Identifier Namespaces in Mathematical Notation

Master Thesis by

Alexey Grigorev

Advisers: Moritz Schubotz, Juan Soto

Supervisor: Prof. Dr. Volker Markl

Outline

1. **Motivation**
2. Namespace Discovery
3. Implementation
4. Evaluation
5. Conclusions

[[Namespace]]

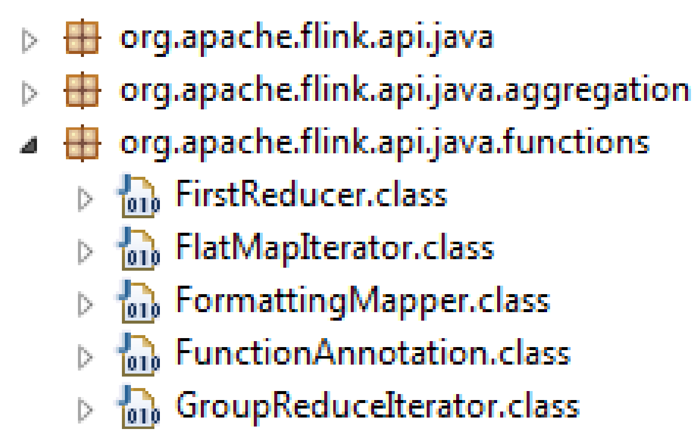
“In programming, **namespaces** are employed to **group symbols and identifiers** around a particular functionality and to **avoid name collisions** between multiple identifiers that share the same name”

No namespaces (C, old PHP)


```
$foo = new Zend_CodeGenerator_Php_Class();
```

http://framework.zend.com/apidoc/1.11/

With namespaces (C++, Java, C#, Python)




```
import o.a.f.api.java.ExecutionEnvironment;
ExecutionEnvironment.getExecutionEnvironment()
```

https://flink.apache.org/

Namespaces in Mathematics

What is it?

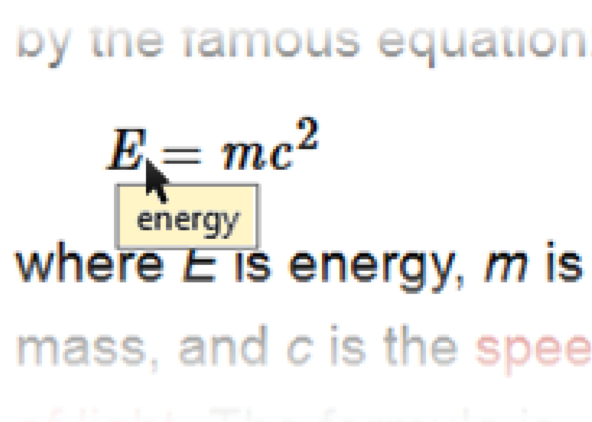
$$E = mc^2$$

Energy

Expected Value

Elimination matrix

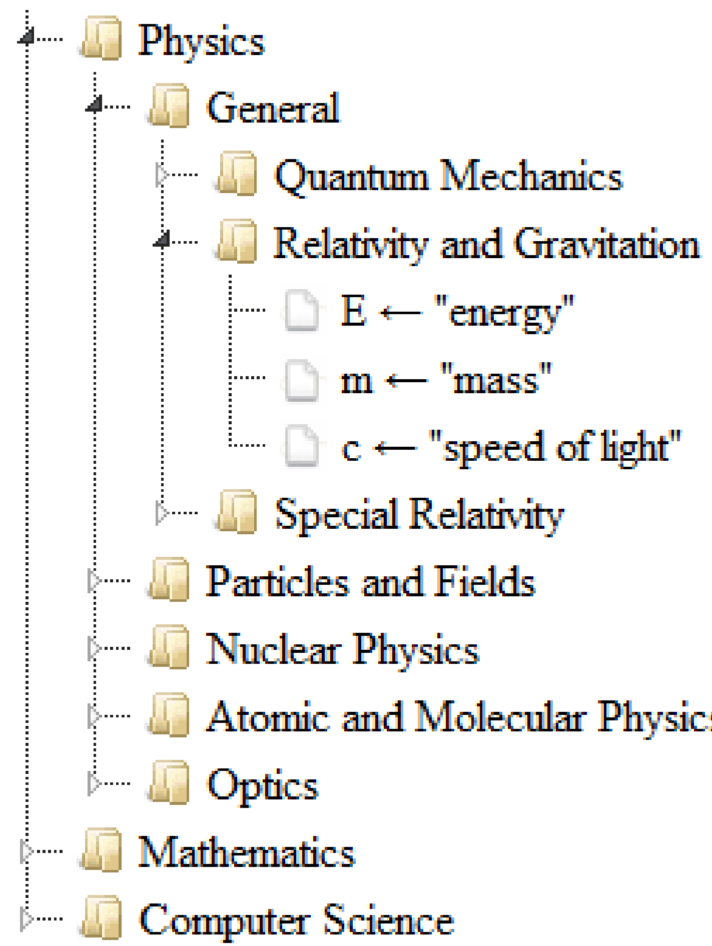
- Can resolve it by introducing namespaces to Mathematics
  - `import Physics/General/Relativity and Gravitation/{E, m, c}`
- It can give:
  - identifier disambiguation
  - better user experience
  - additional context



credit: [MLP]

Namespaces in Mathematics

- Problem: How to organize identifiers into namespaces?
- Manual assignment would take a lot of time
- Our approach: employ **automatic namespace discovery** from a collection of documents



"energy" "mass" "speed of light"

$$E = mc^2$$

```
import Physics/General/Relativity and Gravitation/{E, m, c}
```

Outline

1. Motivation
2. **Namespace Discovery**
3. Implementation
4. Evaluation
5. Conclusions

Definition Extraction

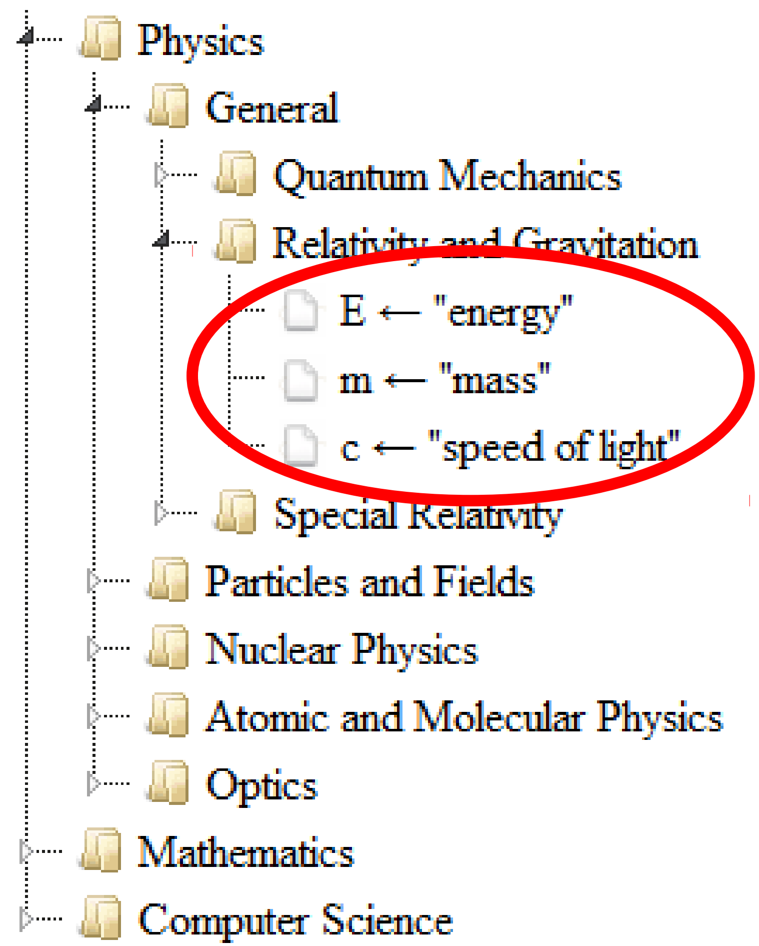
How to get the definitions? Extract them!

