Australian National University

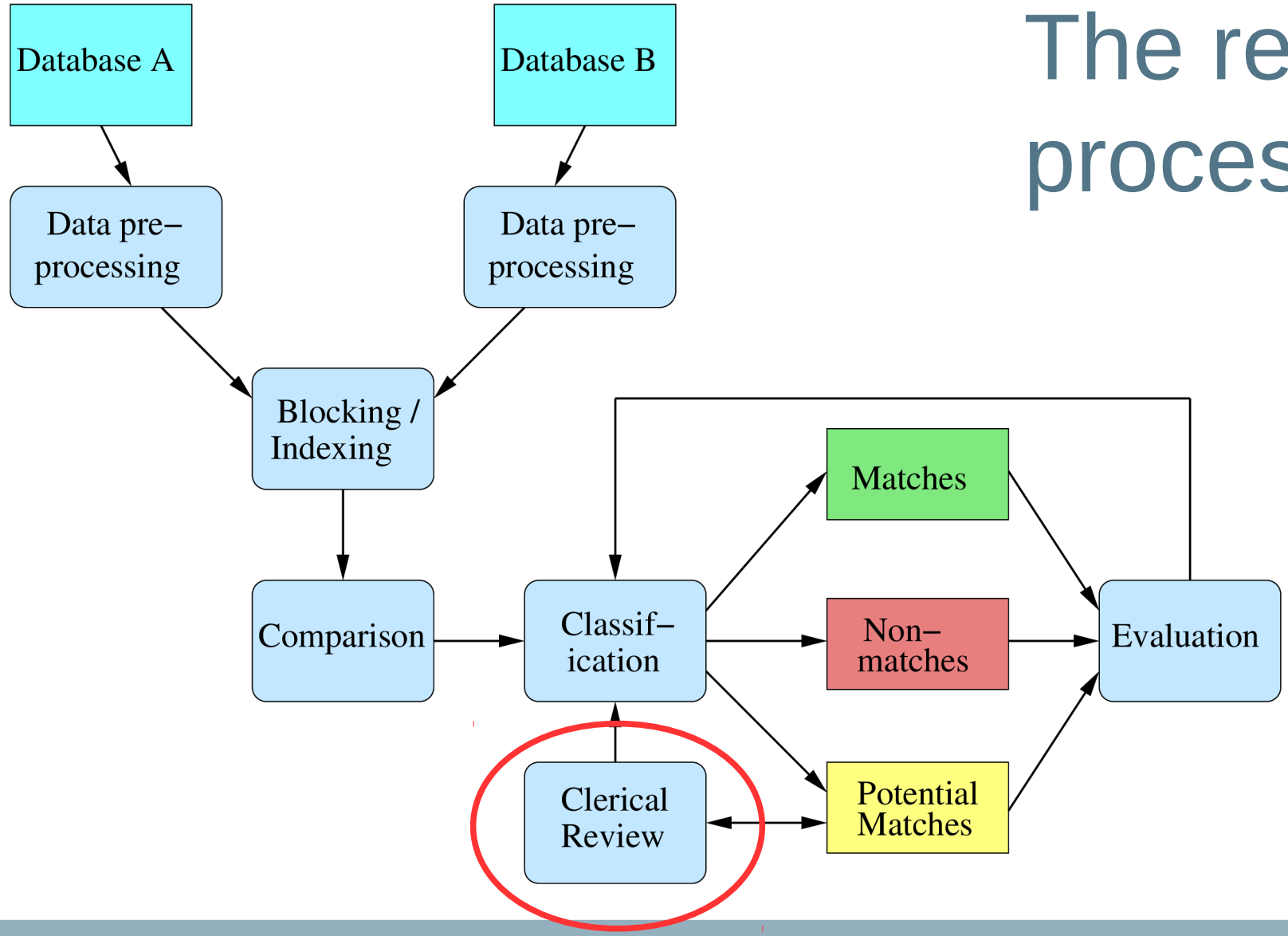
COMP3430 / COMP8430
Data wrangling

Lecture 20: Record linkage evaluation (2)
(Lecturer: Peter Christen)

Lecture outline

- Clerical review
- Test databases and benchmarks
- Synthetic test data

The record linkage process



Why clerical review?

