

---

# Left-over Men and Women in Post-One-Child-Policy China

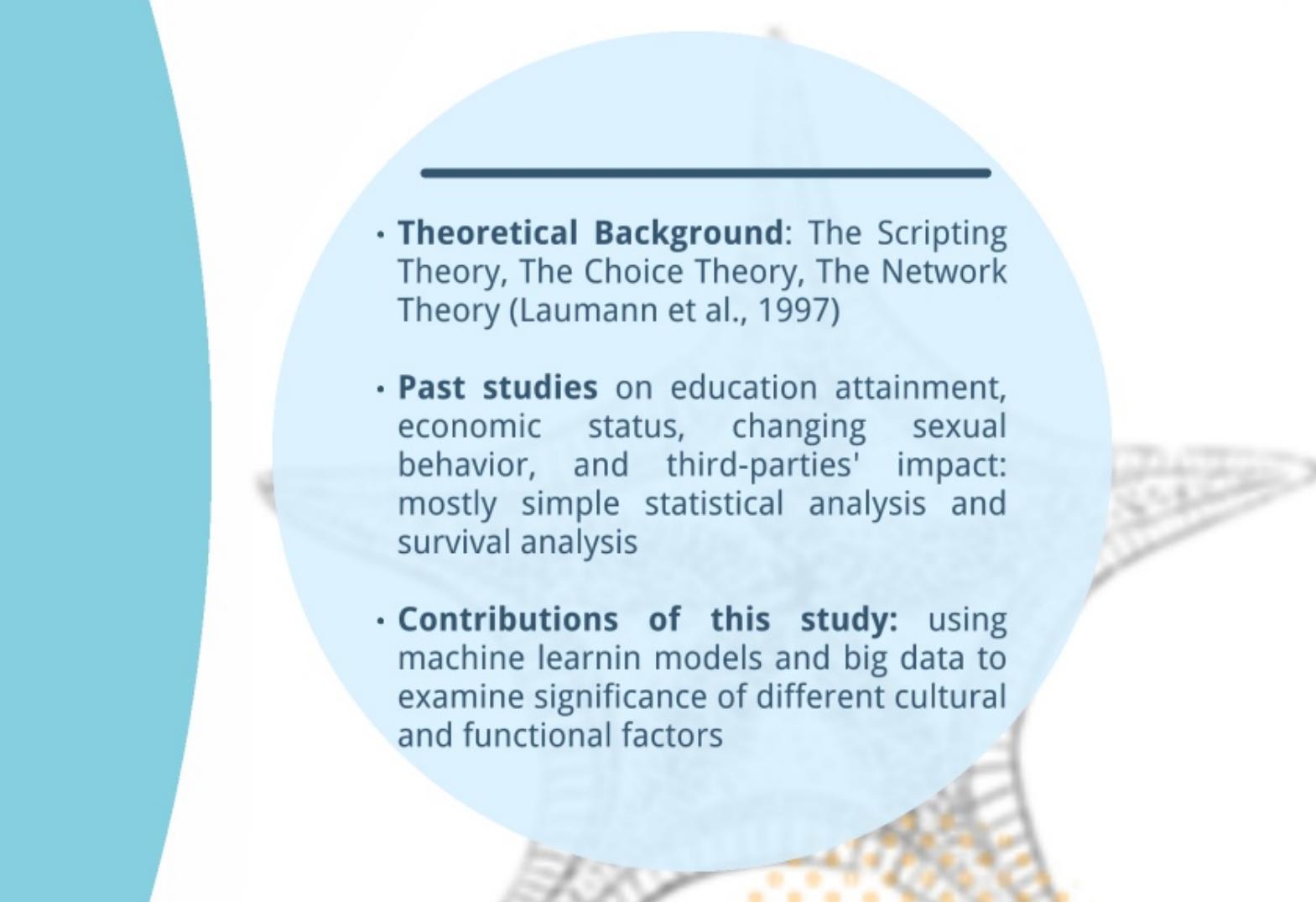


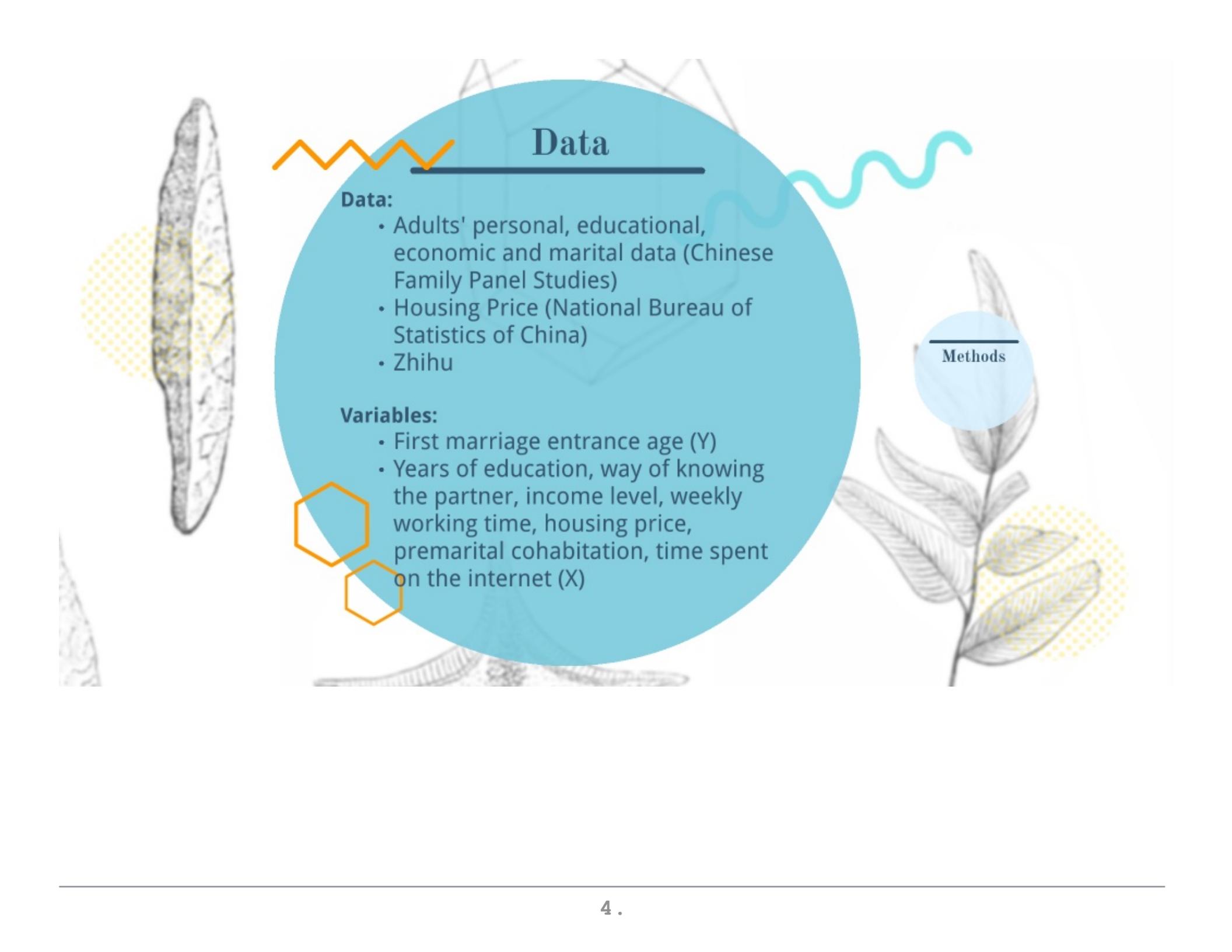
# Introduction

This study will explore functional and cultural factors that cause Chinese people to postpone their age of first marriage after the change in birth control policy.



Literatures & Contributions

- 
- 
- **Theoretical Background:** The Scripting Theory, The Choice Theory, The Network Theory (Laumann et al., 1997)
  - **Past studies** on education attainment, economic status, changing sexual behavior, and third-parties' impact: mostly simple statistical analysis and survival analysis
  - **Contributions of this study:** using machine learning models and big data to examine significance of different cultural and functional factors



## Data

### Data:

- Adults' personal, educational, economic and marital data (Chinese Family Panel Studies)
- Housing Price (National Bureau of Statistics of China)
- Zhihu

### Variables:

- First marriage entrance age (Y)
- Years of education, way of knowing the partner, income level, weekly working time, housing price, premarital cohabitation, time spent on the internet (X)



## Methods

## Methods

---

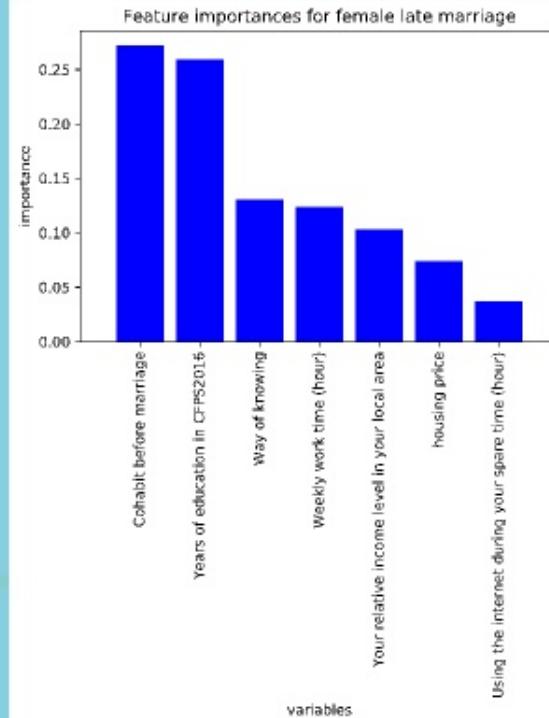
### Machine Learning:

- Feature importance
- Regressor vs Classifier based on prediction power
- Random Forest vs SVM based on prediction power

### Text Analysis:

- Crawling top answers to 3 Zhihu questions
- Jieba for extracting Chinese words
- Word Cloud for High-frequency words

## Feature Importance



## Model Comparison

## Random Forest vs SVC

Table 1: Performance comparison between Random Forest Classifier and SVC

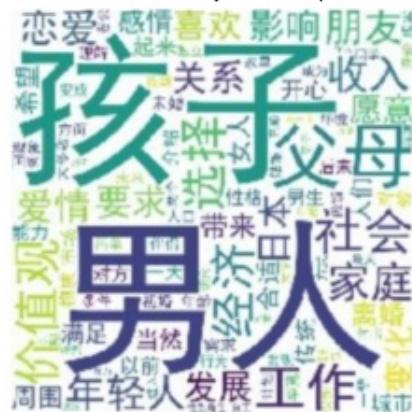
	Female Late Marriage		Male Late Marriage	
	Random Forest Classifier	SVC	Random Forest Classifier	SVC
Accuracy	0.6669	0.4972	0.6616	0.5872
Informedness	0.3798	-0.0055	0.3231	0.1744
Sensitivity	0.4138	0.0344	0.5517	0.5172
Specificity	0.92	0.96	0.7714	0.6571

- Classifier > Regressor
- Random Forest > SVM

## Zhihu Questions

1. "Are there a great many young people who no longer want to get married? Why?"
2. "What are left-over women insisting on?"
3. "What are left- over men insisting on?"

Patterns of answers to Question 1, min 100 likes



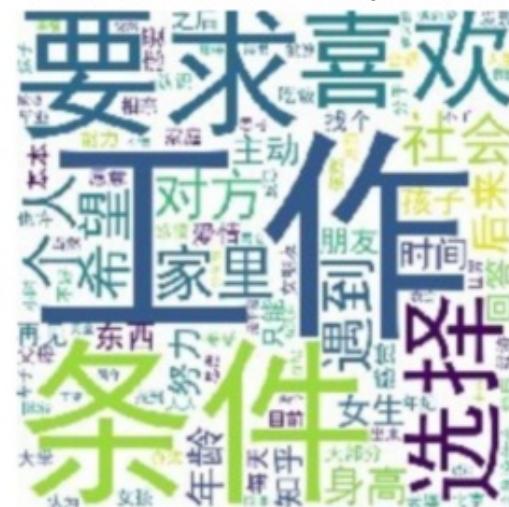
Gender-Specific

# Women vs Men

Patterns of answers to Question2



Patterns of answers to Question3



# Conclusion

- Better accuracy if treated as a classification problem; Random Forest performs better than SVM
- Pre-marital Cohabitation is an important predictor
- Child-rearing is a big concern
- Gender differences in cultural scripting



Thank you!

Nice work!

Who is up next?

# Fiscal Decentralization and Economic Development in China - Before and After the Tax-Sharing Reform

Faculty Advisor: Luis Martinez

Fangfang Wan

Masters in Computational Social Science, University of Chicago

# Research Question

- ❑ In the last 40 years, especially after the tax-sharing reform in 1994, what is the role of fiscal decentralization in economic development in China?
- ❑ Lin and Liu (2000) conducted research on the role of fiscal decentralization on economic development in China using data in 1980s and 1990s (before 1994)

# Background

- ❑ Before 1978: All collected by central government and then redistributed to local governments
- ❑ 1978 – 1994: Local governments retained a fixed portion of local revenue
- ❑ After 1994: Tax sharing system – tax distribution between central and local governments by tax categories

# Key Literatures

- ❑ Lin and Liu (2000)
- ❑ De Valk (1990) viewed fiscal decentralization as a cause of better economic development
  - ❑ fiscal decentralization increases effectiveness and efficiency of economic development
  - ❑ local governments have more and better information about local needs, so they can distribute money more efficiently.
- ❑ Bahl and Linn (1992) held the point of view that fiscal decentralization is resulted from economic development.

# Data

- National Oceanic and Atmospheric Administration (NOAA) (2014)
  - Nightlight data as proxy for economic growth
- National Bureau of Statistics of China
  - Other variables

# Model

Table 1: Variable abbreviations and definitions

<b>Variable</b>	<b>Definition</b>
<b>GGDP</b>	Growth rate of real per capita GDP (%)
<b>NL</b>	Night Light Proxy
<b>FD</b>	Fiscal Decentralization: calculated in 2 ways: local government revenue/central government revenue, and local government expenditure/central government expenditure
<b>NL</b>	Night Light Proxy
<b>POPSHR</b>	Rural population (%)
<b>TPOP</b>	Total population (in thousands)
<b>FPMP</b>	Relative price of farm products to nonfarm product: the ratio of state's real procurement price index for farm products to real price index of manufacture goods in rural area
<b>NSOESH</b>	Share of Non-SOEs' output in the total industrial output (%)
<b>GI</b>	Growth rate of per capita fixed asset investment (in real term) (%) (Lin and Liu, 2000)

$$\begin{aligned}
 \text{NL}_{it} \text{ or } \text{GGDP}_{it} = & \beta_0 + \beta_1 \text{FD}_{it} + \beta_2 \text{NSOESH}_{it} + \beta_3 \text{GI}_{it} \\
 & + \beta_4 (\text{FISCAP})_{it} + \beta_5 \text{FPMP}_{it} + \beta_6 \text{POPSHR}_{it} + \mu_i + \lambda_t \\
 \text{where } i \text{ is province, } t \text{ is time, and } i = 1, \dots, N; t = 1, \dots, T.
 \end{aligned}$$

# Nightlight & Economic Activities



# Random-effect Models

- ❑ LGR/CGR and LGE/CGE have contrasting effects on economic growth.
- ❑ Revenue ratios exhibit positive effects, while expenditure ratios show negative effects.
- ❑ The growth rate of real per capita fixed assets investments have strong positive correlation with GGDP, but its correlation with nightlight is not significant.
- ❑ Higher rural population ratio results in significantly lower nightlight measurements, and areas with more total population have brighter nightlights.

LGR: Local Government Revenue

CGR: Central Government Revenue

LGE: Local Government Expenditure

CGE: Central Government Expenditure

# Fixed-effect models

## ❑ Motivation

- ❑ Remove portion of the effect from time varying controls that is not related to the dependent variable – for example, GI may just be increasing but not necessarily be correlated with economic growth/nightlight proxy
- ❑ Therefore, provincial and time fixed-effect dummy variables were included
- ❑ Conclusion: almost the same trend as in Random effect models

# Conclusion

- ❑ No obvious relationship between fiscal decentralization and economic growth
- ❑ Disentanglement between revenue and public service responsibilities for local governments

Nice work!

Who is up next?



# Goal Attainment in C4P

---

Jingwen (Fiona) Fan

Advisor: Alex Tate



# CCP & C4P

---

- Trade-off between specialization (Hospitalist) and continuity of care
- Comprehensive Care Program (CCP), Comprehensive Care, Community and Culture Program (C4P)
  - Recruits only high risk patients identified by machine learning algorithms
  - Have the same primary care physician care for both inpatient hospitalization and in outpatient clinics
  - C4P has additional community programs targeted at unmet needs

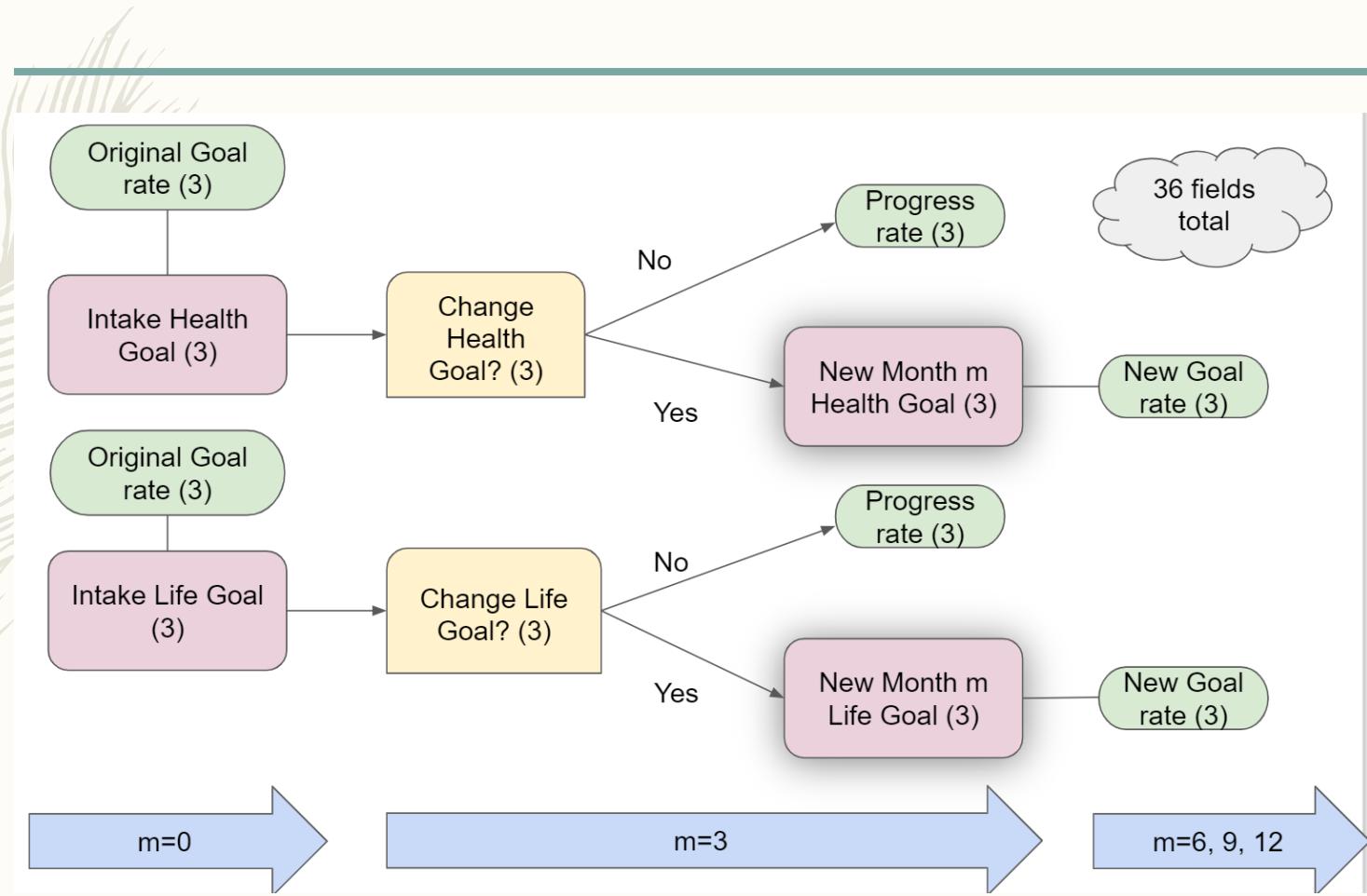


# CCP & C4P

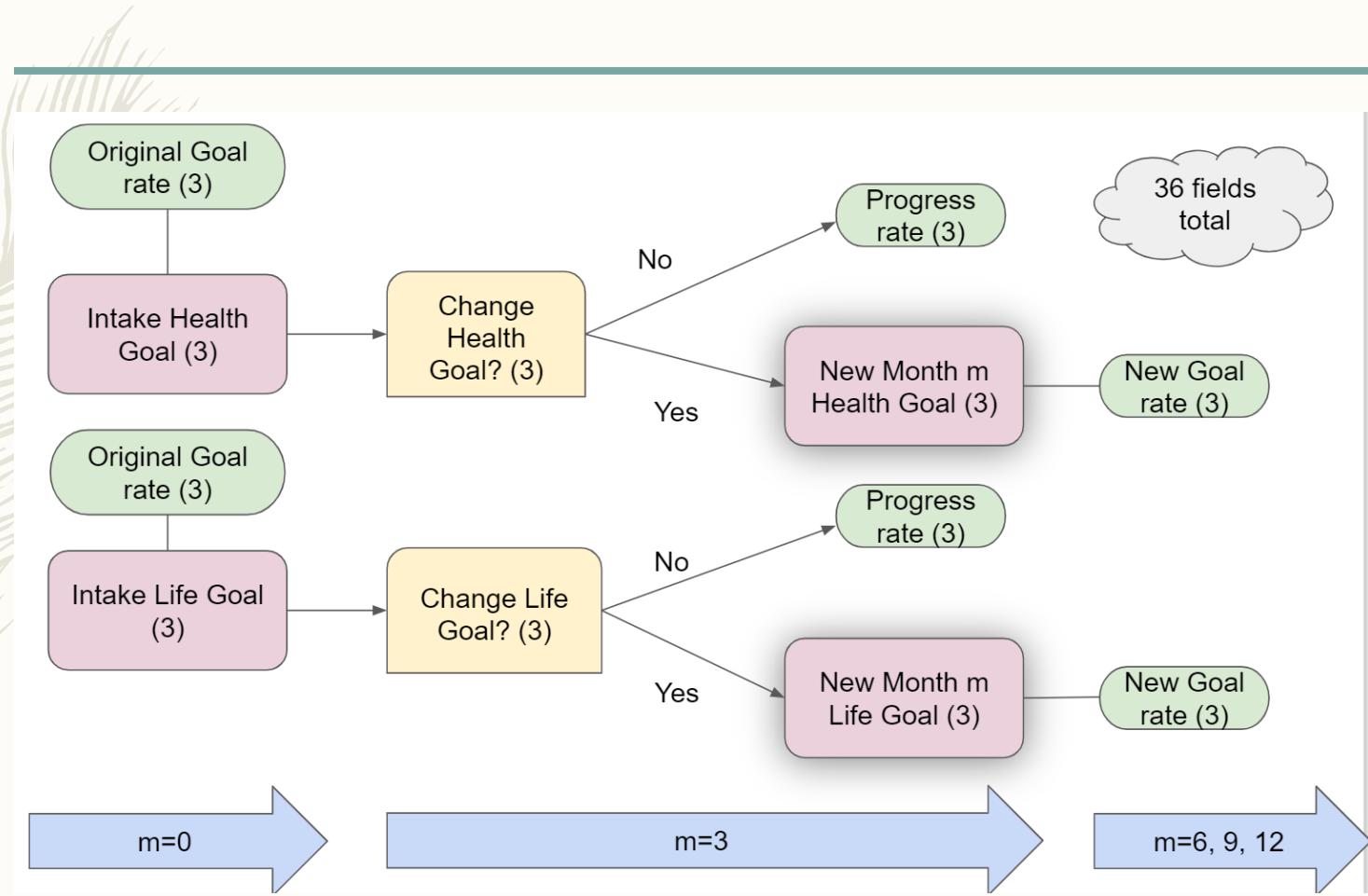
---

- Trade-off between specialization (Hospitalist) and continuity of care
- Comprehensive Care Program (CCP), Comprehensive Care, Community and Culture Program (C4P)
  - Recruits only high risk patients identified by machine learning algorithms
  - Have the same primary care physician care for both inpatient hospitalization and in outpatient clinics
  - C4P has additional community programs targeted at unmet needs
- Question: What goals do patients participating in C4P have? How do the categories of goals affect their outcomes?

# Goal Attainment



# Goal Attainment





# Mixed Effect Pattern Mixture Model

---

Drop:

CCCCCCCC  
CCCCCCC D  
CCCCCCC R  
CCCCCC D D  
CCCCCC O O  
CCCCCC RC  
CCCCCR RD  
CCCCCR RR

- $$Y_{i,j} = \beta_0 + \beta_1 C4P_i + \beta_2 \sqrt{round_j} + \beta_3 (C4P_i * \sqrt{round_j}) + \beta_0^D Drop_i + \beta_1^D (Drop_i C4P_i) + \beta_2^D (Drop_i * \sqrt{round_j}) + \beta_3^D (Drop_i * C4P_i * \sqrt{round_j}) + v_{0i} + v_{1i} \sqrt{round_j} + \epsilon_{ij}$$
  - i: subjects, j: observation
- $\beta_0 \sim \beta_3$  are for completers,  $\beta_0^D \sim \beta_3^D$  are how dropouts differ from completers
- $\beta_3$  is the variable of interest.



# Mixed Effect Pattern Mixture Model

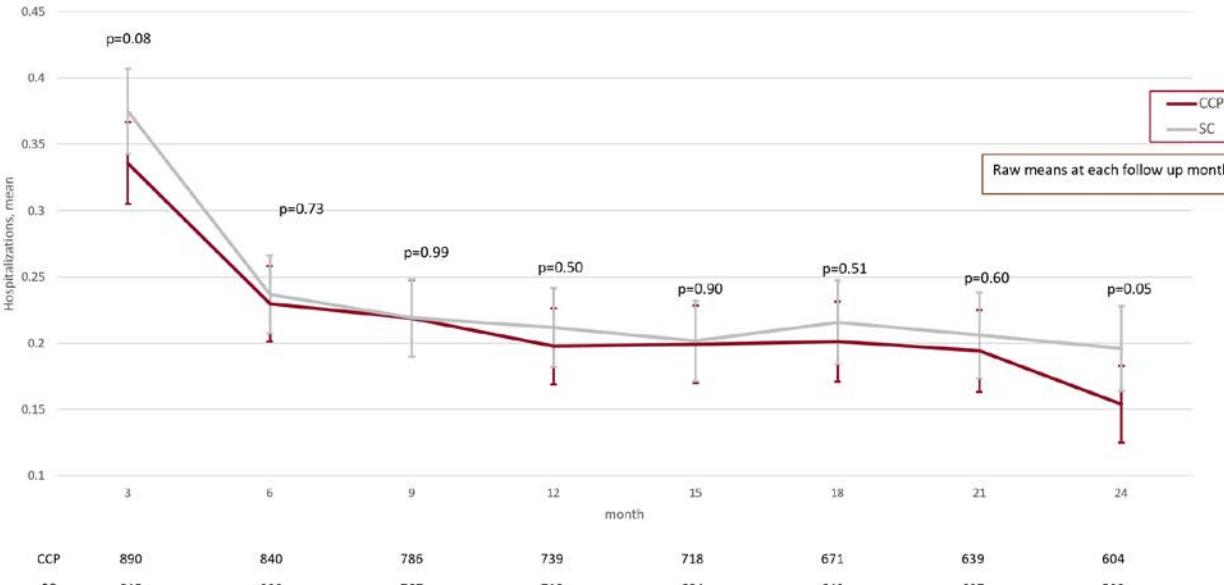
---

Drop:

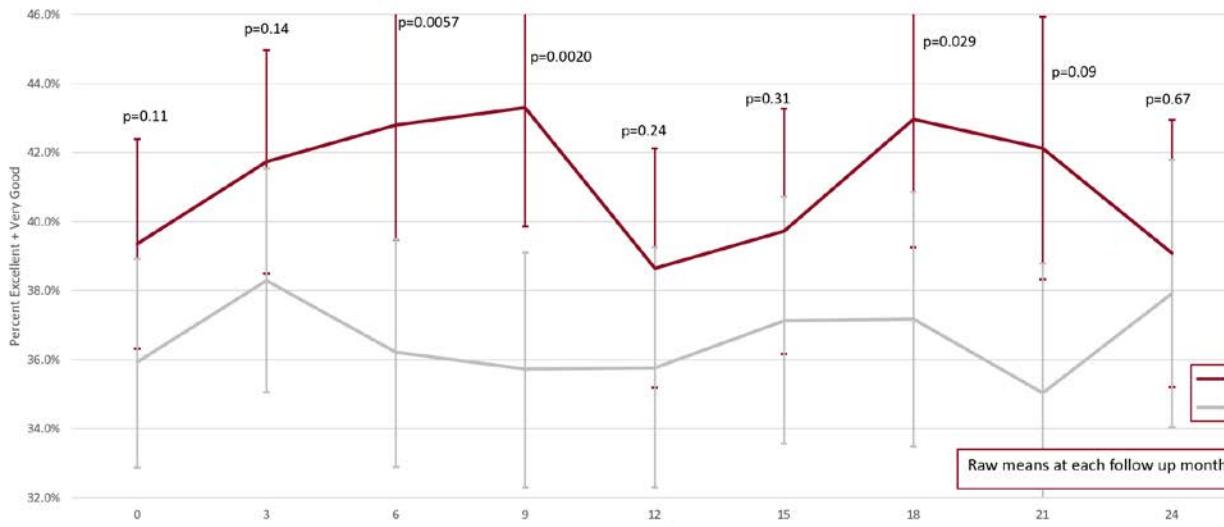
CCCCCCCC  
CCCCCCC D  
CCCCCCC R  
CCCCCC D D  
CCCCCC O O  
CCCCCC RC  
CCCCCR RD  
CCCCCR RR

- $Y_{i,j} = \beta_0 + \beta_1 C4P_i + \beta_2 \sqrt{round_j} + \beta_3 (C4P_i * \sqrt{round_j}) + \beta_0^D Drop_i + \beta_1^D (Drop_i C4P_i) + \beta_2^D (Drop_i * \sqrt{round_j}) + \beta_3^D (Drop_i * C4P_i * \sqrt{round_j}) + v_{0i} + v_{1i} \sqrt{round_j} + \epsilon_{ij}$ 
  - i: subjects, j: observation
- $\beta_0 \sim \beta_3$  are for completers,  $\beta_0^D \sim \beta_3^D$  are how dropouts differ from completers
- $\beta_3$  is the variable of interest.

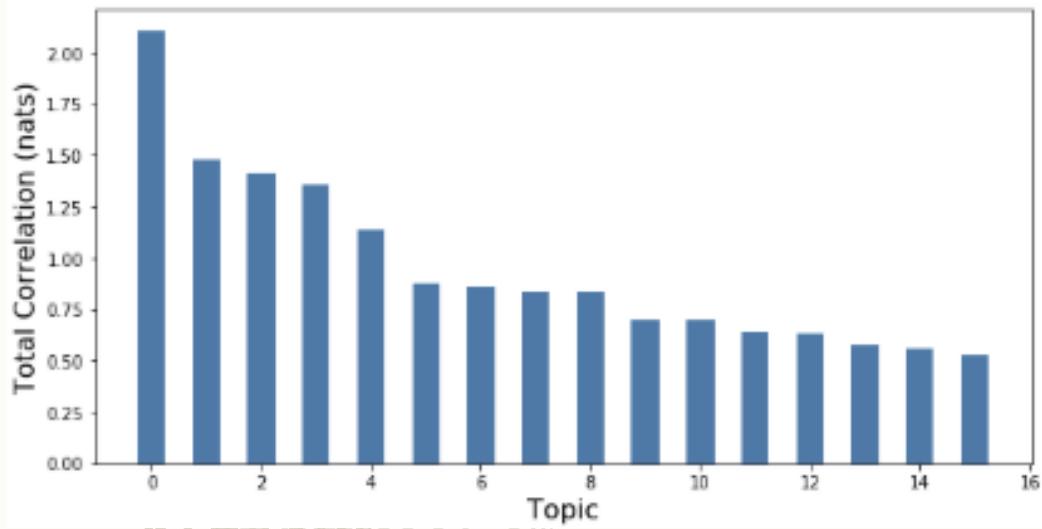
# Follow-up Hospitalizations



# Mental Health Rating



# Topic Modeling for Goals in C4P



- Use Latent Dirichlet Allocation for Topic Modeling
- 16 ideal topics for both Health and Life goals reported by patients
  - Top health topic: control blood sugar, control pain, able to XX again, lose XX pounds
  - Top life topic: reunite with children, family members, work, travel, live long



# Still to-do: Goal Attainment with MEPM

---

- $$Y_{i,j} = \beta_0 + \beta_1 Goal_i + \beta_2 \sqrt{round_j} + \beta_3 (Goal_i * \sqrt{round_j}) + \beta_0^D Drop_i + \beta_1^D (Drop_i Goal_i) + \beta_2^D (Drop_i * \sqrt{round_j}) + \beta_3^D (Drop_i * Goal_i * \sqrt{round_j}) + v_{0i} + v_{1i} \sqrt{round_j} + \epsilon_{ij}$$
- Getting insignificant results due to small sample size

Nice work!

Who is up next?

# The Aesthetics of Knowledge Consumption:

**Investigating Stylistics and Representation in Online Science Communication**

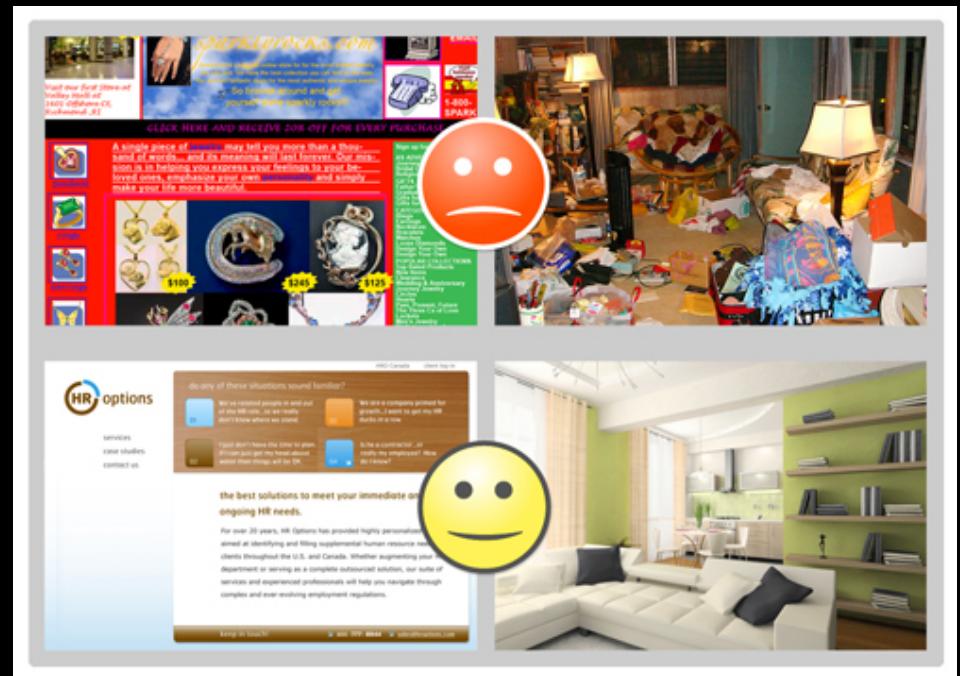
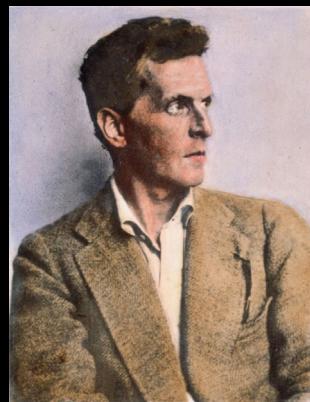
Leeson Hoay

**Advisor: A/P Hoyt Long** - Knowledge Lab | Department of East Asian Languages and Civilizations

**Additional Thanks: A/P Patrick Jagoda** - Department of English | Department of Cinema and Media Studies

# FOUNDATIONS AND RESEARCH QUESTION

- “Knowledge is in the end based on acknowledgement.”/“Ethics and Aesthetics are one.” - Wittgenstein
- Value and Aesthetics are inextricable (Gombrich, 1960)
  - **Especially in communicative acts – science communication is not spared from this!**



# Talking about Aesthetics in Science

- The effects of aesthetics and writing style has been widely studied in marketing/commerce/design
  - Well-designed visuals promote attention (Markovic, 2012)
  - And increase perceived reliability (Robins & Holmes 2008, Alsudani & Casey 2009, Goering et. al 2011)
- In education
  - Well-written, readable texts result in better motivation and problem-solving (Walkington et. al)
- Science Communication – to advance the knowledge society, to increase interest in science and encourage knowledge sharing
  - General public outside of scientific circles consume scientific content **through media**
  - The “network” (Jagoda, 2016)

# Science Communication

- Studies in science communication have been traditionally focused on contextual/social factors, and public literacy (Nisbet and Scheufele 2009 and others)
  - Less emphasis on actual technique and quantifiable metrics
  - Calls by sociologists over recent years to develop more formalized examinations of hermeneutics and knowledge diffusion (Declich & d'Andrea 2005, Leydesdorff 2009, Nielsen 2013)
- **Are computational measures of aesthetics and style robust enough in predicting human readers' perceptions along similar measures?**
- **Can we 'open the door', so to speak, to more empirical pathways through which science communication can be assessed?**

# Defining ‘Aesthetics’ and Stylistics

- **Visual Aesthetics**
  - Colorfulness (Reinecke et. al 2013)
  - Screen Balance (Ngo et. al 2000, Altaboli & Lin 2011)
- **Textual Style**
  - Readability (De-jargonizer, Flesch-Kincaid)
  - Uncertainty/Hedging (Vincze et. al 2014)
- **How do these correlate with readers’ perception of:**
  - Aesthetic Design
  - Colorfulness
  - Tidiness
  - Reliability
  - Readability
  - Enjoyment

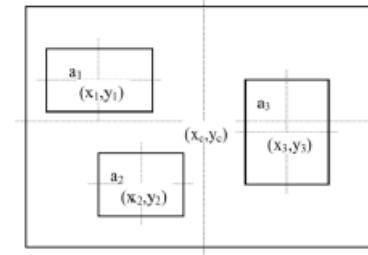


Figure 5: A screen in equilibrium (Ngo et. al, 2000).

might not be  
wonder trying to say  
maybe  
could be perhaps  
**probably**  
to some degree  
some confidence  
probably true

‘Hedge words’

# Study Design

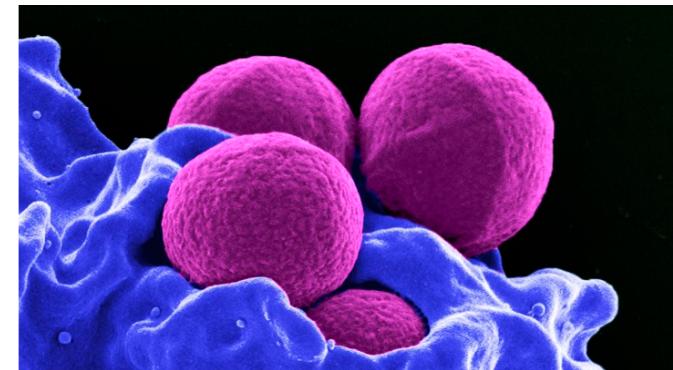
- Science Article extracted from The Atlantic
  - Manipulated on Colorfulness (High/Med/Low)
  - Screen Balance (Balanced/Not Balanced)
  - Readability (High/Low)
  - Level of Uncertainty (High/Low)
- 250 MTurk Participants
  - With demographic controls

SCIENCE

## In Bacteria, Persistence Leads to Resistance

No, this is not a metaphor.

ED YONG FEB 9, 2017



MRSA (REUTERS)

The threat of drug-resistant bacteria grows more pressing with every year. These microbes can shrug off the most potent antibiotics, including some [drugs of last resort](#). Some bacteria have become resistant to [all of our available drugs](#). Scary stuff, but bacteria don't have to *resist* antibiotics to defy them. There is another way—a much simpler, very common, and largely unappreciated one.

MMG  
Mercy Medical Group

Need an appointment?  
Schedule a time convenient for you.

BOOK TODAY

# Study Design (Tools)

- **EBImage (R)** to assess and manipulate visual aesthetics
  - aPixel Manipulation, Element Distance
- **De-jargonizer** (Rakedzon et. al, 2017)
  - Readability model trained on 250k BBC articles
- Flesch-Kincaid Reading Ease
- Uncertainty Classifier (Vincze et. al, 2014)

```
In [51]: cls = Classifier(binary = True)
pred_list = cls.predict("The scientists may have found something that seems to hint at the existence of dark matter.")
print(pred_list)
print("")

total = 0
for j in pred_list:
    if j == 'U':
        total+=1
print("Uncertainty score: ", total/len(pred_list))

['C', 'C', 'U', 'C', 'C', 'C', 'U', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C']

Uncertainty score: 0.11764705882352941
```

Uncertainty Scoring: Hedge Words ÷ Total Words

Colorfulness is operationalized by first defining the opposing color spaces:

$$rg = R - G \quad (2)$$

$$yb = \frac{R - G}{2} - B \quad (3)$$

Then define the standard deviation( $\mu$ ) and mean( $\sigma$ ), before computing the colorfulness metric  $C$  (Hassler and Susstrunk, 2003):

$$\mu_{rgyb} = \sqrt{\mu_{rg}^2 + \mu_{yb}^2} \quad (4)$$

$$\sigma_{rgyb} = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} \quad (5)$$

$$C = \sigma_{rgyb} + 0.3 \times \mu_{rgyb} \quad (6)$$

Operationalizing Colorfulness

# Synonym Replacement

- Using Wordnet's `synset.lemma_names()` to obtain synonym lemmas of a word, then checking it against the word rarity model borrowed from Rakedzon et. al 2017 -> replace rare words with more common words

```
In [94]: def replacer(sent, ref):
    ls_rep = sent.split(" ")
    new_sent = []
    for i in ls_rep:
        if i in list(ref[0]):
            if(int(ref.loc[ref[0]] == i)[1]) < 600:
                syn = wn.synsets(i)[0]
                for j in syn.lemma_names():
                    if j in list(ref[0]):
                        if(int(ref.loc[ref[0]] == j)[1]) < 600:
                            pass
                        else:
                            new_sent.append(j)
                            break
                    else:
                        new_sent.append(j)
                        break
            else:
                new_sent.append(i)
        else:
            new_sent.append(i)

    return new_sent
```

```
In [98]: sent1 = "The apoptosis observed in the sample was unusual."
print("Original sentence: " + sent1)
print("")
print("New sentence: " + " ".join(replacer(sent1, bbc)))
```

Original sentence: The apoptosis observed in the sample was unusual.

