Information Systems Group

Hasso Plattner Institut | Universität Potsdam

Workshop „Datenreinigung“ Auswertung

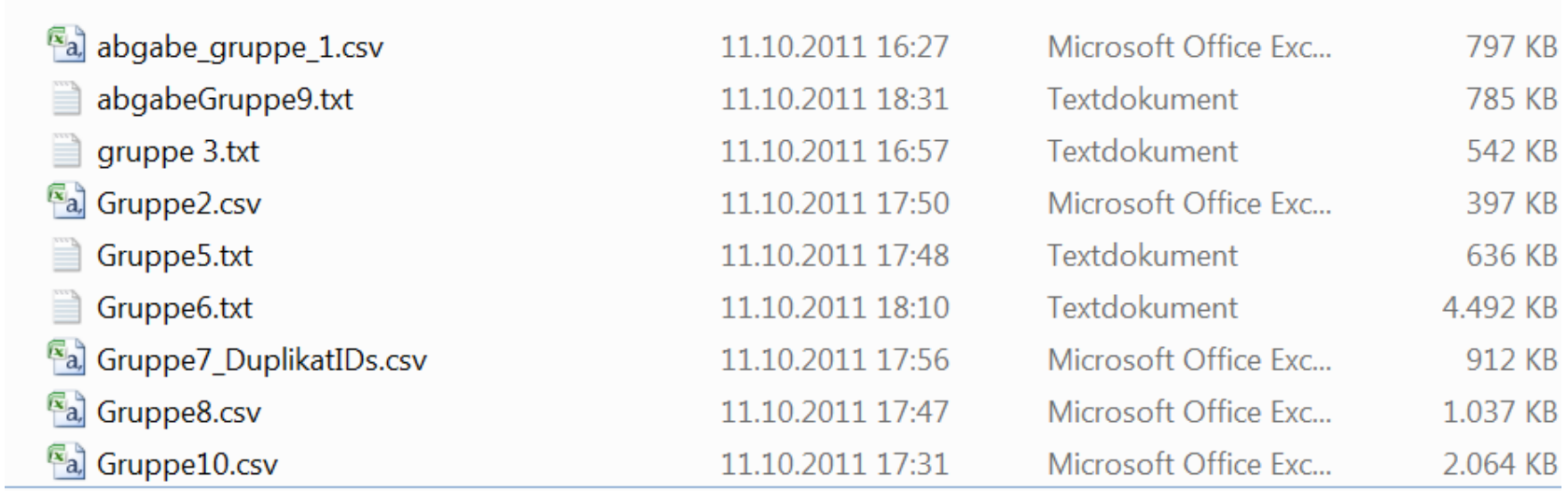
12.10.2011

Felix Naumann
Joachim Schmid
Uwe Draisbach

Reihenfolge der Vorträge

Gruppe 1	Henning, Grapentin
Gruppe 2	Eckert, Gurol
Gruppe 3	Marten, Klinger
Gruppe 5	Bunk, Bergmann
Gruppe 6	Schulze, Honauer
Gruppe 7	Noffke, Petrick
Gruppe 8	Lehmann, Ramson
Gruppe 9	Spivak, Swart
Gruppe 10	Würz, Gimbatschki

Dateigrößen



abgabe_gruppe_1.csv	11.10.2011 16:27	Microsoft Office Exc...	797 KB
abgabeGruppe9.txt	11.10.2011 18:31	Textdokument	785 KB
gruppe 3.txt	11.10.2011 16:57	Textdokument	542 KB
Gruppe2.csv	11.10.2011 17:50	Microsoft Office Exc...	397 KB
Gruppe5.txt	11.10.2011 17:48	Textdokument	636 KB
Gruppe6.txt	11.10.2011 18:10	Textdokument	4.492 KB
Gruppe7_DuplikatIDs.csv	11.10.2011 17:56	Microsoft Office Exc...	912 KB
Gruppe8.csv	11.10.2011 17:47	Microsoft Office Exc...	1.037 KB
Gruppe10.csv	11.10.2011 17:31	Microsoft Office Exc...	2.064 KB

SQL

- create table muster.musterl (id1 integer, id2 integer)

Korrektur

- SELECT * FROM muster.GRUPPE10 A1, muster.GRUPPE10 A2 WHERE A1.ID1 = A2.ID2 and A2.ID1 = A1.ID2 AND A1.ID1 > A1.ID2

Positives

- `SELECT COUNT(*) FROM (SELECT DISTINCT * FROM MUSTER.GRUPPE1)`

True positives

- ```
  SELECT COUNT(*) FROM
       (( (SELECT ID1, ID2 FROM muster.Gruppe1) INTERSECT
                 (SELECT ID1, ID2 FROM muster.Musterl)) UNION
  ( (SELECT ID2, ID1 FROM muster.Gruppe1) INTERSECT
                 (SELECT ID1, ID2 FROM muster.Musterl))
  AS TP
  ```

