Information Extraction: Coreference and Relation Extraction

Lecture #20

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

Information Extraction: Coreference and Relation Extraction

Lecture #20

Computational Linguistics

CMPSCI 591N, Spring 2006

University of Massachusetts Amherst

Andrew McCallum

What is “Information Extraction”

As a family of techniques:

Information Extraction =
segmentation + classification + association + clustering

October 14, 2002, 4:00 a.m. PT

For years, Microsoft Corporation CEO Bill Gates railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“

Richard Stallman, founder of the Free Software Foundation, countered saying…

* Microsoft Corporation
CEO
Bill Gates

* Microsoft

* Gates

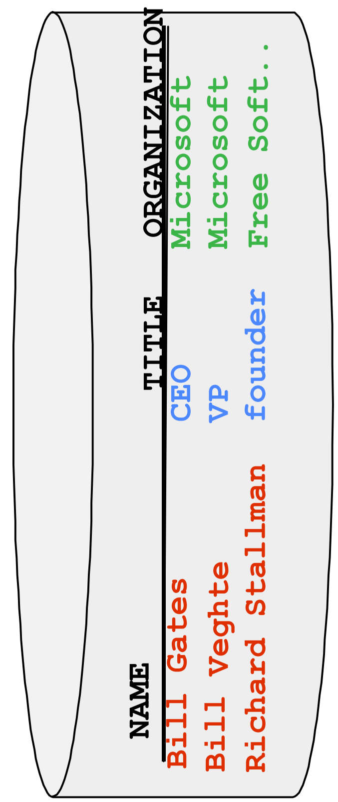
* Bill Veghte
Microsoft

VP

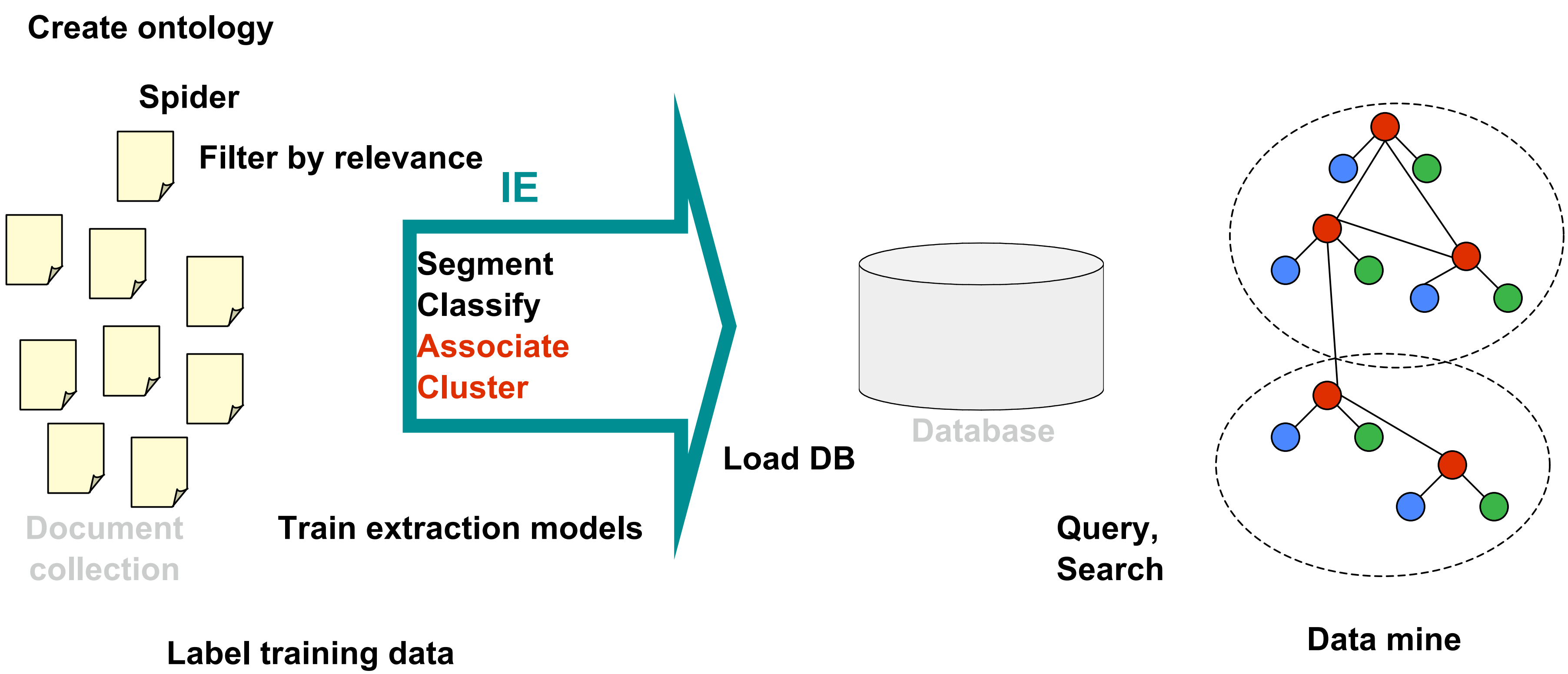
Richard Stallman

founder

Free Software Foundation



IE in Context



Main Points

Co-reference

- How to cast as classification [Cardie]
- Joint resolution [McCallum et al]
- Canopies (time permitting..)

Coreference Resolution

AKA "record linkage", "database record deduplication", "citation matching", "object correspondence", "identity uncertainty"

Input

News article, with named-entity "mentions" tagged

Today Secretary of State Colin Powell met with . . . .
. . . . he . . . . . .
. . . . Condoleezza Rice . . . . .
. . . . Mr Powell . . . . .she . . . .
. . . . Powell . . . .
. . . President Bush . . . .
. . . . Rice . . . .
. . . . . . Bush . . . .
. . . .
. . . .

Output

Number of entities, *N = 3*

#1

Secretary of State Colin Powell
he
Mr. Powell
Powell

#2

Condoleezza Rice
she
Rice

#3

President Bush
Bush

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband,
King George VI, into a viable monarch. Logue,
a renowned speech therapist, was summoned to help
the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

> Queen Elizabeth set about transforming her husband,
> King George VI, into a viable monarch. Logue,
> a renowned speech therapist, was summoned to help
> the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

Noun Phrase Coreference

Identify all noun phrases that refer to the same entity

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

IE Example: Coreference

SAN SALVADOR, 15 JAN 90 (ACAN-EFE) -- [TEXT] ARMANDO CALDERON SOL, PRESIDENT OF THE NATIONALIST REPUBLICAN ALLIANCE (ARENA), THE RULING SALVADORAN PARTY, TODAY CALLED FOR AN INVESTIGATION INTO ANY POSSIBLE CONNECTION BETWEEN THE MILITARY PERSONNEL IMPLICATED IN THE ASSASSINATION OF JESUIT PRIESTS.

"IT IS SOMETHING SO HORRENDOUS, SO MONSTROUS, THAT WE MUST INVESTIGATE THE POSSIBILITY THAT THE FMLN (FARABUNDO MARTI NATIONAL LIBERATION FRONT) STAGED THESE MURDERS TO DISCREDIT THE GOVERNMENT," CALDERON SOL SAID.

SALVADORAN PRESIDENT ALFREDO CRISTIANI IMPLICATED FOUR OFFICERS, INCLUDING ONE COLONEL, AND FIVE MEMBERS OF THE ARMED FORCES IN THE ASSASSINATION OF SIX JESUIT PRIESTS AND TWO WOMEN ON 16 NOVEMBER AT THE CENTRAL AMERICAN UNIVERSITY.

Why It’s Hard

Many sources of information play a role

- head noun matches
  - IBM *executives* = the *executives*
- syntactic constraints
  - John helped himself to...
  - John helped him to…
- number and gender agreement
- discourse focus, recency, syntactic parallelism, semantic class, world knowledge, …

Why It's Hard

- No single source is a completely reliable indicator
  - number agreement
    - the assassination = these murders
- Identifying each of these features automatically, accurately, and in context, is hard
- Coreference resolution subsumes the problem of pronoun resolution…

A Machine Learning Approach

- Classification
  - given a description of two noun phrases, NP_i and NP_j, classify the pair as *coreferent* or *not coreferent*

coref ? *coref* ?

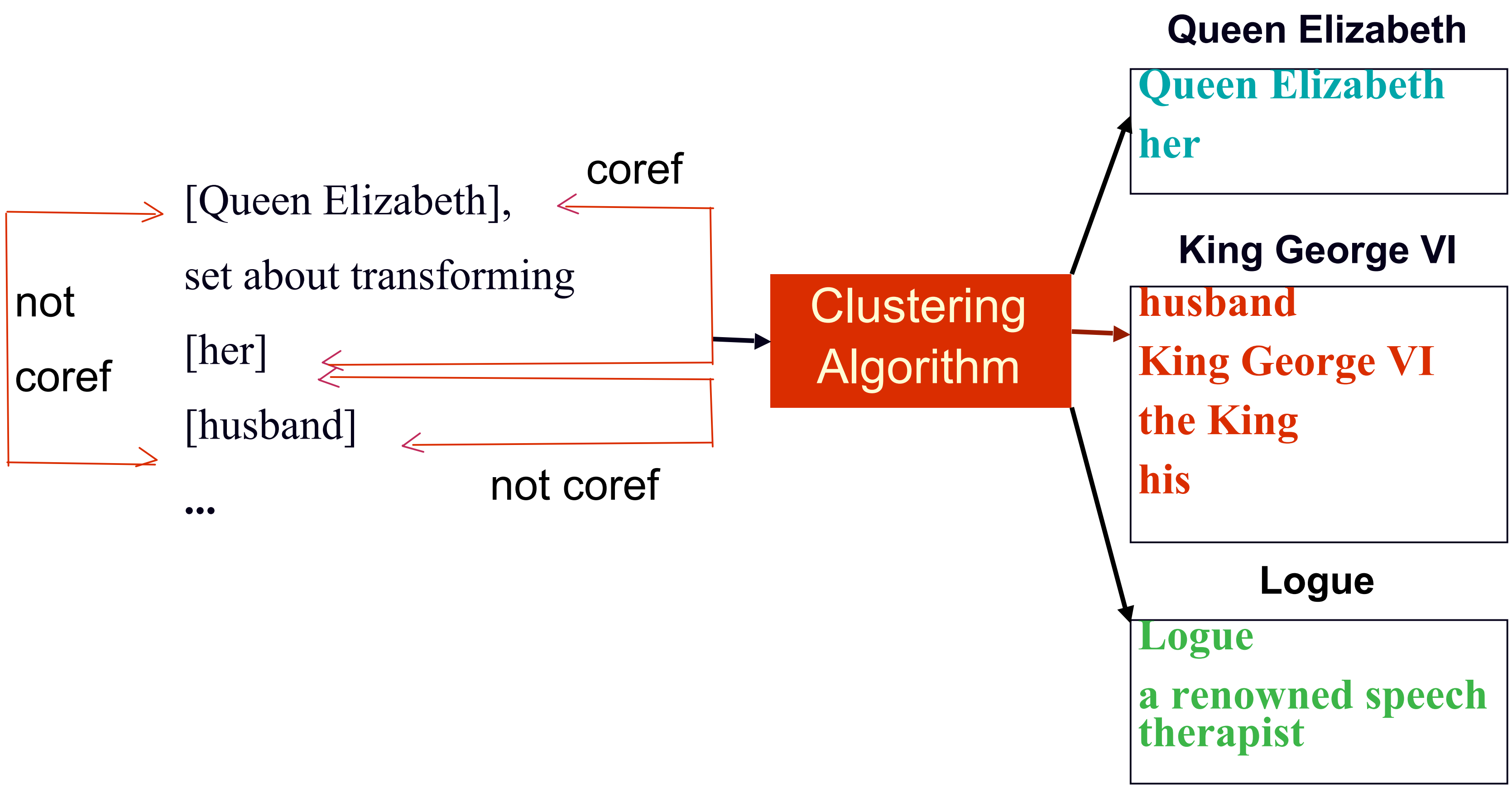
[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Aone & Bennett [1995]; Connolly et al. [1994]; McCarthy & Lehnert [1995]; Soon et al. [2001]; Ng & Cardie [2002]; …

