COMP20008
Elements of data processing

Semester 1 2020

Lecture 7: Data (Record) Linkage

Today

- What is data linkage, when and why is it needed and by whom?
- What are some challenges?
  - How to define similarity
  - Scalability
  - How to merge/group records
- Thanks to
  - Dr Ben Rubinstein for use of lecture materials on movies example

Data linkage: what is it?

- We collect information about entities such as people, products, images, songs, …
- Combining, grouping, matching electronic records of the same real-word entities.
- Two hospitals H1 and H2 wants to know if same patients visited both hospitals.

PatTbl

PatientID	Name	DOB	Age	Gender	StreetAddress	Suburb	Postcode
P1273489	John Smith	8/10/1960	51	M	8/42 Miller Street	Melbourne	3011
Q6549234-2	Mick Meyer	30/01/1948	63	M	10 Port Road	Ferny Grove	7004
P7693427-8	Joanna Smith	12/11/1984	27	F	76 George Crest	Sydeny	2020

AdmittedPatients

PID	Surname	GivenName	BirthDate	Sex	AID
25198	Smith	Jo Anna	19841112	1	A347
55642	Smith	John W.	19601008	0	A135
15907	Meier	Michael	19480101	0	A810
99801	Meyer	Mike	19790320	0	A135

Addresses

AID	Street	Location
A135	42 Miller St	3000 Melbourne
A347	16 George Crs	2000 Sydney
A810	PO Box 553	7000 Brisbane

Example from Data Matching book by Christen

Applications: security

Match data about people scheduled to fly to Australia by plane, with information across different databases, to identify high risk passengers before boarding. Databases with information such as

- Previous visits to Australia
- Previous visa applications/cancellations
- Crime databases …

A famous senator Sen. Ted Kennedy told the Senate Judiciary Committee in 2004 that he had been stopped and interrogated on at least five occasions as he attempted to board flights at several different airports. A Bush administration official explained to the Washington Post that Kennedy had been held up because the name "T. Kennedy" had become a popular pseudonym among terror suspects.

http://edition.cnn.com/2015/12/07/politics/no-fly-mistakes-cat-stevens-ted-kennedy-john-lewis/

