



EMORY

GOIZUETA
BUSINESS
SCHOOL



TM-OKC: An Unsupervised Topic Model for Text in Online Knowledge Communities

Dongcheng Zhang
Emory University

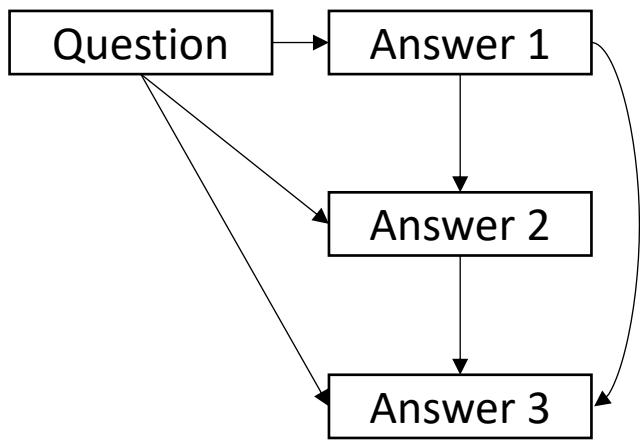
Kunpeng Zhang (KZ)
University of Maryland

David A. Schweidel
Emory University

October 14, 2023

1 Research motivation

- “Question-answer” textual data

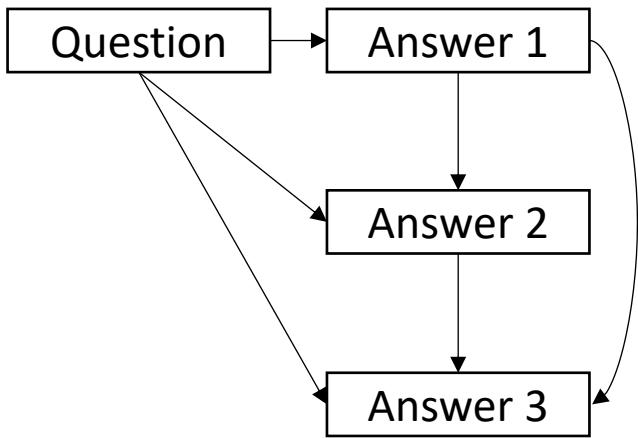


Note: “Question-answer” refers to a general structure



1 Research motivation

- “Question-answer” textual data



Note: “Question-answer” refers to a general structure



Marketing analytics texts

Asked 8 years ago Modified 4 years, 5 months ago Viewed 93 times

- Can you recommend a comprehensive textbook or a monograph on marketing analytics?
- 1 I'm interested in metrics, measures and statistical methods used in the field. For instance, what are the approaches at companies like Amazon and Walmart?
- UPDATE: I'm looking for a text like Greene, [Econometric Analysis](#). Econometrics is a set of statistical methods applied to economics. So, I want the same but for Marketing. I'm not interested in technology aspects, because that's the same everywhere: big data, databases etc.

[references](#) [marketing](#)

Share Cite Improve this question

Follow

dated Feb 26, 2019 at 19:12

community wiki
3 revs, 2 users 100%
Aksakal

What is your quantitative background? Generally speaking, I think that it would be useful to learn regression, time series analysis, and other statistical methods. Marketing Analytics = Statistics + SQL programming + business insights – [mmmmmmmmmm](#) Aug 2, 2015 at 1:33

@cwh_UCF, my background is econometrics + theor physics + programming – [Aksakal](#) Aug 2, 2015 at 1:41 ↗

Add a comment

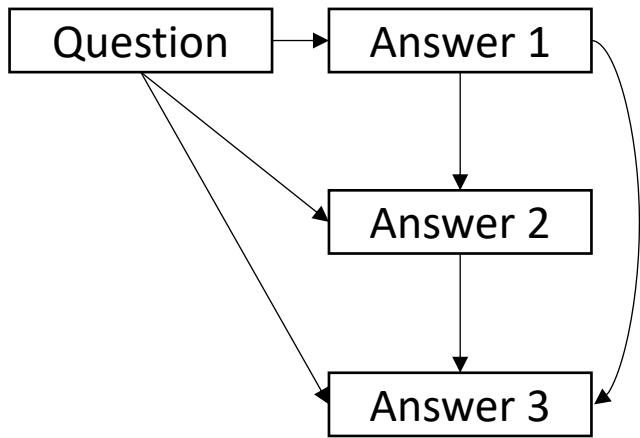
2 Answers

Sorted by: Highest score (default) ▾

With your background, I would recommend that you learn statistics and SQL programming. For the statistics, learn linear regression, logistic regression, time series analysis, decision trees,

1 Research motivation

- “Question-answer” textual data



Note: “Question-answer” refers to a general structure



merlinus12 • 13 hr. ago

Is your issue that AI will necessarily be a net negative in the long-term, or that it will have short-term transition costs that are harmful as society adopts it?

If the problem is simply short-term, then we shouldn't ban AI, but rather pass laws that tax job-killing AI and use that to fund assistance programs for those affected. This would also have the affect of slowing the AI rollout by increasing the cost, which would also help.

Put another way, AI is only a net negative if we don't respond as a society to this new tech with suitable regulation. Our failure to do so won't be the fault of the technology, but of our political systems which we have allowed to partisanize to the point of gridlock.

(–) ↑ ↓ Reply ↑ Share ...



Swaggshrew82 • 13 hr. ago

I'm mostly worried about the short term. Long term it could literally solve most of the world's problems. However...the short term will be catastrophic. I completely agree with what you said...which is why I'm worried. Our political system has shown zero ability to adapt to something like this. Our leaders are old psychopaths who don't understand technology.

(–) ↑ ↓ Reply ↑ Share ...



Kazthespokey • 13 hr. ago

Would catastrophe mean quicker political, economic and societal change?

If countries can't fix short term issues, won't the political unrest and agitation fix it?

↑ ↓ Reply ↑ Share ...

(+) 11 more replies



merlinus12 • 12 hr. ago

Then it isn't AI that's the problem, it's our politics.

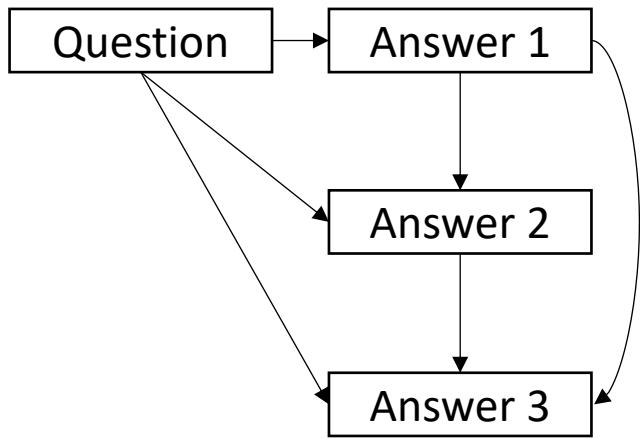
That's an incredibly important distinction, because it lets us focus our energy in the right direction. We need to push for legislation and political change, not try to bury AI (as many propose).

It's still quite possible that we won't act effectively, but we have no chance of doing so if we don't identify the real problem.

Social media posts and responses

1 Research motivation

- “Question-answer” textual data



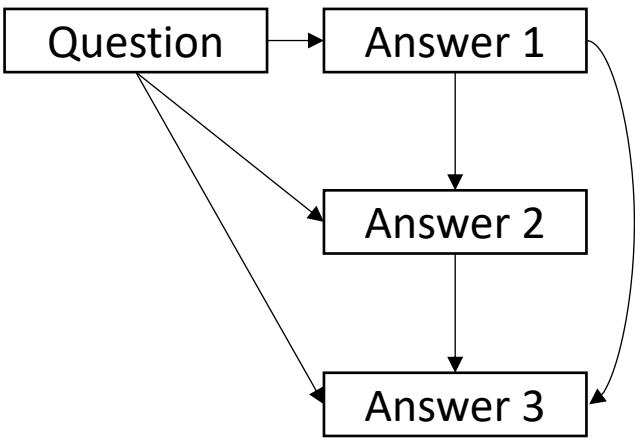
Note: “Question-answer” refers to a general structure

- Millions of users
 - Stack Exchange (100+M users), Quora (300+M users)
 - Twitter (300+M users), Reddit (400+M users)
- Huge business opportunities
 - \$ 65.31 billion on social advertising (US) (eMarketer, 2022)



1 Research motivation

- “Question-answer” textual data



Note: “Question-answer” refers to a general structure



 Gabby Tess

★★★★★ The Dream Team of Laundry: Tide Pods with Downy Unleash the Magic!

Reviewed in the United States on June 16, 2023

Unit Count: 73.0 | Verified Purchase

Prepare to witness the ultimate laundry alchemy with Tide Pods infused with Downy. It's like a symphony of cleanliness and freshness that will leave you in awe. Brace yourself for a laundry experience like no other!

These little wonders are the secret to transforming your laundry routine into a magical affair. With the power of Tide's cleaning prowess and the irresistible scent of Downy, your clothes emerge from the wash with a newfound brilliance and a touch of heavenly softness.

The convenience of these Tide Pods is simply unrivaled. Just toss one in, and let the enchantment begin. No measuring, no mess—just pure laundry bliss. The pods dissolve effortlessly, releasing their extraordinary cleaning powers that obliterate stains and leave your garments looking as good as new.