A Machine Learning Approach

- Clustering
  - coordinates pairwise coreference decisions



Machine Learning Issues

- Training data creation
- Instance representation
- Learning algorithm
- Clustering algorithm

Training Data Creation

- Creating training instances
  - texts annotated with coreference information
  - one instance $inst(NP_i, NP_j)$ for each pair of NPs
    - assumption: NP_i precedes NP_j
    - feature vector: describes the two NPs and context
    - class value:

coref	pairs on the same coreference chain
not coref	otherwise

Instance Representation

- 25 features per instance
  - lexical (3)
    - string matching for pronouns, proper names, common nouns
  - grammatical (18)
    - pronoun, demonstrative (the, this), indefinite (it is raining), …
    - number, gender, animacy
    - appositive (george, the king), predicate nominative (a horse is a mammal)
    - binding constraints, simple contra-indexing constraints, …
    - span, maximalnp, …
  - semantic (2)
    - same WordNet class
    - alias
  - positional (1)
    - distance between the NPs in terms of # of sentences
  - knowledge-based (1)
    - naïve pronoun resolution algorithm

Learning Algorithm

- RIPPER (Cohen, 1995)
  C4.5 (Quinlan, 1994)
  - rule learners
    - input: set of training instances
    - output: coreference classifier

- Learned classifier
  - input: test instance (represents pair of NPs)
  - output: classification
    confidence of classification

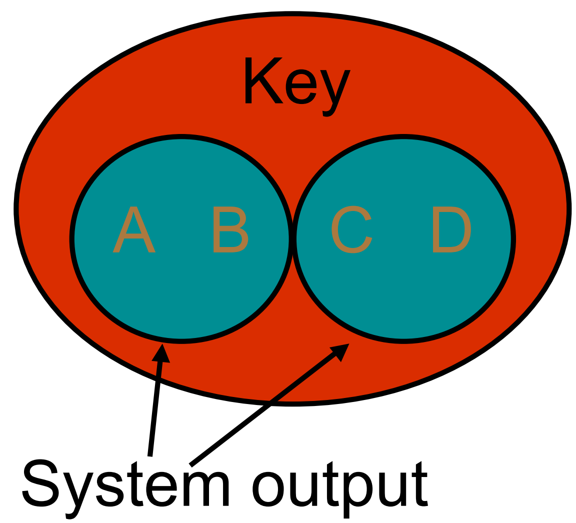
Clustering Algorithm

- Best-first single-link clustering
  - Mark each NP_j as belonging to its own class: $NP_j \in c_j$
  - Proceed through the NPs in left-to-right order.
    - For each NP, NP_j, create test instances, $inst(NP_i, NP_j)$, for all of its preceding NPs, NP_i.
    - Select as the antecedent for NP_j the highest-confidence coreferent NP, NP_i, according to the coreference classifier (or none if all have below .5 confidence);

      Merge c_j and c_j .

Evaluation

- MUC-6 and MUC-7 coreference data sets
- documents annotated w.r.t. coreference
- 30 + 30 training texts (dry run)
- 30 + 20 test texts (formal evaluation)
- scoring program
  - recall
  - precision
  - F-measure: 2PR/(P+R)
- Types
  - MUC
  - ACE
  - Bcubed
  - Pairwise

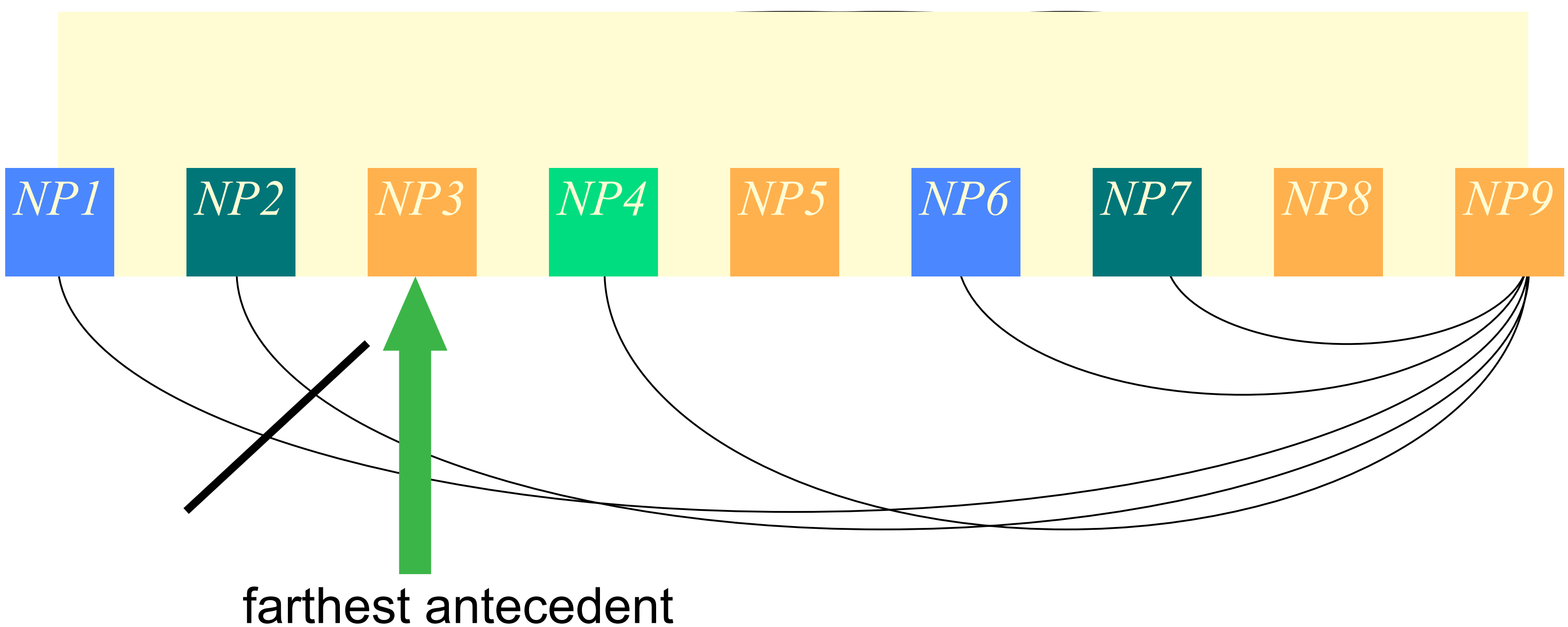


Baseline Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	**52.4**	27.2	86.3	**41.3**
Worst MUC System	36	44	40	52.5	21.4	30.4
Best MUC System	59	72	65	56.1	68.8	61.8

Problem 1

- Coreference is a rare relation
  - skewed class distributions (2% positive instances)
  - *remove some negative instances*



Problem 2

- Coreference is a discourse-level problem
  - different solutions for different types of NPs
    - proper names: string matching and aliasing
  - inclusion of “hard” positive training instances
  - *positive example selection*: selects easy positive training instances (cf. Harabagiu *et al.* (2001))

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, the renowned speech therapist, was summoned to help the King overcome his speech impediment...

Problem 3

- Coreference is an equivalence relation
  - loss of transitivity
  - need to tighten the connection between classification and clustering
  - *prune learned rules w.r.t. the clustering-level coreference scoring function*

coref ? *coref ?*

[Queen Elizabeth] set about transforming [her] [husband], ...

not coref ?