False positives

- ```
  SELECT COUNT(*) FROM
       ( SELECT * FROM Gruppe1 EXCEPT (SELECT ID1, ID2 FROM Musterl)) AS FP
  ```

False negatives

- ```
  SELECT COUNT(*) FROM
       ( SELECT * FROM MUSTERVIEW EXCEPT (SELECT * FROM Gruppe1)) AS FN
  ```
  ```

SQL für das kombinierte Team

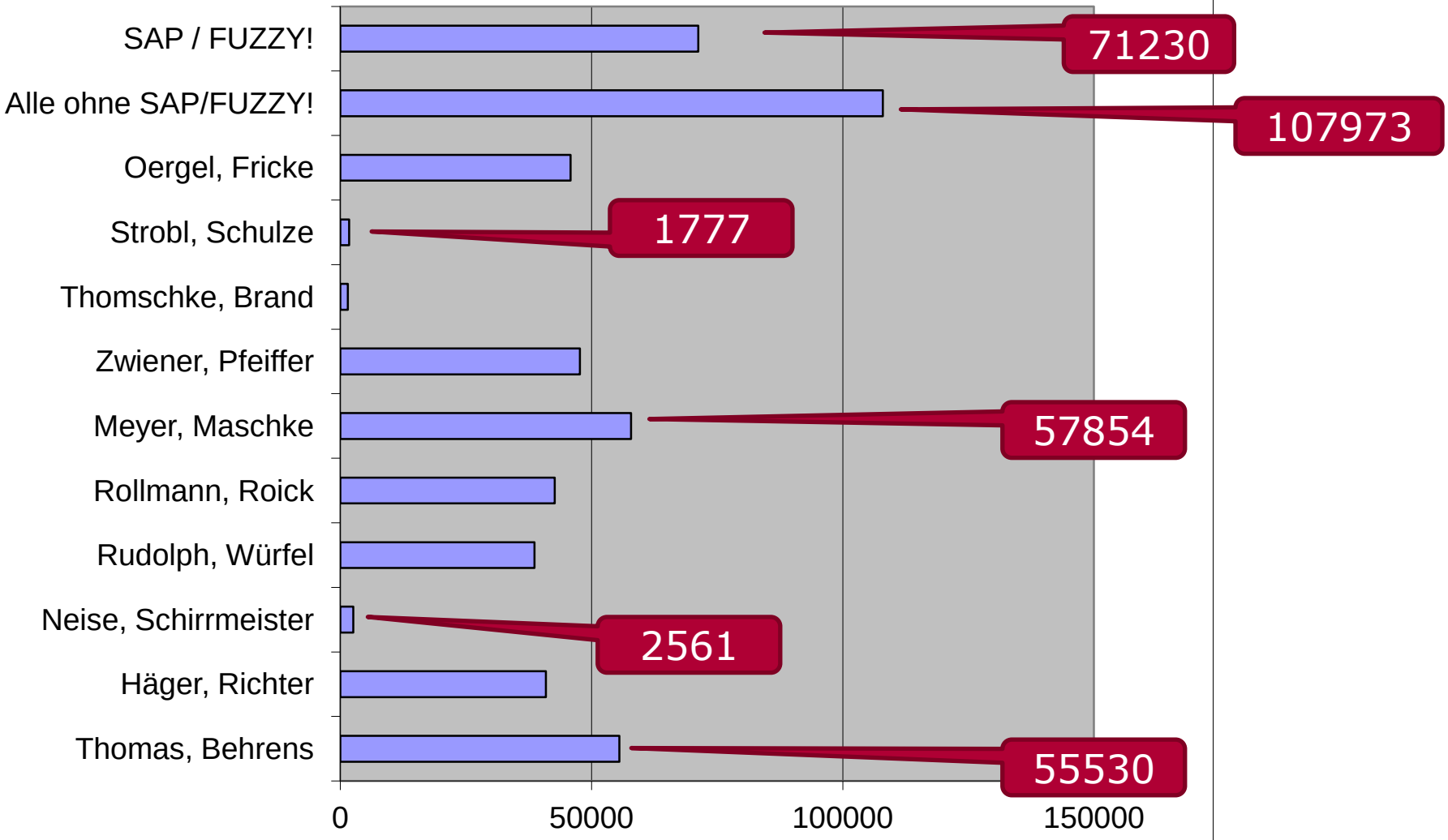

```
SELECT COUNT(*) FROM (SELECT DISTINCT * FROM
(SELECT * FROM MUSTER.GRUPPE1) UNION
(SELECT * FROM MUSTER.GRUPPE2) UNION
(SELECT * FROM MUSTER.GRUPPE3) UNION
(SELECT * FROM MUSTER.GRUPPE4) UNION
(SELECT * FROM MUSTER.GRUPPE5) UNION
(SELECT * FROM MUSTER.GRUPPE7) UNION
(SELECT * FROM MUSTER.GRUPPE8) UNION
(SELECT * FROM MUSTER.GRUPPE10) UNION
(SELECT * FROM MUSTER.GRUPPE9) UNION
(SELECT * FROM MUSTER.GRUPPE6)
)

