Reading China:

Predicting Policy Change with Machine Learning

Julian TszKin Chan
(Bates White)

Weifeng Zhong
(AEI)

March 15, 2019
Boston University Pi-day Econometrics Conference

The views expressed here are solely those of our own and do not represent the views of the American Enterprise Institute, Bates White Economic Consulting, or their other employees.

Predicting policy change: why?

- China's industrialization: product of gov't direction.
- Opaque system make prediction difficult... until now.

Policy Change Index (PCI) for China:

- *leading* indicator of policy moves;
- quarterly, 1951 – present.

How to predict policy changes?

Build a machine learning algorithm to

- "read" the *People's Daily*;
- detect changes in how it prioritizes policy issues.

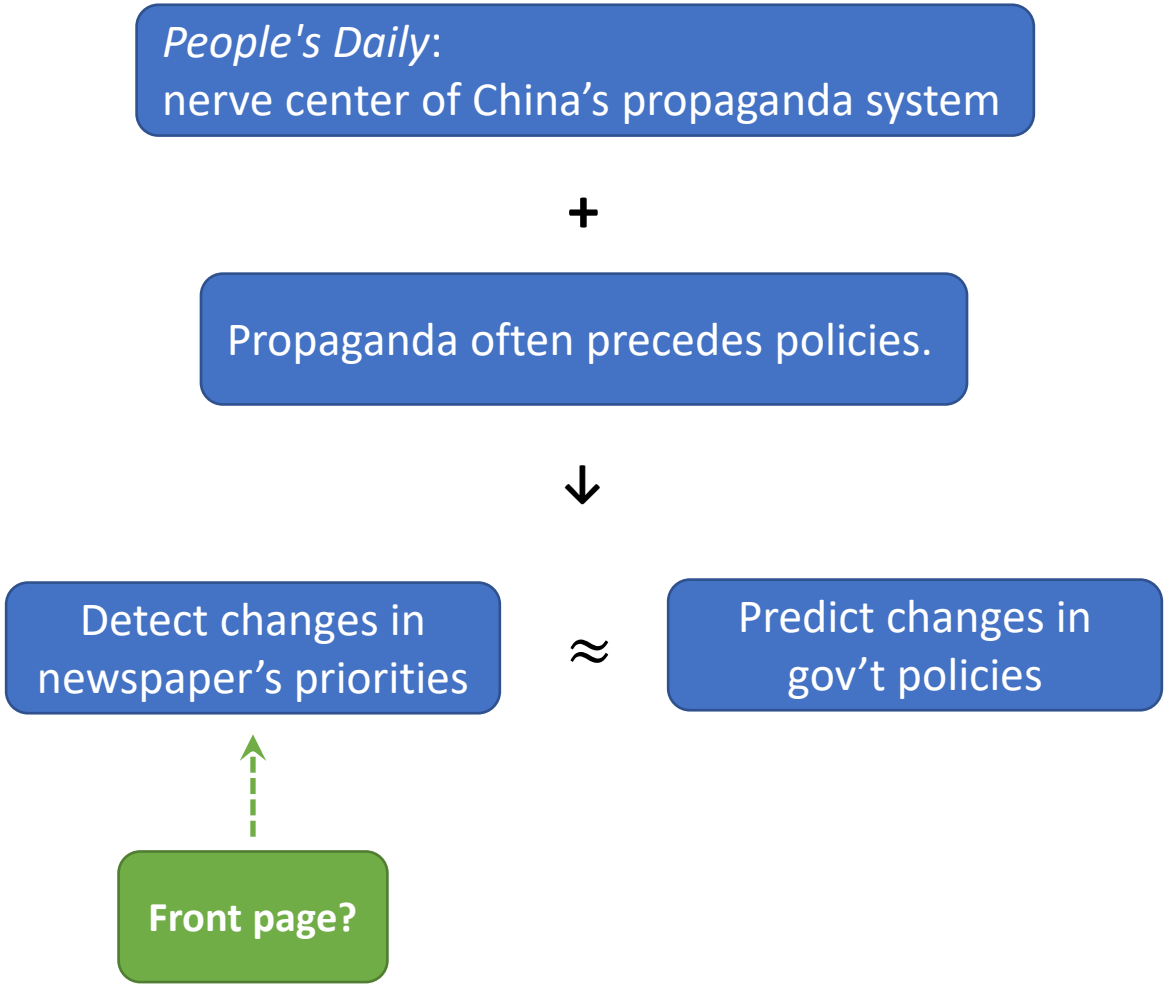
Official newspaper, 1946-present

Source of predictive power

The Leninist tradition:

- “[T]he whole task of the Communists is to be able to convince the backward elements.”
- Necessary “to transform the press... into a serious organ for the economic education of the mass of the population.”

