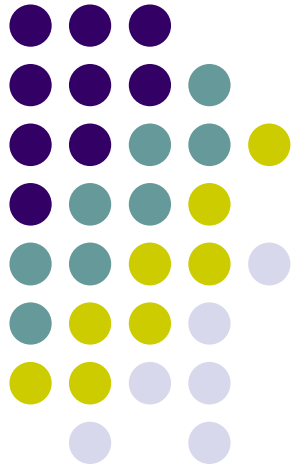
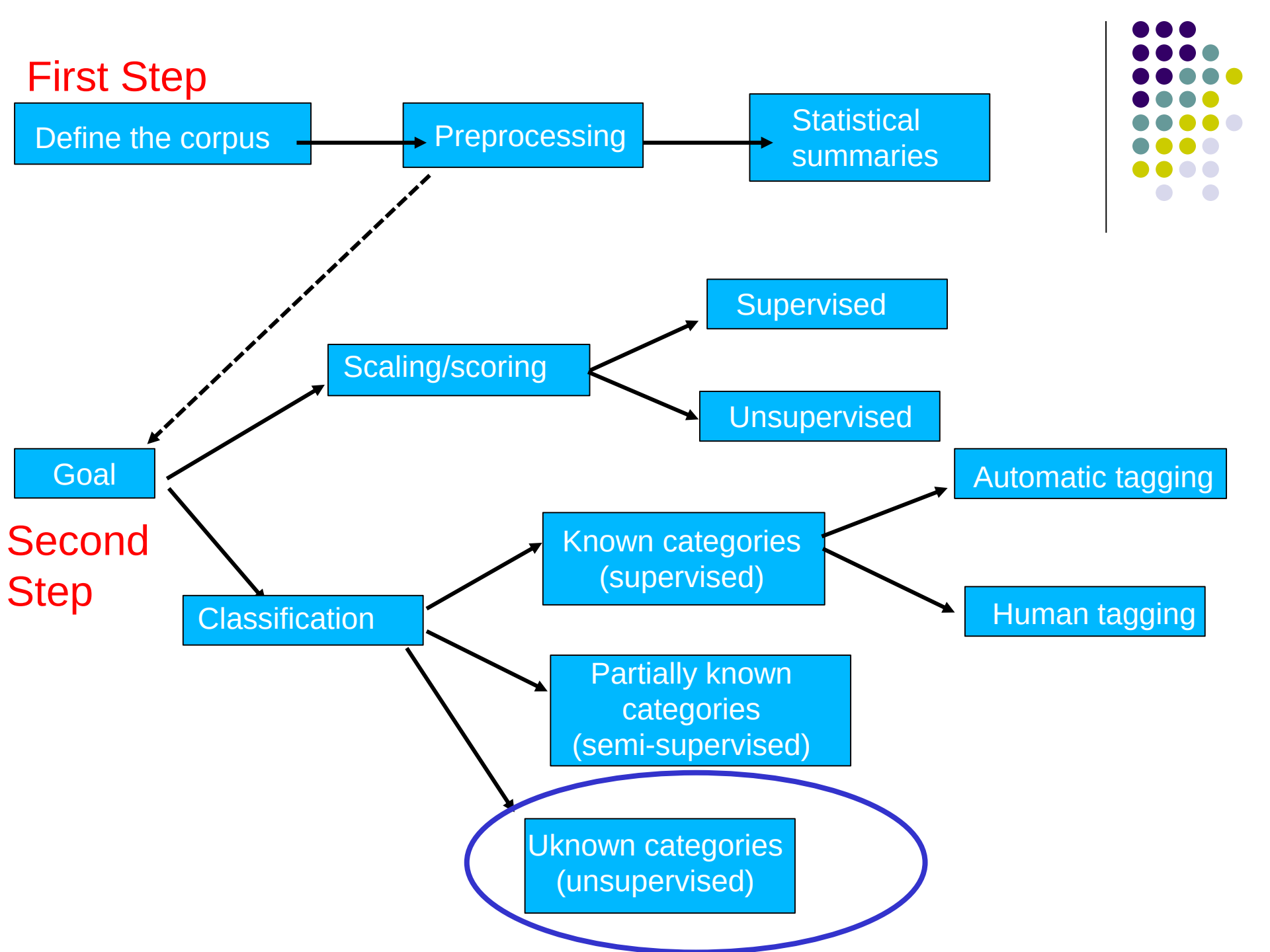
Big Data Analytics

