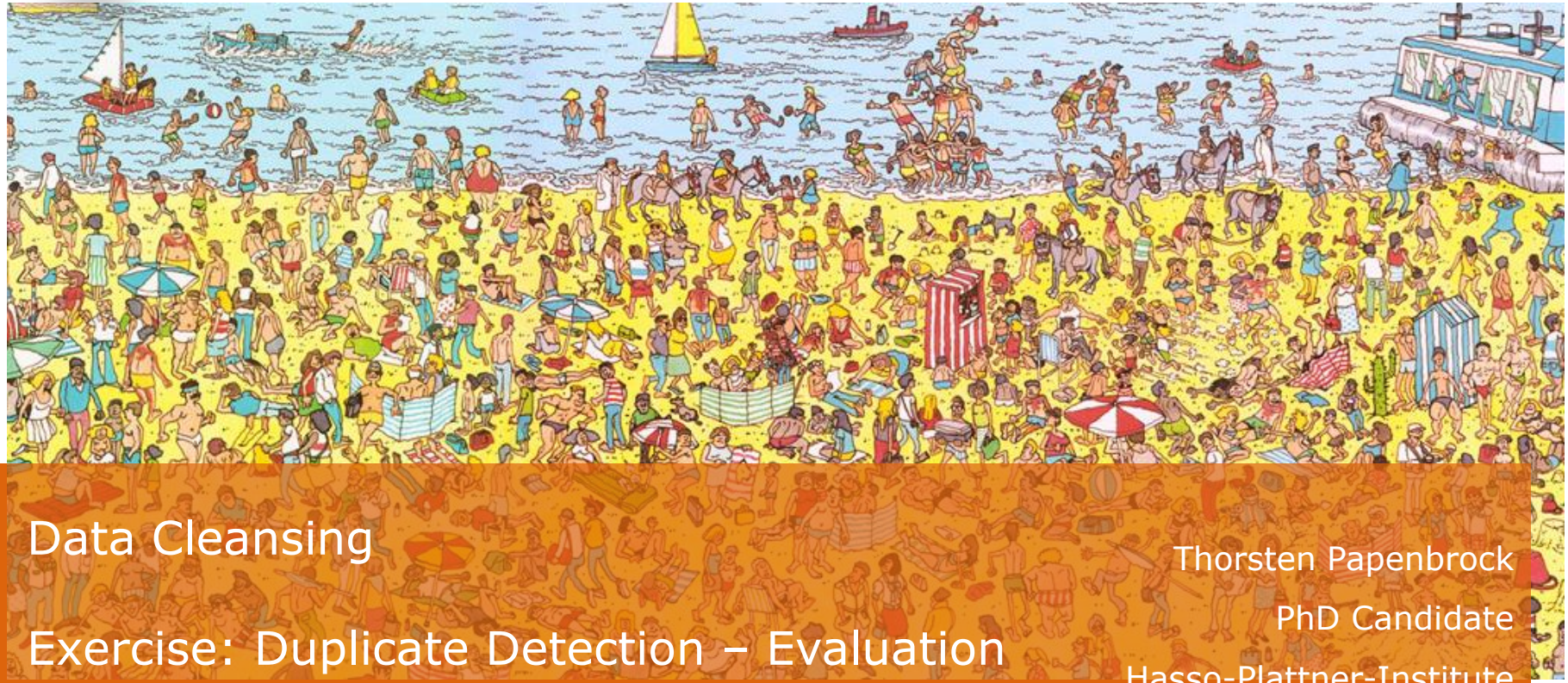
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Data Cleansing

Exercise: Duplicate Detection – Evaluation

Thorsten Papenbrock
PhD Candidate
Hasso-Plattner-Institute

Exercise 4
Our evaluation metrics

1. **Duplicates**

$$Duplicates = False\ Positives + True\ Positives$$

2. **Precision**

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3. **Recall**

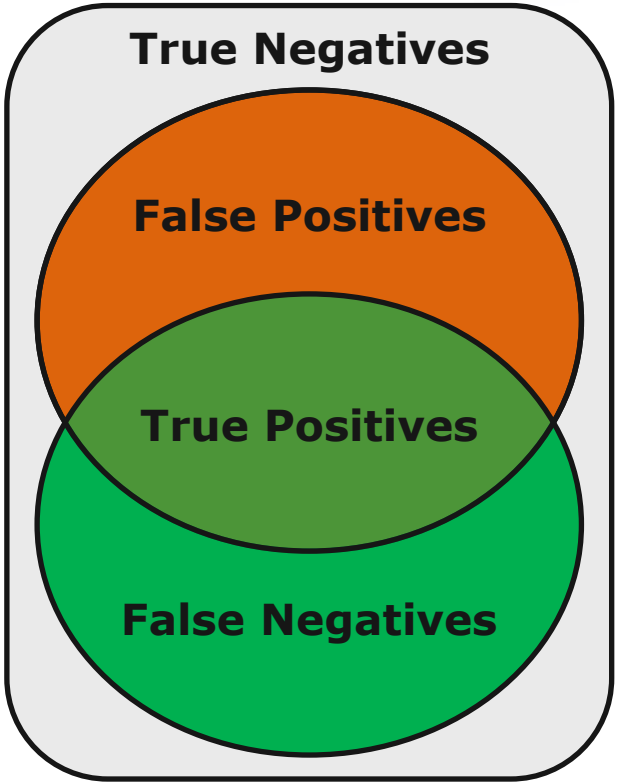
$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

4. **F-measure**

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall}$$

5. **F_β-measure**

$$F_\beta measure = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$



Exercise 4
F-measures

Hasso Plattner Institut

Rijsbergen's Effectiveness Measure

$$E_\alpha = 1 - \frac{1}{\frac{\alpha}{Precision} + \frac{1-\alpha}{Recall}}$$

F_α-Measure

$$F_\alpha = \frac{1}{\frac{\alpha}{Precision} + \frac{1-\alpha}{Recall}} \quad | * Precision * Recall$$

$$= \frac{Precision * Recall}{\alpha * Recall + (1-\alpha) * Precision} \quad | \ \alpha = \frac{1}{1+\beta^2}$$

