

Topic Modeling and Text Mining

Workshop Session at 2025 Rawley Conference

Pei-Ying Chen

November 6, 2025

About Me

Current Position

Assistant Professor, Digital Scholarship & Sciences Librarian, Research Partnerships, University Libraries, University of Nebraska-Lincoln (August 2025–)

Education

- **Ph.D., Information Science** (minor: Sociology), Indiana University Bloomington, IN, USA
- **M.S., Applied Statistics**, Indiana University Bloomington, IN, USA
- **M.A., History** (concentration: Science, Technology, Society), National Tsing Hua University, Hsinchu, Taiwan
- **B.A., Foreign Languages and Literatures** (double major: Sociology), National Taiwan University, Taipei, Taiwan

Agenda

Lecture (20–30mins)

- Gentle Intro to Text Mining and Topic Modeling
- Historian's Applications & Critique: Jo Guldi
- Comparing Topic Modeling Methods: LDA / BERTopic / TopicGPT
- UNL Libraries (Digital) Collections
- Important Reminders & Pitfalls
- References, Resources, & Further Readings

Hands-on Session (50–60 mins)

- A small corpus built from the *The Daily Nebraskan*: Editorial, Letter to the Editor, Opinion
- Run topic modelings using `topicmodels` for LDA & `stm` for Structural Topic Modeling

Lecture

Last update: November 9, 2025

Welcome & Warm-Up

- Who here has used **R** or **Python** before?
 - Beginner / never used
 - Some hands-on
 - Intermediate
 - Expert
- Have you heard of **text mining**?
- Have you ever run **topic modeling**?

What is Text Mining?

- Computational methods for analyzing **large collections of text**
- Common goals
 - Patterns
 - Themes/Topics
 - Sentiment
 - Named entities

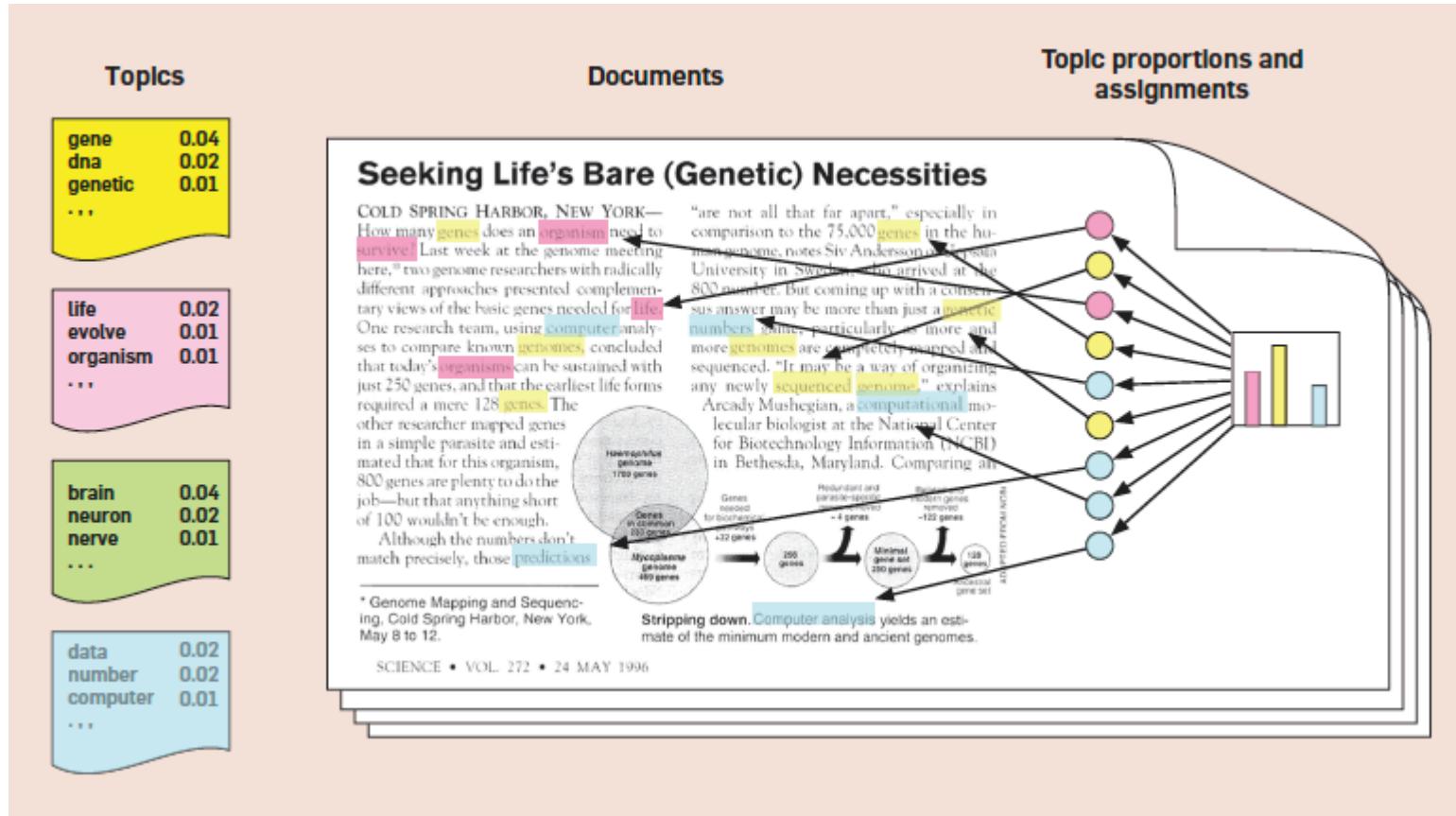
What is Topic Modeling?

