A Tutorial on Population Informatics using Big Data

Peter Christen[1], Hye-Chung Kum[2], Qing Wang[1], and Dinusha Vatsalan[1]

[1] **Research School of Computer Science,**
The Australian National University, Canberra, Australia

[2] **School of Public Health and College of Engineering,**
Texas A&M University, College Station, TX, USA

Contacts: **peter.christen@anu.edu.au** / kum@tamu.edu /
qing.wang@anu.edu.au / dinusha.vatsalan@anu.edu.au

This work is partially funded by the Australian Research Council (ARC)
under Discovery Projects DP130101801 and DP160101934.

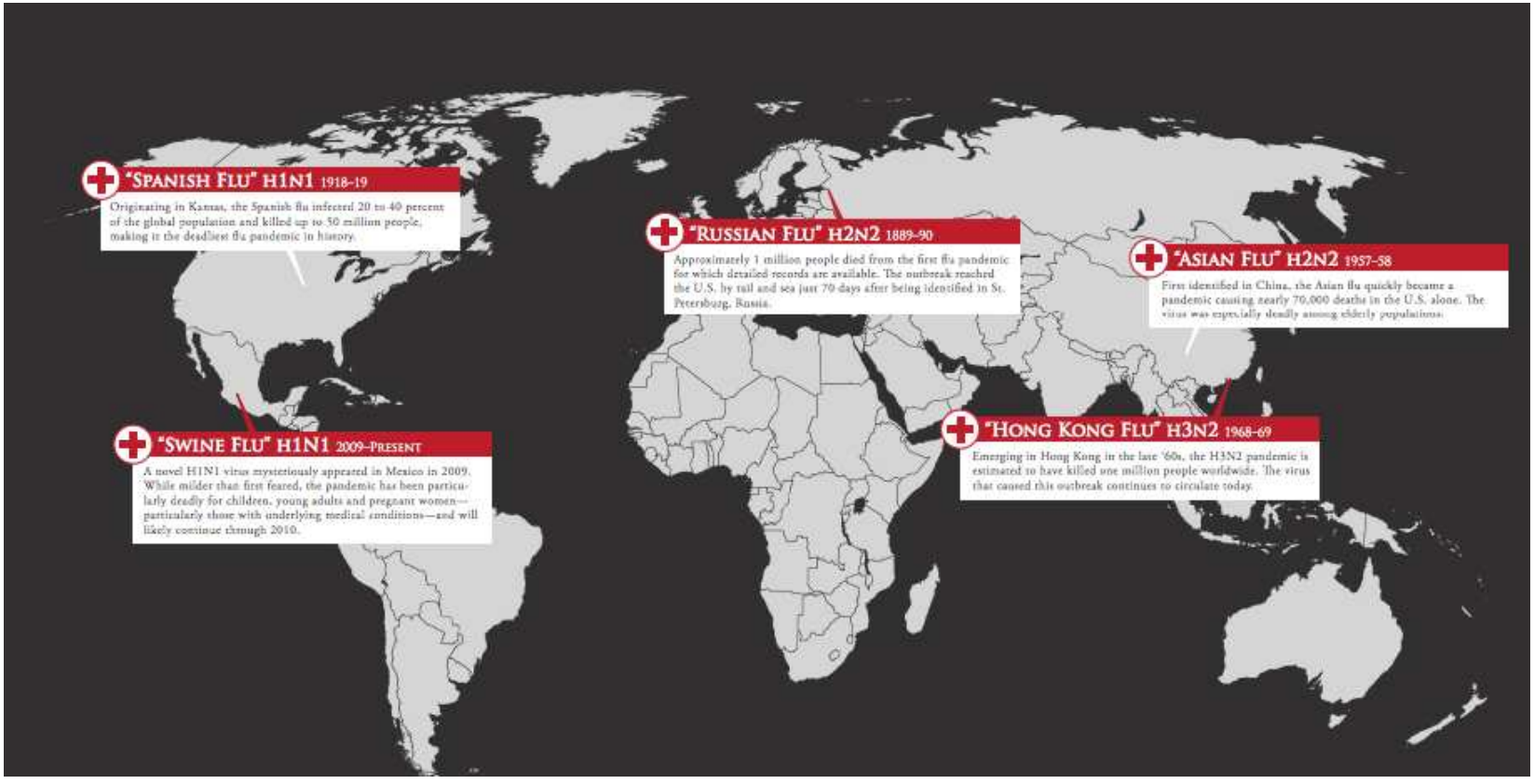
Motivation (1)

- Large amounts of data are being collected both by organisations in the private and public sectors, as well as by individuals
- Much of these data are about people, or they are generated by people
  - Financial, shopping, and travel transactions
  - Electronic health records
  - Tax, social security, and census records
  - Emails, tweets, SMSs, Facebook posts, etc.
- Analysing such data can provide huge benefits to businesses, governments and researchers

Motivation (2)

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To enrich data with additional information
  - To improve data quality
- Lack of unique entity identifiers means that linking is often based on personal information
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- Population informatics is concerned with studies of society (groups of people)

Motivating example: Health surveillance (1)



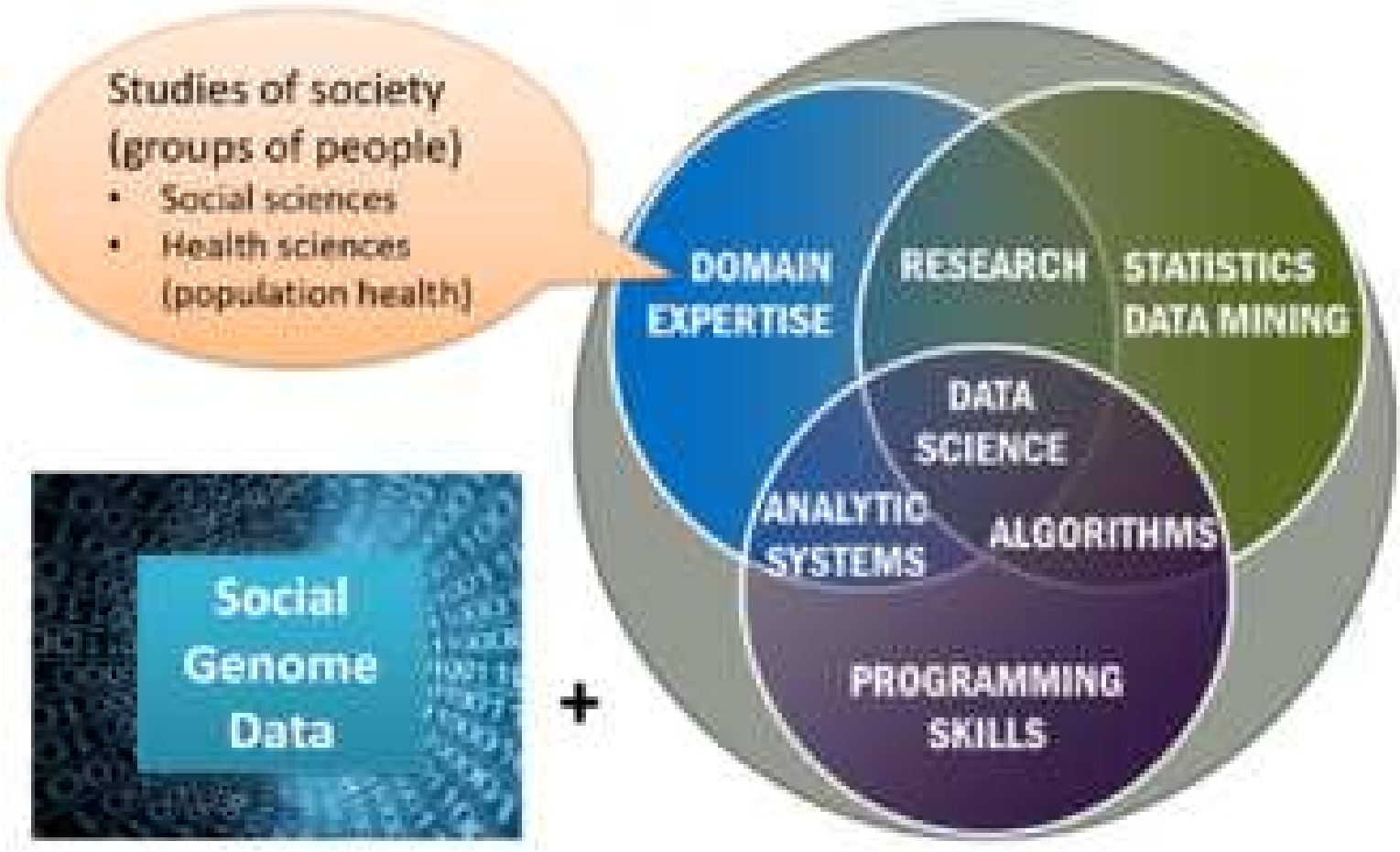
Motivating example: Health surveillance (2)

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns in symptoms, ideally in real time
- Data from many different sources will need to be collected continuously (including immigration and travel records; doctors, emergency and hospital admissions; drug purchases in pharmacies; social network data; possibly even animal health data)
- Privacy concerns arise if such data are stored and linked at a central location
- Such data sets are **large**, **dynamic**, **complex**, and **distributed**, and they require **linking** and **analysis** in near **real time**

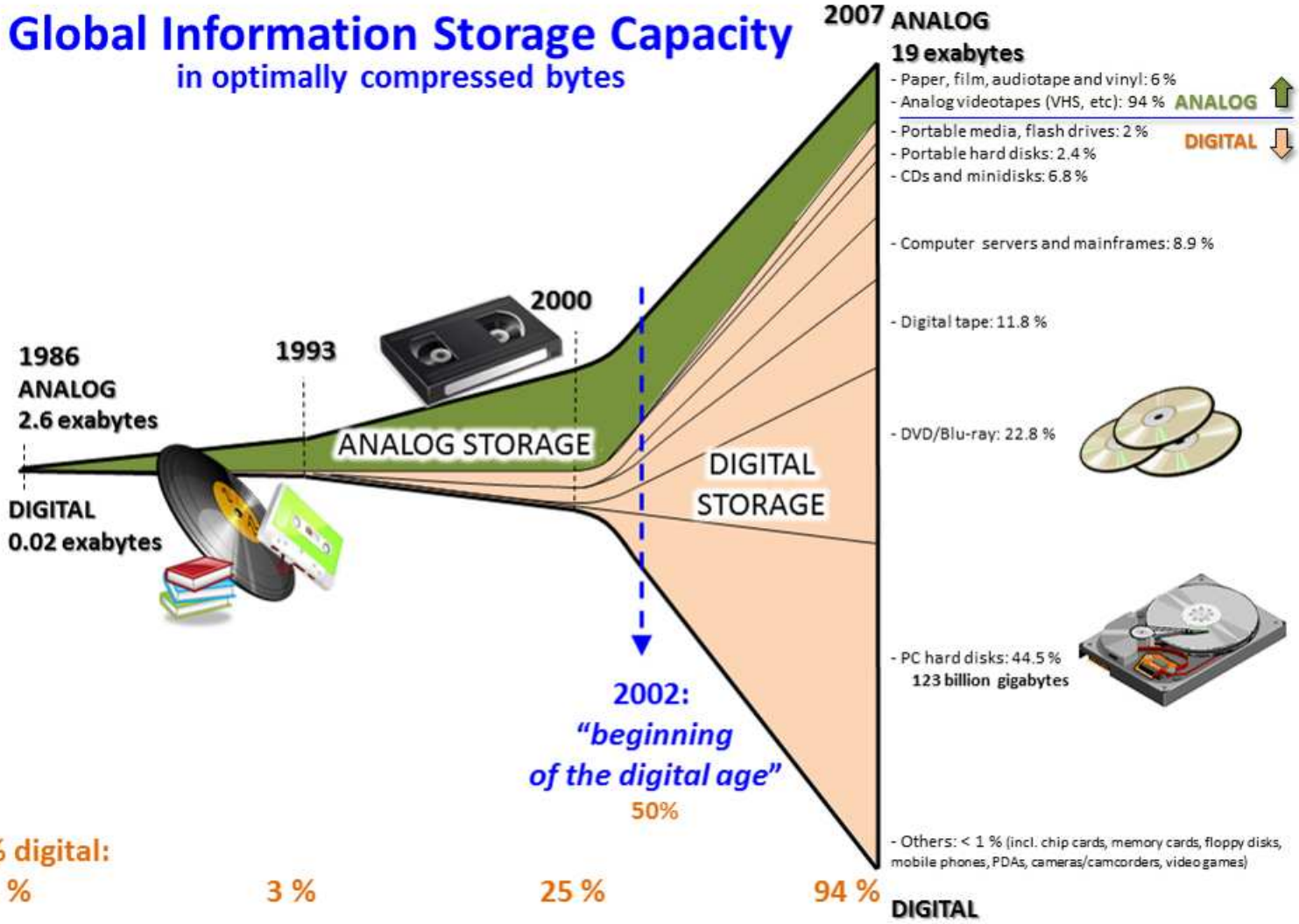
Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

What is Population informatics?



The digital society



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

What is a social genome?

- The collection of data about members of a society that are captured in ever-larger and ever-more complex databases (such as government administrative data, operational data, social media data etc.)

- Almost everything we do is recorded digitally

The cost of the digital society

- There is no turning back!
- Personal information is already being used without meaningful consent
  - Marketing: Target predicting pregnancies based on buying patterns
  - Campaigning: Obama 2012
  - Intelligence: Edward Snowden data leakage, Australian government metadata collection
- Facebook, Gmail, LinkedIn, etc. are *free*! Are they?

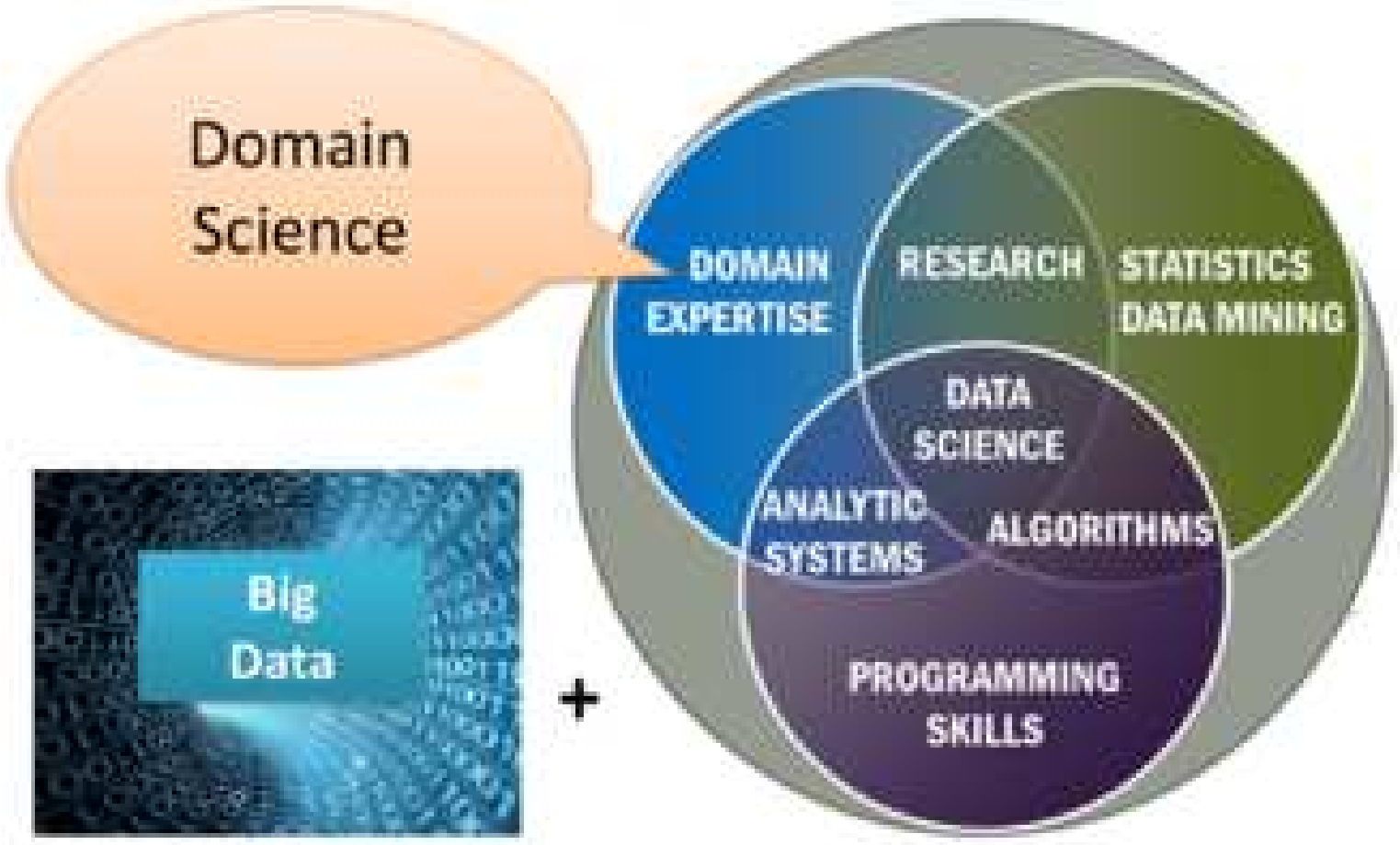
Why not reap the benefits too?

