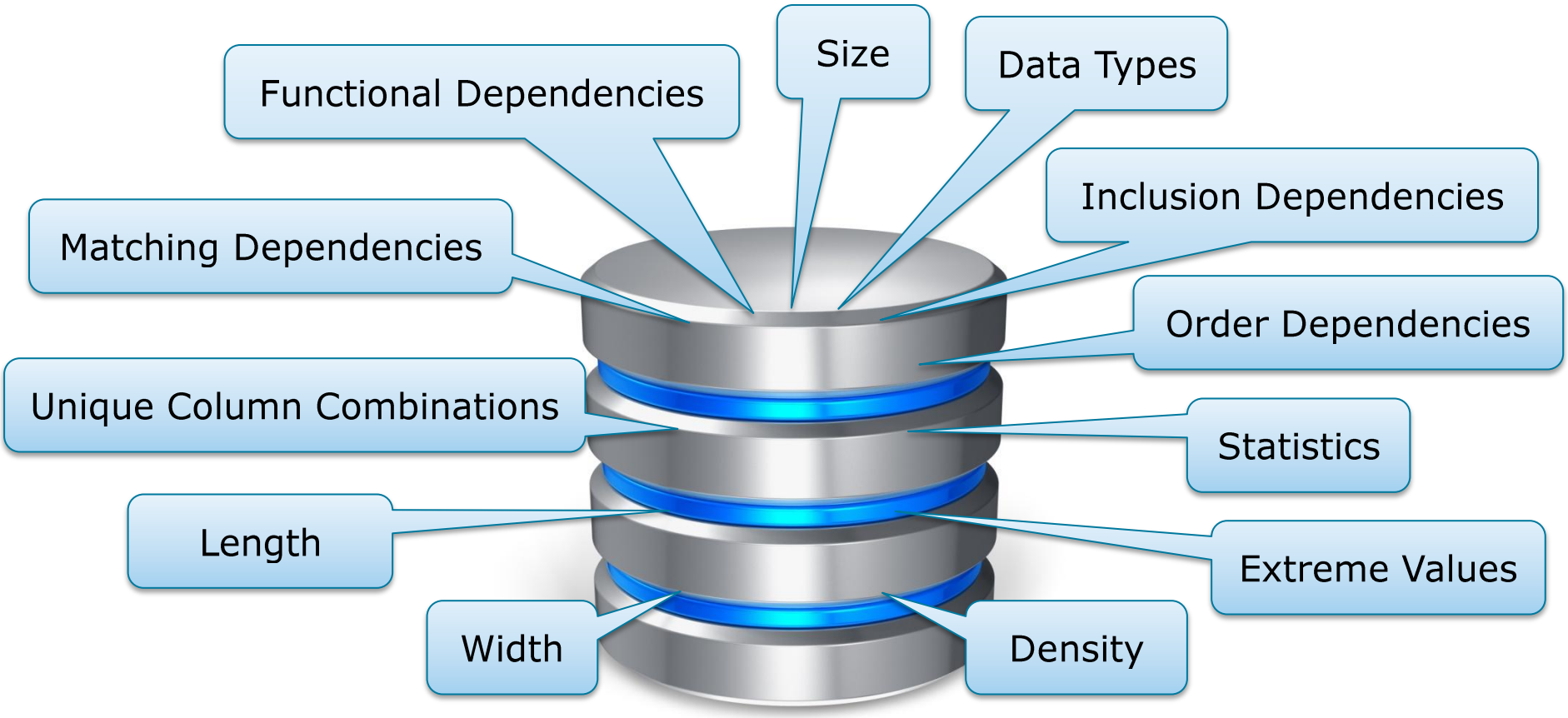
Metanome

HPI Hasso Plattner Institut

Holistic Data Profiling: Simultaneous Discovery of Various Metadata

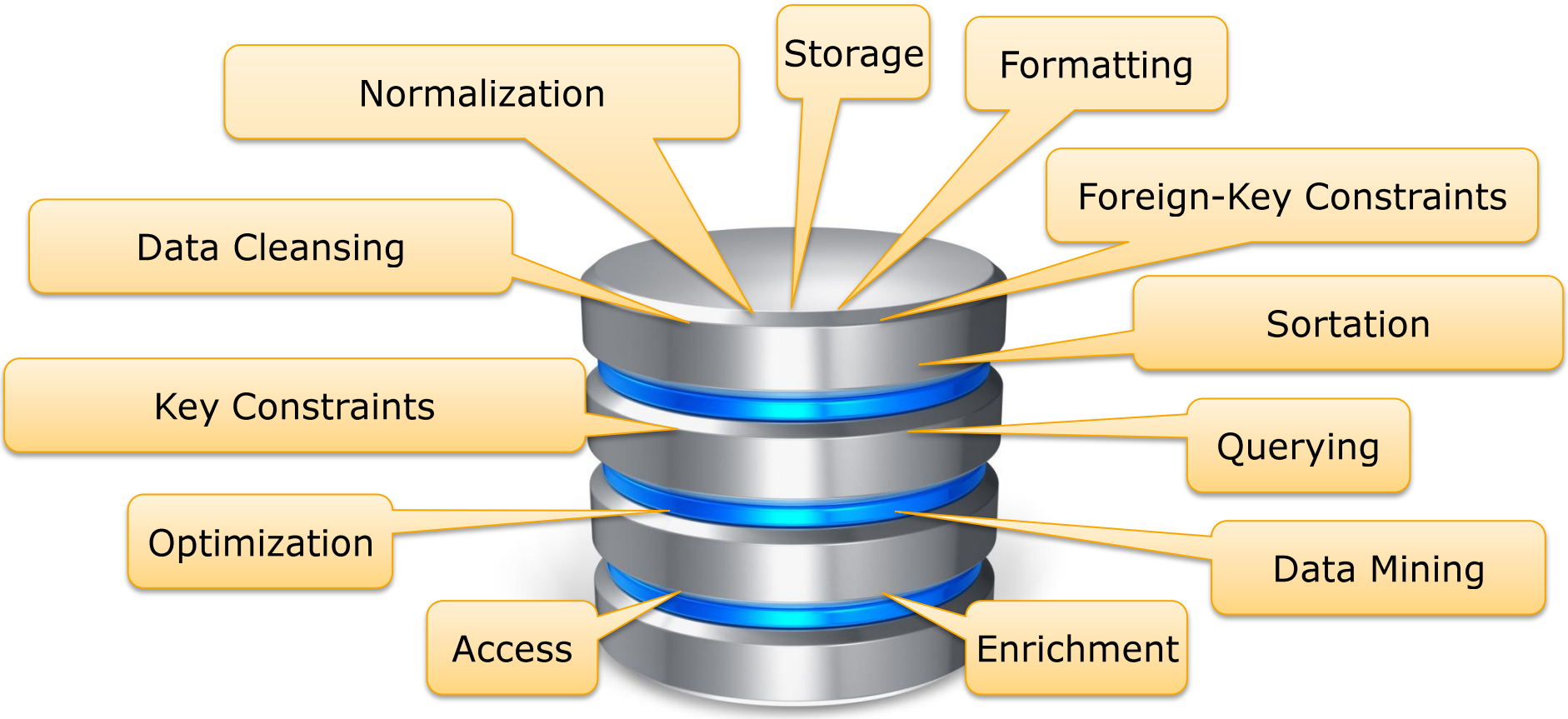
Ehrlich, Roick, Schulze, Zwiener, **Thorsten Papenbrock**, Naumann

Motivation
Metadata



Metadata is data about data.

Motivation
Metadata



Metadata is important to use, maintain and modify datasets.

Motivation
Metadata

Metadata is often missing and must be re-discovered: Data Profiling.

Motivation
Metadata

Each metadata type comes with its own discovery algorithm.

Motivation
Idea

Idea: One discovery algorithm for all metadata types.

