The NIST Digital Repository of Mathematical Formulae and Scalable Math Search

Howard Cohl*, Moritz Schubotz[§], Marje McClain*, Bonita Saunders*, Cherry Zou[§§], Azeem Mohammed[§§], Alex Danoff[†], Jimmy Li[‡], *Shraeya Madhu*[§§], *Claude Zou*[§§], *Akash Shah, Yusuf Ameri*

* Applied and Computational Mathematics Division, NIST, Gaithersburg, Maryland, U.S.A.

[§] Database Systems and Information Management Group, Technische Universität Berlin, Germany

[§§] Poolesville High School, Poolesville, Maryland, U.S.A.

[†] Thomas S. Wootton High School, Rockville, Maryland, U.S.A.

[‡] Richard Montgomery High School, Rockville, Maryland, U.S.A.

APPLIED & COMPUTATIONAL MATHEMATICS DIVISION SEMINAR SERIES

2015-07-24

Technische Universität Berlin

Table of contents

Technische Universität Berlin

Collecting information on special functions at NIST

- 1964: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables (AMS 55)
  - Milton Abramowitz and Irene Stegun (A&S), editors
  - 1064 Pages (book), most highly cited NIST (NBS) publication
  - definitions, identities, approximations, plots and numerical tables
- 2010: Digital Library of Mathematical Functions
  - NIST Handbook of Mathematical Functions as successor of A&S
  - F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, editors.
  - 2012 US Department of Commerce Gold Medal (Bruce Miller, Bonita Saunders, Marje McClain, Abdou Youssef, Brian Antonishek)
  - 968 Pages (printed version), HTML version
  - Links, Math Search, Info boxes, interactive 2D and 3D graphics
- 2013: Digital Repository of Mathematical Formulae

Technische Universität Berlin

NIST Digital Repository of Mathematical Formulae Goals

DRMF compendium of individual formulae for a mathematically literate audience

1. OPSF **community interaction** for mathematicians and scientists
2. **expandability** from the literature (support the DLMF)
3. context-free full **semantic** information
4. **user friendly** perspective
5. **searchable** mathematics
6. modern **MathML** tools

Technische Universität Berlin

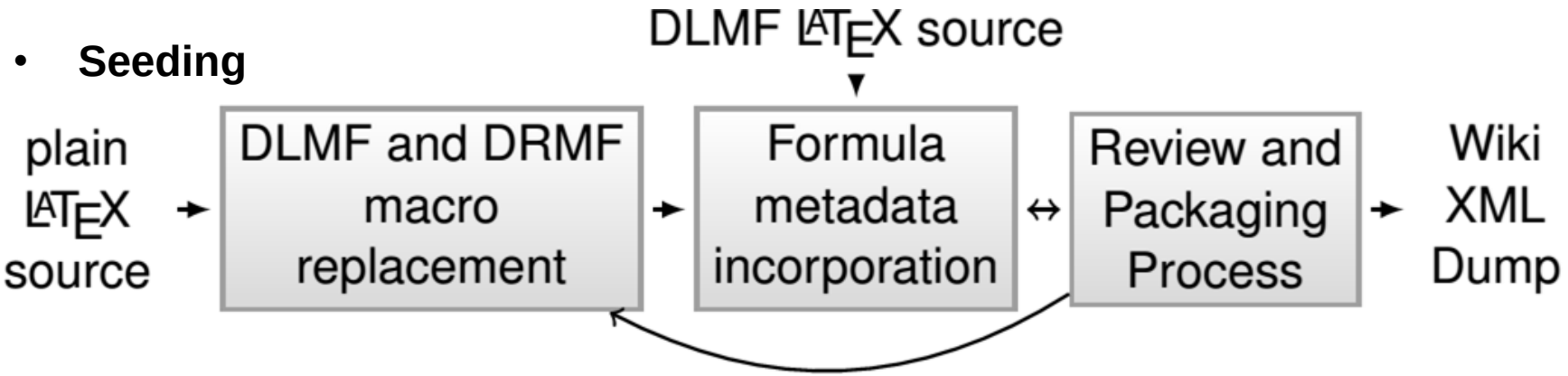
First three DRMF seeding projects

- **2013** 167 DLMF Formulae Home Pages
  - 533 semantic DLMF Macros (developed by Bruce Miller)
  - Content MathML with LaTeXML's implicit Content Dictionaries
  - estimated effort 10 min / Formula Home Page
- **2014** 1469 additional KLS (OP) Formulae Home Pages
  - Additional 153 semantic DRMF macros
  - estimated effort 5 min / Formula Home Page
- **2015** ~5000 eCF, BMP Formula Home Pages
  - Non LaTeX input
  - desired effort 1 min / Formula Home Page

