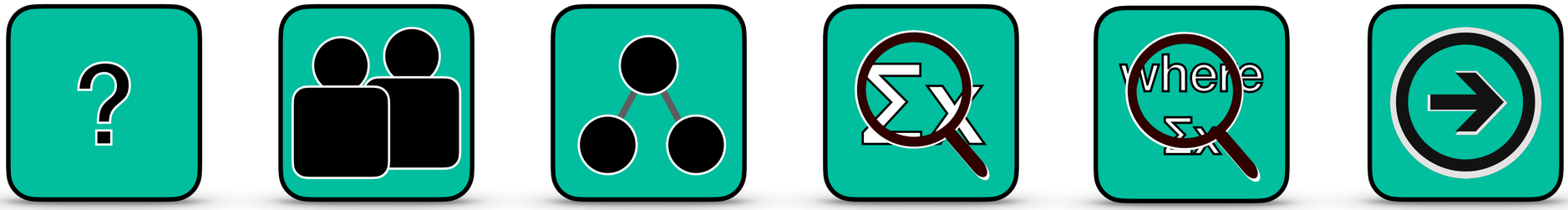


Mathematical Information Retrieval: Searching with Formulas and Text

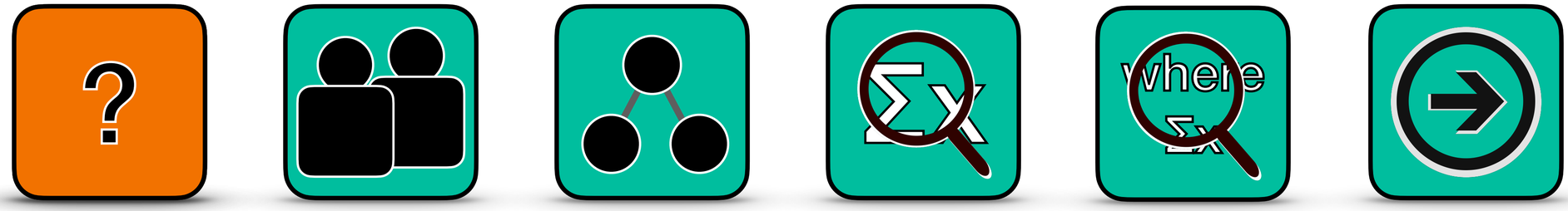
Richard Zanibbi

Department of Computer Science
Rochester Institute of Technology

dprl
Document and Pattern Recognition Lab

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE



Mathematical Information Retrieval:

Some Preliminaries

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

Mathematical Information

Q. When is information in documents *mathematical*?

A. When one of the following is true:

Focus of this talk

1. Defines or describes **computations and their mathematical properties**

   e.g., operations, arguments, their values or types, relationships to problems and/or models

2. Defines or describes **mathematical concepts**

   e.g., theorems, lemmas, ‘inverse’, problem and model formalization (e.g., information retrieval)

3. Defines or describes **relationships between concepts and/or computations**

   e.g., **proofs**, derivations, explanations for and expositions on interactions of model properties (e.g., for retrieval)

Notation vs. Information

Q. What information does a formula carry?

$$N$$

What does this formula represent?

Notation vs. Information

Q. What information does a formula carry?

$$\log N - \log n_i$$

What does this compute?

Notation vs. Information

Excerpt, from *Understanding Inverse Document Frequency* (Robertson, 2004)

2 The basic formula

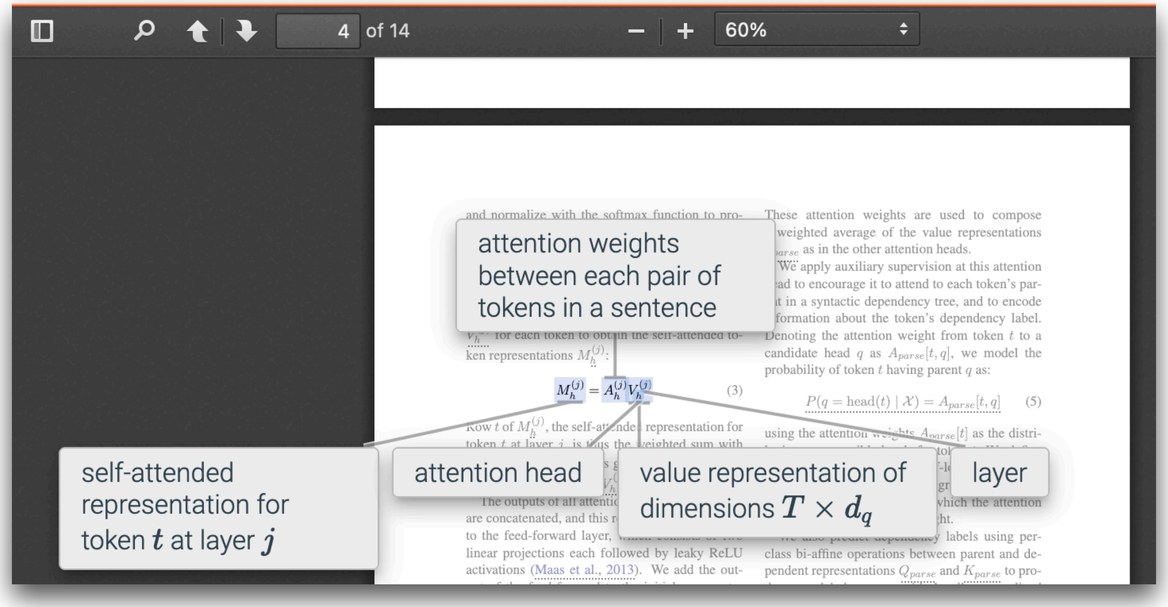
Assume there are N documents in the collection, and that term t_i occurs in n_i of them. (What might constitute a 'term' is not of concern to us here, but we may assume that terms are words, or possibly phrases or word stems. 'Occurs in' is taken as shorthand for 'is an index term for', again ignoring all the difficulties or subtleties of either automatic indexing from natural language text, or human assignment of index terms.) Then the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

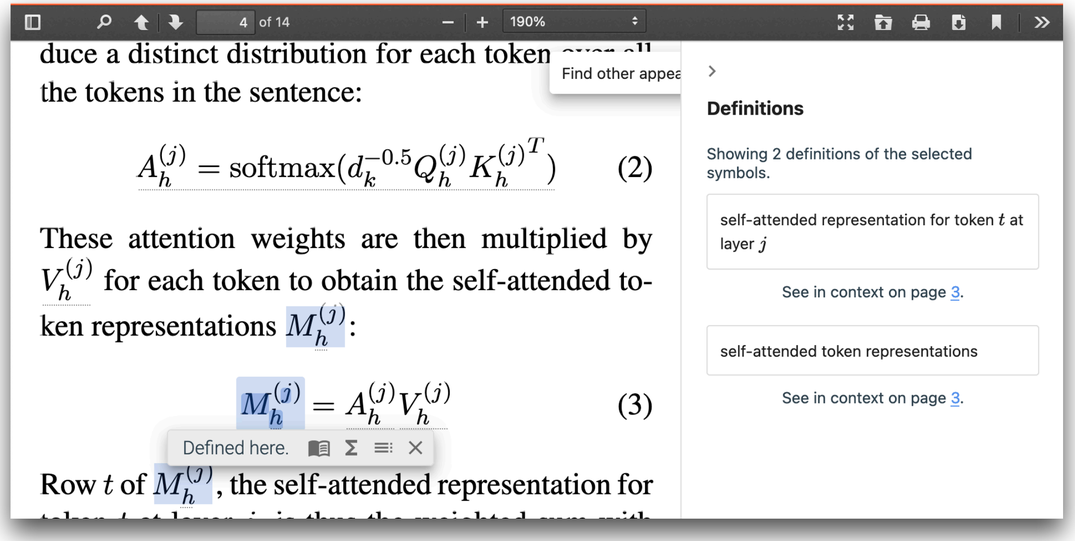
$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