- The ability to answer questions about human populations *in near real time* using *distributed data sets* that are *large*, *complex*, *dynamic* and *diverse* has the potential to transform social, behavioural, economic, and health sciences.
- Could lead to more informed and effective policy decisions and allocations of public resources
  - What is the long term impact of moving to managed care?
  - What effect does teacher pay in middle school have on college grades?
- Answers could easily be derived from relevant data sets

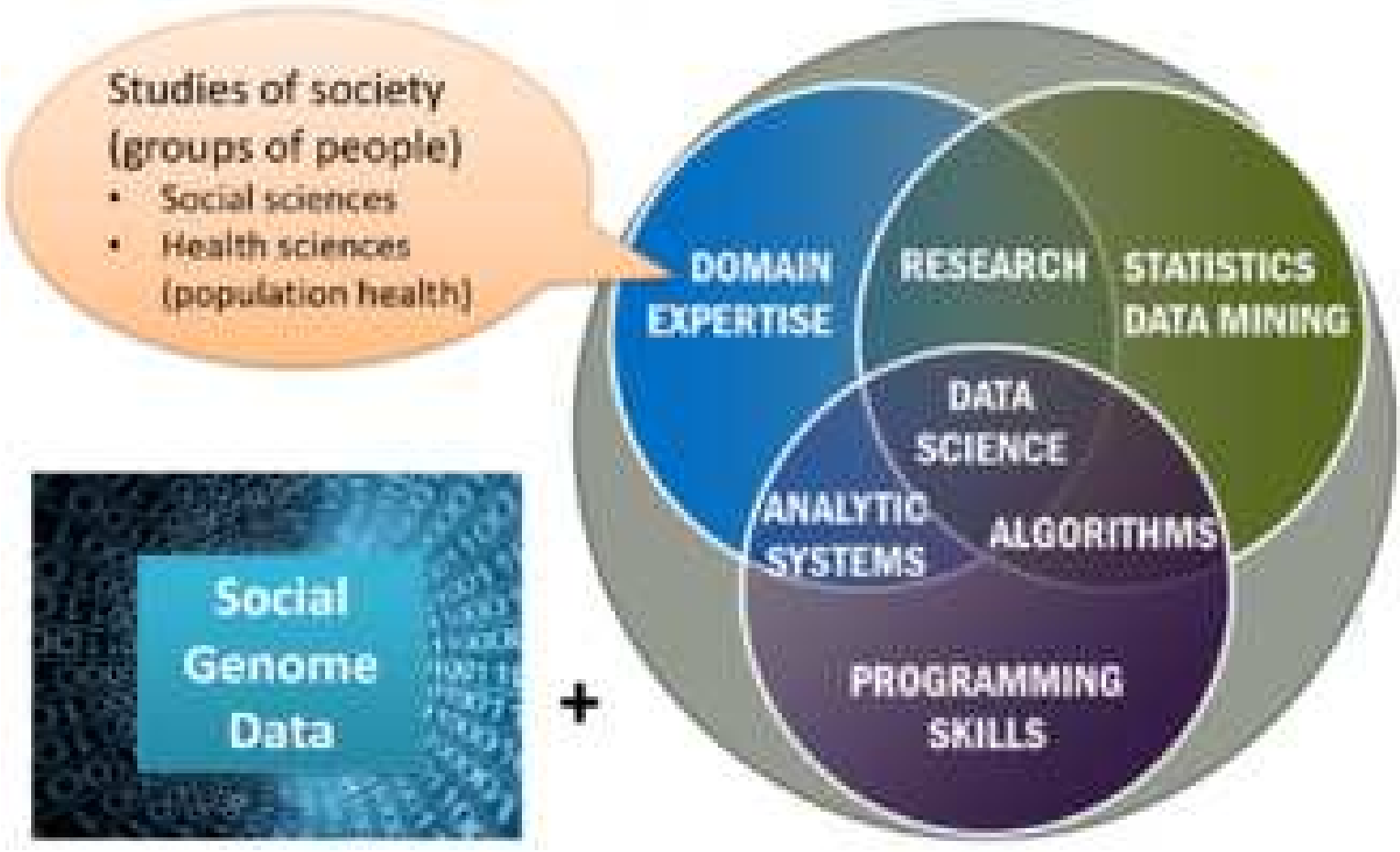
How do we reap the benefits too?

- Population informatics overarching question:
  - How can we use the abundance of existing digital data about people, aka Big Data (such as government administrative data, electronic health records, etc.) to support accurate evidence based decisions for policy, management, legislation, evaluation, and research,
  - while protecting the confidentiality of individual subjects of the data?

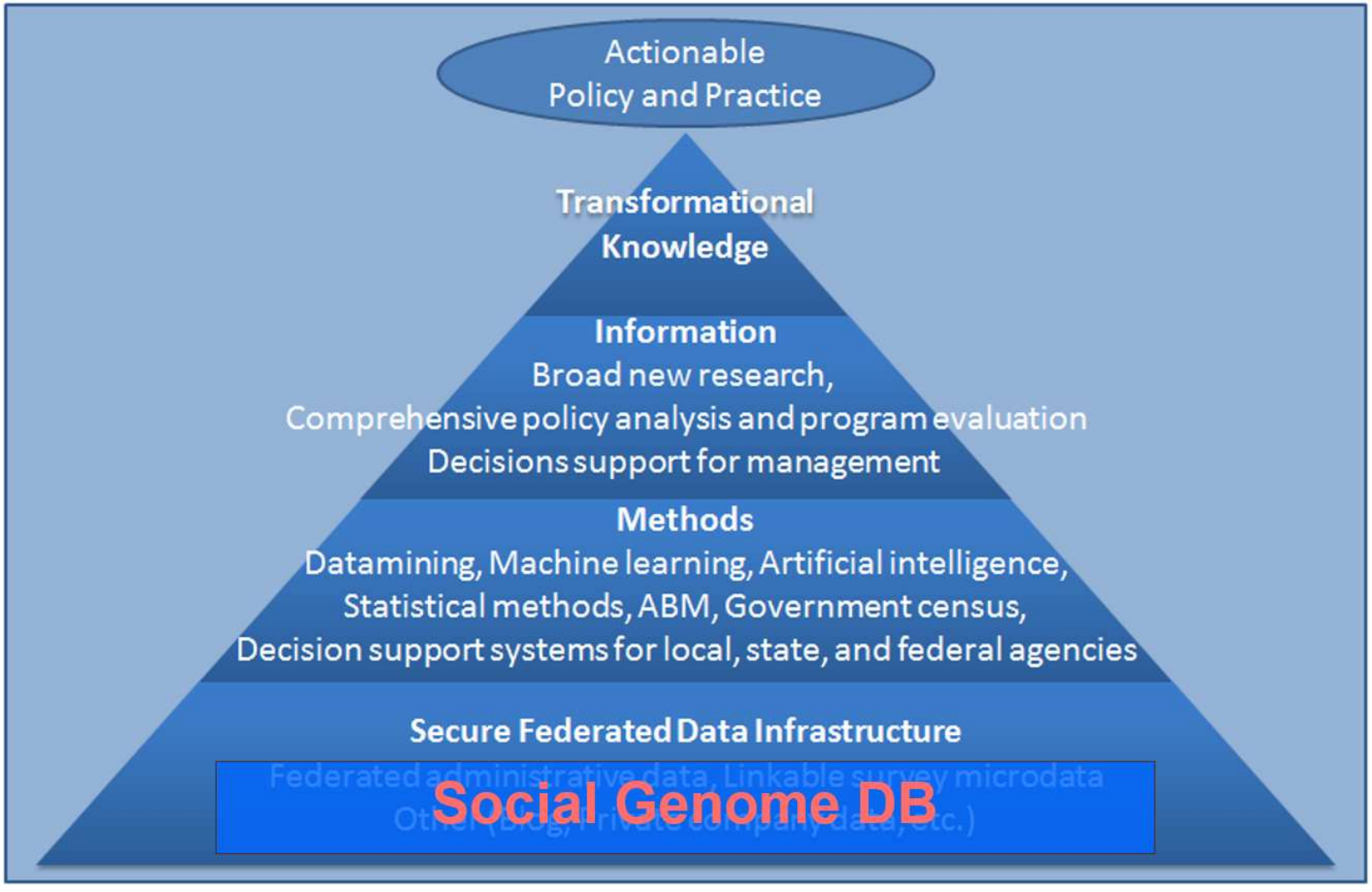
Primary methodology: Data science



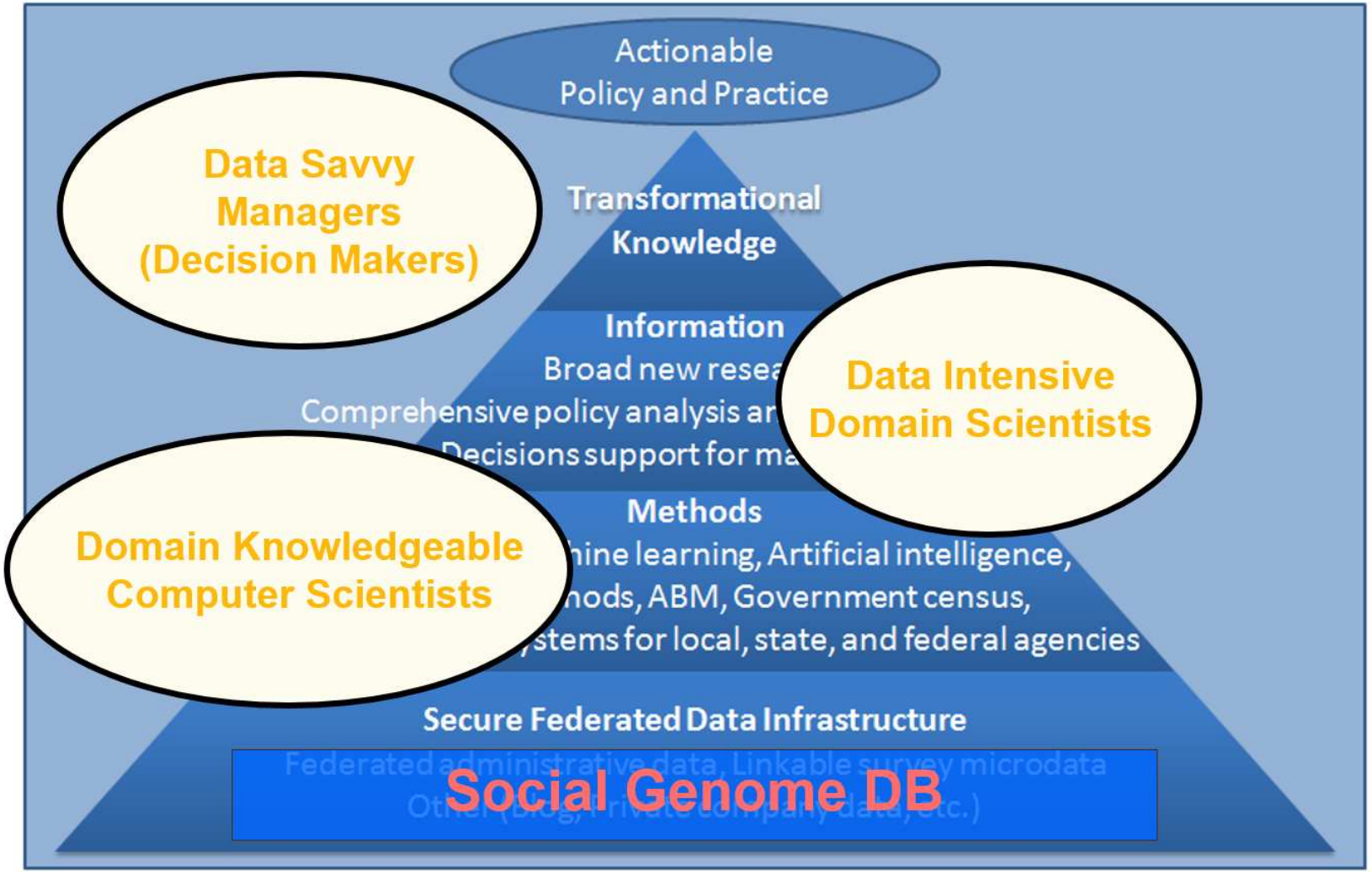
Population informatics



Population informatics hierarchy



Three types of data scientists



Is it new?

- Data Science for Social Good (DSSG)
- Computational social science
  - Population informatics
  - Simulations (such as agent based modelling)
- Population informatics
  - Business analytics
  - Social computing: social network data analysis
  - Policy informatics
  - Computational journalism
  - Computational transportation
  - Computational epidemiology
  - ...

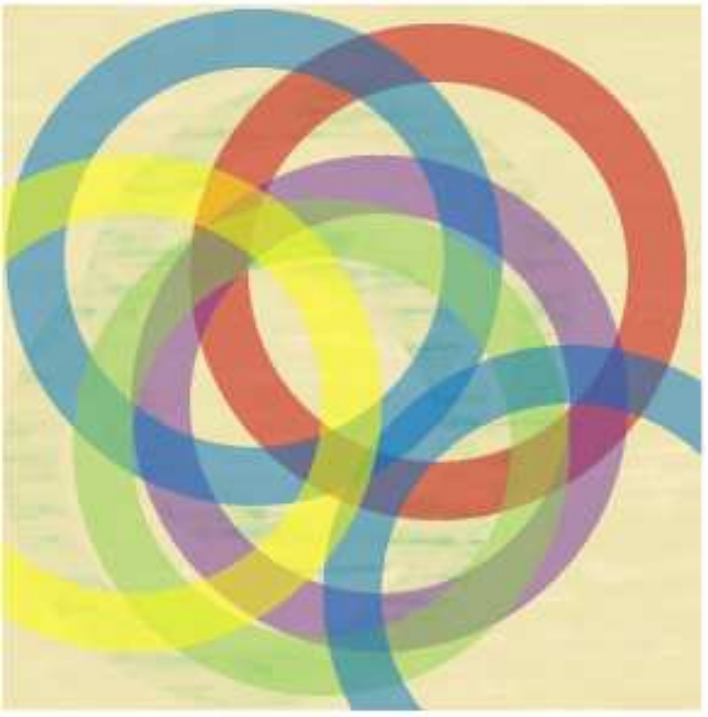
Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Traditional science vs Data science

Traditional science:
Start from nowhere

Data science:
Start from everywhere



Consumer Price Index (CPI)

- A statistical estimate constructed using the prices of a sample of representative items whose prices are collected periodically
- Calculated by most national statistical agencies

  For example, every year since 1913 by the U.S. Bureau of Labor Statistics
- The annual percentage change in a CPI is used as a measure of inflation
- A frequently asked question:

  What is the real level of inflation?

Inflation

- There was a widespread claim:
  - The Argentine government has been manipulating the official inflation indexes since 2007
- Questions:
  - Is it possible to evaluate this claim?
  - If possible, how can we evaluate it?

Inflation

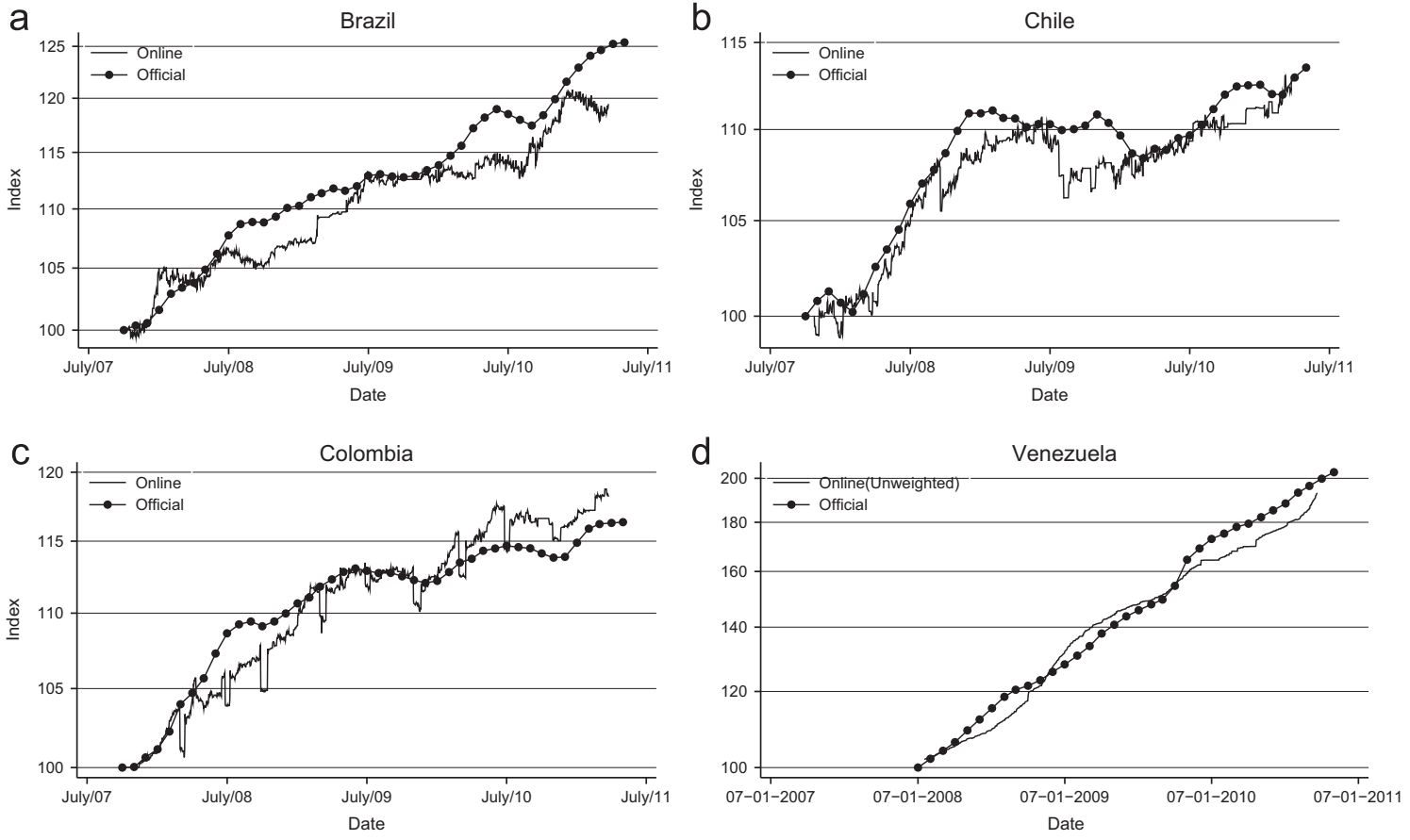
- There was a widespread claim:
  - The Argentine government has been manipulating the official inflation indexes since 2007
- Questions:
  - Is it possible to evaluate this claim?
  - If possible, how can we evaluate it?
- Possible approach:
  - Construct price indexes with online data to obtain alternative inflation estimates in countries where official estimates have lost their credibility

