RIT
Rochester Institute of Technology

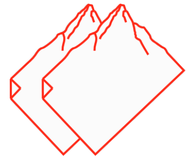
dprl
Document and Pattern Recognition Lab

Information Retrieval:
Searching with Formulas and Text, and a Framework for Characterizing Search by Humans & Systems

Richard Zanibbi

Director, Document and Pattern Recognition Lab
Professor of Computer Science
Rochester Institute of Technology, USA

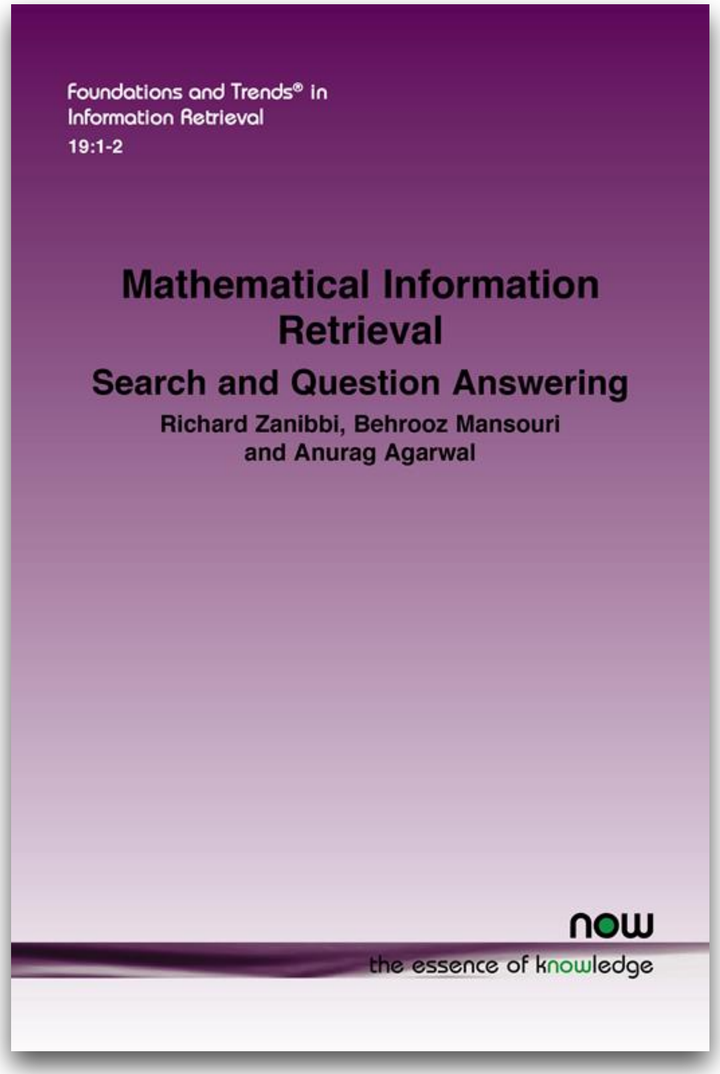
May 28, 2026



SSDA 2026
Summer School on Document Analysis

Mathematical Information Retrieval:
Search and Question Answering

Richard Zanibbi, Behrooz Mansouri, and Anurag Agarwal



Foundations and Trends allowed the final PDF to be released for free! Link to PDF:

SSDA 2026
Summer School on
Document Analysis

Interpreting & Representing Math Formulas

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

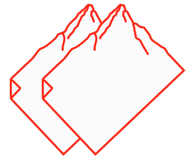
dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on
Document Analysis

An Example Technical Document Excerpt

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

Unless already known in the intended context, text and prior knowledge are *needed* to interpret formulas



SSDA 2026
Summer School on Document Analysis

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

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

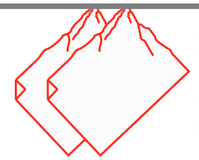
Unless already known in the intended context, text and prior knowledge are *needed* to interpret formulas

Example 2.1: Inverse Document Frequency (*IDF*)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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



SSDA 2026
Summer School on Document Analysis

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

Example 2.1: Inverse Document Frequency (IDF)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Textual context / information

Variables

Functions and Operators

Additional Context

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

Example 2.1: Inverse Document Frequency (IDF)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variable and function definitions

... Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variables

Functions and Operators

Additional Context

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

Example 2.1: Inverse Document Frequency (IDF)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variable and function definitions

... Assume there are N documents in the collection, and that term t_i occurs in n_i of them ... the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variables: placeholders for a set of values, similar to common nouns

- The text identifies N as the number of documents in a collection. N is like a common noun, because the collection is not specified.
- The text defines t_i as any term appearing n_i times in a collection, with shared identifier i, *e.g.*, (t_3, n_3) could be `('weight', 11)`.

Functions and Operators

Additional Context

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

Example 2.1: Inverse Document Frequency (IDF)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them . . . the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variable and function definitions

. . . Assume there are N documents in the collection, and that term t_i occurs in n_i of them . . . the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variables: placeholders for a set of values, similar to common nouns

- The text identifies N as the number of documents in a collection. N is like a common noun, because the collection is not specified.
- The text defines t_i as any term appearing n_i times in a collection, with shared identifier i, *e.g.*, (t_3, n_3) could be `('weight', 11)`.

Functions & Operators: create new from given values, like verbs

- log: log function with unspecified base.
- $idf(t_i)$: aside from the unspecified log base, a concrete function in Equation (1). The text says this gives a weight for term t_i.
- Division ($\frac{\cdot}{\cdot}$), application ($idf(\cdot)$, $\log \cdot$), and equivalence ($=$) appearing in Equation (1) that are not defined in the excerpt.

Additional Context

An Example Technical Document Excerpt

Taken from “Understanding Inverse Document Frequency” by Robertson (2004)

Example 2.1: Inverse Document Frequency (IDF)

Excerpt from Robertson (2004)

...Assume there are N documents in the collection, and that term t_i occurs in n_i of them . . . the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variable and function definitions

. . . Assume there are N documents in the collection, and that term t_i occurs in n_i of them . . . the measure proposed by Sparck Jones, as a weight to be applied to term t_i, is essentially

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

Variables: placeholders for a set of values, similar to common nouns

- The text identifies N as the number of documents in a collection. N is like a common noun, because the collection is not specified.
- The text defines t_i as any term appearing n_i times in a collection, with shared identifier i, *e.g.*, (t_3, n_3) could be `('weight', 11)`.

Functions & Operators: create new from given values, like verbs

- log: log function with unspecified base.
- $idf(t_i)$: aside from the unspecified log base, a concrete function in Equation (1). The text says this gives a weight for term t_i.
- Division ($\frac{\cdot}{\cdot}$), application ($idf(\cdot)$, $\log \cdot$), and equivalence ($=$) appearing in Equation (1) that are not defined in the excerpt.

Additional context:

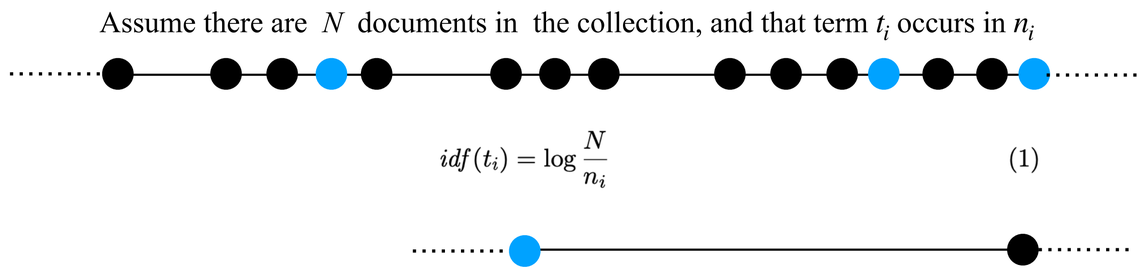
- The text indicates Spärck Jones introduced the idf formula in a different, unspecified form (Jones, 1972).

Formulas in Text (Symbol Layout Trees)

Sequence of Tokens and Symbol Layout Trees

Example 2.5: Text tokens, formula tokens, and visual formula structure

Sequence of text (black) and formula (blue) tokens



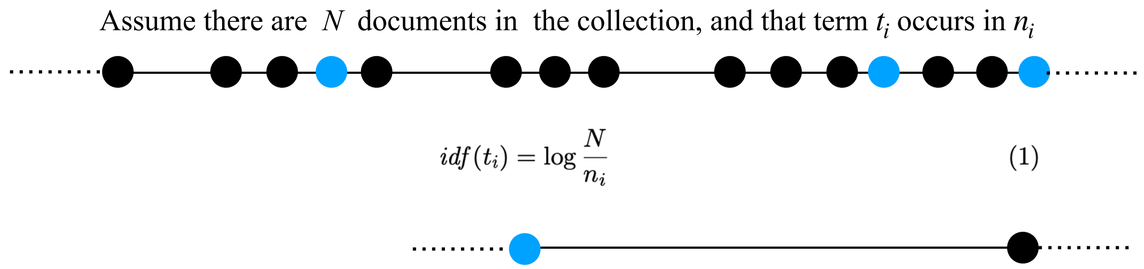
Symbol Layout Tree (SLT)

Formulas in Text (Symbol Layout Trees)

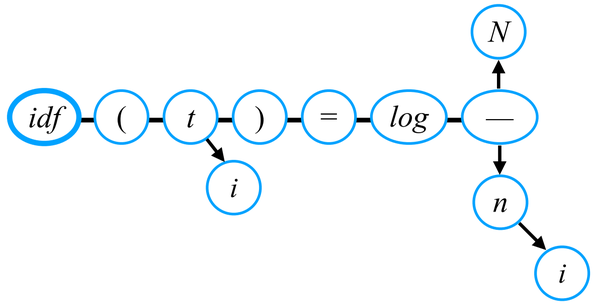
Sequence of Tokens and Symbol Layout Trees

Example 2.5: Text tokens, formula tokens, and visual formula structure

Sequence of text (black) and formula (blue) tokens



Symbol Layout Tree (SLT) for Equation (1)

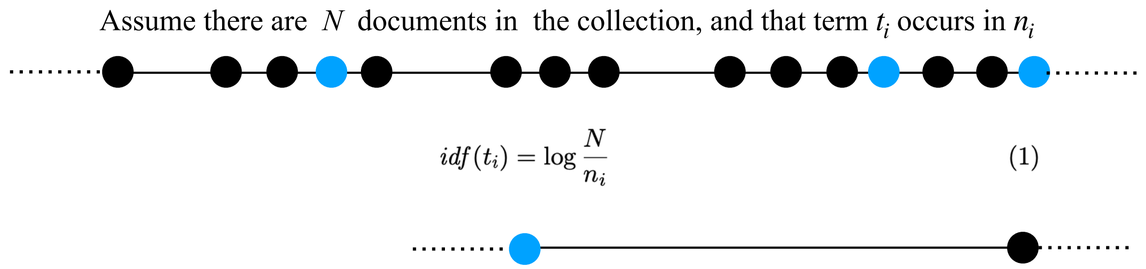


Formulas in Text (Symbol Layout Trees)

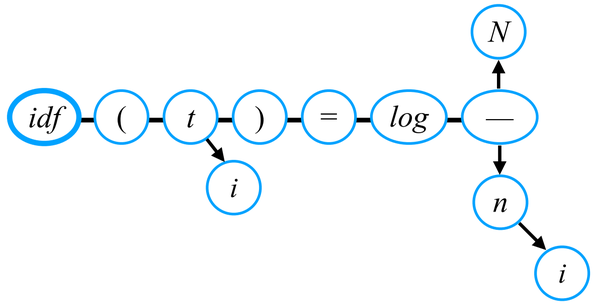
Sequence of Tokens and Symbol Layout Trees

Example 2.5: Text tokens, formula tokens, and visual formula structure

Sequence of text (black) and formula (blue) tokens



Symbol Layout Tree (SLT) for Equation (1)



Symbol Layout Trees

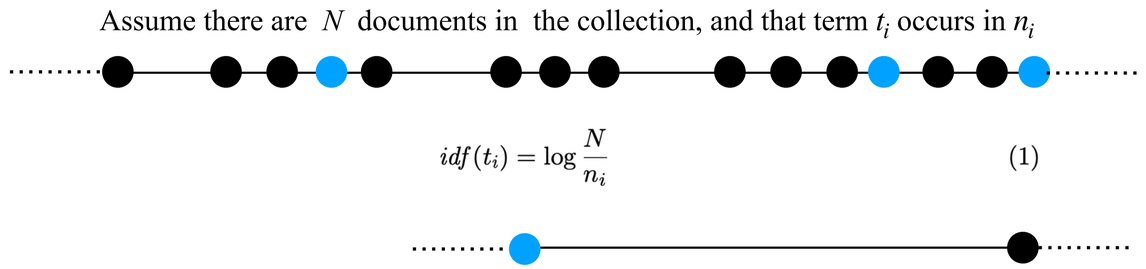
Represent formula **appearance**

Formulas in Text (Symbol Layout Trees)

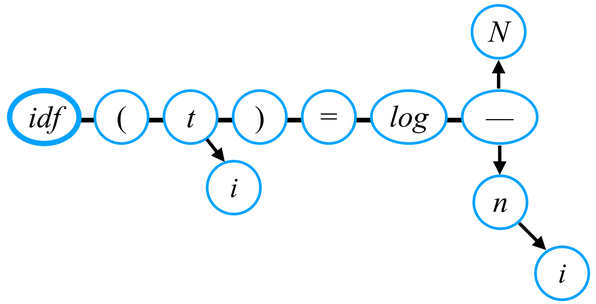
Sequence of Tokens and Symbol Layout Trees

Example 2.5: Text tokens, formula tokens, and visual formula structure

Sequence of text (black) and formula (blue) tokens



Symbol Layout Tree (SLT) for Equation (1)



Symbol Layout Trees

Represent formula **appearance**

Representation

Symbols on writing lines;

Writing lines can be *nested* at symbols (e.g., subscript)

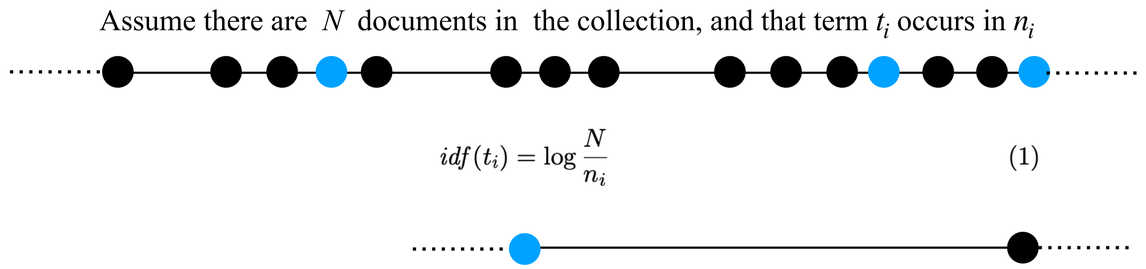
SSDA 2026
Summer School on Document Analysis

Formulas in Text (Symbol Layout Trees)

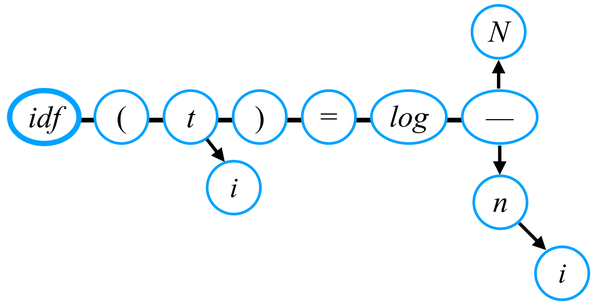
Sequence of Tokens and Symbol Layout Trees

Example 2.5: Text tokens, formula tokens, and visual formula structure

Sequence of text (black) and formula (blue) tokens



Symbol Layout Tree (SLT) for Equation (1)



Symbol Layout Trees

Represent formula **appearance**

Representation

Symbols on writing lines;

Writing lines can be *nested* at symbols (e.g., subscript)

Example 2.7: *idf* formula in LaTeX and Python code

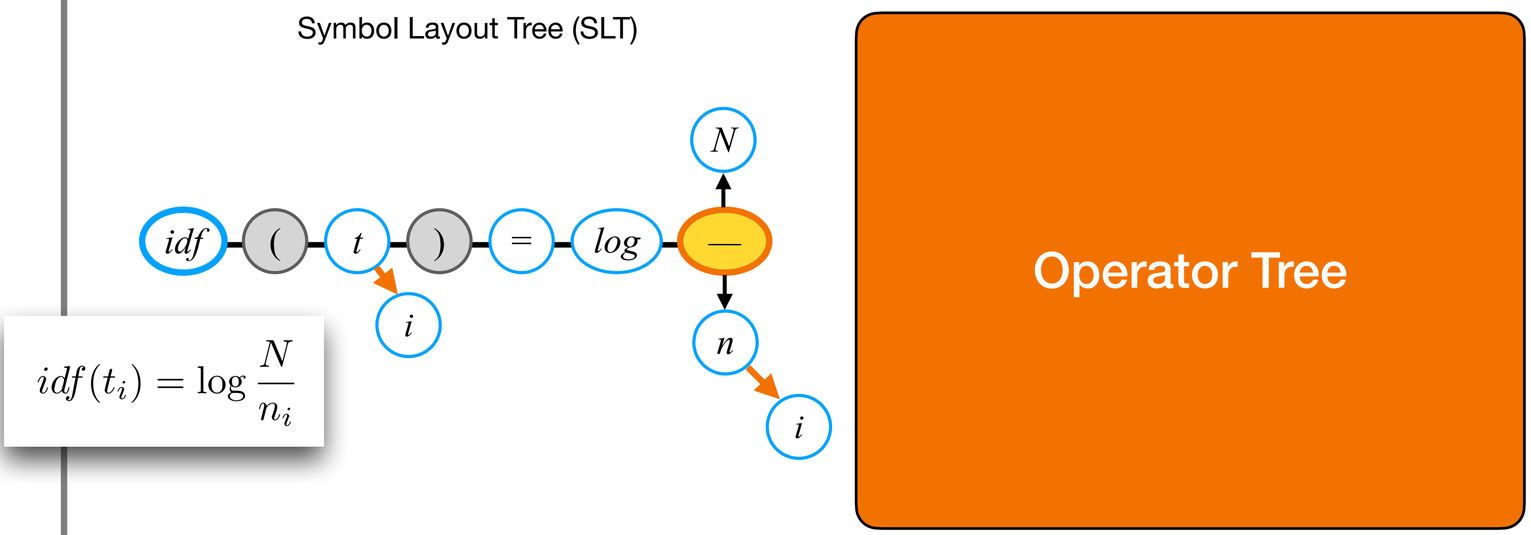
LaTeX: Symbol Layout Tree representation

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

Formulas as Operator Trees (OPTs)

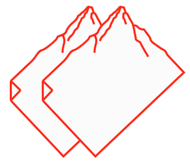
Operation Hierarchy = Math operator syntax

Example 2.6: Translating a Symbol Layout Tree to an Operator Tree



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

Grey nodes in the SLT indicate parentheses removed in the OPT, where they are redundant. SLT subscript edges and fraction line are replaced by `sub` and `divide` nodes in the OPT.

