Distributed Computing

Data Profiling at Scale

Dr. Thorsten Papenbrock

Information Systems Group, HPI

Dr. Thorsten Papenbrock

Education

2017	PhD	HPI	Potsdam
2013	Master	HPI	Potsdam
2010	Bachelor	HPI	Potsdam
2007	Abitur	KvG	Mettingen

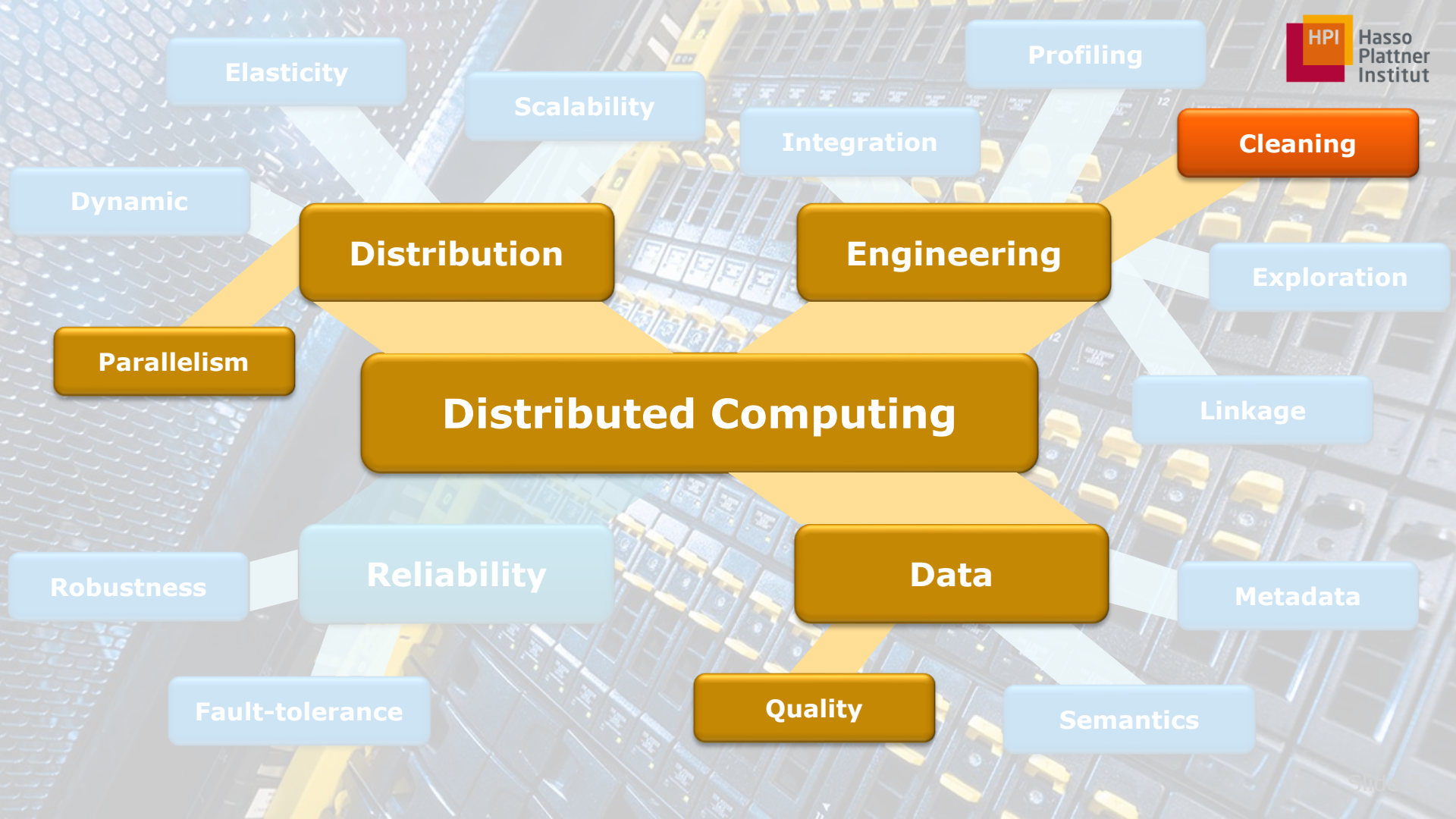
Industry

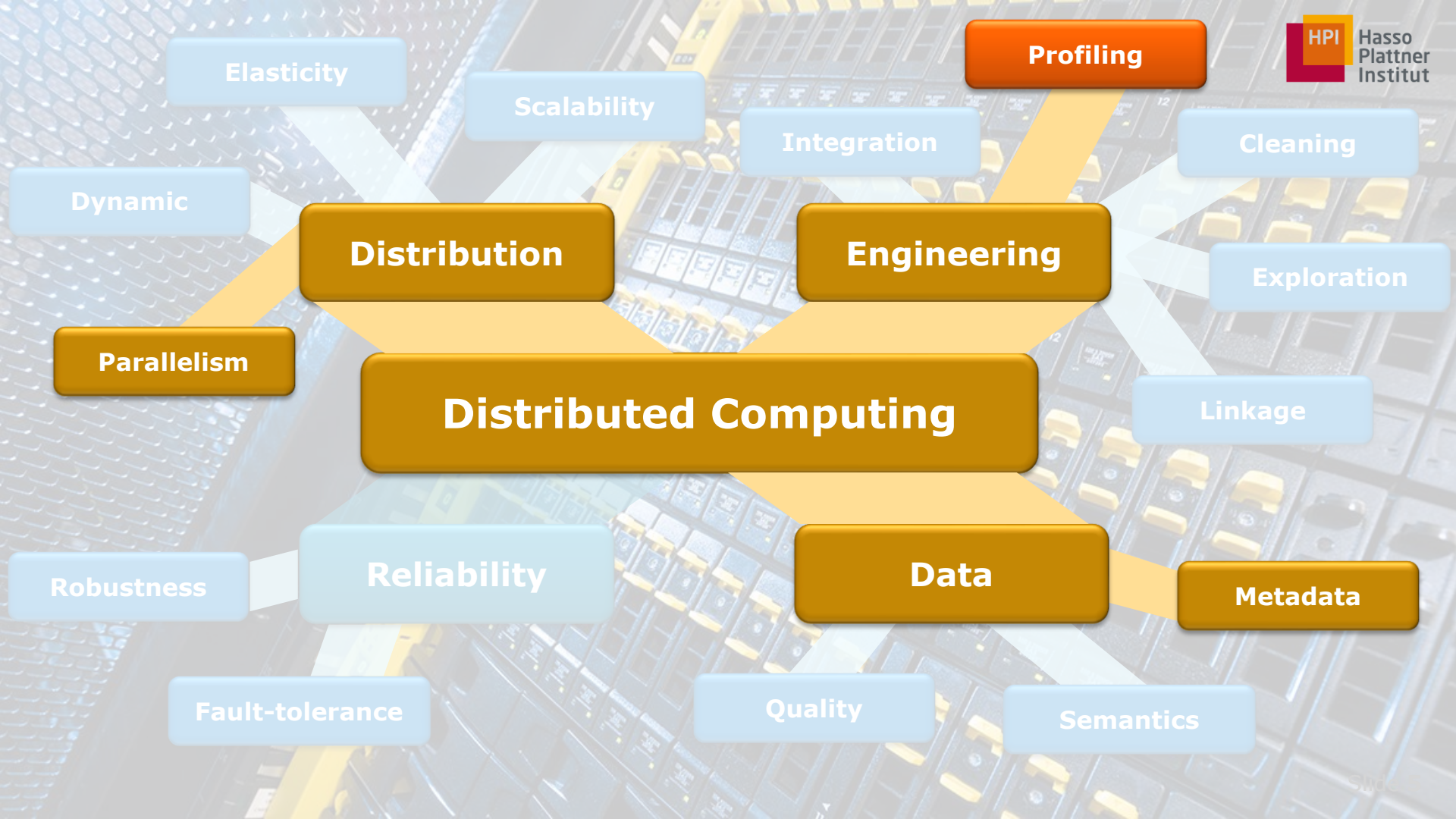
2014	Intern	QCRI	Qatar
2011/12	Intern	SAP	Belfast
2010/11	Project	BBF	Munich
2009	Intern	SAP	Walldorf

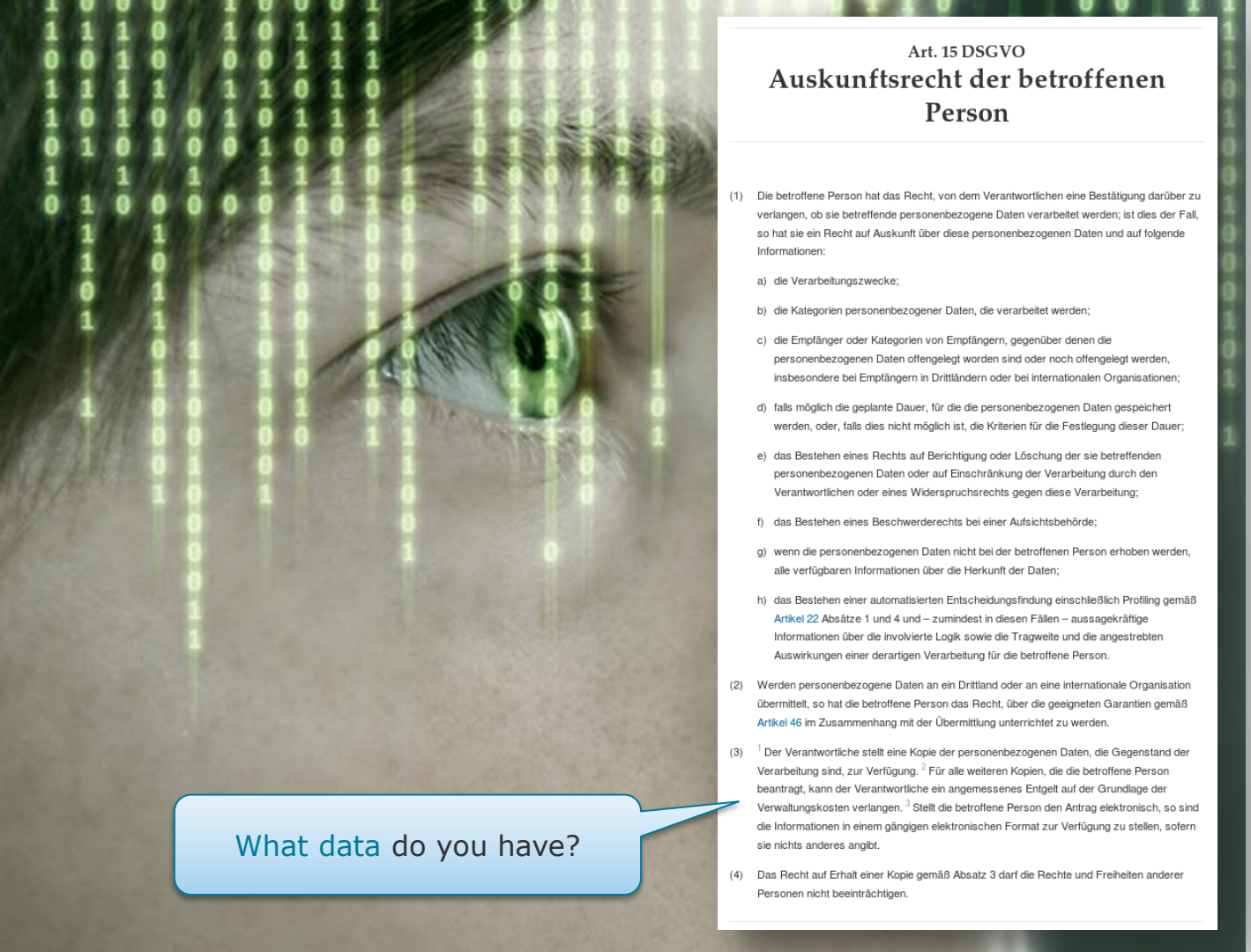
Research Interests

- Distributed Computing
- Data Profiling and Cleaning
- Information Systems and Databases


Distributed Computing
Distribution
Engineering
Reliability
Data
Elasticity
Scalability
Profiling
Integration
Cleaning
Dynamic
Exploration
Parallelism
Linkage
Robustness
Metadata
Fault-tolerance
Quality
Semantics


Elasticity
Scalability
Profiling
Integration
Cleaning
Dynamic
Distribution
Engineering
Exploration
Parallelism
Distributed Computing
Linkage
Reliability
Data
Robustness
Metadata
Fault-tolerance
Quality
Semantics


Profiling
Elasticity
Scalability
Integration
Cleaning
Dynamic
Distribution
Engineering
Exploration
Parallelism
Distributed Computing
Linkage
Reliability
Robustness
Data
Metadata
Fault-tolerance
Quality
Semantics



Art. 15 DSGVO
Auskunftsrecht der betroffenen Person

(1) Die betroffene Person hat das Recht, von dem Verantwortlichen eine Bestätigung darüber zu verlangen, ob sie betreffende personenbezogene Daten verarbeitet werden; ist dies der Fall, so hat sie ein Recht auf Auskunft über diese personenbezogenen Daten und auf folgende Informationen:

a) die Verarbeitungszwecke;

b) die Kategorien personenbezogener Daten, die verarbeitet werden;

c) die Empfänger oder Kategorien von Empfängern, gegenüber denen die personenbezogenen Daten offengelegt worden sind oder noch offengelegt werden, insbesondere bei Empfängern in Drittländern oder bei internationalen Organisationen;

d) falls möglich die geplante Dauer, für die die personenbezogenen Daten gespeichert werden, oder, falls dies nicht möglich ist, die Kriterien für die Festlegung dieser Dauer;

e) das Bestehen eines Rechts auf Berichtigung oder Löschung der sie betreffenden personenbezogenen Daten oder auf Einschränkung der Verarbeitung durch den Verantwortlichen oder eines Widerspruchsrechts gegen diese Verarbeitung;

f) das Bestehen eines Beschwerderechts bei einer Aufsichtsbehörde;

g) wenn die personenbezogenen Daten nicht bei der betroffenen Person erhoben werden, alle verfügbaren Informationen über die Herkunft der Daten;

h) das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling gemäß Artikel 22 Absätze 1 und 4 und – zumindest in diesen Fällen – aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person.

