



Clean-IT Sustainable Data Profiling

Dr. Thorsten Papenbrock
Information Systems Group
Hasso Plattner Institute

Motivation

Sustainable Computing



Sustainable computing means (i.a.) saving energy.

Saving energy means minimizing work.

Minimizing work means avoiding unnecessary computation steps, such as superfluous, needlessly complex or redundant calculations (anything that does not effectively contribute to the final result).

Avoiding unnecessary computation steps decreases the runtime (hence, we usually measure runtime as a sustainability indicator).

**Sustainable
Data Profiling**

Thorsten Papenbrock

Chart 2

Data Profiling Example

Discovering Keys

Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina				
37	Vulpix	Ninetails				
37	Vulpix	Ninetails				
38	Ninetails	null				

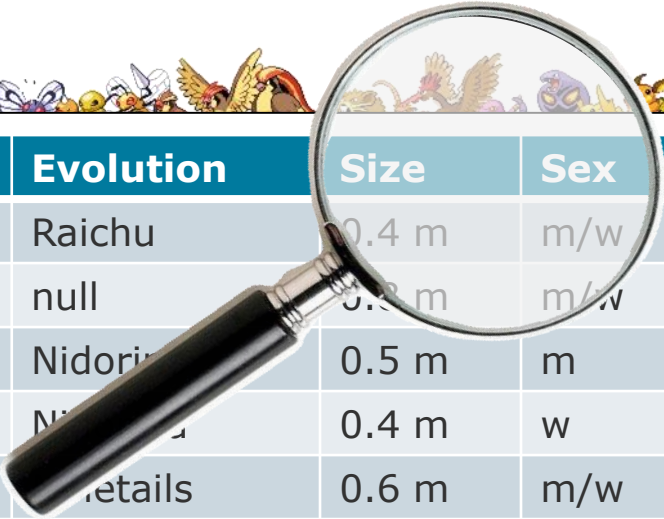
Given a relational instance r of schema R , a **unique column combination** X with $X \subseteq R$ is valid in r , iff $\forall t_i, t_j \in r, i \neq j : t_i[X] \neq t_j[X]$.

A database **key** is a set of attributes whose values uniquely identify every tuple in the table.

Such keys are also referred to as **unique column combinations** (UCCs).

Data Profiling Example

Discovering Keys



Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorina	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails	null	1.1 m	m/w	fire	water

Keys are ...

- **important** for data management, query answering, machine learning etc.
- often not explicitly stored and, hence, **missing**.

Data Profiling Example

Discovering Keys

Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails	null	1.1 m	m/w	fire	water
63	Abra	Kadabra	0.9 m	m/w	psychic	ghost
64	Kadabra	Alakazam	1.3 m	m/w	psychic	ghost
65	Alakazam	null	1.5 m	m/w	psychic	ghost
150	Mewtwo	null	2.0 m	null	psychic	ghost

Data Profiling Example

Discovering Keys

Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails	null	1.1 m	m/w	fire	water
63	Abra	Kadabra	0.9 m	m/w	psychic	ghost
64	Kadabra	Alakazam	1.3 m	m/w	psychic	ghost
65	Alakazam	null	1.5 m	m/w	psychic	ghost
150	Mewtwo	null	2.0 m	null	psychic	ghost

Data Profiling Example

Discovering Keys

Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails	null	1.1 m	m/w	fire	water
63	Abra	Kadabra	0.9 m	m/w	psychic	ghost
64	Kadabra	Alakazam	1.3 m	m/w	psychic	ghost
65	Alakazam	null	1.5 m	m/w	psychic	ghost
150	Mewtwo	null	2.0 m	null	psychic	ghost

Data Profiling Example

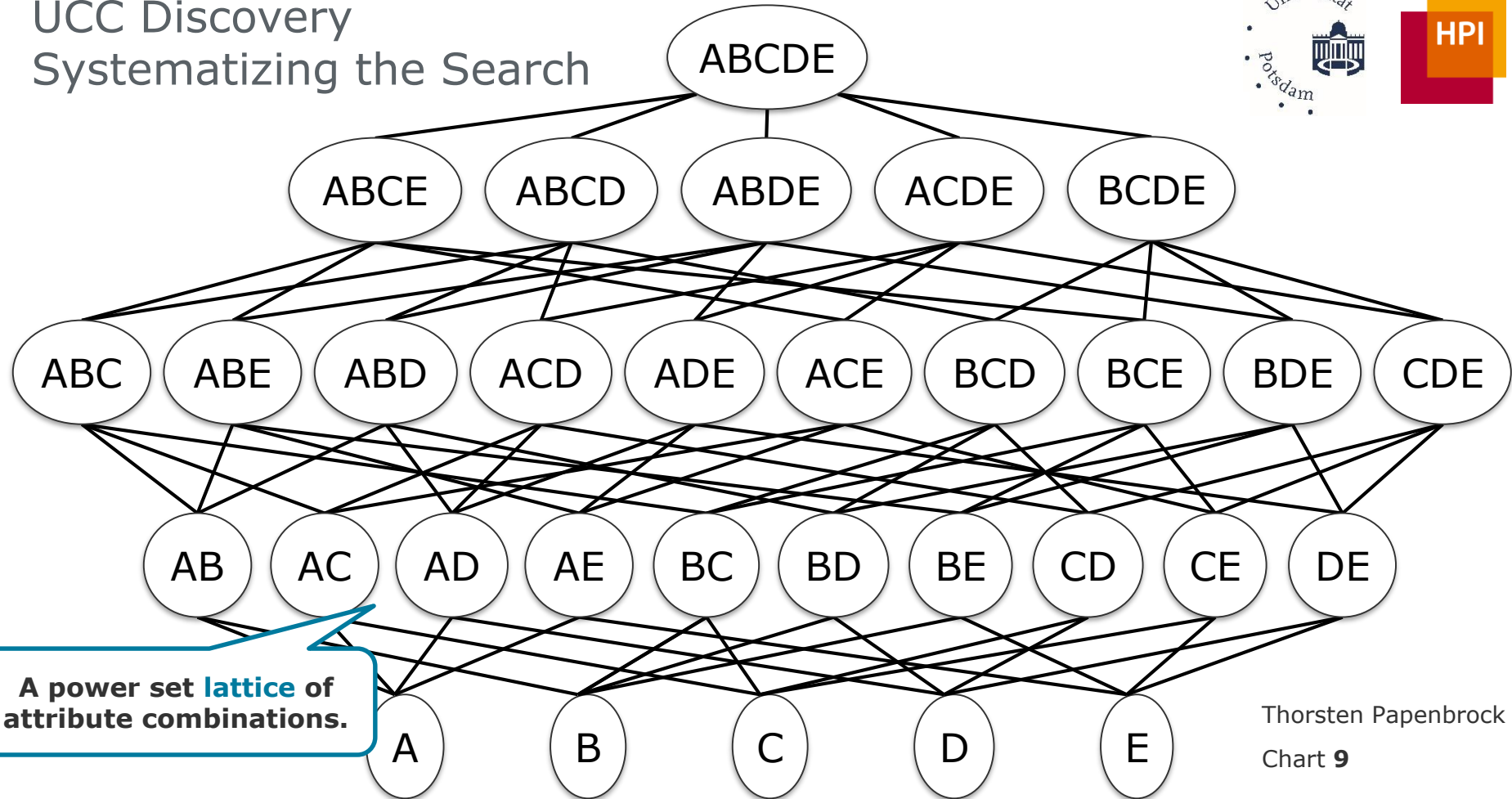
Discovering Keys

Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix			/w	ice	fire
37	Vulpix			/w	fire	water
38	Ninetails	null	1.1 m	m/w	fire	water
63	Abra	Kadabra	0.9 m	m/w	psychic	ghost
64	Kadabra	Alakazam	1.3 m	m/w	psychic	ghost
65	Alakazam	null	1.5 m	m/w	psychic	ghost
150	Mewtwo	null	2.0 m	null	psychic	ghost