Billion Prices Project (BPP)

- A project started by professors Alberto Cavallo and Roberto Rigobon at MIT
- BPP covers daily price fluctuations of $\sim$5 million items sold by $\sim$300 online retailers in more than 70 countries
- BPP collects price information from hundreds of online retailers on a daily basis, using a technique called “web scraping", which uses:
  - HTML tags to locate relevant information about a product and store it in a database
  - The web address or URL of the page to classify products into standardised categories

Evaluation results – Latin american countries (1)

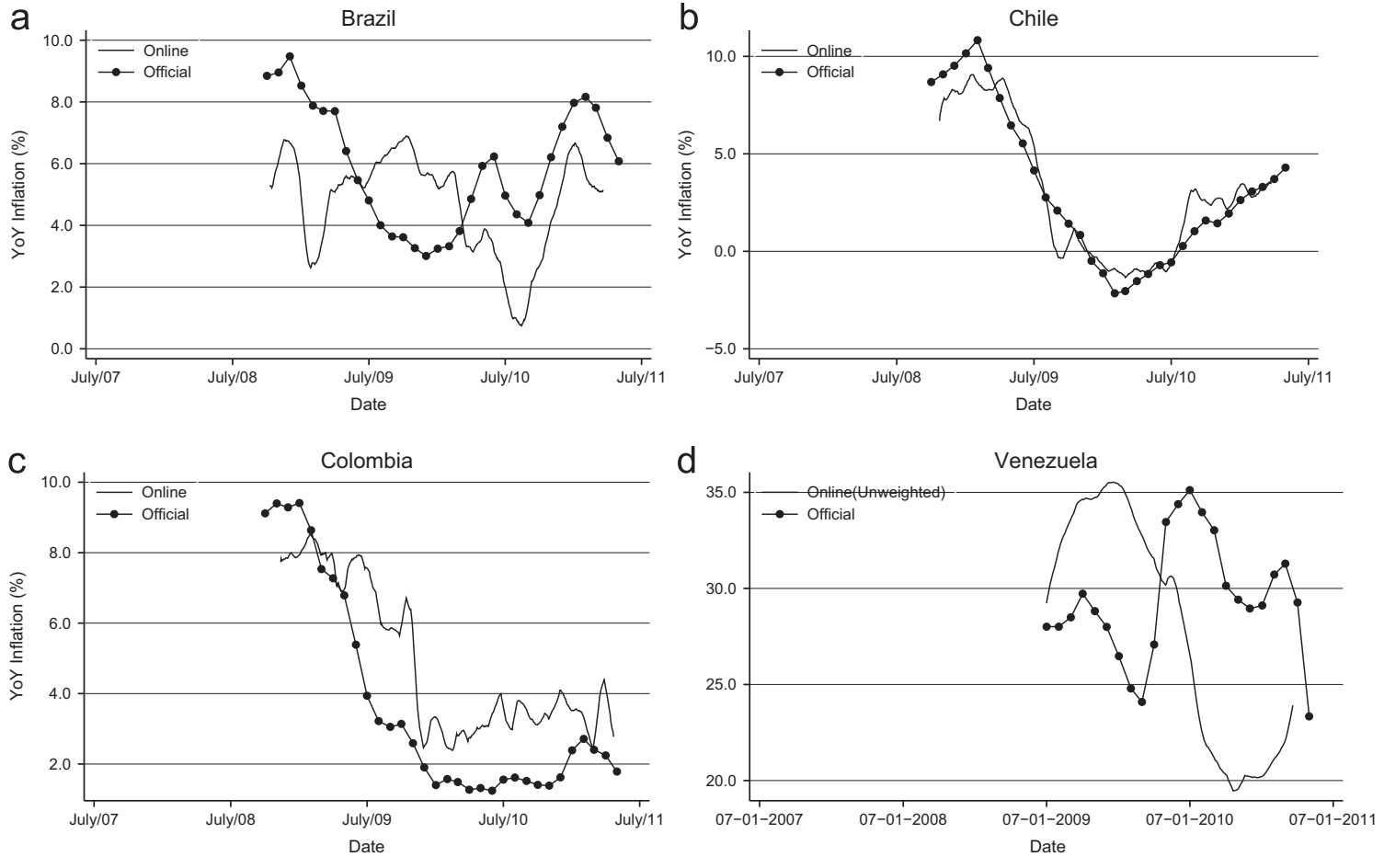
- Online and official price indexes:



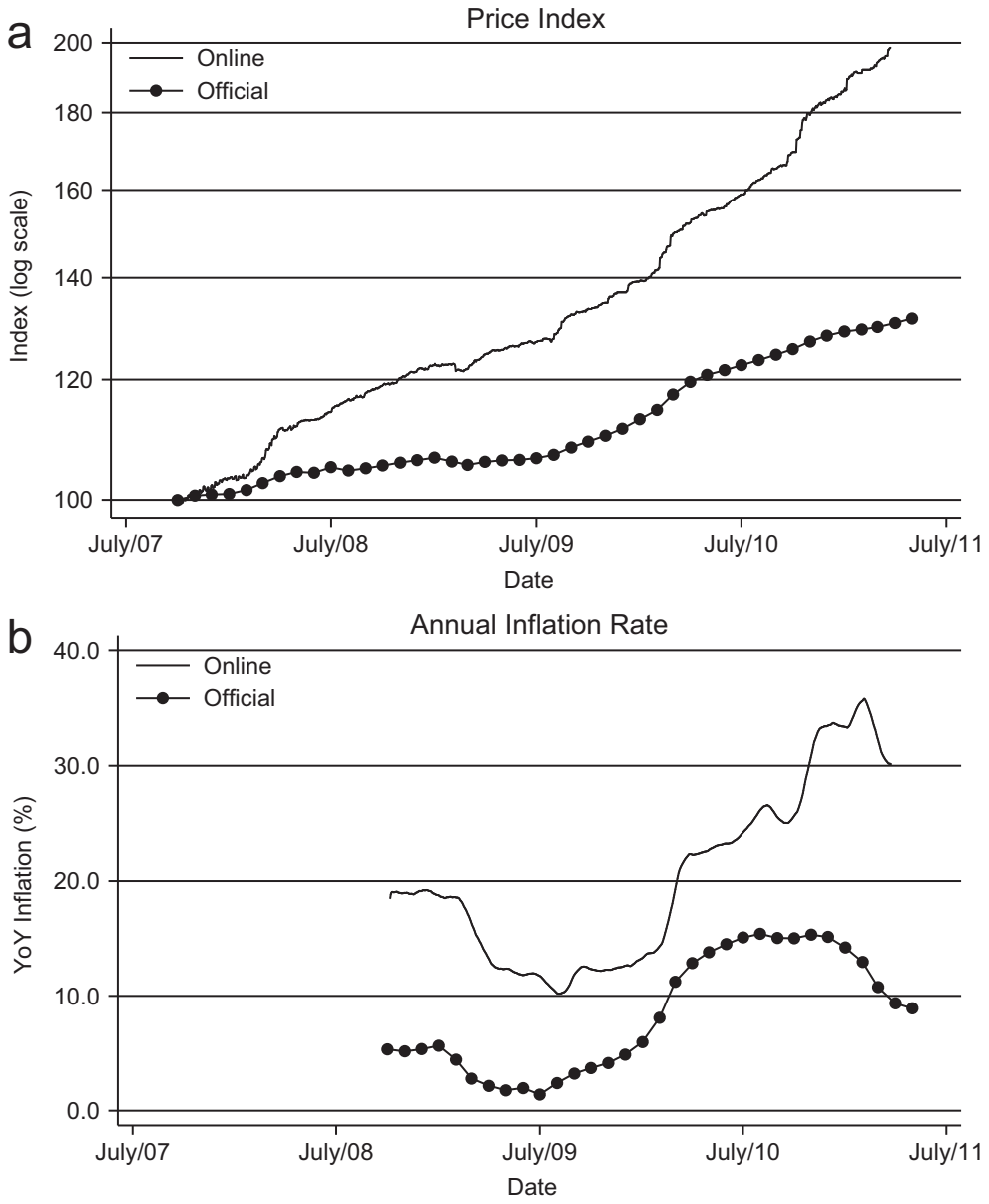
Source: [Cavallo, Journal of Monetary Economics, 2013] (also following two slides)

Evaluation results – Latin american countries (2)

- Online and official annual inflation rates:



Evaluation results – Argentina



Comparison

	Traditional Science	Data Science
Common	Use statistics to model from the data points (number of data does matter)	
Focus	Usually more about causation	STRONG correlation
Measurement	• Mostly Based on theory (deductive) decide what to measure - green only • With out seeing the other colors • Slow iterative process to discovery	• Iterate between deductive (theory based top down) and inductive (data based bottom up) reasoning to figure out what to measure : can see the other colors, so use existing data to compare • Different from fishing for results or atheoretical • Faster iteration to discovery
Measurement Error	Reduce/minimize by designing experiments well	Know what it is, adjust for it as best as possible. Usually use data that exist
Bias	Random Points, oversampling	Validation is very important: be careful not to over fit to the data, Know the bias
Main issue	Are there enough points to get the full picture?	Is the data clean enough? Is the data representative ? Sensitivity analysis

Source: [Kum, 2015, Annual Conference on Society for Social Work and Research, https://secure.sswr.org/2015RMW3.pdf]

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Knowledge base synthesis

- **Knowledge base synthesis** is the process of ingestion, disambiguation and enrichment of entities from a variety of structured and unstructured data sources
  - Sheer scale of the data
    $\Rightarrow$ Hundreds of millions of entities daily
  - Diverse domains
    $\Rightarrow$ From hundreds of data sources
  - Diverse requirements
    $\Rightarrow$ Multiple tenants, such as Locals, Movies, Deals, and Events in (for example) the Yahoo! website

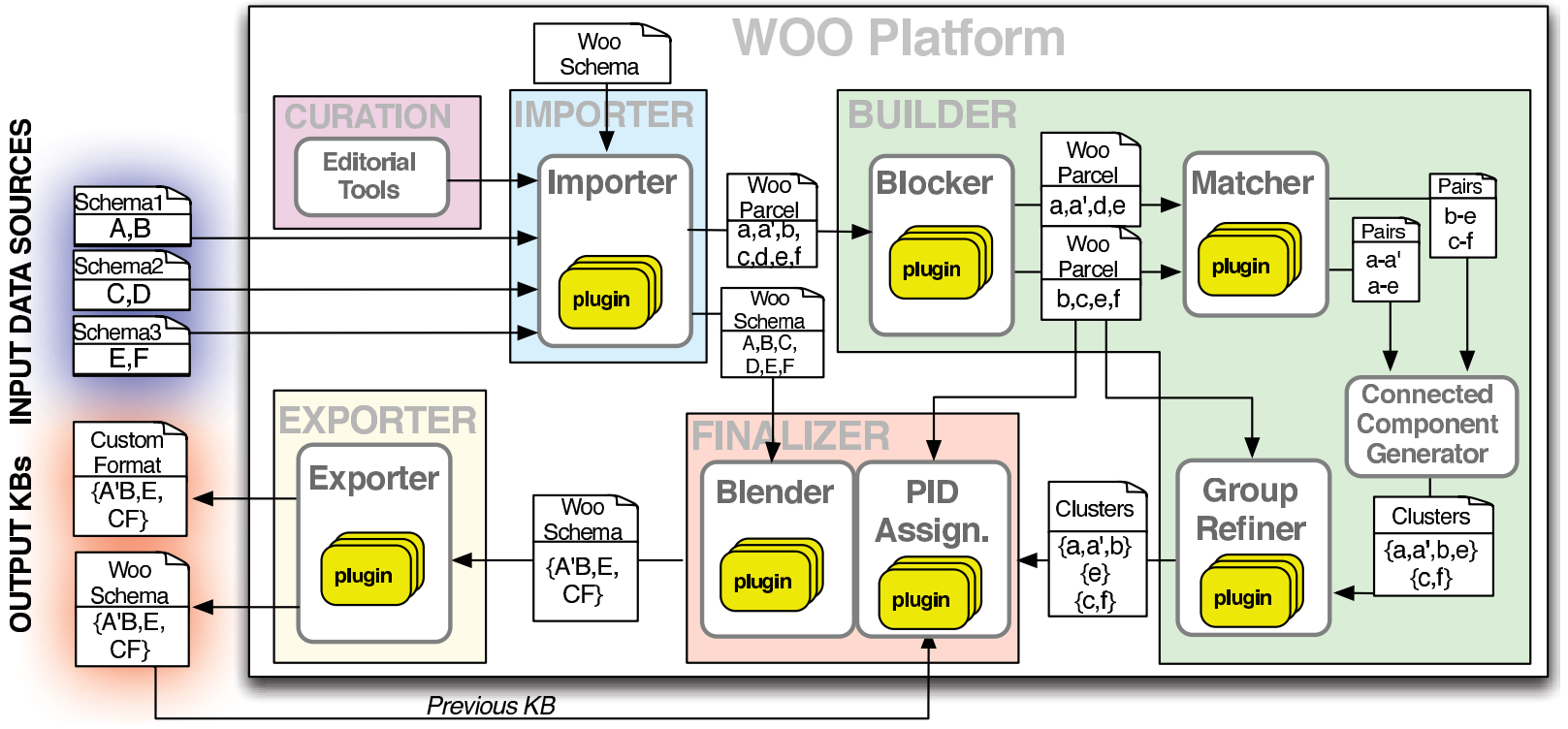
Challenges

- **Size**
  - Large data size
  - Heterogeneous input formats and schemas
  - Diverse data quality
- **Domains**
  - Multiple domains
  - Heterogeneous output formats and schemas
- **Across time and space**
  - History-aware knowledge-bases
  - Persistence
  - Incremental maintenance

Web Of Object (WOO)

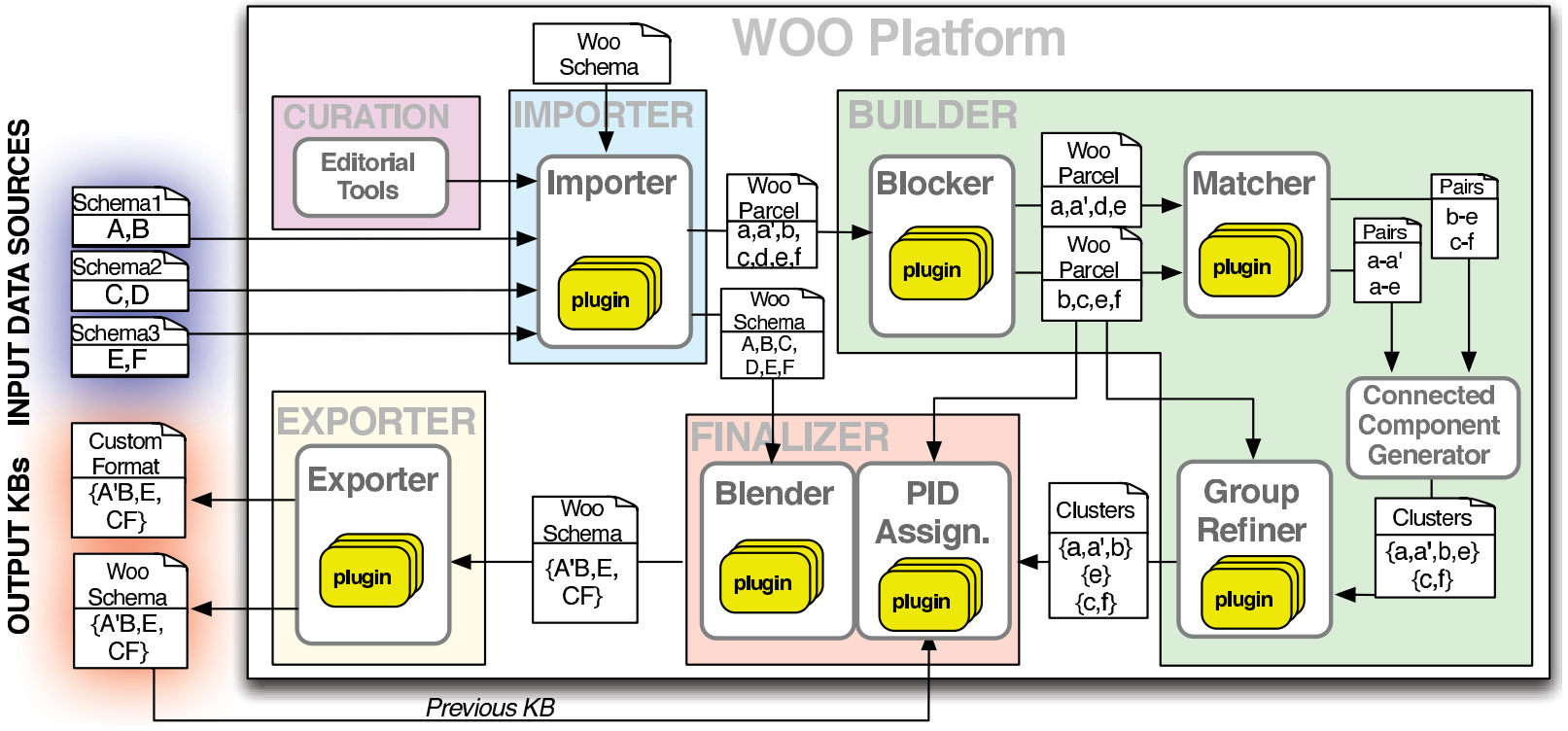
- **Goal**: To enable various products in Yahoo! to synthesise knowledge-bases of entities relevant to their domains [Bellare et al., VLDB, 2013]
- **Desiderata**:
  - *Coverage*: the fraction of real-world entities
  - *Accuracy*: information must be accurate
  - *Linkage*: the level of connectivity of entities
  - *Identifiability*: one and only one identifier for a real-world entity
  - *Persistence/content continuity*: variants of the same entity across time must be linked
  - *Multi-tenant*: be useful to multiple portals