(2) Werden personenbezogene Daten an ein Drittland oder an eine internationale Organisation übermittelt, so hat die betroffene Person das Recht, über die geeigneten Garantien gemäß Artikel 46 im Zusammenhang mit der Übermittlung unterrichtet zu werden.

(3) [1] Der Verantwortliche stellt eine Kopie der personenbezogenen Daten, die Gegenstand der Verarbeitung sind, zur Verfügung. [2] Für alle weiteren Kopien, die die betroffene Person beantragt, kann der Verantwortliche ein angemessenes Entgelt auf der Grundlage der Verwaltungskosten verlangen. [3] Stellt die betroffene Person den Antrag elektronisch, so sind die Informationen in einem gängigen elektronischen Format zur Verfügung zu stellen, sofern sie nichts anderes angibt.

(4) Das Recht auf Erhalt einer Kopie gemäß Absatz 3 darf die Rechte und Freiheiten anderer Personen nicht beeinträchtigen.

What data do you have?

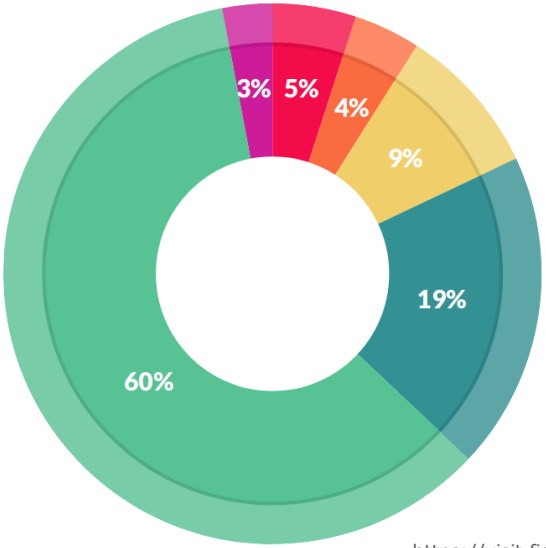
Knowledge Discovery

Knowledge Discovery

Many companies do not know what data they have!

- Decentralized storage and retrieval
- Heterogeneous data formats and systems
- Unconnected sources
- Lack of metadata and integrity constraints
- Different access rights
- Data quality issues
- Complicated business processes
- Data backups and archives
- Data acquisition and sharing
- …

CrowdFlower Data Science Report 2016



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

~80% on data preparation!

https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

Knowledge Discovery

Data Analytics

Data scientists spend most of their time on data preparation!

- Multiple, heterogeneous data sources
- Lack of metadata and documentation
- Data quality issues
- Data acquisition and sharing
- …

Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy and Li Fei-Fei, Stanford University, TPAMI, 2015

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."

AI systems learn what they see and understand

AI systems learn erroneous, non-interpretable behavior!

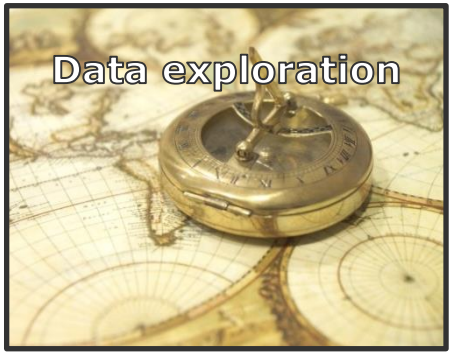
- Data quality issues
- Insufficient training data
- Heterogeneous data formats and systems
- Lack of metadata and documentation
- …

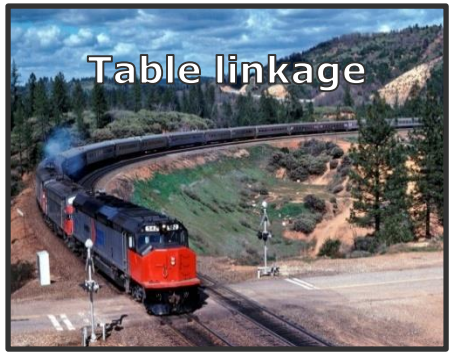
Knowledge Discovery

Data Analytics

AI Systems

Data Engineering for Data Science

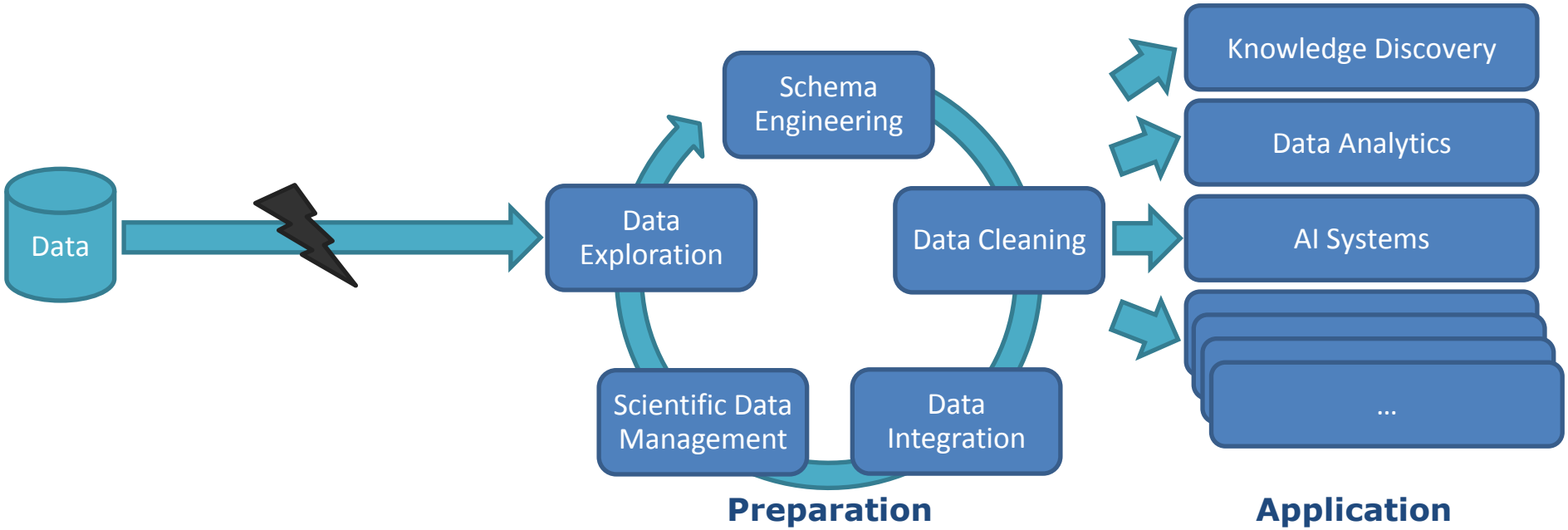




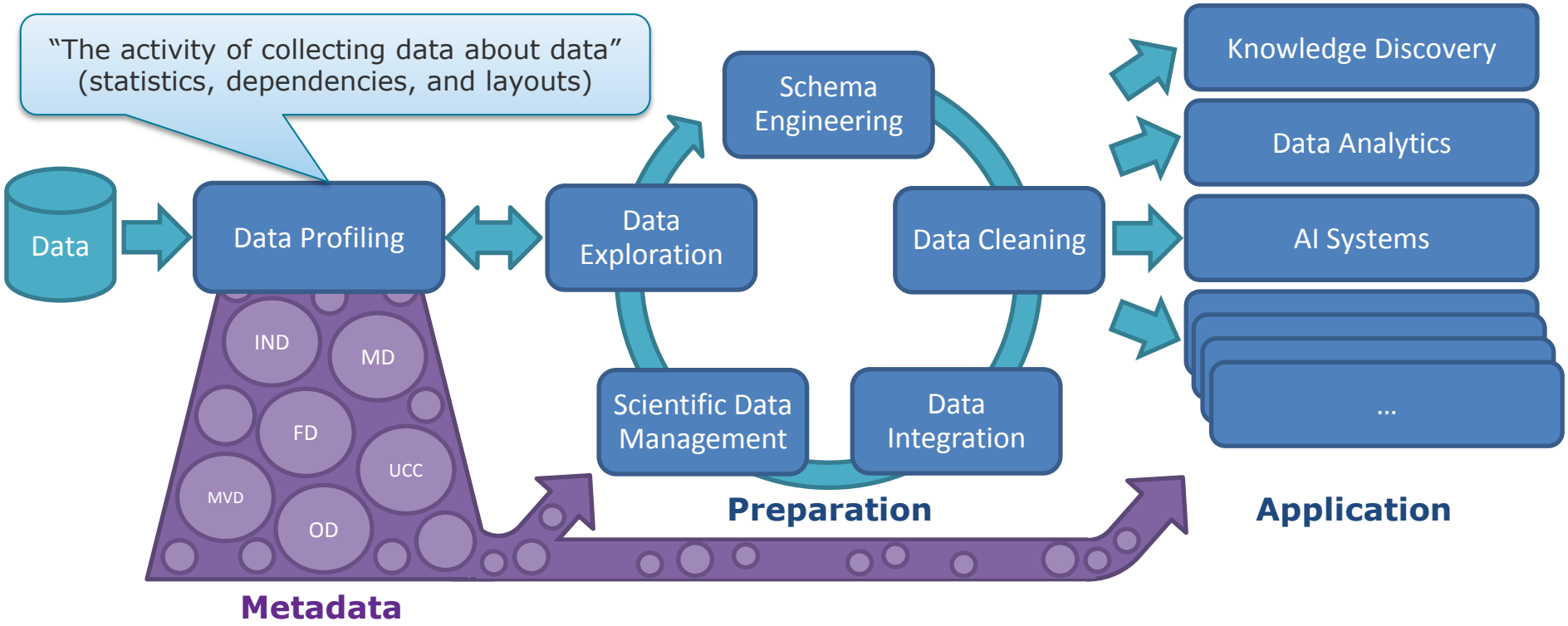
- Knowledge Discovery
- Data Analytics
- AI Systems
- ...

Application

Data Engineering for Data Science



Data Engineering for Data Science

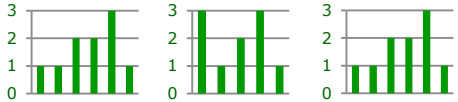


Data Profiling

ID	Name	Evolution	Location	Sex	Weight	Size	Type	Weak	Strong	Special
25	Pikachu	Raichu	Viridian Forest	m/w	6.0	0.4	electric	ground	water	false
27	Sandshrew	Sandslash	Route 4	m/w	12.0	0.6	ground	gras	electric	false
29	Nidoran	Nidorino	Safari Zone	m	9.0	0.5	poison	ground	gras	false
32	Nidoran	Nidorina	Safari Zone	w	7.0	0.4	poison	ground	gras	false
37	Vulpix	Ninetails	Route 7	m/w	9.9	0.6	fire	water	ice	false
38	Ninetails	null	null	m/w	19.9	1.1	fire	water	ice	true
63	Abra	Kadabra	Route 24	m/w	19.5	0.9	psychic	ghost	fighting	false
64	Kadabra	Alakazam	Cerulean Cave	m/w	56.5	1.3	psychic	ghost	fighting	false
130	Gyarados	null	Fuchsia City	m/w	235.0	6.5	water	electric	fire	false
150	Mewtwo	null	Cerulean Cave	null	122.0	2.0	psychic	ghost	fighting	true

http://bulbapedia.bulbagarden.net

Data Profiling

density #null = 3, %null = 30 | **ranges** min = 0.4, max = 2.0 | **aggregations** sum = 14.3, avg = 1.43 | **distributions** 

format

ID	Name	Evolution	Location	Sex	Weight	Size	Type	Weak	Strong	Special
25	Pikachu	Raichu	Viridian Forest	m/w	6.0	0.4	electric	ground	water	false
27	Sandshrew	Sandslash	Route 4	m/w	12.0	0.6	ground	gras	electric	false
29	Nidoran	Nidorino	Safari Zone	m	9.0	0.5	poison	ground	gras	false
32	Nidoran	Nidorina	Safari Zone	w	7.0	0.4	poison	ground	gras	false
37	Vulpix	Ninetails	Route 7	m/w	9.9	0.6	fire	water	ice	false
38	Ninetails	null	null	m/w	19.9	1.1	fire	water	ice	true
63	Abra	Kadabra	Route 24	m/w	19.5	0.9	psychic	ghost	fighting	false
64	Kadabra	Alakazam	Cerulean Cave	m/w	56.5	1.3	psychic	ghost	fighting	false
130	Gyarados	null	Fuchsia City	m/w	235.0	6.5	water	electric	fire	false
150	Mewtwo	null	Cerulean Cave	null	122.0	2.0	psychic	ghost	fighting	true
INTEGER	CHAR(16)	CHAR(16)	CHAR(32)	CHAR(3)	FLOAT	FLOAT	CHAR(8)	CHAR(8)	CHAR(8)	BOOLEAN

size # = 10

data types

Data Profiling

inclusion dependencies
Pokemon.Location ⊆ Location.Name

functional dependencies
Type → Weak

ID	Name	Evolution	Location	Sex	Weight	Size	Type	Weak	Strong	Special
25	Pikachu	Raichu	Viridian Forest	m/w	6.0	0.4	electric	ground	water	false
27	Sandshrew	Sandslash	Route 4	m/w	12.0	0.6	ground	gras	electric	false
29	Nidoran	Nidorino	Safari Zone	m	9.0	0.5	poison	ground	gras	false
32	Nidoran	Nidorina	Safari Zone	w	7.0	0.4	poison	ground	gras	false
37	Vulpix	Ninetails	Route 7	m/w	9.9	0.6	fire	water	ice	false
38	Ninetails	null	null	m/w	19.9	1.1	fire	water	ice	true
63	Abra	Kadabra	Route 24	m/w	19.5	0.9	psychic	ghost	fighting	false
64	Kadabra	Alakazam	Cerulean Cave	m/w	56.5	1.3	psychic	ghost	fighting	false
130	Gyarados	null	Fuchsia City	m/w	235.0	6.5	water	electric	fire	false
150	Mewtwo	null	Cerulean Cave	null	122.0	2.0	psychic	ghost	fighting	true