[[Mass–energy equivalence]]

“The equivalence of energy *E* and mass *m* is reliant on the speed of light *c* and is described by the famous equation:

$$E = mc^2$$ ”

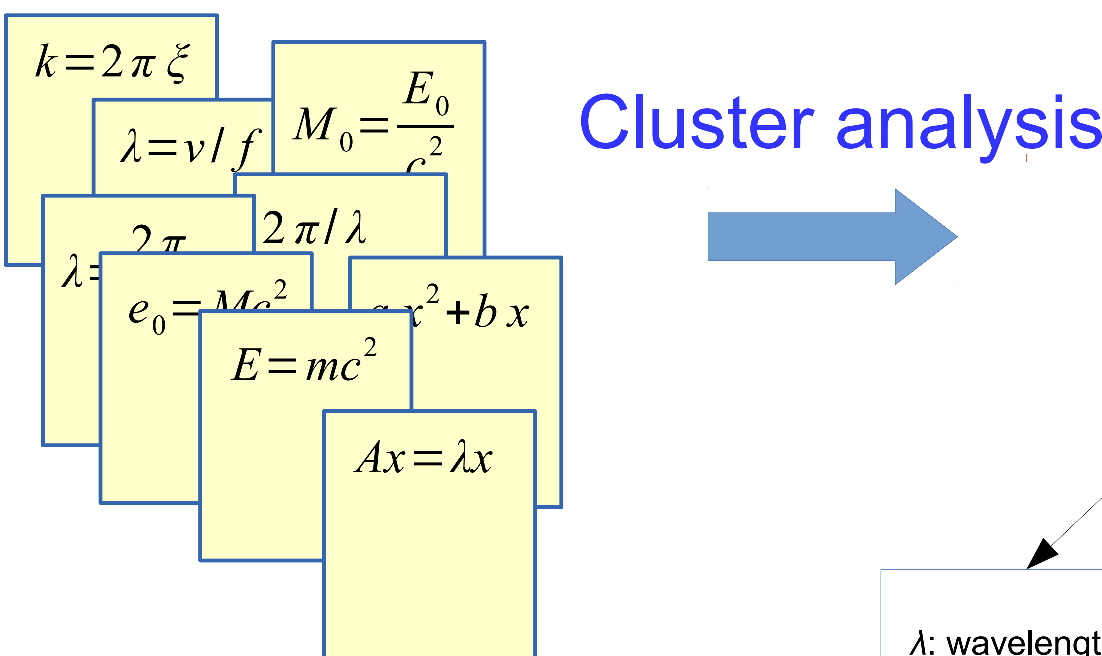
ID	Definition
E	energy
m	mass
c	speed of light

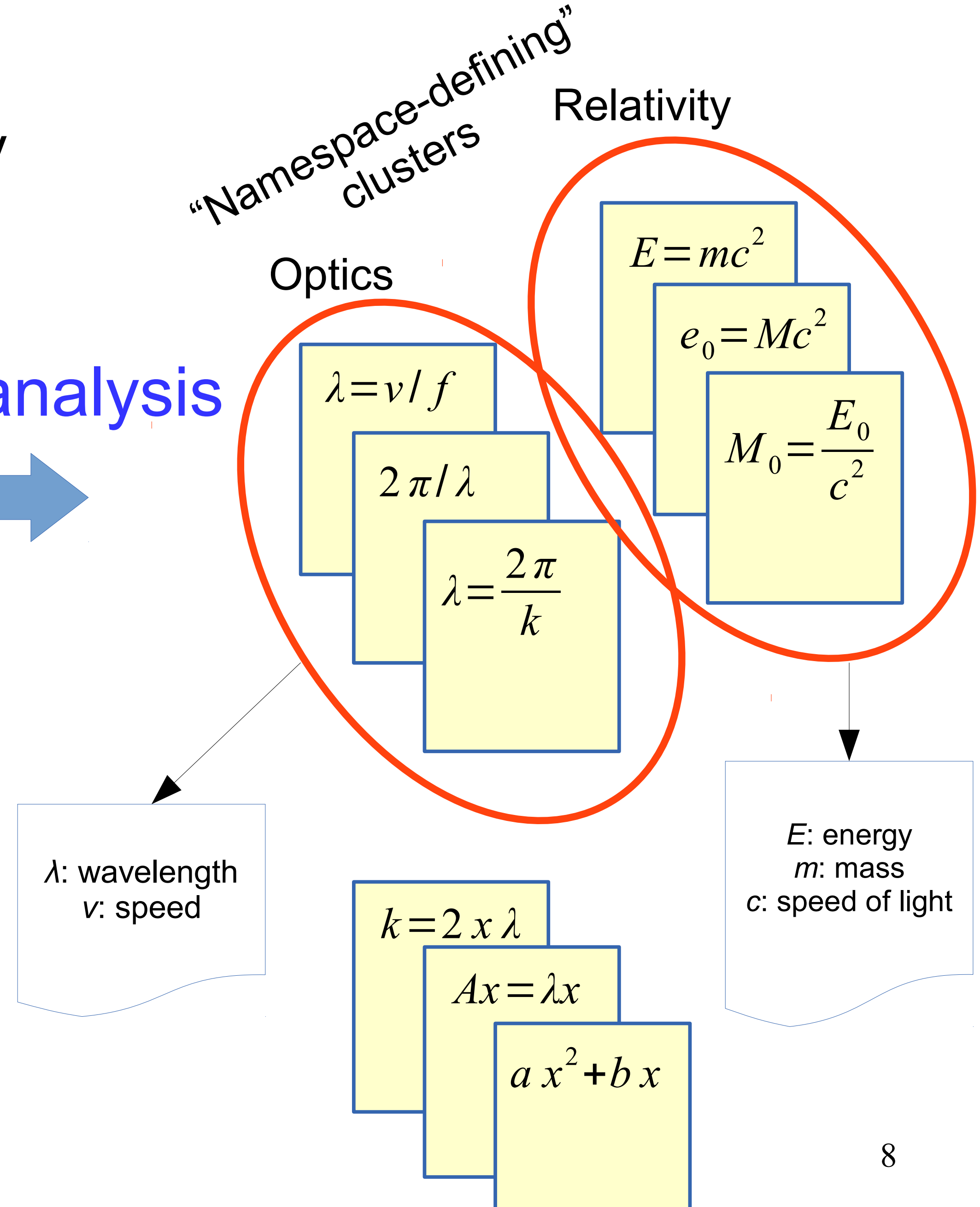


[MLP]