F_β-Measure

$$F_\beta = \frac{Precision * Recall}{\frac{1}{1+\beta^2} * Recall + \left(1 - \frac{1}{1+\beta^2}\right) * Precision} \quad | * (1+\beta^2)$$

$$= (1+\beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

Exercise 4
F-measure in literature

HPI Hasso Plattner Institut

1. F_β-measure

$$F_\beta measure = (1 + \beta^2) * \frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

Most frequently used in literature

- $F_1 measure = Fmeasure$
- $\beta \in R$

2. F_α-measure

$$F_\alpha measure = (1 + \alpha) * \frac{Precision * Recall}{\alpha * Precision + Recall}$$

Sometimes used for simplifcation

- $F_1 measure = Fmeasure$
- $0 < \alpha < \infty$

3. F_α-measure

$$F_\alpha measure = \frac{Precision * Recall}{(1 - \alpha) * Precision + \alpha * Recall}$$

Used in lecture for intuition

- $F_{½} measure = Fmeasure$
- $0 < \alpha < 1$

Result						
alatar						
AlexoFredDuplicates						
cpdd						
DDBotheJoerkeReissaus						
dd_finke_dullweber						
dennis_marius						
DJ_DD						
dpdc-cnms-dude						
dreamteamdd						
Dubstep						
DuDePerchykSchmidt						
dude_kirsten_zwerg						
dude_scheffer_zoellner_results						
dupDect						
duplicateDetector						
duplicates_grundke_wiese						
DUP_SPIRO						
FiftyShadesOfDuplicates						
FrohnOttoDuDe						
HorLehDupDec						
HuhnHuhn						
iSwoosh						
JungRohloffDd						
klinger_marten_dude						
LucieKerstinDD						
MMDupDec						
MyDC						
ParseAndMatch						
REMdup						
RT_dupDec						
SBMMDEDUP						
smart_duplicat_cat						
Tsun12DuDe						
YuckDude						
GOLD						
INTERSECTION						
UNION						
VOTING						

Data Profiling with Metanome

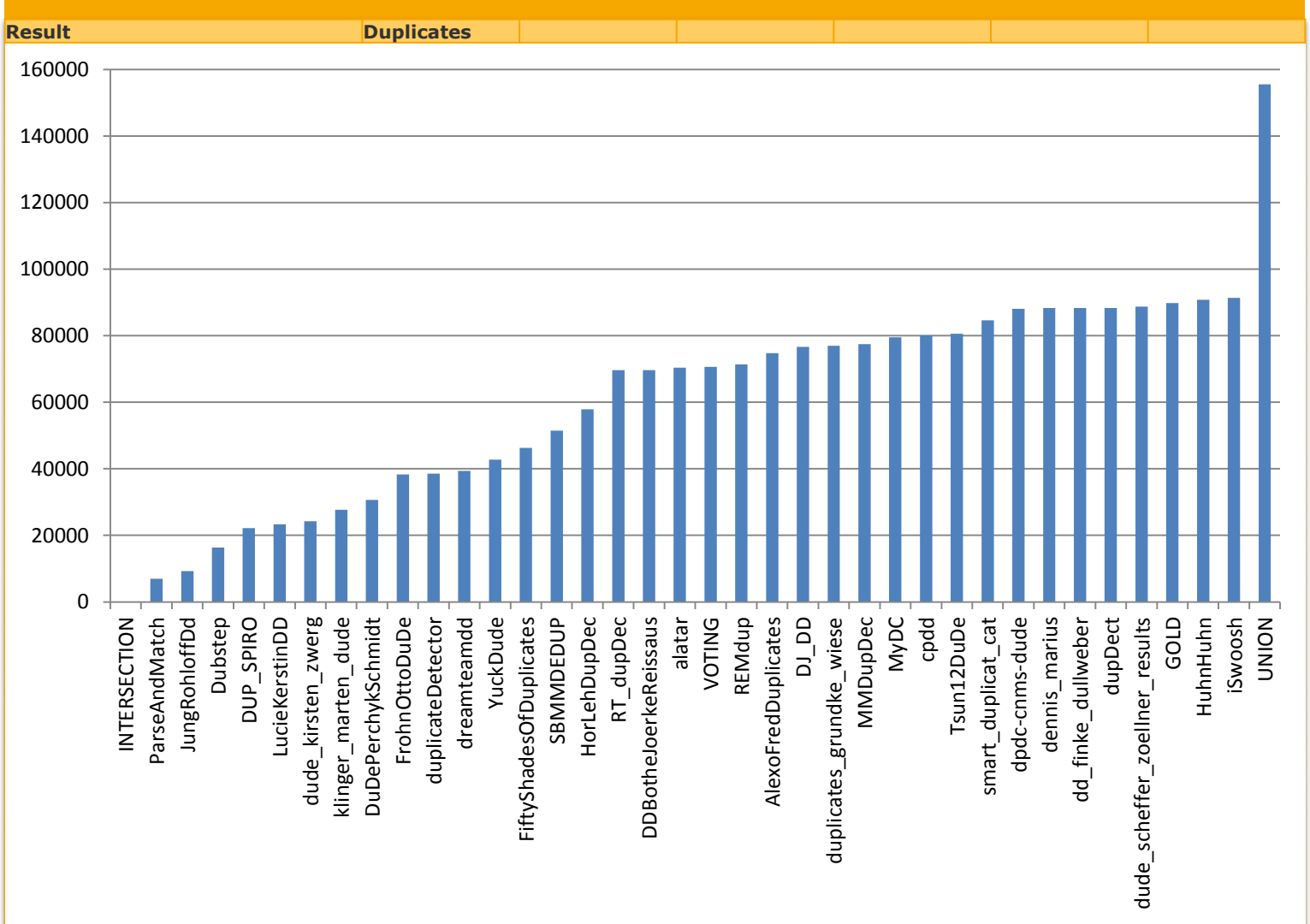
ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **5**

