Learning Structural Changes from Text Data

Weifeng Zhong, Ph.D.

American Enterprise Institute

February 5, 2019

IMPAQ International

Text as data

- Voluminous in the digital era.
- But unstructured and sequential.

- How to detect structural changes from text?

- Proposal: machine learning — *with a twis*t.

Two papers

1. “Reading China” (w/ J. Chan).
    - Predicts policy changes in China.
2. “Opinionated News?” (w/ J. Chan & S. Slavov).
    - Quantifies subjectivity in American journalism.

- Different subjects/languages, same method.

1. “Reading China”

Predicting policy change: why?

- China’s industrialization: a product of gov’t direction.
- Opaque system makes prediction prohibitively difficult.

Predicting policy change: why?

- China’s industrialization: a product of gov’t direction.
- Opaque system makes prediction prohibitively difficult...

  ... until now.

- We construct the first predictive algorithm for China’s policy shifts.

Predicting policy change: how?

Build a neural network algorithm to

- “read” the *People's Daily*;
- detect structural changes in its priorities.

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.”
- Necessity “to transform the press… into a serious organ for the economic **education** of the mass of the population.”

Source of predictive power

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

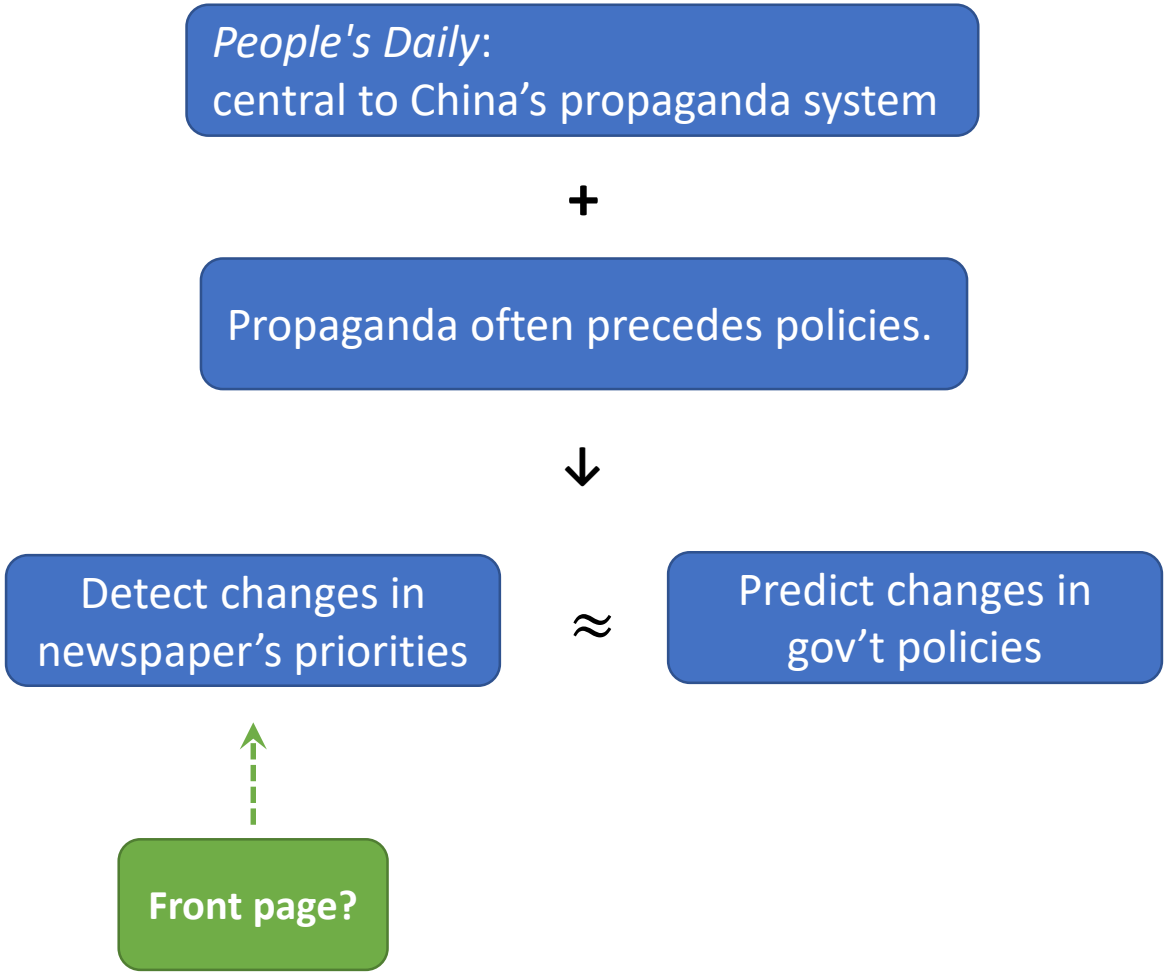
+

Propaganda often precedes policies.

