorize

Porter Stemmer

Counter Vectorizer

Reduce Dimensions

Truncated Singular Value Decomposition

Classify

Logistic Regression

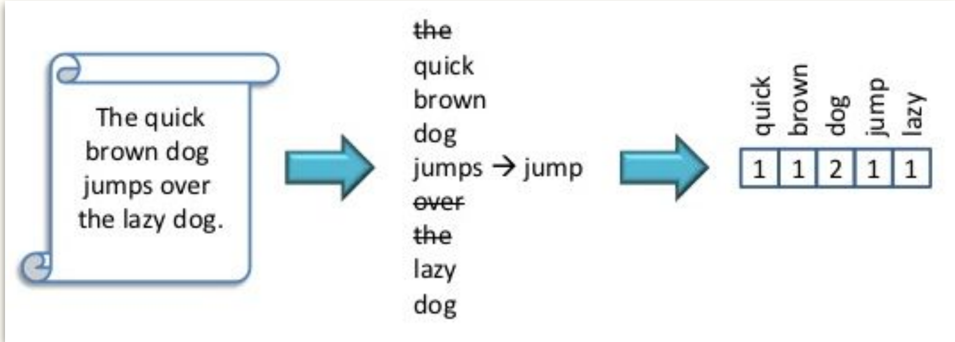
Label

1: Illicit

0: Licit

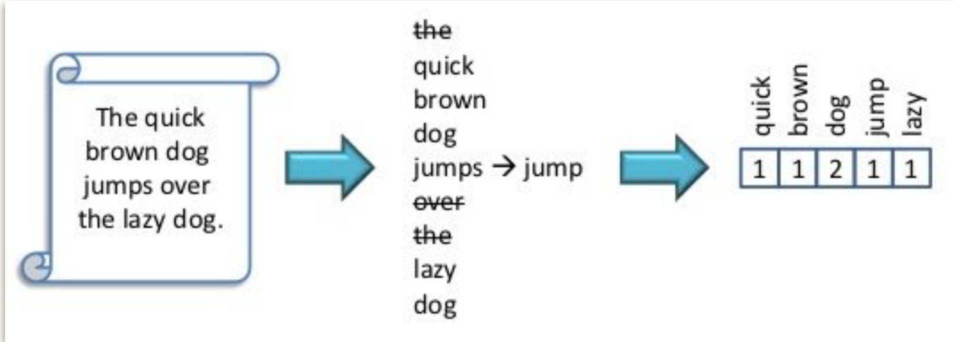
Bag-of-Words (a.k.a. CountVectorizer)

- Tally words
- Order of words doesn't matter



Bag-of-Words (a.k.a. CountVectorizer)

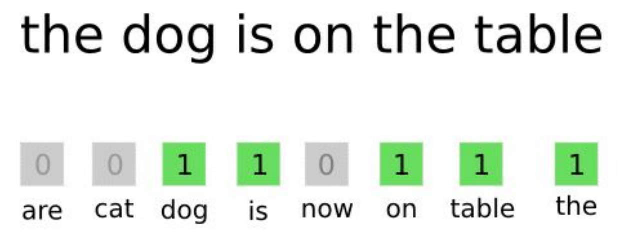
- Tally words
- Order of words doesn't matter



The quick brown dog jumps over the lazy dog.

~~the~~
quick
brown
dog
jumps → jump
~~over~~
~~the~~
lazy
dog

quick	brown	dog	jump	lazy
1	1	2	1	1



the dog is on the table

are	cat	dog	is	now	on	table	the
0	0	1	1	0	1	1	1

Stemming

Example:

Massage, massaged, massages -> massag

- Reduces vocabulary size of BoW

Stemming

Example:
Massage, massaged, massages -> massag

- Reduces vocabulary size of BoW

- **Raw:**
  "Happy we went there."
  "The place smelled."
  "Spoke English there."

- **Concatenate:** "Happy we went there The place smelled Spoke English there"

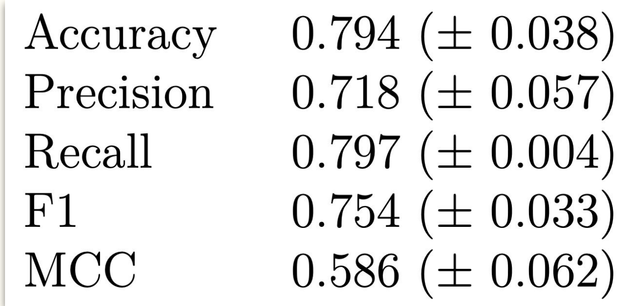
- **Stem:** "happi we went there the place smell spoke english there"

Evaluation: 5-fold cross-validation

- 80% train, 20% test
- Run five times

Result: ~80% accuracy

Metrics



Accuracy	0.794 (± 0.038)
Precision	0.718 (± 0.057)
Recall	0.797 (± 0.004)
F1	0.754 (± 0.033)
MCC	0.586 (± 0.062)

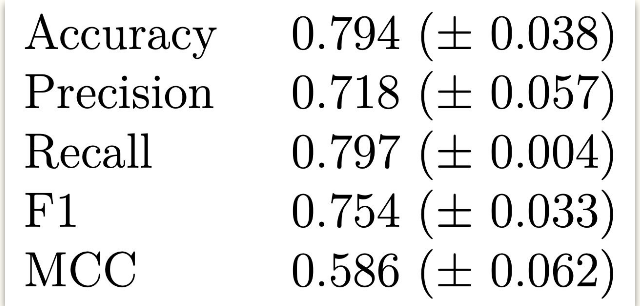
MCC - Matthew’s correlation coefficient

Requires good performance on both classes

● -1 or 1	perfect correlation
● 0	random chance

Result: ~80% accuracy

Metrics



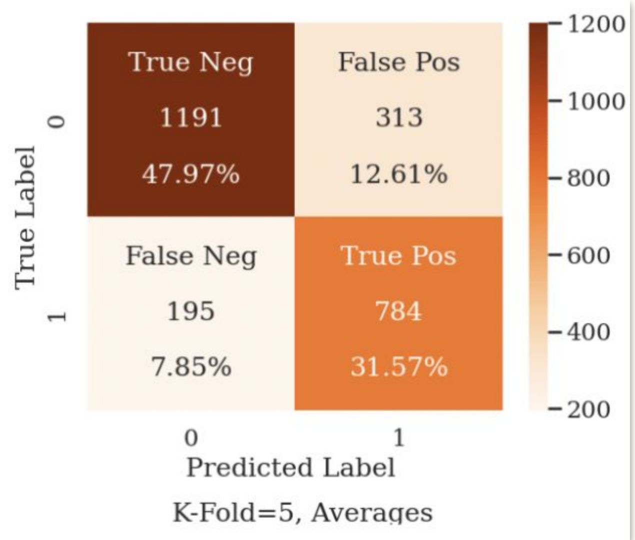
Accuracy	0.794 (± 0.038)
Precision	0.718 (± 0.057)
Recall	0.797 (± 0.004)
F1	0.754 (± 0.033)
MCC	0.586 (± 0.062)

MCC - Matthew's correlation coefficient

Requires good performance on both classes

-1 or 1	perfect correlation
0	random chance

Confusion Matrix



True Label \ Predicted Label	0	1
0	True Neg 1191 47.97%	False Pos 313 12.61%
1	False Neg 195 7.85%	True Pos 784 31.57%

K-Fold=5, Averages

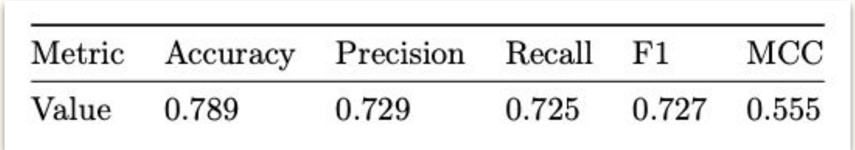
Can we improve? with bigger model

- Transformers architecture
- 66 million parameters
  vs 14,000

- End-to-end model
  Input: Text
  Output: Prediction

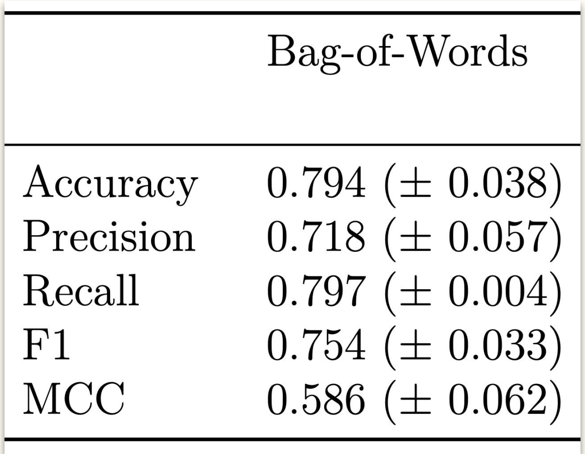
Result: ~80% accuracy still

DistilBERT



Metric	Accuracy	Precision	Recall	F1	MCC
Value	0.789	0.729	0.725	0.727	0.555

Previous results:



	Bag-of-Words
Accuracy	0.794 (± 0.038)
Precision	0.718 (± 0.057)
Recall	0.797 (± 0.004)
F1	0.754 (± 0.033)
MCC	0.586 (± 0.062)

Hard problem -- likely noise ceiling on data

Summary: Google Places works

- Bag-of-Words: 80% accuracy
- DistilBERT: 80% accuracy

Caveats:

- Not all places in Rubmaps are illicit

The Google Places Dataset: Illegal Activity in Plain Sight

Part I

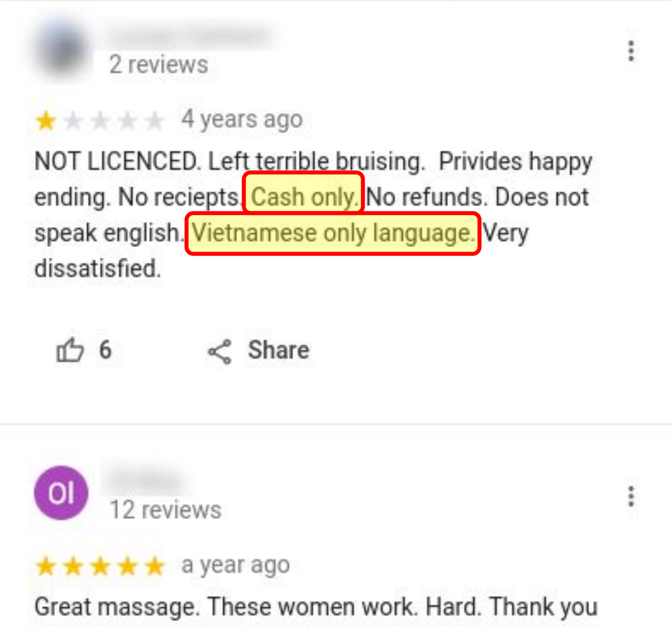
Sections

1. Introduction
2. Classifier
3. Vulnerability Insights with Named Entity Recognition
4. Fairness with Ablation & Business Hours

Vulnerabilities

- Language barrier, racism
- Labor regulations: hours and pay

Does this vary between illicit and legal locations?



[redacted]
2 reviews

★☆☆☆☆ 4 years ago

NOT LICENCED. Left terrible bruising. Privides happy ending. No reciepts. Cash only. No refunds. Does not speak english. Vietnamese only language. Very dissatisfied.

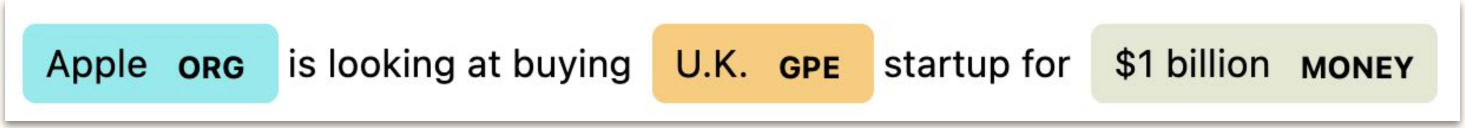
[redacted]
12 reviews

★★★★★ a year ago

Great massage. These women work. Hard. Thank you

Named Entity Recognition

Pre-defined list of tags



Apple **ORG** is looking at buying U.K. **GPE** startup for $1 billion **MONEY**

Labor:

- MONEY: monetary values, including units

Ethnicity:

- NORP: Nationalities or religious or political groups
- LANGUAGE: Any named language

Labor & Opening Hours

Metadata:

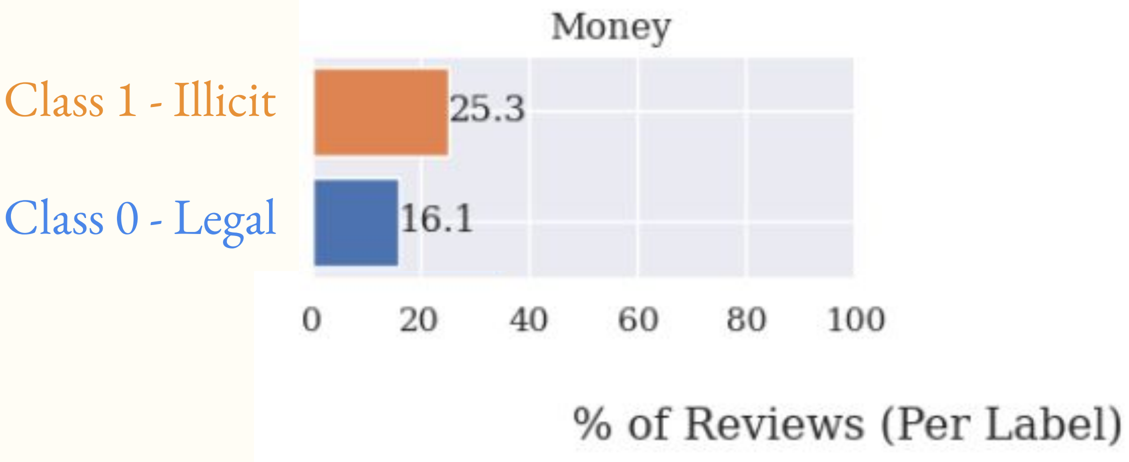
Business opening hours

- Business open 7 days a week
- Open until 9PM or later

Features

- MONEY
- ETHNICITY
- OPEN LATE
- OPEN 7 DAYS

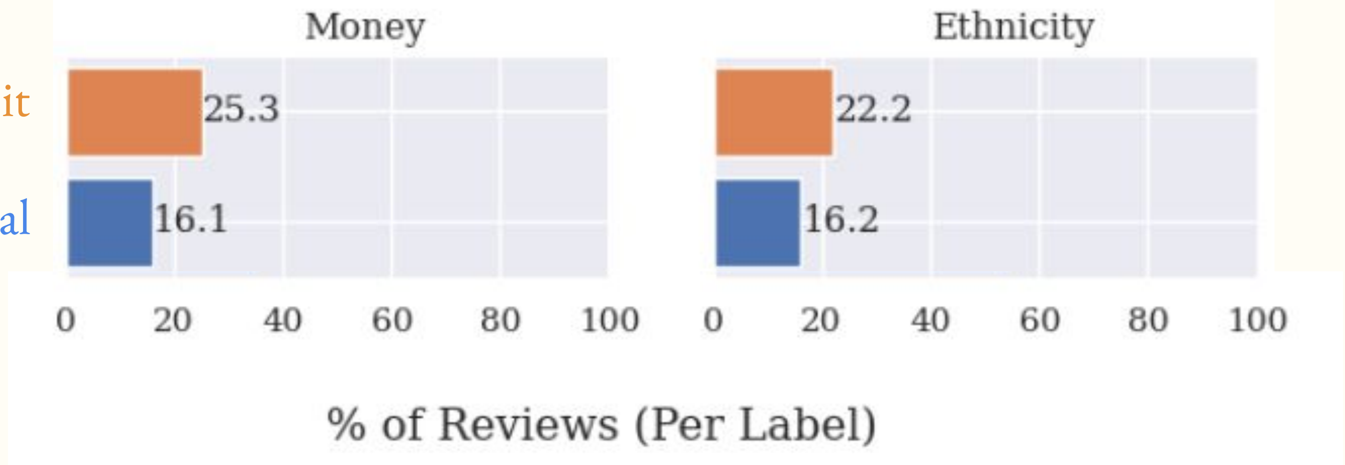
Distribution of NER Features



Distribution of NER Features

Class 1 - Illicit

Class 0 - Legal

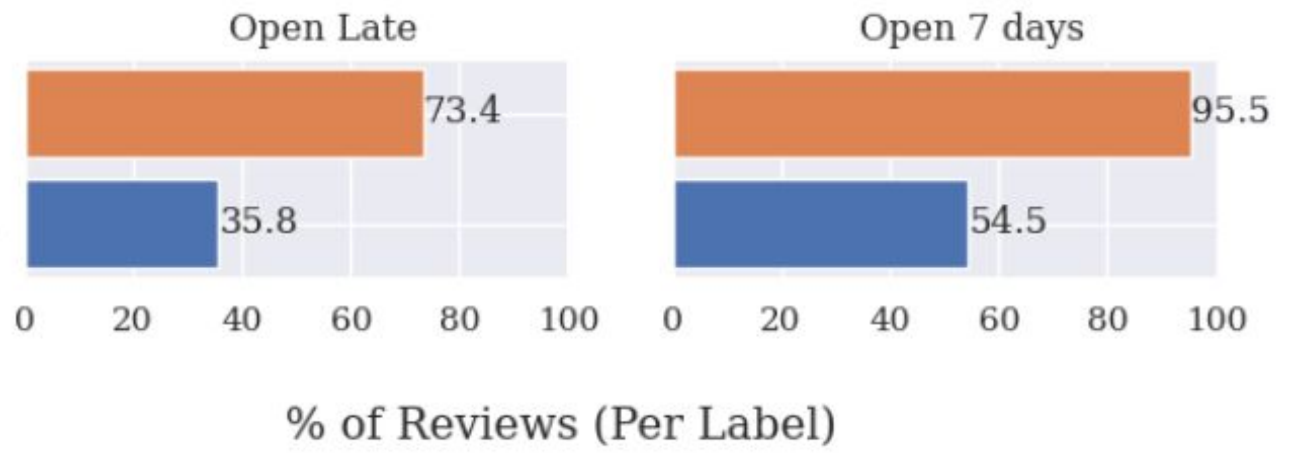


Distribution of business hours features

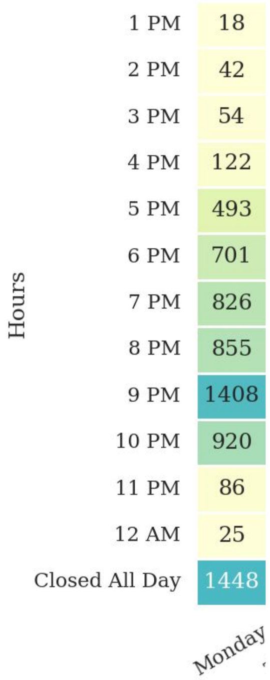
Distribution of business hours features

Class 1 - Illicit

Class 0 - Legal

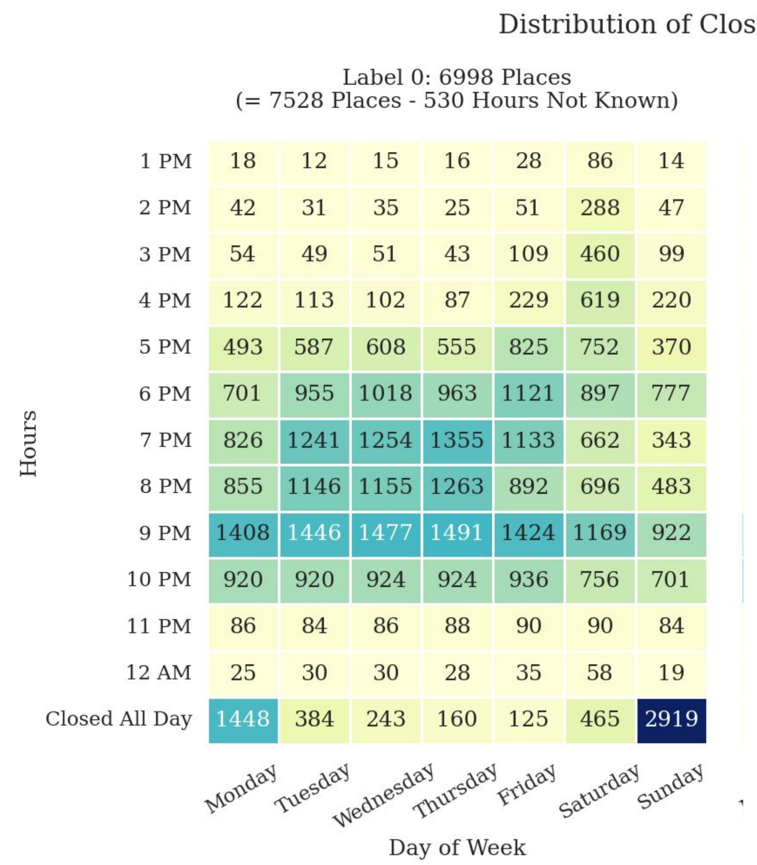


Closer look



Hours	Monday
1 PM	18
2 PM	42
3 PM	54
4 PM	122
5 PM	493
6 PM	701
7 PM	826
8 PM	855
9 PM	1408
10 PM	920
11 PM	86
12 AM	25
Closed All Day	1448