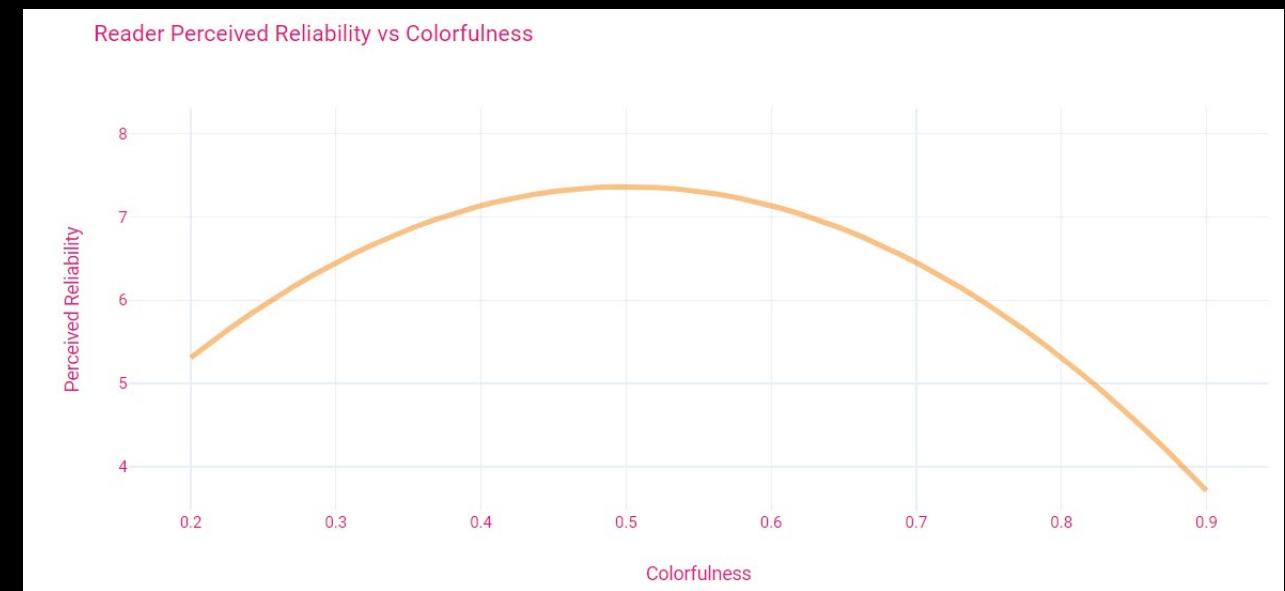
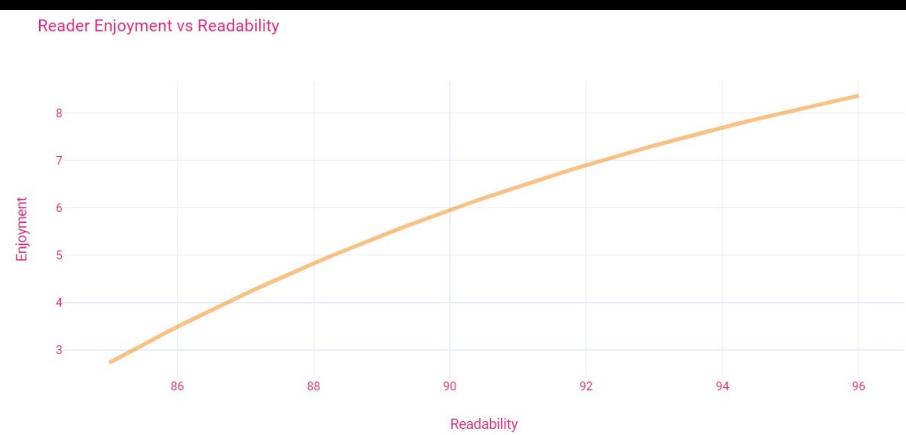
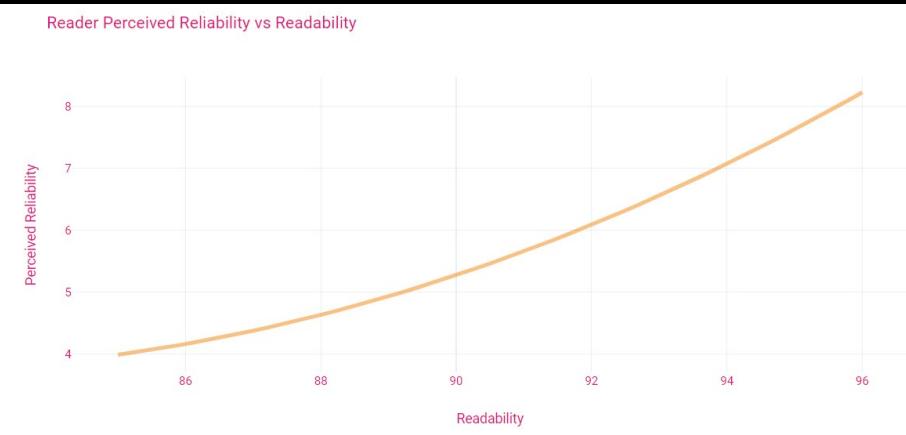
New sentence: The programmed\_cell\_death observed in the sample was unusual.

# Study Design

- On a scale of 1-9, how would you **rate the design** of the webpage?
- On a scale of 1-9, how **colorful** was the webpage?
- On a scale of 1-9, how **tidy** was the webpage layout?
- On a scale of 1-9, how **difficult** was the article to read?
- On a scale of 1-9, how **reliable** do you think was the information in the article you just read?
- On a scale of 1-9, how much did you **enjoy reading** the article?

\* Participants' 'linger time' were also recorded.

# Results



NS relationship between linger time and readability.  
Possible hypothesis: If it is easy to read, then it will take less time. If it is difficult to read, participants may give up on digesting the obscure portions?

Nice work!

Who is up next?



# PREDICTING ICU TRANSFERS IN HOSPITALIZED CHILDREN

Yangyang Dai

Advisor: Dr. Anoop Mayampurath



## BACKGROUND

- An ICU transfer is associated with:
  - Increased mortality
  - Increased neurological complications
  - Increased hospital cost

- Currently, University of Chicago Medicine uses the Pediatric Early Warning Score (PEWS) to recognize risk of ICU transfer.
- Subjective

## VITAL SIGN MODEL

- Study Details

- Pediatric patients (age < 18 years) admitted to the general ward during years 2009-2018.
- 38,199 patients, 1,375 (3.7%) experienced outcome - ICU transfer
- Variables: six vital signs (heart rate, temp etc.). plus patient characteristics (age, gender etc.)

- Train: 2009 – 2014 Test: 2015 – 2018
- 10-fold cross validation
- Performance
  - VS Model predicted ICU transfer better than PEWS (AUC 0.78 vs 0.72, P < 0.01 12 hours in advance of event).

# OBJECTIVE

- To investigate if
  - Adjusting for collinearity improved prediction
    - (used LASSO)
  - Adjusting for non-linear trends in vital signs improves prediction
    - (used RCS)
  - Adjusting for complex interactions between variables improves predictions
    - (used gradient boosting)

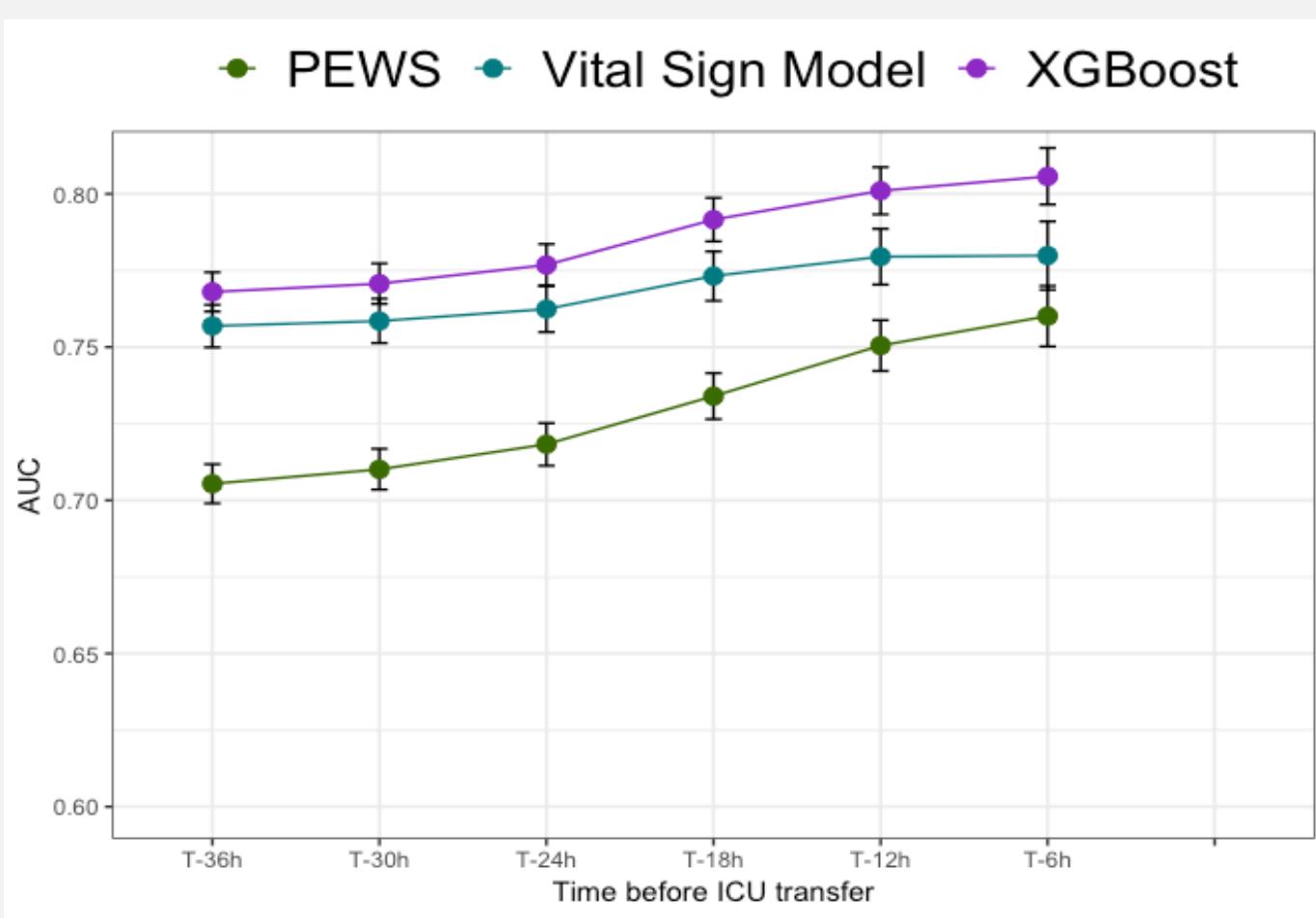
## PATIENTS' STATISTICS

	Patient admissions with ICU transfer (n=1,375)	Patients admissions without ICU transfer (n=36,824)
<b>Age yrs, mean (sd)</b>	<b>6 (6)*</b>	<b>7(6)</b>
<b>Female, n (%)</b>	<b>632 (46)</b>	<b>16,699 (45)</b>
<b>Race, n (%)</b>		
Black	<b>821 (60)</b>	<b>21,756 (59)</b>
White	<b>387 (28)</b>	<b>10,623 (29)</b>
Other	<b>167 (12)</b>	<b>4,445 (12)</b>
<b>Hispanic, n (%)</b>	<b>183 (13)</b>	<b>4,308 (12)</b>
<b>Mortality, n (%)</b>	<b>44 (3)*</b>	<b>24 (0.07)</b>
<b>Hospital length of stay days, median (IQR)</b>	<b>9 (5, 18)*</b>	<b>2 (1, 4)</b>
<b>Admit Location, n (%)</b>		
Ward	<b>244 (18)*</b>	<b>4,893 (13)</b>
ED	<b>664 (48)</b>	<b>18,916 (51)</b>
Other	<b>467 (34)</b>	<b>13,015 (36)</b>

\* p<0.01

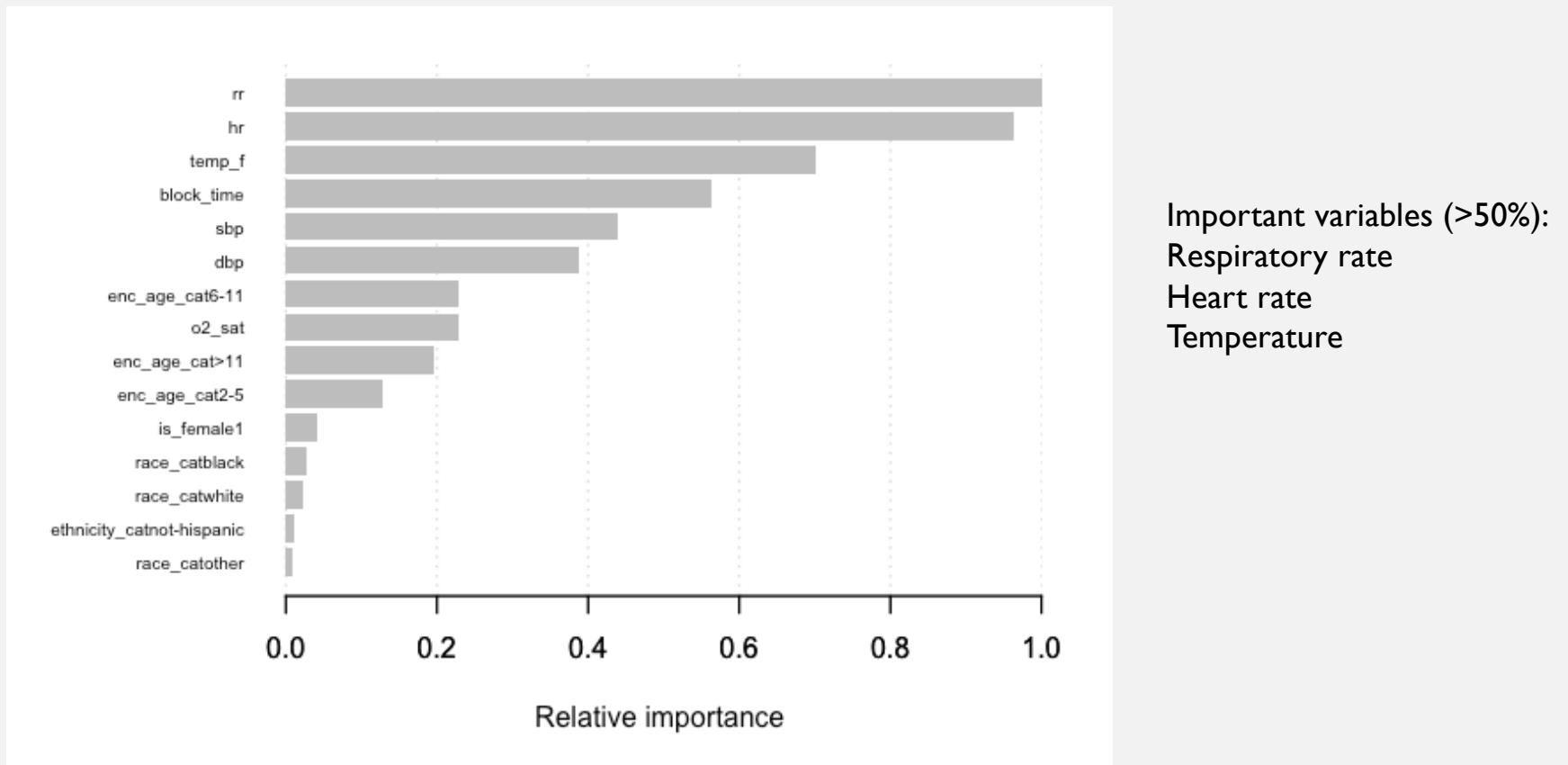
## TABLE OF AUCS

Model	PICU AUC
VS Model	0.7798
Lasso	0.7798
rCS	0.7798
XGBoost	0.801



XG boost predicts ICU transfer better and earlier than VS model and PEWS

### Variable importance plot – Gradient Boosting Model



Nice work!

Who is up next?



The  
image  
part with  
relations

MACSS Conference  
Lightning Talk

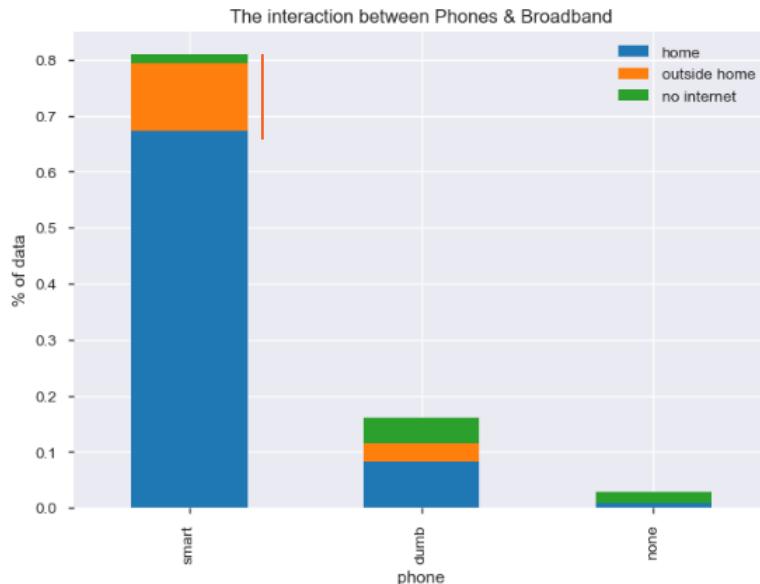
Laurence (w4rner) Warner  
[w4rner.tech](http://w4rner.tech)

Advisor: Prof. Ben Soltoff

Research Topic:

'Smartphone Dependence'

Smartphone + no Broadband => Smartphone dependent



Perspective 1:  
Smartphone Dependence  
as lack of alternative  
devices.

The image part with relations

Research Question:  
Which US demographic  
factors are associated  
with smartphone  
dependence?

Data: Pew Research 2018

Contributor: Darian  
Bailey



The image part with relations

```
Optimization terminated successfully.  
Current function value: 0.337199  
Iterations 7
```

#### Logistic Regression

Dep. Variable:	dep	No. Observations:	1561			
Model:	Logit	Df Residuals:	1550			
Method:	MLE	Df Model:	10			
Date:	Mon, 30 Jul 2018	Pseudo R-squ.:	0.1563			
Time:	19:29:09	Log-Likelihood:	-526.37			
converged:	True	LL-Null:	-623.85			
		LLR p-value:	1.807e-36			
	coef	std err	z	P> z	[0.025	0.975]
const	2.4003	0.446	5.384	0.000	1.526	3.274
sex	-0.2434	0.163	-1.494	0.135	-0.563	0.076
age	-0.3054	0.072	-4.264	0.000	-0.446	-0.165
educ	-0.2246	0.054	-4.151	0.000	-0.331	-0.119
non_hisp	-0.8502	0.234	-3.636	0.000	-1.309	-0.392
inc	-0.2428	0.038	-6.466	0.000	-0.316	-0.169
white	0.4396	0.305	1.440	0.150	-0.159	1.038
black	0.7382	0.353	2.092	0.036	0.047	1.430
asian	0.0807	0.584	0.138	0.890	-1.063	1.225
other	1.7867	0.770	2.320	0.020	0.277	3.296
native	1.3406	0.563	2.382	0.017	0.237	2.444

Statistically significant negative relationships:

- Intuitive: Age, Income
- Notable: Education



The  
image  
part with  
relations

# Too Small to Bridge the Digital Divide: Demographic Inequalities in Smartphone Technology Uptake. \*

Laurence Warner<sup>†</sup>

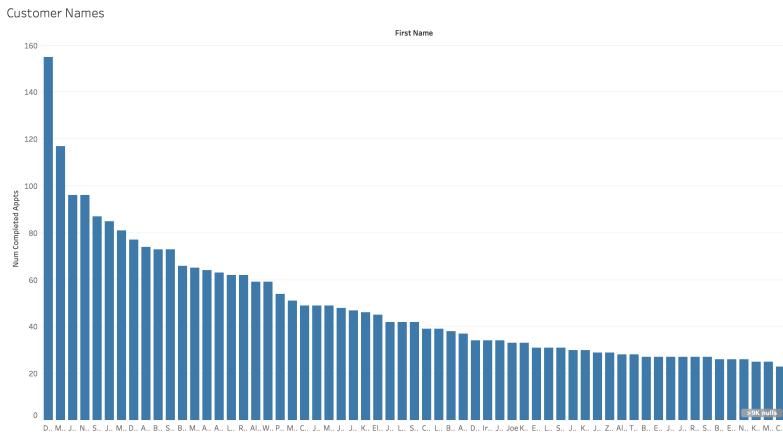
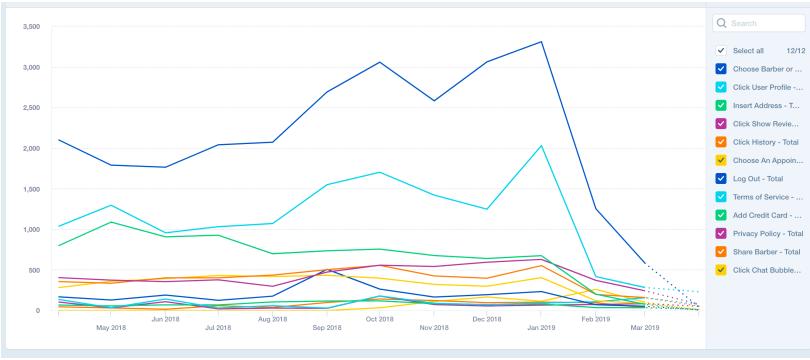
July 31, 2018

## 1 Abstract

Advances in communication technology has revolutionized American society. Internet, high-speed broadband and mobile phone technology have all swept across society to near-ubiquity. However, there was a 'digital divide' in speed of uptake across socioeconomic groups in society. Moreover, there still exists a 'usage gap' in which social groups derive differential benefits from new technology. Internet-enabled smartphones represent a huge technological advance in the way we access the internet, and some detect a bright future for mobile-only internet. Whilst smartphone ownership in general is correlated with high socioeconomic status, there is a growing portion of the country who rely upon this technology and are deemed 'smartphone dependent'. This paper shows that smartphone dependence is most prevalent amongst already disadvantaged groups. Considering sociological theories of inequality, this could forebode a widening digital divide.

Keywords: Technology, Internet, Demographics, Sociology, Statistical Inference, Limited Dependent Variable Models





The image part with relations

Supplementary Data:



30k users.

**SHORTCUT**

Advice: Prof. Paolo  
Parigi, Stanford  
University



The  
image  
part with  
relations

Perspective 2: Smartphone Dependence as behavioral addiction

### **First Week**

**Flip Phone:**

**The**

**Smartphone**

**Detox**

|

Research Question: Is Digital Detox effective at reducing smartphone usage?

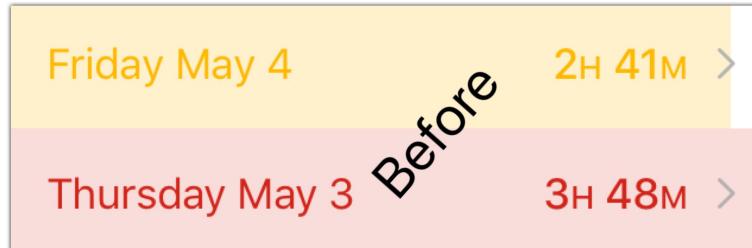
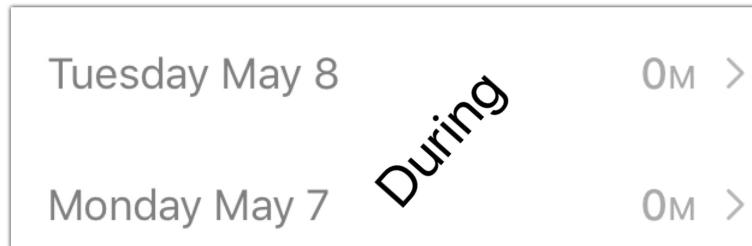
Data: First Week Flip Phone

Contributors: Bhargav Desikan & Abe Pandit

---

 The  
image  
part with  
relations

# Screen Time Tracking



The image part with relations

# Pilot Study

“During the three days without my phone, I found myself constantly tapping around for my phone and having heart palpitations until I realized my phone was safe and sound in a locked box. I found myself checking the time using wall clocks instead of checking the time on my phone. I don't wear watches. While doing my AP Spanish homework, I couldn't just look up a vocabulary word if the definition had slipped my mind. I had to think of a synonym or completely rewrite the phrase. My productivity levels skyrocketed, all three days I finished almost all of my homework whilst at school. Overall, I do think the absence of my phone increased my level of productivity and I was able to accomplish more.”

“The phone detox challenge was very hard and challenging through out my school day. The detox really made me realize the things I can't do without my phone. I participated four out of the five days of the challenge. My average screen time before the detox was 5 hours and 29 minutes per day. After the challenge, my screen time reduced 2 hours, to 3 hours and 44 minutes per day. One challenge that I faced was communication. It was very hard to contact my friends because I could not call them to find out where they were. Also, in some of my classes we have lots of spare time to get on our phones and because I was phoneless it forced me to be more productive and do work for other classes, which brought my grades up. Another challenge was checking my grades for my classes, but the hardest part of being with out a phone was checking the time. Although I know how to read the wall clocks in my classes, my first resort for checking the time is my phone. Apps that I missed the most was Instagram and YouTube, which was the apps that took up most of my screen time. The detox did not make me associate with different people, but it really made me focus on my work. Although the detox made me focus more on my work and be more attentive in class, I feel as if our generation revolves around phones because they are necessities throughout our everyday lives.”

Average daily screen time (hours)		
Week Before	Week After	
1	2	0.25
2	4	0
3	5.3	4
4	5.75	3.2
5	4	2.75
6	6.5	5.5
Average	4.6	2.6
% Reduction		43



The  
image  
part with  
relations

# Questions?

Suggestions?



The  
image  
part with  
relations

# Questions?

Suggestions?

Slide Design:



Nice work!

Who is up next?

# Size of Local Labor Market and Productivity: A Quantitative Analysis of China

Xiang Zhang

Thesis Advisor: Professor Richard Hornbeck

- Economic activity is spatially concentrated (Duranton and Puga[2004]). Higher density of local economic activities creates productivity advantages (Glaeser and Gottlieb[2009], Greenstone, Hornbeck, and Moretti[2010])
- Larger labor markets generate agglomeration economy
  - ▶ The quality of matching between firms and workers is higher in larger labor markets (Diamond[1982], Helsley and Strange[1990])
  - ▶ Concentration of human capital generates positive externality and increases productivity (Lucas[1988], Glaeser[1999], Moretti[2004])

# Introduction

- Larger labor markets generate agglomeration economy
  - ▶ The quality of matching between firms and workers is higher in larger labor markets (Diamond[1982], Helsley and Strange[1990])
  - ▶ Concentration of human capital generates positive externality and increases productivity (Lucas[1988], Glaeser[1999], Moretti[2004])
- Migration changes size of local labor markets, but also changes the composition of workers
  - ▶ Number of migrant workers in China: 37 million in 1997, 145 million in 2009, and 245 million in 2017 (National Bureau of Statistics), the majority of whom are low-skill workers
- Question: Do expanded local labor market size increases or decreases productivity?

# Conceptual Framework

- The production function of city  $c$  in year  $t$  can be written in the following CES function

$$Y_{ct} = K_{ct}^\alpha \left[ (A_{ct}^H H_{ct})^{\frac{\sigma-1}{\sigma}} + (A_{ct}^L L_{ct})^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}(1-\alpha)}$$

where  $K_{ct}$  is the capital stock,  $H_{ct}$  and  $L_{ct}$  represent the number of high-skill and low-skill workers respectively, and  $A_{ct}$  is the total factor productivity

- Rewrite the production function as

$$Y_{ct} = K_{ct}^\alpha [A_{ct}\phi_{ct}(H_{ct} + L_{ct})]^{(1-\alpha)}$$

where

$$\phi_{ct} = [(\beta_{ct}h_{ct})^{\frac{\sigma-1}{\sigma}} + ((1 - \beta_{ct})(1 - h_{ct}))^{\frac{\sigma-1}{\sigma}}]^{\frac{\sigma}{\sigma-1}}$$

- $A_{ct}\phi_{ct}$  represents the local productivity
- In this setting,  $\beta_{ct}$  captures the relative productivity between high-skill and low-skill workers,  $h_{ct}$  is the fraction of high-skill workers (Peri[2010])

# Conceptual Framework

- Improved matching quality and knowledge spillover imply that productivity is an increasing function of labor market size

$$\frac{\partial A_{ct}\phi_{ct}}{\partial N_{ct}} > 0$$

- A higher proportion of high-skill workers will raise the overall productivity

$$\frac{\partial A_{ct}\phi_{ct}}{\partial h_{ct}} > 0$$

# Measurement of City Productivity

- Assume a Cobb-Douglas productivity function at firm-level, I estimate city-level TFP by running the following regression using firm data (Greenstone, Hoenbeck, and Moretti [2010], Hornbeck and Moretti[2019])

$$\log(Y_{it}) = \beta_1 \log(L_{it}) + \beta_2 \log(K_{it}) + \beta_3 \log(M_{it}) + \gamma_{ct} + \varepsilon_{it}$$

where  $Y_{it}$  is the total output of firm  $i$  in year  $t$ ,  $L_{it}$  is the number of employee,  $K_{it}$  is the value of capital stock, and  $M_{it}$  is the material inputs

- The estimated city-by-year fixed effect  $\gamma_{ct}$  represents the average TFP of all firms in city  $c$  in year  $t$
- I also use the following methods to estimate city TFP to check the robustness
  - ▶ Olley and Pakes Method (Olley and Pakes[1996])
  - ▶ Cost-sharing Method
  - ▶ Levinsohn and Petrin Method (Levinsohn and Petrin[2003])

# Methodology and Instrumental Variables

- To analyze the effects of labor market size on productivity, I estimate the following specification

$$\log(A_{ct}) - \log(A_{ct-N}) = \alpha(\log(L_{ct}) - \log(L_{ct-N})) + \gamma_p + \varepsilon_{ct}$$

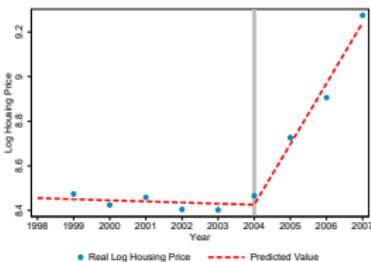
where  $A_{ct}$  is the productivity of city  $c$  in year  $t$ ,  $L_{ct}$  is population size,  $\gamma_p$  is the region fixed effect, and  $\varepsilon_{ct}$  is error term

- Taking the first-difference could attenuate endogeneity issue, but we still have factors that might correlate with both changes in TFP and changes in labor market size
- First IV: structural break in housing price (Charles, Hurst, and Matthew[2018])
- Second IV: shift-share IV

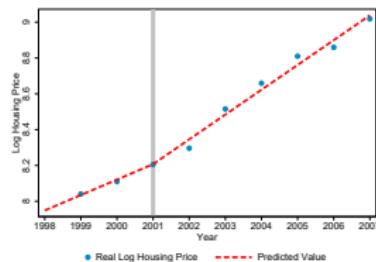
# First IV: Structural Break in Housing Price

- Housing boom in China largely increases the housing price, but the magnitude varies (Chen and Wen[2018], Glaeser et al.[2017])
- Abnormal changes in housing price could be a result of speculative investment behavior, but not changes in fundamentals (Scheinkman and Xiong 2003, Brueckner et al. 2016)
- I estimate the following equation for each city, the  $\lambda_k$  represents the magnitude of structural break in housing price

$$P_k^H(t) = \alpha + \tau_k t + \lambda_k(t - t_k^*) \mathbb{1}\{t > t_k^*\} + \varepsilon_{kt}$$



(a) Beijing: Huge Jump



(b) Shanghai: Medium Jump

## Second IV: Shift-share IV

- In shift-share IV, I instrument for population change during 2000 and 2010 using predicted population change (Altoji and Card[1991], Shimer[2001], Borjas[2003], Saiz[2007, 2010], Hornbeck and Moretti[2019])

$$IV_c = \alpha_{c,1990-2000} \times \Delta_{2000-2010}$$

where  $\alpha_{c,1990-2000}$  is the share of net population change during 1990-2000 of city  $c$ , and  $\Delta_{2000-2010}$  is the total net population change for all cities in China from 2000 to 2010

# Data

- Firm Data: Annual Survey of Manufacturing Firms 1998-2007
  - ▶ Census of all manufacturing firms with more than 5 million Yuan sales value (600 thousand USD in 2000)
  - ▶ Sample size increased from 160 thousand in 1998 to 340 thousand in 2007
- City Population Data
  - ▶ Covers all 334 cities in China
  - ▶ Statistical yearbook of all provinces
  - ▶ 1990, 2000, and 2010 Census data bulletin
- City Housing Price Data
  - ▶ Covers 208 cities from 2000 to 2008 (some cities also have data for 1998 and 1999)
  - ▶ Statistical yearbook of all provinces

# First IV Results: Effects of Labor Market Size on TFP

- Increase in population size lowers TFP
- The magnitude of negative effects gets smaller over time, suggesting labor market is responsive to influx of population

	(1) 2004-2007	(2) 2003-2007	(3) 2002-2007	(4) 2001-2007	(5) 2000-2007
Log Pop Difference: 2004-2007	-3.185*** (0.682)				
Log Pop Difference: 2003-2007		-1.876*** (0.631)			
Log Pop Difference: 2002-2007			-2.009*** (0.697)		
Log Pop Difference: 2001-2007				-1.637*** (0.542)	
Log Pop Difference: 2000-2007					-1.587*** (0.552)
Observations	208	208	208	206	207
F	3.784	3.801	3.819	3.803	3.867
Prob > F	0.0007	0.0007	0.0007	0.0007	0.0006

Standard errors in parentheses

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$

Nice work!

Who is up next?

# Understanding the Posts and Comments on SuicideWatch Subreddit

---

Lerong Wang

# Research Question

- What are some common suicidal risk factors implied by the original posts on SuicideWatch Subreddit?
- How do different linguistic characteristics affect the popularity of a given comment?

# Past Literature

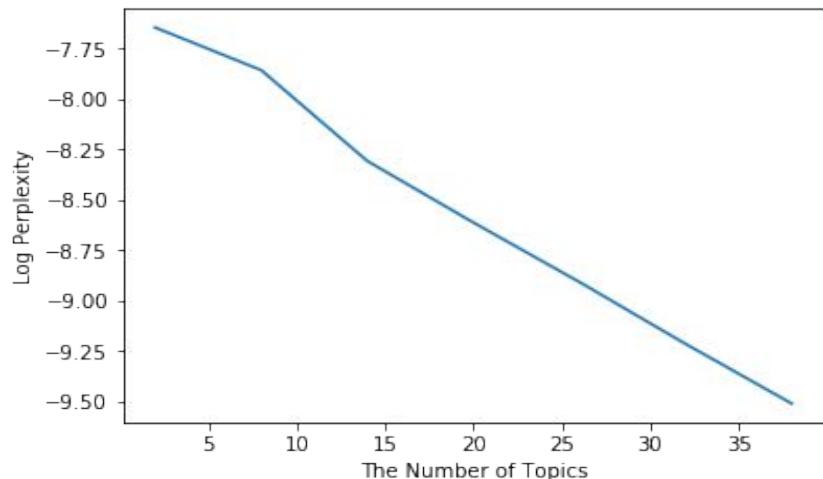
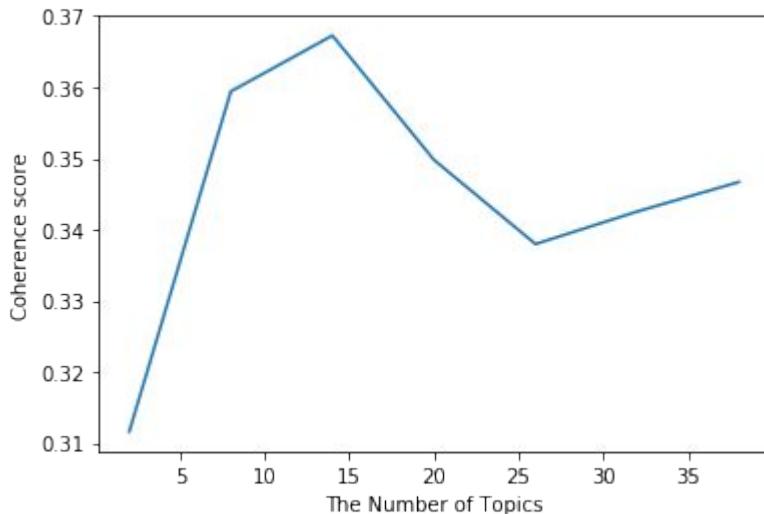
- An emphasis on issues related to depression, anxiety and detecting suicide ideation from social media posts
- A recent study (Grant et al., 2018) performed topic modeling on over 130000 original posts collected from SuicideWatch Subreddit: Word2Vec language models with k-means clustering

# Data

- Google BigQuery
- Post data were collected monthly from April 2018 to October 2018
- 24207 posts in total
- Performed tokenization, lemmatization on the posts using SpaCy
- Comment data were collected from December 2018
- 26967 comments in total
- Used LIWC (linguistic inquiry and word count) to get the semantic categories of words

# Methodology

- Topic modeling: Latent Dirichlet allocation (LDA) to discover topics from posts
- Choosing optimal number of topics
  - Coherence Score
  - Perplexity



# LDA results

Topic	Terms
topic1	end, die, love, kill, hate, keep, care, ever, always, give
topic2	depression, suicidal, anxiety, deal, scared, afraid, reach, past, struggle, depressed
topic3	mother, father, idk, stab, bill, daughter, stomach, provide, reject, ass
topic4	final, somebody, difficult, steal, helpful, subreddit, peaceful, gut, garage, replace
topic5	attempt, man, entire, mental, survive, vent, physical, hotline, pull, site
topic6	work, job, amp, money, collge, move, pay, able, experience, situation
topic7	sleep, night, wake, eat, bed, eye, drug, drink, morning, doctor
topic8	friend, talk, school, guy, close, relationship, girl, high, stuff, fail
topic9	last, back, leave, still, parent, family, home, well, mom, old
topic10	blood, painful, overdose, wrist, planet, slit, beg, fire, painless, fighting

Terms	Potential risk factors
depression, anxiety, scared, afraid, struggle, depressed	mental health conditions
mother, father, idk, stab, bill, daughter, stomach, provide, reject	relationship problems
attempt, man, entire, mental, survive, vent, physical, hotline	previous suicide attempts
work, job, money, college, pay, able, situation	financial difficulties
sleep, night, wake, eat, bed, eye, drug, drink, doctor	sleeping difficulties
friend, school, guy, close, relationship, girl, high, fail	school difficulties
last, back, leave, parent, family, home, mom, old	family violence/discord
blood, painful, overdose, wrist, slit, beg, fire, painless, fighting	drug abuse disorder

# How do different linguistic characteristics affect the popularity of a given comment?

- Outcome variable: comment score = ups - downs
- Choosing independent variables: 93 variables about semantic categories from LIWC
- Choudhury & De (2014). Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity.

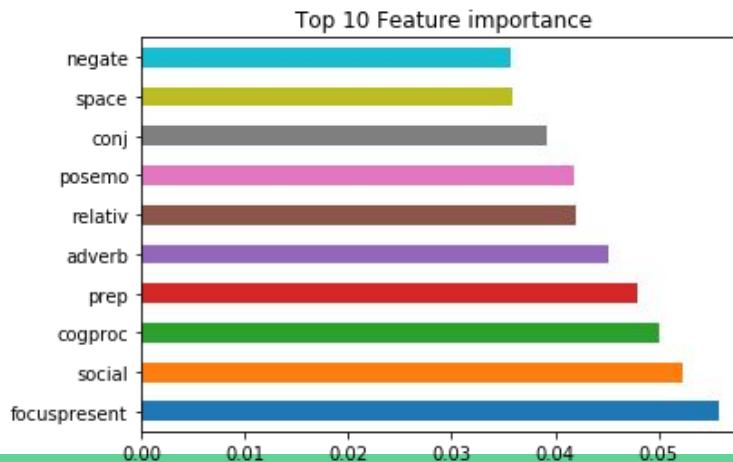
Independent variables				
1st person pronoun	Conjunction	Motion	Sexual	
2nd p. pronoun	Death	Negation	Social	
3rd p. pronoun	Discrepancy	Neg. emotion	Space	
Achievement	Exclusion	Numbers	Swear	
Adverbs	Health	Perception	Tense	
Assent	Home	Pos. emotion	Tentative	
Bio	Inclusion	Preposition	Time	
Body	Ingestion	Quantitative	Work	
Cause	Inhibition	Relationships		
Certainty	Leisure	Relativity		
Cognitive	Money	Religion		

# Linear regression vs. Random Forest vs. Decision Tree

- Linear regression

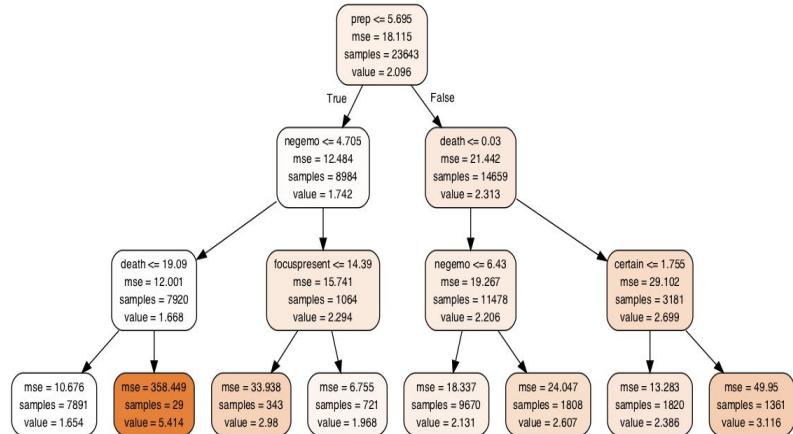
Significant features: discrepancy, cause, death, negative emotion, preposition, focus on present, focus on past are significant at the level of 0.05

- Random forest



# Linear regression vs. Random Forest vs. Decision Tree

- Decision Tree



Model	MSE
Linear Regression	16.128
Random Forest	15.191
Decision Tree	15.226

# Future Work

- Conduct PCA on the semantic categories from LIWC
- Cross-validation
- Speed up LDA
- More insights on the interpretations and correlations between topics

Nice work!

Who is up next?

# Participatory Logics in Digital Meme Culture: A Case Study of Reddit Dankmemes Sub-community

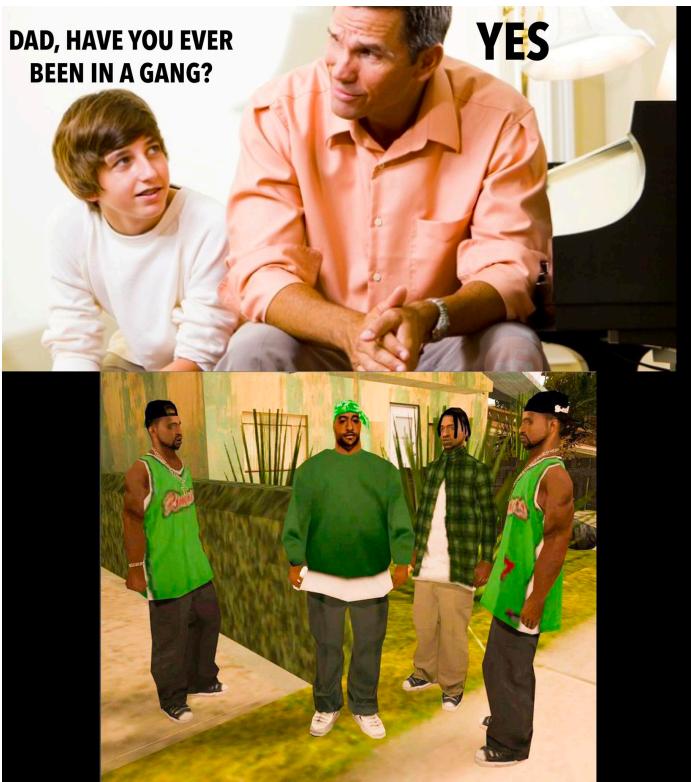


By Weiwei Zheng

Supervised by John Levi Martin

# Literature Review & Motivation

- Reddit r/dankmemes



↑ Alpha\_Demon2002 1.4k points · 9 hours ago  
↓ Duck genesis Evangelion.  
Reply Give Award Share Report Save

↑ dcxr OC Memer 485 points · 8 hours ago  
↓ Duck life 5  
Reply Give Award Share Report Save

↑ Eineg\_Htims\_Lliv WTF 240 points · 7 hours ago  
↓ Duck life 3 gang rise up  
Reply Give Award Share Report Save

↑ GalaxyMettaton 91 points · 5 hours ago  
↓ i hated grinding in that game  
Reply Give Award Share Report Save

↑ Sriracha-Lord 31 points · 3 hours ago  
↓ Coolmathgames intensifies  
Reply Give Award Share Report Save

1 more reply

↑ Ryan----- 52 points · 4 hours ago  
↓ Anyone here from coolmathgames  
Reply Give Award Share Report Save

↑ C-h-r-i-s-p-y-y-y-y 24 points · 4 hours ago  
↓ Was duck life 5 the dungeon runner one  
Reply Give Award Share Report Save

# Research Questions:

1. How is the language usage/topicality of memes related to their image patterns?
2. How are the textual and visual features of memes related to the discussion patterns in the comment?
3. What is a possible predictive model on the popularity of memes in this subreddit?