↓

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

Source of predictive power

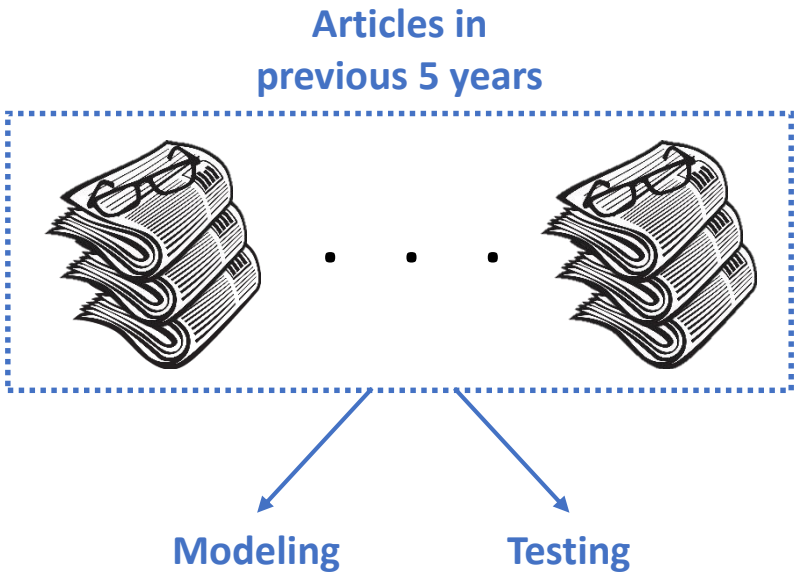


Method

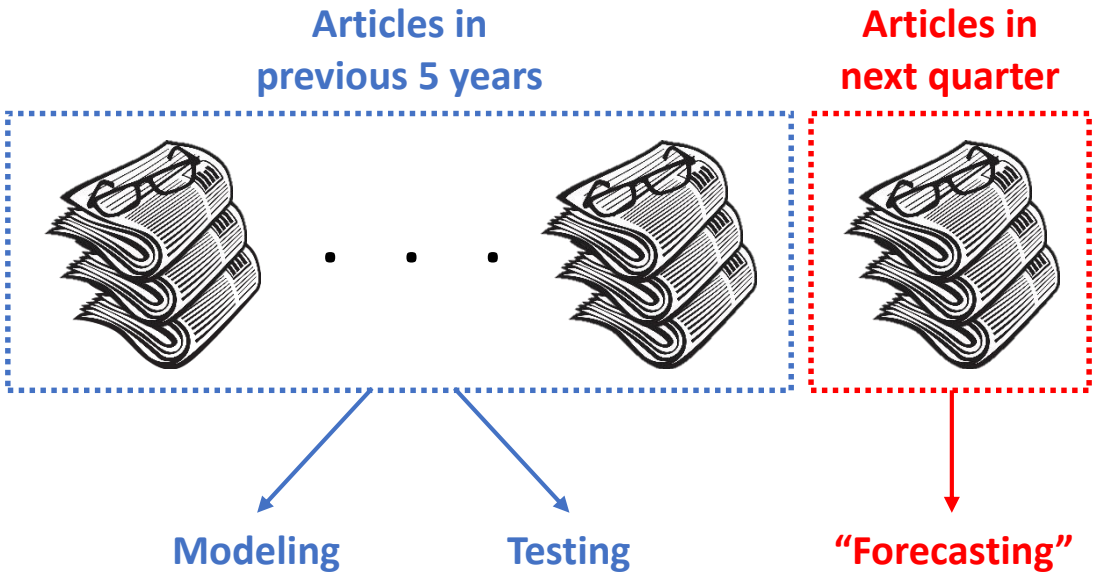
Train a **front-page** classifier by

“reverse-engineering” the editor’s mind.