Closer look

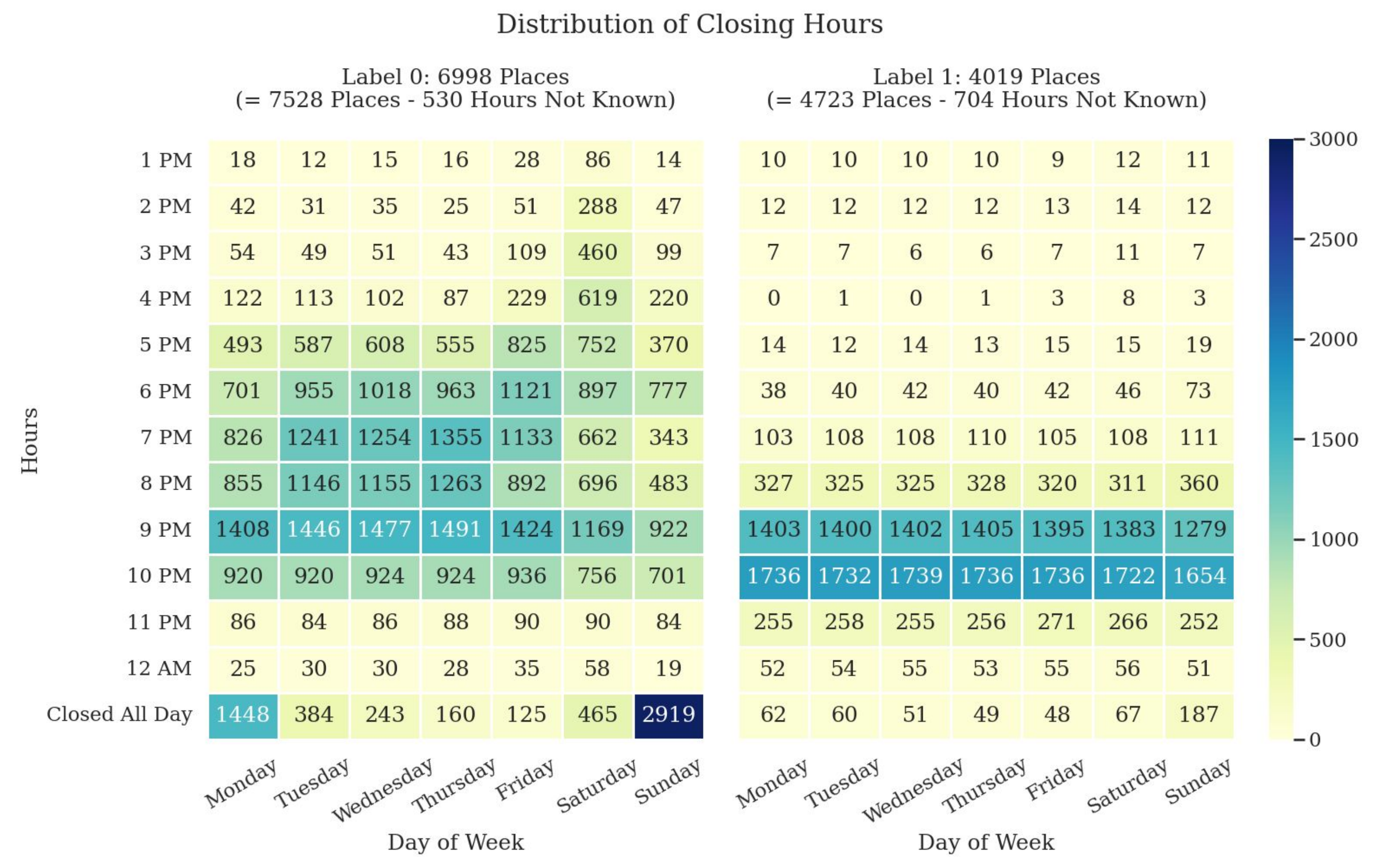


Distribution of Closing Hours

Label 0: 6998 Places
(= 7528 Places - 530 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	18	12	15	16	28	86	14
2 PM	42	31	35	25	51	288	47
3 PM	54	49	51	43	109	460	99
4 PM	122	113	102	87	229	619	220
5 PM	493	587	608	555	825	752	370
6 PM	701	955	1018	963	1121	897	777
7 PM	826	1241	1254	1355	1133	662	343
8 PM	855	1146	1155	1263	892	696	483
9 PM	1408	1446	1477	1491	1424	1169	922
10 PM	920	920	924	924	936	756	701
11 PM	86	84	86	88	90	90	84
12 AM	25	30	30	28	35	58	19
Closed All Day	1448	384	243	160	125	465	2919

Closer look



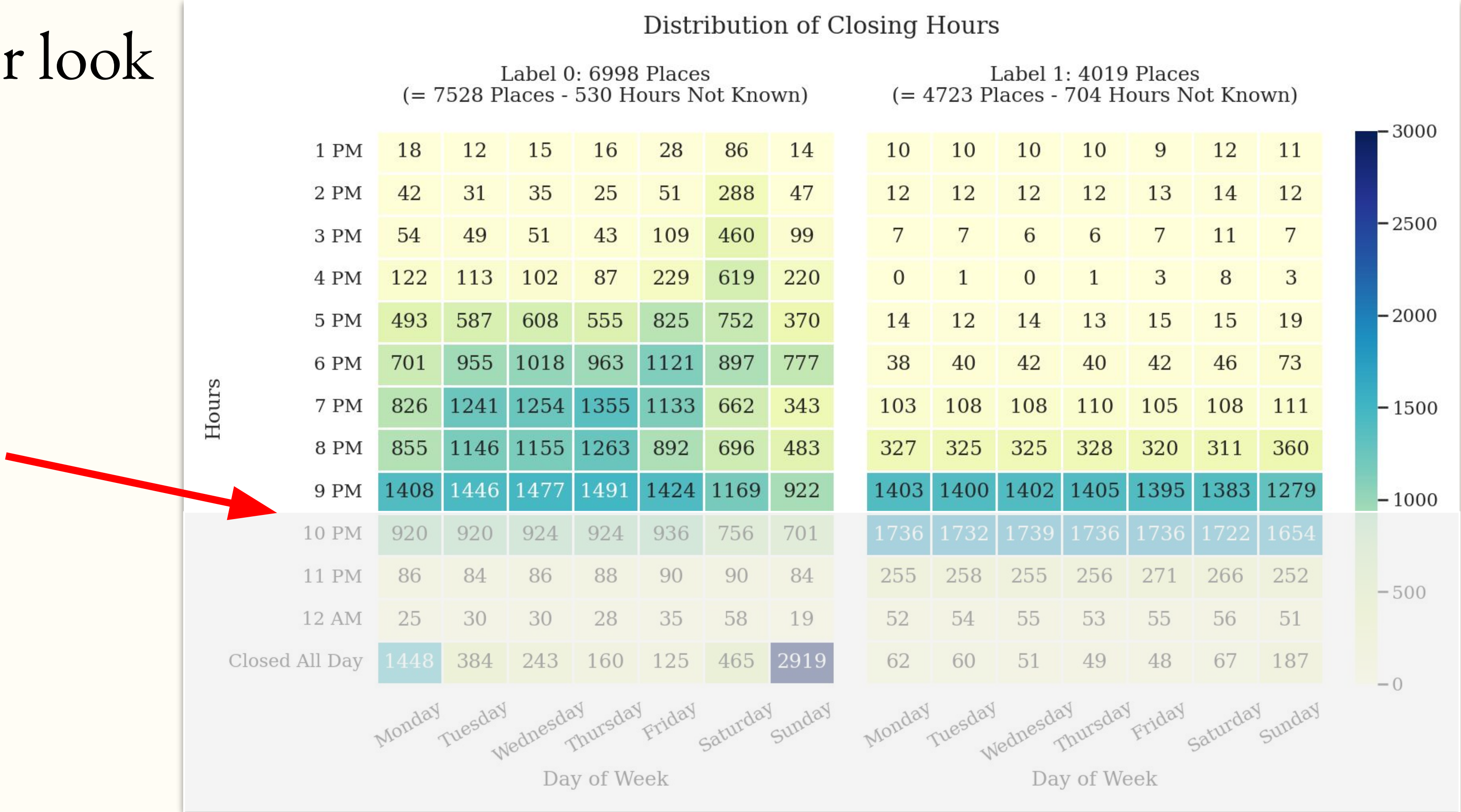
Label 0: 6998 Places
(= 7528 Places - 530 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	18	12	15	16	28	86	14
2 PM	42	31	35	25	51	288	47
3 PM	54	49	51	43	109	460	99
4 PM	122	113	102	87	229	619	220
5 PM	493	587	608	555	825	752	370
6 PM	701	955	1018	963	1121	897	777
7 PM	826	1241	1254	1355	1133	662	343
8 PM	855	1146	1155	1263	892	696	483
9 PM	1408	1446	1477	1491	1424	1169	922
10 PM	920	920	924	924	936	756	701
11 PM	86	84	86	88	90	90	84
12 AM	25	30	30	28	35	58	19
Closed All Day	1448	384	243	160	125	465	2919

Label 1: 4019 Places
(= 4723 Places - 704 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	10	10	10	10	9	12	11
2 PM	12	12	12	12	13	14	12
3 PM	7	7	6	6	7	11	7
4 PM	0	1	0	1	3	8	3
5 PM	14	12	14	13	15	15	19
6 PM	38	40	42	40	42	46	73
7 PM	103	108	108	110	105	108	111
8 PM	327	325	325	328	320	311	360
9 PM	1403	1400	1402	1405	1395	1383	1279
10 PM	1736	1732	1739	1736	1736	1722	1654
11 PM	255	258	255	256	271	266	252
12 AM	52	54	55	53	55	56	51
Closed All Day	62	60	51	49	48	67	187

Closer look



Distribution of Closing Hours

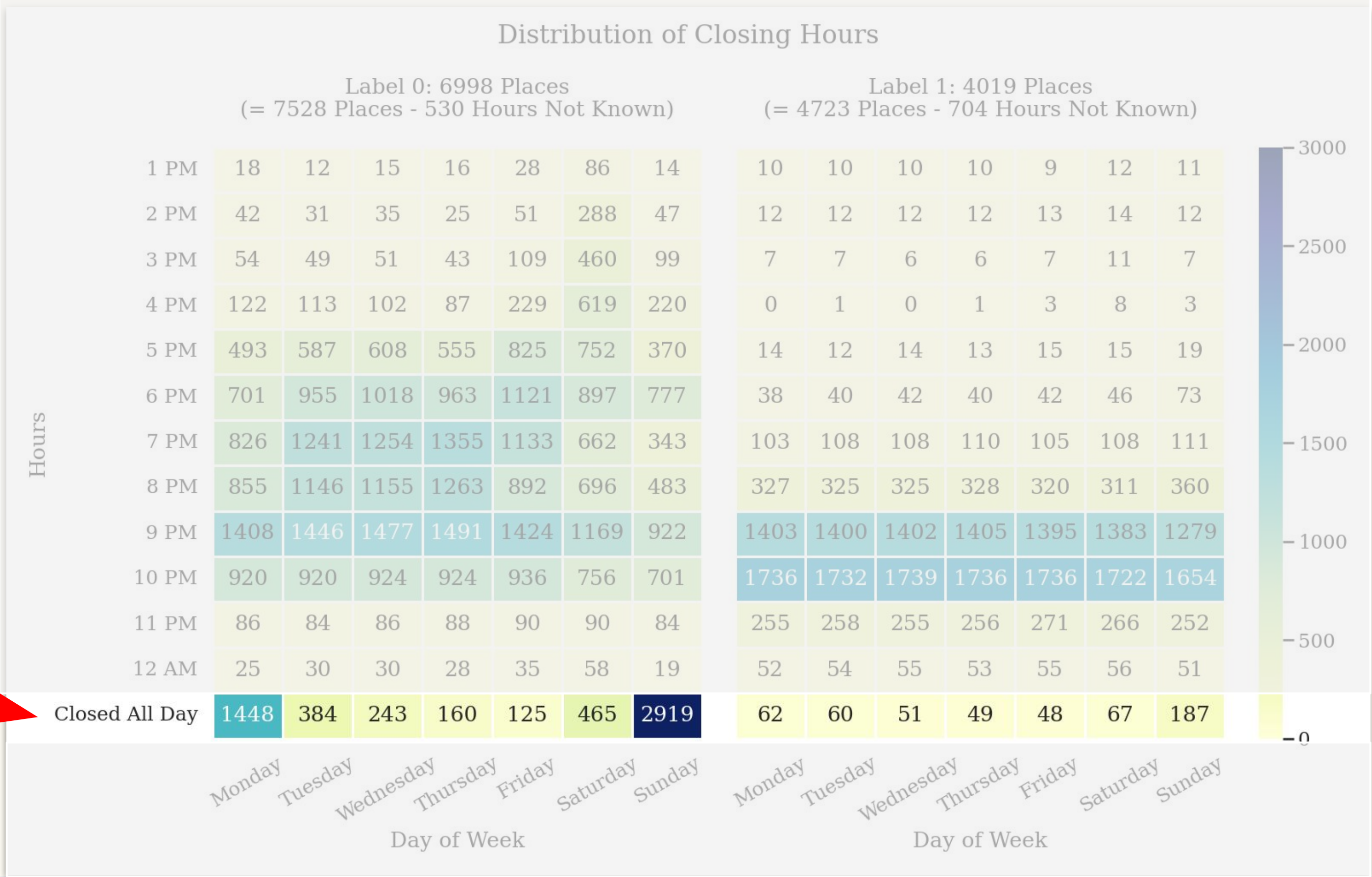
Label 0: 6998 Places
(= 7528 Places - 530 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	18	12	15	16	28	86	14
2 PM	42	31	35	25	51	288	47
3 PM	54	49	51	43	109	460	99
4 PM	122	113	102	87	229	619	220
5 PM	493	587	608	555	825	752	370
6 PM	701	955	1018	963	1121	897	777
7 PM	826	1241	1254	1355	1133	662	343
8 PM	855	1146	1155	1263	892	696	483
9 PM	1408	1446	1477	1491	1424	1169	922
10 PM	920	920	924	924	936	756	701
11 PM	86	84	86	88	90	90	84
12 AM	25	30	30	28	35	58	19
Closed All Day	1448	384	243	160	125	465	2919

Label 1: 4019 Places
(= 4723 Places - 704 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	10	10	10	10	9	12	11
2 PM	12	12	12	12	13	14	12
3 PM	7	7	6	6	7	11	7
4 PM	0	1	0	1	3	8	3
5 PM	14	12	14	13	15	15	19
6 PM	38	40	42	40	42	46	73
7 PM	103	108	108	110	105	108	111
8 PM	327	325	325	328	320	311	360
9 PM	1403	1400	1402	1405	1395	1383	1279
10 PM	1736	1732	1739	1736	1736	1722	1654
11 PM	255	258	255	256	271	266	252
12 AM	52	54	55	53	55	56	51
Closed All Day	62	60	51	49	48	67	187

Closer look



Distribution of Closing Hours

Label 0: 6998 Places
(= 7528 Places - 530 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	18	12	15	16	28	86	14
2 PM	42	31	35	25	51	288	47
3 PM	54	49	51	43	109	460	99
4 PM	122	113	102	87	229	619	220
5 PM	493	587	608	555	825	752	370
6 PM	701	955	1018	963	1121	897	777
7 PM	826	1241	1254	1355	1133	662	343
8 PM	855	1146	1155	1263	892	696	483
9 PM	1408	1446	1477	1491	1424	1169	922
10 PM	920	920	924	924	936	756	701
11 PM	86	84	86	88	90	90	84
12 AM	25	30	30	28	35	58	19
Closed All Day	1448	384	243	160	125	465	2919

Label 1: 4019 Places
(= 4723 Places - 704 Hours Not Known)

Hours	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1 PM	10	10	10	10	9	12	11
2 PM	12	12	12	12	13	14	12
3 PM	7	7	6	6	7	11	7
4 PM	0	1	0	1	3	8	3
5 PM	14	12	14	13	15	15	19
6 PM	38	40	42	40	42	46	73
7 PM	103	108	108	110	105	108	111
8 PM	327	325	325	328	320	311	360
9 PM	1403	1400	1402	1405	1395	1383	1279
10 PM	1736	1732	1739	1736	1736	1722	1654
11 PM	255	258	255	256	271	266	252
12 AM	52	54	55	53	55	56	51
Closed All Day	62	60	51	49	48	67	187

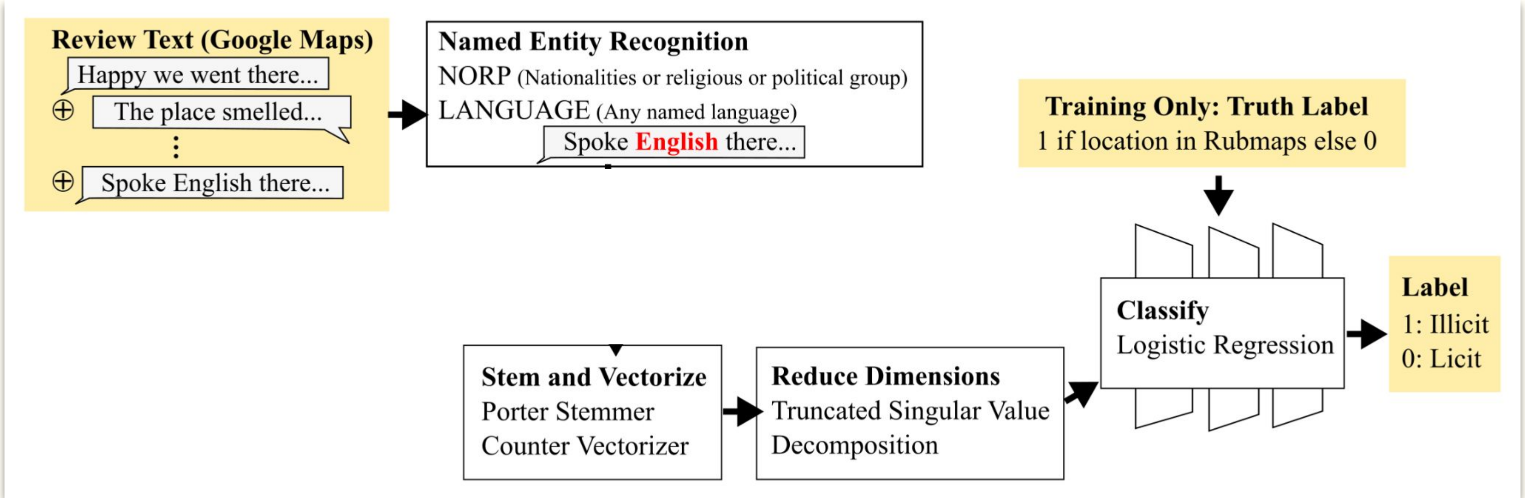
The Google Places Dataset: Illegal Activity in Plain Sight

Part I

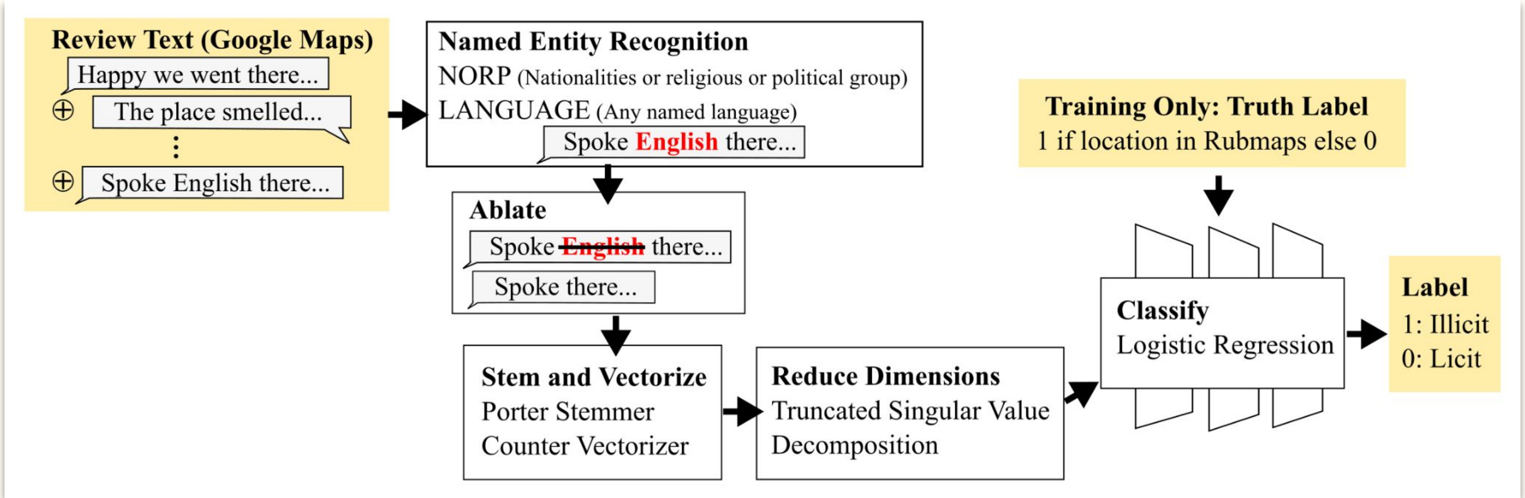
Sections

1. Introduction
2. Classifier
3. Vulnerability Insights with Named Entity Recognition
4. Fairness with Ablation & Business Hours

Fairness by Ablation (Removing Text)



Fairness by Ablation (Removing Text)



Review Text (Google Maps)

Happy we went there...

⊕ The place smelled...

⋮

⊕ Spoke English there...

Named Entity Recognition

NORP (Nationalities or religious or political group)

LANGUAGE (Any named language)

Spoke **English** there...

Ablate

Spoke ~~English~~ there...

Spoke there...

