Reading China:

Predicting Policy Change with Machine Learning

Julian TszKin Chan (Bates White Economic Consulting)

Weifeng Zhong (Mercatus Center at George Mason University)

Federal Reserve Board (9/18/2019)

The views expressed here are the authors' and do not represent the views of Bates White, the Mercatus Center, or their other employees.

What's product of this research?

Policy Change Index (PCI) is a machine learning program that predicts China's policy moves by "reading" this:

人民日報
RENMIN RIBAO

习近平应约同美国总统特朗普通电话

习近平同志主持会议并作重要讲话

党的十九届一中全会产生中央领导机构

习近平任中共中央总书记中央军委主席

习近平李克强栗战书汪洋王沪宁赵乐际韩正任中央政治局常委

State-run media tell us a lot — about the state!

- Example: *People's Daily* ≈ the *Pravda* of China

- The primary mission is not to report news, but to propagate official viewpoints and signal policies — often *in advance*
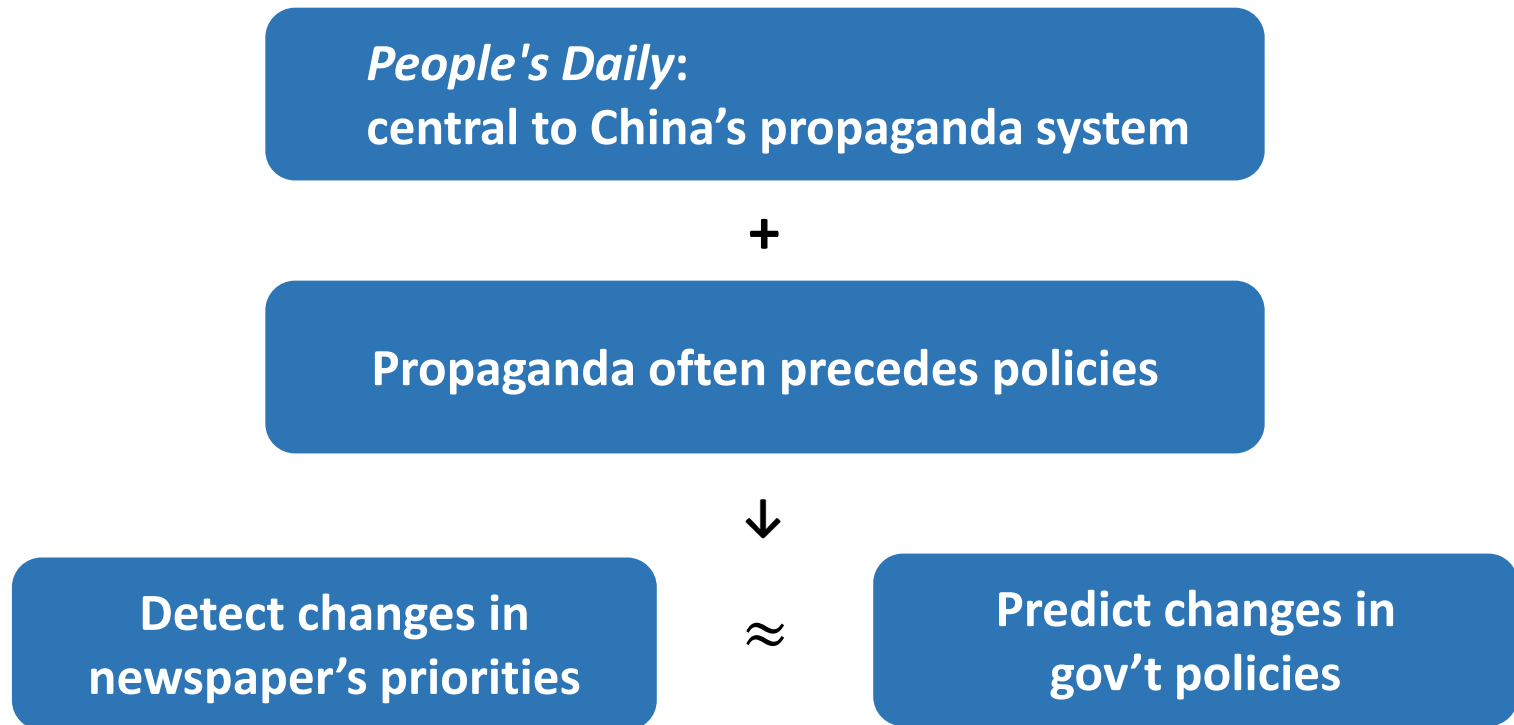
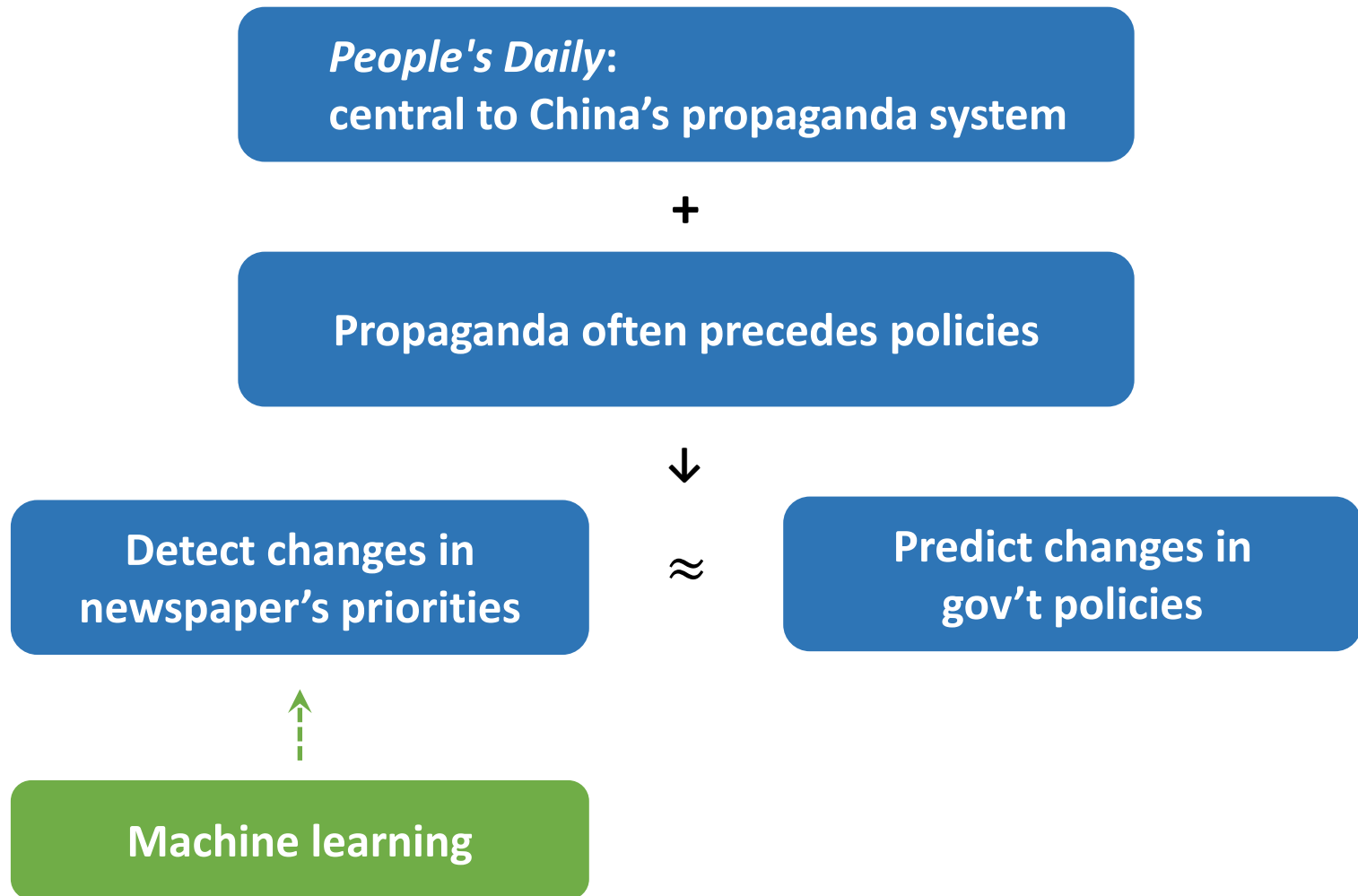# State-run media tell us a lot — about the state!

