

# **AD VALOREM REVENUE AND**

## **SCHOOL QUALITY:**

PREDICTING SCHOOL PERFORMANCE IN OKLAHOMA PUBLIC SCHOOLS USING NEURAL NETWORKS, PCA AND REGRESSION

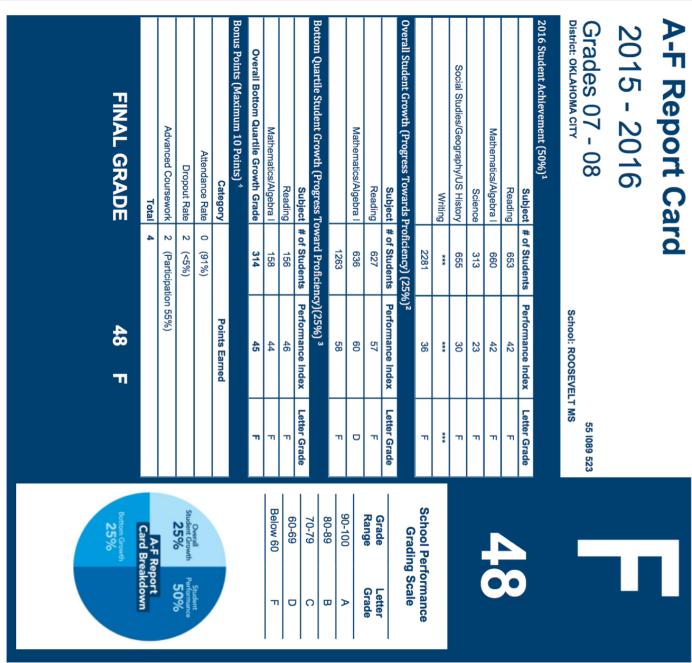
Tom Curran

MACSS Project Proposal

April 4, 2018

# RESEARCH QUESTION:

- Is a school district's ad valorem revenue the primary driver of school quality in Oklahoma?



## RESEARCH SUPPORTS CONVENTIONAL WISDOM

- **More money means more and better teachers:**
  - “Students of less effective teachers experienced reading achievement gains of one third of a standard deviation less than that of students with effective teachers. In mathematics the differences was slightly less than one half a standard deviation.” (Stronge et al 2008)
- **More money means smaller class sizes, even when quality teachers are not available**
  - *Class Size Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size (Jepsen et al 2007)*

## RESEARCH ALSO DISPROVES CONVENTIONAL WISDOM:

- **School quality and outcomes of students is not related to student teacher ratio or spending:**
  - “...studies find that, on balance, improving school resources such as the pupil-teacher ratio or per pupil spending do not improve students performance on standardized achievement tests” (Eide et al 1997)
  - *The Effects of Class Size on Student Achievement: New Evidence from Population Variation* (Hoxby 2008)
- **A students performance is not related to school quality but more reliant on parents and family:**
  - “Most previous research on effects of school has concluded that the effect of school or teacher quality on academic achievement is less than that of family background or other characteristics” (Heyneman et al 1983)
- **School quality doesn't matter at all:**
  - Does School Quality Matter? Evidence From The National Longitudinal Survey of Youth ([Betts 1995](#))

## GOALS & CONTRIBUTION OF RESEARCH:

- **Goals:**
  - Understand how different economic and sociological factors interact with each other to influence school quality
  - Create predictive model for school quality using Oklahoma's existing A-F Framework
  - Employ model to create and enhance existing policy levers
- **Research Contribution:**
  - Very little school quality program evaluation literature employs neural networks
  - Using neural networks and other theoretical approaches means model is not constrained to a subject domain
  - Captures the complex interactions of variables that eludes other predictive or explanatory models
  - Brings together diverse but related data sets to develop a more robust models

## WHY THIS PROJECT IS INTERESTING

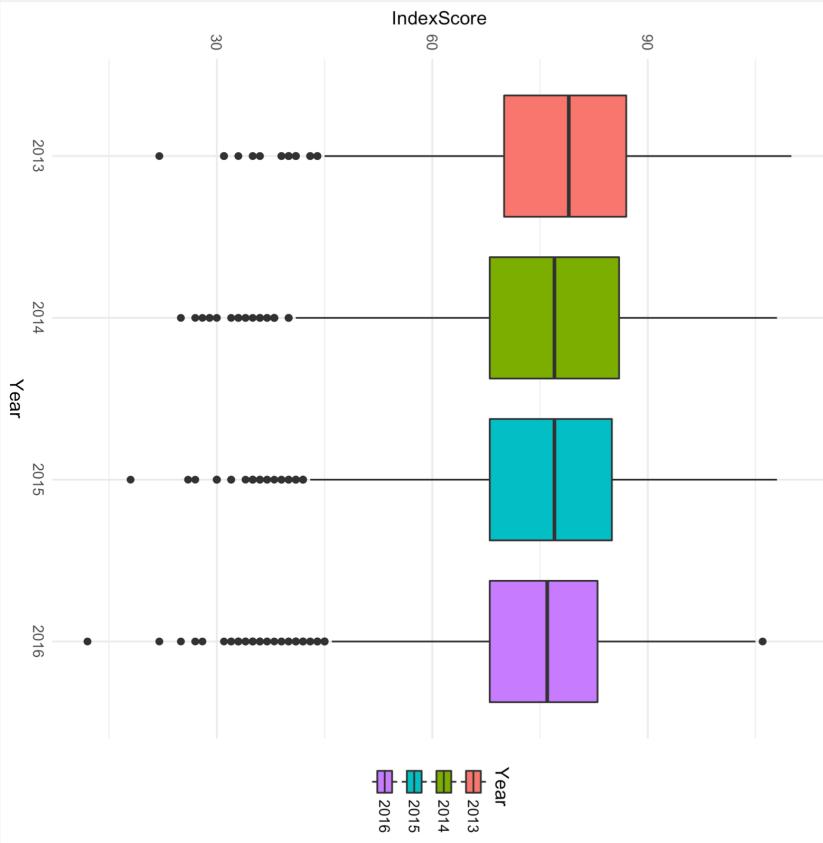
- Oklahoma is ranked #39 in education
- What's the Matter with Oklahoma – Economist Article from January 30<sup>th</sup> 2018
- Budget Crisis forces 4 day school week
- As of April 2, 2018 Teachers in Oklahoma are on strike – similar to West Virginia
- Ad Valorem Taxes make up a large percentage of school district's revenue
- Lots of economic activity in Oklahoma
  - Scott Pruitt, former AG of Oklahoma, opened Oklahoma to fracking for oil and gas causing a surge in revenues for Oklahoma counties and added land valuation
  - Google building server farms in Oklahoma (Minco County)

## DATA AND SOURCES

- Oklahoma has a robust reporting index that offers rich variety of data source:
- **Office of Educational Quality and Accountability District and School Profile Reports**
- Reports will provide basis for non financial dimensions to analysis
- Included Data:
  - ADM, ADA, % Attending College, Average ACT Score, Suspension Rates, Average Income, Average Property Value, % of parents attending Parent Teacher Conference, juvenile arrests, Free and Reduced Lunch
- **Oklahoma Cost Accounting System District Revenue Summary Reports:**
  - Detailed break down of state and federal incomes for school budgets
  - Will act as the main source of calculating ad valorem revenue as well as any other revenue source
  - Data available from 2006 to 2017
- **Oklahoma A-F Reporting Index:**
  - Contains A-F Index Scores and inputs into A-F Score for 2013 - 2016

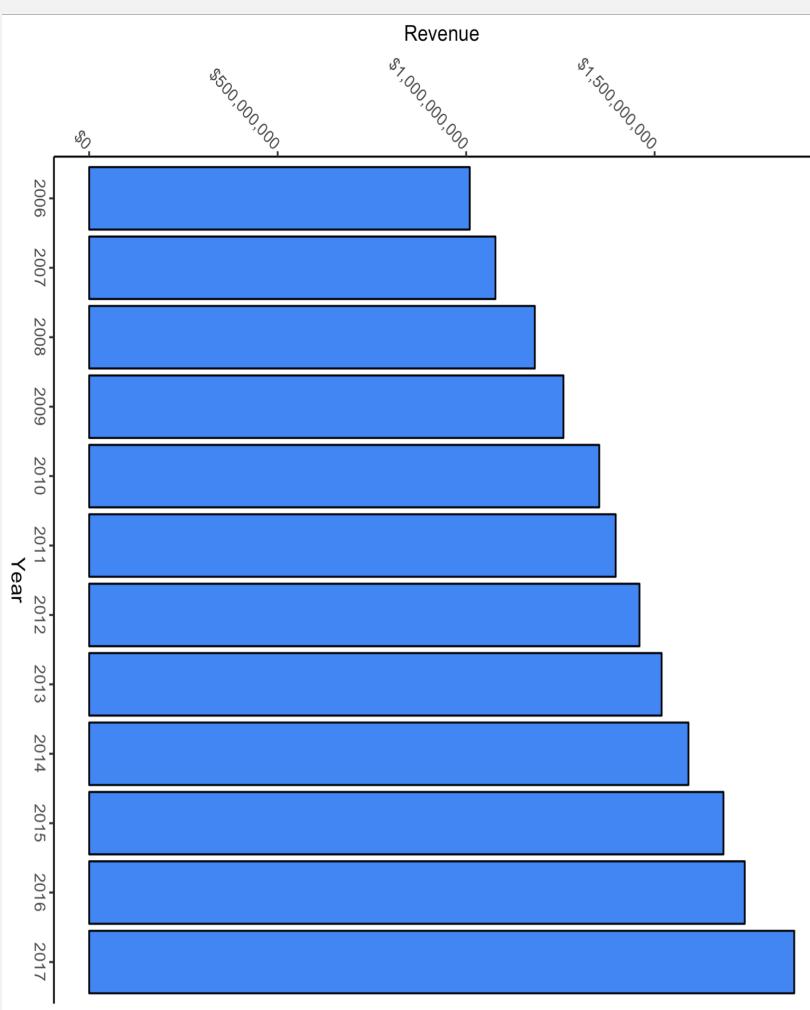
## SOURCE: OKSDE A-F REPORT INDEX

Distribution of A-F Index Scores



## SOURCE: OCAS DISTRICT REVENUE SUMMARY REPORT

Total Ad Valorem Revenue FY 2006 - 2017



## **THEORY & METHODS:**

- Use three different methods to test model:
  - **Linear Regression**
  - Predict A-F Index Score using linear model
- **Principal Component Analysis (PCA)**
  - Dimensionality Reduction combined with Regression methods
- **Artificial Neural Network**
  - Capture hidden layers, non-linear relationships and interaction effects that illustrates the complexity that goes into measuring school quality.
- Using these theories and methods I will evaluate the performance to see which one provides the most accurate prediction of school quality

## COMPUTATIONAL TOOLS:

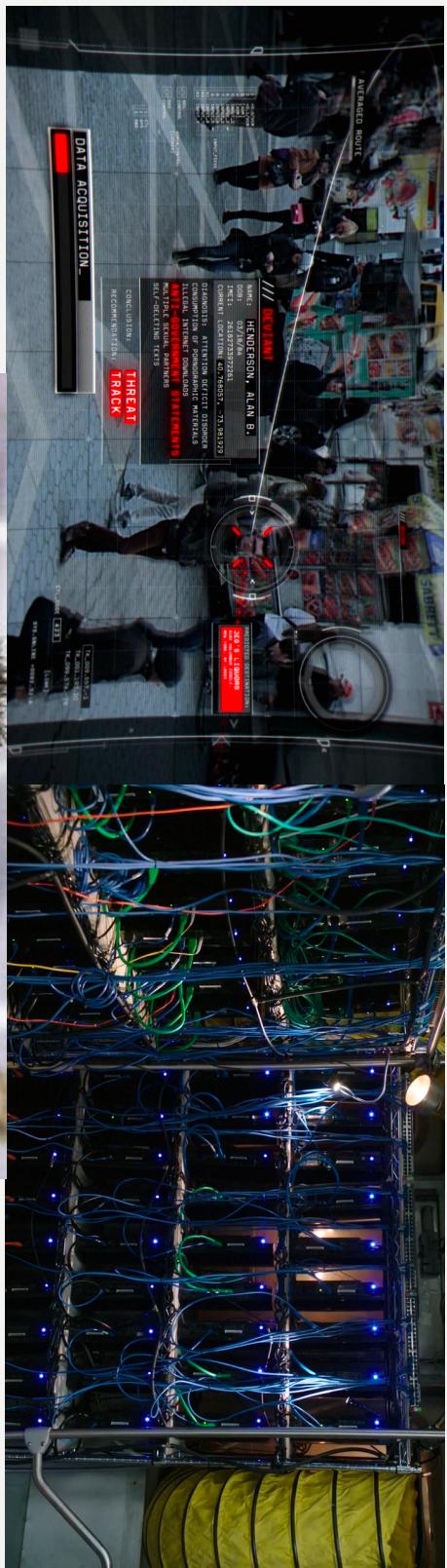
- **MySQL**
- Use database hosted by the Oklahoma Public School Resource Center to aggregate and manage necessary data
- **Python**
- Data Cleaning and Management – SQLAlchemy, Pandas, NumPy
- Multinomial Logistic Regression – Scikit learn LogisticRegression()
- Principal Component Analysis – Scikit Learn Decomposition()
- Artificial Neural Network – Tensor Flow, PyTorch

# PREDICT CRIME USING SPATIAL AND TEMPORAL APPROACH

A COMPARISON STUDY BETWEEN CHICAGO  
AND SEATTLE

Yangyang Dai

# RESEARCH BACKGROUND AND IDEAS



A crime will happen!!!

When ??  
Where ??



# AVAILABLE STUDIES AND RESOURCES

- Example studies:
- Predict London crime rates, predict crime types in Denver, Los Angeles and etc.
- Eg.A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi and A. Pentland, 'Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data', CoRR, vol. 14092983, 2014.
- Methods overview:
  - SVM: using this method is still slow and computationally expensive
  - Fuzzy time series: works well on binary data, but not so much on more levels
  - Artificial neural network (ANN): relatively accurate, but takes a long time in training phases
  - Unsupervised method by using multivariate time series based on parametric Minkowski model and dynamic time wrapping (DTW), however, this method is difficult to handle a missing value in order to get more accurate results
  - Bayesian Network: it may lead to some deviation during conducted the experiment, so the more factors unrelated to the geography should be considered to improve the accuracy of the model.
  - decision tree: eg, the result of crime cases has been classified into two classes such as neutral and danger. However, this method does not work well on all type of datasets.
  - logistic regression: the limitation of this approach is difficult to identify the probability of burglary activity and specific locations.

# AVAILABLE STUDIES AND RESOURCES

- Contribution and focuses:
- mid-west vs. west coast
- the newly developed vs. Early developed
- Crime city vs. civilized city

# CRIME DATA

- Mid-West
  - With relatively long history, early-developed
  - infamous for the crime, gangs
- Chicago crime – 2001 to present
  - extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system
  - 22 attributes (id, date, block, x, y coordinate etc.), 6.57M instances of crimes
  - <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

ID	:	Case Num	:	Date	↓	:	Block	:	IUCR	:	Primary Type	:	Description	:	Locati
11267718	B201047			03/26/2018 11:55:00 PM			06XXX S WESTERN AVE		0860		THEFT		RETAIL THEFT		GAS ST.
11267724	B201048			03/26/2018 11:45:00 PM			007XX S WELLS ST		1330		CRIMINAL TRESPASS		TO LAND		CONST STREET
11267725	B201049			03/26/2018 11:45:00 PM			117XX S THROOP ST		0910		MOTOR VEHICLE THEFT		AUTOMOBILE		STREET
11267753	B201053			03/26/2018 11:40:00 PM			064XX W IRVING PARK RD		1330		CRIMINAL TRESPASS		TO LAND		GROC STREET
11267705	B201038			03/26/2018 11:35:00 PM			068XX S ASHLAND AVE		502P		OTHER OFFENSE		FALSE/STOLEN/ALTERED TRIP		STREET
11267737	B201040			03/26/2018 11:30:00 PM			031XX S KEELER AVE		0326		ROBBERY		AGGRAVATED VEHICULAR HIJACKING		STREET
11267784	B201062			03/26/2018 11:30:00 PM			040XX W MAYPOLE AVE		031A		ROBBERY		ARMED: HANDGUN		STREET
11267753	B201044			03/26/2018 11:19:00 PM			039XX W 63RD ST		1812		NARCOTICS		POSS: CANNABIS MORE THAN 30GMS		STREET
11267813	B201046			03/26/2018 11:15:00 PM			032XX W 109TH PL		0486		BATTERY		DOMESTIC BATTERY SIMPLE		RESIDE
11268448	B202026			03/26/2018 11:15:00 PM			007XX W 47TH ST		1320		CRIMINAL DAMAGE		TO VEHICLE		STREET
11267714	B201023			03/26/2018 11:10:00 PM			071XX S ASHLAND AVE		0320		ROBBERY		STRONGARM - NO WEAPON		STREET
11267792	B201032			03/26/2018 11:08:00 PM			077XX S CORNELI AVE		1310		CRIMINAL DAMAGE		TO PROPERTY		RESIDE
11267701	B201028			03/26/2018 11:05:00 PM			033XX W WARNER AVE		1320		CRIMINAL DAMAGE		TO VEHICLE		STREET
11268384	B201039			03/26/2018 11:05:00 PM			056XX S WABASH AVE		0620		BURGLARY		UNLAWFUL ENTRY		APART

# CRIME DATA

## West coast

- Newly developed, tech-oriented

## Seattle crime – 2010 to 2017

- all the Police responses to 9-1-1 calls within the city

Hundred Block Location	::	District/Sec	::	Zone/B	::	Census Tr	::	Longitude	::	Latitude	::	Incident Location
3 AV S / S WASHINGTON ST	K	K3		920.2014		-122.330271593		47.600875809		(47.600875809°, -122.330271593°)		
20XX BLOCK OF 15 AV W	Q	Q1		5802.2003		-122.37613941		47.636336049		(47.636336049°, -122.37613941°)		
6 AV / YESLER WY	K	K3		9200.1002		-122.326350868		47.601708802		(47.601708802°, -122.326350868°)		
86XX BLOCK OF 24 AV SW	F	F2		11401.2005		-122.363172642		47.525585666		(47.525585666°, -122.363172642°)		
135XX BLOCK OF 23 AV NE	L	L1		2005017		-122.304248161		47.727498035		(47.727498035°, -122.304248161°)		
63XX BLOCK OF 29 AV SW	F	F1		10700.4001		-122.369833395		47.546493546		(47.546493546°, -122.369833395°)		
MARTIN LUTHER KING JR WY S / S GENESSEE ST	R	R2		10001.3006		-122.295370641		47.563805602		(47.563805602°, -122.295370641°)		
CALIFORNIA AV SW / SW ALASKA ST	W	W2		10500.4003		-122.386778535		47.561104368		(47.561104368°, -122.386778535°)		
24XX BLOCK OF AURORA AV N	Q	Q2		600.2045		-122.346642846		47.641419741		(47.641419741°, -122.346642846°)		
43XX BLOCK OF S FERDINAND ST	R	R3		10300.1002		-122.278843236		47.558197565		(47.558197565°, -122.278843236°)		
68XX BLOCK OF 30 AV NE	U	U3		3800.1004		-122.29516869		47.678505415		(47.678505415°, -122.29516869°)		
1XX BLOCK OF NW 81 ST	I	I2		2900.1014		-122.359311741		47.687664142		(47.687664142°, -122.359311741°)		
26XX BLOCK OF S DEARBORN ST	G	G3		8900.4011		-122.298714668		47.595534943		(47.595534943°, -122.298714668°)		
30XX BLOCK OF E UNION ST	C	C3		8800.1000		-122.293157653		47.612937729		(47.612937729°, -122.293157653°)		

39jp?category=Public-Safety&view\_name=Seattle-Police-Department-911-Incident-Response

# METHODS – FRAMEWORK

- Ideas:
- Build models using Chicago and Seattle crime data, train and test the two models, and predict the crime in each two cities
- Compare the temporal and spatial crime distributions
- Understand the differences and similarities between two cities
  - Crime distribution over the 12 months, Weekly distribution, daily distribution
  - Location or neighborhood differences (apart, street, sidewalk, school...)
  - Crime type differences

# METHODS – DATA PREPARATION

- Data cleaning: missing values
- Data reduction: eg. possible dimension reduction
- Data integration: eg. unify attribute naming, adopt military time system
- Data Transformation and Discretization:
  - create `crime_type`, `crime_time` variables, group into smaller subsets if too many, into hour intervals

# METHODS – MODELS AND ANALYSIS

- **Apriori Algorithm**
    - to find all possible crime frequent patterns regardless of the committed crime type
    - come up with a list of all crime hotspots along with its related frequent time
  - **Naïve Bayesian Classifier**
    - assumes the independent effect between attribute value
    - sklearn
    - Multinomial Naïve Bayes, crime type prediction
  - **Decision Tree Classifier**
    - sklearn
    - entropy function for the information gain to measure the quality of the split
- Evaluation:
- prediction accuracy
  - Running time

# POTENTIAL RESULTS

- Spatial and temporal hot spots in Chicago and Seattle
- Comparison of crime patterns
- Predicted Crime type in a given location within a give period of time
- Questions?