Lecture 5 – Part 1

Unsupervised classification methods: the strucutural topic model




First Step
Define the corpus
Preprocessing
Statistical summaries
Supervised
Scaling/scoring
Unsupervised
Goal
Automatic tagging
Second Step
Known categories (supervised)
Classification
Human tagging
Partially known categories (semi-supervised)
Uknown categories (unsupervised)

Reference

- ✓ Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Luca, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G. Rand. 2014. Structural Topic Models for Open-Ended Survey Response, *American Journal of Political Science*, 58(4), 1064-1082
- ✓ Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley(2014). STM: R Package for Structural Topic Models, *Journal of Statistical Software*, https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

Classification methods

Structural Topic Model (STM) innovates on Topic models in two different ways:

First: topic proportions (θ) are allowed to be **correlated**: this is a reasonable assumption given that in documents topics discussed are correlated!

For example, if a manifesto contains discussion of Topic X (e.g. administrative reform), the probabilities that it will also contain discussion of Topics Y (e.g. curbing public works) and Z (e.g. reducing the number of Lower House members), are not independent of each other, but correlated

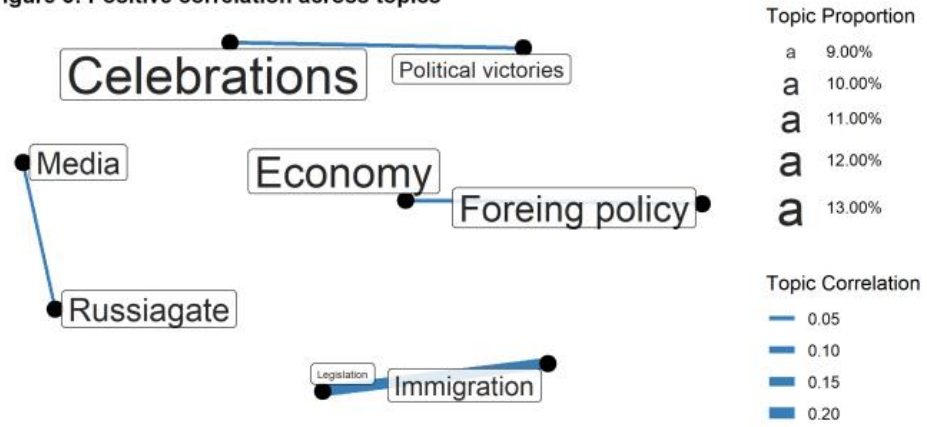
In this sense, STM fits a Correlated Topic Model (rather than a LDA)

Classification methods

Graphical depictions of the (*positive*) correlation between topics provide insight into the organizational structure at the corpus level

In essence, the model identifies when two topics are likely to co-occur (by focusing on positive correlation) within a document

Figure 3: Positive correlation across topics



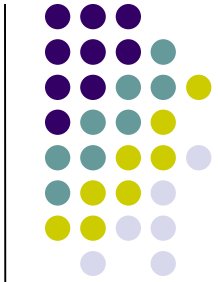
Source: Results from a Structural Topic Model on @realDonaldTrump Twitter account

Classification methods

Second: as we already know, in all topic models the analyst estimates for each document the proportion of words attributable to each topic, providing a measure of *topic prevalence*

The model also calculates the words most likely to be generated by each topic, which provides a measure of *topical content*

