Research and Documentation Centre

On personal data minimization & algorithmic fairness

Mortaza S. Bargh

Workshop at Surfnet, 1 June 2023, afternoon session

Wetenschappelijk onderzoeks- en kennisinstituut
voor het ministerie van Justitie en Veiligheid

Introduction

What is responsible ML (responsible AI)

Important issues/concerns of AI/ML according to [CHO20]

- Security concerns
- Explainability (and interpretability) concerns
- Fairness concerns

Algorithmic fairness

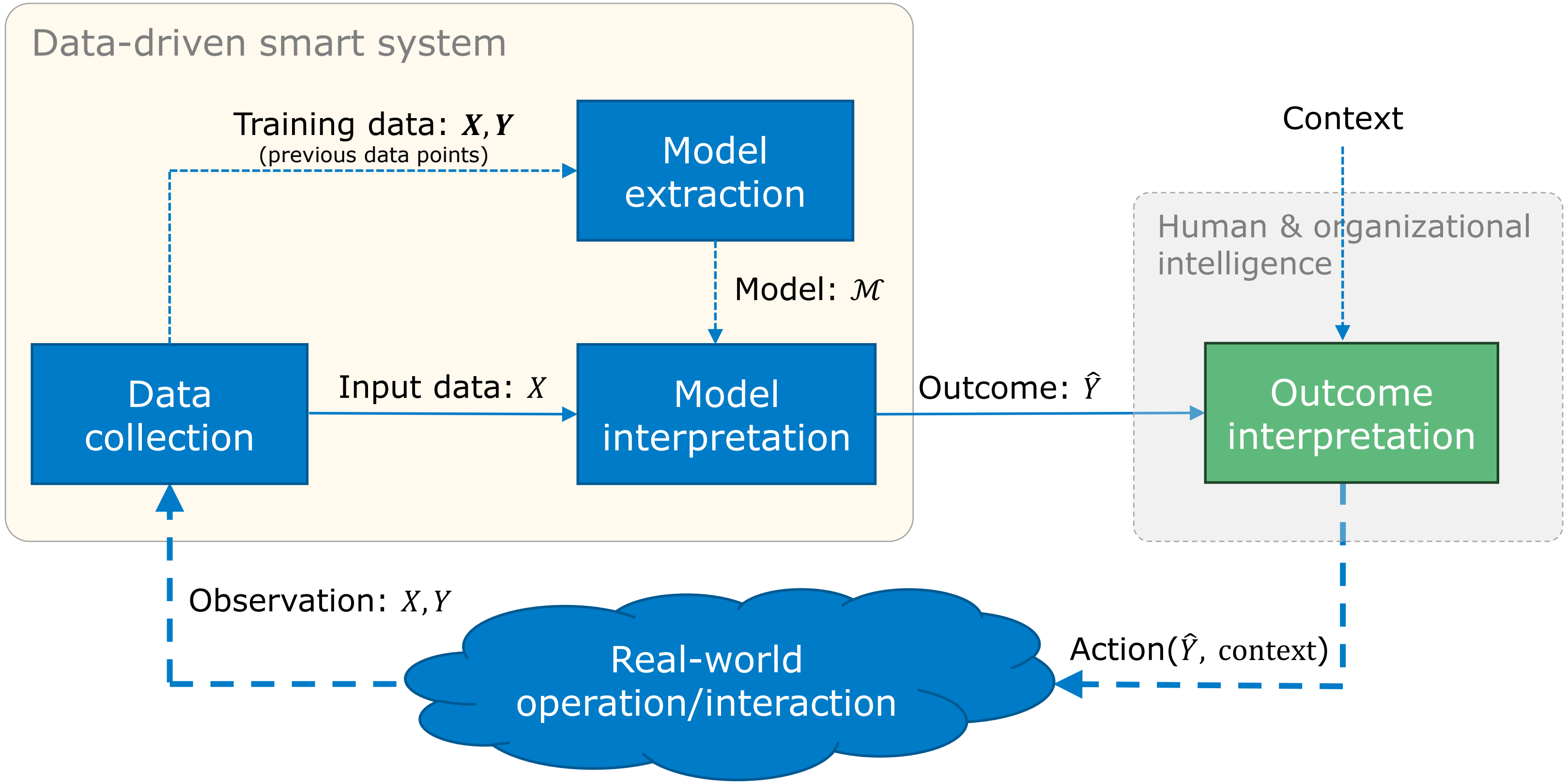
Taxonomy of algorithmic fairness concepts

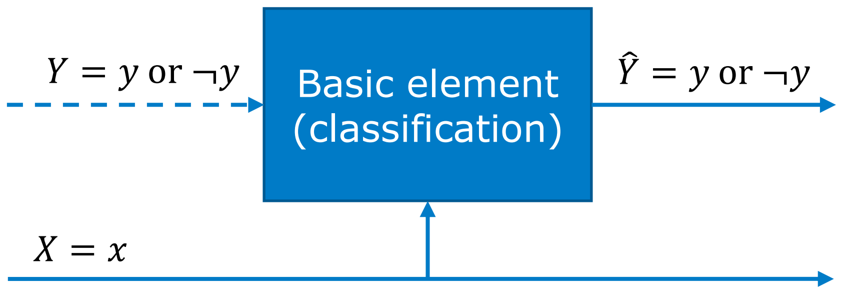
- **I – Philosophical and social-science concepts**
  - Based on egalitarianism
  - Based on organizational justice
  - ...
- **II – Formal (mathematical) definitions**
  - Statistical measures
  - Similarity-based measures
  - Causal reasoning
  - Preference-based measures