Stem and Vectorize

Porter Stemmer

Counter Vectorizer

Reduce Dimensions

Truncated Singular Value Decomposition

Training Only: Truth Label

1 if location in Rubmaps else 0

Classify

Logistic Regression

Label

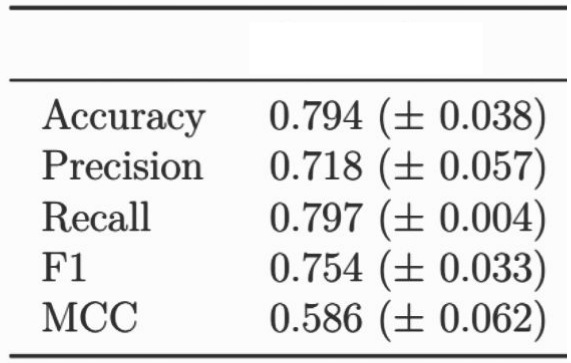
1: Illicit

0: Licit

Ablation Results

5-fold cross-validation

Before

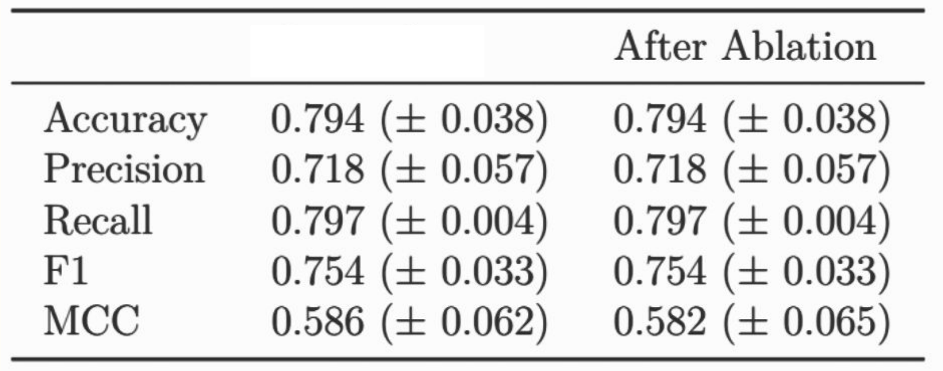


Accuracy	0.794 (± 0.038)
Precision	0.718 (± 0.057)
Recall	0.797 (± 0.004)
F1	0.754 (± 0.033)
MCC	0.586 (± 0.062)

Ablation Results

5-fold cross-validation

Before

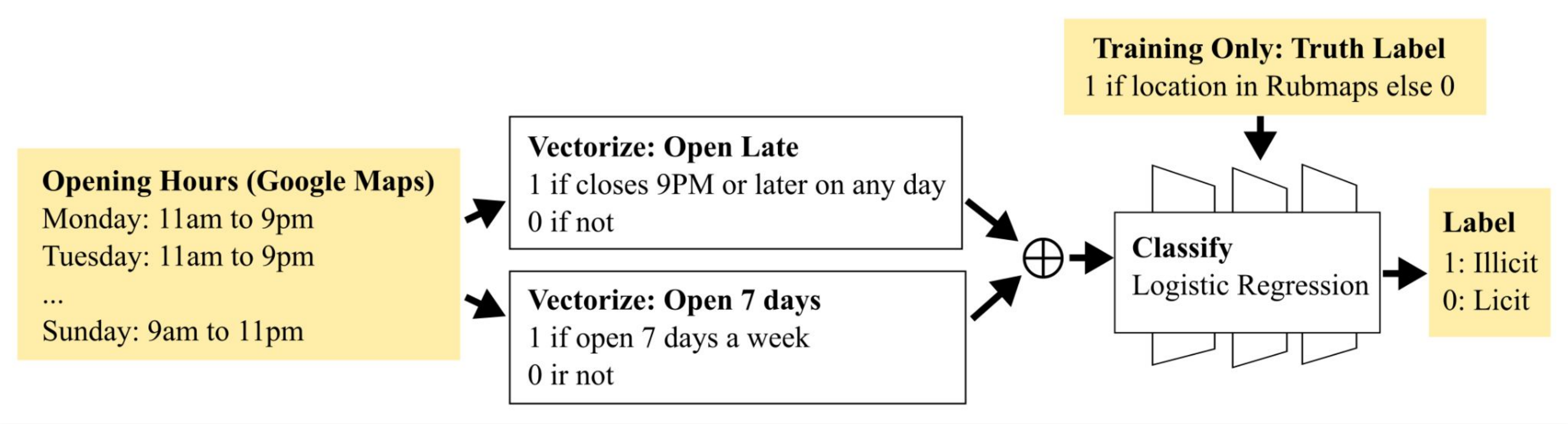


		After Ablation
Accuracy	0.794 (± 0.038)	0.794 (± 0.038)
Precision	0.718 (± 0.057)	0.718 (± 0.057)
Recall	0.797 (± 0.004)	0.797 (± 0.004)
F1	0.754 (± 0.033)	0.754 (± 0.033)
MCC	0.586 (± 0.062)	0.582 (± 0.065)

- About the same, ~80%

Hours only?

-



Training Only: Truth Label
1 if location in Rubmaps else 0

Opening Hours (Google Maps)
Monday: 11am to 9pm
Tuesday: 11am to 9pm
…
Sunday: 9am to 11pm

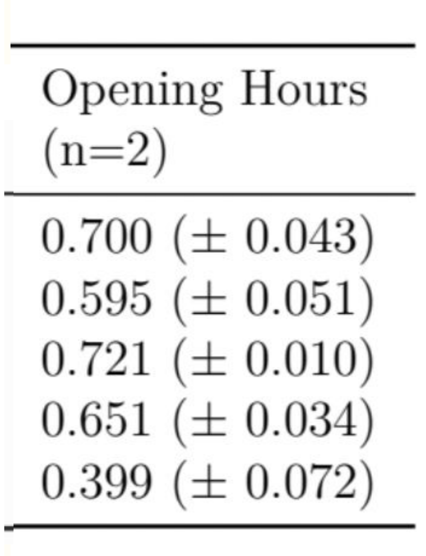
Vectorize: Open Late
1 if closes 9PM or later on any day
0 if not

Vectorize: Open 7 days
1 if open 7 days a week
0 ir not

Classify
Logistic Regression

Label
1: Illicit
0: Licit

Hours only - Results



Opening Hours (n=2)
0.700 (± 0.043)
0.595 (± 0.051)
0.721 (± 0.010)
0.651 (± 0.034)
0.399 (± 0.072)

- Accuracy ~70%

The Google Places Dataset: Illegal Activity in Plain Sight

Part I

Sections

1. Introduction
2. Classifier
3. Vulnerability Insights with Named Entity Recognition
4. Fairness with Ablation & Business Hours

Part I. Summary

- **Bag-of-Word Classifier, DistilBERT**
  Both ~80% accuracy

- **Vulnerability Insights with Named Entity Recognition**
  More likely mention ethnicity, cash, open longer hours / closed less often

- **Fairness with Ablation & Business Hours**
  Ablation: Almost no change
  Business hours: ~70% accuracy

Questions?

break

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Demand

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Sections

1. Introduction
2. Case Study: Domains

3. Case Study: Buyer psychology

   Aside: Acronym expansion
4. Buyer insights
5. Hackathon

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
   - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
    - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

Demand-Side

Rubmaps

- Paywall for reviews
- Login required for forum

Enter…

Demand-Side

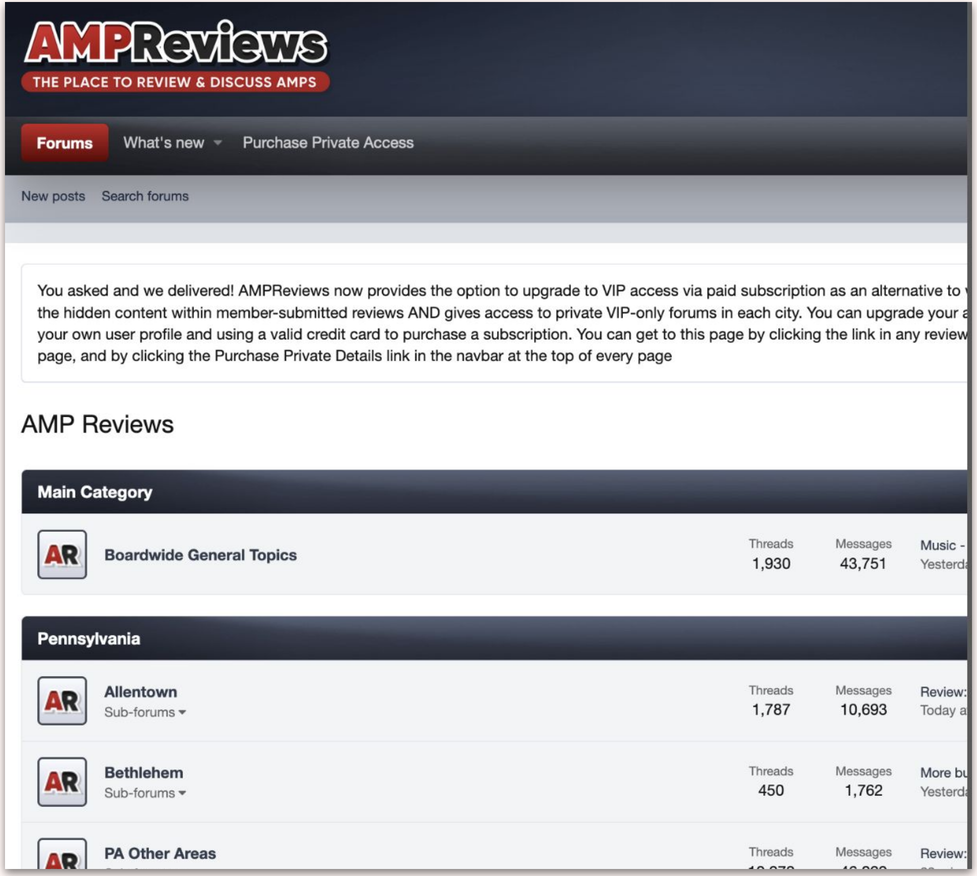
Rubmaps

- Paywall for reviews
- Login required for forum

Enter...

AMPReviews Discussion Forum

- Public: internet archive Wayback Machine



You asked and we delivered! AMPReviews now provides the option to upgrade to VIP access via paid subscription as an alternative to the hidden content within member-submitted reviews AND gives access to private VIP-only forums in each city. You can upgrade your your own user profile and using a valid credit card to purchase a subscription. You can get to this page by clicking the link in any review page, and by clicking the Purchase Private Details link in the navbar at the top of every page

AMP Reviews

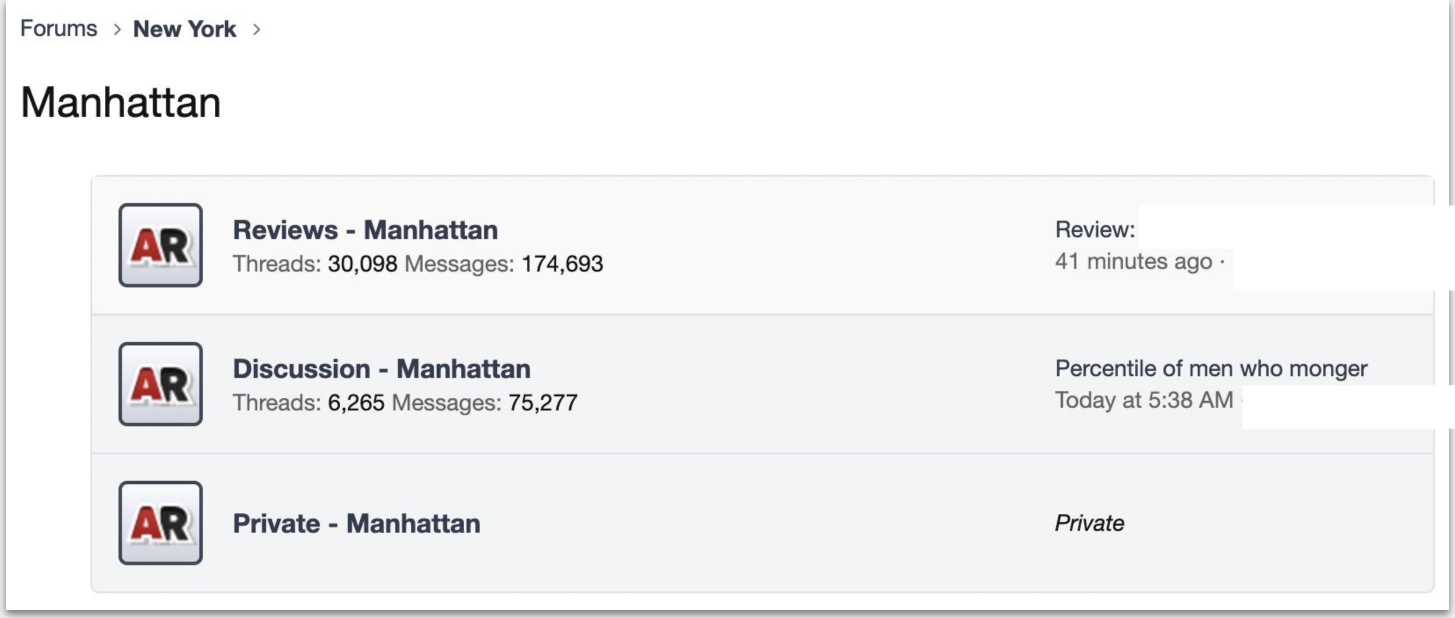
Main Category			
Boardwide General Topics	Threads 1,930	Messages 43,751	Music - Yesterd
Pennsylvania			
Allentown Sub-forums ▾	Threads 1,787	Messages 10,693	Review: Today a
Bethlehem Sub-forums ▾	Threads 450	Messages 1,762	More bu Yesterd
PA Other Areas	Threads	Messages	Review:

Since 2018

- ~620,000 posts
  - 621,636
- ~27,000 users
  - 26,928
- 12 states, ~90k individual visits
  - 90,824

Within each category…

Reviews / Discussions / Private



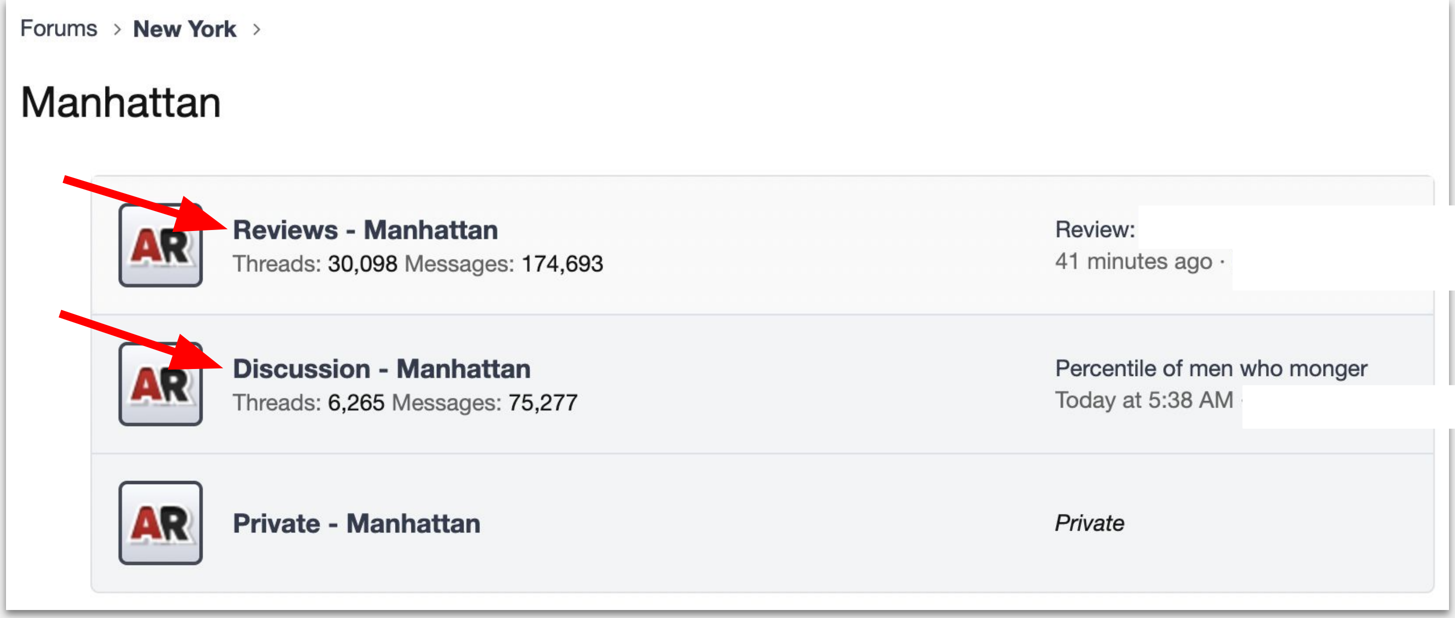
Forums › **New York** ›

Manhattan

Reviews - Manhattan Threads: 30,098 Messages: 174,693	Review: [redacted] 41 minutes ago · [redacted]
Discussion - Manhattan Threads: 6,265 Messages: 75,277	Percentile of men who monger Today at 5:38 AM [redacted]
Private - Manhattan	*Private*

Within each category…

Reviews / Discussions / Private



Forums › **New York** ›

Manhattan

Reviews - Manhattan
Threads: 30,098 Messages: 174,693

Review: [redacted]
41 minutes ago · [redacted]

Discussion - Manhattan
Threads: 6,265 Messages: 75,277

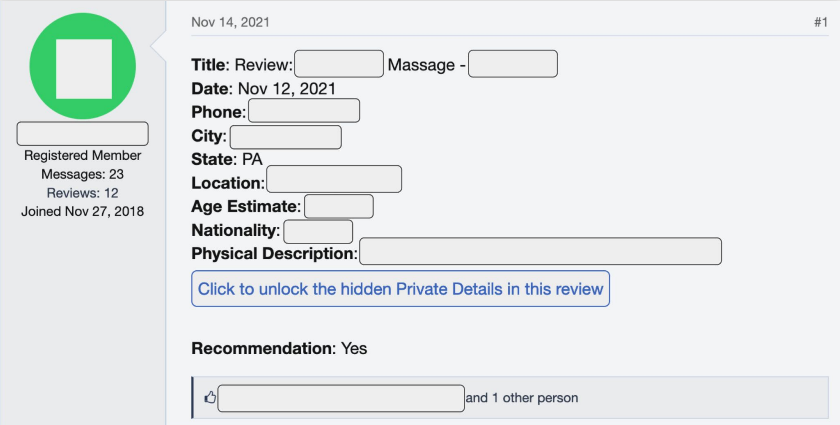
Percentile of men who monger
Today at 5:38 AM · [redacted]

Private - Manhattan

Private

Post types

Reviews - Semi-structured first post



Nov 14, 2021 #1

[redacted]
Registered Member
Messages: 23
Reviews: 12
Joined Nov 27, 2018

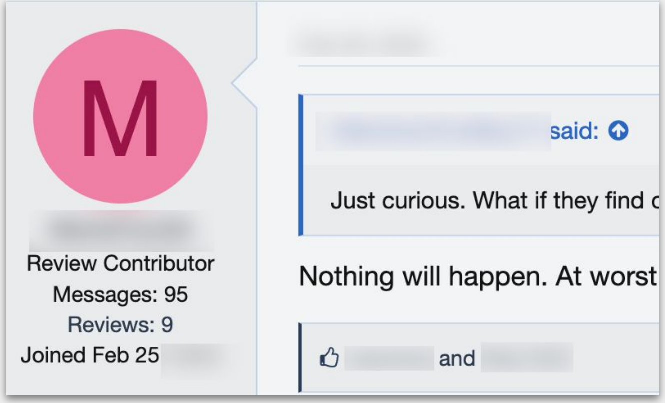
Title: Review: [redacted] Massage - [redacted]
Date: Nov 12, 2021
Phone: [redacted]
City: [redacted]
State: PA
Location: [redacted]
Age Estimate: [redacted]
Nationality: [redacted]
Physical Description: [redacted]

Click to unlock the hidden Private Details in this review

Recommendation: Yes

[redacted] and 1 other person

Posts



[redacted]

[redacted]
Review Contributor
Messages: 95
Reviews: 9
Joined Feb 25 [redacted]

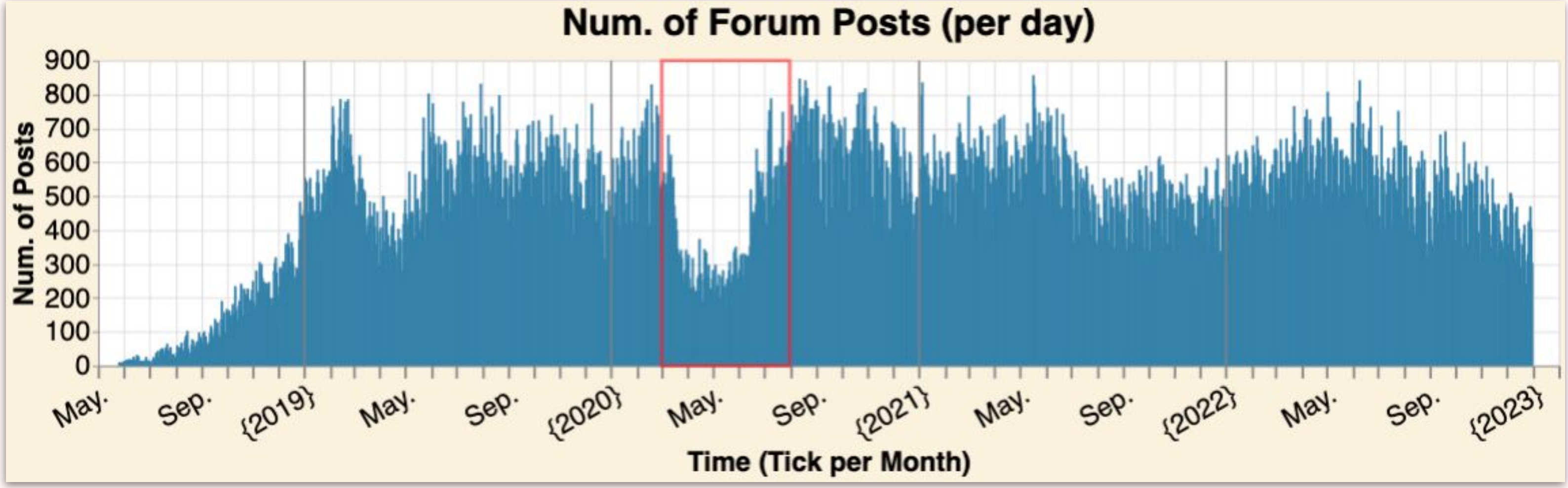
[redacted] said:

Just curious. What if they find c

Nothing will happen. At worst

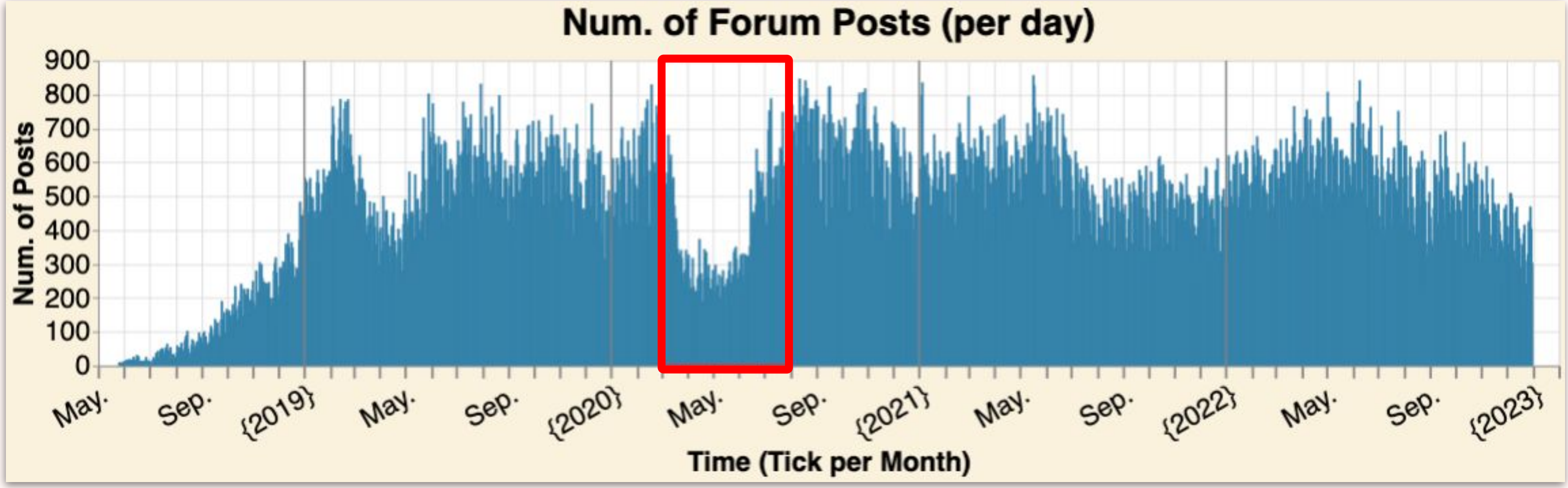
[redacted] and [redacted]

