Methods for classifying nonprofit organisations according to their field of activity:
A report on semi-automated methods based on text

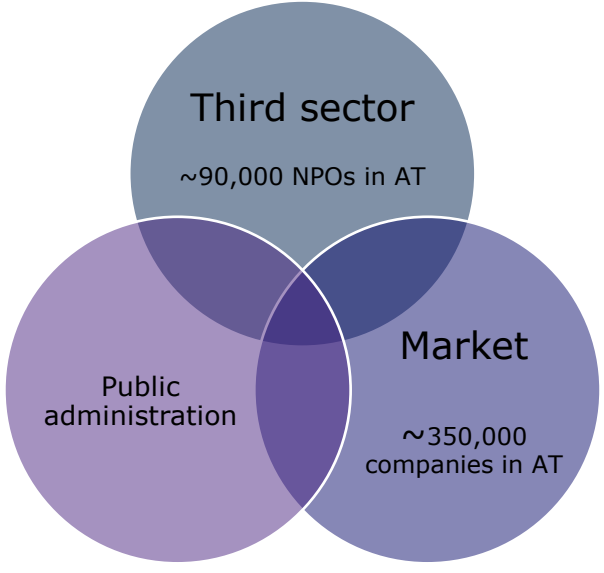
uRos 2020 - 8th International conference on the Use of R in Official Statistics
02.12.2020

Julia Litofcenko, Dominik Karner, Florentine Maier
Institute for Nonprofit-Management

Outline of the presentation

- Motivation
  - Nonprofit organizations not systematically integrated into public statistics
  - Aim: Mapping the sector according to area of activity
- Empirical setting
  - Registry of associations in Austria
- Methods
  - Classification with a rule-based or dictionary approach
  - Classification with Naive Bayes, Lasso regression and decision trees
- Findings
- Conclusion
  - NPOs can satisfactorily be classified according to areas of activity based on names only with semi-automated approaches

Motivation



Sectoral model of society

- Third sector: Nonprofit organizations (NPOs) following ideational goals
- Definition (Salamon & Anheier, 1992: 12f):
  - Formal: institutionalized to some extent
  - Private: institutionally different from government
  - Non-profit-distributing: not returning profits generated to their owner or directors
  - Self-governing: equipped to control their own activities
  - Voluntary: involving some meaningful degree of voluntary participation, either in the actual conduct of the agency's activities or in the management of affairs
- Not systematically integrated into public statistics in Austria and most countries, although recommend by the UN statistical division (United Nations, 2018)

Sports, culture and arts, social clubs

David Hasselhoff
Fanclub Graz

Startsei

Social services & membership organizations

Business & professional associations, interest groups, political associations

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

VOEB
VERBAND ÖSTERREICHISCHER ENTSORGUNGSBETRIEBE
Gemeinsam Ressourcen sichern

Österreichischer Verband
für Fischereiwirtschaft und Aquakultur

Hauptverband des Österreichischen Buchhandels

ÖSTERREICHISCHER BIOMASSE-VERBAND
AUSTRIAN BIOMASS ASSOCIATION

verband österreichischer software industrie

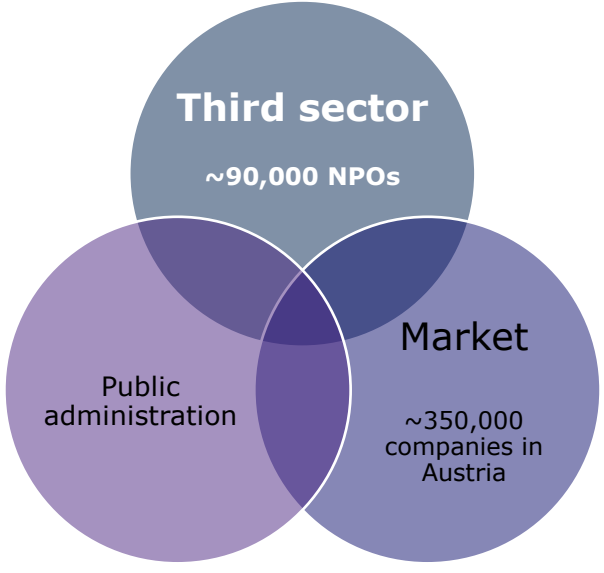
Österreichischer Versicherungsmaklerring

