Metanome

HPI Hasso Plattner Institut

A Hybrid Approach to Functional Dependency Discovery

Thorsten Papenbrock and Felix Naumann

Motivation
Functional Dependencies

Name	Surname	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	Cone	01069	Dresden	Orosz
Thomas	Cone	14482	Jakobs	
Mike	Moore	60329	Frankfurt	Feldmann

Definition FD: **X → A**

- All values in X uniquely define the values in A.
- If $t_1[X] = t_2[X]$, then $t_1[A] = t_2[A]$.

Postcode → City
Postcode → Mayor

Motivation
Functional Dependencies

Name	Surname	Postcode
Thomas	Miller	14482
Sarah	Miller	14482
Peter	Smith	60329
Jasmine	Cone	01069
Thomas	Cone	14482
Mike	Moore	60329

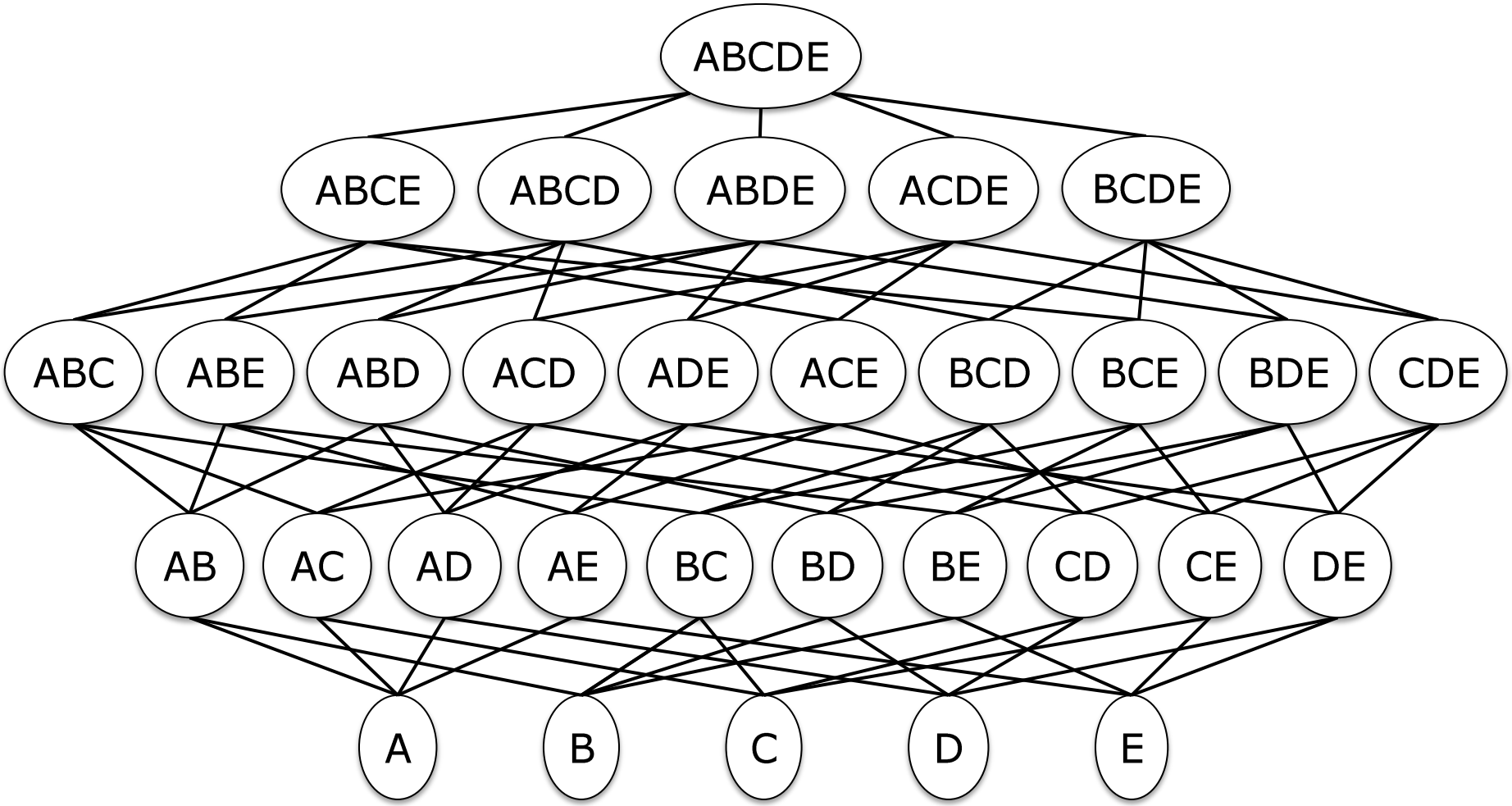
Postcode	City	Mayor
14482	Potsdam	Jakobs
60329	Frankfurt	Feldmann
01069	Dresden	Orosz

Definition FD: **X → A**

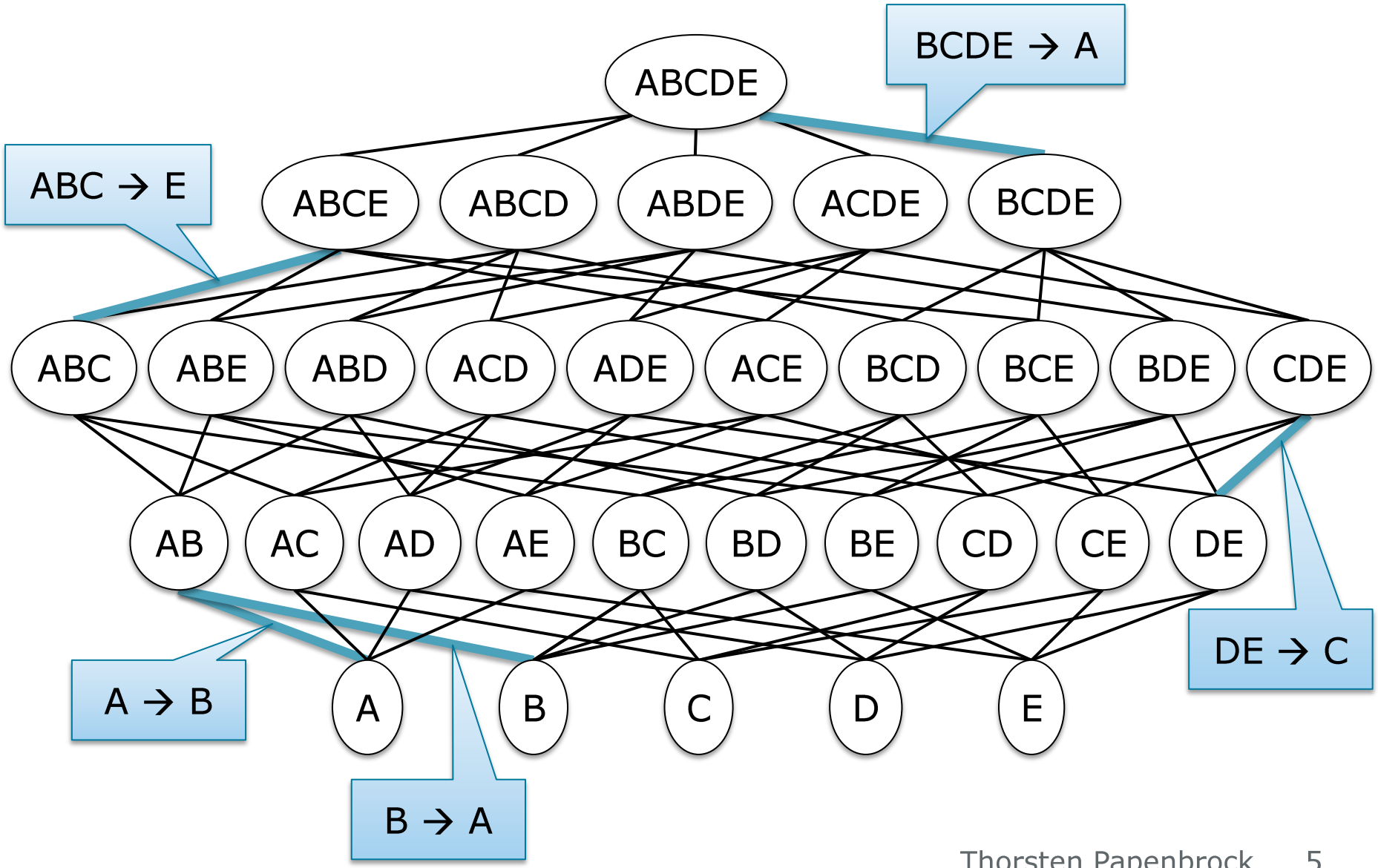
- All values in X uniquely define the values in A.
- If $t_1[X] = t_2[X]$, then $t_1[A] = t_2[A]$.

