Office for National Statistics

Data Linkage at ONS

Shelley Gammon

Summary

- What linkage do we do?
- Case study: 2011 Census
- Challenges

National Statistics Code of Practice

- The Protocol on The Use of Administrative Sources for Statistical Purposes

“Good practices will maximise opportunities for the use of administrative data, cross analysis of sources and exchange and re-use of data, to avoid duplicating requests for information.”

What data do we link

- Addresses
  - e.g. geo-referencing of admin data
- Businesses
  - Inter-Departmental Business Register
- Persons
  - Internal migration
  - Population estimation

Data linkage example - Census

- No population register in the UK, so we do a census to count the population
- But ...

  Not everyone completes a census form.

  Some people less likely to be counted e.g.
  - Young men
  - Babies
  - Non-white British people
  - People living in high density housing

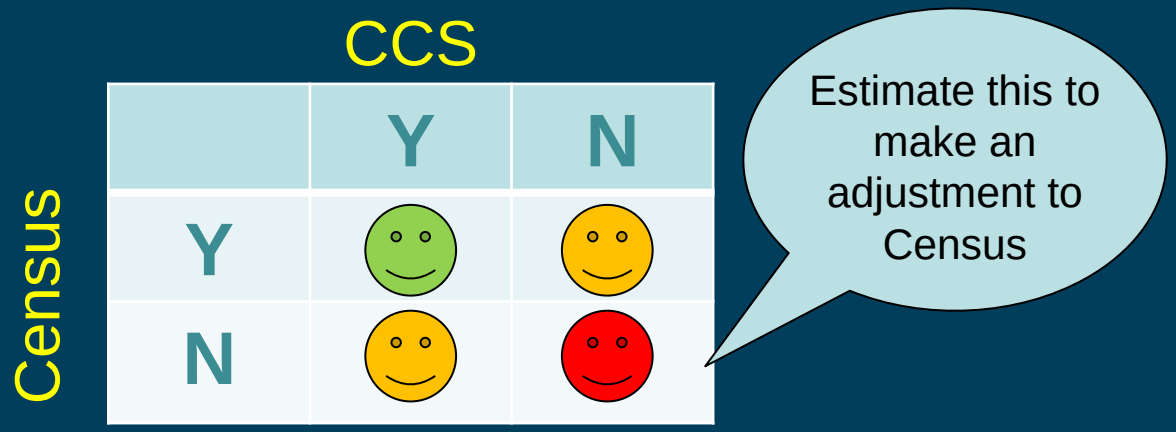
Data linkage example – Census

- In order to estimate coverage and work out who was missed from the census, we do a Census Coverage Survey (CCS)
- CCS – 1% sample survey asking similar questions to the Census
- Run 6-8 weeks after the Census

Data linkage example – Census

Match Census to Census Coverage Survey (CCS)

- to find out:



- Quality requirements very high
  - ➢ we are interested in subgroups
  - ➢ the matching results from the sample are scaled up to population

2011 Census to CCS linkage

1. Exact linkage
2. Probabilistic (Fellegi Sunter)
    - very high threshold
    - zero false positives
3. Clerical evaluation of high-medium scoring candidate pairs
4. Clerical searching for resolution of unmatched records

2011 Census to CCS matching

- 70% of all person matches were made automatically
- clerical resource equivalent to 30 FTE / 30 weeks
- Very high quality:
  >99.9% precision,
  >99.25% recall

The future…

- Census in 2021 will rely on data linkage
  - Census to CCS
  - Increased use of admin data
  - QA of population estimates
  - Address register
- Beyond 2021 – will we need a census?
  - Can we use administrative data to estimate
    i) population size
    ii) population characteristics ?

What have we done so far?

- Extensive feasibility research into administrative data matching (Beyond 2011)
- Produced population estimates by linking:
    - NHS Patient Register (PR),
    - DWP Customer Information System (CIS) and
    - Higher Education Statistics Agency (HESA) data

What have we done so far? (2)

- Focused on the development of automated matching algorithms:
  - Required to link large admin datasets (100m + records)
  - Adopted a ‘pseudonymisation’ approach as part of a tactical solution to preserve privacy in the linkage process
- ONS is reviewing its longer term approach to privacy and data linkage but continue to build on methodological research to date

Beyond 2011 Methods

- Beyond 2011 matching was based around automated linkage

  No common identifier so matching on name, sex date of birth and address

- Three stages to the algorithm:

  (1) Deterministic matching: Linking records on a series of 'match-keys' using exact or partial agreement on a combination of fields

