



© Xeeldek (www.sketchport.com)

A Hybrid Approach for Efficient Unique Column Combination Discovery

Thorsten Papenbrock and Felix Naumann



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



Not unique!
See also: Kanto, Johto, Kalos

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

Not unique!

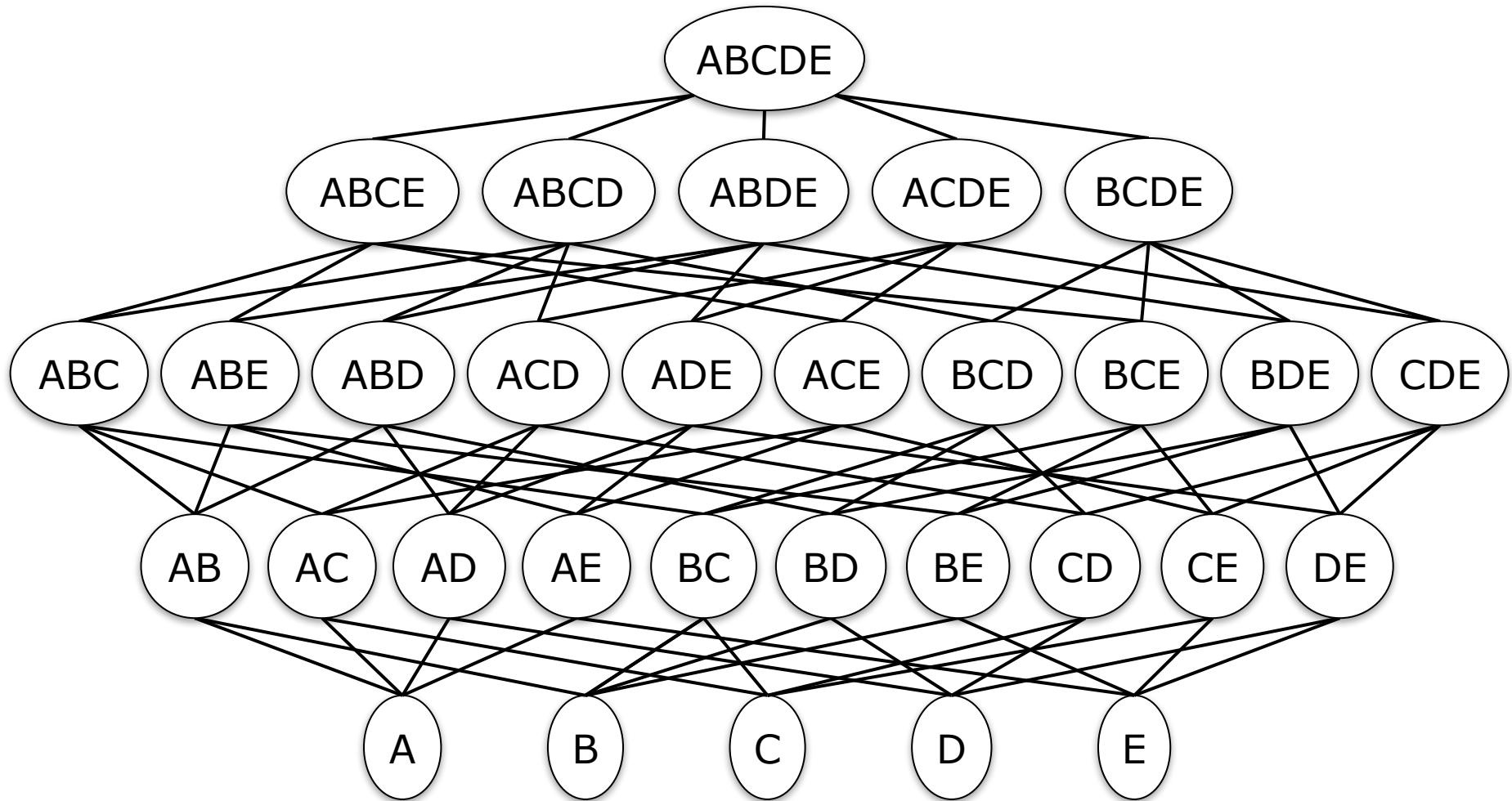
Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails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

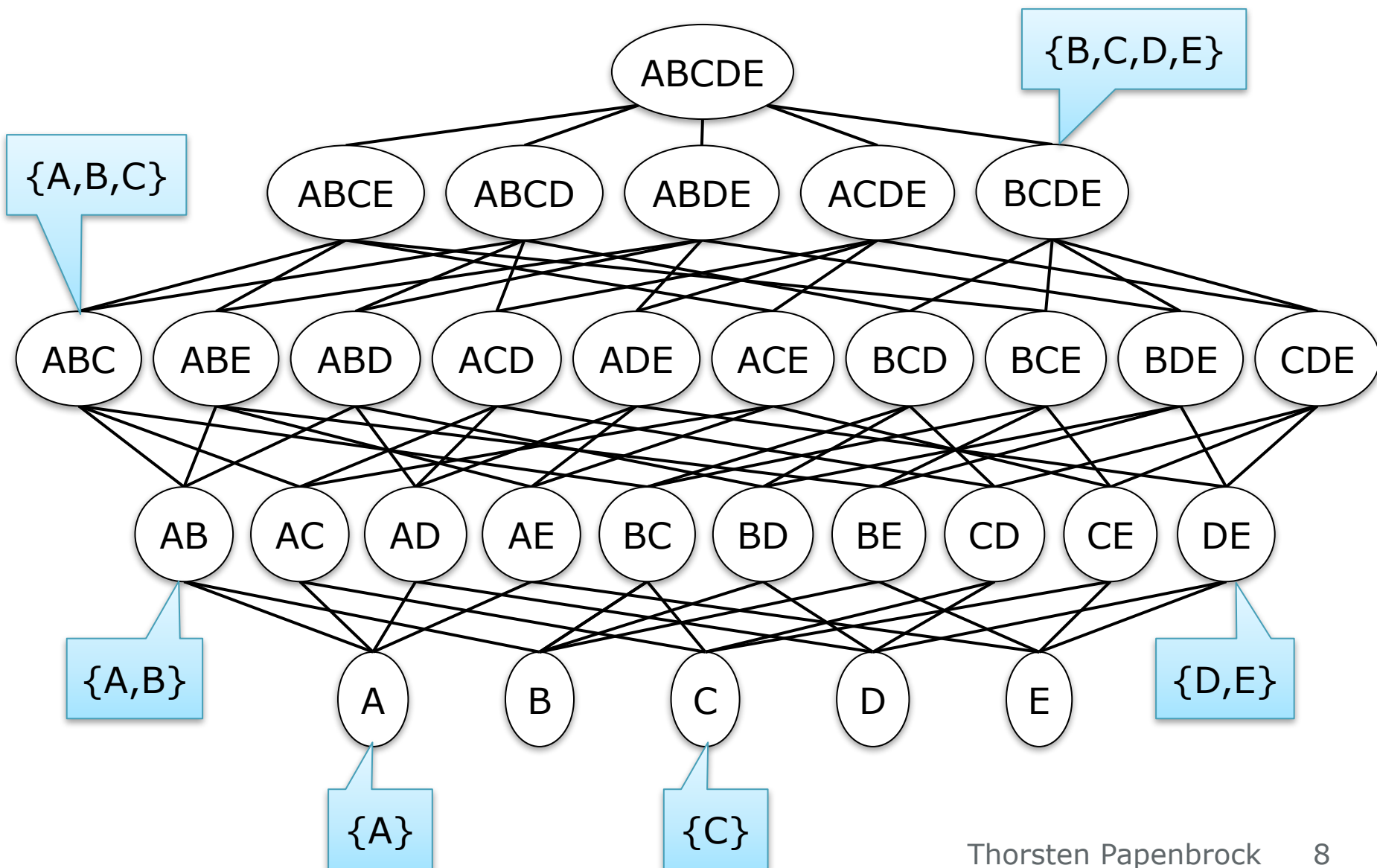
Not unique!

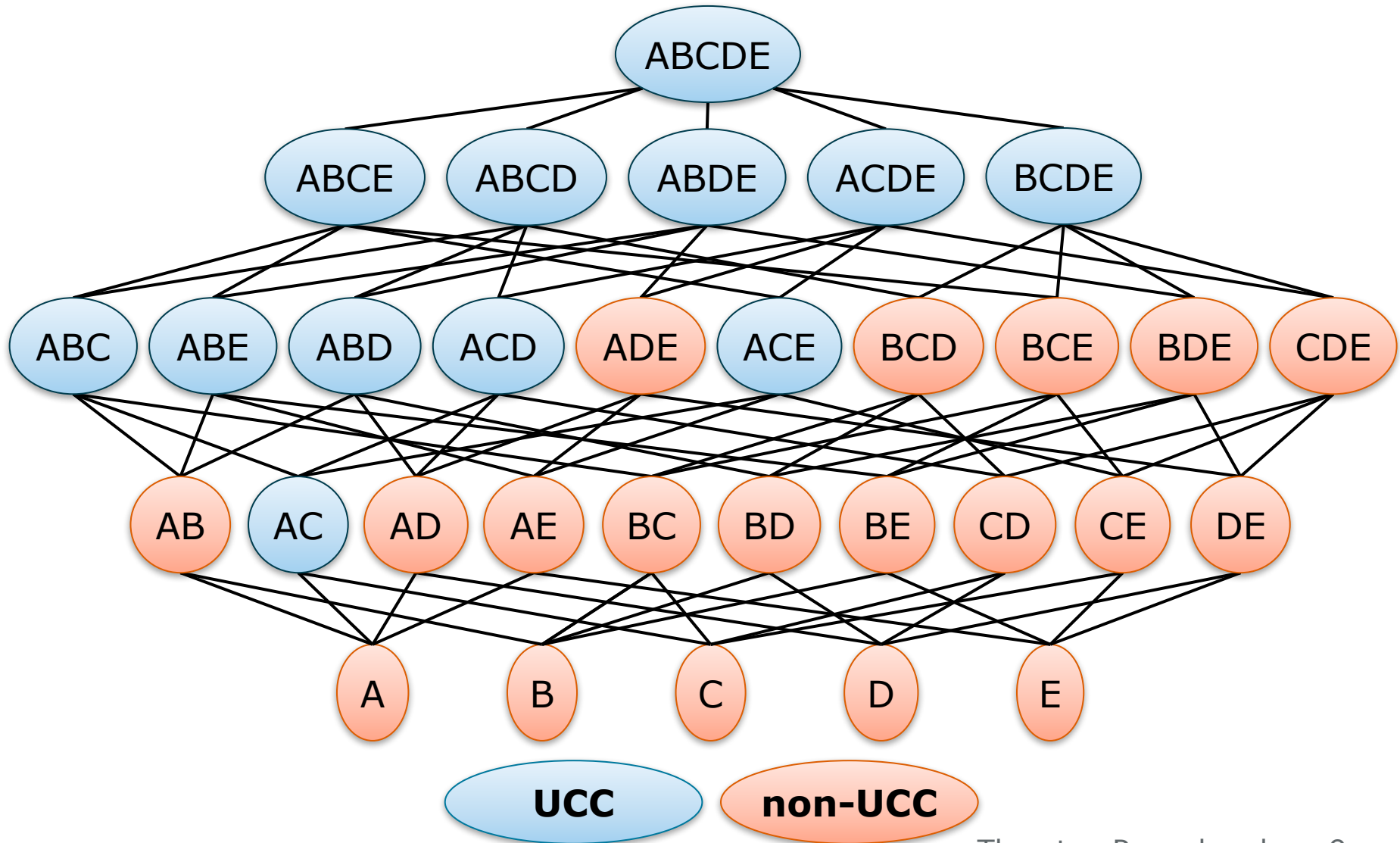
Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails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

