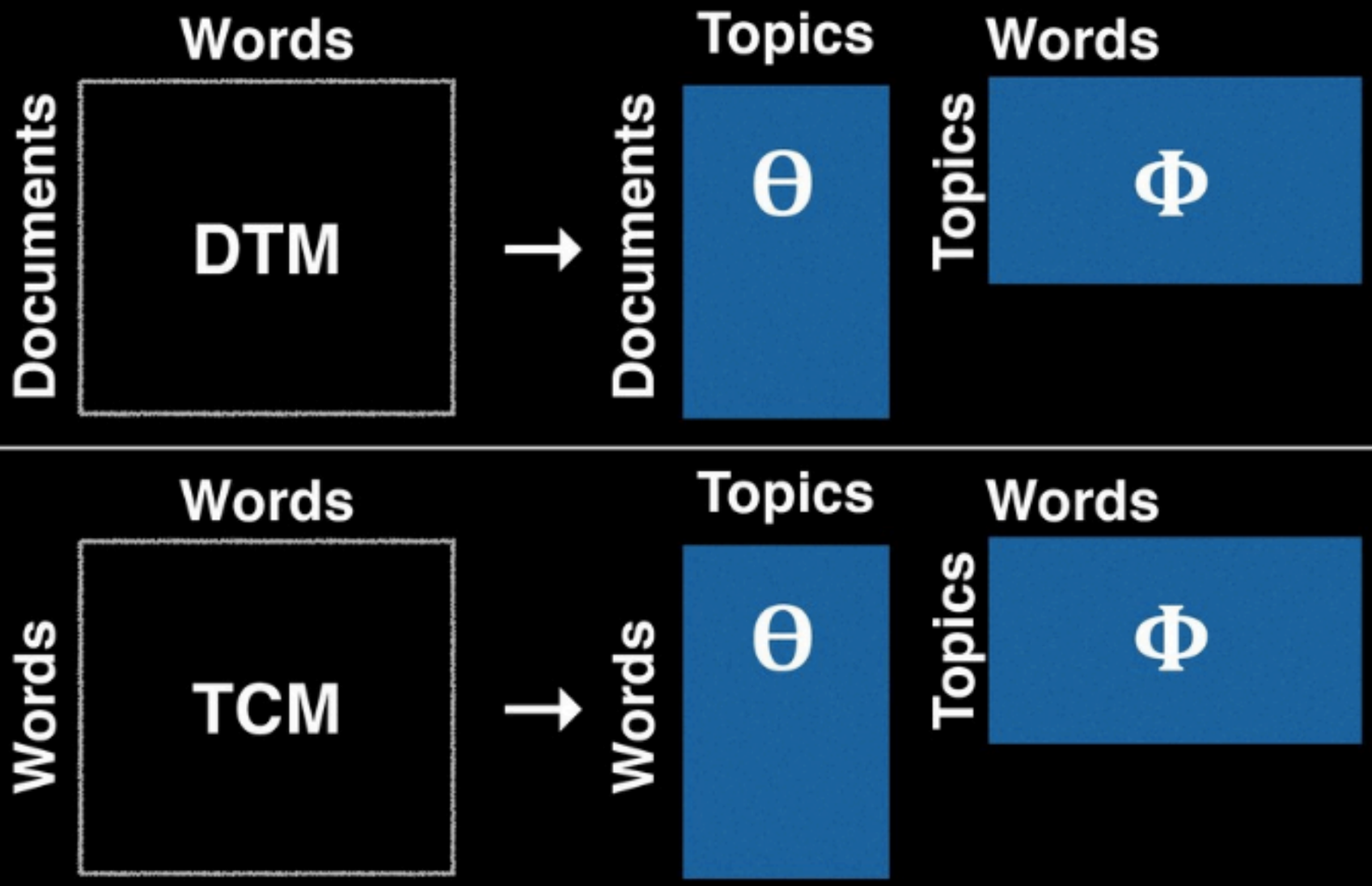
Project Update

Tommy Jones
27 March 2019

Agenda

- Background
- What I was going to do
- What’s got me concerned
- Pivot

Background

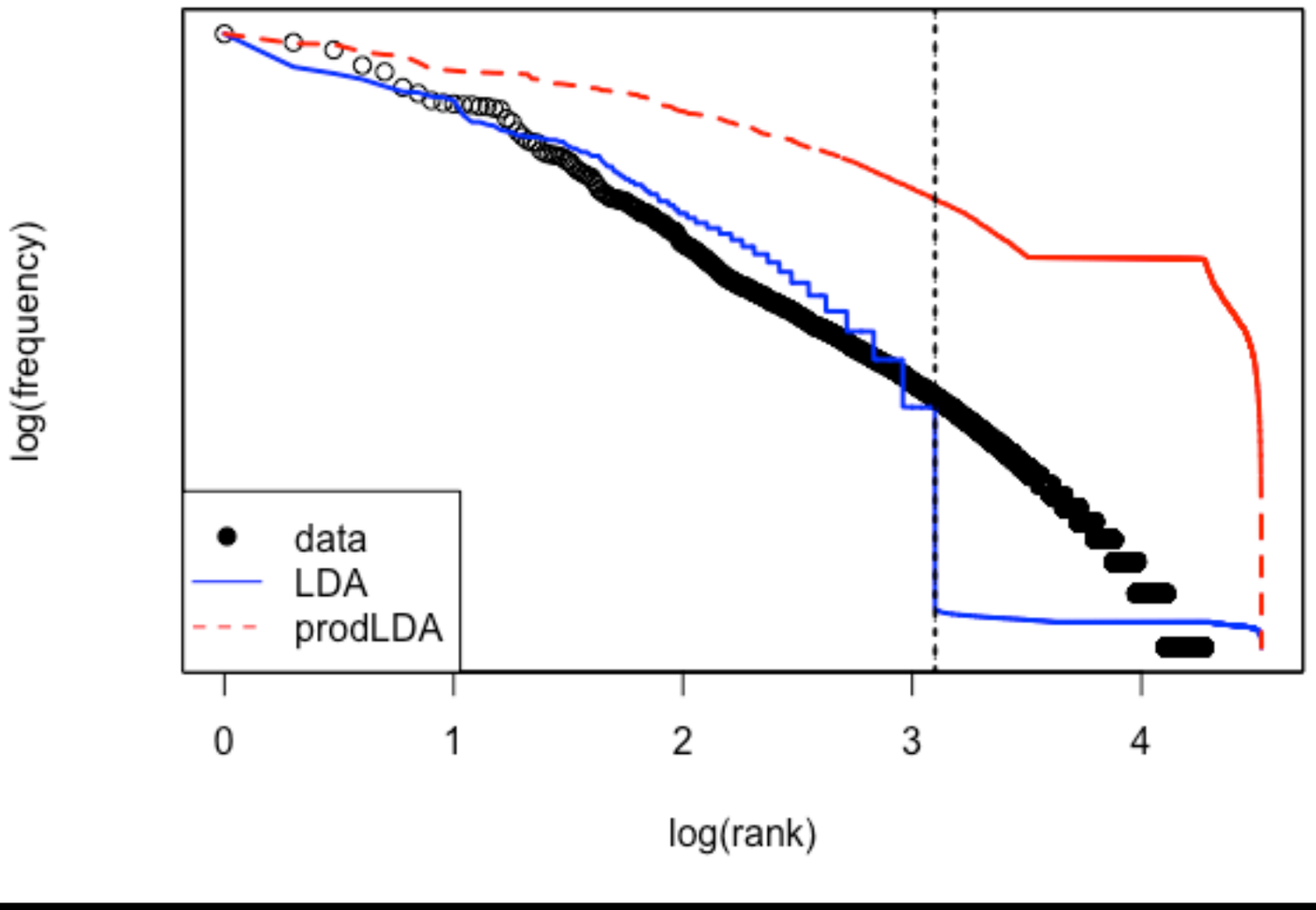

Words
Documents
DTM
Topics
Documents
θ
Words
Topics
Φ
Words
Words
TCM
Topics
Words
θ
Words
Topics
Φ

What I was going to do

- Transfer parameters from prodLDA to LDA
- 20 Newsgroups data set
- Success based on convergence speed and coherence

What’s got me concerned

- Mechanically it works
- The resulting topics are very different
- LDA has lower coherence but…
- R-squared for prodLDA is terrible
- How can we trust a model that doesn’t fit the data well?
- Means that coherence isn’t a be-all-end-all



	prodLDA	LDA
Median Coherence	0.14	0.10
Mean Coherence	0.20	0.14
R-Squared	-0.09	0.22

	prodLDA	LDA
t_1	si, ijs, lavrencic, borut, stefan	ca, game, apr, sport, rec
t_2	god, tartar, rose, began, tartars	time, people, good, back, make
t_3	roger, gant, hirschbeck, players, nhl	rutgers, christian, god, religion, cs
t_4	medical, cancer, clinical, gopher, bit	space, nasa, gov, sci, apr
t_5	ic, uka, imperial, bielefeld, ira	news, de, uiuc, cso, cs
t_6	gun, vancouver, homicide, rates, crime	talk, politics, misc, alt, guns
t_7	atheism, god, atheists, religious, belief	graphics, mail, pub, image, ftp
t_8	programmer, switch, security, privacy, mbeckman	cs, news, cmu, forsale, misc
t_9	geb, anthony, landreneau, gordon, ozonehole	cancer, medical, university, gun, research
t_10	terry, ersys, mjr, astros, cincinnati	comp, windows, cs, net, au

Pivot

- Turn this into an argument for why we need a better way to evaluate topic models/text embeddings?
- A “takedown” of the prodLDA paper would be easy
  - It would also be cheap and misleading…
- Not sure I have a clean story yet…