Results

	MUC-6			MUC-7		
	R	P	F	R	P	F
Baseline	40.7	73.5	52.4	27.2	86.3	41.3
NEG-SELECT	46.5	67.8	55.2	37.4	59.7	46.0
POS-SELECT	53.1	80.8	64.1	41.1	78.0	53.8
NEG-SELECT + POS-SELECT	63.4	76.3	69.3	59.5	55.1	57.2
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	69.5	54.2	76.3	63.4

- Ultimately: large increase in F-measure, due to gains in recall

Comparison with Best MUC Systems

	MUC-6			MUC-7		
	R	P	F	R	P	F
NEG-SELECT + POS-SELECT + RULE-SELECT	63.3	76.9	**69.5**	54.2	76.3	**63.4**
Best MUC System	59	72	**65**	56.1	68.8	**61.8**

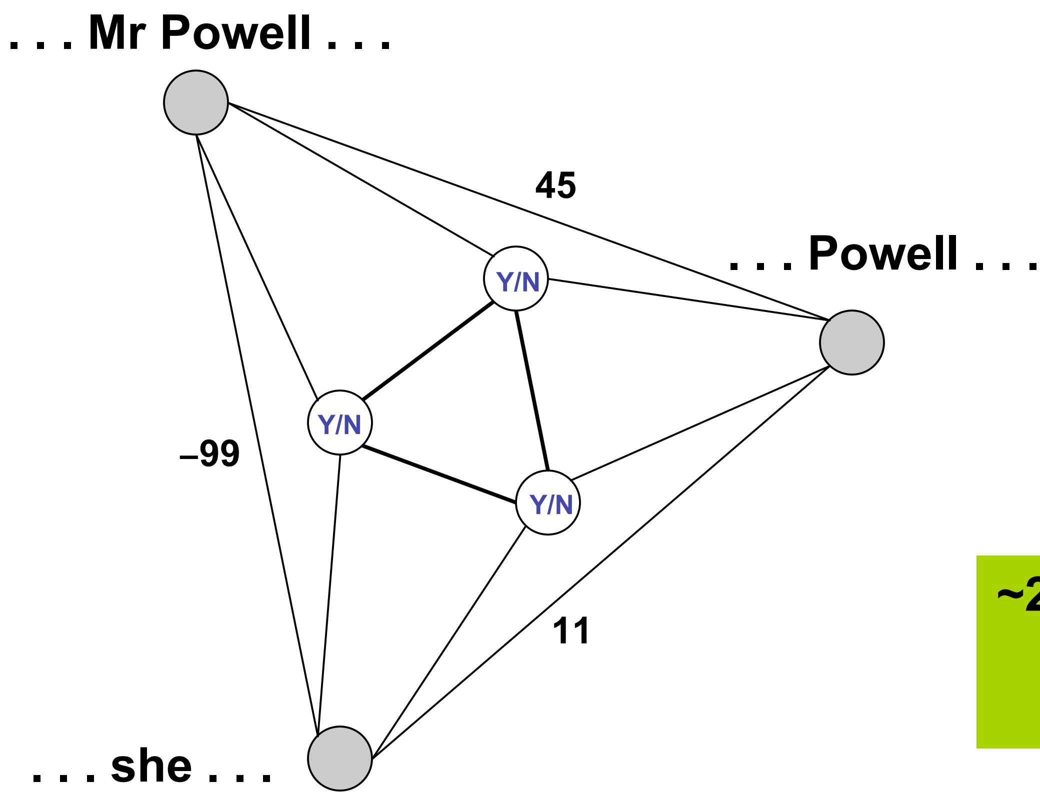
Main Points

Co-reference

- How to cast as classification [Cardie]
- **Joint resolution [McCallum et al]**

Joint co-reference among all pairs
<u>Affinity Matrix CRF</u>

"Entity resolution"
"Object correspondence"



~25% reduction in error on co-reference of proper nouns in newswire.

Inference:
Correlational clustering graph partitioning

[Bansal, Blum, Chawla, 2002]

[McCallum, Wellner, IJCAI WS 2003, NIPS 2004]

Coreference Resolution

AKA "record linkage", "database record deduplication", "citation matching", "object correspondence", "identity uncertainty"

Input

News article, with named-entity "mentions" tagged

Today Secretary of State Colin Powell met with . . . .
. . . . he . . . . . .
. . . . Condoleezza Rice . . . . .
. . . . Mr Powell . . . . .she . . . .
. . . . Powell . . . .
. . . President Bush . . . .
. . . . Rice . . . .
. . . . . . Bush . . . .
. . . .
. . . .

Output

Number of entities, *N = 3*

#1
- Secretary of State Colin Powell
- he
- Mr. Powell
- Powell

#2
- Condoleezza Rice
- she
- Rice

#3
- President Bush
- Bush

Inside the Traditional Solution

Pair-wise Affinity Metric

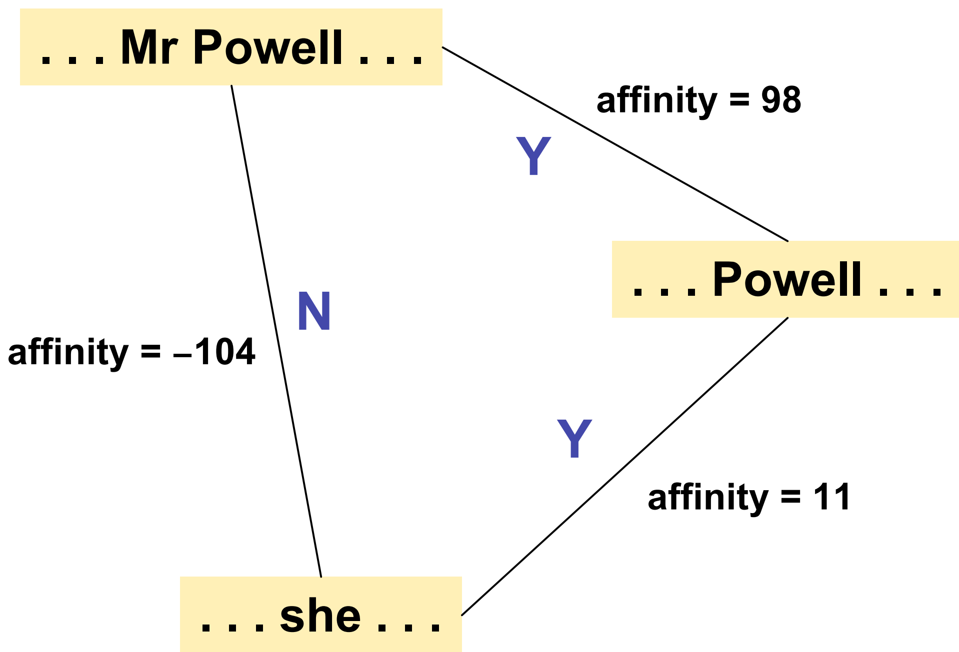
Mention (3) **Mention (4)**

. . . Mr Powell . . . **Y/N?** **. . . Powell . . .**

N	Two words in common	29
Y	**One word in common**	**13**
Y	**"Normalized" mentions are string identical**	**39**
Y	**Capitalized word in common**	**17**
Y	**> 50% character tri-gram overlap**	**19**
N	< 25% character tri-gram overlap	-34
Y	**In same sentence**	**9**
Y	**Within two sentences**	**8**
N	Further than 3 sentences apart	-1
Y	**"Hobbs Distance" < 3**	**11**
N	Number of entities in between two mentions = 0	12
N	Number of entities in between two mentions > 4	-3
Y	**Font matches**	**1**
Y	**Default**	**-19**

OVERALL SCORE = 98 > threshold=0

The Problem



Pair-wise merging decisions are being made independently from each other

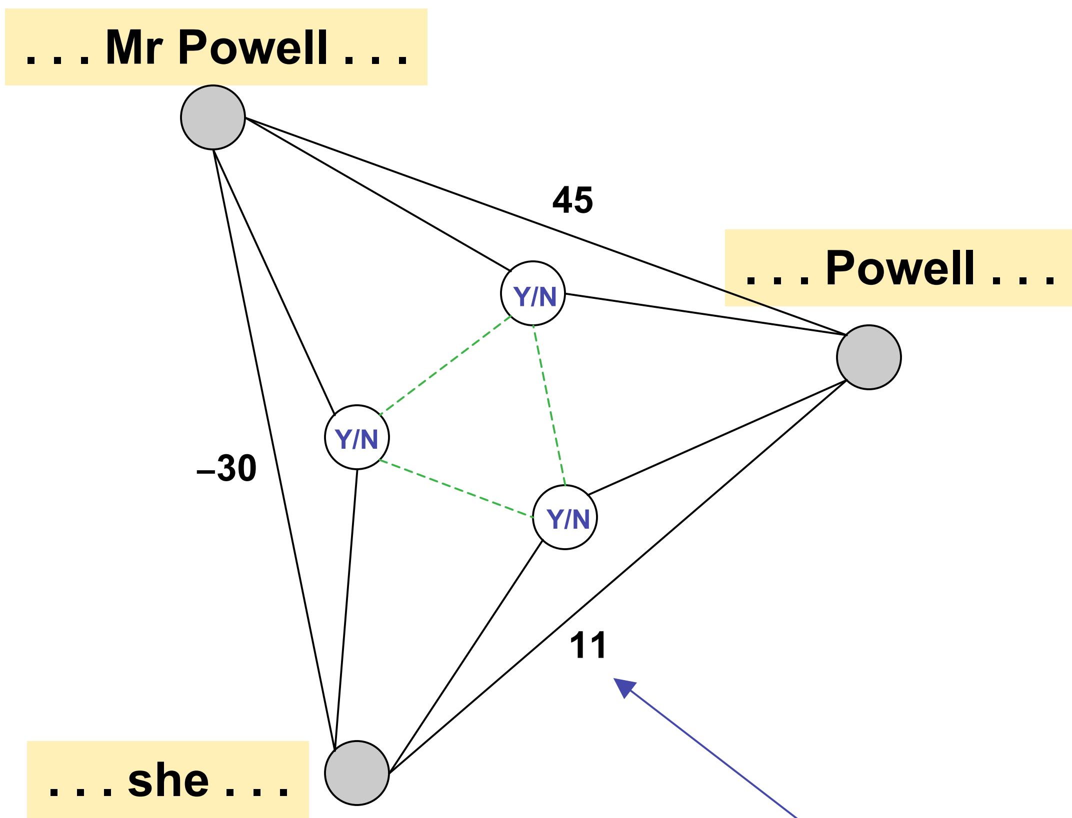
They should be made in relational dependence with each other.

Affinity measures are noisy and imperfect.

