Augmenting Mathematical Formulae for More Effective Querying & Presentation

Moritz Schubotz

Motivation

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

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

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

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

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~~

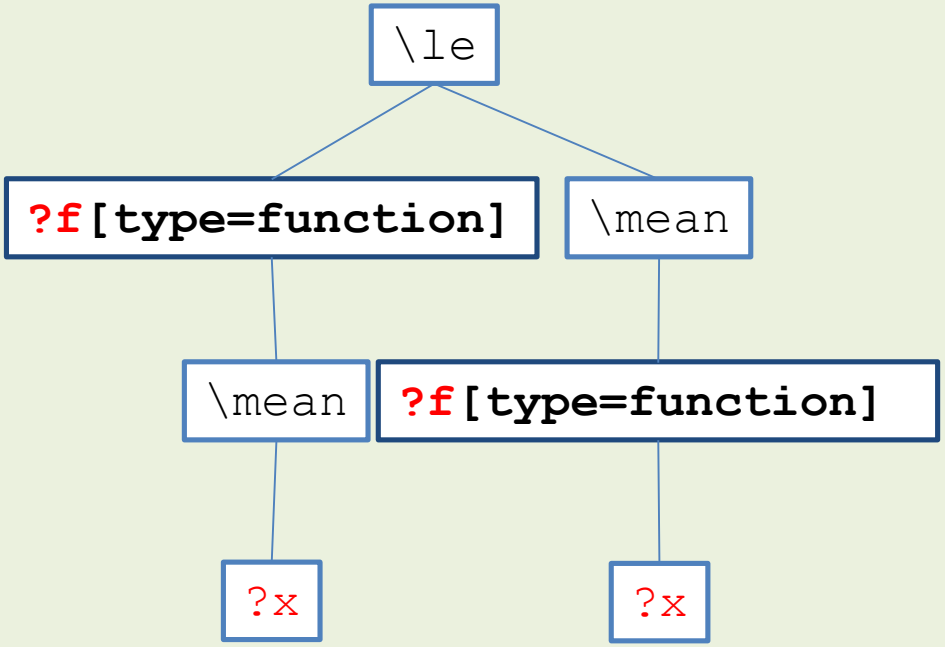
```
$\frac 1 \mean ?x \le
\mean \frac 1 ?x$
```

```
<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>…
```

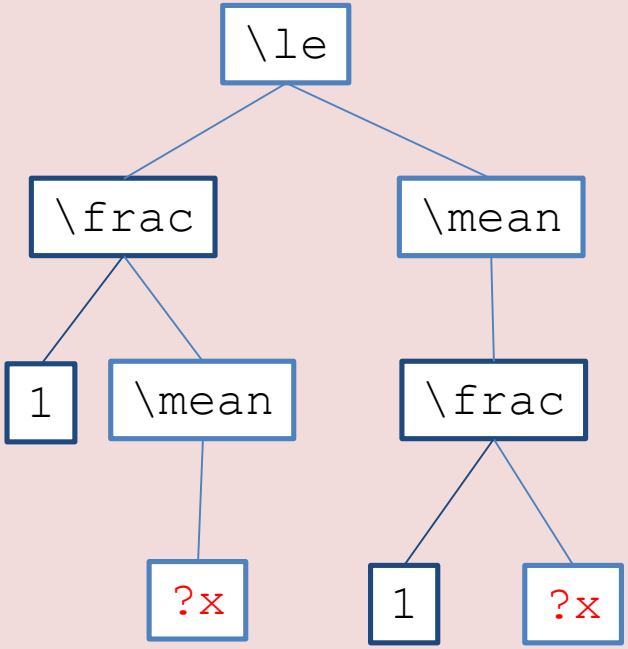
NTCIR-11
Math-2
WMC-D1

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

```
$?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>
        <divide>
        <cn type="integer">1</cn>
        <qvar>x</qvar></apply></apply>…
```

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

Not trivial

Solution 1 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

But there are diverse information needs

1. Definition look-up
2. Explanation look-up
3. Proof look-up
4. Application look-up
5. Computation assistance
6. General formula search

, and the data looks like that

https://ja.wikipedia.org/wiki/イェンゼンの不等式

ブックの新規作成
PDF 形式でダウンロード
印刷用バージョン

ツール
リンク元
関連ページの更新状況
ファイルをアップロード
特別ページ
この版への固定リンク
ページ情報
ウィキデータ項目
このページを引用

他言語版
العربية
Български
Català
Čeština
Deutsch
English
Español

$$\int_{-\infty}^{\infty} f(y(x))p(x)dx \geq f\left(\int_{-\infty}^{\infty} y(x)p(x)dx\right)$$

ルベーグ積分論の観点では、離散の場合も連続の場合も同一に見做せる。

証明は、f の $\int_{-\infty}^{\infty} y(x)p(x)dx$ における接線を g とおいて、常に $g(x)$ が $f(x)$ よりも小さいことを使えばよい。

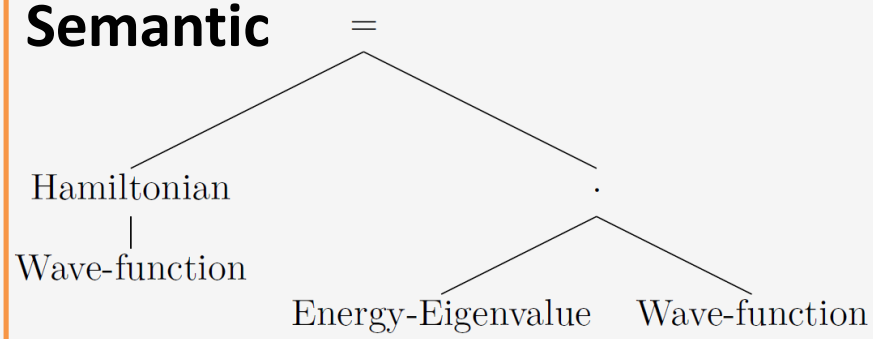
統計学において、式の下限を評価するさいに、一定の役割を担っている。例えば、カルバックライブラーダイバージェンスが常に 0 より大きいことを証明するときに用いられる。$p(x)$ が確率密度関数の場合を考えると、イェンゼンの不等式は次のように書ける。

$$E[f(y)] \geq f(E[y])$$

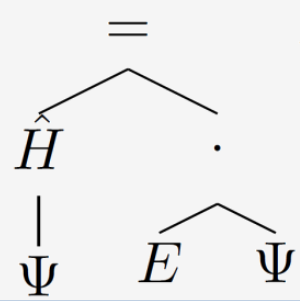
なお、イェンゼンの不等式から、相加相乗平均の不等式などを導くこともできる。

Levels of Abstraction

Semantic



Content



Presentation

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

Overview

Image → Image pro-cessing → Presen-tation → Structure detection → Content → Entity Linkage → Semantic