# Hypotheses

1. Abstract vs Realistic memes – Insulting/Sarcasm vs Plain languages
2. Realistic + Insulting memes – deeper graph; back and forth
3. Textual and visual of memes are useful features for clustering – from cluster to popularity
  - topicality, sentiment, usage of sarcasm +
  - colorfulness, entropy , (?categories)

$$\text{Entropy} = H = -\sum p_i \ln(p_i)$$

# Data

1) Web-scraping – Pushshift + Praw  
70 thousand posts – 2017.1 – 2019. 2

2) Feature extraction –  
Text – Google Cloud Vision API  
– Topic modeling – Gensim API  
Image – Google Cloud Vision API / PIL - Image  
Discussion pattern – NetworkX

# Method

- Pearson correlation
- Logistic regression – Inference
- Lasso & Ridge regression – Prediction

# Result I

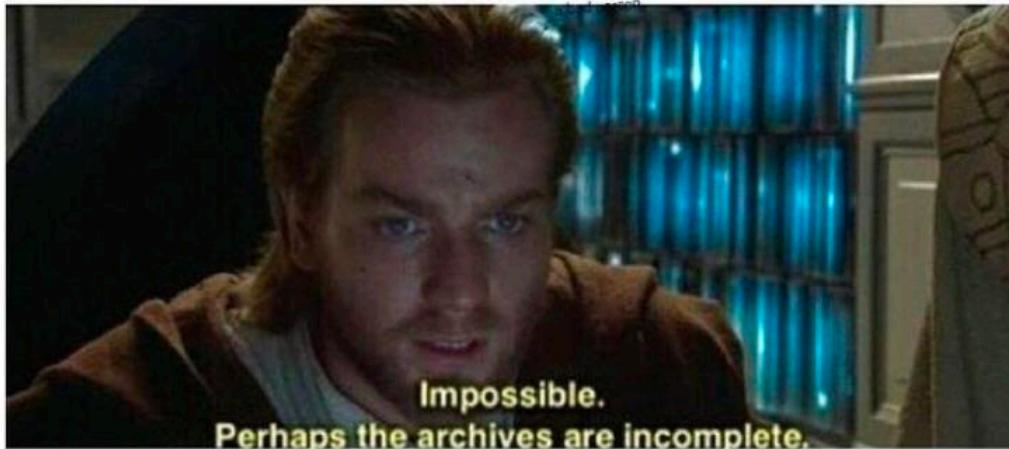
- everywhere is insult...
- discriminative & insulting wording – realistic memes

Script \*contains N-word  
White actors

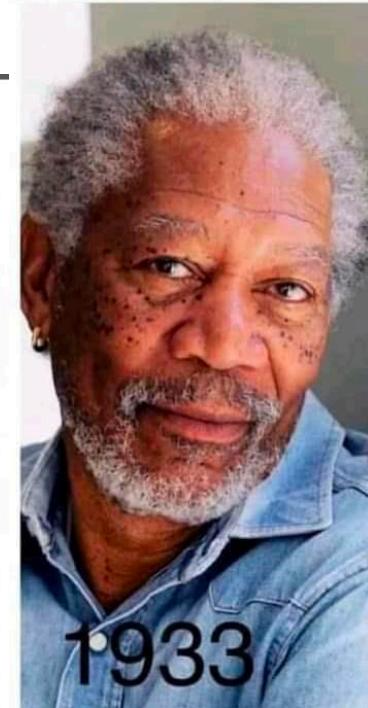


---

When a mod doesn't have  
gay porn in his search history



this nigga been old his whole life .



# Result II

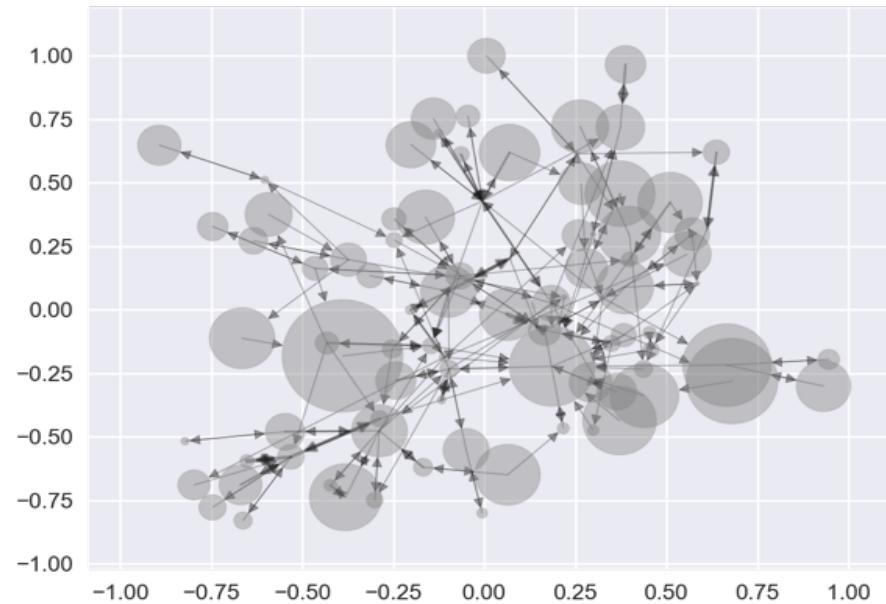
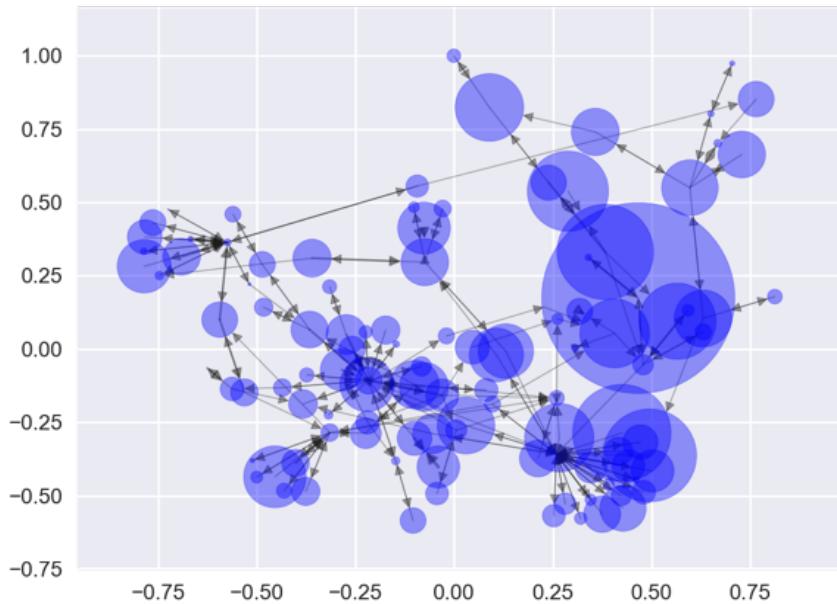
topic 6	topic 16	topic 21
free	make	remember
steal	mean	watch
sex	woman	take
library	body	learn
master	man	tv
man	water	state
teacher	child	fucking
want	fish	feel
eat	girl friend	girl friend
max	cat	japanese

	abstract	realistic	t-statistics	p-value
topic 6	0.0166	0.034	-3.448	0.000
topic 16	0.027	0.042	-2.1983	0.028
topic 21	0.0029	0.033	2.3477	0.019

variables	df	deviance	AIC	LRT	p-value
full model		1559.8	1575.8		
link karma	1	1563.2	1577.2	3.35	0.07
comment karma	1	1562.7	1576.7	2.92	0.09
# nods	1	1563.4	1577.4	3.62	0.06
density	1	1561.8	1575.8	2.02	0.15
# weak components	1	1563.7	1577.7	3.84	0.05
#strong components	1	1563.9	1577.9	4.09	0.04
mean comment link karma	1	1562.7	1576.2	2.90	0.09

variables	estimate	std error	z value	p-value
(intercept)	1.315	2.42 e <sup>-1</sup>	5.43	5.51 e <sup>-8</sup>
link karma	1.528e <sup>-7</sup>	9.173 e <sup>-8</sup>	1.72	0.085
comment karma	-1.727e <sup>-6</sup>	9.879 e <sup>-7</sup>	-1.75	0.081
# nodes	-8.02e <sup>-3</sup>	4.218 e <sup>-3</sup>	-1.9	0.057
density	1.577	1.15	1.4	0.170
# weak components	-3.66e <sup>-3</sup>	0.184	-2.0	0.047
#strong components	9.428e <sup>-3</sup>	4.702 e <sup>-3</sup>	2.0	0.045
mean comment link karma	-5.426e <sup>-6</sup>	3.115 e <sup>-6</sup>	-1.7	0.082

# Result III



# Next step

- Detecting sarcasm + Insult words – be more specific
- Build image classifiers with simple machine learning tools
  - KNN etc.
- Explore more possibilities from the comments structure  
  
(If I could get a PhD?)



 r/Showerthoughts 9h  
If you were born with your legs coming out first, for a moment, you wore your mom as a hat.  
Mindblowing

▲ 3.8k ▾ 130 Share

Nice work!

Who is up next?

# Local Context and Older Adults' Movement Patterns in Chicago

Tyler Amos  
18 April 2019  
Advisor: KA Cagney

***What is the relationship between local context and individual movement patterns?***

# Motivation

- Neighbourhood effects research (e.g., Sampson 2012)
  - Where you live shapes social, health outcomes
- But individuals don't just exist *in*, they select *into* a given context
  - To understand neighbourhood effects, we must improve our understanding of mobility patterns

# Hypotheses

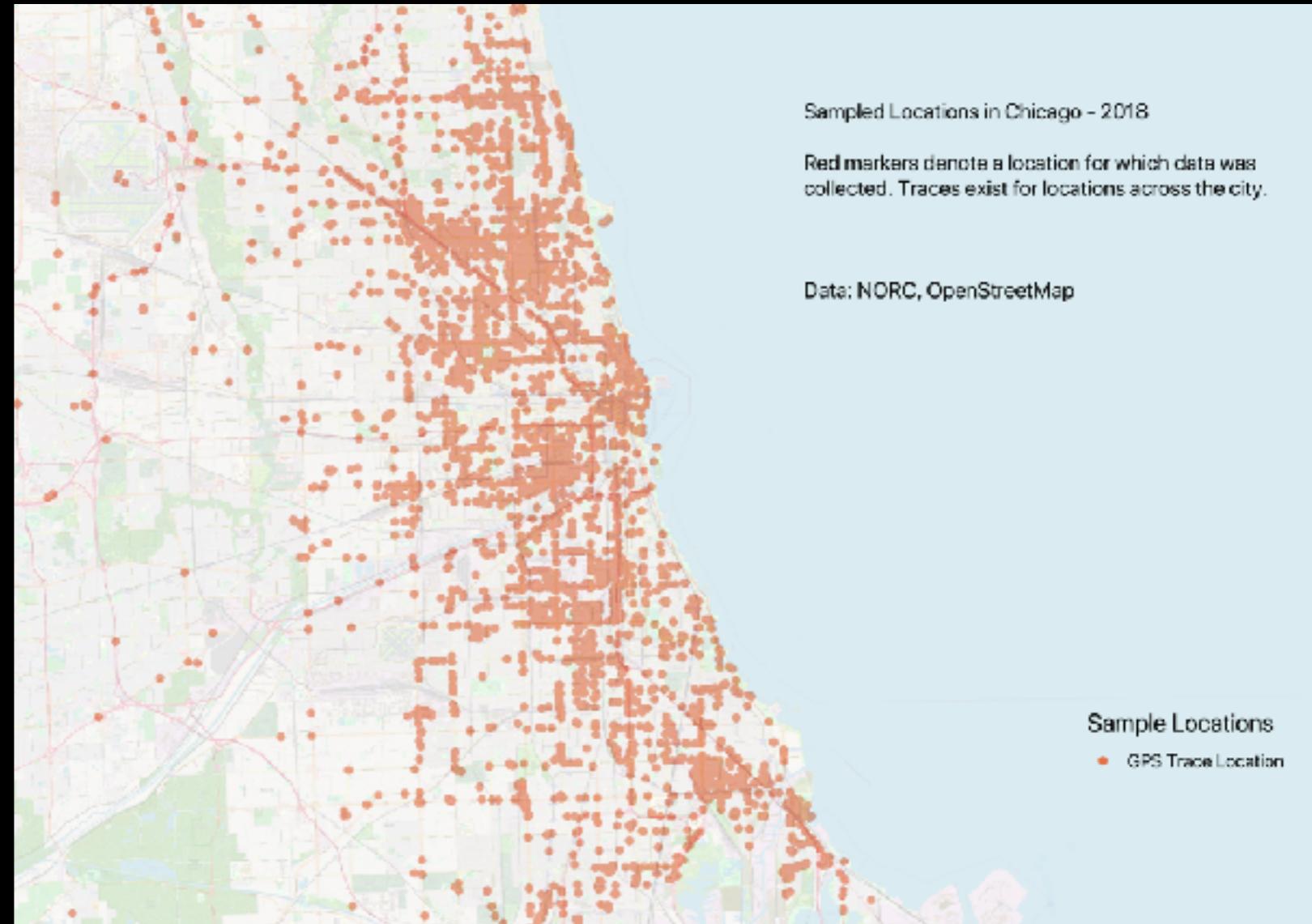
- (1) Measures of activity space derived from observed movement patterns (i.e., Autocorrelated Kernel Density Estimation) will provide a different picture of individual exposure to local context than administrative boundaries.
- (2) Individuals' daily movement patterns are associated with (i) the immediate spatio-temporal context in which movement takes place, and (ii) individuals' membership in social groups (e.g., race, gender).

# Contributions

- How best to model activity space
- Explores links between mobility and neighbourhood effects
- Connect neighbourhood effects/EMA-based literature with movement modelling methods from ecology

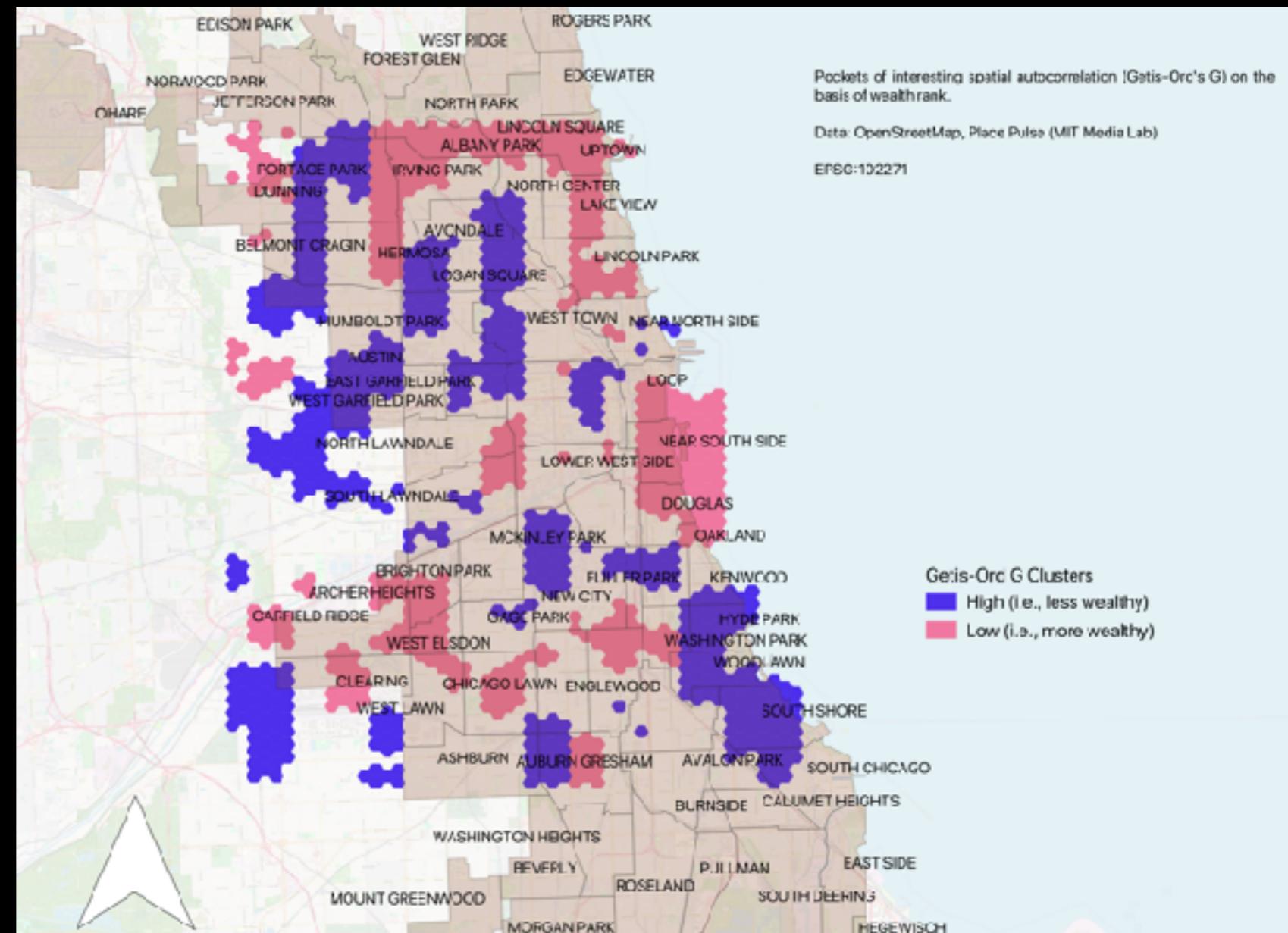
# EMA/GPS Traces

- GPS tracks collected for a sample of seniors in Chicago from NORC (PI K. Cagney)
  - 449 individuals, 1 week period in spring 2018



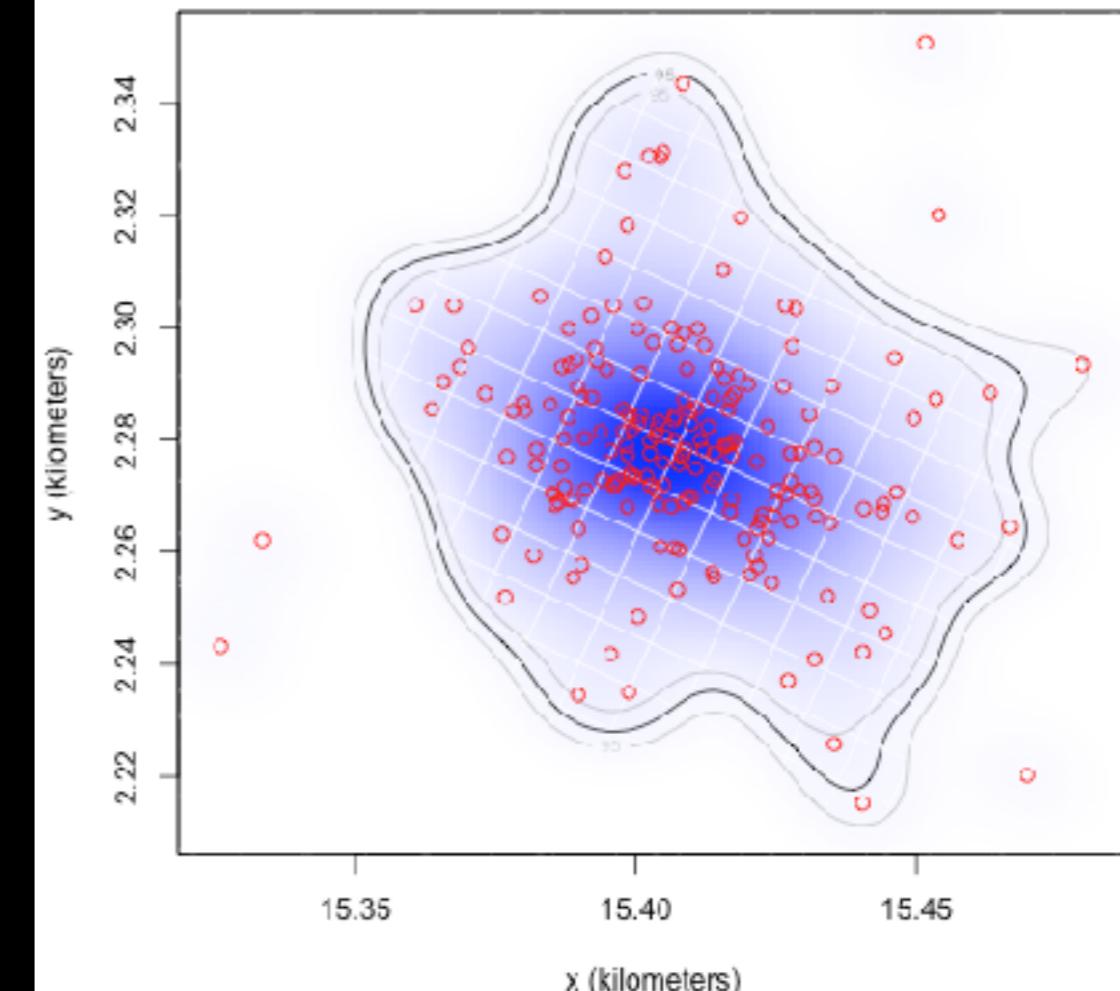
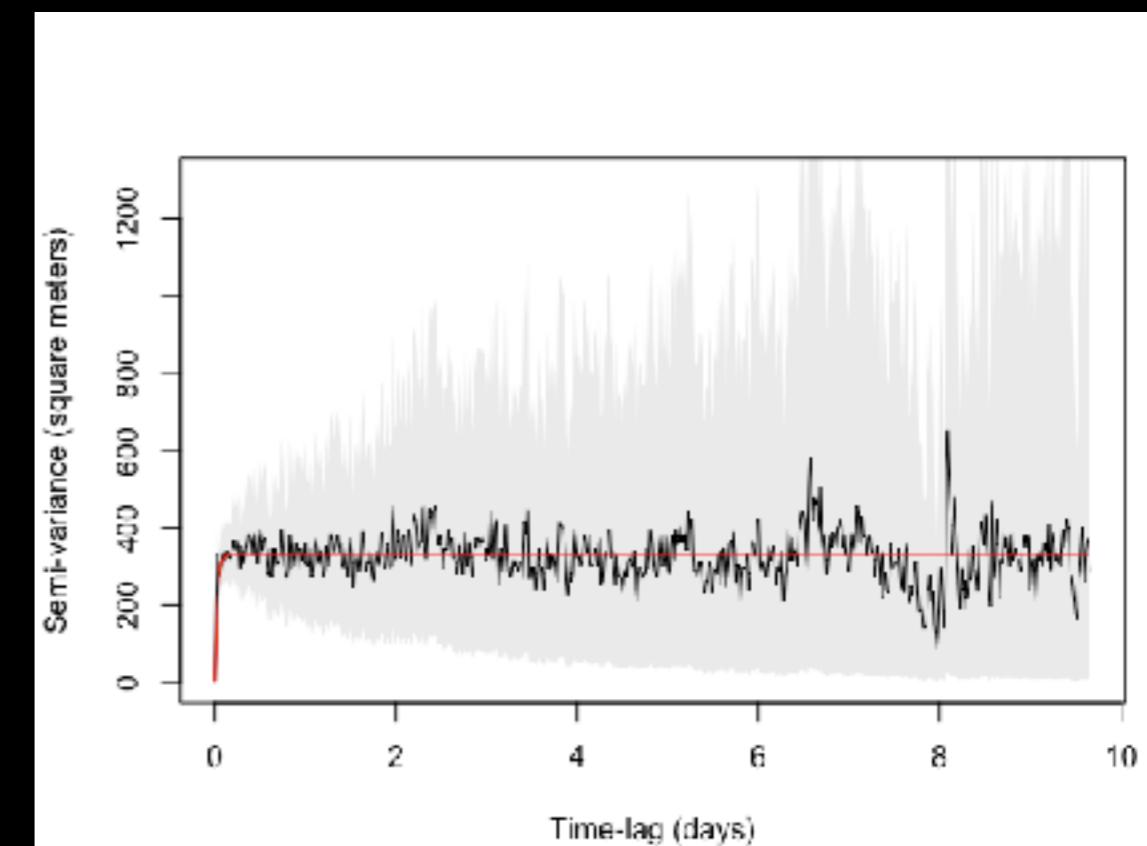
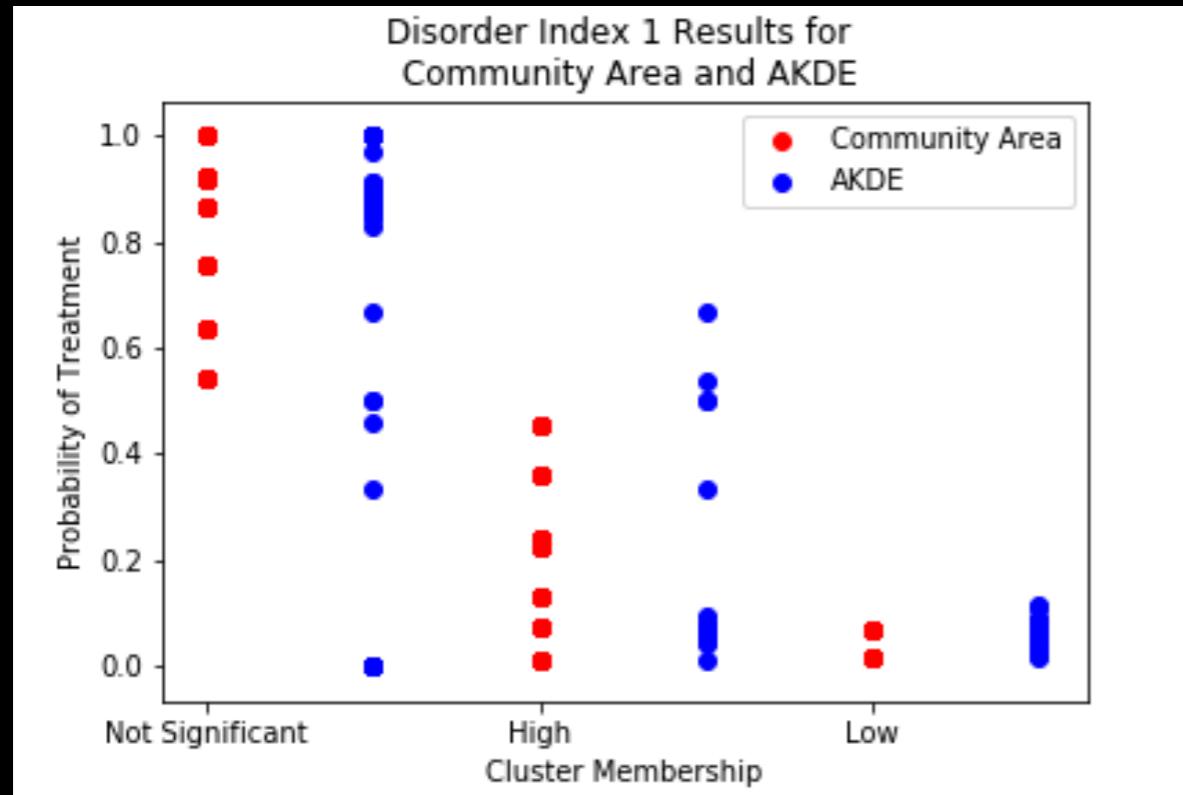
# Place Pulse

- Place Pulse ratings of Google StreetView scenes
  - Rankings along 6 dimensions (e.g., safety, depressing)
  - IDW interpolation, Multidimensional Scaling to reduce to 2 gridded indices, clustered with Getis-Ord Statistic



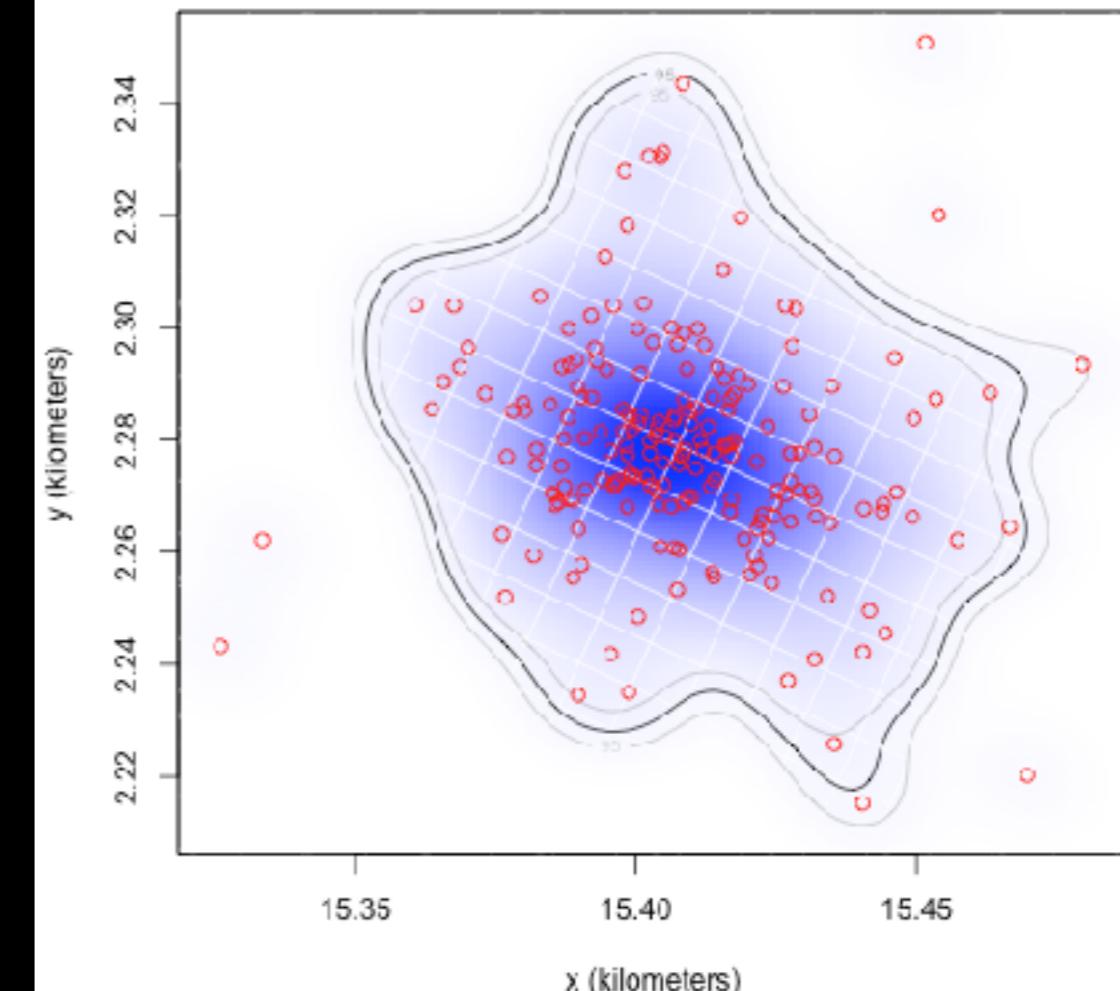
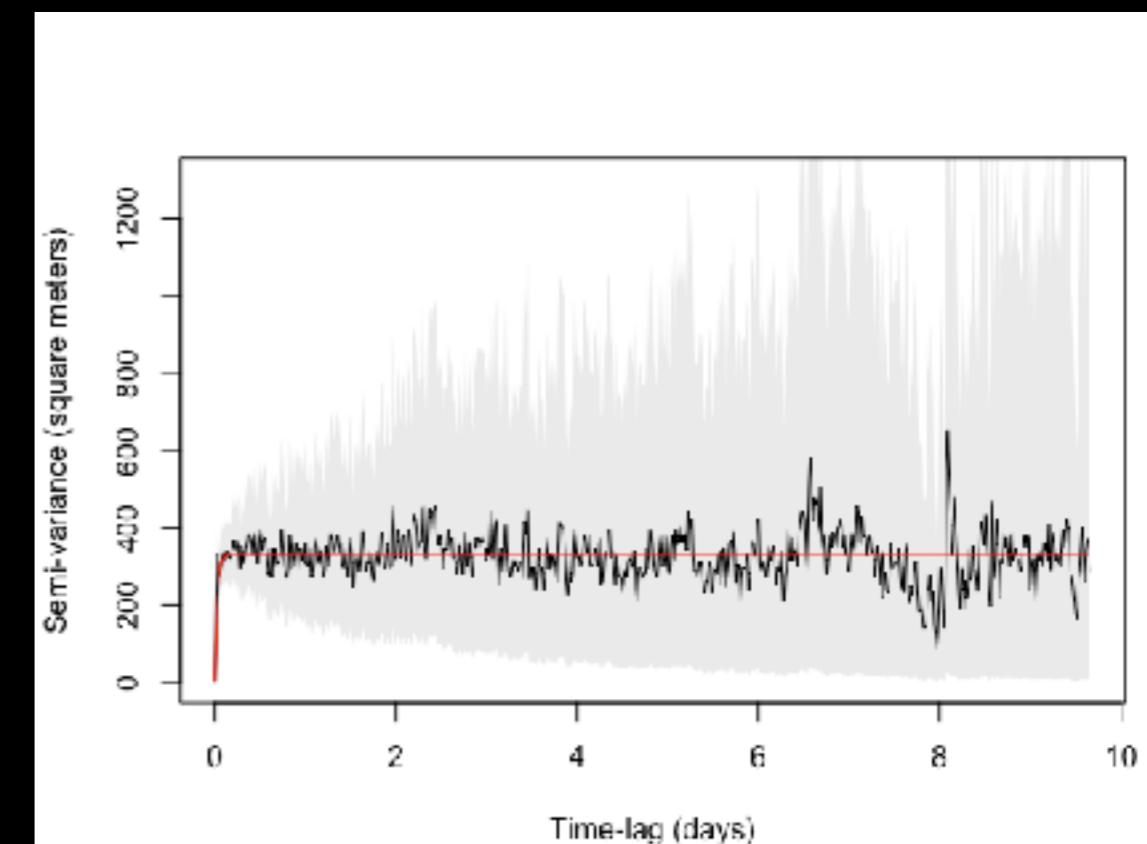
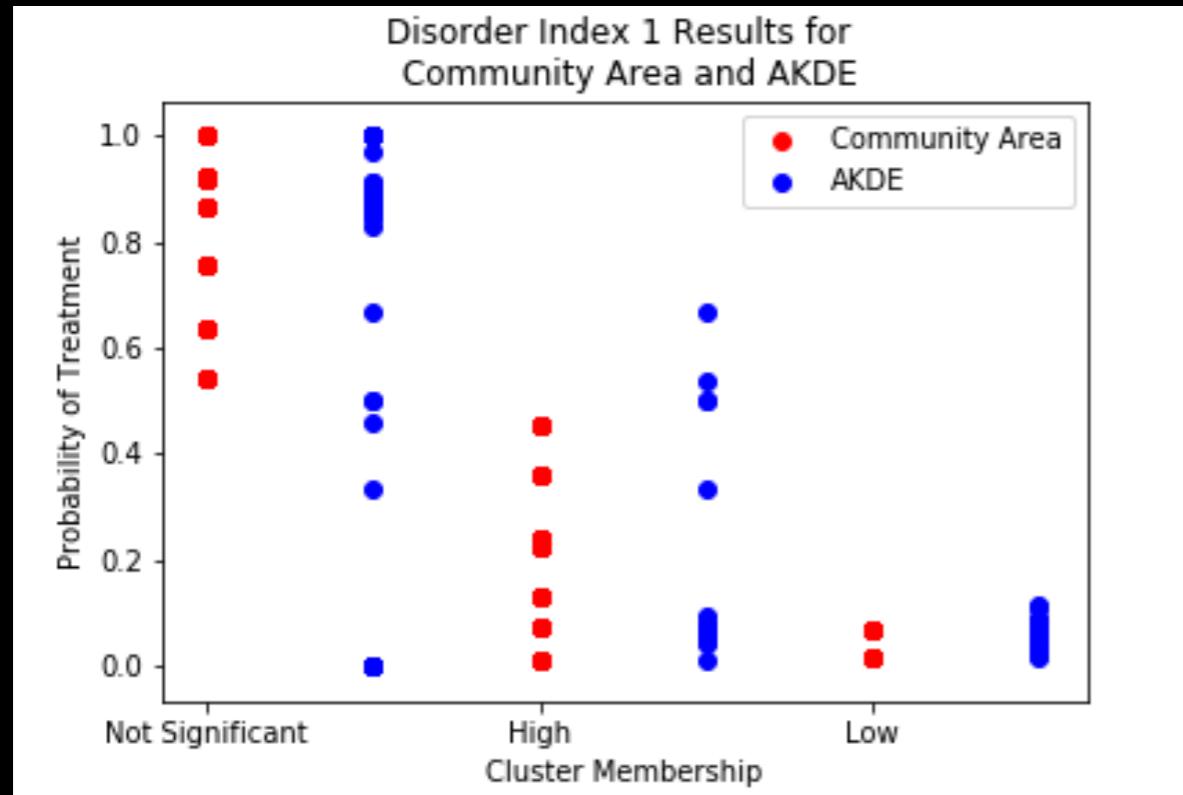
# H1: Supported

- Autocorrelated Kernel Density Estimation (Fleming et al. 2015)
  - Fit a set of GPS tracks to a random walk model
  - Generate a hypothesized home range from the fitted model



# H1: Supported

- Autocorrelated Kernel Density Estimation (Fleming et al. 2015)
  - Fit a set of GPS tracks to a random walk model
  - Generate a hypothesized home range from the fitted model



## H2: No Clear Finding (Yet)

- Hidden Markov Models to identify movement types (Whoriskey et al. 2017)
- Movement is an unobserved process, split into two types of movement (sedentary vs active)
  - Using turning angles, speed, bearing, contextual factors can specify probability of transitioning between types of movement

$$P(m_t | m_{t-1}) = f(x_1, x_2, \dots, x_k)$$

- When a study participant is in what mode should be driven by membership in groups and context (time of day, etc.) according to H2
  - No clear findings thus far

# Limitations, Next Steps

- HMM Methods are numerically unstable
  - Increase number of simulations, tweak parameters
- AKDE and HMM use random walk models built for ecological applications
- Data quality and extent issues with Place Pulse
  - Incorporate MapsCorps data on types of opportunities available in Chicago (e.g., restaurants, churches)

Nice work!

Who is up next?

# Assessing the Impact of Critical Mass in Chicago Public Schools

Kevin Sun

Thesis Advisor: Benjamin Soltoff

MA Computational Social Science

The University of Chicago

**Research Question:** What is the impact of teacher racial diversity on student educational outcomes?

# Key Definitions

**Research Question:** What is the impact of teacher racial diversity on student educational outcomes?

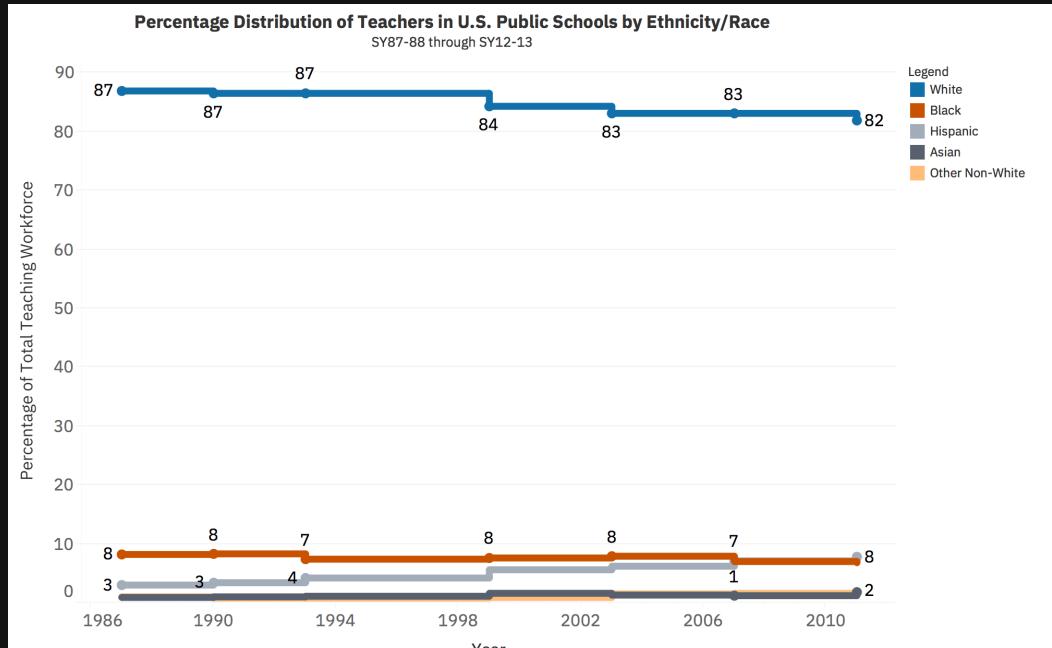
## Critical Mass

A minimum amount or threshold required to realize a certain outcome

## Student Outcomes

Measures of student achievement or success, e.g. test data (attainment or growth), graduation rates, **college persistence rates**

# Motivation



Data Source: National Center for Education Statistics

Racial makeup of public school teachers (K-12) has remained relatively stagnant

Public School students have been majority non-white since 2015.