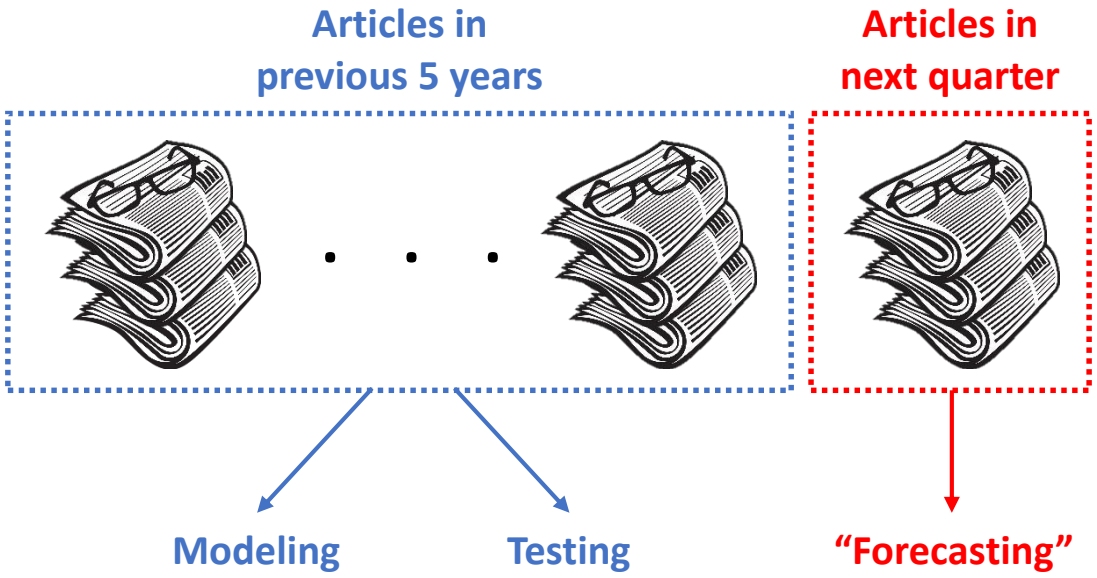
Method



Method


Articles in
previous 5 years
Articles in
next quarter
Modeling
Testing
“Forecasting”

Method



$$\text{Policy Change Index} = \left| \text{Testing performance} - \text{"Forecasting" performance} \right|$$

PCI $\gg 0 \Rightarrow$ Structural difference in data-generating process

Method: data structure

```
sample_df
```

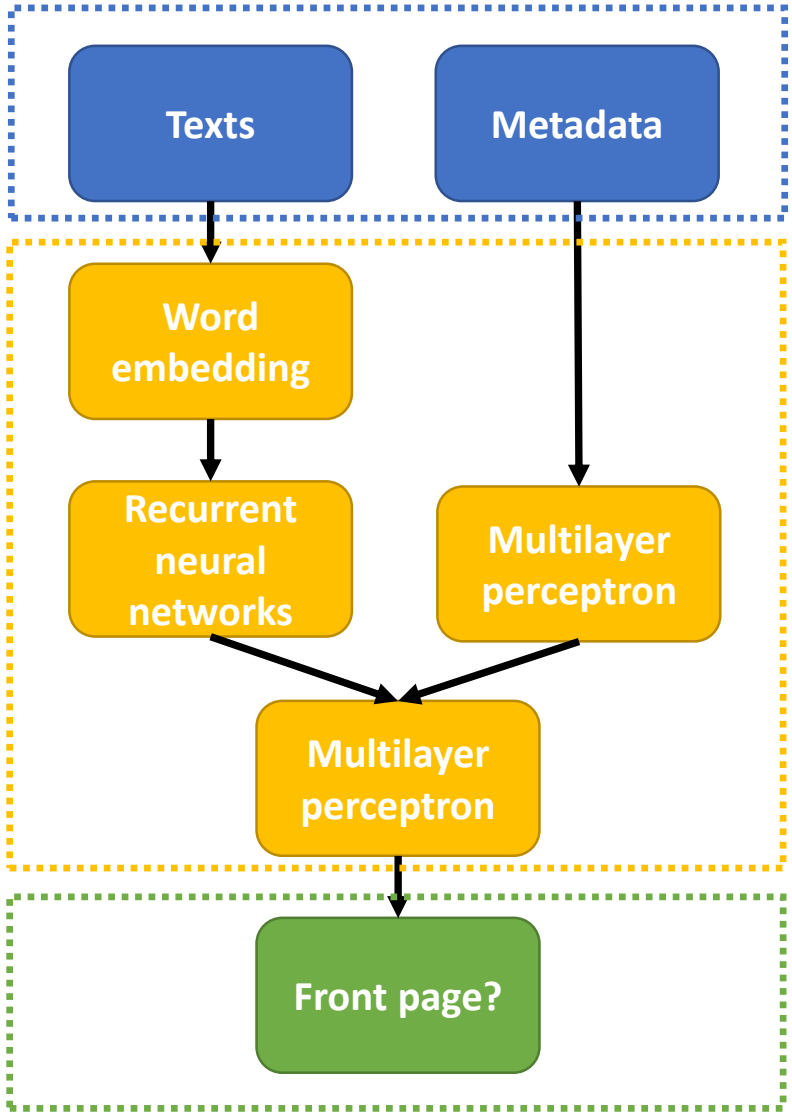
id	date	page	title	body
20181202525	2018-12-31 00:00:00	1.0	今年粮食总产量13158亿斤　农业农村发展取得新成绩，乡村振兴开局良好	本报北京12月30日电　（记者高云才、郁静娴）记者从30日召开的全国农业农村厅局长会议获...
20181202526	2018-12-31 00:00:00	1.0	新时代改革再出发的重要里程碑　——写在党的十八届三中全会召开五周年之际	本报评论员　改革开放40年来，我们党带领人民绘就了一幅波澜壮阔、气势恢宏的历史画卷。其...
20181202527	2018-12-31 00:00:00	1.0	改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　林丽鹏　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...

Method: 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　林丽鹏　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...

