Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Divide & Conquer-based Inclusion Dependency Discovery

Thorsten Papenbrock, Sebastian Kruse, Jorge-Arnulfo Quianè-Ruiz, Felix Naumann

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Motivation

Unary Inclusion Dependencies (INDs)

Book

Title	Author	Price	Pages	Published
Database Systems	Ullman	214	1203	2007
Algorithms in Java	Sedgewick	130	768	2002
3D Computer Graphics	Watt	20	570	1999

Name ⊆ Title

Lending

ID	Name	Location	Student	Course
42	Database Systems	A-1.2	Miller	DBS 1
88	Database Systems	B-2.2	Miller	PT 1
73	Database Systems	A-1.2	Smith	DPDC
69	Algorithms in Java	C-E.1	Miller	PT 1
13	Algorithms in Java	C-E.1	Smith	DPDC

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Motivation

N-ary Inclusion Dependencies (INDs)

Student

Name	Lecture	Credit	Semester	Verified
Miller	DBS 1	20	2	false
Miller	PT 1	15	2	false
Smith	DPDC	10	6	true

Student, Course ⊆ Name, Lecture

Lending

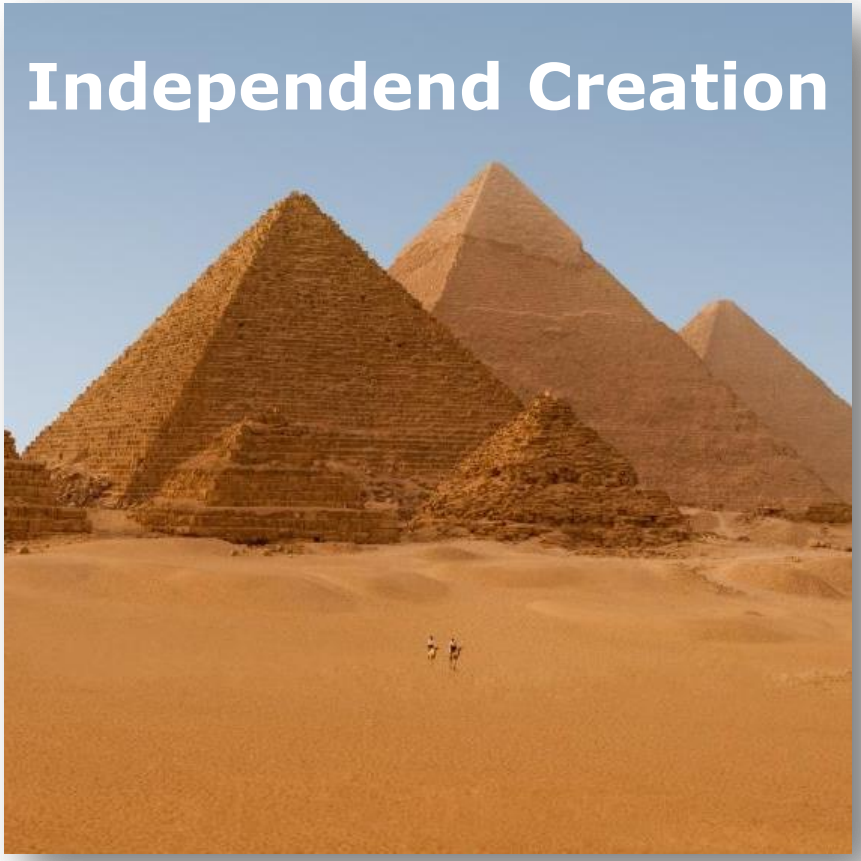
ID	Name	Location	Student	Course
42	Database Systems	A-1.2	Miller	DBS 1
88	Database Systems	B-2.2	Miller	PT 1
73	Database Systems	A-1.2	Smith	DPDC
69	Algorithms in Java	C-E.1	Miller	PT 1
13	Algorithms in Java	C-E.1	Smith	DPDC

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

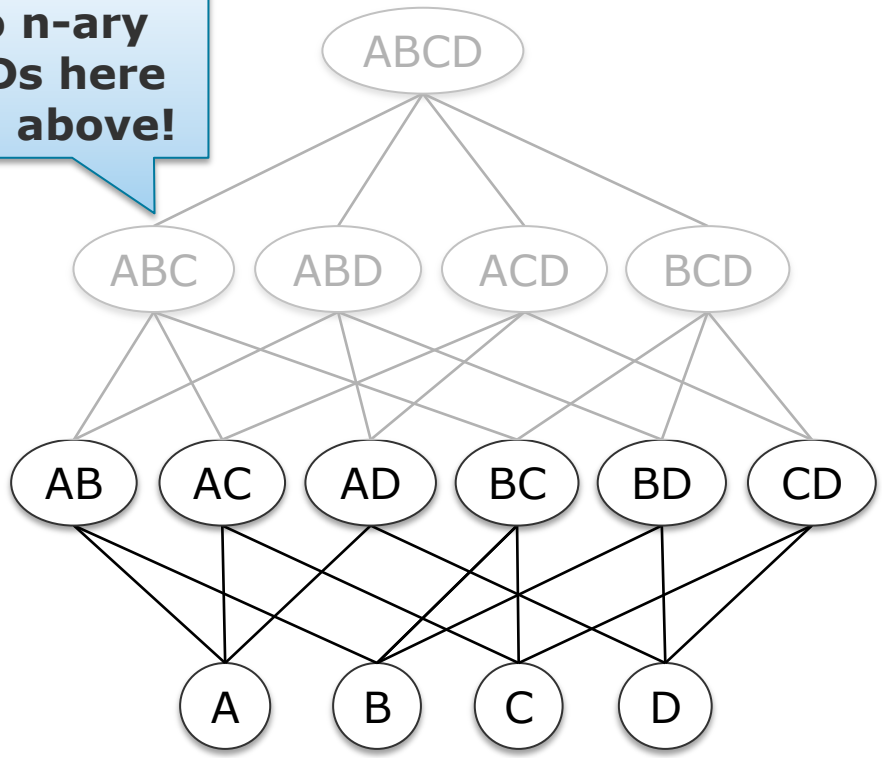
HPI Hasso Plattner Institut

Motivation
Necessity for IND Discovery



Motivation
IND Detection Search Space

No n-ary INDs here and above!