→ UCCs, FDs, and INDs

Motivation
Idea

plowing

seeding

fertiliseing

Metanome

Agenda
Overview

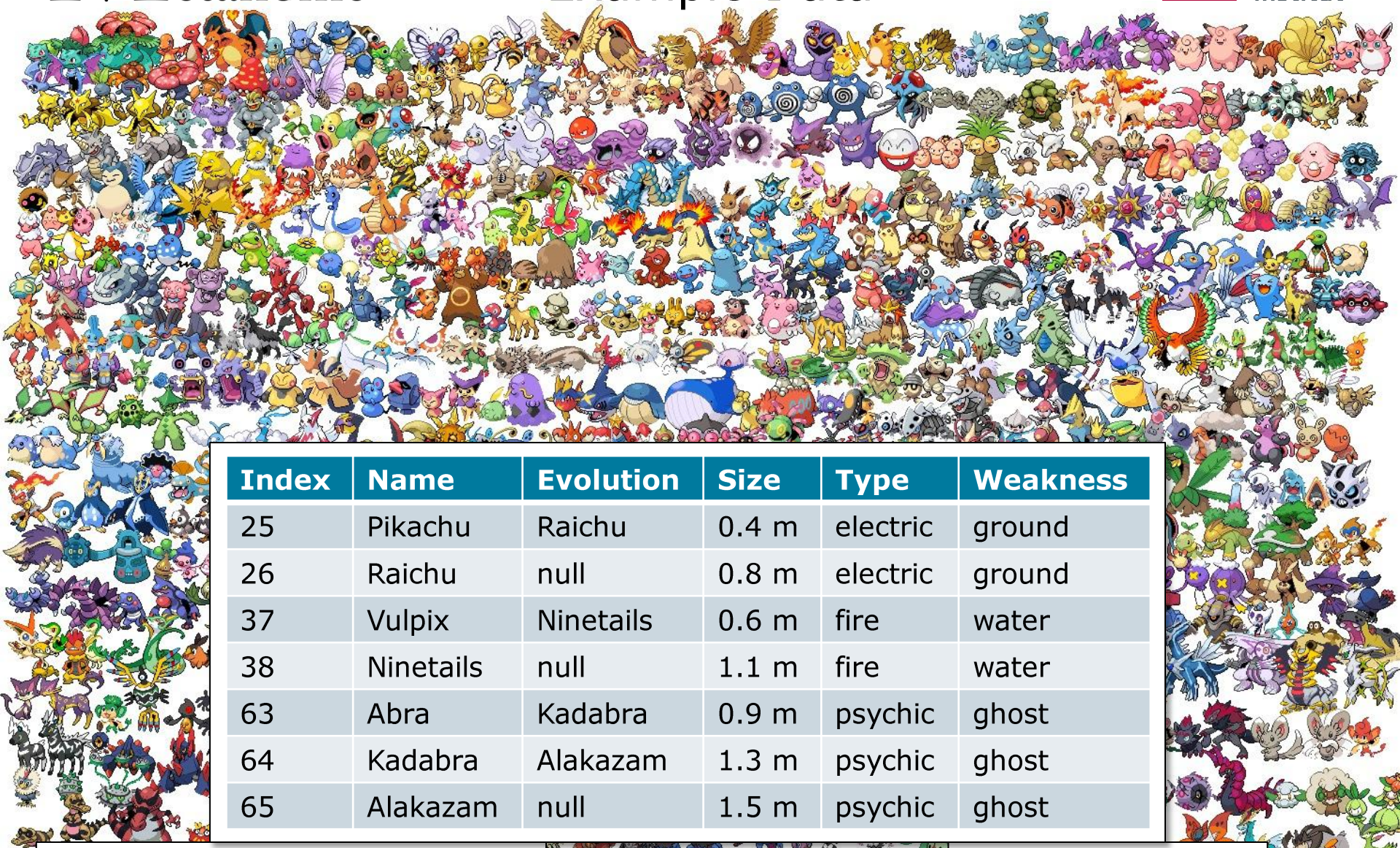
HPI Hasso Plattner Institut

Inter-Task Pruning Rules

The MUDS Algorithm

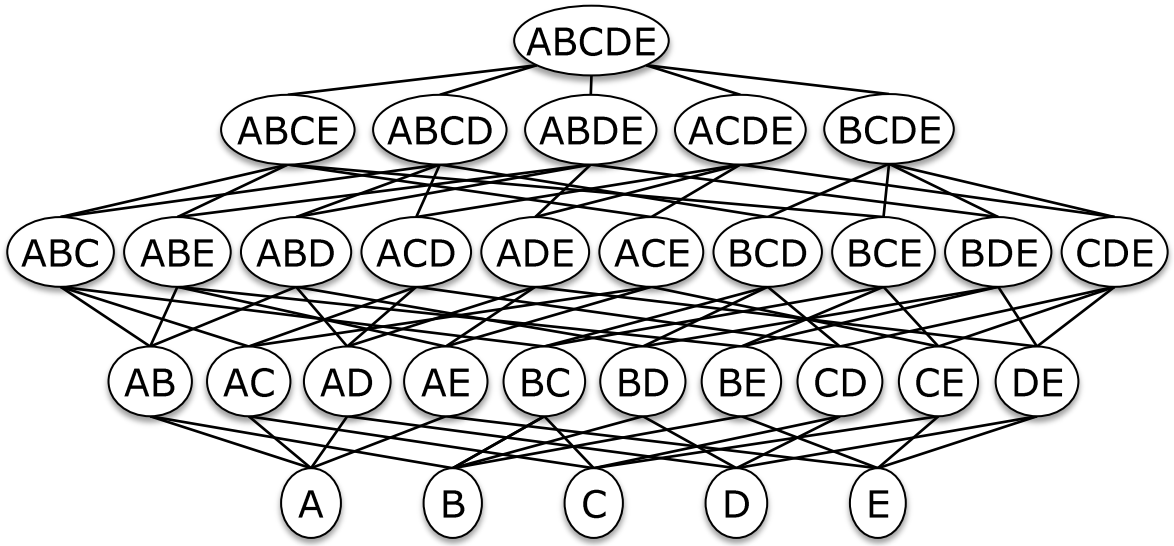
Experimental Evaluation

Inter-Task Pruning Rules
Example Data



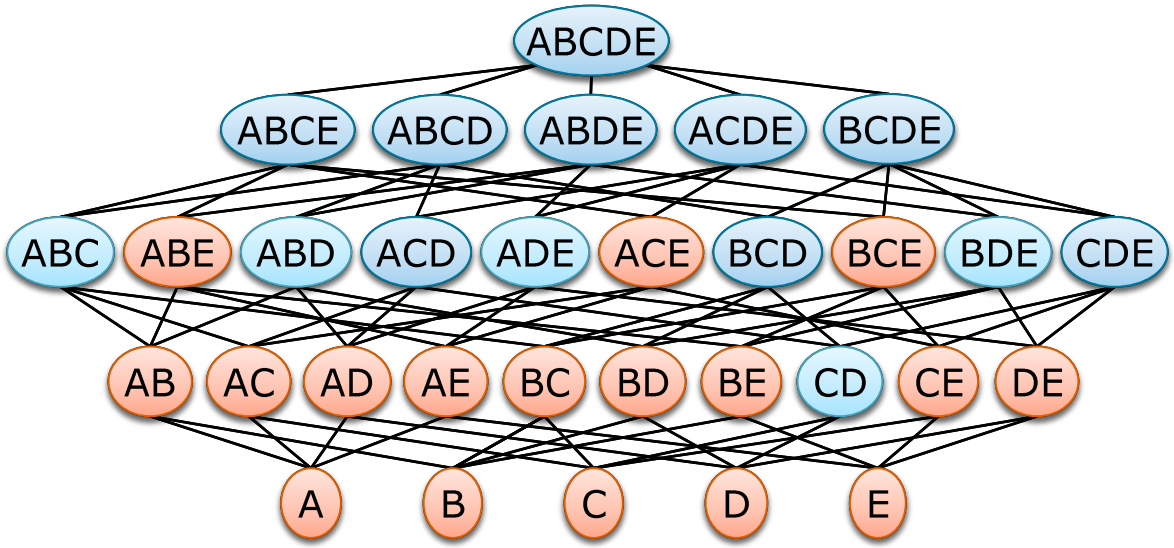
Index	Name	Evolution	Size	Type	Weakness
25	Pikachu	Raichu	0.4 m	electric	ground
26	Raichu	null	0.8 m	electric	ground
37	Vulpix	Ninetails	0.6 m	fire	water
38	Ninetails	null	1.1 m	fire	water
63	Abra	Kadabra	0.9 m	psychic	ghost
64	Kadabra	Alakazam	1.3 m	psychic	ghost
65	Alakazam	null	1.5 m	psychic	ghost

Inter-Task Pruning Rules
Search Space



Index	Name	Evolution	Size	Type	Weakness
25	Pikachu	Raichu	0.4 m	electric	ground
26	Raichu	null	0.8 m	electric	ground
37	Vulpix	Ninetails	0.6 m	fire	water
38	Ninetails	null	1.1 m	fire	water
63	Abra	Kadabra	0.9 m	psychic	ghost
64	Kadabra	Alakazam	1.3 m	psychic	ghost
65	Alakazam	null	1.5 m	psychic	ghost

Inter-Task Pruning Rules
Unique Column Combinations



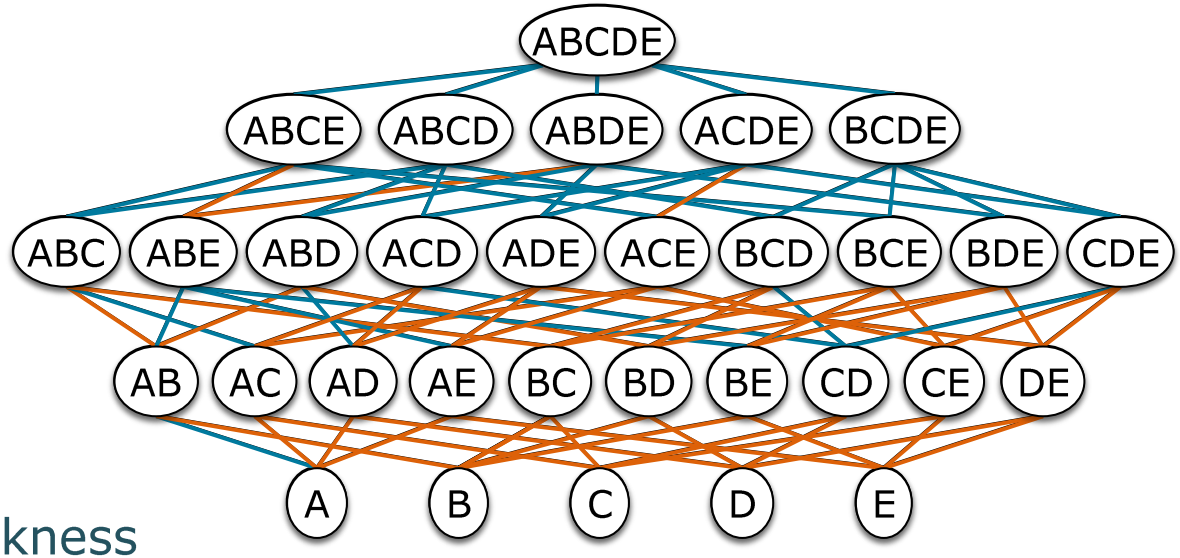
UCC

non-UCC

UCCs:
{Index}
{Name}

Index	Name	Evolution	Size	Type	Weakness
25	Pikachu	Raichu	0.4 m	electric	ground
26	Raichu	null	0.8 m	electric	ground
37	Vulpix	Ninetails	0.6 m	fire	water
38	Ninetails	null	1.1 m	fire	water
63	Abra	Kadabra	0.9 m	psychic	ghost
64	Kadabra	Alakazam	1.3 m	psychic	ghost
65	Alakazam	null	1.3 m	psychic	ghost

Inter-Task Pruning Rules
Functional Dependencies



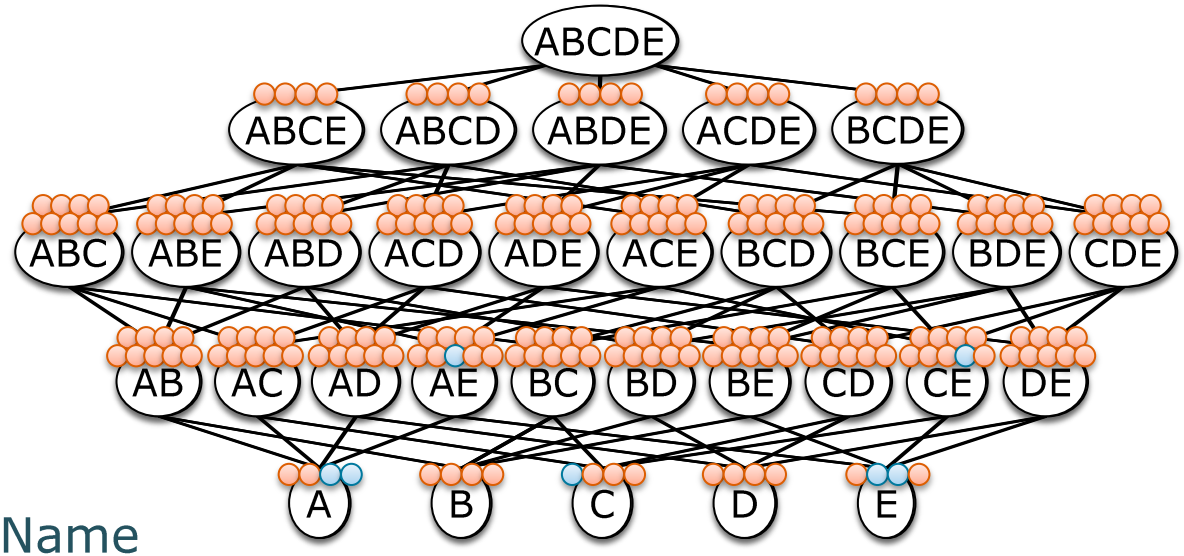
FD

non-FD

FDs:
Type → Weakness

Index	Name	Evolution	Size	Type	Weakness
25	Pikachu	Raichu	0.4 m	electric	ground
26	Raichu	null	0.8 m	electric	ground
37	Vulpix	Ninetails	0.6 m	fire	water
38	Ninetails	null	1.1 m	fire	water
63	Abra	Kadabra	0.9 m	psychic	ghost
64	Kadabra	Alakazam	1.3 m	psychic	ghost
65	Alakazam	null	1.5 m	psychic	ghost