{Name, Sex}
unique column combinations

Weight ↓ Size
order dependencies

Weak ≠ Strong
denial constraints

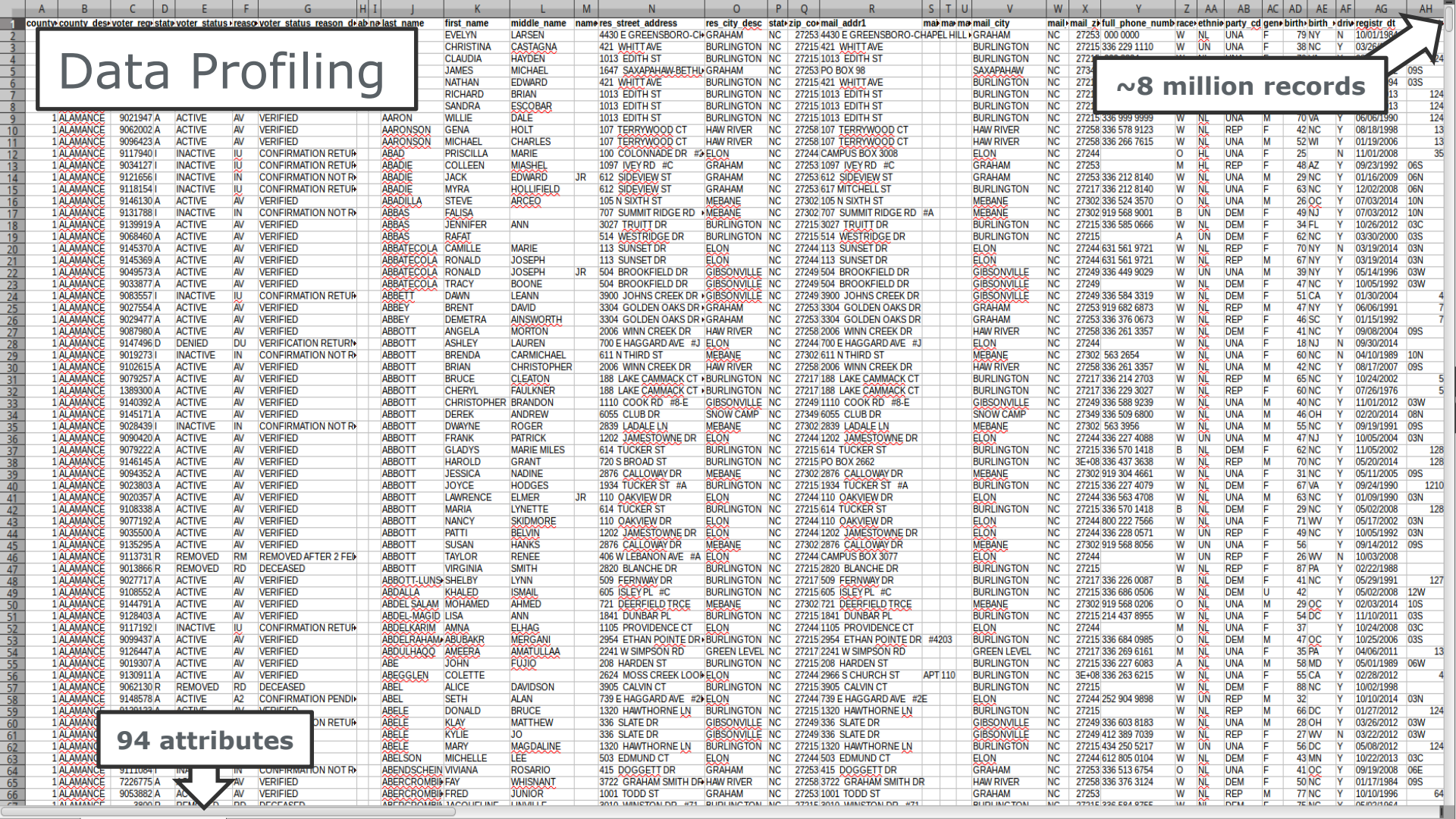
Data Profiling

ID	Name	Evolution	Location	Sex	Weight	Size	Type	Weak	Strong	Special
25	Pikachu	Raichu	Viridian Forest	m/w	6.0	0.4	electric	ground	water	false
27	Sandshrew	Sandslash	Route 4	m/w	12.0	0.6	ground	gras	electric	false
29	Nidoran	Nidorino	Safari Zone	m	9.0	0.5	poison	ground	gras	false
32	Nidoran	Nidorina	Safari Zone	w	7.0	0.4	poison	ground	gras	false
37	Vulpix	Ninetails	Route 7	m/w	9.9	0.6	fire	water	ice	false
38	Ninetails	null	null	m/w	19.9	1.1	fire	water	ice	true
63	Abra	Kadabra	Route 24	m/w	19.5	0.9	psychic	ghost	fighting	false
64	Kadabra	Alakazam	Cerulean Cave	m/w	56.5	1.3	psychic	ghost	fighting	false
130	Gyarados	null	Fuchsia City	m/w	235.0	6.5	water	electric	fire	false
150	Mewtwo	null	Cerulean Cave	null	122.0	2.0	psychic	ghost	fighting	true

Data Profiling

~8 million records

94 attributes



Agenda

Advances in Data Profiling

Challenges in Distributed Computing

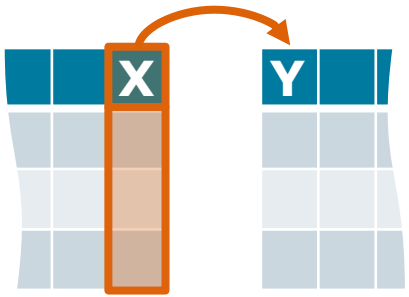
Dataset	Columns [#]	Rows [#]	Size [KB]	FDs [#]	Tane [7]	Fun [14]	Fd_Mine [21]	Dfd [1]	Dep-Miner [12]	FastFDs [20]	Fdep [6]
iris	5	150	5	4	1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**
balance-scale	5	625	7	1	1.2	**0.1**	0.2	0.3	0.3	0.3	0.2
chess	7	28,056	519	1	2.9	1.1	3.8	**1.0**	174.6	164.2	125.5
abalone	9	4,177	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

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

Table 1: Runtimes in seconds for several real-world datasets

[*Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms*, T. Papenbrock et. al., VLDB, 2015]

Inclusion Dependencies



Foreign key relationships

Definition: Given two relational instances r_i and r_j for the schemata R_i, and R_j, respectively. The inclusion dependency $R_i[X] \subseteq R_j[Y]$ (short $X \subseteq Y$) with $X \subseteq R_i$, $Y \subseteq R_j$ and $|X| = |Y|$ is valid, iff $\forall\, t_i[X] \in r_i\ ,\ \exists t_j[Y] \in r_j : t_i[X] = t_j[Y]$.

"All values in X are also contained in Y"

HPI Hasso Plattner Institut

Name	Type	Equatorial diameter	Mass	Orbital radius	Orbital period	Rotation period	Confirmed moons	Rings	Atmosphere
Mercury	Terrestrial	0.382	0.06	0.47	0.24	58.64	0	no	minimal
Venus	Terrestrial	0.949	0.82	0.72	0.62	−243.02	0	no	CO_2, N_2
Earth	Terrestrial	1.000	1.00	1.00	1.00	1.00	1	no	N_2, O_2, Ar
Mars	Terrestrial	0.532	0.11	1.52	1.88	1.03	2	no	CO_2, N_2, Ar
Jupiter	Giant	11.209	317.8	5.20	11.86	0.41	67	yes	H_2, He
Saturn	Giant	9.449	95.2	9.54	29.46	0.43	62	yes	H_2, He
Uranus	Giant	4.007	14.6	19.22	84.01	−0.72	27	yes	H_2, He
Neptune	Giant	3.883	17.2	30.06	164.8	0.67	14	yes	H_2, He

Complexity: $O(n^2-n)$
for n attributes

Example:
10 attributes ~ 90 checks
1,000 attributes ~ 999,000 checks

- Name ⊆ Type ?
- Name ⊆ Equatorial_diameter ?
- Name ⊆ Mass ?
- Name ⊆ Orbital_radius ?
- Name ⊆ Orbital_period ?
- Name ⊆ Rotation_period ?
- Name ⊆ Confirmed_moons ?
- Name ⊆ Rings ?
- Name ⊆ Atmosphere ?

- Type ⊆ Name ?
- Type ⊆ Equatorial_diameter ?
- Type ⊆ Mass ?
- Type ⊆ Orbital_radius ?
- Type ⊆ Orbital_period ?
- Type ⊆ Rotation_period ?
- Type ⊆ Confirmed_moons ?
- Type ⊆ Rings ?
- Type ⊆ Atmosphere ?

- Mass ⊆ Name ?
- Mass ⊆ Type ?
- Mass ⊆ Equatorial_diameter ?
- …

HPI Hasso Plattner Institut

Name	Type	Equatorial diameter	Mass	Orbital radius	Orbital period	Rotation period	Confirmed moons	Rings	Atmosphere
Mercury	Terrestrial	0.382	0.06	0.47	0.24	58.64	0	no	minimal
Venus	Terrestrial	0.949	0.82	0.72	0.62	−243.02	0	no	CO_2, N_2
Earth	Terrestrial	1.000	1.00	1.00	1.00	1.00	1	no	N_2, O_2, Ar
Mars	Terrestrial	0.532	0.11	1.52	1.88	1.03	2	no	CO_2, N_2, Ar
Jupiter	Giant	11.209	317.8	5.20	11.86	0.41	67	yes	H_2, He
Saturn	Giant	9.449	95.2	9.54	29.46	0.43	62	yes	H_2, He
Uranus	Giant	4.007	14.6	19.22	84.01	−0.72	27	yes	H_2, He
Neptune	Giant	3.883	17.2	30.06	164.8	0.67	14	yes	H_2, He

Planet	Rotation Period	Revolution Period
Mercury	58.6 days	87.97 days
Venus	243 days	224.7 days
Earth	0.99 days	365.26 days
Mars	1.03 days	1.88 years
Jupiter	0.41 days	11.86 years
Saturn	0.45 days	29.46 years
Uranus	0.72 days	84.01 years
Neptune	0.67 days	164.79 years
Pluto	6.39 days	248.59 years

Symbol	Unicode	Glyph
Sun	U+2609	☉
Moon	U+263D	☽
Moon	U+263E	☾
Mercury	U+263F	☿
Venus	U+2640	♀
Earth	U+1F728	🜨
Mars	U+2642	♂
Jupiter	U+2643	♃
Saturn	U+2644	♄
Uranus	U+2645	♅
Uranus	U+26E2	⛢
Neptune	U+2646	♆
Eris	≈ U+2641	♁
Eris	≈ U+29EC	⧬
Pluto	U+2647	♇
Pluto	not present	--
Aries	U+2648	♈
Taurus	U+2649	♉
Gemini	U+264A	♊
Cancer	U+264B	♋
Leo	U+264C	♌
Virgo	U+264D	♍
Libra	U+264E	♎
Scorpio	U+264F	♏
Sagittarius	U+2650	♐
Capricorn	U+2651	♑
Capricorn	U+2651	♑
Aquarius	U+2652	♒
Pisces	U+2653	♓
Conjunction	U+260C	☌
...	...	...

Planet	Synodic period	Synodic period (mean)	Days in retrograde
Mercury	116	3.8	~21
Venus	584	19.2	41
Mars	780	25.6	72
Jupiter	399	13.1	121
Saturn	378	12.4	138
Uranus	370	12.15	151
Neptune	367	12.07	158

Planet	Mean distance	Relative mean distance
Mercury	57.91	1
Venus	108.21	1.86859
Earth	149.6	1.3825
Mars	227.92	1.52353
Ceres	413.79	1.81552
Jupiter	778.57	1.88154
Saturn	1,433.53	1.84123
Uranus	2,872.46	2.00377
Neptune	4,495.06	1.56488
Pluto	5,869.66	1.3058

Sign	House	Domicile	Detriment	Exaltation	Fall	Planetary Joy
Aries	1st House	Mars	Venus	Sun	Saturn	Mercury
Taurus	2nd House	Venus	Pluto	Moon	Uranus	Jupiter
Gemini	3rd House	Mercury	Jupiter	N/A	N/A	Saturn
Cancer	4th House	Moon	Saturn	Jupiter	Mars	Venus
Leo	5th House	Sun	Uranus	Neptune	Mercury	Mars
Virgo	6th House	Mercury	Neptune	Pluto, Mercury	Venus	Saturn
Libra	7th House	Venus	Mars	Saturn	Sun	Moon
Scorpio	8th House	Pluto	Venus	Uranus	Moon	Saturn
Sagittarius	9th House	Jupiter	Mercury	N/A	N/A	Sun
Capricorn	10th House	Saturn	Moon	Mars	Jupiter	Mercury
Aquarius	11th House	Uranus	Sun	Mercury	Neptune	Venus
Pisces	12th House	Neptune	Mercury	Venus	Pluto, Mercury	Moon

Planet	Calculated (in AU)	Observed (in AU)	Perfect octaves	Actual distance
Mercury	0.4	0.387	0	0
Venus	0.7	0.723	1	1.1
Earth	1	1	2	2
Mars	1.6	1.524	4	3.7
Asteroid belt	2.8	2.767	8	7.8
Jupiter	5.2	5.203	16	15.7
Saturn	10	9.539	32	29.9
Uranus	19.6	19.191	64	61.4
Neptune	38.8	30.061	96	-96.8
Pluto	77.2	39.529	128	127.7