Technische Universität Berlin

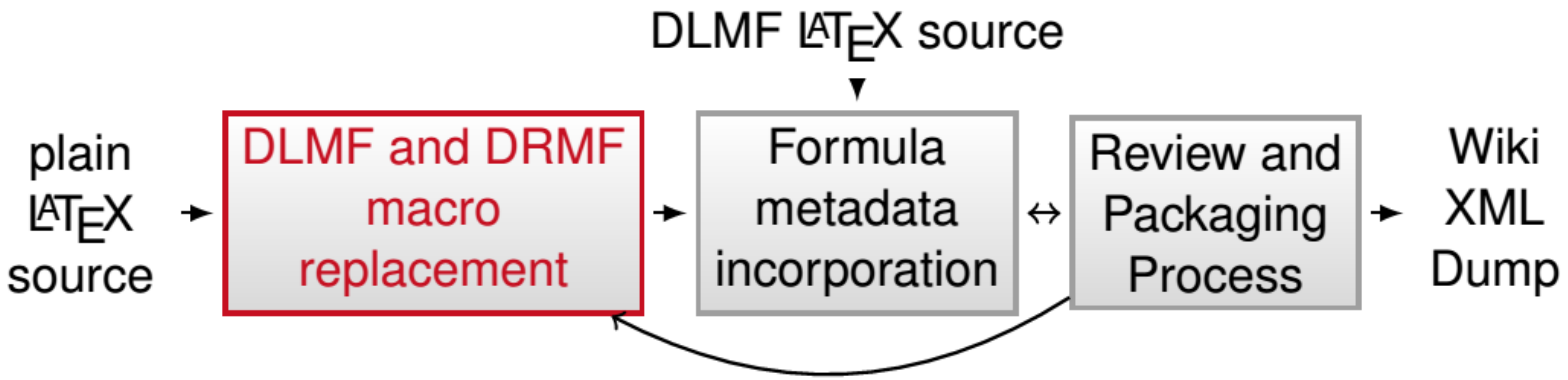
Current DRMF Implementation

- **Platform**: MediaWiki with Math and MathSearch extensions
  - 2 Table of Contents Pages
  - 76 Lists of Formulas Pages
  - 124 Definition Pages for DRMF macros
  - 1636 Formula Home Pages
- **Seeding**