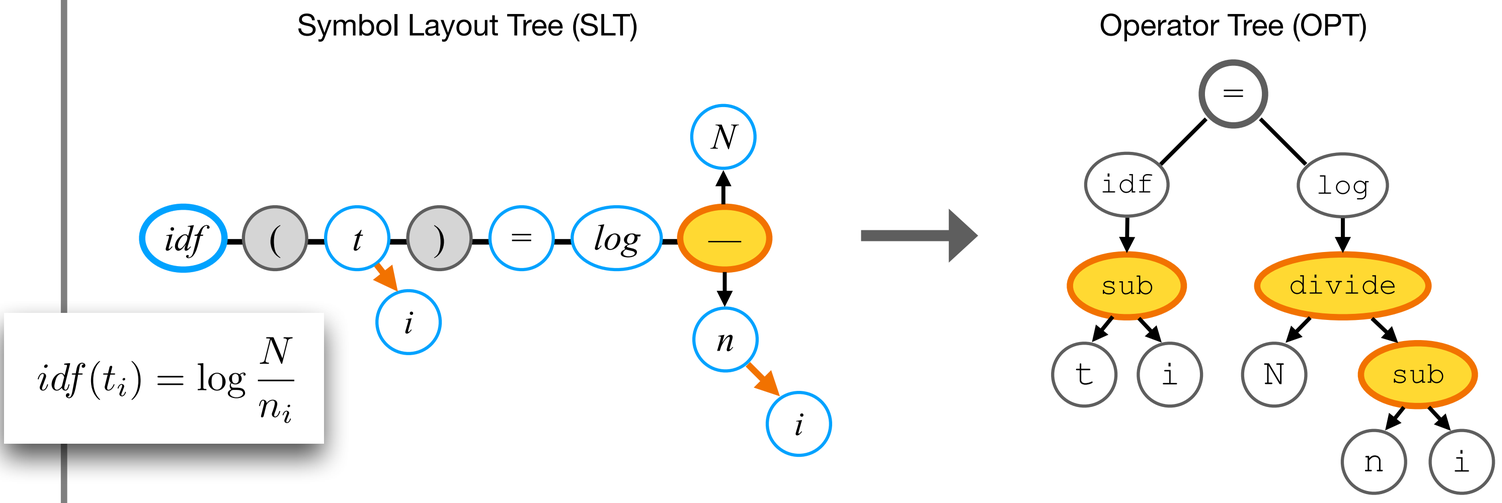


SSDA 2026
Summer School on
Document Analysis

Formulas as Operator Trees (OPTs)

Operation Hierarchy = Math operator syntax

Example 2.6: Translating a Symbol Layout Tree to an Operator Tree



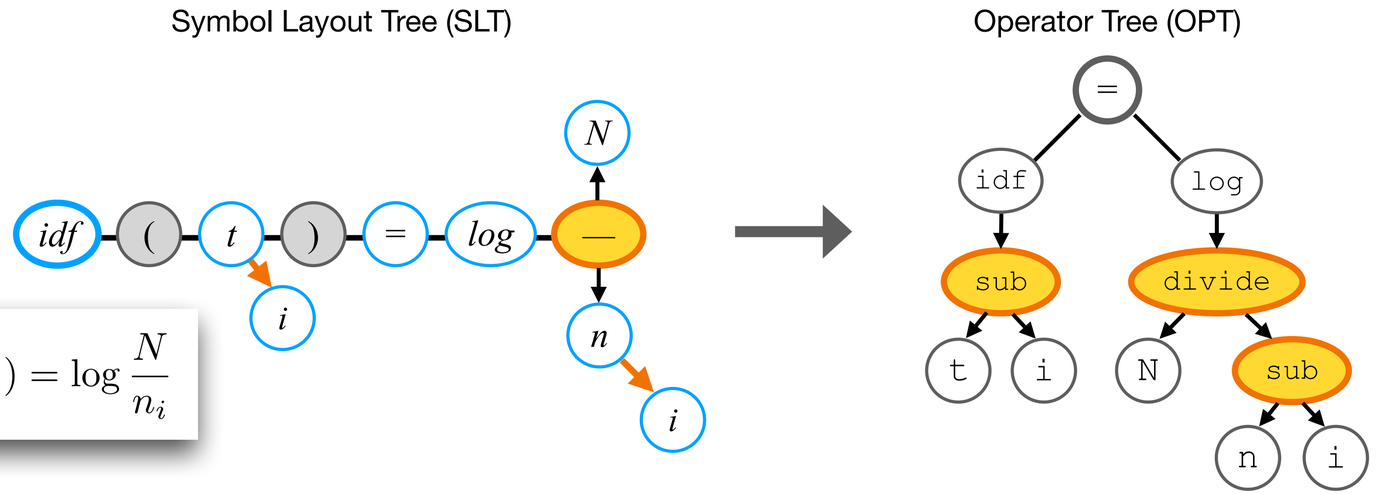
$$idf(t_i) = \log \frac{N}{n_i}$$

Grey nodes in the SLT indicate parentheses removed in the OPT, where they are redundant. SLT subscript edges and fraction line are replaced by `sub` and `divide` nodes in the OPT.

Formulas as Operator Trees (OPTs)

Operation Hierarchy = Math operator syntax

Example 2.6: Translating a Symbol Layout Tree to an Operator Tree



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

Grey nodes in the SLT indicate parentheses removed in the OPT, where they are redundant. SLT subscript edges and fraction line are replaced by `sub` and `divide` nodes in the OPT.

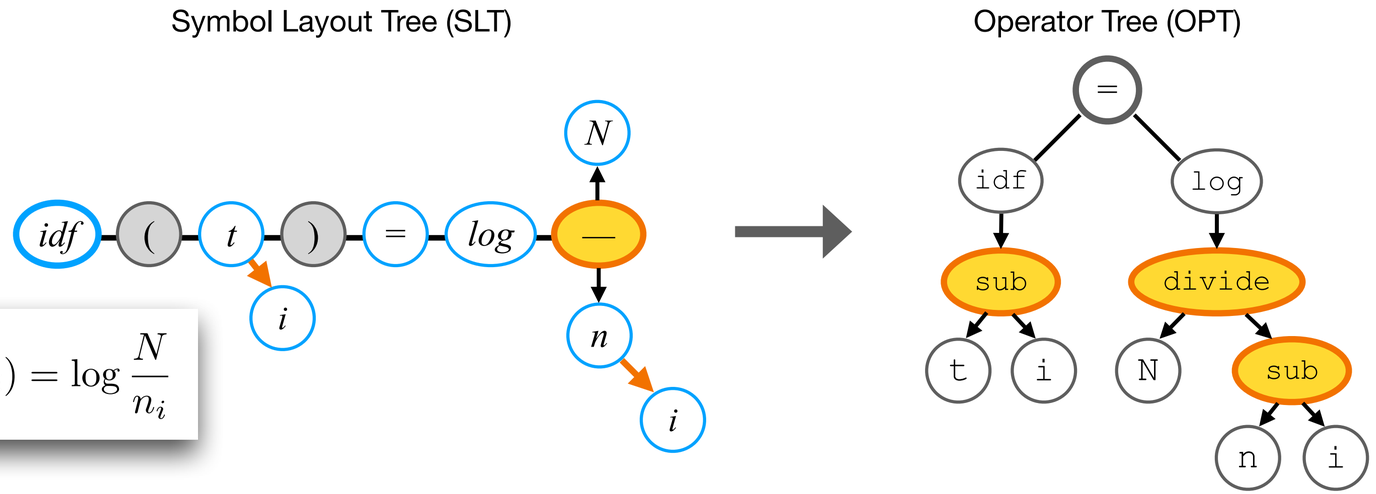
Operator Trees

Represent **operations** on provided augments (**operands**)

Formulas as Operator Trees (OPTs)

Operation Hierarchy = Math operator syntax

Example 2.6: Translating a Symbol Layout Tree to an Operator Tree



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

Grey nodes in the SLT indicate parentheses removed in the OPT, where they are redundant. SLT subscript edges and fraction line are replaced by `sub` and `divide` nodes in the OPT.

Operator Trees

Represent **operations** on provided augments (**operands**)

Representation

Last operation at root

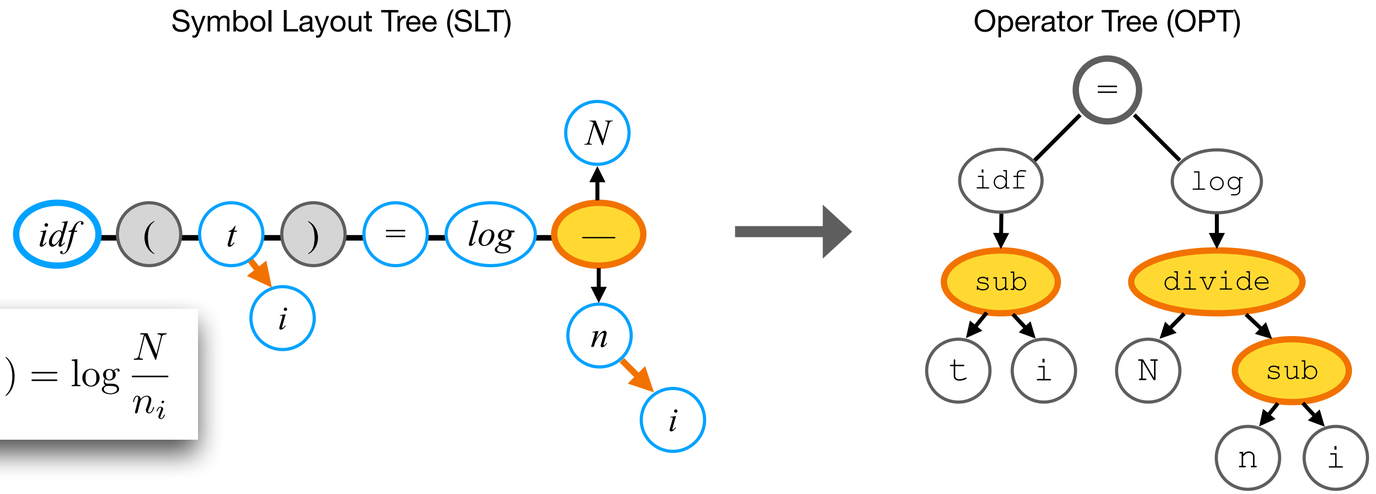
Operands are children of a node

Evaluation: **bottom-up**

Formulas as Operator Trees (OPTs)

Operation Hierarchy = Math operator syntax

Example 2.6: Translating a Symbol Layout Tree to an Operator Tree



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

Grey nodes in the SLT indicate parentheses removed in the OPT, where they are redundant. SLT subscript edges and fraction line are replaced by `sub` and `divide` nodes in the OPT.

Operator Trees

Represent **operations** on provided augments (**operands**)

Representation

Last operation at root

Operands are children of a node

Evaluation: **bottom-up**

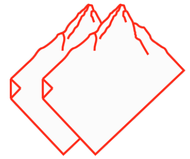
Syntactic NOT Semantic

Semantics needs more information

SSDA 2026
Summer School on Document Analysis

Formulas in the “Wild”

i.e., real-world document collections

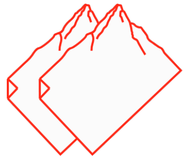


SSDA 2026
Summer School on
Document Analysis

Formulas in the “Wild”

i.e., real-world document collections

Formula representations **overwhelmingly provided as SLTs**
(e.g., LaTeX, Presentation MathML)



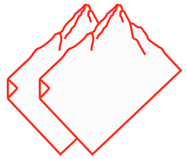
SSDA 2026
Summer School on
Document Analysis

Formulas in the “Wild”

i.e., real-world document collections

Formula representations **overwhelmingly provided as SLTs**
(e.g., LaTeX, Presentation MathML)

Translating SLTs to OPTs requires an expression grammar



SSDA 2026
Summer School on
Document Analysis

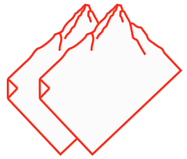
Formulas in the “Wild”

i.e., real-world document collections

Formula representations **overwhelmingly provided as SLTs**
(e.g., LaTeX, Presentation MathML)

Translating SLTs to OPTs requires an expression grammar

…but there is no ‘standard’ grammar or type/definition set for ‘math’



SSDA 2026
Summer School on
Document Analysis

Formulas in the “Wild”

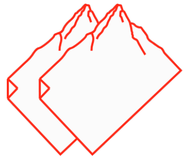
i.e., real-world document collections

Formula representations **overwhelmingly provided as SLTs**
(e.g., LaTeX, Presentation MathML)

Translating SLTs to OPTs requires an expression grammar

…but there is no ‘standard’ grammar or type/definition set for ‘math’

- Operations, variables, and constants can be defined multiple ways
- Operation argument structure varies by context
- Many formulas presented ambiguously, or incompletely in docs

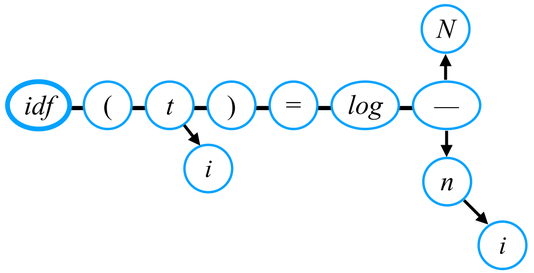


SSDA 2026
Summer School on
Document Analysis

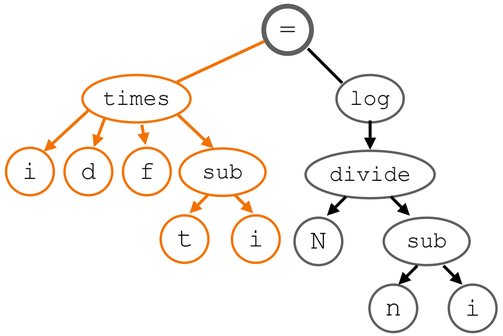
LaTeXML SLT → OPT Translation

Example 2.8: MathML generated from LaTeX using LaTeXML

idf is undefined in LaTeXML[a] and so i, d, and f are treated as variables.



Presentation MathML (SLT)



Content MathML (OPT)

Presentation MathML (SLT)

```
<math xmlns="http://.../MathML">
  <mi>i</mi>
  <mi>d</mi>
  <mi>f</mi>
  <mo stretchy="false">(</mo>
  <msub>
    <mi>t</mi>
    <mi>i</mi>
  </msub>
  <mo stretchy="false">)</mo>
  <mo>=</mo>
  <mi>log</mi>
  <mo>&#x2061;</mo>
  <mfrac>
    <mi>N</mi>
    <msub>
      <mi>n</mi>
      <mi>i</mi>
    </msub>
  </mfrac>
</math>
```

Content MathML (OPT)

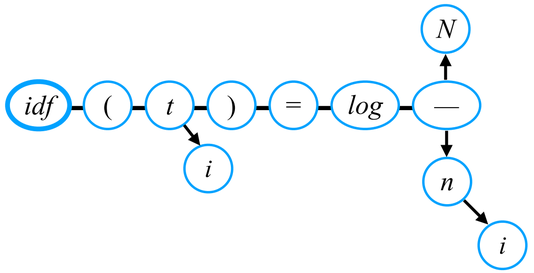
```
<math xmlns="http://.../MathML">
<apply>
  <eq/>
  <apply>
    <times/>
    <ci>i</ci>
    <ci>d</ci>
    <ci>f</ci>
    <msub>
      <ci>t</ci>
      <ci>i</ci>
    </msub>
  </apply>
  <apply>
    <log/>
    <apply>
      <divide/>
      <ci>N</ci>
      <msub>
        <ci>n</ci>
        <ci>i</ci>
      </msub>
    </apply>
  </apply>
</apply>
</math>
```

[a] https://math.nist.gov/~BMiller/LaTeXML

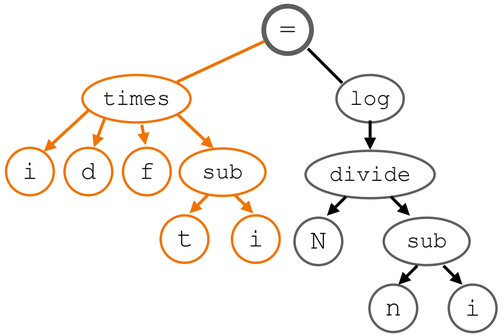
LaTeXML SLT → OPT Translation

Example 2.8: MathML generated from LaTeX using LaTeXML

idf is undefined in LaTeXML[a] and so i, d, and f are treated as variables.



Presentation MathML (SLT)



Content MathML (OPT)

Presentation MathML (SLT)

```
<math xmlns="http://.../MathML">
  <mi>i</mi>
  <mi>d</mi>
  <mi>f</mi>
  <mo stretchy="false">(</mo>
  <msub>
    <mi>t</mi>
    <mi>i</mi>
  </msub>
  <mo stretchy="false">)</mo>
  <mo>=</mo>
  <mi>log</mi>
  <mo>&#x2061;</mo>
  <mfrac>
    <mi>N</mi>
    <msub>
      <mi>n</mi>
      <mi>i</mi>
    </msub>
  </mfrac>
</math>
```

Content MathML (OPT)

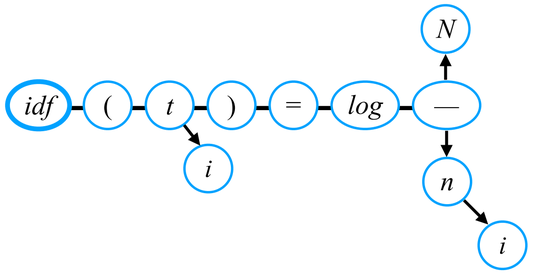
```
<math xmlns="http://.../MathML">
<apply>
  <eq/>
  <apply>
    <times/>
    <ci>i</ci>
    <ci>d</ci>
    <ci>f</ci>
    <msub>
      <ci>t</ci>
      <ci>i</ci>
    </msub>
  </apply>
  <apply>
    <log/>
    <apply>
      <divide/>
      <ci>N</ci>
      <msub>
        <ci>n</ci>
        <ci>i</ci>
      </msub>
    </apply>
  </apply>
</apply>
</math>
```

[a] https://math.nist.gov/~BMiller/LaTeXML

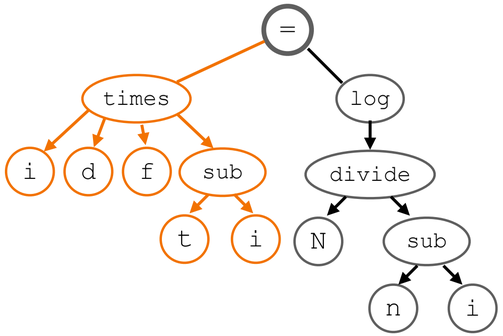
LaTeXML SLT → OPT Translation

Example 2.8: MathML generated from LaTeX using LaTeXML

idf is undefined in LaTeXML[a] and so i, d, and f are treated as variables.



Presentation MathML (SLT)



Content MathML (OPT)

Presentation MathML (SLT)

```
<math xmlns="http://.../MathML">
  <mi>i</mi>
  <mi>d</mi>
  <mi>f</mi>
  <mo stretchy="false">(</mo>
  <msub>
    <mi>t</mi>
    <mi>i</mi>
  </msub>
  <mo stretchy="false">)</mo>
  <mo>=</mo>
  <mi>log</mi>
  <mo>&#x2061;</mo>
  <mfrac>
    <mi>N</mi>
    <msub>
      <mi>n</mi>
      <mi>i</mi>
    </msub>
  </mfrac>
</math>
```

Content MathML (OPT)

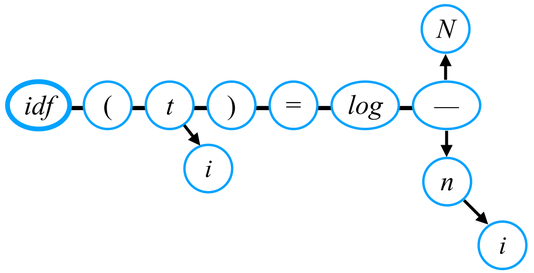
```
<math xmlns="http://.../MathML">
<apply>
  <eq/>
  <apply>
    <times/>
    <ci>i</ci>
    <ci>d</ci>
    <ci>f</ci>
    <msub>
      <ci>t</ci>
      <ci>i</ci>
    </msub>
  </apply>
  <apply>
    <log/>
    <apply>
      <divide/>
      <ci>N</ci>
      <msub>
        <ci>n</ci>
        <ci>i</ci>
      </msub>
    </apply>
  </apply>
</apply>
</math>
```

[a] https://math.nist.gov/~BMiller/LaTeXML

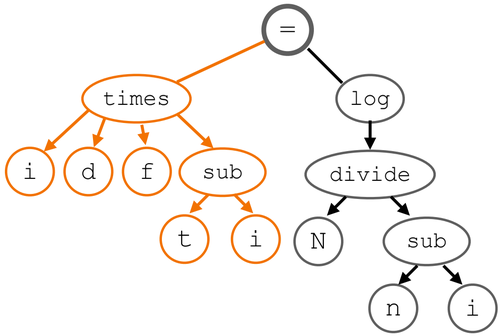
LaTeXML SLT → OPT Translation

Example 2.8: MathML generated from LaTeX using LaTeXML

idf is undefined in LaTeXML[a] and so i, d, and f are treated as variables.