Postcode → City
Postcode → Mayor

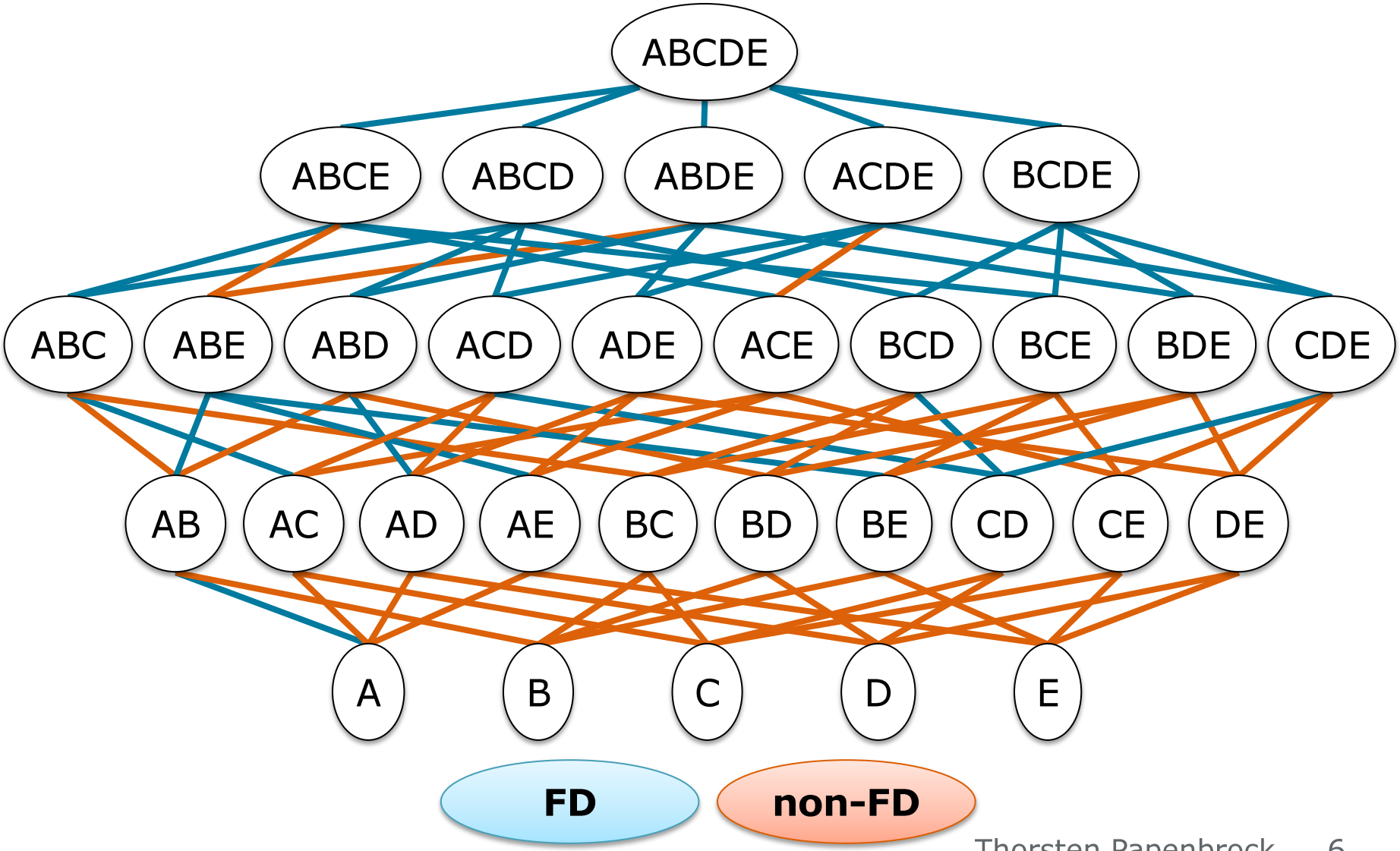
Motivation
FD Discovery



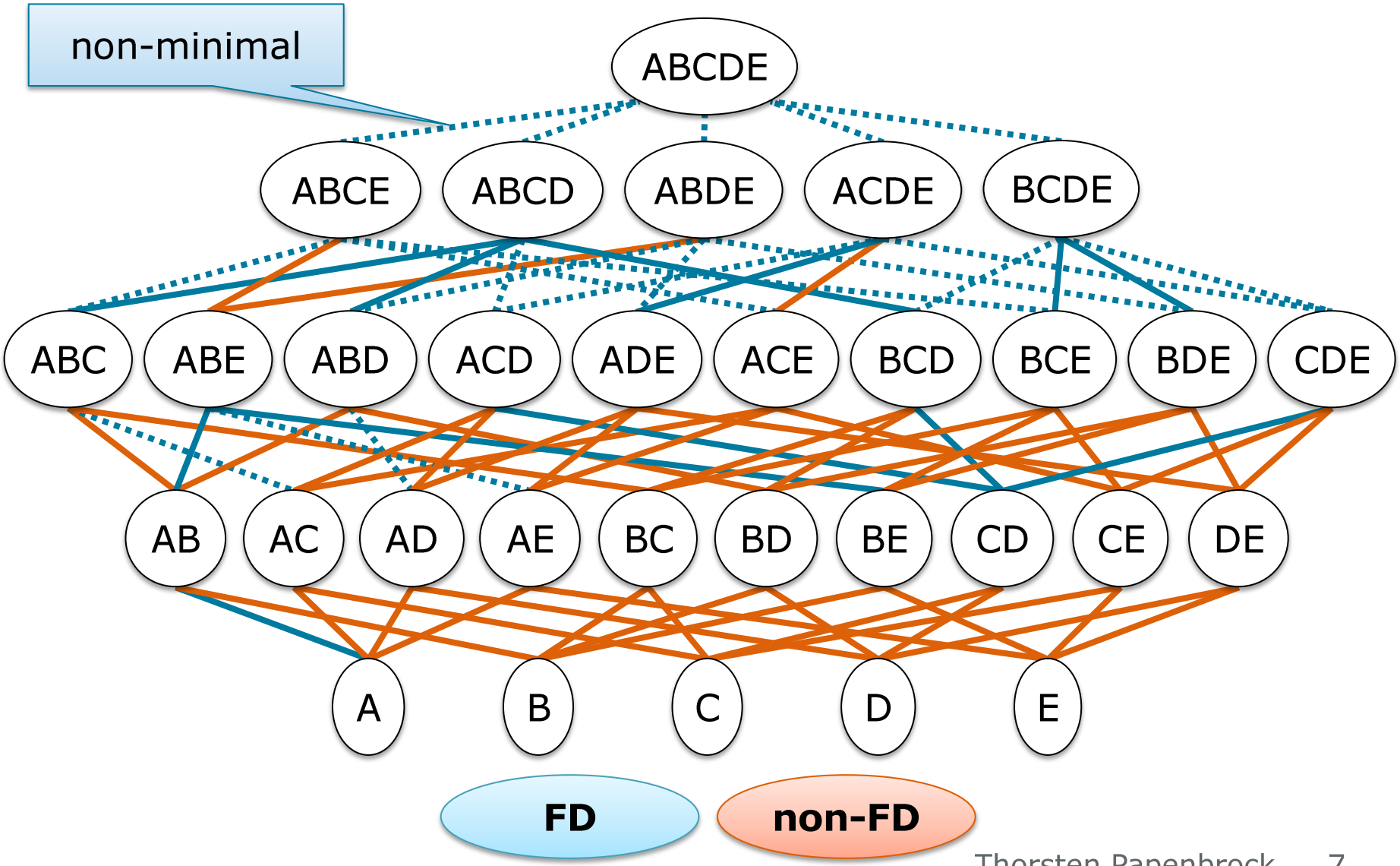
Motivation
FD Discovery



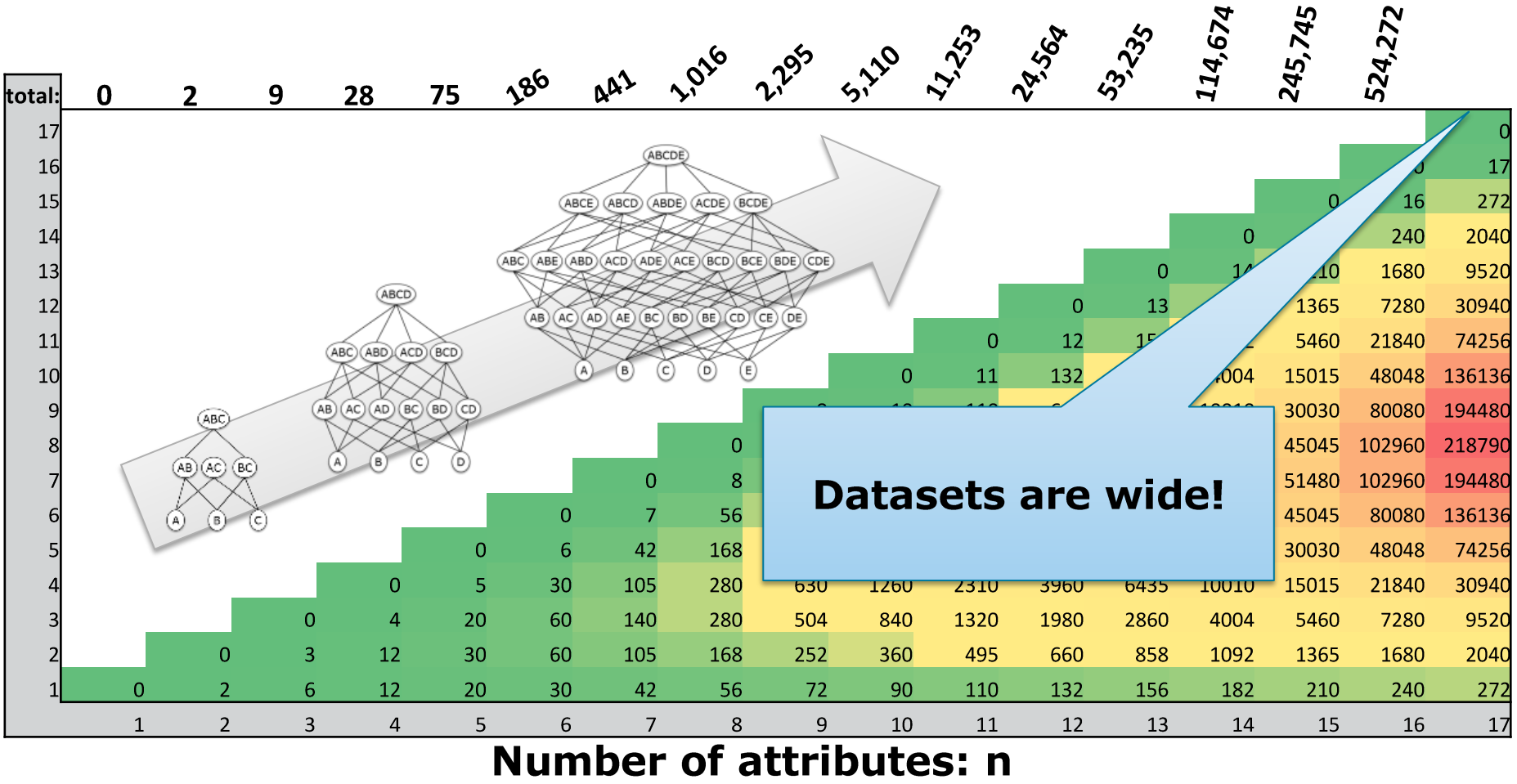
Motivation
FD Discovery



Motivation
FD Discovery



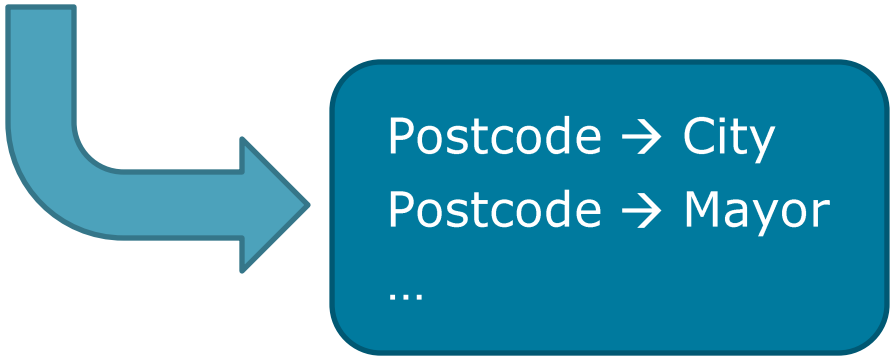
Motivation
FD Discovery



