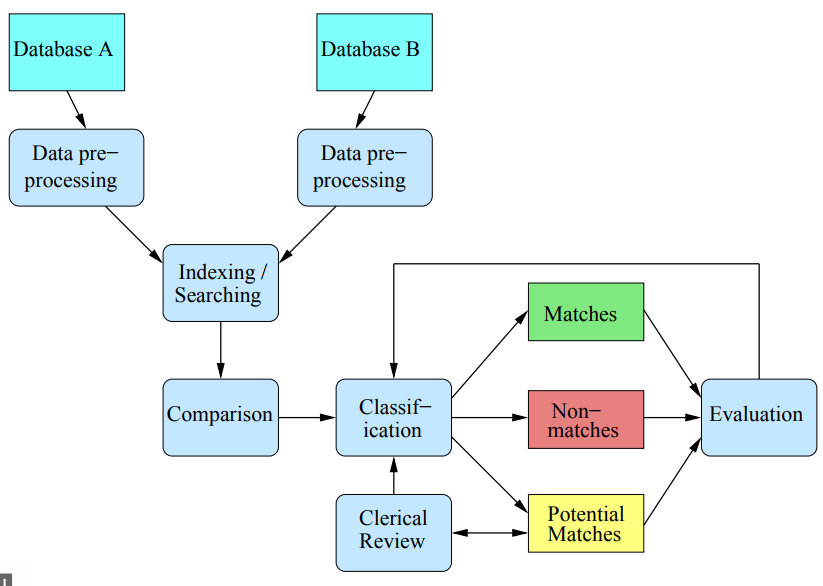
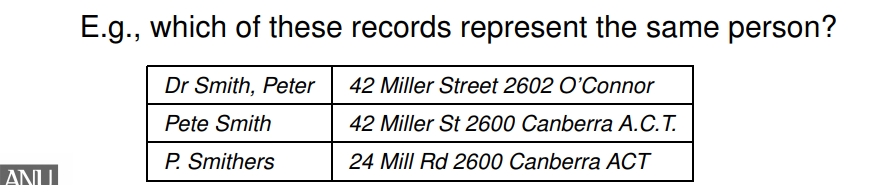
**RECORD** **LINKER :**

**WHAT IS RECORD LINKAGE?**

****

* The process of linking records that represent the same entity in one or more databases (patient, customer, business name, etc.)
* Also known as data matching, entity resolution, object identification, duplicate detection, identity uncertainty, merge-purge, etc.
* Major challenge is that unique entity identifiers are not available in the databases to be linked (or if available, they are not consistent or change over time )



**APPLICATION :**

* Remove duplicates in one data set (de-duplication)
* Merge new records into a larger master data set
* Create patient or customer oriented statistics (for example for longitudinal studies)
* Clean and enrich data for analysis and mining
* Geo-code matching (with reference address data)
* Widespread use of record linkage :
* Immigration, taxation, social security, census
* Fraud, crime, and terrorism intelligence
* Business mailing lists, exchange of customer data
* Health, social science, and data science research

**USE CASE** :

* Find new clients who come from insurance comparison services → Commission
* Find duplicates in existing files (acquisitions)

**Elaboration** :

The insurer has a website in France where clients like you and me can purchase an insurance for their home or car. Clients may come to this website through a quote comparison website such as LesFurets.com or LeLynx.com. When this is the case, the comparison website should be retributed for bringing a new client.

***First issue is***, the client might make a first quote comparison then, only a couple of weeks later, proceed with the subscription. The commission should still be paid, though. That’s why we have to compare the listings of clients – the listing from the insurer and the listing from the quote comparison website – and link records together: that’s what we call *Record Linkage*.

***Second issue is***, neither the insurer nor the comparison website are willing to provide their full listing to each other. They want to keep this confidential and, as a workaround, they provide anonymized data.

A secondary use case is when we have to find duplicates in a listing of clients. This can be done using the same engine and that explains why this problem can carry several names: De-duplication, Entity Resolution…

**RECORD LINKAGE CHALLENGE :**

* No unique entity identifiers are available (use approximate (string) comparison functions)
* Real world data are dirty (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
* Scalability to very large databases (naïve comparison of all record pairs is quadratic; some form of blocking, indexing or filtering is needed)
* No training data in many record linkage applications (true match status not known)
* Privacy and confidentiality (because personal information is commonly required for linking)

**CHALLENGES FOR DATA SCIENCE :**

* Often the aim is to create “social genomes” for individuals by linking large population databases
* Knowing how individuals and families change over time allows for a diverse range of studies (fertility, employment, education, health, crimes, etc.)
* Different challenges for historical data compared to contemporary data, but some are common :
* Database sizes (computational aspects)
* Accurate match classification
* Sources and coverage of population databases

**CHALLENGES OF HISTORICAL DATA :**

* Low literacy (recording errors and unknown exact values), no address or occupation standards
* Large percentage of a population had one of just a few common names (‘John’ or ‘Mary’)
* Households and families change over time
* Immigration and emigration, birth and death
* Scanning, OCR, and transcription errors

**CHALLENGES FOR PRESENT-DAY DATA :**

* These data are about living people, and so privacy is of concern when data are linked between organizations
* Linked data allow analysis not possible on individual databases (potentially revealing highly sensitive information)
* Modern databases contain more details and more complex types of data (free-format text or multimedia)
* Data are available from different sources (governments, businesses, social network sites, the Web)
* Major questions: Which data are suitable?

Which can we get access to?

Some useful links:

* <https://stackoverflow.com/questions/16381133/using-machine-learning-to-de-duplicate-data>
* <http://www.idir.st-andrews.ac.uk/wp-content/uploads/2015/07/christen2015summer-v-slides.pdf>