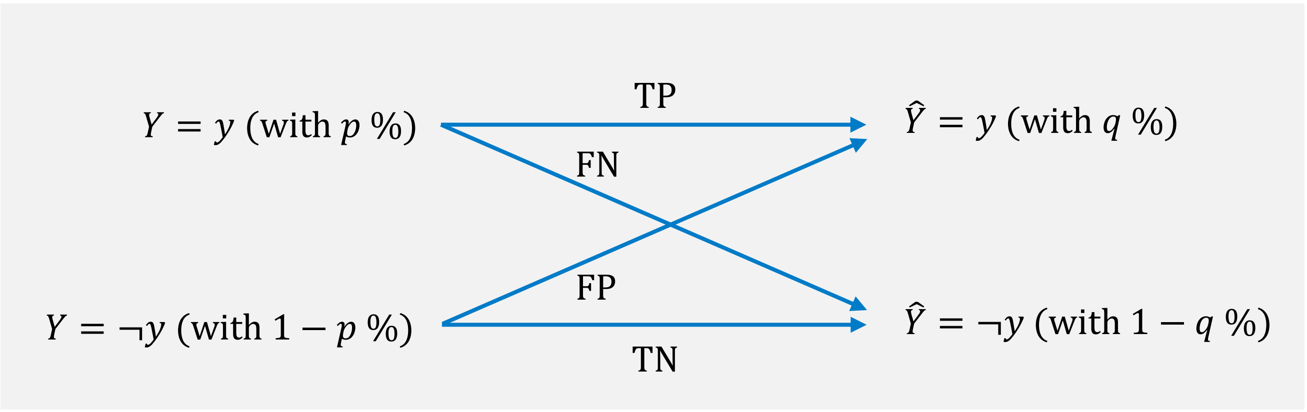
Formal fairness: Statistical measures

Formal and mathematical concepts

Data driven smart systems




$Y = y$ or $\neg y$
Basic element (classification)
$\hat{Y} = y$ or $\neg y$
$X = x$


$Y = y$ (with p %)
$Y = \neg y$ (with $1 - p$ %)
TP
FN
FP
TN
$\hat{Y} = y$ (with q %)
$\hat{Y} = \neg y$ (with $1 - q$ %)

From Section 3 of [STA21]
See also [VER18]

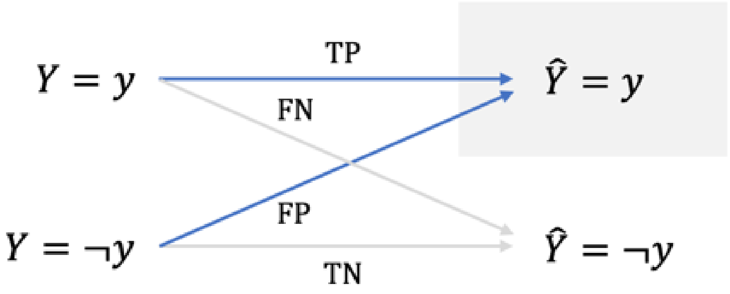
(a) Statistical measures

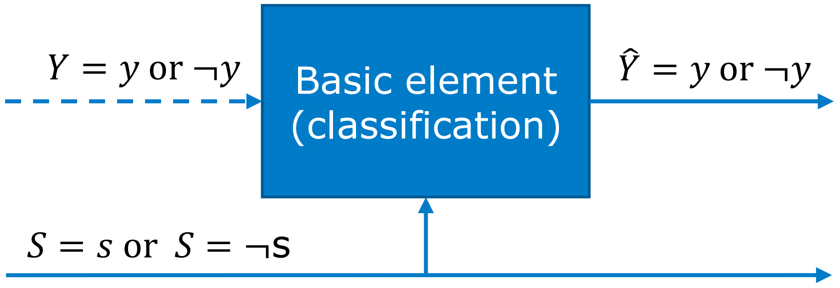
Statistical measures are based on different calibrations of predicted probabilities, predicted outcomes, and actual outcomes

Outline:
- Classical metrics (12 measures)
- Fairness ones (13 measures)

- Statistical parity = group fairness = equal acceptance rate = benchmarking
- Conditional statistical parity
- Predictive parity = outcome test
- False positive rate balance = predictive equality
- False negative rate balance = equal opportunity
- Equalized odds = conditional procedure accuracy = disparate mistreatment
- Conditional use accuracy
- Overall accuracy equality
- Treatment equality
- Test fairness = calibration = matching conditional frequencies
- Well calibration
- Balance for positive class
- Balance for negative class