Technische Universität Berlin

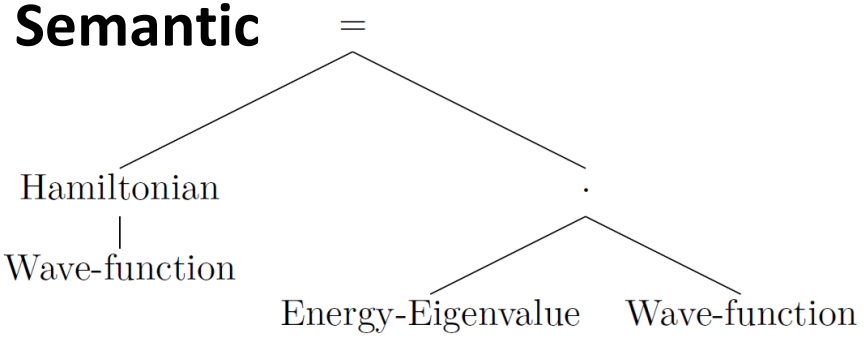
Macro replacement



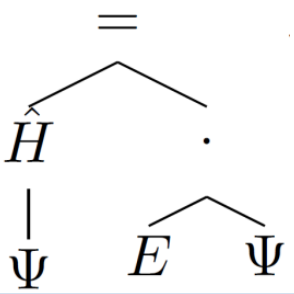
Technische Universität Berlin

Levels of Abstraction for Mathematical Formula

Semantic



Content



Presentation

$$\hat{H}\Psi = E\Psi$$

Technische Universität Berlin

(Bruce Miller) LaTeXML (DLMF) Semantic LaTeX macros

Name	Type	Rendering	LaTeX	semantic LaTeX
Trigonometric sine	f	$\sin z$	`\sin z`	`\sin@@{z}`
Euler gamma	f	$\Gamma(z)$	`\Gamma(z)`	`\EulerGamma@{z}`
Jacobi polynomial	p	$P_n^{(\alpha,\beta)}(x)$	`P_n^{(\alpha, \beta)}(x)`	`\Jacobi{\alpha}{\beta}{n}@{x}`
little q-Laguerre polynomial	p	$p_n(x;a\|q)$	`p_n(x;a|q)`	`\littleqLaguerre{n}@{x}{a}{q}`

Technische Universität Berlin

Semantic macro **breakdown**

- **689 semantic LaTeX macros (unpublished)**
  - 533 macros from DLMF
  - 156 additional DRMF macros using the LaTeXML framework
- **Semantic LaTeX macro properties**:
  - lengths between 1 and 26 characters (median length is 8 characters)
  - Strict Naming conventions
    - individual's names capitalized
    - abbreviations utilized
    - macro names correspond with object names
    - Roman numerals

Technische Universität Berlin

Semantic LaTeX Macro Glossary – csv format

For each macro we store:

- Example Macro calling sequence
- Name of object described by macro
- Object description
- Brief summary and description of calling options
- Link to URL giving precise definition

Glossary.csv used in generation of symbols lists within Wikitext and for statistical purposes.

Technische Universität Berlin

Macro Replacements for Generic LaTeX Source Datasets

For the 3 chapters of KLS (OP) as well as the KLSadd LaTeX source

- **89** semantic macros were replaced
- a total of **3308** times
- represented by **774** (LOC) of regular expressions written by Cherry Zou

Currently the six most common replacements are:

1. *q*-Pochhammer symbol – replaced **659** times
2. Euler gamma function – replaced **266** times
3. *q*-hypergeometric function – replaced **237** times
4. Pochhammer symbol – replaced **205** times
5. Racah polynomial – replaced **117** times
6. cosine function – replaced **82** times

Technische Universität Berlin

Example of complexity of the problem – KLSadd dataset

- After processing of the LaTeX input, only formulas remain.
- Current metadata: are constraint and substitution annotations.
- Future metadata
  - bibliographic metadata
  - references to KLS formulae
  - errata information
  - formula comments and notes
  - symmetries in parameters
  - proofs, etc...

Technische Universität Berlin

Ongoing NIST Student Summer Projects

- **Yusuf Ameri** - DRMF Standard MediaWiki Text Search Integration
- **Akash Shah** - DRMF Formula Search MediaWiki Frontend
- **Shraeya Madhu** - DRMF DLMF Seeding Project
- **Claude Zou** - DRMF Semantic LaTeX Mathematical Operators
- **Jimmy Li** - DRMF Formula Search Backend & NTCIR-12 Competition
- **Moritz Schubotz** - What is "good" content MathML?

Technische Universität Berlin

SCALABLE MATH SEARCH

Motivation for MathSearch

Proposition 2 (expectation value of waiting time times tunnel rate) *For every PDF* $f(x)$ *in means of definition* (0.1) *the inequality*

$$\langle x \rangle \left\langle \frac{1}{x} \right\rangle \geq 1 \tag{0.10}$$

is valid.

