Australian National University

COMP3430 / COMP8430
Data wrangling

Lecture 21: Advanced record linkage techniques

(Lecturer: Peter Christen)

Lecture outline

- Group linkage
- Collective linkage techniques
- Active learning
- Geocode matching
- Linking temporal and dynamic data

(Much of this lecture is based on our research – for papers see Peter's homepage: http://users.cecs.anu.edu.au/~christen/publications.html)

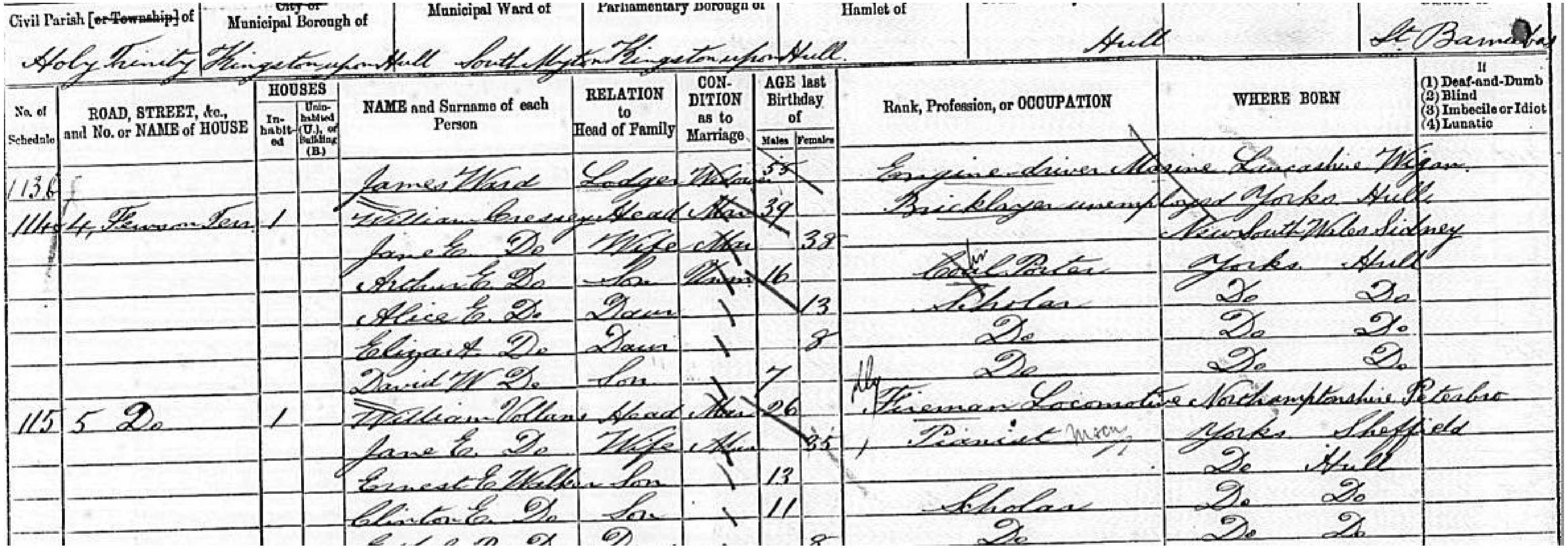
Group linkage (1)

- Traditional (probabilistic) record linkage considers individual record pairs, and classifies each pair individually
- In some applications we have groups of records
  - People living in the same household (for example in census databases)
  - Publications written by a group of co-authors

- Group linkage algorithms make use of such information to improve linkage quality
  - First they generally calculate similarities between individual records
  - Then calculate group similarities based on graph algorithms

Group linkage (2)

- Example: Linking households in historical census databases
  (PhD thesis by Zhichun (Sally) Fu, ANU 2014)

