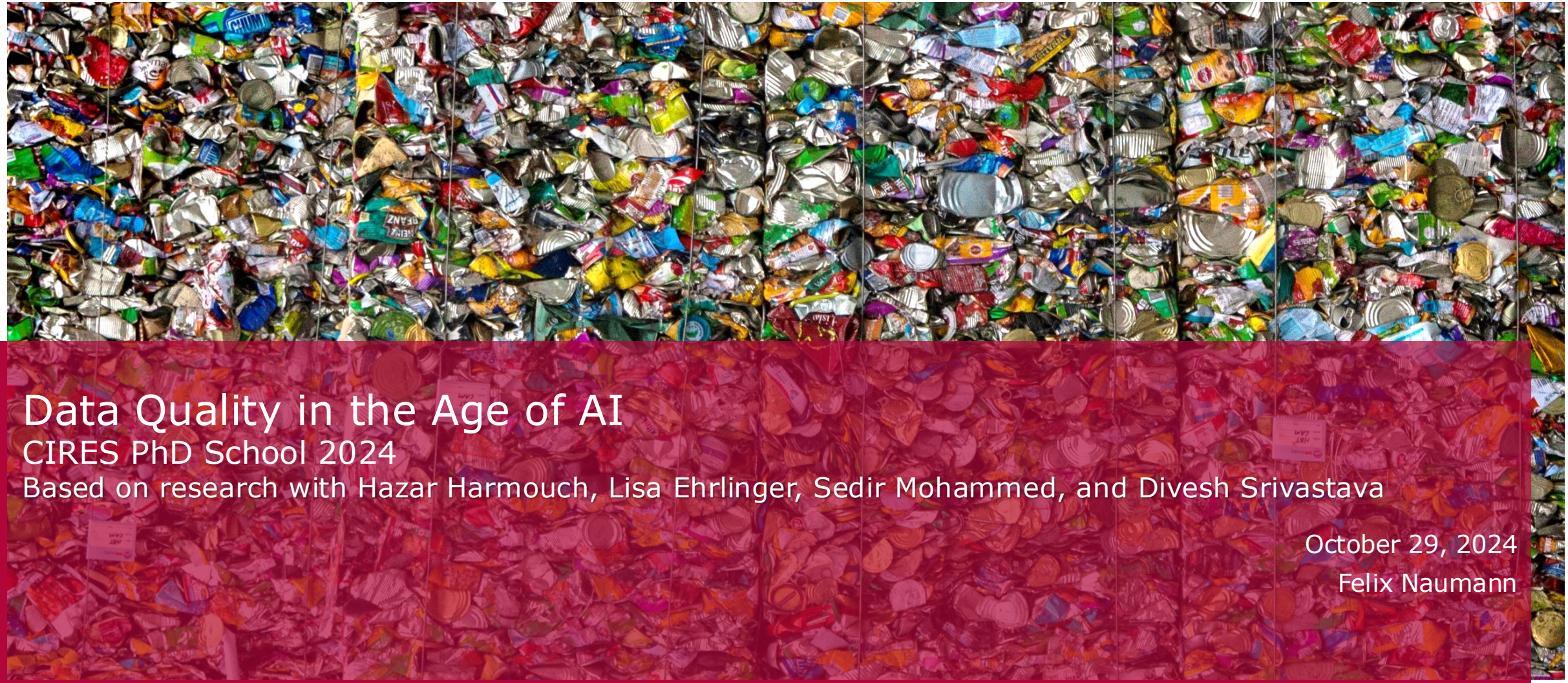

Datorios
things go
better
with
Cake
Small Mistake, Huge Difference.
Ensure Data Quality.

Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Data Quality in the Age of AI

CIRES PhD School 2024

Based on research with Hazar Harmouch, Lisa Ehrlinger, Sedir Mohammed, and Divesh Srivastava

October 29, 2024

Felix Naumann

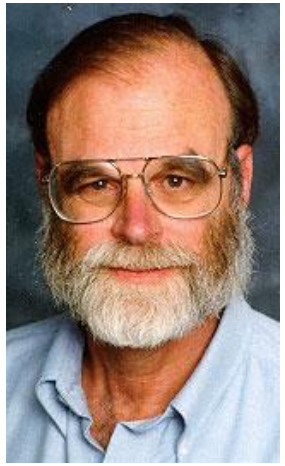
https://www.flickr.com/photos/136315829@N03/27640684175

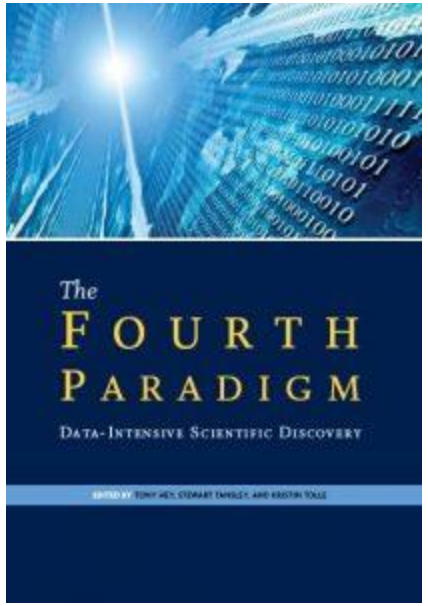
The Fourth Paradigm of Science

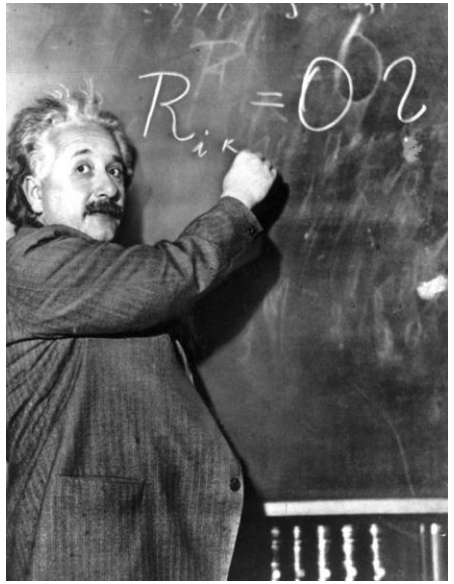
1. Empirical and experimental
2. Theoretical
3. Computational
4. Data-intensive

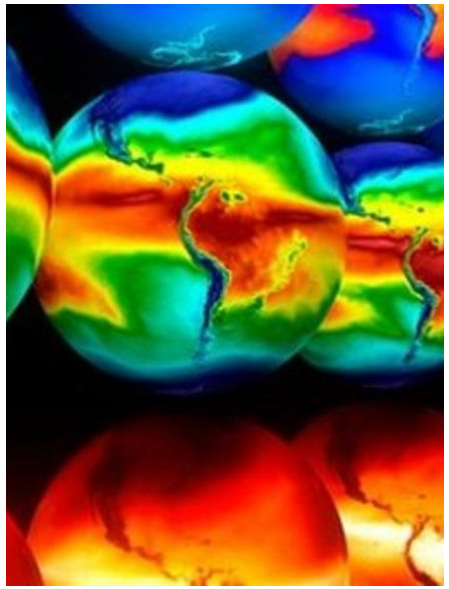
5. Intelligence-driven and knowledge-centric

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





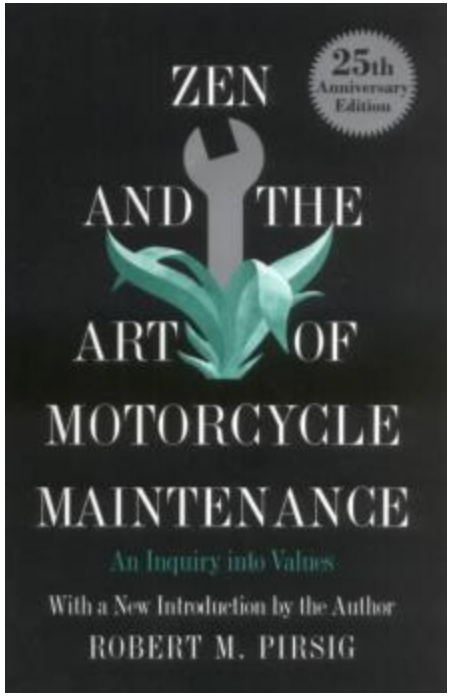




Quality

"Even though quality cannot be defined, you know what it is."