Namespace Discovery

Want to find groups of documents that use identifiers in the same way

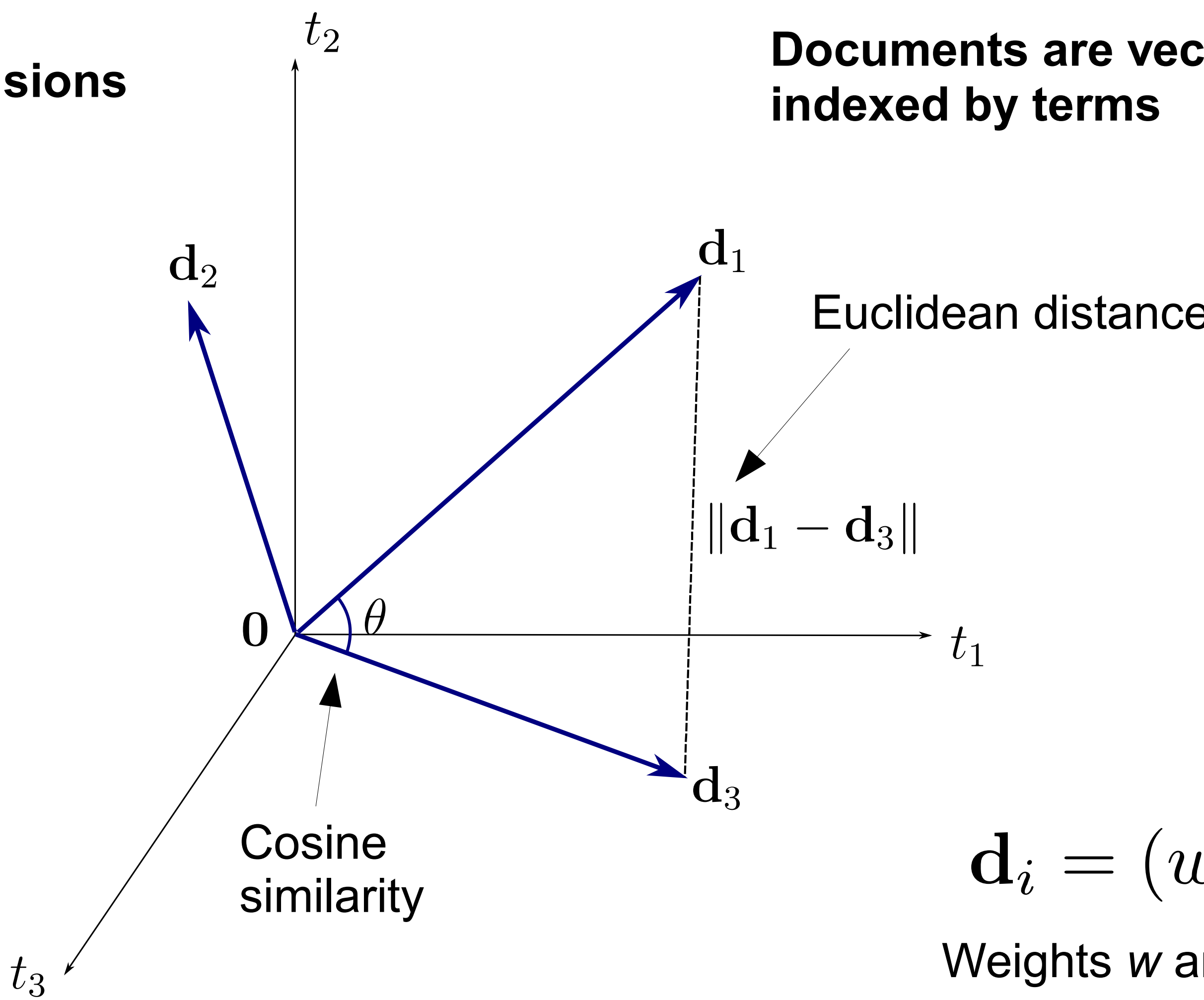




Vector Space Model (VSM)

Terms are dimensions

Documents are vectors indexed by terms



$$\mathbf{d}_i = (w_1, w_2, w_3)$$

Weights *w* are TF or TF-IDF

[IR]

Identifier VSM

d_1

	Definition
E	energy
m	mass
c	speed of light

d_2

	Definition
m	mass
c	speed of light

d_3

	Definition
E	energy

d_4

	Definition
m	integer
c	constant

Build identifier-document matrix

TF of terms

	d_1	d_2	d_3	d_4
E	1	0	1	0
m	1	1	0	1
c	1	1	0	1

No definitions

	d_1	d_2	d_3	d_4
E	1	0	1	0
m	1	1	0	1
c	1	1	0	1
energy	1	0	1	0
mass	1	1	0	0
speed of light	1	1	0	0
integer	0	0	0	1
constant	0	0	0	1

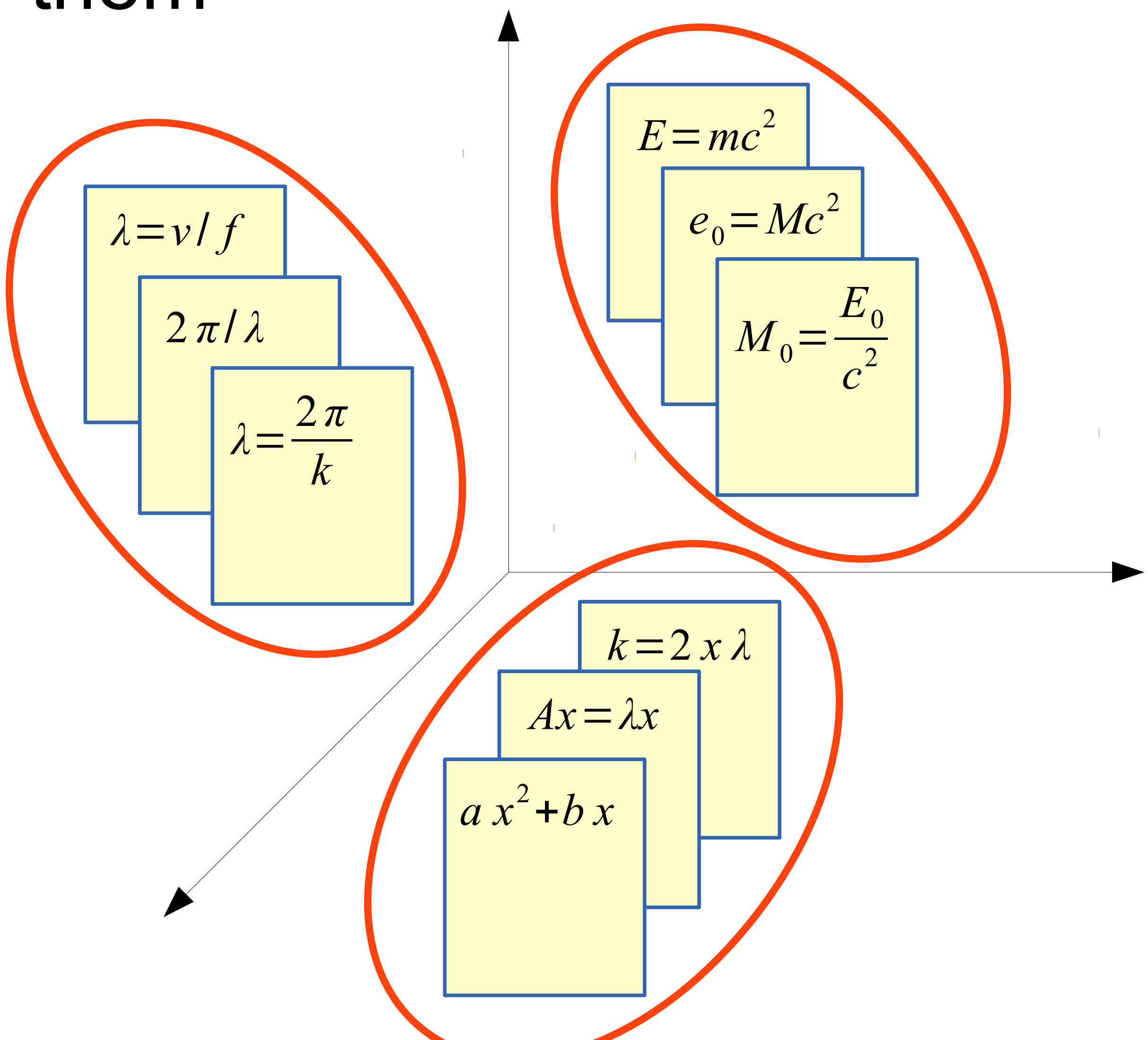
"Weak" association

	d_1	d_2	d_3	d_4
E_energy	1	0	1	0
m_mass	1	1	0	0
c_speed of light	1	1	0	0
m_integer	0	0	0	1
c_constant	0	0	0	1

"Strong" association

Document Clustering

Once documents are represented using vectors we can cluster them



We can employ:

- K-Means [IR]
- DBSCAN [SNN]
- LSA [LSI]

Outline

1. Motivation
2. Namespace Discovery
3. **Implementation**
4. Evaluation
5. Conclusions

Implementation

Definition extraction → Namespace discovery →

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

our contribution

Images: [[File:Wikipedia-logo-en-big.png]] https://flink.apache.org/material.html
http://www.freeflagicons.com http://scikit-learn.org/stable/

Physics
General
Particles and Fields
Nuclear Physics
Optics
Mathematics
Computer Science