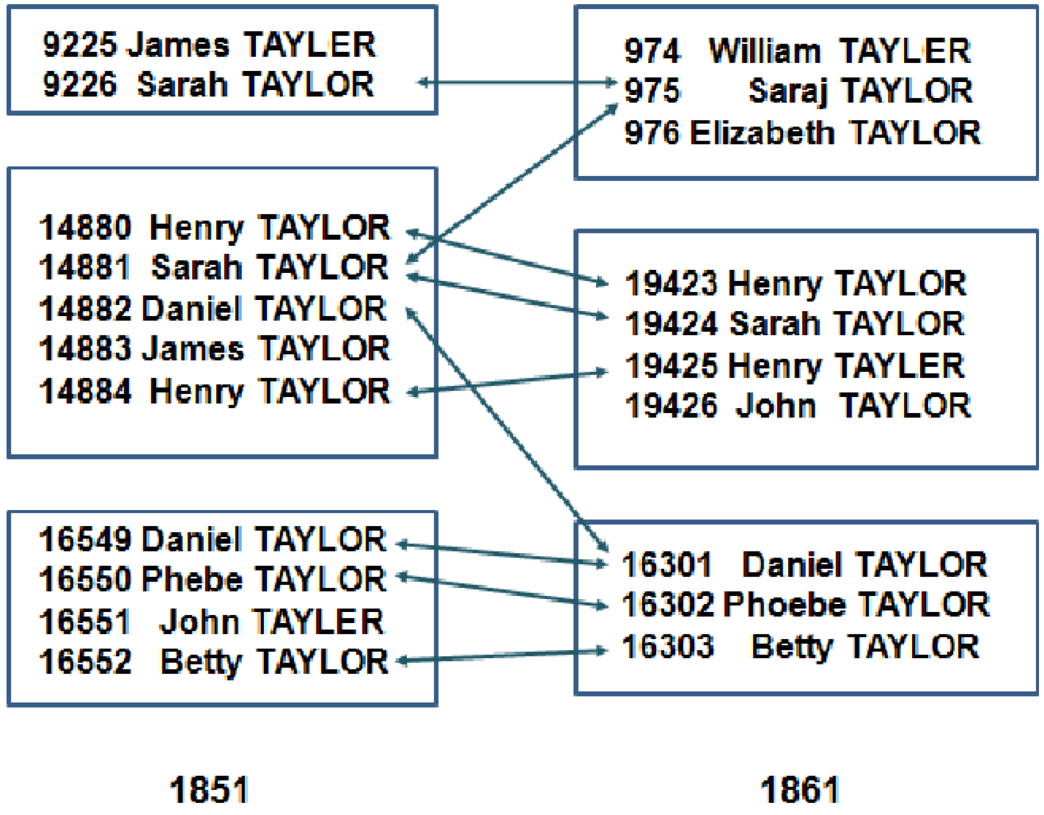


Civil Parish [~~or Township~~] of: Holy Trinity | City or Municipal Borough of: Kingston upon Hull | Municipal Ward of: South Myton | Parliamentary Borough of: Kingston upon Hull | Hamlet of: Hull | St Barnabas

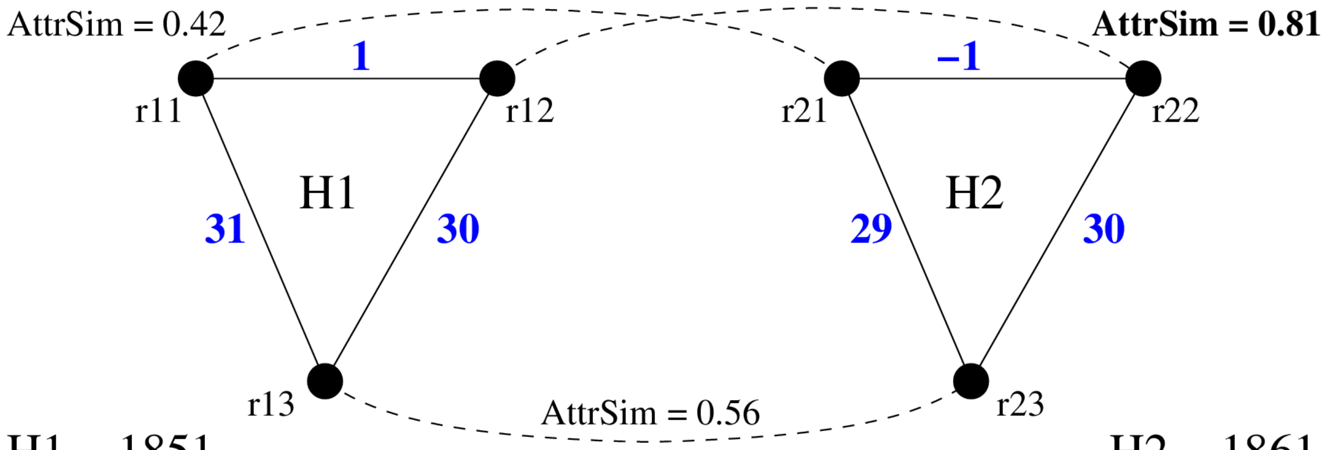
No. of Schedule	ROAD, STREET, &c., and No. or NAME of HOUSE	HOUSES In-habited	HOUSES Uninhabited (U.), or Building (B.)	NAME and Surname of each Person	RELATION to Head of Family	CONDITION as to Marriage	AGE last Birthday of Males	AGE last Birthday of Females	Rank, Profession, or OCCUPATION	WHERE BORN	If (1) Deaf-and-Dumb (2) Blind (3) Imbecile or Idiot (4) Lunatic
113				James Ward	Lodger	Widow	55		Engine driver Marine	Lincolnshire Wigan	
114	4, Fenton Ter.	1		William Cressey	Head	Mar	39		Bricklayer unemployed	Yorks Hull	
				Jane E. Do	Wife	Mar		38		New South Wales Sidney	
				Arthur E. Do	Son	Unm	16		Coal Porter	Yorks Hull	
				Alice E. Do	Dau			13	Scholar	Do Do	
				Eliza A. Do	Dau			8	Do	Do Do	
				David W. Do	Son		7		Do	Do Do	
115	5 Do	1		William Vollans	Head	Mar	26		Fireman Locomotive	Northamptonshire Peterboro	
				Jane E. Do	Wife	Mar		25	Pianist	Yorks Sheffield	
				Ernest E. Walker	Son		13			Do Hull	
				Clinton E. Do	Son		11		Scholar	Do Do	
				[illegible]	[illegible]			8	Do	Do Do	