Integrated Queries

- Math
- Text
- Meta-data

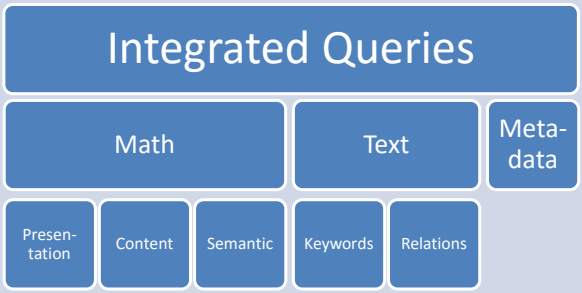
Math: Presen-tation, Content, Semantic

Text: Keywords, Relations

Completed Research

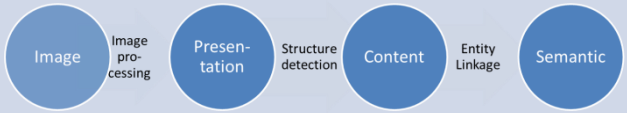
Querying

- Making Math Searchable in Wikipedia (CICM 2012)
- Evaluation of Similarity-Measure Factors for Formulae (NTCIR 2015)
- Wikipedia Subtask at NTCIR 11 (SIGIR 2015)
- Exploring the single-brain barrier (NTCIR 2016)



Processing

- Mathematical Language Processing (CICM 2014)
- Digital Repository of Mathematical Formulae (CICM 2014 coauthor)
- Growing the DRMF with generic LaTeX sources
- Mathoid: Accessible Math Rendering for Wikipedia (CICM 2014)
- Semantification of Identifiers in Mathematics for Better Math Information Retrieval (SIGIR 2016)



Scalibility

- Applying Stratosphere for Big Data Analytics (coauthor, BTW 2013)
- Querying large Collections of Mathematical Publications (NTCIR 2013 with Marcus Leich)

Mathoid: Robust, Scalable, Fast and Accessible Math Rendering for Wikipedia

Let be a probability space, X an integrable real-valued random variable and φ a convex function. Then:

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$

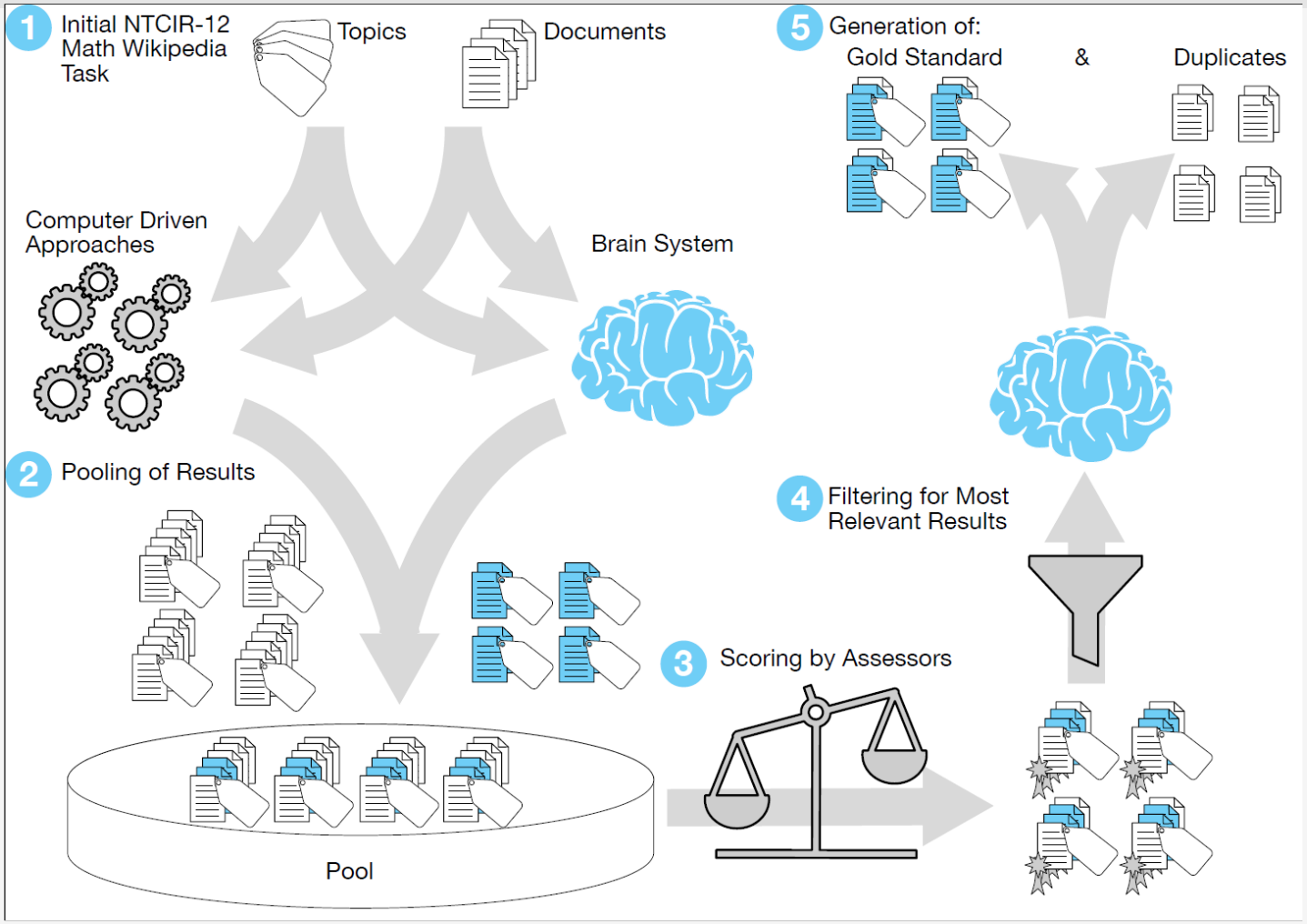
- **convex function (Q319913, NDL ID 00573442)**
- subclass of function
- ja:凸関数