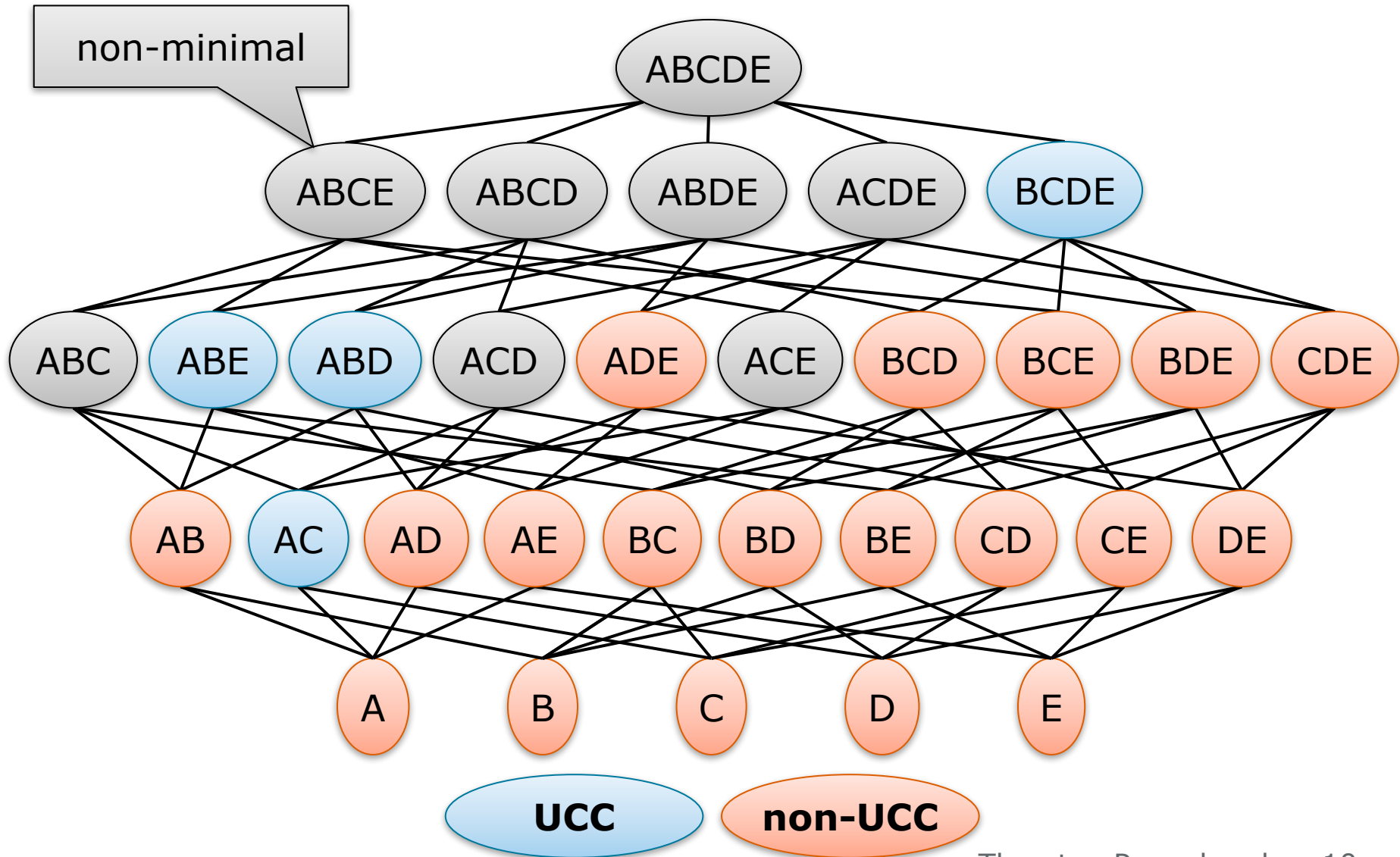
Unique: {Name, Sex, Type}

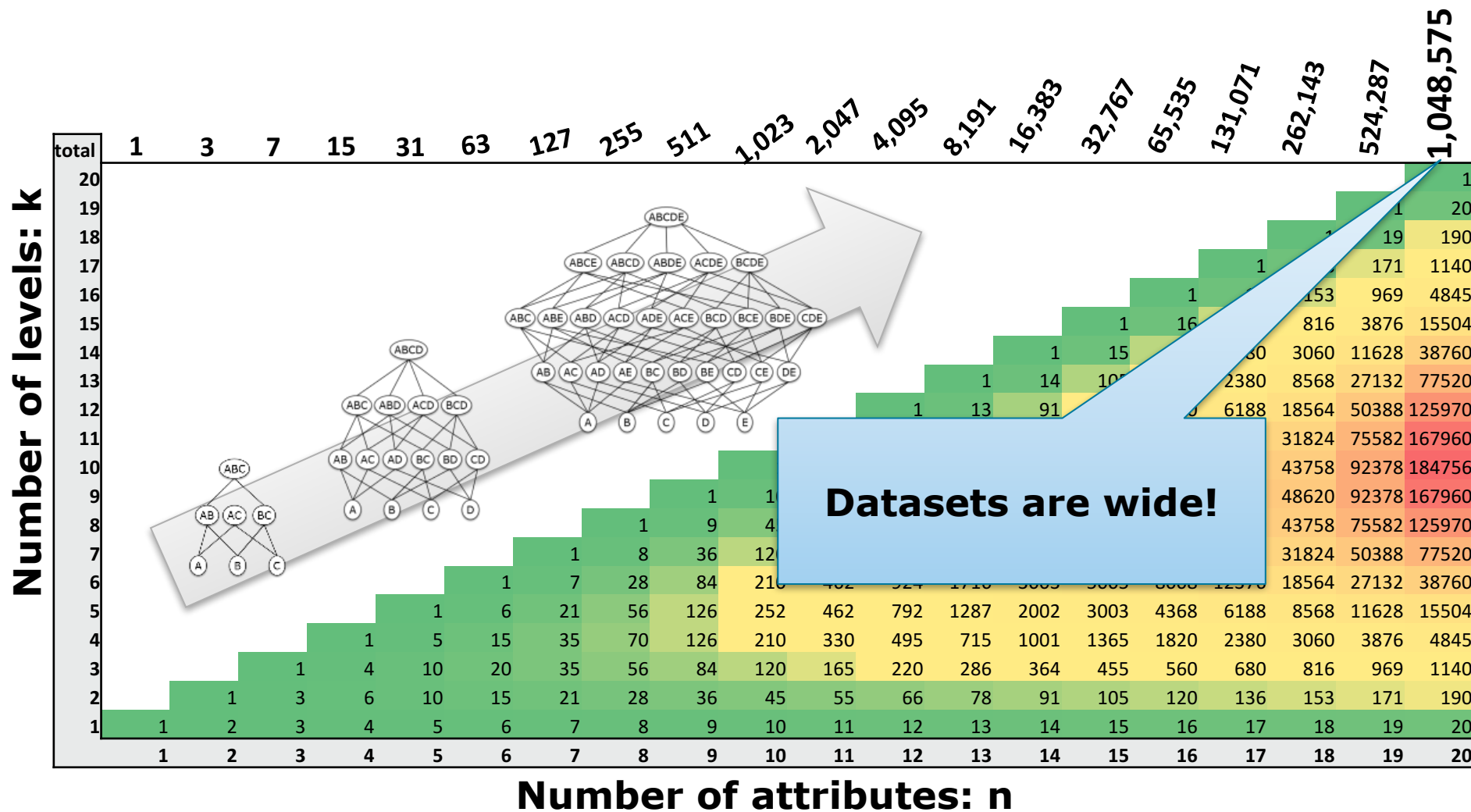
Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails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





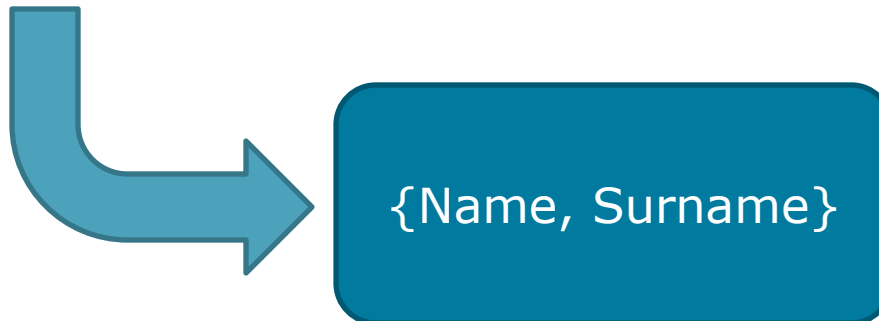






<u>Name</u>	<u>Surname</u>	Postcode	City	Mayor
Thomas	Miller	14482	Potsdam	Jakobs
Sarah	Miller	14482	Potsdam	Jakobs
Peter	Smith	60329	Frankfurt	Feldmann
Jasmine	Cone	01069	Dresden	Orosz
Thomas	Cone	14482	Potsdam	Jakobs
Mike	Moore	60329	Frankfurt	Feldmann

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



ncvoter_1024001r.19c.csv - LibreOffice Calc

File Edit View Insert Format Tools Data Window Help

Liberation Sans 10

AL71 f(x) Σ =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
924934	9882331	thorpe	i	grainfield		559	h	108	barrenigan pl	carey	nc	27511	nc	01.01.1982	2011.10				
924934	9882333	robert	i	gib		644m	h	6612	coech light cir	hensh	nc	27613	nc	01.01.1982	2011.10				
924934	9882337	stake	i	kurtz		500	h	5004	denwood ct	hensh	nc	27609	nc	01.01.1982	2011.10				
924934	9882346	stake	i	wayne		644m	h	3904	winged thistle ct	hensh	nc	27617	nc	01.01.1982	2011.10				
924937	9882347	haley	c	essler		600m	h	916	petebrook dr	hensh	nc	27609	nc	01.01.1982	2011.10				
924938	9882348	hendrix	c	essler		638	h	916	petebrook dr	hensh	nc	27609	nc	01.01.1982	2011.10				
924939	9882351	virginia	i	newman		800m	h	35	springnoe ct	hensh	nc	27615	nc	01.01.1982	2011.10				
924940	9882352	robert	i	newman		500m	h	632	denwood ct	hensh	nc	27617	nc	01.01.1982	2011.10				
924941	9882355	mark	c	hulbert		520m	h	2031	mcclerty st	hensh	nc	27608	nc	01.01.1982	2011.10				
924942	9882356	marie-elaine	i	winlow		540	h	1383	mayfield dr	hensh	nc	27608	nc	01.01.1982	2011.10				
924943	9882360	beth	b	hendrix		489	h	4008	perthcroft pl	hensh	nc	27536	nc	01.01.1982	2011.10				
924943	9882367	henry	b	hendrix		377m	h	35	springnoe ct	hensh	nc	27615	nc	01.01.1982	2011.10				
924945	9882370	joyce	marie	hennings		589	b	1722	crag burn in	hensh	nc	27604	nc	01.01.1982	2011.10				
924946	9882381	helen	i	hennings		720	h	4177	penetration dr	hensh	nc	27609	nc	01.01.1982	2011.10				
924947	9882382	john	i	crenshaw		460m	h	9612	batons creek rd	hensh	nc	27615	nc	01.01.1982	2011.10				
924948	9882385	rich	iv	fox		720	h	801	e chatham st	carey	nc	27511	nc	01.01.1982	2011.10				
924949	9882390	justin	iv	johnson		639	h	208	hennelwood st	hensh	nc	27609	nc	01.01.1982	2011.10				
924950	9882395	jill	h	santaly		600	h	6412	winthrop dr	hensh	nc	27612	nc	01.01.1982	2011.10				
924951	9882396	harriet	emmett	hennery		430m	h	1613	ingradyway in	hensh	nc	27603	nc	01.01.1982	2011.10				
924952	9882397	scott	a	hammer		520m	h	12900	richmond run dr	hensh	nc	27614	nc	01.01.1982	2011.10				
924953	9882398	william	h	johnson		510m	h	311	edwards dr	hensh	nc	27615	nc	01.01.1982	2011.10				
924954	9882399	margaret	h	holten		844	h	1416	cool spring rd	hensh	nc	27614	nc	01.01.1982	2011.10				
924955	9882400	james	h	holten		644m	h	1410	cool spring rd	hensh	nc	27614	nc	01.01.1982	2011.10				
924956	9882403	john	h	hodge		610m	h	1300	holway hollow in	hensh	nc	27610	nc	01.01.1982	2011.10				
924957	9882410	elodie	chavis	freeman		610m	h	103	forewood dr	gamer	nc	27529	nc	01.01.1982	2011.10				
924958	9882412	alan	h	brunson		500m	h	3217	banks st	hensh	nc	27603	nc	01.01.1982	2011.10				
924959	9882413	hennetta	iv	threalt		489	b	908	knolls hill pl	knightsdale	nc	27445	nc	01.01.1982	2011.10				
924960	9882415	patrick	a	johnson		540	h	2141	springridge dr #4	hensh	nc	27615	nc	01.01.1982	2011.10				
924961	9882418	claudia	b	clavis		620	h	2740	community dr	hensh	nc	27610	nc	01.01.1982	2011.10				
924962	9882419	johnson	i	johnson		520	h	307	king st	hensh	nc	27601	nc	01.01.1982	2011.10				
924963	9882420	tylvestre	i	davis		670m	h	2740	community dr	hensh	nc	27610	nc	01.01.1982	2011.10				
924964	9882423	claire	i	lee		510m	h	5113	edwards dr	hensh	nc	27615	nc	01.01.1982	2011.10				
924965	9882441	claire	iv	lightner-shape		470	h	6530	glendower st	hensh	nc	27613	nc	01.01.1982	2011.10				
924966	9882442	catherine	iv	metrick		630	h	1337	swallow dr	hensh	nc	27606	nc	01.01.1982	2011.10				
924967	9882451	isa	h	tyne		600	h	313	wentworth ct	hensh	nc	27610	nc	01.01.1982	2011.10				
924968	9882452	guendolyn	h	lucas		640	h	2900	poole st	hensh	nc	27610	nc	01.01.1982	2011.10				
924969	9882453	john	h	johnson		540m	h	1141	brampton st	hensh	nc	27610	nc	01.01.1982	2011.10				
924970	9882454	alicia	h	edrick		620	h	311	kenrick dr #311	gamer	nc	27529	nc	01.01.1982	2011.10				
924971	9882455	johnella	h	mc		510	h	7204	solums ct	hensh	nc	27610	nc	01.01.1982	2011.10				
924972	9882457	call	h	huller		480	b	223	walstop st	hensh	nc	27610	nc	01.01.1982	2011.10				
924973	9882458	william	a	hennelwood		700m	h	1408	andy woods dr	hensh	nc	27610	nc	01.01.1982	2011.10				
924974	9882459	richelle	h	williams		540	h	3205	canfield way	hensh	nc	27613	nc	01.01.1982	2011.10				
924975	9882461	alexia	h	brunson		480m	h	4307	hennelwood dr	hensh	nc	27610	nc	01.01.1982	2011.10				
924976	9882462	christy	b	burgin		540	h	3746	brunch st	hensh	nc	27610	nc	01.01.1982	2011.10				
924977	9882463	greg	h	je		580	a	3705	carriage minor ct	wendell	nc	27591	nc	01.01.1982	2011.10				
924978	9882464	john	g	deem-powell		570	h	114	padgett ct	carey	nc	27516	nc	01.01.1982	2011.10				
924979	9882468	viola	h	cunningham		510	h	1333	wentworth ct	hensh	nc	27610	nc	01.01.1982	2011.10				
924980	9882469	edna	willson	h		600	h	3800	lyon dr	hensh	nc	27616	nc	01.01.1982	2011.10				
924981	9882490	larry	id	clay		600m	h	4230	wendell blvd	wendell	nc	27591	nc	01.01.1982	2011.10				
924982	9882497	leah	h	puflin		480	h	910	williams st	hensh	nc	27610	nc	01.01.1982	2011.10				
924983	9882701	marguerite	iv	denison		470	h	5404	hounds ear pl	hensh	nc	27606	nc	01.01.1982	2011.10				
924984	9882702	sue	a	douglas		600m	h	2804	malvern ct	hensh	nc	27608	nc	01.01.1982	2011.10				
924985	9882704	marvin	h	brancosone		600m	h	2618	wellis ave	hensh	nc	27610	nc	01.01.1982	2011.10				
924986	9882705	terrie	p	douglas		600m	h	2804	malvern ct	hensh	nc	27613	nc	01.01.1982	2011.10				
924987	9882706	edith	h	huller		590m	h	3008	devonwood ct	hensh	nc	27609	nc	01.01.1982	2011.10				
924988	9882715	marilyn	p	lemond		740	h	6725	brookholow dr	hensh	nc	27615	nc	01.01.1982	2011.10				
924989	9882717	georgia	g	mcclain		720	h	2409	village grove rd	hensh	nc	27613	nc	01.01.1982	2011.10				
924990	9882724	christina	g	harris		510	h	8516	sleepy creek dr	hensh	nc	27613	nc	01.01.1982	2011.10				
924991	9882725	david	h	harris		510m	h	8516	sleepy creek dr	hensh	nc	27613	nc	01.01.1982	2011.10				
924992	9882729	robert	i	wolckelmann		580m	h	11600	apallosa run w	hensh	nc	27613	nc	01.01.1982	2011.10				
924993	9882734	christie	h	hennelwood		480	h	4820	carroll forest cir	hensh	nc	27610	nc	01.01.1982	2011.10				
924994	9882730	douglas	k	miller		500m	h	3008	tree meadow in	hensh	nc	27603	nc	01.01.1982	2011.10				
924995	9882764	john	p	haley		670m	h	113	star flower ct	apex	nc	27529	nc	01.01.1982	2011.10				
924996	9882765	william	m	powell		470	h	4702	barfield rd	hensh	nc	27610	nc	01.01.1982	2011.10				
924997	9882770	frances	p	dulley		670	h	119	baton dr	carey	nc	27518	nc	01.01.1982	2011.10				
924998	9882814	broderick	a	broderick		540m	h	7409	poole st	hensh	nc	27610	nc	01.01.1982	2011.10				
924999	9882835	larry	h	lyde		540m	h	1213	e martin st	hensh	nc	27610	nc	01.01.1982	2011.10				
925000	9882840	john	h	johnson		630	h	3008	kansas st	hensh	nc	27610	nc	01.01.1982	2011.10				
925001	9882846	teresa	a	mcroy		540m	h	8408	morgans way	hensh	nc	27613	nc	01.01.1982	2011.10				
925002	9882848	john	h	mcroy		540m	h	3001	lyon rd	hensh	nc	27610	nc	01.01.1982	2011.10				
925003	9882849	gerald	thomas	baugh		680m	h	408	haley dr	hensh	nc	27606	nc	01.01.1982	2011.10				
925004	9882850	william	h	sanders		490	h	7912	lyon mill rd	hensh	nc	27606	nc	01.01.1982	2011.10				
925005	9882851	john	h	sanders		510m	h	7912	lyon mill rd	hensh	nc	27606	nc	01.01.1982	2011.10				
925006	9882856	caryn	beaty	dumes		480	b	3210	cotton grove run	hensh	nc	27610	nc	01.01.1982	2011.10				
925007	9882857	john	h	jones		700m	h	3810	bosco ct	hensh	nc	27606	nc	01.01.1982	2011.10				
925008	9882887	gloria	h	burns		720	h	2000	sierra dr	hensh	nc	27608	nc	01.01.1982	2011.10				
925009	9882888	john	b	carver		540	h	5216	west waks dr	hensh	nc	27610	nc	01.01.1982	2011.10				
925010	9882892	rebecca	m	johnson		500m	h	114	dorchester pines ct	carey	nc	27511	nc	01.01.1982	2011.10				
925011	9882893	phyllis	h	johnson		510	h	4213	curry ridge ct	hensh	nc	27610	nc	01.01.1982	2011.10				
925012	9882896	grady	h	patterson		540m	h	12208	chastain dr	hensh	nc	27614	nc	01.01.1982	2011.10				
925013	9882897	john	h	hennelwood		480	h	4820	carroll forest cir	hensh	nc	27610	nc	01.01.1982	2011.10				
925014	9882909	hannah	h	hennelwood		480	h	104	hennelwood ct	carey	nc	27511	nc	01.01.1982	2011.10				
925015	9882916	john	a	whitstone		600m	h	2604	whistling gull run	apex	nc	27529	nc	01.01.1982	2011.10				
925016	9882917	john	a	cooper		620	h	208	willowood ct	hensh	nc	27606	nc	01.01.1982	2011.10				
925017	9882945	john	a	mcclain		590m	h	2613	castalia dr	hensh	nc	27607	nc	01.01.1982	2011.10				
925018	9882947	ella	p	laundon		710	h	5200	westgate ct	hensh	nc	27606	nc						