Source of predictive power

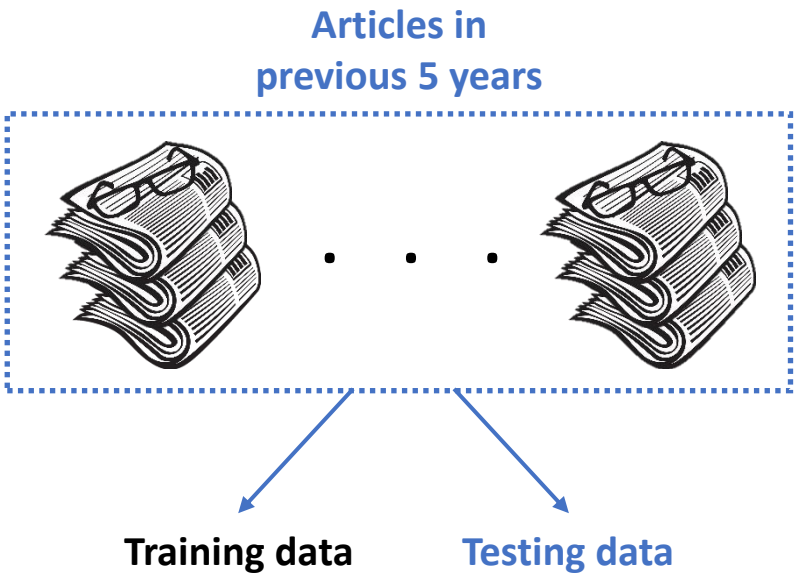


Method

Imagine an avid reader of the *People's Daily* who

1. reads recent articles (i.e., x);
2. forms a paradigm (i.e., $f(.)$) about what content "should" be on the front page (i.e., y);
3. tests the paradigm on new articles.