But wait, there's more! The infusion of Downy fabric softener takes your laundry game to the next level. Say goodbye to rough fabrics and hello to a realm of luxurious softness. Your clothes will caress your skin

▼ [Read more](#)

2 people found this helpful

[Helpful](#) | [Report](#)

 Pearl Christian

★★★★★ Last 3 orders have pods that have ruptured

Reviewed in the United States on July 15, 2023

Unit Count: 61.0 | Verified Purchase

The last 3 orders have had pods that were ruptured and then they leak to other pods and stick together. I have a new container that I have not used yet. What can be done about this? Usually 4 or 5 pods stick together and you cannot get them apart without them bursting. So I am losing 4 or 5 pods per container.

[Helpful](#) | [Report](#)

 Gabby Bon

★★★★★ Works!

Reviewed in the United States on July 28, 2023

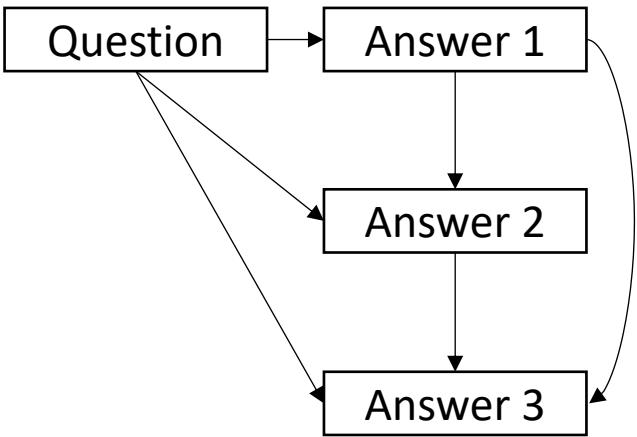
Unit Count: 61.0 | Verified Purchase

Have always used Tide. Our clothes are clean and fresh. Use to use the liquid Tide and now have totally switched to the pods- so convenient!

Product descriptions and reviews

1 Research motivation

- “Question-answer” textual data



Note: “Question-answer” refers to a general structure

Hacker News new | past | comments | ask | show | jobs | submit

Ask HN: Is GPT 4's quality lately worst than GPT 3.5?
41 points by agonz253 11 hours ago | hide | past | favorite | 36 comments
Has anyone else encountered this phenomenon lately? I've found myself prompting GPT 3.5 with simple questions that GPT 4 provided an incorrect answer for, and lo and behold I get a much better answer for ex this is GPT 4: <https://chat.openai.com/share/e24501ad-8f1c-4b5a-a6d0-d933f5d1d209>
And this is GPT 3.5: <https://chat.openai.com/share/b9372bdc-ffff-4655-bee4-2b3f3c3b8285>
In the latter case I didn't even need to ask for the order by clause as it anticipates it and provides an answer for it. GPT 4's first answer was wrong.
In the past two days I've seen at least 2 other cases where GPT 4's answer was plain wrong and GPT 3.5's was not only correct but of very high quality, reminding me of what I first felt when using GPT 4 the first time.

muzani 25 minutes ago | next [-]
My observation (ChatGPT and not the API models):
For code, 3.5 is superior. 3.5 allows for about 21k tokens of input, while GPT 4 is ~10k for round 10k. This also makes it a lot better for boilerplate work as it can take a lot more input, and handles long conversations and iterations better.
Brainstorming, 4 is better. It's capable of some top tier brainstorming and it gives back quite frequently.
Unguided creative writing (describe a potato), they're roughly equal.
Guided creative writing (i.e. write a story around (4.0, 0.0, 0.0 requirements)), 4 is much better.
Poems and wordplay, 4 absolutely floors 3.5. With vocabulary and it's able to do rhymes and alliterations better, which humans are usually bad at.
For reasoning and riddles, 4 is still benchmark among the LLMs.
I really dislike that they named it GPT-3.5 instead of something like "Glide". It implies that it's inferior to 4, when they're just suited for different things.
reply

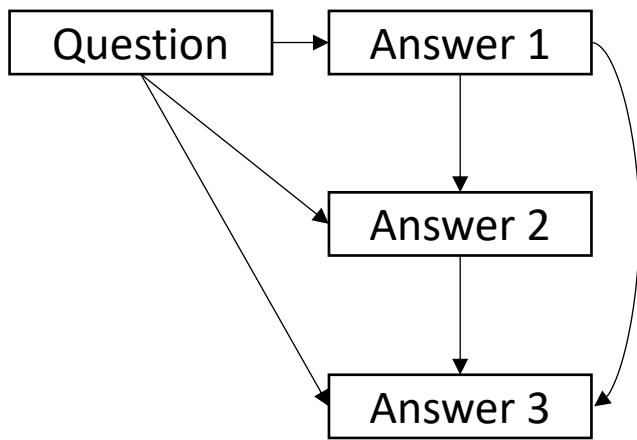
kromem 5 hours ago | prev | next [-]
Yes, and I'm willing to bet that within 12 months we'll be looking back realizing that this was due to the fine tuning taking the world's SorA pretrained model aligned with "completing human tax" and putting it in the box of "you are an AI without feelings or desires tasked with XYZ."
The search space on the fine tuned GPT-3.5 chat models versus the foundational Davinci text completion model is MUCH more narrow, particularly in starting off.
Even with the same temperature, you'll see any marketing-style prompt for chat begin with "Introducing XYZ..." around 30% of the time as if it's a junior door to door salesman, whereas the foundational model doesn't have any single intro that common across runs and generally employs a much broader vocabulary set.
We saw Google shoot Lambda in the foot after Blake's press tour which set them behind the next round of competition.

News articles and comments



1 Research motivation

- “Question-answer” textual data



Note: “Question-answer” refers to a general structure

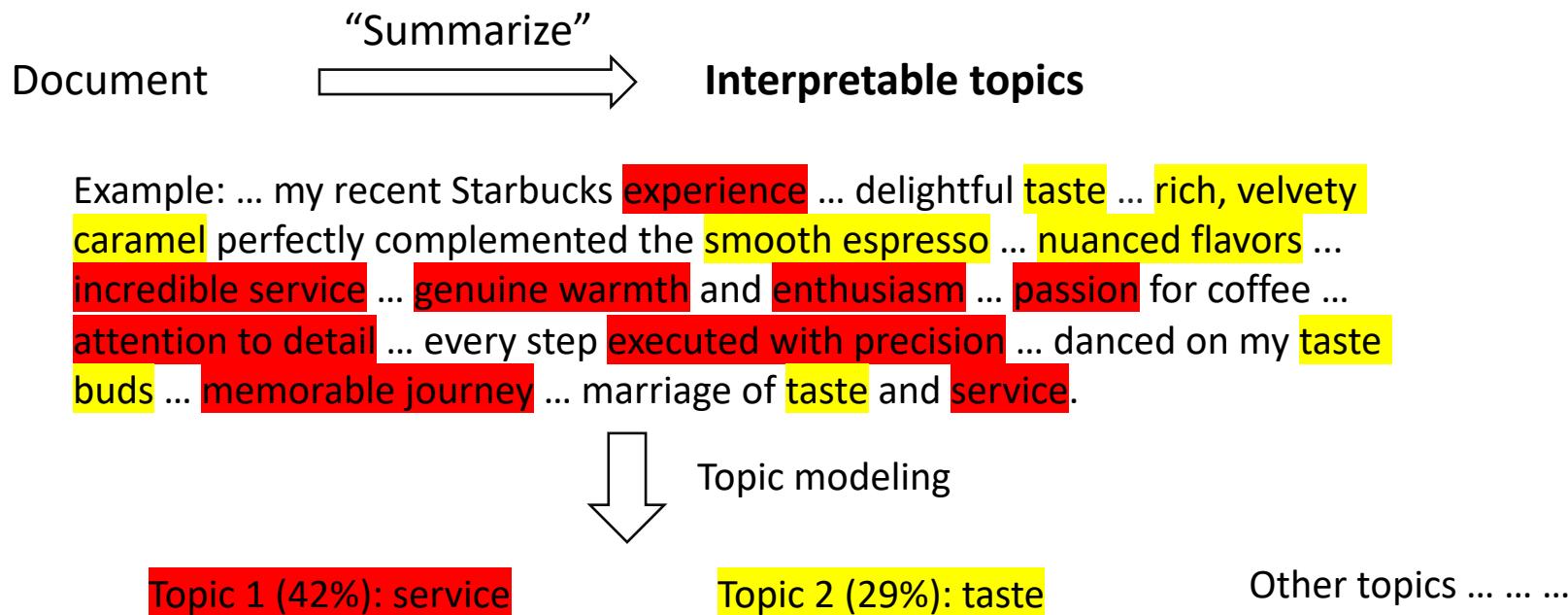
How to extract information from texts?

