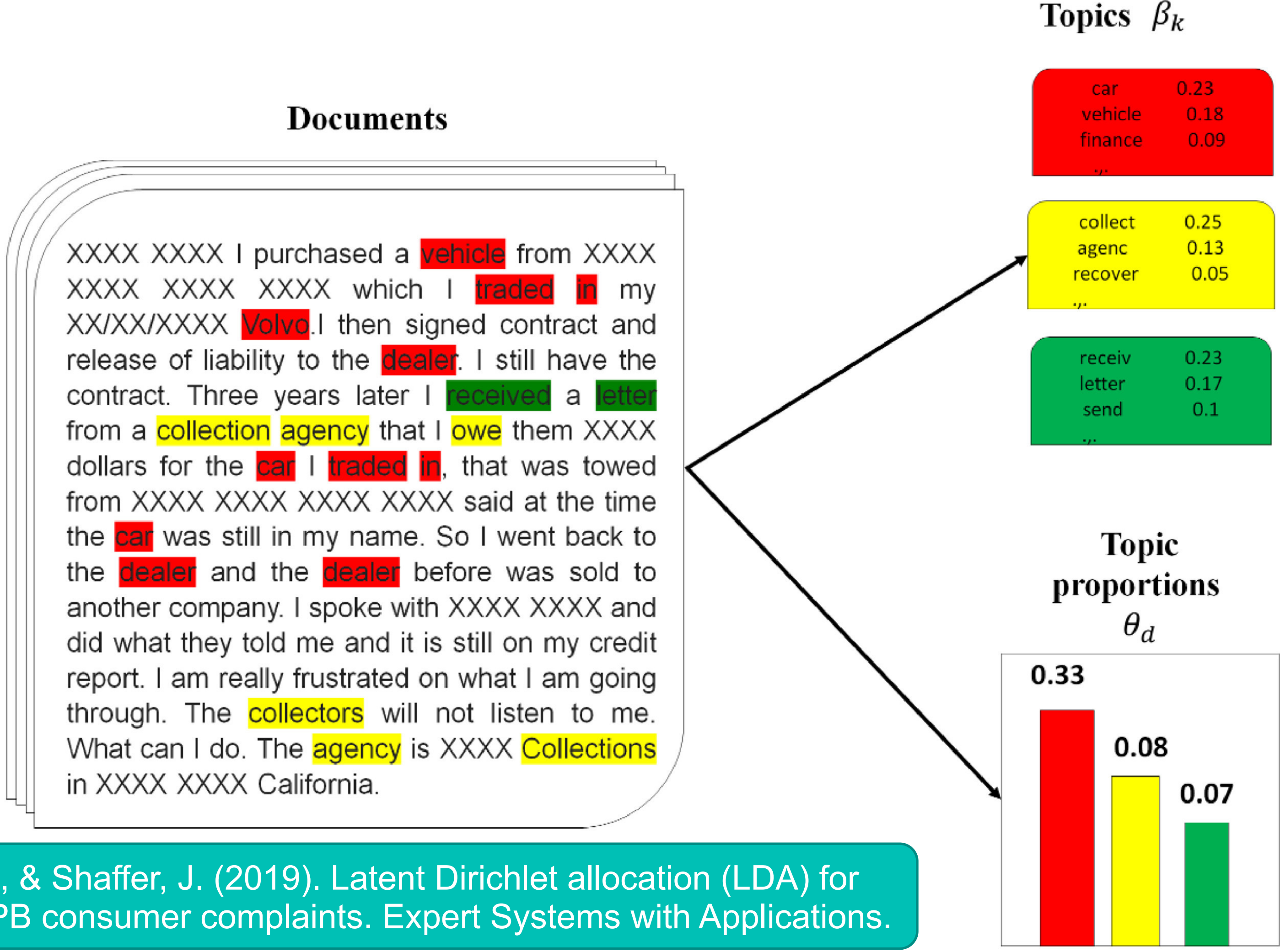
3.3 Applying Topic Models

Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

Example of LDA Complaint Topics



Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. Expert Systems with Applications.

Complaint Topics

From Bastani et al. (2019).

Extracted topics for the CFPB consumer complaints using LDA.

