Metanome

HPI Hasso Plattner Institut

Functional Dependency Discovery:
An Experimental Evaluation of Seven Algorithms

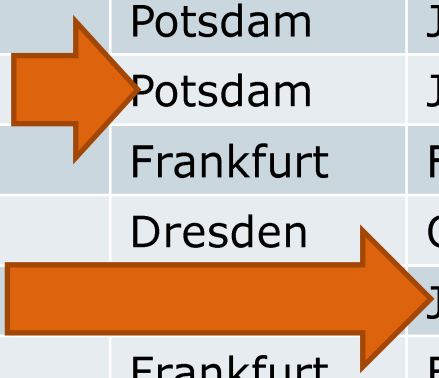
Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert,
Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, Felix Naumann

Metanome

HPI Hasso Plattner Institut

Motivation
Functional Dependencies and Normalization

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

Metanome

HPI Hasso Plattner Institut

Motivation
Functional Dependencies and Normalization

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

Less Memory Consumption
Less Anomaly Vulnerability

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

Metanome

HPI Hasso Plattner Institut

Motivation

Very large Tables e.g. uniprot

name	references_pubmed	keywords_gene_ontology	protein_existence	virus_host_ncbi	sequence	sequence_length	sequence_mass	sequence_checksum	sequence_modified	taconomic	taxonomic_lineage_protein_names	gene_orf	cross_reference_cross_refe
017L_FRG3G	15165820	Complete proteome; Reference	predicted	8295; 8316; 8404; 3034	METMSDYSKEVSE	502	53469	A747EE6F952CBAD7	2004-07-19+01:00	654924	Viruses > dsDNA viruse Uncharacterized p	FV3-017L	AY548484; AAT0967(YP_031595.1
040R_FRG3G	15165820	Complete proteome; Reference	predicted	8295; 8316; 8404; 3034	MIRALCTIVLIAAG\	182	19577	FCF64F420039CF35	2004-07-19+01:00	654924	Viruses > dsDNA viruse Uncharacterized p	FV3-040R	AY548484; AAT0969(YP_031618.1
057L_IIV3													
078L_FRG													
1001R_AS													
11014_AS													
1107L_AS													
124R_IIV3													
14331_CA													
14335_AR													
1433B_VI													
1433T_BO													
1433_NEC													
14KL_BRU													
17KD_RIC													
1A11_ORY													
1A1D_BU													
1A1D_PSE													
1A43_HUI													
1B07_HUI													
1B67_HUI													
1FEH_AEC													
242L_IIV6													
2A5D_HU													
2AAB_PIG													
2ABD_DA													
2B_PEBV													
2SS1_BRA													
3001L_AS													
348R_IIV6													
36015_AS													
3603L_AS													
385L_IIV6													
3BHS_HO													
3DHQ1_A													
3DHQ_SC													
3HAO_CH													
3HAO_XA													
3MGA_BA													
3MGH_BO													
3MGH_CL													
3MGH_FR													
3MGH_MY													
3MGH_RH													
3MGH_ST													
3MG_HUN													
441R_IIV6													
4CL4_ORY													
4EBP3_HU													
50510_AS													
5056R_AS													
5E5_RAT													
5HT1B_RA													
5HT2A_PI													
5HT5A_HU													
5NT3L_MOUSE	15489334; 16141072; 1946	Acetylation; Alternative splicing	evidence at transcript level		MKATVLMRQPGR\	292	33578	3AE26BCE06833085	2008-04-08+01:00	10090	Eukaryota > Metazoa > Cytosolic 5'-nucleotidase III-li		BC015307; AAH1530 NP_080837.3
60A_DROVI	8688461	Cytokine; Disulfide bond; Glyco	inferred from homology		MTASLVVLPSLWLI	436	49999	C744B4AE58796692	1997-11-01Z	7244	Eukaryota > Metazoa > Protein 60A		U48595; AAC47262.1; Genomic_D

accession, name, references_pubmed, keywords, gene_ontology, protein_existence, virus_host_ncbi, sequence, sequence_length, sequence_mass, sequence_checksum, sequence_modified, taconomic_identifier, taxonomic_lineage, protein_names, gene_orf, cross_reference_embl, cross_reference_refseq, cross_reference_geneid, cross_reference_protclustdb, feature_chain, feature_compositionally_biased_region, feature_signal_peptide, gene_ordered_locus, comment_function, comment_similarity, cross_reference_interpro, cross_reference_pfam, feature_glycosylation_site, cross_reference_go, feature_transmembrane_region, feature_coiled-coil_region, comment_subcellular_location, gene_primary, gene_synonym, cross_reference_pir, cross_reference_proteinmodelportal, cross_reference_smr, cross_reference_dip, cross_reference_intact, cross_reference_mint, cross_reference_string, cross_reference_paxdb, cross_reference_pride, cross_reference_ensemblmetazoa, cross_reference_kegg, cross_reference_ucsc, cross_reference_ctd, cross_reference_wormbase, cross_reference_eggnog, cross_reference_genetree, cross_reference_hogenom, cross_reference_inparanoid, cross_reference_ko, cross_reference_oma, cross_reference_nextbio, cross_reference_arrayexpress, cross_reference_gene3d, cross_reference_panther, cross_reference_pirsf, cross_reference_prints, cross_reference_smart, cross_reference_supfam, cross_reference_prosite, feature_sequence_conflict, comment_subunit, comment_interaction, cross_reference_ipi, cross_reference_unigene, cross_reference_ensemblplants, cross_reference_genomereviews, cross_reference_tair, cross_reference_genevestigator, cross_reference_germonline, feature_modified_residue, cross_reference_ensembl, cross_reference_hovergen, cross_reference_orthodb, feature_site, cross_reference_patric, cross_reference_biocyc, feature_lipid_moiety-binding_region, cross_reference_ec, cross_reference_gramene, cross_reference_unipathway, comment_catalytic_activity, comment_cofactor, comment_pathway, comment_induction, cross_reference_hamap, cross_reference_tigrfams, feature_active_site, cross_reference_pdb, cross_reference_pdbsum, cross_reference_evolutionarytrace, feature_helix, feature_strand, feature_turn, cross_reference_dmdm, cross_reference_genecards, cross_reference_hgnc, cross_reference_mim, cross_reference_nextprot, cros[illegible]_domain, feature_region_of_interest, feature_disulfide_bond, feature_sequence_variant, comment_ptm, comment[illegible]armgkb, cross_reference_reactome, cross_reference_genomernai, cross_reference_bgee, cross_reference_h[illegible]ssp, cross_reference_phosphosite, cross_reference_dnasu, cross_reference_phylomedb, cross_reference_path[illegible]nce_motif, feature_splice_variant, comment_alternative_products, comment_tissue_specificity, feature_non-termin[illegible]re_non-consecutive_residues, cross_reference_prodom, feature_initiator_methionine, feature_binding_site, [illegible]site, comment_sequence_caution, cross_reference_ensemblbacteria, cross_reference_genolist, cor[illegible], comment_online_information, feature_nucleotide_phosphate-binding_region, feature_mutagenesis_site[illegible], cross_reference_flybase, feature_propeptide, feature_transit_peptide, cross_reference_tuberculist, cros[illegible]merops, comment_developmental_stage, cross_reference_arachnoserver, comment_toxic_dose, cross_reference_cazy, cross_reference_h[illegible]ross_reference_cygd, cross_reference_sgd, cross_reference_peptideatlas, cross_reference_drugbank, cross_reference_tcdb, comment_biophysicoche[illegible]properties, cross_reference_promex, cross_reference_pombase, cross_reference_swiss-2dpage, cross_reference_echobase, cross_reference_ecogene, cross_r[illegible]rence_rgd, cross_reference_pmap-cutdb, comment_caution, feature_zinc_finger_region, feature_cross-link, cross_reference_vectorbase, feature_unsure_residue, feature_dna-binding_region, cross_reference_pseudocap, cross_reference_orphanet, comment_disease, cross_reference_ensemblprotists, cross_reference_dictybase, cross_reference_world-2dpage, cross_reference_xenbase, feature_calcium-binding_region, comment_mass_spectrometry, feature_peptide, cross_reference_genefarm, cross_reference_peroxibase, cross_reference_maizegdb, cross_reference_allergome, comment_allergen, cross_reference_glycosuitedb, cross_reference_cornea-2dpage, cross_reference_dosac-cobs-2dpage, cross_reference_ogp, cross_reference_reproduction-2dpage, cross_reference_siena-2dpage, cross_reference_ucd-2dpage, cross_reference_agd, cross_reference_bindingdb, cross_reference_phossite, cross_reference_legiolist, feature_intramembrane_region, cross_reference_cgd, cross_reference_disprot, comment_rna_editing, comment_biotechnology, cross_reference_conoserver, cross_reference_pmma-2dpage, cross_reference_mycoclap, comment_pharmaceutical, cross_reference_phci-2dpage, cross_reference_eupathdb, cross_reference_leproma, cross_reference_rebase, cross_reference_compluyeast-2dpage, feature_non-standard_amino_acid, cross_reference_aarhus/ghent-2dpage, cross_reference_euhcvdb, cross_reference_rat-heart-2dpage, cross_reference_pptasedb, cross_reference_anu-2dpage, cross_reference_2dbase-ecoli

223 attributes!

