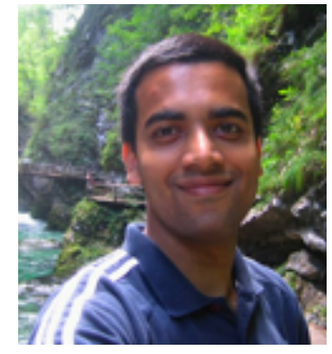
A Tutorial on

Graph-based Semi-Supervised Learning Algorithms for Speech and Spoken Language Processing



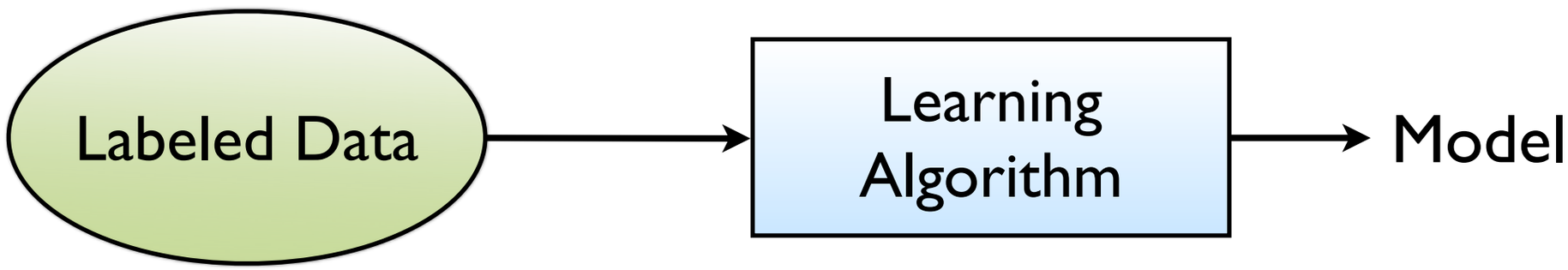
Amarnag Subramanya
(Google Research)

Partha Pratim Talukdar
(Carnegie Mellon University)

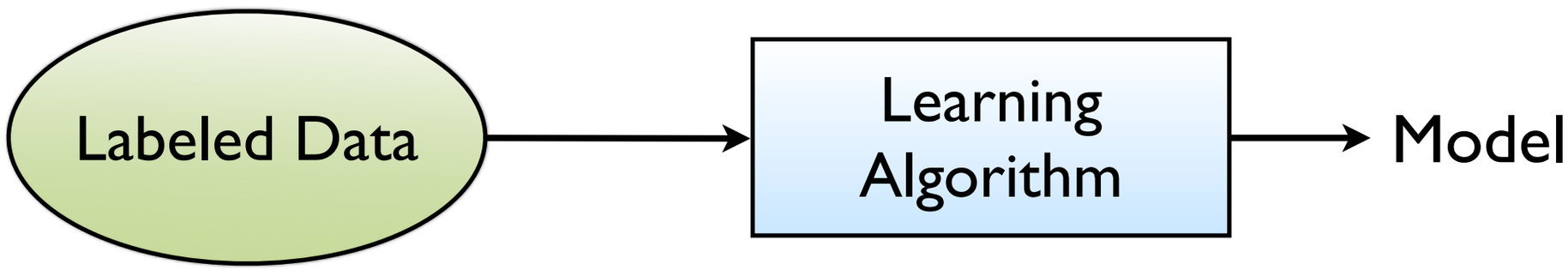
ICASSP 2013, Vancouver

http://graph-ssl.wikidot.com/

Supervised Learning



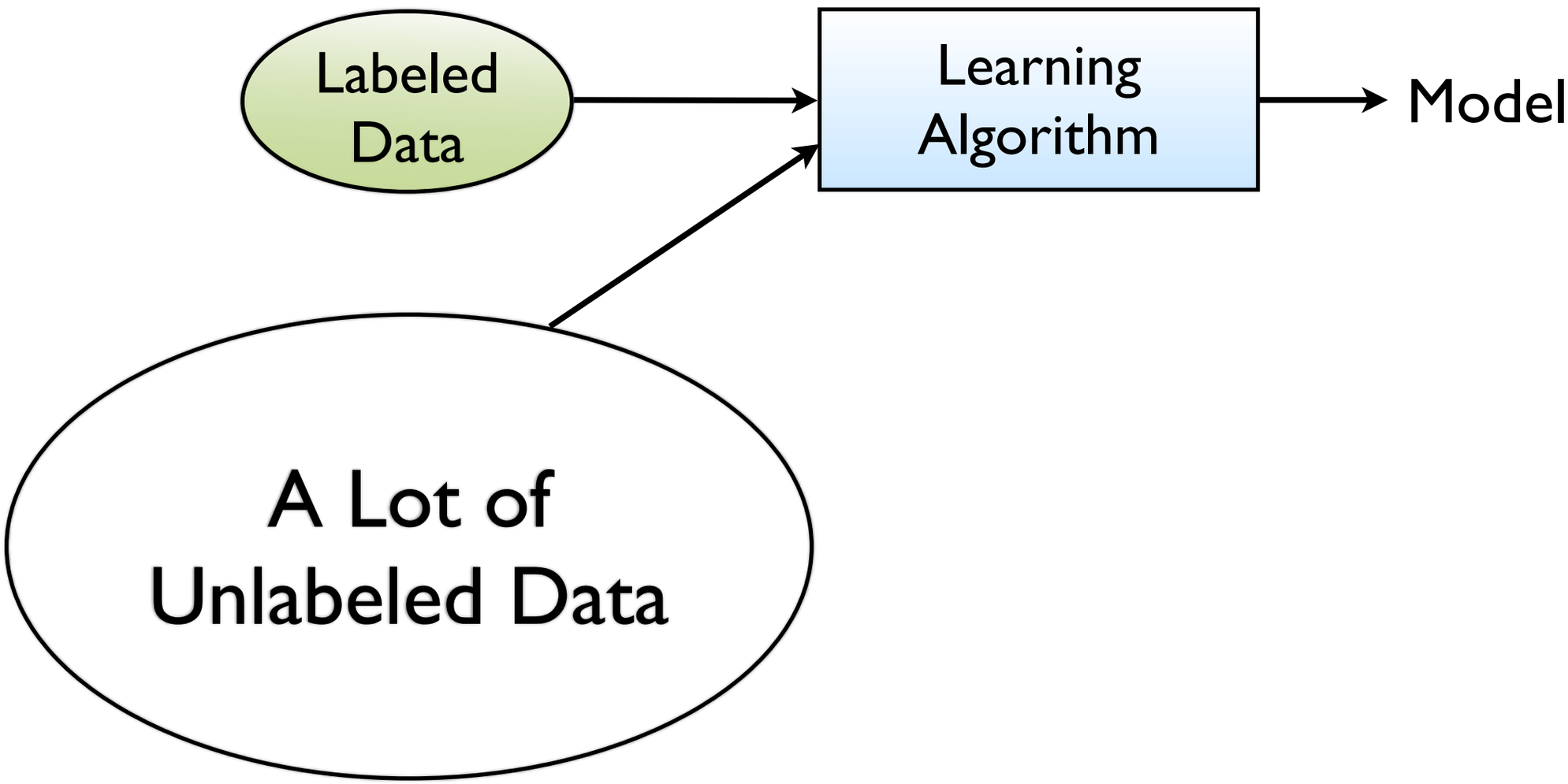
Supervised Learning



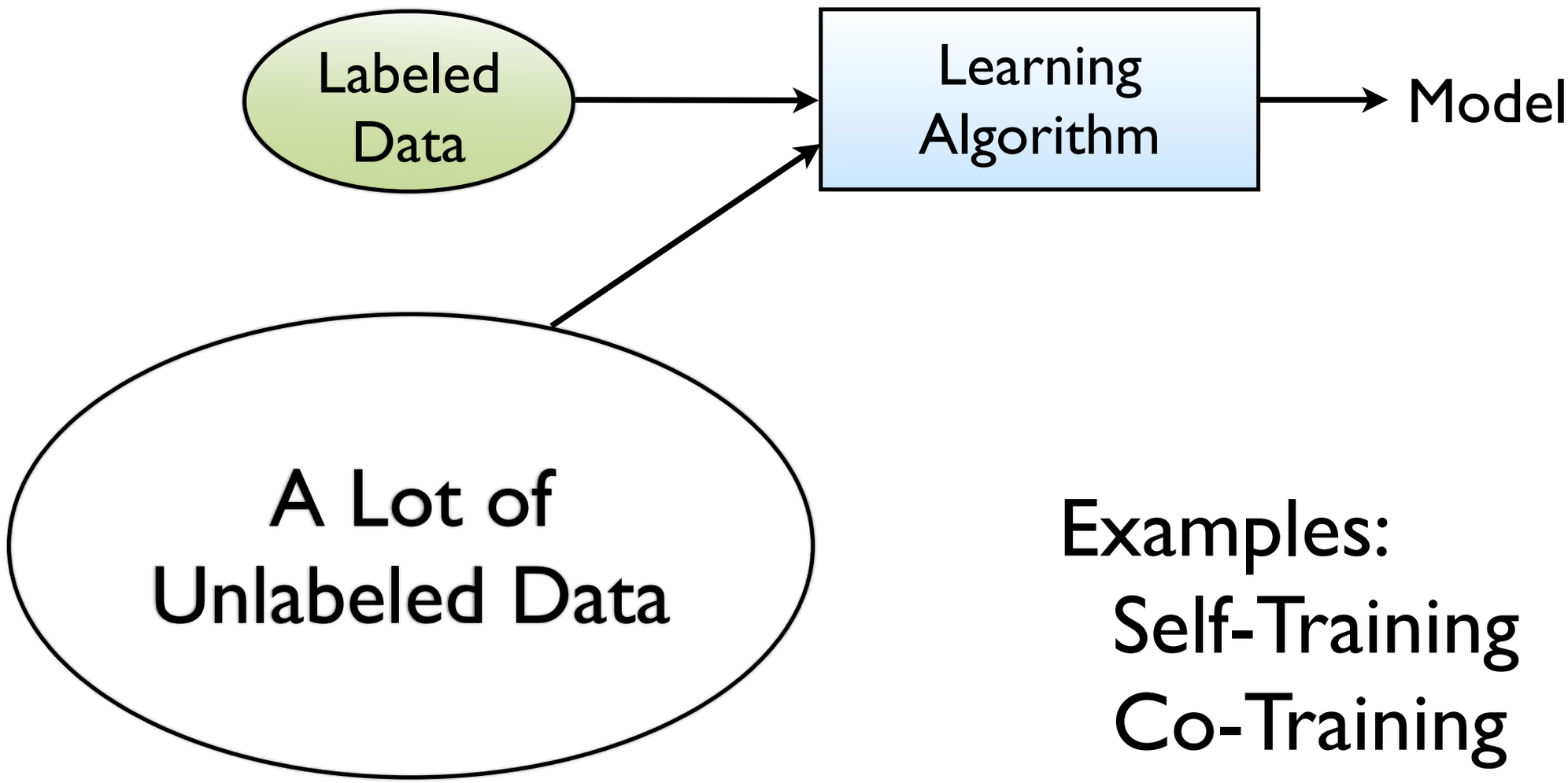
Examples:

- Decision Trees
- Support Vector Machine (SVM)
- Maximum Entropy (MaxEnt)

Semi-Supervised Learning (SSL)



Semi-Supervised Learning (SSL)



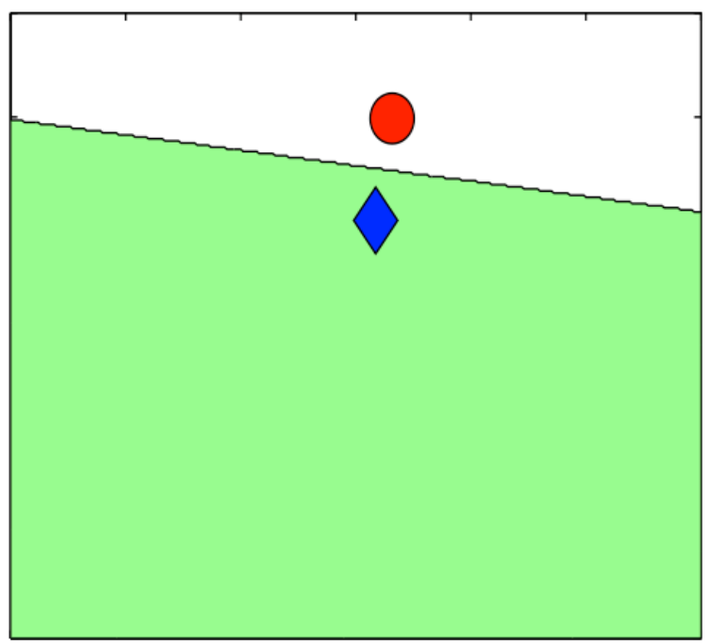
Examples:
- Self-Training
- Co-Training

Why SSL?

How can unlabeled data be helpful?

Why SSL?

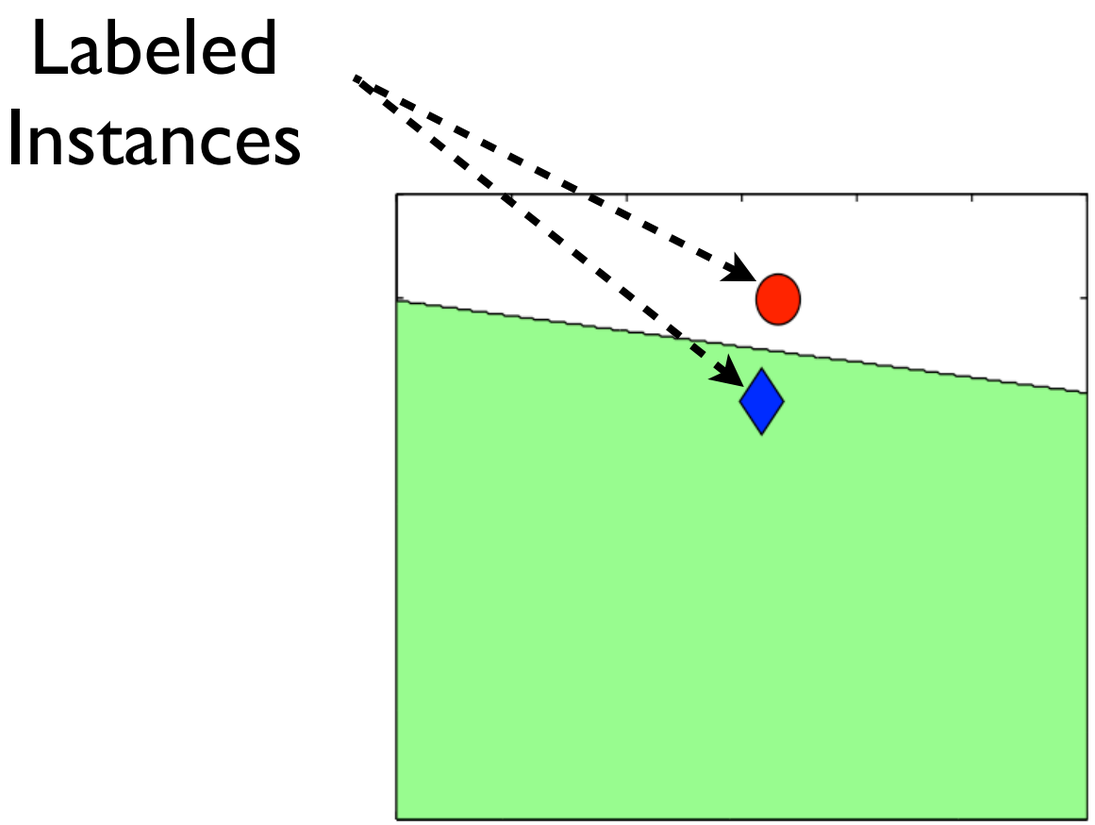
How can unlabeled data be helpful?



Without Unlabeled Data

Why SSL?

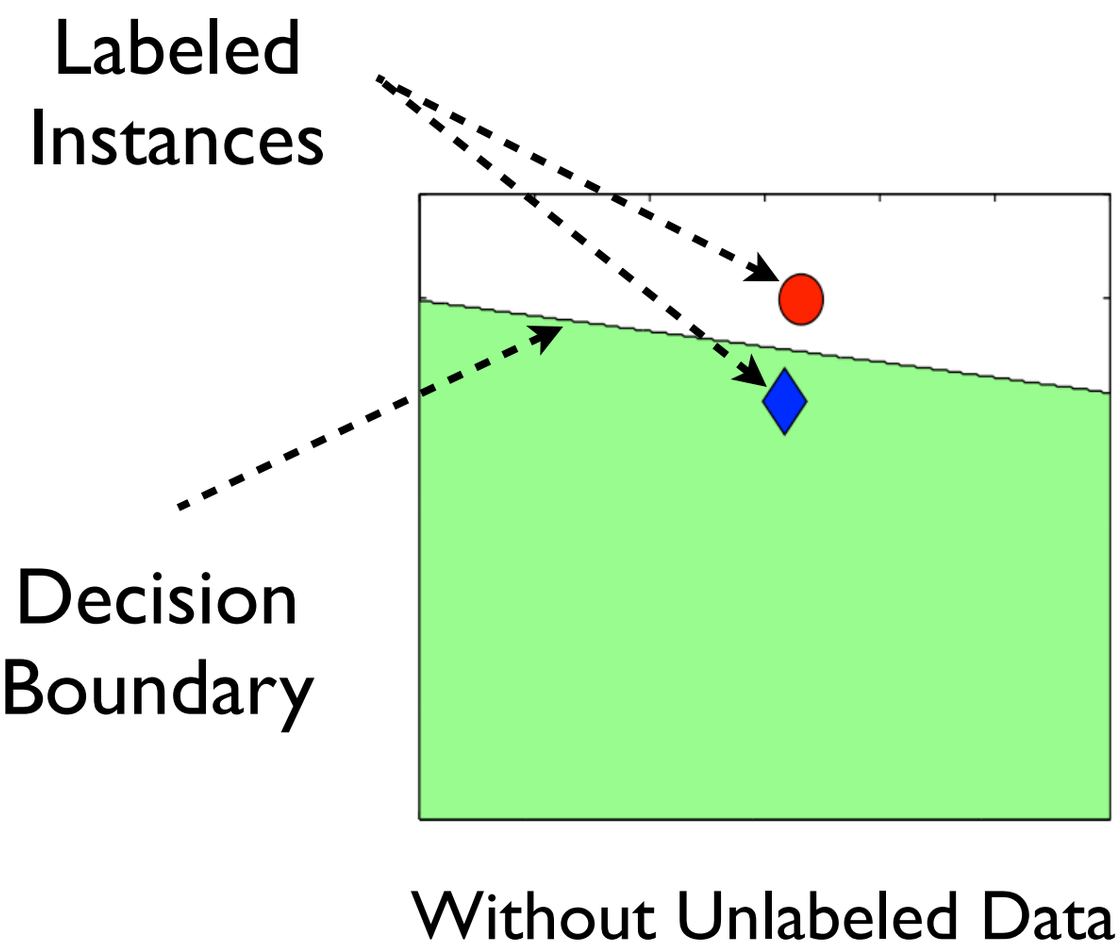
How can unlabeled data be helpful?



Without Unlabeled Data

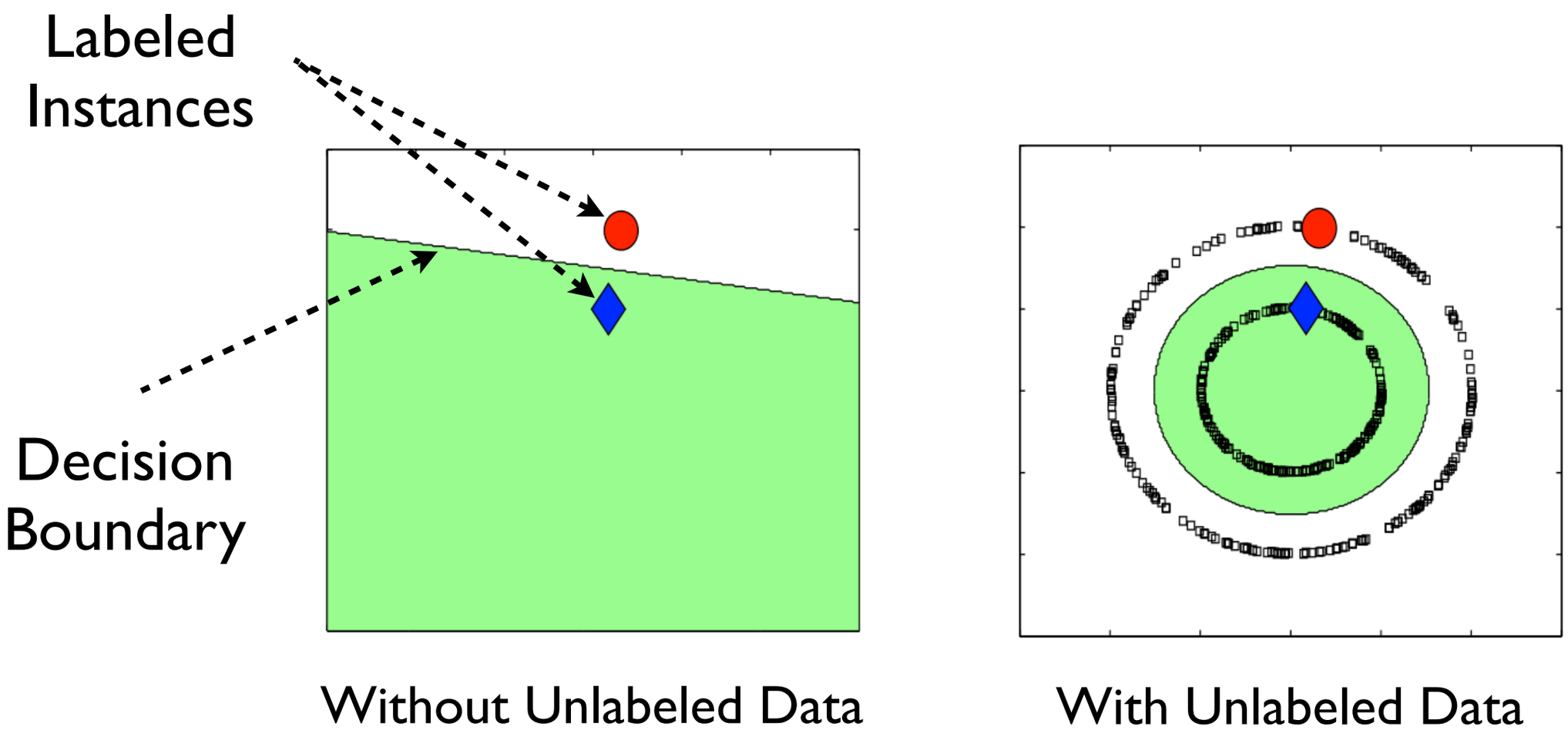
Why SSL?

How can unlabeled data be helpful?



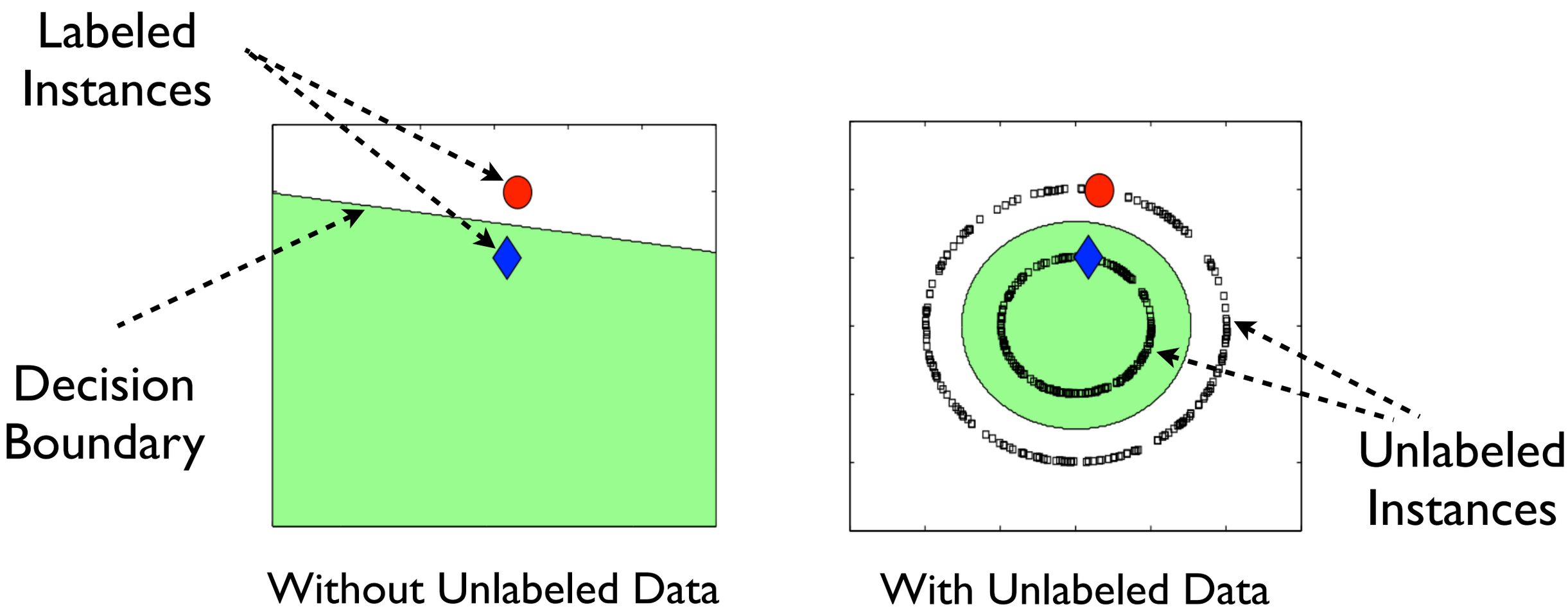
Why SSL?

How can unlabeled data be helpful?



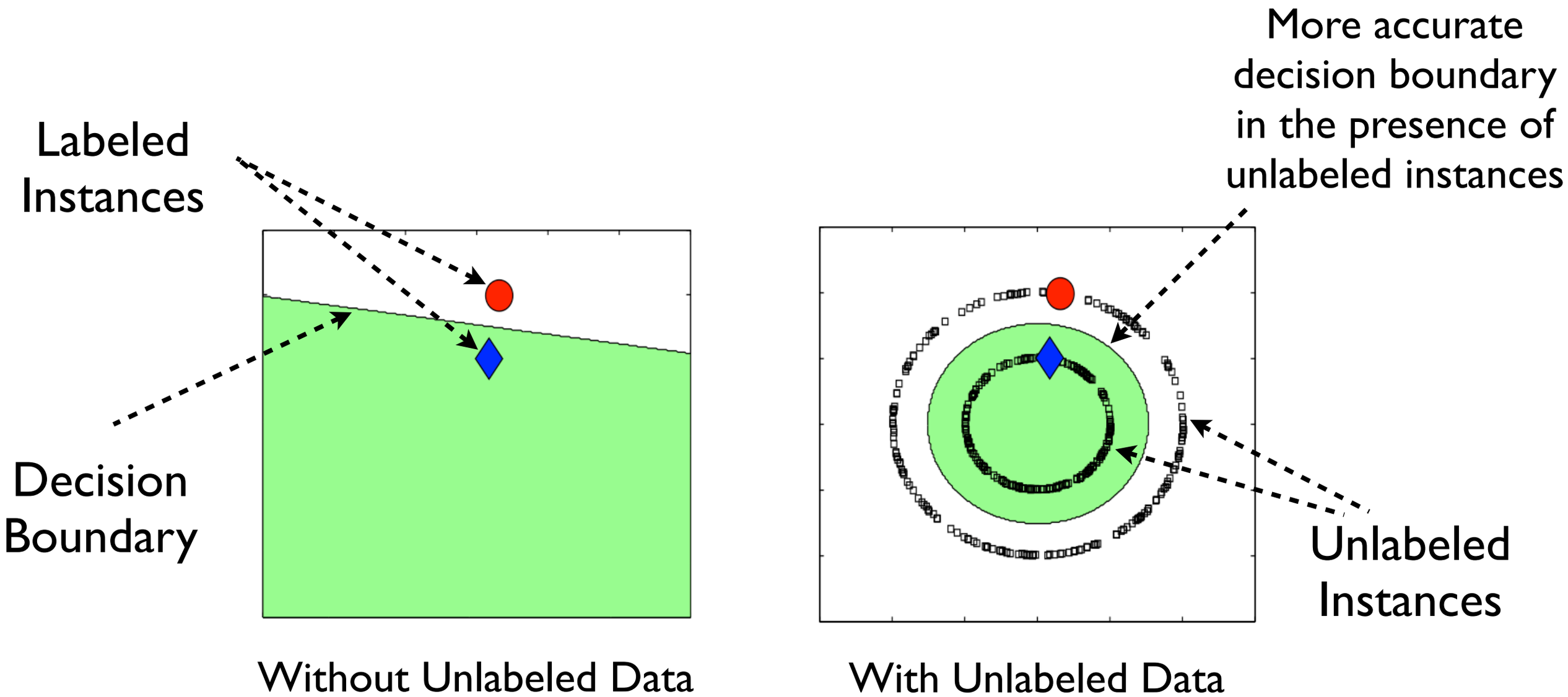
Why SSL?

How can unlabeled data be helpful?



Why SSL?

How can unlabeled data be helpful?



Example from [Belkin et al., JMLR 2006]

Inductive vs Transductive

Inductive vs Transductive

Supervised
(Labeled)

Semi-supervised
(Labeled + Unlabeled)

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)		
Semi-supervised (Labeled + Unlabeled)		

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	
Semi-supervised (Labeled + Unlabeled)		

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)		

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	Label Propagation, MAD, MP, TACO, ...

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	Label Propagation, MAD, MP, TACO, ...

Most Graph SSL algorithms are non-parametric

(i.e., # parameters grows with data size)

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	Label Propagation, MAD, MP, TACO, ...

Focus of this tutorial

Most Graph SSL algorithms are non-parametric

(i.e., # parameters grows with data size)

Inductive vs Transductive

	Inductive (Generalize to Unseen Data)	Transductive (Doesn't Generalize to Unseen Data)
Supervised (Labeled)	SVM, Maximum Entropy	X
Semi-supervised (Labeled + Unlabeled)	Manifold Regularization	Label Propagation, MAD, MP, TACO, ...

Focus of this tutorial

Most Graph SSL algorithms are non-parametric

(i.e., # parameters grows with data size)

See Chapter 25 of SSL Book: http://olivier.chapelle.cc/ssl-book/discussion.pdf

Why Graph-based SSL?

Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...

Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...
- Uniform representation for heterogeneous data

Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...
- Uniform representation for heterogeneous data
- Easily parallelizable, scalable to large data

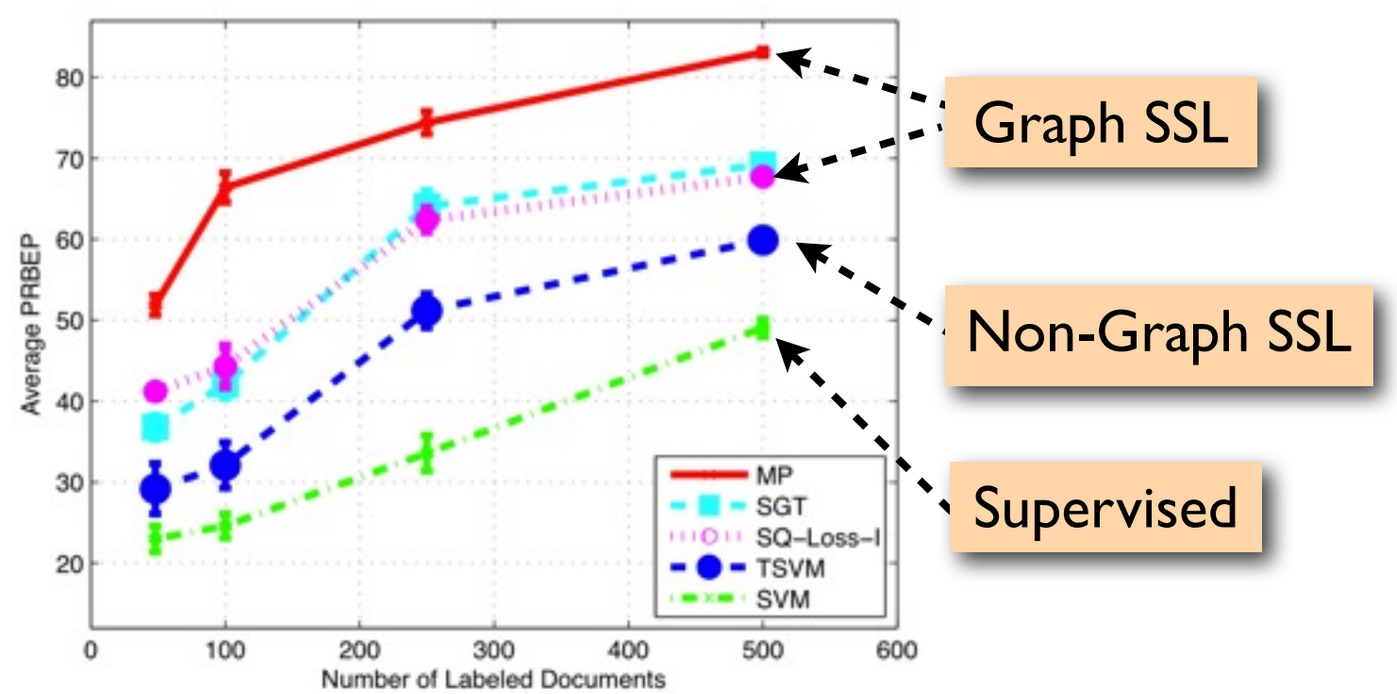
Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...
- Uniform representation for heterogeneous data
- Easily parallelizable, scalable to large data
- Effective in practice

Why Graph-based SSL?

- Some datasets are naturally represented by a graph
  - web, citation network, social network, ...
- Uniform representation for heterogeneous data
- Easily parallelizable, scalable to large data
- Effective in practice

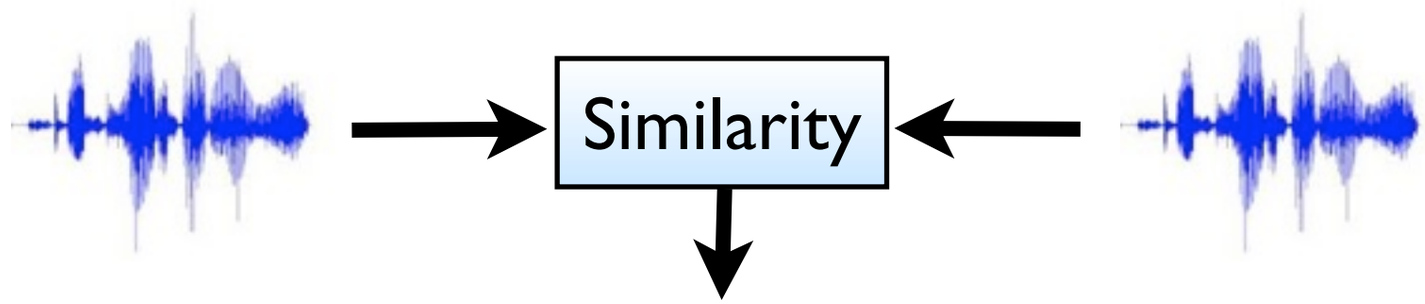
Text Classification



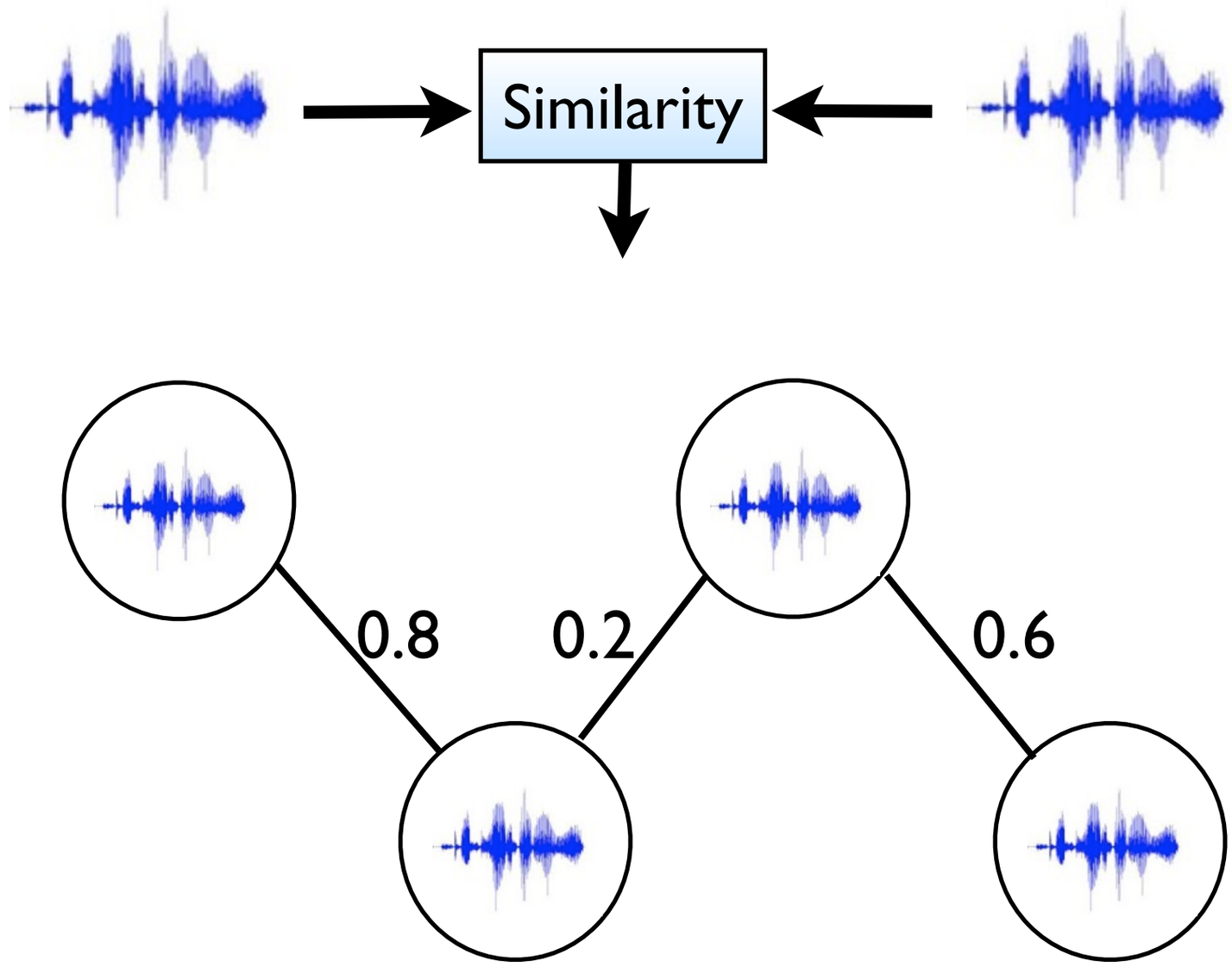
Graph-based SSL

Graph-based SSL

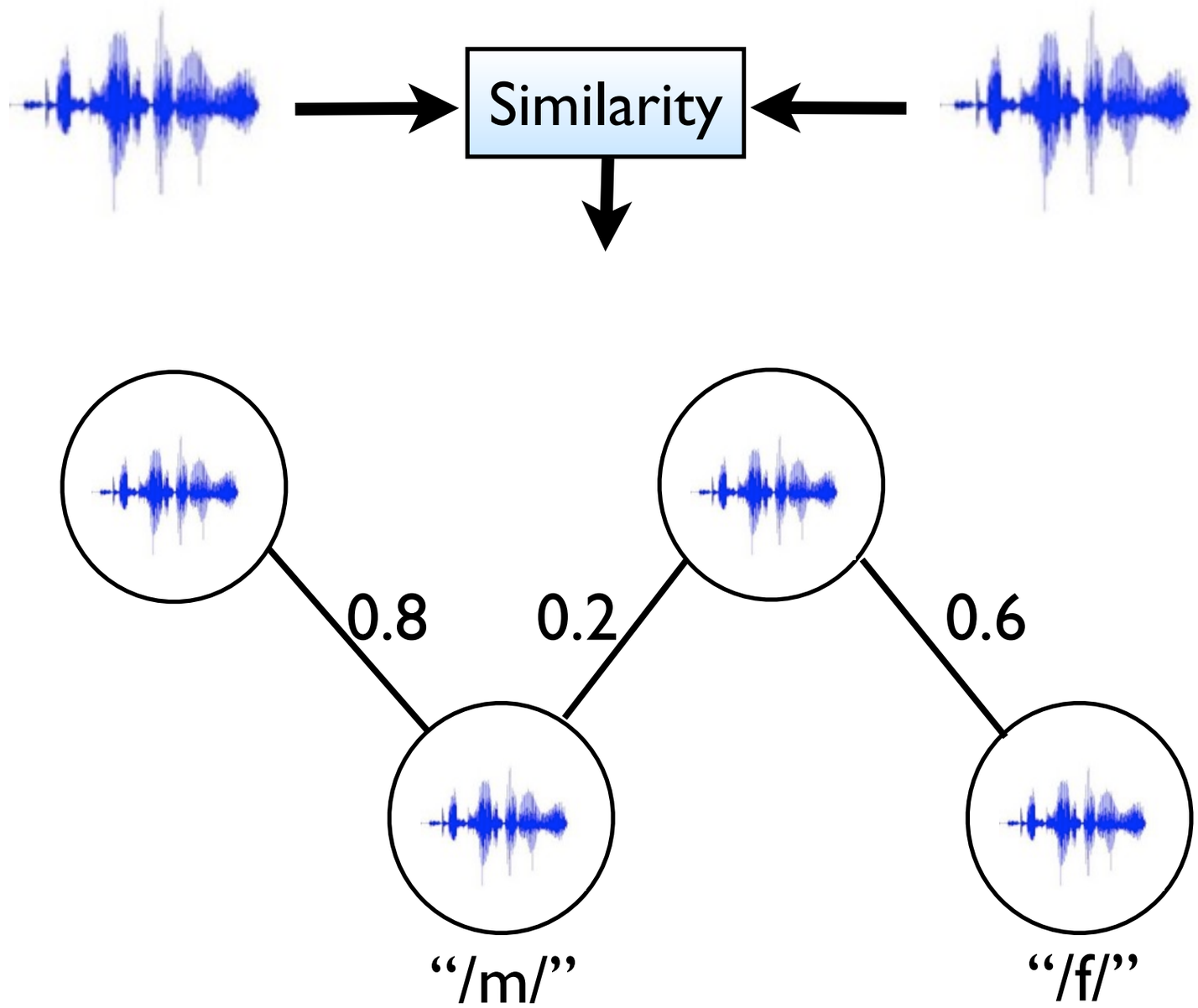
Graph-based SSL



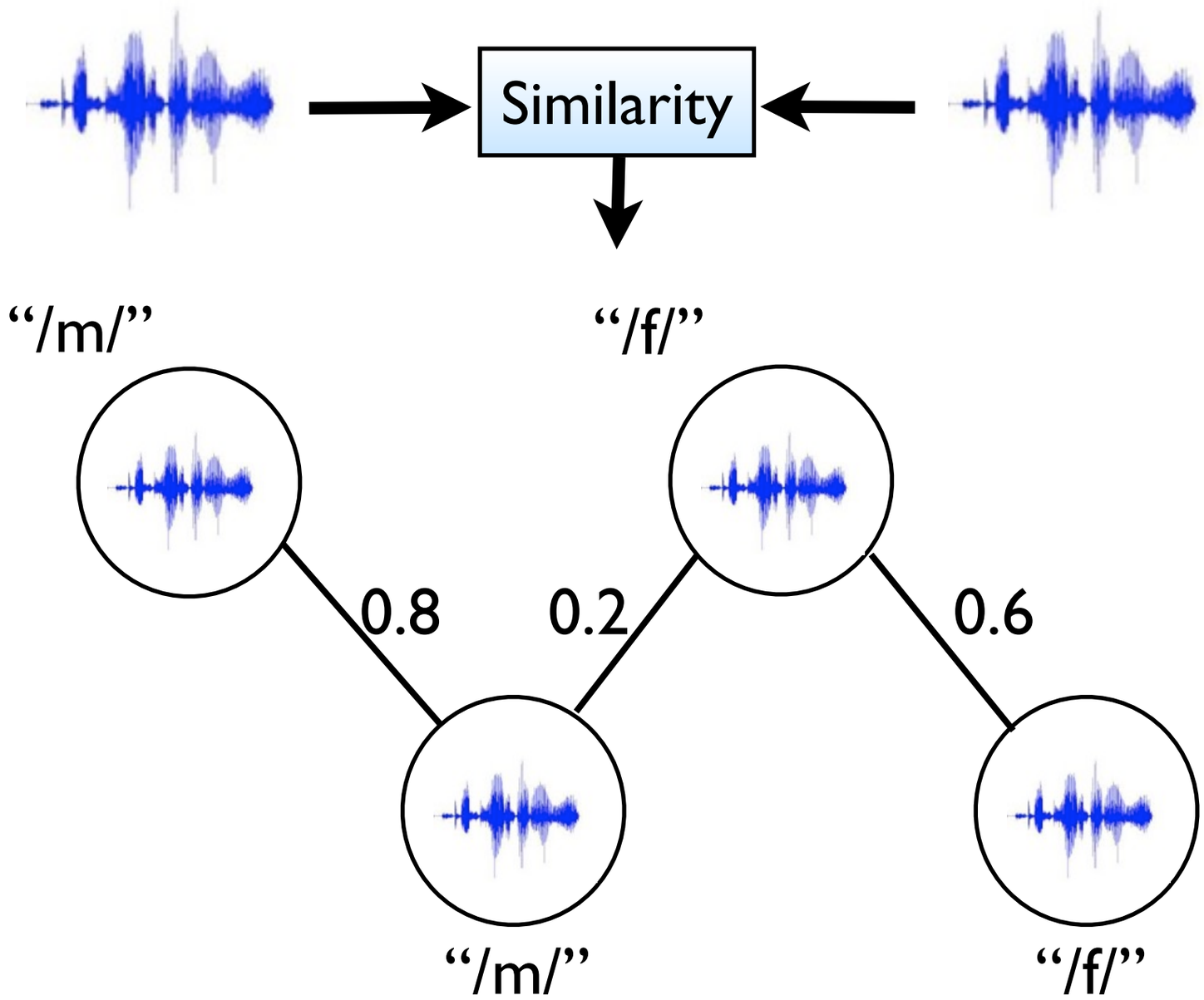
Graph-based SSL



Graph-based SSL



Graph-based SSL



Graph-based SSL

Graph-based SSL

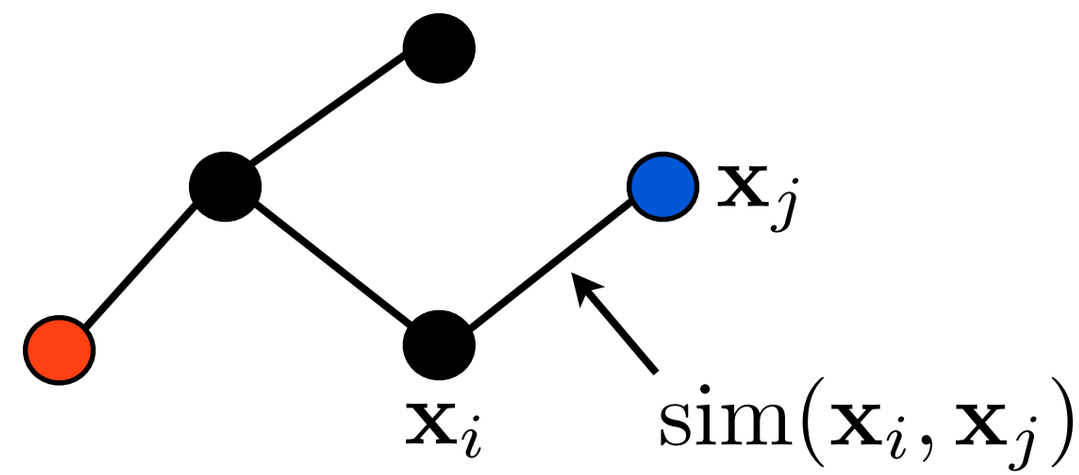
Smoothness Assumption

If two instances are similar according to the graph, then output labels should be similar

Graph-based SSL

Smoothness Assumption

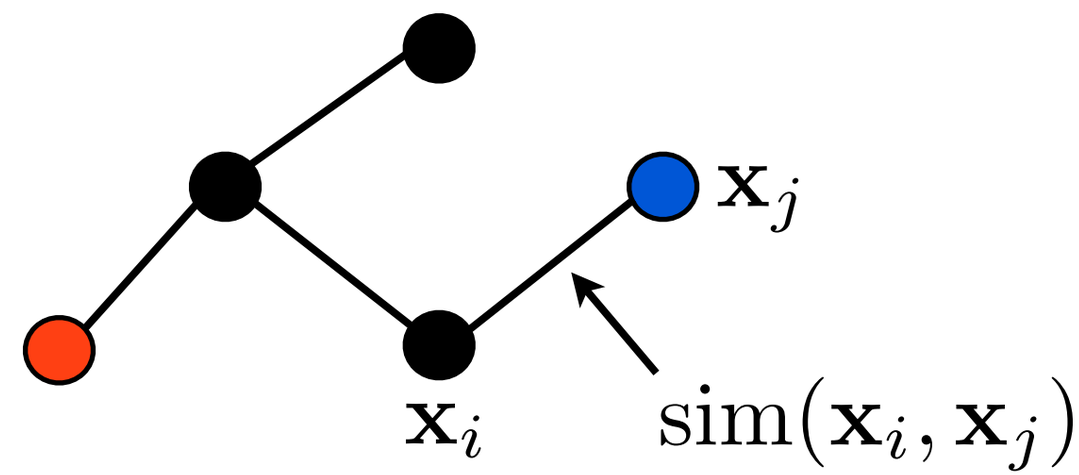
If two instances are similar according to the graph, then output labels should be similar