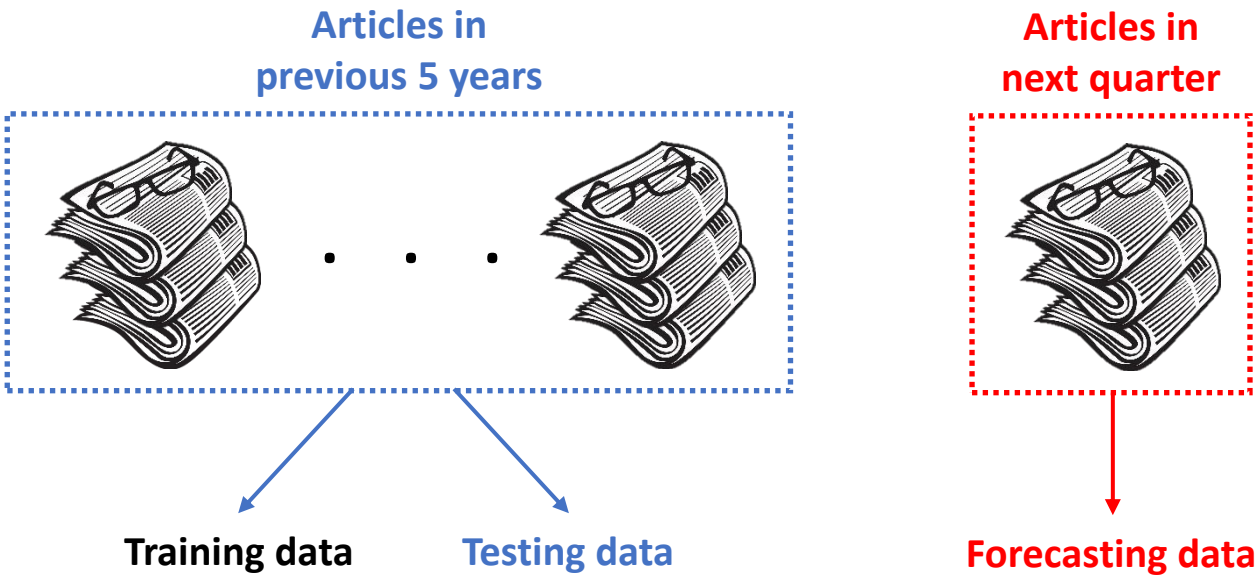
Model: building a front-page classifier



$$\hat{y}_{i_t,\mathrm{t}} = \hat{f}_T\left(x_{i_t,t}\right)$$

where $t = T - 20, \ldots, T - 1;\ i_t \in Training$

Model: building a front-page classifier

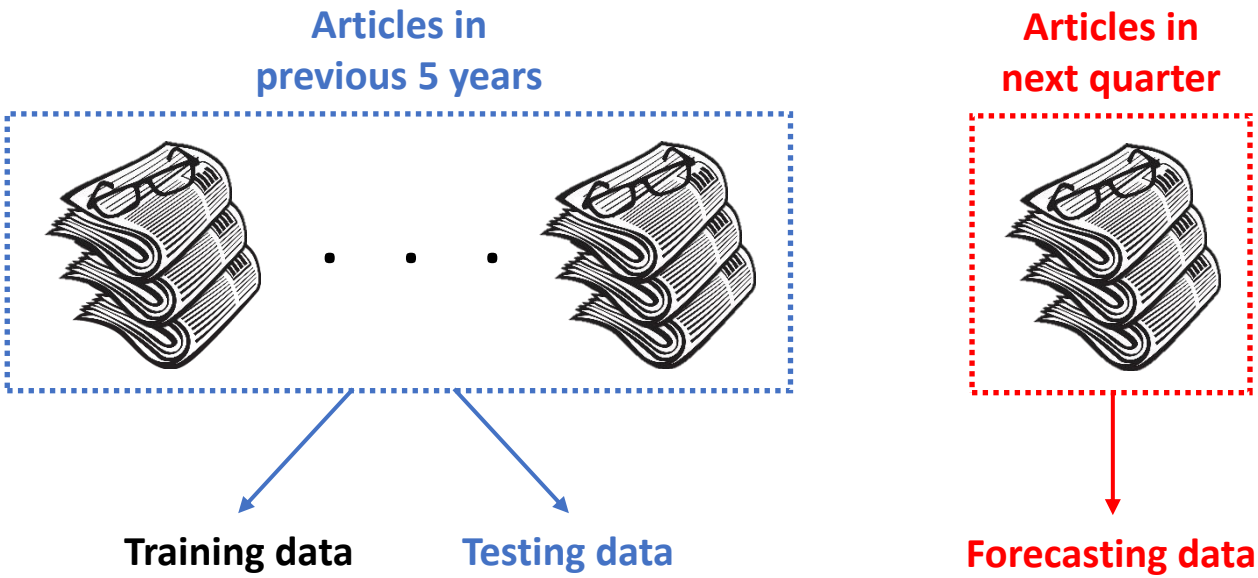


$$\hat{y}_{i_t,\mathrm{t}} = \hat{f}_T\left(x_{i_t,t}\right)$$

$$\hat{y}_{i_T,\mathrm{T}} = \hat{f}_T\left(x_{i_T,T}\right)$$

where $t = T-20, \ldots, T-1;\ i_t \in Training$

Model: building a front-page classifier



$$\hat{y}_{i_t,\mathrm{t}} = \hat{f}_T\left(x_{i_t,t}\right) \qquad \hat{y}_{i_T,\mathrm{T}} = \hat{f}_T\left(x_{i_T,T}\right)$$

where $t = T-20, \ldots, T-1;\ i_t \in Training$

Policy Change Index at period T = | **Test performance** − **"Forecast" performance** |