A photograph of a woman with long dark hair, wearing a white and blue striped hoodie, holding a small black puppy. She is looking down at the puppy. In the background, a man in a red and white striped shirt is looking up at her with a surprised expression.

# WHY HAVE THE CHINESE BECOME MORE TOLERATE ON PUPPY LOVE SINCE LATE 2000S: A FUNCTIONAL ANALYSIS

---

YILUN DAI

# RESEARCH QUESTION: INTRODUCTION

- The Chinese Definition of puppy love: 早恋 (zao lian, early love), most commonly refer to dating and having a love relationship before college
- Puppy love has been a taboo from 1950s to early 2000s
- However, there has been an increasing tolerance on puppy love since late 2000s
- Why are Chinese people changing their opinion?

## RESEARCH QUESTION: PAST LITERATURE

---

- Shen, Y. (2015). Too young to date! The origins of zaolian (early love) as a social problem in 20th-century China. *History Of Science*, 53(1), 86.
- Provided structural explanation for this taboo in 20<sup>th</sup> century
- Purpose: to prevent early marriage, early birth and school drop-outs
- Qualitative analysis on political documents
- Wang, J. (2013, August 31). Puppy love no longer taboo [Electronic version]. *Shanghai Daily*
- Gave cultural explanations on the change: many parents themselves experienced puppy love and controlling parents, and therefore are more understanding

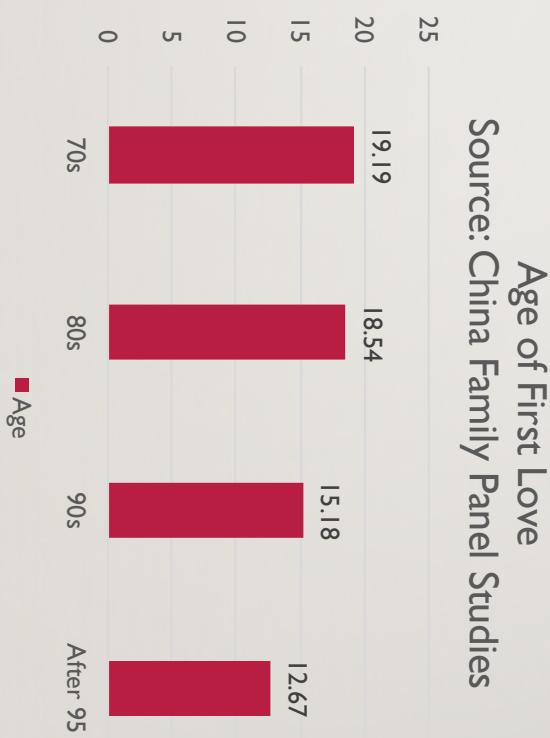
## RESEARCH QUESTION: THIS PROJECT

---

- Will provide functional and structural explanation on the recent change in attitude towards puppy love
- Will use computational methods and quantitative analysis

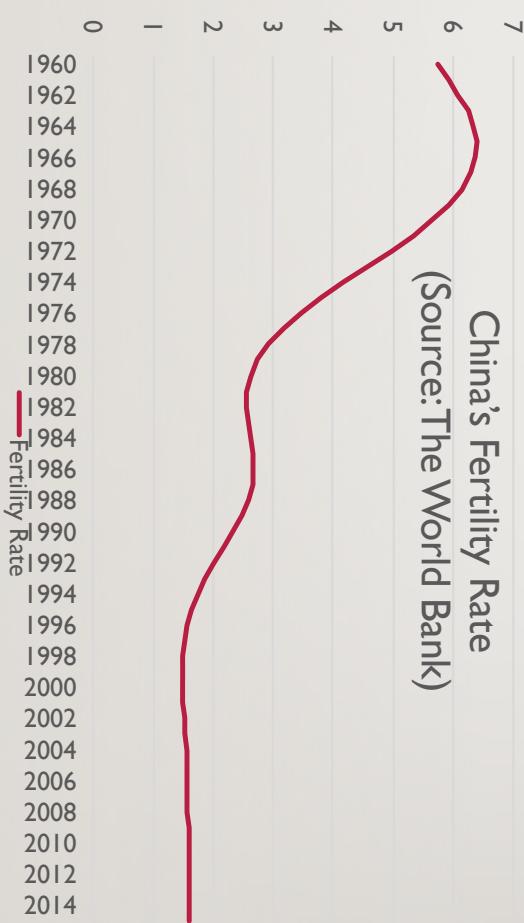
## DATA USED

- Love relationships and marital status statistics from China Family Panel Studies (PKU and i.baihe.com, 2015)
- Movie and TV series data from Douban
- Chinese Fertility rate and marriage rate statistics from the World Bank
- Study abroad statistics: National Bureau of Statistics of China



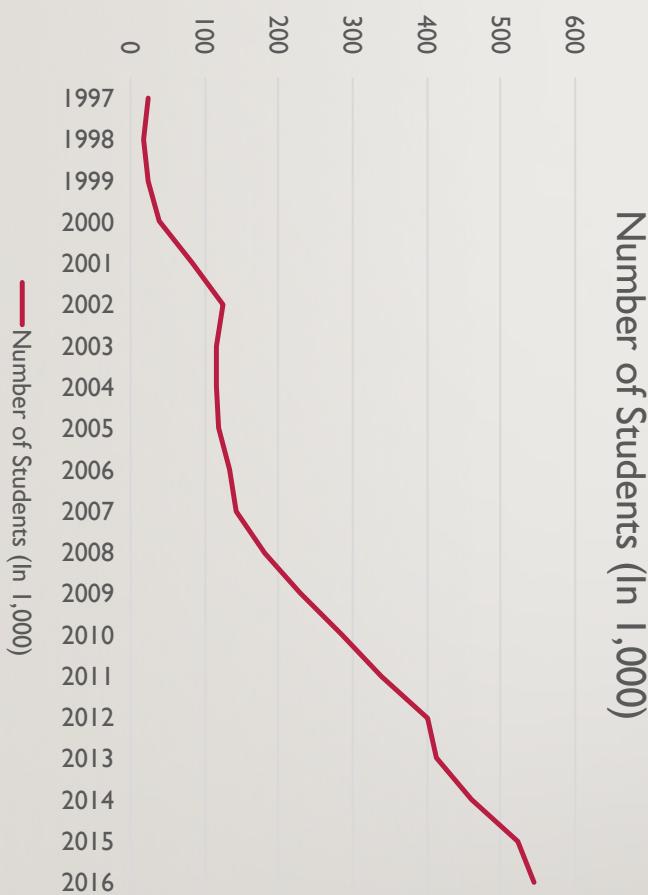
## DATA USED: FERTILITY RATE

---



- Plunging fertility rate since 1990
- One Child Policy became effective in 1980s not 90s
- little change even after the “second child policy”

## DATA USED: NUMBER OF STUDENTS STUDYING ABROAD



- Form of education is diversifying
- College Entrance Exam (Gaokao) is no longer the only way to higher education
- Score is no longer the one and only standard for “good student”

## METHODS AND TOOLS

---

- Web crawling: fetching TV series data from Douban
  - Tool: BeautifulSoup in Python
- Digital Survey on first love age, opinions on marriage, and childbirth
- Time Series Analysis: constructing VAR models with exogenous variables and conduct Granger Causality test
  - Tool: R studio

## THEORIES USED:

---

- VAR model with exogenous variables (Christopher Sims, 1980):
  - A multivariate time series process that includes both exogenous and endogenous variables
  - $Y(t) = \sum_{j=1}^p A_{11,j}X(t-j) + \sum_{j=1}^p A_{12,j}Y(t-j) + E_1(t)$
- Granger Causality:
  - “If a signal  $X_1$  “Granger causes” a signal  $X_2$ , then past values of  $X_1$  should contain information that helps predict  $X_2$  above and beyond the information contained in past values of  $X_2$  alone” (Anil Seth, 2007). Theory developed by Clive Granger in 1960s.
  - $X_1(t) = \sum_{j=1}^p A_{11,j}X_1(t-j) + \sum_{j=1}^p A_{12,j}X_2(t-j) + E_1(t)$
  - $X_2(t) = \sum_{j=1}^p A_{21,j}X_1(t-j) + \sum_{j=1}^p A_{22,j}X_2(t-j) + E_2(t)$

## HYPOTHETICAL RESULT

---

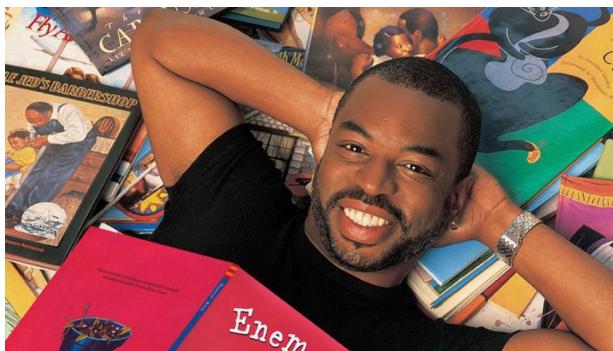
- The number of TV series that have puppy love contents is increasing
- The declining fertility rate and marriage rate, and the increasing number of students studying abroad granger cause the increasing tolerance of puppy love in China

# Take a Look!: Investigating the Relative Contributions of Children's Books and Child-Directed Speech

Joseph Denby

Perspectives on Computational Research

April 4, 2018



## Introduction

# How do we learn language?

or, more specifically,

From where do children glean English language rules/use  
and what does each source contribute respectively?

# Background

- Extensive work on child-directed speech (CDS)
  - Aspects of CDS predict vocabulary skill (Rowe 2008; Rowe 2012)
  - CDS linguistic construction differs markedly from standard speech (Cameron-Faulkner et al., 2001)
- Recent work investigates children's books as important source
  - e.g., Whitehurst et al., 1988; Montag et al., 2015; Montag et al., 2017

## Setup / Objective

- Caregiver speech and picture books are two prominent sources of linguistic input for children
- Research has historically neglected the latter
- Important to assess their relative (unique?) contributions

Montag et al.(2015), 2

What language-learning data might early picture books provide that everyday conversations do not?

# Objective

Question(s) for this project:

Are there substantive differences in the content of child-directed speech and age-appropriate children's books? If so, what?

# Procedure

- Exploratory Content Analysis on Speech and Text Corpora
  - ① Relative POS usage across time
  - ② Average sentence complexity (using tree parsing)
  - ③ Lexical diversity through type-token ratio (TTR)
  - ④ ...etc.

# Corpora – Language Development Project (LDP)

- UChicago-based initiative to document parent-child interactions with a socioeconomically-diverse sample ( $n = 102$ )
- Ecological check-ins between ages 14 - 58 months at four month intervals
- Analysis draws from transcripts of caregiver speech

subject	age	p_chat
0	22	1 play with Mommy's hand . no . no, no, no . nop...
1	22	2 what are you crawling for ? stand up . what ...
2	22	3 you want to show her how you do a head+stand ?...
3	22	4 how do flowers taste ? yeah . do you eat flow...
4	22	5 what ? you brought it down because you said...
5	22	6 do do do . &xxx get the foot book ? where is ...
6	22	7 right ? are you a mysterious girl ? no on...

# Corpora – Books

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, 26, 1489-1496.

- Alexander and the Terrible, Horrible, No Good, Very Bad Day* by Judith Viorst  
*Angelina Ice Skates* by Katharine Holabird  
*Are You My Mother?* by P. D. Eastman  
*Arnie the Doughnut* by Laurie Keller  
*Arthur Writes a Story* by Marc Brown  
*A Bad Case of Stripes* by David Shannon  
*Bark, George* by Jules Feiffer  
*Bear Wants More* by Karma Wilson  
*The Berenstain Bears and the Green-Eyed Monster* by Stan Berenstain and Jan Berenstain  
*The Berenstain Bears Forget Their Manners* by Stan Berenstain and Jan Berenstain  
*Blueberries for Sal* by Robert McCloskey  
*Bread and Jam for Frances* by Russell Hoban  
*Brown Bear, Brown Bear, What Do You See?* by Bill Martin, Jr.  
*Bunny Party* by Rosemary Wells  
*Caps for Sale* by Esphyr Slobodkina  
*The Carrot Seed* by Ruth Krauss  
*The Cat in the Hat* by Dr. Seuss  
*Charlie and the New Baby* by Ree Drummond  
*Chicka Chicka Boom Boom* by Bill Martin, Jr., and John Archambault  
*Chrysanthemum* by Kevin Henkes  
*How Do Dinosaurs Say Good Night?* by Jane Yolen and Mark Teague  
*How to Train a Train* by Jason Carter Eaton  
*If You Give a Moose a Muffin* by Laura Joffe Numeroff  
*If You Give a Mouse a Cookie* by Laura Joffe Numeroff  
*I'm a Big Sister* by Joanna Cole  
*The Keeping Quilt* by Patricia Polacco  
*Knuffle Bunny* by Mo Willems  
*Ladybug Girl at the Beach* by David Soman and Jacky Davis  
*Lilly's Purple Plastic Purse* by Kevin Henkes  
*Little Blue Truck Leads the Way* by Alice Schertle  
*The Little Engine That Could* by Watty Piper  
*The Little House* by Virginia Lee Burton  
*Llama Llama Home With Mama* by Anna Dewdney  
*Llama Llama Red Pajama* by Anna Dewdney  
*The Lorax* by Dr. Seuss  
*Love You Forever* by Sheila McGraw  
*Madeline* by Ludwig Bemelmans  
*Maisy Goes Camping* by Lucy Cousins  
*Maisy Goes to the Library* by Lucy Cousins  
*Make Way for Ducklings* by Robert McCloskey  
*Mike Mulligan and His Steam Shovel* by Virginia Lee Burton  
*Miss Rumphius* by Barbara Cooney  
*The Napping House* by Audrey Wood  
*No, David!* by David Shannon  
*Oh, the Places You'll Go* by Dr. Seuss

# Projections

- Replicate previous work showing that books exhibit higher TTR (with different speech corpus)
- Uncover meaningful distinctions between syntactic make-up of books vs. speech
  - Hopefully highlight specific benefits of children's books (and speech)

Thank you!

# Questions?

**Subsidies and Secession Demands:  
Text Analysis of Regional Parliaments in Spain and  
the United Kingdom 1999-2017**

By  
Dan Gamarnik

# Research Question

- Why do secessionist demands happen in rich, relatively affluent, non-deprived, democratic (RRANDD) minority regions?
  - Specifically demands by *government officials* in ethnic minority, regional governments.
  - Not protests or violence.
- Use computational text analysis of regional parliaments for secessionist demands of Spain (Catalonia, Basque Country) and the UK (Scotland, Wales)

# Literature Review

- Previous ethnic conflict literature has trouble addressing why RRANDD regions make demands.
- Relative deprivation theory (Horowitz 1981; Gurr 1970):
  - Ethnic groups rebel when they are poorer than the rest, discriminated against or under state violence
  - None of this is true in RRANDD regions
- Relational Materialist theory:
  - Robert Hale (2008) argues regions secede when economically they will be better off
  - Estimates show that in all regions they will be the same or worse off if they were independent.

# Literature Review (cont).

- Globalization theory (Hopkins 2014):
  - Argues that regions threaten to secede because of economic integration and increased austerity (ie spending cuts)
  - But cannot explain why some regions try to secede and others do not in the same country.
- No theory has fully explained this phenomenon yet.

# Theory

- Secession is result of elite and public discontent *specific* to each region.
  - Conventional story: funding cuts lead to anger among public.
  - Fiscal appeasement theory suggests that funding is actually *co-opting* officials from seceding and thus funding cuts is ending co-optation.
- Many factors can be analyzed in a model for this.

# Model

- $Y_{\text{Demands}} = B_0 + B_1 \text{SecessionPolls} + B_2 \text{NationalistVote} + B_3 \text{RegionalFunding} + \epsilon$

# Model (cont).

- Also considering doing an instrumental variable for when central government tax revenue
  - Based on logic of bargaining (Putnam 1988) it assumes that regional officials not demand it if the government doesn't have funding to give them.
  - But, if it is co-optation, (and they want their own state) then they will not reduce demands.
    - If goal is to start their own country, not funding.

# Research Design

- Use computational text analysis on minority parliament speeches in Spain and the UK
  - Looks at *local* parliaments of each region.
  - Use “secessionist” regions (Catalonia, Scotland) that had referendums for it
  - Also use “control” minority regions which did not have secessionist referendums (Basque Country, Wales).

# Research Design (cont.).

- Text analysis will look for secessionist stem words like “secession” and “independence” in both the majority and regional minority language (English/Spanish, Catalan/Gaelic/Welsh)
- This is the *dependent* variable
- Will use a word count of these phrases by month from 1999 to 2017.
  - N = 228
- These are proxies for number of demands made in these parliaments
- Because demand phrases are rare events I will use a poisson regression.

# Conclusion

- I will use text analysis and instrumental variables to measure secessionist demands in RANDD regions of Europe.
- It will use parliamentary speeches and test various ways for how funding might affect secessionist demands.

# Using Telematics and Insurance Data to Predict Accident Risk: Evidence from Beijing

Kanyao Han

## Research Question

How can we improve accident risk prediction in this digital era?