Presentation MathML (SLT)



Content MathML (OPT)

Presentation MathML (SLT)

```
<math xmlns="http://.../MathML">
  <mi>i</mi>
  <mi>d</mi>
  <mi>f</mi>
  <mo stretchy="false">(</mo>
  <msub>
    <mi>t</mi>
    <mi>i</mi>
  </msub>
  <mo stretchy="false">)</mo>
  <mo>=</mo>
  <mi>log</mi>
  <mo>&#x2061;</mo>
  <mfrac>
    <mi>N</mi>
    <msub>
      <mi>n</mi>
      <mi>i</mi>
    </msub>
  </mfrac>
</math>
```

Content MathML (OPT)

```
<math xmlns="http://.../MathML">
<apply>
  <eq/>
  <apply>
    <times/>
    <ci>i</ci>
    <ci>d</ci>
    <ci>f</ci>
    <msub>
      <ci>t</ci>
      <ci>i</ci>
    </msub>
  </apply>
  <apply>
    <log/>
    <apply>
      <divide/>
      <ci>N</ci>
      <msub>
        <ci>n</ci>
        <ci>i</ci>
      </msub>
    </apply>
  </apply>
</apply>
</math>
```

[a] https://math.nist.gov/~BMiller/LaTeXML

Formula Semantics from OPT + Interpretation Context

e.g., in Python

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

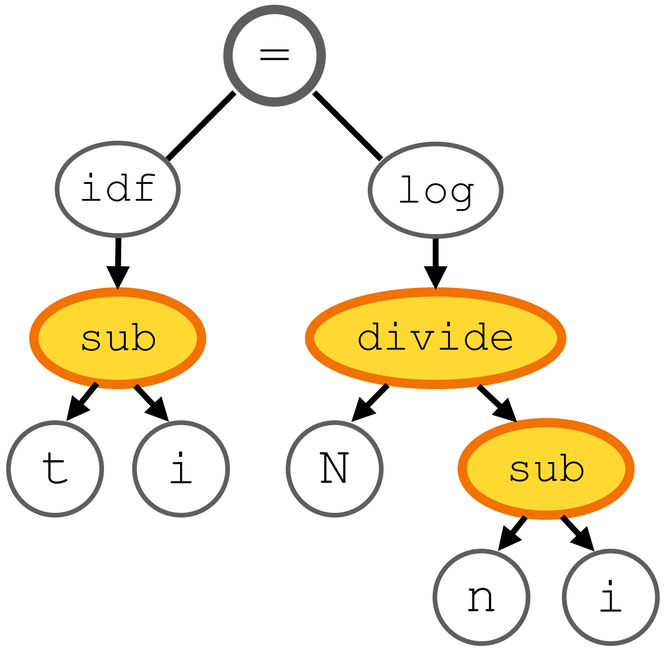
Example 2.7: *idf* formula in LaTeX and Python code

Python
Operator Tree Code, v1

Python
Operator Tree Code, v2

Python Output (for v1)

Operator Tree (OPT)



026
ool on
nalysis

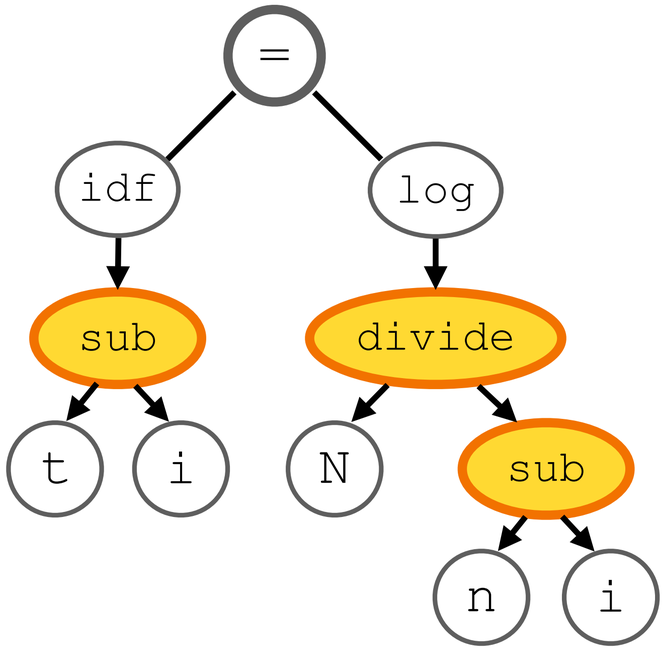
Formula Semantics from OPT + Interpretation Context

e.g., in Python

Interpretation Context

1. **= is assignment; *subscript* is array access**
2. **Select known ops from Python functions**
3. **Select var types & precision**

Operator Tree (OPT)



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

Example 2.7: *idf* formula in LaTeX and Python code

Python
Operator Tree Code, v1

Python
Operator Tree Code, v2

Python Output (for v1)

026
ool on
nalysis

Formula Semantics from OPT + Interpretation Context

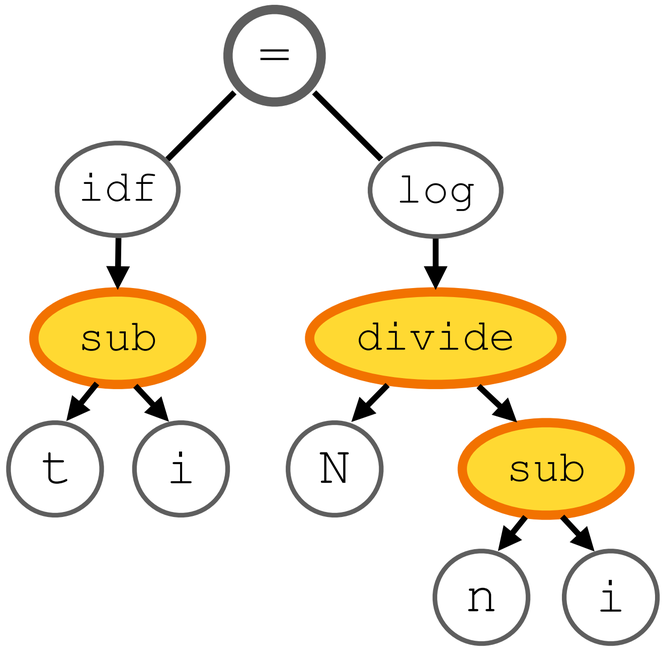
e.g., in Python

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

Interpretation Context

1. **= is assignment; *subscript* is array access**
2. **Select known ops from Python functions**
3. **Select var types & precision**

Operator Tree (OPT)



Example 2.7: *idf* formula in LaTeX and Python code

Python: Two Operator Tree representations

```
import math
t_all = [ "inverse", "document", "frequency" ]
n_all = [ 2, 100, 20 ]
D = 100
def idf(i, t, n, N):
    idf_weight = math.log( N / n[i] )
    return( t[i], idf_weight )
```

Python Operator Tree Code, v2

Python Output (for v1)

026
ool on
nalysis

Formula Semantics from OPT + Interpretation Context

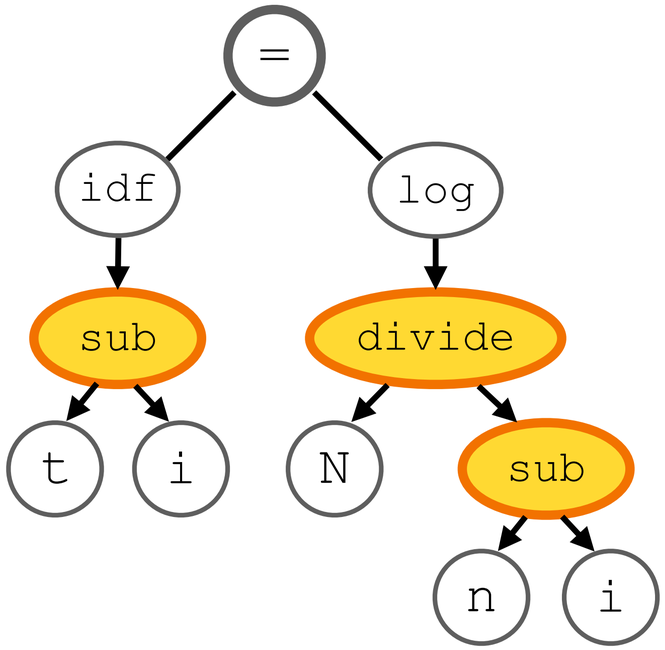
e.g., in Python

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

Interpretation Context

1. **= is assignment; *subscript* is array access**
2. **Select known ops from Python functions**
3. **Select var types & precision**

Operator Tree (OPT)



Example 2.7: *idf* formula in LaTeX and Python code

Python: Two Operator Tree representations

```
import math
t_all = [ "inverse", "document", "frequency" ]
n_all = [ 2, 100, 20 ]
D = 100
def idf(i, t, n, N):
    idf_weight = math.log( N / n[i] )
    return( t[i], idf_weight )

# Prefix form: ops before args to match OPT RHS
def divide(a, b): return a / b
def sub(a, b): return a[b]
def idf_prefix(i, t, n, N):
    idf_weight = math.log( divide( N, sub(n, i)) )
     return( sub(t, i), idf_weight )
```

Python Output (for v1)

026
ool on
nalysis

Formula Semantics from OPT + Interpretation Context

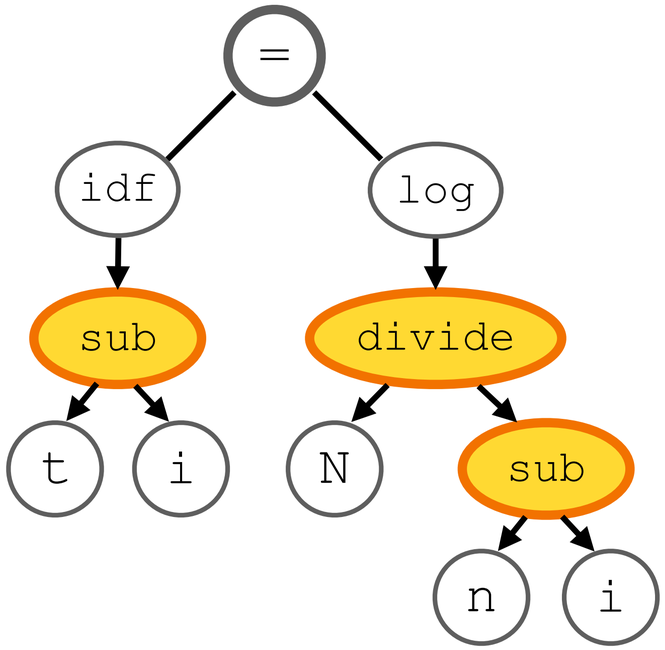
e.g., in Python

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

Interpretation Context

1. **= is assignment; *subscript* is array access**
2. **Select known ops from Python functions**
3. **Select var types & precision**

Operator Tree (OPT)



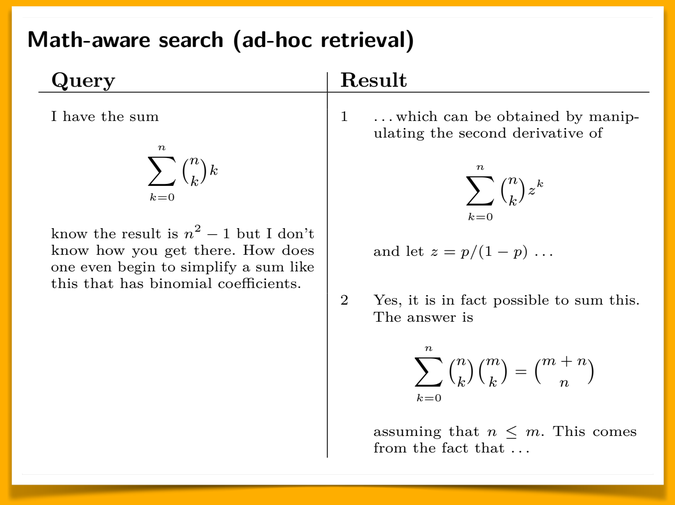
Example 2.7: *idf* formula in LaTeX and Python code

Python: Two Operator Tree representations

```
import math
t_all = [ "inverse", "document", "frequency" ]
n_all = [ 2, 100, 20 ]
D = 100
def idf(i, t, n, N):
   idf_weight = math.log( N / n[i] )
   return( t[i], idf_weight )

# Prefix form: ops before args to match OPT RHS
def divide(a, b): return a / b
def sub(a, b): return a[b]
def idf_prefix(i, t, n, N):
    idf_weight = math.log( divide( N, sub(n, i)) )
    return( sub(t, i), idf_weight )

for i in range(len(t_all)):
   print(i, idf(i, t_all, n_all, D))

OUTPUT:  0 ('inverse', 3.912023005428146)
         1 ('document', 0.0)
         2 ('frequency', 1.6094379124341003)
```

026
ool on
nalysis

Retrieving Formulas
&
(Multimodal) Math-Aware Search



Math-aware search (ad-hoc retrieval)

Query	Result
I have the sum $$\sum_{k=0}^{n} \binom{n}{k} k$$ know the result is $n^2 - 1$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	1 . . . which can be obtained by manipulating the second derivative of $$\sum_{k=0}^{n} \binom{n}{k} z^k$$ and let $z = p/(1-p)$ . . .
	2 Yes, it is in fact possible to sum this. The answer is $$\sum_{k=0}^{n} \binom{n}{k}\binom{m}{k} = \binom{m+n}{n}$$ assuming that $n \leq m$. This comes from the fact that . . .

dprl
Document and Pattern Recognition Lab

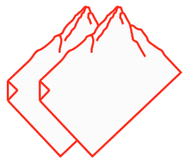
SSDA 2026
Summer School on
Document Analysis

Formula Search Tasks

Formula Similarity

Formula Similarity
w. Wildcards

Contextualized
Formula Search



SSDA 2026
Summer School on
Document Analysis

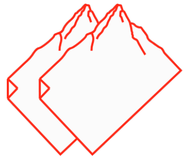
Formula Search Tasks

Formula similarity search

Query	Result 1	2	3	4	5
$y = \frac{a+bx}{b-x}$	$y = \frac{a+bx}{c+x}$	$y = a + bx$	$y = \frac{a-bx}{c-x}$	$y = \frac{a+bx}{x+c}$	$g(x) = \frac{x}{x-a}$

Formula Similarity
w. Wildcards

Contextualized
Formula Search



SSDA 2026
Summer School on
Document Analysis

Formula Search Tasks

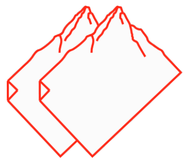
Formula similarity search

Query	Result 1	2	3	4	5
$y = \frac{a+bx}{b-x}$	$y = \frac{a+bx}{c+x}$	$y = a + bx$	$y = \frac{a-bx}{c-x}$	$y = \frac{a+bx}{x+c}$	$g(x) = \frac{x}{x-a}$

Formula similarity with wildcards (?w) (Aizawa *et al.*, 2013)

Query	Result	
$\frac{?\mathbf{f}(?\mathbf{v} + ?\mathbf{d}) - ?\mathbf{f}(?\mathbf{v})}{?\mathbf{d}}$	1	$g'(cx) = \lim_{h \to 0} \frac{\mathbf{g}(\mathbf{cx} + \mathbf{h}) - \mathbf{g}(\mathbf{cx})}{\mathbf{h}}$
	...	...

Contextualized
Formula Search



SSDA 2026
Summer School on
Document Analysis

Formula Search Tasks

Formula similarity search

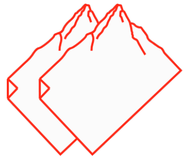
Query	Result				
	1	2	3	4	5
$y = \frac{a+bx}{b-x}$	$y = \frac{a+bx}{c+x}$	$y = a + bx$	$y = \frac{a-bx}{c-x}$	$y = \frac{a+bx}{x+c}$	$g(x) = \frac{x}{x-a}$

Formula similarity with wildcards (?w) (Aizawa *et al.*, 2013)

Query	Result	
$\frac{?\mathbf{f}(?\mathbf{v} + ?\mathbf{d}) - ?\mathbf{f}(?\mathbf{v})}{?\mathbf{d}}$	1	$g'(cx) = \lim_{h \to 0} \frac{\mathbf{g(cx + h)} - \mathbf{g(cx)}}{\mathbf{h}}$
	...	...

Contextualized formula search

Query	Result
I have the sum $\boxed{\sum_{k=0}^{n} \binom{n}{k} k}$ know the result is $n^2 - 1$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	1 ... which can be obtained by manipulating the second derivative of $\boxed{\sum_{k=0}^{n} \binom{n}{k} z^k}$ and let $z = p/(1-p)$ ... 2 Yes, it is in fact possible to sum this. The answer is $\boxed{\sum_{k=0}^{n} \binom{n}{k} \binom{m}{k} = \binom{m+n}{n}}$ assuming that $n \leq m$. This comes from the fact that ...



SSDA 2026
Summer School on
Document Analysis

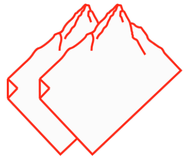
Math-Aware Search

Searching with Formulas & Text (‘Multimodal’)

Ad-hoc Retrieval

Math Question Answering

Word Problems



SSDA 2026
Summer School on
Document Analysis

Math-Aware Search

Searching with Formulas & Text ('Multimodal')

Math-aware search (ad-hoc retrieval)

Query	Result
I have the sum $$\sum_{k=0}^{n} \binom{n}{k} k$$ know the result is $n^2 - 1$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	1 ...which can be obtained by manipulating the second derivative of $$\sum_{k=0}^{n} \binom{n}{k} z^k$$ and let $z = p/(1-p)$ ...
	2 Yes, it is in fact possible to sum this. The answer is $$\sum_{k=0}^{n} \binom{n}{k} \binom{m}{k} = \binom{m+n}{n}$$ assuming that $n \leq m$. This comes from the fact that ...

Math Question Answering

Word Problems

SSDA 2026
Summer School on
Document Analysis

Math-Aware Search

Searching with Formulas & Text ('Multimodal')

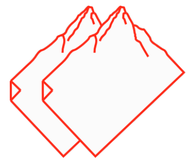
Math-aware search (ad-hoc retrieval)

Query	Result
I have the sum $$\sum_{k=0}^{n} \binom{n}{k} k$$ know the result is $n^2 - 1$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	1 ...which can be obtained by manipulating the second derivative of $$\sum_{k=0}^{n} \binom{n}{k} z^k$$ and let $z = p/(1-p)$ ...
	2 Yes, it is in fact possible to sum this. The answer is $$\sum_{k=0}^{n} \binom{n}{k} \binom{m}{k} = \binom{m+n}{n}$$ assuming that $n \leq m$. This comes from the fact that ...