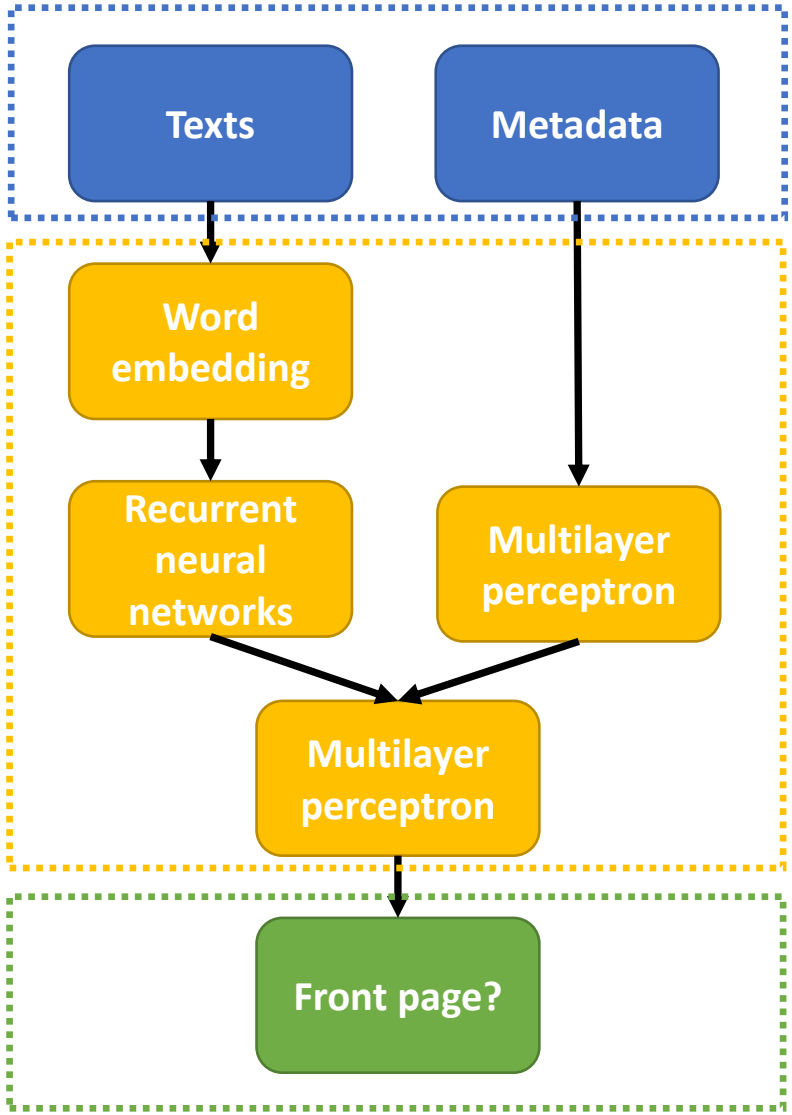
$$= \left| F1\left(\left(Y^{test}, \hat{f}_T(X^{test})\right)\right) - F1\left(\left(Y^{forecast}, \hat{f}_T\left(X^{forecast}\right)\right)\right) \right|$$

Data

```
sample_data
```

	date	year	month	day	page	title	body	id
0	2018-10-01	2018	10	1	1	习近平在会见四川航空"中国民航英雄机组"全体成员时强调　学习英雄事迹 弘扬英雄精神 将非凡英...	中共中央总书记、国家主席、中央军委主席习近平专门邀请四川航空"中国民航英雄机组"全体成员...	2018100000
1	2018-10-01	2018	10	1	1	烈士纪念日向人民英雄敬献花篮仪式在京隆重举行　习近平李克强栗战书汪洋王沪宁赵乐际韩正王岐山出席	9月30日上午，党和国家领导人习近平、李克强、栗战书、汪洋、王沪宁、赵乐际、韩正、王岐山...	2018100001
2	2018-10-01	2018	10	1	1	庆祝中华人民共和国成立69周年　国务院举行国庆招待会　习近平栗战书汪洋王沪宁赵乐际王岐山等...	9月30日晚，国务院在北京人民大会堂举行国庆招待会，热烈庆祝中华人民共和国成立六十九周年...	2018100002
3	2018-10-01	2018	10	1	2	习近平就印度尼西亚中苏拉威西省地震海啸向印尼总统佐科致慰问电	新华社北京9月30日电　9月30日，国家主席习近平就印度尼西亚中苏拉威西省发生强烈地震及...	2018100003
4	2018-10-01	2018	10	1	2	在庆祝中华人民共和国成立六十九周年招待会上的致辞　中华人民共和国国务院总理　李克强　（二〇...	各位来宾、各位朋友、同志们：　今天，我们隆重庆祝中华人民共和国成立六十九周年。新中国波澜壮...	2018100004
5	2018-10-01	2018	10	1	2	用奋斗成就复兴伟业(社论)　——热烈庆祝中华人民共和国成立69周年	时间的年轮，刻印下奋斗者的足迹。当10月的阳光照耀大地，我们迎来了人民共和国69岁华诞。...	2018100005
6	2018-10-01	2018	10	1	2	国务院印发《决定》　进一步压减工业产品生产许可证管理目录和简化审批程序	新华社北京9月30日电　经李克强总理签批，国务院日前印发《关于进一步压减工业产品生产许可...	2018100006
7	2018-10-01	2018	10	1	2	谱写新时代乡村全面振兴新篇章　——论学习习近平总书记关于实施乡村振兴战略重要讲话精神	本报评论员　乡村振兴既是一场攻坚战，更是一场持久战，必须坚定信心，脚踏实地，苦干实干	2018100007