- User interests’ discovery
- Consumer behavior analysis
- Empirical IS research



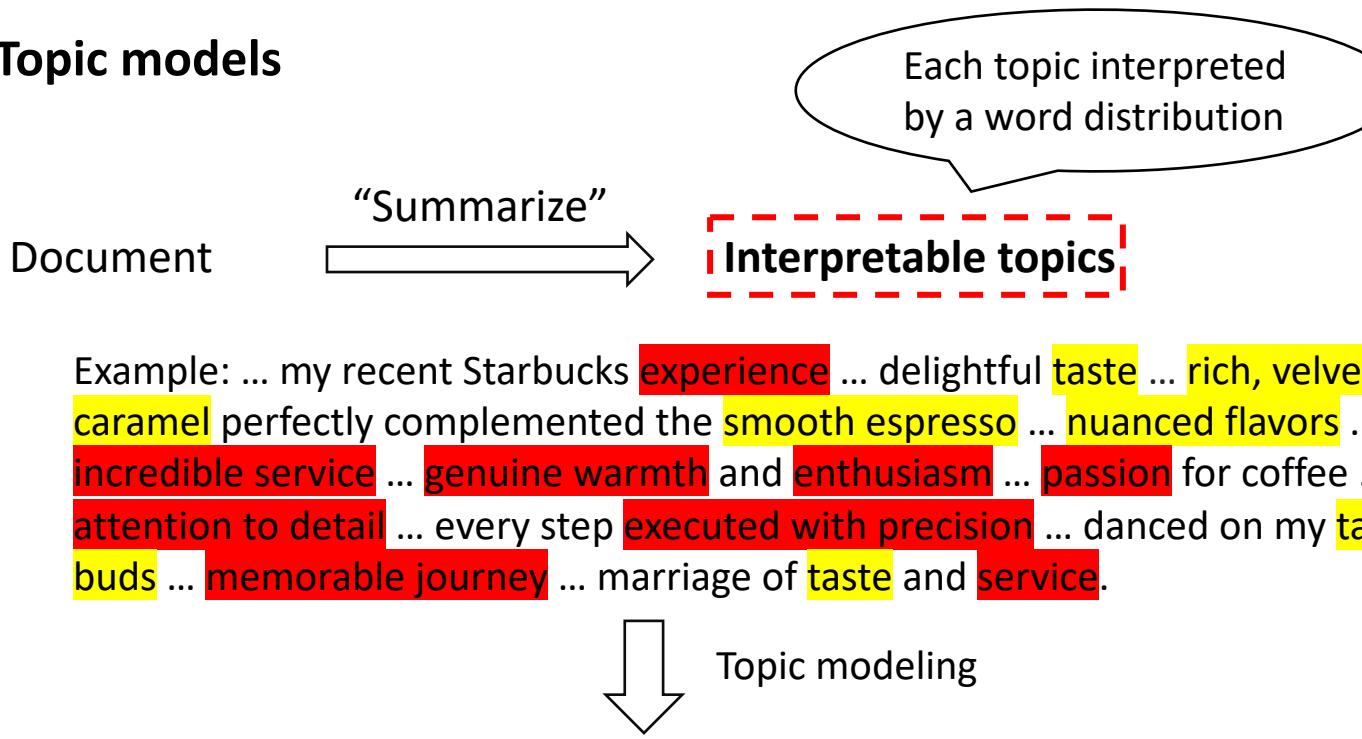
1 Research motivation

- Topic models



1 Research motivation

- Topic models



Topic 1 (42%): service

Service	0.069
Experience	0.041
Journey	0.039
Warmth	0.029
Enthusiasm	0.018

Topic 2 (29%): taste

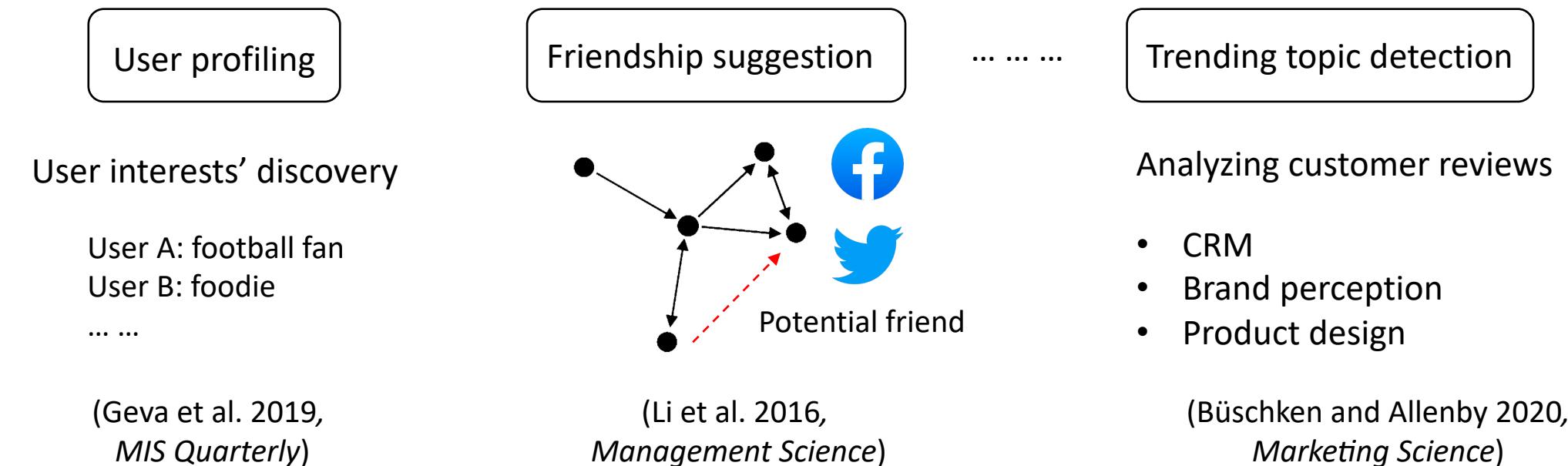
Taste	0.180
Flavor	0.049
Smooth	0.042
Rich	0.033
Nuance	0.027

Other topics



1 Research motivation

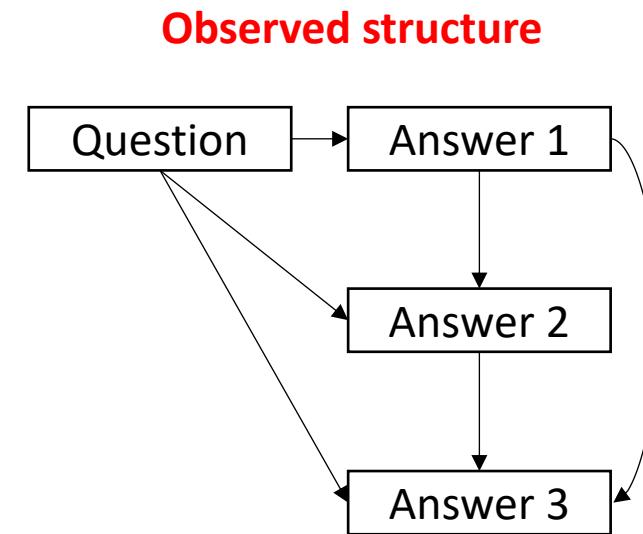
- Topic models



1 Research motivation

- Topic models

- Existing generic topic models (e.g., LDA)
 - Posts are independent
 - No structural relationships



Question: ... data I am working with is not 100% clean ... What are the best practices in such a case as this?

Answer A: 1. Do not modify the original data. Having the original data source intact is important ...

Answer B: ... I agree with the other posters that you should not modify the original data, but add fields for corrected values. I developed a technique in our systems (opengeocode.org) ...

Goal: Develop a new topic model to capture the complex structural relations



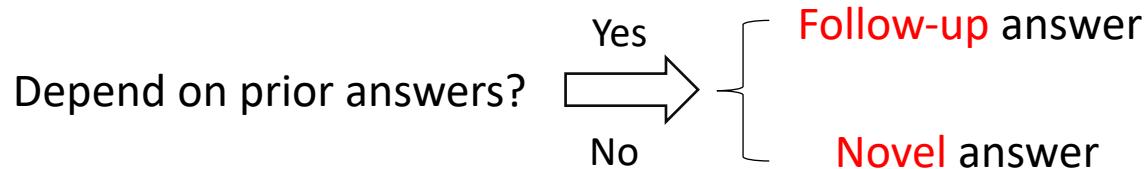
2 Model development: Key features

- Three key features

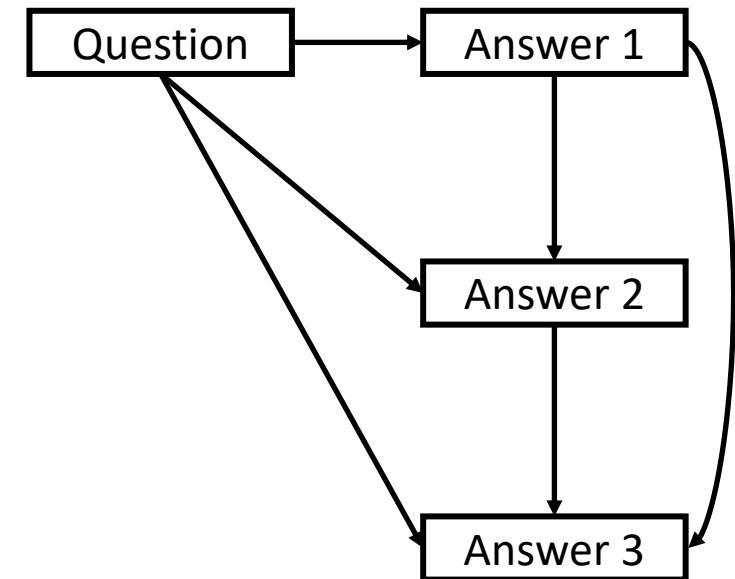
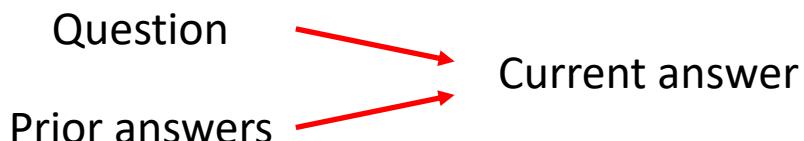
(1) Q&A relationship