Surrounding text and other **context** needed to interpret formulas

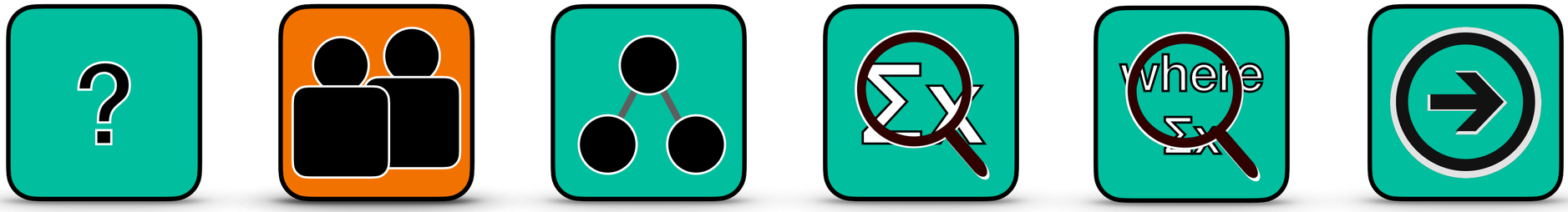
Navigation Aids for Math in Technical Documents

The ScholarPhi System (Head, Hearst, et. al, CHI 2021)





Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols.
Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst, ACM CHI 2021



The User Perspective:

Expert and Non-Expert Use Cases

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

A Study of Mathematical Experts

Zhao et al., JCDL 2008

Studied a small group of professors, graduate students and librarians
@ Math Department of the National University of Singapore

Most participants could not identify a scenario where formula search is useful!

- Formulas often named (e.g., ‘Pythagorean theorem,’ ‘entropy’)
- Formulas overly specific for some information needs (e.g., concepts)
- Inconvenient to enter formulas using **methods known to the participants**
    - Graphical editors, string-based editors (e.g., for LaTeX)

A Study of Mathematical Non-Experts

Wangari et al., SIGIR 2014

> ***Task 3***: *Your classmate is struggling with binomial coefficients. Find one or more resources to help explain to your classmate how to find the value of* $\binom{4}{2}$.

Participants. 16 1st/2nd year undergraduates

Tasks. 4 tasks, rotated through conditions

Conditions (in order - training after Step 2):

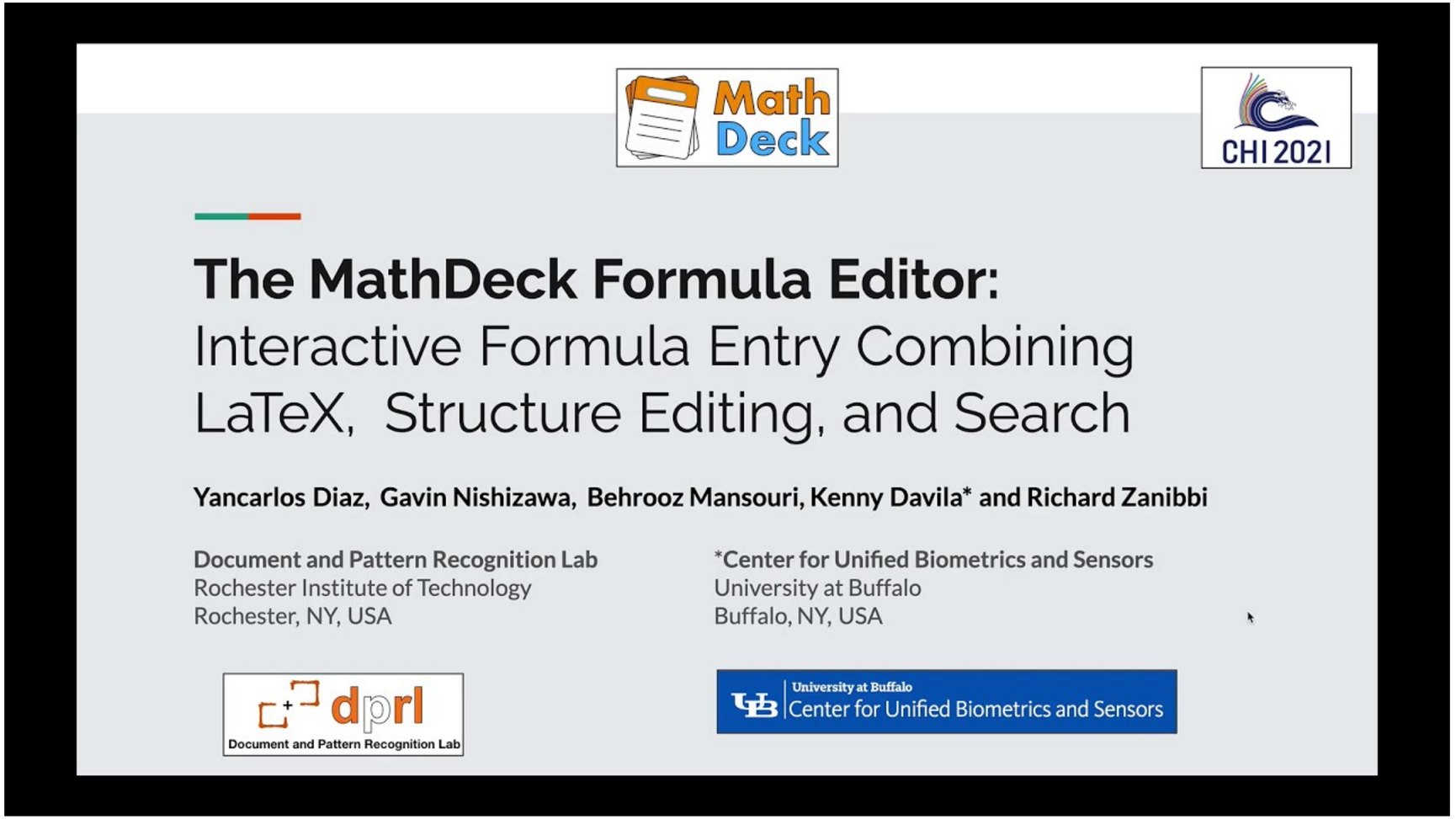
1. Text books, notes, websites, and/or online search
2. Online search using standard search engines
3. Online search using only the m_{in} interface
4. Online search with option of using m_{in}

Summary of Findings

- Self-reported success: little difference between conditions 1/2 & 3/4
- No participants used formula string encodings for search
- Handwritten input appreciated, esp. for Task 3, despite errors
- Provides support for handwritten formula / visual entry may help bridge a query formulation gap for non-expert users

