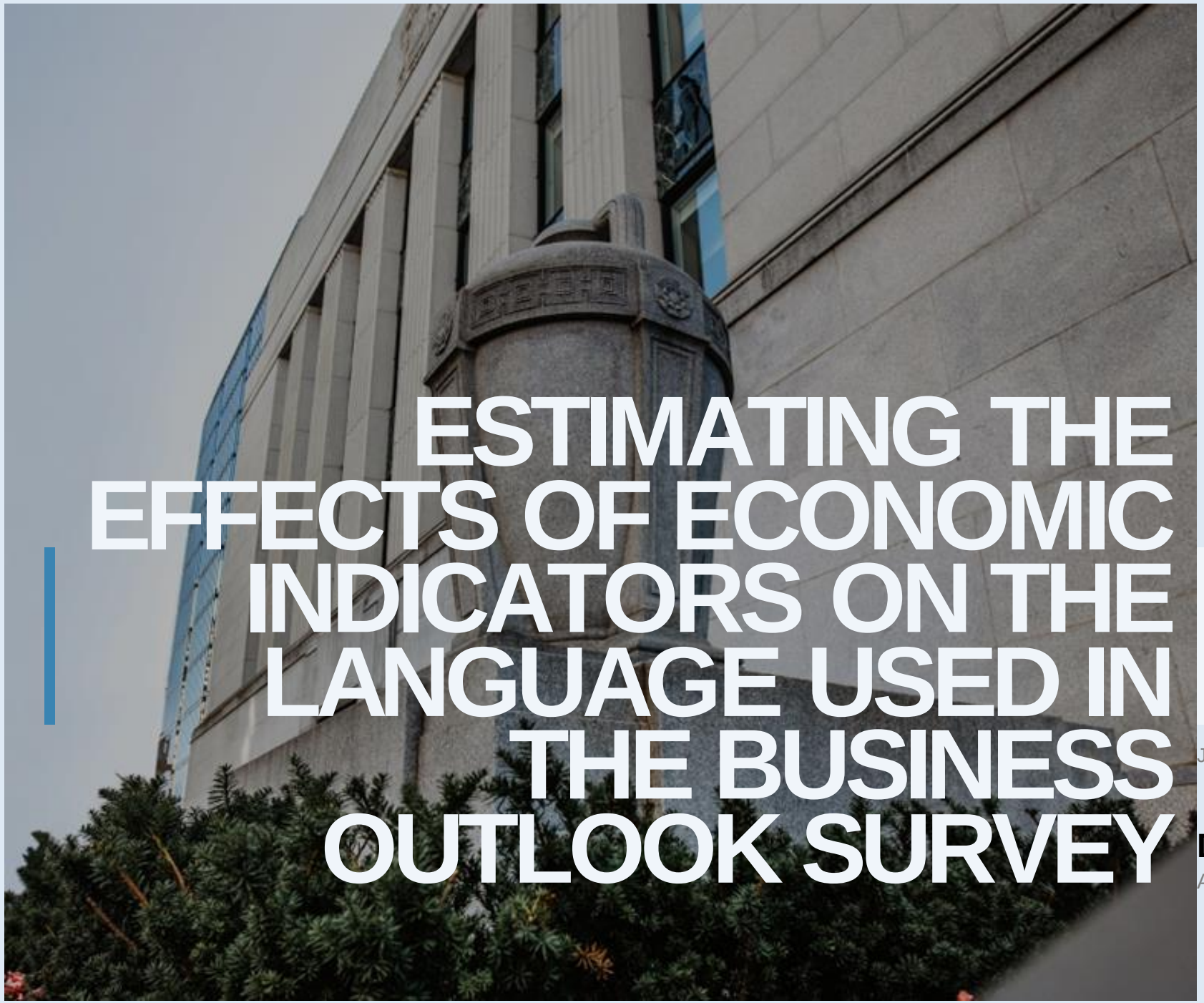
IFC Satellite Seminar on "Granular data: new horizons and challenges for central banks"

Estimating the effects of economic indicators on the language used in the business outlook survey[1]

Dave Campbell, Charita Koya and Naveen Rai,
Bank of Canada

[1] This contribution was prepared for the IFC Satellite Seminar held at the ISI 64th World Statistics Congress, co-organised with the Bank of Canada in Ottawa, Canada, on 15 July 2023. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Canada, the BIS, the IFC or the other central banks and institutions represented at the event.