Exploring the single-brain barrier

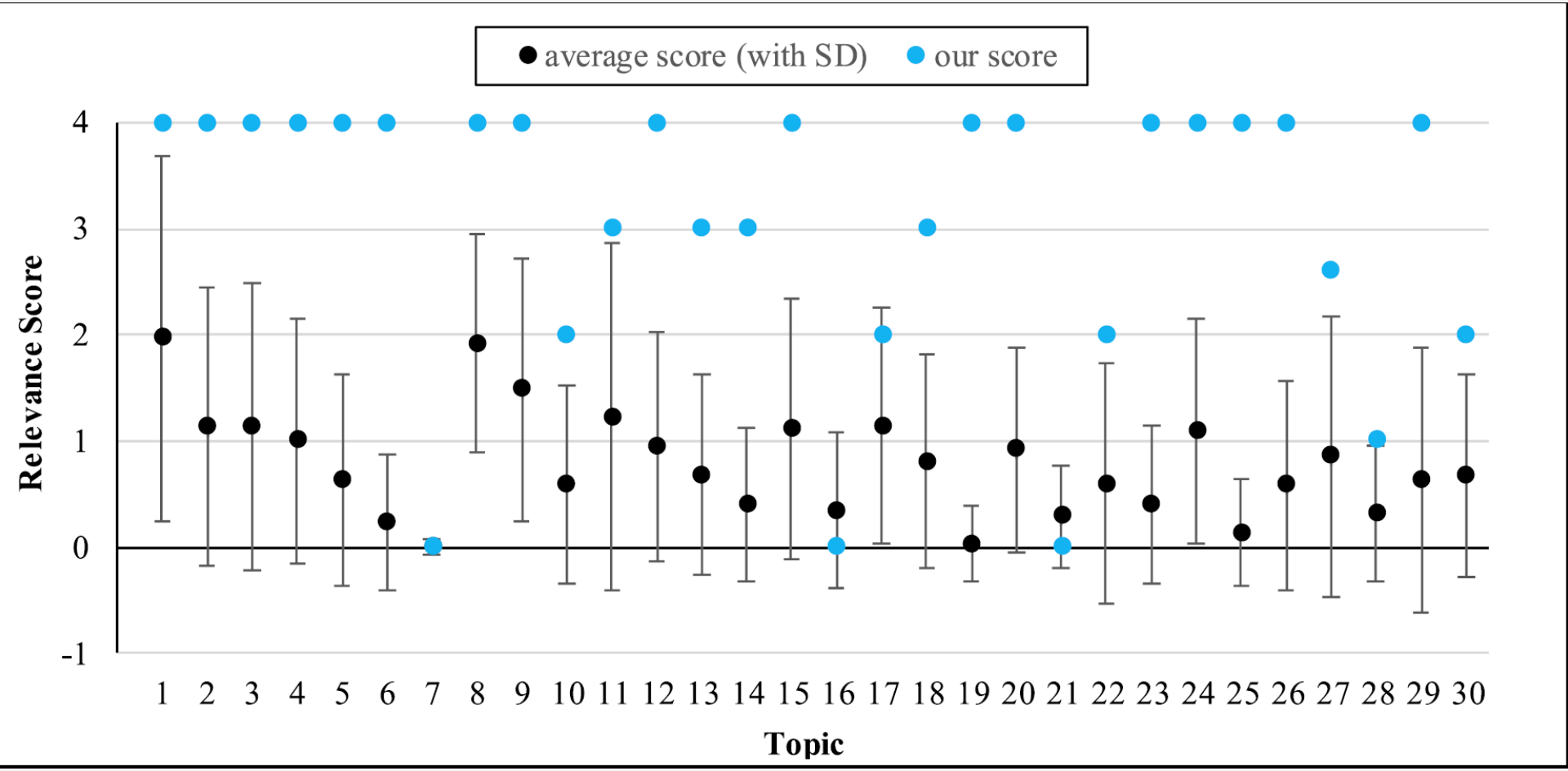
- *"one-brain barrier"* [1]
  - Metaphor: relevant knowledge to conduct math research needs to be co-located in one brain

- Goals of our contribution to NTCIR12:
  - Create a point of reference w.r.t. to this barrier for a trained mathematician
  - Compare the performance of a human to MIR systems and analyse characteristic strengths and weaknesses
  - Derive insights to improve MIR systems
  - Combine the relevant results of the human and the MIR systems to create a gold standard

Exploring the single-brain barrier



Exploring the single-brain barrier



Exploring the single-brain barrier

- Strengths of MIR systems:
  - Definition lookup queries
  - Application lookup
- Weaknesses of MIR systems
  - Low precision
  - No unified query language to specify query type
- Gold standard dataset can help to develop a math-aware search engine for Wikipedia

Semantification of Identifiers in Mathematics for Better Math Information Retrieval

- First step to enable computer to understand mathematicians notations
- Focus on identifiers
- Extract identifier semantics by combining math and
- Use computers to find relevant mathematics
- Computers must understand semantics in math to provide needed information

Math Augmentation Approach

1 Detect formulae

In physics, mass–energy **equivalence** is a concept formulated by Albert Einstein that explains the relationship between mass and energy. It states every mass has an energy equivalent and vice versa—expressed using the formula

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the

2 Extract identifiers

$$E = mc^2$$

3 Find identifiers

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the

4 Find definiens candidates

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number

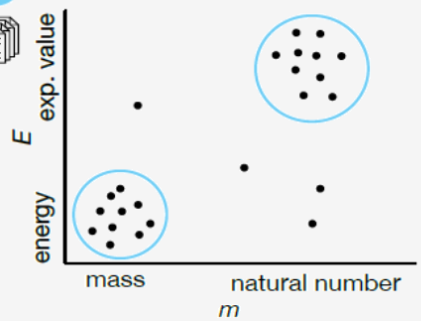
5 Score all identifier-definiens pairs

	E	m	c	m
energy	2	4	13	22
physical system	5	1	10	20
mass	10	2	5	15
speed of light	19	11	2	5
vacuum	21	14	5	3

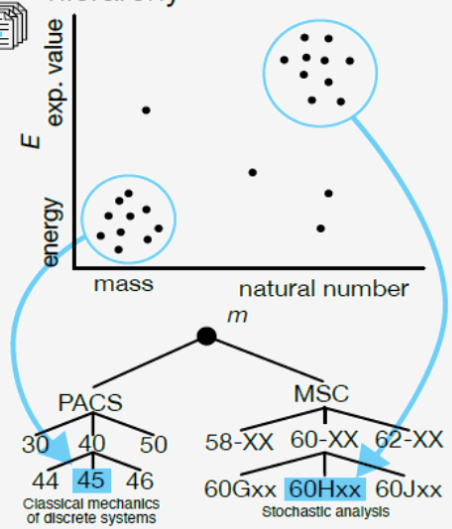
6 Generate feature vectors

$$\begin{array}{r} \\ \text{E\_energy} \\ \text{E\_expected\_value} \\ \text{m\_mass} \\ \text{m\_natural\_number} \\ \vdots \end{array} \begin{array}{c} d_1 \\ \begin{pmatrix} 2 \\ - \\ 2 \\ - \\ \vdots \end{pmatrix} \end{array} \begin{array}{c} d_2 \\ \begin{pmatrix} - \\ 2 \\ - \\ 0 \\ \vdots \end{pmatrix} \end{array} \begin{array}{c} \dots \\ \\ \dots \\ \\ \\ \end{array}$$