26th of February 2011

I. POSITIVITY OF FANO FACTOR PARAMETERS

For every convex function $f(x)$, we have according to the Jenson inequality

$$f(\langle x \rangle) \leq \langle f(x) \rangle \tag{1}$$

That means that

$$\langle x \rangle^{-k} \leq \langle x^{-k} \rangle. \tag{2}$$

Especially $k = 1$ leads to the fact that $\alpha \geq 0$. 28th of March

Technische Universität Berlin

Example $\frac{1}{\langle x \rangle} \leq \left\langle \frac{1}{x} \right\rangle$

1. Different forms e.g.

   $\langle x \rangle \left\langle \frac{1}{x} \right\rangle \geq 1$

2. Different notations e.g.

   $\int_X f(x) x dx = \langle x \rangle$

3. Exact match seldom
4. Ambiguity in syntax e.g.

   $E\Psi = \hat{H}\Psi$

5. No TeX-function mean

```
$\frac 1 \mean ?x \le
\mean \frac 1 ?x$
<apply>
  <leq/>
  <apply>
    <divide/>
    <cn type="integer">1</cn>
    <apply>
      <mean/>
      <qvar>x</qvar></apply></apply>
  <apply>
    <mean/>
      <apply>
        <divide>
        <cn type="integer">1</cn>
       <qvar>x</qvar></apply></apply>…
```

NTCIR-11 Math-2 WMC-D1

Technische Universität Berlin

Example $\frac{1}{\langle x \rangle} \leq \left\langle \frac{1}{x} \right\rangle$

1. Different forms e.g.

   $\langle x \rangle \left\langle \frac{1}{x} \right\rangle \geq 1$

2. Different notations e.g.

   $\int_X f(x) x dx = \langle x \rangle$

3. Exact match seldom
4. ~~Ambiguity in syntax e.g.~~

   ~~$E\Psi = \hat{H}\Psi$~~

5. ~~No TeX-function mean~~


```
$\frac 1 \mean ?x \le
\mean \frac 1 ?x$
```

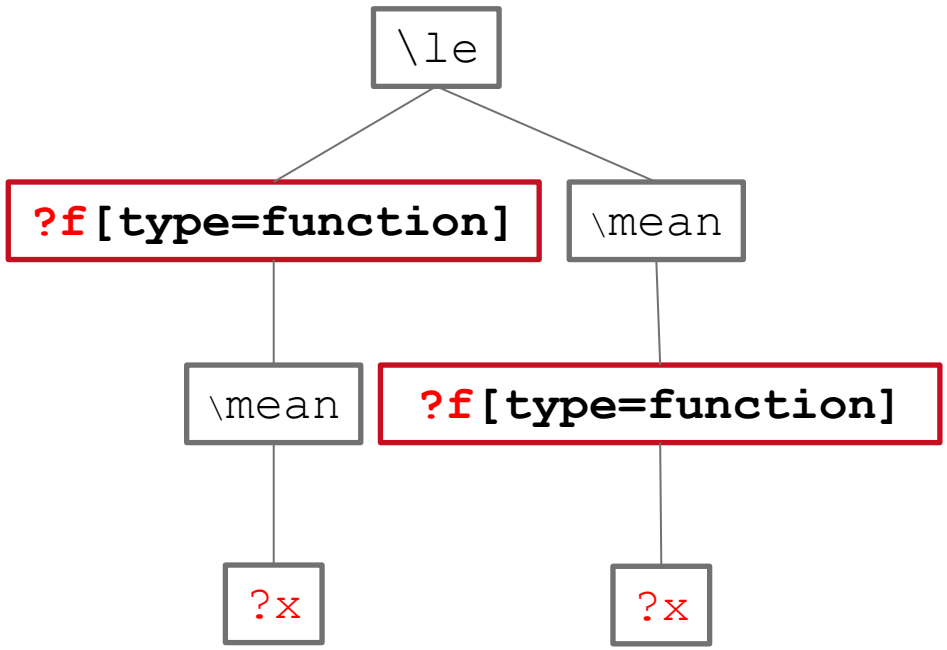
```
<apply>
  <leq/>
  <apply>
    <divide/>
    <cn type="integer">1</cn>
    <apply>
      <mean/>
      <qvar>x</qvar></apply></apply>
  <apply>
    <mean/>
      <apply>
        <divide>
        <cn type="integer">1</cn>
       <qvar>x</qvar></apply></apply>…
```