Robert Pirsig



Data Errors by Data Quality Researchers

Vandalism in Wikipedia Tables

No.		Mayor	Took Office	Left Office	Prior Experience	Deputy Mayor
62		Mel Lastman	January 1, 1998	November 30, 2003	Mayor of North York (1969–1997)	Case Ootes
63		David Miller	December 1, 2003	November 30, 2010	City Councillor for Parkdale-High Park (1994–2003)	Joe Pantalone
64		Rob				

−	l —Non-Hispanic ll **31**.7% ll **37**.9% ll 59.0% <ref name="fifteen">From 15% sample</ref> ll 91.2%	+	l —Non-Hispanic ll **20**.7% ll **21**.9% ll 59.0% <ref name="fifteen">From 15% sample</ref> ll 91.2%
	l-		l-
−	l [[African American|Black or African American]] ll **32**.9% ll **39**.1% ll **32**.7% ll 8.2%	+	l [[African American|Black or African American]] ll **50**.9% ll **49**.1% ll **42**.7% ll 8.2%

Example for vandalism in Wikipedia tables: Tampering with the proportions of ethnic minorities. [https://en.wikipedia.org/w/index.php?title=Chicago&diff=prev&oldid=654893961]

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

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
- Poor AI system performance

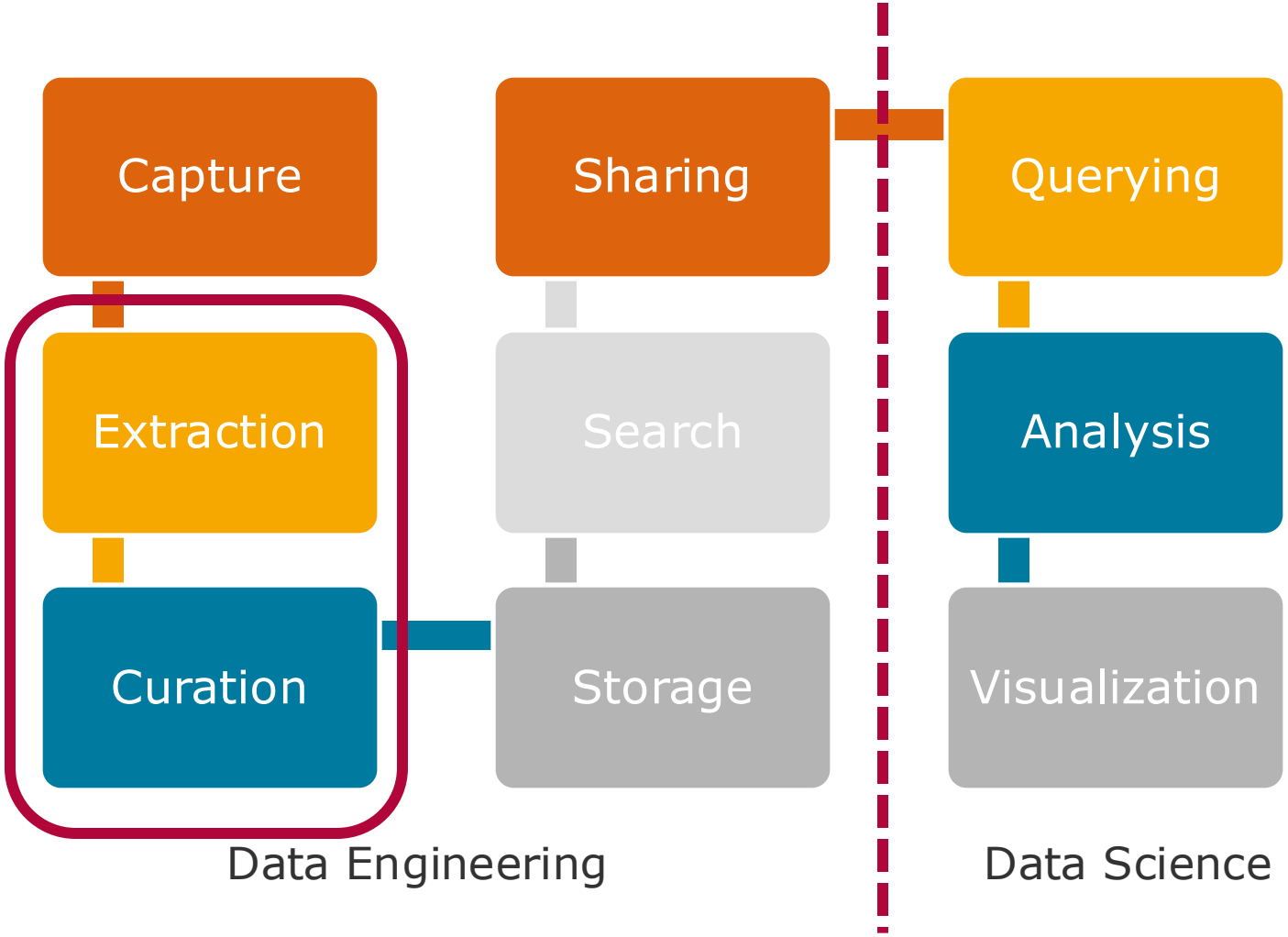
IRS might be after you — to mail you a check

Incorrect addresses stall nearly 1,500 Tennessee refunds

By BONNA de la CRUZ
Staff Writer

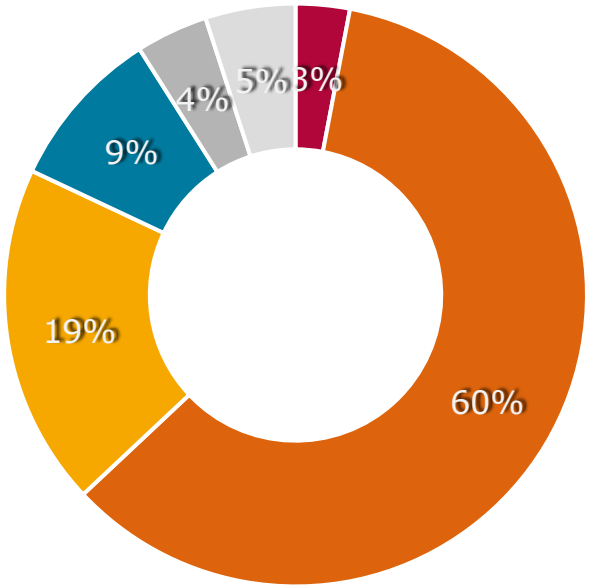
Now that Tilcia L. Menifee knows that she'll be getting $500 in a tax refund from Uncle Sam, she can do some Christmas shopping, she said.

Data Science Pipeline



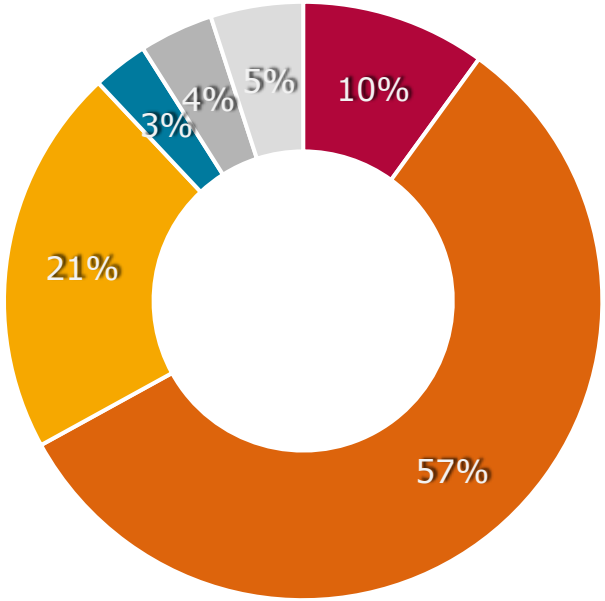
Data preparation in reality

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%
- Others: 5%

What is the least enjoyable part of data science?