State-run media tell us a lot — about the state!

- The Soviet tradition:
  - "[T]he whole task of the Communists is to be able to *convince* the backward elements." (Lenin, 1920)
  - It is a fundamental necessity "to transform the press… into a serious organ for the *economic education* of the mass of the population." (Lenin, in early USSR years)

State-run media tell us a lot — about the state!

People's Daily:
central to China's propaganda system

+

Propaganda often precedes policies

↓

Detect changes in
newspaper's priorities

≈

Predict changes in
gov't policies

State-run media tell us a lot — about the state!

People's Daily:
central to China's propaganda system

+

Propaganda often precedes policies

↓

Detect changes in newspaper's priorities   ≈   **Predict changes in gov't policies**

↑

Machine learning

Policy Change Index (PCI) project

- Data: *People's Daily* text, 1946-present (2 million articles).

- Goal: Develop a machine learning program to detect changes in the newspaper's priorities.

- Output: the **Policy Change Index (PCI)** for China.

Where's the added value?

- A tech'ed-up version of "Kremlinology."

- Deciphering propaganda content has been complicated due to
    - Opaque system;
    - Manipulated data;
    - Difficulty in validation;…

- Machine learning — with some twists — can help!

Data structure

sample_df

id	date	page	title	body
20181202525	2018-12-31 00:00:00	1.0	今年粮食总产量13158亿斤　农业农村发展取得新成绩，乡村振兴开局良好	本报北京12月30日电　（记者高云才、郁静娴）记者从30日召开的全国农业农村厅局长会议获…
20181202526	2018-12-31 00:00:00	1.0	新时代改革再出发的重要里程碑　——写在党的十八届三中全会召开五周年之际	本报评论员　改革开放40年来，我们党带领人民绘就了一幅波澜壮阔、气势恢宏的历史画卷。其…
20181202527	2018-12-31 00:00:00	1.0	改革引领中国经济新航向（在习近平新时代中国特色社会主义思想指引下——新时代新作为新篇章）　…	本报记者　许志峰　吴秋余　王　珂　林丽鹂　巨轮向前，离不开航标指引方向。　　1978年…
20181202528	2018-12-31 00:00:00	1.0	产量连续六年超1200亿斤，实现"十五连丰"　黑龙江 粮食再丰收，粮仓更稳固	本报哈尔滨12月30日电　（记者方圆）此刻的黑土地，正在冬眠蓄力。今年黑龙江粮食又是一…
20181202529	2018-12-31 00:00:00	1.0	伟大的变革——庆祝改革开放40周年大型展览　现场观众达223万人次　展览将持续到2019年…	据新华社北京12月30日电　"伟大的变革——庆祝改革开放40周年大型展览"自11月13日…
20181202530	2018-12-31 00:00:00	1.0	航线不停、景区不休、生意不闲、游客不断　新疆 旅游无淡季，今冬更红火	本报乌鲁木齐12月30日电　（记者杨明方、阿尔达克）一年之中半年是降雪期，冬季雪深超过3…
20181202531	2018-12-31 00:00:00	1.0	图片报道	徜徉知识海洋，感受阅读乐趣。十二月三十日是元旦小长假的第一天，全国各地的特色文化活动丰富…
20181202532	2018-12-31 00:00:00	2.0	创造新的更大奇迹　——党的十八届三中全会召开五周年述评	"全面发力、多点突破、蹄疾步稳、纵深推进"——庆祝改革开放40周年大会上，习近平总书记用…
20181202534	2018-12-31 00:00:00	3.0	开启中国特色大国外交新征程（2018年度国际特别报道）　——国务委员兼外交部长王毅回顾20…	本报记者　赵嘉鸣　管克江　杜尚泽　焦　翔　王海林　国务委员兼外交部长王毅12月29日接…
20181202535	2018-12-31 00:00:00	3.0	二〇一七年岁末，莫斯科大学的中国留学生们欣喜地收到了习近平主席的回信；一年来，牢记习主席的…	本报驻俄罗斯记者　吴　焰　"一年前，习近平主席给留学生回信，鼓励我们胸怀大志，刻苦学习…
20181202536	2018-12-31 00:00:00	3.0	外交部发言人就中美建交40周年发表谈话	本报北京12月30日电　外交部发言人陆慷30日就中美建交40周年发表谈话。　　陆慷说，2…

Constructing the PCI-China

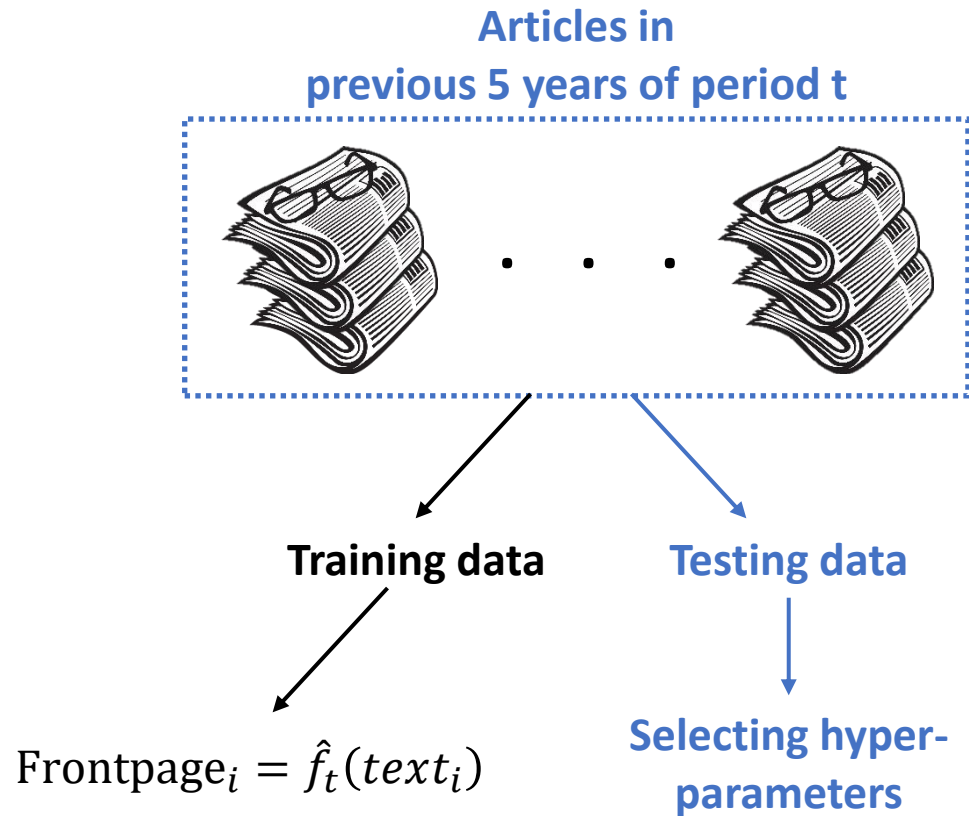# Start with something *seemingly* trivial