Graph-based SSL

Smoothness Assumption

If two instances are similar according to the graph, then output labels should be similar



- Two stages
  - Graph construction (if not already present)
  - Label Inference

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Graph Construction

- Neighborhood Methods
  - k-NN Graph Construction (k-NNG)
  - e-Neighborhood Method
- Metric Learning
- Other approaches

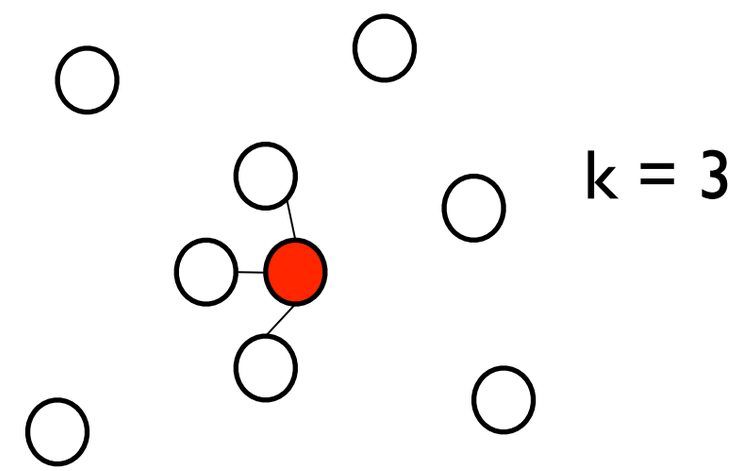
Neighborhood Methods

Neighborhood Methods

- k-Nearest Neighbor Graph (k-NNG)
  - add edges between an instance and its k-nearest neighbors

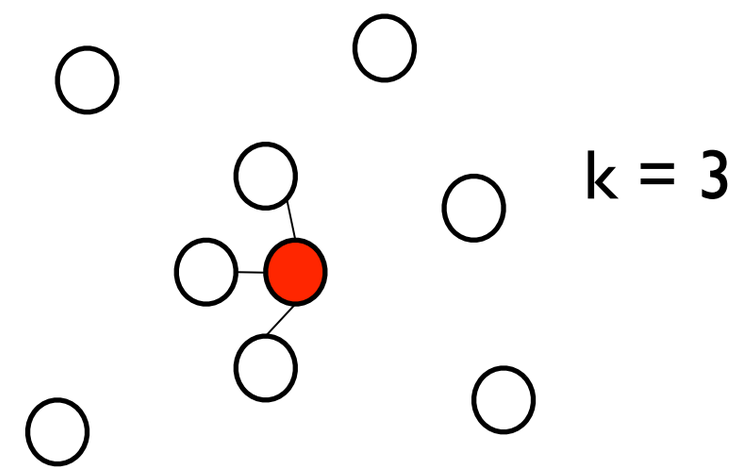
Neighborhood Methods

- k-Nearest Neighbor Graph (k-NNG)
  - add edges between an instance and its k-nearest neighbors



Neighborhood Methods

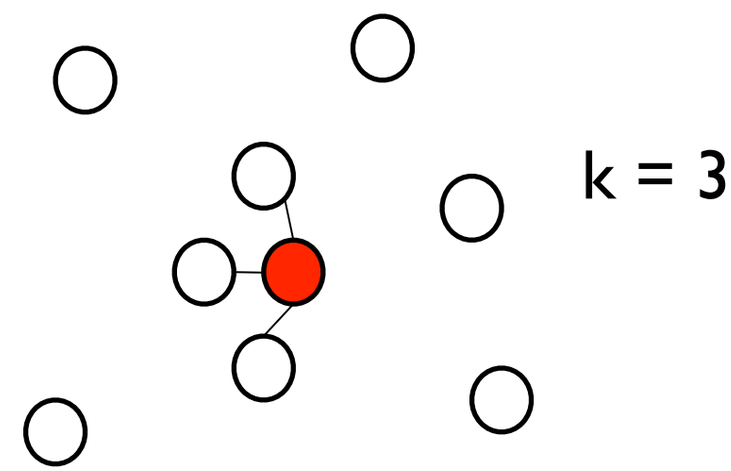
- k-Nearest Neighbor Graph (k-NNG)
  - add edges between an instance and its k-nearest neighbors



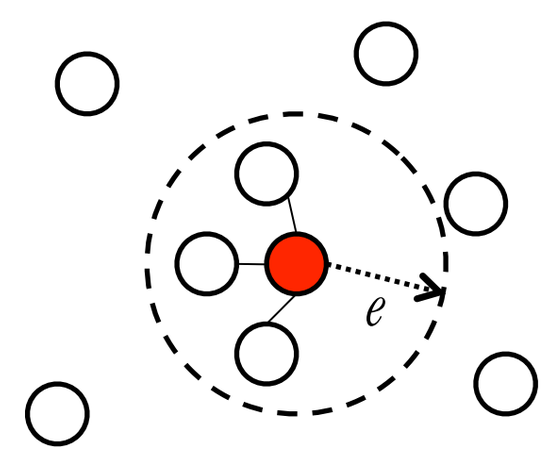
- e-Neighborhood
  - add edges to all instances inside a ball of radius e

Neighborhood Methods

- k-Nearest Neighbor Graph (k-NNG)
  - add edges between an instance and its k-nearest neighbors



- e-Neighborhood
  - add edges to all instances inside a ball of radius e



Issues with k-NNG

Issues with k-NNG

- Not scalable (quadratic)

Issues with k-NNG

- Not scalable (quadratic)
- Results in an asymmetric graph

Issues with k-NNG

- Not scalable (quadratic)
- Results in an asymmetric graph
  - b is the closest neighbor of a, but not the other way

Issues with k-NNG

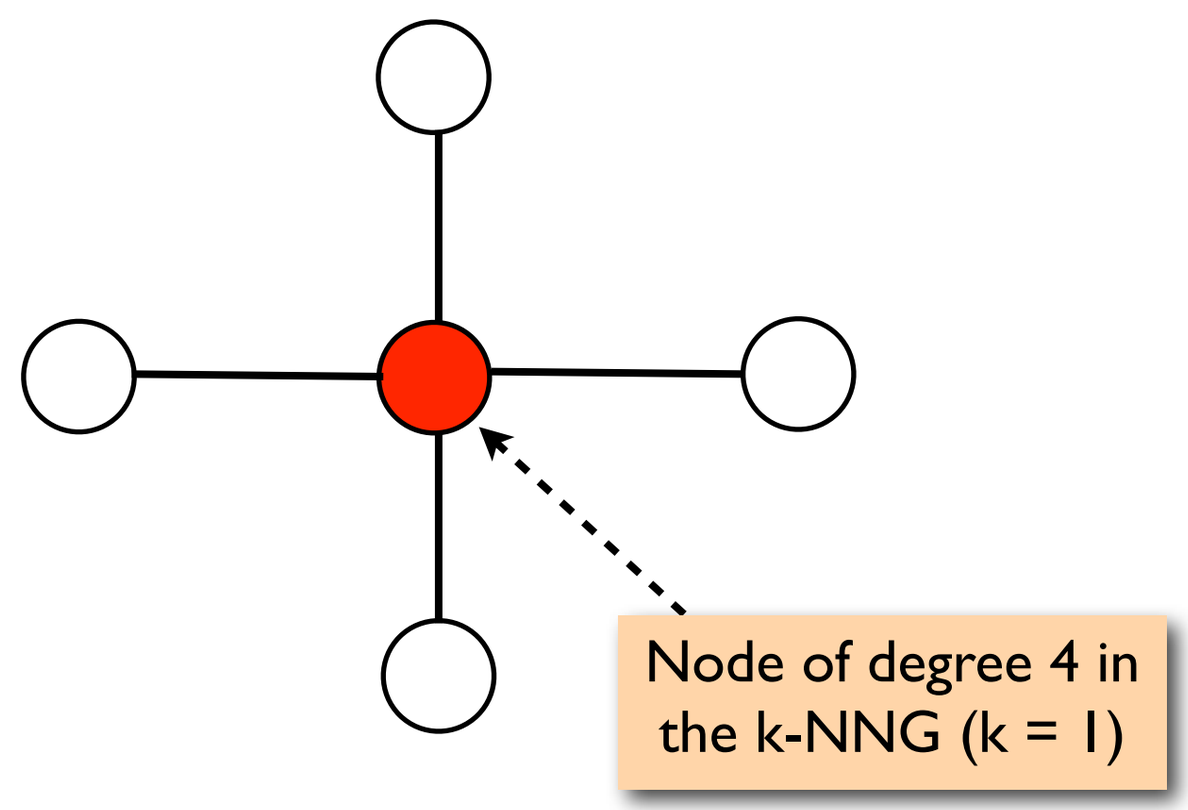
- Not scalable (quadratic)
- Results in an asymmetric graph
  - b is the closest neighbor of a, but not the other way

- Results in irregular graphs
  - some nodes may end up with higher degree than other nodes

Issues with k-NNG

- Not scalable (quadratic)
- Results in an asymmetric graph
  - b is the closest neighbor of a, but not the other way

- Results in irregular graphs
  - some nodes may end up with higher degree than other nodes



Issues with ϵ-Neighborhood

Issues with e-Neighborhood

- Not scalable

Issues with e-Neighborhood

- Not scalable
- Sensitive to value of e : not invariant to scaling

Issues with e-Neighborhood

- Not scalable
- Sensitive to value of e : not invariant to scaling
- Fragmented Graph: disconnected components

Issues with e-Neighborhood

- Not scalable
- Sensitive to value of e : not invariant to scaling
- Fragmented Graph: disconnected components

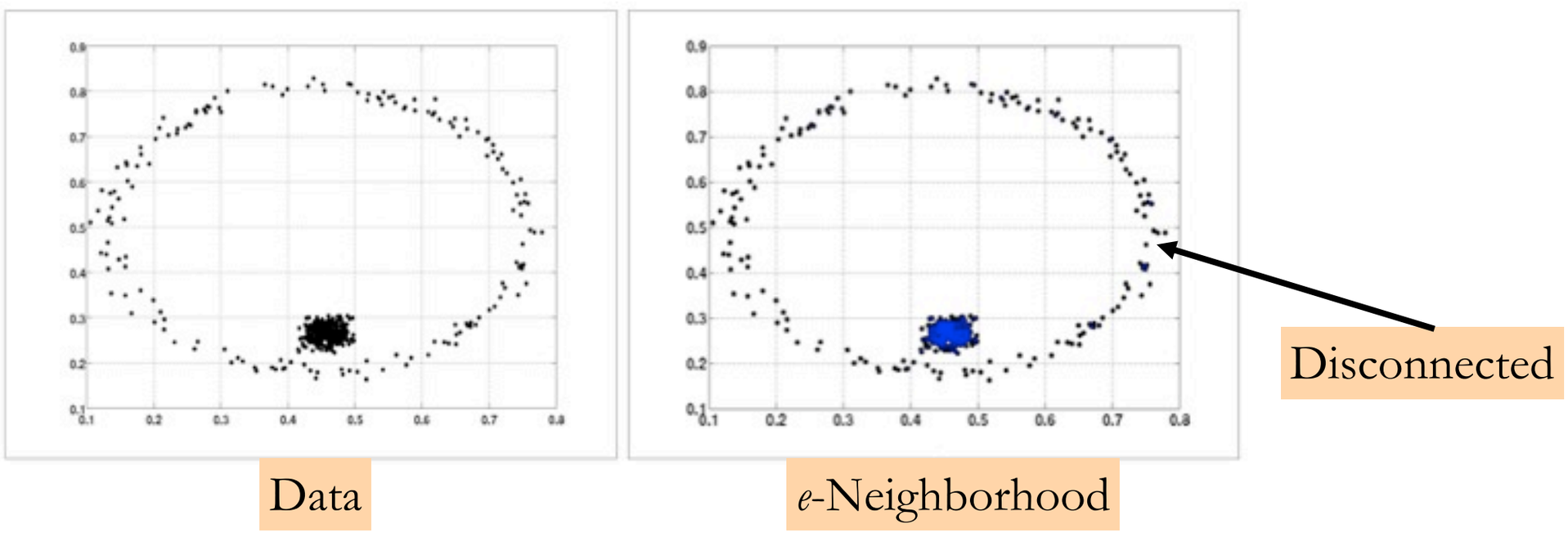
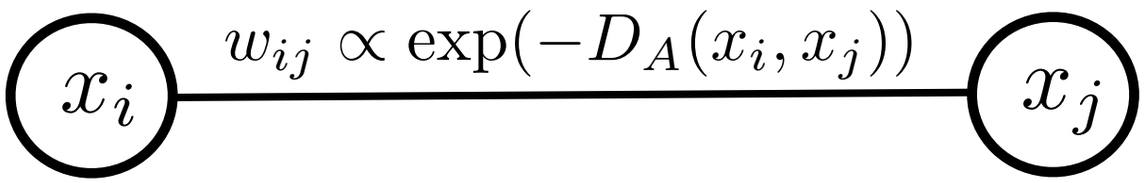


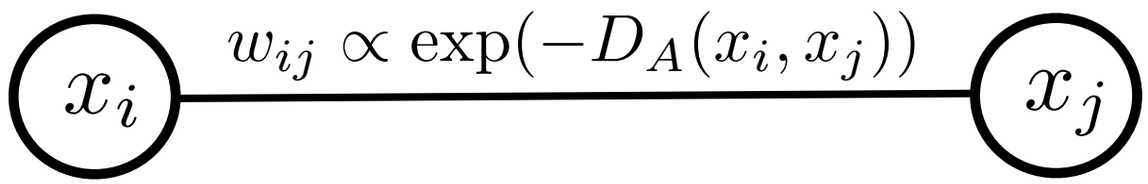
Figure from [Jebara et al., ICML 2009]

Graph Construction using Metric Learning

Graph Construction using Metric Learning



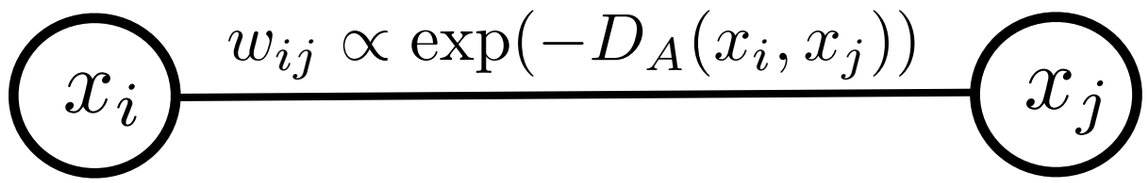
Graph Construction using Metric Learning



$$D_A(x_i, x_j) = (x_i - x_j)^T A (x_i - x_j)$$

Estimated using Mahalanobis metric learning algorithms

Graph Construction using Metric Learning



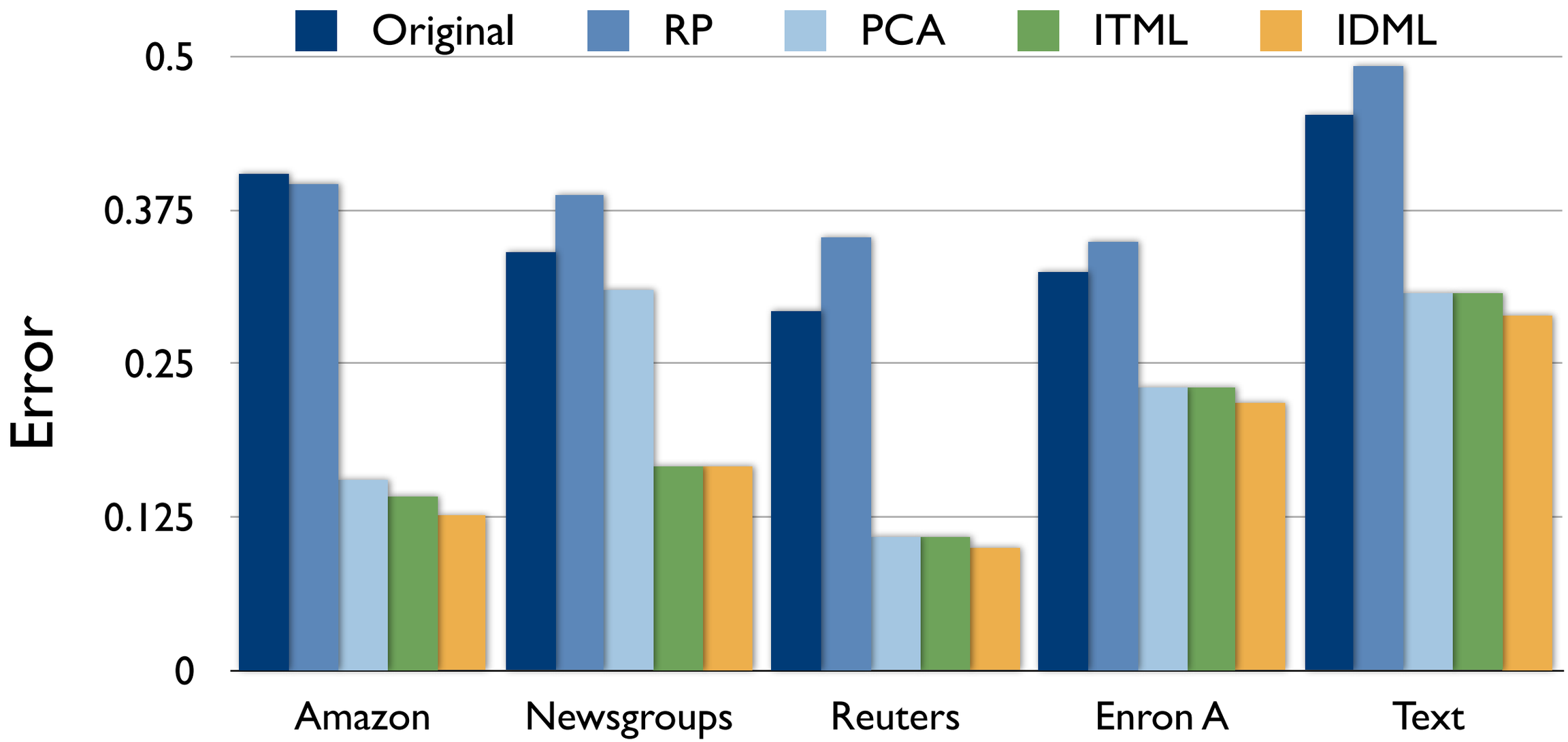
$$D_A(x_i, x_j) = (x_i - x_j)^T A (x_i - x_j)$$

Estimated using Mahalanobis metric learning algorithms

- Supervised Metric Learning
  - ITML [Kulis et al., ICML 2007]
  - LMNN [Weinberger and Saul, JMLR 2009]
- Semi-supervised Metric Learning
  - IDML [Dhillon et al., UPenn TR 2010]

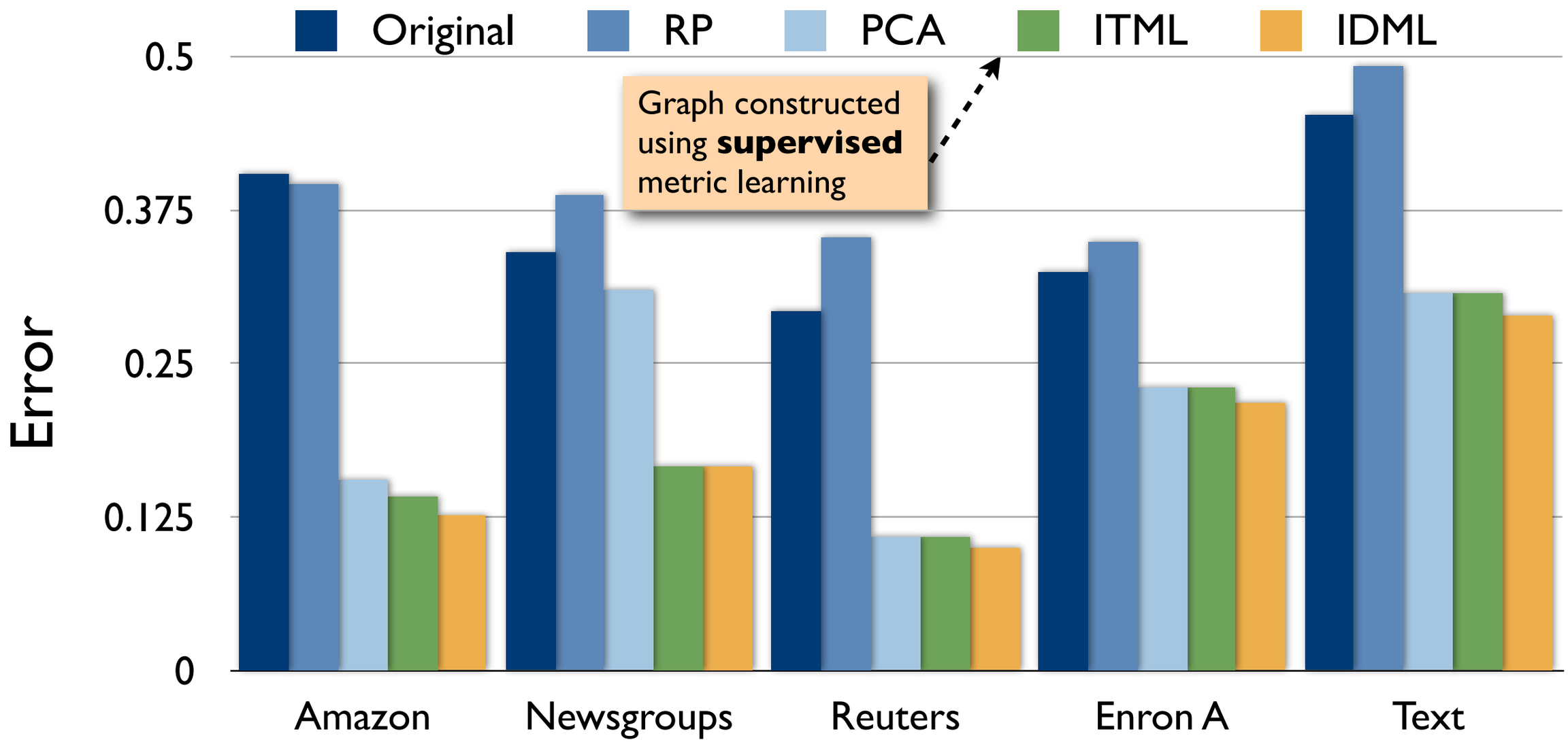
Benefits of Metric Learning for Graph Construction

Benefits of Metric Learning for Graph Construction



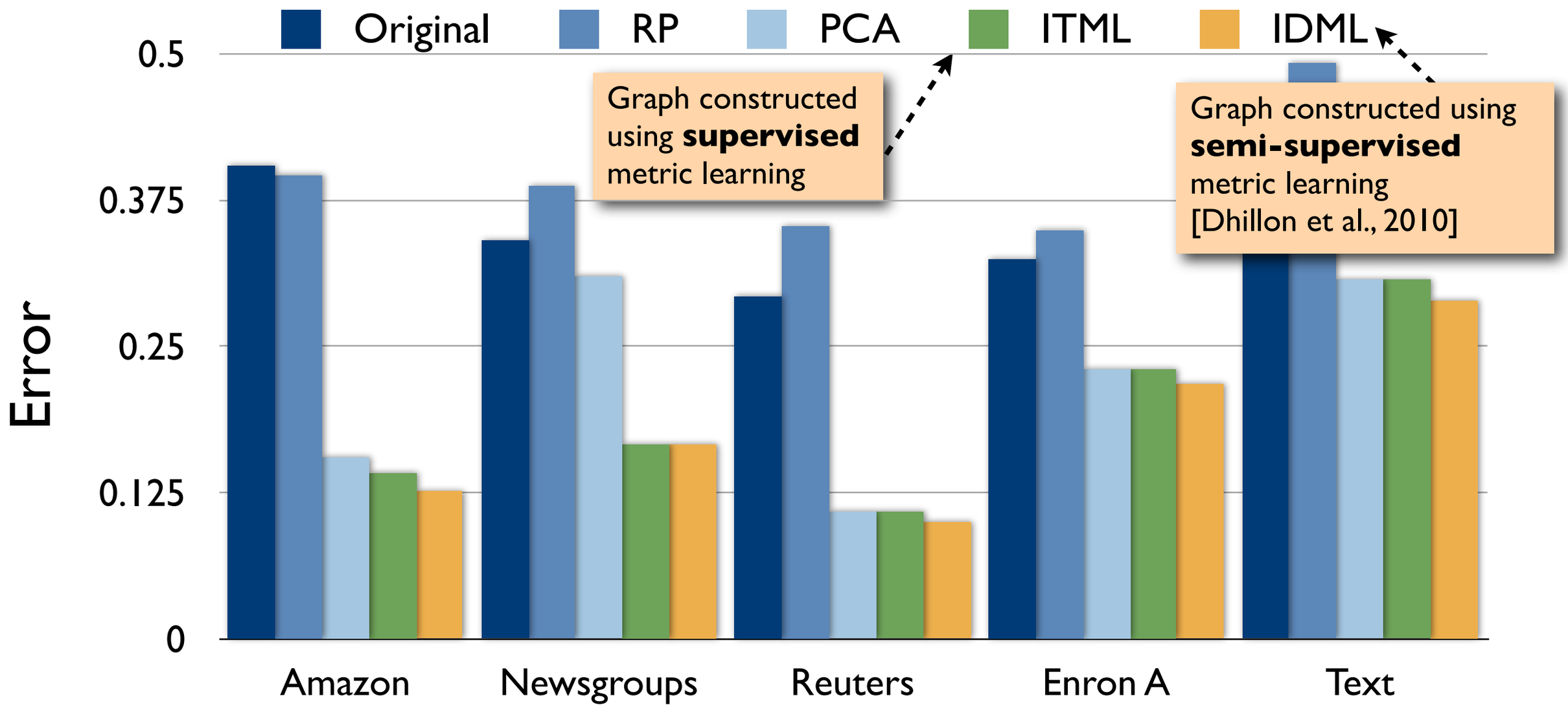
100 seed and1400 test instances, all inferences using LP

Benefits of Metric Learning for Graph Construction



100 seed and1400 test instances, all inferences using LP

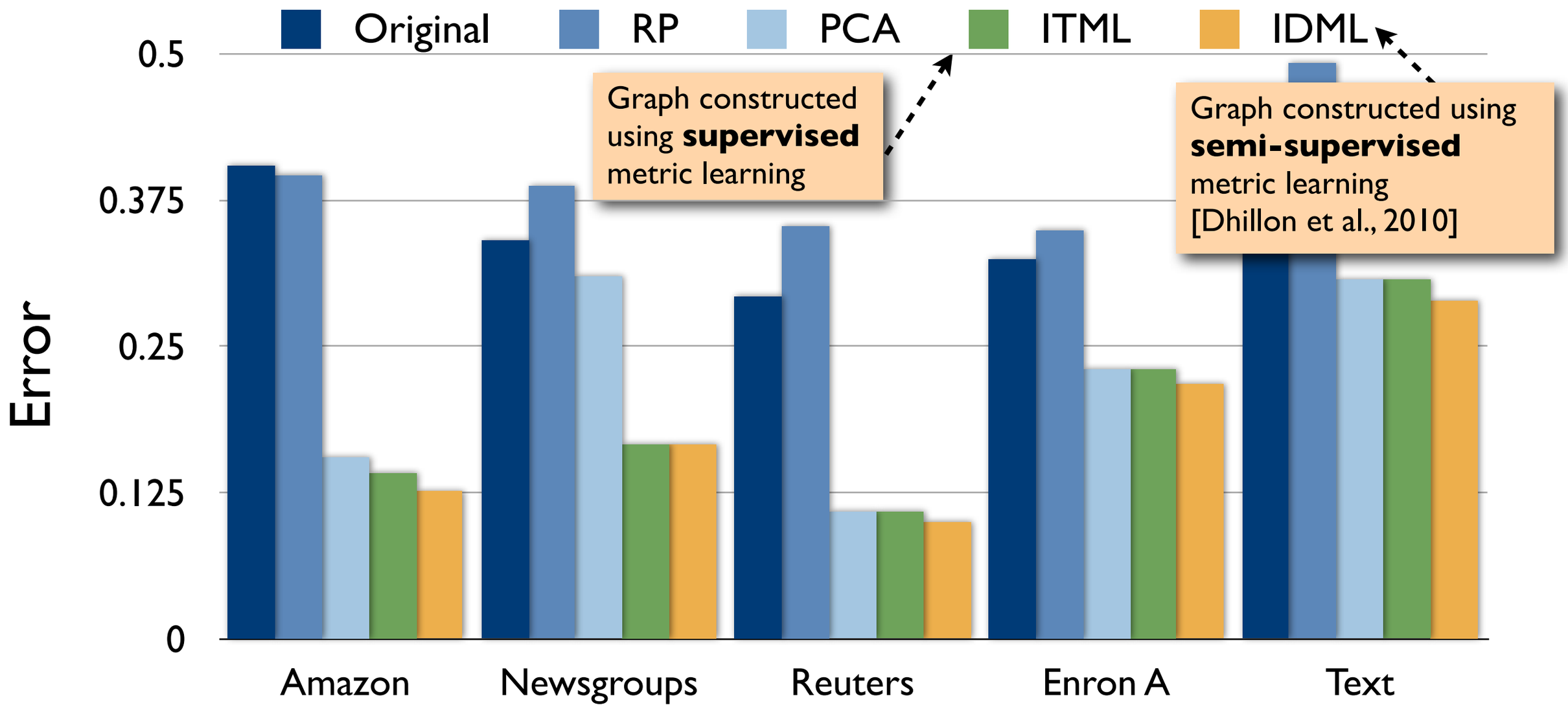
Benefits of Metric Learning for Graph Construction



100 seed and1400 test instances, all inferences using LP

[Dhillon et al., UPenn TR 2010]

Benefits of Metric Learning for Graph Construction



100 seed and1400 test instances, all inferences using LP

Careful graph construction is critical!

[Dhillon et al., UPenn TR 2010]

Other Graph Construction Approaches

- Local Reconstruction
  - Linear Neighborhood [Wang and Zhang, ICML 2005]
  - Regular Graph: b-matching [Jebara et al., ICML 2008]
  - Fitting Graph to Vector Data [Daitch et al., ICML 2009]
- Graph Kernels
  - [Zhu et al., NIPS 2005]

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Graph Laplacian

Graph Laplacian

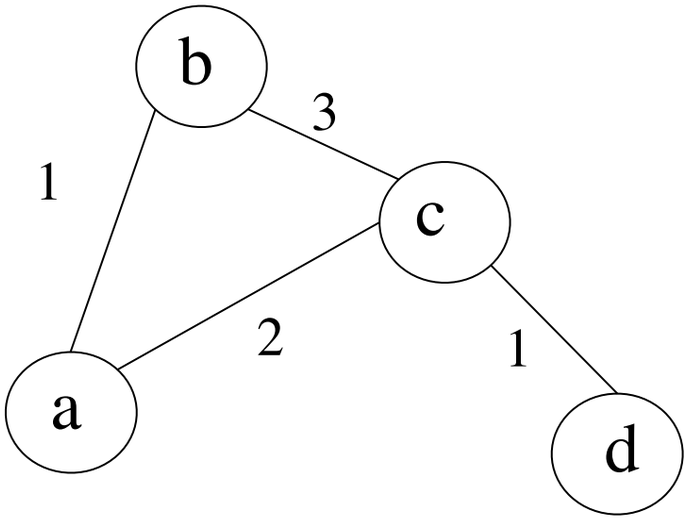
- Laplacian (un-normalized) of a graph:

$$L = D - W, \text{where } D_{ii} = \sum_j W_{ij}, \; D_{ij(\neq i)} = 0$$

Graph Laplacian

- Laplacian (un-normalized) of a graph:

$$L = D - W, \text{where } D_{ii} = \sum_j W_{ij}, \; D_{ij(\neq i)} = 0$$

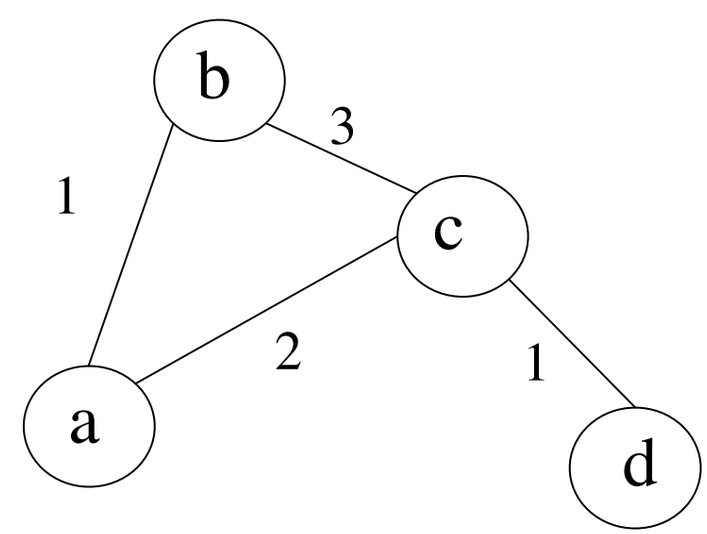


Graph Laplacian

- Laplacian (un-normalized) of a graph:

$$L = D - W, \text{where } D_{ii} = \sum_j W_{ij}, \; D_{ij(\neq i)} = 0$$

$$\begin{array}{c c} & \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{array}{c} a \\ b \\ c \\ d \end{array} & \begin{pmatrix} 3 & -1 & -2 & 0 \\ -1 & 4 & -3 & 0 \\ -2 & -3 & 6 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \end{array}$$



Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

$$f^T L f = \sum_{i,j} W_{ij} (f_i - f_j)^2$$

Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Measure of Non-Smoothness

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$

Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Vector of scores for single label on nodes

Measure of Non-Smoothness

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$

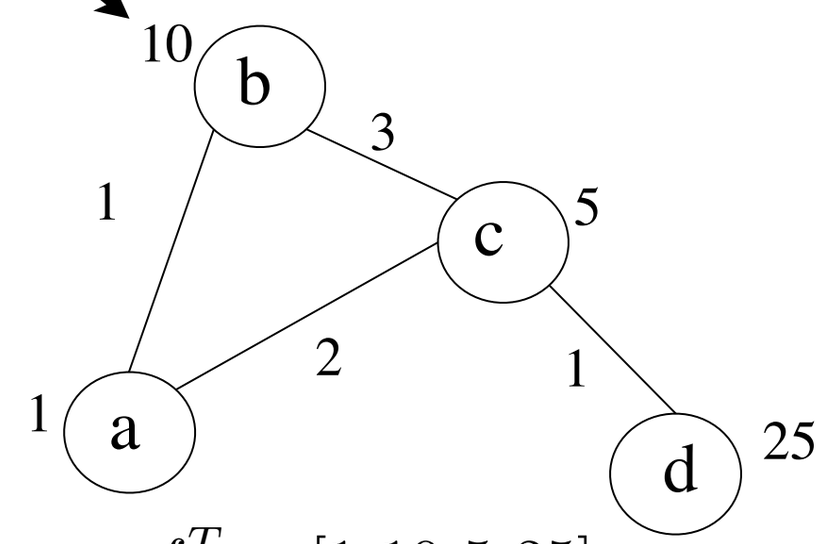
Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Vector of scores for single label on nodes

Measure of Non-Smoothness

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$



$f^T = [1\ 10\ 5\ 25]$

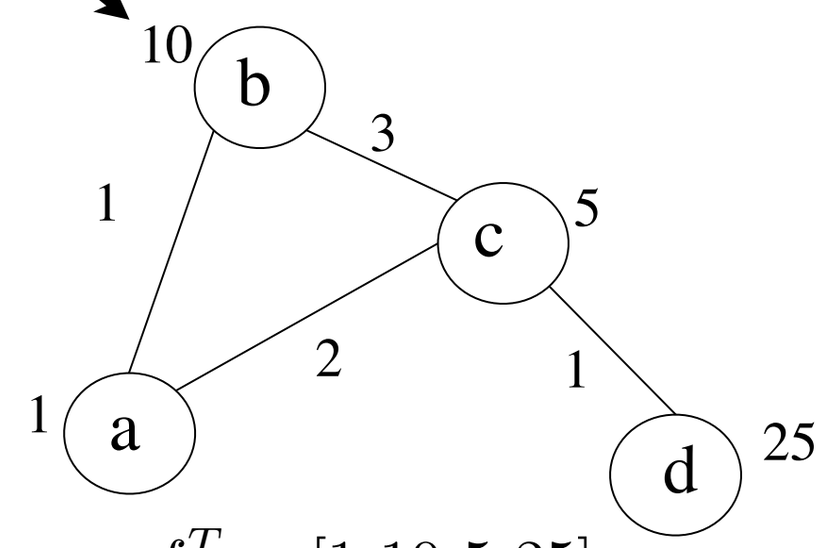
Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Measure of Non-Smoothness

Vector of scores for single label on nodes

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$



$f^T = [1\ 10\ 5\ 25]$

$f^T L f = 588$

Not Smooth

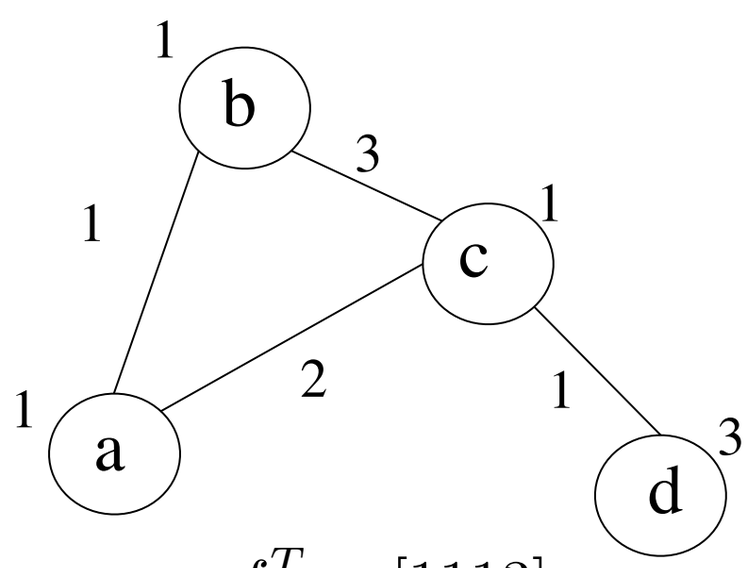
Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

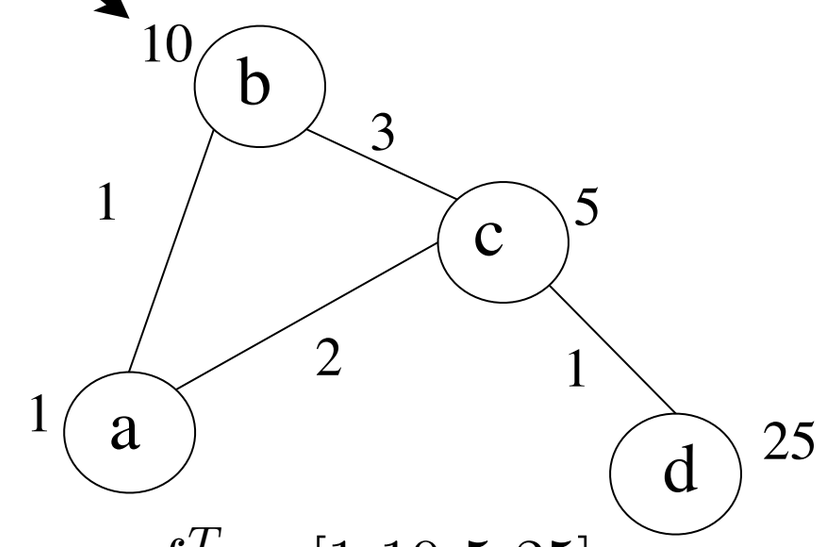
Vector of scores for single label on nodes

Measure of Non-Smoothness

$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$



$f^T = [1113]$
$f^T L f = 4$



$f^T = [1\ 10\ 5\ 25]$
$f^T L f = 588$

Not Smooth

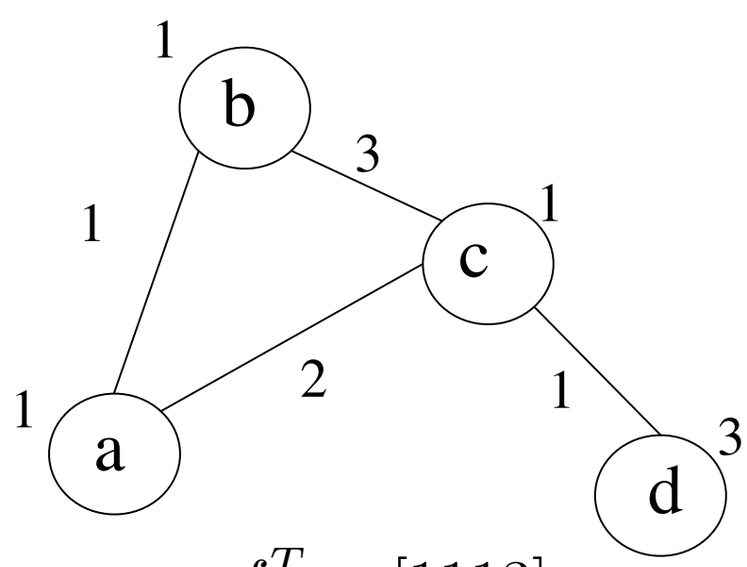
Graph Laplacian (contd.)