The WOO architecture (1)



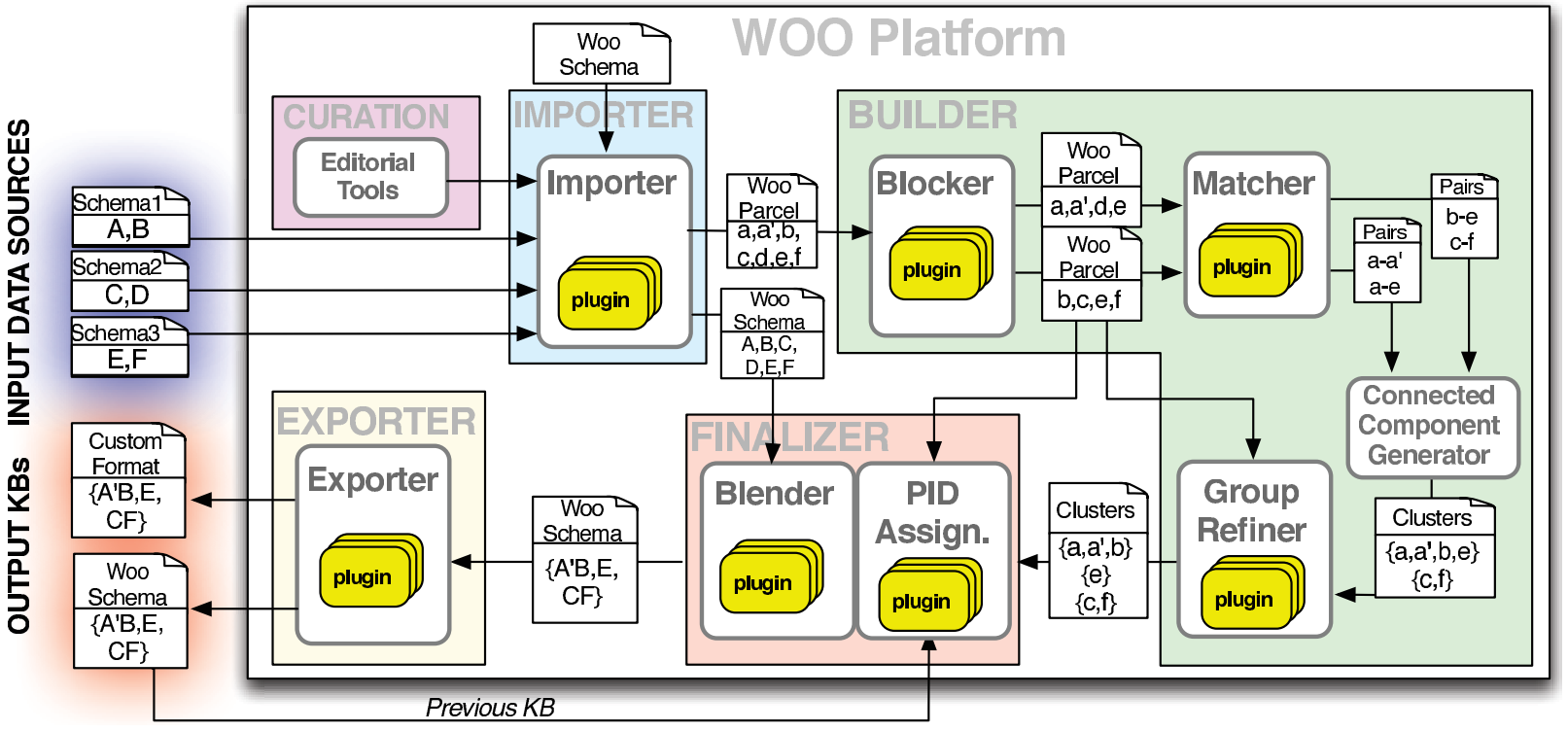
Source: [Bellare et al., VLDB, 2013]

The WOO architecture (2)



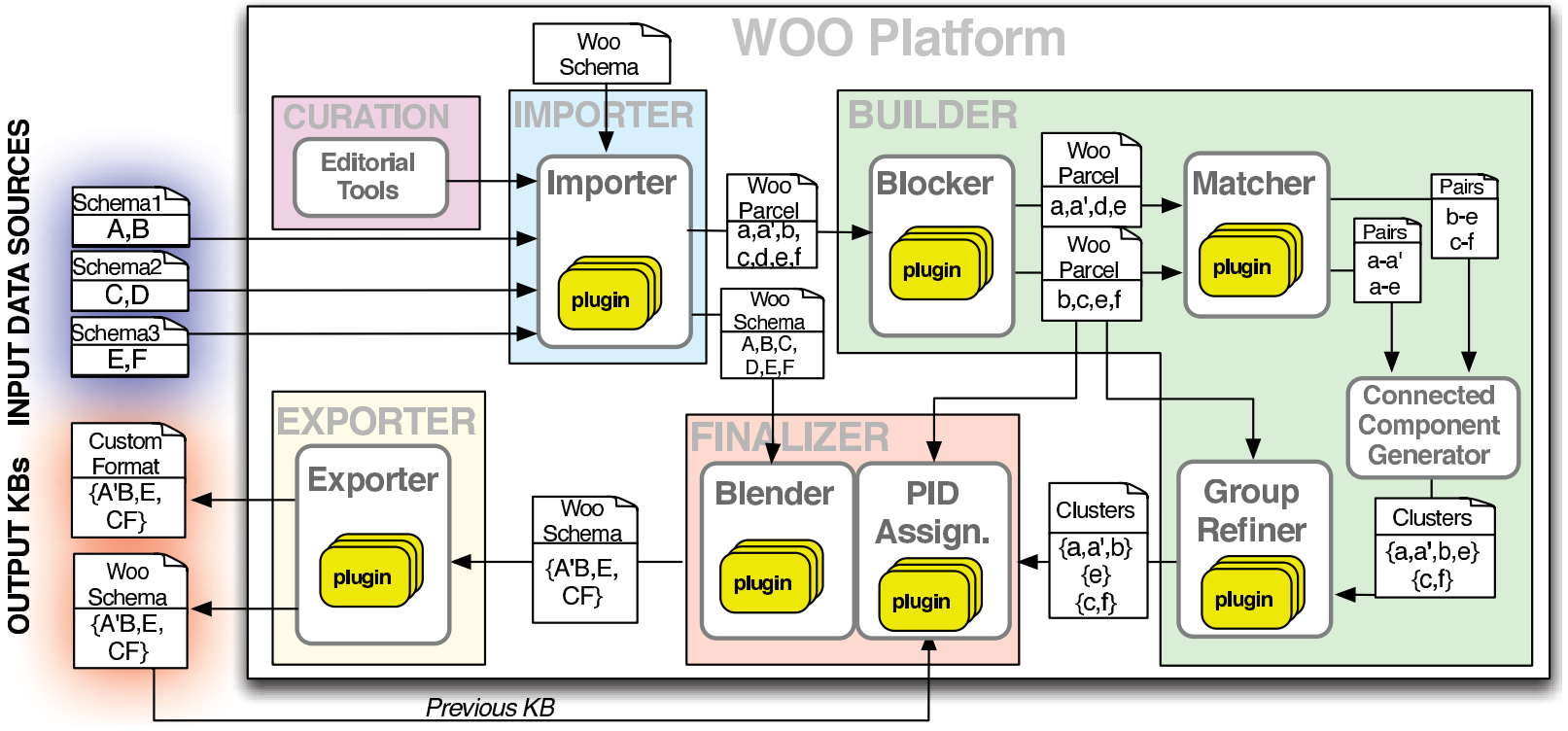
- **Importer** takes a collection of data sources as input (like XML feeds, RDF content, Relational Databases, or other custom formats)

The WOO architecture (3)



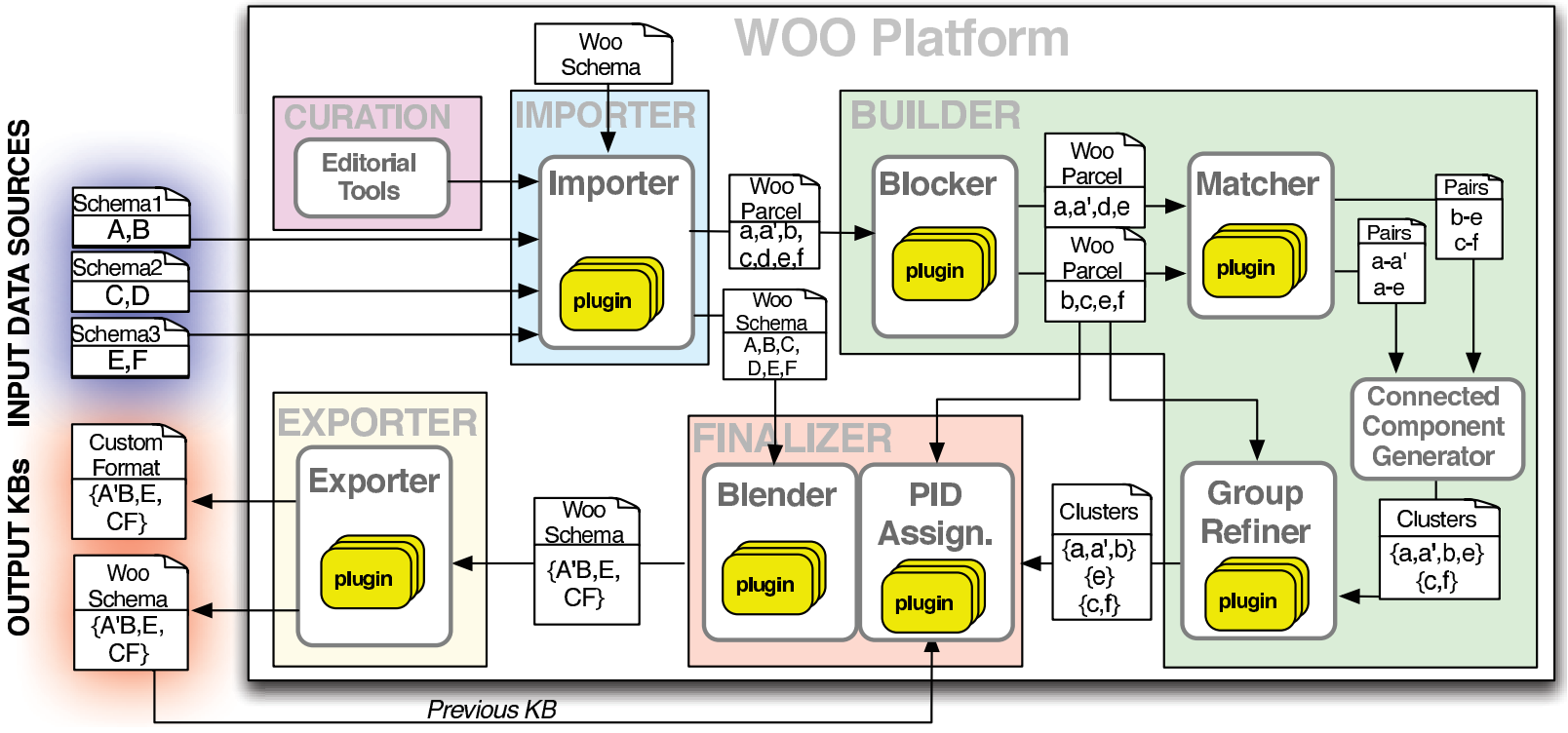
- Each data source is converted into a common format called *the WOO schema*
- The *WOO Parcel*, containing only the attributes needed for matching, is pushed to the **Builder**

The WOO architecture (4)



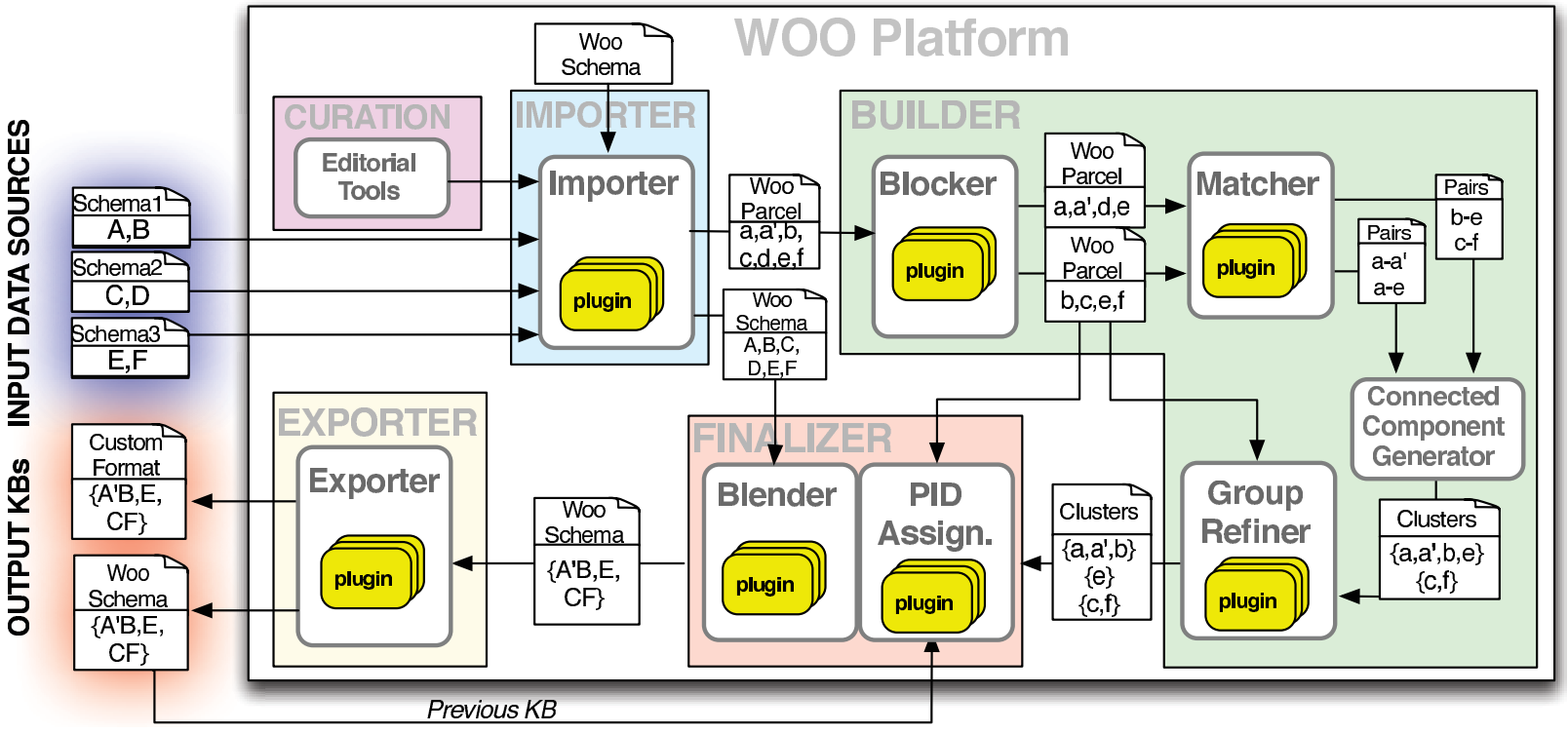
- **Builder** performs the entity deduplication and produces a clustering decision, including (1) *blocker*, (2) *matcher*, (3) *connected component generator*, and (4) *group refiner*

The WOO architecture (5)