ESTIMATING THE EFFECTS OF ECONOMIC INDICATORS ON THE LANGUAGE USED IN THE BUSINESS OUTLOOK SURVEY

BANK OF CANADA
BANQUE DU CANADA

Charita Koya, Naveen Rai

July 2023

Dave Campbell
Assistant Director Data Science Applications, Bank of Canada

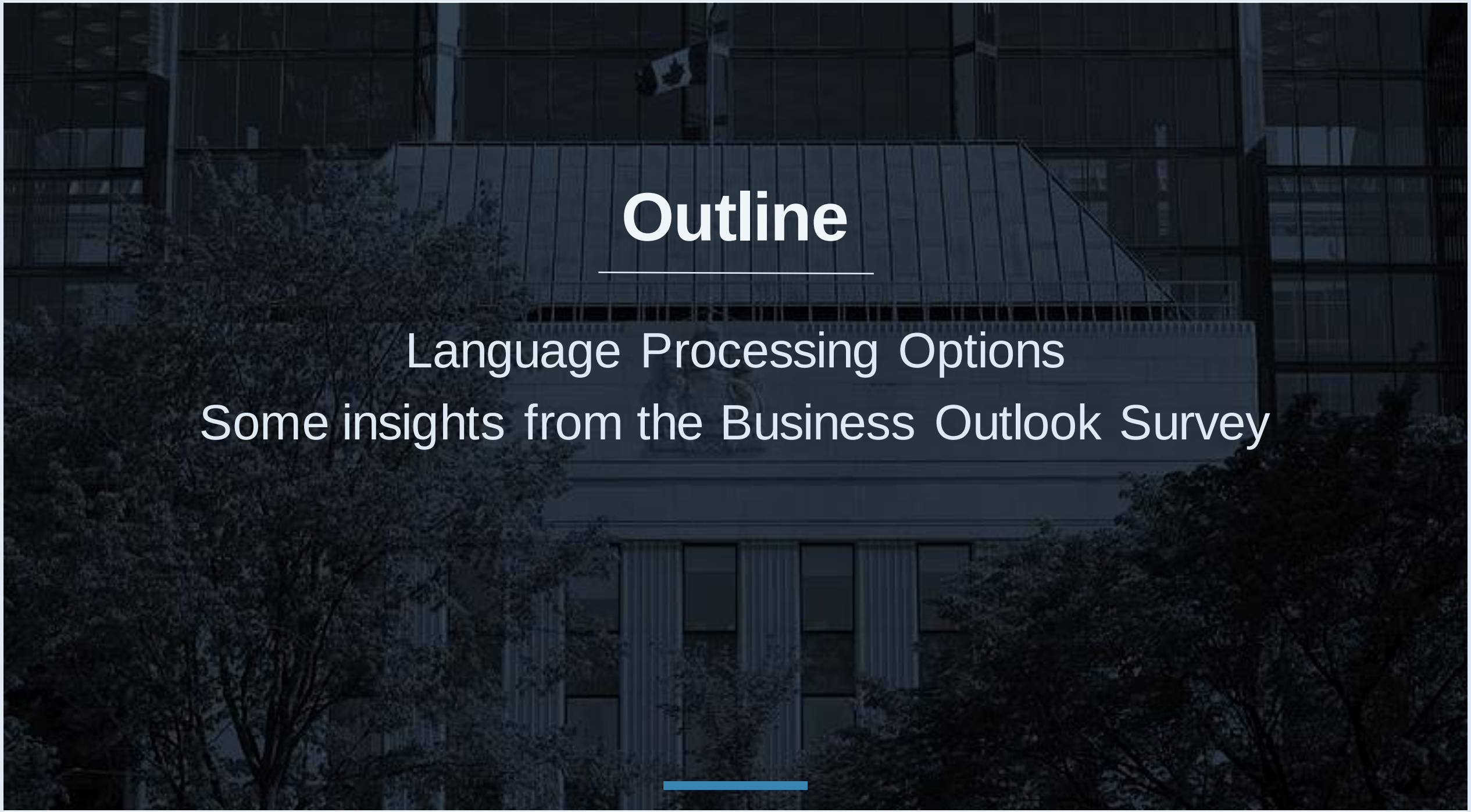
Business Outlook Survey

https://www.bankofcanada.ca/publications/bos/business-outlook-survey-data/

"Four times a year, our regional offices interview business leaders from about 100 firms to gather a broad range of economic perspectives ..."

Stories are gathered about wages, labour, sales, investment, and more

How is the language of stories impacted by the state of the economy?



Outline

Language Processing Options

Some insights from the Business Outlook Survey

Regression on Language Models

Using the topics of discourse (X) to estimate economic variables (Y)

$$Y = f(X_{text}) + \varepsilon$$

Estimate Y and X together or condition on ‘known’ X

Algorithm results are unstable or unidentifiable to rotation

Regression of Language Models

How do Consumer Price Index (X) impact the discourse around investment in the Business Outlook Survey (Y)?

$$Y_{text} = f(X) + \varepsilon$$

Y needs to be meaningfully measured to quantify stability

Converting Text into Numbers

Language Processing = {data acquisition, pre-processing, algorithm}

Each response to a labour question from 2009 - 2022 is a document

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0

- After pre-processing construct a Term x Document matrix
- #Terms = 8,677
- #Documents = 16,776
- Define the ‘word space’ of the text

Decomposing the Text Stories into Substructure

- term-document matrix N_{docs} X N_{terms}.

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0

- Extract Substructures = Topics of discourse based on how words are used together

Non-Negative Matrix Factorization with Anchor Words

Algorithm basics:

Idea: Decompose the Term x Document matrix into [Term xTopic] and [Topic x Document] matrices