- L is positive semi-definite (assuming non-negative weights)
- Smoothness of prediction f over the graph in terms of the Laplacian:

Vector of scores for single label on nodes

Measure of Non-Smoothness

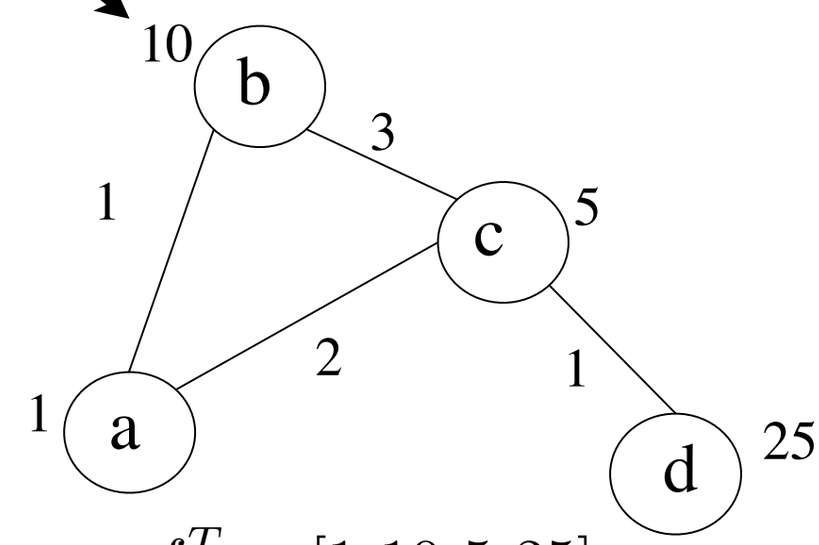
$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$



$f^T = [1113]$

$f^T L f = 4$

Smooth



$f^T = [1\ 10\ 5\ 25]$

$f^T L f = 588$

Not Smooth

Relationship between Eigenvalues of the Laplacian and Smoothness

$$Lg = \lambda g$$

$$g^T L g = \lambda g^T g$$

$$g^T L g = \lambda$$

Relationship between Eigenvalues of the Laplacian and Smoothness

Eigenvector of L

Eigenvalue of L

$$Lg = \lambda g$$

$$g^T L g = \lambda g^T g$$

$$g^T L g = \lambda$$

Relationship between Eigenvalues of the Laplacian and Smoothness

Eigenvector of L

Eigenvalue of L

$$Lg = \lambda g$$

$$g^T L g = \lambda \boxed{g^T g}$$

= 1, as eigenvectors are are orthonormal

$$g^T L g = \lambda$$

Relationship between Eigenvalues of the Laplacian and Smoothness

Eigenvector of L

Eigenvalue of L

$$Lg = \lambda g$$

$$g^T L g = \lambda \boxed{g^T g}$$

= 1, as eigenvectors are are orthonormal

$$g^T L g = \lambda$$

Measure of Non-Smoothness (previous slide)

Relationship between Eigenvalues of the Laplacian and Smoothness

Eigenvector of L

Eigenvalue of L

$$Lg = \lambda g$$

$$g^T L g = \lambda \boxed{g^T g}$$

= 1, as eigenvectors are are orthonormal

$$g^T L g = \lambda$$

Measure of Non-Smoothness (previous slide)

If an eigenvector is used to classify nodes, then the corresponding eigenvalue gives the measure of non-smoothness

Spectrum of the Graph Laplacian

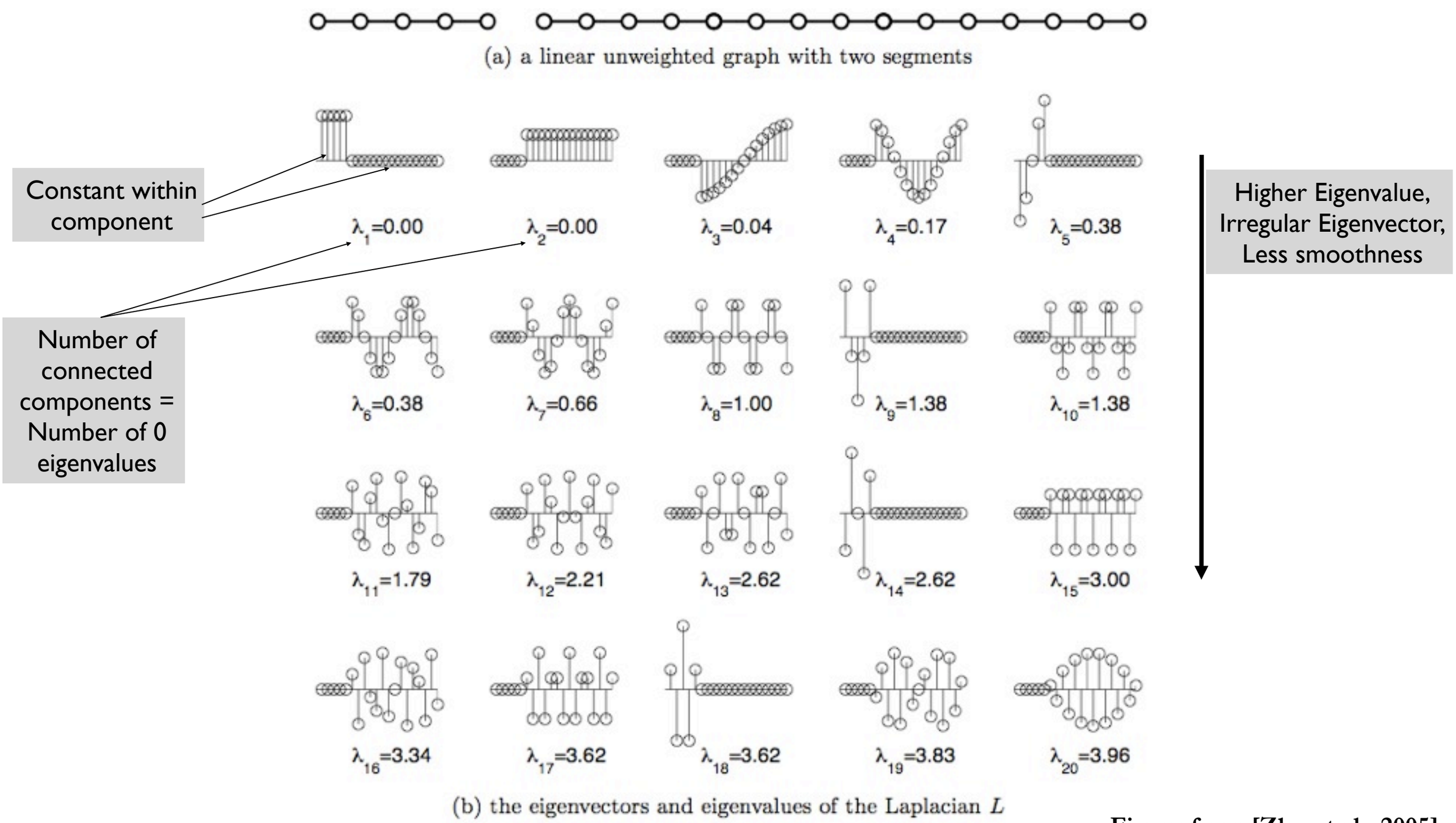
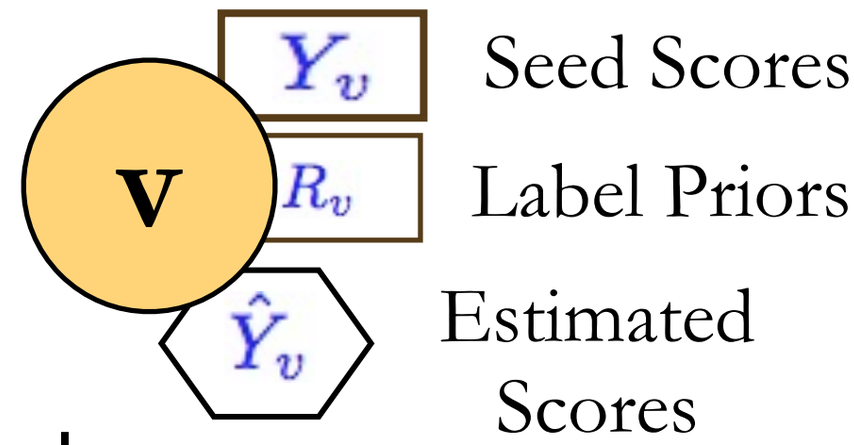


Figure from [Zhu et al., 2005]

Notations

$\hat{Y}_{v,l}$: score of estimated label l on node v



$Y_{v,l}$: score of seed label l on node v

$R_{v,l}$: regularization target for label l on node v

S : seed node indicator (diagonal matrix)

W_{uv} : weight of edge (u, v) in the graph

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Notations

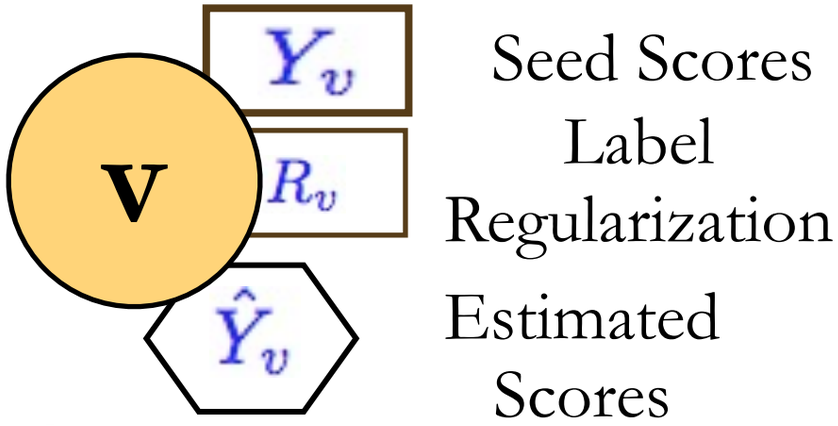
$\hat{Y}_{v,l}$: score of estimated label l on node v

$Y_{v,l}$: score of seed label l on node v

$R_{v,l}$: regularization target for label l on node v

S : seed node indicator (diagonal matrix)

W_{uv} : weight of edge (u, v) in the graph



LP-ZGL [Zhu et al., ICML 2003]

$$\arg\min_{\hat{Y}} \sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

Graph Laplacian

such that $Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1$

LP-ZGL [Zhu et al., ICML 2003]

Smooth

$$\arg\min_{\hat{Y}} \boxed{\sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2} = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

Graph Laplacian

such that $Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1$

LP-ZGL [Zhu et al., ICML 2003]

Smooth

$$\arg\min_{\hat{Y}} \boxed{\sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2} = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

Graph Laplacian

such that $\boxed{Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1}$

Match Seeds (hard)

LP-ZGL [Zhu et al., ICML 2003]

Smooth

$$\arg\min_{\hat{Y}} \boxed{\sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2} = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

Graph Laplacian

such that $\boxed{Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1}$

Match Seeds (hard)

- **Smoothness**
  - two nodes connected by an edge with high weight should be assigned similar labels

LP-ZGL [Zhu et al., ICML 2003]

Smooth

$$\arg\min_{\hat{Y}} \boxed{\sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2} = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l$$

Graph Laplacian

such that $\boxed{Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1}$

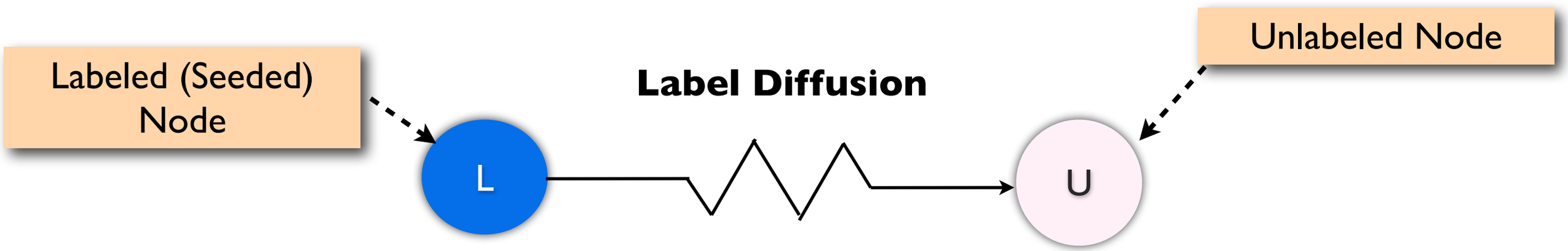
Match Seeds (hard)

- **Smoothness**
  - two nodes connected by an edge with high weight should be assigned similar labels
- Solution satisfies harmonic property

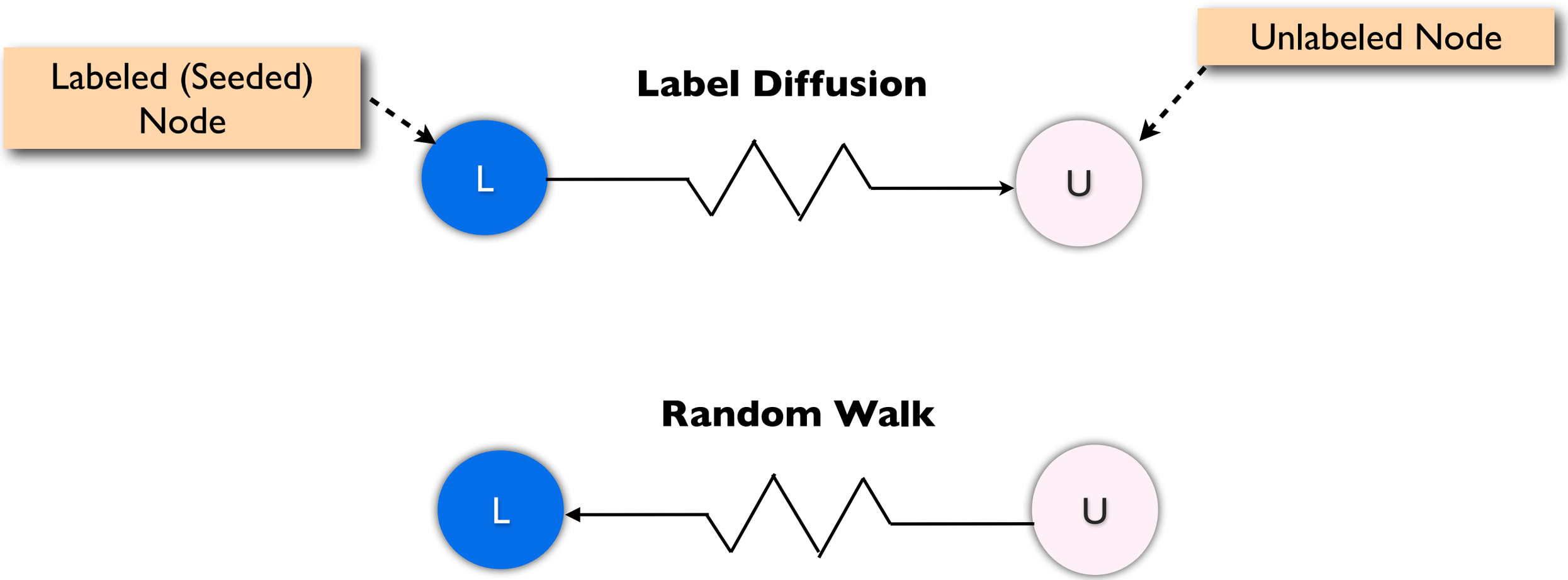
Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Two Related Views

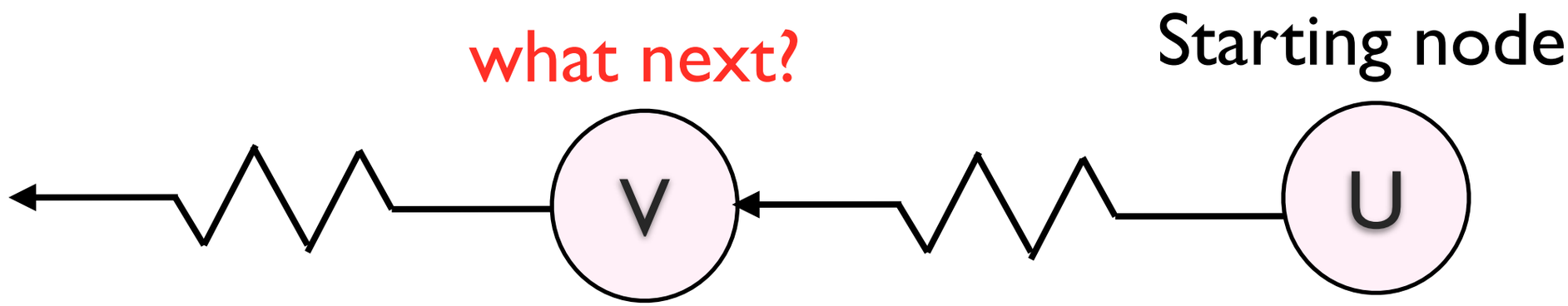


Two Related Views

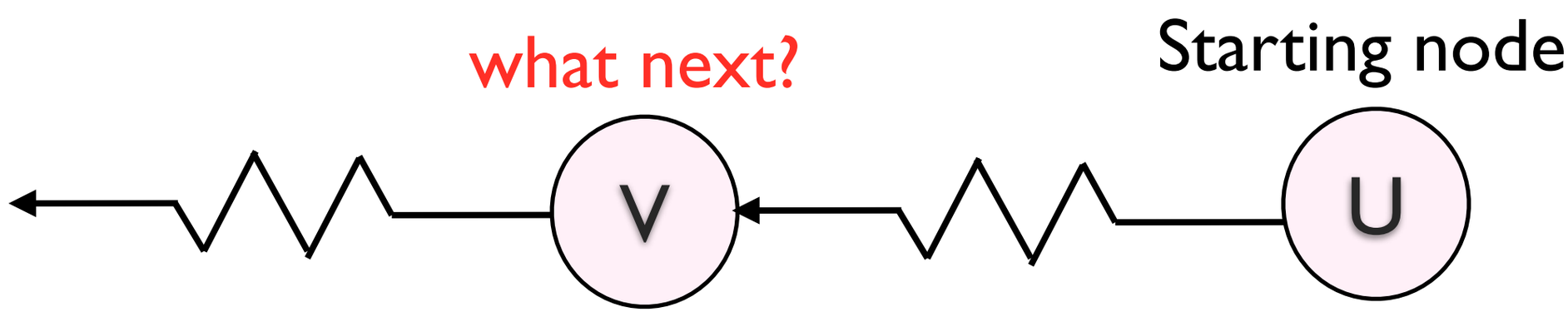


Random Walk View

Random Walk View



Random Walk View



- Continue walk with probability $\mathbf{p_v^{cont}}$
- Assign V's seed label to U with probability $\mathbf{p_v^{inj}}$
- Abandon random walk with probability $\mathbf{p_v^{abnd}}$
  - assign U a dummy label

Discounting Nodes

Discounting Nodes

- Certain nodes can be unreliable (e.g., high degree nodes)
  - do not allow propagation/walk through them

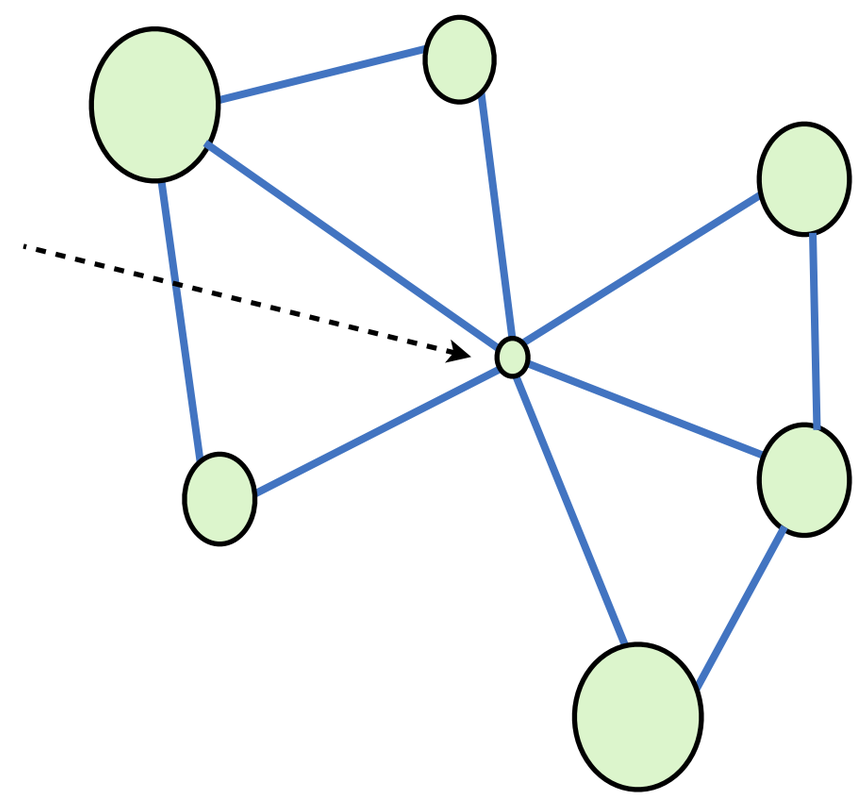
Discounting Nodes

- Certain nodes can be unreliable (e.g., high degree nodes)
  - do not allow propagation/walk through them
- Solution: increase abandon probability on such nodes:

Discounting Nodes

- Certain nodes can be unreliable (e.g., high degree nodes)
  - do not allow propagation/walk through them
- Solution: increase abandon probability on such nodes:

$$\mathbf{p_v^{abnd}} \propto \mathrm{degree(v)}$$



Redefining Matrices

$$W'_{uv} = p_u^{cont} \times W_{uv}$$

New Edge Weight

$$S_{uu} = \sqrt{p_u^{inj}}$$

$$R_{u\top} = p_u^{abnd}$$, and 0 for non-dummy labels

Dummy Label

Modified Adsorption (MAD)

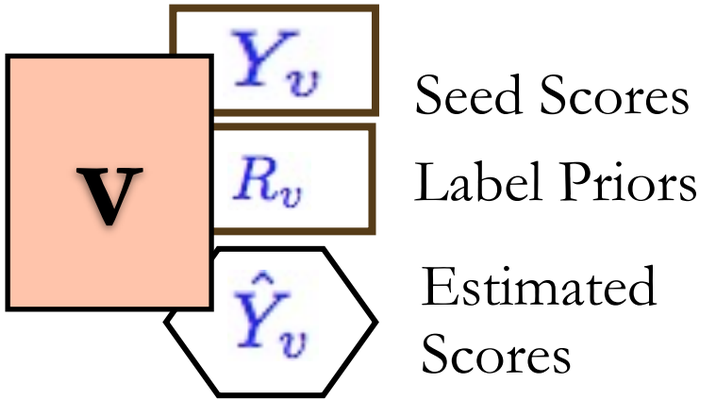
[Talukdar and Crammer, ECML 2009]

Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



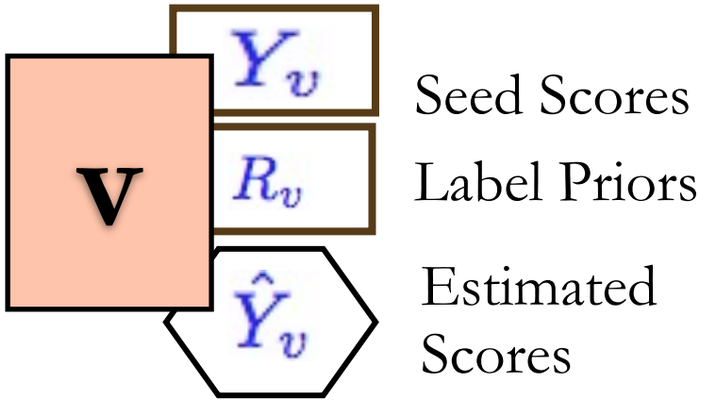
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

Match Seeds (soft)

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2} + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

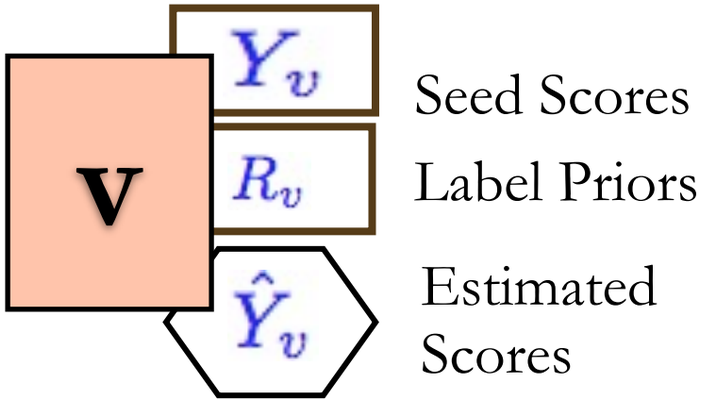


Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\overbrace{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}}^{\text{Match Seeds (soft)}} + \mu_1 \overbrace{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}}^{\text{Smooth}} + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v

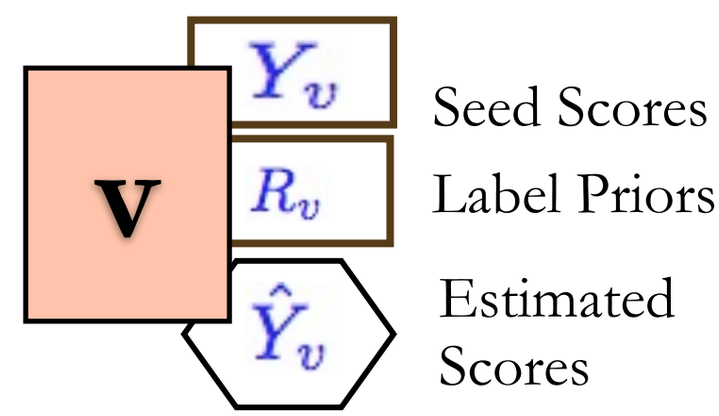


Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\underbrace{\boxed{\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2}}_{\text{Match Seeds (soft)}} + \mu_1 \underbrace{\boxed{\sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2}}_{\text{Smooth}} + \underbrace{\boxed{\mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2}}_{\text{Match Priors (Regularizer)}} \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} = \boldsymbol{W}'^{\top} + \boldsymbol{W}'$ is the symmetrized weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



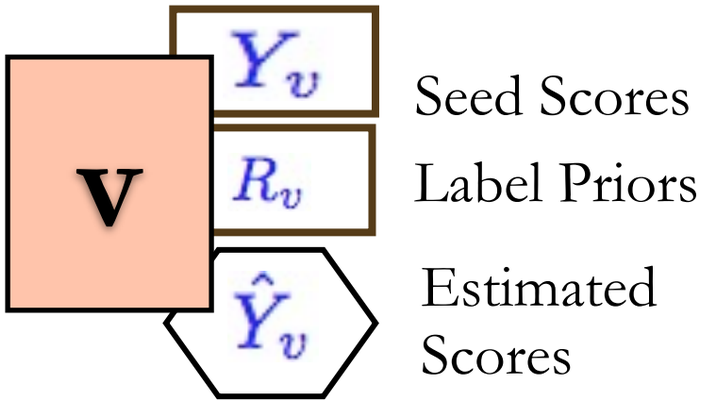
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

Match Seeds (soft) — Smooth — Match Priors (Regularizer)

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} =$ for *none-of-the-above* label ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



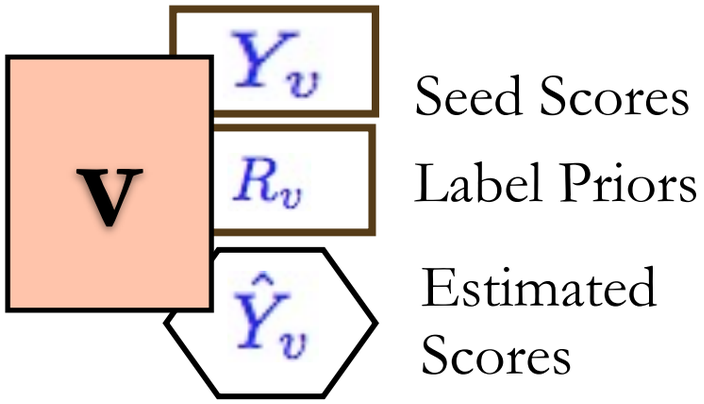
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

Match Seeds (soft) — Smooth — Match Priors (Regularizer)

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} =$ for *none-of-the-above* label ... ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



MAD has extra regularization compared to LP-ZGL [Zhu et al, ICML 03]; similar to QC [Bengio et al, 2006]

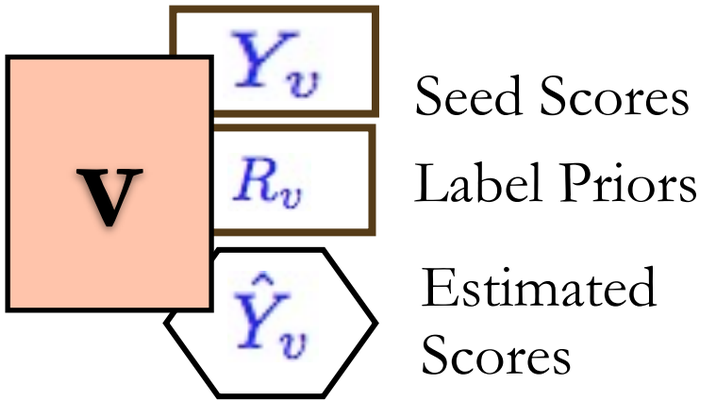
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

Match Seeds (soft) — Smooth — Match Priors (Regularizer)

$$\arg\min_{\hat{\boldsymbol{Y}}} \sum_{l=1}^{m+1} \left[\|\boldsymbol{S}\hat{\boldsymbol{Y}}_l - \boldsymbol{S}\boldsymbol{Y}_l\|^2 + \mu_1 \sum_{u,v} \boldsymbol{M}_{uv}(\hat{\boldsymbol{Y}}_{ul} - \hat{\boldsymbol{Y}}_{vl})^2 + \mu_2 \|\hat{\boldsymbol{Y}}_l - \boldsymbol{R}_l\|^2 \right]$$

- m labels, $+1$ dummy label
- $\boldsymbol{M} =$ for *none-of-the-above* label ... ed weight matrix
- $\hat{\boldsymbol{Y}}_{vl}$: weight of label l on node v
- $\boldsymbol{Y}_{vl}$: seed weight for label l on node v
- $\boldsymbol{S}$: diagonal matrix, nonzero for seed nodes
- $\boldsymbol{R}_{vl}$: regularization target for label l on node v



MAD's Objective is Convex

MAD has extra regularization compared to LP-ZGL [Zhu et al, ICML 03]; similar to QC [Bengio et al, 2006]

Solving MAD Objective

Solving MAD Objective

- Can be solved using matrix inversion (like in LP)
  - but matrix inversion is expensive

Solving MAD Objective

- Can be solved using matrix inversion (like in LP)
  - but matrix inversion is expensive
- Instead solved exactly using a system of linear equations (Ax = b)
  - solved using Jacobi iterations
  - results in iterative updates
  - guaranteed convergence
  - see [Bengio et al., 2006] and [Talukdar and Crammer, ECML 2009] for details

Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal

$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$

$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}^{\dagger}$

$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$

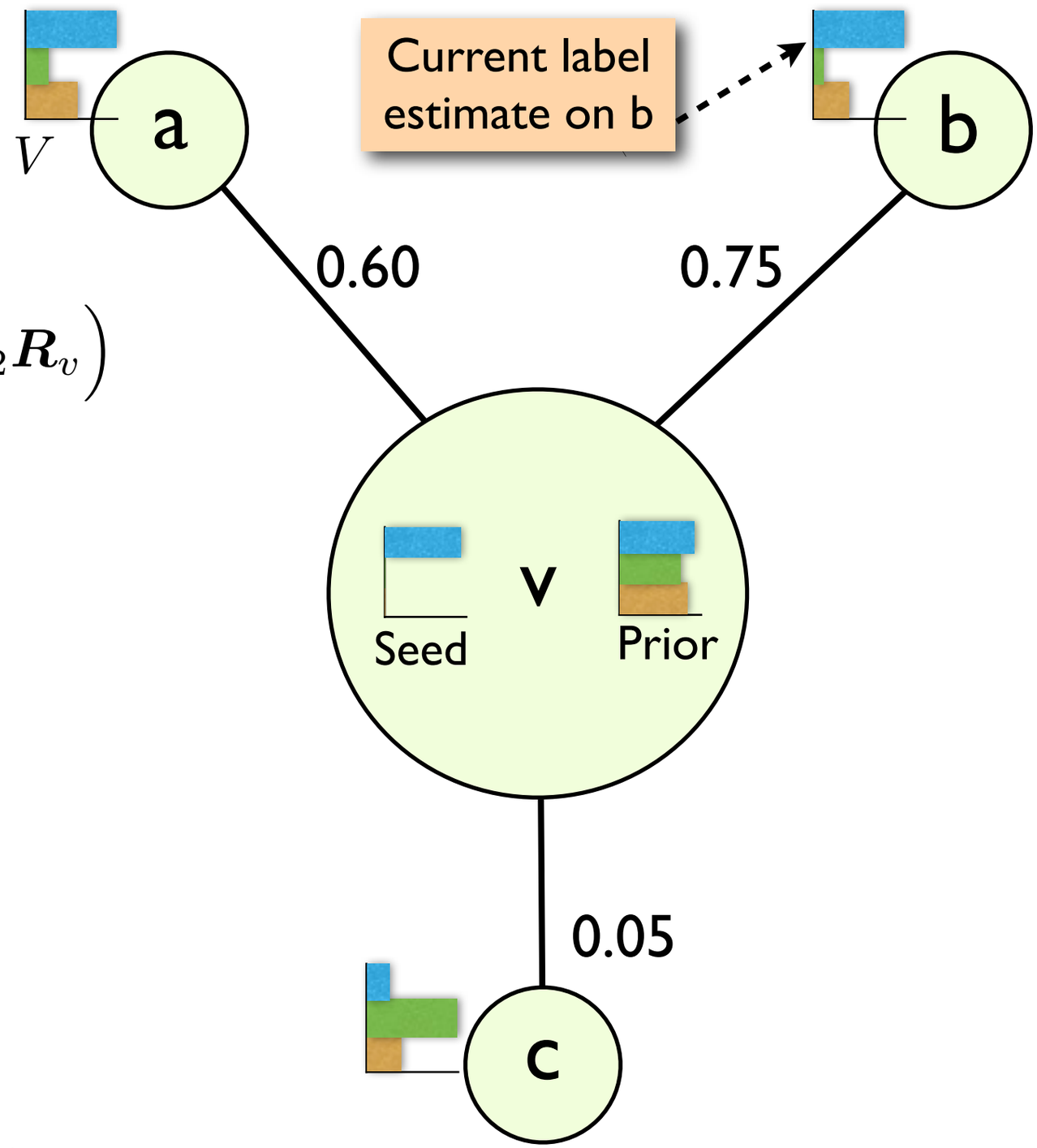
repeat

  for all $v \in V$ **do**

    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v} \left((\boldsymbol{S}\boldsymbol{Y})_v + \mu_1 \boldsymbol{M}_{v\cdot} \hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v \right)$

  end for

until convergence



Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal

$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$

$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}^{\dagger}$

$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$

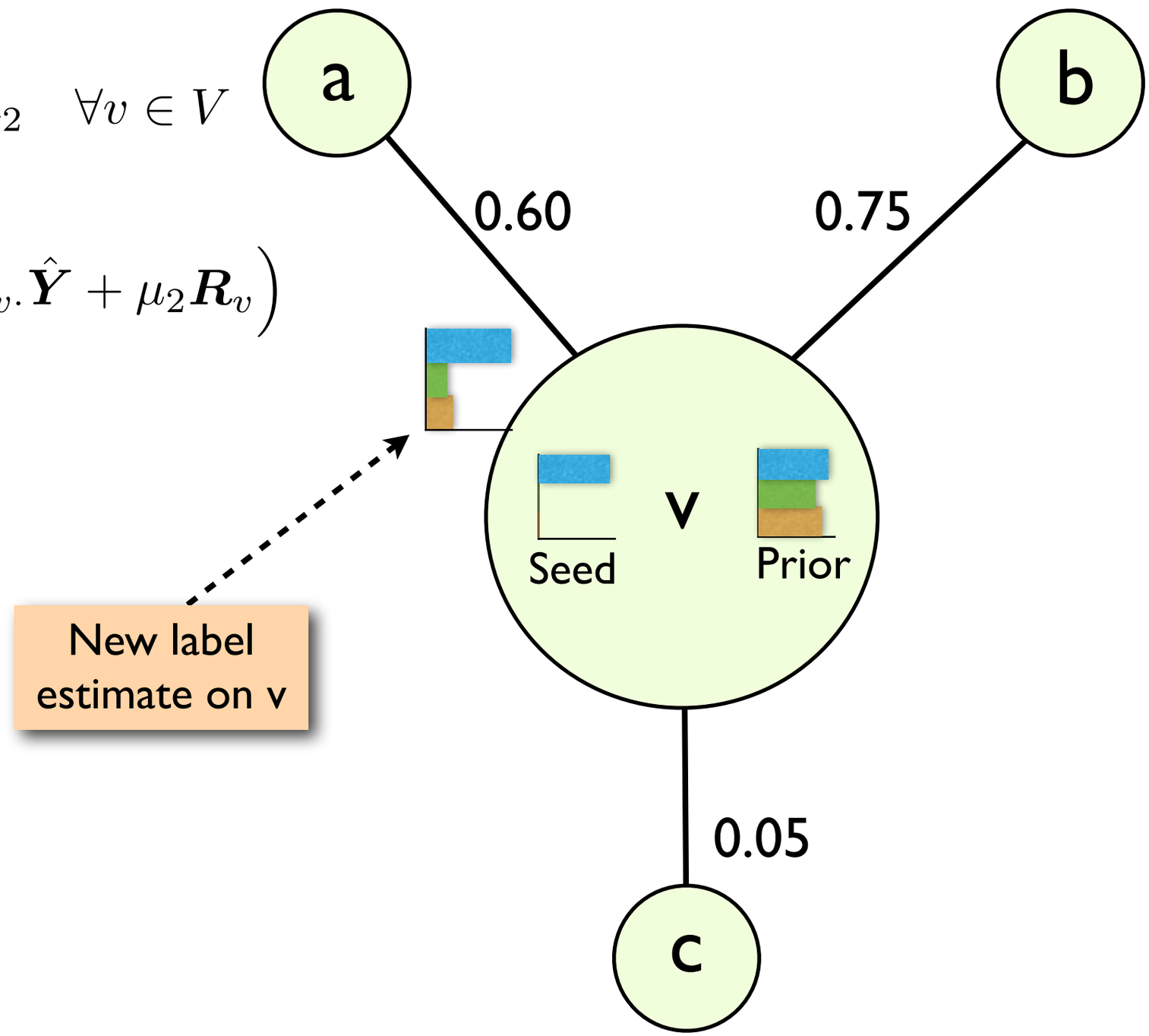
repeat

  for all $v \in V$ **do**

    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v} \left((\boldsymbol{S}\boldsymbol{Y})_v + \mu_1 \boldsymbol{M}_{v\cdot} \hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v \right)$

  end for

until convergence



Solving MAD using Iterative Updates

Inputs $\boldsymbol{Y}, \boldsymbol{R} : |V| \times (|L|+1)$, $\boldsymbol{W} : |V| \times |V|$, $\boldsymbol{S} : |V| \times |V|$ diagonal

$\hat{\boldsymbol{Y}} \leftarrow \boldsymbol{Y}$

$\boldsymbol{M} = \boldsymbol{W}' + \boldsymbol{W}^{\dagger}$

$Z_v \leftarrow \boldsymbol{S}_{vv} + \mu_1 \sum_{u \neq v} \boldsymbol{M}_{vu} + \mu_2 \quad \forall v \in V$

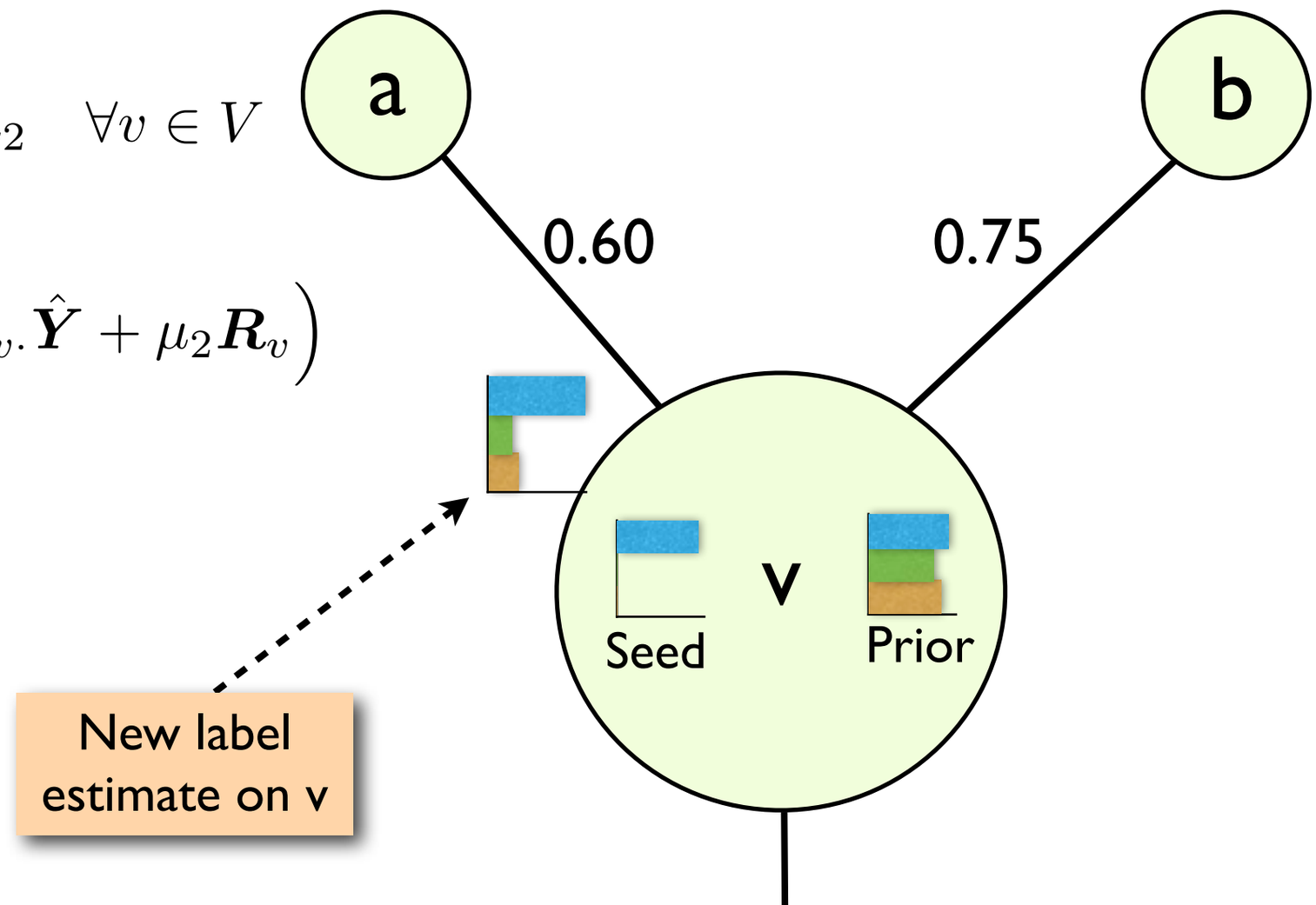
repeat

  for all $v \in V$ **do**

    $\hat{\boldsymbol{Y}}_v \leftarrow \frac{1}{Z_v}\left((\boldsymbol{SY})_v + \mu_1 \boldsymbol{M}_{v\cdot}\hat{\boldsymbol{Y}} + \mu_2 \boldsymbol{R}_v\right)$

  end for

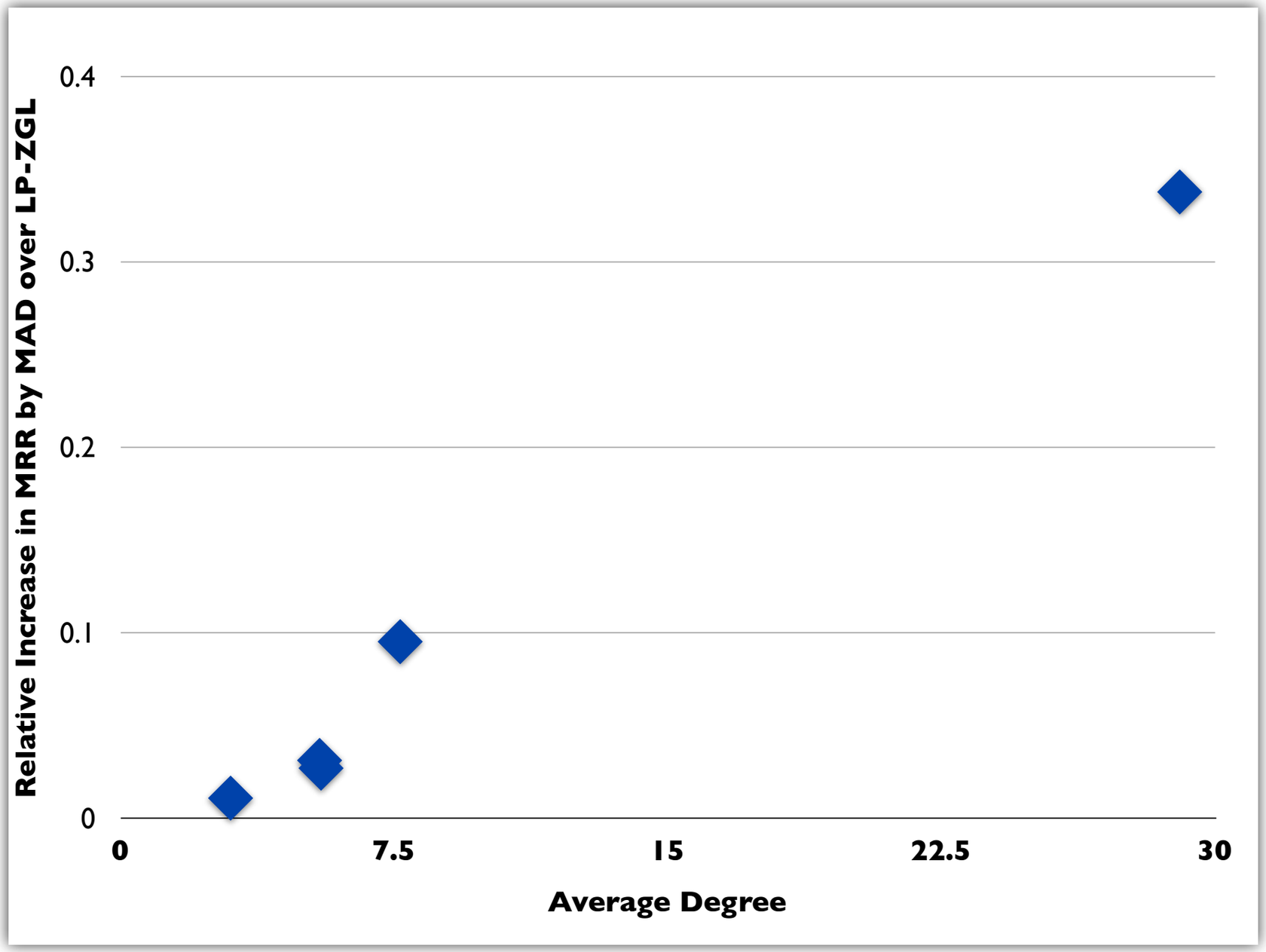
until convergence



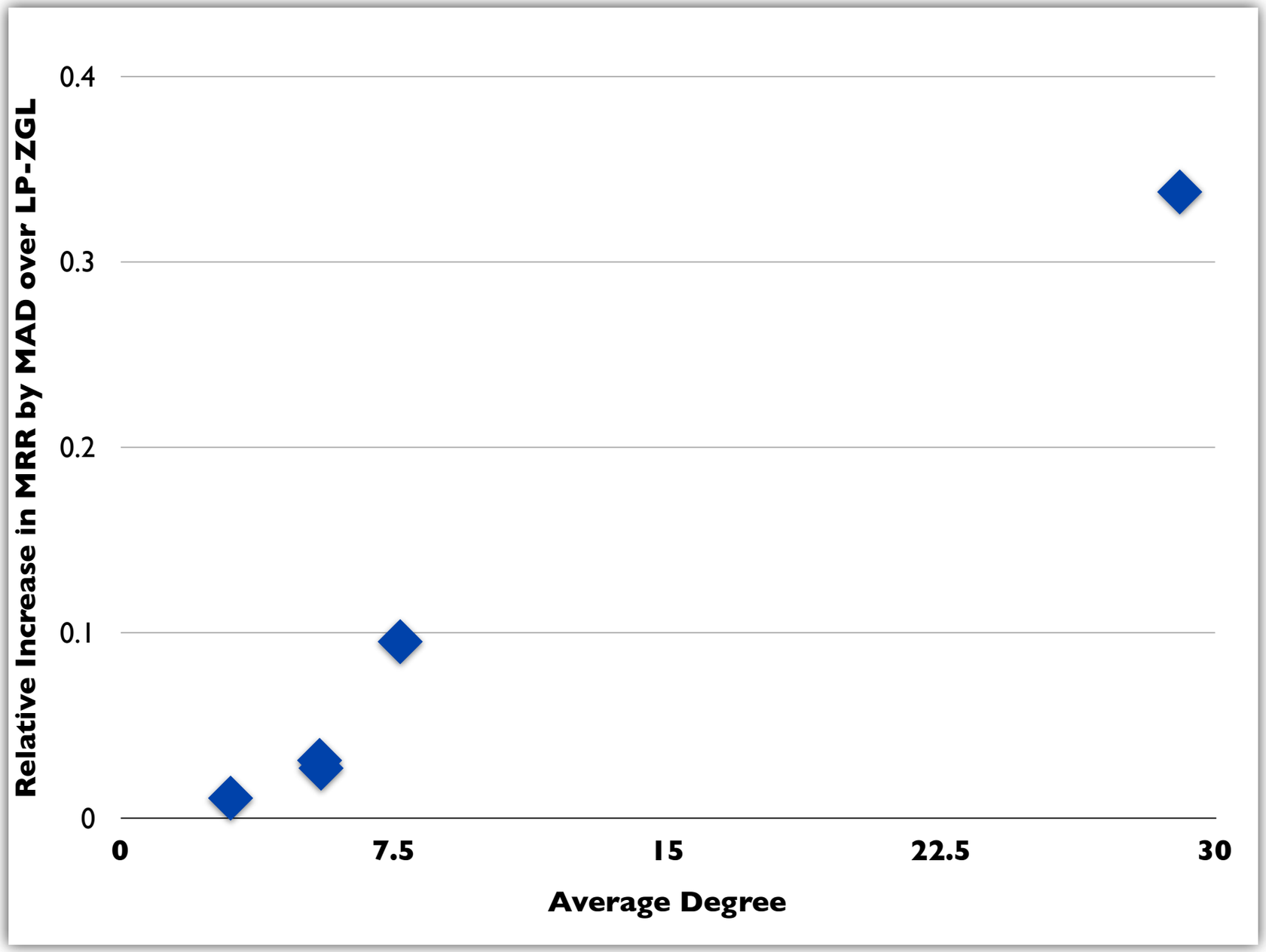
- Importance of a node can be discounted
- Easily Parallelizable: Scalable (more later)

When is MAD most effective?

When is MAD most effective?



When is MAD most effective?



MAD is particularly effective in denser graphs, where there is greater need for regularization.