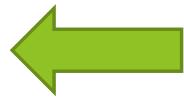
Telematics Device

What?  
How?

Driving behavioral data

## Traditional method in Insurance Industry

Customer Records → Logistic regression (or other regression)



Demographic features: Gender, Age, Job...

Vehicle features: Price, Type, Equipment...

Self-report driving mileage

Previous claim record

$$\log\left(\frac{P(y_i = 1 | d_{ij}, v_{ik}, m_i, c_i)}{1 - P(y_i = 1 | d_{ij}, v_{ik}, m_i, c_i)}\right) = \alpha + \sum_{j=1}^J \beta_j d_{ij} + \sum_{k=1}^K \gamma_k v_{ik} + \delta m_i + \eta c_i + \epsilon_i$$

## Traditional Method: Drawbacks

- ▶ Actual user ≠ customer in insurance record (like family car).
- ▶ Demographic features are not usually good indicators (Jin, Deng & Jiang, 2018).
- ▶ Self-reported records, such as annual driving mileage(White, 1976), are usually not exactly same as the actual ones.

A solution: combining telematics data and traditional insurance data.

# Data! Data! Data!

- ▼ A confidential car insurance dataset from insurance company: 150,000 observations

can be merged by vehicle id number

- ▼ A confidential telematics dataset from telematics company (10,000 cars over 3 months):

In-vehicle sensor data: acceleration, hard brake, actual mileage...

GPS data: averagely each car contains over 10,000 GPS observation

# An example of GPS data structure of a car

For ethical and confidential reason, I don't display some identifiable variables.

VIN <chr>	lon <chr>	lat <chr>	time <S3: POSIXct>
116.365120	39.953669	2016-01-01 09:50:06	
116.364989	39.953702	2016-01-01 09:50:39	
116.364955	39.953041	2016-01-01 09:50:51	
116.364960	39.952391	2016-01-01 09:51:03	
116.365124	39.951912	2016-01-01 09:51:16	
116.365120	39.951150	2016-01-01 09:51:26	
116.365201	39.950516	2016-01-01 09:51:35	
116.365211	39.949626	2016-01-01 09:52:10	
116.365295	39.948980	2016-01-01 09:52:19	
116.365168	39.948152	2016-01-01 09:52:40	

1-10 of 22,573 rows

Previous 1 2 3 4 5 6 ... 100 Next

## Extracting data from GPS (Some examples)

- ▼ Actual demographic features
  - Where the car owner live ➔ Economic status
- ▼ Driving behavior
  - Night driving, urban driving, etc.
- ▼ Driving environment
  - Familiarity with the road (how often a driver/drivers driving in one area/road)
  - Various road conditions in which a driver/drivers driving the car (I also have some types of road conditions data).

Method: Spatial data aggregation

# Modeling

## ► Response:

Self-reported claim in insurance data



Accident or not

Claim amount



Accident loss

## ► Features:

Traditional features in insurance data

Driving behavior in in-vehicle sensor data

Demography, behavior and environment in GPS data

# Model Selection

In data-driven research, there is no golden standard for model selection. Trying different algorithm and parameter.

- ▶ Classification (self-reported claim):  
Neural network often performs other algorithm in accident prediction  
(Paefgen et al. , 2013)
- ▶ Regression (claim amount):  
Lasso and elastic net are frequently used when there is dozens of features.

## So what?

- ▶ For insurance company: improving pricing strategy based on telematics data



- ▶ For telematics company: risk scoring service based on both classification and regression

- ▶ For academics: most similar research are just based on in-vehicle sensor data. The driving behavior and environment information extracted from GPS data have not been given enough attention. It has large potential for prediction.

# Is Coffee Shop an Indicator of Gentrification and Crime?

---

JIE HENG

# Gentrification and Crime

---

- What is gentrification:

“The process by which higher income households displace lower income [households] of a neighborhood, changing the **essential character and flavor of that neighborhood**”
- Criminological theory and the association between gentrification and crime:
  - Broken windows thesis
  - Civic communities perspective, Routine activities theory, Defended communities thesis, Social disorganization
  - Gentrification starts, crime rates goes down

# Gentrification and Crime

---

- How to identify gentrification:
  - “Coffee shop capture a more subtle cultural process of neighborhood change that might not be captured by such census indicators.”--- Richard Lloyd
- Papachristos, Smith & Fugiero. 2011. “More Coffee, Less Crime? The Relationship between Gentrification and Neighborhood Crime Rates in Chicago, 1991 to 2005”
  - Chicago is a racially segregated city
  - Asian neighborhood

# Gentrification and Crime

---

- How to identify gentrification:
  - “Coffee shop capture a more subtle cultural process of neighborhood change that might not be captured by such census indicators.”--- Richard Lloyd
- Papachristos, Smith & Fugiero. 2011. “More Coffee, Less Crime? The Relationship between Gentrification and Neighborhood Crime Rates in Chicago, 1991 to 2005”
  - Chicago is a racially segregated city
  - Asian neighborhood

# Research Questions:

---

1. Is there a relation between the number of coffee shops and crime associated with gentrification in a city with high ethno-racial diversity (New York)?
2. Whether the number of coffee shops of neighborhoods and other features of neighborhoods(race, income) are associated with crimes(murder, robbery)?

# Data

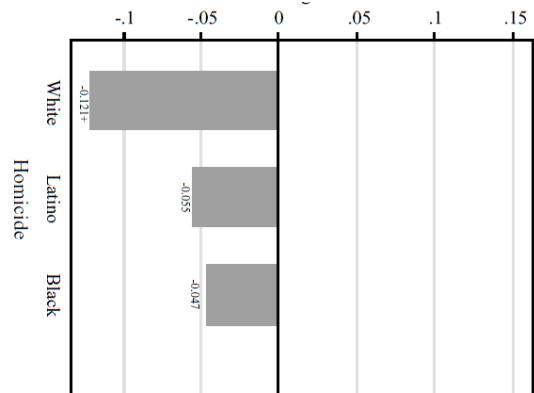
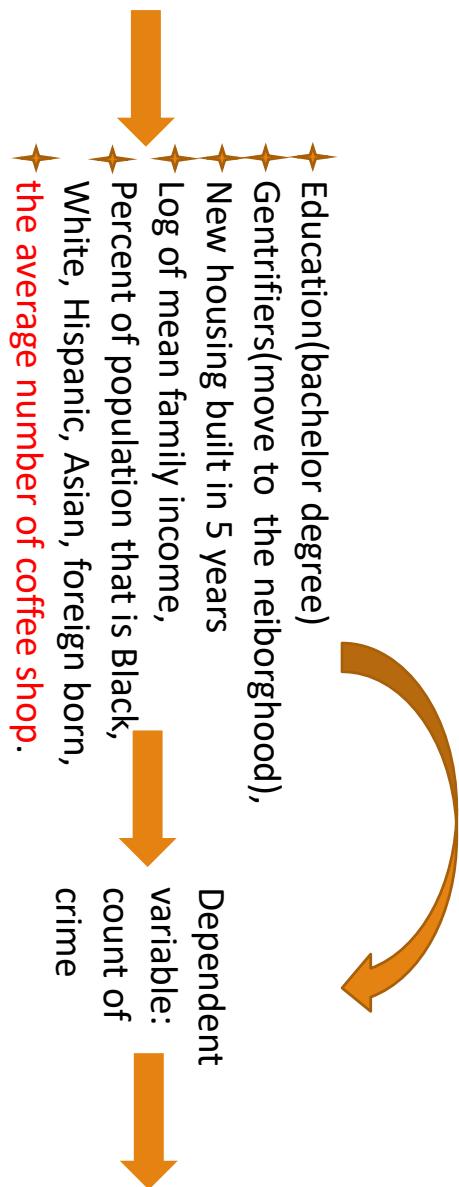
---

- The number of Coffee shops in NY:
- Sidewalk Café Licenses and Applications
- Legally Operating Businesses
- Grand Street BID Business Directory
- Income, population, education, race, housing built census
- Crime data:
- NYPD Complaint Data to Current



# Methods

Independent variables



Maps-- observation

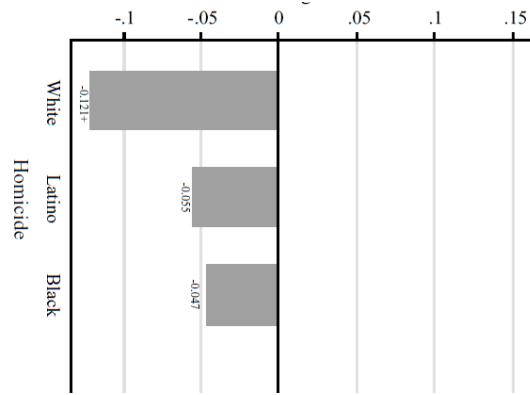
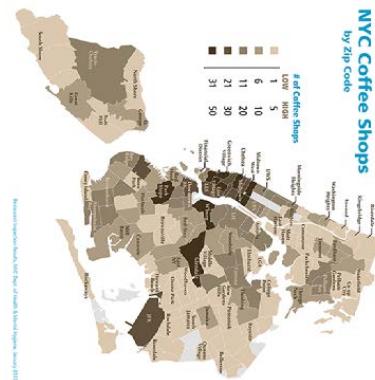
PCA/factor analysis

Compare models to  
predict crime

Crime type & coffee  
shops

# Methods

Independent variables



Maps-- observation

PCA/factor analysis

Compare models to  
predict crime

Crime type & coffee  
shops

## Challenges

---

1. Data cleaning
2. Divide the crime data to neighborhoods according to the GPS coordinates

Thank you

# THE AESTHETICS OF KNOWLEDGE CONSUMPTION:

[A Study of Textual and Graphical Forms in Online Science Communication]

Project Proposal  
Leoson Hoay

# RESEARCH QUESTION

- Can aesthetic measures of science web articles predict the readership and reader linger time of the publication/website that the articles belong to? (Observational)

...and/or:

- Can aesthetic measures on web articles predict readers' ratings of scientific content/websites, and their interest in the aforementioned content?  
(Survey/Experimental)

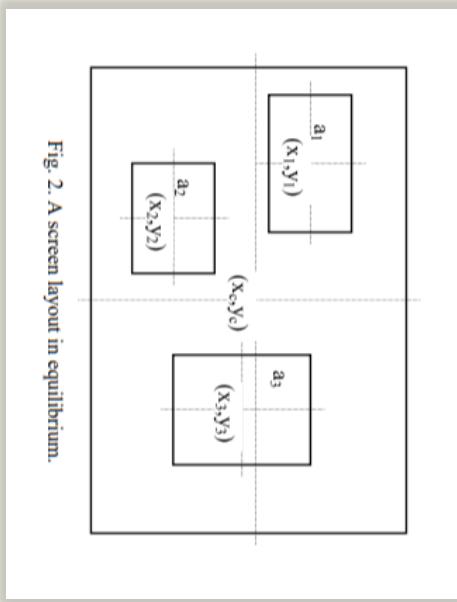
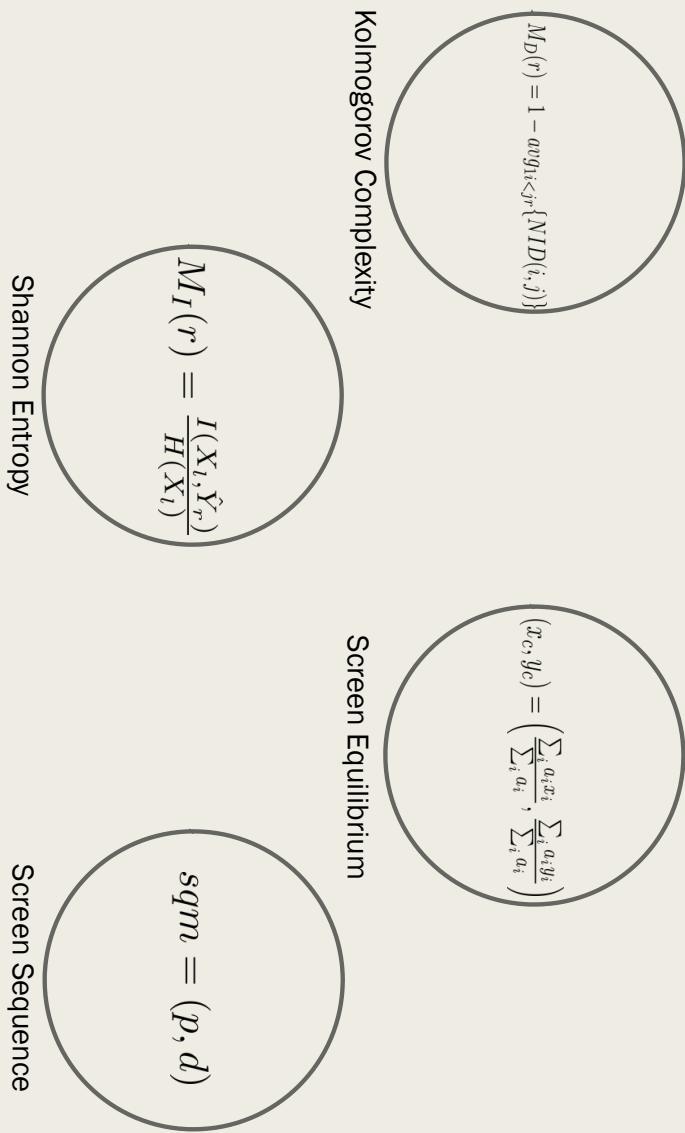
# FOUNDATIONS

- “Ethics and Aesthetics are one.” / “Knowledge is in the end based on acknowledgement.” – Ludwig Wittgenstein (1914 - 1916, 1953)
  - **Value and Aesthetics are inextricable** (Gombrich, 1960)
  - Build on previous studies in HCI (Human-Computer Interaction) and knowledge aesthetics
- **Defining and Quantifying “Aesthetics”**
  - “Formal notions” relating a reader to the content
  - Form and Function
- **Text Aesthetics: Semantic Consistency** (Tang, Qin and Liu, 2015)
- **Layout Aesthetics: HCI/UX Literature** – Pixel Fields, Screen Balance, Entropy, Complexity, Gestalt Unity, Edge Density, etc. (Machado et. al 2015, Rigau et. al 2007, Ngo et. al 2000 and others)



# BITS AND PIECES

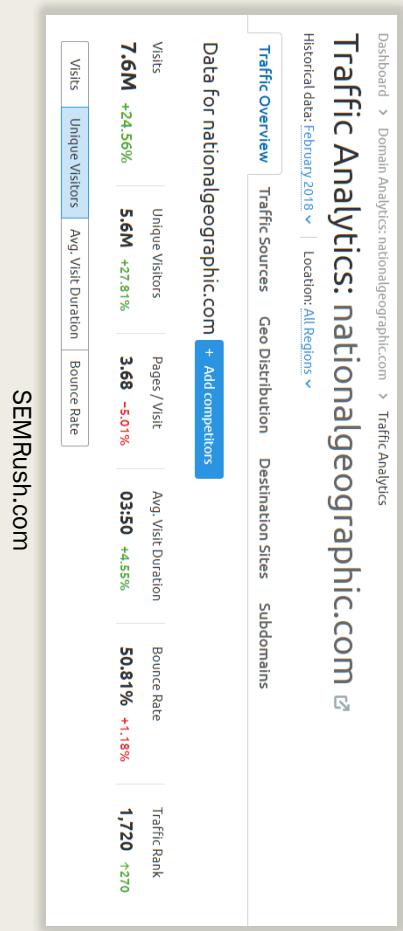
- Text Aesthetics: Semantic Consistency
- Layout Aesthetics (6 measures): (+Color Distribution, +Edge Density):



# DATA SOURCE(S)

- **Aesthetics:** Science Website Article Layouts and Text Content
  - Popular American web publications – National Geographic, BBC Earth, Nature, WIRED, New Scientist, etc. (Include global publications?)
  - Layouts: *PhantomJS* to scrape screenshots of article pages
- **Readership, Linger Time:** Website Metrics
  - Domain Data
  - *Estimated Data (SimilarWeb, SEMRush)*

# DATA SOURCES

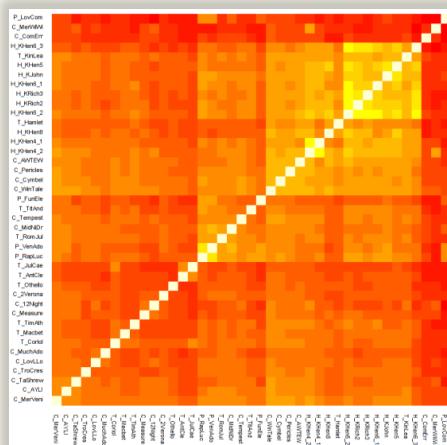


33 articles x 30 publications = 990 data points

American Council on Science and Health, RealClearScience (2017)

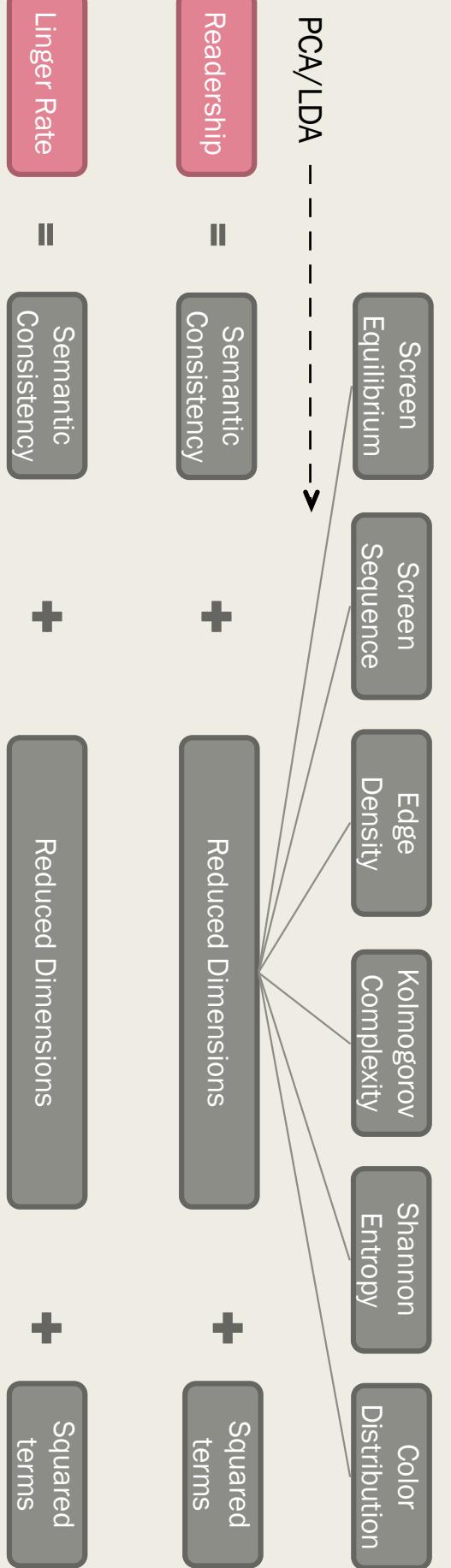
## METHODS

- Article Text Consistency: Gensim + Doc2Vec/Word2Vec
    - Trained vector space of documents from each individual science media outlet used to calculate individual article similarity
    - ‘Document Congruence’ for each science article formulated as the inverse of the document distance from the vector space
    - Other models: Cosine Similarity, WMD (Word Mover’s Distance)
  - Article Webpage Aesthetics:
    - *EBlImage* in R for pixel analysis, *scikit-image* for clustering, *OCROpus* for layout analysis, *OpenCV* for almost everything else



# METHODS

- Supervised Learning



# EXPECTED FINDINGS

- Model should be able to predict readership relatively well, maybe not as well for linger rate
  - While observing that higher aesthetic scores are usually correlated with higher readership and longer visit duration
  - **Expected Challenges:** Unknown requirements for n-power (larger dataset may be needed), extreme non-linearity, inaccuracies in traffic estimation
- Survey/Experimental: Aesthetic scores based on HCI and UI principles should predict readers' ratings of content/websites and initial interest level well

Virtue alone cannot save the world. Only actions will.

