Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Assessing deanonymisation risk in unstructured text

Prof. Dr. Stephanie Evert

AnGer project team

Korpus- und Computerlinguistik

Friedrich-Alexander-Universität Erlangen-Nürnberg

Funded by the European Union

NextGenerationEU

With funding from the:

Federal Ministry of Research, Technology and Space

Deanonymisation risk

Crucial issue for **privacy-preserving data publishing** (PPDP, see Fung et al. 2010)

Quantitative measures

– **re-identification risk** = #linked records / #records (Manzanares-Salor et al. 2024: 4044)

– **k-anonymity** = candidate set cannot be reduced to less than *k* entities

– flip side: **information loss** due to privacy models

Most research focuses on **statistical databases**

– full data set and background knowledge assumed to be known → can compute *k*-anonymity etc.

– approaches often based on perturbation of numerical data

– but not applicable to **unstructured text** (except artifical tasks such as Wikipedia pages of well-known 20th-century actors, cf. Manzanares-Salor et al. 2024: 4060 f.; Manzanares-Salor & Sánchez 2025)

Anonymisation of unstructured text

Anonymisation of **unstructured text** poses entirely different challenges:

– may contain **direct identifiers** (personally identifying information = **PII**)

– contains wealth of further information that might contribute to deanonymisation = **quasi-identifiers**

– pseudo-identifiers are not formally recognisable (unlike attributes in a database, see Lison et al. 2021)

– background knowledge cannot easily be quantified → impossible to estimate *k*-anonymity etc.

– often about "ordinary people" with little information available online (our use case: **court verdicts**)

– information loss difficult to quantify and depends on application (e.g. *DeLorean* irrelevant for case)

On the late afternoon of Wednesday, Oct 29th retired scientist Dr Emmett Brown came driving down from Beachy Head in his silver DeLorean at high speed. Missing a turn, he crashed through the garden fence at 25 Baslow Rd, causing massive damage to …

Masking strategies

- **redaction** (critical spans deleted) → text difficult to read, maximal information loss
- **initials** (surprisingly common) → leaks information, PII spans become quasi-identifiers
- **randomised initials** → much better protection, but potential inconsistencies (e.g. random dates)
- **realistic surrogates** → natural text, suitable as LLM input, dates shifted to remain consistent
  (not pseudonymisation if mapping is discarded after anonymisation)

On the late afternoon of Wednesday, Nov 12th, retired scientist Prof John Cage came driving down from Capitol Hill in his silver BMW at high speed. Missing a turn, he crashed through the garden fence at 9 High St, causing massive damage to …

Deanonymisation risk of unstructured text

1) **PII** text span **not detected**

– assumption: always leads to full reidentification of entity

– deanonymisation risk = recall of PII detection (more precisely: % of texts with ≥ 1 FN)

2) **Quasi-identifiers** not masked

– combination of multiple quasi-identifiers with background knowledge may enable reidentification

– difficult to quantify: empirical success rate of human adversaries using Web searches (or recently LLMs)

