Political Analysis of Social Media Data

Topic Models

Instructor:	Gregory Eady
Office:	18.2.10
Office hours:	Fridays 13-15

Today

- Topic models
- Video lectures & exercises

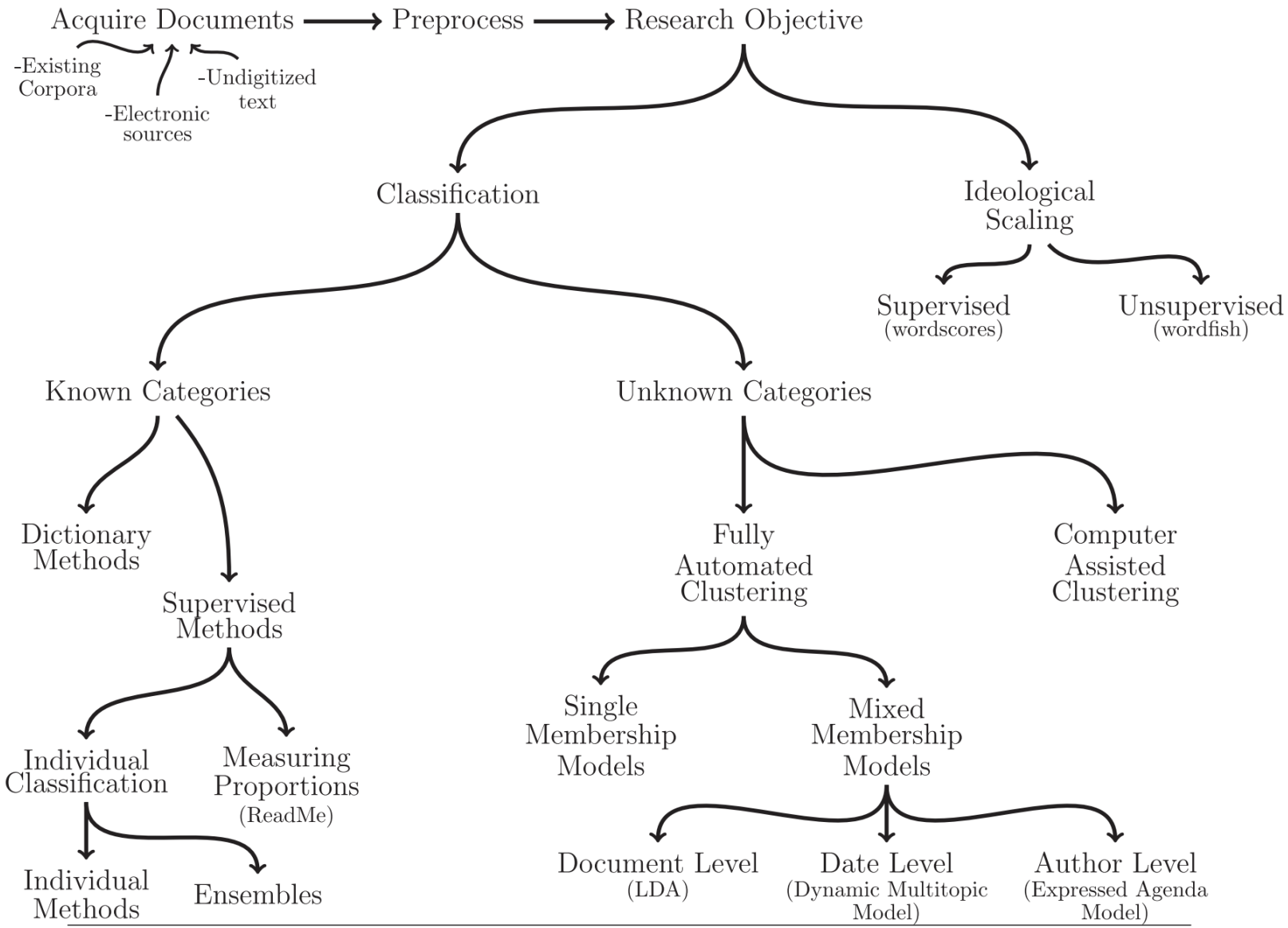


Fig. 1 An overview of text as data methods.

Recall text-as-data pre-processing choices:

1. **Punctuation**: Spaces & special characters (e.g. $, %, &)
2. **Numbers**: Sometimes informative (e.g. Section 423 of the U.S. Code); other times not
3. **Lowercasing**: Sometimes informative (e.g. "Trump" the president, versus "trump" the verb)
4. **Stopwords**: Common function words, e.g. "the," "and", "it," and "she," or domain-specific ones "congress"

Recall text-as-data pre-processing choices:

5. **Stemming**: Reducing a word to its root form
    - e.g. "party," "partying," and "parties" all share a common stem "parti"
6. **n-Grams**: treat multiple words as single "tokens". As bi-grams (2) or tri-grams (3), or more
    - e.g. "national" means something much different when combined with "debt" or "defense", ("national defense" versus "national debt")
7. **Infrequently used terms**: Remove very infrequent or frequent terms (e.g. remove words that occur in fewer than 0.5-1% of documents)

