Accelerating Column Generation with Flexible Dual Optimal Inequalities with application to Entity Resolution

Vishnu Lokhande
University of Wisconsin-Madison*

Shaofei Wang
University of Pennsylvania

Maneesh Singh
Verisk

Julian Yarkony
Verisk

SERVE | ADD VALUE | INNOVATE

*work done during internship at Verisk

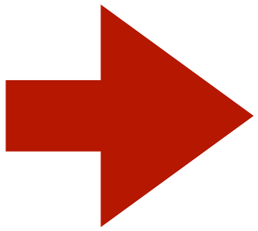
Verisk

The Entity Resolution Problem

Fields

Observations

Name	Place	SSN
Geoff Hinton	Toronto	556
Yoshua Bengio	Montreal	223
Geoffrey Hinton	Ontario	556
Yann LeCun	NYC	361
Hinton	Canada	-
Y. Lecun	New York	367



Name	Place	SSN
Yoshua Bengio	Montreal,	223
Geoffrey	Toronto, Canada	556
Yann LeCun	NYC, USA	367

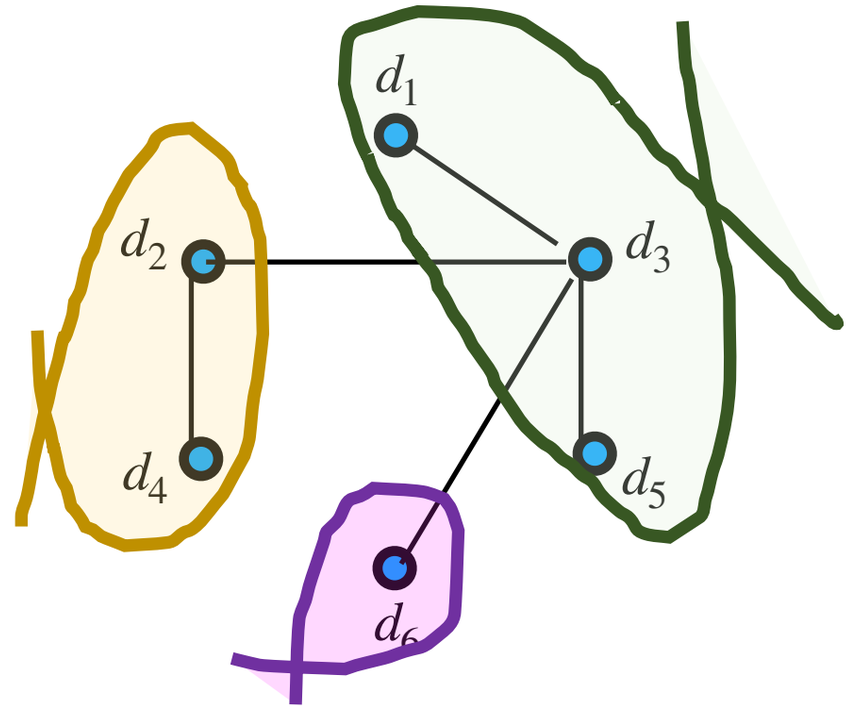
Entity resolution seeks to construct a surjection from observations in input dataset to real world entities

Entity Resolution as Graph Co-relation Clustering

Fields

Observations

Name	Place	SSN	
Geoff Hinton	Toronto	556	d_1
Yoshua Bengio	Montreal	223	d_2
Geoffrey Hinton	Ontario	556	d_3
Yann LeCun	NYC	361	d_4
Hinton	Canada	-	d_5
Y. Lecun	New York	367	d_6



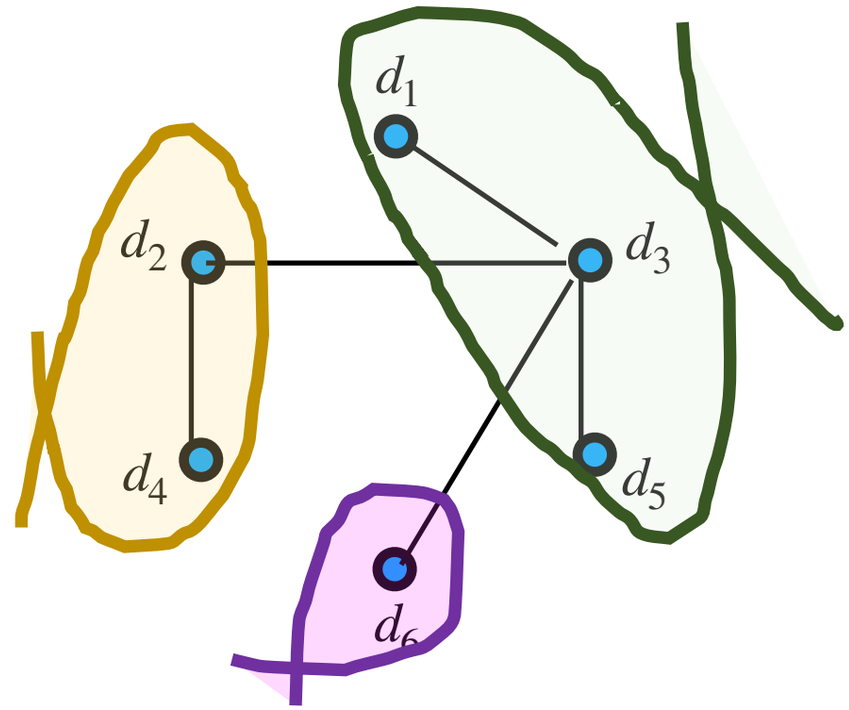
Embed observations into feature space and perform clustering !