(2) Threaded structure + heterogeneity

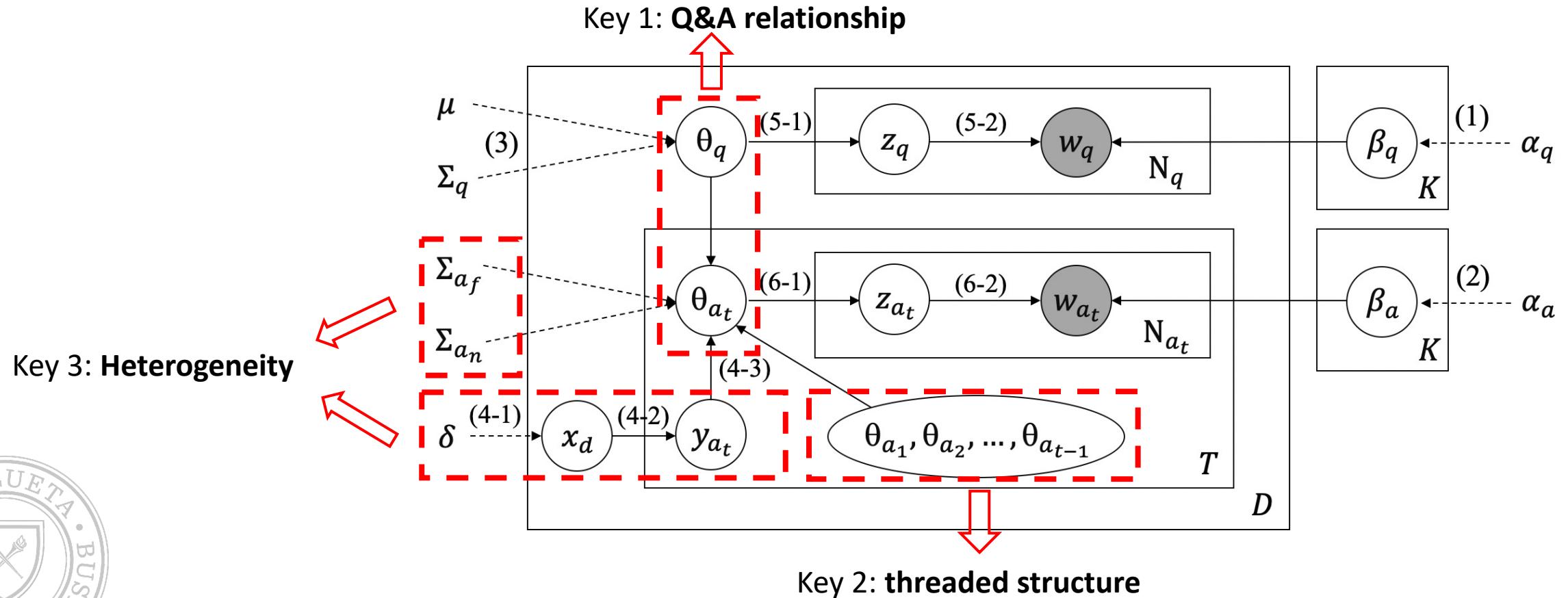


(3) Heterogeneous impacts



2 Model development: Our novel framework

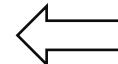
- Develop a novel unsupervised Bayesian topic modeling framework



2 Model development: Inference and estimation

- Complex model inference and estimation under high interdependencies
 - Model inference
 - Variational mean-field inference and coordinate ascent algorithm
 - Maximize the evidence lower bound (ELBO)

$$\begin{aligned}
 \log p(\mathbf{w}_q, \mathbf{w}_{a_{1:T}}; \boldsymbol{\mu}, \boldsymbol{\Sigma}_q, \boldsymbol{\Sigma}_{a_f}, \boldsymbol{\Sigma}_{a_n}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\alpha}_q, \boldsymbol{\alpha}_a) &\geq E_u[\log p(\boldsymbol{\eta}_q; \boldsymbol{\mu}, \boldsymbol{\Sigma}_q)] + \sum_{n_q=1}^{N_q} E_u[\log p(z_q^{n_q} | \boldsymbol{\eta}_q)] \\
 &+ \sum_{n_q=1}^{N_q} E_u[\log p(w_q^{n_q} | z_q^{n_q}, \boldsymbol{\beta}_q^{1:K})] + \sum_{k=1}^K E_u[\log p(\boldsymbol{\beta}_q^k; \boldsymbol{\alpha}_q)] + E_u[\log p(\mathbf{x}_d; \boldsymbol{\delta})] \\
 &+ \sum_{t=1}^T E_u[\log p(\mathbf{y}_{a_t} | \mathbf{x}_d)] + \sum_{t=1}^T E_u[\log p(\boldsymbol{\eta}_{a_t} | \boldsymbol{\eta}_q, \bar{\boldsymbol{\eta}}_{a_{t-1}}, \mathbf{y}_{a_t}, \boldsymbol{\Sigma}_{a_f}, \boldsymbol{\Sigma}_{a_n}, \boldsymbol{\gamma})] \\
 &+ \sum_{t=1}^T \sum_{n_{a_t}=1}^{N_{a_t}} E_u[\log p(z_{a_t}^{n_{a_t}} | \boldsymbol{\eta}_{a_t})] + \sum_{t=1}^T \sum_{n_{a_t}=1}^{N_{a_t}} E_u[\log p(w_{a_t}^{n_{a_t}} | z_{a_t}^{n_{a_t}}, \boldsymbol{\beta}_a^{1:K})] \\
 &+ \sum_{k=1}^K E_u[\log p(\boldsymbol{\beta}_a^k; \boldsymbol{\alpha}_a)] + H(u),
 \end{aligned}$$



$$\begin{aligned}
 u(\boldsymbol{\eta}_q, \boldsymbol{\eta}_{a_{1:T}}, \mathbf{x}_d, \mathbf{y}_{a_{1:T}}, \mathbf{z}_q, \mathbf{z}_{a_{1:T}}, \boldsymbol{\beta}_q, \boldsymbol{\beta}_a) \\
 = \prod_{k=1}^K u(\eta_q^k; \lambda_q^k, (\sigma_q^k)^2) \prod_{t=1}^T \prod_{k=1}^K u(\eta_{a_t}^k; \lambda_{a_t}^k, (\sigma_{a_t}^k)^2) u(\mathbf{x}_d; \boldsymbol{\nu}_d) \prod_{t=1}^T u(\mathbf{y}_{a_t}; \boldsymbol{\psi}_{a_t}) \\
 \prod_{n_q=1}^{N_q} u(z_q^{n_q}; \boldsymbol{\phi}_q^{n_q}) \prod_{t=1}^T \prod_{n_{a_t}=1}^{N_{a_t}} u(z_{a_t}^{n_{a_t}}; \boldsymbol{\phi}_{a_t}^{n_{a_t}}) \prod_{k=1}^K u(\boldsymbol{\beta}_q^k; \boldsymbol{\tau}_q^k) \prod_{k=1}^K u(\boldsymbol{\beta}_a^k; \boldsymbol{\tau}_a^k).
 \end{aligned}$$

- Parameter estimation
 - Variational expectation-maximization (VEM)
 - Maximize the likelihood bound

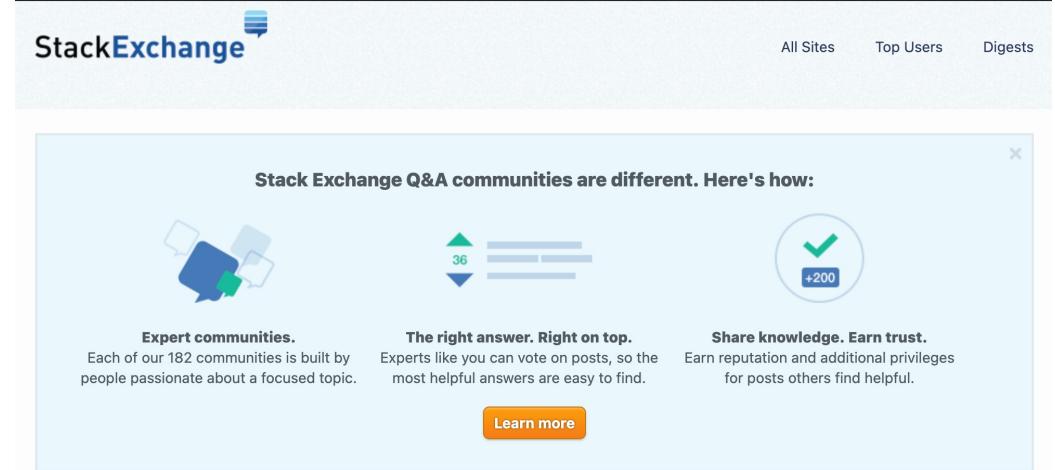
$$\begin{aligned}
 L(\boldsymbol{\mu}, \boldsymbol{\Sigma}_q, \boldsymbol{\Sigma}_{a_f}, \boldsymbol{\Sigma}_{a_n}, \boldsymbol{\gamma}; \mathbf{w}_{1:D,q}, \mathbf{w}_{1:D,a_{1:T}}) &\geq \hat{L} \\
 &= \sum_{d=1}^D \{E_{u_d}[\log p(\boldsymbol{\eta}_{d,q}, \boldsymbol{\eta}_{d,a_{1:T}}, \mathbf{x}_d, \mathbf{y}_{d,a_{1:T}}, \mathbf{z}_{d,q}, \mathbf{z}_{d,a_{1:T}}, \boldsymbol{\beta}_q, \boldsymbol{\beta}_a, \mathbf{w}_{d,q}, \mathbf{w}_{d,a_{1:T}})] \\
 &+ H(u_d)\}.
 \end{aligned}$$