VIRTUE ALONE CANNOT SAVE THE WORLD. ONLY ACTIONS WILL.



Knowledge

Presentation/Representation

This presentation was presented at:  
Saleh 247, The University of Chicago  
Chicago, Illinois

United States of America  
Earth  
The Solar System  
The Milky Way  
The Local Group  
Virgo Supercluster  
The Known Universe





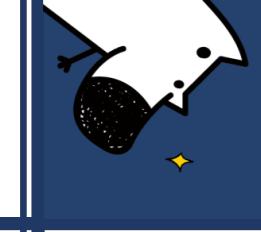
# User Behavior in Q&A community: an analysis of Zhihu



Andi Liao

2018/04/04

# Brief Intro to Zhihu



- Similar with Quora
  - Q&A Community - Follower & Following, Upvotes & Downvotes
  - User Interface - Navigate Bar, Topics
  - Recommend System - Feed, Invitation

The screenshot shows the Zhihu app interface. At the top, there is a navigation bar with tabs for 首页 (Home), 发现 (Discover), and 话题 (Topics). A search bar says '搜索你感兴趣的内容...' (Search for what you're interested in). Below the navigation bar is a list of topics and articles. One prominent article is titled '如何评价美剧《亿万》(Billions) 第三季第一集 (S03E01)？' (How to evaluate the first episode of the third season of the TV show Billions?). The article has 13 upvotes and 23 comments. Other visible sections include '我的收藏' (My Collection), '我的邀请' (My Invitations), and '我的礼物' (My Gifts). The bottom of the screen shows a feed of posts from users like Ben Peters and Mahnoor Wasi, along with a sidebar for 'Quora'.

知乎 首页 发现 话题 搜索你感兴趣的内容... 换回

来自话题: 美剧

陈亮 HKUST Finance

如何评价美剧《亿万》(Billions) 第三季第一集 (S03E01)？

一如既往的高信息量，在一集之内基本把本季几个主要角色的情况都交代清楚了，是个不错的开头，但感觉没有前两季的第一集精彩。目前来看这一季会变成三方混战：... 阅读全文 ▼

▲ 13 ▼ 23 条评论 ▶ 分享 ★ 收藏 ● 感谢 ...

来自话题: 历史

周伯通 Robot & Software

有哪些堪称“大力出奇迹”的事物？？

Space X的猎鹰重型火箭啊！说说配置：三个“9引擎”内核，猎鹰重型火箭共有三个引擎核心，每个核心都是九个梅林引擎的集群。这些引擎为猎鹰重型提供动力，使其成为世界上最具成本效益的重型运载火箭。配... 阅读全文 ▼

108 23条评论 ▶ 分享 ★ 收藏 ● 感谢 ...

我的收藏 8 Feeds Edit

我的邀请 49 Top Stories

我的礼物 Bookmarked Answers

社区服务中心 Links

版权服务中心

公共编辑动态

Quora Home Answer Notifications Search Quora Add Question

University of Chicago Topic you follow

Should I go to UChicago or UCLA? I want to study undergrad economics, but I also want to have some fun, and I'm worried that I won't at UChicago.

Mahnoor Wasi, studies History & Human Biology at University of California, Los Angeles (2019)

Updated Fri · Upvoted by Cyrus Paech, B.A. English Literature & Music, University of Chicago (2020)

The University of Chicago is famous for economics. The co-founder of the University of Chicago, John D. Rockefeller, was a brilliant economist and industrialist himself. And ever since John D. Rockef...

刘青山 · 知乎指南 · 知乎协议 · 应用 · 工作

申请开通知乎机构号

侵权举报 · 网上有害信息举报专区

违法和不良信息举报: 010-82716601

儿童色情信息举报专区

联系我们 © 2018 知乎

Improve Your Feed

Visit your feed

Follow 29 more topics

Find your friends on Quora

Upvote 5 more good answers

Ask your first question

Add 3 credentials

Answer a question

What's really going inside the admissions room for Ivies?

Ben Peters, Volunteer Admissions Interviewer, Cornell University Answered Jun 26, 2017 · Upvoted by Justin Shelby, B.A. Classics, University of Chicago (2010)

Vintage scotch is poured. Harvard sits at the head of the table, and calls for a toast. "To the Ivies!" The others respond. "To the Ivies!" Sips are taken, pinkies raised high. Brown already looks drunk... (more)

# Brief Intro to Zhihu

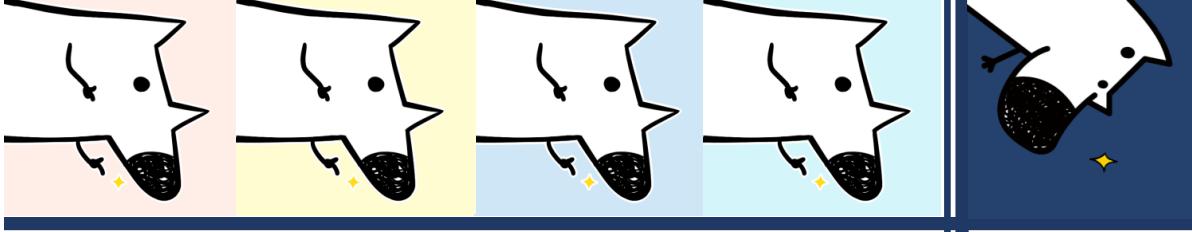


- Difference with Quora:

- User Identity - Any Name
  - Upvote - Agree
  - Top Writer - Centered

# Research Question

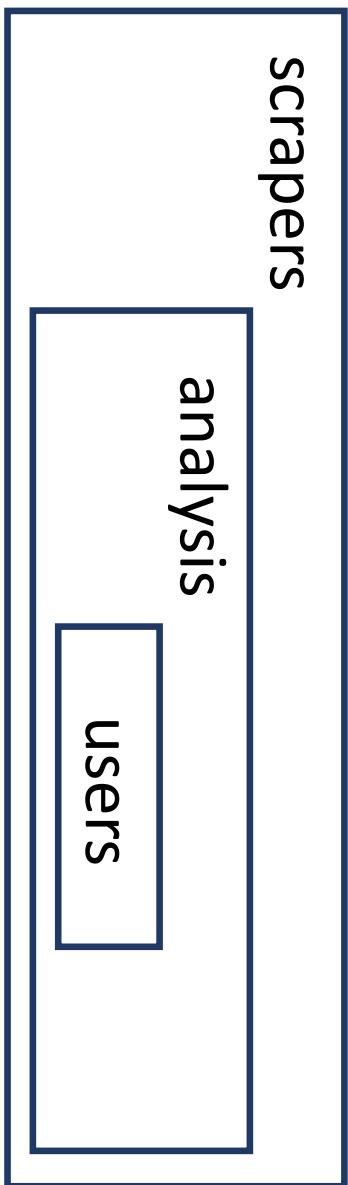
- User Behavior & Interaction in Zhihu Community
- Clustering:
  - Is it possible to cluster users based on data of their activities?
- Prediction:
  - Is it possible to predict a new user to be a top writer based on data of his activities?





# Why It Matters?

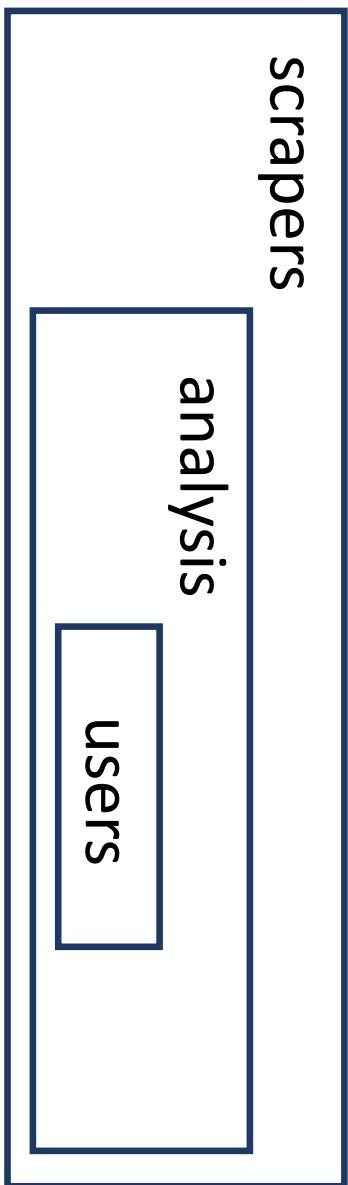
- Mixture: knowledge sharing + social
  - Venmo: Finance + Social
  - Weibo: Nickname + Comment
- Many scrapers, but few analysis focused on users in depth





# Why It Matters?

- Mixture: knowledge sharing + social
  - Venmo: Finance + Social
  - Weibo: Nickname + Comment
- Many scrapers, but few analysis focused on users in depth





# Data Source

- Scraper: <https://github.com/MatrixSeven/ZhihuSpider> - Java
  - Backup: <https://github.com/7sDream/zhihu-oauth> - Python unofficial API

- Data form: Mysql database

Follower	User	User Information
user_name	user_id	company, job
follower_name	index_url	education
update_time	token	answer, question
...	...	...

# Possible Theory

- Social and Interaction Graph
  - Degree-distribution: following, follower
  - Clustering coefficient: similarity
  - Reciprocity and balance: symmetric relationship
  - Assortativity: tend to connect similar nodes in the network
  - Tie Strength: interaction frequency

- Reference:

- Zhang, X., Tang, S., Zhao, Y., Wang, G., Zheng, H., & Zhao, B. Y. (2017). Cold Hard E-Cash: Friends and Vendors in the Venmo Digital Payments System. In *ICWSM* (pp. 387-396).
- Wang, T., Chen, Y., Wang, Y., Wang, B., Wang, G., Li, X., ... & Zhao, B. Y. (2016). The power of comments: fostering social interactions in microblog networks. *Frontiers of Computer Science*, 10(5), 889-907.
- Wang, G., Gill, K., Mohanlal, M., Zheng, H., & Zhao, B. Y. (2013, May). Wisdom in the social crowd: an analysis of quora. In Proceedings of the 22nd international conference on World Wide Web (pp. 1341-1352). ACM.

# Method & Tool

- Clustering & Prediction
  - K-mean++ clustering
  - Random forest classifiers
- Challenge
  - Supervised learning methods without labels
  - Categorical variables exist
- Reference
  - Patil, S., & Lee, K. (2016). Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors. Social network analysis and mining, 6(1), 5.

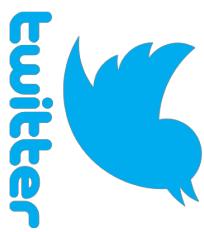
# Expectation



WIKIPEDIA  
The Free Encyclopedia



Quora



facebook

Non-Social

Social

4/4/2018

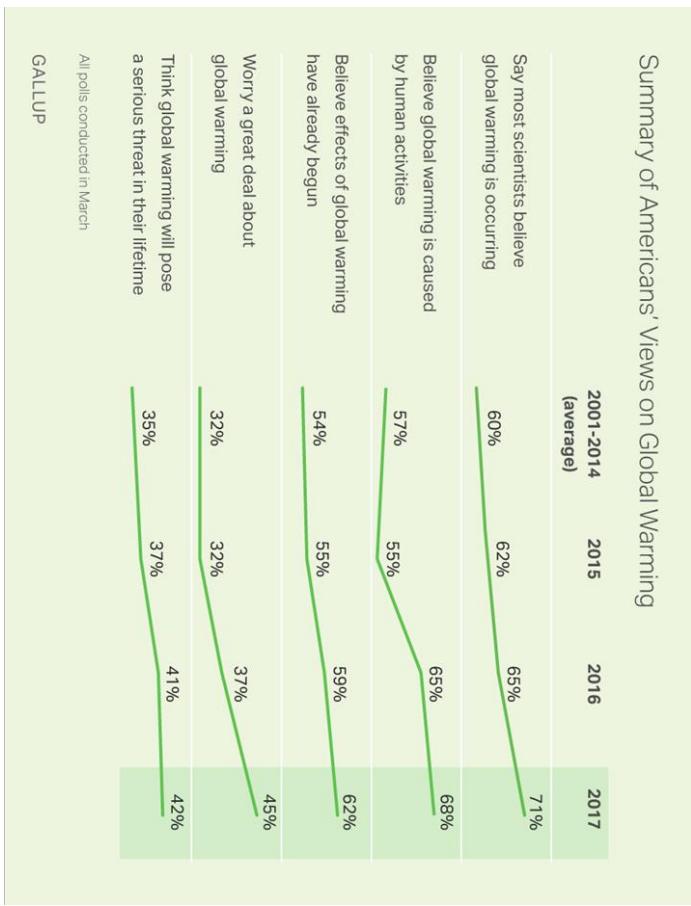
MACS 30200 PROJECT PROPOSAL

# How Abnormal Temperatures Affect Climate Change Attitudes and Behaviors

Kris Nichols

# The Threat of Climate Change

- Polls indicate that Americans are increasingly concerned about climate change.
- Why is there a disconnect between the personal danger a person feels from climate change and the existential danger of climate change?



# Perceived Low Risk of Personal Harm

- Research has indicated a number of potential cognitive biases that may be contributing to this such as skepticism in science, an identity protection mechanism, framing effects, and cognitive dissonance.
- To further investigate how individuals react when personally confronted with the effects of climate change, researchers have begun to investigate how individuals climate change attitudes may fluctuate with abnormal weather.

(Brooks, Oxley, Vedlitz, Zahran, Lindsey, 2014; Poortinga, Spence, Whitmarsh, Capstick, & Pidgeon, 2011; Nerlich, Koteyko, & Brown, 2009; Spence, & Pidgeon, 2010)

# Potential Problems in the Literature

- Methodological:
  - Non-response bias in national survey
  - Lack of precision in instruments
  - Observer-expectancy bias
- Content
  - What does concern really indicate?
    - Particularly in the face of proposed identity-protection mechanisms
  - Lack of insight into mechanisms
  - Introduces schism between behavior and attitudes.
    - Is a person who answers “Not Concerned” on a survey, but googles “climate change debunked” shortly after taking it really not concerned?

# Proposed Project

- Investigate how abnormal temperatures affect number of Google searches for climate change cross Democratic, Republican, and swing states may alter climate change behavior as measured by Google search data.
- Allow for greater clarity as to potential mechanisms
- Allow for comparison between conservative and liberal states
- Allow for greater ability to interpret concern

# Google Search Data

- Arguably, Google search data at its best is data that represents true, unaltered behavior and motivation.
- If someone is concerned about climate change we should see higher activity for climate change Google searches.
- Those who are looking to affirm their disbelief in climate change will also be visible with this data.

**Table 1**  
Signal-to-noise ratio in Google search terms.

Term	Underlying variable	t-Stat	R <sup>2</sup>
God	Percent believe in god	8.45	0.65
Gun	Percent own gun	8.94	0.62
African American(s)	Percent Black	13.15	0.78
Hispanic	Percent Hispanic	8.71	0.61
Jewish	Percent Jewish	17.08	0.86

Notes: The t-stat and R<sup>2</sup> are from a regression with the normalized search volume of the word(s) in the first column as the independent variable and measures of the value in the second column as the dependent variable. The normalized search volume for all terms is from 2004 to 2007. All data are at the state level. Percent Black and Percent Hispanic are from the American Community Survey, for 2008; the Jewish population is from 2002, gun ownership from 2001, and belief in God from 2007. Jewish data are missing one observation (South Dakota); belief in God data are missing for 10 states. The data for belief in God, percent Jewish, and percent owning guns can be found at <http://pewforum.org/how-religious-is-your-state-.aspx>, <http://www.jewishvirtuallibrary.org/jsource/US-Israel/usjewpop.html>, and <http://www.washingtonpost.com/wp-srv/health/interactives/guns/ownership.html>, respectively.

## Other Data

- I would also like to include the effect of *media* as a parameter in this project.
- No study to date has investigated the role of media in proliferating fears about climate change in the context of abnormal weather and this could be an important motivator of people's fears – or the lack thereof.
- This parameter will likely be realized through the scraping of articles on climate change and running sentiment analysis on these articles.

# Model: ARIMA Models

- ARIMA models attempt to describe the movements in a stationary time series as a function of what are called "autoregressive and moving average" parameters
- AR: Autoregressive part of the model
  - Forecast the variable of interest using a linear combination of past values of the variable
- MA: Moving Average part of the model
  - A moving average model uses past forecast errors in a regression-like model
- I: Integrated or "Differencing"
  - Subtracting previous values d times
- ARIMAX
  - Allows for covariates

# Proposed Models

- I will utilize three competing models for this project:
  - A Seasonal ARIMA model which maps Google searches for climate change onto a seasonal series.
  - A Seasonal ARIMAX model while will use Google searches for climate change “debunked” or other disproval terms as a covariate.
  - Finally, a Seasonal ARIMAX model which will use the media parameter as a covariate with the Google data to measure how the effect of media fits into this model.
- The ARIMA methodology will allow for analysis over a period of time while allowing for the flexibility of “Seasons” which capture the cyclic volatility of some seasons containing more abnormal weather than others.

## Hypotheses

- Predict that regardless of political party, that states with high amounts of abnormal weather will have significantly more general Google searches regarding climate change and global warming.
- Furthermore, I predict that when Google searches for “climate change not real” or “global warming not real” are included in analyses that these data will control for the majority of variation in Republican states.
- This study will suggest a cognitive dissonance mechanism by which participants initially Google “climate change not real” to soothe their dissonance.