Build an algorithm that classifies **front-page** articles

("reverse-engineering" the editor's mind).
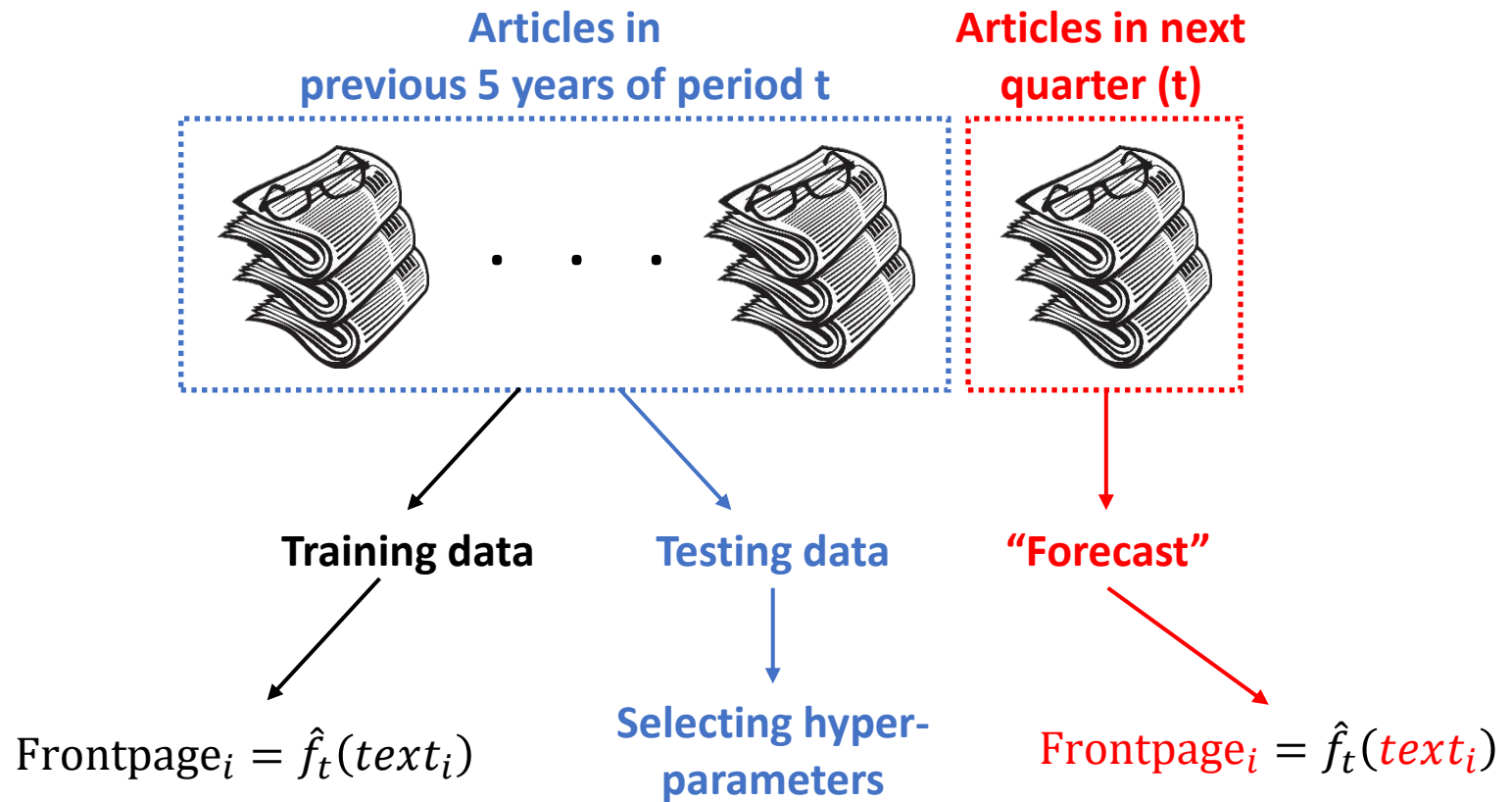
Start with something *seemingly* trivial

Build an algorithm that classifies **front-page** articles
("reverse-engineering" the editor's mind).
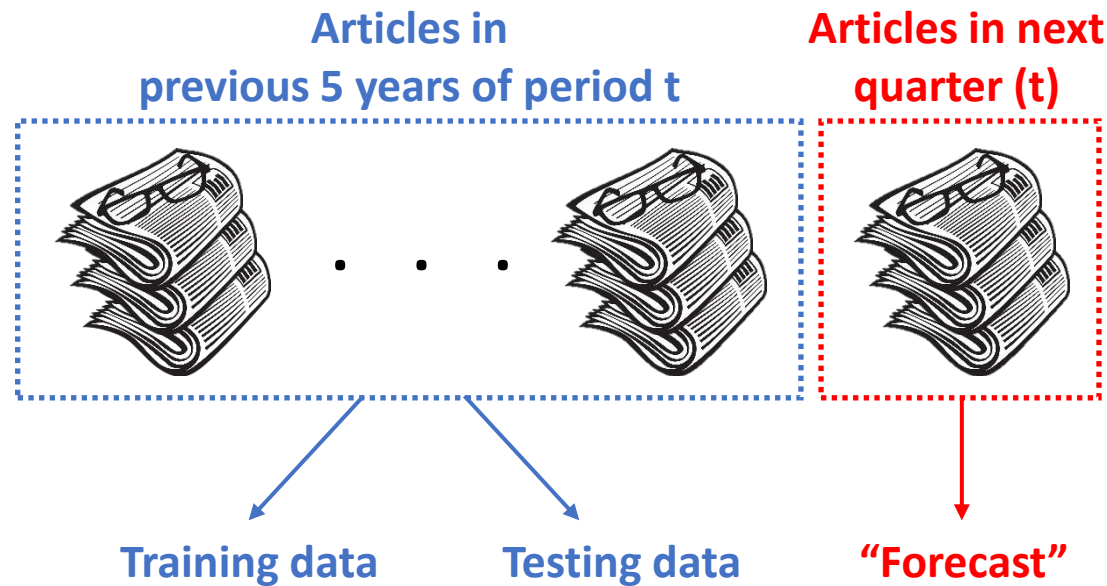
But what's the point?