Metanome

HPI Hasso Plattner Institut

Motivation

Very large Tables e.g. uniprot

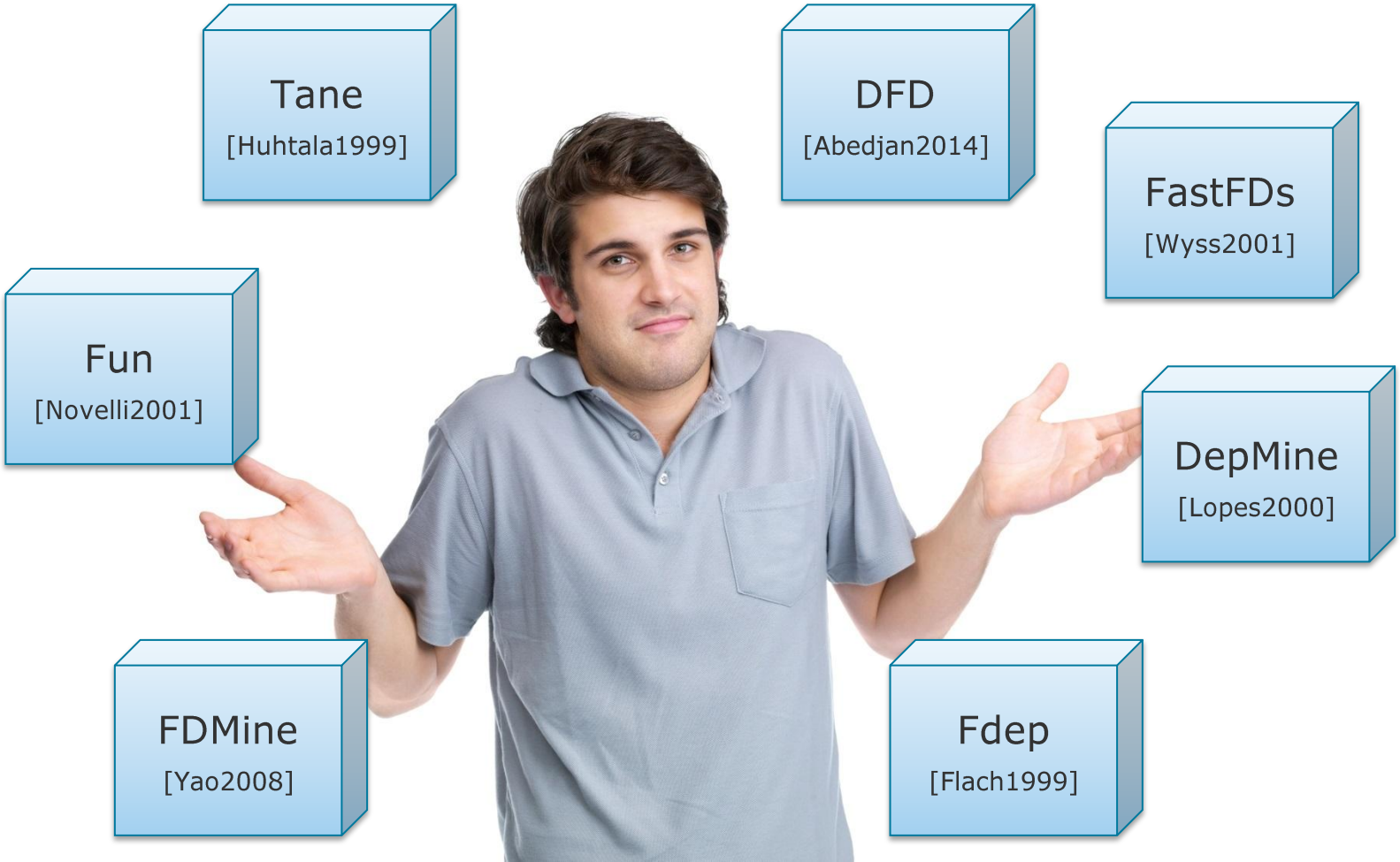
name	references_pubmed	keywords	gene_ontology	protein_existence	virus_host_ncbi	sequence	sequence_length	sequence_mass	sequence_checksum	sequence_modified	taconomic_	taxonomic_lineage	protein_names	gene_orf	cross_reference	cross_refe
017L_FRG3G	15165820	Complete proteome; Reference		predicted	8295; 8316; 8404; 30343	METMSDYSKEVSE	502	53469	A747EE6F952CBAD7	2004-07-19+01:00	654924	Viruses > dsDNA viruse	Uncharacterized p	FV3-017L	AY548484; AAT0967(	YP_031595.1
040R_FRG3G	15165820	Complete proteome; Reference		predicted	8295; 8316; 8404; 30343	MIRALCTIVLIAAG\	182	19577	FCF64F420039CF35	2004-07-19+01:00	654924	Viruses > dsDNA viruse	Uncharacterized p	FV3-040R	AY548484; AAT0969(	YP_031618.1
057L_IIV3	16912294	Complete proteome; Reference		predicted	7163; 7183; 42431; 3105	MFKIYRTSCMGQH	130	14846	0A95A328C7ABA865	2006-07-11+01:00	345201	Viruses > dsDNA viruse	Uncharacterized p	IIV3-057L	DQ643392; ABF8208	YP_654629.1
078L_FRG3G	15165820	Complete proteome; Reference		predicted	8295; 8316; 8404; 30343	MSIGETFAISAHPE	212	23985	920D63B5324F6A2F	2004-07-19+01:00	654924	Viruses > dsDNA viruse	Uncharacterized p	FV3-078L	AY548484; AAT0973(	YP_031657.1
1001R_ASFWA		Complete proteome		inferred from homology	6937; 9823; 41426; 855	MVRLFRNPIKCIFY	124	15327	C1EC5CC5B6D3E2BB	2009-05-05+01:00	561444	Viruses > dsDNA viruse	Protein MGF 100-1R		AY261366; NOT_ANNOTATED_CD	
11014_ASFK5		Complete proteome; Signal		inferred from homology	6937; 9823; 41426; 855	MKVLLGLLLGYSVLI	120	13822	4ECDA03DD16874D9	2009-05-05+01:00	561445	Viruses > dsDNA viruse	Protein MGF 110-14L		AY261360; NOT_ANNOTATED_CD	
1107L_ASFM2		Complete proteome; Glycoprote		inferred from homology	6937; 9823; 41426; 855	MLVIILGVIGLLASS	125	14238	956AFED9727E2A28	2009-05-05+01:00	10500	Viruses > dsDNA viruse	Protein MGF 110-7L		AY261361; NOT_ANNOTATED_CD	
124R_IIV3	16912294	Coiled coil; Complete proteome		predicted	7163; 7183; 42431; 3105	MESTDVNHSTAW/	236	25428	CDEB4CF424704DCA	2006-07-11+01:00	345201	Viruses > dsDNA viruse	Uncharacterized p	IIV3-124R	DQ643392; ABF8215	YP_654696.1
14331_CAEEL	7926802; 9851916; 167776	Complete proteome; Cytoplasm		evidence at protein level		MSDTVEELVQRAKI	248	28191	ABBE0DA27D9341AF	2008-07-22+01:00	6239	Eukaryota > Metazoa >	14-3-3-like proteir	M117.2	Z73910; CAA98138.1	NP_502235.1
14335_ARATH	7972511; 9276953; 111307	Complete proteome; Phosphop		evidence at protein level		MSSDSSREENVYL/	268	30182	ECB35B7E53F688FF	1996-10-01+01:00	3702	Eukaryota > Viridiplant	14-3-3-like proteir	F1N13_190	AY059115; AAL15221	NP_568325.1
1433B_VICFA				evidence at transcript level		MASTKDRENFVYI/	261	29524	2981BA7AE4C7E314	1995-11-01Z	3906	Eukaryota > Viridiplant	14-3-3-like protein B		Z48505; CAA88416.1; mRNA	
1433T_BOVIN	16305752	Acetylation; Complete proteom		evidence at transcript level		MEKTELIQKAKLAE(	245	27764	175534325E9E37C4	2005-10-11+01:00	9913	Eukaryota > Metazoa >	14-3-3 protein theta		BC102840; AAI02841	NP_001071
1433_NEOCA	8992315			evidence at transcript level		MAEEIKNLRDEYVY	266	30650	6DA471FE7ECC7C04	1996-11-01Z	29176	Eukaryota > Alveolata	14-3-3 protein homolog		U31542; AAC47012.1; mRNA	
14KL_BRUO2		Cell membrane; Complete prote		inferred from homology		MNSFRKTCAGALAI	147	16823	7CEE6F1FB2BB38DB	2007-07-10+01:00	444178	Bacteria > Proteobacte	Lectin-like protein BA14k		CP000709; ABQ6233(	YP_00125768
17KD_RICTY	2768201; 15317790	Cell membrane; Cell outer mem		inferred from homology		MKLLSKVMILALAA	159	16539	A9893F26492727A9	2004-11-09Z	257363	Bacteria > Proteobacte	17 kDa surface antigen		AE017197; AAU0427	YP_067758.1
1A11_ORYSI	1438312; 8389618; 156852	Ethylene biosynthesis; Fruit rip		evidence at transcript level		MVSQVVAEEKPQLI	487	53139	561D262125D7759E	2007-06-12+01:00	39946	Eukaryota > Viridiplant	1-aminocycloprop	Osl_012955	CM000128; EAY91722.1;	Genomic_
1A1D_BURCH		Complete proteome; Hydrolase		inferred from homology		MNLQRFPRYPLTFC	338	36627	92D48388C116C83C	2006-11-28Z	331272	Bacteria > Proteobacte	1-aminocyclopropane-1-carbo		CP000459; ABK10113	YP_837006.1
1A1D_PSEUD	1885510	3D-structure; Direct protein seq		evidence at protein level		MNLQRFPRYPLTFC	338	36672	01FC286177012FDF	1993-04-01+01:00	74568	Bacteria > Proteobacte	1-aminocyclopropane-1-carbo		M73488; AAA25689.1;	Genomic_
1A43_HUMAN	1431115; 8475492; 105193	Complete proteome; Disulfide		evidence at transcript level		MAVMAPRTLVLLLS	365	41033	213FC69745FB4B0B	1993-04-01+01:00	9606	Eukaryota > Metazoa >	HLA class I histocompatibility		AH005526; AAB65408.1;	Genomic
1B07_HUMAN	518865; 2320591; 2700944	Complete proteome; Direct prot		evidence at protein level		MLVMAPRTVLLLLS	362	40460	5E5A7BDE031403D6	1991-02-01Z	9606	Eukaryota > Metazoa >	HLA class I histocompatibility		AJ401222; CAC10402	NP_005505.2
1B67_HUMAN	7517584; 21269460	Complete proteome; Disulfide		evidence at protein level		MLVMAPRTVLLLLS	362	40342	3F6A17FC10230F70	1996-11-01Z	9606	Eukaryota > Metazoa >	HLA class I histocompatibility		AH011166; AAL18235.1;	Genomic
1FEH_AEGSP		Disulfide bond; Glycoprotein; G		inferred from homology		MAQAWAFLLPLLVI	595	66492	32B3378BFFE12FD2	2008-11-25Z	4573	Eukaryota > Viridiplant	Fructan 1-exohydrolase		FJ184993; ACI16119.1;	Genomic_
242L_IIV6	11448171; 17239238	Complete proteome; Reference		predicted	6997; 6999; 7108; 5860	MEEKNGWNIYMV	99	11255	4A18A431F3FC2986	2001-12-01Z	176652	Viruses > dsDNA viruse	Putative GIY-YIG d	IIV6-242L	AF303741; AAK8210	NP_149705.1
2A5D_HUMAN	8566219; 8703017; 918026	Alternative splicing; Complete		evidence at protein level		MPYKLKKEKEPPKV/	602	69992	F15F71AF4E565387	1996-11-01Z	9606	Eukaryota > Metazoa >	Serine/threonine-protein pho		BC001175; AAH0117	NP_851308.1
2AAB_PIG		Complete proteome; Reference		evidence at transcript level		NSAGAAAPGTGPV	602	66214	D71EAD46C2B7BB						34931; CAA84403.1;	mRNA
2ABD_DANRE		Cell cycle; Cell division; Comple		evidence at transcript level		MAGVGGGNDFQV	447	51716	9FC2E4A0EFE0E6E						C066723; AAH6672	NP_998045.1
2B_PEBV	2388830; 2491690; 122022	Virion		inferred from homology	3885; 3888	MTNWATLWPNDF	255	29559	2655D9EC58A8754						51828; CAA36126.1	NP_040352.1