- A family of techniques that analyze “bags” or groups of words together
- Captures **broader context** instead of just word counts
- Helps find **latent themes** in text collections
- Use cases
 - Literature & historical corpora
 - News & social media
- cf. cluster analysis: 1 document <> 1 cluster
 - Each document may contain **multiple topics** with different probabilities
 - Use iterative Bayesian algorithms

Latent Dirichlet Allocation (LDA)

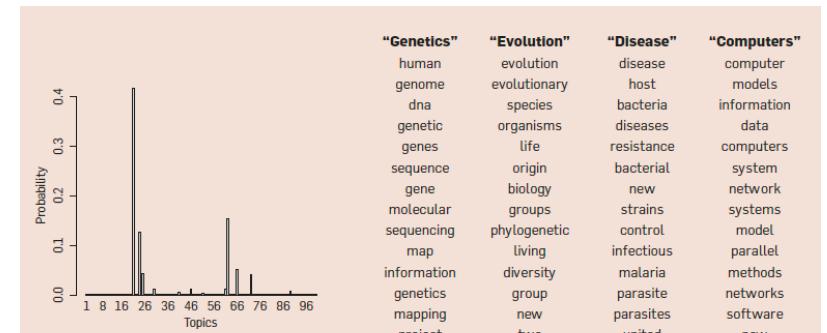
- Most common algorithm for topic models => infer the hidden topic structure using the observed documents (Blei, 2012)
- Principles (Silge & Robinson, 2017)
 - Every document = a **mixture of topics**
 - Every topic = a **mixture of words** => words can be shared between topics
- Outputs
 - **Top words** for each topic
 - e.g. Topic 1 (Genetics): human, genome, dna, genetic, gene; Topic 2 (Evolution): evolution, evolutionary, species, organisms, life; Topic 3 (Disease): disease, host, bacteria, deceases, resistance
 - **Topic probabilities** for each document
 - Document A => 0.57 Topic 1, 0.35 Topic 2, 0.08 Topic 3

An Illustration



(Blei, 2012)

Last update: November 9, 2025



Limitations

- “**Bag of words**” assumption: ignore word order and semantic meaning
- “**Reading tea leaves**” problem ([Chang et al., 2009](#)): # topics and interpretation are subjective/arbitrary
- **Needs human judgment:** validation and expertise required

Historian's Application and Critique: Jo Guldi

Parliament's Debates about Infrastructure (Guldi, 2019)