Method: modelling

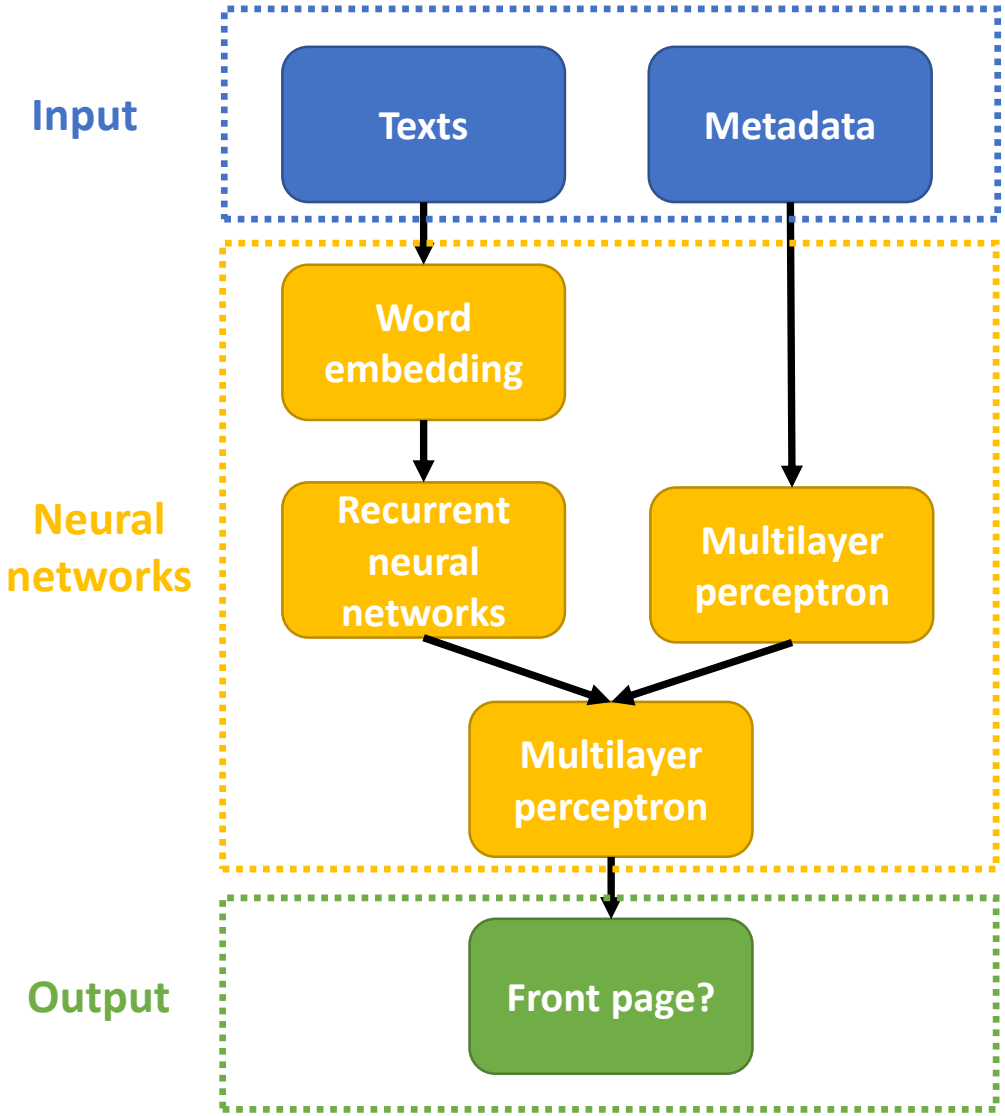


x : each article as an observation.

f : a complicated function.