Inter-Task Pruning Rules
Inclusion Dependencies



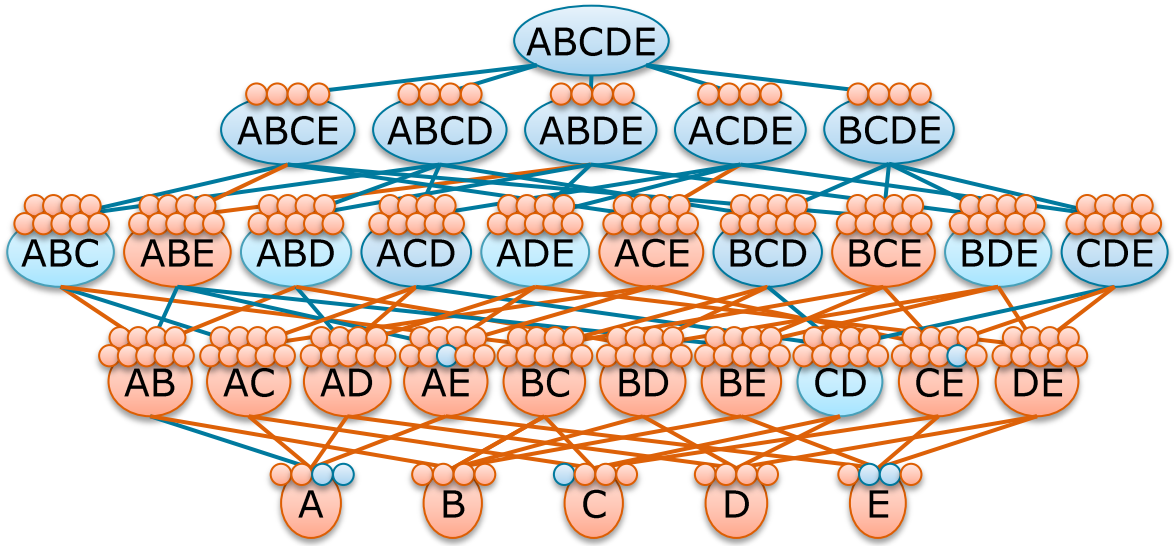
IND

non-IND

INDs: Evolution ⊆ Name

Index	Name	Evolution	Size	Type	Weakness
25	Pikachu	Raichu	0.4 m	electric	ground
26	Raichu	null	0.8 m	electric	ground
37	Vulpix	Ninetails	0.6 m	fire	water
38	Ninetails	null	1.1 m	fire	water
63	Abra	Kadabra	0.9 m	psychic	ghost
64	Kadabra	Alakazam	1.3 m	psychic	
65	Alakazam	null	1.5 m	ps	

Inter-Task Pruning Rules
Holistic Discovery



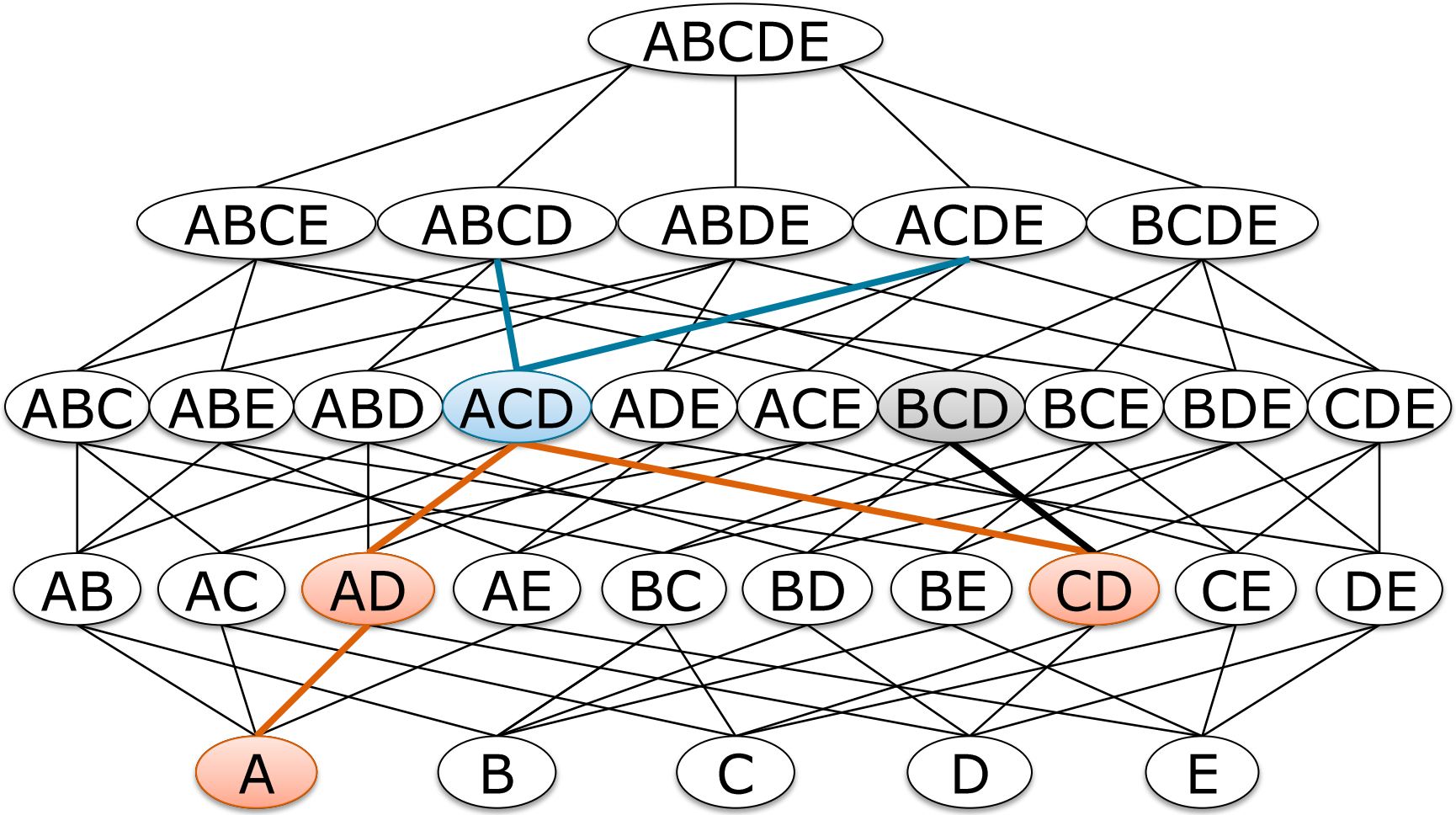
Holistic discovery of UCCs, FDs, and unary INDs

Why only unary INDs?

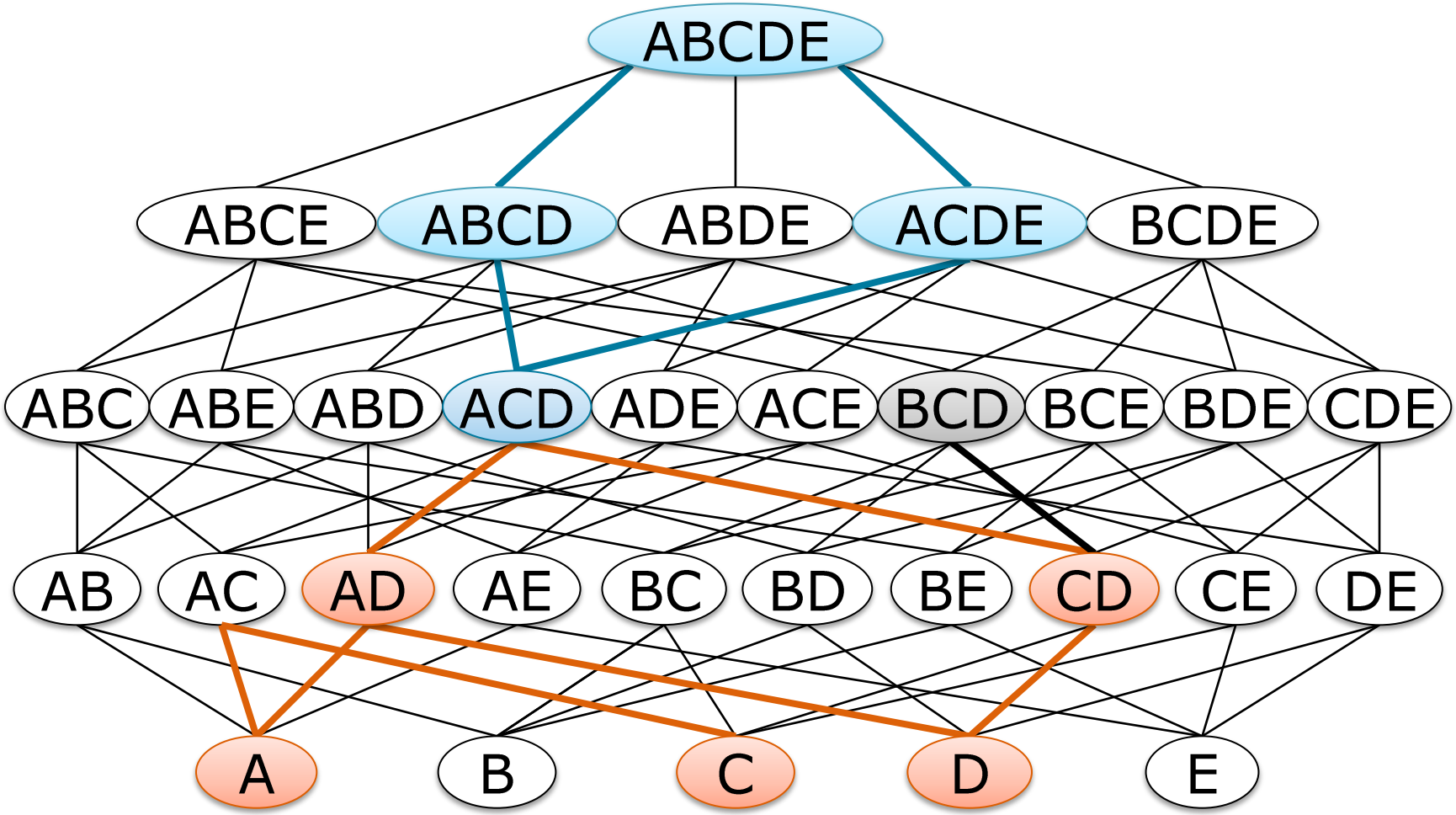
1. INDs: valid in lower lattice, invalid in upper lattice
   UCCs/FDs: invalid in lower lattice, valid in upper lattice
2. INDs: test values (positions don't matter)
   UCCs/FDs: test positions (values don't matter)