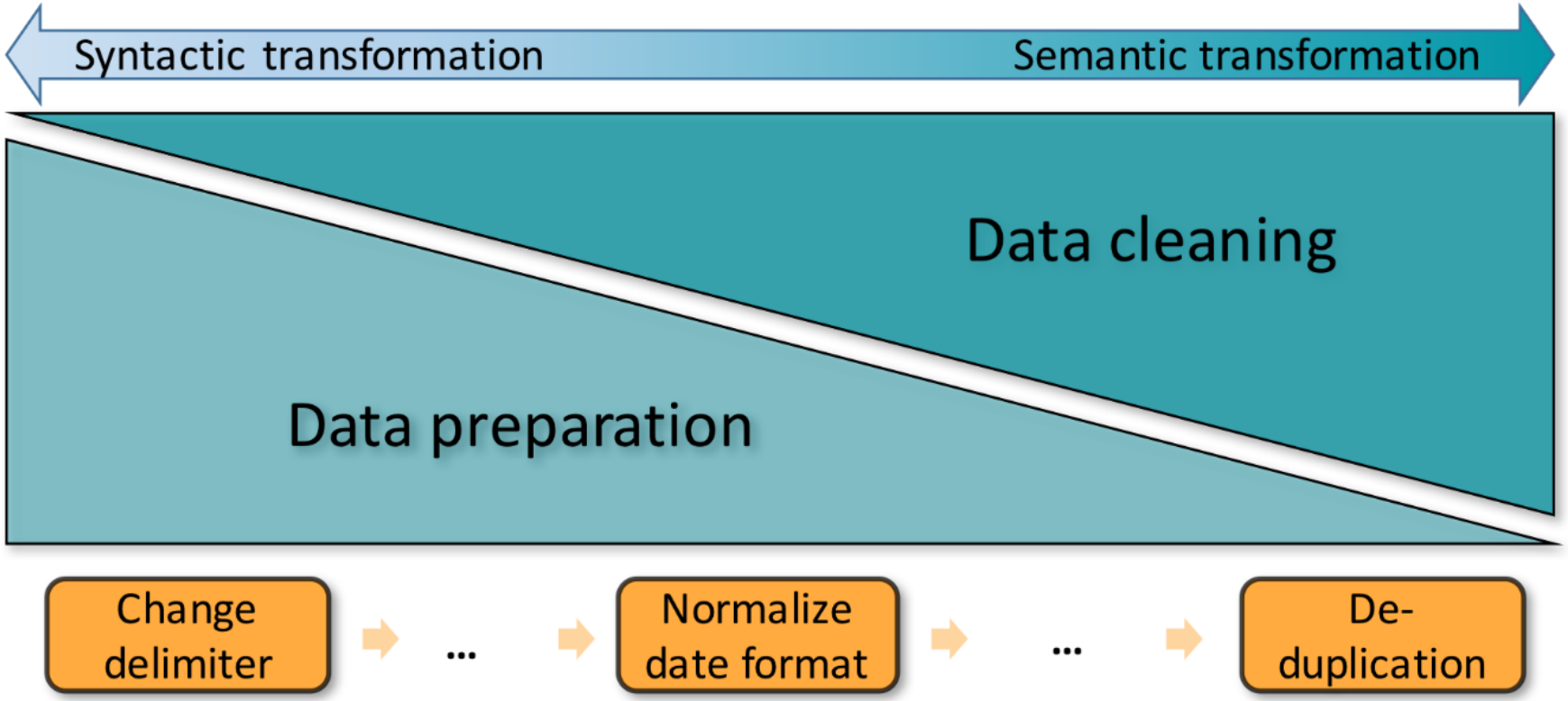


- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Others: 5%

"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/

Data Preparation vs. Data Cleaning

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



Agenda

1. **Data and Information Quality Research**
2. Data Preparation
3. Data Quality and AI Systems
4. Data Quality Assessment

Zooming into Information Quality

1

15

179

Fitness for use

Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

Ability to be Joined With	Ability to Download	Ability to Identify Errors	Ability to Upload
Acceptability	Access by Competitors	Accessibility	Accuracy
Adaptability	Adequate Detail	Adequate Volume	Aesthetics
Age	Aggregatability	Alterability	Amount of Data
Auditable	Authority	Availability	Believability
Breadth of Data	Brevity	Certified Data	Clarity
Clarity of Origin	Clear Data Responsibility	Compactness	Comparability
Competitive Edge	Completeness	Comprehensiveness	Compressibility
Concise	Conciseness	Confidentiality	Conformity
Consistency	Content	Context	Continuity
Convenience	Correctness	Corruption	Cost
Cost of Accuracy	Cost of Collection	Creativity	Critical
Current	Customizability	Data Hierarchy	Data Improves Efficiency
Data Overload	Definability	Dependability	Depth of Data
Detail	Detailed Source	Dispersed	Distinguishable Updated Files
Dynamic	Ease of Access	Ease of Comparison	Ease of Correlation
Ease of Data Exchange	Ease of Maintenance	Ease of Retrieval	Ease of Understanding
Ease of Update	Ease of Use	Easy to Change	Easy to Question
Efficiency	Endurance	Enlightening	Ergonomic
Error Free	Expandability	Expense	Extendibility

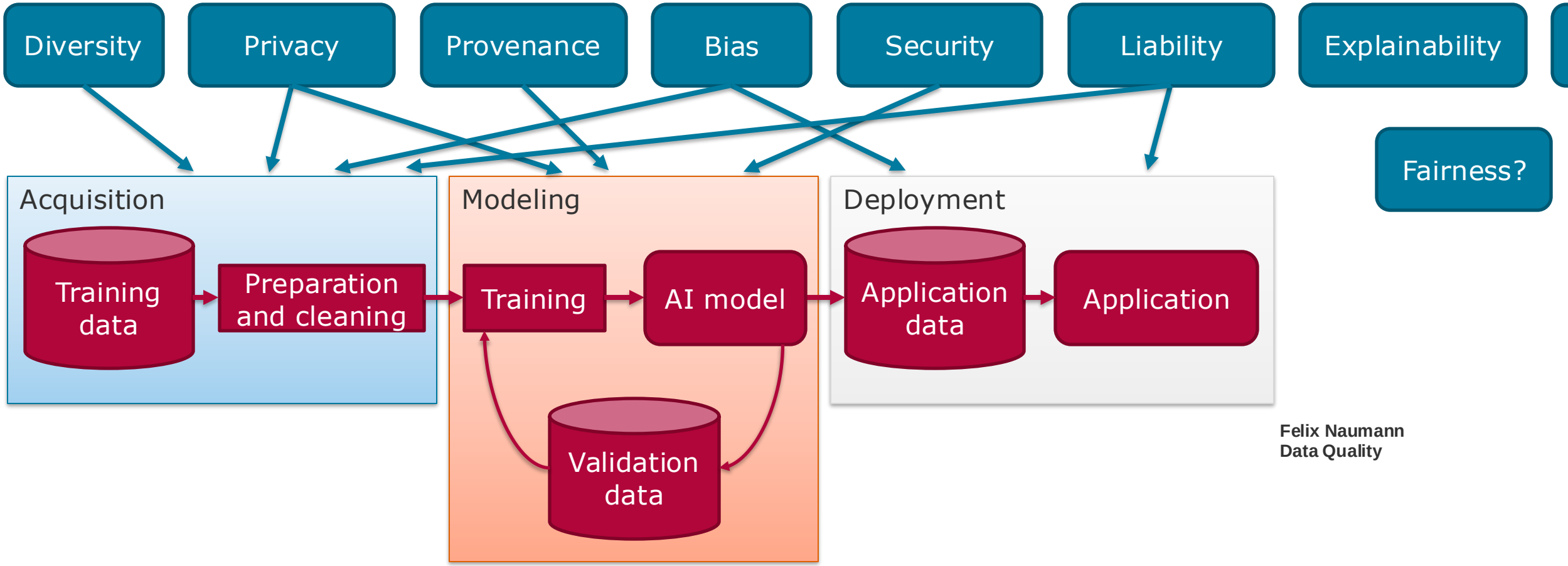
179 Dimensions

Extensibility	Extent	Fabrication	Flawlessness
Flexibility	Form of Presentation	Formal	Integrity
Friendliness	Generality	Habit	Historical Comparability
Importance	Inconsistencies	Integration	Integrity
Interactive	Interesting	Level of Abstraction	Level of Standardization
Localized	Logically Connected	Manageability	Manipulable
Measurable	Medium	Meets Requirements	Minimality
Modularity	Narrowly Defined	No Lost Information	Normality
Novelty	Objectivity	Optimality	Orderliness
Origin	Parsimony	Partitionability	Past Experience
Pedigree	Personalized	Pertinent	Portability
Preciseness	Precision	Proprietary Nature	Purpose
Quantity	Rationality	Redundancy	Regularity of Format
Relevance	Reliability	Repetitive	Reproducibility
Reputation	Resolution of Graphics	Responsibility	Relevability
Revealing	Reviewability	Rigidity	Robustness
Scope of Info	Secrecy	Security	Self Correcting
Semantic Interpretation	Semantics	Size	Source
Specificity	Speed	Stability	Storage
Synchronization	Time-independence	Timeliness	Traceable
Translatable	Transportability	Unambiguity	Unbiased
Understandable	Uniqueness	Unorganized	Up-to-Date
Usable	Usefulness	User Friendly	Valid
Value	Variability	Variety	Verifiable
Volatility	Well-Documented	Well-Presented	

Felix Naumann
Data Quality

Wang, R. Y. & Strong, D. M.
Beyond Accuracy: What data quality means to data consumers
Management of Information Systems, 1996, *12(4)*, 5-34