2SS1_BRARC	15062997	Antibiotic; Antimicrobial; Direct		evidence at protein level		PAQPFRFPKHPQGF	20	2297	F376FB58E8F05EA							
3001L_ASFP4		Complete proteome; Host mem		inferred from homology	6937; 9823; 41426; 855	MVSLTTCCLKNIVN	268	31373	1D3A0FC97BF2072						Y261363; NOT_ANNOTATED_CD	
348R_IIV6	11448171; 17239238	Complete proteome; Reference		predicted	6997; 6999; 7108; 5860	MTHKYQILYDTLIS	98	11803	12FE387C7CDB365						F303741; AAK8220	NP_149811.1
36015_ASFP4		Complete proteome		inferred from homology	6937; 9823; 41426; 855	MVLVEFLTGFFVLY	276	31684	8B891898E4BA2D0						Y261363; NOT_ANNOTATED_CD	
3603L_ASFP4		ANK repeat; Complete proteom		inferred from homology	6937; 9823; 41426; 855	MQPSTLQTLTKKAL	356	41802	79C515A55B10D05						Y261363; NOT_ANNOTATED_CD	
385L_IIV6	11448171; 17239238	Complete proteome; Reference		predicted	6997; 6999; 7108; 5860	MNISKTKSVSLVYEI	254	28243	382F63B58E073E9						F303741; AAK8224	NP_149848.1
3BHS_HORSE	11133676	Complete proteome; Endoplasr		evidence at transcript level		MAGWSCLVTGAG	373	42411	837A04E3C681417						89666; BAA32698.1	NP_001075
3DHQ1_ASPFU	14998527; 16372009	Complete proteome; Lyase; Qui		inferred from homology		MGKSILLINGPNLN	150	16042	9C0A9DC3A17A5D8B	2004					AAHF01000004; EAL	XP_752527.1
3DHQ_SCLS1	21876677	Complete proteome; Lyase; Qui		inferred from homology		MPIKKILLLNGPNLN	154	17142	D8CC05B04DF20989			665079	Eukaryota > Fungi > Dil	Catabolic 3-dehyd	SS1G_14177	CH476650; EDO0030	XP_001584
3HAO_CHAGB		Complete proteome; Cytoplasm		inferred from homology		MIGPPVNLPKWLEI	186	21002	6A9F7228501838E1			306901	Eukaryota > Fungi > Dil	3-hydroxyanthrani	CHGG_0184	CH408029; EAQ9360	XP_00122106
3HAO_XANOM		Complete proteome; Dioxygena		inferred from homology		MLIPPINLHAWIEE	176	20196	4274767A24F2B08	07-11+01:00		342109	Bacteria > Proteobacte	3-hydroxyanthranilate 3,4-dio		AP008229; BAE6905	YP_451331.1
3MGA_BACSU	8376346; 9384377	Complete proteome; DNA dama		evidence at protein level		MTWHEVNDVIVIT	303	35634	17BEEFC78457A	1994-10-01+01:00		224308	Bacteria > Firmicutes >	DNA-3-methyladenine glycosy		AL009126; CAB1195(	NP_388061.1
3MGH_BORAP	16914037; 22123755	Complete proteome; DNA dama		inferred from homology		MDRYFFLQDASTV	186	21287	53D0E55B8	2006-09-05+01:00		390236	Bacteria > Spirochaete	Putative 3-methyladenine DN/		CP002933; AEL69648	YP_709868.1
3MGH_CLOBK	18060065	Complete proteome; DNA dama		inferred from homology		MRLTRDFYAKDAR	203	23581	24D1E	2008-04-29+01:00		498213	Bacteria > Firmicutes >	Putative 3-methyladenine DN/		CP000939; ACA4385(	YP_001780
3MGH_FRATW	17895988	Complete proteome; DNA dama		inferred from homology		MNNLEAILRLKTID	193	21894	B669A55BB9A301B7	2007-05-01+01:00		418136	Bacteria > Proteobacte	Putative 3-methyladenine DN/		CP000608; ABO4687(	YP_001121
3MGH_MYCLB	19881526	Complete proteome; DNA dama		inferred from homology		MRSPRRCKICAVSA	214	22313	6F97A6F363468528	2009-03-03Z		561304	Bacteria > Actinobacte	Putative 3-methyladenine DN/		FM211192; CAR7144	YP_00250359
3MGH_RHIME	11474104; 11481431	Complete proteome; DNA dama		inferred from homology		MHRLTSDVDFFAR	185	19974	7DE6283605FA4A11	2001-12-01Z		266834	Bacteria > Proteobacte	Putative 3-methyla	SMb20709	AL591985; CAC49816	NP_437956.1
3MGH_STAAN	11418146	Complete proteome; DNA dama		inferred from homology		MDFVNNDTRQIAK	202	22771	231650C02B31F4E1	2004-10-11+01:00		158879	Bacteria > Firmicutes >	Putative 3-methyladenine DN/		BA000018; BAB4343	NP_375457.1
3MG_HUMAN	1645538; 1874728; 192437	3D-structure; Alternative splicir		evidence at protein level		MVTPALQMKKPKQ	298	32869	BEA8C4CB250D572B	2008-09-23+01:00		9606	Eukaryota > Metazoa >	DNA-3-methyladenine glycosy		M99626; AAB46421.1;	NP_002425.2
441R_IIV6	11448171; 17239238	Complete proteome; Reference		predicted	6997; 6999; 7108; 5860	MDKLISKKFSEFQY	116	13818	34F3219AE0377BBB	2001-12-01Z		176652	Viruses > dsDNA viruse	Uncharacterized p	IIV6-441R	AF303741; AAK8230	NP_149904.1
4CL4_ORYSJ	12869764; 16100779; 1808	ATP-binding; Alternative splicir		evidence at transcript level		MGSMAAAAEAAQ	559	60029	DAC957A75C3E45BA	2004-10-11+01:00		39947	Eukaryota > Viridiplant	Probable 4-couma	P0460H04.3	AK068985; NOT_ANI	NP_0010582
4EBP3_HUMAN	9593750; 15489334	Complete proteome; Phosphop		evidence at protein level		MSTSTSCPIPGGRC	100	10873	FD29A7F3829E546E	1998-08-01+01:00		9606	Eukaryota > Metazoa >	Eukaryotic translation initiati		BC073751; AAH7375	NP_003723.1
50510_ASFWA		Complete proteome		inferred from homology	6937; 9823; 41426; 855	MFSLQELCRKNIYIL	543	63286	38BCE270F46E0E49	2009-05-05+01:00	561444	Viruses > dsDNA viruse	Protein MGF 505-10R		AY261366; NOT_ANNOTATED_CD	
5056R_ASFP4		ANK repeat; Complete proteom		inferred from homology	6937; 9823; 41426; 855	MFSLQDLCRKNIFF	518	60773	BDD47FA734FDAB29	2009-05-05+01:00	561443	Viruses > dsDNA viruse	Protein MGF 505-6R		AY261363; NOT_ANNOTATED_CD	
5E5_RAT	8537300	Complete proteome; DNA-bindi		evidence at transcript level		MFLEVADLKDGLW	825	86831	AF667FE2FD555BDF	1996-11-01Z		10116	Eukaryota > Metazoa >	5E5 antigen		D37934; BAA07153.1;	mRNA
5HT1B_RABIT	8543023; 8878052	Cell membrane; Complete prote		evidence at transcript level		MEEPGAQCAPPLA/	390	43496	C22EBC077C6C897D	1996-02-01Z		9986	Eukaryota > Metazoa >	5-hydroxytryptamine receptor	U6		NP_0010762
5HT2A_PIG	7656980; 7794950	Cell membrane; Complete prote		evidence at transcript level		MDVLCEENTSLSSP	470	52676	8231AD580BEE2A8F	1996-10-01+01:00		9823	Eukaryota > Metazoa >	5-hydroxytryptamine receptor	Z48		NP_999382.1
5HT5A_HUMAN	7988681; 15489334; 16959	Cell membrane; Complete prote		evidence at protein level		MDLPVNLTSFSLST	357	40255	92F8A78C69169790	1996-02-01Z		9606	Eukaryota > Metazoa >	5-hydroxytryptamine receptor	BC112021; AAI12022	NP_076917.1
5NT3L_MOUSE	15489334; 16141072; 1946	Acetylation; Alternative splicing		evidence at transcript level		MKATVLMRQPGR	292	33578	3AE26BCE06833085	2008-04-08+01:00		10090	Eukaryota > Metazoa >	Cytosolic 5'-nucleotidase III-li	BC015307; AAH1530	NP_080837.1
60A_DROVI	8688461	Cytokine; Disulfide bond; Glyco		inferred from homology		MTASLVVLPSLWLI	436	49999	C744B4AE58796692	1997-11-01Z		7244	Eukaryota > Metazoa >	Protein 60A		U48595; AAC47262.1;	Genomic_D