7 Cluster feature vectors



8 Map clusters to subject hierarchy



(1) Extract formulae

In physics, **mass–energy equivalence** is a concept formulated by Albert Einstein that explains the relationship between mass and energy. It states every mass has an energy equivalent and vice versa—expressed using the formula

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the

(2) Extract identifiers

$$E = mc^2$$

(3) Find identifiers

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the

(4) Find definiens candidates

$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number

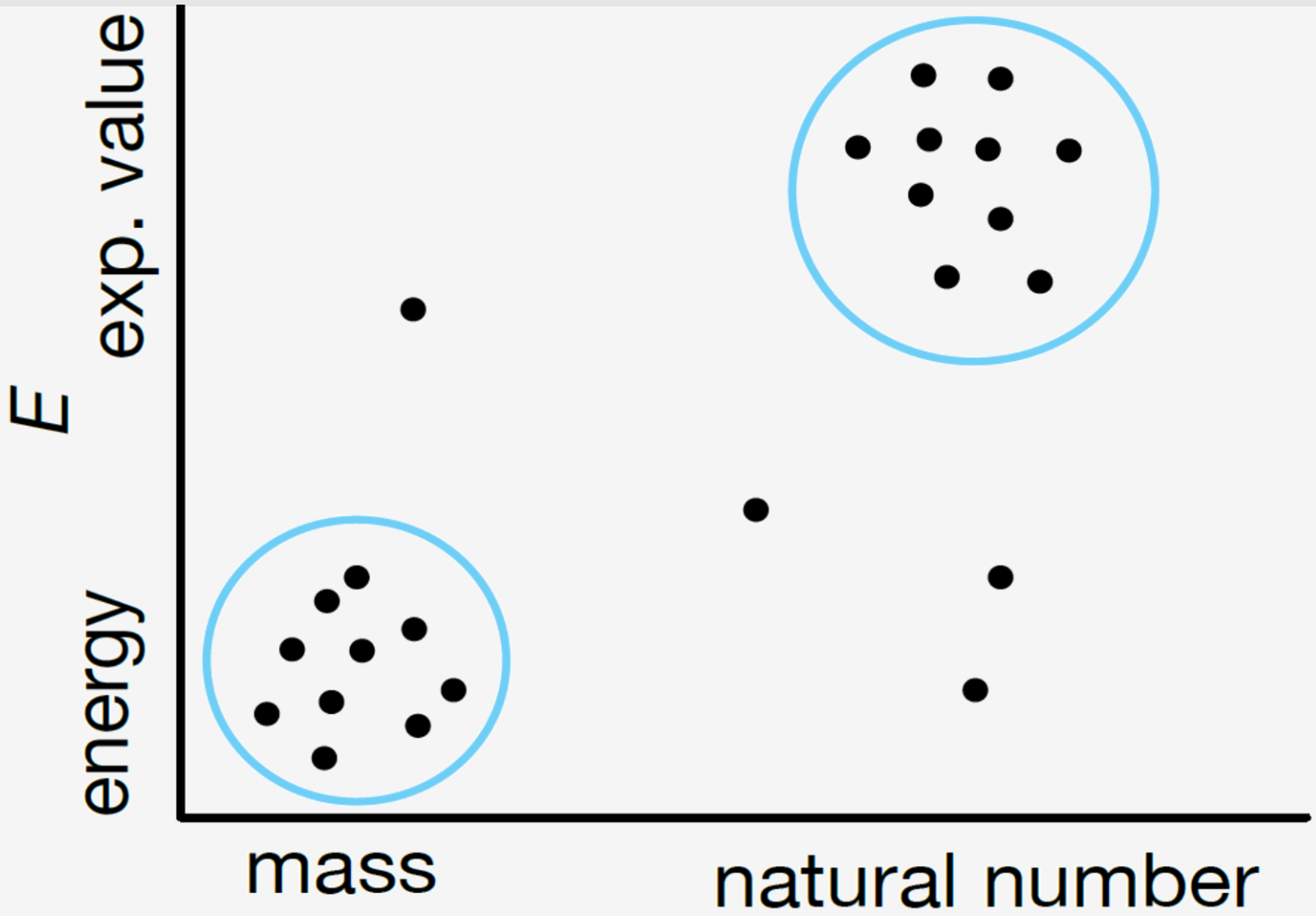
(5) Score all identifer-definiens pairs

	E	m	c	m
energy	2	4	13	22
physical system	5	1	10	20
mass	10	2	5	15
speed of light	19	11	2	5
vacuum	21	14	5	3