New AI-specific Data Quality Dimensions



28 DQ Dimensions

Uniqueness
Relevance
Privacy
Understandability
Accuracy
Security
Believability
Consistent representation
Added value
Fairness
Representativity
Precision
Diversity
Reputation
Documentation
Transparency
Consistency
Reliability
Balance
Traceability
Completeness
Portability
Accessibility
Amount
Recoverability
Timeliness
Compliance
Efficiency
Ease of manipulation

Felix Naumann
Data Quality

https://arxiv.org/pdf/2403.00526

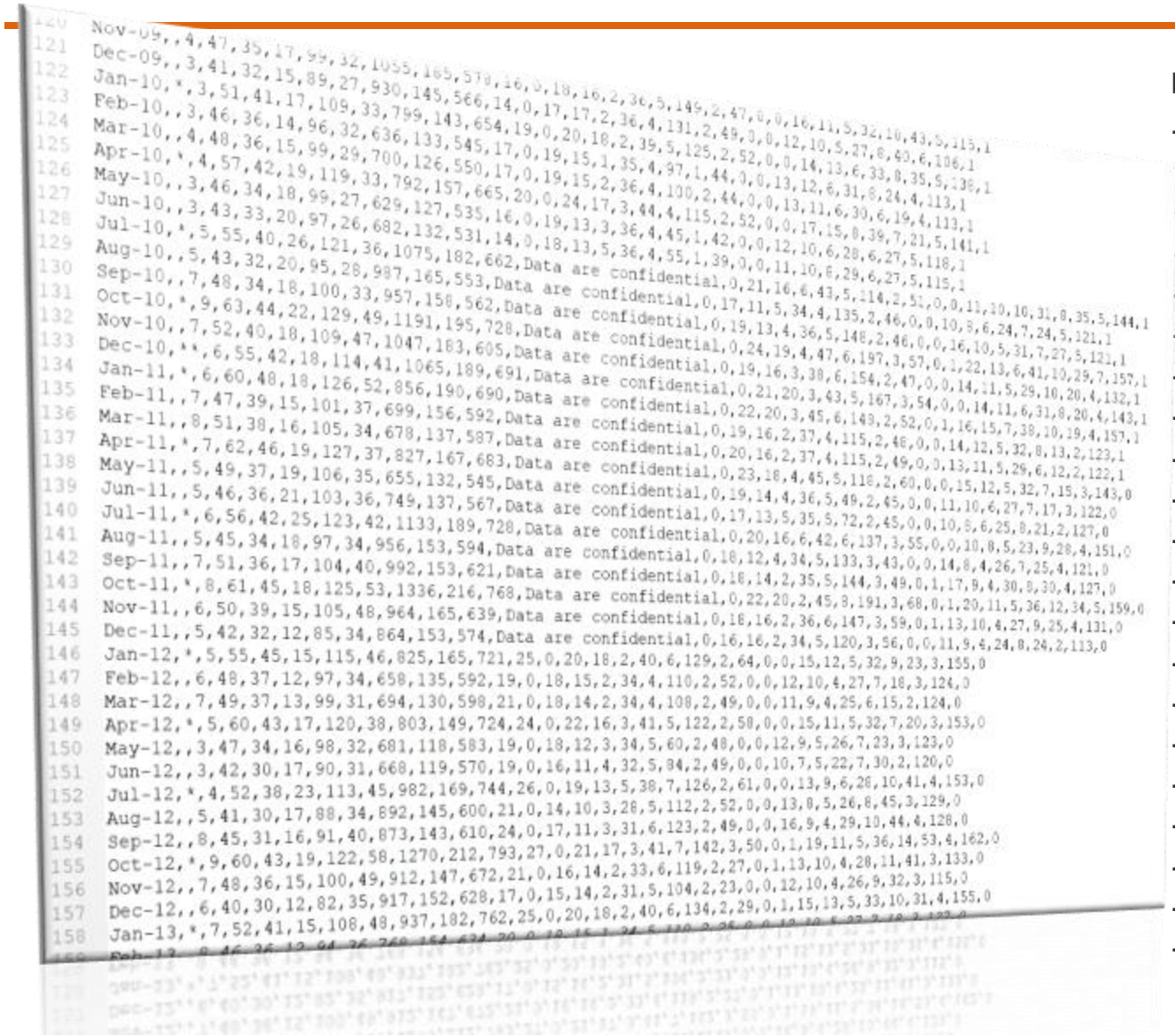
Agenda

1. Data and Information Quality Research
2. **Data Preparation**
3. Data Quality and AI Systems
4. Data Quality Assessment

```
"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"""
"17,""Paavo Johannes Aaltonen"",""M"",28,175,64,""Finland"",""FIN"",""1952 Summer"",1952,""Summer"",""Helsinki"",""Gymnastics"",""Gymnastics Men's Individual All-Around"",NA"
```

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**	**73**	**3**	**3**	**35**	1	8	26	**4**	**38**	29	
258	2011 Q2	Total	*Total	**3.456**	**3.963**	**5.156**	**2.130**	**3.963**	**4.488**	154	1.525	2.809	**1.326**	**668**	301	
259	2011 Q2	Africa	*Total Africa	**703**	**611**	**970**	**377**	**611**	**811**	28	357	426	**326**	**159**	62	
260	2011 Q2	Americas	*Total Americas	**343**	**1.367**	**652**	**301**	**1.367**	**643**	55	194	394	**42**	**9**	0	
261	2011 Q2	Asia	*Total Asia	**1.790**	**888**	**2.892**	**1.006**	**888**	**2.526**	61	812	1.653	**784**	**366**	193	
262	2011 Q2	Europe	*Total Europe	**512**	**638**	**356**	**418**	**638**	**318**	9	92	217	**94**	**38**	7	
263	2011 Q2	Middle East	*Total Middle East	**98**	**192**	**240**	**22**	**192**	**144**	1	47	96	**76**	**96**	39	
264	2011 Q2	Oceania	*Total Oceania	**4**	**153**	**38**	**3**	**153**	**38**	0	16	22	**1**	**0**	0	
265	2011 Q2	Other	*Total Other	**6**	**114**	**8**	**3**	**114**	**8**	0	7	1	**3**	**0**	0	
266	2011 Q2	Asia	Afghanistan	**296**	**70**	**69**	**17**	**70**	**4**	0	3	1	**279**	**65**	34	
267	2011 Q2	Europe	Albania	**100**	**187**	**25**	**53**	**187**	**19**	0	11	8	**47**	**6**	2	
268	2011 Q2	Africa	Algeria	**49**	**32**	**42**	**17**	**32**	**21**	0	11	10	**32**	**21**	3	
269	2011 Q2	Oceania	American Samoa	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	**0**	**0**	0	
270	2011 Q2	Europe	Andorra	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	**0**	**0**	0	
271	2011 Q2	Africa	Angola	**7**	**19**	**12**	**1**	**19**	**4**	0	0	4	**6**	**8**	4	
272	2011 Q2	Americas	Anguilla (British)	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	**0**	**0**	0	
273	2011 Q2	Americas	Antigua and Barbuda	**0**	**6**	**1**	**0**	**6**	**1**	0	1	0	**0**	**0**	0	
274	2011 Q2	Americas	Argentina	**4**	**30**	**4**	**4**	**30**	**4**	1	2	1	**0**	**0**	0	
275	2011 Q2	Europe	Armenia	**2**	**3**	**1**	**1**	**3**	**0**	0	0	0	**1**	**1**	0	
276	2011 Q2	Americas	Aruba	**0**	**0**	**0**	**0**	**0**	**0**	0	0	0	**0**	**0**	0	
277	2011 Q2	Oceania	Australia	**1**	**120**	**24**	**1**	**120**	**24**	0	8	16	**0**	**0**	0	
278	2011 Q2	Europe	Austria	**2**	**1**	**0**	**2**	**1**	**0**	0	0	0	**0**	**0**	0	

Data Preparation for AI: The Challenge



My data won't load …

… because nobody bothered to use escape symbols.

… because ` is not a proper quotation symbol.

… because the maximum line length is exceeded.

… because there is a header row.

… because there is no header row.

… because the first line is the table-name.

… because some lines are empty.

… because it is encoded in CP-1252.

… because columns are shifted every ten rows.

… because a numeric column contains a string in line 590450.

… because some lines are two fields shorter.

… because Ümlauts are not supported.

… because someone added footnotes.

… because who uses § as a delimiter?

… because the file contains multiple tables.

… because tab and space are not the same thing.

… because someone added a comment in line 3.

… because – is not -.

… because it is split across multiple files.

… because headers are repeated every 80 lines.

… because the file ends mid-row.