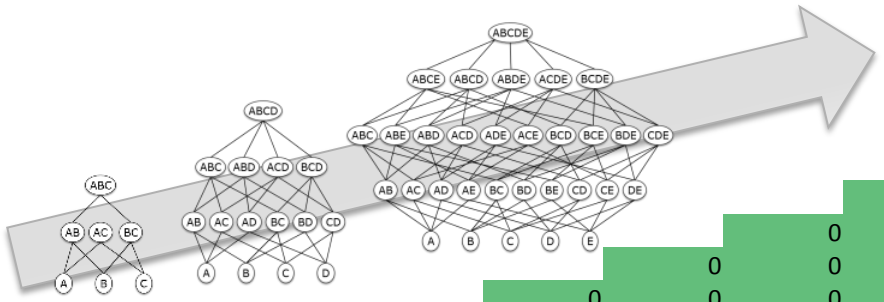
Test every combination with all other combinations of same size!

$AB \subseteq CD$ $AB \subseteq DC$
$AC \subseteq BD$ $AC \subseteq DB$
$AD \subseteq BC$ $AD \subseteq CB$
$BC \subseteq AD$ $BC \subseteq DA$
$BD \subseteq AC$ $BD \subseteq CA$
$CD \subseteq AB$ $CD \subseteq BA$

$A \subseteq B$ $A \subseteq C$ $A \subseteq D$
$B \subseteq A$ $B \subseteq C$ $B \subseteq D$
$C \subseteq A$ $C \subseteq B$ $C \subseteq D$
$D \subseteq A$ $D \subseteq B$ $D \subseteq C$

Motivation
IND Detection Search Space

IND candidates



Number of levels: k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	total
15															0	0
14														0	0	0
13													0	0	0	0
12												0	0	0	0	0
11											0	0	0	0	0	0
10										0	0	0	0	0	0	0
9									0	0	0	0	0	0	0	0
8								0	0	0	0	0	0	0	0	0
7							0	0	0	0	0	0	0	17297280	259459200	259476960
6						0	0	0	0	0	0	665280	8648640	60540480	302702400	372556800
5					0	0	0	0	0	30240	332640	1995840	8648640	30270240	90810720	132087120
4				0	0	0	0	1680	15120	75600	277200	831600	2162160	5045040	10810800	19221600
3			0	0	0	120	840	3360	10080	25200	55440	110880	205920	360360	600600	1372500
2		0	0	12	60	180	420	840	1512	2520	3960	5940	8580	12012	16380	53016
1	0	2	6	12	20	30	42	56	72	90	110	132	156	182	210	1120
total	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	

Number of attributes: n

Unary IND detection:
$O(n^2)$
for n attributes

N-ary IND detection:
$O(2^n \cdot n!)$
for n attributes

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation

HPI Hasso Plattner Institut

Motivation
Related Work

- **Bell**: S. Bell and P. Brockhausen. *Discovery of data dependencies in relational databases*. Technical report, Universität Dortmund, 1995
- **Koeller**: A. Koeller and E. A. Rundensteiner. *Discovery of high-dimensional inclusion dependencies*. In ICDE, 2002
- **Zigzag**: F. D. Marchi and J.-M. Petit. *Zigzag: A new algorithm for mining large inclusion dependencies in databases*. In ICDM, 2003
- **SPIDER**: J. Bauckmann, U. Leser, and F. Naumann. *Efficiently computing inclusion dependencies for schema discovery*. In ICDE Workshops, 2006 (best for **unary** INDs)
- **MIND**: F. D. Marchi, S. Lopes, and J.-M. Petit. *Unary and n-ary inclusion dependency discovery in relational databases*. JIIS, 32:53–73, 2009 (best for **n-ary** INDs)
- **Clim**: F. D. Marchi. *CLIM: Closed inclusion dependency mining in databases*. In ICDM Workshops, 2011