→ N-ary IND search differs; use specialized algorithms (BINDER, MIND)

Inter-Task Pruning Rules
Random Walk

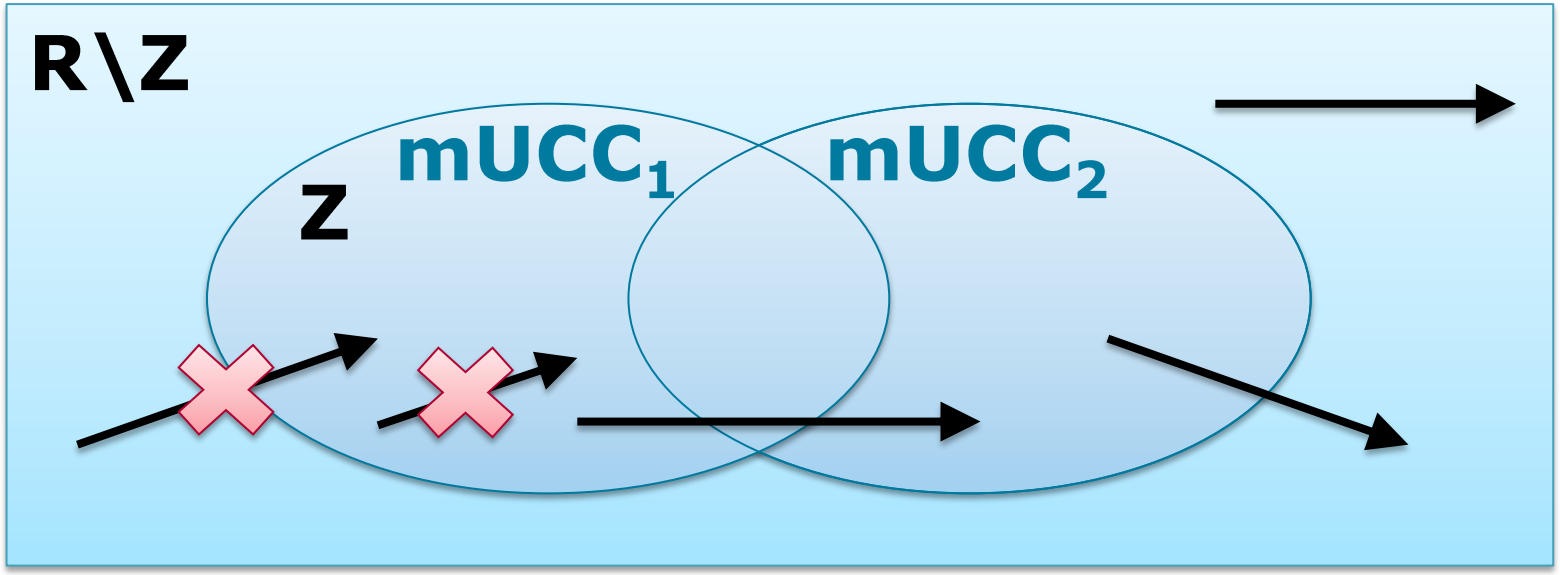


Inter-Task Pruning Rules
Random Walk



Inter-Task Pruning Rules
Novel Pruning Rules

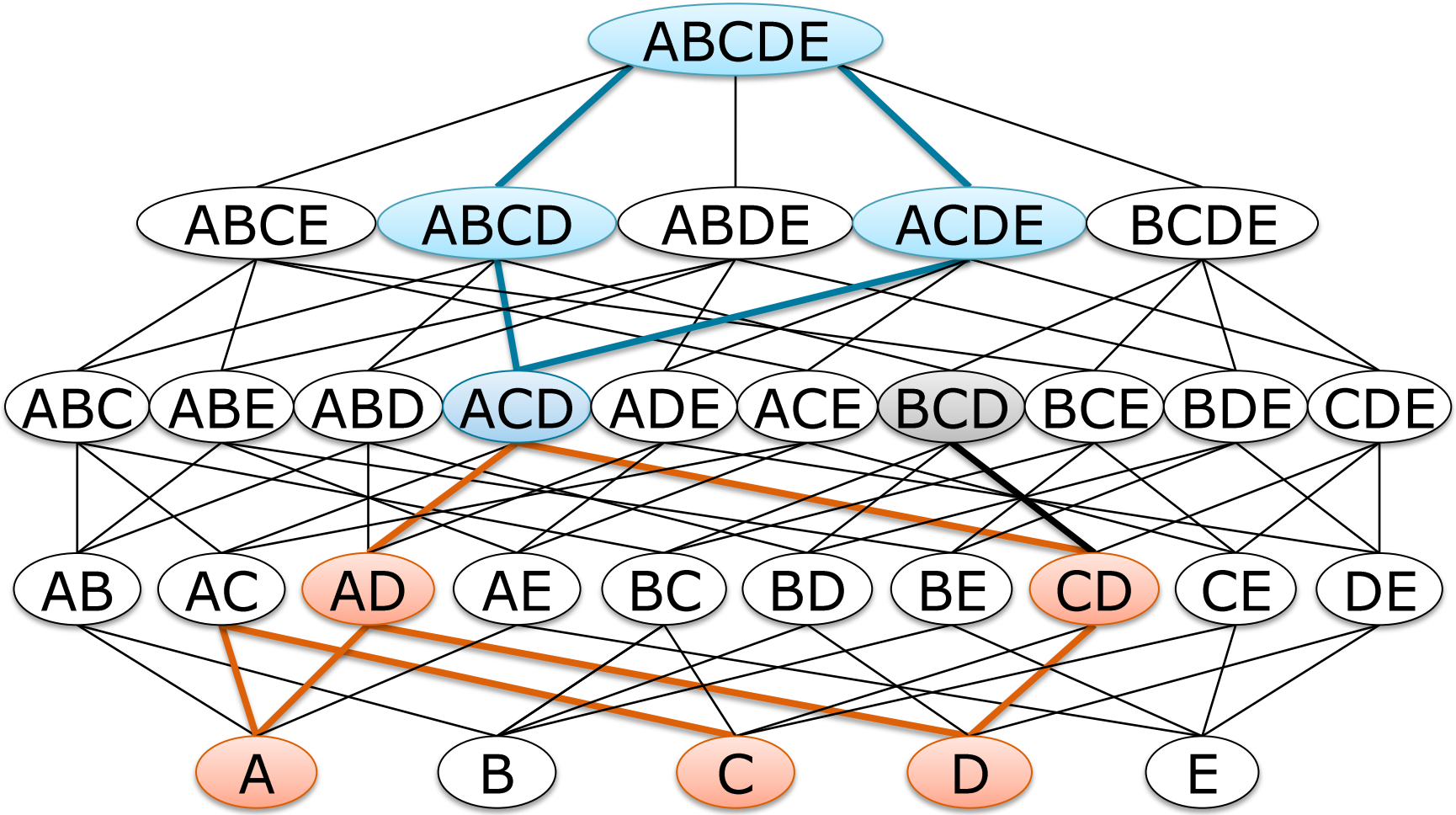
R:



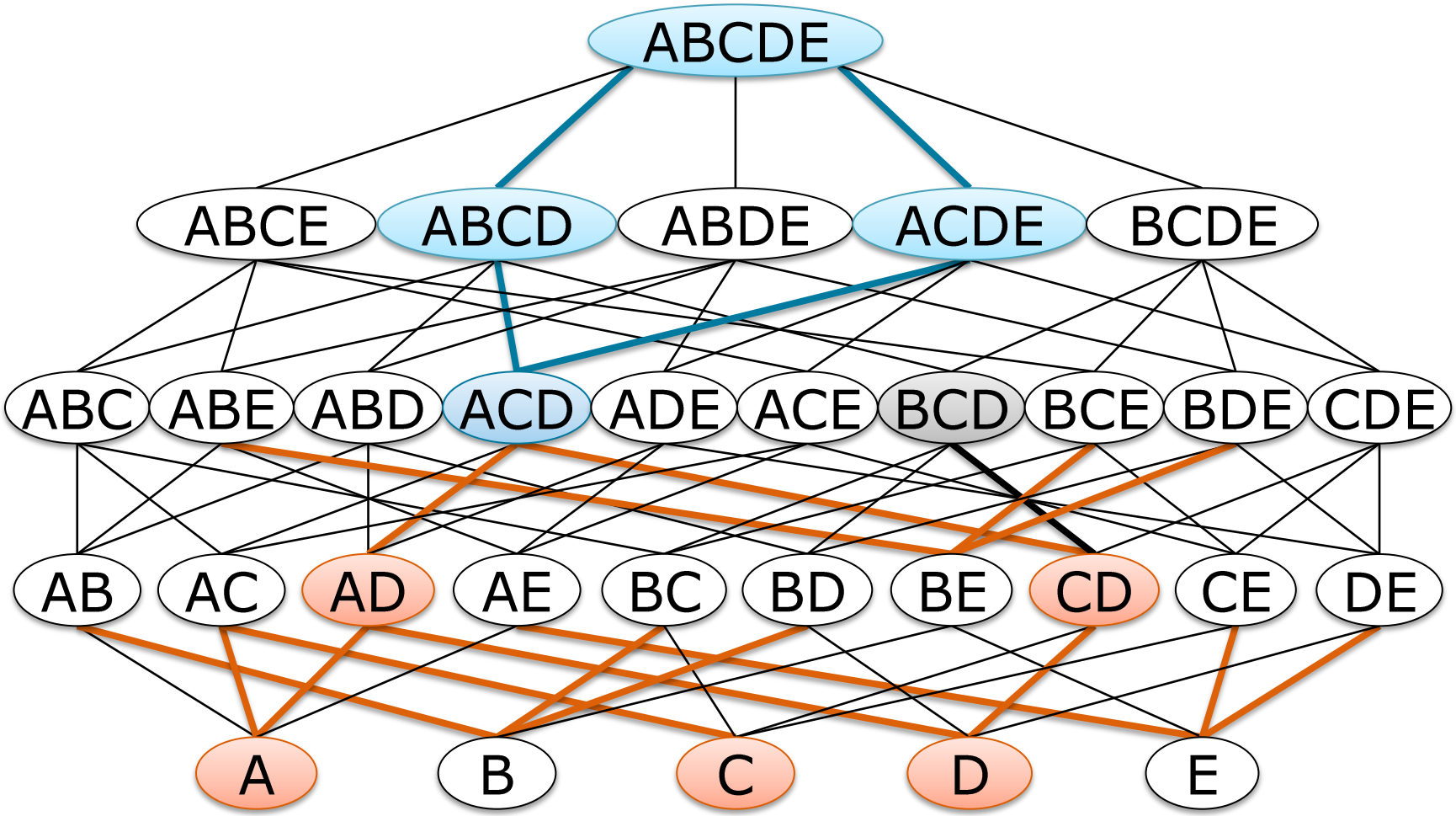
Pruning Rules:

1. Columns from R\Z cannot determine a column in a minimal UCC.
2. Columns from a minimal UCC cannot determine a column in this UCC.