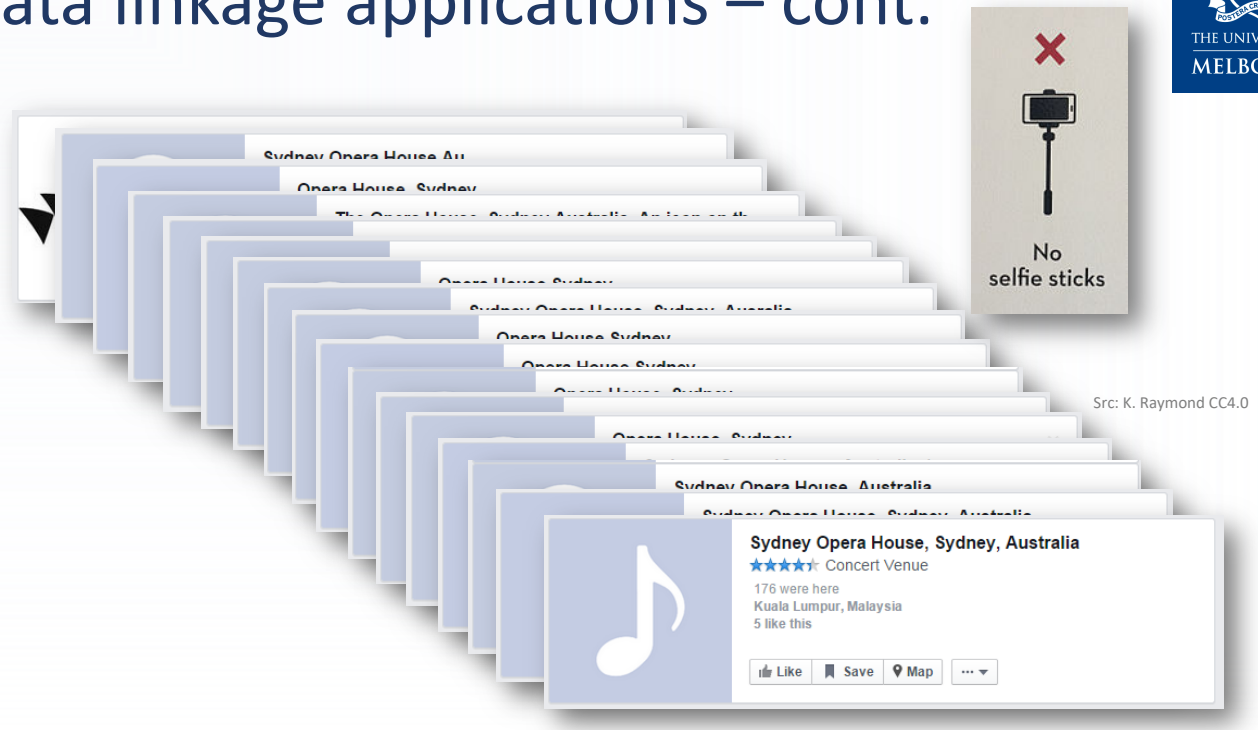
Applications: business

- Two businesses collaborate with each other for a marketing campaign. Need a combined database of individuals to target
- Bob moves into a new home and wishes to be connected to electricity provider. For verification, provider matches the address Bob supplies against its “master” list of street addresses. Not always reliable!
- Online shopping comparison
  - Is product X in Store A the same as product Y in Store B?
  - www.shopbot.com.au

Facebook likes

Rotten Tomato

Data linkage applications – cont.



Src: K. Raymond CC4.0

Centrelink: “Robo-debt” collection

Sydney Morning Herald 10/4/17

'Not reasonable or fair' Ombudsman slams Centrelink's robo-debt scheme

- Data matching using Centrelink data and Tax office data
- System checks for “discrepancies” in income
- Example data matching issue
  - Welfare recipient reports to Centrelink working for a company with its trading name. Tax office records show a different registered company name.
  - Failure to match between the two names triggered conclusion that some income was not being declared
  - Automated notice …

Examples of problematic matching

- Centrelink

Centrelink Income:
Jane Doe

May'16: Maccas $7,000
June'16: Maccas $4,000

ATO Income:
Jane Doe

2015-16: McDonald's $11,000

Discrepancy detected – potential undeclared income
⇒ **Automated** process triggered -> *letter to Jane Doe*

- *A famous senator Sen. Ted Kennedy told the Senate Judiciary Committee in 2004 that he had been stopped and interrogated on at least five occasions as he attempted to board flights at several different airports. A Bush administration official explained to the Washington Post that Kennedy had been held up because the name "T. Kennedy" had become a popular pseudonym among terror suspects.*

Record linkage – terminology

Combine related/equivalent records across sources (sometimes within a single source)

Studied across communities –different terminology

- Statistics: Record linkage [Dunn'46]
- Databases: Entity resolution, data integration, deduplication
- Natural Language Processing: coreference resolution, named-entity recognition

…meaning and scope varies

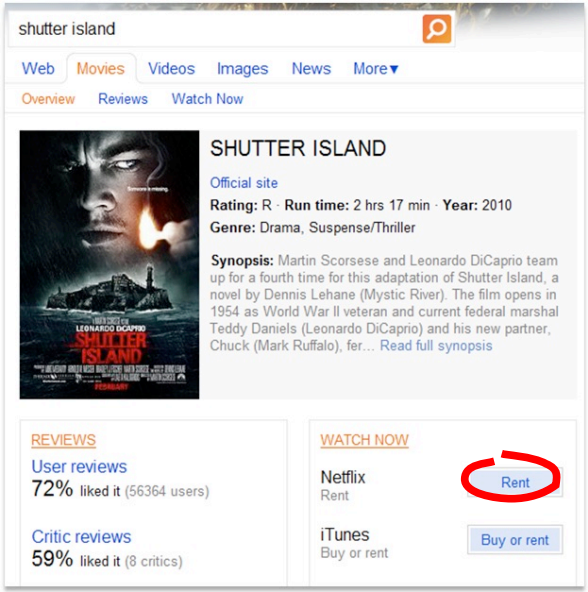
Example: Bing's "Movies Vertical"