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery

N-ary IND Discovery

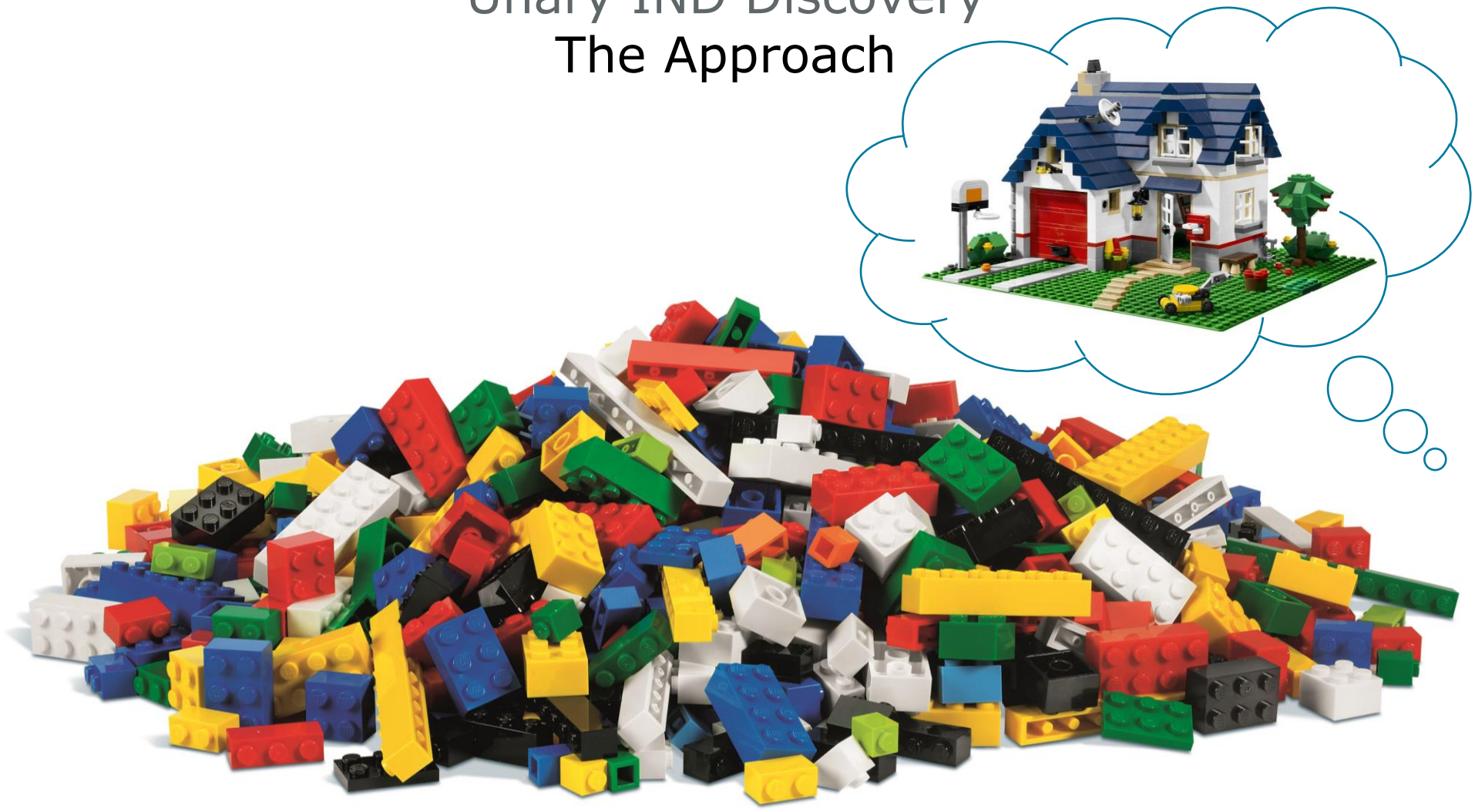
Experimental Evaluation

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery
The Approach



Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery
The Approach

No preparation

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