M

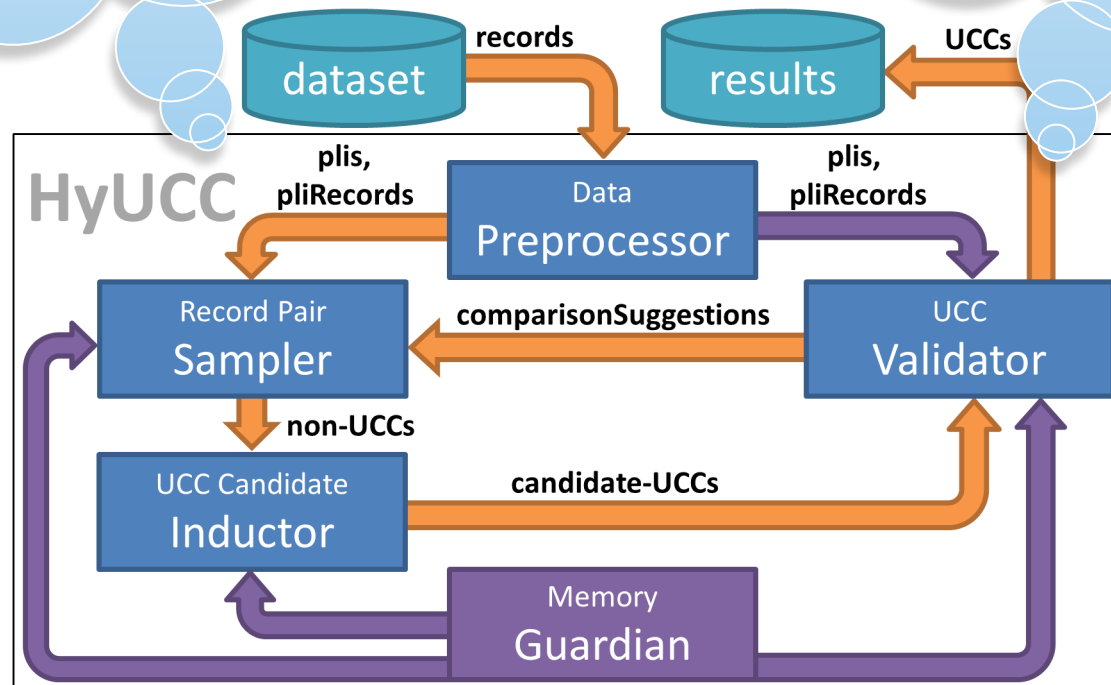
Idea Going Hybrid

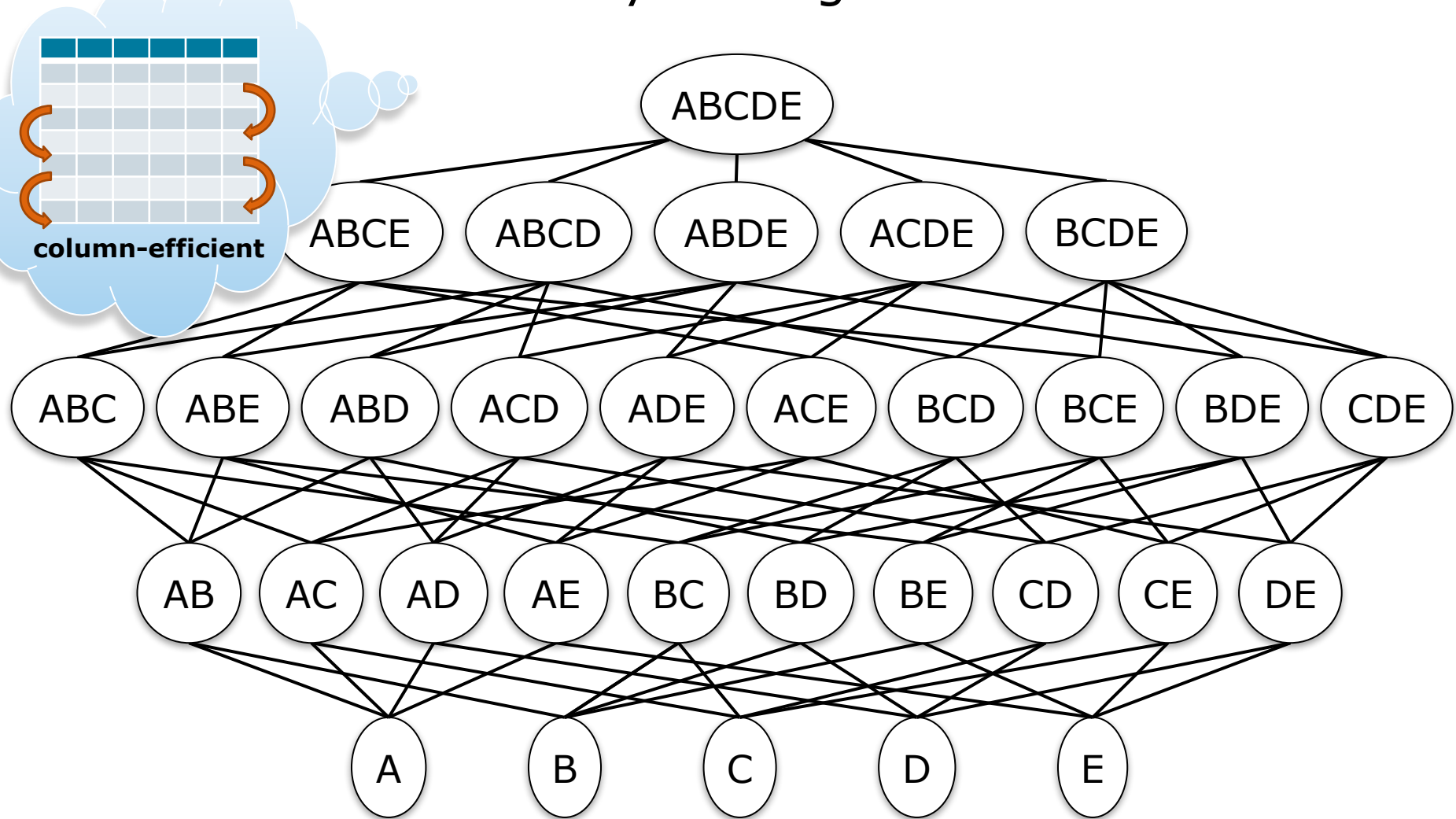
HPI Hasso
Partner
Institut

Same idea as of:
Thorsten Papenbrock, Felix Naumann,
A Hybrid Approach to Functional Dependency Discovery.
SIGMOD Conference 2016: 821-833

