Advances in linking sensitive and complex data

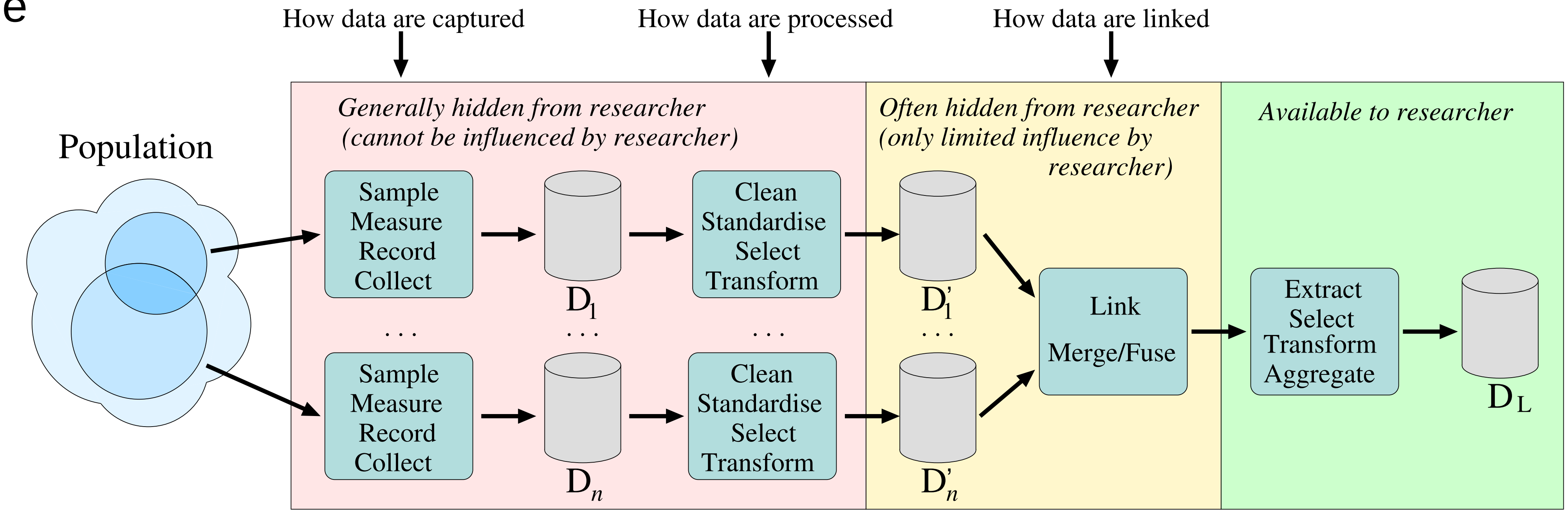
Addressing data quality, vulnerabilities, and scalability

Peter Christen and Charini Nanayakkara

School of Computing, the Australian National University; and
Scottish Centre for Administrative Data Research, the University of Edinburgh

Data quality of administrative data

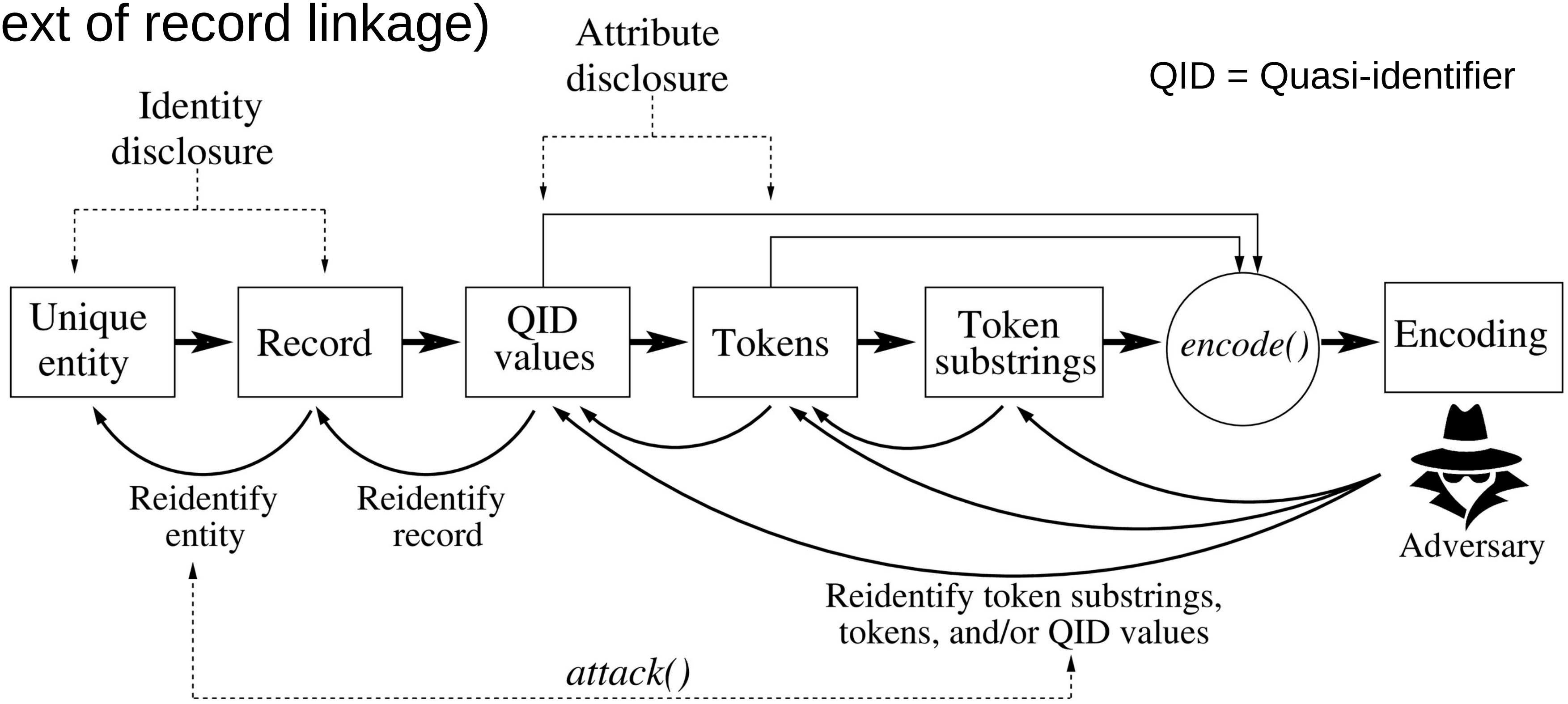
- Much of the administrative data workflow is outside the control of researchers
- Various misconceptions due to the social nature of data collection, processing, and linkage