Group linkage (3)

- Calculate household similarities using Jaccard or weighted similarities (based on pair-wise links)
- Promising results on UK Census data from 1851 to 1901 (town of Rawtenstall, around 17,000 to 31,000 records)



Graph-matching based on household structure



H1 – 1851

ID	Address	SN	FN	Age
r11	goodshaw	smith	john	**32**
r12	goodshaw	smith	mary	**31**
r13	goodshaw	smith	anton	**1**

H2 – 1861

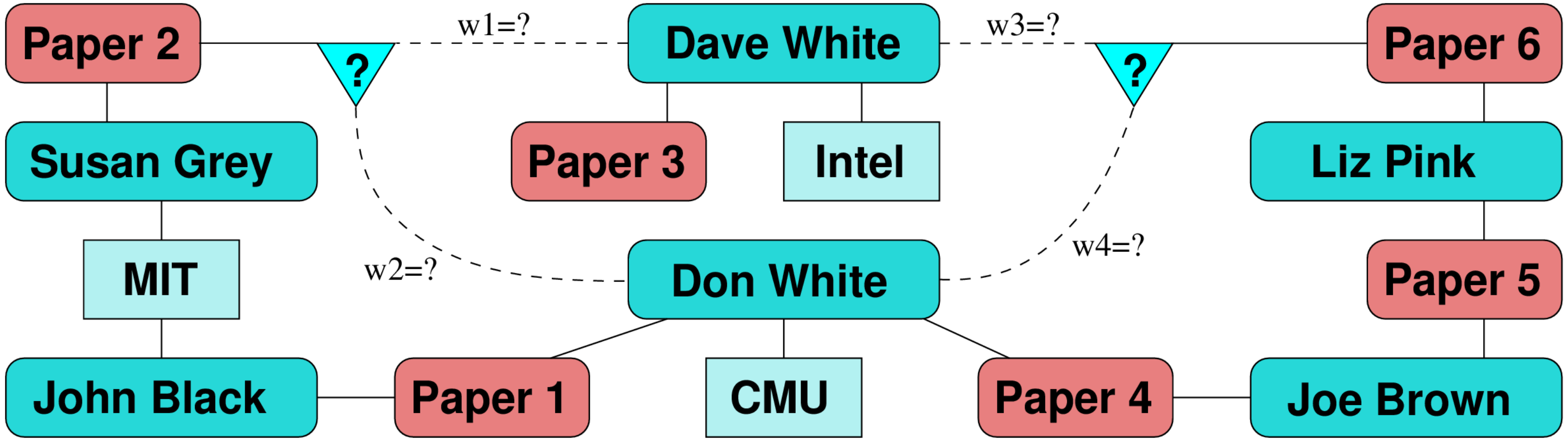
ID	Address	SN	FN	Age
r21	goodshaw	smith	jack	**39**
r22	goodshaw	smith	marie	**40**
r23	goodshaw	smith	toni	**10**

