Workshop 6

COMP20008 Elements of Data Processing

Learning outcomes

By the end of this class, you should be able to:

- explain how record comparisons are used to perform data linkage and duplicate detection
- explain the impact of blocking on accuracy and efficiency of data linkage
- produce advanced visualisations in Python: scatter matrix plots and parallel coordinates plots

Q1: Data linkage

What is the maximum number of matching record pairs? m.

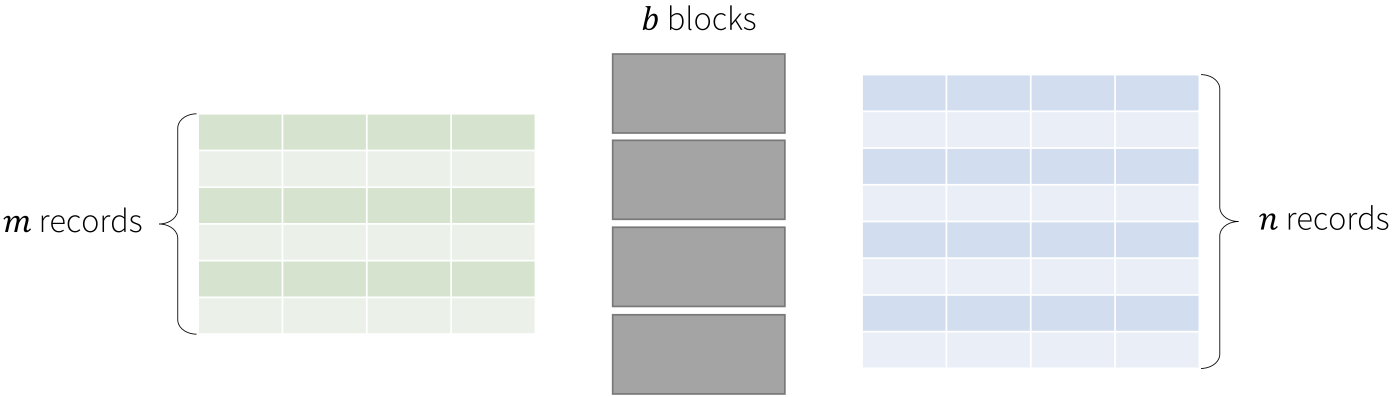
How many non-matching record pairs are there in this case? $nm - m$

Q1: Data linkage

What is the maximum number of matching record pairs? m

How many non-matching record pairs are there in this case? $m(n-1)$

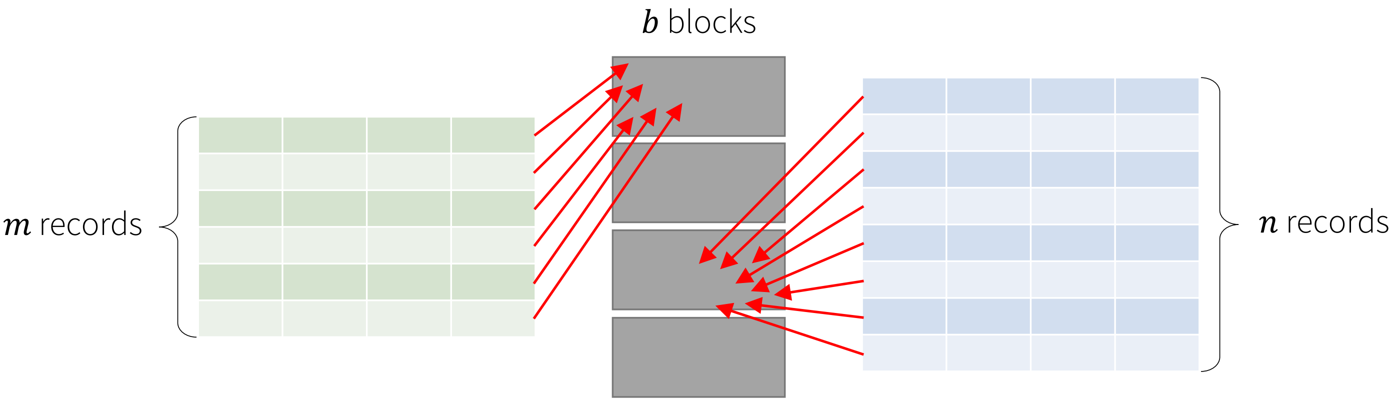
Q1: Data linkage



What's the *smallest* number of comparisons we could expect to make using blocking?