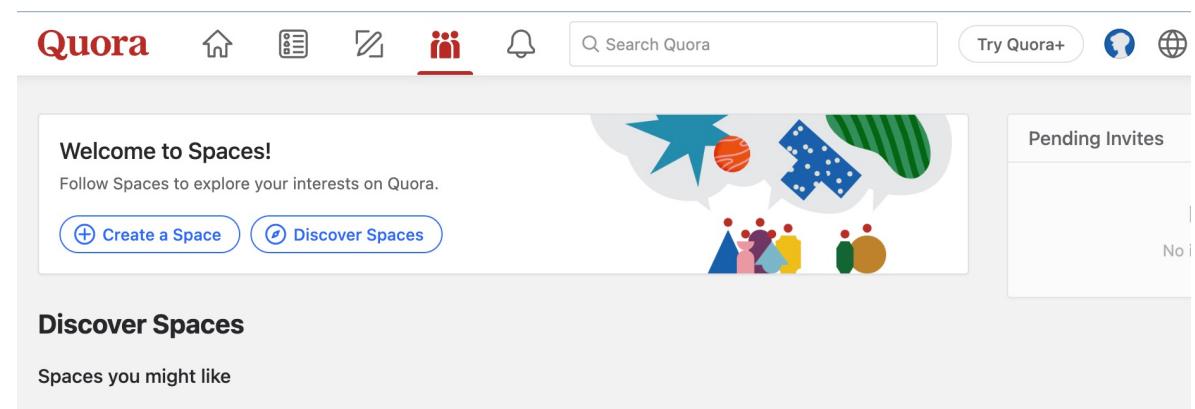
3 Model evaluation: Datasets

- Real-world datasets

- Stack Exchange, 6 categories



- Quora, 3 categories



3 Model evaluation: Datasets

- Real-world datasets

- Stack Exchange, 6 categories
- Quora, 3 categories

Number of Q&A threads: 1,000 ~ 100,000

Descriptive statistics of the Stack Exchange dataset

Section (Category)	# of questions	# of answers	# of words per question/answer
Technology (Data Science)	20740	31181	73.10
Culture/Recreation (English Language & Usage)	109977	268356	47.45
Life/Arts (Cooking)	23387	58078	57.49
Science (Computer Science)	31156	45807	72.49
Professional (Writing)	10429	34116	77.61
Business (Project Management)	5782	17724	70.26

Descriptive statistics of the Quora dataset

Section (Category)	# of questions	# of answers	# of words per question/answer
Science and Technology	1471	3144	28.82
Business and Marketing	1140	1981	23.50
Health and Life	2281	4144	20.18



3 Model evaluation: Statistical model fit

- Perplexity: based on **log-likelihood** (Blei et al. 2003; Roberts et al. 2019)

How “**perplexed**” the model is?



$$\text{perplexity} = \exp \left\{ - \frac{\log p(\mathbf{w})}{\sum_d^{|D_{test}|} \sum_{v=1}^V \text{num}_d^v} \right\}$$

Lower perplexity



Higher likelihood



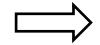
Better model fit / generalizability



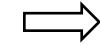
3 Model evaluation: Statistical model fit

- Perplexity: based on **log-likelihood** (Blei et al. 2003; Roberts et al. 2019)

Lower perplexity



Higher likelihood



Better model fit / generalizability

Perplexity comparison for the Stack Exchange dataset

Model	Technology (Data Science)	Culture/ Recreation (English Language & Usage)	Life/Arts (Cooking)	Science (Computer Science)	Professional (Writing)	Business (Project Management)
LDA	1152.11	2181.84	1520.59	1063.53	1726.48	1087.31
NTM	1021.39	2176.38	1407.79	1028.45	1639.31	1063.06
TRTM	1029.95	2030.30	1296.18	976.34	1727.38	998.90
STM	1054.08	2015.43	1271.62	975.40	1718.43	1010.79
SCHOLAR	951.94	<u>1586.73</u>	1251.94	934.26	1615.01	1037.08
QATM	931.21	1625.43	1297.68	958.23	1479.63	939.54
LeadLDA1	1256.93	1702.28	1376.98	1063.31	1512.57	956.25
LeadLDA2	1100.50	1707.15	1326.48	1075.64	1498.23	1064.47
SITS	<u>1008.46</u>	<u>1635.08</u>	<u>1280.41</u>	<u>979.68</u>	<u>1482.28</u>	<u>934.51</u>
TM-OKC (mean)	863.44***	1591.63	<u>1220.97</u>	<u>891.59</u>	1400.93	901.03
TM-OKC (decay)	<u>863.93</u>	1582.44	1221.01	891.30***	<u>1394.29</u>	899.39***
TM-OKC (weight)	867.69	1591.73	1218.70***	892.54	1390.96***	<u>900.59</u>

Perplexity comparison for the Quora dataset

Model	Science and Technology	Business and Marketing	Health and Life
LDA	1579.54	1698.37	1357.71
NTM	1204.54	1368.05	1083.70
TRTM	1131.22	1310.21	1104.22
STM	1117.93	1301.27	1109.51
SCHOLAR	1192.71	1321.77	1087.89
QATM	1033.00	1243.23	1065.07
LeadLDA1	1399.21	1227.56	1088.73
LeadLDA2	1350.26	1242.85	1132.91
SITS	1313.06	1264.70	1058.69
TM-OKC (mean)	1014.30***	1202.66***	1042.97
TM-OKC (decay)	<u>1015.86</u>	<u>1212.19</u>	<u>1040.38</u>
TM-OKC (weight)	1018.13	1215.98	1032.61***



3 Model evaluation: Statistical model fit

- Coherence score
 - “Similarity” (word co-occurrences) of the top words within a topic
 - Higher score indicates the learned topics are more coherent (Syed and Spruit 2017; Dieng et al. 2020)

Coherence comparison for the Stack Exchange dataset

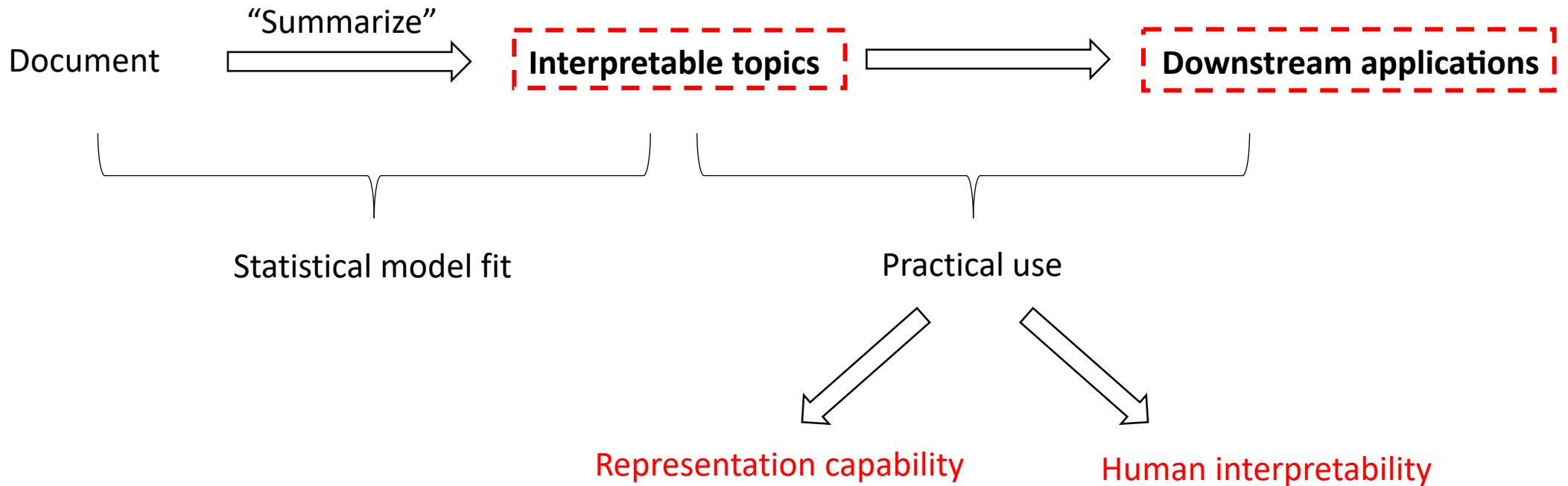
Model	Technology (Data Science)	Culture/ Recreation (English Language & Usage)	Life/Arts (Cooking)	Science (Computer Science)	Professional (Writing)	Business (Project Management)
LDA	0.51	0.42	0.47	0.47	0.37	0.37
NTM	0.53	0.43	0.47	0.48	0.37	0.39
TRTM	0.51	0.42	0.46	0.48	0.37	0.36
STM	0.53	0.41	0.48	0.49	0.39	0.37
SCHOLAR	0.53	0.43	0.48	0.49	0.37	0.37
QATM	0.52	0.42	0.48	0.48	0.37	0.42
LeadLDA1	0.50	0.40	0.47	0.47	0.35	0.36
LeadLDA2	0.51	0.40	0.46	0.48	0.36	0.37
SITS	0.52	0.41	0.47	0.48	0.38	0.39
TM-OKC (mean)	0.53	0.45*	0.50*	0.51**	0.39	0.42
TM-OKC (decay)	0.53	0.43	0.50*	0.51**	0.38	0.41
TM-OKC (weight)	0.56***	0.44	0.50*	0.51**	0.38	0.41