Entity Resolution as Graph Co-relation Clustering

Fields

Observations

Name	Place	SSN	
Geoff Hinton	Toronto	556	d_1
Yoshua Bengio	Montreal	223	d_2
Geoffrey Hinton	Ontario	556	d_3
Yann LeCun	NYC	361	d_4
Hinton	Canada	-	d_5
Y. Lecun	New York	367	d_6



Embed observations into feature space and perform clustering !

Traditional Approaches
- Hierarchical Clustering
- Connected Components
- Star Clustering

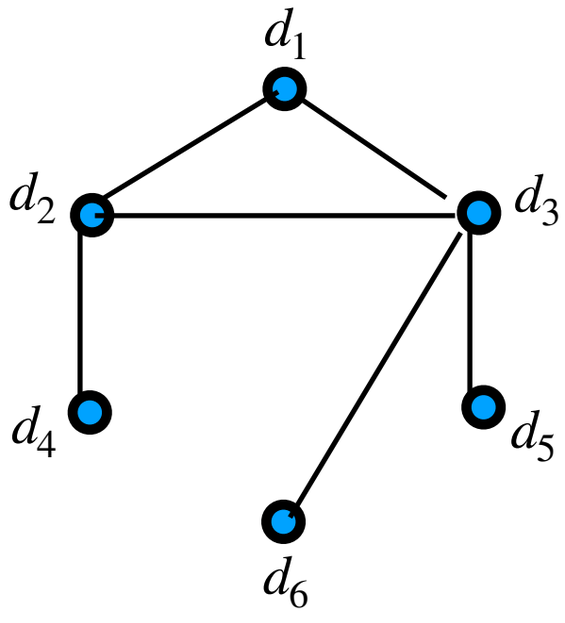
- Slow
- Clusters overlap
- Based on Heuristics

Our Approach
- Significant Speed-ups
- Non-overlapping Clusters
- Provable Guarantees

Observations, Hypothesis, Cost Terms, Etc.

Set of Observations $(\mathscr{D})$

- d_1
- d_2
- d_3
- d_4
- d_5
- d_6



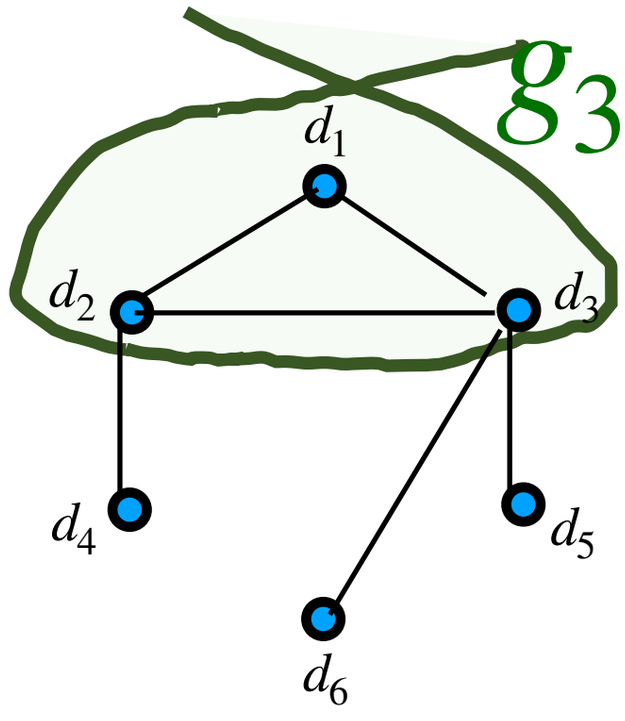
Set of Hypothesis $(\mathscr{G})$

$g_1 = \{d_1, d_2\}$

$g_2 = \{d_3, d_4\}$

$g_3 = \{d_1, d_2, d_3\}$

$g_{2^6} = \{d_1, d_2, d_3, d_4, d_5, d_6\}$



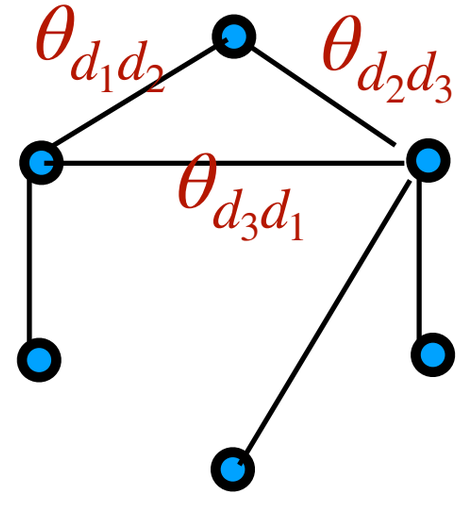
Cost Terms

$\Gamma_{g_1} = \theta_{d_1 d_2}$

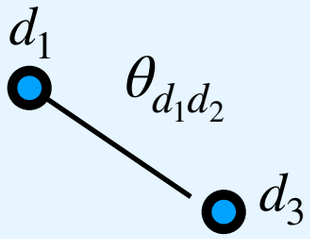
$\Gamma_{g_2} = \theta_{d_3 d_4}$

$\Gamma_{g_3} = \theta_{d_1 d_2} + \theta_{d_2 d_3} + \theta_{d_3 d_1}$

$\Gamma g_{2^6} = \theta_{d_1 d_2} + \theta_{d_2 d_3} + \ldots + \theta_{d_6 d_1}$



