Scientific Discovery as Link Prediction in Influence and Citation Graphs

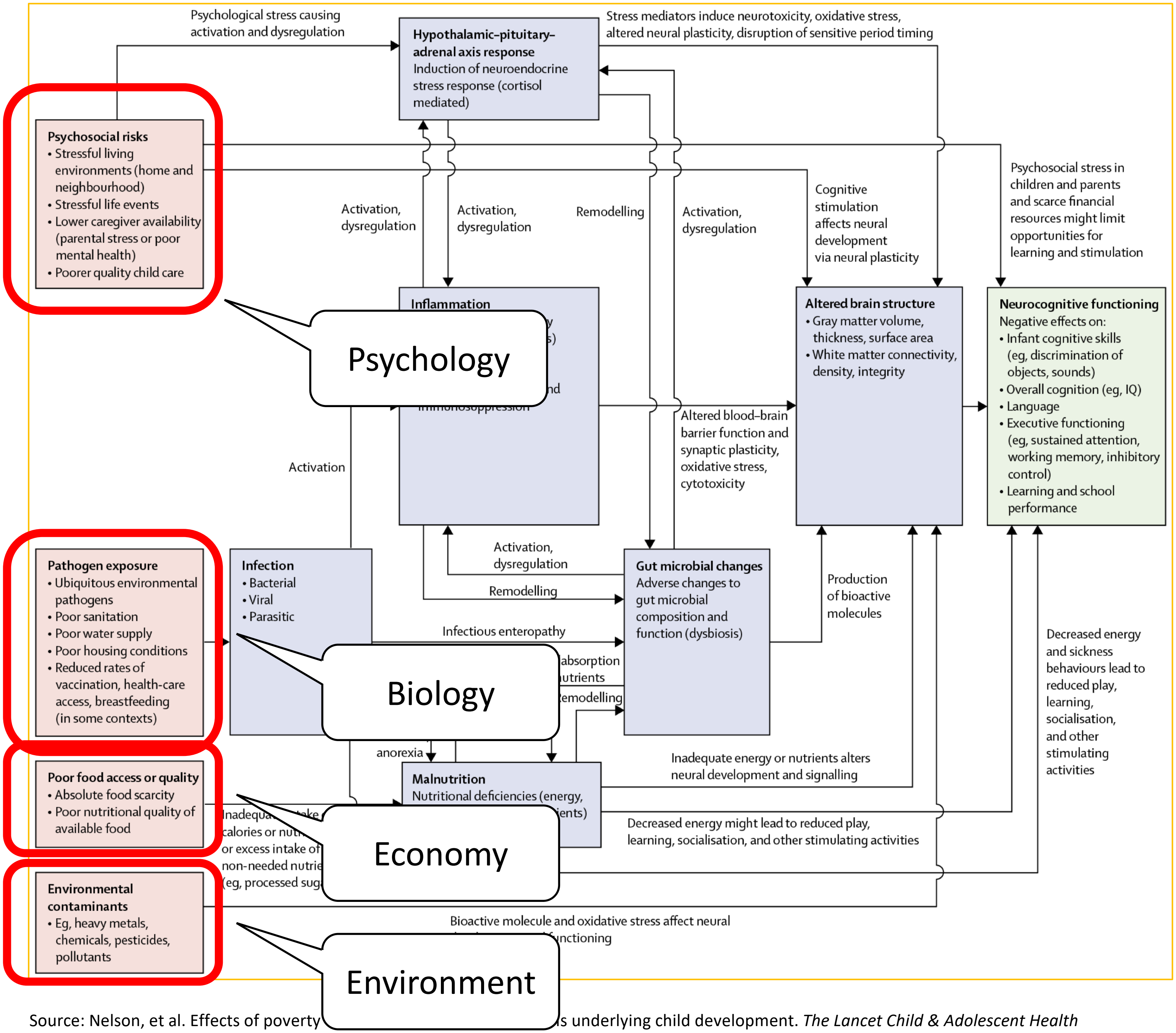
Fan Luo, Marco Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu

Text Graphs Workshop, June 6, 2018

THE UNIVERSITY OF ARIZONA

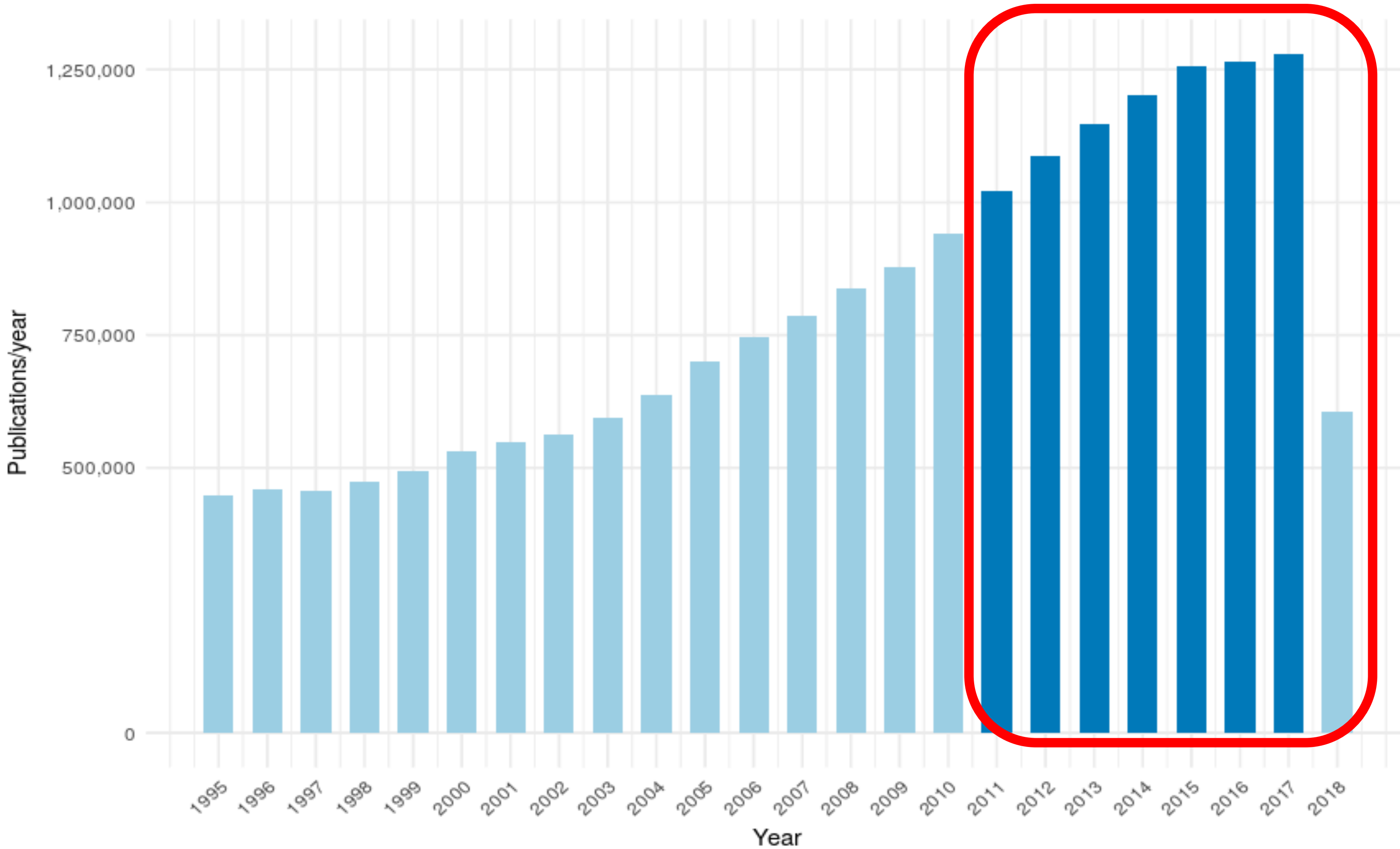
Computational Language Understanding

Background



Source: Nelson, et al. Effects of poverty … s underlying child development. *The Lancet Child & Adolescent Health* https://doi.org/10.1016/S2352-4642(17)30024-X

Publications indexed by PubMed each year since 1995



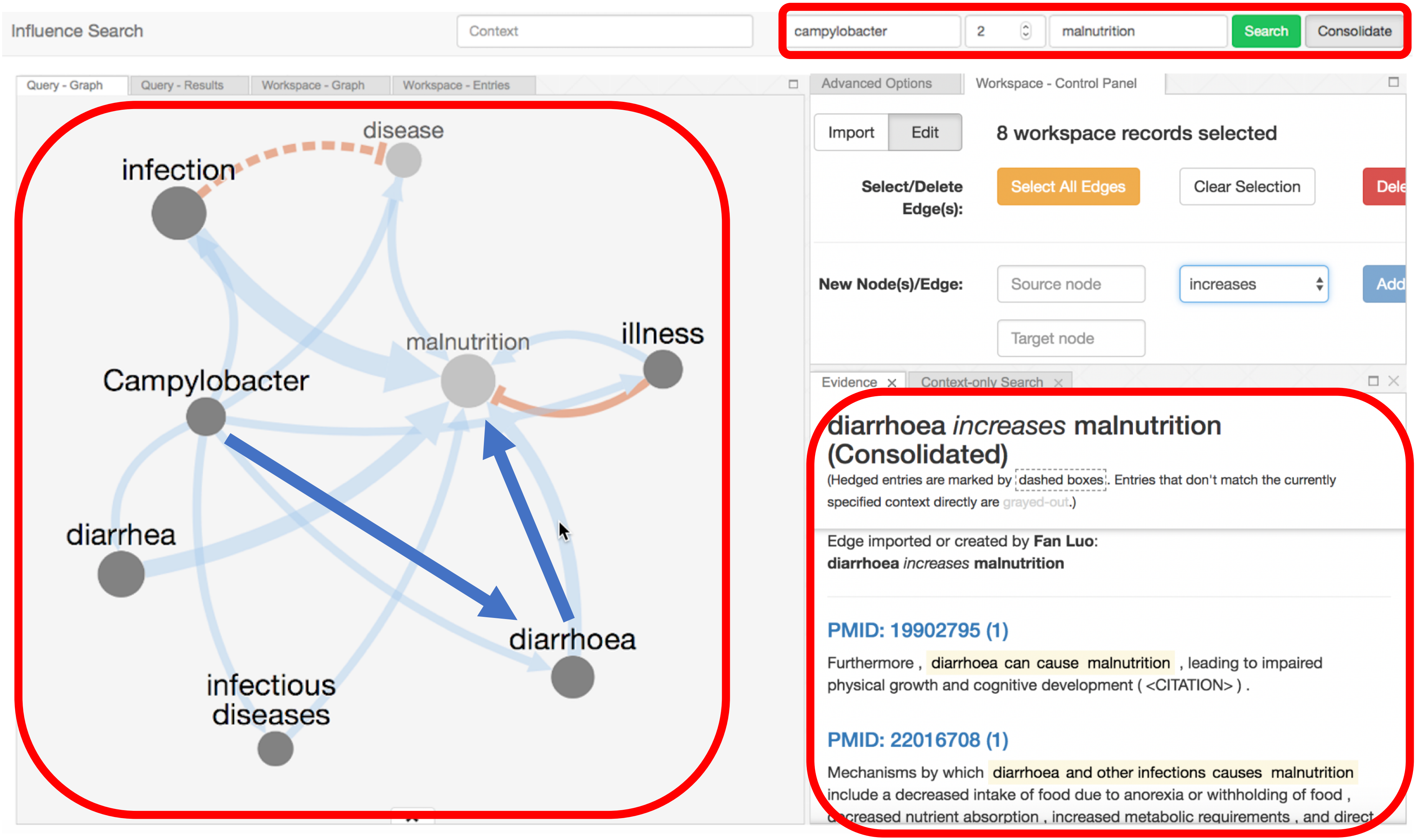
If humans cannot keep up, machines must help!