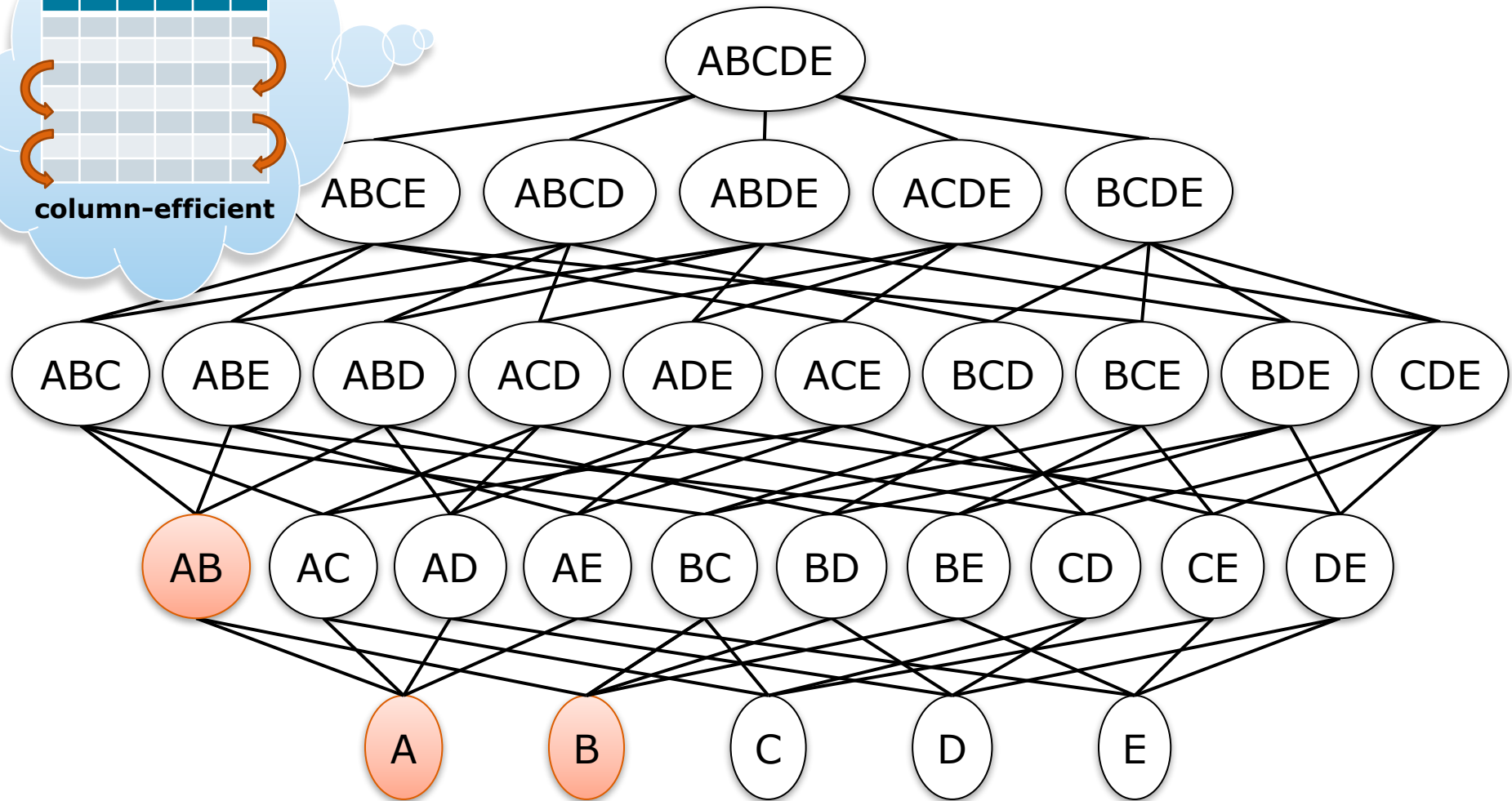
column-efficient

row-efficient

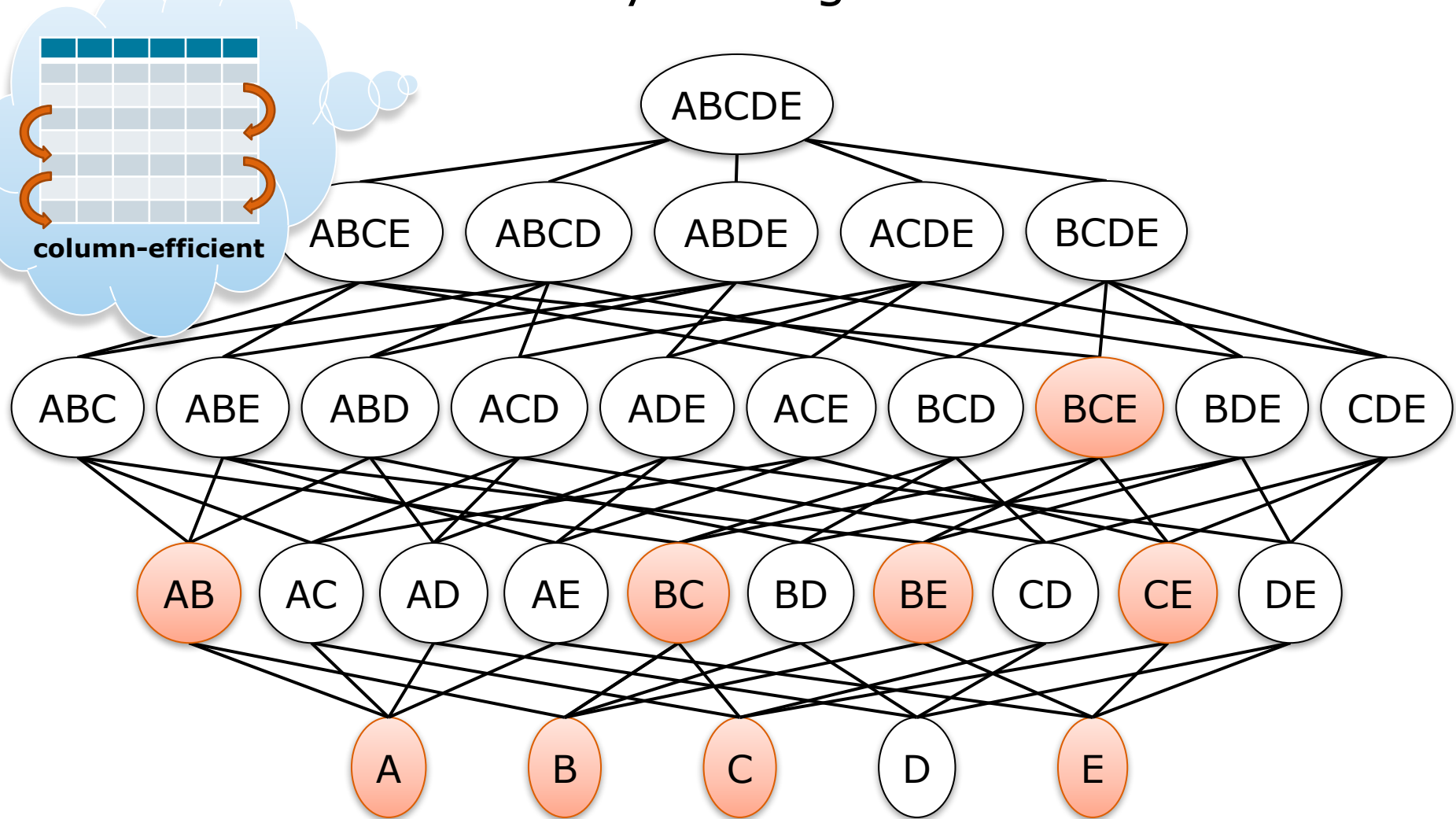




$\neg\{A,B\}$

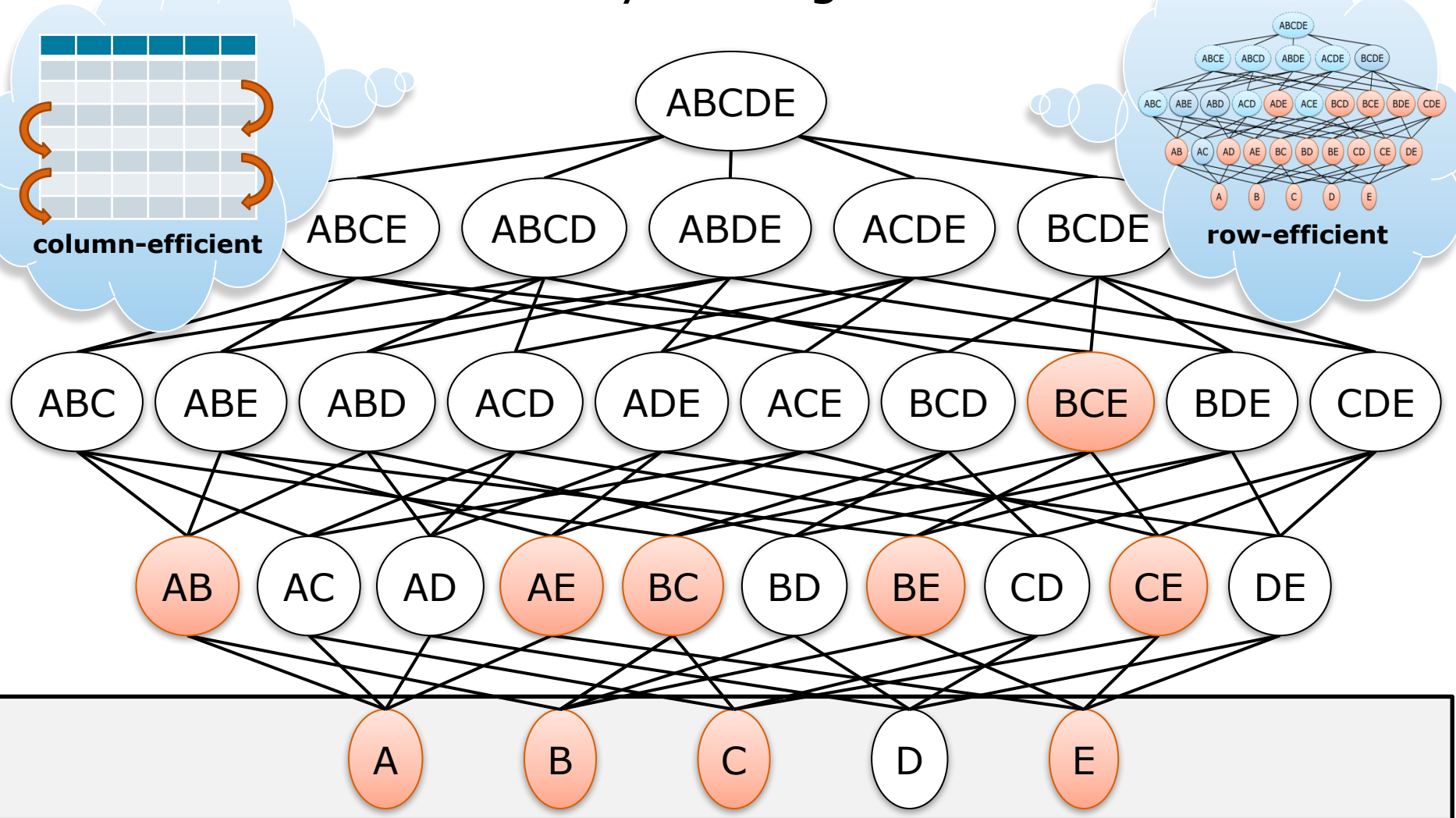


$\neg\{B,C,E\}$



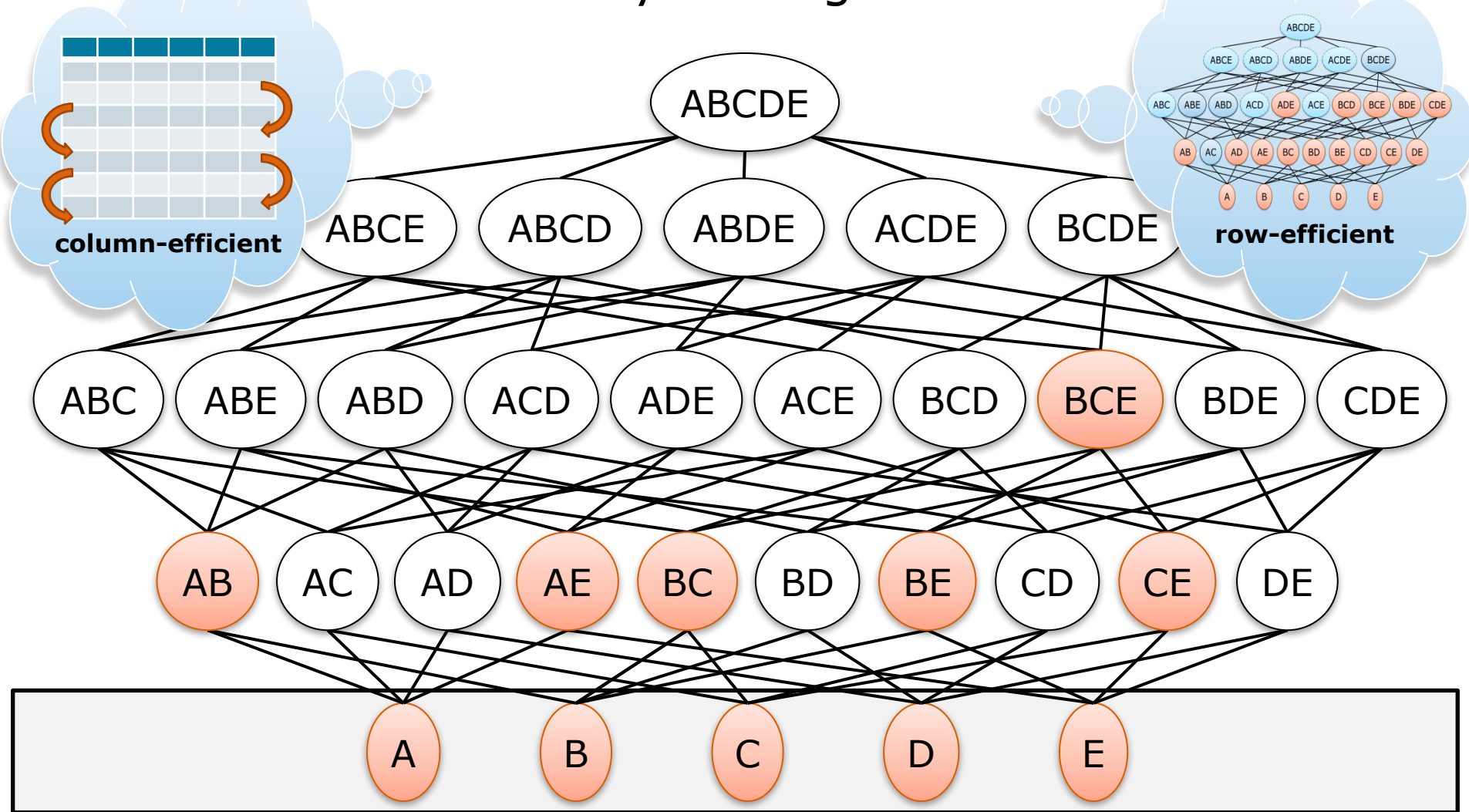
$\neg\{A,E\}$

HyUCC The Hybrid Algorithm

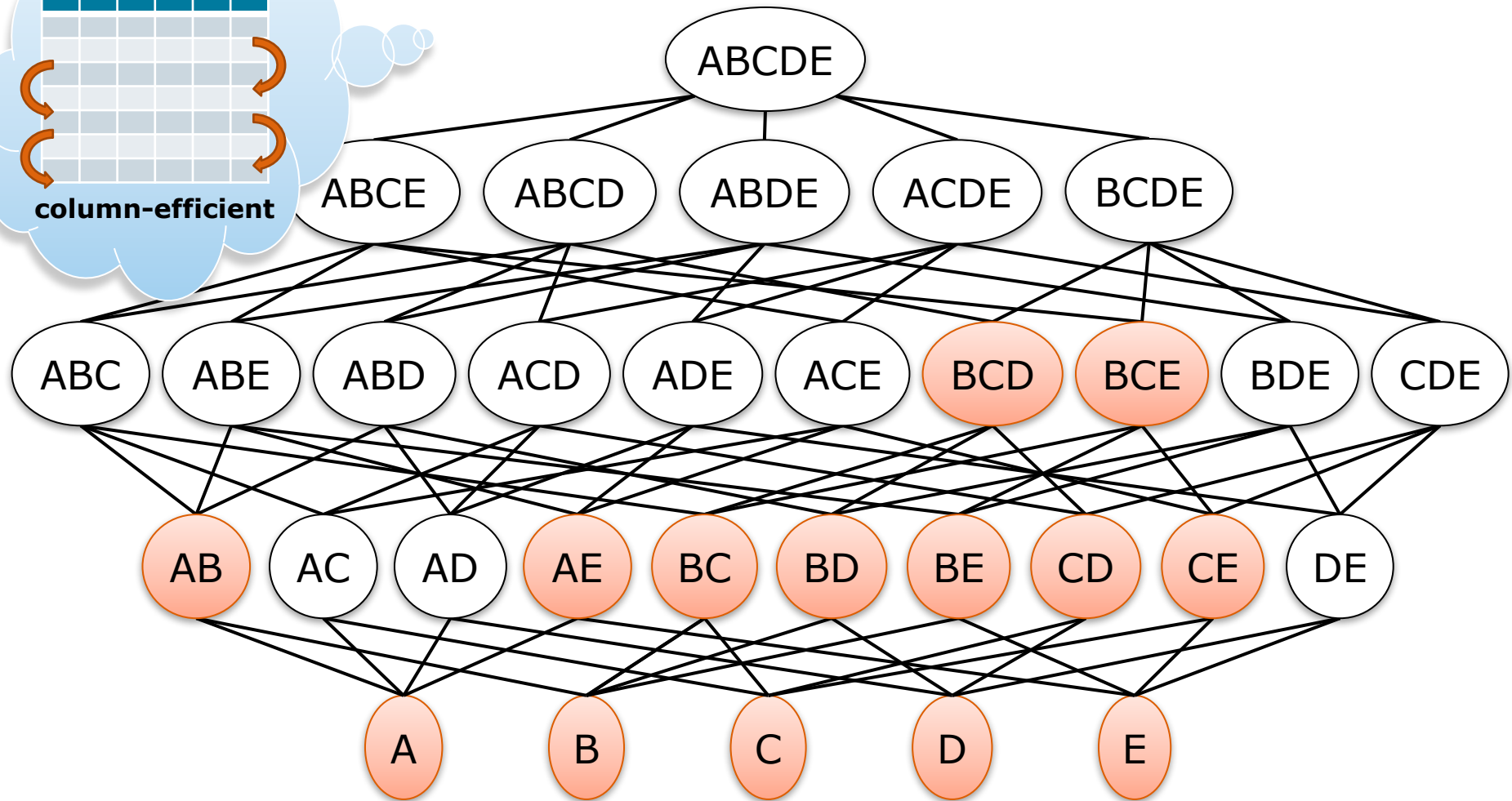


$\neg\{A, E\}$

HyUCC The Hybrid Algorithm

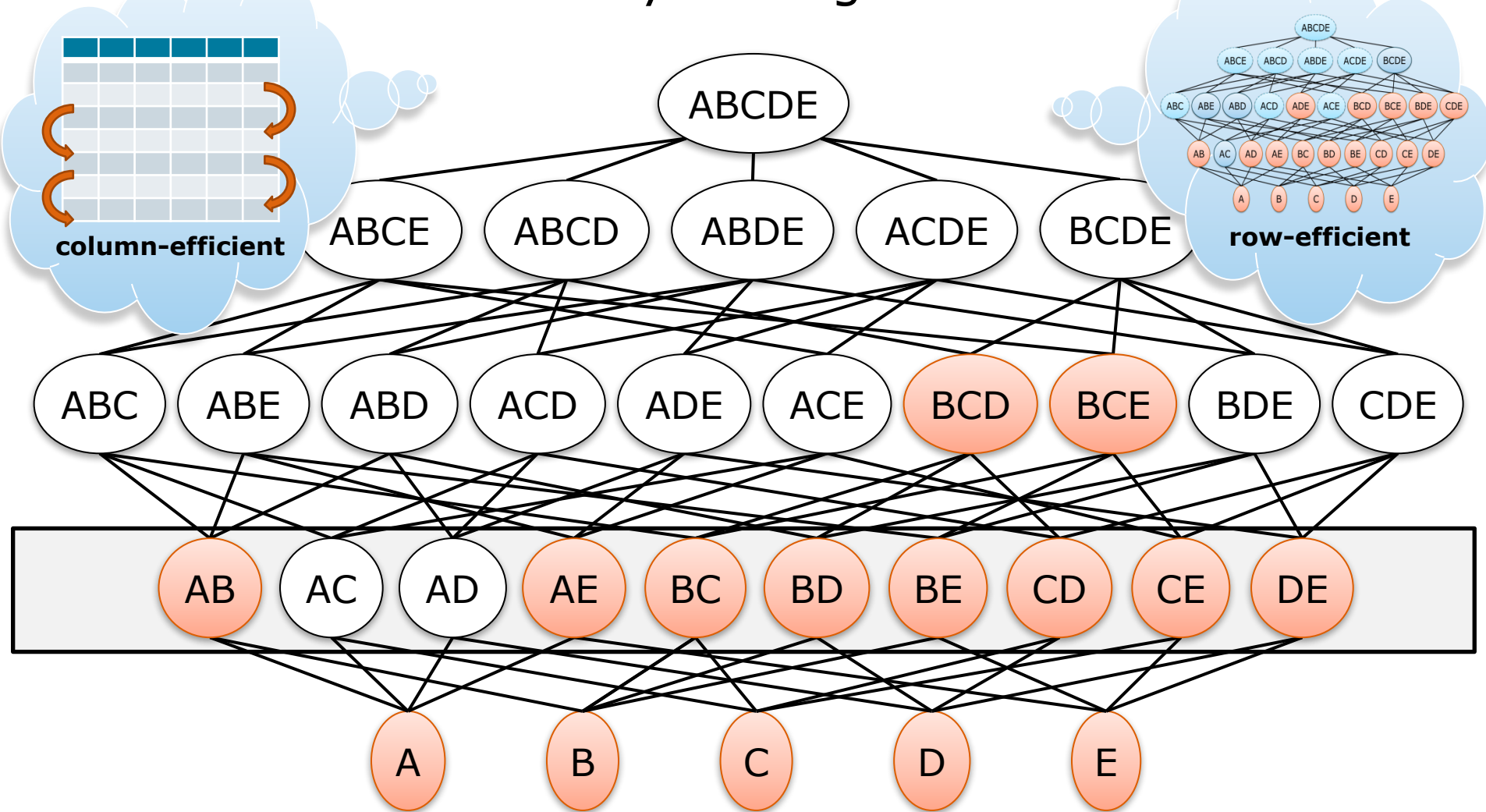