Exploratory Data Analysis (EDA)



Downtick from COVID

Exploratory Data Analysis (EDA)



Downtick from COVID

The Forum Dataset: Two Case Studies and a Hackathon

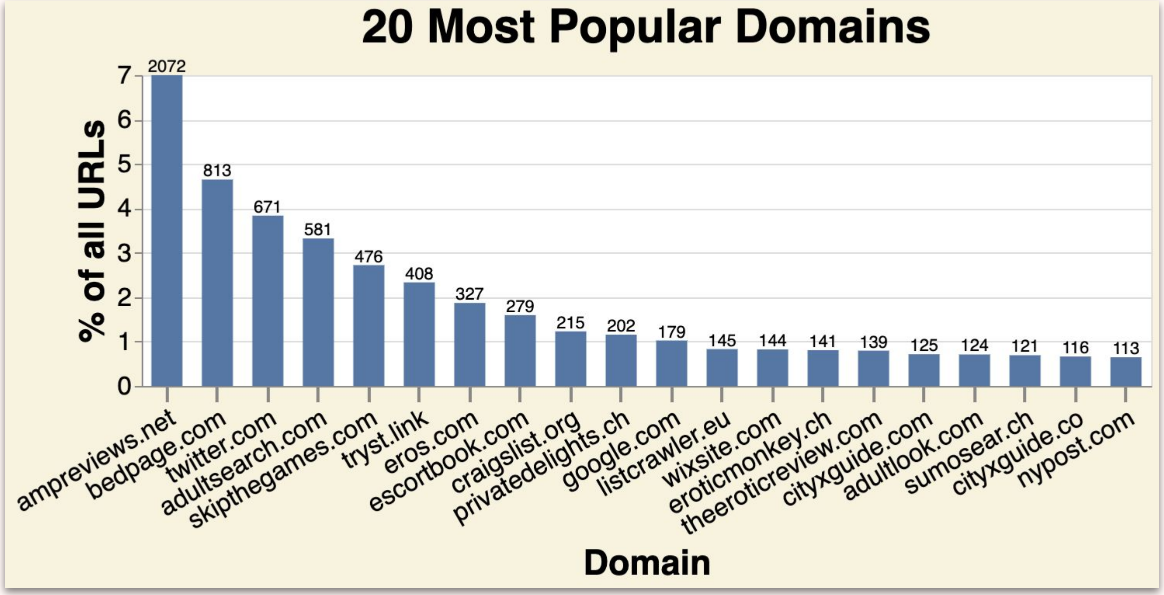
Part II

Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
   - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

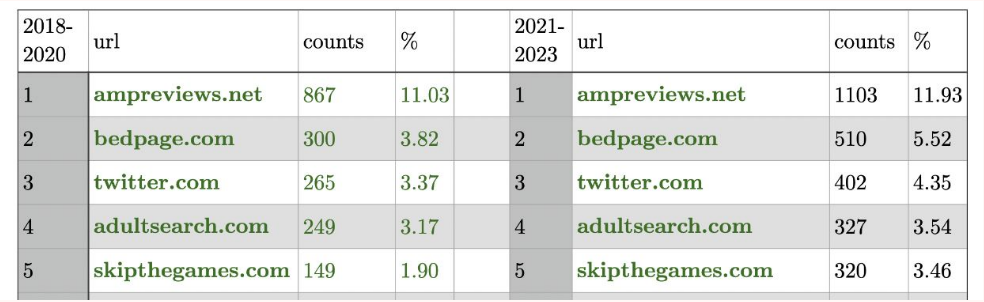
Monitoring: What are the top domains?

- Extract from HTML tags
- Prune to top-level domain



Split by Time

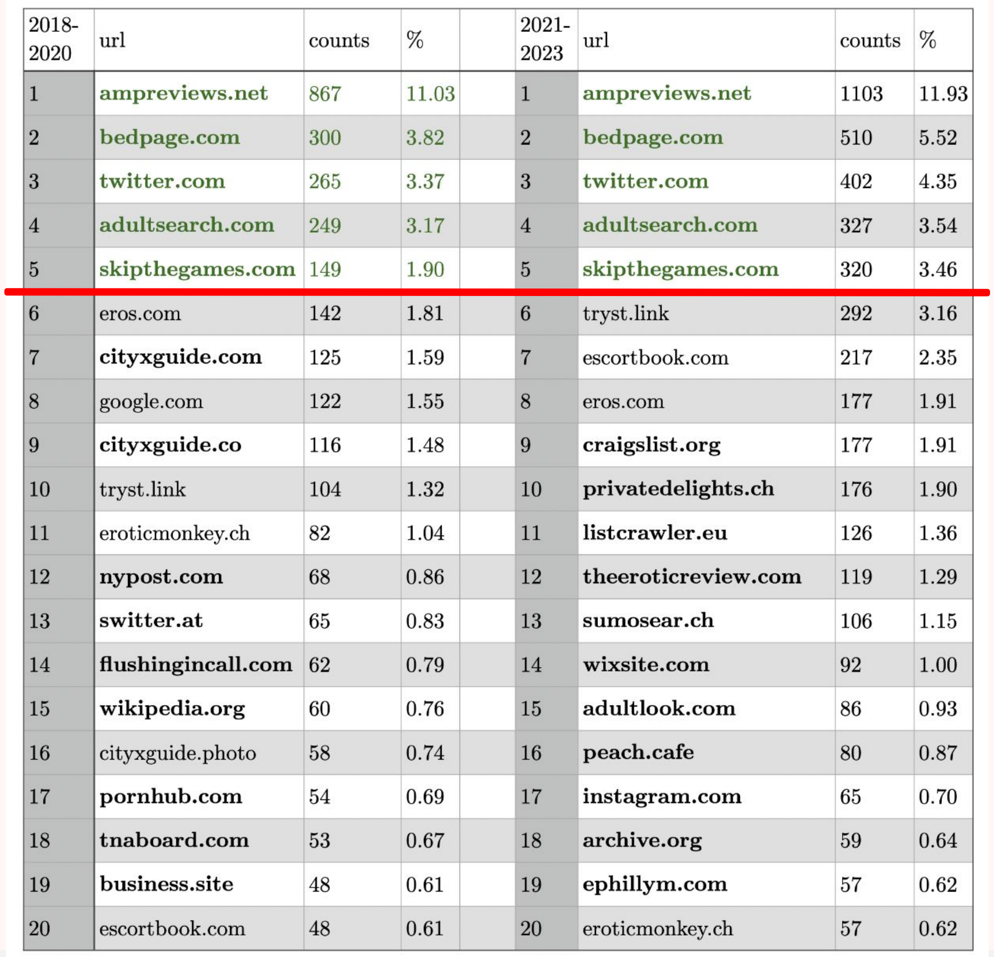
- Top 5 remain stable



2018-2020	url	counts	%		2021-2023	url	counts	%
1	**ampreviews.net**	867	11.03		1	**ampreviews.net**	1103	11.93
2	**bedpage.com**	300	3.82		2	**bedpage.com**	510	5.52
3	**twitter.com**	265	3.37		3	**twitter.com**	402	4.35
4	**adultsearch.com**	249	3.17		4	**adultsearch.com**	327	3.54
5	**skipthegames.com**	149	1.90		5	**skipthegames.com**	320	3.46

Split by Time

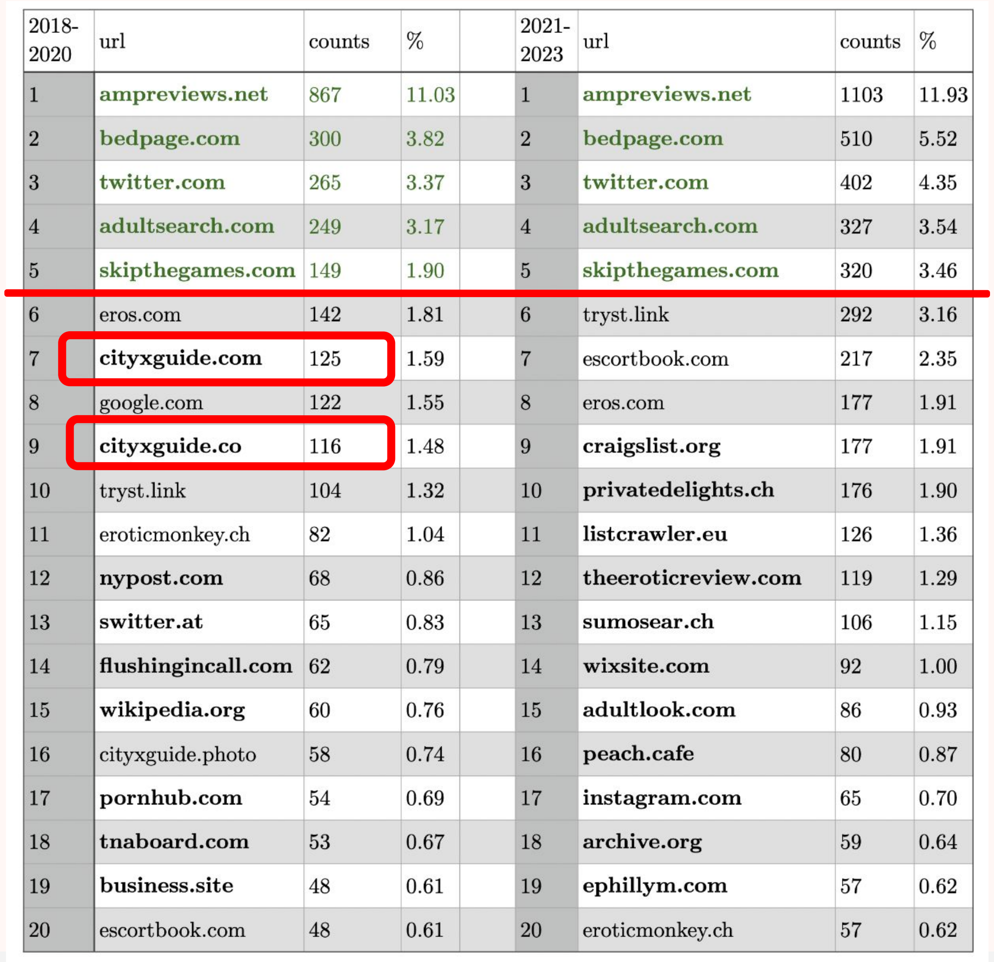
- Top 5 remain stable



2018-2020	url	counts	%		2021-2023	url	counts	%
1	**ampreviews.net**	867	11.03		1	**ampreviews.net**	1103	11.93
2	**bedpage.com**	300	3.82		2	**bedpage.com**	510	5.52
3	**twitter.com**	265	3.37		3	**twitter.com**	402	4.35
4	**adultsearch.com**	249	3.17		4	**adultsearch.com**	327	3.54
5	**skipthegames.com**	149	1.90		5	**skipthegames.com**	320	3.46
6	eros.com	142	1.81		6	tryst.link	292	3.16
7	**cityxguide.com**	125	1.59		7	escortbook.com	217	2.35
8	google.com	122	1.55		8	eros.com	177	1.91
9	**cityxguide.co**	116	1.48		9	**craigslist.org**	177	1.91
10	tryst.link	104	1.32		10	**privatedelights.ch**	176	1.90
11	eroticmonkey.ch	82	1.04		11	**listcrawler.eu**	126	1.36
12	**nypost.com**	68	0.86		12	**theeroticreview.com**	119	1.29
13	**switter.at**	65	0.83		13	**sumosear.ch**	106	1.15
14	**flushingincall.com**	62	0.79		14	**wixsite.com**	92	1.00
15	**wikipedia.org**	60	0.76		15	**adultlook.com**	86	0.93
16	cityxguide.photo	58	0.74		16	**peach.cafe**	80	0.87
17	**pornhub.com**	54	0.69		17	**instagram.com**	65	0.70
18	**tnaboard.com**	53	0.67		18	**archive.org**	59	0.64
19	**business.site**	48	0.61		19	**ephillym.com**	57	0.62
20	escortbook.com	48	0.61		20	eroticmonkey.ch	57	0.62

Split by Time

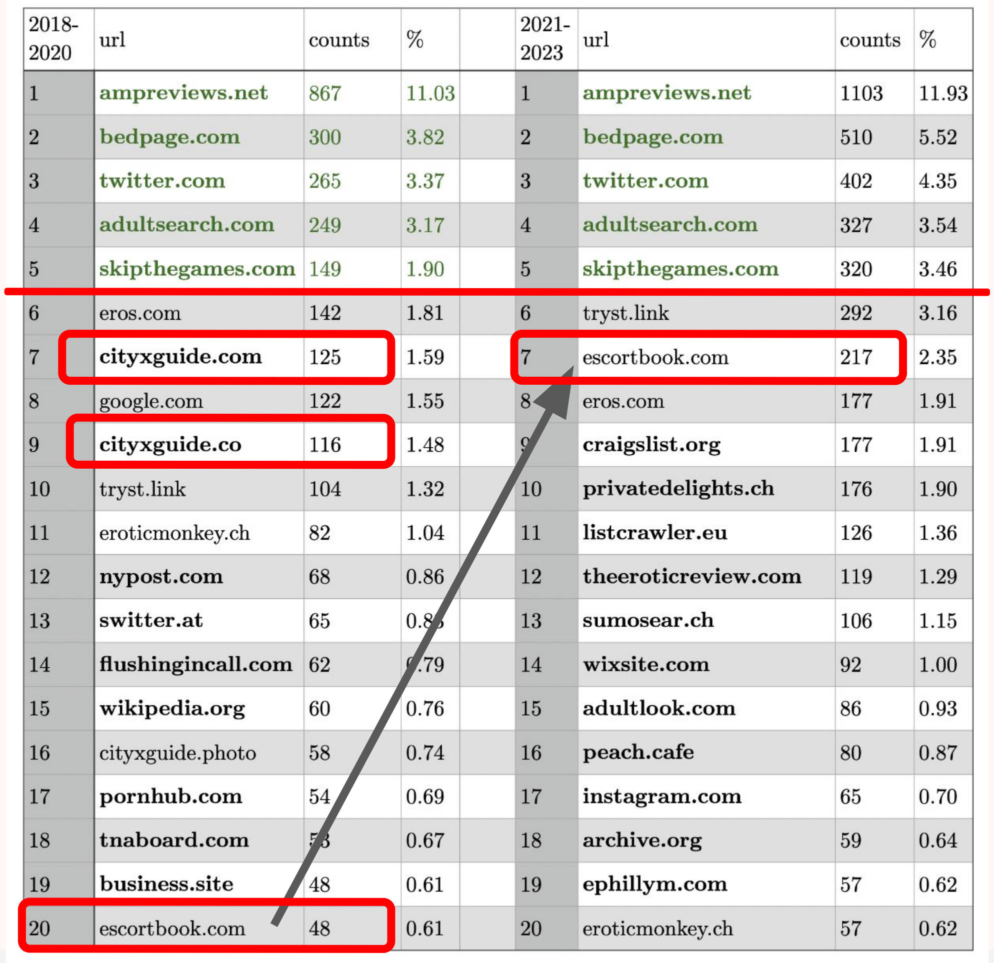
- Top 5 remain stable
- Cityxguide disappears



2018-2020	url	counts	%		2021-2023	url	counts	%
1	**ampreviews.net**	867	11.03		1	**ampreviews.net**	1103	11.93
2	**bedpage.com**	300	3.82		2	**bedpage.com**	510	5.52
3	**twitter.com**	265	3.37		3	**twitter.com**	402	4.35
4	**adultsearch.com**	249	3.17		4	**adultsearch.com**	327	3.54
5	**skipthegames.com**	149	1.90		5	**skipthegames.com**	320	3.46
6	eros.com	142	1.81		6	tryst.link	292	3.16
7	**cityxguide.com**	125	1.59		7	escortbook.com	217	2.35
8	google.com	122	1.55		8	eros.com	177	1.91
9	**cityxguide.co**	116	1.48		9	**craigslist.org**	177	1.91
10	tryst.link	104	1.32		10	**privatedelights.ch**	176	1.90
11	eroticmonkey.ch	82	1.04		11	**listcrawler.eu**	126	1.36
12	**nypost.com**	68	0.86		12	**theeroticreview.com**	119	1.29
13	**switter.at**	65	0.83		13	**sumosear.ch**	106	1.15
14	**flushingincall.com**	62	0.79		14	**wixsite.com**	92	1.00
15	**wikipedia.org**	60	0.76		15	**adultlook.com**	86	0.93
16	cityxguide.photo	58	0.74		16	**peach.cafe**	80	0.87
17	**pornhub.com**	54	0.69		17	**instagram.com**	65	0.70
18	**tnaboard.com**	53	0.67		18	**archive.org**	59	0.64
19	**business.site**	48	0.61		19	**ephillym.com**	57	0.62
20	escortbook.com	48	0.61		20	eroticmonkey.ch	57	0.62

Split by Time

- Top 5 remain stable
- Cityxguide disappears
- Escortbook moves up



2018-2020	url	counts	%		2021-2023	url	counts	%
1	**ampreviews.net**	867	11.03		1	**ampreviews.net**	1103	11.93
2	**bedpage.com**	300	3.82		2	**bedpage.com**	510	5.52
3	**twitter.com**	265	3.37		3	**twitter.com**	402	4.35
4	**adultsearch.com**	249	3.17		4	**adultsearch.com**	327	3.54
5	**skipthegames.com**	149	1.90		5	**skipthegames.com**	320	3.46
6	eros.com	142	1.81		6	tryst.link	292	3.16
7	**cityxguide.com**	125	1.59		7	escortbook.com	217	2.35
8	google.com	122	1.55		8	eros.com	177	1.91
9	**cityxguide.co**	116	1.48		9	**craigslist.org**	177	1.91
10	tryst.link	104	1.32		10	**privatedelights.ch**	176	1.90
11	eroticmonkey.ch	82	1.04		11	**listcrawler.eu**	126	1.36
12	**nypost.com**	68	0.86		12	**theeroticreview.com**	119	1.29
13	**switter.at**	65	0.83		13	**sumosear.ch**	106	1.15
14	**flushingincall.com**	62	0.79		14	**wixsite.com**	92	1.00
15	**wikipedia.org**	60	0.76		15	**adultlook.com**	86	0.93
16	cityxguide.photo	58	0.74		16	**peach.cafe**	80	0.87
17	**pornhub.com**	54	0.69		17	**instagram.com**	65	0.70
18	**tnaboard.com**	53	0.67		18	**archive.org**	59	0.64
19	**business.site**	48	0.61		19	**ephillym.com**	57	0.62
20	escortbook.com	48	0.61		20	eroticmonkey.ch	57	0.62

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
    - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

Reducing Demand with Buyer Psychology

- What are the top concerns of buyers?
  - Law Enforcement
  - STDs

    …
  - Relationships?

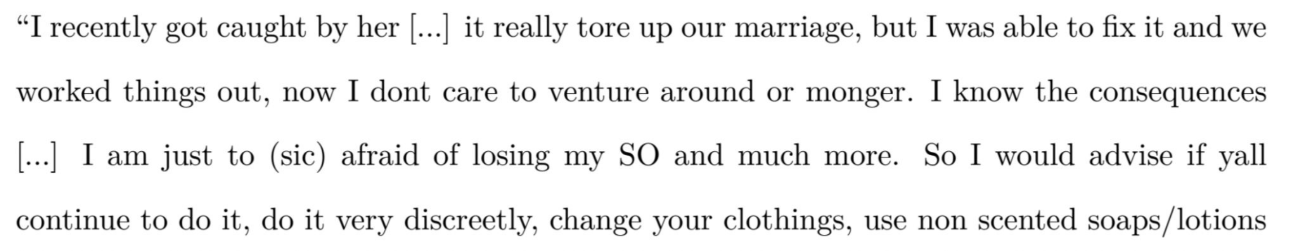
- Claims: most are married

Reducing Demand with Buyer Psychology

- What are the top concerns of buyers?
    - Law Enforcement
    - STDs

      …
    - Relationships?