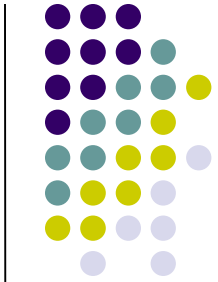
Classification methods



However, in standard LDA, the document collection is assumed to be **unstructured**; that is, each document is assumed to arise from the same data-generating process irrespective of additional information (about the corpus) the analyst might possess. And that shouldn't be always the case…

Suppose for example that you have reasons to believe that the **age** of a text's author affects the probability to discuss about a given topic rather than some other alternatives. Or the probability of using some words rather than others to discuss about a given topic. How to *incorporate such information (i.e., such structure)* in the analysis?

Classification methods



You cannot do that via standard Topic Models…but yes in a STM! That's precisely the second (and crucial) innovation of a STM compared to a LDA Topic Model

In a STM each document can have its own prior distribution over topics according to the document-level variables you decide to include in the fitted topic model (i.e., **topical prevalence** – the *thetas* – can be affected by the covariates you include in the topic model), rather than sharing a global mean

Same things happen for **topical content**, i.e., the *betas* of your fitted topic model

Classification methods

That is…

…rather than assuming that **topic prevalence** (i.e., the frequency with which a topic is discussed) and **topical content** (i.e., the words used to discuss a topic) prior distribution **are constant** across all documents...

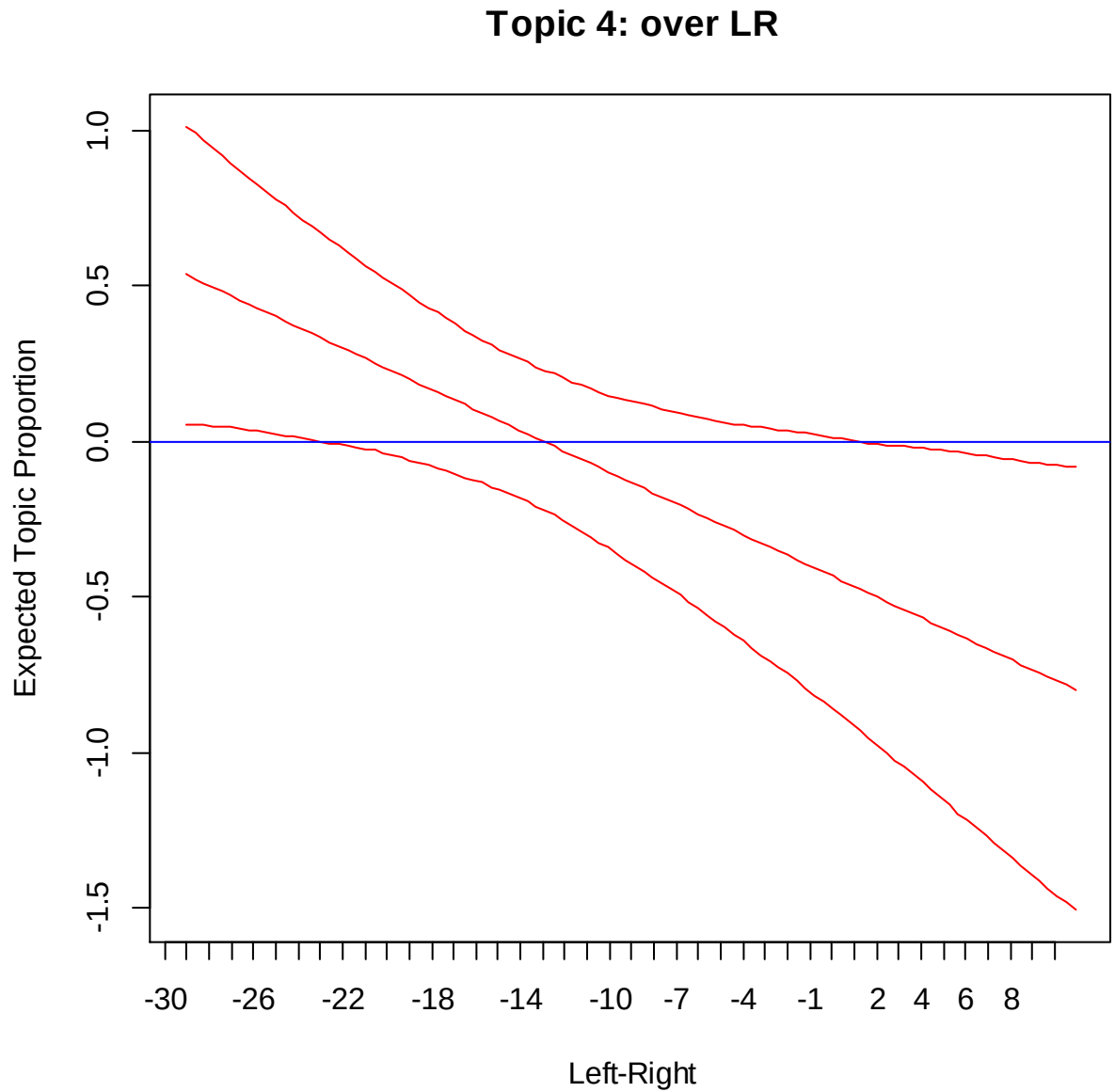
…the analyst can incorporate covariates over which we might expect to see variance directly when estimating the topics