$\neg\{B,C,D\}$

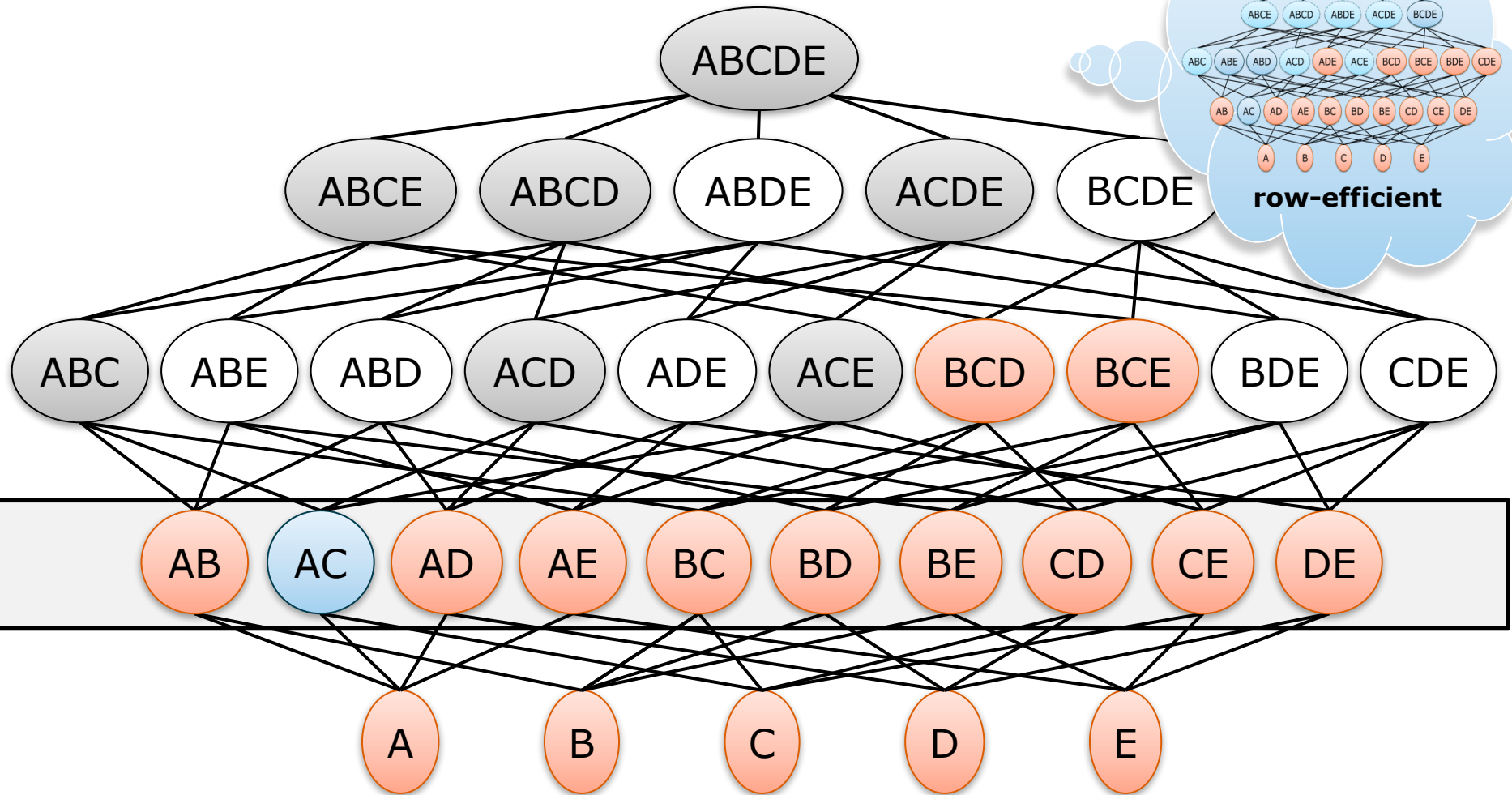


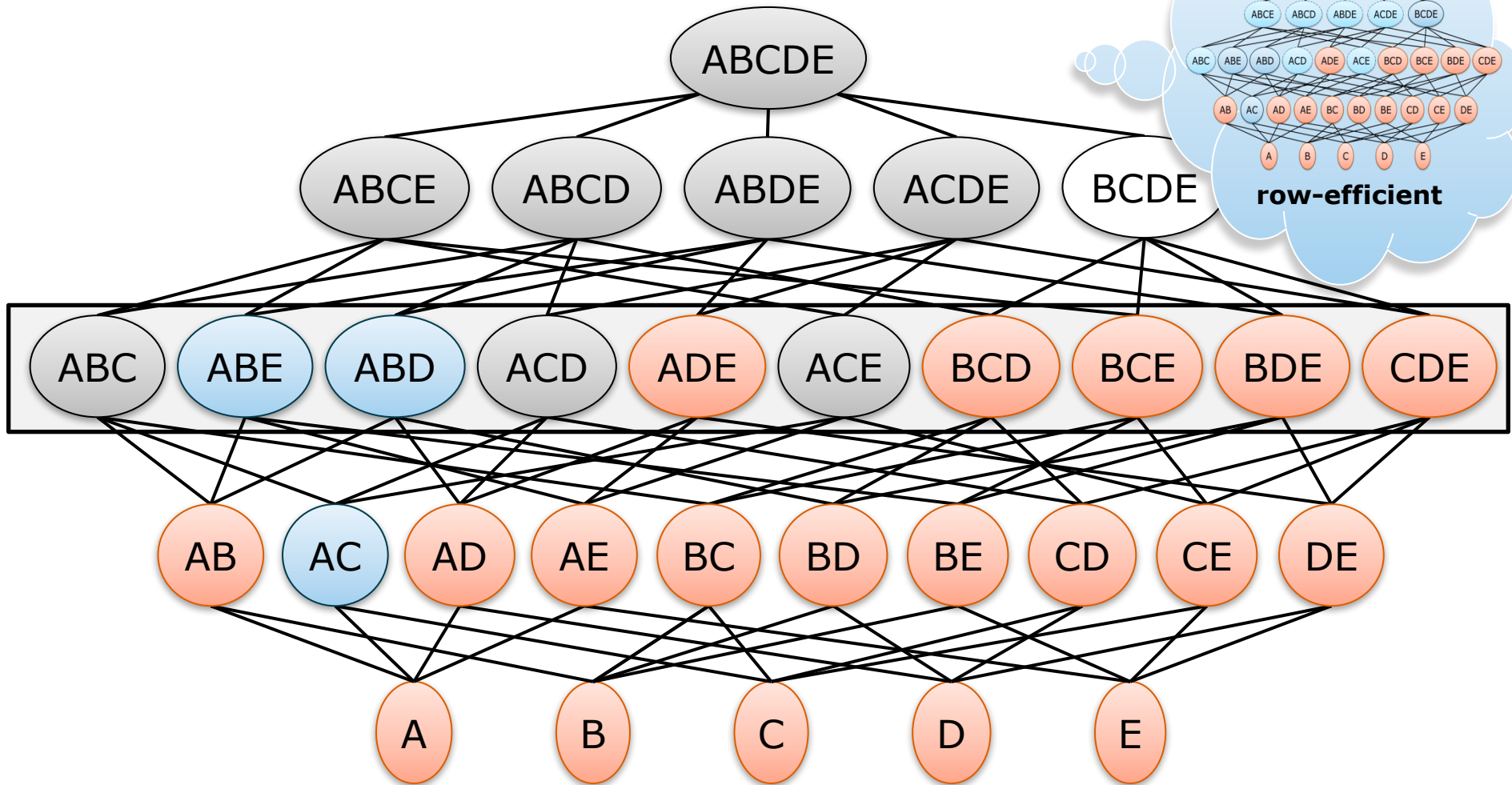
$\neg\{D,E\}$

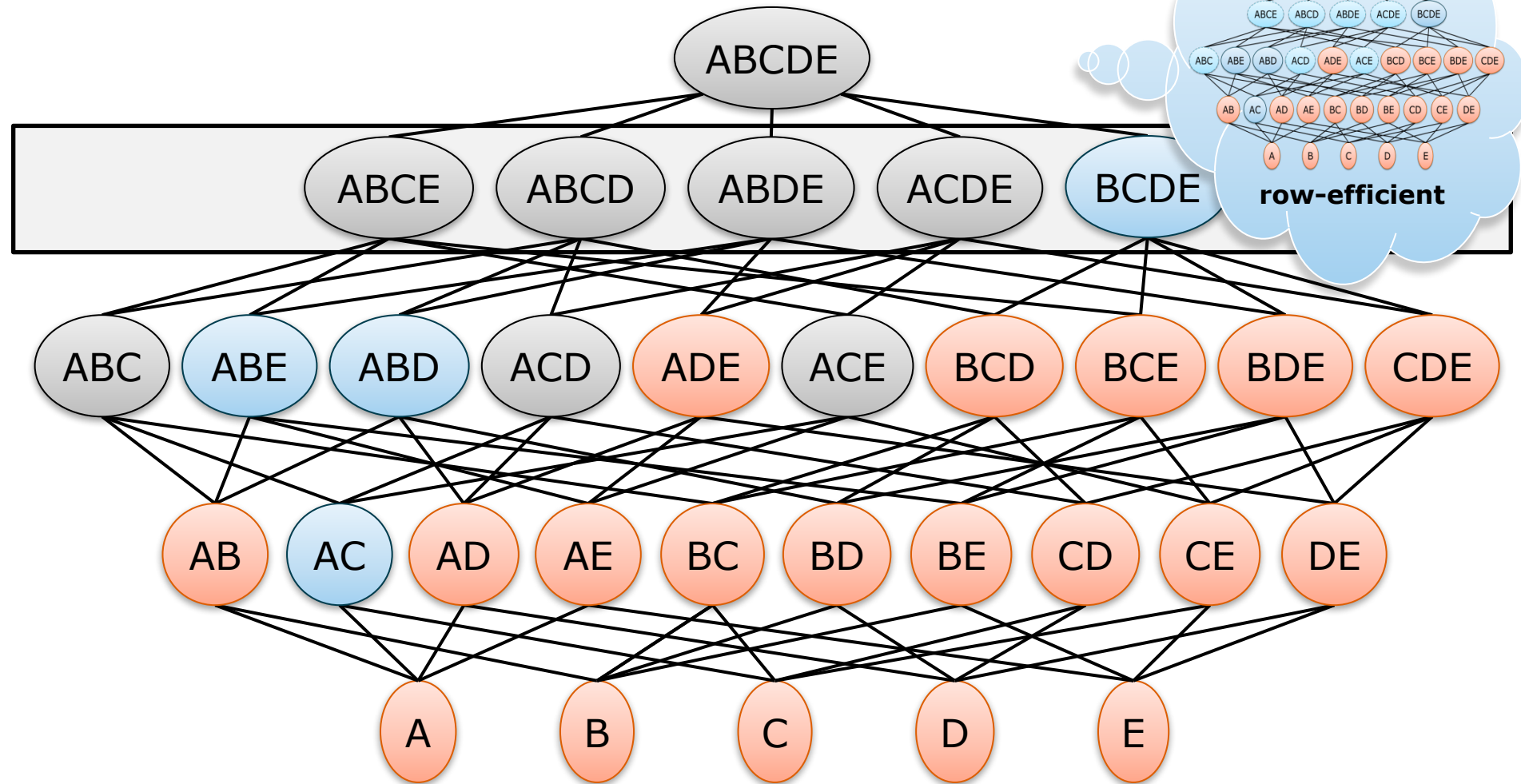
HyUCC The Hybrid Algorithm



HyUCC The Hybrid Algorithm







Runtime in seconds on several real-world datasets

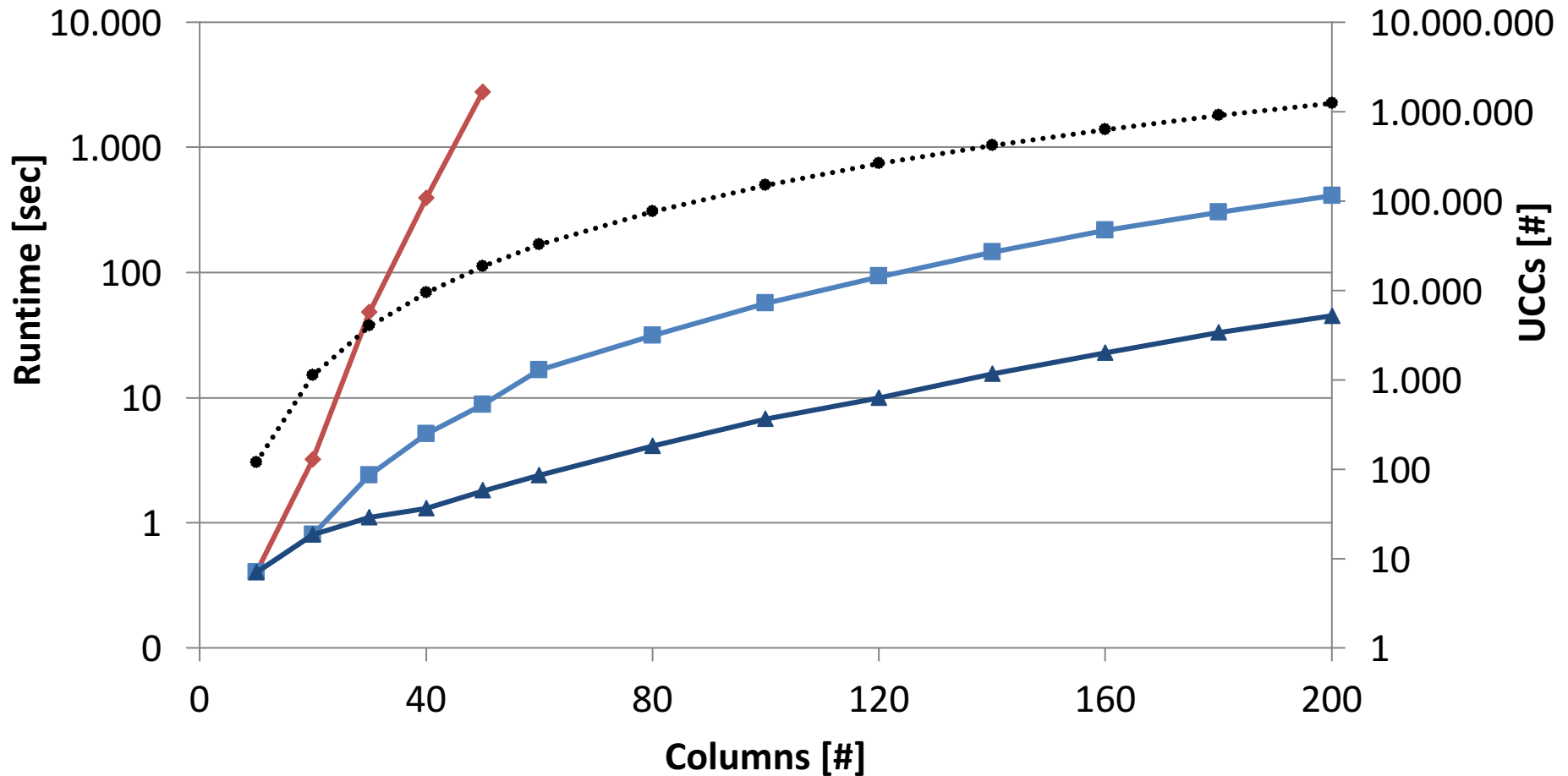
32 threads

Dataset	Cols [#]	Rows [#]	Size [MB]
ncvoter	19	8 m	1,263.5
hepatitis	20	155	0.1
horse	27	368	0.1