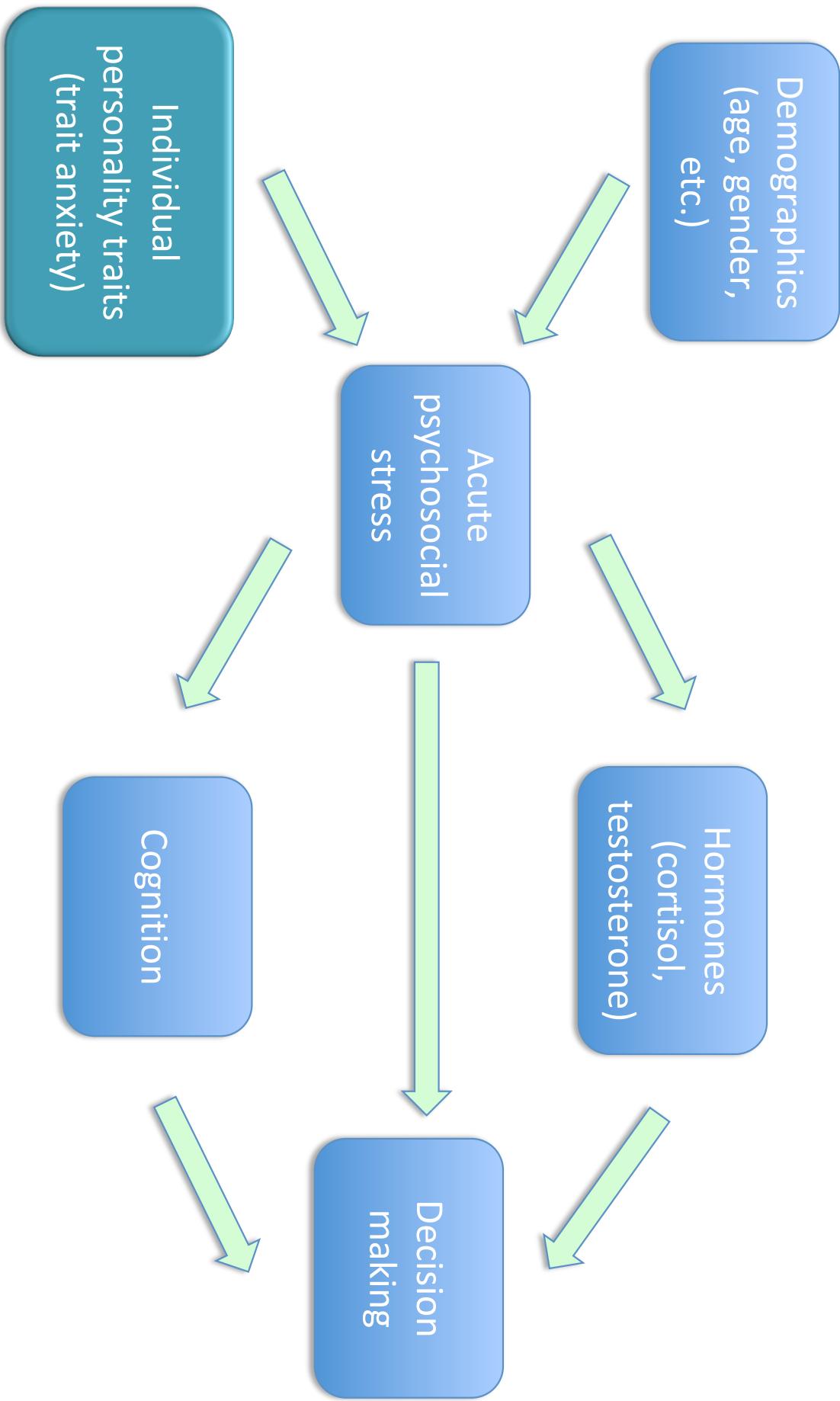
# A digital survey investigation of the construct validity of the Trait Anxiety Inventory in the UChicago community

Nora Nickels

Perspectives of Computational Research

Spring 2018

# Background



# Research Question

- In my study, to save money/time, some of these trait questionnaires are taken outside of the lab.
- RQ: How strong is the construct validity of the Trait Anxiety Inventory in my dissertation research population; specifically, does setting, time of day, and mood relate to trait anxiety responses of UChicago community members when the T.A.I. is completed outside of a controlled laboratory setting?

# What do we know?

- The State-Trait Anxiety Inventory
  - Long standing, frequently used, sensitive, valid (Spielberger, 1989; Chapman & Cox, 1977)
  - Retest correlations show reliability (Spielberger et al., 1983)
- Benefits of online questionnaire vs. in person questionnaire (Murthy, 2008)

# What don't we know?

- Risks of digital questionnaire
  - In person, have more environmental control
- Is the trait anxiety inventory susceptible to this risk?
  - Exaneous effects of mood, time of day, and setting
- Stability of responses of specific UChicago student population
  - Stressful environment

# Model / Theory

- In person vs. digital survey distribution
- If the TAI measures trait anxiety as a stable trait, then trait anxiety scores should not be statistically significantly related to setting factors that are not stable, such as:
  - Time of day
  - Mood
- Setting (where the survey was taken)

# Methods

- Qualtrics survey (digitally administered)
  - Administer the TAI
  - Administer post-survey questions:
    - Time of day
    - Setting
    - Mood
- Qualtrics
  - Software that enables users to collect and analyze data online
  - Benefits of digital design

# Methods

- Recruitment:
  - Goal: To test questionnaire validity within UChicago sample
  - Population: UChicago community
  - Sampling frame: UChicago community members accessible via listservs, Marketplace, Facebook
  - Sample: actual respondents of recruitment
  - Target: 200 respondents

# Analyses

- Descriptive statistics:
    - Distribution of anxiety scores
  - Regression model:
    - **Exogenous variables:**
      - Setting, time of day, mood
    - **Endogenous variable:**
      - Trait anxiety score
- Connection to research question: If extraneous factors of setting do not negatively affect questionnaire responses, we should see no relationship between the extraneous / exogenous variables and anxiety scores.

# Questions?

## References:

- Chapman, C. R., and Cox, G. B. (1977). Determinants of anxiety in elective surgery patients. In C. D. Spielberger and I. G. Sarason (Eds.), *Stress and anxiety* (Vol. 4, pp. 269–290). Washington, DC: Hemisphere/Wiley
- Murthy, D. (2008). Digital ethnography: An examination of the use of new technologies for social research. *Sociology*, 42(5), 837-855.
- Spielberger, C. D. (1989). *State-Trait Anxiety Inventory: a comprehensive bibliography*. Palo Alto, CA: Consulting Psychologists Press
- Spielberger, C. D., Vagg, P. R., Barker, L. R., Donham, G. W. & Westberry, L. G. (1980). The factor structure of the State-Trait Anxiety Inventory. In I. G. Sarason and C. D. Spielberger (Eds.), *Stress and anxiety* (Vol. 7, pp. 95–109). New York: Hemisphere/Wiley

# PREDICTING COLLEGE RETENTION RATES: AN APPLICATION OF THE CRITICAL MASS THEORY

Kevin Sun

Wednesday, April 4, 2018

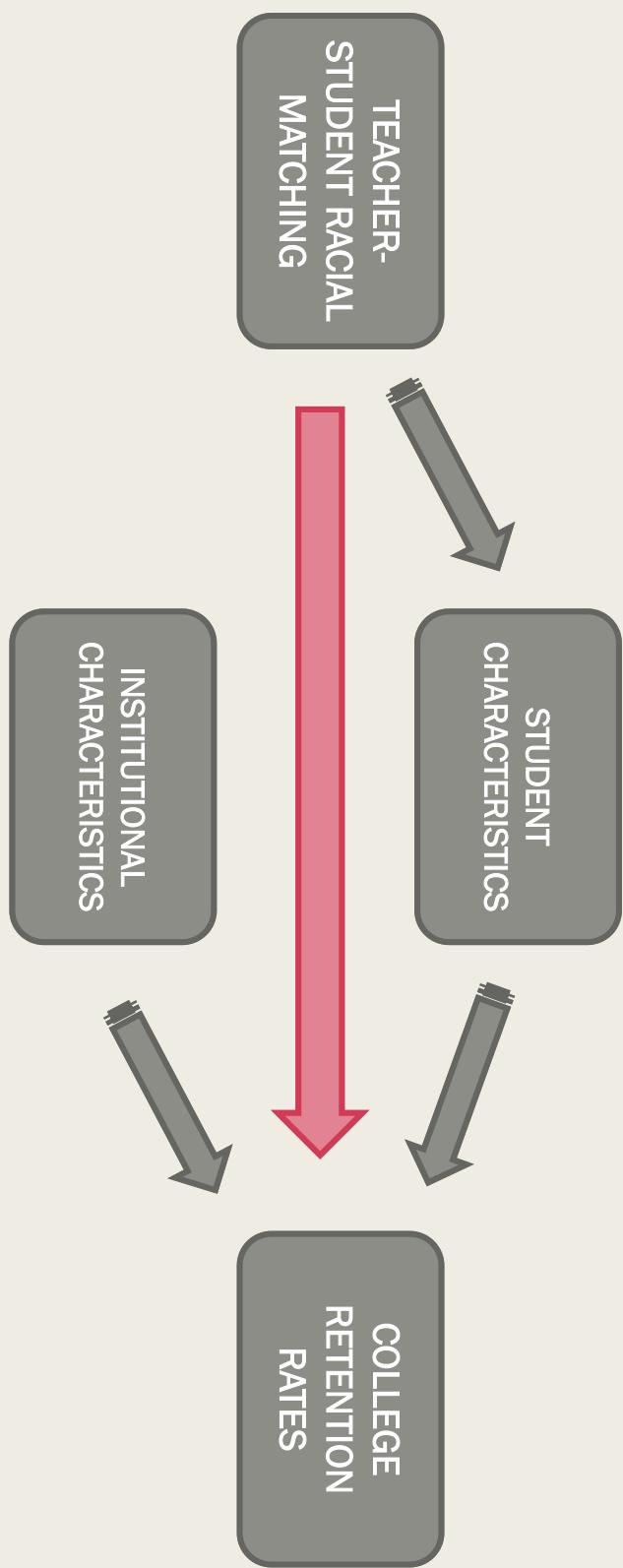
MA Computational Social Science  
Research Proposal

## RESEARCH QUESTION:

To what extent can **college retention**-rates  
be predicted by **bureaucratic representation**  
at the K-12 level?

- **College retention:** students who return to college their second year
- **Bureaucratic representation:** Teachers & administrators sharing demographic similarities with students

# THE LAY OF THE LAND:



## THE LITERATURE: On Retention

- **High School Achievement** (Astin, Korn, Green 1987)
- **Integration into academic and social community** (Tinto 1975, 1987)
- **Demographic Traits** (Astin 1975; St. John, Hu, Simmons, Musoba 2001)
- **Psychological Traits** (Trapmann, Hell, Hirn, Schuler 2007; Oswald, Schmitt, Kim, Ramsay, Gillepsie 2004)

# THE LITERATURE: Racial Matching

- **Assessment of student behavior & disciplinary outcomes**  
(Bates & Glick 2013; Lindsay & Hart 2017)
- **Expectation of student potential** (Gershenson, Holt, Papageorge 2016; McGrady and Reynolds 2012; Dee 2005)
- **Representation of non-white students in gifted programs**  
(Grissom, Rodriguez, Kern 2017)
- **Math & reading achievement** (Dee 2004; Clotfelter, Ladd, and Vigdor 2007)
- **Student perception of non-white teachers** (Cherng & Halpin 2016)

## THE DATA:

- Teacher & Administration Demographics: Chicago Public Schools
  - Individual teachers at each school
    - Impute race/ethnicity of each teacher
    - *NamSor*
    - *ethniclr*
- Other Demographics & School-Level Data: Chicago Public Schools
- College Attendance & Persistence Rates: National Student Clearinghouse

## METHODS & MODELS

$Retention = \beta_0 + \beta_1(BureaucraticRepresentation) + \beta_2(DemographicControls)$

## METHODS & MODELS

- OLS
- Decision Tree
- Random Forest

## ANTICIPATED CHALLENGES & EXTENSIONS:

- Imputing race based on names
- Analysis on every school district in the U.S.
- Ideal: student-teacher racial matching
- Hypothesis: A “critical mass” of teachers/administrators of color at a school will be associated with higher college retention/persistence levels in that school’s graduates

# DECODING CENSORSHIP ON RUSSIAN SOCIAL MEDIA USING FACEBOOK AND VK

ALEXANDER TYAN

MACSS 32000

# RESEARCH QUESTION

WHAT ARE THE CENSORSHIP CRITERIA FOR SOCIAL MEDIA POSTS IN RUSSIA?

## BACKGROUND AND MOTIVATION

- VK.COM
  - MOST VISITED IN RUSSIA (BY TRAFFIC) (SIMILARWEB, 2018)
  - SECOND MOST VISITED SOCIAL NETWORK GLOBALLY (BY TRAFFIC) (SIMILARWEB, 2017)
  - USED BY 40% OF RUSSIANS (LEVADA CENTER, 2017)
  - USED BY 53% OF 18-24 YEAR-OLDS (IBID)
- SOCIAL MEDIA TRENDS
  - GROWING AUDIENCE, DRIVEN BY 18-25 YEAR-OLDS (IBID)
  - 10% GROWTH IN THE LAST 4 YEARS (IBID)

## MORE BACKGROUND AND MOTIVATION

- EX. 2011 DUMA ELECTION PROTESTS
- VK PENETRATION INCREASES CHANCES OF PROTEST AND NUMBER OF PROTESTERS (ENIKOPOV ET AL, 2015)
- SOCIAL MEDIA HELPS OVERCOME COLLECTIVE ACTION PROBLEMS (IBID)
- SOCIAL MEDIA DISSEMINATE PROTEST INFORMATION (IBID; WHITE AND McALLISTER, 2013)
- GOVERNMENTS MAY USE DIFFERENT CRITERIA FOR CENSORSHIP (KING ET AL, 2013, 2017)

## MORE BACKGROUND AND MOTIVATION

- EX. 2011 DUMA ELECTION PROTESTS
- VK PENETRATION INCREASES CHANCES OF PROTEST AND NUMBER OF PROTESTERS (ENIKOPOV ET AL, 2015)
- SOCIAL MEDIA HELPS OVERCOME COLLECTIVE ACTION PROBLEMS (IBID)
- SOCIAL MEDIA DISSEMINATE PROTEST INFORMATION (IBID; WHITE AND McALLISTER, 2013)
- GOVERNMENTS MAY USE DIFFERENT CRITERIA FOR CENSORSHIP (KING ET AL, 2013, 2017)

# CONTRIBUTION

- GENERAL CONTRIBUTION TO UNDERSTANDING OF ONLINE PROTEST MOBILIZATION
- ESTABLISH METHODOLOGY TO TRACK THE EVOLUTION OF CENSORSHIP IN RUSSIA
- FILL THE GAP IN SYSTEMATIC KNOWLEDGE OF ONLINE CENSORSHIP IN RUSSIA

RESEARCH DESIGN AND MODEL

A screenshot of the official Facebook page for Alexei Navalny (@navalny). The page features a large profile picture of Navalny at the top right. Below it, there's a post from him with a photo of him holding a yellow marker, with the caption 'Таинственная распаковка К вам есть вопросы' (Mysterious Unboxing - You have questions). The post has received many likes and comments. To the left of this post, there's another one from 'Очень Эзд' (Very Ezd) with a photo of Navalny and the caption 'Today at 4:23pm'. Further down, there's a post from 'Маргарита Катерина' (Margarita Katerina) with a photo of Navalny and the caption 'Today at 3:19pm'. On the far left, there's a thumbnail for a photo album. The right side of the screen shows a sidebar with various links like 'Алексей Навальный', 'Status', 'Photo/Video', 'Write something on this Page...', 'Search for posts on this Page', 'Visitor Posts', 'Like - Comment', 'Like - Share', 'Interesting pages 11', 'Send gift', 'All posts', and 'Comments'.

## RESEARCH DESIGN AND MODEL

- API/WEBSRAPING OF ALEXEY NAVALNY VK AND FACEBOOK POSTS (PUBLIC)
- NLP ANALYSIS (TOPIC MODELLING?) AND CLASSIFICATION:
- POST ATTRIBUTES (TEXT CONTENT, TYPES OF POSTS) ~ PROBABILITY OF CENSORSHIP

## ANTICIPATED LIMITATIONS AND CHALLENGES

- EXTERNAL VALIDITY
- CONFOUNDING VARIABLES AND OTHER SOURCES OF CENSORSHIP
- TECHNICAL HURDLES

## ANTICIPATED LIMITATIONS AND CHALLENGES

- EXTERNAL VALIDITY
- CONFOUNDING VARIABLES AND OTHER SOURCES OF CENSORSHIP
- TECHNICAL HURDLES

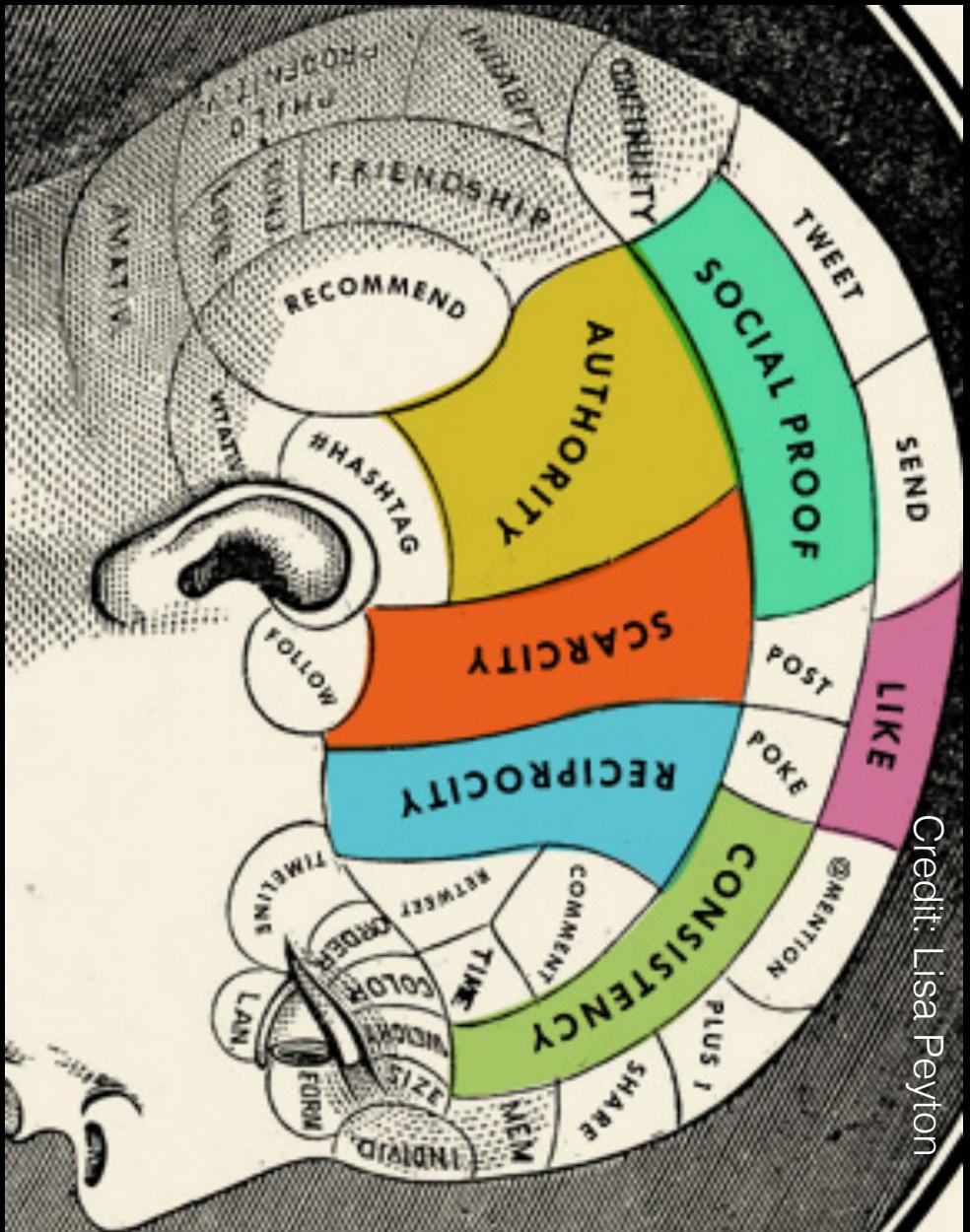
## ANTICIPATED LIMITATIONS AND CHALLENGES

- EXTERNAL VALIDITY
- CONFOUNDING VARIABLES AND OTHER SOURCES OF CENSORSHIP
- TECHNICAL HURDLES

# (Social) Media Psychology

Predicting frequency of social media use from personality traits

Credit: Lisa Peyton



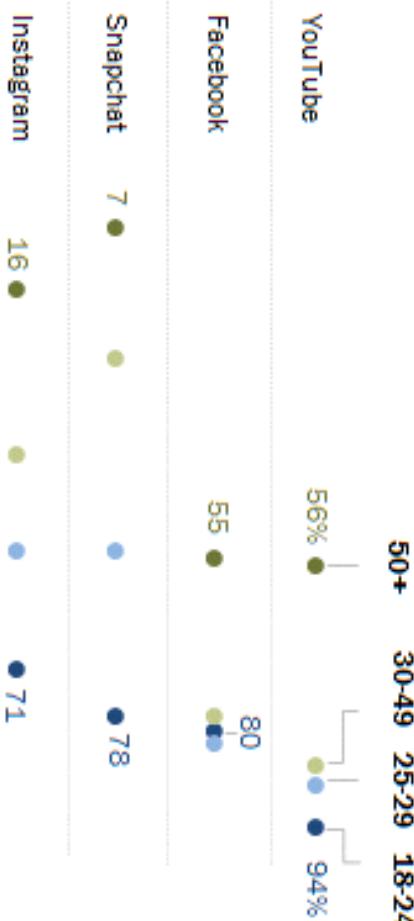
Media psychologists study the interplay between media and behavior.

