COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 13: Data linkage

Announcements

- Project Phase 1 marks will be released Friday at 7pm.
- Project Phase 2 was released on Monday 3rd September.
- Consultation sessions about Project Phase 2
  - Fri 14/09/2018 Room 09.02 Doug McDonell 1:00pm-2:00pm
  - Wed 19/09/2018 Room 07.02 Doug McDonell 10:30am-11:30am
- Phase 3 Oral Presentation Schedule is on the LMS
  - Deadline for changes is on Monday 24/09/2018

Today

- What is data linkage, when and why is it needed and by whom?
- What are some challenges?
  - How to define similarity
  - Efficiency
    - Need for blocking
- Thanks to
  - Ben Rubinstein for use of lecture materials on movies example

Data Linkage: What is it?

- Combining related/equivalent records across data sources
  - Information relating to the same entity (e.g. a person or place) is connected
  - E.g. Two hospitals H1 and H2 want to link the same patients

H1

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

H2

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

Data Linkage Applications: Health

- Match hospital data and death register data
  - Find mortality of certain diseases
- Match hospital data and job occupation data
  - Find correlations between occupation and disease susceptibility

Centrelink: "Robo-debt" collection

'Not reasonable or fair' Ombudsman slams Centrelink's robo-debt scheme Sydney Morning Herald 10/4/17

- 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 names two triggered conclusion that some income was not being declared
  - Automated notice …

Example of problematic (lack of) matching

Centrelink Income: Jane Doe

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

Tax office Income: Jane Doe

2015-16: McDonald's $11,000

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

Data matching act

- Centrelink didn't use any tax file numbers
  - Link Jane Doe using name, date of birth, historical addresses
- Consequently not subject to
  - *Data- matching Program (Assistance and Tax) Act 1990*
- Instead followed *voluntary Guidelines on Data Matching in Australian Government Administration*

Centrelink cont.

- Recent ombudsman's report
  - http://www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf
    - Centrelink launched a new online compliance intervention (OCI) system for raising and recovering debts.
    - The OCI matches the earnings recorded on a customer's Centrelink record with historical employer-reported income data from the Australian Taxation Office (ATO). Parts of the debt raising process previously done manually by compliance officers within DHS are now done using this automated process.

Applications: Bibliographic databases

- Research repositories contain millions of publications, each with one or more authors
- Government want to measure researcher and university productivity. E.g. Count the number of publications per academic
  - How do we know if the same person authored two different publications? E.g. Rui Zhang in
    - **Rui Zhang**, Raymond Chiong, Zbigniew Michalewicz, Pei-Chann Chang: Sustainable scheduling of manufacturing and transportation systems. European Journal of Operational Research 248(3): 741-743 (2016)
    - Chuanwen Li, Yu Gu, Jianzhong Qi, **Rui Zhang**, Ge Yu: A safe region based approach to moving KNN queries in obstructed space. Knowl. Inf. Syst. 45(2): 417-451 (2015)

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

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.

Applications: Security

- Identity matching: Applicant for a bank loan has their identity matched against trusted sources: voter registration lists, drivers licence database, ….

Applications: Business

- Two businesses collaborate with each other for a marketing campaign. Need a combined database of individuals to target
  - Need to identify if two or more records refer to the same individual
- Geospatial data
  - 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

Matching a database against itself

- Business wishes to carry out an advertising campaign.
  - Has a large database of customers
- The customer database changes over time, people move address, change their names.
- Duplicate records about individuals – business wishes to know if the same person appears more than once
- E.g. All the following are the same entity
  - Dr James Bailey, Department of Computing, Kings College London, james@dcs.kcl.ac.uk
  - Dr James Bailey Department of Computer Science, The University of Melbourne, jbailey@cs.mu.oz.au
  - Dr James Bailey, Department of Computer Science and Software Engineering, The University of Melbourne, jbailey@csse.unimelb.edu.au
  - Professor James Bailey, Department of Computing and Information Systems, The University of Melbourne, baileyj@unimelb.edu.au