The MathDeck Search Interface

Diaz et al., CHI 2021

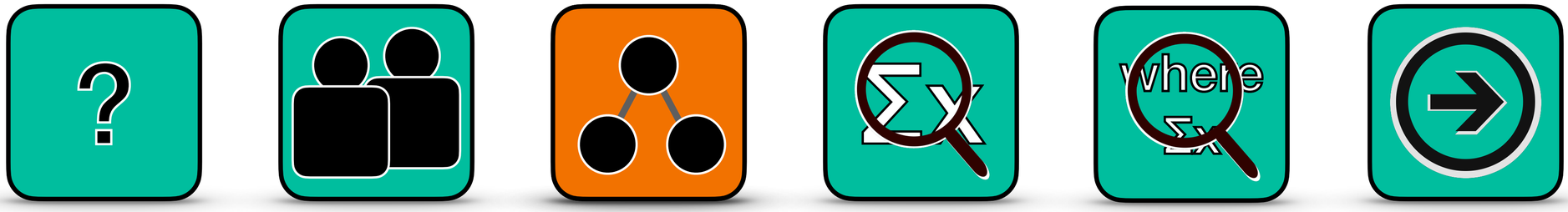


https://mathdeck.org (demo) Video: https://youtu.be/XfXQhwIQlbc

Adapting Broder's Taxonomy of Information Needs

Broder, SIGIR Forum 2002

NAVIGATIONAL	Find a specific resource ('known item' retrieval)
Examples:	Web page (e.g., for formula entry)
	Document (e.g., Book, Technical Paper)
	Video
	Audio recording
TRANSACTIONAL	Find online resources for use/interaction
Examples:	Formula entry
	Evaluating and plotting a formula
	Simplification of a formula
	Interactive theorem proving
INFORMATIONAL	Find information for a topic or question
Examples:	How to compute an expression (e.g., integral)
	Symbol and operation definitions (e.g., ζ, $\binom{n}{k}$)
	Concept name(s) associated with a formula
	When is a function not differentiable?
	Who was Gauss?
	P = NP



Encoding and Storing Math:
Formula Representation and Indexing

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

Document Excerpt

Understanding Inverse Document Frequency (Robertson, 2004)

2 The basic formula

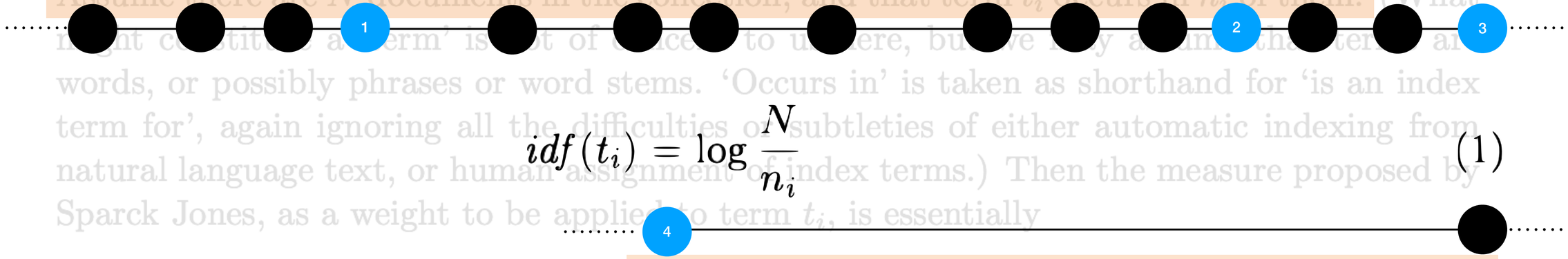
Assume there are N documents in the collection, and that term t_i occurs in n_i of them. (What might constitute a 'term' is not of concern to us here, but we may assume that terms are words, or possibly phrases or word stems. 'Occurs in' is taken as shorthand for 'is an index term for', again ignoring all the difficulties or subtleties of either automatic indexing from natural language text, or human assignment of index terms.) Then the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

Document Excerpt

Sequence of Word and Formula Tokens

Assume there are N documents in the collection, and that term t_i occurs in n_i



$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

Math Formula Representations

A Taxonomy

Operator Trees (OPTs)

How to **evaluate** formulas, from hierarchy of mathematical operations (i.e., operation syntax)

- Operation/relation precedence, associativity, commutativity explicit in tree
- **Examples**: Content MathML, prefix notation (e.g., math expressions in Lisp)

Symbol Layout Trees (SLTs)

How to **draw** formulas, from (1) spatial arrangement of symbols on writing lines, and (2) font/formatting instructions

- **Examples**: LaTeX, Presentation MathML

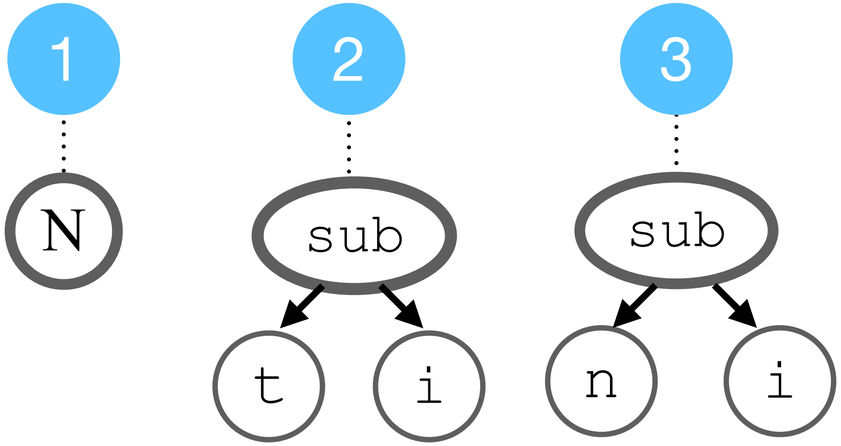
Visual Representations

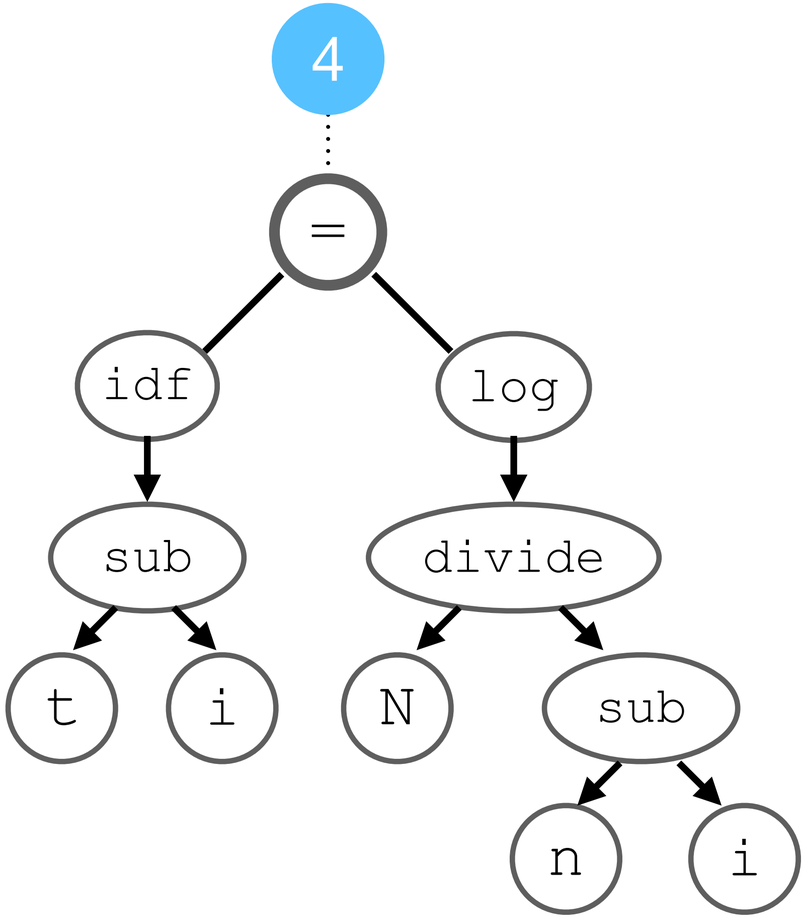
Describe appearance of formulas, without using (1) named symbols and/or (2) named spatial relationships

