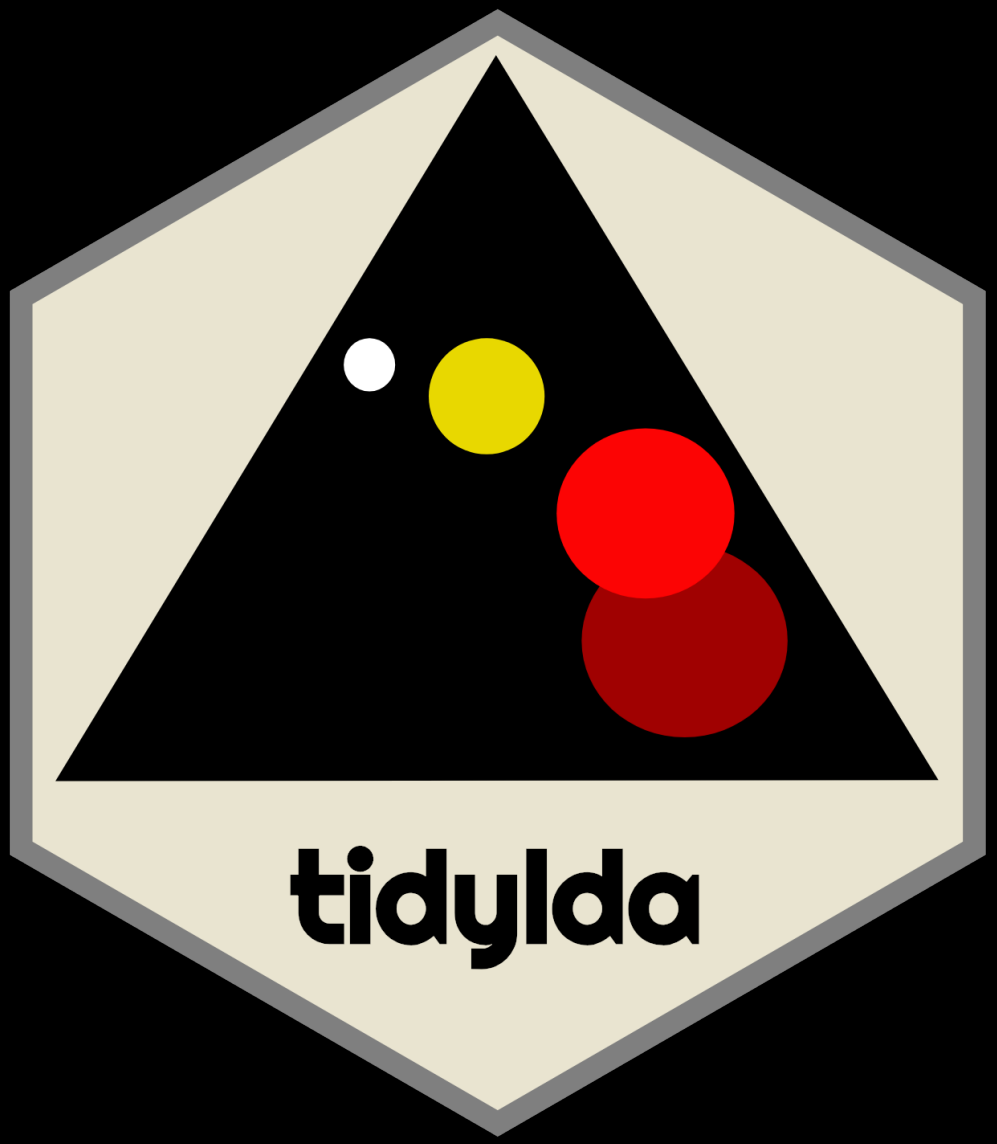
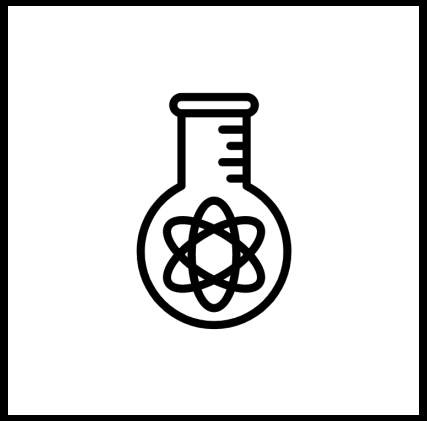


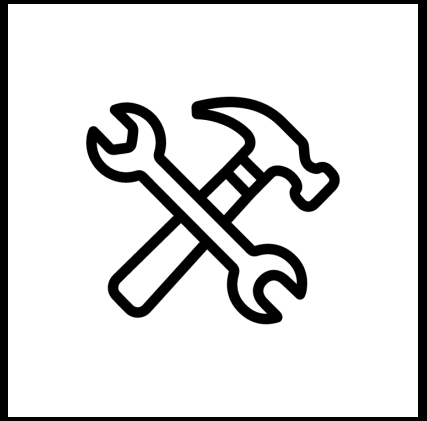
Tommy Jones

Data Science DC
June 14, 2022

How to Take Over the World



Science!