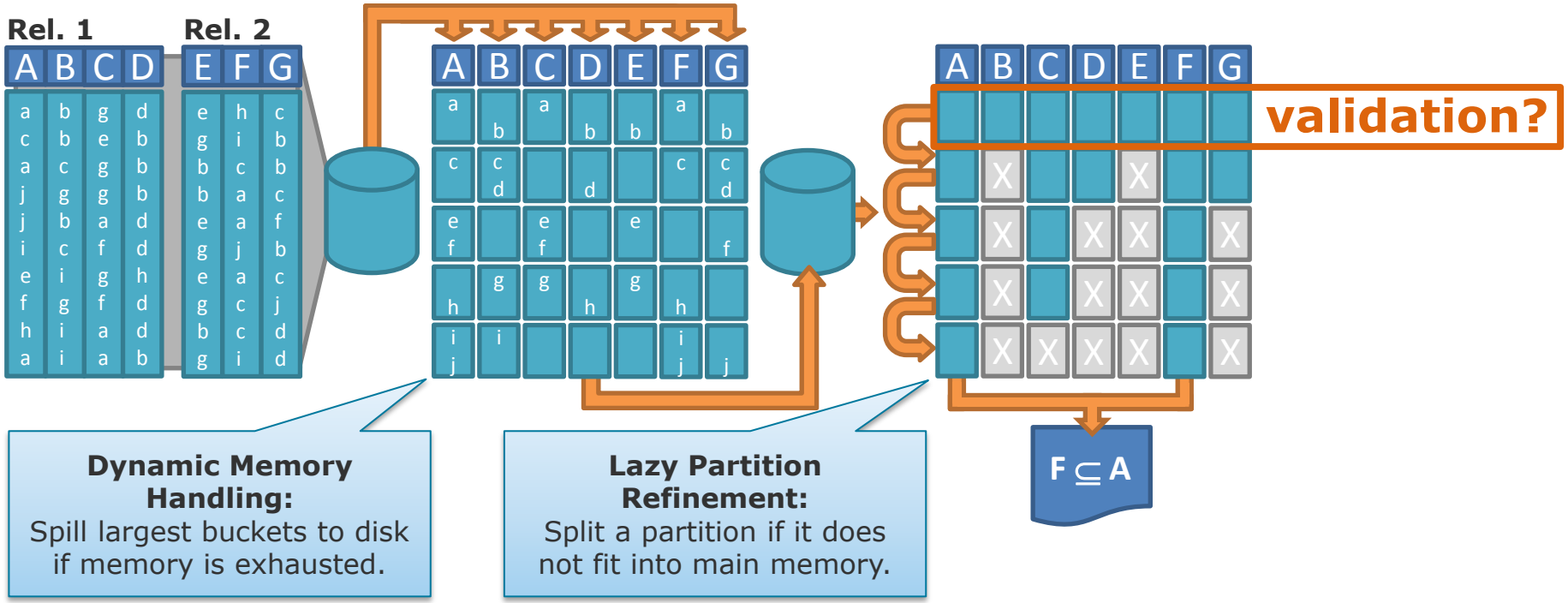
http://webdatacommons.org/webtables/index.html

HPI Hasso Plattner Institut

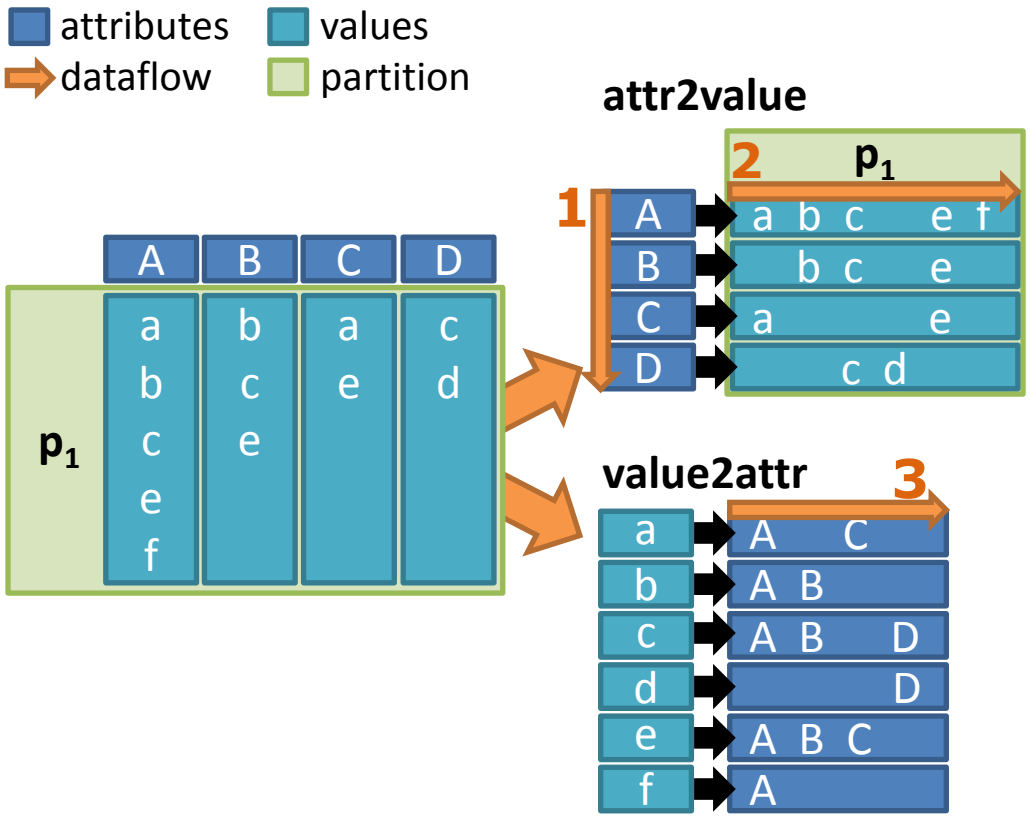
attributes values
dataflow ignored

Divide

Conquer



HPI Hasso Plattner Institut

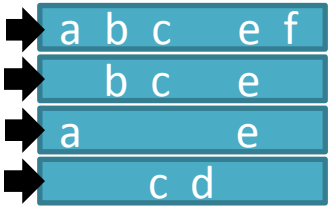


Validation algorithm

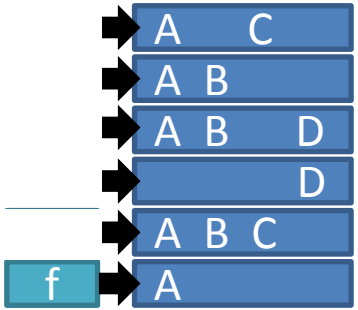
1. Iterate attributes
2. Iterate values
3. If value2attr entry exists
   - Intersect candidates with this list
   - Remove value2attr entry
   - If attribute removed from all candidates
     - Remove entry from attr2value

HPI Hasso Plattner Institut

attr2value



value2attr



Never tested! →

1. Iterate attributes
2. Iterate values
3. If value2attr entry exists
   - ➢ Intersect candidates with this list
   - ➢ Remove value2attr entry
   - ➢ If attribute removed from all candidates
     - ➢ Remove entry from attr2value

look up	A	B	C	D
	B,C,D	**A,C,D**	**A,B,D**	**A,B,C**

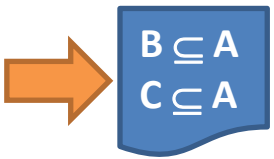


Table 6: uIND performance on real-world datasets (minutes)

Datasets	Bell & Brockhausen	DeMarchi	Spider	Spider BF	S-indd	Binder	Size	Attributes	uInds	nInds
SCOP	0.14	**0.04**	0.08	0.07	0.09	0.07	16 MB	22	39	36
CATH	0.11	**0.02**	0.05	0.05	0.05	0.04	16 MB	25	51	81
CENSUS	1.05	0.97	0.15	0.18	0.17	**0.14**	112 MB	42	39	89
WIKIPEDIA	TL	**1.02**	1.47	1.45	1.56	1.22	540 MB	14	2	0
BIOSQL	4.98	**0.76**	1.30	1.43	1.48	1.41	560 MB	77	348	507
WIKIRANK	2.90	**0.73**	1.53	1.19	1.44	1.23	697 MB	29	99	103
LOD	0.34	**0.25**	0.45	0.43	0.41	0.30	830 MB	41	258	unknown
ENSEMBL	23.52	6.87	3.04	3.50	3.70	**2.39**	836 MB	130	364	100
TESMA	TL	4.53	3.30	**2.85**	4.75	4.27	1 GB	114	2	0
TPC-H 1	17.79	**1.96**	3.88	3.10	3.58	2.96	1 GB	61	96	8
TPC-H 10	TL	ML	44.43	35.06	36.54	**28.21**	10 GB	61	97	11
MUSICBRAINZ	TL	136.03	61.42	130.15	106.26	**45.69**	27 GB	1 054	49 829	unknown

TL: time limit of 4h exceeded ML: memory limit of 83G exceeded

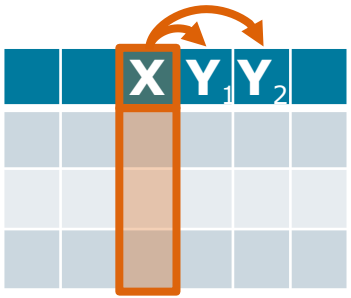
Table 7: nIND performance on real-world datasets in minutes

Datasets	Binder	Find2	Mind	Mind2	ZigZag
SCOP	**0.30**	0.37	0.36	1.85	0.50
CATH	4.18	3.24	**3.15**	29.16	3.23
CENSUS	**0.68**	2.28	2.01	N.A.	3.37
WIKIPEDIA	**1.40**	1.50	1.48	1.47	1.54
BIOSQL	**3.65**	5.51	5.36	TL	4.88
WIKIRANK	8.10	2.99	8.89	TL	**2.96**
ENSEMBL	**6.31**	8.33	7.13	TL	206.94
ghaIND	177.65	13.28	27.94	13.45	**13.13**
TESMA	4.35	3.39	**3.38**	8.79	3.39
TPC-H 1	9.97	7.49	**7.01**	TL	12.37
TPC-H 10	**121.09**	TL	TL	TL	TL

TL: time limit of 4h exceeded ML: memory limit of 83G exceeded

[*Inclusion Dependency Discovery: An Experimental Evaluation of Ten Algorithms*, F. Dürsch, A. Stebner, F. Windheuser, M. Fischer, T. Friedrich, N. Strelow, T. Bleifuß, H. Harmouch, L. Jiang, T. Papenbrock, F. Naumann, *submitted to VLDB*]

Functional Dependencies



Definition: Given a relational instance r for a schema R. The functional dependency $X \rightarrow A$ with $X \subseteq R$ and $A \in R$ is valid in r, iff $\forall t_i, t_j \in r : t_i[X] = t_j[X] \Rightarrow t_i[Y] = t_j[Y]$.

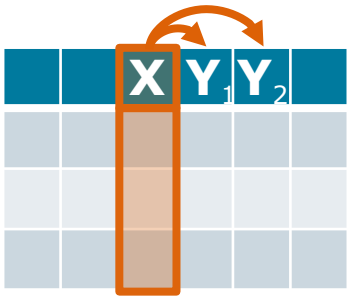
"The values in X functionally define the values in Y"

ID	Name	Size	Type	Weak	Strong	GYM	Leader	Reward
25	Pikachu	0.4	electric	ground	water	Vermillion	Lt. Surge	Thunder
26	Raichu	0.8	electric	ground	water	Vermillion	Lt. Surge	Thunder
29	Nidoran	0.5	poison	ground	gras	Viridian	Giovanni	Earth
37	Vulpix	0.6	fire	water	ice	null	null	null
38	Ninetails	1.1	fire	water	ice	null	null	null
63	Abra	0.9	psychic	ghost	fighting	null	null	null
64	Kadabra	1.3	psychic	ghost	fighting	Saffron	Sabrina	Marsh
65	Alakazam	1.5	psychic	ghost	fighting	Saffron	Sabrina	Marsh
150	Mewtwo	2.0	psychic	ghost	fighting	null	null	null

Type → Weak, Strong

GYM → Leader, Reward

Functional Dependencies



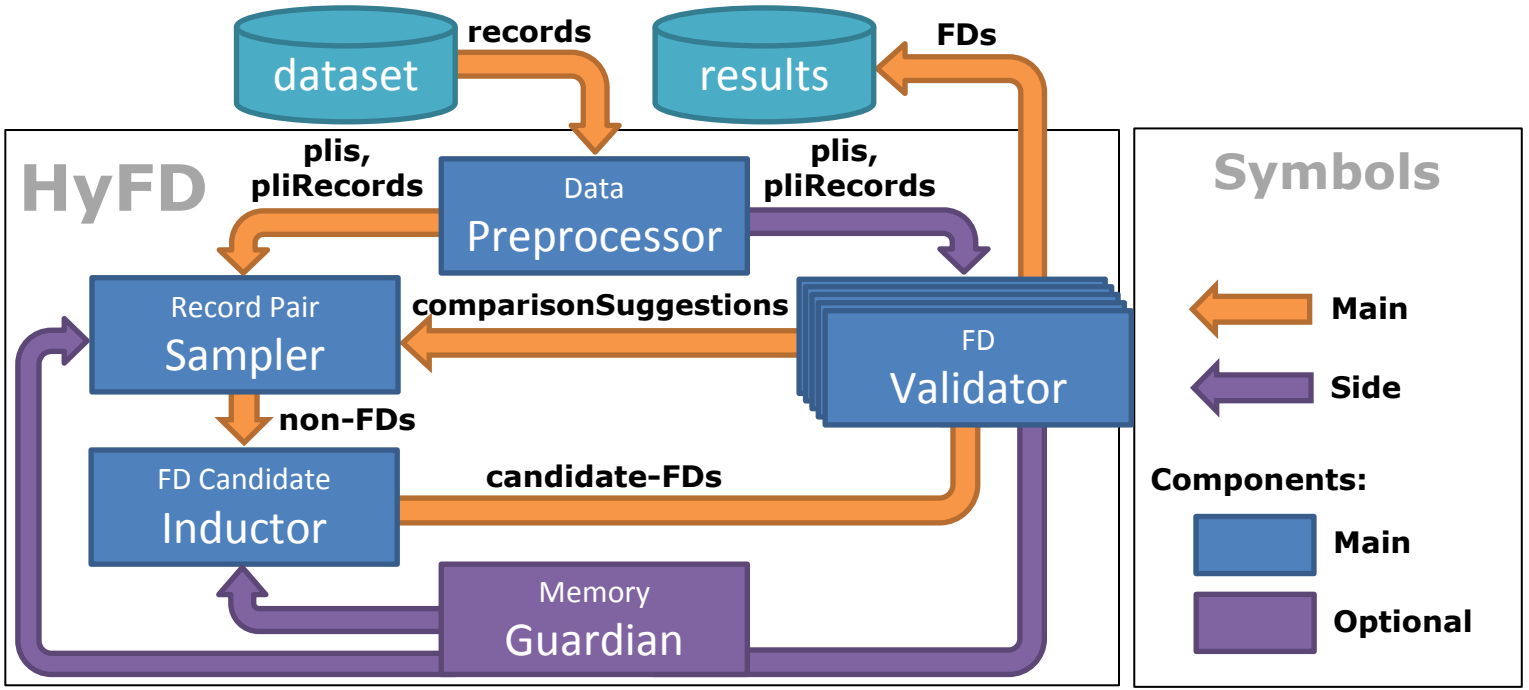
Definition: Given a relational instance r for a schema R. The functional dependency $X \rightarrow A$ with $X \subseteq R$ and $A \in R$ is valid in r, iff $\forall t_i, t_j \in r : t_i[X] = t_j[X] \Rightarrow t_i[Y] = t_j[Y]$.