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Transduction Algorithm with Confidence (TACO) [Orbach and Crammer, ECML 2012]

Transduction Algorithm with Confidence (TACO) [Orbach and Crammer, ECML 2012]

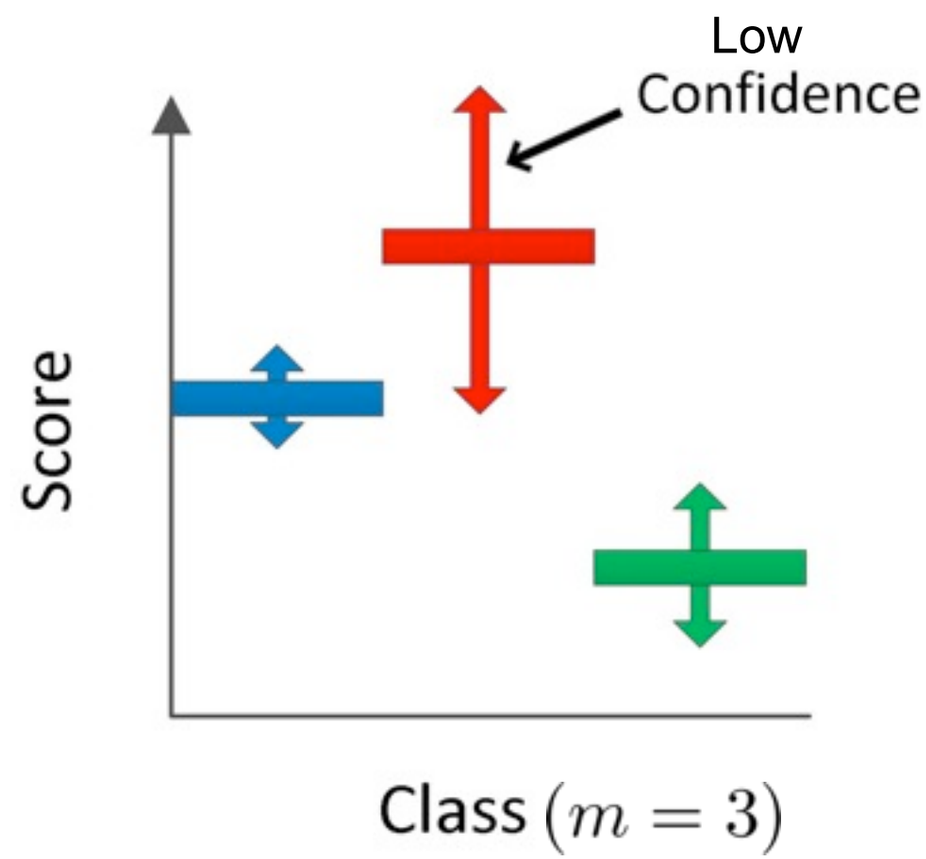
- Main lesson from MAD: discount *bad* (high degree) nodes

Transduction Algorithm with Confidence (TACO) [Orbach and Crammer, ECML 2012]

- Main lesson from MAD: discount *bad* (high degree) nodes
- TACO generalizes notion of bad, adds per-node, per-class **confidence**

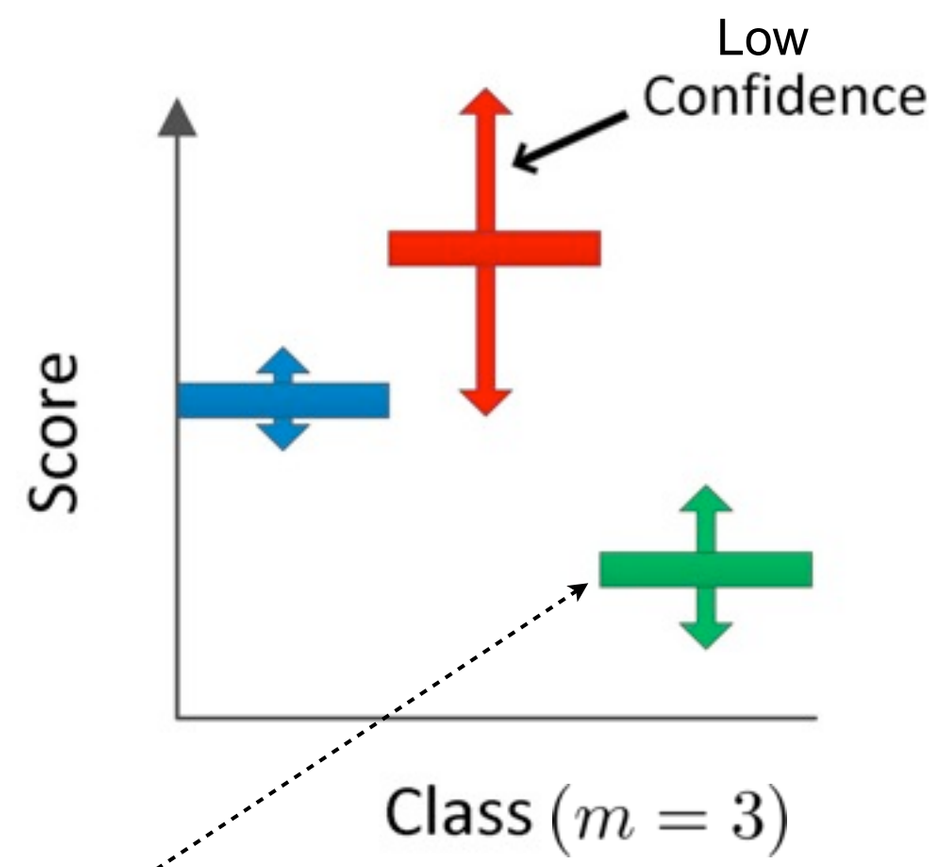
Transduction Algorithm with Confidence (TACO) [Orbach and Crammer, ECML 2012]

- Main lesson from MAD: discount *bad* (high degree) nodes
- TACO generalizes notion of bad, adds per-node, per-class **confidence**



Transduction Algorithm with Confidence (TACO) [Orbach and Crammer, ECML 2012]

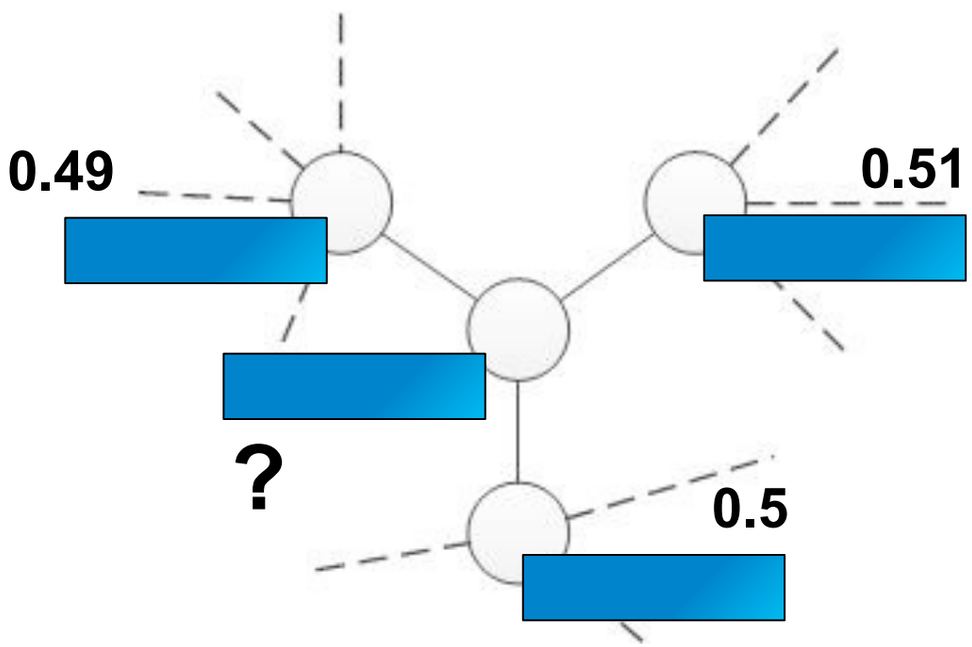
- Main lesson from MAD: discount *bad* (high degree) nodes
- TACO generalizes notion of bad, adds per-node, per-class **confidence**

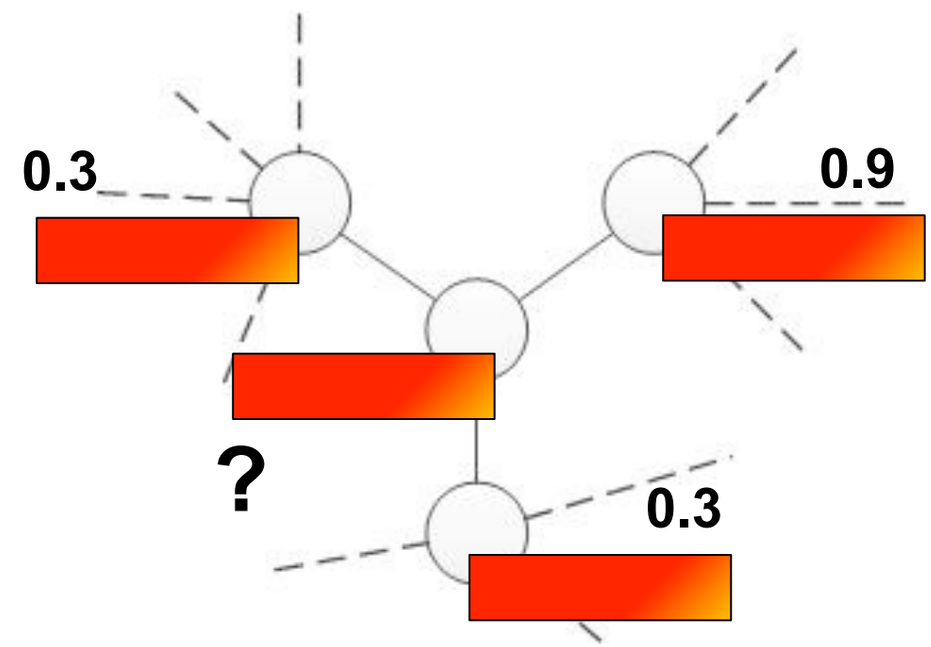


Label Scores: $\boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i,1} & \dots & \mu_{i,m} \end{bmatrix} \in \mathbb{R}^m$

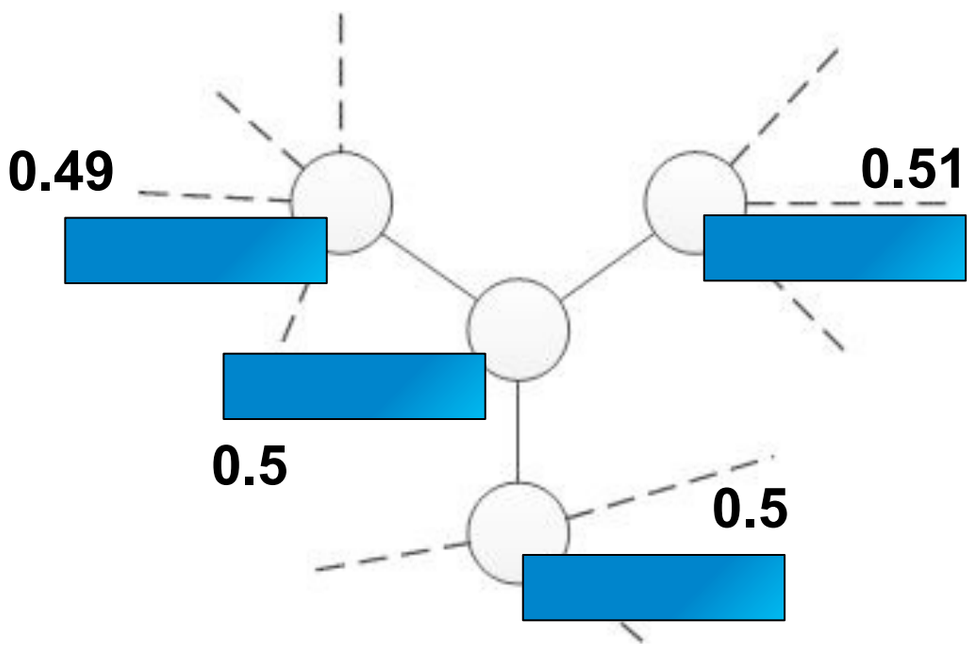
Confidence: $\boldsymbol{\sigma}_i = \begin{bmatrix} \sigma_{i,1} & \dots & \sigma_{i,m} \end{bmatrix} \in \mathbb{R}^m$

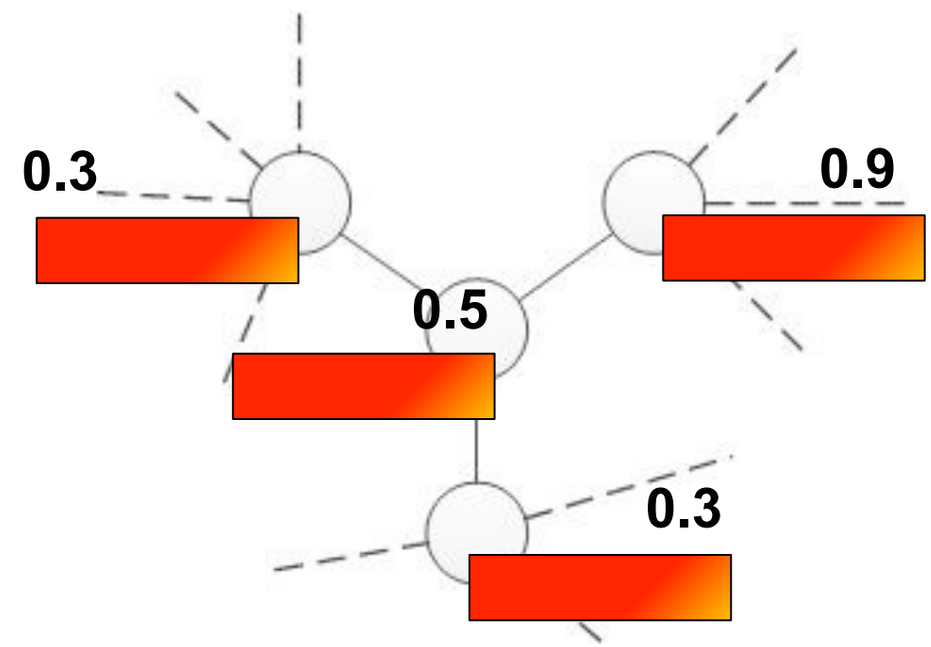
Introducing Confidence



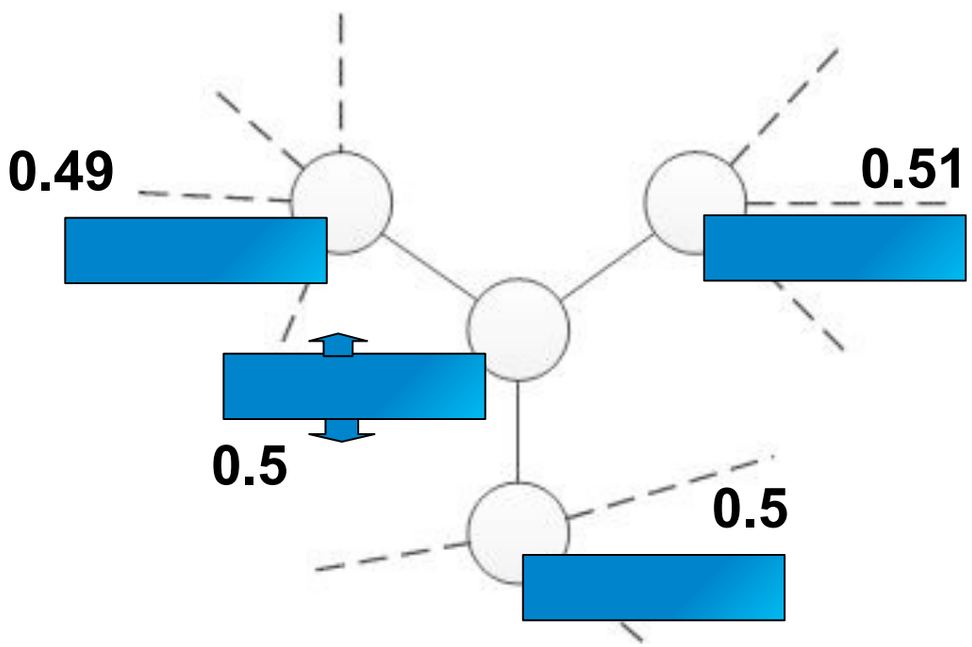


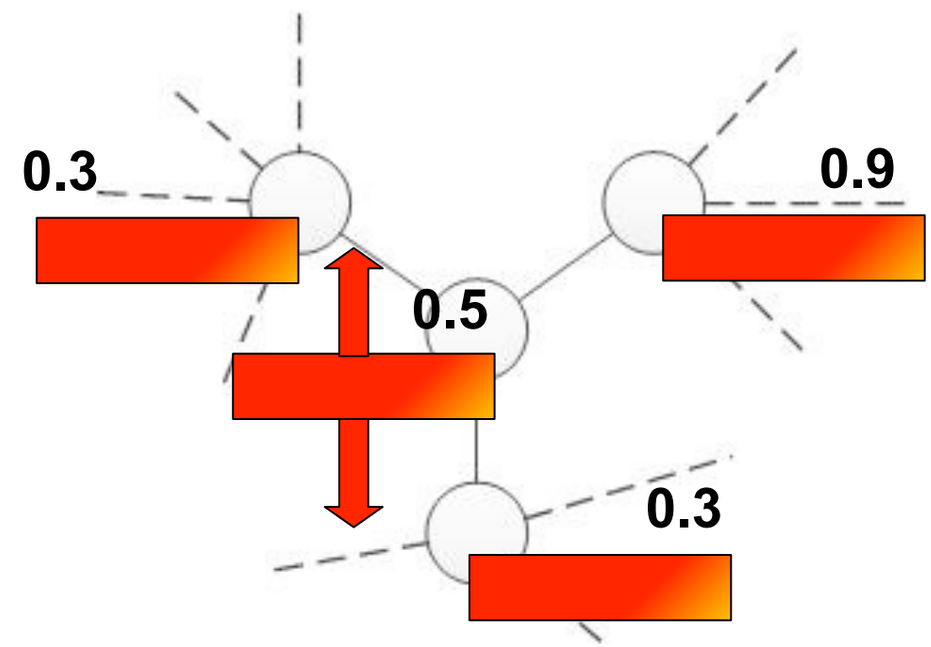
Introducing Confidence



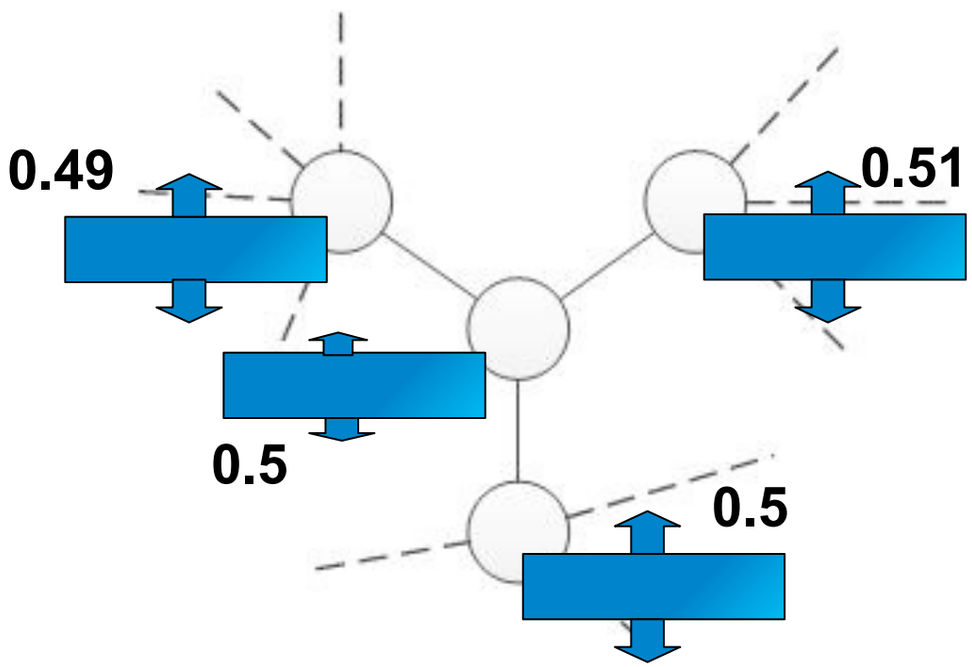


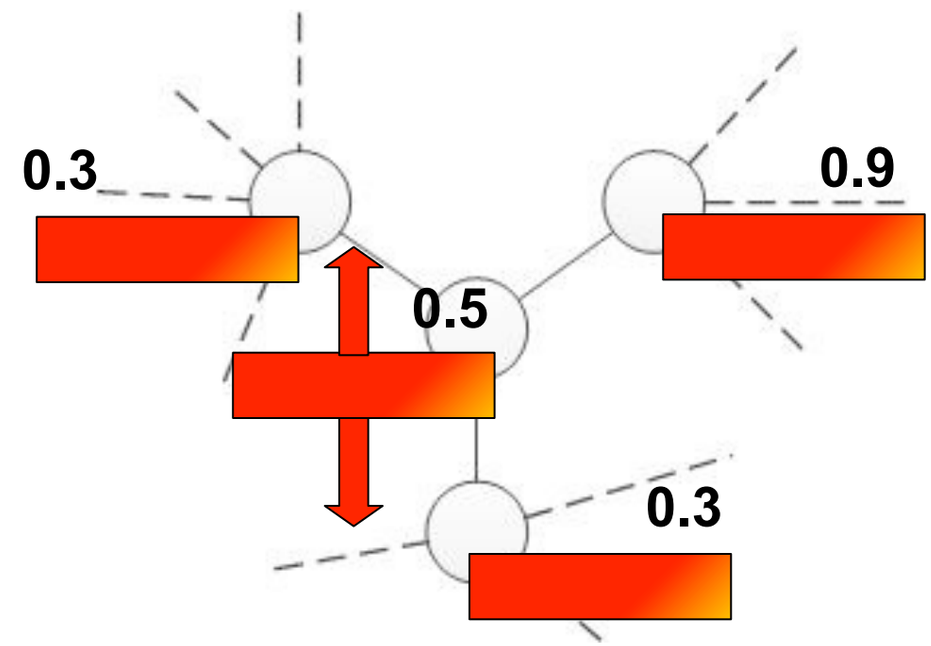
Introducing Confidence



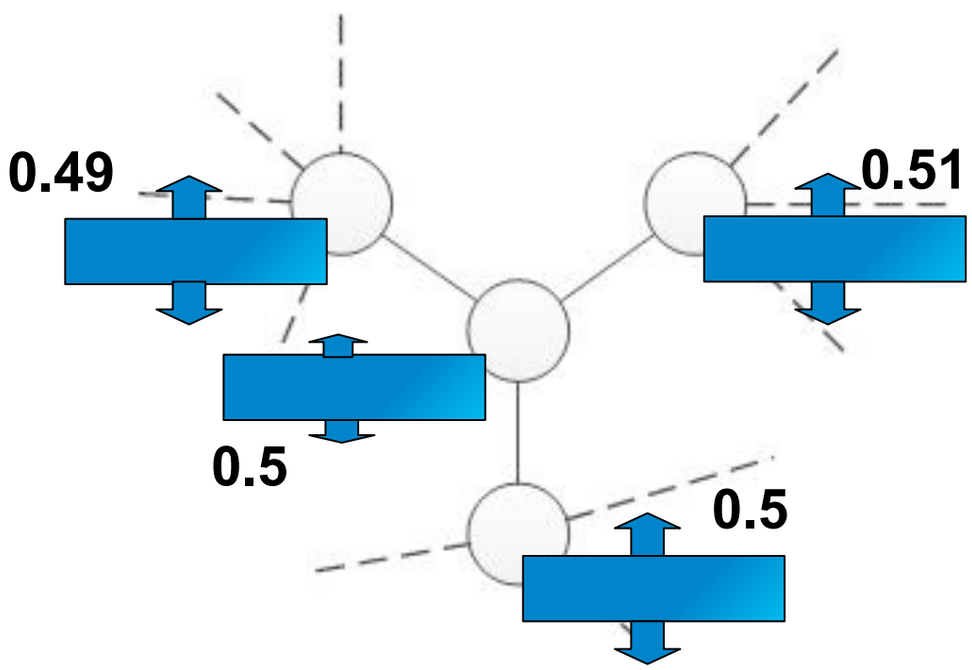


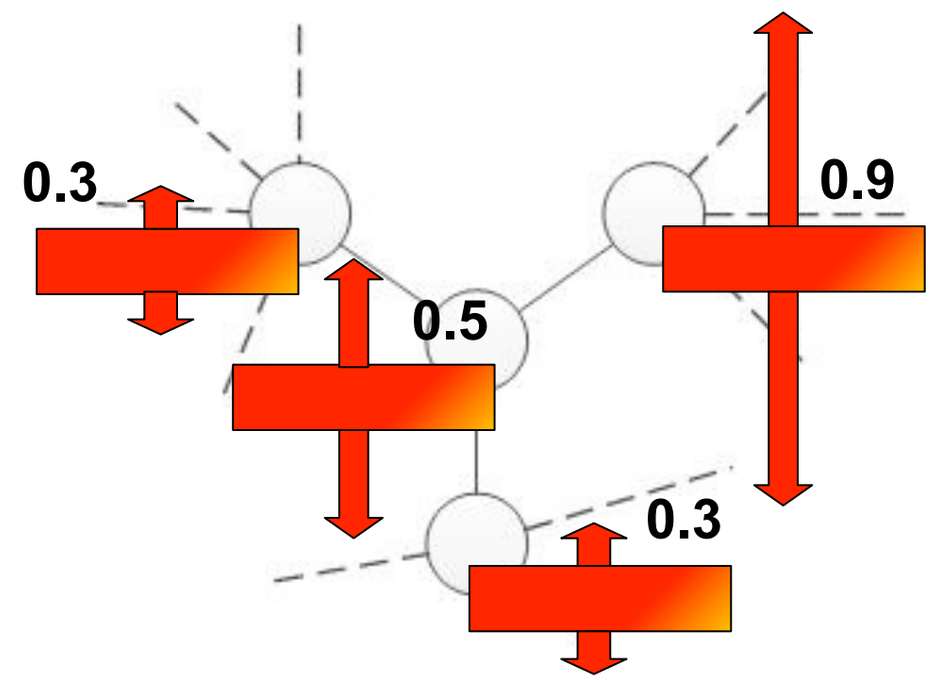
Introducing Confidence



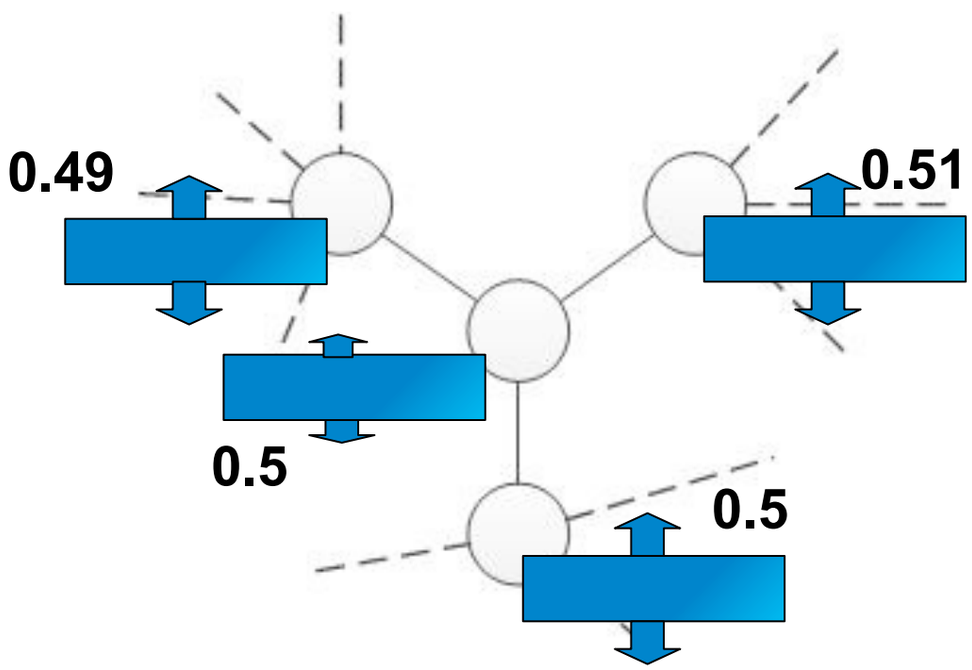


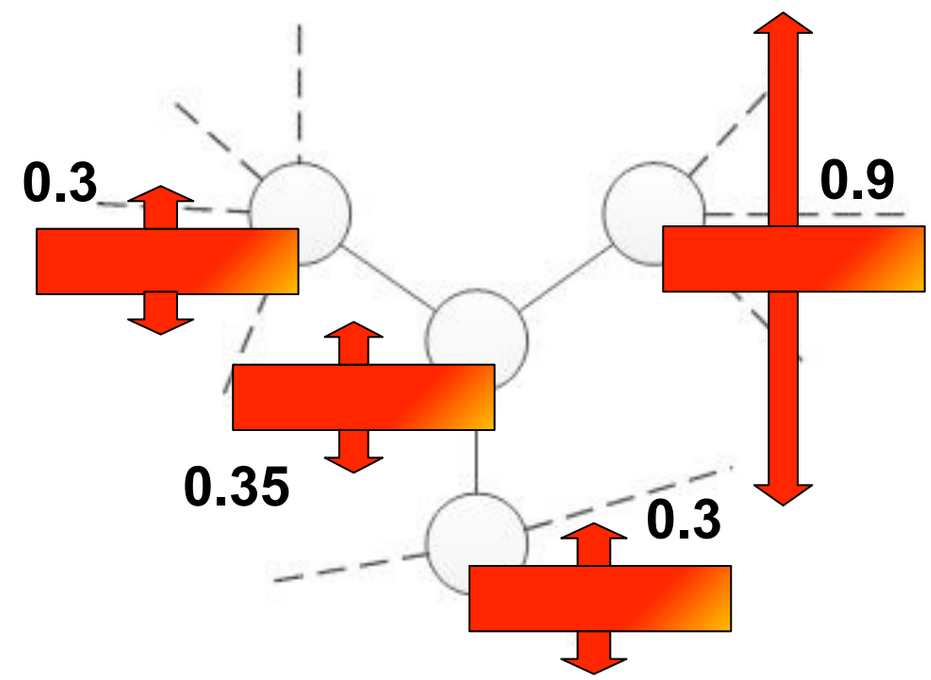
Introducing Confidence



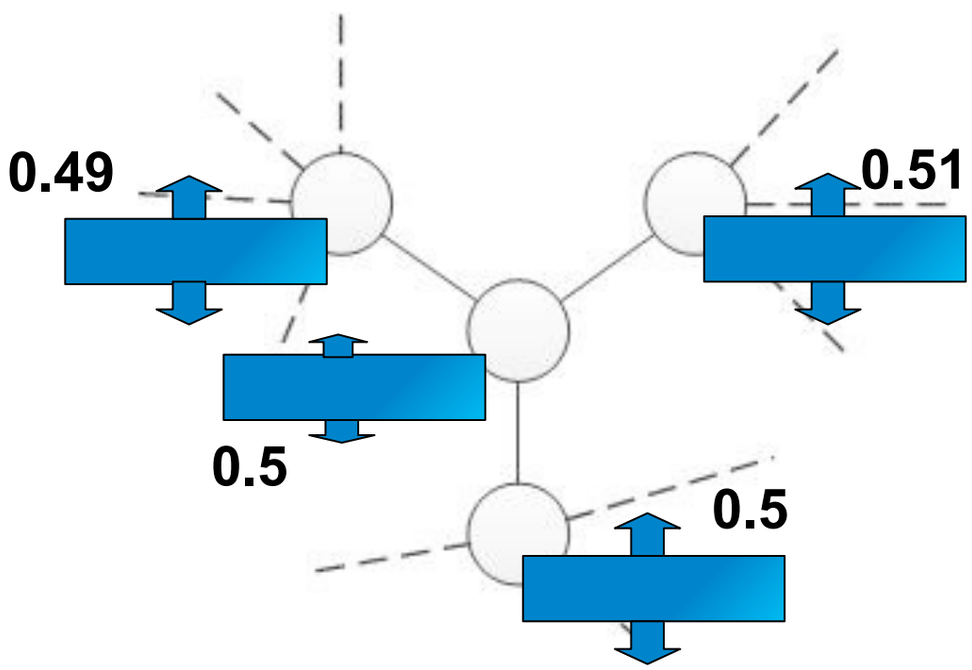


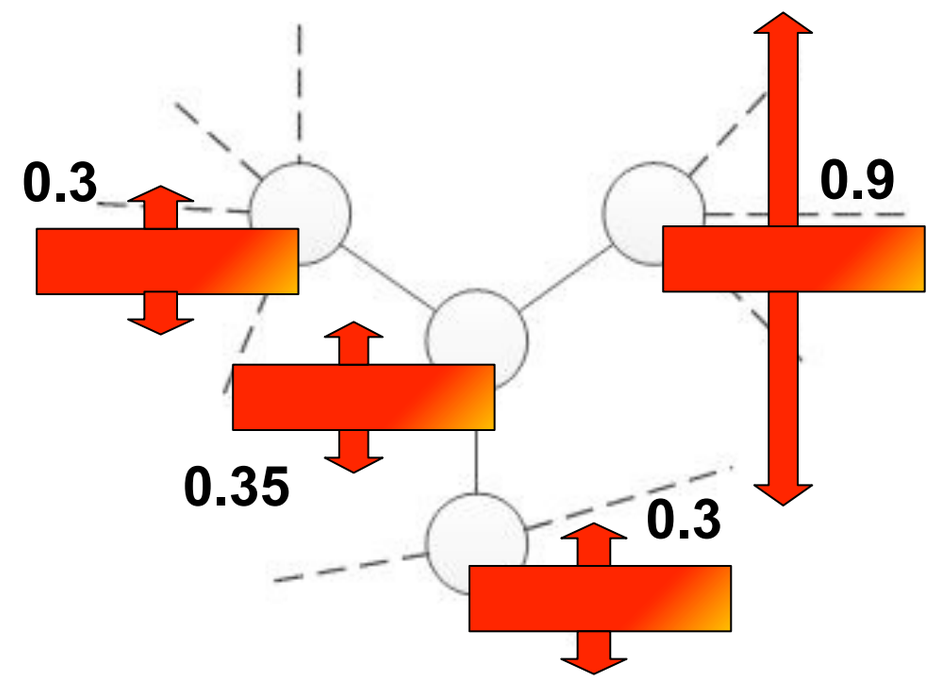
Introducing Confidence





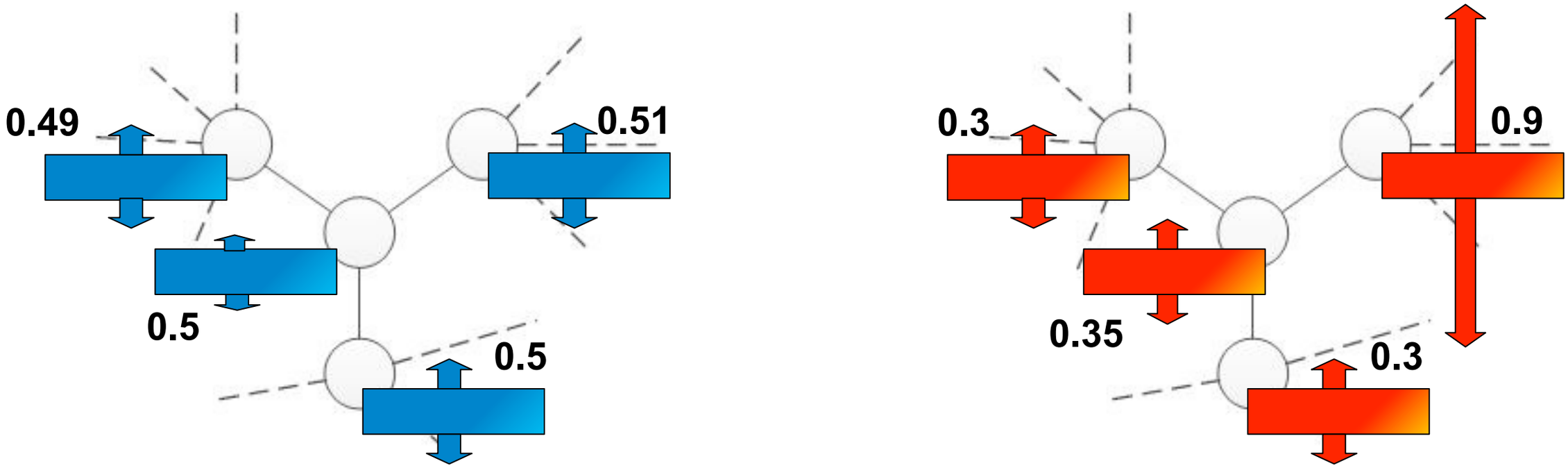
Introducing Confidence





- Neighborhood disagreement => low confidence

Introducing Confidence



- Neighborhood disagreement => low confidence
- Lower the effect of poorly estimated (low confidence) scores

TACO Objective

$$\underset{\{\boldsymbol{\mu}_i\},\{\boldsymbol{\sigma}_i\}}{\arg\min} \overbrace{\sum_{i,j=1}^{n} w_{i,j} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)}^{\text{manifold term}} + \overbrace{\sum_{i=1}^{n_l} D\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)}^{\text{labeled term}} + \overbrace{\sum_{i=1}^{n} R\left(\boldsymbol{x}_i\right)}^{\text{regularization}}$$

TACO Objective

$$\underset{\{\boldsymbol{\mu}_i\},\{\boldsymbol{\sigma}_i\}}{\arg\min} \overbrace{\sum_{i,j=1}^{n} w_{i,j} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)}^{\text{manifold term}} + \overbrace{\sum_{i=1}^{n_l} D\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)}^{\text{labeled term}} + \overbrace{\sum_{i=1}^{n} R\left(\boldsymbol{x}_i\right)}^{\text{regularization}}$$

$$D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sum_{r=1}^{m} \left(\frac{1}{\sigma_{i,r}} + \frac{1}{\sigma_{j,r}}\right) \left(\mu_{i,r} - \mu_{j,r}\right)^2$$

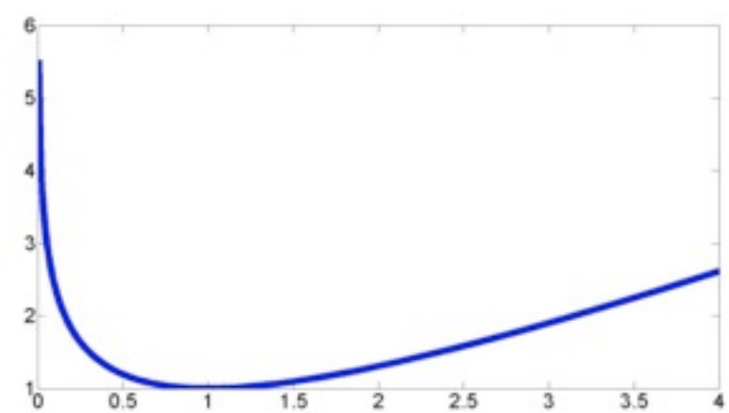
TACO Objective

$$\underset{\{\boldsymbol{\mu}_i\},\{\boldsymbol{\sigma}_i\}}{\arg\min} \overbrace{\sum_{i,j=1}^{n} w_{i,j} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)}^{\text{manifold term}} + \overbrace{\sum_{i=1}^{n_l} D\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)}^{\text{labeled term}} + \overbrace{\sum_{i=1}^{n} R\left(\boldsymbol{x}_i\right)}^{\text{regularization}}$$

$$D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sum_{r=1}^{m} \left(\frac{1}{\sigma_{i,r}} + \frac{1}{\sigma_{j,r}} \right) \left(\mu_{i,r} - \mu_{j,r}\right)^2$$

$$R\left(\boldsymbol{x}_i\right) = \alpha \sum_{r=1}^{m} \sigma_{i,r} - \beta \sum_{r=1}^{m} \log \sigma_{i,r}$$

$$\alpha x - \beta \log\left(x\right)$$



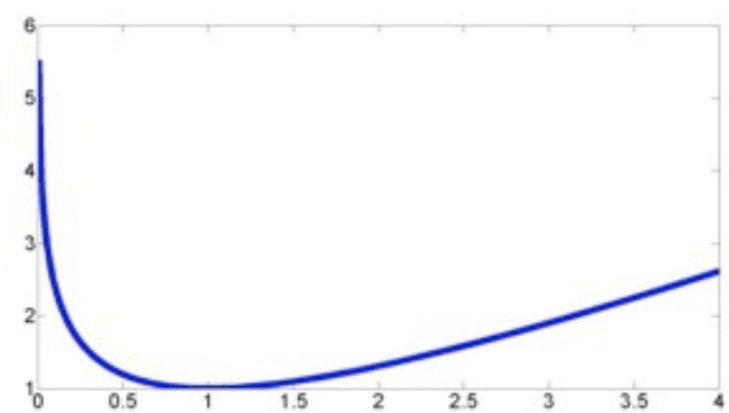
TACO Objective

$$\underset{\{\boldsymbol{\mu}_i\},\{\boldsymbol{\sigma}_i\}}{\arg\min} \overbrace{\sum_{i,j=1}^{n} w_{i,j} D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)}^{\text{manifold term}} + \overbrace{\sum_{i=1}^{n_l} D\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)}^{\text{labeled term}} + \overbrace{\sum_{i=1}^{n} R\left(\boldsymbol{x}_i\right)}^{\text{regularization}}$$

$$D\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \sum_{r=1}^{m} \left(\frac{1}{\sigma_{i,r}} + \frac{1}{\sigma_{j,r}} \right) \left(\mu_{i,r} - \mu_{j,r}\right)^2$$

$$R\left(\boldsymbol{x}_i\right) = \alpha \sum_{r=1}^{m} \sigma_{i,r} - \beta \sum_{r=1}^{m} \log \sigma_{i,r}$$

$$\alpha x - \beta \log\left(x\right)$$



- Convex objective
- Iterative solution

TACO: Iterative Algorithm

Parameters: $\alpha, \beta > 0$ (controls regularization)
- $\gamma > 0$ (labeled confidence)

Input: Graph $G = (V, E, W)$ and prior labeling $\boldsymbol{y}_i$ for each $v_i \in V$

Initialize: $\boldsymbol{\mu}_i = \mathbf{0}$ and $\boldsymbol{\sigma}_i = \mathbf{1}$ for all $v_i \in V$

Repeat updates:

- For $v_i \in V$: [$C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)$ is the objective]

$$\frac{\partial C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)}{\partial \boldsymbol{\mu}_i} = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_i = \ldots$$

$$\frac{\partial C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)}{\partial \boldsymbol{\sigma}_i} = 0 \quad \Rightarrow \quad \boldsymbol{\sigma}_i = \ldots$$

Until convergence

Output: Estimated scores $\{\boldsymbol{\mu}_i\}$

TACO: Iterative Algorithm

Parameters: $\alpha, \beta > 0$ (controls regularization)
- $\gamma > 0$ (labeled confidence)

Input: Graph $G = (V, E, W)$ and prior labeling $\boldsymbol{y}_i$ for each $v_i \in V$

Initialize: $\boldsymbol{\mu}_i = \mathbf{0}$ and $\boldsymbol{\sigma}_i = \mathbf{1}$ for all $v_i \in V$

Repeat updates:

- For $v_i \in V$: [$C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)$ is the objective]

$$\frac{\partial C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)}{\partial \boldsymbol{\mu}_i} = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_i = \ldots$$

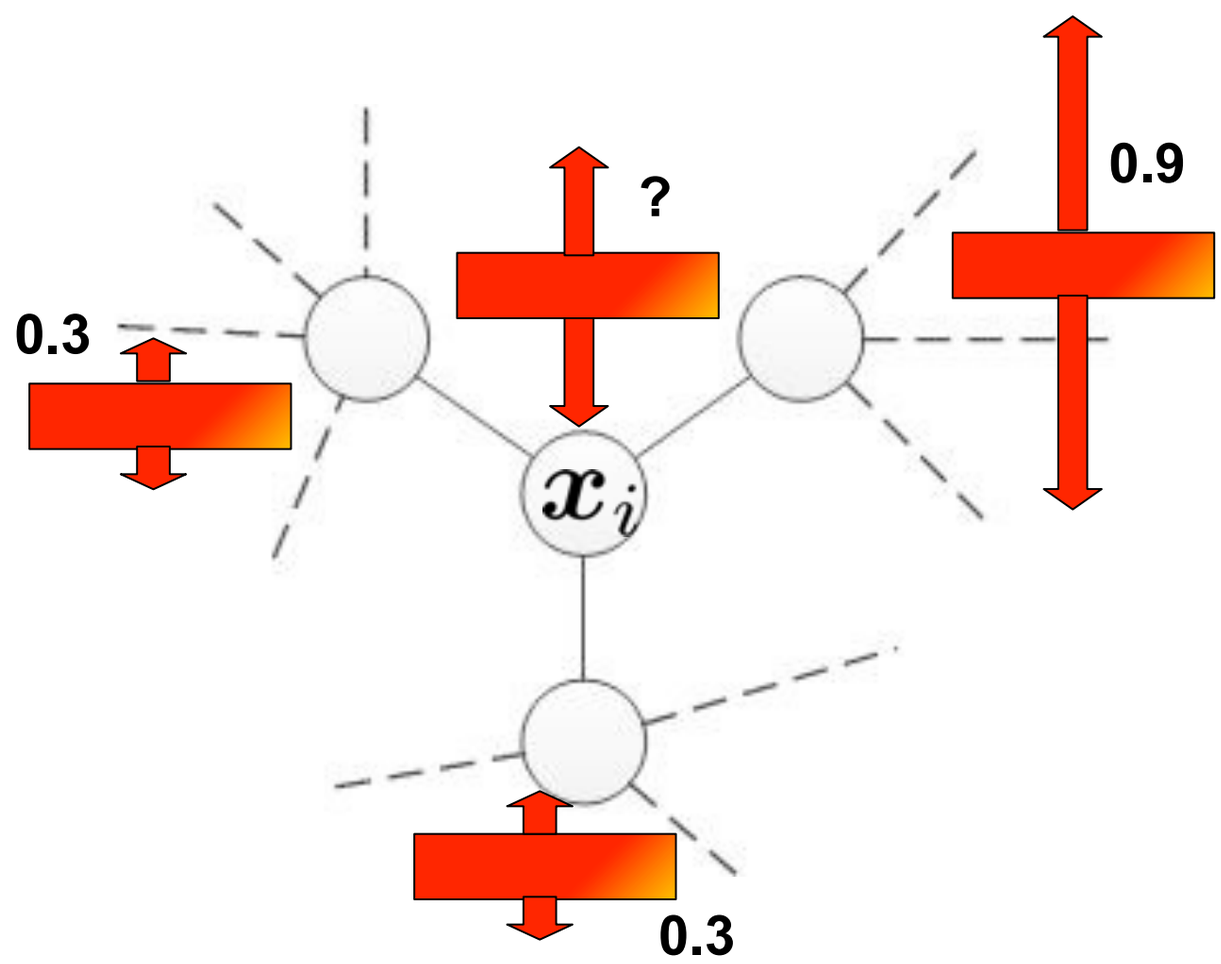
$$\frac{\partial C\left(G, \{\boldsymbol{\mu}_j\}, \{\boldsymbol{\sigma}_j\}\right)}{\partial \boldsymbol{\sigma}_i} = 0 \quad \Rightarrow \quad \boldsymbol{\sigma}_i = \ldots$$

Until convergence

Output: Estimated scores $\{\boldsymbol{\mu}_i\}$

TACO: Score Update

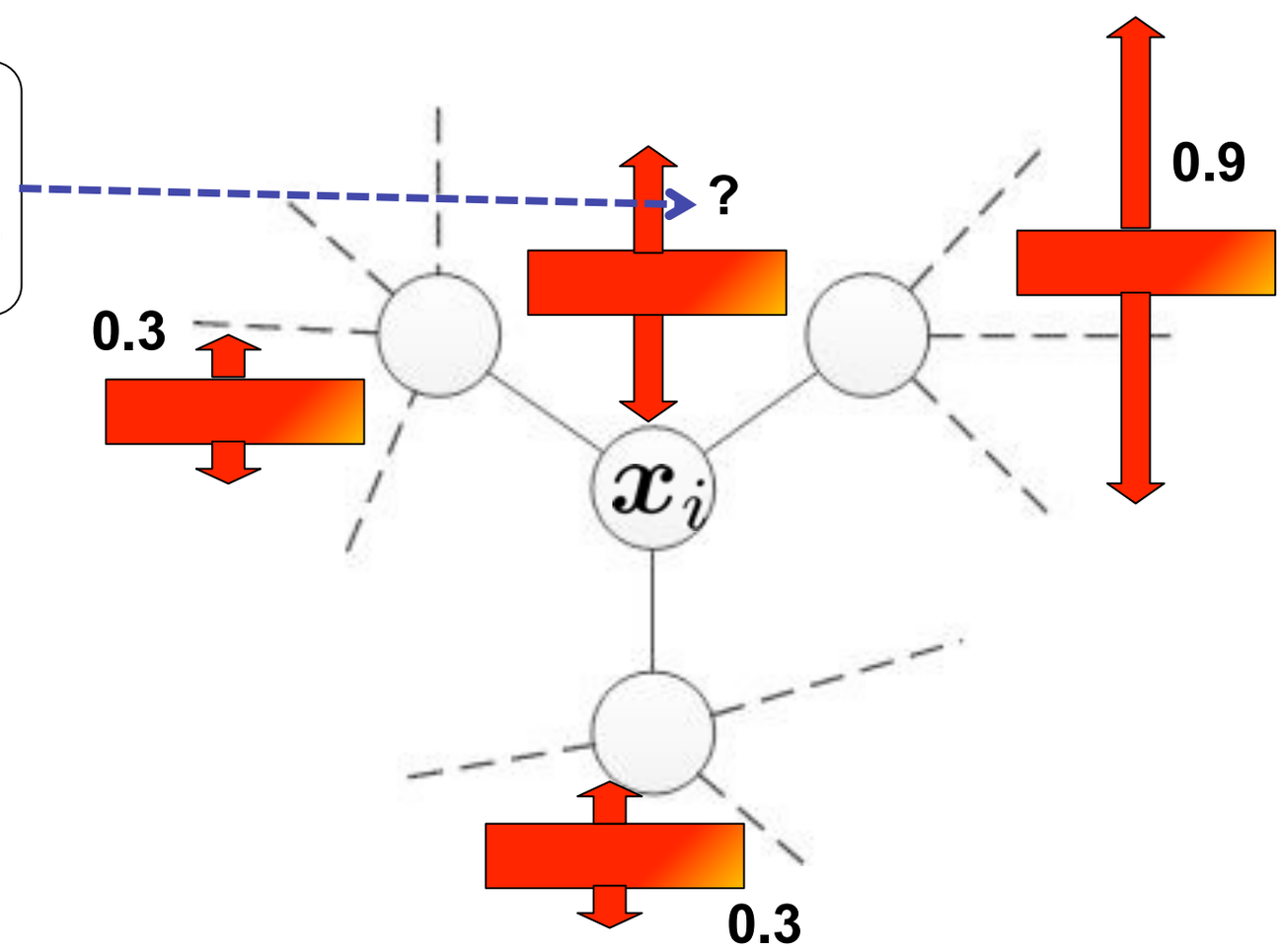
$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$



TACO: Score Update

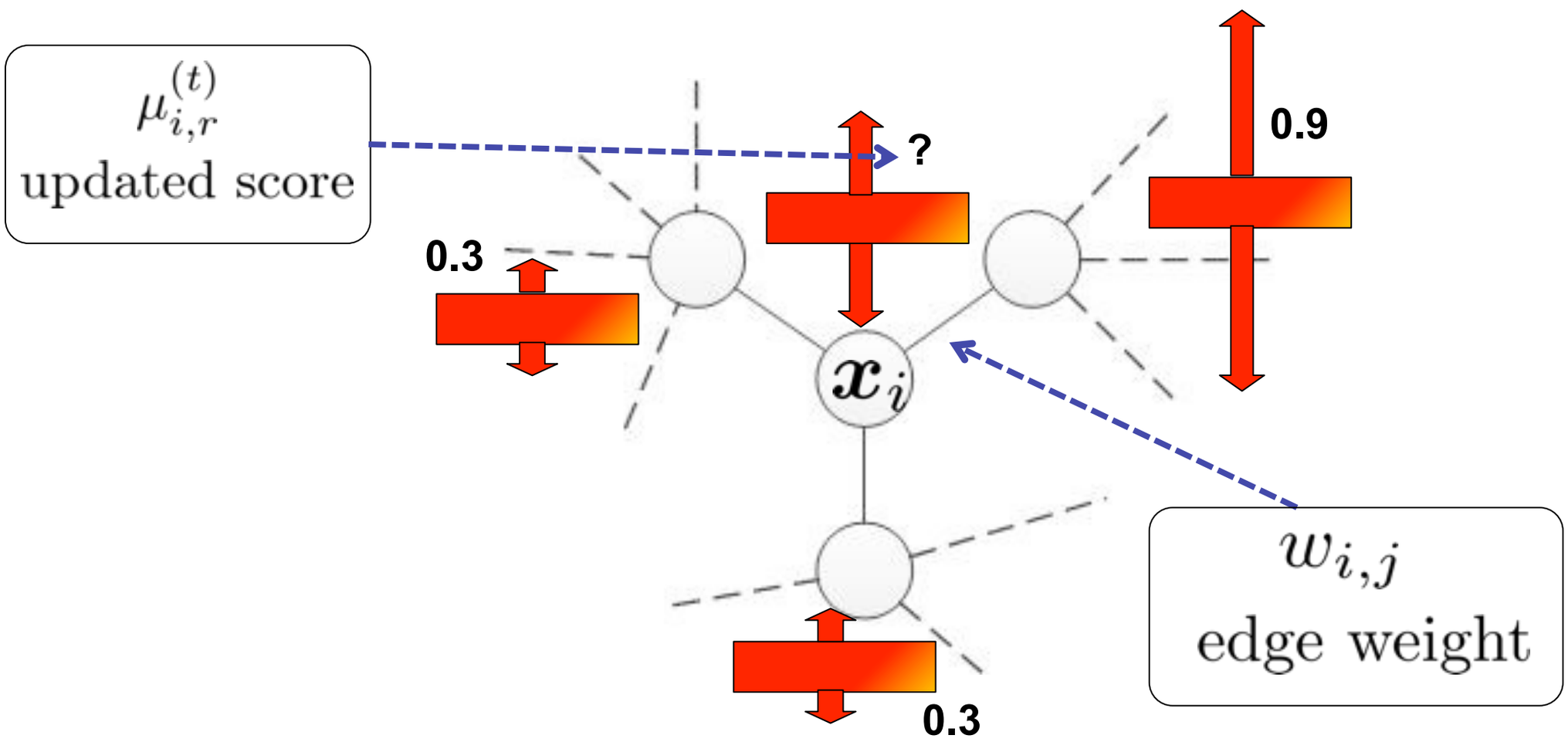
$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$

