Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Algorithms for Pattern Mining

Relim

- Midterm Presentation -

Thomas Stening
Thorsten Papenbrock

Agenda

Relim Algorithm

Performance and Result Analysis

Future Work

Agenda

Relim Algorithm

Performance and Result Analysis

Future Work

Relim Algorithm - Use Case

Medical Claim Analysis:

- 113000 transactions (= number of patients)
- 46 different item values (= different claims)
- 610934 items (= number of claims)
  → ~5.4 items / transaction

Questions:

- Which claims occure often together?
- Are there any claims that correlate each other?

Relim Algorithm - Use Case

Claim	Description
AMI	Acute myocardial infarction
APPCHOL	Appendicitis
ARTHSPIN	Arthropathies
CANCRA	Cancer A
CANCRB	Cancer B
CANCRM	Ovarian and metastatic cancer
CATAST	Catastrophic conditions
CHF	Congestive heart failure
COPD	Chronic obstructive pulmonary disorder
FLaELEC	Fluid and electrolyte
FXDISLC	Fractures and dislocations
GIBLEED	Gastrointestinal bleeding
GIOBSENT	Gastr. Inflam. bowel disease and obstruction
GYNEC1	Gynecology
GYNECA	Gynecologic cancers
HEART2	Other cardiac conditions
HEART4	Atherosclerosis and peripheral vascular disease
HEMTOL	Non-malignant hematologic
HIPFX	Hip fracture
INFEC4	All other infections
LIVERDZ	Liver disorders
METAB1	Diabetic ketoacidosis and related metabolic

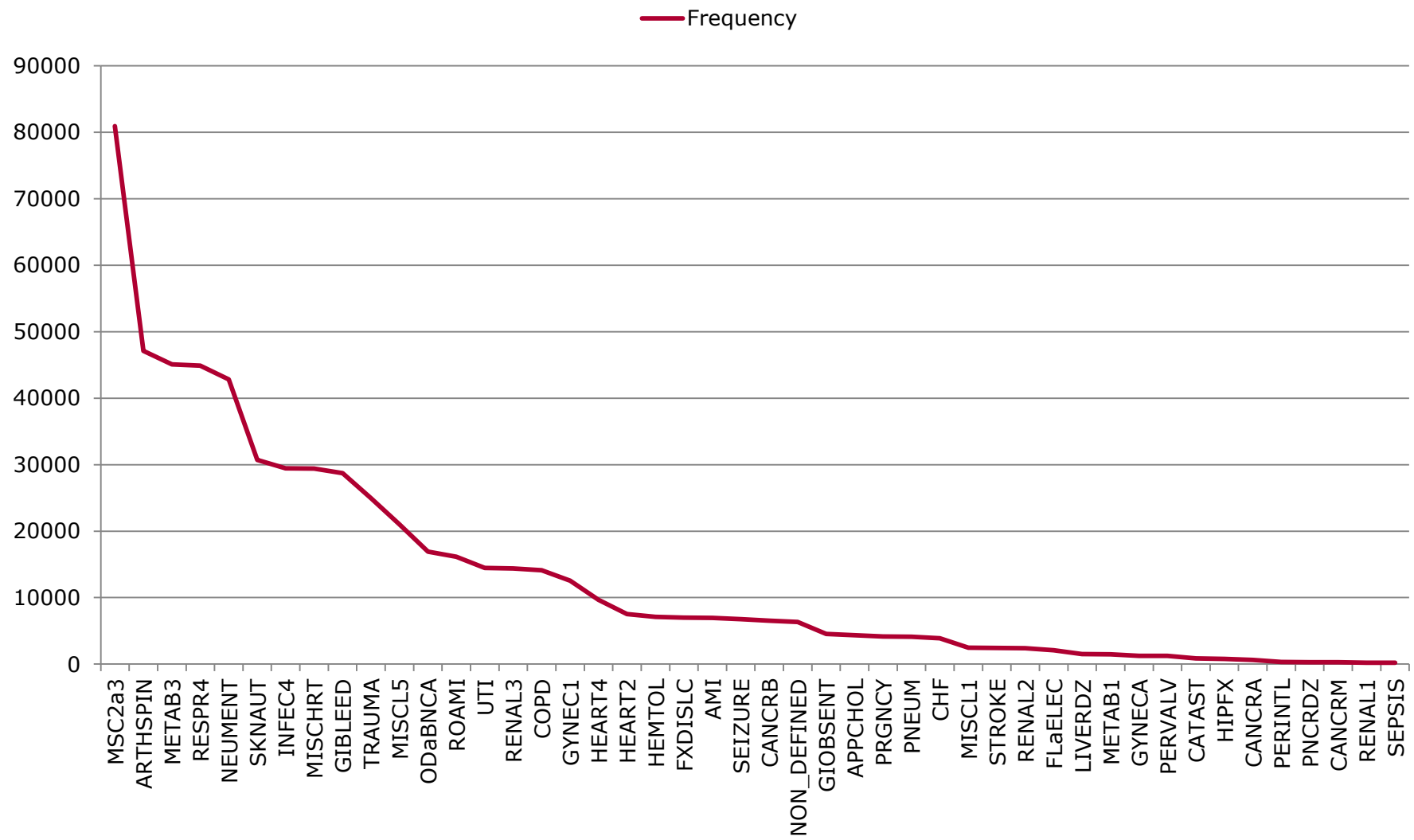
Claim	Description
METAB3	Other metabolic
MISCHRT	Miscellaneous cardiac
MISCL1	Miscellaneous 1
MISCL5	Miscellaneous 3
MSC2a3	Miscellaneous 2
NEUMENT	Other neurological
ODaBNCA	Ingestions and benign tumors
PERINTL	Perinatal period
PERVALV	Pericarditis
PNCRDZ	Pancreatic disorders
PNEUM	Pneumonia
PRGNCY	Pregnancy
RENAL1	Acute renal failure
RENAL2	Chronic renal failure
RENAL3	Other renal
RESPR4	Acute respiratory
ROAMI	Chest pain
SEIZURE	Seizure
SEPSIS	Sepsis
SKNAUT	Skin and autoimmune disorders
STROKE	Stroke
TRAUMA	All other trauma
UTI	Urinary tract infections