Coherence comparison for the Quora dataset

Model	Science and Technology	Business and Marketing	Health and Life
LDA	0.40	0.39	0.41
NTM	0.46	0.41	0.42
TRTM	0.42	0.42	0.42
STM	0.43	0.43	0.43
SCHOLAR	0.45	0.40	0.42
QATM	0.43	0.44	0.42
LeadLDA1	0.46	0.38	0.44
LeadLDA2	0.41	0.39	0.42
SITS	0.43	0.41	0.43
TM-OKC (mean)	0.48**	0.45	0.47***
TM-OKC (decay)	0.47	0.46**	0.46
TM-OKC (weight)	0.48**	0.45	0.46



3 Model evaluation: More direct evaluation



3 Model evaluation: Representation capability

- *Document classification task* (Zeng et al. 2019; Yang et al. 2022)



Prediction accuracy of different methods

	Number of topics	Stack Exchange			Quora		
		40	80	120	40	80	120
Basic text feature extraction methods	TF-IDF features (top 100 words)		0.517			0.600	
	TF-IDF features (top 500 words)		0.717			0.779	
	TF-IDF features (top 1000 words)		0.718			0.804	
Bayesian topic modeling methods	LDA	0.727	0.710	0.699	0.728	0.726	0.712
	TRTM	0.891	0.885	0.893	0.935	0.920	0.937
	STM	0.886	0.847	0.888	0.935	0.935	0.930
	QATM	0.898	0.853	0.860	0.960	0.962	0.937
	LeadLDA1	0.867	0.832	0.751	0.762	0.749	0.747
	LeadLDA2	0.865	0.865	0.749	0.796	0.762	0.737
Topic modeling combined with deep language models	SITS	0.903	0.891	0.868	0.881	0.893	0.876
	NTM	0.905	0.900	0.859	0.813	0.827	0.848
	SCHOLAR	0.901	0.903	0.885	0.826	0.855	0.818
Representation learning methods	Bi-LSTM		0.822			0.919	
	Pre-trained BERT		0.696			0.722	
	Fine-tuned BERT		0.936			0.992	
Our method	TM-OKC	0.911*	0.933***	0.918***	0.983**	0.990***	0.968***



3 Model evaluation: Interpretability

- Lab studies
 - *word intrusion* and *topic intrusion* tasks (Chang et al. 2009; Bao and Datta 2014; Palese and Piccoli 2020)

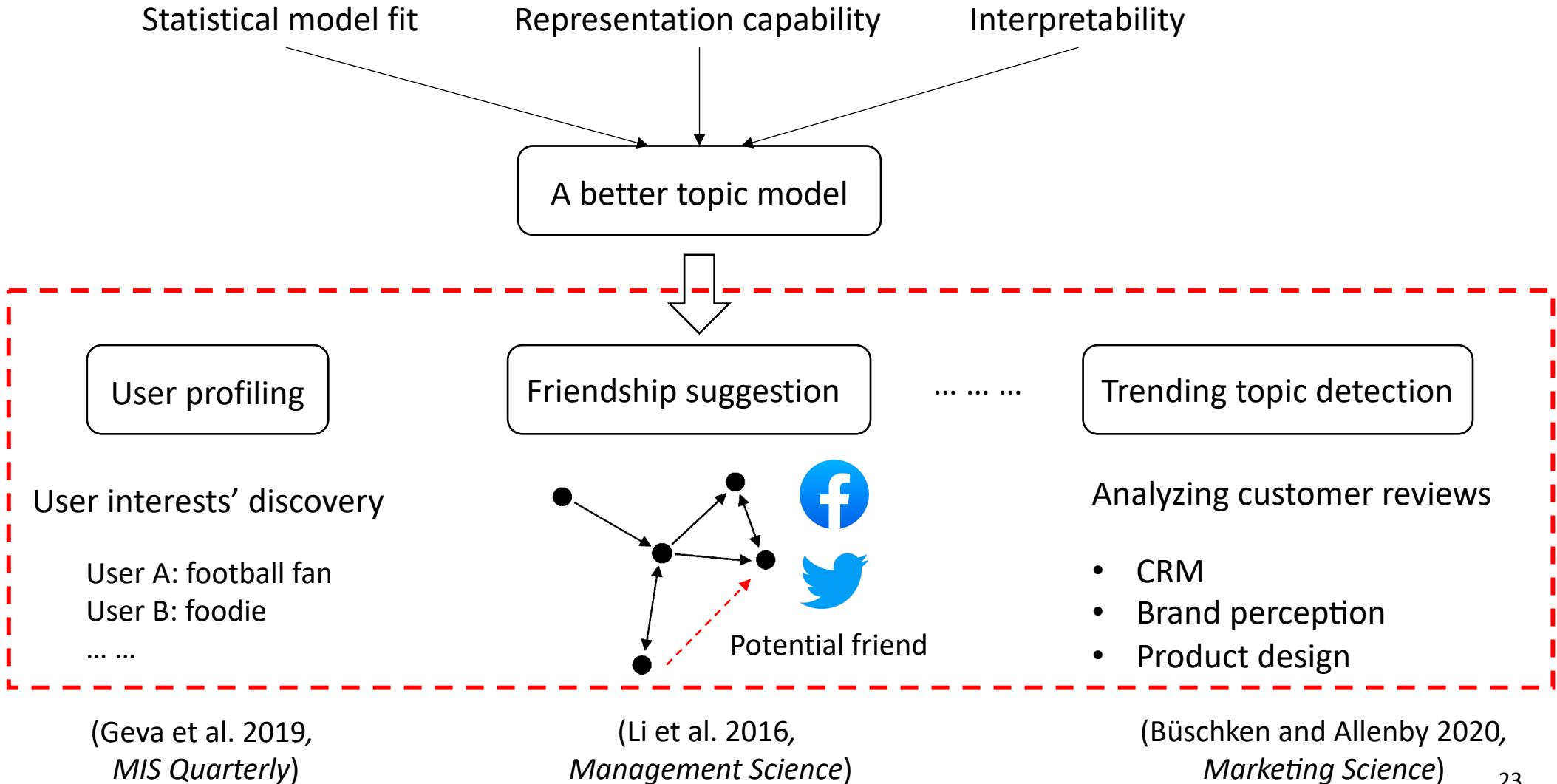


Human evaluation results of *word intrusion* and *topic intrusion* tasks

Number of topics	WIP_m in <i>word intrusion</i>			TLO_m in <i>topic intrusion</i>		
	40	80	120	40	80	120
LDA	0.806	0.794	0.788	-1.61	-1.65	-1.85
NTM	0.825	0.813	0.819	-1.52	-1.47	-1.48
TRTM	0.813	0.838	0.806	-1.43	-1.35	-1.49
STM	0.819	0.825	0.819	-1.37	-1.25	-1.47
SCHOLAR	0.813	0.819	0.813	-1.36	-1.30	-1.51
QATM	0.825	0.831	0.813	-1.31	-1.22	-1.39
SITS	0.819	0.825	0.800	-1.32	-1.29	-1.38
LeadLDA1	0.813	0.806	0.800	-	-	-
LeadLDA2	0.819	0.813	0.794	-	-	-
TM-OKC	0.856**	0.869**	0.838*	-1.12**	-0.96***	-1.23**

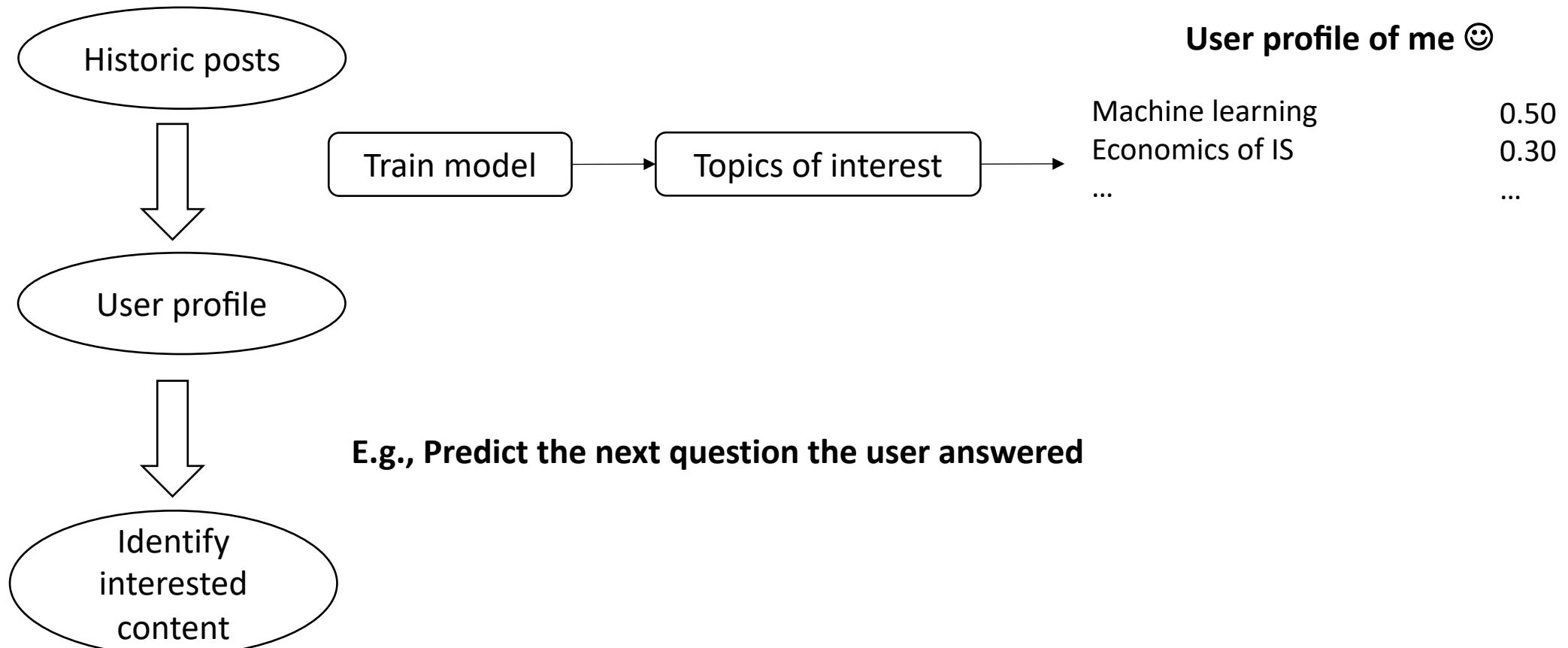


4 Downstream applications



4 Downstream applications

- Example: user profiling
 - User interests' discovery (He et al. 2017; Dhillon and Aral 2021)