- **Examples**: line-of-sight graphs, spatial symbol partitions (e.g., PHOC), images (e.g., PDF, png)

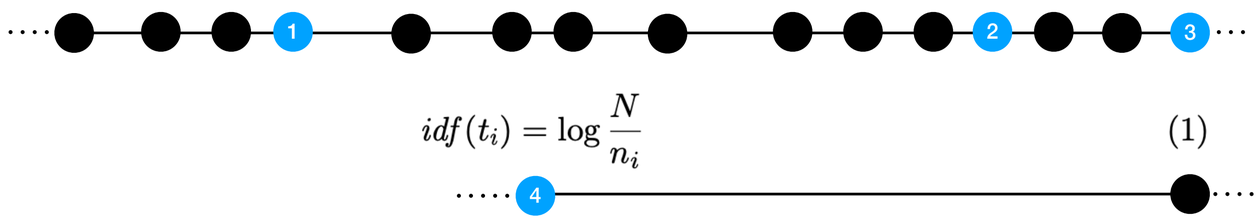
Operator Trees

Expression Syntax: How to Evaluate a Formula





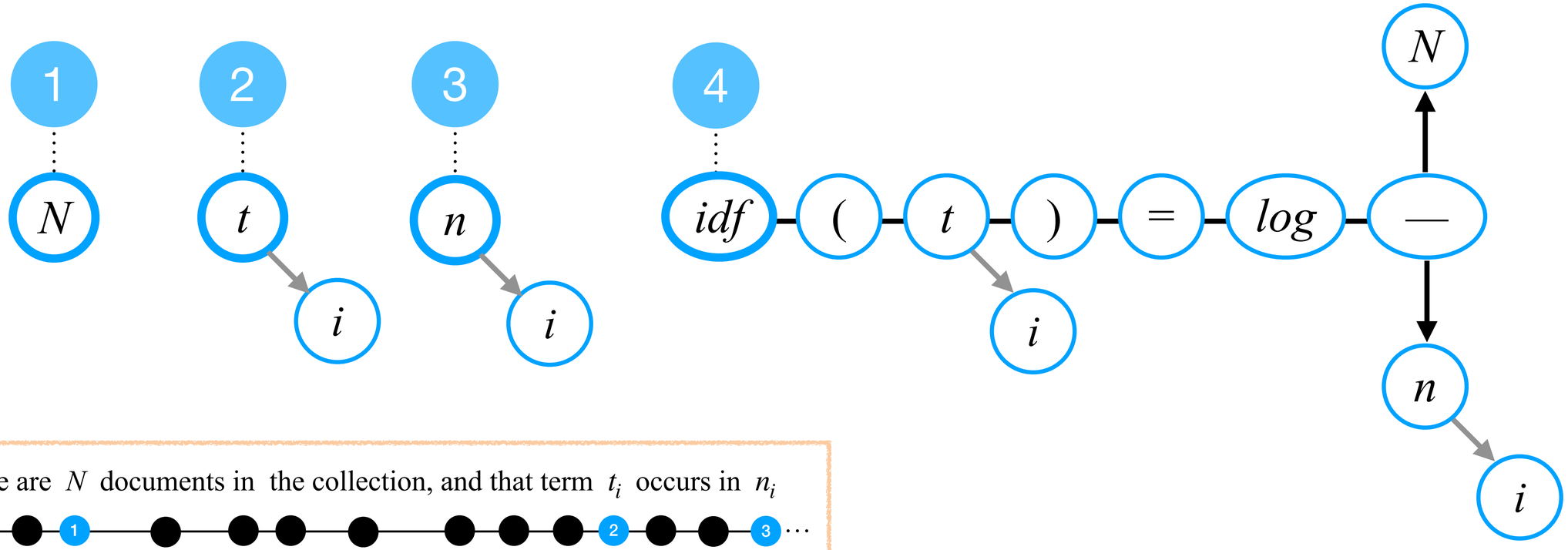
Assume there are N documents in the collection, and that term t_i occurs in n_i



$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

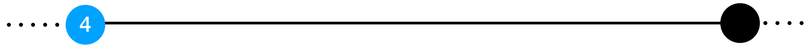
Symbol Layout Trees

Symbols and Spatial Relationships: How to Draw a Formula



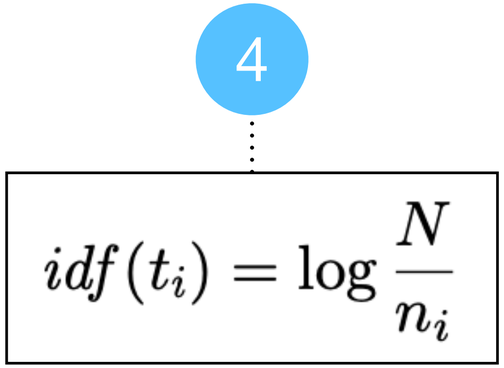
Assume there are N documents in the collection, and that term t_i occurs in n_i

$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

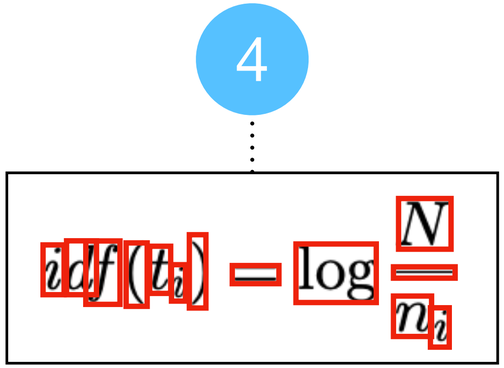


Visual Representations

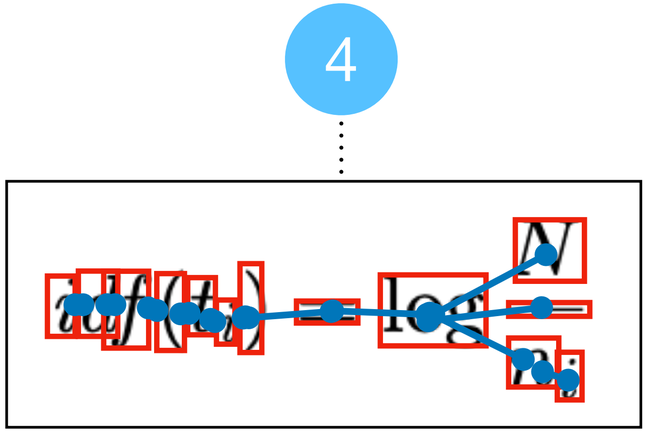
Describe Appearance without Named Symbols and/or Named Relationships



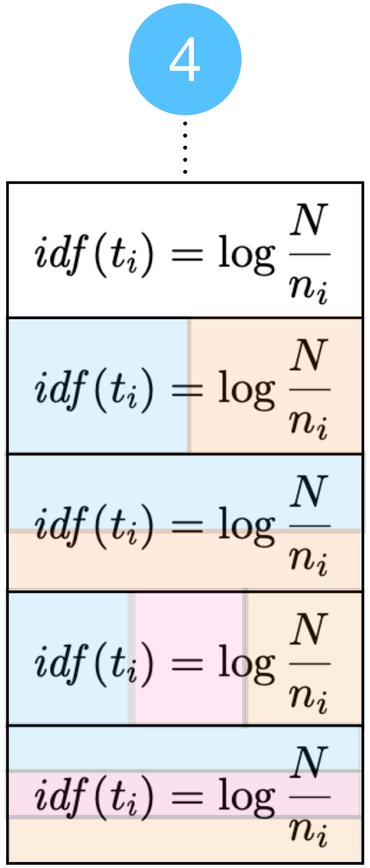
Raster Image (Pixels)



