ARX

A Comprehensive Tool for Anonymizing Biomedical Data

Fabian Prasser, Florian Kohlmayer, Klaus A. Kuhn

Chair of Biomedical Informatics
Institute of Medical Statistics and Epidemiology

Rechts der Isar Hospital
Technische Universität München

Today‘s presenters

Florian Kohlmayer

Fabian Prasser

- Computer scientists, background in IT security and database systems
- Research assistants at the Chair for Biomedical Informatics at TUM
- Core-developers of ARX

Today‘s agenda

- **Introduction**
- **Demonstration**
- **Questions & answers**

Motivation: Data sharing in biomedical research

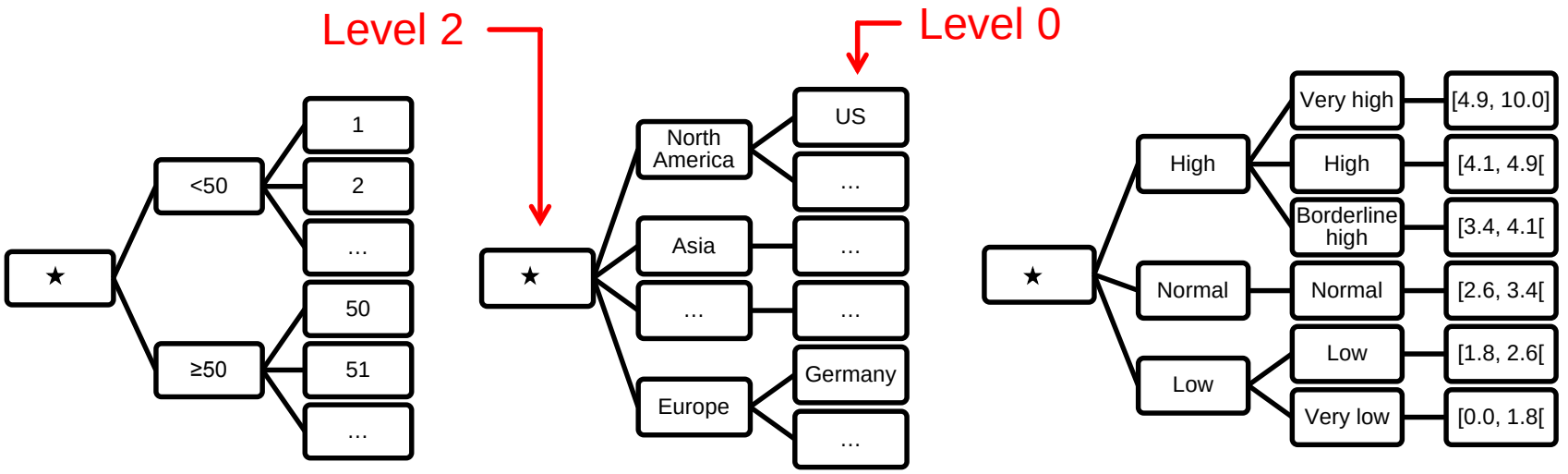
- **Data sharing is a core element of biomedical research**
  - **Wellcome Trust:** Sharing research data to improve public health [1]
  - **OECD:** Principles and guidelines for access to research data from public funding [2]
- **Disclosure of data may lead to harm for individuals**
  - Data may be person-related and highly sensitive
- **Large body of laws & regulations mandates privacy protection**
  - **US:** HIPPA Privacy Rule
  - **EU:** European Data Protection Regulation
  - **DE:** German Federal Data Protection Act
- **Safeguards**
  - Access control, policies, agreements, …
  - De-identification / anonymization

Overview: De-identification / anonymization

- **Controlling interactive data analysis**
  - **Subject:** query results, …
  - **Methods:** differential privacy, query-set-size control, …
  - **Implementations:** Fuzz, PINQ, Airavat, HIDE
- **Masking identifiers in unstructured data**
  - **Subject:** clinical notes, …
  - **Methods:** machine learning, regular expressions, …
  - **Implementations:** MIST, MITdeid, NLM Scrubber
- **Transforming structured data** *(focus of ARX)*
  - **Subject:** tabular data, …
  - **Methods:** generalization, suppression, …
  - **Implementations:** ARX, sdcMicro, PARAT