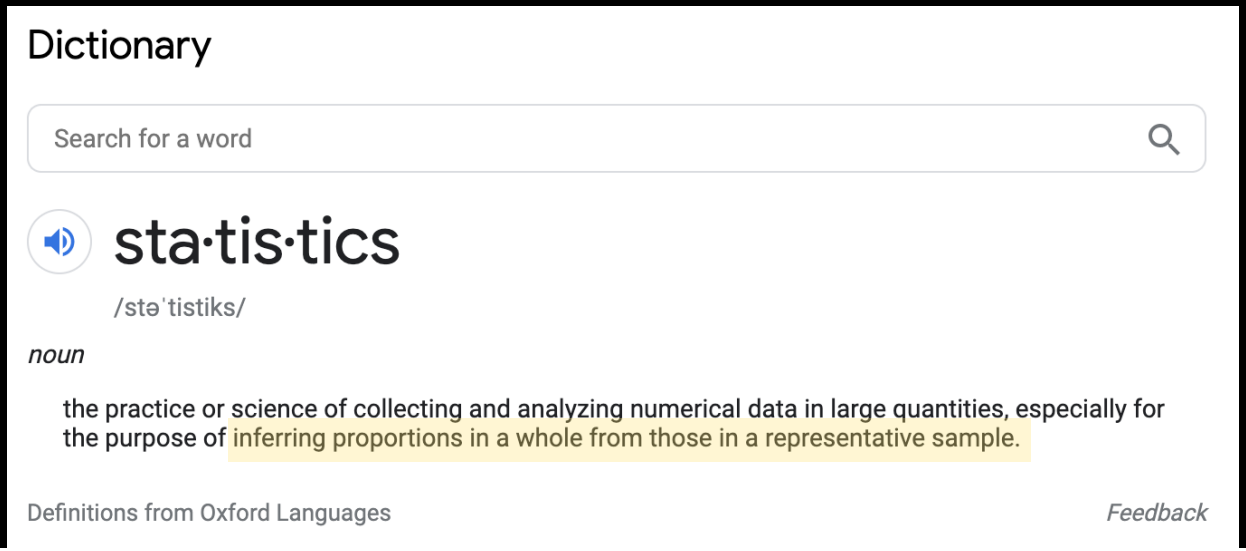


Tools!

Science = “Corpus Statistics”

Task-based
vs.
Inference on populations

- Sample composition
- Experimental design
- Model (mis)specification
- Estimate population parameters
- Uncertainty quantification
- etc…



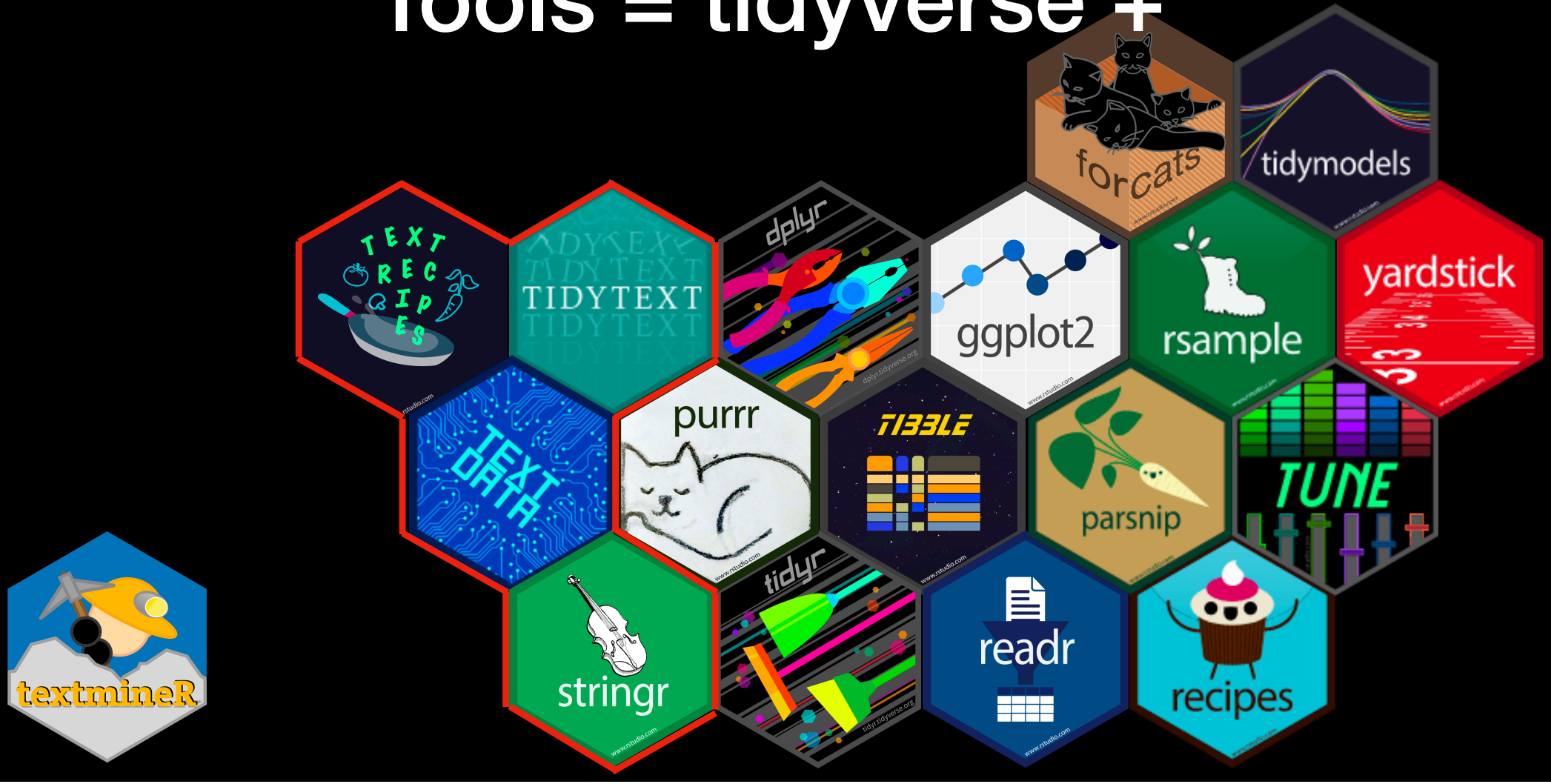
Latent Dirichlet Allocation for Corpus Statistics