More on the Cost Function for Entity Resolution



$\theta_{d_1d_2}$:= **The similarity measure of** d_1 **and** d_2

$G_{dg} = 1$ **indicates** $d \in g$

$$\Gamma_g = \sum_{d_1,d_2} \theta_{d_1d_2} G_{d_1g} G_{d_2g}$$

The structural property of our problem suggests that most pairs of observations are unassociated

$$\theta_{d_1d_2} = \infty$$

Smaller Γ_g $\Longrightarrow$ **Smaller hypothesis cost** $\Longrightarrow$ **Good Cluster**

Larger Γ_g $\Longrightarrow$ **Larger hypothesis cost** $\Longrightarrow$ **Bad Cluster**

The Minimum Weight Set Packing Formulation

$\gamma_g = 1$ **Hypothesis g is selected**

$\gamma_g = 0$ **Hypothesis g is NOT selected**

$$\min_{\gamma \in \{0,1\}} \sum_{g \in \mathcal{G}} \Gamma_g \gamma_g$$

$$\sum_{g \in \mathcal{G}} G_{dg} \gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$

1. Non-overlapping hypothesis
2. Minimum total cost

Introducing Column Generation

$$\min_{\gamma \in \{0,1\}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g$$

1. MWSP is NP-Hard (Karp 1972)

$$\sum_{g \in \mathcal{G}} G_{dg} \gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$

2. $\mathcal{G}$ is too large to enumerate

Introducing Column Generation

$$\gamma \geq 0 \quad \min_{\gamma \in \{0,1\}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g$$

1. MWSP is NP-Hard (Karp 1972)

$$g \in \hat{\mathcal{G}} \quad \sum_{g \in \mathcal{G}} G_{dg} \gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$

2. $\mathcal{G}$ is too large to enumerate

Column Generation

Construct a small sufficient subset $\hat{\mathcal{G}}$ such that an optimal solution exists using only the hypothesis in $\hat{\mathcal{G}}$

The Column Generation Algorithm

ITERATE

STEP 1 (SOLVE): Relax the ILP and restrict it to $\hat{\mathcal{G}} \subset \mathcal{G}$

Restricted Primal

$$\min_{\gamma \geq 0} \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g$$

$$\sum_{g \in \hat{\mathcal{G}}} G_{dg} \gamma_g \leq 1 \quad \forall d \in \mathcal{D}$$

$\Leftrightarrow$

Restricted Dual

$$\max_{\lambda_d \leq 0} \sum_{d \in \mathcal{D}} \lambda_d$$

$$\Gamma_g - \sum_{d \in \mathcal{D}} G_{dg} \lambda_d \geq 0 \quad \forall g \in \hat{\mathcal{G}}$$

STEP 2 (GROW): Find g* from the pricing problem $\hat{\mathcal{G}} \leftarrow \hat{\mathcal{G}} \cup g^*$

The Pricing Problem

$$g^* \leftarrow \min_{g \in \mathcal{G}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg}$$

Pricing Problem made Efficient

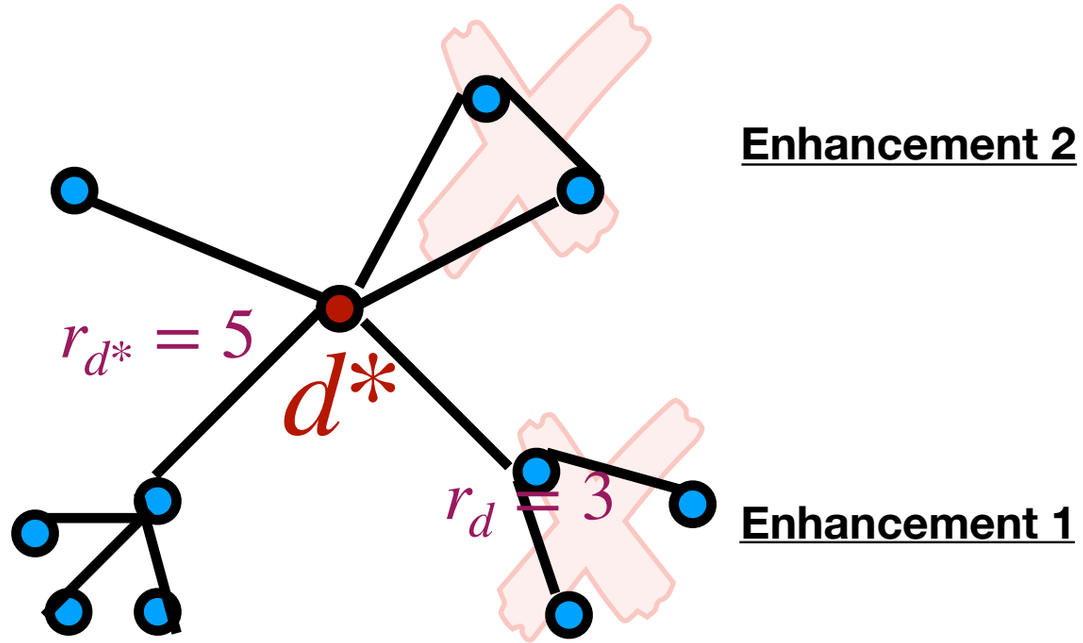
The Pricing Problem

$$\min_{g \in \mathcal{G}} \Gamma_g - \sum_{d \in \mathcal{D}} \lambda_d G_{dg}$$

Expensive operation as it involves min over $\mathcal{G}$

IDEA! Break down the problem into smaller subproblems and solve them in parallel.