– controlled experiments are almost impossible

3) **Masking techniques**

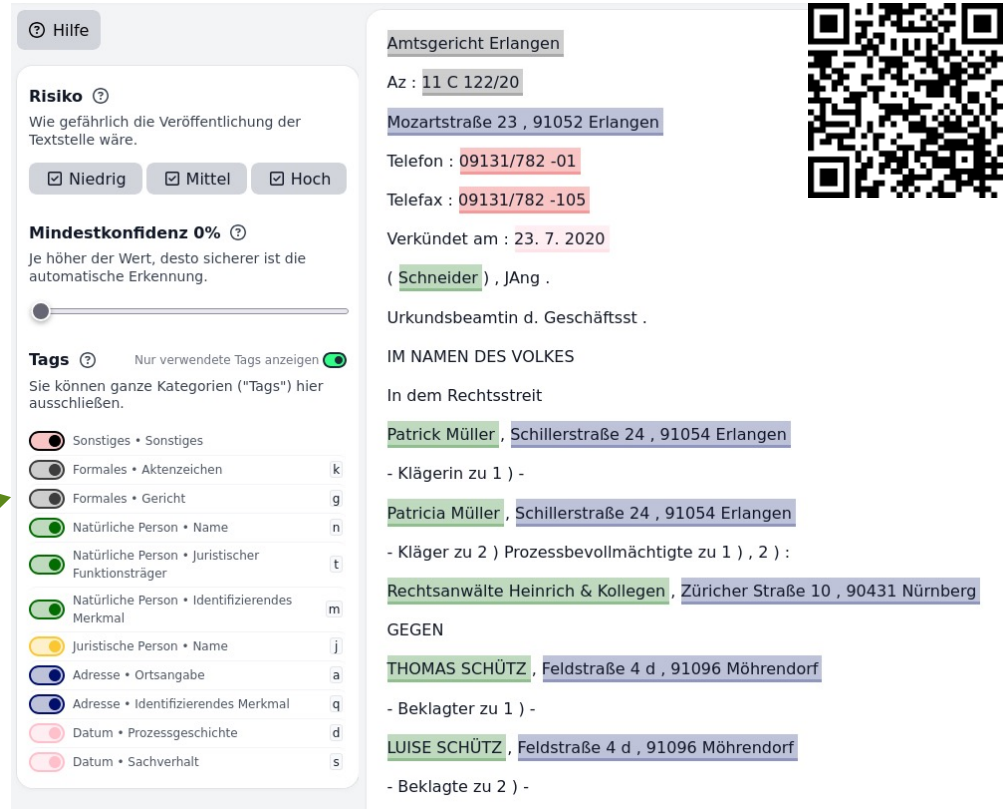
– simple techniques such as initials leak information, but also randomised initials (*the D&B shop*)

– realistic surrogates are considered safe, but challenging for quasi-identifiers

LeAK & AnGer: Anonymisierung von Gerichtsurteilen

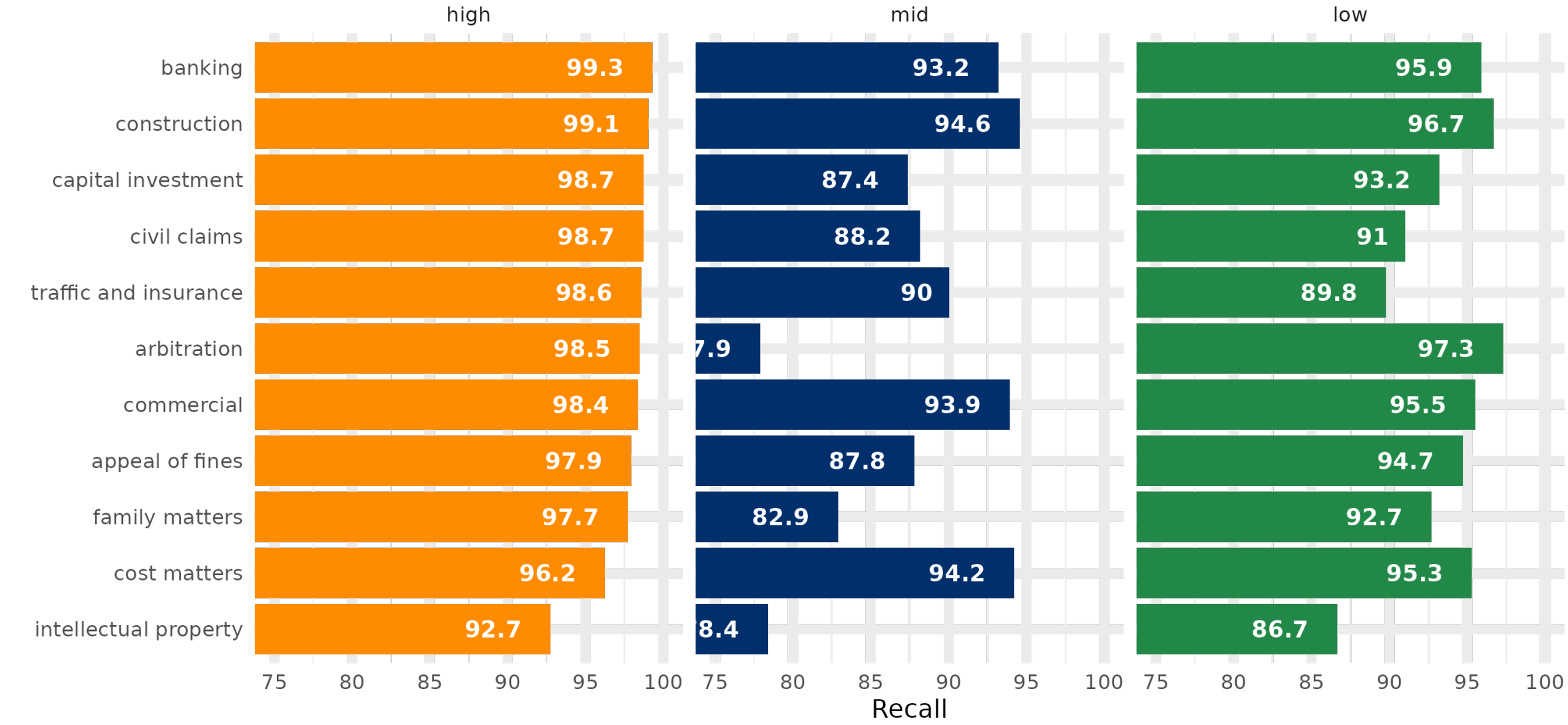
Project leads: Prof. Dr. Stephanie Evert, Prof. Dr. Axel Adrian (FAU Erlangen-Nürnberg)