- Applies topic modeling and dynamic topic modeling to synthesize the historical record, with different degrees of granularity for analysis: 4 / 10 / 100 / 500 / 1000 topics
- **80/20 rule:** 80% of the findings conform to what scholars already know, but the remaining 20% may reveal new patterns
 - e.g., the word “Shannon” appeared earlier in a discussion of river improvement and drainage, than the more familiar Thames in the same topic (p. 7)
- Institutional divides: Army vs. Navy; Conservatives vs. Liberals
- Geographical and temporal/dialectical shifts

TABLE 10

PUNITIVE AND REDEMPTIVE LANDSCAPES OF TECHNOLOGY IN COLONIAL DUBLIN, BELFAST, CORK, AND EDINBURGH, FROM "IMPERIAL METROPOLE" TOPIC (LIGHT SHADING = PUNITIVE LANDSCAPES; DARK SHADING = REDEMPTIVE LANDSCAPES)

<i>1800</i>	<i>1810</i>	<i>1820</i>	<i>1830</i>	<i>1840</i>	<i>1850</i>	<i>1860</i>	<i>1870</i>	<i>1880</i>	<i>1890</i>	<i>1900</i>
dublin-	dublin-	dublin-	dublin-	dublin-	dublin-	dublin-	dublin-	dublin-	dublin-	dublin-
insurrect-		institute-		ships-	hospital-	society-	edinburgh-	university-	university-	college-
rebellion-		society-		college-	mail-	edinburgh-	college-	castle-	college-	field-
street-		chamber-		corporation-	hospitals-	garden-	belfast-	corporation-	green-	green-
rebel-		school-		university-	college-	institute-	science-	barracks-	corporation-	university-
murder-		police-		gaol-	packet-	castle-	newspaper-			college-
depot-					hospital-	belfast-	cork-			castle-
castle-										museum-
riot-										

The Dangerous Art of Text Mining (Guldi, 2023)

[W]ithout help from the humanities, data science can distort the past and lead to perilous errors.