$y = f(x)$

Method: modelling

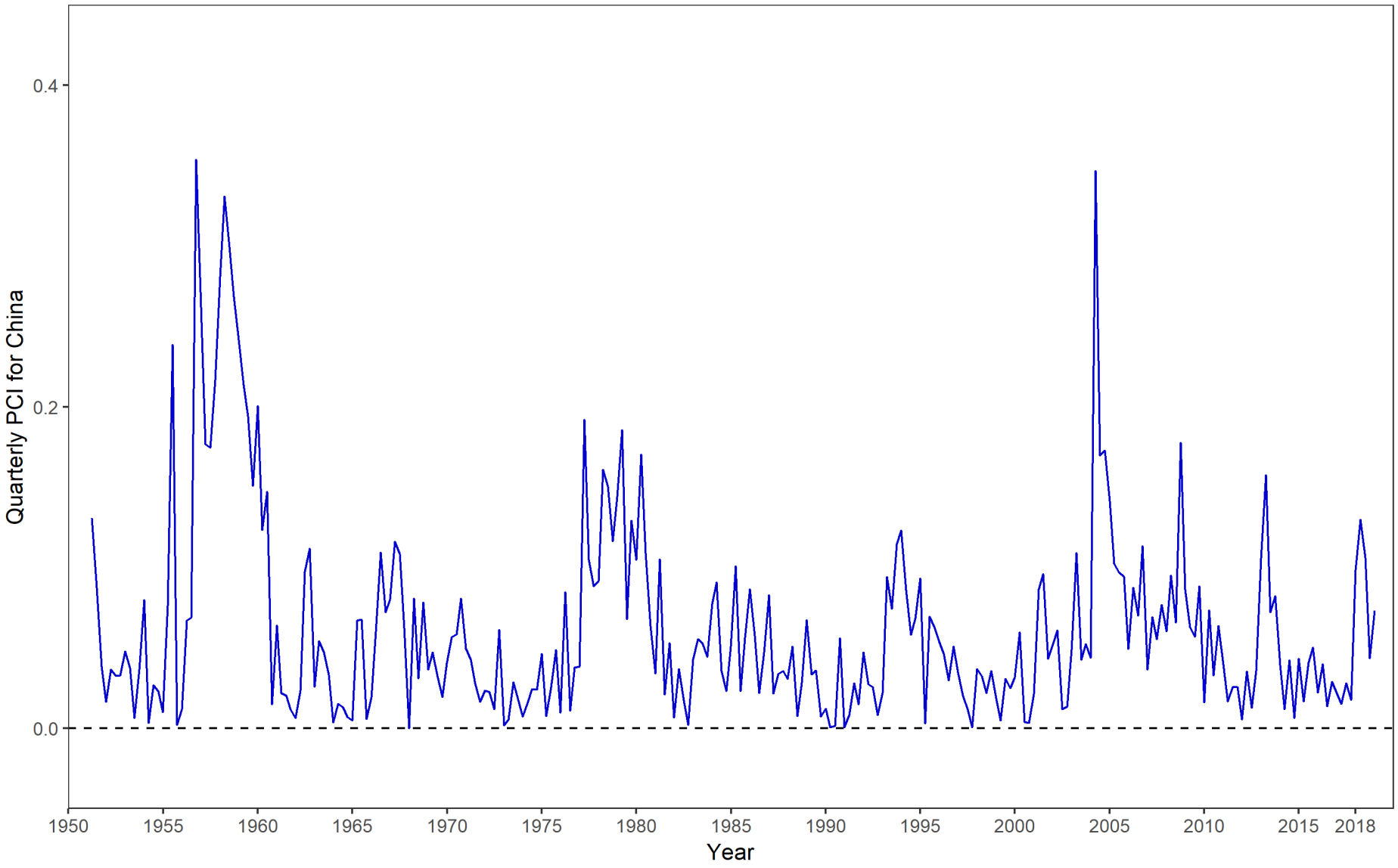


x : each article as an observation.

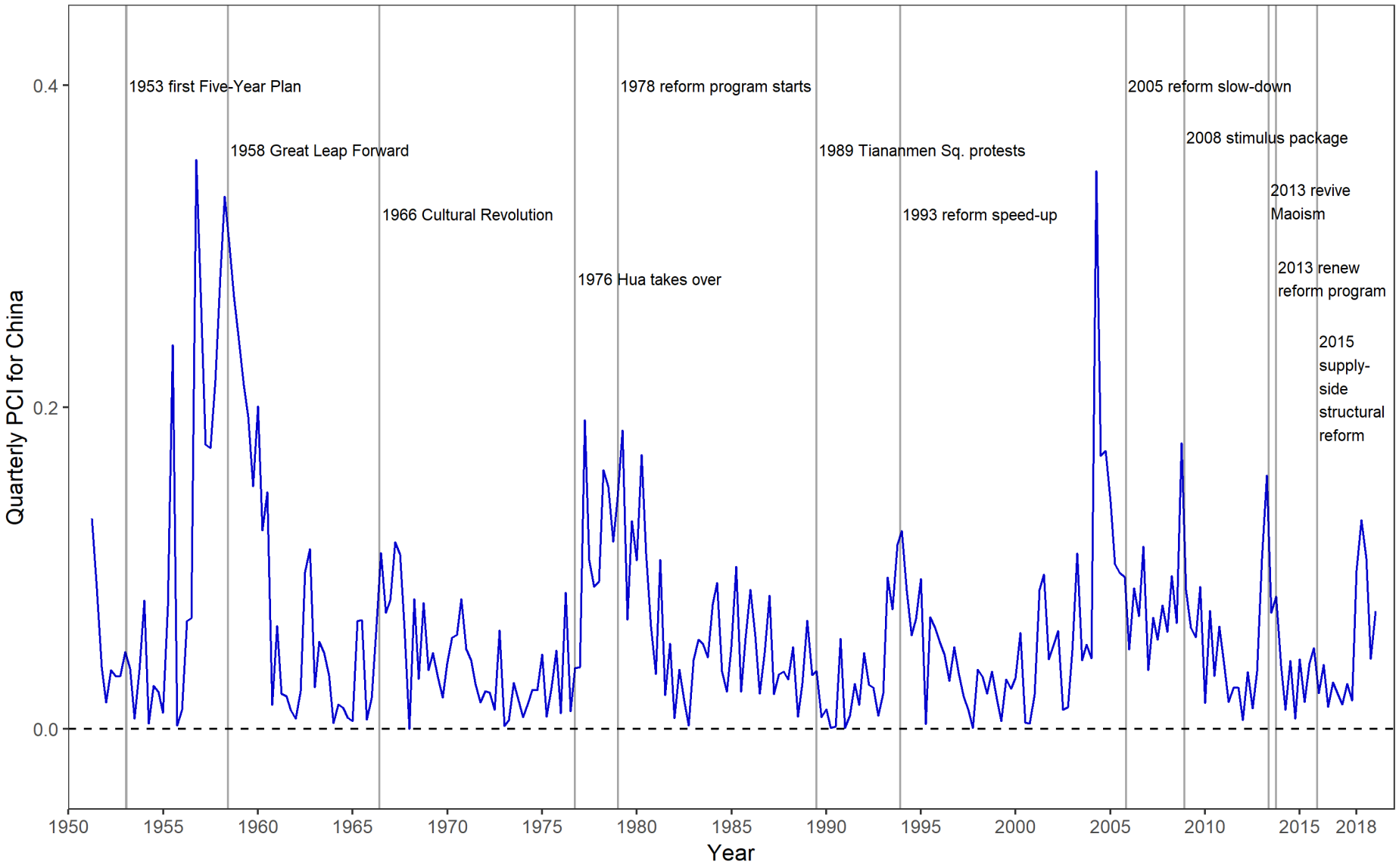
f : a complicated function.

