上海交通大学国际与公共事务学院第十届

“暑期社会科学方法班” （政治学与国际关系）

文本分析与因果推断

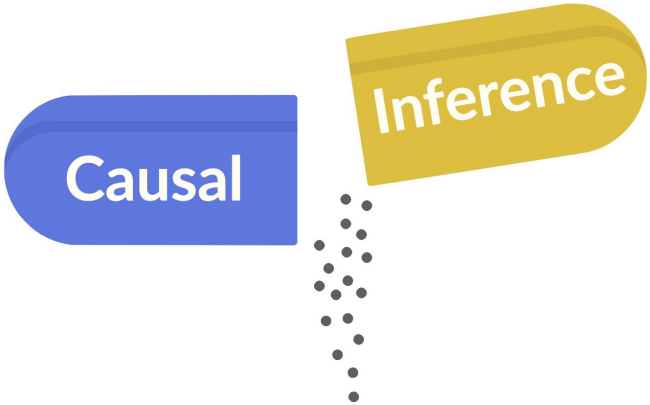
上海交通大学 国务学院 国际关系系 助理教授

付 舒

2024年8月30日

今天

1. 自我介绍
2. 文本分析回顾
3. 文本与因果推断
4. 自由讨论



特色

1. 文本分析如何在研究中**应用**，方法服务于研究问题！
2. 文本分析与因果推断结合
3. 分享一些code

1. **自我介绍**

2. 文本分析回顾

3. 文本与因果推断

4. 自由讨论

自我介绍：付舒

工作经历

- 2023 ~　上海交通大学　国务学院国际关系系　助理教授
- 2021 ~ 2023　芝加哥大学　政治学系与本科生院　讲师

教育背景

- 2015 ~ 2021　芝加哥大学　政治学系　博士　美国政治 & 量化方法
- 2008 ~ 2015　清华大学　国际关系学系　本科、硕士

研究兴趣

美国政治制度（总统、国会、官僚）

- The Filibuster and Legislative Discussion *Journal of Politics* (Fu & Howell 2023)
- Particularism or Policy *Political Research Quarterly* (Fu 2023)
- Direct Appeals of First Ladies *Presidential Studies Quarterly* (Fu & Savel 2020)
- Interbranch Messaging *Presidential Studies Quarterly* (Fu & Howell 2020)
- U.S. Ambassadors and Home-State Trade *World Politics* (Kim & Fu 2025)
- Moderates on Capitol Hill (Fu)

美国选举

- Primaries and Congressional Polarization (Fowler & Fu)
- 特朗普与美国民粹主义 （付舒）
- 问责与选拔：美国选举政治的逻辑谬误（付舒、罗兆天）

中美关系

- 美国涉台舆论分析 《世界经济与政治》 （张传杰、付舒 2012）

研究方法

- **Evidence-based**：定量分析、因果推断、量化文本分析

- 方法训练
  - 博弈论模型
  - 定量方法
    - 数学与统计　Math and Statistics
    - 计算社会科学　Computational Social Sciences
    - 回归分析　Linear Models
    - 最大似然估计　MLE/Model Based Inference
    - 因果推断　Causal Inference (Hong + Grimmer + Fowler)
    - 量化文本分析　Text-as-Data

- 开设：《定量研究方法II：因果推断》

1. 自我介绍
2. **文本分析回顾**
3. 文本与因果推断
4. 自由讨论

社会科学中的文本分析*

基于机器学习（machine learning）的文本分析（Text as Data）

- 发现（Discovery）
- 测量（Measurement）
- 因果推断（Causal Inference）

机器学习与文本分析很强大，但仍要意识到其局限性。

* 特别致谢Justin Grimmer的Text-as-Data课程

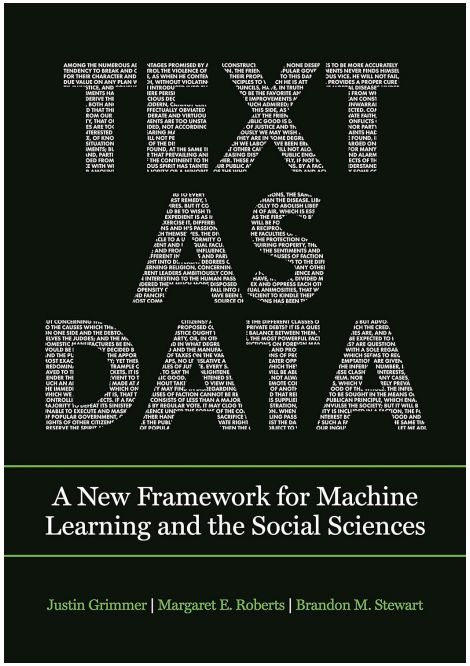
文本分析的兴起

近年来，文本分析（Text-as-Data Methods）方兴未艾

- 2000年之前：
  - 贵（0.1元/条短信）
  - 编码者一人一样
  - 计算机算力达不到

- 2000年之后：
  - 便宜（2023: << $0.0001 / mb）
  - 编码由统计模型决定
  - 计算机算力强大
  - **AI（ChatGPT、文心一言等）**

✓ 系统性地分析文本数据成为可能



为什么需要?

- 社会生活中，文字无处不在
  - 法律
  - 条约
  - 媒体报道
  - 选举活动
  - 新闻稿
  - 政治家发言
  - Twitter/X
  - …

- 文字数量巨大，浩如烟海!

为什么需要?

不仅仅因为文字量巨大（not just for “big data”）

- 手动把100个文件进行分类
  - Bell(n) = 把n个文件分组的组合数
  - Bell(2) = 2 (AB, A B)
  - Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
  - Bell(5) = 52
  - Bell(100)

为什么需要?

不仅仅因为文字量巨大（not just for “big data”）

- 手动把100个文件进行分类
  - Bell(n) = 把n个文件分组的组合数
  - Bell(2) = 2 (AB, A B)
  - Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
  - Bell(5) = 52
  - Bell(100) $\approx 4.75 \times 10^{115}$

雇700万个助研

工作加班加点（24/7/365）

$\approx 1.54 \times 10^{84} \times$ (14,000,000,000) 年

机器学习方法哪怕
对于小问题来说，
都十分有用!

量化文本分析能做什么？

- 基于机器学习的文本分析能够**提升阅读效率**

- Justin Grimmer关于麦子的比喻：
  - 分析各种麦子：理解含义
    - 人类很强大，但是计算机不太行
  - 整理麦堆：比较、整理、分类
    - 人类很挣扎，但是计算机很好！

量化文本分析**不能**做什么？

- **永远不能取代人的阅读**
- 不能设计一种全面的语言统计模型
- 不能设计一个单一的工具并应用在所有任务上

声明

- 跨学科：计算机领域 Natural Language Processing (NLP)、统计学、语言学、社会科学等，今天关注政治学
- 今天所讲的量化文本分析，都是基于英语（后面介绍的一些概念暂不做翻译）
- 文本分析在政治学的初创应用，甚至是从研究中文开始 (King, Pan, Roberts, 2013, APSR)
- 不同语言有不同的特色，对分析的前期处理要求不同
- 但是，一旦我们把**文字**转化成**数字**，后期的处理是一样的

基本概念

- (text) corpus — a large and structured set of texts for analysis
- document — each of the units of the corpus (e.g. a FB post)
- tokens — any word – so token count is total words

Document 1: `A corpus is a set of documents.`

Document 2: `This is the 2nd document in the corpus.`

This is a corpus with 2 documents, where each document is a sentence. The first document has 7 tokens. The second has 8 tokens.