- **Goal**: fully automatic anonymisation → only way to publish as many as 1.5 million German court verdicts / year
- **Challenge**: extremely high recall (> 99%) for PII mandatory, deep text anonymisation also has to cover quasi-identifiers
- **Approach**: fine-tuning of pretrained LLMs for span identification and categorisation (multi-task)
- **Gold standard**: unusally high quality essential for evaluation and training (6 annotators / adjudicators for each text)
- **Result**: automatic anonymisation is **feasible**, but needs training or adaptation for each legal domain & court type
- **Team**: Bao Minh Doan Dang, Philipp Heinrich, Michael Keuchen, Mahdi Mantash, Daniel Odorfer, Melanie Rosa, Pei-Yu Shen, Naveed Unjum, Julian Werner, Leonardo Zilio + student assistants as annotators

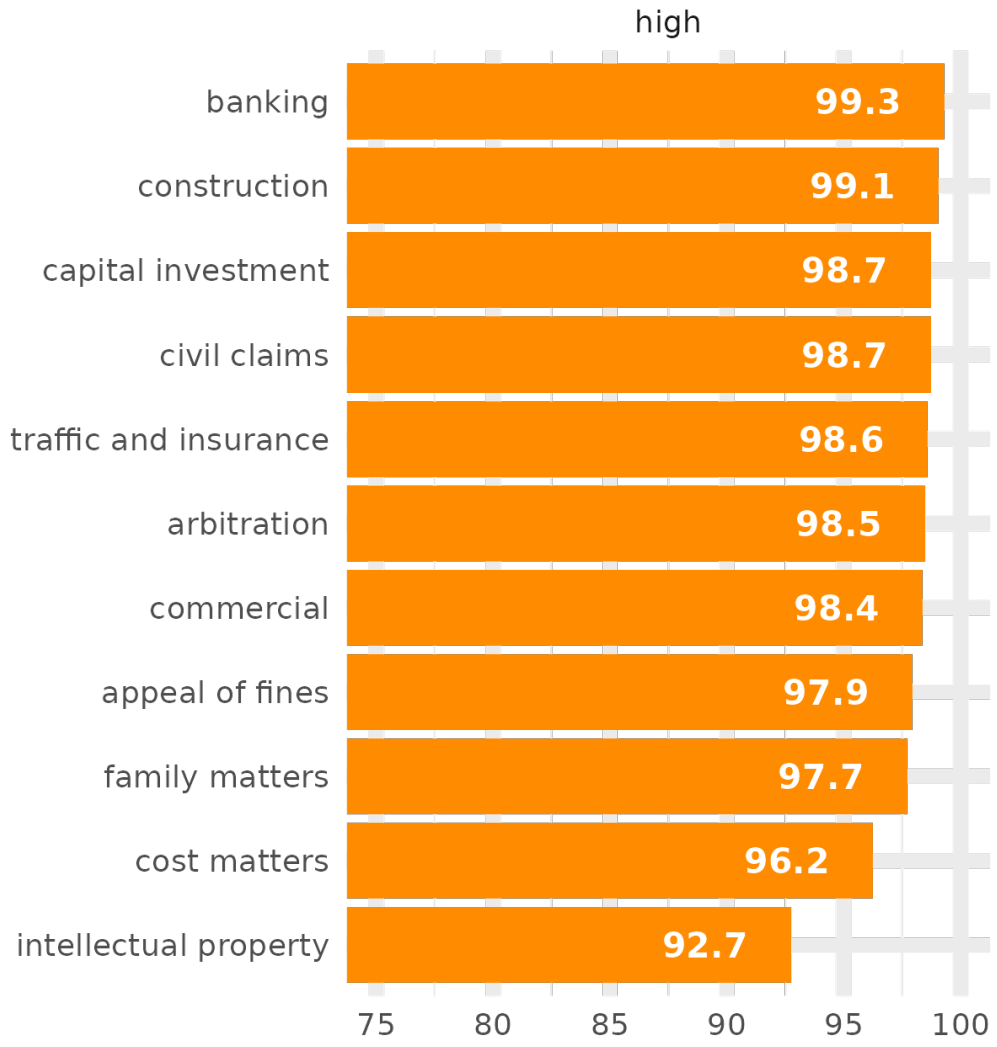


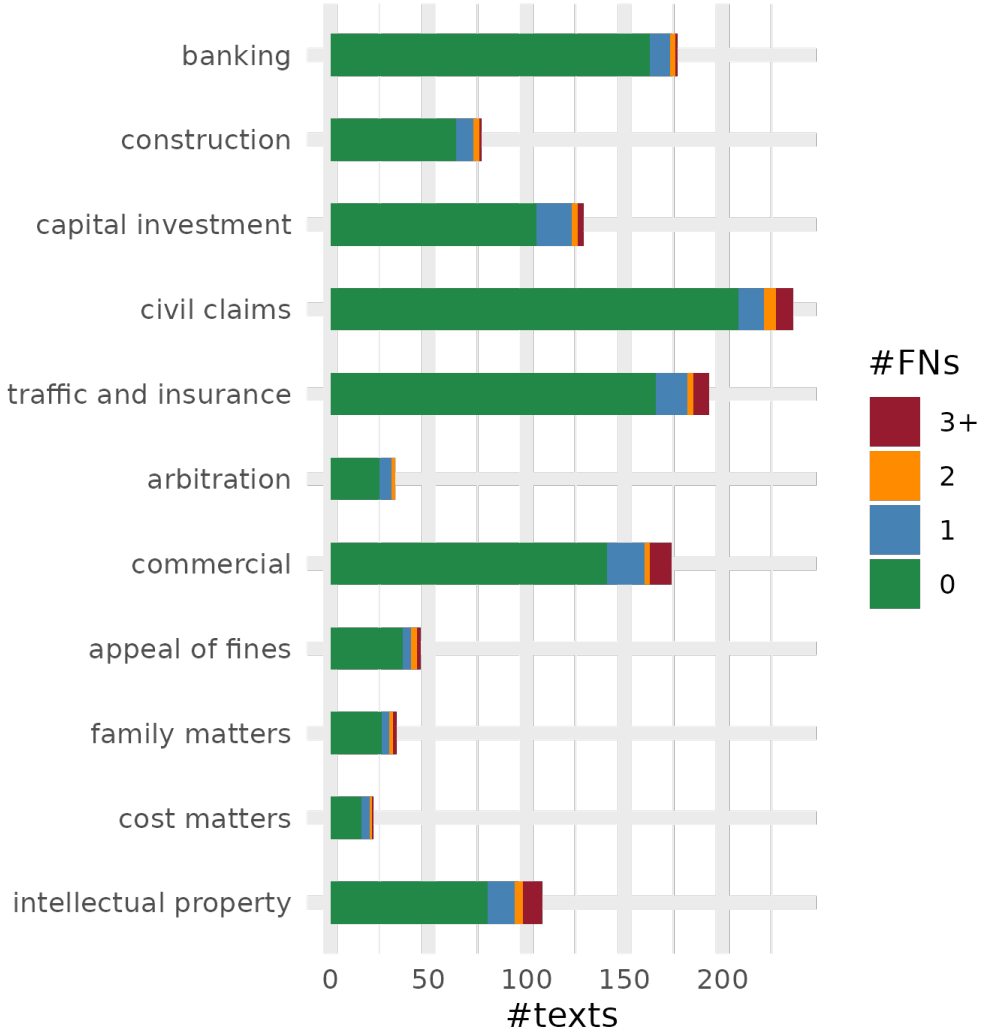
Domain (AG)	Precision	Recall	Recall PII
tenancy law	97.04%	96.05%	**98.90%**
traffic law	97.41%	97.38%	**99.11%**