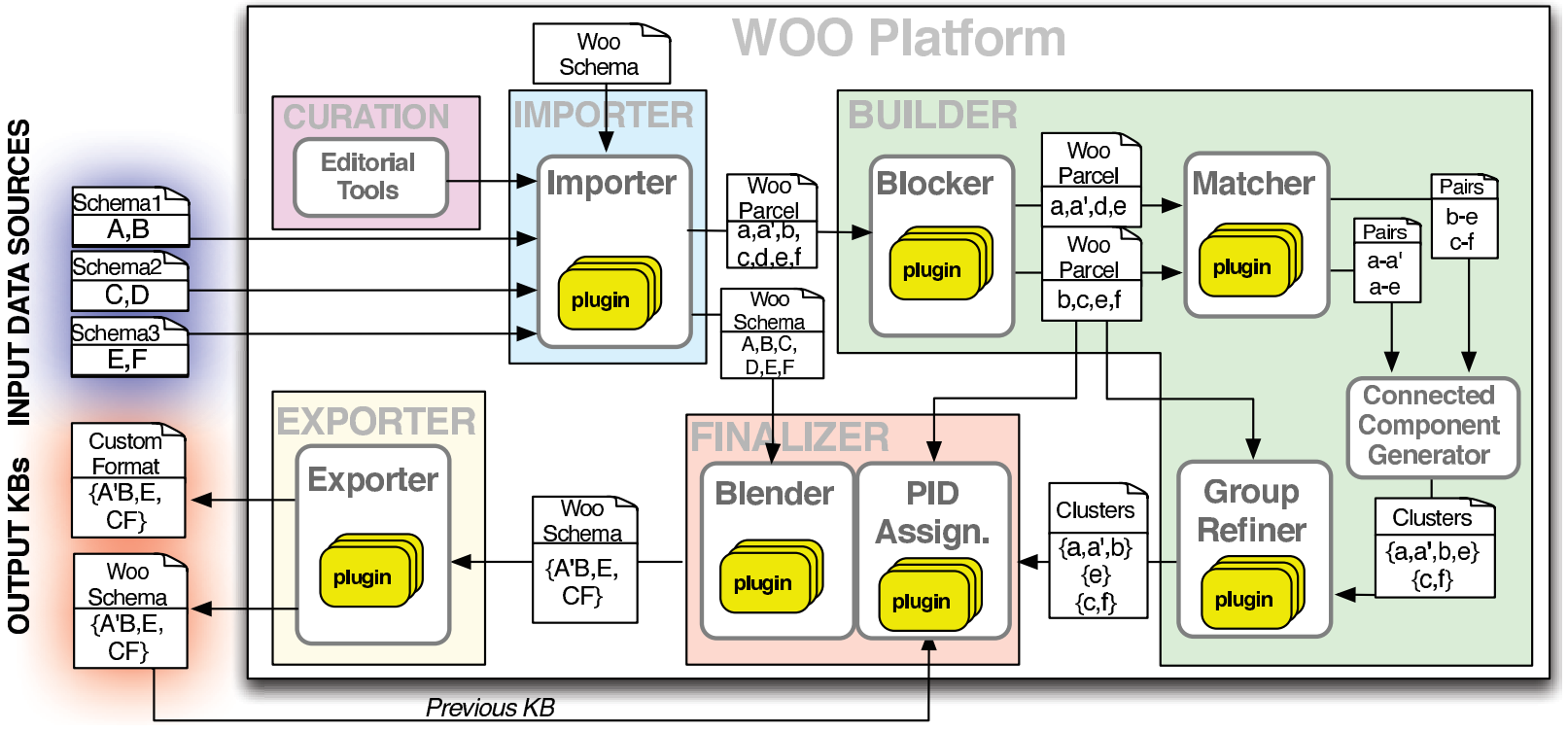
- **Finaliser** is responsible for handling the persistence of object identifiers and the blending of the attributes of the (potentially many) entities that are being merged

The WOO architecture (6)



- **Exporter** generates a fully integrated and de-duplicated knowledge-base, both in a format consistent with the WOO schema and in any custom format

The WOO architecture (7)



- **Curation** enables domain experts to influence the system behaviour through a set of GUIs, such as: forcing or disallowing certain matches between entities, or by editing attribute values

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Privacy, confidentiality, security

- Privacy: Don't Ask
- Confidentiality: Don't Tell
- Important for population informatics
  - Data governance
  - Ethics of data use
- Privacy: Not technical, but social
- Security: Tools used for privacy
  - Access control (who has access to what resource?)
  - Authentication (who are you?)
  - Encryption

Privacy is a budget constrained problem

- Differential privacy proves each query leads to some privacy loss while providing some utility in terms of data analysis (adding random noise to a database to minimise risk of identifying its records)
- The goal is to achieve the maximum utility under a fixed privacy budget

Information accountability

- Very clear transparency in the use of the data
- Disclosure: Declared in writing, so when something goes wrong the right people are held accountable (data use agreements)
- It works! Primary method used to protect financial data (credit report system)
- Internet: Crowdsourced auditing (public access *Institutional Review Board*, IRB)
- Logs and audits: What to log, how to keep tamperproof log

Privacy-by-design (1)

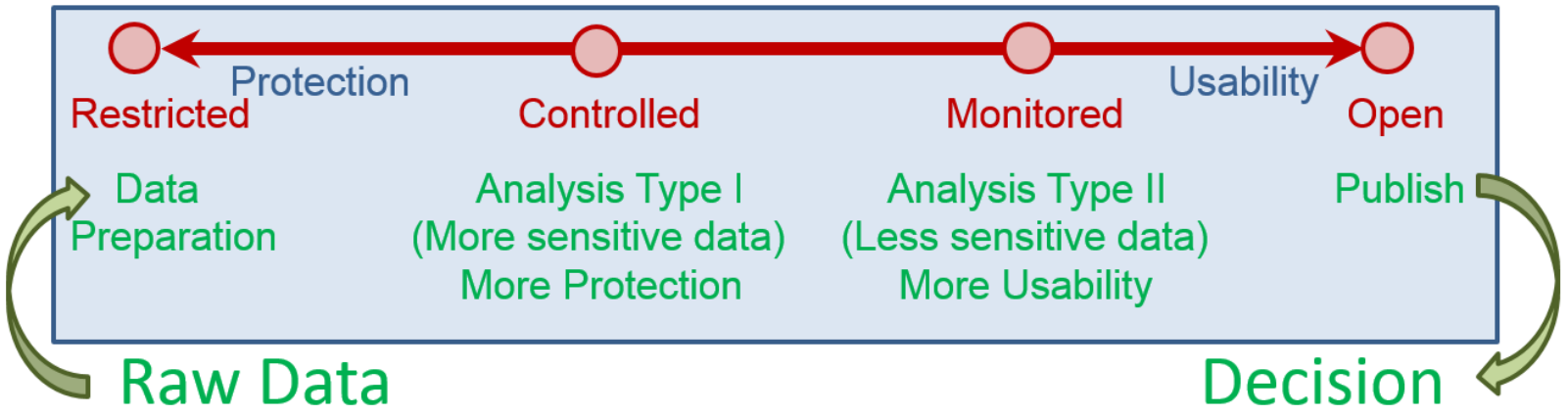
- Goes beyond the narrow view of privacy as anonymity
- Attempts to meaningfully design privacy principles and protection into the full system

  (from the beginning of the development process to deployment, use, and ultimate disposal)
- Personal data are hazardous but valuable research material

Privacy-by-design (2)

- Personal data are hazardous but valuable research material
  - Important to have proper systems in place that give protection
  - But allow for continued research in a safe manner (de-identified when possible)
  - All hazardous material need standards
    - Safe environments to handle them in, like closed computer server system lab
    - Proper handling procedures, like what software are allowed to run on the data
    - Safe containers to store them (database system)

System of access models



- Goal: To design an information system that can enforce the varied continuum from one end to the other such that one can balance privacy and usability as needed to turn data into decisions for a given task

Privacy protection mechanism

Access	Restricted Access	Controlled Access	Monitored Access	Open Access
Protection Approach	Physical restriction to access	Lock down VM (limit what you can do on the system)	Information accountability	Disclosure Limitation
Monitoring Use	All use on & OFF the computer is monitored	All use on the computer is monitored		Trust
IRB	Full IRB approved	Full IRB approved	IRB Exempt (register)	Terms of Use
R1:Crypto-graphic Attack	Very Low Risk	Low Risk. Would have to break into VM	High Risk	NA
R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	

Comparison of risk and usability

		Restricted Access	Controlled Access	Monitored Access	Open Access
Usability	U1.1: Software (SW)	Only preinstalled data integration & tabulation SW. No query capacity	Requested and approved statistical software only	Any software	Any software
	U1.2: Data	No outside data allowed But PII data	Only preapproved outside data allowed	Any data	Any data
	U2: Access	No Remote Access	Remote Access	Remote Access	Remote Access
Risk	R1:Crypto-graphic Attack	Very Low Risk	Low Risk. Would have to break into VM.	High Risk	NA
	R2: Data Leakage	Very Low Risk. Memorize data and take out	Physical data leakage (Take a picture of monitor)	Electronically take data off the system.	NA

Disclosure limitations

- The broad array of methods used to protect confidentiality of statistical data
- Filter the raw data to block what is revealed
- Disclosure-limiting masking: Transformations of the data whereby there is a specific functional relationship (possibly stochastic) between the masked values and the original data
  - Summarisation / generalisation
  - Suppression
  - Swapping
  - Add noise
  - Simulated data

Privacy as contextual integrity

- Helen Nissenbaum (NYU Law School)
  [Washington Law Review, 2004]
- A conceptual framework for understanding privacy expectations and their implications developed in the literature on law, public policy, and political philosophy
- Privacy protection / Violation
  - Social norms of expectation (on use, sharing, etc.)
  - Due diligence
  - Quantifying harm: loss of job

Tea/coffee break

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Linking and integrating databases

- Linking and integrating records that represent the same entity in one or more databases improves data quality and enriches data for further analysis
- Known as *record linkage*, *data matching*, *entity resolution*, *object identification*, *merge-purge*, *duplicate detection*, and various other names
- Linked data empower efficient and quality data analysis and mining
- Record linkage is required in many applications:

  (health-care, government services, crime and fraud detection, national security, business applications, and more recently *population informatics*)

Applications of record linkage

- Health data mining and analytics
  (epidemiological or adverse drug reaction studies)
- National security and crime investigation
  (effective identification of fraud, crime, or terrorism suspects)
- Population informatics
  (for example, the *Beyond 2011* program by the Office of National Statistics (ONS) in the UK aimed at producing population and socio-demographics statistics for the UK by using record linkage)
- Business mailing lists (de-duplication of customer databases for effective marketing)
- Geocode matching (with reference address data)

Challenges of record linkage

- **Scalability**: Every record from a database potentially needs to be compared with all records from other databases
- **Linkage quality**: Unique entity identifiers are not available in the databases to be linked

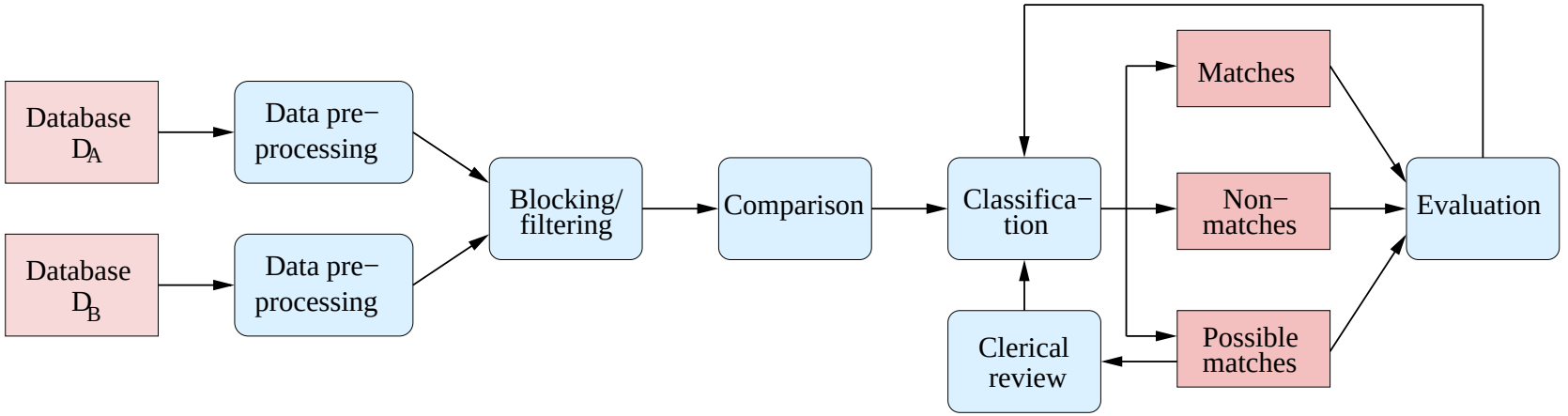
  Approximate matching of personal identifiers (such as names and addresses) is required

  For example, which records represent the same person?

Dr Smith, Peter	*42 Miller Street 2602 O'Connor*
Pete Smith	*42 Miller St 2600 Canberra A.C.T.*
P. Smithers	*24 Mill Rd 2600 Canberra ACT*

- **Privacy**: A big concern when using sensitive personal information across organisations

The record linkage process



- Databases are pre-processed (cleaned and standardised)
- Scalability is addressed by blocking/filtering
- Candidate pairs are compared and classified (into matches, non-matches, and possible matches)
- Clerical review is conducted on possible matches
- Results are evaluated in terms of complexity and quality

Record linkage techniques (1)
Blocking / filtering

- **Blocking** groups records according to a criteria (blocking key) such that records from the same group need to be compared
- **Filtering** prunes non-matches based on similarity-dependent characteristics (such as lengths, prefixes, etc.)
- Results in candidate record pairs/sets to be compared
- Various techniques have been developed

  (including standard (phonetic) blocking, sorted neighbourhood, and canopy clustering; and filtering techniques such as PP-Join)

Record linkage techniques (2)
Comparison

- **Exact matching** of record pairs, if a *unique identifier* of high quality is available: precise, robust, stable over time

  (for example *Social security* or *Medicare* numbers)

  In the absence of a unique identifier, exact matching of identifying attributes, such as names, does not provide accurate matching due to data errors and variations

- **Approximate matching** employs comparison functions that provide a numerical similarity for a compared record pair (between 0 and 1)

- Various comparison functions have been developed (such as edit distance, Jaro-Winkler, Jaccard and Dice coefficients)

Record linkage techniques (3) Classification

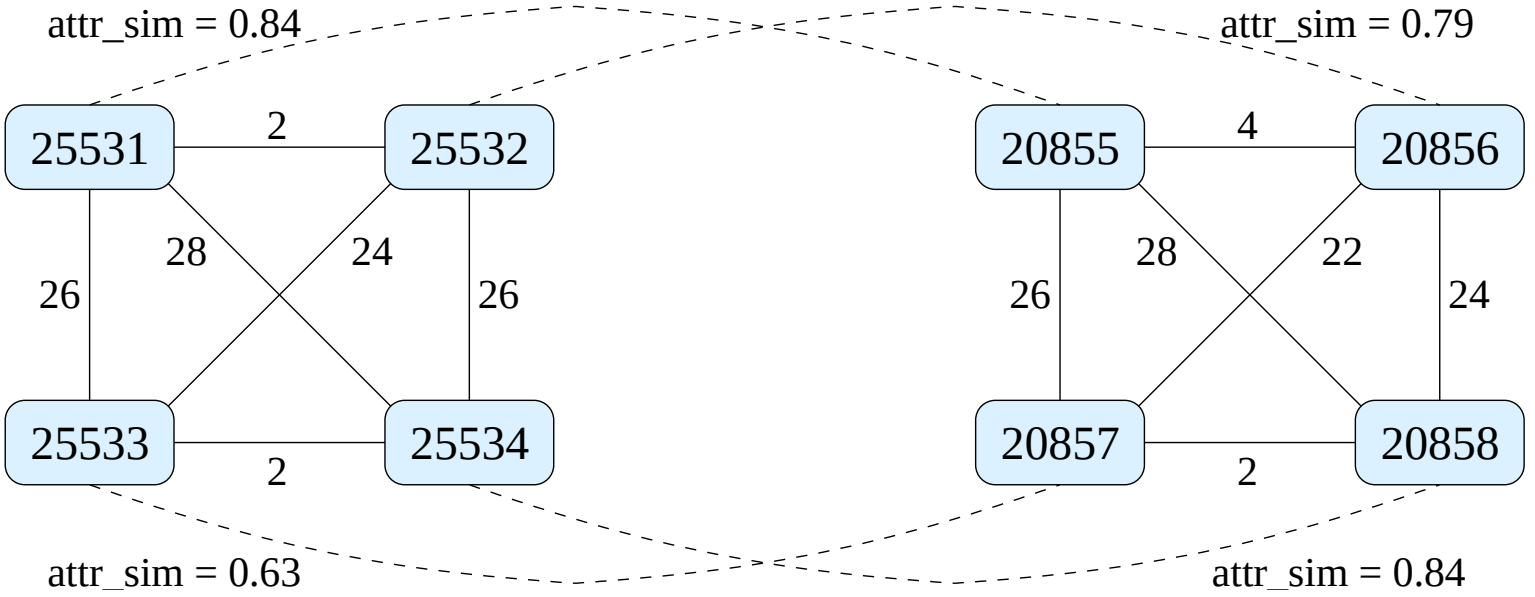
- **Rule-based** or **threshold-based** uses a set of rules or similarity thresholds to classify the compared record pairs (into matches, non- matches, and possible matches)
- **Probabilistic record linkage** [Fellegi and Sunter, 69]
  - Uses personal identifying attributes for linking (such as names, addresses and dates of birth)
  - Calculates match weights for attributes
- **Machine learning**
  - Supervised learning requires training data (record pairs with known true match status)
  - Unsupervised learning (clustering)
  - Active learning or semi-supervised

Advanced classification: Active learning and group linkage

- **Active learning**
  - Semi-supervised by human-machine interaction
  - Overcomes the problem of supervised learning that requires training data
  - Selects a sample of record pairs to be manually classified (budget constraints)
  - Trains and improves the classification model using manually labelled data
- **Group linkage**
  - Conducts pair-wise linking of individual records
  - Calculates group similarities using Jaccard or weighted similarities (based on pair-wise similarities)