Model: topic weight within document = $f(X) + \varepsilon$

Inference on *f* comes from Bootstrap

Non-negative Matrix Factorization: “eigen-ish” decomposition

- term-document matrix N_{docs} X N_{terms}. Decomposed

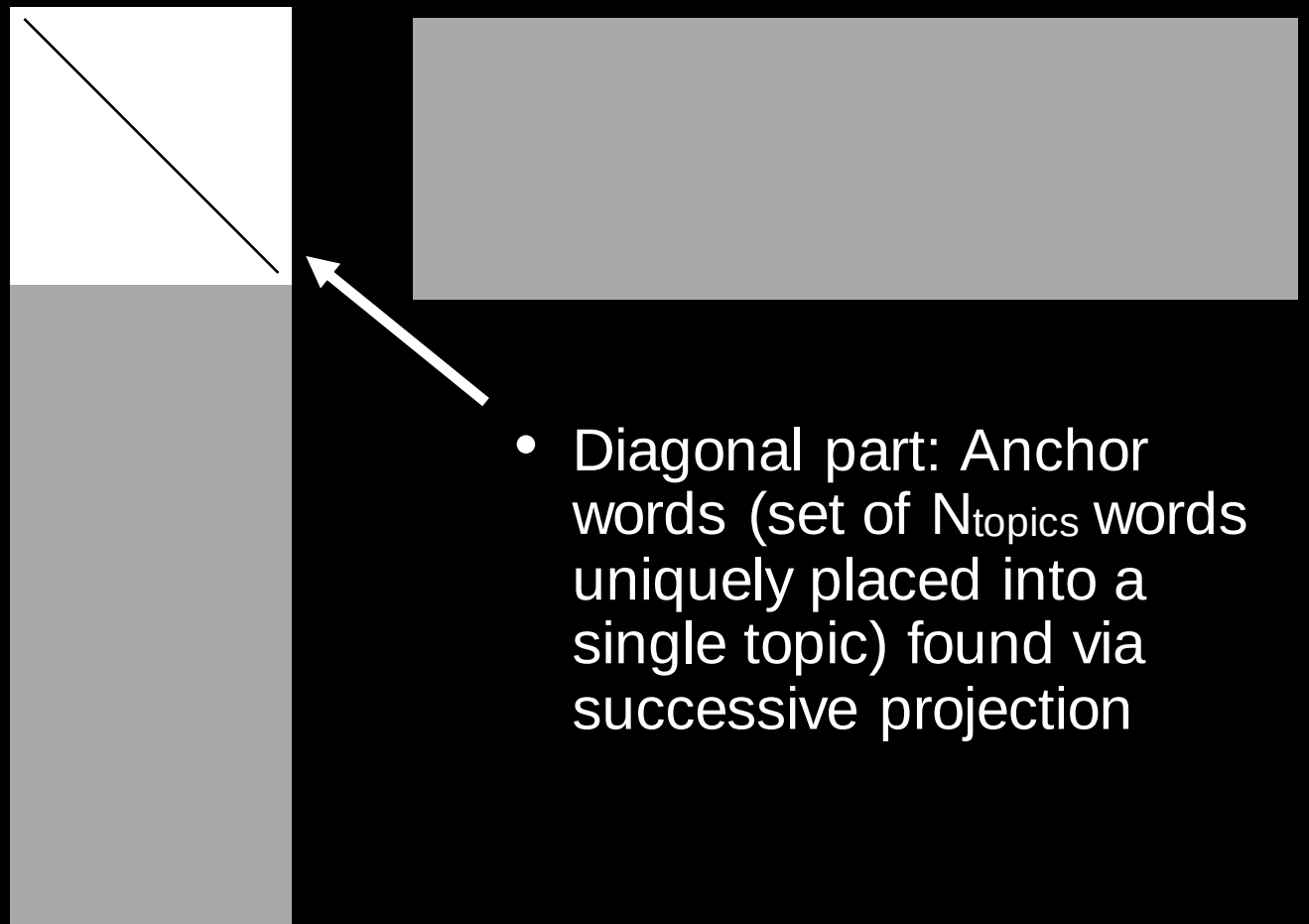
=

terms X topics (N_{terms} X N_{topics})

topics X document (N_{topics} X N_{docs})

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0

=



- Diagonal part: Anchor words (set of N_{topics} words uniquely placed into a single topic) found via successive projection

Non-negative Matrix Factorization: “eigen-ish” decomposition

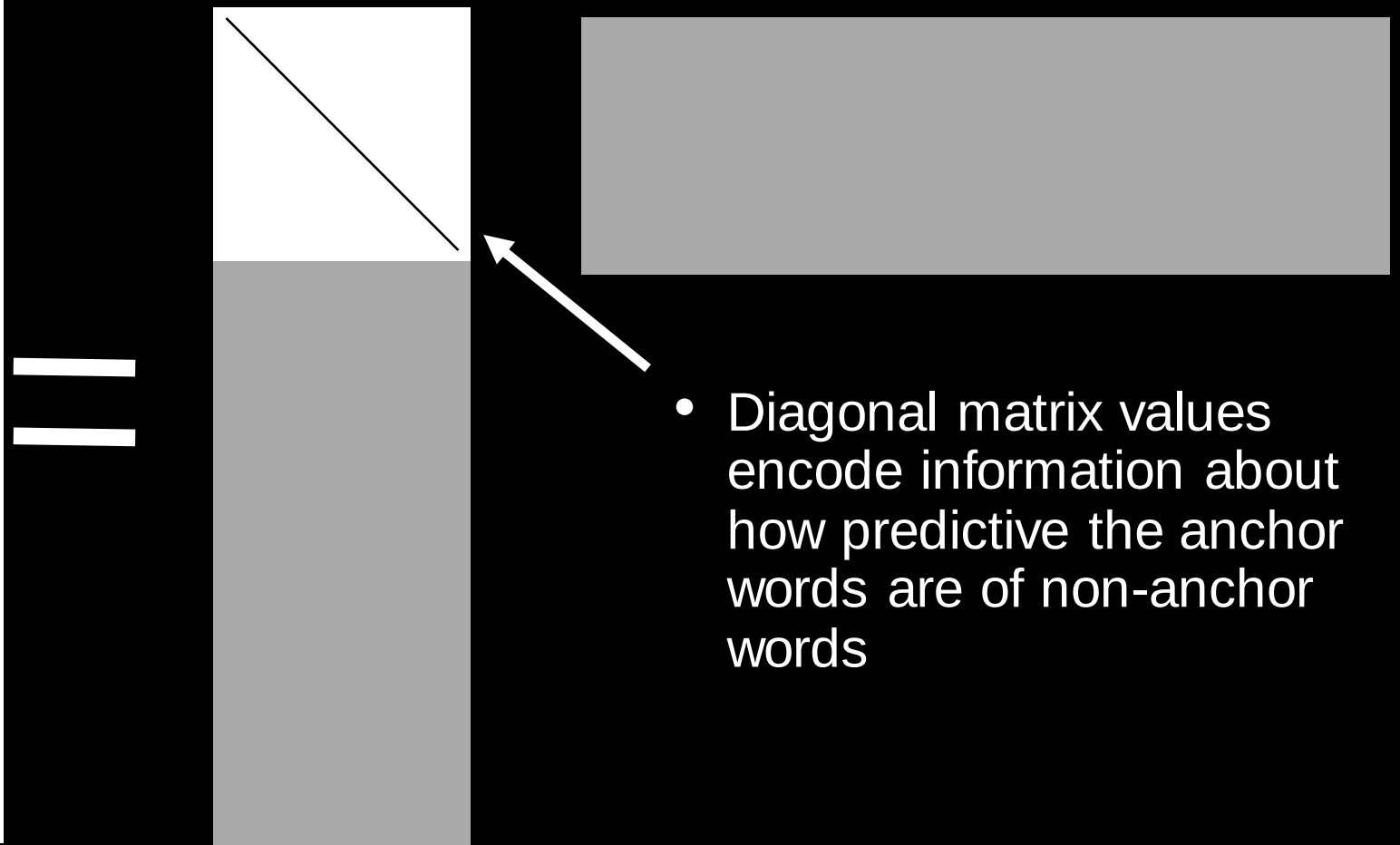
- term-document matrix N_{docs} X N_{terms}. Decomposed