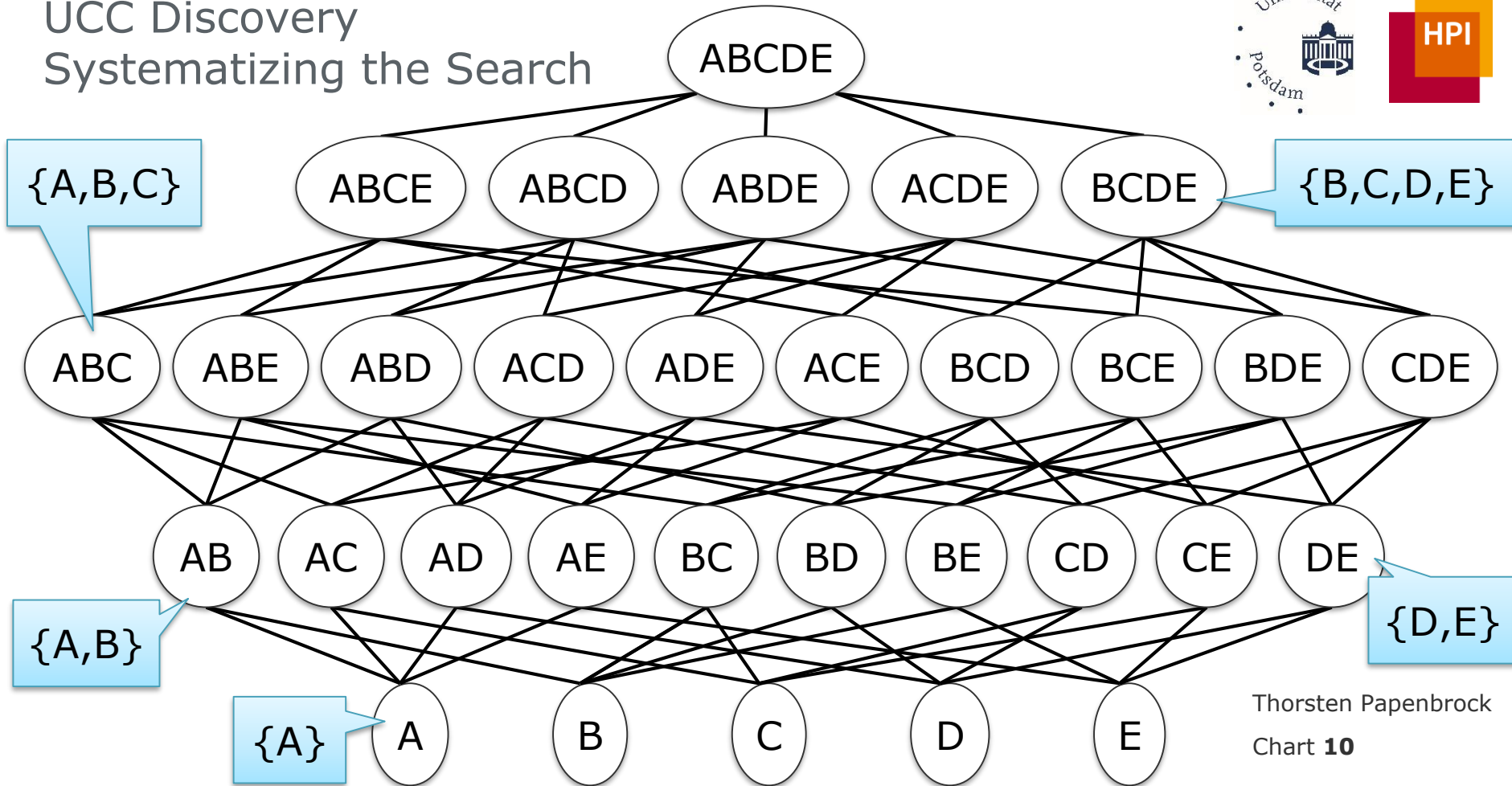
Unique: {Name, Sex, Type}

UCC Discovery

Systematizing the Search

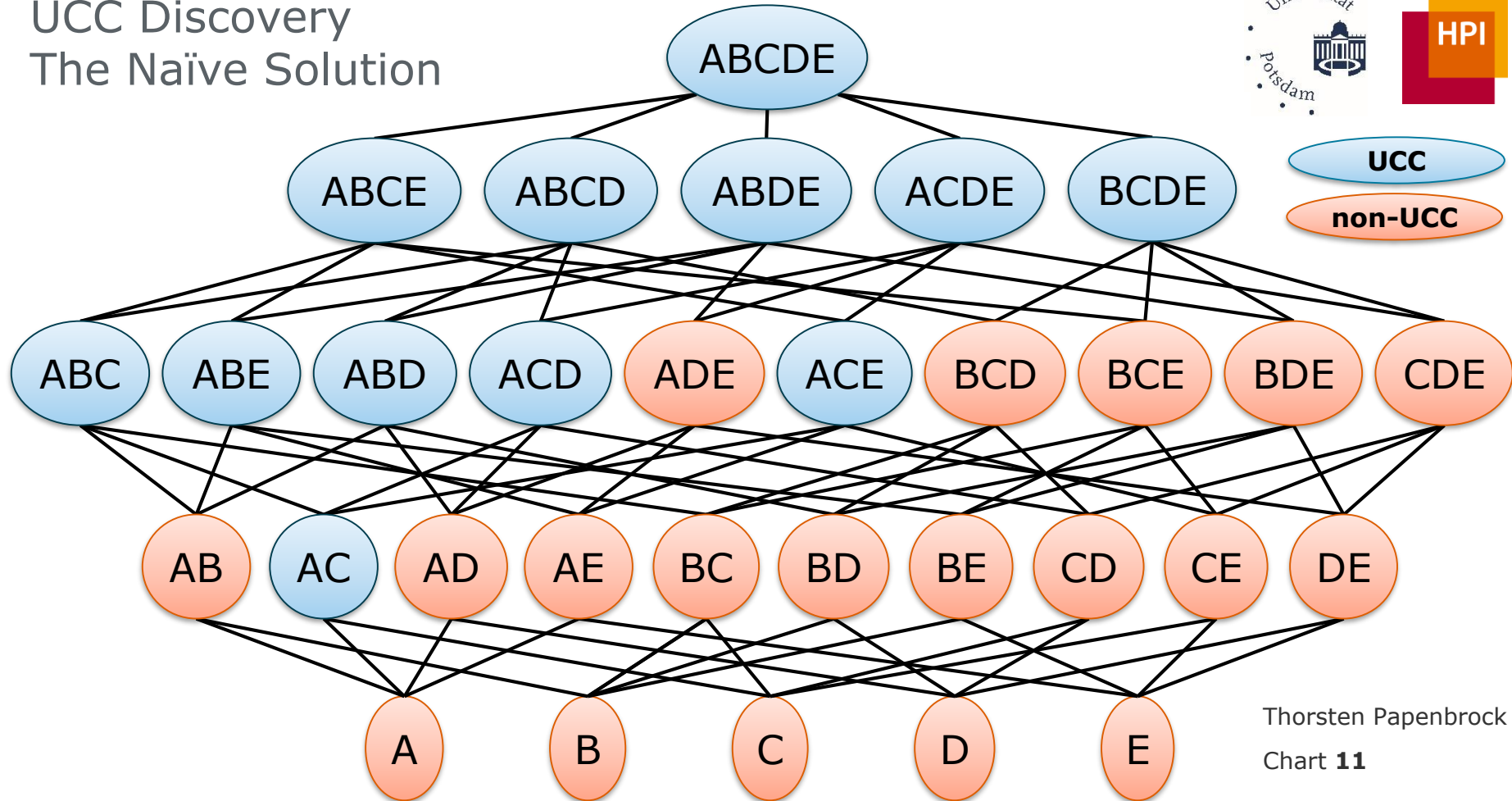


UCC Discovery Systematizing the Search



UCC Discovery

The Naïve Solution

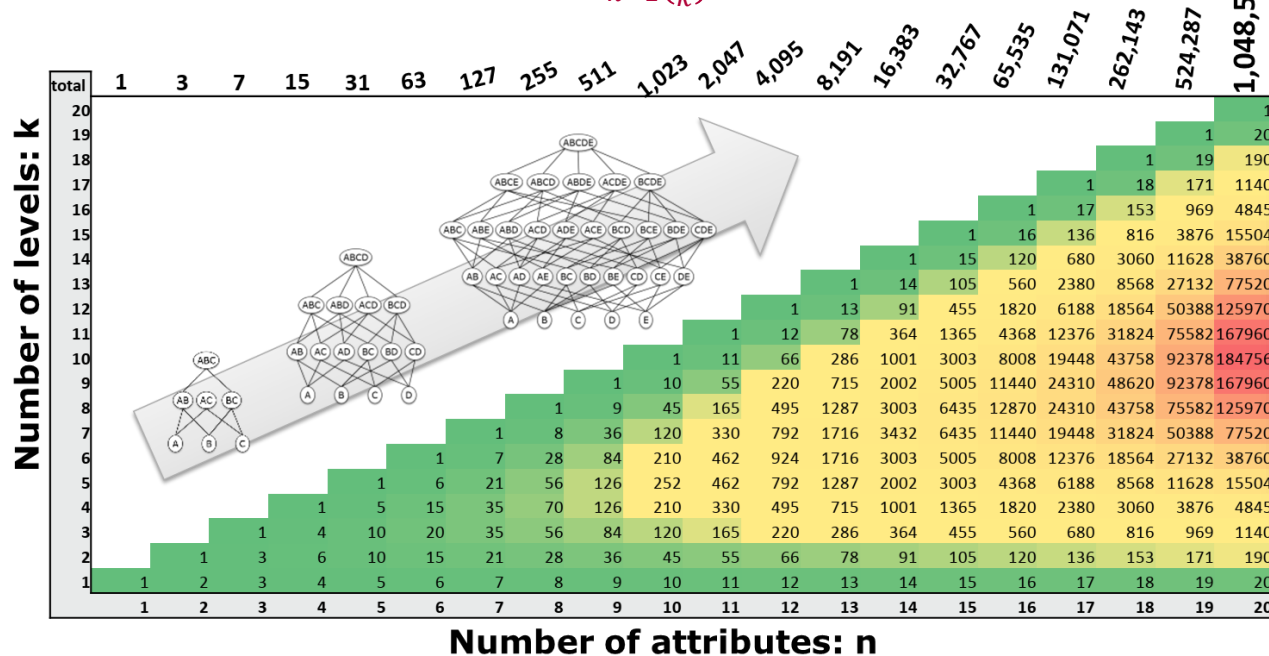


UCC Discovery

The Naïve Solution

ABCDE

Number of UCC candidates: $\sum_{k=1}^n \binom{n}{k} = 2^n - 1$