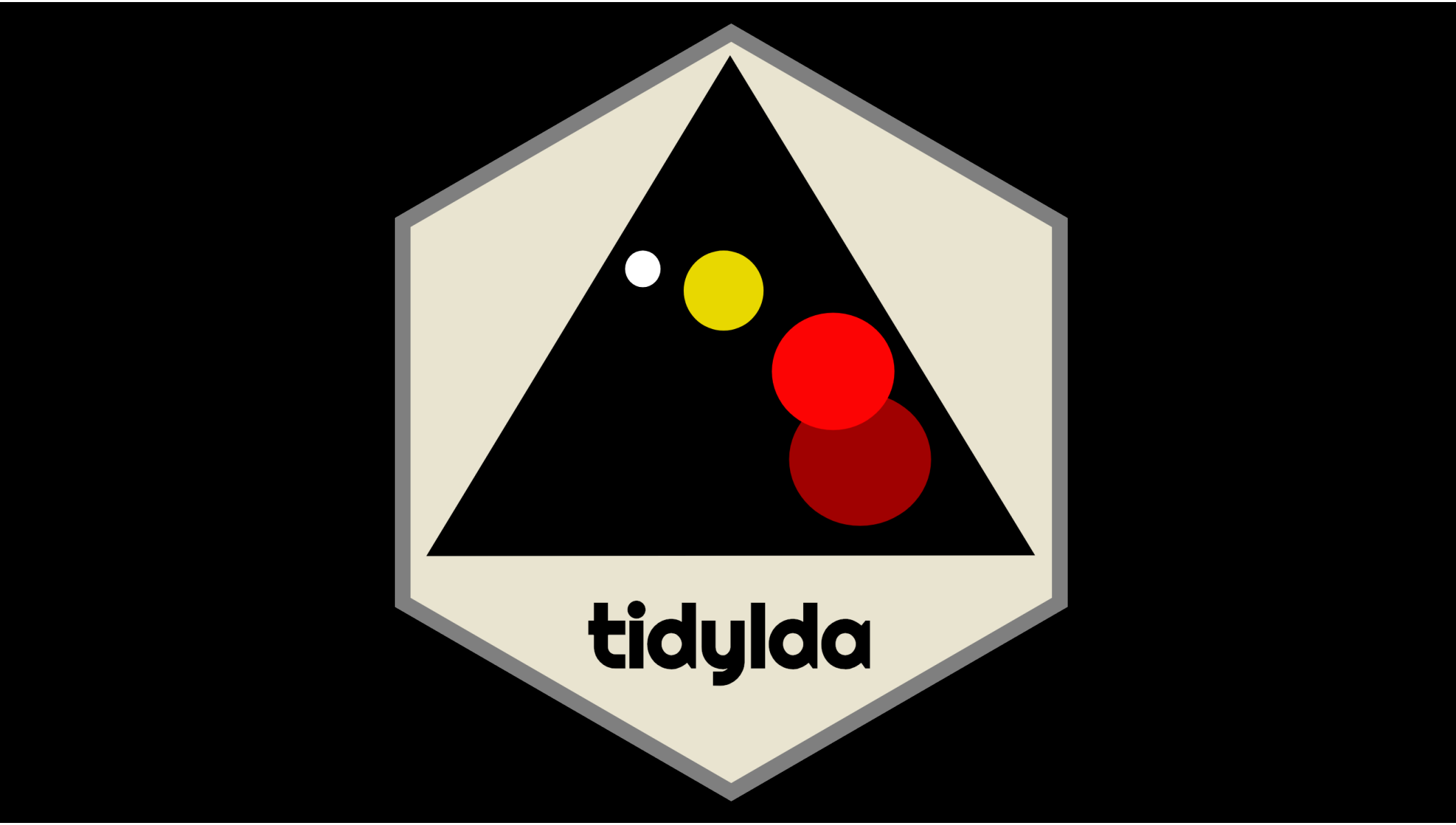
- **LDA embeds text into a probability space**
  Well-defined and interpretable relationships

- **LDA is a Bayesian parametric model**
  Uncertainty quantification, diagnostics, etc. in line with established statistical practice