Result	Duplicates
INTERSECTION	**18**
ParseAndMatch	6953
JungRohloffDd	9298
Dubstep	16376
DUP_SPIRO	22144
LucieKerstinDD	23341
dude_kirsten_zwerg	24269
klinger_marten_dude	27716
DuDePerchykSchmidt	30675
FrohnOttoDuDe	38301
duplicateDetector	38549
dreamteamdd	39329
YuckDude	42708
FiftyShadesOfDuplicates	46246
SBMMDEDUP	51496
HorLehDupDec	57902
RT_dupDec	69646
DDBotheJoerkeReissaus	69662
alatar	70372
VOTING	**70613**
REMdup	71335
AlexoFredDuplicates	74780
DJ_DD	76594
duplicates_grundke_wiese	76971
MMDupDec	77450
MyDC	79523
cpdd	80071
Tsun12DuDe	80605
smart_duplicat_cat	84602
dpdc-cnms-dude	88091
dennis_marius	88296
dd_finke_dullweber	88309
dupDect	88342
dude_scheffer_zoellner_results	88731
GOLD	**89784**
HuhnHuhn	90764
iSwoosh	91370
UNION	**155491**

HPI Hasso Plattner Institut

Data Profiling with Metanome

ThorstenPapenbrock, PhD Candidate, 17th November, 2014
Chart **6**



HPI Hasso Plattner Institut

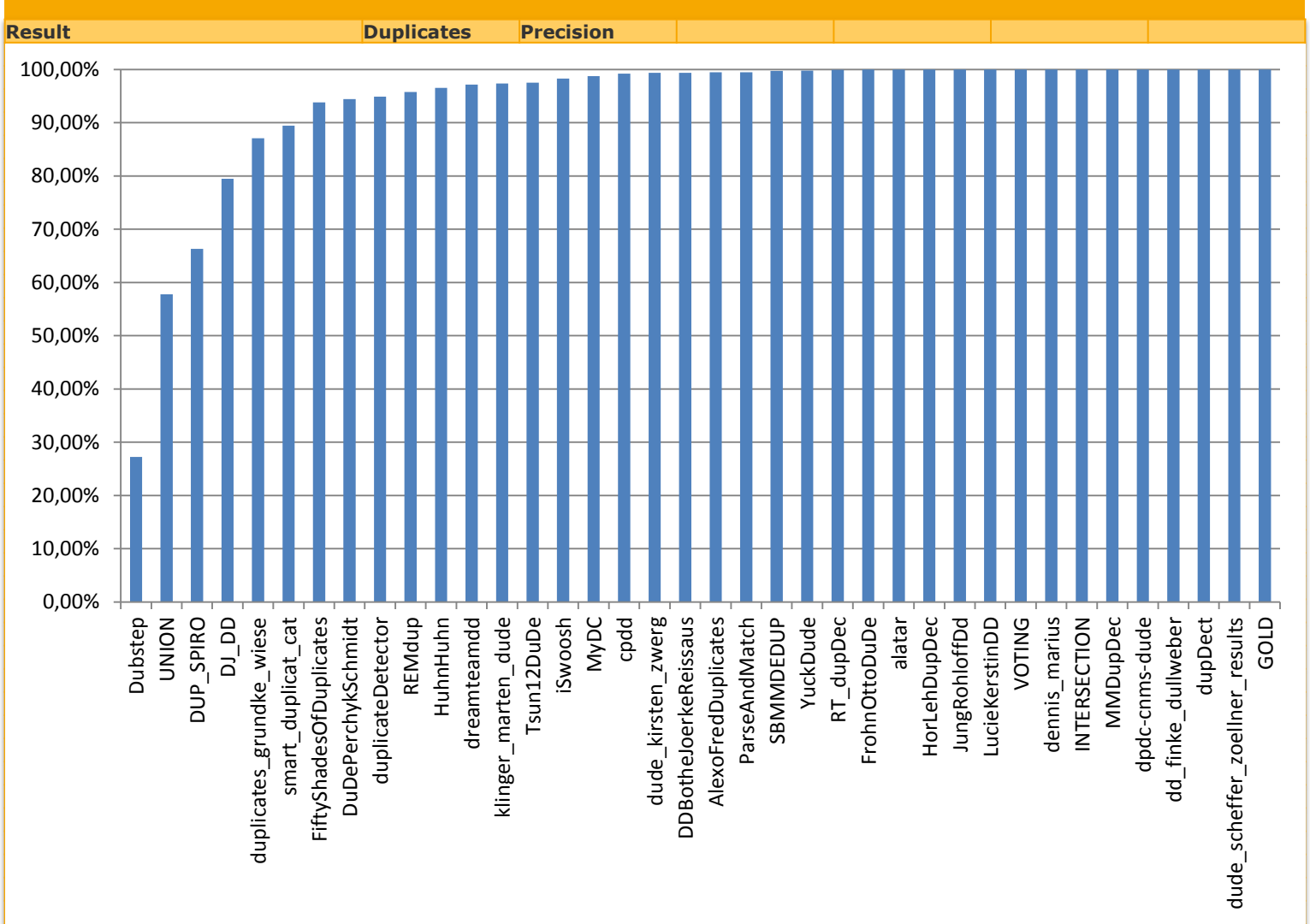
Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **7**

Result	Duplicates	Precision
Dubstep	16376	27,25%
UNION	**155491**	**57,74%**
DUP_SPIRO	22144	66,28%
DJ_DD	76594	79,45%
duplicates_grundke_wiese	76971	87,08%
smart_duplicat_cat	84602	89,46%
FiftyShadesOfDuplicates	46246	93,80%
DuDePerchykSchmidt	30675	94,44%
duplicateDetector	38549	94,90%
REMdup	71335	95,76%
HuhnHuhn	90764	96,52%
dreamteamdd	39329	97,16%
klinger_marten_dude	27716	97,37%
Tsun12DuDe	80605	97,51%
iSwoosh	91370	98,26%
MyDC	79523	98,74%
cpdd	80071	99,20%
dude_kirsten_zwerg	24269	99,34%
DDBotheJoerkeReissaus	69662	99,38%
AlexoFredDuplicates	74780	99,47%
ParseAndMatch	6953	99,48%
SBMMDEDUP	51496	99,70%
YuckDude	42708	99,77%
RT_dupDec	69646	99,93%
FrohnOttoDuDe	38301	99,96%
alatar	70372	99,98%
HorLehDupDec	57902	99,99%
JungRohloffDd	9298	99,99%
LucieKerstinDD	23341	99,99%
VOTING	**70613**	**100,00%**
dennis_marius	88296	100,00%
INTERSECTION	**18**	**100,00%**
MMDupDec	77450	100,00%
dpdc-cnms-dude	88091	100,00%
dd_finke_dullweber	88309	100,00%
dupDect	88342	100,00%
dude_scheffer_zoellner_results	88731	100,00%
GOLD	**89784**	**100,00%**