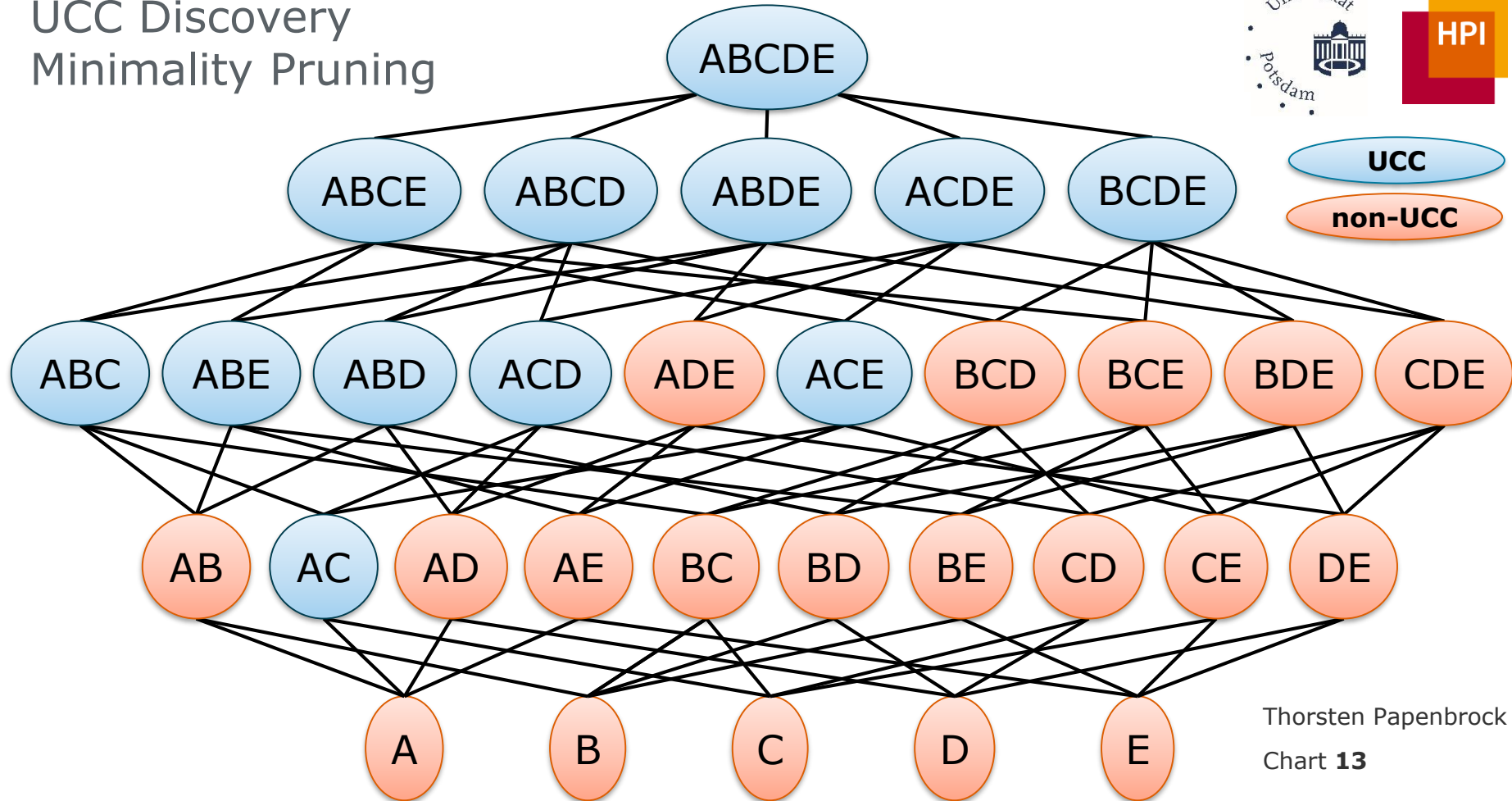


UCC

non-UCC

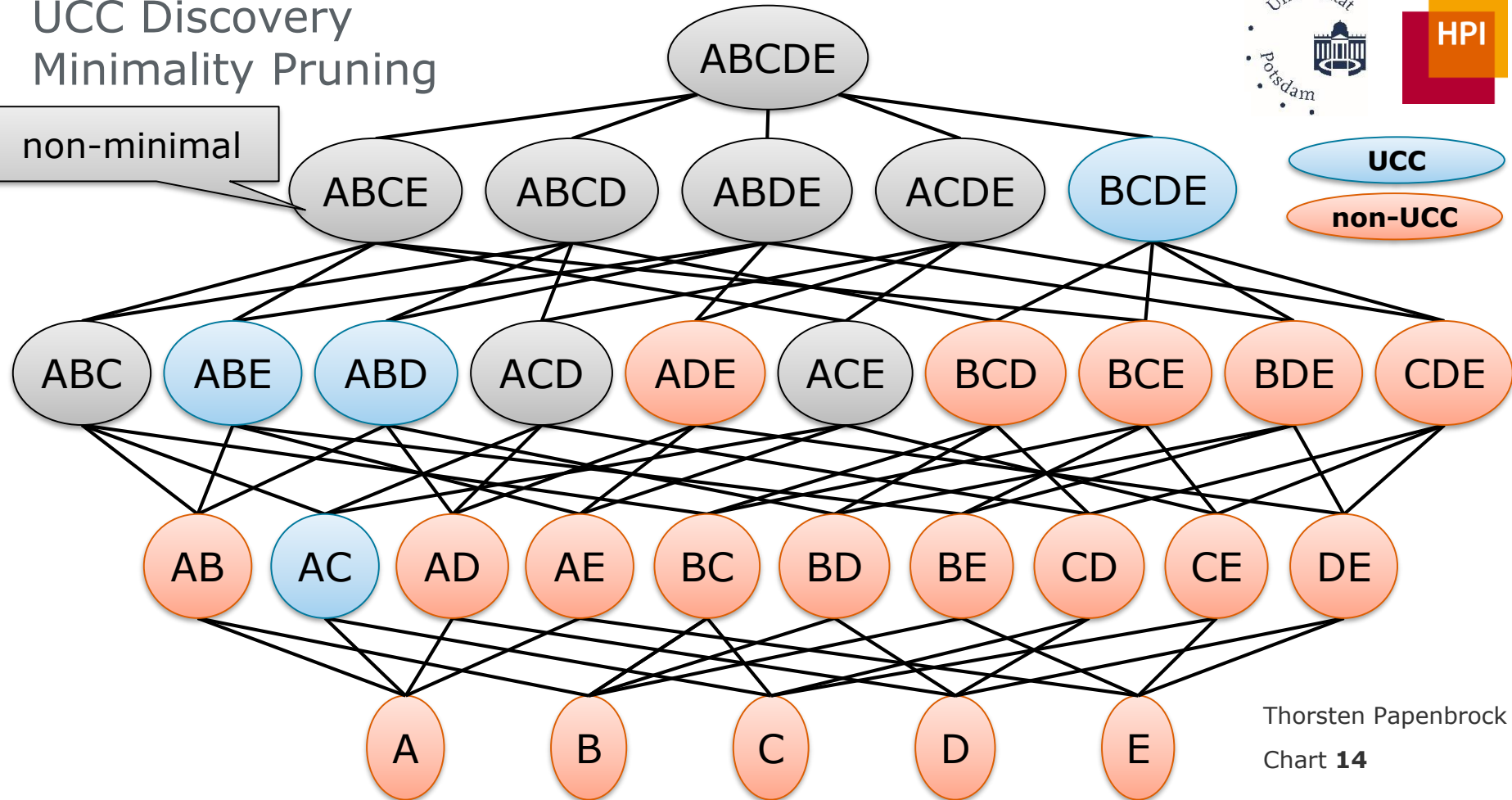


UCC Discovery Minimality Pruning



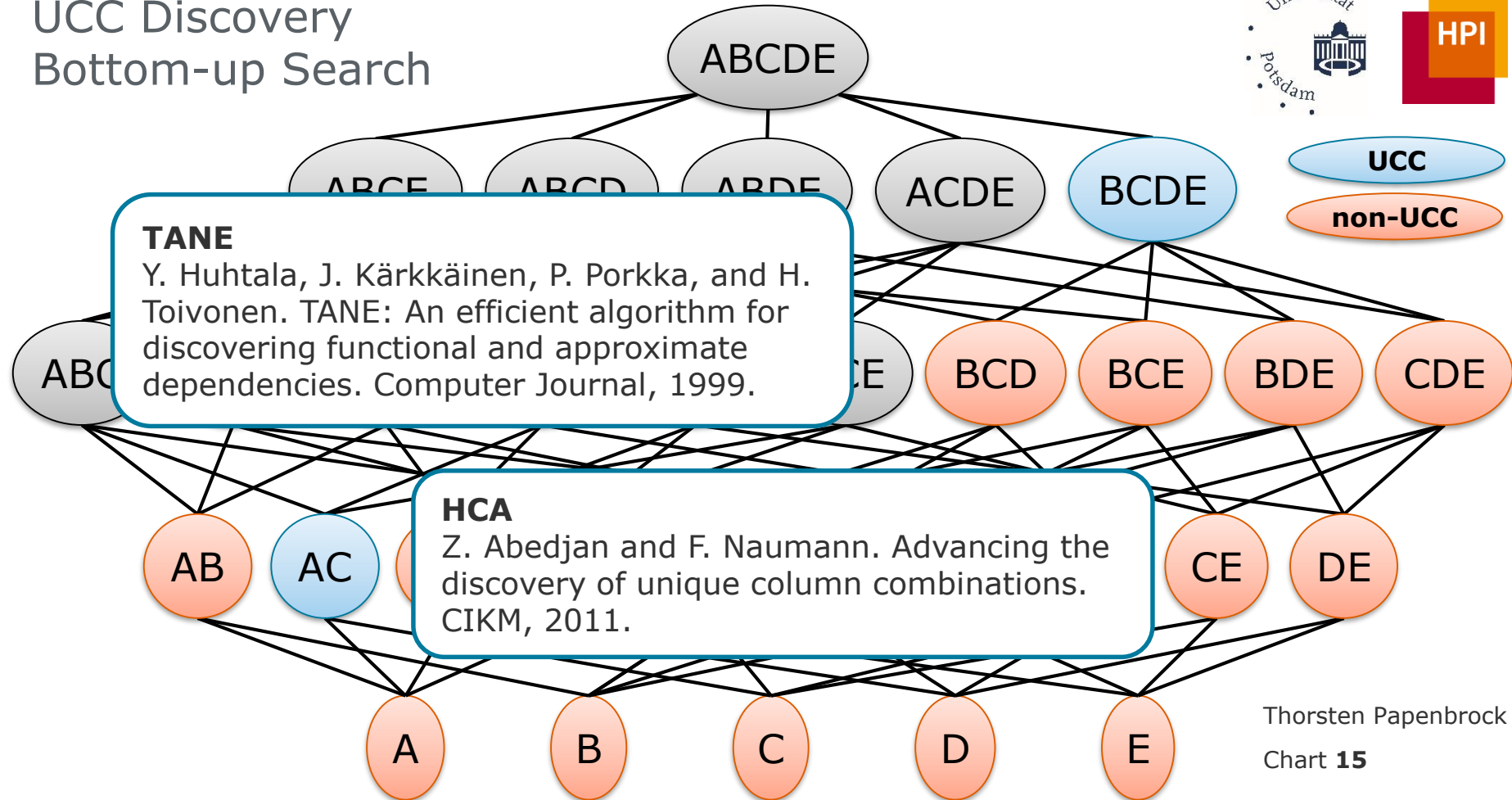
UCC Discovery Minimality Pruning

non-minimal

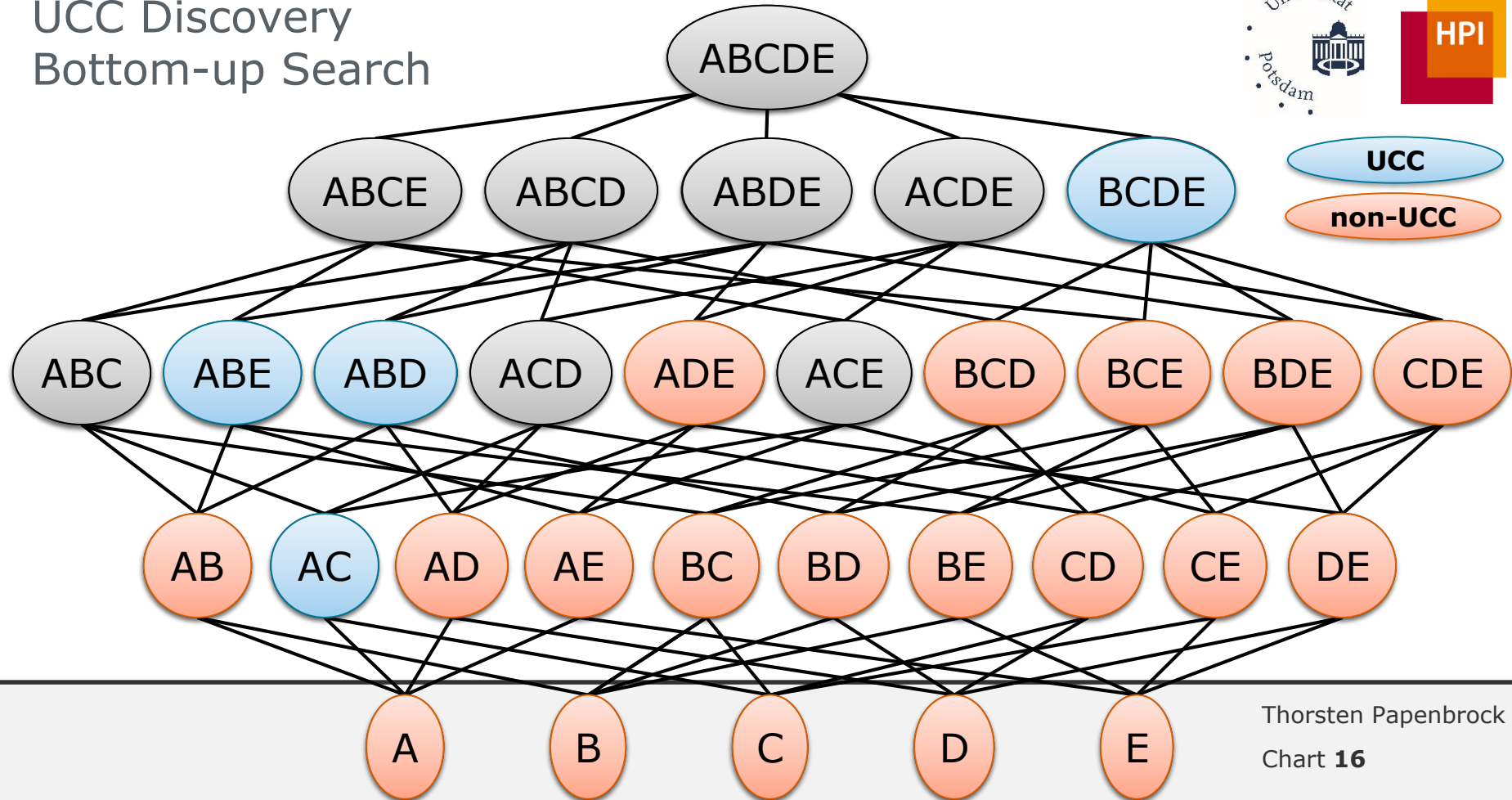


UCC Discovery

Bottom-up Search

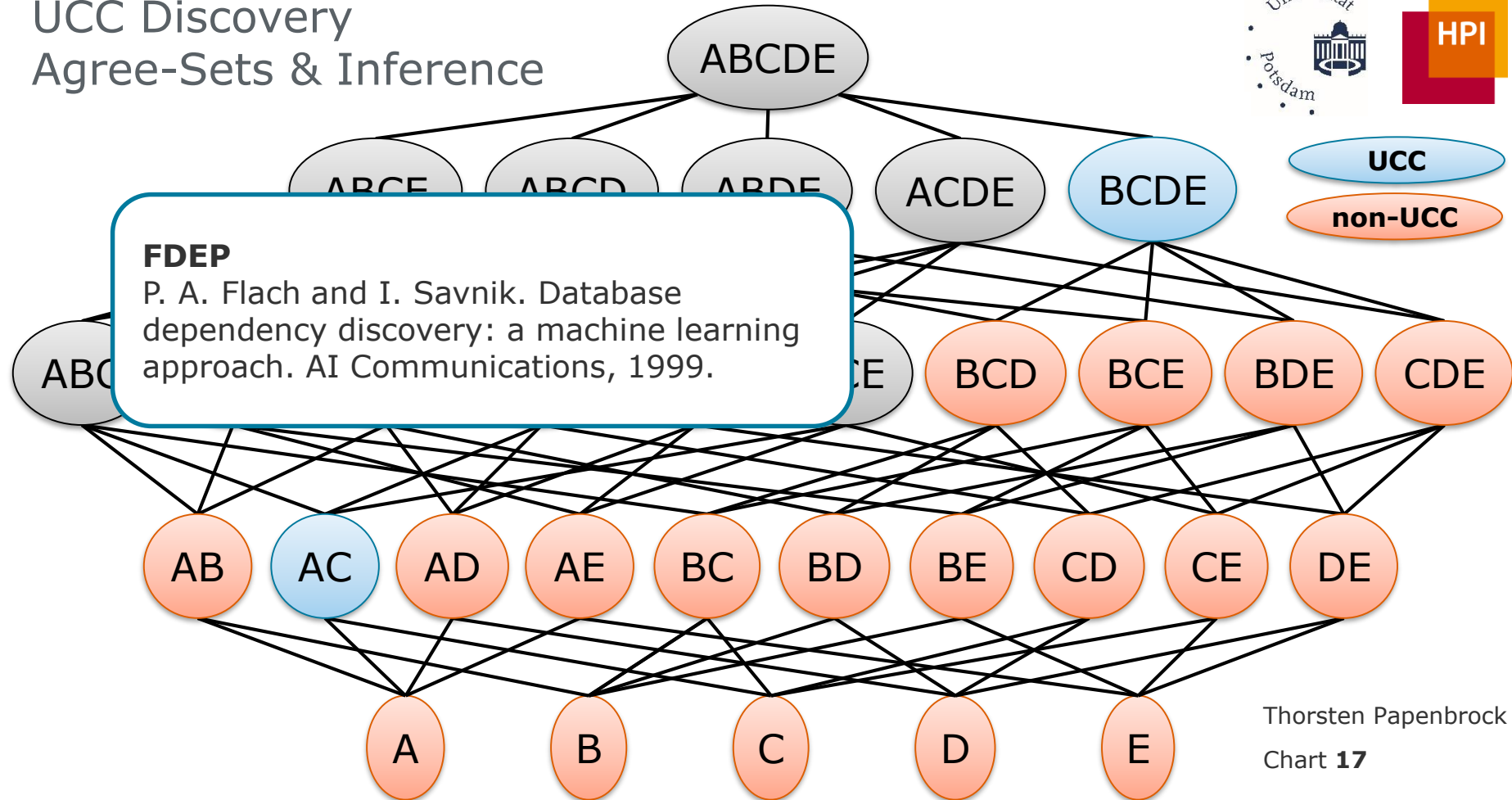


UCC Discovery Bottom-up Search



UCC Discovery

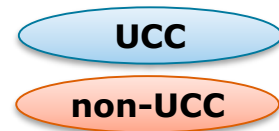
Agree-Sets & Inference



UCC Discovery

Agree-Sets & Inference

<u>Name</u>	<u>Surname</u>	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

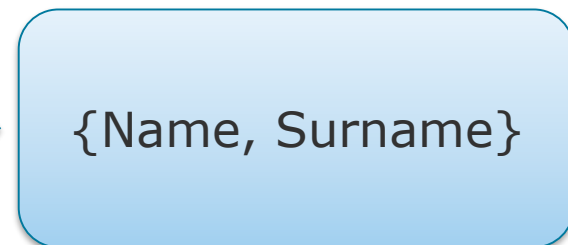


- $\neg\{\text{Surname, Postcode, City, Mayor}\}$
- $\neg\{\text{Name, Postcode, City, Mayor}\}$
- $\neg\{\text{Surname}\}$