Math Question Answering (Mansouri *et al.*, 2022a)

Query	Result
What does it mean for a matrix to be Hermitian?	A matrix is Hermitian if it is equal to its transpose conjugate.

Word Problems



SSDA 2026
Summer School on
Document Analysis

Math-Aware Search

Searching with Formulas & Text ('Multimodal')

Math-aware search (ad-hoc retrieval)

Query	Result
I have the sum $$\sum_{k=0}^{n} \binom{n}{k} k$$ know the result is $n^2 - 1$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.	1 ...which can be obtained by manipulating the second derivative of $$\sum_{k=0}^{n} \binom{n}{k} z^k$$ and let $z = p/(1-p)$ ... 2 Yes, it is in fact possible to sum this. The answer is $$\sum_{k=0}^{n} \binom{n}{k}\binom{m}{k} = \binom{m+n}{n}$$ assuming that $n \leq m$. This comes from the fact that ...

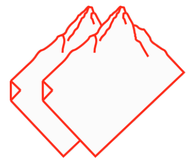
Math Question Answering (Mansouri *et al.*, 2022a)

Query	Result
What does it mean for a matrix to be Hermitian?	A matrix is Hermitian if it is equal to its transpose conjugate.

Math word problems

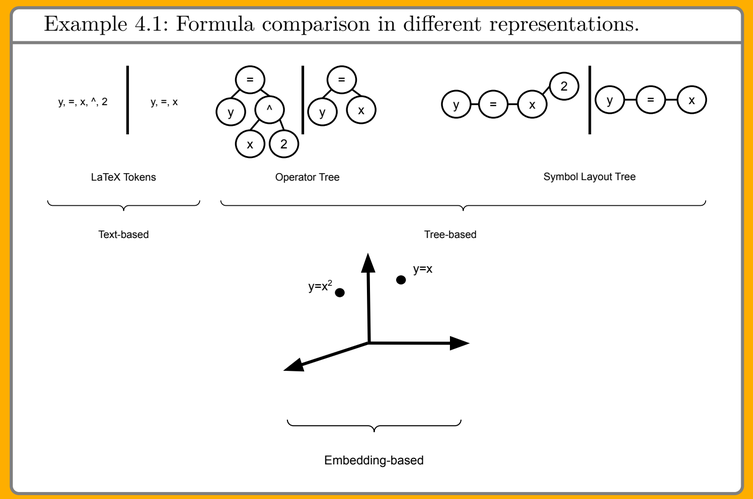
Query	Result	
	Equation	*Answer*
Sarah has 5 pens, David has 3 pens. How many pens do they have?	$x = 5 + 3$	8
Find two consecutive integers whose sum is 7.	$x + (x + 1) = 7$	3, 4

From MathQA (Amini *et al.*, 2019) & Dolphin18K (Huang *et al.*, 2016)



SSDA 2026
Summer School on
Document Analysis

Representing Math Formulas for Retrieval

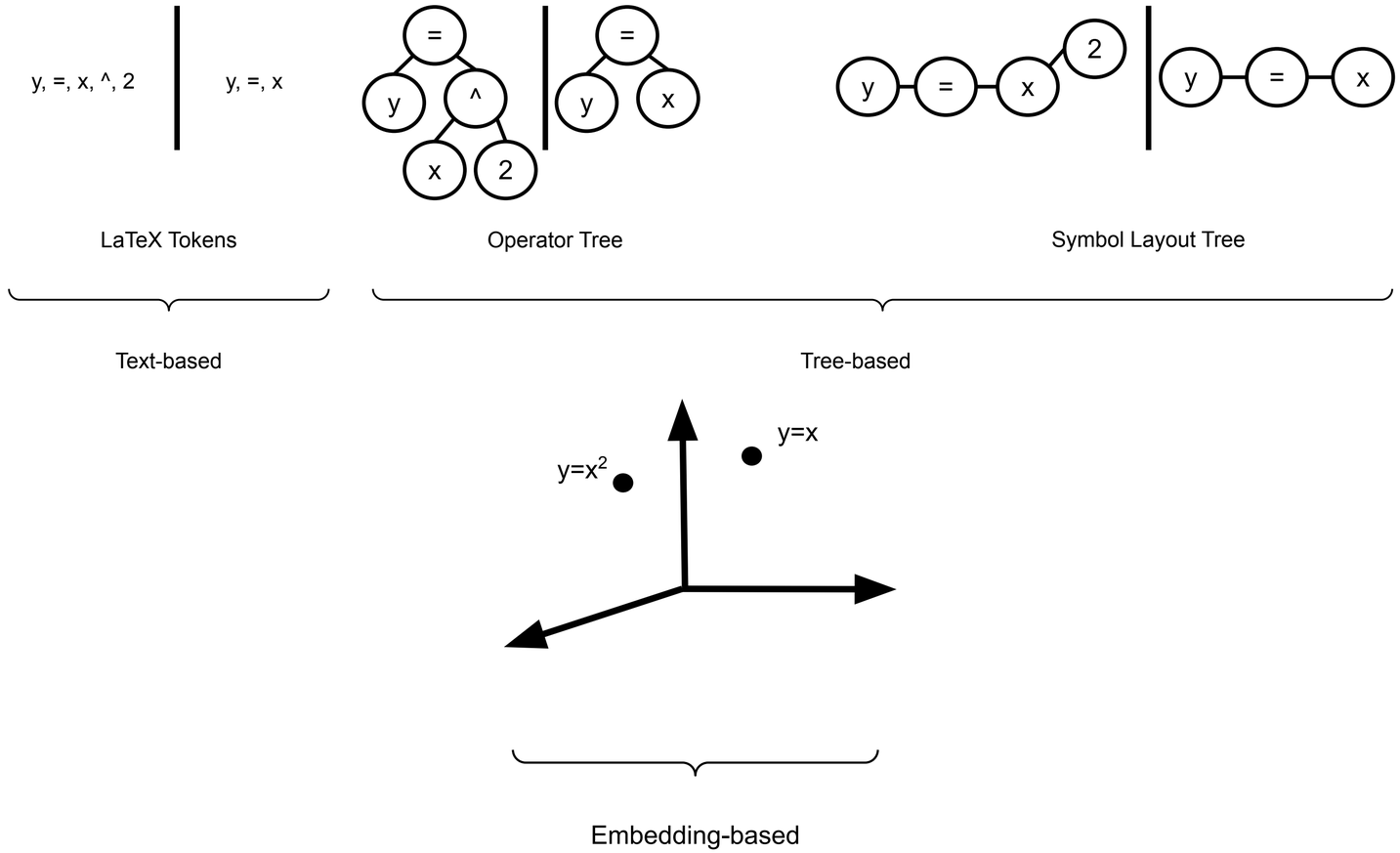


dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on
Document Analysis

Common Representations for Formula Retrieval

Example 4.1: Formula comparison in different representations.

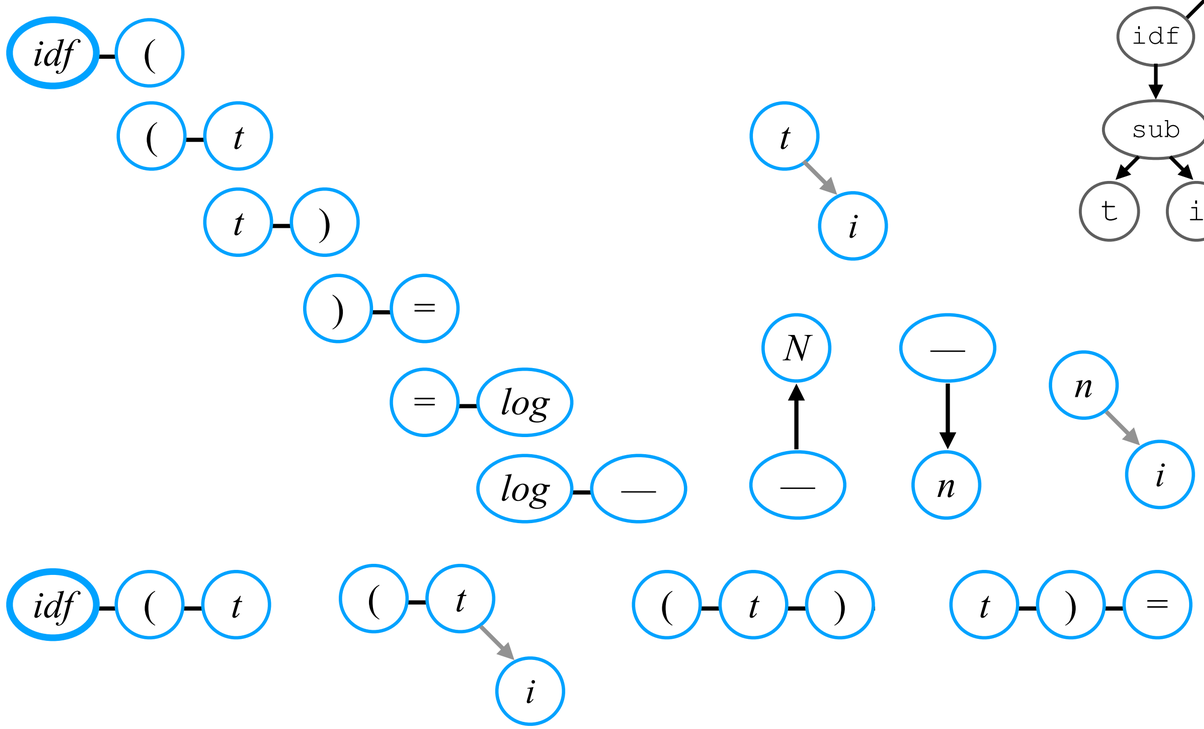


Our choice of representation(s) limit the **types of information**, and **patterns** within information types that can be searched **effectively**

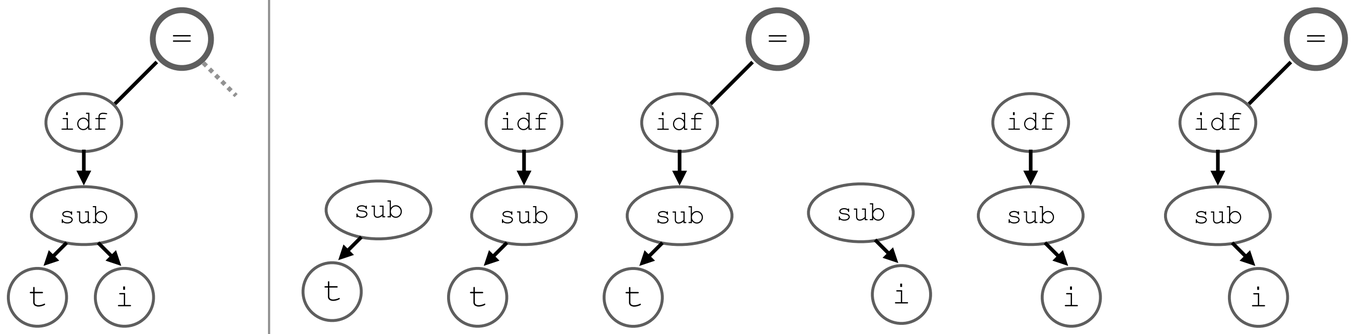
SSDA 2026
Summer School on Document Analysis

Path-Based Indexing

Symbol Layout Tree Paths (directed)



Operator Tree Paths (leaf-root)



**Also used to create inputs to many embedding-based retrieval models*

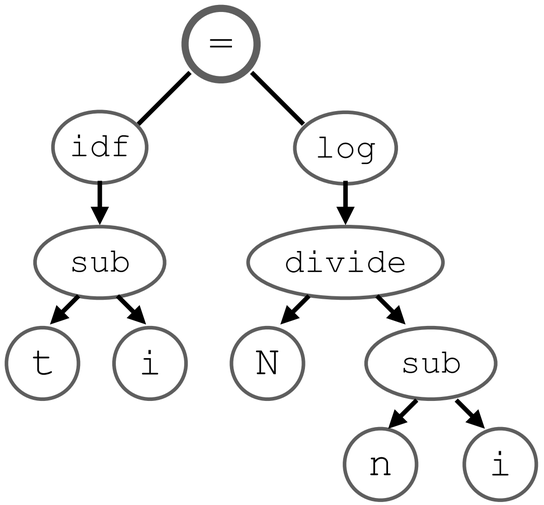
SSDA 2026
Summer School on Document Analysis

Common Abstractions for Formula Search

Variable Enumeration and Types

Example 2.10: OPT Variable Enumeration and Symbol Types

Original Operator Tree



Variable Enumeration

Assigning Types Symbols/Nodes

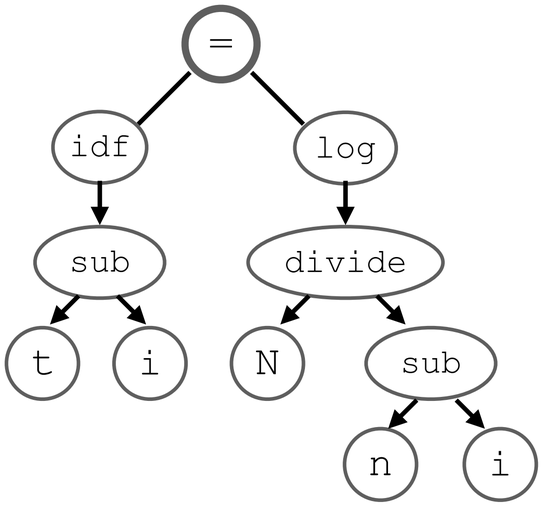
SSDA 2026
Summer School on Document Analysis

Common Abstractions for Formula Search

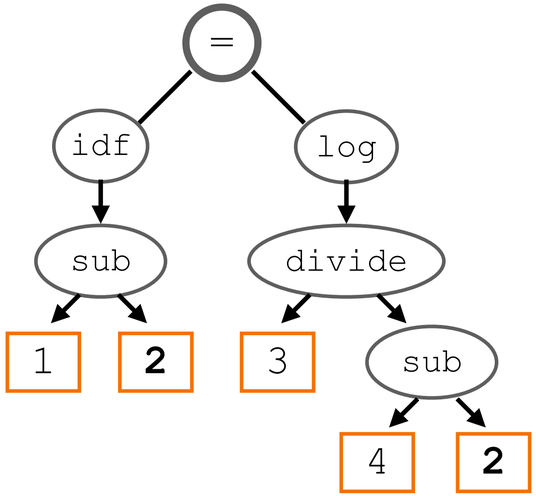
Variable Enumeration and Types

Example 2.10: OPT Variable Enumeration and Symbol Types

Original Operator Tree



Enumerated Variables



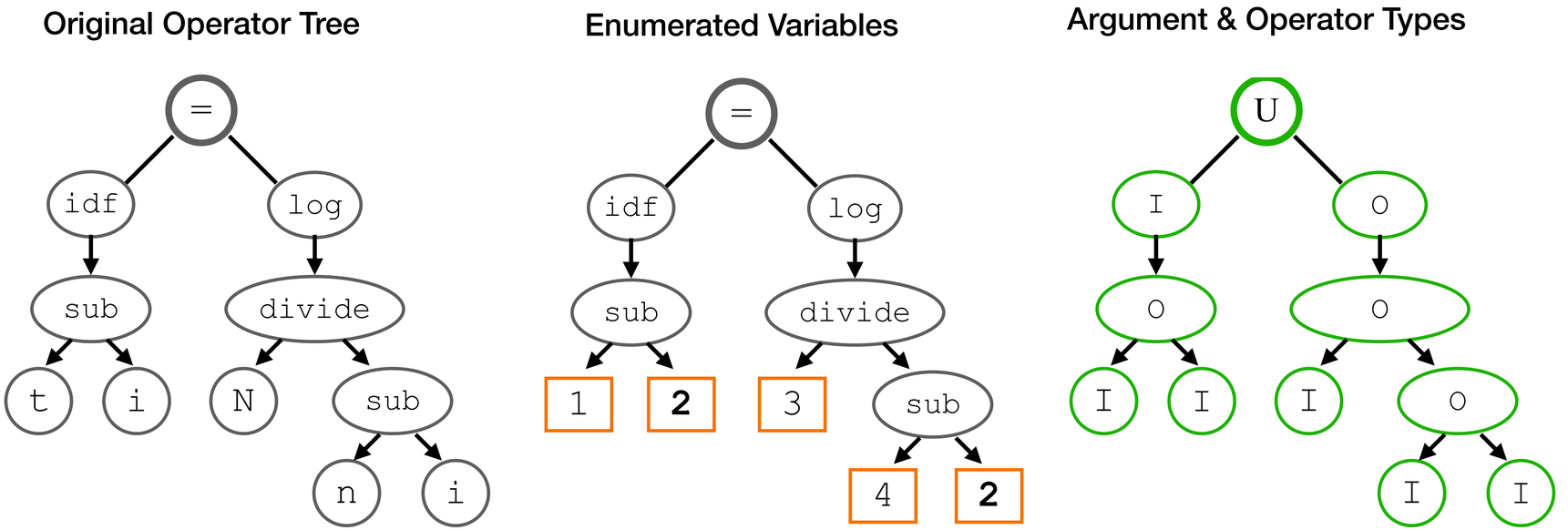
Assigning Types
Symbols/Nodes

SSDA 2026
Summer School on
Document Analysis

Common Abstractions for Formula Search

Variable Enumeration and Types

Example 2.10: OPT Variable Enumeration and Symbol Types



U: Operator w. **unordered** args

O: Operator w. **ordered** args

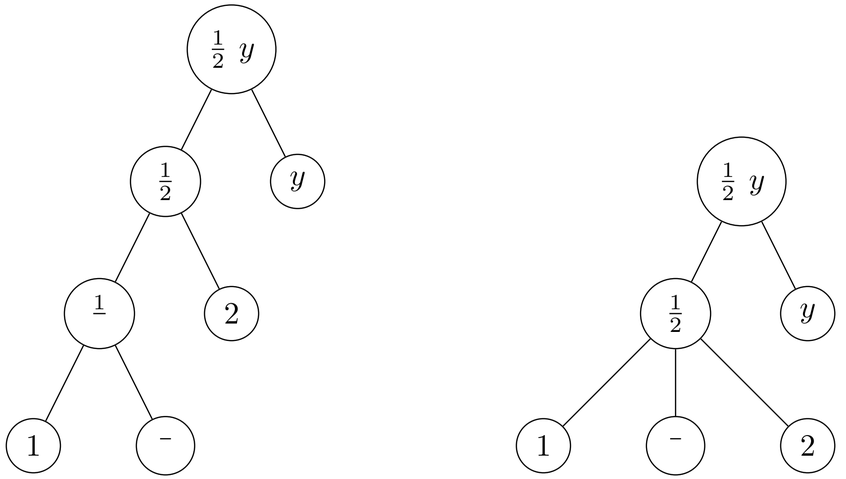
I: Identifier

SSDA 2026
Summer School on Document Analysis

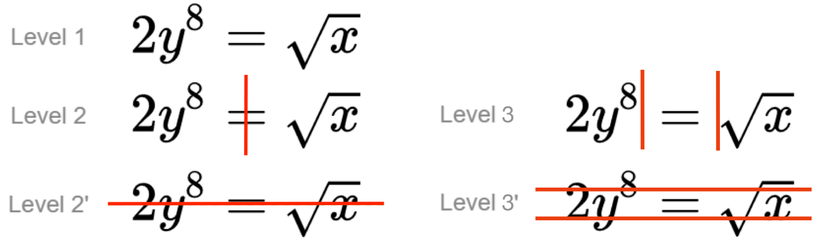
Spatial Formula Representations

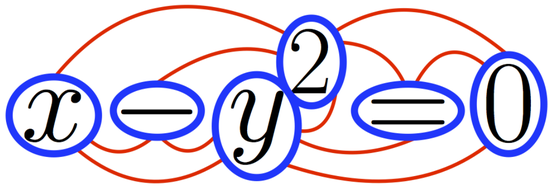
Some examples

XY-Cut Trees (left: Recursive, right: Standard)