A sub-problem is indexed with a central node d^* and has at most one neighborhood



Enhancement 1: Remove leaves with lower neighborhood cardinality than the centre node

Enhancement 2: Remove leaves that form superfluous subproblems

TWO VARIANTS

Exact Pricing: Formulate and solve an Integer Linear Program

Heuristic Pricing: Terminate the pricing early and solve approximately (QPBO)

Varying Dual Optimal Inequalities

Dual

$$-\Xi_d \leq \max_{\lambda_d \leq 0} \sum_{d \in \mathcal{D}} \lambda_d$$

$$\Gamma_g - \sum_{d \in \mathcal{D}} G_{dg} \lambda_d \geq 0$$

Reduce feasible region while preserving an optimal solution provably

Primal

$$\min_{\substack{\gamma \geq 0 \\ \xi \geq 0}} \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g + \sum_{d \in \mathcal{D}} \Xi_d \xi_d$$

$$-\xi_d + \sum_{g \in \hat{\mathcal{G}}} G_{dg} \gamma_g \leq 1$$

Relax constraints and add a soft penalty
Penalty provably inactive at termination

$\xi_d :=$ **Number of violation for** d

$\Xi_d \geq$ **Increase in cost of removing** d **from** g **or subsets of** g

Flexible Dual Optimal Inequalities (F-DOIs)

$$\min_{\substack{\gamma \geq 0 \\ \xi \geq 0}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g + \sum_{d \in \mathcal{D}} \Xi_d \xi_d$$

$$-\xi_d + \sum_{g \in \hat{\mathcal{G}}} G_{dg} \gamma_g \leq 1$$

$$\min_{\substack{\gamma \geq 0 \\ \xi \geq 0}} \quad \sum_{g \in \hat{\mathcal{G}}} \Gamma_g \gamma_g + \sum_{d \in \mathcal{D}} \sum_{g \in \hat{\mathcal{G}}} \Xi_{dg} \xi_{dg}$$

$$-\xi_{dg} + \sum_{g \in \hat{\mathcal{G}}} G_{dg} \gamma_g \leq 1$$

$$\xi_{dg} \leq G_{dg} \gamma_g$$

*** In Symbols**

$\xi_d :=$ **Number of violation for** d

$\xi_{dg} :=$ **Number of violation for** d **in** g

Augment the primal problem with ξ_{dg}

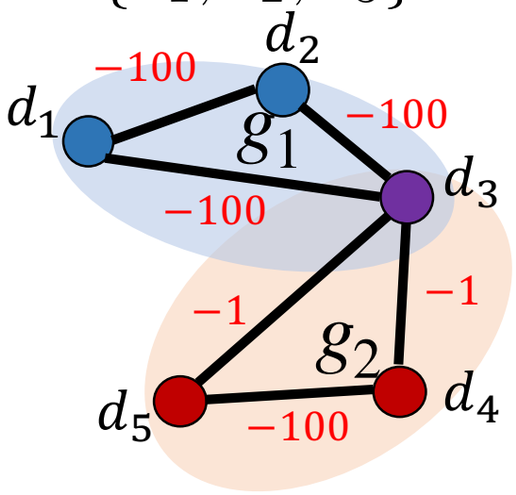
Lower Objective Achieved !
CG converges faster

In practice, quantize and round up Ξ_{dg} **for computational tractability**

*** In words**

F-DOIs : An example

$\mathcal{D}_{g_1} = \{d_1, d_2, d_3\}$



$\mathcal{D}_{g_2} = \{d_3, d_4, d_5\}$

	Feasible?	Optimal at current iter?	
No-DOI	✗	✗	g_1 **selected** **Cost** $= -300$
Varying DOI	✓	✗	g_1 **and** g_2 **selected** **Cost** $= -300 - 102 + 200 = -202$
F-DOI	✓	✓	g_1 **and** g_2 **selected** **Cost** $= -300 - 102 + 2 = -400$