# Existing Research

## Critical masses of ...

... women in corporate or political spheres (Kanter 1977; Dahlerup 1988); women in police departments (Meier et al. 2006)

... African-Americans in local elected offices (Meier et al. 1991; Goode and Baldwin 2005)

... people of color in managerial positions (Choi 2013)

## In Schools...

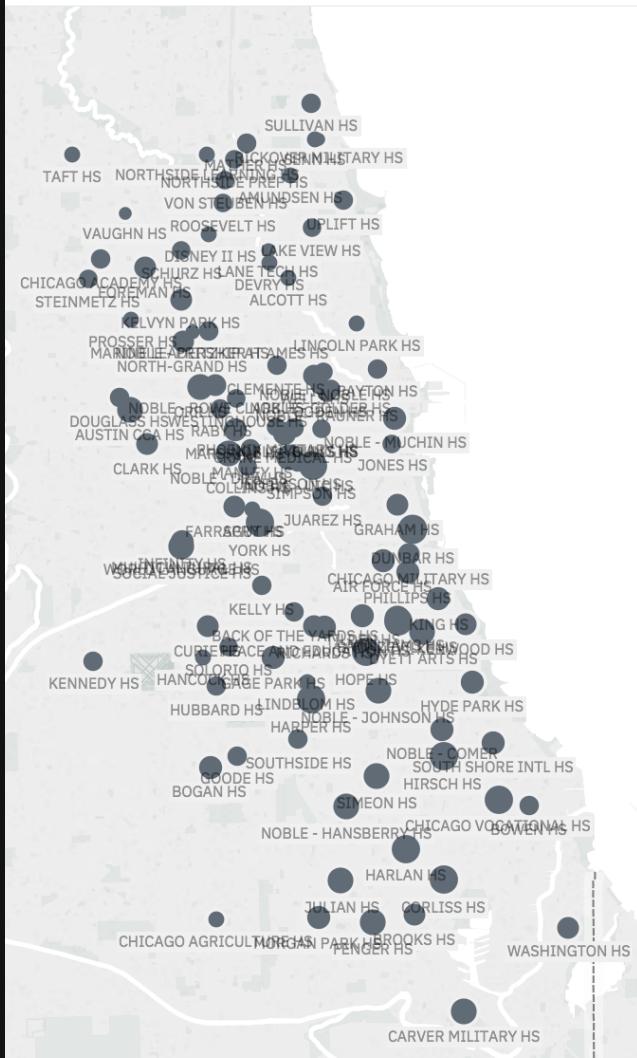
... latinx school administrators linked to improved outcomes to latinx students (Meier 1993)

... teachers of color expectations of students of color; role-model effect; culturally relevant curriculum (Grissom et al. 2015; Nicholson-Crotty 2011; Dee 2005; Cole 1986)

# Data

## Location of Schools in City of Chicago

Size of point indicates critical mass percentage



Data Source: City of Chicago

## This paper's contribution:

Examine the effects of critical mass at a school-level on longer-term student outcomes such as college retention rates

## Data Sources:

1. Teacher Demographics: Freedom of Information Act (FOIA) for Chicago Public High Schools (CPS); Requested from a charter school network
2. Student Outcomes Data: open data from CPS

107 Public and Charter High Schools

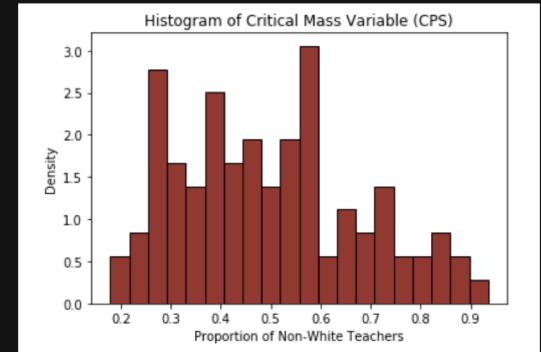
- 5,940 classroom teachers employed
  - ~50% self-identify as white
  - ~25% self-identify as black
  - ~16% self-identify as latinx
  - ~4% self-identify as Asian
- 85,743 students enrolled
  - CPS Students are 90% non-white

# Building the Model

**Dependent Variable:**  
College Persistence  
Rate of graduates from  
each high school

**Independent Variable:**  
The percentage of non-white teachers in a given high school (critical mass)

**Controls:**  
Student Race/Ethnicity  
Demographics; Free-Reduced Lunch Rates (FRL); English Language Learners (ELL), Special Education (SPED)



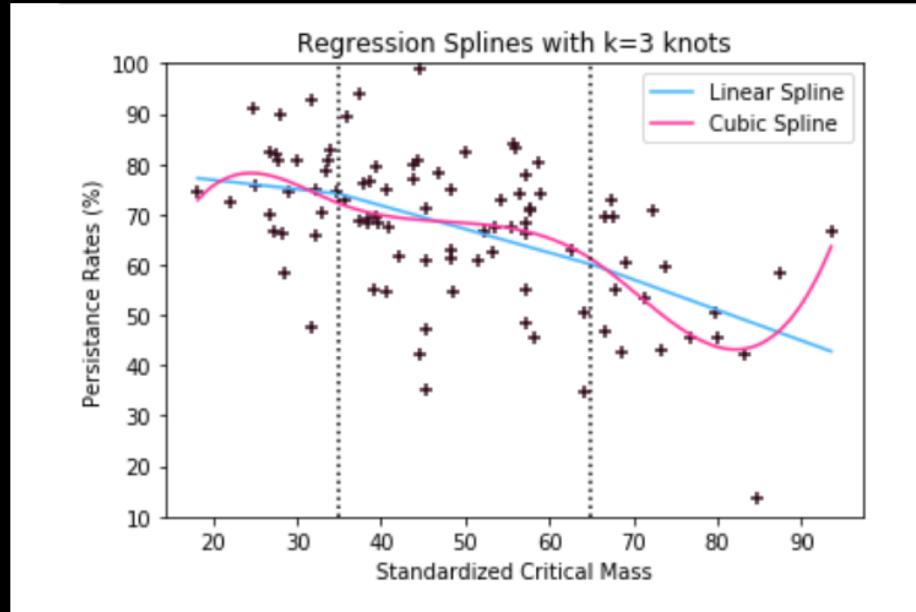
# OLS Regression

$$\begin{aligned} Persistence = & \beta_0 + \beta_1 CriticalMass + \beta_2 ELL + \beta_3 SPED + \beta_4 FRL \\ & + \beta_5 White + \beta_6 Black + \beta_7 Native \\ & + \beta_{8,2} Latinx + \beta_{9,2} MultiRace + \beta_{10,2} Asian + \beta_{11,2} HIPI + \varepsilon \end{aligned}$$

	Coefficient	Standard Error	p-value
<b>Intercept</b>	61.0935	86.162	0.480
<b>Critical mass</b>	-0.0972	0.090	0.285
<b>White</b>	0.2493	0.886	0.779
<b>Black</b>	0.7255	0.889	0.417
<b>Native</b>	2.6210	4.462	0.559
<b>Am/Alaskan</b>			
<b>Latinx</b>	0.8458	0.899	0.350
<b>Multiracial</b>	1.5121	2.129	0.480
<b>Asian</b>	1.35	1.045	0.217
<b>HI/PI</b>	-6.2264	5.330	0.246
<b>ELL</b>	-0.0872	0.173	0.617
<b>SPED</b>	-0.1950	0.169	0.252
<b>FRL</b>	-0.7199	0.206	0.001

# Regression Splines

$$Persistence = \begin{cases} \beta_{01} + \beta_{11} CriticalMass + \beta_{21} ELL + \beta_{31} SPED \\ \quad + \beta_{41} FRL + \beta_{51} White + \beta_{61} Black + \beta_{71} Native \\ \quad + \beta_{81} Latinx + \beta_{91} MultiRace + \beta_{10,1} Asian + \beta_{11,1} HIPI + \varepsilon & \text{if } CriticalMass \leq 35 \\ \\ \beta_{02} + \beta_{12} CriticalMass + \beta_{22} ELL + \beta_{32} SPED \\ \quad + \beta_{42} FRL + \beta_{52} White + \beta_{62} Black + \beta_{72} Native \\ \quad + \beta_{82} Latinx + \beta_{92} MultiRace + \beta_{10,2} Asian + \beta_{11,2} HIPI + \varepsilon & \text{if } 35 > CriticalMass < 65 \\ \\ \beta_{03} + \beta_{13} CriticalMass + \beta_{23} ELL + \beta_{33} SPED \\ \quad + \beta_{43} FRL + \beta_{53} White + \beta_{63} Black + \beta_{73} Native \\ \quad + \beta_{83} Latinx + \beta_{93} MultiRace + \beta_{10,3} Asian + \beta_{11,3} HIPI + \varepsilon & \text{if } 65 \leq CriticalMass \end{cases}$$



# Some Reflections

## Other Student Outcomes...

- PSAT/SAT Attainment/Growth
- Suspension/Expulsion Rates

## Also...

- Small  $n$
- Questionable reliability of college persistence data
- Non-representativeness of Chicago teaching workforce relative to nation & previous studies
- Racial Solidarity assumed (Fields and Fields 2015; Reed 2006)

## **Next Step: Multivariate Adaptive Regression Spline (MARS)**

$$f(X) = \beta_0 + \sum_{M=1}^M \beta_m h_m(X)$$
$$=$$
$$\hat{\beta}_{M+1} h_l(X) \times (X_j - t)_+ + \hat{\beta}_{M+2} h_l(X) \times (t - X_j)_+, h_l \in M$$

Nice work!

Who is up next?

# COMMUNITY OF CRYPTOCURRENCY

A study of the differences and similarities within and between online cryptocurrency communities using natural language processing techniques

Main research question: What are the social and cultural differences between the online communities of r/bitcoin and r/ethereum?

By: Bethany Bailey

Advisor: Karin Knorr Cetina

# RESEARCH QUESTIONS

- Two sub-questions:
  - Macro analysis that studies the difference in topics between the two communities:
    - Are the values of the founders and maintainers of individual coins obvious in the discussions in these coin communities? If so, how?
  - Micro analysis that looks at individual posts to see how individual affiliation and belief vary across and within cryptocurrency communities
    - Overall community
    - Individuals engaged in specific discussions

# DATA

- What?
  - Reddit post data from r/bitcoin and r/Ethereum
  - Collected using PSAW (wrapper for the PushShift API)
- When?
  - Part 1 (Topics): July 2015 – December 2018, split into two chunks
  - Part 2 (Affiliation): January 2014 – December 2018
- How much?

Subreddit	# Posts	# Tokens	# Characters	Avg Tokens Per Post
r/Bitcoin				
Q32015-Q42016	19128	2167896	10109673	113.34
Q12017-Q42018	80816	10074037	46434109	124.65
Q32015-Q42016	7709	930062	4367778	120.65
Q12017-Q42018	13138	2359031	11102099	179.56

# METHODS

- Part 1: How do topics differ between the communities? Is there a change over time?
  - Use topic modeling (Latent Dirichlet Allocation) in gensim library in python
  - Compare two time periods: pre-2017 and post-2017 (to look at effect of market activity)
- Part 2: How do individuals' affiliations with different online communities vary within the general cryptocurrency population and based on engagement in topics of discussion
  - Find posts that contain keywords about important topics
  - Using sentiment analysis in TextBlob (bag-of-words) and NLTK (naïve Bayes classifier) to find people with extreme sentiment towards these topics
  - Find affiliation with other subreddits (operationalized by posting on other subreddits during the same time period)

# RESEARCH PART I

- Topics classified as follows:

Time Period	Topic #	Bitcoin Topic	Ethereum Topic
Q32015-Q42016	1	Mining	Exchange/Platform
	2	Market	Transaction
	3	Transaction	Technology
	4	Market	Decentralization/Community
	5	Transaction	Platform
	6	Mining	Transaction
	7	Trading	Mining
	8	Technology	Security/Technology
	9	Transaction	Transaction
	10	General Bitcoin	Technology
Q12017-Q42018	1	Markets	Transaction
	2	Mining Community	Security/Technology
	3	Trading/Business	Exchange
	4	Time	Community
	5	Mining	Technology/Transaction
	6	Market	Transaction/Platform
	7	Mobile Exchange	Technology Communication
	8	Trading Transaction	Community/Conferences
	9	Market/Trading	Community
	10	Exchange/Transaction	Technology

- Found that Bitcoin topics were much more focused on trading and markets, whereas Ethereum topics were much more focused on technology
- Topics changed and diverged over time

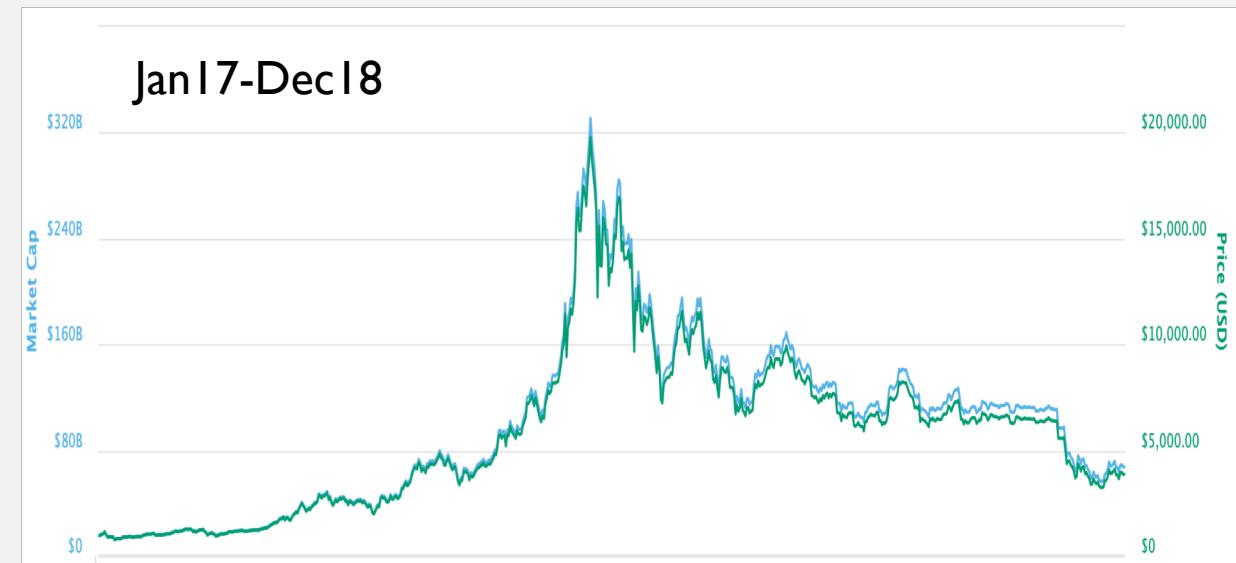
# DIFFERENCE IN BITCOIN FOCUS (MARKETS/TRADING) AND ETHEREUM FOCUS (TECHNOLOGY): WHY?

- Differences represent values of founders/maintainers:
  - Bitcoin: focus is on the coin and maintaining its tradability
  - Ethereum: means to realizing a technological platform – the founders “considered Ethereum to simply be fuel for the operation and build-out of the Ethereum platform” ([TheStreet](#), 2018)

Time Period	Topic #	Bitcoin Topic	Ethereum Topic
Q32015-Q42016	1	Mining	Exchange/Platform
	2	Market	Transaction
	3	Transaction	Technology
	4	Market	Decentralization/Community
	5	Transaction	Platform
	6	Mining	Transaction
	7	Trading	Mining
	8	Technology	Security/Technology
	9	Transaction	Transaction
	10	General Bitcoin	Technology
Q12017-Q42018	1	Markets	Transaction
	2	Mining Community	Security/Technology
	3	Trading/Business	Exchange
	4	Time	Community
	5	Mining	Technology/Transaction
	6	Market	Transaction/Platform
	7	Mobile Exchange	Technology Communication
	8	Trading Transaction	Community/Conferences
	9	Market/Trading	Community
	10	Exchange/Transaction	Technology

# CHANGE OVER TIME COINCIDE WITH MARKET VOLATILITY?

## BITCOIN PRICE



source: coinmarketcap

- Over time, Bitcoin topics became more market and exchange focused
- Maybe market focus in topics is related to greater volatility and growth Bitcoin
  - Bitcoin quadrupled from Oct-Dec 2017: \$4,300-\$19,900
  - Interestingly, ethereum only doubled in the same time period

## RESEARCH PART 2: MORE GRANULAR ANALYSIS OF AFFILIATION

- What are the affiliations of posters from r/Bitcoin and r/Ethereum, and how do they differ?
  - Affiliation operationalized by posting on another subreddit in the five year period studied.
- Interesting findings:
  - Ethereum members were generally more active in other cryptocurrency communities
  - Even though the involvement in highly popular subreddits was the same
  - Surprisingly, though topic modeling showed a difference in technological involvement, this analysis shows the same level across communities

Subreddit	% r/Bitcoin	% r/ethereum
AskReddit	20.6	19.68
CryptoCurrency	19.11	35.82
btc	11.98	19.06
funny	11.52	10.93
Showrethoughts	11.25	11.6
ethereum	9.01	0
pics	8.69	8.08
videos	8.3	8.35
gaming	8.26	7.99
explainlikeimfive	7.37	6.92
personalfinance	6.95	7.16
askscience	6.85	7.34
buildapc	6.82	7.26
litecoin	6.64	9.6
techsupport	6.27	6.49
aww	5.89	5.69
ethtrader	5.68	34.09
todayilearned	5.55	5.78
dogecoin	5.41	6.94
BitcoinMarkets	5.32	7.3
mildlyinteresting	5.17	4.6
pcmasterrace	5.14	5.02
legaladvice	4.82	4.55
The_Donald	4.77	4.92
Music	4.73	4.74
trees	4.63	3.36
news	4.38	4.31
Fitness	4.03	4.13
CircleofTrust	4.01	4.26
CryptoMarkets	4.0	9.35

## RESEARCH PART 2: MORE GRANULAR ANALYSIS OF AFFILIATION

- What are the affiliations of posters from r/Bitcoin and r/Ethereum, and how do they differ?
  - Affiliation operationalized by posting on another subreddit in the five year period studied.
- Interesting findings:
  - Ethereum members generally more active in other cryptocurrency communities
  - Even though the involvement in highly popular reditts was the same
  - Surprisingly, though topic modeling showed a difference in technological involvement, this analysis shows the same level across communities

Subreddit	% r/Bitcoin	% r/ethereum
AskReddit	20.6	19.68
CryptoCurrency	19.11	35.82
btc	11.98	19.06
funny	11.52	10.93
Showrethoughts	11.25	11.6
ethereum	9.01	0
pics	8.69	8.08
videos	8.3	8.35
gaming	8.26	7.99
explainlikeimfive	7.37	6.92
personalfinance	6.95	7.16
askscience	6.85	7.34
buildapc	6.82	7.26
litecoin	6.64	9.6
techsupport	6.27	6.49
aww	5.89	5.69
ethtrader	5.68	34.09
todayilearned	5.55	5.78
dogecoin	5.41	6.94
BitcoinMarkets	5.32	7.3
mildlyinteresting	5.17	4.6
pcmasterrace	5.14	5.02
legaladvice	4.82	4.55
The_Donald	4.77	4.92
Music	4.73	4.74
trees	4.63	3.36
news	4.38	4.31
Fitness	4.03	4.13
CircleofTrust	4.01	4.26
CryptoMarkets	4.0	9.35

# CONCLUSION

- The cryptocurrency community is not monolithic
  - LDA results show differences between the communities' values
  - Affiliation varies between the populations
- Applying this type of analysis to other coin communities might help researchers better understand this confusing subject
  - Classify coins into “types”
  - Better understand the level of diversity across similar communities
  - Understand what drives price movements in different communities
  - Understand what cryptocurrencies are – technology? currency? investment?

Nice work!

Who is up next?

# Endogenous Income Lottery, Risk Sharing and Negative Assortative Matching

Xinyu Cao<sup>1</sup>

CSS Workshop, Spring 2019

---

<sup>1</sup>Advisor: Rick Evans

# Motivation

- ① The analysis of matching pattern started with Beck (1973) and Shubik (1971). In general, those works rely on transferable utility (TU) assumption.
- ② Mazzocco (2004) analyzed a condition if ISHARA holds, then we can treat the matching utility as TU holds.

# Motivation

- ① The analysis of matching pattern started with Beck (1973) and Shubik (1971). In general, those works rely on transferable utility (TU) assumption.
- ② Mazzocco (2004) analyzed a condition if ISHARA holds, then we can treat the matching utility as TU holds.
- ③ Legros and Newman (2007) consider a risk sharing problem in which individual has different risk aversion (ordered by Arrow-Pratt measure) in which each couple's joint income can be either high or low.
- ④ Chiappori and Reny (2016) give a general treatment for income risk involving an arbitrary number of states of the world and for general risk averse preferences.
- ⑤ My paper extend Chiappori and Reny (2016) in a sense in their paper, the income risk is ex-ante identical for any men and women, my paper consider a situation where the income lottery is endogenous choosen.

- ① Agent: a group of men  $w \in \{1, \dots, n\}$  and a group of women  $m \in \{1, \dots, n\}$ .
- ② Income lottery:  $w_i$ , which will generate a pair of income lottery with a summary statistics  $(r(w_i), V[w_i])$
- ③ Variance and mean trade-off frontier of income lottery:

$$r(w_i) = f(V[w_i]) \quad (\text{A1})$$

such that the function  $f$  satisfies  $f'(x) > 0, f''(x) < 0$ .

- ④ Agent has a quadratic utility function

$$u(w) = (w - b)^2 \quad (\text{A2})$$

## Model 2

- ① Now Consider a two stage game, in first stage the men  $m$  and women  $w$  choose their desired income lottery  $w_m, w_w$
- ② In the second stage, men and women matched together.
- ③ Is the NAM stable matching results still holds?

## Unique NAM Stable Matching

For general stable matching mean variance lottery choose such that satisfies (A1), and a quadratic preference (A2). If we order the men and women by their risk preference (measured by Arrow - Pratt sense), There is a unique negative assortative stable matching.

## Exmaple 1

- ① Suppose that There are  $n$  men and women, each have a quadratic utility is given by

$$u(w_i) = aE[w_i] - b_i V[w_i] \quad (\text{A1}')$$

- ② in the first stage the agent is free to choose a variance of mean-variance lottery where the mean of the lottery is given by

$$r(w_i) = r_f + \beta V[w_i], \forall V[w_i] \in [0, 1] \quad (\text{A2}')$$

in this case we have

$$\begin{aligned} u(w_i) &= aE[w_i] - b_i V[w_i] \\ &= aE[r_f + \beta V[w_i]] - b_i V[w_i] \\ &= ar_f + (a\beta - b_i)V[w_i] \end{aligned}$$

## Example 2

From the above function we can see that the optimization is

$$U_{ij}(v_j) = \max_{r'_i, V'_i} u(r'_i, V'_i) = ar_f + (a\beta - b_i)V'_i$$

such that  $u(r_j, V_j) \geq v_j$

$$r_i = r_f + \beta V_i$$

$$r_j = r_f + \beta v_j$$

$$r_i + r_j = r'_i + r'_j$$

$$V_i + V_j = V'_i + V'_j$$

## Example 3

so this optimization is optimized if  $\max\{a\beta - b_i, a\beta - b_j\} > 0$ , then  $V_i = V_j = 1$ , and  $\max\{a\beta - b_i, a\beta - b_j\} < 0$ , then  $V_i = V_j = 0$ . Now, we shall consider the following three cases

- ①  $a\beta - b_i > a\beta - b_j > 0$ , the optimal is achieved when  
 $V_i = V_j = V'_i = V'_j = 1$
- ②  $a\beta - b_i > 0 > a\beta - b_j$ , the optimal is achieved when  
 $V_i = V_j = 1, V'_i = 2, V'_j = 0$
- ③  $0 > a\beta - b_i > a\beta - b_j$ , the optimal is achieved when  
 $V_i = V_j = V'_i = V'_j = 0$

## Example 4

From above derivation we can see that

$$U_{11}(v_1) > U_{12}(v_2) \Rightarrow U_{21}(v_1) > U_{22}(v_2)$$

then by theorem 1 of Chaipori and Reny (2016), NAM results holds. Intuitively, it's just the most risk averse men will be matched with the least risk averse women, because the least risk averse women are willing to pay risk at a higher price while the most risk averse men are willing to pay more.

Nice work!

Who is up next?

# The Impact of Transportation Infrastructure on Economic Growth: Evidence from China

Shuting Chen

Advisor: Dr. Richard Evans

## ● Motivation

- Developing countries have spent enormous amount of investment on transportation infrastructure
- Transportation infrastructure has been treated as a key to facilitating economic growth

- **Motivation - cont.**

- Limited empirical analysis of examining the impact of transportation infrastructure on economic growth at the sub-national level, especially for developing countries

- **Research Question**

- Exploring the impact/causality of transportation infrastructure on regional economic growth, based on economic outcomes for 178 non-metropolitan prefecture cities in China from 1997 to 2011
- Whether having better access to transportation serves as engines of possible economic growth
- Hypothesis: Holding everything else constant, cities having better access to transportation are more likely to experience higher economic growth

# Literature Review and Contributions

## • Literature Review

- Banerjee et al. (2012) and Faber (2009): study the impact of overall transportation or recently constructed highways in China on economic development; adopt “straight line” identification strategy to deal with the endogeneity of transportation networks
- Storeygard (2016): investigate the role of transportation in determining the income growth across sub-Saharan African cities using night lights data

## • Contributions

- Enrich empirical work by examining the impact of transportation on economic growth at the sub-national level
- Using both official economic data and night lights data to mitigate the effect of measurement errors

# Literature Review and Contributions

## • Literature Review

- Banerjee et al. (2012) and Faber (2009): study the impact of overall transportation or recently constructed highways in China on economic development; adopt “straight line” identification strategy to deal with the endogeneity of transportation networks
- Storeygard (2016): investigate the role of transportation in determining the income growth across sub-Saharan African cities using night lights data

## • Contributions

- Enrich empirical work by examining the impact of transportation on economic growth at the sub-national level
- Using both official economic data and night lights data to mitigate the effect of measurement errors

# Methods

- “Straight line” identification strategy
  - Draw a straight line from one provincial capital city to the nearest provincial capital city/Treaty Port
  - Compute the nearest geographic distance from each prefecture city to a constructed straight line - exogenous variation for access to transportation
- Estimation functions:

$$TI_{cpt} = \alpha \ln D_{cp} + \omega X_{ct} + \gamma_p + \delta_t + \varepsilon_{cpt} \quad (1)$$

$$\Delta \ln y_{cpt} = \beta \Delta \ln \hat{TI}_{cpt} + \omega \Delta X_{ct} + \Delta \delta_t + \Delta \varepsilon_{cpt} \quad (2)$$

$TI_{cpt}$ : transportation infrastructure of city  $c$  in province  $p$  in year  $t$

$D_{cp}$ : distance to the nearest straight line for city  $c$  in province  $p$

$y_{cpt}$ : economic outcome,  $X_{ct}$ : city-year fixed effects

$\gamma_p$ : province fixed effects,  $\delta_t$ : year fixed effects,  $\varepsilon_{cpt}$ : error term

- **Night Lights Data**

- Collected by U.S. Air Force Defense Meteorological Satellite Program (DMSP); became globally digital available in 1992
- Grid-based datasets: every 30 arc-second pixel has been labeled by a digital number (0 - 63) - intensity of lights
- 1997 - 2011: 26 satellite-year datasets, including 110 million pixels for 178 prefecture cities in China

- **Chinese Government Economic Data**

- Provincial Statistical Yearbooks: published annually by each province in China
- Using city-level data for 178 prefecture cities in 15 provinces: per capita GDP, population, land area, length of highway, length of railway

# Data - cont.

## • Spatial Data

- GIS maps based on 2010 China Prefecture Population Census Data
- “Straight line” identification strategy

Figure 5: Straight Lines and Transportation Infrastructure



# Data Analysis and Results

- Night Lights Data

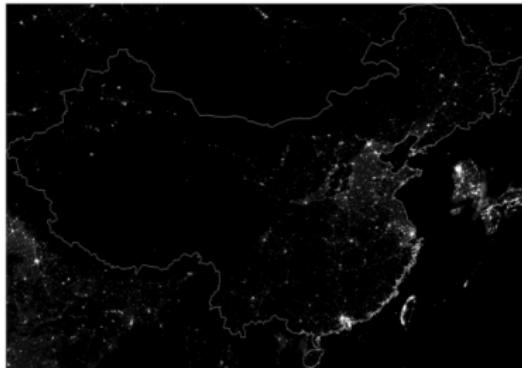


Figure 1: Lights at Night in China, 1997

Source: DMSP data collected by the United States Air Force.



Figure 2: Lights at Night in China, 2011

Source: DMSP data collected by the United States Air Force.

- Example of the extracted night lights data - Jiangsu Province, 1997

OBJECTID	Province	Prefecture	City	Num of Pixels	Total NTL	Prop Zero	DN - Min	DN - Mean	DN - Max
171	Jiangsu		Wuxi	6291	84038	0.089334	0	13.358449	59
172	Jiangsu		Xuzhou	15844	110300	0.158167	0	6.961626	61
163	Jiangsu		Changzhou	6008	59961	0.082723	0	9.980193	59
169	Jiangsu		Suzhou	11547	140323	0.136053	0	12.152334	58
167	Jiangsu		Nantong	12264	102877	0.01484	0	8.388536	55
165	Jiangsu		Lianyungang	10471	63890	0.159679	0	6.101614	59
164	Jiangsu		Huai'an	13946	62674	0.340528	0	4.494048	56
173	Jiangsu		Yancheng	21434	99440	0.23682	0	4.639358	57
174	Jiangsu		Yangzhou	9173	73695	0.119481	0	8.033904	58
175	Jiangsu		Zhenjiang	5256	52588	0.046613	0	10.005327	56
170	Jiangsu		Taizhou	7983	63939	0.048603	0	8.009395	56

# Results

- First stage:

		Dependent Variable: $TI_{cpt}$			
		Length of Highway		Length of Railway	
		(1)	(2)	(1)	(2)
In Dist to Line	-0.3115	-0.2422	-0.0159	-0.0017	
	(0.0803)	(0.0813)	(0.0037)	(0.0033)	
Land Area		0.0153		0.0024	
		(0.0034)		(0.0001)	
Obs	2135	2135	1411	1411	
Adj. $R^2$	0.065	0.075	0.769	0.819	

- Second stage:

		Dependent Variable: $\ln \hat{T}I_{cpt}$				
		In Per Capita GDP		In Night Lights		
		(1)	(2)	(1)	(2)	
In Fitted Highway <sub>cpt</sub>	0.2335	0.2335	0.4694	0.4323	In Fitted Railway <sub>cpt</sub>	
	(0.0243)	(0.0243)	(0.0329)	(0.0325)	0.1081	0.0690
Land Area		$4.95 * 10^{-7}$		$-3.63 * 10^{-6}$	Land Area	
		(5.08 * 10 <sup>-7</sup> )		(6.81 * 10 <sup>-7</sup> )		$-4.89 * 10^{-7}$
Obs	2135	2135	2135	2135	Obs	1409
Adj. $R^2$	0.098	0.097	0.254	0.263	Adj. $R^2$	0.099

# Results

- First stage:

		Dependent Variable: $\hat{T}I_{cpt}$			
		Length of Highway		Length of Railway	
		(1)	(2)	(1)	(2)
In Dist to Line	-0.3115	-0.2422	-0.0159	-0.0017	
	(0.0803)	(0.0813)	(0.0037)	(0.0033)	
Land Area		0.0153		0.0024	
		(0.0034)		(0.0001)	
Obs	2135	2135	1411	1411	
Adj. $R^2$	0.065	0.075	0.769	0.819	

- Second stage:

		Dependent Variable: $\ln \hat{T}I_{cpt}$				
		In Per Capita GDP		In Night Lights		
		(1)	(2)	(1)	(2)	
In Fitted Highway <sub>cpt</sub>	0.2335	0.2335	0.4694	0.4323	In Fitted Railway <sub>cpt</sub>	
	(0.0243)	(0.0243)	(0.0329)	(0.0325)	0.1081	0.0690
Land Area		$4.95 * 10^{-7}$		$-3.63 * 10^{-6}$	Land Area	
		(5.08 * 10 <sup>-7</sup> )		(6.81 * 10 <sup>-7</sup> )		$-4.89 * 10^{-7}$
Obs	2135	2135	2135	2135	Obs	1409
Adj. $R^2$	0.098	0.097	0.254	0.263	Adj. $R^2$	0.099

Nice work!

Who is up next?

# PATHWAYS OF GENTRIFICATION: ANALYSIS OF GENTRIFIERS IN RACIAL DIMENSION, A COMPARATIVE STUDY OF GENTRIFICATION TRANSITION IN THREE U.S. CITIES

Advisor: Kevin Credit  
Jie Heng

# RESEARCH QUESTIONS:

What are the pathways of neighborhood gentrification since the 1990s in three cities?

For different racial groups, do demographic characteristics matter in their neighborhood selection?

# PREVIOUS STUDIES

- “White-only” gentrification: Gentrification is often described as a “racial turnover” (Wilson, 1992, p.123), stressing on the White affluent gentrifiers and the displacement of poor Black people.
- Since the 1990s, the minorities’ influence in gentrification has increased (Hwang, 2016) and the traditional prejudice towards Blacks are softened (Ellen, 2000)
  - Asians are the pioneers of gentrification in Seattle (Hwang, 2016)

# DATA AND METHODS

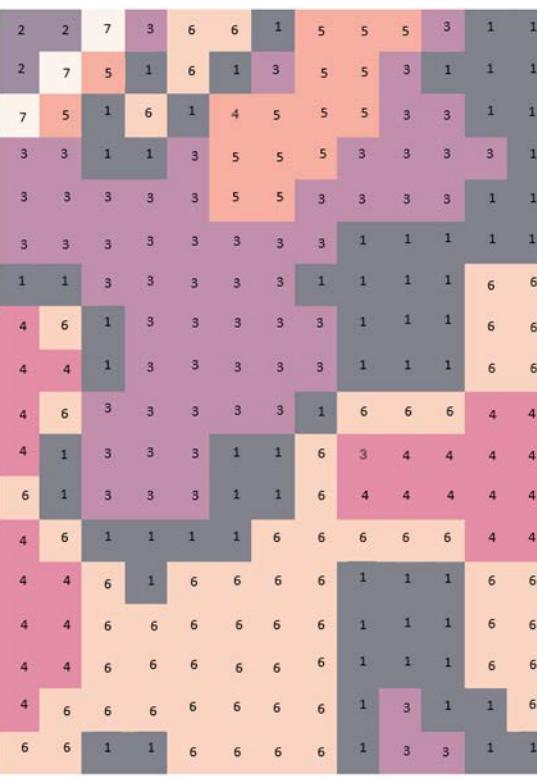
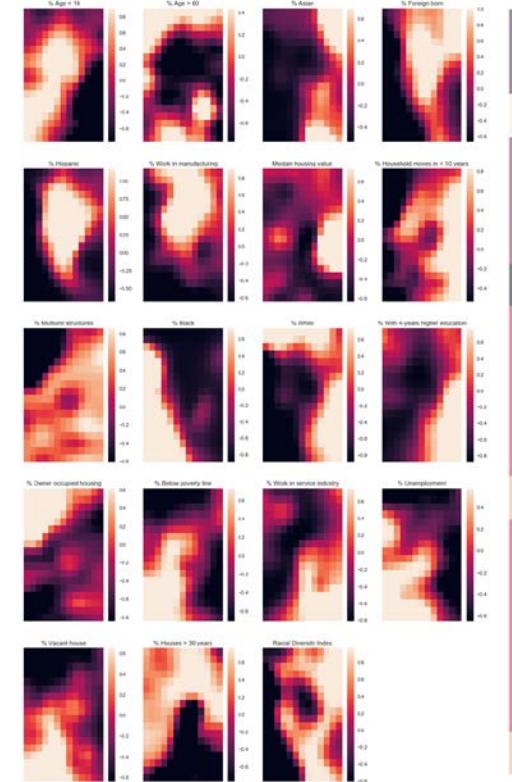
- Census Tract 1990, 2000, 2010
- Calculate racial diversity rate
- Monetary values are inflation adjusted and expressed in 2015 dollars
- Crosswalk files provided by Longitudinal Tract Data Base (LTDB)

Variables	Mean			SD		
	Chicago	Seattle	NYC	Chicago	Seattle	NYC
<b>Demographic</b>						
% Persons' age under 18	24.90	16.44	23.56	9.26	7.14	6.81
% Persons' age above 60	14.90	16.84	17.21	7.02	5.91	6.84
% non-Hispanic White	33.71	69.06	38.36	32.73	22.71	33.38
% non-Hispanic Black	38.08	9.50	26.39	42.25	11.62	32.43
% Asian	4.39	14.29	10.62	8.04	12.57	13.43
% Hispanic	23.23	5.36	23.28	27.87	4.04	20.99
% Foreign born	18.34	15.94	35.08	16.53	10.59	16.36
Racial Diversity Rate	0.32	0.41	0.44	0.22	0.18	0.19
<b>Socioeconomic</b>						
% Persons with at least 4-year college degree	10.43	20.47	9.42	10.09	9.37	6.90
% Unemployed	5.84	3.41	4.48	3.42	1.70	2.38
% Work in manufacturing industry	13.69	9.94	7.56	9.14	4.96	5.37
% Work in service industry	47.41	53.03	27.87	15.47	14.35	15.95
% Below poverty level	21.42	12.35	17.34	14.98	9.09	12.18
<b>Housing</b>						
% Multiunit structures	68.84	40.64	72.71	26.64	26.97	25.25
Median home value	162,047.08	218,634.06	260,668.07	185,240.68	182,926.42	204,932.73
% Structures built more than 30 years ago	82.46	70.88	81.61	15.66	14.75	18.45
% Vacant housing	10.07	5.38	5.98	6.92	3.02	4.04
% Owner occupied housing	39.90	49.67	36.26	21.57	20.40	22.36
% Household moves into a unit less than 10 years ago	57.93	65.46	53.59	14.11	10.51	10.65

# DATA AND METHODS

Delmelle (2017)

# 1. Classifying Neighborhoods

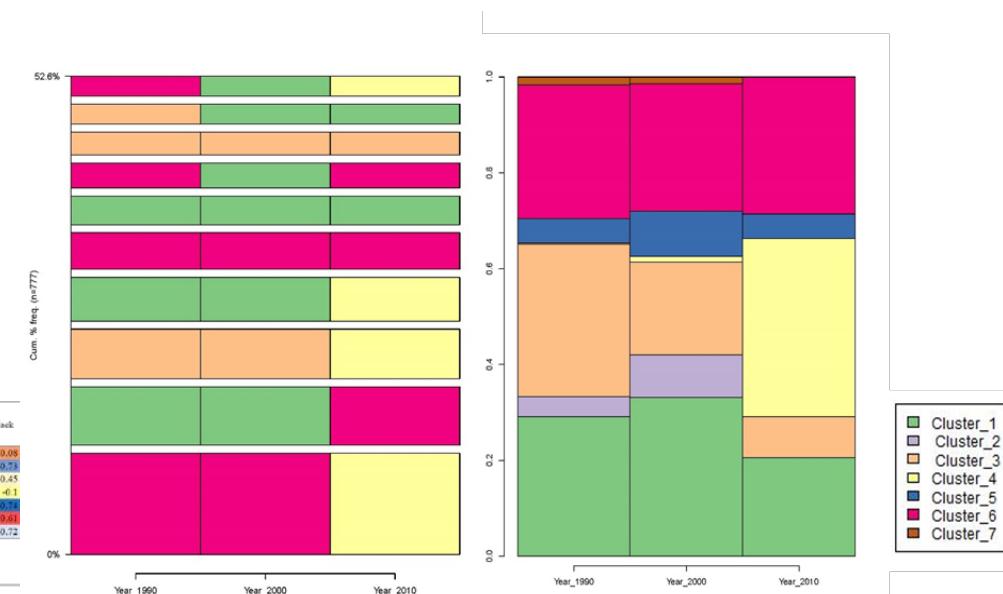


## Cluster

- 1 Young, blue collar, poor
  2. Employed, few vacant housings, adult, American, wealthy, White
  3. Old people, low percent of people working in service industry, some blue collar, fairly high percent of Asians
  4. Owner occupied, foreign born, Hispanic, poor, low percent of new movers and multi-housing, service industry, unemployed, old house
  5. Asian, few Black, racial diverse
  6. Black singularity, vacant house, less educated, low Asian
  7. White, educated, renters, new in-movers, multi-structured housing, adults, low percent of Hispanic

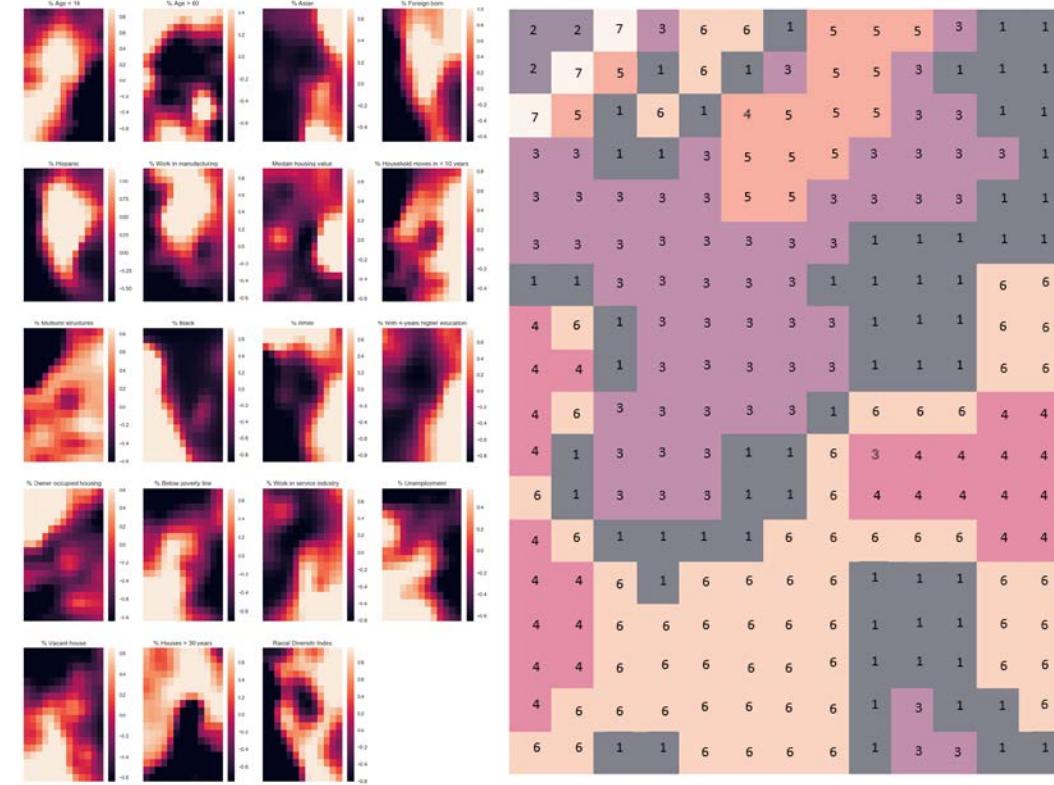
	Cluster	% Age < 18	% Age > 60	% Asian	% Foreign	% Hispanic	Work in manufacturing	Median housing value	% Household income > \$100k	% Median evictions	% Black
1	0	0.46	-0.18	0.19	0.06	0.1	0.18	-0.15	0.11	0.21	0.09
2	1	-1.57	-0.77	0.63	-0.52	-0.47	-0.53	0.35	0.18	0.63	-0.79
3	2	-0.43	0.66	0.24	0.28	-0.03	0.07	0.15	0.11	0	-0.44
4	3	0.24	0.15	-0.23	0.54	0.71	-0.12	-0.06	-0.01	-0.14	0
5	4	-0.17	0.01	0.41	-0.02	-0.03	0.03	0.59	0.82	0.41	0.01
6	5	0.35	-0.11	-0.05	-0.15	-0.11	0.03	-0.33	-0.07	-0.11	0.66
7	6	-0.49	-0.59	0.69	-0.39	-0.24	-0.26	1.11	1.41	0.9	-0.77
	Cluster	% White	% High school or less	% Owner occupied houses	% Below poverty line	% Work in service industry	Unemployment rate	% Vacant houses	Rental vacancies > 30 years	Racial Diversity	Median income
1	0	-0.29	-0.23	-0.21	0.04	-0.16	0.04	0.19	-0.02	0.07	0.07
2	1	1.33	1.99	0.31	-0.6	-0.27	-0.11	-0.56	0.56	0.16	0.28
3	2	0.77	0.11	0.19	-0.31	-0.31	-0.49	-0.3	-0.52	-0.26	0.28
4	3	-0.41	-0.27	0.48	-0.06	0.72	0.56	0.06	0.66	-0.11	0.11
5	4	0.88	1.09	-0.22	-0.63	0.3	-0.76	-0.42	0.09	0.79	0.08
6	5	-0.06	-0.35	0	0.26	0	0.39	0.2	0.18	-0.01	0.01
7	6	1.35	2.01	-0.55	-0.06	0.18	-0.17	-0.49	-0.34	0.15	0.15

## 2. Sequence and Cluster Analysis



# DATA AND METHODS

## 1. Classifying Neighborhoods

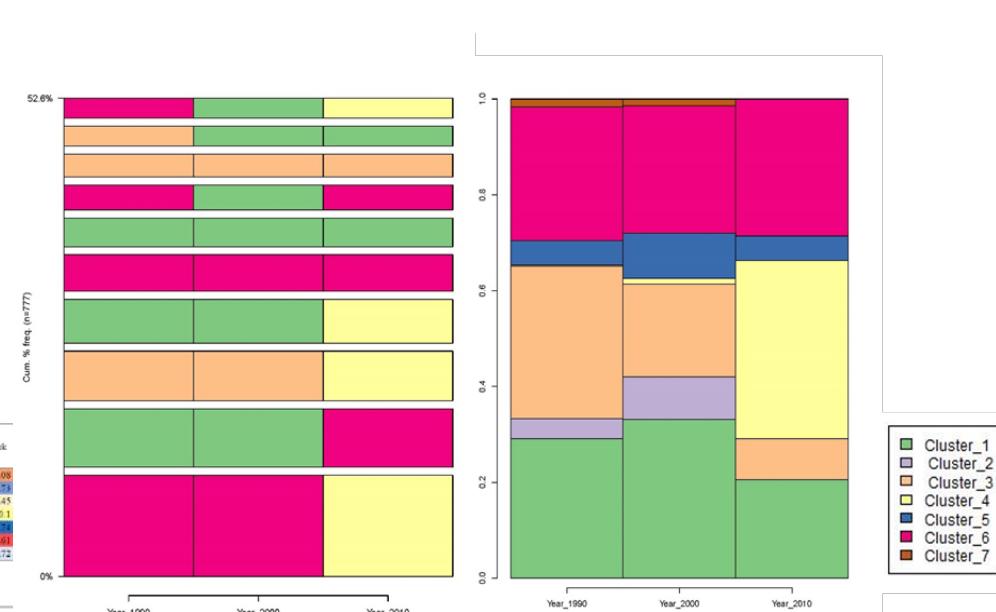


### Cluster:

- Young, blue collar, poor
- Employed, few vacant housings, adult, American, wealthy, White
- Old people, low percent of people working in service industry, some blue collar, fairly high percent of Asians
- Owner occupied, foreign born, Hispanic, poor, low percent of new movers and multi-housing, service industry, unemployed, old house
- Asian, few Black, racial diverse
- Black singularity, vacant house, less educated, low Asian
- White, educated, renters, new in-movers, multi-structured housing, adults, low percent of Hispanic