"The values in X functionally define the values in Y"

ID	Name	Size	Type	GYM
25	Pikachu	0.4	electric	Vermillion
26	Raichu	0.8	electric	Vermillion
29	Nidoran	0.5	poison	Viridian
37	Vulpix	0.6	fire	null
38	Ninetails	1.1	fire	null
63	Abra	0.9	psychic	null
64	Kadabra	1.3	psychic	Saffron
65	Alakazam	1.5	psychic	Saffron
150	Mewtwo	2.0	psychic	null

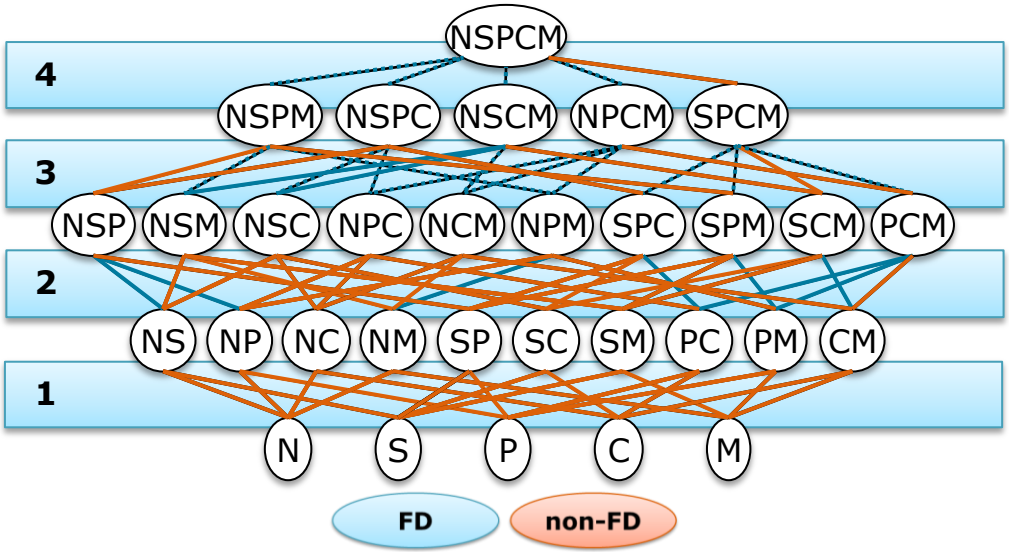
Type	Weak	Strong
electric	ground	water
poison	ground	gras
fire	water	ice
psychic	ghost	fighting

GYM	Leader	Reward
Vermillion	Lt. Surge	Thunder
Viridian	Giovanni	Earth
Saffron	Sabrina	Marsh



[*A Hybrid Approach to Functional Dependency Discovery*, T. Papenbrock, F. Naumann, SIGMOD, 2016]

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



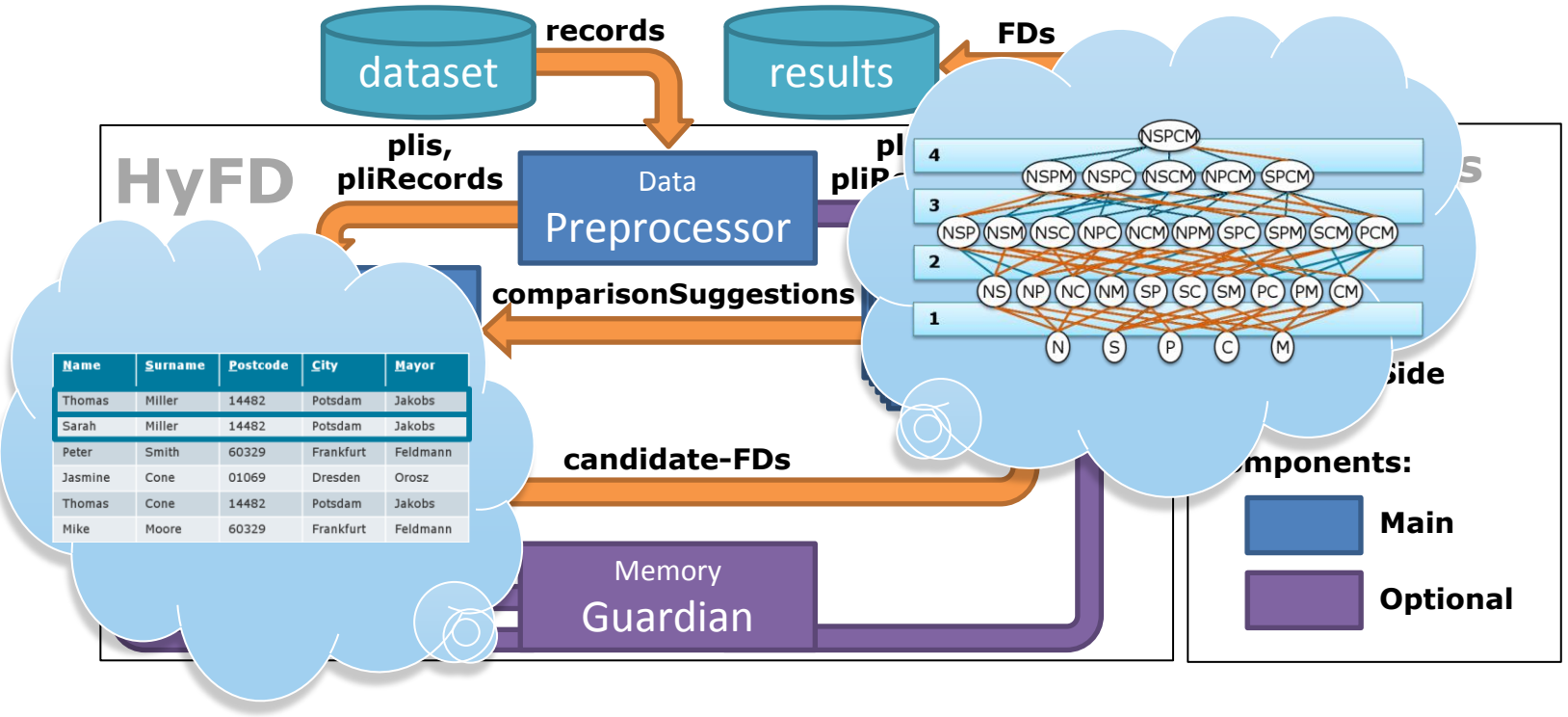
- Surname, Postcode, City, Mayor ↛ Name
- Name, Postcode, City, Mayor ↛ Surname
- Surname ↛ Name, Postcode, City, Mayor

Postcode → City

Postcode → Mayor

…

[*A Hybrid Approach to Functional Dependency Discovery*, T. Papenbrock, F. Naumann, SIGMOD, 2016]



[*A Hybrid Approach to Functional Dependency Discovery*, T. Papenbrock, F. Naumann, SIGMOD, 2016]

HPI Hasso Plattner Institut

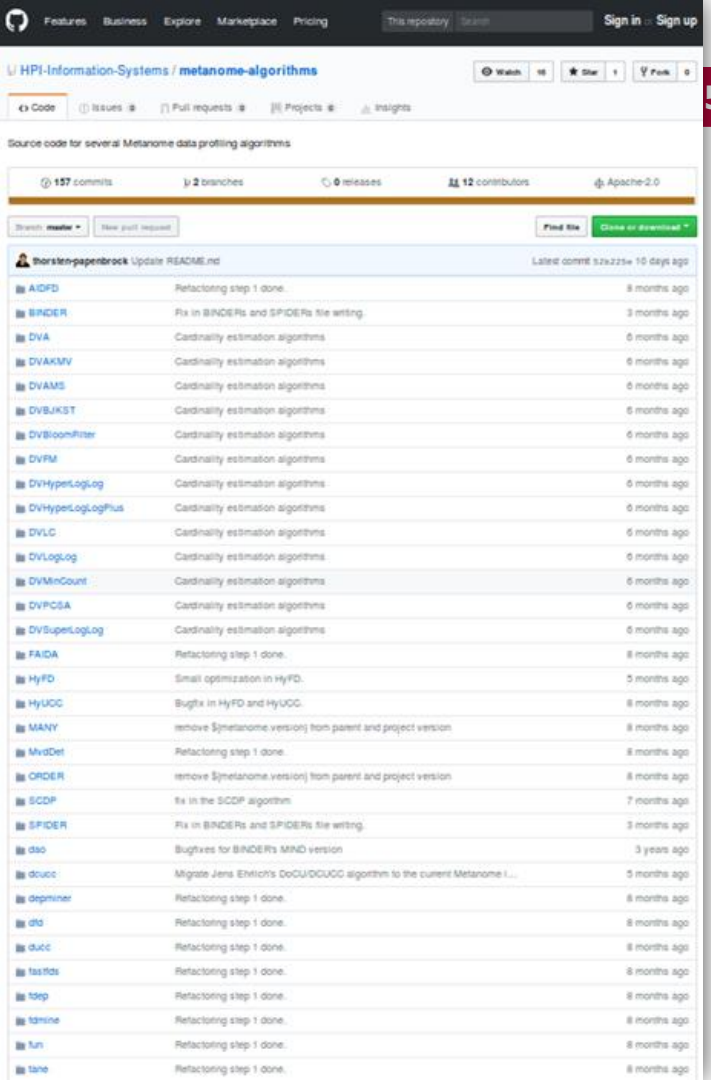
Dataset	Cols [#]	Rows [#]	Size [KB]	FDs [#]	TANE	FUN	FD_MINE	DFD	DEP-MINER	FASTFDS	FDEP	HYFD
iris	5	150	5	4	1.1	**0.1**	0.2	0.2	0.2	0.2	**0.1**	**0.1**
balance-scale	5	625	7	1	1.2	**0.1**	0.2	0.3	0.3	0.3	0.2	**0.1**
chess	7	28,056	519	1	2.9	1.1	3.8	1.0	174.6	164.2	125.5	**0.2**
abalone	9	4,177	187	137	2.1	0.6	1.8	1.1	3.0	2.9	3.8	**0.2**
nursery	9	12,960	1,024	1	4.1	1.8	7.1	0.9	121.2	118.9	46.8	**0.5**
breast-cancer	11	699	20	46	2.3	0.6	2.2	0.8	1.1	1.1	0.5	**0.2**
bridges	13	108	6	142	2.2	0.6	4.2	0.9	0.5	0.6	0.2	**0.1**
echocardiogram	13	132	6	527	1.6	0.4	69.9	1.2	0.5	0.5	0.2	**0.1**
adult	14	48,842	3,528	78	67.4	111.6	531.5	5.9	6039.2	6033.8	860.2	**1.1**
letter	17	20,000	695	61	260.0	529.0	7204.8	6.0	1090.0	1015.5	291.3	**3.4**
ncvoter	19	1,000	151	758	4.3	4.0	ML	5.1	11.4	1.9	1.1	**0.4**
hepatitis	20	155	8	8,250	12.2	175.9	ML	326.7	5576.5	9.5	0.8	**0.6**
horse	27	368	25	128,727	457.0	TL	ML	TL	TL	385.8	7.2	**7.1**
fd-reduced-30	30	250,000	69,581	89,571	**41.1**	77.7	ML	TL	377.2	382.4	TL	513.0
plista	63	1,000	568	178,152	ML	ML	ML	TL	TL	TL	26.9	**21.8**
flight	109	1,000	575	982,631	ML	ML	ML	TL	TL	TL	216.5	**53.4**
uniprot	223	1,000	2,439	>2,437,556	ML	ML	ML	TL	TL	TL	ML	**>5254.7**

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

[*A Hybrid Approach to Functional Dependency Discovery*, T. Papenbrock, F. Naumann, SIGMOD, 2016]

Dataset	Cols [#]	Rows [#]	Size [MB]	FDs [#]	HyFD [s/m/h/d]
TPC-H.lineitem	16	6 m	1,051	4 k	4 m
PDB.POLY_SEQ	13	17 m	1,256	68	3 m
PDB.ATOM_SITE	31	27 m	5,042	10 k	64 m
SAP_R3.ZBC00DT	35	3 m	783	211	2 m
SAP_R3.ILOA	48	45 m	8,731	16 k	8 h
SAP_R3.CE4HI01	65	2 m	649	2 k	10 m
NCVoter.statewide	71	1 m	561	5 m	31 h
CD.cd	107	10 k	5	36 k	3 s