Redundancy!

Metanome

HPI Hasso Plattner Institut

Motivation
Functional Dependencies Discovery



Metanome

HPI Hasso Plattner Institut

FD Algorithm Classes

Experimental Evaluation

Extrapolation of Results

Metanome

HPI Hasso Plattner Institut

FD Algorithm Classes

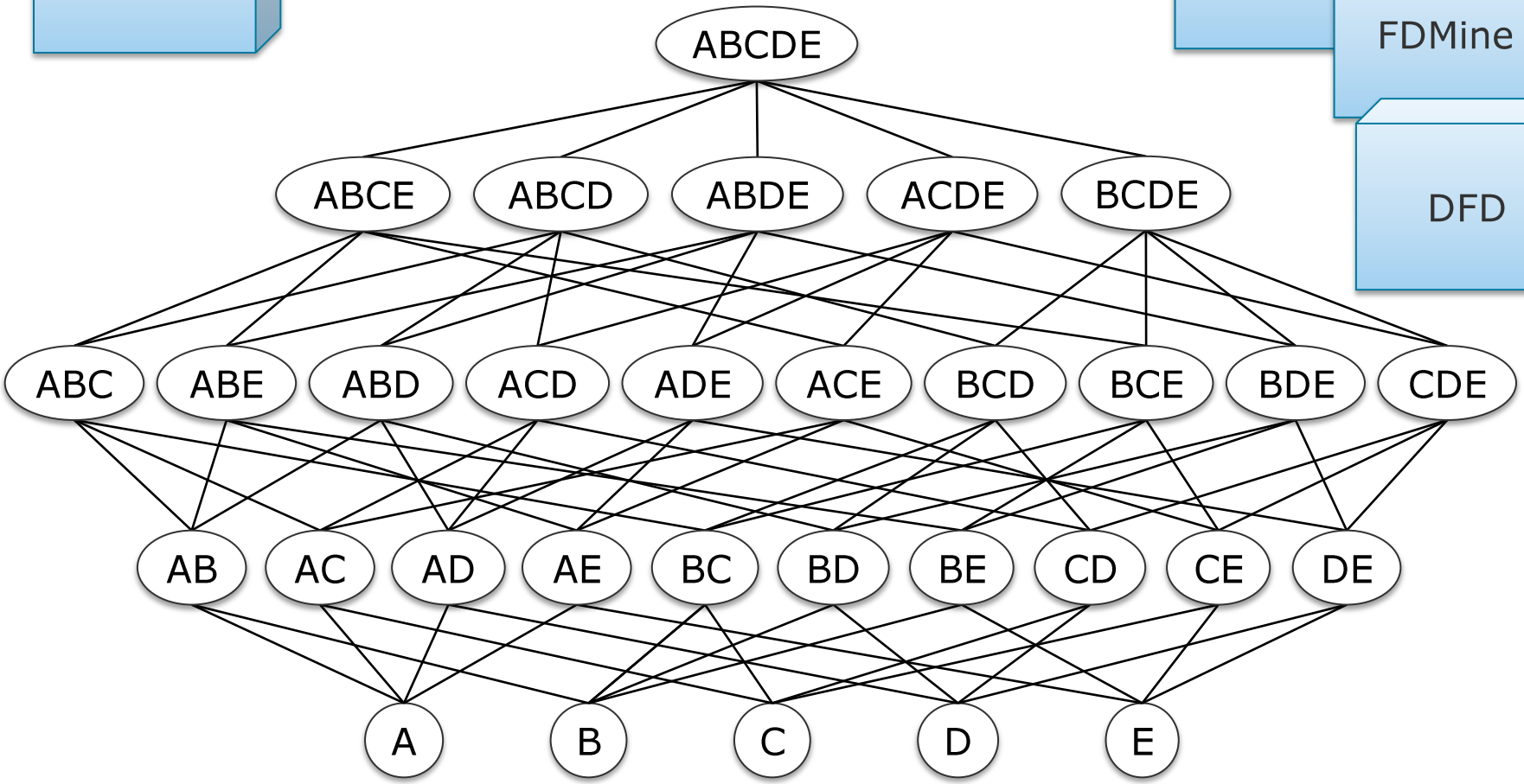
Lattice Traversal Algorithms

Tane

Fun

FDMine

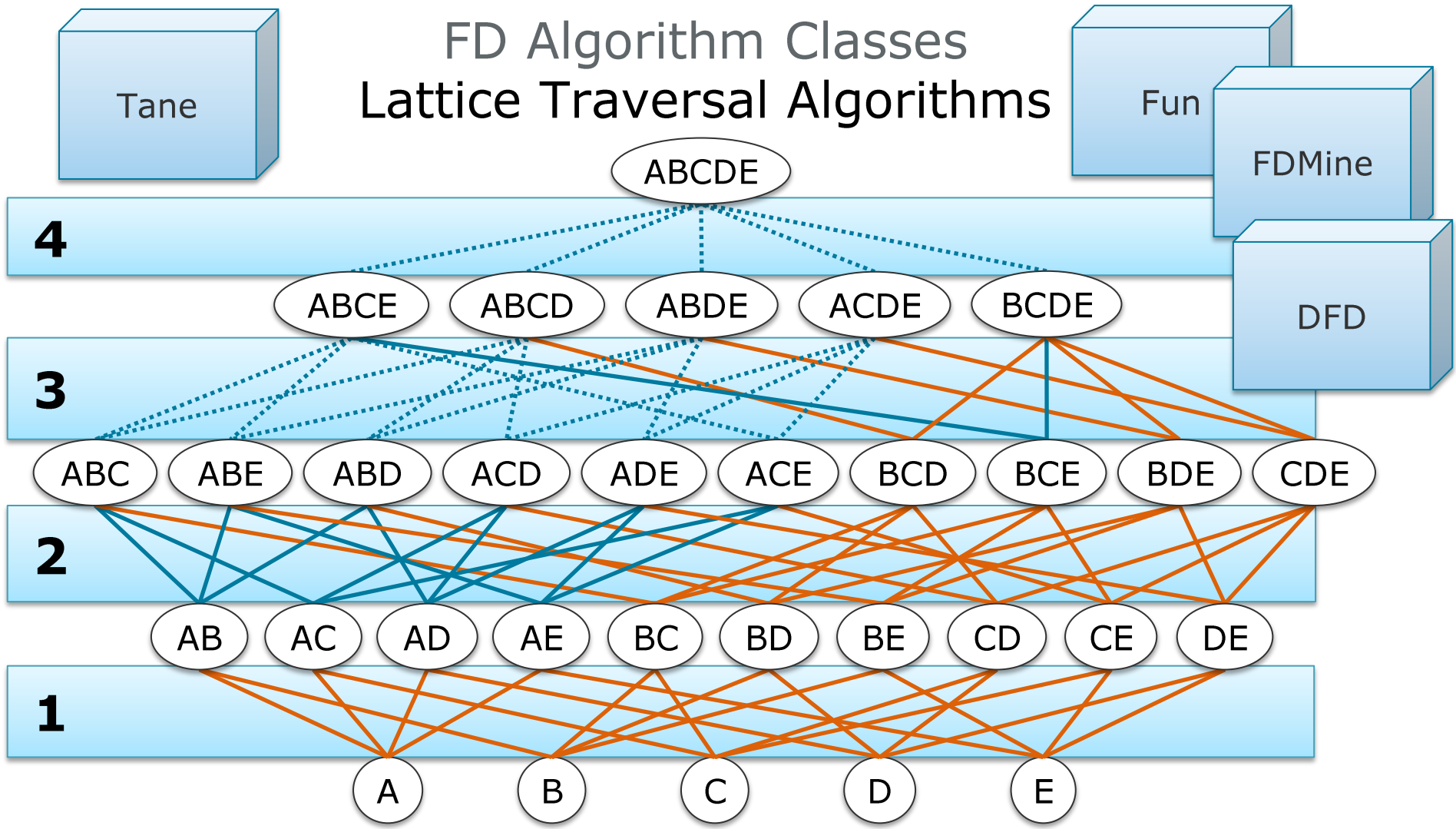
DFD



Metanome

HPI Hasso Plattner Institut

FD Algorithm Classes
Lattice Traversal Algorithms



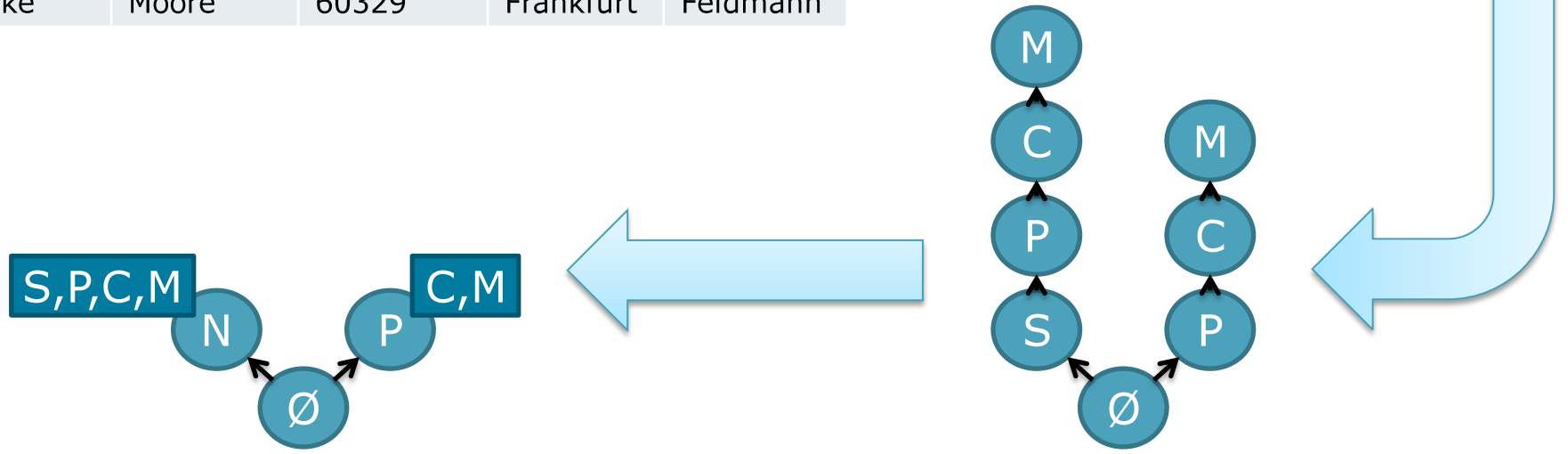
invalid FD (non-FD) minimal FD non-minimal FD