=

terms X topics (N_{terms} X N_{topics})

topics X document (N_{topics} X N_{docs})

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0



- Diagonal matrix values encode information about how predictive the anchor words are of non-anchor words

Non-negative Matrix Factorization: “eigen-ish” decomposition

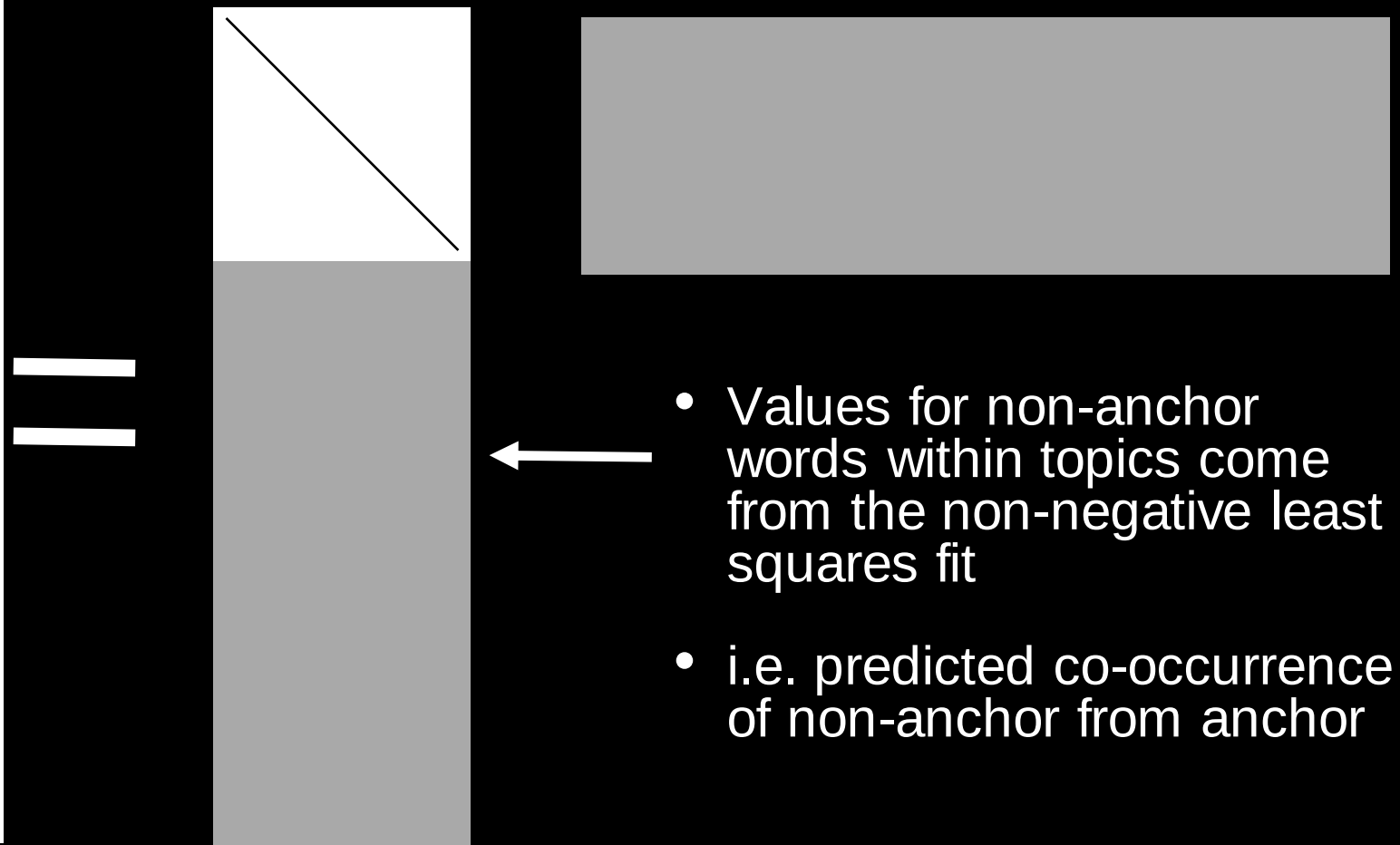
- term-document matrix N_{docs} X N_{terms}. Decomposed

=

terms X topics (N_{terms} X N_{topics})

topics X document (N_{topics} X N_{docs})

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0



- Values for non-anchor words within topics come from the non-negative least squares fit
- i.e. predicted co-occurrence of non-anchor from anchor

Non-negative Matrix Factorization: “eigen-ish” decomposition

- term-document matrix N_{docs} X N_{terms}. Decomposed

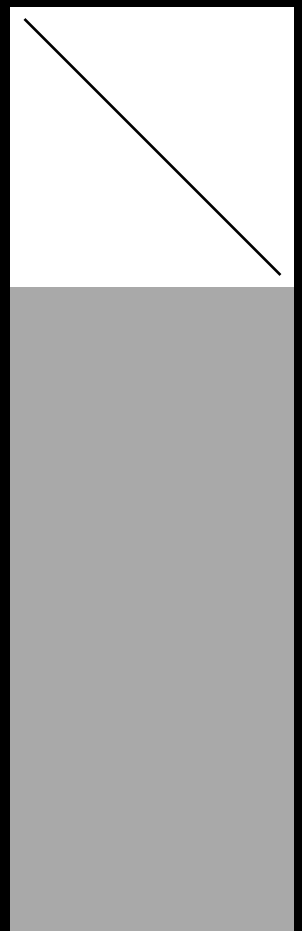
=

terms X topics (N_{terms} X N_{topics})

topics X document (N_{topics} X N_{docs})