- One graph per household, find best matching graphs using both record attribute and structural similarities
- Edge attributes are information that does not change over time (like age differences)

Collective linkage techniques (1)

- Group and graph techniques generally still are based on pair-wise similarities, and they classify each group individually
  - Still have the problem of possibly violating transitivity
- Recently developed collective techniques aim to find an overall based linkage solution
  - Generally based on some form of clustering (grouping) of records, where each cluster should contain all records about the same entity
  - These approaches take relationships into account when calculating similarities (not just attributes)
  - Generally lead to improved linkage quality, but at much higher computational costs

Collective linkage techniques (2)



(A1, Dave White, Intel)
(A2, Don White, CMU)
(A3, Susan Grey, MIT)
(A4, John Black, MIT)
(A5, Joe Brown, unknown)
(A6, Liz Pink, unknown)

(P1, John Black / Don White)
(P2, Sue Grey / D. White)
(P3, Dave White)
(P4, Don White / Joe Brown)
(P5, Joe Brown / Liz Pink)
(P6, Liz Pink / D. White)

Adapted from: [Kalashnikov and Mehrotra, ACM TODS, 2006]

Active learning (1)

- Supervised classification techniques for record linkage generally result in improved linkage quality
- However, training data in the form of true matches and true non-matches are rarely available in practice
- They have to be manually generated, which is generally difficult both in terms of cost and quality

- Two challenges stand out:
  1. How can we ensure *good* examples (record pairs) are selected for training?
  2. How can we minimise the user's burden of labeling examples?

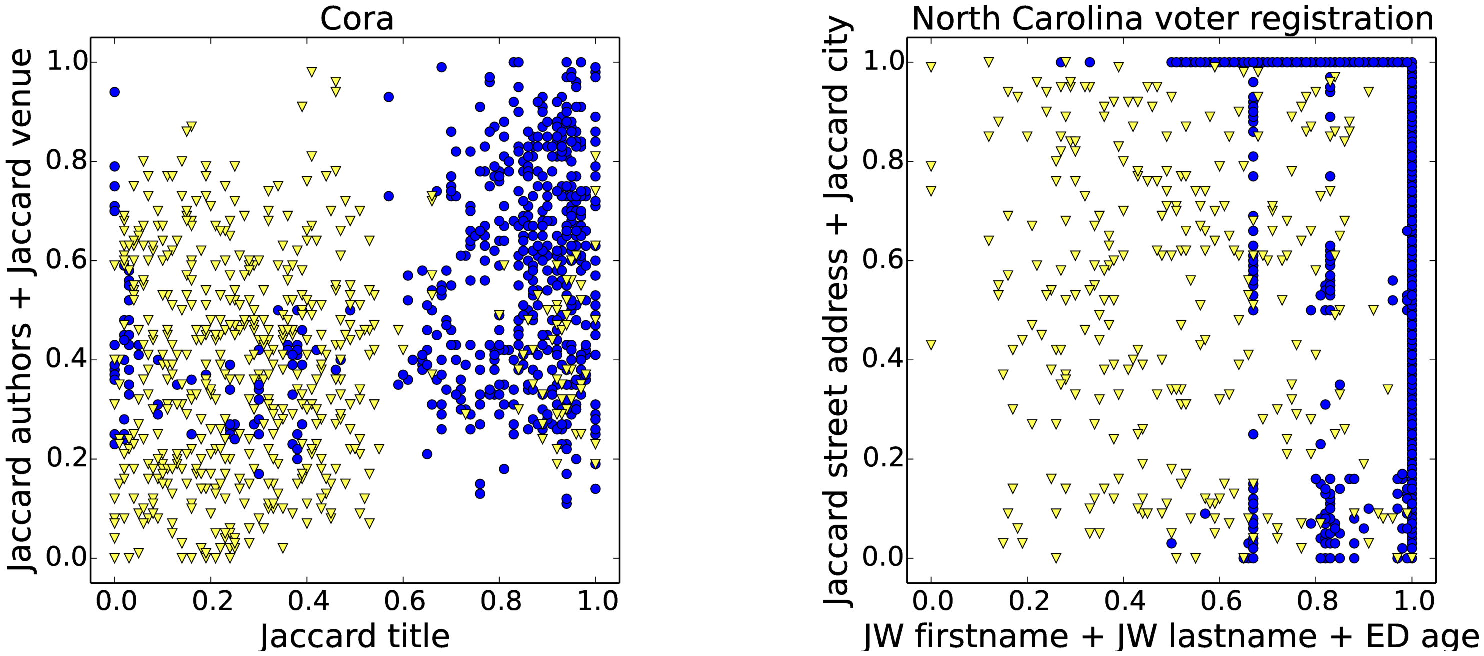
Active learning (2)