Example metrics

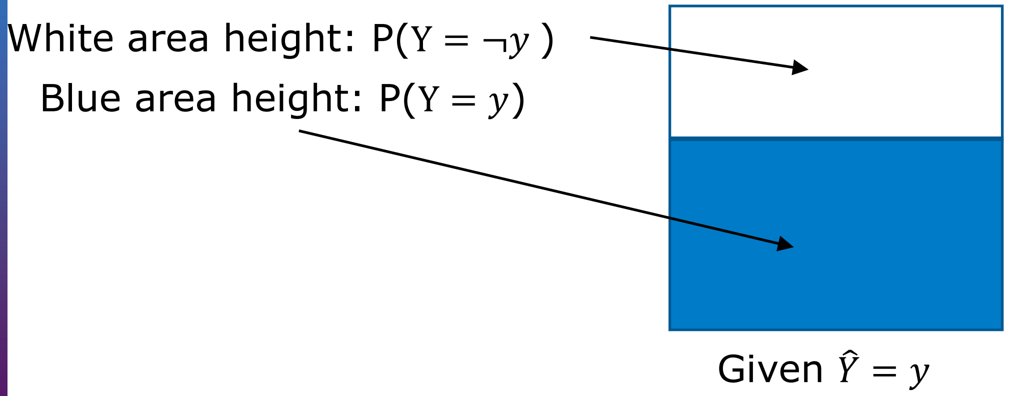


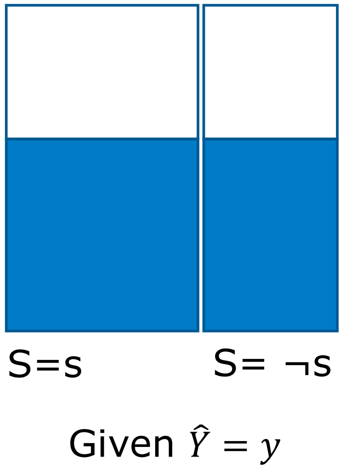


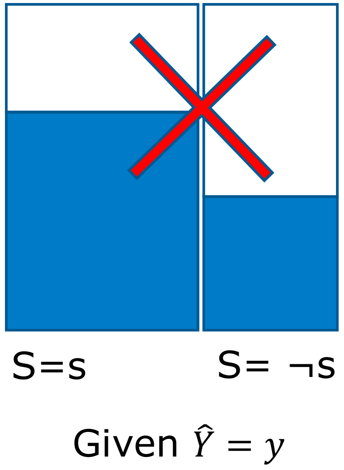
Classical statistical metrics	**Related fairness metrics**
Positive Predictive Value (PPV) or precision or correct acceptance $Pr(Y = y \mid \hat{Y} = y) = \frac{TP}{TP + FP}$	Predictive parity or outcome test $Pr(Y = y \mid \hat{Y} = y, S = s) = Pr(Y = y \mid \hat{Y} = y, S = \neg s)$

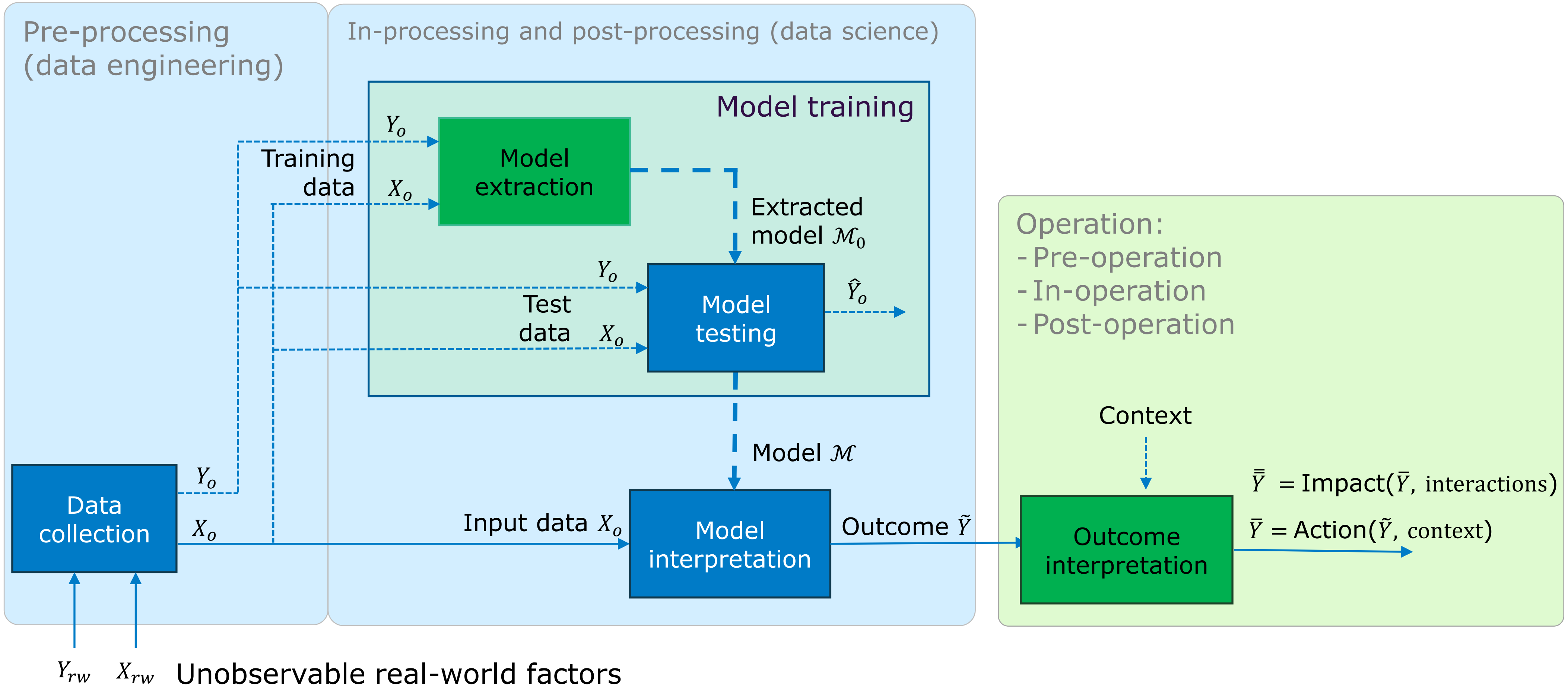
Predictive parity or outcome test

- Both protected and unprotected groups have equal PPV – the probability of a subject with positive predictive value to truly belong to the positive class
- Example: Both male and female applicants with a good predicted credit score having actually a good credit score