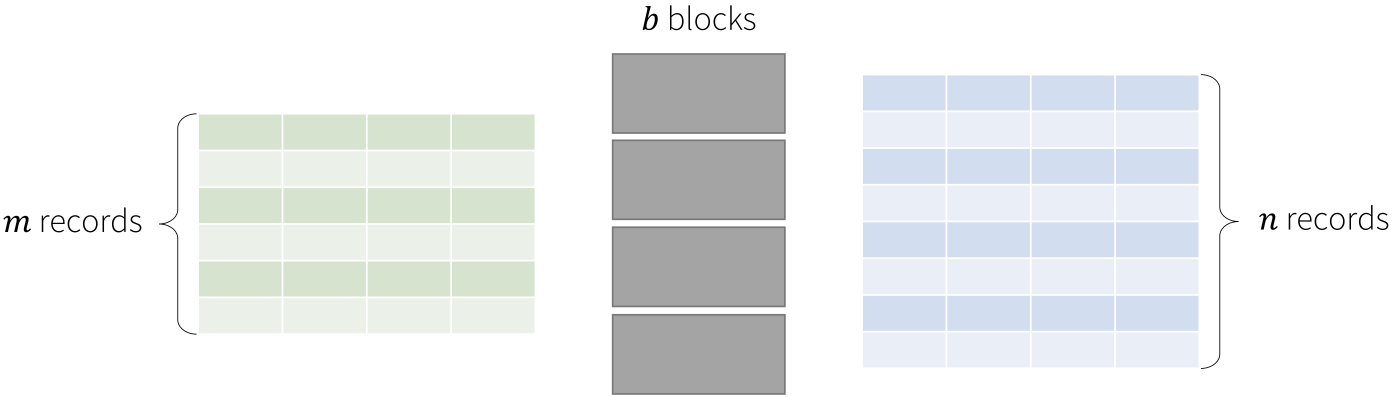
0.

Q1: Data linkage



What's the *smallest* number of comparisons we could expect to make using blocking? None

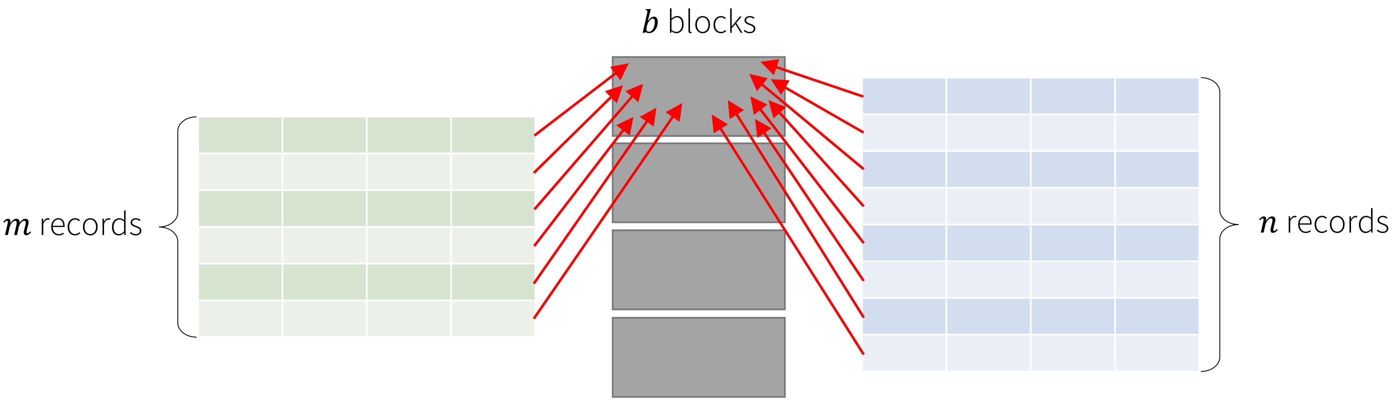
Q1: Data linkage



What's the *largest* number of comparisons we could expect to make using blocking?