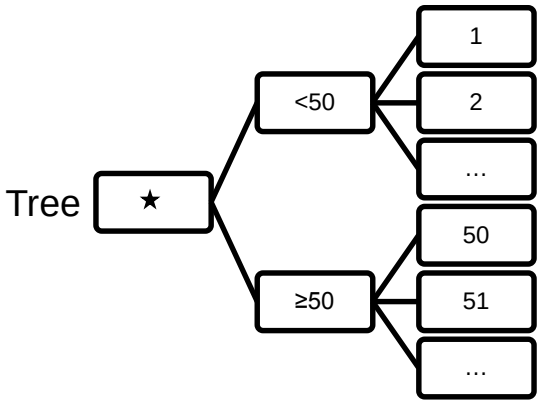
Background: Transforming structured data

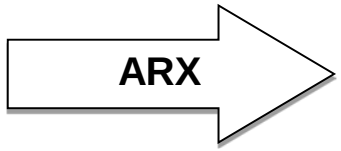
- **Basic idea:** transform datasets in such a way that they adhere to a set of formal privacy guarantees
- **Typical transformations**
  - **Generalization:** Germany → Europe (often applied to individual values)
  - **Suppression:** Germany → * (often applied to whole entries)
- **Example generalization hierarchies**



Background: Transforming structured data (cont.)

- **Representation of gen. hierarchies**

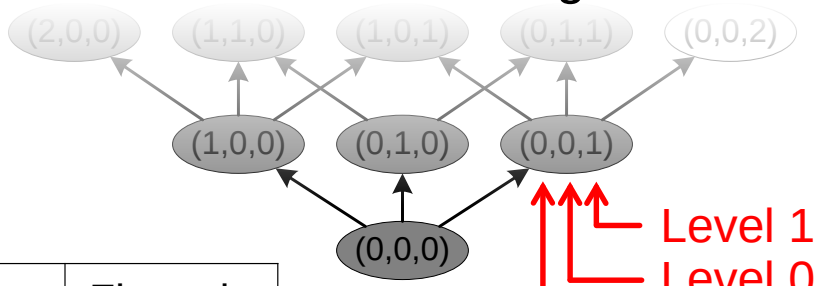




Tabular representation

Level 0	Level 1	Level 2
1	<50	*
2	<50	*
…	<50	*
50	≥50	*
51	≥50	*
…	≥50	*

- **Full-domain generalization:** all values of an attribute are generalized to the same level of the associated generalization hierarchy
- **Search space:** combinations of all generalization levels (lattice)



Schema:

Age	Gender	Zip code

Level 1 for attribute “zip code”
Level 0 for attribute “gender”
Level 0 for attribute “age”

Background: Privacy models

- **Well-known models:** k-anonymity, ℓ-diversity, t-closeness, δ-presence
- **Example:** k-anonymity
  - Proposed by Samarati and Sweeney in 1998 [3]
  - Attacker model: linkage via a set of quasi-identifiers (identity disclosure)
  - Mitigated by: building groups of indistinguishable data entries
  - Adherence can be achieved with generalization and suppression

Age	Gender	Zip code
34	male	81667
45	female	81675
34	female	81931
45	male	81925
70	female	81931
70	male	81931
66	male	80931

2-anonymity

Generalization (1,1,2)

Age	Gender	Zip code
< 50	*	816**
< 50	*	816**
< 50	*	819**
< 50	*	819**
≥ 50	*	819**
≥ 50	*	819**
*	*	*

Suppression

Background: k-Anonymity

Original dataset

Age	Gender	Zip code
34	male	81667
45	female	81675
34	female	81931
45	male	81925
70	female	81931
70	male	81931
66	male	80931

Background knowledge, e.g. voter list

Age	Gender	Zip code	Name
45	female	81675	Alice
45	male	81925	Bob
70	male	81931	Charlie
⋮	⋮	⋮	⋮

2-anonymous dataset

Age	Gender	Zip code
< 50	*	816**
< 50	*	816**
< 50	*	819**
< 50	*	819**
≥ 50	*	819**
≥ 50	*	819**
*	*	*

?

Background knowledge, e.g. voter list

Age	Gender	Zip code	Name
45	female	81675	Alice
45	male	81925	Bob
70	male	81931	Charlie
⋮	⋮	⋮	⋮

Challenge: Tool support

- **Situation:** anonymization of structured data is frequently recommended (laws, regulations, guidelines) but in practice it is only used rarely
- **Main reasons**
  - Lack of understanding of opportunities and limitations
  - Lack of ready-to-use tools
- **Non-trivial:** implementing useful tools is challenging
- **Usefulness has many dimensions**
  - Ability to balance data utility with privacy requirements
  - Support a broad spectrum of privacy methods, transformation techniques and methods for measuring and analyzing data utility
  - Performance and scalability
  - Intuitive visualization and parameterization of all process steps
  - Provide methods to end-users as well as programmers
  - In an integrated and harmonized manner
  - Openness