STAATLICH ERMÄCHTIGTE AUSBILDUNGSEINRICHTUNG
erster österreichischer berufstauchlehrer verband

Die Stimme der Gemeinnützigen
IGO

EQUIS ACCREDITED AACSB ACCREDITED ASSOCIATION AMBA ACCREDITED

Motivation

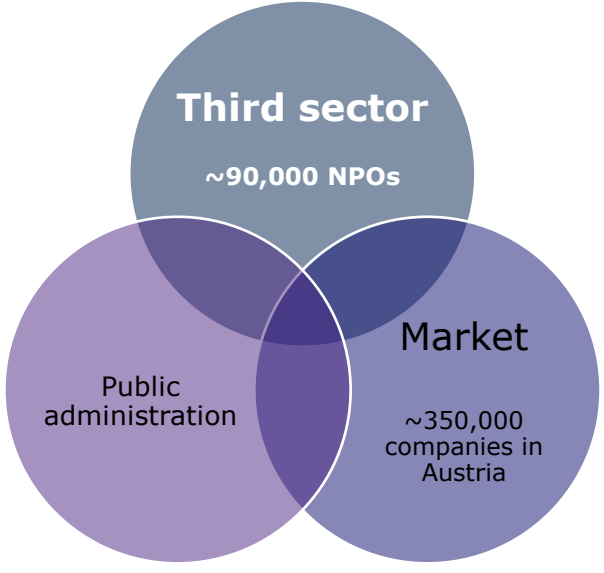


Aim: Mapping the sector according to area of activity

Scattered data:

- Registry of associations: ~90,000
- Commercial register: ~1,200
- Payroll tax statistics: ~12,000 NPOs as employers
- 11 databases for ~500 foundations
- Occasional survey data (max. 12,000 participants)

Motivation



Aim: Mapping the sector according to area of activity

Scattered data:

- **Registry of associations: ~90,000**
- Commercial register: ~1,200
- Payroll tax statistics: ~12,000 NPOs as employers
- 11 databases for ~500 foundations
- Occasional survey data (max. 12,000 participants)

Empirical setting

Registry of associations

- Ministry for the Interior
- Not publicly available
- Not digitized

Copy from business information publisher *Compass Verlag GmbH*

- Collects data for financial institutions
- Digitized:
  - **Association's name**
  - Address
  - Founding year
  - Legal representatives
  - **NO bylaws or mission statements**
- Not completely up to date (most recent two years not reliable)

Empirical setting

Area of activity

- International Classification of Nonprofit Organizations (ICNPO)
  (Salamon and Anheier, 1992)
- Internationally comparable, similar to NACE

ICNPO (Sub-)group number	ICNPO (Sub-)group name
1 000	Culture and recreation
1 100	Culture and arts
1 200	Sports
1 300	Other recreation and social clubs
2 000	Education and research
3 000	Health
4 000	Social services
5 000	Environment
6 000	Development and housing
7 000	Law, advocacy and politics
8 000	Philanthropic intermediaries and voluntarism promotion
9 000	International
10 000	Religion
11 000	Business and professional associations, unions
12 000	Not elsewhere classified

Method

Machine learning (ML) approaches as a common starting point
(e.g. Naïve-Bayes classifiers, decision trees, regression methods, neural networks)

Good results with (Fisher, 2016; Lepere-Schloop, 2017; Ma, 2020)

i. Large training samples
ii. Long, high quality texts

Problems

i. Training sample needs to be constructed manually -
Best Model, trained on 3.333 cases, classifies only 49% of NPOs correctly
ii. Long texts: Not available/ quality issues

Catch-22

Constructing a benchmark sample

Performance of manual human coding (Litofcenko, Karner, Maier, 2020)