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **8**



Hasso Plattner Institut

Data Profiling with Metanome

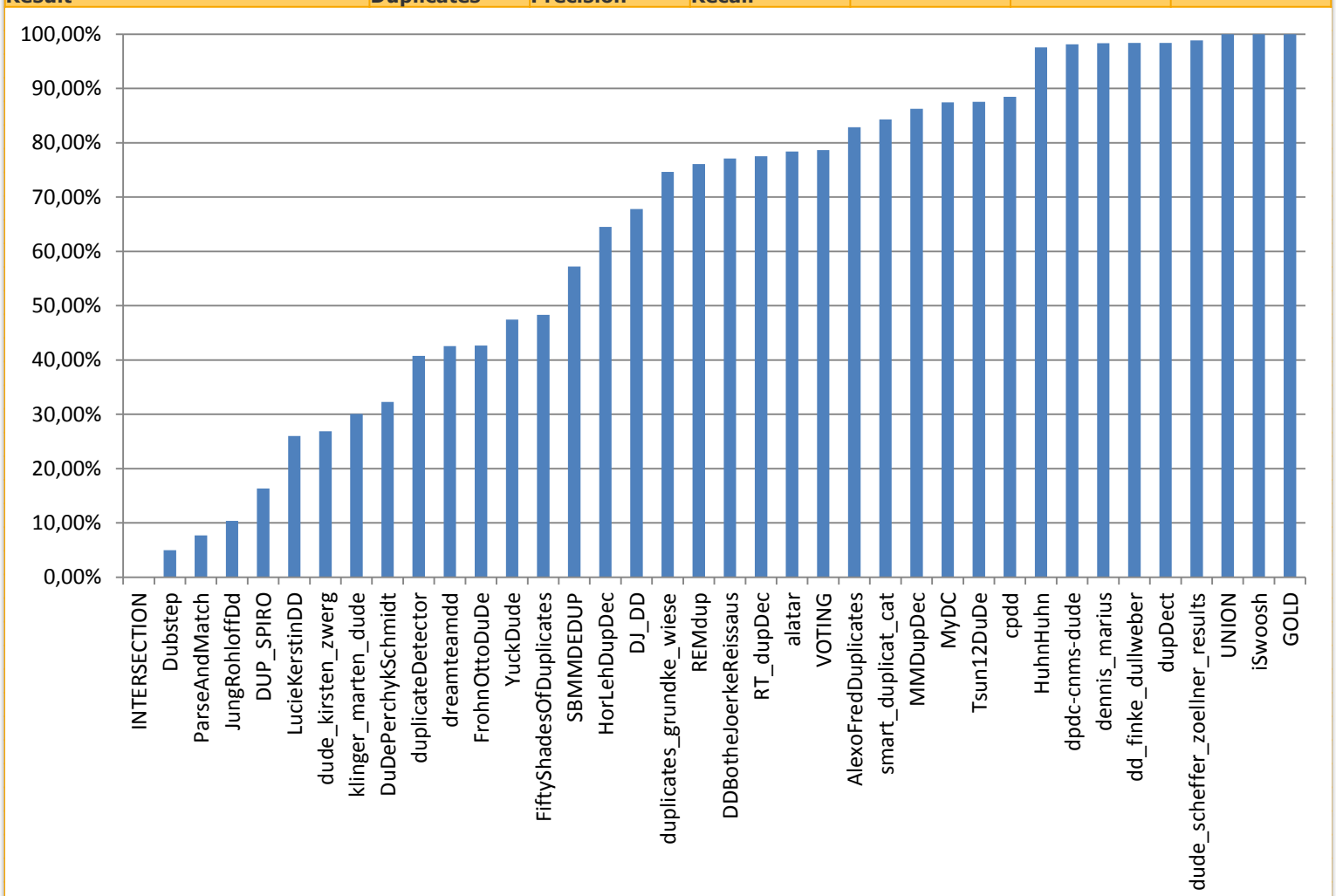
ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **9**

Result	Duplicates	Precision	Recall			
INTERSECTION	**18**	**100,00%**	**0,02%**			
Dubstep	16376	27,25%	4,97%			
ParseAndMatch	6953	99,48%	7,70%			
JungRohloffDd	9298	99,99%	10,35%			
DUP_SPIRO	22144	66,28%	16,35%			
LucieKerstinDD	23341	99,99%	25,99%			
dude_kirsten_zwerg	24269	99,34%	26,85%			
klinger_marten_dude	27716	97,37%	30,06%			
DuDePerchykSchmidt	30675	94,44%	32,26%			
duplicateDetector	38549	94,90%	40,75%			
dreamteamdd	39329	97,16%	42,56%			
FrohnOttoDuDe	38301	99,96%	42,64%			
YuckDude	42708	99,77%	47,46%			
FiftyShadesOfDuplicates	46246	93,80%	48,31%			
SBMMDEDUP	51496	99,70%	57,19%			
HorLehDupDec	57902	99,99%	64,48%			
DJ_DD	76594	79,45%	67,78%			
duplicates_grundke_wiese	76971	87,08%	74,65%			
REMdup	71335	95,76%	76,08%			
DDBotheJoerkeReissaus	69662	99,38%	77,11%			
RT_dupDec	69646	99,93%	77,51%			
alatar	70372	99,98%	78,36%			
VOTING	**70613**	**100,00%**	**78,64%**			
AlexoFredDuplicates	74780	99,47%	82,85%			
smart_duplicat_cat	84602	89,46%	84,29%			
MMDupDec	77450	100,00%	86,26%			
MyDC	79523	98,74%	87,46%			
Tsun12DuDe	80605	97,51%	87,54%			
cpdd	80071	99,20%	88,47%			
HuhnHuhn	90764	96,52%	97,57%			
dpdc-cnms-dude	88091	100,00%	98,11%			
dennis_marius	88296	100,00%	98,34%			
dd_finke_dullweber	88309	100,00%	98,36%			
dupDect	88342	100,00%	98,39%			
dude_scheffer_zoellner_results	88731	100,00%	98,83%			
UNION	**155491**	**57,74%**	**100,00%**			
iSwoosh	91370	98,26%	100,00%			
GOLD	**89784**	**100,00%**	**100,00%**			