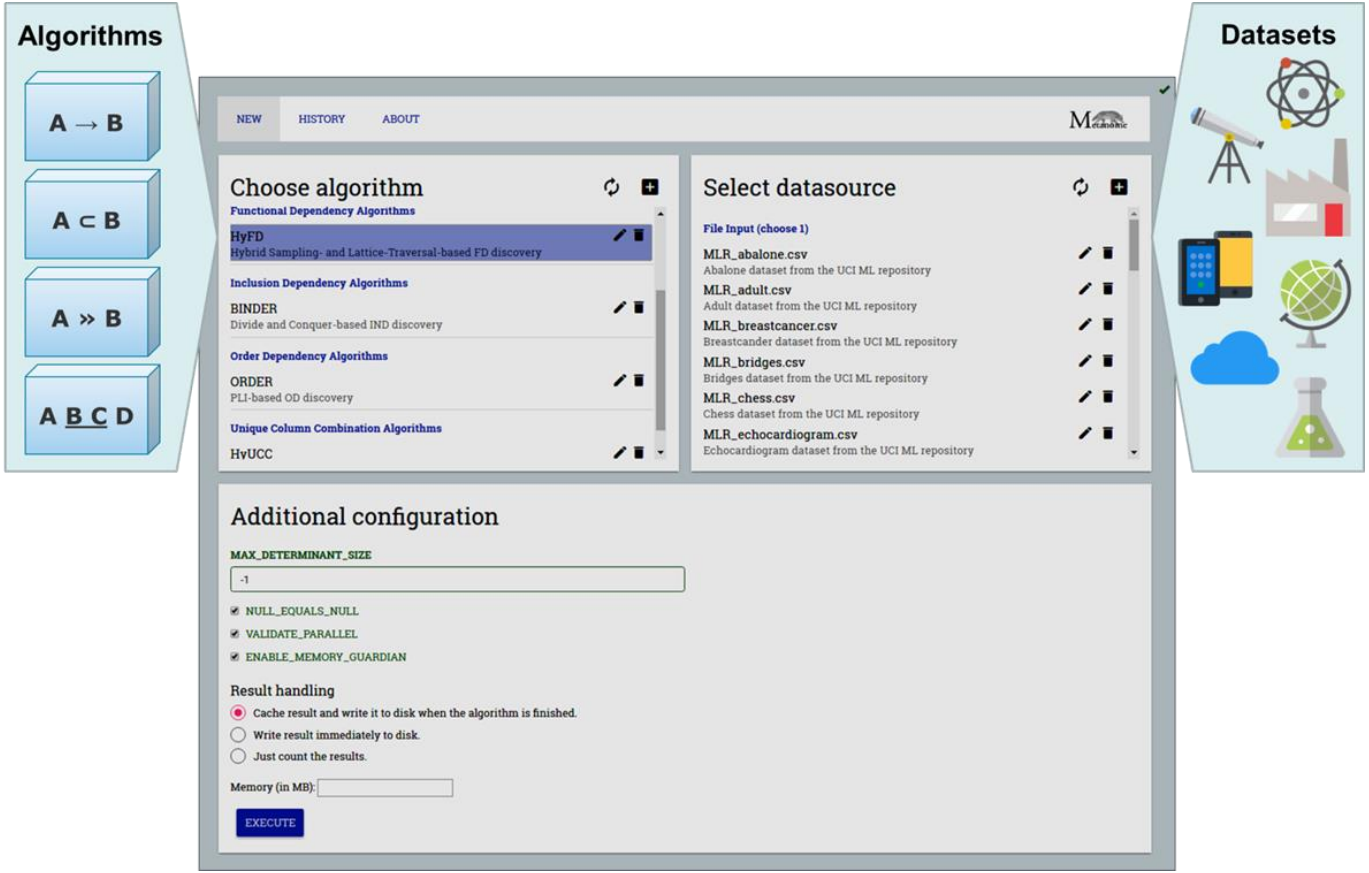
[*A Hybrid Approach to Functional Dependency Discovery*, T. Papenbrock, F. Naumann, SIGMOD, 2016]



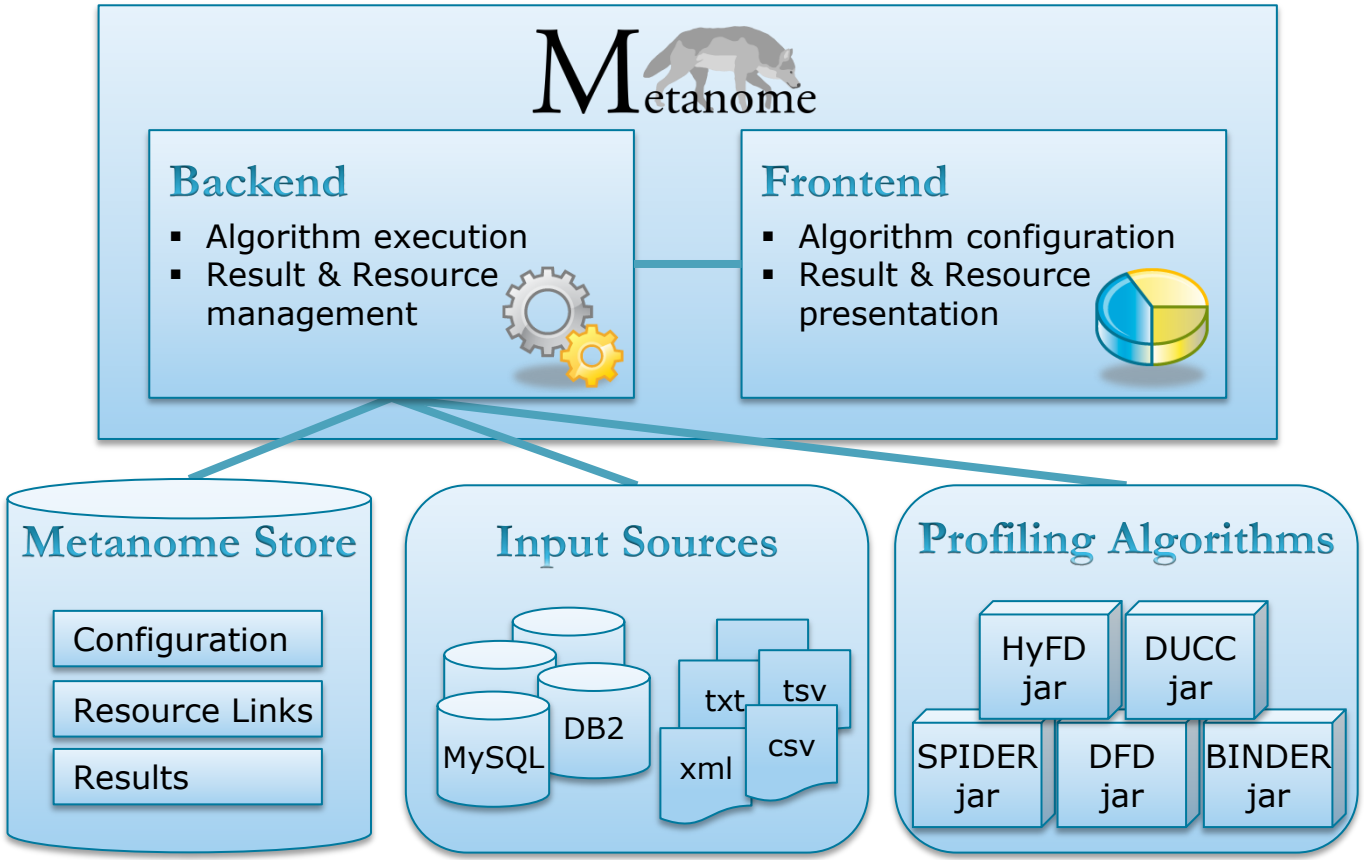
5 | 2016 | 2017 | 2018

HPI Hasso Plattner Institut

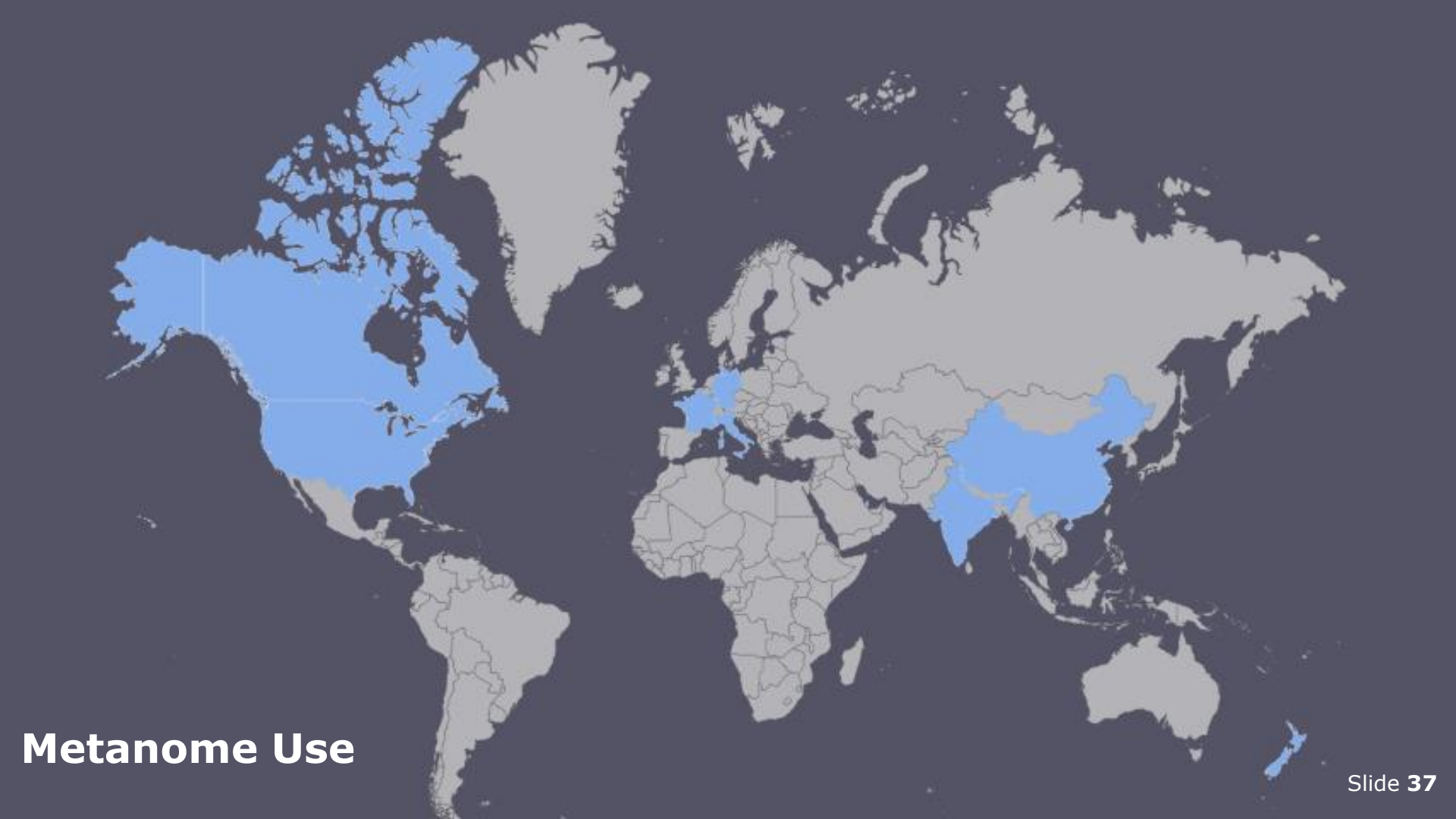
Dependency	Algorithms (exact)	Algorithms (approximate)
Unique Column Combination (UCC)	2	0
Inclusion Dependency (IND)	5	2
Functional Dependency (FD)	8	2
Order Dependency (OD)	1	0
Matching Dependency (MD)	2	0
Multi-valued Dependency (MvD)	1	0
Denial Constraints (DC)	1	0
Statistics	1	13
	21	**17**



[*Data Profiling with Metanome*, T. Papenbrock, T. Bergmann, M. Finke, J. Zwiener, F. Naumann, VLDB, 2015]



[*Data Profiling with Metanome*, T. Papenbrock, T. Bergmann, M. Finke, J. Zwiener, F. Naumann, VLDB, 2015]



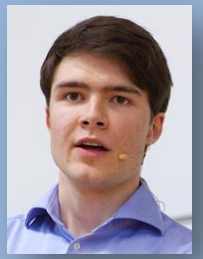
Metanome Use

Metanome

Algorithm Research

- Anja Jentzsch (RDF)
- Arvid Heise (UCC)
- Fabian Tschirschnitz (IND)
- Felix Naumann (Research lead)
- Hazar Harmouch (Single Column Profiling)
- Jens Ehrlich (Conditional UCC)
- Jorge-Arnulfo (UCC, IND)
- Maximilian Grundke (Conditional FD)
- Moritz Finke (Approximate FD/IND)
- Philipp Langer (OD)
- Philipp Schirmer (MVD)
- Sebastian Kruse (IND, partial FD; Metadata Store)
- Thorsten Papenbrock (IND, UCC, FD, …; Metanome)
- Tim Draeger (MVD)
- Tobias Bleifuß (DC)
- Ziawasch Abedjan (UCC)

Tool Development

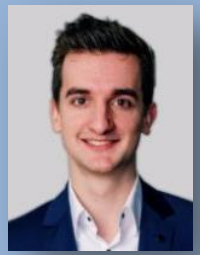


Jakob Zwiener
(Backend & Frontend)

Claudia Exeler
(Frontend)

Tanja Bergmann
(Backend & Frontend)

Moritz Finke
(Backend)



Carl Ambroselli
(Frontend)

Maxi Fischer
(Backend & Frontend)

Vincent Schwarzer
(Backend)

2019 **DynFD: Functional Dependency Discovery in Dynamic Datasets**
P. Schirmer, T. Papenbrock, S. Kruse, D. Hempfing, T. Mayer, D. Neuschäfer-Rube, F. Naumann (**EDBT**)
An Actor Database System for Akka
S. Schmidl, F. Schneider, T. Papenbrock (**BTW**)