The Complete Pipeline

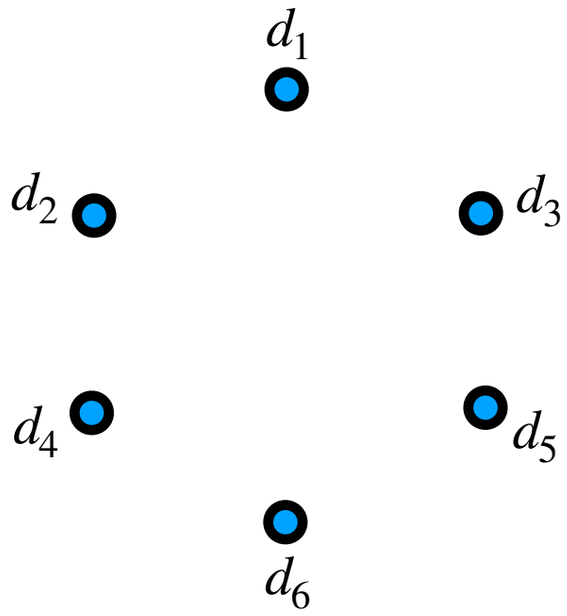
Input Dataset

Name	Place	Age	
Geoff Hinton	Toronto	56	d_1
Yoshua	Montreal	55	d_2
Geoffrey	Ontario	69	d_3
Yann LeCun	NYC	61	d_4
Hinton	Canada	78	d_5
Y. Lecun	New York	94	d_6

The Complete Pipeline

Input Dataset

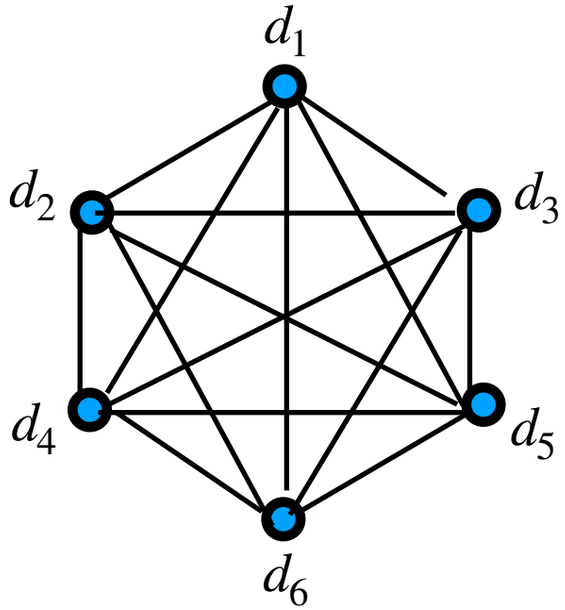
Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94



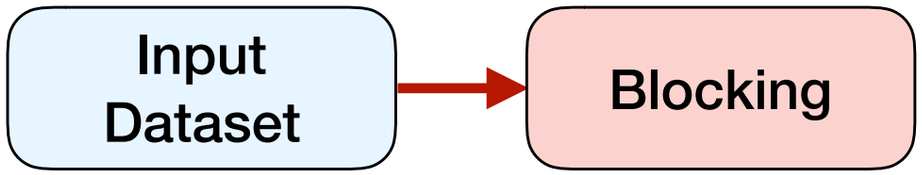
The Complete Pipeline

Input Dataset

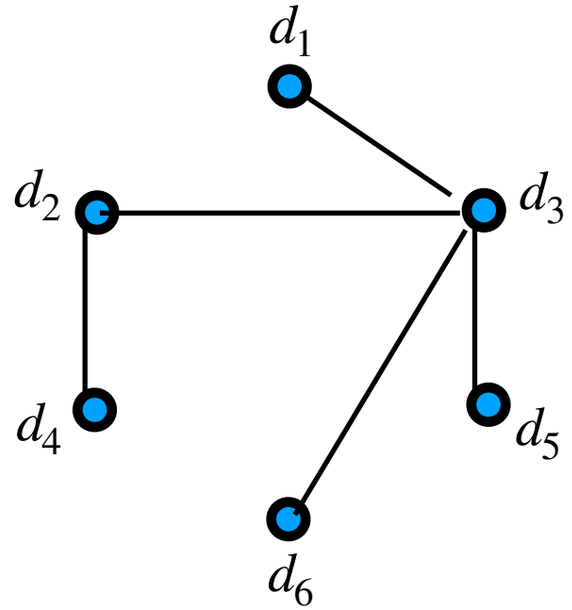
Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94