基本概念

- stems — words with suffixes removed (using set of rules)
- lemmas — canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)

word	win	winning	wins	won	winners
stem	win	win	win	won	winner
lemma	win	win	win	win	win

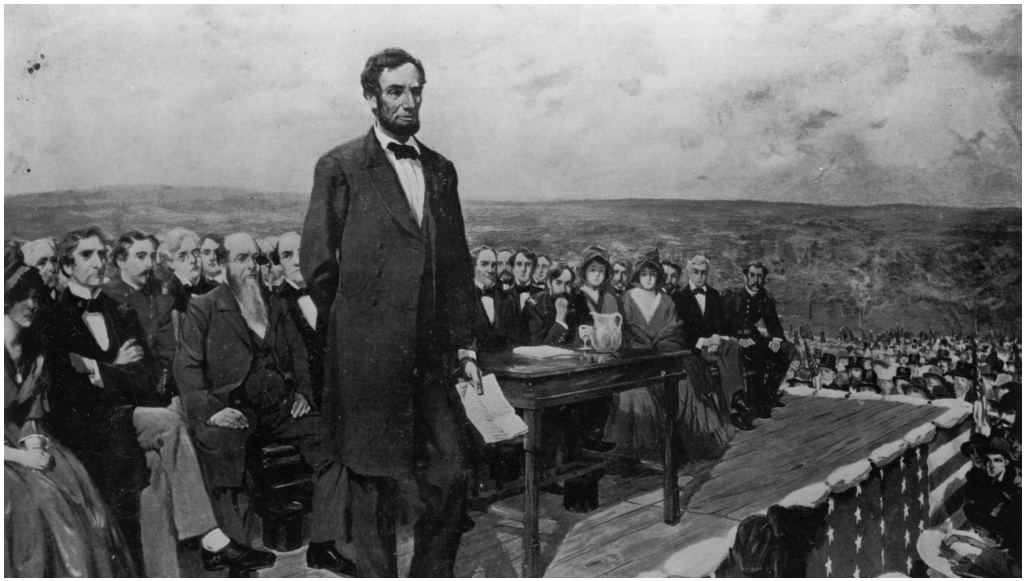
- stop words — words that are designated for exclusion from any analysis of a text, e.g., a, an, the, it, be, because, …

前期处理 Preprocessing

One (of many) recipe for preprocessing: retain useful information

1. Remove capitalization, punctuation
2. Discard word order (Bag of Words Assumption)
3. Discard stop words
4. Create Equivalence Class: Stem, Lemmatize, or synonym
5. Discard less useful features (depends on application)
6. Other reduction, specialization

举例：Gettysburg Address



- 葛底斯堡演说，是亚伯拉罕·林肯总统于1863年11月19日在宾夕法尼亚州葛底斯堡发表的一次著名演说
- 这次演说是在葛底斯堡战役结束后的全国公墓奉献仪式上进行的，目的是纪念在这场内战中牺牲的士兵
- 葛底斯堡演说以其简短、深刻和感人的特点而著名，全文只有约272个字，却对美国历史和文化产生了深远的影响

Original Text

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure.

Remove capitalization, punctuation
Discard word order (bag of words)

```
now we are engaged in a great civil war
testing whether that nation or any nation so
conceived and so dedicated can long endure
```

Discard stop words

now we are engaged in a great civil war
testing whether that nation or any nation so
conceived and so dedicated can long endure

Keep the word stems

now we are engaged in a great civil war
testing whether that nation or any nation so
conceived and so dedicated can long endure

Count vector: each element counts occurrence of stems

now we are engaged in a great civil war
testing whether that nation or any nation so
conceived and so dedicated can long endure

unigram	civil	conceive	dedicate	endure	engage	great	long	nation	test	war
count	1	1	1	1	1	1	1	2	1	1

How Could This Possibly Work?

Speech is

- Ironic or Sarcastic:
  - I'm absolutely confident that Joe Biden is in perfect health and will easily make it to November.
- Subtle in negations:
  - Hope this is not inappropriate, …
  - I ain't got no mood tonight.
- Order Dependent:
  - Peace, no more war
  - War, no more peace

How Could This Possibly Work?

Three answers

1. **It might not**: **Validation** is critical (task specific)

2. **Central Tendency in Text**: Words often imply what a text is about
   - `war, civil, dead, died, lives, nation`
   - Likely to be used repeatedly: create a theme for an article

3. **Human supervision**: Human judgement (coders) helps methods identify subtle relationships between words and outcomes of interest

文本分析在政治学的常用方法

- 发现 Discovery
  - 主题模型 Topic Models
- 测量 Measurement
  - 字典方法 Dictionary Methods

*

* 王宇： “大语言模型在政治科学的应用”

主题模型

- 主题模型是对文本进行**分类**的一种方法
- 目的是发现文本中的主题/议题

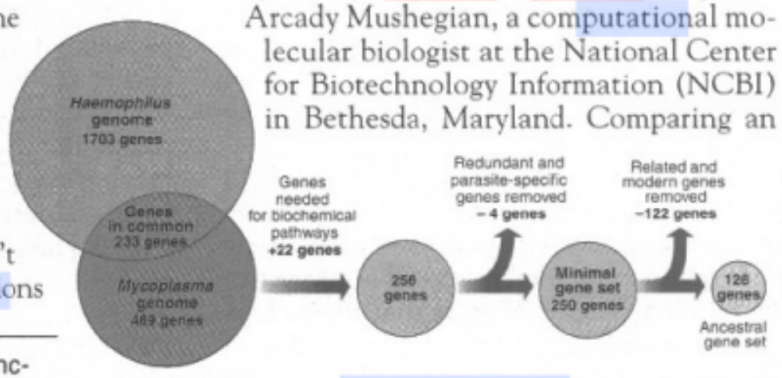
- Latent Dirichlet Allocation (LDA) 是一种比较流行的主题模型
  - Unsupervised
  - It treats each document as a mixture of topics, and each topic as a mixture of words.

Latent Dirichlet Allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

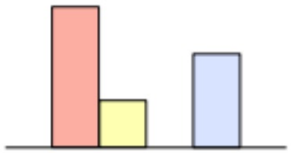
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.