# (Social) Media Psychology

## Social media usage is a diverse phenomenon.

**Social platforms like Snapchat and Instagram are especially popular among those ages 18 to 24**

% of U.S. adults in each age group who say they use ...



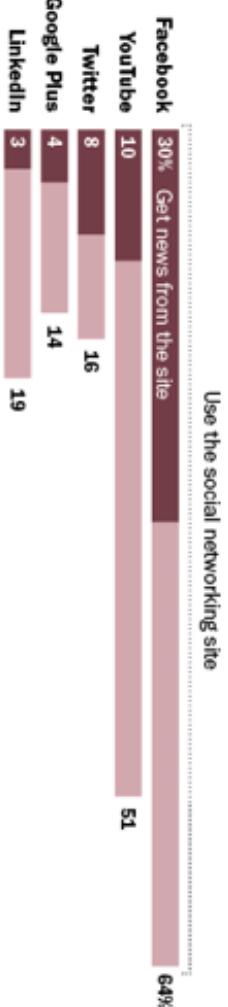
Twitter 14 ● 56% 73% 78% 80% 45%

YouTube 56% 51% 41% 30%

Source: Survey conducted Jan. 3-10, 2018.  
“Social Media Use in 2018”

PEW RESEARCH CENTER

Percent of U.S. adults who use each social networking site and percent of U.S. adults who get news from each social networking site

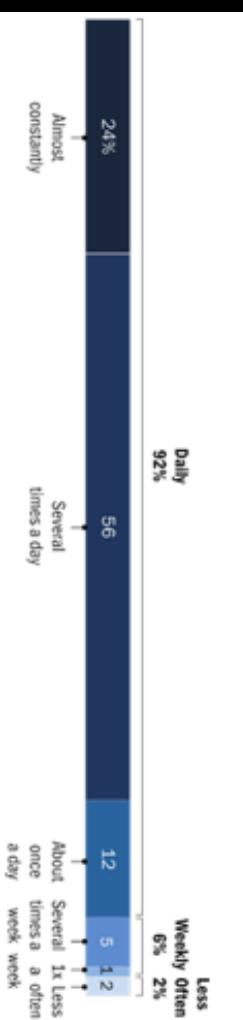


Note: The percent of U.S. adults who get news on Pinterest and Vine each account to less than one percent.  
Facebook News Survey, Aug. 21-Sept. 2, 2013 (N=5,173)

PEW RESEARCH CENTER

**Frequency of Internet Use by Teens**

% of teens ages 13 to 17 who use the internet with the following frequencies

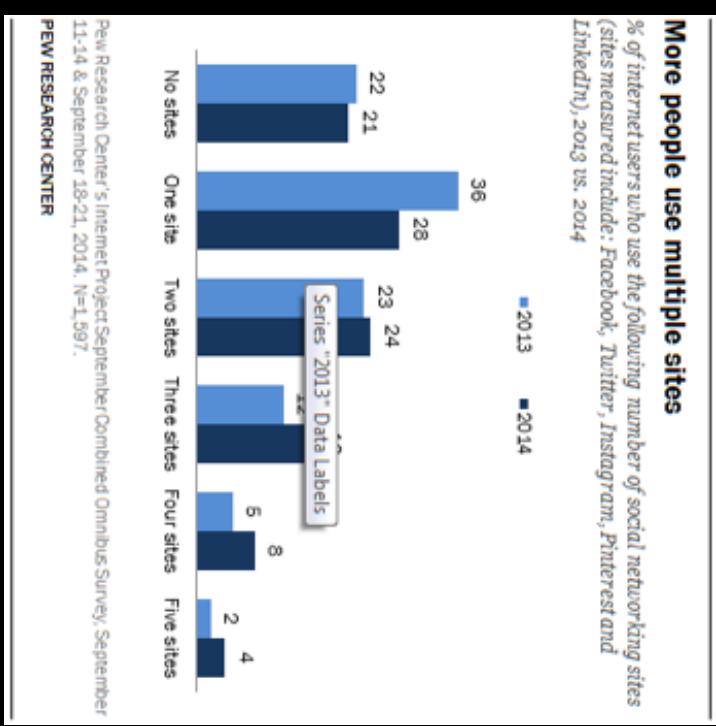
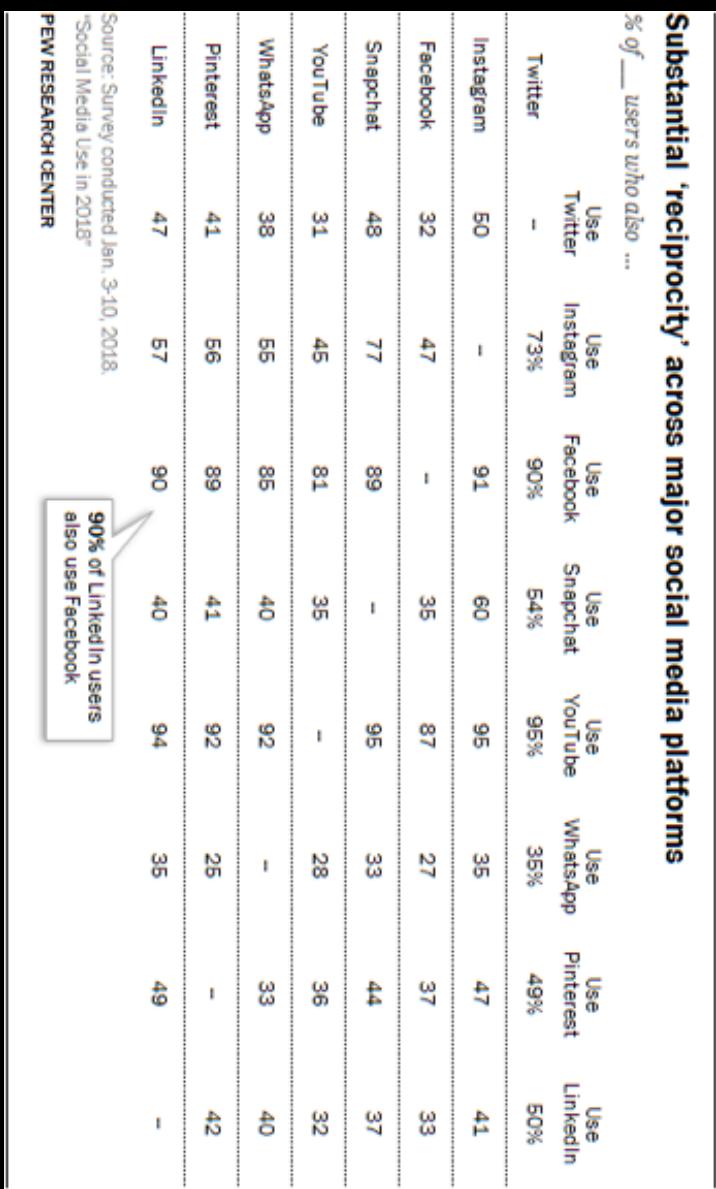


Source: Pew Research Center’s Teens Relationships Survey, Sept. 25-Oct. 9, 2014 and Feb. 10-Mar. 16, 2015. (n=1,016 teens ages 13 to 17).

PEW RESEARCH CENTER

# (Social) Media Psychology

Social media usage is a diverse phenomenon.



Why do different people use different kinds of social media platforms in different ways?

Let's use psychology to understand the origins of diversity in social media usage.

How do different kinds of people use different kinds of social media platforms in different ways?

Personality Differences

Demographic Differences

How can individual and personality differences explain patterns of social media usage?

# Past Research

Demographic Survey + Personality Survey + Social Media Use Survey

## A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage

David John Hughes <sup>a,\*</sup>, Moss Rowe <sup>a,b</sup>, Mark Batye <sup>a</sup>, Andrew Lee <sup>a</sup>

<sup>a</sup> Psychometrics at Work Research Group, Manchester Business School East, The University of Manchester, United Kingdom

<sup>b</sup> Department of Psychology, University of Bath, United Kingdom

**Table 4**  
Model summaries and fit statistics for latent variable regression models.

Model	R	$\beta$	$\chi^2$	df	CFI
Twitter Info					
A: Sociability	10.1	-.318**	5.245	5	1.000
B: Sociability & Need for Cognition &	17.5	-.284**	25.896	18	.980
C: Sociability & Need for Cognition	20.8	-.273**	29.419	23	.992
Age					
Twitter Social					
D: Conscientiousness	8.5	-.291**	n/a	0	n/a
E: Conscientiousness & Openness	12.3	-.248**	18.271	7	.975
F: Conscientiousness & Sociability	10.1	-.201**	11.640	12	1.00
Facebook Info					
G: Sociability	11.8	.343**	9.660	5	.995
H: Sociability & Need for cognition	13.8	.332**	43.396	18	.979
I: Sociability & Need for Cognition &	15.8	-.142*	47.053	23	.981
Age					
Facebook Social					
J: Sociability	2.4	.156**	3.945	5	1.000
K: Sociability & Neuroticism	4.8	.161**	19.75	18	.998
L: Sociability & Neuroticism	9.4	.162**	22.867	23	1.000
Age					

Note: All factor indicator loadings are >.07.

\* p < .05.

\*\* p < .001.

# Past Research

## Demographic Survey + Personality Survey + Social Media Use Survey

### Who interacts on the Web?: The intersection of users' personality and social media use

Teresa Correa \* , Amber Willard Hinsley, Homero Gil de Zúñiga

Center for Journalism & Communication Research, School of Journalism, University of Texas at Austin, USA

**Table 4**  
Regression on social media use by age.

	Young adults (18-29)				Adults (30 and older)			
	Model 1		Model 2		Model 1		Model 2	
	Beta	p value	Beta	p value	Beta	p value	Beta	p value
Gender	.05	.63	.12	.001	-.05	.13	-.03	.4
Race	-.26	.01	-.23	.10	-.09	.01	-.12	.001
Education	.01	.93	.08	.8	-.02	.57	-.03	.4
Income	-.05	.66	-.10	.000	.02	.55	.03	.52
Life satisfaction	.004	.97	.000	.50	-.09	.01	-.08	.05
R <sup>2</sup>	6.9%			2%				
Extraversion		.31	.005				.14	.000
Emotional stability		-.15	.18				-.15	.004
Openness		.06	.56				.08	.03
R <sup>2</sup>		18.4%					6%	

What about other social media platforms, other personality traits and cross-platform usage?

Nearly 56% of all American internet users use more than one social media platform (Pew, 2015)

# Present Research I

What personality differences, specifically for adolescents, predict social media usage?

Does personality predict usage of social media applications other than Facebook and Twitter?

What about personality traits other than the big five?

Do these also predict social media usage?

What about personality traits other than the big five?

Do these also predict social media usage?

**Can we predict cross-platform social media usage from personality data?**

# Computation I

1540 students (mean age=18.83) from UT Austin, enrolled in an online class, completed the survey as a part of the class. Two sub-samples from Fall 16 and Spring 2017 were collapsed to create one large dataset.

- Big-Five Inventory
- Dirty Dozen Measure of the Dark Triad
- Attachment Style-Questionnaire
- CES-Depression Questionnaire
- Social Connectedness Questionnaire
- Rosenberg Self-Esteem Questionnaire
- Demographics
- Individual Differences
- 11 Social Media Use Items

**We expect that a users' personality traits (especially the big-five inventory) will be able to predict social media usage, even after accounting for the variance explained by demographic variables.**

# Computation II

1540 students (mean age=18.83) from UT Austin, enrolled in an online class, completed the survey as a part of the class. Two sub-samples from Fall 16 and Spring 2017 were collapsed to create one large dataset.

- Big-Five Inventory
- Dirty Dozen Measure of the Dark Triad
- Attachment Style-Questionnaire
- CES-Depression Questionnaire
- Social Connectedness Questionnaire
- Rosenberg Self-Esteem Questionnaire
- Demographics
- Individual Differences
- 11 Social Media Use Items

**We expect that a users' personality traits (especially the big-five inventory) will be able to predict social media usage, even after accounting for the variance explained by demographic variables.**

# Computation

## Study 1

1. Use stepwise and hierarchical regression techniques to model individual social media use items as the exogenous variable and personality traits as the endogenous variable.

2. Explore better modeling strategies for social media use variables from personality data.

## Study 2

1. Run Principal Component Analysis on social media use variables to shed light on underlying structure in cross-platform social media usage.
2. Use stepwise and hierarchical regression techniques to model factor scores from step 1 as exogenous variables, with personality and demographic information as endogenous variables.

# RESTAURANT

## ATTRIBUTES AND HOW THEY AFFECT YELP

### RATINGS

MACS 30200

Fangfang Wan



# Why it's interesting?

- We can see what makes a restaurant highly rated – valuable for restaurant owners
- We can see if a rating is fair in our own perspective – valuable for guests



Find tacos, cheap dinner, Max's

Near Chicago, IL

11 Restaurants

Y Nightlife

X Home Services

Write a Review

Even

## Au Cheval

Claimed



48/74 reviews

[View Details](#)

[Edit](#)

[Write](#)

[Share review](#)

[Embed review](#)

[Compliment](#)

[Send message](#)

[Follow Michelle E.](#)

[Photo](#)

[United](#)

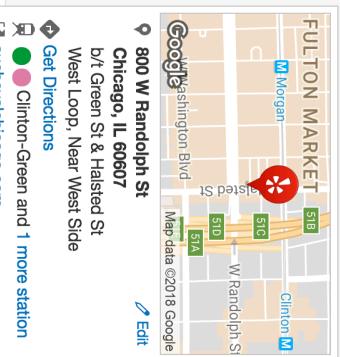
[20 photos](#)

[11 reviews](#)

[3 friends](#)

[May date ©2018 Google](#)

[Edit](#)



[Photo](#)

[United](#)

[20 photos](#)

[11 reviews](#)

[3 friends](#)

[May date ©2018 Google](#)

[Edit](#)



**Michelle E.**

Chicago, IL

3/31/2018

I have heard about the legend of these burgers for years. Finally pulled the trigger and went for lunch. The line was no joke. I had heard about the wait, but was hoping that maybe today it would be shorter; it was a 1 hr and 15 minute wait. However you can see your spot in line and wait via an app of you would like.

The waitstaff was quick to take our order. Everything arrived quickly. I got the single cheeseburger with egg and bacon along with the friends with Mornay and aioli sauces to split with my mothers. It was plenty of food and the cheeseburger was stacked high with the bacon and the egg. This is definitely a napkin meal as I had on to my cheeseburger to prevent the toppings from sliding out.

All the items I ordered were super good. The fries with the aioli were great. The Mornay sauce I could have done without. The pickles on the cheeseburger, and the side pickle, were all very good. Loved how the red onions on it were chopped super fine- just enough to add flavor with the fear of biting into a huge chunk.

If I am ever just craving the cheeseburger and fries, may try one of the Small Cheval restaurants. But I would miss the aioli...



"You have to get it "au cheval" style with a [fried egg](#) and bacon (which was more like pork belly)." in 592 reviews

\$2 Fried Egg



# Contribution

- Add evidence to business researches
- Provide guidance on how to improve Yelp rating for restaurants, and then revenue (Luca, 2016)

# Literature Review

- Luca, M. (2016). Reviews, reputation, and revenue: The case of Yelp.com
- Byers, J. W., Mitzenmacher, M., & Zervas, G. (2012, June). The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 248-265). ACM.

# Data

- Web scraping from Yelp html.

## Au Cheval Claimed

    4873 reviews 

\$\$\$ • Bars, Burgers, American (Traditional)



📍 800 W Randolph St

 Edit

Chicago, IL 60607

b/t Green St & Halsted St

West Loop, Near West Side

 Get Directions

● Clinton-Green and 1 more station

[auchevalchicago.com](#)

 Send to your Phone

### More business info

Takes Reservations No

Delivery No

Take-out No

Accepts Credit Cards Yes

Accepts Apple Pay No

Accepts Android Pay No

Good For Dinner

Parking Valet, Street

Bike Parking Yes

Good for Kids No

Good for Groups No

Attire Casual

Ambience Trendy

Noise Level Loud

Music Background

Good For Dancing No

Alcohol Full Bar

Best Nights Mon, Tue, Wed

Outdoor Seating No

Wi-Fi No

Has TV No

Drive-Thru No

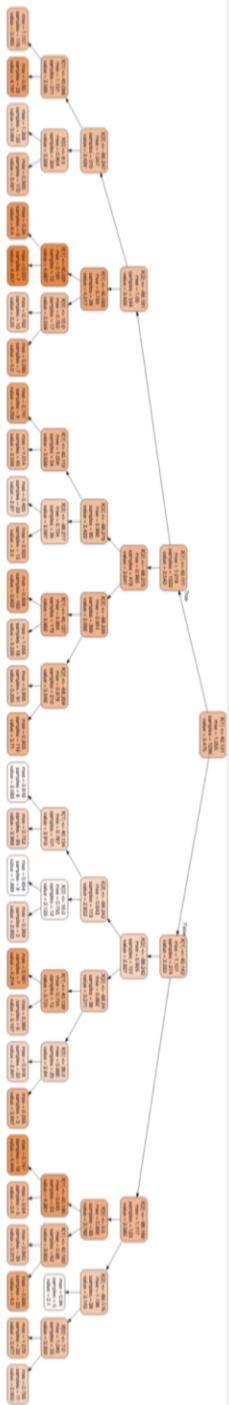
Caters No

Has Pool Table No

Gender Neutral Restrooms Yes

Data Plot (all businesses in IL as an example.  
Data from

<https://www.kaggle.com/yelp-dataset/yelp-dataset/data>



# Theory

Not exactly a theory in strict sense –

How do restaurant owners' behaviors affect Yelp rating?