- Active learning is the process of combining a supervised machine learning classifier with manual classification

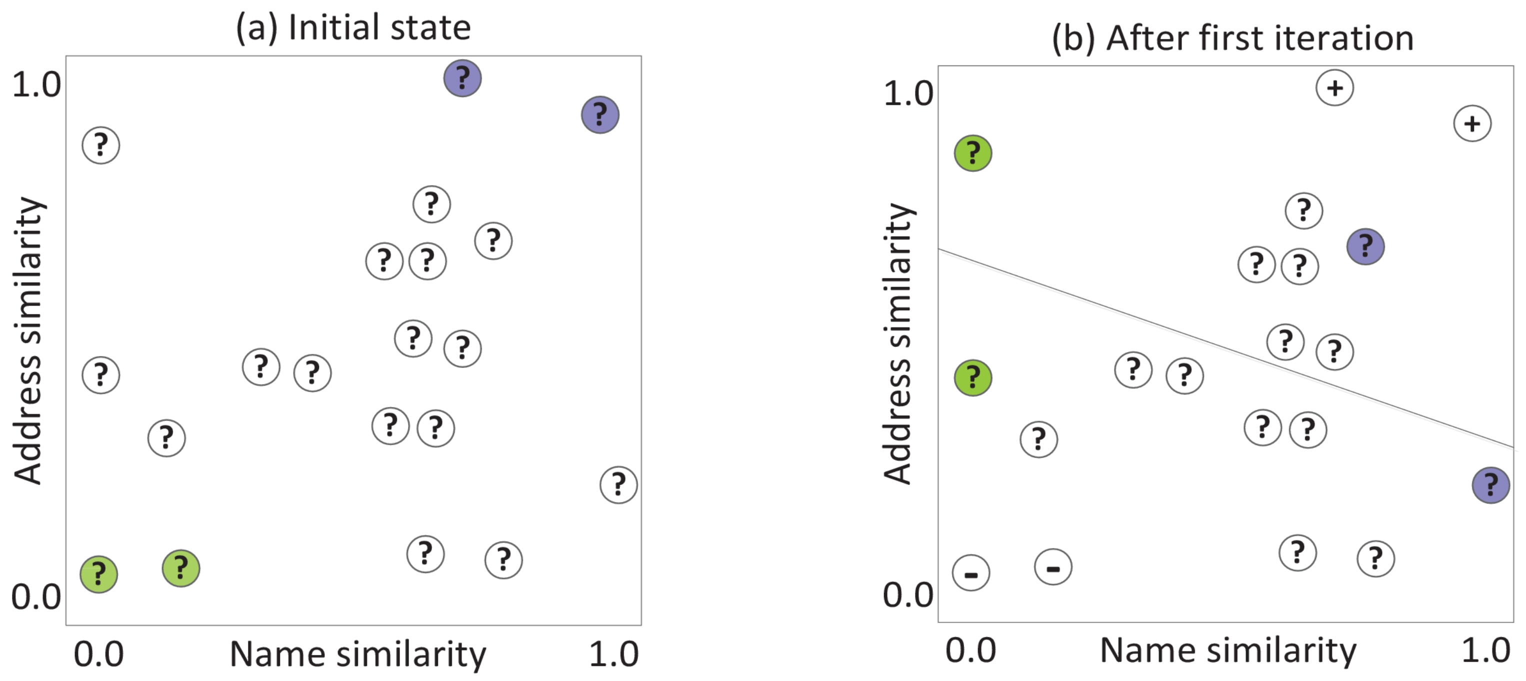
- An iterative process where
  1) A machine learning classifier is trained on some initial training data (possibly already available or manually generated)
  2) A set of difficult to classify training examples (record pairs) are given to a domain expert for manual classification and added to the training set
  3) An improved machine learning classifier is trained
  4) The process is repeated until (a) high enough linkage quality is achieved or (b) a *budget* for the amount of manual classifications possible is reached