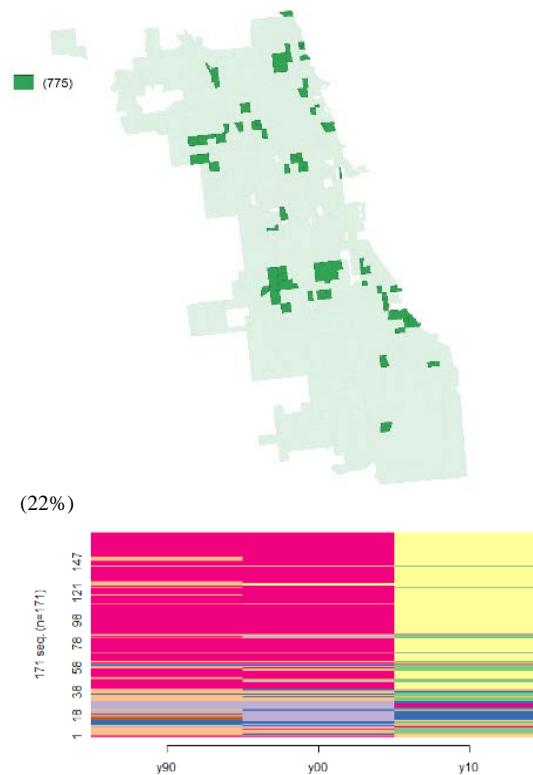
Cluster	% Age < 18	% Age > 60	% Asian	% Foreign born	% Hispanic	Work in manufacturing	Median housing value	Household moves in 10 years	% Multifamily structures	% Black
0	-0.46	-0.18	0.16	0.06	0.1	0.18	-0.15	0.11	0.21	0.08
1	-1.57	-0.77	0.03	-0.42	-0.47	-0.65	0.25	1.38	0.62	-0.73
2	-0.43	0.66	0.24	0.23	-0.3	0.07	0.15	0.11	0	-0.45
3	0.24	0.15	-0.23	0.34	0.75	-0.16	0.05	0.61	0.16	-0.1
4	-0.29	-0.40	0.41	0.2	-0.05	-0.35	0.59	0.82	0.47	-0.1
5	0.35	-0.11	-0.62	-0.55	-0.11	0.03	0.25	0.27	-0.11	0.61
6	-0.49	-0.59	-0.69	-0.39	-0.72	-0.56	1.11	1.41	0.3	-0.72
Cluster	% White > 6 years higher education	% Owner occupied housing	% Below poverty line	% Work in service industry	% Unemployment	% Vacant houses	% Houses > 30 years	Racial Diversity Index		
0	-0.23	-0.25	-0.24	0.3	-0.16	0.04	0.19	-0.02	0.07	
1	1.33	1.98	-0.33	0.8	0.27	0.04	0.19	0.16		
2	0.79	0.17	0.19	-0.4	-0.4	-0.49	-0.3	-0.52	0.25	
3	-0.01	-0.37	-0.56	-0.06	-0.72	0.64	0.06	-0.69		
4	0.88	1.09	-0.22	-0.63	0.2	-0.76	-0.42	0.69	0.78	
5	-0.6	-0.33	0	0.26	0	0.39	0.7	0.15	-0.41	
6	1.35	2.01	-0.35	-0.82	0.18	-0.67	-0.49	0.81	0.15	

## 2. Sequence and Cluster Analysis

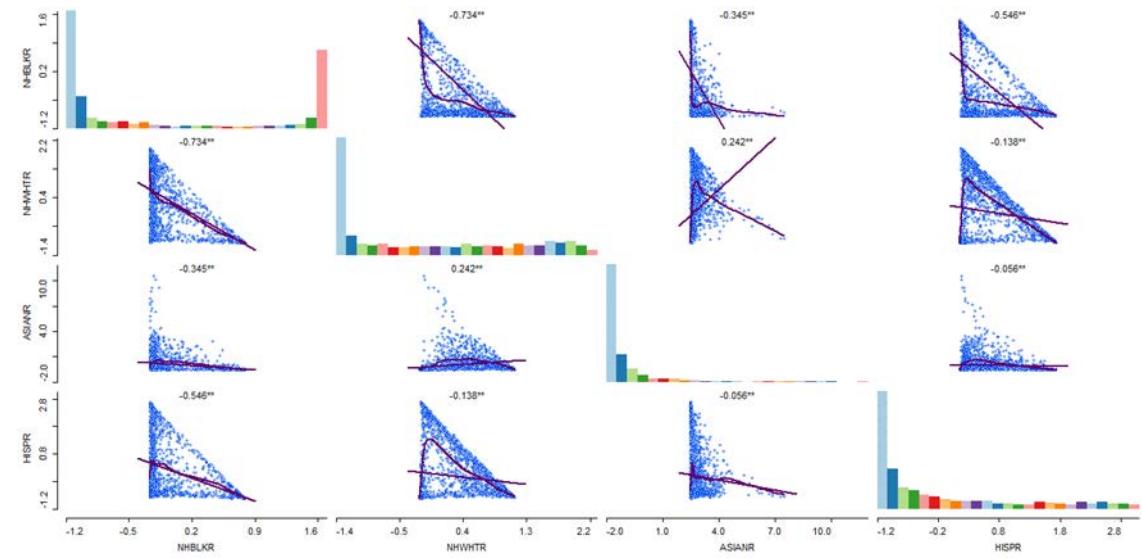


# DATA AND METHODS

## 3. Classifying Sequence and Mapping Tractrices of Neighborhoods



## 4. Associations among racial groups



# RESULTS

- Gentrification
- Marginal Gentrification
- Incumbent Gentrification
- Upgrading
- Incumbent Upgrading

## Chicago

- Group 1: Diversification, White and Asian Gentrification
- Group 2: Hispanic Older Stable Suburban
- Group 3: Prosperous Downtown and Incumbent Upgrading Neighborhoods
- Group 4: White Flight
- Group 5: White and Asian Flight and Struggling Black and Hispanic Neighborhoods
- Group 6: Black and Hispanic Older Stable Suburban
- Group 7: Struggling Black and Hispanic Neighborhoods

## Seattle

- Group 1: Decline in Socioeconomic Status
- Group 2: White Flight, Diversification and Struggling Neighborhoods
- Group 3: Mixed-raced Gentrification
- Group 4: White Flight and Decline in Socioeconomic Status
- Group 5: Diversification and Stable Prosperous
- Group 6: Incumbent Upgrading Mixed-race Neighborhoods
- Group 7: Upgrading Mixed-race Neighborhoods

# RESULTS

- Gentrification
- Marginal Gentrification
- Incumbent Gentrification
- Upgrading
- Incumbent Upgrading

## Chicago

- Group 1: Diversification, White and Asian Gentrification
- Group 2: Hispanic Older Stable Suburban
- Group 3: Prosperous Downtown and Incumbent Upgrading Neighborhoods
- Group 4: White Flight
- Group 5: White and Asian Flight and Struggling Black and Hispanic Neighborhoods
- Group 6: Black and Hispanic Older Stable Suburban
- Group 7: Struggling Black and Hispanic Neighborhoods

## Seattle

- Group 1: Decline in Socioeconomic Status
- Group 2: White Flight, Diversification and Struggling Neighborhoods
- Group 3: Mixed-raced Gentrification
- Group 4: White Flight and Decline in Socioeconomic Status
- Group 5: Diversification and Stable Prosperous
- Group 6: Incumbent Upgrading Mixed-race Neighborhoods
- Group 7: Upgrading Mixed-race Neighborhoods

# RESULTS

## Seattle

- Group 1: Decline in Socioeconomic Status
- Group 2: White Flight, Diversification and Struggling Neighborhoods
- Group 3: Mixed-raced Gentrification
- Group 4: White Flight and Decline in Socioeconomic Status
- Group 5: Diversification and Stable Prosperous
- Group 6: Incumbent Upgrading Mixed-race Neighborhoods
- Group 7: Upgrading Mixed-race Neighborhoods

- Gentrification
- Marginal Gentrification
- Incumbent Gentrification
- Upgrading
- Incumbent Upgrading

## New York City

- Group 1: White Upgrading
- Group 2: White Gentrification
- Group 3: Black Flight and Decline in Socioeconomic Status
- Group 4: White Gentrification and Young Urban
- Group 5: Decline in Socioeconomic Status
- Group 6: Stable Struggling Black and Hispanic Neighborhoods
- Group 7: Diversification and Children
- Group 8: White Incumbent Upgrading
- Group 9: Incumbent Upgrading and Diversification
- Group 10: Decline in Socioeconomic Status, Diversification and Older People
- Group 11: Black and Hispanic Neighborhoods and Decline in Socioeconomic Status
- Group 12: Decline in Socioeconomic Status and Diversification

Nice work!

Who is up next?

# Estimating Propensity Scores and Testing Overlap using Support Vector Machines

Ari Boyarsky

Advisor: Prof. Alex Torgovitsky (Department of Economics)

## ► Motivation

- Matching methods are common across the social sciences. They are powerful but require strong assumptions, in particular that,

$$\mathbb{E}[Y_0|D = 1, X] = \mathbb{E}[Y_0|D = 0, X] \quad (1)$$

Heckman et al. (1998) and region of common support,

$$S = \text{Supp}(X|D = 1) \cap \text{Supp}(X|D = 0) \quad (2)$$

- Propensity score matching attempts to simplify this. Using  $p(X) := P[D = 1|X]$  (usually with logistic regression). In this case, overlap requires that,

$$p(X) \in (0, 1) \quad (3)$$

Such that treatment outcome is not perfectly predicted.

- Heinrich et al. (2010), Harder et al. (2011), Imbens (2014)
- But, these methods may misidentify overlap (i.e., a nonlinear boundary). This paper provides a test to resolve this issue.

## Proposed Methodology

- If we want to be sure that we have common support we should employ a methodology that maximizes predictive power.
- Using Support Vector Machines (SVM) with radial basis function kernels allows us to find nonlinear boundaries in a high dimensional space and assess the fit using standard machine learning model estimates. Primal problem,

$$\min_w \quad \frac{1}{2} \langle w, w \rangle + C \sum_{i=1}^N \xi_i \quad (4a)$$

$$\text{s.t.} \quad y_i(\langle w, \phi(x_i) \rangle + b) \geq 1 - \xi_i \quad (4b)$$

$$\xi_i \geq 0 \quad \forall i = 1, \dots, N \quad (4c)$$

$$K(u, v) = \langle \phi(u), \phi(v) \rangle = e^{-\gamma \|u-v\|_2^2} \quad (5)$$

- K-fold cross validation error to evaluate fit.

## Proposed Methodology

- We can also use Platt (1999) transforms to calculate  $p(X)$ ,

$$P(y = 1|f_i) = p(x_i) = \frac{1}{1 + \exp(Af(x_i) + B)} \quad (6)$$

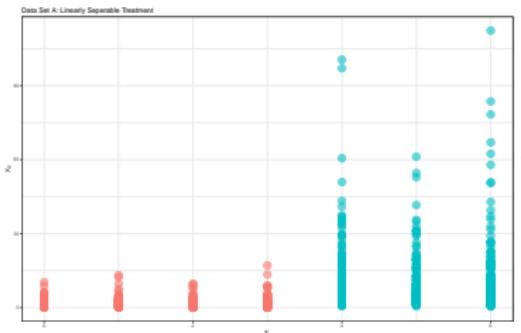
$A$  and  $B$  are fit with maximum likelihood estimation using the initial data set,

$$\min_{A,B} - \sum_i^N \frac{y_i + 1}{2} \log(p(x_i)) + (1 - \frac{y_i + 1}{2}) \log(1 - p(x_i)) \quad (7)$$

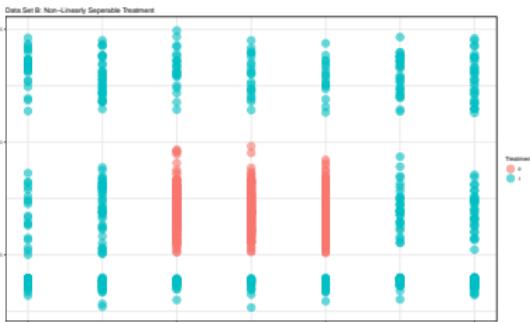
- We should pay particular attention to observations close to the separating hyperplane i.e.  $p(X) \in [0.4, 0.6]$ .
  - Trade off between within sample accuracy and generality.
- If we have an imbalanced data set we can use weights to adjust for this (like in LaLonde (1986)).

# Simulations of Test Data

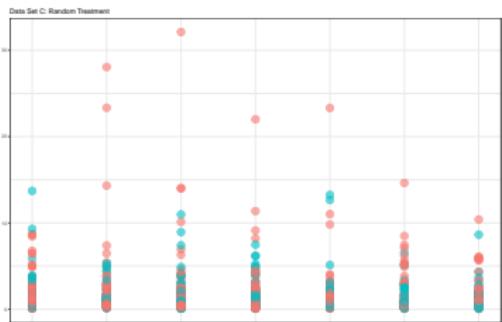
(a) Convex Separation



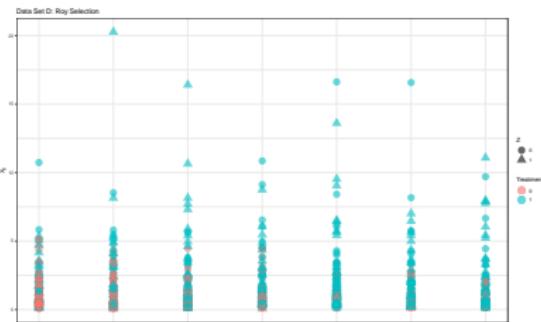
(b) Nonlinear Separation



(c) Random Treatment



(d) Roy Selection



## Simulational Results: SVM

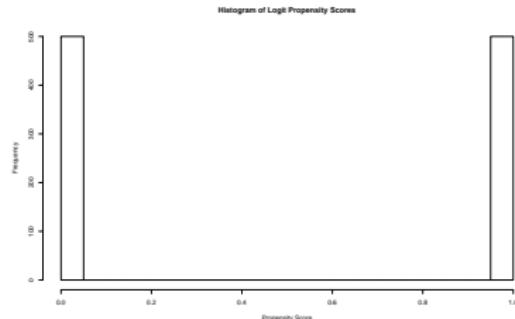
Table: SVM Overlap Tests of Test Data Sets (A)-(D)

Data Set	(A)	(B)	(C)	(D)
% Fixed Cross Validation Error	0.210	0.196	33.133	16.190
$K = 10$ CV Accuracy Rate	99.6	100	56.6	74.5
Brier Score	0.0	0.0	0.48	0.17
% Margin (0.1 Width)	0	0	0	5.5
% Margin (0.2 Width)	0	0	80.7	14.5
Kernel	Radial	Radial	Radial	Radial
Gamma	1	2	1	0.01
Cost	32	256	4	4
N	1,000	1,000	1,000	1,000

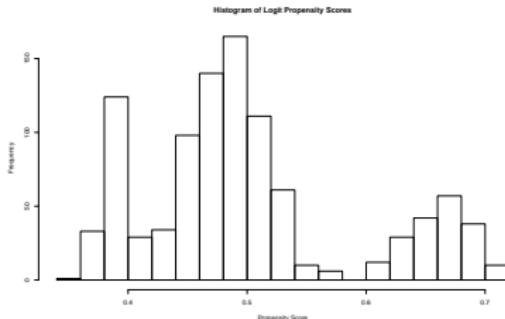
*Note:* Fixed Cross Validation Error is computed by cross validating the entire data set against gamma values  $\{0.01, 0.1, 0.5, 1, 1.5, 2\}$  and cost parameters  $\{4, 8, 16, 32, 64, 128, 256, 512, 1024\}$  using K-fold cross validation error with a fixed test and validation sample.

# Simulational Results: Logistic Regression

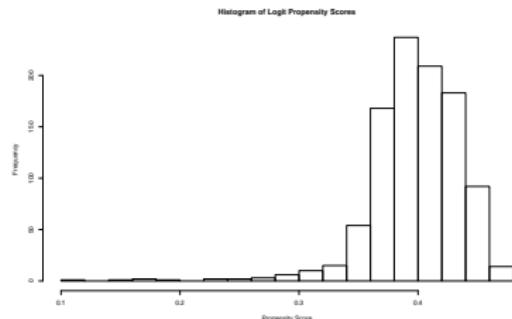
(a) Convex Separation



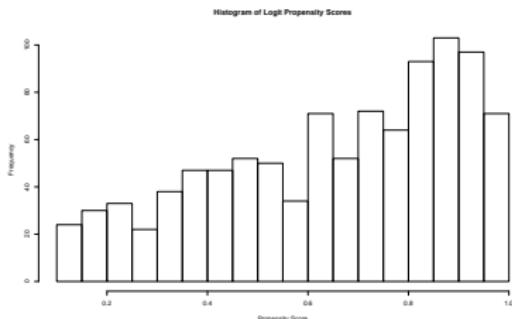
(b) Nonlinear Separation



(c) Random Treatment

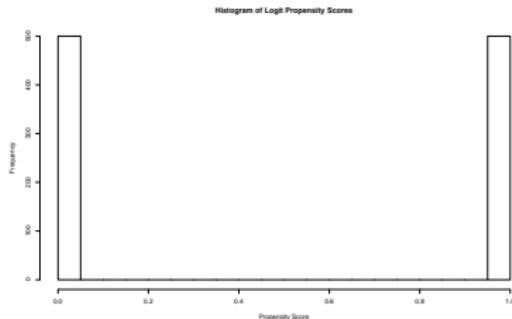


(d) Roy Selection

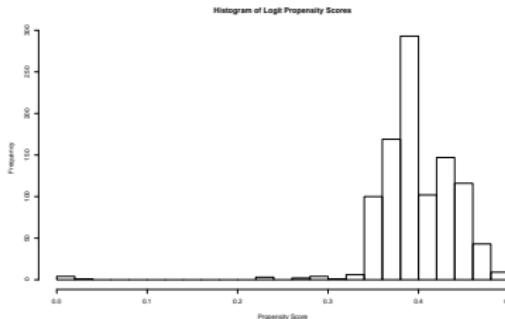


# Simulational Results: Logistic Regression (Higher Order Terms)

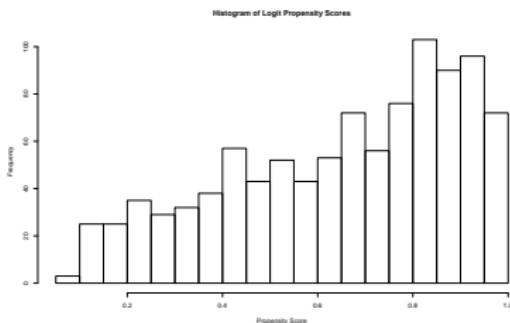
(b) Nonlinear Separation



(c) Random Treatment



(d) Roy Selection



## Application: LaLonde (1986) Data

- ▶ Dehejia and Wahba (1999) – Issues with sample and variable selection (Smith and Todd (2005))
- ▶ Could there also be issues with overlap?
  - ▶ DW (1999): Compute  $p(X)$ , assess overlap/trim data to increase overlap, match.
  - ▶ We follow the same procedure but use an SVM.
  - ▶ After trimming such that  $p(X) \in (0.05, 1)$  we have only 178 observations, the same specification with logistic regression yields 561 observations.
  - ▶ The data set is unbalanced 118 treated, only 49 control, 146 observations with propensity scores in 0.7 – 0.72 is suggestive of overlap.
  - ▶ However, we should weight data to address imbalance. So,

$$G(x) = \text{sgn}[wf(x) + b] \quad (8)$$

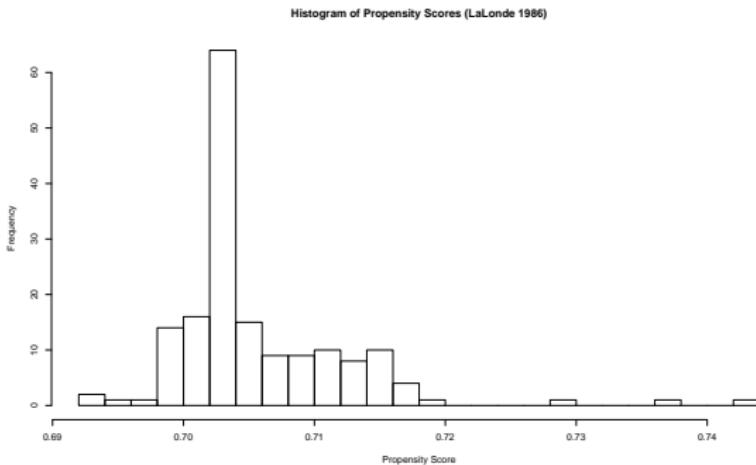
Where  $w$  is vector of inverse frequency weights ( $k = 2$  classes),

$$w_j = \frac{n}{kn_j} \quad (9)$$

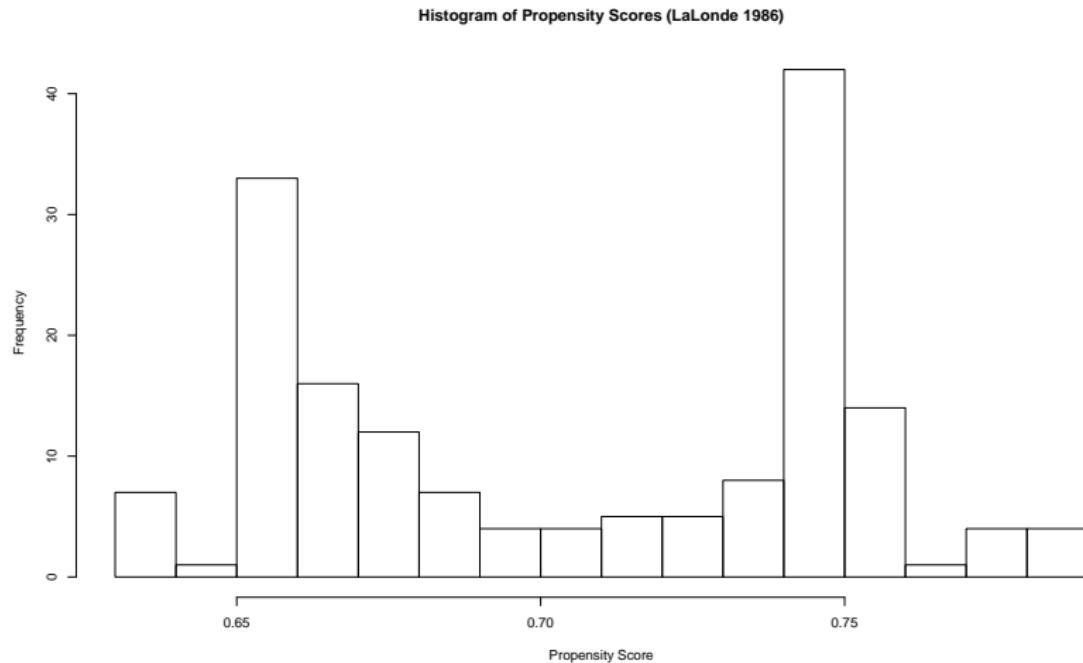
# LaLonde (1986) SVM Results (before weighting)

	(1)
% Cross Validation Error	0.1984
% Margin (0.1 Width)	0
% Margin (0.2 Width)	0
Kernel	Radial
Gamma	0.5
Cost	4
N	178

*Note:* Training Cross Validation Error is computed by cross validating the entire data set against gamma values  $\{0.01, 0.1, 0.5, 1\}$  and cost parameters  $\{4, 8, 16, 32, 64, 128, 256\}$  using the K-fold cross validation such that we split the data into K subsets, train the SVM on every other subset, and compute the average test error for each subset. This specification predicts D using: age, education, no degree, black, married, Hispanic, RE74, RE75, u74, u75, education\*RE74.



# LaLonde (1986) SVM Results (after weighting)



*Note:* Applying inverse frequency weighting yields a control weight of 6.816, and a treated weight of 2.831.

So there is overlap, but in a much smaller support than considered by DW (1999). ATT of about 1643.908 matching Imbens (2014).

Nice work!

Who is up next?

# **ONCE UPON A TIME IN CALIFORNIA**

**TITLE OF THESIS: "HOW MUCH SHOULD YOU FINE PEOPLE FOR WATERING THEIR LAWNS?"**

**RUIXUE LI**

**ADVISOR: LUDOVICA GAZZE**

**18 APR 2019**

# **BACKGROUND**

- California has been having water issues
- So one city made it illegal for people to use water...
- ...during certain times of day and on certain days
- And they recently started giving out fines automatically based on electronic water meter data



# **BACKGROUND**

- California has been having water issues
- So one city made it illegal for people to use water...
- ...during certain times of day and on certain days
- And they recently started giving out fines automatically based on electronic water meter data



# **BACKGROUND**

- **Urban labs partnered with the city to run a RCT to implement and evaluate the the program**
- **Research question: what's the best way to implement the policy?**

# **EXPERIMENT**

- Over 100,000 households
- 12 groups
- 9 treatment groups: 3 different amounts of fine, 3 different levels of threshold
- 3 control groups: receive visual inspection, 3 different amounts of fine
- July to September, 2018

# **EXPERIMENT**

- Over 100,000 households
- 12 groups
- 9 treatment groups: 3 different amounts of fine, 3 different levels of threshold
- 3 control groups: receive visual inspection, 3 different amounts of fine
- July to September, 2018

# EXPERIMENT

Visual  
Inspection

300 gal/hr

500 gal/hr

700 gal/hr

100%

100%

100%

100%

50%

50%

50%

50%

25%

25%

25%

25%

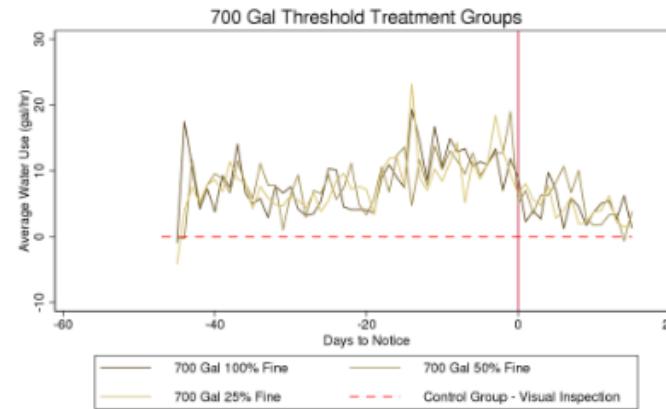
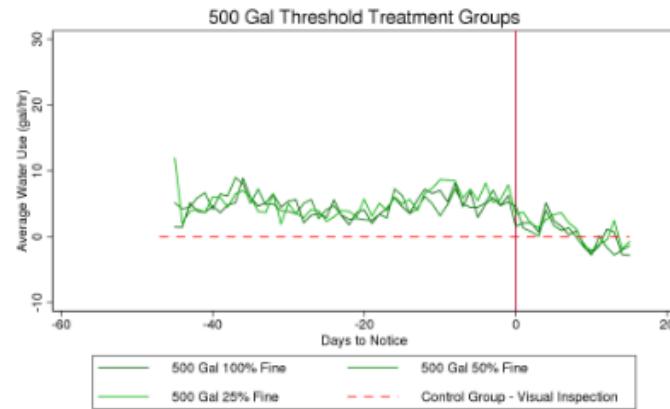
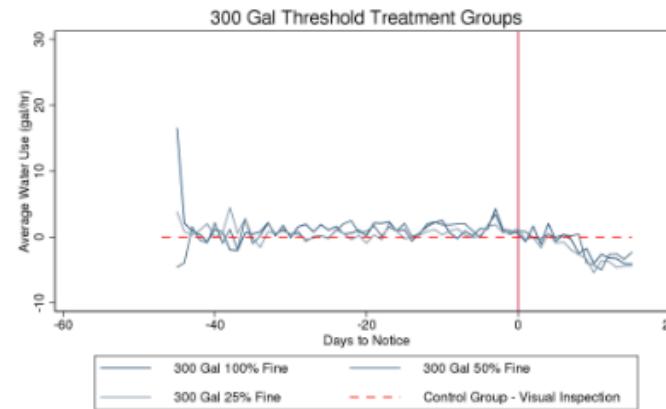
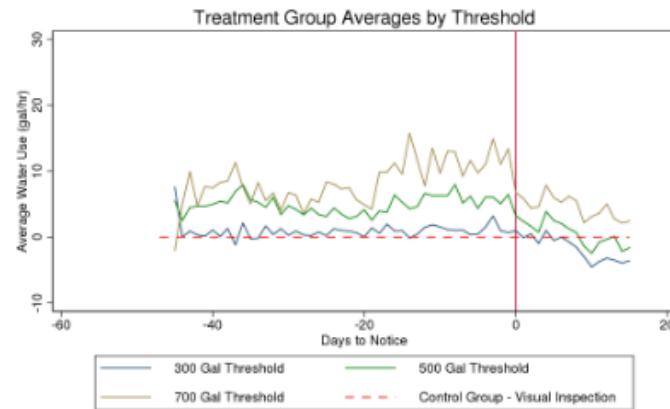
# **DATA**

- **Hourly household level water used (gal/hr)**
- **Amount and date of the fines**

# SOME RESULTS

## Event Study for Average Hourly Water Use during Prohibited Hours

July 17th 2018 to August 2nd 2018



# **SOME TAKEAWAYS**

- **Hire enough software engineers (or MACSS students)**
- **...and make sure that they test their code**
- **Running an experiment or implementing a policy is hard**

Nice work!

Who is up next?

# Integrating Delay Discounting and Heuristics to Better Explain Intertemporal Choice

Xi Chen

Thesis Advisor: Dr. Oleg Urminsky

MACSS Thesis Lightning Presentation

April 18, 2019

# Introduction

Decisions involving consequences at different time points are referred to as *Intertemporal Choice* (Frederick, Loewenstein, & O'Donoghue, 2002).

Heuristic models can outperform traditional delay discounting models. (Ericson, White, Laibson, & Cohen, 2015).

## Research Questions:

- ▶ Comparing heuristic models and delay discounting models – which are better explaining intertemporal choices, and why?
- ▶ By integrating discounting and heuristics, can we find new insights in modeling intertemporal choice?

# Literature

## Delay Discounting Models

- ▶ Exponential model:  $L(a(x_2\delta^{t_2} - x_1\delta^{t_1}))$
- ▶ Hyperbolic model:  $L(a(x_2(1 + \alpha t_2)^{-1} - x_1(1 + \alpha t_1)^{-1}))$
- ▶ Quasi-Hyperbolic model /  $\beta - \delta$  discounting model:  
 $L(a(x_2\beta^{I(t_2>0)}\delta^{t_2} - x_1\beta^{I(t_1>0)}\delta^{t_1}))$

## Heuristic Models

- ▶ Tradeoff model:  $L(a((\log(1 + \gamma_x x_2)/\gamma_x - \log(1 + \gamma_x x_1))/\gamma_x - k(\log(1 + \gamma_t t_2)/\gamma_t - \log(1 + \gamma_t t_1)/\gamma_t)))$
- ▶ DRIFT model:  
 $L(\beta_0 + \beta_1(x_2 - x_1) + \beta_2 \frac{x_2 - x_1}{x_1} + \beta_3 \left(\left(\frac{x_2}{x_1}\right)^{\frac{1}{t_2 - t_1}} - 1\right) + \beta_4(t_2 - t_1))$
- ▶ ITCH model:  
 $L(\beta_I + \beta_{xA}(x_2 - x_1) + \beta_{xR} \frac{x_2 - x_1}{x^*} + \beta_{tA}(t_2 - t_1) + \beta_{tR} \frac{t_2 - t_1}{t^*})$

# Experiments & Data

## Dataset I

- ▶ Ericson, White, Laibson & Cohen (2015)
- ▶ 940 participants; each participant answered 25 MEL questions
- ▶ 5 conditions: delay vs. speedup framing
- ▶ Money range: \$0.01 to \$100,000.00
- ▶ Time range: 0 weeks to 6 weeks

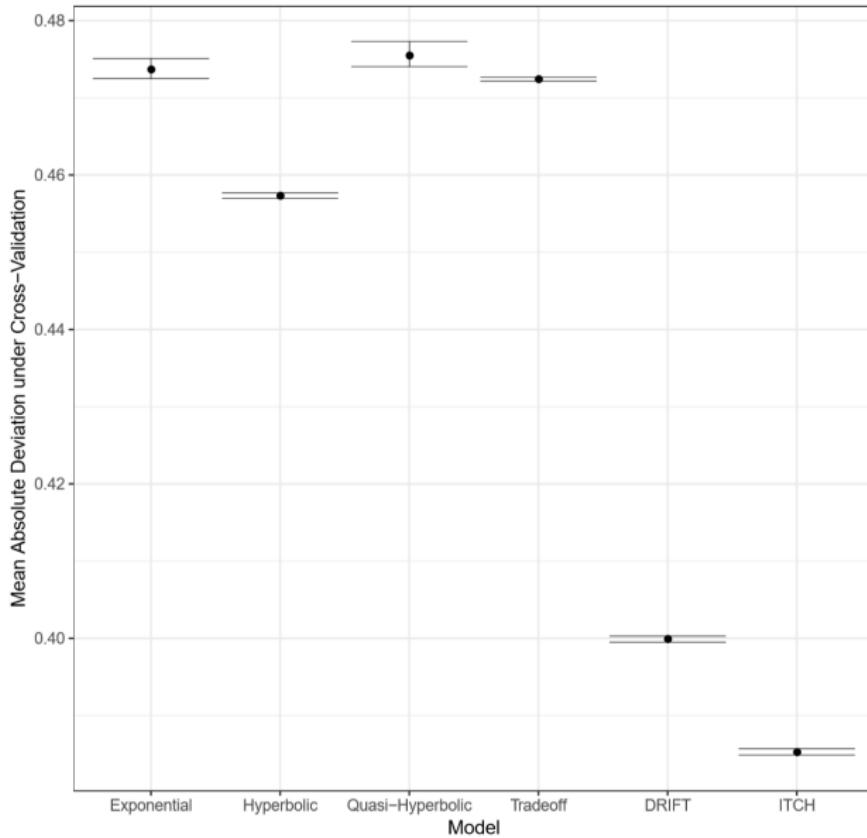
## Dataset II

- ▶ 377 participants; each participant answered 44 MEL questions
- ▶ 4 repeated questions
- ▶ Money range: \$0.01 to \$100,000.00 (more amounts)
- ▶ Time range: 0, 1, 2 ... 365 days ... 30 years (wider range)

# Method

- ▶ Binary outcome - Logistic regression:  
$$L(x) = (1 + e^{-x})^{-1}$$
- ▶ Generalized Linear Models
- ▶ Maximum Likelihood Estimation
- ▶ Cross-validation techniques
- ▶ Error metrics: Mean Absolute Deviation, AIC, BIC

## Results (Dataset II)



# Results (Dataset I)

ITCH model:

$$L(\beta_I + \beta_{xA}(x_2 - x_1) + \beta_{xR} \frac{x_2 - x_1}{x^*} + \beta_{tA}(t_2 - t_1) + \beta_{tR} \frac{t_2 - t_1}{t^*})$$

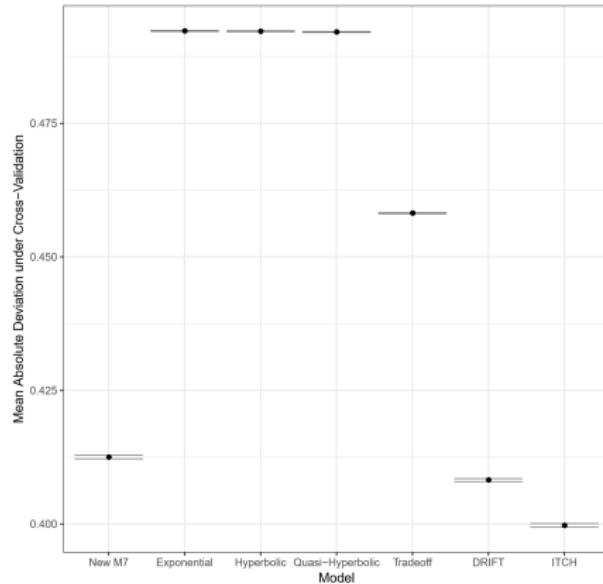
Model Manipulation: Removing ...	Model Fit (Mean Absolute Deviations/MAD)					
	Pooled Data	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5
Baseline / ITCH	0.3997	0.3213	0.3899	0.3158	0.4527	0.4410
Relative Time Term	0.4034	0.3225	0.3924	0.3172	0.4619	0.4460
Absolute Time Term	0.4034	0.3240	0.3948	0.3197	0.4562	0.4440
Relative Money Term	0.4524	0.3963	0.4578	0.3938	0.4761	0.4779
Absolute Money Term	0.4063	0.3320	0.3933	0.3277	0.4572	0.4464
Relative Terms	0.4554	0.3976	0.4592	0.3950	0.4839	0.4829
Absolute Terms	0.4099	0.3350	0.3974	0.3327	0.4603	0.4493
Constant Term	0.4318	0.4097	0.4199	0.4029	0.4585	0.4466

## New Models (Dataset I)

New Model 6:  $\beta_1(v_2 - v_1) + \beta_2 \frac{v_2 - v_1}{v^*}$ ,  $v_1 = x_1 \delta^{t_1}$ ,  $v_2 = x_2 \delta^{t_2}$

New Model 7:  $\beta_1(v_2 - v_1) + \beta_2 \frac{v_2 - v_1}{v^*} + \beta_3(d_2 - d_1)$

New Model 8:  $\beta \frac{v_2 - v_1}{v^*}$



## New Models (Dataset II)

New Model 10:  $\alpha + \beta_1(x_2 - x_1) + \beta_2(t_2 - t_1)$

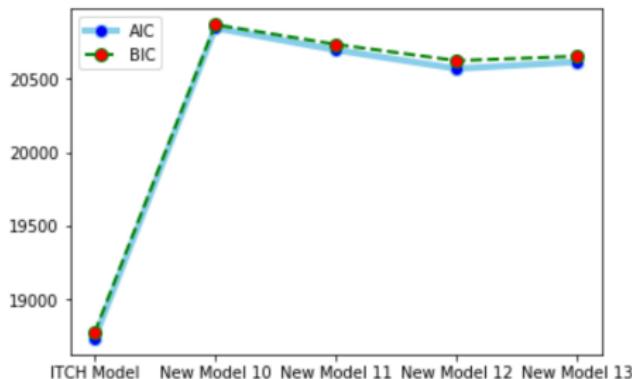
New Model 11:  $\alpha + \beta_1(x_2 - x_1) + \beta_2(t_2 - t_1) + \beta_3x_1 + \beta_4t_1$

New Model 12:

$\alpha + \beta_1(x_2 - x_1) + \beta_2(t_2 - t_1) + \beta_3x_1 + \beta_4t_1 + \beta_5(x_2 - x_1)x_1 + \beta_6(t_2 - t_1)t_1$

New Model 13:

$\alpha + \beta_1(x_2 - x_1) + \beta_2(t_2 - t_1) + \beta_3(x_2 - x_1)x_1 + \beta_4(t_2 - t_1)t_1$



# Discussion

- ▶ Heuristics models capture some important characteristics of intertemporal choice that the standard economic models haven't.
- ▶ People's decisions do seem to incorporate relative judgments.
- ▶ Integrating discounting and heuristics may be a promising way to develop better intertemporal choice models.

## Future Direction:

- ▶ Parameter recovery - simulating data from models
- ▶ Heterogeneity - individual level modeling

Nice work!

Who is up next?

# How to Predict Spatial Distribution of Airbnb in Cities: A Machine Learning Method

Mengchen Shi

Advised by: Dr. Richard Evans

M.A. in Computational Social Science  
The University of Chicago

April 18, 2019

# Research Question

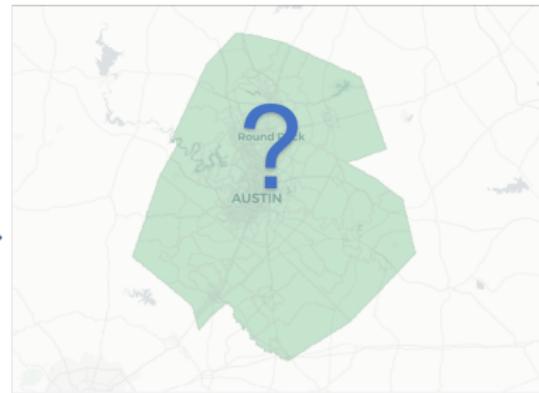
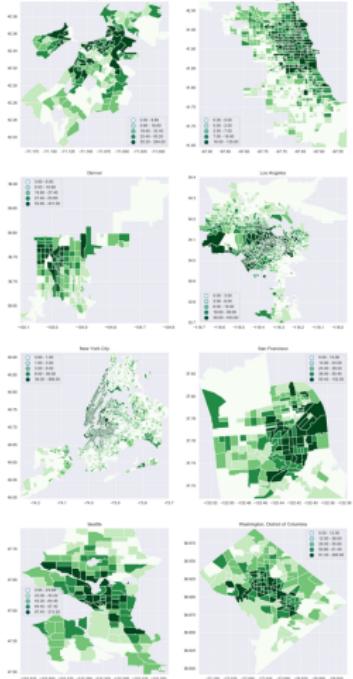
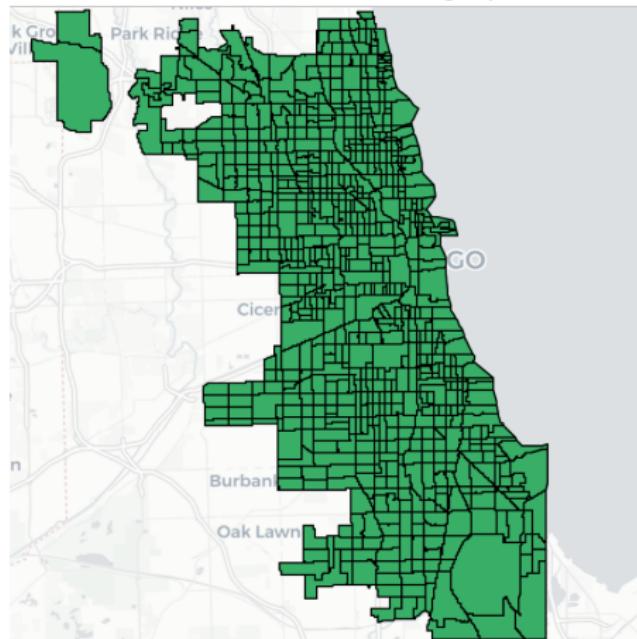


Figure: Predict Airbnb Listings Distribution in a City

# Spatial Units and Neighbors

801 Census Tracts in Chicago (2010 Census)



Queen contiguity-based Neighbors



# Data

- United States Census Bureau: Census data and spatial units
- Inside Airbnb: a website periodically publishes snapshots of Airbnb listings founded by Murray Cox
- OpenStreetMap: a collaborative project to create a free editable map of the world

# Variables of Interest

Category	Variable	Description
Airbnb	num_list	Number of Airbnb listings in a given tract
Geographic	distance	Distance a tract to downtown
	hotel	number of hotels in the area
	poi	Number of point of interests in the area.
	trans	Number of public transportation infrastructure
	pop_den	Population density in the tract
Economic	unemp	Proportion of unemployed residents
	log_inc	Log(Median of household income in an area)
	log_hvalue	Log(Median of housing value in an area)
	owner	Proportion of owner-occupied properties
	poverty	Proportion below poverty level
Social	edu	Proportion of residents with an advanced degree
	young	Proportion of people aged between 20 and 34
	race	Race diversity represented by Gini-Simpson index

# Variables of Interest

## Variables about neighbors

Suppose a census tract has  $m$  neighbors, and we have  $n$  variables.