Agree-Sets

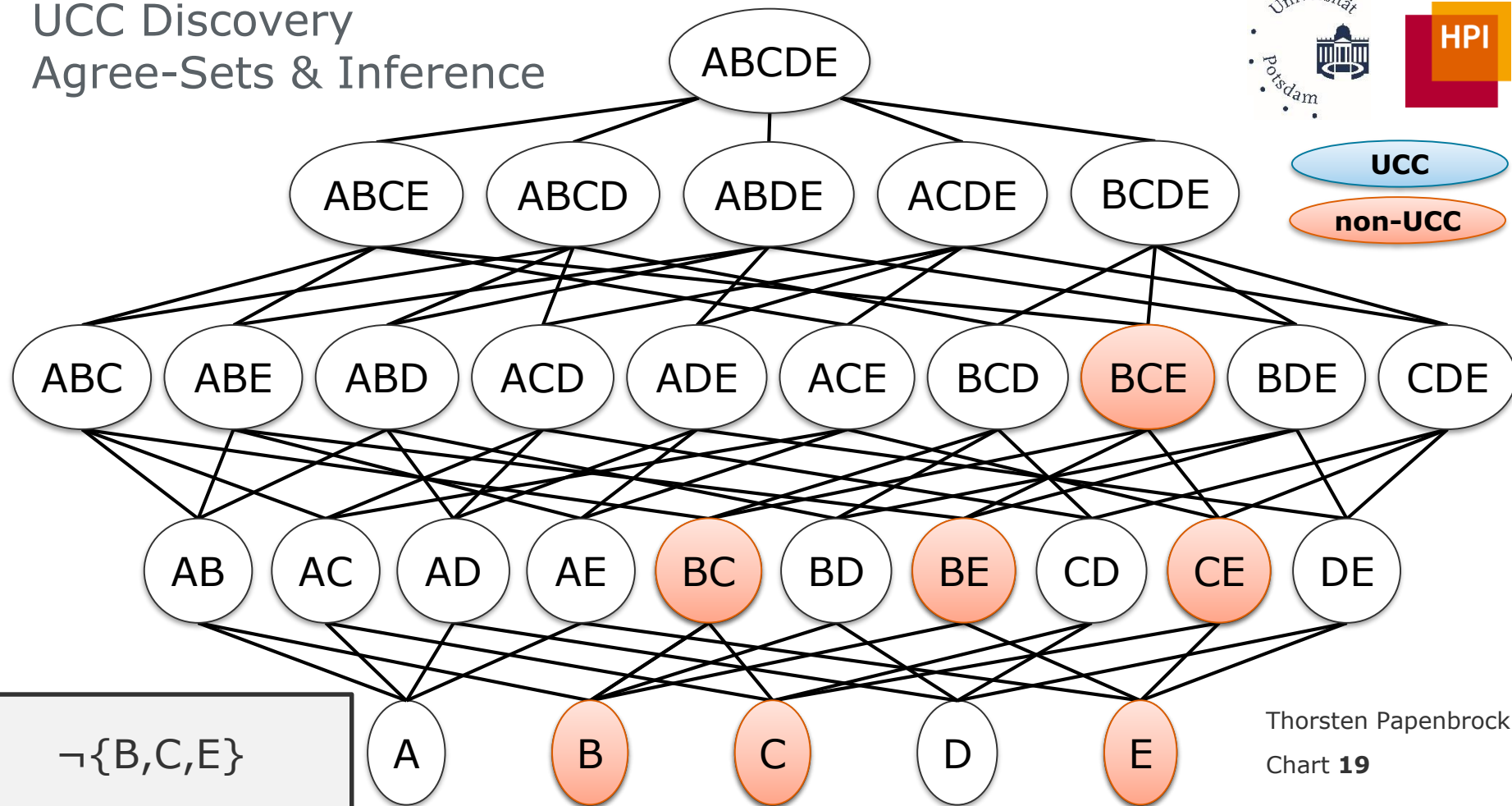
Inference



UCC Discovery

Agree-Sets & Inference

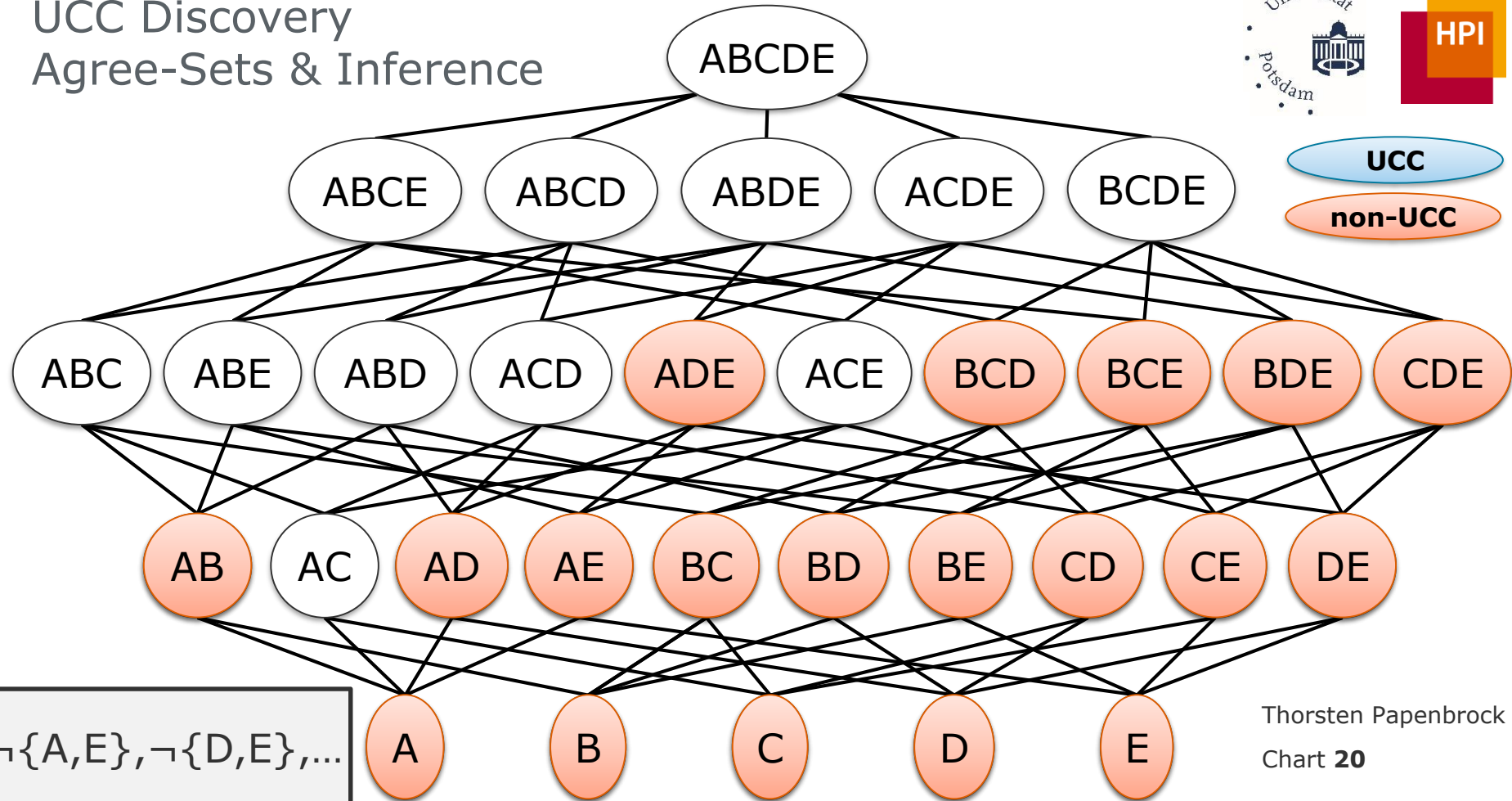
UCC
non-UCC



UCC Discovery

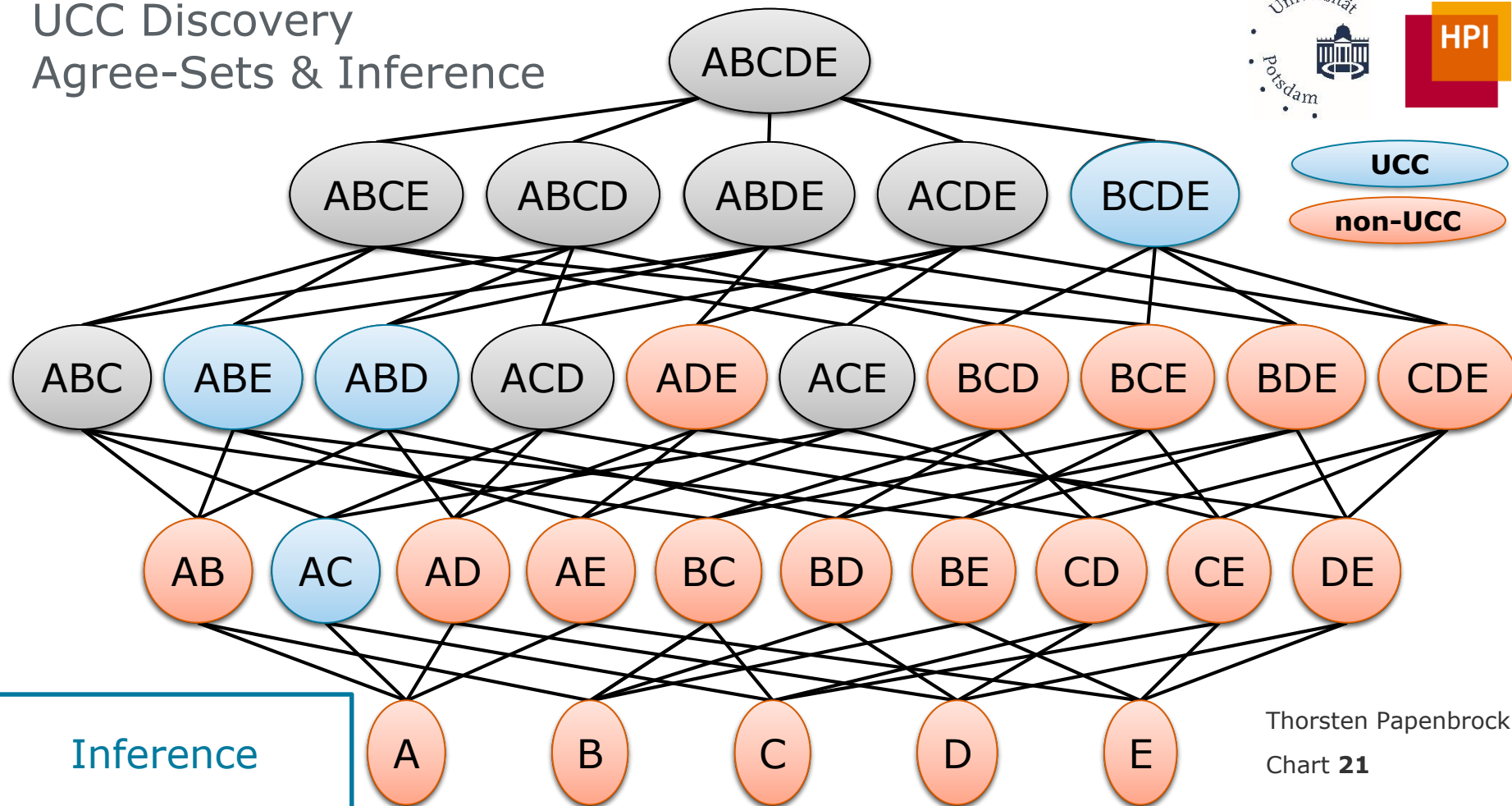
Agree-Sets & Inference

UCC
non-UCC

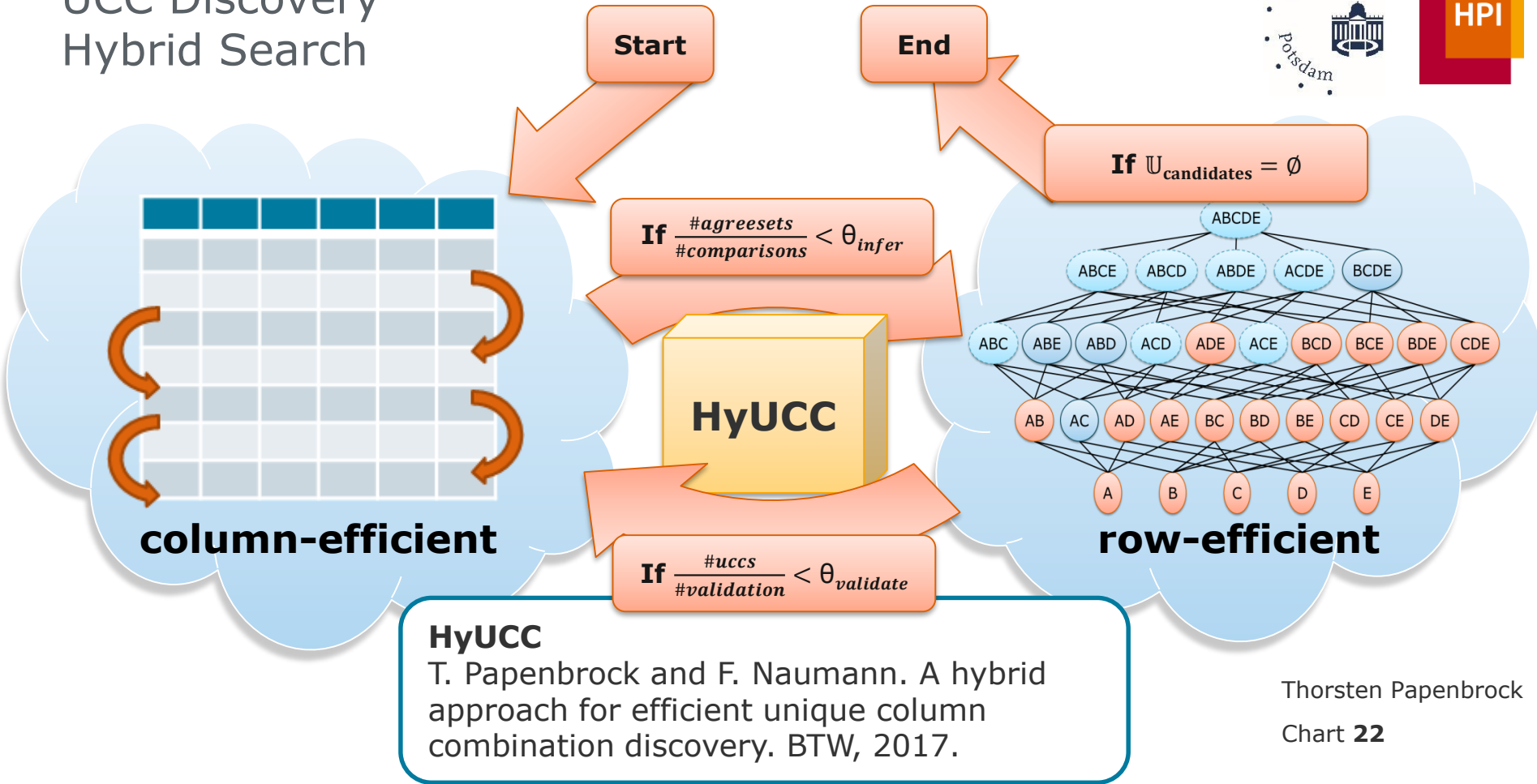


UCC Discovery

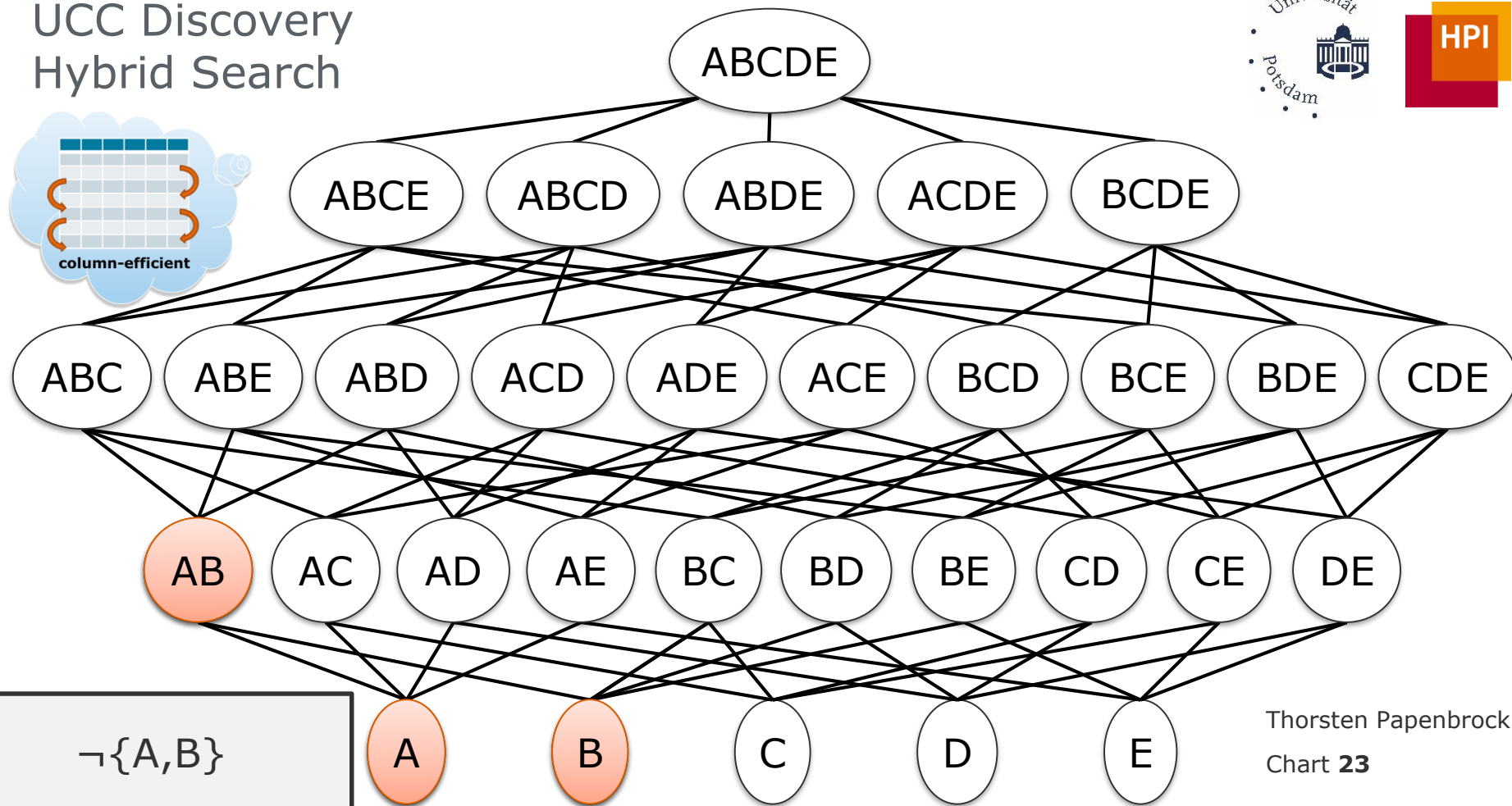
Agree-Sets & Inference



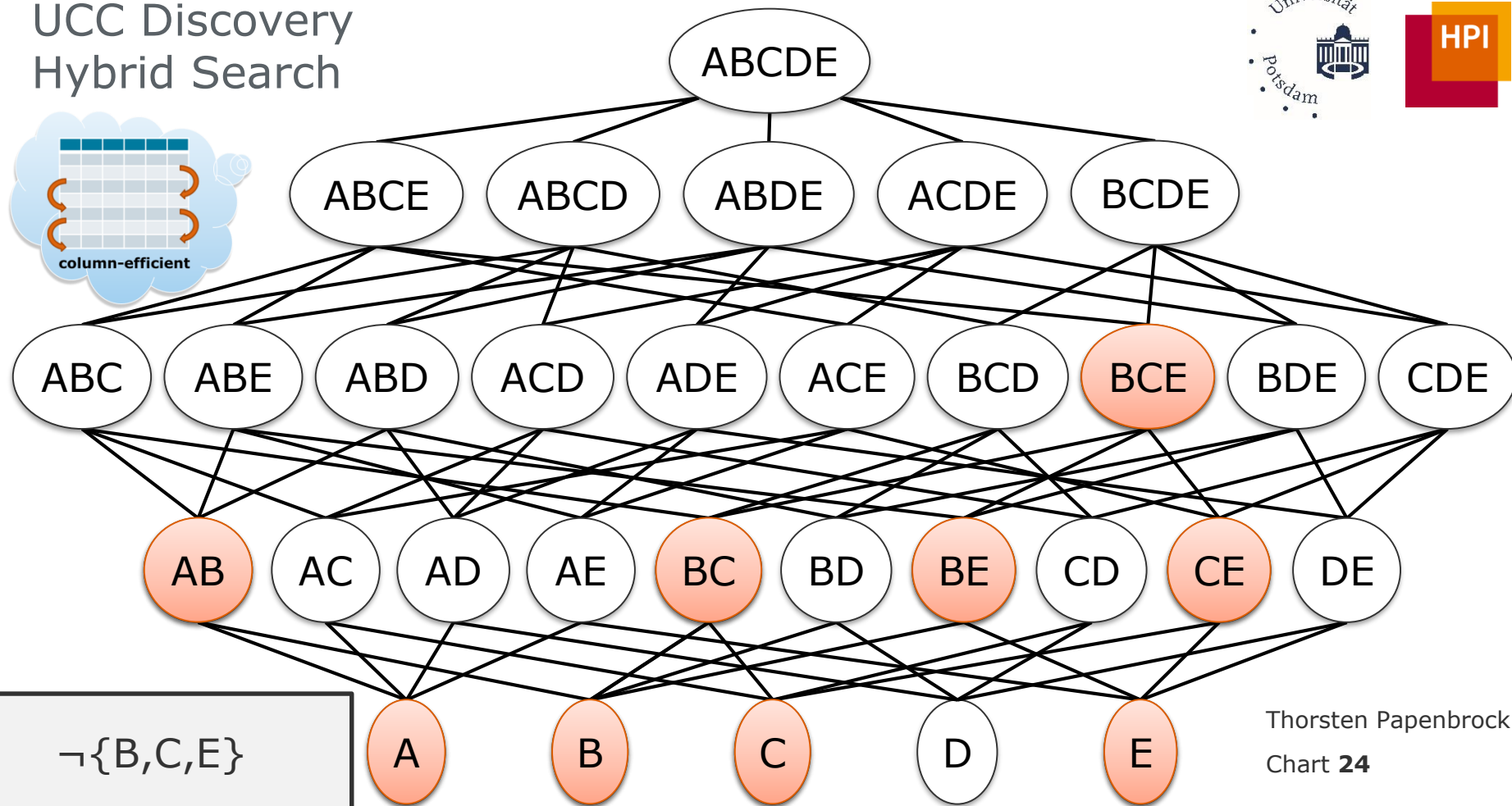
UCC Discovery Hybrid Search



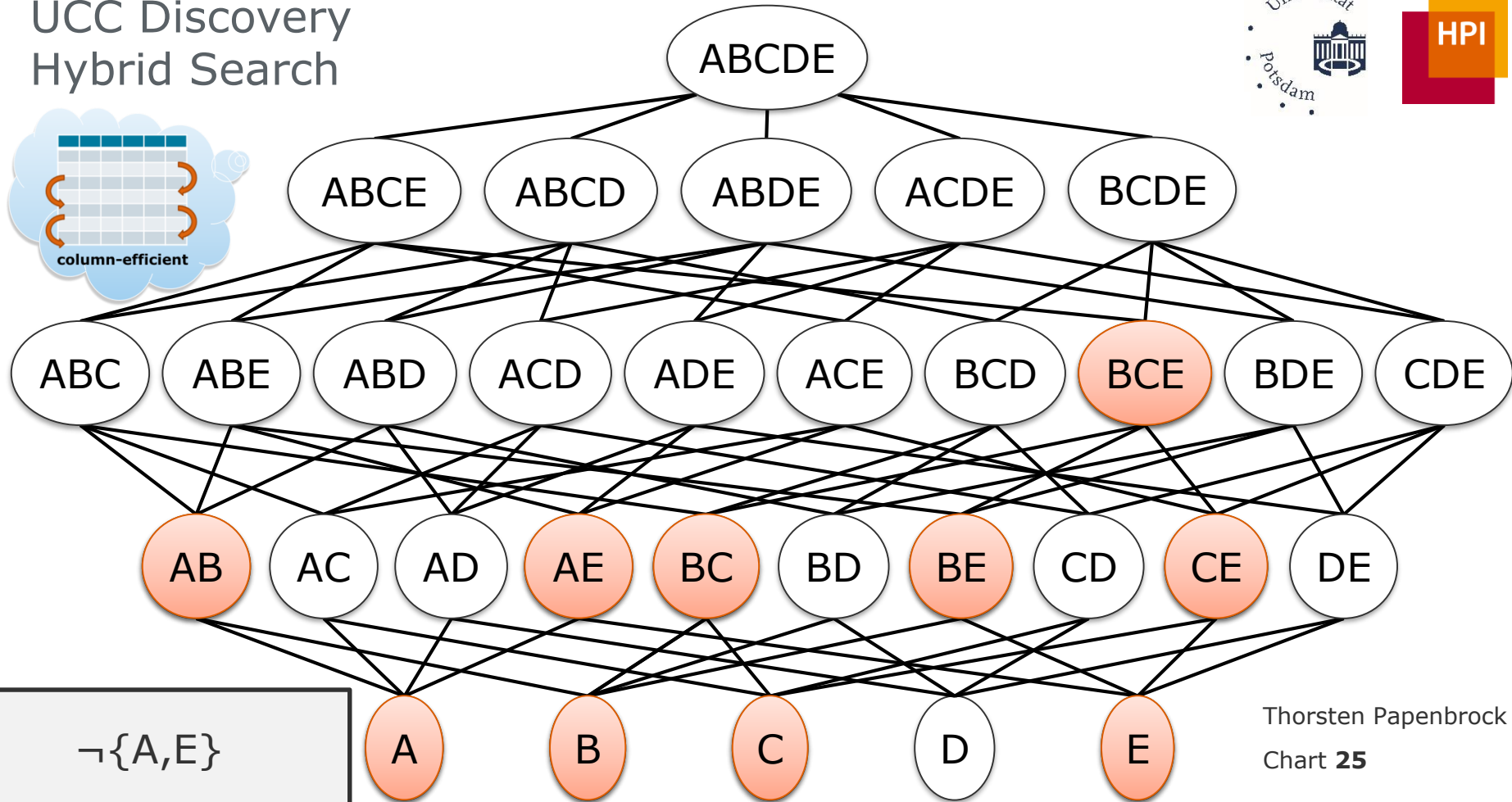
UCC Discovery Hybrid Search



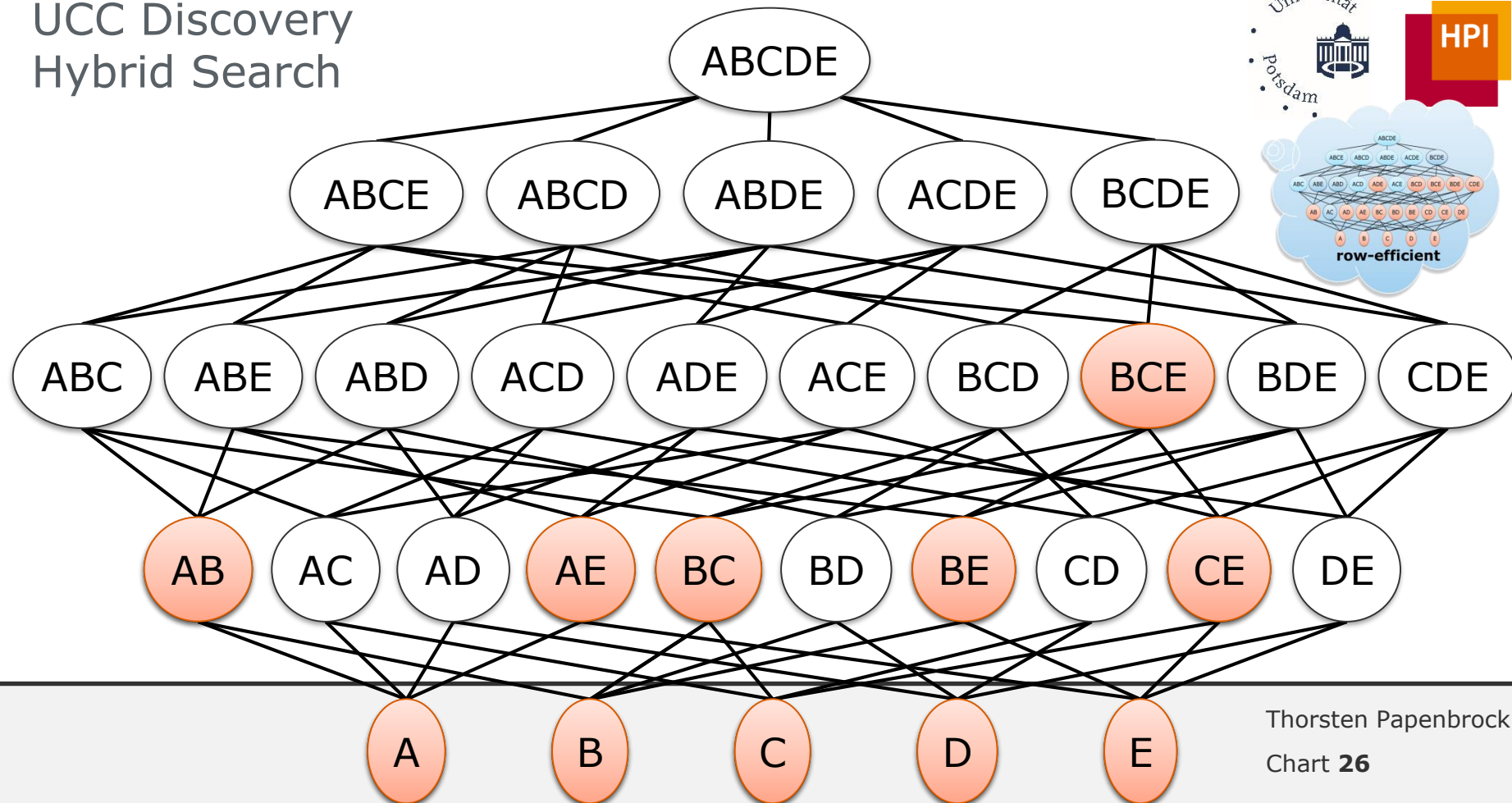
UCC Discovery Hybrid Search



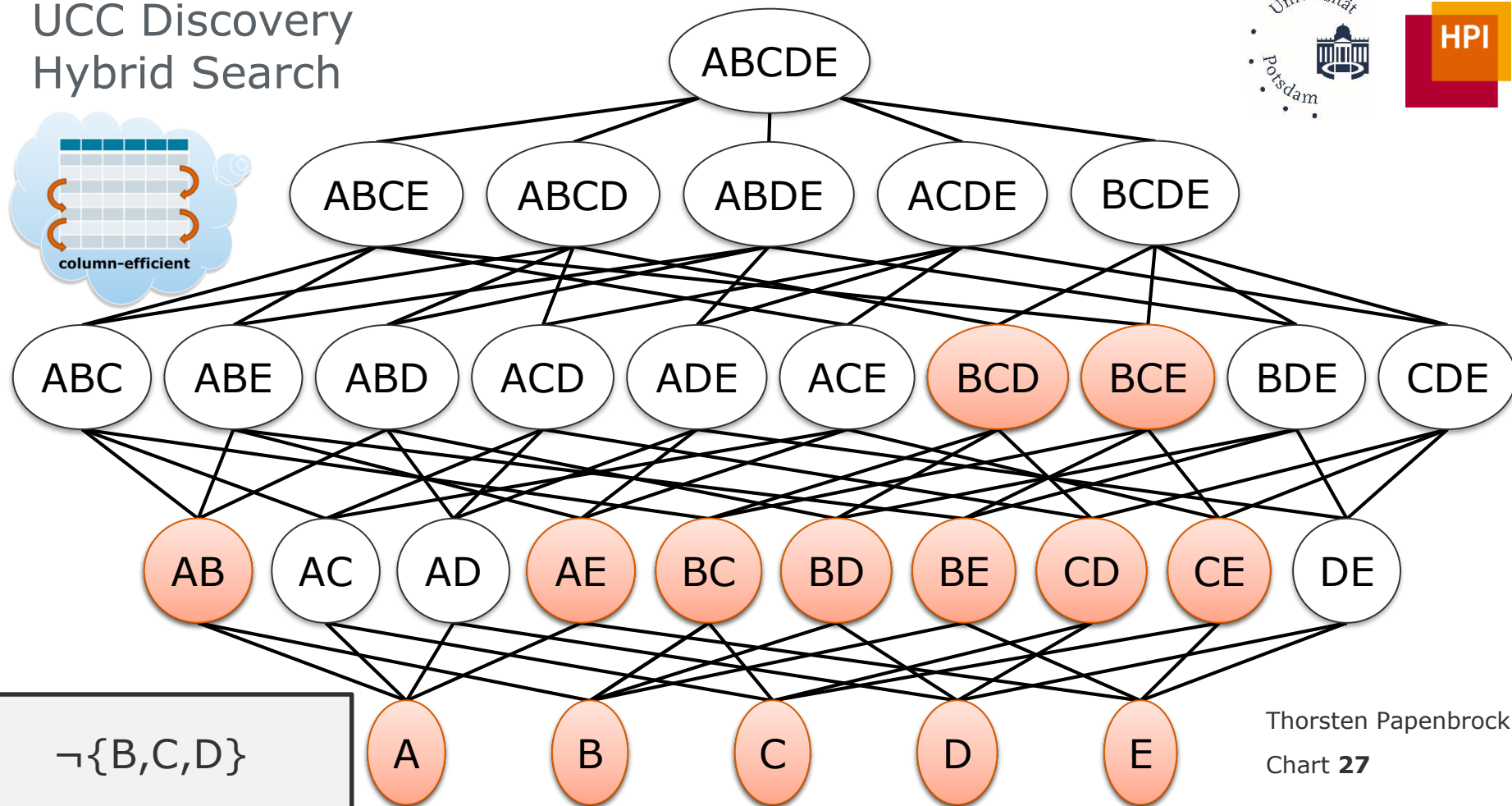
UCC Discovery Hybrid Search



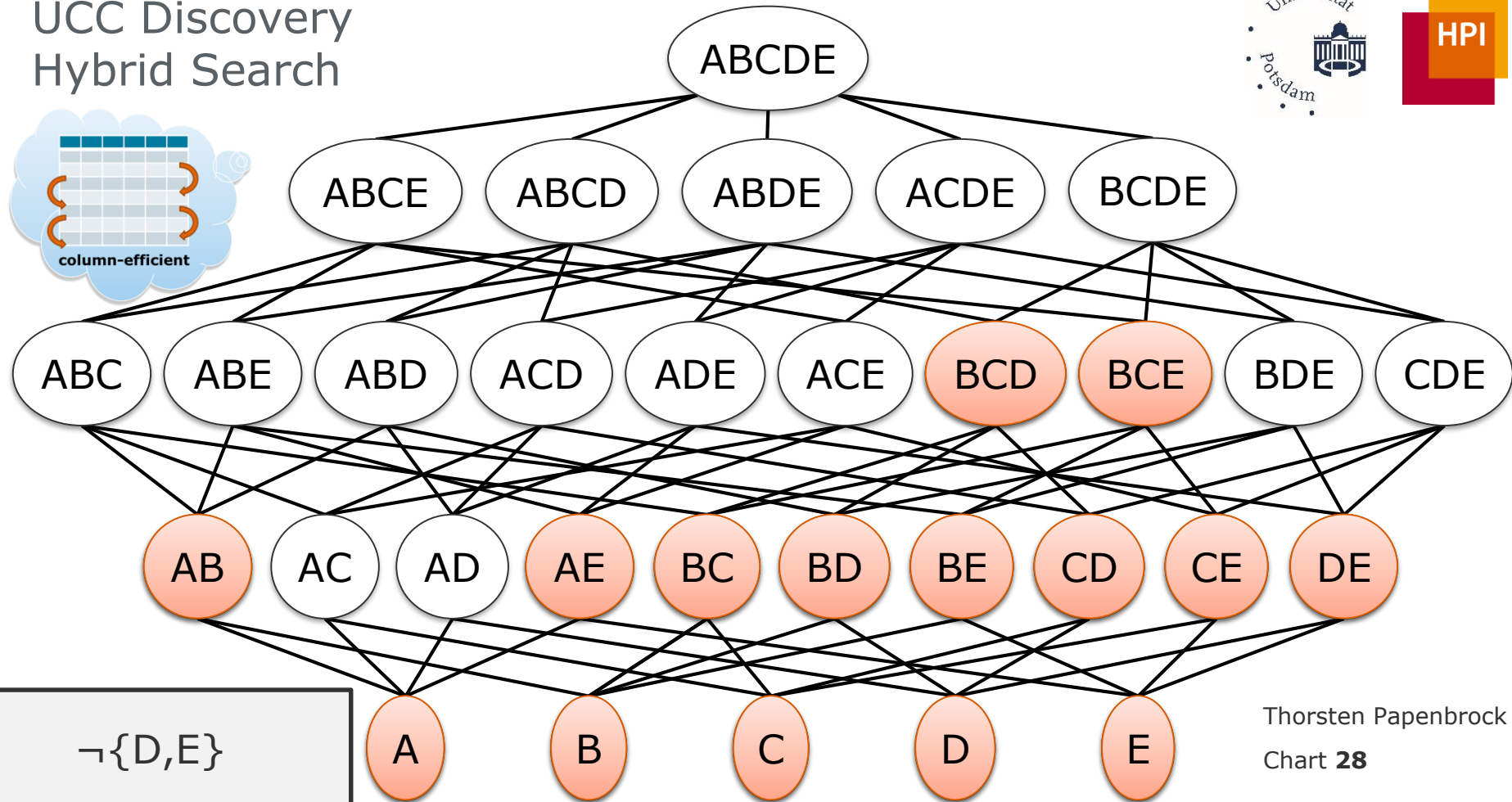
UCC Discovery Hybrid Search



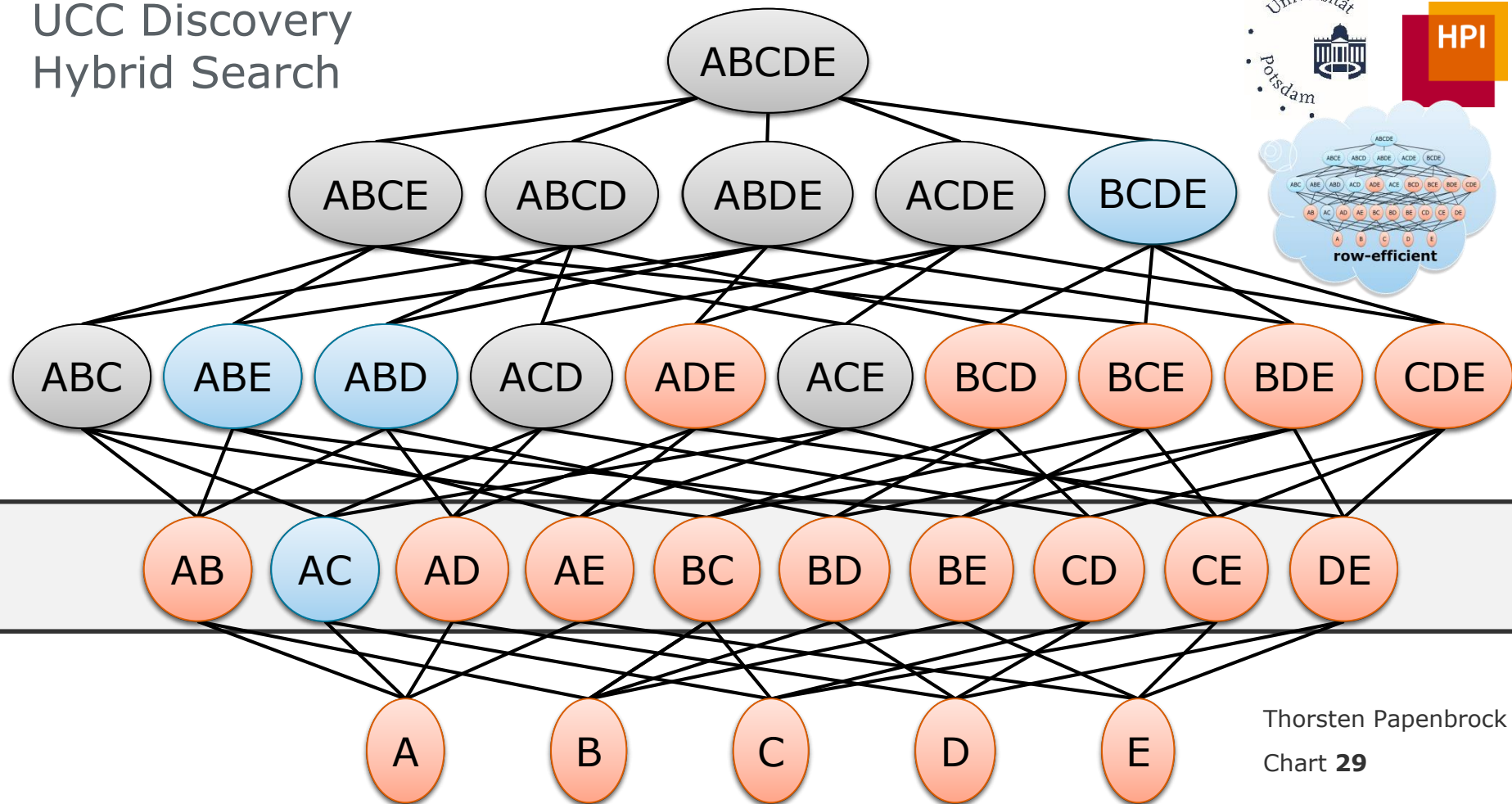
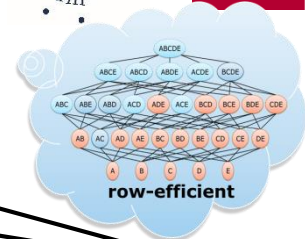
UCC Discovery Hybrid Search



UCC Discovery Hybrid Search



UCC Discovery Hybrid Search