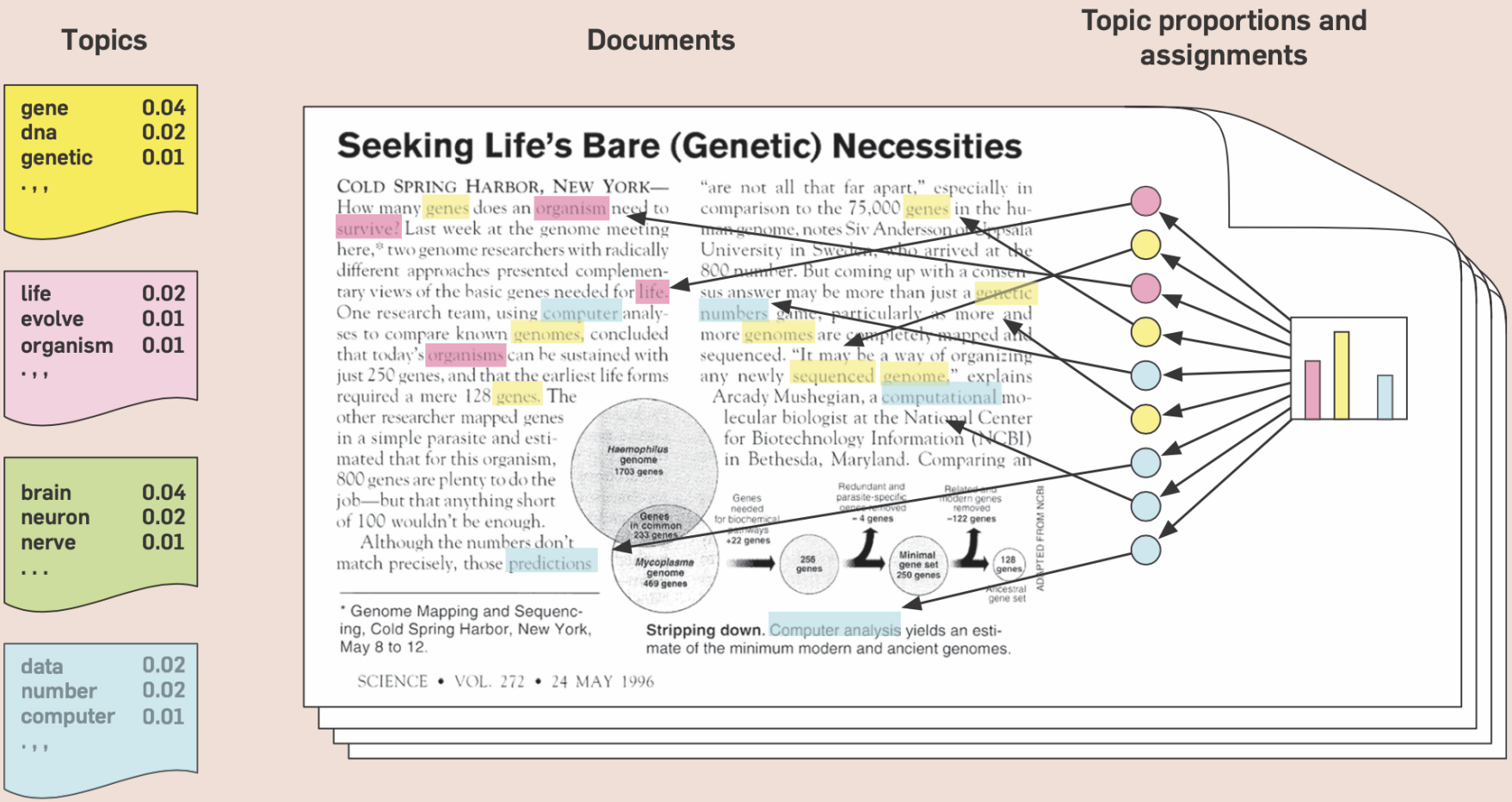
Topic models:

- An unsupervised model for discovering the latent topics / themes in a set of documents
- "Unsupervised" because we don't have any labels for the topics of any documents
- Thus we only need the text of the documents themselves i.e. no human annotators or pre-existing labels (similar to unsupervised models for ideology)

Topic models:

- Classical topic model is called a Latent Dirichlet Allocation (LDA) model
  - "latent" because the topics are unobservable (i.e. unlabeled)
  - "dirichlet" because the model relies on a Dirichlet distribution
- Is a "generative model" in the sense that we posit a simple process by which documents are created, and set up a model to capture that process
- M documents, where documents are distributions over topics.
- K topics, where topics are distributions over words.

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



How do LDA models assume documents are generated?