The framework

**Articles in
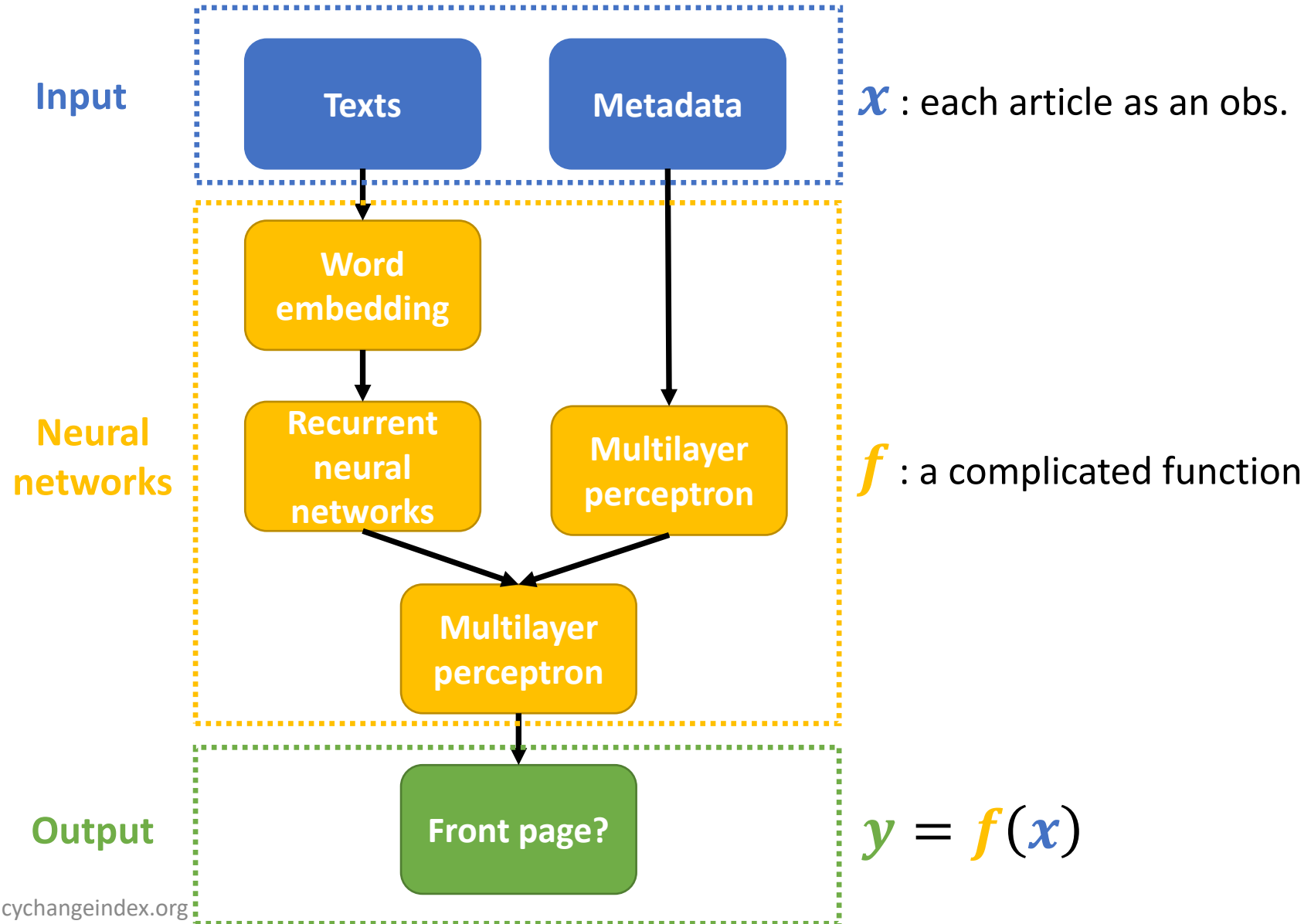previous 5 years of period t**

Training data   **Testing data**

$$\text{Frontpage}_i = \hat{f}_t(text_i)$$

**Selecting hyper-
parameters**

The framework

**Articles in
previous 5 years of period t**

**Articles in next
quarter (t)**

Training data

Testing data

"Forecast"

$$\text{Frontpage}_i = \hat{f}_t(text_i)$$

**Selecting hyper-
parameters**

$$\text{Frontpage}_i = \hat{f}_t(text_i)$$

The framework

Articles in previous 5 years of period t

Articles in next quarter (t)
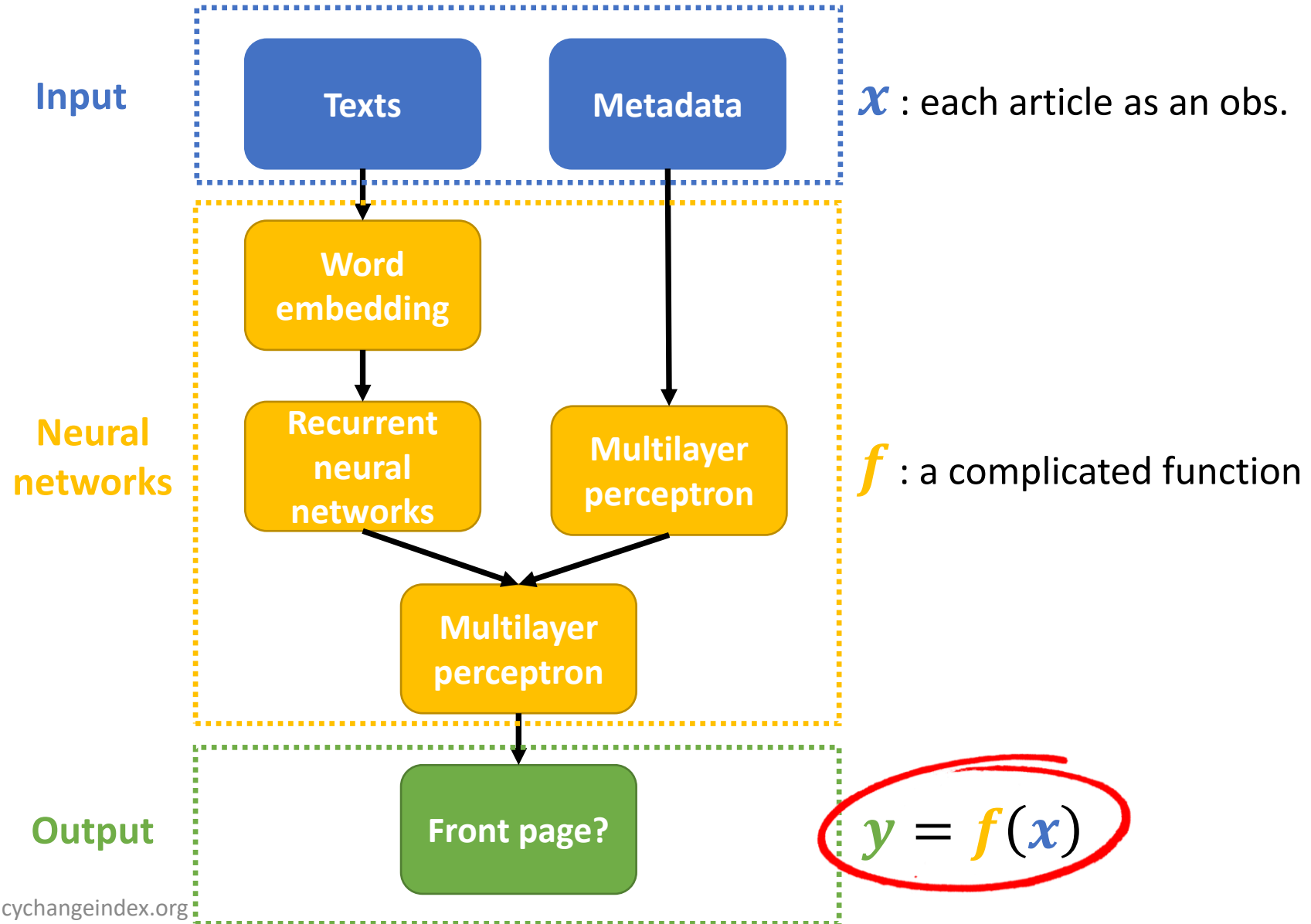

Training data **Testing data** **"Forecast"**

Policy Change Index of period _t_ $= \left| \text{Test performance} - \text{"Forecast" performance} \right|$