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **10**

Result	Duplicates	Precision	Recall



HPI Hasso Plattner Institut

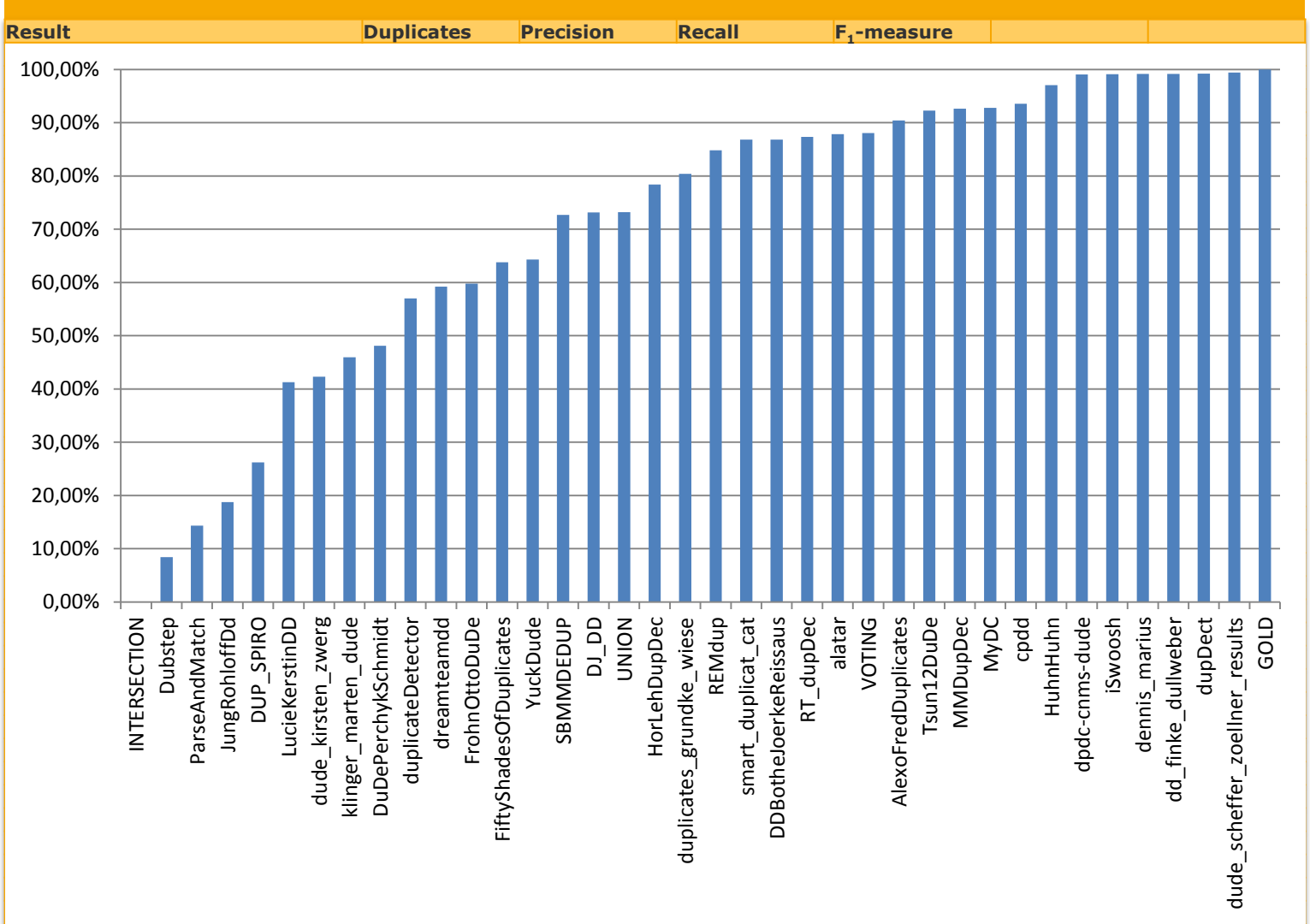
Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **11**

Result	Duplicates	Precision	Recall	F_1-measure		
INTERSECTION	**18**	**100,00%**	**0,02%**	**0,04%**		
Dubstep	16376	27,25%	4,97%	8,41%		
ParseAndMatch	6953	99,48%	7,70%	14,30%		
JungRohloffDd	9298	99,99%	10,35%	18,77%		
DUP_SPIRO	22144	66,28%	16,35%	26,22%		
LucieKerstinDD	23341	99,99%	25,99%	41,26%		
dude_kirsten_zwerg	24269	99,34%	26,85%	42,28%		
klinger_marten_dude	27716	97,37%	30,06%	45,93%		
DuDePerchykSchmidt	30675	94,44%	32,26%	48,10%		
duplicateDetector	38549	94,90%	40,75%	57,01%		
dreamteamdd	39329	97,16%	42,56%	59,19%		
FrohnOttoDuDe	38301	99,96%	42,64%	59,78%		
FiftyShadesOfDuplicates	46246	93,80%	48,31%	63,78%		
YuckDude	42708	99,77%	47,46%	64,32%		
SBMMDEDUP	51496	99,70%	57,19%	72,68%		
DJ_DD	76594	79,45%	67,78%	73,15%		
UNION	**155491**	**57,74%**	**100,00%**	**73,21%**		
HorLehDupDec	57902	99,99%	64,48%	78,40%		
duplicates_grundke_wiese	76971	87,08%	74,65%	80,39%		
REMdup	71335	95,76%	76,08%	84,79%		
smart_duplicat_cat	84602	89,46%	84,29%	86,80%		
DDBotheJoerkeReissaus	69662	99,38%	77,11%	86,84%		
RT_dupDec	69646	99,93%	77,51%	87,31%		
alatar	70372	99,98%	78,36%	87,86%		
VOTING	**70613**	**100,00%**	**78,64%**	**88,04%**		
AlexoFredDuplicates	74780	99,47%	82,85%	90,40%		
Tsun12DuDe	80605	97,51%	87,54%	92,25%		
MMDupDec	77450	100,00%	86,26%	92,62%		
MyDC	79523	98,74%	87,46%	92,76%		
cpdd	80071	99,20%	88,47%	93,52%		
HuhnHuhn	90764	96,52%	97,57%	97,04%		
dpdc-cnms-dude	88091	100,00%	98,11%	99,05%		
iSwoosh	91370	98,26%	100,00%	99,12%		
dennis_marius	88296	100,00%	98,34%	99,16%		
dd_finke_dullweber	88309	100,00%	98,36%	99,17%		
dupDect	88342	100,00%	98,39%	99,19%		
dude_scheffer_zoellner_results	88731	100,00%	98,83%	99,41%		
GOLD	**89784**	**100,00%**	**100,00%**	**100,00%**		

