HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Bad Files, Bad Data, Bad Results: Data Quality and Data Preparation

CAiSE 2021 – keynote

July 2, 2021

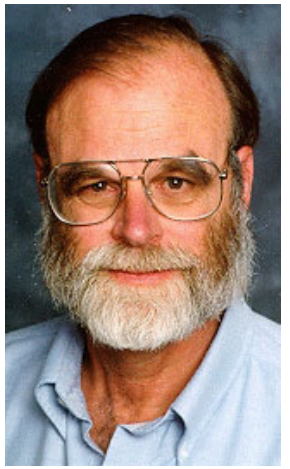
Felix Naumann

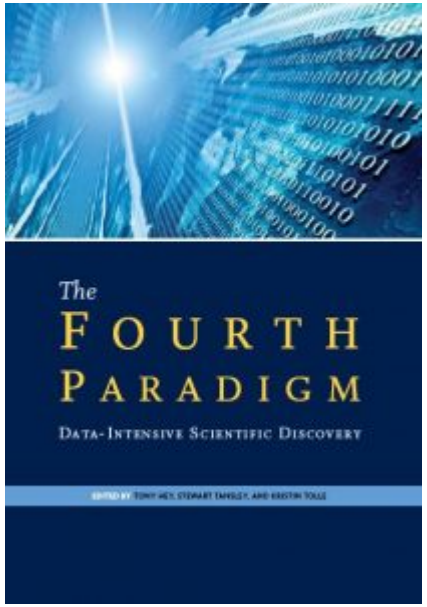
https://www.flickr.com/photos/136315829@N03/27640684175

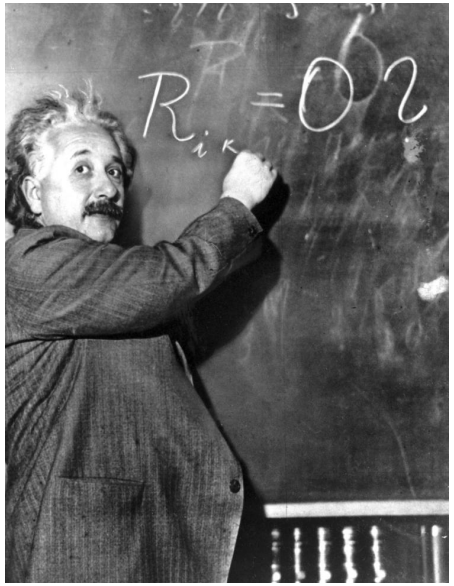
The Fourth Paradigm of Science

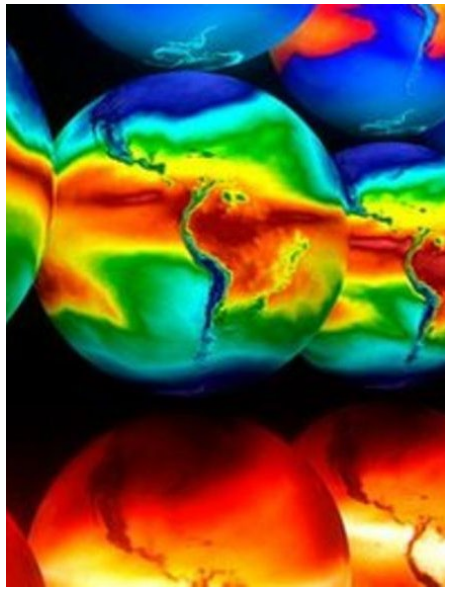
1. Empirical and experimental
2. Theoretical
3. Computational
4. Data-intensive

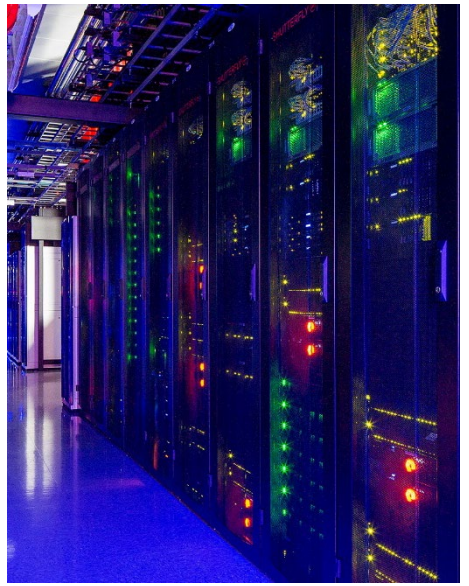
We have to do better producing tools to support the whole research cycle - from data capture and data curation to data analysis and data visualization. Jim Gray







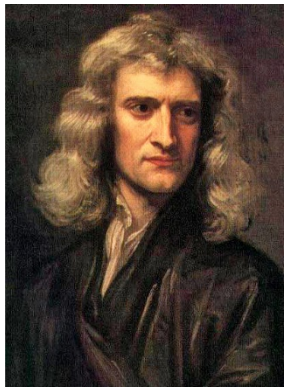




Big Data = Science?

- The End of Theory: The Data Deluge Makes the Scientific Method Obsolete (Chris Anderson, Wired, 2008)
  - *All models are wrong, but some are useful.* (George Box)
  - *All models are wrong, and increasingly you can succeed without them.* (Peter Norvig)
- Before Big Data: Correlation is not causation!
- With Big Data: Who cares?
  - Petabytes allow us to say: "Correlation is enough."

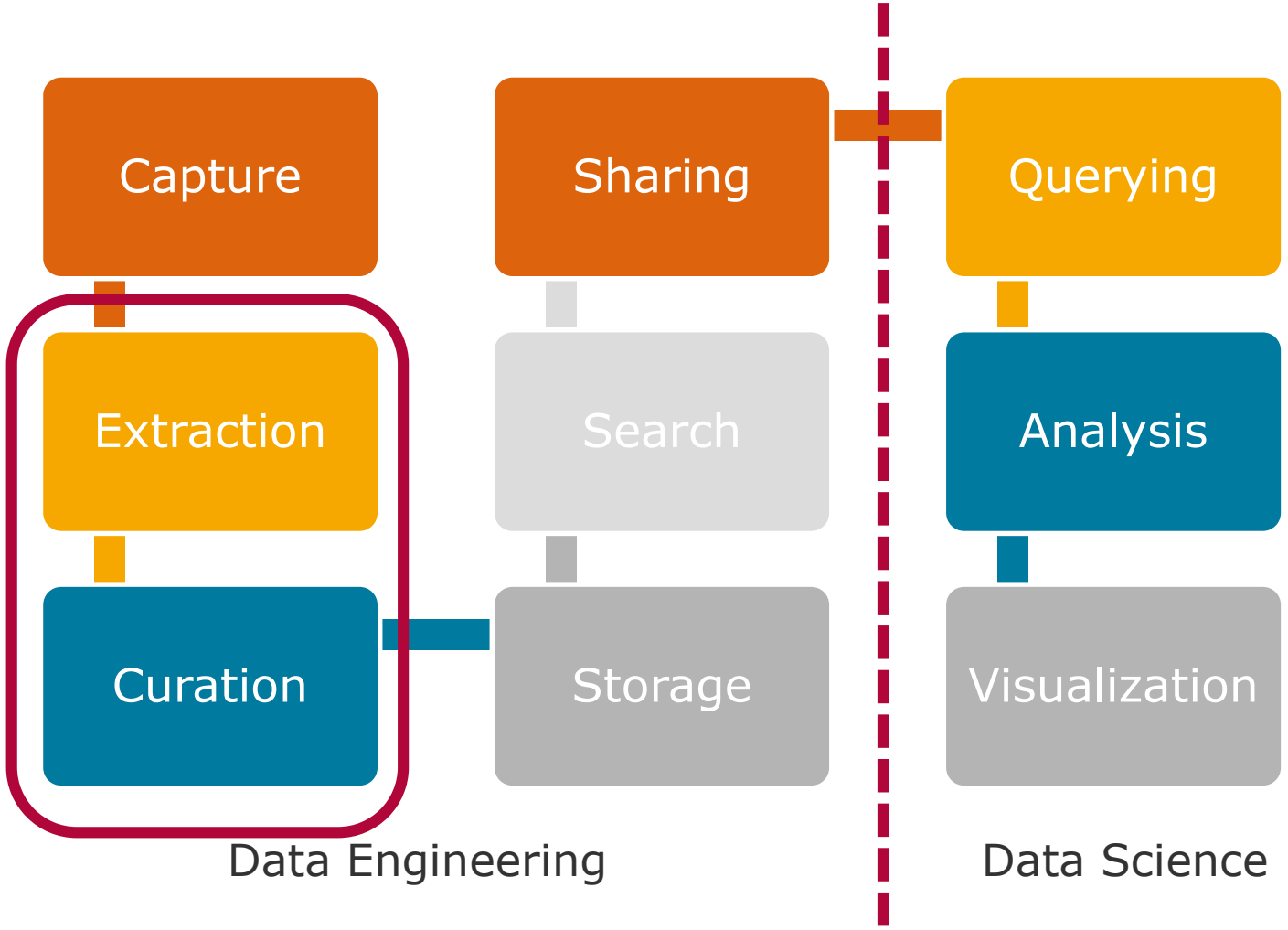
$$F = G\frac{m_1 m_2}{r^2}$$



vs.

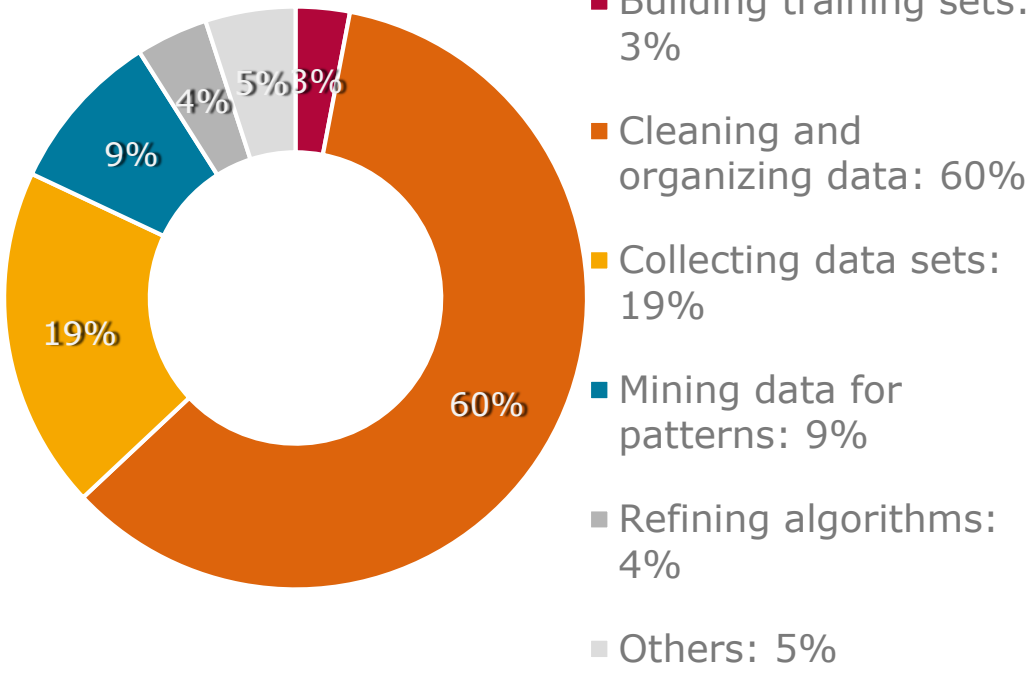
Felix Naumann
Bad Files, Bad Data 2021

Data Science Pipeline

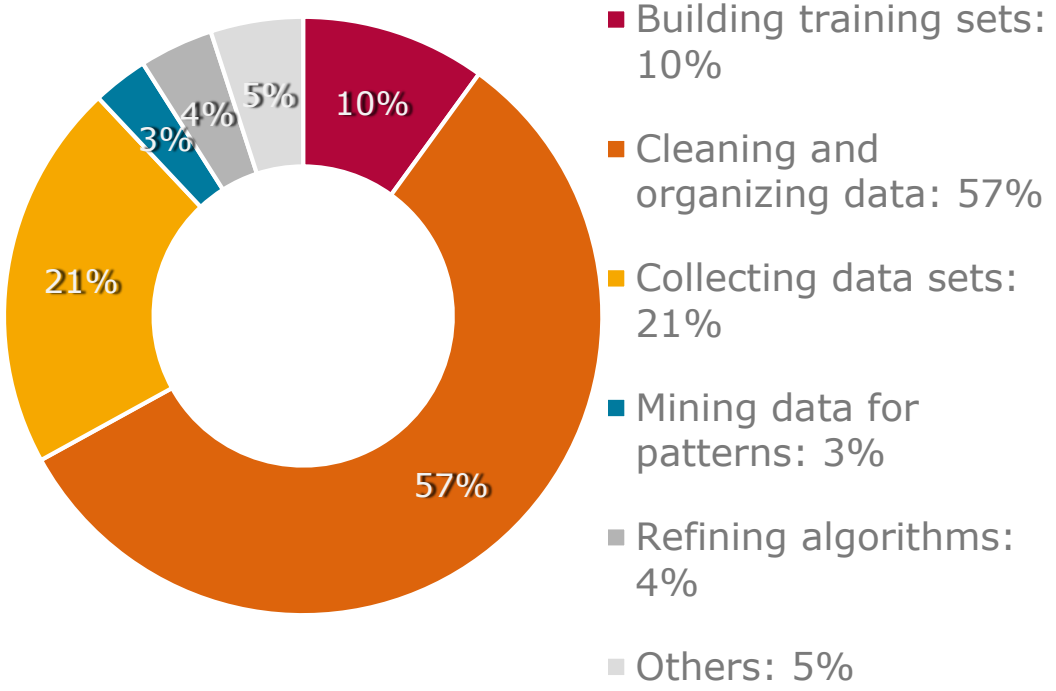


Data preparation in reality

What data scientists spend the most time doing?



What is the least enjoyable part of data science?



"Cleaning Data: Most Time-Consuming, Least Enjoyable Data Science Task", Gil Press, Forbes, March 23rd, 2016
http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

Overview

1. **Bad Files**
2. Data Preparation
3. Bad Data
4. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

```
"ID,""Name"",""Sex"",""Age"",""Height"",""Weight"",""Team"",""NOC"",""Games"",""Year"",""Season"",""City"",""Sport"",""Event"",""Medal"""
"1,""A Dijiang"",""M"",24,180,80,""China"",""CHN"",""1992 Summer"",1992,""Summer"",""Barcelona"",""Basketball"",""Basketball Men's Basketball"",NA"
"2,""A Lamusi"",""M"",23,170,60,""China"",""CHN"",""2012 Summer"",2012,""Summer"",""London"",""Judo"",""Judo Men's Extra-Lightweight"",NA"
"3,""Gunnar Nielsen Aaby"",""M"",24,NA,NA,""Denmark"",""DEN"",""1920 Summer"",1920,""Summer"",""Antwerpen"",""Football"",""Football Men's Football"",NA"
"4,""Edgar Lindenau Aabye"",""M"",34,NA,NA,""Denmark/Sweden"",""DEN"",""1900 Summer"",1900,""Summer"",""Paris"",""Tug-Of-War"",""Tug-Of-War Men's Tug-Of-War"",""Gold"""
"5,""Christine Jacoba Aaftink"",""F"",21,185,82,""Netherlands"",""NED"",""1988 Winter"",1988,""Winter"",""Calgary"",""Speed Skating"",""Speed Skating Women's 500 metres"",NA"
"5,""Christine Jacoba Aaftink"",""F"",21,185,82,""Netherlands"",""NED"",""1988 Winter"",1988,""Winter"",""Calgary"",""Speed Skating"",""Speed Skating Women's 1,000 metres"",NA"
"5,""Christine Jacoba Aaftink"",""F"",25,185,82,""Netherlands"",""NED"",""1992 Winter"",1992,""Winter"",""Albertville"",""Speed Skating"",""Speed Skating Women's 500 metres"",NA"
"5,""Christine Jacoba Aaftink"",""F"",25,185,82,""Netherlands"",""NED"",""1992 Winter"",1992,""Winter"",""Albertville"",""Speed Skating"",""Speed Skating Women's 1,000 metres"",NA"
"5,""Christine Jacoba Aaftink"",""F"",27,185,82,""Netherlands"",""NED"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Speed Skating"",""Speed Skating Women's 500 metres"",NA"
"5,""Christine Jacoba Aaftink"",""F"",27,185,82,""Netherlands"",""NED"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Speed Skating"",""Speed Skating Women's 1,000 metres"",NA"
"6,""Per Knut Aaland"",""M"",31,188,75,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 10 kilometres"",NA"
"6,""Per Knut Aaland"",""M"",31,188,75,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 50 kilometres"",NA"
"6,""Per Knut Aaland"",""M"",31,188,75,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 10/15 kilometres Pursuit"",NA"
"6,""Per Knut Aaland"",""M"",31,188,75,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 4 x 10 kilometres Relay"",NA"
"6,""Per Knut Aaland"",""M"",33,188,75,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 10 kilometres"",NA"
"6,""Per Knut Aaland"",""M"",33,188,75,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 30 kilometres"",NA"
"6,""Per Knut Aaland"",""M"",33,188,75,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 10/15 kilometres Pursuit"",NA"
"6,""Per Knut Aaland"",""M"",33,188,75,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 4 x 10 kilometres Relay"",NA"
"7,""John Aalberg"",""M"",31,183,72,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 10 kilometres"",NA"
"7,""John Aalberg"",""M"",31,183,72,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 50 kilometres"",NA"
"7,""John Aalberg"",""M"",31,183,72,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 10/15 kilometres Pursu
"7,""John Aalberg"",""M"",31,183,72,""United States"",""USA"",""1992 Winter"",1992,""Winter"",""Albertville"",""Cross Country Skiing"",""Cross Country Skiing Men's 4 x 10 kilometres Relay
"7,""John Aalberg"",""M"",33,183,72,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 10 kilometres"",NA"
"7,""John Aalberg"",""M"",33,183,72,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 30 kilometres"",NA"
"7,""John Aalberg"",""M"",33,183,72,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 10/15 kilometres Pursu
"7,""John Aalberg"",""M"",33,183,72,""United States"",""USA"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Cross Country Skiing"",""Cross Country Skiing Men's 4 x 10 kilometres Relay
"8,""Cornelia """"Cor"""" Aalten (-Strannood)"",""F"",18,168,NA,""Netherlands"",""NED"",""1932 Summer"",1932,""Summer"",""Los Angeles"",""Athletics"",""Athletics Women's 100 metres"",NA"
"8,""Cornelia """"Cor"""" Aalten (-Strannood)"",""F"",18,168,NA,""Netherlands"",""NED"",""1932 Summer"",1932,""Summer"",""Los Angeles"",""Athletics"",""Athletics Women's 4 x 100 metres Relay"",NA"
"9,""Antti Sami Aalto"",""M"",26,186,96,""Finland"",""FIN"",""2002 Winter"",2002,""Winter"",""Salt Lake City"",""Ice Hockey"",""Ice Hockey Men's Ice Hockey"",NA"
"10,""Einar Ferdinand """"Einari"""" Aalto"",""M"",26,NA,NA,""Finland"",""FIN"",""1952 Summer"",1952,""Summer"",""Helsinki"",""Swimming"",""Swimming Men's 400 metres Freestyle"",NA"
"11,""Jorma Ilmari Aalto"",""M"",22,182,76.5,""Finland"",""FIN"",""1980 Winter"",1980,""Winter"",""Lake Placid"",""Cross Country Skiing"",""Cross Country Skiing Men's 30 kilometres"",NA"
"12,""Jyri Tapani Aalto"",""M"",31,172,70,""Finland"",""FIN"",""2000 Summer"",2000,""Summer"",""Sydney"",""Badminton"",""Badminton Men's Singles"",NA"
"13,""Minna Maarit Aalto"",""F"",30,159,55.5,""Finland"",""FIN"",""1996 Summer"",1996,""Summer"",""Atlanta"",""Sailing"",""Sailing Women's Windsurfer"",NA"
"13,""Minna Maarit Aalto"",""F"",34,159,55.5,""Finland"",""FIN"",""2000 Summer"",2000,""Summer"",""Sydney"",""Sailing"",""Sailing Women's Windsurfer"",NA"
"14,""Pirjo Hannele Aalto (Mattila-)"",""F"",32,171,65,""Finland"",""FIN"",""1994 Winter"",1994,""Winter"",""Lillehammer"",""Biathlon"",""Biathlon Women's 7.5 kilometres Sprint"",NA"
"15,""Arvo Ossian Aaltonen"",""M"",22,NA,NA,""Finland"",""FIN"",""1912 Summer"",1912,""Summer"",""Stockholm"",""Swimming"",""Swimming Men's 200 metres Breaststroke"",NA"
"15,""Arvo Ossian Aaltonen"",""M"",22,NA,NA,""Finland"",""FIN"",""1912 Summer"",1912,""Summer"",""Stockholm"",""Swimming"",""Swimming Men's 400 metres Breaststroke"",NA"
"15,""Arvo Ossian Aaltonen"",""M"",30,NA,NA,""Finland"",""FIN"",""1920 Summer"",1920,""Summer"",""Antwerpen"",""Swimming"",""Swimming Men's 200 metres Breaststroke"",""Bronze"""
"15,""Arvo Ossian Aaltonen"",""M"",30,NA,NA,""Finland"",""FIN"",""1920 Summer"",1920,""Summer"",""Antwerpen"",""Swimming"",""Swimming Men's 400 metres Breaststroke"",""Bronze"""
"15,""Arvo Ossian Aaltonen"",""M"",34,NA,NA,""Finland"",""FIN"",""1924 Summer"",1924,""Summer"",""Paris"",""Swimming"",""Swimming Men's 200 metres Breaststroke"",NA"
"16,""Juhamatti Tapio Aaltonen"",""M"",28,184,85,""Finland"",""FIN"",""2014 Winter"",2014,""Winter"",""Sochi"",""Ice Hockey"",""Ice Hockey Men's Ice Hockey"",""Bronze"""
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Individual All-Around"",""Bronze"""
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Team All-Around"",""Gold"""
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Floor Exercise"",NA"
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Horse Vault"",""Gold"""
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Parallel Bars"",NA"
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Horizontal Bar"",NA"
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Rings"",NA"
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1948 Summer"",1948,""Summer"",""London"",""Gymnastics"",""Gymnastics Men's Pommelled Horse"",""Gold"""
```