Wikipedia
The Free Encyclopedia

Definition extraction

Namespace discovery

wikiFilter — Filter pages with <math>

XPath — Extract identifiers

Standford NLP — POS Tagging

MLP Ranking — Rank definition candidates

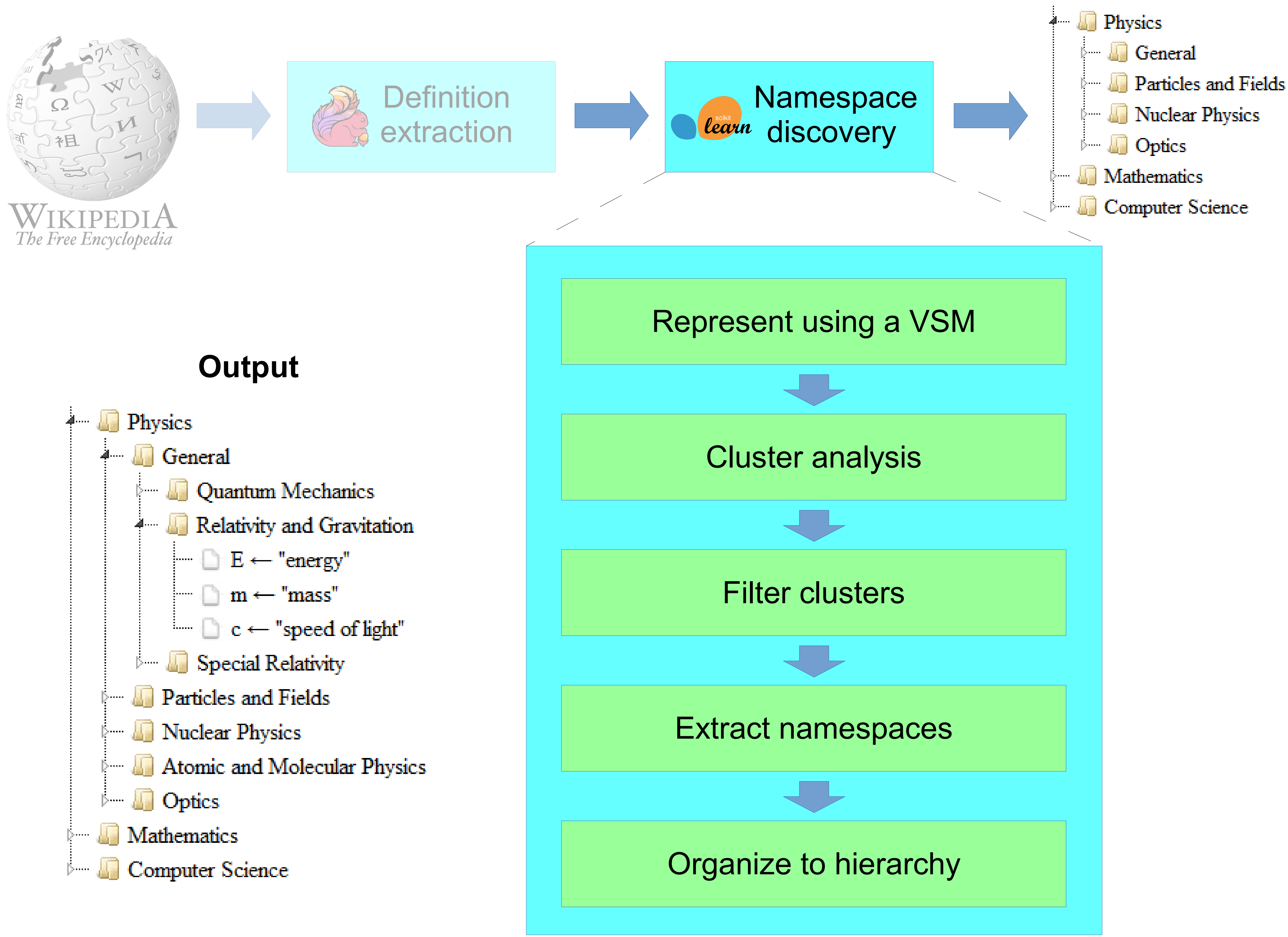
Output

[[Mass-energy equivalence]]

IDs: [E, m, c, μ, ...]

	Def	Score
E	enegy	0.99
m	mass	0.99
c	speed of light	0.90

[MLP] Based on open source project by Robert Pagel


WIKIPEDIA
The Free Encyclopedia
Definition extraction
scikit learn
Namespace discovery
Physics
General
Particles and Fields
Nuclear Physics
Optics
Mathematics
Computer Science
Represent using a VSM
Cluster analysis
Filter clusters
Extract namespaces
Organize to hierarchy
Output
Physics
General
Quantum Mechanics
Relativity and Gravitation
E ← "energy"
m ← "mass"
c ← "speed of light"
Special Relativity
Particles and Fields
Nuclear Physics
Atomic and Molecular Physics
Optics
Mathematics
Computer Science