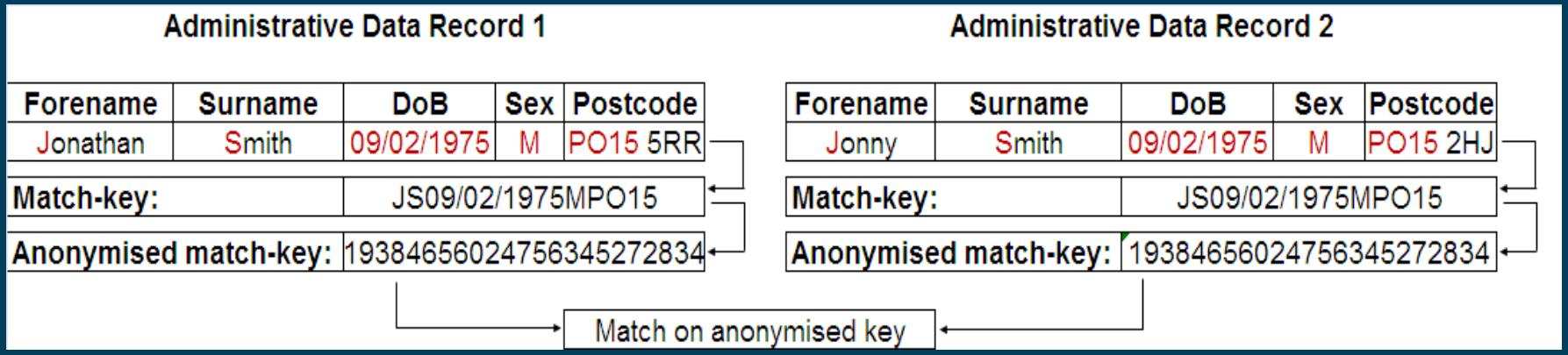
  (2) Logistic regression matching: Using clerically matched training data to model matching decisions for more complex cases

  (3) Associative matching: linking individuals based on collectively resolving matches within a household

- The three stages are designed to run sequentially

Match-Keys

- **Match-keys:** used to resolve common minor inconsistencies between match-fields on two datasets



Key	Type	Unique records on PR
1	Forename, Surname, DoB, Sex, Postcode	100.0%
2	Forename initial , Surname initial, DoB, Sex, Postcode District	99.6%
3	Forename bi-gram, Surname bi-gram, DoB, Sex, Postcode Area	99.4%
4	Forename initial, DoB, Sex, Postcode	99.8%
5	Surname initial, DoB, Sex, Postcode	99.4%
6	Forename, Surname, Age, Sex, Postcode Area	99.5%

What did we learn from B2011?

- Match rates between key admin datasets were high
  - 95% of PR records link to the CIS
  - 97% of School Census records link to the PR
  - Formed the basis of a Statistical Population Dataset (SPD)