UCC Discovery

Hitting Set Enumeration

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



~~Agree-Sets~~
Difference-Sets

➤ {Name}

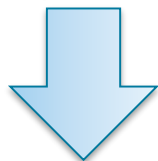
GORDIAN

Y. Sismanis, P. Brown, P. J. Haas, and B. Reinwald. GORDIAN: Efficient and scalable discovery of composite keys. PVLDB, 2006.

UCC Discovery

Hitting Set Enumeration

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



~~Agree-Sets~~
Difference-Sets

- {Name}
- ~~{Name, Postcode, City, Mayor}~~

GORDIAN

Y. Sismanis, P. Brown, P. J. Haas, and B. Reinwald. GORDIAN: Efficient and scalable discovery of composite keys. PVLDB, 2006.

UCC Discovery

Hitting Set Enumeration

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

slow

~~Agree-Sets~~
Difference-Sets

- {Name}
- {Surname}

fast

GORDIAN

Y. Sismanis, P. Brown, P. J. Haas, and B. Reinwald. GORDIAN: Efficient and scalable discovery of composite keys. PVLDB, 2006.

{Name, Surname}

fast

Hitting Set Enumeration

UCC Discovery

Hitting Set Enumeration

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

slow

~~Agree-Sets~~
Difference-Sets

- {Name}
- {Surname}

fast

HPIVvalid

J. Birnick, T. Bläsius, T. Friedrich, F. Naumann, T. Papenbrock, M. Schirneck.
Hitting Set Enumeration with Partial Information for Unique Column Combination Discovery. PVLDB, 2020.

{Name, Surname}

fast

Hitting Set Enumeration

UCC Discovery

Hitting Set Enumeration

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

sample

Sample Difference-Sets

- {Name}
- {Surname}

fast

fast

Hitting Set Enumeration

HPIVvalid

J. Birnick, T. Bläsius, T. Friedrich, F. Naumann, T. Papenbrock, M. Schirneck.
Hitting Set Enumeration with Partial Information for Unique Column Combination Discovery. PVLDB, 2020.

fast

incomplete

{Name, Surname}

wrong

{Name, Surname}

fast

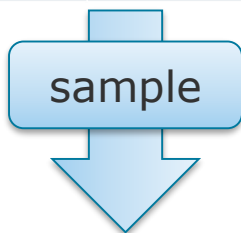