The Complete Pipeline

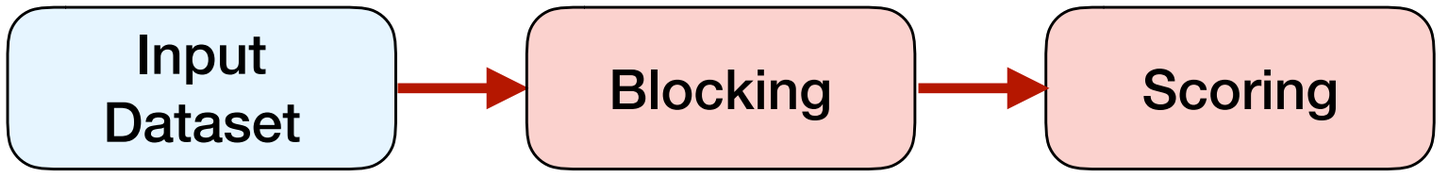


Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94

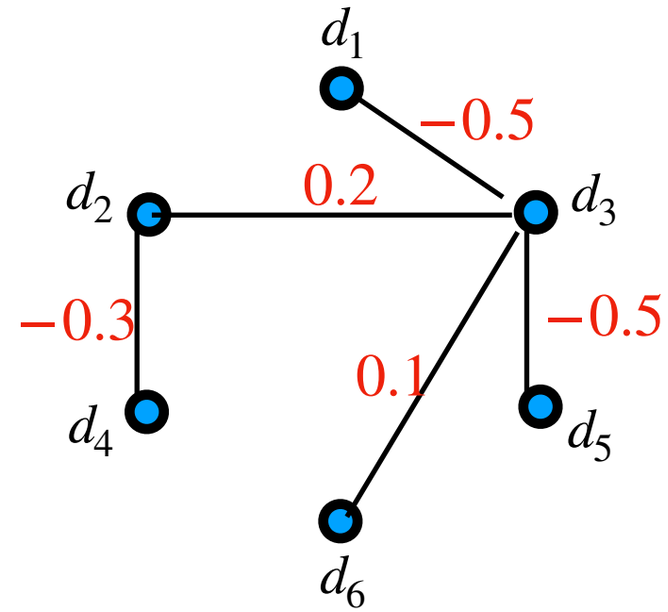


Blocking: Remove obvious non-matches by comparing a certain key attribute
Scoring: A score represents how close one observation is to another.
Clustering: With MWSP and F-DOIs

The Complete Pipeline

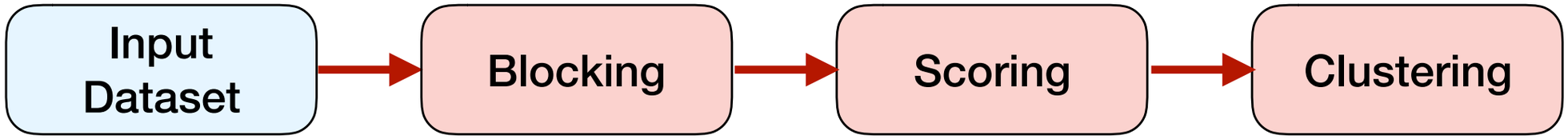


Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94

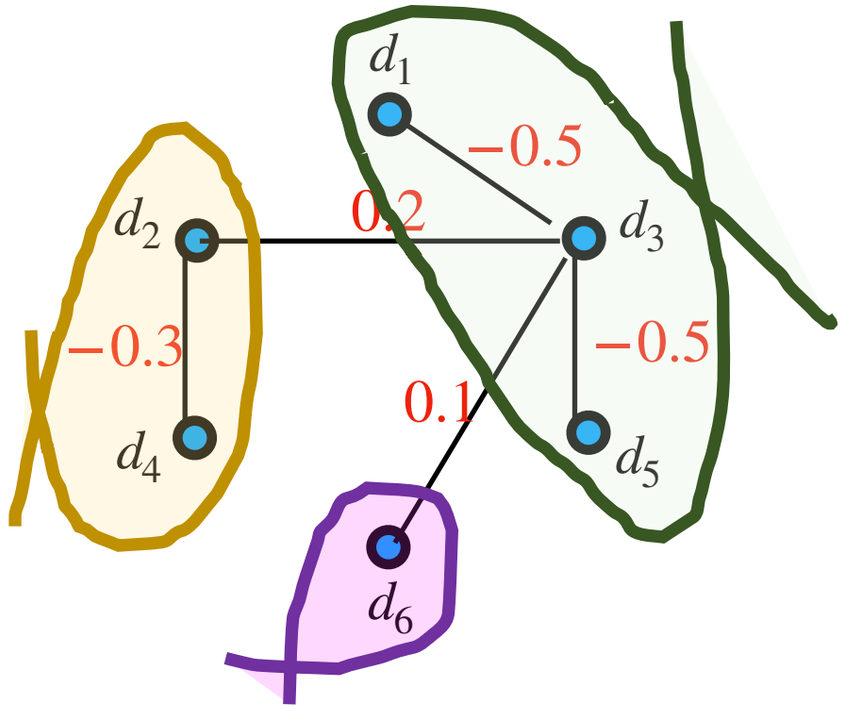


Blocking: Remove obvious non-matches by comparing a certain key attribute
Scoring: A score represents how close one observation is to another.
Clustering: With MWSP and F-DOIs