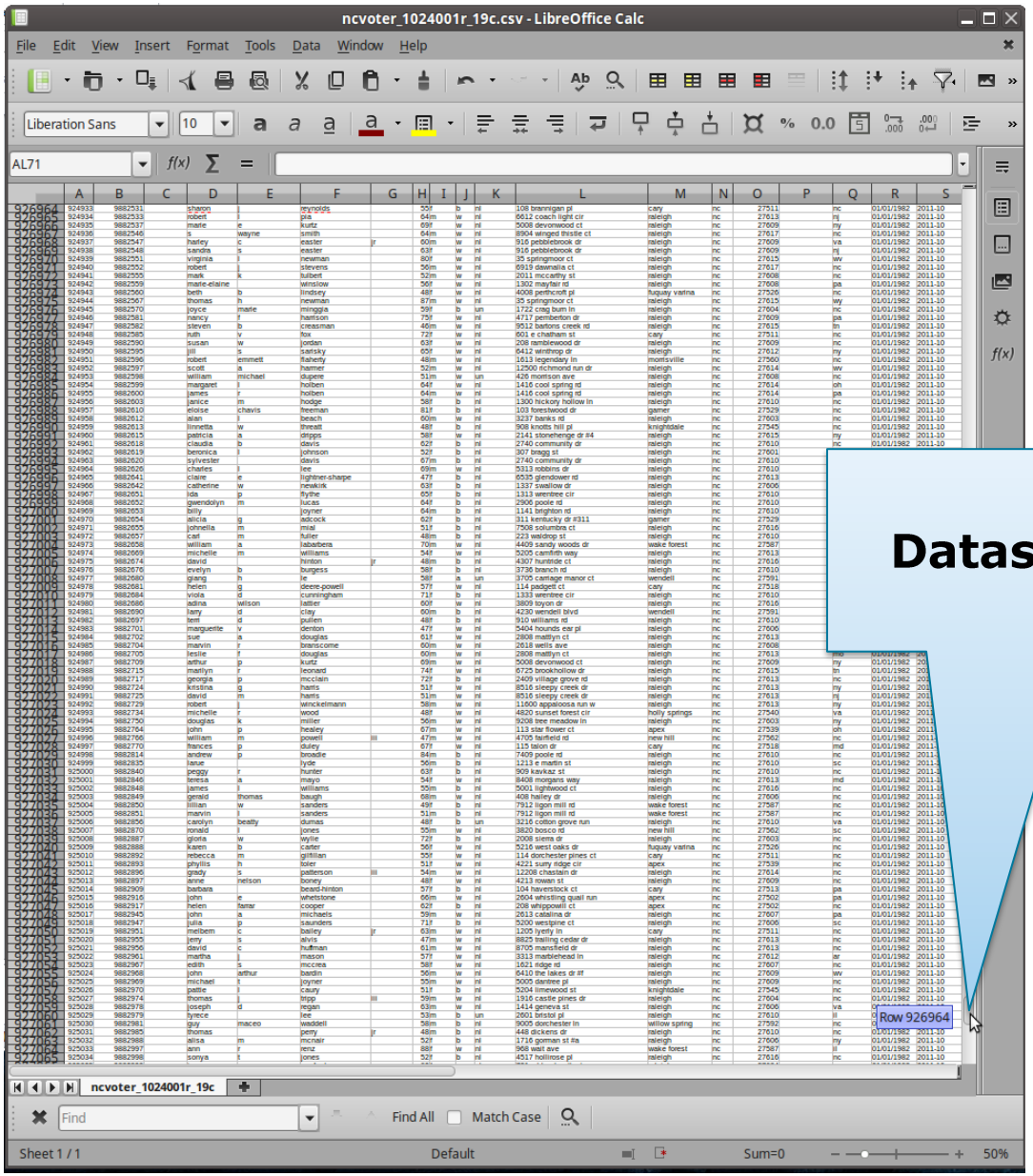
Motivation
FD Discovery Alternative

Name	Surname	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	Cone	01069	Dresden	Orosz
Thomas	Cone	14482	Potsdam	Jakobs
Mike	Moore	60329	Frankfurt	Feldmann

- Surname, Postcode, City, Mayor ↛ Name
- Name, Postcode, City, Mayor ↛ Surname
- Surname ↛ Name, Postcode, City, Mayor