- The traditional (probabilistic) record linkage process classifies record pairs into the class of **potential matches** if no clear decision can be made
- These record pairs are given to a domain expert for manual classification
- Manual classification requires inspection of the attribute values of a record pair, and possibly also external information
  - Other records about the same entities from related databases
  - Information found on the Internet
  - Information from the linkage process (how many other similar records are there – is a record pair unique or highly connected?)

Example clerical review (1)

RECORD ID	SURNAME	GIVEN NAME	GENDER	AGE	Similarity
a116	Stephens	**Sally**	F	**16**	75%
b342	Stephens	**S**	F	**18**	

Match

Non–Match

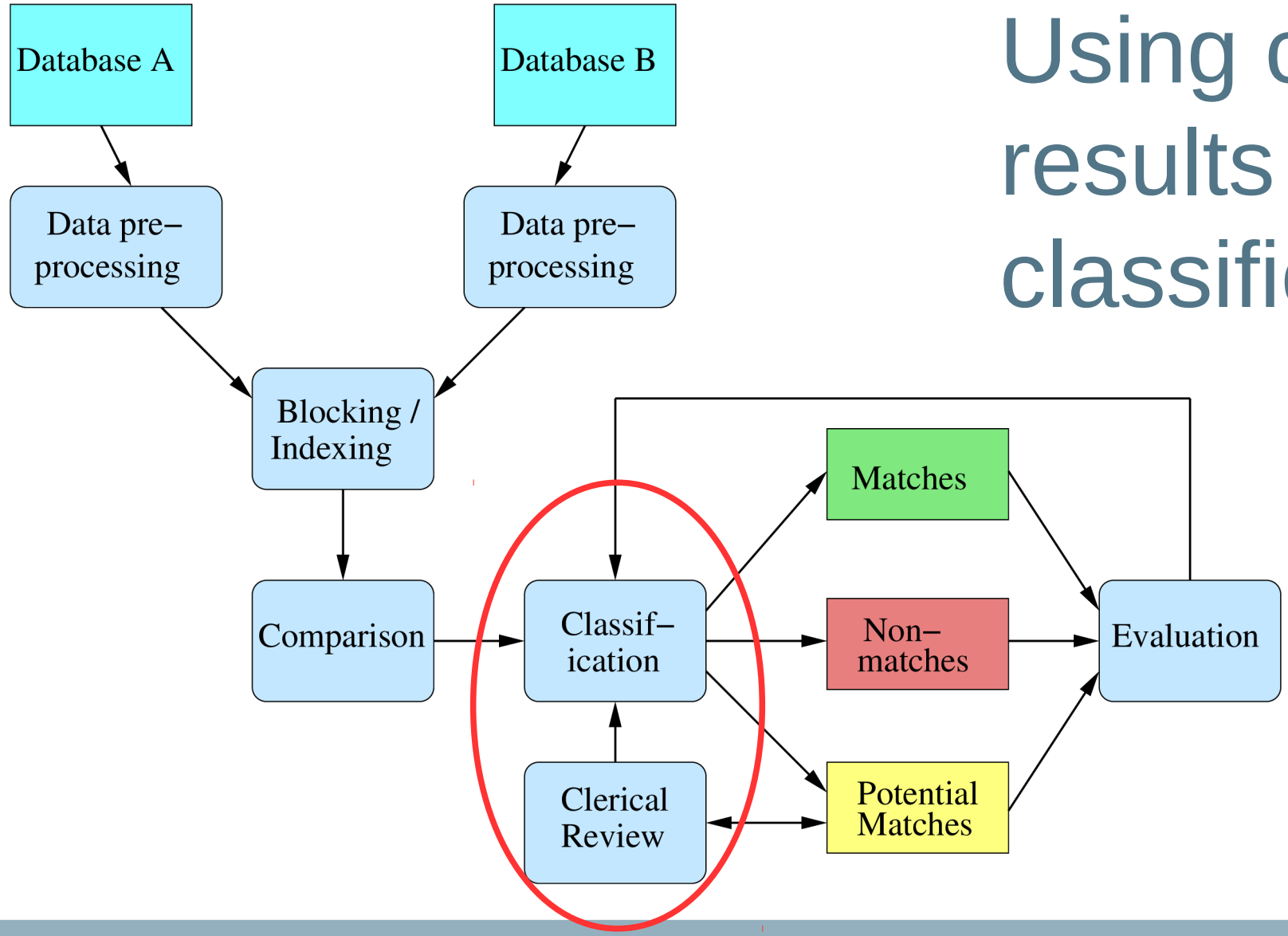
- **Question**: *Do these two records refer to the same person?*
  Is this a good way to support clerical review?