主题模型

- 主题模型有各种各样，有无监督的（unsupervised）
  - (Vanilla) Latent Dirichlet Allocation (Blei, Ng, and Jordon 2003)
  - Structural Topic Models (Roberts, Stewart, and Tingley 2017)
  - …
- 有监督的（supervised）
  - Keyword Assisted Topic Models (Eshima, Imai, Sasaki 2020)
  - …
- **主题模型的核心**是把文本内的主题分类（K），测量每个主题的占比（prevalence），展示每个主题的高频词（frequent words）

字典方法 Dictionary Methods

比如，测量有多快乐:

- 一个国家人民多快乐?
- 一首歌多快乐??
- 一个Twitter内容有多快乐?

我们可以用字典方法（Dictionary Methods）

字典方法 Dictionary Methods

Dodds and Danforth (2009) 采用字典方法测量快乐

- Affective Norms for English Words (ANEW)
- On a scale of 1-9, how happy does this word make you?
  - Happy : triumphant (8.82), paradise (8.72), love (8.72)
  - Neutral: street (5.22), paper (5.20), engine (5.20)
  - Unhappy : cancer (1.5), funeral (1.39), rape (1.25), suicide (1.25)

- Happiness for text i (with word j having happiness θ_j and document frequency X_{ij})

$$Happiness_i = \frac{\sum_{j=1}^{J} \theta_j X_{ij}}{\sum_{j=1}^{J} X_{ij}}$$

字典方法 Dictionary Methods

Lyrics for
Michael Jackson's Billie Jean

"She was more like a beauty queen
from a movie scene.
⋮
And mother always told me,
be careful who you love.
And be careful of what you do
'cause the lie becomes the truth.
Billie Jean is not my lover,
She's just a girl who claims
that I am the one.
⋮

ANEW words	v_k	f_k
k=1. love	8.72	1
2. mother	8.39	1
3. baby	8.22	3
4. beauty	7.82	1
5. truth	7.80	1
6. people	7.33	2
7. strong	7.11	1
8. young	6.89	2
9. girl	6.87	4
10. movie	6.86	1
11. perfume	6.76	1
12. queen	6.44	1
13. name	5.55	1
14. lie	2.79	1

$$v_{\text{text}} = \frac{\sum_k v_k f_k}{\sum_k f_k}$$

$v_{\text{Billie Jean}} = 7.1$

$v_{\text{Thriller}} = 6.3$

$v_{\text{Michael Jackson}} = 6.4$

字典方法 Dictionary Methods

GUESS: The happiest song on *Thriller*?

P.Y.T. (Pretty Young Thing) (This is the right answer!)

举例：第一夫人演讲 (Fu and Savel 2020)

PRESIDENTIAL STUDIES QUARTERLY
CENTER FOR THE STUDY OF THE PRESIDENCY AND CONGRESS

Article | Full Access

Policy without Partisanship: The Direct Appeals of First Ladies

Shu Fu, Meg Savel

First published: 10 September 2020 | https://doi.org/10.1111/psq.12678

举例：第一夫人演讲 (Fu and Savel 2020)

- 研究问题：第一夫人在美国政治中扮演什么样的角色？第一夫人如何与民众沟通？
- 传统观点：不参政，主妇角色，apolitical，hostess
- 新的理论：
  - 第一夫人讨论政策，go personal
  - 第一夫人讨论政策，go purple
  - “总统说官话，夫人说人话”
- 数据：第一夫人的所有公开演讲（N=1,264）
  - 希拉里·克林顿
  - 劳拉·布什
  - 米歇尔·奥巴马
- 方法：LDA + Dictionary

数据前期处理 Preprocessing

遵循量化文本分析的标准：

- 去掉无实质内容的发言
- 去掉标点、数字和其他符号
- 全部小写
- 保留词根
- 去掉stop words（a, an, the, he, she, …）
- **去掉具体情况的无意义词语**
  - “mr”, “mrs”, “ms”, “obama”, “barack”, “michelle”, “audience”, “laughter”, “applause”, “ve”

数据前期处理 Preprocessing

```
# Discard punctuation, capitalization, and split the text on white space
# Apply the Porter Stemmer to the tokenized documents
for z in MichelleLauraClintonRemarks:
    z[4] = re.sub(r'[\n]', ' ', z[4])
    z[4] = z[4].lower()
    z[4] = re.sub(r'\W+', ' ', z[4])
    z[4] = word_tokenize(z[4])
    z[4] = map(pt.stem, z[4])

# Discard the stemmed stop words
for z in MichelleLauraClintonRemarks:
    z[4] = [x for x in z[4] if x not in stop_words_stemmed]

# bigram
for z in MichelleLauraClintonRemarks:
    bigrams = nltk.bigrams(z[4])
    z.append(bigrams)

for z in MichelleLauraClintonRemarks:
    temp = map(bituple, z[5])
    z.append(temp) # stored in z[7]
```