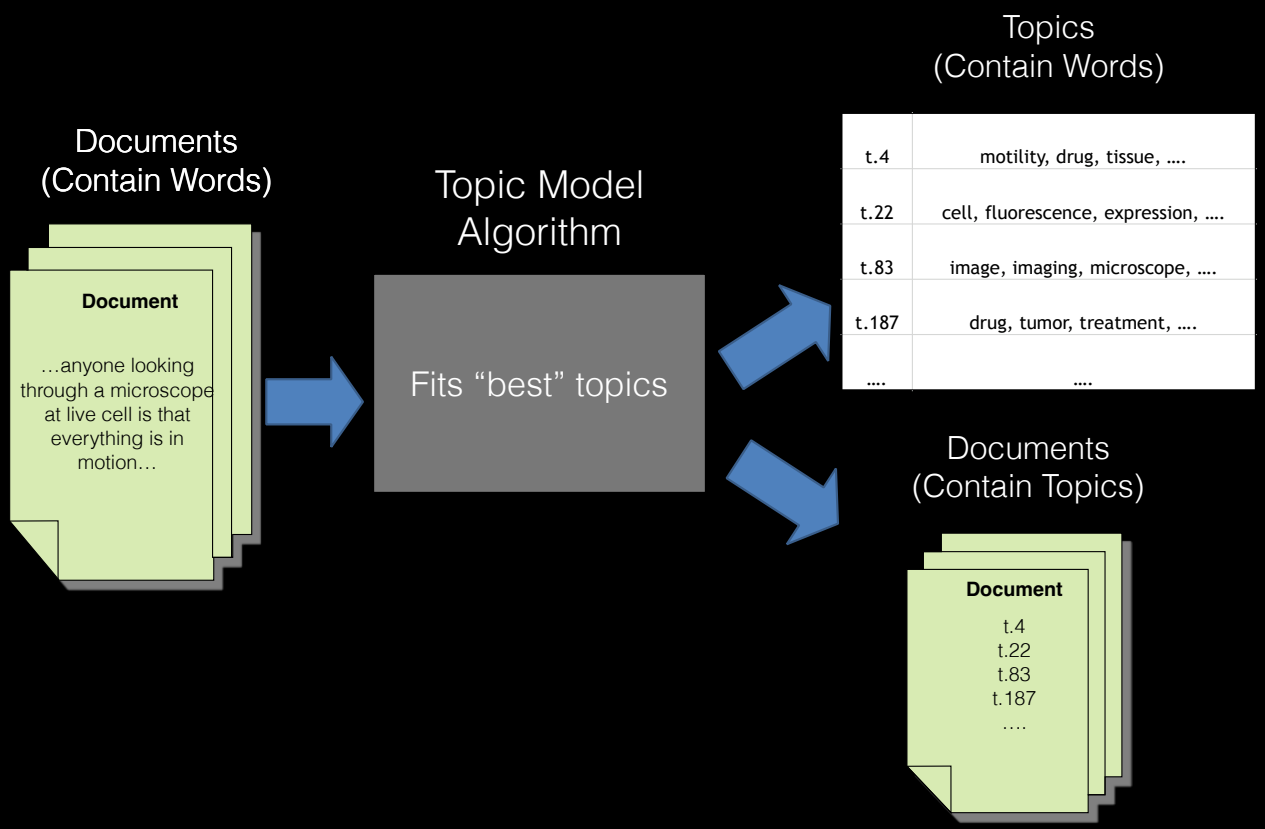
- **LDA is a generative model of language**
  Study the LDA-DGP to build better models and diagnose pathological misspecification

Tools = tidyverse +

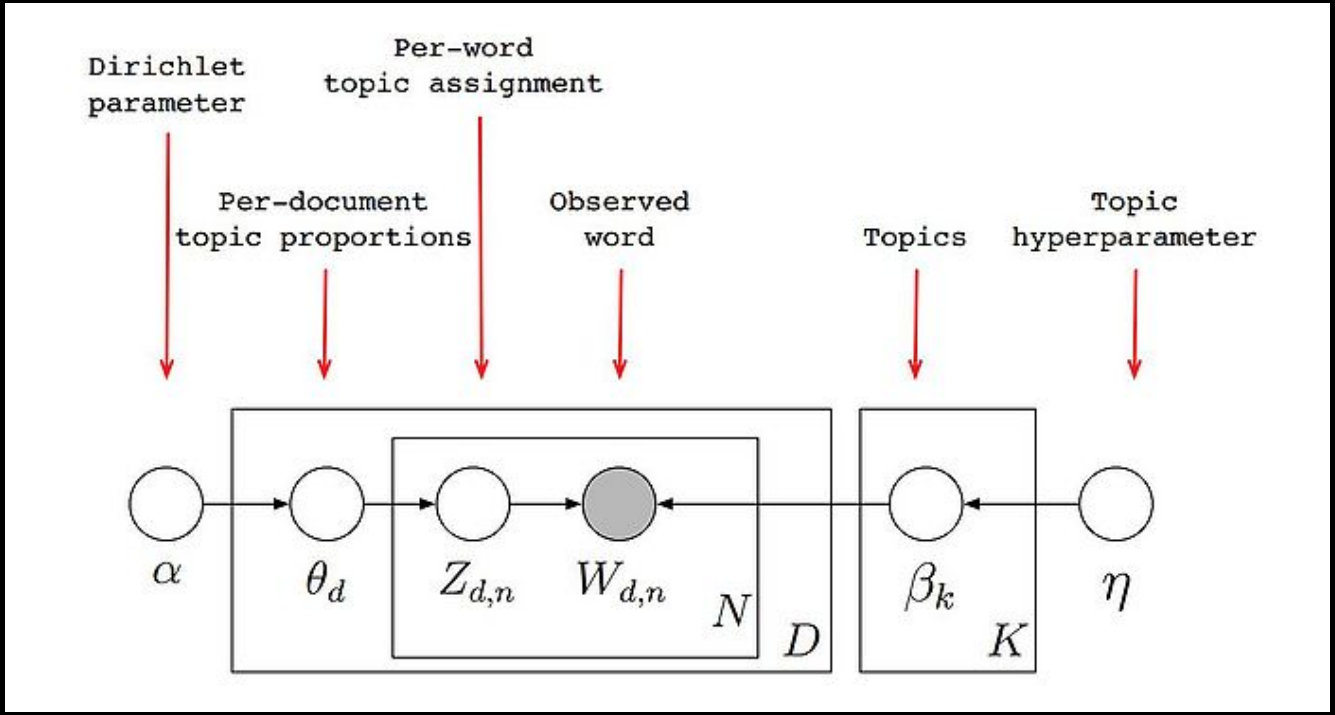



tidylda

Quick Review of LDA I



Quick Review of LDA II



- η, α: priors
- β, θ: parameters of interest
- w: observed word frequencies
- z: latent topic frequencies