Model



x : each article as an observation.

f : map article to whether it is on front page

$$y = f(x)$$

State of the art

BERT (Devlin, et al. 2018)

Machine learning algorithm is performing as good as human (88%) on language tests, such as:

On stage, a woman takes a seat at the piano. She

a) *sits on a bench as her sister plays with the doll.*

b) *smiles with someone as the music plays.*

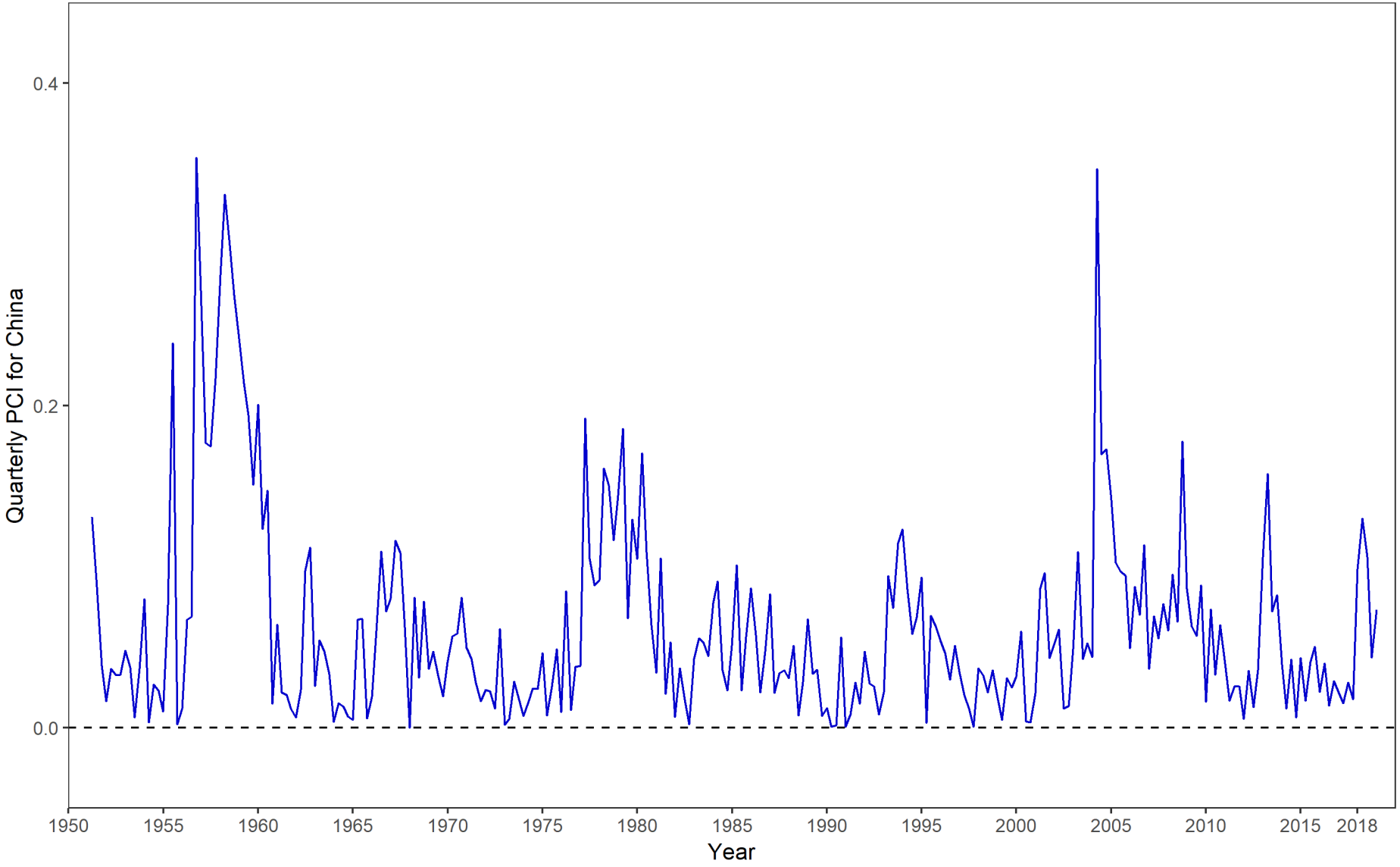
c) *is in the crowd, watching the dancers.*