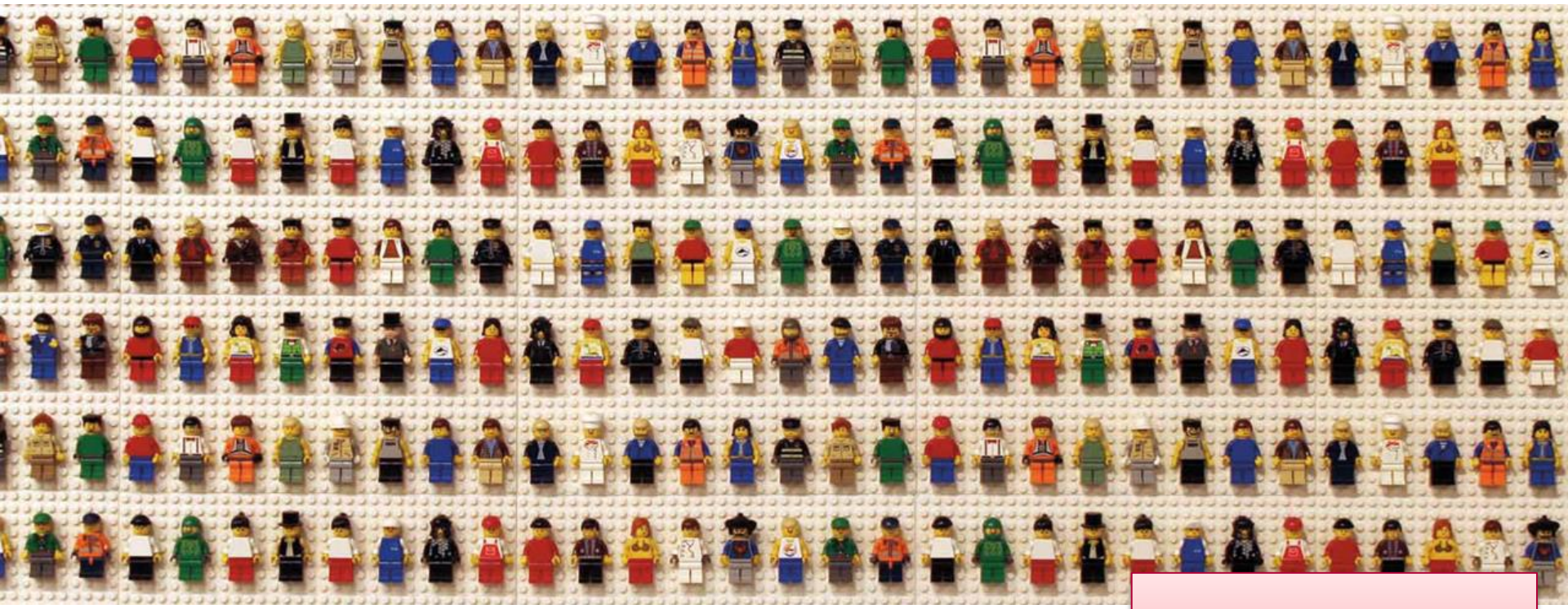
Unary IND Discovery
The Approach

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery
The Approach



Sorting

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery
The Approach

Bucketing

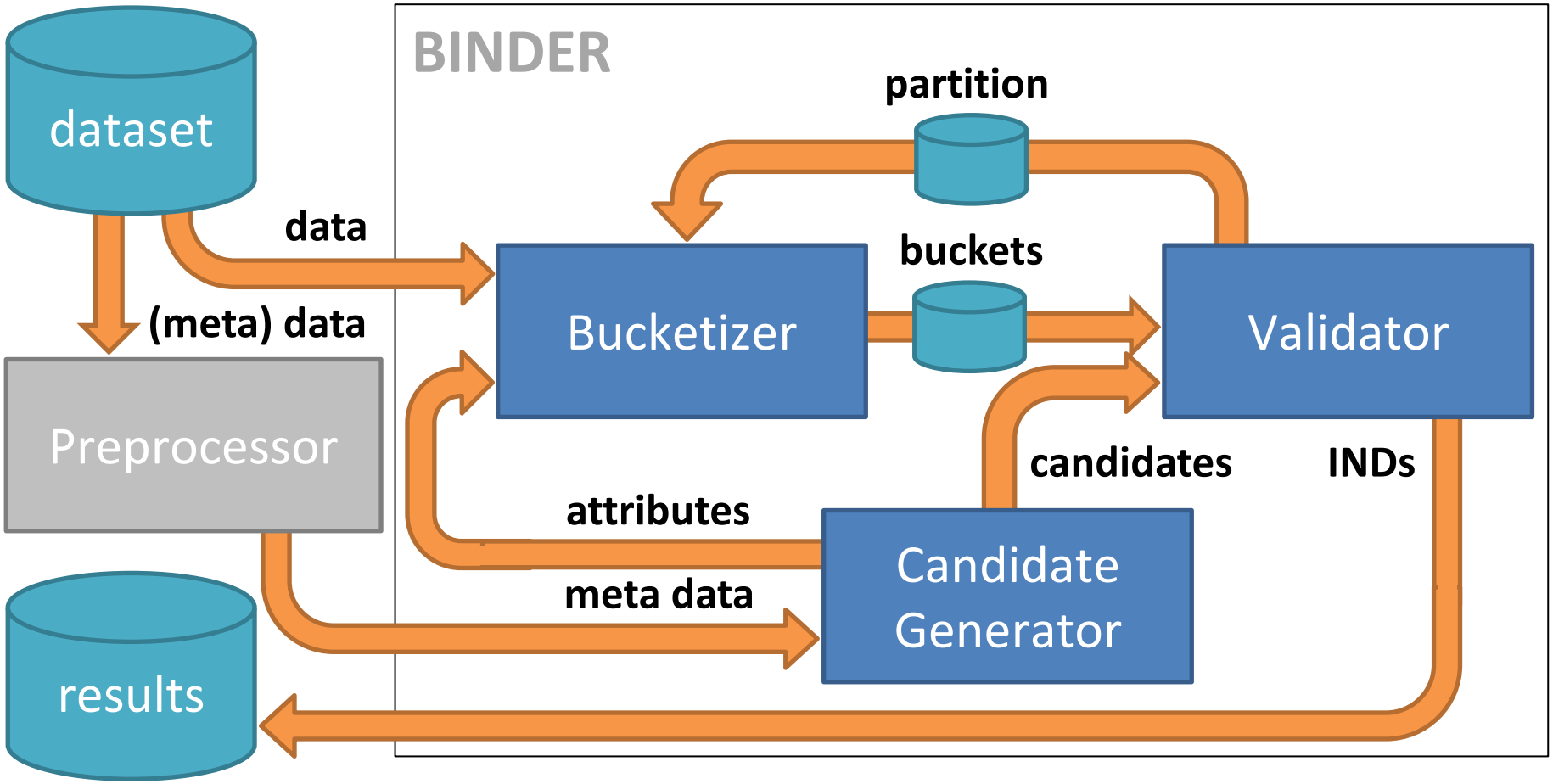
Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