Motivation
FD Discovery Alternative

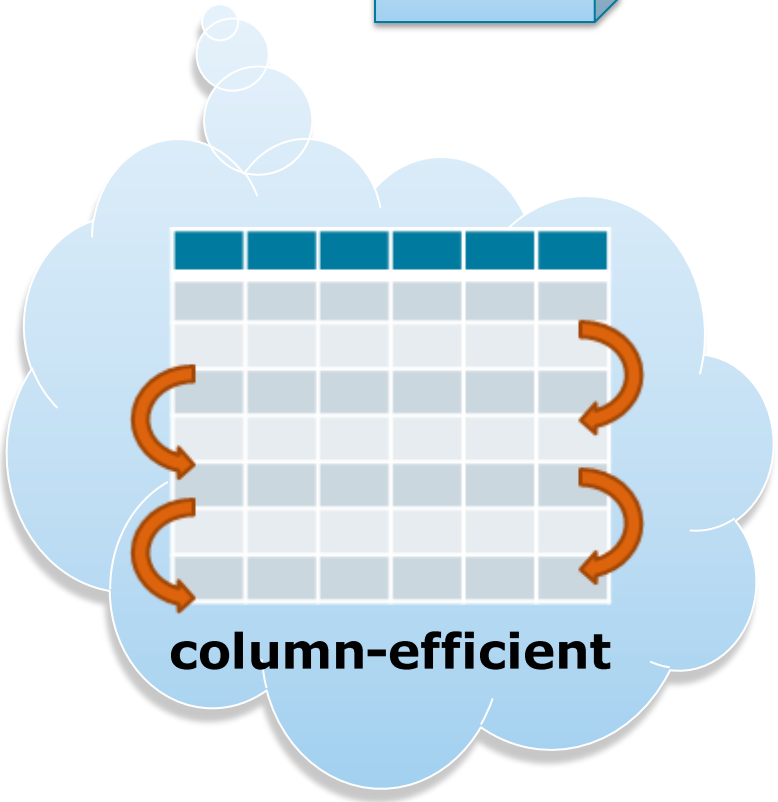


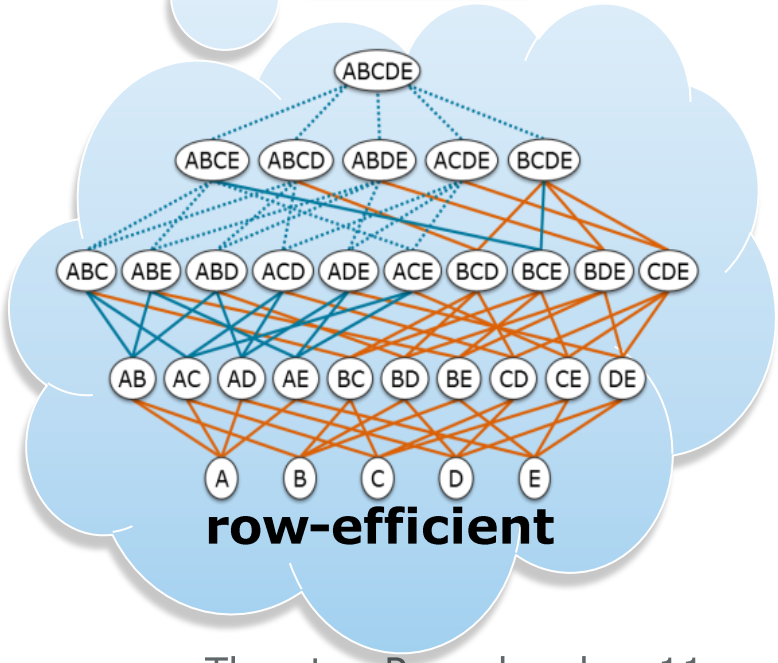
Idea
Going Hybrid

FD Discovery Algorithms

[Flach1999]
Fdep
[Lopes2000]
DepMiner
[Wyss2001]
FastFDs

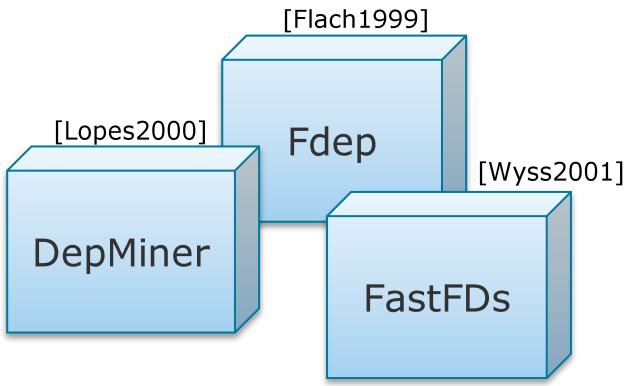
[Huhtala1999]
Tane
[Yao2008]
FDMine
[Novelli2001]
Fun
DFD
[Abedjan2014]

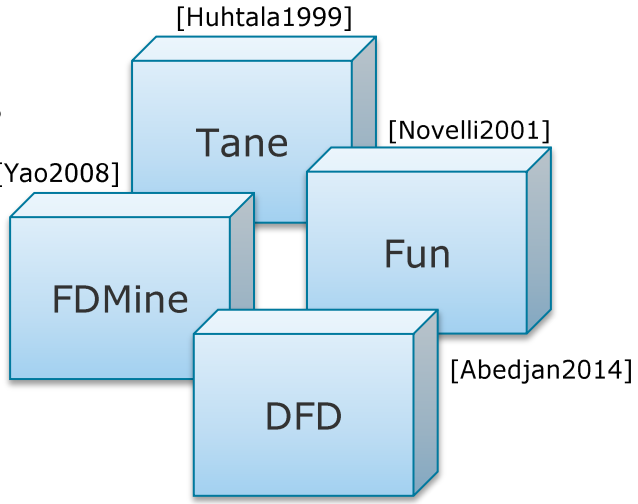




Idea
Going Hybrid

FD Discovery Algorithms





© http://4electric.de

© http://indiancarsbikes.in

Metanome

Idea
Going Hybrid

HPI Hasso Plattner Institut

FD Discovery Algorithms

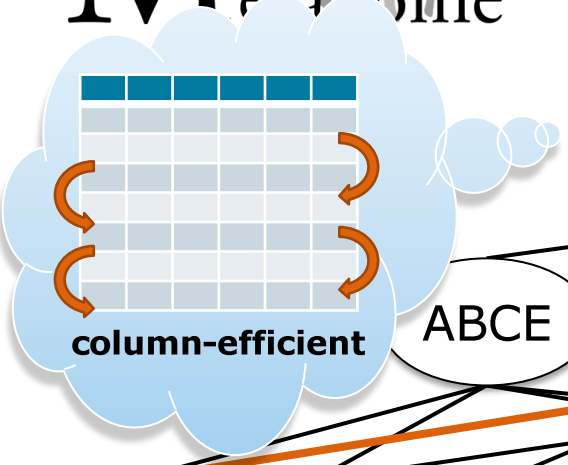
[Flach1999]
Fdep
[Lopes2000]
DepMiner
[Wyss2001]
FastFDs

HyFD

[Huhtala1999]
Tane
[Novelli2001]
Fun
[Yao2008]
FDMine
DFD
[Abedjan2014]