2018 **Data Profiling – Synthesis Lectures on Data Management**
Z. Abedjan, L. Golab, F. Naumann, T. Papenbrock (**Morgan & Claypool**)

2017 **Detecting Inclusion Dependencies on Very Many Tables**
F. Tschirschnitz, T. Papenbrock, F. Naumann (**TODS**)
Data-driven Schema Normalization
T. Papenbrock, F. Naumann (**EDBT**)
A Hybrid Approach for Efficient Unique Column Combination Discovery
T. Papenbrock, F. Naumann (**BTW**)
Fast Approximate Discovery of Inclusion Dependencies
S. Kruse, T. Papenbrock, C. Dullweber, M. Finke, M. Hegner, M. Zabel, C. Zöllner, F. Naumann (**BTW**)

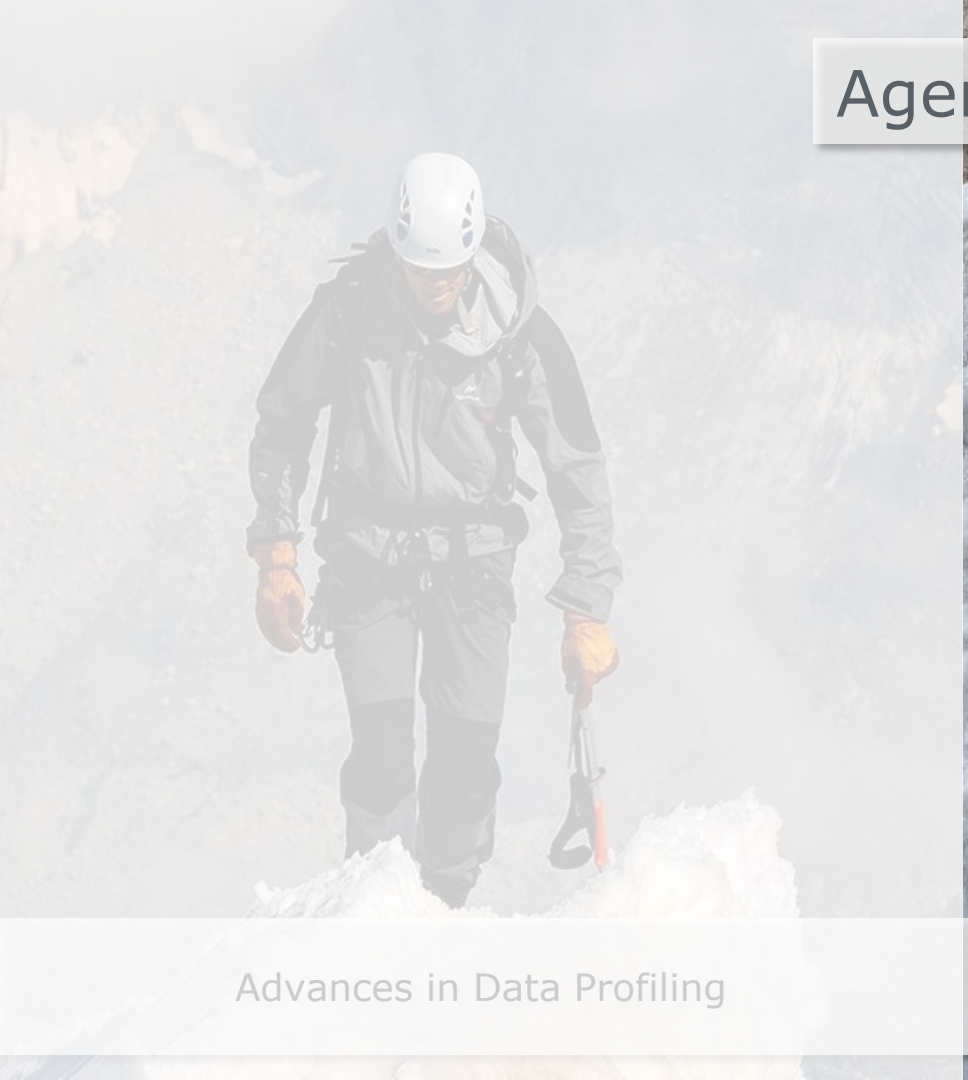
2016 **A Hybrid Approach to Functional Dependency Discovery**
T. Papenbrock, F. Naumann (**SIGMOD**)
Data Anamnesis: Admitting Raw Data into an Organization
S. Kruse, T. Papenbrock, H. Harmouch, F. Naumann (**IEEE Data Engineering Bulletin**)
Holistic Data Profiling: Simultaneous Discovery of Various Metadata
J. Ehrlich, M. Roick, L. Schulze, J. Zwiener, T. Papenbrock, F. Naumann (**EDBT**)
RDFind: Scalable Conditional Inclusion Dependency Discovery in RDF Datasets
S. Kruse, A. Jentzsch, T. Papenbrock, Z. Kaoudi, J. Quiané-Ruiz, F. Naumann (**SIGMOD**)
Approximate Discovery of Functional Dependencies for Large Datasets
T. Bleifuß, S. Bülow, J. Frohnhofen, J. Risch, G. Wiese, S. Kruse, T. Papenbrock, F. Naumann (**CIKM**)

2015 **Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms**
T. Papenbrock, J. Ehrlich, J. Marten, T. Neubert, J. Rudolph, M. Schönberg, J. Zwiener, F. Naumann (**VLDB**)
Data Profiling with Metanome
T. Papenbrock, T. Bergmann, M. Finke, J. Zwiener, F. Naumann (**VLDB**)
Divide & Conquer-based Inclusion Dependency Discovery
T. Papenbrock, S. Kruse, J. Quiané-Ruiz, F. Naumann (**VLDB**)
Scaling Out the Discovery of Inclusion Dependencies
S. Kruse, T. Papenbrock, F. Naumann (**BTW**)
Progressive Duplicate Detection
T. Papenbrock, A. Heise, F. Naumann (**TKDE**)

2013 **Ein Datenbankkurs mit 6000 Teilnehmern**
F. Naumann, M. Jenders, T. Papenbrock (**Informatik-Spektrum**)
Duplicate Detection on GPUs
B. Forchhammer, T. Papenbrock, T. Stening, S. Viehmeier, U. Draisbach, F. Naumann (**BTW**)

2011 **BlackSwan: Augmenting Statistics with Event Data**
J. Lorey, F. Naumann, B. Forchhammer, A. Mascher, P. Retzlaff, A. Zamani Farahani, S. Discher, C. Fähnrich, S. Lemme, T. Papenbrock, R. C. Peschel, S. Richter, T. Stening, S. Viehmeier (**CIKM**)

Agenda



Advances in Data Profiling

Challenges in Distributed Computing

Query Optimization
(with Jan Kossmann, EPIC Chair)

Data Cleaning
(with Ioannis Koumarelas, IS Chair)

Application

Matching Dependency Discovery
(with Philipp Schirmer, Bakdata)

Denial Constraint Discovery
(with Eduardo Pena, IS Chair)

Discovery

Distributed UCC Discovery
(with Martin Schirneck, AE Chair)

Distributed FD Discovery
(with Felix Naumann, IS Chair)

Distribution

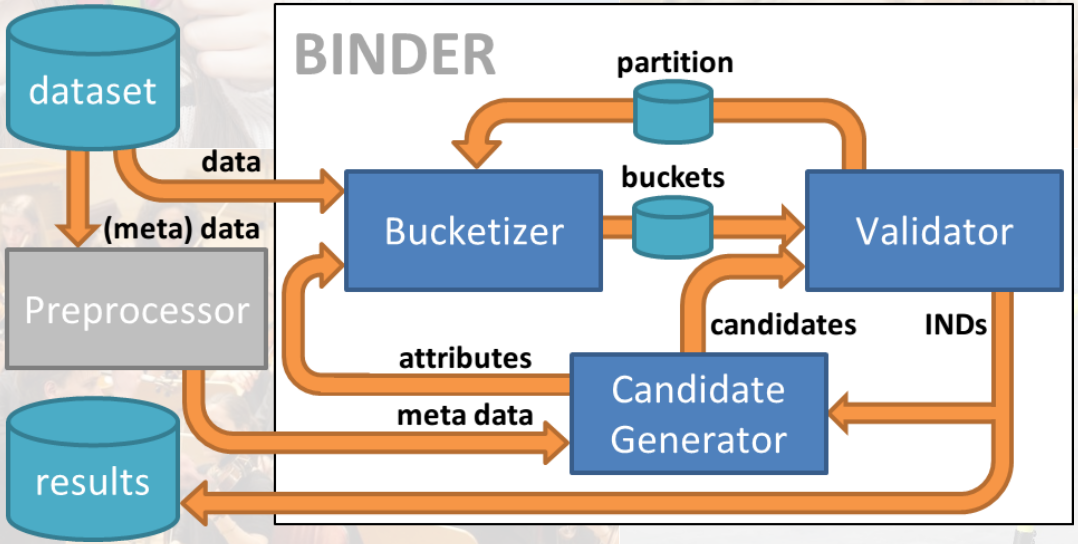
Reliable Distributed Systems
(MA seminar, WS 2019)

Mining Streaming Data
(MA seminar, WS 2019)

Robustness

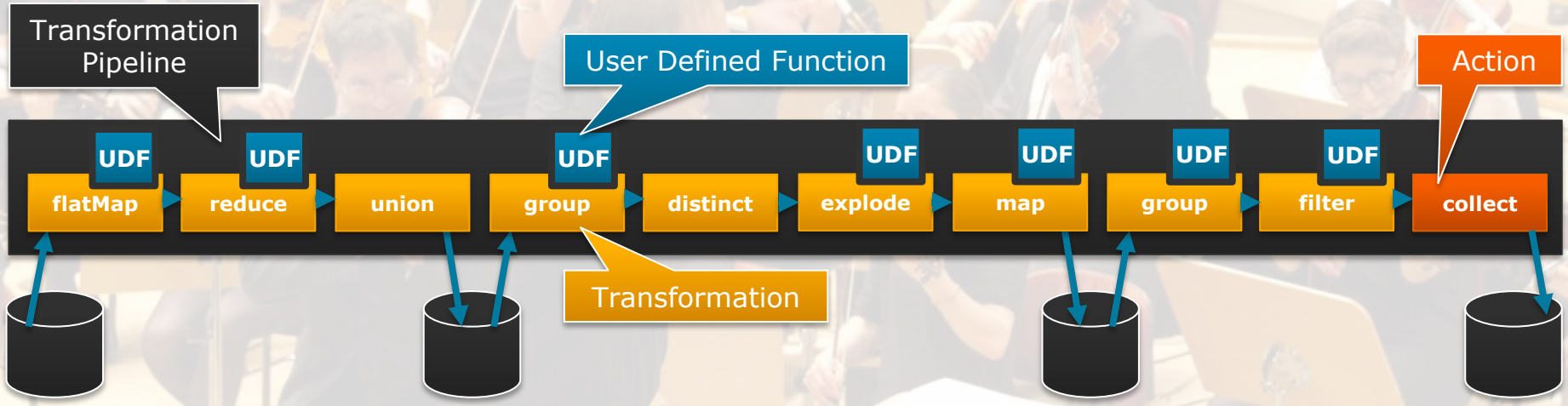
The Standard Batch Approach

Apache Flink



Distribution

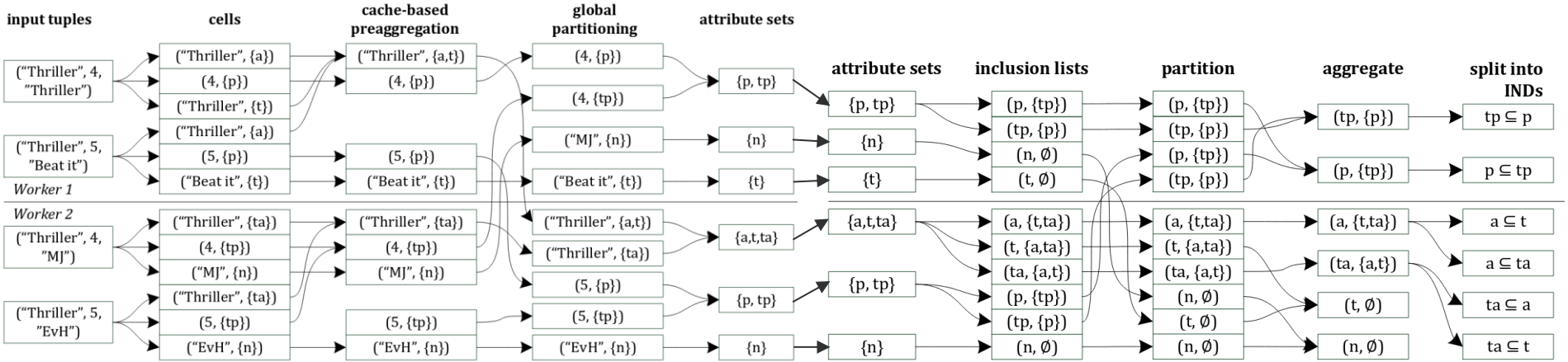
The Standard Batch Approach



[*Scaling Out the Discovery of Inclusion Dependencies*, S. Kruse, T. Papenbrock, F. Naumann, BTW, 2015]