zbc00dt	35	3 m	783.0
ncvoter_c	71	100 k	55.7
ncvoter_s	71	7 m	4,167.6
flight	109	1 k	0.6
uniprot	120	1 k	2.1
isolet	200	6 k	12.9

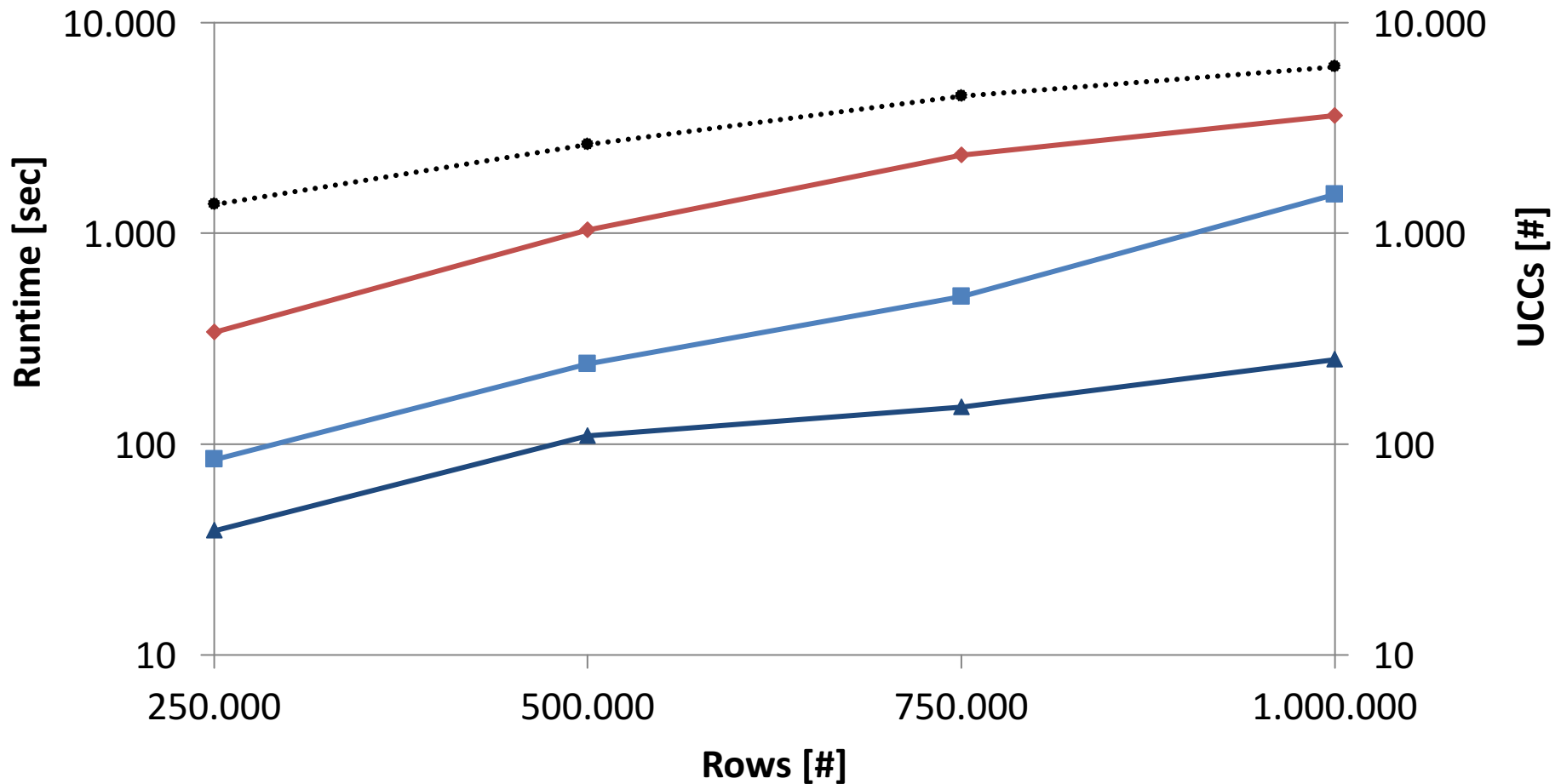
DUCC: Arvid Heise; Jorge-Arnulfo Quiane-Ruiz; Ziawasch Abedjan; Anja Jentzsch; Felix Naumann: *Scalable Discovery of Unique Column Combinations*. In: Proceedings of the International Conference on Very Large Databases (VLDB), 2013.