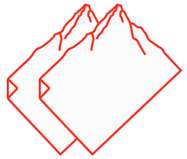


Pyramidal Histogram of Characters (XY-PHOC)





Appearance (LOS)



SSDA 2026
Summer School on Document Analysis

Tangent-V Demonstration:

Handwritten Formula Search & Content-Based Navigation in Lecture Videos

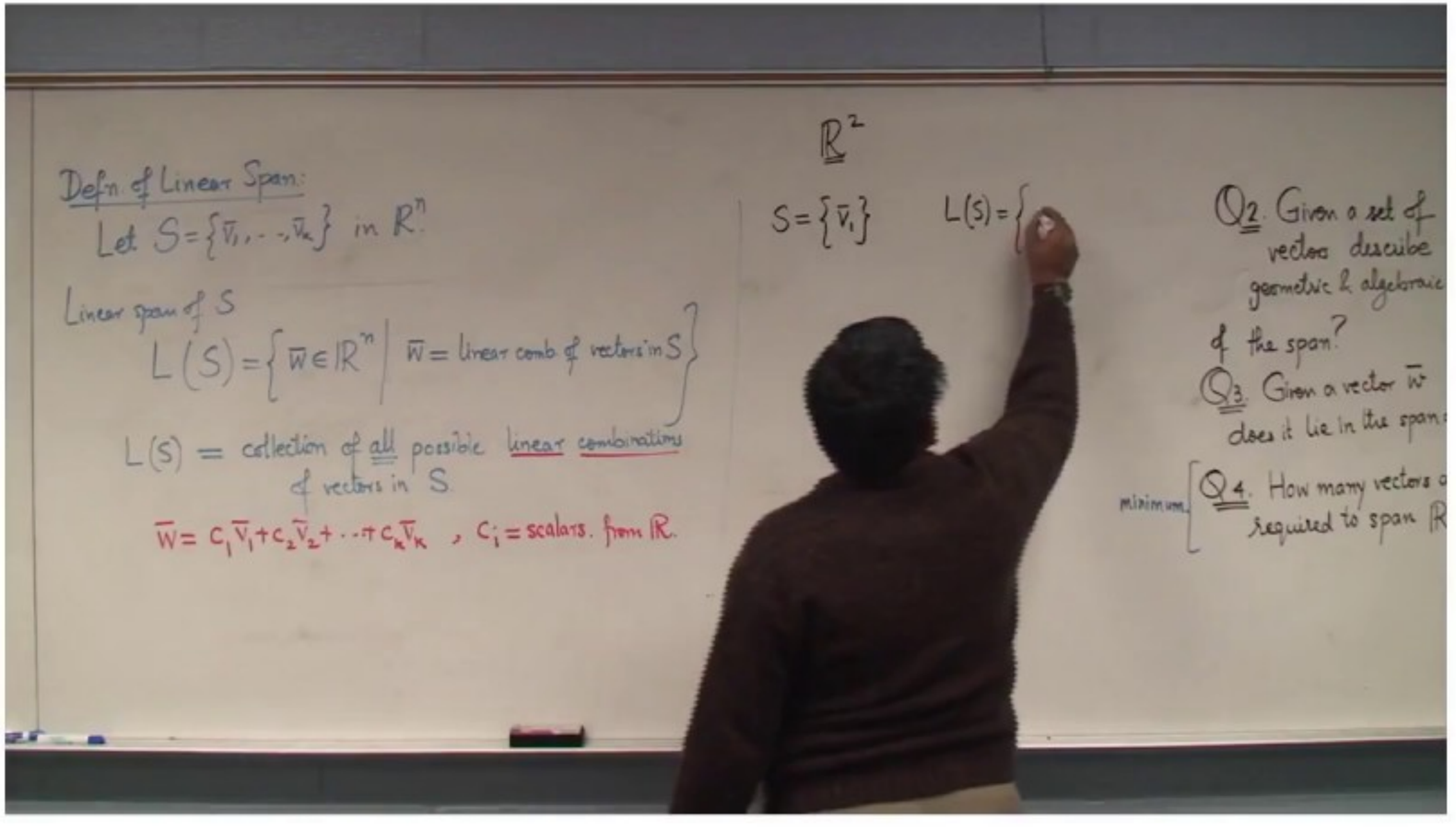
Kenny Davila, Richard Zanibbi:
Visual Search Engine for Handwritten and Typeset Math in Lecture Videos and LATEX Notes. ICFHR 2018: 50-55

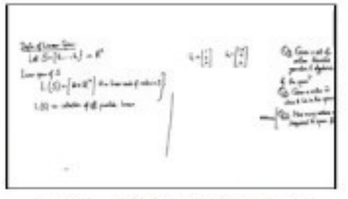
dprl
Document and Pattern Recognition Lab

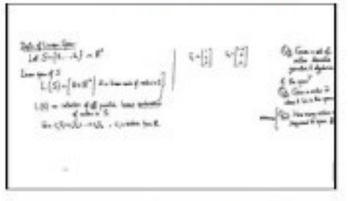
SSDA 2026
Summer School on Document Analysis

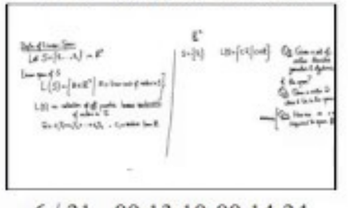
View Lecture - NM_lecture_05

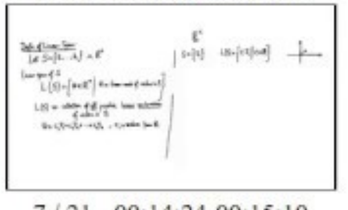
Normal | Binary | Content




4 / 21 - 00:09:59-00:11:33


5 / 21 - 00:11:33-00:13:10

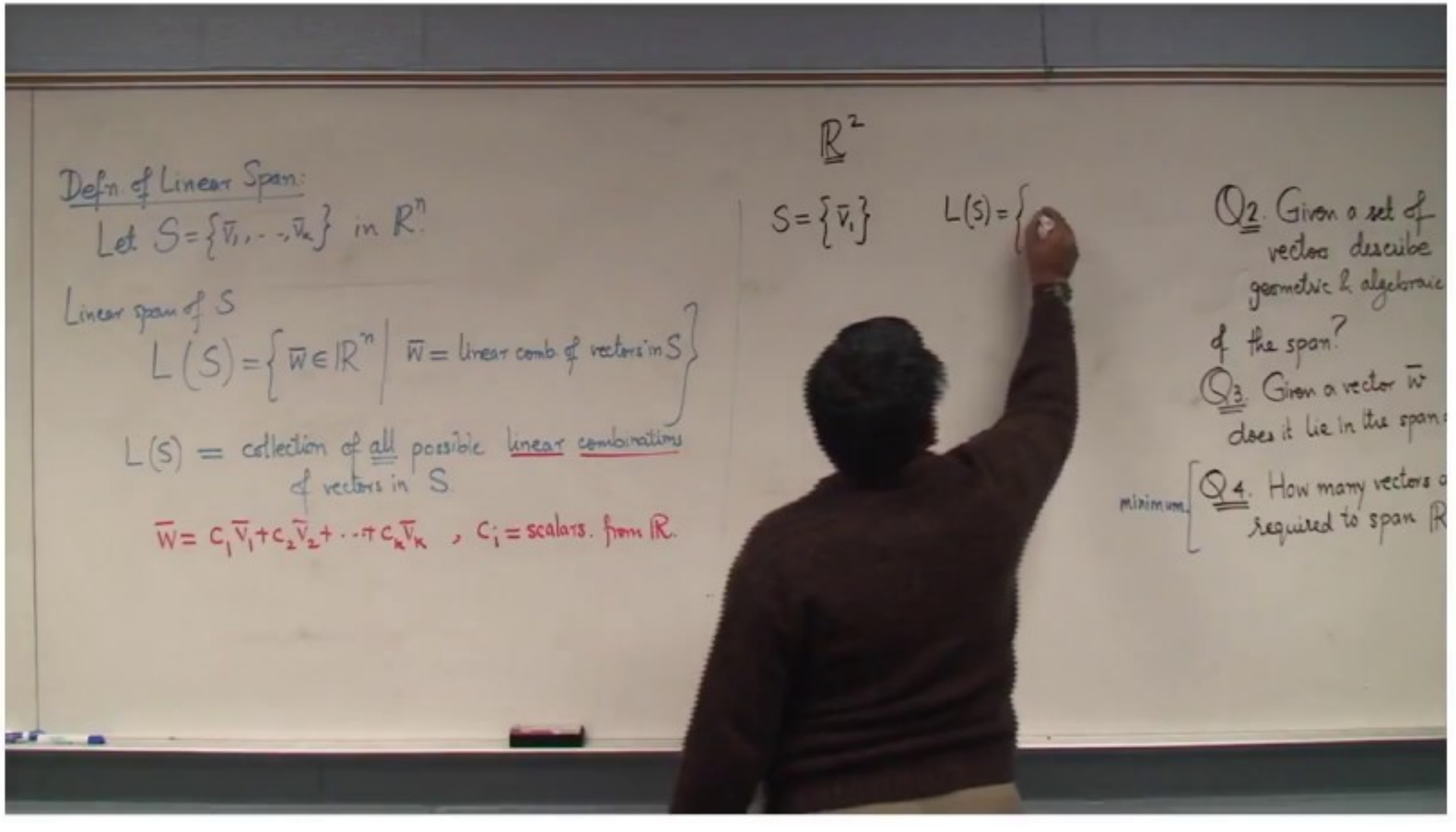

6 / 21 - 00:13:10-00:14:24

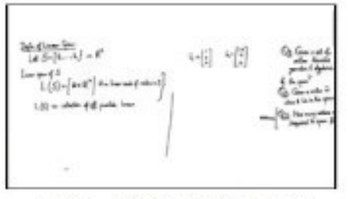

7 / 21 - 00:14:24-00:15:10

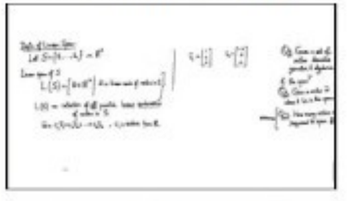
<< Prev Next >>

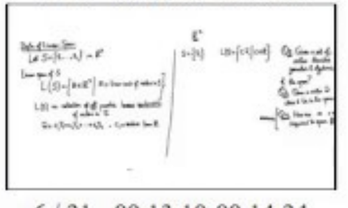
View Lecture - NM_lecture_05

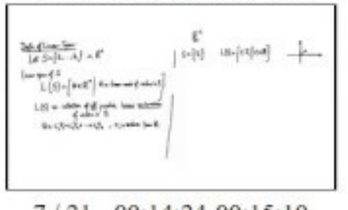
Normal | Binary | Content




4 / 21 - 00:09:59-00:11:33

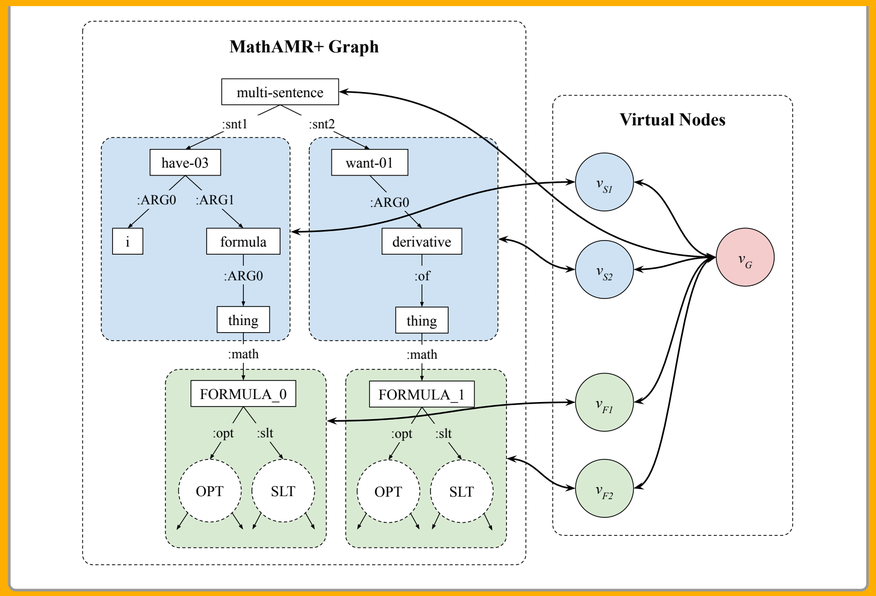

5 / 21 - 00:11:33-00:13:10


6 / 21 - 00:13:10-00:14:24


7 / 21 - 00:14:24-00:15:10

<< Prev Next >>

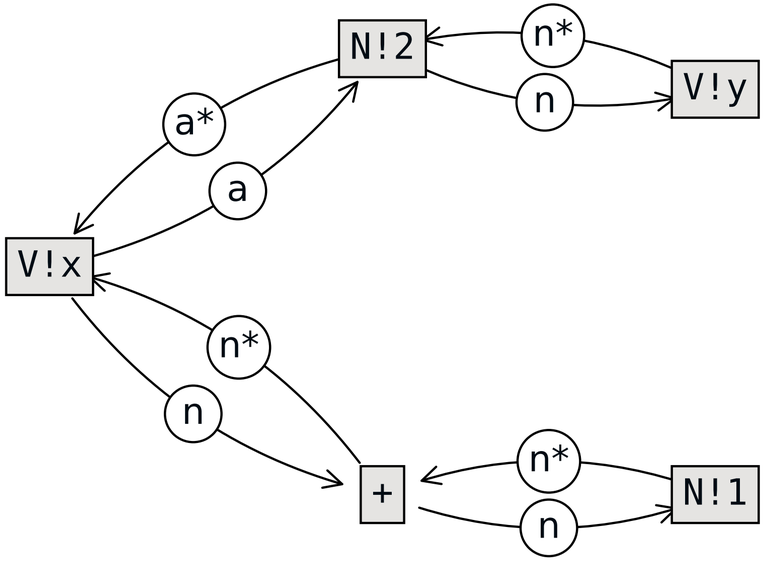
Graph Neural Networks for Formula & Text+Formula Retrieval



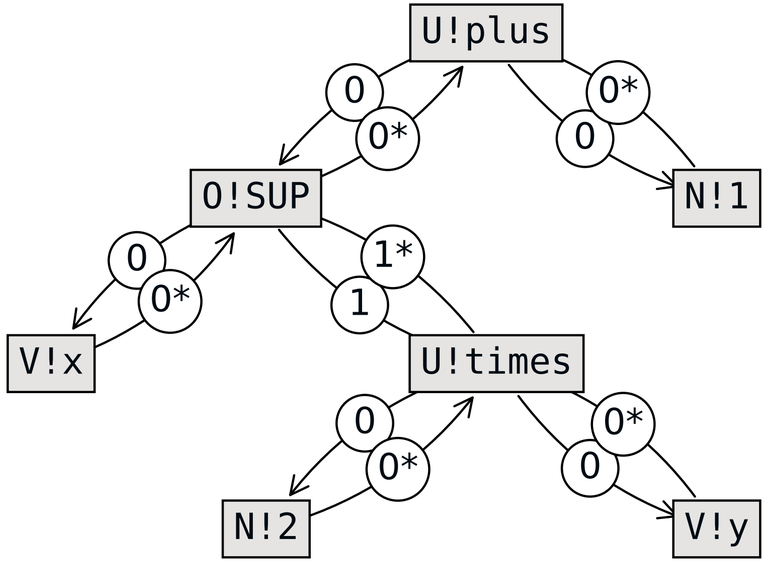
dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on Document Analysis

Node-Based Retrieval Using Graph Neural Nets



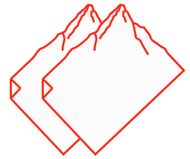
(a) Symbol Layout Graph (SLG)



(b) Operator Graph (OPG)

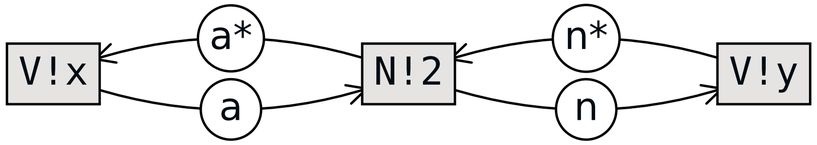
$$x^{2y} + 1$$

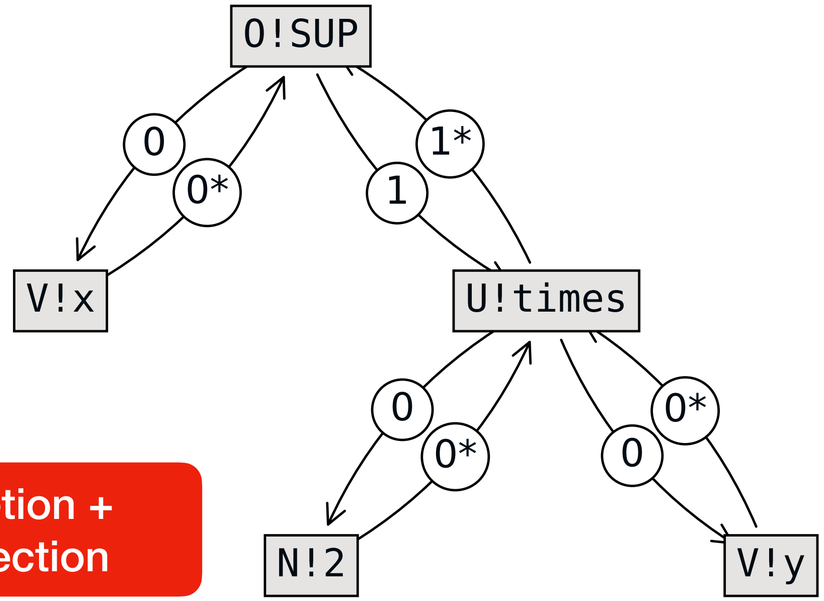
Bryan Amador, Richard Zanibbi:
Math Formula Graph Retrieval Using Contrastive Learning Over Visual and Semantic Embeddings.
ICTIR 2025: 230-237



SSDA 2026
Summer School on
Document Analysis

Node-Based Retrieval Using Graph Neural Nets





Random Node Deletion +
Larger Subtree Selection

(c) SLG sub-expression

(d) OPG sub-expression

Bryan Amador, Richard Zanibbi:
Math Formula Graph Retrieval Using Contrastive Learning Over Visual and Semantic Embeddings.
ICTIR 2025: 230-237

SSDA 2026
Summer School on
Document Analysis

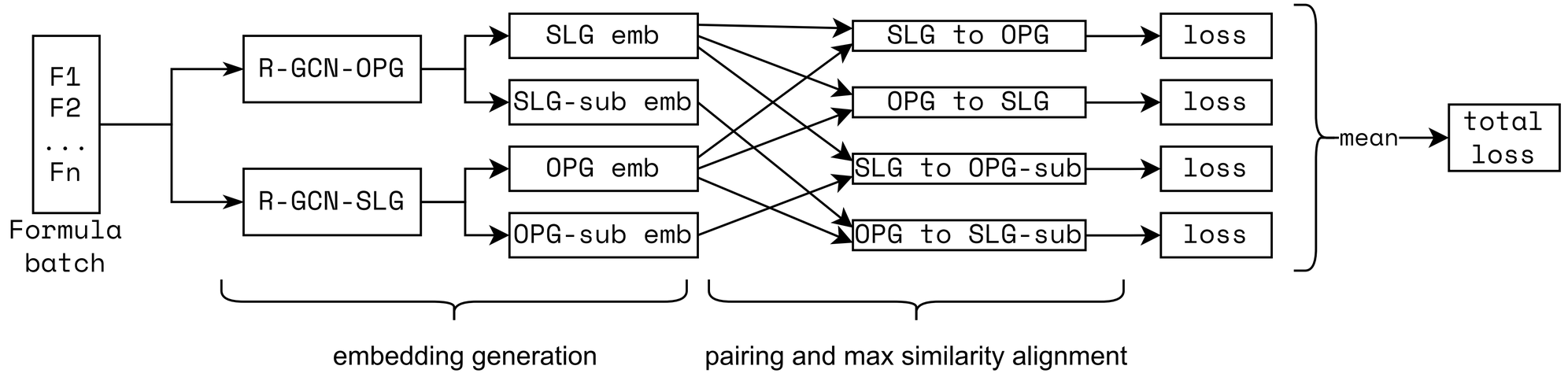
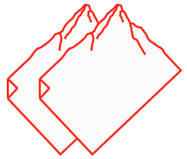


Figure 3: Training with contrastive learning across SLG and OPG nodes using the `Outer` set of graph augmentations. Positive and negative node pairs are created by maxsim alignment of nodes from the first to second type in each graph pairing. This occurs both for complete formulas (i.e., whole OPGs/SLGs), and for complete formulas in one representation vs. sub-expressions in the other representation (e.g., (SLG, OPG-sub). The final loss is averaged across the four graph type pairings.