Previous Work: Influence Search

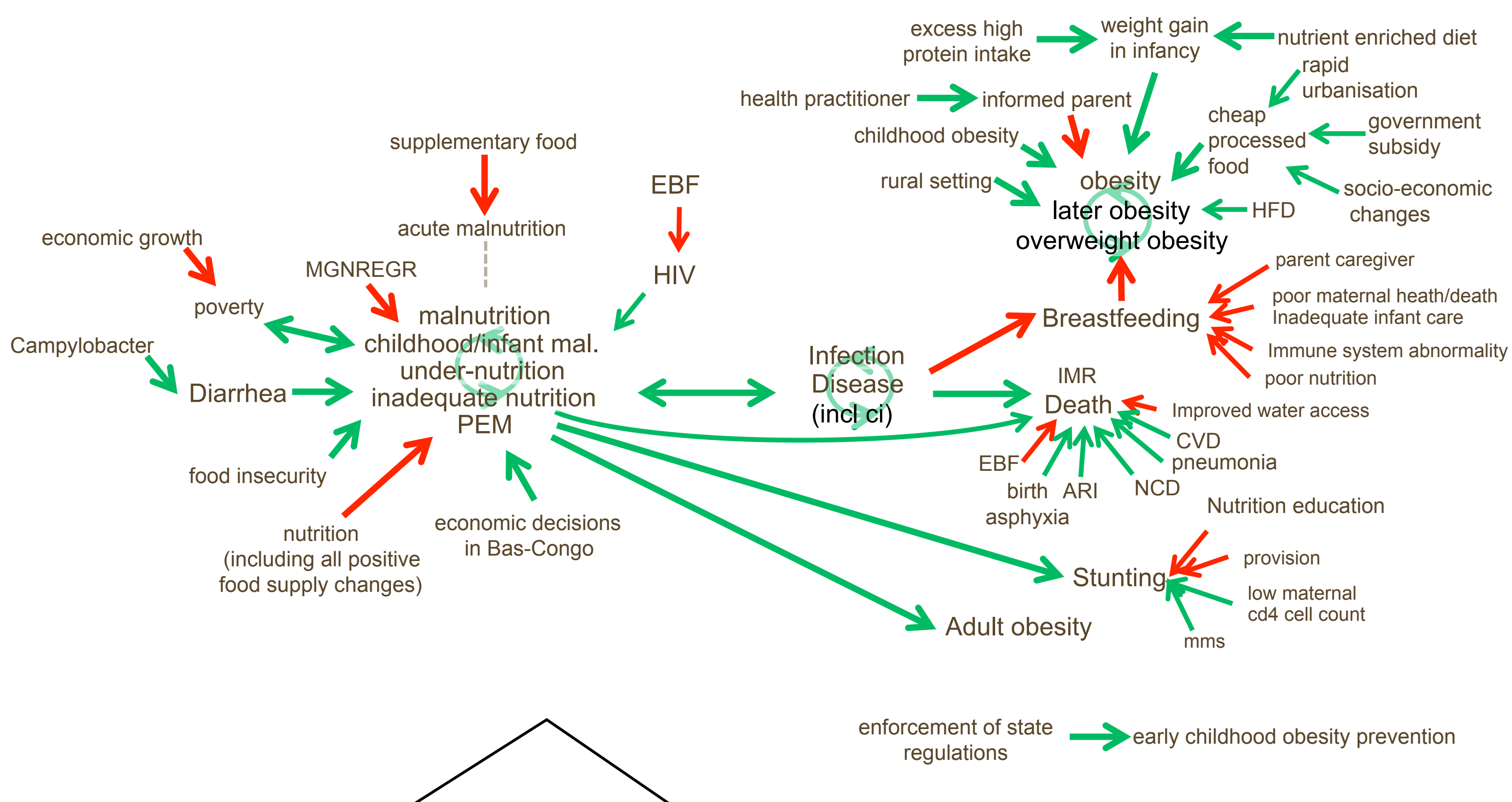
- We implemented a machine reading system focused on **influence** statements in **children's health literature**

1. Large-scale automated reading with Reach discovers new cancer driving mechanisms.
2. Swanson linking revisited: Accelerating literature-based discovery across domains using a conceptual influence graph.

Influence Search



Use Case



Constructed in 2 days (human + machine);
Normally, it takes 1 month (human alone).

Model courtesy of: Lyn Powell, HBGDki-qPM team

Motivation

Past vs. Future

- This system can only search **past**, published facts
- No information about what comes **next in science**…

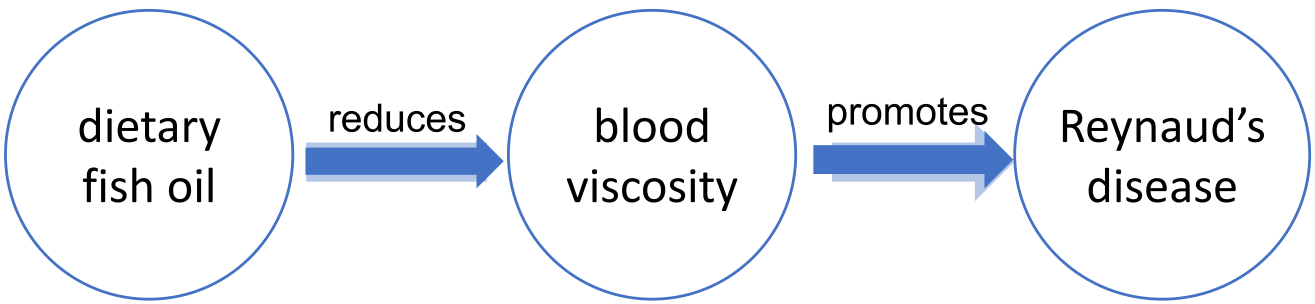
Definition

- **White spaces** in science
  - + Topics that are insufficiently studied, but
  - + May lead to important scientific discoveries

Our Contributions

1. White space discovery = **link prediction over the influence graph**

- Predict whether an influence link will be added to the graph



- Binary classification task:
  - **positive**, if the influence relation will be added to the influence graph in the future;
  - **negative,** otherwise

Swanson, D.R. Undiscovered public knowledge. The Library Quarterly, 56 (2), 1986.

Our Contributions

1. White space discovery = **link prediction over the influence graph**

   - Predict whether an influence link will be added to the graph

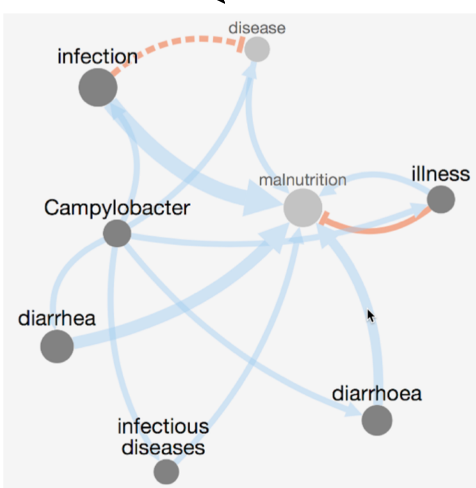
- Binary classification task:
  - positive, if the influence relation will be added to the influence graph in the future;
  - **negative,** otherwise

Swanson, D.R. Undiscovered public knowledge. The Library Quarterly, 56 (2), 1986.