$\mu_{i,r}^{(t)}$
updated score



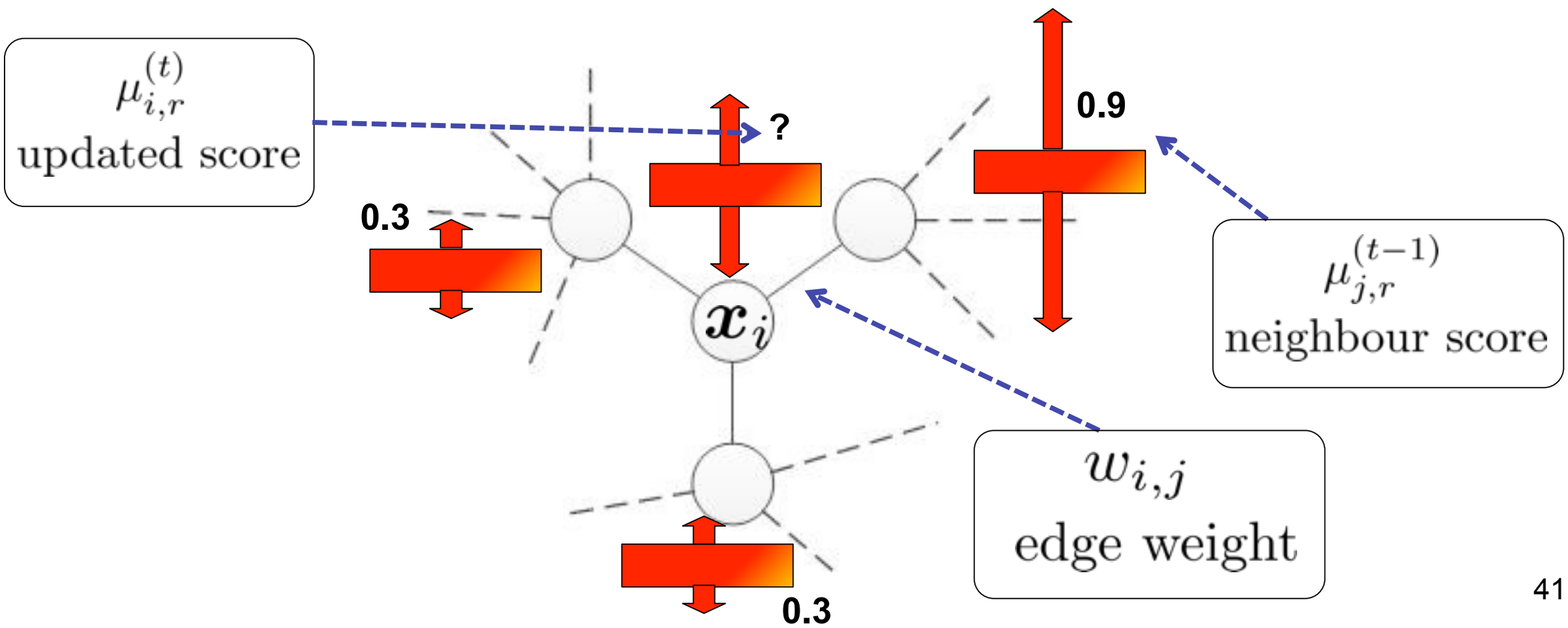
TACO: Score Update

$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$



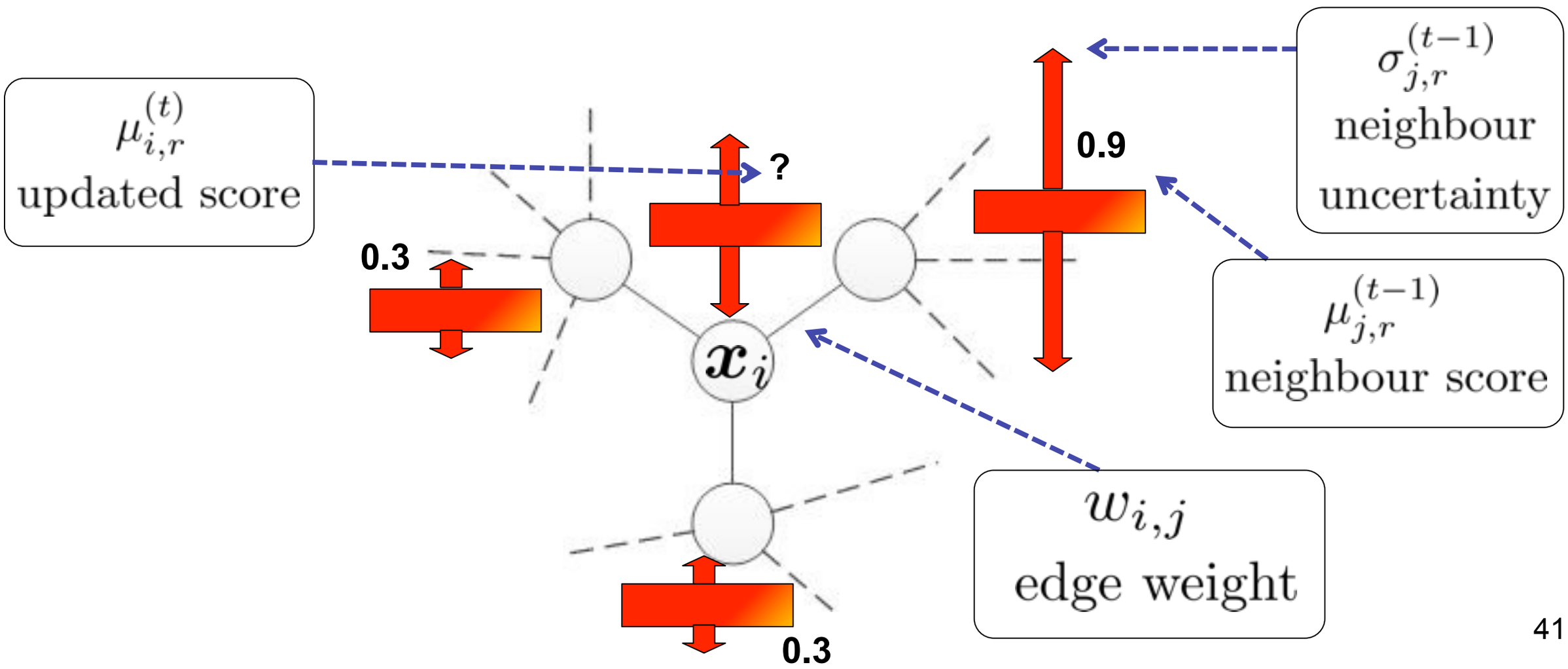
TACO: Score Update

$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$



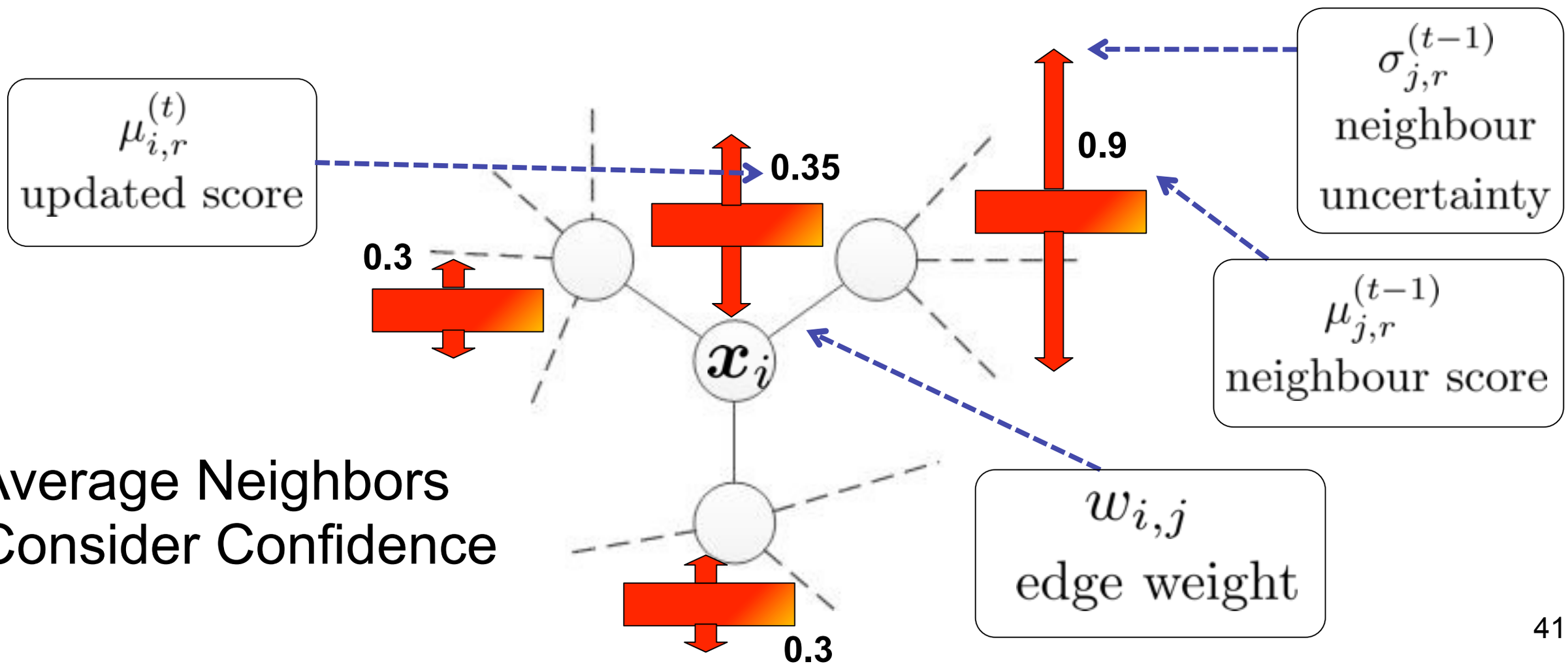
TACO: Score Update

$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$



TACO: Score Update

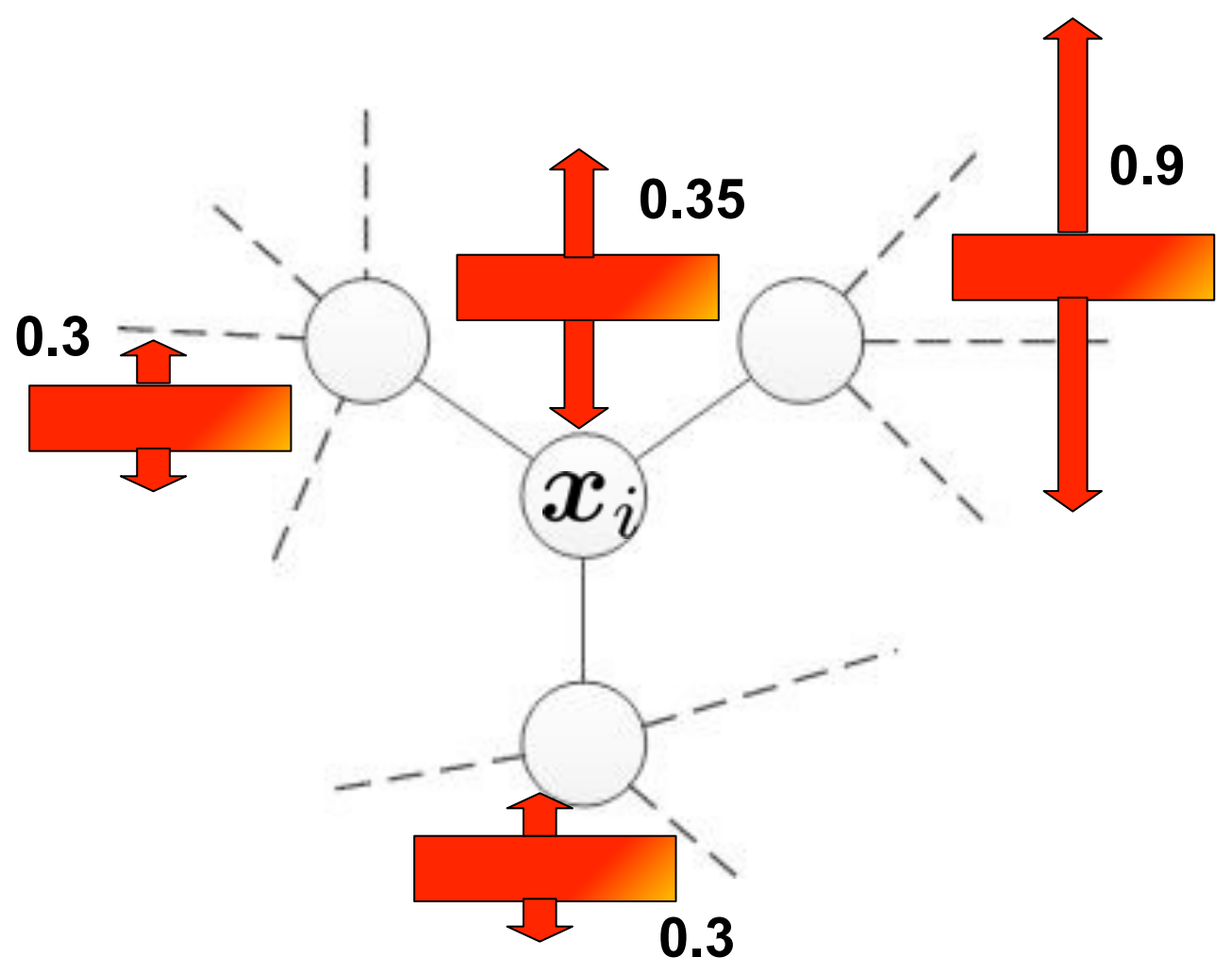
$$\mu_{i,r}^{(t)} \leftarrow \frac{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) \mu_{j,r}^{(t-1)} + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right) y_{i,r}}{\sum_{\substack{j \neq i \\ j=1}}^{n} w_{i,j} \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\sigma_{j,r}^{(t-1)}} \right) + \delta_l(i) \left(\frac{1}{\sigma_{i,r}^{(t-1)}} + \frac{1}{\gamma} \right)}$$



- Average Neighbors
- Consider Confidence

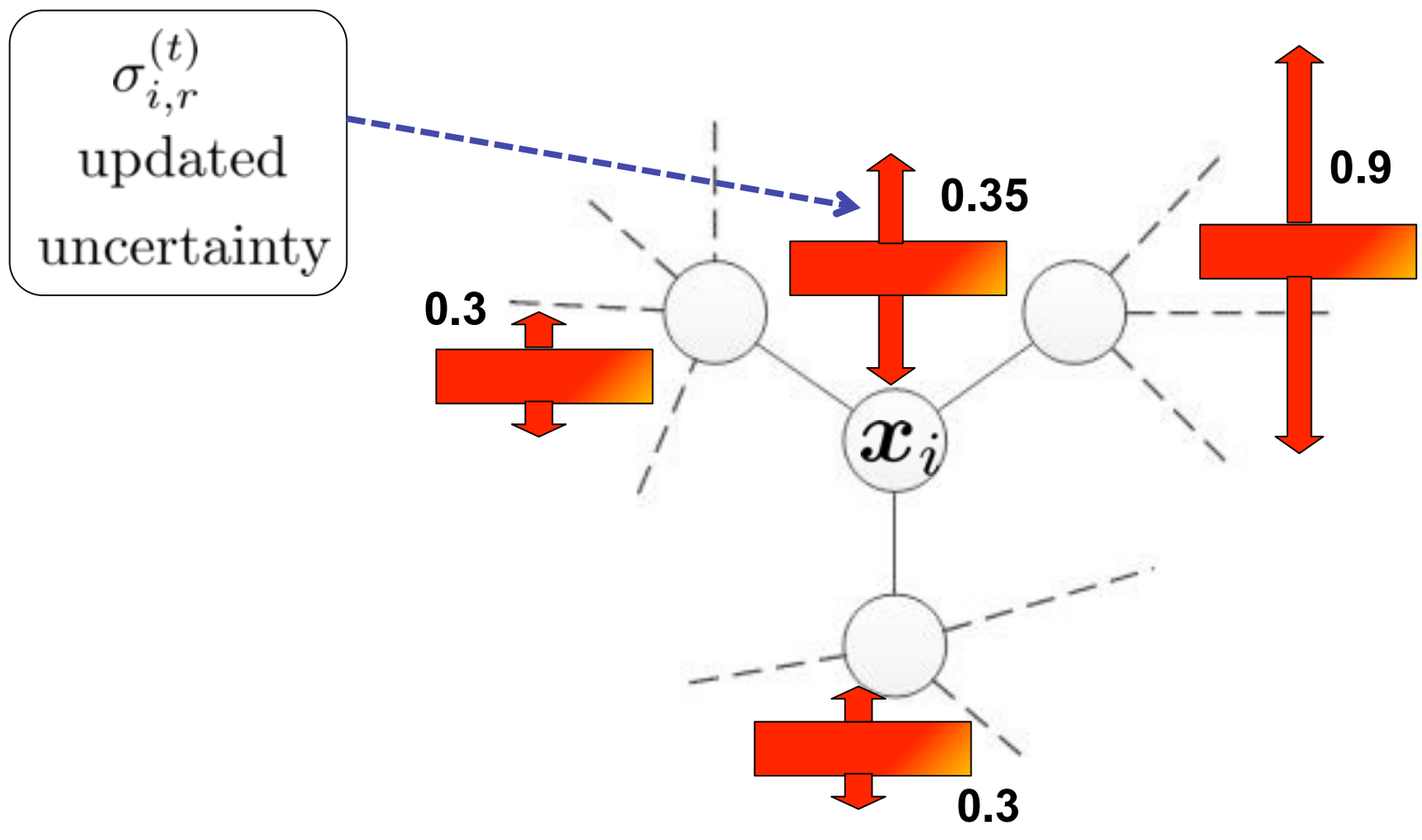
TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



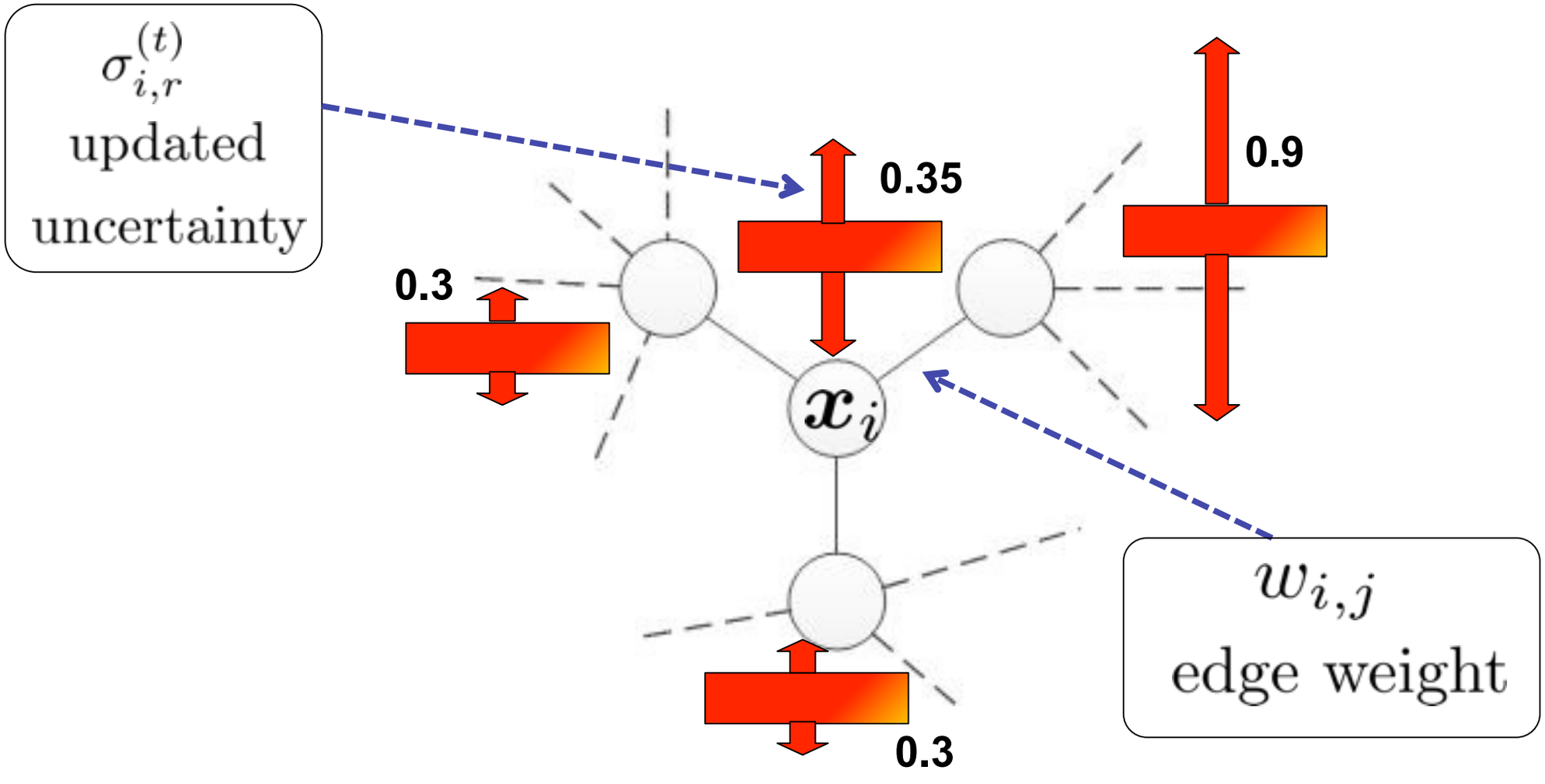
TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



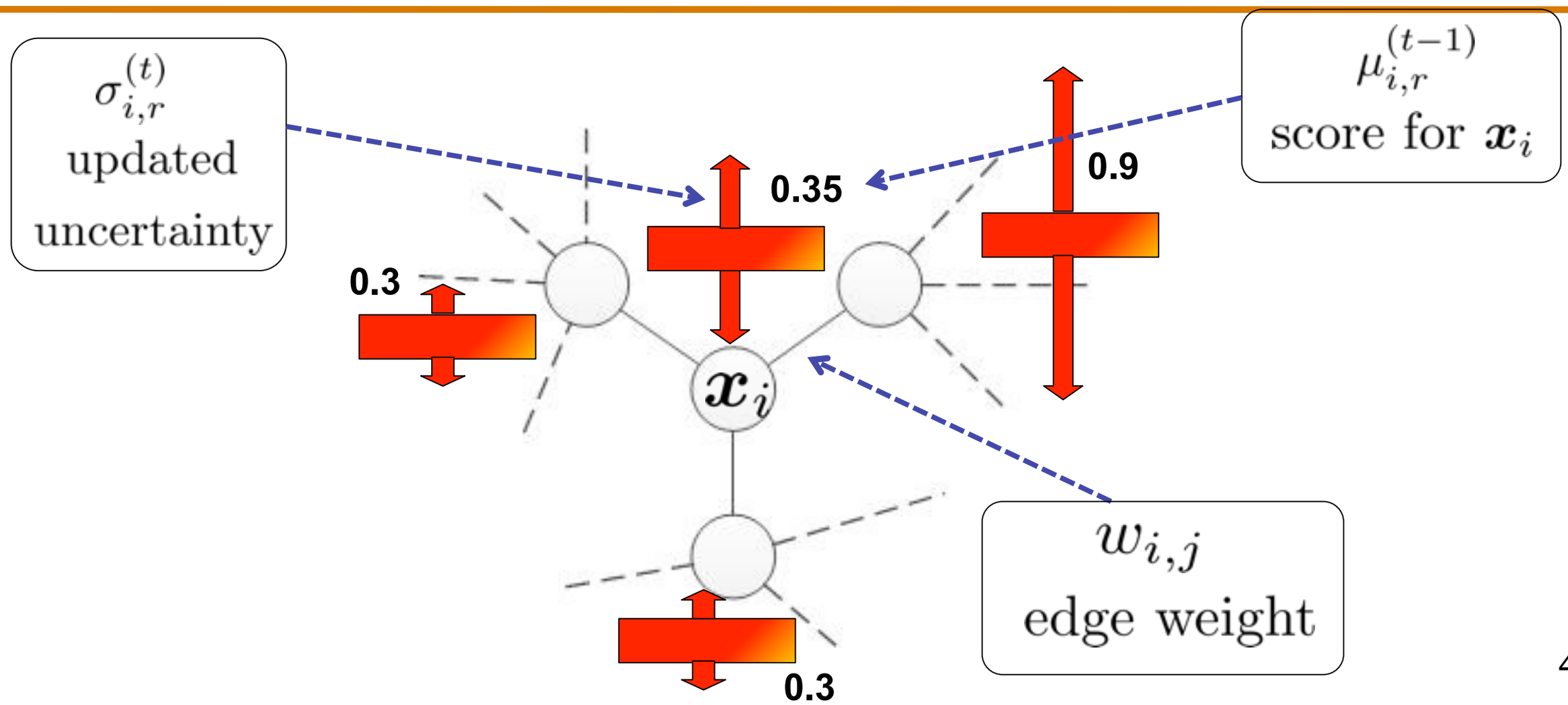
TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



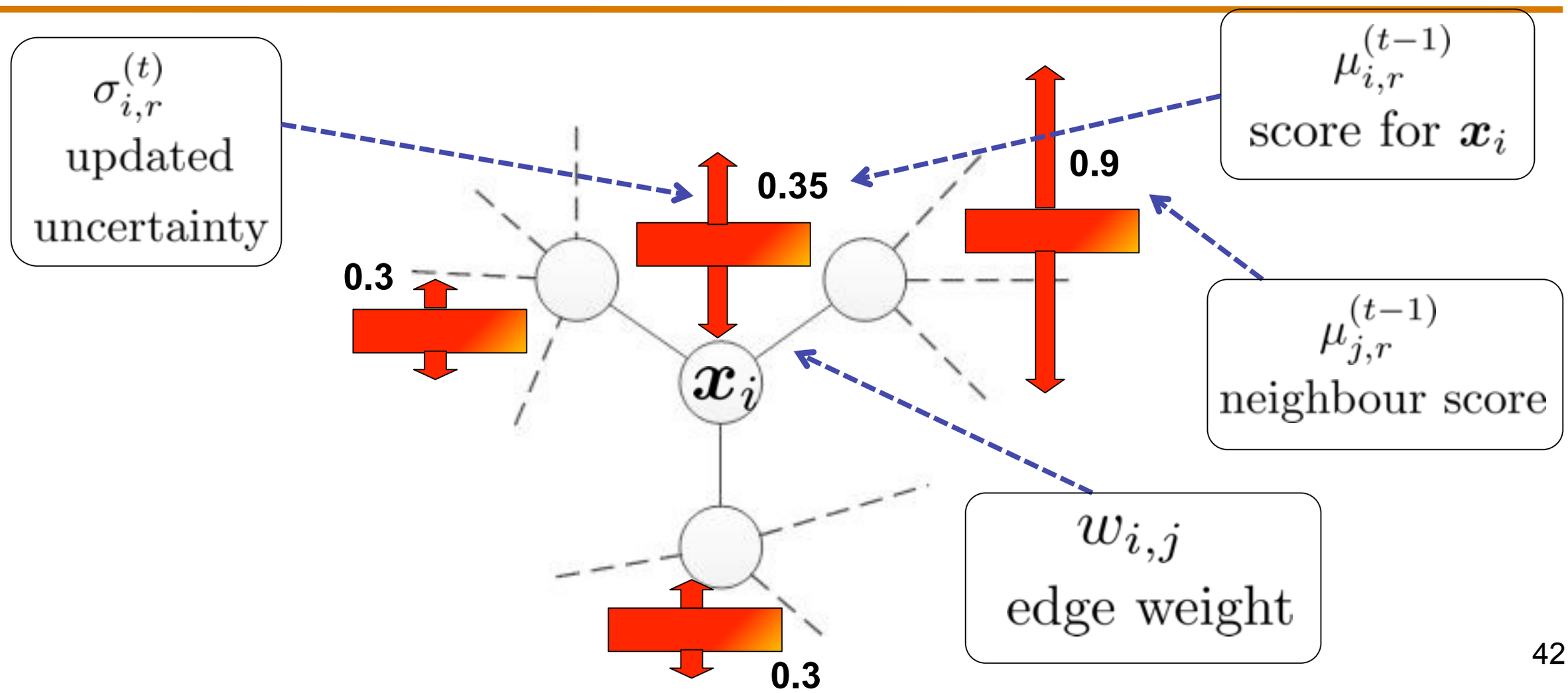
TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



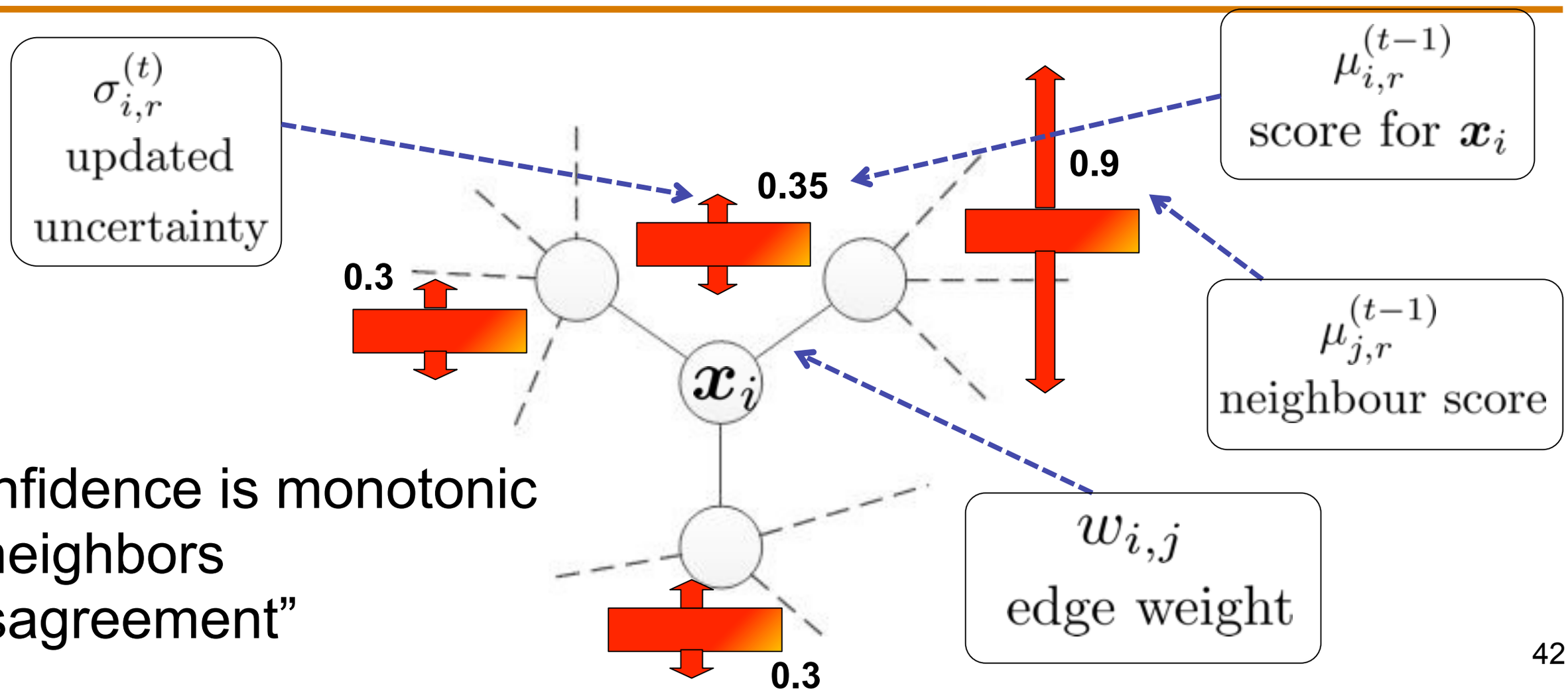
TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



TACO: Confidence Update

$$\sigma_{i,r}^{(t)} \leftarrow \frac{\beta}{2\alpha} + \frac{1}{2\alpha}\sqrt{\beta^2 + 2\alpha\left[\sum_{j=1}^{n} w_{i,j}\left(\mu_{i,r}^{(t-1)} - \mu_{j,r}^{(t-1)}\right)^2 + \delta_l(i)\left(\mu_{i,r}^{(t-1)} - y_{i,r}\right)^2\right]}$$



Confidence is monotonic in neighbors "disagreement"

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Manifold Regularization

[Belkin et al., JMLR 2006]

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} V(y_i, f(x_i)) + \beta \; f^T L f + \gamma ||f||_K^2$$

Manifold Regularization

[Belkin et al., JMLR 2006]

Training Data Loss

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \beta \; f^T L f + \gamma ||f||_K^2$$

Loss Function
(e.g., soft margin)

Manifold Regularization

[Belkin et al., JMLR 2006]

Training Data Loss

Smoothness Regularizer

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \boxed{\beta \; f^T L f} + \gamma ||f||_K^2$$

Loss Function (e.g., soft margin)

Laplacian of graph over labeled and unlabeled data

Manifold Regularization

[Belkin et al., JMLR 2006]

Training Data Loss

Smoothness Regularizer

Regularizer (e.g., L2)

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \boxed{\beta \; f^T L f} + \boxed{\gamma ||f||_K^2}$$

Loss Function (e.g., soft margin)

Laplacian of graph over labeled and unlabeled data

Manifold Regularization

[Belkin et al., JMLR 2006]

Training Data Loss

Smoothness Regularizer

Regularizer (e.g., L2)

$$f^* = \arg\min_f \frac{1}{l} \sum_{i=1}^{l} \boxed{V(y_i, f(x_i))} + \boxed{\beta \; f^T L f} + \boxed{\gamma ||f||_K^2}$$

Loss Function (e.g., soft margin)

Laplacian of graph over labeled and unlabeled data

Trains an <u>inductive</u> classifier which can generalize to unseen instances

Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Transduction with Confidence
  - Manifold Regularization
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} D_{KL}(r_i||p_i) + \mu \sum_{i,j} w_{ij} D_{KL}(p_i||p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t. } \sum_y p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

Divergence on seed nodes

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} w_{ij} D_{KL}(p_i||p_j) - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t.} \ \sum_y p_i(y) = 1, \ p_i(y) \geq 0, \ \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i||p_j)} - \nu \sum_{i=1}^{n} H(p_i)$$

$$\text{s.t.} \sum_y p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

Entropic Regularizer

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i||p_j)} - \boxed{\nu \sum_{i=1}^{n} H(p_i)}$$

$$\text{s.t.} \sum_y p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_y p_i(y) \log p_i(y)$$

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

Entropic Regularizer

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i||p_j)} - \boxed{\nu \sum_{i=1}^{n} H(p_i)}$$

$$\text{s.t.} \sum_y p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_y p_i(y) \log p_i(y)$$

Measure Propagation (MP)

[Subramanya and Bilmes, EMNLP 2008, NIPS 2009, JMLR 2011]

C_{KL}

Divergence on seed nodes

Smoothness (divergence across edge)

Entropic Regularizer

$$\arg\min_{\{p_i\}} \sum_{i=1}^{l} \boxed{D_{KL}(r_i||p_i)} + \mu \sum_{i,j} \boxed{w_{ij} D_{KL}(p_i||p_j)} - \boxed{\nu \sum_{i=1}^{n} H(p_i)}$$

$$\text{s.t.} \sum_y p_i(y) = 1, \; p_i(y) \geq 0, \; \forall y, i$$

Seed and estimated label distributions (normalized) on node *i*

Normalization Constraint

KL Divergence

$$D_{KL}(p_i||p_j) = \sum_y p_i(y) \log \frac{p_i(y)}{p_j(y)}$$

Entropy

$$H(p_i) = -\sum_y p_i(y) \log p_i(y)$$

C_{KL} is convex (with non-negative edge weights and hyper-parameters)

MP is related to Information Regularization [Corduneanu and Jaakkola, 2003]

Solving MP Objective

- For ease of optimization, reformulate MP objective:

CMP

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

Solving MP Objective

- For ease of optimization, reformulate MP objective:

CMP

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

Solving MP Objective

- For ease of optimization, reformulate MP objective:

CMP

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

$$w'_{ij} = w_{ij} + \alpha \times \delta(i, j)$$

Solving MP Objective

- For ease of optimization, reformulate MP objective:

CMP

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

$$w'_{ij} = w_{ij} + \alpha \times \delta(i, j)$$

Encourages agreement between p_i and q_i

$$\operatorname*{argmin}_{p \in \Delta^n} C_{KL}(p) = \lim_{\alpha \to \infty} \operatorname*{argmin}_{p,q \in \Delta^n} C_{MP}(p, q)$$

Solving MP Objective

- For ease of optimization, reformulate MP objective:

C_{MP}

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

$$w'_{ij} = w_{ij} + \alpha \times \delta(i, j)$$

Encourages agreement between p_i and q_i

$$\operatorname*{argmin}_{p \in \Delta^n} C_{KL}(p) = \lim_{\alpha \to \infty} \operatorname*{argmin}_{p,q \in \Delta^n} C_{MP}(p, q)$$

C_{MP} is also convex

(with non-negative edge weights and hyper-parameters)

Solving MP Objective

- For ease of optimization, reformulate MP objective:

C_{MP}

$$\arg\min_{\{p_i, q_i\}} \sum_{i=1}^{l} D_{KL}(r_i || q_i) + \mu \sum_{i,j} w'_{ij} D_{KL}(p_i || q_j) - \nu \sum_{i=1}^{n} H(p_i)$$

New probability measure, one for each vertex, similar to p_i

$$w'_{ij} = w_{ij} + \alpha \times \delta(i,j)$$

Encourages agreement between p_i and q_i

$$\operatorname*{argmin}_{p \in \Delta^n} C_{KL}(p) = \lim_{\alpha \to \infty} \operatorname*{argmin}_{p,q \in \Delta^n} C_{MP}(p,q)$$

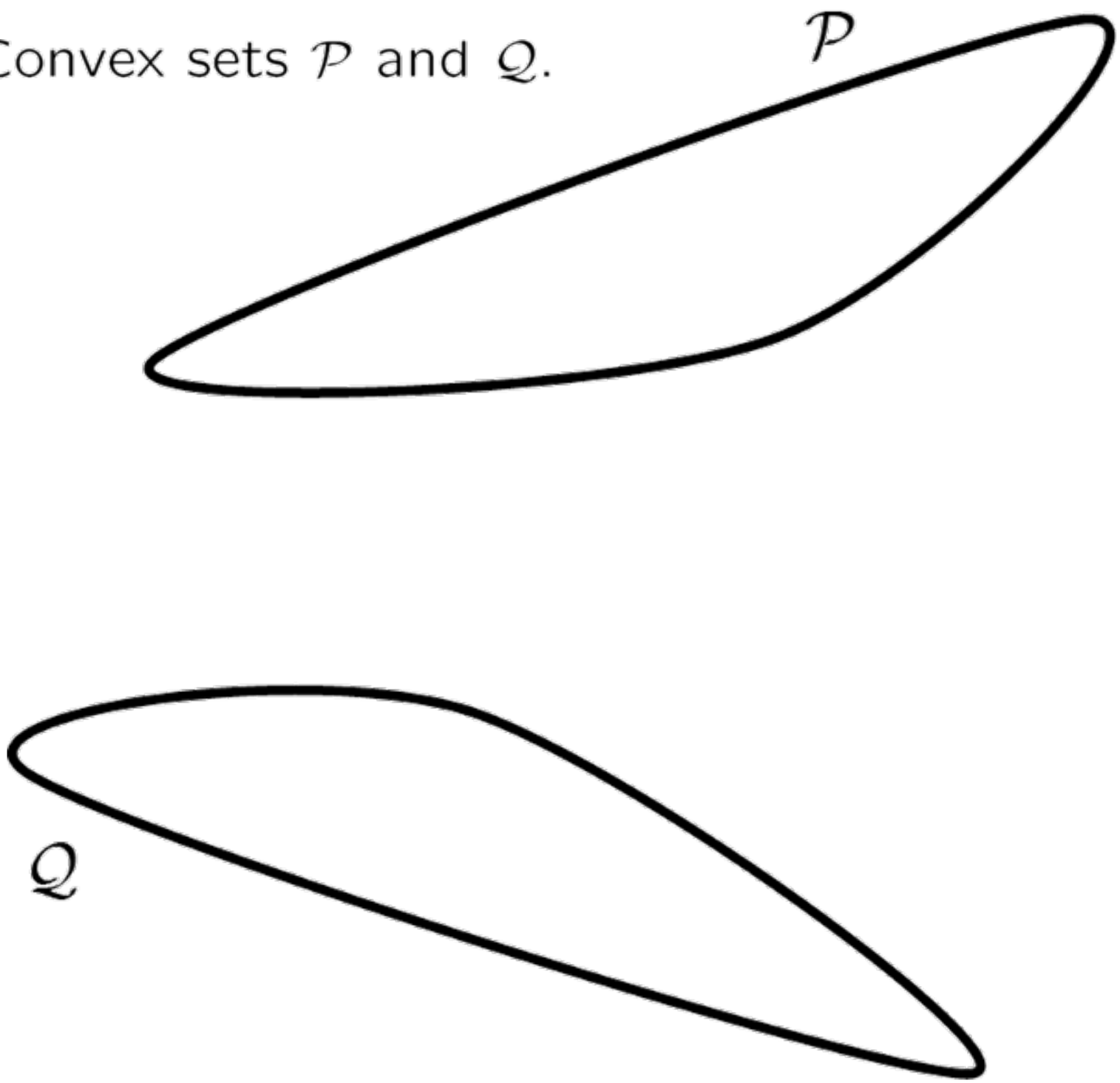
C_{MP} is also convex

(with non-negative edge weights and hyper-parameters)

C_{MP} can be solved using Alternating Minimization (AM)

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



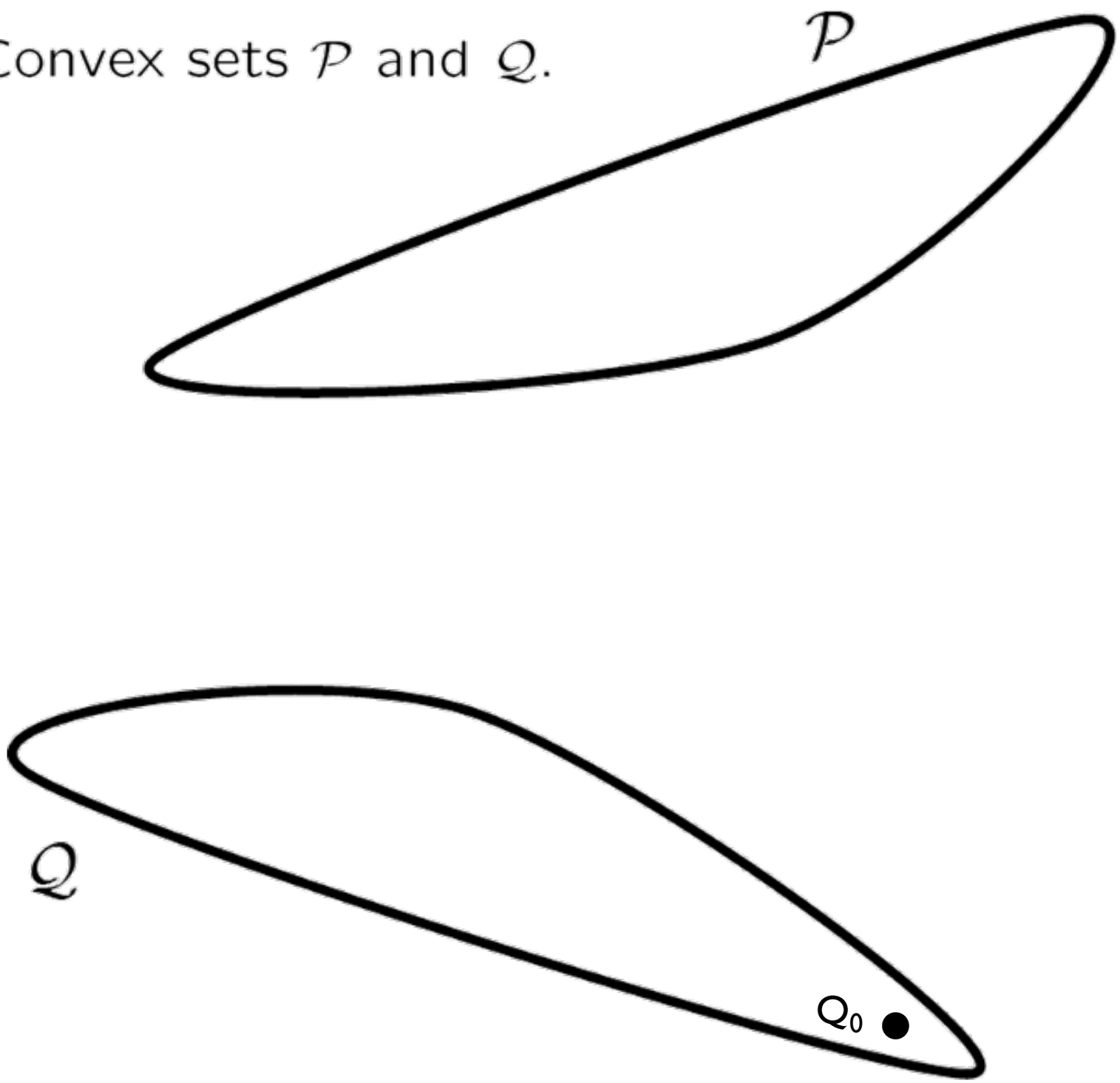
Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.

