

# Algorithmen und Analyse auf bibliographischen Daten

peterr und Lusy

9. November 2011

# Eigenschaften des Datensatzes

- enthält ca. 706 000 Einträge
- mit 19 verschiedenen Themengebieten
- nur der Themenbereich Physik wird in Themengruppen unterteilt
- 11 Einträge ohne Informationen
- die Publikationen haben im Durchschnitt 1.3 und maximal 9 Themen

# Aufbau der Datensatz

## Header

```
<identifier>oai:arXiv.org:0704.0001</identifier>  
<timestamp>2007-07-24</timestamp>  
<setSpec>physics:physics</setSpec>  
<setSpec>math</setSpec>
```

## Metadaten

```
<dc:title>Titel des Papers</dc:title>  
<dc:creator>Author 1</dc:creator>  
<dc:creator>Author 2</dc:creator>  
<dc:subject>Physics - Optics</dc:subject>  
<dc:subject>Mathematics - Combinatorics</dc:subject>  
<dc:description>Description</dc:description>  
<dc:description>Comment</dc:description>  
<dc:date>2007-04-02</dc:date>  
<dc:date>2007-07-24</dc:date>  
<dc:type>text</dc:type>  
<dc:identifier>http://arxiv.org/abs/0704.0001</dc:identifier>  
<dc:identifier>Phys.Rev.D76:013009,2007</dc:identifier>
```

# Parsen der Daten

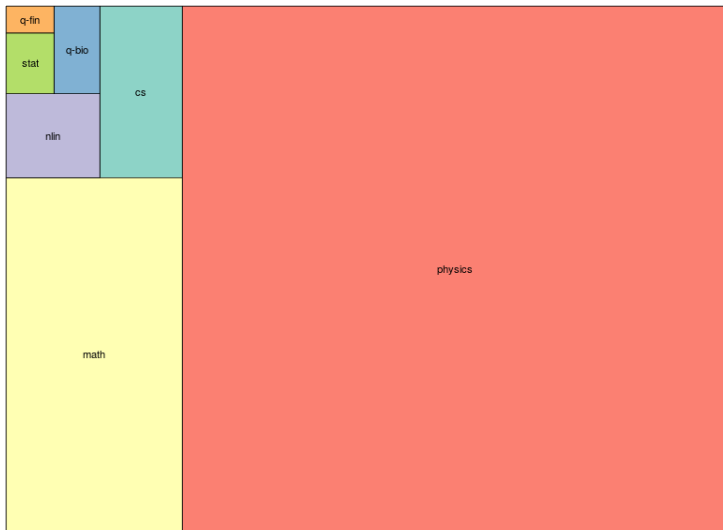
- Parser in Python geschrieben
- kompletter Datensatz in den Speicher
  - Overhead des XML-Parser nicht beachtet
- iterativer Ansatz <sup>1</sup>
- benötigt ca. 70 Sekunden für 1.2 GB

---

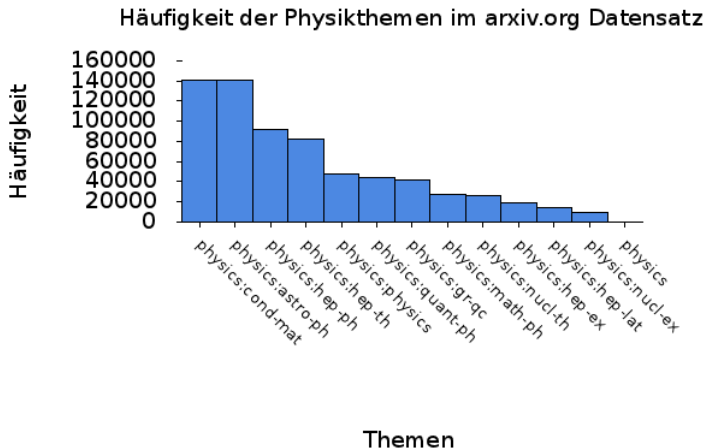
<sup>1</sup><http://www.ibm.com/developerworks/xml/library/x-hiperfparse/>

# Verteilung der Themen

Verteilung der Themen im arxiv.org Datensatz

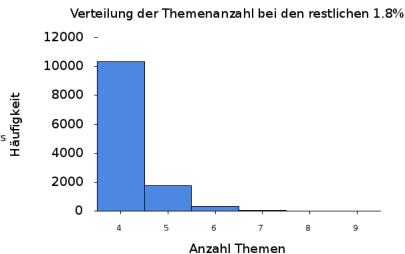
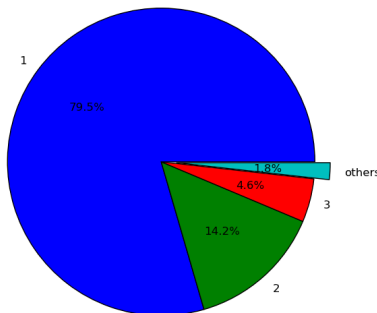


# Aufschlüsselung von physics



# Häufigkeit von Themen pro Publikation

Häufigkeit der Anzahl von Themen im arxiv.org Datensatz



# Was sind Assoziationsregeln?

- Beschreibung
- Regel der Form "Wenn Menge A, dann Menge B"
- Kenngrößen
  - Support - relative Häufigkeit der Regel in den Daten
  - Konfidenz -
  - Lift



# Assoziationsregeln - aller Themen

	Regel	Support	Konfidenz	Lift
	$\text{math} \implies \text{stat}$	0.6%	64%	3.0
	$\text{physics:math-ph} \implies \text{math}$	3.8 %	100%	4.7
	$\text{physics:hep-th, physics:math-ph} \implies \text{math}$	0.9 %	100%	4.7
	$\text{math, physics:hep-th} \implies \text{physics:math-ph}$	0.9 %	63%	16.3
	$\text{physics:gr-qc, physics:hep-th} \implies \text{physics:hep-th}$	0.6 %	72 %	6.1
	$\text{physics:gr-qc, physics:hep-th} \implies \text{physics:astro-ph}$	0.6 %	70 %	3.5
	$\text{physics:gr-qc, physics:astro-ph} \implies \text{physics:hep-th}$	0.9 %	50 %	4.3
	$\text{physics:astro-ph, physics:hep-th} \implies \text{physics:gr-qc}$	0.9 %	74 %	12.4

Support: 0.5 % und Konfidenz 50 %

# Assoziationsregeln - Oberthemen

Regel	Support	Konfidenz	Lift
$\emptyset \implies \text{physics}$	78%	78%	1.0
$\text{stat} \implies \text{math}$	0.6 %	63 %	3.0
$\text{nlin} \implies \text{physics}$	1.3 %	50 %	0.64
$\text{math, nlin} \implies \text{physics}$	0.4 %	83 %	1.1

Support: 0.1 % und Konfidenz 50 %

# Probleme

- mehrere Datumsangaben

# Weitere Analysen

- Aufschlüsselung der Themenbereiche
- Algorithmus implementieren?
  - AIS-Algorithmus
  - Apriori-Algorithmus
  - FPGrowth
- Entwicklung in Abhängigkeit von der Zeit

# Beschreibung

First Item Description of first item

Second Item Description of second item

Third Item Description of third item

Forth Item Description of forth item