AnGer: 11 legal domains at higher regional court (OLG)



PII deanonymisation risk





Deanonymisation risk from quasi-identifiers

Experiment: **empirical deanonymisation** by human adversaries

– 10 verdicts from district court (AG) attacked by 6 human adversaries (time limit: 35 minutes / verdict)

– verdicts from AG gold standard chosen to contain large number of quasi-identifiers with elevated risk

Results:

– no successful reidentification of a natural person

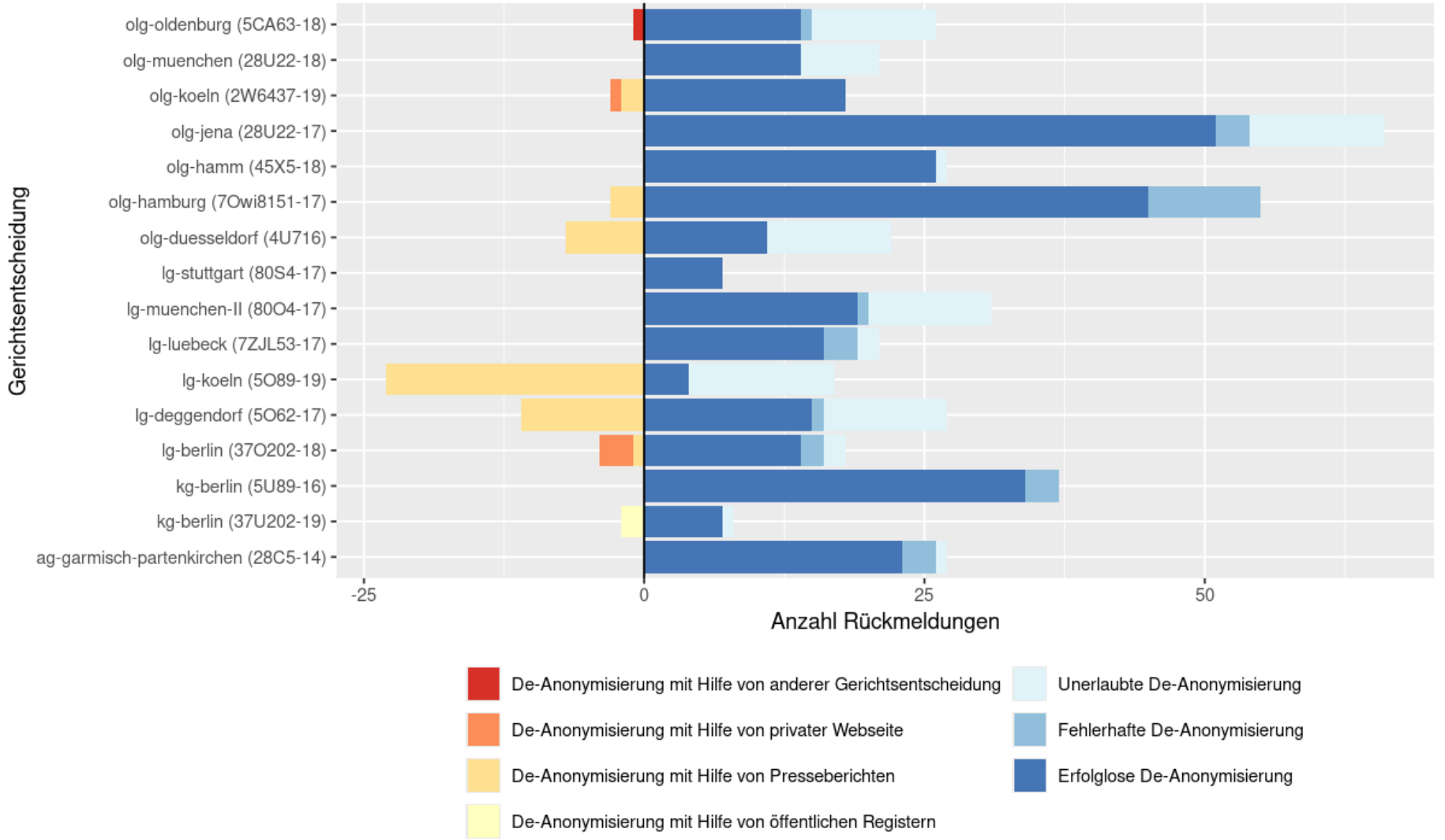
– many spurious deanonymisation results (existing family name, existing street name or address, *I+D-Versicherung* → *R+V-Versicherung*, press reports of similar accident exactly one month off)