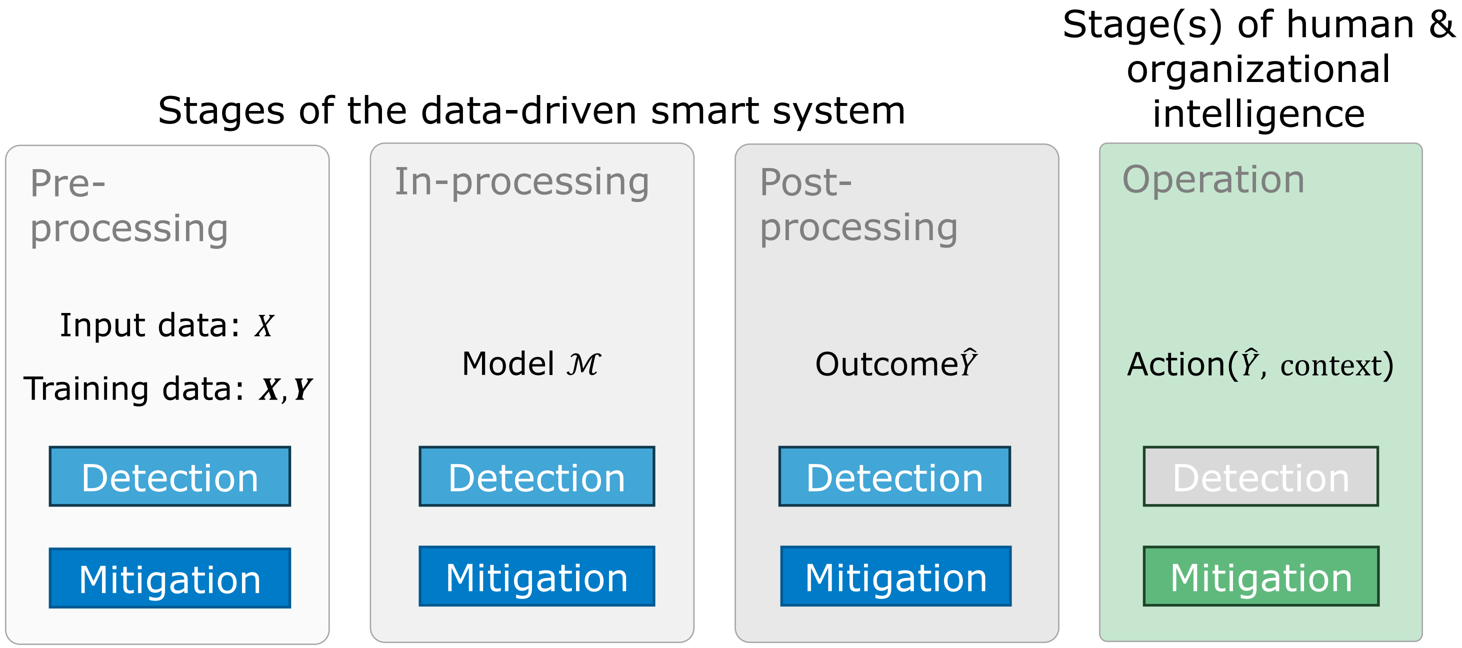






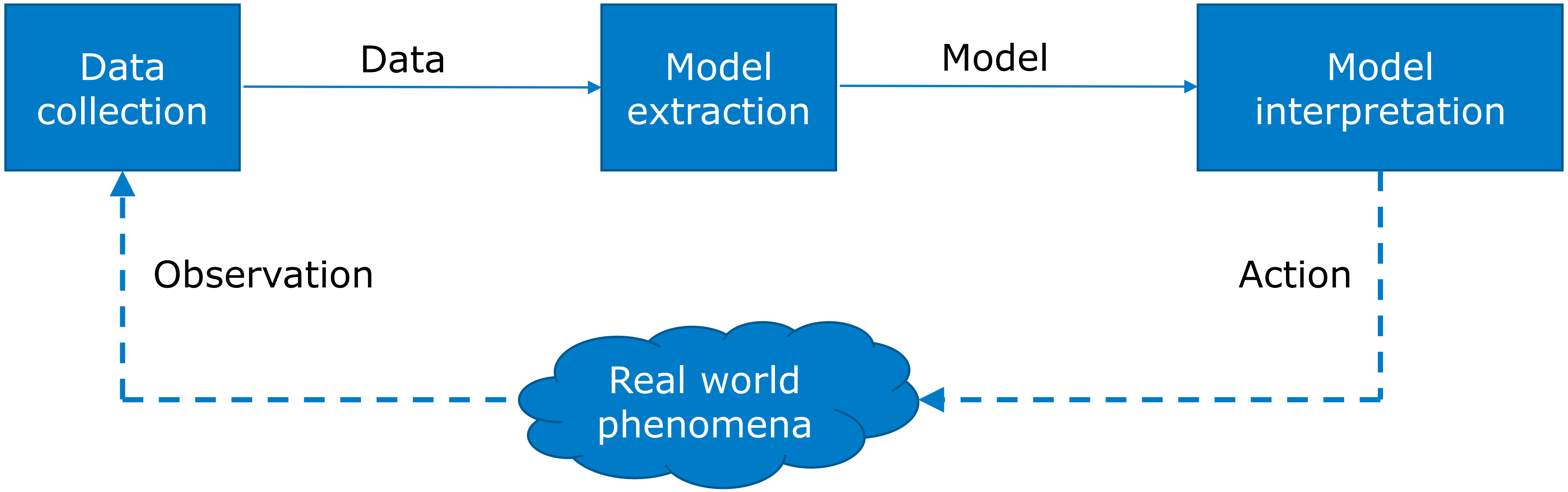

Pre-processing
(data engineering)
In-processing and post-processing (data science)
Model training
Training data
Y_o
X_o
Model extraction
Extracted model $\mathcal{M}_0$
Test data
Y_o
X_o
Model testing
$\hat{Y}_o$
Model $\mathcal{M}$
Data collection
Y_o
X_o
Input data X_o
Model interpretation
Outcome $\tilde{Y}$
Operation:
- Pre-operation
- In-operation
- Post-operation
Context
Outcome interpretation
$\bar{\bar{Y}}$ = Impact($\bar{Y}$, interactions)
$\bar{Y}$ = Action($\tilde{Y}$, context)
Y_{rw}
X_{rw}
Unobservable real-world factors