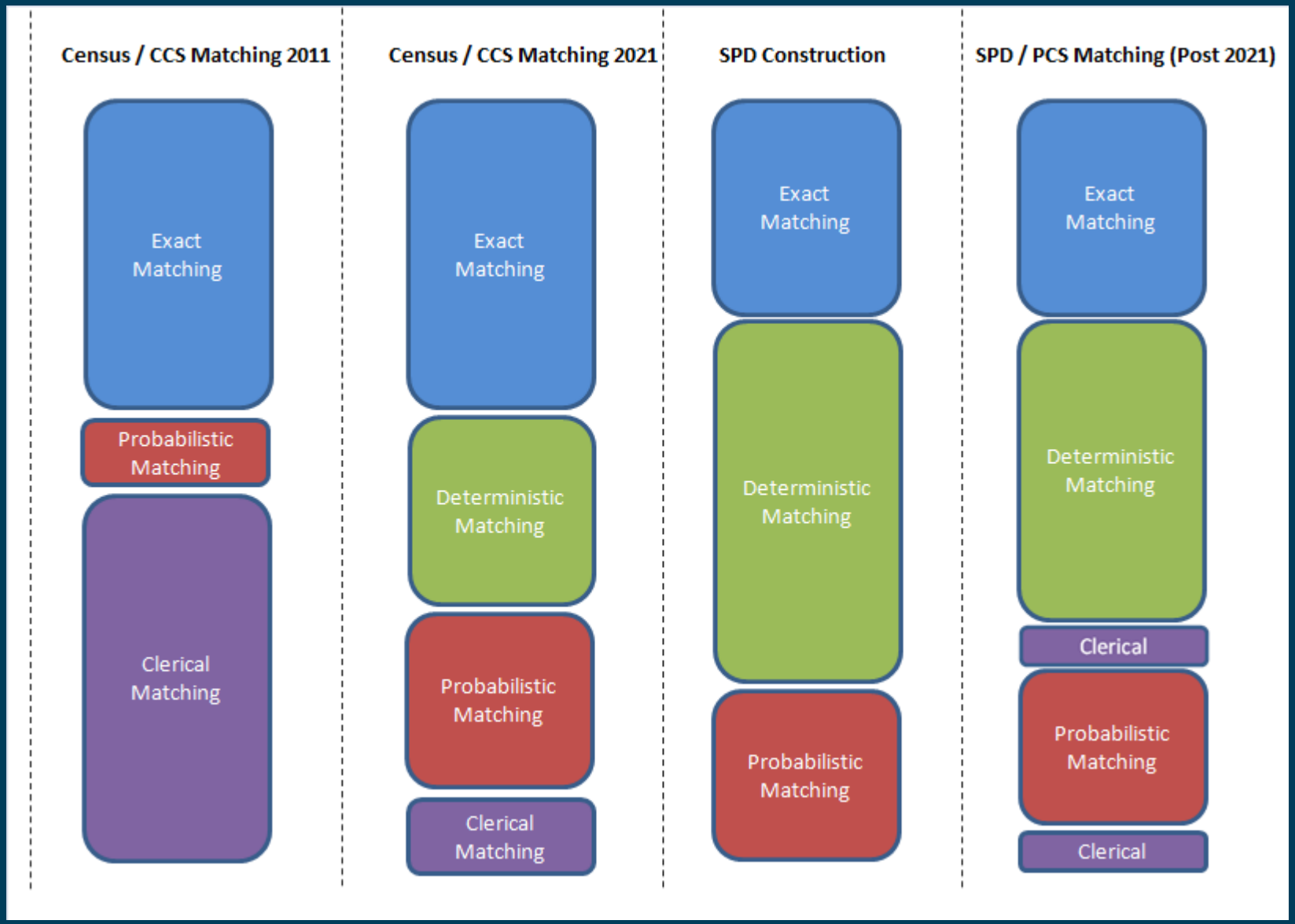
- Large volumes of administrative and census data can be linked efficiently using automated methods:
  - **BUT** trade off between quality and scale of matching

What did we learn from B2011? (2)

- A pseudonymisation approach restricts the implementation of some methods:
  - Probabilistic matching e.g. threshold setting
  - Clerical matching

- The quality of linkage delivered through auto-matching will not be high enough for:
  - Coverage adjustment (for example, linking records for dual system estimation)
  - Multivariate estimates of population characteristics

Matching Strategy for 2021



Issues for data linkage

- Multiple sources of data
- Large data sets
- Updates required – changing data over time
- Need to target clerical resource
- Master linked data file – viewed as ‘truth’
- Different matching requirements for different users
- Inconsistent clusters
- Improved speed (e.g. more automation)

Methodology current research areas

- Increasing the level of linkage automation, whilst maintaining stringent quality standards
- Duplicate pair methods
  - for threshold setting and
  - resolution of multiple links above the threshold
- Graph databases
  - for management of linked data

Current research areas (2)

- Sampling procedures to assess the quality of a data linkage project
- Which are the best string comparators?
  - for partial agreement e.g. in a Fellegi Sunter model
- Machine learning techniques for data linkage
- Optimisation of matchkeys

Increased automation of linkage

In 2011 Census to CCS matching:

- automatic match rate: 70% of all person matches
- clerical resource equivalent to 30 FTE / 30 weeks
- quality of <0.01% FP rate, <0.25% FN rate