Through this, we can obtain measures of how our treatment condition systematically affects how often a topic is discussed (prevalence)!

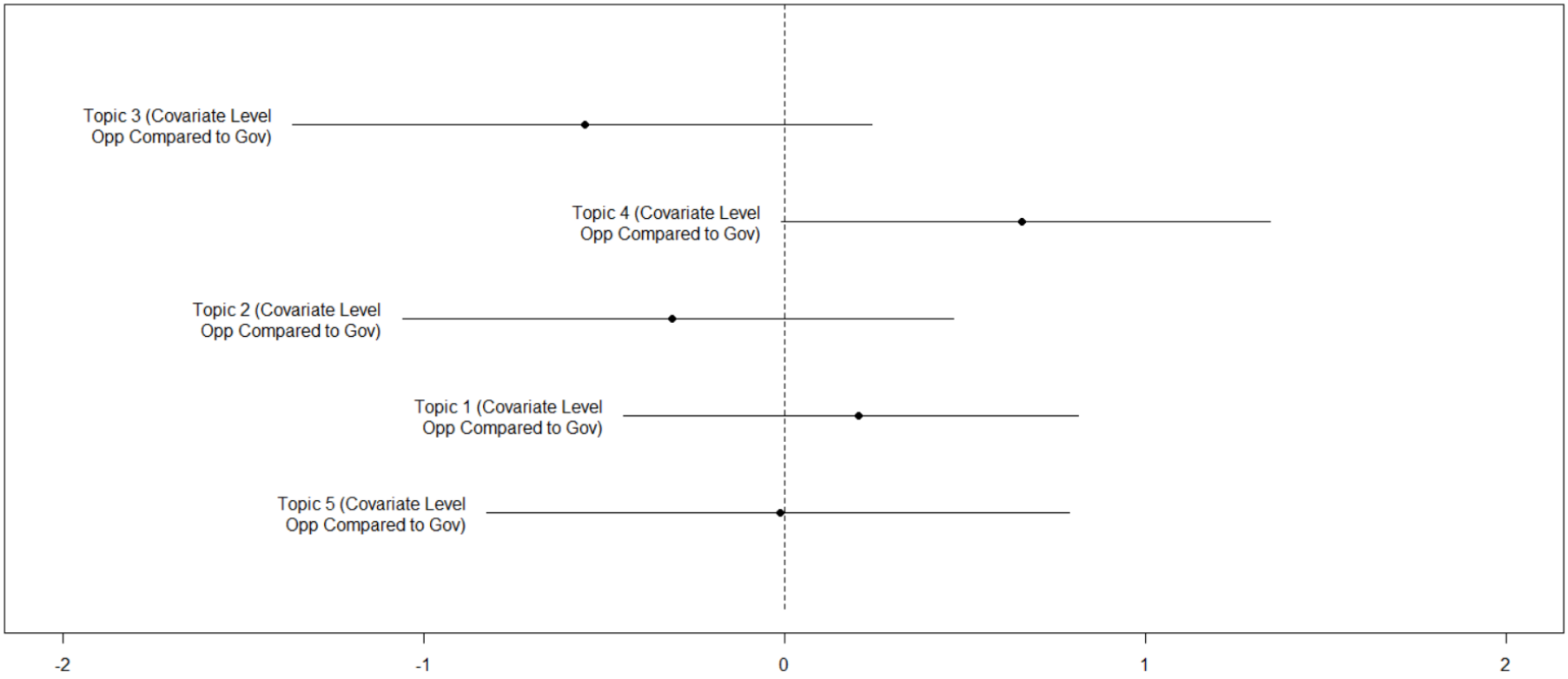
Classification methods

- for example, do documents of **left