LDA

Document-Term Matrix

Most frequent 2000 unigrams + 200 bigram

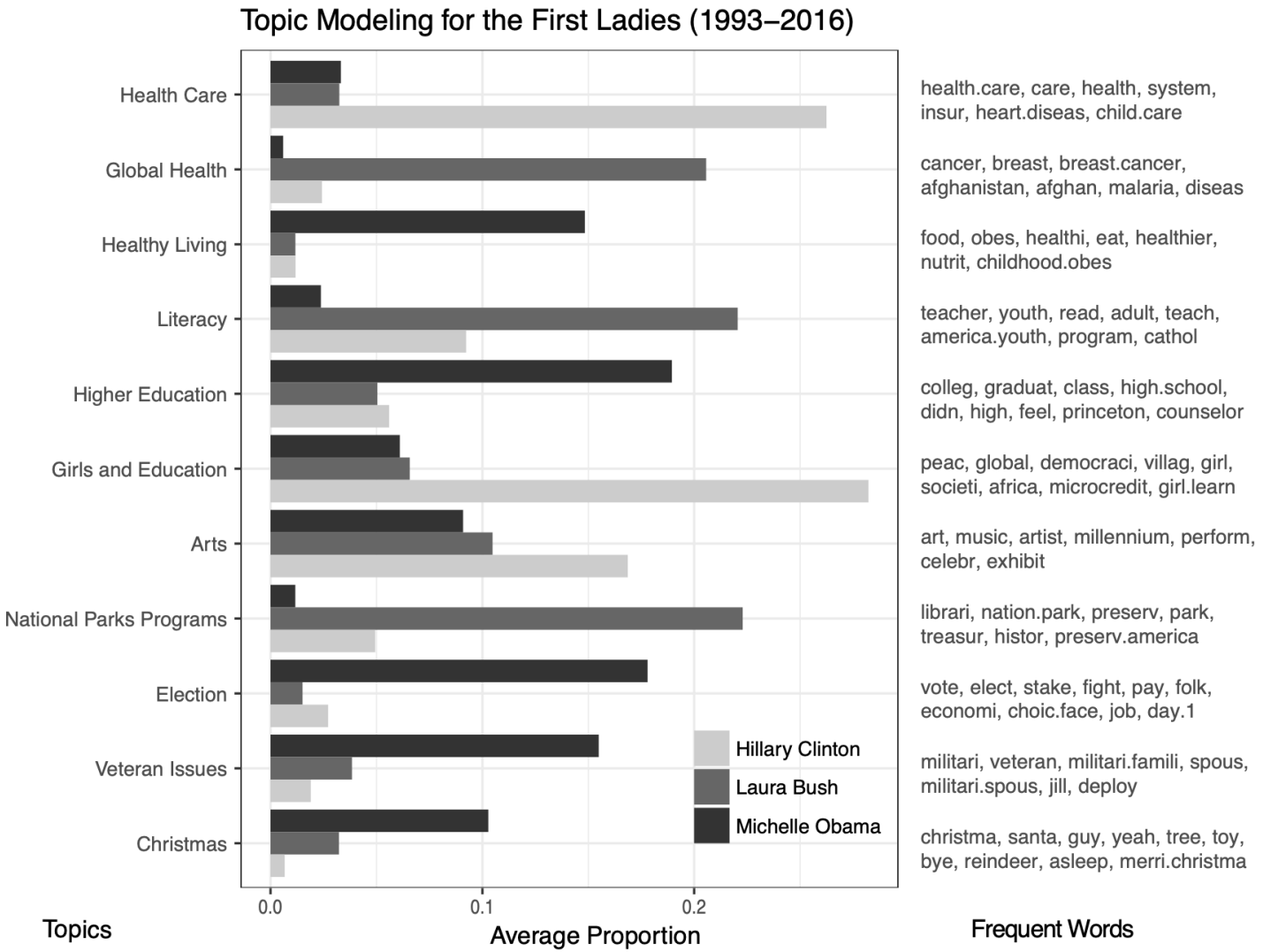
year	month	day	speaker	abil	abort	abroad	absolut	abus	academ	...	21st centuri	african american	america treasur
2012	Nov	3	MichelleObama	0	0	0	3	0	0		0	0	0
2016	May	17	MichelleObama	0	0	0	6	0	1		0	0	0
2010	Jul	7	MichelleObama	0	0	0	0	1	0		0	0	0
2010	Jun	4	MichelleObama	0	0	0	0	0	0		0	0	0
2012	Oct	10	MichelleObama	0	0	0	3	0	0		0	0	0
2011	Jan	28	MichelleObama	1	0	0	0	0	0		0	0	0
2013	Apr	23	MichelleObama	0	0	1	2	0	0		0	0	0
2013	Sep	12	MichelleObama	0	0	0	0	0	0		0	0	0
2011	Mar	15	MichelleObama	0	0	0	1	0	0		1	0	0

```
MLHout <- prepDocuments(MLHdocuments, colnames(MLHdtm2))
MLHstm.out <- stm(MLHout$documents, MLHout$vocab, K = 11, init = 'Spectral')

MLHtopic <- labelTopics(MLHstm.out)

# Percentage of Speeches in each topic model
apply(MLHstm.out$theta, 2, mean)
```

LDA结果：Go Personal



字典方法

- 字典方法：“Fightin’ Words” 模型
  （Monroe, Colaresi, and Quinn 2009）

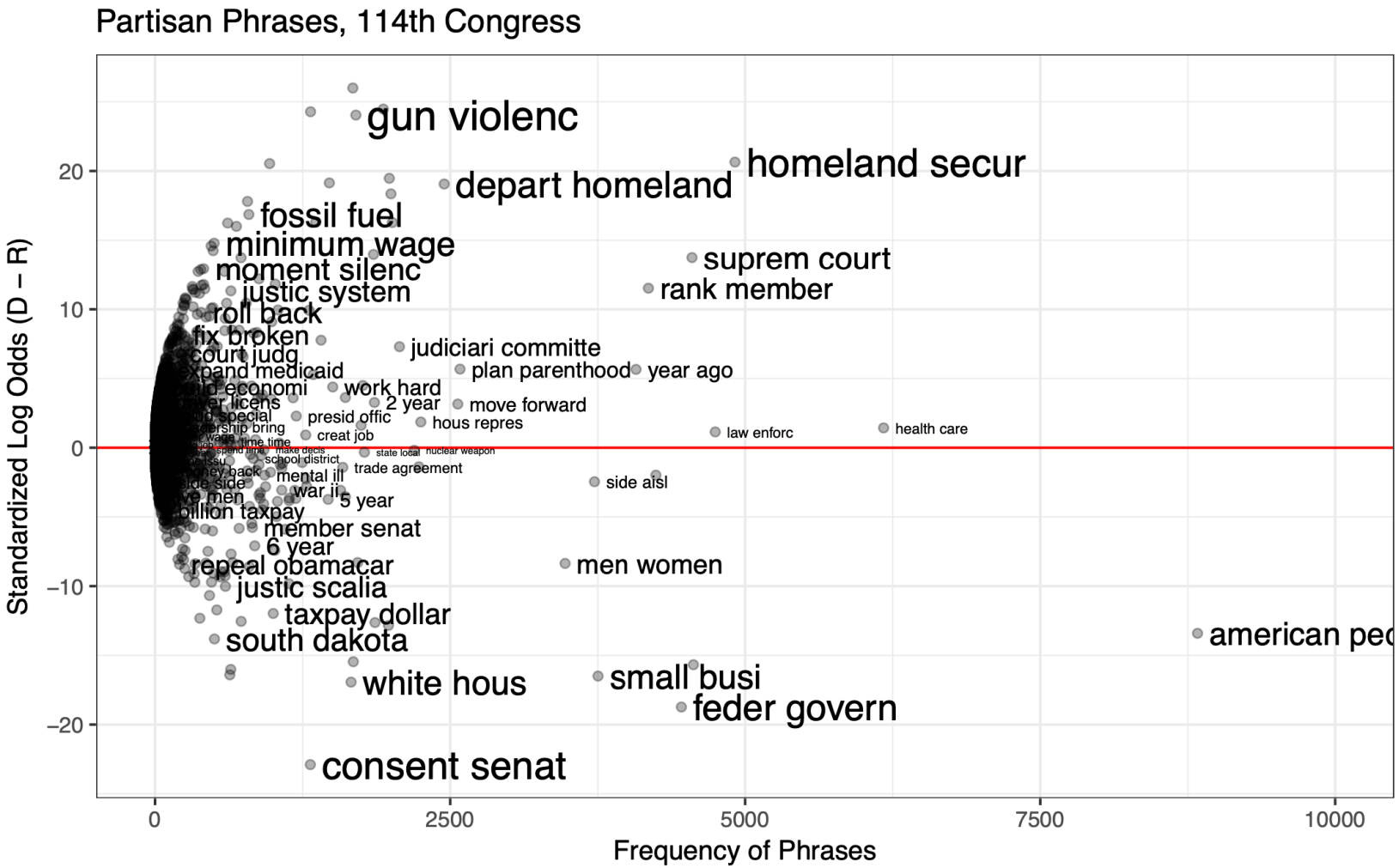
- 数据：国会议事大厅实录
- 简言之，经常被民主（共和）党人使用的词组带有民主（共和）党腔调。

phrase	D	R
tax loophol	2681	529
tax loss	173	173
tax lot	78	184

- 词组 Bigrams (two-word phrases)

- 建立一个“党派腔调字典”