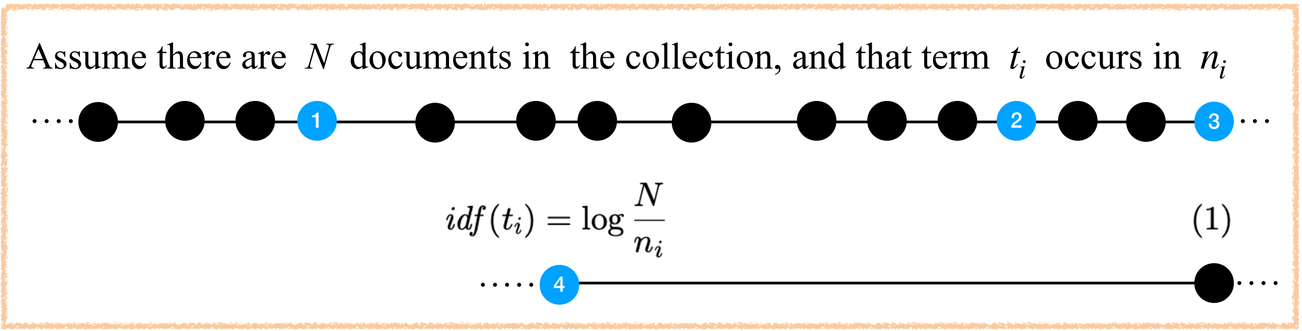
Vector Image (Instructions, e.g. PDF)



Line-of-Sight Graph (on Symbols)



Pyradmidal Histogram of Characters (PHOC)



Indexing Formulas for Search

Q. Where do searchable formulas come from?

Web Pages

Tagged math (e.g., in Math Stack Exchange)

Formulas between LaTeX delimiters (e.g., *$ … $* for MathJax)

Stand-Alone Documents

PDF documents (born-digital and scanned/OCR'd)

Word processing and presentation files (e.g., LaTeX, Word, PowerPoint)

Videos

e.g., from Math courses, technical lectures

Tools Available (e.g., SymbolScraper for PDF):
https://www.cs.rit.edu/~dprl/software.html

Formula Index Types

Two Common Approaches, both Usable w. OPT, SLT, or Visual Representations

1. Symbolic Inverted Index ('Sparse vectors' over a vocabulary of symbols and relationships)

Maps **identifier tuple keys** to **inverted lists** of formulas containing the key (+ opt. position, etc.)

e.g., 'ζ' ➜ [f1, f5, f100, …]
(x, 2, times) ➜ [f1, f3, f99, …]
(x, k, +, 2, (*super, hor, hor*)) ➜ [f1, f3, f99, …] (**key:** x^{k+2})

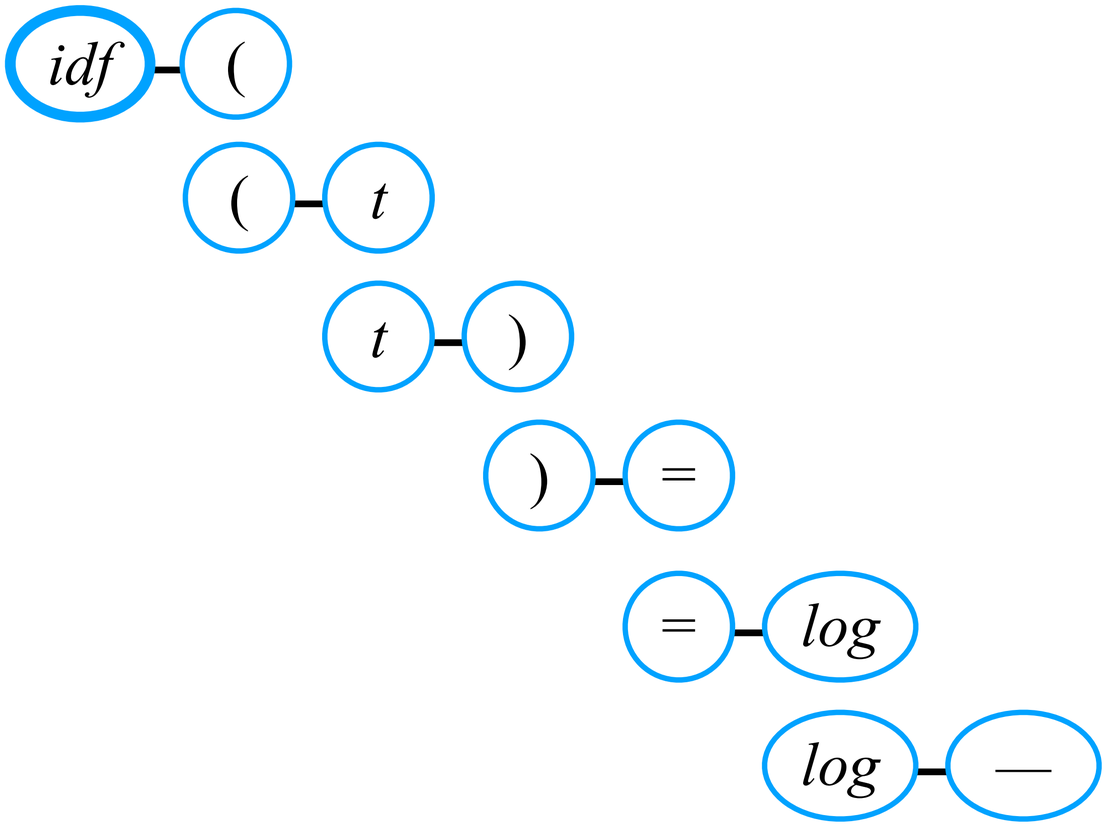
***2.* Spatial Inverted Index ('Dense vectors' in Euclidean space)**

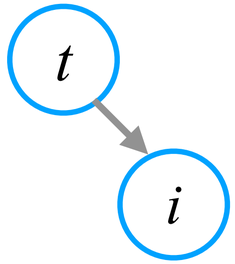
Maps **vector keys** to formulas/sub-formulas close in the embedding space using nearest neighbor methods (e.g., faiss)

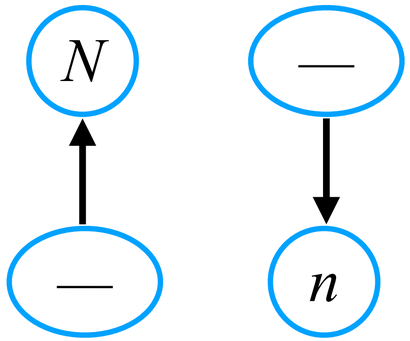
e.g., 'ζ' ➜ [0, 1, 0, 0, …,] ➜ [f1, f5, f100, …]
(x, 2, times) ➜ [0.3, -0.1, 0.99, ….] ➜ [f1, f3, f99, …]
(x, k, +, 2, (*super, hor, hor*)) ➜ [2, 1.43, -0.6, …] ➜ [f1, f3, f99, …] (**key:** x^{k+2})