- Claims: most are married



"I recently got caught by her [...] it really tore up our marriage, but I was able to fix it and we worked things out, now I dont care to venture around or monger. I know the consequences [...] I am just to (sic) afraid of losing my SO and much more. So I would advise if yall continue to do it, do it very discreetly, change your clothings, use non scented soaps/lotions

Preliminary Investigation

- Traditional method:
  Manual coding

- New method:
  Word embeddings / topic models

- Word2Vec

Word2Vec

- Predict context of words
- e.g. given "The cat is " → red, black

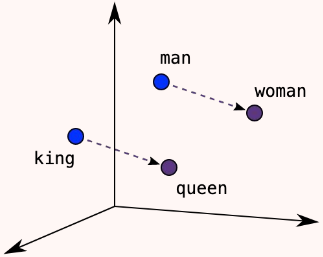
Trained on my data

- Custom pre-processing
  - Strip punctuation (S.O. → SO)
  - Kept capitalization (SO ≠ so)
  - Fewer stop words (and, the, so)
  - Kept words of length two

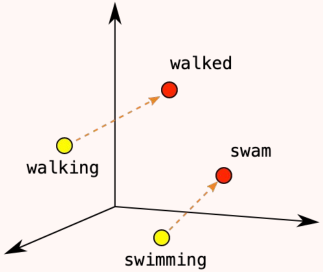
Word2Vec

- Predict context of words
- e.g. given "The cat is " → red, black

Vectors: Distance and direction



Male-Female



Verb Tense

Trained on my data

- Custom pre-processing
    - Strip punctuation (S.O. → SO)
    - Kept capitalization (SO ≠ so)
    - Fewer stop words (and, the, so)
    - Kept words of length two

Seed words

- **Negative sentiments:**
    - **worry**
    - **anxiety**

Seed words

- **Negative sentiments:**
  - **worry**
  - **anxiety**
- Hypothesized concerns:
  - SO (for significant other)
  - marriage
- Other concerns:
  - LEO (for law enforcement officer)
  - STD (sexually transmitted disease)
- Control words:
  - provider
  - parking
  - table

Seed words

- **Negative sentiments:**
    - **worry**
    - **anxiety**
- Hypothesized concerns:
    - SO (for significant other)
    - marriage
- Other concerns:
    - LEO (for law enforcement officer)
    - STD (sexually transmitted disease)
- Control words:
    - provider
    - parking
    - table

- Hypothesis:

distance_worry(marriage)

≈

distance_worry(LEO)

>

distance_worry(table)

Seed words

- **Negative sentiments:**
  - **worry**
  - **anxiety**
- Hypothesized concerns:
  - SO (for significant other)
  - marriage
- Other concerns:
  - LEO (for law enforcement officer)
  - STD (sexually transmitted disease)
- Control words:
  - provider
  - parking
  - table

- Hypothesis:

distance_worry(marriage)

≈

distance_worry(LEO)

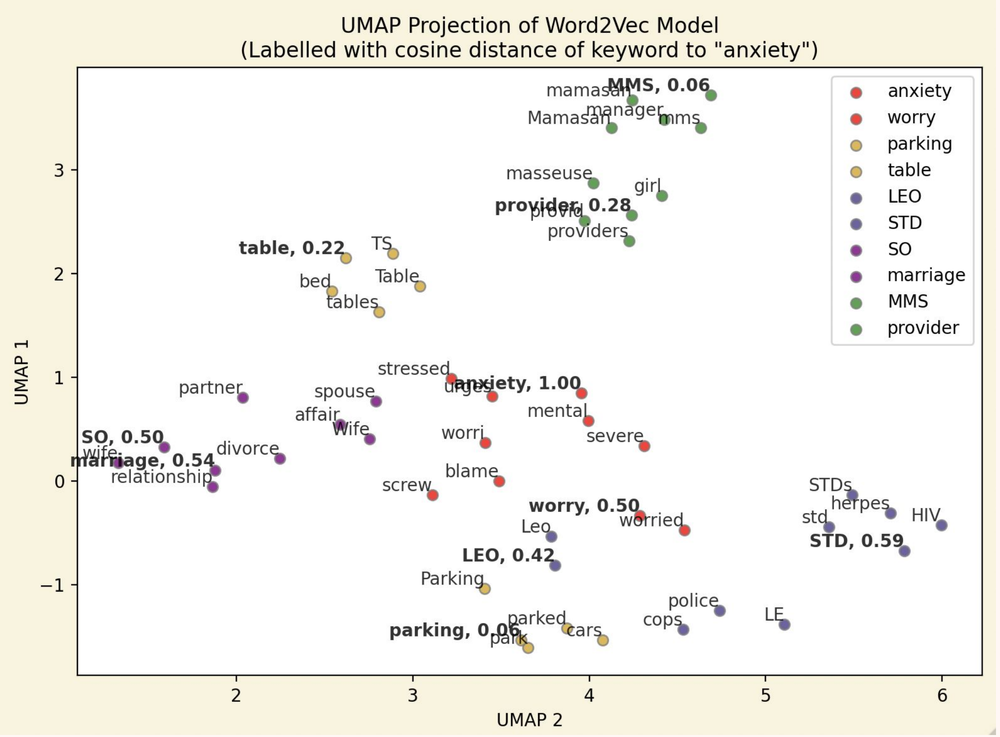
>

distance_worry(table)

For visualization:

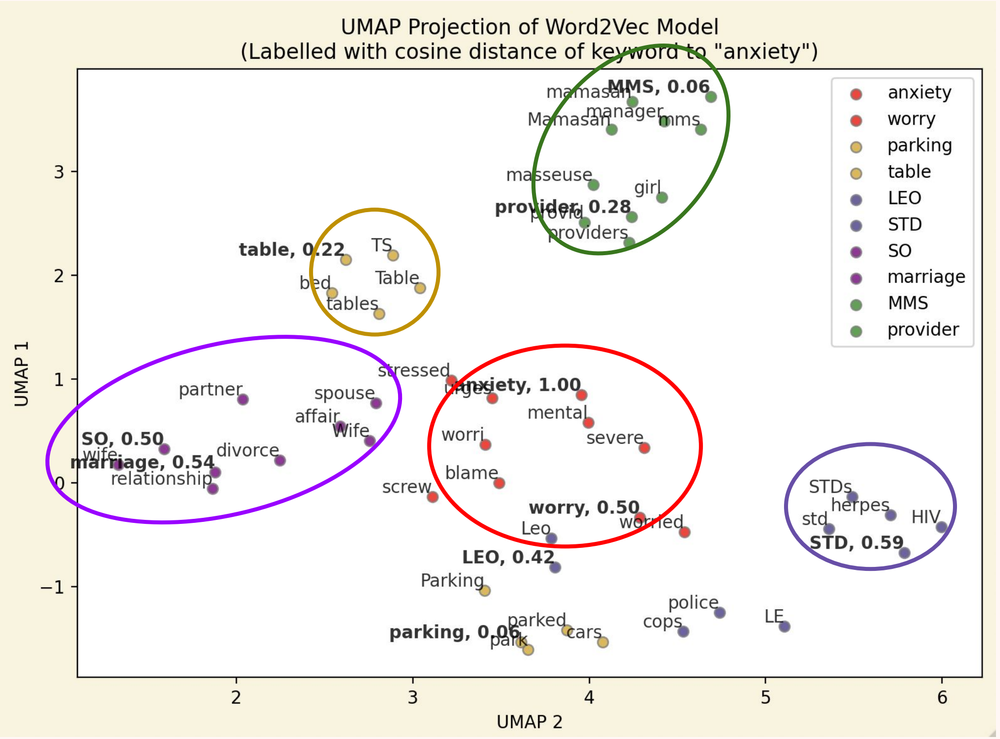
- Find closest 5 words in word embedding space

Project to lower dimension



UMAP Projection of Word2Vec Model
(Labelled with cosine distance of keyword to "anxiety")

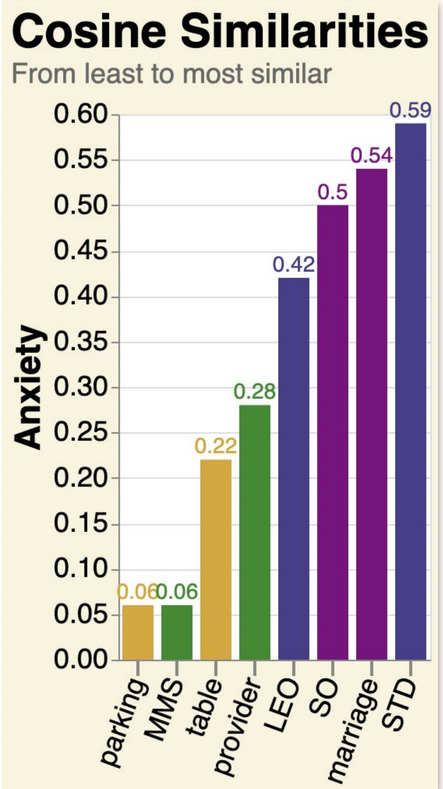
Project to lower dimension



Not visually conclusive

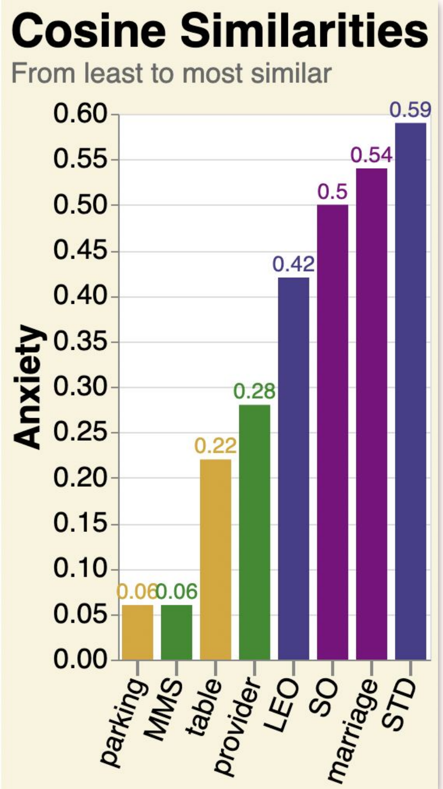
Directly graph seed words

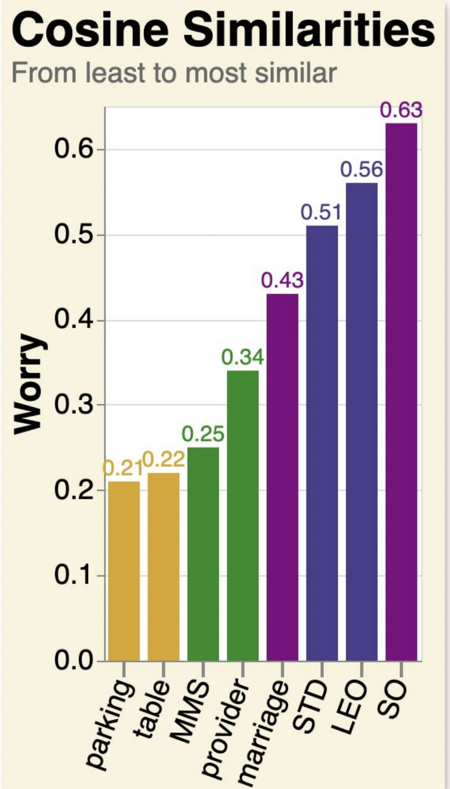
- Distance from "anxiety", "worry"



Directly graph seed words

- Distance from "anxiety", "worry"





The Forum Dataset: Two Case Studies and a Hackathon

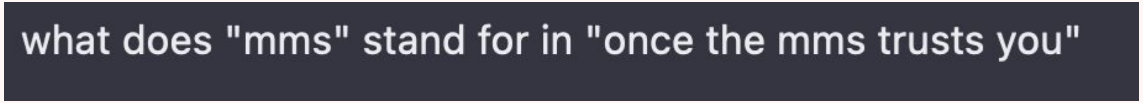
Part II

Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
   Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

Aside: Acronyms!

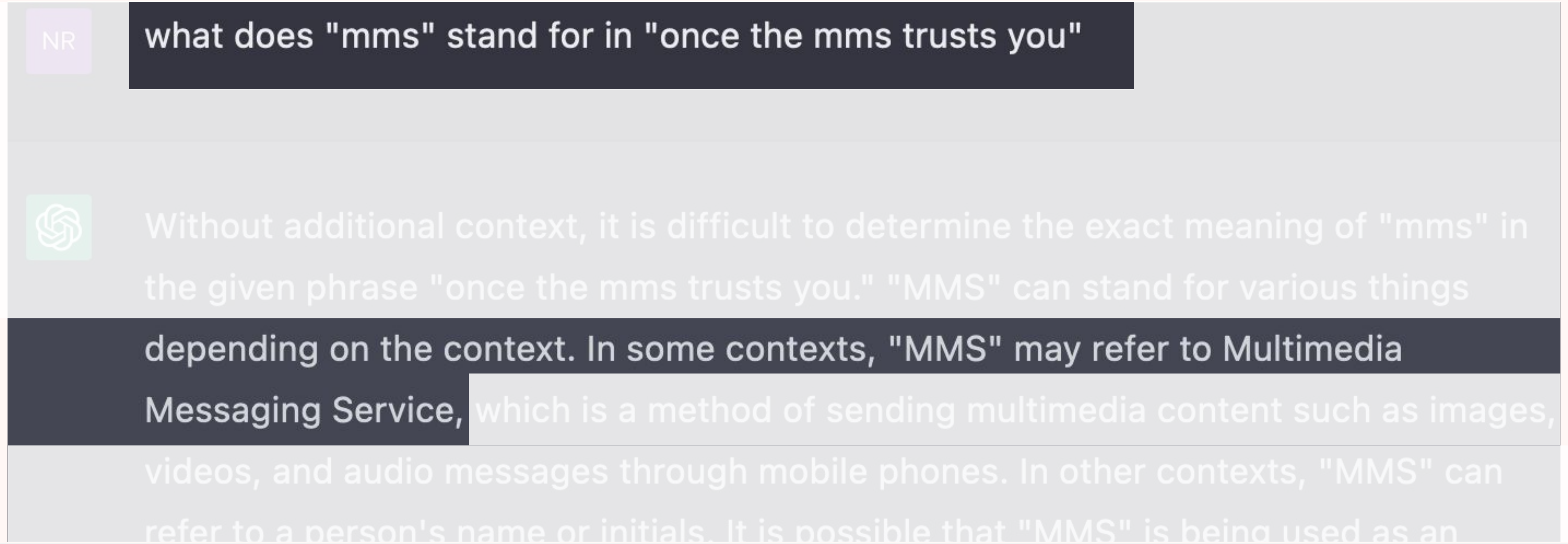
A lot of acronyms used:



what does "mms" stand for in "once the mms trusts you"

Aside: Word embeddings are interesting

A lot of acronyms used:



what does "mms" stand for in "once the mms trusts you"

Without additional context, it is difficult to determine the exact meaning of "mms" in the given phrase "once the mms trusts you." "MMS" can stand for various things depending on the context. In some contexts, "MMS" may refer to Multimedia Messaging Service, which is a method of sending multimedia content such as images, videos, and audio messages through mobile phones. In other contexts, "MMS" can refer to a person's name or initials. It is possible that "MMS" is being used as an

Word2Vec embeddings

Nearest to "MMS":

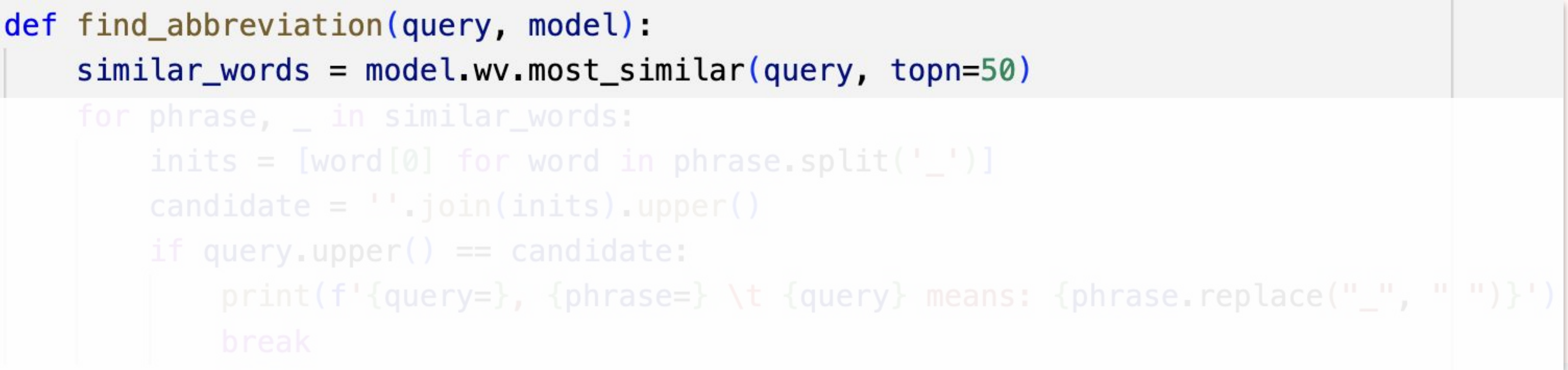
- mms = 0.89
- **mamasan = 0.83,**
- manager = 0.80,
- Mamasan = 0.76

Retrain word2vec with bigrams

Bigrams are constructed with underscores

- happy_ending
  <u>**h**</u>appy <u>**e**</u>nding
  HE

- Can expand two-letter initialisms, e.g. HE



```
def find_abbreviation(query, model):
    similar_words = model.wv.most_similar(query, topn=50)
    for phrase, _ in similar_words:
        inits = [word[0] for word in phrase.split('_')]
        candidate = ''.join(inits).upper()
        if query.upper() == candidate:
            print(f'{query=}, {phrase=} \t {query} means: {phrase.replace("_", " ")}')
            break
```

The Forum Dataset: Two Case Studies and a Hackathon

Part II

Sections

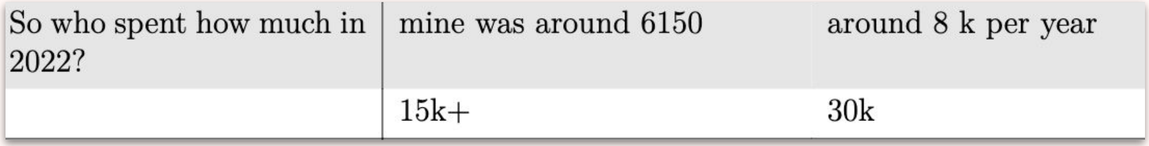
1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
    - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

Buyer Insights

- How much are buyers spending?
- How frequently do they go?

Buyer Insights

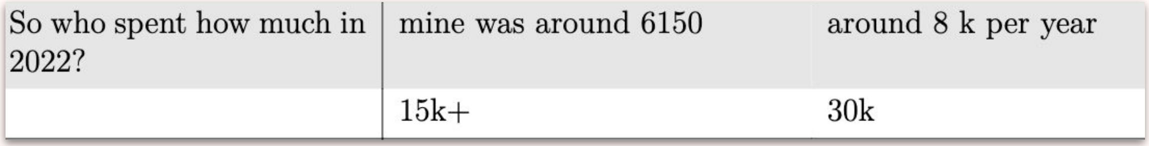
- How much are buyers spending?
- How frequently do they go?



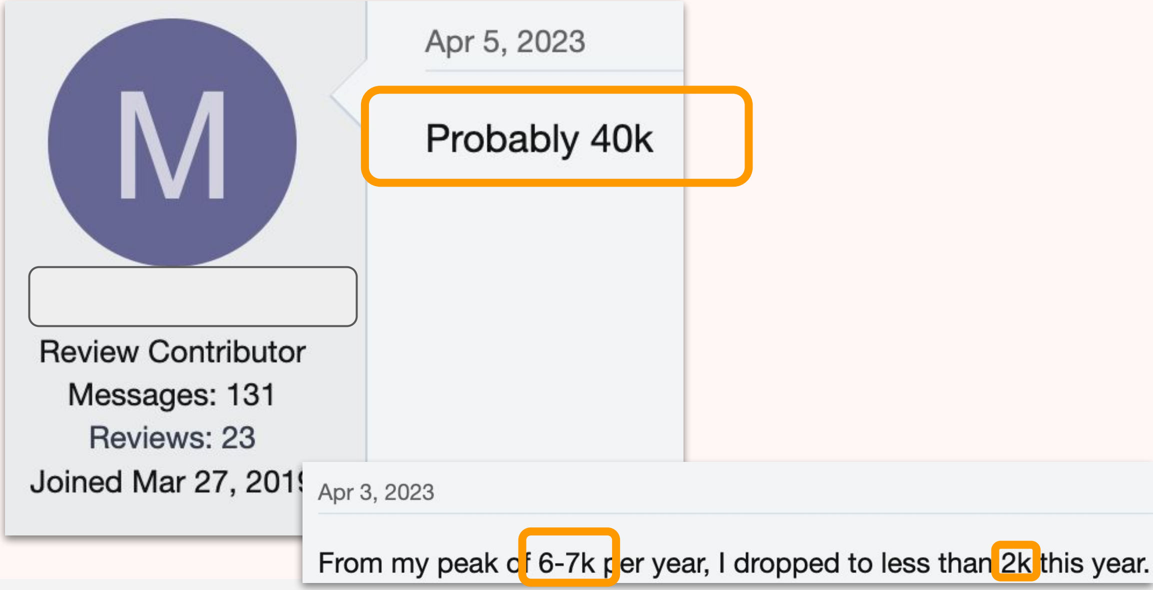
So who spent how much in 2022?	mine was around 6150	around 8 k per year
	15k+	30k

Buyer Insights

- How much are buyers spending?
- How frequently do they go?



So who spent how much in 2022?	mine was around 6150	around 8 k per year
	15k+	30k



Review Contributor
Messages: 131
Reviews: 23
Joined Mar 27, 201

Apr 5, 2023

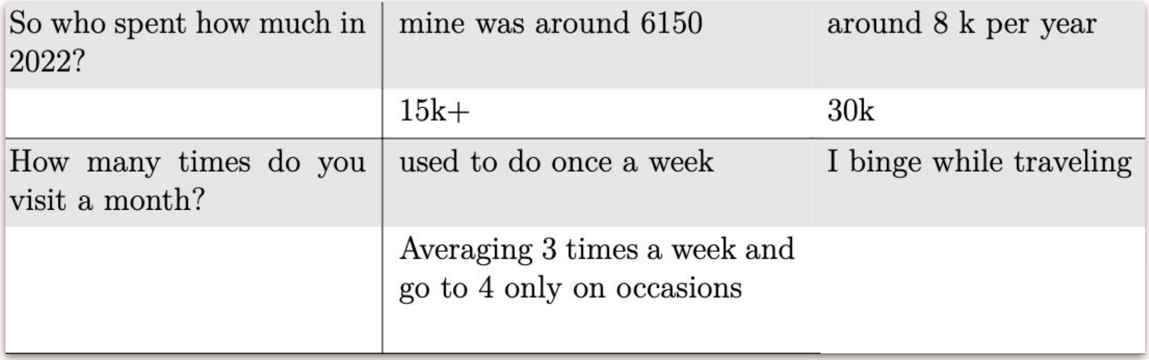
Probably 40k

Apr 3, 2023

From my peak of 6-7k per year, I dropped to less than 2k this year.

Buyer Insights

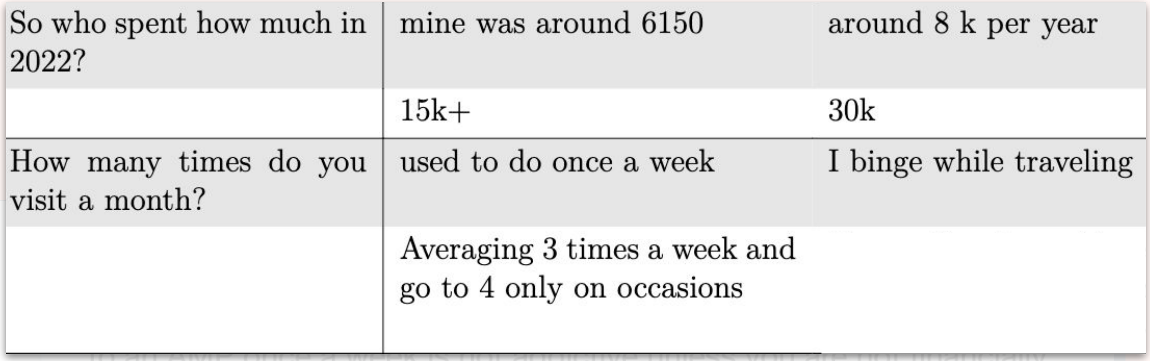
- How much are buyers spending?
- How frequently do they go?