The Standard Batch Approach

Apache Flink

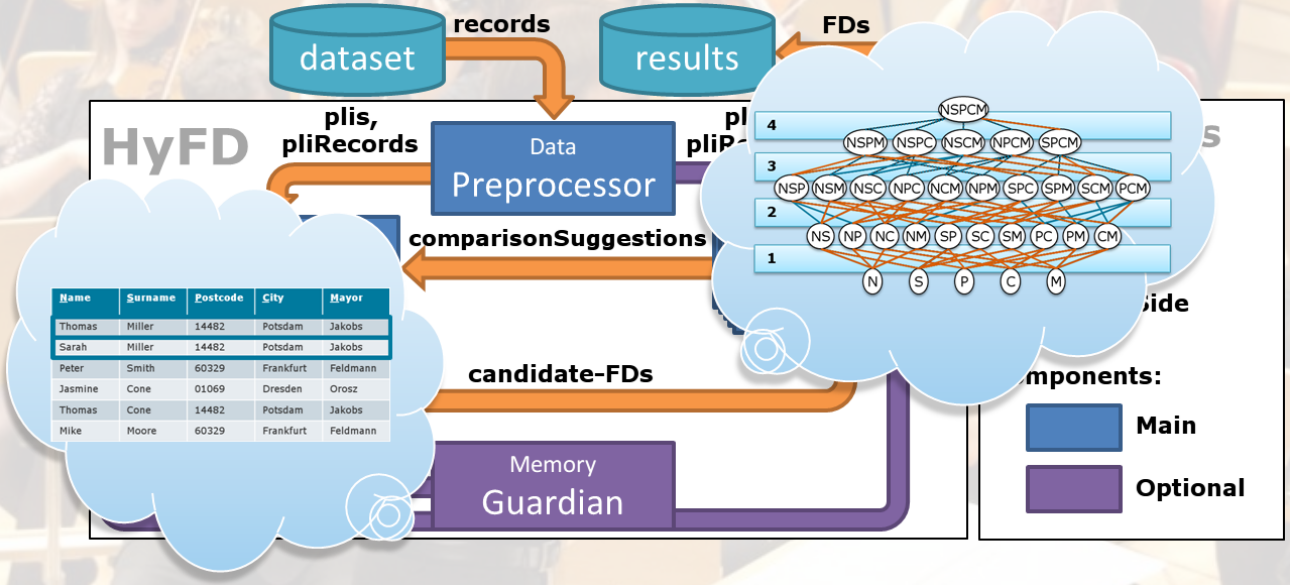


Distribution

[*Scaling Out the Discovery of Inclusion Dependencies*, S. Kruse, T. Papenbrock, F. Naumann, BTW, 2015]

The Standard Batch Approach

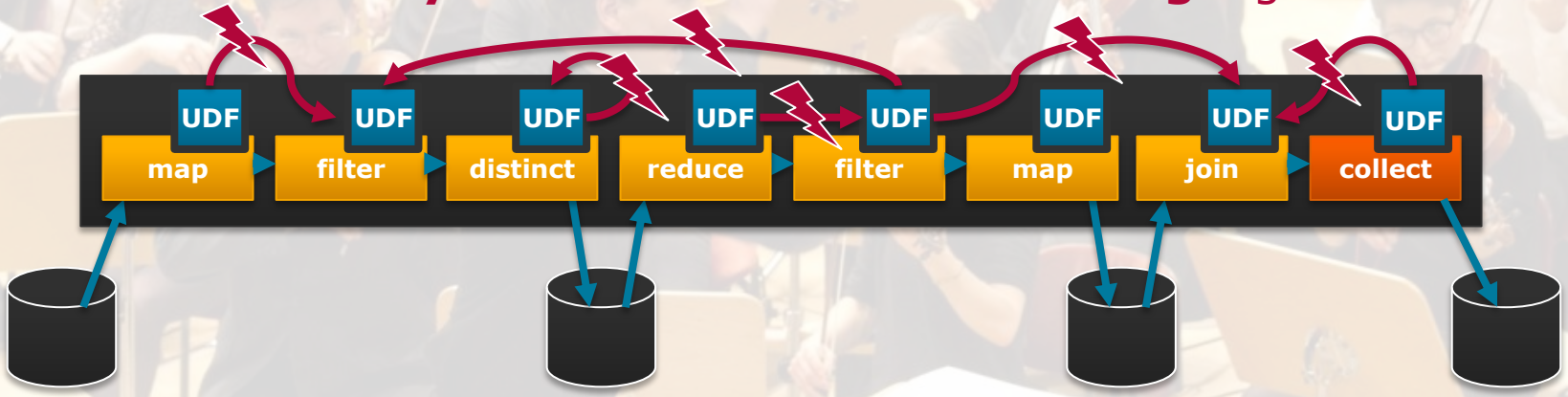
Apache Flink



The Standard Batch Approach

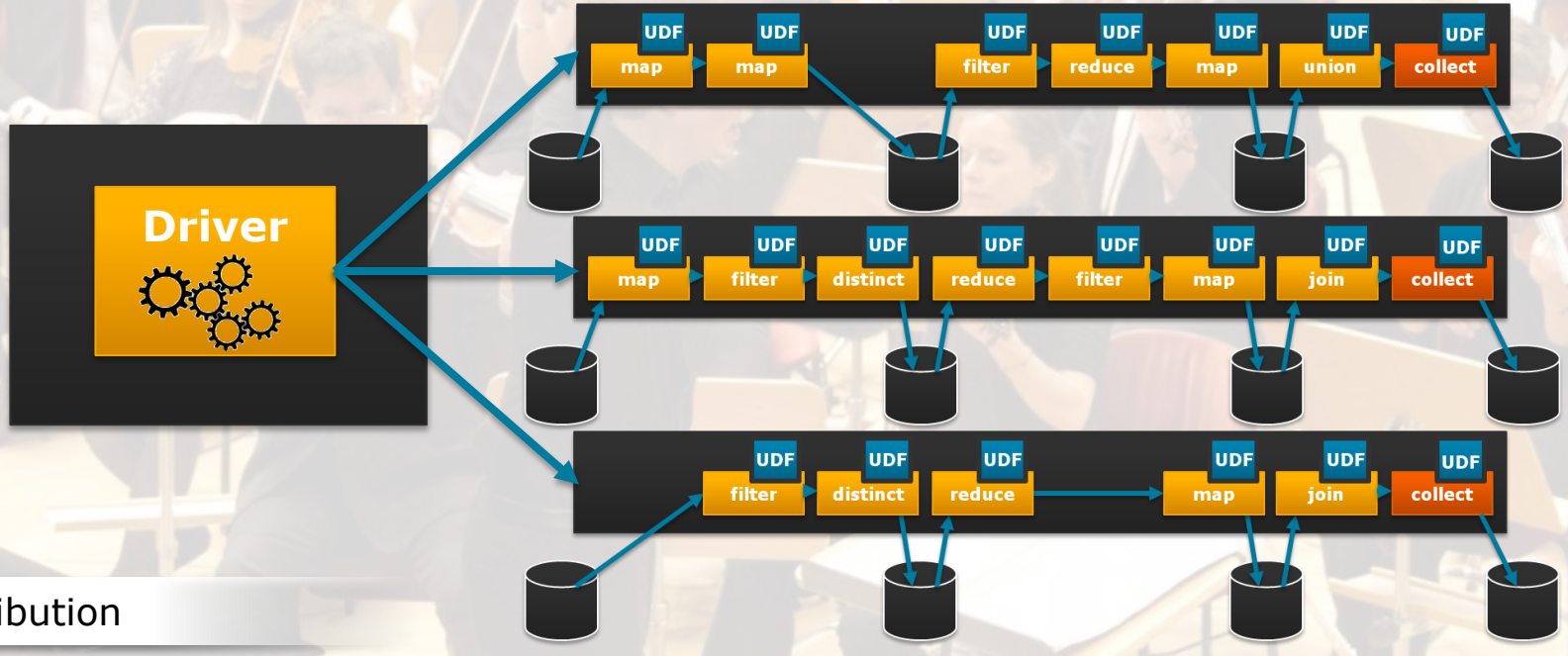
Apache Flink

dynamic behavior and **branching** logic



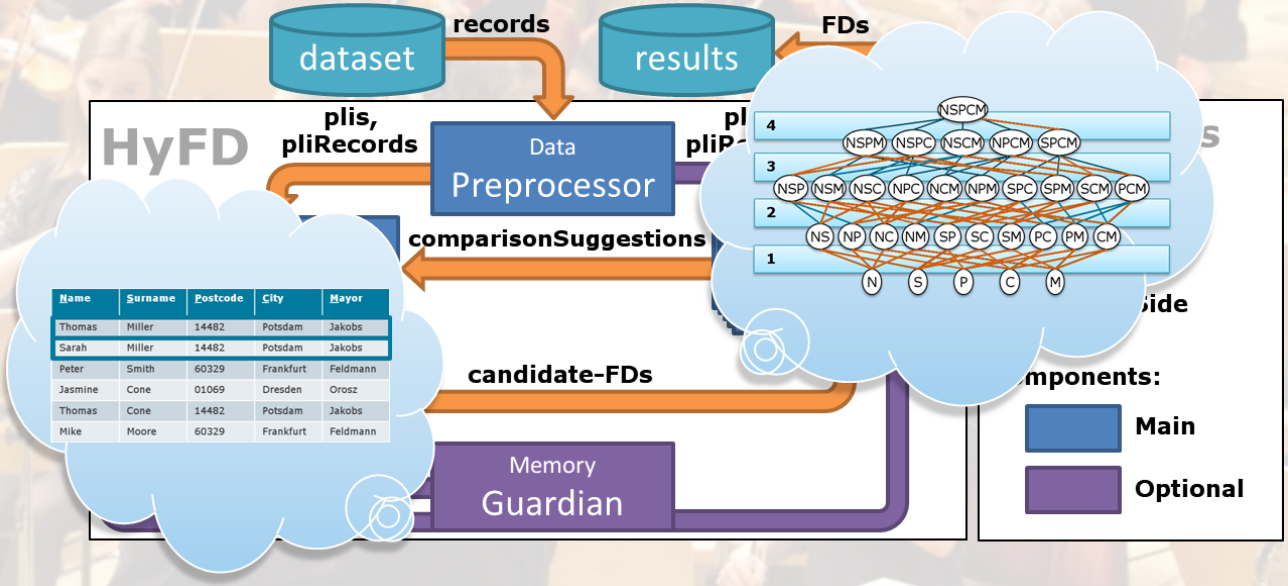
The Standard Batch Approach

Apache Flink

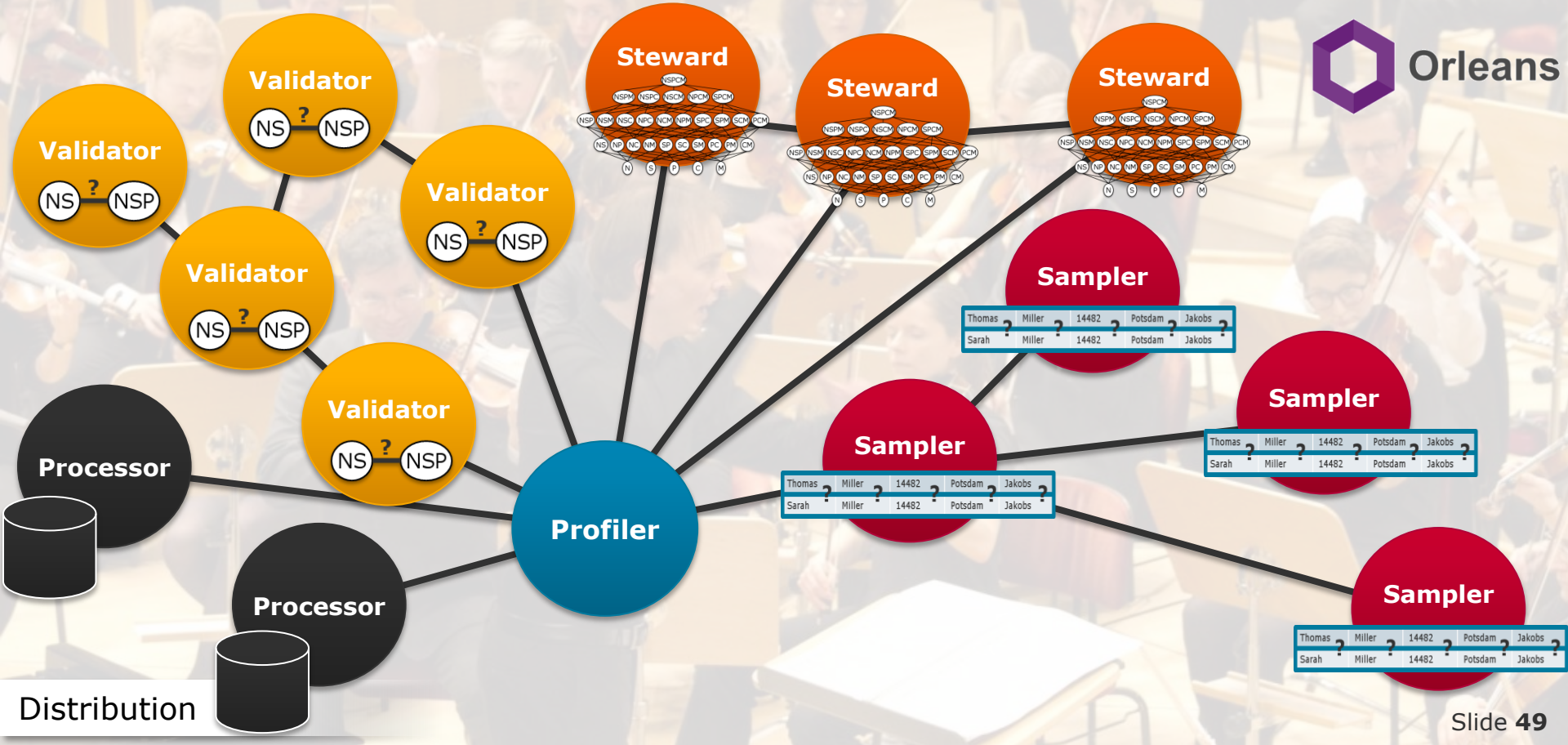


Distribution

The Low-Level Message-Passing Approach



The Low-Level Message-Passing Approach



Distributed Computing

- Distribution
  - Elasticity
  - Dynamic
  - Scalability
- Engineering
  - Integration
  - Profiling
  - Cleaning
  - Exploration
  - Linkage
- Parallelism
- Reliability
  - Robustness
  - Fault-tolerance
- Data
  - Quality
  - Semantics
  - Metadata