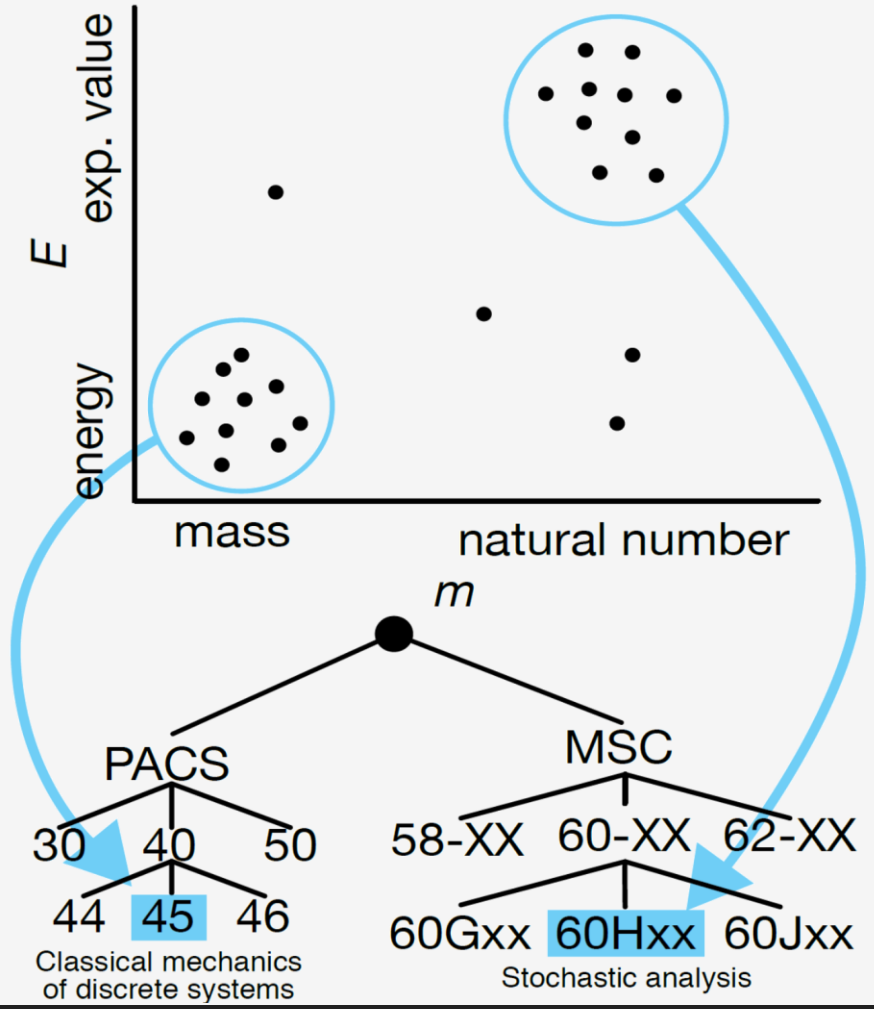
(6) Generate feature vectors

$$
\begin{array}{r}
\\
\text{E\_energy} \\
\text{E\_expected\_value} \\
\text{m\_mass} \\
\text{m\_natural\_number} \\
\vdots
\end{array}
\begin{array}{c}
d_1 \\
\begin{pmatrix} 2 \\ - \\ 2 \\ - \\ \vdots \end{pmatrix}
\end{array}
\begin{array}{c}
d_2 \\
\begin{pmatrix} - \\ 2 \\ - \\ 0 \\ \vdots \end{pmatrix}
\end{array}
\begin{array}{c}
\cdots \\
\\
\cdots \\
\\
\\
\end{array}
$$

(7) Cluster feature vectors



(8) Map clusters to subject hierarchy



Wikipedia Subtask at NTCIR 11

- NTCIR 11 Wikipedia dataset*
- 30k Wikipedia Articles
- 280k Formulae
- 100 queries

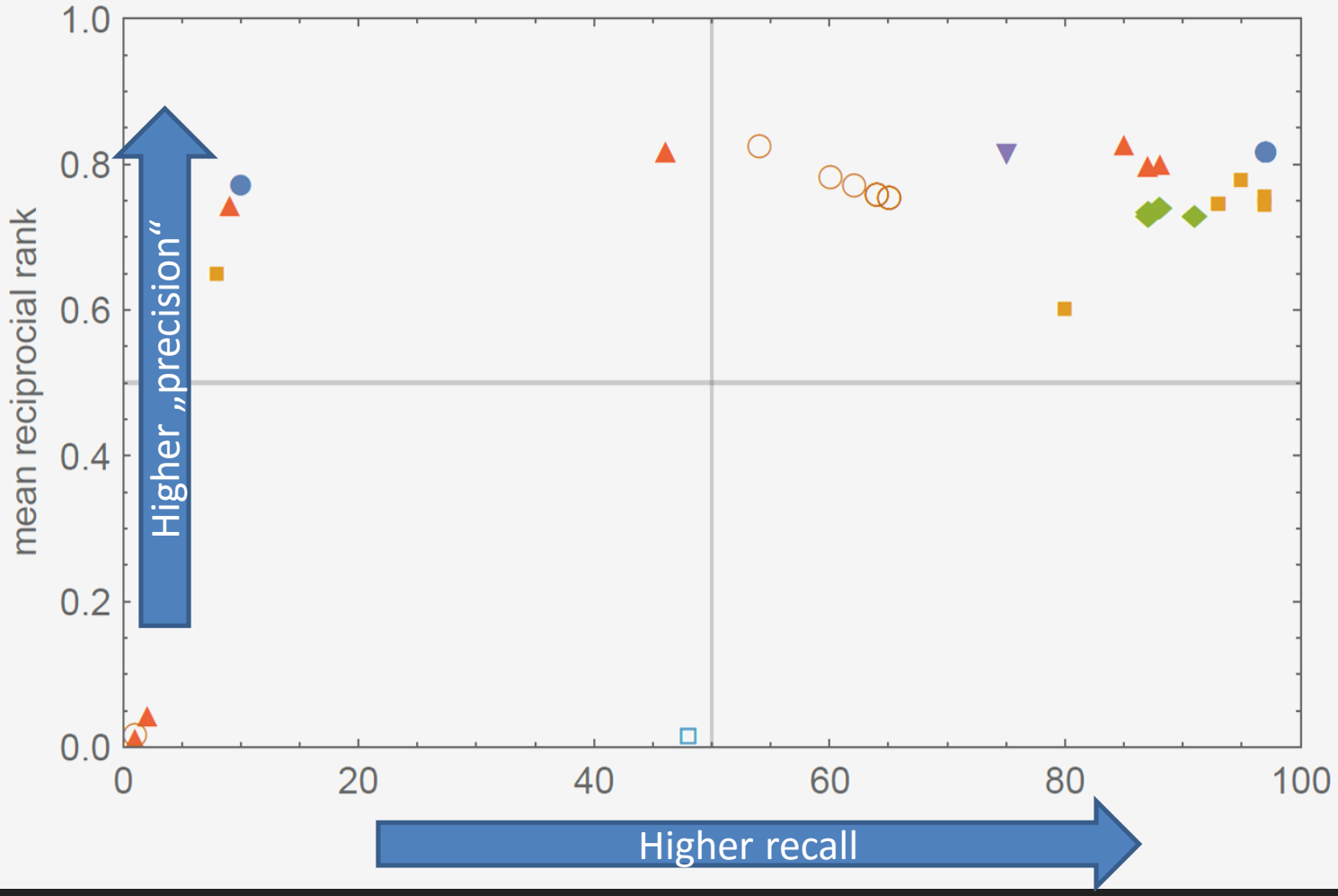
*) Moritz Schubotz, Abdou Youssef, Volker Markl, and Howard S. Cohl. 2015. Challenges of Mathematical Information Retrieval in the NTCIR-11 Math Wikipedia Task. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '15). ACM, New York, NY, USA, 951-954.