Data driven smart systems

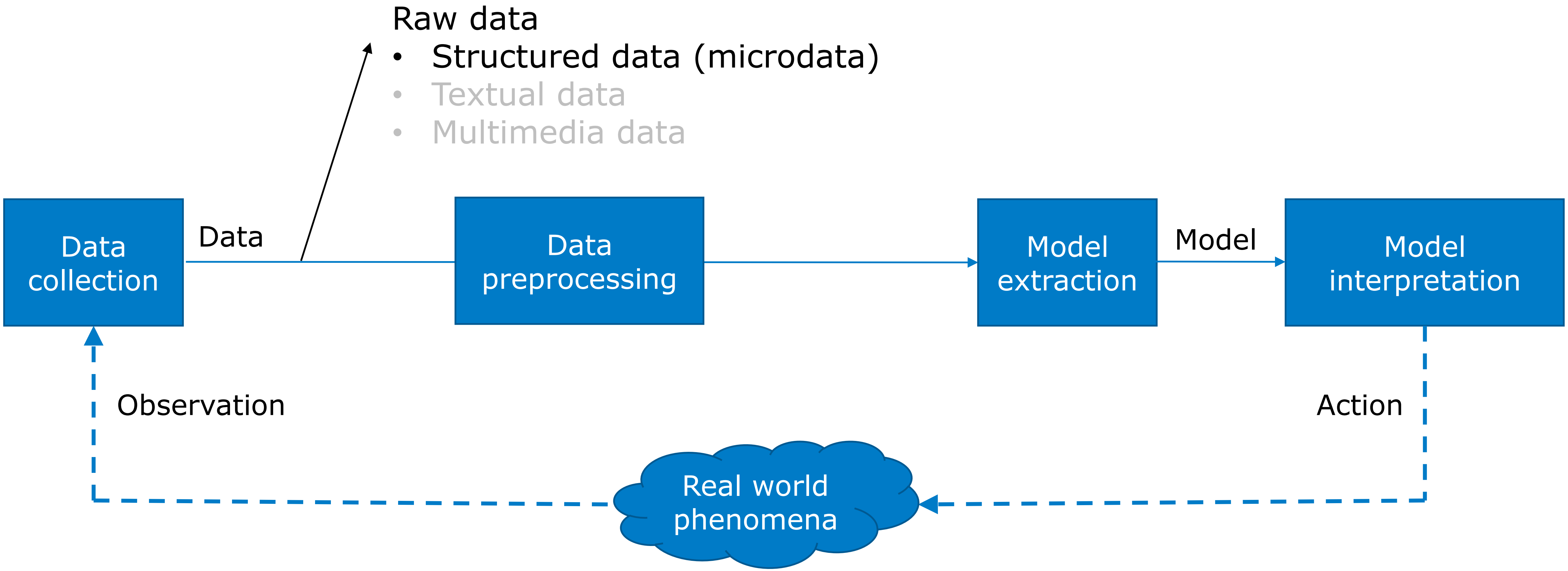


Applying together with k-anonymity

Data driven (or AI) applications



Data driven (or AI) applications



Example of a microdata set

A data set collected at a hospital

name	job	sex	age	disease	height (cm)
Bob	dancer	male	35	hepatitis	184
Fred	writer	male	38	HIV	180
Doug	dancer	male	38	Flu	210
Alice	engineer	female	30	Flu	172
Cathy	engineer	female	33	HIV	170
Emily	physician	female	31	HIV	169
Gladys	lawyer	female	31	HIV	171

Microdata protection: De-identification

Example

Explicit
Identifier
EID

name	job	sex	age	disease	height (cm)
Bob	dancer	male	35	hepatitis	184
Fred	writer	male	38	HIV	180
Doug	dancer	male	38	Flu	210
Alice	engineer	female	30	Flu	172
Cathy	engineer	female	33	HIV	170
Emily	physician	female	31	HIV	169
Gladys	lawyer	female	31	HIV	171

Microdata protection: De-identification

Example: Here via removal (other methods: suppression and replacement with pseudo-IDs)

name	job	sex	age	disease	height (cm)
	dancer	male	35	hepatitis	184
	writer	male	38	HIV	180
	dancer	male	38	Flu	210
	engineer	female	30	Flu	172
	engineer	female	33	HIV	170
	physician	female	31	HIV	169
	lawyer	female	31	HIV	171

Microdata protection: Applying SDC

Example

Quasi
Identifiers
QIDs

name	job	sex	age	disease	height (cm)
	dancer	male	35	hepatitis	184
	writer	male	38	HIV	180
	dancer	male	38	Flu	210
	engineer	female	30	Flu	172
	engineer	female	33	HIV	170
	physician	female	31	HIV	169
	lawyer	female	31	HIV	171

Traditionally some of them are
protected through blindness

Methods for protecting QIDs

Generalization

- To replace some values with a parent value in the taxonomy of an attribute
- Example: Age: 34 → [30, 40)

Suppression

- To replace the values of QIDs with a meaningless character
- Age: 34 → ***

Microdata protection: Applying SDC

Example

name	job	sex	age	disease	height (cm)
	dancer	male	35	hepatitis	184
	writer	male	38	HIV	180
	dancer	male	38	Flu	210
	engineer	female	30	Flu	172
	engineer	female	33	HIV	170
	physician	female	31	HIV	169
	lawyer	female	31	HIV	171