d*) nervously sets her fingers on the keys.*

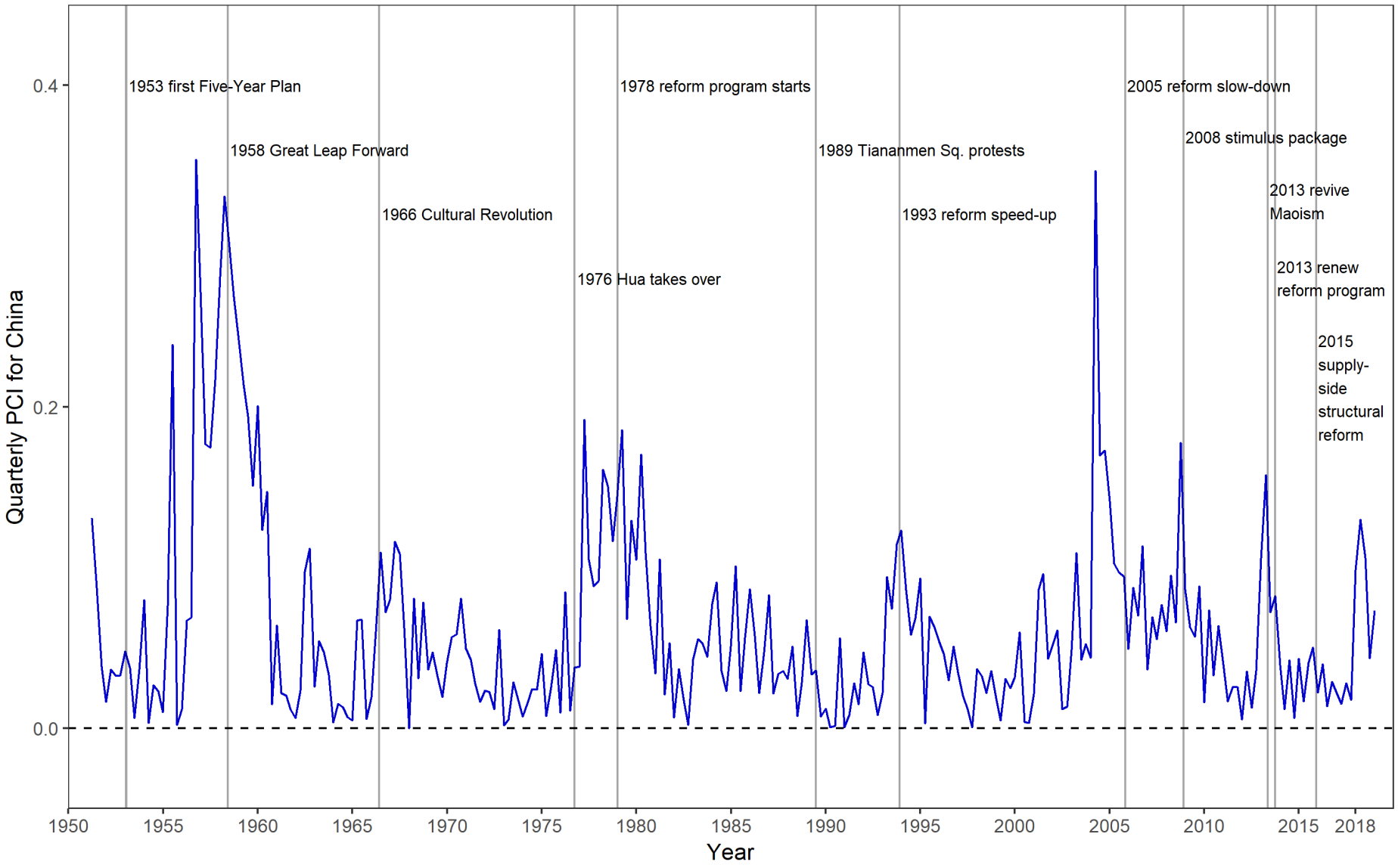
P.S. The algorithm is not trained to perform those tests.