How data are captured
How data are processed
How data are linked
Population
Generally hidden from researcher (cannot be influenced by researcher)
Often hidden from researcher (only limited influence by researcher)
Available to researcher
Sample Measure Record Collect
D_1
Clean Standardise Select Transform
D'_1
. . .
Sample Measure Record Collect
D_n
Clean Standardise Select Transform
D'_n
Link
Merge/Fuse
Extract Select Transform Aggregate
D_L

P. Christen and R. Schnell (2023): *Big Data is not the New Oil: Common Misconceptions about Population Data*, International Journal of Population Data Science, 8(1).

Vulnerabilities of sensitive data

- Data privacy and confidentiality is generally via statistical disclosure control methods
- What is ‘vulnerable’ data is less explored (especially in the context of record linkage)

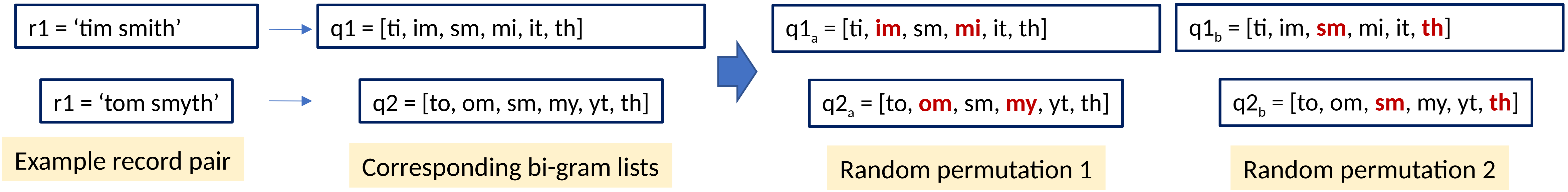


QID = Quasi-identifier

A. Vidanage, P. Christen, T. Ranbaduge, and R. Schnell (2023): *A Vulnerability Assessment Framework for Privacy-Preserving Record Linkage*, Under review, ACM Transactions on Privacy and Security.

Scalability of linking complex data

- Traditional ‘hand-crafted’ linkage techniques require domain expertise (such as knowledge about suitable attributes and their data quality)
- We use data agnostic methods such as *Locality Sensitive Hashing* (LSH), as developed for large-scale search engines and text mining systems
- Combining spatial and temporal constraints with LSH based filtering can substantially improve the scalability of linking complex data



r1 = ‘tim smith’ → q1 = [ti, im, sm, mi, it, th]

r1 = ‘tom smyth’ → q2 = [to, om, sm, my, yt, th]

$q1_a$ = [ti, **im**, sm, **mi**, it, th]

$q2_a$ = [to, **om**, sm, **my**, yt, th]

$q1_b$ = [ti, im, **sm**, mi, it, **th**]

$q2_b$ = [to, om, **sm**, my, yt, **th**]

Example record pair

Corresponding bi-gram lists

Random permutation 1

Random permutation 2

Because $q1_b$ and $q2_b$ are the same ([sm, th]) , r1 and r2 will be compared

C. Nanayakkara and P. Christen (2022): *Locality Sensitive Hashing with Temporal and Spatial Constraints for Efficient Population Record Linkage*, ACM International Conference on Information and Knowledge Management.

Conclusions and future directions

- The use of administrative data brings novel challenges for researchers (data quality, accessibility / sensitivity, large-scale processing)
- We need to ensure researchers are aware of any possible pitfalls
  - Understand the provenance of their data, and any limitations due to data cleaning, processing, and linkage outside of their control
  - Properly learn new methods (using traditional methods on larger data sets might not provide the best results)
  - Understand the limits of evaluation of linkage or classification accuracy
- We also need to better understand the limits of administrative data
  - Administrative data cannot answer the same questions as survey data (David Hand (2018): *Administrative data often tell us what people are and what they do, not what they say they are and what they claim to do.*)