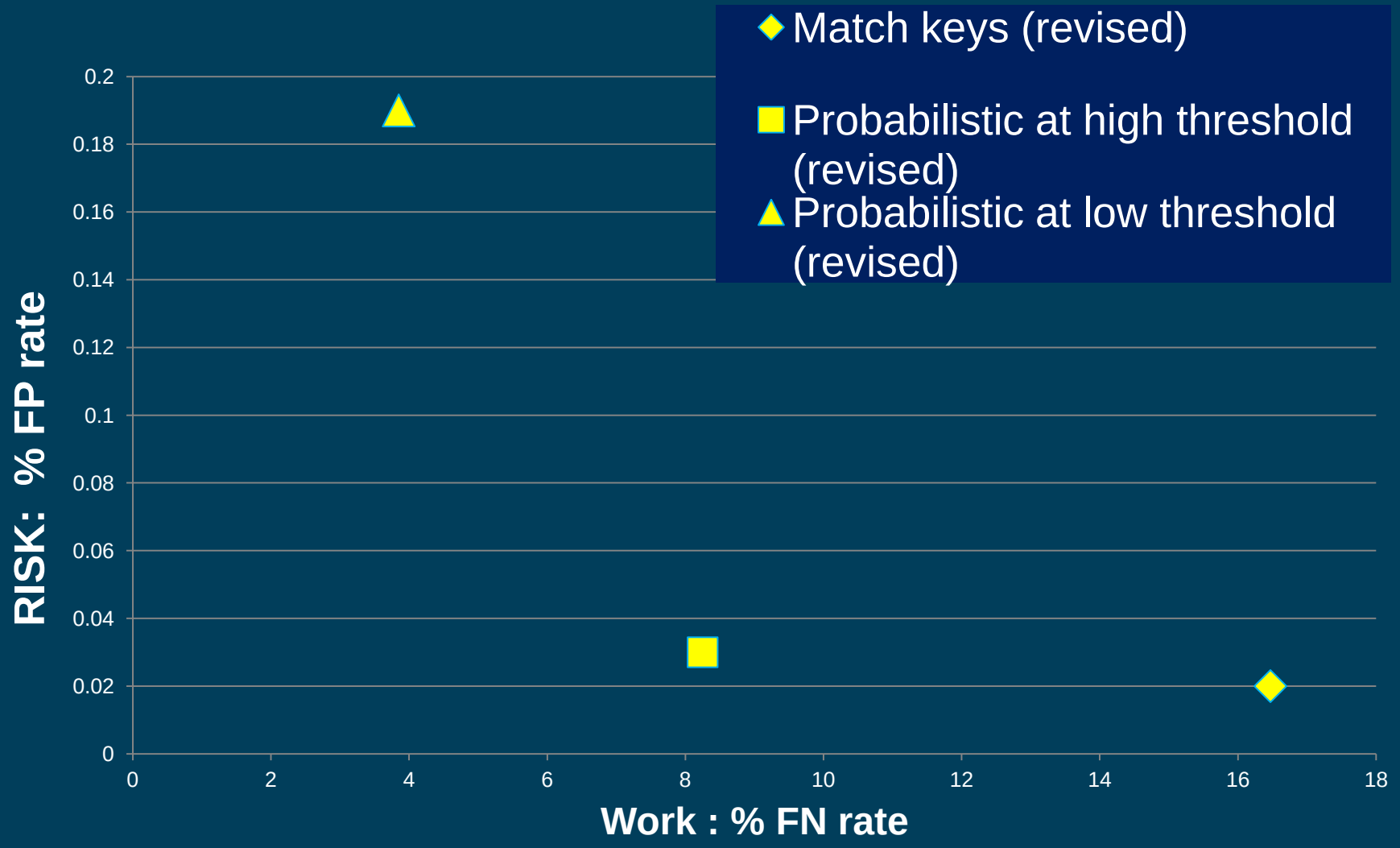
Can automated matching be increased in 2021 without incurring unacceptable numbers of false positives?

Using linked 2011 Census-CCS data for testing and comparison

Automated matching methods

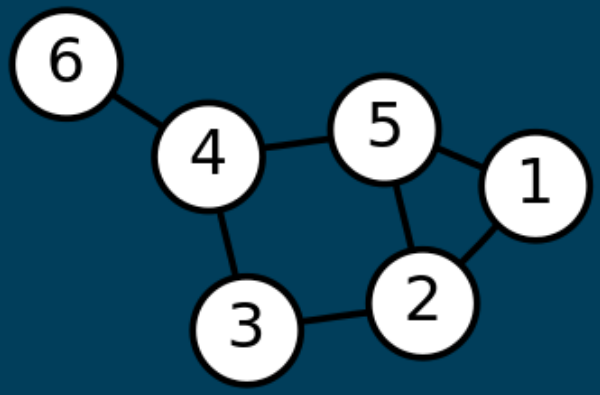
- Exact and deterministic matching
  - Matchkeys
- Probabilistic matching
  - Fellegi Sunter
  - Using partial agreement on names
  - Automatic de-duplication of links above threshold

Automated matching results



What is a graph database?

- A graph database stores data points and also the relationships between the data.