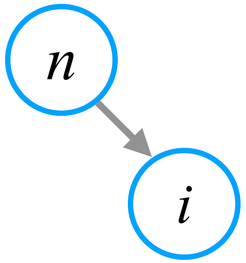
Keys for SLT Path Length 1 (All Edges)

$$idf(t_i) = \log \frac{N}{n_i}$$

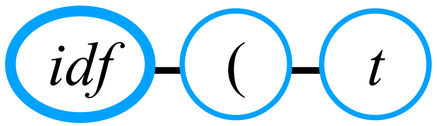


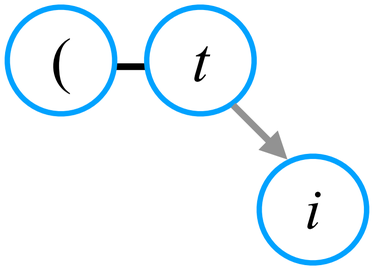


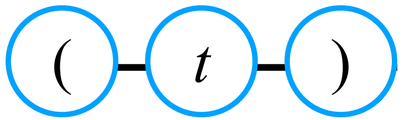


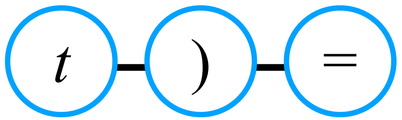


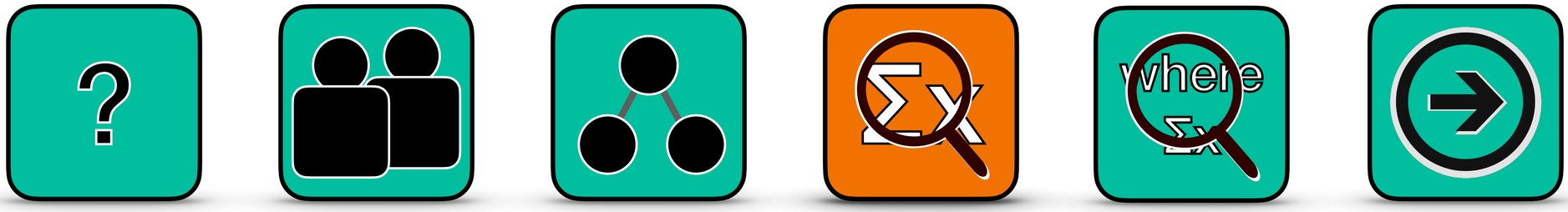
Keys for SLT Path Length 2 (Partial Edge List)











Formula Search:

Query-by-Expression

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

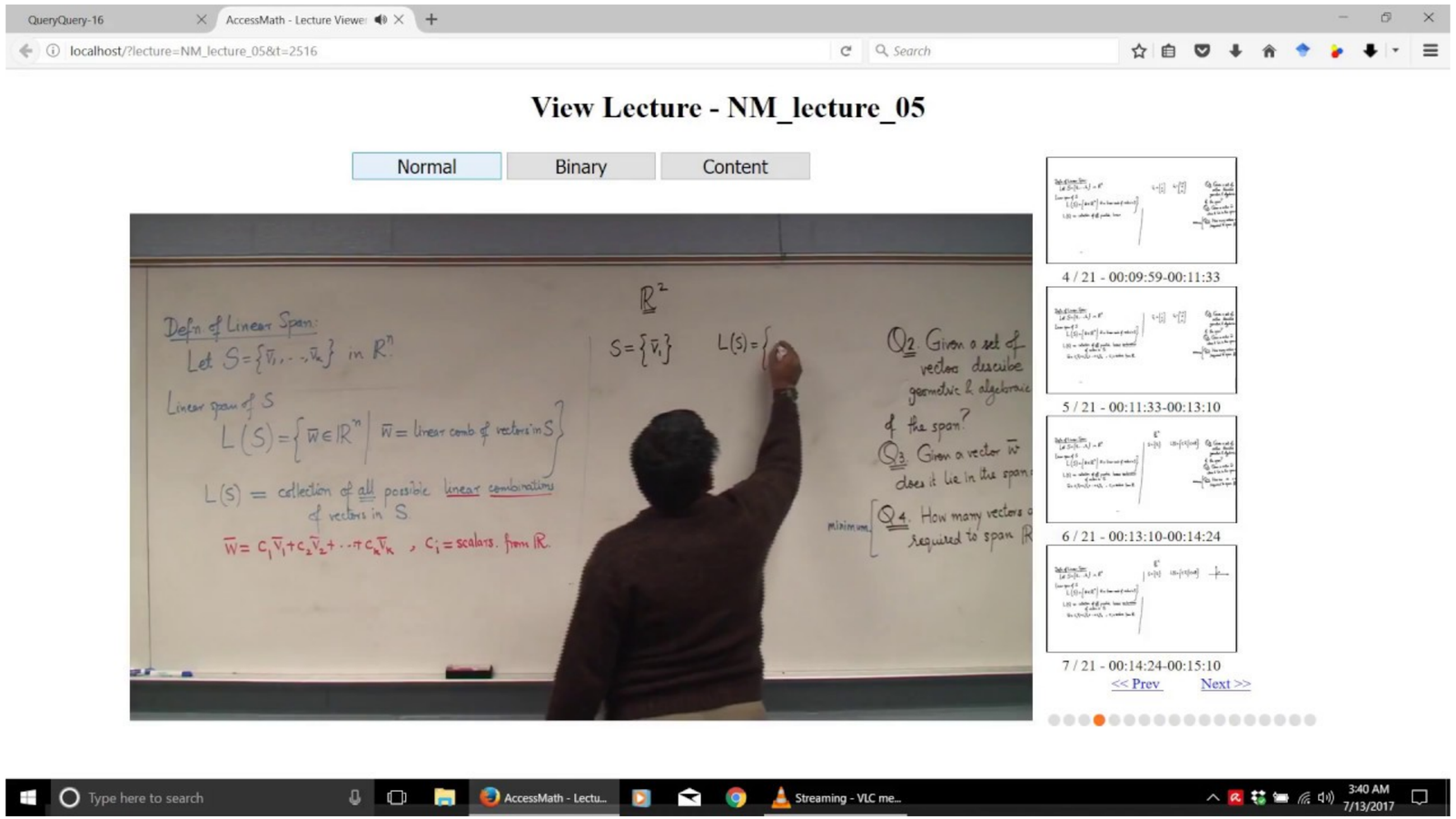
Formula Search

Basic Strategy

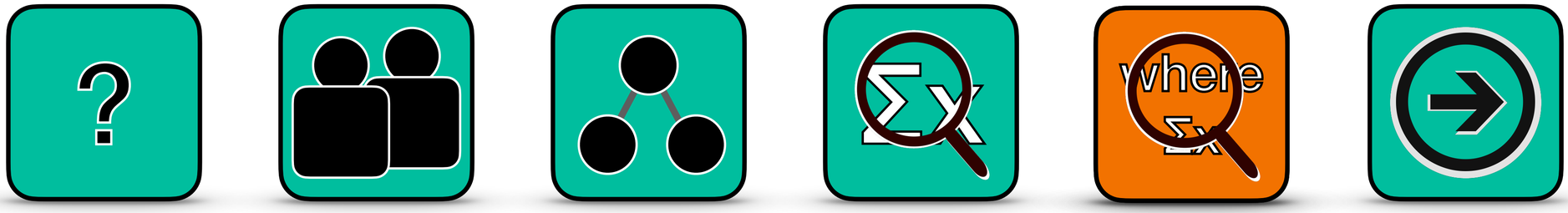
1. **Convert query into desired representation(s)** (OPT, SLT, Visual)
2. **Decompose representation(s) into search keys** (e.g., tuples, vectors)
3. **Lookup keys** in symbolic/spatial inverted index/indices & **compile matches**
4. **Score** matched formulas (usually, apply score constraints to **prune matches)**
   - e.g., cosine similarity for embedded dense vectors for each match
   - e.g., TF-IDF or BM25 for sparse vectors
   - e.g., in general, models designed to collect matches and score using a sum/accumulator