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **12**



Hasso Plattner Institut

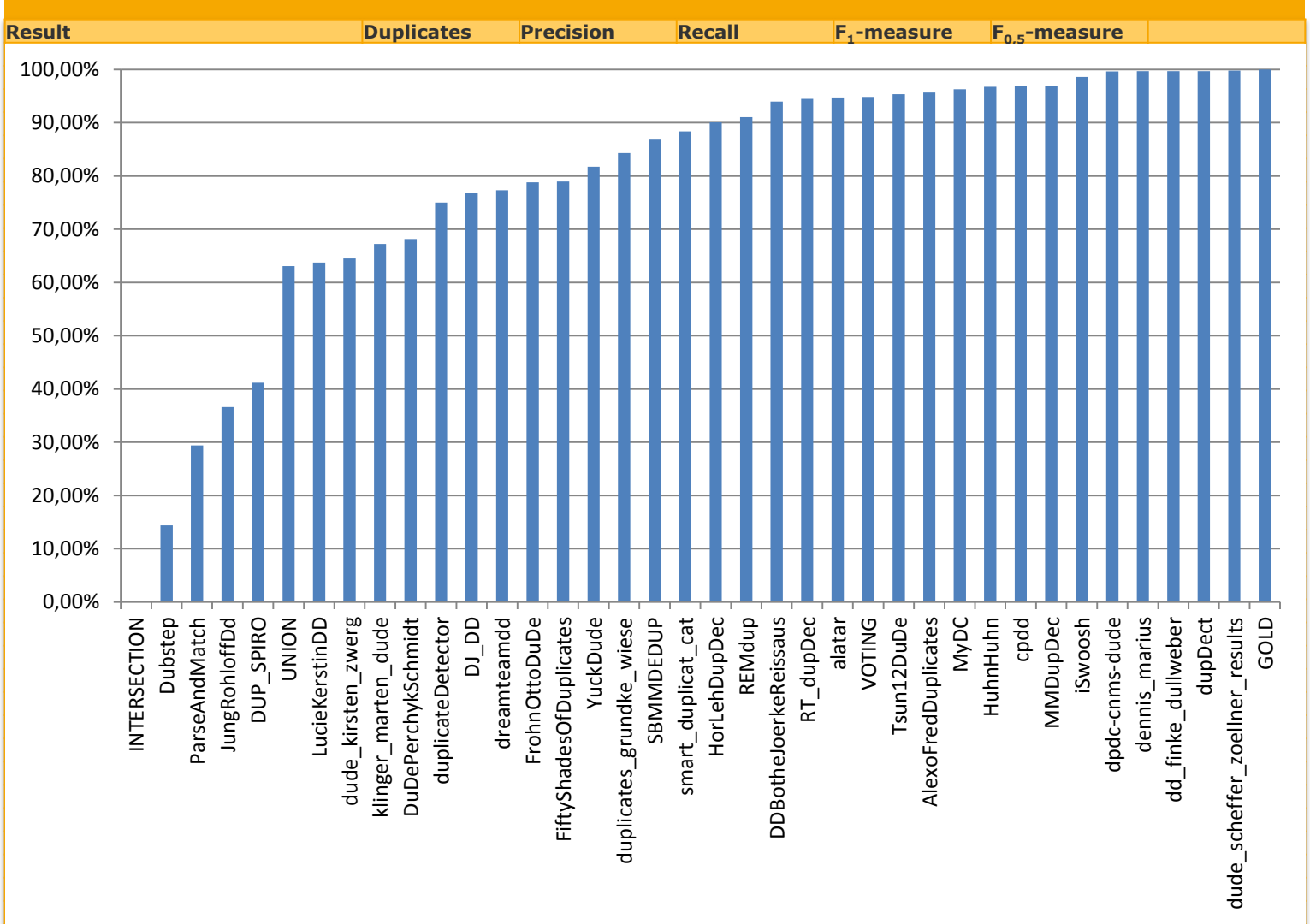
Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **13**