Relim Algorithm - Use Case

Claim	Description
AMI	Acute myocardial infarction
APPCHOL	Appendicitis
ARTHSPIN	Arthropathies
CANCRA	Cancer A
CANCRB	Cancer B
CANCRM	Ovarian and metastatic cancer
CATAST	Catastrophic conditions
CHF	Congestive heart failure
COPD	Chronic obstructive pulmonary disorder
FLaELEC	Fluid and electrolyte
FXDISLC	Fractures and dislocations
GIBLEED	Gastrointestinal bleeding
GIOBSENT	Gastr. Inflam. bowel disease and obstruction
GYNEC1	Gynecology
GYNECA	Gynecologic cancers
HEART2	Other cardiac conditions
HEART4	Atherosclerosis and peripheral vascular disease
HEMTOL	Non-malignant hematologic
HIPFX	Hip fracture
INFEC4	All other infections
LIVERDZ	Liver disorders
METAB1	Diabetic ketoacidosis and related metabolic

Claim	Description
METAB3	Other metabolic
MISCHRT	Miscellaneous cardiac
MISCL1	Miscellaneous 1
MISCL5	Miscellaneous 3
MSC2a3	Miscellaneous 2
NEUMENT	Other neurological
ODaBNCA	Ingestions and benign tumors
PERINTL	Perinatal period
PERVALV	Pericarditis
PNCRDZ	Pancreatic disorders
PNEUM	Pneumonia
PRGNCY	Pregnancy
RENAL1	Acute renal failure
RENAL2	Chronic renal failure
RENAL3	Other renal
RESPR4	Acute respiratory
ROAMI	Chest pain
SEIZURE	Seizure
SEPSIS	Sepsis
SKNAUT	Skin and autoimmune disorders
STROKE	Stroke
TRAUMA	All other trauma
UTI	Urinary tract infections

Relim Algorithm - Use Case



Relim Algorithm - Motivation

- **Aim:**
  Test and optimize an alternative algorithm to todays most common algorithms

- **Comparison:**

Name	Simplicity	Performance
Apriori	Easier	Slower
Eclat	Harder	Similar (*)
FP-Growth	Harder / Similar	Faster

Relim Algorithm – Datastructure Generation

1. Load transactions (in memory)

a d f
c d e
b d
a b c d
b c
a b d
b d e
b c e g
c d f
a b d

Relim Algorithm – Datastructure Generation

1. Load transactions (in memory)
2. Count item frequencies

} 1. Iteration

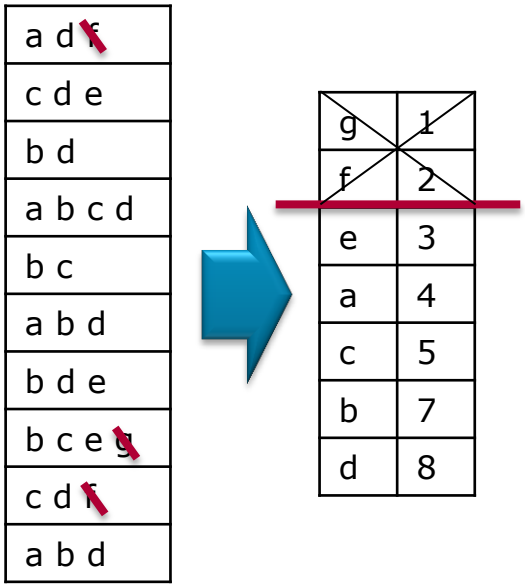
a d f
c d e
b d
a b c d
b c
a b d
b d e
b c e g
c d f
a b d

g	1
f	2
e	3
a	4
c	5
b	7
d	8

Relim Algorithm – Datastructure Generation

1. Load transactions (in memory)
2. Count item frequencies
3. Delete all rare items from the transactions

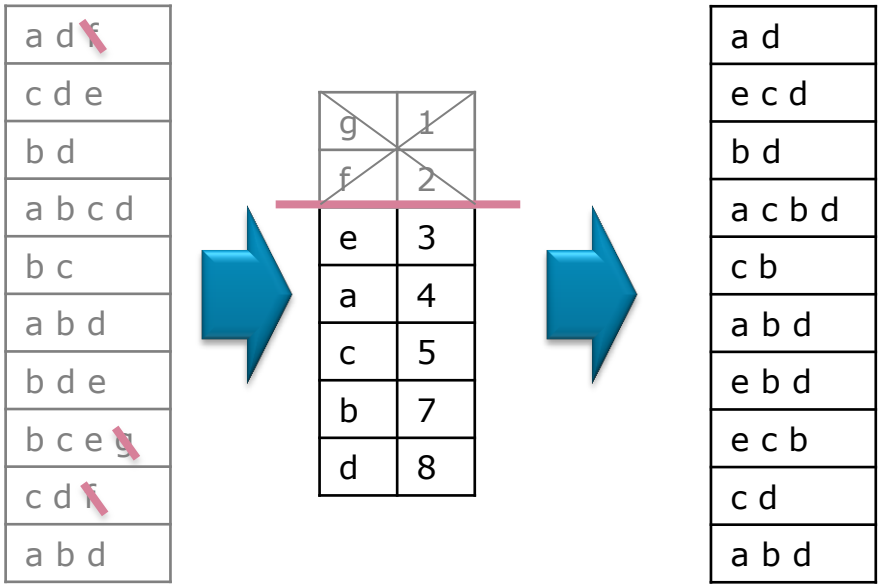
1. Iteration



Relim Algorithm – Datastructure Generation

1. Load transactions (in memory)
2. Count item frequencies
3. Delete all rare items from the transactions
4. Sort each transaction according the items frequency