Validation

Thorsten Papenbrock
Chart 34

UCC Discovery

Hitting Set Enumeration

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



Sample Difference-Sets

- {Name}
- {Surname}

fast

fast

If a hitting set was no UCC

HPIVValid

J. Birnick, T. Bläsius, T. Friedrich, F. Naumann, T. Papenbrock, M. Schirneck.
Hitting Set Enumeration with Partial Information for Unique Column Combination Discovery. PVLDB, 2020.

fast

{Name, Surname}

Validation

{Name, Surname}

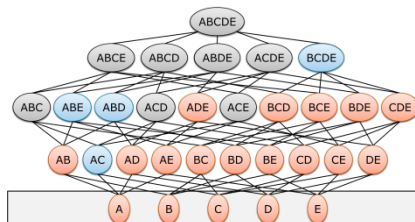
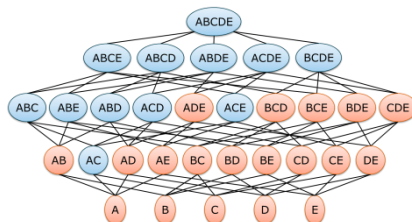
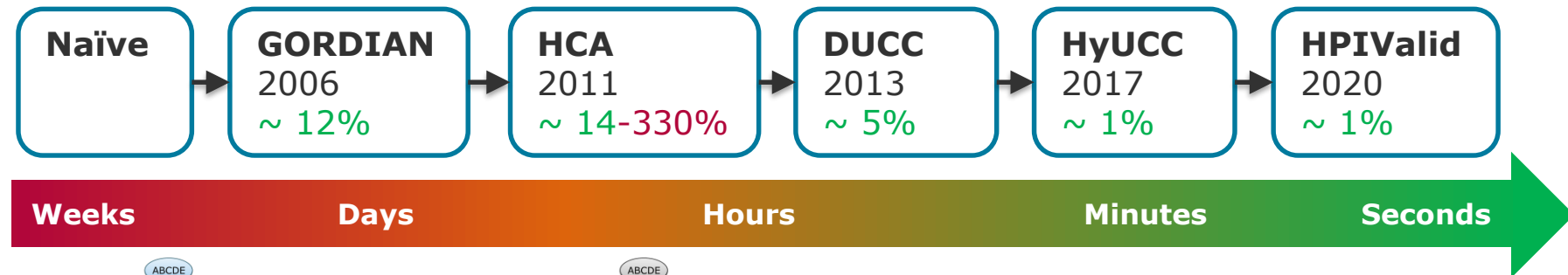
fast

Hitting Set Enumeration

Thorsten Papenbrock
Chart 35

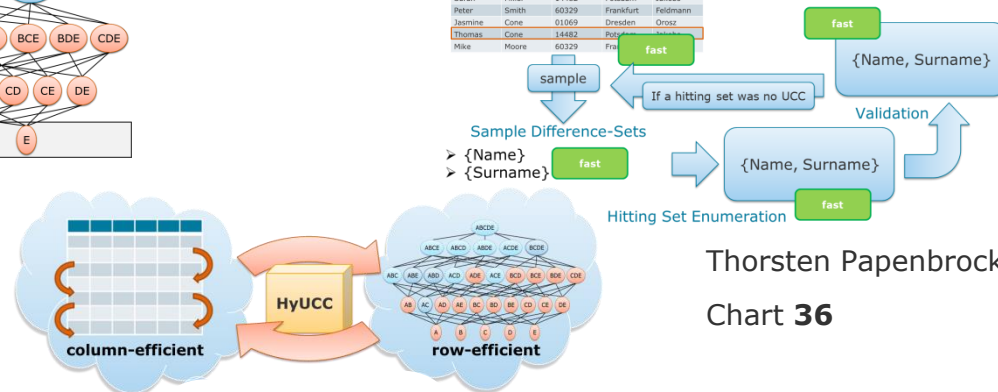
UCC Discovery Evaluation

Expected runtime for 100 MB of data:



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

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



Thorsten Papenbrock
Chart 36