Challenge: Related software

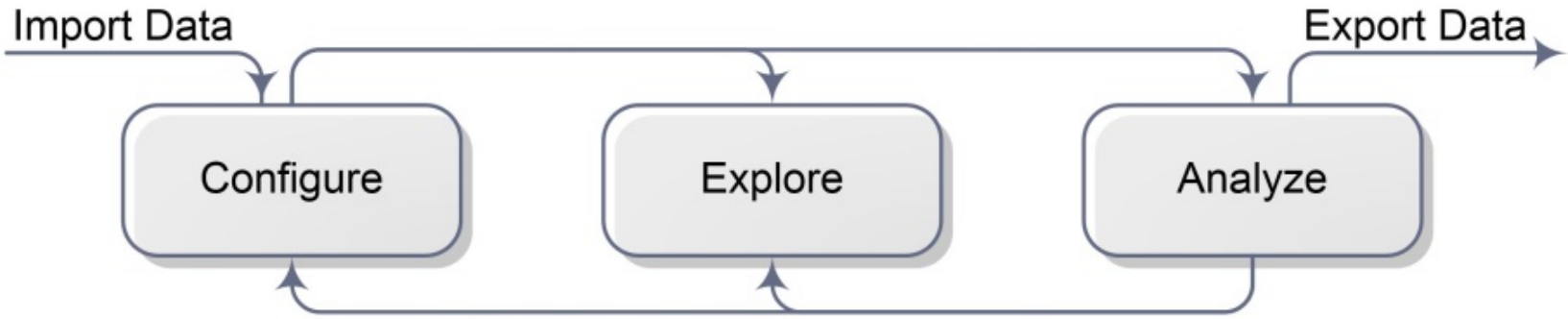
- **sdcMicro**
  - Cross-platform open source software implemented in “R”
  - Collection of a set of methods, not an integrated application
  - Different types of recoding models and risk models
  - Minimalistic graphical user interface
- **µArgus**
  - Closed source software for MS Windows
  - Methods comparable to sdcMicro but more comprehensive user interface
  - Development has ceased
- **PARAT**
  - Commercial tool for MS Windows
  - Powerful graphical interface
  - Methods implemented overlap with methods implemented in ARX
  - Centered around a risk-based approach
- **More comprehensive list:** http://arx.deidentifier.org/related-software/

ARX: Highlights

- **Flexible transformation methods:** generalization and suppression in a parameterizable and utility-driven manner
- **Multiple privacy models:** k-anonymity, ℓ-diversity (three variants), t-closeness (two variants) and δ-presence, as well as arbitrary combinations
- **Multiple methods for measuring data utility:** automatically as well as manually
- **Optimality**: classification of the complete solution space
- **Functional generalization rules:** support for continuous and discrete variables
- **Highly scalable:** several million data entries on commodity hardware
- **Comprehensive cross-platform Graphical User Interface:** wizards, visualization of the solution space, analysis of data utility
- **Application Programming Interface:** full-blown Java library

ARX: Anonymization workflow

- Iteratively refine the anonymization process
- Supported by the scalability of our framework
- Three (potentially repeating) steps