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003, ICML]



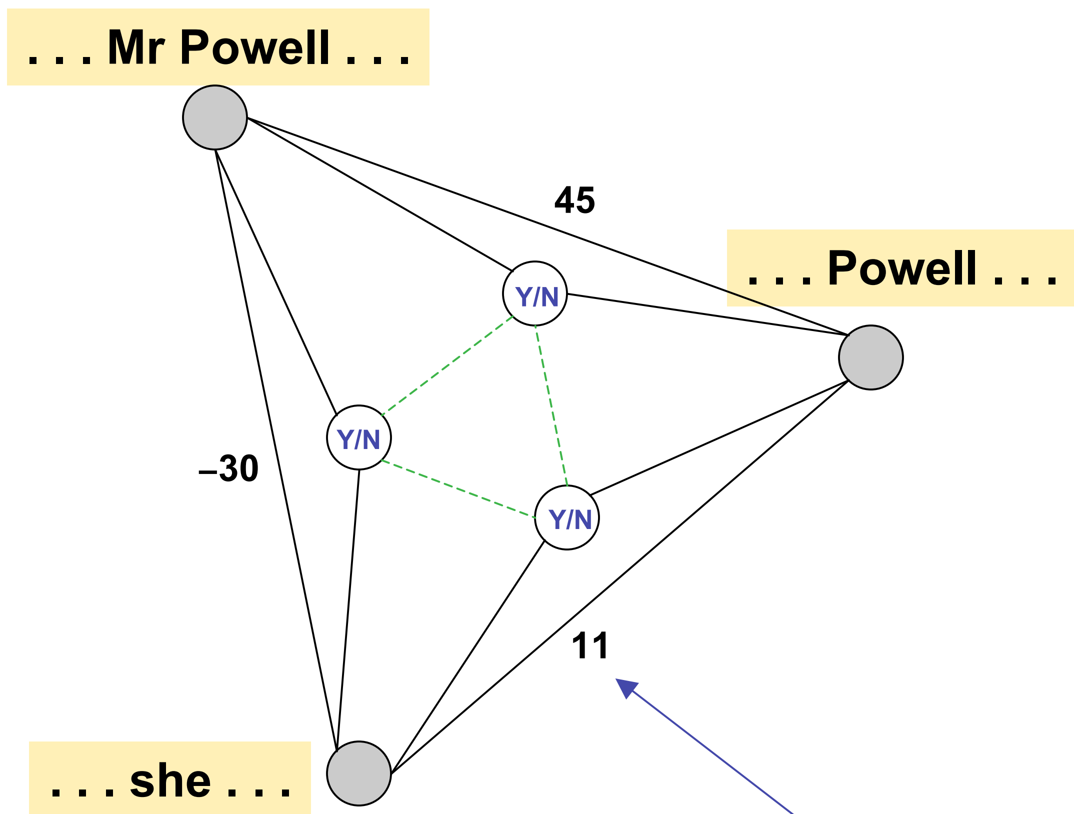
Make pair-wise merging decisions in dependent relation to each other by
- calculating a joint prob.
- including all edge weights
- adding dependence on consistent triangles.

$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \overbrace{\sum_{l} \lambda_l f_l(x_i, x_j, y_{ij})} + \sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik}) \right)$$

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003]



Make pair-wise merging decisions in dependent relation to each other by
- calculating a joint prob.
- including all edge weights
- adding dependence on consistent triangles.

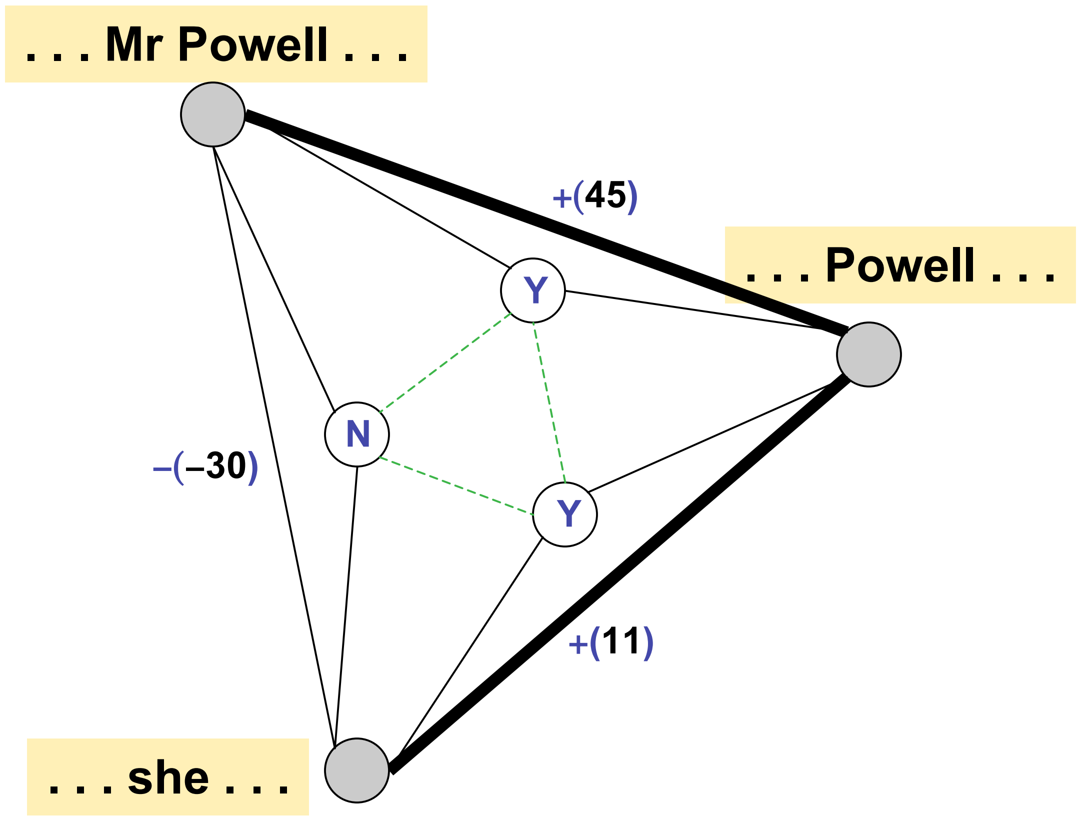
$-\infty$

$$P(\vec{y}\,|\,\vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \overbrace{\sum_{l} \lambda_l f_l(x_i, x_j, y_{ij})} + \sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik}) \right)$$

A Markov Random Field for Co-reference

(MRF)

[McCallum & Wellner, 2003]

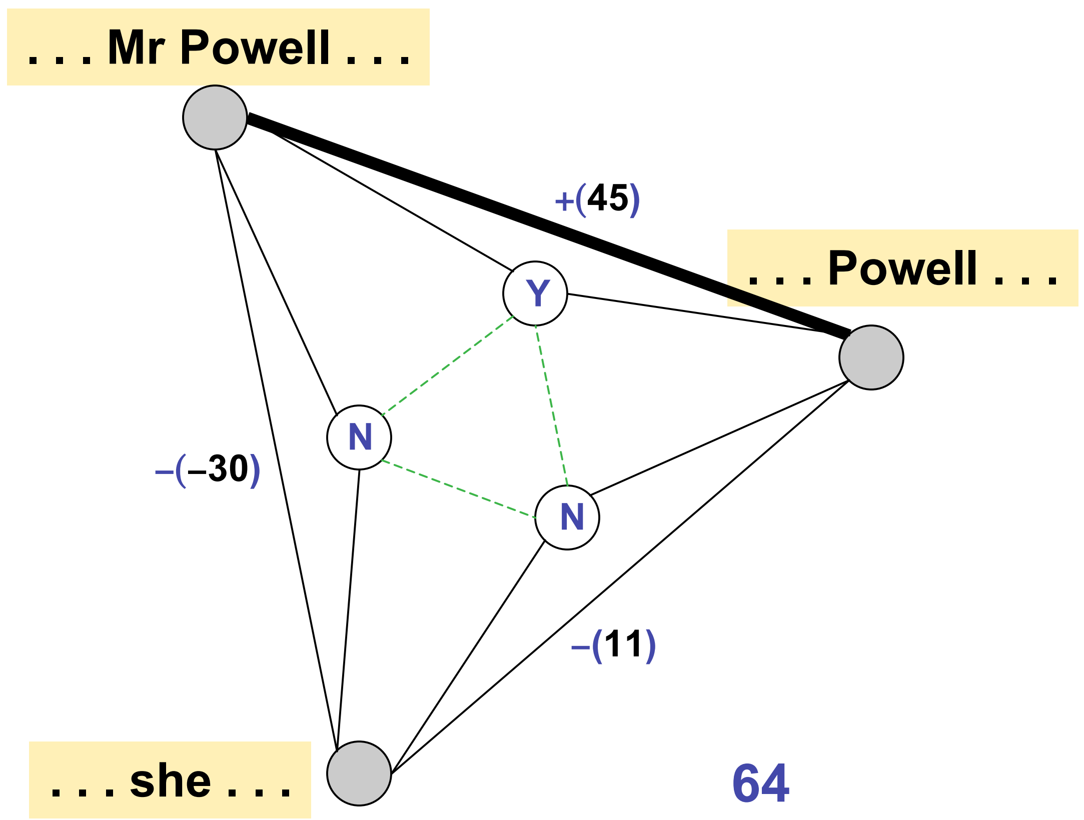


$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) + \overbrace{\sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik})}^{-\text{infinity}} \right)$$

A Markov Random Field for Co-reference

(MRF)

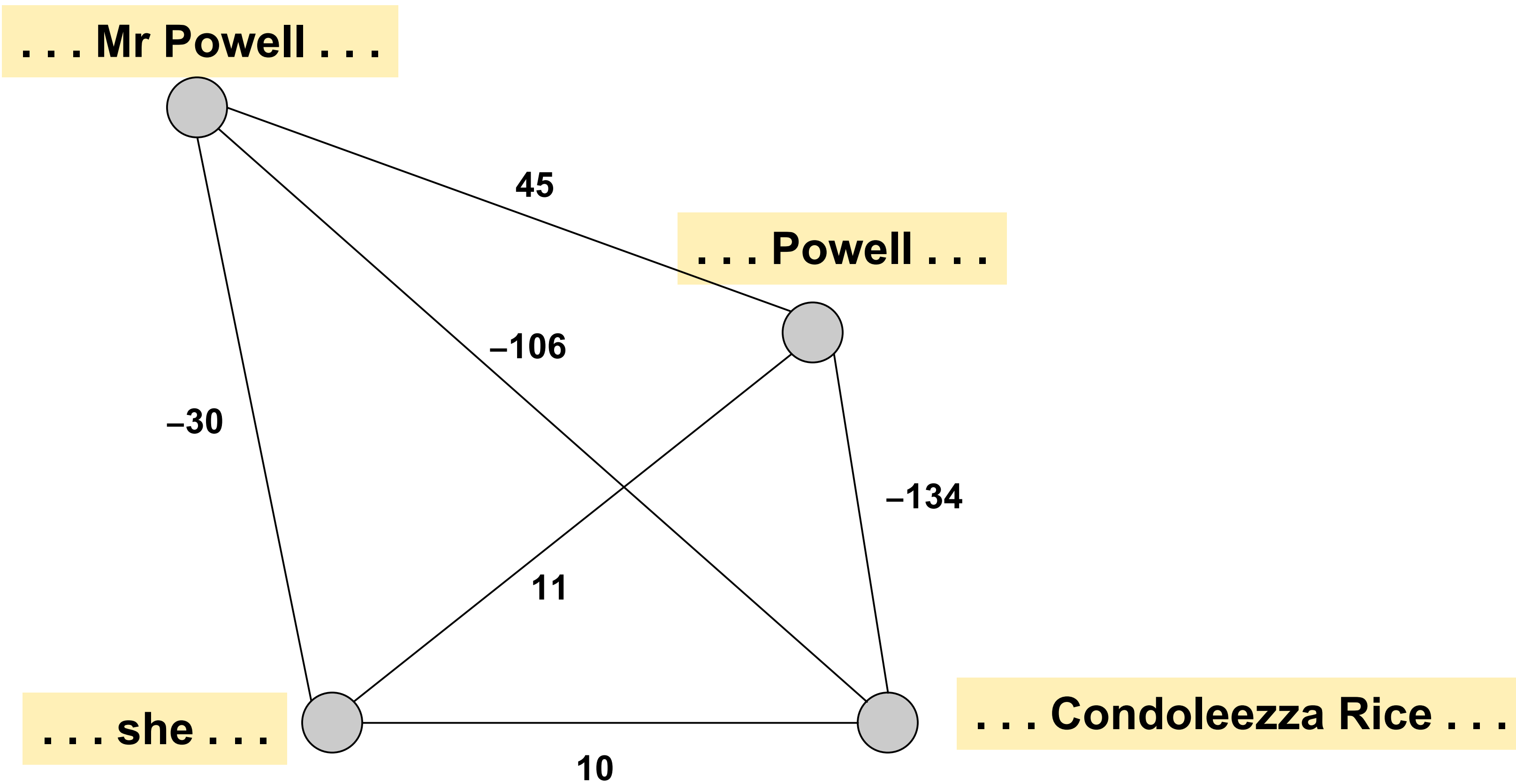
[McCallum & Wellner, 2003]