Neighbor mean:

$$\frac{\sum_{j=1}^{m_i} X_{ij}}{m}, i = 1, 2, \dots, n; j = 1, 2, \dots, m$$

Neighbor maximum:

$$\text{Max}(X_{ij}), i = 1, 2, \dots, n; j = 1, 2, \dots, m$$

Neighbor minimum:

$$\text{Min}(X_{ij}), i = 1, 2, \dots, n; j = 1, 2, \dots, m$$

# Spatial Autocorrelation

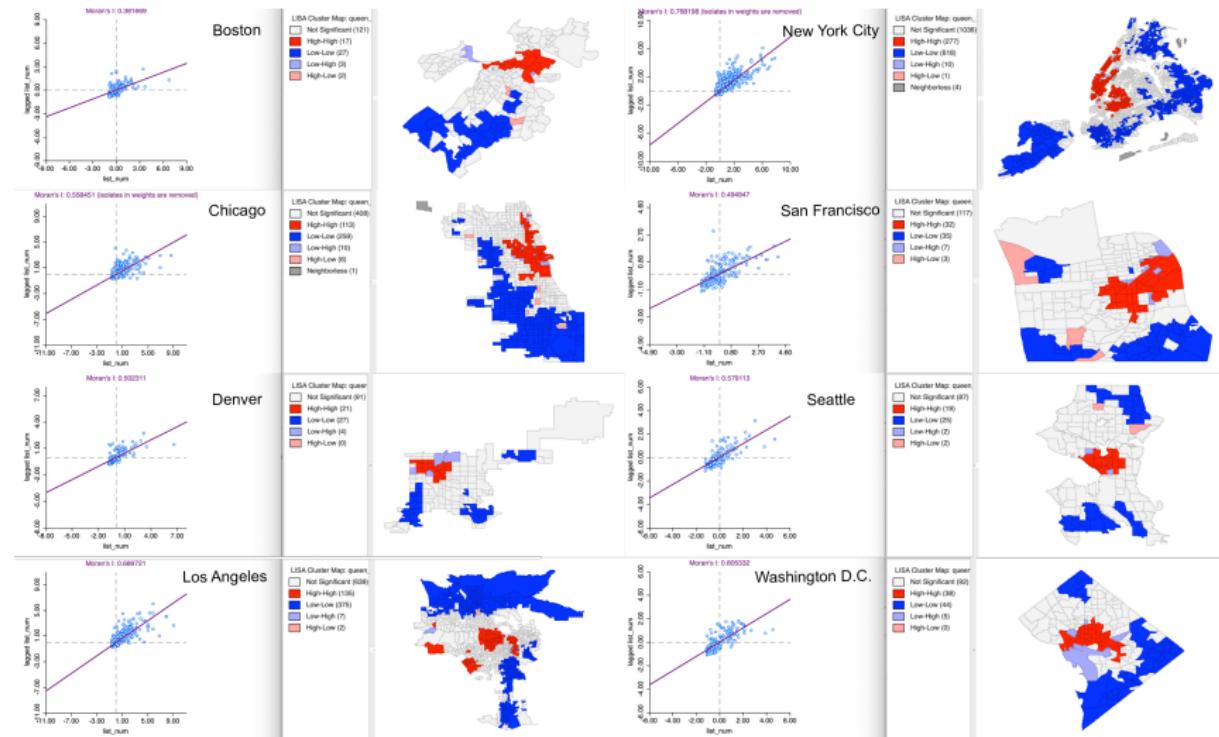


Figure: Local Moran's I and Clusters

# Regression Results

[scale=0.5]

Table: RMSE of Lasso, Random Forest, and XGBoost Regression

Model	OLS	Random Forest	XGBoost
	Spatial v.s.(non-spatial)		
All 8 cities	31.49 (32.37)	26.89 (27.94)	25.39 (27.62)
Boston	37.69 (26.78)	32.93 (31.95)	33.69 (33.61)
Chicago	10.01 (9.20)	9.76 (9.70)	9.84 (9.34)
Denver	32.93 (27.10)	35.04 (35.63)	37.38 (40.39)
Los Angeles	34.77 (34.66)	30.69 (33.39)	32.40 (33.87)
New York City	28.60 (28.73)	23.01 (24.53)	21.96 (24.55)
San Francisco	22.49 (19.95)	21.67 (22.46)	20.66 (22.88)
Seattle	30.92 (27.90)	37.35 (37.35)	36.61 (37.71)
Washington, D.C.	36.18 (27.18)	30.69 (32.58)	31.66 (34.12)

# Feature Importance

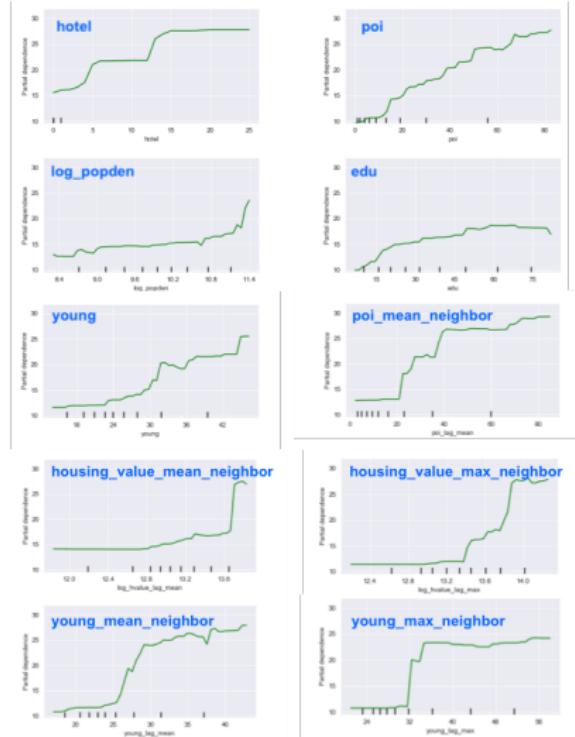
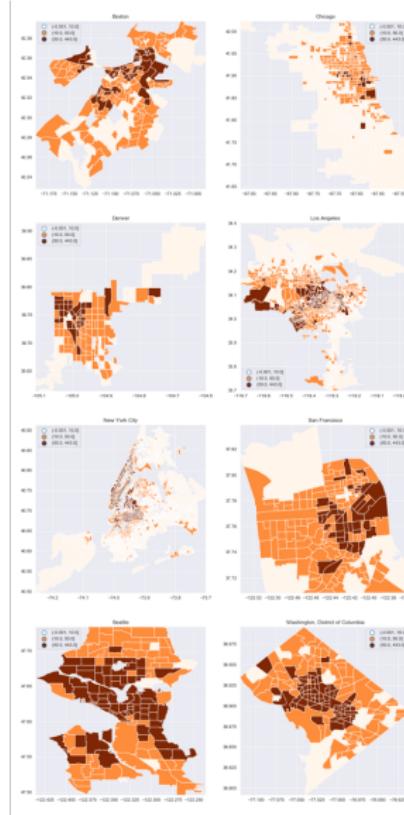
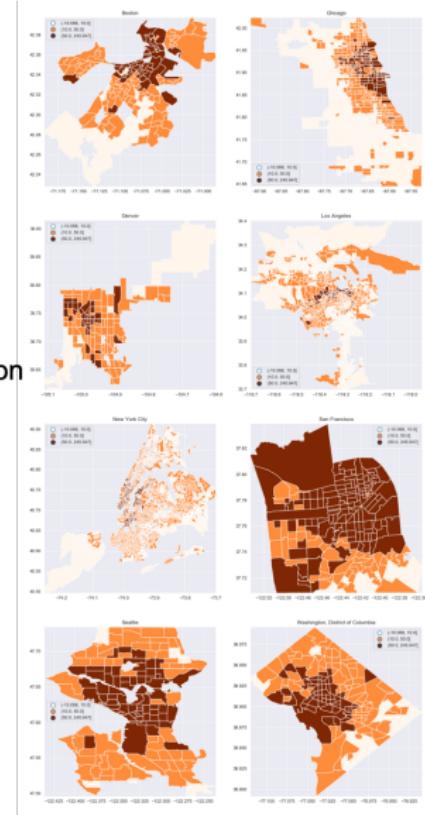


Figure: Most important features selected by XGBoost regression

# XGBoost Prediction Results



Out of sample prediction



Nice work!

Who is up next?

verb – present participle

noun phrase

[ Studying Language Development

within a Text Classifier Framework ]

preposition

indefinite  
article

noun phrase

Flora Zhang

Proper nouns - names

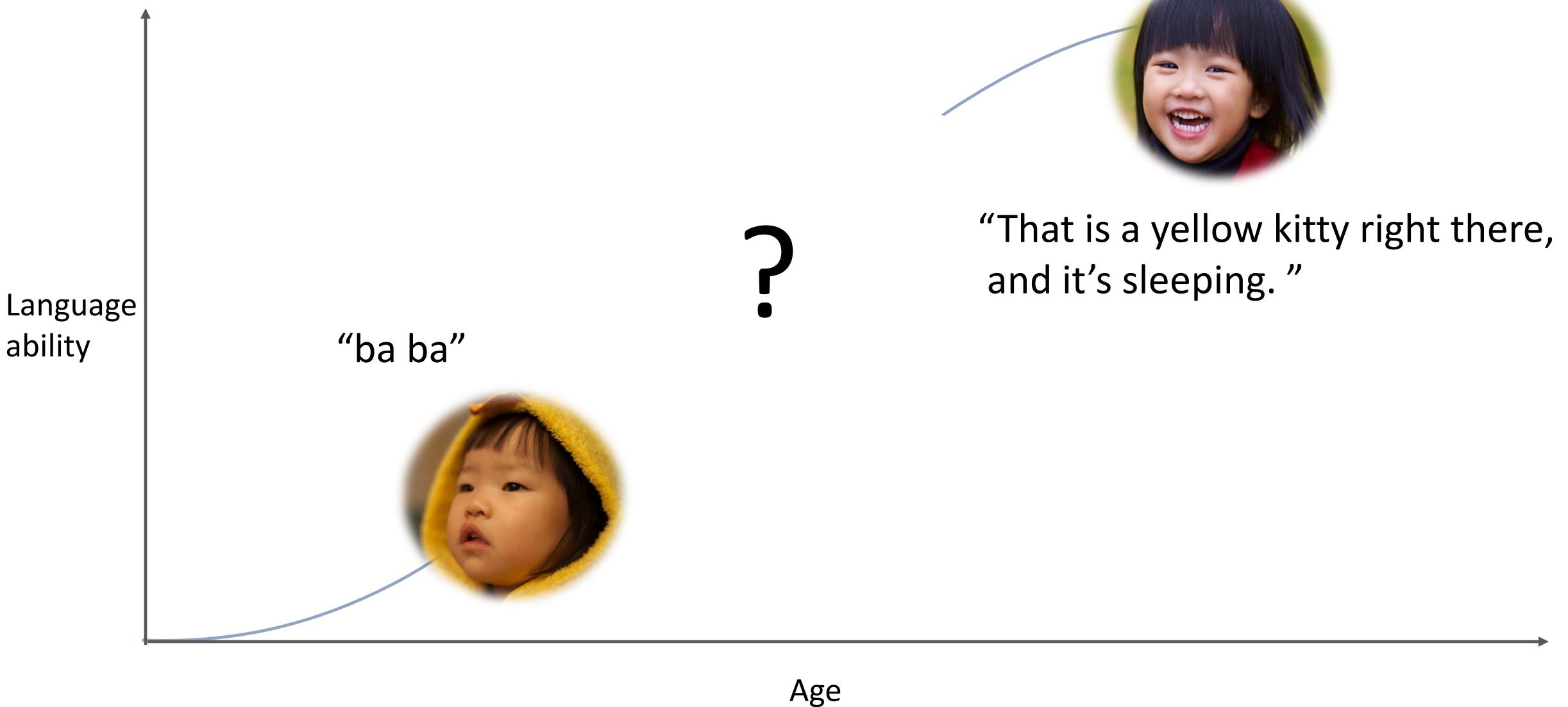
[ Who do you think said these sentences?

... and, how can you tell? ]

“That is a yellow kitty right there,  
and it’s sleeping.”

“ba ba”

# [ How did children become competent speakers of a language? ]



Historically, the focus has been on

Topically

Number of words  
(Huttenlocher et al., 1991; Fernal et al., 2013)

Specific syntactic structures  
(Brown, 1973; Bowerman, 2014)

... ...

Methodologically

Small number of data points  
(Nice, 1925; Brown, 1973)

Observation at one or a few time points  
(Johnson et al., 2005; Yuan & Fisher, 2009)

Data collection  
(Nice, 1925; Brown, 1973; Parra et al, 2011)

Historically, the focus has been on

## Current project

Topically

- Number of words
- Specific syntactic structures
- ....



**The construction of language:**  
- grammar  
- words  
- communicative signals

Methodologically

- Small number of data points
- Observation at one time point
- Data collection



## Data from the Language Development Project (Goldin-Meadow et al., 2014)

- 60 children and their parents
- racially and socio-economically diverse
- Across 5 years, from 14 to 58 months
- Transcriptions of recordings

# [ Build a text classifier for each time point in our data ]

At time  $t$ , **Input:** a sentence

**Output:** child OR parent

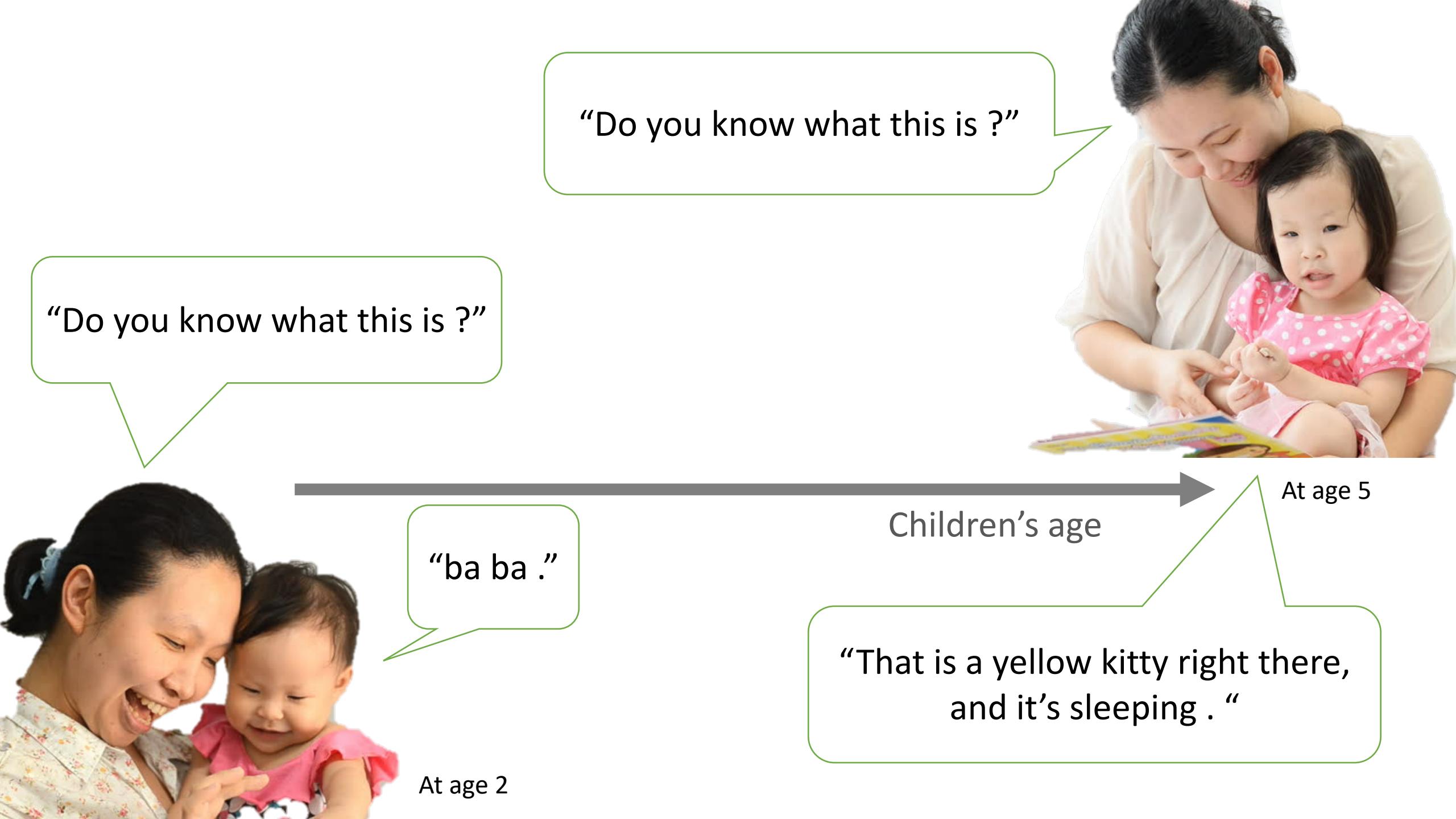
For a given sentence  $S$ ,

what is the probability that a parent said  $S$ ?

what about the child's probability?

$S$  is represented as a composition of different linguistic features

# of words, # of verbs, # nouns,  
whether it is a question, ... ..., in  $S$



Yellow	Verb
Cyan	Pronoun
Magenta	Noun
#	Number of words
Green	Question



“Do you know what this is ?”

“Do you know what this is ?”

Children’s age

“ba ba .”

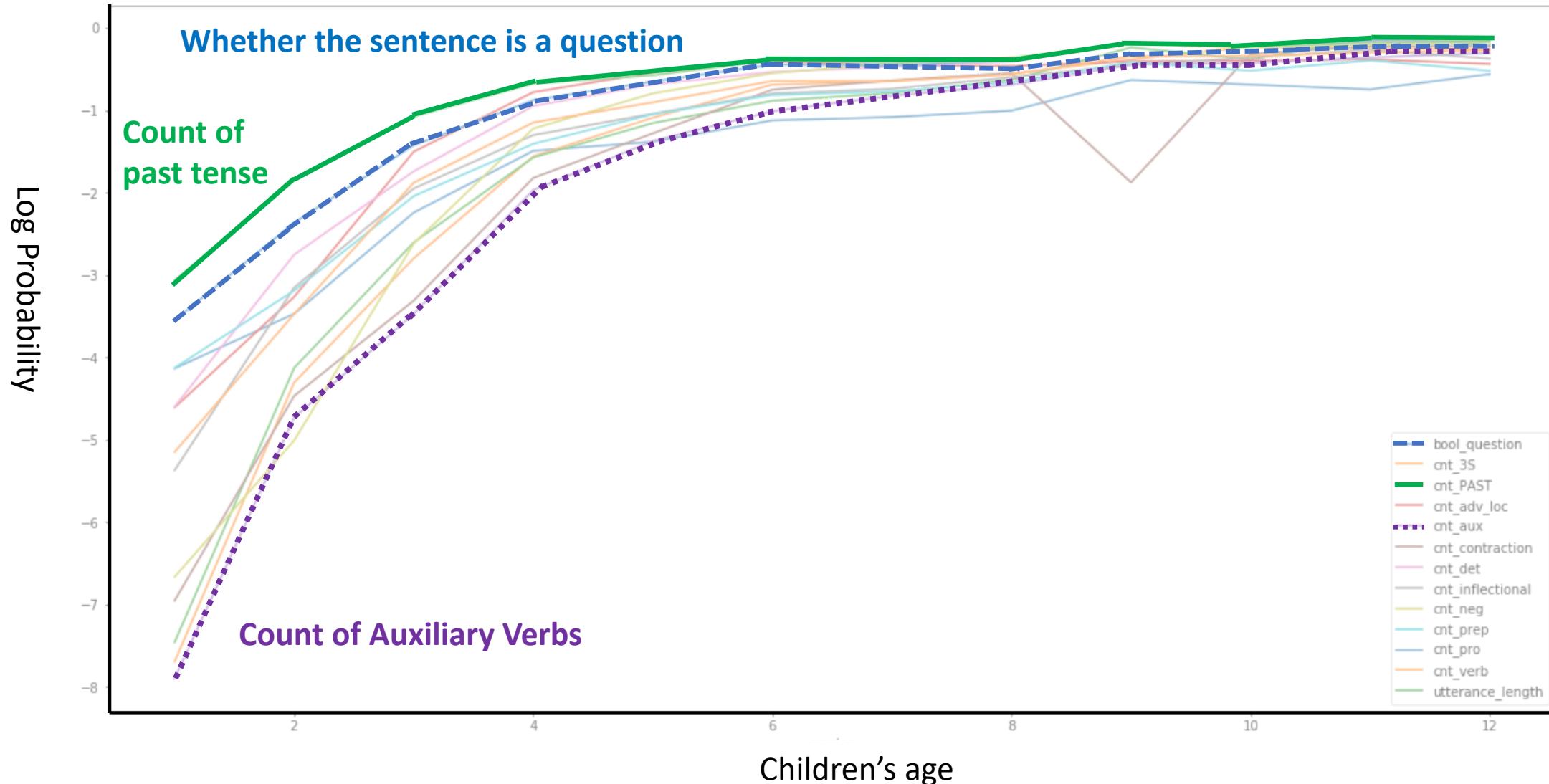
At age 2



At age 5

“That is a yellow kitty right there,  
and it’s sleeping .”

[ “meaningless” features of language  
can be particularly informative and a crucial part of a mature language]





Advisor: Daniel Yurovsky



Collaborator: Allyson Ettinger

# [ Thank you ]



Members of CAL Lab

Nice work!

Who is up next?

# How does parental speech affect early language learning?

---



Parents' Linguistic Alignment Predicts  
Children's Language Development  
Joseph Denby  
Advisor: Dan Yurovsky

How do kids learn language so quickly?

Learning a language is easy?



Kids are smarter than adults?



Kids get input tuned specifically for them?

Linguistic Tuning Hypothesis (Snow, 1972)

# How do we measure this?

## Linguistic Alignment

(Pennebaker et al., 2015; etc.)

Features of interlocutors' language  
dynamically interact over time

Parent: I don't know. I'll have to think about it.

Child: I was going to do the people across street.

Parent: across the street?

Child: yeah.

## How do we measure this?

We can borrow formal notions of conversational dynamics to investigate linguistic tuning!

- Specifically, dynamics at the syntactic level

---

Parent: I don't know. I'll have to think about it.

Child: I was going to do the people across street.

Parent: across the street?

Child: yeah.

---

# Language Development Project

(Goldin-Meadow et al., 2014)

In-home conversations between parents and children from across Chicago

- 59 parent-child pairs  
14-58 months old at 4 month intervals

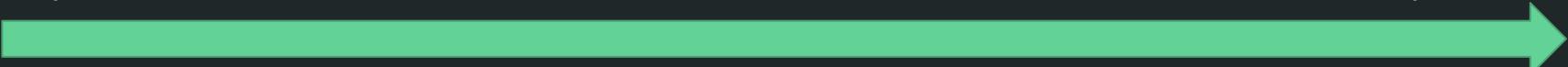


1 year

Estimate alignment over time and measure relationship with vocabulary



5 years



# Hierarchical Alignment Model

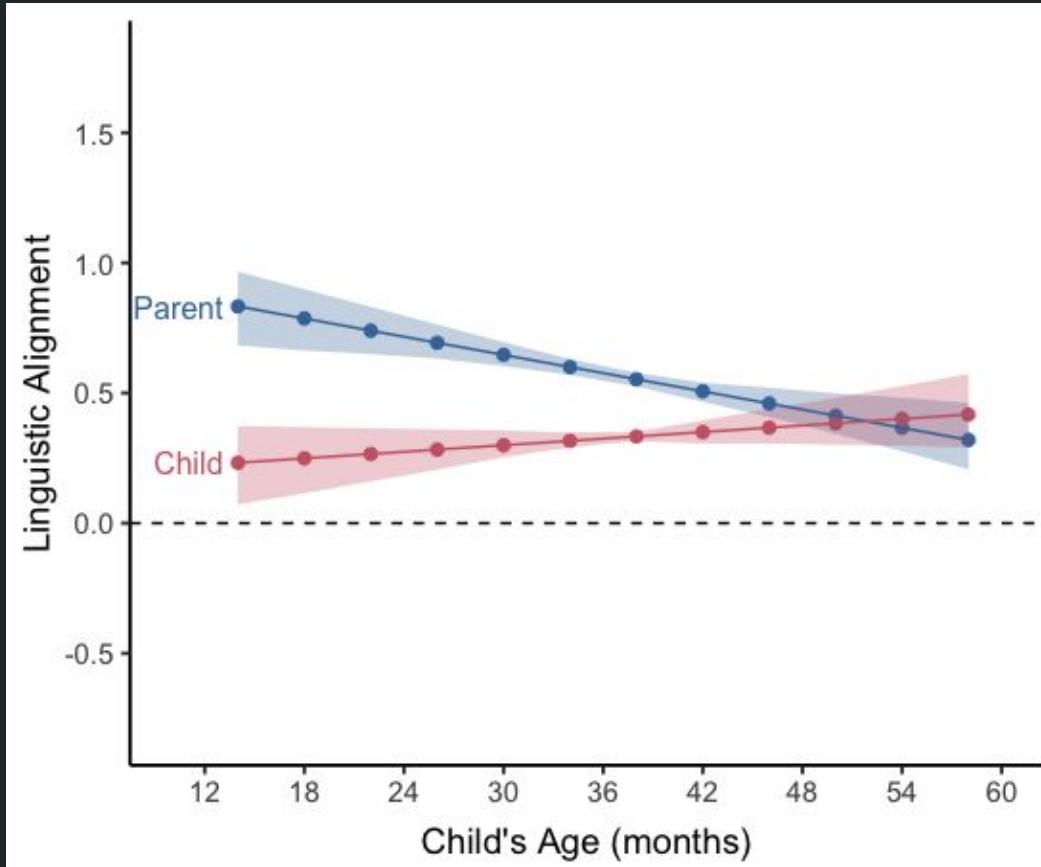
(Yurovsky et al., 2016)

1. Alignment = increase in odds of using a syntactic category after your partner uses it
2. Use alignment to predict vocabulary (alongside demographic variables)



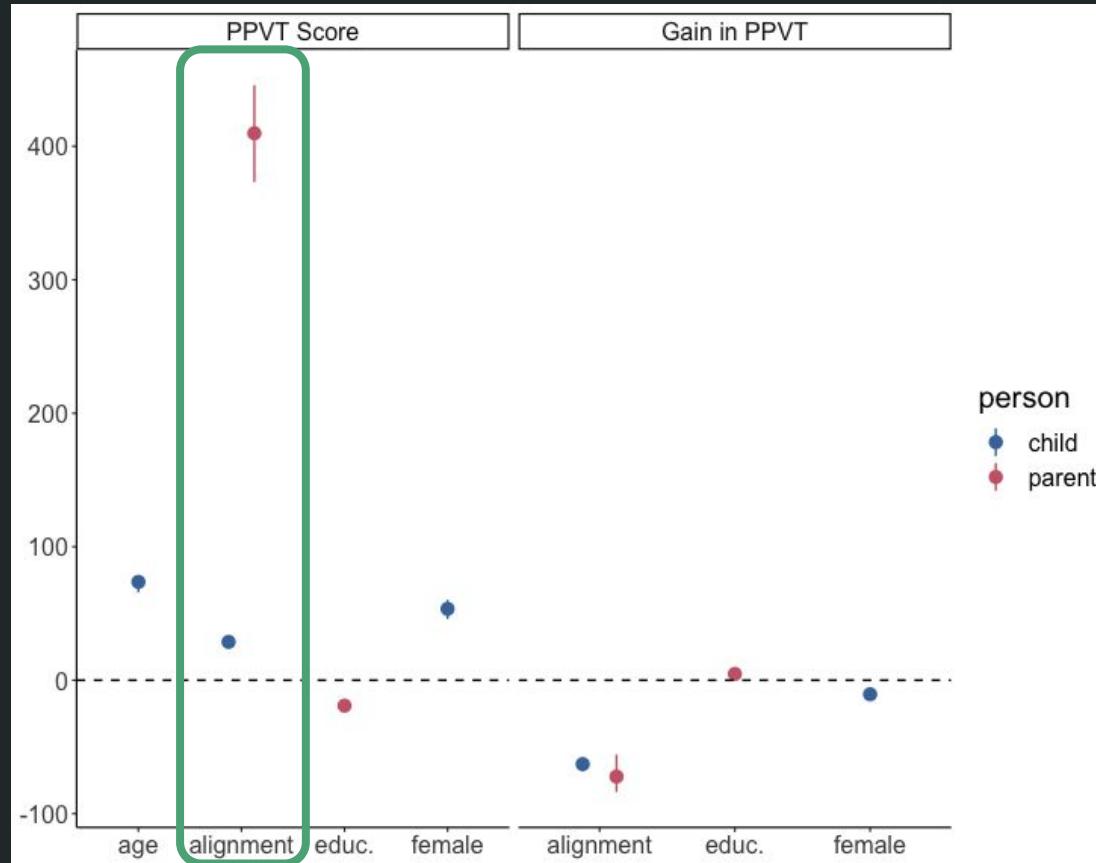
Estimated using Stan, a Bayesian MCMC sampler

Alignment exists between parents and children (and vice versa)



# Parental alignment predicts vocabulary development

Controlling for  
demographics  
(e.g., education,  
gender)



# Summary

- Presence of syntactic alignment is replicated in a new dataset
- Moreover, alignment effects on development are significant
  - Indicates that alignment affects language development beyond demographics
- Future work
  - Tinkering with alignment model
  - Extending to / aggregating more outcome measures



# Thanks!

---

Nice work!

Who is up next?

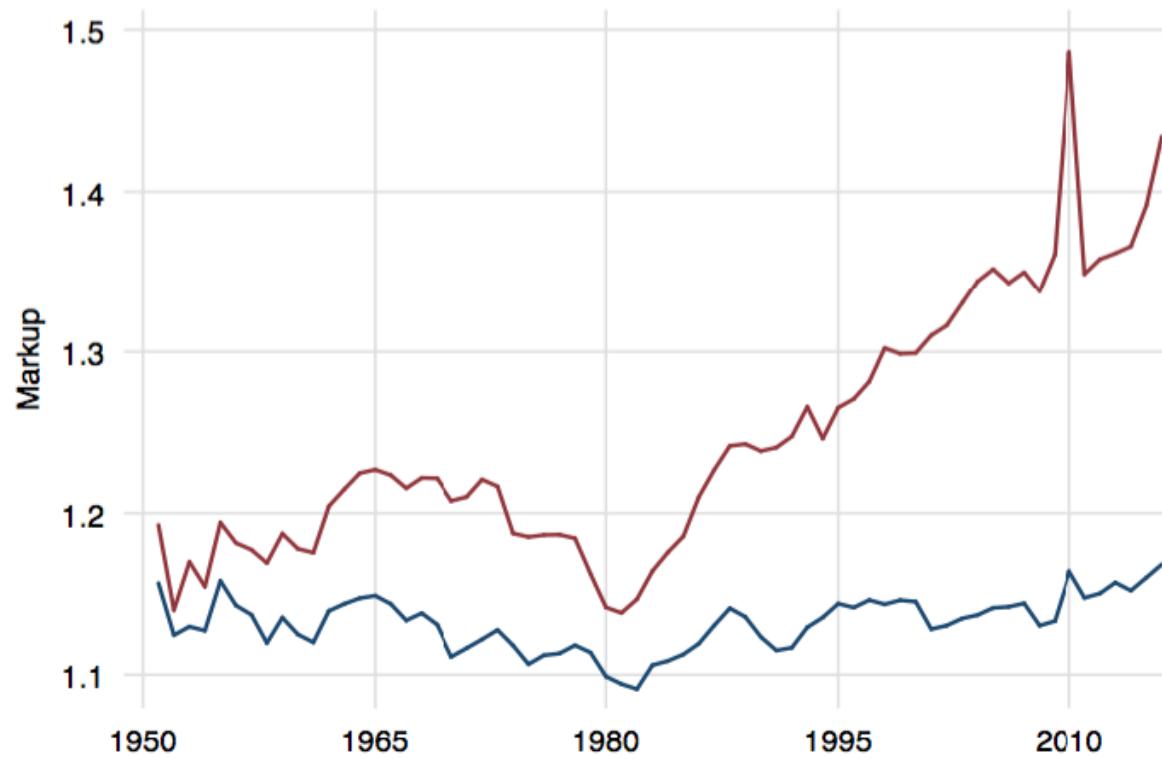
# The Rise of Market Power or Transformation of Production?

*Anhua Chen*

*Prepared For MACSS Lightening Talk Night*

# A Rise in Markup, or Not?

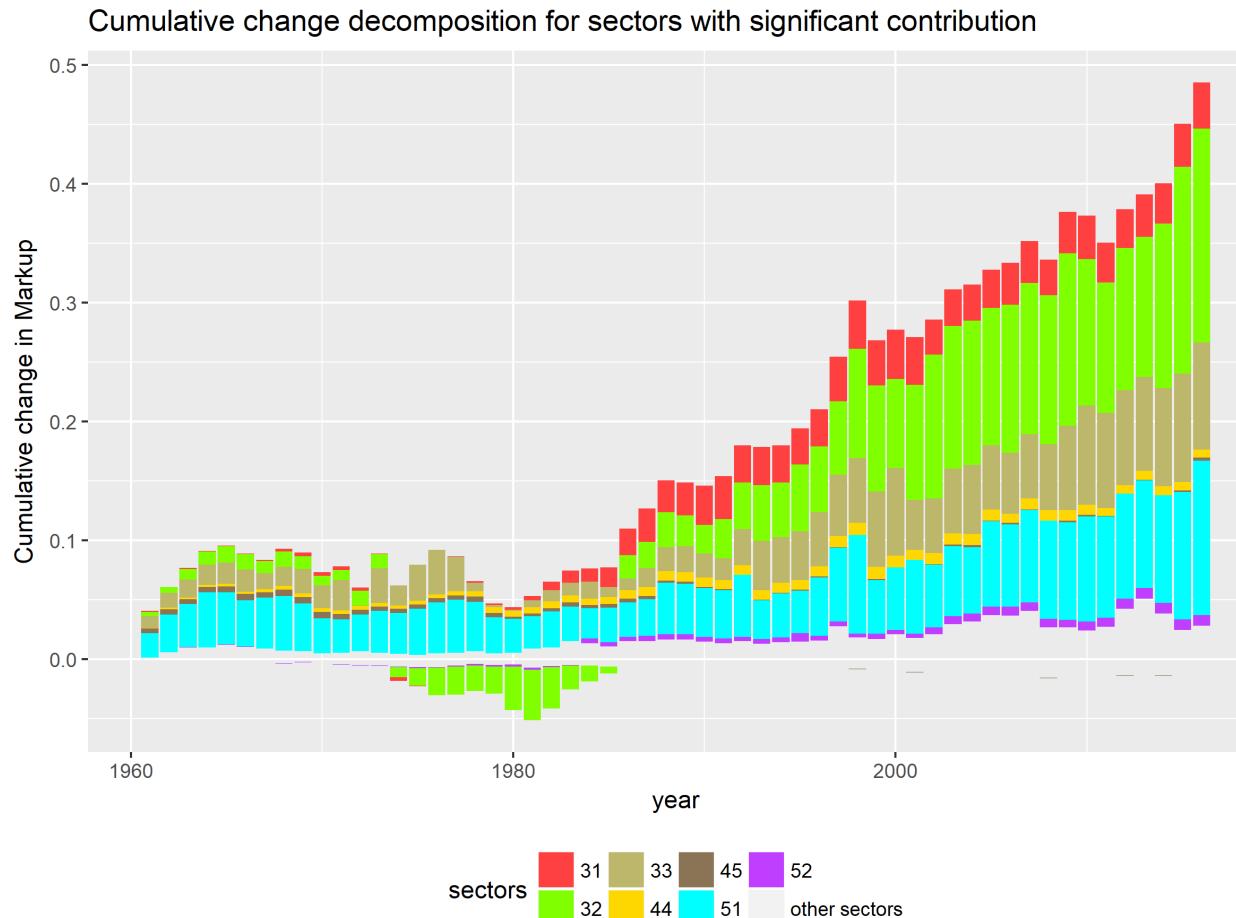
- Markup = Price / Marginal Cost
- The answer is No, when you account for transformation of production



Source: Traina (2018)

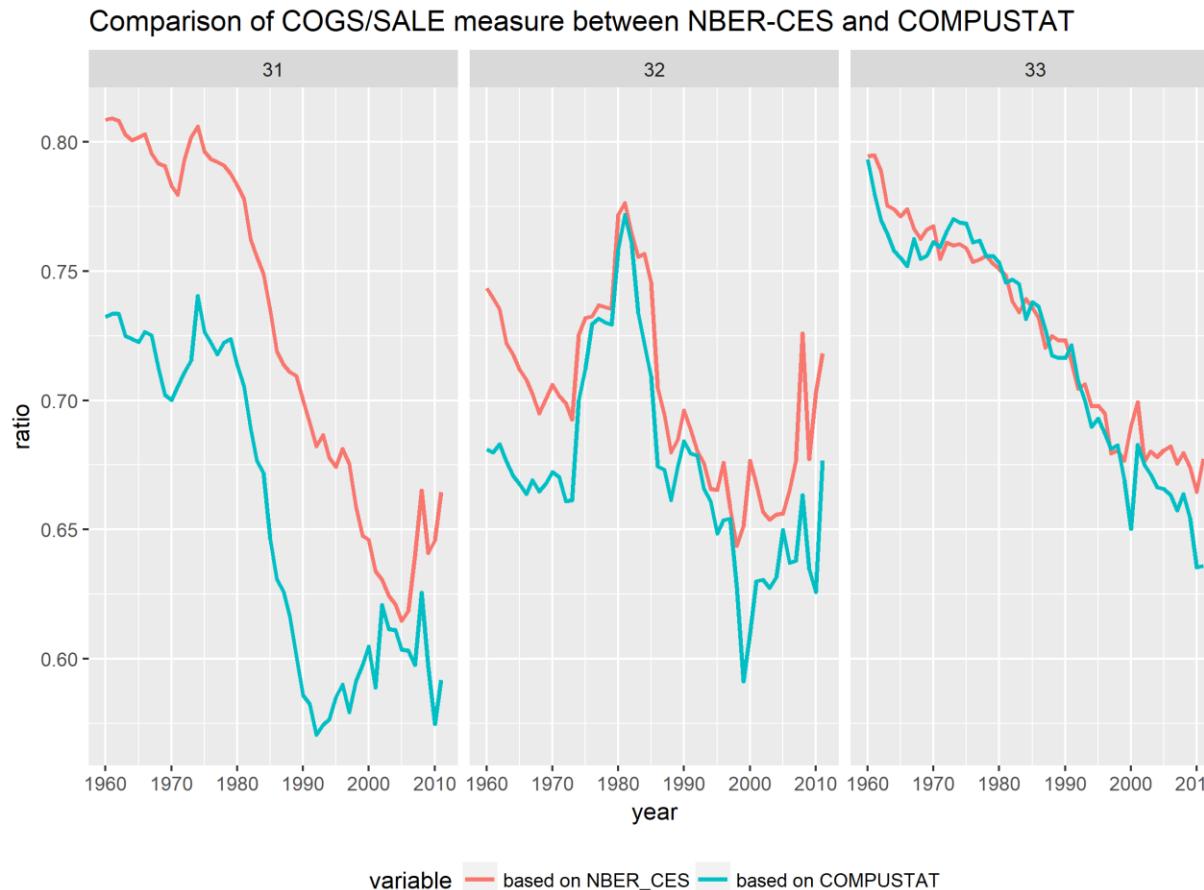
# Where did the transformation of production happen the most?

- Manufacturing



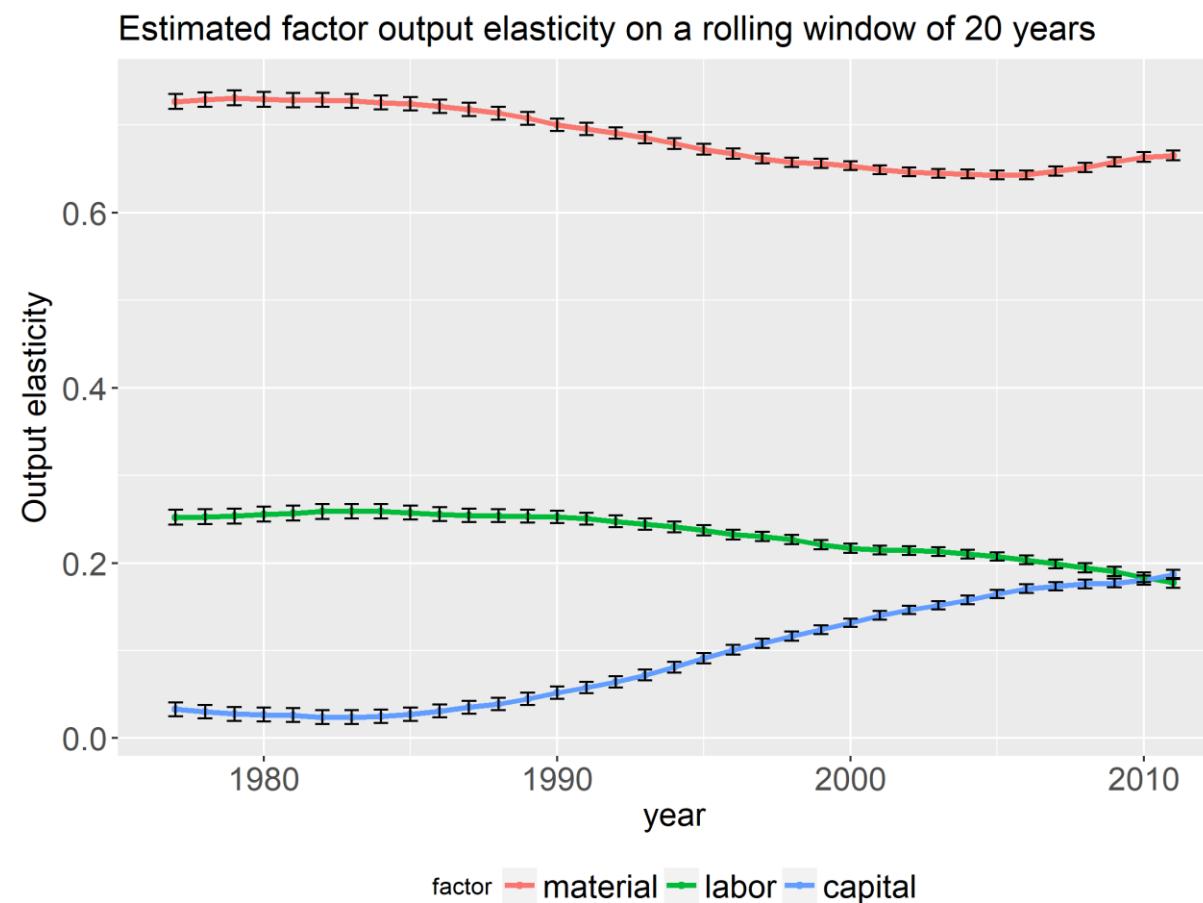
# Pairing Compustat with Census Manufacturing Data