Historical Studies

- Social scientists tracking life courses of individuals over decades.
  - Census in 1870
  - Census in 1880
  - Census in 1890
  - Census in 1900
  - ….
- Challenge
  - Addresses weren't standardised
  - Frequency distributions of names highly clustered
    - ``Not uncommon in the mid-nineteenth century England that more than 10% of the population had the given name 'John' and more than 10% of the population had the given name 'Mary' ''

Australian Census 2016

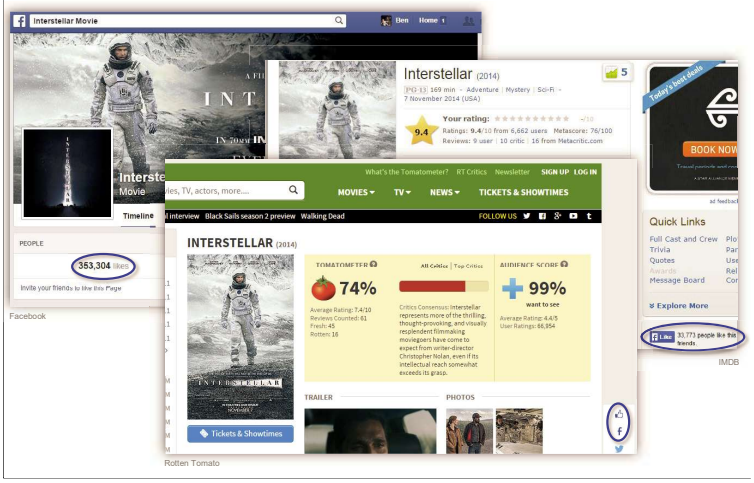
- Controversy last year over decision to retain people's names and addresses from census, for up to 4 years (rather than 18 months)
- From ABS website http://www.abs.gov.au/websitedbs/censushome.nsf/home/privacy

The benefits of retaining names and address in the Census are significant. Names and addresses will be used by the ABS to generate anonymous keys that can be used to combine existing data sets to create richer and more valuable statistics for Australia.

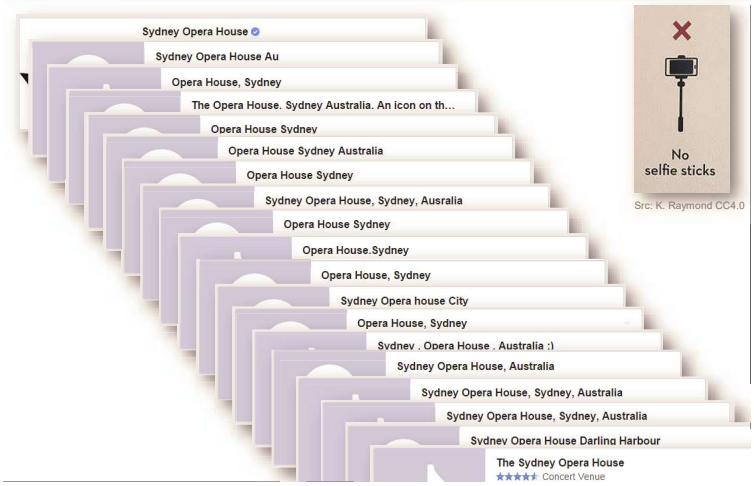
The new data sets, containing no names and addresses, will improve the lives of Australians by:

- *better informing decisions, policies and services in important areas like health, education, infrastructure and the economy*
- *enabling greater use of existing data and reducing the burden on individuals to provide data that is already available*
- *providing additional insights and more confidence in decisions, particularly for the most vulnerable and challenging policy areas.*

Facebook – Likes, Interests



Facebook – Places



Today's Learning Objectives

Understanding what the record linkage problem is

Ability to outline where record linkage is applied

Appreciation of why record linkage can be tricky

Can describe basic approaches to record linkage, such as the methodology of blocking

Challenges

- How to efficiently do linkage when matching two large databases
  - Blocking
- How to define similarity between records?
- How to maintain privacy when doing data linkage? (later lectures)
  - Why is privacy important?
  - An example method for privacy preserving linkage