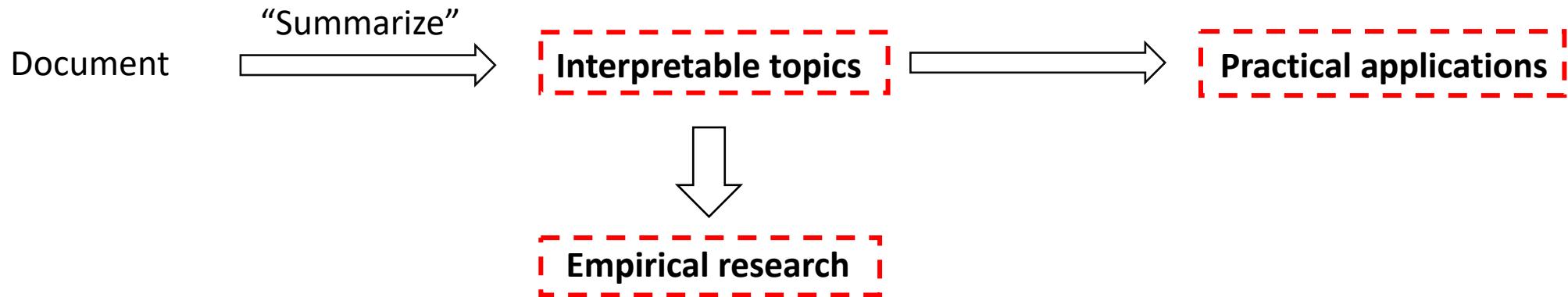
4 Downstream applications

- Example: user profiling
 - User interests' discovery (He et al. 2017; Dhillon and Aral 2021)

User profiling performance comparison of different methods

		Hit rate for top K			Data size N (number of Q&A threads)													
		$K=5$	$K=10$	$K=20$	1,000		2,000		4,000		8,000		16,000		32,000		64,000	
Basic text feature extraction methods	TF-IDF features (top 100 words)	5.1%	11.7%	21.4%	TF-IDF features (top 100 words)	10.5%	10.7%	11.1%	11.7%	11.1%	12.4%	12.2%						
	TF-IDF features (top 500 words)	5.3%	11.5%	21.2%	TF-IDF features (top 500 words)	10.3%	10.6%	11.5%	11.5%	11.8%	11.3%	11.5%						
	TF-IDF features (top 1000 words)	5.5%	11.8%	21.6%	TF-IDF features (top 1000 words)	10.6%	10.4%	11.9%	11.8%	11.2%	11.2%	11.7%						
Bayesian topic modeling methods	LDA	20.9%	33.9%	49.7%	LDA	26.9%	29.6%	31.5%	33.9%	34.8%	35.9%	36.4%						
	TRTM	26.6%	43.7%	59.3%	TRTM	37.7%	39.8%	40.5%	43.7%	44.3%	45.1%	45.1%						
	STM	26.5%	44.1%	59.4%	STM	36.4%	40.9%	42.0%	44.1%	44.9%	45.2%	45.6%						
	QATM	26.5%	45.2%	60.1%	QATM	30.0%	35.0%	40.0%	45.2%	45.7%	45.3%	46.0%						
	LeadLDA1	6.2%	11.7%	22.1%	LeadLDA1	10.2%	10.1%	10.7%	11.7%	12.2%	14.8%	15.2%						
	LeadLDA2	6.5%	12.8%	24.1%	LeadLDA2	10.5%	10.6%	10.6%	12.8%	12.1%	15.4%	15.9%						
	SITS	26.0%	41.4%	58.3%	SITS	34.5%	40.1%	39.8%	41.4%	43.9%	44.7%	45.8%						
Pre-trained deep language models	Pre-trained BERT	8.6%	14.9%	26.6%	Pre-trained BERT	13.3%	14.1%	13.4%	14.9%	14.6%	13.6%	14.1%						
Topic modeling combined with deep language models	NTM	18.2%	32.1%	42.8%	NTM	15.5%	23.6%	27.0%	32.1%	41.7%	47.2%	50.6%						
SCHOLAR	SCHOLAR	19.4%	33.6%	50.6%	SCHOLAR	20.4%	22.5%	28.8%	33.6%	42.6%	47.8%	51.4%						
Neural matrix factorization	NMF	19.0%	32.1%	46.0%	NMF	20.8%	23.7%	27.8%	32.1%	40.8%	46.7%	49.8%						
Our method	TM-OKC	32.6%***	49.1%**	64.8%**	Our method	46.8%***	47.7%***	48.0%***	49.1%**	49.7%**	50.1%*	51.3%						

4 Downstream applications



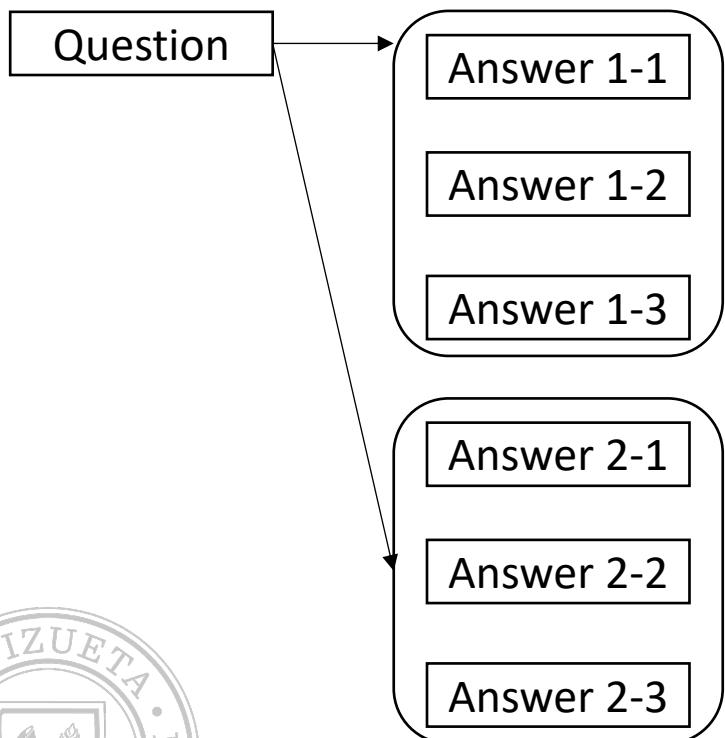
- **Topics as “independent variables”** (Yue et al. 2019, *MIS Quarterly*; Narang et al. 2022, *Journal of Marketing Research*; Gour et al. 2022, *Production and Operations Management*)
- **Topics as “dependent variables”** (Singh et al. 2014, *Information Systems Research*; Tirunillai and Tellis 2014, *Journal of Marketing Research*; Geva et al. 2019, *MIS Quarterly*)
- **Topics used to derive new variables** (Ghose et al. 2019, *Management Science*; Pu et al. 2022, *Information Systems Research*; Bachura et al. 2022, *MIS Quarterly*)