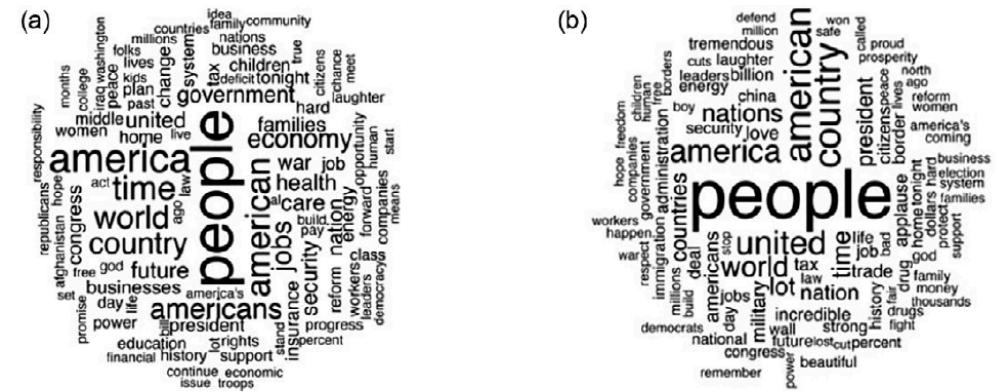
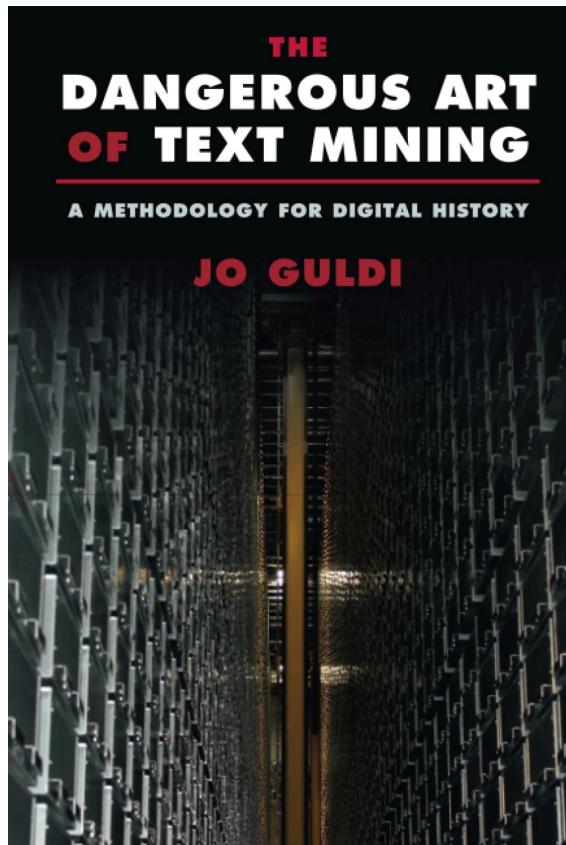
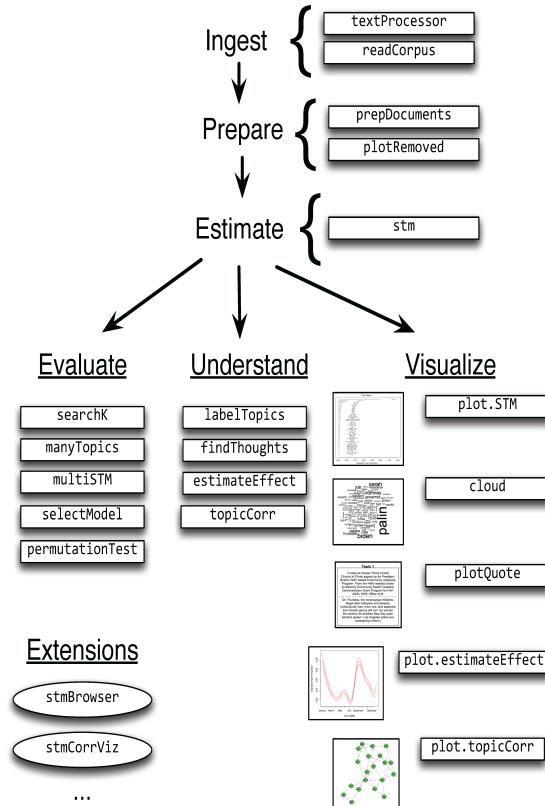


Figure 0.2 Data-driven visualizations that obscure rather than illuminate meaningful differences: Wordclouds for the official speeches of a) Barack Obama and b) Donald Trump during their presidencies. Source: “Presidential Speeches” dataset, <https://millercenter.org/the-presidency/presidential-speeches>. Visualized by author.