	Docs				
Terms	15950	6515	1281	14677	11726
people	0	2	3	2	2
labour	0	0	3	1	0
staff	2	0	0	0	4
work	0	2	2	5	1
year	3	0	0	0	5
employees	2	0	0	2	0
wage	1	1	0	0	1
demand	0	1	0	1	0
workers	0	0	0	0	1
increase	2	0	0	0	1
covid	0	2	1	1	0
get	0	1	1	1	2
shortages	0	1	2	0	0
back	0	0	0	2	1
due	0	0	1	0	0
higher	2	0	0	0	0
last	1	1	0	0	5
skilled	0	1	1	0	1
market	6	0	1	0	0
months	0	2	1	1	0

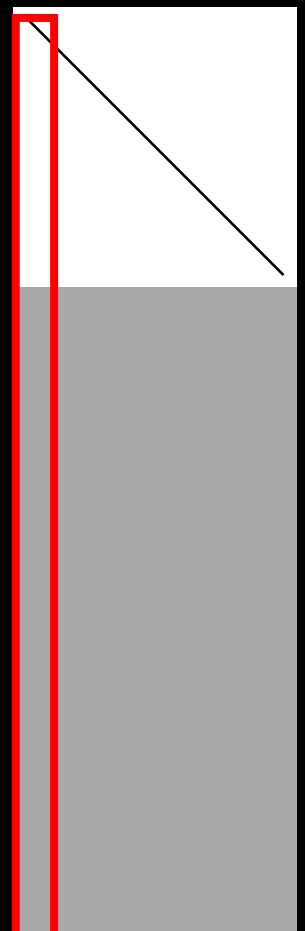
=

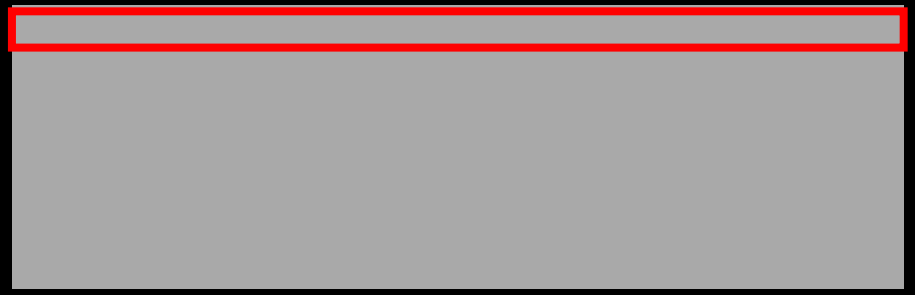


- The Topic intensity per document is found by projecting the Term-Document matrix into the anchor space

Non-negative Matrix Factorization (NMF)

- term-document matrix N_{docs} X N_{terms}. Decomposed = terms X topics (N_{terms} X N_{topics}) topics X document (N_{topics} X N_{docs})
  - Use Term x Topic to define what we are measuring