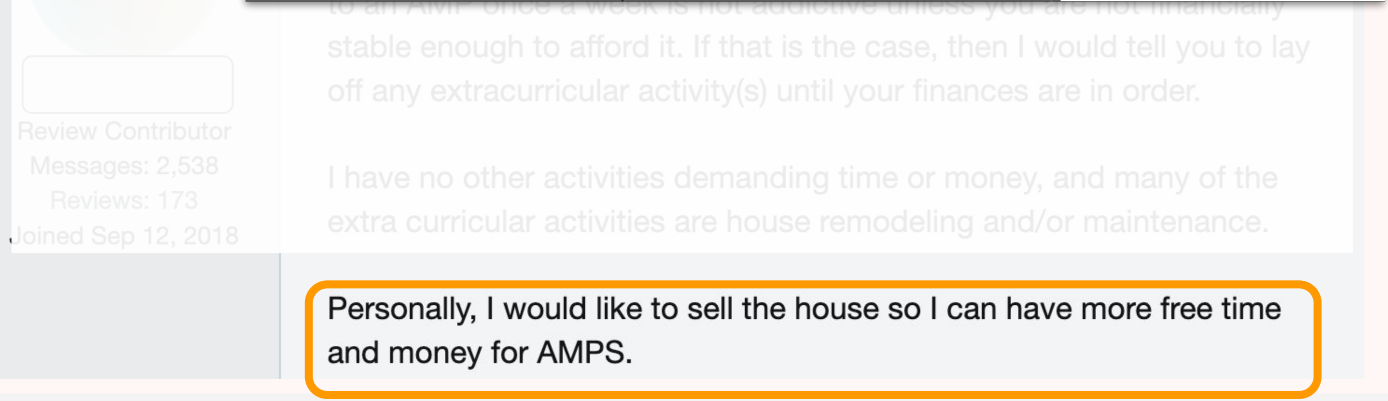
So who spent how much in 2022?	mine was around 6150	around 8 k per year
	15k+	30k
How many times do you visit a month?	used to do once a week	I binge while traveling
	Averaging 3 times a week and go to 4 only on occasions	

Buyer Insights

- How much are buyers spending?
- How frequently do they go?



So who spent how much in 2022?	mine was around 6150	around 8 k per year
	15k+	30k
How many times do you visit a month?	used to do once a week	I binge while traveling
	Averaging 3 times a week and go to 4 only on occasions	



Review Contributor
Messages: 2,538
Reviews: 173
Joined Sep 12, 2018

stable enough to afford it. If that is the case, then I would tell you to lay off any extracurricular activity(s) until your finances are in order.

I have no other activities demanding time or money, and many of the extra curricular activities are house remodeling and/or maintenance.

Personally, I would like to sell the house so I can have more free time and money for AMPS.

Buyer Demographics

- Income

  Above average
- Age
  - “I am almost 40 and have been mongering since my early 20s”
  - “If they are 45, they are still 25 years younger than me”
  - “been there a couple times when I was in grad school”
- Occupation
  - “I work: in finance and logistics, in a medical setting, in pharmaceutical consulting, in tech
  - “Being in banking for over 30 years [...]”

Buyer Stereotypes

Commonly white male, but ...

- Ethnicity
    - “I’m ½ Chinese/Korean” “I'm Hispanic”
- Gender
    - “I saw a female monger on rubmaps that contributes reviews quite often”

Usernames

有

田

스

오

이

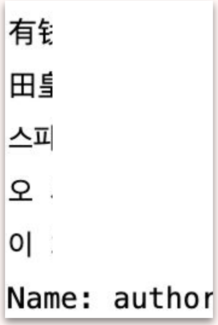
Name: author

Buyer Stereotypes

Commonly white male, but ...

- Ethnicity
    - “I’m ½ Chinese/Korean” “I'm Hispanic”
- Gender
    - “I saw a female monger on rubmaps that contributes reviews quite often”
- Relationship status
    - “married once , never again. [...] I have my adult children & grand-children”
    - “If I had to guess, 85%+ of mongers are married or in a committed relationship”

Usernames



有

田

스

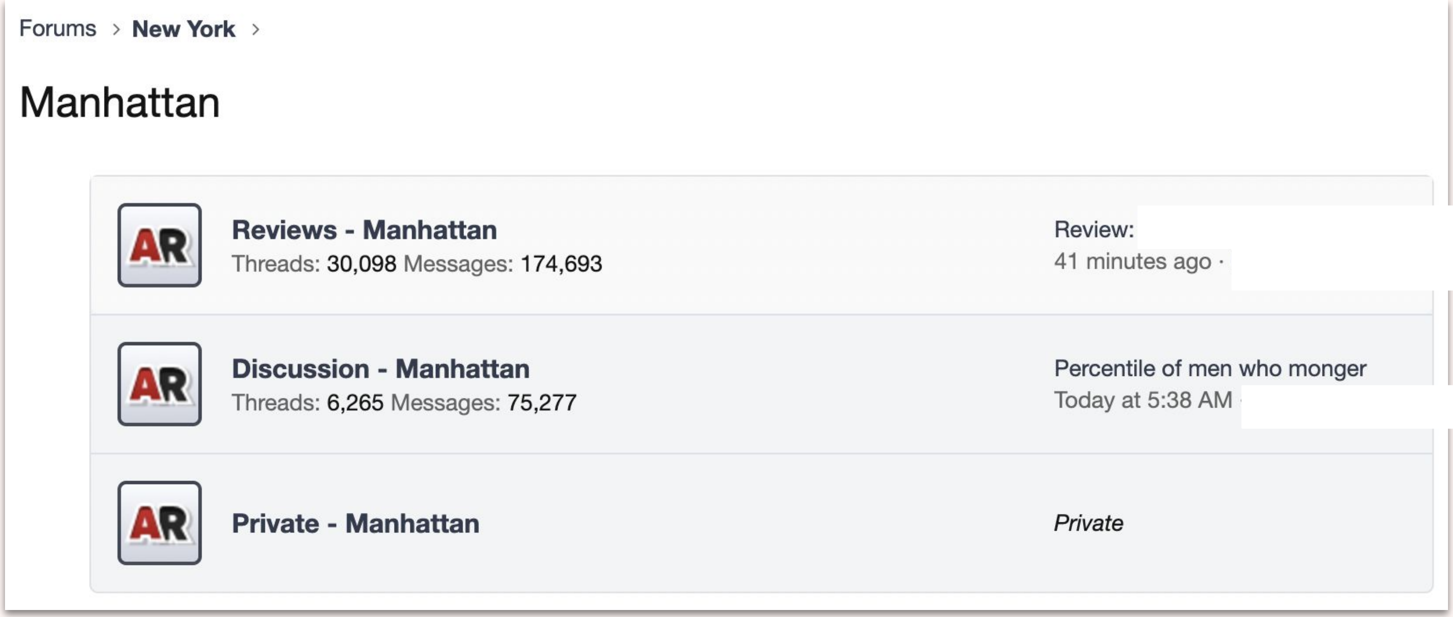
오

이

Name: author

There is more information…

Reviews / Discussions / Private



Forums › **New York** ›

Manhattan

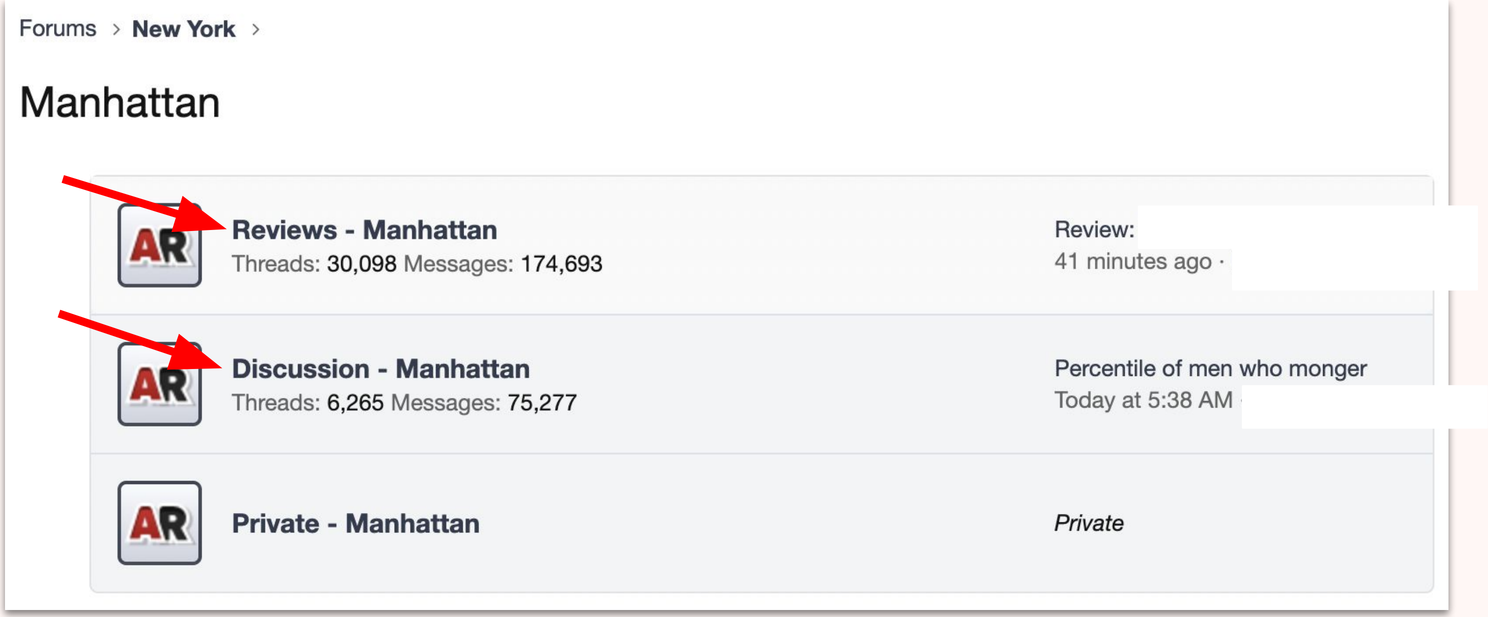
Reviews - Manhattan
Threads: 30,098 Messages: 174,693
Review: [redacted]
41 minutes ago ·

Discussion - Manhattan
Threads: 6,265 Messages: 75,277
Percentile of men who monger
Today at 5:38 AM · [redacted]

Private - Manhattan
Private

There is more information…

Reviews / Discussions / Private



Forums › **New York** ›

Manhattan

Reviews - Manhattan
Threads: 30,098 Messages: 174,693

Review: [redacted]
41 minutes ago ·

Discussion - Manhattan
Threads: 6,265 Messages: 75,277

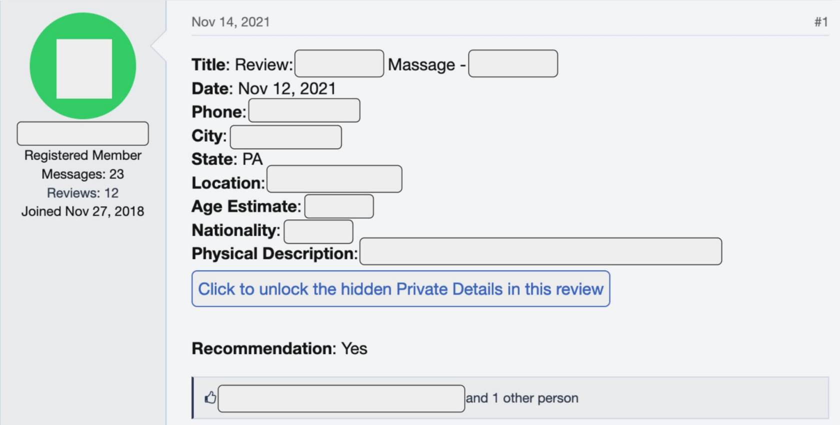
Percentile of men who monger
Today at 5:38 AM · [redacted]

Private - Manhattan

Private

Post types

Reviews - Semi-structured first post



[redacted]
Registered Member
Messages: 23
Reviews: 12
Joined Nov 27, 2018

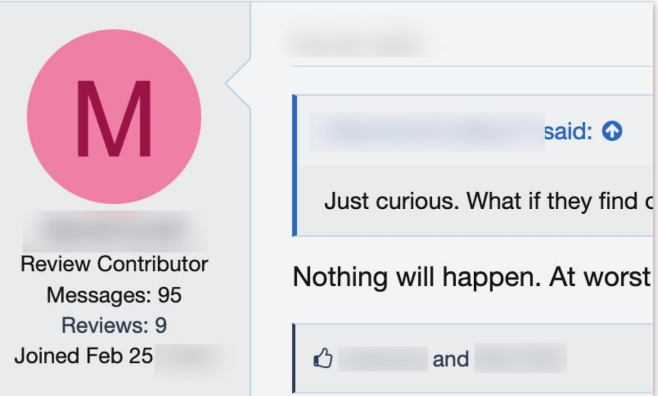
Nov 14, 2021 #1

Title: Review: [redacted] Massage - [redacted]
Date: Nov 12, 2021
Phone: [redacted]
City: [redacted]
State: PA
Location: [redacted]
Age Estimate: [redacted]
Nationality: [redacted]
Physical Description: [redacted]

Recommendation: Yes

[redacted] and 1 other person

Posts



[redacted]
Review Contributor
Messages: 95
Reviews: 9
Joined Feb 25 [redacted]

[redacted] said:

Just curious. What if they find c

Nothing will happen. At worst

[redacted] and [redacted]

The Forum Dataset: Two Case Studies and a Hackathon

Part II

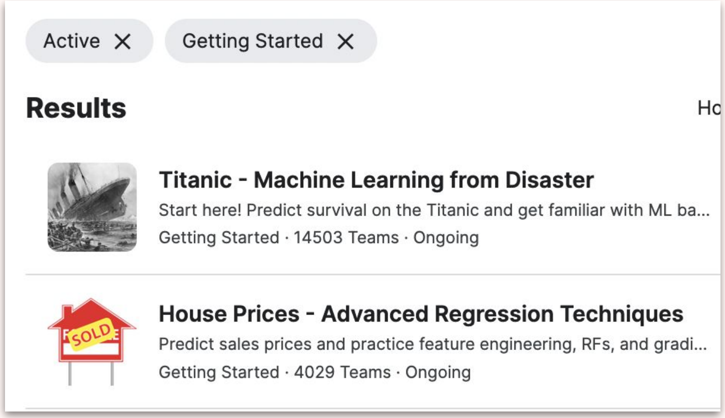
Sections

1. Introduction
2. Case Study: **Monitoring** with domain extraction
3. Case Study: **Reducing demand** with buyer psychology
    - Aside: Acronym expansion
4. Buyer insights
5. **Raising public awareness** with a hackathon

Raising Awareness with Hackathons

- Kaggle Competitions

- First step: Open datasets!
    - MPForum dataset at: https://kaggle.com/datasets/34ab6a6b2f6166fe59b77815e3922f1f835770d08bd827babc7bb4e9d31bbd4b



Results

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML ba...

Getting Started · 14503 Teams · Ongoing

House Prices - Advanced Regression Techniques

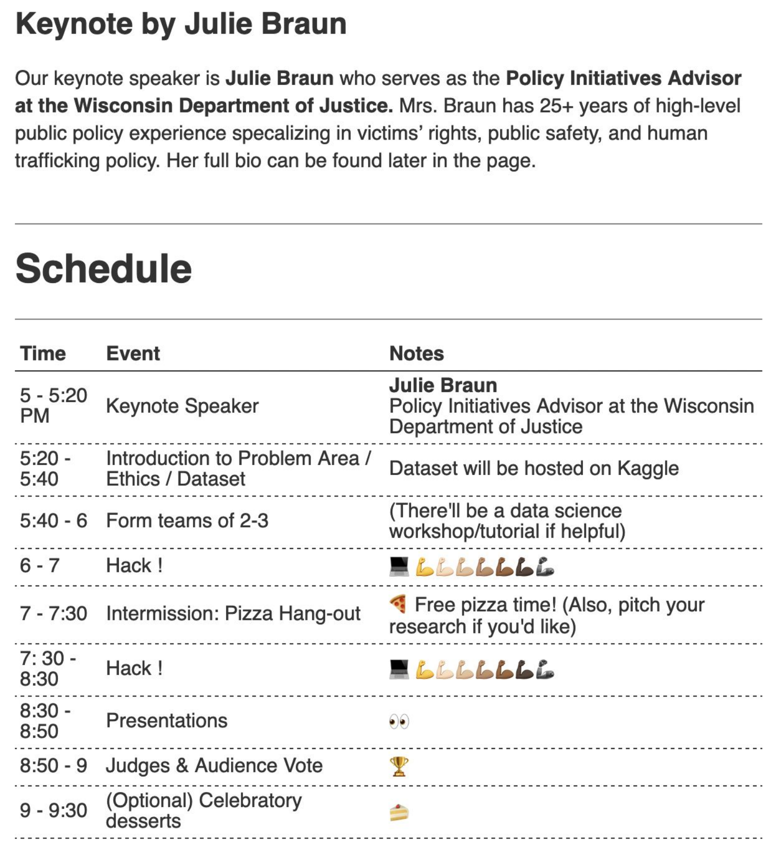
Predict sales prices and practice feature engineering, RFs, and gradi...

Getting Started · 4029 Teams · Ongoing

May 8th, 2023

Hackathon trial run

Keynote Speaker: Julie Braun



Keynote by Julie Braun

Our keynote speaker is **Julie Braun** who serves as the **Policy Initiatives Advisor at the Wisconsin Department of Justice.** Mrs. Braun has 25+ years of high-level public policy experience specalizing in victims' rights, public safety, and human trafficking policy. Her full bio can be found later in the page.

Schedule

Time	Event	Notes
5 - 5:20 PM	Keynote Speaker	**Julie Braun** Policy Initiatives Advisor at the Wisconsin Department of Justice
5:20 - 5:40	Introduction to Problem Area / Ethics / Dataset	Dataset will be hosted on Kaggle
5:40 - 6	Form teams of 2-3	(There'll be a data science workshop/tutorial if helpful)
6 - 7	Hack !	💻 💪💪🏻💪🏼💪🏽💪🏾💪🏿🦾
7 - 7:30	Intermission: Pizza Hang-out	🍕 Free pizza time! (Also, pitch your research if you'd like)
7: 30 - 8:30	Hack !	💻 💪💪🏻💪🏼💪🏽💪🏾💪🏿🦾
8:30 - 8:50	Presentations	👀
8:50 - 9	Judges & Audience Vote	🏆
9 - 9:30	(Optional) Celebratory desserts	🍰

Next HackHT: Oct 21st, 2023 Kickoff

Go to hack4fem.github.io to get email updates

- Sign-ups open at the end of September

Also looking for co-organizers :)

Existing Collaborators

HARVARD UNIVERSITY

Massachusetts Institute of Technology

THENETWORK

Traffik Analysis Hub

IBM

WI Department of Justice
Office of Crime Victim Services

My Research Journey & Conclusion

Part III

Sections

1. My research journey:
   | Scotiabank | Digger Finger |
   | Fiducial Force Sensor |
2. Call-to-Action
3. Acknowledgements

My Research Journey & Conclusion

Part III

Sections

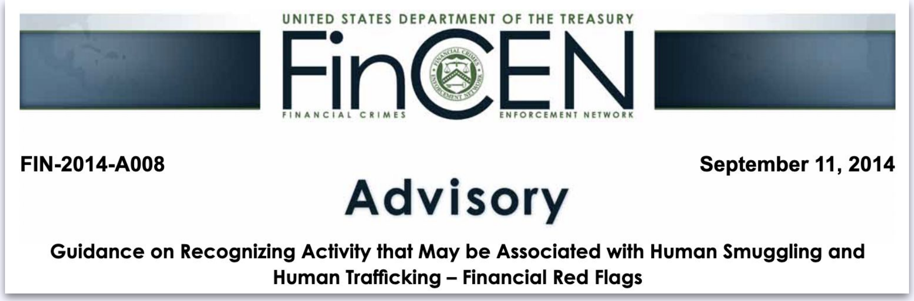
1. My research journey:
   | Scotiabank | Digger Finger |
   | Fiducial Force Sensor |
2. Call-to-Action
3. Acknowledgements

Banks & Synthetic Transaction Data

- Bank Regulation
    - Hundred of millions of dollars in fines
- Use rules-based system
- 98% false positives
- Adopt machine learning
    - Synthetic data: good for class balance (low incidence outliers)
    - Good for continuing education
    - and Hackathons
- Evaluate graph algorithm -- vary label sparsity

Agent-Based Models

Two Agent Types, Normal and Suspicious



UNITED STATES DEPARTMENT OF THE TREASURY

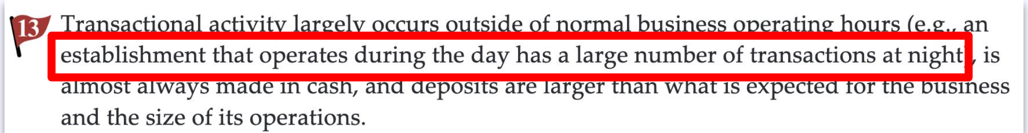
FinCEN

FINANCIAL CRIMES ENFORCEMENT NETWORK

FIN-2014-A008 **September 11, 2014**

Advisory

Guidance on Recognizing Activity that May be Associated with Human Smuggling and Human Trafficking – Financial Red Flags



13 Transactional activity largely occurs outside of normal business operating hours (e.g., an establishment that operates during the day has a large number of transactions at night, is almost always made in cash, and deposits are larger than what is expected for the business and the size of its operations.