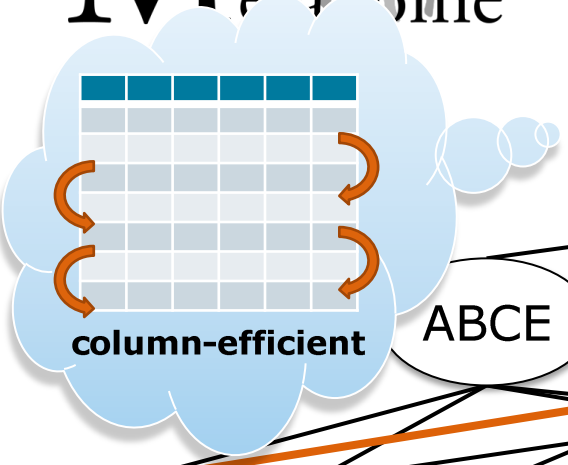
© www.expresswaytoyota.com

The Hybrid Algorithm



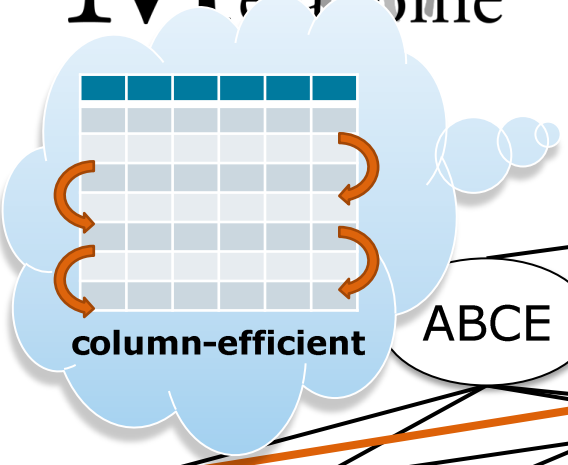
ABC ↛ DE

HyFD
The Hybrid Algorithm



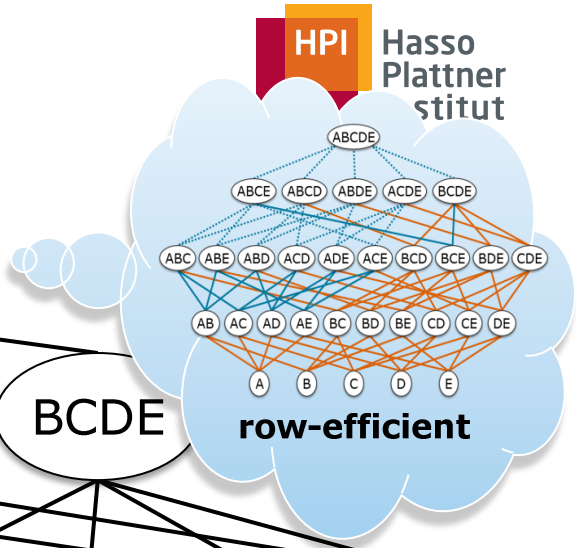
AD ↛ BCE

The Hybrid Algorithm

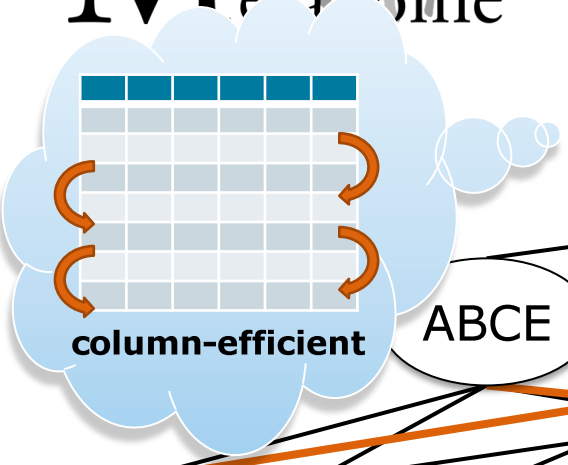


BE ↛ ACD

HyFD
The Hybrid Algorithm

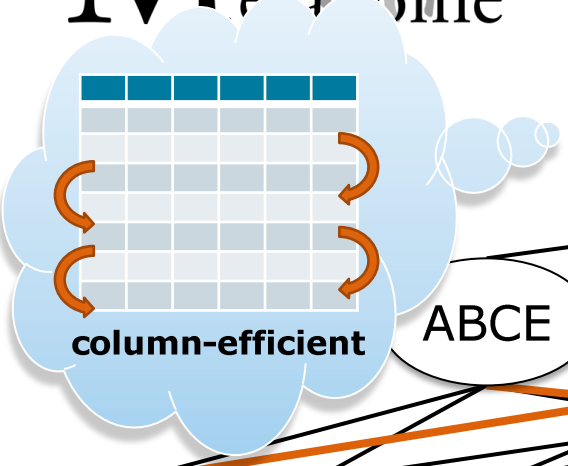


The Hybrid Algorithm



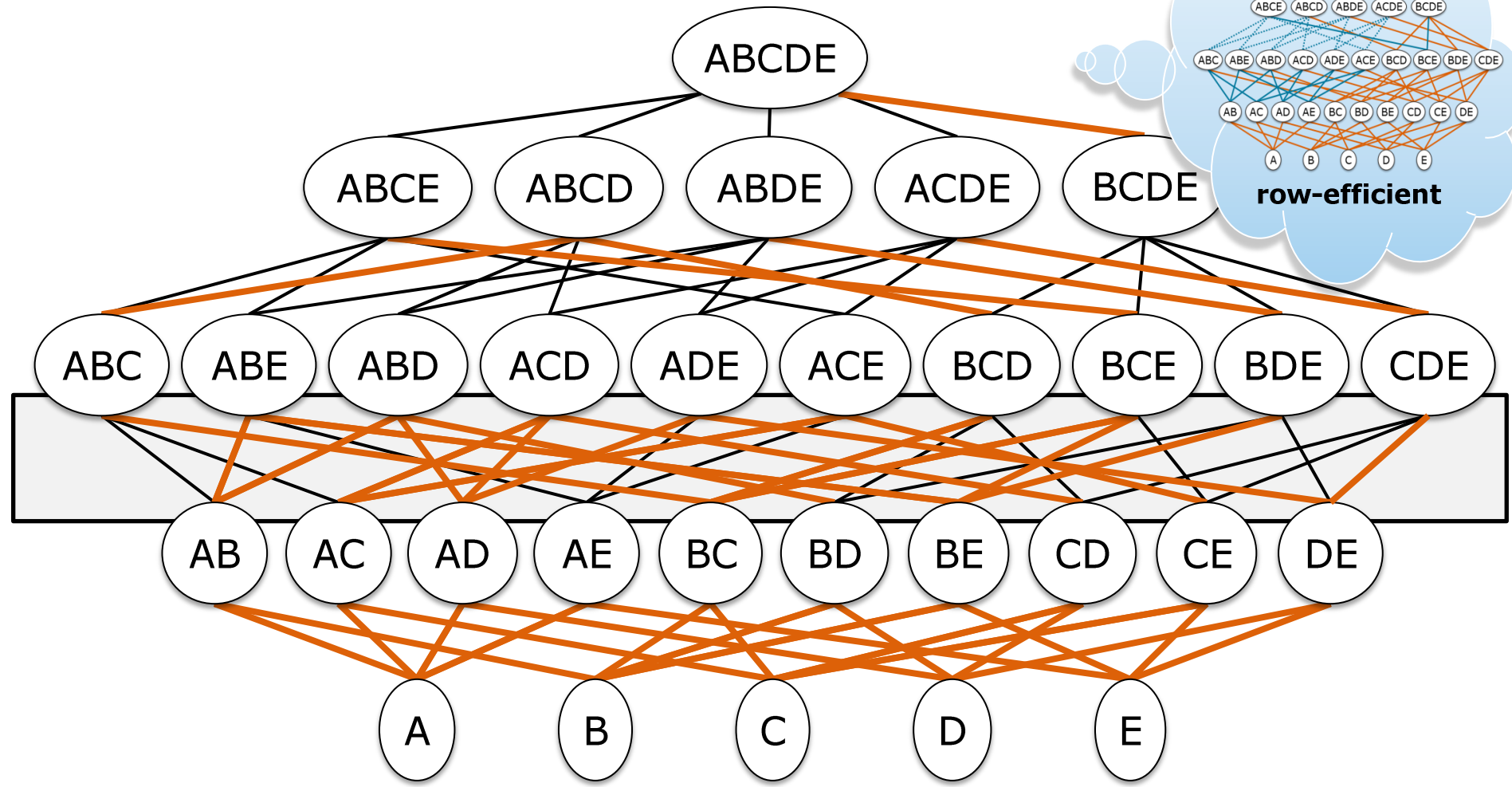
BCDE ↛ A

The Hybrid Algorithm



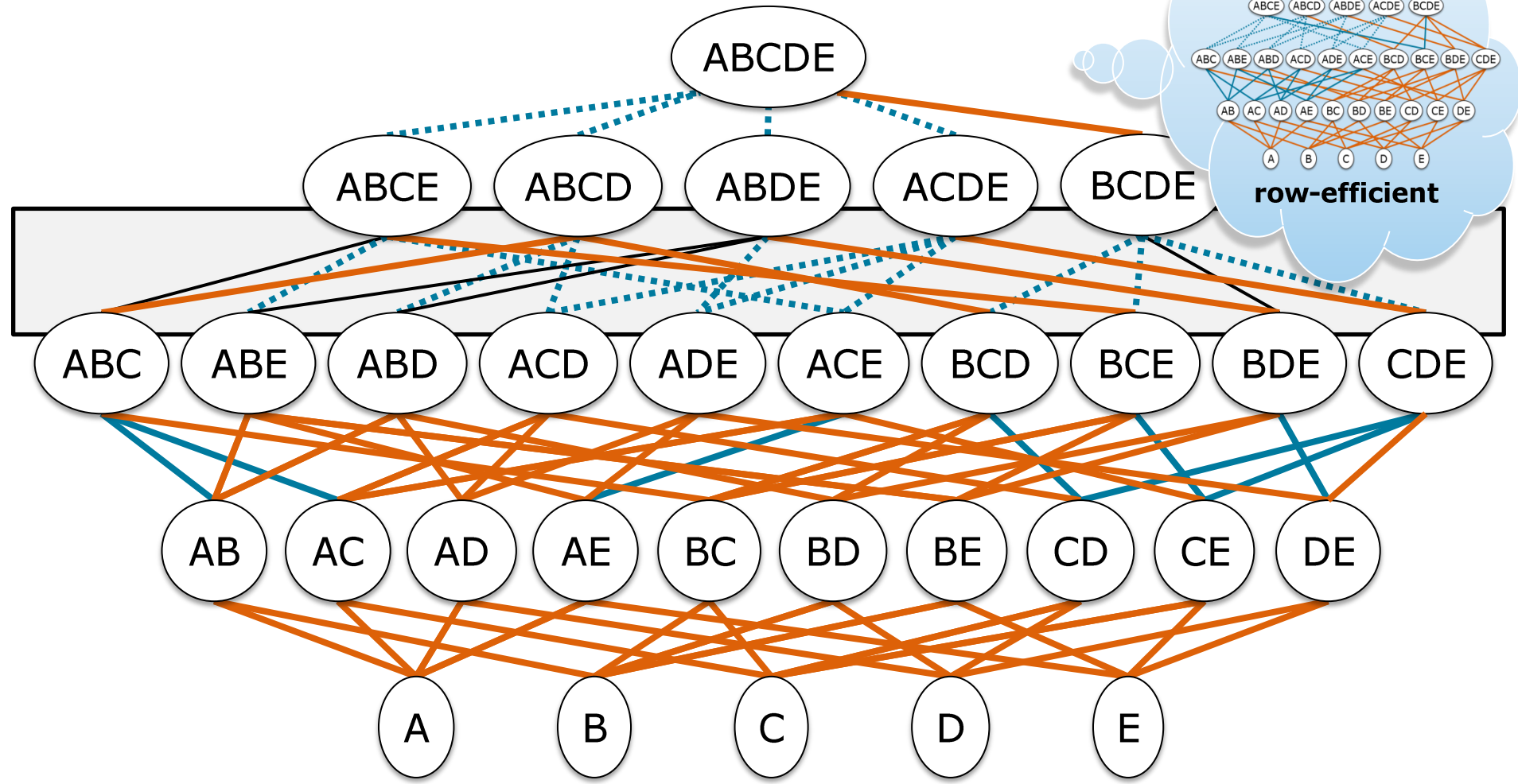
DE ↛ C

HyFD
The Hybrid Algorithm



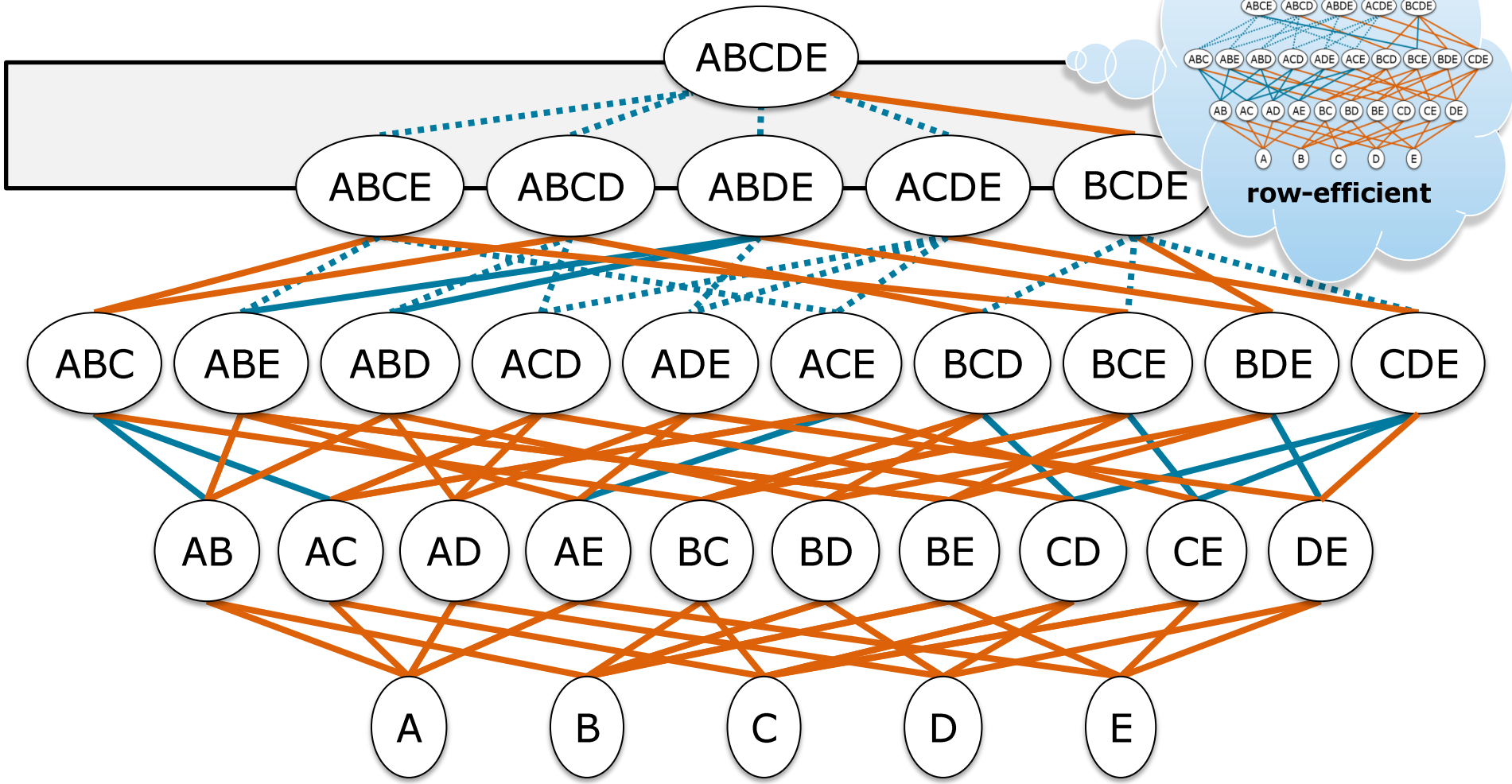
Metanome

HyFD
The Hybrid Algorithm

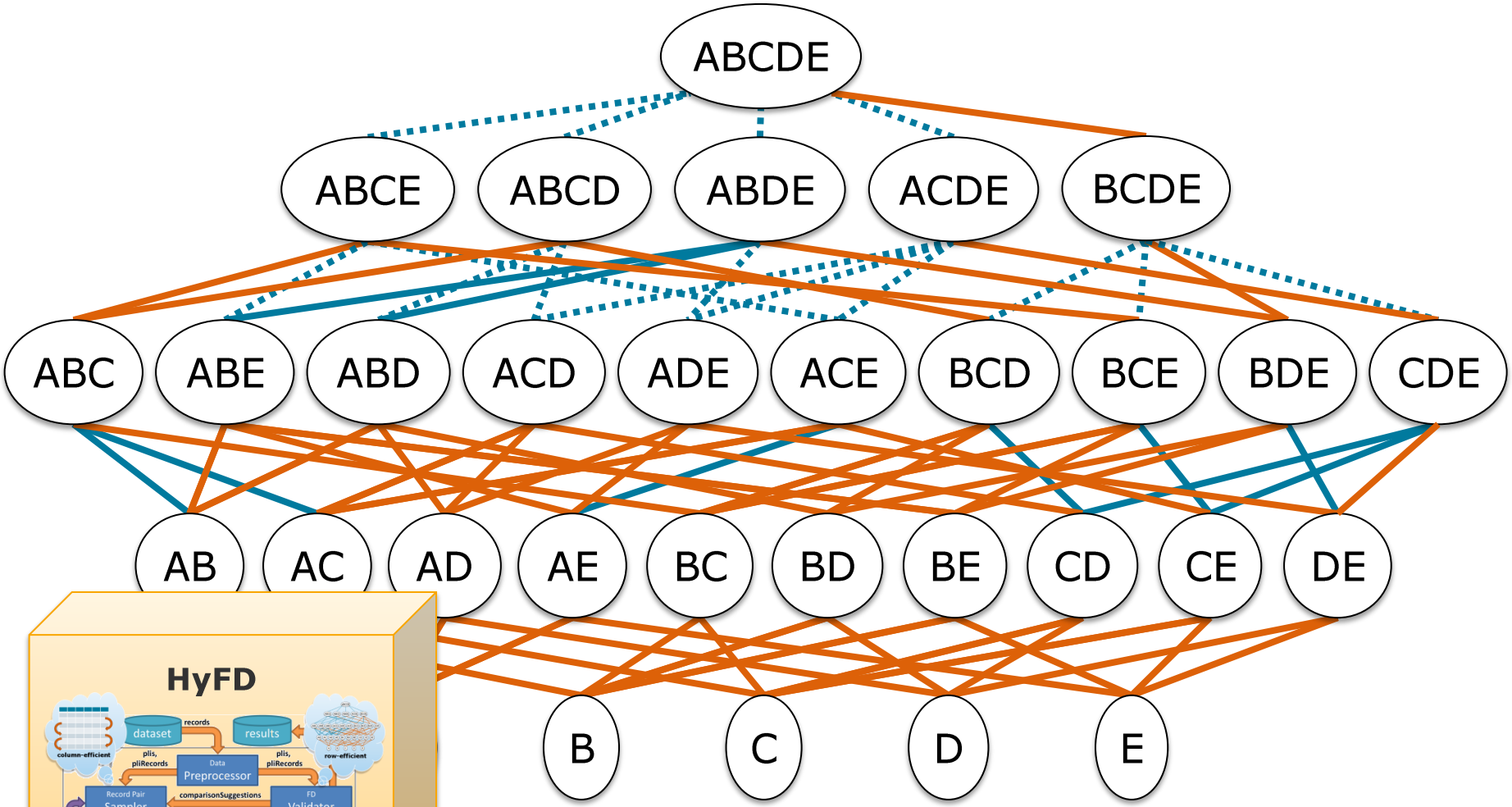


Metanome

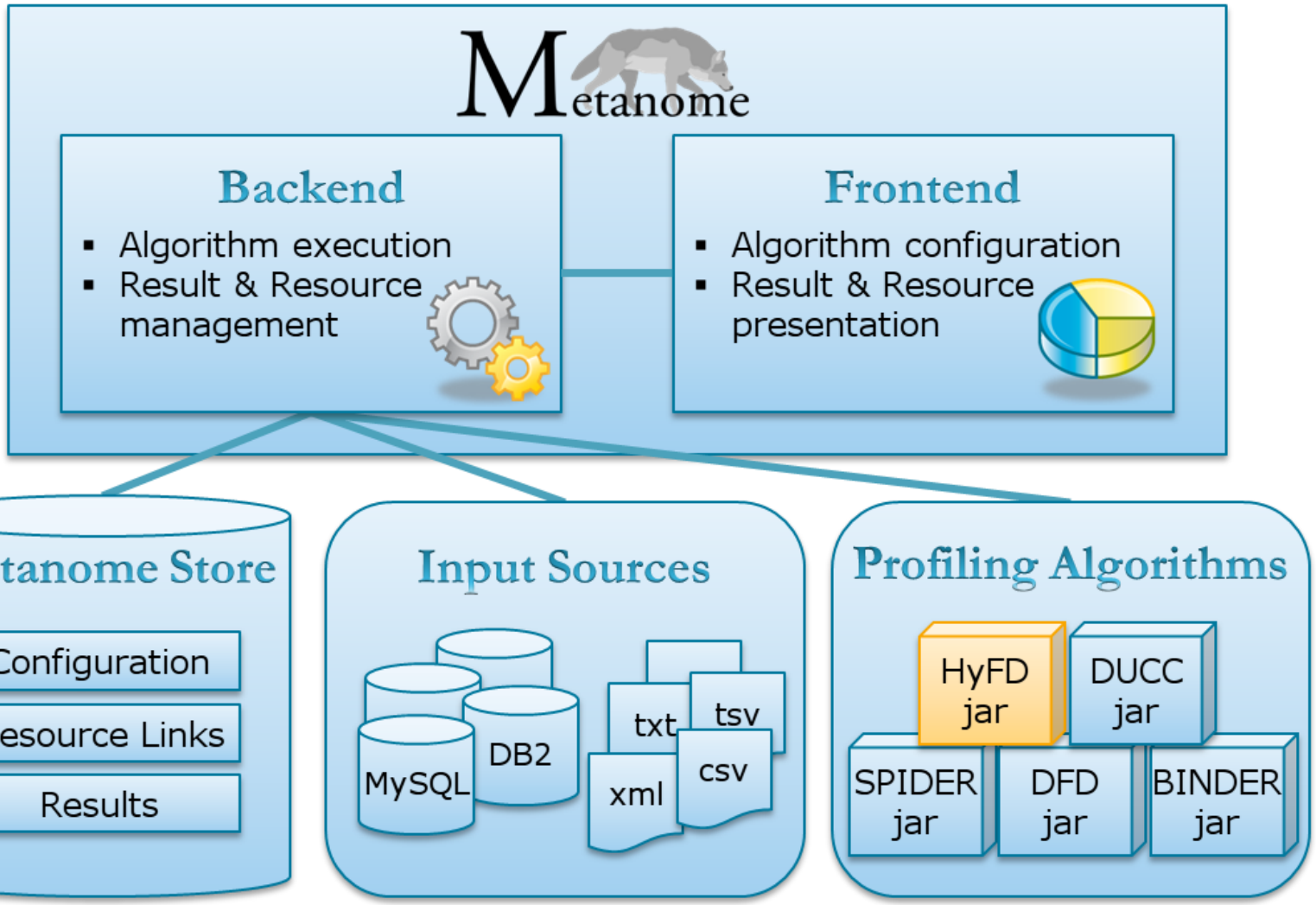
HyFD
The Hybrid Algorithm