Monotonicity of similarities

- Assumption of most active learning techniques for record linkage: The higher the overall similarity between records is the more likely they are true matches
- In practice, monotonicity does generally not hold

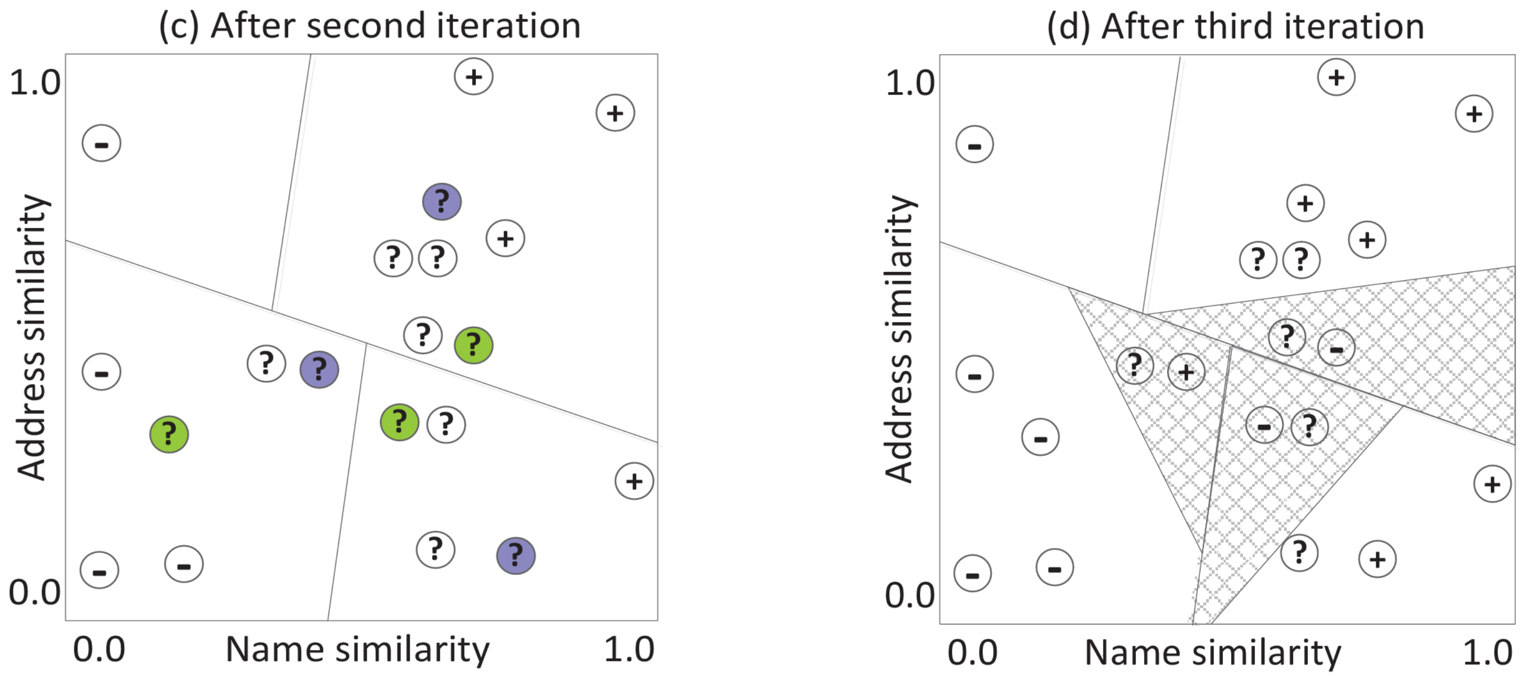


Adaptive and interactive training data selection (1)



- Our approach exploit the cluster structure of similarity vectors calculated from compared record pairs
- In each iteration, a selected set of record pairs is manually classified

Adaptive and interactive training data selection (2)