Results

Result: PCI


0.4
0.2
0.0
Quarterly PCI for China
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2018
Year

Result: PCI — with ground truth

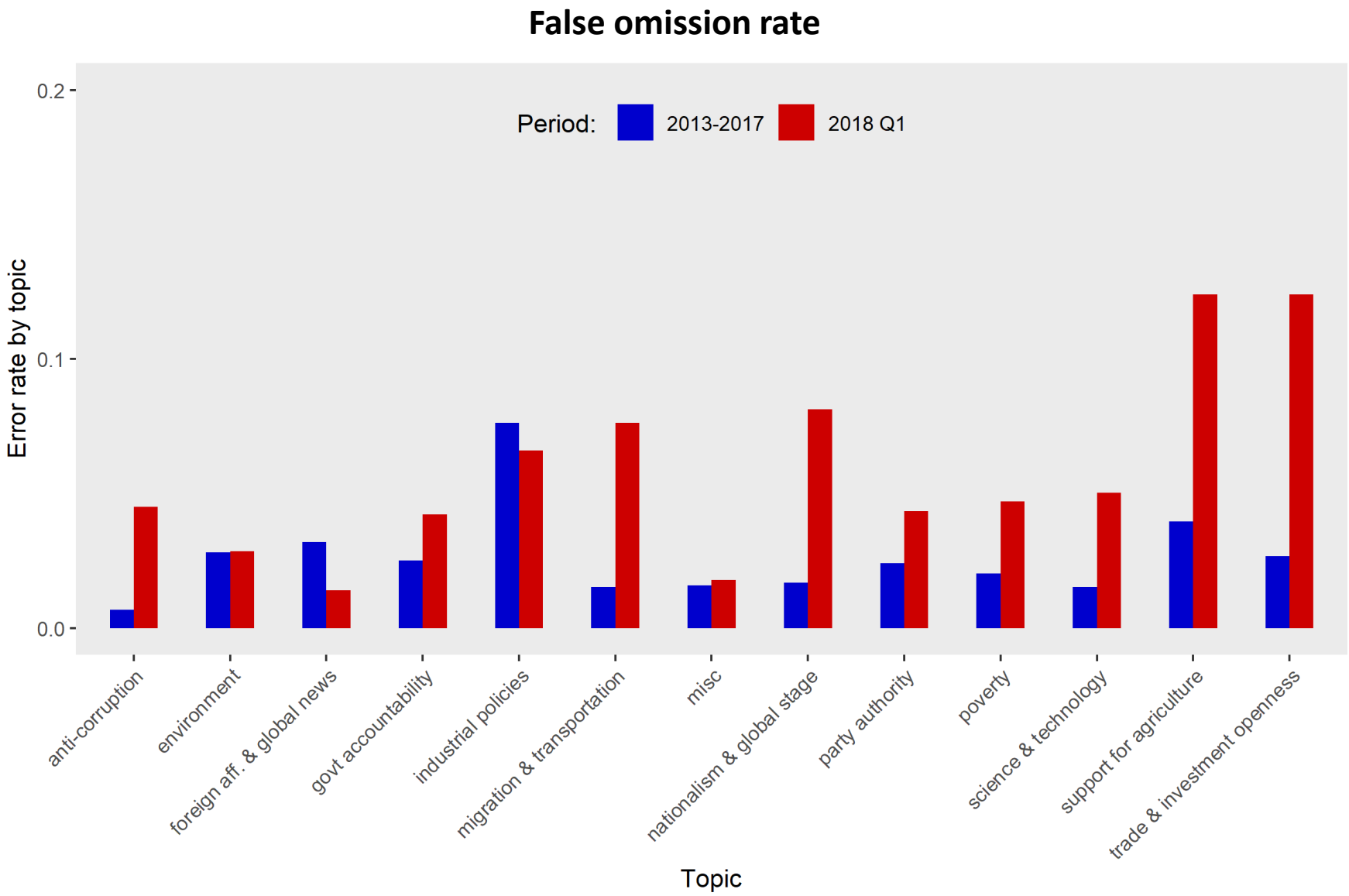


Understanding substance of change

		Classified on front page?	
		No	Yes
Front page?	No	√	false positives
	Yes	false negatives	√

- Content of *mis*-classified articles has policy substance.
  - False positive: new policies
  - False negative: policies that are phasing out

The 2018 Q1 uptick



Discussion

Supervised learning

$$mapping : X \rightarrow Y$$

- Trained on $\{x_i, y_i\}_{i \in training}$.
- Goal: from $\{x_j\}_{j \in new}$, to predict $\{y_j\}_{j \in new}$.

- Challenge: need lots of training data.

Understanding policy priority:

an *in*feasible approach

$$g : \{(Article, FrontPage)\} \rightarrow \{(Policy, Priority)\}$$

- With the learned function g:
  - $g(\text{"pvt sector is important"}, \text{front page}) = (\text{reform}, \text{high priority})$;
  - $g(\text{"central planning is great"}, \text{front page}) = (\text{reform}, \text{low priority})$; ...

- But where are the training data?

Understanding policy priority:

a feasible approach

- Think of policy priorities as a latent variable:

$$f_{\{(Policy,Priority)\}} : \{Article\} \rightarrow \{FrontPage\}$$

- Lots of training data to learn each function f.
- Difference in function $\Rightarrow$ difference in priorities.
- “Language-free!”

Discussion

- Adversarial attack
  - If the Chinese government knows that we can detect their policy change based on the newspaper, would they change their behavior to avoid detection?