1	Table rv.03.q: Removals and voluntary departures by country of nationality and type															
2	Quarter	Geographical region	Country of nationality	Total enforced removals	Total Refused entry at port and subsequently departed	Total voluntary departures	Total non-asylum enforced removals	Non-asylum cases: Refused entry at port and subsequently departed	Total non-asylum voluntary departures	Non-asylum cases: Assisted Voluntary Returns	Non-asylum cases: Notified voluntary departures	Non-asylum cases: Other confirmed voluntary departures	Total asylum enforced removals	Total asylum voluntary departures	Asylum cases: Assisted Voluntary Returns	As ca No volu depar
237	2011 Q1	Europe	Turkey	48	39	79	22	39	74	0	23	51	26	5	0	
238	2011 Q1	Europe	Turkmenistan	2	3	10	1	3	9	0	0	9	1	1	0	
239	2011 Q1	Americas	Turks and Caicos Islands (British)	0	0	0	0	0	0	0	0	0	0	0	0	
240	2011 Q1	Oceania	Tuvalu	0	0	0	0	0	0	0	0	0	0	0	0	
241	2011 Q1	Africa	Uganda	24	3	58	7	3	50	4	8	38	17	8	7	
242	2011 Q1	Europe	Ukraine	53	63	39	46	63	36	0	12	24	7	3	2	
243	2011 Q1	Middle East	United Arab Emirates	0	1	5	0	1	5	0	0	5	0	0	0	
244	2011 Q1	Americas	United States	14	472	91	14	472	88	0	35	53	0	3	3	
245	2011 Q1	Americas	Uruguay	1	3	0	0	3	0	0	0	0	1	0	0	
246	2011 Q1	Europe	Uzbekistan	22	1	46	20	1	46	0	8	38	2	0	0	
247	2011 Q1	Oceania	Vanuatu	0	0	0	0	0	0	0	0	0	0	0	0	
248	2011 Q1	Europe	Vatican City	0	0	0	0	0	0	0	0	0	0	0	0	
249	2011 Q1	Americas	Venezuela	2	45	5	0	45	5	0	3	2	2	0	0	
250	2011 Q1	Asia	Vietnam	249	24	77	190	24	72	2	10	60	59	5	3	
251	2011 Q1	Americas	Virgin Islands (British)	0	0	0	0	0	0	0	0	0	0	0	0	
252	2011 Q1	Americas	Virgin Islands (US)	0	0	0	0	0	0	0	0	0	0	0	0	
253	2011 Q1	Oceania	Wallis and Futuna	0	0	0	0	0	0	0	0	0	0	0	0	
254	2011 Q1	Africa	Western Sahara	2	0	0	2	0	0	0	0	0	0	0	0	
255	2011 Q1	Middle East	Yemen	2	0	2	2	0	2	0	2	0	0	0	0	
256	2011 Q1	Africa	Zambia	3	3	27	3	3	26	5	6	15	0	1	1	
257	2011 Q1	Africa	Zimbabwe	7	3											
258	2011 Q2	Total	*Total	3.456	3.963	5										
259	2011 Q2	Africa	*Total Africa	703	611											
260	2011 Q2	Americas	*Total Americas	343	1.367											
261	2011 Q2	Asia	*Total Asia	1.790	888	2										
262	2011 Q2	Europe	*Total Europe	512	638											
263	2011 Q2	Middle East	*Total Middle East	98	192											
264	2011 Q2	Oceania	*Total Oceania	4	153											
265	2011 Q2	Other	*Total Other	6	114											
266	2011 Q2	Asia	Afghanistan	296	70											
267	2011 Q2	Europe	Albania	100	187											
268	2011 Q2	Africa	Algeria	49	32											
269	2011 Q2	Oceania	American Samoa	0	0											
270	2011 Q2	Europe	Andorra	0	0											
271	2011 Q2	Africa	Angola	7	19											
272	2011 Q2	Americas	Anguilla (British)	0	0											
273	2011 Q2	Americas	Antigua and Barbuda	0	6											
274	2011 Q2	Americas	Argentina	4	30											
275	2011 Q2	Europe	Armenia	2	3											
276	2011 Q2	Americas	Aruba	0	0											
277	2011 Q2	Oceania	Australia	1	120											
278	2011 Q2	Europe	Austria	2	1	0	2	1	0	0	0	0	0	0	0	

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Research Note

Open data: Quality over quantity

Shazia Sadiq [a], Marta Indulska [b,*]

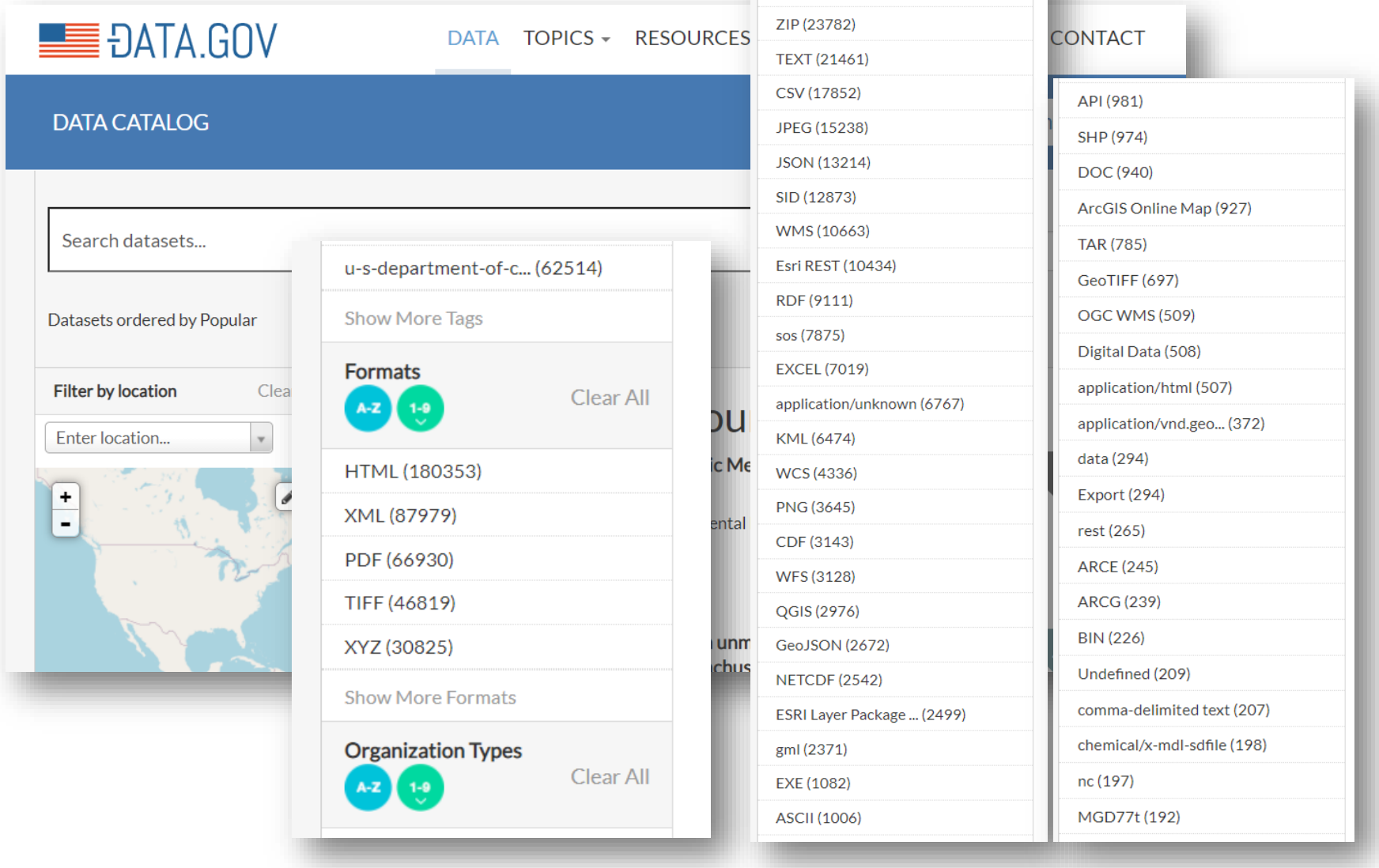
[a] School of Information Technology and Electrical Engineering, The University of Queensland, St Lucia, QLD 4072, Australia
[b] UQ Business School, The University of Queensland, St Lucia, QLD 4072, Australia

```
Nov-09,,4,47,35,17,99,32,1055,185,578,16,0,18,16,2,36,5,149,2,47,0,0,16,11,5,32,10,43,5,115,1
Dec-09,,3,41,32,15,89,27,930,145,566,14,0,17,17,2,36,4,131,2,49,0,0,12,10,5,27,8,40,6,106,1
Jan-10,*,3,51,41,17,109,33,799,143,654,19,0,20,18,2,39,5,125,2,52,0,0,14,13,6,33,8,35,5,138,1
Feb-10,,3,46,36,14,96,32,636,133,545,17,0,19,15,1,35,4,97,1,44,0,0,13,12,6,31,8,24,4,113,1
Mar-10,,4,48,36,15,99,29,700,126,550,17,0,19,15,2,36,4,100,2,44,0,0,13,11,6,30,6,19,4,113,1
Apr-10,*,4,57,42,19,119,33,792,157,665,20,0,24,17,3,44,4,115,2,52,0,0,17,15,8,39,7,21,5,141,1
May-10,,3,46,34,18,99,27,629,127,535,16,0,19,13,3,36,4,45,1,42,0,0,12,10,6,28,6,27,5,118,1
Jun-10,,3,43,33,20,97,26,682,132,531,14,0,18,13,5,36,4,55,1,39,0,0,11,10,8,29,6,27,5,115,1
Jul-10,*,5,55,40,26,121,36,1075,182,662,Data are confidential,0,21,16,6,43,5,114,2,51,0,0,11,10,10,31,8,35,5,144,1
Aug-10,,5,43,32,20,95,28,987,165,553,Data are confidential,0,17,11,5,34,4,135,2,46,0,0,10,8,6,24,7,24,5,121,1
Sep-10,,7,48,34,18,100,33,957,158,562,Data are confidential,0,19,13,4,36,5,148,2,46,0,0,16,10,5,31,7,27,5,121,1
Oct-10,*,9,63,44,22,129,49,1191,195,728,Data are confidential,0,24,19,4,47,6,197,3,57,0,1,22,13,6,41,10,29,7,157,1
Nov-10,,7,52,40,18,109,47,1047,183,605,Data are confidential,0,19,16,3,38,6,154,2,47,0,0,14,11,5,29,10,20,4,132,1
Dec-10,**,6,55,42,18,114,41,1065,189,691,Data are confidential,0,21,20,3,43,5,167,3,54,0,0,14,11,6,31,8,20,4,143,1
Jan-11,*,6,60,48,18,126,52,856,190,690,Data are confidential,0,22,20,3,45,6,148,2,52,0,1,16,15,7,38,10,19,4,157,1
Feb-11,,7,47,39,15,101,37,699,156,592,Data are confidential,0,19,16,2,37,4,115,2,48,0,0,14,12,5,32,8,13,2,123,1
Mar-11,,8,51,38,16,105,34,678,137,587,Data are confidential,0,20,16,2,37,4,115,2,49,0,0,13,11,5,29,6,12,2,122,1
Apr-11,*,7,62,46,19,127,37,827,167,683,Data are confidential,0,23,18,4,45,5,118,2,60,0,0,15,12,5,32,7,15,3,143,0
May-11,,5,49,37,19,106,35,655,132,545,Data are confidential,0,19,14,4,36,5,49,2,45,0,0,11,10,6,27,7,17,3,122,0
Jun-11,,5,46,36,21,103,36,749,137,567,Data are confidential,0,17,13,5,35,5,72,2,45,0,0,10,8,6,25,8,21,2,127,0
Jul-11,*,6,56,42,25,123,42,1133,189,728,Data are confidential,0,20,16,6,42,6,137,3,55,0,0,10,8,5,23,9,28,4,151,0
Aug-11,,5,45,34,18,97,34,956,153,594,Data are confidential,0,18,12,4,34,5,133,3,43,0,0,14,8,4,26,7,25,4,121,0
Sep-11,,7,51,36,17,104,40,992,153,621,Data are confidential,0,18,14,2,35,5,144,3,49,0,1,17,9,4,30,8,30,4,127,0
Oct-11,*,8,61,45,18,125,53,1336,216,768,Data are confidential,0,22,20,2,45,8,191,3,68,0,1,20,11,5,36,12,34,5,159,0
Nov-11,,6,50,39,15,105,48,964,165,639,Data are confidential,0,18,16,2,36,6,147,3,59,0,1,13,10,4,27,9,25,4,131,0
Dec-11,,5,42,32,12,85,34,864,153,574,Data are confidential,0,16,16,2,34,5,120,3,56,0,0,11,9,4,24,8,24,2,113,0
Jan-12,*,5,55,45,15,115,46,825,165,721,25,0,20,18,2,40,6,129,2,64,0,0,15,12,5,32,9,23,3,155,0
Feb-12,,6,48,37,12,97,34,658,135,592,19,0,18,15,2,34,4,110,2,52,0,0,12,10,4,27,7,18,3,124,0
Mar-12,,7,49,37,13,99,31,694,130,598,21,0,18,14,2,34,4,108,2,49,0,0,11,9,4,25,6,15,2,124,0
Apr-12,*,5,60,43,17,120,38,803,149,724,24,0,22,16,3,41,5,122,2,58,0,0,15,11,5,32,7,20,3,153,0
May-12,,3,47,34,16,98,32,681,118,583,19,0,18,12,3,34,5,60,2,48,0,0,12,9,5,26,7,23,3,123,0
Jun-12,,3,42,30,17,90,31,668,119,570,19,0,16,11,4,32,5,84,2,49,0,0,10,7,5,22,7,30,2,120,0
Jul-12,*,4,52,38,23,113,45,982,169,744,26,0,19,13,5,38,7,126,2,61,0,0,13,9,6,28,10,41,4,153,0
Aug-12,,5,41,30,17,88,34,892,145,600,21,0,14,10,3,28,5,112,2,52,0,0,13,8,5,26,8,45,3,129,0
Sep-12,,8,45,31,16,91,40,873,143,610,24,0,17,11,3,31,6,123,2,49,0,0,16,9,4,29,10,44,4,128,0
Oct-12,*,9,60,43,19,122,58,1270,212,793,27,0,21,17,3,41,7,142,3,50,0,1,19,11,5,36,14,53,4,162,0
Nov-12,,7,48,36,15,100,49,912,147,672,21,0,16,14,2,33,6,119,2,27,0,1,13,10,4,28,11,41,3,133,0
Dec-12,,6,40,30,12,82,35,917,152,628,17,0,15,14,2,31,5,104,2,23,0,0,12,10,4,26,9,32,3,115,0
```

Hasso Plattner Institut

Data Sources – Data Formats

- Data lakes
- Open (government) data
- Instrumented processes
- Sensor data
- Experimental output
- Database exports
- Excel



Overview

1. Bad Files
2. **Data P-r_e\p+a|r¶a.t/i~o-n**
3. Bad Data
4. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Data Preparation: Tasks and Tools

- Data discovery
- Data validation
- Data structuring
- Data enrichment
- Data filtering
- Data cleaning

- And for data scientists
  - Feature selection
  - Feature extraction

Categories	Available features	Data preparation tools						
		Altair	Paxata	SAP	SAS	Tableau	Talend	Trifacta
Data discovery	Locate missing values (nulls)	✓	✓	✓	✓	✓	✓	✓
	Locate outliers		✓		✓			✓
	Search by pattern	✓	✓	✓	✓	✓	✓	✓
	Sort data	✓	✓	✓	✓	✓	✓	✓
Data validation	Compare values (selection and join)	✓	✓	✓		✓	✓	✓
	Check data range	✓	✓	✓		✓	✓	✓
	Check permitted characters							✓
	Check column uniqueness	✓	✓	✓		✓	✓	✓
	Find type-mismatched data		✓	✓		✓	✓	✓
	Find data-mismatched datatypes		✓				✓	✓
Data structuring	Change column data type		✓	✓	✓	✓	✓	✓
	Delete column			✓	✓	✓	✓	✓
	Detect & change encoding						✓	✓
	Pivot / unpivot					✓		✓
	Rename column						✓	✓
	Split column						✓	✓
	Transform by example [13]							✓
Data enrichment	Assign semantic data type							
	Calculate column using ex							
	Discover & merge extern							
	Duplicate column							
	Generate primary key col							
	Join & union							
	Merge columns							✓
	Normalize numeric values	✓						✓
Data filtering	Delete/keep filtered rows	✓	✓					✓
	Delete empty and invalid rows	✓	✓	✓				✓
	Extract value parts	✓			✓			✓
	Filter with regular expressions							✓
Data cleaning	Change date & time format	✓	✓	✓	✓	✓	✓	✓
	Change letter case	✓	✓	✓	✓	✓	✓	✓
	Change number format	✓	✓	✓	✓	✓	✓	✓
	Deduplicate data	✓	✓	✓	✓		✓	✓
	Delete by pattern	✓	✓		✓	✓	✓	✓
	Edit & replace cell data	✓	✓	✓	✓	✓	✓	✓
	Fill empty cells	✓	✓				✓	✓
	Remove extra whitespace	✓	✓	✓	✓	✓	✓	✓
	Remove diacritics			✓				
	Standardize strings by pattern		✓	✓	✓	✓	✓	✓
	Standardize values in clusters		✓	✓	✓	✓	✓	✓