WIKIPEDIA
The Free Encyclopedia

Definition extraction

Namespace discovery

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

Cluster analysis

Filter clusters

Extract namespaces

Organize to hierarchy

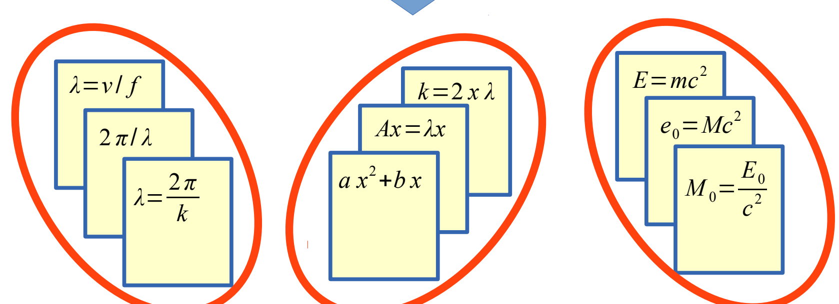
scikit learn

`TfidfVectorizer(min_df=2)`

`Kmeans` and `MiniBatchKMeans`

DBSCAN

`randomized_svd` and `NMF`



WIKIPEDIA
The Free Encyclopedia

Definition extraction → Namespace discovery

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

Cluster analysis

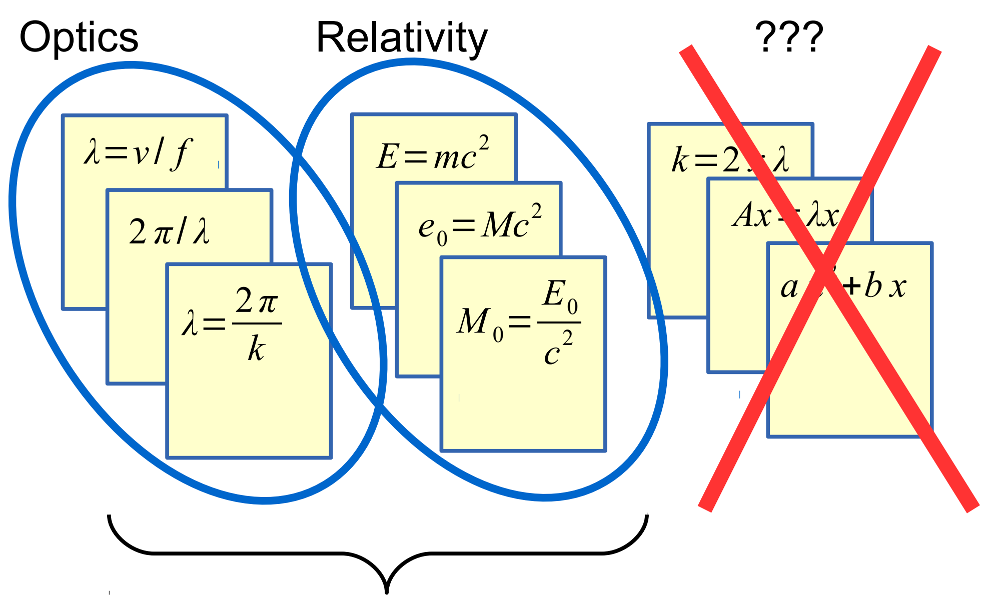
Filter clusters

Extract namespaces

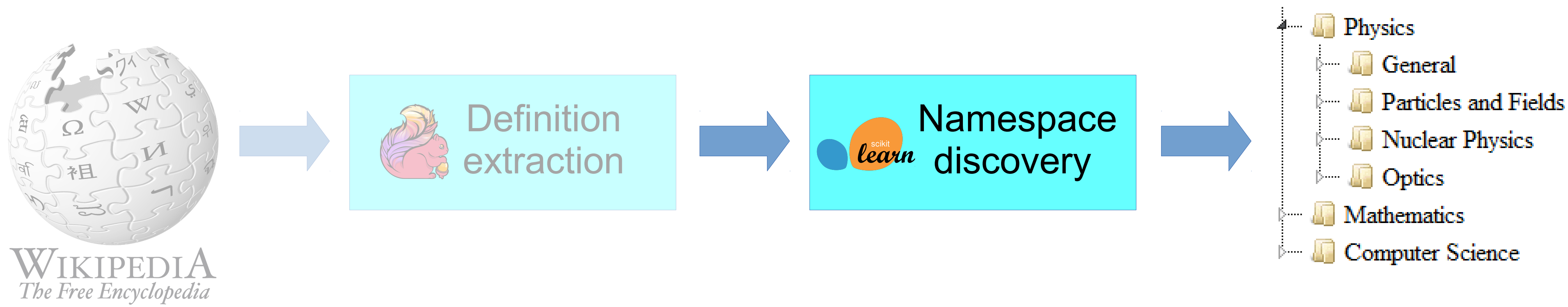
Organize to hierarchy

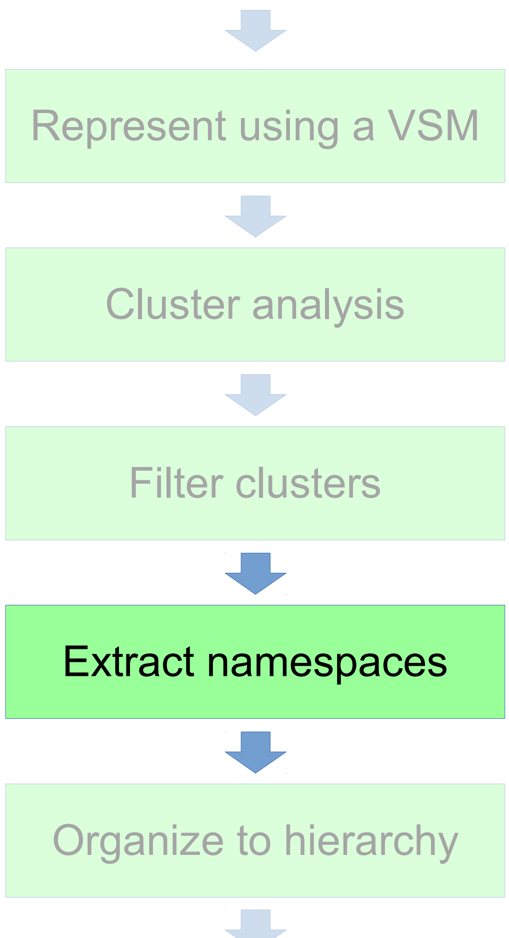
All obtained clusters are "*homogenous*": within-cluster similarity is maximal.

We keep those clusters whose documents correspond to the the same category. Otherwise, we discard uncategorised clusters.



"namespace-defining" clusters





$$E=mc^2$$

	Def	S
E	energy	0.99
m	mass	0.99
c	speed of light	0.90
c	speed of light	1.00

$$E_0=M_0c^2$$

	Def	S
E_0	energy	0.90
M_0	mass	0.99
c	speed of light	0.90

$$M_0=\frac{E_0}{c^2}$$

	Def	S
E_0	energy	0.99
M_0	mass	0.95
c	speed of light	0.90
c	energy	0.80

Relativity

	Def	S
E	energy	0.99
m	mass	0.99
c	speed of light	3.70
e_0	energy	0.99
M_0	mass	1.94
E_0	energy	1.89
~~c~~	~~energy~~	~~0.80~~

WIKIPEDIA
The Free Encyclopedia

Definition extraction → Namespace discovery →

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

Cluster analysis

Filter clusters

Extract namespaces

Organize to hierarchy

$$E=mc^2$$

	Def	S
E	energy	0.99
m	mass	0.99
c	speed of light	0.90
c	speed of light	1.00

$$E_0=M_0c^2$$

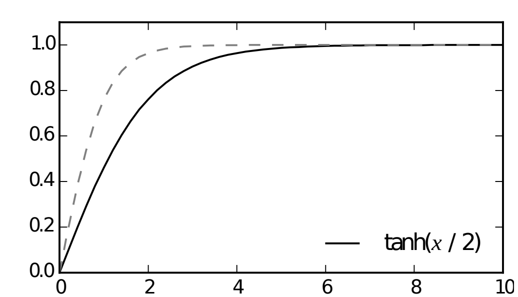
	Def	S
E_0	energy	0.90
M_0	mass	0.99
c	speed of light	0.90

$$M_0=\frac{E_0}{c^2}$$

	Def	S
E_0	energy	0.99
M_0	mass	0.95
c	speed of light	0.90
c	energy	0.80

Relativity

	Def	S
E	energy	0.46
m	mass	0.46
c	speed of light	0.87
e_0	energy	0.46
M_0	mass	0.60
E_0	energy	0.57
c	energy	0.20



Definition extraction → Namespace discovery →

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

Cluster analysis

Filter clusters

Extract namespaces

Organize to hierarchy

	Def	S
E	energy	0.99
m	mass	0.99
c	speed of light	3.70
c	speed of light in vacuum	0.99
m	mass	1.94
m	total mass	1.89

→

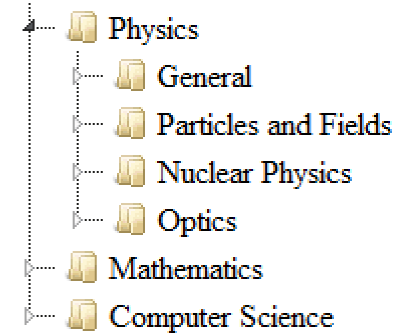
	Def	S
E	energy	0.99
c	***speed of light**	4.69
m	***mass**	4.82

Fuzzy grouping

FuzzyWuzzy https://github.com/seatgeek/fuzzywuzzy

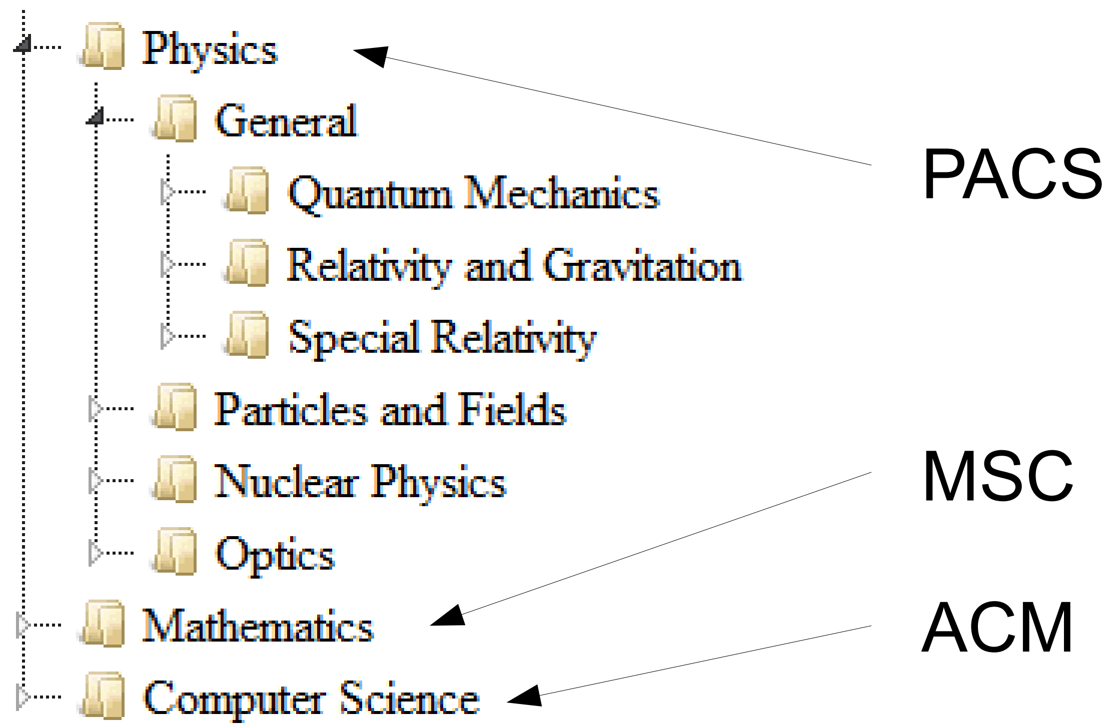
WIKIPEDIA
The Free Encyclopedia

Definition extraction → Namespace discovery →



- Represent using a VSM
- Cluster analysis
- Filter clusters
- Extract namespaces
- **Organize to hierarchy**

A reference hierarchy: drawn from what source?