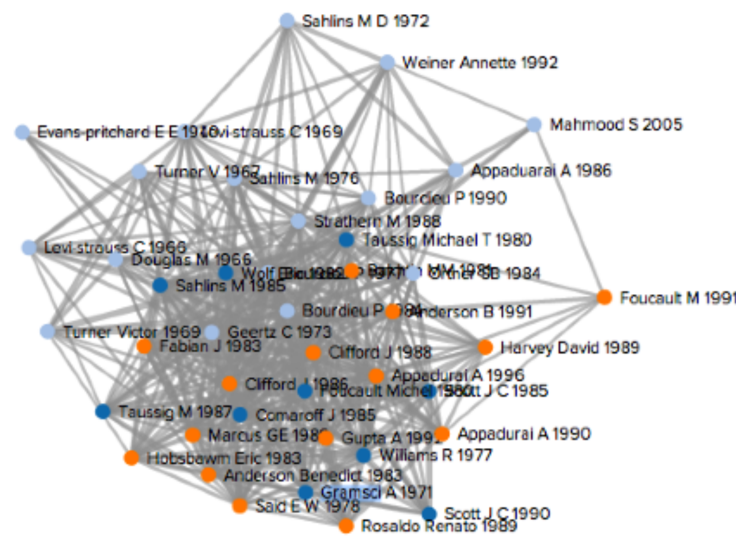
Our Contributions

2. Features from multiple graphs!

Influence graph (to understand influence connectivity)



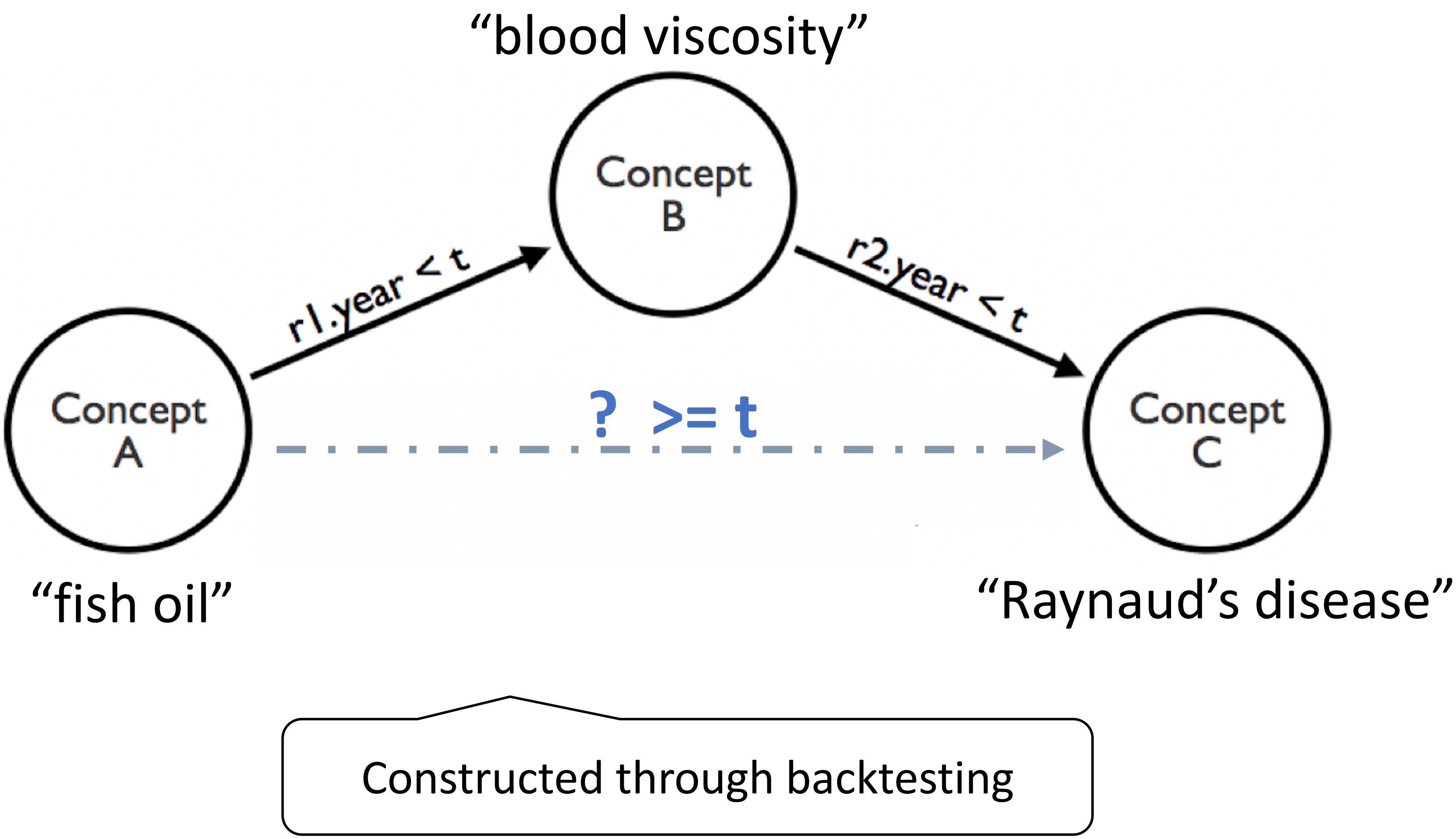
Citation graph (to understand community overlap)