1. Iteration

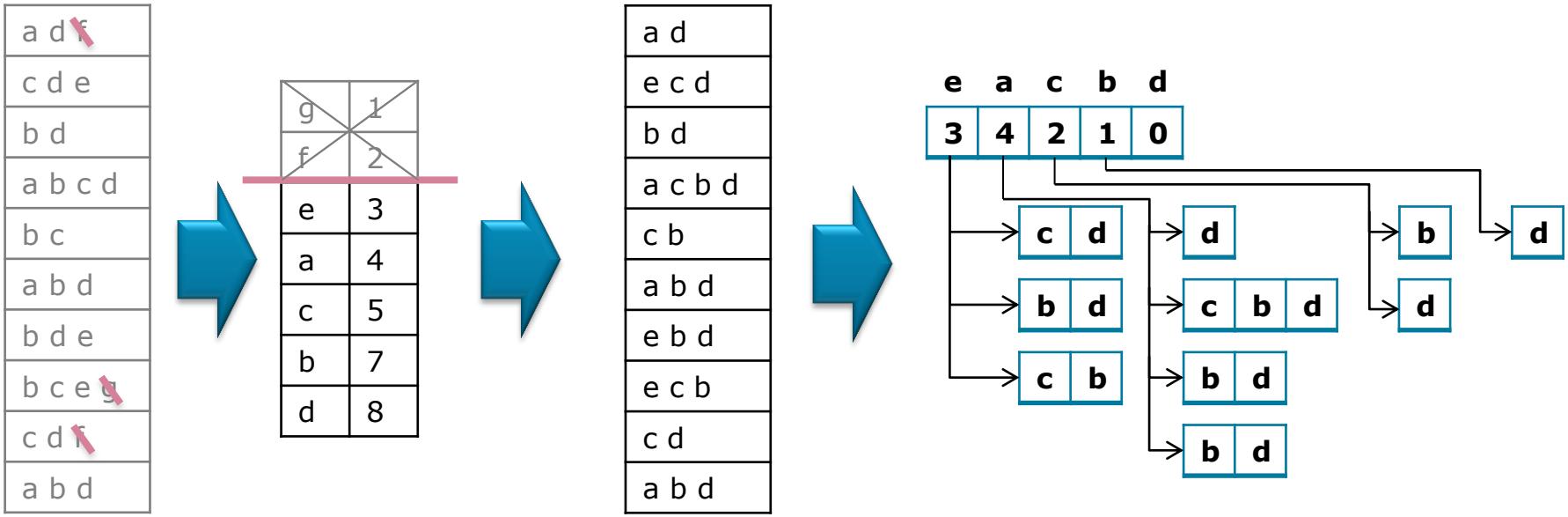


Relim Algorithm – Datastructure Generation

1. Load transactions (in memory)
2. Count item frequencies
3. Delete all rare items from the transactions
4. Sort each transaction according the items frequency
5. Create Relim datastructure

1. Iteration

2. Iteration

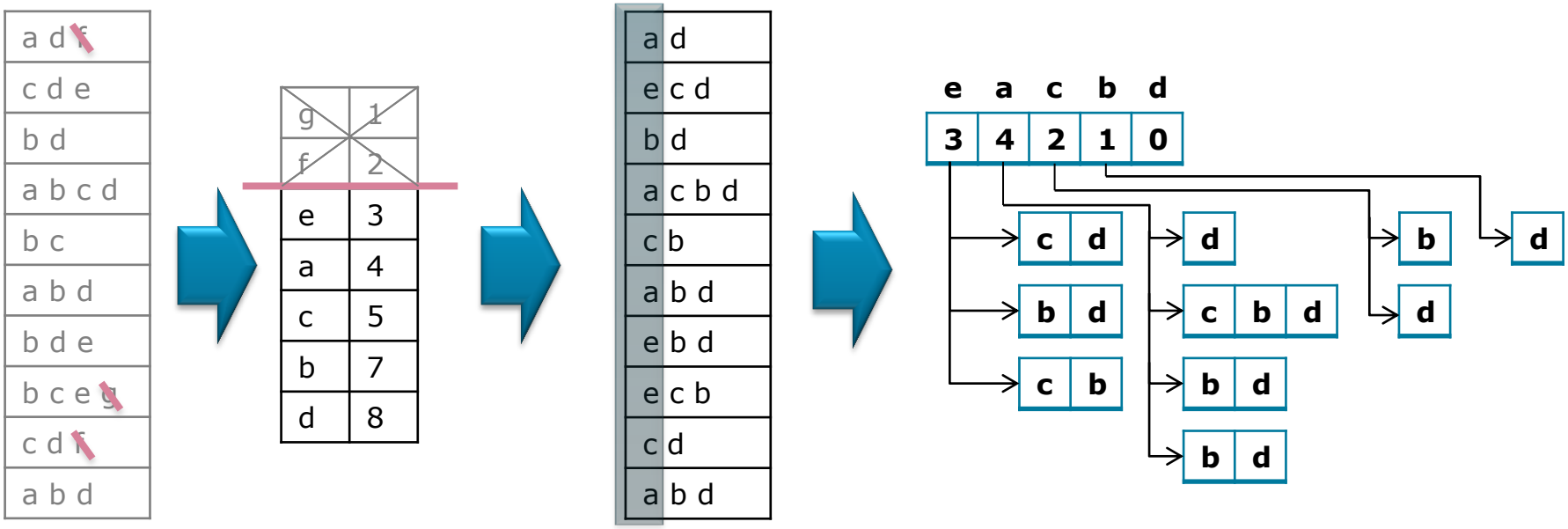


Relim Algorithm – Datastructure Generation

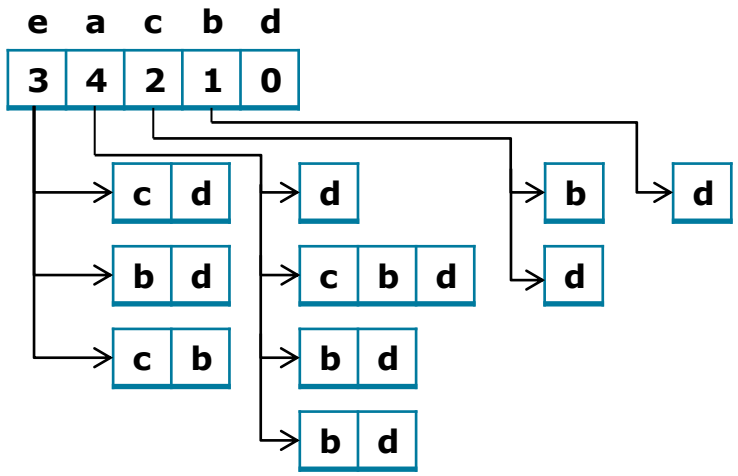
1. Load transactions (in memory)
2. Count item frequencies
3. Delete all rare items from the transactions
4. Sort each transaction according the items frequency
5. Create Relim datastructure