$$y = f(x)$$

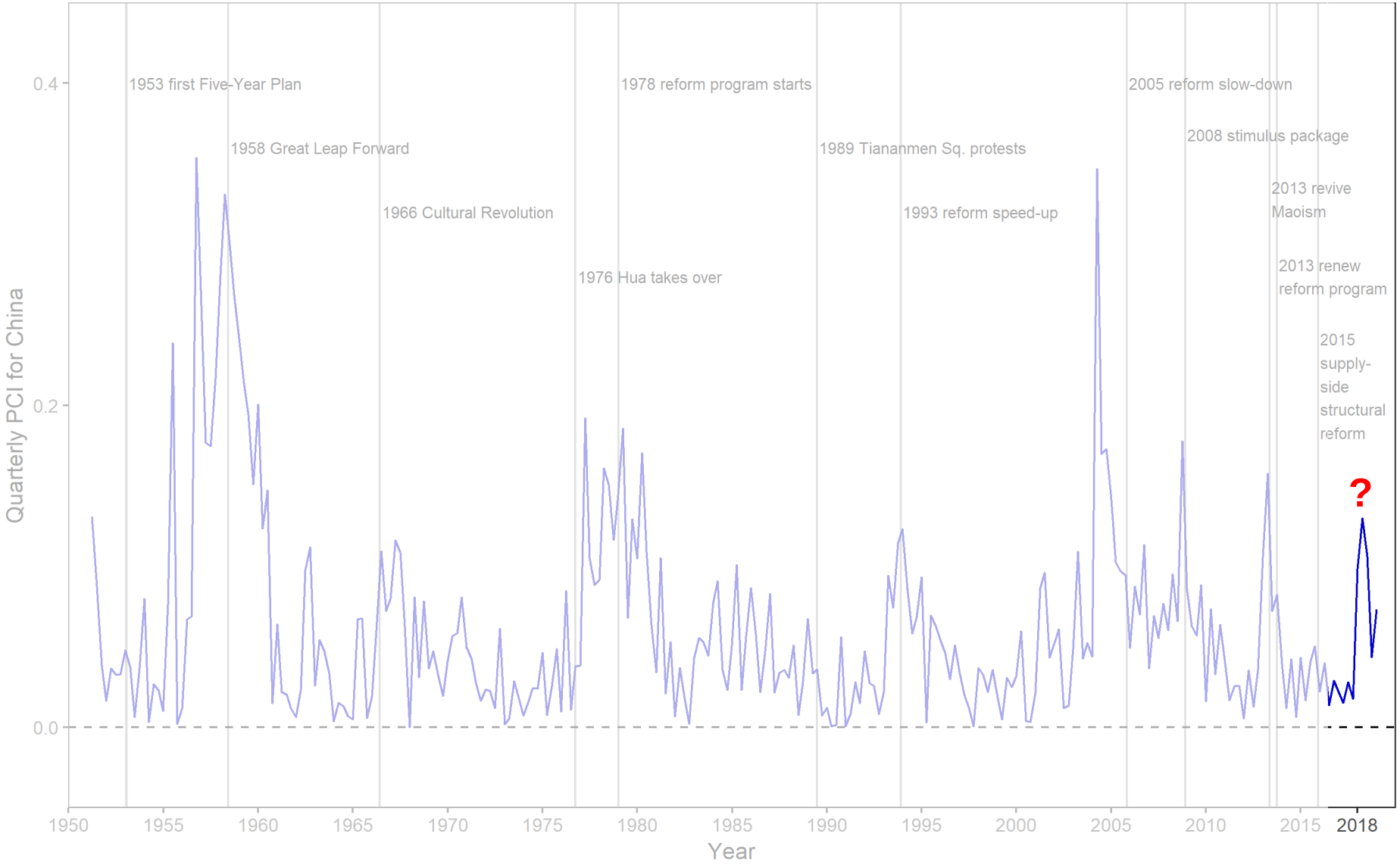
Result: PCI


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



Result: PCI — going forward



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.