Felix Naumann
Bad Files, Bad Data 2021

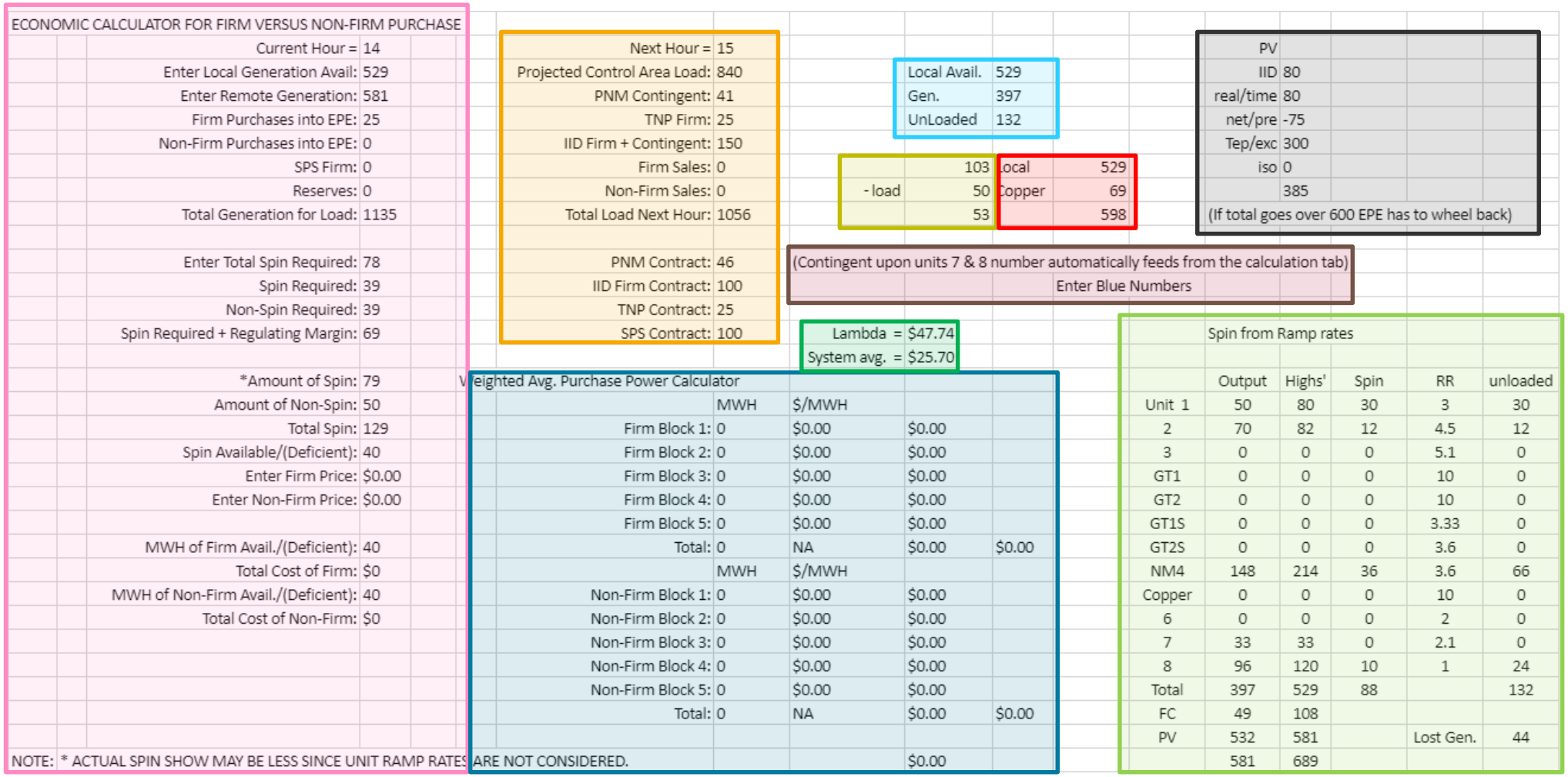
Selected Data Preparation Projects – Bringing Order to Files

- Mondrian
  - Dissecting multi-table files
- ExtracTable
  - Parsing visually delimited files
- Suragh
  - Identifying ill-formed records
- Strudel
  - Classify cell-types
- AggreCol
  - Identify aggregation cells
- Pollock
  - A data ingestion benchmark

Felix Naumann
Bad Files, Bad Data 2021

Mondrian
Dissecting multi-table files

Mondrian: Multitable Spreadsheets



ECONOMIC CALCULATOR FOR FIRM VERSUS NON-FIRM PURCHASE	
Current Hour =	14
Enter Local Generation Avail:	529
Enter Remote Generation:	581
Firm Purchases into EPE:	25
Non-Firm Purchases into EPE:	0
SPS Firm:	0
Reserves:	0
Total Generation for Load:	1135
Enter Total Spin Required:	78
Spin Required:	39
Non-Spin Required:	39
Spin Required + Regulating Margin:	69
*Amount of Spin:	79
Amount of Non-Spin:	50
Total Spin:	129
Spin Available/(Deficient):	40
Enter Firm Price:	$0.00
Enter Non-Firm Price:	$0.00
MWH of Firm Avail./(Deficient):	40
Total Cost of Firm:	$0
MWH of Non-Firm Avail./(Deficient):	40
Total Cost of Non-Firm:	$0

NOTE: * ACTUAL SPIN SHOW MAY BE LESS SINCE UNIT RAMP RATES ARE NOT CONSIDERED.

Next Hour =	15
Projected Control Area Load:	840
PNM Contingent:	41
TNP Firm:	25
IID Firm + Contingent:	150
Firm Sales:	0
Non-Firm Sales:	0
Total Load Next Hour:	1056
PNM Contract:	46
IID Firm Contract:	100
TNP Contract:	25
SPS Contract:	100

Local Avail.	529
Gen.	397
UnLoaded	132

	103	Local	529
- load	50	Copper	69
	53		598

PV	
IID	80
real/time	80
net/pre	-75
Tep/exc	300
iso	0
	385

(If total goes over 600 EPE has to wheel back)

(Contingent upon units 7 & 8 number automatically feeds from the calculation tab)

Enter Blue Numbers

Lambda =	$47.74
System avg. =	$25.70

Weighted Avg. Purchase Power Calculator

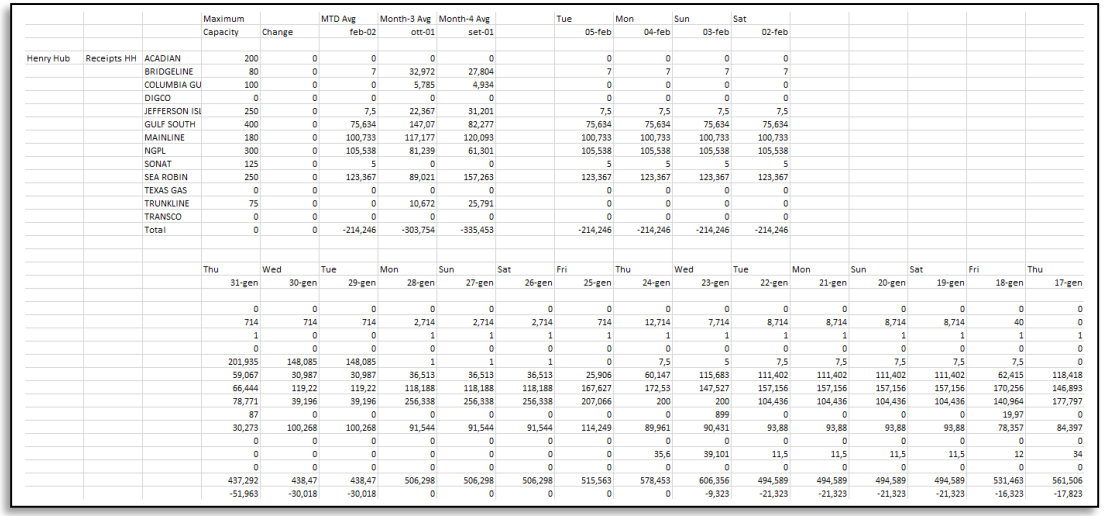
	MWH	$/MWH		
Firm Block 1:	0	$0.00	$0.00	
Firm Block 2:	0	$0.00	$0.00	
Firm Block 3:	0	$0.00	$0.00	
Firm Block 4:	0	$0.00	$0.00	
Firm Block 5:	0	$0.00	$0.00	
Total:	0	NA	$0.00	$0.00
	MWH	$/MWH		
Non-Firm Block 1:	0	$0.00	$0.00	
Non-Firm Block 2:	0	$0.00	$0.00	
Non-Firm Block 3:	0	$0.00	$0.00	
Non-Firm Block 4:	0	$0.00	$0.00	
Non-Firm Block 5:	0	$0.00	$0.00	
Total:	0	NA	$0.00	$0.00
			$0.00	

Spin from Ramp rates

	Output	Highs'	Spin	RR	unloaded
Unit 1	50	80	30	3	30
2	70	82	12	4.5	12
3	0	0	0	5.1	0
GT1	0	0	0	10	0
GT2	0	0	0	10	0
GT1S	0	0	0	3.33	0
GT2S	0	0	0	3.6	0
NM4	148	214	36	3.6	66
Copper	0	0	0	10	0
6	0	0	0	2	0
7	33	33	0	2.1	0
8	96	120	10	1	24
Total	397	529	88		132
FC	49	108			
PV	532	581		Lost Gen.	44
	581	689			

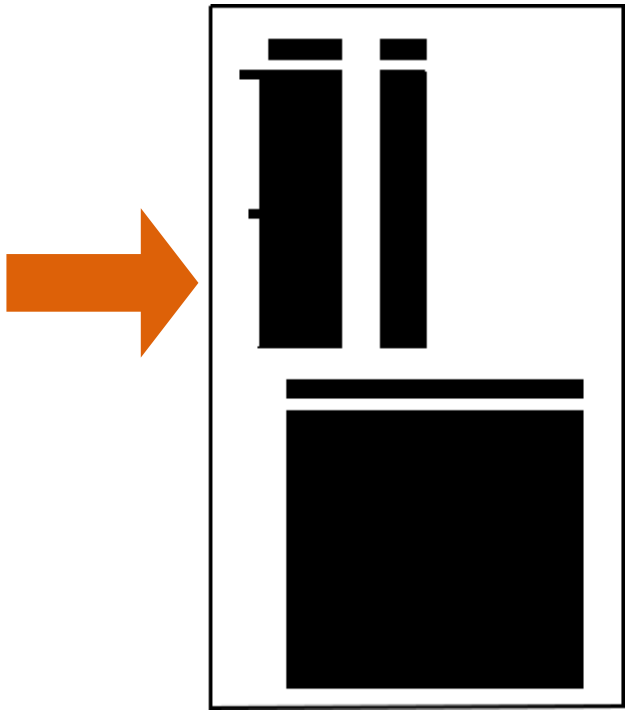
Felix Naumann
Bad Files, Bad Data 2021

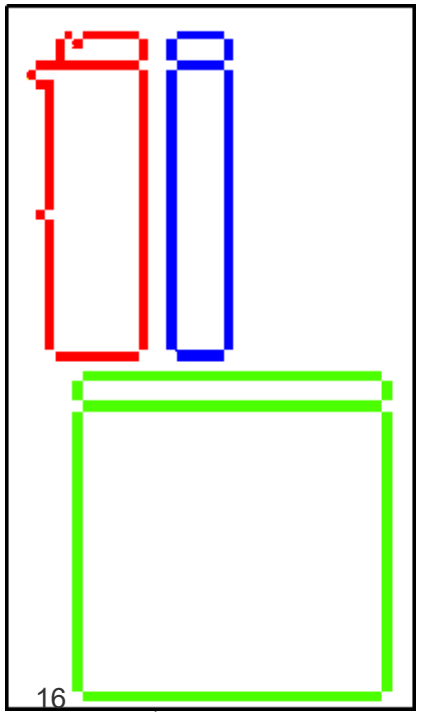
Mondrian: Clustering-based Table Recognition



			Maximum Capacity	Change	MTD Avg feb-02	Month-3 Avg ott-01	Month-4 Avg set-01		Tue 05-feb	Mon 04-feb	Sun 03-feb	Sat 02-feb
Henry Hub	Receipts HH	ACADIAN	200	0	0	0	0		0	0	0	0
		BRIDGELINE	80	0	7	32,972	27,804		7	7	7	7
		COLUMBIA GU	100	0	0	5,785	4,934		0	0	0	0
		DIGCO	0	0	0	0	0		0	0	0	0
		JEFFERSON ISL	250	0	7,5	22,367	31,201		7,5	7,5	7,5	7,5
		GULF SOUTH	400	0	75,634	147,07	82,277		75,634	75,634	75,634	75,634
		MAINLINE	180	0	100,733	117,177	120,093		100,733	100,733	100,733	100,733
		NGPL	300	0	105,538	81,239	61,301		105,538	105,538	105,538	105,538
		SONAT	125	0	5	0	0		5	5	5	5
		SEA ROBIN	250	0	123,367	89,021	157,263		123,367	123,367	123,367	123,367
		TEXAS GAS	0	0	0	0	0		0	0	0	0
		TRUNKLINE	75	0	0	10,672	25,791		0	0	0	0
		TRANSCO	0	0	0	0	0		0	0	0	0
		Total	0	0	-214,246	-303,754	-335,453		-214,246	-214,246	-214,246	-214,246

Thu 31-gen	Wed 30-gen	Tue 29-gen	Mon 28-gen	Sun 27-gen	Sat 26-gen	Fri 25-gen	Thu 24-gen	Wed 23-gen	Tue 22-gen	Mon 21-gen	Sun 20-gen	Sat 19-gen	Fri 18-gen	Thu 17-gen
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
714	714	714	2,714	2,714	2,714	714	12,714	7,714	8,714	8,714	8,714	8,714	40	0
1	0	0	1	1	1	1	1	1	1	1	1	1	1	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
201,935	148,085	148,085	1	1	1	0	7,5	5	7,5	7,5	7,5	7,5	7,5	0
59,067	30,987	30,987	36,513	36,513	36,513	25,906	60,147	115,683	111,402	111,402	111,402	111,402	62,415	118,418
66,444	119,22	119,22	118,188	118,188	118,188	167,627	172,53	147,527	157,156	157,156	157,156	157,156	170,256	146,893
78,771	39,196	39,196	256,338	256,338	256,338	207,066	200	200	104,436	104,436	104,436	104,436	140,964	177,797
87	0	0	0	0	0	0	0	899	0	0	0	0	19,97	0
30,273	100,268	100,268	91,544	91,544	91,544	114,249	89,961	90,431	93,88	93,88	93,88	93,88	78,357	84,397
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	35,6	39,101	11,5	11,5	11,5	11,5	12	34
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
437,292	438,47	438,47	506,298	506,298	506,298	515,563	578,453	606,356	494,589	494,589	494,589	494,589	531,463	561,506
-51,963	-30,018	-30,018	0	0	0	0	0	-9,323	-21,323	-21,323	-21,323	-21,323	-16,323	-17,823

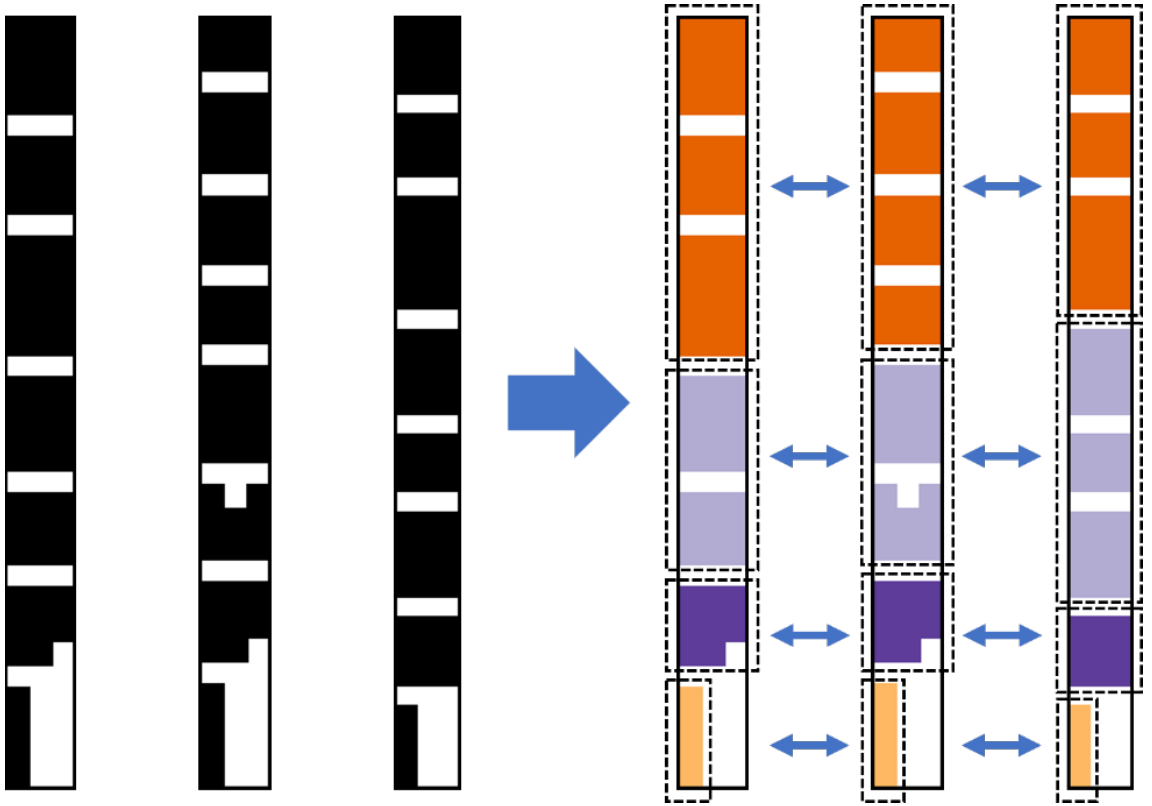




1. Render spreadsheet as image
2. Recognize elements
3. Cluster elements into tables

Mondrian: Layout Templates

- Different files may contain the same region layouts
- Same layout – different shapes of regions
- Template: a set of equivalent layouts



Felix Naumann
Bad Files, Bad Data 2021

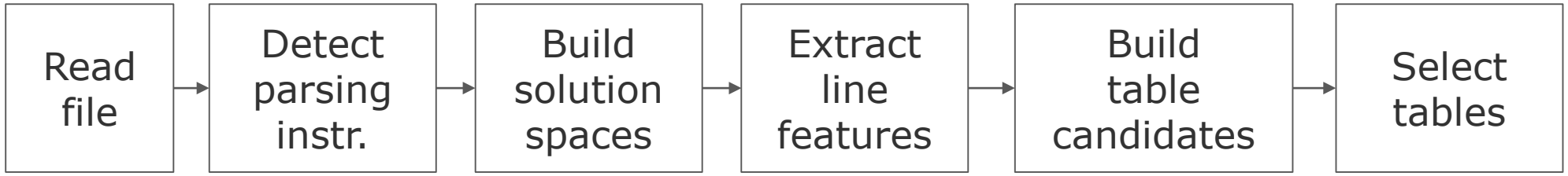
ExtracTable
Parsing visually delimited files

ExtracTable: Bad Files – Worse Files

min	max	num	dist	mean	std	comment
1.8	1.8	1	0	1.5	0	N
20	60	40	1	40	15	cab
0	0	1	0			
0	1	10	2			
0.01	0.01	1	0			
0.009	0.009	1	0			
0.2	7	40	1			
-0.35	-0.35	1	2			
-0.15	-0.15	1	0			
0.5	0.5	1	0			
0.2	0.2	1	0			
0.01	0.01	1	0			
27.947	27.947	1	0			
7.04345	7.04345	1	0			
146.691	146.691	1	0			
1	1	1	0			

```
#   OBIA4RTM config file for setting up Prospect4SAIL
#
#   Typical values (taken from J Gomez-Dans on https://pypi.org/project/prosail/)
#
#   ================================================================================
#   | Parameter   | Description of parameter          | Units        |Typical min | Typical max |
#   |-------------|-----------------------------------|--------------|------------|-------------|
#   |   N         | Leaf structure parameter          | N/A          | 0.8        | 2.5         |
#   |  cab        | Chlorophyll a+b concentration     | ug/cm2       | 0          | 80          |
#   |  caw        | Equivalent water thickiness       | cm           | 0          | 200         |
#   |  car        | Carotenoid concentration          | ug/cm2       | 0          | 20          |
#   |  cbrown     | Brown pigment                     | NA           | 0          | 1           |
#   |  cm         | Dry matter content                | g/cm2        | 0          | 200         |
#   |  lai        | Leaf Area Index                   | N/A          | 0          | 10          |
#   |  lidfa      | Leaf angle distribution           | N/A          | -          | -           |
#   |  lidfb      | Leaf angle distribution           | N/A          | -          | -           |
#   |  psoil      | Dry/Wet soil factor               | N/A          | 0          | 1           |
#   |  rsoil      | Soil brigthness factor            | N/A          | -          | -           |
#   |  hspot      | Hotspot parameter                 | N/A          | -          | -           |
#   |  tts        | Solar zenith angle                | deg          | 0          | 90          |
#   |  tto        | Observer zenith angle             | deg          | 0          | 90          |
#   |  phi        | Relative azimuth angle            | deg          | 0          | 360         |
#   | typelidf    | Leaf angle distribution type      | Integer      | -          | -           |
#   ================================================================================
#
#   You can enter your values below -> make sure not to alter the overall structure of this
#   template -> otherwise bad things might happen
#
#   Further Explainations:
#
#   min: Minimum Value of Parameter
#   max: Maximum Value of Parameter (in case min=max, the parameter will not be retrieved)
#   num: in case min!=max, the number of samples to be drawn for the specific parameter
```