parties (or opposition politicians)** discuss more about a given topic than documents of right parties?

Classification methods



Classification methods



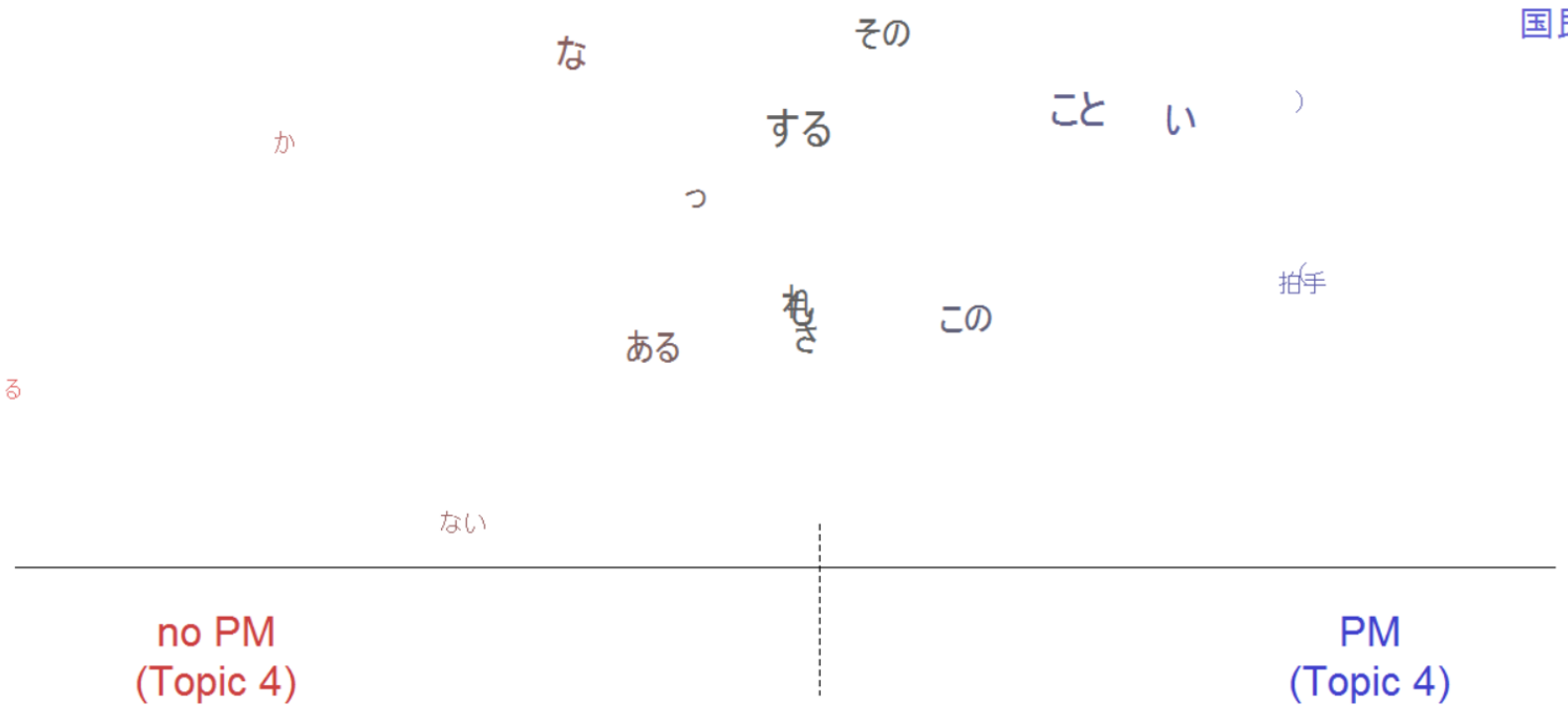
Reported coefficient:
«opposition – government»