Inter-Task Pruning Rules
Random Walk

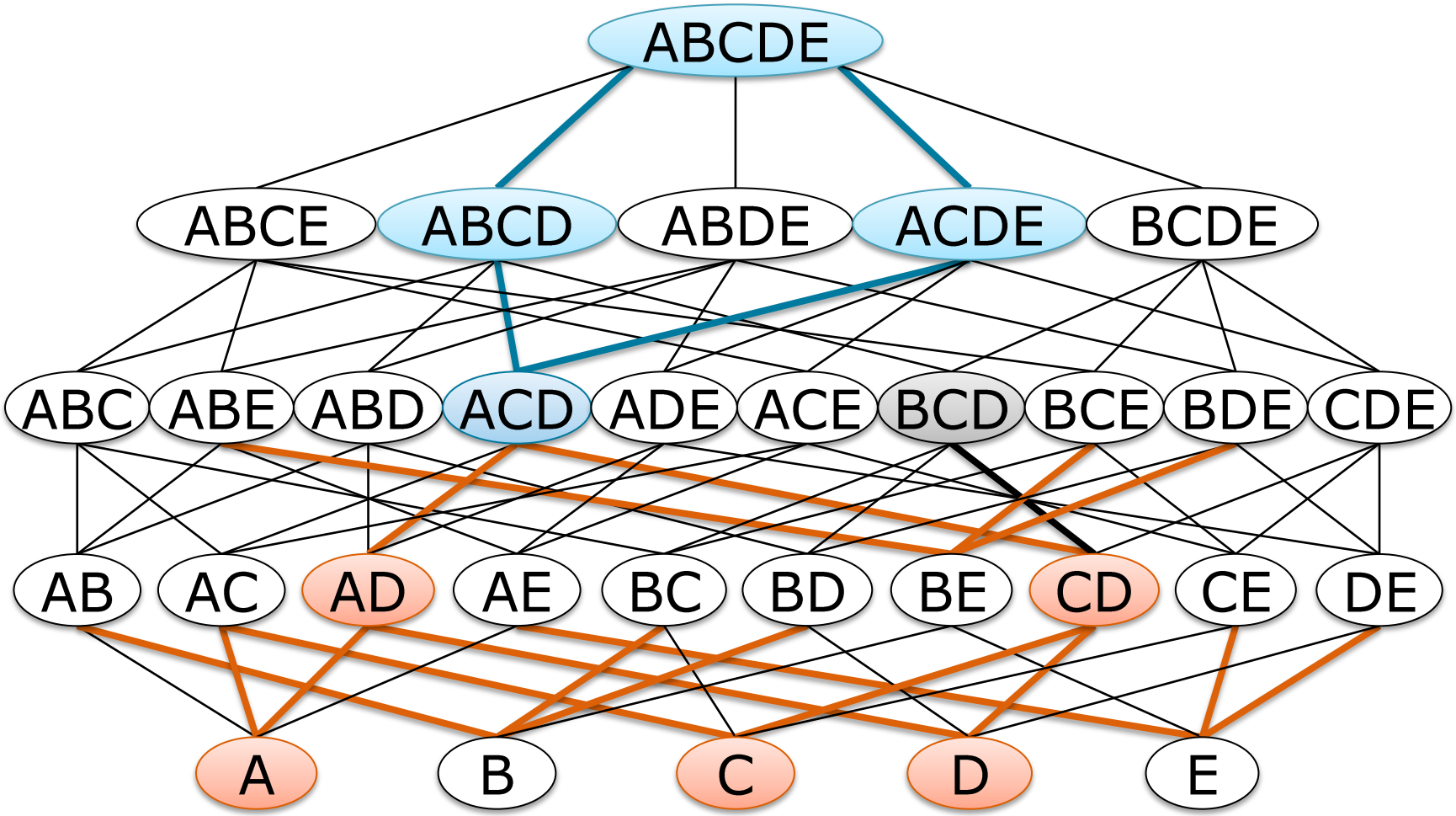


Inter-Task Pruning Rules
Random Walk



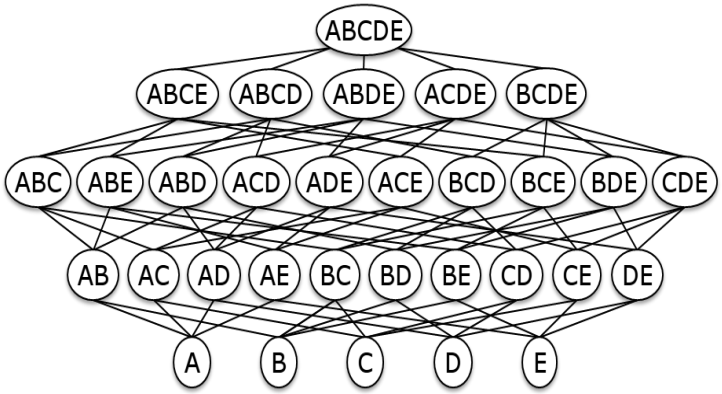
1. "Columns from R\Z cannot determine a column in a minimal UCC."

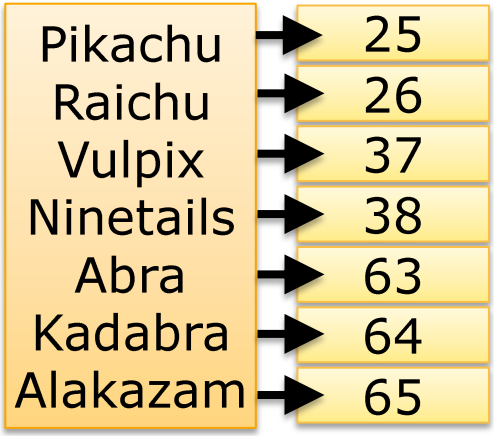
Inter-Task Pruning Rules
Random Walk



2. "Columns from a minimal UCC cannot determine a column in this UCC."

Inter-Task Pruning Rules
Sharing Resources





One I/O-Pass:
- Allocating memory
- Reading bytes
- Parsing lines

One search lattice:
- Candidates
- Random walker
- Pruning information

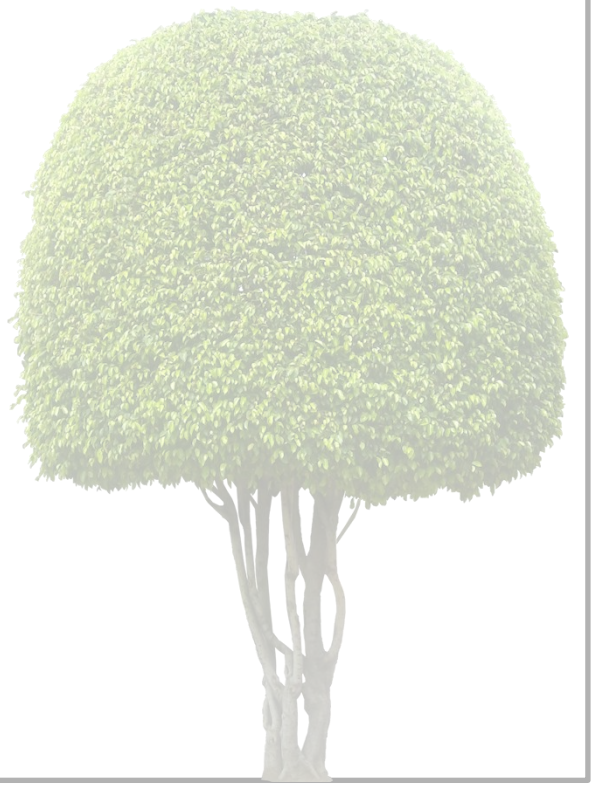
Shared data structures:
- Position List Indices
- Sorted value lists
- Column representations