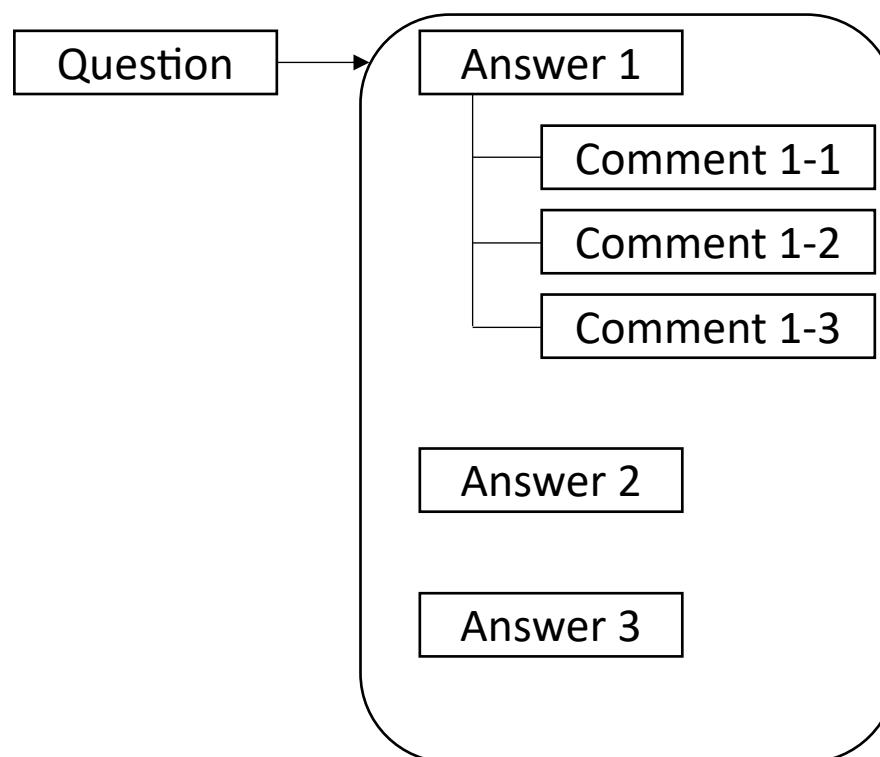
5 Model extensions



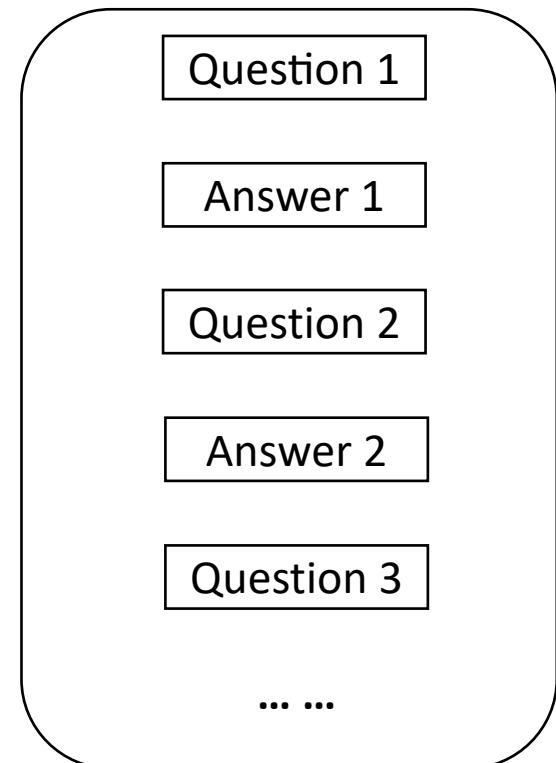
Scenario 1



Scenario 2

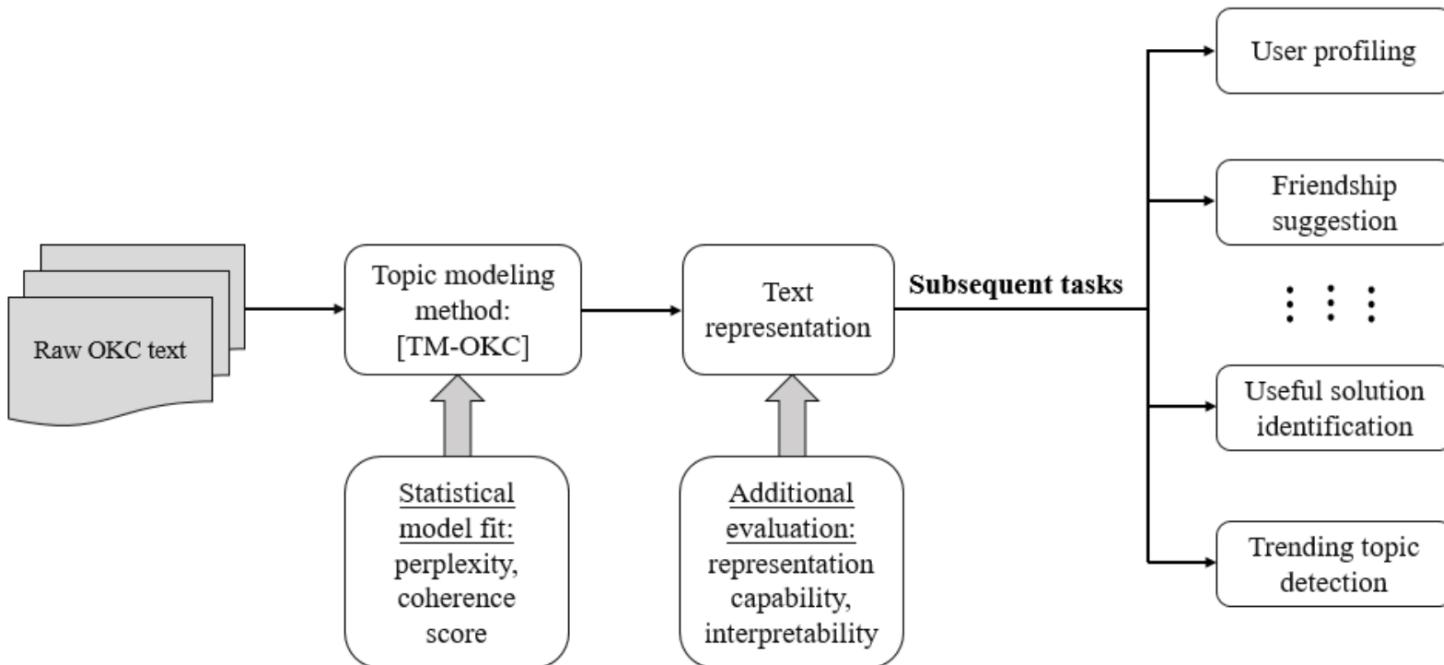


Scenario 3



6 Summary

- Develop a novel unsupervised Bayesian topic modeling framework
 - Complex Q&A relations and threaded structures among texts with heterogeneity
- Superior performance
 - Statistical model fit, representation capability, and human interpretability
- Implications to IS researchers and practitioners in downstream empirical studies and practical tasks

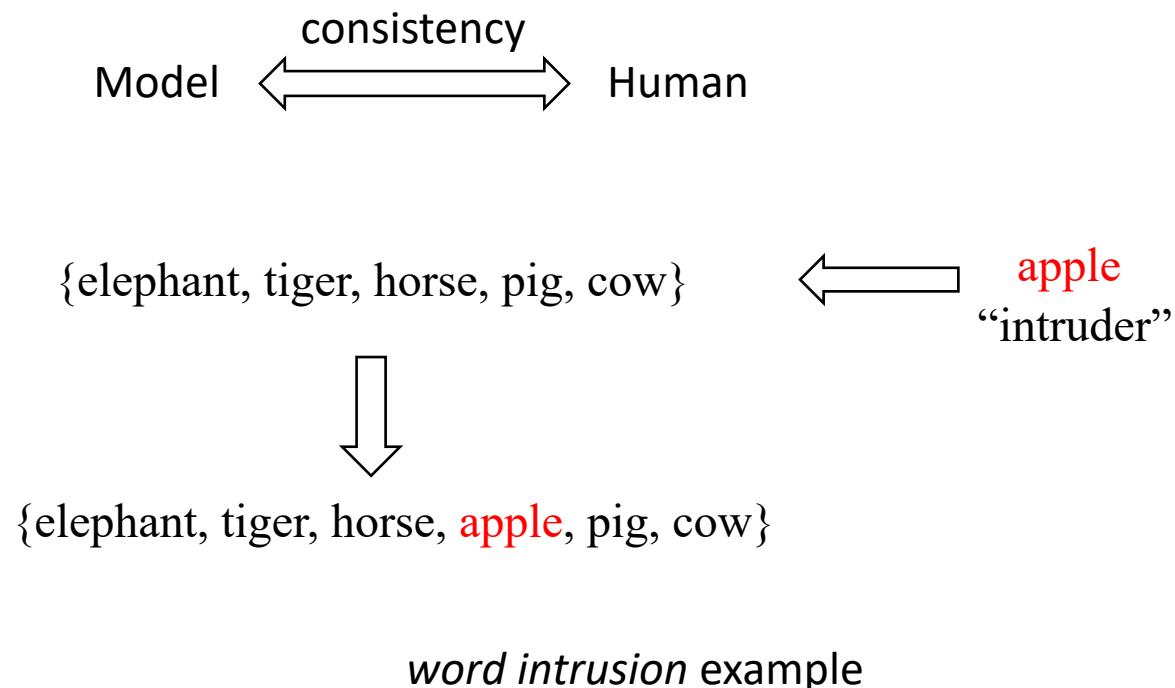


Thank you very much!



Appendix: Human interpretability

- Well-known lab studies
 - *Word intrusion* and *topic intrusion* tasks (Chang et al. 2009; Bao and Datta 2014; Palese and Piccoli 2020)
 - Amazon Mechanical Turk (MTurk)



Appendix: Human interpretability

- Well-known lab studies
 - *Word intrusion* and *topic intrusion* tasks (Chang et al. 2009; Bao and Datta 2014; Palese and Piccoli 2020)
 - Amazon Mechanical Turk (MTurk)

1/20

Not every model is able to learn sample-by-sample or incrementally. However, in scikit-learn, there're some models which have `partial_fit` method: Incremental fit on a batch of samples ... You can just search for methods name in sklearn's documentation ... Also, you can use Random Forest and set number of samples (or sample ratio) per tree is small to fit the memory. Or use Dask and Dask ML to fit your data in memory.

model	data	train	test	feature	value	predict	class
memory	access	address	map	block	bit	store	device
inform	method	news	govern	descript	like	medium	office
problem	number	algorithm	time	sum	log	function	frac

topic intrusion example