党派腔调字典


Partisan Phrases, 114th Congress
Standardized Log Odds (D – R)
Frequency of Phrases
20
10
0
−10
−20
0
2500
5000
7500
10000
gun violenc
homeland secur
depart homeland
fossil fuel
minimum wage
moment silenc
justic system
suprem court
rank member
roll back
fix broken
judiciari committe
court judg
xpand medicaid
plan parenthood
year ago
work hard
2 year
move forward
presid offic
hous repres
creat job
law enforc
health care
time time
state local
nuclear weapon
school district
trade agreement
mental ill
side aisl
war ii
5 year
billion taxpay
member senat
6 year
repeal obamacar
men women
justic scalia
taxpay dollar
south dakota
american pe
white hous
small busi
feder govern
consent senat

计算第一夫人演讲的党派腔调

因变量赋值：

- 根据党派腔调字典，加权平均，计算第一夫人的演讲 i 党派腔调

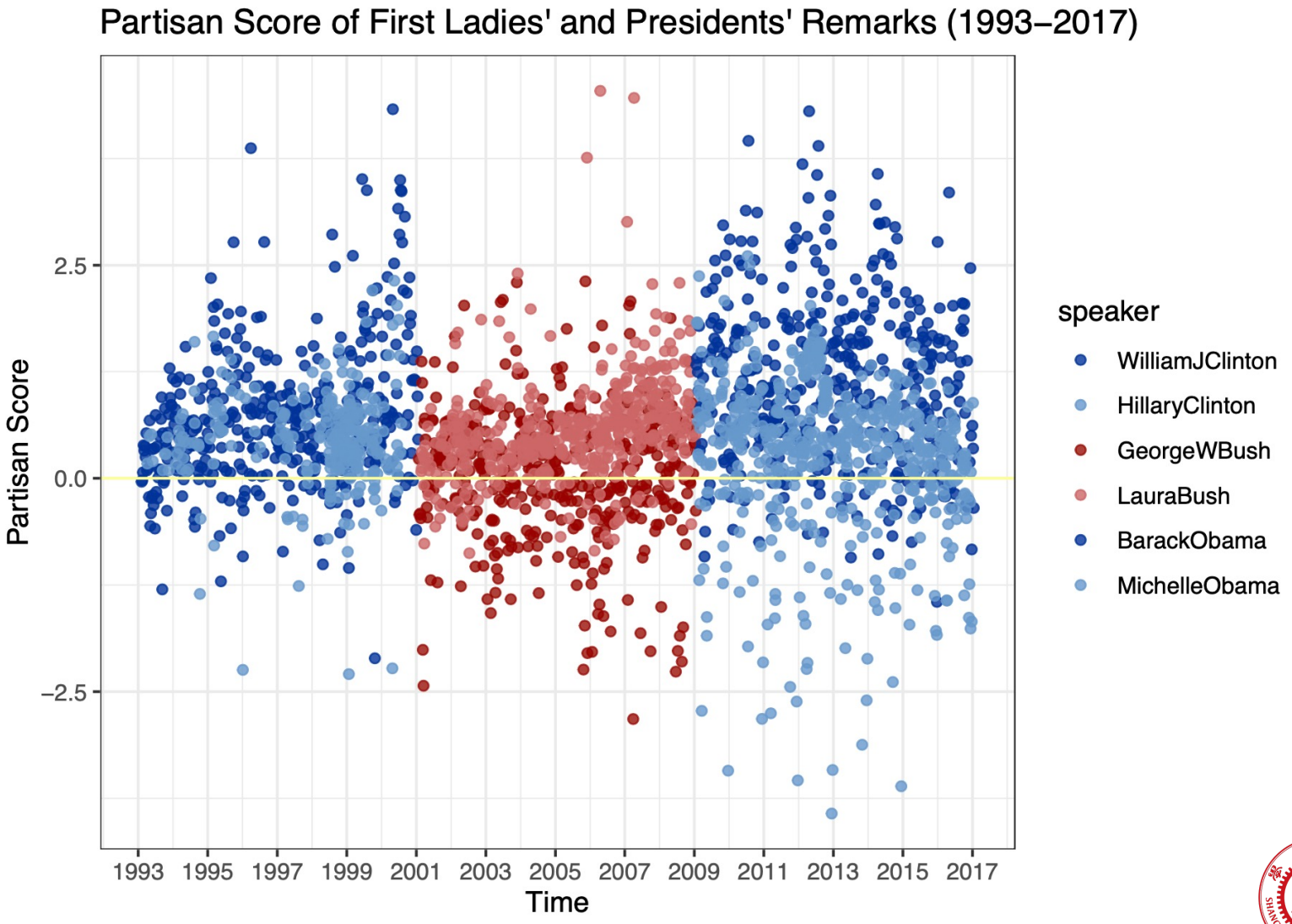
$$\text{Partisan Score}_i = \frac{\sum_{j=1}^{J} \theta_j X_{ij}}{\sum_{j=1}^{J} X_{ij}}$$

- 党派腔调是个连续变量

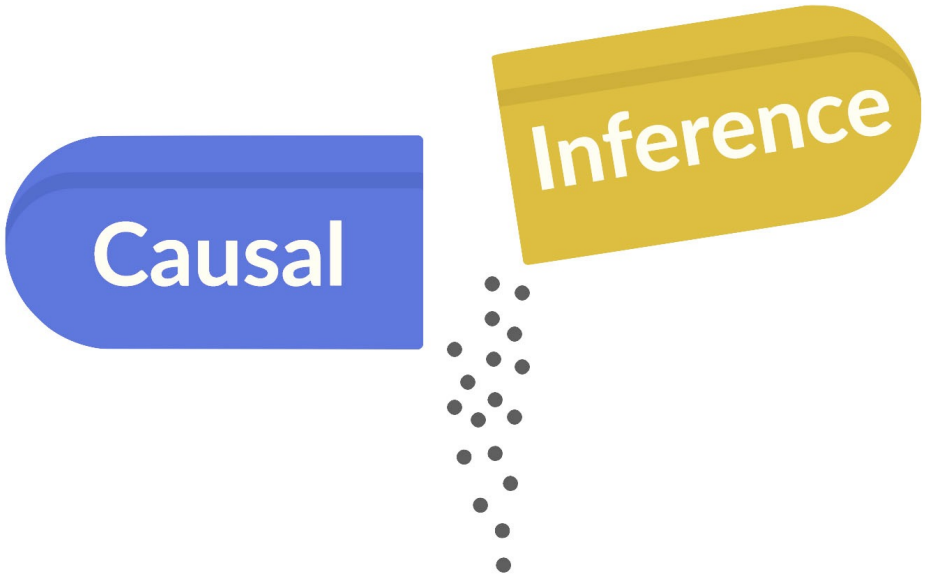
$$\text{Partisan Score}_i > 0 \quad \text{民主党腔}$$
$$\text{Partisan Score}_i < 0 \quad \text{共和党腔}$$
$$\text{Partisan Score}_i \approx 0 \quad \text{模糊}$$