HyFD: Thorsten Papenbrock; Felix Naumann: *A Hybrid Approach to Functional Dependency Discovery*. In: Proceedings of the International Conference on Management of Data (SIGMOD). 2016.

isolet with 6238 rows

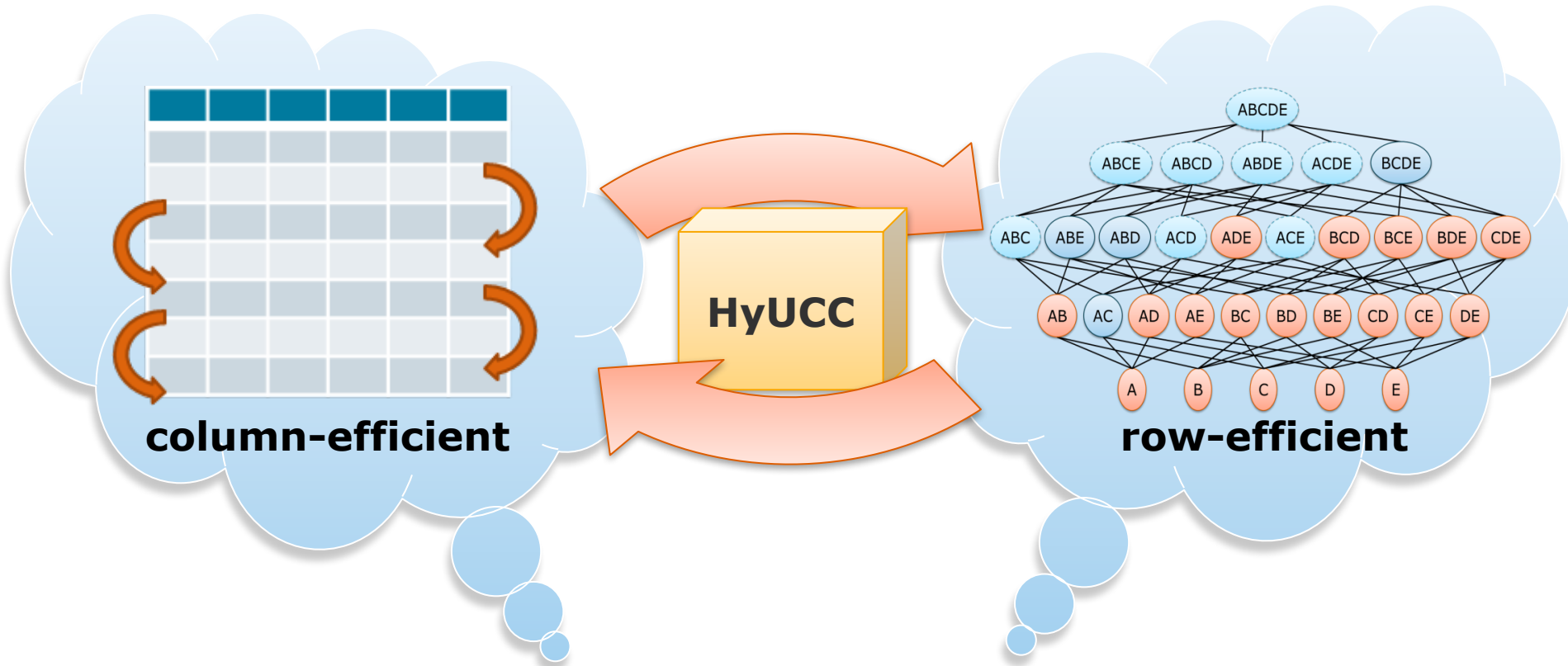


nvcoter_s with 71 columns



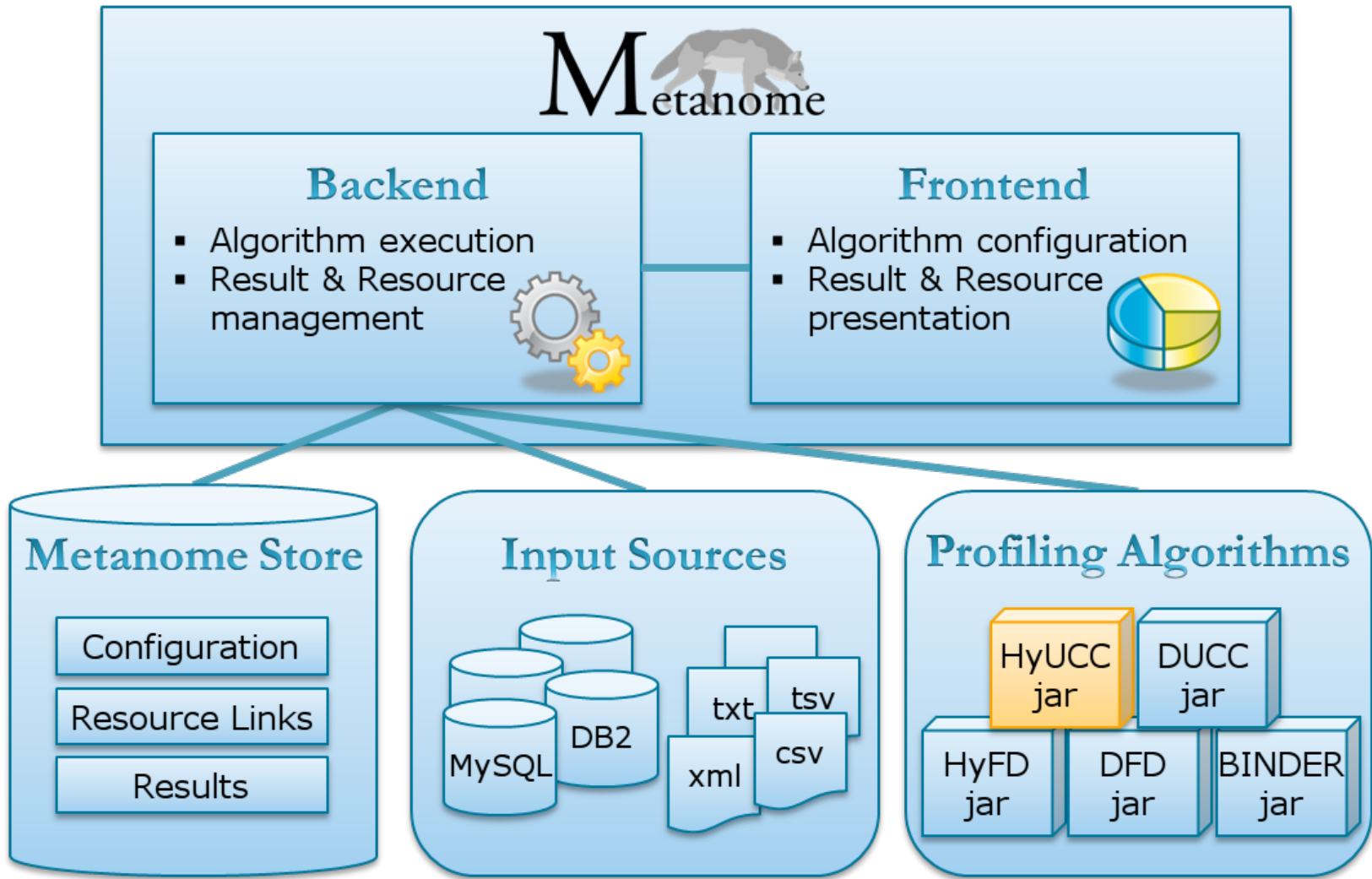
Unique: {Name, Sex, Type}

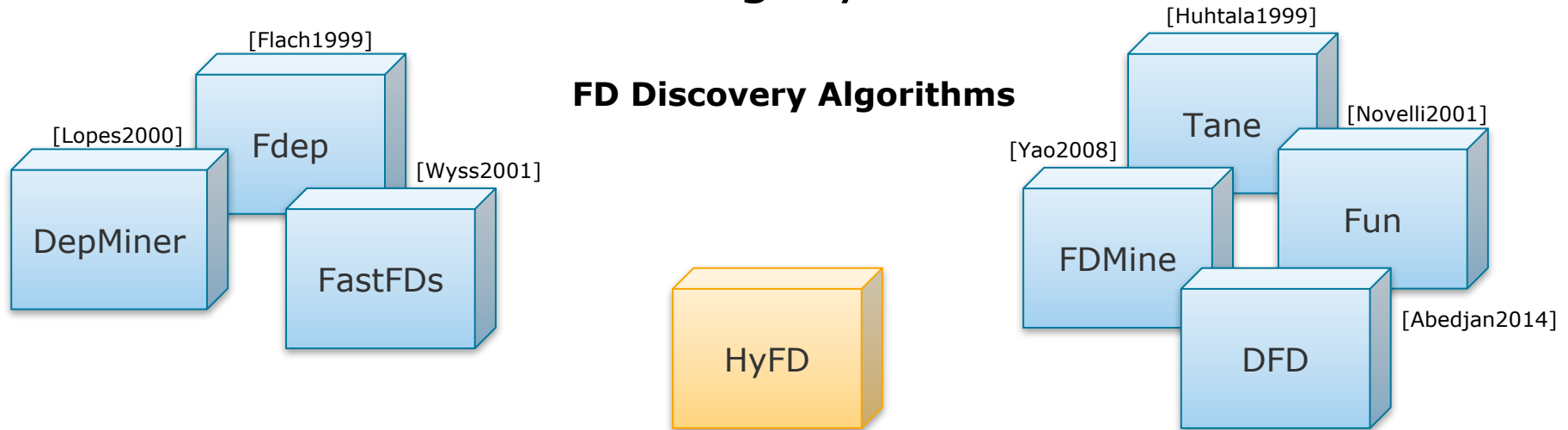
Index	Name	Evolution	Size	Sex	Type	Weakness
25	Pikachu	Raichu	0.4 m	m/w	electric	ground
26	Raichu	null	0.8 m	m/w	electric	ground
29	Nidoran	Nidorino	0.5 m	m	poison	ground
32	Nidoran	Nidorina	0.4 m	w	poison	ground
37	Vulpix	Ninetails	0.6 m	m/w	ice	fire
37	Vulpix	Ninetails	0.6 m	m/w	fire	water
38	Ninetails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



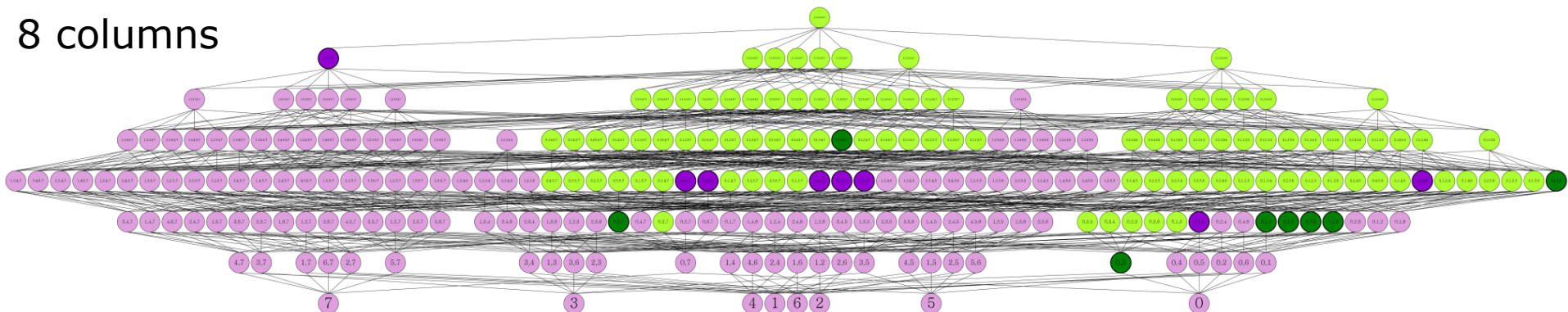
A Hybrid Approach for Efficient Unique Column Combination Discovery