ExtracTable: Process



Topic	Student	Supervision
Extracting Plain Tables from Text	Leonardo Hübscher	Felix Naumann, Lan Jiang
Distributed Duplicate Detection on Streaming-Data	Jakob Köhler	Thorsten Papenbrock
Multi-Aspect Embeddings for Fiction Novels	Lasse Kohlmeyer	Ralf Krestel, Tim Repke
Generating Rap Lyrics with Flow and Rhythm	Noel Danz	Ralf Krestel, Tim Repke

```
00000111111111111111111111111111111111111111111111000000011111111111000000000001111111111111
00000000001000001000000100001000011111111111111111110000000010000000011000001000000001000100000
00000000001000000000100000000100100000000000001110000010000001111111000000001000000000011111
00000000001000000000100010000001000000111111111100000100000000011110000100000001000100000 1
00000000010001000001000100010001000000111111111100001000111111111000010000000010001000001
```

Suragh
Identifying ill-formed records

Ill-formed Records Abort Data Loading

1	NAME	LAST NAME	JOBTITLE	DESCR	HIRE_DT	ANNUAL_RT	GROSS
2	Aaron	Kareem D	Utilities Inst Repair I	A50550	08/27/2018	32470	25743.94
3	Aaron	Patricia G	Office Services II	A03031	10/24/1979	60200	57806.13
4	Abadir	Adam O	Council Technician	A02002	12/12/2016	64823	64774.11
5	Abaku	Aigbolosimuan O	Police Officer	A99094	04/17/2018	53640	59361.55
6	Abbeduto	Mack	Assistant State Attorney	A29011	05/22/2017	68562	61693.59
7	Abbott	Ethan N	Recreation Arts Instructor	A68002	04/11/2018	33280	26156.48
8	AbbottCole	Michelle	Operations Officer III	A90005	11/28/2014	75110	75529.99
9	Abdal Rahim	Naim A	Fire Pump Operator Suppression	A64120	03/30/2011	69595	82132.61
10	Abdi	Ezekiel W	Police Sergeant	A99160	06/14/2007	93284	122992.1
11	Abdul Adl	Attrice A	Radio Dispatcher Sheriff	A38410	09/02/1999	50079	58459.78
12	Abdul Aziz	Hajr E	Swimming Pool Operator	P04002	06/01/2017	28554	15807.64
13	Abdul Aziz	Yaqub M	Swimming Pool Operator	P04002	06/01/2017	28554	6417.9
14	Abdul Saboor	Dana N	Paralegal	A99393	04/13/1998	57857	35715.2
15	Abdul	Jalil	Engineer I	A50101	07/17/2017	64505	65577.5
16	Abdul-Jabbar	Bushra A	Social Service Coordinator	A65028	04/14/2008	46395	46359.55
17	Abdul-Khaliq	Amahl	Recreation Leader II	A04005	06/06/2019	32131	5344.5
18	Abdullah	Beverly A	Office Support Specialist III	A06004	12/01/1986	41757	44522.96
19	Abdullahi	Sharon M	911 Operator	A64604	Wednesday, 6 October 2004	56322	53751.88
20	Abdullateef	Muhammed L	Supt of Public Bldg Repair	A85001	05/09/2019	78000	9000
21	Abdulrahman	Mustafa H	Police Officer Trainee	A99416	12/28/2018	53512	25347.99
22	Abdul Saboor	Jamillah	Printer Library	A75055	07/27/2009	44584	42555.83
23	Abdunafi	Karim	Community Aide	A04015	06/13/2019	24960	9888.55
24	Abdur-Rahman	Diane	Office Services I	A03092	03/27/2017	25363	25649.94
25	Abdurrahman	Saleh Z	Lifeguard I	P04002	06/04/2019	23920	1245.5
26	Abebe	Miraf E	Auditor II	A24002	Saturday, 2 June 2012	67236	60600.92
27	Abend Kollin	Emily L	Fleet Quality Control Analyst	A85301	01/05/2017	47828	49133.11
28	Abid	Amal	Engineer II	A49102	12/02/2013	71774	71630.83
29	Abid	Paula	Recreation Arts Instructor	A04009	06/25/2007	24960	3591.96

Suragh: Row Patterns – Outlier Rows

```
SFY,Fund_Source,Age,Total Children Receiving CHDP Services,Total,...
2008-2009,All,0,557757,24.34%,"$21,840,767 ",38.03%
2008-2009,All,1,314994,13.75%,"$7,262,306 ",12.64%
                    ...
2008-2009,All,10,55674,2.43%,"$1,146,066 ",2.00%SFY,Fund_Source,...
2008-2009,All,0,557757,24.34%,"$21,840,767 ",38.03%
2008-2009,All,1,314994,13.75%,"$7,262,306 ",12.64%
                    ...
2008-2009,All,Unknown,1063,0.05%,$0 ,0.00%
2008-2009,All,TOTAL,2291689,100.00%,"$57,436,517 ",100.00%
2008-2009,FFS,0,169699,41.75%,"$14,450,511 ",45.16%
2008-2009,FFS,1,48823,12.01%,"$5,157,938 ",16.12%
2008-2009,FFS,2,29241,7.19%,"$2,019,526 ",6.31%
                    ...
2008-2009,FFS,UNKNOWN,0,0.00%,$0 ,0.00%
2008-2009,FFS,TOTAL,406504,100.00%,"$31,995,243 ",100.00%
2008-2009,GFS,0,91612,33.09%,"$7,363,856 ",31.33%
2008-2009,GFS,1,22723,8.21%,"$2,061,059 ",8.77%
                    ...
```

2008-2009	All	<SEQD>	<SEQD>	<SEQD>.<SEQD>%	"$<NUM> "	<SEQD>.<SEQD>%
2008-2009	<UL><UL><UL>	<SEQD>	<SEQD>	<SEQD>.<SEQD>%	"$<NUM> "	<SEQD>.<SEQD>%

Felix Naumann
Bad Files, Bad Data 2021

Strudel
Classify cell-types

Strudel: Verbose CSV Files

Arrest Table						
Arrests for Drug Abuse Violations						
Percent Distribution by Region, 2007						
Drug abuse violations		United States total	Northeast	Midwest	South	West
Total1		100	100	100	100	100
Sale/Manufacturing:	Total	17.5	22.5	18.3	17.1	15
	Heroin or cocaine and their derivatives	7.9	14.2	6.2	7.9	5.5
	Marijuana	5.3	5.7	7.7	4.6	4.7
	Synthetic or manufactured drugs	1.5	1.1	1.1	2.6	0.7
	Other dangerous nonnarcotic drugs	2.8	1.6	3.3	2	4.2
Possession:	Total	82.5	77.5	81.7	82.9	85
	Heroin or cocaine and their derivatives	21.5	22.3	14.7	22.8	22.7
	Marijuana	42.1	44.2	53.1	47.9	29.6
	Synthetic or manufactured drugs	3.3	2.3	3.2	4.3	2.8
	Other dangerous nonnarcotic drugs	15.6	8.6	10.7	7.8	29.9
1 Because of rounding, the percentages may not add to 100.(						

Metadata

Header

Group header

Aggregation

Notes

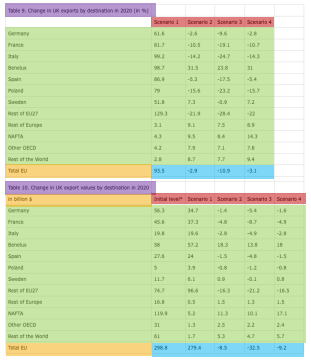
Data

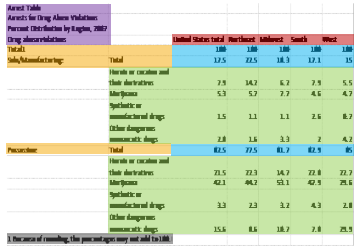
Metadata
Header
Group header
Data
Aggregation
Notes

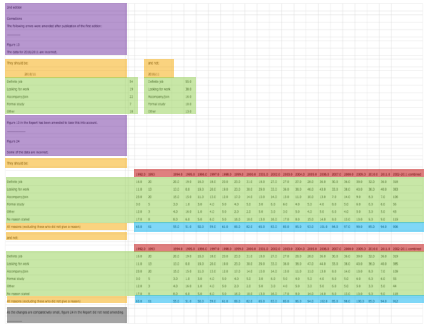
Felix Naumann
Bad Files, Bad Data 2021

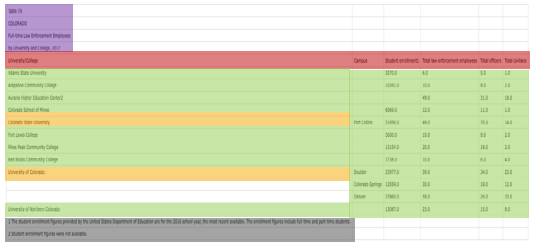
Crime In the US (CIUS): https://ucr.fbi.gov/crime-in-the-u.s

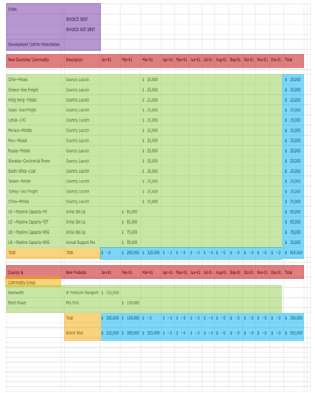
Strudel: Structure Diversity in Verbose CSV Files

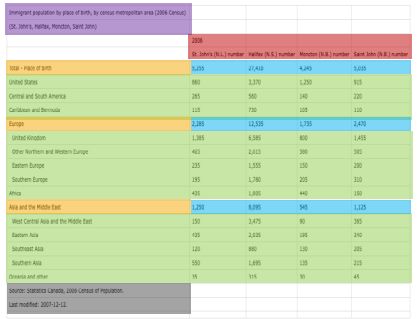


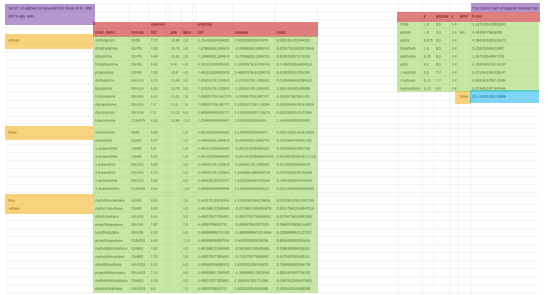










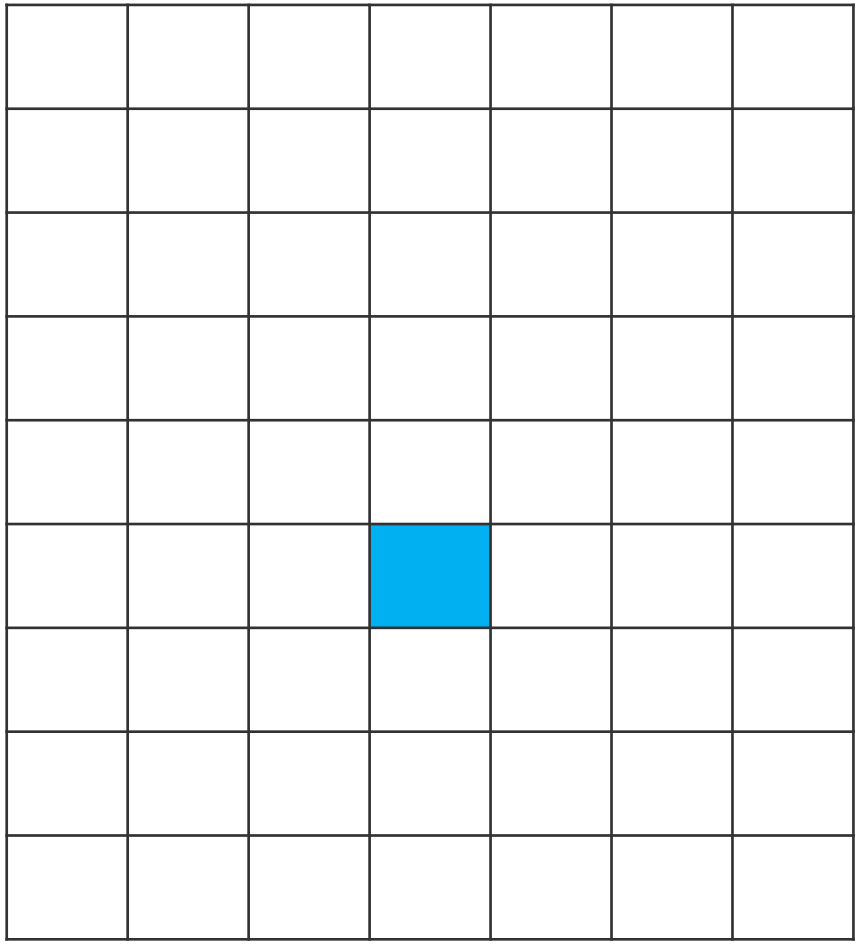


Metadata
Header
Group header
Data
Aggregation
Notes

Felix Naumann
Bad Files, Bad Data 2021

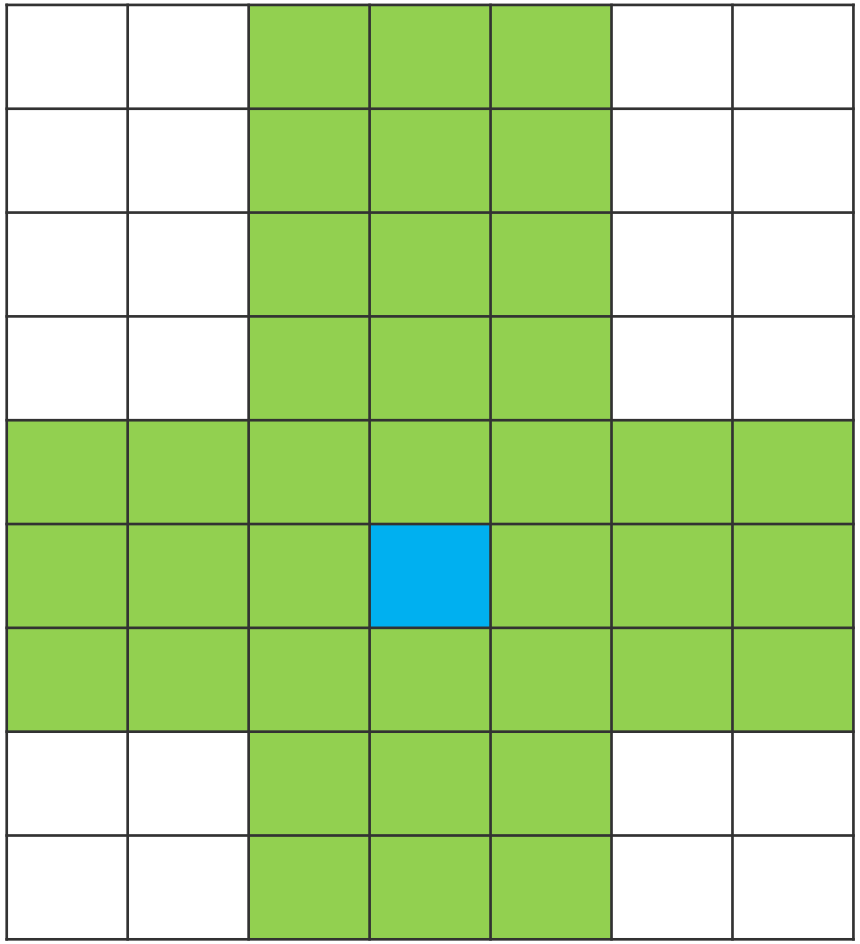
Strudel: Cell Classification Features

- Content features
  - Cell value length
  - Cell data type
  - Cell row/column index (position)
  - Cell has aggregation keywords (e.g., “total”)
  - Row/column of this cell has aggregation keywords



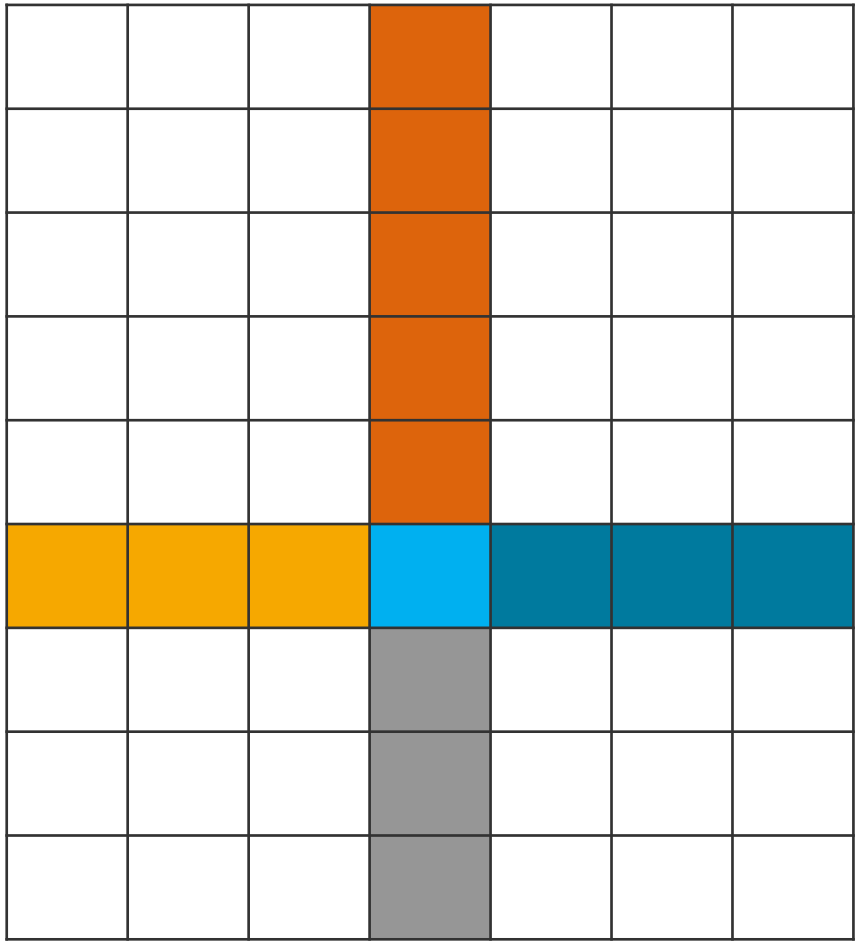
Strudel: Cell Classification Features

- Contextual features
  - Is preceding/following row empty
  - Is left/right column empty
  - Empty cell ratio in the row/column
  - Block size (4-connected component incl. cell)
  - Neighbor cell value length
  - Neighbor cell data type



Strudel: Cell Classification Features

- Computational feature
  - Is the value of this cell an aggregation of other cells in the same row or column?