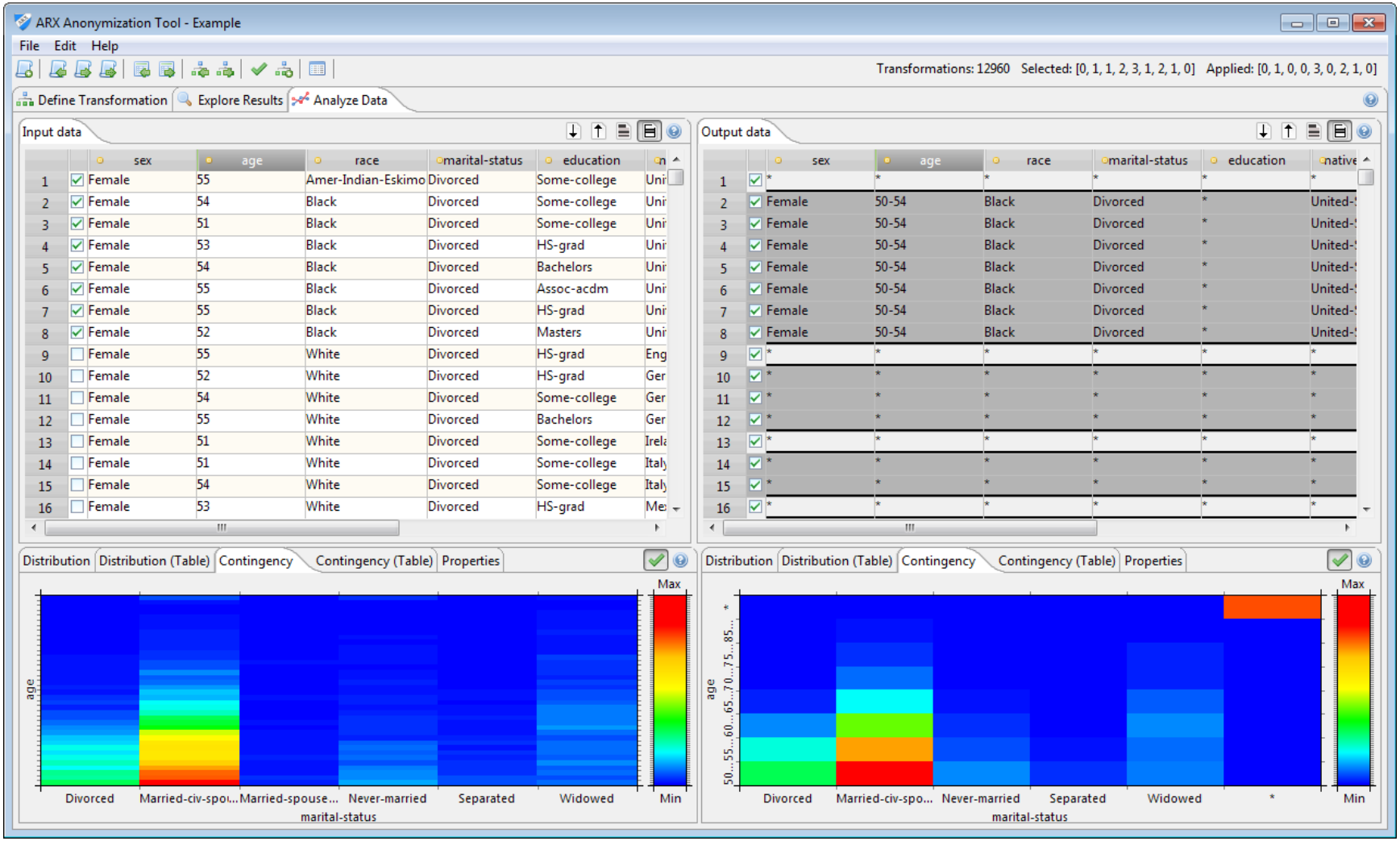


- Create and edit rules
- Define privacy guarantees
- Parameterize coding model
- Configure utility measure

- Filter and analyze the solution space
- Organize transformations

- Compare and analyze input and output

ARX: Demo

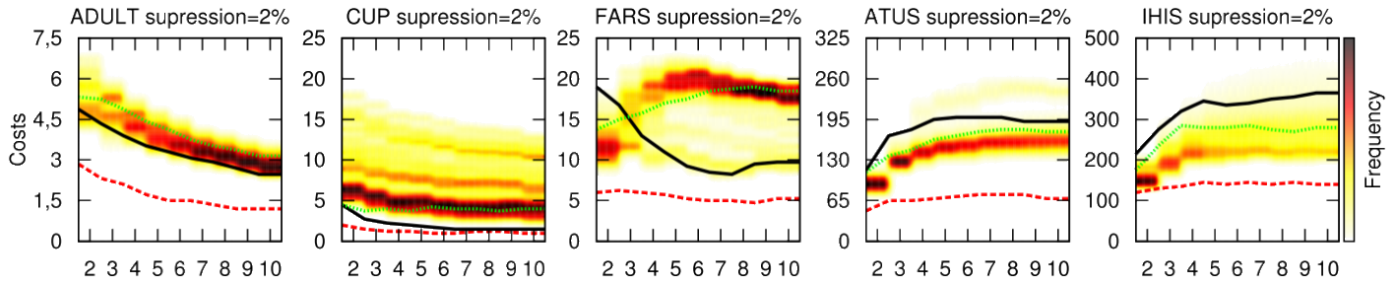

ARX Anonymization Tool - Example
File Edit Help
Transformations: 12960 Selected: [0, 1, 1, 2, 3, 1, 2, 1, 0] Applied: [0, 1, 0, 0, 3, 0, 2, 1, 0]
Define Transformation
Explore Results
Analyze Data
Input data
Output data
sex age race marital-status education
Distribution Distribution (Table) Contingency Contingency (Table) Properties
Max
Min
age
marital-status
Divorced
Married-civ-spou...
Married-spouse...
Never-married
Separated
Widowed