Bing Movies adding
entity actions to entity cards

Need to link movie records from Bing and Netflix

Easy problem only if there's an attribute with unique value per record

- If there's a "**key**" then use "**database join**"



Not so fast! The linkage challenge

No keys

Noisy values

Missing attributes

Usually remove diacritics (why?)
déjà → deja

19. **Deja Vu** (2006)

7.0/10

After a ferry is bombed in New Orleans, an A.T.F. agent joins a unique investigation using experimental surveillance technology to find the bomber, but soon finds himself becoming obsessed with one of the victims. (126 mins.)

Director: Tony Scott
Stars: Denzel Washington, Paula Patton, Jim Caviezel, Val Kilmer
Add to Watchlist

20. **Déjà Vu** (1997)

7.0/10

L.A. shop owner Dana and Englishman Sean meet and fall in love at first sight, but Sean is married and Dana is to marry her business partner Alex. (117 mins.)

Director: Henry Jaglom
Stars: Victoria Foyt, Stephen Dillane, Vanessa Redgrave, Glynis Barber
Add to Watchlist

" This is just one of many " - wunderdunder

"Unrelated works with same titles" http://www.imdb.com/list/ls000397020/

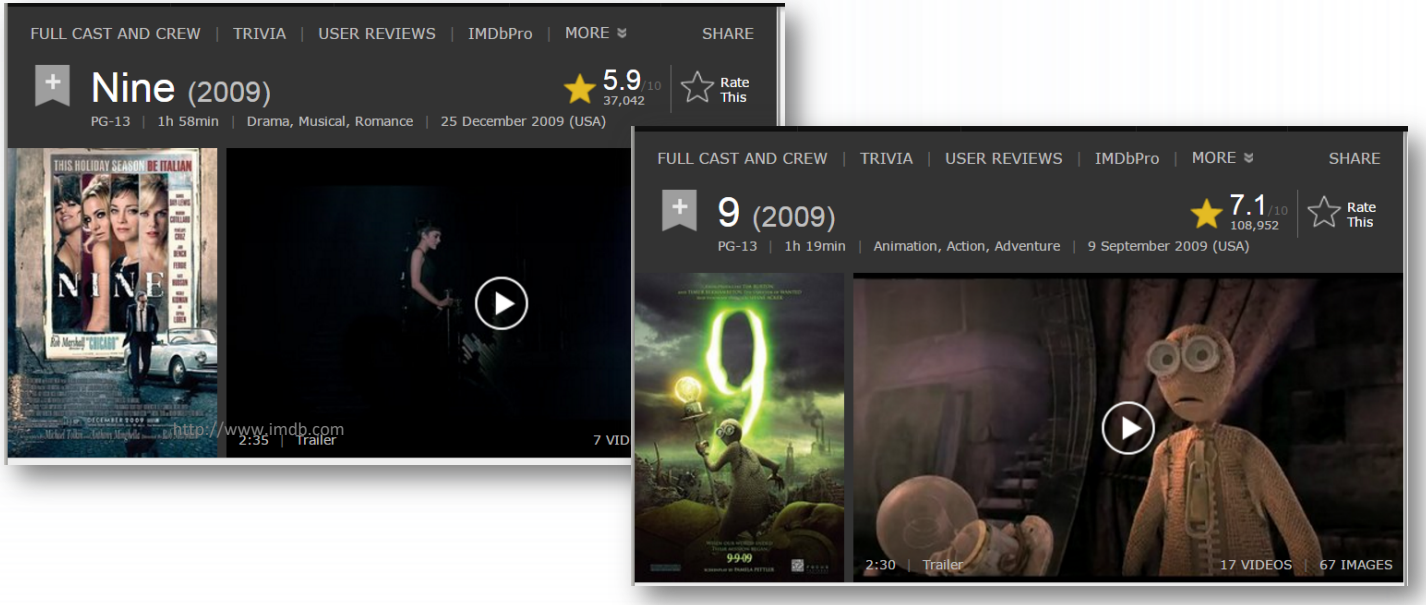
Concatenate “Title, Year”: surely a key?