Record Linkage: Brief History

Combine related/equivalent records across sources

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

Running Example: Building a Search Engine's "Movies Vertical"

Why: Bing Movies adding entity actions to entity cards

What: Need to link movie records from Bing and Netflix

How: 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

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

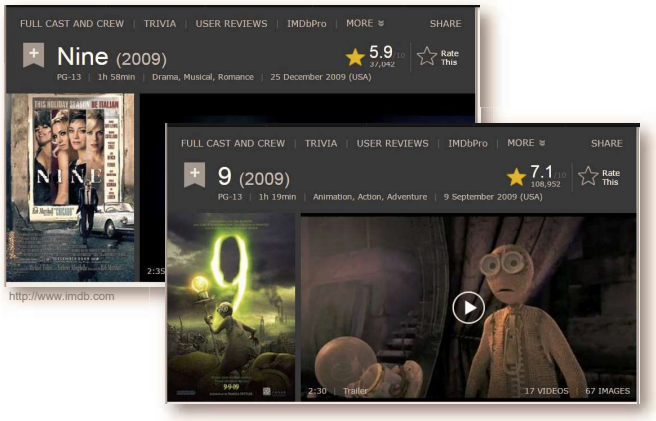
Concatenate "Title,Year": Surely a Key?

"Deja Vu,2006" ←

"Deja Vu,1997" ←

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

Concatenate “Title,Year”: Surely a Key?



Typical Pipeline

	Title	Year	Directors	Cast	Runtime	…
	Come Fly With Me	2004	Peter Kagan	Michael Bublé	63	…
	Michael Jordan: Come Fly With Me	1989		Michael Jordan, Jay Thomas	42	…

Prep → Block → Score → Match

Typical Pipeline

	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

Clean records

Typical Pipeline

	Title	…
	come fly with me	…
	michael jordan come fly with me	…

We’ll need to compare/”score” many pairs of records for similarity

If we compare source 1’s n records against ***all*** m of source 2’s: $m \cdot n$ work

One MSFT problem had **1b** records

Prep → Block → Score → Match → Merge

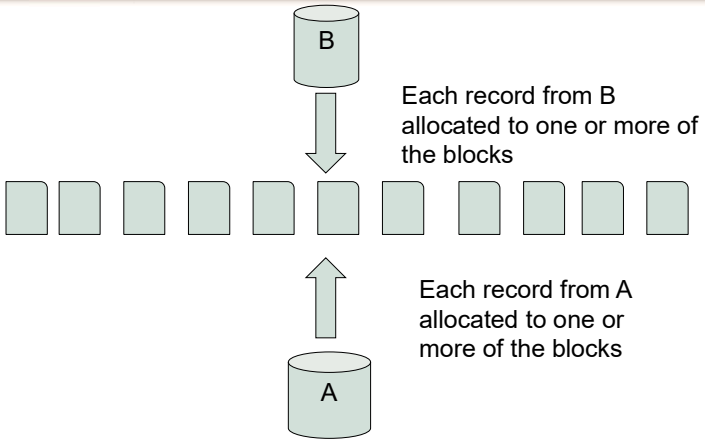
Represent complex records as simple values (blocks); Only score records with simple value (block) in common.

Typical Pipeline

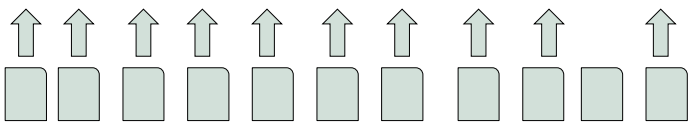
	Title	…
	come fly with me	…
	michael jordan come fly with me	…

Prep | Block | Score | Match | Merge

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



Within each block, compare the records from A against those from B and find those that match



If two records are not assigned to the same block, it means we believe they are not a match

Blocking example <``Ghost busters'', 2016>

Prep | Block | Score | Match | Merge

What blocks to use and which block(s) contain above record?