Metanome

Agenda
Overview

HPI Hasso Plattner Institut

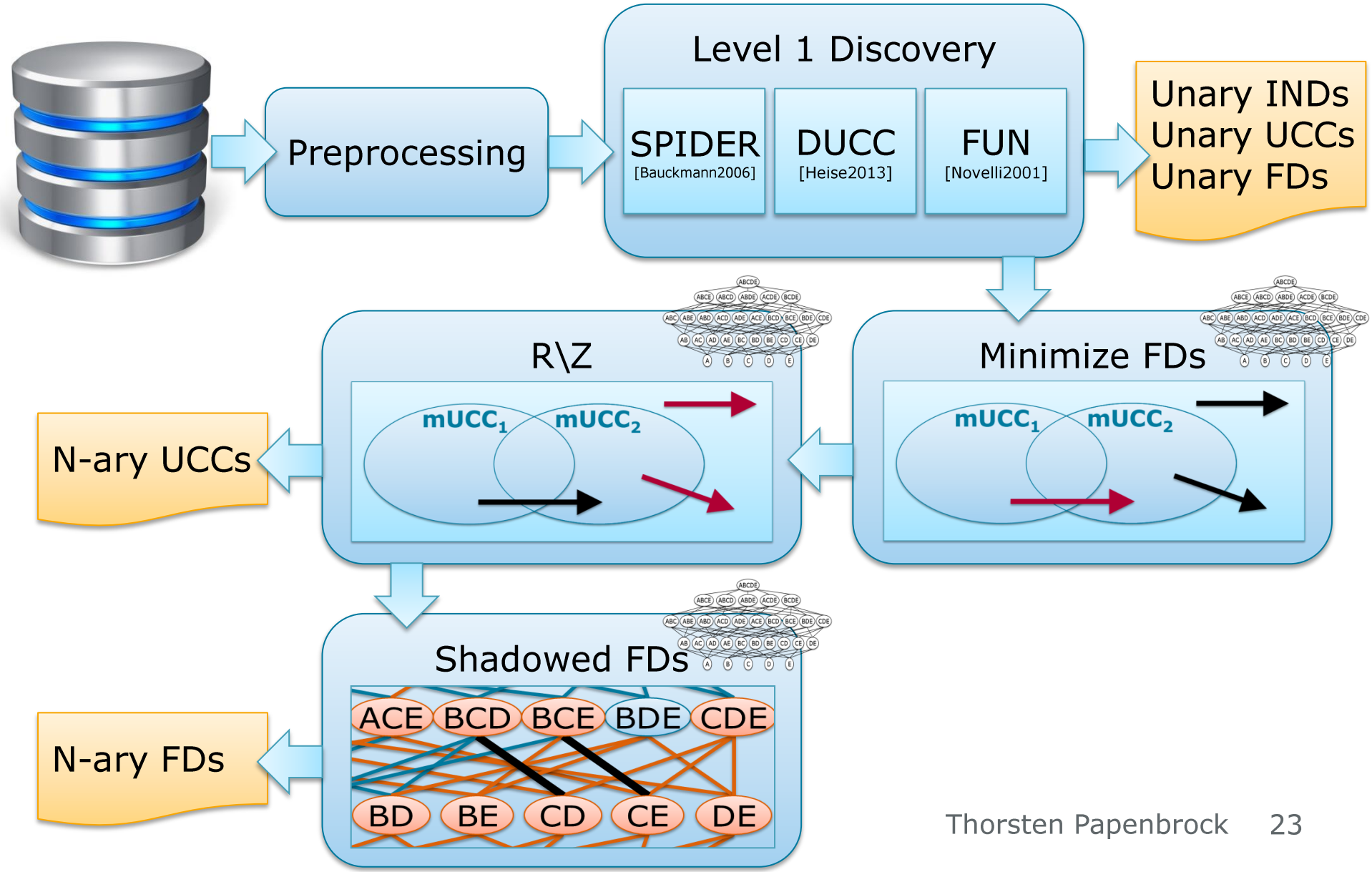
Inter-Task Pruning Rules



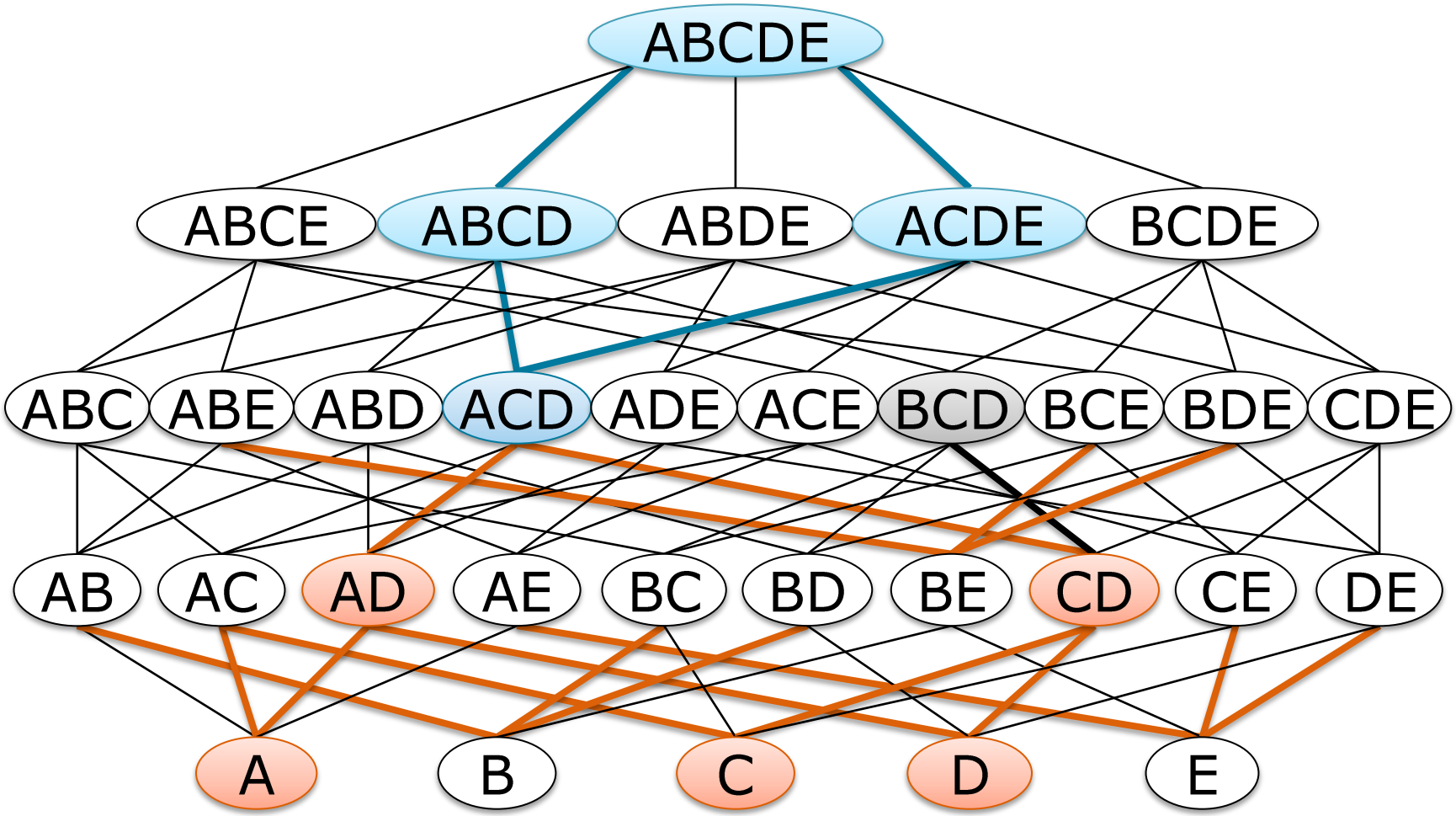
The MUDS Algorithm

Experimental Evaluation

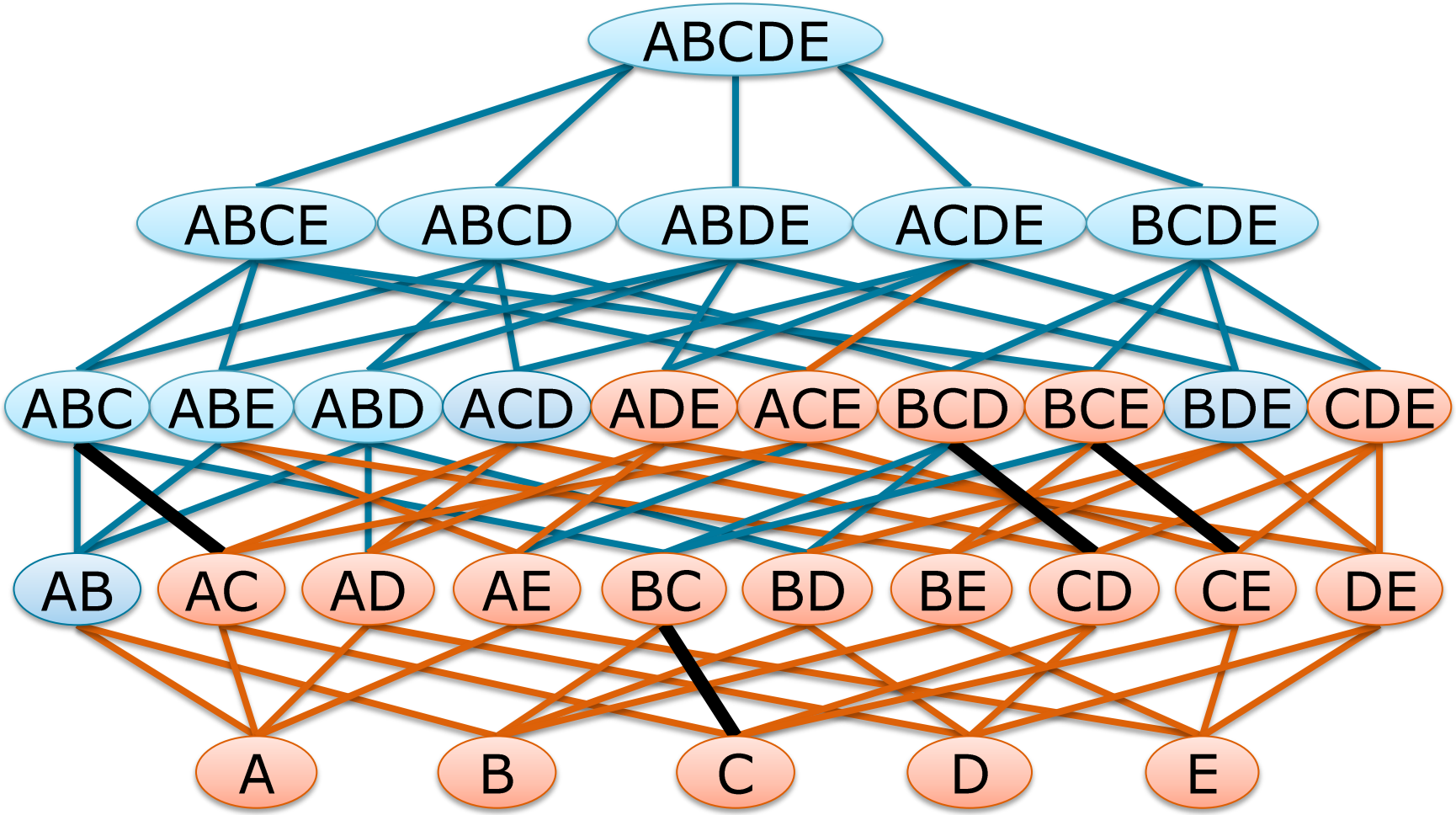
The MUDS Algorithm
Algorithm Overview



The MUDS Algorithm
Shadowed FDs

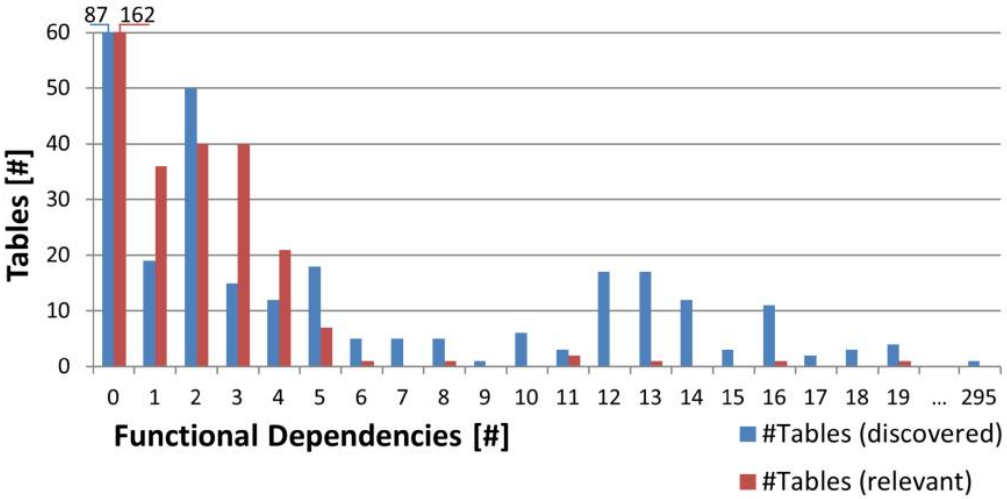


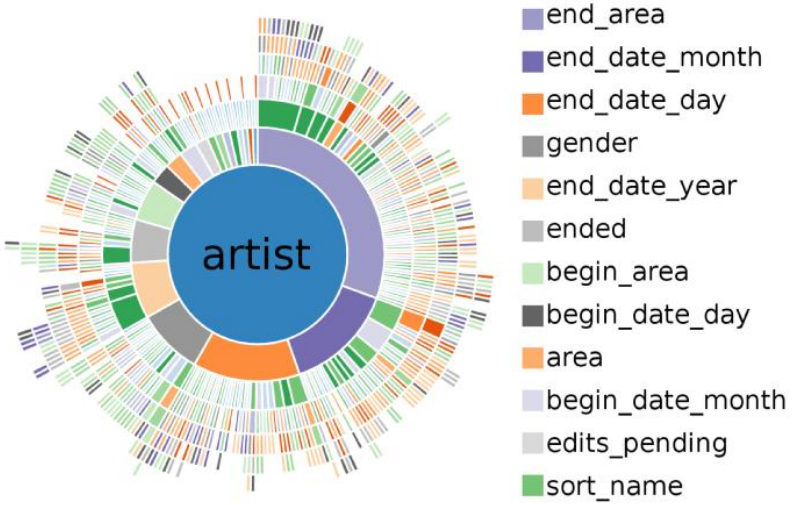
The MUDS Algorithm
Shadowed FDs



The MUDS Algorithm
Results on MusicBrainz

Metadata type	#Results
Unique Column Combinations (UCCs)	675
Functional Dependencies (FDs)	4,193
Inclusion Dependencies (INDs)	381,927