ID	Topics	Label
0	Payment late made due make appli month past day miss	Payment/late payment
1	Receiv letter sent send mail state request email document notic	Communication
2	Loan student borrow privat navient repay lender default defer forbear	Loan/Student Loan
3	Car vehicl financ dealership dealer ticket book drive trade truck	Auto Loan/Dealership
4	Servic custom repres manag transfer spoke depart supervisor cancel speak	Customer Service
5	Check cash advanc clear return wrote flagstar seiz present payabl	Check
6	File complaint cfpb case complain respond clerk district bsi compliant	CFPB
7	Home hous equiti repair inspect buy door damag sell valu	Home Equity
8	Call phone number person stop time answer messag harass work	Harassment
9	Credit report remov bureau show neg correct inform agenc transunion	Credit Reporting
10	Co program school class colleg signer enrol student region educ	Education

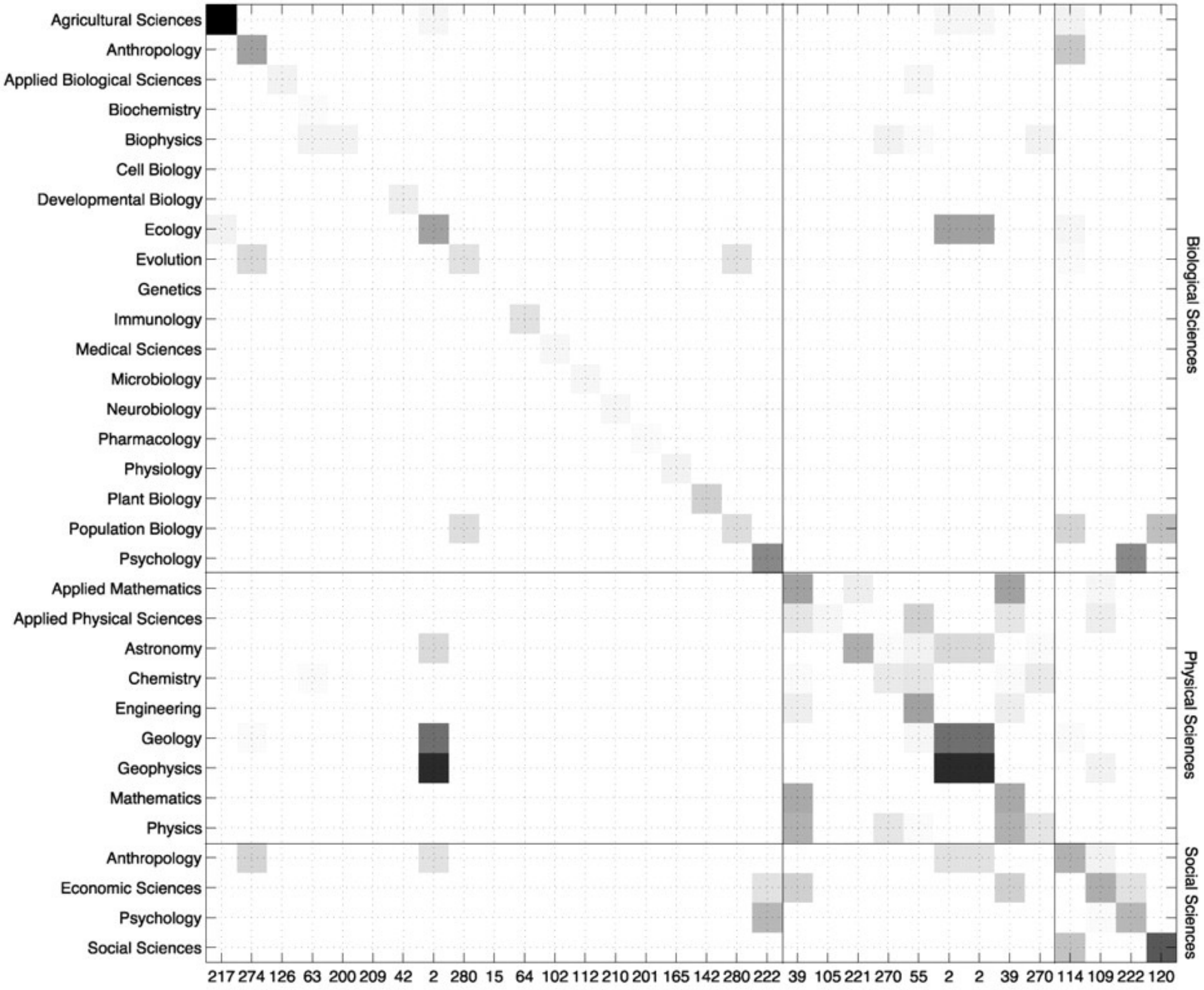
Topic Trends over Time

From Bastani et al. (2019).