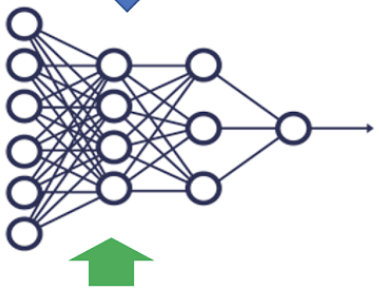
Demo Time

A Paradigm Shift in NLP

Very large corpus of unlabeled text

Self-supervised learning task

Pre-trained language model

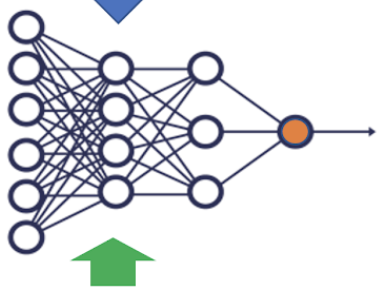


Self-supervision = more data

	text	labels
1	Argentina rose as the successor state of the Viceroyal…	x
2	It is also a founding member of the Mercosur block, h…	x
3	Some of the largest sea ports are La Plata-Ensenada, …	x
4	Argentina has its own satellite programme, nuclear p…	x
5	The public universities of Buenos Aires, Córdoba, La P…	x
6	The valleys have a climate that is similar to the tempe…	y
7	There are, on average, 300 days per year of sunshine.	y
8	Average daily peak insolation varies from 1150 W/m2…	y
9	The driest months tend to be January and February, a…	x

Small corpus of labeled text

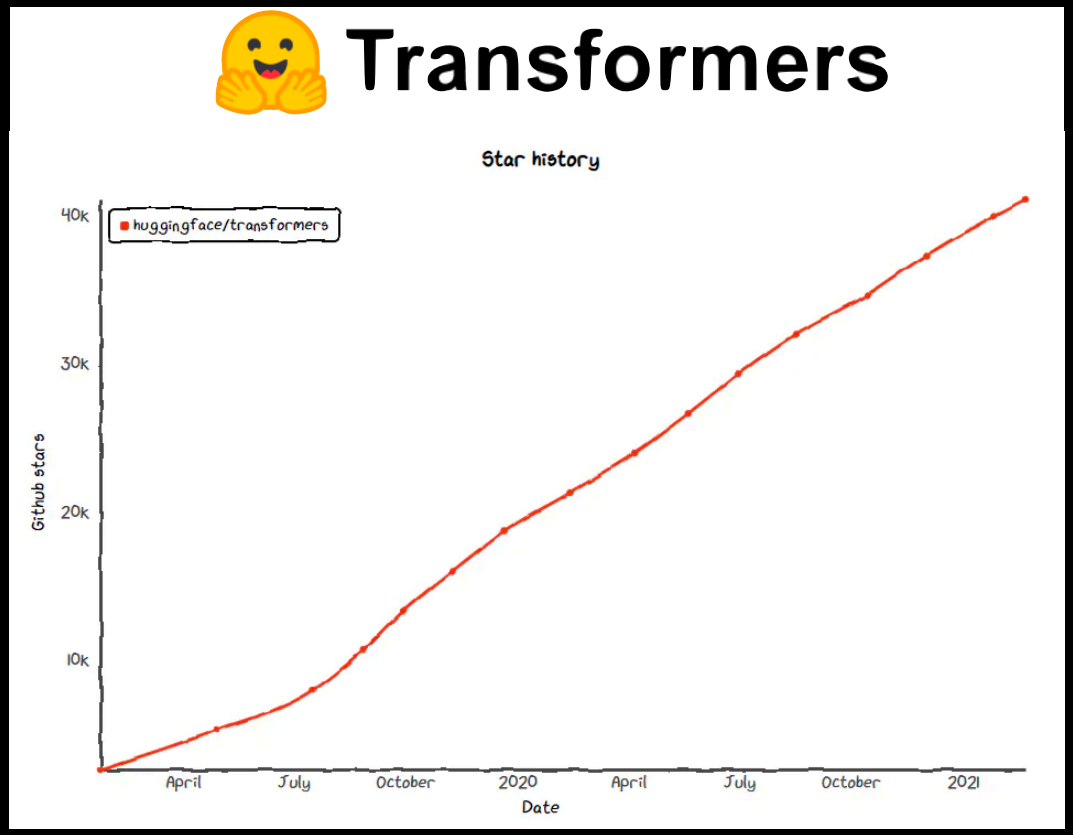
Fully supervised learning task



Fine-tuned, task specific language model

Pre-training = need less labeled data for a better model

Reign of the Pre-Trained Language Model



Fine Tuning LDA for Transfer Learning

$$\overrightarrow{\beta}_k^{(t)} \sim \textbf{Dirichlet}\left(Cv_k^{(t)} + a^{(t)} \cdot \omega_k^{(t)} \cdot \mathbb{E}\left[\overrightarrow{\beta}_k^{(t-1)} \right] \right)$$

Posterior | **Data** | **Weight** | **Base Posterior**

Prior

Demo Time

When does tLDA converge?

- tLDA has one tuning parameter: $a^{(t)}$
- Do certain values of $a^{(t)}$ allow/not allow tLDA to converge at all?
- Do certain values of $a^{(t)}$ allow/not allow tLDA to converge towards the correct distribution?

Revisiting the key equation

$$\overrightarrow{\beta}_k^{(t)} \sim \textbf{Dirichlet}\left(Cv_k^{(t)} + a^{(t)} \cdot \omega_k^{(t)} \cdot \mathbb{E}\left[\overrightarrow{\beta}_k^{(t-1)}\right]\right)$$

Posterior | **Data** | **Weight** | **Base Posterior**

Prior

Simulation Experiment: Setup

- 128 data generating distributions
- 100 corpora of 10,000 documents sampled from each
- For each corpus:
  - Train a model on the first 100 documents
  - Iteratively add 100 documents to update the model
- Do the above for a in {0.2, 0.4, …, 1, 1.2, 1.4, …, 2}
- Does the model converge? Does it converge towards the data generating distribution?