FD Algorithm Classes
Dependency Induction Algorithms

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

fdep

N,P,C,M ↛ S



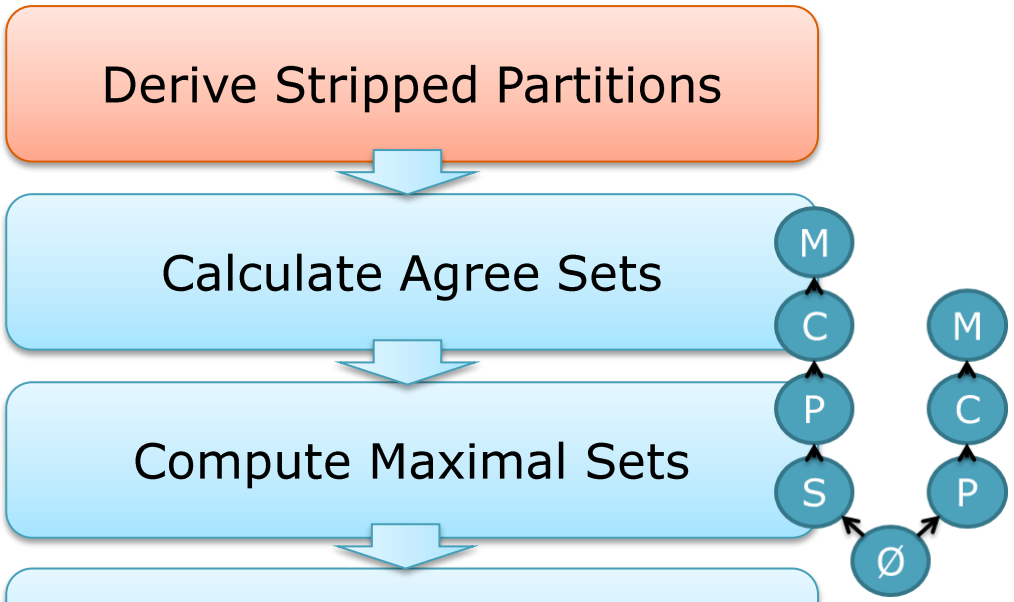
Negative Cover

Positive Cover

Metanome

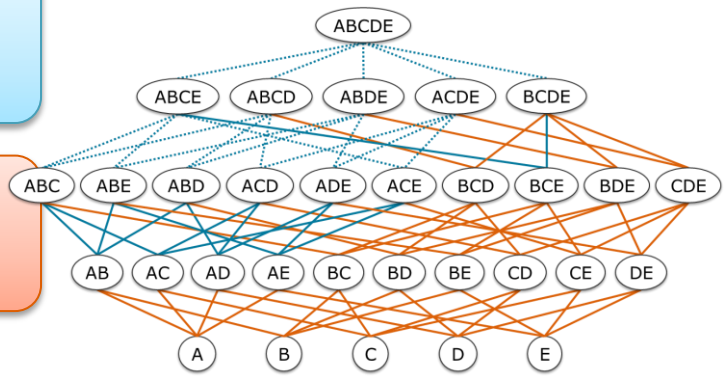
HPI Hasso Plattner Institut

FD Algorithm Classes
Difference- and Agree-Set Algorithms



DepMiner

FastFDs



Metanome

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

HPI Hasso Plattner Institut

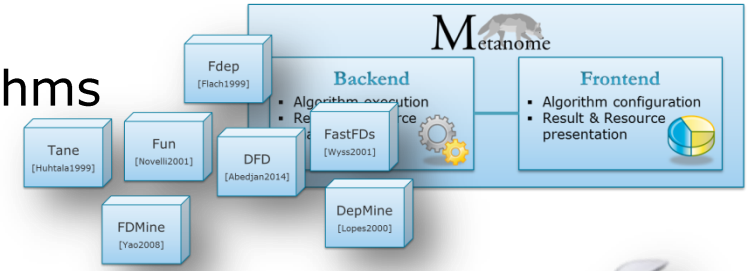
FD Algorithm Classes

Experimental Evaluation

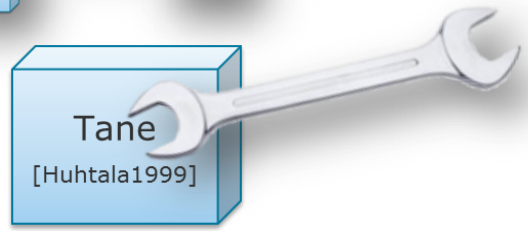
Extrapolation of Results

Experimental Evaluation
Contributions

- We **re-implemented** all seven algorithms for the Metanome profiling tool.



- We **amended** the algorithms where original descriptions were sparse.



- We **evaluated** the algorithms under comparable conditions.

- We **analyzed** and extrapolated the experimental results.