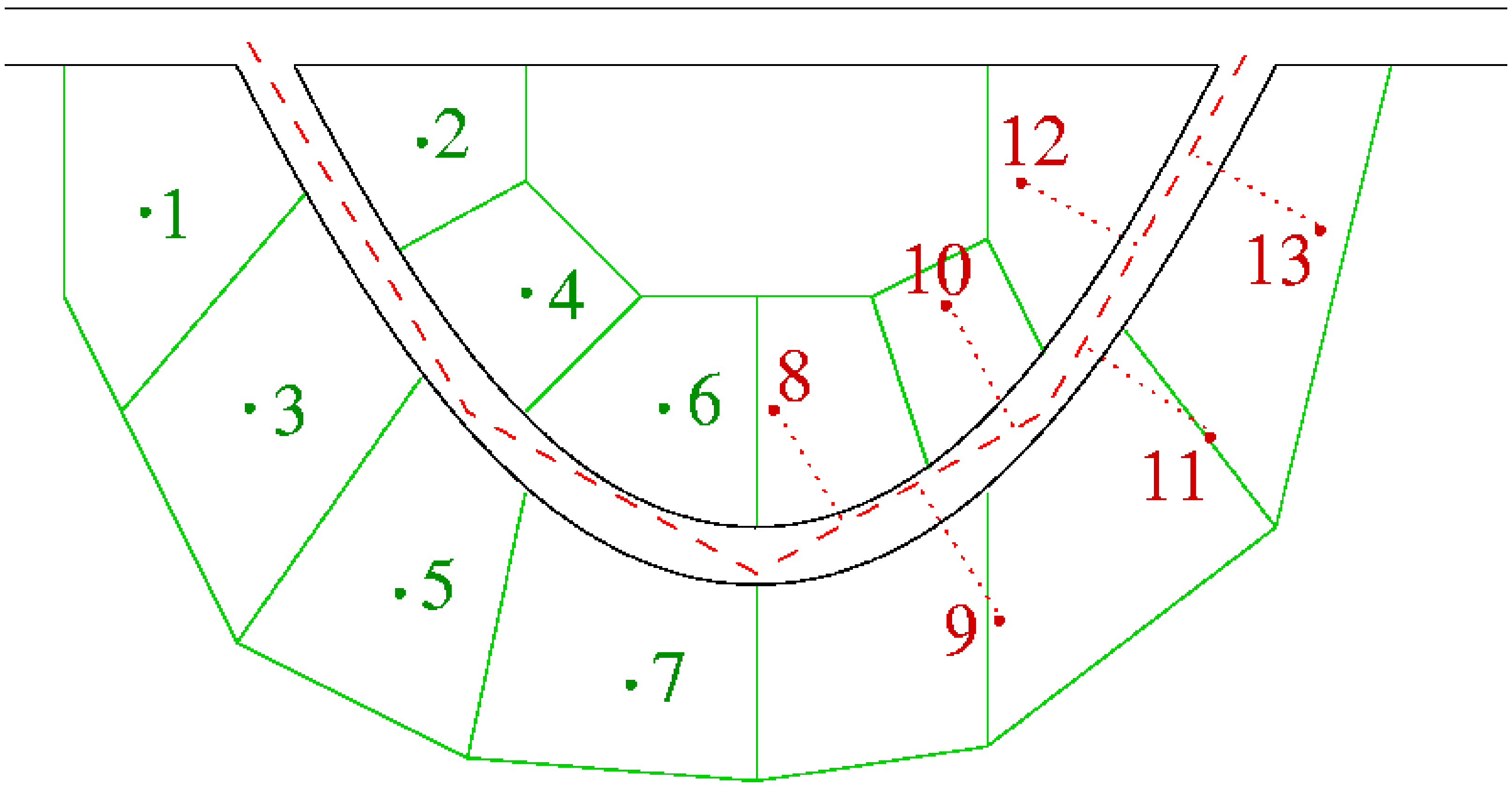


- We recursively split the set of similarity vectors to find pure enough sub-sets for training
- We select clusters into the training set if they have a minimum purity, otherwise they are inserted into a queue for further recursive splitting