1. Iteration

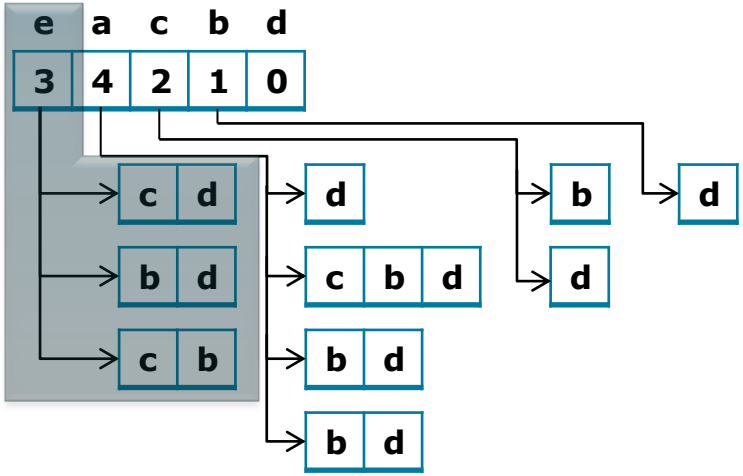
2. Iteration



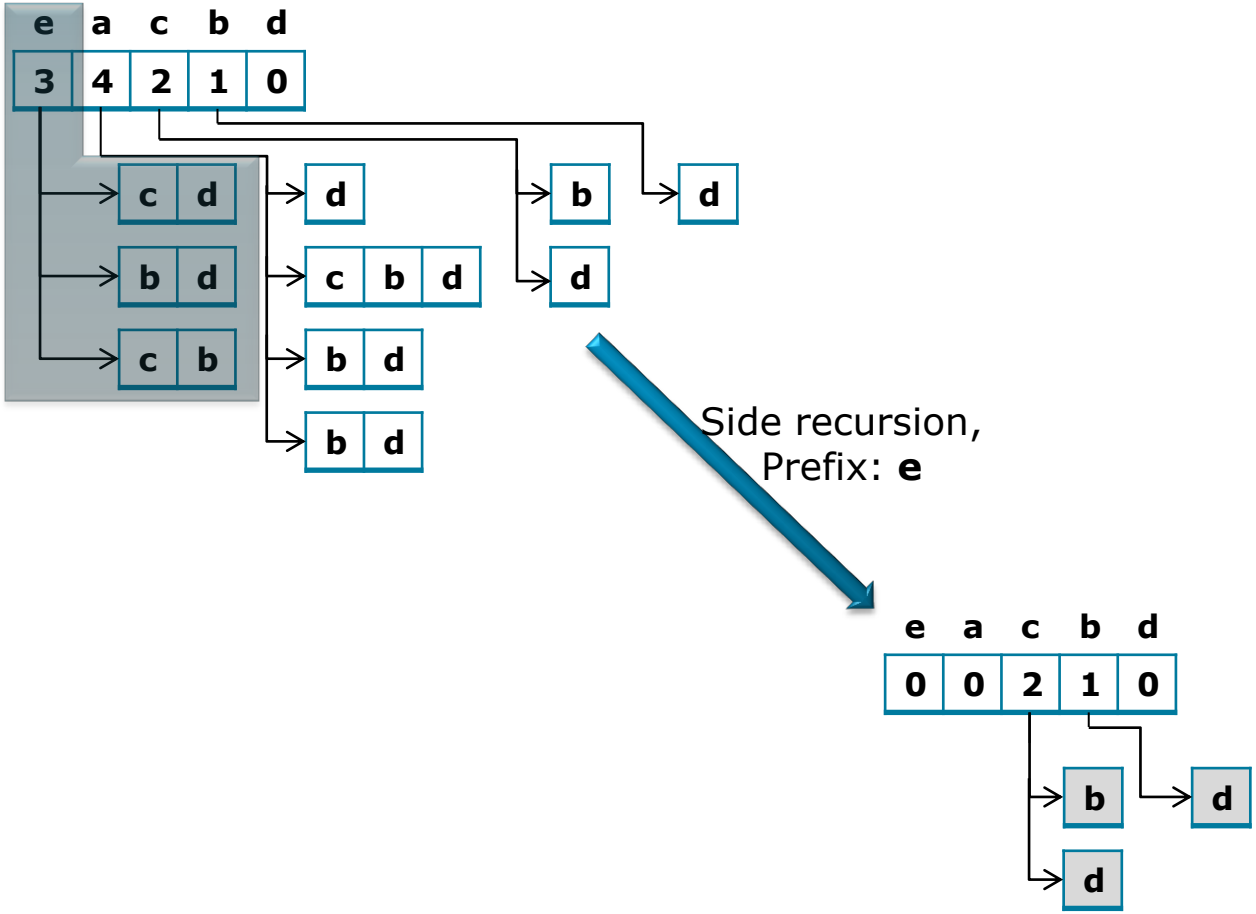
Relim Algorithm – Recursive Tree Processing