"Deja Vu,2006" ←

"Deja Vu,1997" ←

“Unrelated works with same titles” http://www.imdb.com/list/ls000397020/

Concatenate “Title, Year”: surely a key?



Pre-processing

	Title	Year	Directors	Cast	Runtime	...
	come fly with me	2004	peter kagan	michael buble	63	...
	michael jordan come fly with me	1989		michael jordan, jay thomas	42	...

Prep → Block → Score → Match → Merge

Simple pre-processing to clean records

Pairwise comparison

- Need to compare two records to determine if they are to be linked
- Calculate similarity between a pair of records

```
If similar:
   link pair
```

- High complexity – we will need to compare/"score" many pairs of records for similarity.

Complexity – pairwise comparison

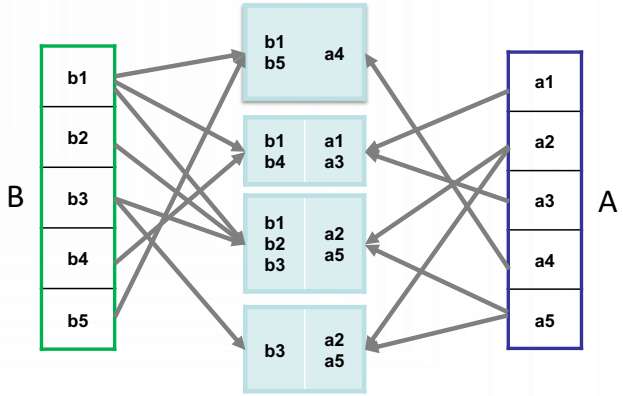
- Two databases A (m records) and B (n records)
  - How many comparisons? $m \times n$
  - One MSFT problem had **1b** records
- Complexity $O(n^2)$
  - If $m == n == 1b$ (10^9) records
  - $10^9 \times 10^9$ number of comparisons;
  - It will take **a long time** to compute 10^{18} pairs
- Naïve whole-database all-pairs comparison doesn't scale.

Can we do better?

Blocking – scalable record linkage

- Maps complex records to simple values (blocks).
- The simple values are **blocking-keys**
- Same blocking methods applied to both A and B.

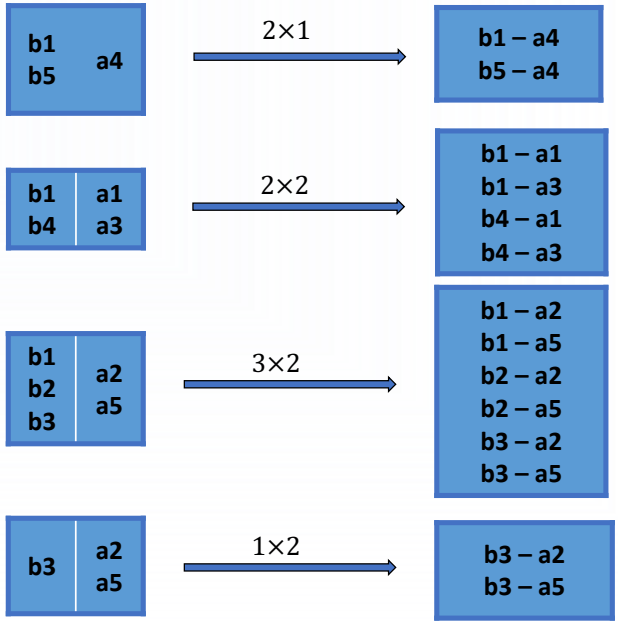
Each record is allocated to **one or more of the blocks**