Tangent-V: Video Search w. Inverted Index on LOS Edges

AccessMath (Davila & Zanibbi, ICFHR 2018)



https://www.youtube.com/watch?v=gn24qo1MLN0



Multimodal Search:

Searching with Formulas and Text

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

ARQMath

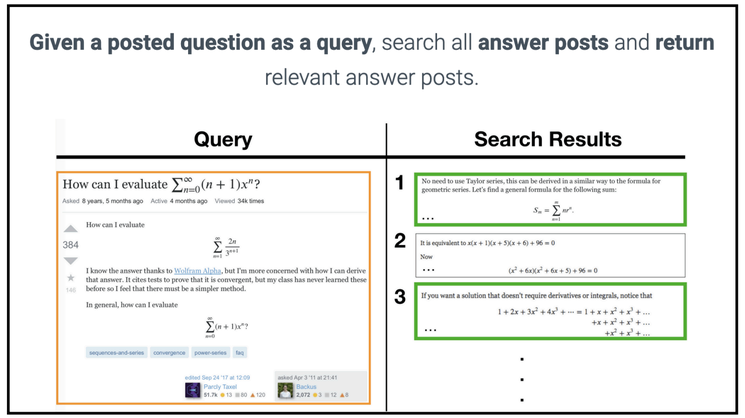
Answer Retrieval for Questions on Math (Mansouri et al., CLEF 2022)

https://www.cs.rit.edu/~dprl/ARQMath

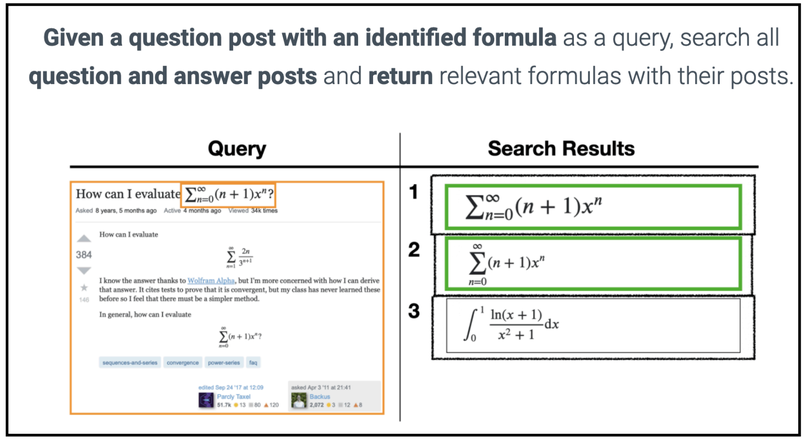
Shared task (lab/competition) held at CLEF 2020-2022

New benchmark for math QA in Math Stack Exchange posts + contextualized formula search (200+ test queries and result evaluations for Tasks 1 & 2)

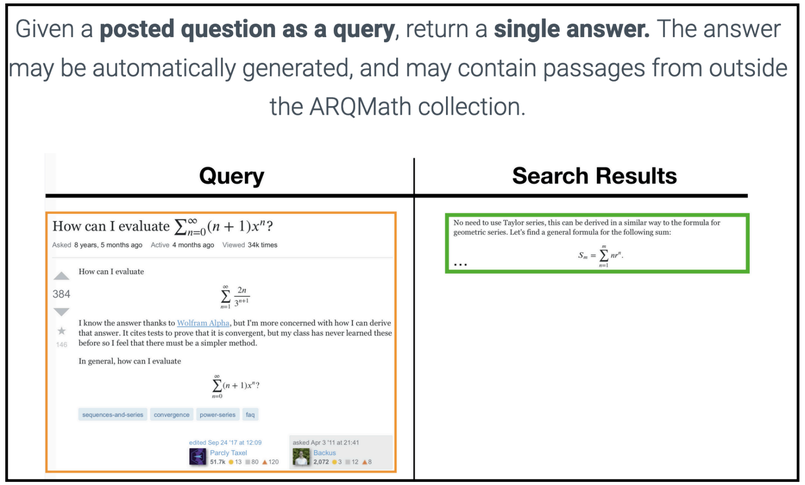
Task 1



Task 2



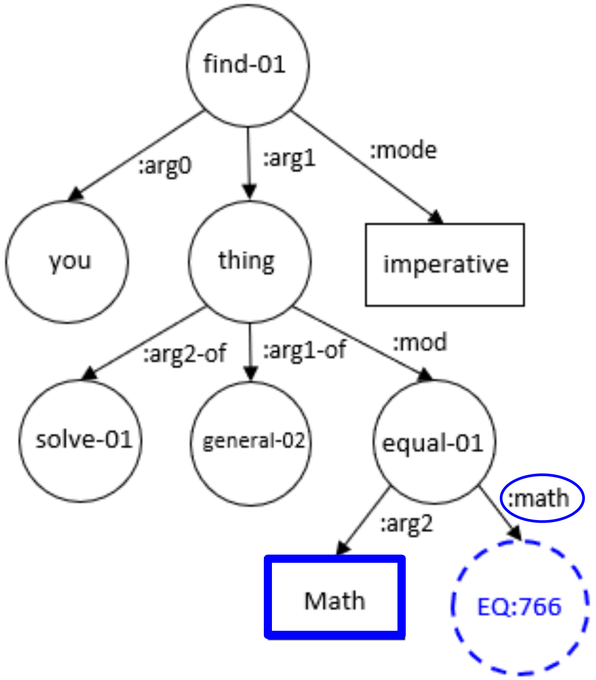
Task 3



MathAMR for "Find $x^n + y^n + z^n$ general solution"