SELECT COUNT(*) FROM (
 (SELECT ID1, ID2 FROM Gruppe1)
INTERSECT
 (SELECT ID1, ID2 FROM Musterl)
)
AS TP
```

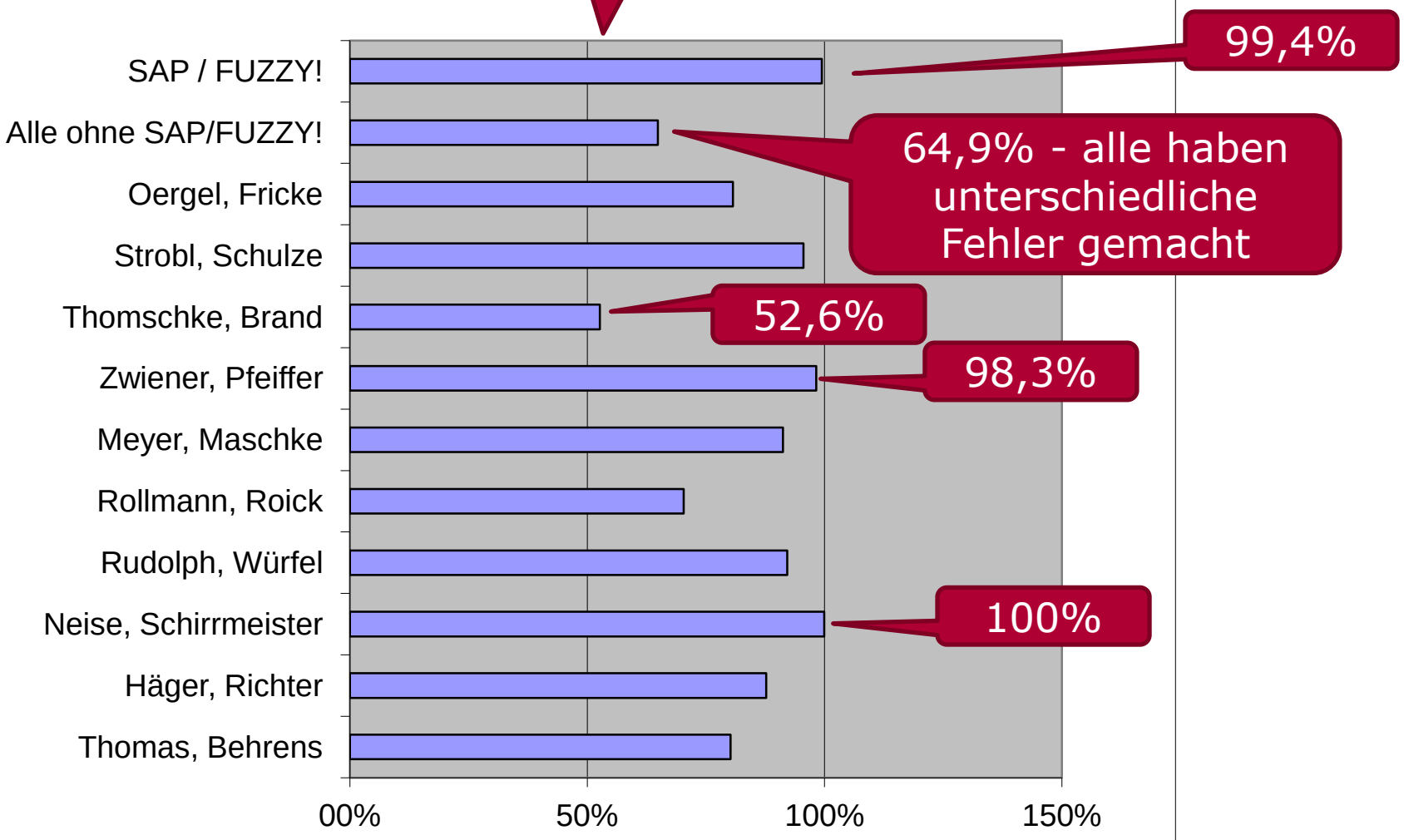
```
WITH POSITIVES AS (SELECT DISTINCT * FROM
(SELECT * FROM GRUPPE1) UNION
(SELECT * FROM GRUPPE2) UNION
(SELECT * FROM GRUPPE3) UNION
(SELECT * FROM GRUPPE5) UNION
(SELECT * FROM GRUPPE6) )