AggreCol
Identify aggregation cells

AggreCol: Aggregations in CSV Files

$$\%\ \text{Change 2003 vs. 2002} = \frac{\text{FY2003} - \text{FY2002}}{\text{FY2002}}$$

MICROS Systems, Inc.								
Financial Summary								
	% Change							
Income Statement Data	2003 vs. 2002	FY2003	FY2002	FY2001	FY2000	FY1999	FY1998	FY1997
Hardware Revenue	2.2%	$137,013	$134,121	$116,058	$152,186	$155,237	$126,974	$102,816
Software Revenue	17.8%	$71,251	$60,484	$55,873	$66,290	$63,317	$57,744	$45,985
Service Revenue	11.2%	$191,927	$172,558	$154,845	$143,378	$118,525	$97,200	$79,368
Total Revenue	9.0%	$400,191	$367,163	$326,776	$361,854	$337,079	$281,918	$228,169
Memo Item:								
Maintenance Revenue (included in Service Revenue)	13.9%	$113,274	$99,467	$87,007	$65,628	$54,953	$45,908	$37,38
Hardware Gross Profit	2.3%	$38,977	$38,116	$40,683	$51,462	$50,670	$43,947	$39,267
Hardware Gross Profit %	-	28.4%	28.4%	35.1%	33.8%	32.6%	34.6%	38.2%
Software Gross Profit	11.5%	$54,045	$48,457	$46,875	$51,349	$52,138	$47,235	$37,464
Software Gross Profit %	-4.2 Points	75.9%	80.1%	83.9%	77.5%	82.3%	81.8%	81.5%
Service Gross Profit	16.5%	$105,538	$90,564	$76,472	$71,741	$61,367	$46,455	$39,447
Service Gross Profit %	+2.5 Points	55.0%	52.5%	49.4%	50.0%	51.8%	47.8%	49.7%
Total Gross Profit	12.1%	$198,560	$177,137	$164,028	$174,552	$164,175	$137,637	$116,178
Gross Profit %	+1.4 Points	49.6%	48.2%	50.2%	48.2%		48.8%	50.9%

$$\text{Hardware GP \%} = \frac{\text{Hardware GP}}{\text{Hardware Revenue}}$$

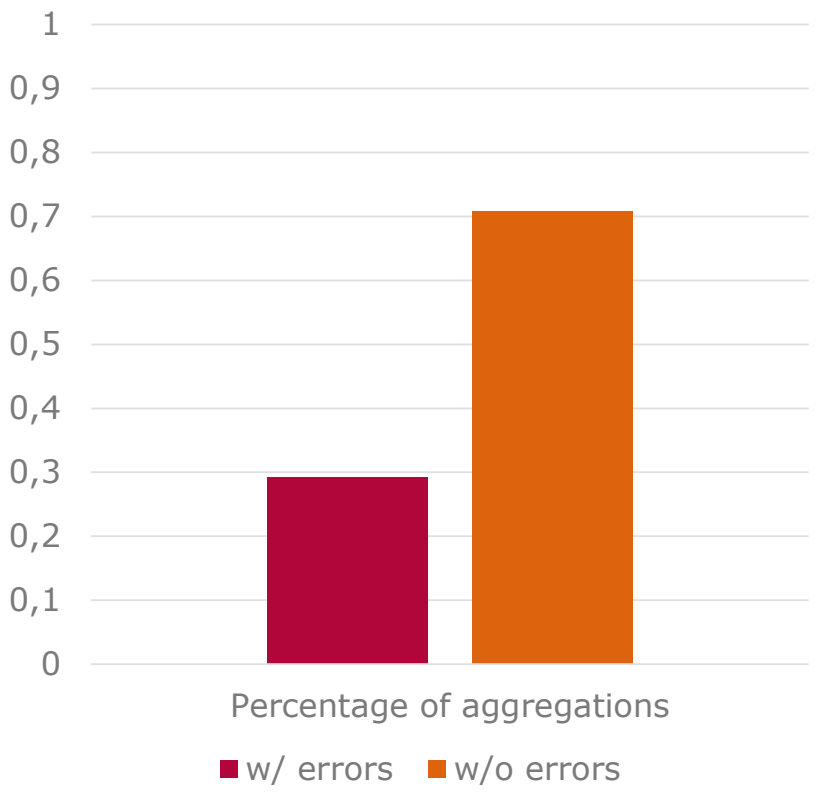
$$\text{Service GP \%} = \text{FY2003} - \text{FY2002}$$

$$\text{Total Gross Profit} = \text{Hardware GP} + \text{Software GP} + \text{Service GP}$$

Felix Naumann
Bad Files, Bad Data 2021

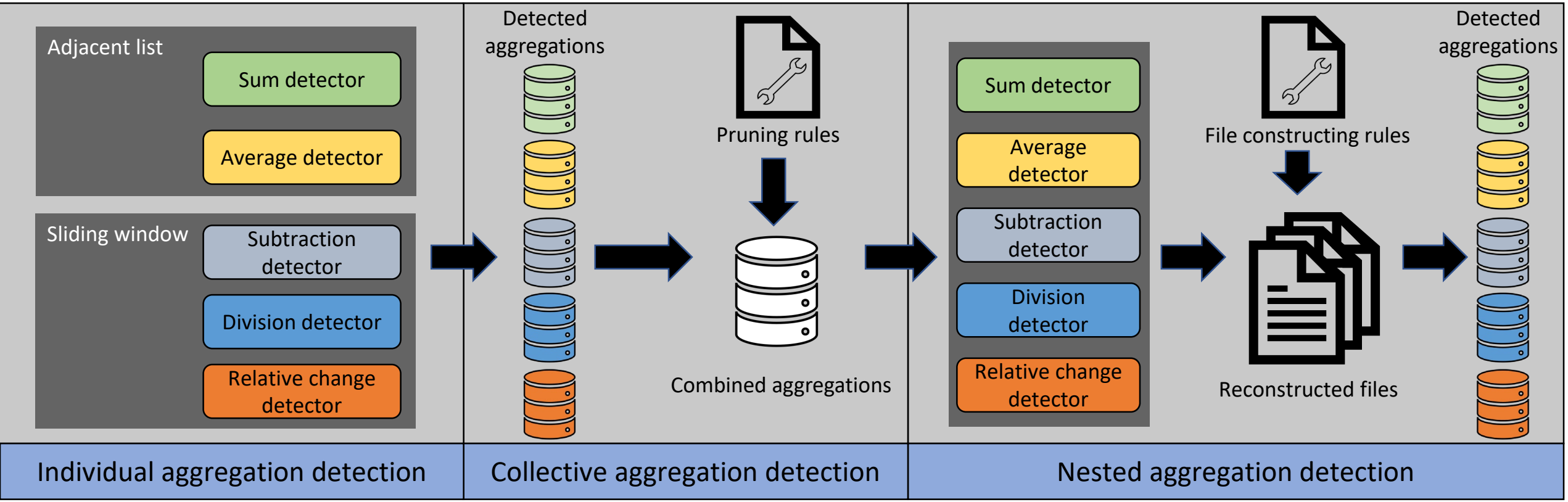
AggreCol: Errors in Aggregations

- Error level
  - Difference between the true number and the observed number
  - Highest observed error level: 37.5%
  - 29% of all aggregation have error level > 0%



Felix Naumann
Bad Files, Bad Data 2021

AggreCol: Process



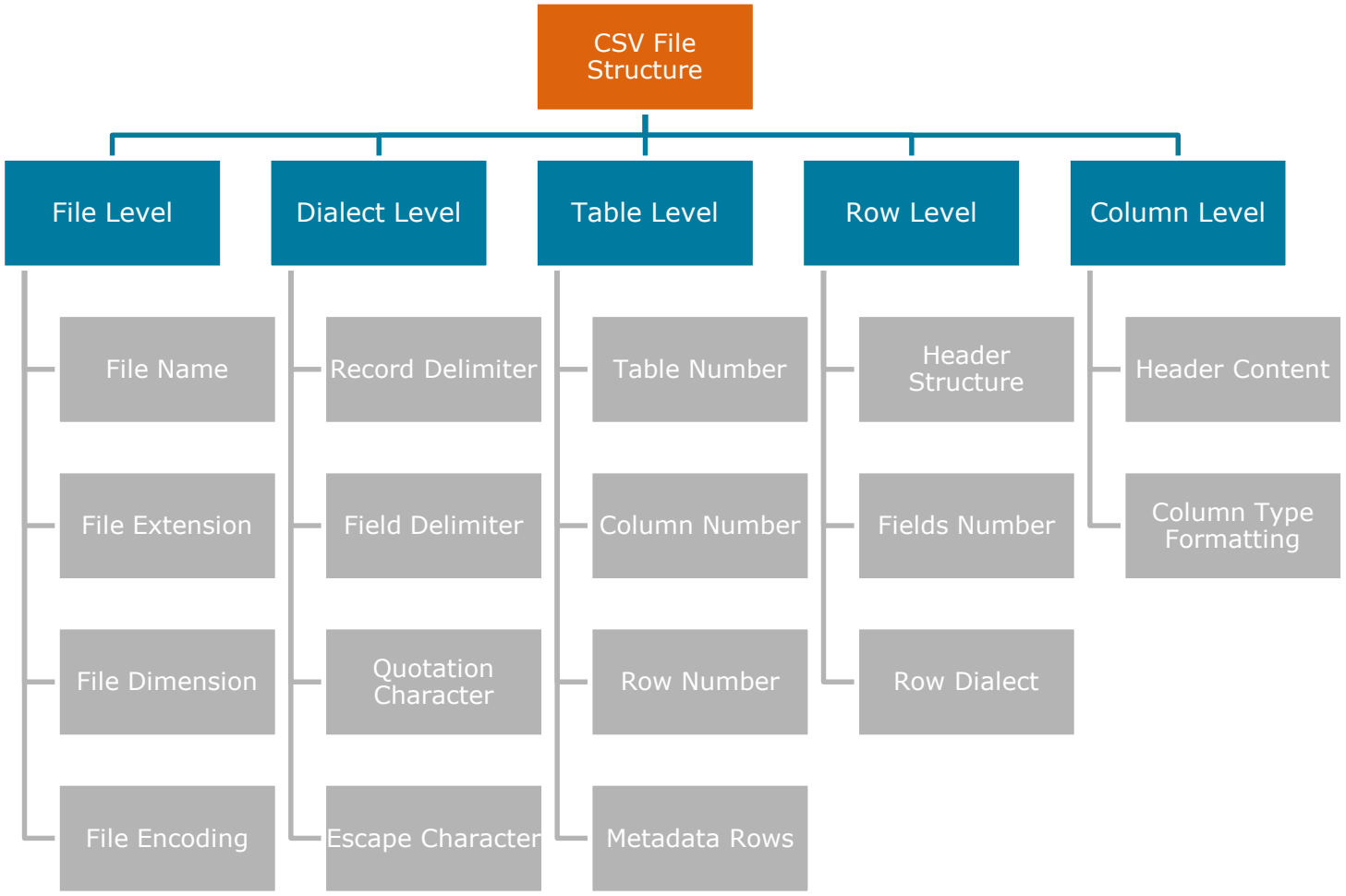
Pollock
A data ingestion benchmark

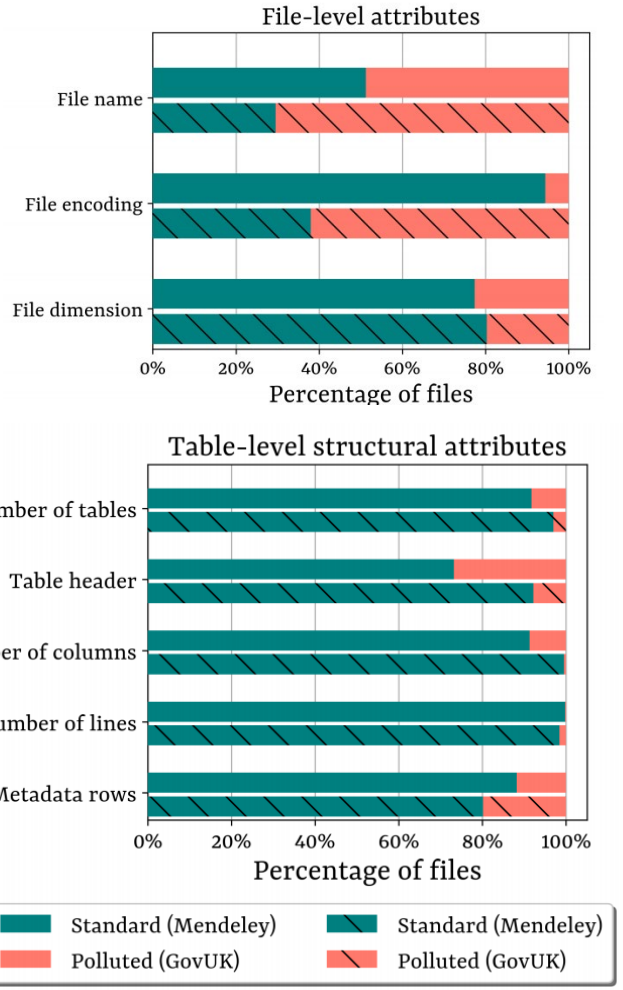
Pollock: Benchmarking Ingestion Ability

Systems under test:
- Programming framework (Pandas)
- Spreadsheet software (Libreoffice)
- Database tool (MySQL)
- Data Visualization (Tableau)

```
Python 3.8.5 (default, Sep  3 2020, 21:29:08) [MSC v.1916 64 bi
Type "help", "copyright", "credits" or "license" for more infor
>>> import pandas as pd
>>> pd.read_csv("11-708-data-nlss-2009-1.csv")
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "C:\Users\User\miniconda3\envs\pollution\lib\site-packages\pandas\io\parsers.py", line 686, in read_csv
    return _read(filepath_or_buffer, kwds)
  File "C:\Users\User\miniconda3\envs\pollution\lib\site-packages\pandas\io\parsers.py", line 458, in _read
    data = parser.read(nrows)
  File "C:\Users\User\miniconda3\envs\pollution\lib\site-packages\pandas\io\parsers.py", line 1196, in read
    ret = self._engine.read(nrows)
  File "C:\Users\User\miniconda3\envs\pollution\lib\site-packages\pandas\io\parsers.py", line 2155, in read
    data = self._reader.read(nrows)
  File "pandas\_libs\parsers.pyx", line 847, in pandas._libs.parsers.TextReader.read
  File "pandas\_libs\parsers.pyx", line 862, in pandas._libs.parsers.TextReader._read_low_memory
  File "pandas\_libs\parsers.pyx", line 918, in pandas._libs.parsers.TextReader._read_rows
  File "pandas\_libs\parsers.pyx", line 905, in pandas._libs.parsers.TextReader._tokenize_rows
  File "pandas\_libs\parsers.pyx", line 2042, in pandas._libs.parsers.raise_parser_error
pandas.errors.ParserError: Error tokenizing data. C error: Expected 25 fields in line 97, saw 27
```

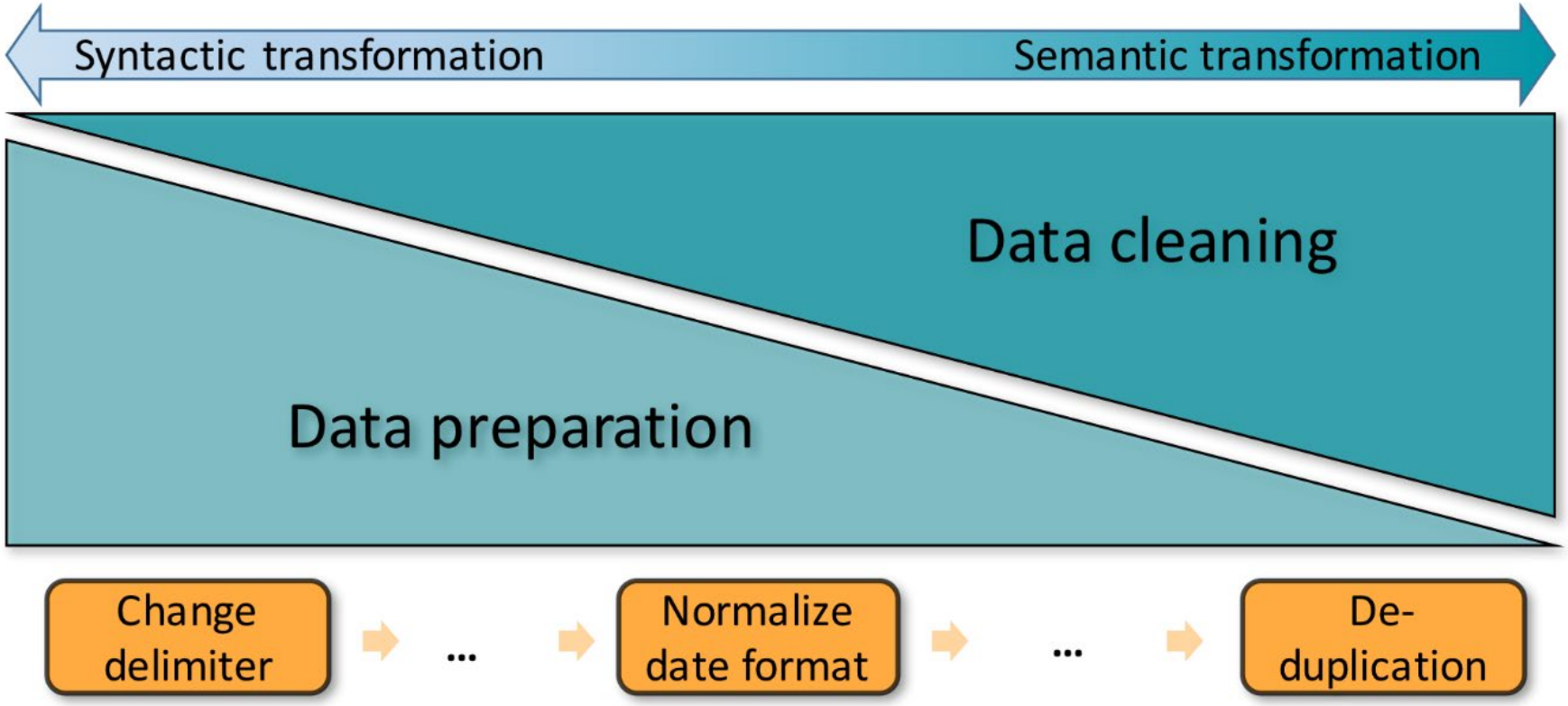
Pollock: Benchmark Dimensions





Data Preparation vs. Data Cleaning

- Data preparation adds syntactic and structural value
- Data cleaning adds semantic value



Felix Naumann
Bad Files, Bad Data 2021

Overview

1. Bad Files
2. Data Preparation
3. **Bad Data**
4. Data Cleaning

https://unsplash.com/photos/vGefUiWm0xI

Real-world data is raw and dirty

488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brirreny spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 brirtany spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brirttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 brirttney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 b		
2696 britteny spears	26 brinity spears	9 britrney spears	5 g		
1807 briney spears	26 britenay spears	9 brtiny spears	5 s		
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 b		
1479 brintey spears	26 brittan spears	9 brtny spears	4 b		
1479 britanny spears	26 brittne spears	9 brytny spears	4 b		
1338 britiny spears	26 btittany spears	9 rbitney spears	4 b		
1211 britnet spears	24 beitney spears	8 birtiny spears	4 b		
1096 britiney spears	24 birteny spears	8 bithney spears	4 b		
991 britaney spears	24 brightney spears	8 brattany spears	4 b		
991 britnay spears	24 brintiny spears	8 breitny spears	4 b		
811 brithney spears	24 britanty spears	8 breteny spears	4 b		
811 brtiney spears	24 britenny spears	8 brightny spears	4 b		
664 birtney spears	24 britini spears	8 brintay spears	4 b		
664 brintney spears	24 britnwy spears	8 brinttey spears	4 b		
664 briteney spears	24 brittni spears	8 briotney spears	4 b		
601 bitney spears	24 brittnie spears	8 britanys spears	4 b		
601 brinty spears	21 biritney spears	8 britley spears	4 b		
544 brittaney spears	21 birtany spears	8 britneyb spears	4 b		
544 brittnay spears	21 biteny spears	8 britnrey spears	4 b		
364 britey spears	21 bratney spears	8 britnty spears	4 b		
364 brittiny spears	21 britani spears	8 brittner spears	4 b		
329 brtney spears	21 britanie spears	8 brottany spears	4 b		
269 bretney spears	21 briteany spears	7 baritney spears	4 b		
269 britneys spears	21 brittay spears	7 birntey spears	4 b		
244 britne spears	21 brittinay spears	7 biteney spears	4 b		
244 brytney spears	21 brtany spears	7 bitiny spears	4 b		
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany spears	2 britneyh spears
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears

DQ-Problems: Effects

- Incorrect prices in inventory retail databases
  - Costs for consumers 2.5 billion $
  - 80% of barcode-scan-errors to the disadvantage of consumer
- IRS 1992: almost 100,000 tax refunds not deliverable
- 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous.
- US-Postal Service: of 100,000 mass-mailings up to 7,000 undeliverable due to incorrect addresses

IRS might be after you — to mail you a check

Incorrect addresses stall nearly 1,500 Tennessee refunds

By BONNA de la CRUZ
Staff Writer

Now that Tilcia L. Menifee knows that she'll be getting $500 in a tax refund from Uncle Sam, she can do some Christ-mas shopping, she said.