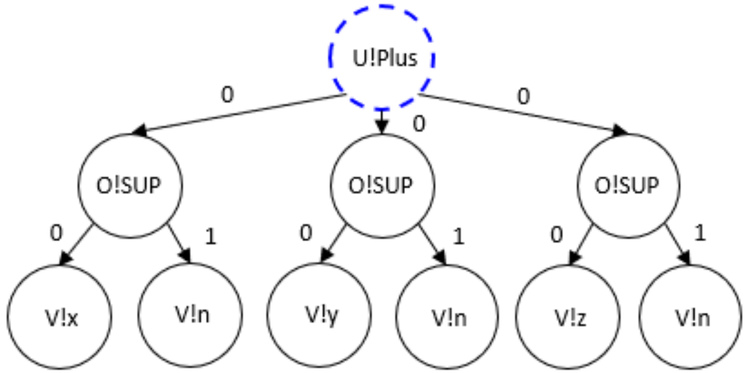
AMR Tree

Find *EQ:766* general solution



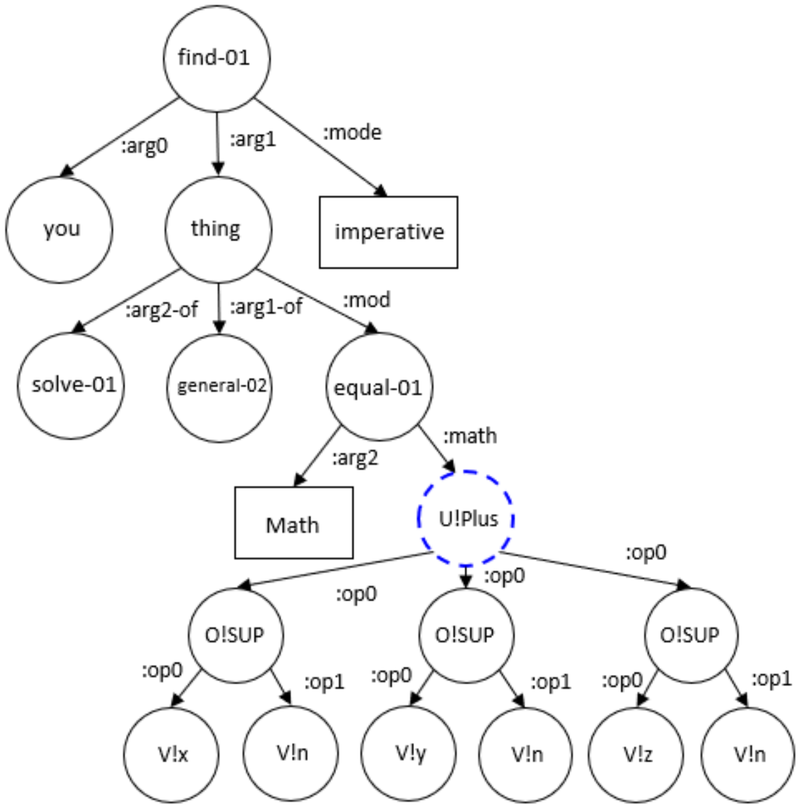
Operator Tree

$x^n + y^n + z^n$

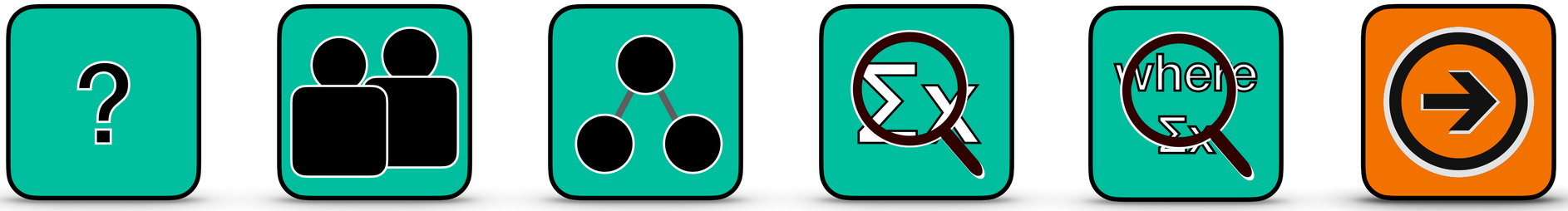


MathAMR Tree

Find $x^n + y^n + z^n$ general solution



Mansouri et al., CIKM 2022



Closing Thoughts on Math IR:

Taking Stock and Moving Forward

Topos Institute Colloquium Series, Oct 27, 2022

TOPOS INSTITUTE

Notation vs. Information

Excerpt, from *Understanding Inverse Document Frequency* (Robertson, 2004)

*Meadows and Freitas, arXiv 2022

2 The basic formula

Assume there are N documents in the collection, and that term t_i occurs in n_i of them. (What might constitute a 'term' is not of concern to us here, but we may assume that terms are words, or possibly phrases or word stems. 'Occurs in' is taken as shorthand for 'is an index term for', again ignoring all the difficulties or subtleties of either automatic indexing from natural language text, or human assignment of index terms.) Then the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

$$idf(t_i) = \log \frac{N}{n_i} \qquad (1)$$

Surrounding text and other **context** needed to interpret formulas

Example NLP Tasks

Important Future Directions (Meadows and Freitas, arXiv 2022)

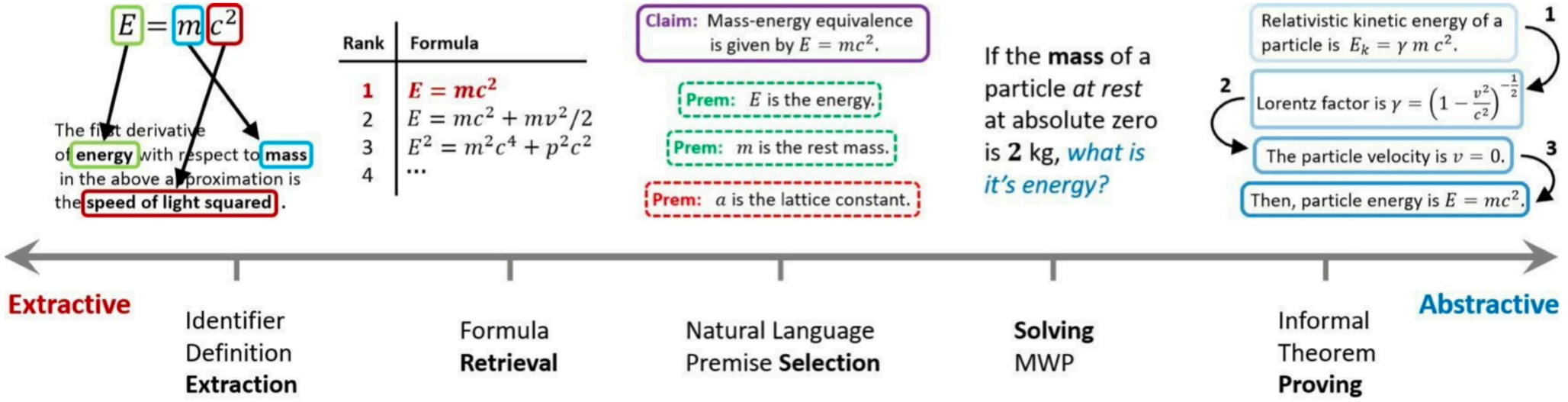
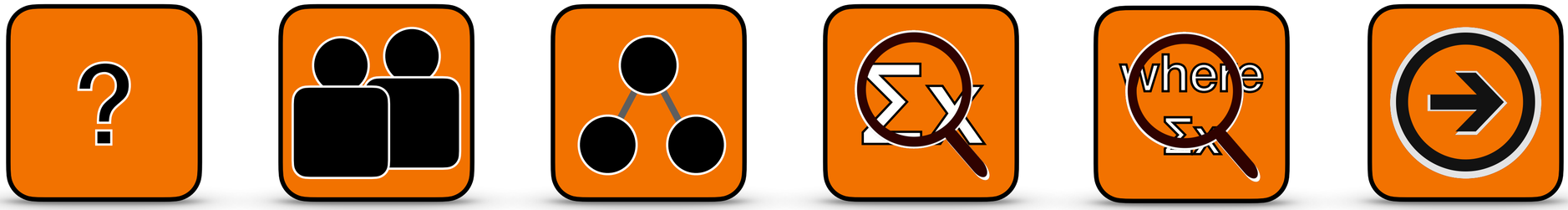


Figure 1: *Extractive* tasks are closer to the lexical and surface-level expression of the text while *abstractive* tasks tend to require the integration of symbolic-level and abstract reasoning.



Thank you! Acknowledgements:

A sincere thank you to Valeria de Paiva and the Topos Institute for the invitation to give this talk.

We also thank the NSF and Alfred P. Sloan Foundation for providing financial support for this work

MathSeer Project Collaborators

Anurag Agarwal, C. Lee Giles, Douglas W. Oard, Jian Wu

MathSeer Advisory Board

Marti Hearst, Frank Tompa, Michael Kohlhase

Students

Behrooz Mansouri, Shaurya Rohatgi (Penn State), Kenny Davila, Ayush Kumar Shah, Abhisek Dey, Matt Langsenkamp, Bryan Manrique Amador, Robin Avenoso, Yancarlos Diaz, Gavin Nishizawa, Mahshad Mahdavi, JP Ramissini, Alex Keller, Jennifer Liu, Abhishai Dmello, Parag Mali

Document and Pattern Recognition Lab

https://www.cs.rit.edu/~dprl

ALFRED P. SLOAN FOUNDATION