Example clerical review (2)

RECORD ID	SURNAME	GIVEN NAME	GENDER	AGE	Similarity
a116	Stephens	**Sally**	F	**16**	75%
b342	Stephens	**S**	F	**18**	

Clear match | Likely match | Likely non–match | Clear non–match

Using clerical review results to improve classification



Clerical review results to improve classification

- A major challenge with record linkage classification is the lack of ground truth data in many applications
  - Preventing the use of supervised machine learning classification techniques
- The clerical review process generates training data
  - Of the difficult to classify cases
  - These can be used in a process called *active learning* which combines manual with machine learning based classification
- Recent research has investigated crowed-based approaches to clerical review (using systems such as *Amazon's Mechanical Turk*)
  - Pay small amount of money for each manual classification
  - Applicable only in situations where the data are not sensitive

Test databases and benchmarks (1)

- To evaluate a record linkage system or software, it needs to be employed on a set of suitable databases from the same domain
  - Similar in size and characteristics to the databases it will be deployed on
  - Ground truth needs to be available in these test databases to evaluate blocking and linkage quality (pairs completeness, pairs quality, precision, recall, etc.)
- It is generally very difficult to obtain such test databases
  - Because in record linkage we often need personal details of people
  - Linking product or bibliographic databases (which are publicly available) will not be helpful for linking personal data

Test databases and benchmarks (2)

- In other research areas (databases, data mining, machine learning, information retrieval) there are publicly available test data collections or even benchmarking systems
  - Transaction Processing Performance Council (TPC)
  - UCI Machine Learning and KDD Repository
  - Text Retrieval Conferences (TREC)

- In record linkage, individual research groups provide small test data collections

Generating and using synthetic data (1)