- Same Trend (**Production VS Markup**) captured in both data



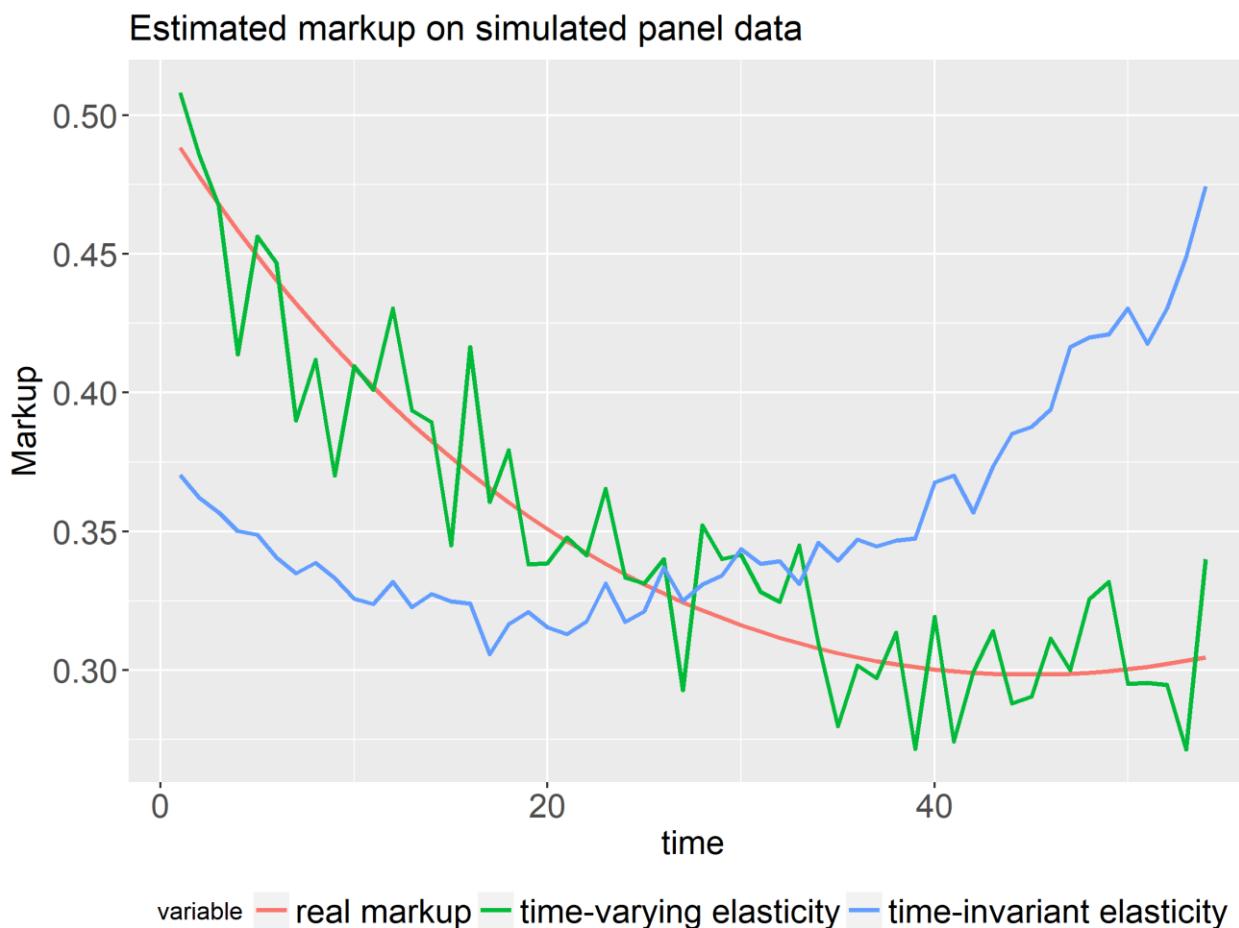
# Less Labor and More Capital

## - Capital Matters



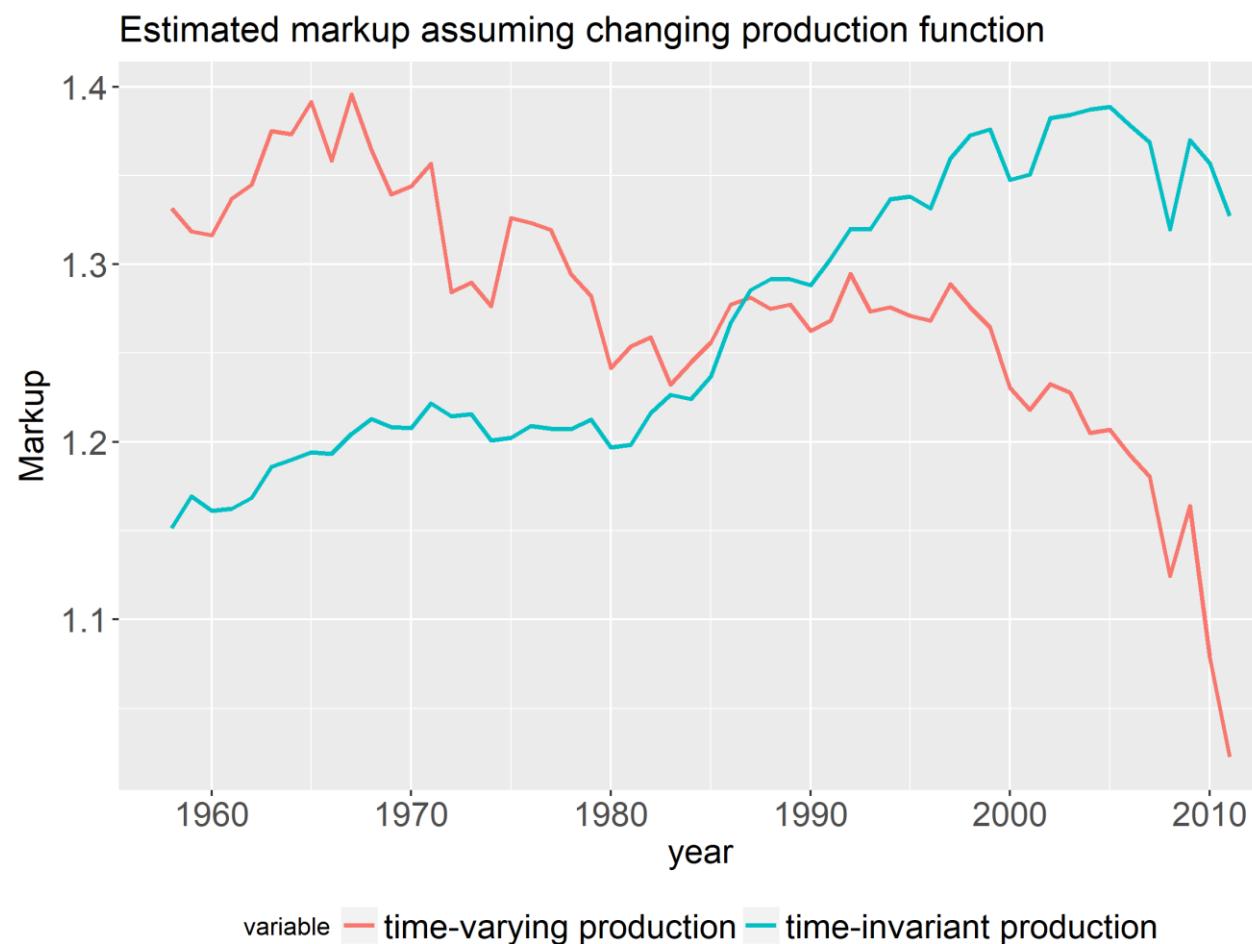
# Monte-Carlo Simulation

- A response to declining relative price of Capital



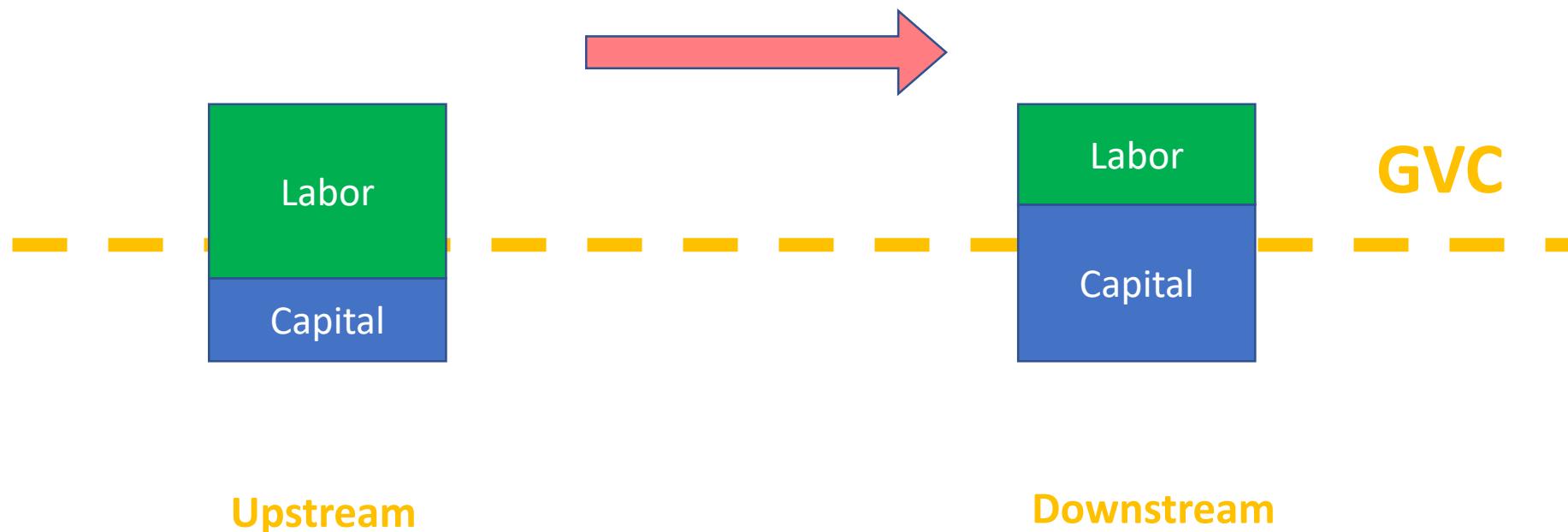
# A Decreasing, Not Increasing Mfg. Markup

- When accounting for a transforming production process



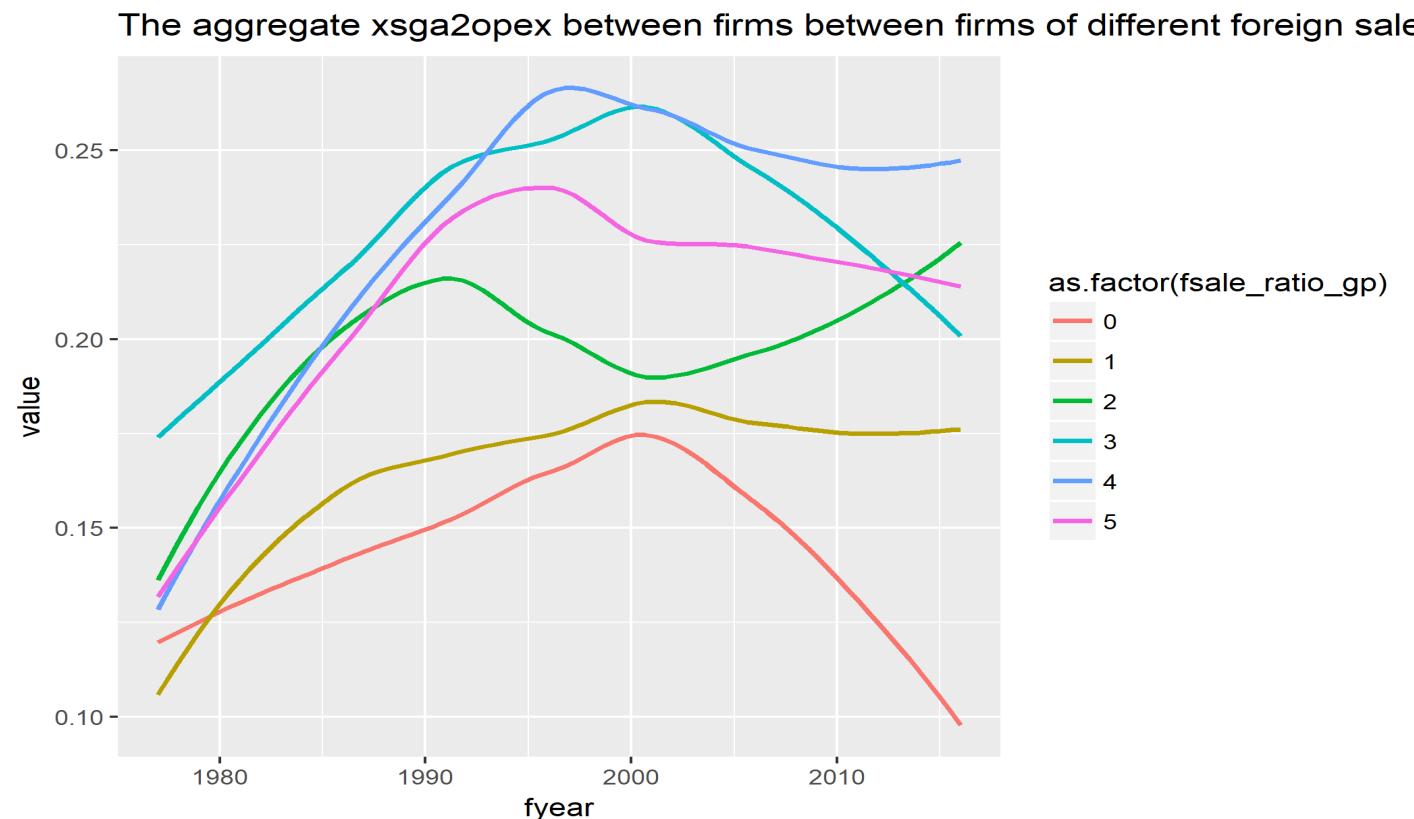
# Some Other Explanations: GVC

- Move down along the Global Value Chain :  
a more capital/intangible-intensive model



# Some Other Explanations: GVC -- Evidence

- Using Compustat-segment dataset, paired with text-mining on 10K
- Firms with larger share of foreign sales tend to see larger extent of transformation of production



# Policy Implication

- Call for Anti-trust policy might be too-premature
- Instead, more focus should be shifted towards:
  - Education Policy: in response to the shift of production factors
  - Global tax-policy coordination: in response to US industry's changing position on GVC

Nice work!

Who is up next?

# Income Inequality, Social Disorganization, and Urban Crime

Jiang Wang

*Advisor:* Dr. Benjamin Soltoff

# Research Question

- ❖ How do within-group income inequality and social disorganization help predict crime incidents in Chicago?
- ❖ Unit of analysis: Census tract in Chicago city (US Census 2010)
- ❖ Within-group income inequality: The Gini coefficient of a census tract
- ❖ Social disorganization (Shaw and Mckay, 1942):
  - ❖ Poverty
  - ❖ Heterogeneity
  - ❖ Physical dilapidation
  - ❖ High transient population
  - ❖ Other social illnesses

# What contributes to urban crimes?

- ❖ Individual (Biological) characteristics
- ❖ (Socioeconomic) Unemployment and poverty
- ❖ Cultural factors (Family, peer, community)
  
- ❖ and more .....

# Some Crime Theories

- ❖ Economic
  - ❖ Income inequality
  - ❖ Opportunity\*
- ❖ Sociological
  - ❖ Social disorganization
  - ❖ Relative Deprivation
  - ❖ Class conflict\*
  - ❖ Strain
  - ❖ Routine activity\*
  - ❖ Collective efficacy
  - ❖ Broken Window

# Data

- ❖ Chicago crime data (2012-2016), point-aggregated
- ❖ Chicago ACS 2012-2016, census tract level
- ❖ Unit of analysis: Chicago 2010 census tracts, excluding partial city census tracts
  - ❖ 787 census tracts in total

# Method

- ❖ Cliff-Ord autoregressive model with spatial autoregressive disturbances (SARAR)

$$\ln Y_i = \lambda W \ln Y_i + \sum_j \alpha_j X_{ij} + u_i$$

where  $u_i = \rho W u_i + \varepsilon_i \quad \forall \varepsilon_i \text{ i.i.d.}$

- ❖ Removes autocorrelation in the dependent variable as well as the error term
- ❖ Spatial weight: Queen contiguity

# Variables

Variable	Definition	Description
Crime Rate	The natural log of the annual reported crime incidents per 1000 population per census tract	Dependent
Income	The natural log of aggregated median household income	Control
Unemployment	% of population that are in the civilian labor force but are unemployed	Control
Education	% of population aged over 25 that have a high school degree or higher	Control
Income Inequality	The estimated income Gini coefficient	Inequality
Poverty	% of households that are below the poverty line	Social disorganization
Foreign	% of foreign-born population	Social disorganization
Vacancy	% of housing units that are vacant	Social disorganization
Renter	% of housing units that are occupied by renters	Social disorganization

# Tentative Results

- ❖ Significant: Unemployment, Income, Income Inequality, foreign, vacancy
- ❖ Interestingly,
  - ❖ The impact of income and income inequality reduces after introducing social disorganization variables
  - ❖ Vacancy is very influential
  - ❖ The coefficient of foreign is negative (contrary to Shaw and Mckay's findings)

# Limitations

- ❖ Does not discuss the issue: “Ascription” vs. “Achievement”
- ❖ Does not discuss the effect of the choice of unit of analysis
- ❖ ... and more

# Further Improvements

- ❖ Interactions between units of analysis
- ❖ Try different of analysis
- ❖ Leveraging new datasets to test other theories
  - ❖ E.g. Street disorganization
- ❖ Using PYSTAL package

Nice work!

Who is up next?



---

# Spatial Linguistics Variations in Northeastern U.S. Based on Geo-Tagged Tweets

---



Andi Liao  
Advisor: Luc Anselin  
2019/04/18



# Research Question

How do linguistic features derived from  
lexical alternation pairs  
local words      average score for state  
geo-tagged Tweets vary within the Northeast part  
of U.S.?  
latitude + longitude + text info      to overcome data sparsity



# Boundary Clarification

## Content

- **No NLP!!!**
- **Spatial Linguistics**
  - Method: Spatial Data Science
  - Concept: Geo-Linguistics

## Task

- **Explore spatial linguistics variations at state level**
- **Predict geo-locations using text information**



# Relevant Literature: Geo-Linguistics Patterns

## Spatial

- Spatial distribution of lexical alternation pairs

Mom-Mother

- Multivariate mapping approach using 13 principal components
- Regionalization methods for constrained hierarchical clustering and partitioning

- Spatial variations of African American Vernacular English

- Mapping around 30 common nonstandard spellings on Twitter
- Subregions align with movement patterns during the Great Migrations
- Huang, Guo, Kasakoff & Grieve (2016); Jones(2015)

## Temporal

- Diffusion of lexical changes

- An autoregressive model of word frequencies to demonstrate the linguistic influence between American cities
- The network is helpful in identifying geographical and demographical factor that drives the spread of lexical innovation

- Linguistics evolvement in urban areas using frequently used terms

- A logistics regression model consisting of geographical and demographical predictors
- Absolute difference of the percentage of African Americans was the most powerful indicator of linguistics transmit
- Eisenstein, O'Connor, Smith, and Xing (2012, 2014)



# Relevant Literature: Geo-Prediction Models

- **A probabilistic framework via Tweet contents**
  - Trained a local word classifier
  - Constructed a lattice-based neighborhood smoothing model to balance cities and words of various distributions
  - Both local word filtering and smoothing have positive impact on prediction accuracy, and with location estimators, 51% of Twitter users can be placed within 100 miles of their actual locations at the city level
- **Decomposing lexical variation as regional and topical variation**
  - Constructed a prediction model with the assumption that regions and topics interact to shape observed lexical frequencies
  - The model can identify words with high regional affinity as well as geographically-coherent linguistic regions
- **A multi-elemental location inference method**
  - Combing text contents, profile location and place labelling
  - The model can successfully predict 87% of Tweets locations at the average distance error of 12.2 km
  - Cheng, Caverlee and Lee (2010) ; Eisenstein, O'Connor, Smith and Xing (2010) ; Laylavi, Rajabifard and Kalantari (2016)

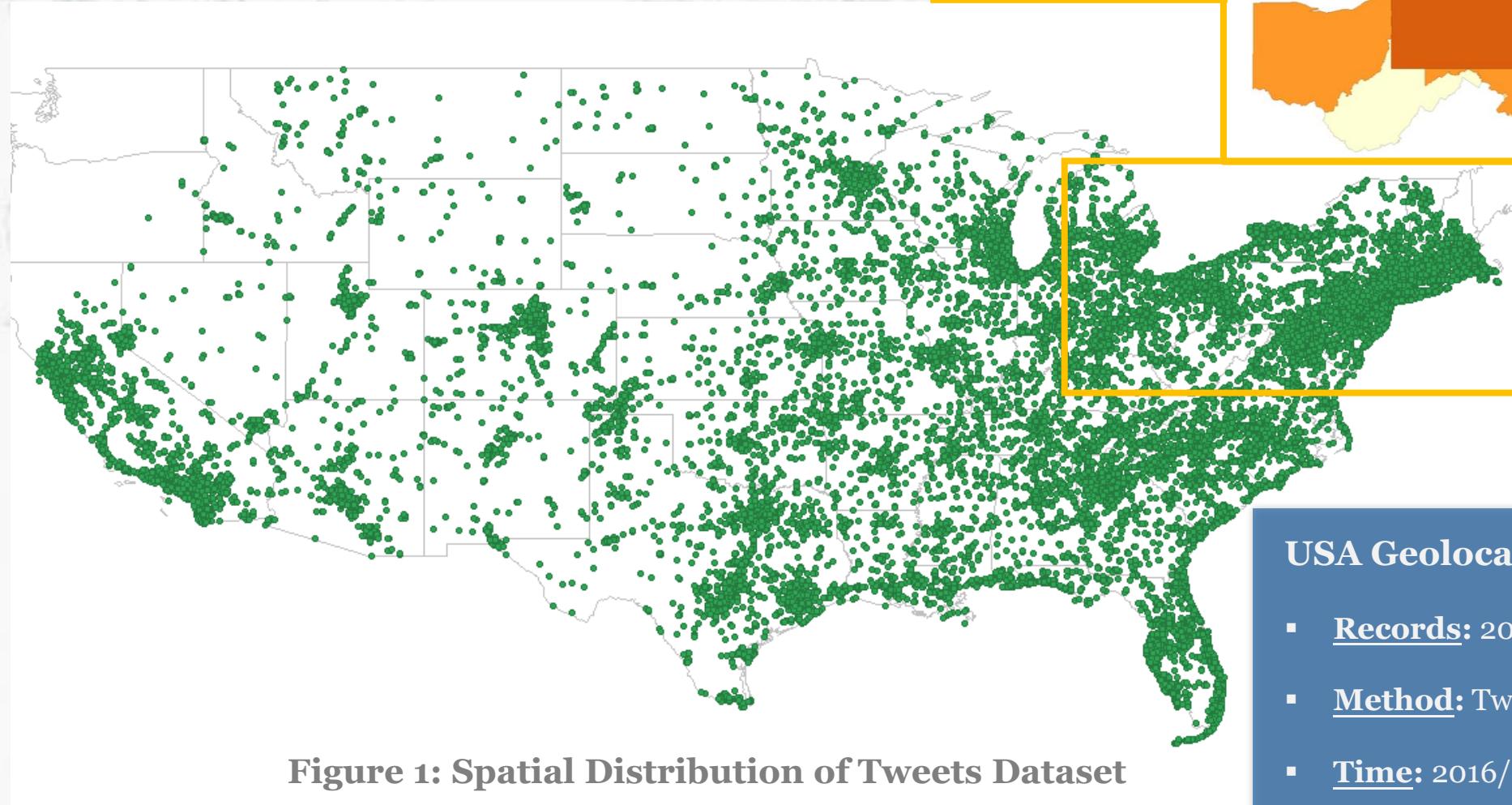
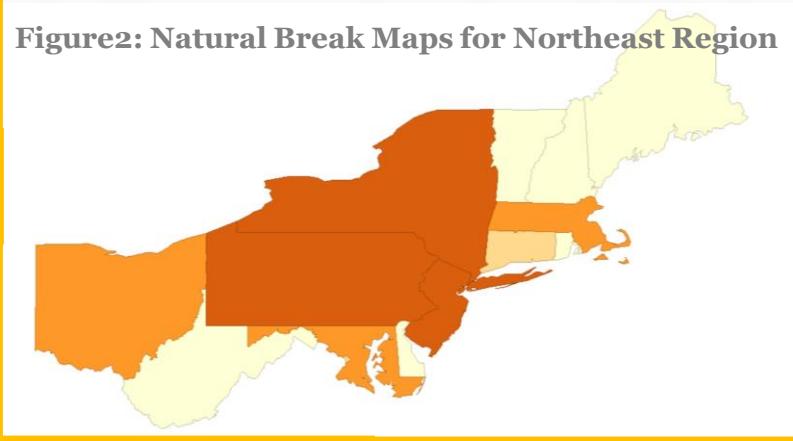


Figure2: Natural Break Maps for Northeast Region

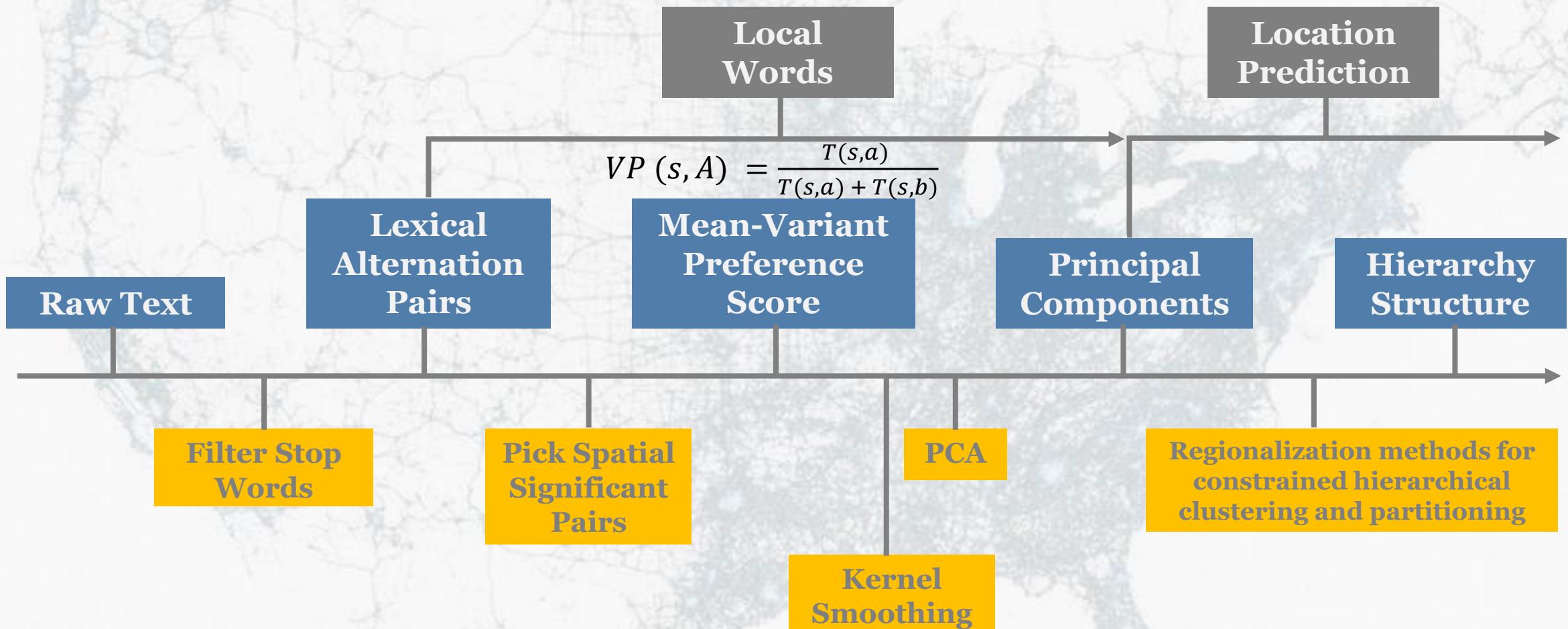


### USA Geolocated Twitter Dataset

- Records: 204,820 observations
- Method: Twitter API
- Time: 2016/04/14-16
- Source: <http://followthehashtag.com/>

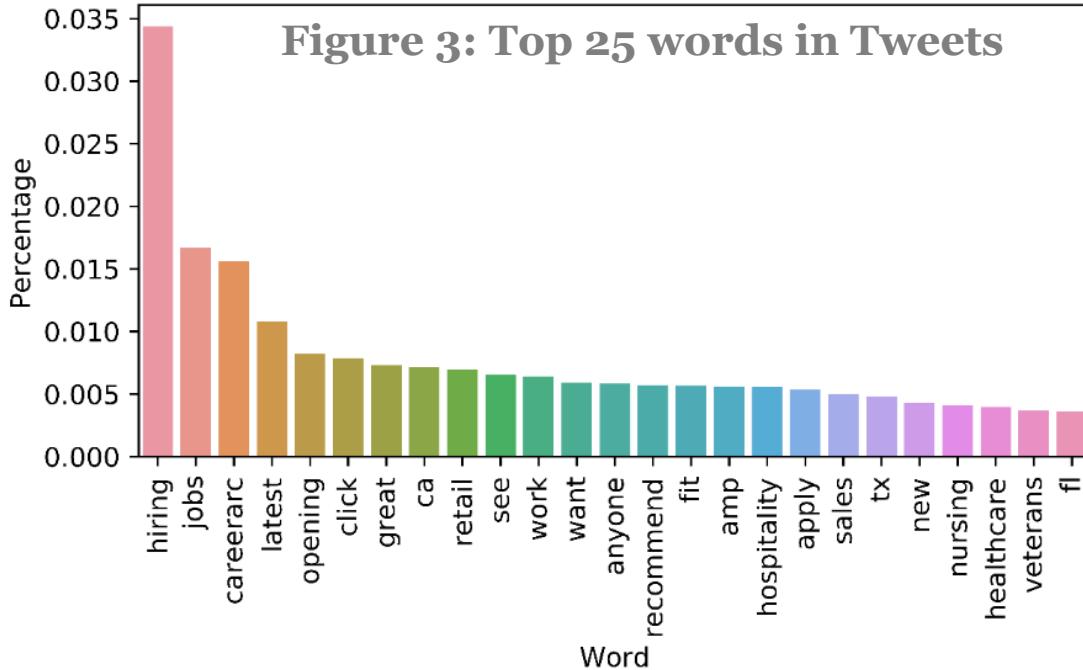


# Method

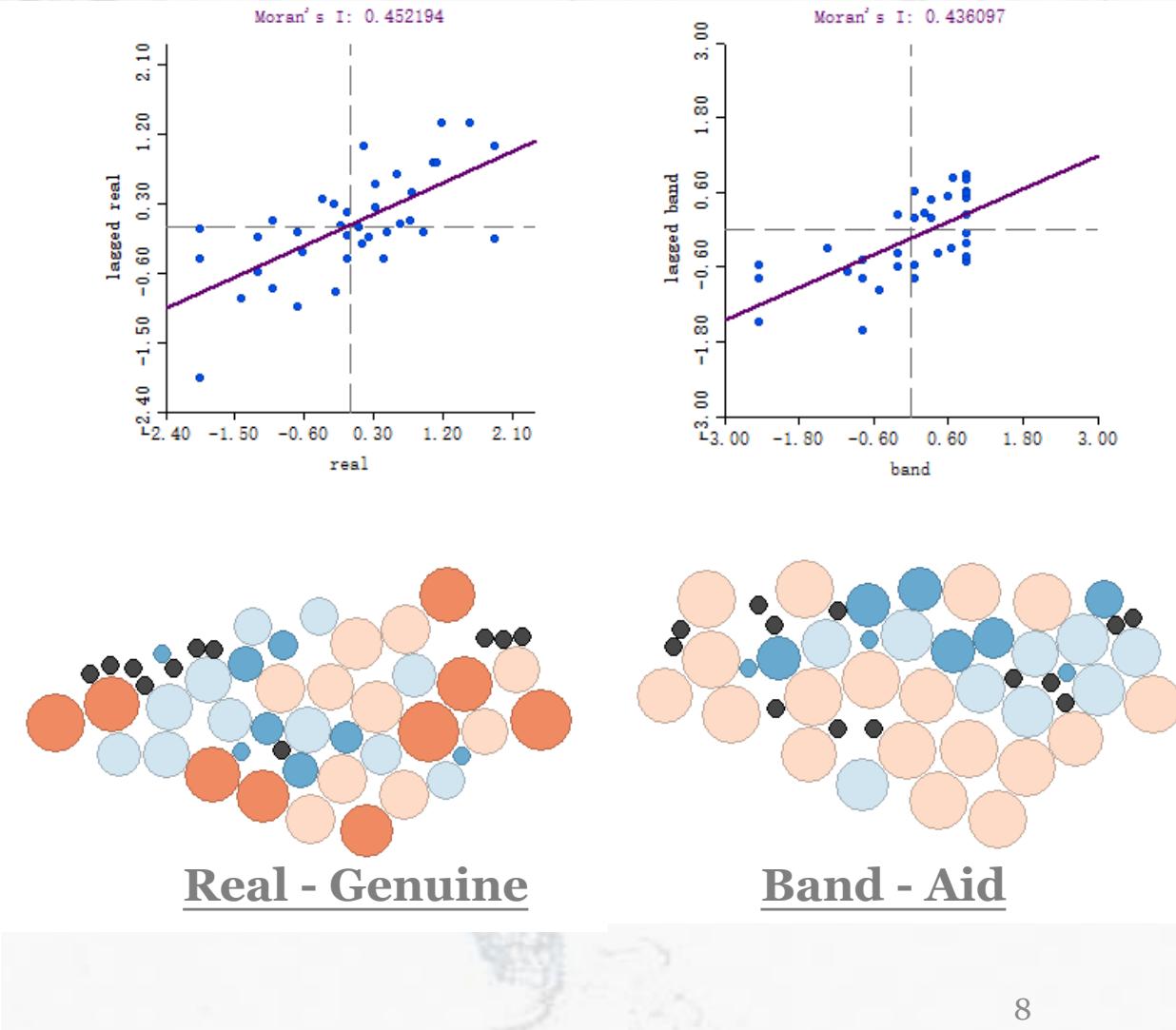




# Results



	Total Variances Explained	Singular Values
PC1	0.97181456	99.11797571
PC2	0.00711633	8.48181549
PC3	0.00201042	4.50821371
PC4	0.00178227	4.2447043
PC5	0.00168624	4.12876605





# Discussion

## Issue

- **Data Sparsity**
  - Not enough data for each user or county
  - Might overlook existing patterns
- **What to include in local words**
  - Too similar for each state
  - Prediction model failed
- **We are where we tweet?**
  - Geo-locations can cheat
  - Reply on self-report

## Contribution

- Used spatial methods to study linguistics topics
- The Northeast region does have prefer words compared to the country



---

# Spatial Linguistics Variations in Northeastern U.S. Based on Geo-Tagged Tweets

---

Thank you!



Nice work!

Who is up next?

# THE “MERE REMINDER” EFFECT OF VISUALLY SALIENT CALORIE LABELING

Ling Dai

Thesis Advisor: Oleg Urminsky

## BACKGROUND

- Preventing obesity at the population level has been a challenge for policy making
- Recent focus of policy making on calorie labeling:
  - Many policy makers believe that calorie labeling reduce people's calorie consumption through providing information
- Inconsistent results of the effectiveness of calorie labeling from past studies:
  - Non-restaurant settings: (-58.16 kcal;  $p = 0.01$ )
  - Restaurant setting: (-6.70 kcal;  $p = 0.331$ )

## RESEARCH QUESTION

- Can calorie labeling effectively reduce people's calorie consumption?
  - Average total calories per transaction
- Does effective calorie labeling reduce calorie overconsumption through providing additional information? (Or through prompting people to think about their own health and diet?)
- Through influencing choice architecture? (or spend?)
  - Avg. Calorie per USD spent
  - Choice architecture (% of bottled sugar drinks)

## DESIGN: A 9-WEEK EXPERIMENT

- At four cafeterias on the campus of the University of Chicago: Harris, GCIS, Stuart, and Law
- 4 “poster weeks”, each followed by a week of “washout” period + 1 coupon week:
  - “poster weeks”: Week 1, 3, 5, 8
  - “washout weeks”: Week 2, 4, 7, 9
  - coupon week: Week 6
- Posters are exhibited at the cafeterias on a rotational basis to eliminate the confounding effect of the cafeterias

## 4 SETS OF SIGNAGE

### Do you know?\*

Total Per-Meal (3 meals per day)  
**Calorie recommended** is typically between  
**650 to 800 Calories.**

\*Depends on age, gender, and activity level. US Department of Agriculture and US Department of Health and Human Services, Washington, 7th ed., 2010

Signage 1

### Do you know?\*



Albacore Tuna Wrap	<i>has</i>
<b>320 Calories</b>	
Turkey & Gouda Wrap	<i>has</i>
<b>500 Calories</b>	
Chicken Caesar Salad	<i>has</i>
<b>190 Calories</b>	

\*Source: UChicago Dining

Signage 2

### Do you know?

**Calorie information** is available for many of the pre-packaged items we carry in this café.

### Do you know?

Do you know how many **Calories** are there in your lunch today?

Signage 4

## DATA PROCESSING

CHK 254            GST 0  
1263 Carmona        142  
TRN 13/24189 FEB02'15 8:05AM  
Law Sch Cafe  
-----  
Main  
1 Coffee Sm        1.69  
XXXXXXXXXXXX0422 xx/xx  
Visa            1.85  
Subtotal        1.69  
Tax            0.16  
Total Paid      1.85  
=====



Raw Data: Electronic Checks

- Data Scraping + Parsing:
  - Subtotal
  - Item Names
  - Item Prices
  - Cafeteria
  - Check#
  - Date
- Calorie information lookup
- Ready for analysis:
  - Check-level and day-level
  - OLS and WLS linear regressions

## RESULTS: CALORIE PER TRANSACTION

	Confounders				
Variable	Model 1	Model 2	Model 3	Model 4	Model 5
<i>Intercept</i>	-10.85	-37.38**	-46.64***	-49.13***	-50.22***
<i>Subtotal</i>	72.28***	79.02***	76.75***	76.67***	76.90***
<i>2014Spring</i>	1.30	2.11	13.98***	12.59***	12.59***
<i>2014Winter</i>	11.22***	12.63***	19.21***	16.32***	16.27***
<i>Coupon</i>	9.42.	8.93	7.10	4.19	4.34
<i>Signage1</i>	-5.92	-5.30.	-5.58.	-9.21*	-9.25* 
<i>Signage2</i>	-0.26	-0.03	-0.44	-1.54	-1.68
<i>Signage3</i>	-0.86	-0.25	-0.24	-1.74	-1.84
<i>Signage4</i>	-7.06	-7.36.	-6.73.	-8.82*	-8.99* 
<i>R</i> <sup>2</sup>	0.821	0.846	0.885	0.892	0.894

\* $p < .05$  \*\* $p < .01$  \*\*\* $p < .001$

## RESULTS: CALORIE PER USD SPENT

Variable	Confounding Variables				
	Model 1	Model 2	Model 3	Model 4	Model 5
<i>Intercept</i>	69.9693***	71.027***	67.1724***	66.2732***	66.282***
<i>2014Spring</i>	0.0939	0.132	2.4906***	2.2266***	2.201***
<i>2014Winter</i>	2.2269***	2.318***	3.6558***	2.9821***	2.943***
<i>Coupon</i>	2.1768.	1.963.	1.6192	1.0654	0.995
<i>Signage1</i>	-1.3487	-1.271	-1.3319.	-2.0912*	-2.13* 
<i>Signage2</i>	0.0701	0.216	0.0556	-0.0888	-0.154
<i>Signage3</i>	-0.2351	-0.206	-0.1743	-0.364	-0.405
<i>Signage4</i>	-1.6327	-1.663.	-1.5242.	-1.8547*	-1.912* 
<i>R</i> <sup>2</sup>	0.111	0.206	0.414	0.454	0.462

\**p* < .05 \*\**p* < .01 \*\*\**p* < .001

## OTHER RESULTS

- Sales / Spend: None of the posters had a statistically significant influence on either total sales of the cafeterias or people's average level of spend during exhibition
- % of bottled sugary drinks: no significant change
  - The choice architecture of hot food / snacks? (follow-up analysis)

## DISCUSSION

- Conclusion:

Effective labeling works primarily as a reminder, by prompting people to consider nutrition rather than by providing new information.

- Implication on policy making:

Instead of focusing on the accuracy of the informational content, providing approximate but salient calorie information may be more effective.

Nice work!

Who is up next?

# Precautionary Labor Supply: Calibrated Income Shocks and Age Heterogeneity

Zunda Xu

The University of Chicago

M.A. in Computational Social Science

*[zunda@uchicago.edu](mailto:zunda@uchicago.edu)*

Advisor: Dr. Richard Evans

April 19, 2019

# Introduction

## Facts:

- ① People don't seem to save enough:

"The thriftlessness of early times was in great measure due to the want of security that those who made provision for the future would enjoy it" (Marshall, Alfred (1920))

- ② Old people save more than young people:

A potential way for young people to self-insure - increasing labor supply in early period.

## Research Question:

Do people substitute potential labor supply for savings as insurance against future income shocks?

## Hypothesis:

When young, people use latent potential labor as "precautionary savings" and need less actual savings.

## Literature related with Endogenous Labor Supply:

### Theoretical Analysis:

- ① Eaton and Rosen (1980): Future labour supply can increase in response to increased wage uncertainty if risk aversion is sufficiently high.
- ② Bodie et al. (1992): Analyze how labour-supply flexibility influences investors' portfolio decisions and find that labor supply flexibility raises precautionary motives when wages are stochastic.
- ③ Floden (2006): Labour-supply flexibility tends to raise saving when future wages are uncertain and that future wage uncertainty tends to raise current labour supply and future leisure.

## Empirical Studies:

- ① Low (2004) uses a calibrated model finds that young workers with much unresolved wage uncertainty work longer hours than old workers with little remaining wage uncertainty.
- ② The empirical relationship between risk and hours of work has been documented to be positive for self-employed men in the USA (Parker et al., 2005), male employees in the USA (Kuhn and Lozano, 2008), and German and US workers (including self-employed) of both sexes (Bell and Freeman, 2001).
- ③ Jessen et al. (2018) conducted an empirical study to quantify the importance of precautionary labour supply and they find that individuals choose an additional 2.8% of their hours of work (i.e. about one week per year) to shield against wage shocks.

# General Model

A model with  $s$ -period-lived agents, endogenous labor supply and stochastic ability.

## Stochastic Ability

Initially, each individual is randomly assigned one of  $J$  discrete ability types  $e_{j,t} \in \epsilon = \{e_1, e_2, \dots, e_J\}$ .

Define a Markov transition matrix  $\Pi(e_{k,t+1}|e_{j,t})$  that gives the probability of being ability type  $e_{k,t+1}$  next period given that you are type  $e_{j,t}$  today with  $k, j \in 1, 2, \dots, J$ .

$$\Pi(e_{k,t+1}|e_{j,t}) = \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \dots & \pi_{1,J} \\ \pi_{2,1} & \pi_{2,2} & \dots & \pi_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{J,1} & \pi_{J,2} & \dots & \pi_{J,J} \end{bmatrix}$$

# General Model

## Disutility of Labor

Model the disutility of labor using a constant Frisch elasticity (CFE) functional form.

$$u(c_t, n_t) = \frac{c_t^{1-\sigma} - 1}{1 - \sigma} - \chi^n \frac{\gamma n_t^{1+\frac{1}{\gamma}}}{1 + \gamma}$$

where  $\sigma \geq 1$  is the coefficient of relative risk aversion on consumption and  $\gamma \geq 0$  is the Frisch elasticity of labor supply.

## Household Problem Budget Constraint

$$c_{s,t} + a_{s+1,t+1} = R_t a_{s,t} + w_t e_{j,t} n_{s,t} \quad \forall j, s, t$$

$$\text{where } a_{1,t}, a_{S+1,t} = 0 \quad \forall t$$

## General Model

The household will choose a sequence of consumption and labor supply to maximize lifetime utility:

$$\max_{\{c_t, n_t\}_{t=0}^T} \sum_{t=0}^T \beta^t u(c_t, n_t)$$

The Bellman equation for this problem is:

$$V_t(e_t, a_t) = \max_{\{c_t, n_t\}} u(c_t, n_t) + \beta E_t V_{t+1}(e_{t+1}, a_{t+1})$$

The FOCs (after applying the envelope conditions) are:

$$u_1(c_t, n_t) = \beta E_t R_{t+1} u_1(c_{t+1}, n_{t+1}) \quad \forall t$$

and

$$e_t w_t u_1(c_t, n_t) = -u_2(c_t, n_t) \quad \forall t$$

# Computational Methodology

Generalized Endogenous Grid Method (Barillas and Fernandez-Villaverde (2007))

1. Create a grid for  $a_T$  and  $e_T$
2. Use the FOC for the choice of labor supply to find  $n_T = n(c_T)$ .
3. Plug  $n(c_T)$  into the budget constraint. Use a root finder to solve for  $c_T$  for every point on the  $(a_T, e_T)$  grid.
4. Interpolate  $c_T(a_T, e_T)$  so that have a function for  $c_T$  for all points.
5. Use the FOC for the choice of  $a_T$  to find  $c_{T-1}$ .
6. Use  $n(c_{T-1})$  to analytically solve for  $n_{T-1}$  for all points on the  $(a_T, e_{T-1})$  grid.
7. Use the budget constraint to solve for a grid of  $a_{T-1}$  that is endogenous to the choice of  $a_T$ .
8. Interpolate  $c_{T-1}(a_{T-1}, e_{T-1})$  so that have a function for  $c_T$  for all points.

# Solutions

## Analytical Solution:

To simplify, we solve a two-period models with four different settings:

- (1) deterministic income & exogenous labor supply
- (2) deterministic income & endogenous labor supply
- (3) stochastic income & exogenous labor supply
- (4) stochastic income & endogenous labor supply

## Numerical Solution:

Parameters	Description	Value
$S$	Life Periods	80
$\beta$	Discount Factor	0.96
$\sigma$	CRRA Coefficient	2.2
$\gamma$	Frisch Elasticity	0.9
$\chi^n$	CFE constant	10

Table: Model Calibration

# Key Results and Conclusions

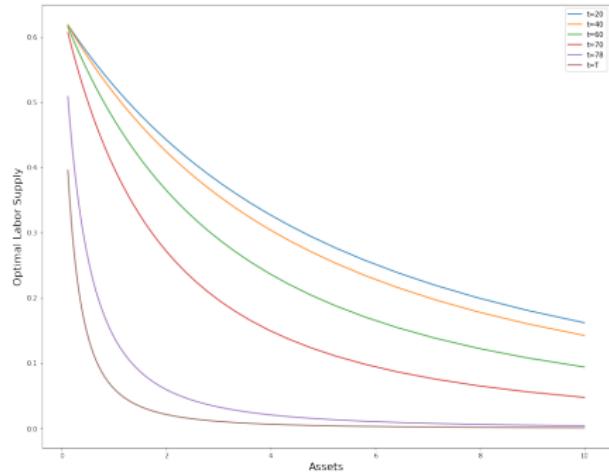


Figure: Different Period's Labor Supply

## Conclusions

Young generations have less actual savings when labor supply are flexible, and they provide more labor supply compared with old generations.

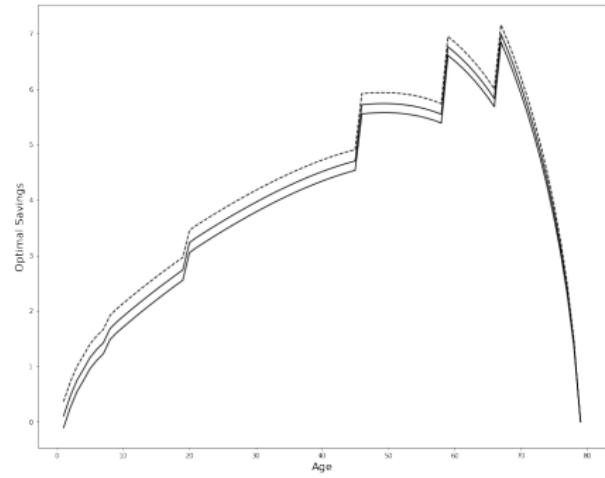


Figure: Optimal Savings over Ages

Nice work!

Who is up next?