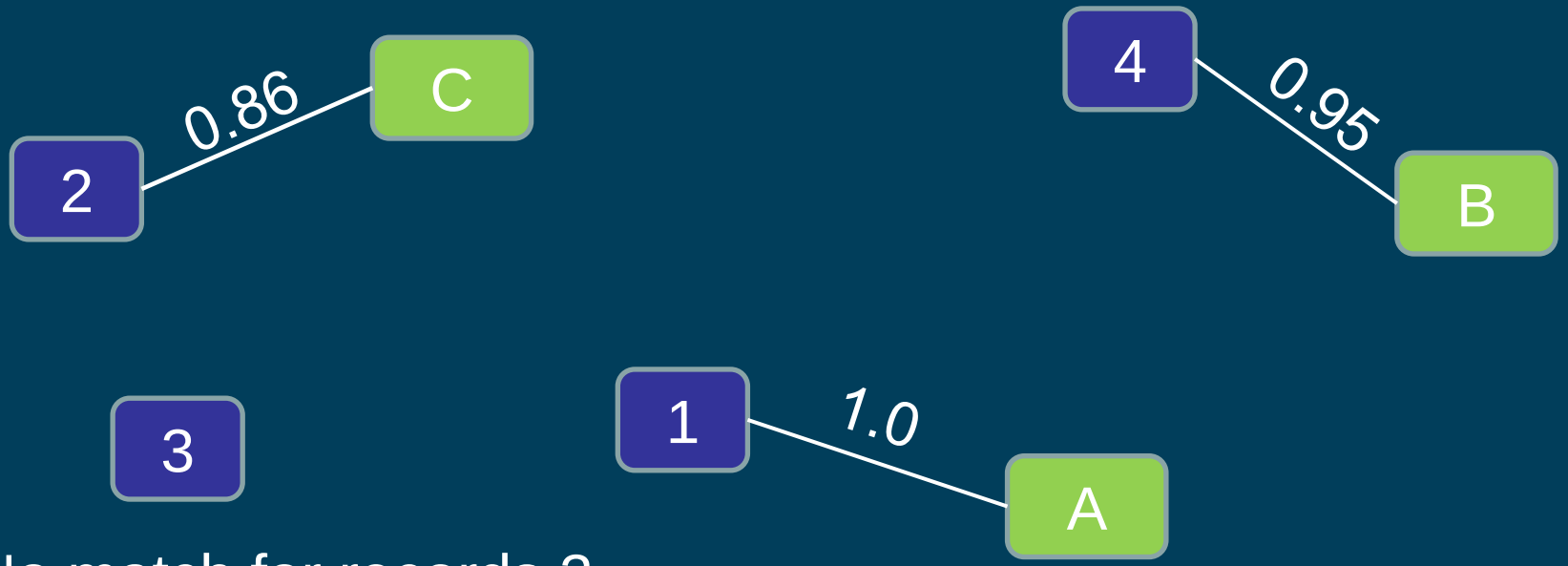
$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Adjacency matrix

- A graph can also be represented as a matrix.
- Graph format storage in certain use cases is more efficient than other forms of storage

How can you use a graph to model linked data?