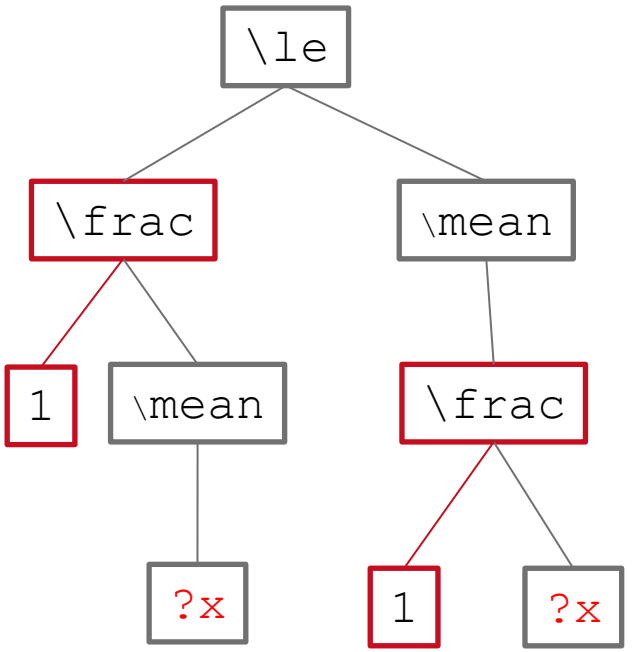
NTCIR-11 Math-2 WMC-D1

Result 1: $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$

```
$?f[type=function] \mean ?x \le
\mean ?f[type=function] ?x $
```

```
$\frac 1 \mean ?x \le
\mean \frac 1 ?x$
```





Result 1: $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$

Technische Universität Berlin

```
$?f[type=function] \mean ?x \le
\mean ?f[type=function] ?x $
```

```
<apply>
  <leq/>
  <apply>
    <qvar type="functi
    <apply>
      <mean/>
      <qvar>x</qvar></
  <apply>
    <mean/>
      <apply>
        <qvar type="function">f</qvar>
        <qvar>x</qvar></apply></apply>…
```

```
$\frac 1 \mean ?x \le
\mean \frac 1 ?x$
```

```
<apply>
  <leq/>
  <apply>
    <apply>
      <mean/>
      <apply></apply>
  <apply>
    <mean/>
      <apply>
        <divide>
        <cn type="integer">1</cn>
        <qvar>x</qvar></apply></apply>…
```

$\frac 1$ -> $?f[type=function]$

Not trivial

Technische Universität Berlin

Solution inexact matches

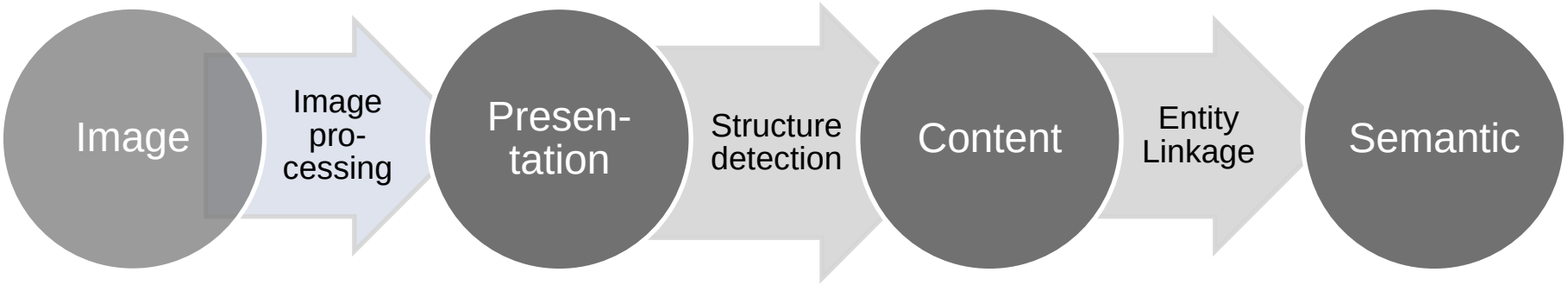
Refined query:

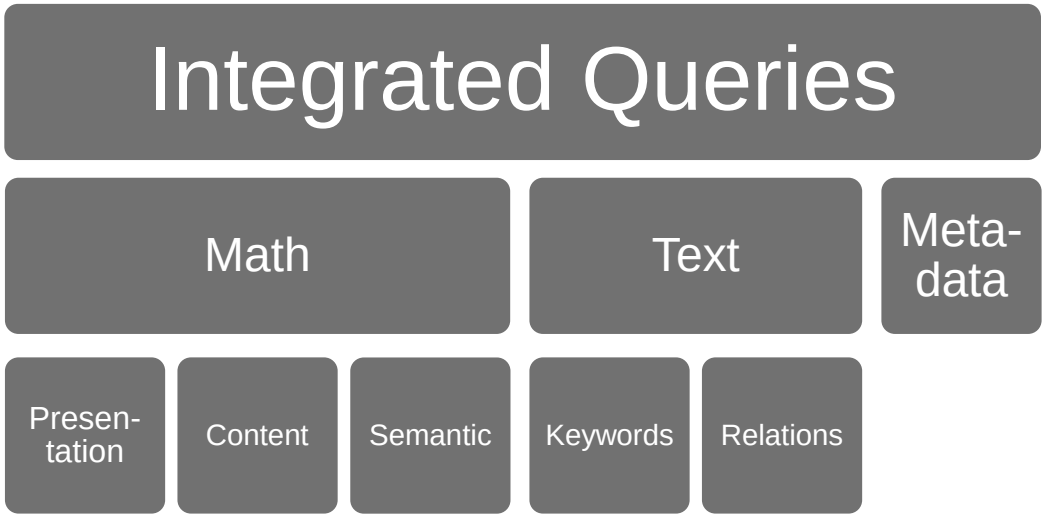
```
$\superconceptOf[
  orderby = editdistance ]{
  \frac 1 \mean ?x \le
  \mean \frac 1 ?x
}$
```

- Computational complexity
- Restriction of the search space
- Check most likely solutions at first