Microdata protection: Applying SDC

Example

name	job	sex	age	disease	height (cm)
	artist	male	35-39	hepatitis	184
	artist	male	35-39	HIV	180
	artist	male	35-39	Flu	210
	engineer	female	30	Flu	172
	engineer	female	33	HIV	170
	physician	female	31	HIV	169
	lawyer	female	31	HIV	171

Microdata protection: Applying SDC

Example

name	job	sex	age	disease	height (cm)
	artist	male	35-39	hepatitis	184
	artist	male	35-39	HIV	180
	artist	male	35-39	Flu	210
	lawyer	female	30	Flu	172
	engineer	female	33	HIV	170
	engineer	female	31	HIV	169
	physician	female	31	HIV	171

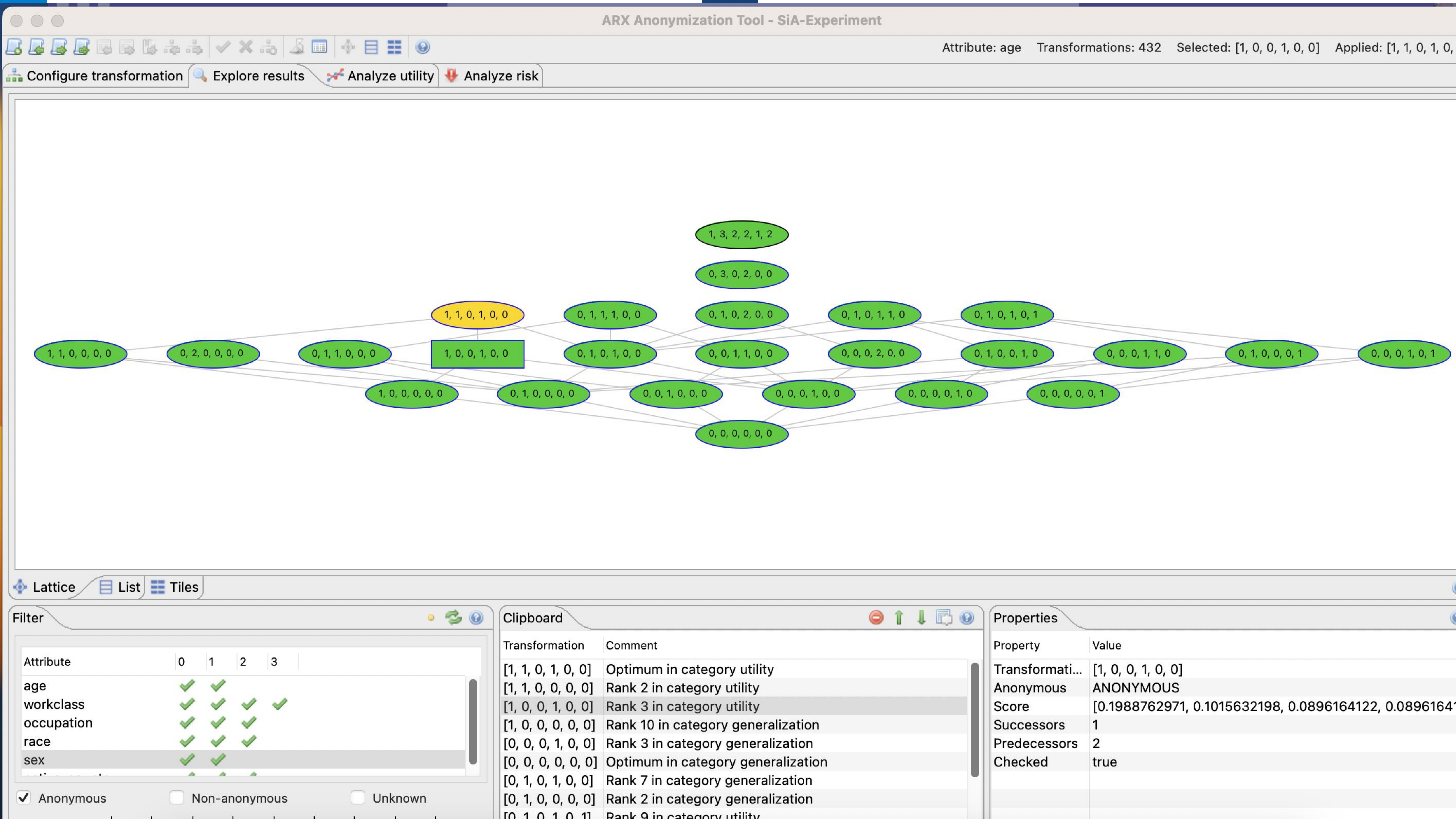
Microdata protection: Applying SDC

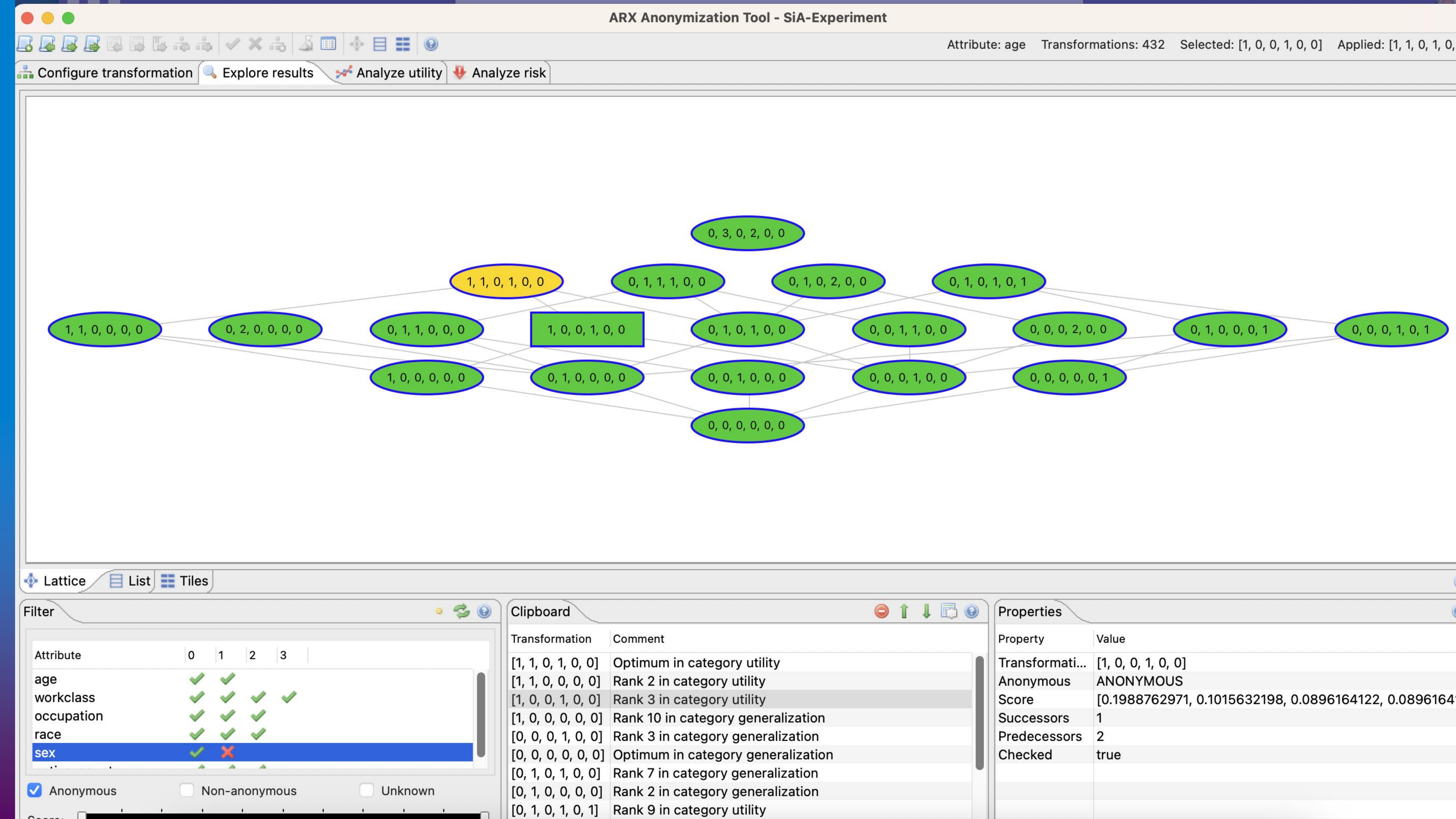
k-anonymity
applied to QIDs

Example

	name	job	sex	age	disease	height (cm)
k=3 Group 1		artist	male	35-39	hepatitis	184
		artist	male	35-39	HIV	180
		artist	male	35-39	Flu	210
k=4 Group 2		profess.	female	30-34	Flu	172
		profess.	female	30-34	HIV	170
		profess.	female	30-34	HIV	169
		profess.	female	30-34	HIV	171

k= min (3, 4) = 3
For this data set


ARX Anonymization Tool - SiA-Experiment
Attribute: age
Transformations: 432
Selected: [1, 0, 0, 1, 0, 0]
Applied: [1, 1, 0, 1, 0,
Configure transformation