- FinCEN: U.S. Financial Crimes Enforcement Network
  https://www.fincen.gov/sites/default/files/advisory/FIN-2014-A008.pdf
- FINTRAC: Financial Transactions and Reports Analysis Centre (Canada)
  https://www.fintrac-canafe.gc.ca/intel/operation/oai-hts-2021-eng

Agents

- Two types: **N**ormal and **S**uspicious

- Vary mean time of day

$$\mu_{hr,N} = 12$$
$$\mu_{hr,S} = 22$$

- Vary homophily:
  Transact more with same type agent

$$P_{S,S} = 0.7$$
$$P_{S,N} = 0.3$$

Agents & Results

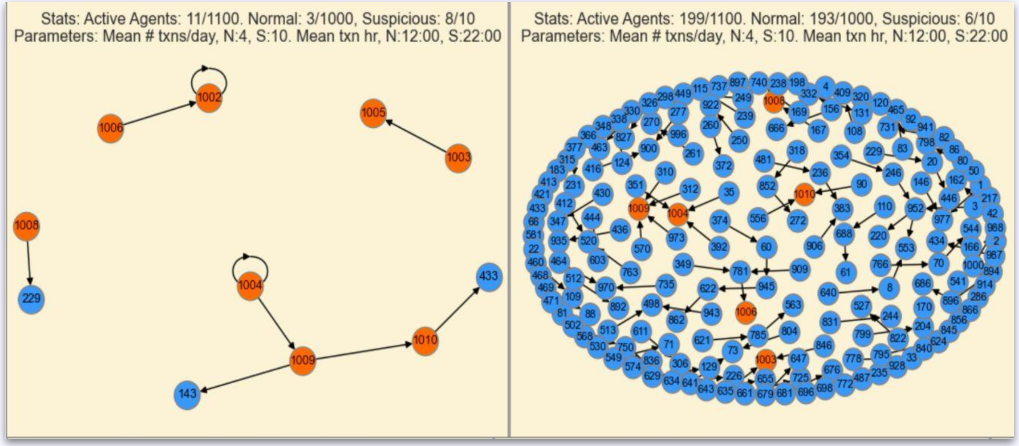
- Two types: **N**ormal and **S**uspicious
- Vary mean time of day

$$\mu_{hr,N} = 12$$
$$\mu_{hr,S} = 22$$

- Vary homophily:
  Transact more with same type agent

$$P_{S,S} = 0.7$$
$$P_{S,N} = 0.3$$

Midnight **8 AM**



Stats: Active Agents: 11/1100. Normal: 3/1000, Suspicious: 8/10
Parameters: Mean # txns/day, N:4, S:10. Mean txn hr, N:12:00, S:22:00

Stats: Active Agents: 199/1100. Normal: 193/1000, Suspicious: 6/10
Parameters: Mean # txns/day, N:4, S:10. Mean txn hr, N:12:00, S:22:00

github.com/nro-bot/fake-banking-data

Digger Finger: GelSight Tactile Sensor for Object Identification Inside Granular Media

Radhen Patel, Branden Romero, Rui Ouyang, Edward Adelson

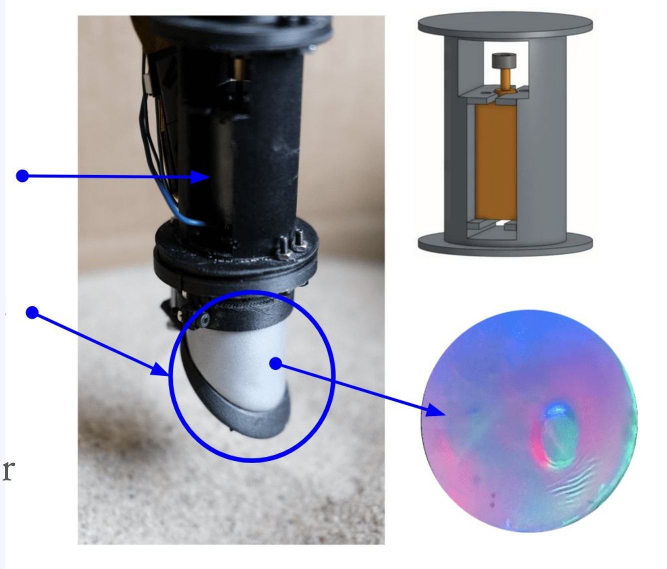
17th International Symposium on Experimental Robotics (ISER) 2020

Digger Finger: GelSight Tactile Sensor for Object Identification Inside Granular Media

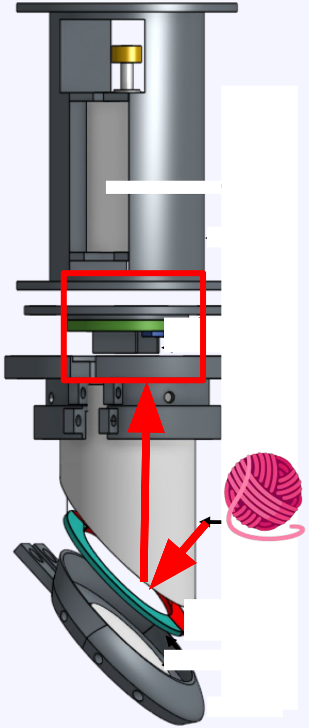
Radhen Patel, Branden Romero, Rui Ouyang, Edward Adelson

17th International Symposium on Experimental Robotics (ISER) 2020

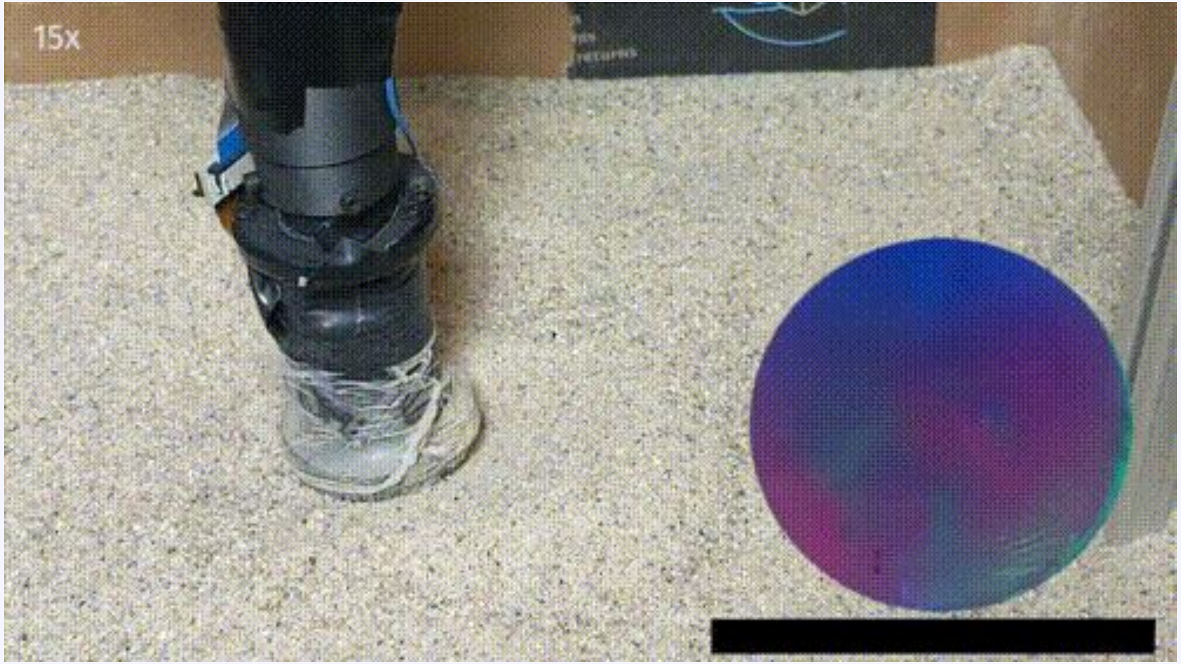
- Small
- Vibrator Motor
- Wedge-shaped
- Integrated sensor

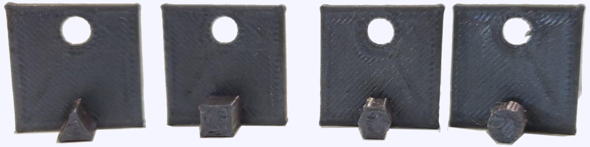


Digger Finger: GelSight Tactile Sensor for Object Identification Inside Granular Media



Digger Finger: GelSight Tactile Sensor for Object Identification Inside Granular Media

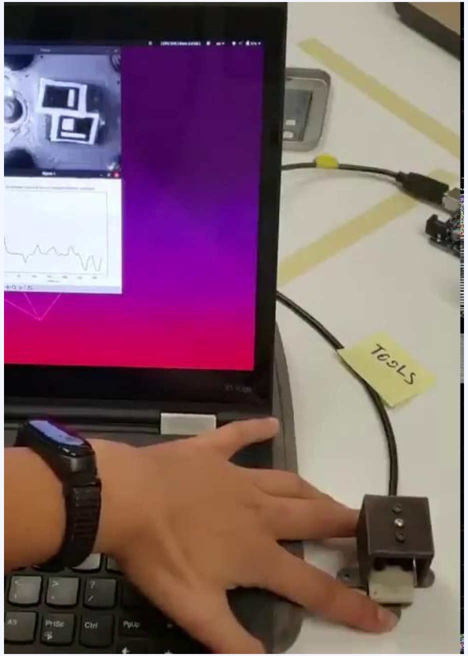




Low-Cost Fiducial-Based 6-Axis Force-Torque Sensor

Rui Ouyang, Robert Howe

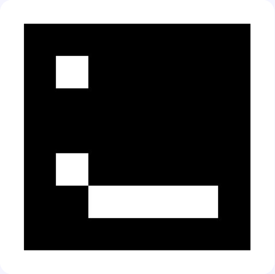
IEEE International Conference on Robotics and Automation (ICRA) 2020



Low-Cost Fiducial-Based 6-Axis Force-Torque Sensor

or "Fiducial Force Sensor" for short

Fiducials

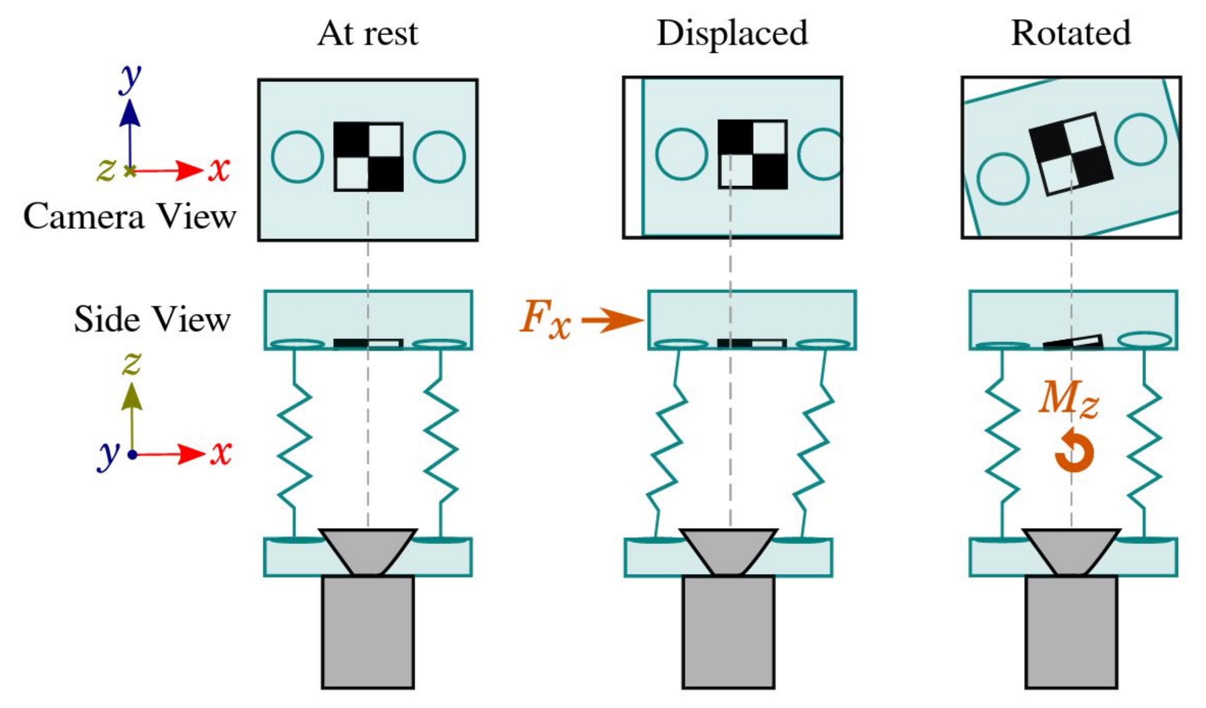


Vector

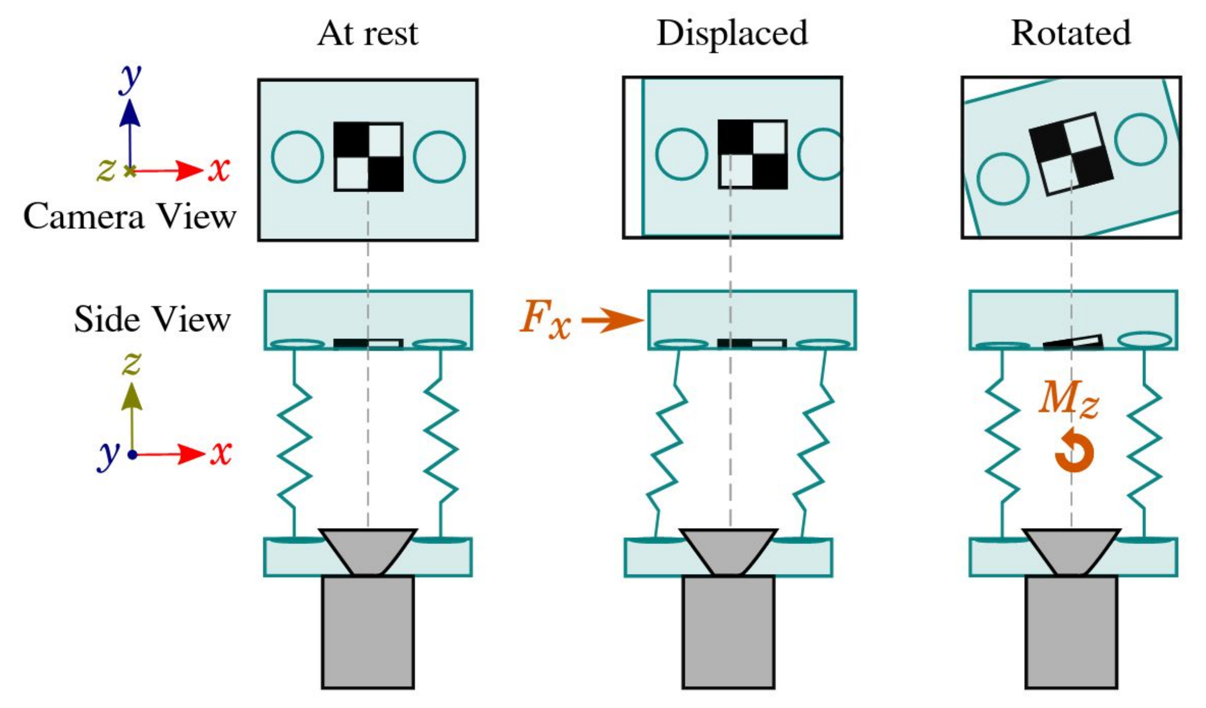
Printed

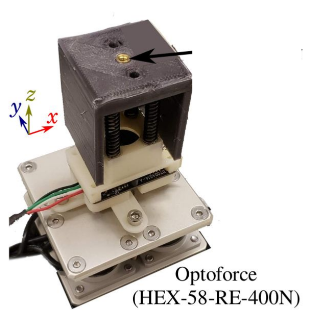
Estimate 6D Pose

Low-Cost Fiducial-Based 6-Axis Force-Torque Sensor



Low-Cost Fiducial-Based 6-Axis Force-Torque Sensor





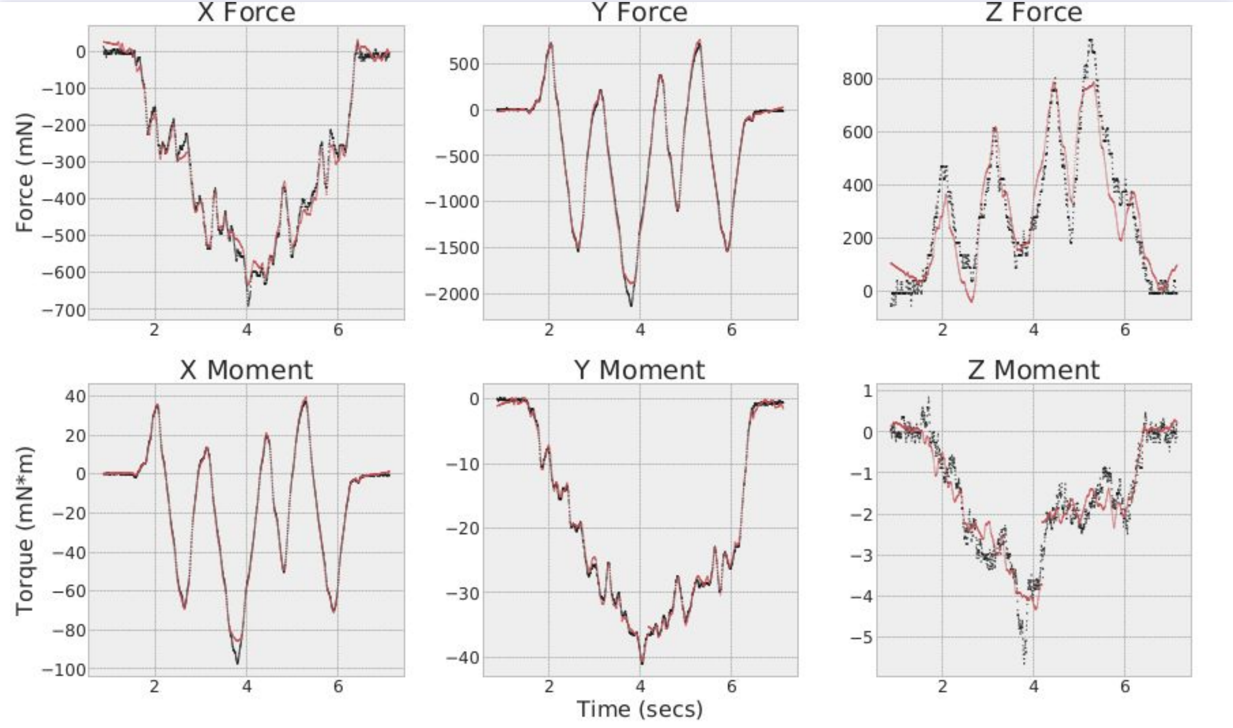
Optoforce
(HEX-58-RE-400N)

$$\begin{bmatrix} F_x \\ F_y \\ F_z \\ M_x \\ M_y \\ M_z \end{bmatrix} = \begin{bmatrix} & & \\ & K_{6\times 6} & \\ & & \end{bmatrix} \begin{bmatrix} D_x \\ D_y \\ D_z \\ D_\theta \\ D_\phi \\ D_\gamma \end{bmatrix} + \begin{bmatrix} B \end{bmatrix}$$

Low-Cost Fiducial-Based 6-Axis Force-Torque Sensor

Black = Ground truth

Red = Fiducial sensor



100x Cheaper: $35 vs. $3,500

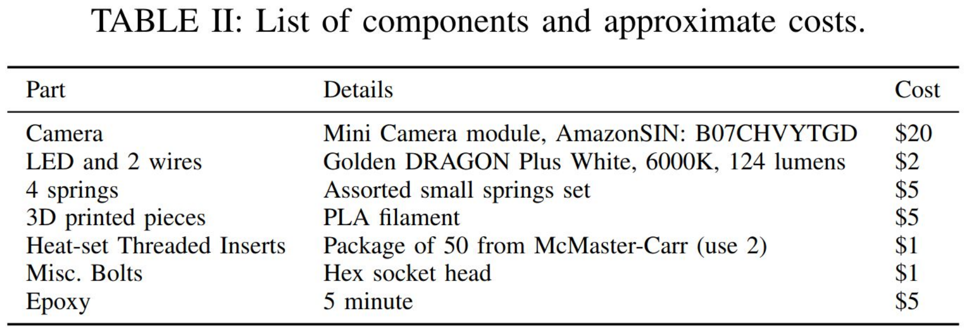
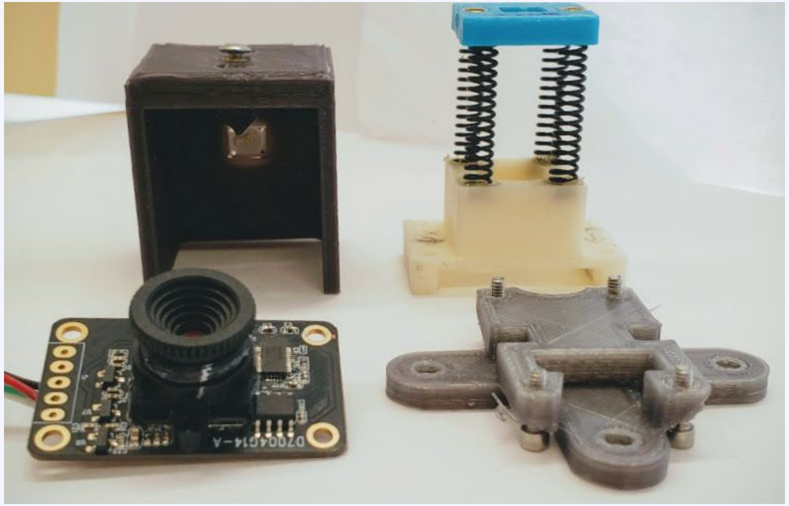


TABLE II: List of components and approximate costs.

Part	Details	Cost
Camera	Mini Camera module, AmazonSIN: B07CHVYTGD	$20
LED and 2 wires	Golden DRAGON Plus White, 6000K, 124 lumens	$2
4 springs	Assorted small springs set	$5
3D printed pieces	PLA filament	$5
Heat-set Threaded Inserts	Package of 50 from McMaster-Carr (use 2)	$1
Misc. Bolts	Hex socket head	$1
Epoxy	5 minute	$5

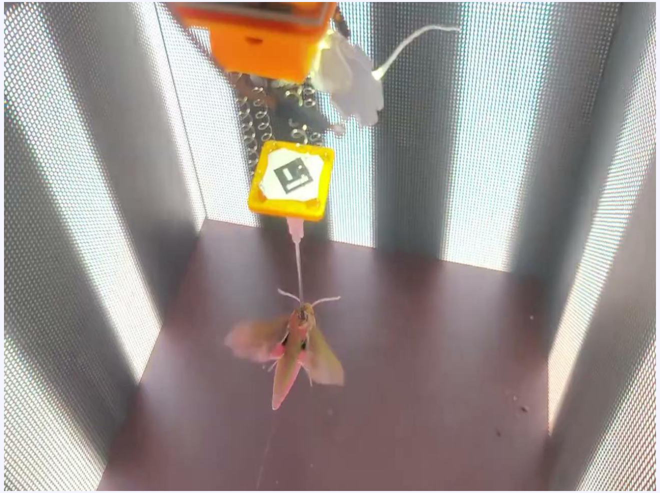


Open-Source Hardware & Software

- Released design files: sites.google.com/view/fiducialforcesensor

Open-Source Hardware & Software

- Released design files: sites.google.com/view/fiducialforcesensor
- Dr. Pavan Kaushik - postdoc at Max Planck Institute of Animal Behaviour - Locust Swarming



My Research Journey & Conclusion

Part III

Sections

1. My research journey:
   | Scotiabank | Digger Finger |
   | Fiducial Force Sensor |
2. Call-to-Action
3. Acknowledgements

Call-to-Action: Research

- **Question**
  To what extend do these forums promote and normalize misogyny and contribute to real-life harm?
- **Motivation**
  Atlanta shootings in 2020
- **Sub-questions**
  Toxicity metric?
  Toxicity spread?
  Link to real life?