Understanding substance of change

What does the 2018 Q1 uptick represent?

- Strengthening party authority;
- Emphasizing nationalism and global leadership;
- Populist policies to boost political support.

Understanding substance of change

What does the 2018 Q1 uptick represent?

- Strengthening party authority;
- Emphasizing nationalism and global leadership;
- Populist policies to boost political support.

⇒ Curb your enthusiasm for US-China trade talks.

More generally

PCIs for other (ex-)Communist regimes, using:

- Soviet Union's *Pravda*
- East Germany's *Neues Deutschland*
- Cuba's *Granma*
- North Korea's *Rodong Sinmun*
- Vietnam's *Nhân Dân*

More generally

PCIs for other (ex-)Communist regimes, using:

- Soviet Union's *Pravda*
- East Germany's *Neues Deutschland*
- Cuba's *Granma*
- North Korea's *Rodong Sinmun*
- Vietnam's *Nhân Dân*

Work in progress

Even more generally

Applicability well beyond using page numbers.

2. “Opinionated News?”

Americans' declining trust in media

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?

Detecting opinionated news

Data: *The New York Times*, 1987-2007.

A “translation” of the PCI method:

- PD articles ⇢ **NYT articles**;
- front-page indicator ⇢ **opinion indicator**;
- other components stay the same.

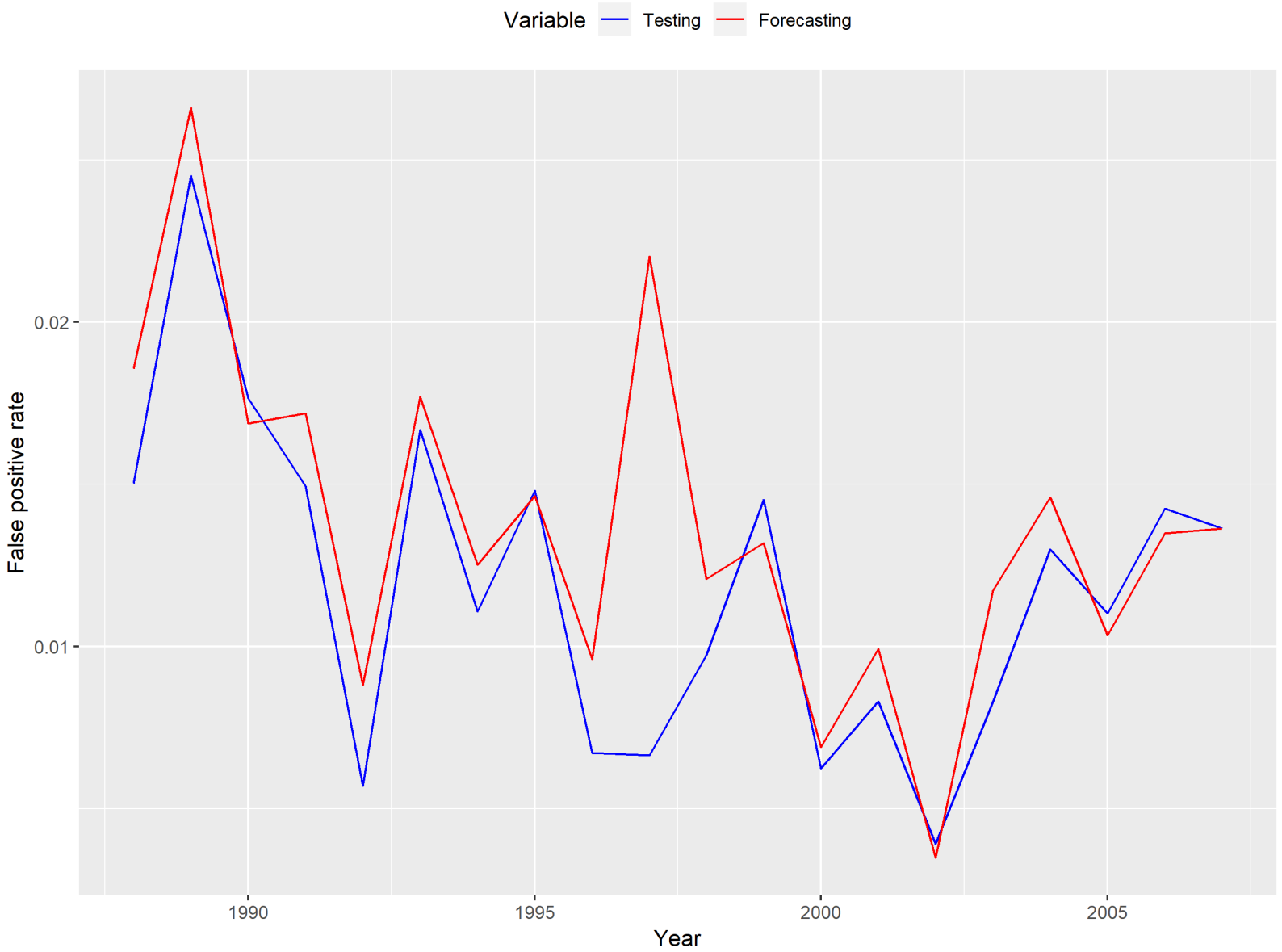
Detecting opinionated news

News misclassified as opinion

⇒ More opinionated than the algorithm “thinks”

Metric: false positive rate of the opinion classifier.