Simulation Experiment: Results I

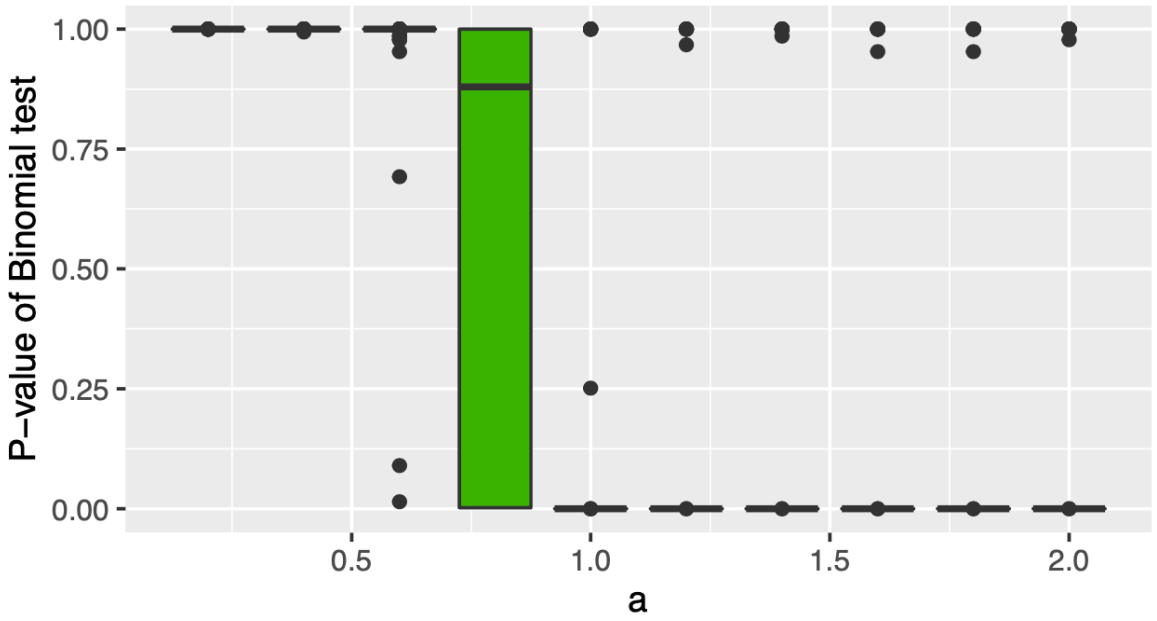


Figure 6: Boxplot of the p-values assessing convergence for each value of a. Below $a = 0.8$ almost all p-values are near one. Above $a = 0.8$ almost all p-values are near zero, with a few exceptions. But at $a = 0.8$ there is high variance in the p-values assessing model convergence.

Simulation Experiment: Results II

Table 1: Effects of a on convergence

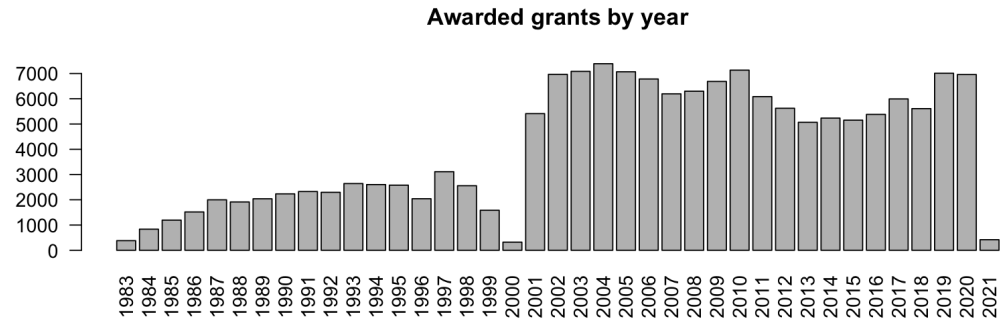
Variable	Convergence		Convergence		Direction		Direction	
Intercept	-6.01	***	-17.09	***	64.59	***	65.19	***
a	5.92	***	28.38	***	1.94	***	8.62	***
Avg. Doc. Length	0.00		-0.01	**	0.01	***	0.02	***
$\sum(\boldsymbol\eta)$	0.00		0.00	**	-0.02	***	-0.06	***
$\sum(\boldsymbol\alpha)$	0.32	***	0.40		-0.18	***	0.01	
a^2			-9.20	***			-2.28	***
$(\text{Avg. Doc. Length})^2$			0.00	**			0.00	***
$\sum(\boldsymbol\eta)^2$			0.00	**			0.00	***
$\sum(\boldsymbol\alpha)^2$			0.02				-0.02	**

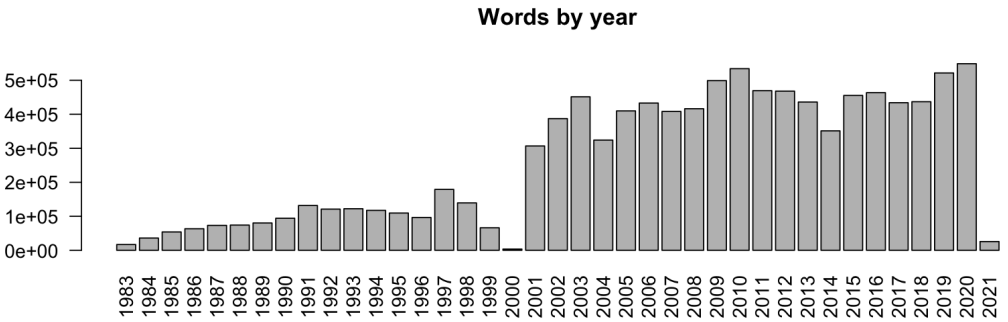
The Data

Small Business Innovation Research (SBIR) Program

Technology R&D grants from US Small Business Admin.