- Definition extraction
- Namespace discovery

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

- Represent using a VSM
- Cluster analysis
- Filter clusters
- Extract namespaces
- **Organize to hierarchy**

PACS

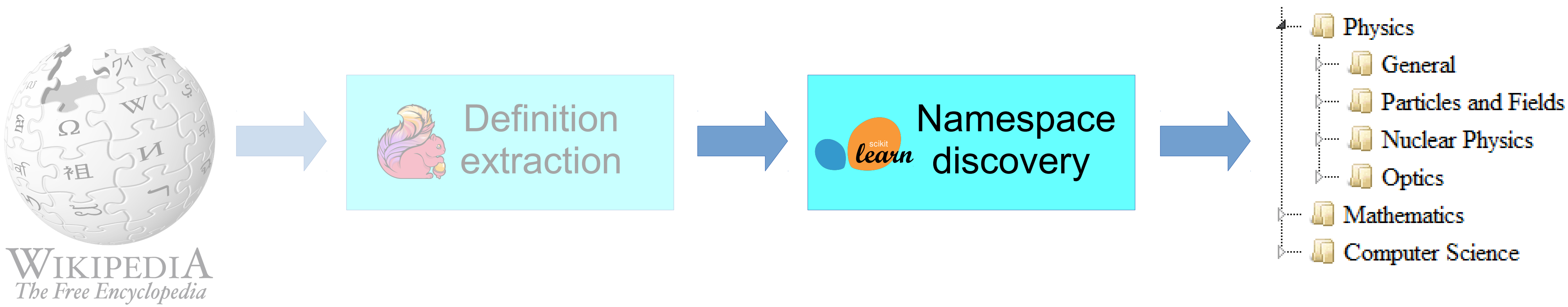
00—General
01. Communication, education, history, and philosophy
02. Mathematical methods in physics
03. Quantum mechanics, field theories, and special relativity
04. General relativity and gravitation
05. Statistical physics, thermodynamics, and no ar dynamical systems
06. Metrology, measurements, and laboratory pr dures
07. Instruments, apparatus
physics and astronomy

10—The Physics of Elemen
11. General theory of fields
12. Specific theories and in
13. Specific reactions and p
14. Properties of specific p

20—Nuclear Physics
21. Nuclear structure
23. Radioactive decay and i
24. Nuclear reactions: gene
25. Nuclear reactions: spec
26. Nuclear astrophysics

04.	**General relativity and gravitation** *(for astrophysical aspects, gravitation; for relativistic aspects of cosmology, see 98.80.Jk; fo 03.30.+p)*
04.20.-q	**Classical general relativity** *(see also 02.40.-k Geometry, diffe topology)*
04.20.Cv	Fundamental problems and general formalism
04.20.Dw	Singularities and cosmic censorship
04.20.Ex	Initial value problem, existence and uniqueness of solutions
04.20.Fy	Canonical formalism, Lagrangians, and variational principles
04.20.Gz	Spacetime topology, causal structure, spinor structure
04.20.Ha	Asymptotic structure
04.20.Jb	Exact solutions
04.25.-g	**Approximation methods; equations of motion**
04.25.D-	Numerical relativity

https://www.aip.org/publishing/pacs/pacs-2010-regular-edition



Represent using a VSM

Cluster analysis

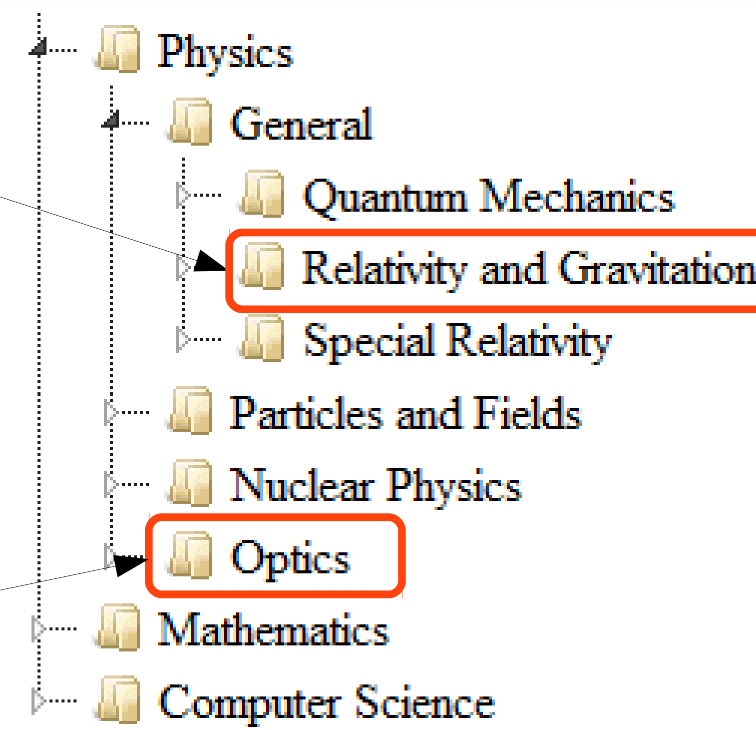
Filter clusters

Extract namespaces

Organize to hierarchy

	Def	S
E	energy	0.99
c	***speed of light**	0.96
m	***mass**	0.87

	Def	S
λ	wavelength	0.99
k	wavenumber	0.89
f	frequency	1.0



Extract keywords from namespaces

Extract keywords from categories

Calculate cosine between them

Outline

1. Motivation
2. Namespace Discovery
3. Implementation
4. **Evaluation**
5. Conclusions

Java Language Processing

How to evaluate the quality?

- Hard! No ground truth, unsupervised settings
- Use data where ground truth is known: source code!

- org.apache.flink.api.java
- org.apache.flink.api.java.aggregation
- org.apache.flink.api.java.functions
  - FirstReducer.class
  - FlatMapIterator.class
  - FormattingMapper.class
  - FunctionAnnotation.class
  - GroupReduceIterator.class

“definition” identifier

```
package org.apache.flink.api.java.functions;

public class FirstReducer<T> implements ... {

  private final int count;
  // ...

  @Override
  public void reduce(Iterable<T> values, Collector<T> out) {
    int emitCnt = 0;

    for (T val : values) {
      out.collect(val);
      // ...
    }
  }
}
```

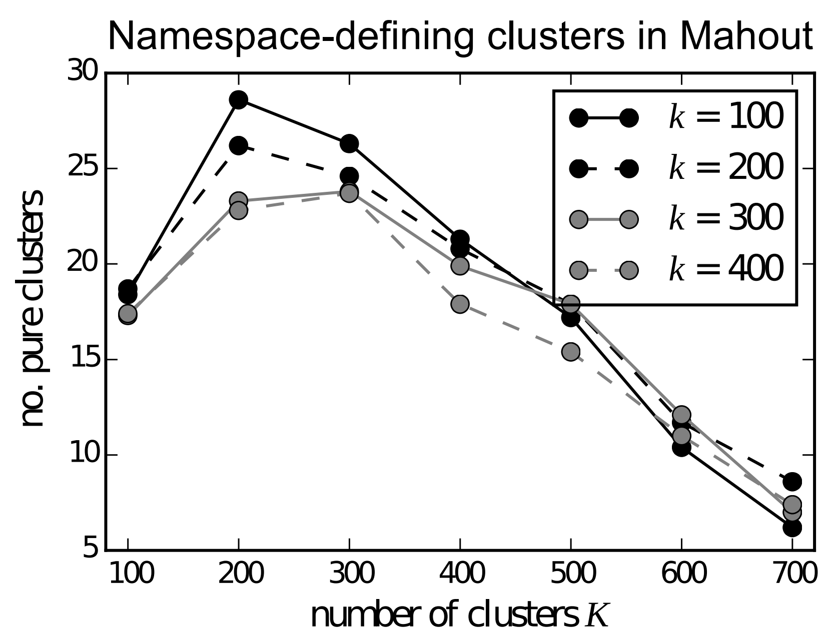

[[File:Java_logo_and_wordmark.svg]]

Java Language Processing

Apache Mahout
- 1560 Java Classes
- 46k variable declarations
- 150 packages

AST Tree
extracted with JavaParser*

Namespace discovery



✓ **Method works**

http://mahout.apache.org/images/
* https://github.com/javaparser/javaparser

Experimental Setup

Definition extraction → Namespace discovery

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

Cluster analysis

Filter clusters

Extract namespaces

Organize to hierarchy

Objective: want to find as many namespace-defining clusters as possible

Cluster is *namespace-defining* if it
- has at least purity *p and*
- contains at least *n* documents

$p = 0.8, n = 5$

Relativity, Gravitation
Gravitation, **Relativity**, Physics
Relativity, Einstein
Physics, **Relativity**
Physics, Gravitation

Purity *p* vs size *n* tradeoff:

- Larger *p* – only pure clusters, smaller *p* – allow some slack
- Larger *n* – only big well-connected clusters are taken into account