返回

字典方法结果：Go Purple



1. 自我介绍
2. 文本分析回顾
3. **文本与因果推断**
4. 自由讨论



文本与因果推断

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

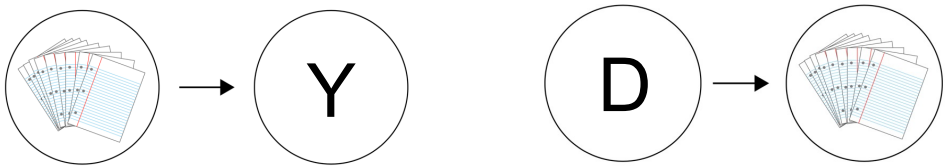
How to make causal inferences using texts

Naoki Egami[1]†, Christian J. Fong[2]†, Justin Grimmer[3,4]*†, Margaret E. Roberts[5,6]*†, Brandon M. Stewart[7,8]*†

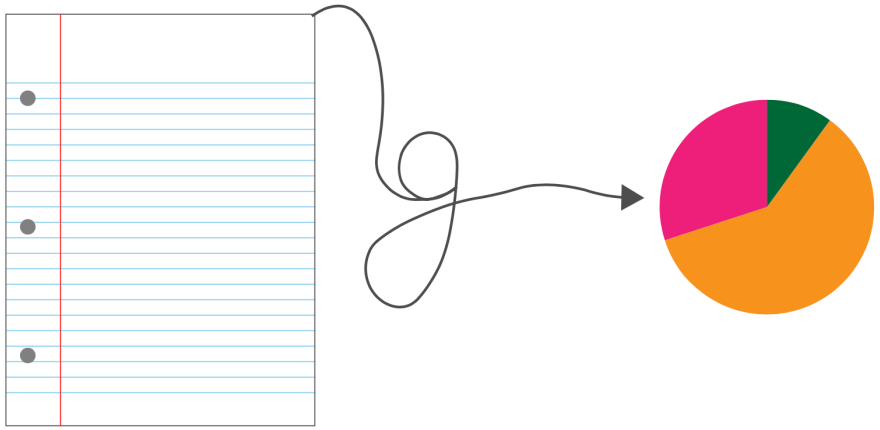
Text as data techniques offer a great promise: the ability to inductively discover measures that are useful for testing social science theories with large collections of text. Nearly all text-based causal inferences depend on a latent representation of the text, but we show that estimating this latent representation from the data creates underacknowledged risks: we may introduce an identification problem or overfit. To address these risks, we introduce a split-sample workflow for making rigorous causal inferences with discovered measures as treatments or outcomes. We then apply it to estimate causal effects from an experiment on immigration attitudes and a study on bureaucratic responsiveness.

文本与因果推断

- 文本可以作为干预变量，也可以作为结果变量

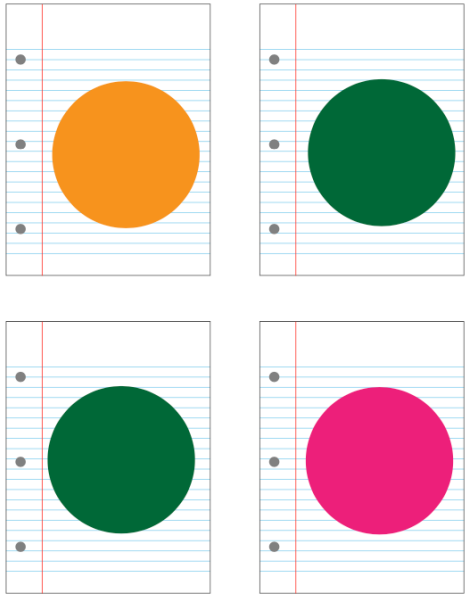


- 任何文本分析的核心是降维，把高纬度降到低维度
- 降维的过程通过一个g函数，也称codebook function
  - 比如，主题模型



g函数

- g函数，降维发现文本潜在的含义（latent representation）
- g函数可以得到不同的形式
  - 分类 Categorical
  - 主题分布 Mixed Membership
  - 零一特征 Binary Features
  - 连续特征 Scales

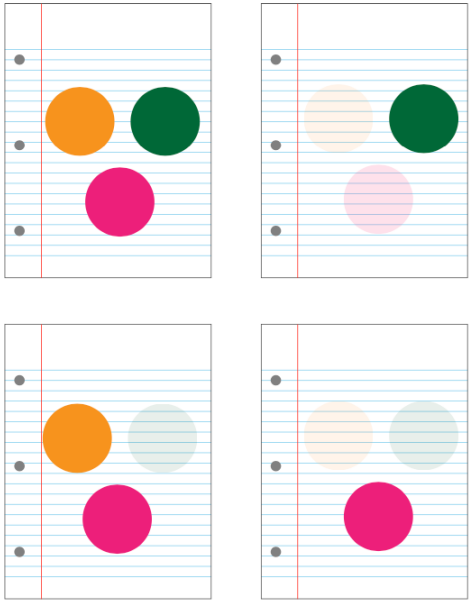


g函数

- g函数，降维发现文本潜在的含义（latent representation）
- g函数可以得到不同的形式
  - 分类 Categorical
  - 主题分布 Mixed Membership
  - 零一特征 Binary Features
  - 连续特征 Scales

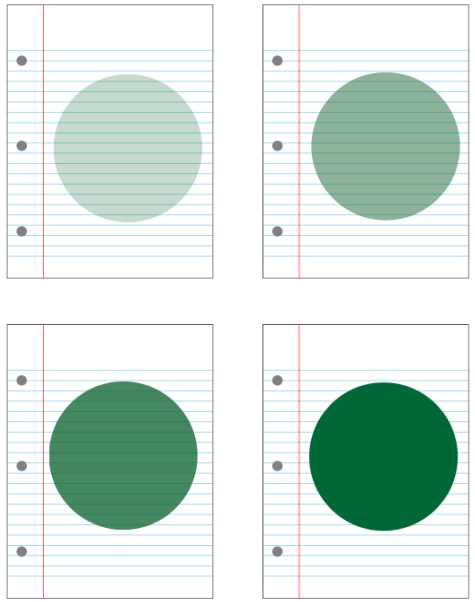
g函数

- g函数，降维发现文本潜在的含义（latent representation）
- g函数可以得到不同的形式
  - 分类 Categorical
  - 主题分布 Mixed Membership
  - 零一特征 Binary Features
  - 连续特征 Scales



g函数

- g函数，降维发现文本潜在的含义（latent representation）
- g函数可以得到不同的形式
  - 分类 Categorical
  - 主题分布 Mixed Membership
  - 零一特征 Binary Features
  - 连续特征 Scales