– success: format of case reference → **insurance company**

– success: index of rents (average cost, area type "red") → **specific city**

Residual deanonymisation risk

- 17 published verdicts of more prominent cases at regional courts
- known to offer attack vectors (previous work)
- deep anonymisation via LeAK/Anger guidelines with realistic surrogates (manual annotation)
- small number of texts with successful attacks
- mostly due to press coverage of the case



Deanonymisation risk due to masking technique

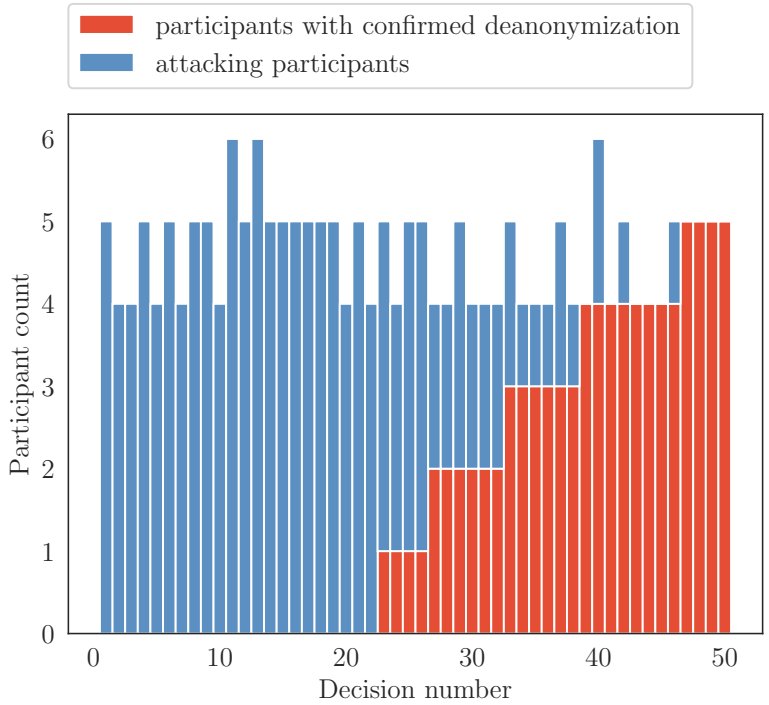
(Deuber et al. 2023)