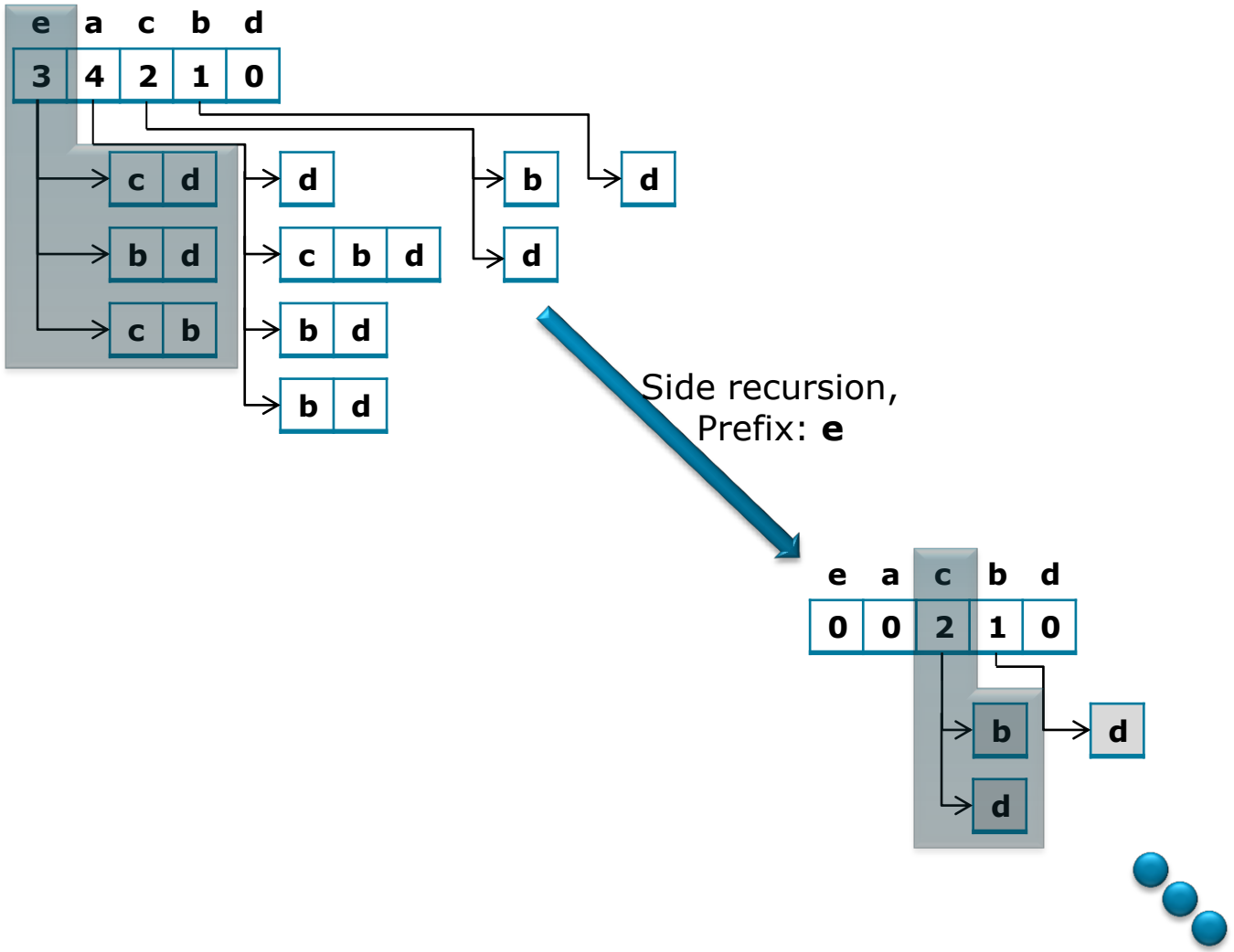
Relim Algorithm – Recursive Tree Processing



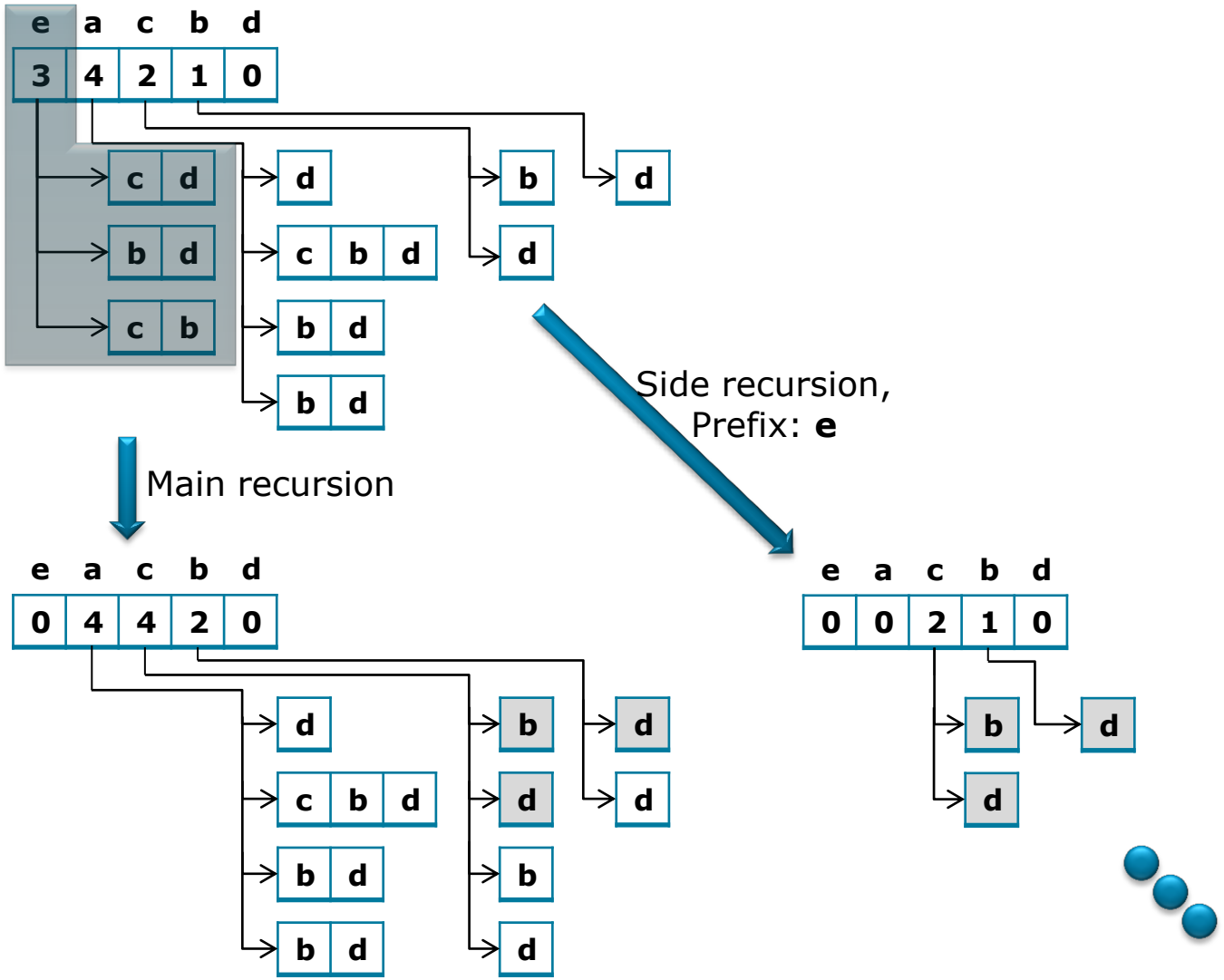
Relim Algorithm – Recursive Tree Processing