文本因果推断的根本问题

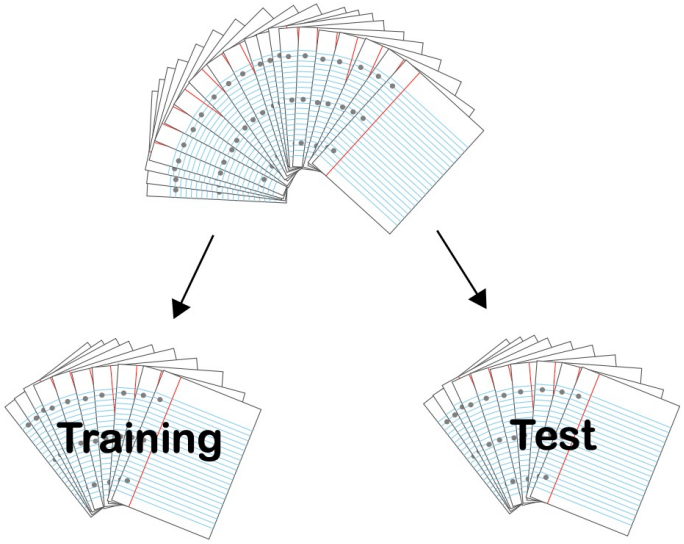
在降维的过程中，会遇到一个不易察觉的微妙的问题：

- Fundamental Problem of Causal Inference with Latent Variables (Egami, Fong, Grimmer, Roberts, and Stewart 2018)
  - 或Analyst-induced SUTVA violation
  - 无法使用相同的观测点，同时发现规律并做因果推断

研究者如何做?

训练–检测分离 Train-Test Split

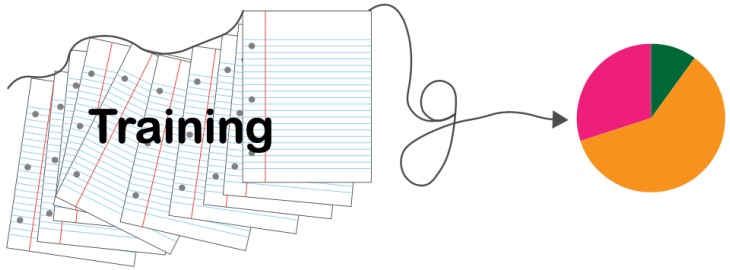
1. 明确训练组，用来发现



研究者如何做?

训练–检测分离 Train-Test Split

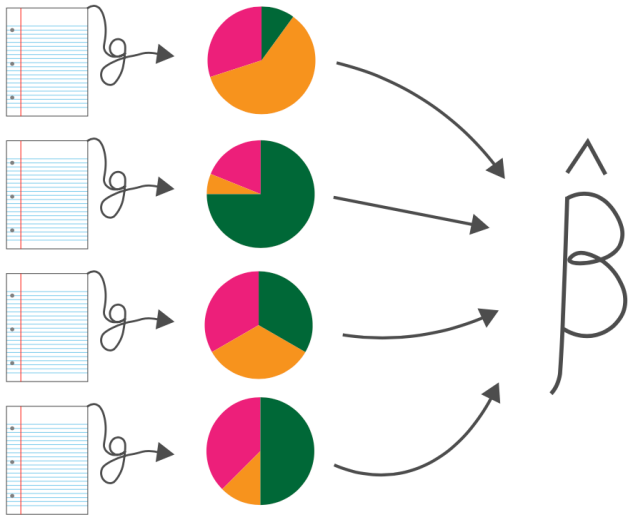
1. 明确训练组，用来发现
2. 用训练组来构建g函数，从而把高维度的文本数据降维到所需的测量值



研究者如何做?

训练–检测分离 Train-Test Split

1. 明确训练组，用来发现
2. 用训练组来构建g函数，从而把高维度的文本数据降维到所需的测量值
3. 确定好g函数后，用检测组估计因果效应



研究者如何做?

训练–检测分离 Train-Test Split

1. 明确训练组，用来发现
2. 用训练组来构建g函数，从而把高维度的文本数据降维到所需的测量值
3. 确定好g函数后，用检测组估计因果效应

Train-Test allows for discovery while avoiding possibilities of fishing.

举例：

- Text as Treatment

特朗普的推特如何影响选民？
(Fong and Grimmer 2023, AJPS)

- Text as Outcome

总统的公开讲话如何影响新闻报道？ (Franco, Grimmer, and Lim 2019)

特朗普的推特 (Fong and Grimmer 2023)

- YouGov
- Survey equal # DEM, IND, REP: read Trump tweet + evaluate (Great, Good, OK, Bad, Terrible)

- Treatment: a supervised topic model (Indian buffet)
- Outcome: Aggregate, create scale [-200, 200]

- Train (66%), Test (33%), clustered by tweet

特朗普的推特 (Fong and Grimmer 2023)