$\text{PCI} \gg 0 \Rightarrow$ Structural difference in the newspaper's priorities
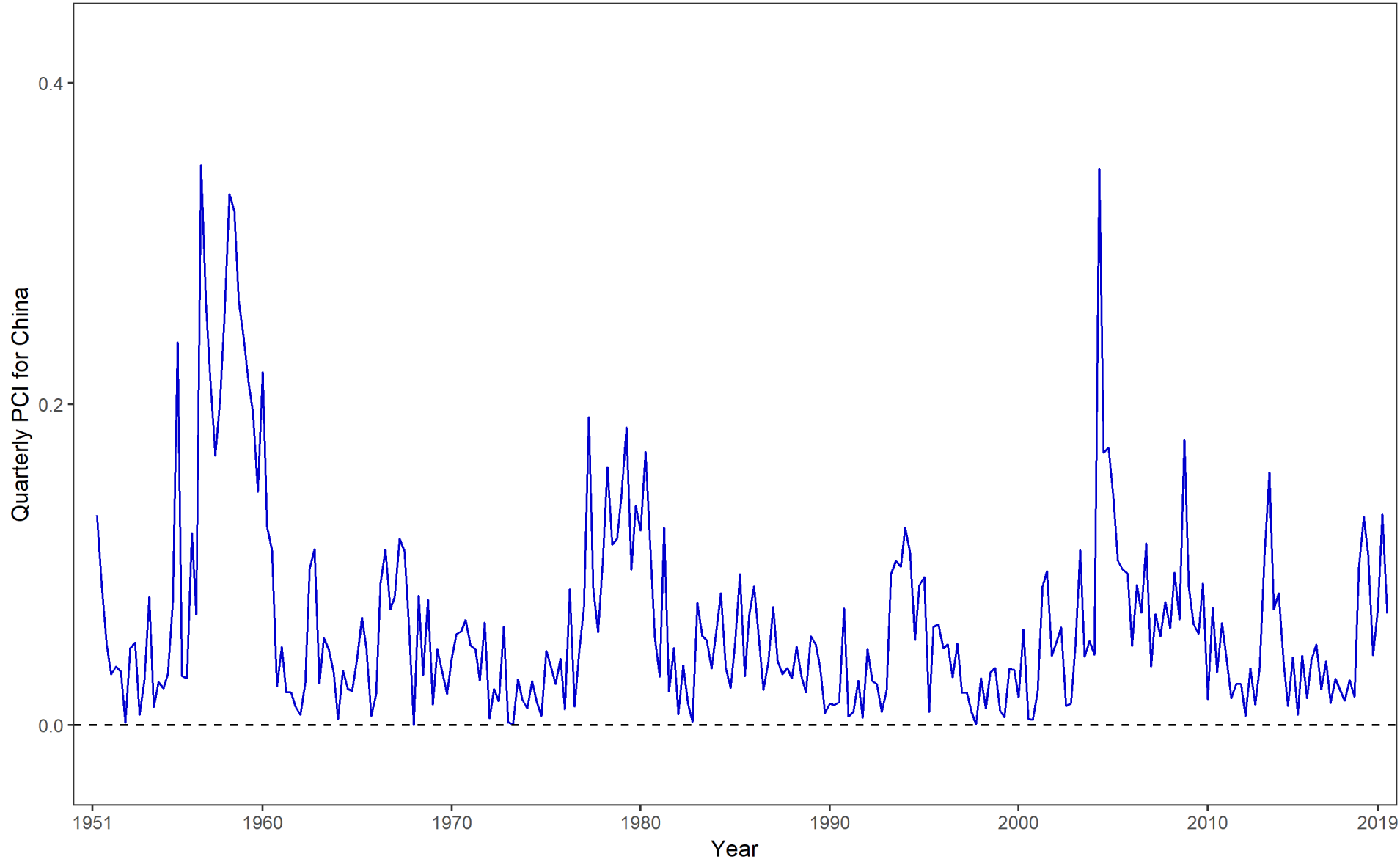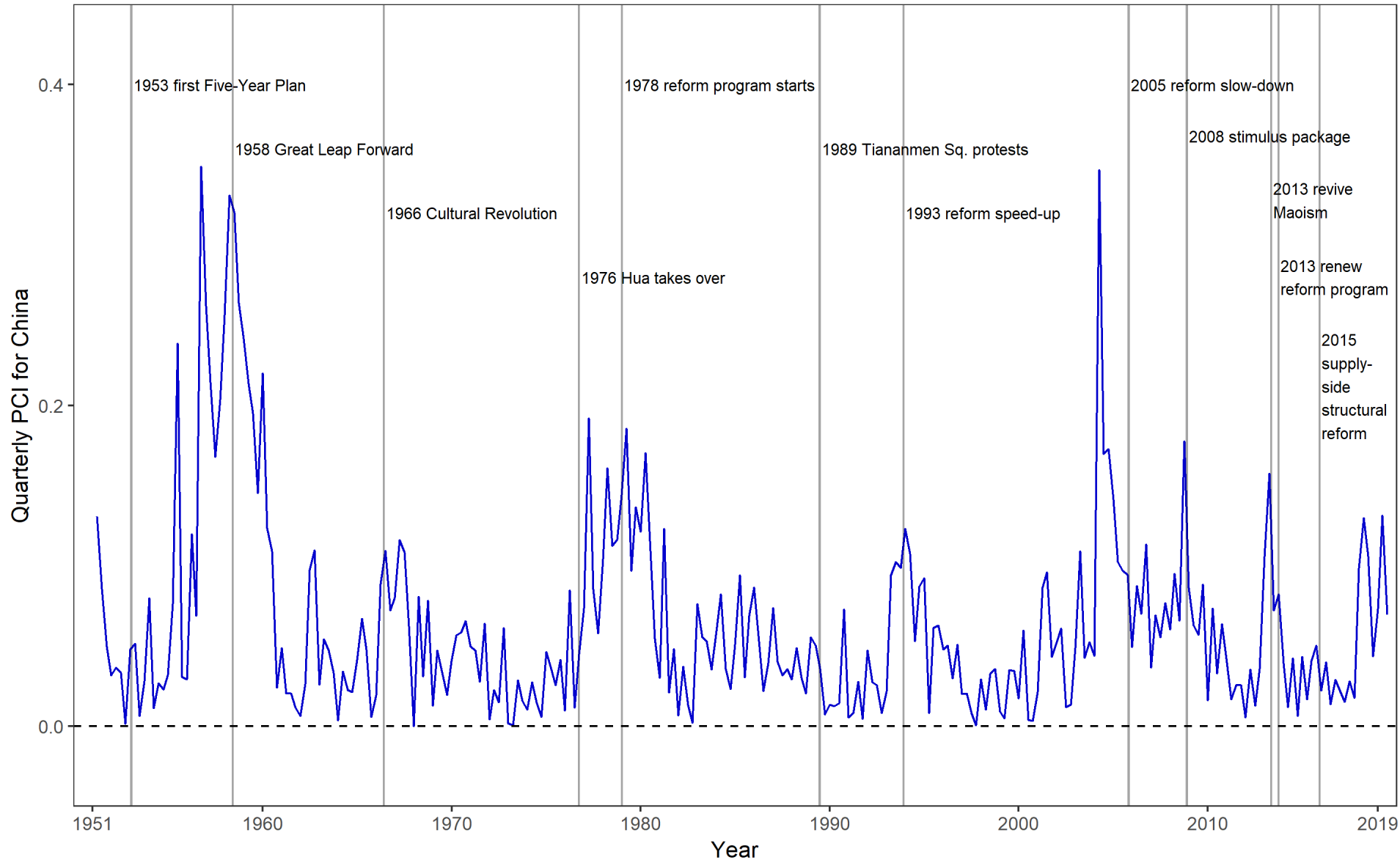
Modeling