Bryan Amador, Richard Zanibbi:
Math Formula Graph Retrieval Using Contrastive Learning Over Visual and Semantic Embeddings.
ICTIR 2025: 230-237



SSDA 2026
Summer School on Document Analysis

Table 2. ARQMath-3 Formula Search Task Results (76 Test Queries). Approach0 is a state-of-the-art formula retrieval model. Text + formula search results are shown for context

Model	NDCG'	MAP'	P'@10
Formula Search			
Approach-0 [4]	0.639	0.501	0.615
Inner+Outter MAX (ours)	**0.701**	**0.505**	0.597
**Inner RRF (ours, updated)*	**0.773**	**0.554**	**0.632**
Text + Formula Search			
Approach0+ColBERT[4]	**0.720**	**0.568**	**0.688**
TanCFT MathAMR [5]	0.640	0.388	0.478

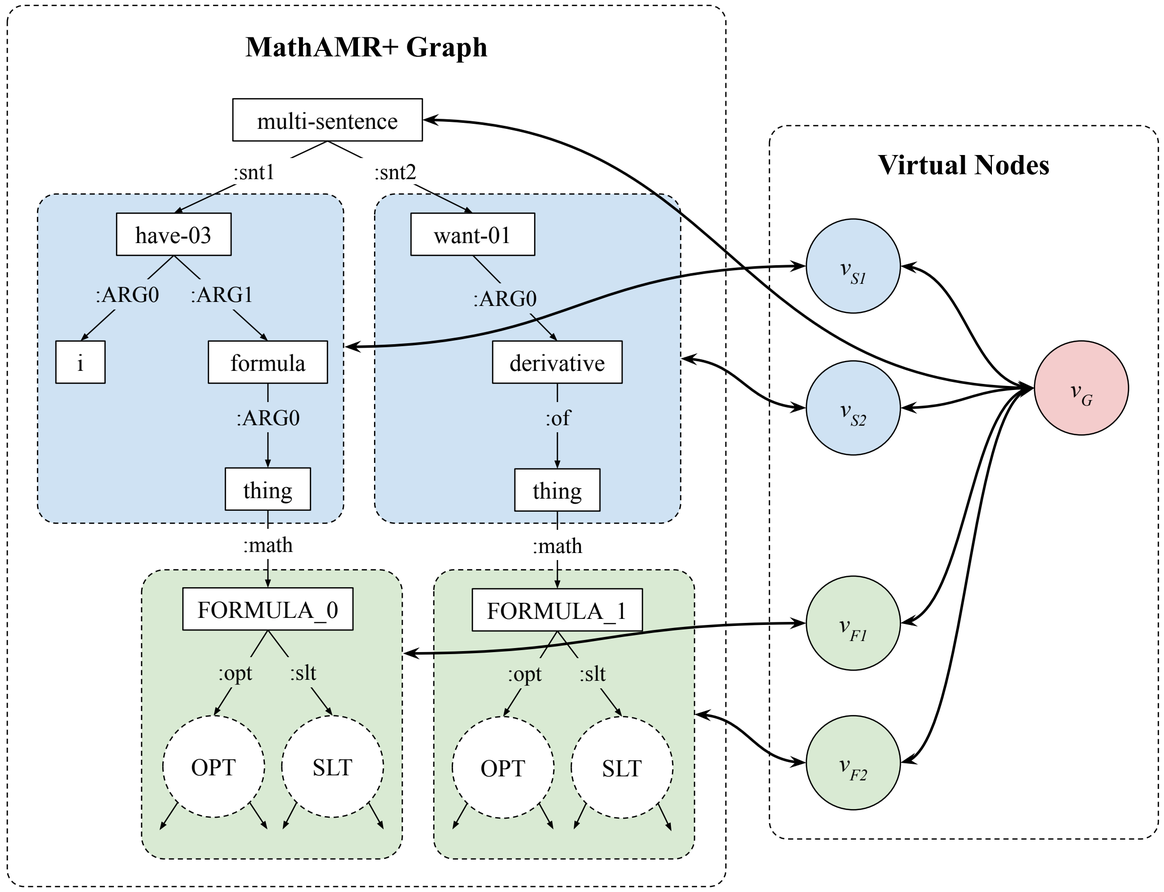
SSDA 2026
Summer School on
Document Analysis

Moving to Text + Formulas

Original Question

"I have the formula $f(x) = x^2 + 1$. What is the derivative of $f(x)$?"

MathAMR+ Graph with Hierarchical Virtual Nodes

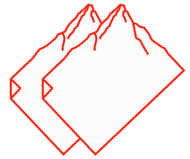


Using **Abstract Meaning Representation (AMR)** graphs produced by encoder-decoder

Formulas replaced by identifiers before AMR generation;
OPT + SLT added as id children

Addition of **virtual nodes** for:

- Whole post
- Each sentence
- Each formula



SSDA 2026
Summer School on Document Analysis

Results on (Multimodal) Answer Retrieval

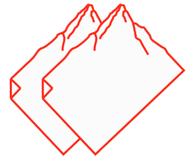
ARQMath-3 Benchmark

Table 4.19: ARQMath-3 Answer Retrieval (Task 1) benchmarking results.

System	nDCG'	MAP'	P'@10	Bpref
Approach0 / Struct.+ColBERT (Coco-MAE) [68]	**0.546**	**0.237**	0.360	0.221
Kassaie et al. / RRF+LLM [23]	0.522	0.195	0.277	–
MSM / Ensemble RRF [14]	0.504	0.157	0.241	0.138
MIRMU / MiniLM+RoBERTa [50]	0.498	0.184	0.267	0.169
Satpute et al. / GPT-4+DPR [52]	0.486	0.219	**0.374**	**0.225**
DPRL / QQ-QA-AMR [36]	0.185	0.040	0.091	–
Ours / MathAMR+	0.334	0.108	0.236	0.163

Embed size: 64
Batch size: 16

Jacob Yoon.
MathAMR+: A Unified Graph Neural Network Framework for Multimodal Mathematical Information Retrieval.
MSc Thesis, RIT, 2026.



SSDA 2026
Summer School on
Document Analysis

Results for Contextualized Formula Retrieval

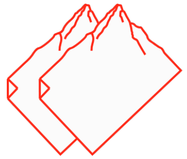
ARQMath-3 Benchmark

Table 4.21: ARQMath-3 Formula Retrieval (Task 2) benchmarking results.

System	nDCG'	MAP'	P'@10	Bpref
L2L / All (Unweighted) [60]	**0.849**	**0.620**	0.680	–
Amador et al. / Inner RRF† [4]	0.773	0.554	0.632	–
Approach0 / Struct.+ColBERT [68]	0.720	0.568	**0.688**	**0.560**
Amador et al. / In+Out [4]	0.701	0.505	0.597	–
DPRL / TanCFT+MathAMR [36]	0.681	0.471	0.617	–
MathDowsers / latex_L8_a040 [20]	0.640	0.451	0.549	0.443
DPRL / MathAMR†1 [36]	0.579	0.367	0.549	–
DPRL / MathAMR2 [35]	0.316	0.160	0.253	—
Ours / MathAMR+	0.507	0.333	0.588	0.503

Embed size: 64
Batch size: 16

† *Unofficial run by the authors*



SSDA 2026
Summer School on Document Analysis

Jacob Yoon.
MathAMR+: A Unified Graph Neural Network Framework for Multimodal Mathematical Information Retrieval.
MSc Thesis, RIT, 2026.

MathDeck Demonstration:

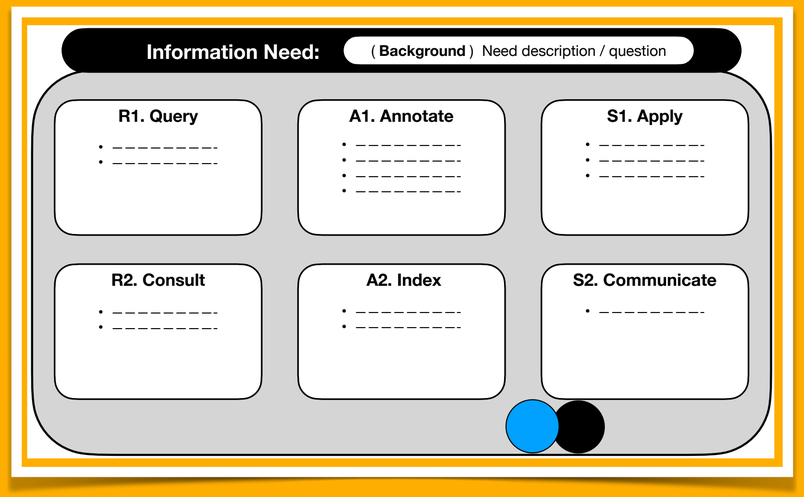
Multimodal Search in PDF Documents

Bryan Amador, Matt Langsenkamp, Abhisek Dey, Ayush Kumar Shah, Richard Zanibbi:
Searching the ACL Anthology with Math Formulas and Text. SIGIR 2023: 3110-3114

dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on
Document Analysis

The ‘Task Jar’ Framework for Human Retrieval: Needs, Tasks, and Sources



dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on Document Analysis

Information Needs

An Example from Math

Example 1.1: Differing Information Needs

Query: What does $a^2 + b^2 = c^2$ represent and how is it useful?

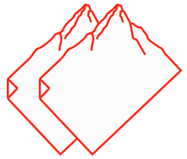
Students

Educators

Researchers

Information need is closely related to the **query intent**, i.e., which information a user hopes a prompt will retrieve

The background of the searcher strongly impacts relevance, in terms of **both pertinence, and interpretability**



SSDA 2026
Summer School on
Document Analysis

Information Needs

An Example from Math

Example 1.1: Differing Information Needs

Query: What does $a^2 + b^2 = c^2$ represent and how is it useful?

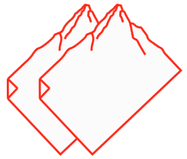
Students might use this query to learn the Pythagoras theorem, and perhaps find an example demonstrating the theorem, and a possible proof.

Educators

Researchers

Information need is closely related to the **query intent**, i.e., which information a user hopes a prompt will retrieve

The background of the searcher strongly impacts relevance, in terms of **both pertinence, and interpretability**



SSDA 2026
Summer School on Document Analysis

Information Needs

An Example from Math

Example 1.1: Differing Information Needs

Query: What does $a^2 + b^2 = c^2$ represent and how is it useful?

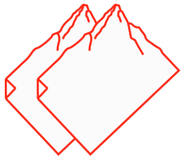
Students might use this query to learn the Pythagoras theorem, and perhaps find an example demonstrating the theorem, and a possible proof.

Educators may have similar interests to students, but may seek additional resources on how to teach this result.

Researchers

Information need is closely related to the **query intent**, i.e., which information a user hopes a prompt will retrieve

The background of the searcher strongly impacts relevance, in terms of **both pertinence, and interpretability**



SSDA 2026
Summer School on Document Analysis

Information Needs

An Example from Math

Example 1.1: Differing Information Needs

Query: What does $a^2 + b^2 = c^2$ represent and how is it useful?

Students might use this query to learn the Pythagoras theorem, and perhaps find an example demonstrating the theorem, and a possible proof.

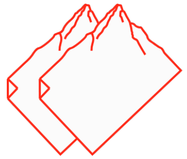
Educators may have similar interests to students, but may seek additional resources on how to teach this result.

Researchers can have very different interests than the other audiences. They may be interested in one or more of the following:

- For a mathematician: Is this true in a general metric space and/or a Hilbert space?
- For a physicist: How is it linked to the probability assignments in quantum mechanics?
- For an IR expert: How is it related to probability assignments in a Hilbert space used in describing interaction for information retrieval?

Information need is closely related to the **query intent**, i.e., which information a user hopes a prompt will retrieve

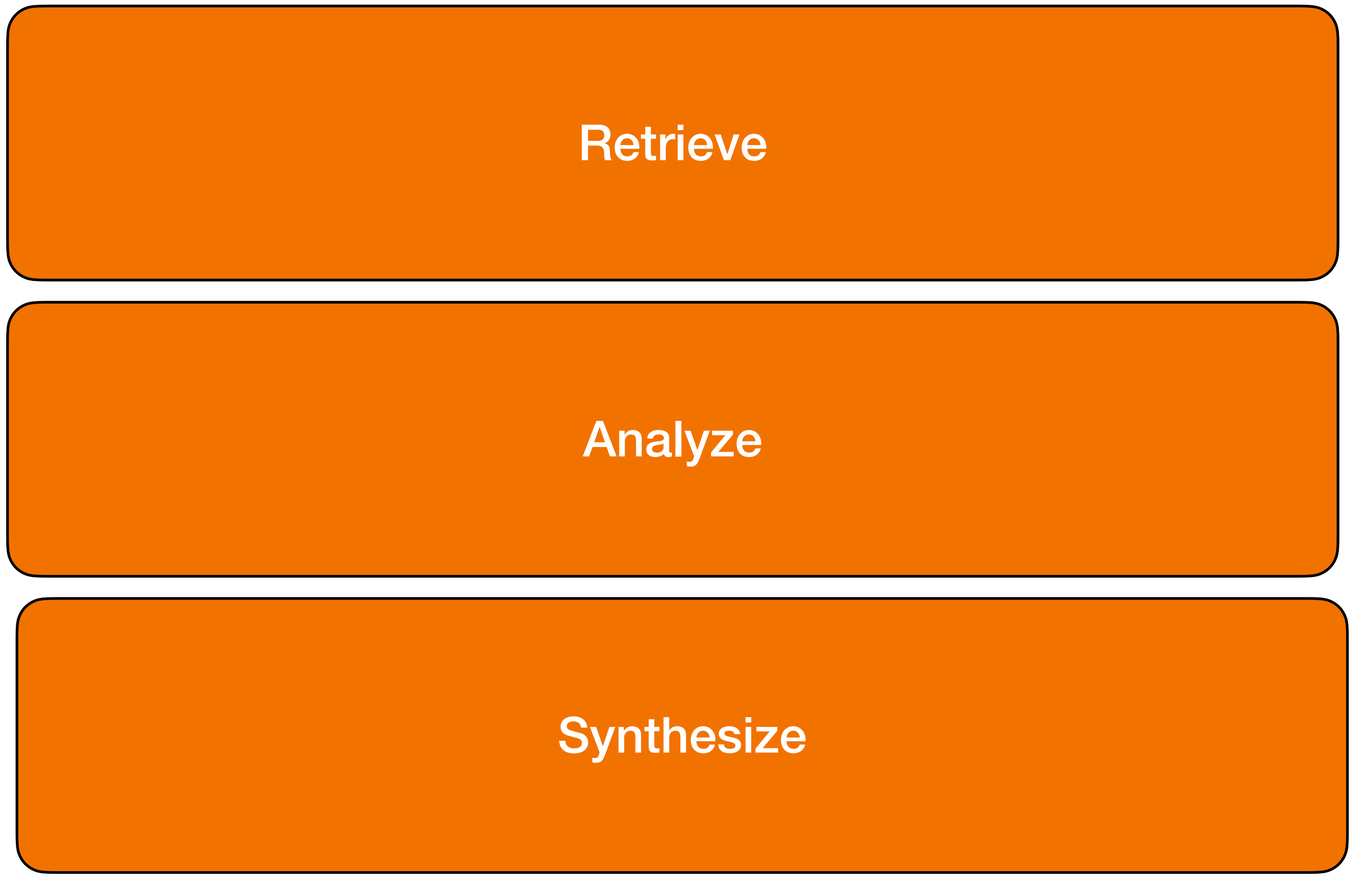
The background of the searcher strongly impacts relevance, in terms of **both pertinence, and interpretability**



SSDA 2026
Summer School on Document Analysis

Information Task Framework

Overview



Retrieve - Analyze - Synthesize

Adapted from **sense-making,** i.e., organizing information on a large topic (e.g., for a term paper on IDF and its applications)

Often, when we find or create information we **create information sources of our own** (e.g., papers, post-its)

Information Task Framework

Overview

Retrieve Information

R1 **Query** to request sources of information
R2 **Consult** and interpret available sources, *examining* and *navigating* within and across sources

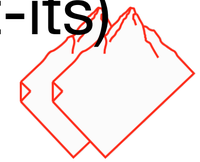
Analyze

Synthesize

Retrieve - Analyze - Synthesize

Adapted from **sense-making,** i.e., organizing information on a large topic (e.g., for a term paper on IDF and its applications)

Often, when we find or create information we **create information sources of our own** (e.g., papers, post-its)



SSDA 2026
Summer School on Document Analysis

Information Task Framework

Overview

Retrieve Information

- R1 **Query** to request sources of information
- R2 **Consult** and interpret available sources, *examining* and *navigating* within and across sources

Analyze Information

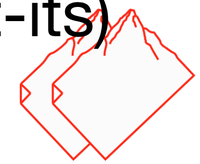
- A1 **Annotate** sources with additional information, *e.g.,* notes, add formula locations
- A2 **Index** sources by organizing them for retrieval

Synthesize

Retrieve - Analyze - Synthesize

Adapted from **sense-making,** i.e., organizing information on a large topic (e.g., for a term paper on IDF and its applications)

Often, when we find or create information we **create information sources of our own** (e.g., papers, post-its)



SSDA 2026
Summer School on Document Analysis

Information Task Framework

Overview

Retrieve Information

- R1 **Query** to request sources of information
- R2 **Consult** and interpret available sources, *examining* and *navigating* within and across sources

Analyze Information

- A1 **Annotate** sources with additional information, *e.g.,* notes, add formula locations
- A2 **Index** sources by organizing them for retrieval

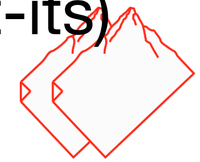
Synthesize Information

- S1 **Apply** available information that we know, have in available sources, or is encoded in algorithms, etc.
- S2 **Communicate** information by creating new sources