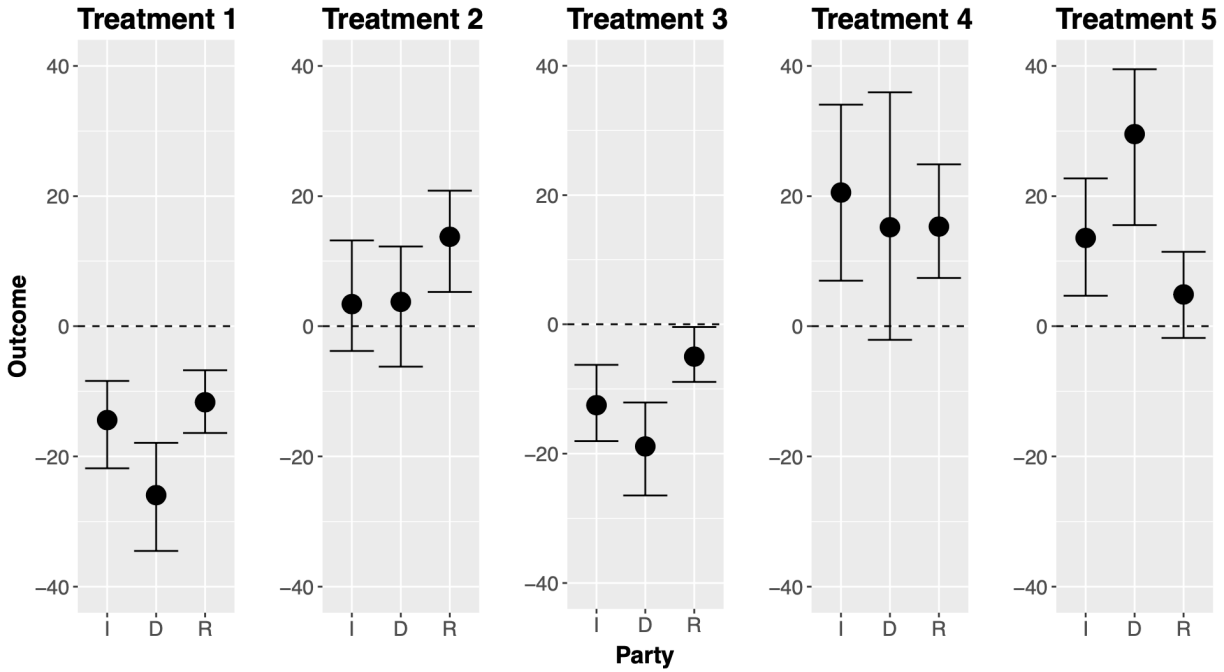


TABLE 5 Words Most Strongly Associated with Treatments

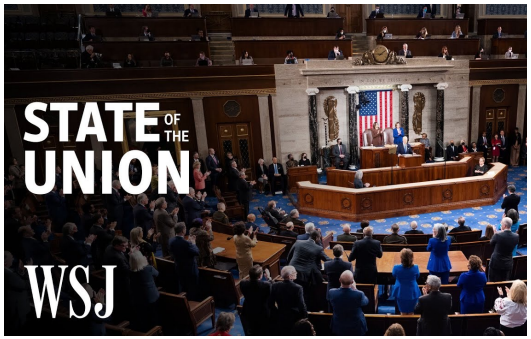
Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
fake	cuts	obamacare	flotus	prime
news	strange	senators	behalf	minister
media	tax	repeal	anthem	korea
cnn	luther	healthcare	melania	north
election	stock	replace	nfl	stock
story	market	republican	flag	market
nbc	alabama	vote	prayers	china
stories	reform	republicans	bless	executive
hillary	record	senate	ready	prayers
clinton	high	north	players	order

Note: Latent treatments were obtained from a supervised Indian Buffet Process. The listed words are the most characteristic of the latent treatment.

总统的公开讲话影响新闻报道

(Franco, Grimmer, and Lim 2019)

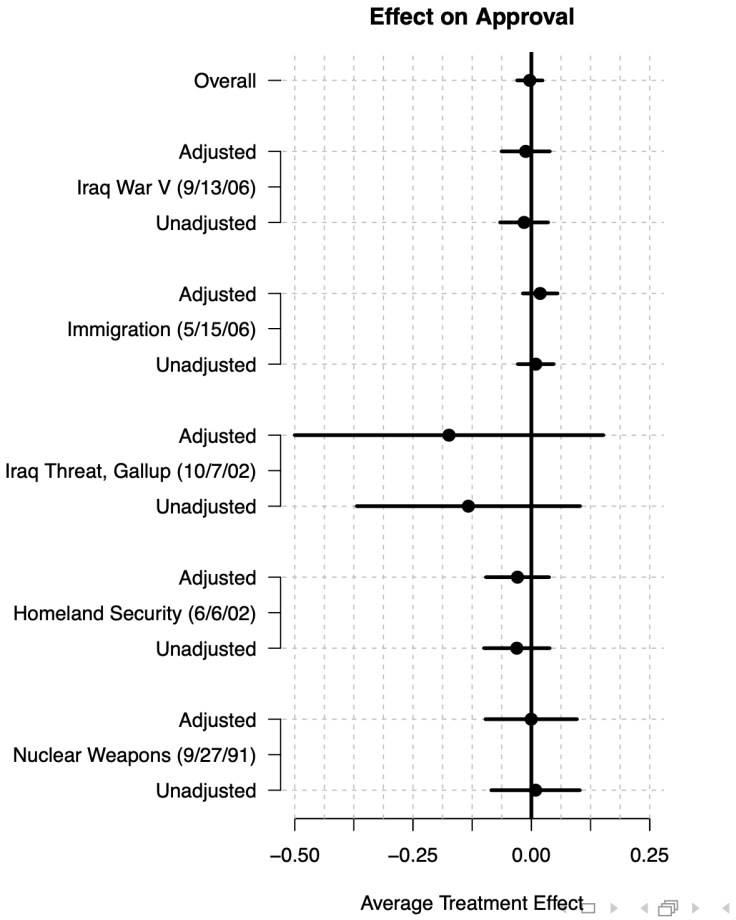
- Treatment: 搜集所有总统的公开演讲（going public）
- Outcome: 演讲前后的总统支持率以及**新闻报道**（text）
  - 10家热门报纸，在总统演讲前后两周是否提到president
- Train (10%), Test (90%)



总统的公开讲话影响新闻报道

(Franco, Grimmer, and Lim 2019)

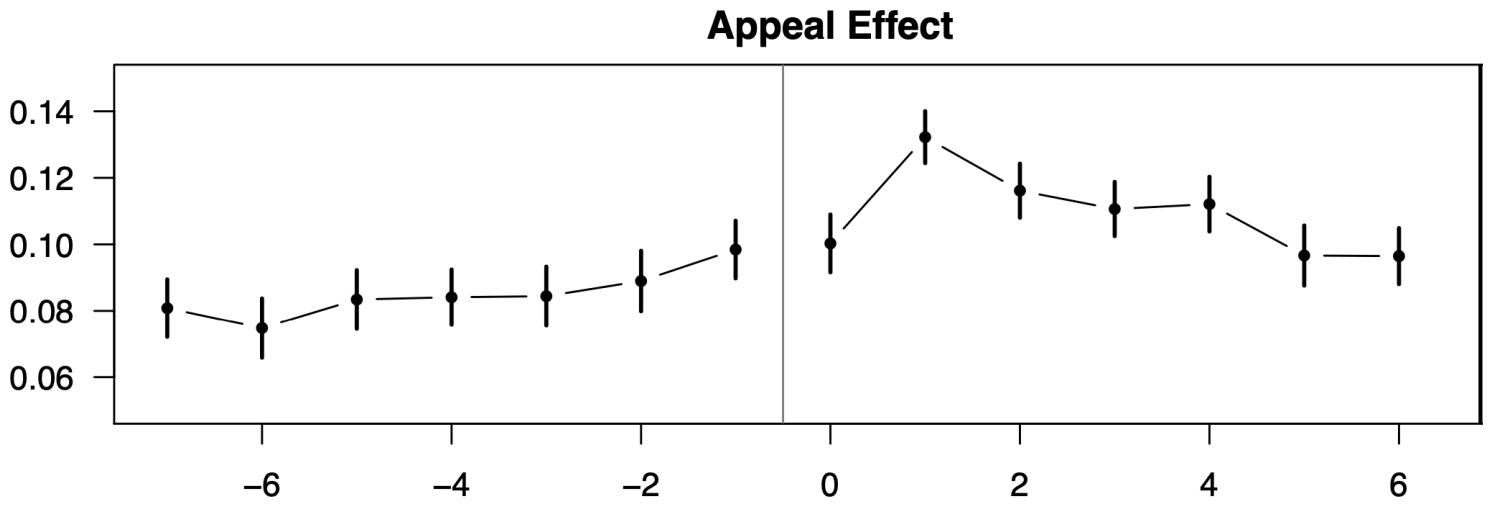
- 总统的公开演讲对其支持率的影响微乎其微



总统的公开讲话影响新闻报道

(Franco, Grimmer, and Lim 2019)

- 总统的公开演讲后对报纸报道有显著影响，尤其在演讲后1天



1. 自我介绍
2. 文本分析回顾
3. 文本与因果推断
4. **自由讨论**

自由讨论

谢谢！

fushu@sjtu.edu.cn

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
国际与公共事务学院 SCHOOL OF INTERNATIONAL AND PUBLIC AFFAIRS