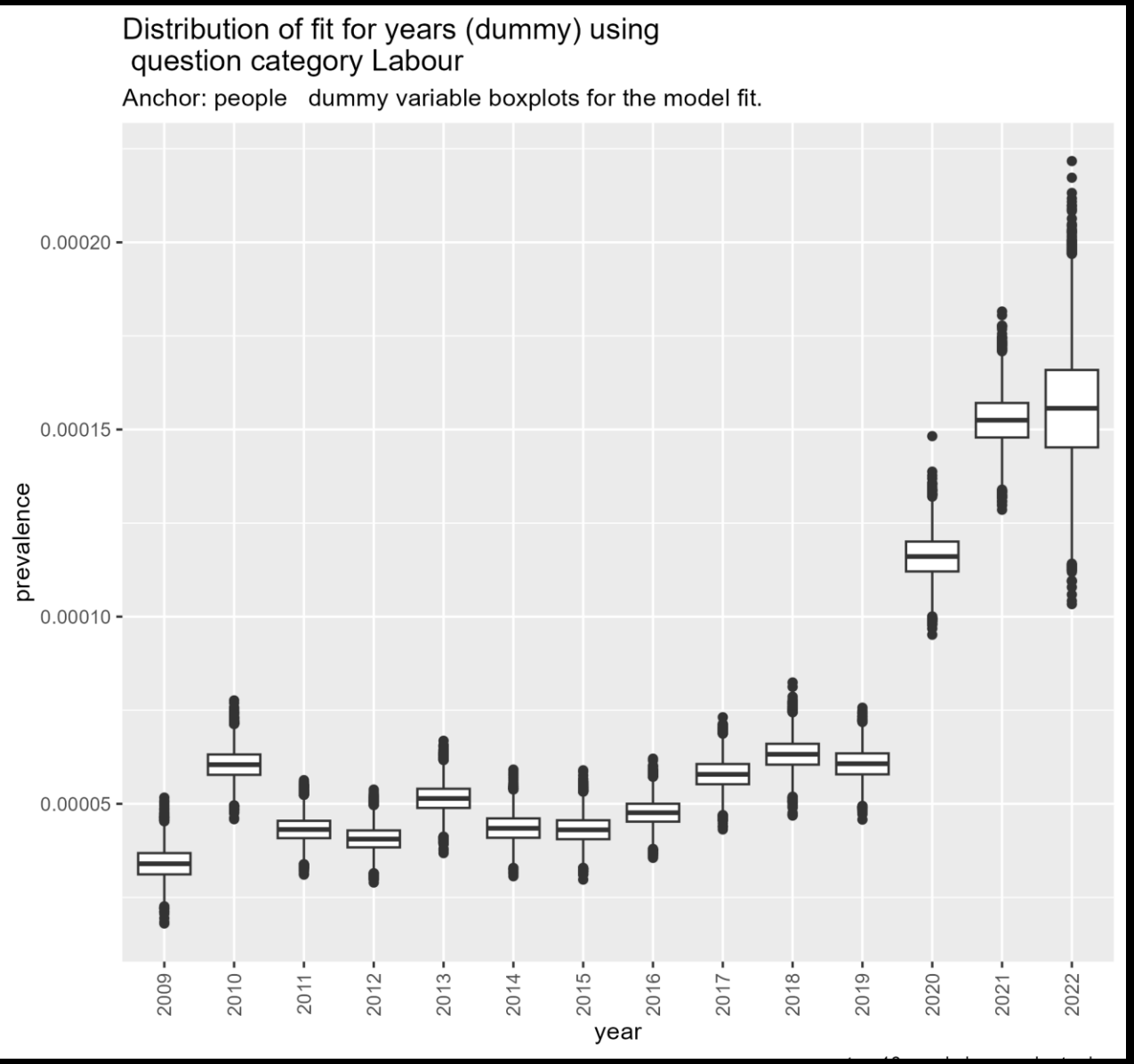


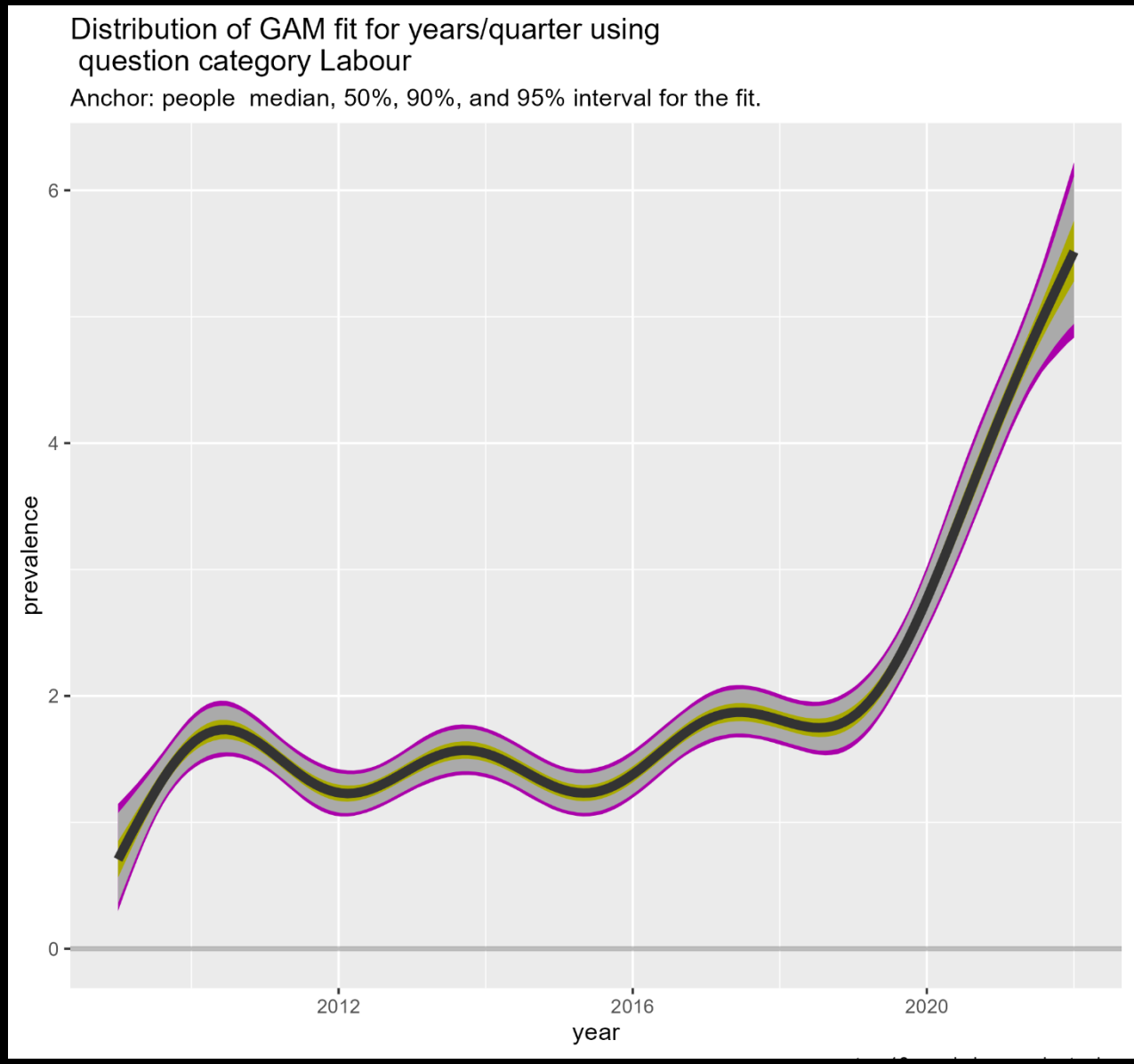
- Within the labour questions:
- Anchor words:
- "people" "staff" "year" "labour" "wage"

people topic:

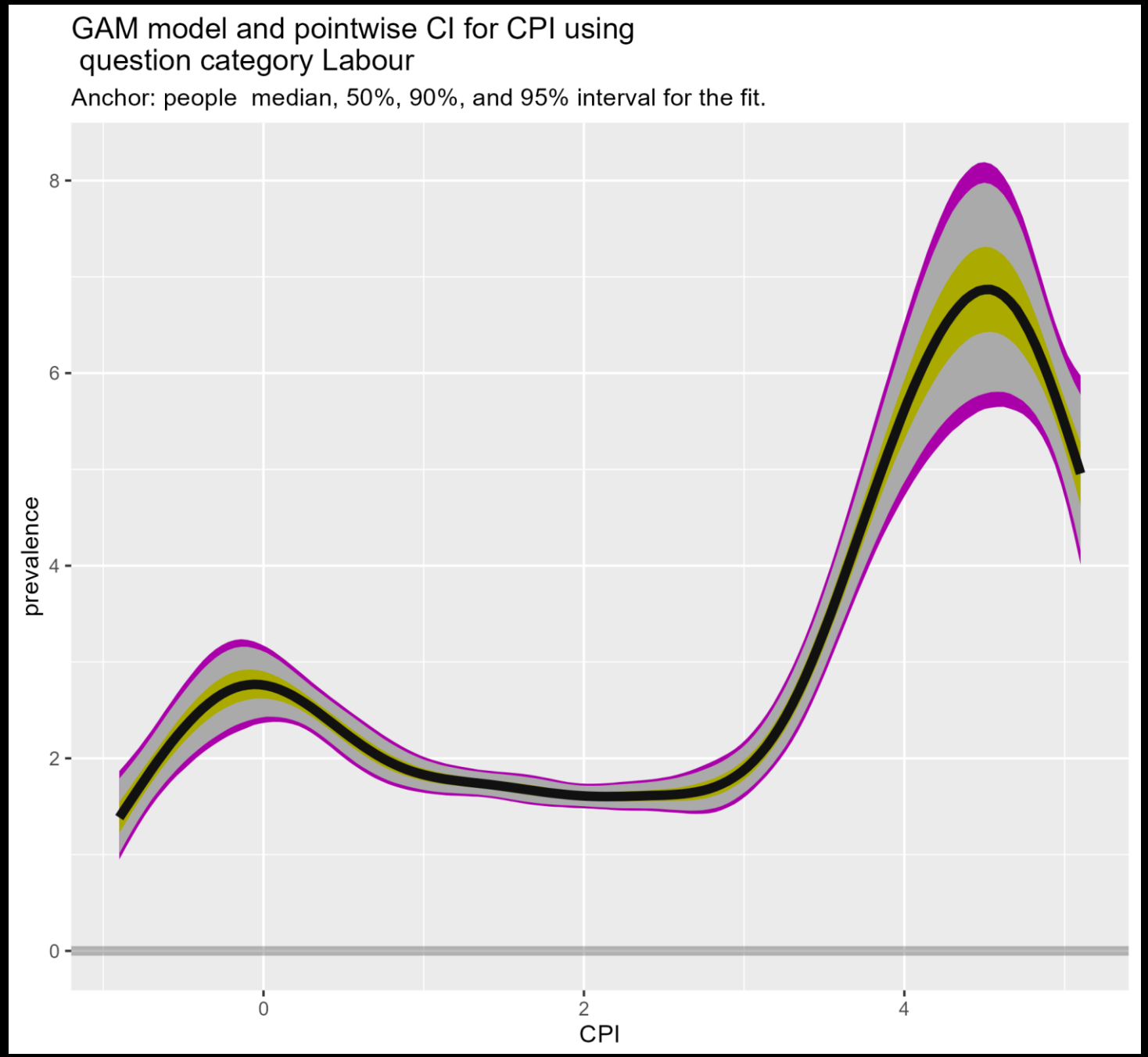
abated, workaround, hires, ghosting, newest, backgrounds, wanted

- Use corresponding row of Topic x Document as our Y
- Economic variables as our X
- Fit model $Y_{text} = f(X) + \varepsilon$
- Inference from bootstrap