Wikipedia Subtask at NTCIR 11

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

arXiv (8 Participants) Wikipedia (7 Participants)

Wikipedia Task results



Gold standard

available from http://mlp.formulasearchengine.com

(1) Van der Waerden's theorem: $W(2,k) > 2^k / k^\varepsilon$

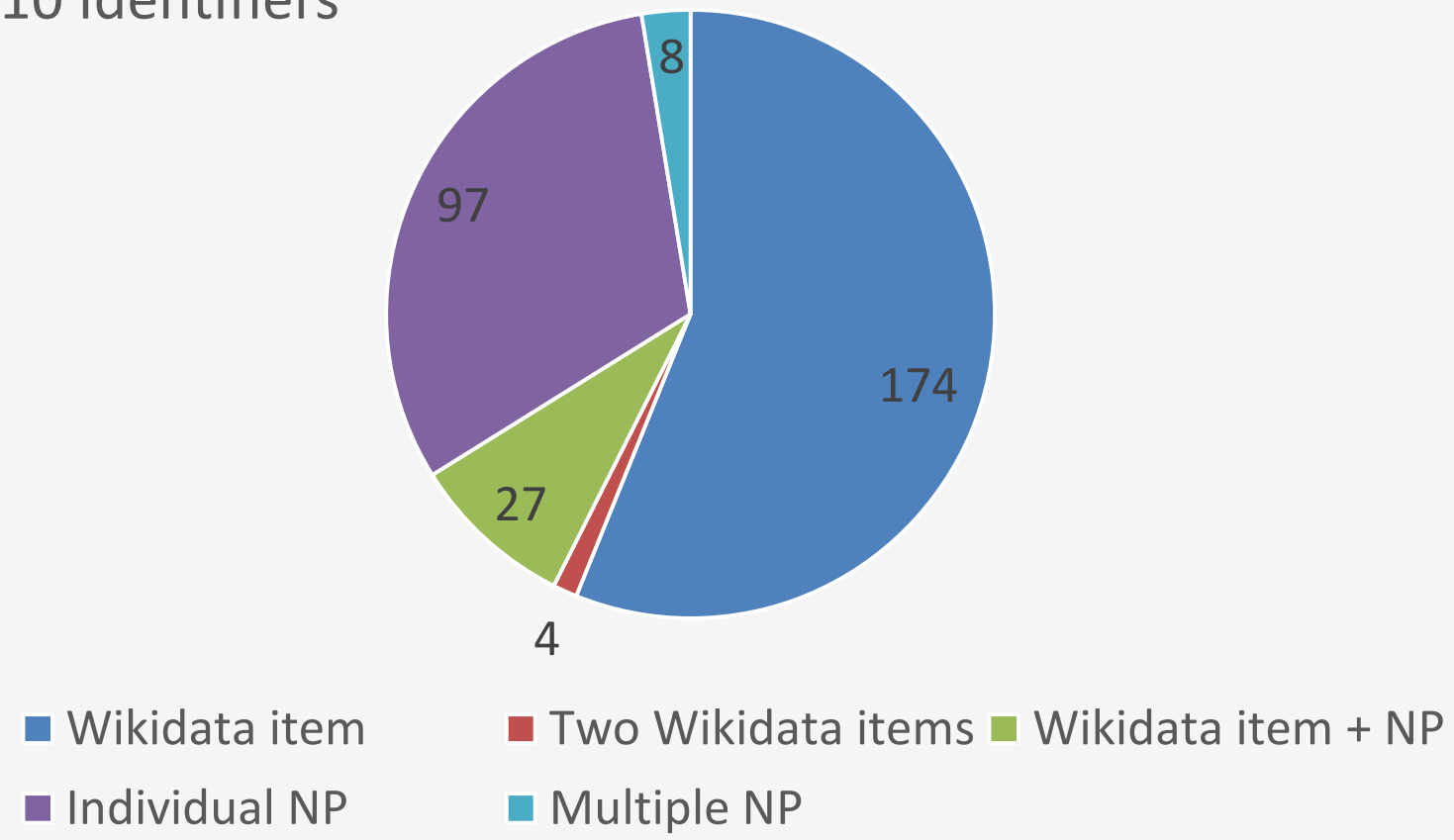
W **Van der Waerden number**
k **integer** : number that can be written without a fractional or decimal component
ε **positive number** (real number...)

(69) Engine efficiency: $\eta = \frac{work\ done}{heat\ absorbed} = \frac{Q1 - Q2}{Q1}$

η **energy efficiency**
Q_1 **heat** (energy)
Q_2 **heat** (energy)

Gold standard details

310 Identifiers



Results: Identifiers

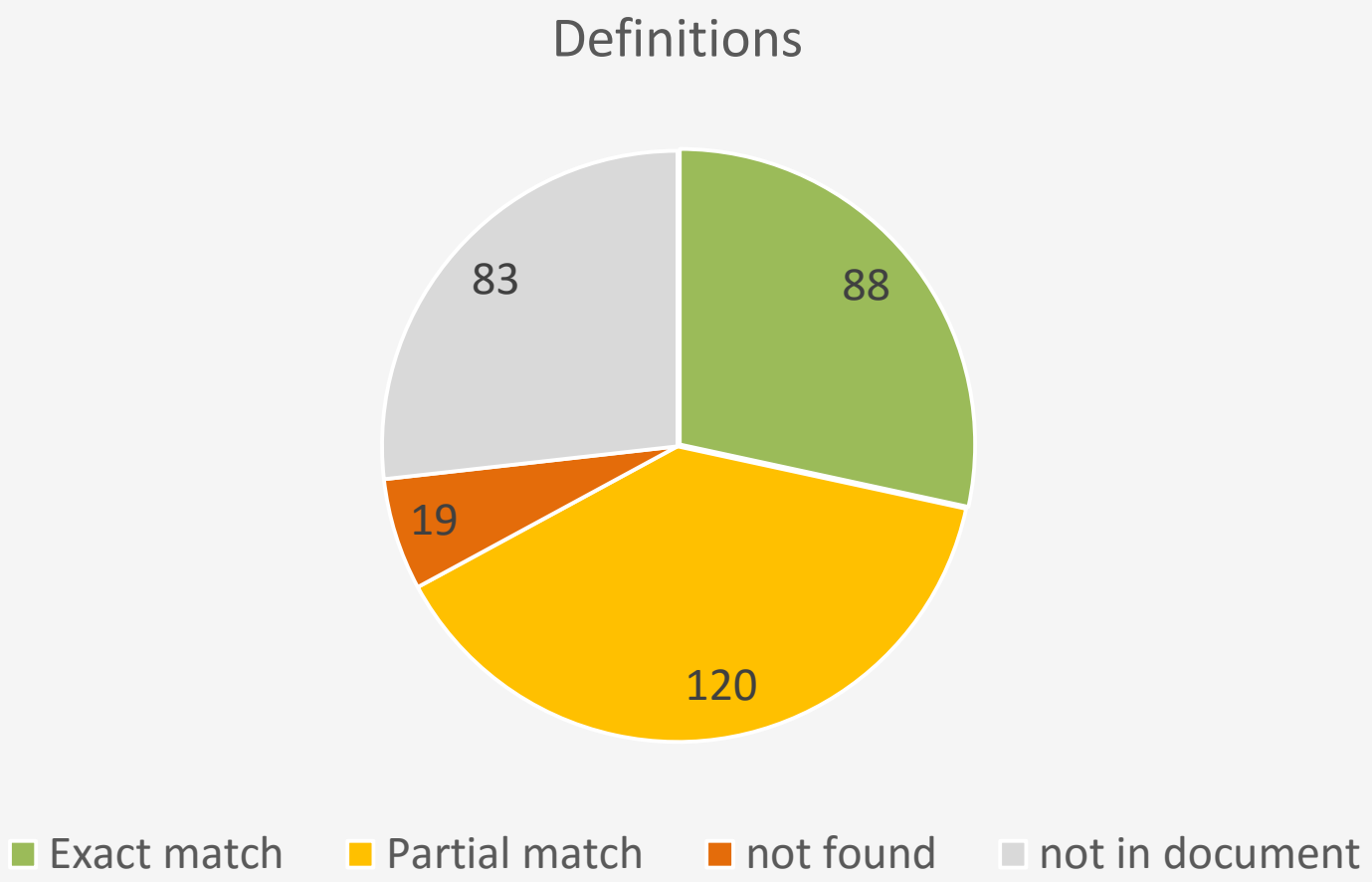
- 294/310 correctly extracted (94.8%)
- 57 false positive (fp)
- Problems
  - Incorrect markup (8fn, 33fp) $\eta = \frac{Q1-Q2}{Q1}$ $Lin(v, v')$
  - Symbols (9fp) $\frac{\mathrm{d}}{\mathrm{d}x}$
  - Sub-super script (3fp, 2fn) σ_y^2
  - Special notation (10fp, 2fn) $\boldsymbol{u} \times \boldsymbol{v} = \epsilon^i{}_{jk} u^j v^k \boldsymbol{e_i}$