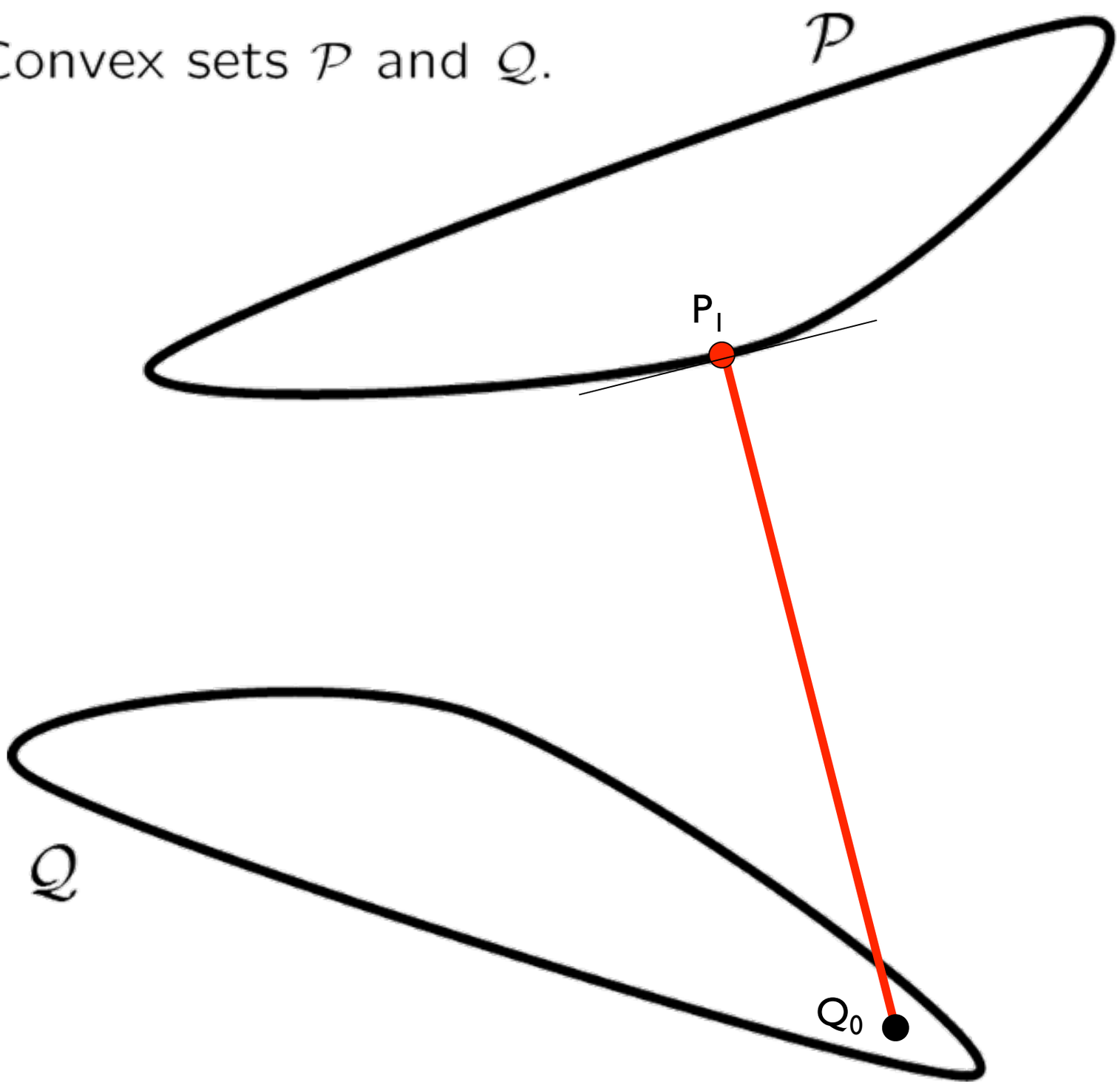
Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$



Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



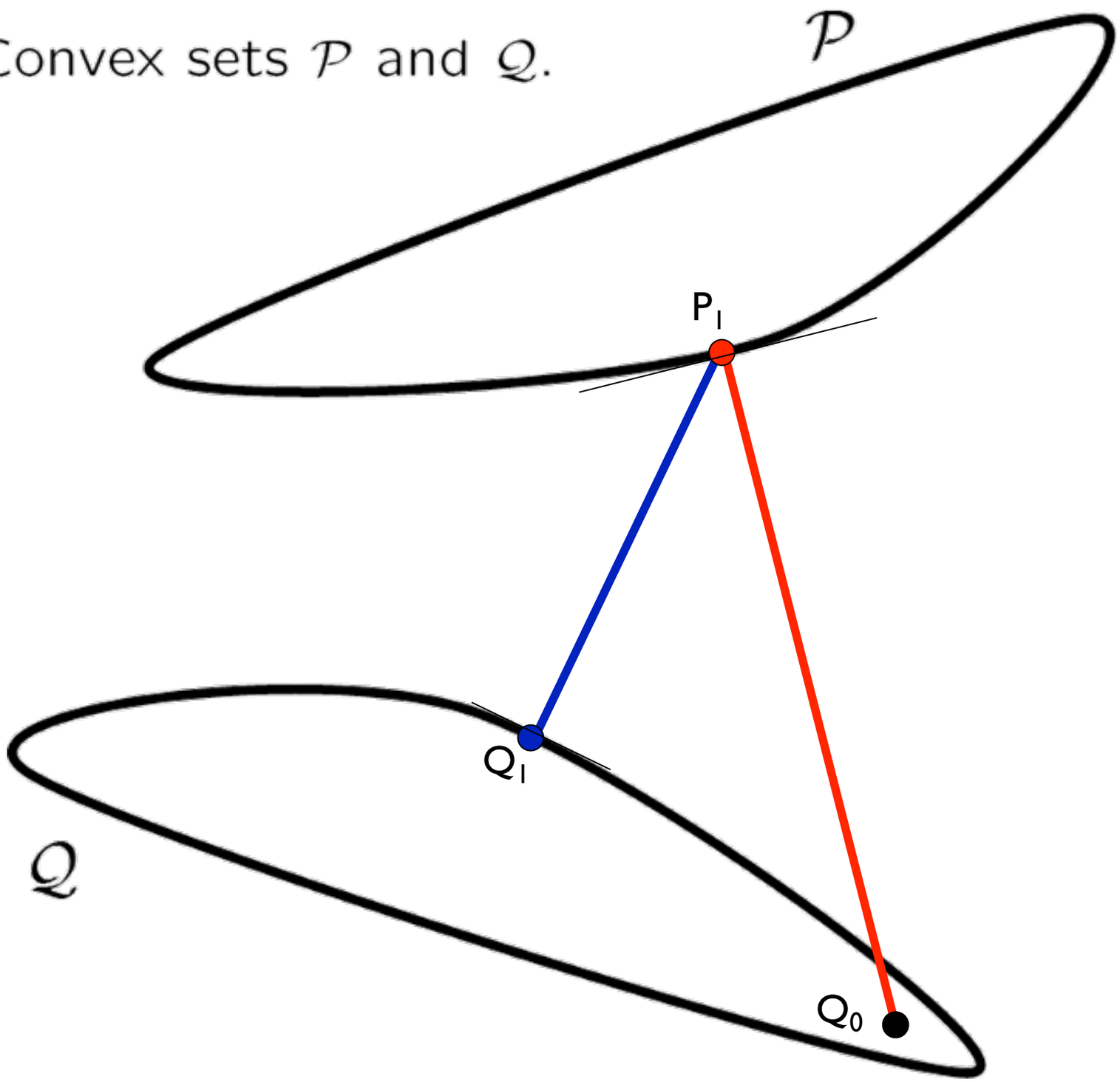
Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \operatorname*{argmin}_P d(P, Q_0)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

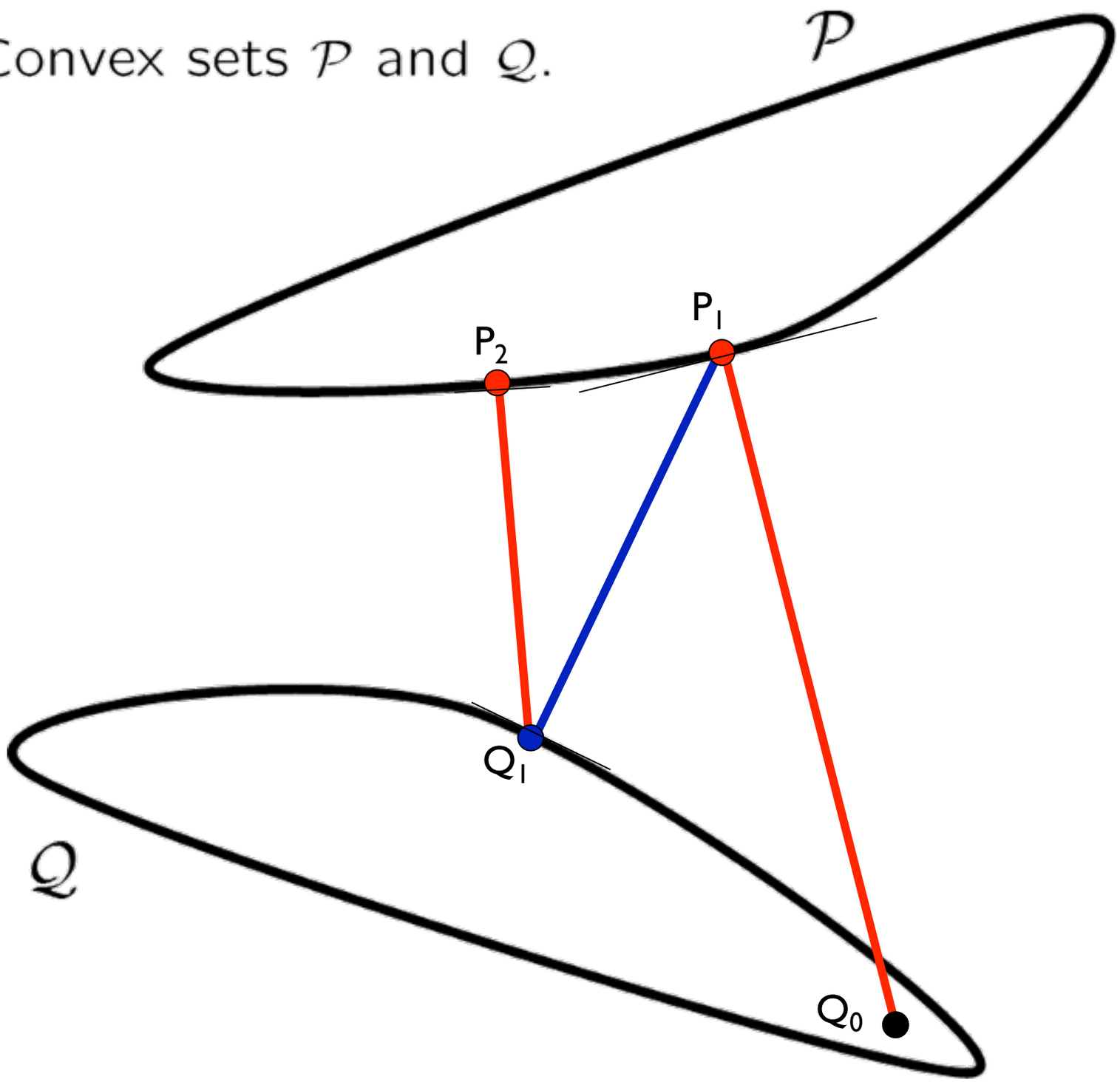
Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \operatorname*{argmin}_P d(P, Q_0)$$

$$Q_1 = \operatorname*{argmin}_Q d(P_1, Q)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

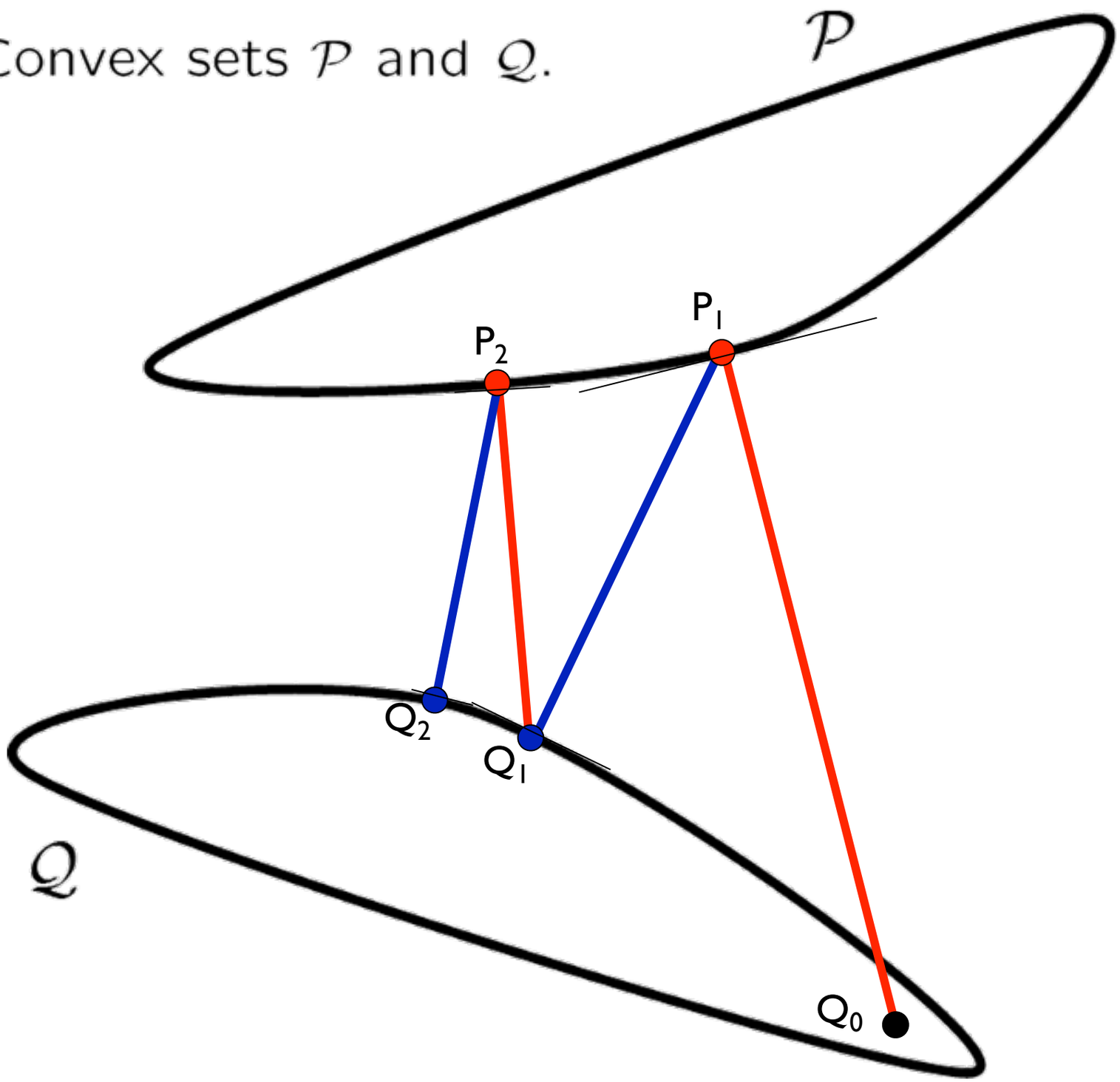
$$P_1 = \operatorname*{argmin}_P d(P, Q_0)$$

$$Q_1 = \operatorname*{argmin}_Q d(P_1, Q)$$

$$P_2 = \operatorname*{argmin}_P d(P, Q_1)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\, d(P, Q_0)$$

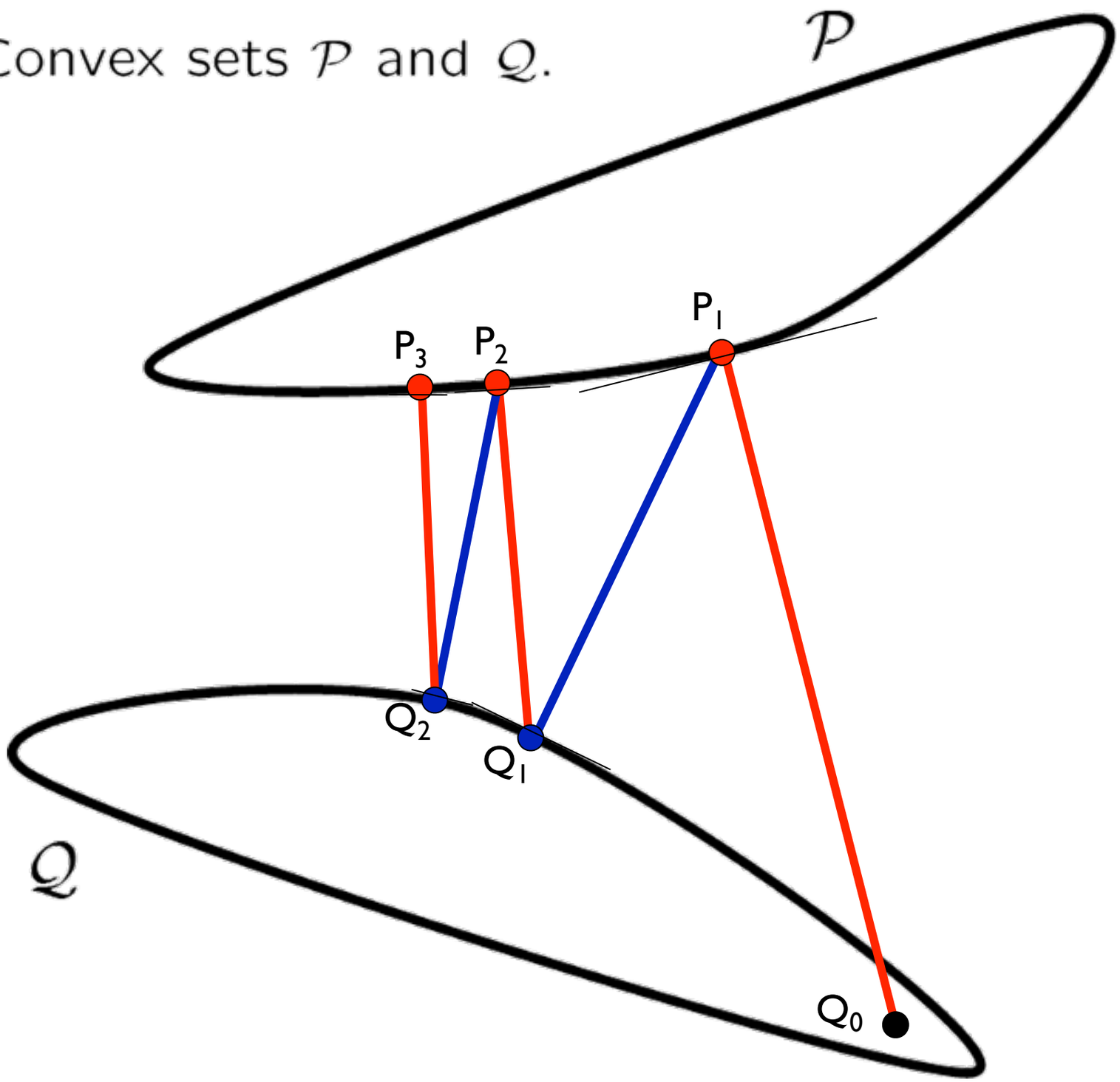
$$Q_1 = \underset{Q}{\operatorname{argmin}}\, d(P_1, Q)$$

$$P_2 = \underset{P}{\operatorname{argmin}}\, d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\, d(P_2, Q)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \operatorname*{argmin}_P d(P, Q_0)$$

$$Q_1 = \operatorname*{argmin}_Q d(P_1, Q)$$

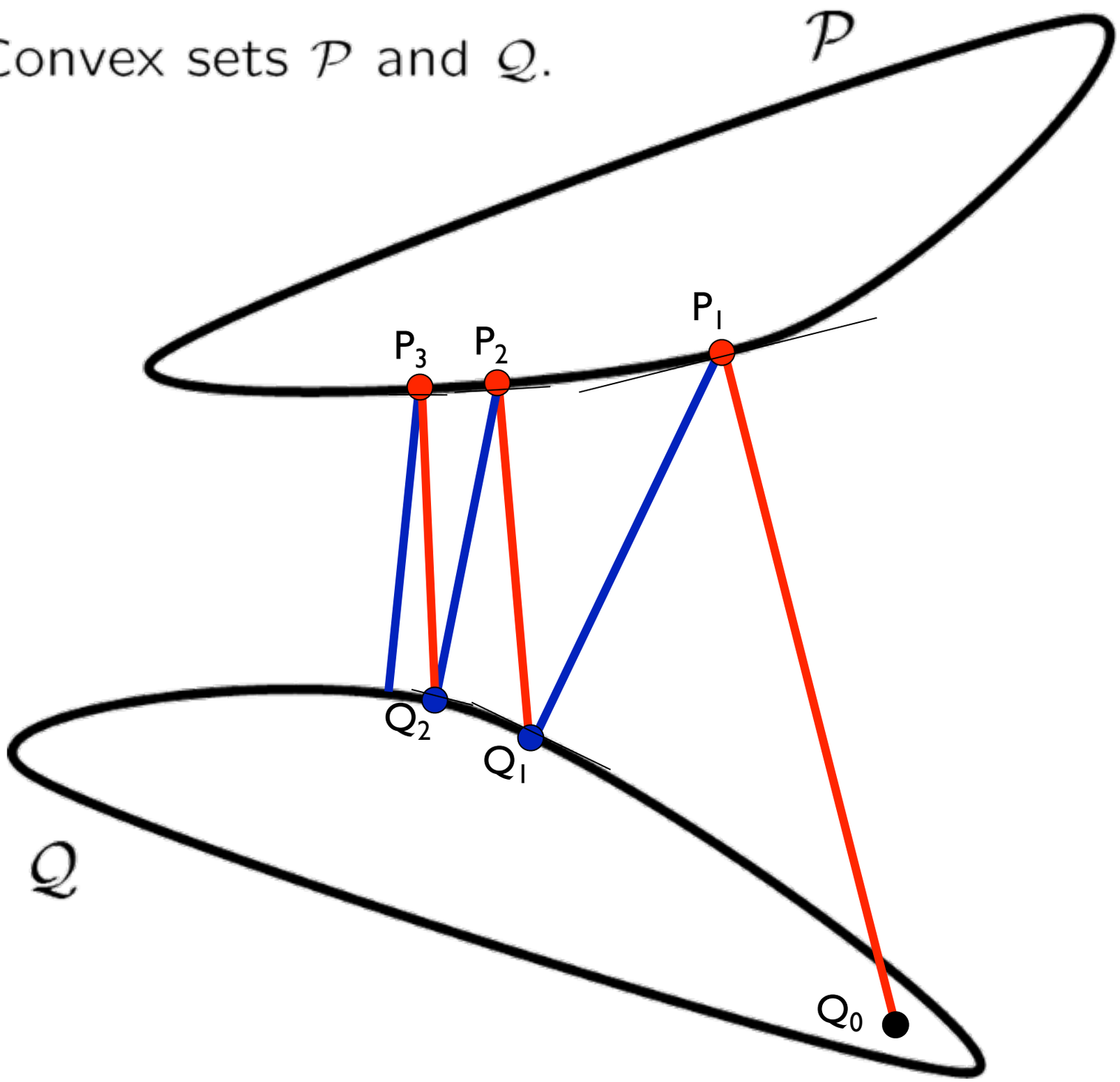
$$P_2 = \operatorname*{argmin}_P d(P, Q_1)$$

$$Q_2 = \operatorname*{argmin}_Q d(P_2, Q)$$

$$P_3 = \operatorname*{argmin}_P d(P, Q_2)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\, d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\, d(P_1, Q)$$

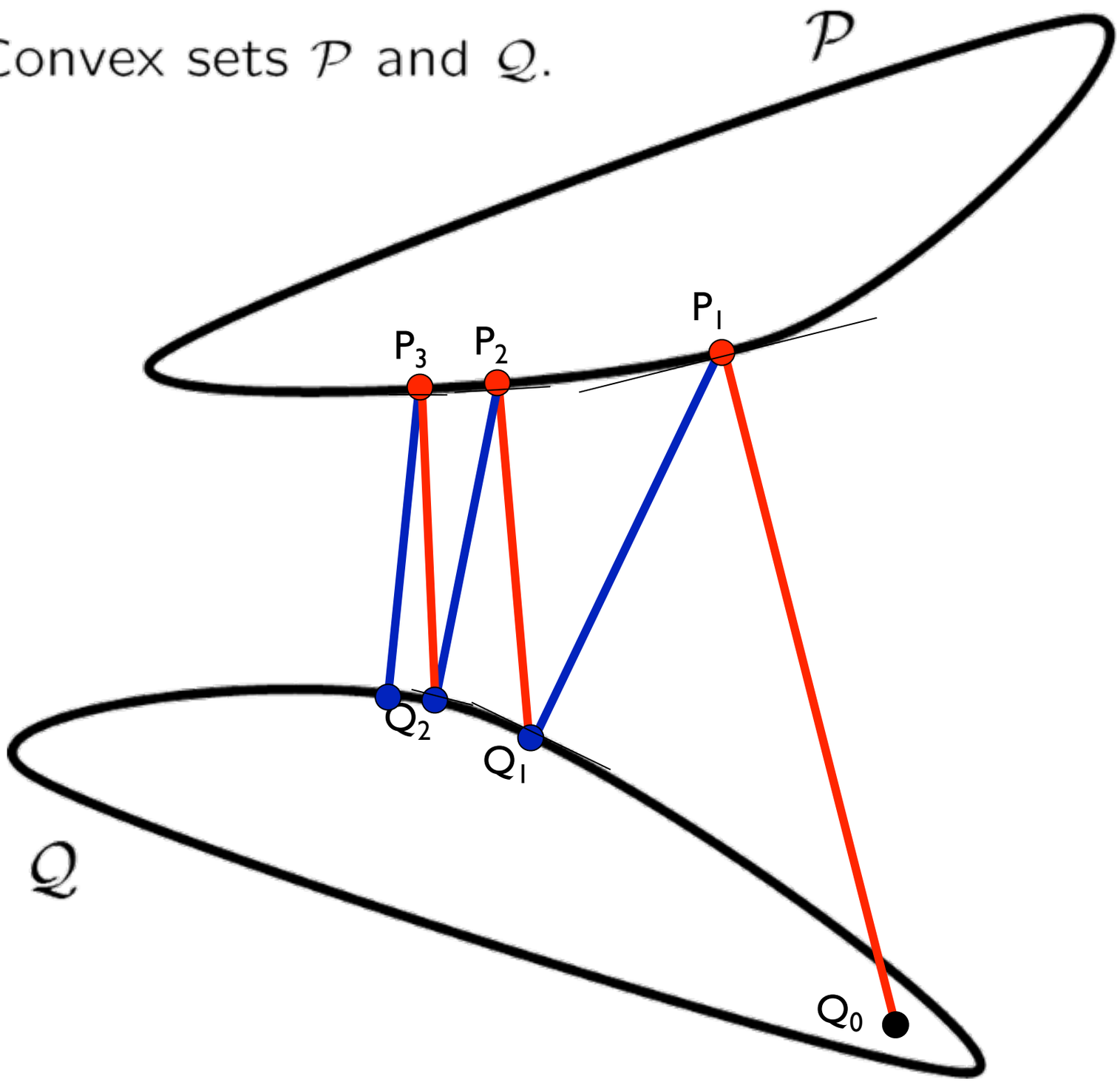
$$P_2 = \underset{P}{\operatorname{argmin}}\, d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\, d(P_2, Q)$$

$$P_3 = \underset{P}{\operatorname{argmin}}\, d(P, Q_2)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\, d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\, d(P_1, Q)$$

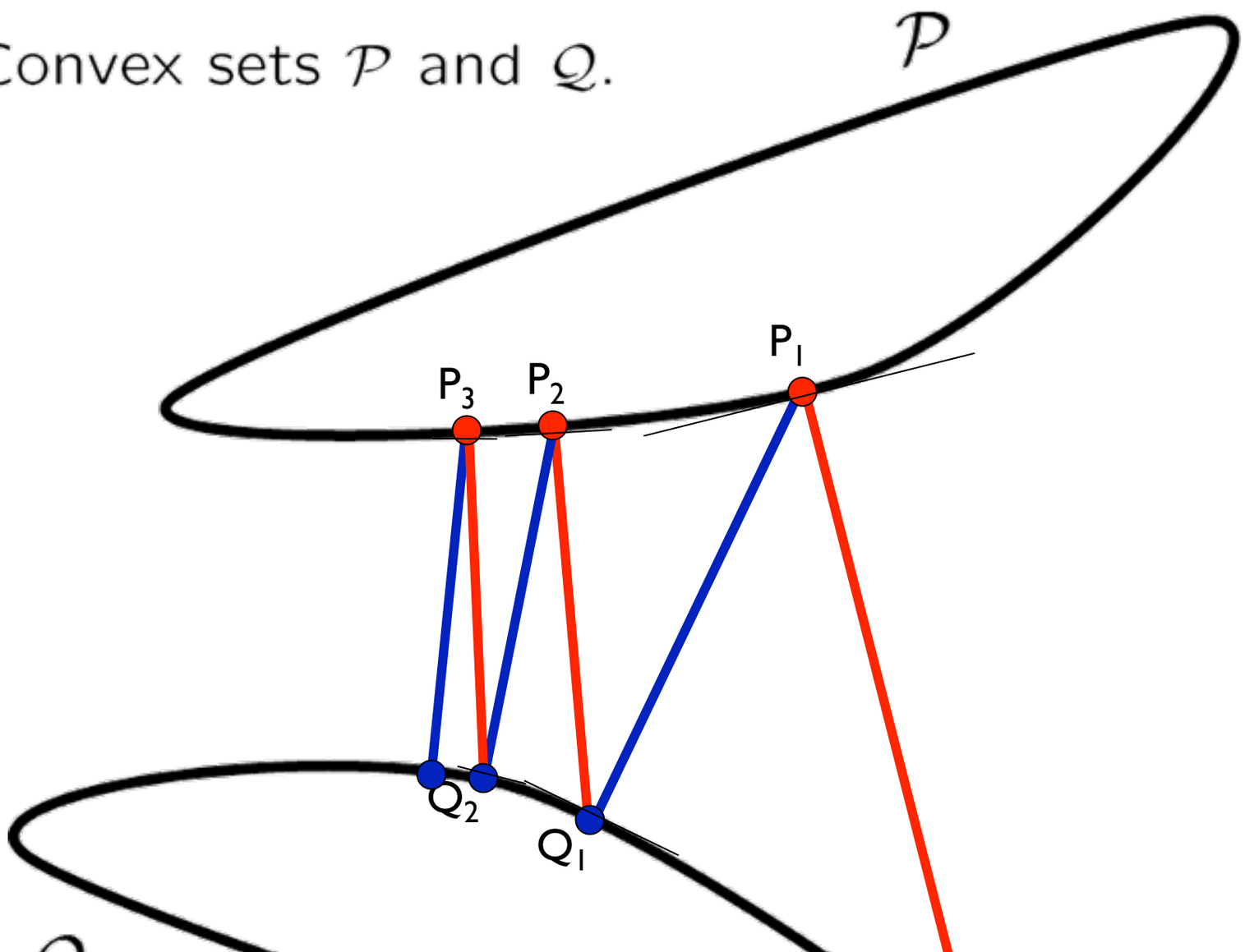
$$P_2 = \underset{P}{\operatorname{argmin}}\, d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\, d(P_2, Q)$$

$$P_3 = \underset{P}{\operatorname{argmin}}\, d(P, Q_2)$$

Alternating Minimization

Convex sets $\mathcal{P}$ and $\mathcal{Q}$.



Given distance $d(P, Q)$ with $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$.

Start with $Q_0 \in \mathcal{Q}$

$$P_1 = \underset{P}{\operatorname{argmin}}\, d(P, Q_0)$$

$$Q_1 = \underset{Q}{\operatorname{argmin}}\, d(P_1, Q)$$

$$P_2 = \underset{P}{\operatorname{argmin}}\, d(P, Q_1)$$

$$Q_2 = \underset{Q}{\operatorname{argmin}}\, d(P_2, Q)$$

C_{MP} satisfies the necessary conditions for AM to converge [Subramanya and Bilmes, JMLR 2011]

Why AM?

Why AM?

Criteria	MOM	AM
Iterative	YES	YES
Learning Rate	Armijo Rule	None
Number of Hyper-parameters	7	1 (α)
Test for Convergence	Requires Tuning	Automatic
Update Equations	Not Intuitive	Intuitive and easily Parallelized

Table 1: There are two ways to solving the proposed objective, namely, the popular numerical optimization tool method of multipliers (MOM), and the proposed approach based on alternating minimization (AM). This table compares the two approaches on various fronts.

Why AM?

Criteria	MOM	AM
Iterative	YES	YES
Learning Rate	Armijo Rule	None
Number of Hyper-parameters	7	1 (α)
Test for Convergence	Requires Tuning	Automatic
Update Equations	Not Intuitive	Intuitive and easily Parallelized

Table 1: There are two ways to solving the proposed objective, namely, the popular numerical optimization tool method of multipliers (MOM), and the proposed approach based on alternating minimization (AM). This table compares the two approaches on various fronts.

$$p_i^{(n)}(y) = \frac{\exp\{\frac{\mu}{\gamma_i}\sum_j w'_{ij}\log q_j^{(n-1)}(y)\}}{\sum_y \exp\{\frac{\mu}{\gamma_i}\sum_j w'_{ij}\log q_j^{(n-1)}(y)\}}$$

$$q_i^{(n)}(y) = \frac{r_i(y)\delta(i \leq l) + \mu\sum_j w'_{ji} p_j^{(n)}(y)}{\delta(i \leq l) + \mu\sum_j w'_{ji}}$$

$$\text{where } \gamma_i = \nu + \mu\sum_j w'_{ij}$$

Performance of SSL Algorithms

	COIL						OPT					
l	10	20	50	80	100	150	10	20	50	80	100	150
k-NN	34.5	53.9	66.9	77.9	79.2	83.5	79.6	83.9	85.5	90.5	92.0	93.8
SGT	40.1	61.2	78.0	88.5	89.0	89.9	90.4	90.6	91.4	94.7	**97.4**	**97.4**
LapRLS	**49.2**	61.4	78.4	80.1	84.5	87.8	89.7	**91.2**	92.3	96.1	**97.6**	**97.3**
SQ-Loss-I	**48.9**	63.0	**81.0**	87.5	89.0	90.9	**92.2**	90.2	**95.9**	**97.2**	**97.3**	**97.7**
MP	47.7	**65.7**	78.5	**89.6**	**90.2**	**91.1**	90.6	**90.8**	94.7	**96.6**	**97.0**	**97.1**

Comparison of accuracies for different number of labeled samples across COIL (6 classes) and OPT (10 classes) datasets

Performance of SSL Algorithms

	COIL						OPT					
l	10	20	50	80	100	150	10	20	50	80	100	150
k-NN	34.5	53.9	66.9	77.9	79.2	83.5	79.6	83.9	85.5	90.5	92.0	93.8
SGT	40.1	61.2	78.0	88.5	89.0	89.9	90.4	90.6	91.4	94.7	**97.4**	**97.4**
LapRLS	**49.2**	61.4	78.4	80.1	84.5	87.8	89.7	**91.2**	92.3	96.1	**97.6**	**97.3**
SQ-Loss-I	**48.9**	63.0	**81.0**	87.5	89.0	90.9	**92.2**	90.2	**95.9**	**97.2**	**97.3**	**97.7**
MP	47.7	**65.7**	78.5	**89.6**	**90.2**	**91.1**	90.6	**90.8**	94.7	**96.6**	**97.0**	**97.1**

Comparison of accuracies for different number of labeled samples across COIL (6 classes) and OPT (10 classes) datasets

Performance of SSL Algorithms

	COIL						OPT					
l	10	20	50	80	100	150	10	20	50	80	100	150
k-NN	34.5	53.9	66.9	77.9	79.2	83.5	79.6	83.9	85.5	90.5	92.0	93.8
SGT	40.1	61.2	78.0	88.5	89.0	89.9	90.4	90.6	91.4	94.7	**97.4**	**97.4**
LapRLS	**49.2**	61.4	78.4	80.1	84.5	87.8	89.7	**91.2**	92.3	96.1	**97.6**	**97.3**
SQ-Loss-I	**48.9**	63.0	**81.0**	87.5	89.0	90.9	**92.2**	90.2	**95.9**	**97.2**	**97.3**	**97.7**
MP	47.7	**65.7**	78.5	**89.6**	**90.2**	**91.1**	90.6	**90.8**	94.7	**96.6**	**97.0**	**97.1**

Comparison of accuracies for different number of labeled samples across COIL (6 classes) and OPT (10 classes) datasets

Graph SSL can be effective when the data satisfies manifold assumption. More results and discussion in Chapter 21 of the SSL Book (Chapelle et al.)

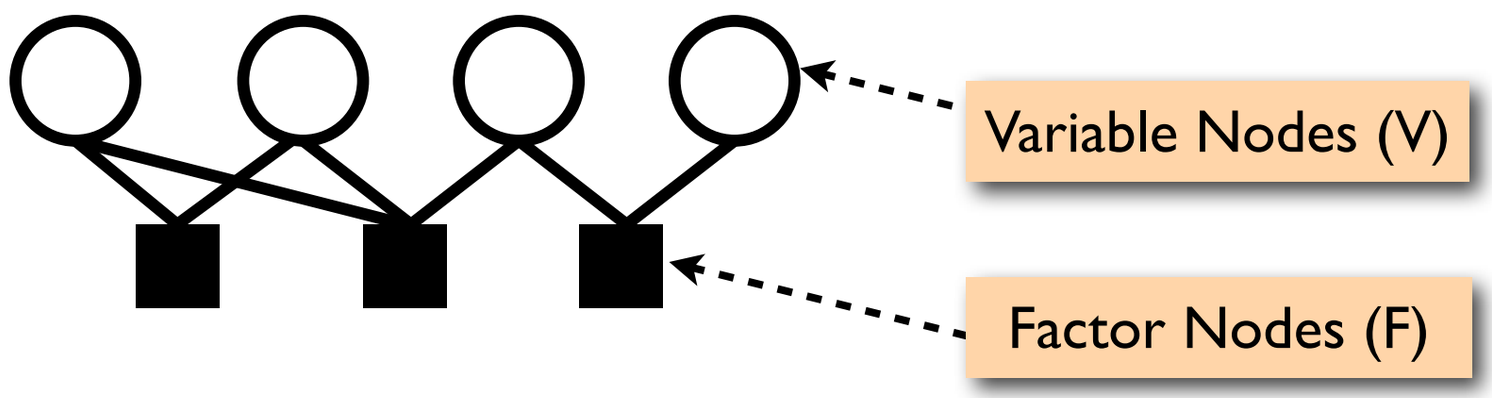
Outline

- Motivation
- Graph Construction
- Inference Methods
  - Label Propagation
  - Modified Adsorption
  - Manifold Regularization
  - Spectral Graph Transduction
  - Measure Propagation
  - Sparse Label Propagation
- Scalability
- Applications
- Conclusion & Future Work

Background: Factor Graphs
[Kschischang et al., 2001]

Factor Graph

- bipartite graph
- variable nodes (e.g., label distribution on a node)
- factor nodes: fitness function over variable assignment



Distribution over all variables' values

$$\log P\left(\{v\}_{v\in V}\right) \quad = \quad -\log Z + \sum_{f\in F} \log \alpha_f \left(\{v\}_{(v,f)\in E}\right)$$

variables connected to factor f

Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]

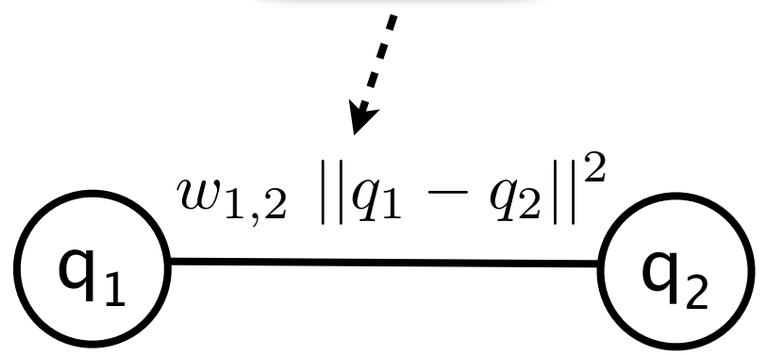
3-term Graph SSL Objective (common to many algorithms)

min Seed Matching Loss (if any) + Edge Smoothness Loss + Regularization Loss

Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]

3-term Graph SSL Objective (common to many algorithms)

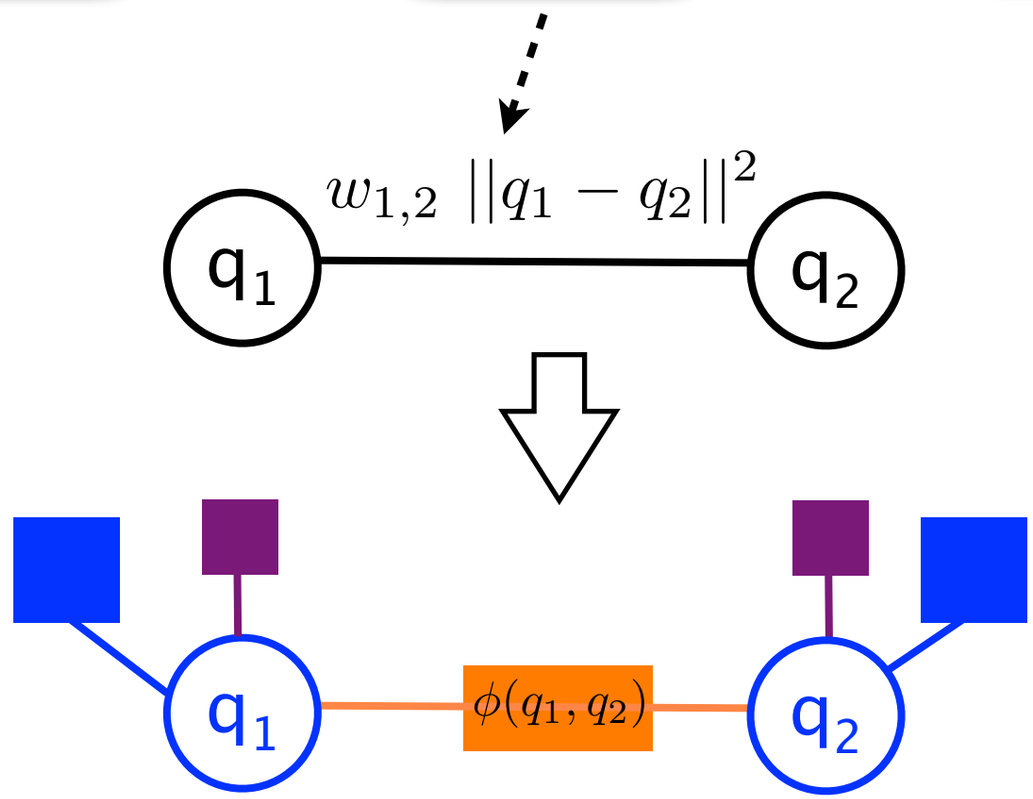
min Seed Matching Loss (if any) + Edge Smoothness Loss + Regularization Loss



Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]

3-term Graph SSL Objective (common to many algorithms)

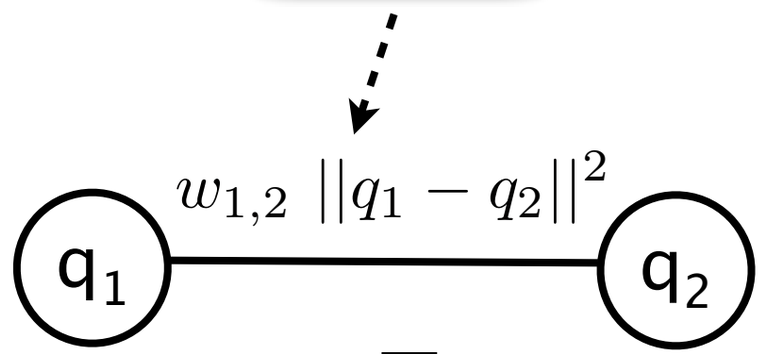
min Seed Matching Loss (if any) + Edge Smoothness Loss + Regularization Loss

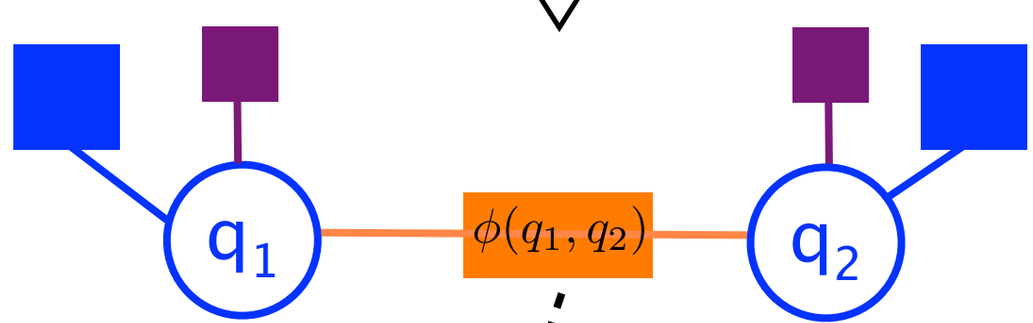


Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]

3-term Graph SSL Objective (common to many algorithms)

min Seed Matching Loss (if any) + Edge Smoothness Loss + Regularization Loss

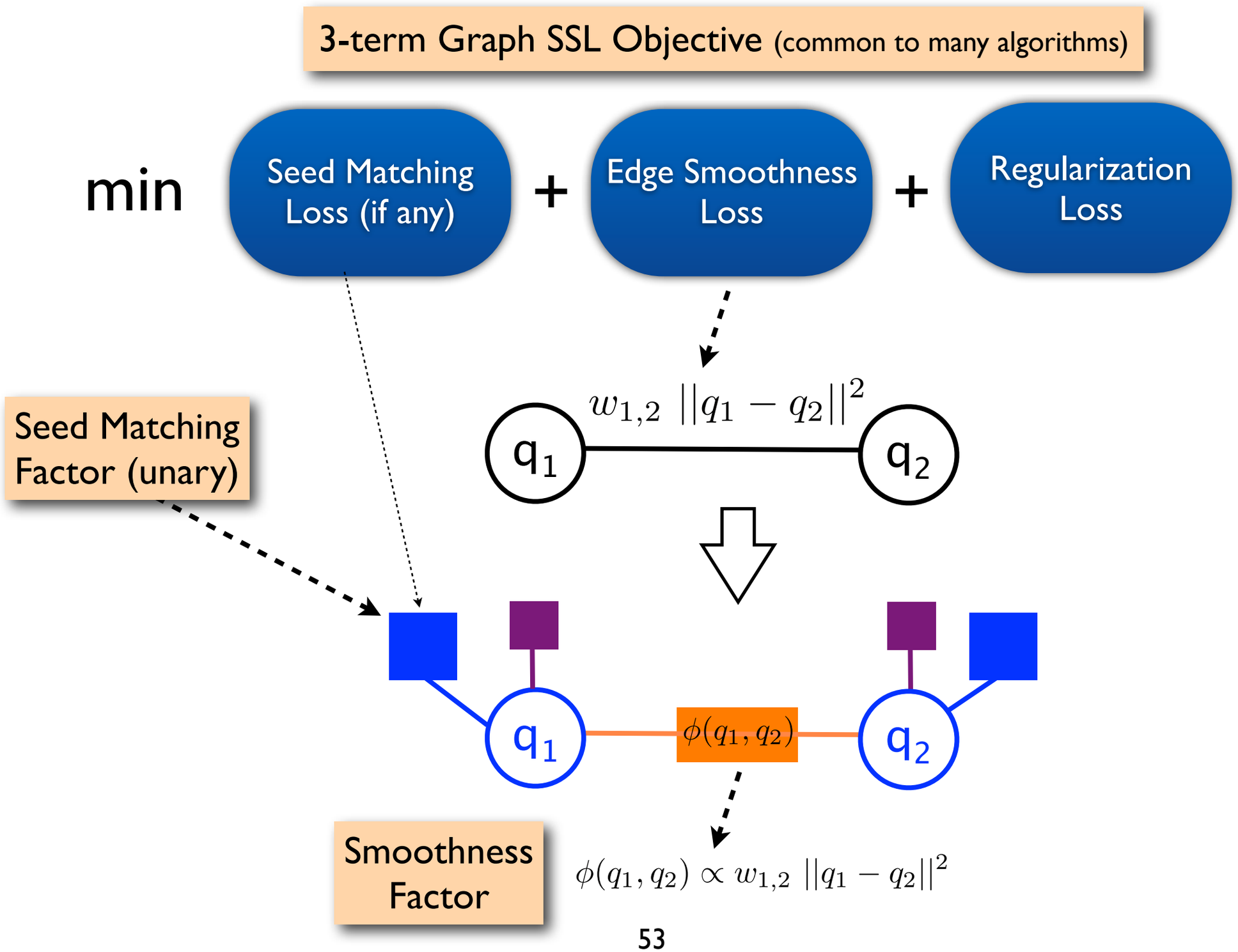




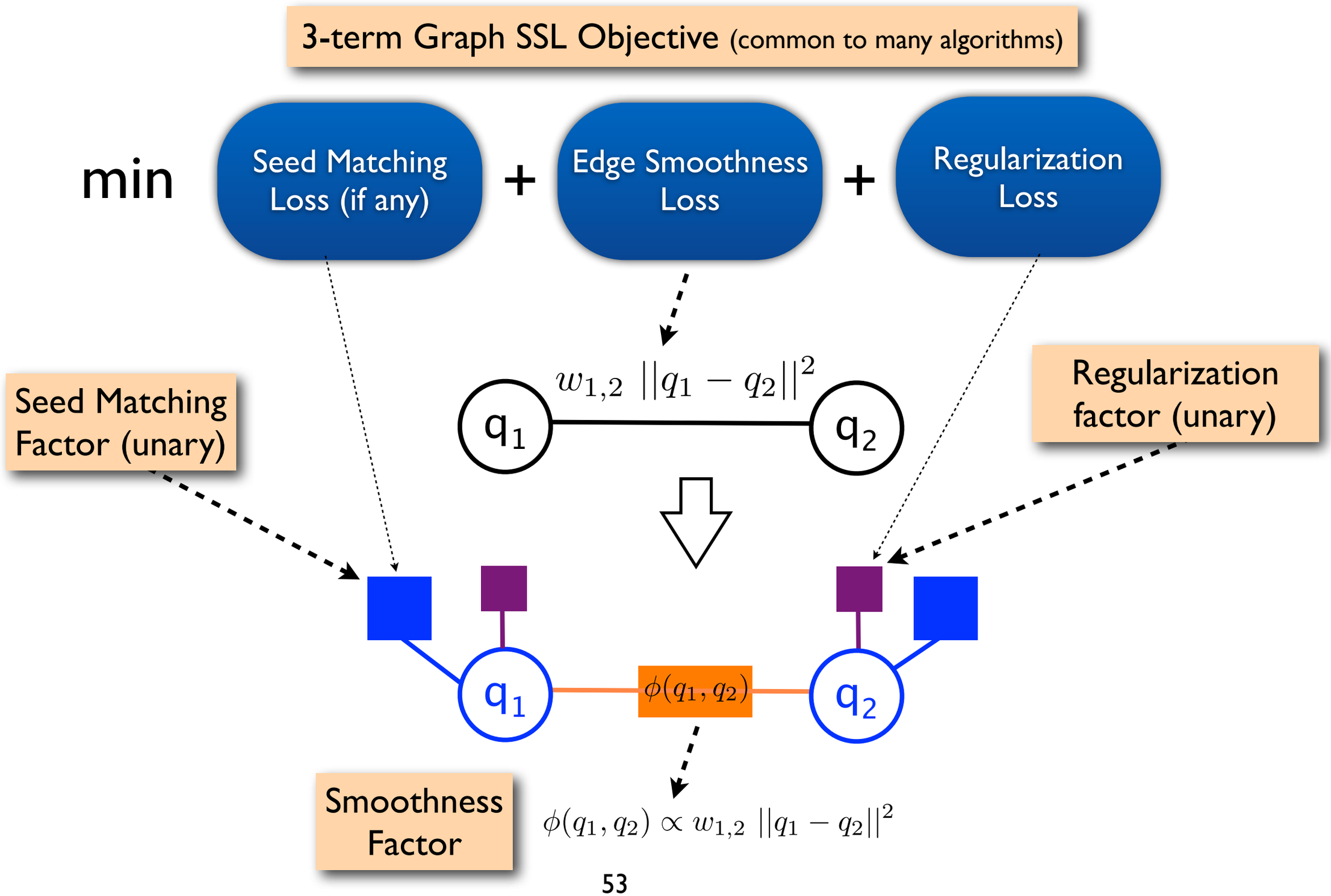
Smoothness Factor

$$\phi(q_1, q_2) \propto w_{1,2}\ ||q_1 - q_2||^2$$

Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]