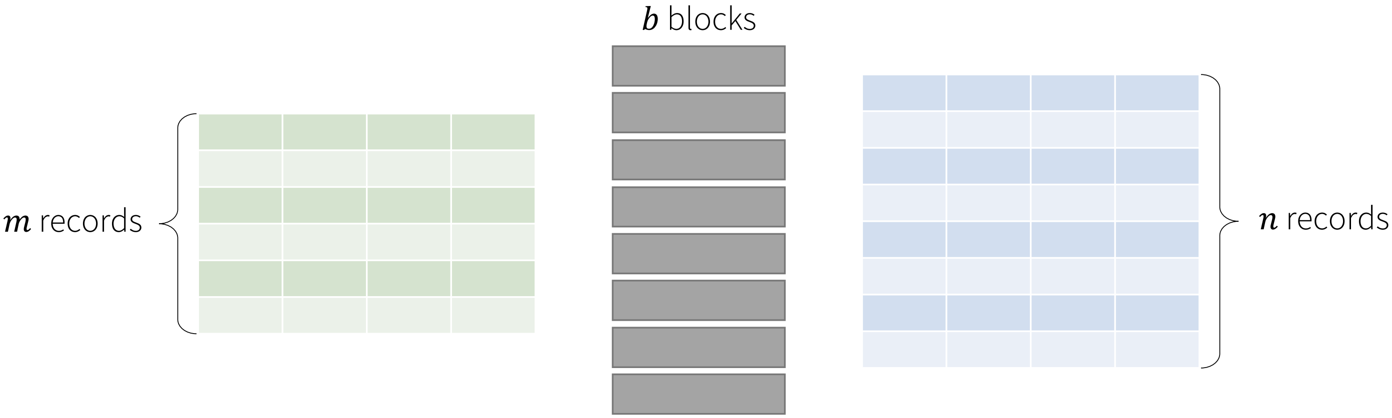
mn

Q1: Data linkage



What's the *largest* number of comparisons we could expect to make using blocking? $n \times m$

Q1: Data linkage

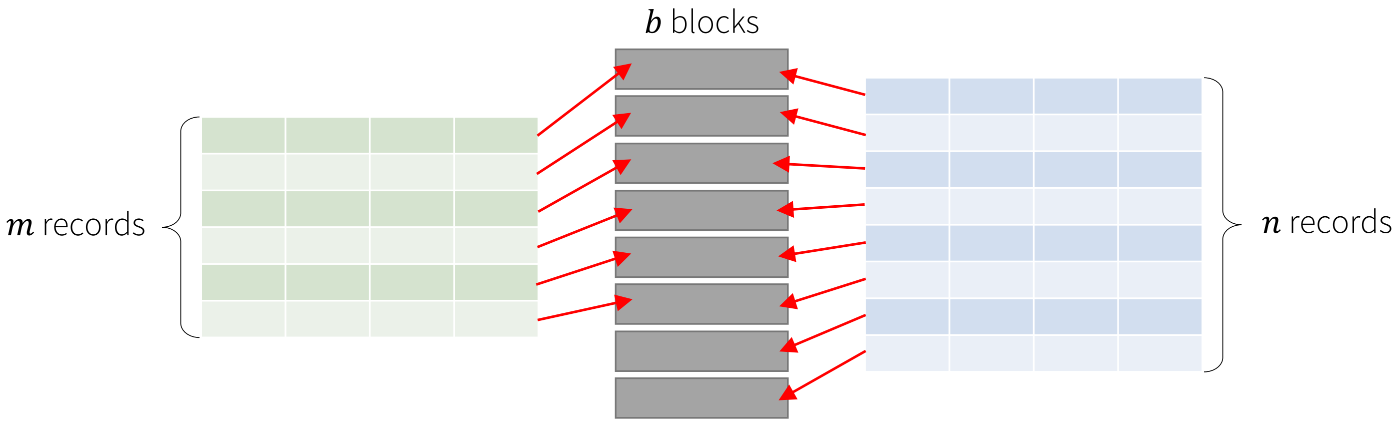


Suppose there are m record matches. What's the *smallest* number of comparisons needed to return all matches, and how many blocks would there be?

m

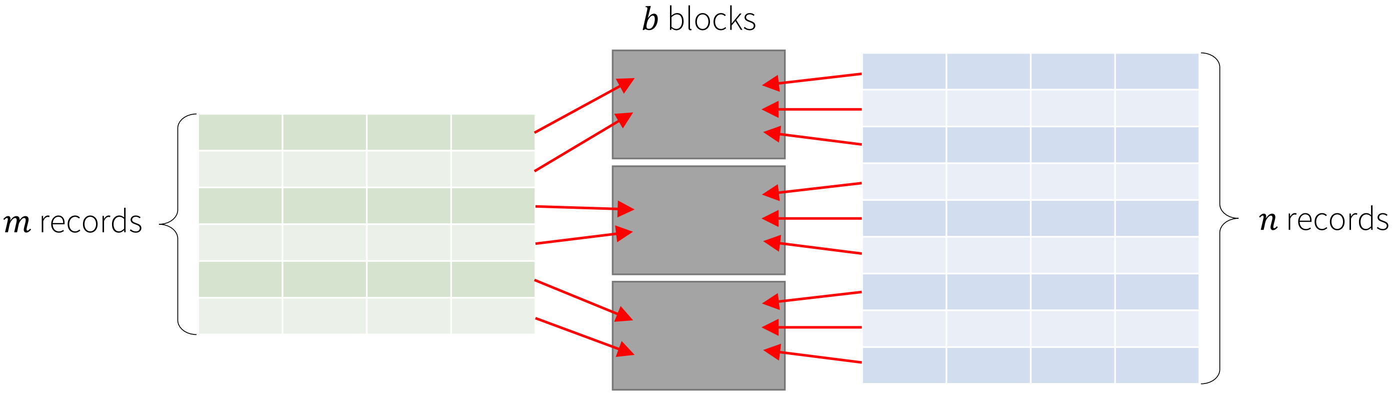
$m+1$. to n block.