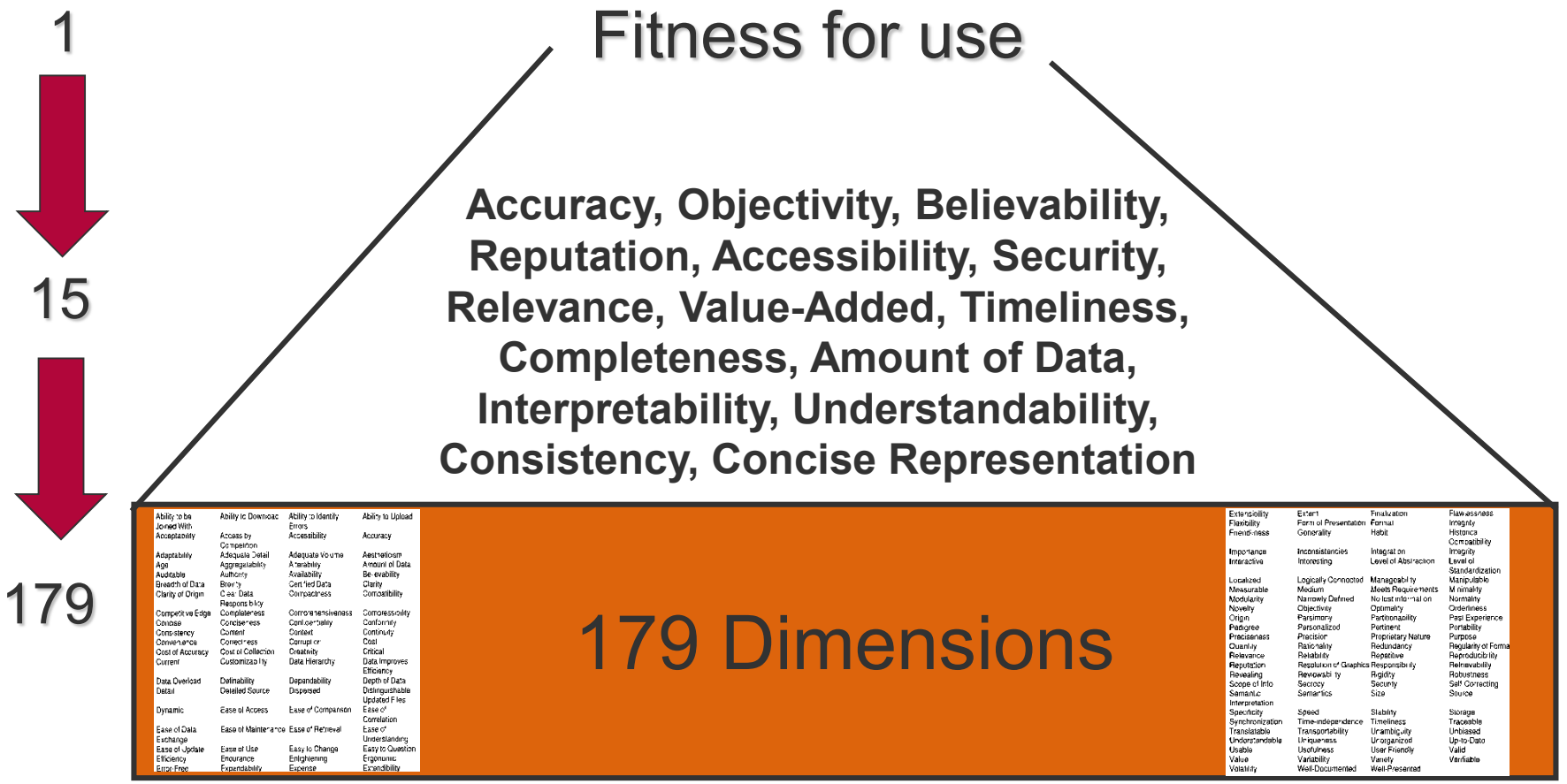
Felix Naumann
Bad Files, Bad Data 2021

Hidden Values / Hidden Value

	Feld						
	Name1	Name2	Name3	City	District	Street	Sum
Mobile phone	41	501	10	0	2677	297	3526
Phone	15	98	6	0	221	9579	9919
Cost center	283	1112	73	2	87	16	1573
Registration ID	11	583	1	1	0	3	599
Delivery ID	55	390	9	0	212	15	681
Department	3711	9997	115	60	439	175	14497
Embargo flag	129	143	2	0	66	9	349
Deletion flag	1028	442	5	36	113	10	1634
Legal form	131700	66136	187	6	64	57	198150
Credit info	0	100	11	0	18	0	129
Commission	216	352	1	2	36	10	617
Construction site	2013	3452	42	5	124	222	5858
Loading point	2923	3808	94	1503	958	3065	12351
Administration	13410	12461	172	19	295	7075	33432
Summe	155535	99575	728	1634	5310	20533	

Source: Joachim Schmid, FUZZY! Informatik AG

Zooming into Information Quality



Felix Naumann
Bad Files, Bad Data 2021

Overview

1. Bad Files
2. Data Preparation
3. Bad Data
4. **Data Cleaning**

https://unsplash.com/photos/vGefUiWm0xI

Error Detection

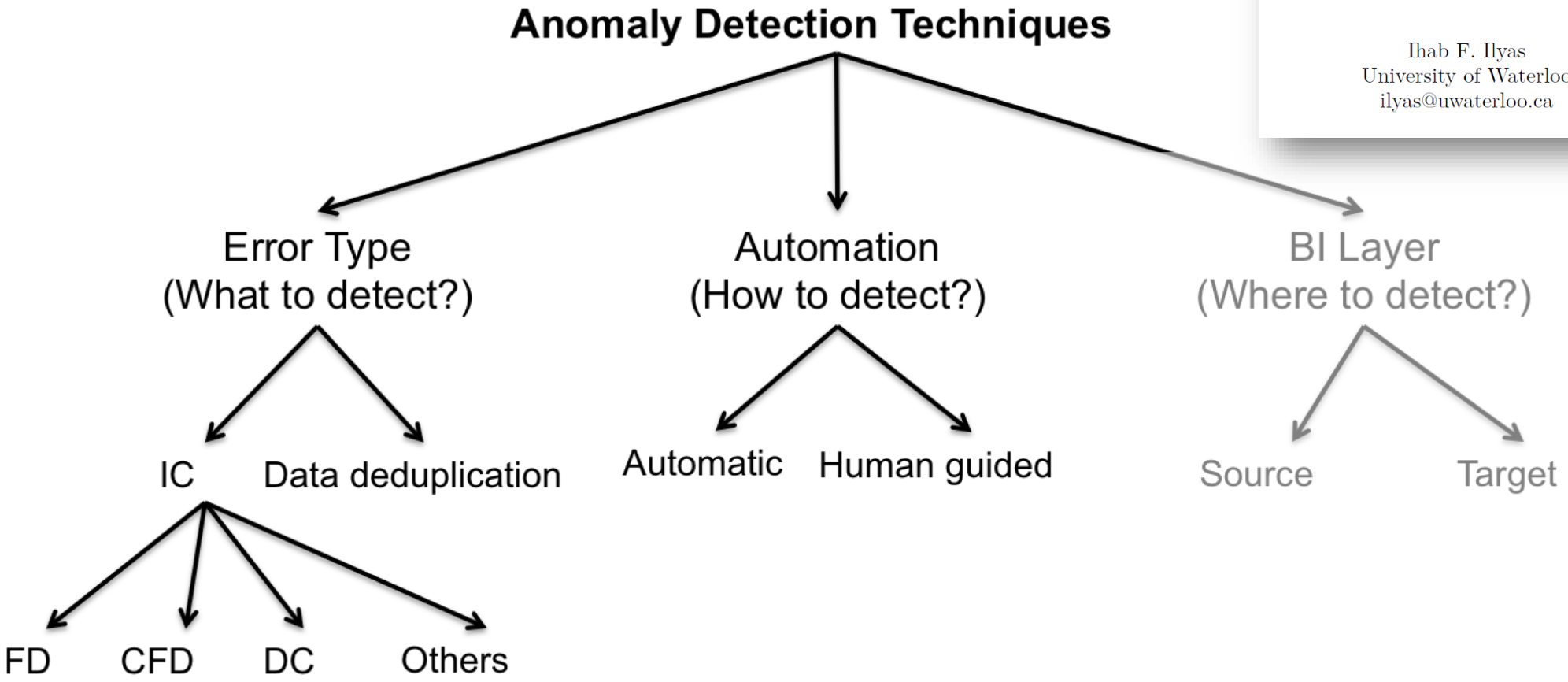
Foundations and Trends® in Databases
Vol. 5, No. 4 (2012) 281–393
© 2015 I. F. Ilyas and X. Chu
DOI: 10.1561/1900000045

now
the essence of knowledge

Trends in Cleaning Relational Data: Consistency and Deduplication

Ihab F. Ilyas
University of Waterloo
ilyas@uwaterloo.ca

Xu Chu
University of Waterloo
x4chu@uwaterloo.ca



Felix Naumann
Bad Files, Bad Data 2021

Data Repairing

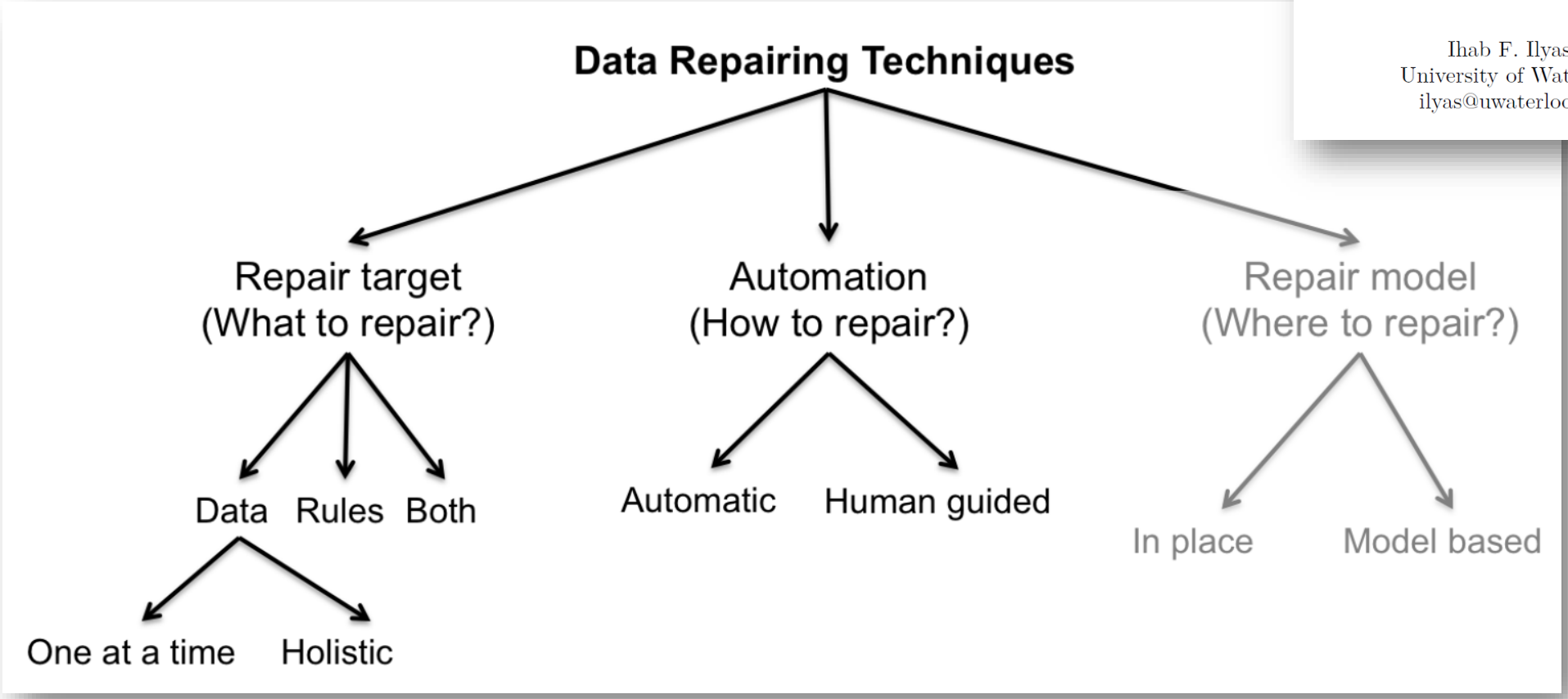
Foundations and Trends® in Databases
Vol. 5, No. 4 (2012) 281–393
© 2015 I. F. Ilyas and X. Chu
DOI: 10.1561/1900000045

now
the essence of knowledge

Trends in Cleaning Relational Data: Consistency and Deduplication

Ihab F. Ilyas
University of Waterloo
ilyas@uwaterloo.ca

Xu Chu
University of Waterloo
x4chu@uwaterloo.ca



Felix Naumann
Bad Files, Bad Data 2021

What to Detect

- Integrity constraints (ICs)
  - Functional dependencies (FDs)
  - Denial constraints (DCs)
    - Functional Dependency ZIP → City:
      $$\forall t_1, t_2 \in R: \neg(t_1.zip = t_2.zip \wedge t_1.city \neq t_2.city)$$
    - Same state, more income, lower tax rate:
      $\forall t_1, t_2 \in R: \neg(t_1.state = t_2.state \wedge t_1.income > t_2.income \wedge t_1.taxRate < t_2.taxRate)$
  - Other constraints
    - Inclusion dependencies, matching dependencies, conditional dependencies…
- Automatic discovery methods: data profiling
- Duplicate detection

What to Repair

- Find violations to integrity constraints
  - Example integrity constraints:
    1. ZIP $\rightarrow$ State
    2. $\forall t_1, t_2 \in R: \neg(t_1.\mathrm{ST} = t_2.\mathrm{ST} \wedge t_1.\mathrm{SAL} > t_2.\mathrm{SAL} \wedge t_1.\mathrm{ROLE} = \mathrm{E} \wedge t_2.\mathrm{ROLE} = \mathrm{M})$
  - Which value is likely to be wrong?

HoloClean: Holistic Data Repairs with Probabilistic Inference

Theodoros Rekatsinas*, Xu Chu†, Ihab F. Ilyas†, Christopher Ré*
* Stanford University and † University of Waterloo

TID	FN	LN	ROLE	ZIP	ST	SAL
t_1	Anne	Nash	E	85376	NY	110
t_2	Mark	White	M	90012	NY	80
t_3	Mark	Lee	E	85376	AZ	75

Felix Naumann
Bad Files, Bad Data 2021

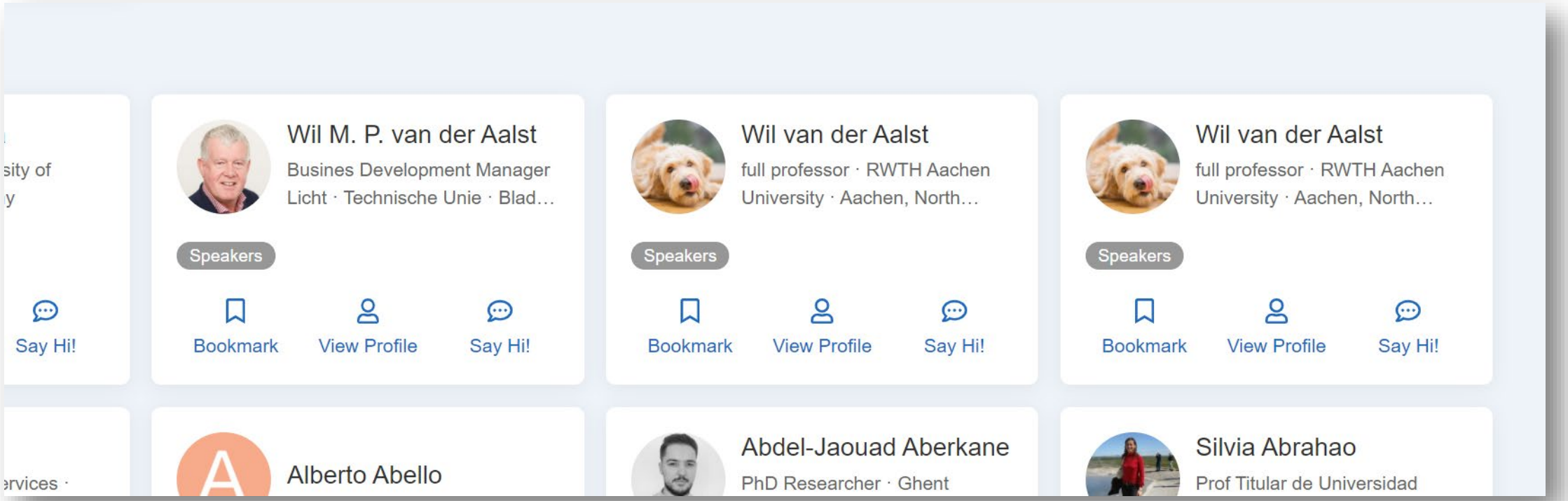
Duplicate Detection

- Duplicate detection is the discovery of multiple representations of the same real-world object.

- Problem 1: Representations are not identical.
  - *Fuzzy duplicates*
  - Solution: Similarity measures and threshold
    - Value- and record-comparisons
    - Domain-dependent or domain-independent

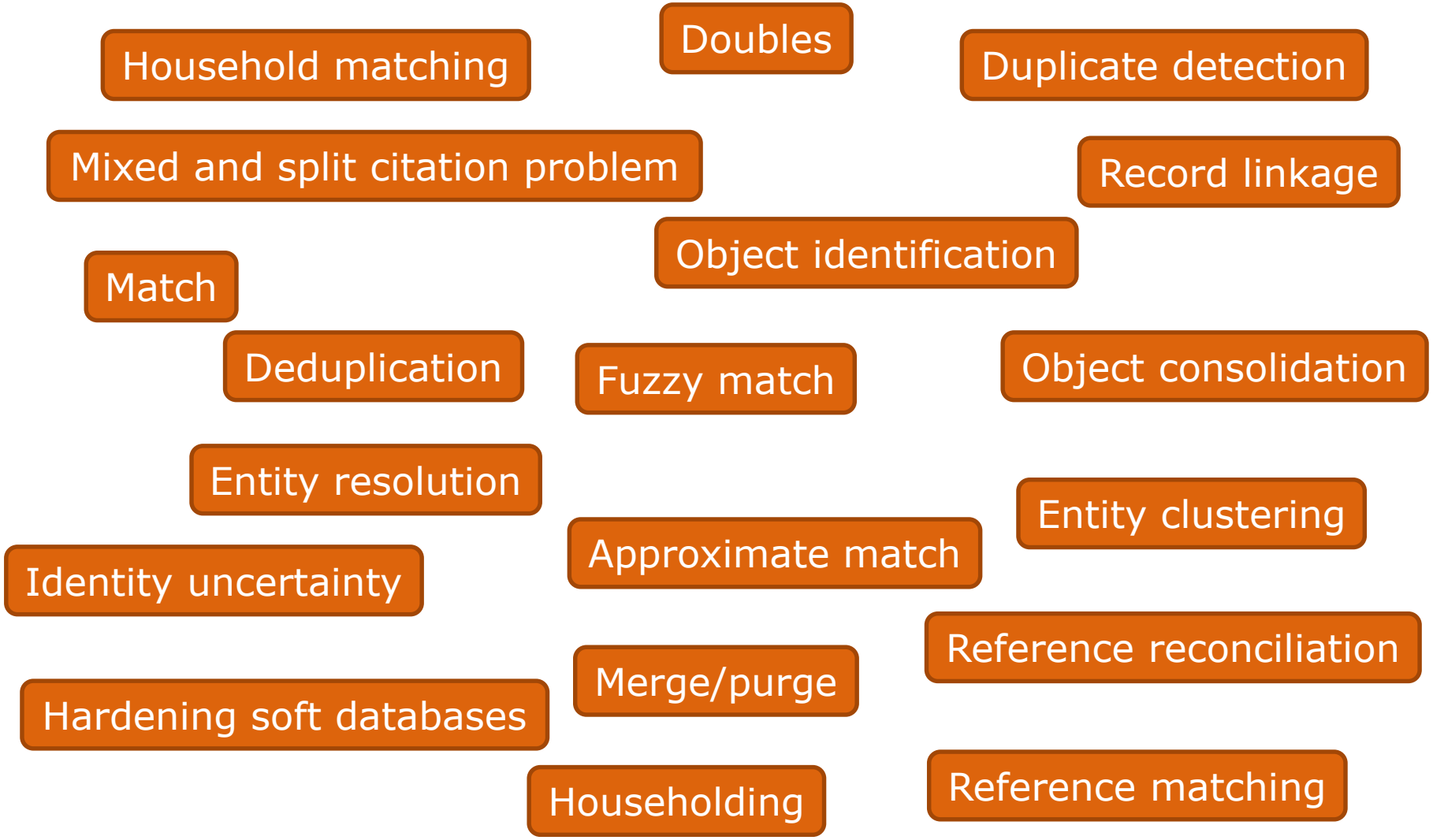
- Problem 2: Data sets are large.
  - $O(n^2)$: Comparison of every pair of records.
  - Solution: Algorithms
    - E.g., avoid comparisons by partitioning.