Dataset

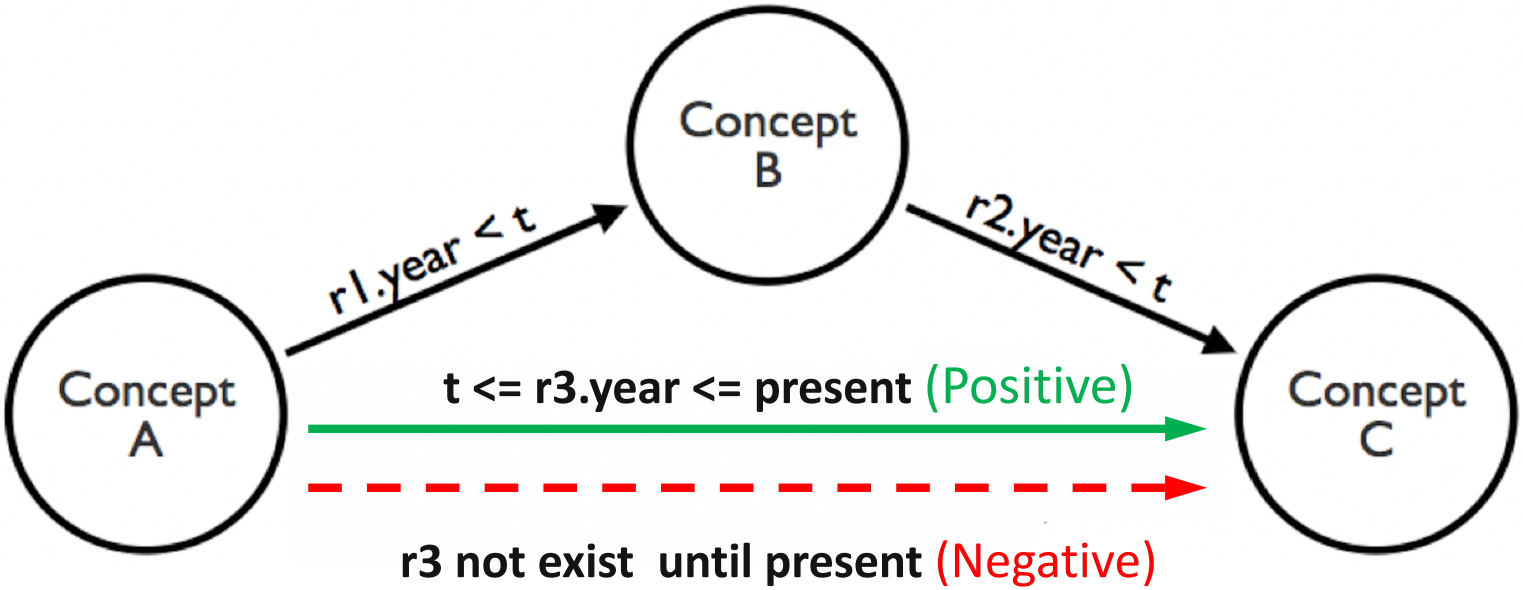
Complication: No "Back to the Future"

Dataset



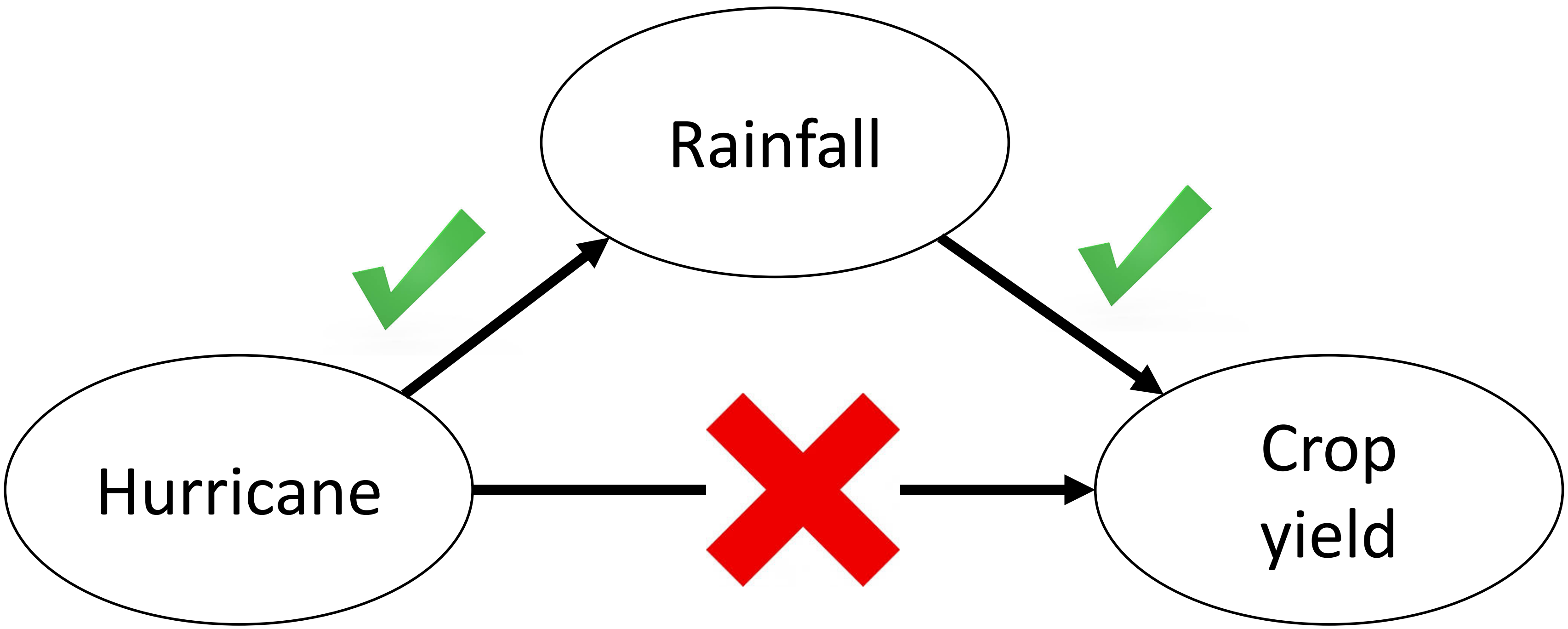
"blood viscosity"
Concept B
r1.year < t
r2.year < t
Concept A
? >= t
Concept C
"fish oil"
"Raynaud's disease"
Constructed through backtesting

Dataset



Concept B
r1.year < t
r2.year < t
Concept A
t <= r3.year <= present (Positive)
Concept C
r3 not exist until present (Negative)

Note: Transitivity Generally Not True!



Missing information impacts non-linear models!