$$P(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{x}}} \exp\left(\underbrace{\sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij})}^{64} + \sum_{i,j,k} \lambda' f'(y_{ij}, y_{jk}, y_{ik}) \right)$$

Inference in these MRFs = Graph Partitioning

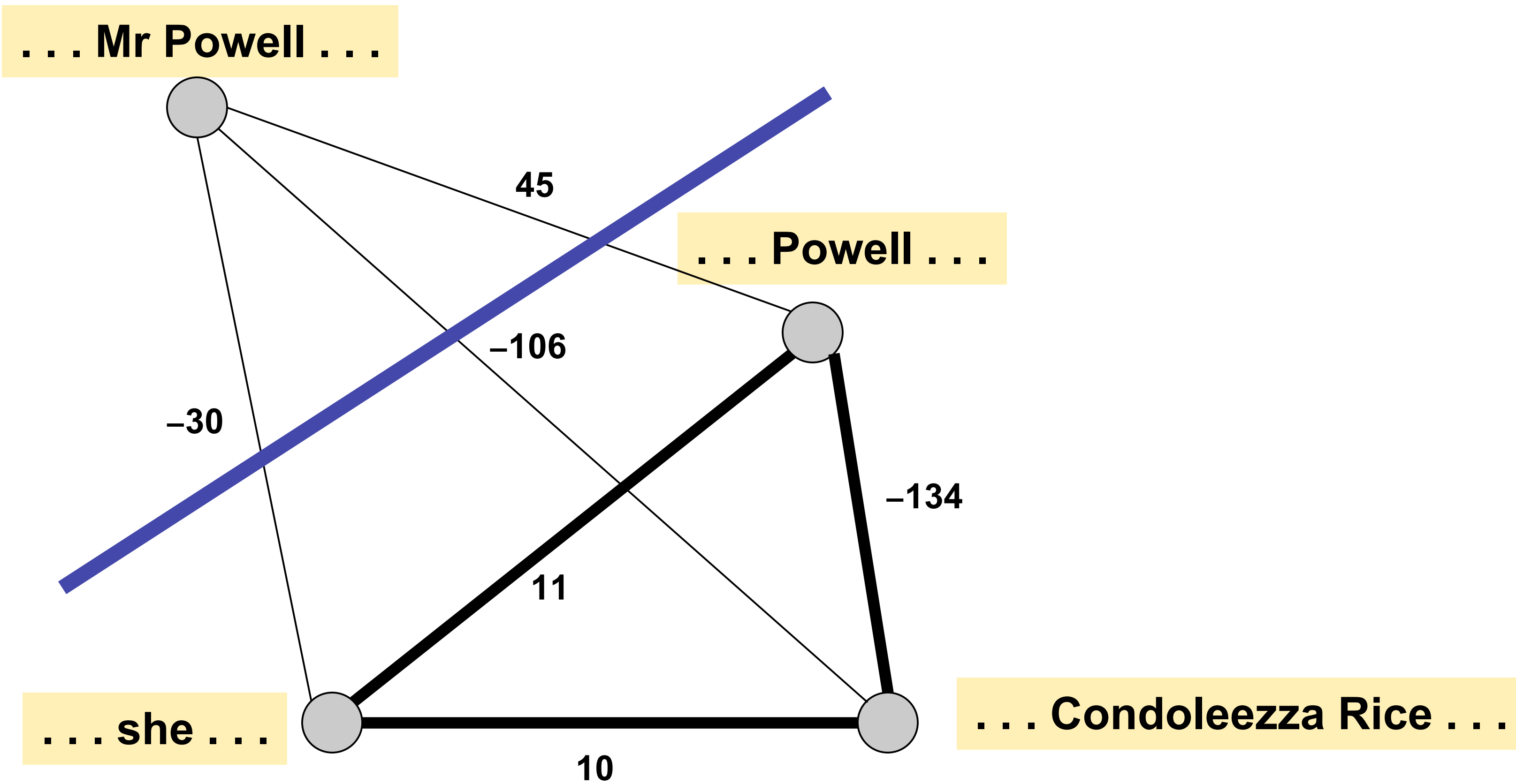
[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]



$$\log\left(P(\vec{y} \mid \vec{x})\right) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}} - \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}}$$

Inference in these MRFs = Graph Partitioning

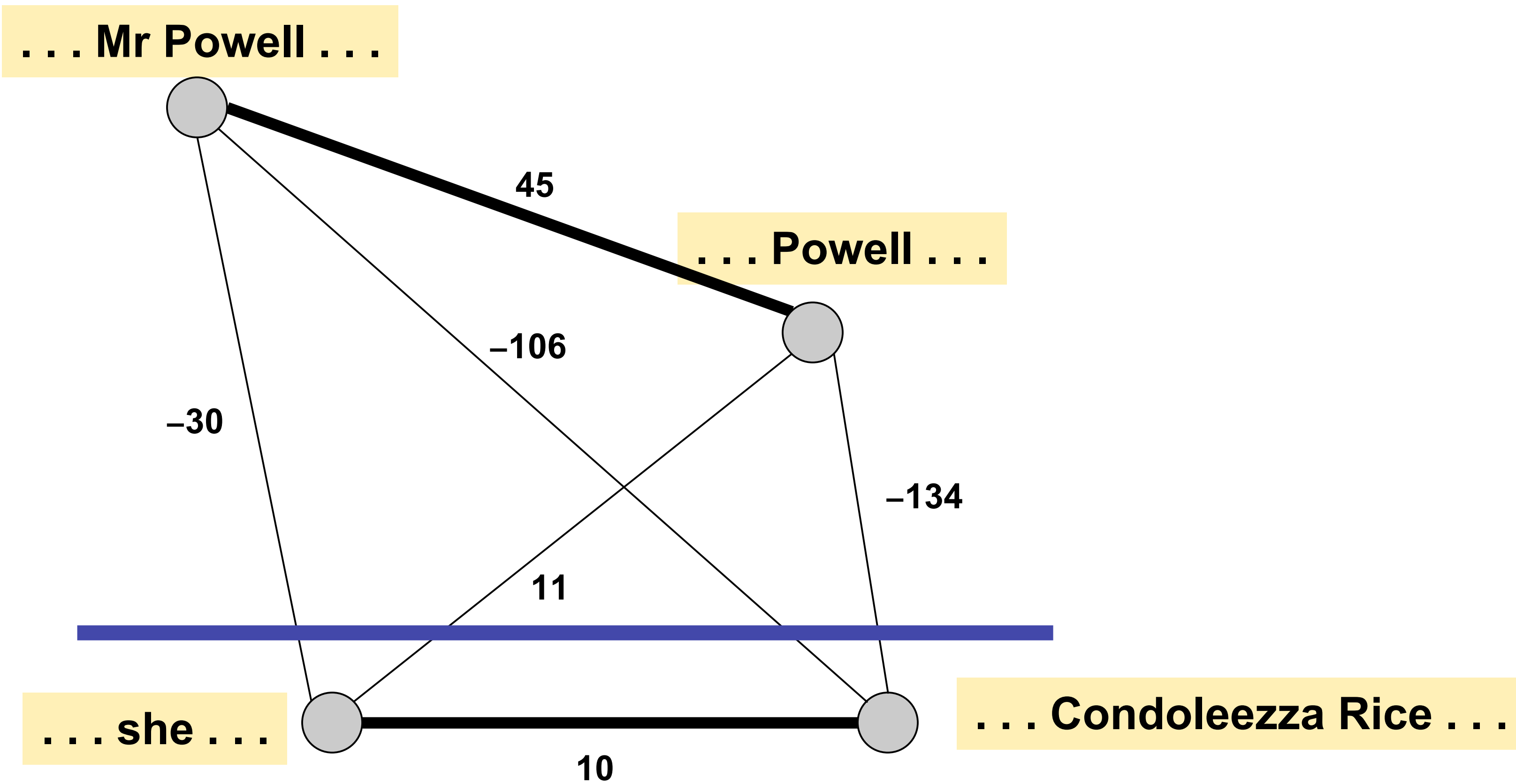
[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]



$$\log\big(P(\vec{y} \mid \vec{x})\big) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{ij} - \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w}_{ij} \quad = -22$$

Inference in these MRFs = Graph Partitioning

[Boykov, Vekler, Zabih, 1999], [Kolmogorov & Zabih, 2002], [Yu, Cross, Shi, 2002]



$$\log\left(P(\vec{y} \mid \vec{x})\right) \propto \sum_{i,j} \sum_{l} \lambda_l f_l(x_i, x_j, y_{ij}) = \sum_{\substack{i,j \text{ w/in} \\ \text{paritions}}} \mathrm{w}_{\mathrm{ij}} + \sum_{\substack{i,j \text{ across} \\ \text{paritions}}} \mathrm{w'}_{\mathrm{ij}} \quad = 314$$