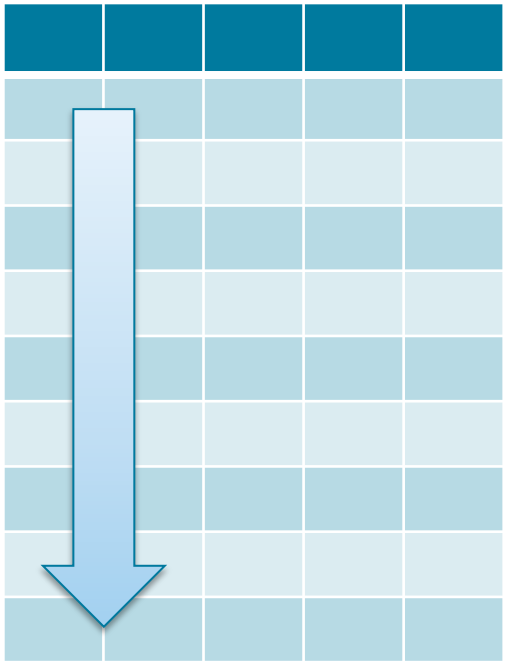
www.metanome.de

Metanome

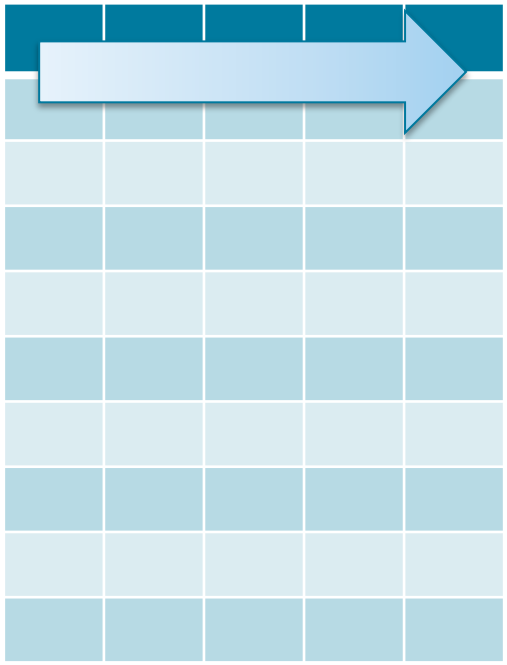
HPI Hasso Plattner Institut

Experimental Evaluation
Evaluation Strategy

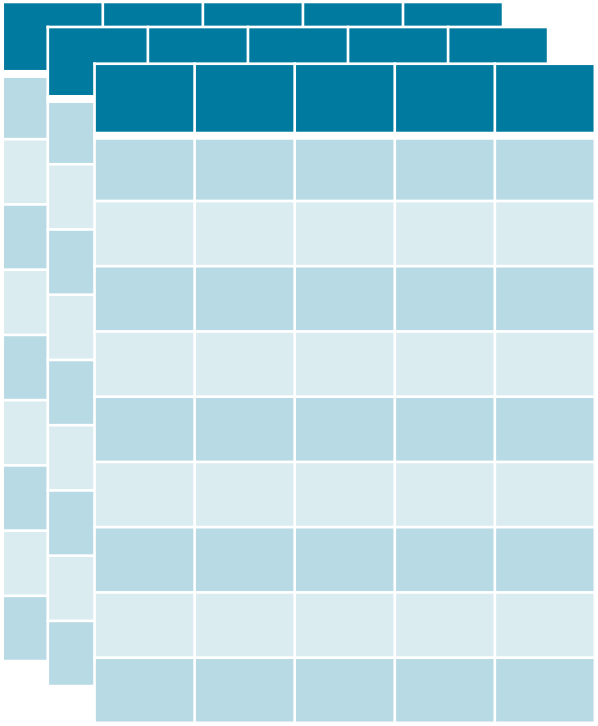
Row Scalability



Column Scalability



Different Datasets

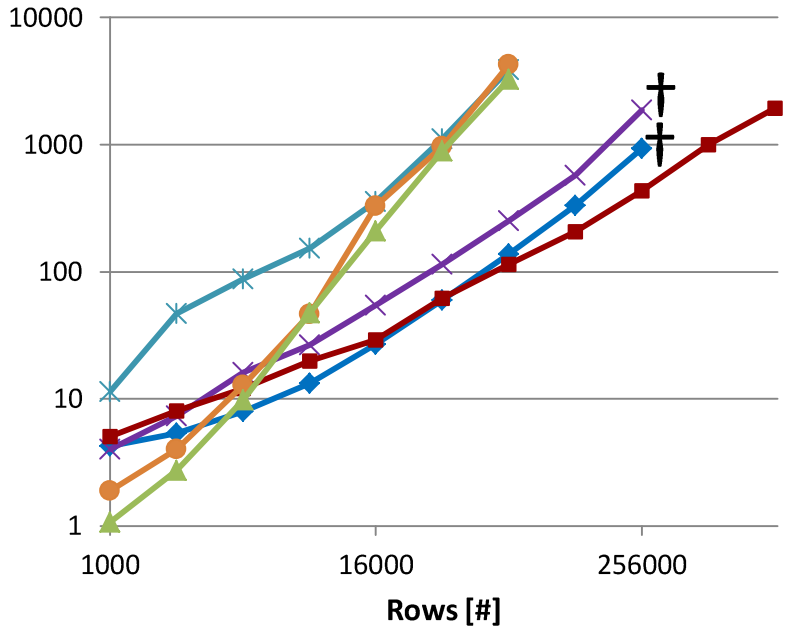


Metanome

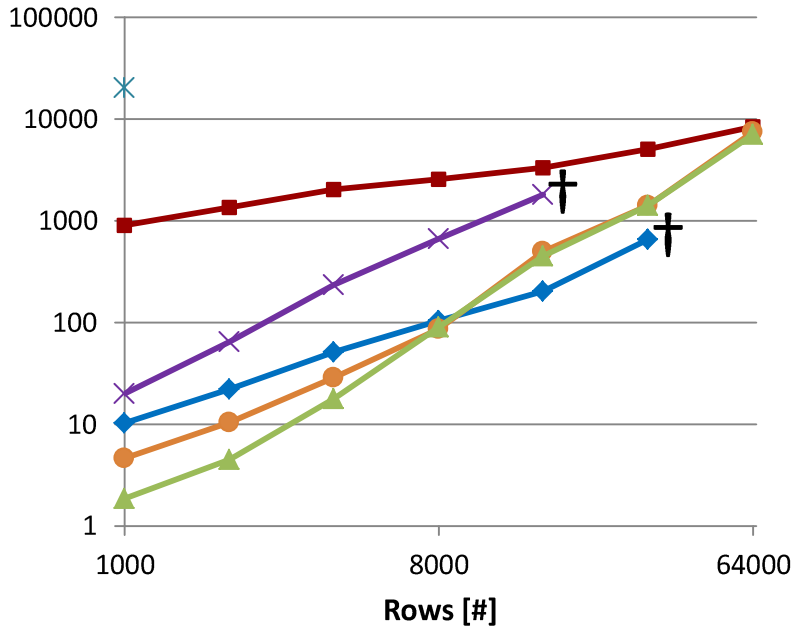
HPI Hasso Plattner Institut

Experimental Evaluation
Row-Scalability

ncvoter dataset (19 columns)



uniprot dataset (30 columns)



Tane FUN FD_Mine DFD
Dep-Miner FastFDs Fdep FDs

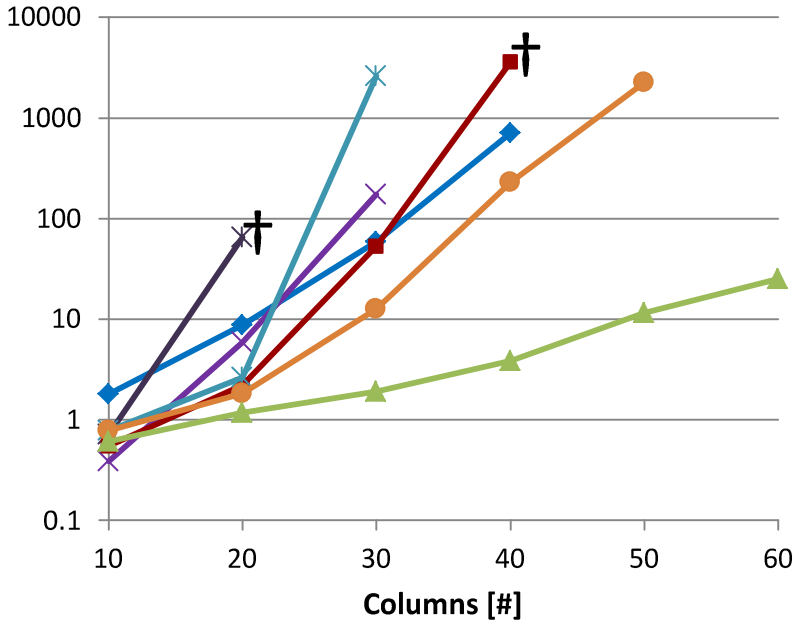
† 100 GB memory exhausted

Metanome

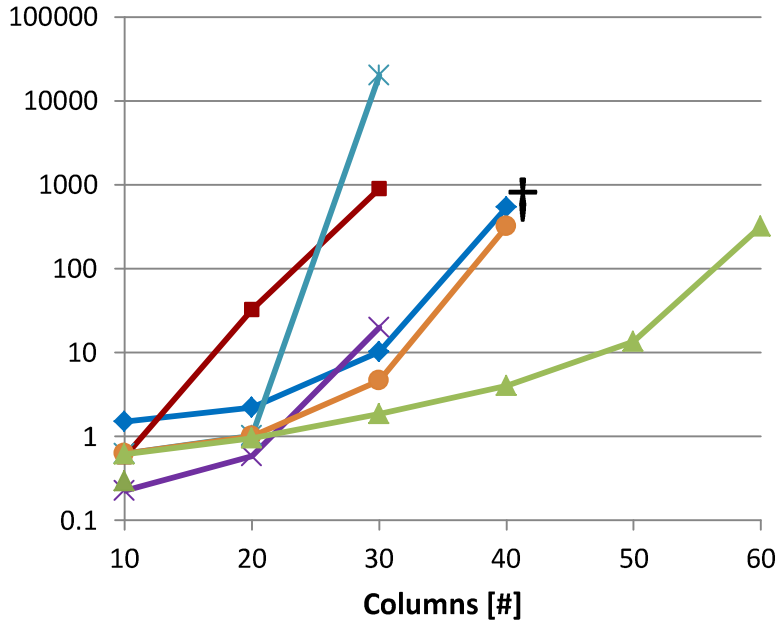
HPI Hasso Plattner Institut

Experimental Evaluation
Column-Scalability

plista dataset (1000 rows)



uniprot dataset (1000 rows)



Tane FUN FD_Mine DFD
Dep-Miner FastFDs Fdep FDs

† 100 GB memory exhausted

Metanome

HPI Hasso Plattner Institut

Experimental Evaluation
Different Datasets

Runtime [sec]

Dataset	Columns [#]	Rows [#]	Size [KB]	FDs [#]	TANE [7]	FUN [15]	FD_MINE [21]	DFD [1]	DEP-MINER [12]	FASTFDS [20]	FDEP [6]
iris	5	150	5	4	1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**