    - That's the purpose of propaganda.
  - What if the Chinese government knew we are reading the newspaper and want to fool us?
    - Human judgement
- Readership is dropping overtime.
  - Government officials are required to read the People's Daily.

Other applications

Other PCI projects

- Text summarization and highlighting — what words/sentences cause misclassification?
- Regional and local PCIs for China, their development implications, etc. (joint w/ W. Cheung).
- PCIs for other (ex-)Communist regimes' policies:
  - Soviet Union's *Pravda* and East Germany's *Neues Deutschland* (joint with w/ E. Melly)
  - North Korea's *Rodong Sinmun* (collecting data)

“Opinionated News?” (joint w/ S. Slavov)

- A wide discrepancy found in 2018:
  - 42% of Americans think the news they see is just commentary and opinion, and
  - only 5% of Americans think that’s useful.
- Q: Is that true? How to detect opinionated news?
- Data: The New York Times, 1987-2007.
  - PD articles ⇢ NYT articles;
  - front-page indicator ⇢ opinion indicator;

Interested in DIY?

- Website: policychangeindex.com (newsletter sign-up)
- Paper: policychangeindex.com/pdf/Reading_China.pdf
- Source code: github.com/PSLmodels/PCI

- A simulated example to show how the PCI works.

References

- Word embeddings
  - Word2sec (Mikolov, et al., 2013)
  - GloVe (Pennington, et al., 2014)
  - ELMo (Peters, et al., 2018)
  - BERT (Devlin, et al. 2018)

- GRU: Cho et al. (2014)

- LSTM: Hochreiter and Schmidhuber (1997)

- Hierarchy model and document classification (Tang et al. 2015, Yang et al., 2016)