- Blue nodes - Records in dataset 1
- Green nodes – records in dataset 2



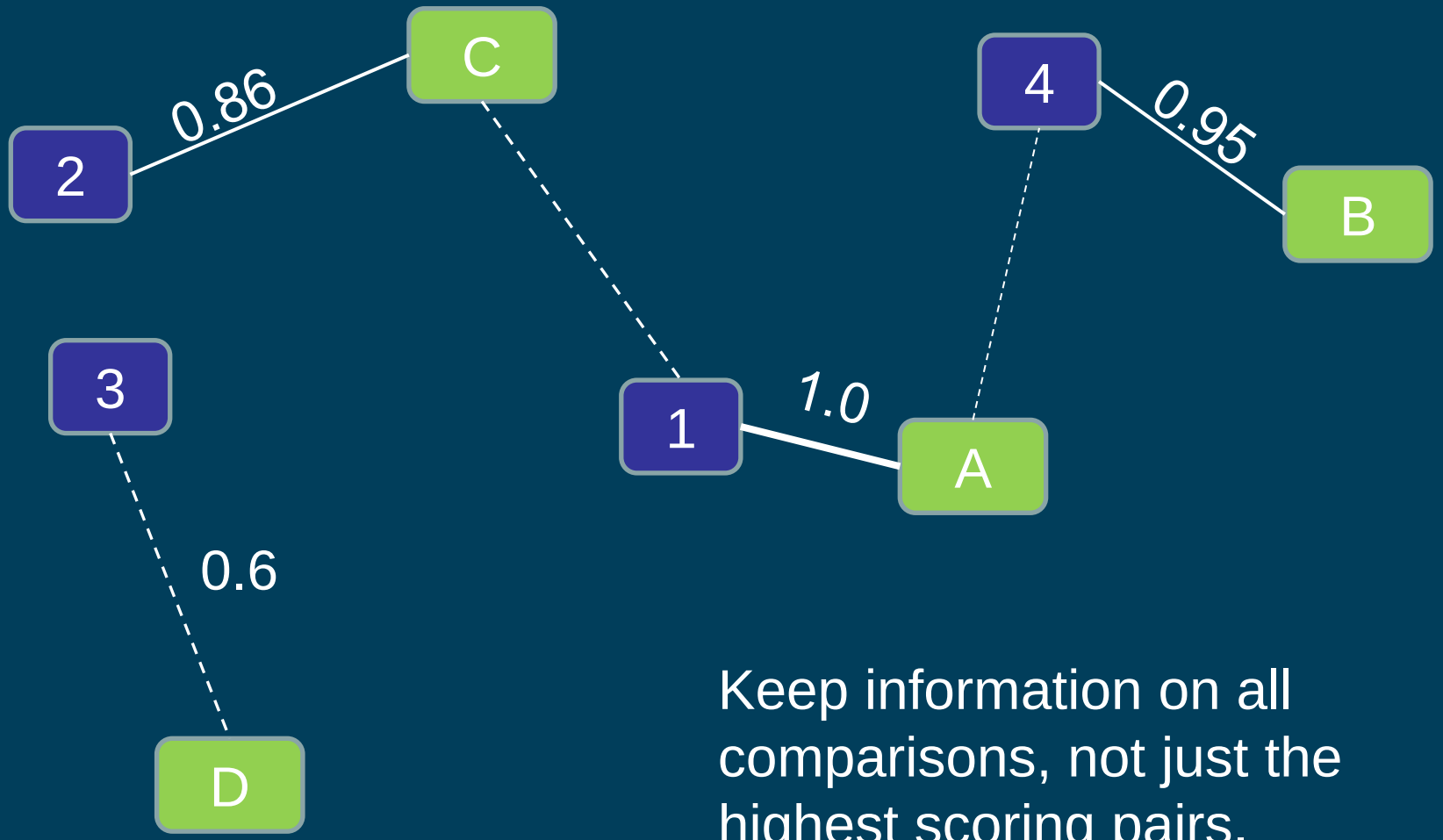
No match for records 3 & D

D

Numbers on the edges represent match score, where 1 is an exact match.

Graphical representation of linked data

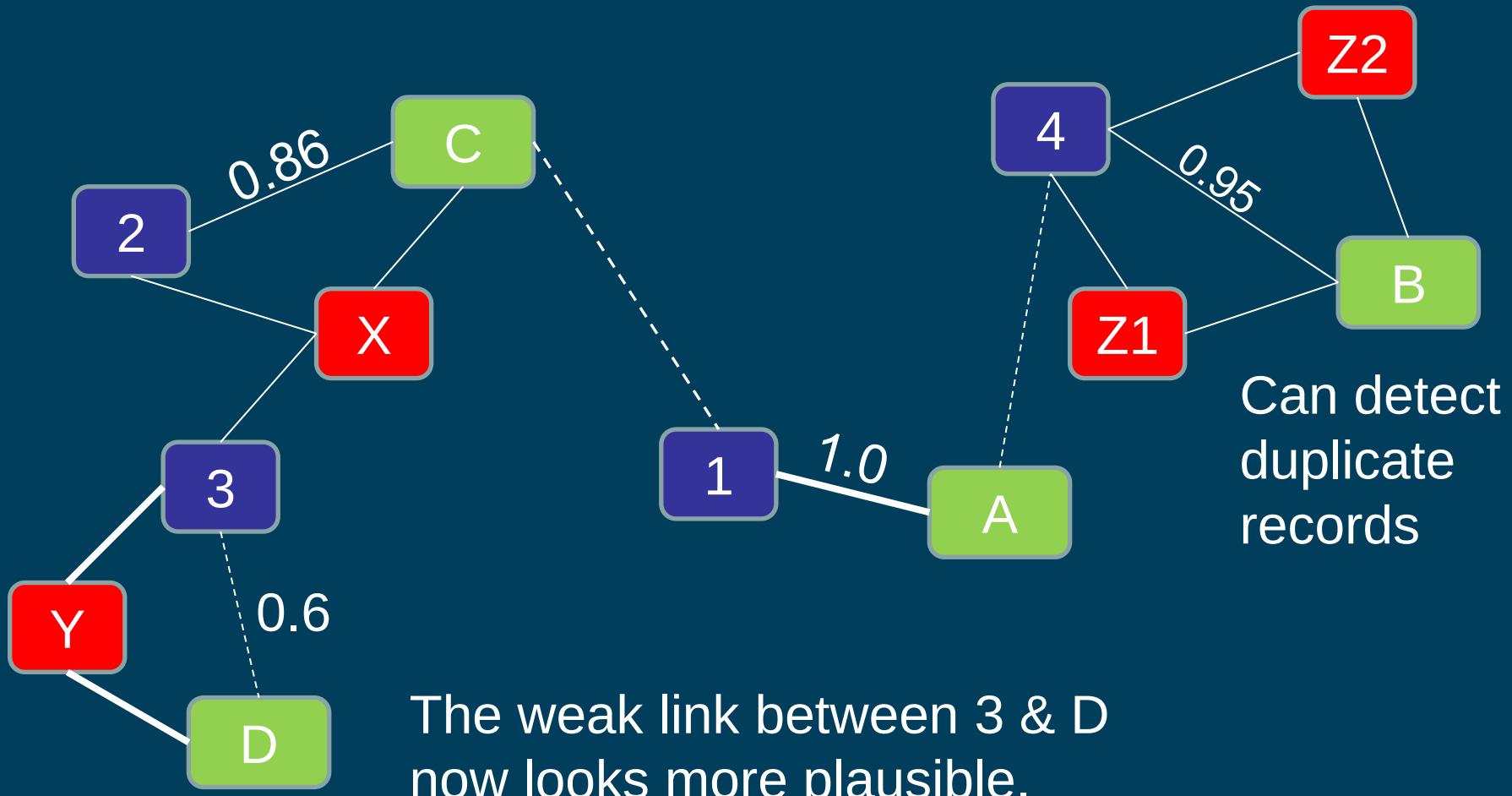
- Add in data on weaker links
- Model strength of linkage score