SELECT COUNT(*) FROM (
(SELECT * FROM POSITIVES)
   INTERSECT
(SELECT ID1, ID2 FROM Musterl)
)
```

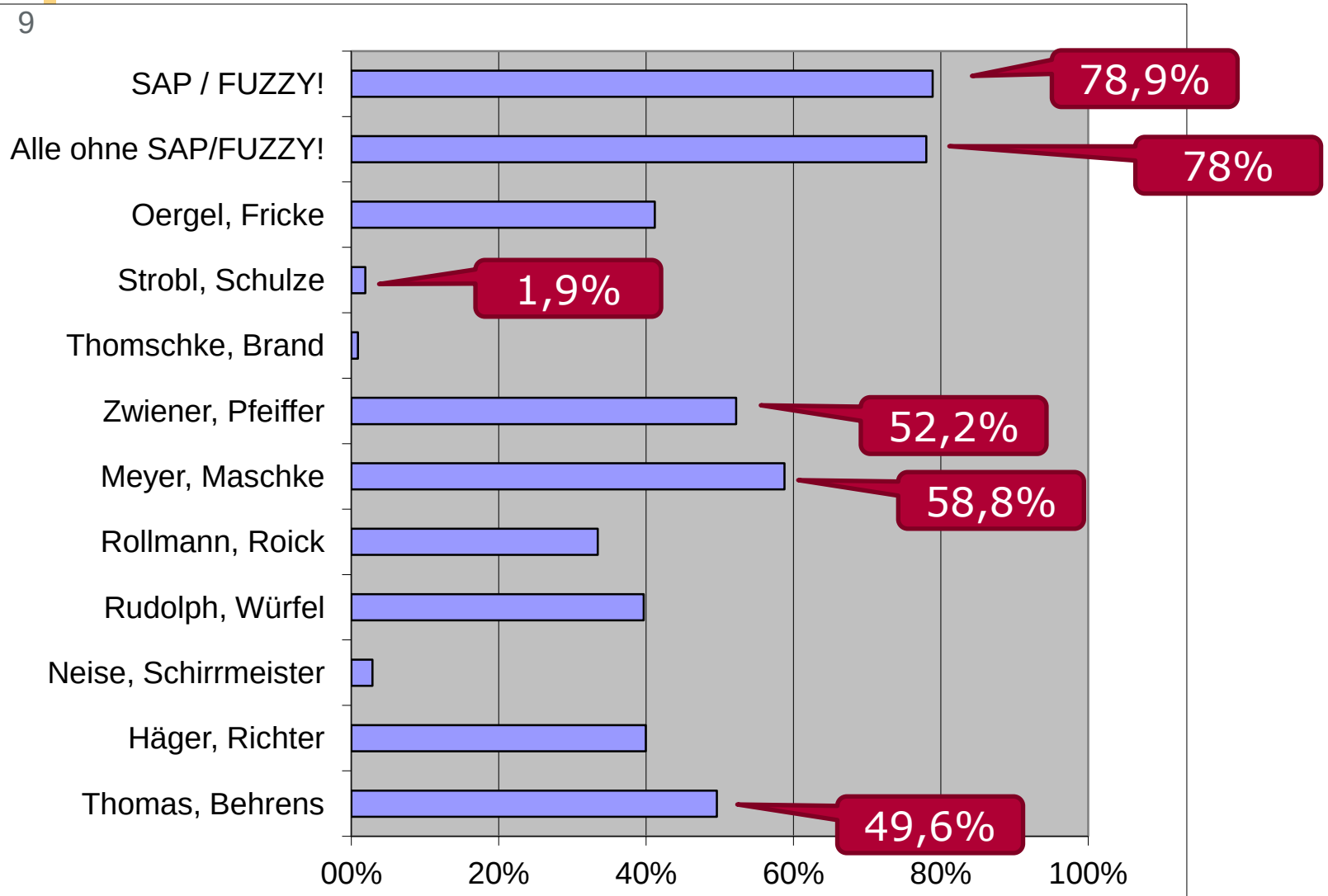
Positives 2010 (Musterlösung: 89783)



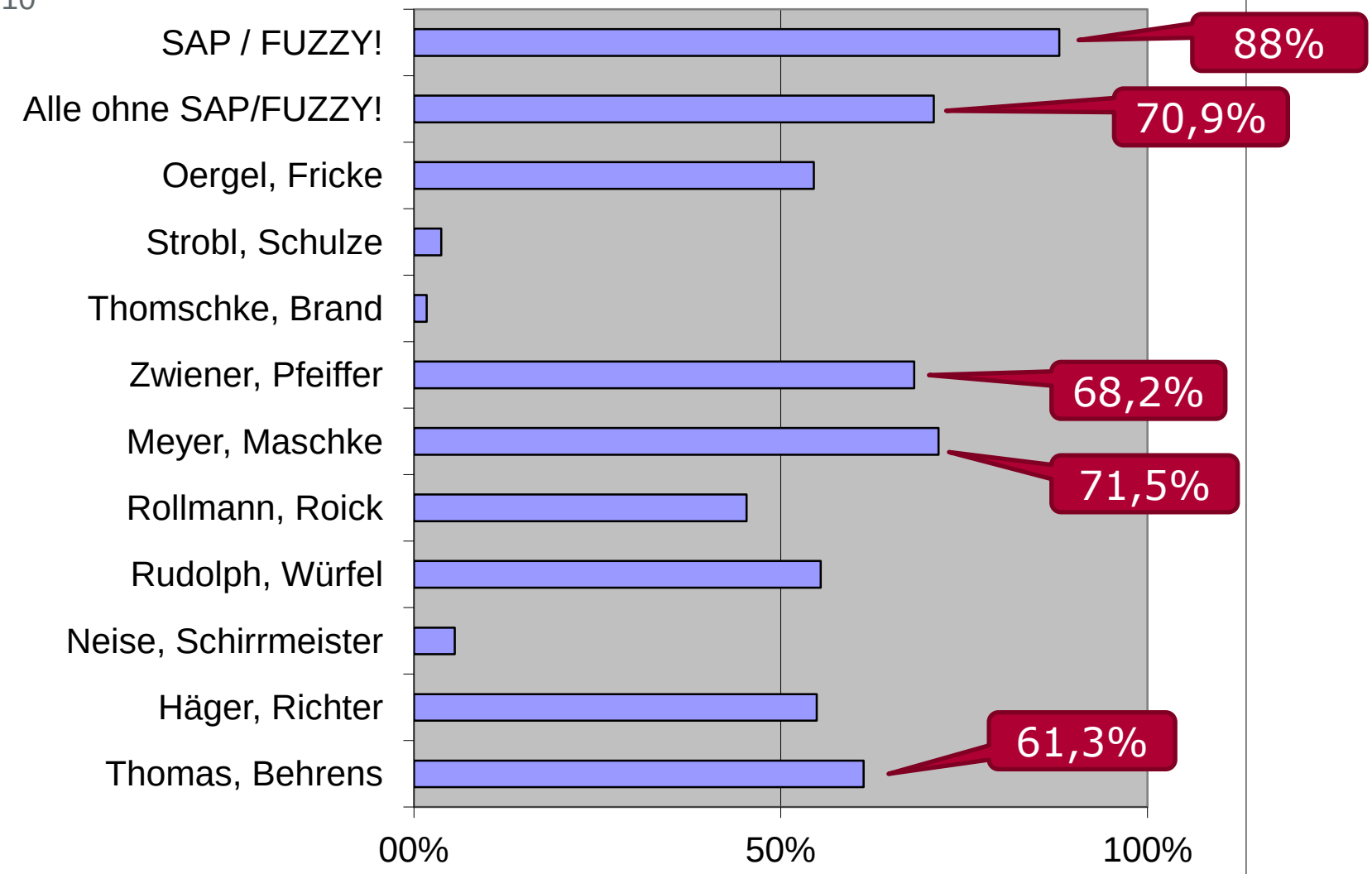
Precision 2010



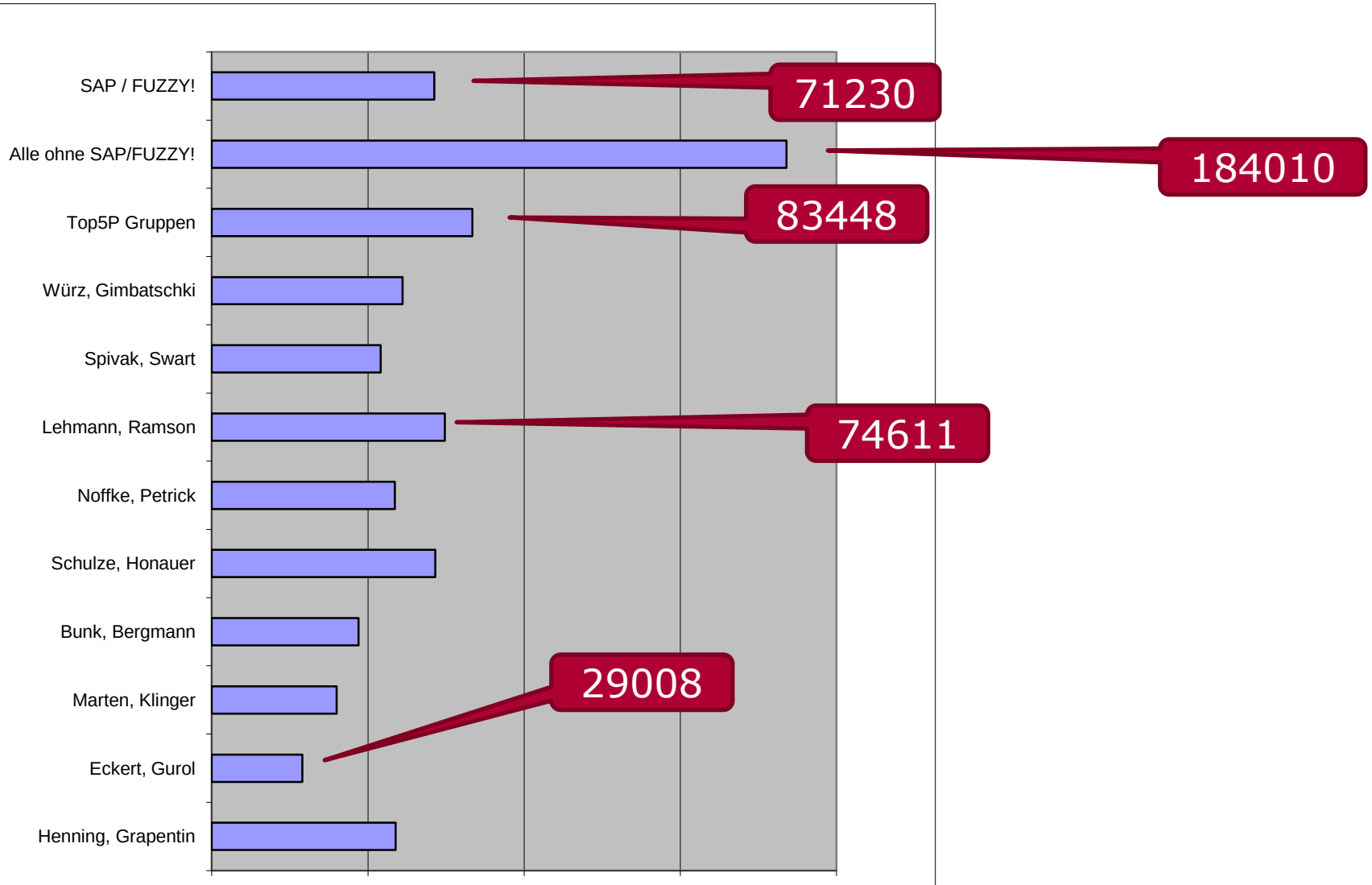