Input

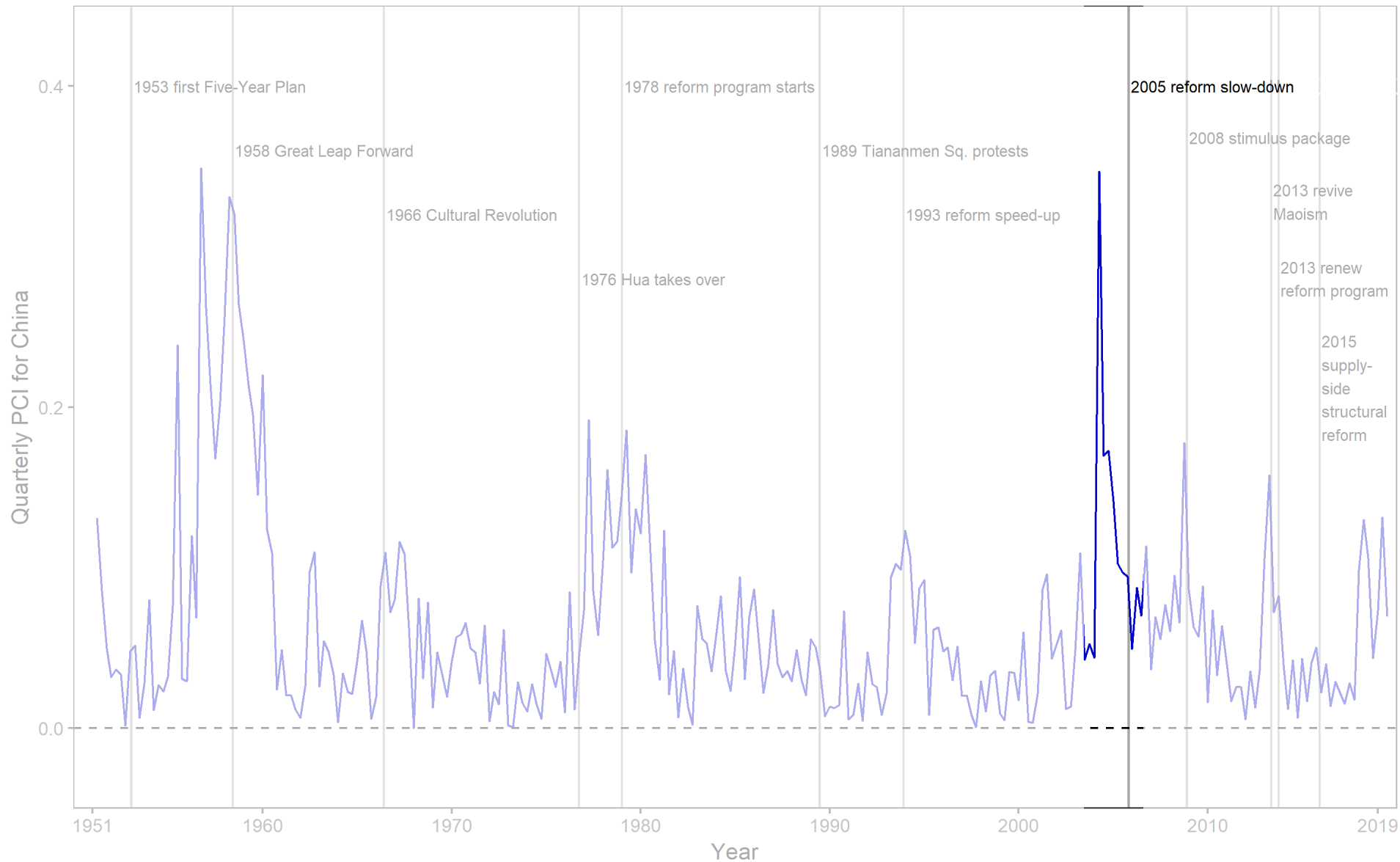Texts   Metadata

x : each article as an obs.

Neural networks

Word embedding

Recurrent neural networks   Multilayer perceptron

Multilayer perceptron

f : a complicated function

Output

Front page?

$y = f(x)$

policychangeindex.org

Modeling

Input
Texts   Metadata

x : each article as an obs.

Neural networks

Word embedding

Recurrent neural networks   Multilayer perceptron

Multilayer perceptron

f : a complicated function

Output

Front page?

$y = f(x)$

Results

PCI-China—detecting policy changes

PCI-China—with ground truth

PCI-China—with ground truth

Quarterly PCI for China (y-axis, values 0.0, 0.2, 0.4)

Year (x-axis, 1951, 1960, 1970, 1980, 1990, 2000, 2010, 2019)