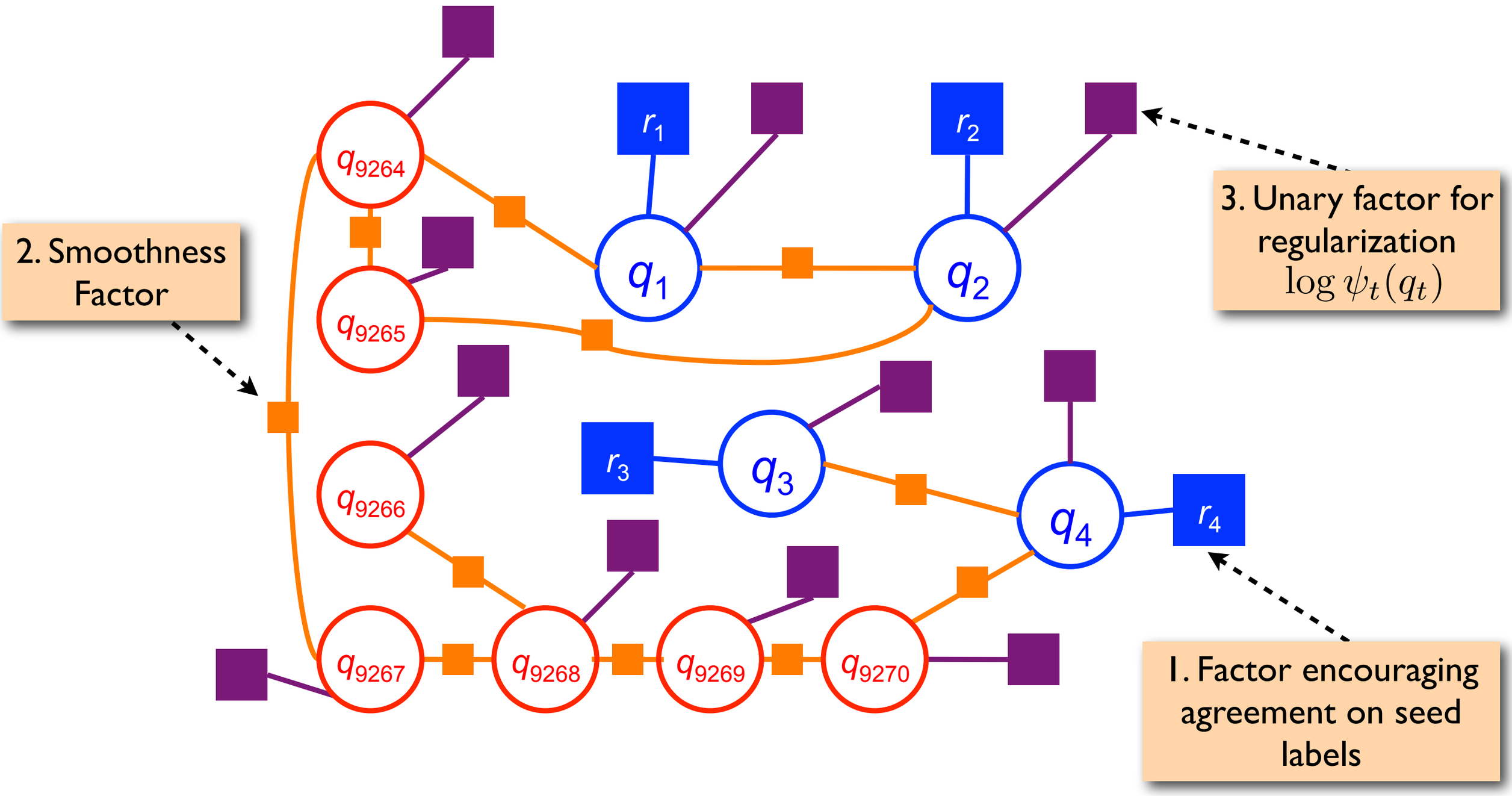


Factor Graph Interpretation of Graph SSL [Zhu et al., ICML 2003] [Das and Smith, NAACL 2012]



Factor Graph Interpretation

[Zhu et al., ICML 2003][Das and Smith, NAACL 2012]



Label Propagation with Sparsity

Label Propagation with Sparsity

Enforce through sparsity inducing unary factor

Label Propagation with Sparsity

Enforce through sparsity inducing unary factor

Lasso (Tibshirani, 1996) $\log \psi_t(q_t) = -\lambda \left\| q_t \right\|_1$

Elitist Lasso (Kowalski and Torrésani, 2009)

$$\log \psi_t(q_t) = -\lambda \left(\left\| q_t \right\|_1 \right)^2$$

Label Propagation with Sparsity

Enforce through sparsity inducing unary factor

Lasso (Tibshirani, 1996) $\log \psi_t(q_t) = -\lambda \left\| q_t \right\|_1$

Elitist Lasso (Kowalski and Torrésani, 2009)

$$\log \psi_t(q_t) = -\lambda \left(\left\| q_t \right\|_1 \right)^2$$

For more details, see [Das and Smith, NAACL 2012]

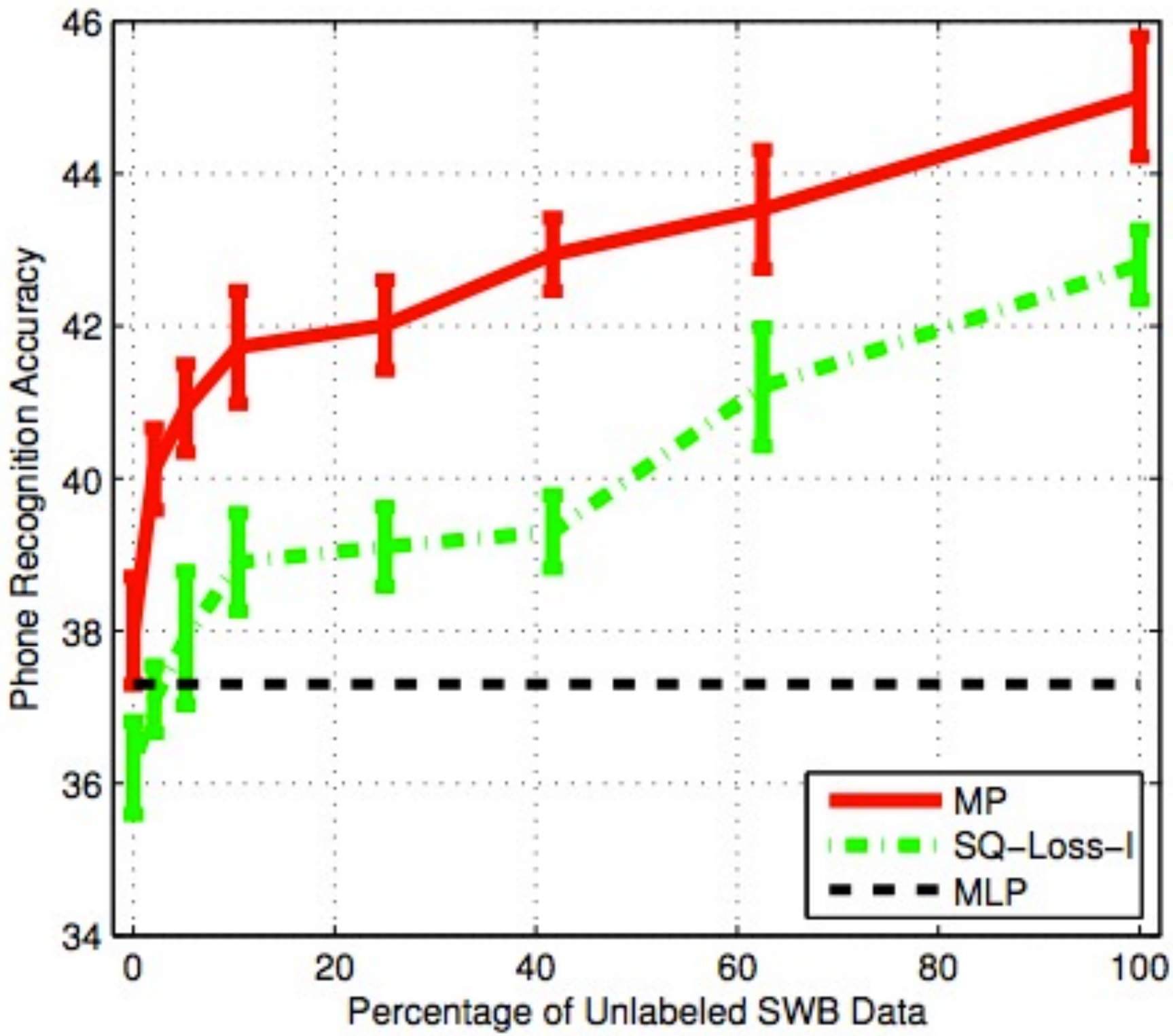
Other Graph-SSL Methods

- Spectral Graph Transduction [Joachims, ICML 2003]
- SSL on Directed Graphs
  - [Zhou et al, NIPS 2005], [Zhou et al., ICML 2005]
- Learning with dissimilarity edges
  - [Goldberg et al., AISTATS 2007]
- Learning to order: GraphOrder [Talukdar et al., CIKM 2012]
- Graph Transduction using Alternating Minimization
  - [Wang et al., ICML 2008]
- Graph as regularizer for Multi-Layered Perceptron
  - [Karlen et al., ICML 2008], [Malkin et al., Interspeech 2009]

Outline

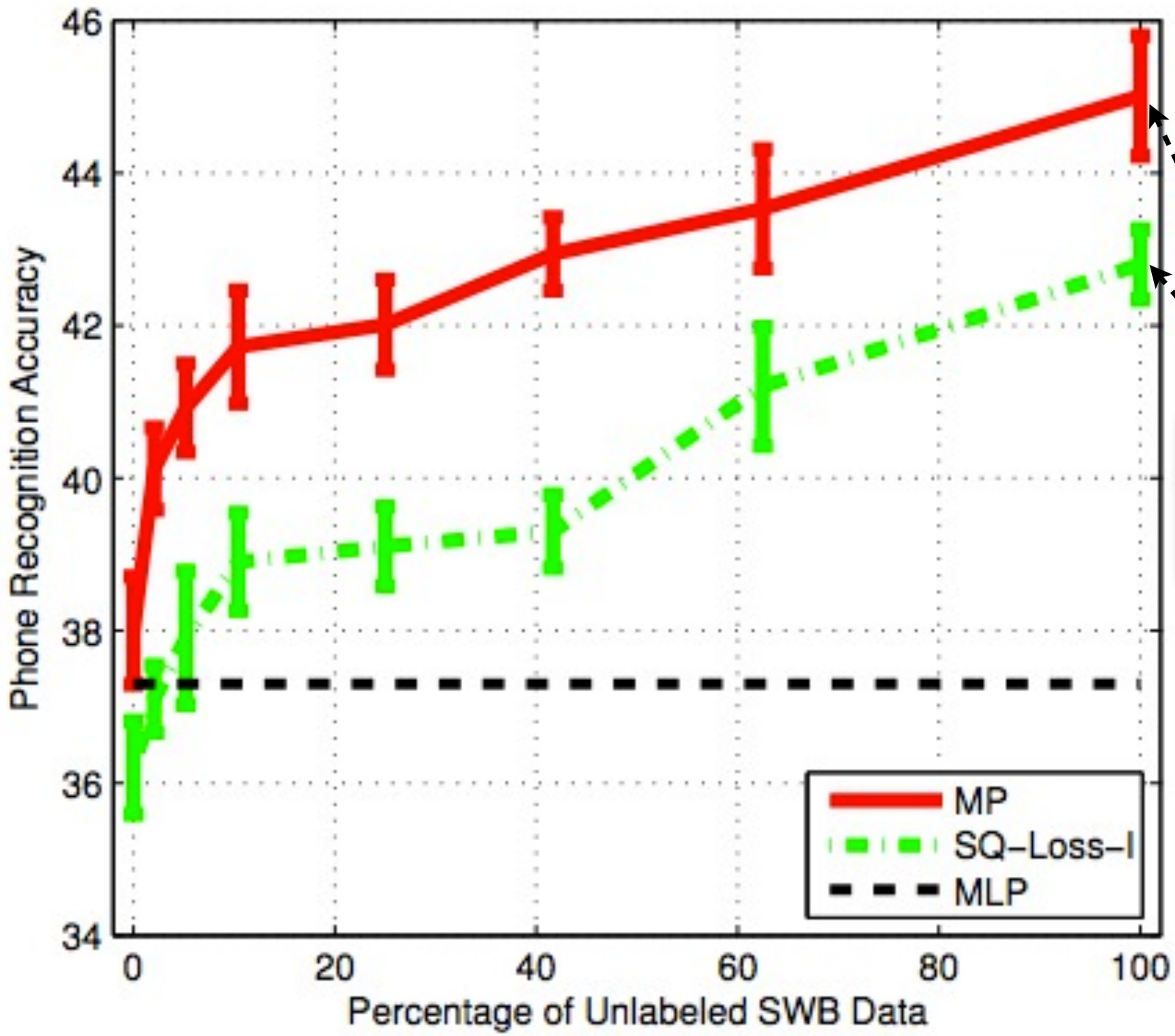
- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

More (Unlabeled) Data is Better Data



[Subramanya & Bilmes, JMLR 2011]

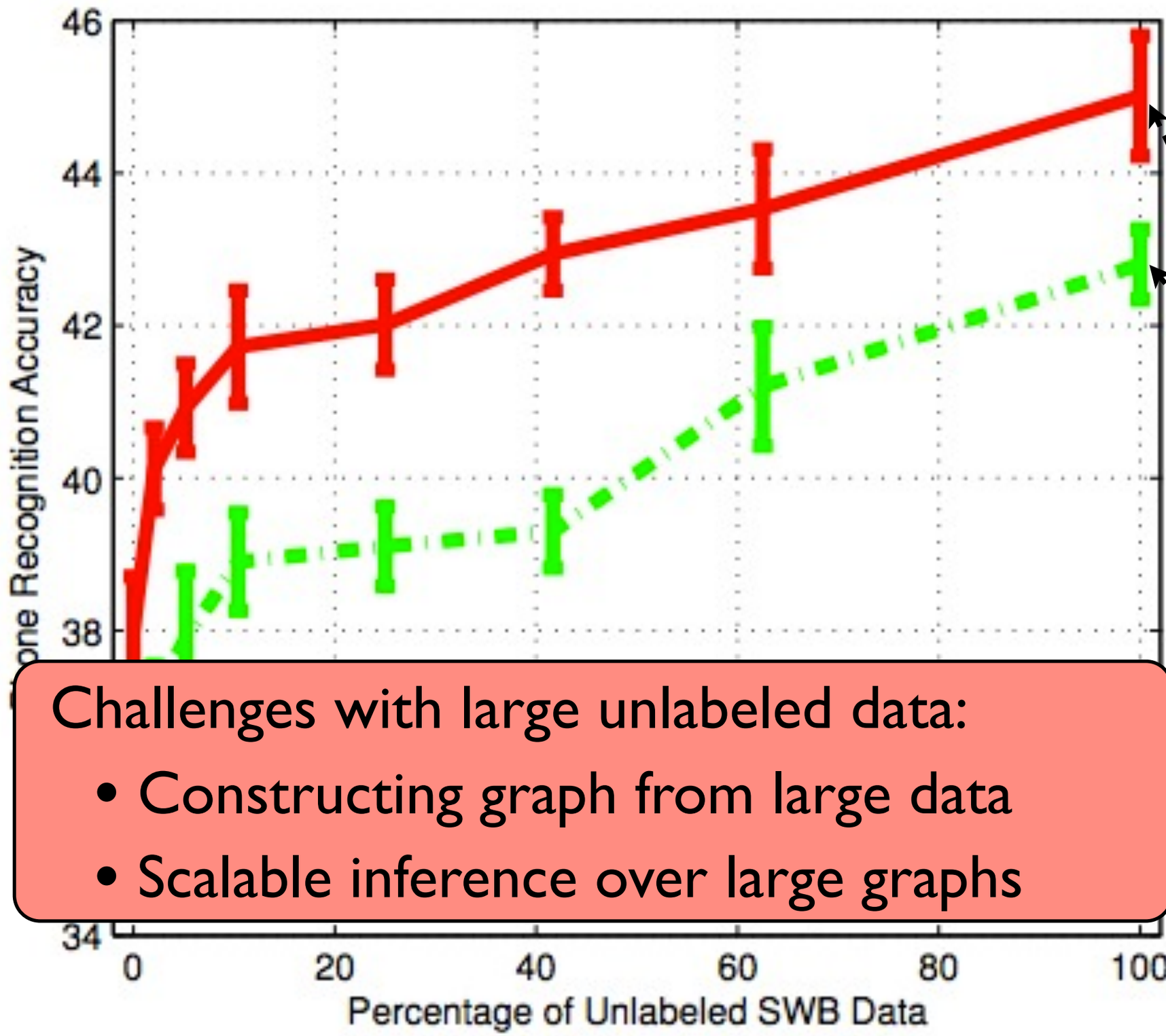
More (Unlabeled) Data is Better Data



Graph with **120m** vertices

[Subramanya & Bilmes, JMLR 2011]

More (Unlabeled) Data is Better Data



Graph with **120m** vertices

Challenges with large unlabeled data:

- Constructing graph from large data
- Scalable inference over large graphs

[Subramanya & Bilmes, JMLR 2011]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

Scalability Issues (I)

Graph Construction

Scalability Issues (I)

Graph Construction

- Brute force (exact) k-NNG too expensive (quadratic)

Scalability Issues (I)

Graph Construction

- Brute force (exact) k-NNG too expensive (quadratic)
  - Approximate nearest neighbor using kd-tree [Friedman et al., 1977, also see http://www.cs.umd.edu/~mount/]

Scalability Issues (II)

Label Inference

- Sub-sample the data
  - Construct graph over a subset of a unlabeled data [Delalleau et al., AISTATS 2005]
  - Sparse Grids [Garcke & Griebel, KDD 2001]

Scalability Issues (II)

Label Inference

- Sub-sample the data
  - Construct graph over a subset of a unlabeled data [Delalleau et al., AISTATS 2005]
  - Sparse Grids [Garcke & Griebel, KDD 2001]
- How about using <u>more computation</u>? (next section)
  - Symmetric multi-processor (SMP)
  - Distributed Computer

Outline

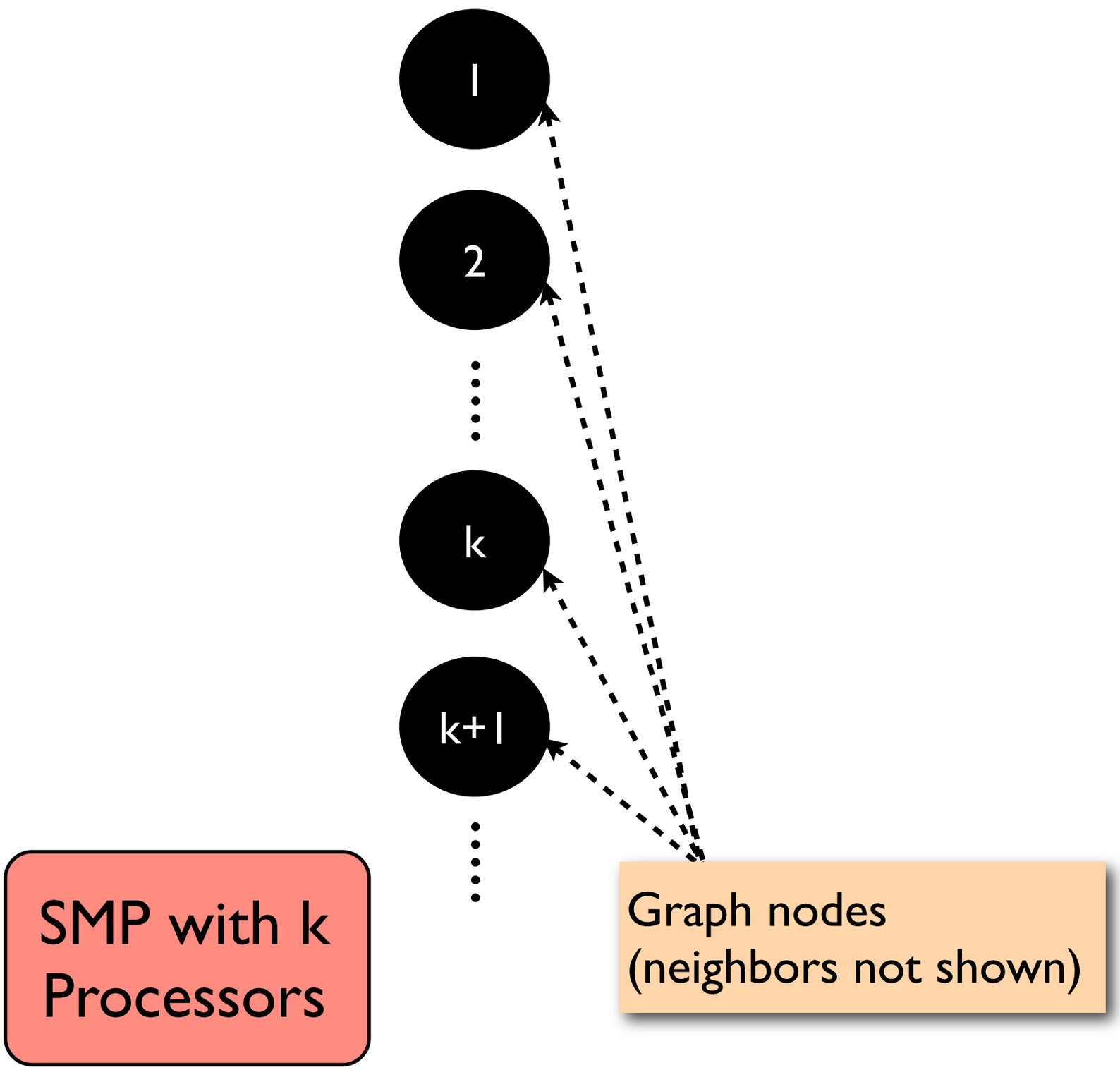
- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Scalability Issues

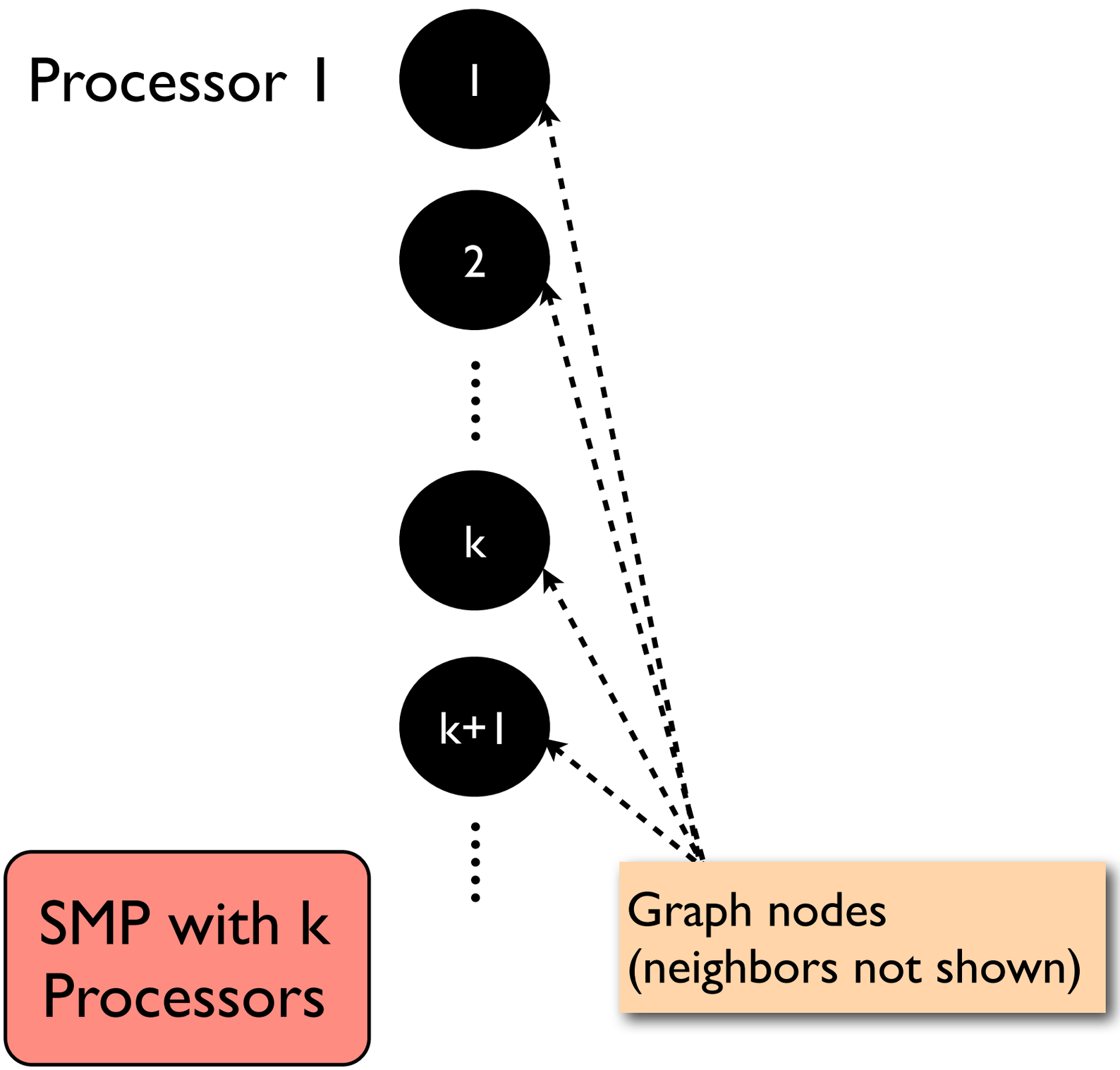
Node reordering
[Subramanya & Bilmes, JMLR 2011; Bilmes & Subramanya, 2011]

MapReduce Parallelization

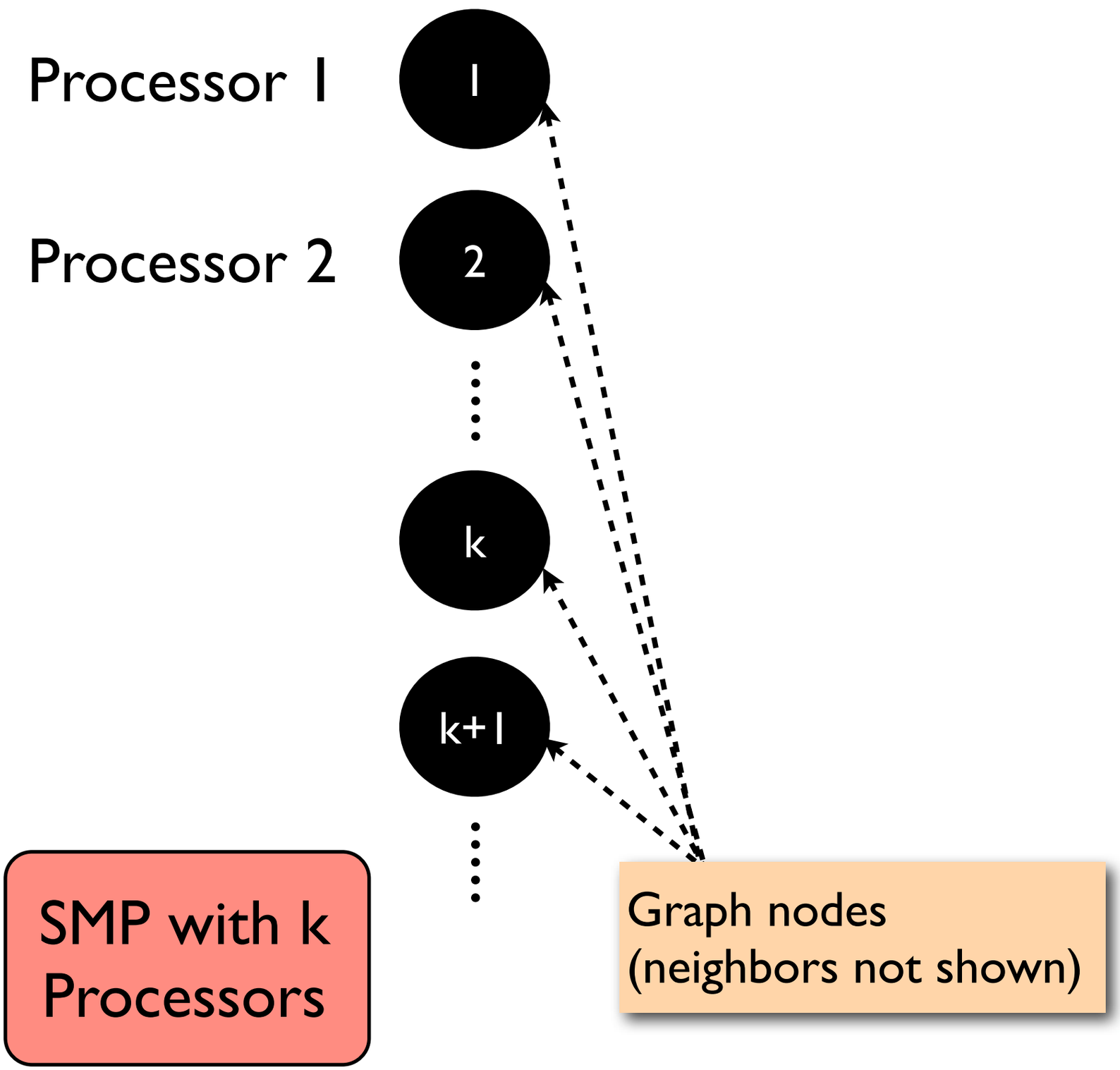
Parallel computation on a SMP



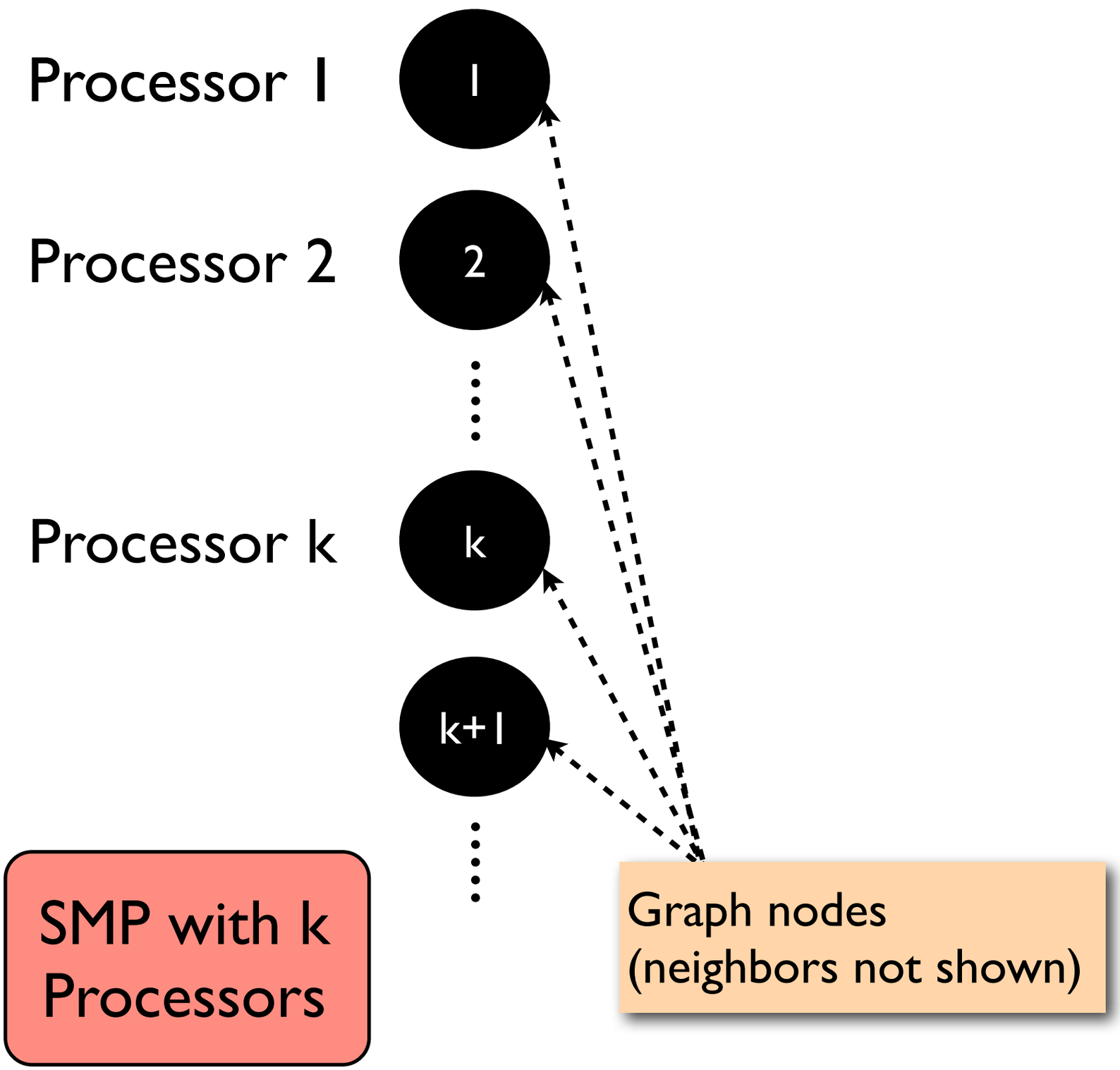
Parallel computation on a SMP



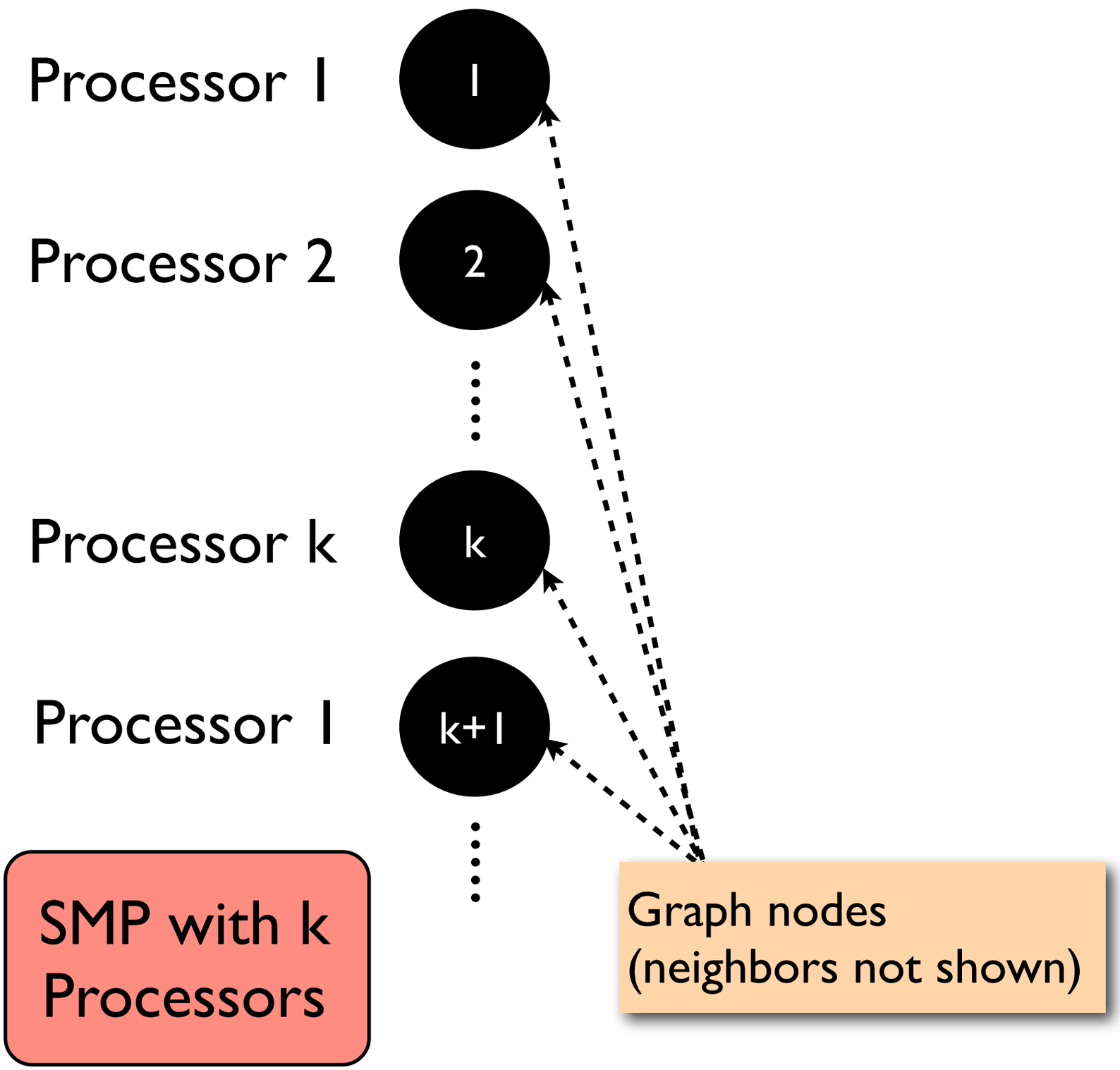
Parallel computation on a SMP



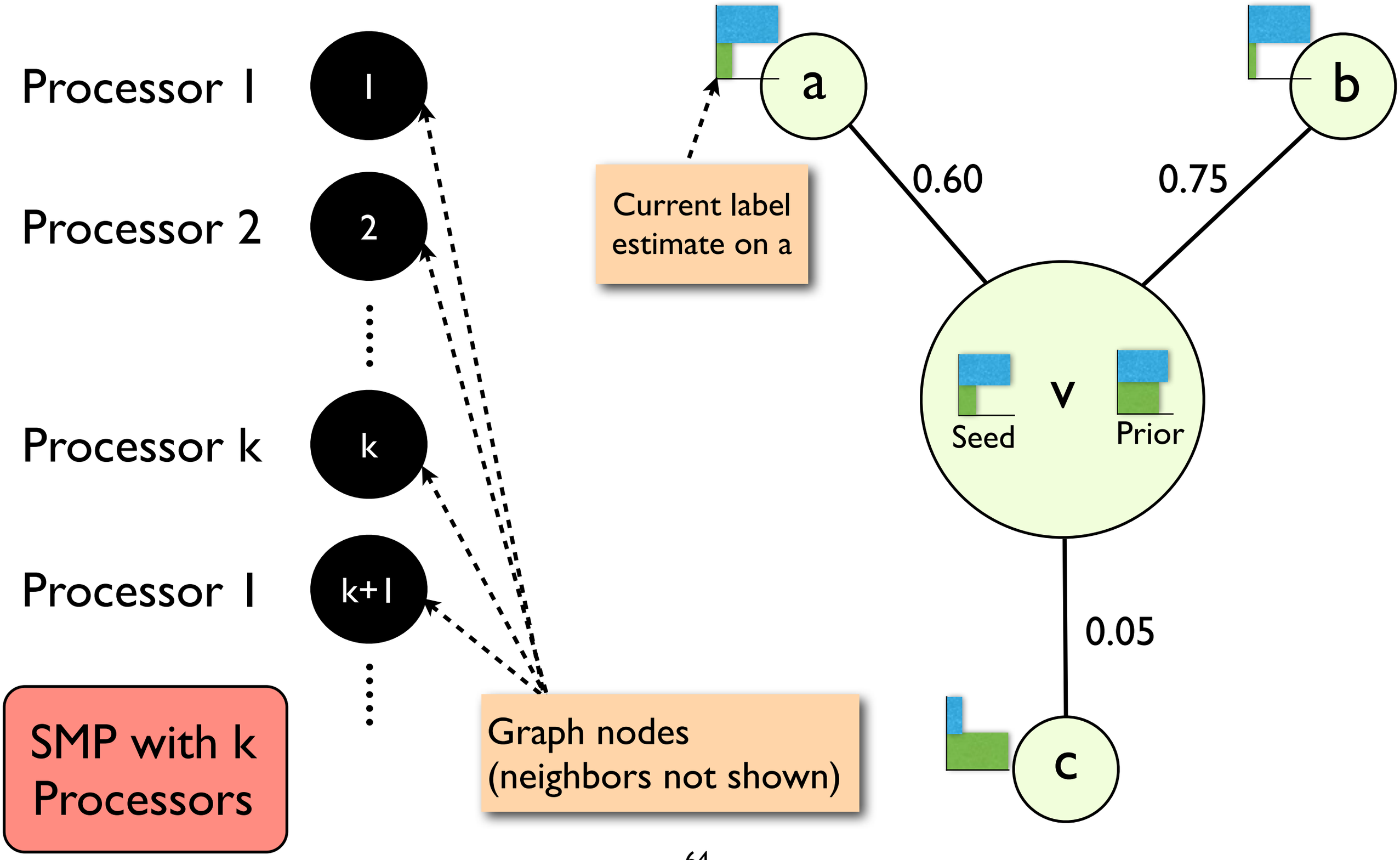
Parallel computation on a SMP



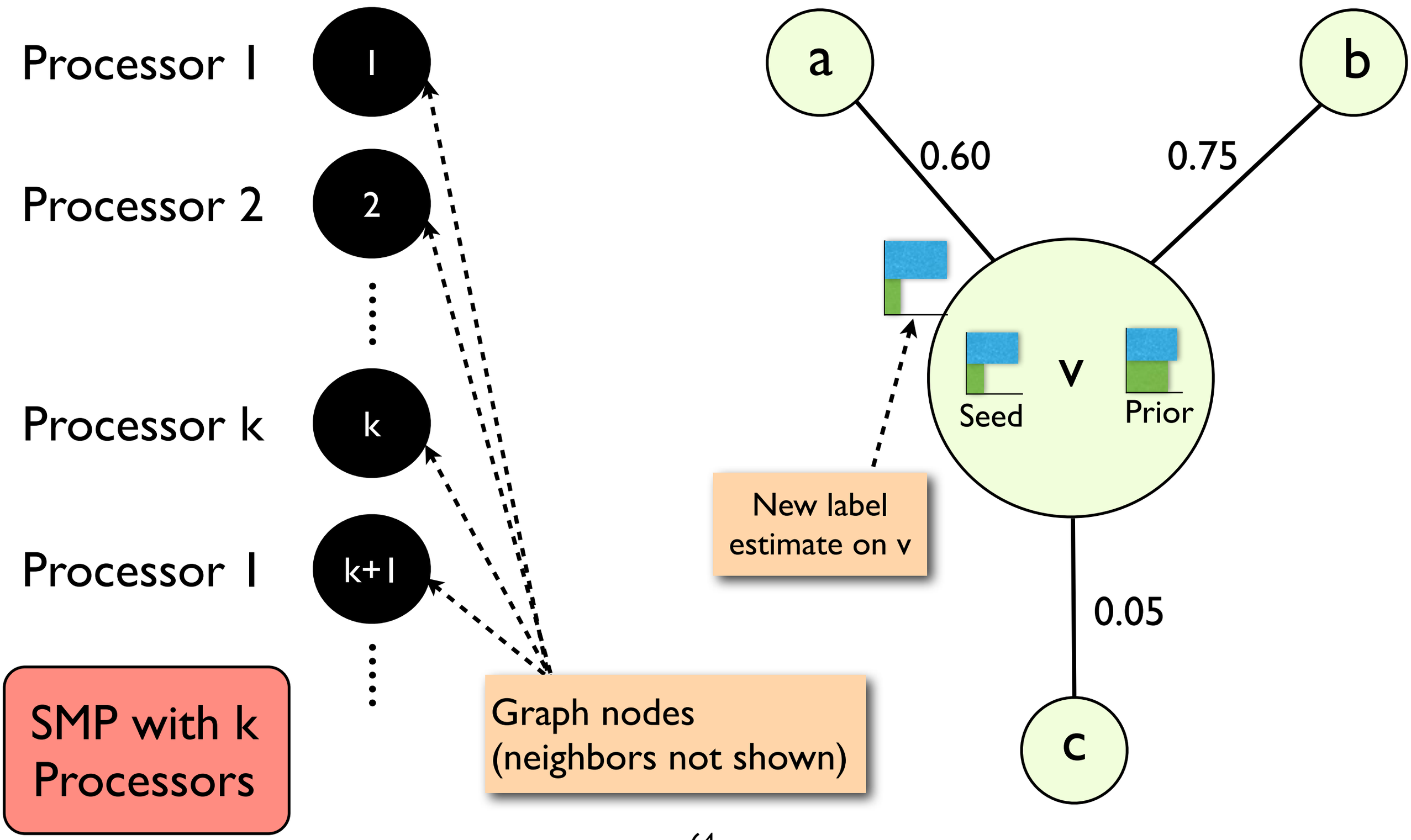
Parallel computation on a SMP



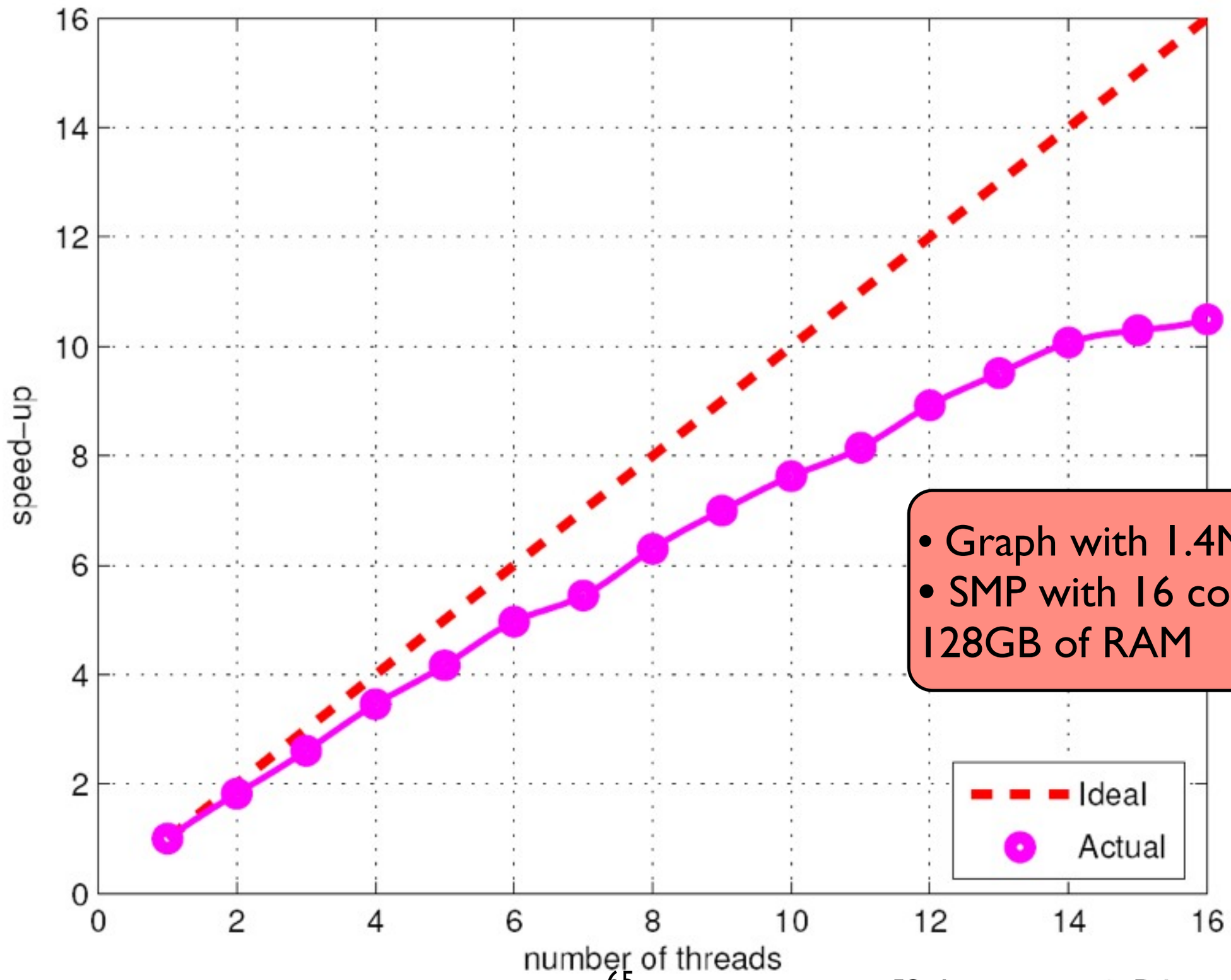
Label Update using Message Passing



Label Update using Message Passing



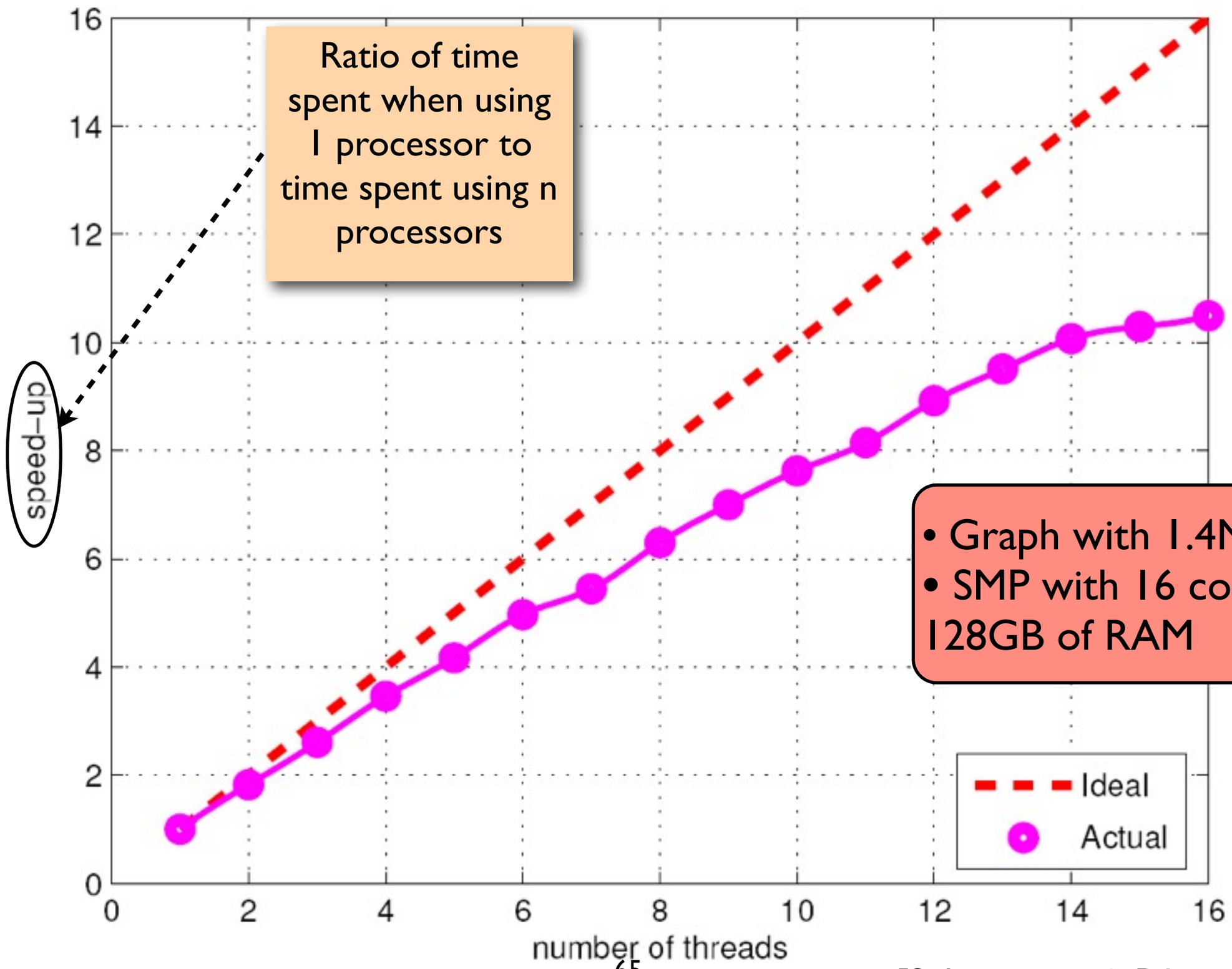
Speed-up on SMP



- Graph with 1.4M nodes
- SMP with 16 cores and 128GB of RAM

[Subramanya & Bilmes, JMLR, 2011]