Dataset

t = 2012

	Positive Examples	Negative Examples
Training	3,011	164,551
Development	1,002	54,850
Testing	1,002	54,850

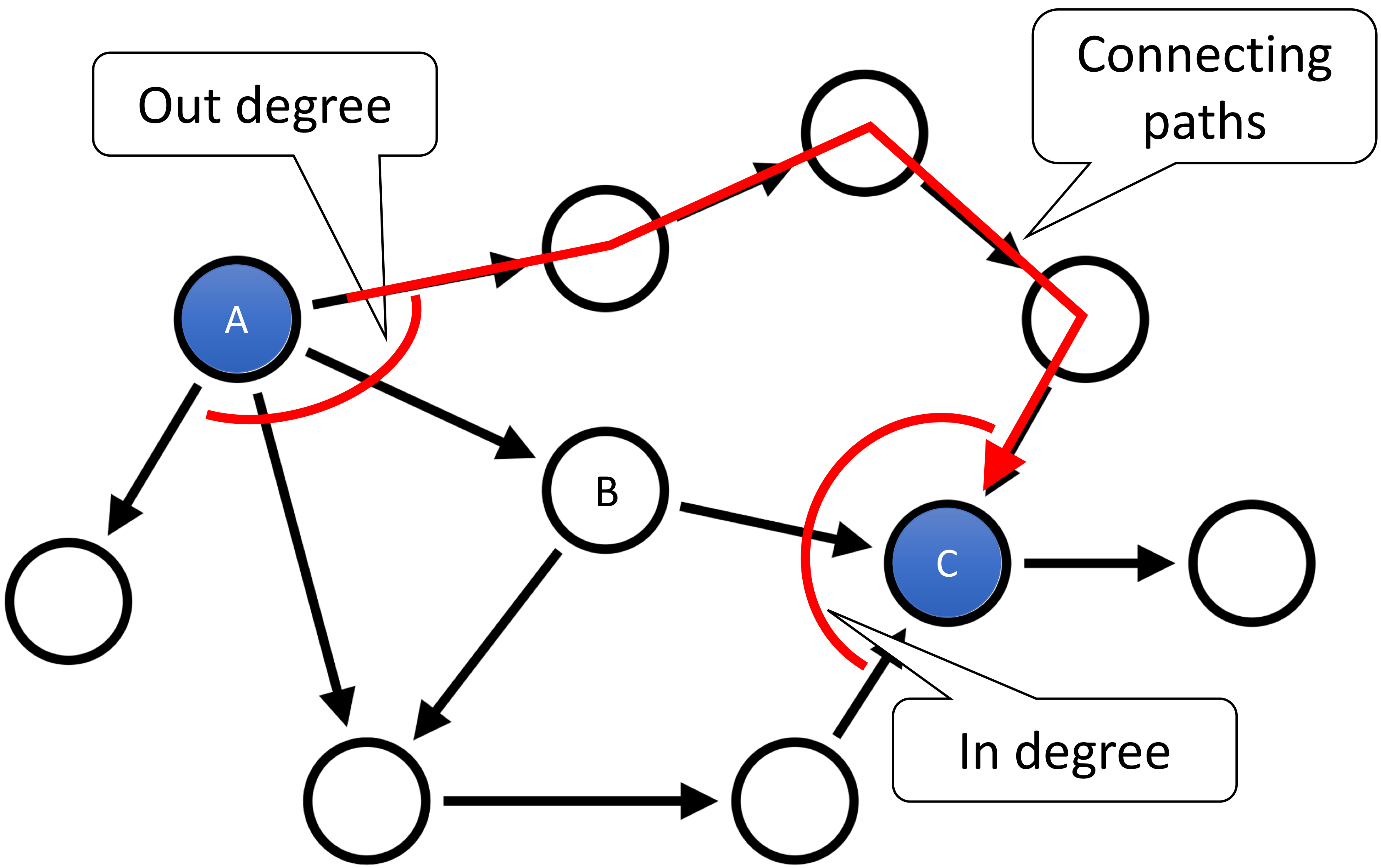
Features

- Extracted from two graphs
- Influence graph (influence relations between concepts)
  - 1,564,748 distinct nodes
  - Connected by 2,395,944 influence relations
- Citation graph (citations between papers)
  - 119K papers
  - 5,523,759 citation links