Results: (4) Find definiens candidates

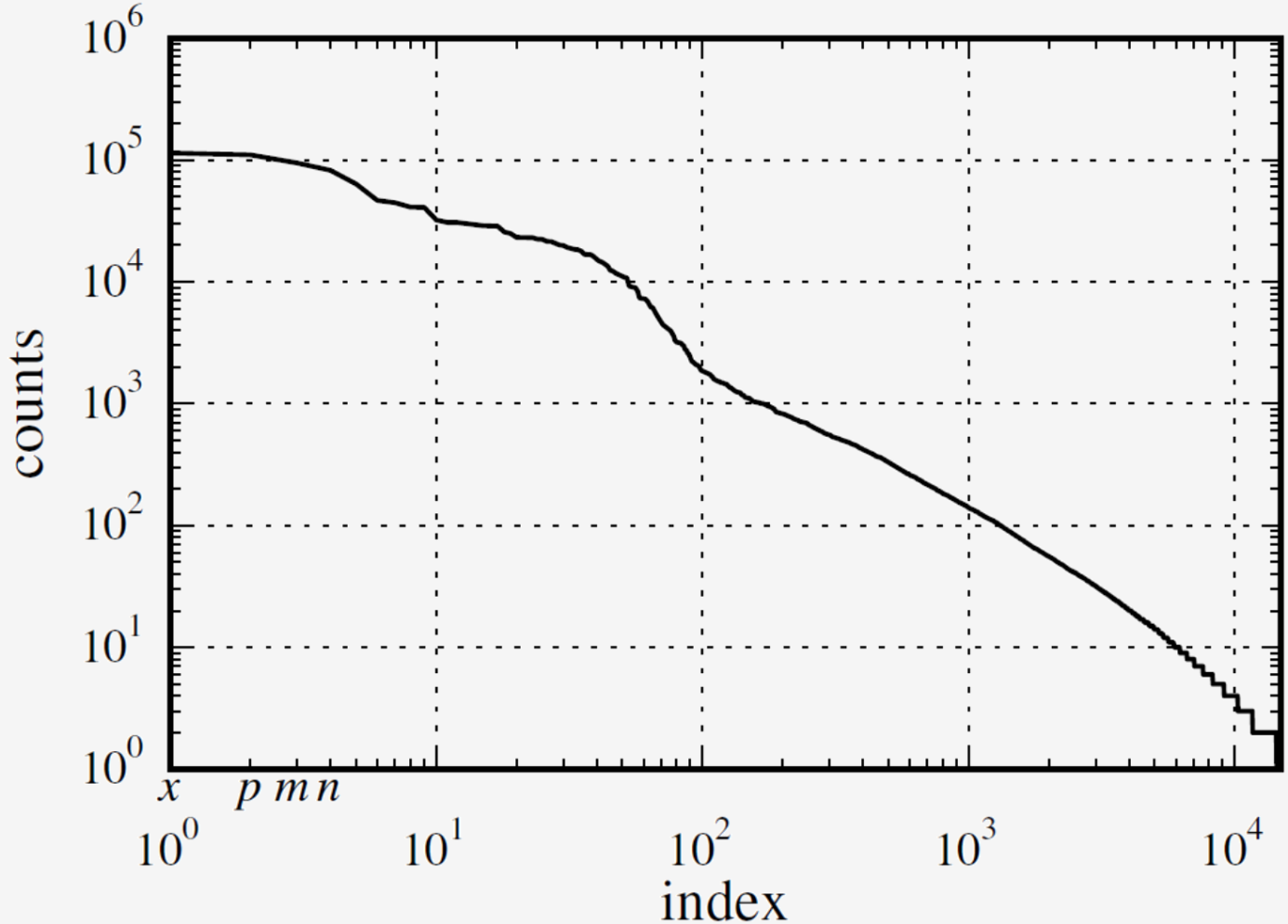
$$E = mc^2$$

where E is the energy of a physical system, m is the mass of the system, and c is the speed of light in a vacuum (about 3×10^8 m/s). In words, energy equals mass multiplied by the speed of light squared. Because the speed of light is a very large number