# The Range of Interaction in Spatial Autoregressive Econometric Models

Ruxin Chen

Thesis Submitted for Master of Computational Social Science

Supervised by Pr.Luc Anselin

2019

# Research Question

- Motivation
  - Tobler's first law of geography: everything is related to everything else, but near things are more related than distant things.
  - Modern spatial econometric models take into account of the correlation for units that are closely located. The range of interaction of models indicates how far two units should be accounted for their impact on each other.
  - Commonly used spatial econometric models rely heavily on the assumption of the range of interaction. However, the spatial dependence structure is an unknown priori.
- Research Question
  - Conduct Monte Carlo simulations to find the effect on inference if the model is mis-specified
  - Summarize the discussion in the most recent literature related to specification and estimation of spatial dependence structure

# Spatial Econometric Model 1: SAR

- The Spatial Autoregressive Model (SAR)

- Model specification

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (1)$$

- Reduced form

$$\mathbf{y} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta} + (\mathbf{I} - \rho \mathbf{W})^{-1} \boldsymbol{\epsilon} \quad (2)$$

- Leontief expansion

$$(\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta} = \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{X} \boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \mathbf{X} \boldsymbol{\beta}^2 + \rho^3 \mathbf{W}^3 \mathbf{X} \boldsymbol{\beta}^3 + \dots \quad (3)$$

- Global Interaction: the value of  $y$  at location  $i$  is determined by the value of  $x$  at location  $i$  and all other locations through their dependence with location  $i$  such dependence structure is specified by the spatial weight matrix  $\mathbf{W}$ .

# Alternative Autoregressive Models

- The Spatial Error Model (SEM)
  - Model specification

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \quad (4)$$

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + \epsilon \quad (5)$$

- The Spatial Durbin Model (SDM)
  - Model specification

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{W}\mathbf{X}\gamma + \epsilon \quad (6)$$

- Problem of Identification

# The Spatial Econometric Model 4: SLX

- The Spatial Lag Model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{W}\mathbf{X}\gamma + \epsilon \quad (7)$$

- Unlike models mentioned previously, SLX imply local interaction: spatial effect is only limited to direct neighbors
- Advantage over SAR
  1. No endogeneity problem. Hence, it can be estimated directly using OLS
  2. Furthermore, it allows the spatial weight matrix to be parameterized – easy to combine with more complicated methods
  3. Parameters are easy to interpret

# Problems for Spatial Autoregressive Models

- Unknown spatial dependence structure
  - Spatial weight matrix should reflect the prior information corresponding to a specific research question, the applied work mainly follows the same routine of specifying this matrix, such as using contiguity, k nearest neighbors (KNN) or some other heuristic procedures, without considering the idiosyncracy for each case
  - Symmetric spatial weight matrix ?
  - Different spatial models imply different assumption on the range of interaction, but it is hard to distinguish them empirically
- Economic Interpretation
  - The omitted variables often are highly correlated over space. McMillen(2012) recognized that SAR is just a form of spatial smoothing and used as a panacea for model misspecification issue

# Monte Carlo Simulation

- What if choose an incorrect spatial model?
  - Specify spatial autoregressive parameters  
 $\rho = (0.1, 0.3, 0.5, 0.7, 0.9, 0.95)$  and  $\frac{\sigma_x}{\sigma_u} = (1, 2, 4)$
  - Conduct 10,000 Monte Carlo simulations for each case
    1. SLX-SAR, SAR-SLX
    2. SDM-SAR, SAR-SDM
    3. SDM-SLX, SLX-SDM
    4. SEM-SDM, SEM-SAR, SEM-SLX
  - Randomly generate  $\mathbf{X}$  and  $\mathbf{u}$ , generate  $\mathbf{y}$  using the reduced form specification for each model
  - Geometry: US counties (3085 counties, mainland excluding Alaska). Generate spatial weight matrix using Queen, Rook contiguity and Block by State
  - Test nulls

$$H_0 : \theta_i = \theta_{i0} \quad (8)$$

Compare the rejection probability with the true DGP

# Simulation Result: SAR-SLX

- Empirical distribution for  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  and the probability of rejection for nulls

Figure A.6: Estimates for  $\beta_1$  for SAR-SLX, Queen Contiguity

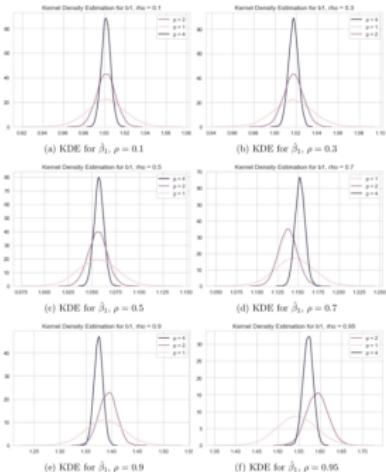


Figure A.5: Estimates for  $\beta_0$  for SAR-SLX, Queen Contiguity

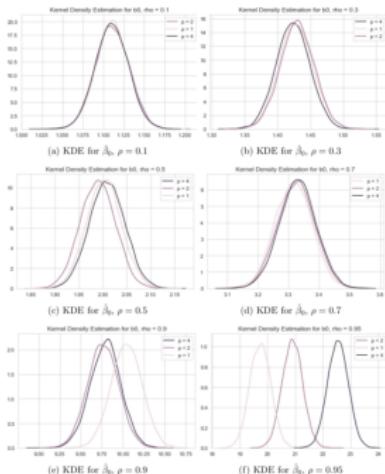


Table A.2: Parameter Estimates: SAR-SLX

	True SAR, Estimated SLX											
	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$	$\beta_0$	$\beta_1$	$\gamma$	$\beta_0$	$\beta_1$	$\gamma$	$\beta_0$	$\beta_1$	$\gamma$
Queen	$\sigma_x = 1$	1.000	0.052	0.689	1.000	0.151	1.000	1.000	0.793	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	0.066	1.000	1.000	0.970	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 1$	1.000	0.050	0.999	1.000	0.486	1.000	1.000	0.890	1.000	1.000	1.000
Rook	$\sigma_x = 2$	1.000	0.054	0.999	1.000	0.437	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	0.066	1.000	1.000	0.975	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 1$	1.000	0.051	0.126	1.000	0.049	0.795	1.000	0.072	0.998	1.000	1.000
Block	$\sigma_x = 2$	1.000	0.050	0.459	1.000	0.056	0.970	1.000	0.135	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	0.051	0.751	1.000	0.070	1.000	1.000	0.421	1.000	1.000	1.000
	$\rho = 0.7$	$\rho = 0.9$	$\rho = 0.95$	$\beta_0$	$\beta_1$	$\gamma$	$\beta_0$	$\beta_1$	$\gamma$	$\beta_0$	$\beta_1$	$\gamma$
Queen	$\sigma_x = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 1$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Rook	$\sigma_x = 2$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 1$	1.000	0.251	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000	1.000
Block	$\sigma_x = 2$	1.000	0.766	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	$\sigma_x = 4$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

# Simulation Result: SLX-SAR

- Empirical distribution for  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  and the probability of rejection for nulls

Figure A.1: Estimates for  $\beta_0$  for SLX-SAR, Rook Contiguity

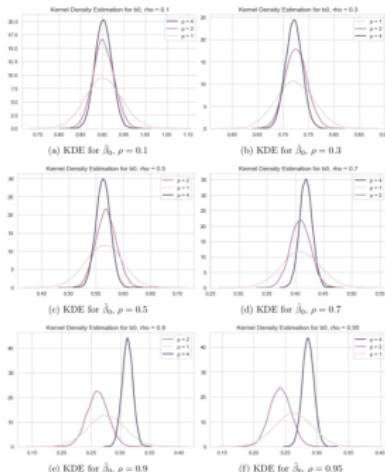


Figure A.2: Estimates for  $\beta_1$  for SLX-SAR, Rook Contiguity

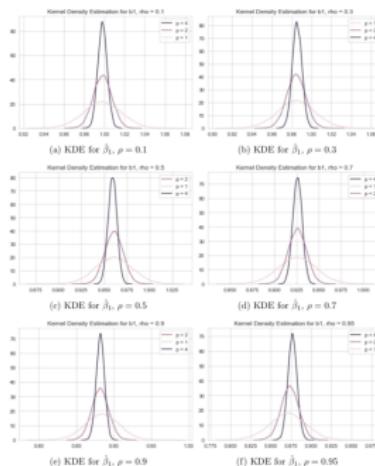


Table A.1: Parameter Estimates: SLX-SAR

	True SLX, Estimated , SAR			
	$\gamma = 0.1$	$\gamma = 0.3$	$\gamma = 0.5$	
$\beta_0$	$\beta_0$	$\beta_0$	$\beta_0$	
Queen	$\sigma_x = 1$	$0.591$	$0.053$	$0.657$
Rook	$\sigma_x = 2$	$0.960$	$0.055$	$0.998$
Block	$\sigma_x = 4$	$0.998$	$0.070$	$1.000$
Queen	$\sigma_x = 1$	$0.597$	$0.051$	$0.668$
Rook	$\sigma_x = 2$	$0.960$	$0.051$	$0.997$
Block	$\sigma_x = 4$	$0.998$	$0.067$	$1.000$
Queen	$\sigma_x = 1$	$0.100$	$0.049$	$0.102$
Rook	$\sigma_x = 2$	$0.308$	$0.047$	$0.277$
Block	$\sigma_x = 4$	$0.633$	$0.049$	$0.708$
$\beta_1$	$\beta_1$	$\beta_1$	$\beta_1$	
Queen	$\sigma_x = 1$	$1.000$	$0.950$	$1.000$
Rook	$\sigma_x = 2$	$1.000$	$1.000$	$1.000$
Block	$\sigma_x = 4$	$1.000$	$1.000$	$1.000$
$\rho$	$\rho$	$\rho$	$\rho$	
Queen	$\sigma_x = 1$	$1.000$	$1.000$	$1.000$
Rook	$\sigma_x = 2$	$1.000$	$1.000$	$1.000$
Block	$\sigma_x = 4$	$1.000$	$1.000$	$1.000$
$\sigma_x$	$\sigma_x$	$\sigma_x$	$\sigma_x$	
Queen	$\sigma_x = 1$	$1.000$	$1.000$	$1.000$
Rook	$\sigma_x = 2$	$1.000$	$1.000$	$1.000$
Block	$\sigma_x = 4$	$1.000$	$1.000$	$1.000$

# Conclusion

- Concluding Remark
  - Models do not deviate severely when the spatial dependence is weak, however, a misspecified model can severely over-reject the null hypothesis when the dependence is sufficiently strong
  - Parameter estimates for the slope coefficient  $\beta_1$  is in general more robust to model misspecification than  $\hat{\beta}_0$ .
  - The SLX models are unable to fully capture the global effects in most spatial autoregressive framework, i.e. SAR and SDM. However, this distortion might be mitigated if the spatial weight matrix reflects the global feature.

Nice work!

Who is up next?

# USING SATELLITE IMAGERY AND CONVOLUTIONAL NEURAL NETS TO UNDERSTAND SLUM MORPHOLOGY – IN LAGOS, NIGERIA

By Cooper Nederhood

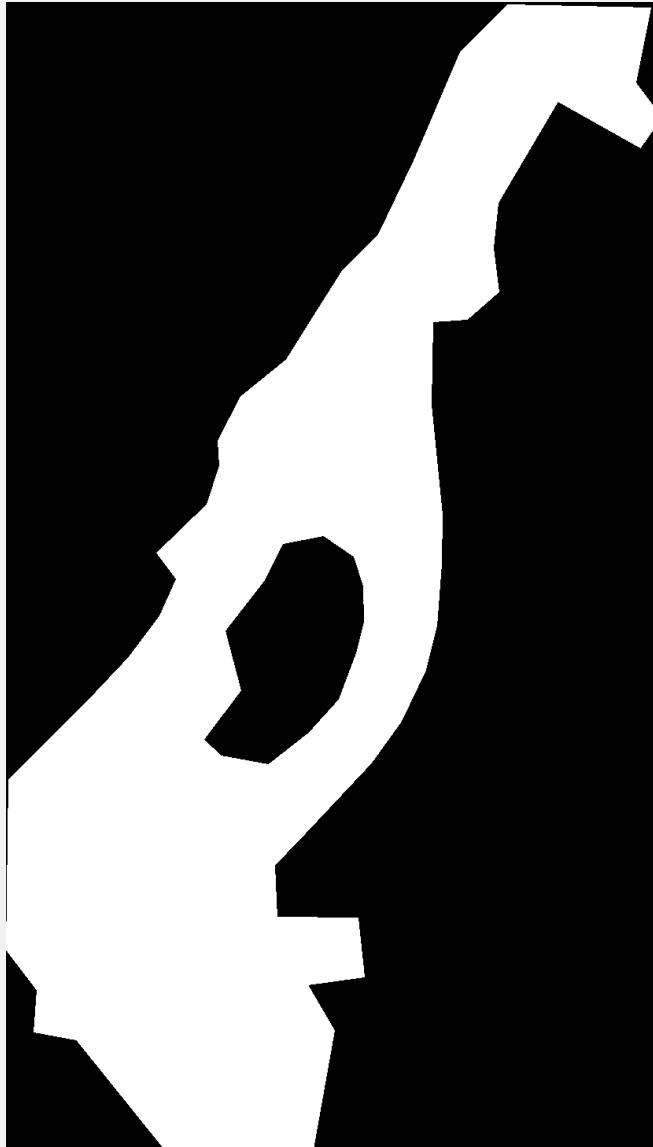
Advisor – Luis Bettencourt

# IDENTIFYING AND MAPPING SETTLEMENTS FROM SATELLITE IMAGERY

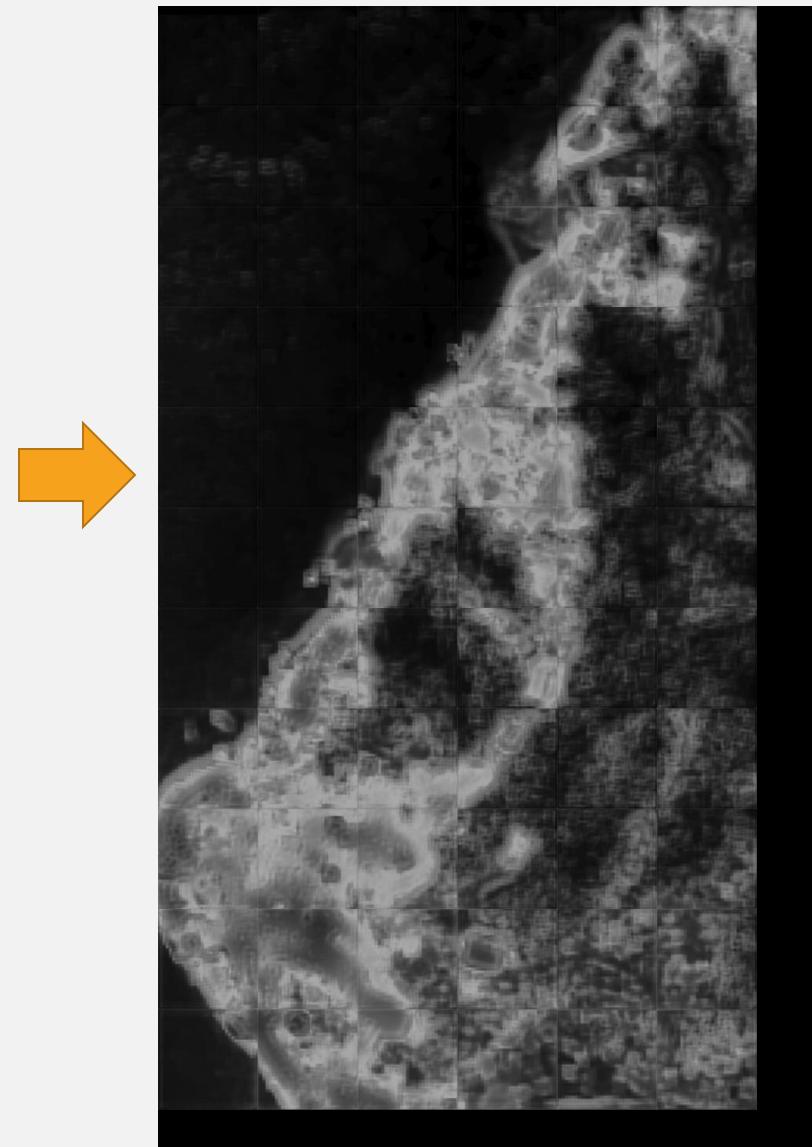
Input Image



Ground Truth Boundaries



Estimated Boundaries

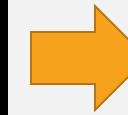
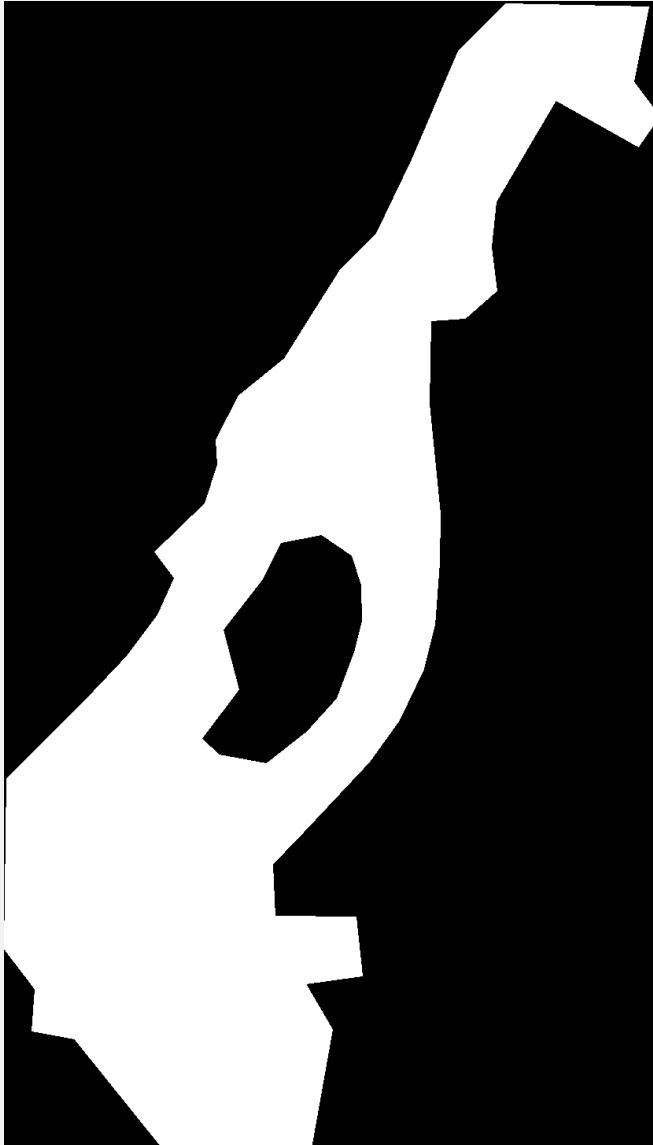


# IDENTIFYING AND MAPPING SETTLEMENTS FROM SATELLITE IMAGERY

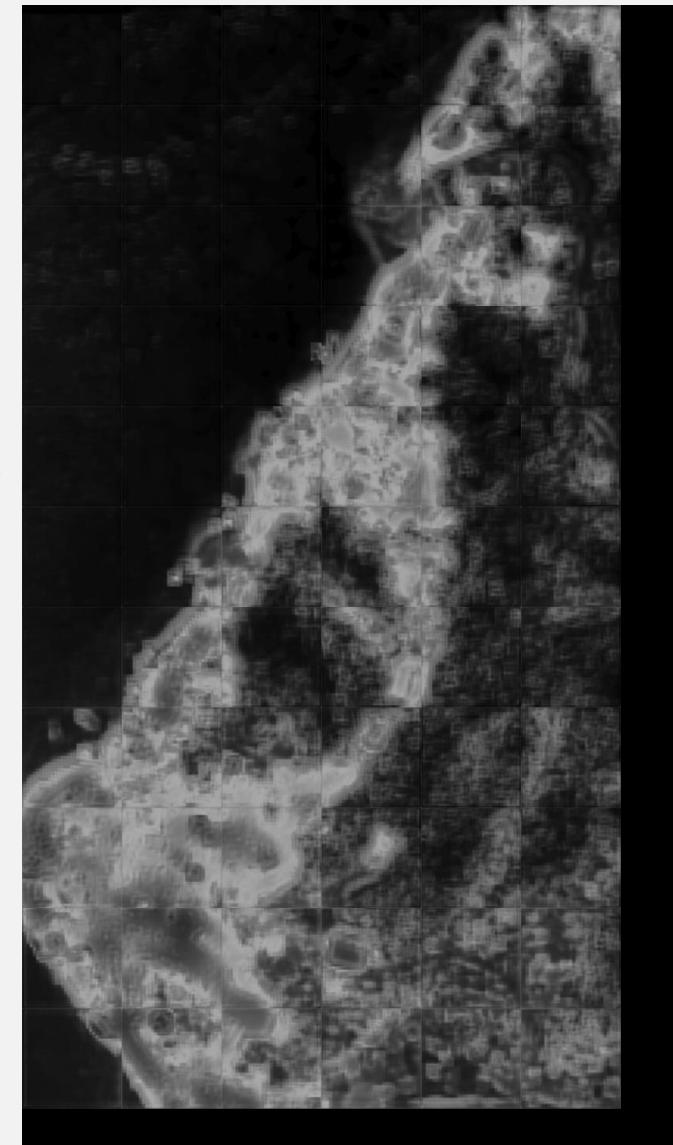
Input Image



Ground Truth Boundaries



Estimated Boundaries



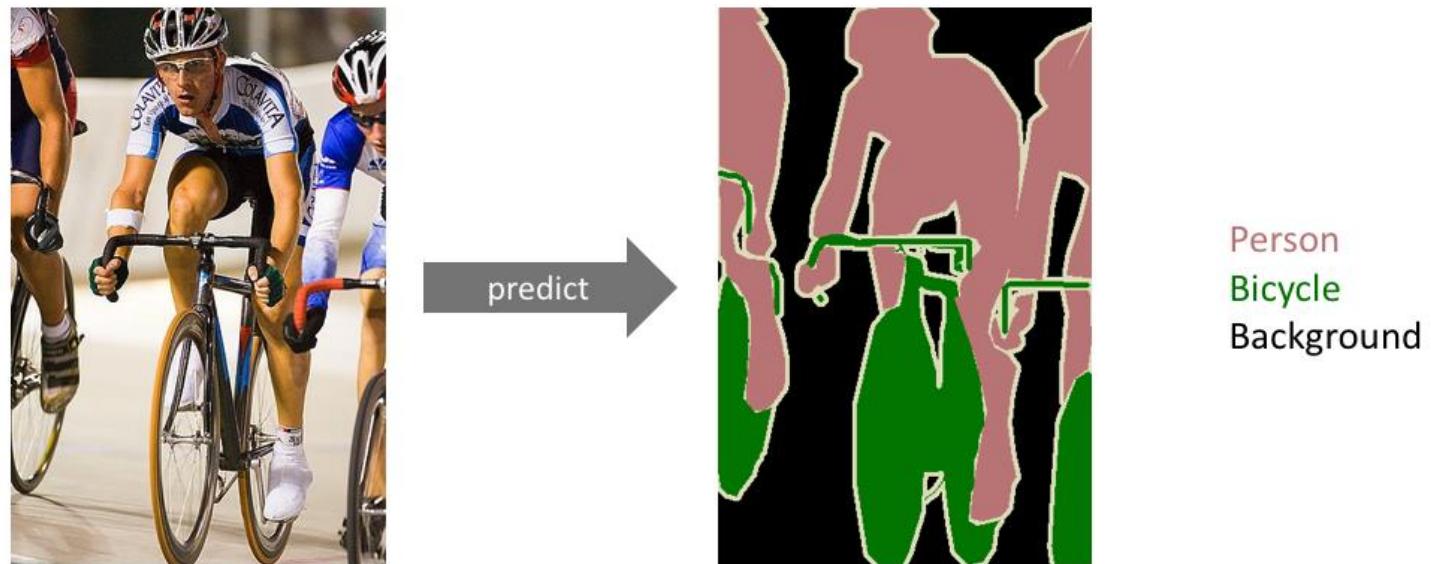
# MOTIVATION

- Most of the world lives in urban areas now
- Rapid urbanization in emerging economies like Lagos, Nigeria
- Mapping can help gain secure land tenure!!!



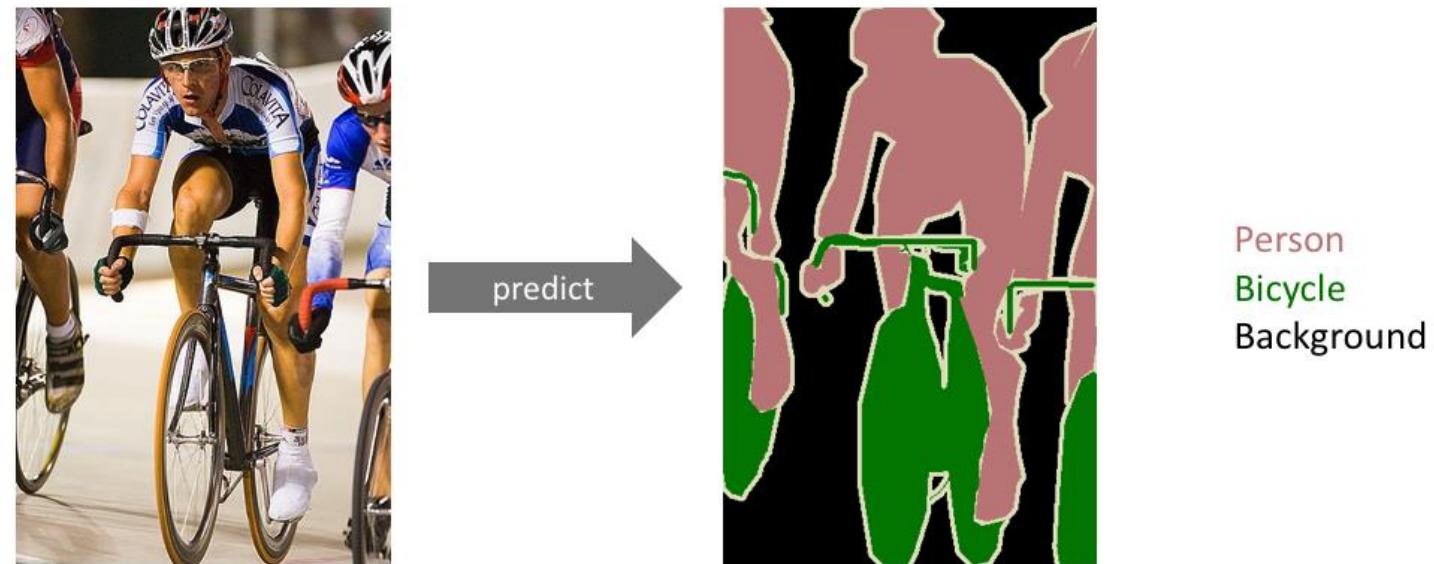
# METHOD: SEMANTIC SEGMENTATION

- Semantic segmentation – predicting a discrete classification for each pixel in the input image
- Use Convolutional Neural Networks
- Original computer vision imagery is very different than satellite imagery
  - To the side vs overhead
  - Large object vs many small objects



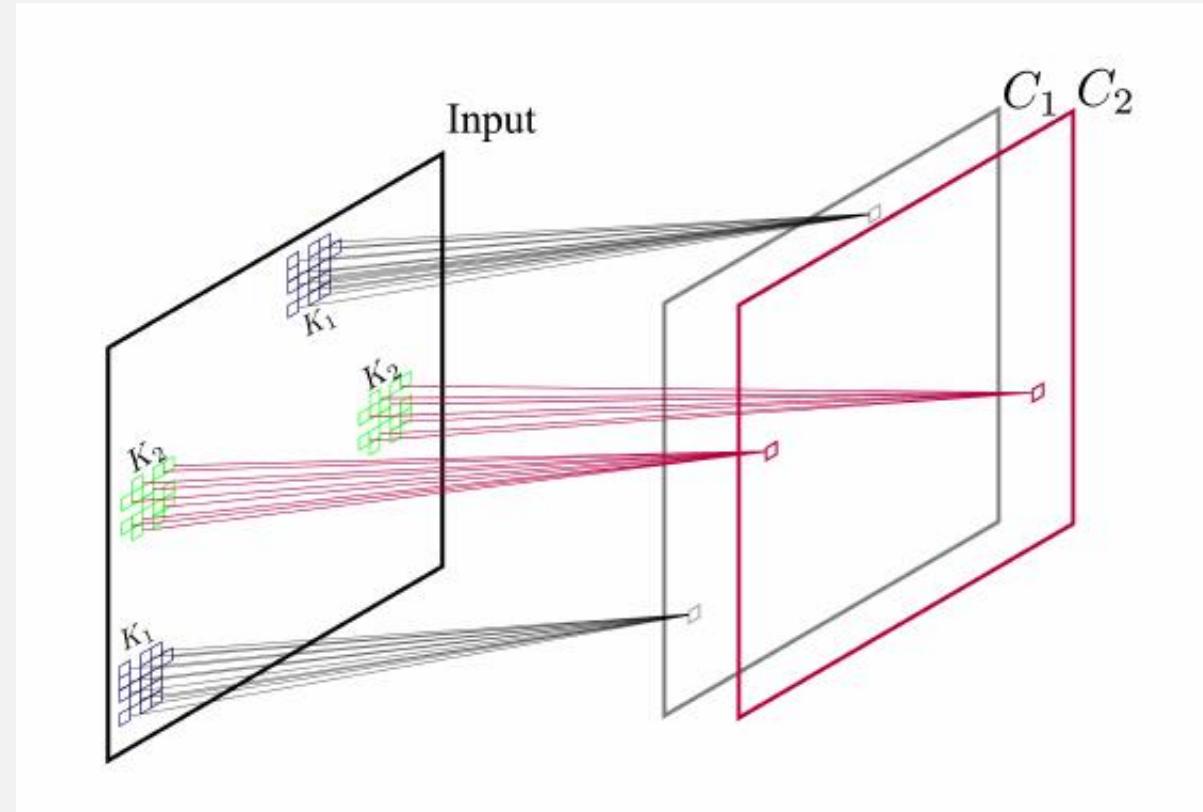
# METHOD – SEMANTIC SEGMENTATION

- Semantic segmentation – predicting a discrete classification for each pixel in the input image
- Use Convolutional Neural Networks
- Original computer vision imagery is very different than satellite imagery
  - To the side vs overhead
  - Large object vs many small objects



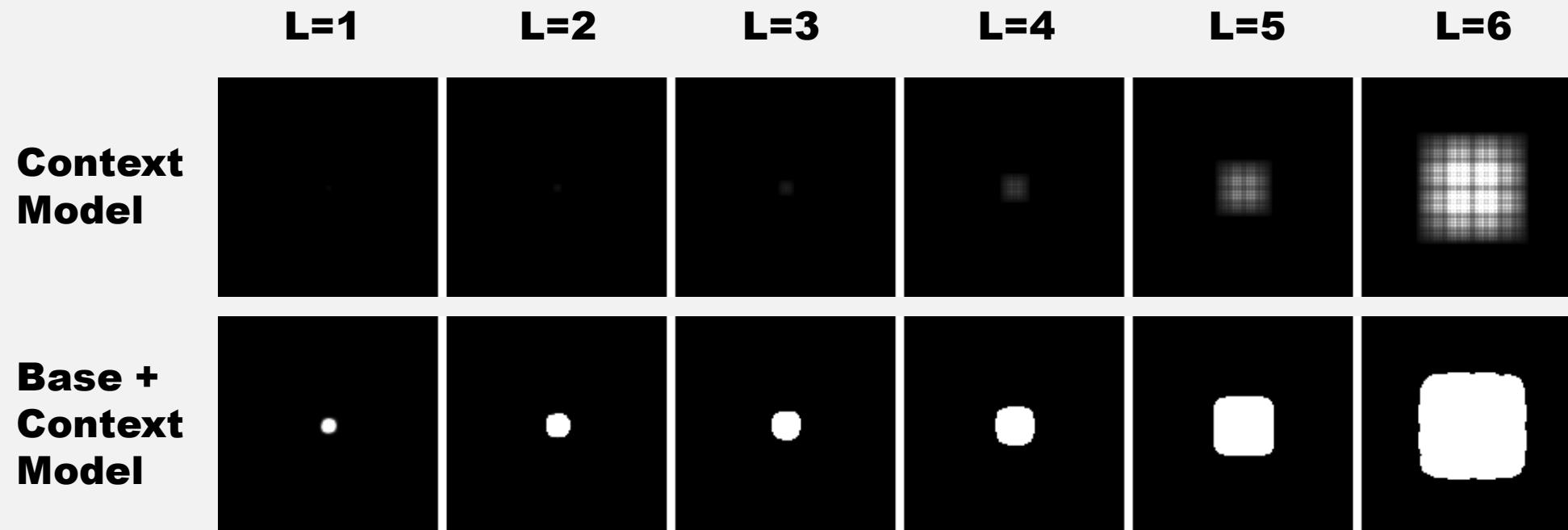
## TEST #1 - EXPANDING THE 'RECEPTIVE FIELD' OF THE SEGMENTATION MODEL

- 'Receptive field' is the set of Input pixels affecting a given Output classification
- Effective Receptive Field << Theoretical Receptive Field
- Large receptive field important in all segmentation
  - Especially important for slum urban vs non-slum urban



## TEST #1 - EXPANDING THE 'RECEPTIVE FIELD' OF THE SEGMENTATION MODEL

- Append 'Context Model' to 'Base Model' to iteratively increase the Effective Receptive Field



## TEST #2 - BANDS BEYOND THE RGB VISIBLE SPECTRUM

- Near infrared bands beyond the human visible spectrum



# CONCLUSION

- Combining deep learning with satellite imagery allows for unprecedented analysis of the built environment
- But adapting deep learning models requires specific understanding of the unique context with satellite imagery

Nice work!

Who is up next?

# SEARCHING FOR NEWS BIAS

WHAT NEWS CHARACTERISTICS PREDICT BIAS BETWEEN NEWS SOURCES?

ALEXANDER TYAN

ADVISOR: ALLYSON ETTINGER (TTIC)

PRECEPTOR: JOSHUA MAUSOLF

# PRIOR RESEARCH

- DIVERSE DEFINITIONS OF BIAS
- DIVERSE MEASURES OF BIAS
- CONFLICTING CONCLUSIONS ABOUT BIAS IN THE MEDIA

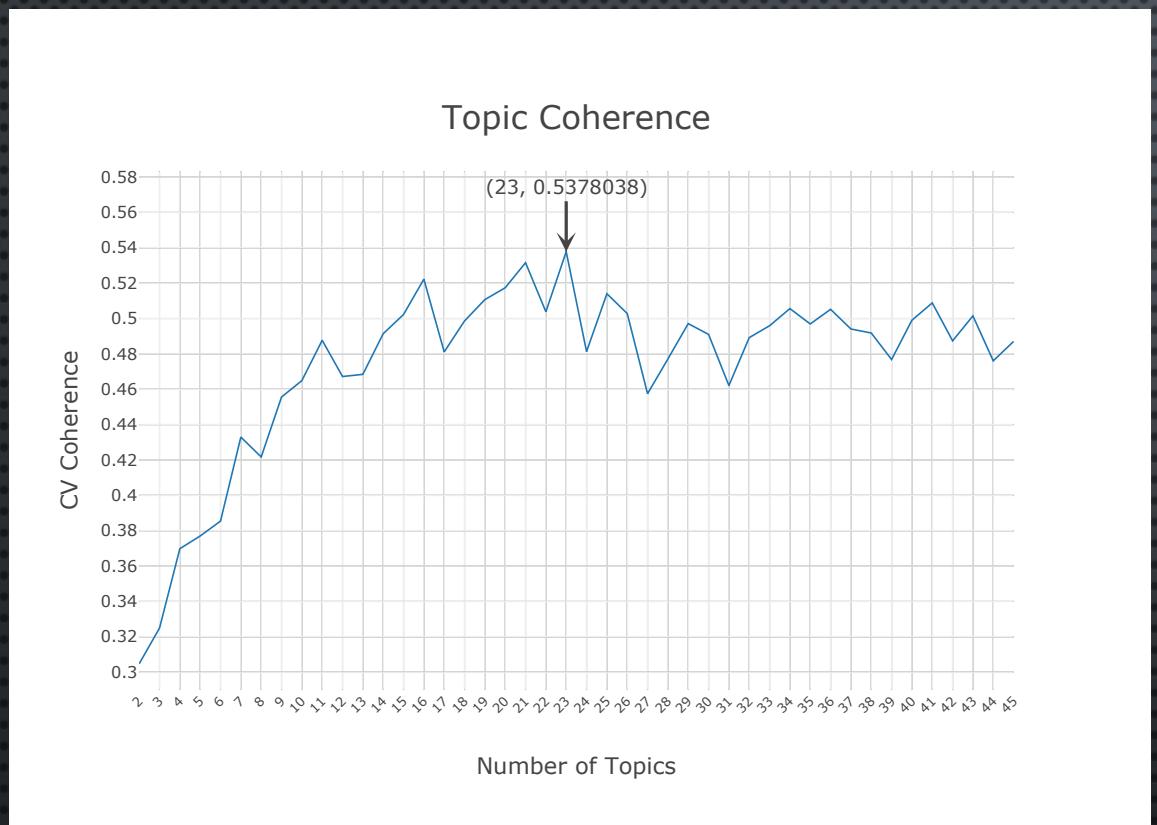
# PRIOR RESEARCH

- DIVERSE DEFINITIONS OF BIAS
- DIVERSE MEASURES OF BIAS
- CONFLICTING CONCLUSIONS ABOUT BIAS IN THE MEDIA
- HYPOTHESIS: NEWS SOURCES DISSIMILARITY CONDITIONAL ON TOPICS

# DATA AND METHODS

- BBC, CNN, RT, FOX NEWS
- RSS FEEDS
- JANUARY – APRIL 2019
- STANDARD CLEANING
- PREEMPTING DATA LEAKAGE

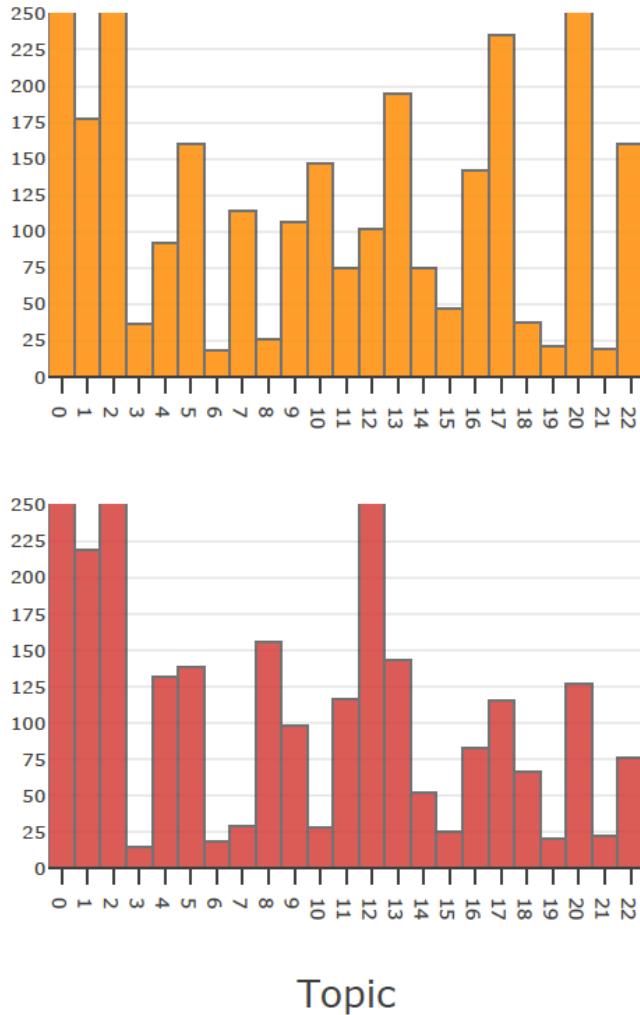
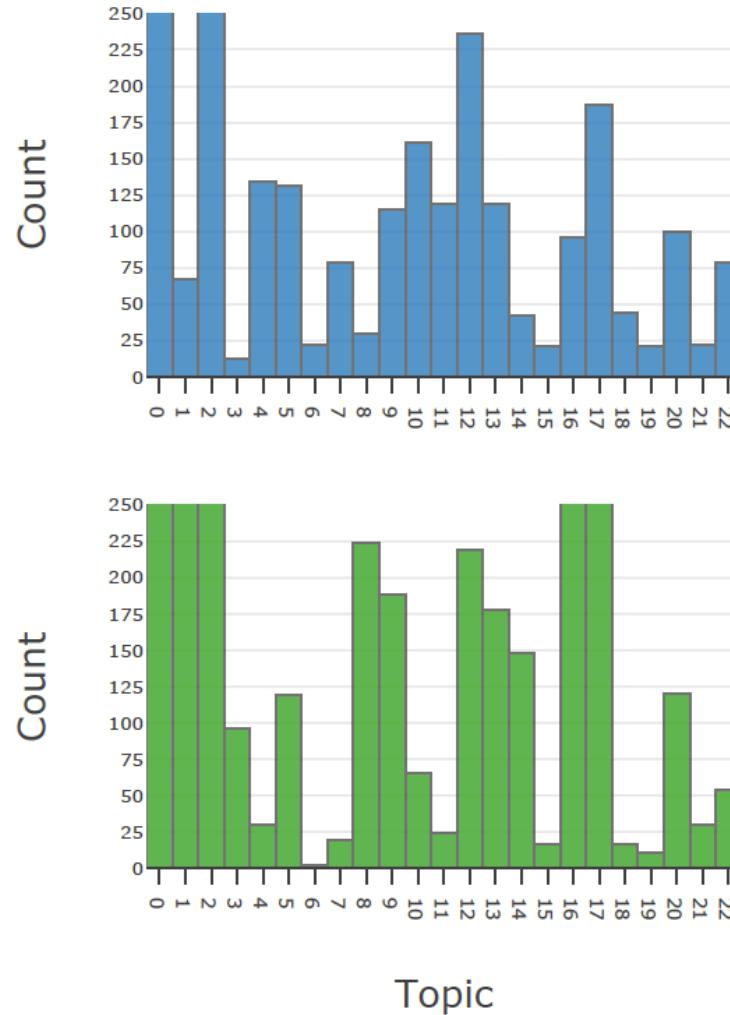
# TOPIC MODELING



→

Topic	Token	Weight
Unknown	say	0.03809549
Unknown	not	0.030164942
Unknown	do	0.017989479
...	...	...
Sports	team	0.027598321
Sports	game	0.024382351
Sports	win	0.017087057
...	...	...
Military Nuclear Technology	north_korea	0.02561758
Military Nuclear Technology	north_korean	0.018099198
Military Nuclear Technology	iran	0.015305956
...	...	...
Political Investigation (Cohen/Virginia)	say	0.050332915
Political Investigation (Cohen/Virginia)	cnn	0.024718568
Political Investigation (Cohen/Virginia)	report	0.01599263
Political Investigation (Cohen/Virginia)	tell	0.01137681
Political Investigation (Cohen/Virginia)	investigation	0.011040452
Political Investigation (Cohen/Virginia)	statement	0.010947393
Political Investigation (Cohen/Virginia)	case	0.009329448
Political Investigation (Cohen/Virginia)	public	0.008804999
Political Investigation (Cohen/Virginia)	cohen	0.008026111

## Topic Distribution by Source



CNN  
Fox News  
RT  
BBC

# TEXT VECTORIZATION AND COSINE DISSIMILARITY

- *Word2Vec GoogleNews Trained Embeddings*
- $dissimilarity = 1 - \frac{\vec{a} * \vec{b}}{|\vec{a}| |\vec{b}|}$   
$$\begin{bmatrix} 0 & .34 & \cdots & .89 & .45 \\ \vdots & & & \ddots & \vdots \\ .45 & .23 & \cdots & .76 & 0 \end{bmatrix}$$

# TEXT VECTORIZATION AND COSINE DISSIMILARITY

- *Word2Vec GoogleNews Trained Embeddings*
- $dissimilarity = 1 - \frac{\vec{a} * \vec{b}}{|\vec{a}| |\vec{b}|}$   
$$\begin{bmatrix} 0 & .34 & \cdots & .89 & .45 \\ \vdots & & & \ddots & \vdots \\ .45 & .23 & \cdots & .76 & 0 \end{bmatrix}$$

# PRELIMINARY RESULTS

topic	target source	count	mean	std	min	25%	50%	75%	max
Crime	Fox News	11160	0.829556933	0.05382656	0.546680808	0.799793571	0.835678518	0.86743091	0.983913898
	RT	6789	0.839204574	0.048338039	0.622709155	0.811861634	0.844044447	0.872817934	0.970934093
US Politics	Fox News	3848	0.883838486	0.053207771	0.616143465	0.859033167	0.89343968	0.919283092	0.986397505
	RT	11856	0.880364607	0.048124528	0.603044808	0.857095465	0.889096975	0.914100915	0.977500319

	Mean Difference	Test Statistic	p-value
CNN vs Fox News	-0.05428	-54.40851	0.000
CNN vs RT	-0.04115	-56.03709	0.000

# FURTHER STEPS

- TOPIC AND WORD EMBEDDING VALIDATION
  - NER
- COSINE-BASED BIAS -> NEURAL NETWORK SINGLE-LABEL, MULTI-CLASS CLASSIFICATION WITHIN TOPICS

That's all Folks!