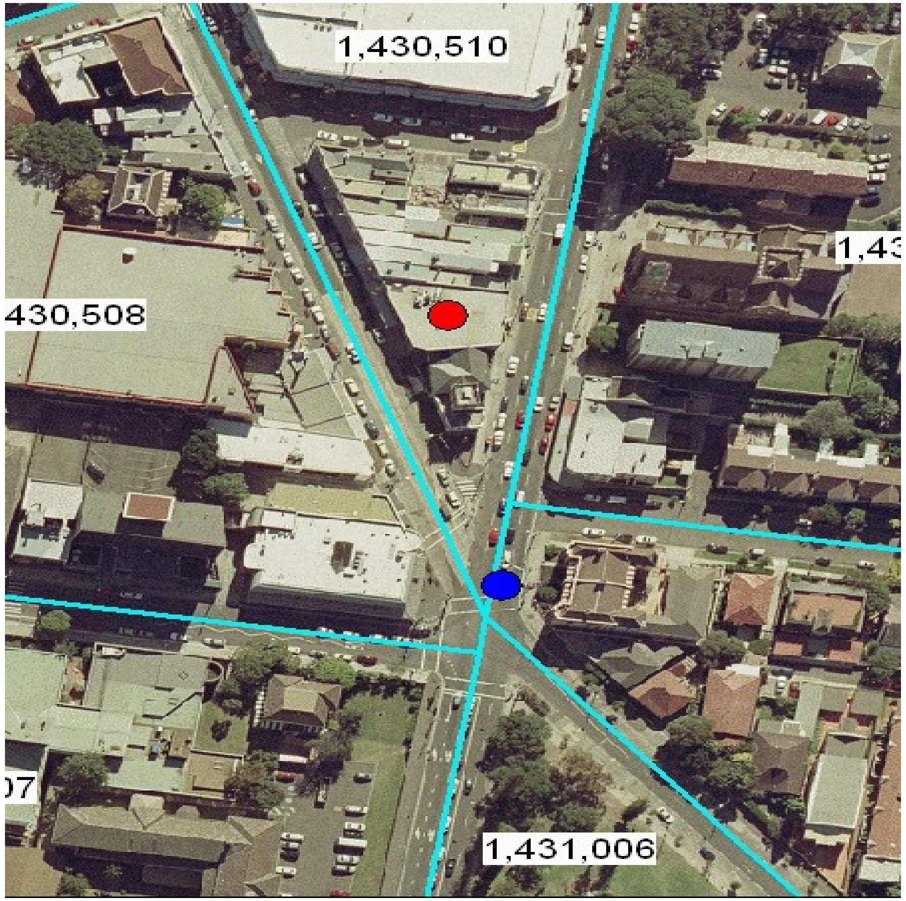
Geocode matching (1)

- Aim is to match addresses against *geocoded* reference data (addresses and their geographic locations: latitudes and longitudes)
- Useful for spatial data analysis / mining and for loading data into geographical information systems
- Matching accuracy is critical for good geocoding (as is accurate geocoded address data)
- Australia has a *Geocoded National Address File (G-NAF)* since early 2004 (all Australian property addresses and their locations)
- Commercial geocoding systems in the past have been based on *street centreline* data

Geocode matching (2)



Geocode matching example



Linking temporal and dynamic data (1)

- So far we assumed the databases to be linked are static and do not contain temporal information
- However, many databases contain time-stamps for all records
  - When a record was added to the databases (a new customer or a new patient)
  - When a record was modified (change of name or address details of a person)
- Approaches to linking temporal data aim to make use of patterns in such changing details

Linking temporal and dynamic data (2)

RecID	EntID	GivenName	Surname	Street	City	Time-stamp
r1	e1	Gale	Miller	13 Main Rd	Sydney	2006-01-21
r2	e2	Peter	O’Brian	43/1 Miller St	Sydeny	2006-02-21
r3	e1	Gail	Miller	11 Town Pl	Hobart	2007-01-28
r4	e1	Gail	Smith	42 Ocean Dr	Perth	2007-07-12
r5	e2	Pete	O’Brien	43 Miller St	Sydney	2008-01-11
r6	e1	Abigail	Smith	42 Ocean Dr	Perth	2008-06-30
r7	e2	Peter	OBrian	12 Nice Tr	Brisbane	2009-01-01
r8	e1	Gayle	Smith	11a Town Pl	Sydney	2009-04-29

- An entity changes address values more often than surname values
- Small variations in values are possible (no actual changes)
- Several entities can have the same value in an attribute

Linking temporal and dynamic data (3)

- Basic ideas of linking temporal and dynamic data are to adjust the similarity weights based on probabilities of attribute values changing over time
  - For example, if two records are five years apart and they have a different address then this doesn't mean they necessarily refer to different people
  - Therefore, a low address similarity is given a small weight in a weighted similarity calculation (as discussed in lecture on classification)
- We calculate temporal agreement and disagreement based on temporal value changes over time
  - For example, address and surname values are more likely to change compared to given name or gender