HyFD
The Hybrid Algorithm

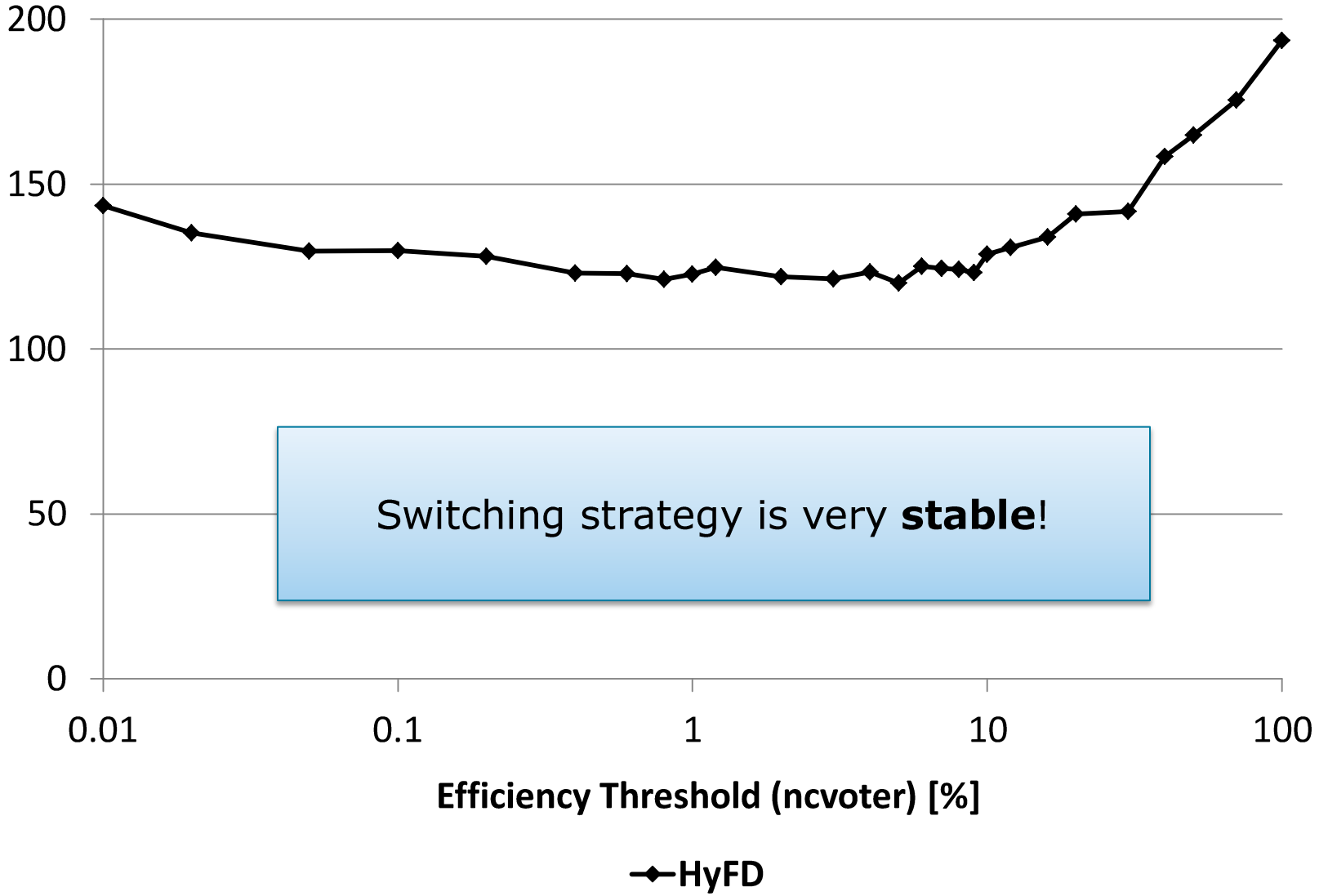


Experimental Evaluation
Metanome Profiling Tool



www.metanome.de

Experimental Evaluation
Switch Criterion



Experimental Evaluation
Datasets

Dataset	Columns [#]	Rows [#]	Size [KB]	FDs [#]	TANE [7]	FUN [15]	FD_MINE [21]	DFD [1]	DEP-MINER [12]	FASTFDS [20]	FDEP [6]	HyFD
iris	5	150	5	4	1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**	**0.1s**
balance-scale	5	625	7	1	1.2	**0.1**	0.2	0.3	0.3	0.3	0.2	**0.1s**
chess	7	28,056	519	1	2.9	1.1	3.8	**1.0**	174.6	164.2	125.5	**0.2s**
abalone	9	4,177	187	137	2.1	**0.6**	1.8	1.1	3.0	2.9	3.8	**0.2s**
nursery	9	1								8.9	46.8	**0.5s**
breast-cancer	11									1.1	**0.5**	**0.2s**
bridges	13									0.6	**0.2**	**0.1s**