Co-reference Experimental Results

[McCallum & Wellner, 2003]

Proper noun co-reference

DARPA ACE broadcast news transcripts, *117 stories*

	Partition F1	Pair F1
Single-link threshold	16 %	18 %
Best prev match [Morton]	83 %	89 %
MRFs	88 %	92 %
	Δerror=30%	Δerror=28%

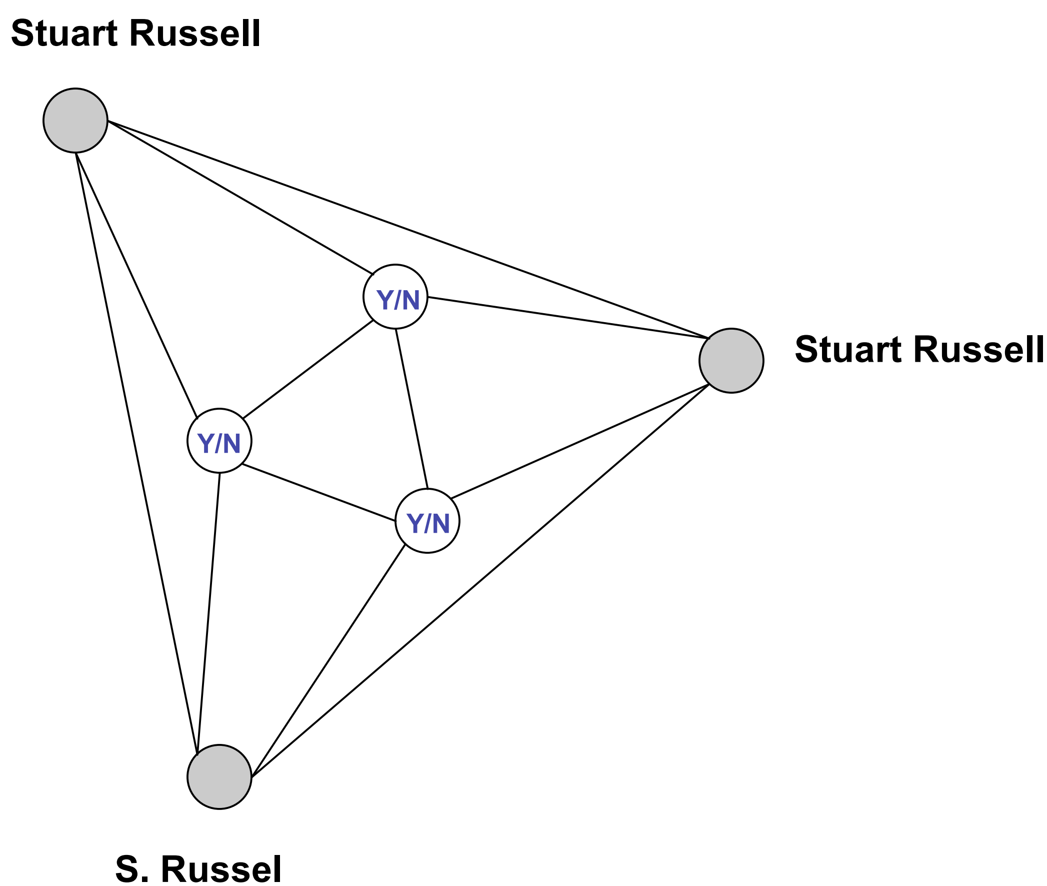
DARPA MUC-6 newswire article corpus, *30 stories*

	Partition F1	Pair F1
Single-link threshold	11%	7 %
Best prev match [Morton]	70 %	76 %
MRFs	74 %	80 %
	Δerror=13%	Δerror=17%

Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]

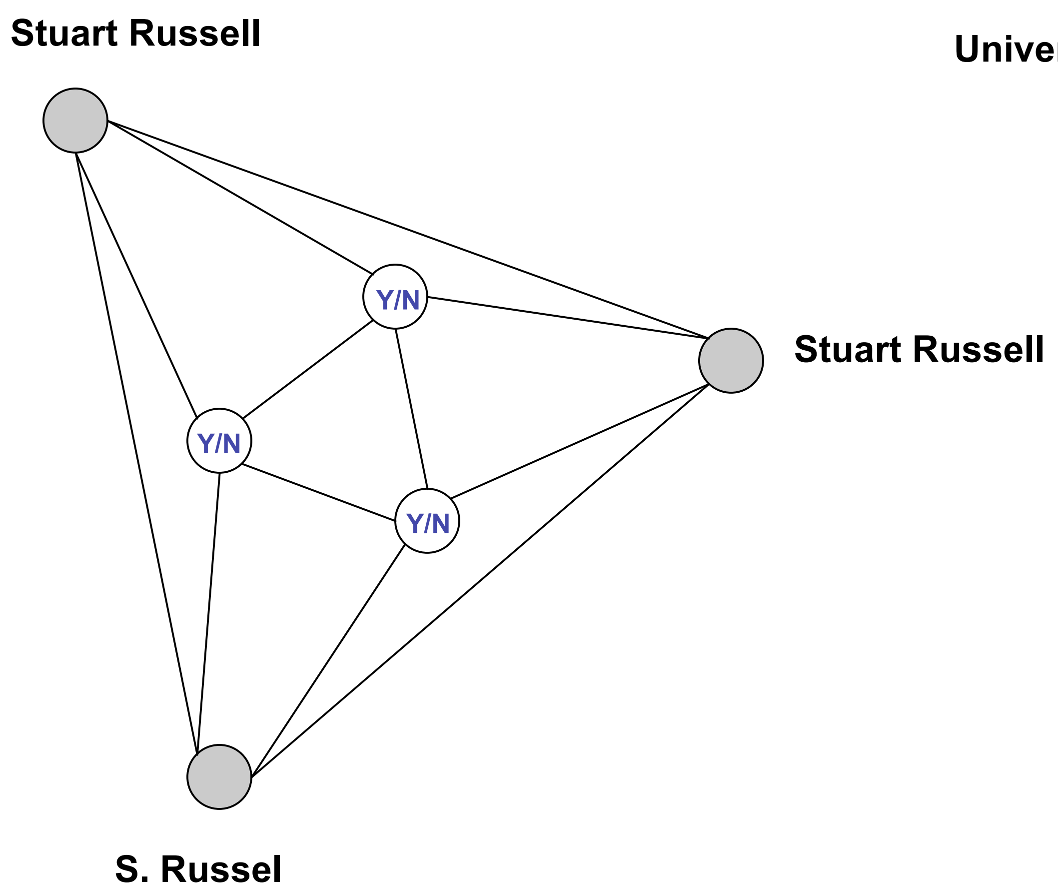
<u>People</u>



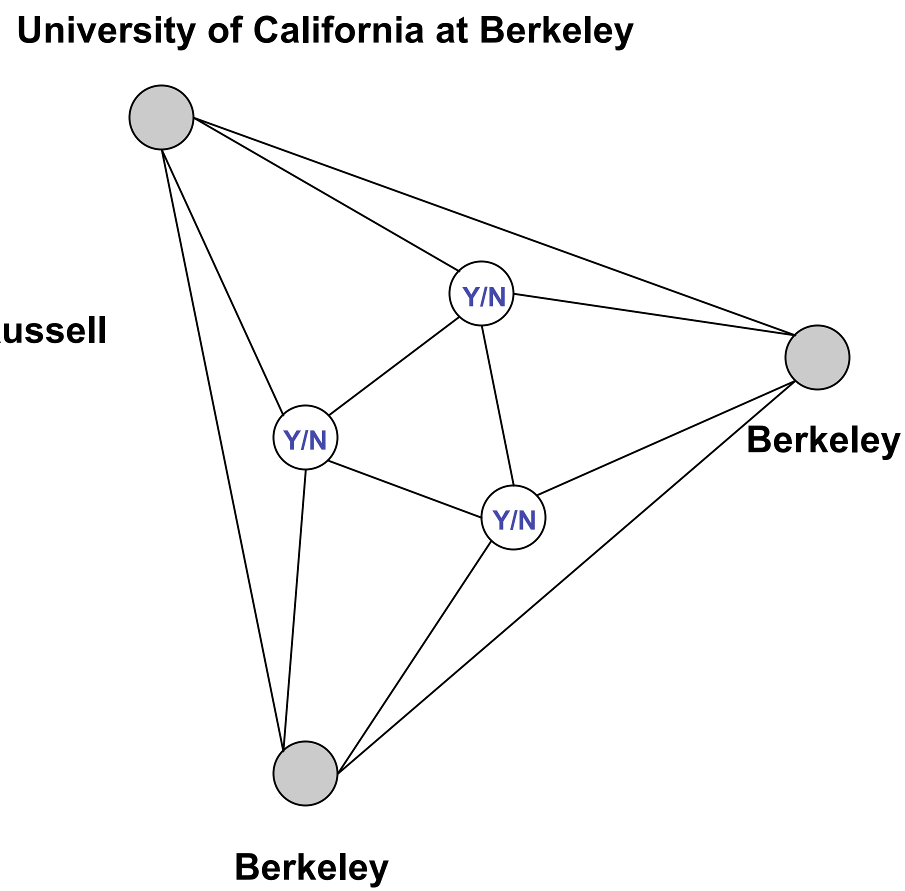
Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]