Origin of Duplicates

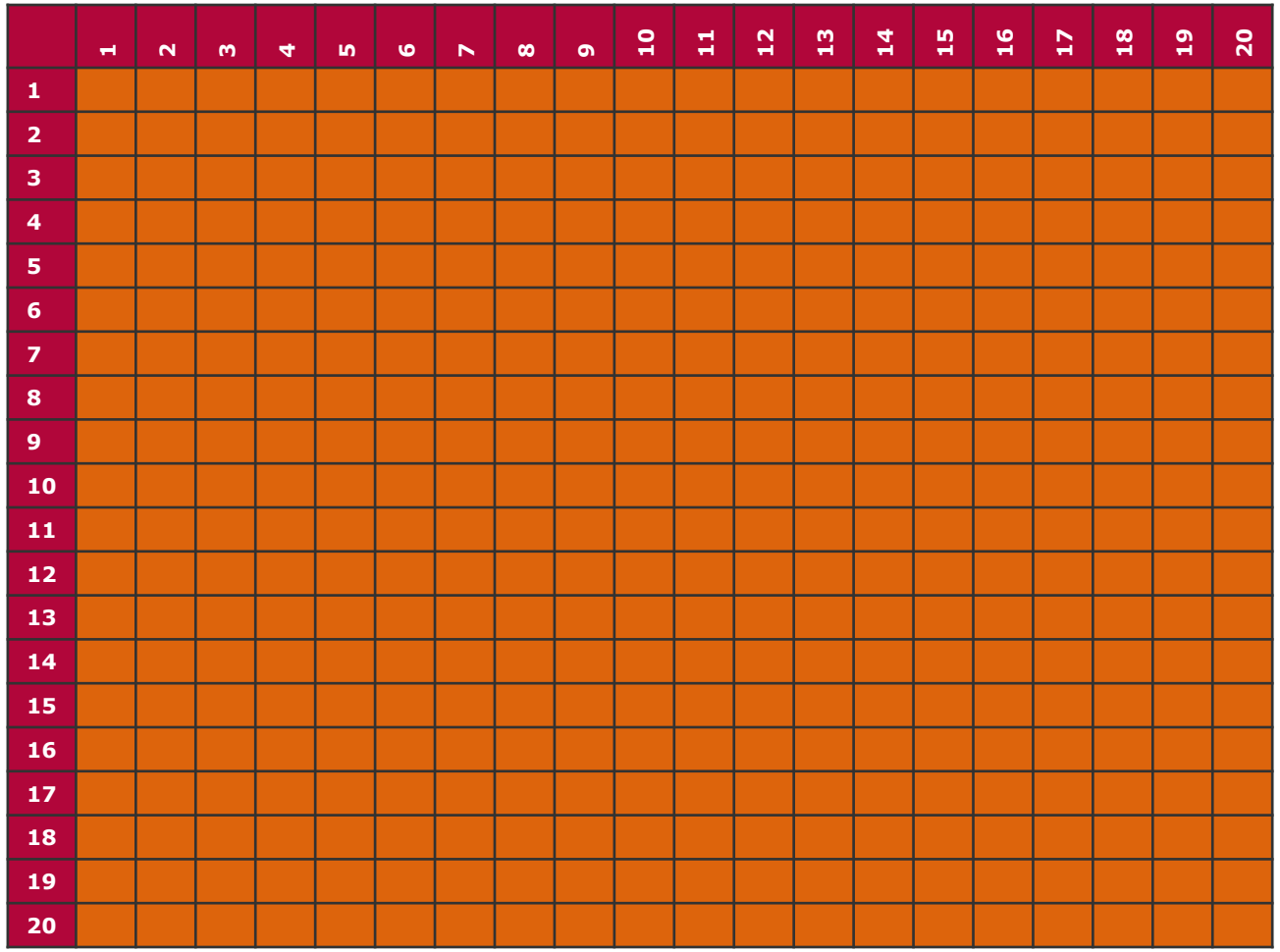


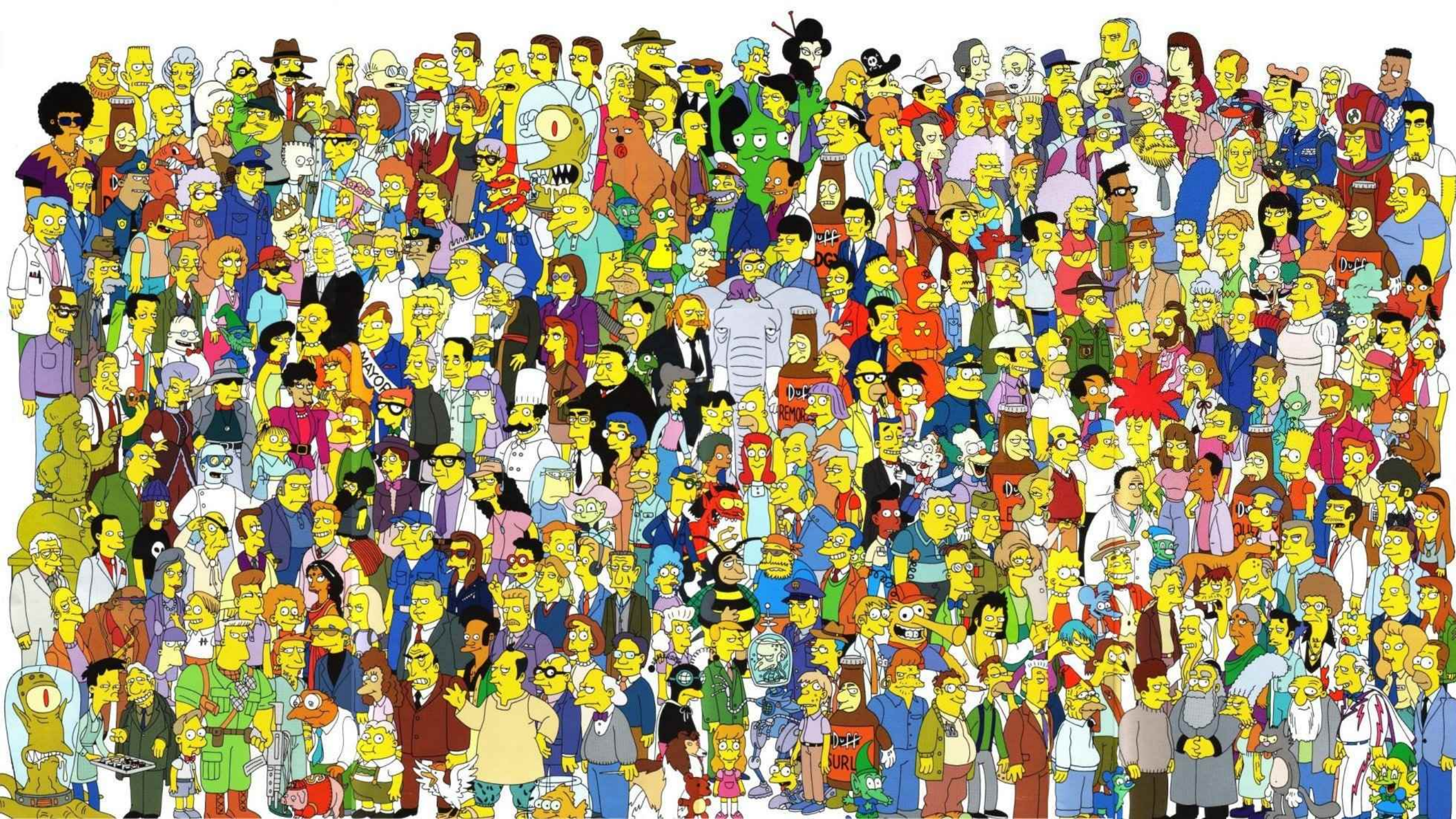
Felix Naumann
Bad Files, Bad Data 2021

Ironically, “Duplicate Detection” has many Duplicates

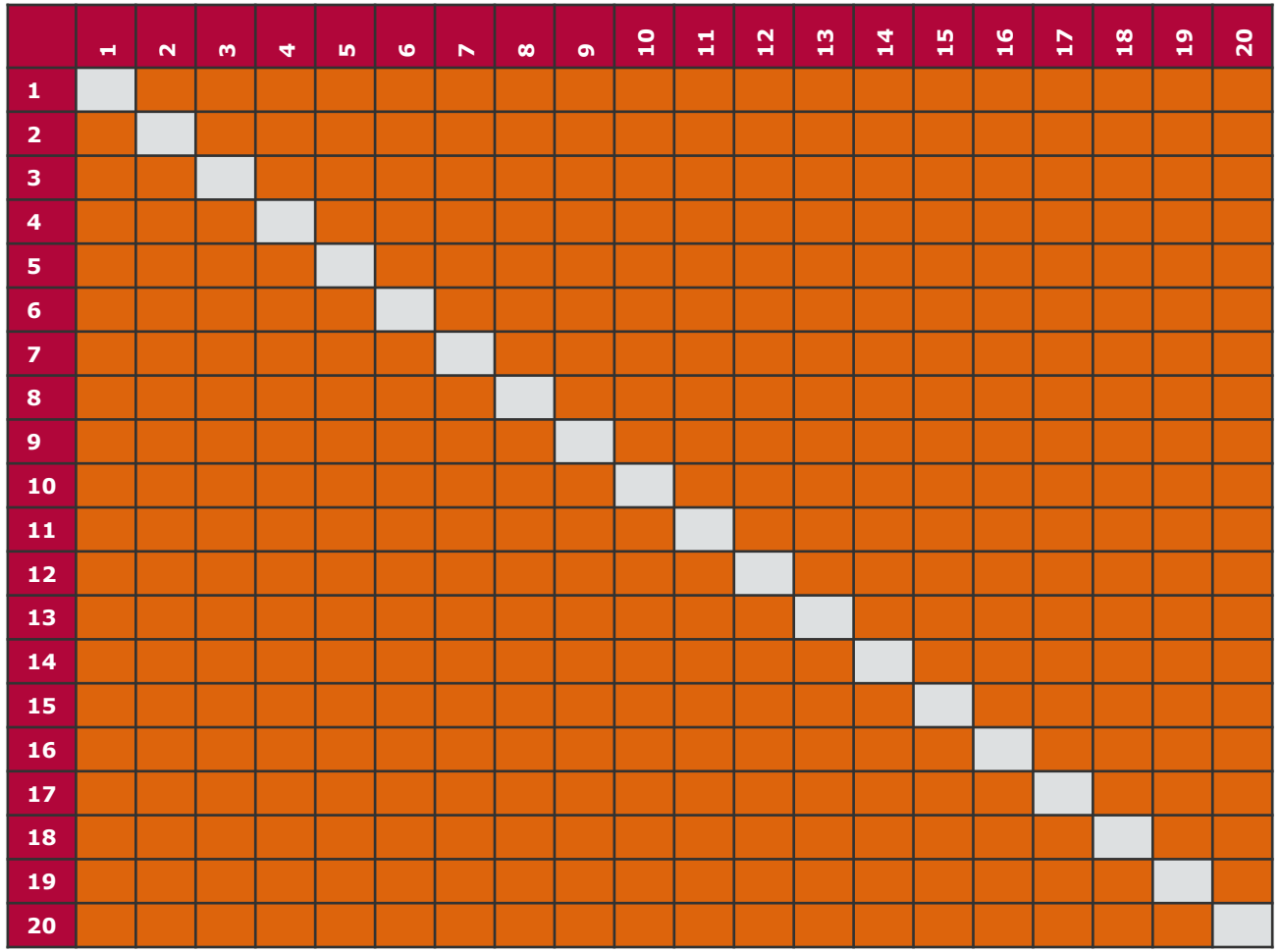


Number of comparisons: All pairs



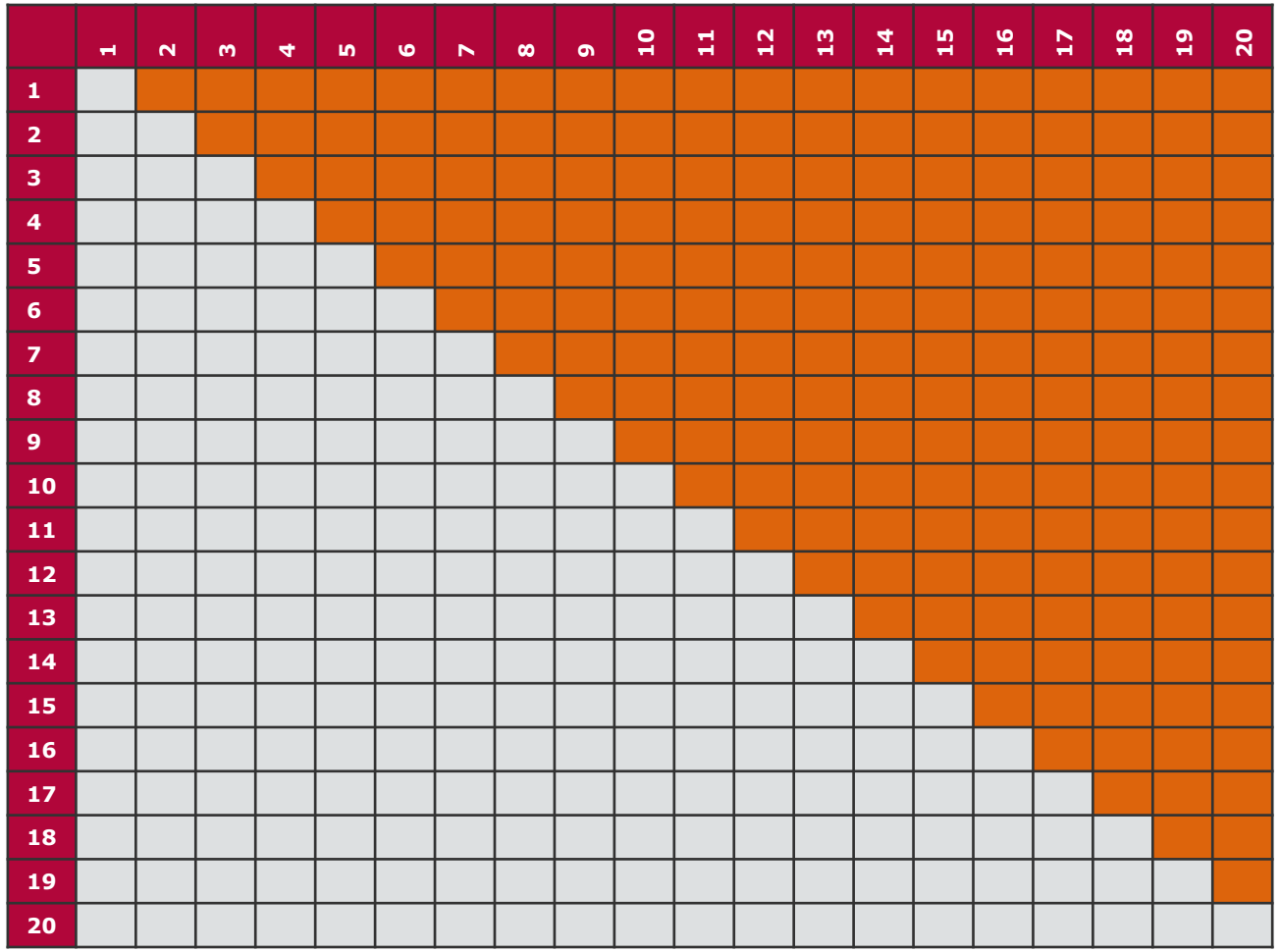

MAYOR
Duff
REMOR
Duff
SURL
Duff

Reflexivity of Similarity

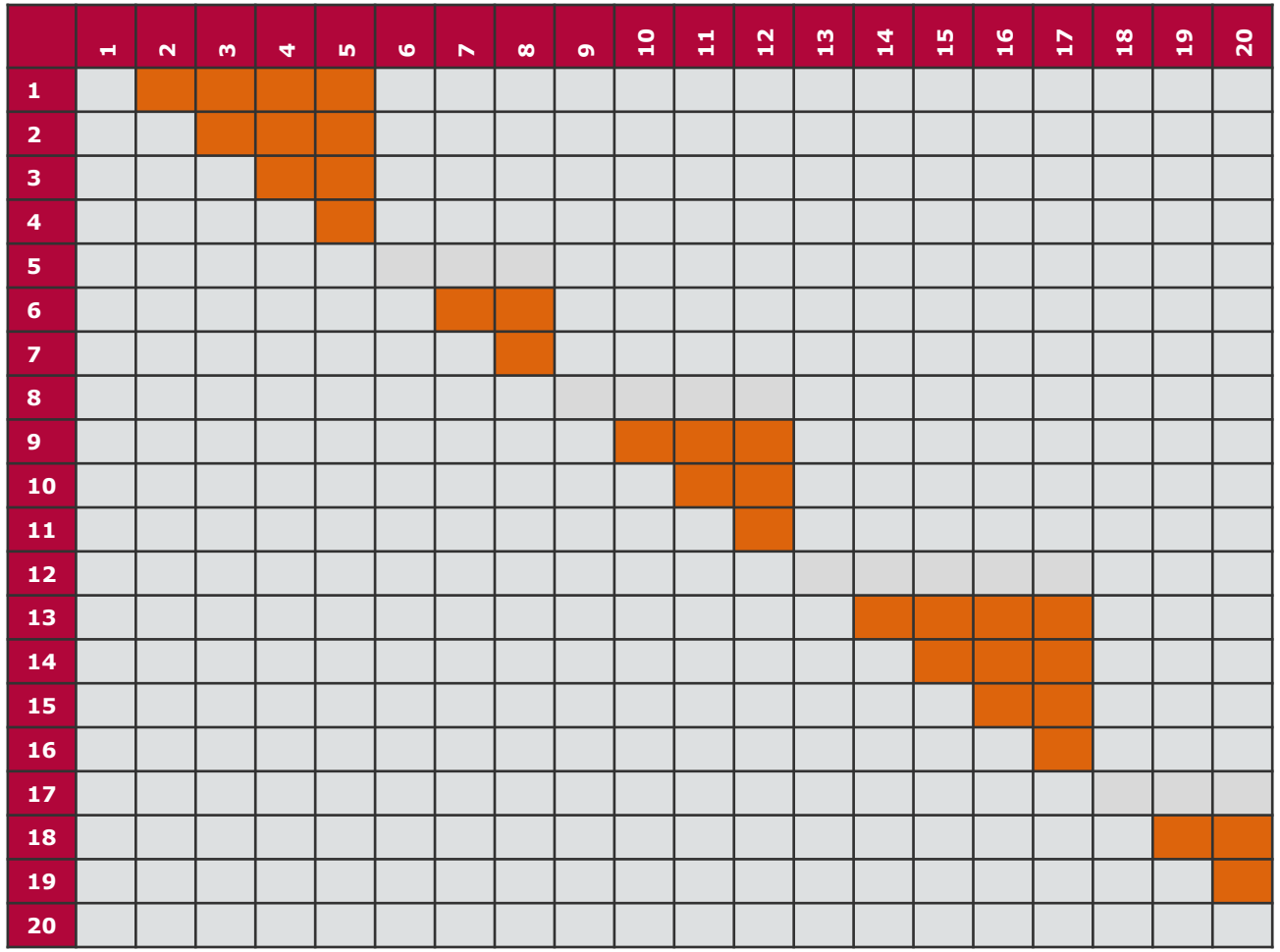


Felix Naumann
Bad Files, Bad Data 2021

Symmetry of Similarity

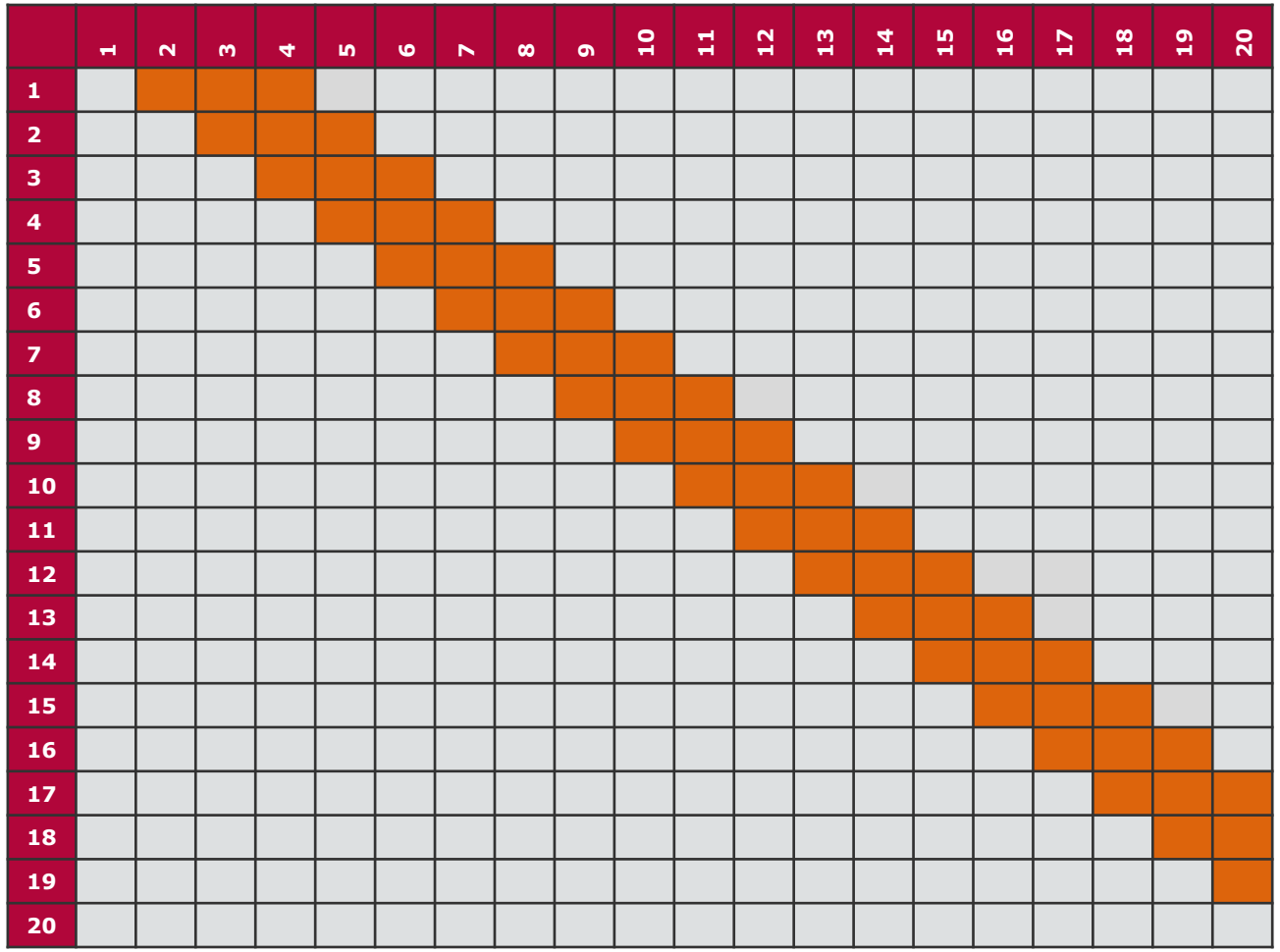


Blocking by zip-code

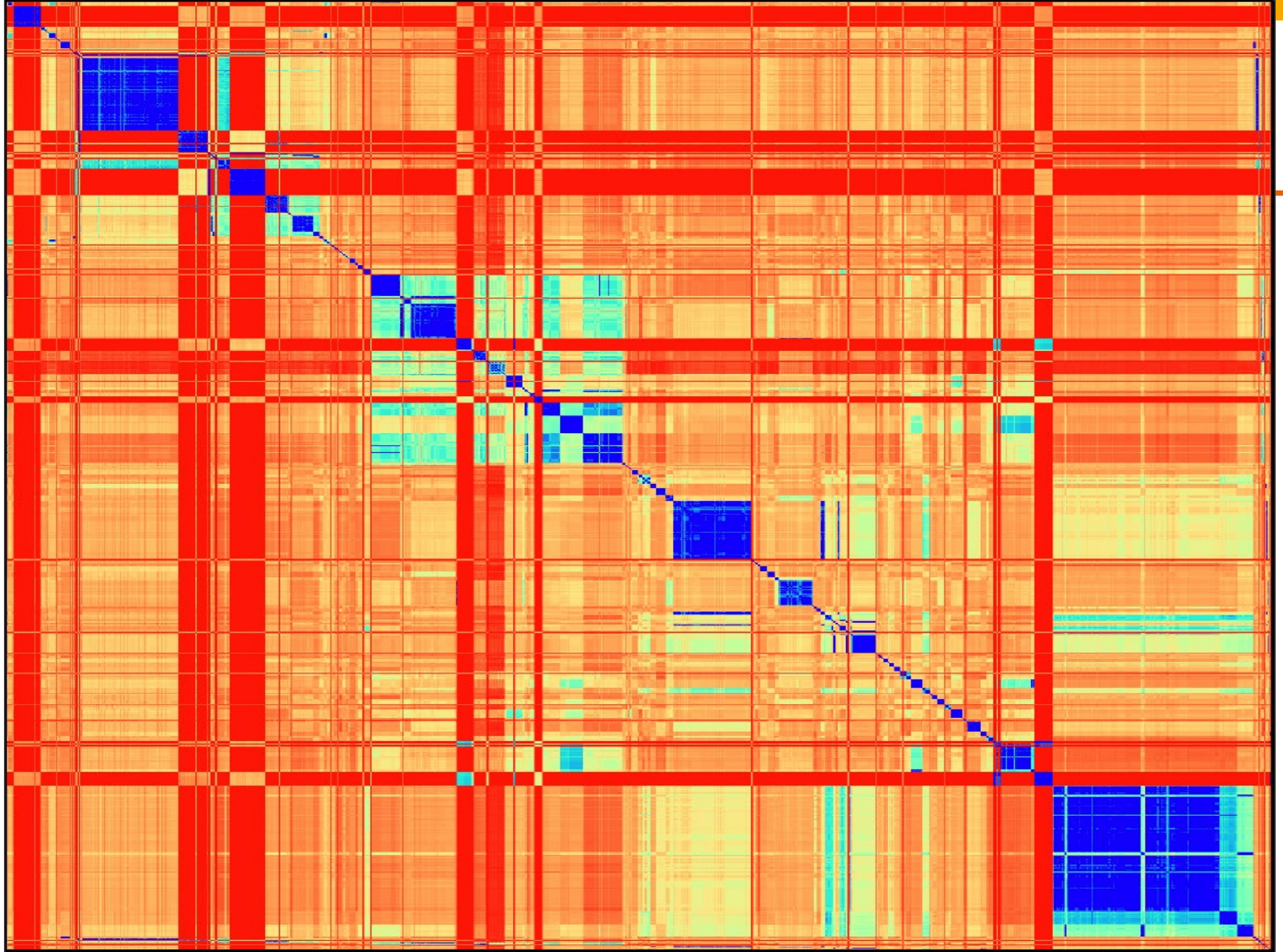


Felix Naumann
Bad Files, Bad Data 2021

Sorting by zip-code

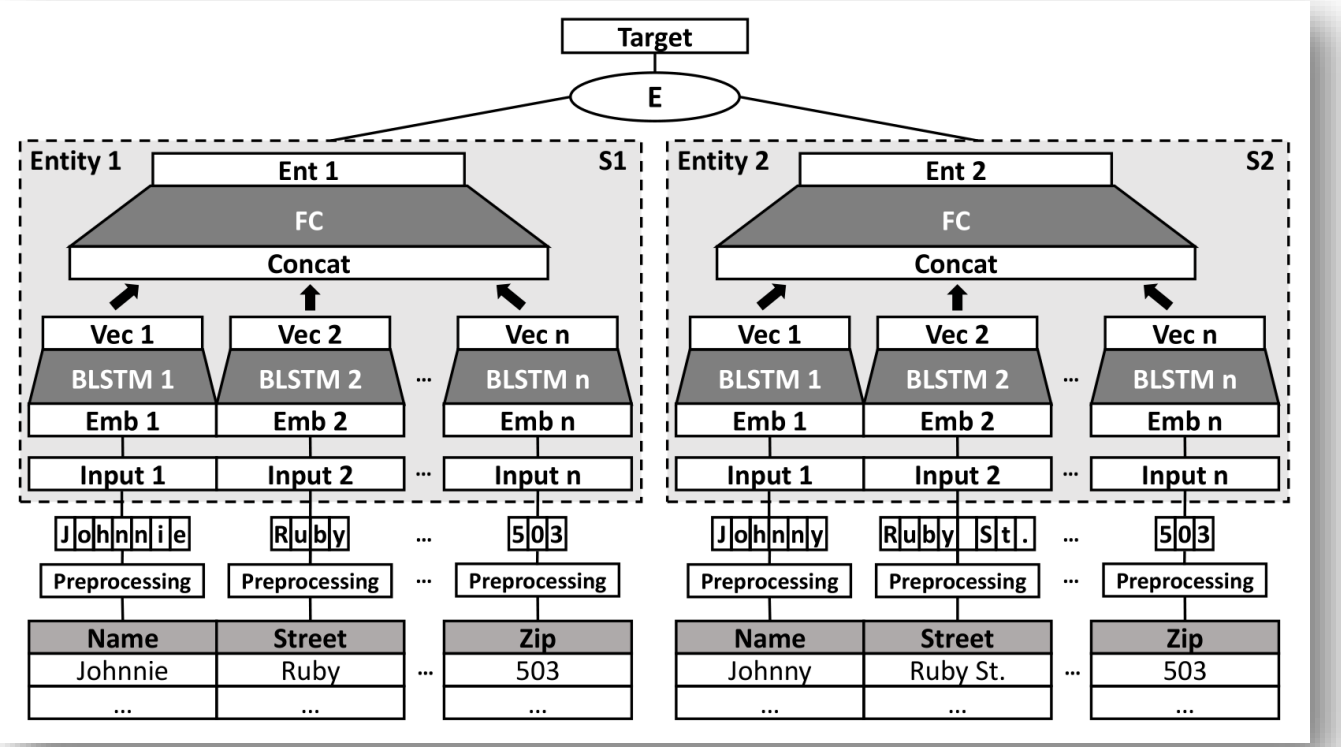


Felix Naumann
Bad Files, Bad Data 2021



Duplicate Detection Today

- Candidate selection
  - Blocking and windowing
  - Adaptive blocking and meta-blocking
- Duplicate classification
  - Rule-based approaches: explainability
  - ML-based approaches
    - Similarity measures as features
    - Learned similarity measure
    - Learned classifier
    - Active learning
- Data Fusion / Record Merging



Summary

1. Bad Files
2. Data Preparation
3. Bad Data
4. Data Cleaning