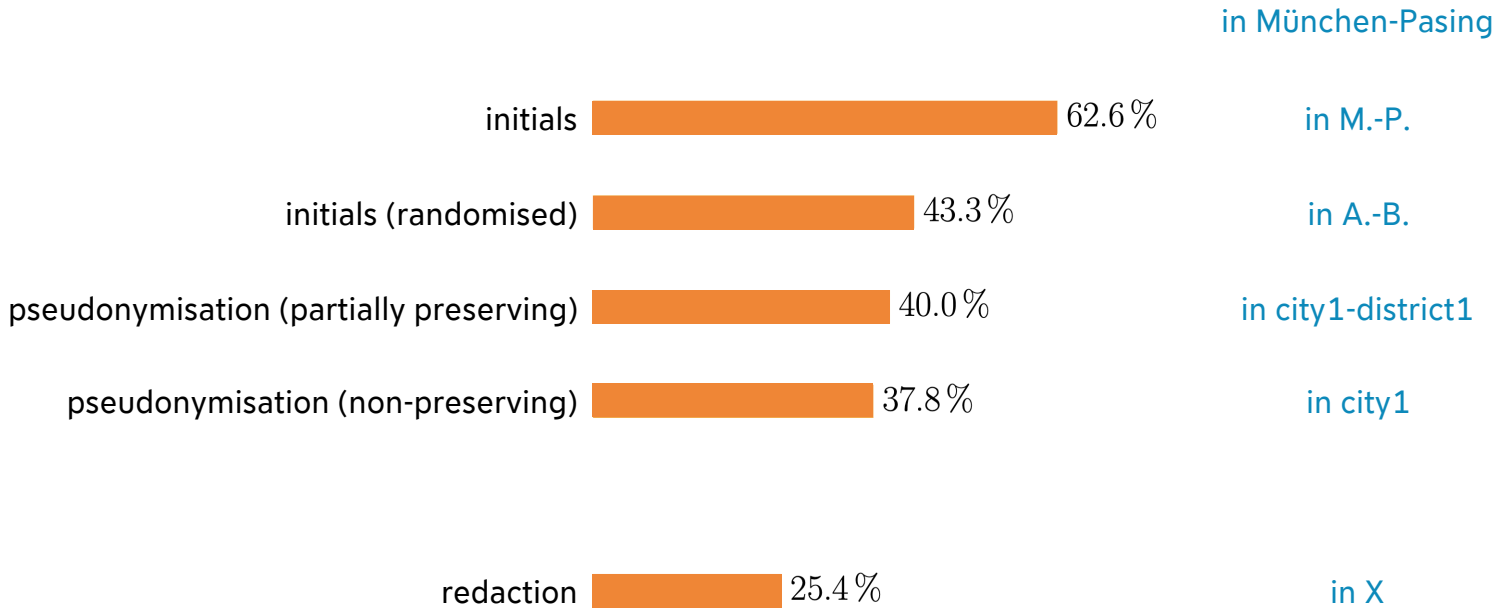
Experiment: **empirical deanonymisation** by human adversaries

– 50 published verdicts using different masking techniques (identified with heuristics)

– strong publication bias towards higher courts (due to random selection of eligible verdicts)

– human adversaries: 54 law students in online experiment | time limit: 35 minutes per verdict





Friedrich-Alexander-Universität
Erlangen-Nürnberg

FAU

Thanks for listening!

https://www.linguistik.phil.fau.de/projects/leak-anger/

References

Adrian, A., Dykes, N., Evert, S., Heinrich, P., and Keuchen, M. (2023). Automatische Anonymisierung von Gerichtsurteilen: Eine Vision scheint realisierbar. In *Rechtsinformatik als Methodenwissenschaft des Rechts – Tagungsband des 26. Internationalen Rechtsinformatik Symposions IRIS 2023*. Editions Weblaw.

Adrian, A., Evert, S., Doan Dang, B. M., Heinrich, P., Mantash, M., Odorfer, D., Rosa, M., Shen, P.-Y., and Werner, J. (2025). Robustheit und Domänenanpassung bei der automatischen Anonymisierung von Gerichtsentscheidungen. *Künstliche Intelligenz und Recht (KIR)*, 2(2): 60–69.

Deuber, D., Keuchen, M., and Christin, N. (2023). Assessing anonymity techniques employed in German court decisions: A de-anonymization experiment. In *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA.

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42(4).

Lison, P., Pilán, I., Sanchez, D., Batet, M., and Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203, Online. Association for Computational Linguistics.

Manzanares-Salor, B. and Sánchez, D. (2025). Enhancing text anonymization via re-identification risk-based explainability. *Knowledge-Based Systems*, 310: 112945.

Manzanares-Salor, B., Sánchez, D., and Lison, P. (2024). Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack. *Data Mining and Knowledge Discovery*, 38(6): 4040–4075