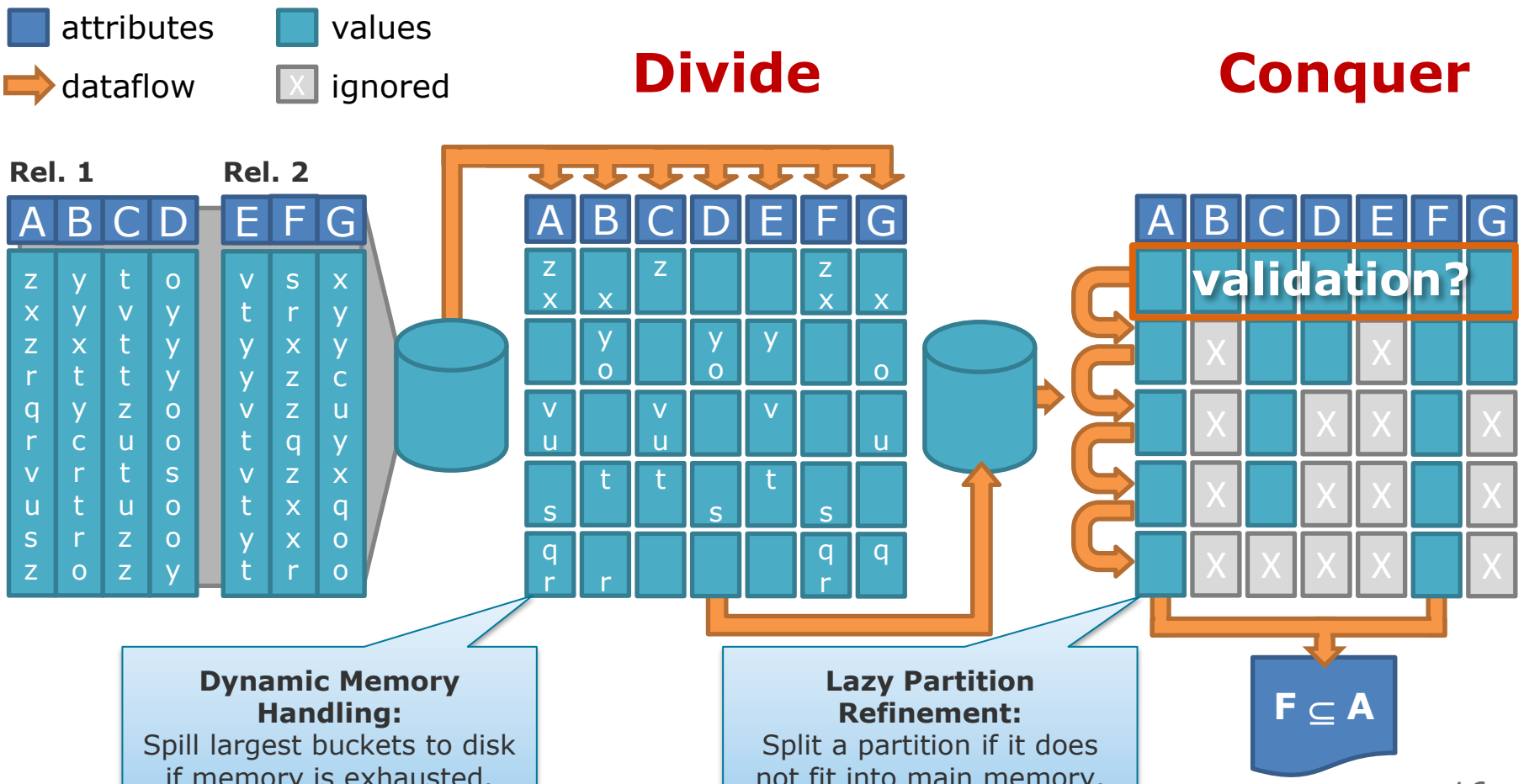
HPI Hasso Plattner Institut

Unary IND Discovery
The Approach

N-ary IND Discovery
The BINDER Algorithm

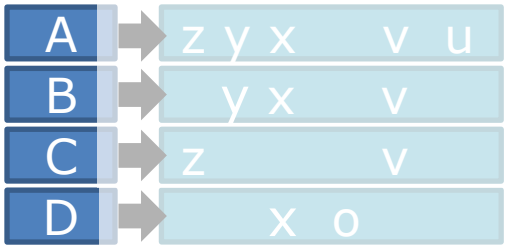


Unary IND Discovery
The BINDER Algorithm

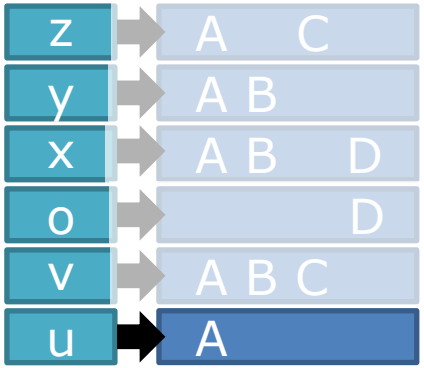


Unary IND Discovery
The BINDER Algorithm

attr2value

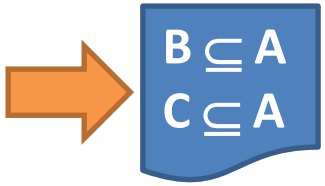


value2attr



Never tested! →

look up	A	B	C	D
	B,C,D	**A,C,D**	**A,B,D**	**A,B,C**
A → z → A,C	**C**	A,C,D	**A**	A,B,C
A → y → A,B	**-**	**A**	A	A,B,C
B → x → A,B,D	**-**	**A**	A	**A,B**
B → v → A,B,C	**-**	**A**	**A**	A,B
D → o → D	-	A	A	-