159,734 abstracts of awarded grants from 1983 - 2021





SBIR Time Series

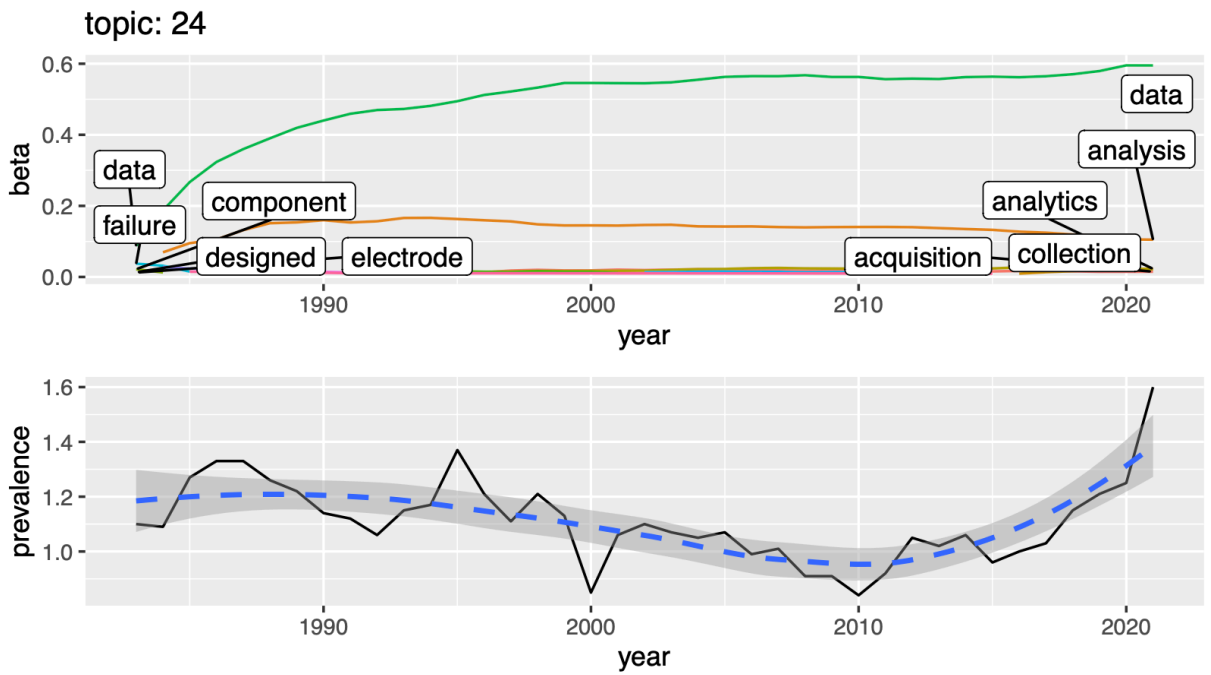


Figure 10: Each topic is a time series of its prevalence and distribution of its words. The top depicts changes in the top 5 words for topic 24. The bottom shows changes in prevalence for topic 24 with actual data in black and a blue loess curve with corresponding shaded confidence region.