Retrieve - Analyze - Synthesize

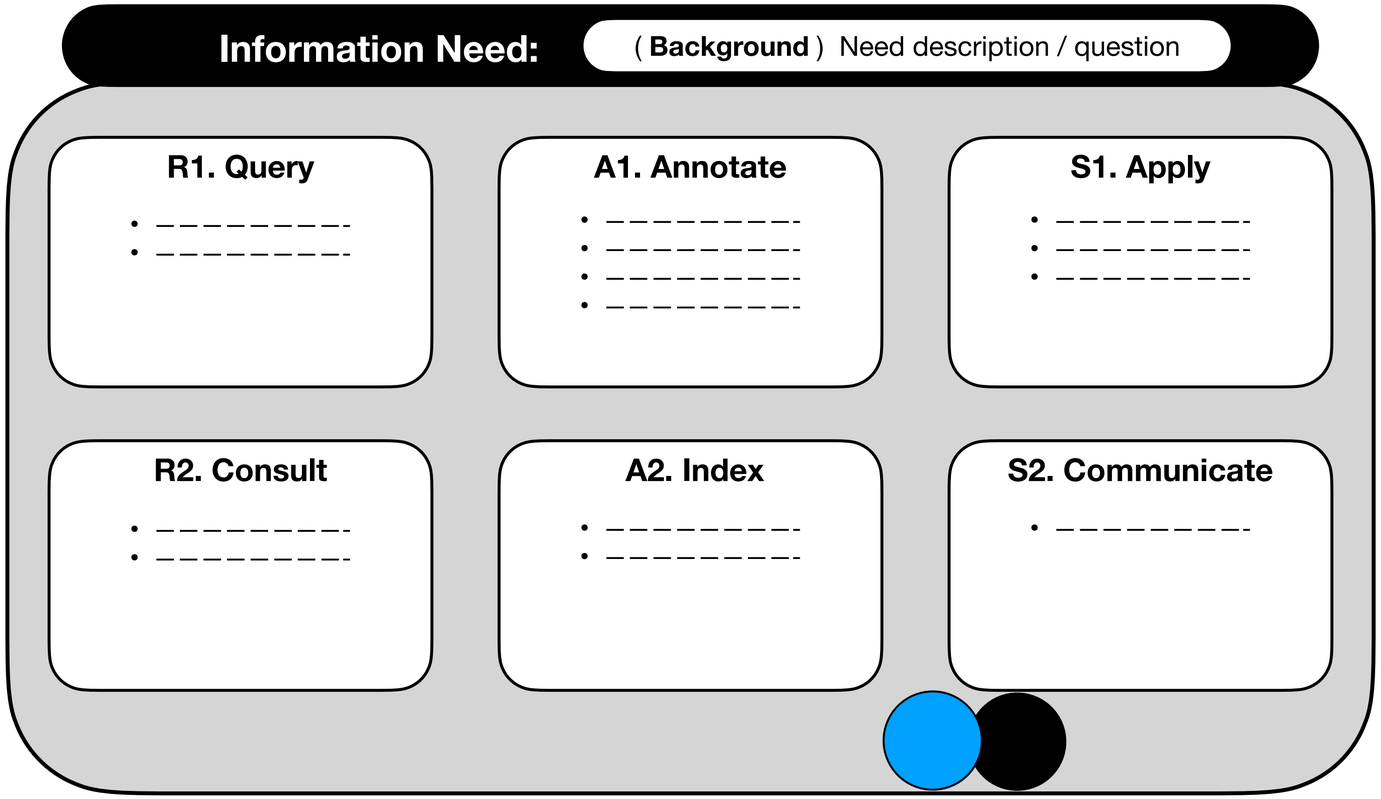
Adapted from **sense-making,** i.e., organizing information on a large topic (e.g., for a term paper on IDF and its applications)

Often, when we find or create information we **create information sources of our own** (e.g., papers, post-its)



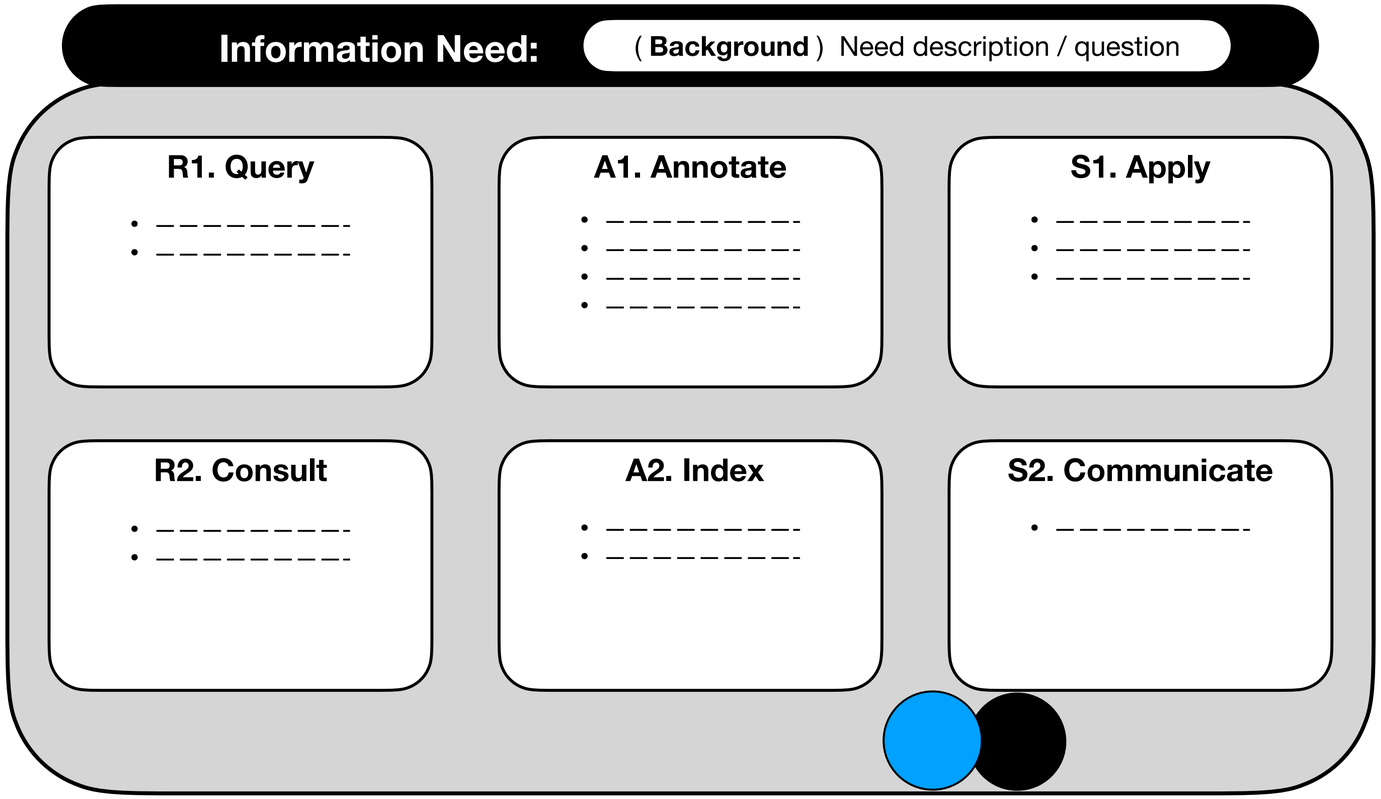
SSDA 2026
Summer School on Document Analysis

Information Need & Task Interaction: The “Task Jar”



SSDA 2026
Summer School on
Document Analysis

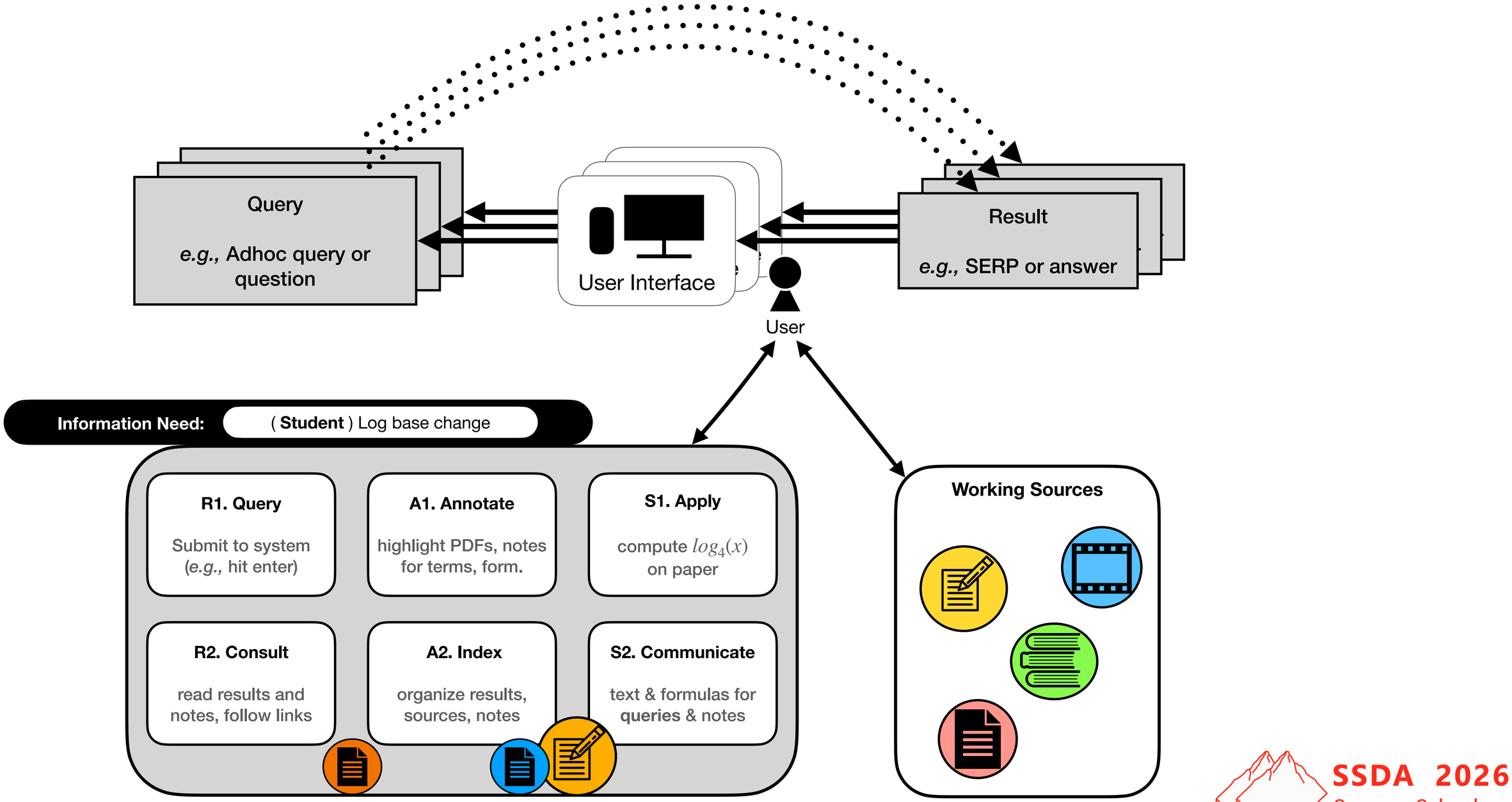
Information Need & Task Interaction: The “Task Jar”



Information sources used in the wild by people are of many different types, e.g.,

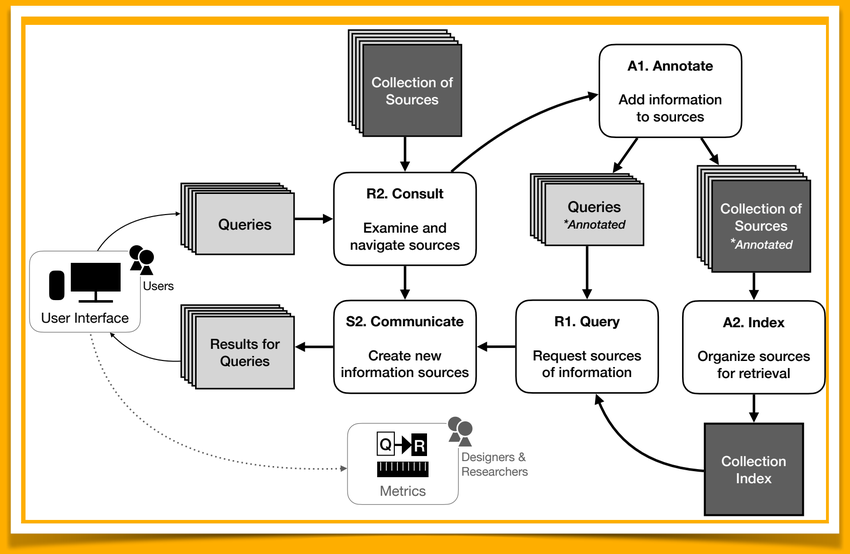
- Books, papers, Video, Audio
- **Queries**
- **Search results & LLM responses**
- Post-its (‘sticky note’)
- Conversations
- Observations
- **Formula chips & cards**
- **Generated whiteboard keyframes**

SSDA 2026
Summer School on Document Analysis



SSDA 2026
Summer School on Document Analysis

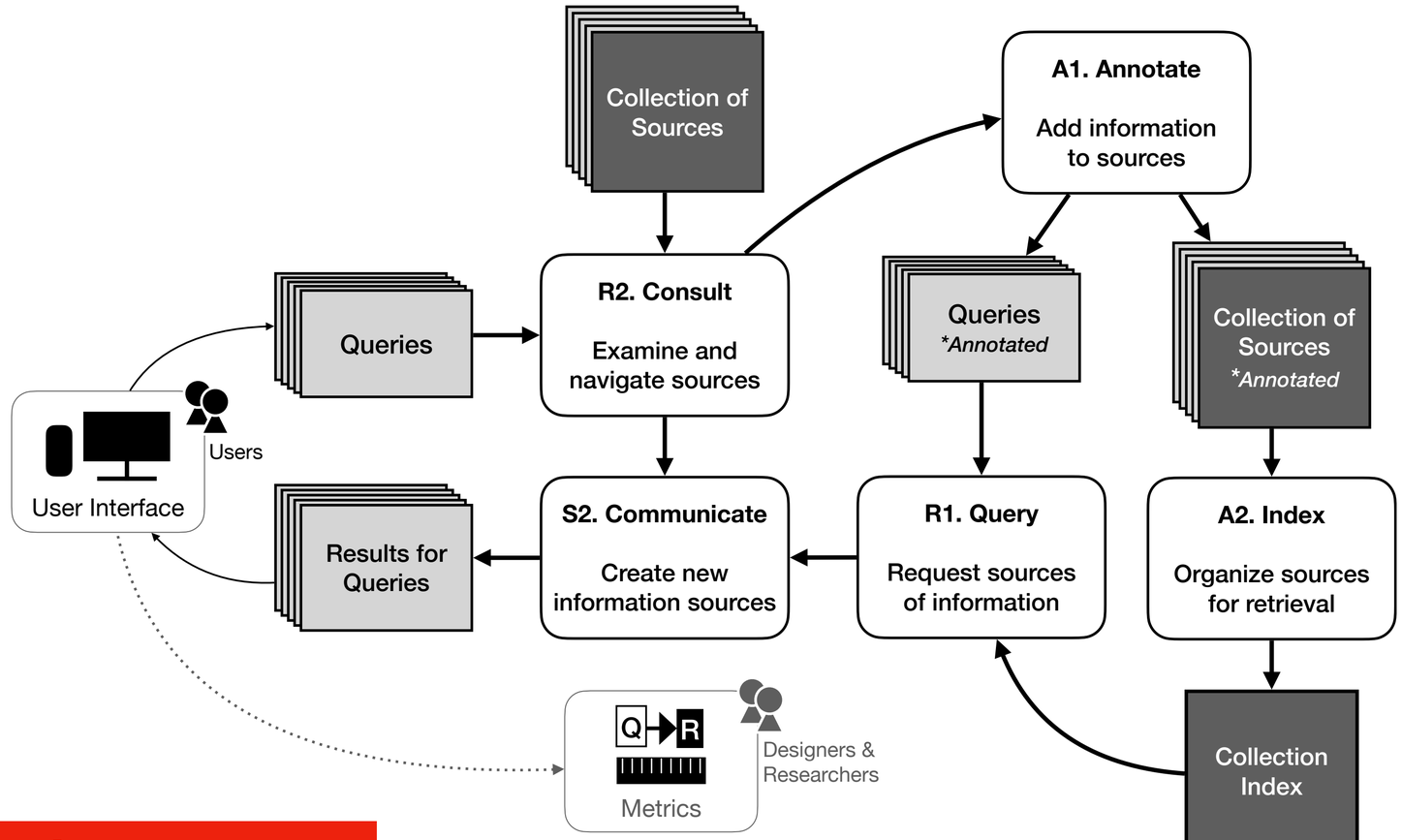
Related Information Task Model for Systems



dprl
Document and Pattern Recognition Lab

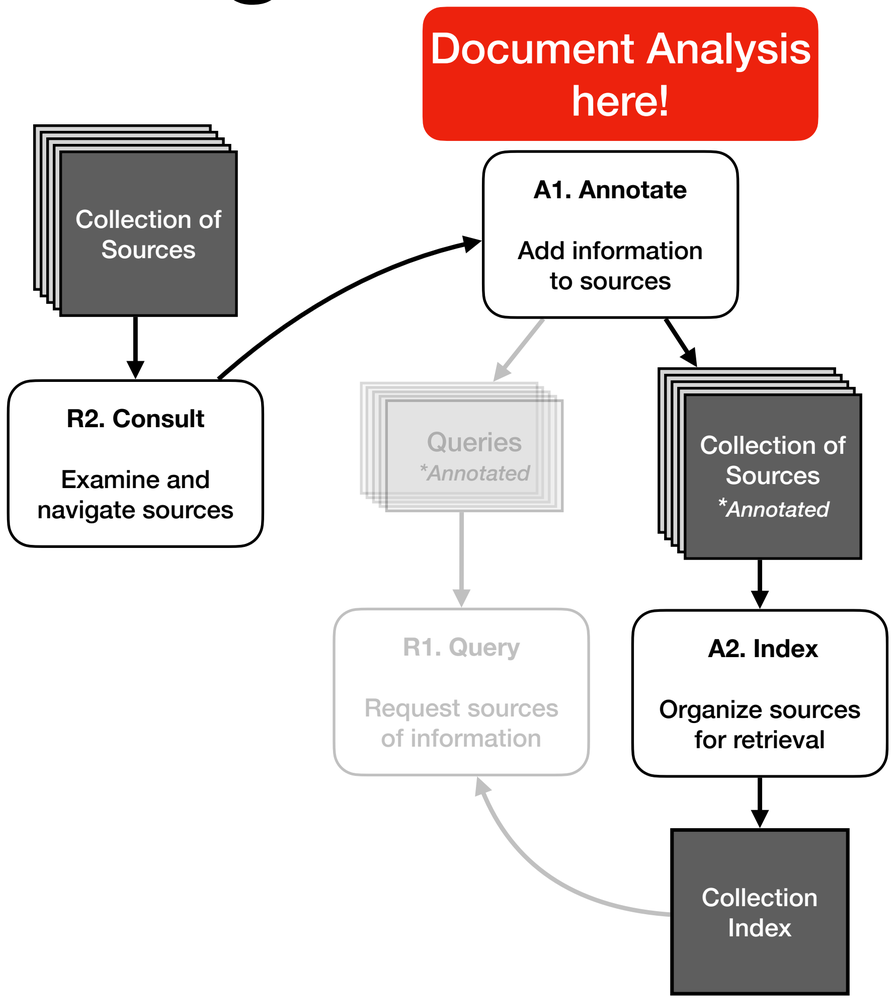
SSDA 2026
Summer School on Document Analysis