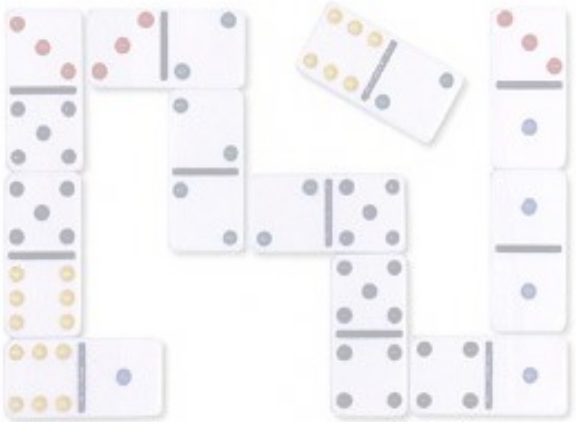


Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

Unary IND Discovery



N-ary IND Discovery

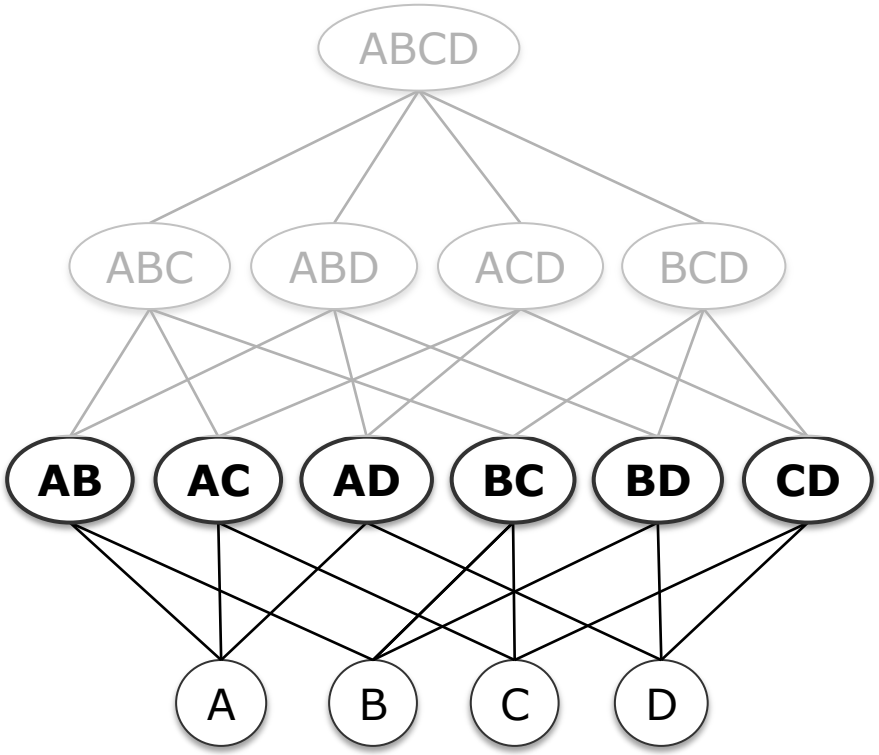
Experimental Evaluation

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

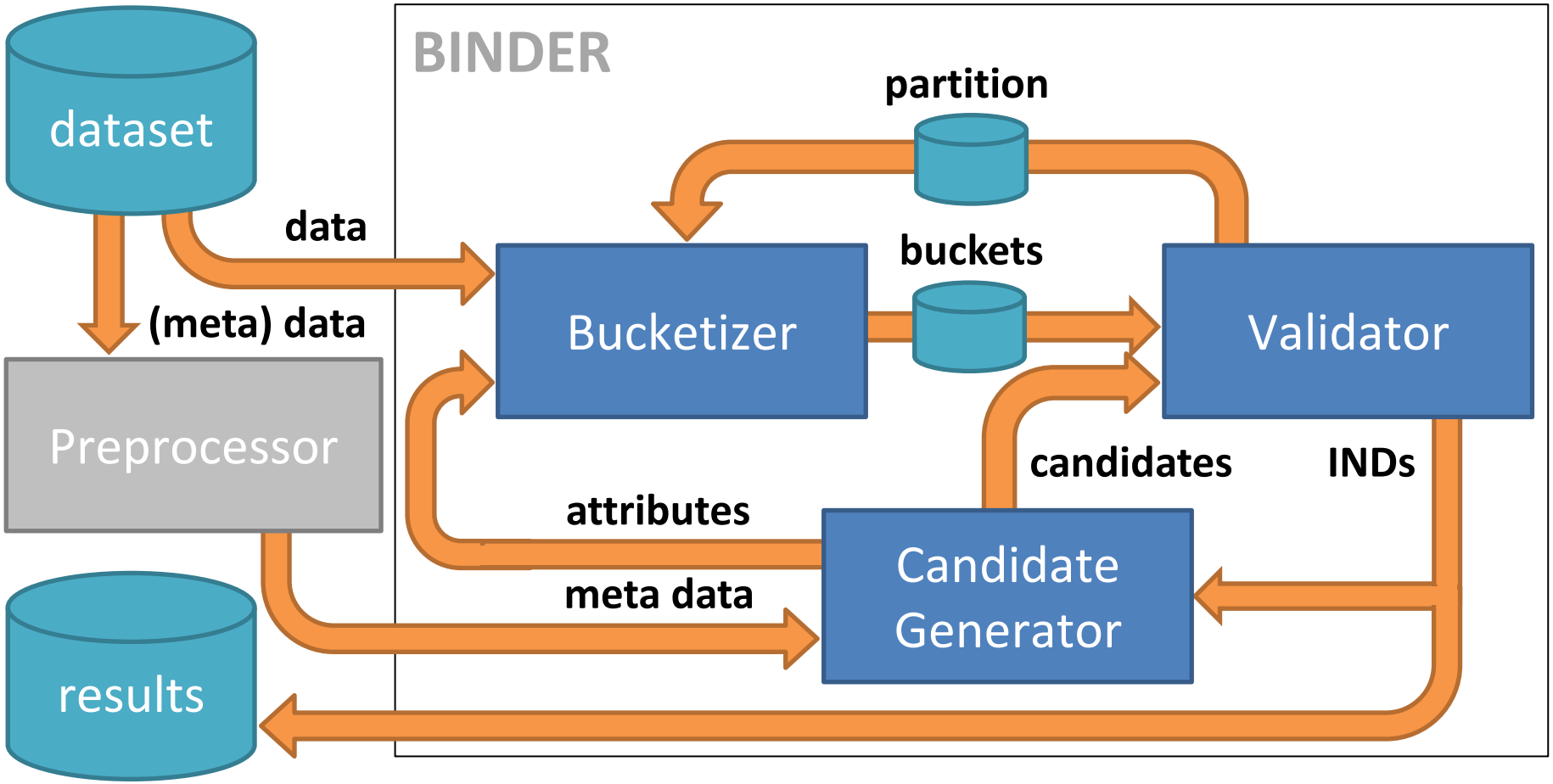
N-ary IND Discovery
The BINDER Algorithm



AB ⊆ CD **AB ⊆ DC**
AC ⊆ BD **AC ⊆ DB**
AD ⊆ BC **AD ⊆ CB**
BC ⊆ AD **BC ⊆ DA**
BD ⊆ AC **BD ⊆ CA**
CD ⊆ AB **CD ⊆ BA**

A ⊆ B A ⊆ C A ⊆ D
B ⊆ A B ⊆ C B ⊆ D
C ⊆ A C ⊆ B C ⊆ D
D ⊆ A D ⊆ B D ⊆ C

N-ary IND Discovery
The BINDER Algorithm



N-ary IND Discovery
The BINDER Algorithm

New Candidate?