Byers, J. W., Mitzenmacher, M., & Zervas, G. (2012, June). The groupon effect on yelp ratings: a root cause analysis. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 248-265). ACM.

# Model and Tentative results

- Model:
  - Decision tree
  - Linear regression
- X: attributes of a restaurant (price level, parking, accept Apple Pay, etc.) y: Yelp rating of a restaurant
- Tentative results:
  - Restaurants that offer parking, with higher price level, offer alcohol, etc. may have higher Yelp ratings.

# Computational tools

- Computational methods: Mainly Python
- Linear Regression
- Tree-based methods
- Web scraping

Thank you!

# **Neighborhood Disadvantage and High School Dropout**

---

MACS30200 Lerong Wang

## Research Question

---

- To what extent do neighborhood disadvantages affect **high school dropout rates**
- Use NYC data
- Neighborhood disadvantage: poverty rate, unemployment rate, crime rate...
- Control individual characteristics such as race and gender

# Motivation

---

Neighborhood Effect: The neighborhood effect is an economic and social science concept that posits that neighborhoods have either a direct or indirect effect on individual behaviors.

# **Why studying high school dropout is important?**

---

- Dropping out from high school is associated with negative employment and life outcomes
- Dropout status has also been linked with poor health, including poor mental health
- Possible policy implications for reducing dropout rates

# Previous Work

---

- Donnelly, Louis. Neighborhood disadvantage and school dropout. Retrieved from  
<https://doi.org/doi:10.7282/T375ZQRD>
- Vartanian, Thomas P., and Philip M. Gleason. "Do Neighborhood Conditions Affect High School Dropout and College Graduation Rates?" *The Journal of Socio-Economics*, vol. 28, no. 1, 1999, pp. 21–41., doi:10.1016/s1053-5357(99)00011-6.
- "Poverty and High School Dropouts." *American Psychological Association*, American Psychological Association,  
[www.apa.org/pi/ses/resources/indicator/2013/05/poverty-dropouts.aspx](http://www.apa.org/pi/ses/resources/indicator/2013/05/poverty-dropouts.aspx).

## My Contributions

---

- Prior research emphasizes on poverty and socio-economic status
- I will take more environmental determinants into account
- Model comparison

## Data

---

- High School Dropout Rate: NYC Department of Education Graduation Outcomes
- Neighborhood Disadvantage: American Community Survey
- Neighborhood Disadvantage: [www.nyc.gov](http://www.nyc.gov)

New York City Department of Education

## District Graduation Rate

District	Category	Cohort Year	Cohort			Total Grads			Total Regents			Advanced Regents			Regents without Advanced			Local			Still Enrolled			Dropout		
			#	% of cohort	#	% of cohort	#	% of grads	#	% of cohort	#	% of grads	#	% of cohort	#	% of grads	#	% of cohort	#	% of grads	#	% of cohort	#	% of grads	#	% of cohort
1	All Students	2013	4 year August	1043	61.3	608	58.3	95.1	215	20.6	33.6	393	37.7	61.5	31	3.0	4.9	258	24.7	121	11.6	7	0.7	7	0.7	
1	All Students	2012	4 year August	1059	65.0	641	60.0	98.3	229	21.4	35.1	412	38.5	63.2	11	1.0	1.7	258	24.1	148	13.8	7	0.7	7	0.7	
1	All Students	2011	4 year August	1128	66.5	59.0	65.3	95.5	166	14.7	25.0	469	41.6	70.5	30	2.7	4.5	303	26.9	143	12.7	10	0.9	10	0.9	
1	All Students	2010	4 year August	1104	59.0	53.4	56.4	51.1	95.6	127	11.5	21.5	437	39.6	74.1	26	2.4	4.4	308	27.9	197	17.8	2	0.2	2	0.2
1	All Students	2009	4 year August	1080	59.7	55.3	56.9	52.7	95.3	154	14.3	25.8	415	38.4	69.5	28	2.6	4.7	283	26.2	189	17.5	2	0.2	2	0.2
1	All Students	2008	4 year August	1128	68.5	60.7	66.1	58.6	96.5	187	16.6	27.3	474	42.0	69.2	24	2.1	3.5	246	21.8	172	15.2	12	1.1	12	1.1
1	All Students	2007	4 year August	1069	64.6	60.4	55.9	52.3	86.5	155	14.5	24.0	404	37.8	62.5	87	8.1	13.5	244	22.8	147	13.8	15	1.4	15	1.4
1	All Students	2006	4 year August	905	56.3	62.2	49.8	55.0	88.5	126	13.9	22.4	372	41.1	66.1	65	7.2	11.5	211	23.3	102	11.3	15	1.7	15	1.7
1	All Students	2005	4 year August	886	56.0	63.2	430	48.5	76.8	115	13.0	20.5	315	35.6	56.3	130	14.7	23.2	217	24.5	86	9.7	7	0.8	7	0.8
1	All Students	2013	4 year June	1043	61.3	58.8	58.8	56.4	95.9	211	20.2	34.4	377	36.1	61.5	25	2.4	4.1	284	27.2	121	11.6	7	0.7	7	0.7
1	All Students	2012	4 year June	1059	62.9	58.8	61.7	57.7	98.1	228	21.3	36.2	389	36.4	61.8	12	1.1	1.9	281	26.3	148	13.8	7	0.7	7	0.7
1	All Students	2011	4 year June	1128	64.9	57.5	62.2	55.1	95.8	166	14.7	25.6	455	40.4	70.3	27	2.4	4.2	319	28.3	143	12.7	10	0.9	10	0.9
1	All Students	2010	4 year June	1104	56.4	51.1	54.3	49.2	96.3	127	11.5	22.5	416	37.7	73.8	21	1.9	3.7	334	30.3	197	17.8	2	0.2	2	0.2
1	All Students	2009	4 year June	1080	56.9	52.7	54.4	50.4	95.6	153	14.2	26.9	391	36.2	68.7	25	2.3	4.4	311	28.8	189	17.5	2	0.2	2	0.2
1	All Students	2008	4 year June	1128	63.9	56.5	62.0	55.0	97.0	184	28.8	436	38.7	68.2	19	1.7	3.0	290	25.7	173	15.3	13	1.2	13	1.2	
1	All Students	2007	4 year June	1069	56.9	53.5	50.0	88.0	153	14.3	25.2	382	35.7	62.8	73	6.8	12.0	280	26.2	148	13.8	16	1.5	16	1.5	
1	All Students	2006	4 year June	905	54.9	60.7	49.3	54.5	89.8	125	13.8	22.8	368	40.7	67.0	56	6.2	10.2	225	24.9	102	11.3	15	1.7	15	1.7
1	All Students	2005	4 year June	886	52.2	58.9	41.9	47.3	80.3	113	12.8	21.6	306	34.5	58.6	103	11.6	19.7	255	28.8	86	9.7	7	0.8	7	0.8
1	All Students	2004	4 year June	756	47.0	62.2	34.7	45.9	73.8	111	14.7	23.6	236	31.2	50.2	123	16.3	26.2	213	28.2	68	9.0	3	0.4	3	0.4
1	All Students	2003	4 year June	603	32.8	54.4	28.5	47.3	86.9	60	10.0	18.3	225	37.3	68.6	43	7.1	13.1	209	34.7	60	10.0	5	0.8	5	0.8
1	All Students	2002	4 year June	381	21.2	55.6	185	48.6	87.3	18	4.7	8.5	167	43.8	78.8	27	7.1	12.7	130	34.1	28	7.3	5	1.3	5	1.3
1	All Students	2001	4 year June	376	252	67.0	205	54.5	81.3	19	5.1	7.5	186	42.5	73.8	47	12.5	18.7	58	15.4	54	14.4	11	2.9	11	2.9
1	All Students	2012	5 year August	1082	753	69.6	727	67.2	96.5	231	21.3	30.7	495	45.8	65.9	26	2.4	3.5	128	11.8	187	17.3	9	0.8	9	0.8

## Methods

---

- Spatial analysis to analyze the pattern of high school dropout rates in NYC: Local spatial autocorrelation using Geoda
- Linear Regression
- Random Forest/Tree-based Method in Python

**Thank you!**



- Questions?

# MACSS Project Proposal

Laurence Warner



## Research Question

Which special events cause the greatest changes in traffic flows in major US cities?

# Data

≡ UBER Movement

FAQS

SUBMIT FEEDBACK

SIGN OUT

## Let's find smarter ways forward

Uber Movement provides anonymized data from over two billion trips to help urban planning around the world

▶ Watch how Movement works

Select a city to explore ▾

# Data

≡ UBER Movement

FAQS

SUBMIT FEEDBACK

SIGN OUT

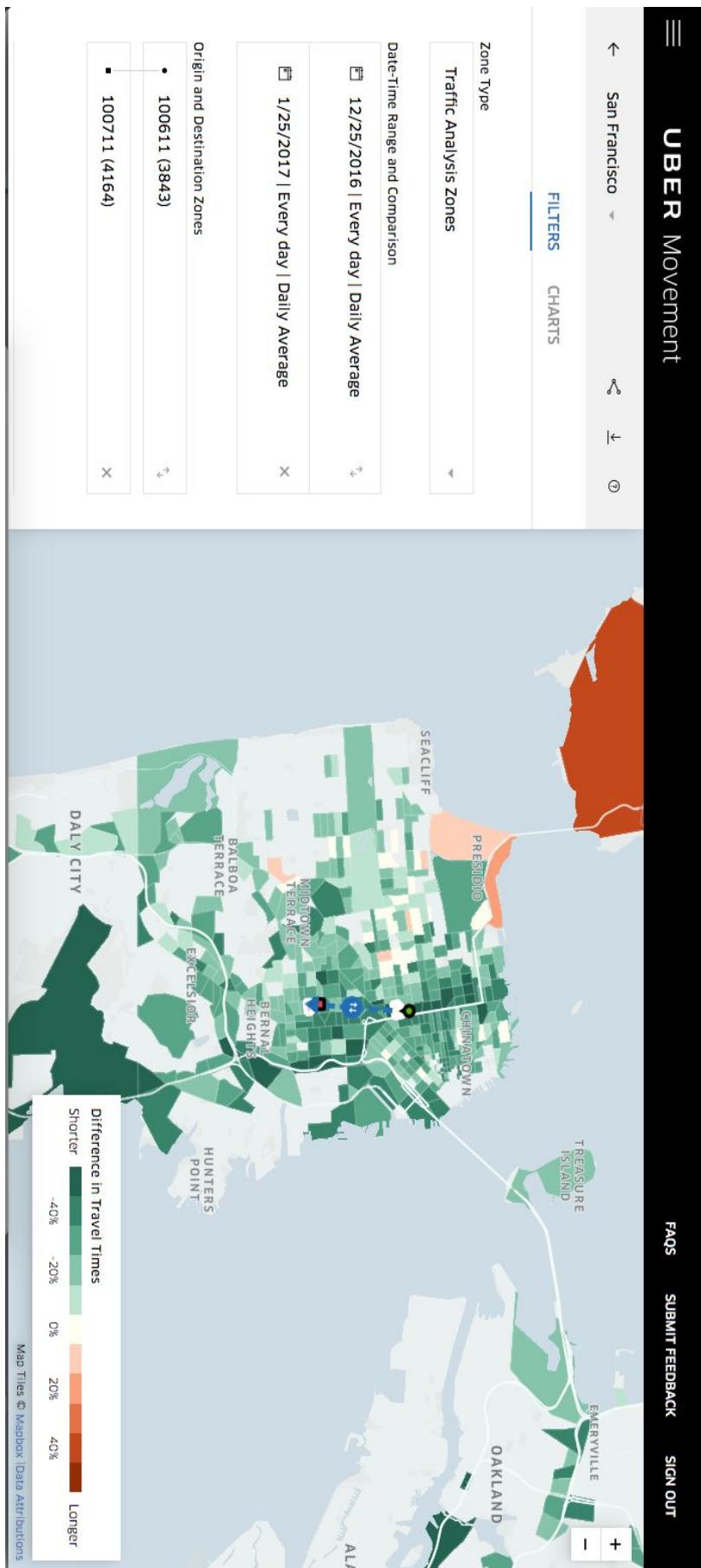
## Let's find smarter ways forward

Uber Movement provides anonymized data from over two billion trips to help urban planning around the world