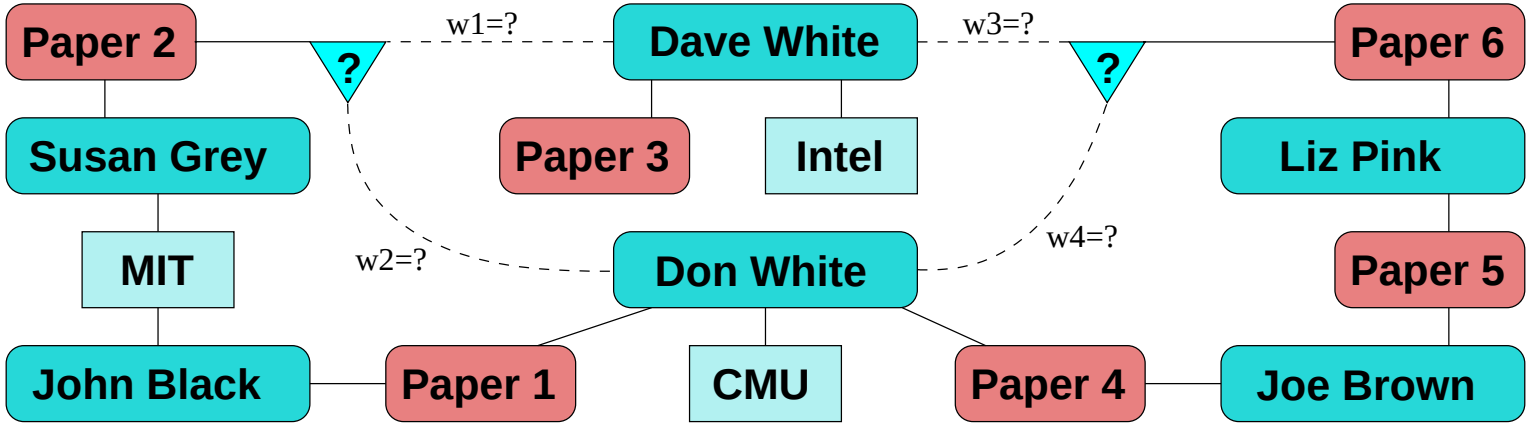
Advanced classification: Graph-based linkage

- Based on structure between groups of records
  (for example linking households from different censuses)
  - One graph per household, finds best matching graphs using both record attribute and structural similarities
  - Edge attributes are information that does not change over time (like age differences)



Advanced classification: Collective entity resolution

- Considers *relational similarities* not just attribute similarities



(A1, Dave White, Intel)	(P1, John Black / Don White)
(A2, Don White, CMU)	(P2, Sue Grey / **D. White**)
(A3, Susan Grey, MIT)	(P3, Dave White)
(A4, John Black, MIT)	(P4, Don White / Joe Brown)
(A5, Joe Brown, unknown)	(P5, Joe Brown / Liz Pink)
(A6, Liz Pink, unknown)	(P6, Liz Pink / **D. White**)

Adapted from: [Kalashnikov and Mehrotra, ACM TODS, 2006]

Clerical review and evaluation

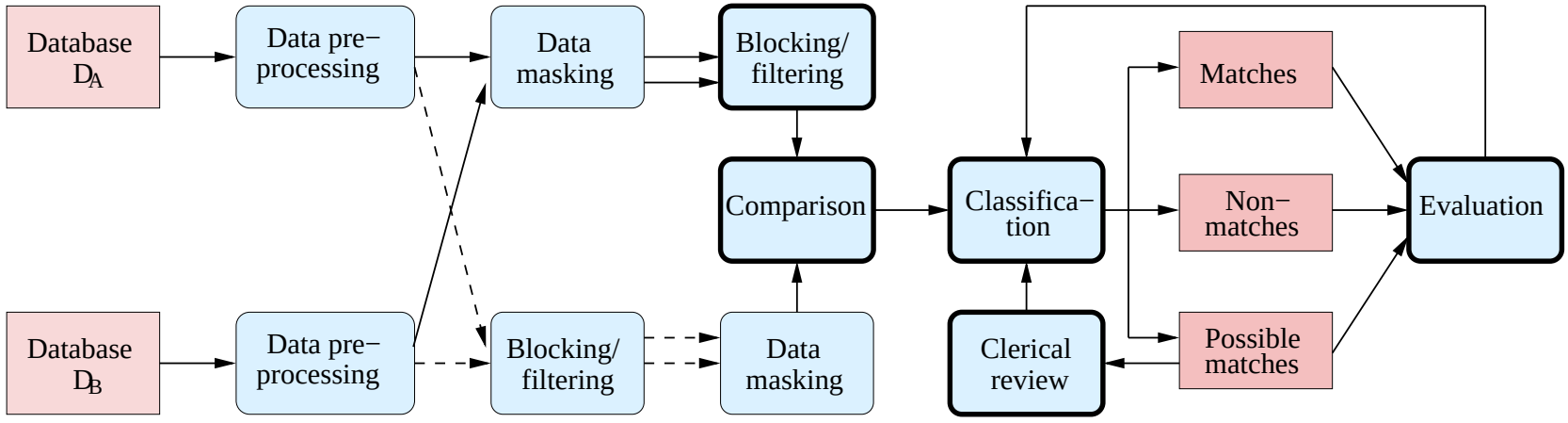
- **Clerical review**
  - Record pairs classified as possible matches need to be manually assessed and classified into matches or non-matches
  - A time-consuming (*budget*) and error-prone process (*accuracy*)
  - Active learning can be used for clerical review
- **Evaluation**
  - Complexity (or scalability) using measures such as run-time, memory consumption, number of comparisons (reduction ratio)
  - Linkage quality using measures such as pairs completeness, precision, recall, and F-score

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Privacy-preserving record linkage (PPRL)

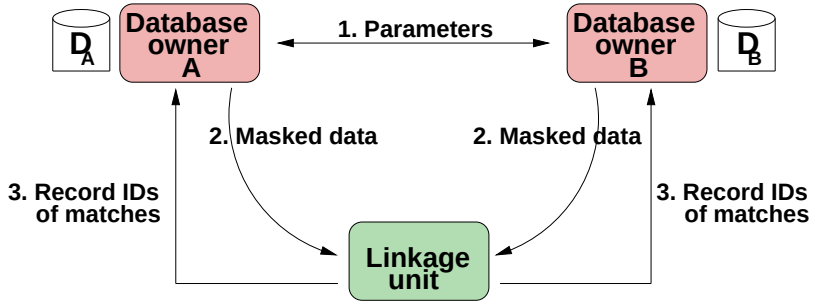
- Objective of PPRL is to perform linkage across organisations using *masked (encoded)* records such that besides certain attributes of the matched records no information about the sensitive source data can be learned by any party involved in the linking, or any external party

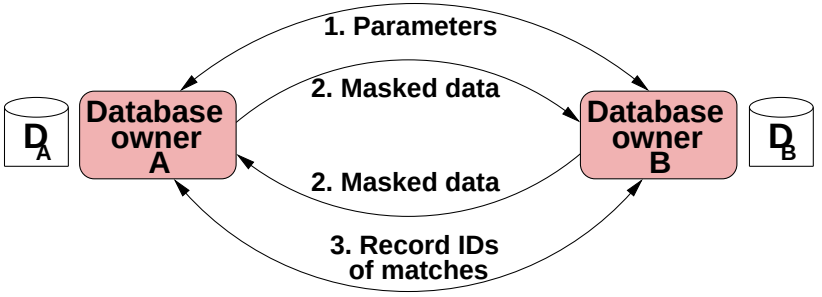


PPRL: An example scenario

- A demographer who aims to investigate how mortgage stress is affecting different people with regard to their mental and physical health
- She will need data from financial institutions, government agencies (social security, health, and education), and private sector providers (such as health insurers)
- It is unlikely she will get access to all these databases (for commercial or legal reasons)
- She only requires access to some attributes of the records that are linked, but not the actual identities of the linked individuals (but personal details are needed to conduct the actual linkage)

PPRL protocols





- Three-party protocols

  Use a linkage unit to conduct or facilitate linkage

- Two-party protocols

  Only the two database owners participate in the linkage

- Multi-party protocols

  Linking records from multiple databases (with or without a linkage unit)

Adversary models

- **Honest-but-curious** (HBC) model assumes that parties follow the protocol while being curious to find about another party's data
  - HBC model does not prevent collusion
  - Most existing PPRL protocols assume HBC model
- **Malicious** model assumes that parties behave arbitrarily (do not follow the protocol)
  - Evaluating privacy under malicious model is difficult
- **Accountable computing and covert model**
  - Allow for proofs if a party has followed the protocol or the misbehaviour can be detected with high probability
  - Lower complexity than malicious and more secure than HBC

Attack methods

- **Dictionary** attacks

  An adversary masks a list of known values using existing masking functions until a matching masked value is identified (a keyed masking approach, like HMAC, can help prevent this attack)

- **Frequency** attacks

  Frequency distribution of masked values is matched with the distribution of known values

- **Cryptanalysis** attack

  A special category of frequency attack applicable to Bloom filter based masking

- **Collusion**

  A set of parties (in multi-party or three-party protocols) collude with the aim to learn about another party's data

PPRL techniques

- First generation (mid 1990s): exact matching only using simple hash encoding
- Second generation (early 2000s): approximate matching but not scalable (PP versions of edit distance and other string comparison functions)
- Third generation (mid 2000s): take scalability into account (often a compromise between PP and scalability, some information leakage accepted)
- Different approaches have been developed for PPRL, so far no clear best technique

  For example based on Bloom filters, embedding space, generalisation, noise addition, or secure multi-party computation (SMC)

PPRL techniques: Secure hash-encoding

- Use a one-way hash function (like SHA) to encode values and then compare hash-codes
- Having only access to hash-codes will make it nearly impossible to learn their original input values
- But dictionary and frequency attacks are possible
- Single character difference in input values results in completely different hash codes

  For example:

  'peter' $\rightarrow$ '101010...100101'

  'pete' $\rightarrow$ '011101...011010'
- Only exact matching is possible

PPRL techniques: Reference values and embedding

- Reference values
  - Values extracted from a publicly available source in the same domain (e.g. telephone directory) or randomly generated values
  - Calculate similarities between private values using the similarities of each private value with the reference value (triangular inequality)
- Embedding space:
  - Embeds records into multi-dimensional space while preserving the distances
  - Difficult to determine the dimension of space and select suitable pivots

PPRL techniques: Noise and differential privacy

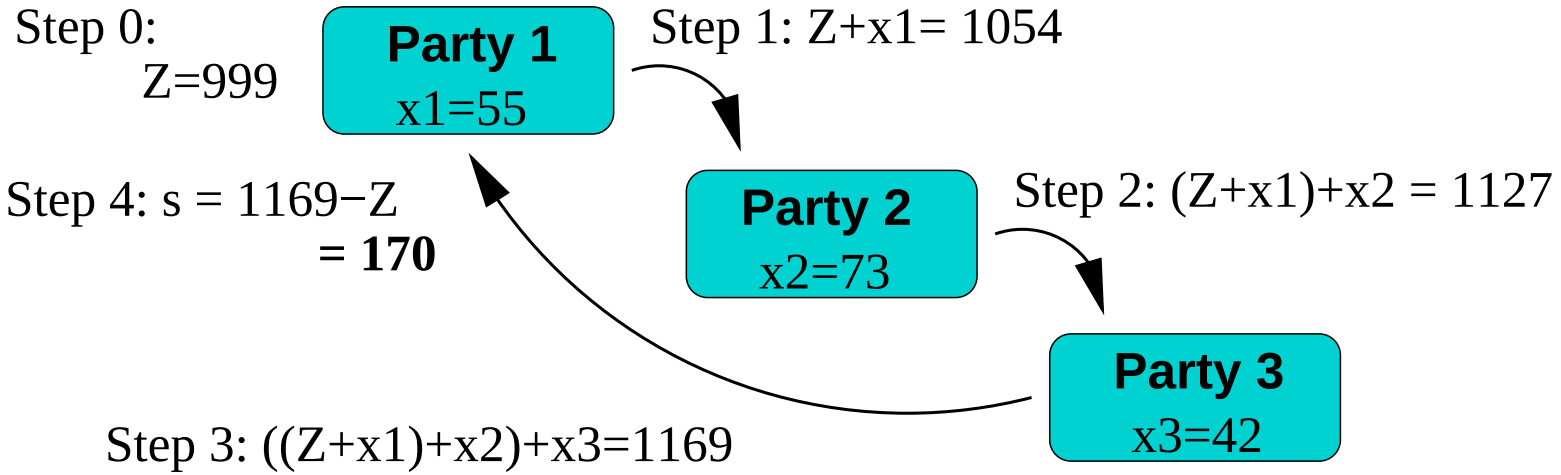
- Noise addition:
  - Extra (fake) records to perturb data
  - Overcomes frequency attack (improves privacy) at the cost of more comparisons and loss in linkage quality (due to false matches)
- Differential privacy:
  - Alternative to noise addition
  - The probability of holding any property on the perturbed database is approximately the same whether or not an individual value is present in the database
  - Magnitude of noise depends on privacy parameter and sensitivity of data

PPRL techniques: Encryption and generalisation

- Generalisation:
  - Generalises the records to overcome frequency attacks
  - For example k-anonymity: ensure every combination of attribute values is shared by at least k records
  - Other techniques are value generalisation hierarchies, top-down specialisation, and binning
- Encryption schemes (SMC):
  - Commutative and homomorphic encryption are used
  - Secure scalar product, secure set intersection, and secure set union are the most commonly used SMC techniques
  - However, many are computationally expensive

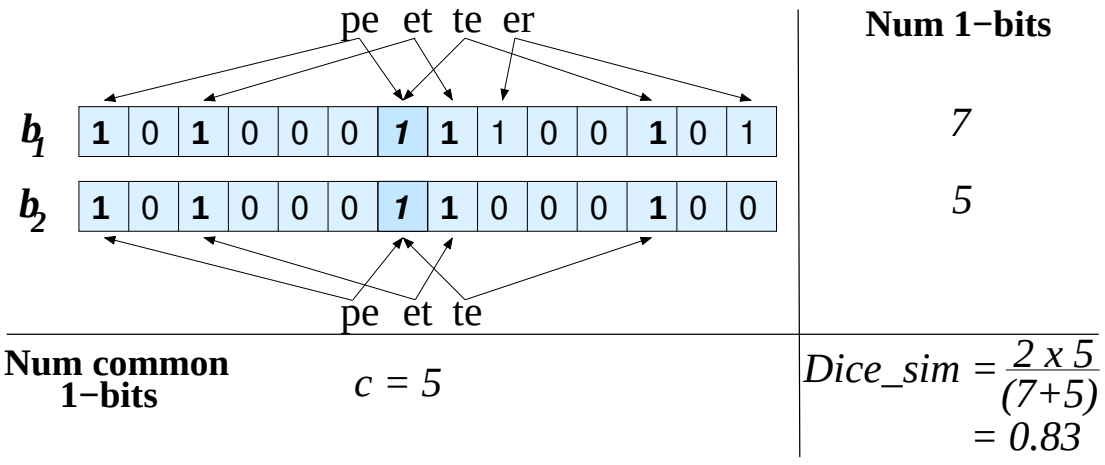
PPRL techniques: Secure multi-party computation

- Compute a function across several parties, such that no party learns the information from the other parties, but all receive the final results
  [Yao, Foundations of Computer Science, 1982]
- Simple example: Secure summation $s = \sum_i x_i$.



PPRL techniques: Bloom filters

- Bloom filters are bit vectors initially set to 0-bits
- Use k hash functions to hash-map a set of elements by setting corresponding k bit positions to 1
- A set of q-grams (string) or neighbour values (numerical) are hash-mapped to allow approximate matching
- Dice similarity of two Bloom filters b_1 and b_2 is $Dice\_sim(b_1, b_2) = \frac{2 \times c}{(x_1 + x_2)}$, where $c = |b_1 \cap b_2|$, $x_i = |b_i|$

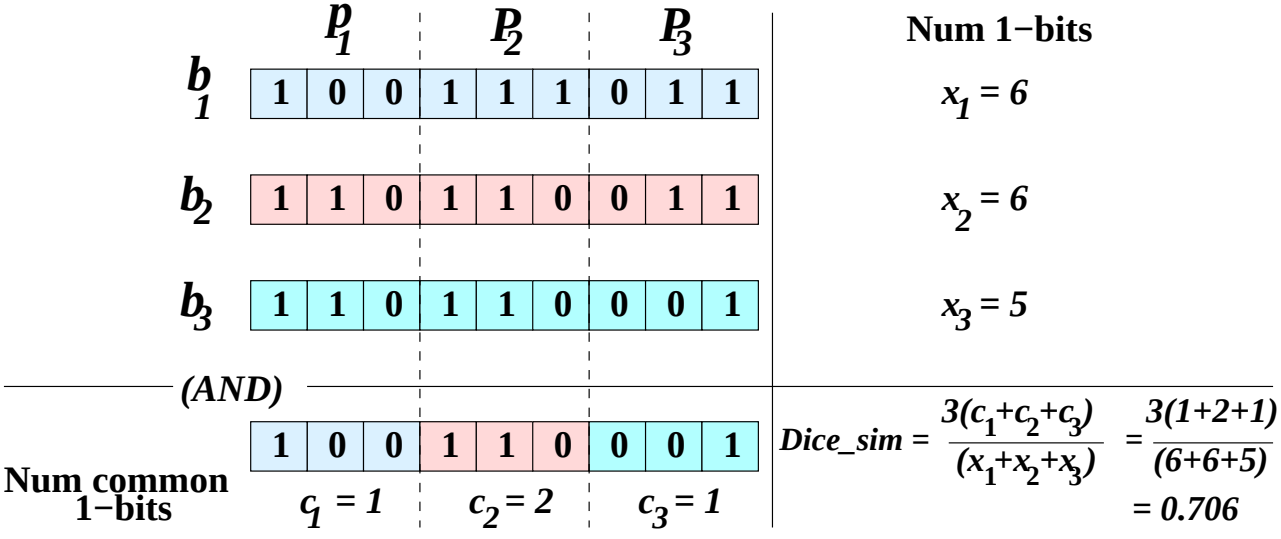


Multi-Party PPRL (1)

- Privacy-preserving linking of multiple databases (more than two sources)
- Example applications:
  - Health outbreak systems require data to be integrated across human health data, travel data, drug data, and animal health data
  - National security applications need to integrate data from law enforcement agencies, Internet service providers, businesses, and financial institutions
- Additional challenges:
  - Exponential complexity with number of sources
  - Increased privacy risk of collusion

Multi-Party PPRL (2)