Our settings: p ≥ 80% and n ≥ 3

Parameter Tuning

Definition extraction → Namespace discovery →

- Physics
  - General
  - Particles and Fields
  - Nuclear Physics
  - Optics
- Mathematics
- Computer Science

Represent using a VSM

- Identifier VSM: no-def, weak, strong
- Weighting: TF, TF-IDF, logTF-IDF

Cluster analysis

- DBSCAN
  - base similarity function, ε, MinPts
- K-Means
  - number of clusters K
- Latent Semantic Analysis
  - matrix decomposition: SVD or NMF
  - rank of reduced matrix k

Filter clusters

Extract namespaces

Organize to hierarchy

Baseline

Definition extraction → Namespace discovery →

Physics
General
Particles and Fields
Nuclear Physics
Optics
Mathematics
Computer Science

~~Represent using a VSM~~

Cluster analysis

Filter clusters

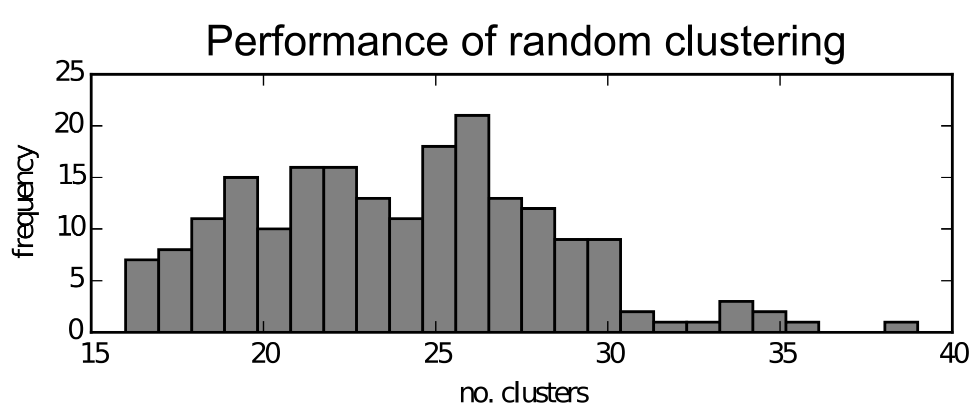
Extract namespaces

Organize to hierarchy

Random cluster assignment

Algorithm:

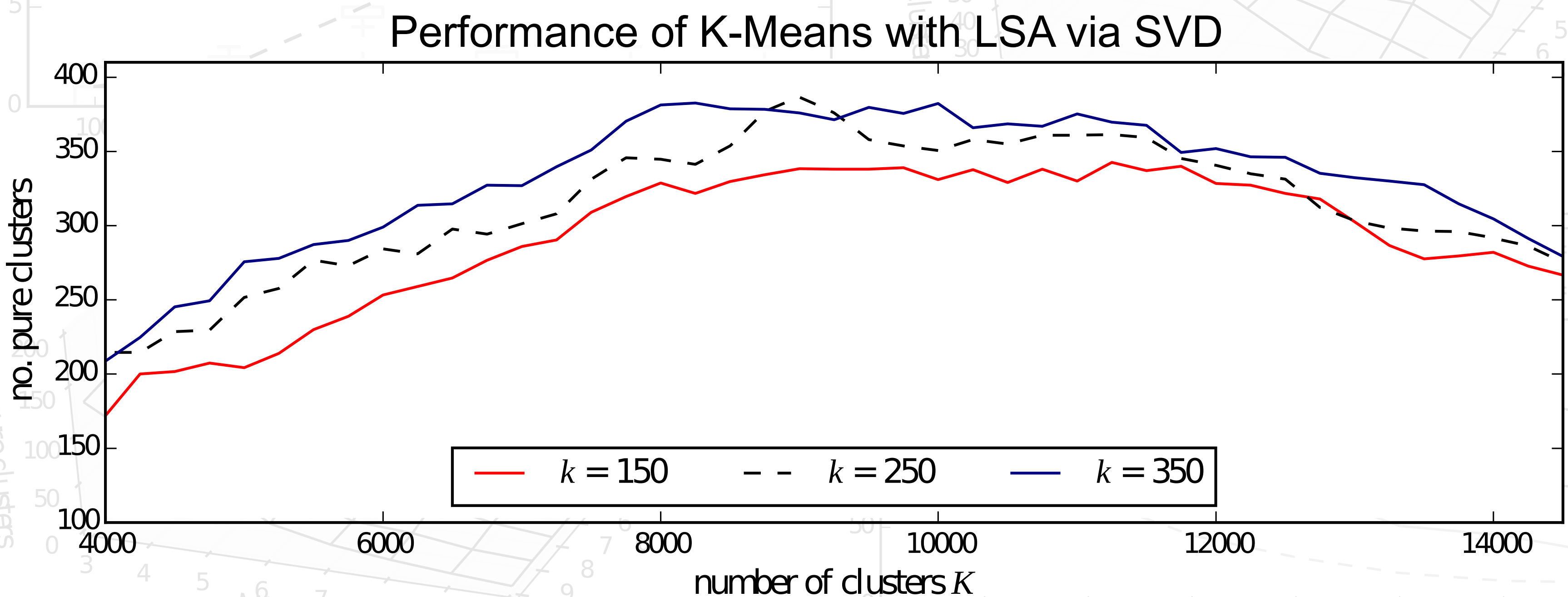
- let k = 0
- take three unseen documents at random
- assign them to cluster k
- increment k
- repeat until no documents left



Parameter Tuning

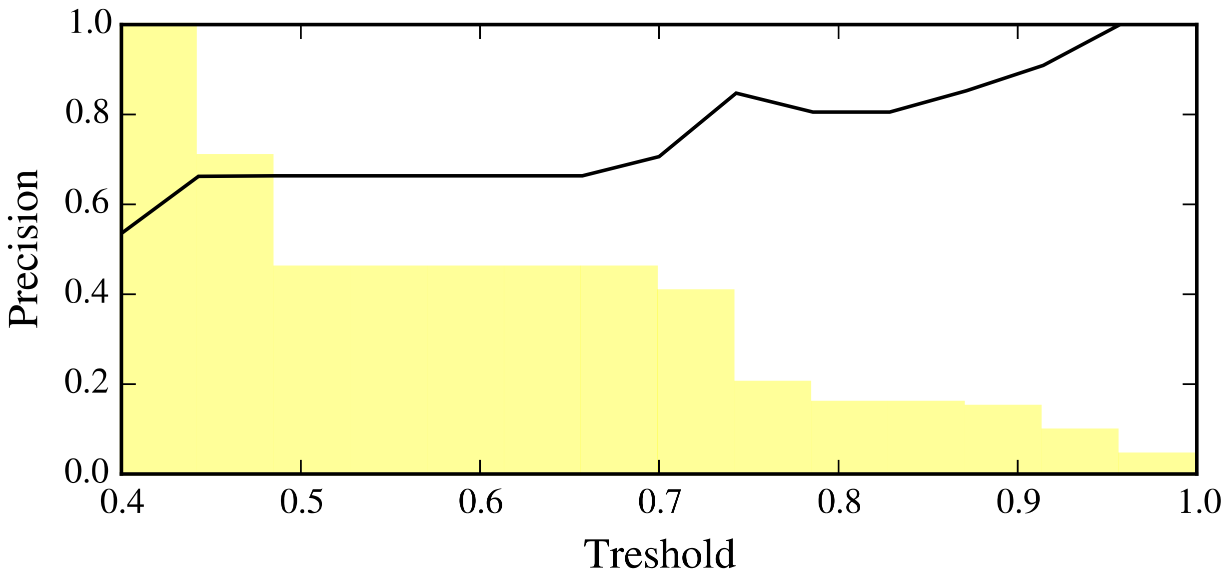
Best result is obtained with:

- Weak association
- LSA via SVD with k = 350 + K-Means with K = 9750



Evaluation & Results

- Results: bitly.com/1fWIbO2
- Evaluation:
  - draw 100 relations at random
  - verify if they are correct or not manually



What is it?