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 expressions	✓	✓	✓	✓	✓	✓	
	Discover & merge external data	✓	✓	✓			✓	✓
	Duplicate column	✓	✓	✓		✓	✓	✓
	Generate primary key column			✓				✓
	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		✓	✓	✓	✓	✓	✓

Selected Data Preparation Projects – Bringing Order to Files

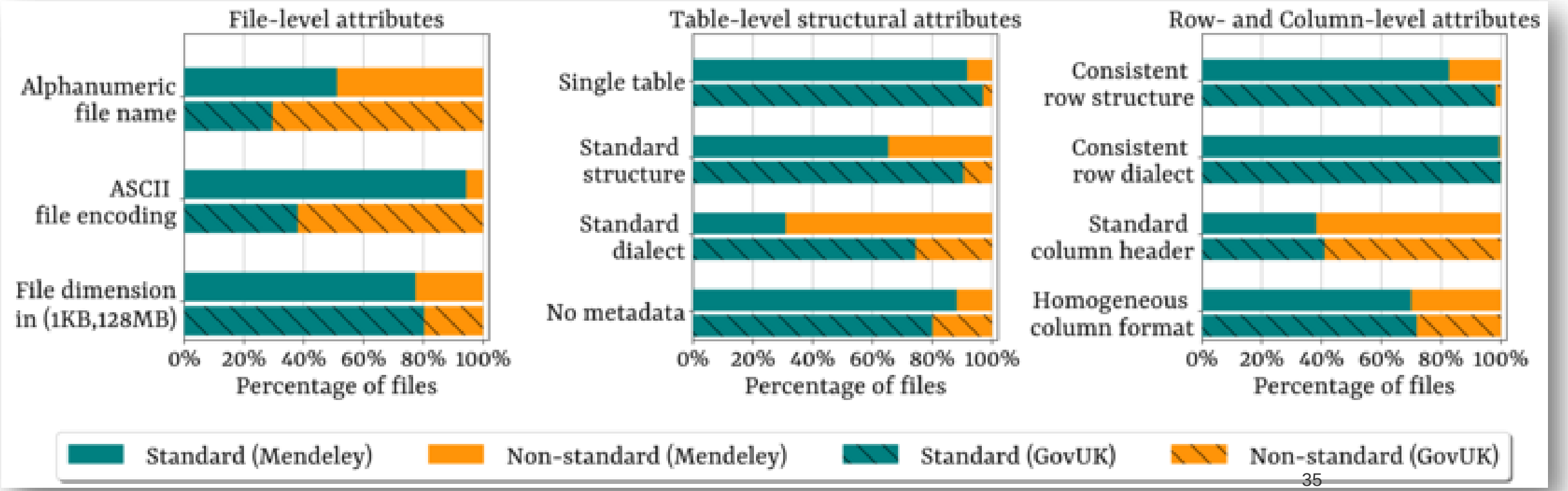
- Mondrian
  - Dissecting multi-table files
- ExtracTable
  - Parsing visually delimited files
- Suragh and Tasheeh
  - Identifying ill-formed records
- Strudel
  - Classify cell-types
- AggreCol
  - Identify aggregation cells
- Pollock benchmark
  - Evaluate data ingestion ability