- **Question**
  How can we automatically extract user insights?
- **Motivation**
  Shift attention from providers to buyers
  Estimate impact of policies
- **Sub-questions**
  What information is present?
  Disambiguating professions?
  Subject of text (first- or third-person)?
  Skew of data (vs. general population)?

Call-to-Action: Public Awareness

- Join the hackathon!
  Oct. 21st
  hack4fem.github.io

- Explore the datasets!
  github.com/nro-bot/imi_forums

- I'll post more details at
  nrobot.dev

- Academic/Institutional Support
  Year-long Fellowships
  (similar to Work of the Future)

More generally --

sites.google.com/view/nlp4positiveimpact

Summary

- The Google Places dataset: illegal activity in plain sight
- The Forum dataset: two case studies and a hackathon
- My research journey:
  | Scotiabank | Digger Finger | Fiducial Force Sensor |]

Future

- nrobot.dev / nouyang@alum.mit.edu
- Jobs: Industry, research - doesn't have to be AI4SG
  (Effective altruism)

Research is collaborative!

Open datasets, open source hardware, open source software, tools for collaboration

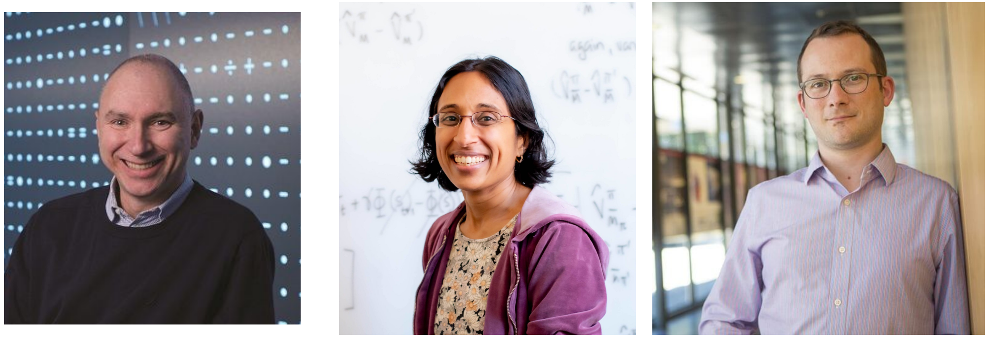
Acknowledgments

My committee!

Professor Roberto Rigobon

Professor Finale Doshi-Velez

Professor David Parkes



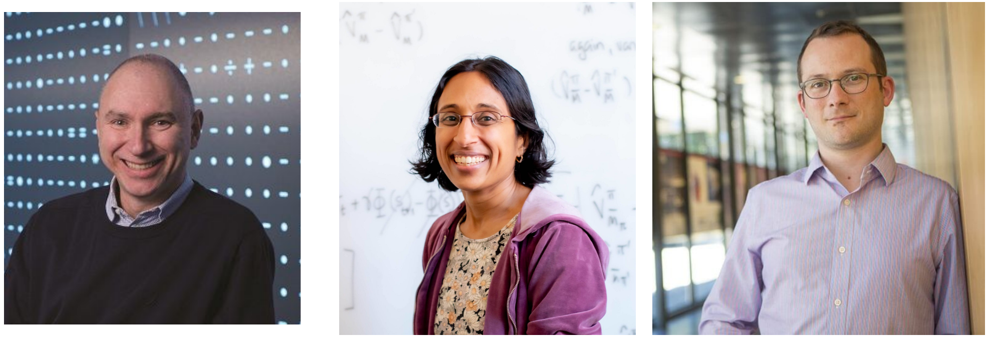
Acknowledgments

My committee!

Professor Roberto Rigobon

Professor Finale Doshi-Velez

Professor David Parkes

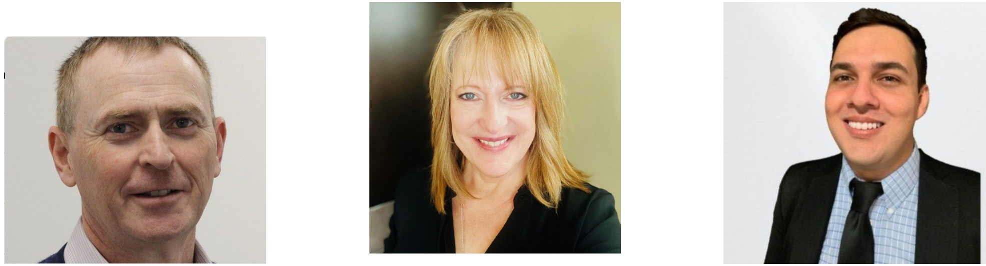


Collaborators

John McGrath, IBM

Julie Braun, WI DoJ

Carlos Garcia, The Network



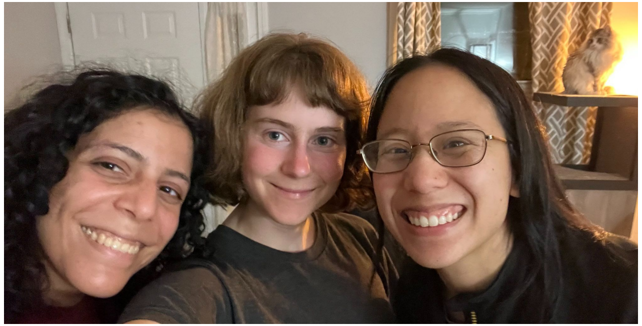
Loyal Friends!

Friends
Marcela Rodriguez, Irina Tolkova
Judy Baek

Roommates
Erons Ohienmhen, Ondřej Bíža
Arianna McQuillen, Gagan Khandate

Partner
Diony Rosa

Friends -- Ilia Lebedev, Sarah Cheng, Lily Zdansky, Cathy Wu, Anvita Pandit



Support Systems

Psychiatrist - Blake Ritter

Dane Kouttron
Julian Merrick
Amber Houghstow
John Aleman
Nick Kirkby

Daniel Gonzalez
Albert Wang
Joao L. A. S. Ramos, Michael
Laura Shumaker, Robin Deits, Amy Qian, Ethan
Rahn, Annie Labine, Ava Chen, Ben Katz, Juliann Ma

Letter Writers

Alexander Wait Zaranek (Curoverse)
Sangbae Kim (Biomimetics)
James Bales (Strobe Lab)
Lucas Janson (Statistics)
Daniel Frey (2.007)
Sanjay Sarma (EdX)
Isaac Chuang (NarwhalEdu)
Ken VanArsdel (Fitbit)

Final Push

- Anna Kreuder
- Erons Ohienmhen
- John Aleman
- Eric Marion
- Ben Hayes
- Mark Goldstein
- Eric Lu
- Cheryl C & Tammi Chen

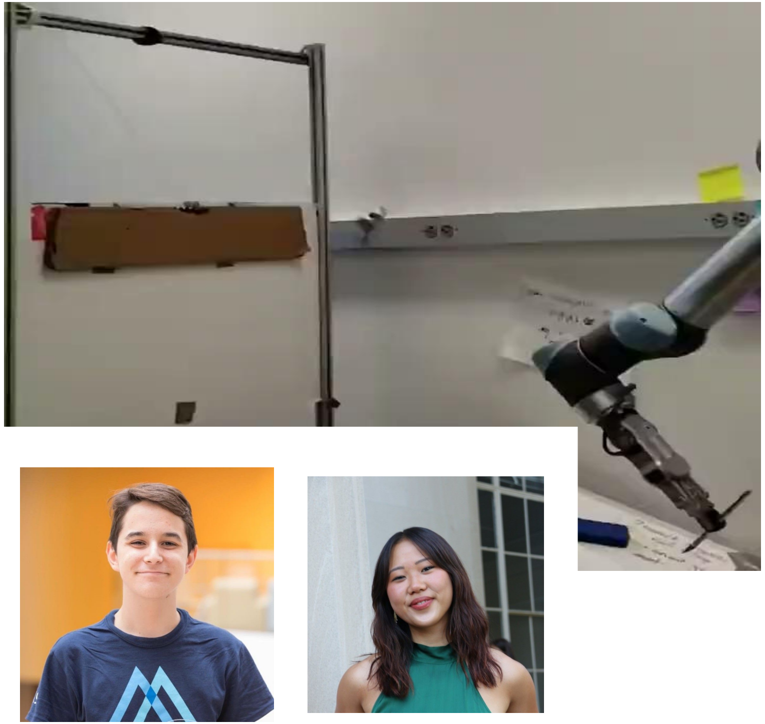
More people!

Admins -- Huge thanks to Dr. John Girash!

UROPs
Santi Cantu
Erin Zhang

Interns
Julian Phillips Kennedy
Danny O'Connor

Janitorial & Security Staff



Labs I went through

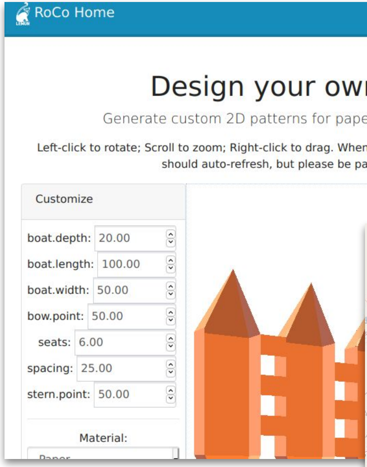
Howe Lab - Prof. Robert Howe
Buse Aktas, Alperen, Yash, Qian
Ted Sirota, James Weaver
Biomimetics Lab - Albert, Joao
Helping Hands Lab - Robert Platt
Adelson Lab - Radhen Patel, Branden Romero, Shaoxiong Wang, Sandra Liu, Felipe Veiga, Edward Adelson, Greg Izatt
MD309 - Anitha Gollamundi, Aaron B.
MD209 - Jialiang, Dor, Mia, Kat
MD121 - David, Eric, Mark, Crystal, Mark York
Prof. Suzanne Smith

More people!

Carrie Chai, Ming, and Elsa Riachi @ Scotiabank

Ankur Mehta

Zoz Brooks



Design your own Trimaran

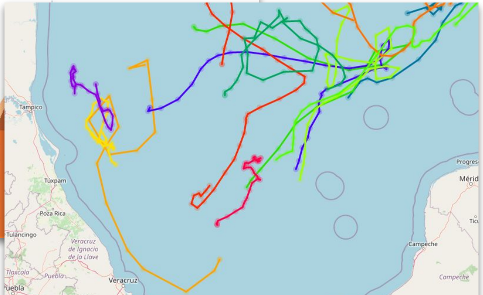
Generate custom 2D patterns for papercraft, lasercutter, or wood.

Left-click to rotate; Scroll to zoom; Right-click to drag. When loading page or changing parameters, desigr
should auto-refresh, but please be patient (wait 1-3 seconds)

Customize

boat.depth: 20.00
boat.length: 100.00
boat.width: 50.00
bow.point: 50.00
seats: 6.00
spacing: 25.00
stern.point: 50.00

Material:



My Parents !

The End

Thank you!

Extra Slides

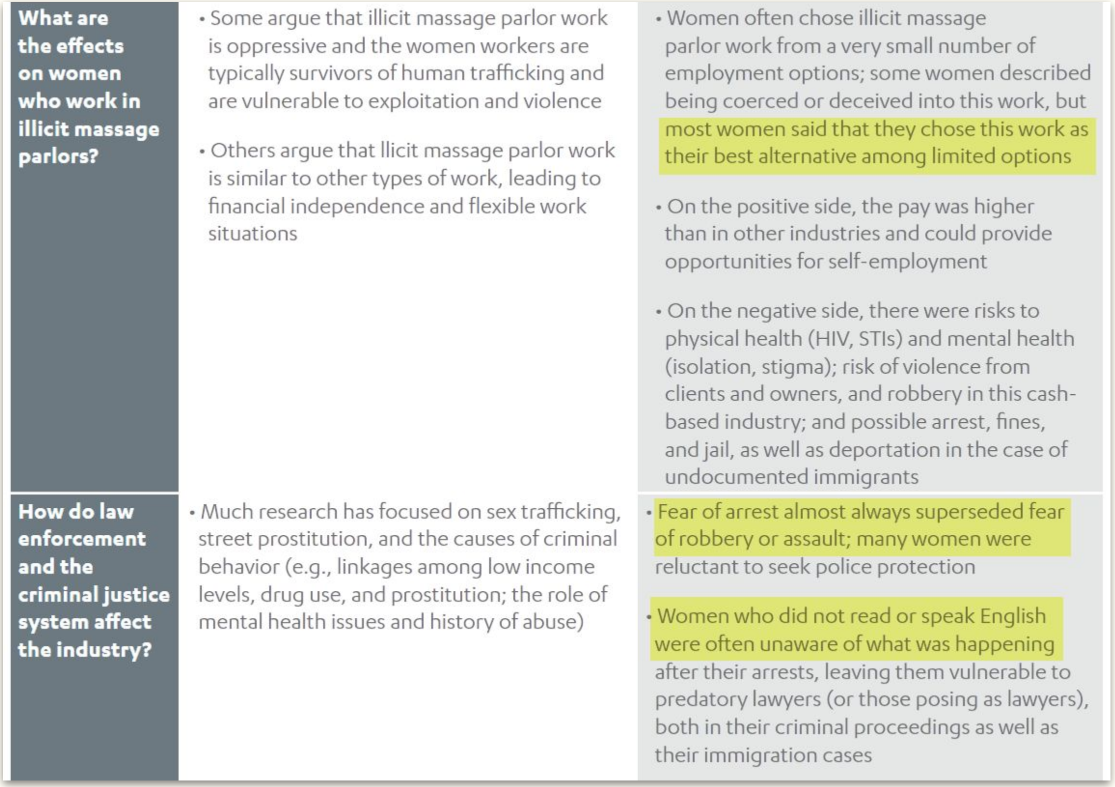
aaari.info/20-12-11chin

Prof.'s Chin, Takahashi

Ethics

- (Almost) all sex work is illegal in US
- Limited by resources to only egregious cases of trafficking
  - However the "quick win" can be to just shut a place down

- Points of View
  - Consult Hacking\\Hustling hackinghustling.org



What are the effects on women who work in illicit massage parlors?	• Some argue that illicit massage parlor work is oppressive and the women workers are typically survivors of human trafficking and are vulnerable to exploitation and violence • Others argue that llicit massage parlor work is similar to other types of work, leading to financial independence and flexible work situations	• Women often chose illicit massage parlor work from a very small number of employment options; some women described being coerced or deceived into this work, but most women said that they chose this work as their best alternative among limited options • On the positive side, the pay was higher than in other industries and could provide opportunities for self-employment • On the negative side, there were risks to physical health (HIV, STIs) and mental health (isolation, stigma); risk of violence from clients and owners, and robbery in this cash-based industry; and possible arrest, fines, and jail, as well as deportation in the case of undocumented immigrants
How do law enforcement and the criminal justice system affect the industry?	• Much research has focused on sex trafficking, street prostitution, and the causes of criminal behavior (e.g., linkages among low income levels, drug use, and prostitution; the role of mental health issues and history of abuse)	• Fear of arrest almost always superseded fear of robbery or assault; many women were reluctant to seek police protection • Women who did not read or speak English were often unaware of what was happening after their arrests, leaving them vulnerable to predatory lawyers (or those posing as lawyers), both in their criminal proceedings as well as their immigration cases

The Network Video (Subject Matter Experts)

The Network Video (Personal Story)

Ethics

Butterfly

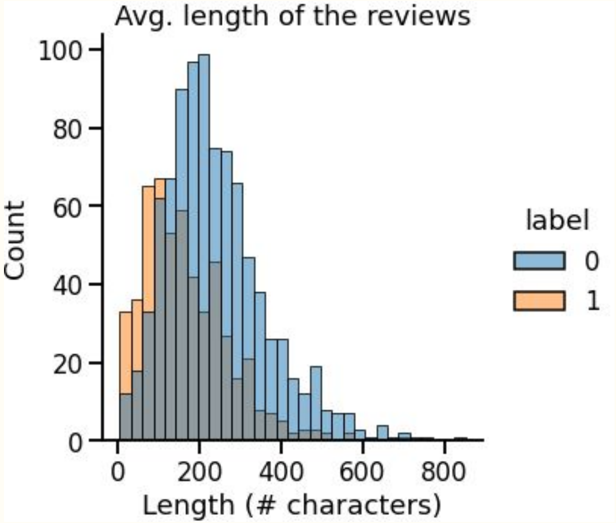
- https://www.butterflysw.org/publication
- Chin & Takahashi; https://aaari.info/20-12-11chin

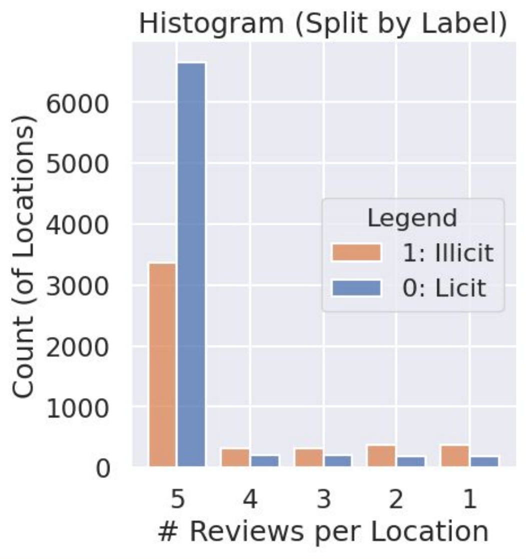
The Network Team

- https://www.thenetworkteam.org/research
- https://www.mass.gov/files/documents/2018/04/30/Polaris%20HT%20IMB%20Report.pdf

“Most victims of illicit massage businesses are women from the mid-thirties to late fifties from China and South Korea”

Happy to discuss more :)

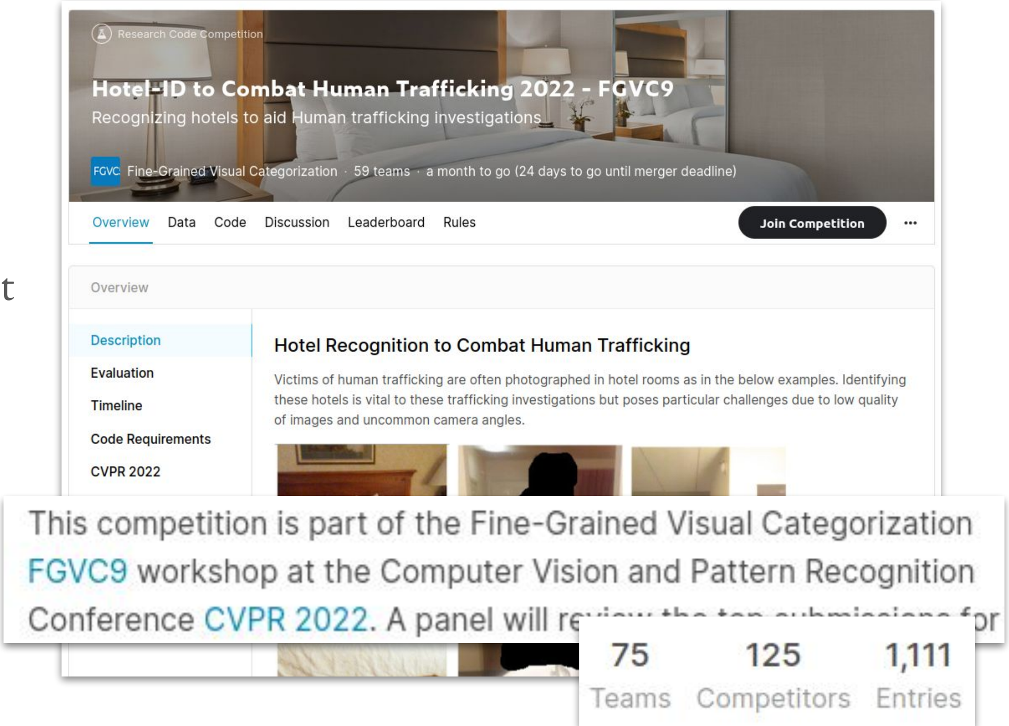

Avg. length of the reviews
Count
100
80
60
40
20
0
0
200
400
600
800
Length (# characters)
label
0
1


Histogram (Split by Label)
Count (of Locations)
6000
5000
4000
3000
2000
1000
0
5
4
3
2
1
# Reviews per Location
Legend
1: Illicit
0: Licit

Meta Thoughts ala Traffik Analysis Hub

How to engage computer science community?

- Datasets (incl. labor trafficking)
- Competitions (Hotels50k kaggle)
- Funding, manpower, institutional support
  - internships, capstones, fellowships, postdoc positions
- Concrete use cases



Hotel-ID to Combat Human Trafficking 2022 - FGVC9

Recognizing hotels to aid Human trafficking investigations

FGVC Fine-Grained Visual Categorization · 59 teams · a month to go (24 days to go until merger deadline)

Hotel Recognition to Combat Human Trafficking

Victims of human trafficking are often photographed in hotel rooms as in the below examples. Identifying these hotels is vital to these trafficking investigations but poses particular challenges due to low quality of images and uncommon camera angles.

This competition is part of the Fine-Grained Visual Categorization FGVC9 workshop at the Computer Vision and Pattern Recognition Conference CVPR 2022. A panel will re

75	125	1,111
Teams	Competitors	Entries

PhD Thesis Defense

🖐 Machine Learning for Tangible Effects: Natural Language Processing for Uncovering the Illicit Massage Industry

Rui Ouyang
nrobot@mit.edu | nrobot.dev

Sept. 5, 2023

📜 Thesis
arxiv.org/abs/2309.03470

Future

If you know folks working in industry in the AI/NLP or robotics spaces in industry, I'm looking for a job!

- It's a bit old but [I have a resume, linked here]
- Preferably: A nice, 500+ person company
  - I prefer a 40 hr work-week
- Prefer: in Boston (in-person at least part of the time)
  - Would consider other places (especially if remote)
- Prefer: hard floor at 175 k
  - Would consider factors like vacation / mission / opportunities to advance / supportive hiring manager etc.

Also if you want to co-work on job applications, let me know :)
[Edited 15 Sep 2023]