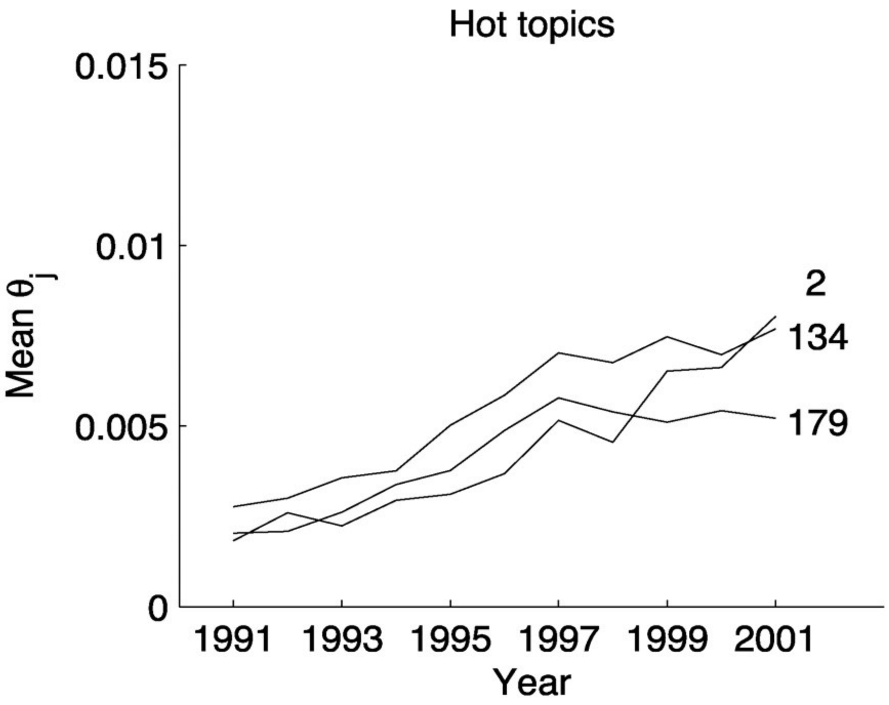
Scientific Abstracts

Mean values of θ for LDA topics for PNAS minor categories, computed from all abstracts published in 2001.

Griffiths, T. and Steyvers, M. Finding Scientific Topics (2004) PNAS.



Hottest topics from 1991 to 2001



2	134	179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

Griffiths, T. and Steyvers, M. Finding Scientific Topics (2004) PNAS.

Figure 2. Top words associated with ailments and topics.

Paul MJ, Dredze M (2014) Discovering Health Topics in Social Media Using Topic Models. PLOS ONE

Non-Ailment Topics						
TV & Movies	**Games & Sports**	**School**	**Conversation**	**Family**	**Transportation**	**Music**
watch	killing	ugh	ill	mom	home	voice
watching	play	class	ok	shes	car	hear
tv	game	school	haha	dad	drive	feelin
killing	playing	read	ha	says	walk	lil
movie	win	test	fine	hes	bus	night
seen	boys	doing	yeah	sister	driving	bit
movies	games	finish	thanks	tell	trip	music
mr	fight	reading	hey	mum	ride	listening
watched	lost	teacher	thats	brother	leave	listen
hi	team	write	xd	thinks	house	sound

Ailments						
	Influenza-like Illness	**Insomnia & Sleep Issues**	**Diet & Exercise**	**Cancer & Serious Illness**	**Injuries & Pain**	**Dental Health**
General Words	better	night	body	cancer	hurts	dentist
	hope	bed	pounds	help	knee	appointment
	ill	body	gym	pray	ankle	doctors
	soon	ill	weight	awareness	hurt	tooth
	feel	tired	lost	diagnosed	neck	teeth
	feeling	work	workout	prayers	ouch	appt
	day	day	lose	died	leg	wisdom
	flu	hours	days	family	arm	eye
	thanks	asleep	legs	friend	fell	going
	xx	morning	week	shes	left	went
Symptoms	sick	sleep	sore	cancer	pain	infection
	sore	headache	throat	breast	sore	pain
	throat	fall	pain	lung	head	mouth
	fever	insomnia	aching	prostate	foot	ear
	cough	sleeping	stomach	sad	feet	sinus
Treatments	hospital	sleeping	exercise	surgery	massage	surgery
	surgery	pills	diet	hospital	brace	braces
	antibiotics	caffeine	dieting	treatment	physical	antibiotics
	fluids	pill	exercises	heart	therapy	eye
	paracetamol	tylenol	protein	transplant	crutches	hospital

Topic Modelling for Social Media

- Collect Tweets by searching for keywords scraped from medical websites (medical lexicons);
- Train a classifier to separate relevant and irrelevant Tweets:
  - *I'm sick of this*
  - *I have Bieber fever!*
- How can we ensure that we find health-related topics?
  - *damn flu, home with a fever watching TV*
  - Ailment topic: *flu*
  - Symptom: *fever*
  - Another topic? *Home, watching, TV*

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS one.