echocardiogram	13									0.5	**0.2**	**0.1s**
adult	14	4								3.8	860.2	**1.1s**
letter	17	2								5.5	291.3	**3.4s**
ncvoter	19									1.9	**1.1**	**0.4s**
hepatitis	20	155	8	8,250	12.2	175.9	ML	326.7	5576.5	9.5	**0.8**	**0.6s**
horse	27	368	25	128,726	457.0	TL	ML	TL	TL	385.8	**7.2**	**7.1s**
fd-reduced-30	30	250,000	69,581	89,571	**41.1**	77.7	ML	TL	377.2	382.4	TL	513.0s
plista	63	1,000	568	178,152	ML	ML	ML	TL	TL	TL	**26.9**	**21.8s**
flight	109	1,000	575	982,631	ML	ML	ML	TL	TL	TL	**216.5**	**53.4s**
uniprot	223	1,000	2,439	unknown	ML	ML	ML	TL	TL	TL	ML	**>5254.7s**

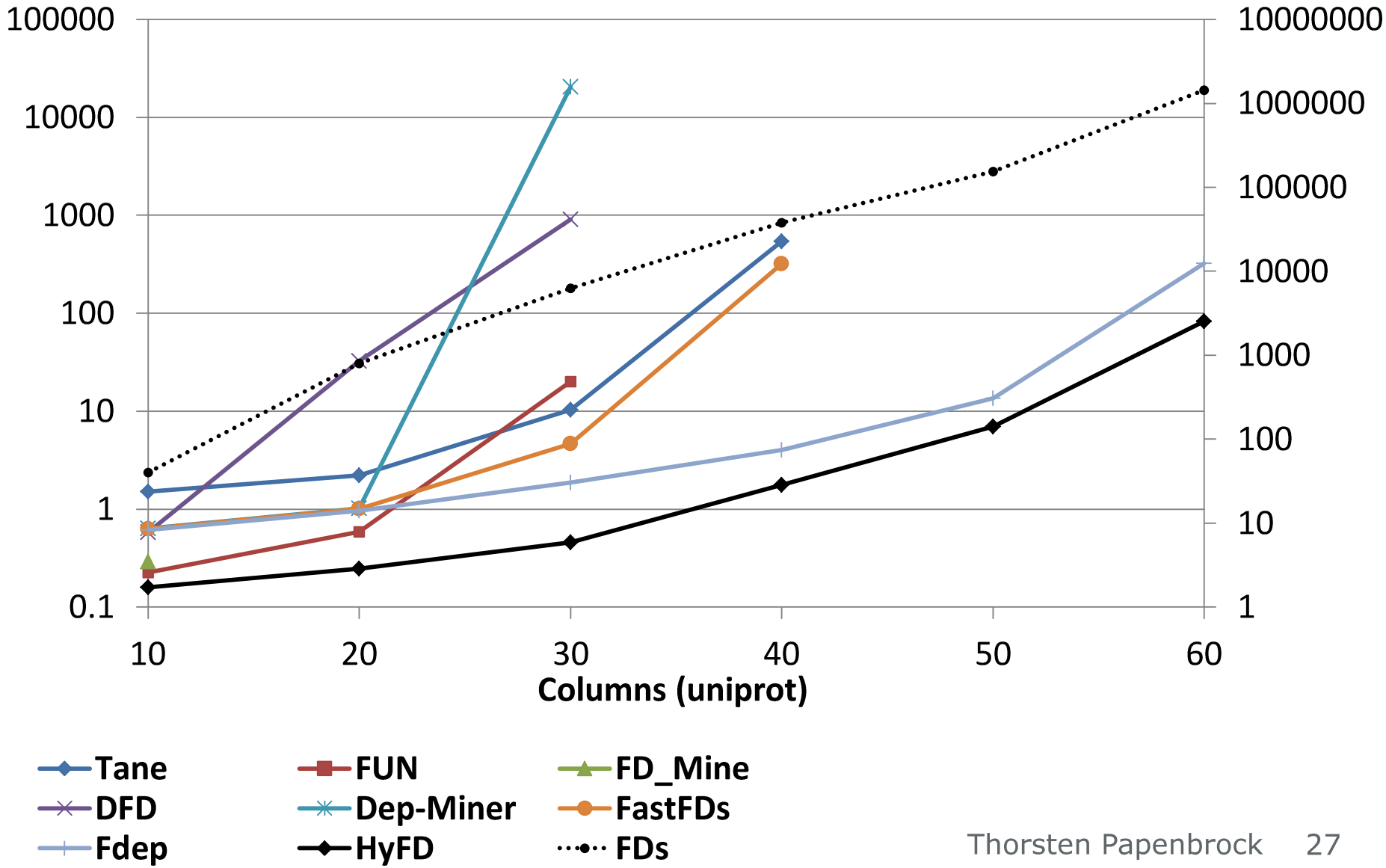
Results larger than 1,000 FDs are only counted **TL**: time limit of 4 hours exceeded **ML**: memory limit of 100GB exceeded