Result	Duplicates	Precision	Recall	F_1-measure	$F_{0,5}$-measure
INTERSECTION	**18**	**100,00%**	**0,02%**	**0,04%**	**0,10%**
Dubstep	16376	27,25%	4,97%	8,41%	14,37%
ParseAndMatch	6953	99,48%	7,70%	14,30%	29,41%
JungRohloffDd	9298	99,99%	10,35%	18,77%	36,61%
DUP_SPIRO	22144	66,28%	16,35%	26,22%	41,14%
UNION	**155491**	**57,74%**	**100,00%**	**73,21%**	**63,07%**
LucieKerstinDD	23341	99,99%	25,99%	41,26%	63,72%
dude_kirsten_zwerg	24269	99,34%	26,85%	42,28%	64,51%
klinger_marten_dude	27716	97,37%	30,06%	45,93%	67,25%
DuDePerchykSchmidt	30675	94,44%	32,26%	48,10%	68,17%
duplicateDetector	38549	94,90%	40,75%	57,01%	74,97%
DJ_DD	76594	79,45%	67,78%	73,15%	76,80%
dreamteamdd	39329	97,16%	42,56%	59,19%	77,32%
FrohnOttoDuDe	38301	99,96%	42,64%	59,78%	78,78%
FiftyShadesOfDuplicates	46246	93,80%	48,31%	63,78%	78,93%
YuckDude	42708	99,77%	47,46%	64,32%	81,75%
duplicates_grundke_wiese	76971	87,08%	74,65%	80,39%	84,27%
SBMMDEDUP	51496	99,70%	57,19%	72,68%	86,80%
smart_duplicat_cat	84602	89,46%	84,29%	86,80%	88,37%
HorLehDupDec	57902	99,99%	64,48%	78,40%	90,07%
REMdup	71335	95,76%	76,08%	84,79%	91,05%
DDBotheJoerkeReissaus	69662	99,38%	77,11%	86,84%	93,95%
RT_dupDec	69646	99,93%	77,51%	87,31%	94,47%
alatar	70372	99,98%	78,36%	87,86%	94,75%
VOTING	**70613**	**100,00%**	**78,64%**	**88,04%**	**94,85%**
Tsun12DuDe	80605	97,51%	87,54%	92,25%	95,34%
AlexoFredDuplicates	74780	99,47%	82,85%	90,40%	95,64%
MyDC	79523	98,74%	87,46%	92,76%	96,26%
HuhnHuhn	90764	96,52%	97,57%	97,04%	96,73%
cpdd	80071	99,20%	88,47%	93,52%	96,85%
MMDupDec	77450	100,00%	86,26%	92,62%	96,91%
iSwoosh	91370	98,26%	100,00%	99,12%	98,61%
dpdc-cnms-dude	88091	100,00%	98,11%	99,05%	99,62%
dennis_marius	88296	100,00%	98,34%	99,16%	99,66%
dd_finke_dullweber	88309	100,00%	98,36%	99,17%	99,67%
dupDect	88342	100,00%	98,39%	99,19%	99,67%
dude_scheffer_zoellner_results	88731	100,00%	98,83%	99,41%	99,76%
GOLD	**89784**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **14**



HPI Hasso Plattner Institut

Data Profiling with Metanome

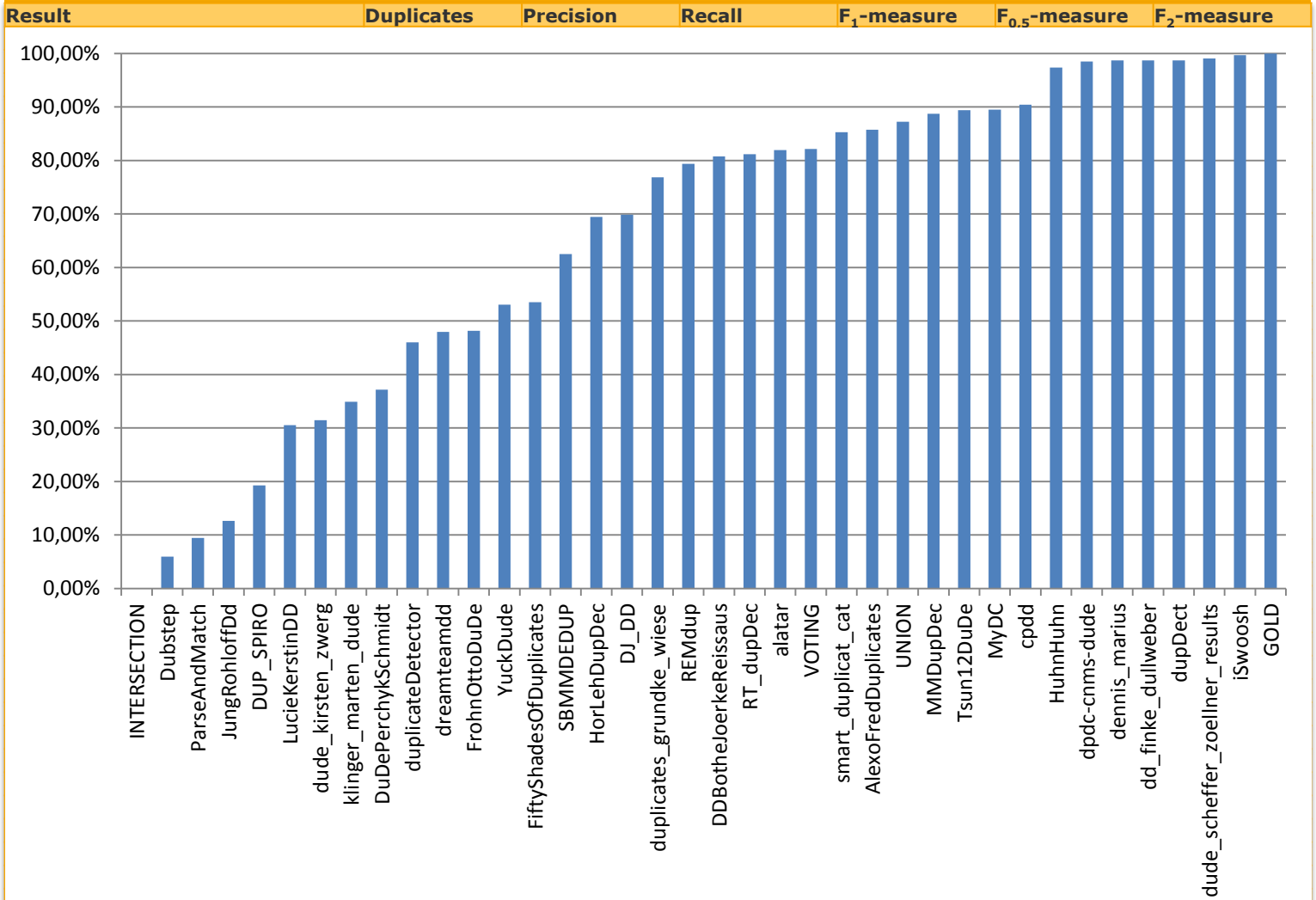
ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **15**