ARX: Facts and credits

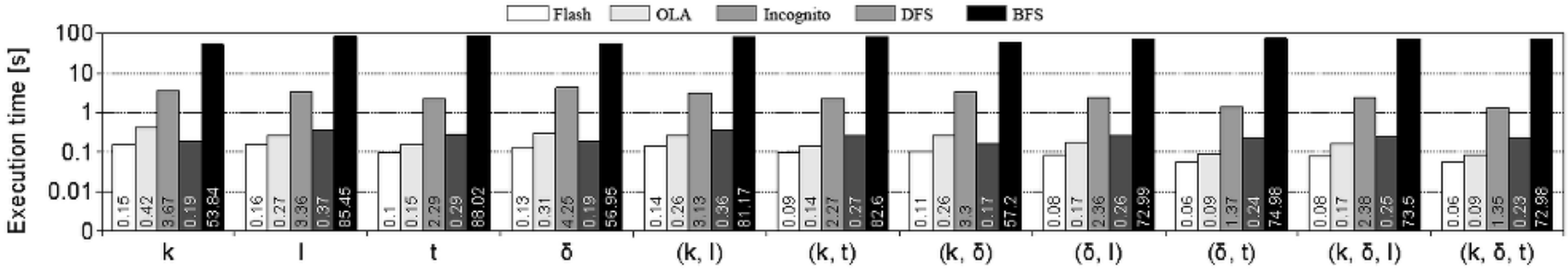
- **Three years of work by two main developers:** Florian Kohlmayer and Fabian Prasser
- **With help from multiple students:** see credits on our website
- **Interdisciplinary cooperation:** Chair for IT Security, Chair for Database Systems, Chair for Biomedical Informatics
- **Code metrics**
  - ARX Core/API: 178 files, 200 classes
    37,332 LOC (16,281 lines of comments)
  - ARX GUI: 174 files, 207 classes
    44,772 LOC (15,062 lines of comments)
  - ARX Tests: 997 JUnit tests
  - Commits: 1,722 commits (since 03/2013)

ARX: Publications

- **Implementation framework:** Proc Int Symp CBMS, 2012 [4]
- **Anonymization algorithm:** Proc Int Conf PASSAT, 2012 [5]



- **Anonymization of distributed data:** J Biomed Inform, 2013 [6]
- **Benchmark for anonymity methods:** Proc Int Symp CBMS, 2014 [7]



- **"White Paper":** AMIA Annu Symp Proc, 2014 [8]

Thank you for your attention! Questions?

- **Disclaimer**
  - Anonymization must be performed by experts
  - Additional safeguards are required (e.g., contractual measures)
- **ARX is open source software**
  - Contributions are welcome, e.g., feature requests, code reviews, criticism, enhancements, questions
- **Future developments:** various projects, especially risk models
- **Resources**
  - **Project website:** http://arx.deidentifier.org
  - **Code repository:** https://github.com/arx-deidentifier/arx
  - **Get in touch**
    - Fabian Prasser (prasser@in.tum.de)
    - Florian Kohlmayer (florian.kohlmayer@tum.de)

References

1. Wellcome Trust. Sharing research data to improve public health. 2013. [cited 2015 Jan 14]. Available at: http://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTDV030690.htm.
2. OECD. Principles and guidelines for access to research data from public funding. 2006. [cited 2015 Jan 14]. Available at: www.oecd.org/sti/sci-tech/38500813.pdf.
3. Pierangela Samarati, Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Proc Symp Res Secur Priv 1998.
4. Florian Kohlmayer*, Fabian Prasser*, Claudia Eckert, Alfons Kemper and Klaus A. Kuhn. Highly efficient optimal k-anonymity for biomedical datasets. Proc Int Symp CBMS 2012.
5. Florian Kohlmayer*, Fabian Prasser*, Claudia Eckert, Alfons Kemper, Klaus. A. Kuhn. Flash: efficient, stable and optimal k-anonymity. Proc Int Conf PASSAT 2012.
6. Florian Kohlmayer*, Fabian Prasser*, Claudia Eckert, Klaus A. Kuhn. A flexible approach to distributed data anonymization. J Biomed Inform 2013.
7. Fabian Prasser*, Florian Kohlmayer*, Klaus A. Kuhn. A benchmark of globally-optimal anonymization methods for biomedical data. Proc Int Symp CMBS 2014.
8. Fabian Prasser*, Florian Kohlmayer*, Ronald Lautenschlaeger, Klaus A. Kuhn. ARX – A comprehensive tool for anonymizing biomedical data. AMIA Annu Symp Proc 2014.

*Equal contributors