Non-informative Prior Distributions

- LDA places priors over the word distribution for each topic:
  - $P(\boldsymbol{\beta_c} | \boldsymbol{\eta_c}) = Dirichlet(\boldsymbol{\eta_c})$,
  - $\boldsymbol{\beta_c}$ is a vector of probabilities for topic *c* for all words in the vocabulary.
- By default, we use **non-informative** priors that do give equal probability to all words in the vocabulary,
  - Use the same **hyperparameter** values, η_{cw}, for all words *w*.

Informative Prior Distributions

- **Informative** priors as a kind of **pretraining**:
  - Collect pages from WebMD on the most popular health topics;
  - Associate each collected page with an LDA topic, *c;*
  - Compute the word frequencies in each WebMD topic $count(w, c)$
  - Set $\eta_{cw} = 0.01 \cdot count(w, c)$
- Result: some topics are biased toward the topics collected from WebMD
- Other topics may be discovered from the social media data

Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. PloS one.

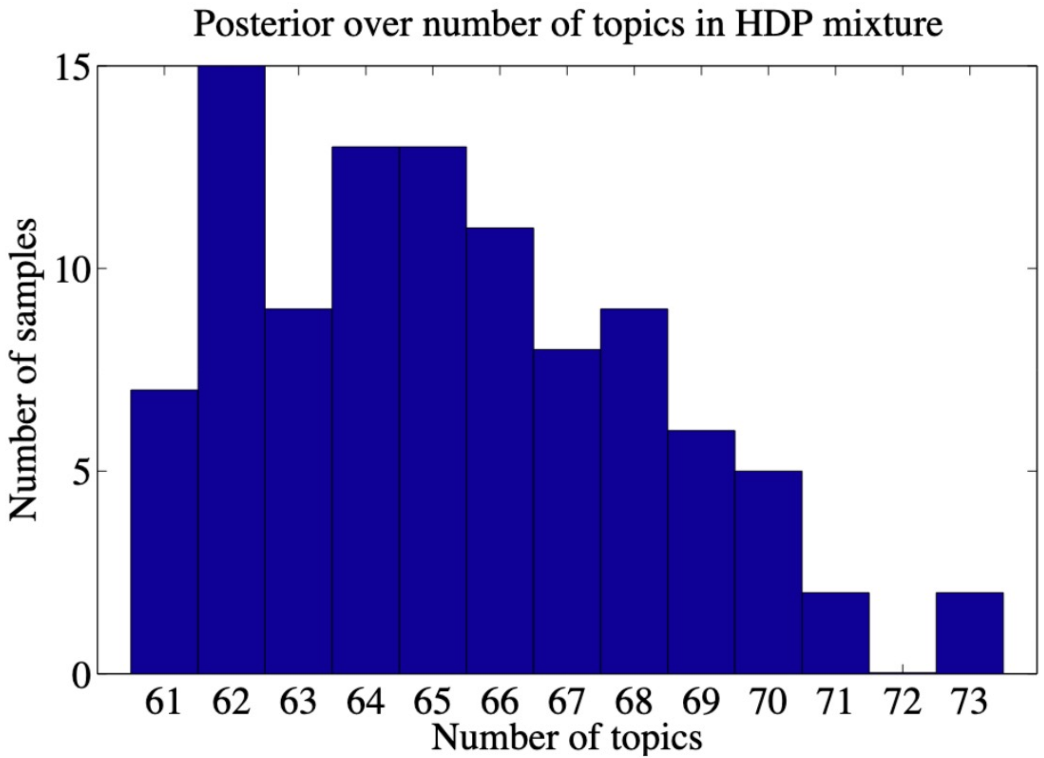
How Many Topics Do I Need?

- For LDA, it has to be fixed before training.
- One possible solution: use HDP instead of LDA…

Hierarchical Dirichlet Process (HDP)

- Generalises the LDA model
- Learns the number of topics needed to model a particular dataset
- Outputs a probability distribution over the number of topics →

Teh, Y. W., et al. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*



Summary

- Applications: tracking consumer complaints, scientific research and health topics
- Social media: filtering and preprocessing for applying topic models
- Number of topics is unclear: HDP avoids the need to specify in advance