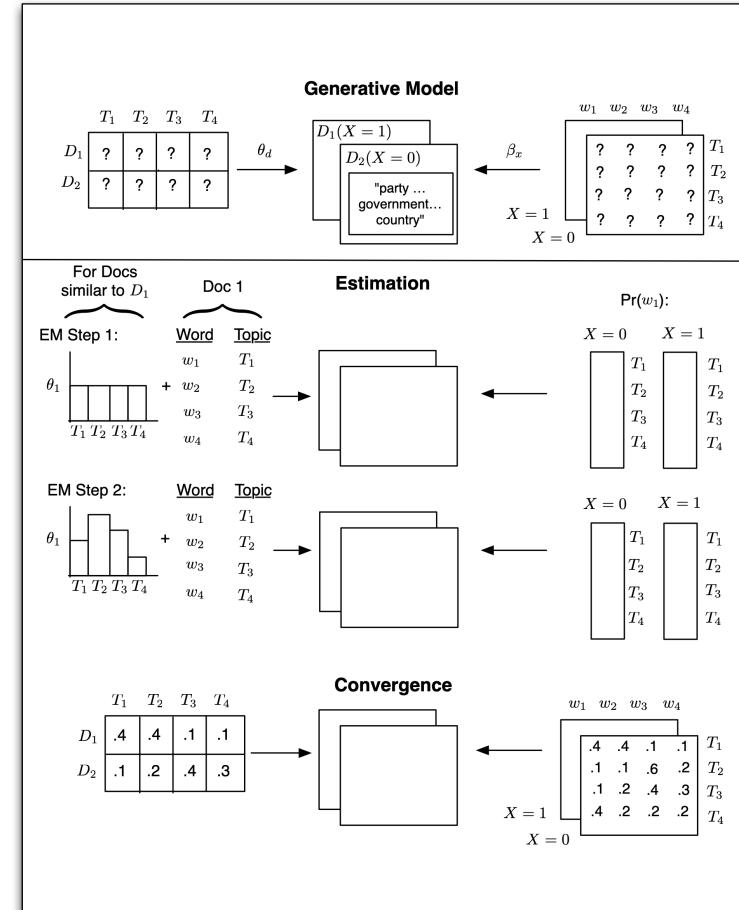
Comparison of Topic Modeling Methods

Feature / Method	LDA (Blei et al., 2003)	BERTopic (Grootendorst, 2022)	TopicGPT (Pham et al., 2024)
Approach	Probabilistic topic model (Bayesian)	Transformer-based embeddings + clustering + class-based TF-IDF	LLM / GPT-based prompting and summarization
Best for	Longer text (hundreds of words per document)	Short text (tweets, headlines), mixed-length corpora	Flexible lengths
Input Representation	Bag of words	Dense semantic embeddings	Natural-language prompts
Output	Topic-word distributions; document-topic probabilities	Human-readable topic labels automatically; cluster assignments	Natural-language topic labels; coherent summaries
Interpretability	Requires human reading of top words	Usually more coherent topics	Often most interpretable
Computational Cost	Low-moderate	Moderate; depends on embedding model	High; relies on LLM inference
Libraries / Tools	topicmodels , stm (R), gensim , scikit-learn (Python), Mallet	bertopic (Python)	custom LLM pipelines
Parameter Tuning	Choose # of topics k	HDBSCAN clustering parameters, dimensionality reduction	Prompt templates; LLM model choice
Advantages	Proven, transparent, reproducible, lightweight	Works well for short texts; coherent topics; automatic naming	Extremely interpretable; strong for messy or domain-specific corpora
Limitations	Bag-of-words misses semantics; topics can be vague	Can struggle with very small corpora or noisy embeddings	Expensive, proprietary models, reproducibility issues

Structural Topic Modeling with `stm`



Flowchart



Diagram

UNL Libraries Collections

- [Digital Humanities Projects](#) @Center for Digital Research in the Humanities (CDRH)
 - [Nebraska Newspapers](#) - full-text available but needs considerable processing
 - [Journals of the Lewis and Clark Expedition Online](#) - high-quality full-text data with comprehensive annotation
- [Notable Collections](#) @Archives & Special Collections

Things to Keep in Mind

- Constructing a **clean corpus** can be labor-intensive
- Topic modeling as a “tool for reading” & a way of making educated guesses
- Human reading is required

References

- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
<https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022. <https://jmlr.org/papers/v3/blei03a.html>
- Chang, J., Gerrish, S., Wang, C., Boyd-graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.
https://proceedings.neurips.cc/paper_files/paper/2009/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. <https://doi.org/10.48550/arXiv.2203.05794>
- Guldi, J. (2019). Parliament’s debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change. *Technology and Culture*, 60(1), 1–33. <https://doi.org/10.1353/tech.2019.0000>
- Guldi, J. (2023). *The dangerous art of text mining: A methodology for digital history*. Cambridge University Press.
<https://doi.org/10.1017/9781009263016>
- Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2024). TopicGPT: A prompt-based topic modeling framework. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2956–2984). Association for Computational Linguistics.
<https://doi.org/10.18653/v1/2024.naacl-long.164>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91, 1–40. <https://doi.org/10.18637/jss.v091.i02>
- Silge, J., & Robinson, D. (2017). *Text mining with R*. O’Reilly Media. <https://www.tidytextmining.com/>
Last update: November 9, 2025