Classification methods

Moreover, we can obtain measures of how the **language** used to discuss the same topic (content)

- for example, when **men** politicians discuss about a particular topic do they use the same words than **female** politicians?

Classification methods



Classification methods

In the STM framework, the researcher has therefore the option to choose covariates to incorporate in the model

These covariates inform either the **topic prevalence** or the **topical content** latent variables with observed information about the respondent

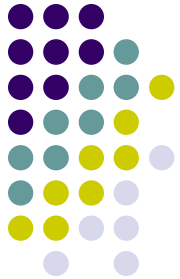
Classification methods

The analyst will want to include a covariate in the topical prevalence portion of the model when she believes that the observed covariate will affect *how much* the respondent is to discuss a particular topic

The analyst also has the option to include a covariate in the topical content portion of the model when she believes that the observed covariate will affect *the words which a respondent uses* to discuss a particular topic

These two sets of covariates can overlap, suggesting that the topic proportion and the way the topic is discussed change with particular covariate values

Classification methods

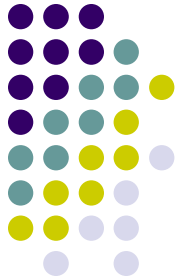


Note the difference with respect to a Topic Model approach

You can still try to control if there is there any relationship between for example the ideology of the writer of a document and the emphasis/salience she devotes in her document(s) towards a particular topic (for example, a topic about social welfare or migrants)

But you can do it…

Classification methods



…only after you have ended to run the Topic Model and obtained the $\theta s, \theta s$ that have been moreover generated by assuming that they arise from the same data-generating process (irrespective of any document-level variables of interest)

On the contrary, STM by relaxing this assumption, allows you to estimate **simultaneously** the θs and the possible impact on their variance across documents of some document-level variables of interest!

Classification methods

The quantities of interest from a **Structural** Topic Model (beyond θ and β_k of any Topic Model)

QOI: Topical Prevalence Covariate Effects

- Level of Analysis: Corpus
- Part of the Model: θ, X
- Description: Degree of association between a document covariate X and the average proportion of a document discussing each topic.
- Example Finding: Subjects receiving the treatment on average devote twice as many words to Topic 2 as control subjects.

Classification methods

The quantities of interest from a **Structural** Topic Model (beyond θ and β_k of any Topic Model)

QOI: Topical Content Covariate Effects

- Level of Analysis: Corpus
- Part of the Model: κ, U
- Description: Degree of association between a document covariate U and the rate of word use within a particular topic.
- Example Finding: Subjects receiving the treatment are twice as likely to use the word "worry" when writing on the immigration topic as control subjects.

STM and R

install.packages("stm", repos='http://cran.us.r-project.org')

install.packages("igraph", repos='http://cran.us.r-project.org')

install.packages("servr", repos='http://cran.us.r-project.org')

devtools::install_github("mroberts/stmBrowser",dependencies=TRUE)