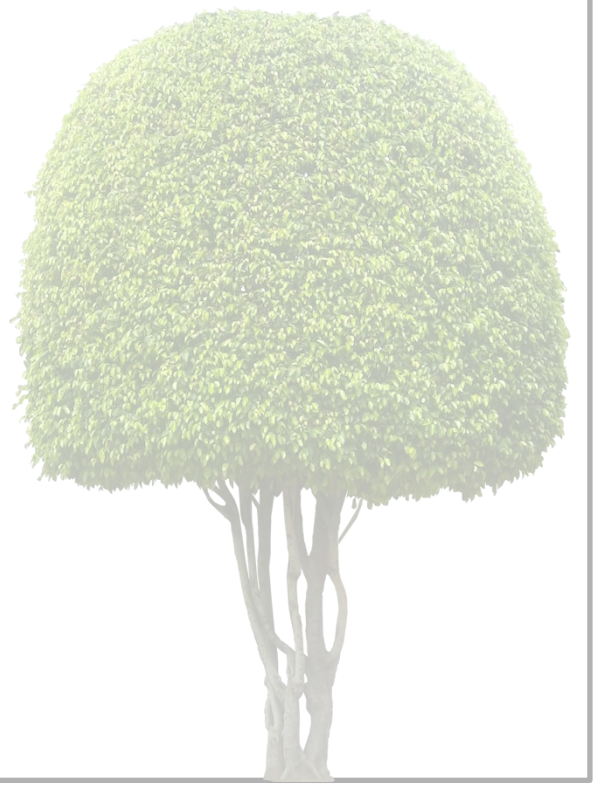


Metanome

Agenda
Overview

HPI Hasso Plattner Institut

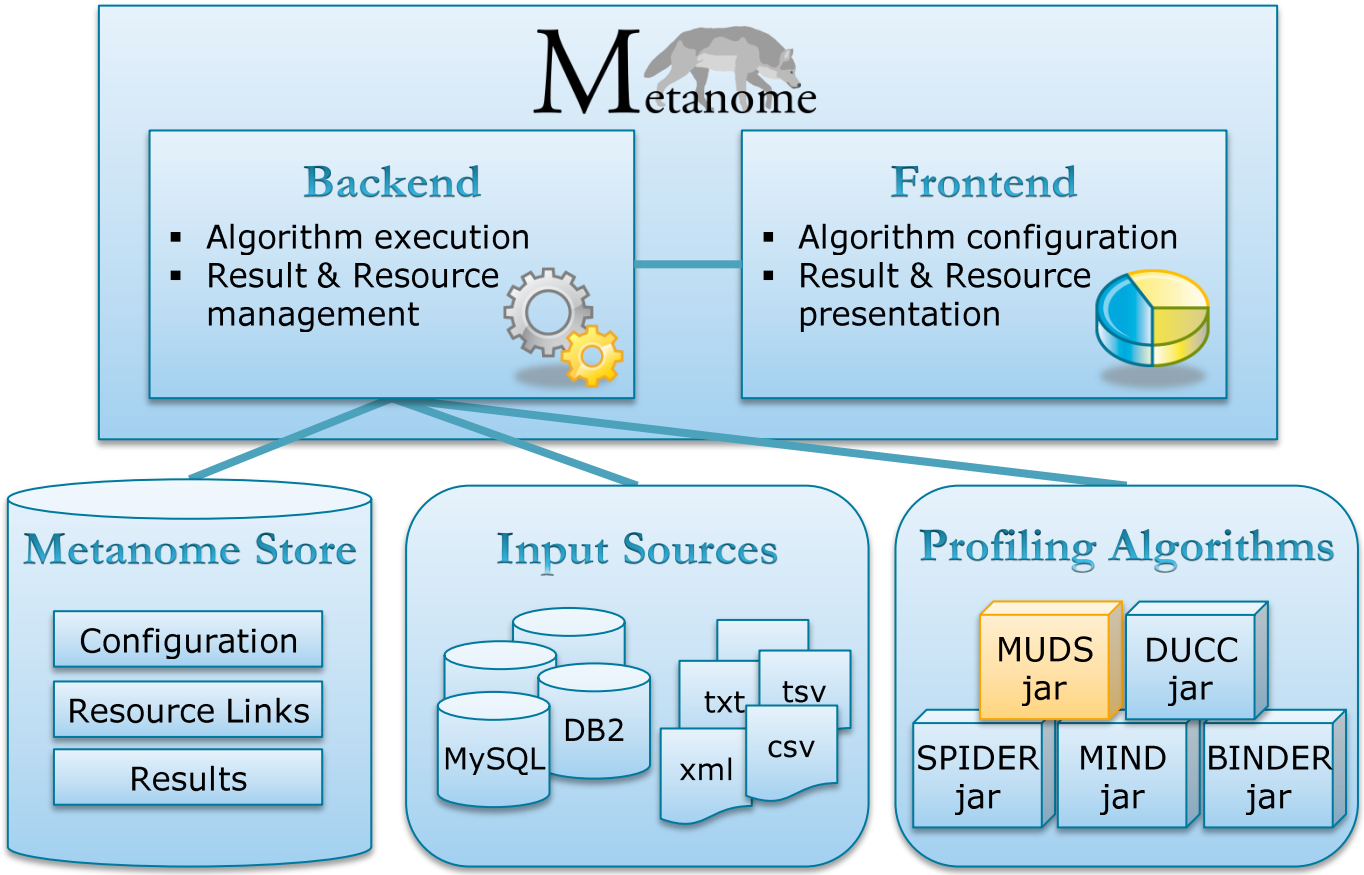
Inter-Task Pruning Rules



The MUDS Algorithm

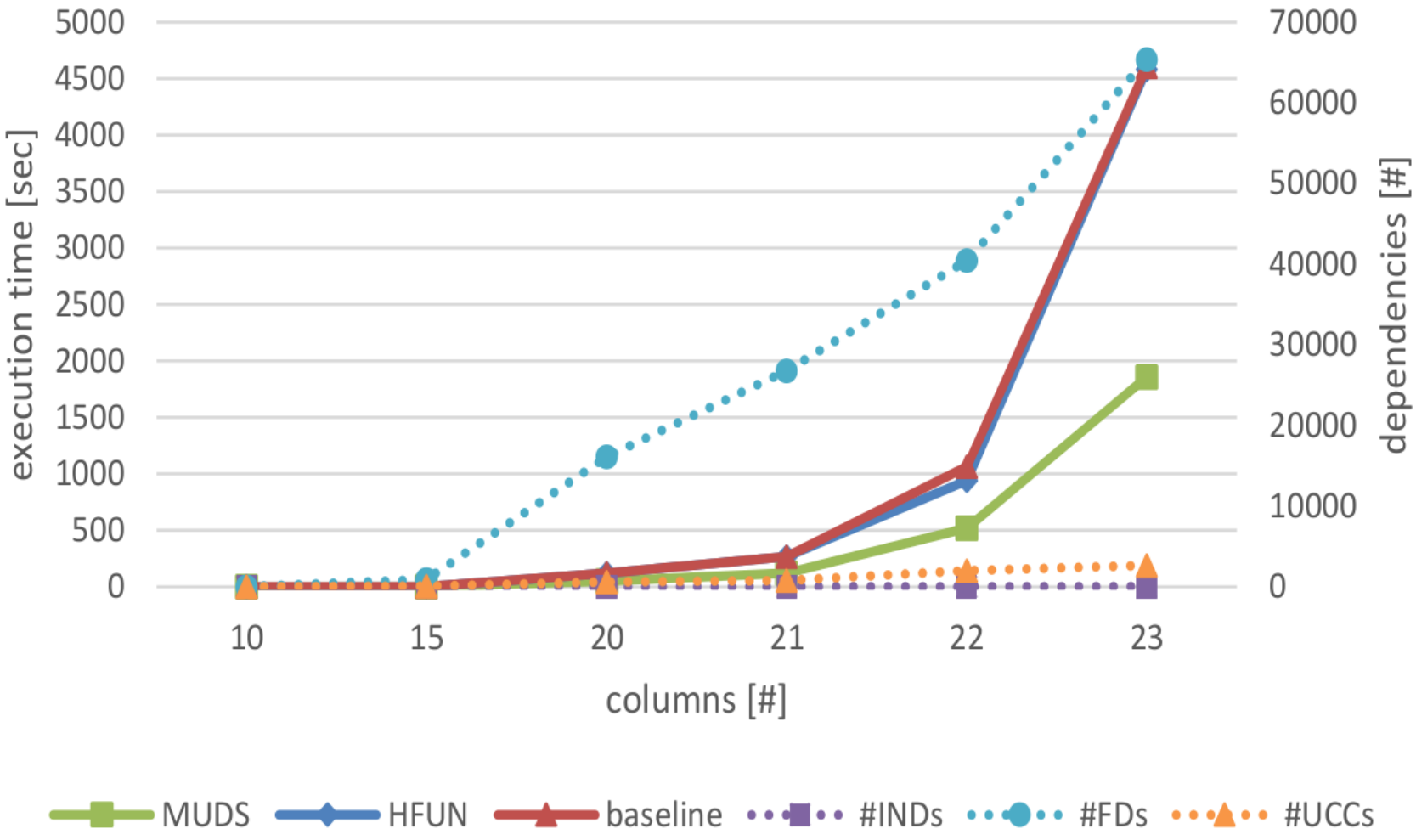
Experimental Evaluation

Experimental Evaluation
Metanome Profiling Tool



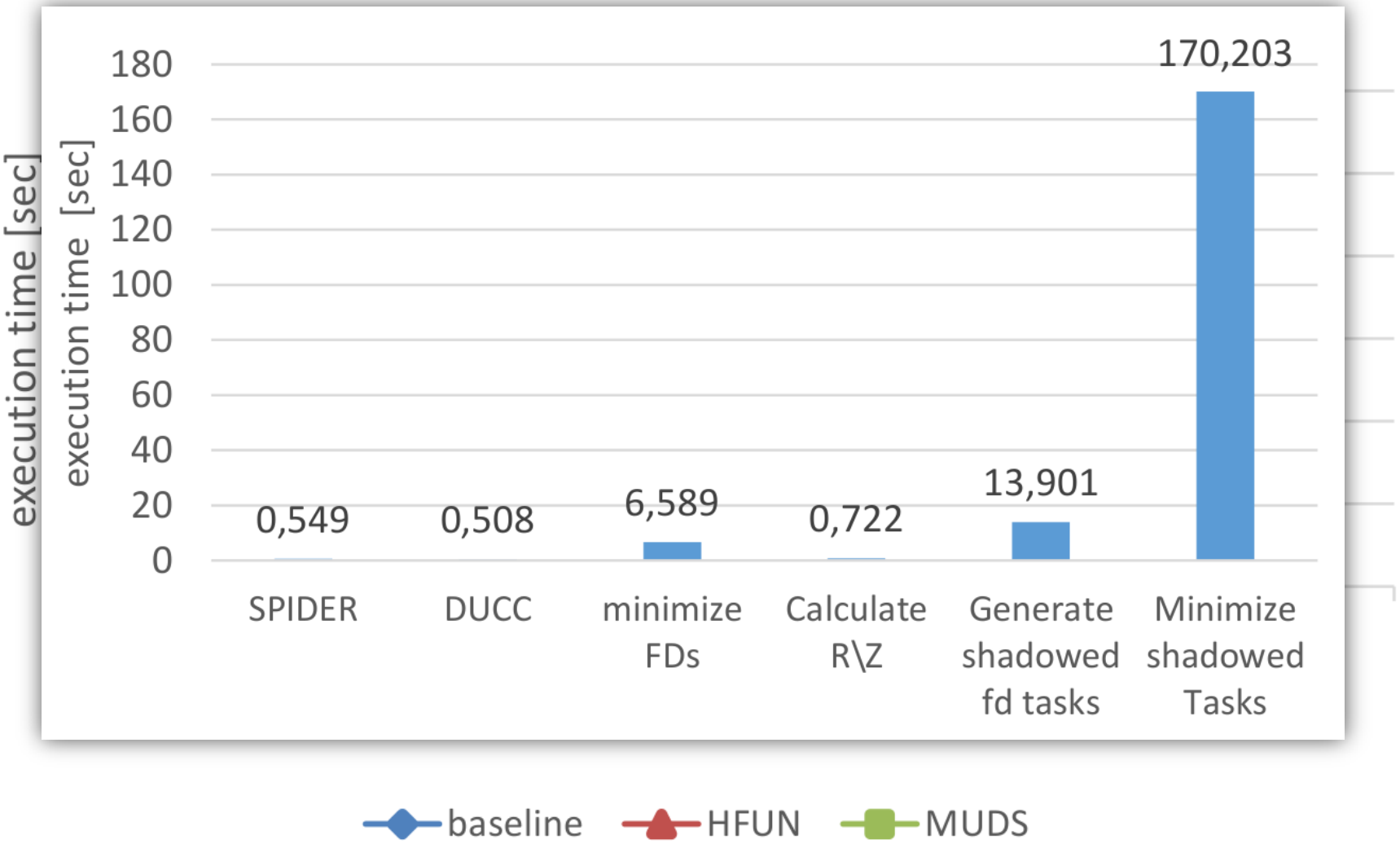
www.metanome.de

Experimental Evaluation
Column Scalability



ionosphere dataset with 351 rows

Experimental Evaluation
Row Scalability



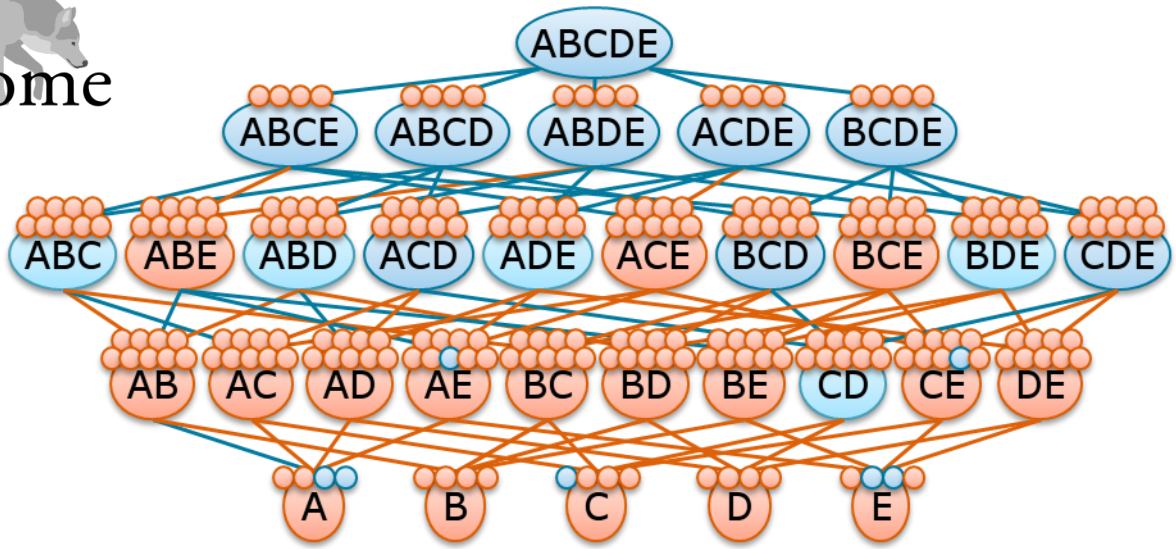
uniprot dataset with 20 columns

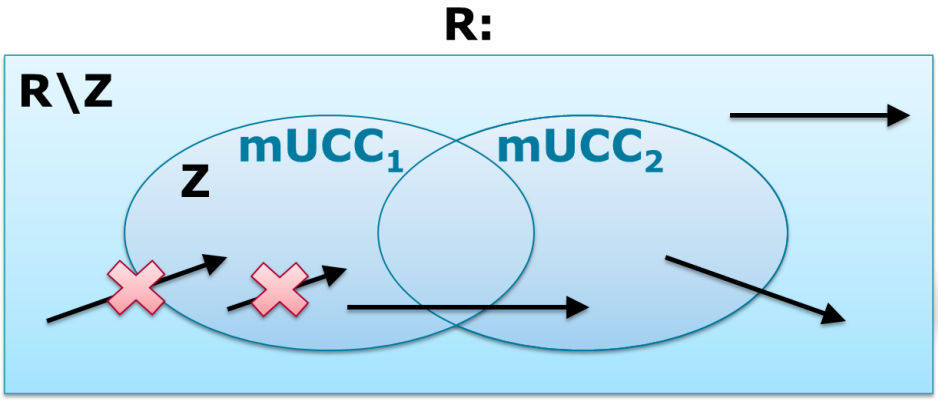
Experimental Evaluation
Datasets

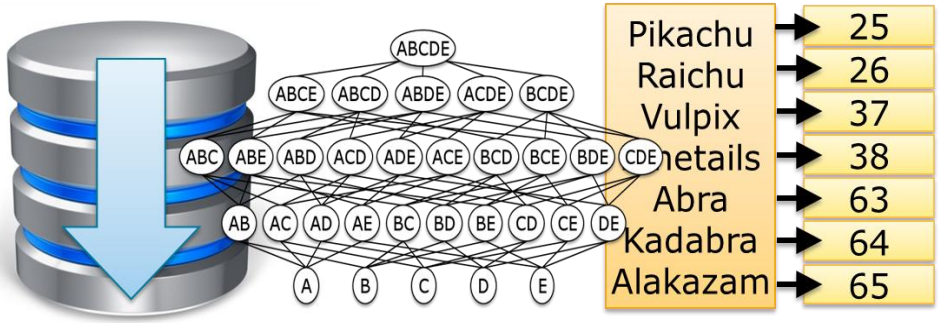
Dataset	Col	Row	FDs	basel.	HFUN	MUDS	TANE
iris	5	150	4	**.1s**	**.1s**	**.1s**	.6s
balance	5	625	1	.3s	**.1s**	**.1s**	.9s
chess	7	28k	1	2.0s	**.9s**	1.5s	2.0s
abalone	9	4k	137	1.3s	**.6s**	1.1s	1.0s
nursery	9	12k	1	2.3s	**1.9s**	3.1s	3.1s
b-cancer	11	699	46	.8s	.6s	**.5s**	1.4s
bridges	13	108	142	.8s	.7s	**.6s**	1.3s
echocard	13	132	538	1.0s	**.6s**	1.6s	.8s
adult	14	48k	78	126s	118s	**9.9s**	81.2s
letter	17	20k	61	706s	636s	**13.2s**	326.0s
hepatitis	20	155	8k	462s	450s	88.1s	**10.9s**

Metanome

HPI Hasso Plattner Institut







Holistic Data Profiling: Simultaneous Discovery of Various Metadata

Ehrlich, Roick, Schulze, Zwiener, **Thorsten Papenbrock**, Naumann