Resources & Further Readings

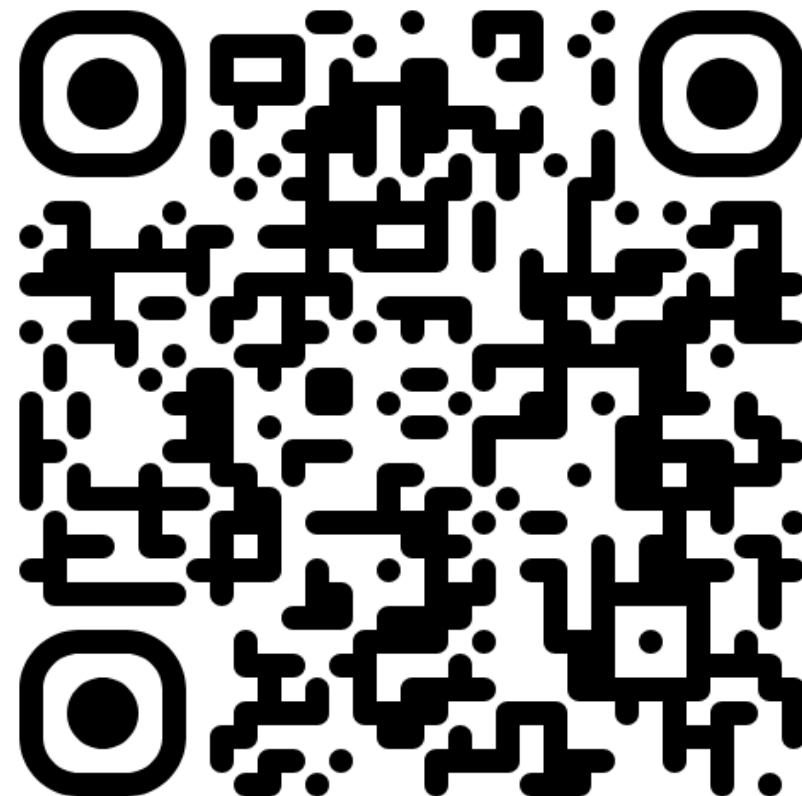
- An Introduction to Text Analysis & Topic Models by Professor Chris Bail of Duke University @The Summer Institutes in Computational Social Science (SICSS)
- Guldi, J. (2024). The revolution in text mining for historical analysis is here. *The American Historical Review*, 129(2), 519–543. <https://doi.org/10.1093/ahr/rhae163>
- Grimmer, J., Robert, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.
<https://press.princeton.edu/books/hardcover/9780691207544/text-as-data>
- Karsdorp, F., Kestemont, M., & Riddell, A. (2021). *Humanities data analysis: Case studies with Python*. Princeton University Press. <https://press.princeton.edu/books/hardcover/9780691172361/humanities-data-analysis>
- Story, D., & Guldi, J. (2025, January 15). Jo Guldi on text mining, AI, and digital history [Broadcast].
<https://www.historians.org/podcast/jo-guldi-on-text-mining-ai-and-digital-history/>
- UC Berkeley Social Science Matrix. (2023, April 12). Jo Guldi: Towards a practice of text mining to understand change over historical time [Video recording]. https://www.youtube.com/watch?v=JI00B_KqH6U

Hands-on Session

Last update: November 9, 2025

Install R, RStudio & Download Code

- R: <https://cloud.r-project.org/>
- RStudio: <https://posit.co/downloads/>
- GitHub Repo: <https://github.com/peiychen/textMining>



Shameless Promotion



GIS Day 2025

- **Time:** Monday, November 10, 2025 1–2 pm
- **Place:** Peterson Room 211, Love Library (City Campus)

Main Activities

- I will present a project titled “**Exploring Nebraska’s Farmers Markets Using an R Shiny Dashboard**,” demonstrating how interactive data visualization can reveal insights into local products, growers, and their connections to space/place.
- Kurt J. Elder, Information Intelligence and Special Project Coordination professional with the City of Lincoln, will showcase the **Healthy Food Access 2025** map, which identifies neighborhoods with limited access to healthy food and helps guide targeted interventions.
- Recognition of this year’s Map Competition winners, with posters on display during the reception.

Light refreshments will be provided. | More details: <https://unl.libguides.com/unlgisday>

Last update: November 9, 2025

Thank You! Questions?

Pei-Ying Chen | pchen12@unl.edu | LLS 218D