Keep information on all comparisons, not just the highest scoring pairs.

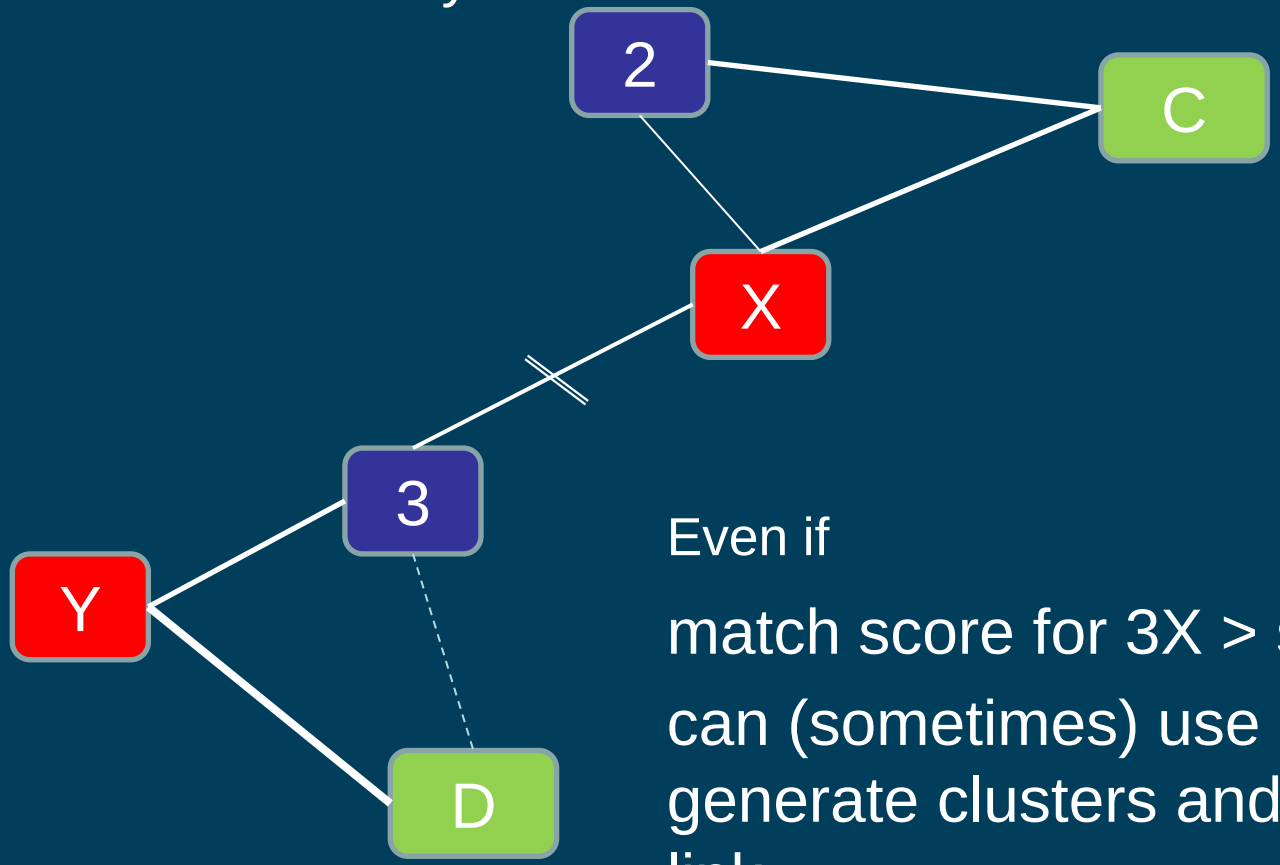
Graphical representation of linked data

- Add a third data source



Graphical representation of linked data

- AUTOMATIC resolution of some clusters of records relating to the same entity



Even if

match score for 3X > score for 2X,

can (sometimes) use cluster metrics to generate clusters and break the 3-X link.