System Retrieval Model



Note that the task types are same as for human searchers

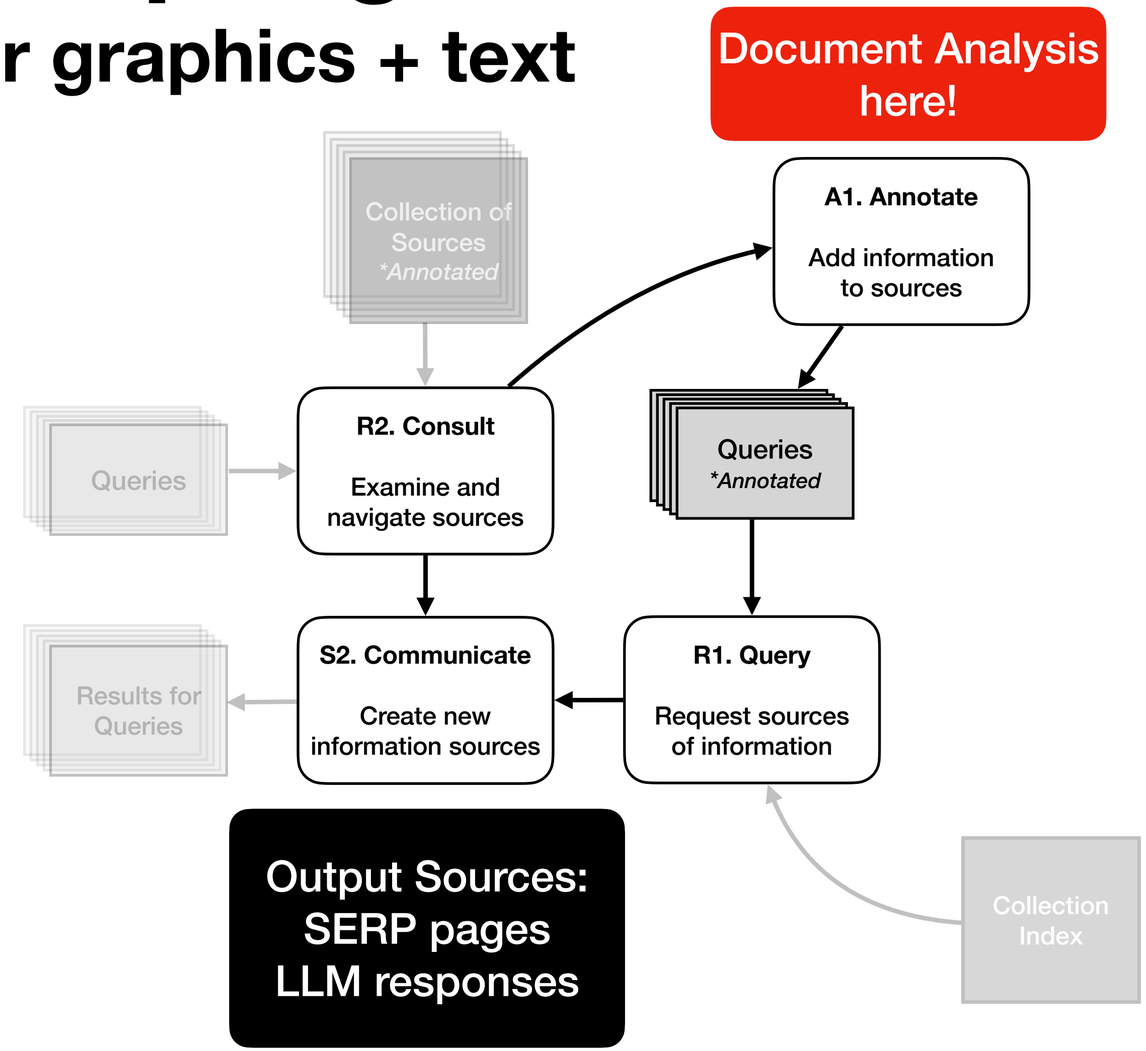
Annotation and Indexing



SSDA 2026
Summer School on Document Analysis

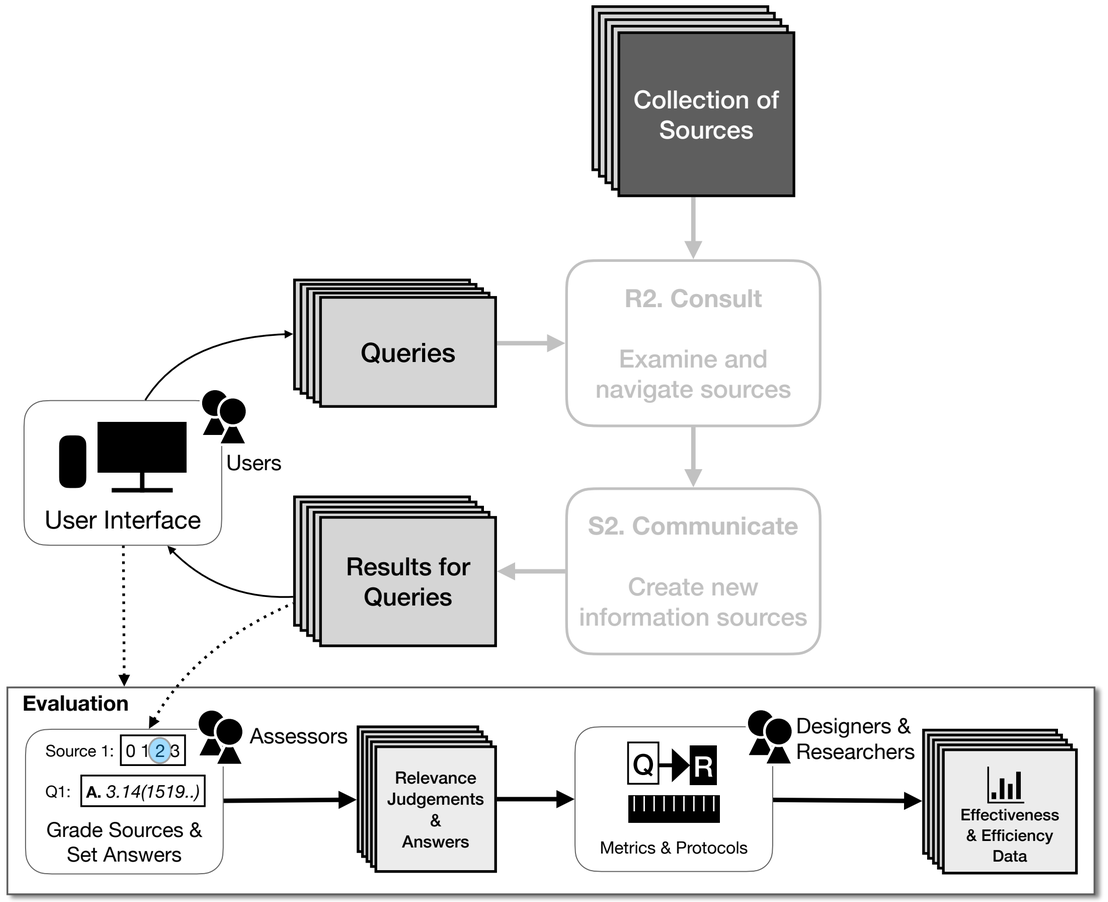
Retrieval / Prompting

e.g., formula, text, or graphics + text

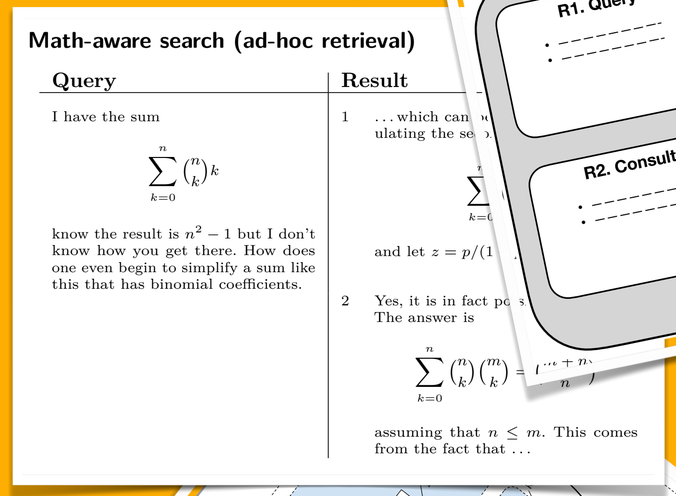


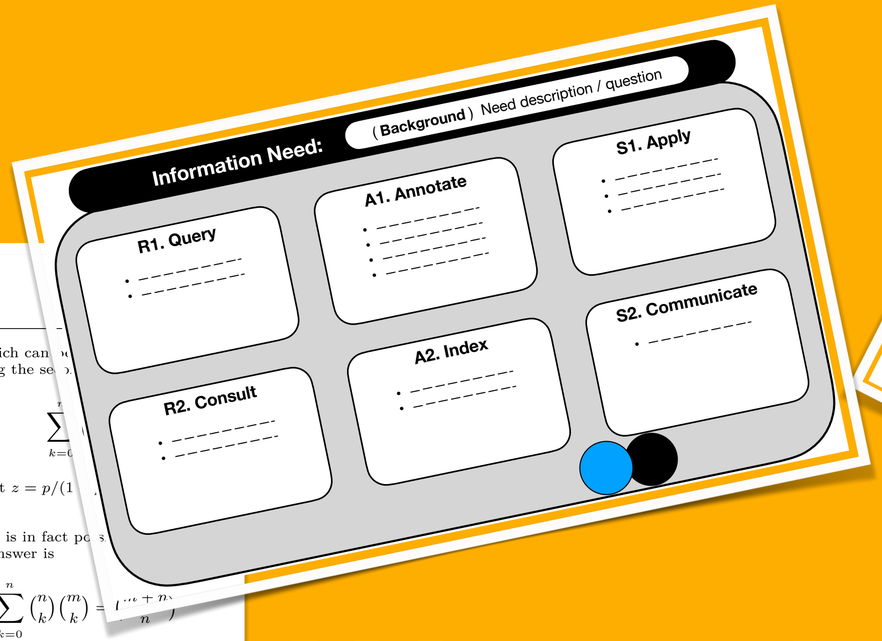
SSDA 2026
Summer School on Document Analysis

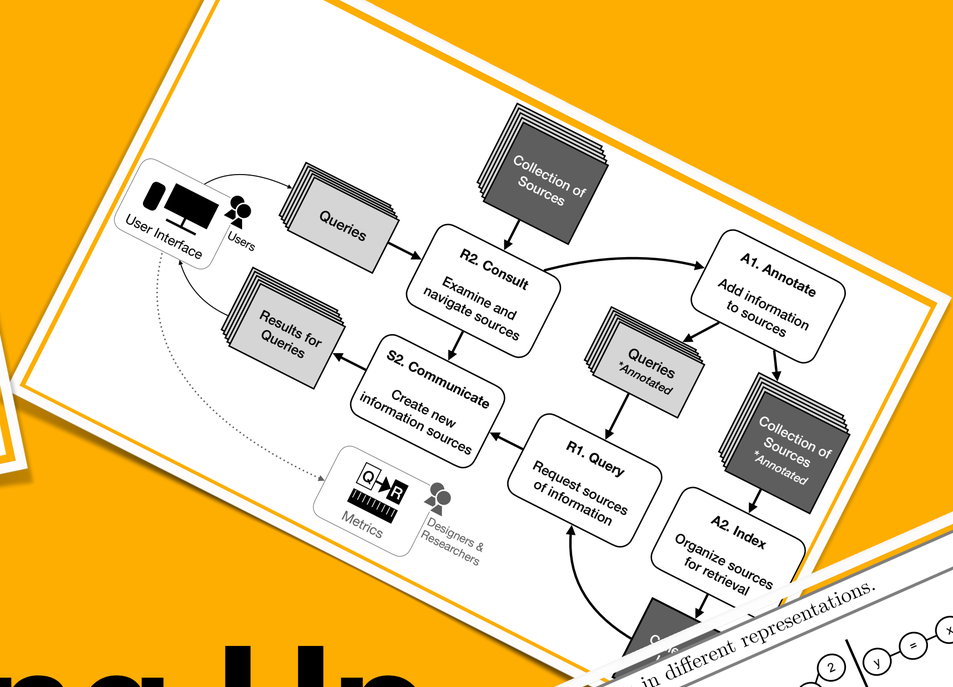
Evaluation

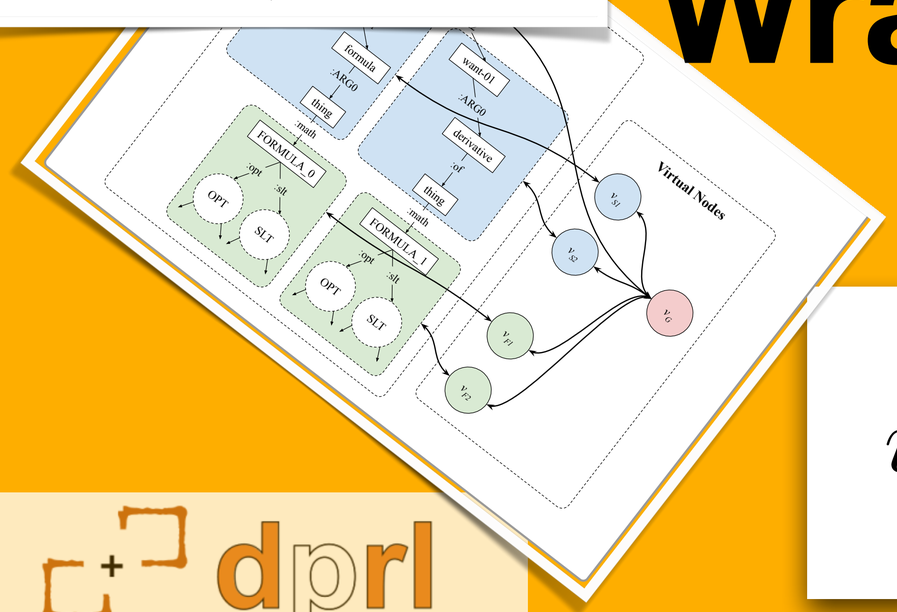


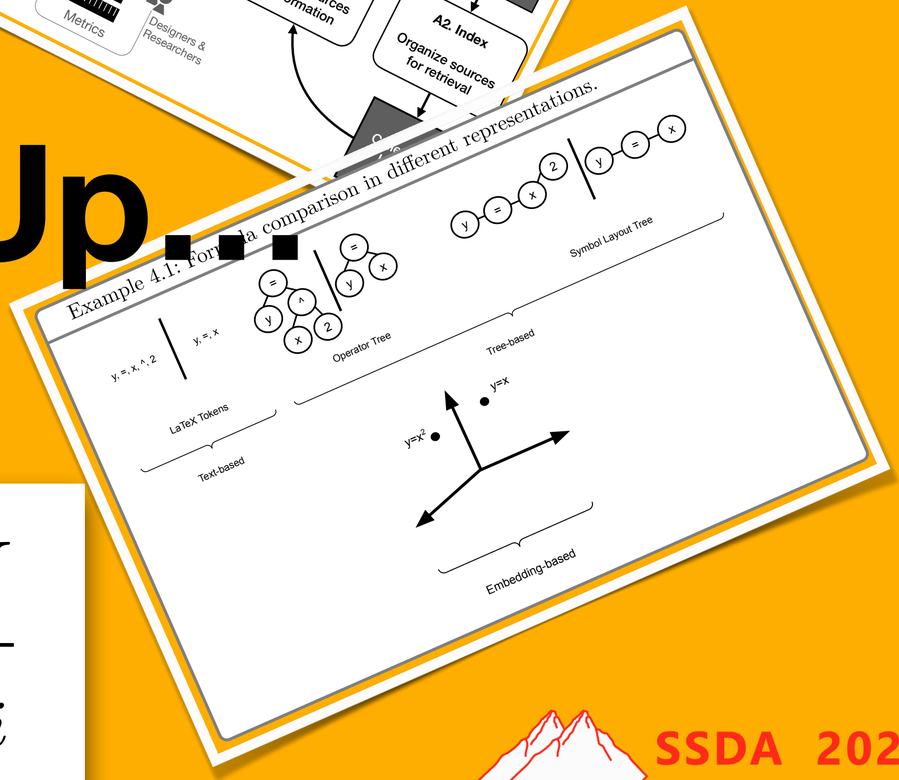
Wrapping Up











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

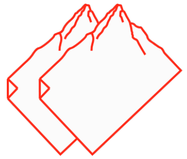
dprl
Document and Pattern Recognition Lab

SSDA 2026
Summer School on Document Analysis

What We've Discussed Today

…however briefly.

- **Interpreting graphics requires context:** from within a document **and** from the reader's context (e.g., background knowledge). *Much is assumed.*
- **Graphics representations** impact both the types of information, and patterns in information that can be readily found in search
- **Well-designed information systems** (LLM, search, other) align well with human information needs & tasks, and provide affordances for them
- **'Task jar'** comprised of six task can characterize and relate information processing by people; the same tasks are implemented in systems
- A key role for **document analysis** in search is **annotating** sources and queries with additional information for use in search



SSDA 2026
Summer School on
Document Analysis

Thank you

RIT Rochester Institute of Technology

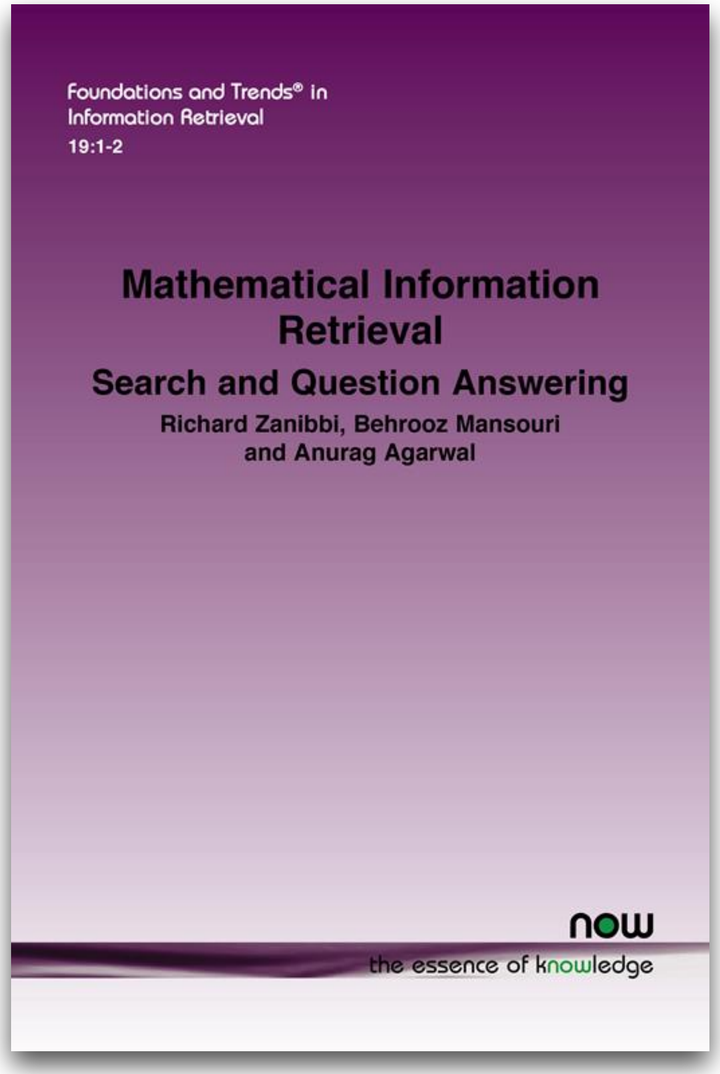
dprl Document and Pattern Recognition Lab

Big thanks to Josep, Dimos, Adria, Artemis, the CVC, and everyone else at this great meeting.

SSDA 2026
Summer School on Document Analysis

Mathematical Information Retrieval:
Search and Question Answering

Richard Zanibbi, Behrooz Mansouri, and Anurag Agarwal



Foundations and Trends allowed the final PDF to be released for free! Link to PDF:

SSDA 2026
Summer School on
Document Analysis