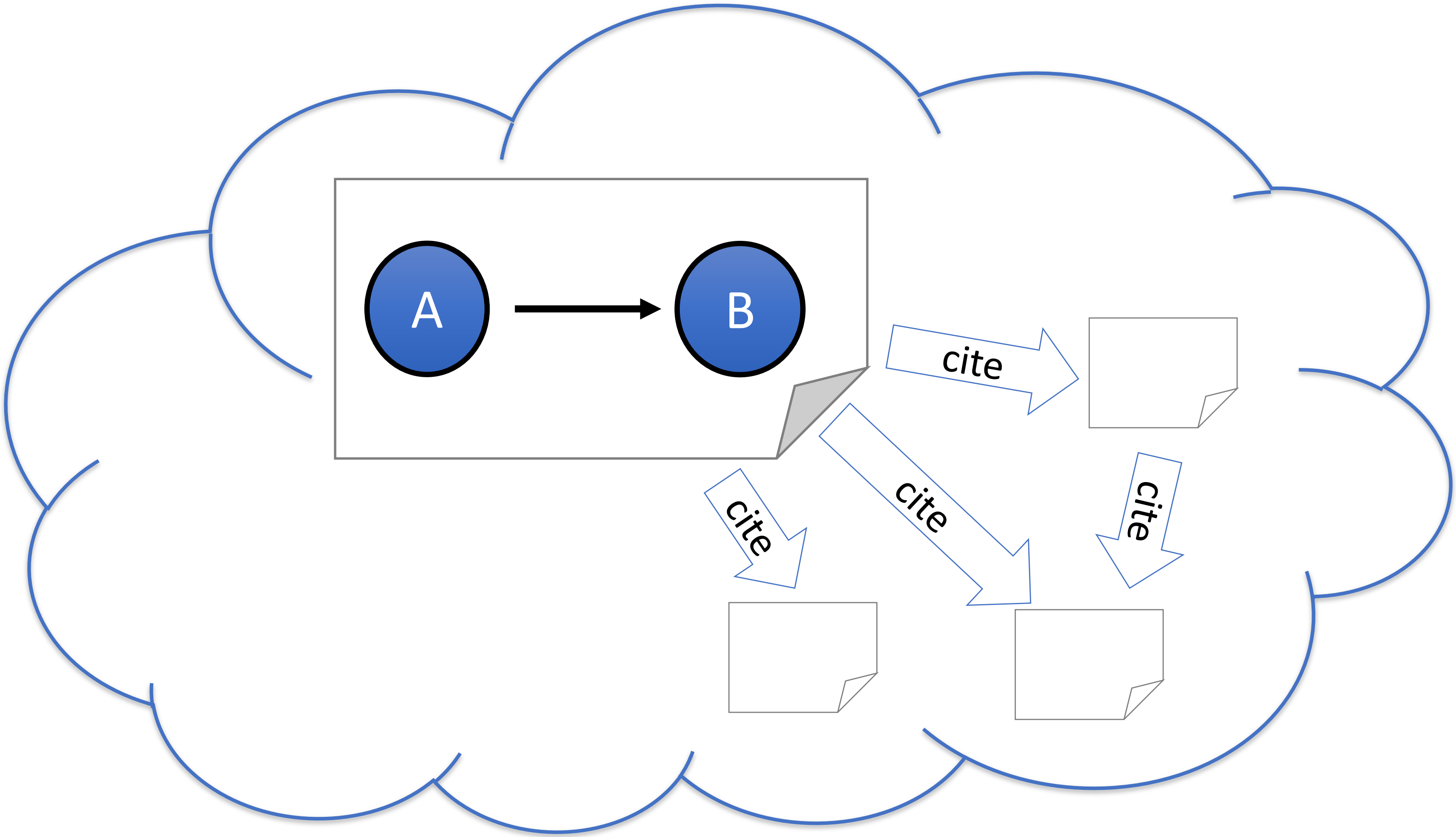
Feature groups

Feature Group	Intuition	From
Connectivity features	The more connected concepts are, the easier is to discover a relation between them	influence graph
Community-based features	The larger the intersection of communities containing the two influence statements, the easier it is to make the connection	citation graph
Information retrieval features	The more distinct a concept or an influence statement is, the harder it is to make a discovery around it	papers containing influence statements

Connectivity Features

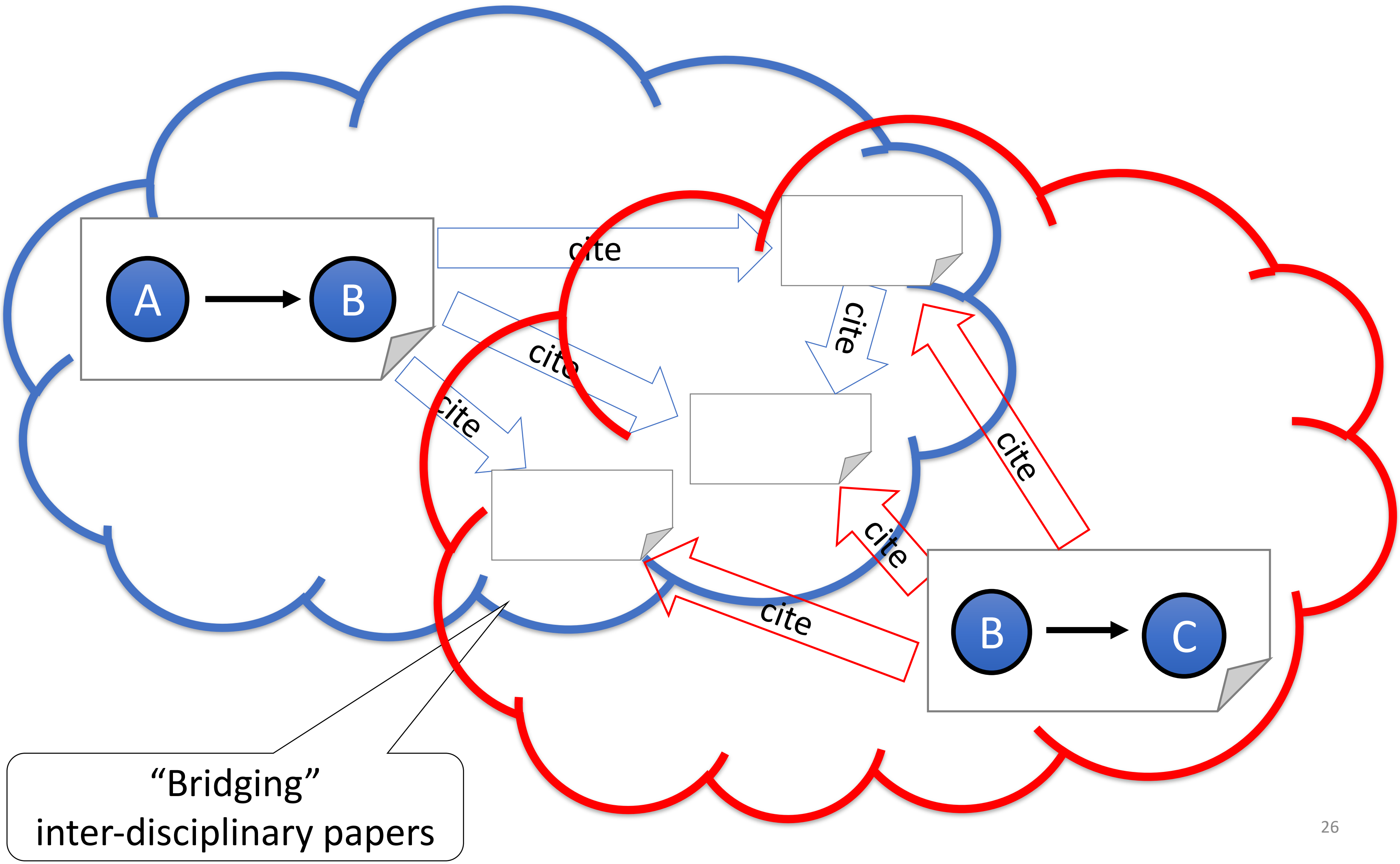


Community-based Features



The communities were detected using the Coda algorithm (Yang et al., 2014)

Community-based Features



Information Retrieval Features

- Inverse document frequency (IDF) score of lemmas in concept A
- IDF score of lemmas in concept B
- IDF score of lemmas in concept C

- Number of papers that mention A $\rightarrow$ B
- Number of papers that mention B $\rightarrow$ C

Evaluation

Evaluation Metrics

Unranked

- $$\text{Precision} = \frac{tp}{tp + fp}$$
- $$\text{Recall} = \frac{tp}{tp + fn}$$
- F1 = harmonic mean of P and R

$$F_1 = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Ranked

- P@10 = how many links predict in top 10 are correct
- MAP = mean average precision