$$E = mc^2$$

Energy

Expected Value

Elimination matrix

	E	m	c	λ	σ	μ
Linear algebra	matrix	matrix	scalar	eigenvalue	related permutation	algebraic multiplicity
General relativity	energy	mass	speed of light	length	shear	reduced mass
Coding theory	encoding function	message	transmitted codeword		natural isomorphisms	
Optics		order fringe	speed of light in vacuum	wavelength	conductivity	permeability
Probability	expectation	sample size		affine parameter	variance	mean vector

http://www.freeflagicons.com

Experiments

- Available on Github:
  - github.com/alexeygrigorev/namespacediscovery
- Software used for experiments
  - Apache Flink 0.8.1
  - numpy 1.9.2, scipy 0.15.1, scikit-learn 0.16.1
  - IPython notebook 3.1.0
- Hardware used for experiments:

Manufacturer:	Samsung Electronics
Rating:	4,6 Windows Experience Index
Processor:	Intel(R) Pentium(R) CPU B950 @ 2.10GHz 2.10 GHz
Installed memory (RAM):	8,00 GB
System type:	64-bit Operating System

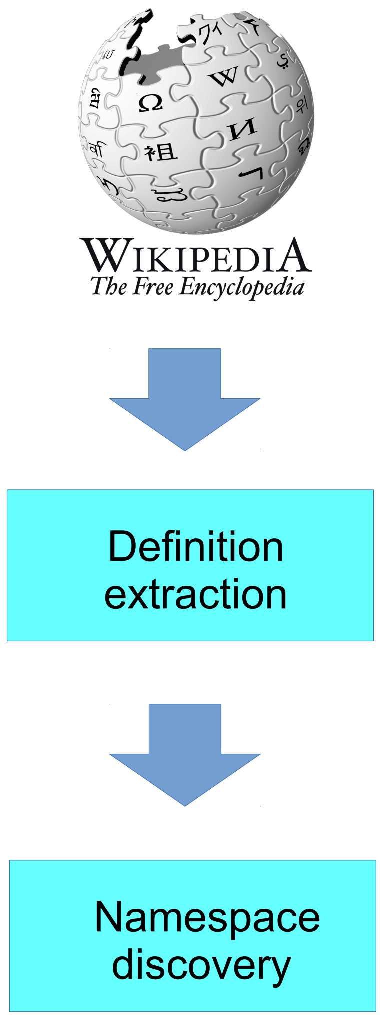
Outline

1. Motivation
2. Namespace Discovery
3. Implementation
4. Evaluation
5. **Conclusions**

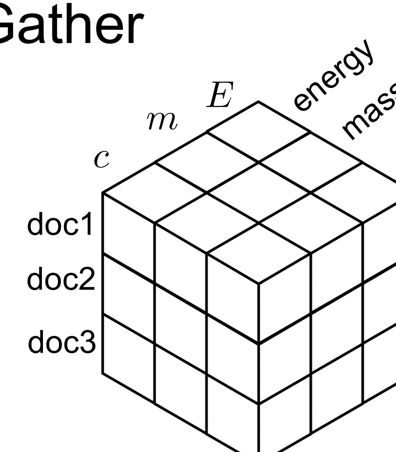
Conclusions

- We are the very first to approach the problem of namespace discovery
- Automatic namespace discovery is possible
- We can employ established methods such as VSM and Document Clustering
- Best result: 414 namespaces, 10 times better than random guessing
- Suitable for other natural languages, besides English

Future Work



- Other datasets:
  - arXiv
  - StackExchange Q/A network: mathematics, cross-validated, physics, …
- ML methods for identifier extraction may give better results
- Other ways to embed definitions: 3-D tensors
- Expect advanced clustering algorithms to perform better
  - Split and Join operations in Scatter/Gather
  - Spectral Clustering
  - Cluster Ensembles
  - Topic Modeling: LDA



Acknowledgments

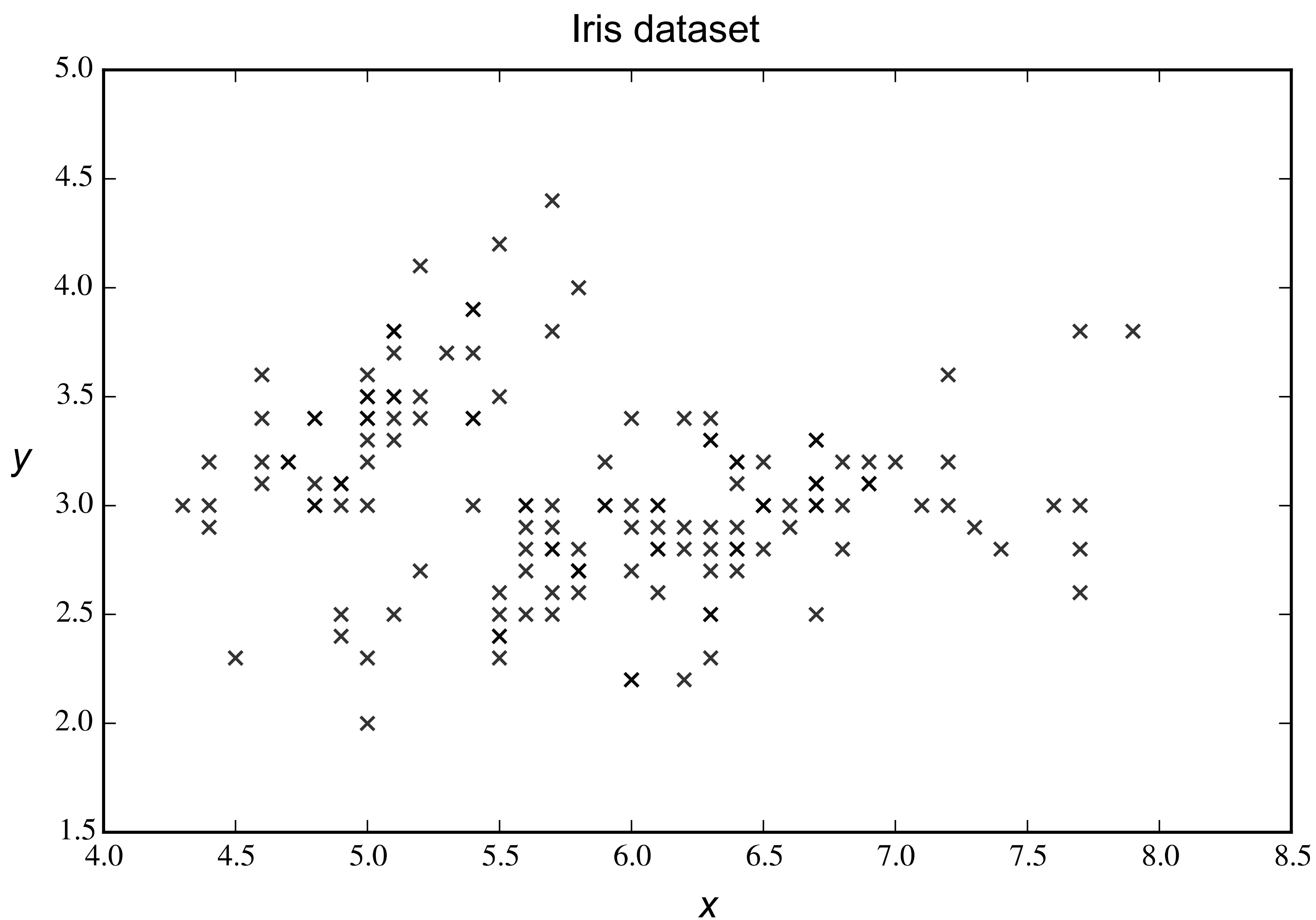
- My adviser Moritz Schubotz
- Sergey Dudoladov and Juan Soto
- All IT4BI teachers from ULB, UFRT, TUB
  - especially teachers of IR and DM courses

References

- [MLP] Pagel, Robert, and Schubotz, Moritz. "Mathematical Language Processing Project.", 2014.
- [IR] Manning, Christopher et al. “Introduction to Information Retrieval”, 2008.
- [SSN] Ertöz, Levent, et al. "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data.", 2003.
- [LSI] Deerwester, Scott, et al. "Indexing by Latent Semantic Analysis.", 1990.

Questions?

Back-up slide: Clustering Algorithms



Back-up slide: LSA

- Natural language data is “noisy”
  - Synonymy: “graph” vs “chart”
  - Polysemy: “trunk” (part of elephant vs part of car)
- Denoise with dimensionality reduction
  - SVD: $D = U\Sigma V^T$ $\quad D \approx U_k \Sigma_k V_k^T$
  - NMF: $D = UV^T$ $\quad D \approx U_k V_k^T$
- Not only denoises but also reveals the latent structure in data