ICNPO Group		True ICNPO (n)	True ICNPO %	Sensitivity of mode of human coders %	Precision of mode of human coders %
1100	Culture	994	20%	92%	94%
1200	Sports	1061	21%	92%	96%
1300	Other Recreation and Social Clubs	909	18%	87%	84%
2000	Education and Research	299	6%	86%	92%
3000	Health	94	2%	70%	80%
4000	Social Services	385	8%	82%	91%
5000	Environment	84	2%	71%	92%
6000	Development and Housing	404	8%	85%	82%
7000	Law, Advocacy and Politics	187	4%	71%	83%
8000	Philanthropic Intermediaries and Voluntarism Promotion	6	0%	50%	100%
9000	International	75	2%	87%	88%
10000	Religion	90	2%	66%	89%
11000	Business and Professional Associations, Unions	350	2%	81%	82%
12000	Not Elsewhere Classified	62	1%	13%	100%
Total		5000	100%	85%	

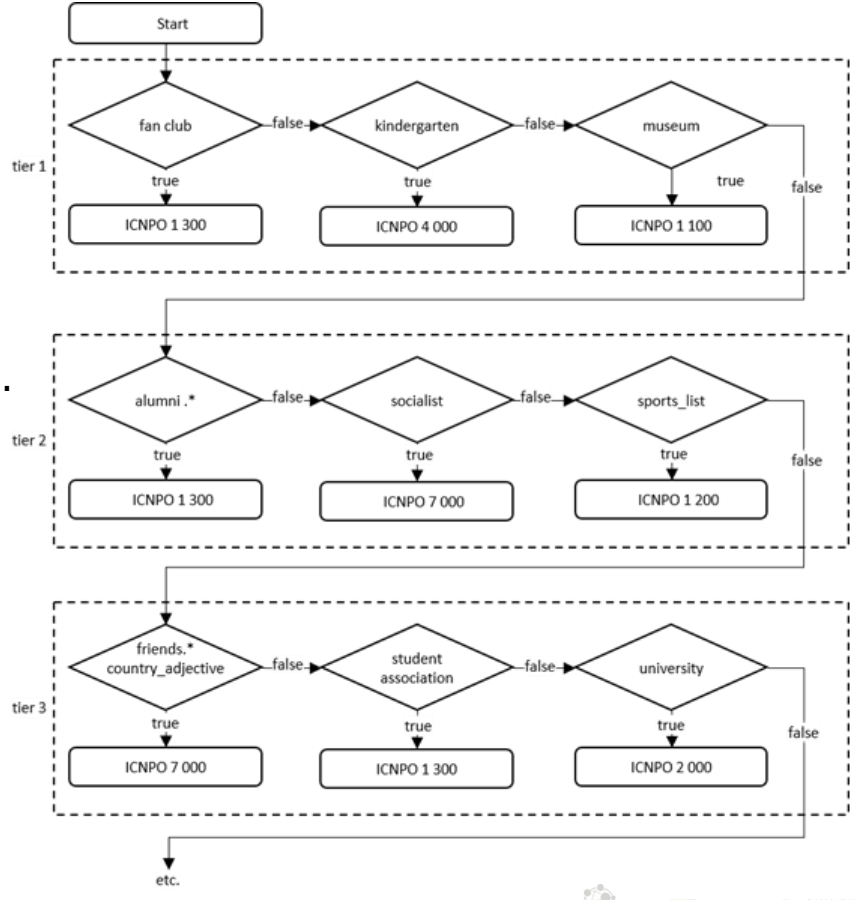
$$\text{Sensitivity} = {}^{TP}/_{(TP+FN)}$$

$$\text{Precision} = {}^{TP}/_{(TP+FP)}$$

Dictionary or rule-based approach

Applicable if (Zhai & Massung, 2016)

i. Categories are clearly defined.
ii. Categories can be relatively easily distinguished based on surface features in the text (e.g., particular words).
iii. Researchers have sufficient domain knowledge to suggest many effective rules.