AUTOMATICALLY

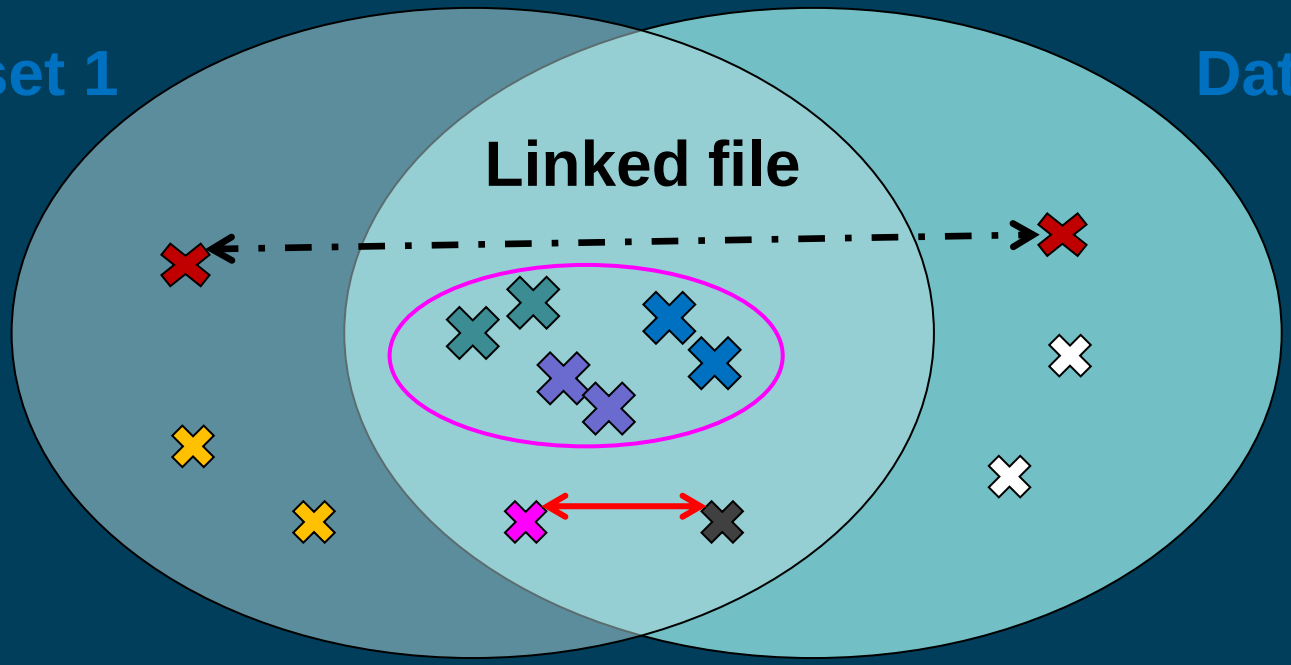
Graph databases (progress)

- Still early stages
- Learning about technology
- Improving linkage quality
- Next step is to use real data - lots of it!
  - Larger data sets (whole population)
  - More data sources

Quality of linkage

- After we have matched two datasets, we need to assess the quality of linkage
- Errors in matching can impact subsequent analyses, especially if there is a bias in matching errors towards certain groups of people
- Informing data users of estimated linkage error enables them to adjust their analysis methods to account for this

How good is our linking?



Correct links

Quality depends on:

1. Missed matches
2. Incorrect links

Sampling to estimate linkage quality (1)

- Linkage is often undertaken in stages

  e.g. exact, deterministic, probabilistic.

- Estimate level of incorrect links (false positives) overall, by stage
  - (& other factors, e.g. age, sex, LA)
- Estimate level of missed matches (false negatives) overall, by linkage probability score (& other factors?)
- How many records do we need to assess clerically?

Sampling for precision and recall (2)

- Sampling approach for clerical checks to estimate false positives and false negatives?
  - Used gold-standard matched data from the 2011 Census to evaluate different approaches.
  - Can stratification improve quality and reduce costs?
- Optimum sampling strategy depends on:
  1. availability of prior information on quality
  2. What level of detail is required in the estimates

Issues for data linkage

- Multiple sources of data
- Large data sets
- Updates required – changing data over time
- Need to target clerical resource
- Master linked data file – viewed as ‘truth’
- Different matching requirements for different users
- Inconsistent clusters
- Improved speed (e.g. more automation)