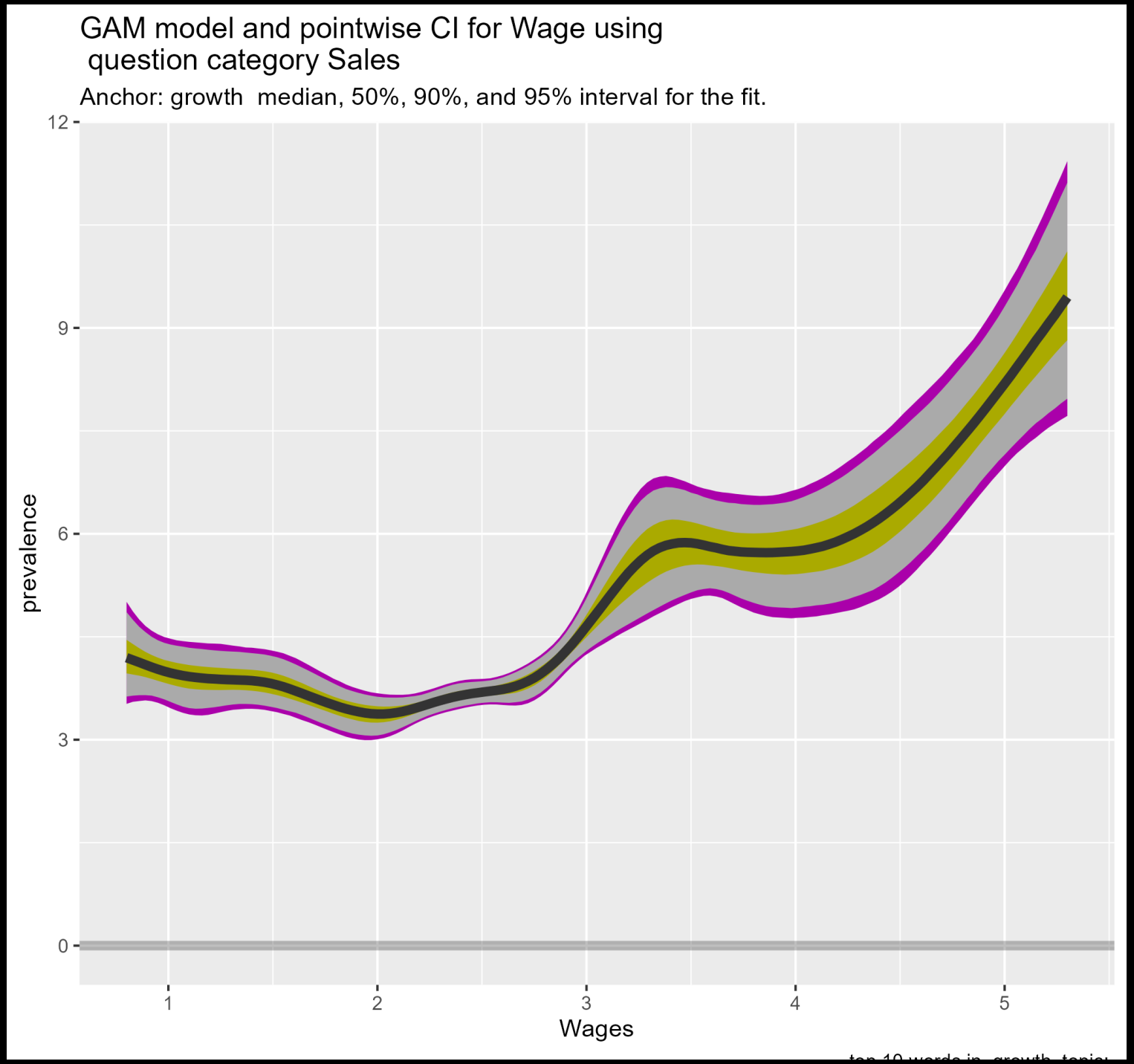

Distribution of fit for years (dummy) using
question category Labour
Anchor: people dummy variable boxplots for the model fit.
prevalence
0.00020
0.00015
0.00010
0.00005
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
year

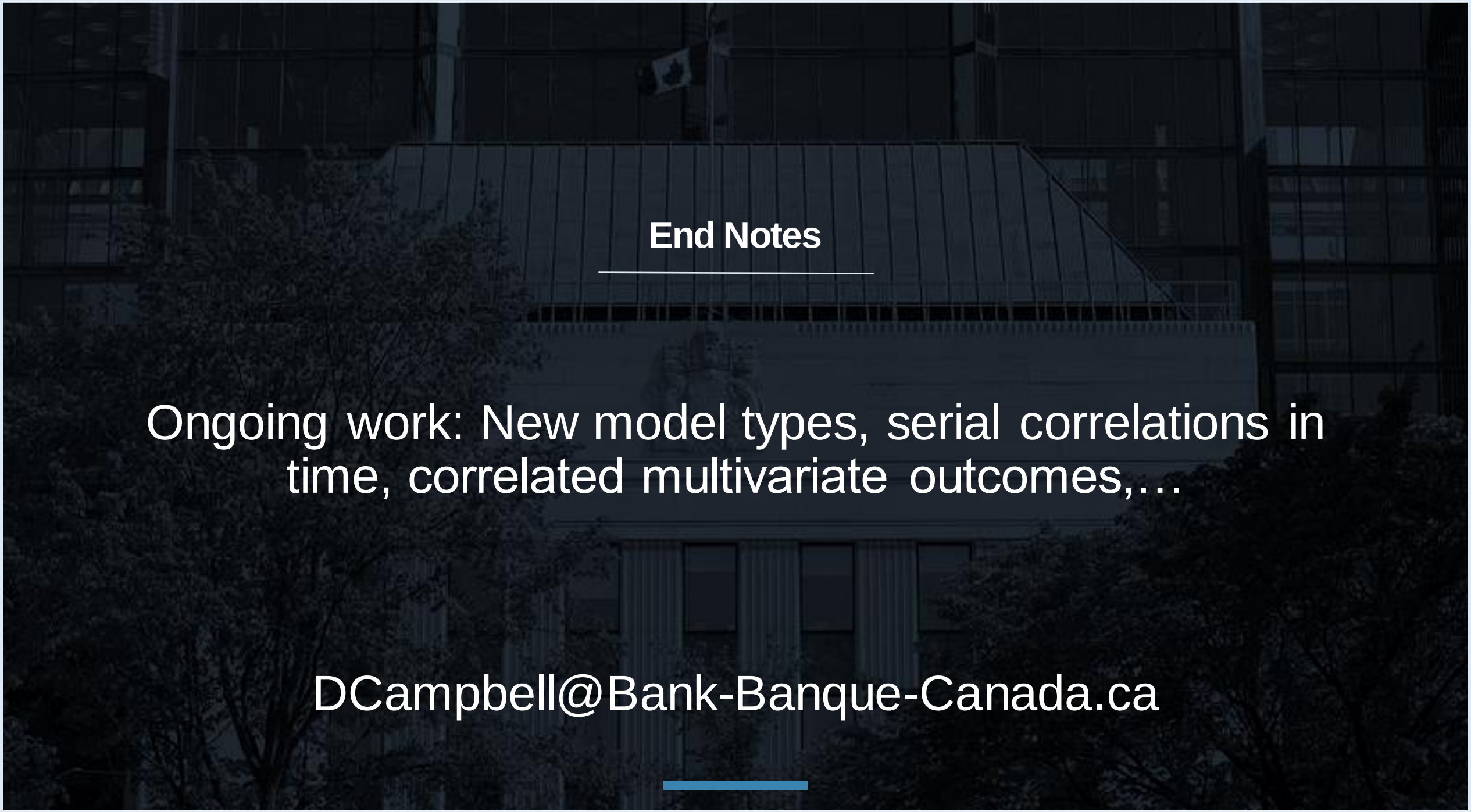

Distribution of GAM fit for years/quarter using
question category Labour
Anchor: people median, 50%, 90%, and 95% interval for the fit.
prevalence
6
4
2
0
2012
2016
2020
year

- X = CPI


GAM model and pointwise CI for CPI using question category Labour
Anchor: people median, 50%, 90%, and 95% interval for the fit.
prevalence
8
6
4
2
0
0
2
4
CPI

- X = Wages


GAM model and pointwise CI for Wage using question category Sales
Anchor: growth median, 50%, 90%, and 95% interval for the fit.
prevalence
12
9
6
3
0
1
2
3
4
5
Wages



End Notes

Ongoing work: New model types, serial correlations in time, correlated multivariate outcomes,…

DCampbell@Bank-Banque-Canada.ca