In the Austrian case:

- 3090 search terms arranged in 211 tiers
- Including wildcard-lists (sports, professions, countries and so forth)

Dictionary or rule-based approach

See https://epub.wu.ac.at/6767/

	A	B	C	D
1	search_term	preliminary_ICNPO_marker	tier	ICNPO_category
2	.*musikfreunde	01100_1	1	1100
3	kapelle	01100_1	1	1100
4	museum	01100_1	1	1100
5	musikfreunde	01100_1	1	1100
6	traditionsverband	01100_1	1	1100
7	.*chor	01100_4	1	1100
8	.*absolventen.*	01300_1	1	1300
9	.*kanarien.*	01300_1	1	1300
10	.*kleintier.*	01300_1	1	1300
11	.*sparclub.*	01300_1	1	1300
12	.*spargemeinschaft	01300_1	1	1300
13	.*sparrunde.*	01300_1	1	1300

```
for (i in 1:nrow(search_terms_r)){
  print(i)
  search_term <- search_terms_r[i, "search_term"]
  preliminary_ICNPO_marker <- search_terms_r[i, "preliminary_ICNPO_marker"]
  df$preliminary_ICNPO_marker[grepl(search_term, df$assoc_name, fixed=FALSE, ignore.case=TRUE) == TRUE & is.na(df$preliminary_ICNPO_marker)] <-
preliminary_ICNPO_marker
}
```

Dictionary or rule-based approach

Performance of rule-based classification
(column percent; figures are rounded; Litofcenko, Karner, Maier, 2020)

		true ICNPO													
		1100	1200	1300	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	% predicted ICNPO
predicted ICNPO	1100	90%	0%	0%	1%	1%	1%	0%	1%	6%	0%	1%	1%	0%	2%	19%
	1200	0%	90%	1%	0%	1%	0%	0%	0%	1%	33%	0%	0%	1%	2%	20%
	1300	1%	2%	86%	0%	2%	3%	4%	1%	3%	0%	1%	0%	2%	2%	17%
	2000	0%	0%	0%	86%	1%	1%	1%	0%	0%	0%	1%	2%	0%	2%	5%
	3000	0%	0%	0%	1%	85%	2%	2%	0%	0%	0%	1%	1%	1%	0%	2%
	4000	1%	0%	1%	1%	2%	85%	2%	1%	3%	17%	9%	1%	1%	2%	8%
	5000	0%	0%	0%	2%	0%	1%	79%	2%	1%	0%	0%	0%	1%	2%	2%
	6000	0%	0%	0%	0%	0%	0%	1%	80%	2%	0%	3%	0%	4%	0%	7%
	7000	0%	0%	1%	1%	0%	1%	0%	0%	64%	0%	5%	0%	1%	0%	3%
	8000	0%	0%	0%	0%	0%	0%	0%	0%	0%	17%	0%	0%	0%	0%	0%
	9000	0%	0%	0%	1%	0%	0%	0%	0%	2%	0%	53%	0%	2%	2%	1%
	10000	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	80%	1%	0%	2%
	11000	0%	0%	0%	1%	2%	0%	2%	2%	5%	0%	1%	1%	74%	2%	6%
	12000	6%	5%	10%	6%	5%	6%	8%	11%	13%	33%	23%	13%	12%	87%	9%
true ICNPO (n)		994	1061	909	299	94	385	84	404	187	6	75	90	350	62	5000
% true ICNPO		20%	21%	18%	6%	2%	8%	2%	8%	4%	0%	2%	2%	7%	1%	100%
precision		96%	98%	92%	94%	78%	85%	67%	93%	82%	100%	66%	90%	89%	12%	

Sensitivity 85%

ML and curated keywords

Improve the quality of input texts for ML - best of both worlds?

-> Decision tree with curated keywords

Original organization name	Curated association name
Studentensport	.*ensport.* .*sport.* .*student.*
GOLD - FINGER : gemeinnütziger Verein zur Förderung der Musikkultur in EUROPA	musikkultur musik.* .*kultur.* .*musi.*
Alumni der Akademie der bildenden Künste Wien	Alumni.* akademie künste.*
Bosniakische Kultur- und Glaubensgemeinschaft Oberland	glaubens.* bosniak.* kultur .