People

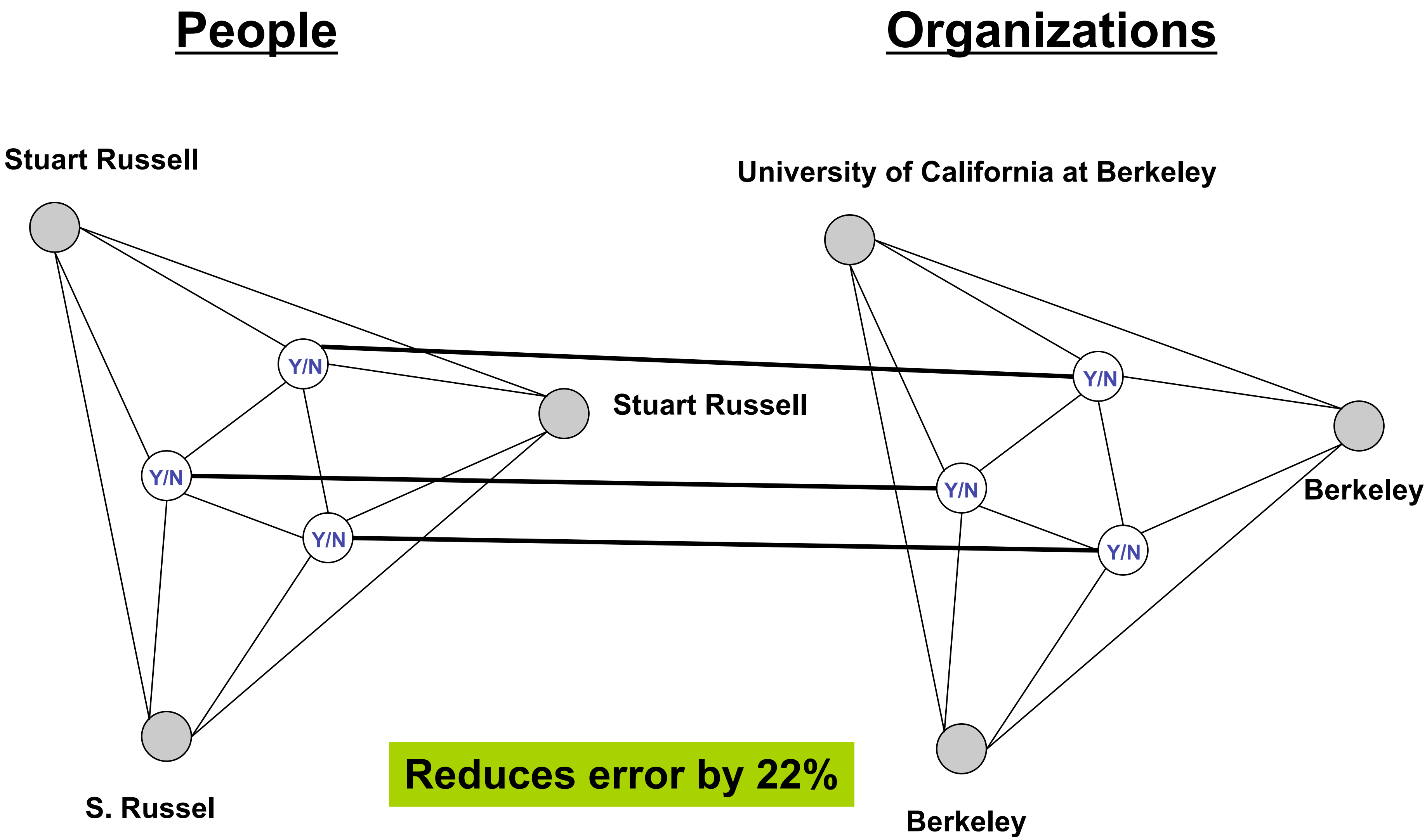


Organizations



Joint Co-reference for Multiple Entity Types

[Culotta & McCallum 2005]



Reduces error by 22%

The Canopies Approach

- Two distance metrics: cheap & expensive
- First Pass
  - very inexpensive distance metric
  - create overlapping canopies
- Second Pass
  - expensive, accurate distance metric
  - canopies determine which distances calculated

Main Points

- Important IE task
- Coreference as classification
- Coreference as CRF
- Joint resolution of different object type

Reference Matching

- Fahlman, Scott & Lebiere, Christian (1989). The cascade-correlation learning architecture. In Touretzky, D., editor, Advances in Neural Information Processing Systems (volume 2), (pp. 524-532), San Mateo, CA. Morgan Kaufmann.

- Fahlman, S.E. and Lebiere, C., "The Cascade Correlation Learning Architecture," NIPS, Vol. 2, pp. 524-532, Morgan Kaufmann, 1990.

- Fahlman, S. E. (1991) The recurrent cascade-correlation learning architecture. In Lippman, R.P. Moody, J.E., and Touretzky, D.S., editors, NIPS 3, 190-205.

The Citation Clustering Data

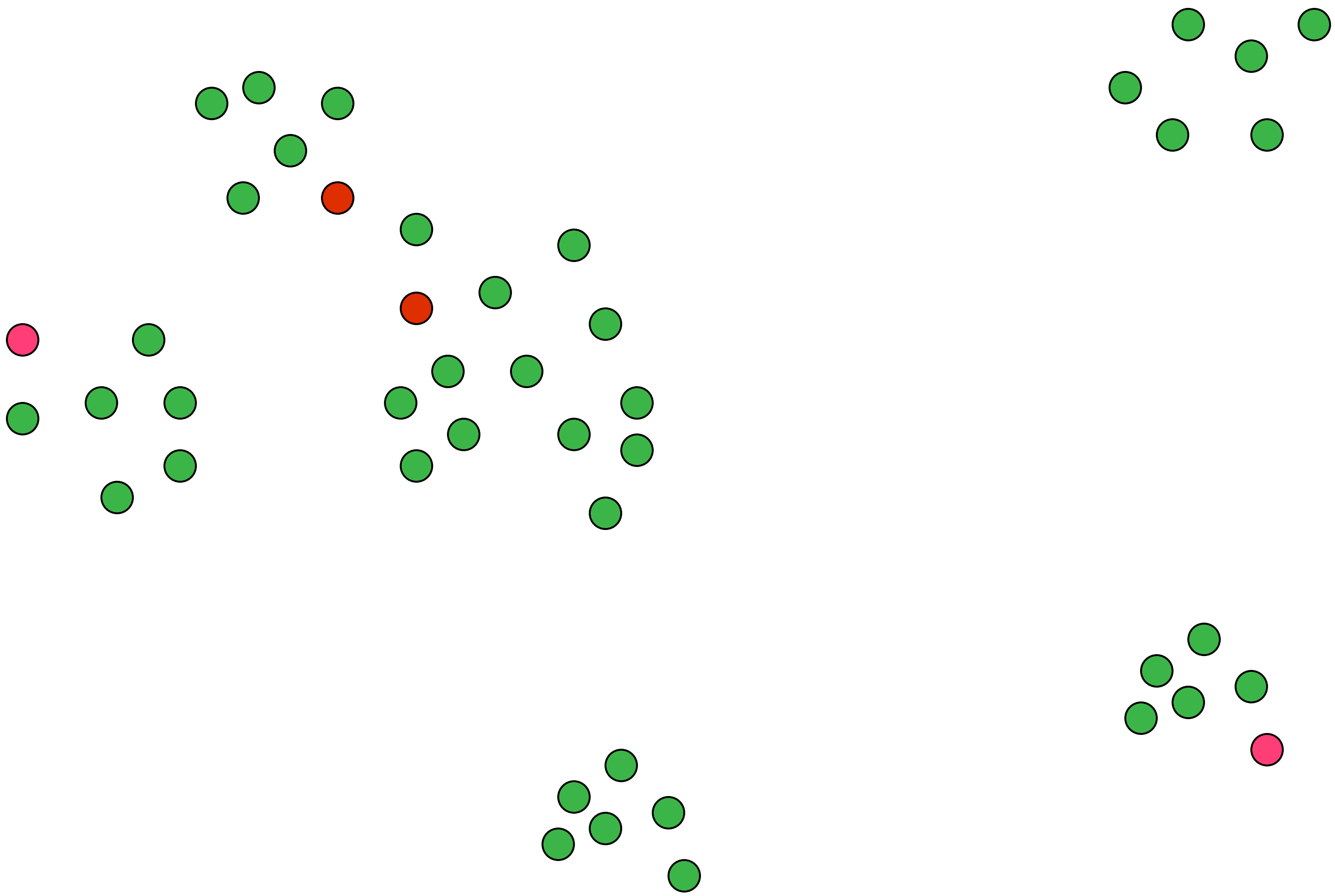
- Over 1,000,000 citations
- About 100,000 unique papers
- About 100,000 unique vocabulary words

- Over 1 trillion distance calculations

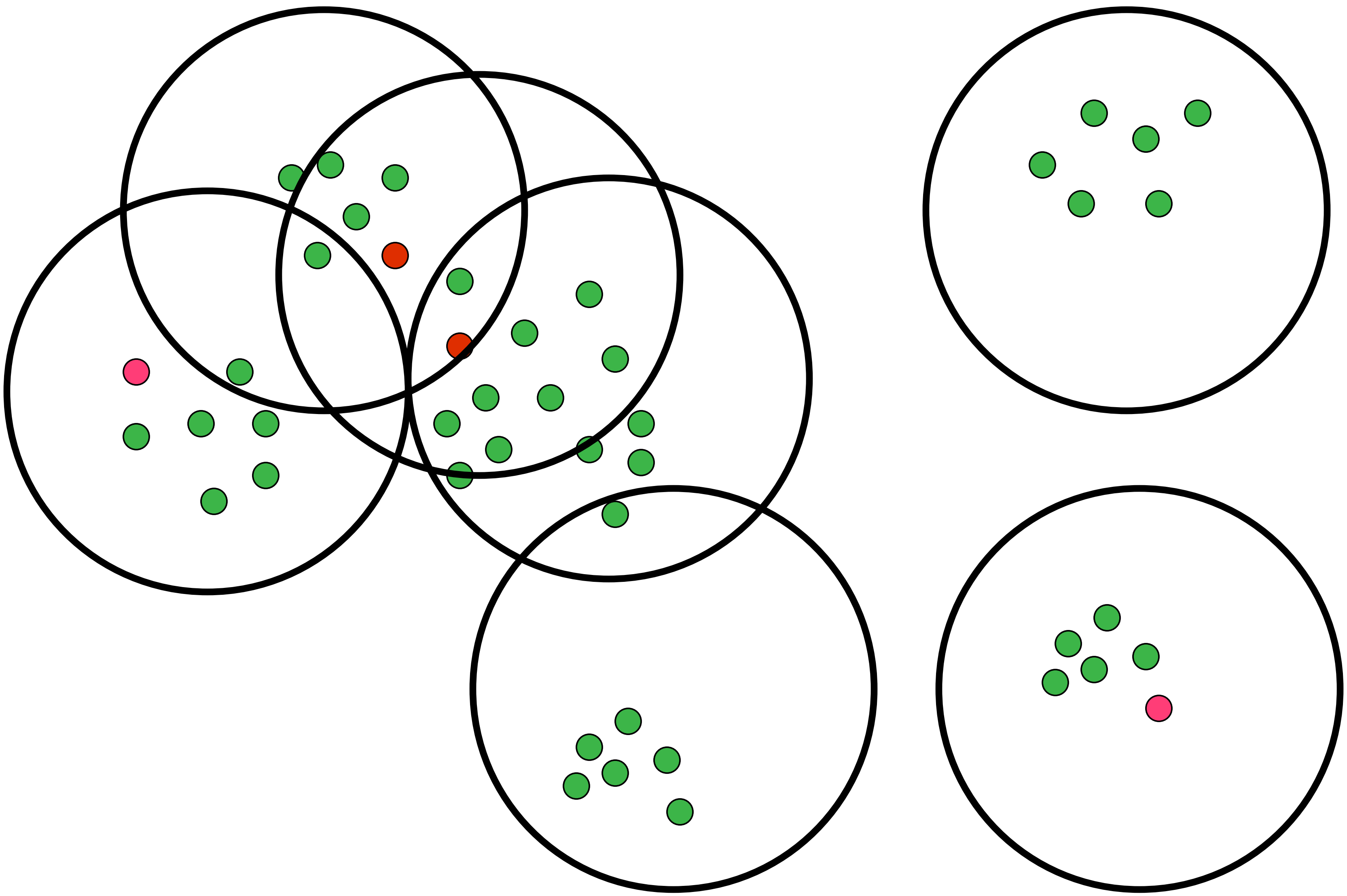
The Canopies Approach

- Two distance metrics: cheap & expensive
- First Pass
  - very inexpensive distance metric
  - create overlapping canopies
- Second Pass
  - expensive, accurate distance metric
  - canopies determine which distances calculated