Pollock: Benchmarking the Ingestion Ability of Systems

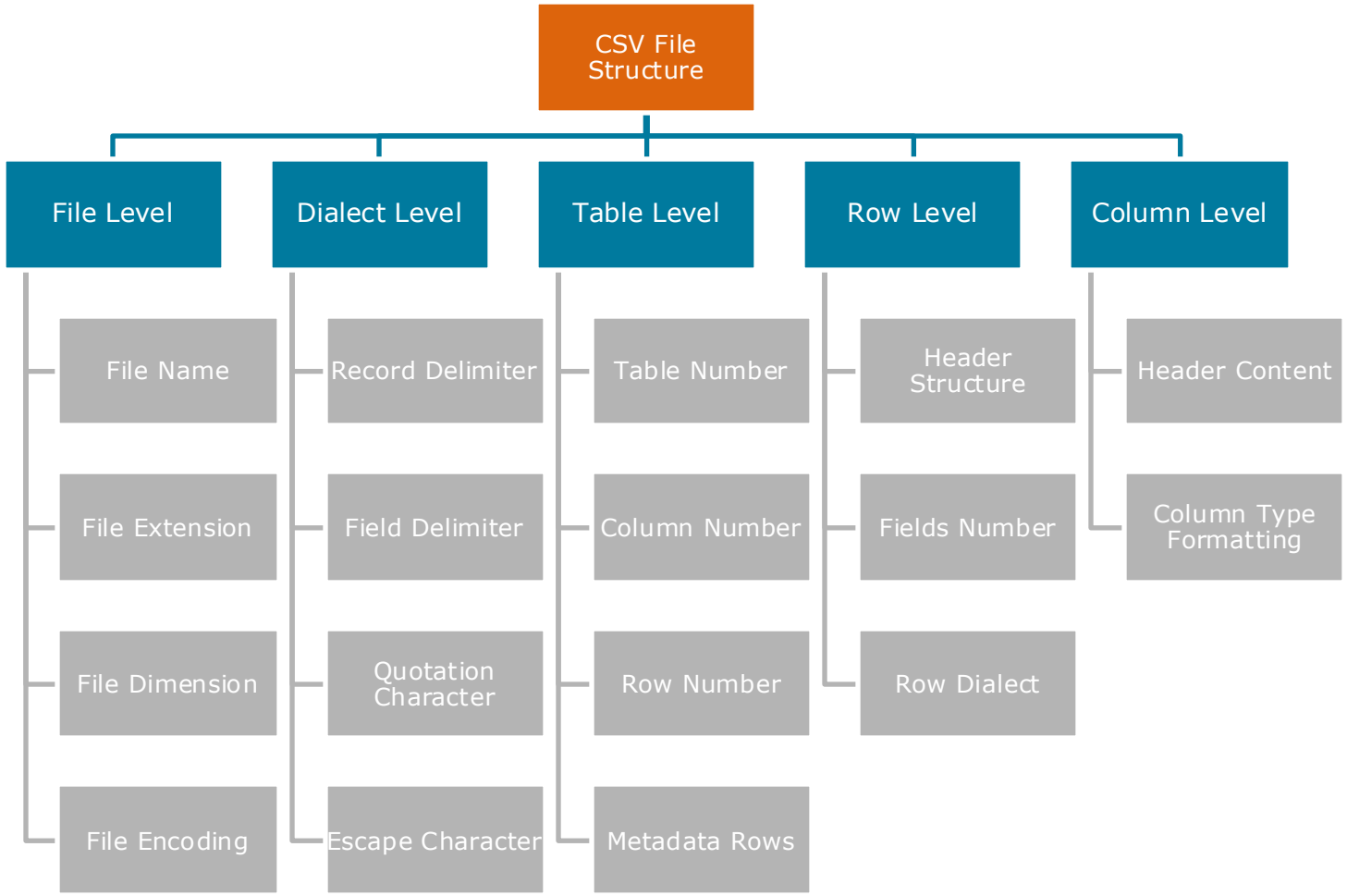
```
Python 3.8.5 (default, Sep  3 2020, 21:29:08) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import pandas as pd
>>> pd.read_csv("11-708-data-nlss-2009-1.csv")
Traceback (most recent call last):
  File "<stdin>", line 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: Raw Data Survey

- Manual Annotation
  - 1,438 random files from GovUK
  - 2,274 random files from Mendeley



Pollock: Benchmark Dimensions and Results



	Pollock score (2 289 +1) files	
	Simple	Weighted
CleverCSV 0.7.4	**9.05**	**9.49**
CSVCommons 1.9.0	**6.63**	**9.29**
Hypoparsr 0.1.0	**3.73**	**4.41**
OpenCSV 5.6	**6.62**	**7.80**
Pandas 1.4.3	**9.88**	**9.75**
PyCsv 3.10.5	**9.71**	**9.47**
RCsv 4.2.1	**7.78**	**6.76**
Univocity 2.9.1	**9.35**	**7.97**
MariaDB 10.9.3	**8.81**	**7.44**
MySQL 8.0.31	**8.88**	**7.45**
PostgreSQL 15.0	**0.14**	**7.33**
SQLite 3.39.0	**9.94**	**9.73**
Calc 7.3.6	**9.75**	**7.52**
SpreadDesktop	**9.79**	**9.29**
SpreadWeb	**9.65**	**9.29**
DataViz	**4.93**	**5.51**

Agenda

1. Data and Information Quality Research
2. Data Preparation
3. **Data Quality and AI Systems**
   - With Hazar Harmouch, Sedir Mohammed et al.
4. Data Quality Assessment

Empirical Measurement of the Effects of Poor Data Quality on ML Results: Measurement Dimensions

Pollutions

- Consistent representation
- Completeness
- Feature accuracy
- Target accuracy
- Uniqueness
- Target balance

Scenarios

- Pollute only training data
- Pollute only test data
- Pollute training and test data

Runs

- 5 runs, average

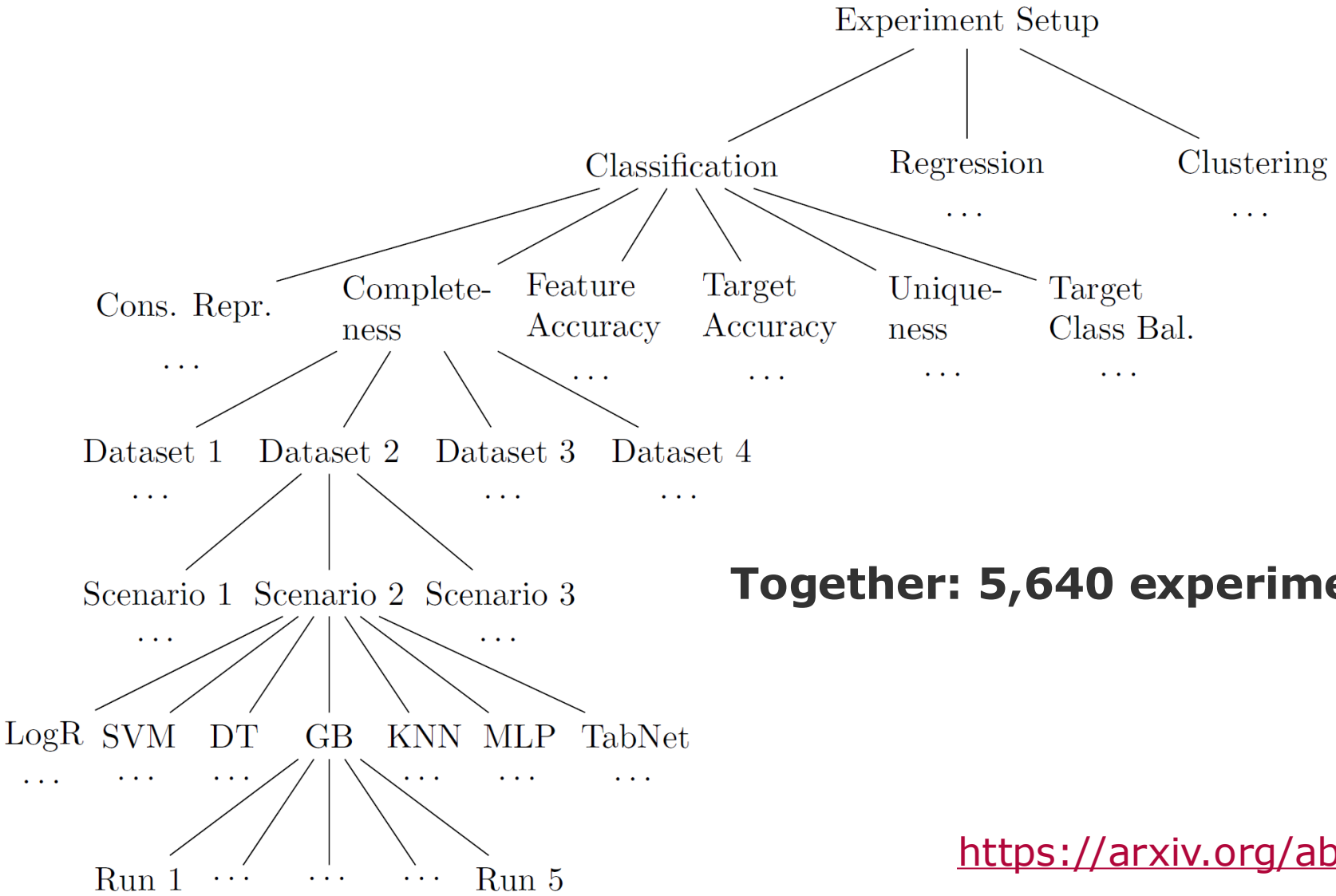
Tasks and algorithms

- Classification
  - LogR, SVM, DT, GB, KNN, MLP
- Clustering
  - GM, k-Means, k-Prototypes, AC, OPTICS
- Regression
  - LR, RR, DT, RF, GB, MLP, TabNet

Datasets

- TelcoChurn, GermanCredit, Contraceptive, COVID
- Houses, IMDB, Cars
- Bank, Covertype, Letter

Empirical Measurement of the Effects of Poor Data Quality on ML Results: Measurement Dimensions

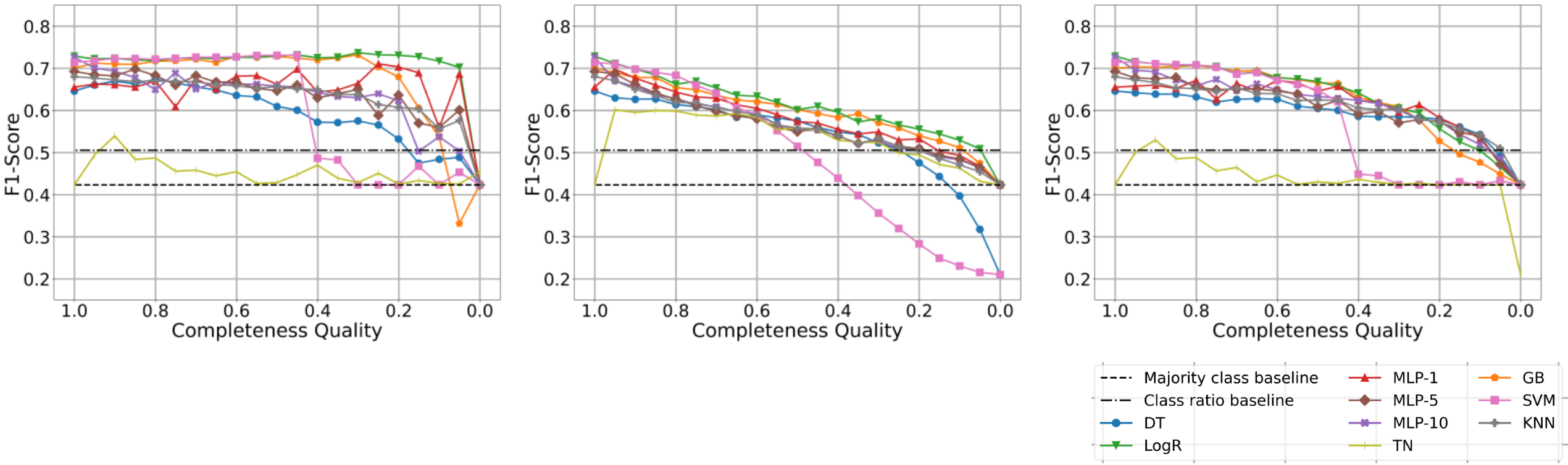


Together: 5,640 experiments

Felix Naumann
Data Quality

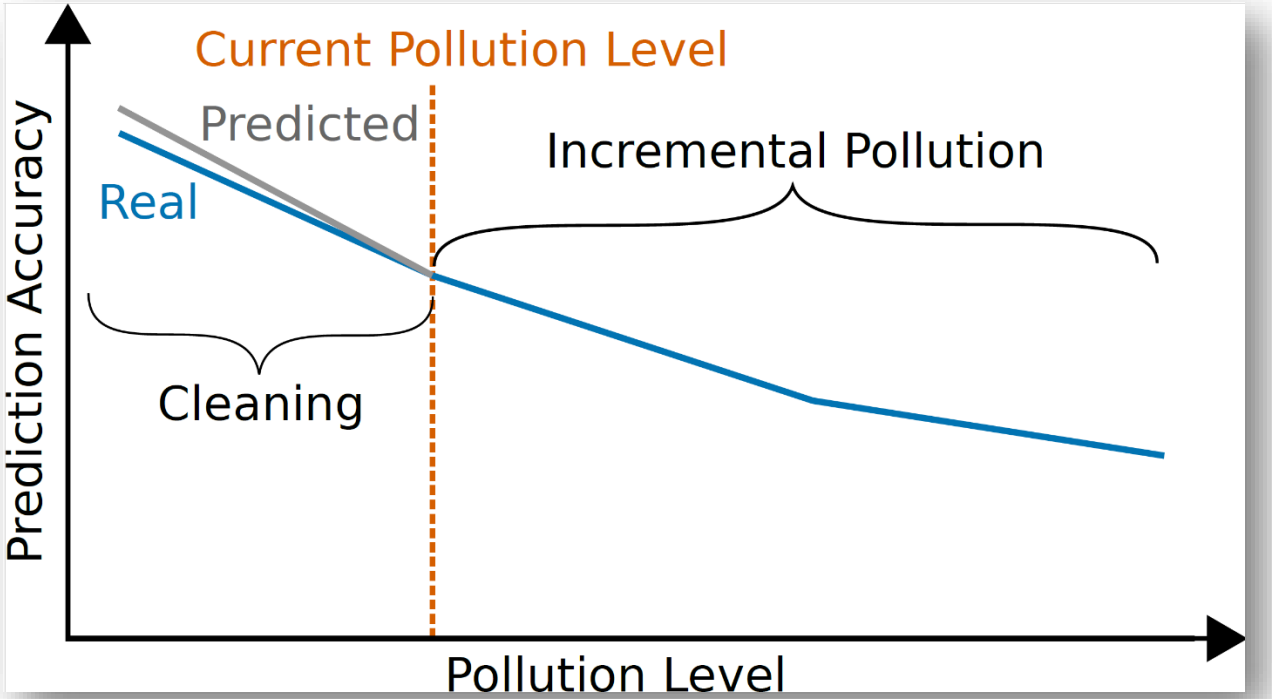