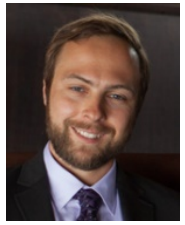
https://unsplash.com/photos/vGefUiWm0xI

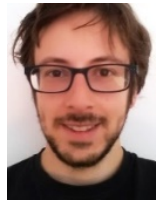
Information Systems Team

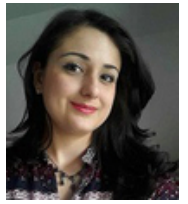

Dr. Thorsten **Papenbrock**


Diana **Stephan**


Prof. Felix **Naumann**

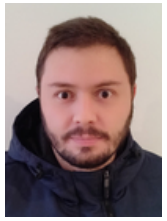

Dr. Ralf **Krestel**


Leon **Bornemann**


Hazar **Harmouch**


Phillip **Wenig**


Tim **Repke**


Alejandro **Sierra-Múnerah**


Michael **Loster**


Sebastian **Schmidl**


Tobias **Bleifuß**

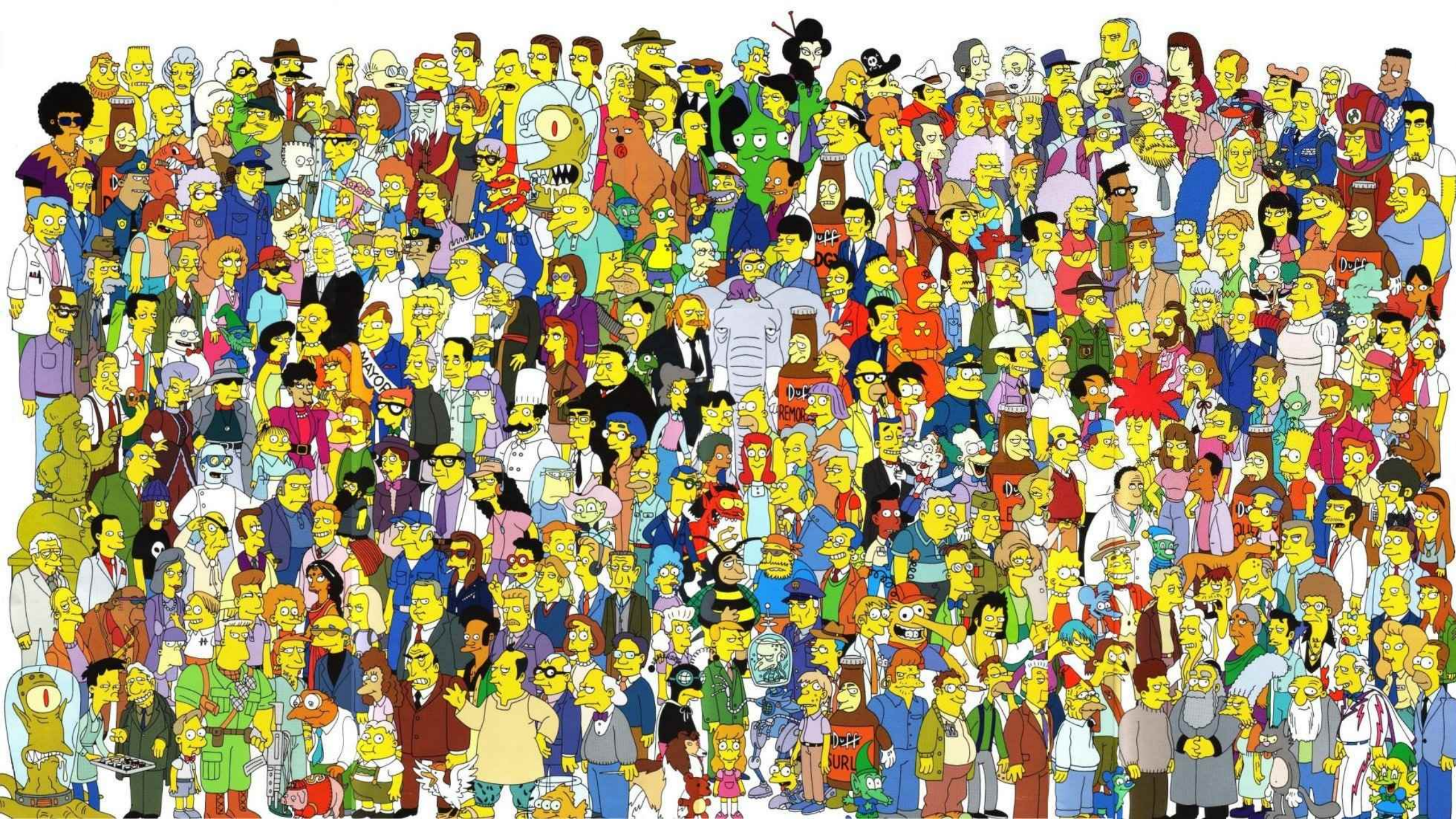

Nitisha **Jain**

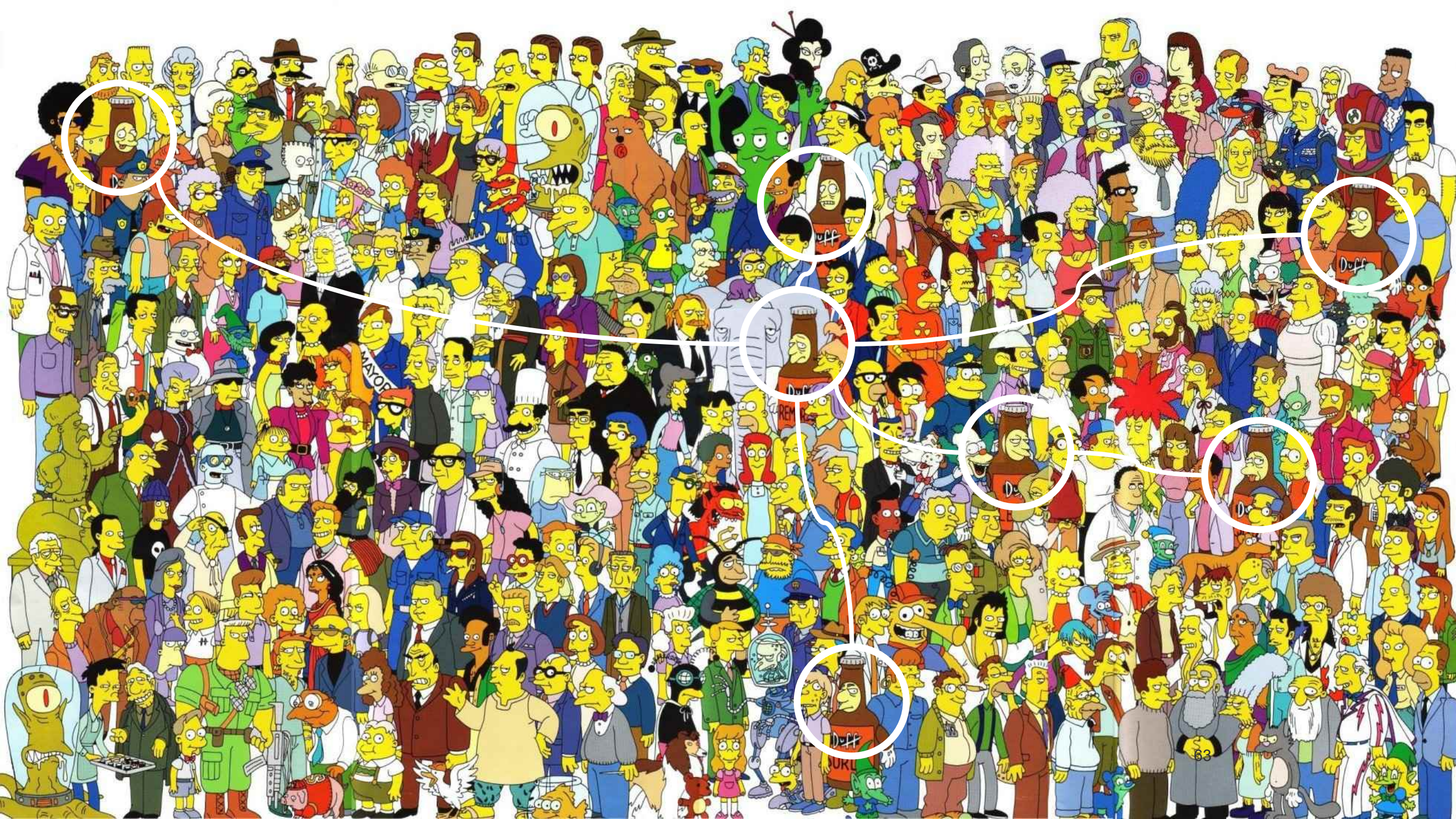

Lan **Jiang**


Mazhar **Hameed**


Gerardo **Vitagliano**

Duplicate Detection
Data Change
Data Fusion
project **DSV**
Entity Search
Data Profiling
Information Integration
project **DataKnoller**
Web Science
Data Scrubbing
project **DataChEx**
Data as a Service
Data Cleansing
Information Quality
Text Mining
Linked Open Data
CSV parsing
Dependency Detection
Knowledge Management for the Arts
Distributed Computing
project **Janus**
Web Data
Entity Recognition
Opinion Mining
Data Preparation
project **Metanome**
Change Exploration


MAYOR
Duff
REMOR
Duff
SURL
Duff


MAYOR
Duff
Duff
Duff
Duff
Duff