Explore results
Analyze utility
Analyze risk
1, 3, 2, 2, 1, 2
0, 3, 0, 2, 0, 0
1, 1, 0, 1, 0, 0
0, 1, 1, 1, 0, 0
0, 1, 0, 2, 0, 0
0, 1, 0, 1, 1, 0
0, 1, 0, 1, 0, 1
1, 1, 0, 0, 0, 0
0, 2, 0, 0, 0, 0
0, 1, 1, 0, 0, 0
1, 0, 0, 1, 0, 0
0, 1, 0, 1, 0, 0
0, 0, 1, 1, 0, 0
0, 0, 0, 2, 0, 0
0, 1, 0, 0, 1, 0
0, 0, 0, 1, 1, 0
0, 1, 0, 0, 0, 1
0, 0, 0, 1, 0, 1
1, 0, 0, 0, 0, 0
0, 1, 0, 0, 0, 0
0, 0, 1, 0, 0, 0
0, 0, 0, 1, 0, 0
0, 0, 0, 0, 1, 0
0, 0, 0, 0, 0, 1
0, 0, 0, 0, 0, 0
Lattice
List
Tiles
Filter
Attribute
0
1
2
3
age
workclass
occupation
race
sex
Anonymous
Non-anonymous
Unknown
Clipboard
Transformation
Comment
[1, 1, 0, 1, 0, 0]
Optimum in category utility
[1, 1, 0, 0, 0, 0]
Rank 2 in category utility
[1, 0, 0, 1, 0, 0]
Rank 3 in category utility
[1, 0, 0, 0, 0, 0]
Rank 10 in category generalization
[0, 0, 0, 1, 0, 0]
Rank 3 in category generalization
[0, 0, 0, 0, 0, 0]
Optimum in category generalization
[0, 1, 0, 1, 0, 0]
Rank 7 in category generalization
[0, 1, 0, 0, 0, 0]
Rank 2 in category generalization
Properties
Property
Value
Transformati...
[1, 0, 0, 1, 0, 0]
Anonymous
ANONYMOUS
Score
[0.1988762971, 0.1015632198, 0.0896164122, 0.0896164
Successors
1
Predecessors
2
Checked
true