Opinionated news? — Not really…



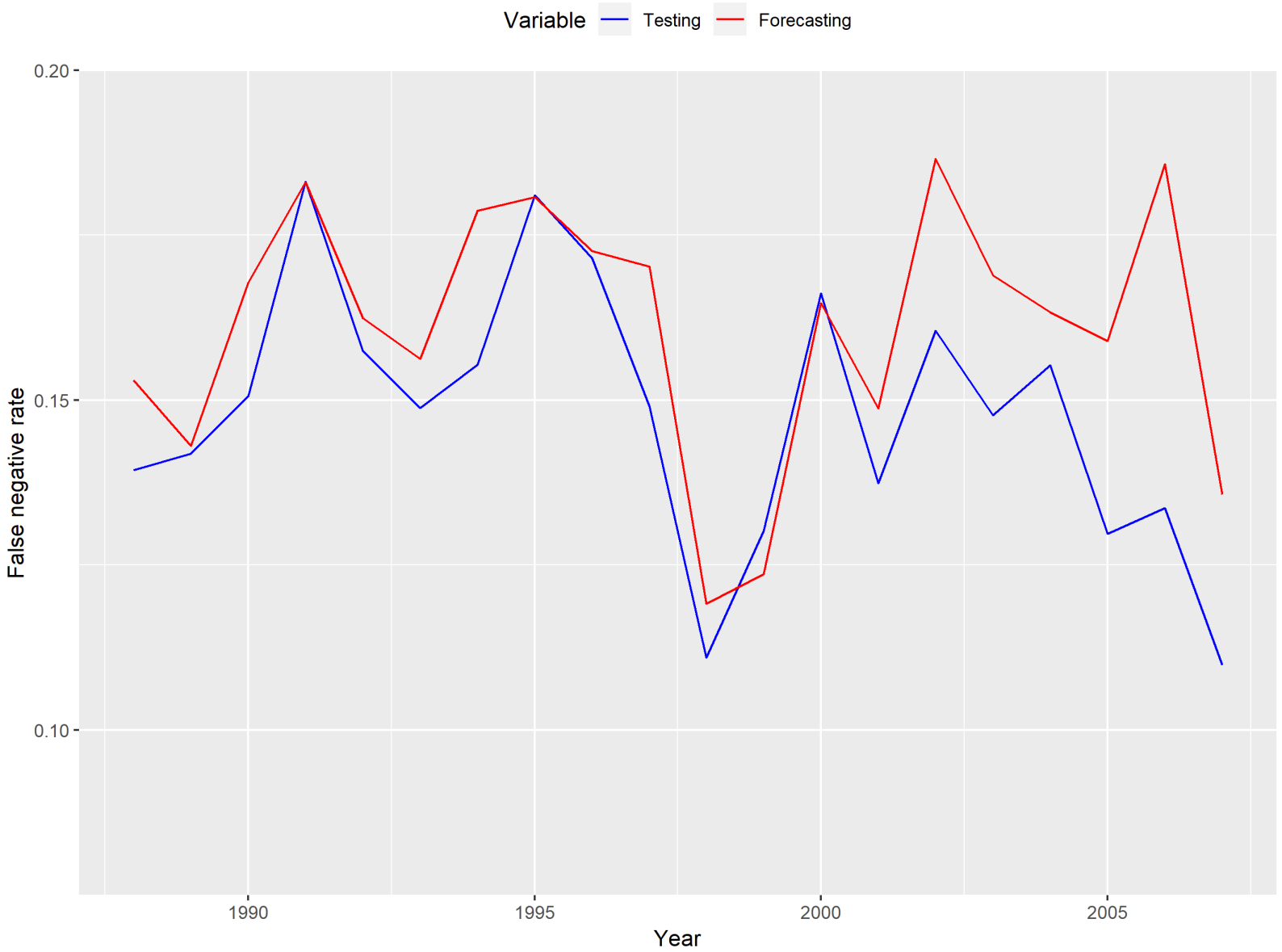
How about the opposite?

Opinion misclassified as news

⇒ More “factual” than the algorithm “thinks”

Metric: false negative rate of the opinion classifier.

Factual commentary? — Yes.



Prelim finding

NYT shows *increasing objectivity* from 2001 to 2007.

Machine learning — with a twist

To recap the common method

The twist:

- Use seemingly trivial metadata as training labels;
- Realize that “forecasting” errors contain information.

Interested in DIY?

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

Interested in DIY?

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

Questions?