SBIR Results

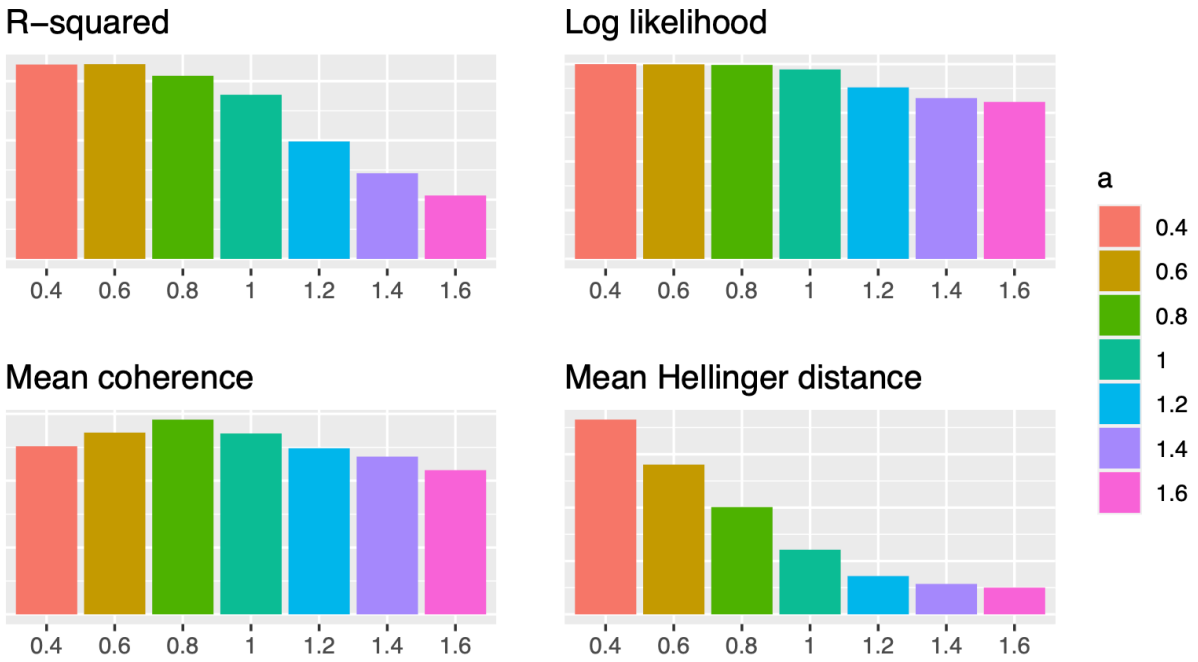
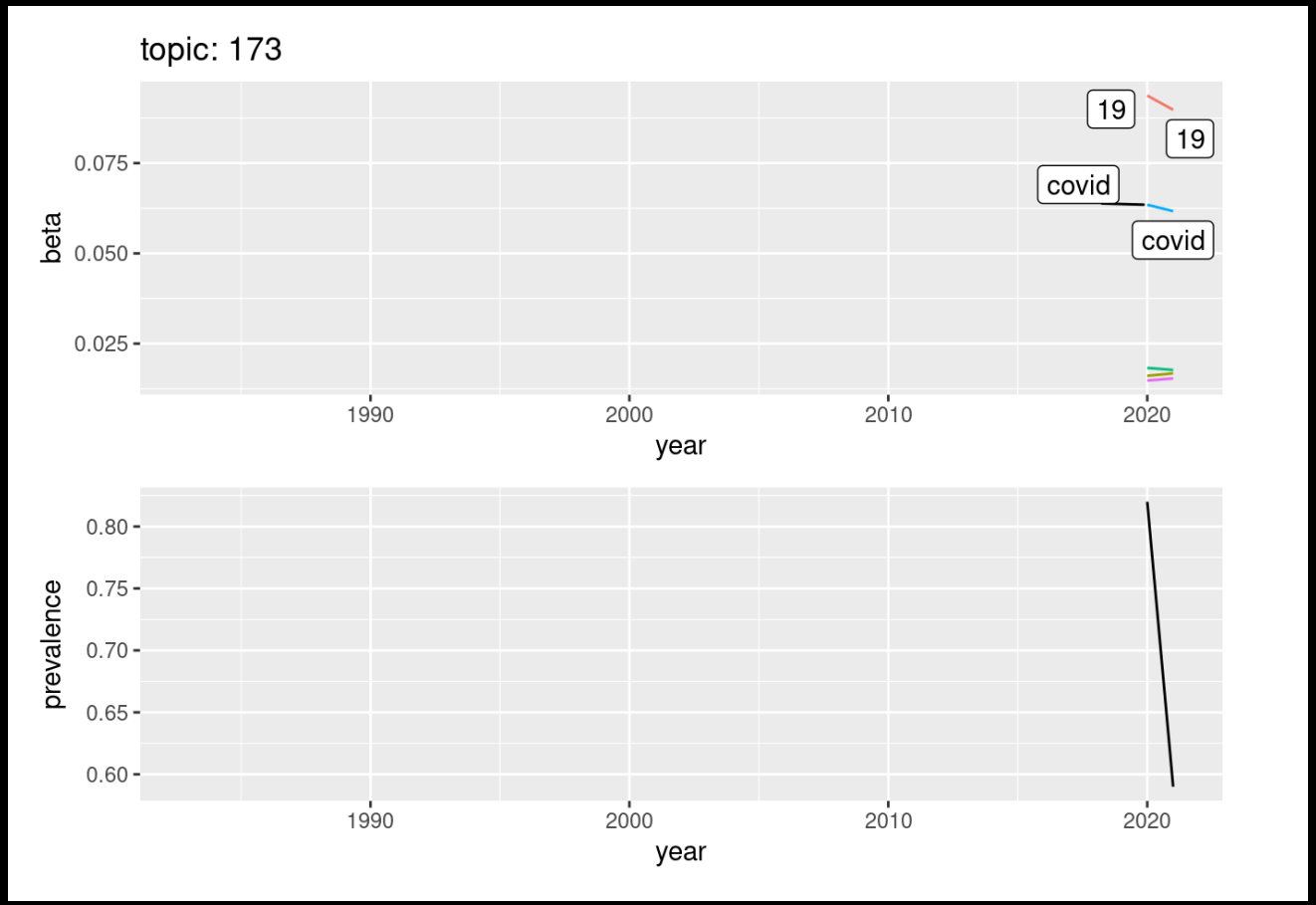


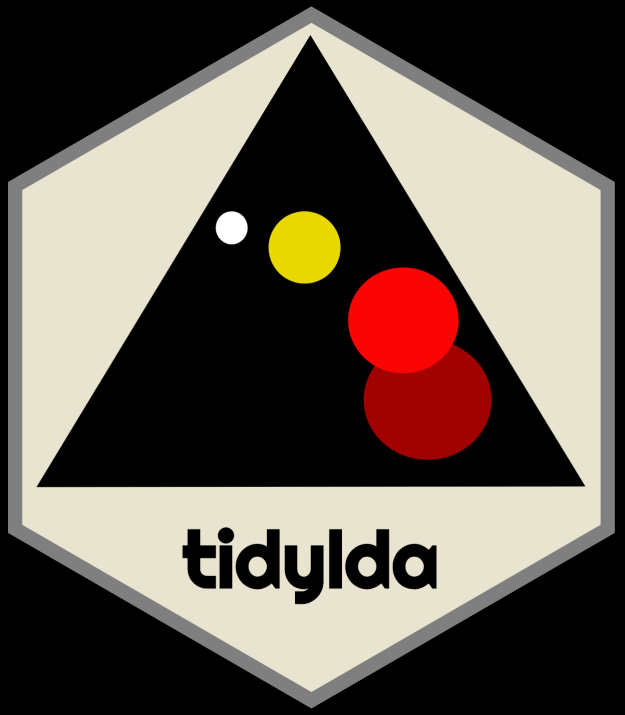
Figure 8: Averaging over all years, the patterns for each statistic are clearer. R^2 and log likelihood are decreasing functions of a, but for values less than $a = 1$, differences are very small. Coherence reaches a small peak at $a = 0.8$. Mean Hellinger distance is also a decreasing function of a as expected.

SBIR Time Series II



Next Steps for tidylda

- Speed and scaleability
- UX for seeding topics
- Helper functions for common tasks
- Fold tidylda into textmineR



Thank You!

jones.thos.w@gmail.com

iqt tojones@iqt.org

@thos_jones

https://github.com/tommyjones