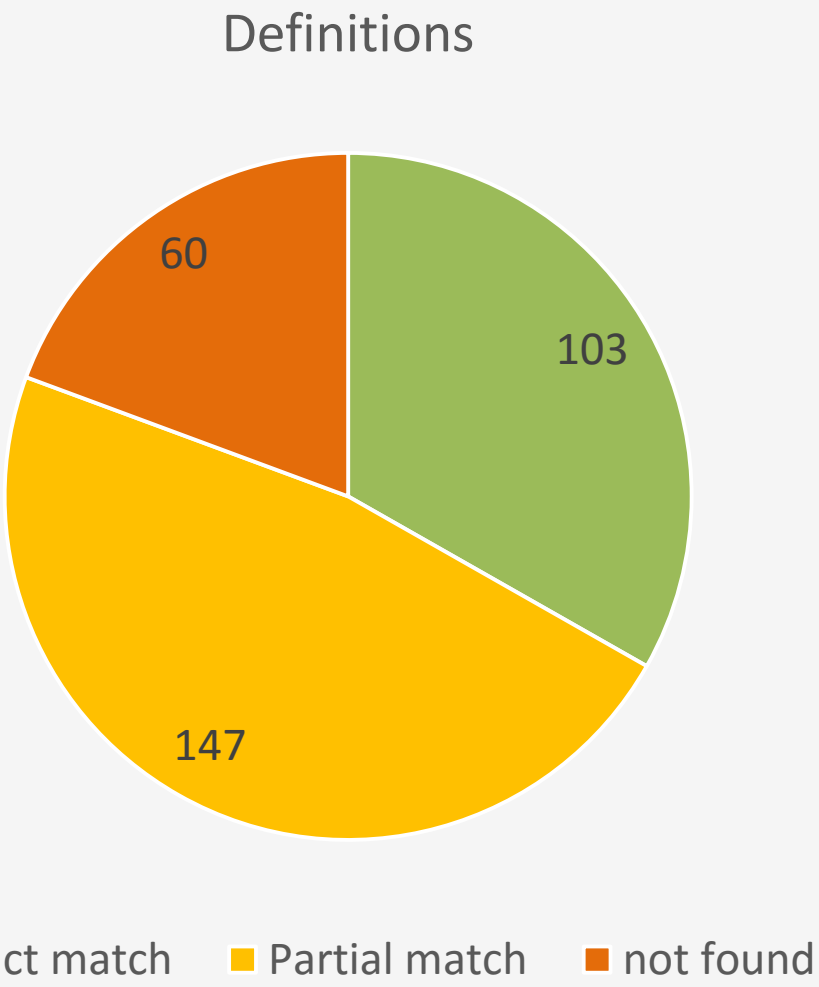
Results: Definitions



Distribution of identifier counts



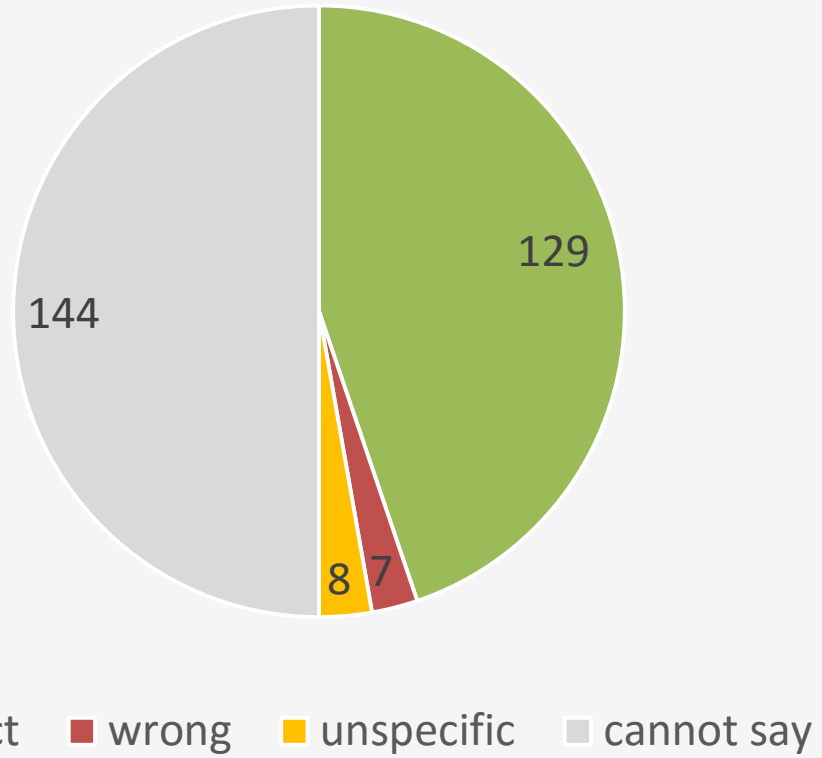
Results: Definitions with Namespace support



Discovered namespaces

- 250 clusters -> 167 mapped to classification schemata
- 5618 definitions with $s > 1$ (2124 Wikidata concepts)
- Evaluate 6 randomly sampled namespaces

Impression from namespace samples



Discovered namespaces

Classical mechanics of discrete systems 45.00 (PACS)
Categories: Physics, Mechanics, Classical mechanics
Purity: 61%, matching score: 31%,
identifiers 103, semantic concepts 50, ✓ 58, ✓ 4, ? 42, ✗ 1
Identifier-definitions:

m **mass** (quantitative measure of a physical object's resistance to acceleration by a force . . .) [$s \approx 29$] ✓

$\mathbf{F}$ **force** (influence that causes an object to change) [$s \approx 25$] ✓

$\mathbf{v}$ **velocity** (rate of change of the position of an object . . . and the direction of that change) [$s \approx 24$] ✓

t **time** (dimension in which events can be ordered the past through the present into the future) [$s \approx 19$] ✓

Stochastic analysis 60Hxx (MSC)
Categories: Stochastic processes, Probability theory
Purity: 92%, matching score: 62%, identifiers 54, semantic concepts 32, ✓ 18, ✓ 0, ? 30, ✗ 0
Identifier-definitions:

a **stochastic process** (. . . random variables) [$s \approx 12$] ✓

X **stochastic process** (. . . random variables) [$s \approx 10$] ✓

...

$\mathbf{E}$ **expected value** [$s \approx 2$] ✓

...

$\mathbb{E}$ **expected value** $s < 1$

v function $s < 1$

- Physics
- Identifiers are significant for formulae

- Mathematics
- Identifiers might be less significant for formulae

Conclusions

- For 10% of the identifiers used in Wikipedia (en) we could assign the associated Wikidata item
- →90% ahead
  - More specific Wikidata items needed
  - Combine data from different language versions
  - Improve recognition rate within a document
- Namespaces for mathematical identifiers could be identified

Next steps

- The identifier information is available from the Wikipedia API today http://en.wikipedia.org/api
- Develop tools to augment the user experience for math in Wikipedia and beyond
  - Tooltips (ongoing)
  - Physical Dimensions (ongoing)
  - Translation to Computer Algebra Systems (started)
  - Math Question and Answering (ongoing)
  - Author assistance (ongoing with WMF)
  - Related formulae search (started)
- →Justification for semantification effort

Contact

Moritz Schubotz (now at Universität Konstanz)

moritz@schubotz.de

+49 7531 88 4438

Mobile: +49 1578 047 1397

www.isg.uni-konstanz.de

www.formulasearchengine.com