1. Both valid:

$A \subseteq C$

$B \subseteq D$

2. Non-trivial:

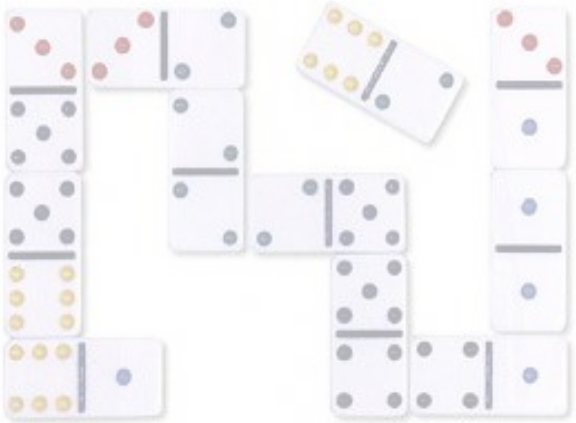
$A \neq B \neq C \neq D$

Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut

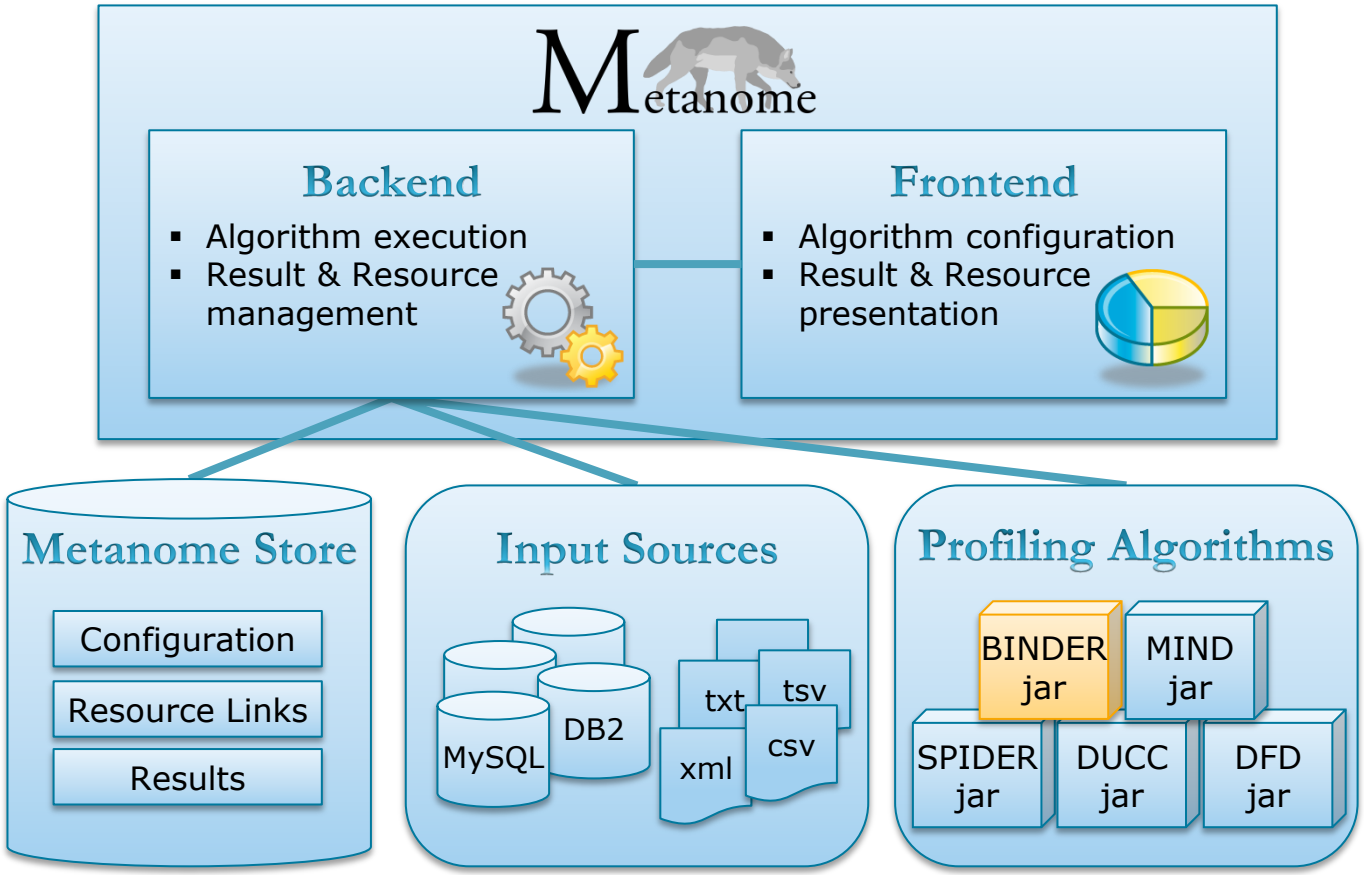
Unary IND Discovery



N-ary IND Discovery

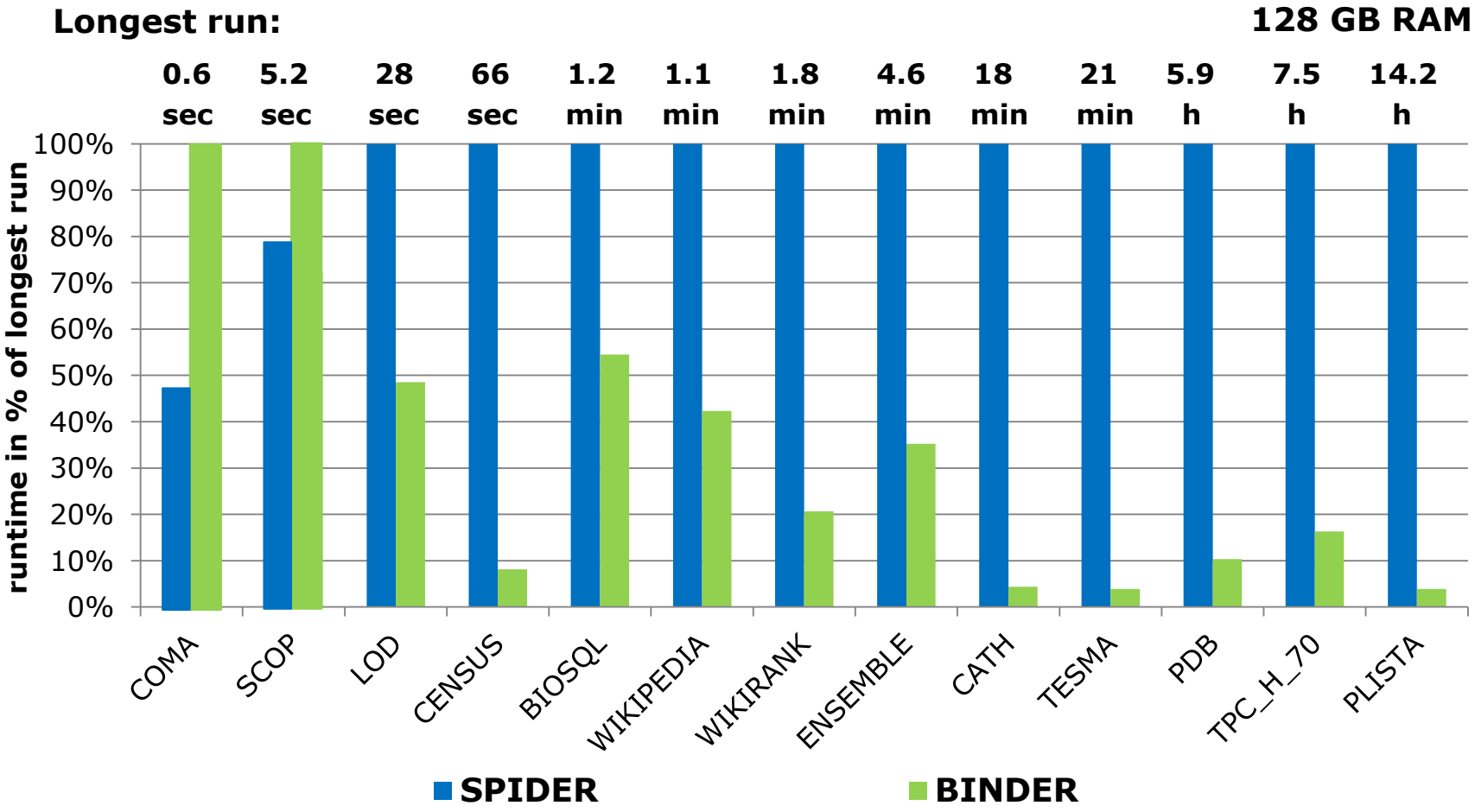
Experimental Evaluation

Experimental Evaluation
The Metanome Profiling Tool



www.metanome.de

Experimental Evaluation
Unary IND Discovery

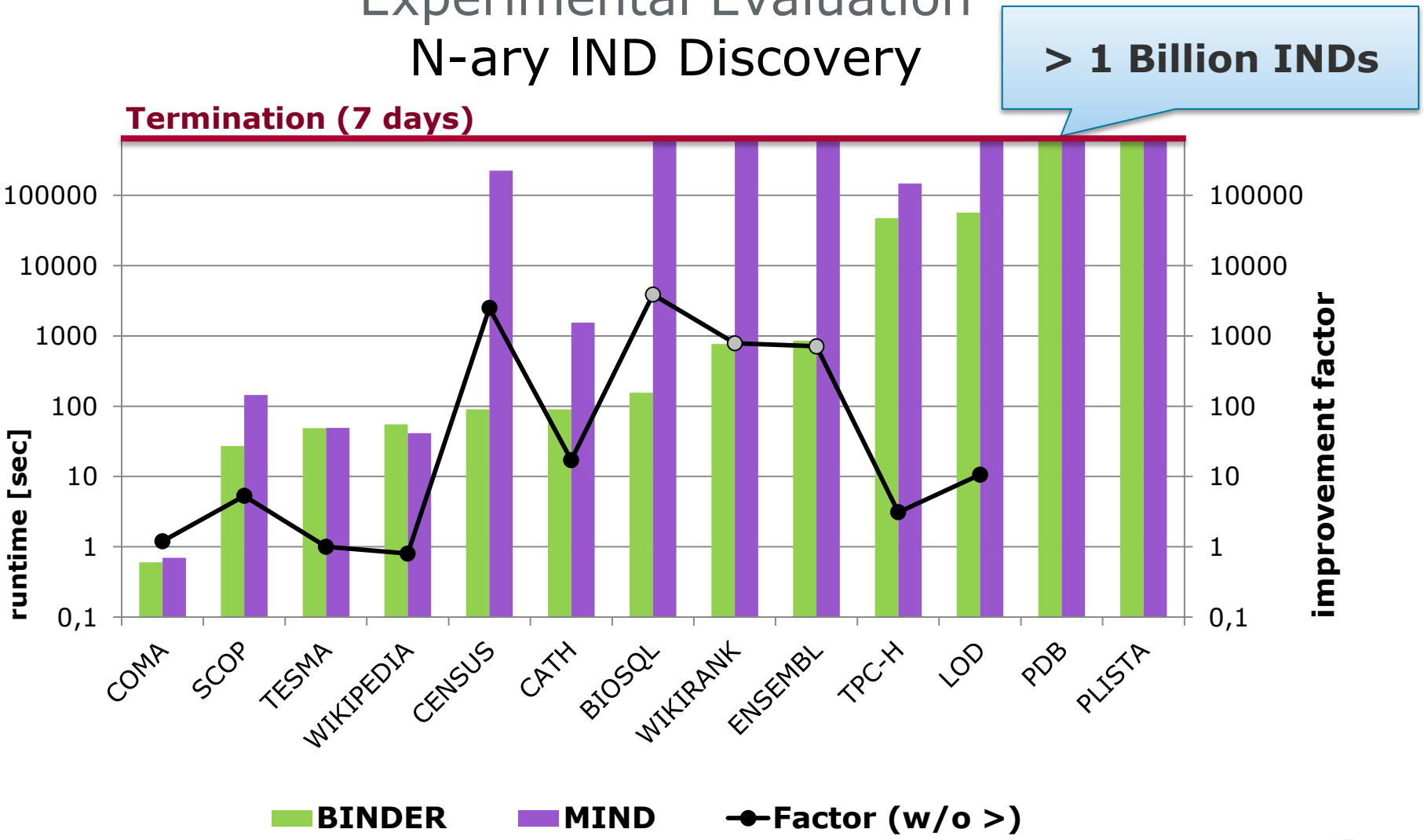


Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation

HPI Hasso Plattner Institut

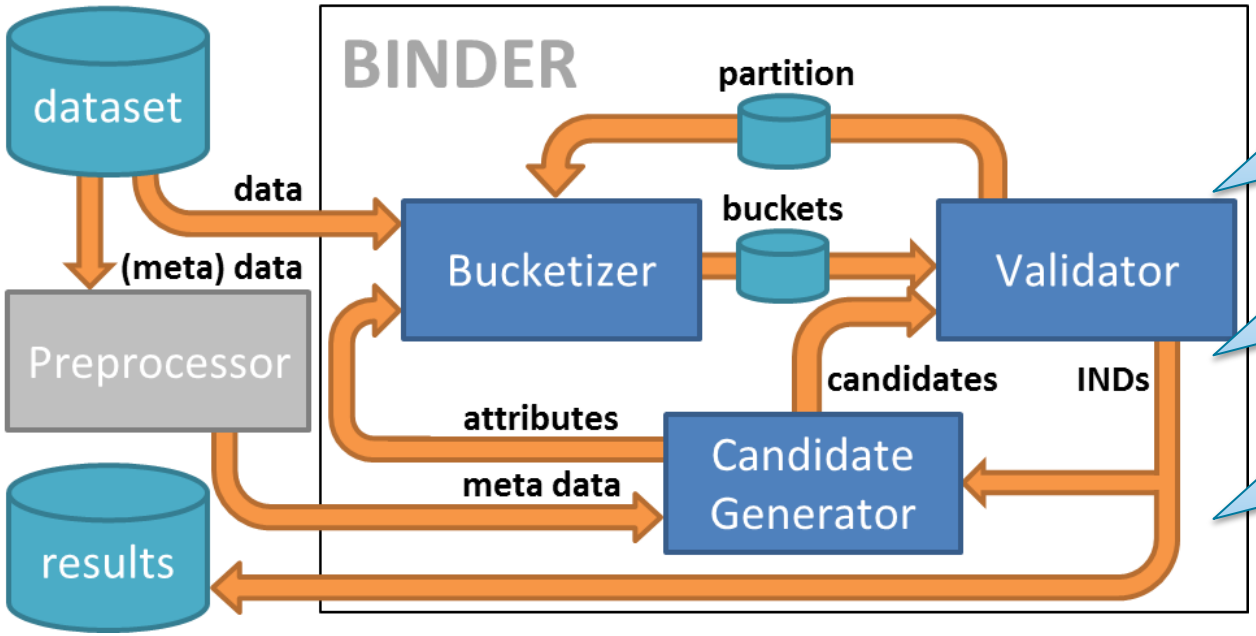
Experimental Evaluation
N-ary IND Discovery



Metanome

معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
Member of Qatar Foundation عضو في مؤسسة قطر

HPI Hasso Plattner Institut



Up to **26x faster** unary IND Discovery

Up to **2500x faster** n-ary IND Discovery

No file-handle or memory limits

Divide & Conquer-based Inclusion Dependency Discovery

Thorsten Papenbrock, Sebastian Kruse, Jorge-Arnulfo Quianè-Ruiz, Felix Naumann