Thorsten Papenbrock and Felix Naumann

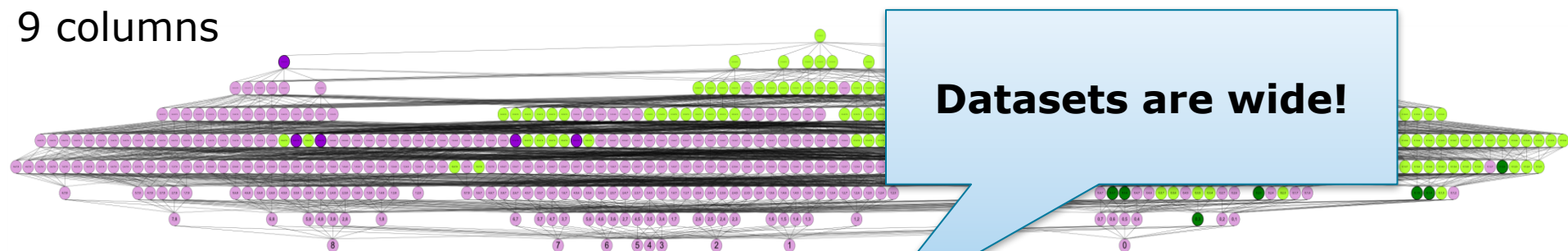




8 columns

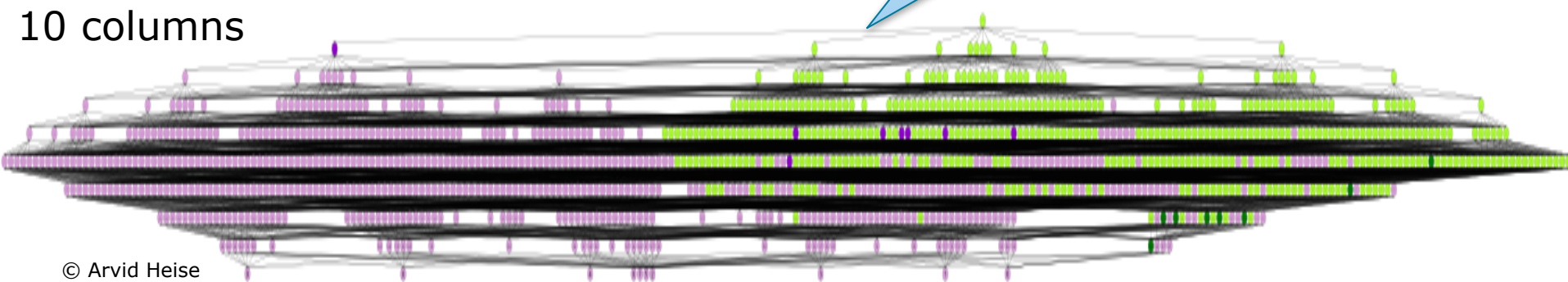


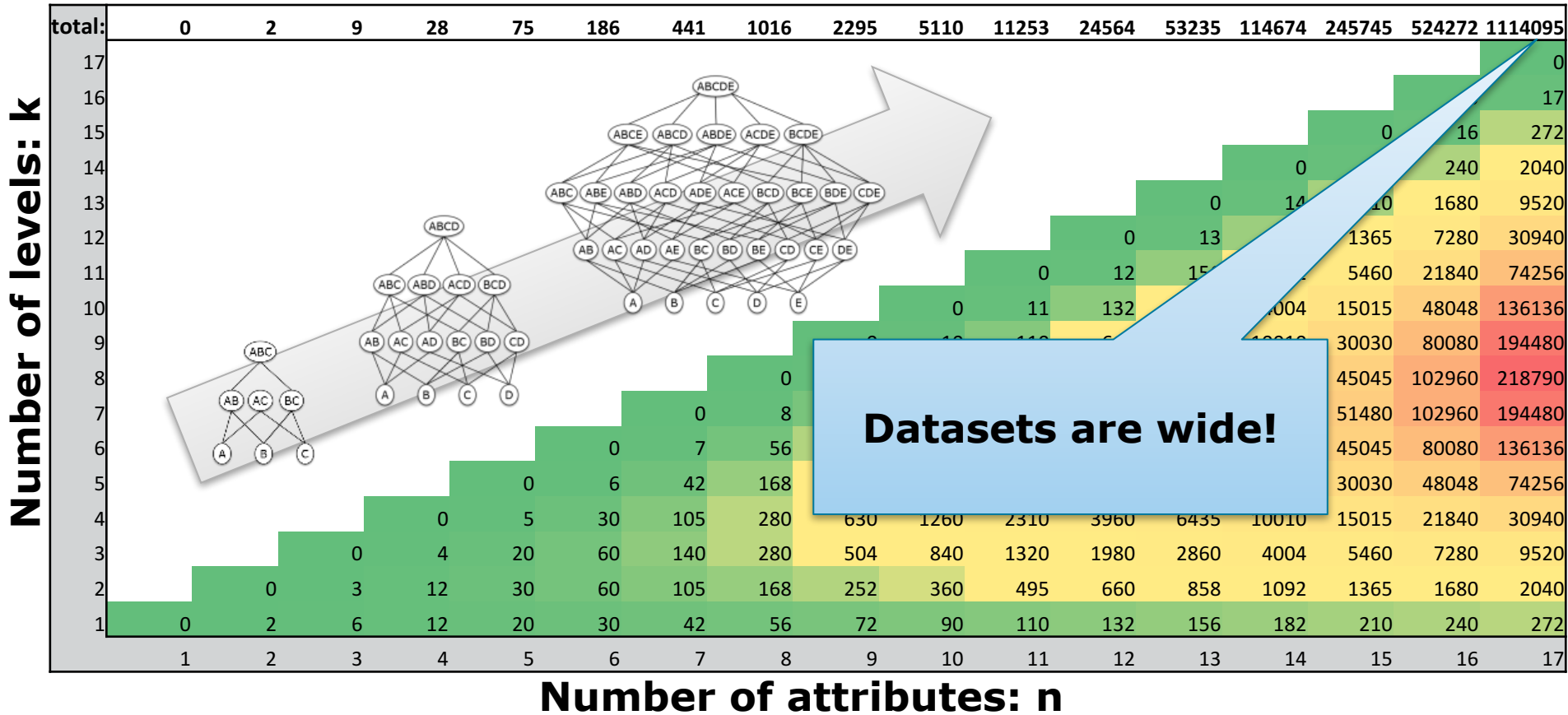
9 columns

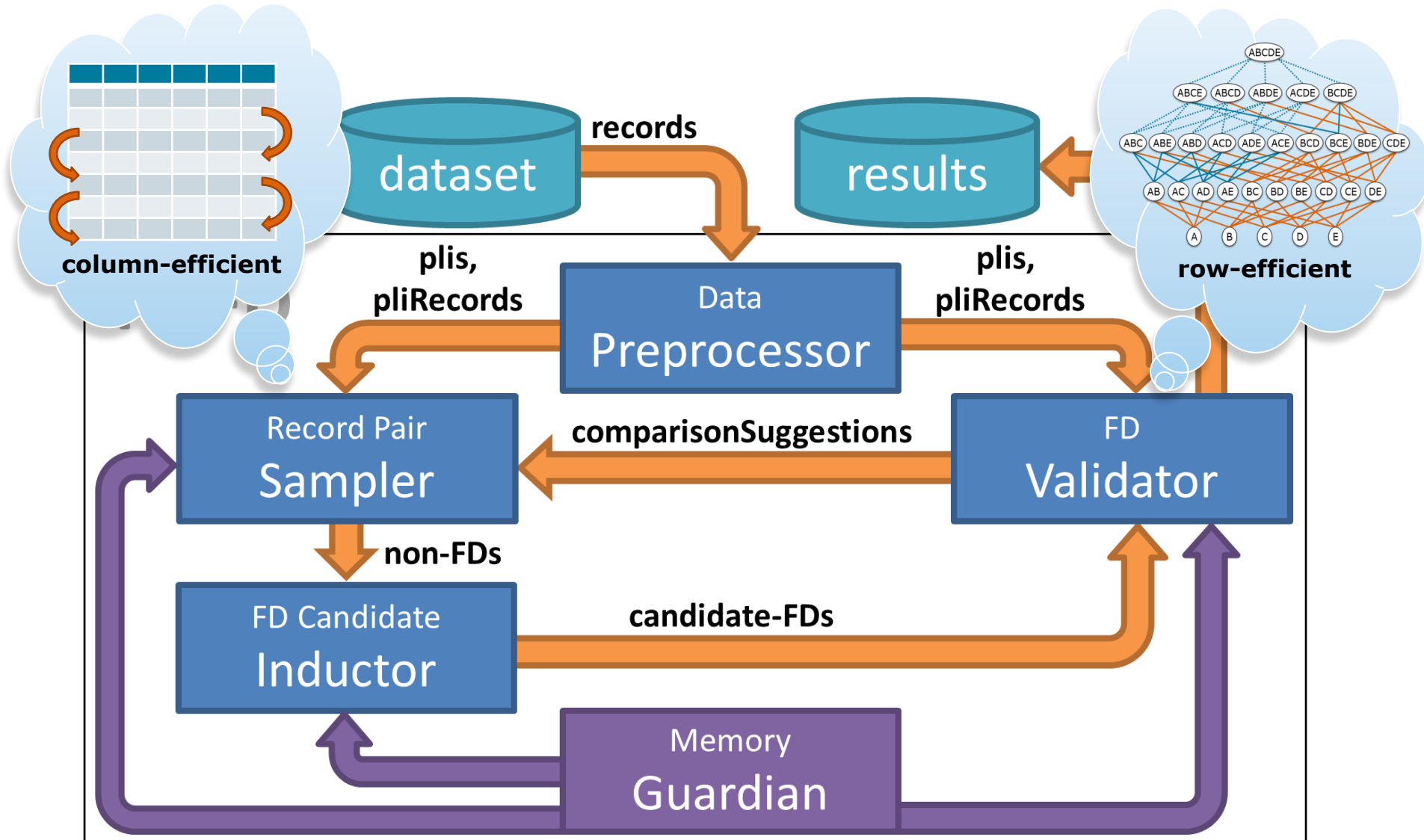


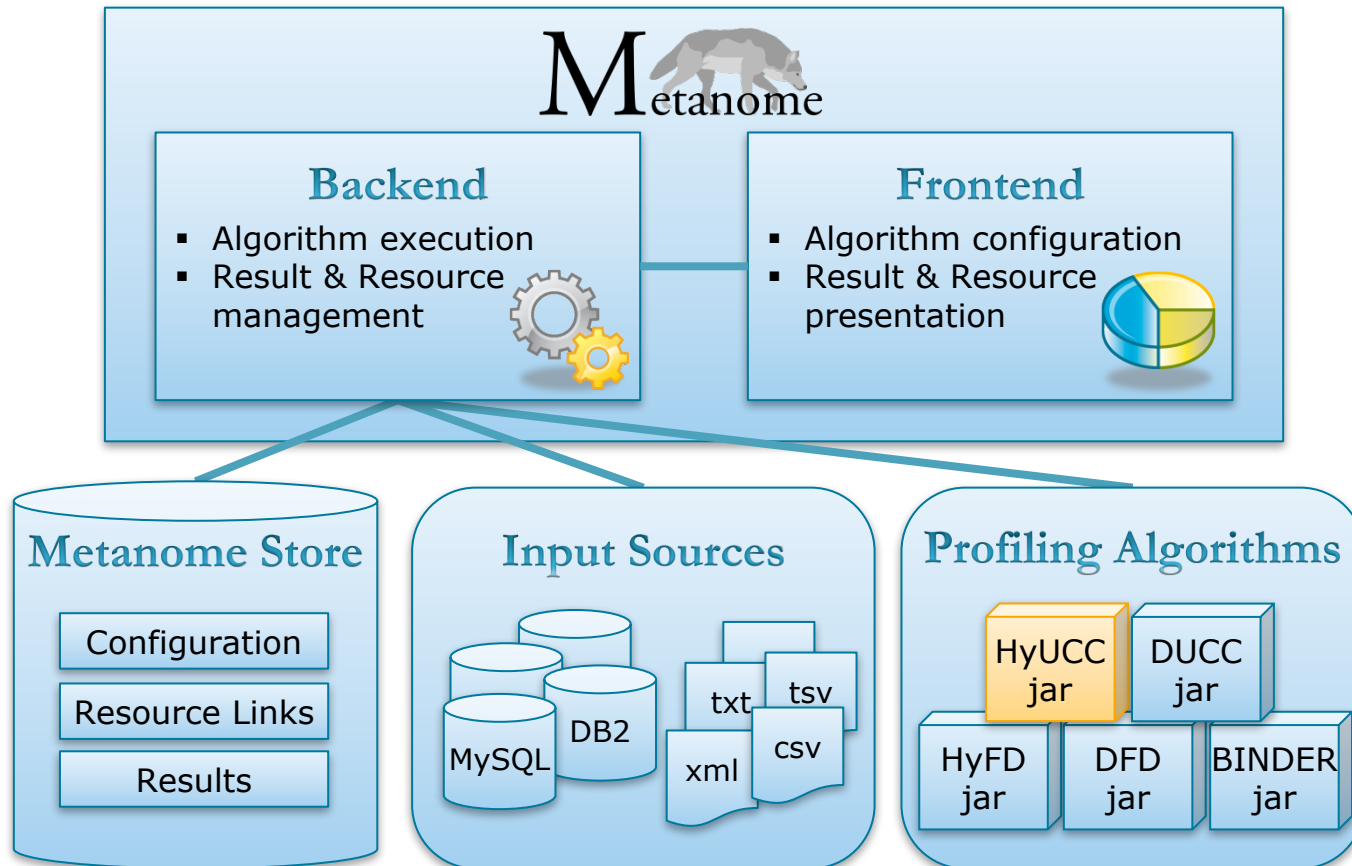
Datasets are wide!

10 columns









www.metanome.de