Illustrating Canopies

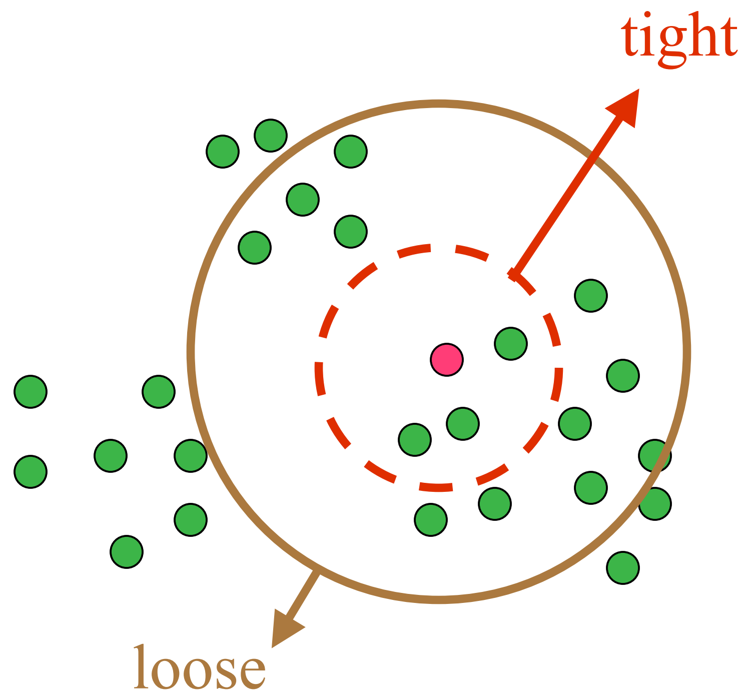


Overlapping Canopies



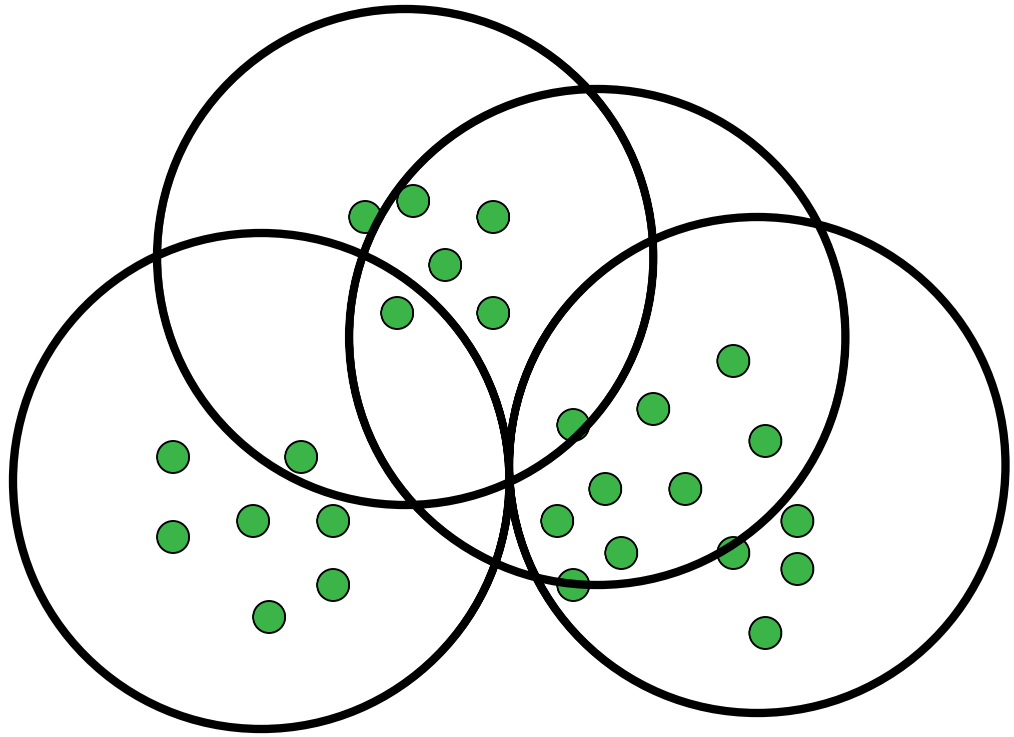
Creating canopies with two thresholds

- Put all points in D
- Loop:
  - Pick a point X from D
  - Put points within
  - K_{loose} of X in canopy
  - Remove points within
  - K_{tight} of X from D



Using canopies with Greedy Agglomerative Clustering

- Calculate expensive distances between points in the same canopy
- All other distances default to infinity
- Sort finite distances and iteratively merge closest



Computational Savings

- inexpensive metric << expensive metric
- # canopies per data point: f (small, but > 1)
- number of canopies: c (large)

- complexity reduction:

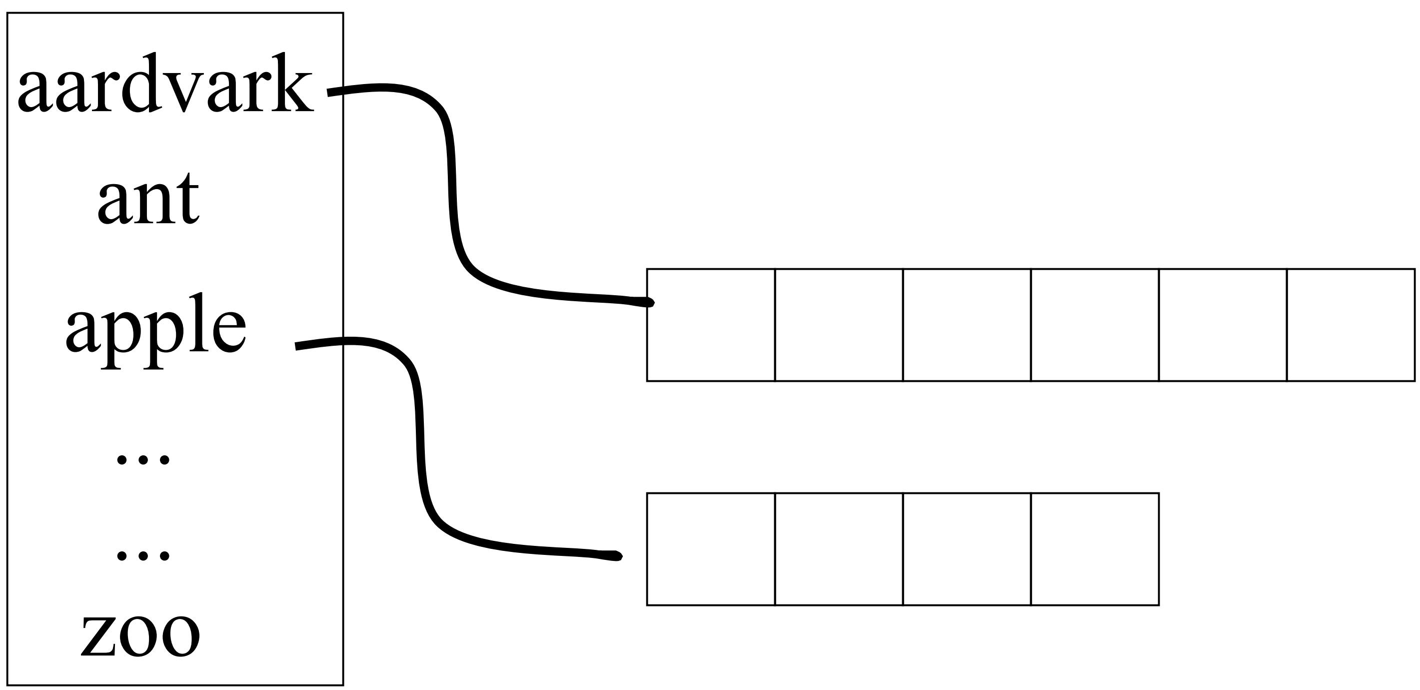
$$O\left(\frac{f^2}{c}\right)$$

The Experimental Dataset

- All citations for authors:
  - Michael Kearns
  - Robert Schapire
  - Yoav Freund
- 1916 citations
- 121 unique papers
- Similar dataset used for parameter tuning

Inexpensive Distance Metric for Text

- Word-level matching (TFIDF)
- Inexpensive using an inverted index



Expensive Distance Metric for Text

- String edit distance
- Compute with Dynamic Programming
- Costs for character:
  - insertion
  - deletion
  - substitution
  - ...

		S	e	c	a	t
	0.0	0.7	1.4	2.1	2.8	3.5
S	0.7	0.0	0.7	1.1	1.4	1.8
c	1.4	0.7	1.0	0.7	1.4	1.8
o	2.1	1.1	1.7	1.4	1.7	2.4
t	2.8	1.4	2.1	1.8	2.4	1.7
t	3.5	1.8	2.4	2.1	2.8	2.4

do Fahlman vs Falman

Experimental Results

	F1	Minutes
Canopies GAC	0.838	7.65
Complete GAC	0.835	134.09
Old Cora	0.784	0.03
Author/Year	0.697	0.03

Add precision, recall along side F1

Main Points

Co-reference

- How to cast as classification [Cardie]
- Measures of string similarity [Cohen]
- Scaling up [McCallum et al]