Technische Universität Berlin

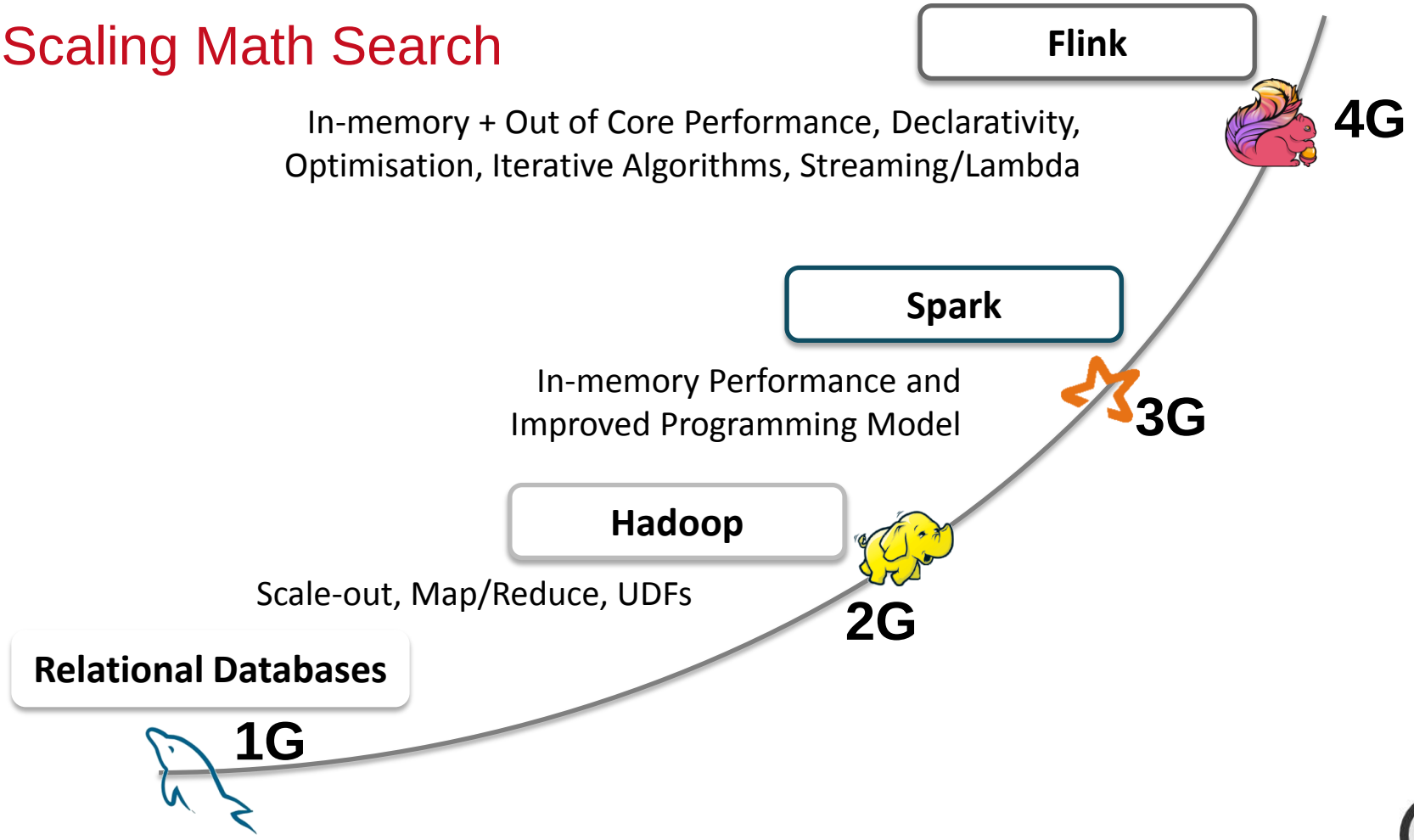
Challenges



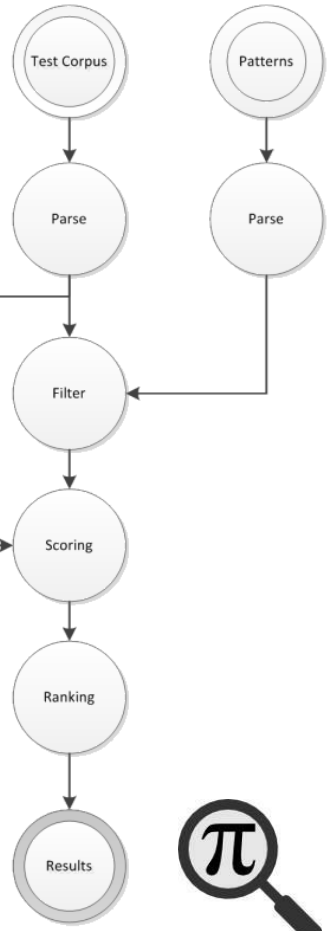


Technische Universität Berlin

Scaling Math Search



Mathosphere: Using Apache Flink for scalable Math Search

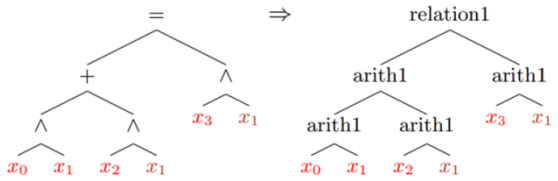


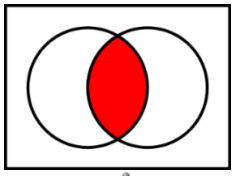
Evaluation

- CICM 2012
  (2 Participants)
- NTCIR 2013 (pilot)
  (6 Participants)
- NTCIR 2014

Evaluation of Similarity Measures Developed by Zhang & Youssef

Measure	Specificity
1 Taxonomic Distance	.91
2 Data Type	.91
3 Match Depth	(level 10) .87 … .93 (level 1)
4 Query Coverage	>0.7 (linear)
5 Formula vs. Equation	.26





Technische Universität Berlin

Conclusion and Summary

- Online Compendium of Mathematical Formulae using MediaWiki, DRMF
- Successes at seeding this compendium with Generic LaTeX sources
- Context free semantics for individual formulae
- Macro replacements
- Implementing Math Searchable archives
- NTCIR 12
- Ongoing: Different quality levels of Mathematical content
- DRMF Demo at http://drmf.wmflabs.org