- Distributed similarity calculation:
  [Vatsalan and Christen, CIKM, 2014]
  - Bloom filters are split into segments such that each party processes a segment to calculate the number of common 1-bits in its segment
  - Secure summation is applied to sum the number of common 1-bits (c_i) and total 1-bits (x_i) in their Bloom filter to calculate the similarity



Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Interactive record linkage

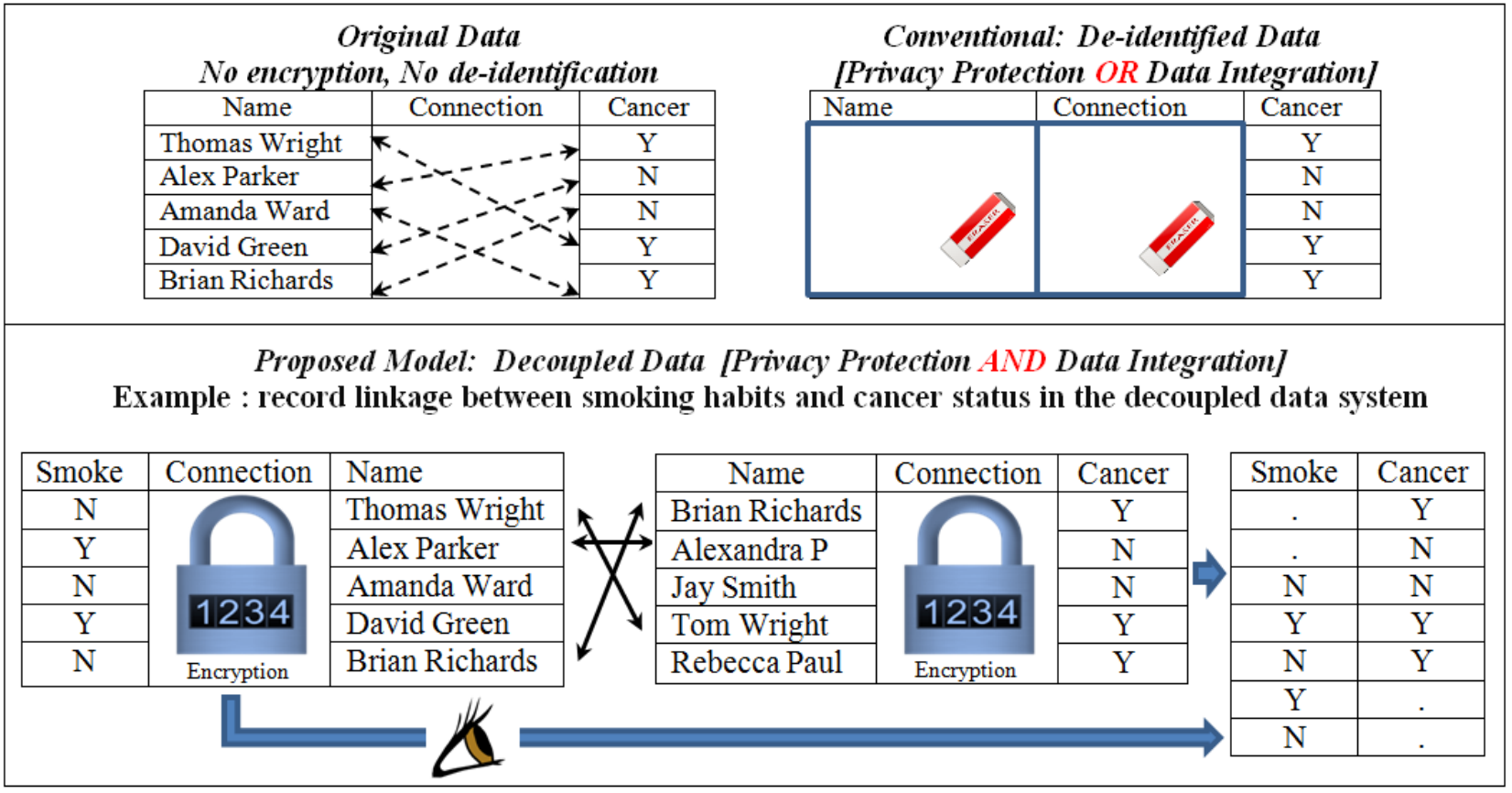
- Record linkage often involves uncertain linkages that must be resolved manually
- Visualisation tools to effectively support interactive record linkage will also guide standardising and cleaning the data
- Interactive record linkage: People tuning and managing errors in the data and false matches from approximate record linkage algorithms
- We define the properly tuned output from a hybrid human-machine data integration system as high quality record linkage

Privacy-preserving interactive record linkage framework

- Privacy: Guarantee against attribute disclosure and minimise identity disclosure
- Interactive:
  - High quality record linkage requires a human computer linkage system where people make refinements to the uncertain linkages resulting from automatic algorithms
  - There is a need to pass on the confidence level of the linkages downstream to the analysis phase for accurate analysis of the linked data

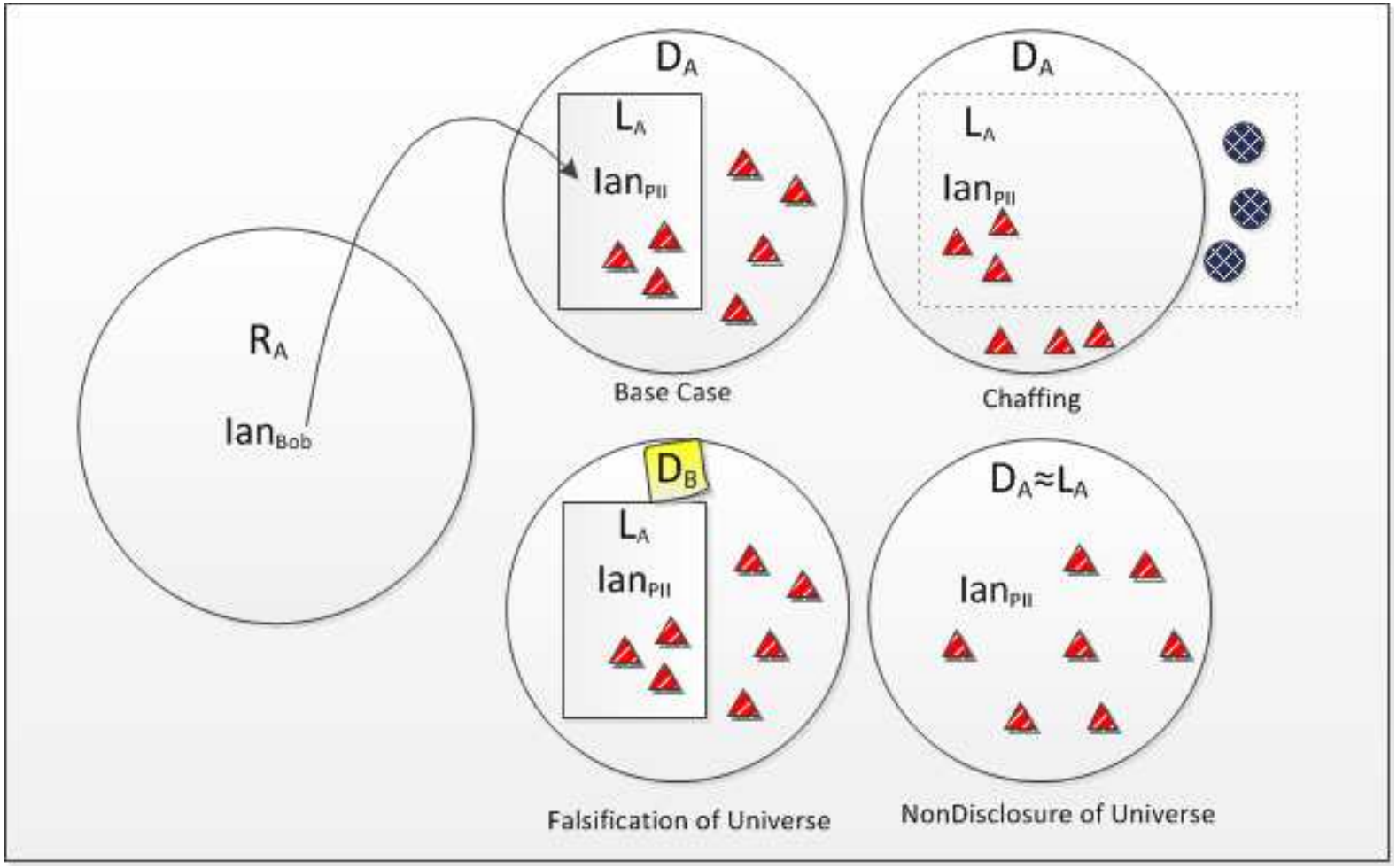
Privacy-preserving interactive record linkage approaches (1)

- Decoupling of tables



Privacy-preserving interactive record linkage approaches (2)

- Chaffing (add noise at record level)



Privacy-preserving interactive record linkage approaches (3)

- Incremental disclosure using a privacy budget

A2b. *Proposed Method*: First screen of On-demand Drill-Down Incremental GUI

D=Different, Diff=More than 1 digit or letter difference
M=Missing (either or both records), TX=Transpose, _=Same,
Type Y/N in the right most column to establish a linkage.
Click on a cell to drill down for more information.

Total Priv Budget: 100%

No.	Fname	Lname	DOB	SSN	Gender	Link?	Info. discl.
11	-------	--------	T/X/----	----------	-	Yes	0%
12	Diff	--D-	--/D-/--D-	----D---	M	No	0%
13	Diff	D-----	--/--/----	------TX--	-	??	0%
14	Diff	Diff	--/TX/----	M	D	??	0%

A2c. *Proposed Method*: Sequence of displays showing on-demand incremental reveal of clear text data.

No.	Fname	Lname	DOB	SSN	Gender	Link?	
13.1	Diff	D-----	--/--/----	------TX--	-	??	Info. discl.: 0% Total Priv Budget: 100%
13.2	Diff	B----- P-----	--/--/----	------TX--	-	??	Info. discl.: 5% Total Priv Budget: 99%
13.3	Diff	Balmer Palmer	--/--/----	------TX--	-	??	Info. discl.: 20% Total Priv Budget: 90%
13.4	rare: L=4, 7 edist=3	Balmer Palmer	--/--/----	------TX--	-	Yes	Info. discl.: 25% Total Priv Budget: 89%
13.5	Tamika Tam	Balmer Palmer	--/--/----	------TX--	-	Yes	Info. discl.: 40% Total Priv Budget: 80%
13.6	Tamika Tam	Balmer Palmer	--/--/----	-------48-- -------84--	-	??	Info. discl.: 45% Total Priv Budget: 78%
13.7	Tamika Tam	Balmer Palmer	--/--/----	-------48-- -------84--			NOT ALLOWED (No change in display)

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Data mining and analysis with linked population data

- Linked data by data integration improve data quality and allows new and valuable knowledge discovery and data mining not possible on individual databases
- Data integration consists of three major aspects [Christen et al., ACM JDIQ, 2014]
  - **Schema matching**: which attributes in two database schemas contain the same type of information
  - **Data matching**: which records in two databases refer to the same entity
  - **Data fusion**: merging matched records into consistent and coherent forms

Uncertainty, bias and error in linkage

- No shared error-free identifying attributes
  - Deterministic record linkage assumes error-free identifying attributes and links records that have exactly matching values in such attributes
  - In multiple data sources, such shared error-free attributes are uncommon
  - When no error-free identifier is available, probabilistic record linkage is used
- Two types of possible errors:
  - **Type I error**: false matches, record pairs identified as matches that are not true matches
  - **Type II error**: false non-matches, record pairs identified as non-matches that are true matches

Bias in linkage – An example

- A study was conducted to link Medicaid claims with evidence of pregnancy to vital records

  [Bronstein et al., Maternal Child Health Journal, 2009]

  - Among the matched records, 13% did not include claims for delivery services

- Selection bias was examined

  - To assess how accurately the evaluation dataset represents all Medicaid-covered pregnancies
  - Use of all pregnancy claims is likely to lower the match rate, as not all women who receive pregnancy-related services are represented by vital records
  - This selection bias introduced by using only evidence of delivery can be avoided by using evidence of pregnancy

Uncertainty and error propagation (1)

- Analytic linking (1)

  [Scheuren and Winkler, Survey Methodology, 1997]

  - Propagate uncertainty in record linkage into the analysis
  - Intended to adjust for biases and uncertainty introduced by linkage errors
  - Account for matching not directly comparable data and the effect of matching error in analyses
  - For example, *date of birth* from one data source and *age* from another source can be matched by estimating a new value, *estimated age*, in the first source

    Similarly, *receipt* and *income* from two different sources can be matched by estimating a new value, *estimated income from receipt*

Uncertainty and error propagation (2)

- Analytic linking (2)
  - Place predictors in one source that can be used to improve matching with the second source
  - Use simple predicted values that may not account for many types of matching error
  - After each matching pass, predictors are refined and improved iteratively
  - Summary representations (in graphs) are successively improved as erroneous data due to false matches are eliminated
  - Matching accuracy can also be improved by targeting outliers and systematic errors in the linked data

Uncertainty and error propagation (3)

- Bias adjustment [Scheuren and Winkler, Survey Methodology, 1993; Lahiri and Larsen, JASA, 2005]
  - Use dependent variable from one source and independent variable from the second source
  - If there is matching error, dependent and independent variables associated with false matches will not correspond as closely as those associated with true matches
  - Bias adjustments are based on probabilities of false match rates [Belin and Rubin, JASA, 1995]
    - Uses a mixture-model with weights for true matches and false matches; EM algorithm is used for fitting mixtures to find posterior modes
    - Variance in associated posterior is due to errors

Uncertainty and error propagation (4)

- Bridging files [Winker, ASA, 1999]
  - A large bridging file can be used to improve matching of two smaller files
  - A bridging file is a large universe file that approximately contains the two smaller files
  - For example, social security administration file of the whole population
  - Although the bridging file does not generally have all information for matching records from the two smaller data files, it has sufficient information for reducing the set of potential matches to small subsets
  - Additional linkage runs on the smaller data files can then lead to higher proportions of matches

Tutorial outline

- Introduction: What is Population Informatics?
- How is Population Informatics different from traditional sciences?
- The *Knowledge Based Platform*
- Privacy frameworks
- Linking and integrating databases
- Privacy-preserving record linkage
- Privacy-preserving interactive record linkage
- Data mining and analysis with linked population data
- Outlook, research directions and conclusions

Outlook, research directions and conclusions (1)

- Big Data challenges (4V's) for population informatics
  - **Volume** – increased volume of population data
  - **Velocity** – constantly updated
  - **Variety** – data from multiple heterogeneous sources
  - **Veracity** – inconsistent, incomplete, and erroneous data
- Scalability of linking large population databases
  - Advanced blocking/indexing/filtering techniques are required
  - Parallelisation or distributed computing

Outlook, research directions and conclusions (2)

- Integration of several databases from multiple sources
  - Most work so far is limited to linking two databases
  - In many real applications data are required from several organisations
  - Pair-wise integration or PPRL does not scale-up
  - Computational efforts increase with more parties
  - Preventing collusion between (sub-groups of) parties becomes more difficult in PPRL
  - Identifying matching records across subsets of parties requires more research in both non-PPRL and PPRL

Outlook, research directions and conclusions (3)

- Linkage of dynamic and temporal population data
  - Population data are dynamic reflecting people's changes over time
  - Temporal information need to be considered
  - How to best store, index, and link dynamic data?
  - How can temporal information improve linkage quality?
  - Real-time linking and analysis
- How to assess linkage quality?
  - Often no truth data available
  - Not possible in PPRL (as this would reveal sensitive information)

Outlook, research directions and conclusions (4)

- For linking historical data, the main challenge is data quality (develop (semi-)automatic data cleaning and standardisation techniques)
- No training data in many situations
  - Employ active learning approaches
  - Visualisation for improved manual clerical review
- Collections of test data sets which can be used by researchers
  - Challenging (impossible?) to have true match status
  - Challenging as most data are either proprietary or sensitive

Outlook, research directions and conclusions (5)

- Frameworks that allow comparative experimental studies
- Develop practical PPRL techniques
  - Standard measures for privacy
  - Advanced blocking/filtering techniques for PPRL
  - Improved advanced classification techniques for PPRL
  - Methods to assess accuracy and completeness
- Multi-party PPRL (a challenging task, due to exponential comparison space and issue of collusion)
- Pragmatic challenge: Collaborations across multiple research disciplines

Thank you for attending our tutorial!

Enjoy PAKDD and your stay in Auckland...

For questions please contact:

peter.christen@anu.edu.au

kum@tamu.edu

qing.wang@anu.edu.au

dinusha.vatsalan@anu.edu.au

References (1)

- Agrawal R, Evfimievski A, and Srikant R: *Information sharing across private databases.* ACM SIGMOD, San Diego, 2005.
- Al-Lawati A, Lee D and McDaniel P: *Blocking-aware private record linkage.* IQIS, Baltimore, 2005.
- Atallah MJ, Kerschbaum F and Du W: *Secure and private sequence comparisons.* WPES, Washington DC, pp. 39–44, 2003.
- Bachteler T, Schnell R, and Reiher J: *An empirical comparison of approaches to approximate string matching in private record linkage.* Statistics Canada Symposium, 2010.
- Barone D, Maurino A, Stella F, and Batini C: *A privacy-preserving framework for accuracy and completeness quality assessment.* Emerging Paradigms in Informatics, Systems and Communication, 2009.
- Belin T. R and Rubin D. B: *A method for calibrating false-match rates in record linkage.* Journal of the American Statistical Association, vol. 90, pp. 694–707, 1995.