1953 first Five-Year Plan

1958 Great Leap Forward

1966 Cultural Revolution

1976 Hua takes over

1978 reform program starts

1989 Tiananmen Sq. protests

1993 reform speed-up

2005 reform slow-down

2008 stimulus package

2013 revive Maoism

2013 renew reform program

2015 supply-side structural reform

PCI for China — going forward
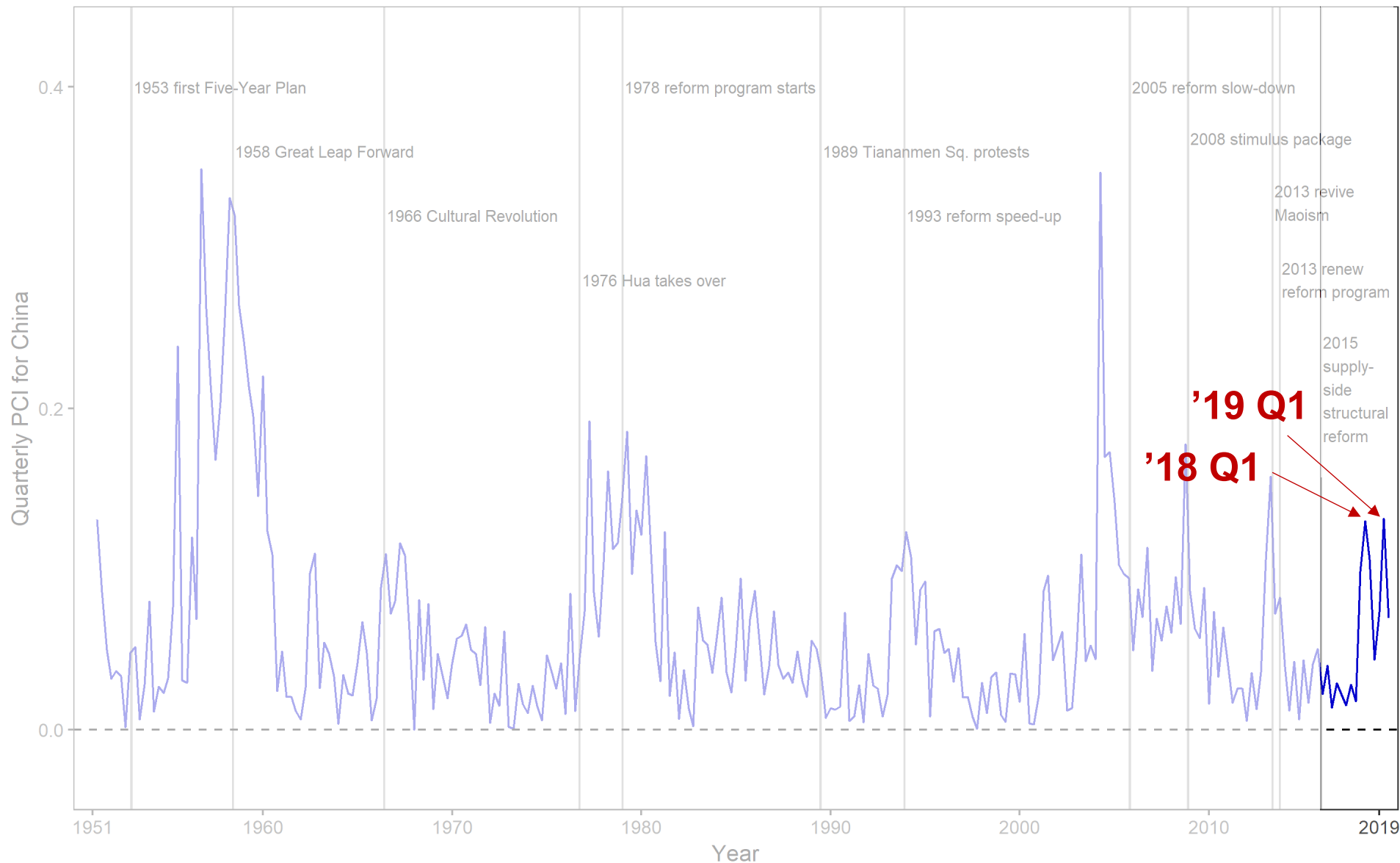
Understanding changes

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

Content of *misclassified* articles has policy substance

- False negative: new policy

- False positive: phase-out policy

Understanding changes — 2018 Q1 uptick
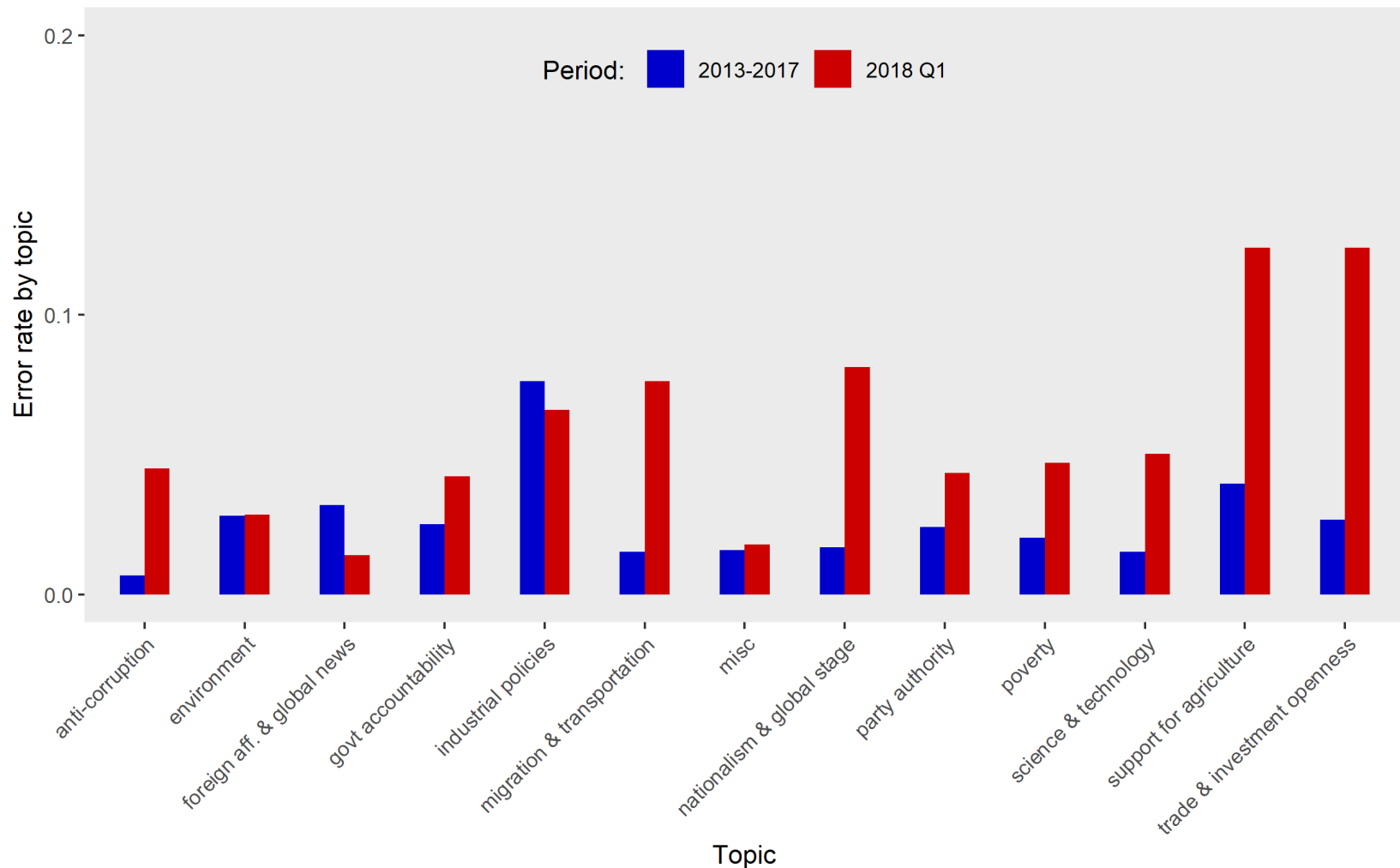
Understanding changes — 2018 Q1 uptick

- 2018 Q1 spike represents:
  - internally: strengthening party authority;
  - externally: nationalism and global leadership;
  - populist policies to boost political support.

- 2019 Q1 uptick: similar patterns.

- Our consistent prediction since January: curb your enthusiasm.

* 2019 Q3 update coming soon!

Discussion

Understanding policy priority:
an *in*feasible approach

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

- With the learned function g:

    - $g(\text{"pvt sector is important"}, \text{front page}) = (\text{reform, high priority});$

    - $g(\text{"central planning is great"}, \text{front page}) = (\text{reform, low priority});$
      ...