Results

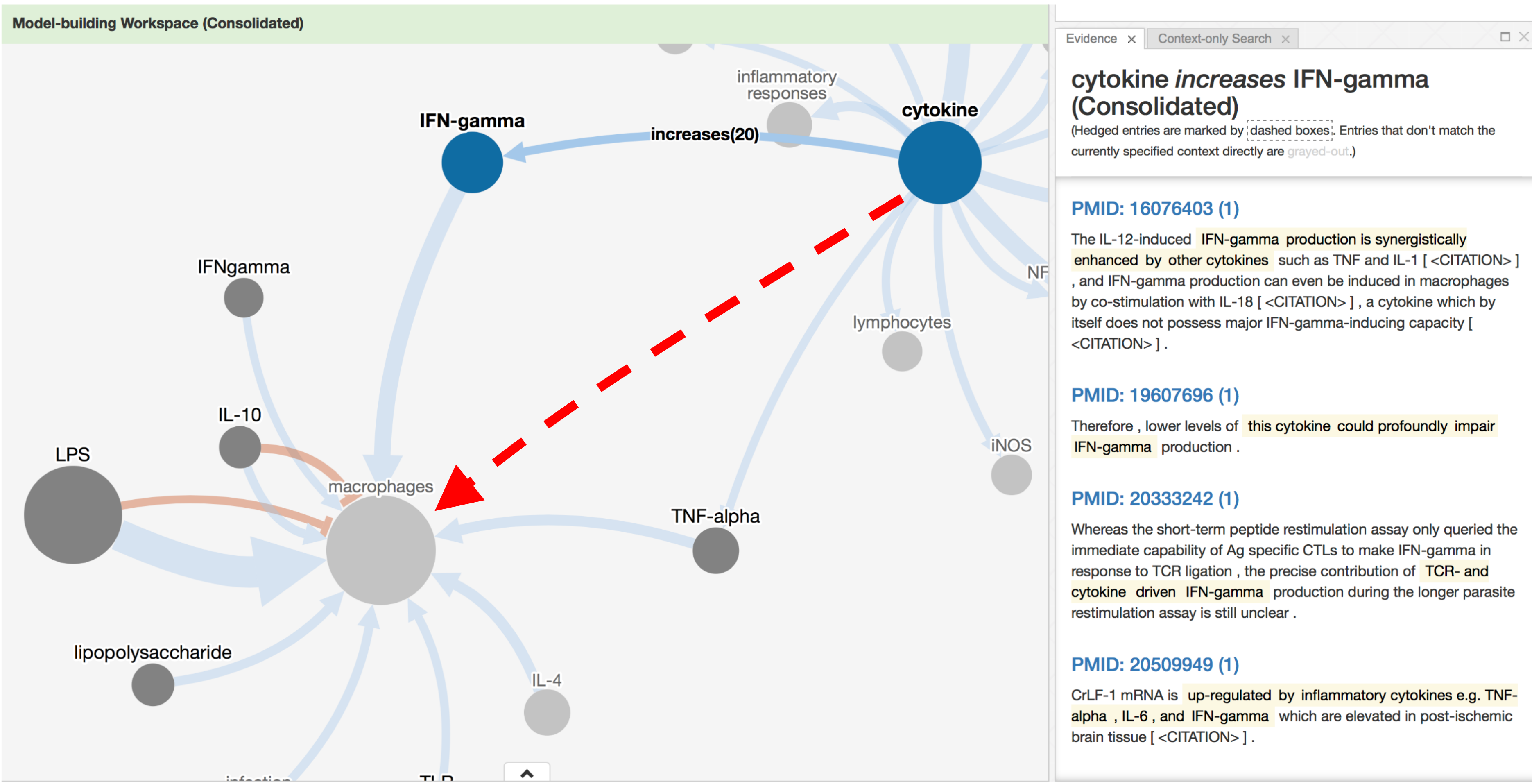
	F1	Precision	Recall	P@10	MAP
Baseline (random)	0.02	0.02	0.02	-	-
Baseline (all positive)	0.035	0.018	**1**	-	-
Neural Network	**0.27**	0.398	0.206	0.8	0.537
AdaBoost	**0.27**	**0.536**	0.178	**0.9**	**0.681**
Random Forest	0.23	0.244	**0.216**	0.5	0.309

All Feature Groups Help

Full model	0.268
— Connectivity features	0.2 - 0.246
— Community-based features	0.226 - 0.232
— Information retrieval features	0.216 - 0.248

F1 scores for feature ablation

What Does the System Predict?



Conclusions

- Novel strategy for the identification of white spaces in scientific knowledge
- Operates over real-world graphs of influence relations and citations
- F1 score of 27 points, and a mean average precision of 68%
- Important to
  Researchers: “What should I research next?”
  Program officers: “What should I fund next?”

Thank you!

Resource Available:

- Data and code
  https://github.com/clulab/releases/tree/master/textgraphs2018-discovery
- Influence search engine
  http://influence.clulab.org/

Fan Luo, Marco Valenzuela-Escárcega, Gus Hahn-Powell, Mihai Surdeanu
{fanluo, marcov, hahnpowell, msurdeanu}@email.arizona.edu

Acknowledgements

- Marco Valenzuela-Escarcega, Gus Hahn-Powell, and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. This work was funded by the Bill and Melinda Gates Foundation HBGDki Initiative.

Appendix

t (year)	positive	negative
2017	-	998,586
2016	319	979,709
2015	1,706	881,689
2014	3,767	696,406
2013	5,208	481,671
2012	**5,015**	**274,251**
2011	3,741	151,460
2010	2,448	73,226
2009	1,521	36,839
2008	782	16,372
2007	444	7,843

Table 1: Total number of positive and negative examples for different values of the threshold t.

All feature groups help

Removed Feature	F1
Full model	0.268
Connectivity Features	
— C_A.outdegree	0.234
— C_A.indegree	0.246
— C_C.outdegree	0.214
— C_C.indegree	0.2
— $C_{\text{inbetween}}$.outdegree	0.22
— $C_{\text{inbetween}}$.indegree	0.234
— $C_{\text{inbetween}}$.avg-idf	0.214
— $r_{\text{inbetween}}$.avg-seen	0.23
— shortest_path_count	0.222
— shortest_path_length	0.204
Community based Features	
— max P($p_{A\to B}$,$p_{B\to C}$) (c=100)	0.226
— min P($p_{A\to B}$,$p_{B\to C}$) (c=100)	0.228
— avg P($p_{A\to B}$,$p_{B\to C}$) (c=100)	0.232
— max P($p_{A\to B}$,$p_{B\to C}$) (c=300)	0.232
— min P($p_{A\to B}$,$p_{B\to C}$) (c=300)	0.23
— avg P($p_{A\to B}$,$p_{B\to C}$) (c=300)	0.232
— Jaccard($\mathbf{p}_{A\to B}$,$\mathbf{p}_{B\to C}$)	0.226
— Inter-citation ratio	0.23
Information retrieval Features	
— C_A.idf	0.248
— C_B.idf	0.216
— C_C.idf	0.22
— r1.seen	0.226
— r2.seen	0.228