balance-scale	5	625	7	1	1.2	**0.1**	0.2	0.3	0.3	0.3	0.2
chess	7	28,056	519	1	2.9	1.1	3.8	**1.0**	174.6	164.2	125.5
abalone	9	4,177	187	137	2.1	**0.6**	1.8	1.1	3.0	2.9	3.8
nursery	9	12,960	1,024	1	4.1	1.8	7.1	**0.9**	121.2	118.9	46.8
breast-cancer	11	699	20	46	2.3	0.6	2.2	0.8	1.1	1.1	**0.5**
bridges	13	108	6	142	2.2	0.6	4.2	0.9	0.5	0.6	**0.2**
echocardiogram	13	132	6	538	1.6	0.4	69.9	1.2	0.5	0.5	**0.2**
adult	14	48,842	3,528	78	67.4	111.6	531.5	**5.9**	6039.2	6033.8	860.2
letter	17	20,000	695	61	260.0	529.0	7204.8	**6.0**	1090.0	1015.5	291.3
ncvoter	19	1,000	151	758	4.3	4.0	ML	5.1	11.4	1.9	**1.1**
hepatitis	20	155	8	8,250	12.2	175.9	ML	326.7	5576.5	9.5	**0.8**
horse	27	368	25	128,726	457.0	TL	ML	TL	TL	385.8	**7.2**
fd-reduced-30	30	250,000	69,581	89,571	**41.1**	77.7	ML	TL	377.2	382.4	TL
plista	63	1,000	568	178,152	ML	ML	ML	TL	TL	TL	**26.9**
flight	109	1,000	575	982,631	ML	ML	ML	TL	TL	TL	**216.5**
uniprot	223	1,000	2,439	unknown	ML	ML	ML	TL	TL	TL	ML

TL: time limit of 4 hours exceeded **ML**: memory limit of 100GB exceeded

1. There are **many FDs** to discover!
2. Algorithms are either **row or column efficient**!
3. No algorithm can handle datasets of **real-world size**!