- 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

Potential applications

- PCIs for other countries – North Korea, Cuba, …

- PCI-China sub-indices:

    - By issue – trade policy sub-index, fiscal policy sub-index, …

    - By country – China-US relations, China-EU relations, …

- Some countries may not control the media

    - But politicians/policymakers still talk!

PCI-Crackdown

Policy Change Index (PCI) for Crackdown

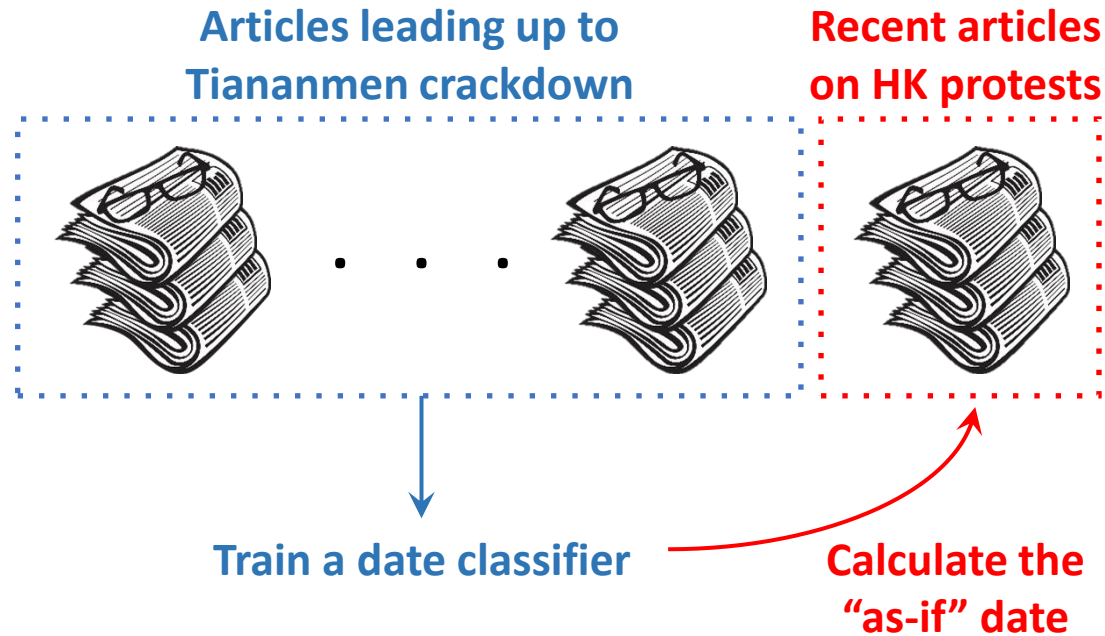**Protest-related People's Daily articles** → **PCI algorithm** → **Predictor of potential crackdown**

(400+ articles on Tiananmen 1989 protest & Hong Kong 2019 protests)
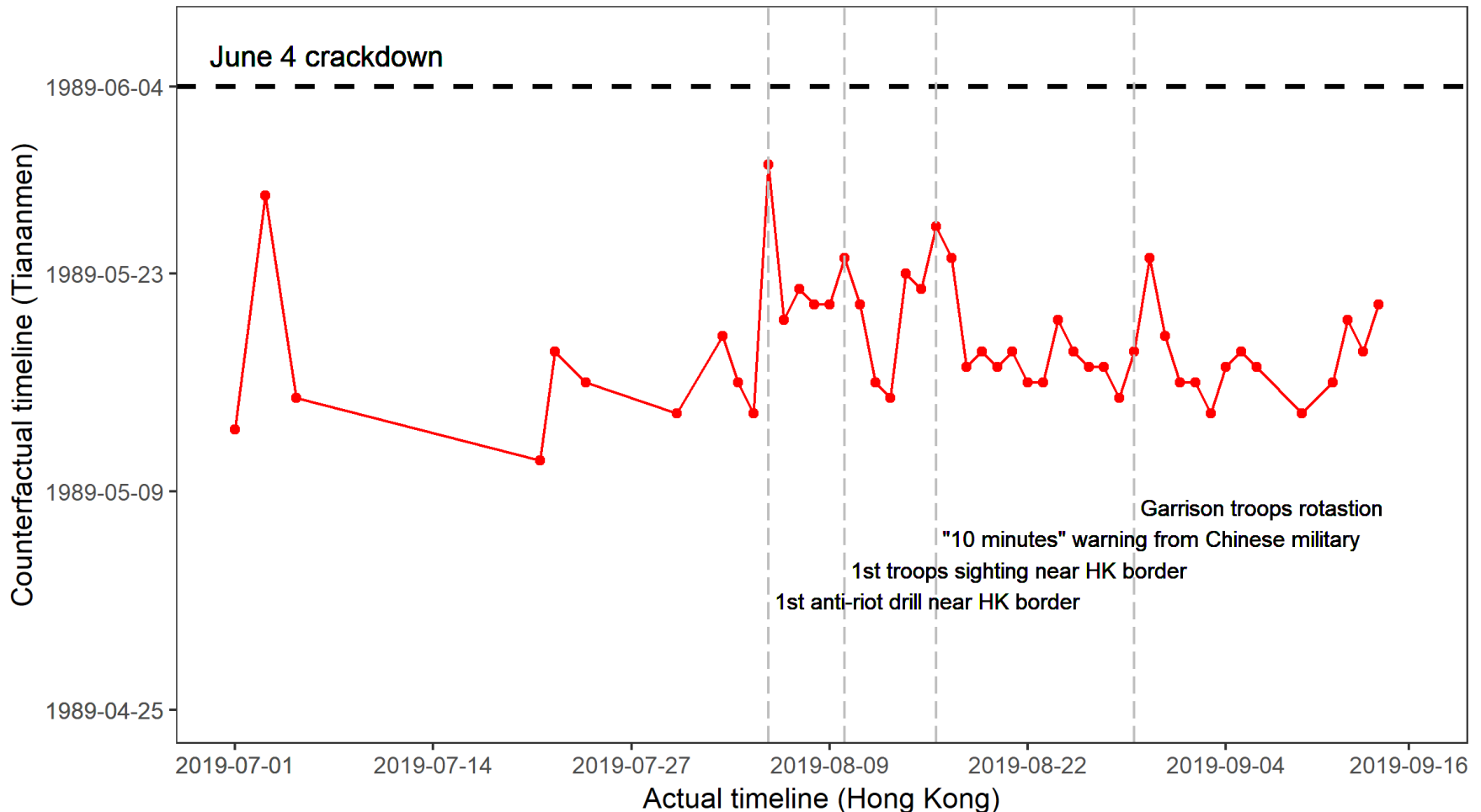
(daily index update)

PCI-Crackdown algorithm

**Articles leading up to
Tiananmen crackdown**

**Recent articles
on HK protests**

. . .

Train a date classifier

**Calculate the
"as-if" date**

PCI-Crackdown : mapping current dates to Tiananmen timeline

PCI-Crackdown: 2019 Hong Kong protests
(Daily index, Jul 01 – Sep 15, 2019)

Interested in DIY?

- Website: policychangeindex.org

- Newsletter: mercatus.org/PCInewsletter

- Repository: github.com/PSLmodels/PCI

- A simulated example to show how the PCI works.

Questions?

State of the art model

- BERT (Devlin, et al. 2018)

- Machine learning algorithm is performing as good as human (88% accuracy) 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.*

- the algorithm is not trained to perform those tests.