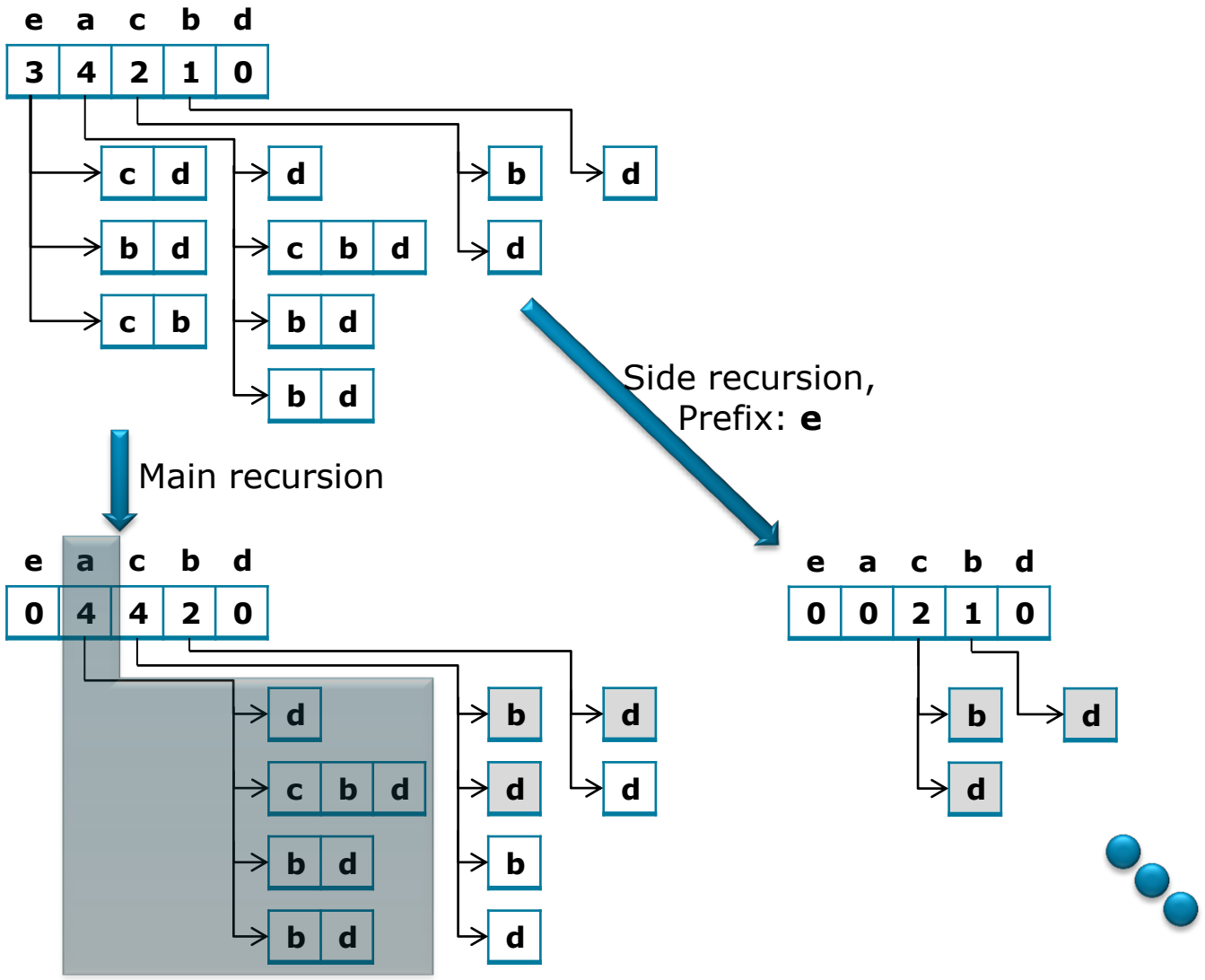
Relim Algorithm – Recursive Tree Processing



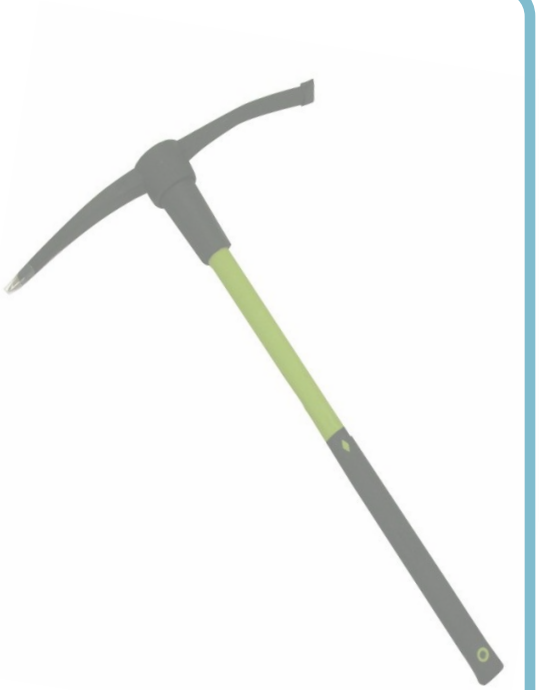
Relim Algorithm – Recursive Tree Processing