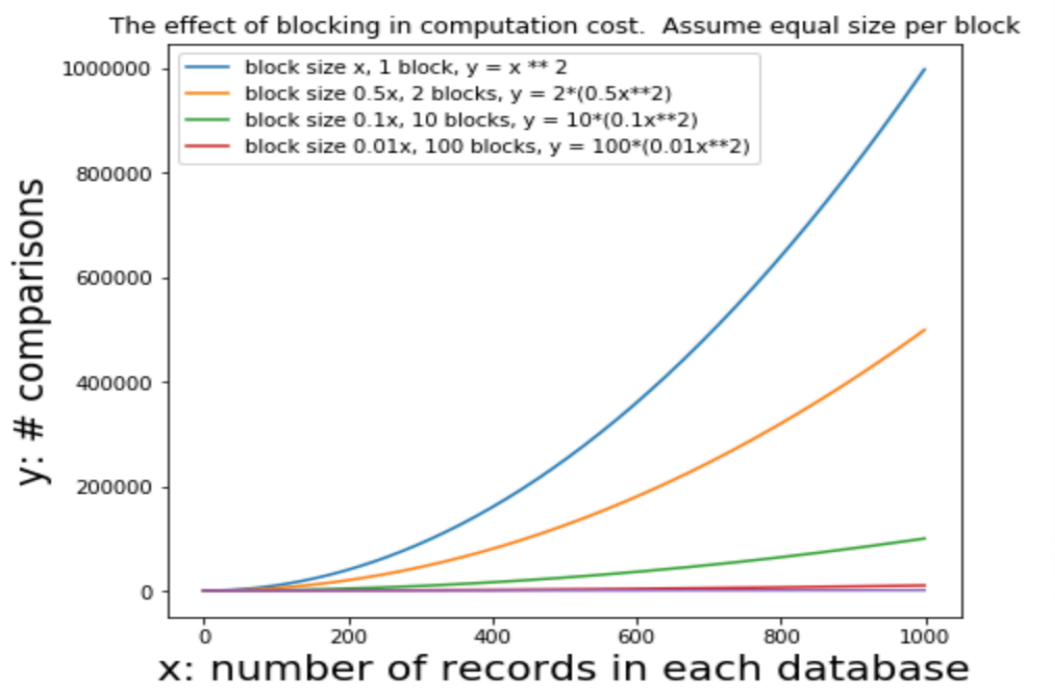
Prep | Block | Score | Match | Merge

Blocking – cont.

- Record-pairs assigned to the same blocks are potential matches.
- Within block pairwise comparisons only
- For example, b5 is only compared with a4, and no other records in A

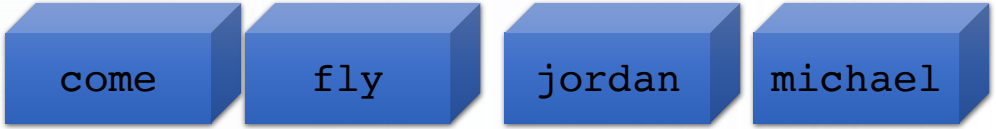


Scalable record linkage with blocking



Blocking example

	Title	...
	come fly with me	...
	michael jordan come fly with me	...



Prep | Block | Score | Match | Merge

Represent complex records as simple values (blocks);
Only score records with simple value (block) in common.