The Complete Pipeline

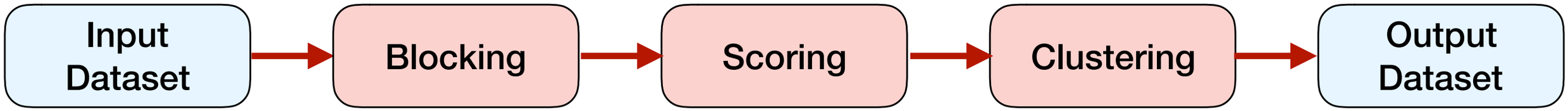


Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94

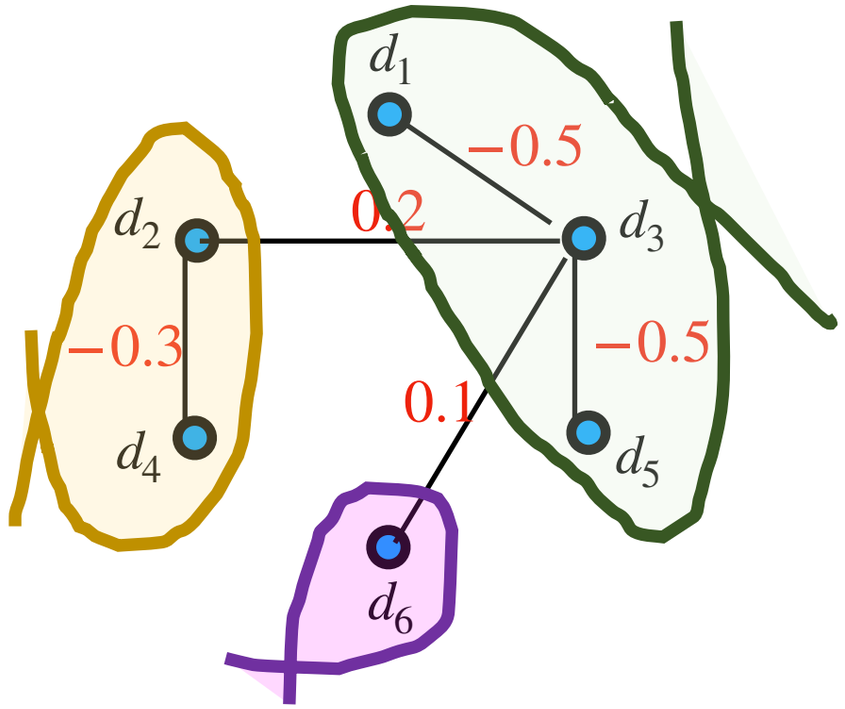


Blocking: Remove obvious non-matches by comparing a certain key attribute
Scoring: A score represents how close one observation is to another.
Clustering: With MWSP and F-DOIs

The Complete Pipeline

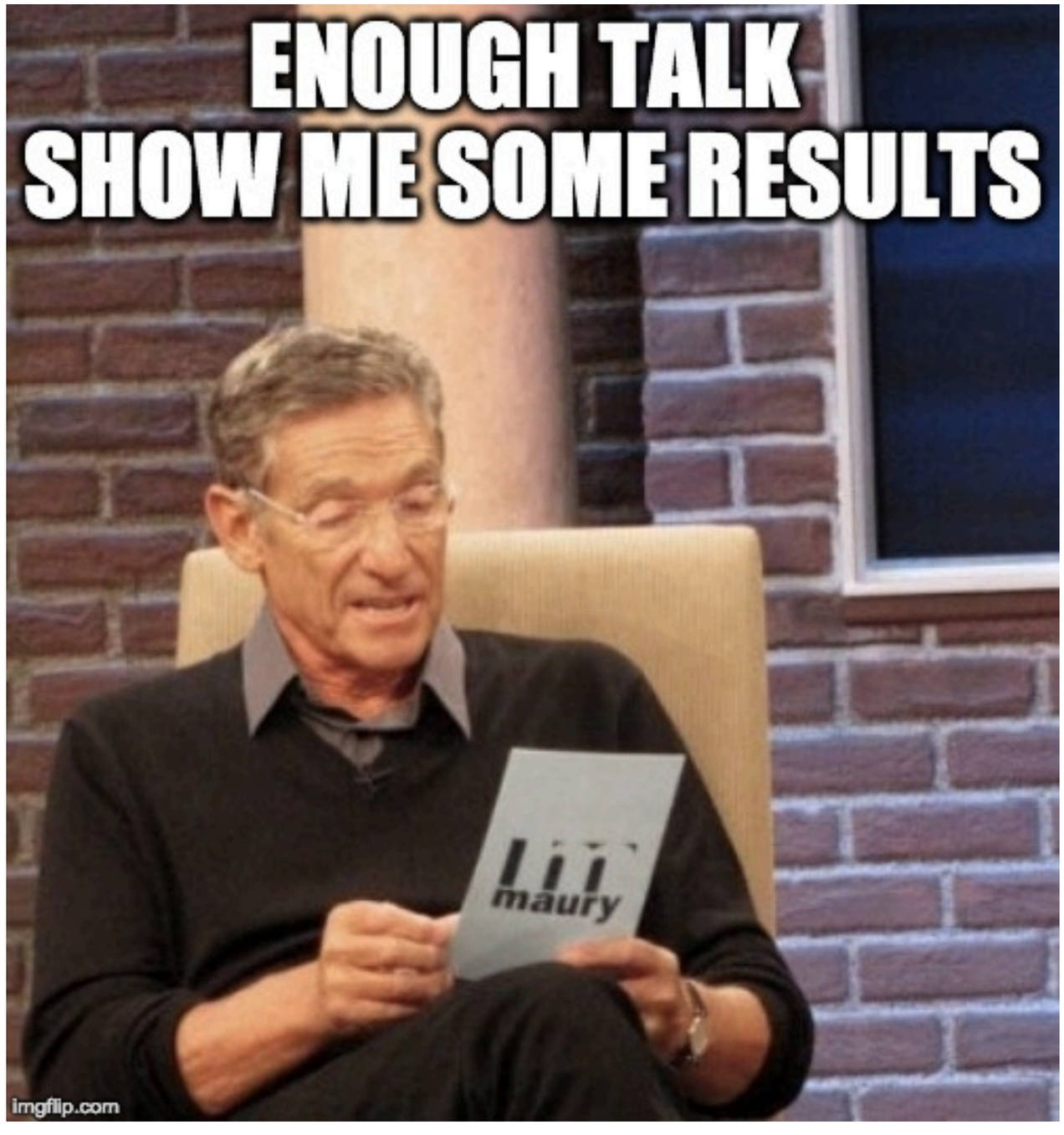


Name	Place	Age
Geoff Hinton	Toronto	56
Yoshua	Montreal	55
Geoffrey	Ontario	69
Yann LeCun	NYC	61
Hinton	Canada	78
Y. Lecun	New York	94