Q: Block on release year being same?

Q: Block on same release year, and title word in common?

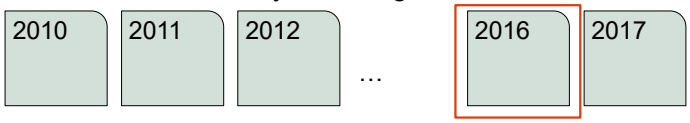
Q: Block on release year being same or ±1?

THE UNIVERSITY OF MELBOURNE | Blocking example <``Ghost busters'', 2016>

Prep | Block | Score | Match | Merge

What blocks to use and which block(s) contain above record?

Q: Block on release year being same?



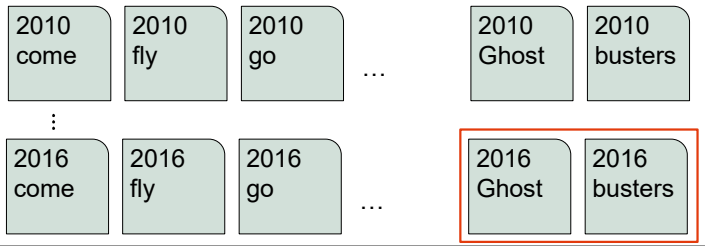
THE UNIVERSITY OF MELBOURNE | Blocking example <``Ghost busters'', 2016>

Prep | Block | Score | Match | Merge

What blocks to use and which block(s) contain above record?

How the blocks defined using year-keyword is different than the one defined using year only?

Q: Block on same release year, and title word in common?



THE UNIVERSITY OF MELBOURNE | Blocking example <``Ghost busters'', 2016>

Prep | Block | Score | Match | Merge

What blocks to use and which block(s) contain above record?

Q: Block on release year being same or ±1?

THE UNIVERSITY OF MELBOURNE | Blocking example <``Ghost busters'', 2016>

Prep | Block | Score | Match | Merge

What blocks to use and which block(s) contain above record?

Q: Block on release year being same or ±1?

Is it correct to put the "Ghost busters", 2016 movie in only two blocks: 2016 and 2017?

Typical Pipeline

	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 : Asses their similarity

Scoring Examples: String Similarity

Prep → Block → Score → Match → Merge

- Jaccard similarity

$$sim(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|}$$

  - Treat each string s_i as a set of words
  - $|\ldots|$ means count number of words in set
  - Set intersection symbol ∩
  - Set union symbol ∪
  - Ex: $Jaccard\_sim$ (“come fly with me” , “michael jordan come fly with me”)= $\frac{4}{6}$

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

Typical Pipeline

	Title	…
	come fly with me	…
	michael jordan come fly with me	…

(0.82, 15, 0, 0, 21)

↓

f: $\mathbf{R}^d \rightarrow [0,1]$ — Need a good *f*

↓

0.1

Prep → Block → Score → Match → Merge

Score record pairs for similarity

More on Score Combination

Prep → Block → Score → Match → Merge

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 weights

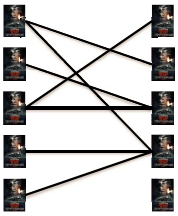
(0.82, 15, 0, 0, 21)

↓

f: $\mathbf{R}^d \rightarrow [0,1]$

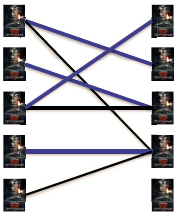
↓

0.1



Prep | Block | Score | Match | Merge

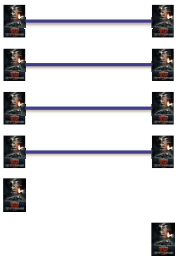
Match "sufficiently similar" records



Threshold
* Match when score > 0.5
- Match when score > θ

Prep | Block | Score | Match | Merge

Match "sufficiently similar" records



Prep | Block | Score | Match | Merge

Merge matched records;
Resolve conflicting attributes

- ✓ 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

THE UNIVERSITY OF MELBOURNE

Acknowledgements

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