References (2)

- Bellare K, Curino C, Machanavajihala A, Mika P, Rahurkar M, and Sane A: *Woo: A scalable and multi-tenant platform for continuous knowledge base synthesis*. VLDB Endowment, 6(11), pp. 1114–1125, 2013.
- Bhattacharya, I and Getoor, L: *Collective entity resolution in relational data.* ACM TKDD, 2007.
- Blakely T, Woodward A and Salmond C: *Anonymous linkage of New Zealand mortality and census data.* ANZ Journal of Public Health, 24(1), 2000.
- Bloom, BH: *Space/time trade-offs in hash coding with allowable errors.* Communications of the ACM, 1970.
- Bonomi L, Xiong Li, Chen R, and Fung B: *Frequent grams based embedding for privacy preserving record linkage.* ACM Information and knowledge management, 2012.
- Bouzelat H, Quantin C, and Dusserre L: *Extraction and anonymity protocol of medical file.* AMIA Fall Symposium, 1996.

References (3)

- Bronstein J.M, Lomatsch C.T, Fletcher D, Wooten T, Lin T, Nugent R, and Lowery C.L: *Issues and biases in matching Medicaid pregnancy episodes to vital records data: the Arkansas experience.* Springer Maternal and child health journal, 13(2), pp. 250–259, 2009.
- Cavallo A: *Online and official price indexes: measuring Argentina's inflation.* Elsevier Journal of Monetary Economics, 60(2), pp. 152–165, 2013.
- Chaytor R, Brown E and Wareham T: *Privacy advisors for personal information management.* SIGIR workshop on Personal Information Management, Seattle, pp. 28–31, 2006.
- Christen P: *Privacy-preserving data linkage and geocoding: Current approaches and research directions.* PADM held at IEEE ICDM, Hong Kong, 2006.
- Christen P: *Geocode Matching and Privacy Preservation.* ACM PinKDD, 2009.
- Christen, P: *A survey of indexing techniques for scalable record linkage and deduplication.* IEEE TKDE, 2012.

References (4)

- Christen, P: *Data matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer, 2012.
- Christen, P: *Preparation of a real voter data set for record linkage and duplicate detection research.* Technical Report, The Australian National University, 2013.
- Christen P and Churches T: *Secure health data linkage and geocoding: Current approaches and research directions.* ehPASS, Brisbane, 2006.
- Christen, P and Goiser, K: *Quality and complexity measures for data linkage and deduplication.* In *Quality Measures in Data Mining.* Springer Studies in Computational Intelligence, vol. 43, 2007.
- Christen P, Vatsalan D, and Verykios V: *Challenges for privacy preservation in data integration.* In *Journal of Data and Information Quality.* ACM, vol. 5, 2014.
- Churches T: *A proposed architecture and method of operation for improving the protection of privacy and confidentiality in disease registers.* BMC Medical Research Methodology, 3(1), 2003.

References (5)

- Churches T and Christen P: *Some methods for blindfolded record linkage.* BMC Medical Informatics and Decision Making, 4(9), 2004.
- Clifton C, Kantarcioglu M, Vaidya J, Lin X, and Zhu MY: *Tools for privacy preserving distributed data mining.* ACM SIGKDD Explorations, 2002.
- Clifton C, Kantarcioglu M, Doan A, Schadow G, Vaidya J, Elmagarmid AK and Suciu D: *Privacy-preserving data integration and sharing.* SIGMOD workshop on Research Issues in Data Mining and Knowledge Discovery, Paris, 2004.
- Dinur I and Nissim K: *Revealing information while preserving privacy.* ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, California, 2003.
- Du W, Atallah MJ, and Kerschbaum F: *Protocols for secure remote database access with approximate matching.* ACM Workshop on Security and Privacy in E-Commerce, 2000.
- Dusserre L, Quantin C and Bouzelat H: *A one way public key cryptosystem for the linkage of nominal files in epidemiological studies.* Medinfo, 8:644-7, 1995.

References (6)

- Durham, EA: *A framework for accurate, efficient private record linkage.* PhD Thesis, Vanderbilt University, 2012.
- Durham, EA, Toth C, Kuzu, M. Kantarcioglu M, and Malin B: *Composite Bloom for secure record linkage.* IEEE Transactions on Knowledge and Data Engineering, 2013.
- Durham, EA, Xue Y, Kantarcioglu M, and Malin B: *Private medical record linkage with approximate matching.* AMIA Annual Symposium, 2010.
- Durham, EA, Xue Y, Kantarcioglu M, and Malin B: *Quantifying the correctness, computational complexity, and security of privacy-preserving string comparators for record linkage.* Information Fusion, 2012.
- DuVall S, Fraser A, Rowe K, Thomas A, and Mineau G: *Evaluation of record linkage between a large healthcare provider and the Utah Population Database.* Journal of the American Medical Informatics Association, 19(E1), pp. E54 – E59, 2012.
- Dwork, C: *Differential privacy.* International Colloquium on Automata, Languages and Programming, 2006.

References (7)

- Elmagarmid AK, Ipeirotis PG and Verykios VS: *Duplicate record detection: A survey.* IEEE TKDE 19(1), pp. 1–16, 2007.
- Fienberg SE: *Privacy and confidentiality in an e-Commerce World: Data mining, data warehousing, matching and disclosure limitation.* Statistical Science, IMS Institute of Mathematical Statistics, 21(2), pp. 143–154, 2006.
- Gostin L.O et al.: *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research.* National Academies Press, 2009.
- Hall R and Fienberg SE: *Privacy-preserving record linkage.* Privacy in Statistical Databases, Springer LNCS 6344, 2010.
- Hertzman C, Pencarrick N, and McGrail K: *Privacy by Design at Population Data BC: a case study describing the technical, administrative, and physical controls for privacy-sensitive secondary use of personal information for research in the public interest.* Journal of the American Medical Informatics Association, 20(1), pp. 25–28, 2013.
- Herzog TN, Scheuren F, and Winkler WE: *Data quality and record linkage techniques.* Springer, 2007.

References (8)

- Holman et al.: *A decade of data linkage in Western Australia: strategic design, applications and benefits of the WA data linkage system.* CSIRO Australian Health Review, 32(4), pp. 766–777, 2008.
- Ibrahim A, Jin H, Yassin AA, and Zou D: *Approximate Keyword-based Search over Encrypted Cloud Data.* IEEE ICEBE, pp. 238–245, 2012.
- Inan A, Kantarcioglu M, Bertino E and Scannapieco M: *A hybrid approach to private record linkage.* IEEE ICDE, Cancun, Mexico, pp. 496–505, 2008.
- Inan A, Kantarcioglu M, Ghinita G, and Bertino E: *Private record matching using differential privacy.* International Conference on Extending Database Technology, 2010.
- Jones JJ, Bond RM, Fariss CJ, Settle JE, Kramer ADI, Marlow C, and Fowler JH: *Yahtzee: An anonymized group level matching procedure.* PloS one, vol. 8, 2013.
- Kang H, Getoor L, Shneiderman B, Bilgic M, and Licamele L: *Interactive entity resolution in relational data: A visual analytic tool and its evaluation.* IEEE Transactions on Visualization and Computer Graphics, 14(5), pp. 999–1014, 2008.

References (9)

- Kantarcioglu M, Jiang W, and Malin B: *A privacy-preserving framework for integrating person-specific databases.* Privacy in Statistical Databases, 2008.
- Kantarcioglu M, Inan A, Jiang W and Malin B: *Formal anonymity models for efficient privacy-preserving joins.* Data and Knowledge Engineering, 2009.
- Karakasidis A and Verykios VS: *Privacy preserving record linkage using phonetic codes.* IEEE Balkan Conference in Informatics, 2009.
- Karakasidis A and Verykios VS: *Advances in privacy preserving record linkage.* E-activity and Innovative Technology, Advances in Applied Intelligence Technologies Book Series, IGI Global, 2010.
- Karakasidis A and Verykios VS: *Secure blocking+secure matching = Secure record linkage.* Journal of Computing Science and Engineering, 2011.
- Karakasidis A, Verykios VS, and Christen P: *Fake injection strategies for private phonetic matching.* International Workshop on Data Privacy Management, 2011.
- Karakasidis A and Verykios VS: *Reference table based k-anonymous private blocking.* Symposium on Applied Computing, 2012.

References (10)

- Karakasidis A and Verykios VS: *A sorted neighborhood Approach to multidimensional privacy preserving blocking.* IEEE ICDM workshop, 2012.
- Karapiperis D and Verykios VS: *A distributed framework for scaling Up LSH-based computations in privacy preserving record linkage.* Balkan Conference in Informatics, 2013.
- Kelman CW, Bass AJ and Holman CDJ: *Research use of linked health data – A best practice protocol.* ANZ Journal of Public Health, 26(3), pp. 251–255, 2002.
- Kum, H.-C., Duncan, D.F. and Stewart, C.J.: *Supporting self-evaluation in local government via Knowledge Discovery and Data mining.* Government Information Quarterly, 26(2), pp. 295-304, 2009.
- Kum H.-C and Ahalt S: *Privacy by design: understanding data access models for secondary data.* AMIA Joint Summits on Translation Science and Clinical Research Informatics, 2013.
- Kum H.-C, Krishnamurthy A, Machanavajjhala A, and Ahalt S: *Social genome: Putting big data to work for population informatics.* IEEE Computer, 2014.

References (11)

- Kum, H.-C., Ahalt, S., and Pathak, D.: *Privacy-Preserving Data Integration Using Decoupled Data.* Security and Privacy in Social Networks, Springer, pp. 225-253, 2013.
- Kum, H.-C., Krishnamurthy, A., Pathak, D., Reiter, M., and Ahalt, S.: *Secure Decoupled Linkage (SDLink) system for building a social genome.* IEEE International Conference on BigData, 2013.
- Kum H.-C, Krishnamurthy A, Machanavajjhala A, Reiter M.K, and Ahalt S: *Privacy preserving interactive record linkage (PPIRL).* Journal of the American Medical Informatics Association, 21(2), pp. 212–220, 2014.
- Kum, H.-C., Stewart, C.J., Rose, R.A. and Duncan, D.F.: *Using big data for evidence based governance in child welfare.* Children and Youth Services, Review (2015), Volume 58, pp. 127-136, ISSN 0190-7409, 2015.
- Kum, H-C: *Applying Data Science to Advance the Health and Welfare of Populations.* Annual Conference on Society for Social Work and Research (SSWR), Invited half day workshop, New Orleans, LA, 2015.

References (12)

- Kuzu M, Kantarcioglu M, Durham EA and Malin B: *A constraint satisfaction cryptanalysis of Bloom filters in private record linkage.* Privacy Enhancing Technologies, 2011.
- Kuzu M, Kantarcioglu M, Inan A, Bertino E, Durham EA and Malin B: *Efficient privacy-aware record integration.* ACM Extending Database Technology, 2013.
- Kuzu M, Kantarcioglu M, Durham EA, Toth C, and Malin B: *A practical approach to achieve private medical record linkage in light of public resources.* Journal of the American Medical Informatics Association, vol. 20, pp. 285–292, 2013.
- Lahiri P and Larsen M: *Regression analysis with linked data.* Journal of the American statistical association, 100(469), pp. 222–230, 2005.
- Lai PK, Yiu SM, Chow KP, Chong CF, and Hui LC: *An efficient Bloom filter based solution for multiparty private matching.* International Conference on Security and Management, 2006.
- Li Y, Tygar JD and Hellerstein JM: *Private matching.* Computer Security in the 21st Century, Lee DT, Shieh SP and Tygar JD (editors), Springer, 2005.

References (13)

- Li F, Chen Y, Luo B, Lee D, and Liu P: *Privacy preserving group linkage.* Scientific and Statistical Database Management, 2011.
- Lyons R et al.: *The SAIL databank: linking multiple health and social care datasets.* BMC Medical Informatics and Decision Making, 9(1), 2009.
- Malin B, Airoldi E, Edoho-Eket S and Li Y: *Configurable security protocols for multi-party data analysis with malicious participants.* IEEE ICDE, Tokyo, pp. 533–544, 2005.
- Malin B and Sweeney L: *A secure protocol to distribute unlinkable health data.* American Medical Informatics Association, Washington DC, pp. 485–489, 2005.
- Mohammed N, Fung BC and Debbabi M: *Anonymity meets game theory: secure data integration with malicious participants.* VLDB Journal, 2011.
- Murugesan M, Jiang W, Clifton C, Si L and Vaidya J: *Efficient privacy-preserving similar document detection.* VLDB Journal, 2010.
- Naumann F and Herschel M: *An introduction to duplicate detection.* Synthesis Lectures on Data Management, Morgan and Claypool Publishers, 2010.

References (14)

- Navarro-Arribas G and Torra V: *Information fusion in data privacy: A survey.* Information fusion, 2012.
- Nissenbaum H: *Privacy as contextual integrity.* Washington Law Rev. 79(1), pp. 19–158, 2004.
- O'Keefe CM, Yung M, Gu L and Baxter R: *Privacy-preserving data linkage protocols.* WPES, Washington DC, pp. 94–102, 2004.
- Pang C, Gu L, Hansen D and Maeder A: *Privacy-preserving fuzzy matching using a public reference table.* Intelligent Patient Management, 2009.
- Quantin C, Bouzelat H and Dusserre L: *Irreversible encryption method by generation of polynomials.* Medical Informatics and The Internet in Medicine, Informa Healthcare, 21(2), pp. 113–121, 1996.
- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *How to ensure data quality of an epidemiological follow-up: Quality assessment of an anonymous record linkage procedure.* International Journal of Medical Informatics, 49, pp. 117–122, 1998.

References (15)

- Quantin C, Bouzelat H, Allaert FAA, Benhamiche AM, Faivre J and Dusserre L: *Automatic record hash coding and linkage for epidemiological follow-up data confidentiality.* Methods of Information in Medicine, Schattauer, 37(3), pp. 271–277, 1998.
- Ravikumar P, Cohen WW and Fienberg SE: *A secure protocol for computing string distance metrics.* PSDM held at IEEE ICDM, Brighton, UK, 2004.
- Scannapieco M, Figotin I, Bertino E and Elmagarmid AK: *Privacy preserving schema and data matching.* ACM SIGMOD, 2007.
- Schadow G, Grannis SJ and McDonald CJ: *Discussion paper: Privacy-preserving distributed queries for a clinical case research network.* CRPIT'14: Proceedings of the IEEE international Conference on Privacy, Security and Data Mining, Maebashi City, Japan, pp. 55–65, 2002.
- Schnell R, Bachteler T and Reiher J: *Privacy-preserving record linkage using Bloom filters.* BMC Medical Informatics and Decision Making, 9(1), 2009.
- Scheuren F, and Winkler W. E: *Regression analysis of data files that are computer matched.* Survey Methodology, vol. 19, pp. 39-58, 1993.

References (16)

- Scheuren F, and Winkler W. E: *Regression analysis of data files that are computer matched, II.* Survey Methodology, vol. 23, pp. 157-165, 1997.
- Schnell R, Bachteler T and Reiher J: *A novel error-tolerant anonymous linking code.* German record linkage center working paper series, 2011.
- Schnell R: *Privacy-preserving record linkage and privacy-preserving blocking for large files with cryptographic keys using multibit trees.* ASA JSM Proceedings, Alexandria, VA, 2013.
- Sweeney L: *Privacy-enhanced linking.* ACM SIGKDD Explorations, 7(2), 2005.
- Trepetin S: *Privacy-preserving string comparisons in record linkage systems: a review.* Information Security Journal: A Global Perspective, 2008.
- Vatsalan D, Christen P and Verykios VS: *An efficient two-party protocol for approximate matching in private record linkage.* AusDM, CRPIT, 2011.
- Vatsalan D and Christen P: *An iterative two-party protocol for scalable privacy-preserving record linkage.* AusDM, CRPIT, vol. 134, 2012.
- Vatsalan D and Christen P: *Sorted nearest neighborhood clustering for efficient private blocking.* PAKDD, Gold Coast, Australia, Springer LNCS vol. 7819, 2013.

References (17)

- Vatsalan D, Christen P and Verykios VS: *A taxonomy of privacy-preserving record linkage techniques.* Journal of Information Systems, 2013.
- Vatsalan D, Christen P and Verykios VS: *Efficient two-party private-blocking based on sorted nearest neighborhood clustering.* CIKM, 2013.
- Vaidya J and Clifton C: *Secure set intersection cardinality with application to association rule mining.* Journal of Computer Security, 2005.
- Verykios VS, Karakasidis A and Mitrogiannis VK: *Privacy preserving record linkage approaches.* International Journal of Data Mining, Modelling and Management, 2009.
- Wartell J and McEwen T: *Privacy in the information age: A Guide for sharing crime maps and spatial data.* Institute for Law and Justice, National Institute of Justice, 188739, 2001.
- Weber SC, Lowe H, Das A and Ferris T: *A simple heuristic for blindfolded record linkage.* Journal of the American Medical Informatics Association, 2012.
- Weitzner D.J et al.: *Information accountability.* ACM Communications, 51(6), pp. 82–87, 2008.

References (18)

- Winkler WE: *Issues with linking files and performing analyses on the merged files.* Section on Government Statistics and Social Statistics, American Statistical Association, pp. 262–265, 1999.
- Winkler WE: *Masking and re-identification methods for public-use microdata: Overview and research problems.* Privacy in Statistical Databases, Barcelona, Springer LNCS 3050, pp. 216–230, 2004.
- Winkler WE: *Overview of record linkage and current research directions.* RR 2006/02, US Census Bureau, 2006.
- Yakout M, Atallah MJ and Elmagarmid AK: *Efficient private record linkage.* IEEE ICDE, 2009.
- Yao, AC: *How to generate and exchange secrets.* Annual Symposium on Foundations of Computer Science, 1986.
- Zhang Q and Hansen D: *Approximate processing for medical record linking and multidatabase analysis.* International Journal of Healthcare Information Systems and Informatics, 2(4), pp. 59–72, 2007.