ML and curated keywords

Performance of decision tree classification with curated organization names (column percent; figures are rounded; Litofcenko, Karner, Maier, 2020)

		true ICNPO													
		1100	1200	1300	2000	3000	4000	5000	6000	7000	8000	9000	10000	11000	12000	% predicted ICNPO
predicted ICNPO	1100	84%	0%	1%	1%	0%	1%	0%	0%	3%	0%	0%	11%	0%	0%	39%
	1200	1%	88%	2%	2%	3%	0%	0%	0%	3%	0%	0%	0%	2%	0%	20%
	1300	13%	11%	91%	10%	10%	14%	12%	31%	40%	100%	19%	17%	19%	100%	13%
	2000	0%	0%	1%	77%	7%	1%	6%	4%	1%	0%	0%	0%	1%	0%	5%
	3000	0%	0%	0%	0%	59%	2%	0%	0%	1%	0%	0%	0%	1%	0%	2%
	4000	1%	1%	1%	3%	14%	75%	3%	4%	1%	0%	0%	3%	3%	0%	7%
	5000	0%	0%	0%	1%	0%	3%	67%	0%	1%	0%	5%	0%	0%	0%	1%
	6000	0%	0%	1%	2%	0%	1%	9%	53%	4%	0%	0%	0%	5%	0%	4%
	7000	0%	0%	0%	2%	0%	2%	0%	1%	42%	0%	0%	6%	0%	0%	2%
	8000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	9000	0%	0%	0%	1%	0%	0%	0%	0%	1%	0%	76%	0%	3%	0%	1%
	10000	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	57%	0%	0%	1%
	11000	1%	0%	2%	1%	7%	1%	3%	7%	3%	0%	0%	6%	66%	0%	6%
	12000	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
true ICNPO (n)		325	358	299	103	29	138	33	112	77	1	21	35	115	21	1667
% true ICNPO		19%	21%	18%	6%	2%	8%	2%	7%	5%	0%	1%	2%	7%	1%	100%
precision		95%	96%	54%	84%	77%	79%	73%	77%	76%	-	70%	91%	74%	-	

Sensitivity 77%

Hypothesis:
Limitations to algorithm based on local optimization in high dimensional spaces
(see also Gentzkow, Kelly, & Taddy, 2019)

Conclusion

NPOs can satisfactorily be classified according to areas of activity based on names only with semi-automated approaches

- Obtained sensitivity: 85%
- Performance not inferior to human coding
- Performance not inferior to classification based on mission/ program statements
- Best resp. only possible solution in most real-world scenarios
- Classification based on a manually generated rule-set surprisingly superior to a decision-tree classification

References

Fisher, I. E., Garnsey, M. R., & Hughes, M. E. (2016). Natural language processing in accounting, auditing and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, *23*(3), 157-214.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535-574.

Lepere-Schloop, M., Zook, S., & Bawole, J. N. (2018). *NGO classification from the bottom-up: Using self-reported data and machine learning to generate categories of NGOs in Ghana*. Paper presented at the ISTR 13th International Conference, Amsterdam.

Litofcenko, J., Karner, D., & Maier, F. (2020). Methods for Classifying Nonprofit Organizations According to their Field of Activity: A Report on Semi-automated Methods Based on Text. *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, *31*(1), 227-237.