Faster on all real-world datasets!

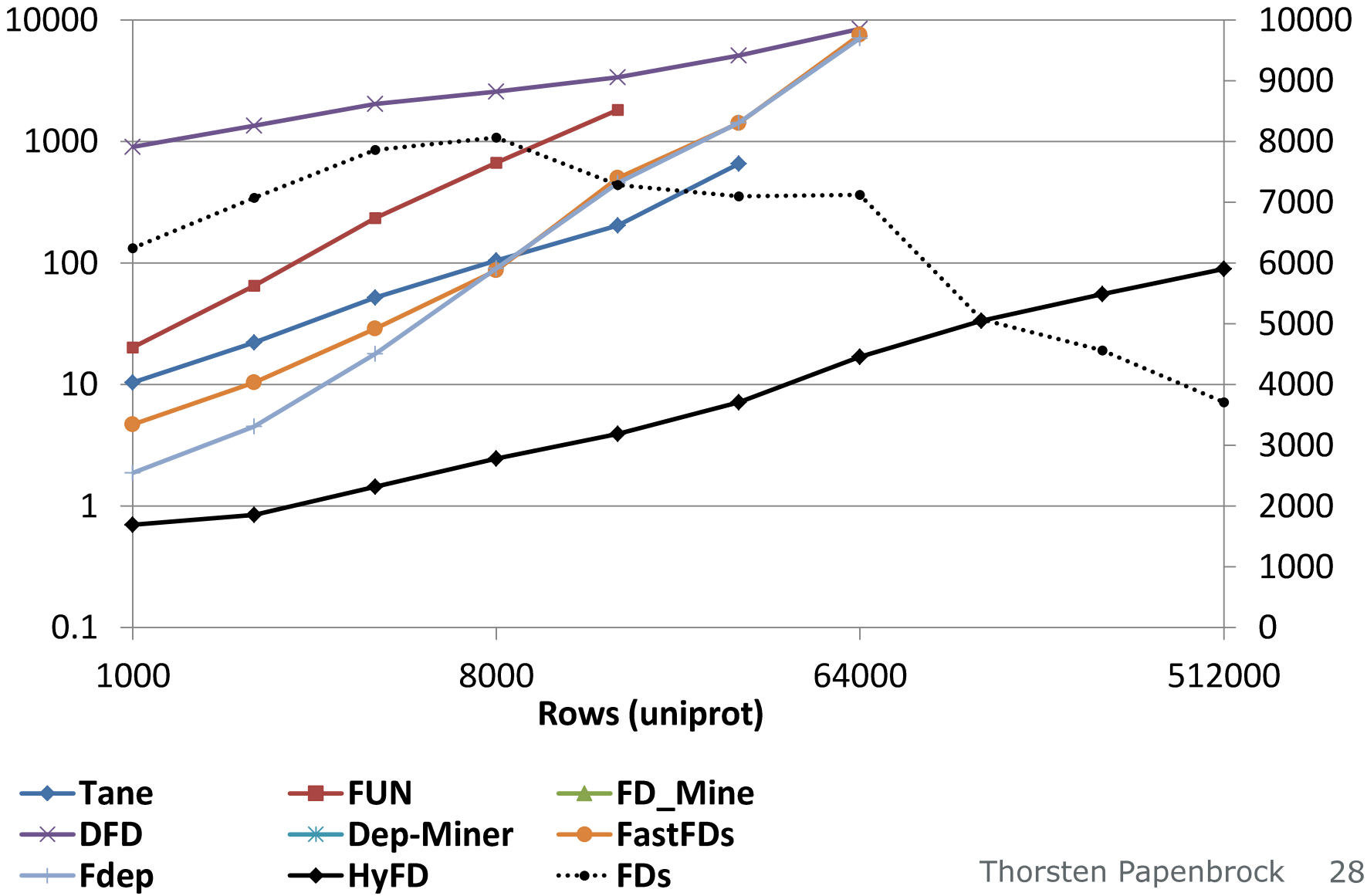
Dataset	Cols [#]	Rows [#]	Size [MB]	FDs [#]	HyFD [s/m/h/d]	
TPC-H.lineitem	16	6 m	1,051	4 k	**39 m**	4 m
PDB.POLY_SEQ						
PDB.ATOM_SITE						
SAP_R3.ZBC00DT						
SAP_R3.ILOA						
SAP_R3.CE4HI01						
NCVoter.statewide						
CD.cd	107	10 k	5	36 k	**5 s**	3 m

Larger datasets than ever before!

Experimental Evaluation
Column Scalability

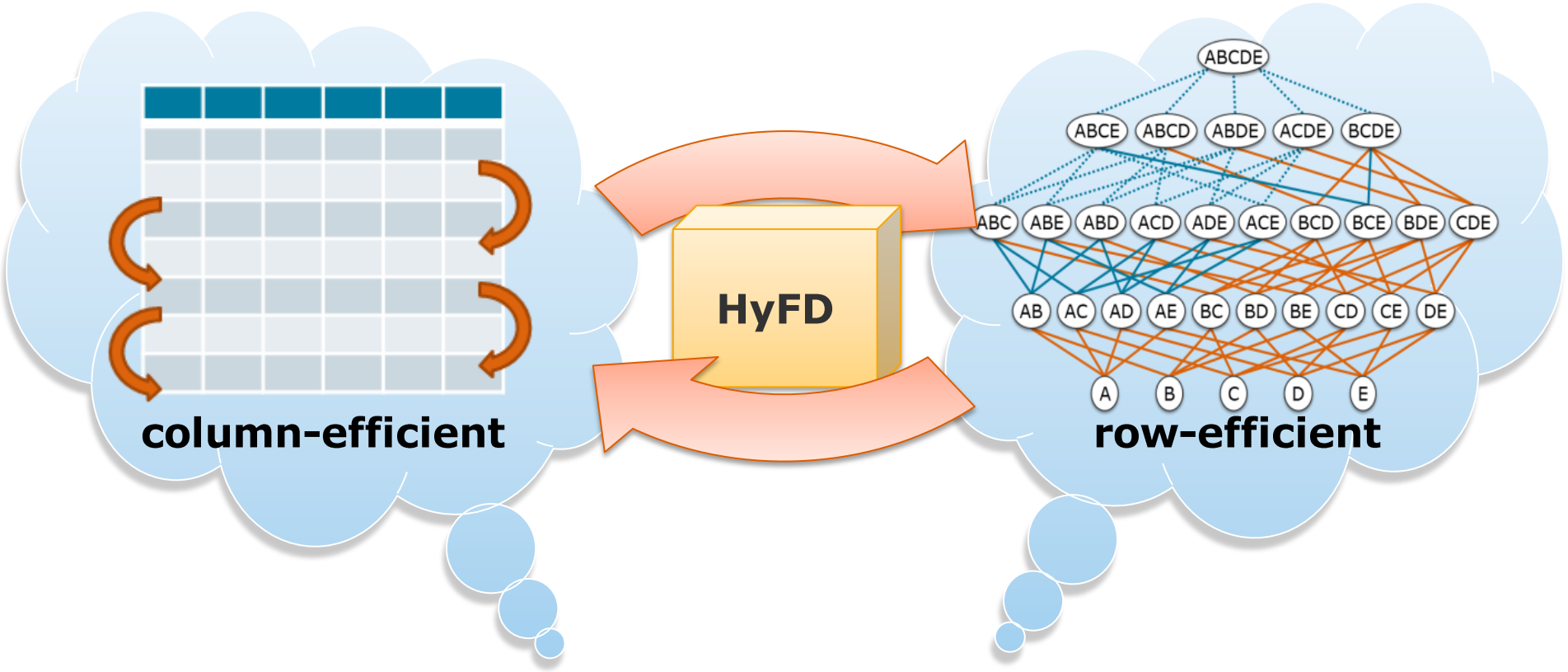


Experimental Evaluation
Row Scalability



Metanome

HPI Hasso Plattner Institut



A Hybrid Approach to Functional Dependency Discovery

Thorsten Papenbrock and Felix Naumann