Relim Algorithm – Recursive Tree Processing



Agenda



Relim Algorithm

Performance and Result Analysis

Future Work

Performance and Result Analysis – Setup

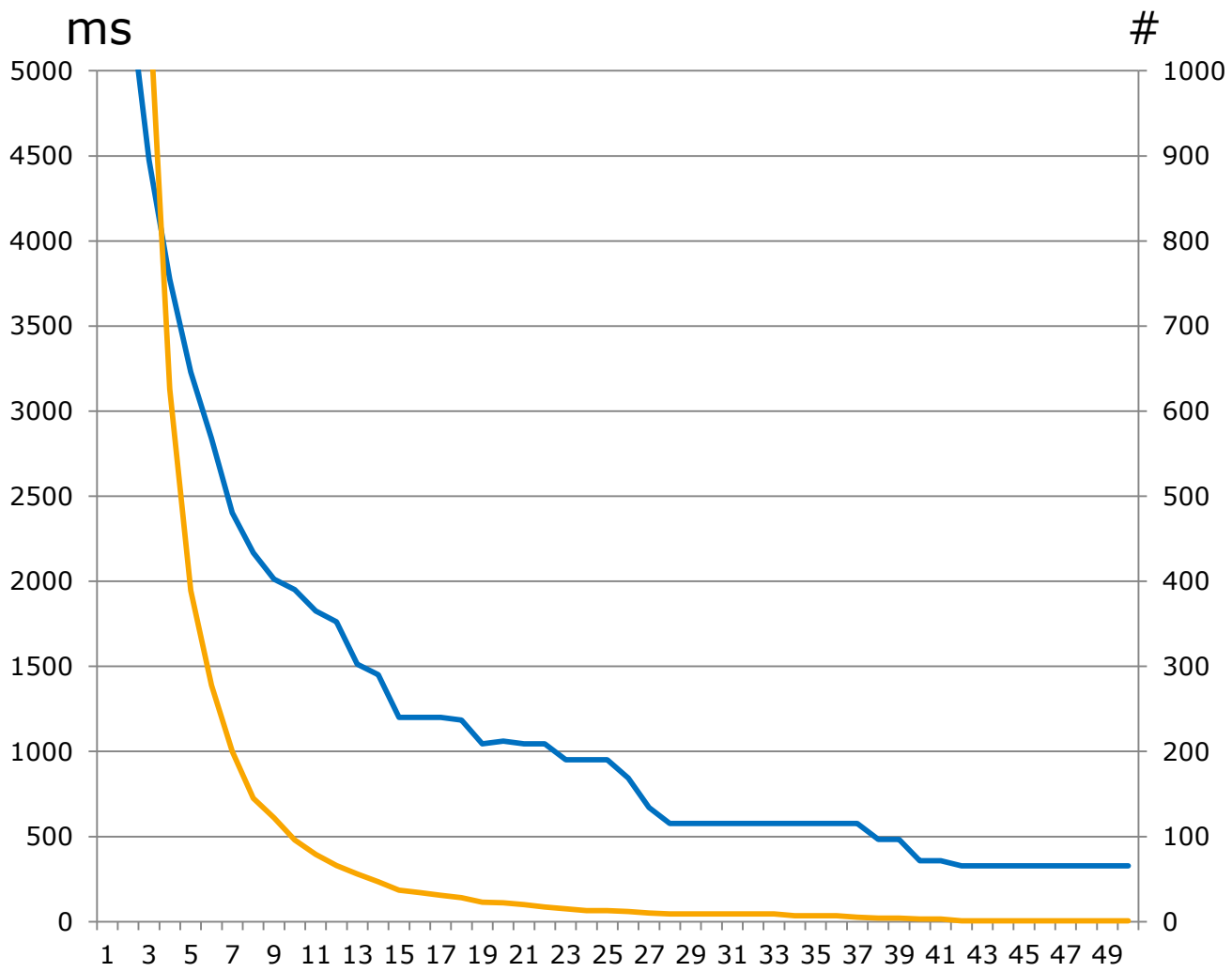
System:

- DELL StudioXPS
- 64 bit Windows 7 Enterprise
- Intel Core i5 M520 2,40 GHz
- 4 GB RAM

Implementierung:

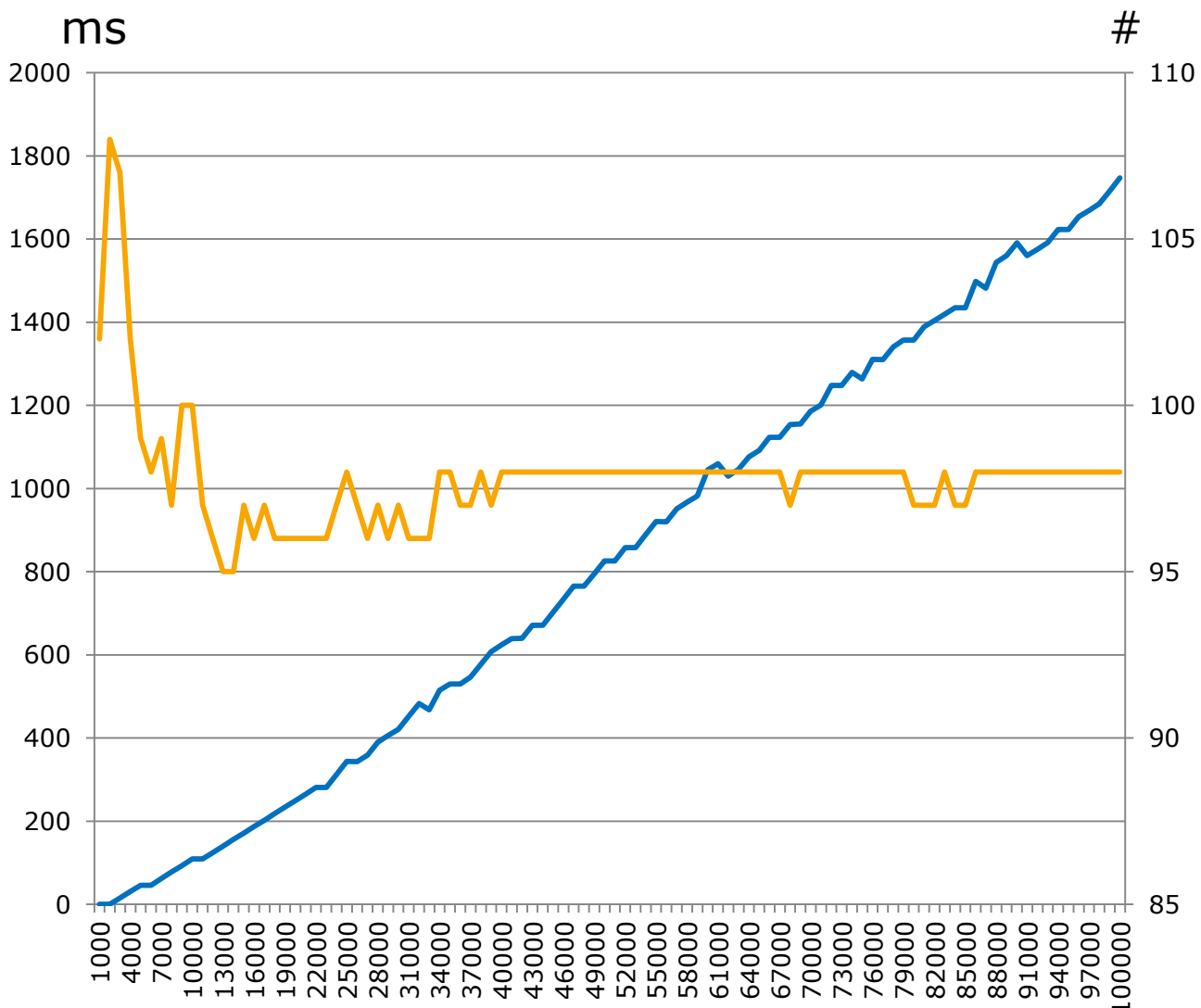
- Java
- Basis-Algorithmus ohne algorithmische Optimierungen