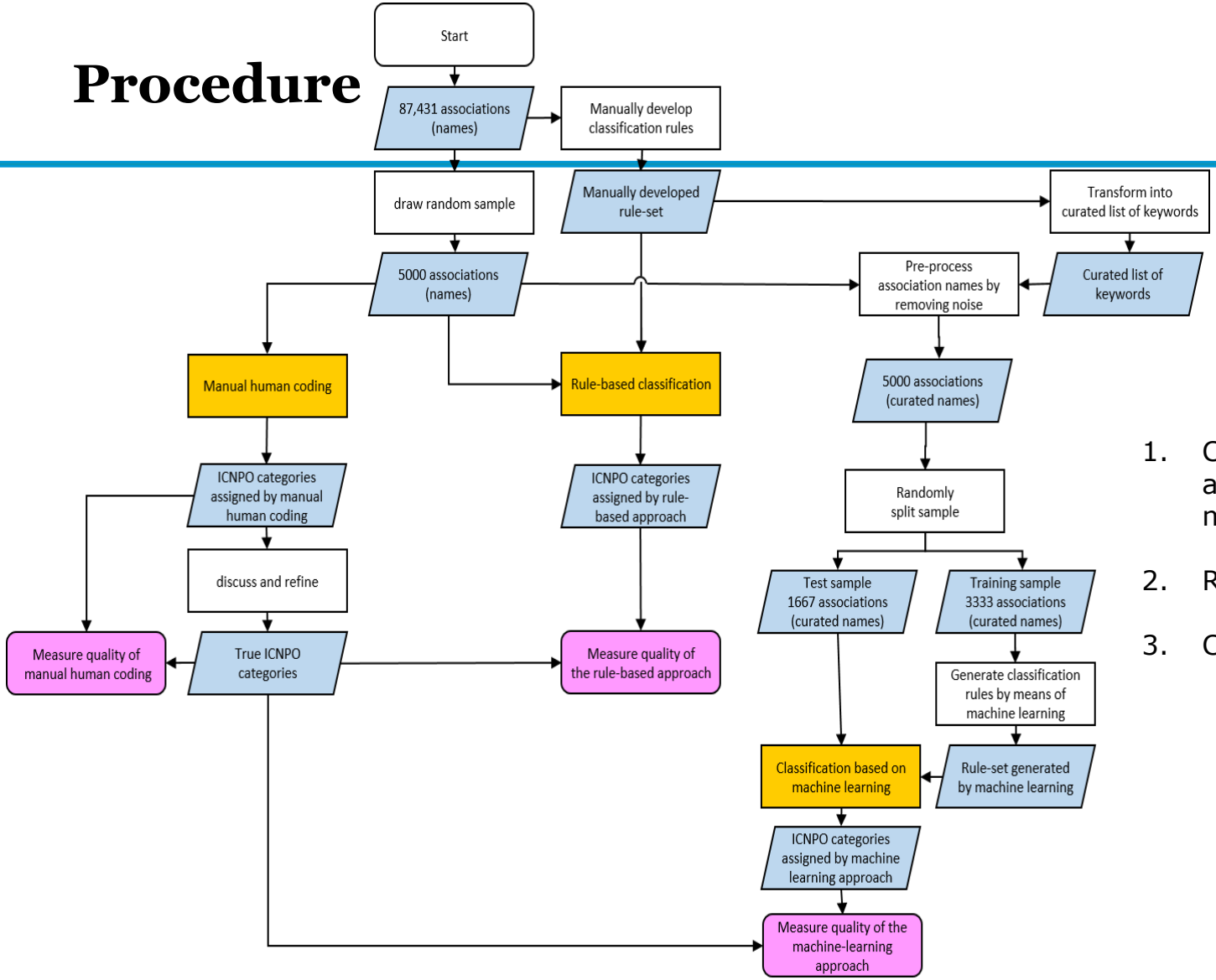
Ma, J. (2020). Automated Coding Using Machine Learning and Remapping the US Nonprofit Sector: A Guide and Benchmark. *Nonprofit and Voluntary Sector Quarterly*, 0899764020968153.

Salamon, L. M., & Anheier, H. K. (1992). In search of the non-profit sector II: The problem of classification. *Voluntas: International Journal of Voluntary and Nonprofit Organizations, 3*(3), 267-309.

United Nations. (2018). Satellite Account on Non-profit and Related Institutions and Volunteer Work. Retrieved from https://unstats.un.org/unsd/nationalaccount/docs/UN_TSE_HB_FNL_web.pdf

Zhai, C. X., & Massung, S. (2016). *Text data management and analysis: a practical introduction to information retrieval and text mining*. New York, NY: Association for Computing Machinery and Morgan & Claypool.

Procedure



1. Constructing a training and test sample through manual coding
2. Rule-based classification
3. Classification based on ML