Name	Place	Ag
Yoshua	Montreal,	55
Geoffrey	Toronto,	69
Yann LeCun	NYC, USA	56

Blocking: Remove obvious non-matches by comparing a certain key attribute
Scoring: A score represents how close one observation is to another.
Clustering: With MWSP and F-DOIs



Results

- Significant speed-ups owing to Flexible DOI
  Upto 60% improvement compared to varying DOI
- Tractable solutions to the pricing problem
  Compute time reduced from 1 hour to 20 secs with efficient pricing
- **SOTA accuracies on two benchmark datasets**

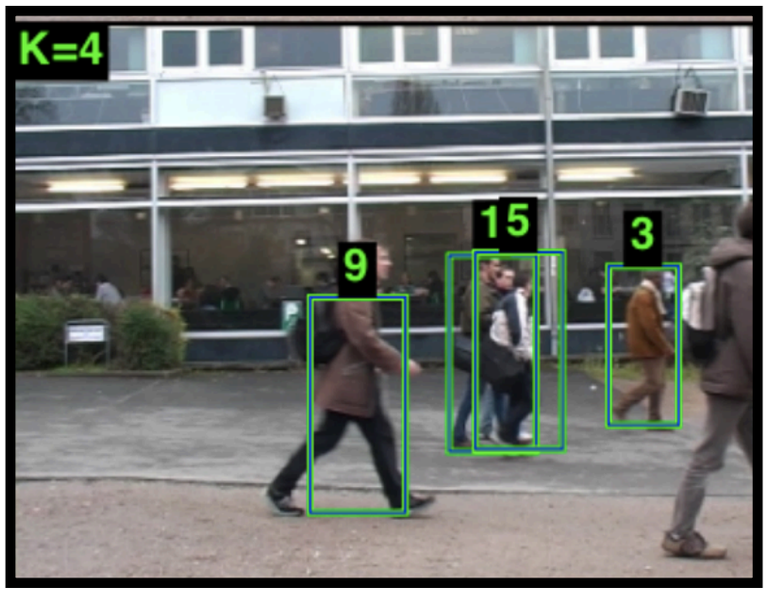
Datasets

Settlements - 3054 observations - 820 clusters

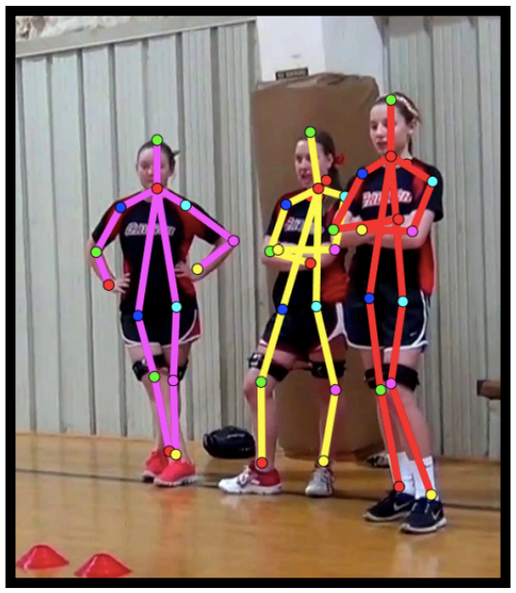
Music 20K - 19375 observations - 10000 clusters

Method	Settlements	Music 20K
ConCom	0.65	0.26
CCPivot	0.90	0.74
Center	0.88	0.66
MergCenter	0.68	0.39
Star1	0.82	0.62
Star2	0.92	0.69
F-MWSP	**0.96**	**0.81**

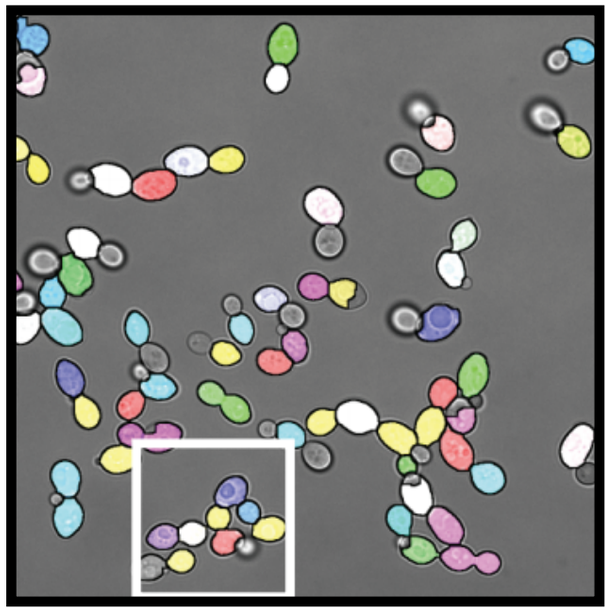
Column Generation on Problems outside Entity Resolution?



Multi-person tracking



Multi-person pose estimation



Multi-cell segmentation

Vishnu Lokhande

Shaofei Wang

Catch us at the Poster Session this evening

Maneesh Singh

Julian Yarkony