https://arxiv.org/abs/2207.14529

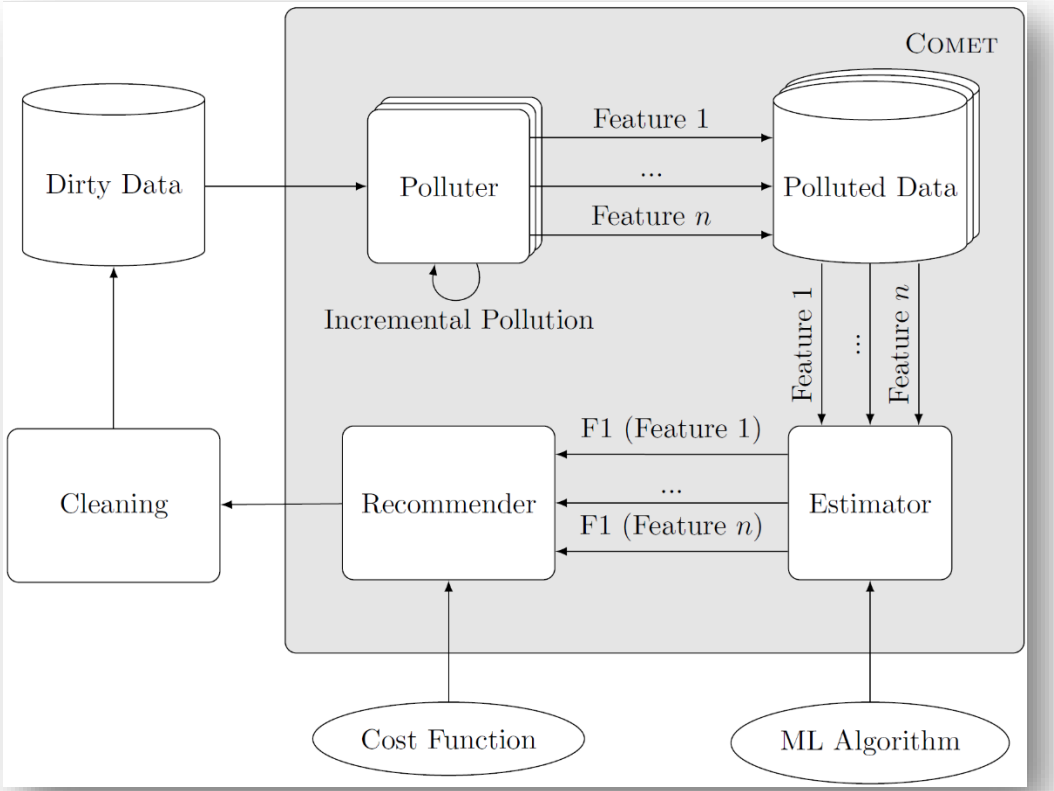
Example Results



Average F1-Score for Classification of the Telco-Churn dataset

So what?
Recommend Data Cleaning Steps





Agenda

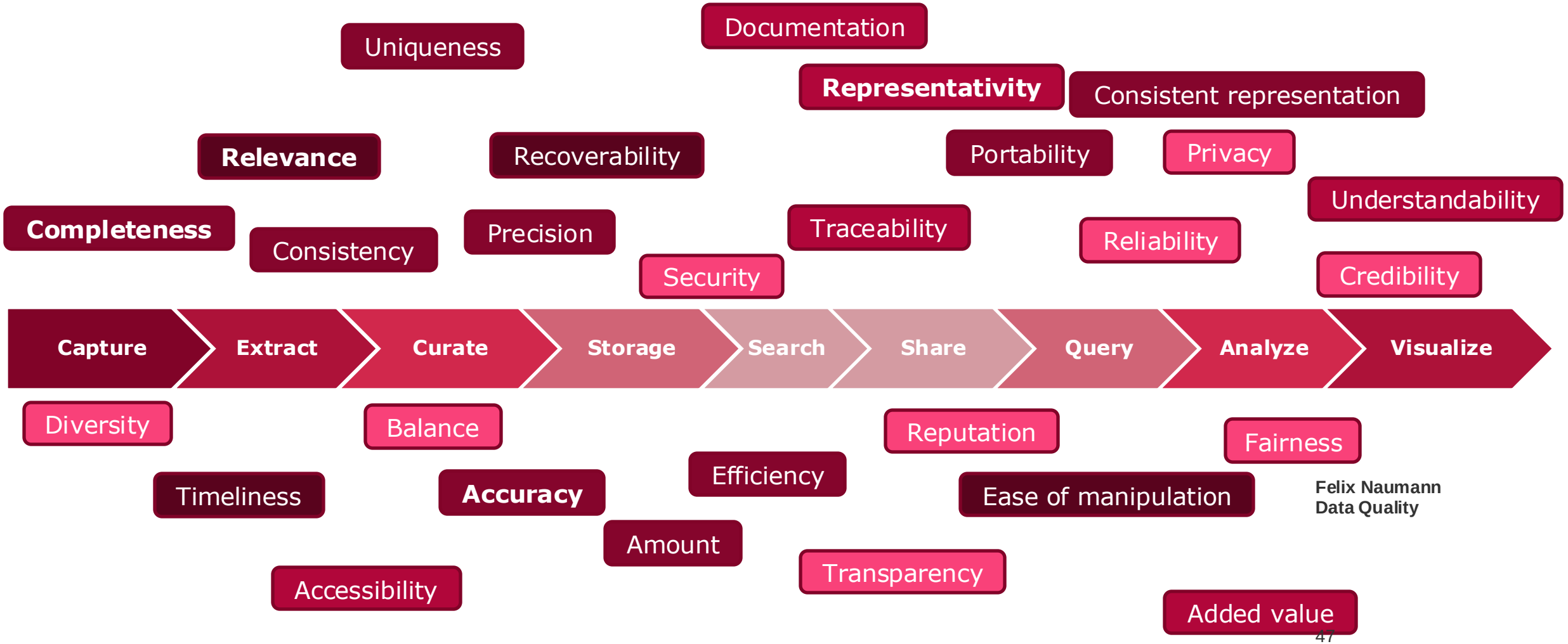
1. Data and Information Quality Research
2. Data Preparation
3. Data Quality and AI Systems
4. **Data Quality Assessment**
   - With Hazar Harmouch, Lisa Ehrlinger, Sedir Mohammed and Divesh Srivastava

European AI Act Article 10 (3): Data and Data Governance

- High-quality data and access to high-quality data plays a vital role in providing structure and in ensuring the performance of many AI systems, especially when techniques involving the training of models are used, with a view to ensure that the high-risk AI system performs as intended and safely and it does not become a source of discrimination prohibited by Union law.
- High-quality data sets for training, validation and testing require the implementation of appropriate data governance and management practices.
- Data sets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system.
- The data sets should also have the appropriate statistical properties, including as regards the persons or groups of persons in relation to whom the high-risk AI system is intended to be used, with specific attention to the mitigation of possible biases in the data sets […].

https://artificialintelligenceact.eu/article/10/

Data Quality along the AI Pipeline



DQ Assessment in the Year 2000

[c5] Felix Naumann, Claudia Rolker:
Assessment Methods for Information Quality Criteria. IQ 2000: 148-162

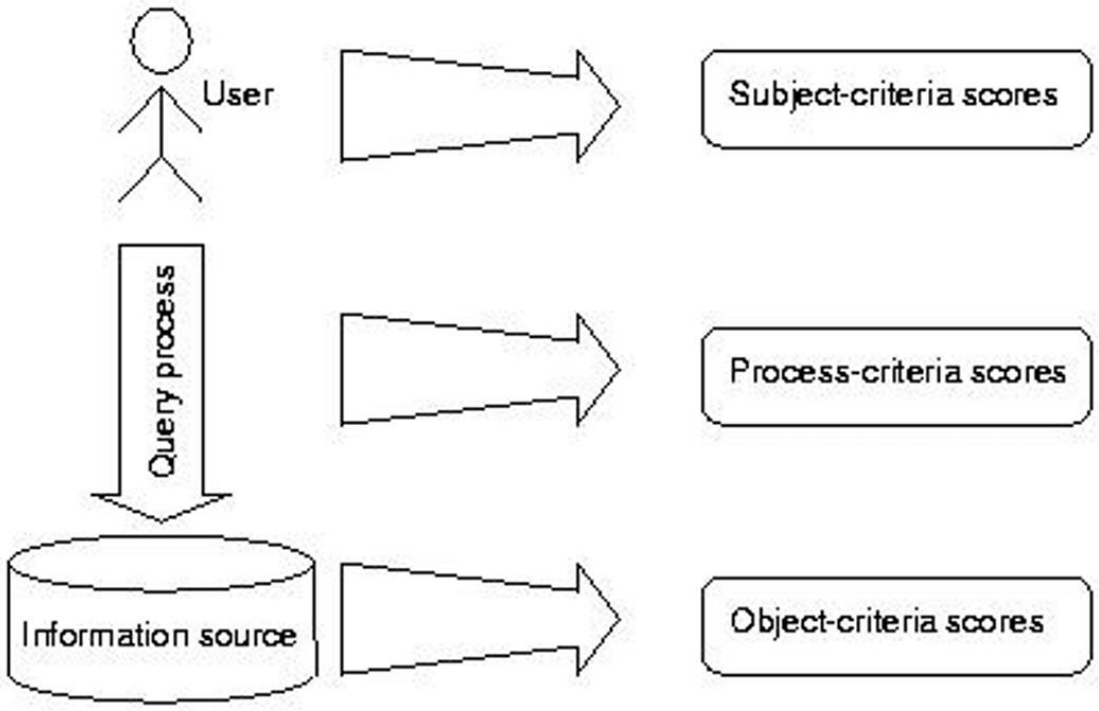
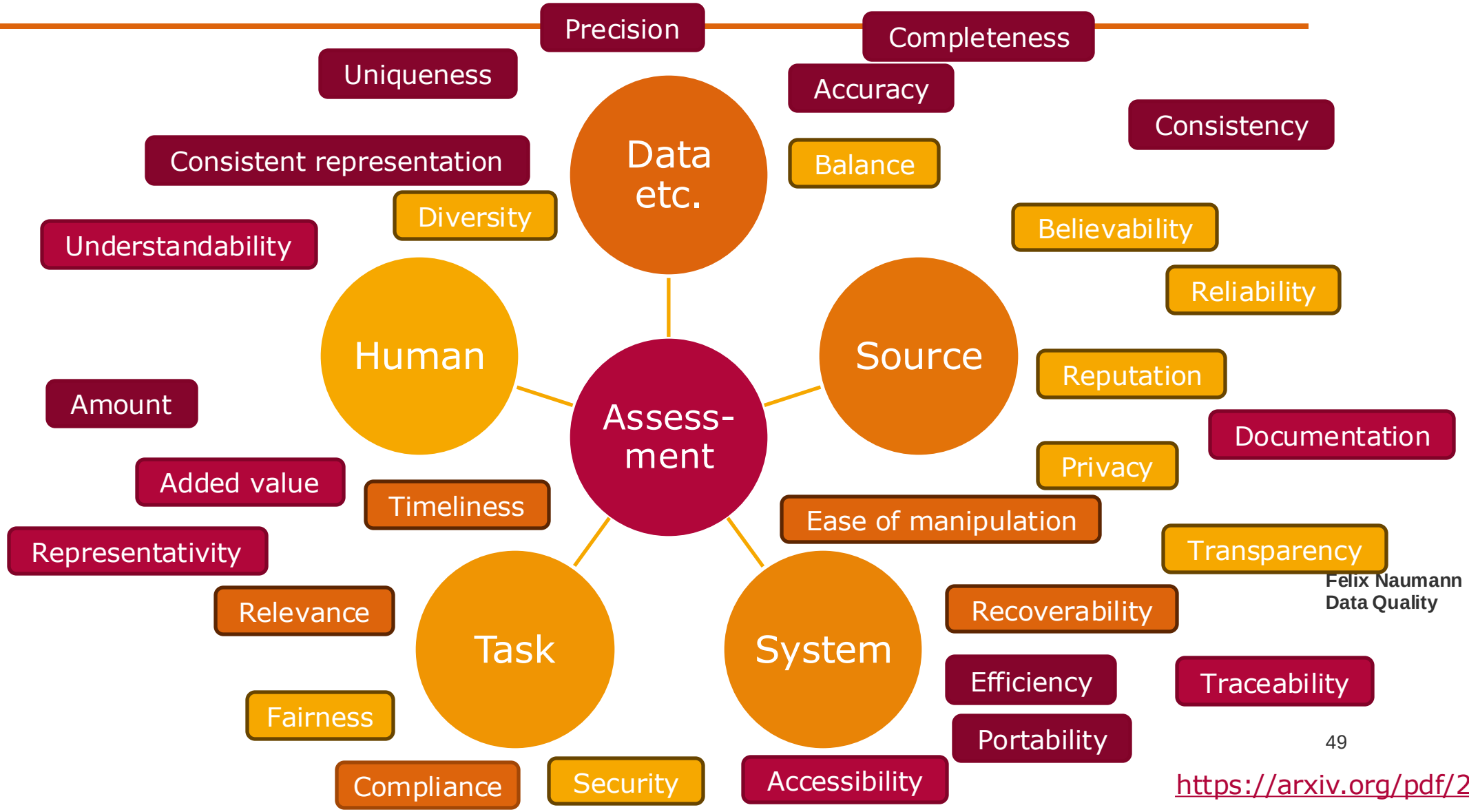


Figure 1: Three sources of IQ criterion scores

Assessment Class	IQ Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy	Continuous user assessment
	Reputation	User experience
	Understandability	User sampling
	Value-Added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer Support	Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

Table 2: Classification of IQ Metadata Criteria

Ingredients for DQ Assessment: Five Facets



https://arxiv.org/pdf/2403.00526

Assessing Data Quality

Recoverability