Q1: Data linkage



Suppose there are m record matches. What's the *smallest* number of comparisons needed to return all matches, and how many blocks would there be? m comparisons; $m + 1$ to n blocks

Q1: Data linkage



How many comparisons are needed if the records are allocated evenly to the blocks?

$\frac{m}{b} \times \frac{n}{b} \times b.$

Q1: Data linkage

How many comparisons are needed if the records are allocated evenly to the blocks?

- Each block has $\frac{m}{b}$ records from the green data set and $\frac{n}{b}$ records from the blue data set
- Number of comparisons within one block: $\frac{m}{b} \times \frac{n}{b} = \frac{mn}{b^2}$
- Total across all blocks: $\frac{mn}{b^2} \times b = \frac{mn}{b}$

Q1: Data linkage

What's the advantage of using a large number of blocks?

efficiency↑ reduce useless comparison

What about a small number of blocks?

precision.

How would your answers change if records are allocated to *multiple blocks* and the block allocation is *uneven*?

block size matter, extreme case block n

with all in one block.

Q1: Data linkage

What's the advantage of using a large number of blocks?

Fewer comparisons ⇒ more efficient ✓.

What about a small number of blocks?

More comparisons ⇒ typically more accurate ✓.

How would your answers change if records are allocated to *multiple blocks* and the block allocation is *uneven*?

Conclusions don't change, except now the block *sizes* determine the number of comparisons

Q1: Data linkage

What's the advantage of allowing records to be allocated to multiple blocks?

precision .

Q1: Data linkage

What's the advantage of allowing records to be allocated to multiple blocks?

- Restricting to a single block is often too inflexible: end up missing matches.
- Consider the movie records example from lectures.
  The release date of a movie may vary across markets. This means we will fail to detect some matches from different markets if we block on release_year. To fix this, we can allocate records to multiple blocks: e.g. release_year and release_year + 1.

Q2: Duplicate detection vs. linkage

Duplicate detection:

- Identify records within a single data set that refer to the same entity
- Can cast as linking a data set with itself

Bill	Joy	US	1954
Sophie	Wilson	UK	1957
Stephen	Cook	US	1939
Radia	Perlman	US	1951
Barbara	Liskov	US	1939
William	Joy	US	1954

Q2: Duplicate detection vs. linkage

Data linkage:

- Identify records across a pair of data sets that refer to the same entity
- Often enforce a 1-to-1 constraint

Bill	Joy	US	1954
Sophie	Wilson	UK	1957
Stephen	Cook	US	1939
Radia	Perlman	US	1951
Barbara	Liskov	US	1939
Donald	Knuth	US	1938

Grace	Hopper	USA	1906
Alan	Turing	UK	1912
Frances	Allen	USA	1932
Peter	Naur	Denmark	1928
William	Joy	USA	1954

Q3: Evaluation of data linkage

Why might two records be incorrectly linked by a data linkage system?

Q3: Evaluation of data linkage

Why might two records be incorrectly linked by a data linkage system?

Two reasons:

1. If there are not enough informative attributes, some records may match exactly when they actually refer to different entities. Linkage is hopeless in this case.
2. System may not be strict enough when considering fuzzy matches. For example, linking a father and son that share the same name and address, but different date of birth.

Q3: Evaluation of data linkage

Evaluation: assess performance by comparing the predicted matches (links) from the system to human-verified truth

Category	Abbr	Description
False positive	FP	Pair of records **linked** that **don’t match**
False negative	FN	Pair of records **not linked** that **do match**
True positive	TP	Pair of records **linked** that **do match**
True negative	TN	Pair of records **not linked** that **don’t match**

What are the relative sizes of these categories?

Q3: Evaluation of data linkage

What are the relative sizes of these categories?

- They should add up to the total number of pairs $\boldsymbol{m} \times \boldsymbol{n}$
- Often represent in a confusion matrix

		Truth	
		Match	Non-match
Predicted	Link	TP (small)	FP (small)
	Non-link	FN (small)	TN (large)

Q3: Evaluation of data linkage

When would minimizing FP be more important?

When would minimizing FN be more important?

Q3: Evaluation of data linkage

When would minimizing FP be more important?

- Cautious about linking: want predicted matches to be correct, even if some matches are missed
- E.g. sending debt notices by linking social security and tax data.

When would minimizing FN be more important?

- Lenient about linking: want to find all matches, even if some predicted matches are incorrect
- E.g. linking child protection data across states. Want to find as many matches as possible. Can use manual review to deal with FP.

Q4: Advanced visualisation

Let’s move to JupyterLab