Performance and Result Analysis – Scaling MinSupport



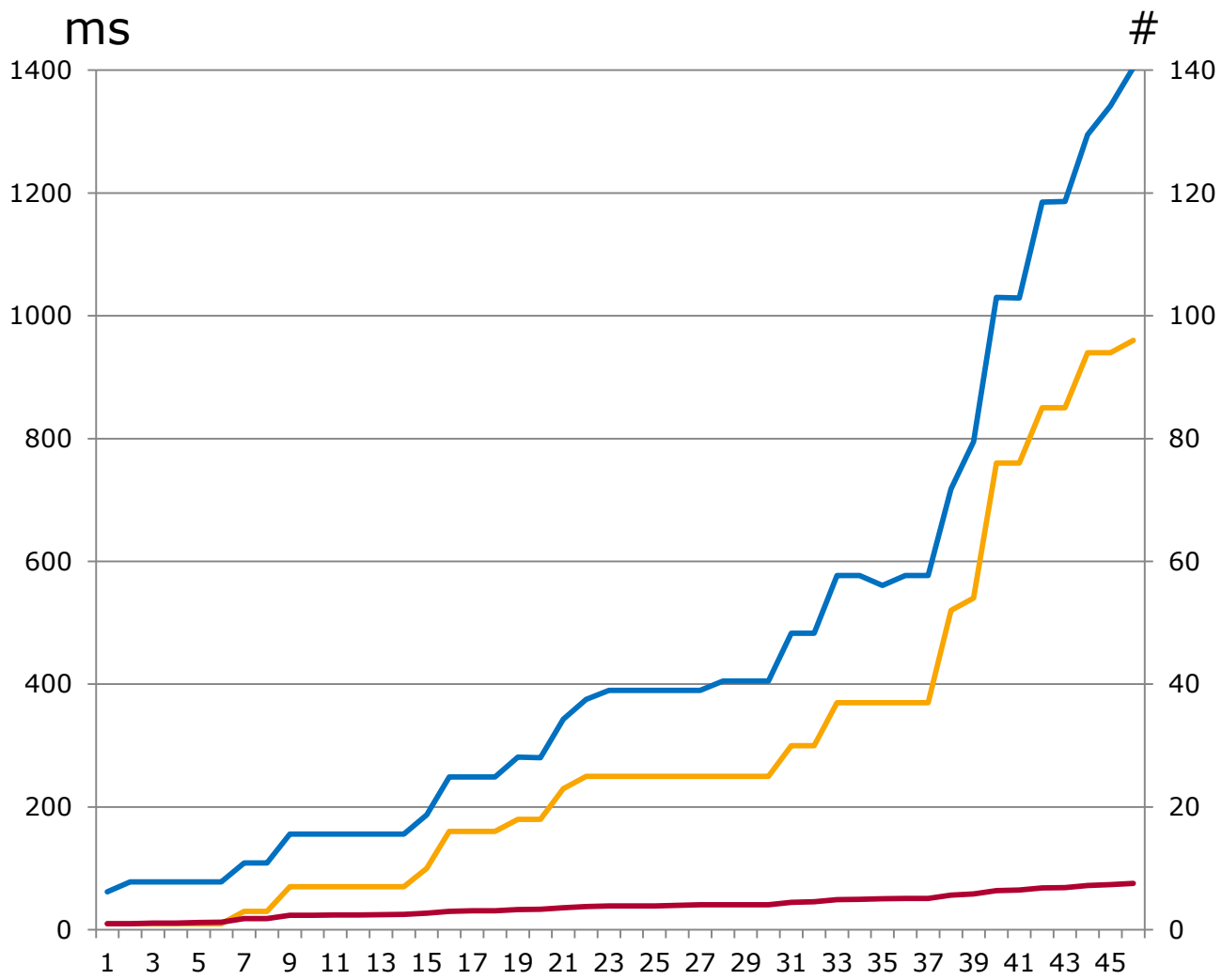
Transactions	113000 #
DifferentItems	46 #
TransactionSize	Ø 5.4 #
MinSupport	- %

Performance and Result Analysis – Scaling Transactions



Transactions	- #
DifferentItems	46 #
TransactionSize	Ø 5.4 #
MinSupport	10 %

Performance and Result Analysis – Scaling DifferentItems



Transactions	80944 #
DifferentItems	- #
TransactionSize	Ø - #
MinSupport	10 %

Performance and Result Analysis – Example Results

[MISCHRT , METAB3 , ARTHSPIN] – 13188 #

- Miscellaneous cardiac → Herzerkrankung
- Other metabolic → Stoffwechselprobleme
- Arthropathies → Gelenkerkrankungen

[RESPR4 , METAB3 , ARTHSPIN] – 13123 #

- Acute respiratory → Atembeschwerden
- Other metabolic → Stoffwechselprobleme
- Arthropathies → Gelenkerkrankungen

[TRAUMA , NEUMENT] – 11826 #

- All other trauma → Trauma
- Other neurological → Nervensystemerkrankungen

Agenda

Relim Algorithm

Performance and Result Analysis

Future Work

Future Work

1. **Fuzzy Datamining:**
   - Some claims might not have been detected or reported, but they are still very frequent
   - Try to find Frequent Item Sets although data is incomplete

2. **Use Case Specialization:**
   - Finding rules in the data
   - Considering the time between claims might deliver better results for the Frequent Claim Sets

3. **Parallelisation:**
   - Transforming recursion steps into thread branches might increase the performance

Sources

Paper: (Links Stand 10.05.12)

- "Keeping Things Simple: Finding Frequent Item Sets by Recursive Elimination", Christian Borgelt, http://www.borgelt.net/relim.html
- "Mining Fuzzy Frequent Item Sets", Xiaomeng Wang, Christian Borgelt, and Rudolf Kruse, http://www.borgelt.net/relix.html

Bilder: (Links Stand 10.05.12)

- http://regrounding.files.wordpress.com/2011/07/doctor-talking-to-patient.jpg
- http://searchtrafficpro.com/wp-content/uploads/2010/02/mining-the-search-query-report.jpg#
- http://www.bogensportwelt.de/bilder/produkte/gross/523208_Fernglas-ZEISS-Conquest-8-x-30-T.jpg
- http://www.wittewarenhandel.de/bilder/156227.jpg