ARX Anonymization Tool - SiA-Experiment
Attribute: age
Transformations: 432
Selected: [1, 0, 0, 1, 0, 0]
Applied: [1, 1, 0, 1, 0,
Configure transformation
Explore results
Analyze utility
Analyze risk
0, 3, 0, 2, 0, 0
1, 1, 0, 1, 0, 0
0, 1, 1, 1, 0, 0
0, 1, 0, 2, 0, 0
0, 1, 0, 1, 0, 1
1, 1, 0, 0, 0, 0
0, 2, 0, 0, 0, 0
0, 1, 1, 0, 0, 0
1, 0, 0, 1, 0, 0
0, 1, 0, 1, 0, 0
0, 0, 1, 1, 0, 0
0, 0, 0, 2, 0, 0
0, 1, 0, 0, 0, 1
0, 0, 0, 1, 0, 1
1, 0, 0, 0, 0, 0
0, 1, 0, 0, 0, 0
0, 0, 1, 0, 0, 0
0, 0, 0, 1, 0, 0
0, 0, 0, 0, 0, 1
0, 0, 0, 0, 0, 0
Lattice
List
Tiles
Filter
Attribute
0
1
2
3
age
workclass
occupation
race
sex
Anonymous
Non-anonymous
Unknown
Clipboard
Transformation
Comment
[1, 1, 0, 1, 0, 0] Optimum in category utility
[1, 1, 0, 0, 0, 0] Rank 2 in category utility
[1, 0, 0, 1, 0, 0] Rank 3 in category utility
[1, 0, 0, 0, 0, 0] Rank 10 in category generalization
[0, 0, 0, 1, 0, 0] Rank 3 in category generalization
[0, 0, 0, 0, 0, 0] Optimum in category generalization
[0, 1, 0, 1, 0, 0] Rank 7 in category generalization
[0, 1, 0, 0, 0, 0] Rank 2 in category generalization
[0, 1, 0, 1, 0, 1] Rank 9 in category utility
Properties
Property
Value
Transformati... [1, 0, 0, 1, 0, 0]
Anonymous ANONYMOUS
Score [0.1988762971, 0.1015632198, 0.0896164122, 0.089616
Successors 1
Predecessors 2
Checked true

Another approach for integrating generalization with fairness

See: Hajian, S., Domingo-Ferrer, J., Farràs, O. (2014). **Generalization-based** privacy preservation and discrimination prevention in data publishing and mining. In Data Mining and Knowledge Discovery, 28(5–6), 1158–1188.

Increases the QID attribute set (also includes the discrimination sensitive attributes)

Causes extra data utility degradation if both (fairness protection and privacy protection) are considered

Takeaways

Topics addressed today

Mentioned a new trend: Algorithmic fairness becomes important

Showed a way integration with personal data minimization (anonymization)

Explained another approach by giving a pointer (NB: extra data utility degradation)

References

References

[ALT’18] Altman, M., Wood, A., & Vayena, E. (2018). A harm-reduction framework for algorithmic fairness. IEEE Security & Privacy, 16(3), 34-45.

[BAY’20] Baylon, C., Berghoff, C., Brunessaux, S., Burdalo, L., Dacquisto, G., Damiani, E., Herpig, S., Louveaux, C., Mistiaen, J., Nguyen, D.C., Polemi, N., Praca, I., Sharkov, G., Slieker, V., Szczekocka, E., Orange Polska SA (15 Dec 2020). Artificial Intelligence Cybersecurity Challenges: Threat Landscape for Artificial Intelligence. Technical report. , The European Union Agency for Cybersecurity (ENISA), https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges .

[BIN’18] Binns, R. (2018). What can political philosophy teach us about algorithmic fairness?. IEEE Security & Privacy, 16(3), 73-80.

[CHO’20] Choraś, M., Pawlicki, M., Puchalski, D., & Kozik, R. (2020, June). Machine learning– the results are not the only thing that matters! what about security, explainability and fairness?. In International Conference on Computational Science (pp. 615-628). Springer, Cham.

Algorithmic fairness references

[CHO’16] Chouldechova, A. (2016). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Proc. Conf. Fairness, Accountability, and Transparency in Machine Learning (FAT ML 16), Oct.; https://arxiv .org/abs/1610.07524.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness Through Awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science, 214–226. https://doi.org/10.1145/2090236.2090255

Gajane, P., & Pechenizkiy, M. (2017). On Formalizing Fairness in Prediction with Machine Learning. ArXiv. http://arxiv.org/abs/1710.03184

[HOL17] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.

[KIL17] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. 2017. Avoiding Discrimination Through Causal Reasoning. In Ad- vances in Neural Information Processing Systems

References

[KUS17] Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in neural information processing systems, 30.

[MON18] Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1-15

[NAB18] R. Nabi and I. Shpitser. 2018. Fair Inference On Outcomes. In the 32de Association for the Advancement of Artificial Intelligence (AAAI) conference.

[RAM18] Ramnarayan, G. (2018). Equalizing Financial Impact in Supervised Learning. arXiv preprint arXiv:1806.09211

[SEL’18] Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham L. Rev., 87, 1085.

References

[STA21] Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. arXiv preprint arXiv:2103.12016.

[VEL'17] Veltheer, M. (2017). Using or Being Used by Algorithms – The Ethical Concerns to be Aware Of, Master Thesis MBA Big Data & Business Analytics, University of Amsterdam.

[VER18] Verma, S., & Rubin, J. (2018). Fairness Definitions Explained. Proceedings of the ACM/IEEE International Workshop on Software Fairness, 1–7. https://doi.org/10.1145/3194770.3194776

[VER20] Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596

[VER21] Verma, S., Dickerson, J., & Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited. arXiv preprint arXiv:2106.07756.

References

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1