Blocking example

A movie record: <“ghost busters”, 2016>

- On one function: release year, result in one block
  - *2016*
- On one function: concatenated values of **release year** and **one title-word:** results in two blocks
  - *2016 ghost*
  - *2016 busters*
- Movies with a difference in release year by one; three functions:
  - **release year**,
  - **release year + 1**,
  - **release year – 1;** results in three blocks:
  - *2015*, *2016*, and *2017*
  - *Redundant blocks, why?*

Common blocking methods for strings

- Token/word (‘come fly with me’)

  `{come, fly, with, me}`
- Phrases/n-words (‘come fly with me’, n=2)

  `{’_ come’, ’come fly’, ’fly with’, ‘with me’, ‘me _’}`
- N-grams (‘Jamie’, n=2)
  - `{_j, ja, am, mi, ie, e_}`
  - blocking-keys are the 2-grams; ‘jamie’ are in 6 blocks
- Prefix, suffix
- soundex
  - Pauline: P450, Paulina: P450,
  - Chris: C620
  - Kris: K620

Blocking methods– design

- Blocking methods should be efficient, e.g. hashing
  - What would be a bad blocking method?

- Trade-off between block sizes: true matches being missed vs computational efficiency
  - Can filter out large blocks
  - Blocks can overlap but avoid redundant blocks
  - Need to ensure that **recall** does not suffer

Measuring blocking method

Given the ground truth (matching record-pairs)

`False positives (fp)`: non-matching pairs in the same block.

`False negatives(fn)`: matching pairs never in the same block.

`True positives (tp)`: matching pairs in the same block.

`True negative (tn)`: non-matching pairs never in the same block.

Total number of pairs with no blocking: $m \times n$ ($== fp + fn + tp + tn$)

Blocking measures:

- `Pair-completeness (PC)`: $tp/(tp + fn)$
- `Reduction-ratio (RR)`: $(fp + tp)/(fp + fn + tp + tn)$

Assess pairwise similarity

Multiple scoring functions on different parts of the records

	Title	Year	Directors	Cast	Runtime	…
	come fly with me	2004	peter kagan	michael buble	63	…
	0.82	15	0	0	21	
	michael jordan come fly with me	1989		michael jordan, jay thomas	42	…

Prep | Block | **Score** | Match | Merge

Comparing two records : Assess their similarity

Scoring similarities – text

What is the similarity score between a text attribute between a pair of records?

Revision on text processing!

Exercise: set-based string similarity

- *Compute dice coefficient similarity between x and y using 2-grams.*
  - *x=james*
  - *y=jamie*

  - $S_x = \{$**`_j, ja, am,`** `me, es, s_`$\}$
  - $S_y = \{$**`_j, ja, am,`** `mi, ie, e_`$\}$
  - $|S_x \cap S_y| = 3, |S_x| = 6\ |S_y| = 6$
  - $sim_{dice}(S_x, S_y) = \frac{2\times|S_x \cap S_y|}{|S_x|+|S_y|} = \frac{2*3}{6+6} = 0.5$
- Why 2-grams? What about 1-gram or 10-grams?

Exercise: direct string similarity

- Edit-distance similarity

  Minimum number of character insertions, deletions, substitutions to go from s_1 to s_2

  - "valuation" –> "revolution"?

		v	a	l	u	a	t	i	o	n
r	e	v	o	l	u	-	t	i	o	n

  - $sim(x, y) = 1 - \frac{d(x,y)}{\max(|x|,|y|)} = 1 - \frac{4}{10} = 0.6$

- Jaro-Winkler similarity
  - Based on edit-distance
  - Favours matches at the start of the strings.

Scoring similarity

Combine the set of similarity scores → final score

	Title	...
	come fly with me	...
	michael jordan come fly with me	...