Portability
Privacy
Uniqueness
Accuracy
Added value
Believability
Consistent representation
Security
Documentation
Fairness
Precision
Balance
Transparency
Timeliness
Measure
Check
Assess
Reliability
Amount
Accessibility
Relevance
Consistency
Traceability
Completeness
Ease of manipulation
Efficiency
reputation
Compliance
Diversity
Understandability
Representativity

Assessment Examples

Completeness

- Values vs. rows vs. columns
- Absolute (counts) or relative (percentages)
- Relative to what? External data needed
- Semantically challenging

Representativity

- vs. balance vs. diversity
- Presence of every value combination
  - Existing values vs. all values
  - Computationally challenging
- Distribution similar to real-world distribution: external data needed

Free-of-errors / Correctness

- Error detection
  - Count at value or row-level
- Business rules
  - Patterns, dependencies, data-types
- Outlier detection
- Validation with external data

Relevance

- …

Understandability

- …

Further Challenges for DQ Assessment

- Ambiguity
  - Many attempts to compile and define DQ dimensions
  - Definitions of the dimensions inherently ambiguous
- Explainability
  - Assessment results explainable to consumers
  - Results traceable to their root cause, to improve quality
- Efficiency
  - Assessment effort and time should be low
- Compliance
  - Fulfill organizational data governance processes
  - Comply to a legal framework, e.g., GDPR or the AI Act
- Scoring
  - Aggregate and normalize assessment results to some numeric scale.
  - Allows comparison across datasets and across time
- Adequacy
  - Is the data of sufficient quality or adequate for the task at hand?

Summary

- Data and Information Quality Research
- Data Preparation
- Data Quality and AI Systems
- Data Quality Assessment

Privacy
Uniqueness
Relevance
Understandability
Security
Accuracy
Added value
Consistent representation
Fairness
Reputation
Diversity
Representativity
Precision
Consistency
Reliability
Believability
Traceability
Completeness
Portability
Accessibility
Documentation
Balance
Timeliness
Amount
Transparency
Recoverability
Efficiency
Compliance
Representativity
Ease of manipulation