Metanome

HPI Hasso Plattner Institut

FD Algorithm Classes

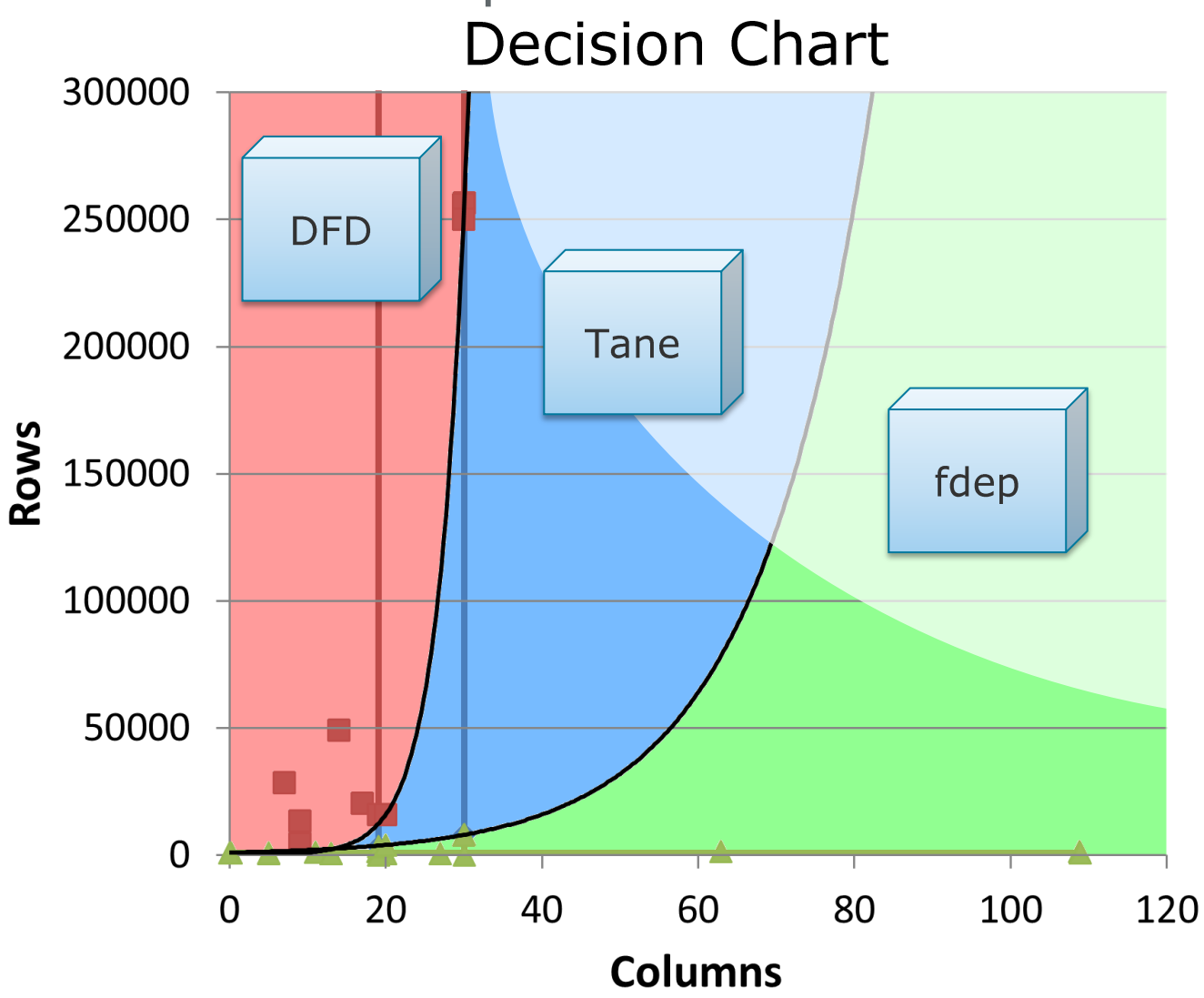
Experimental Evaluation

Extrapolation of Results

Metanome

HPI Hasso Plattner Institut

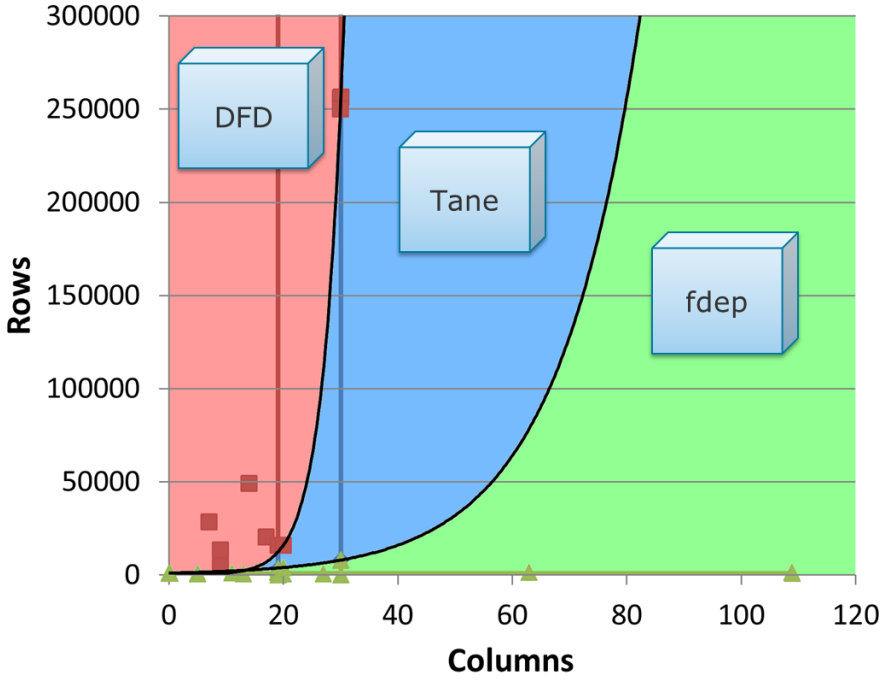
Extrapolation of Results



Metanome

HPI Hasso Plattner Institut

Tane [7]	Fun [15]	Fd_Mine [21]	Dfd [1]	Dep-Miner [12]	FastFDs [20]	Fdep [6]
1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**
1.2	**0.1**	0.2	0.3	0.3	0.3	0.2
2.9	1.1	3.8	**1.0**	174.6	164.2	125.5
2.1	**0.6**	1.8	1.1	3.0	2.9	3.8
4.1	1.8	7.1	**0.9**	121.2	118.9	46.8
2.3	0.6	2.2	0.8	1.1	1.1	**0.5**
2.2	0.6	4.2	0.9	0.5	0.6	**0.2**
1.6	0.4	69.9	1.2	0.5	0.5	**0.2**
67.4	111.6	531.5	**5.9**	6039.2	6033.8	860.2
260.0	529.0	7204.8	**6.0**	1090.0	1015.5	291.3
4.3	4.0	ML	5.1	11.4	1.9	**1.1**
12.2	175.9	ML	326.7	5576.5	9.5	**0.8**
457.0	TL	ML	TL	TL	385.8	**7.2**
41.1	77.7	ML	TL	377.2	382.4	TL
ML	ML	ML	TL	TL	TL	**26.9**
ML	ML	ML	TL	TL	TL	**216.5**
ML	ML	ML	TL	TL	TL	ML



Functional Dependency Discovery:
An Experimental Evaluation of Seven Algorithms

Thorsten Papenbrock, Jens Ehrlich, Jannik Marten, Tommy Neubert,
Jan-Peer Rudolph, Martin Schönberg, Jakob Zwiener, Felix Naumann