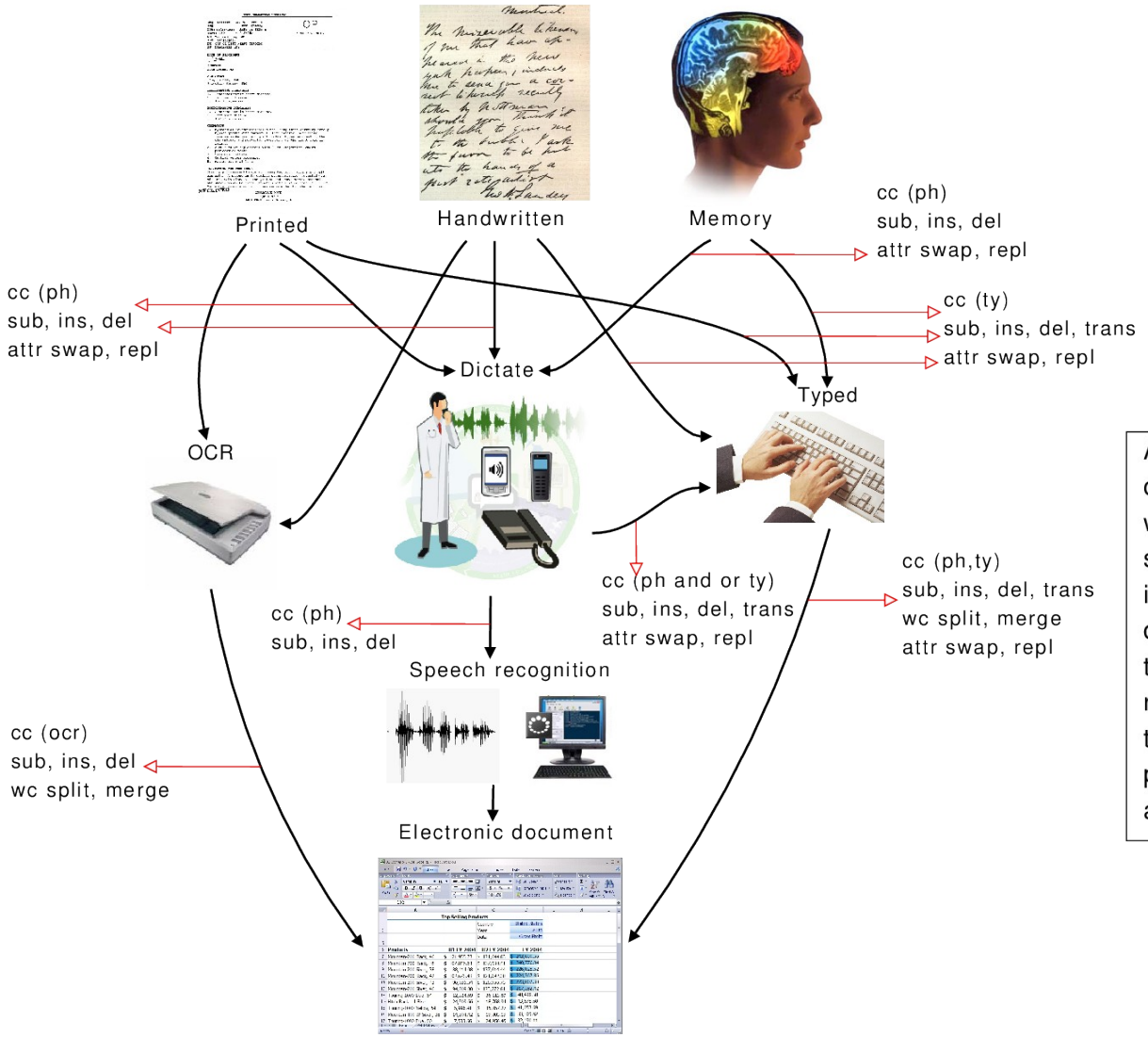
- Privacy issues prohibit publication of real personal information
- De-identified or encrypted data cannot be used for record linkage research or evaluation of record linkage systems
  (because personal details such as real names and addresses are needed)
- Alternatively, create artificial / synthetic databases for research and evaluation, with several advantages
  - Volume and characteristics can be controlled (errors and variations in records, number of duplicates, etc.)
  - It is known which records are duplicates of each other, and so linkage quality can be calculated
  - Data and the data generator program can be published

Generating and using synthetic data (2)

- Creating synthetic data is usually a two step process
  - First generate data based on lookup tables with real values and their distributions (name and address values and their frequencies), and functions that generate values following real distributions (age, salary, blood pressure, etc.), and that model dependencies between attributes
  - Then corrupt the generated data using realistic corruption functions (such as typos, phonetic variations, OCR errors, etc)

- Various data generators have been developed, including at ANU: https://dmm.anu.edu.au/geco (*GeCo* generator / corruptor)

Modelling of variations and errors



Abbreviations:
cc : character change
wc : word change
subs : substitution
ins : insertion
del : deletion
trans : transpose
repl : replace
ty : typographic
ph : phonetic
attr : attribute

Example of generated data

RecID	Age	GivenName	Surname	Street	Suburb
rec-1-org	33	Madison	Solomon	Tazewell Circuit	Beechboro
rec-1-dup-0	33	*Madisoi*	Solomon	Tazewell *Circ*	*Beech Boro*
rec-1-dup-1		Madison	Solomon	Tazewell *Crct*	*Bechboro*
rec-2-org	39	Desirae	Contreras	Maltby Street	Burrawang
rec-2-dup-0	39	Desirae	*Kontreras*	Maltby Street	*Burawang*
rec-2-dup-1	39	*Desire*	Contreras	Maltby Street	*Buahrawang*
rec-3-org	81	Madisyn	Sergeant	Howitt Street	Nangiloc
rec-3-dup-0	*87*	*Madisvn*	Sergeant	*Hovvitt* Street	*Nanqiloc*

- **rec-1**: typing/abbreviations; **rec-2**: phonetic; **rec-3**: OCR
- Generated using the *Febrl* and *GeCo* data generators