(0.82 , 15 , 0 , 0 , 21)

f: $\mathbf{R}^d$ → [0,1]

0.1

Need a good *f*

Prep | Block | Score | Match | Merge

Score record pairs for similarity

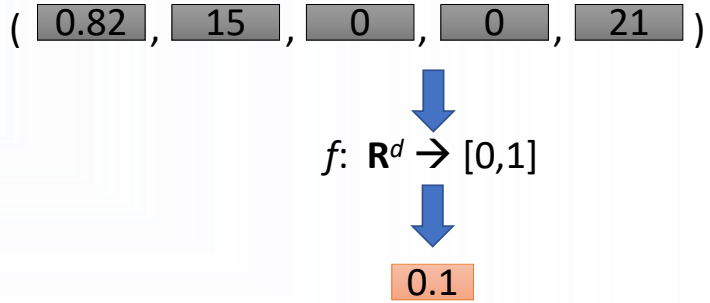
More on score combination

Idea 1: sum up feature scores

Idea 2: +similarities, -dissimilarities

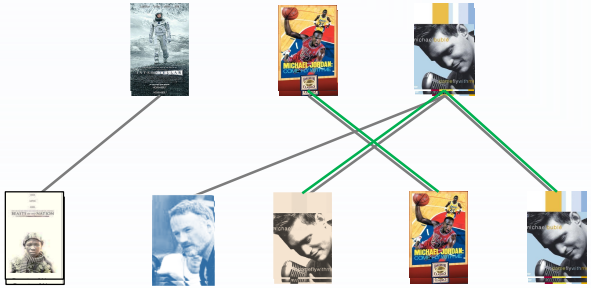
Idea 3: weighted sum

Idea 4: label examples of non/matching record pairs, train a classifier using **machine learning**



Will learn the weight

Match ‘sufficiently’ similar records



pairs compared

sufficiently similar pairs

Threshold θ

Match when final score > θ

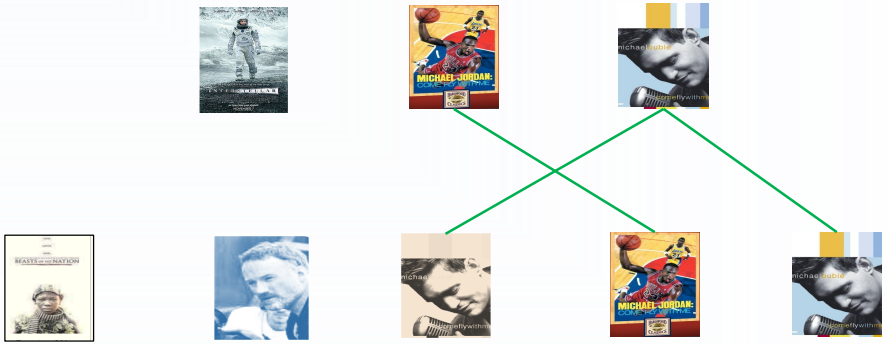
e.g.

threshold θ = 0.5

final score > 0.5

Prep | Block | Score | **Match** | Merge

Merge matched records



matched pairs

Prep | Block | Score | Match | Merge

Merge matched records

- Also needs to resolve conflicting attributes
- False positives and false negatives still exist

Evaluation of record linkage results

False positives (fp): # non-matched pairs that are incorrectly classified as matches.

False negatives(fn): # true matching pairs that are incorrectly classified as non-matches

True positives (tp): # true matching pairs that are classified as matches

True negative (tn): #non-matched pairs that are classified as non-matches

- Precision: $tp/(tp + fp)$
  Proportion of pairs classified as matches that are true matches.

- Recall: $tp/(tp + fn)$
  Proportion of true matching pairs that are classified as matches.

Summary

- ✓Understanding what the record linkage problem is
- ✓Ability to outline where record linkage is applied and why
- ✓Appreciation of why record linkage can be tricky
- ✓Can describe basic approaches to record linkage, such as the methodology of blocking
- ✓Can evaluate blocking methods and record linkage results.

Acknowledgements

- Lecture slides are based on presentation materials created by Ben Rubinstein