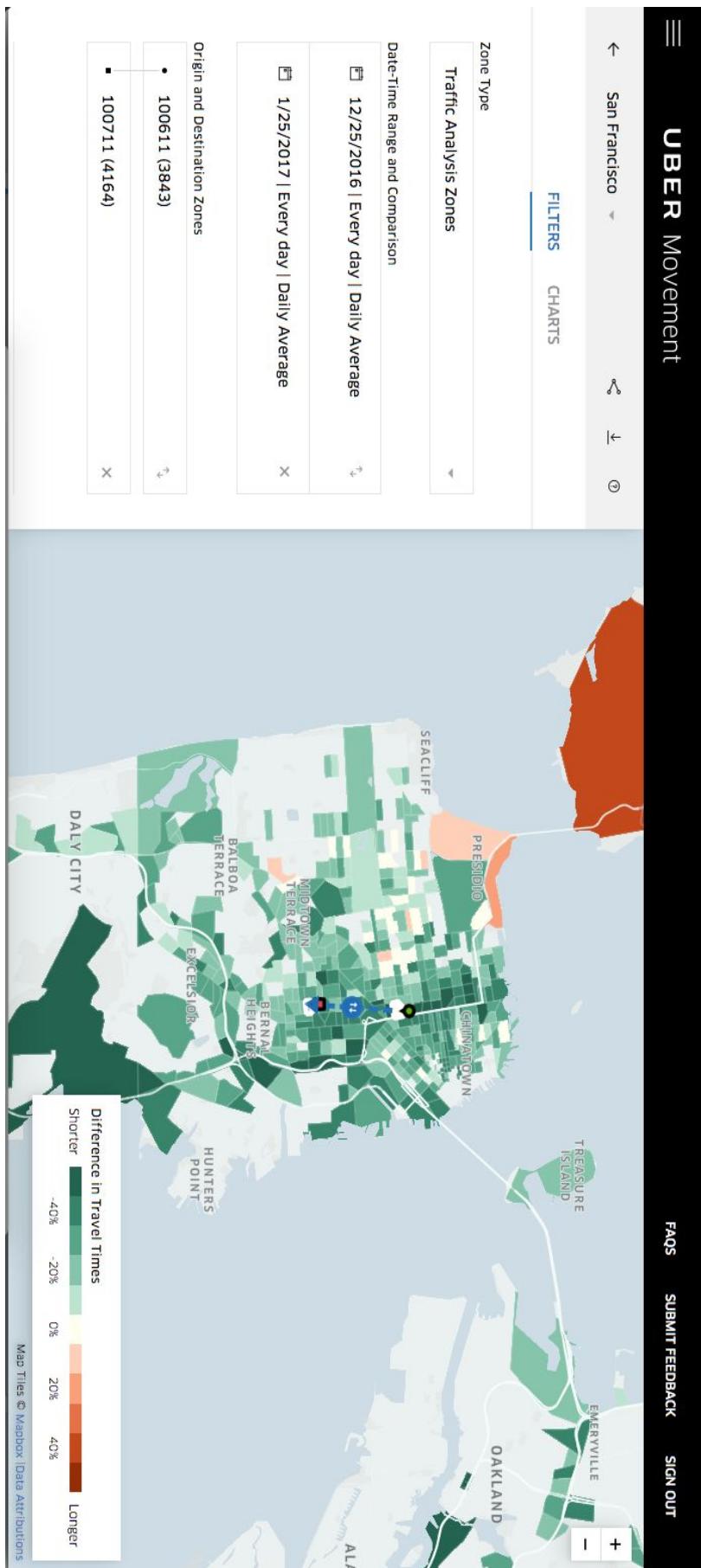
▶ Watch how Movement works

Select a city to explore ▾

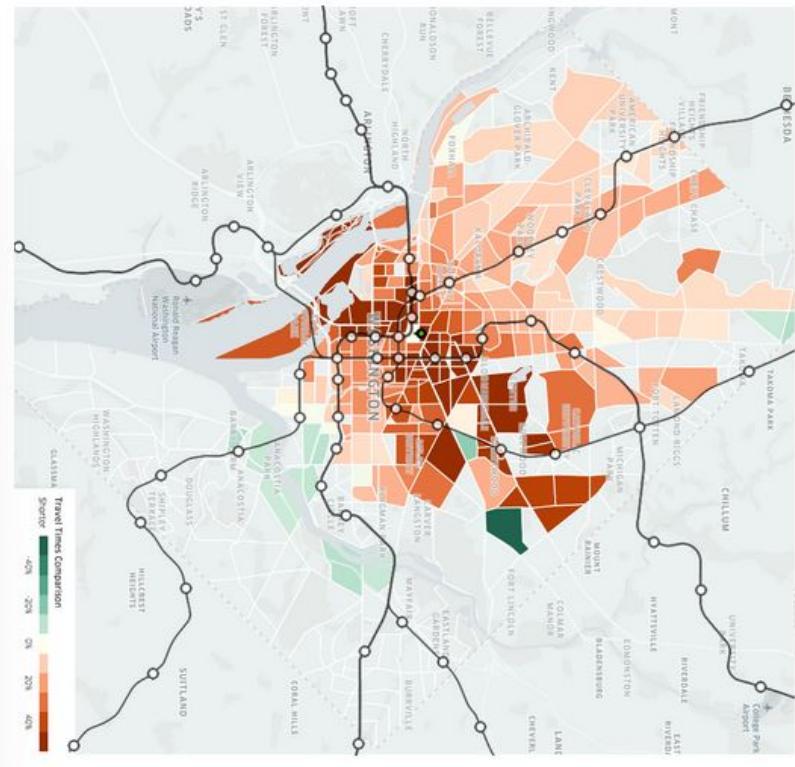
# Public Holidays



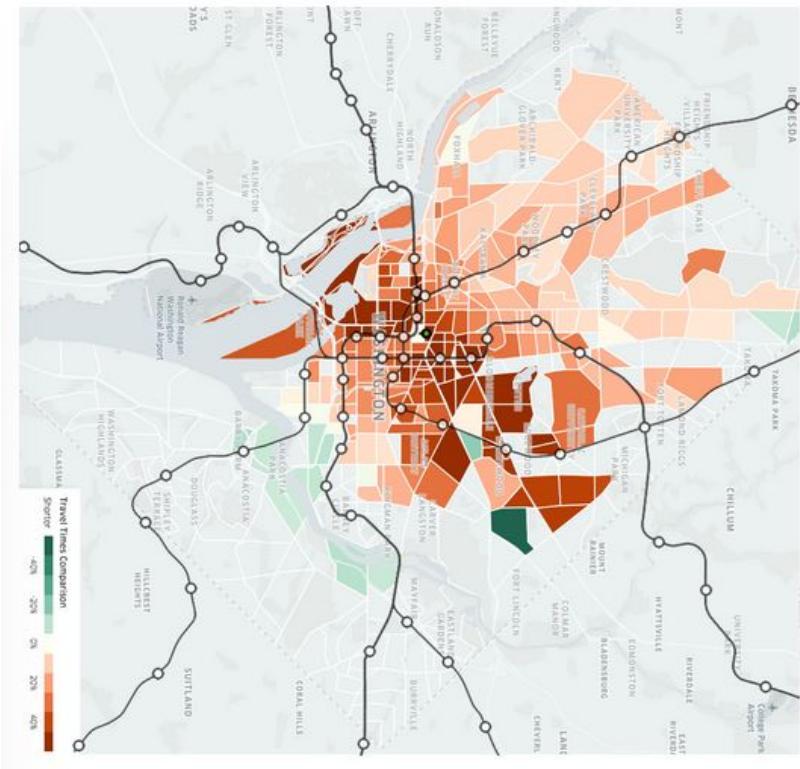
# Public Holidays



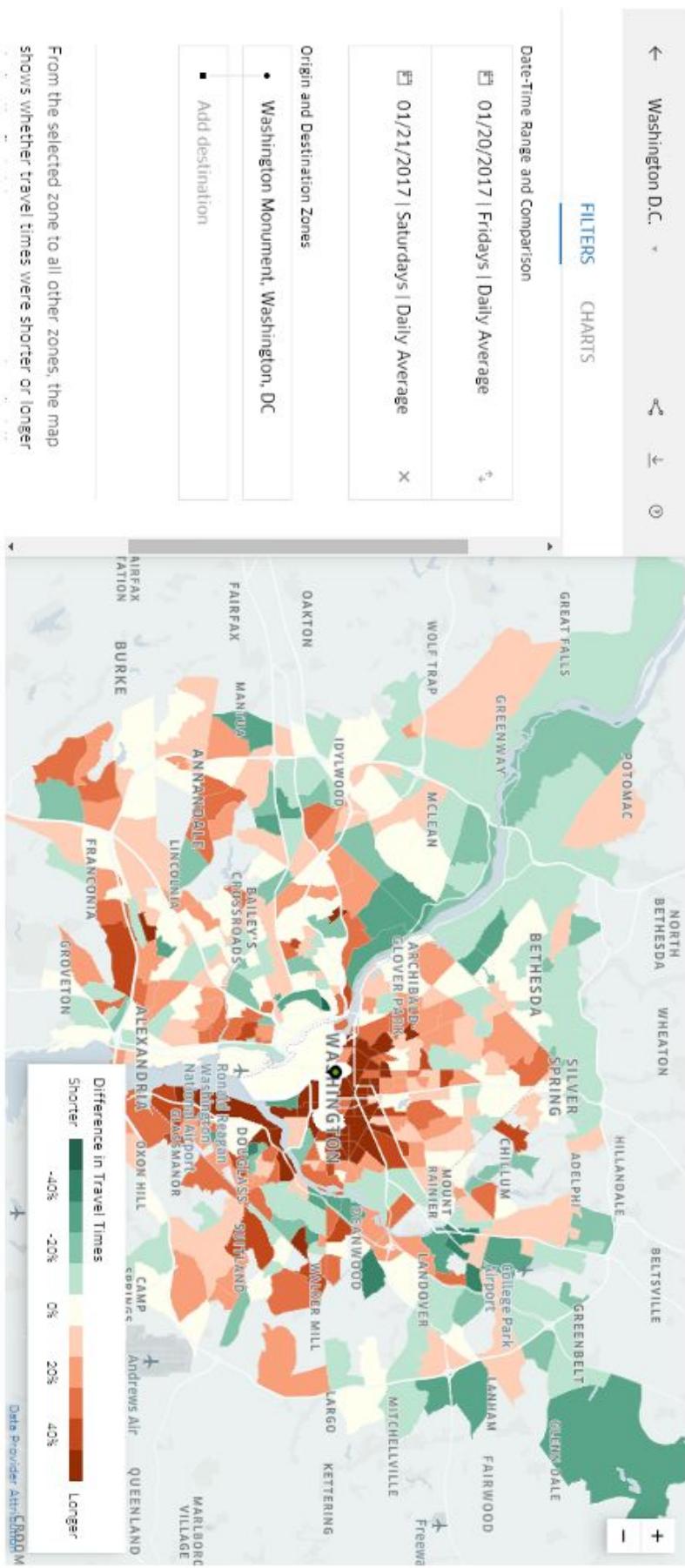
# Transport problems



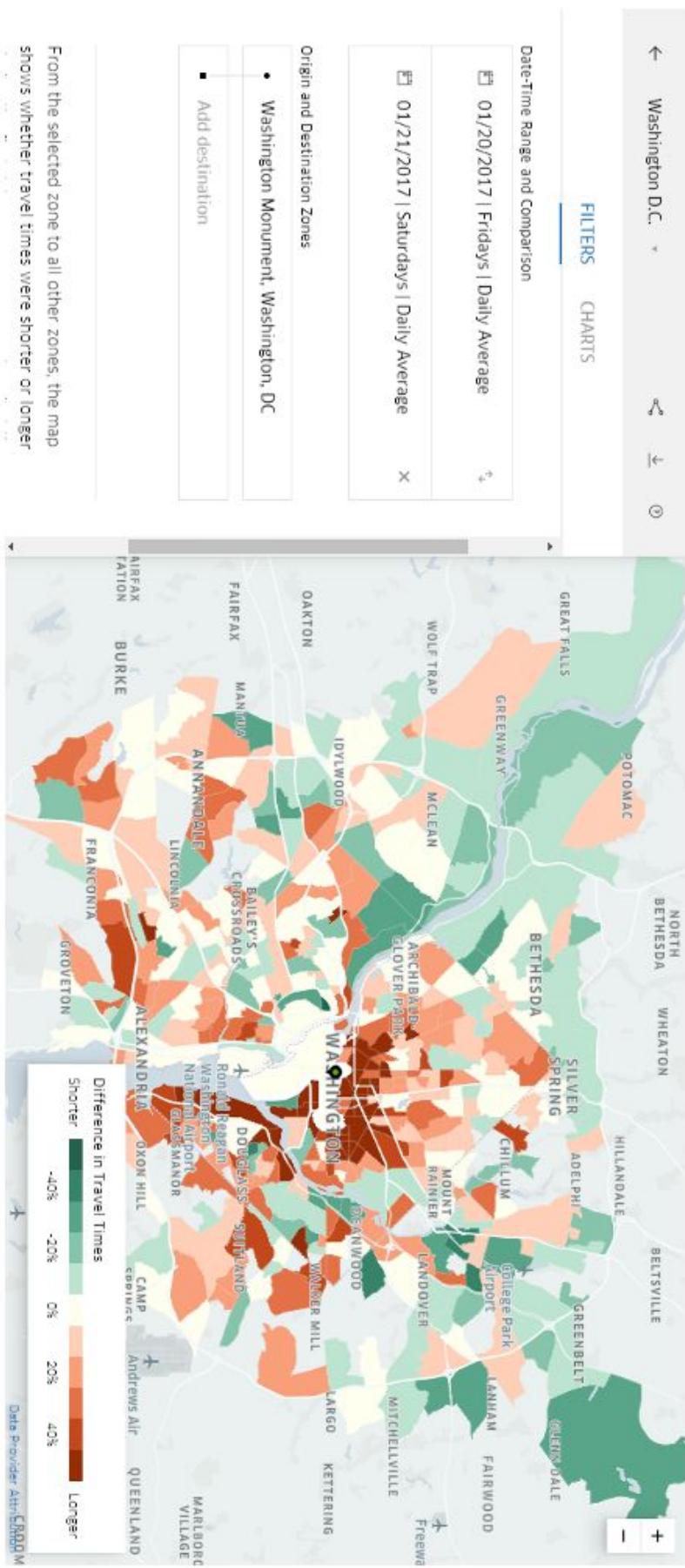
# Transport problems



# Special Events



# Special Events



# **Understanding The Ideological Stances of Chinese Opinion Leaders:**

A Story told by Social Media Data

Yinxian Zhang

Department of Sociology

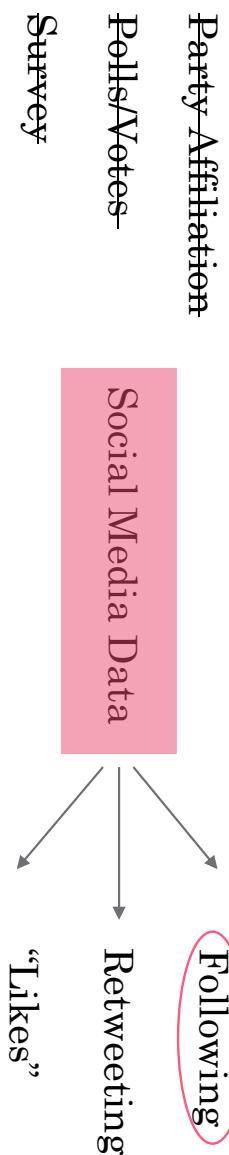
# Research Question One

1. How to measure Chinese opinion leaders' ideological preferences?

- Party Affiliation
- Polls/Vote
- Survey

# Research Question One

1. How to measure Chinese opinion leaders' ideological preferences?

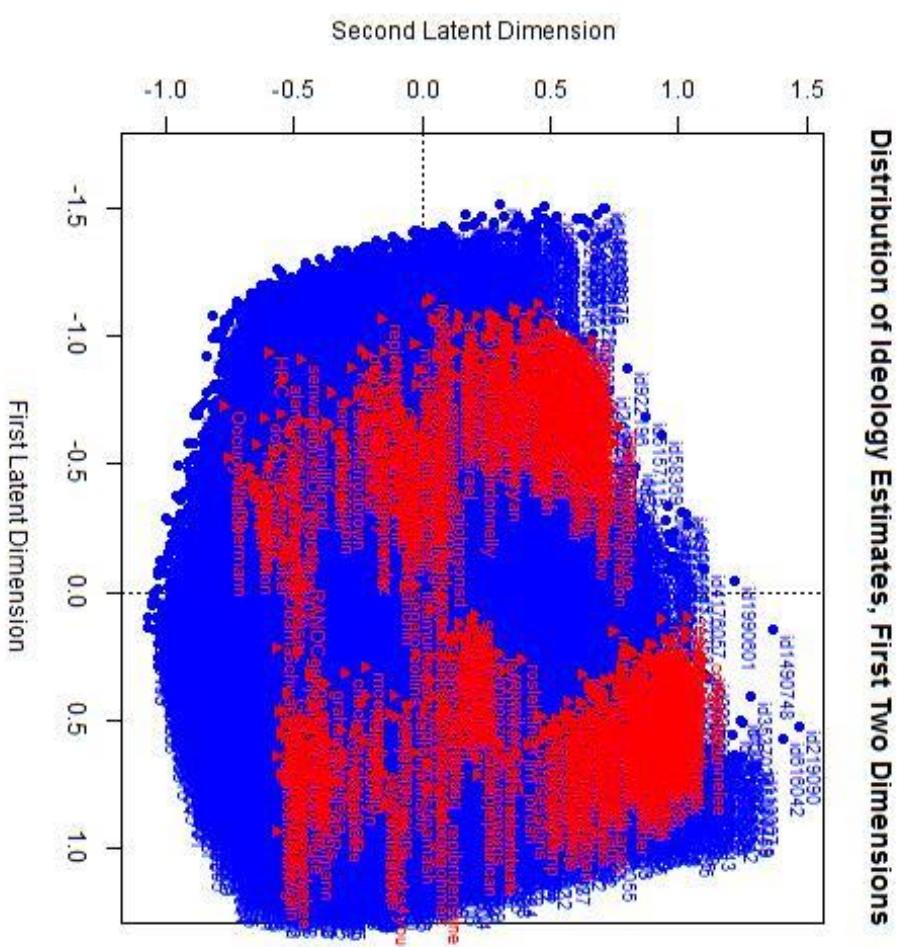


## Method

$$p(Y_{ij}=1 \mid \alpha_i, \beta_j, d_{ij}) = \text{Logit}(\alpha_i + \beta_j - d_{ij}),$$

$\alpha_i$ : the “out degree” of an ordinary user  $i$ ;  
 $\beta_j$ : the “in degree” of an opinion leader  $j$ .

*d: the distance between i and  
ideological space*



**Barberá P**, Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, doi: 10.1177/0956797615594620.

# Research Question Two

2. How to make sense of the ideological clustering?

- It is also a **validation** of the original estimation.

\*\* Note that the connotations of the political “left” and “right” are very different in China from what they mean in western societies.

# Research Question Two

What opinion leaders talk about when they talk about “democracy”?

# Method

Social media posts mentioning “Democracy/民主”

< 140 characters **short texts**

- Topic detection: LDA? v.s. Hierarchical clustering
- Word embedding: W2V model
- Qualitative reading

# Data

- 2.7 billion **posts** produced by over 170 million active users in 2013 on **Weibo**, plus their user profiles (**following relationships**), obtained via APIs.

- 228 opinion leaders identified. [e.g. Ren Zhiqiang(任志强), Sun Liping(孙立平), Wuyue Sanren(五岳散人), Zuoyeben(作业本), Sima Nan(司马南), Hu Xijin(胡锡进), Dai Xu(戴旭), Cai Xiaoxin(蔡小心)...]

# Data

- 2.7 billion **posts** produced by over 170 million active users in 2013 on **Weibo**, plus their user profiles (**following relationships**), obtained via APIs.
- 228 opinion leaders identified. (Zhang, Yinxian, Jiajun Liu and Jirong Wen. *forthcoming*. “Nationalism on Weibo”. *The China Quarterly*. )
- IRB approved.

	Mean	SD	Min	Max
Follower count	2272875	4904651	24799	3.45e+07
Following count	1385.25	895.45	0	3685
Friends count	1143.53	791.16	0	3623
2013 post count	443.21	434.70	30	3092

# Expected outcomes

- 1. Getting ideological positions (and clustering) of Chinese opinion leaders.
  - Easy to extend to ordinary people.
  - **One of the pioneering studies** to measure ideological preferences of Chinese people using observational data.
- 2. Interpreting the ideological stances of different opinion leader clusters.
  - Understanding the specific connotations of “democracy” in the Chinese context.
  - What are people’s opinions towards democracy? → joining in the scholarly efforts to explain the (lack of) democratization process in China.

# Actor, Culture Or Structure? A Predictive Model on News Article Popularity

Weiwei Zheng

# Introduction

## What influences News Popularity?

- Agenda Setting vs. Agenda Building

*Media influences what we think.*

*Transfer of the media agenda is a reciprocal process.*

-- Rogers, Dearing, 1988

- Audience takes an active role in virtual sphere during the process of information diffusion. Human **autonomy matters**.
- News popularity can be predicted by crowd manipulation.

-- *D. Horne and Adali, 2017*

# Introduction

## What shapes human behavior?

- "Duality of structure" : social practice, which is the principal unit of investigation, has both a **structural** and an **agency-component**.

-- Anthony Giddens

- Culture is an important mechanism.
- "A critical site of social action and intervention, where power relations are both established and potentially unsettled." -- Stuart Hall
- Headline negativity, subjectivity and overall sentiment influence news popularity.

-- Reis, Julio, etc., 2015

# Research Questions

**Goal:** To build a predictive model on news popularity

- Can active participants/opinion leaders, news content, and network structure altogether predict the popularity of online news article?
- Hypothesis: Yes.

**Significance of the Topic:**

- 1) Touch on the relationship between network structure and info diffusion (ameliorate communication strategy online).
- 2) Look into the mechanism and process of developing online public opinion.

# Informal Model

**Dependent Variable** – News popularity  
*an aggregate of communication effect and public opinion*

**Predictive Variables** – three dimensions

- Structure
- Audience Network Attributes
- Agency
- Opinion Leaders (number of actors of different roles)
- Culture
- News articles' content, sentiment, media agency
- Confounders – Date of published, total time of view

# Data Sources

- Reddit *an American social news aggregation, web content rating, and discussion website.*

## • API: Praw

### • /r/news, /r/worldnews (top section)

The screenshot shows the top news section of the Reddit homepage. It includes a sidebar with links to other subreddits like /r/intherews, /r/worldnews, and /r/politics. The main content area displays several news items:

- Legend aren't born. They're forged on the open seas. Set sail across a shared world to create your own pirate legend** (promoted by Sea of Thieves)
- Scientist Stephen Hawking has died aged 76** (news.sky.com)
- Self-pawned F.C. Announces Plan to Repeal Net Neutrality** (nytimes.com)
- Apple admits it slows older iPhones, confirming Geekbench report** (net.com)
- Net Neutrality Overturned** (nytimes.com)
- Japanese firm gives non-smokers extra six days holiday to compensate for cigarette break** (independent.co.uk)
- Chester Bennington of Linkin Park commits suicide** (nbc.com)
- #DeleteFacebook Movement Gains Steam After 50 Million Users Have Data Leaked** (newspap.com)

## Python + PRAW Working with Reddit API (p. 1)





# Methods

## Data Processing

- API: Networkx, NLTK
- Average clustering, transitivity, centrality, connectivity, etc.
- Algorithm to differentiate different types of users
- Text Classification (Bayes Classifiers, SVM, sentiment)
- Topic Modeling (Generative Models and LDA)

## Sample size

- now 1,000 top news/worldnews and several hundred comment trees for each)

## Model Selection

- Machine Learning: lasso regression, principal component regression, spline

# Few Points to Mention

## *Algorithm Confounding Issues*

- Probably only top news can be scrapped  
*(Reddit limits number of objects to be returned by Praw within 1000)*
- Unpredictable size of comment data
- Data sources, sample size and methods are subject to change

## *Measurement Issues*

- 1) unable to measure the impact from the offline world
- 2) silent participants are filtered
- 3) unable to study the interaction between different dimensions

## Citations

Buntain, Cody, and Jennifer Golbeck. "Identifying social roles in reddit using network structure." In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 615-620. ACM, 2014.

Horne, Benjamin D., and Sibel Adali. "The impact of crowds on news engagement: A reddit case study." *arXiv preprint arXiv:1703.10570* (2017).

Lippmann, Walter. *Public opinion*. Routledge, 2017.

Mrogers, Everett, and James Wdearing. "Agenda-setting research: Where has it been, where is it going?." *Annals of the International Communication Association* 11, no. 1 (1988): 55-594.

Reis, Julio, Fabricio Benevenuto, P. Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. "Breaking the news: First impressions matter on online news." In *Proceedings of the 9th International AAAI Conference on Web-Blogs and Social Media*. 2015.

Zaman, Tauhid, Emily B. Fox, and Eric T. Bradlow. "A Bayesian approach for predicting the popularity of tweets." *The Annals of Applied Statistics* 8, no. 3 (2014): 1583-1611.