Result	Duplicates	Precision	Recall	F_1-measure	$F_{0,5}$-measure	F_2-measure
INTERSECTION	**18**	**100,00%**	**0,02%**	**0,04%**	**0,10%**	**0,03%**
Dubstep	16376	27,25%	4,97%	8,41%	14,37%	5,94%
ParseAndMatch	6953	99,48%	7,70%	14,30%	29,41%	9,45%
JungRohloffDd	9298	99,99%	10,35%	18,77%	36,61%	12,62%
DUP_SPIRO	22144	66,28%	16,35%	26,22%	41,14%	19,25%
LucieKerstinDD	23341	99,99%	25,99%	41,26%	63,72%	30,51%
dude_kirsten_zwerg	24269	99,34%	26,85%	42,28%	64,51%	31,44%
klinger_marten_dude	27716	97,37%	30,06%	45,93%	67,25%	34,88%
DuDePerchykSchmidt	30675	94,44%	32,26%	48,10%	68,17%	37,16%
duplicateDetector	38549	94,90%	40,75%	57,01%	74,97%	46,00%
dreamteamdd	39329	97,16%	42,56%	59,19%	77,32%	47,95%
FrohnOttoDuDe	38301	99,96%	42,64%	59,78%	78,78%	48,17%
YuckDude	42708	99,77%	47,46%	64,32%	81,75%	53,02%
FiftyShadesOfDuplicates	46246	93,80%	48,31%	63,78%	78,93%	53,50%
SBMMDEDUP	51496	99,70%	57,19%	72,68%	86,80%	62,52%
HorLehDupDec	57902	99,99%	64,48%	78,40%	90,07%	69,41%
DJ_DD	76594	79,45%	67,78%	73,15%	76,80%	69,83%
duplicates_grundke_wiese	76971	87,08%	74,65%	80,39%	84,27%	76,85%
REMdup	71335	95,76%	76,08%	84,79%	91,05%	79,34%
DDBotheJoerkeReissaus	69662	99,38%	77,11%	86,84%	93,95%	80,72%
RT_dupDec	69646	99,93%	77,51%	87,31%	94,47%	81,16%
alatar	70372	99,98%	78,36%	87,86%	94,75%	81,91%
VOTING	**70613**	**100,00%**	**78,64%**	**88,04%**	**94,85%**	**82,15%**
smart_duplicat_cat	84602	89,46%	84,29%	86,80%	88,37%	85,28%
AlexoFredDuplicates	74780	99,47%	82,85%	90,40%	95,64%	85,72%
UNION	**155491**	**57,74%**	**100,00%**	**73,21%**	**63,07%**	**87,23%**
MMDupDec	77450	100,00%	86,26%	92,62%	96,91%	88,70%
Tsun12DuDe	80605	97,51%	87,54%	92,25%	95,34%	89,37%
MyDC	79523	98,74%	87,46%	92,76%	96,26%	89,50%
cpdd	80071	99,20%	88,47%	93,52%	96,85%	90,42%
HuhnHuhn	90764	96,52%	97,57%	97,04%	96,73%	97,36%
dpdc-cnms-dude	88091	100,00%	98,11%	99,05%	99,62%	98,49%
dennis_marius	88296	100,00%	98,34%	99,16%	99,66%	98,67%
dd_finke_dullweber	88309	100,00%	98,36%	99,17%	99,67%	98,68%
dupDect	88342	100,00%	98,39%	99,19%	99,67%	98,71%
dude_scheffer_zoellner_results	88731	100,00%	98,83%	99,41%	99,76%	99,06%
iSwoosh	91370	98,26%	100,00%	99,12%	98,61%	99,65%
GOLD	**89784**	**100,00%**	**100,00%**	**100,00%**	**100,00%**	**100,00%**

HPI Hasso Plattner Institut

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **16**

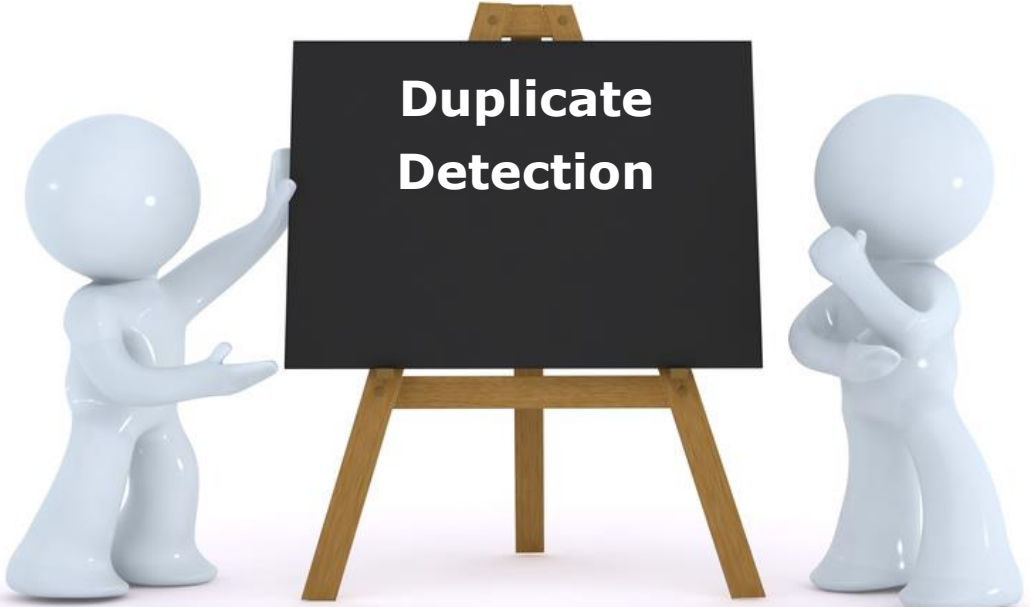


HPI Hasso Plattner Institut

Data Profiling with Metanome

ThorstenPapenbrock,
PhD Candidate,
17th November, 2014
Chart **17**

Exercise 4
Short presentations



Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Data Cleansing

Exercise: Duplicate Detection – Evaluation

Thorsten Papenbrock
PhD Candidate
Hasso-Plattner-Institute