Recall 2010


SAP / FUZZY!
Alle ohne SAP/FUZZY!
Oergel, Fricke
Strobl, Schulze
Thomschke, Brand
Zwiener, Pfeiffer
Meyer, Maschke
Rollmann, Roick
Rudolph, Würfel
Neise, Schirrmeister
Häger, Richter
Thomas, Behrens
78,9%
78%
1,9%
52,2%
58,8%
49,6%
00%
20%
40%
60%
80%
100%

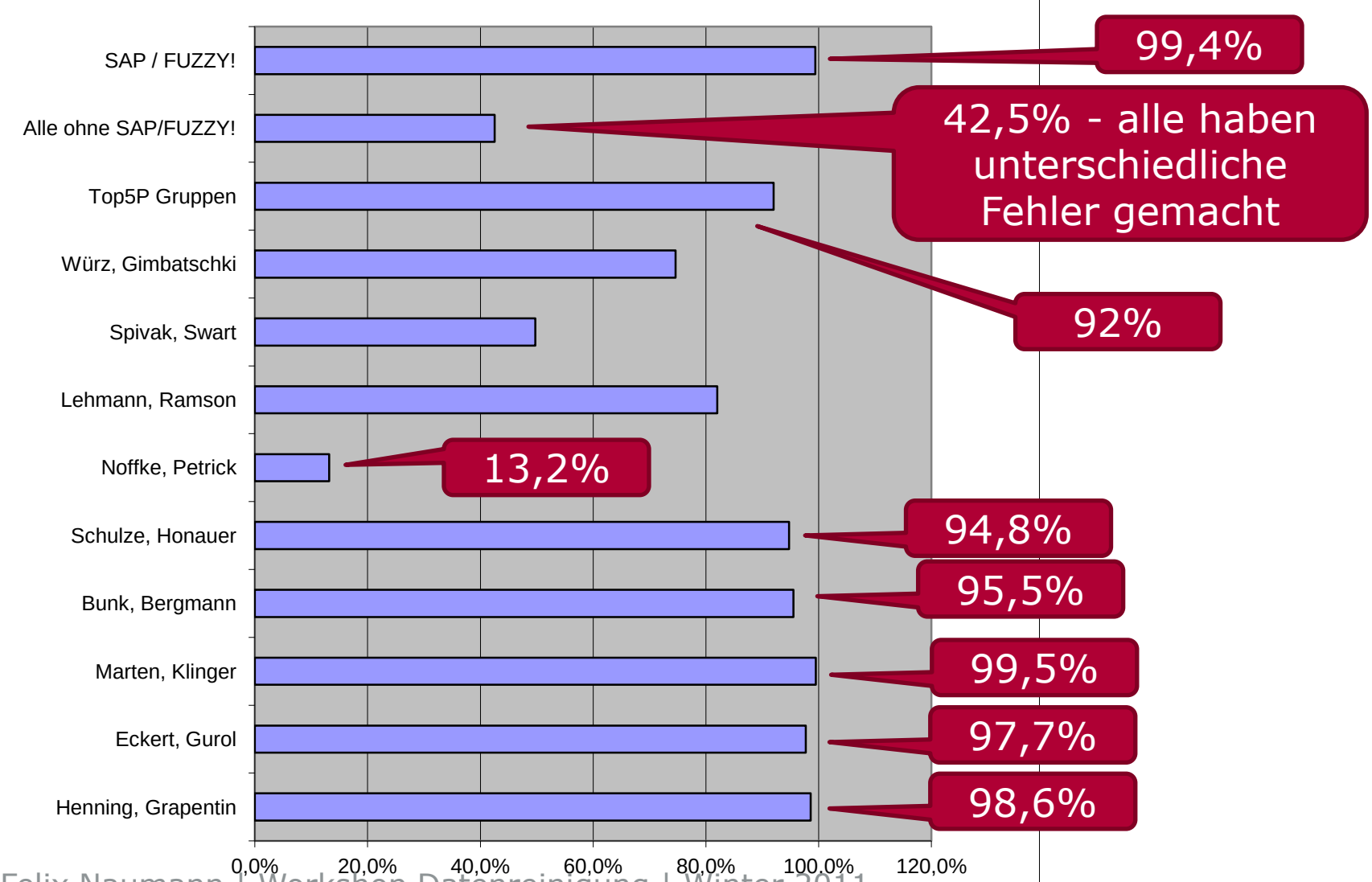
F-Measure 2010



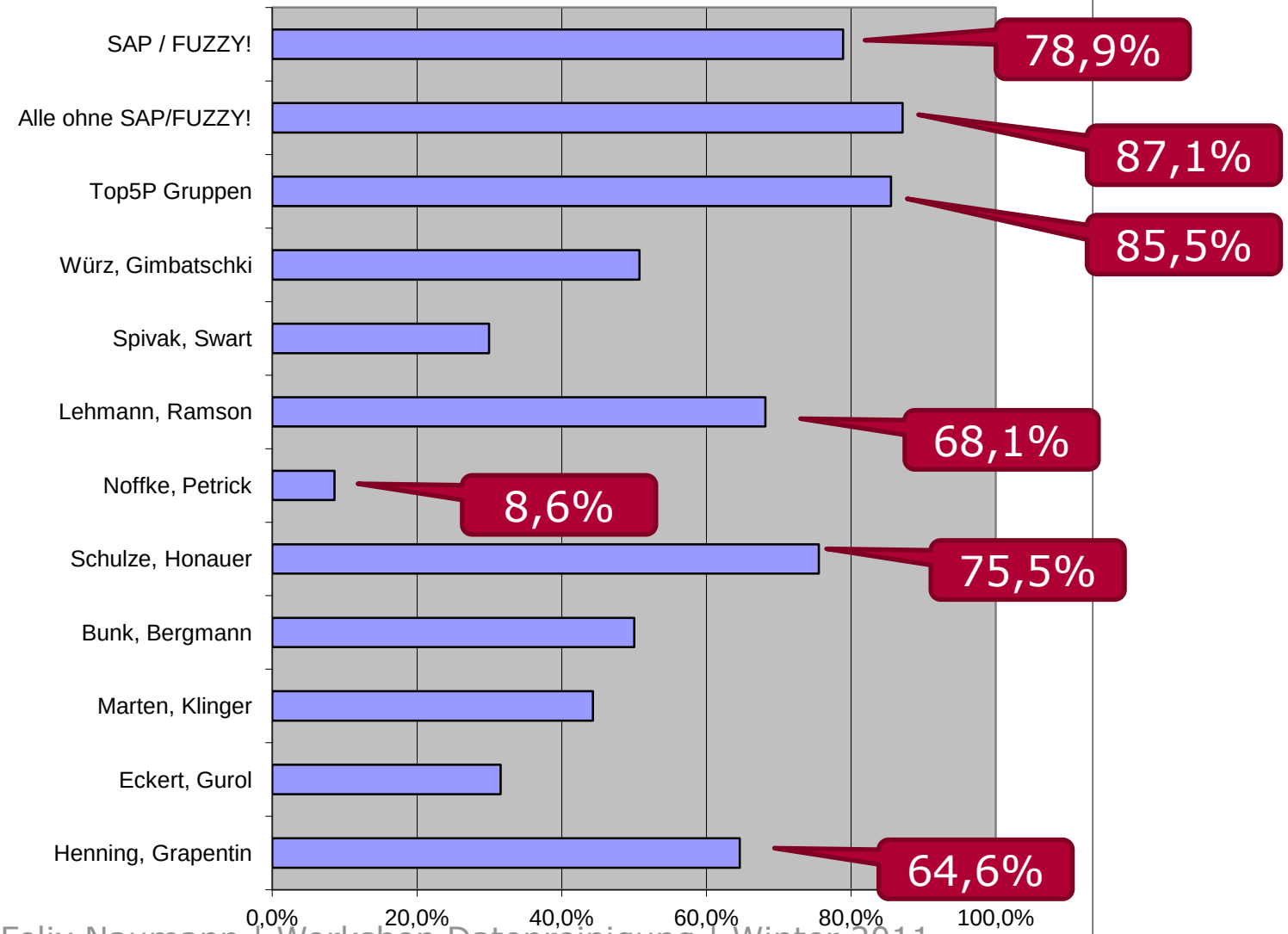
Positives 2011 (Musterlösung: 89783)



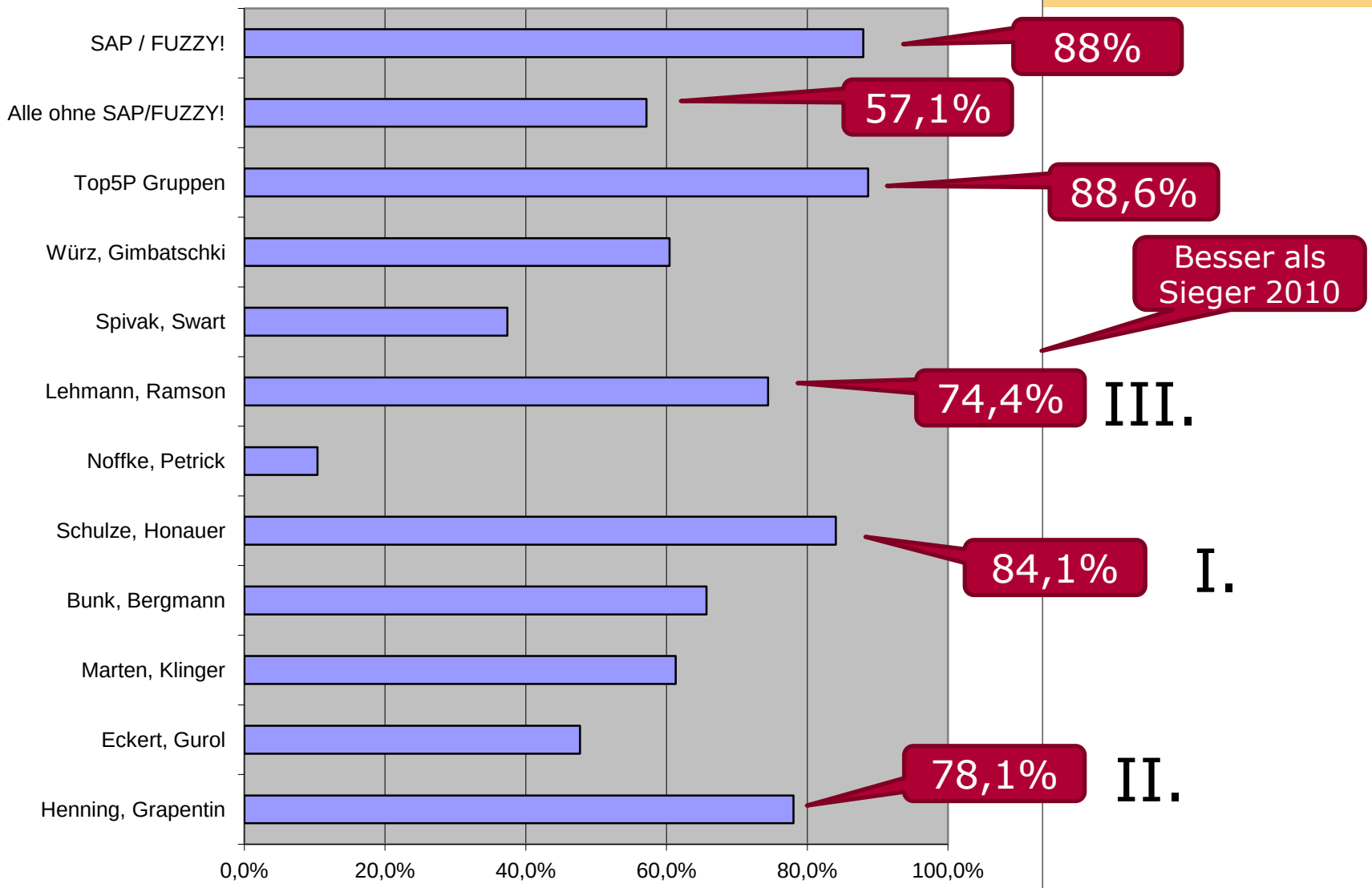
Precision 2011



Recall 2011


SAP / FUZZY!
Alle ohne SAP/FUZZY!
Top5P Gruppen
Würz, Gimbatschki
Spivak, Swart
Lehmann, Ramson
Noffke, Petrick
Schulze, Honauer
Bunk, Bergmann
Marten, Klinger
Eckert, Gurol
Henning, Grapentin
78,9%
87,1%
85,5%
68,1%
8,6%
75,5%
64,6%
0,0%
20,0%
40,0%
60,0%
80,0%
100,0%

F-Measure 2011



Anmerkungen

- Positives
  - Alle haben viel abgegeben
  - Effizienz diesmal kein Problem: typischer Durchlauf < 5min
- Precision
  - Meist erstaunlich gut
  - Einmal 99,5% bei 39976 Abgaben
- Recall
  - Es fehlten teilweise Ideen und neue Ansätze
  - Mehr am Sortierschlüssel variieren