- Are K topics. Are M documents. Are N words.
- Choose $\gamma_m \sim Dirichlet(\alpha_\gamma)$
  - These are just the distribution of topics in each document m (e.g. $\gamma_{m=1} = [0.25, 0.7, 0.05]$)
- Choose $\beta_k \sim Dirichlet(\alpha_\beta)$
  - Distribution of words in a topic k (e.g. $\gamma_{k=1} = [$"refugee" $= 0.15$, "migrant" $= 0.1$, "taxes" $= 0$, ...])
- Now, we'll fill up each document with words...
  - Pick a topic from the document's topic distribution: $z_i \sim Multinomial(\gamma_m)$
  - Pick a word from that topic's word distribution: $w_i \sim Multinomial(\beta_{k=z_i})$

What parameters do we actually care about?

1. γ: a $M \times K$ matrix where columns are the proportions of each topic in each document:

	Topic 1	Topic 2	Topic 3	...	Topic K
Document 1	0.02	0.00	0.10	...	0.45
Document 2	0.00	0.03	0.90	...	0.05
Document 3	0.20	0.01	0.01	...	0.4
...	...	...	...	...	...
Document M	0.30	0.30	0.30	...	0

2. β: a $K \times N$ matrix where columns are the proportions of each word in each topic:

	Word 1	Word 2	Word 3	...	Word N
Topic 1	0.002	0.001	0.000	...	0.009
Topic 2	0.070	0.000	0.004	...	0.002
Topic 3	0.002	0.004	0.001	...	0.011
...	...	...	...	...	...
Topic K	0.001	0.006	0.021	...	0.000

What does the output of topic models look like, and how do I know what a topic means?

- The output are the parameters γ and β (and some incidental parameters like α_γ and α_β)
- But how might we understand those parameters substantively?
- The γ tell you the topics of each document
- The β tell you what words dominate each topic, and by looking at these words *qualitatively*, you can determine an appropriate label for each topic

To understand each topic, look at the distribution of β_i

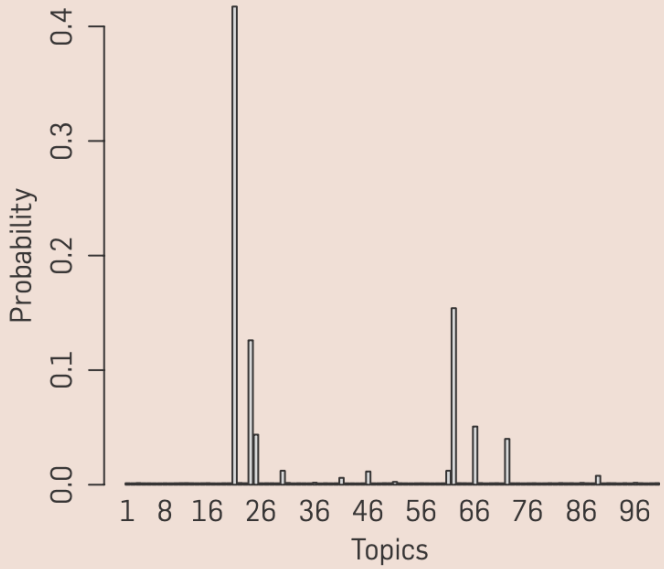
- N is often large (vocabularies are big), so in practice we look at the words with the most weight (the largest βs)

TABLE 3 Topic Keywords for 42-Topic Model

Topic (Short Label)	Keys
1. Judicial Nominations	*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*
2. Constitutional	*case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut*
3. Campaign Finance	*campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit*
4. Abortion	*procedur, abort, babi, thi, life, doctor, human, ban, decis, or*
5. Crime 1 [Violent]	*enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil*
6. Child Protection	*gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school*
7. Health 1 [Medical]	*diseas, cancer, research, health, prevent, patient, treatment, devic, food*
8. Social Welfare	*care, health, act, home, hospit, support, children, educ, student, nurs*
9. Education	*school, teacher, educ, student, children, test, local, learn, district, class*
10. Military 1 [Manpower]	*veteran, va, forc, militari, care, reserv, serv, men, guard, member*
11. Military 2 [Infrastructure]	*appropri, defens, forc, report, request, confer, guard, depart, fund, project*
12. Intelligence	*intellig, homeland, commiss, depart, agenc, director, secur, base, defens*
13. Crime 2 [Federal]	*act, inform, enforc, record, law, court, section, crimin, internet, investig*
14. Environment 1 [Public Lands]	*land, water, park, act, river, natur, wildlif, area, conserv, forest*
15. Commercial Infrastructure	*small, busi, act, highwai, transport, internet, loan, credit, local, capit*

To understand each topic, look at the distribution of γ_m

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Other topic models

- Correlated topic models (Blei and Lafferty 2007)
- Dynamic topic models (Quinn et al. 2010)
- Hierarchical topic models (Grimmer 2010)
- Structural topic models (Roberts et al. 2014)
- Keyword topic models (Eshima et al. Forthcoming)
- BERT topic models (Grootendorst 2022) (an large-language model for topic classification)