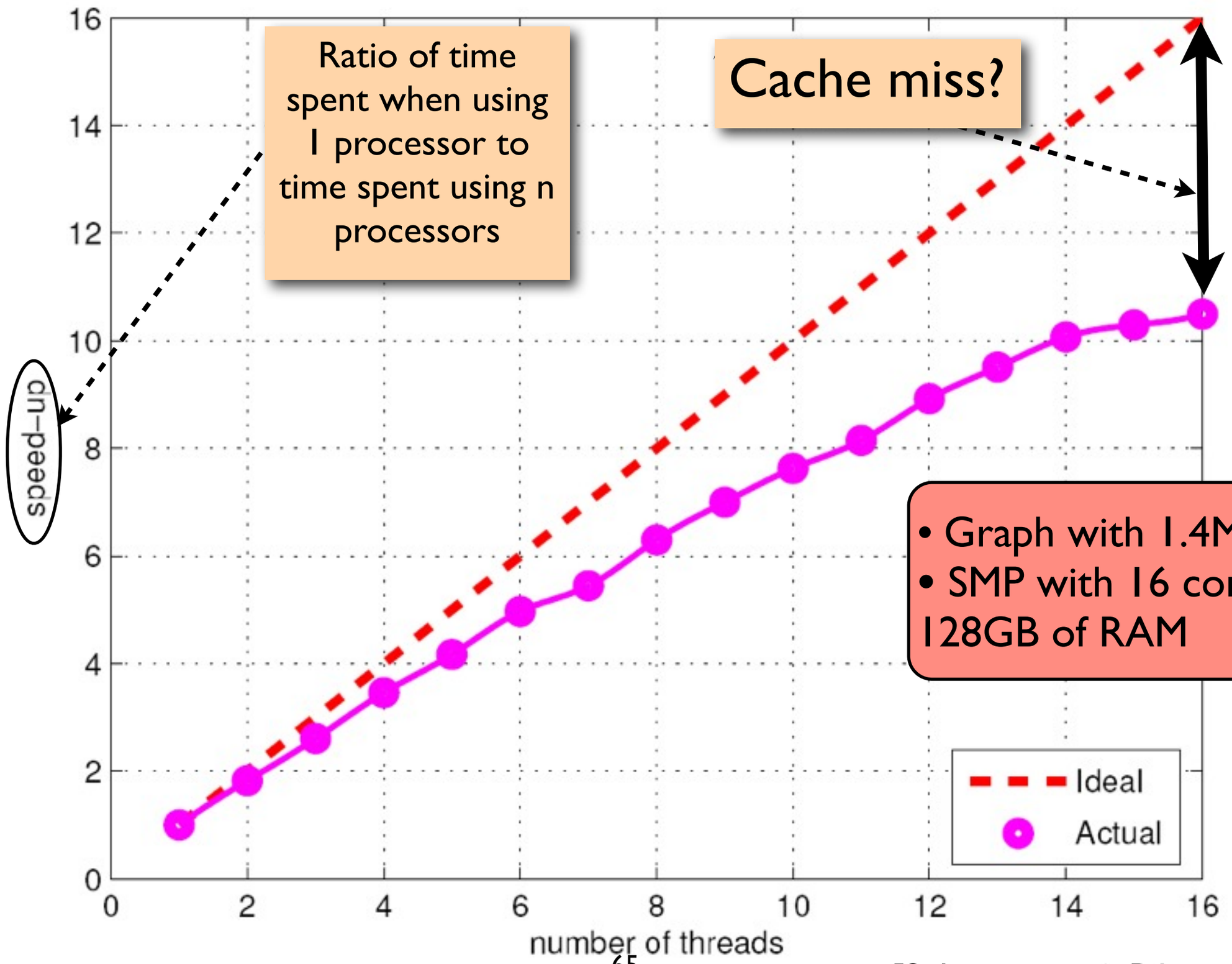
Speed-up on SMP



- Graph with 1.4M nodes
- SMP with 16 cores and 128GB of RAM

[Subramanya & Bilmes, JMLR, 2011]

Speed-up on SMP



- Graph with 1.4M nodes
- SMP with 16 cores and 128GB of RAM

[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm

Input: Graph G = (V, E)

Result: Node ordered graph

1. Select an arbitrary node v
2. while unselected nodes remain do
   - 2.1. select an unselected node v` from among the neighbors' neighbors of v that has maximum overlap with v` neighbors
   - 2.2. mark v` as selected
   - 2.3. set v to v`

[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm

Input: Graph G = (V, E)

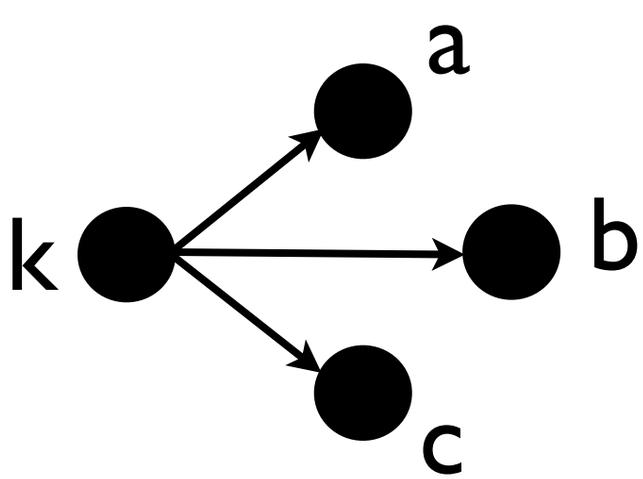
Result: Node ordered graph

1. Select an arbitrary node v
2. while unselected nodes remain do
   2.1. select an unselected node v` from among the neighbors' neighbors of v that has maximum overlap with v` neighbors
   2.2. mark v` as selected
   2.3. set v to v`

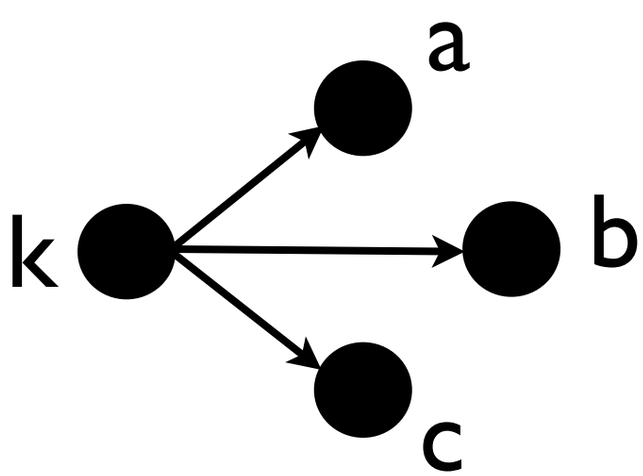
Exhaustive for sparse (e.g., k-NN) graphs

[Subramanya & Bilmes, JMLR, 2011]

Node Reordering Algorithm : Intuition

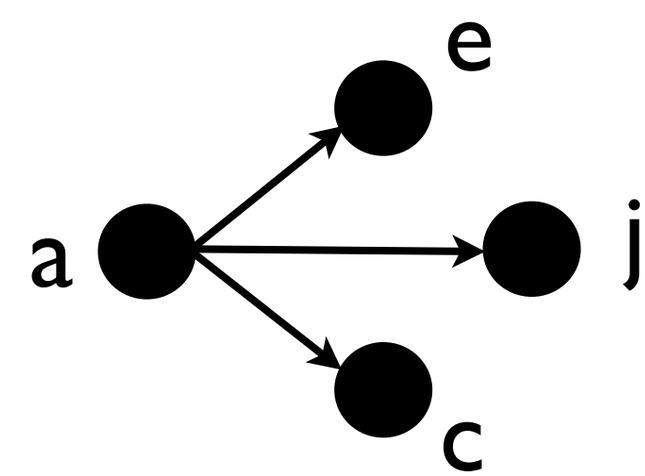


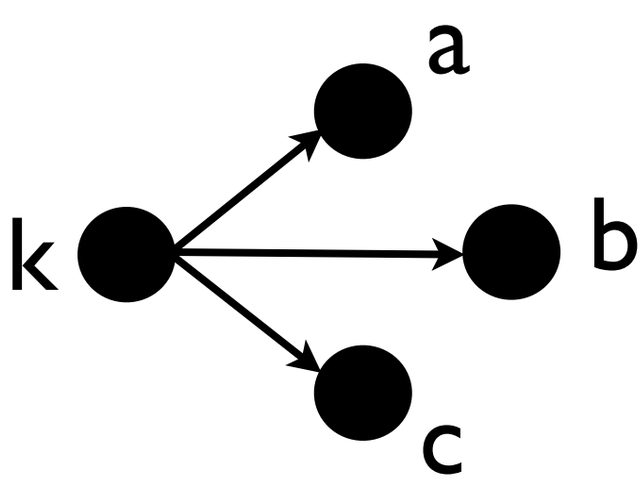
Node Reordering Algorithm : Intuition



Which node should be placed after k to optimize cache performance?

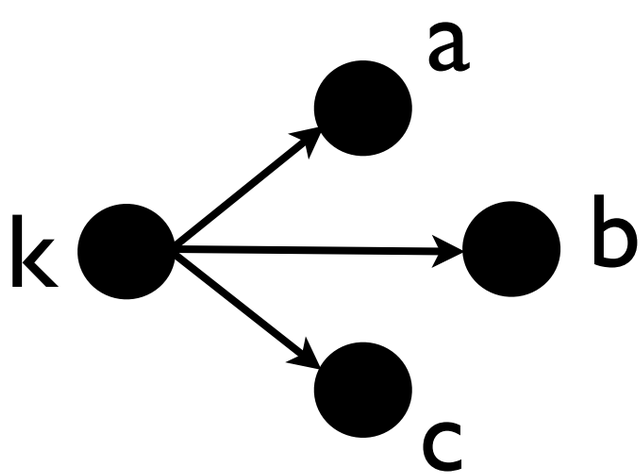
Node Reordering Algorithm : Intuition

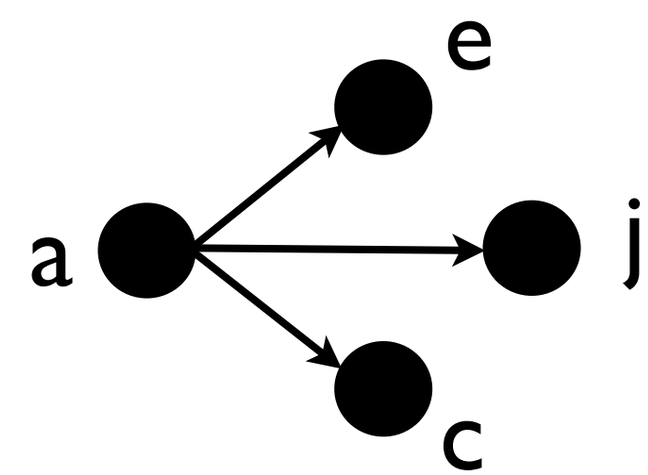


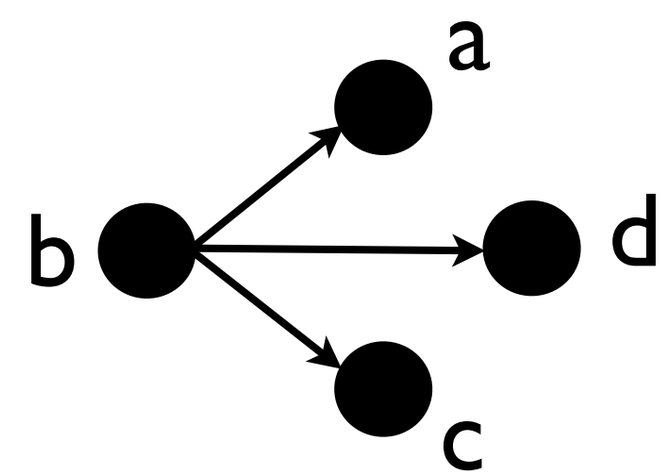


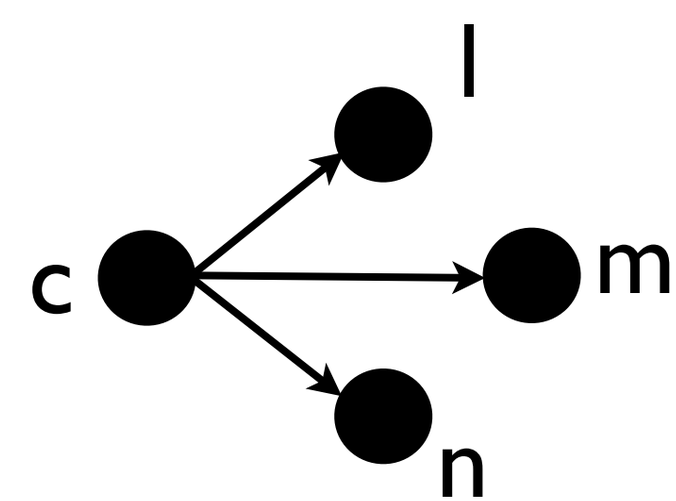
Which node should be placed after k to optimize cache performance?

Node Reordering Algorithm : Intuition



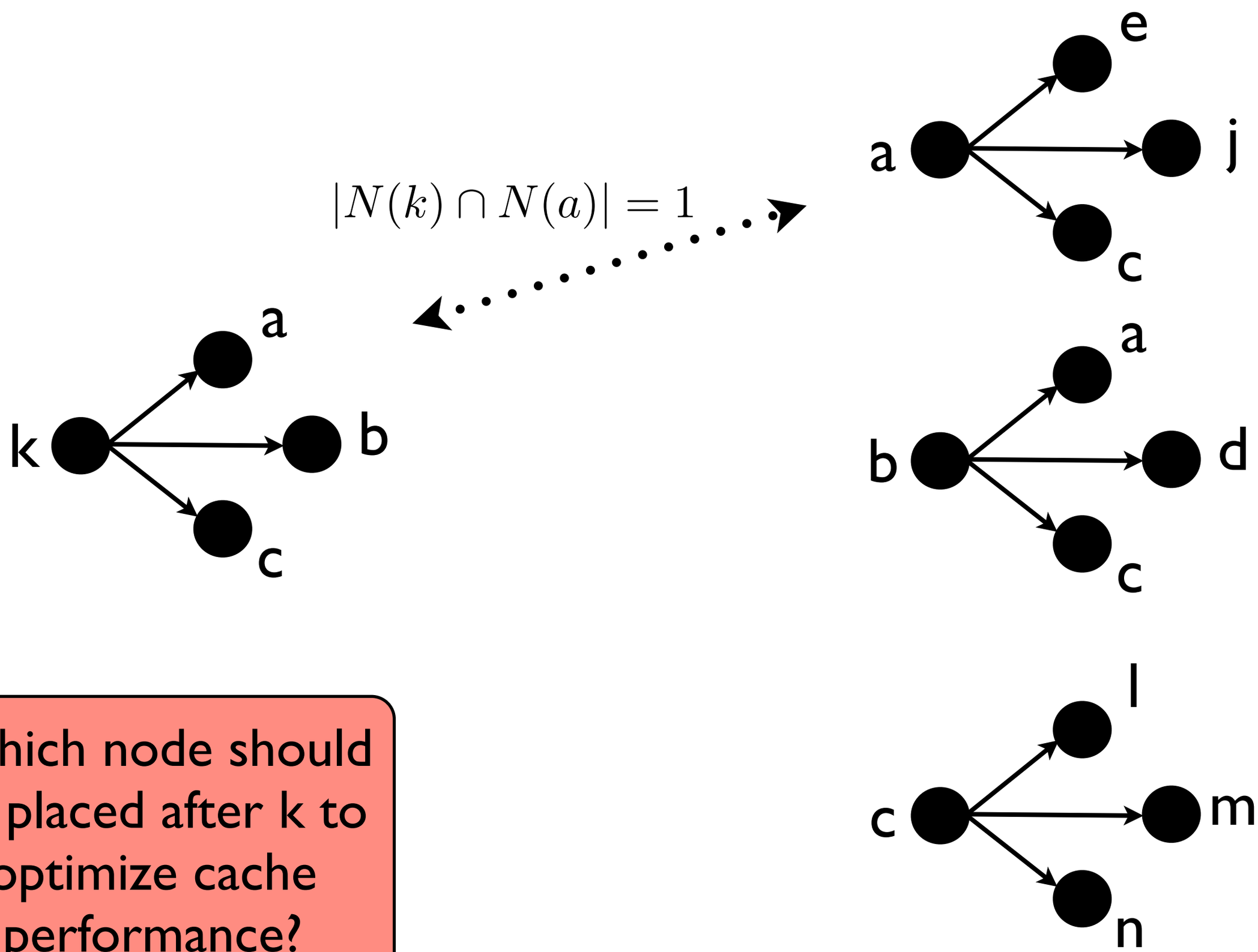






Which node should be placed after k to optimize cache performance?

Node Reordering Algorithm : Intuition

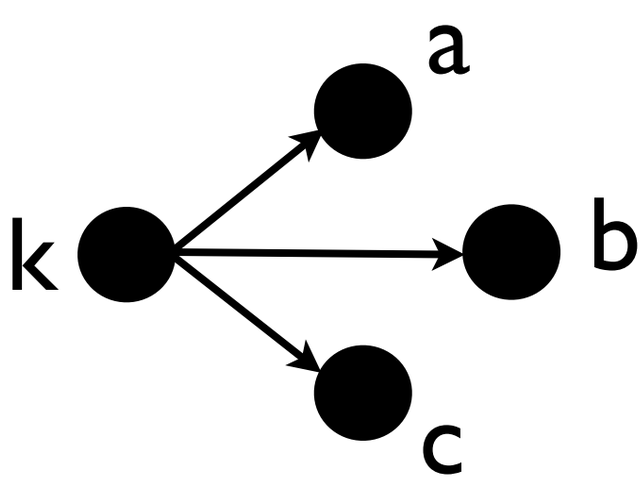


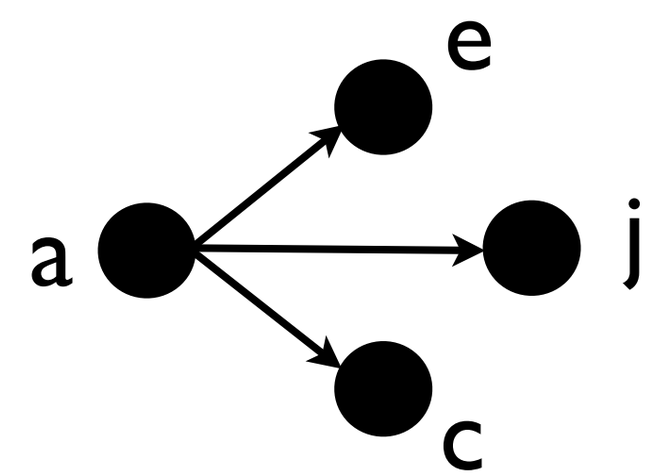
Which node should be placed after k to optimize cache performance?

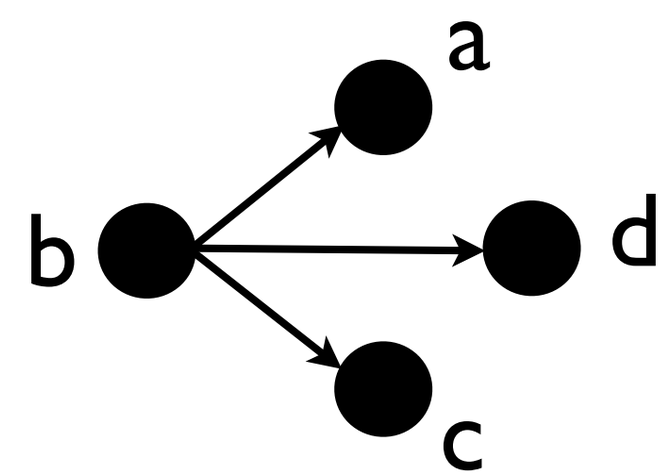
Node Reordering Algorithm : Intuition

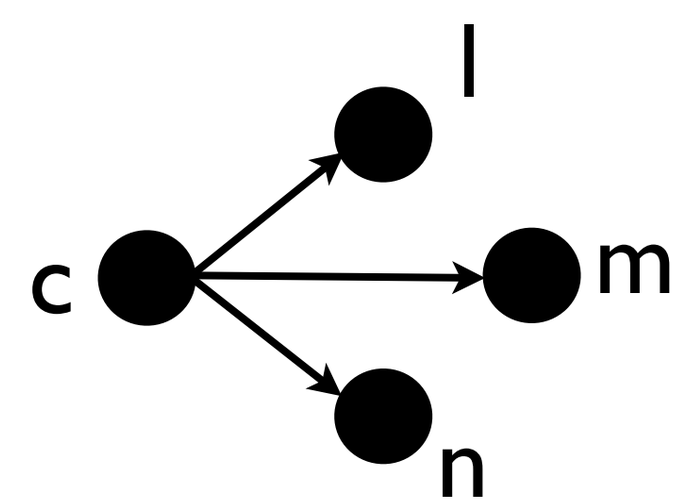
Cardinality of Intersection

$$|N(k) \cap N(a)| = 1$$





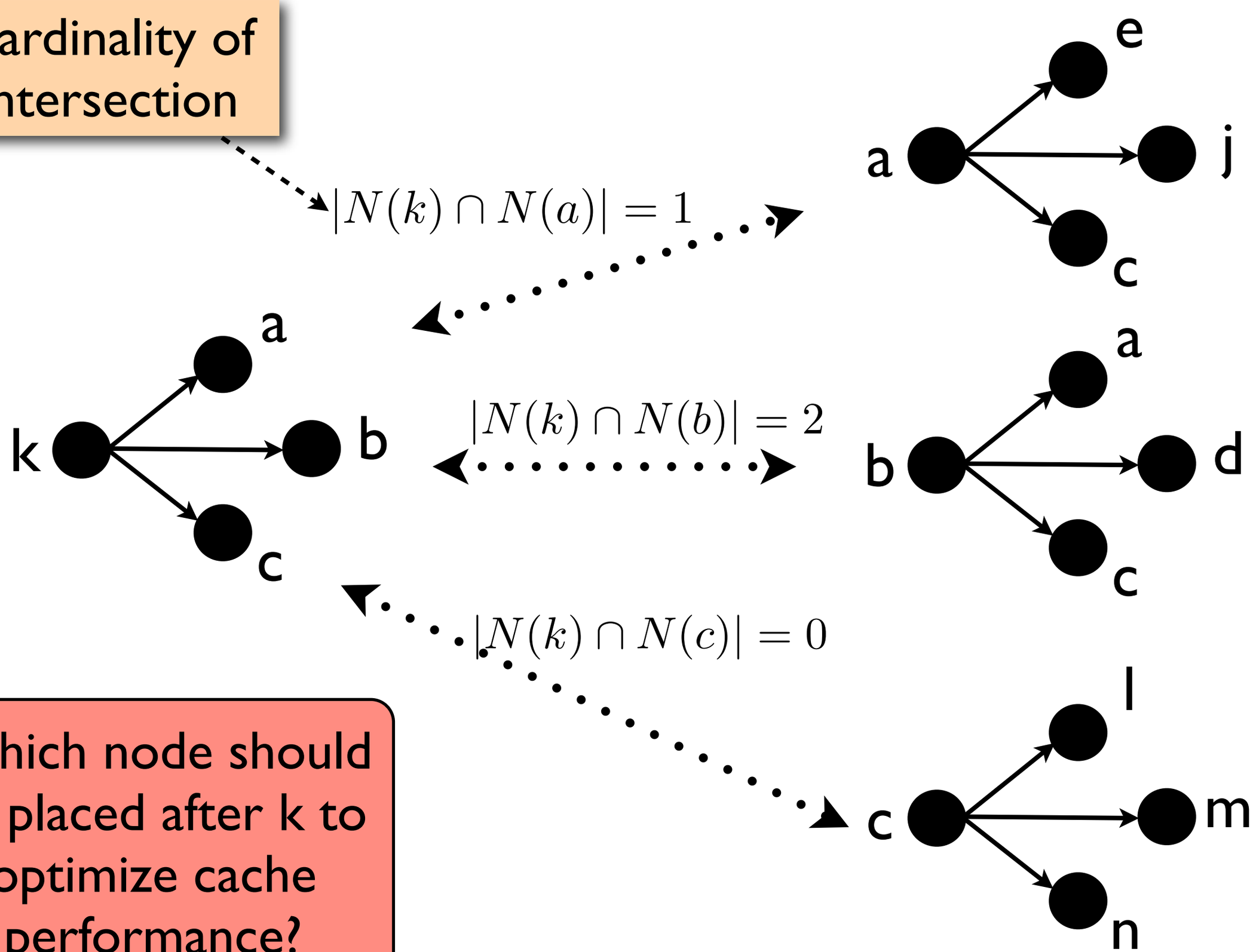




Which node should be placed after k to optimize cache performance?

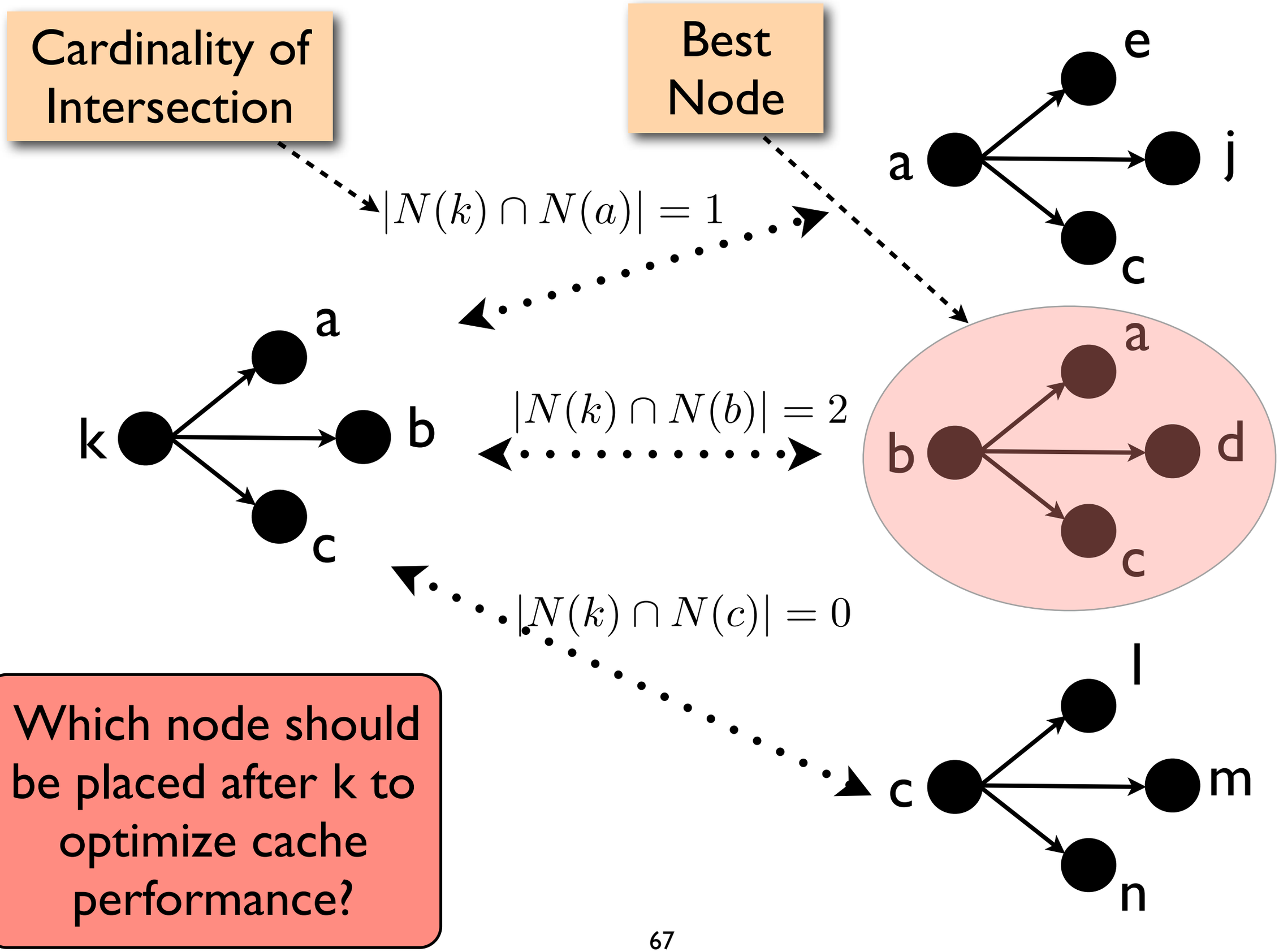
Node Reordering Algorithm : Intuition

Cardinality of Intersection



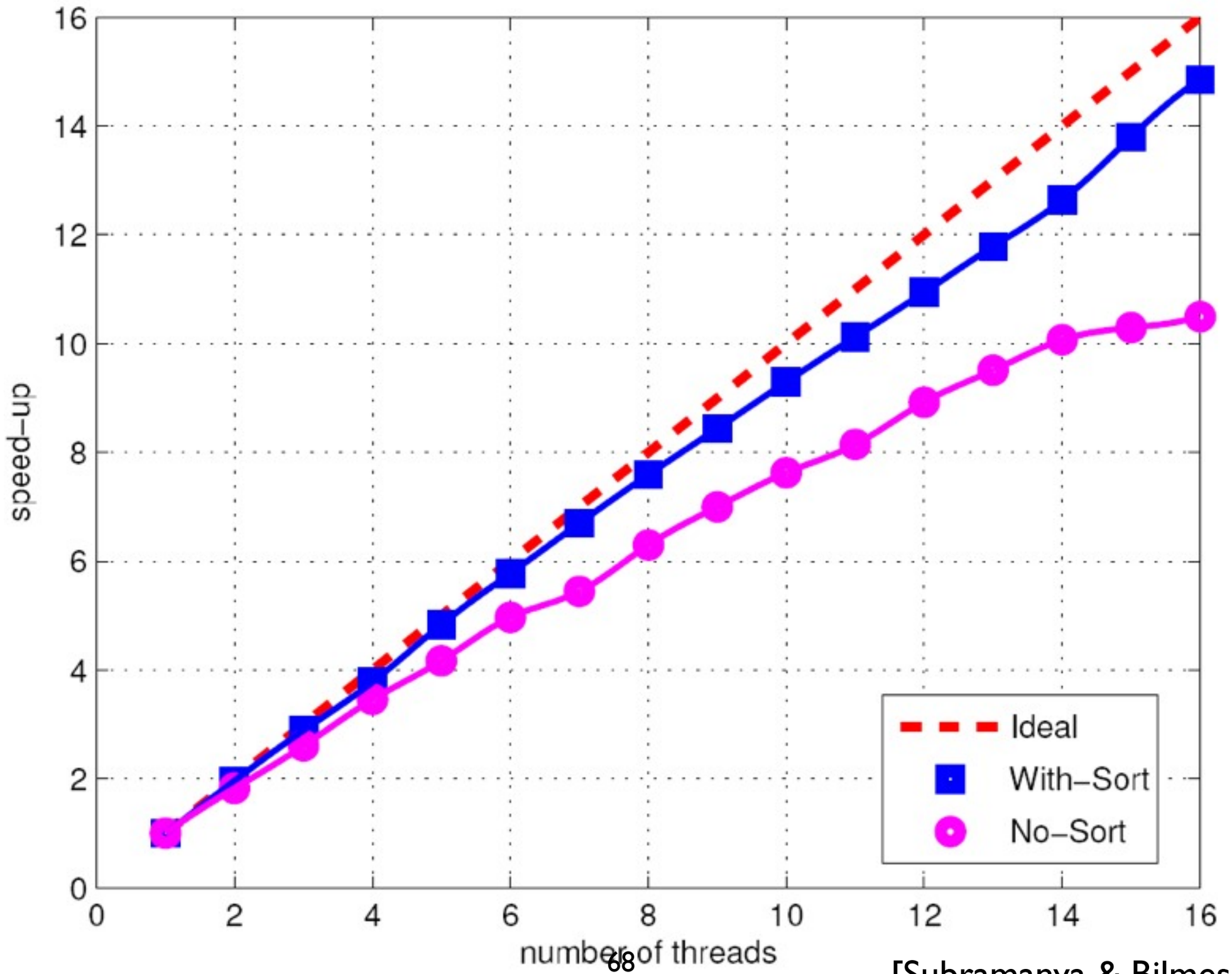
Which node should be placed after k to optimize cache performance?

Node Reordering Algorithm : Intuition



Which node should be placed after k to optimize cache performance?

Speed-up on SMP after Node Ordering

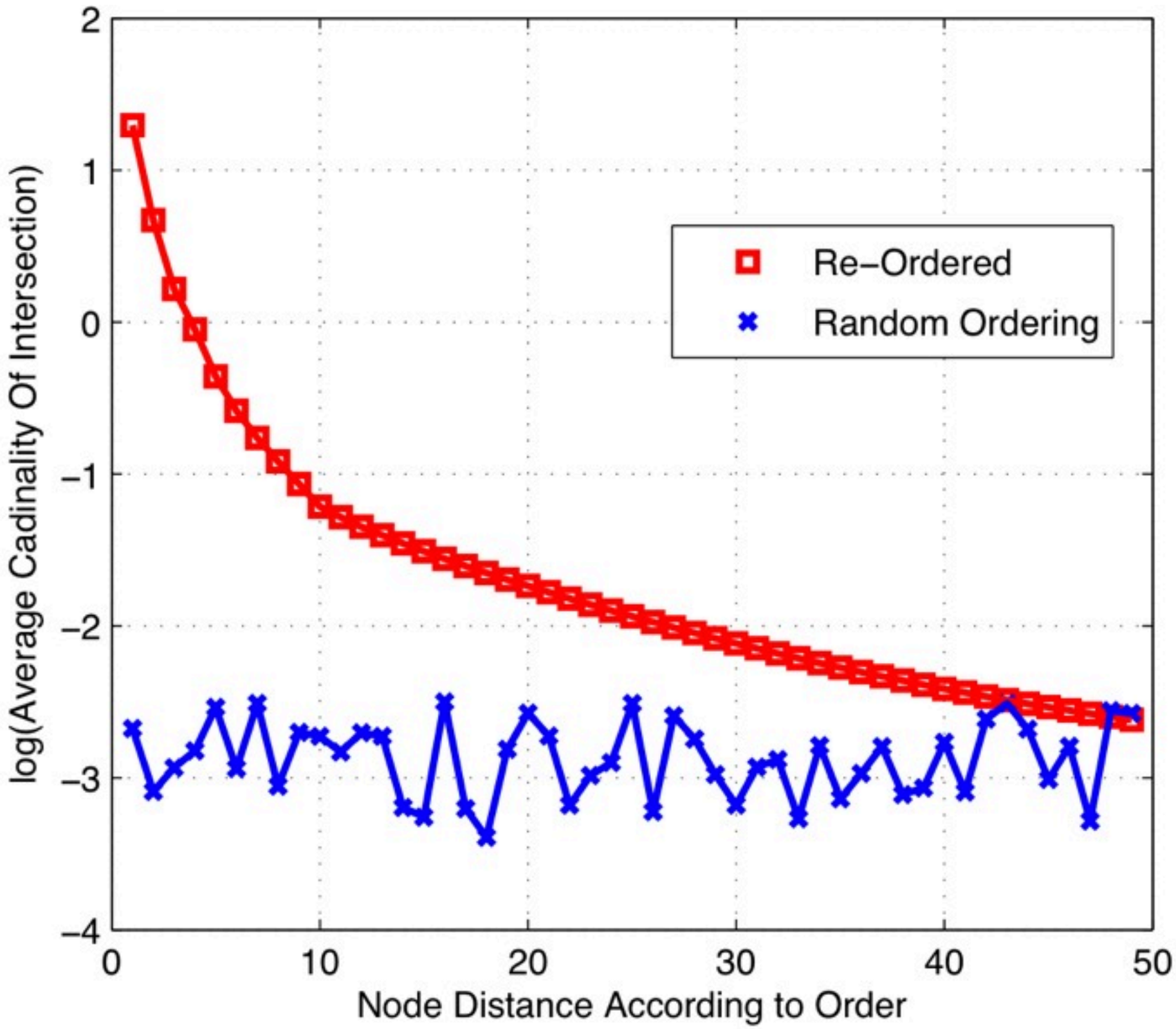


[Subramanya & Bilmes, JMLR, 2011]

Distributed Processing

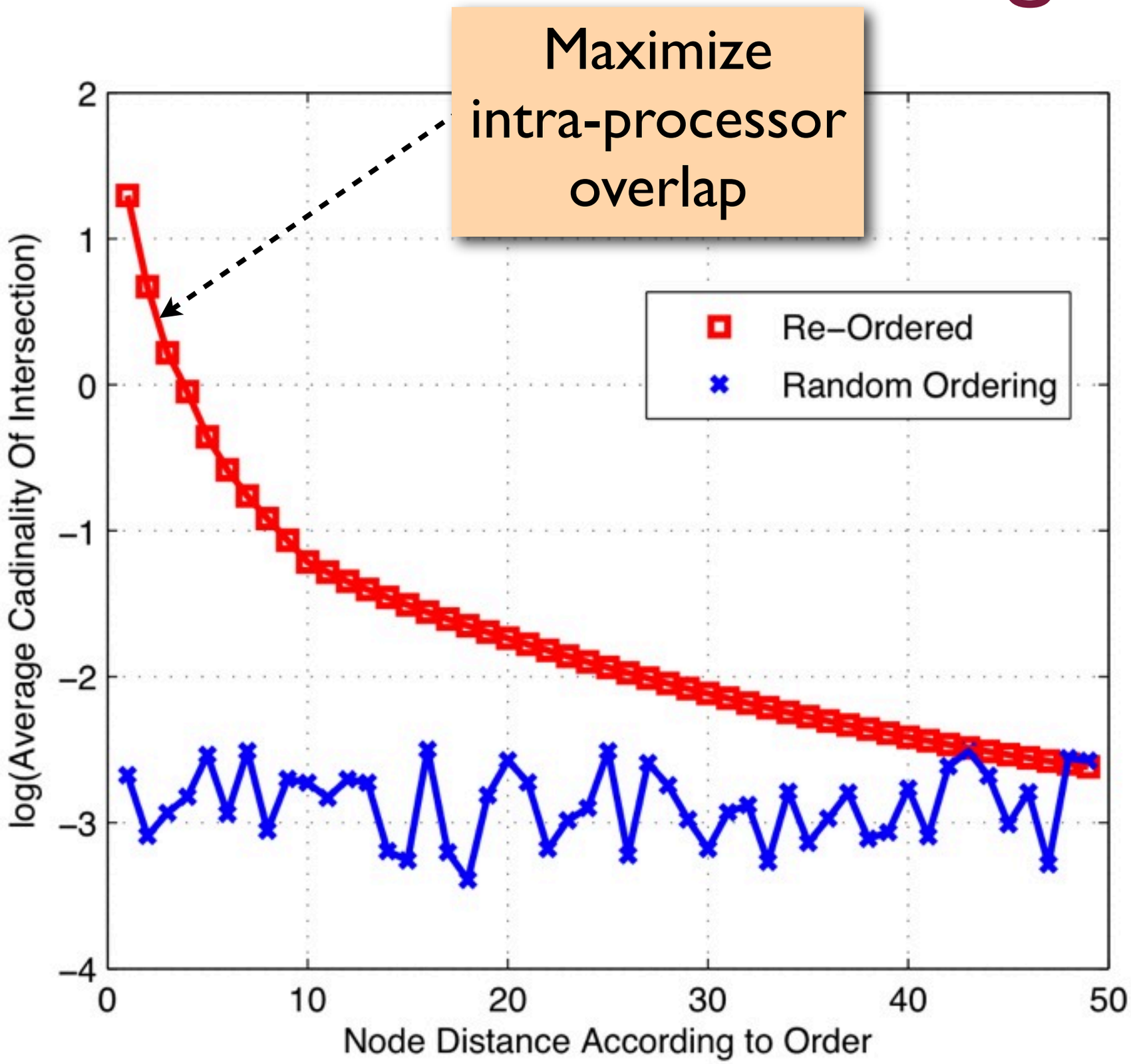
- **Maximize** overlap between consecutive nodes within the same machine
- **Minimize** overlap across machines (reduce inter machine communication)

Distributed Processing



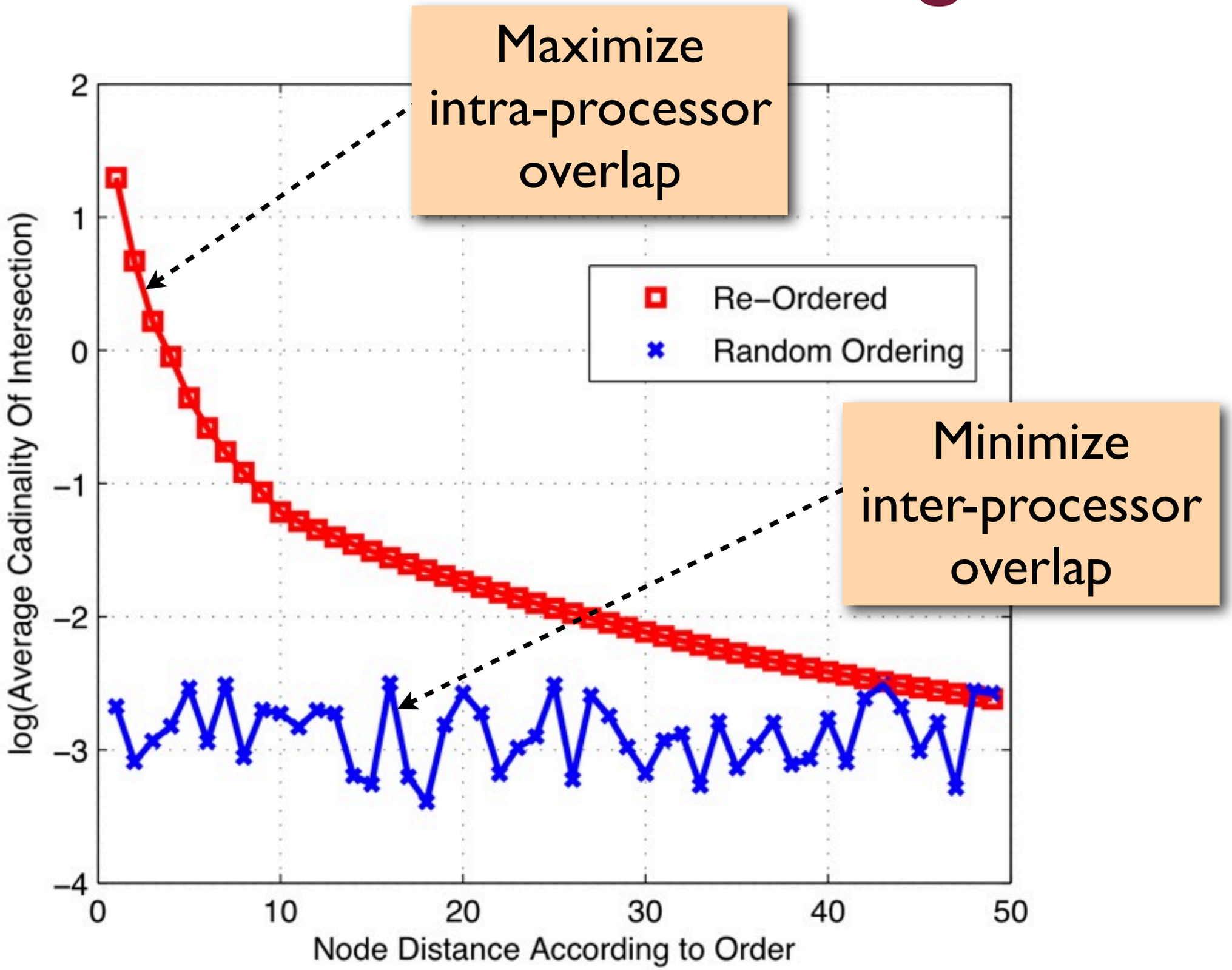
[Bilmes & Subramanya, 2011]

Distributed Processing



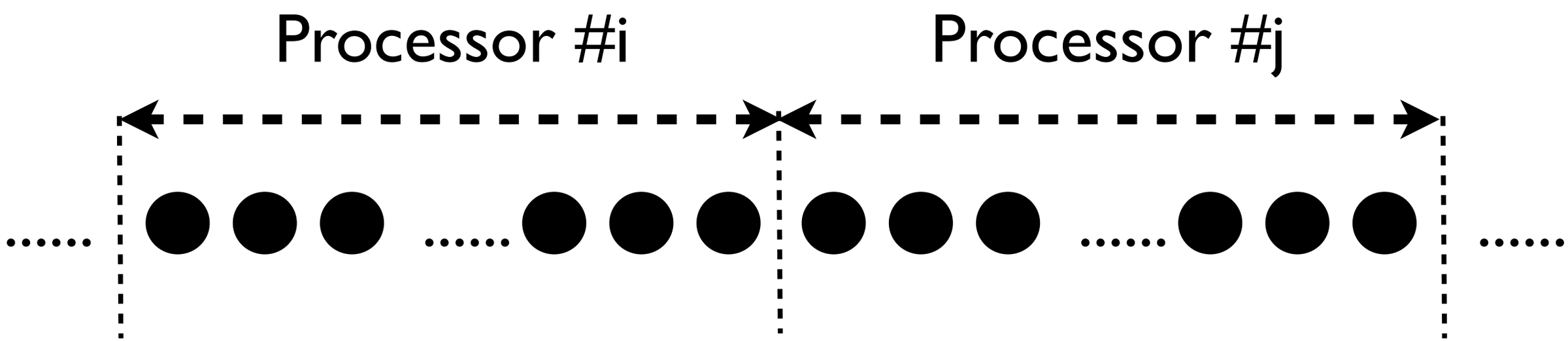
[Bilmes & Subramanya, 2011]

Distributed Processing

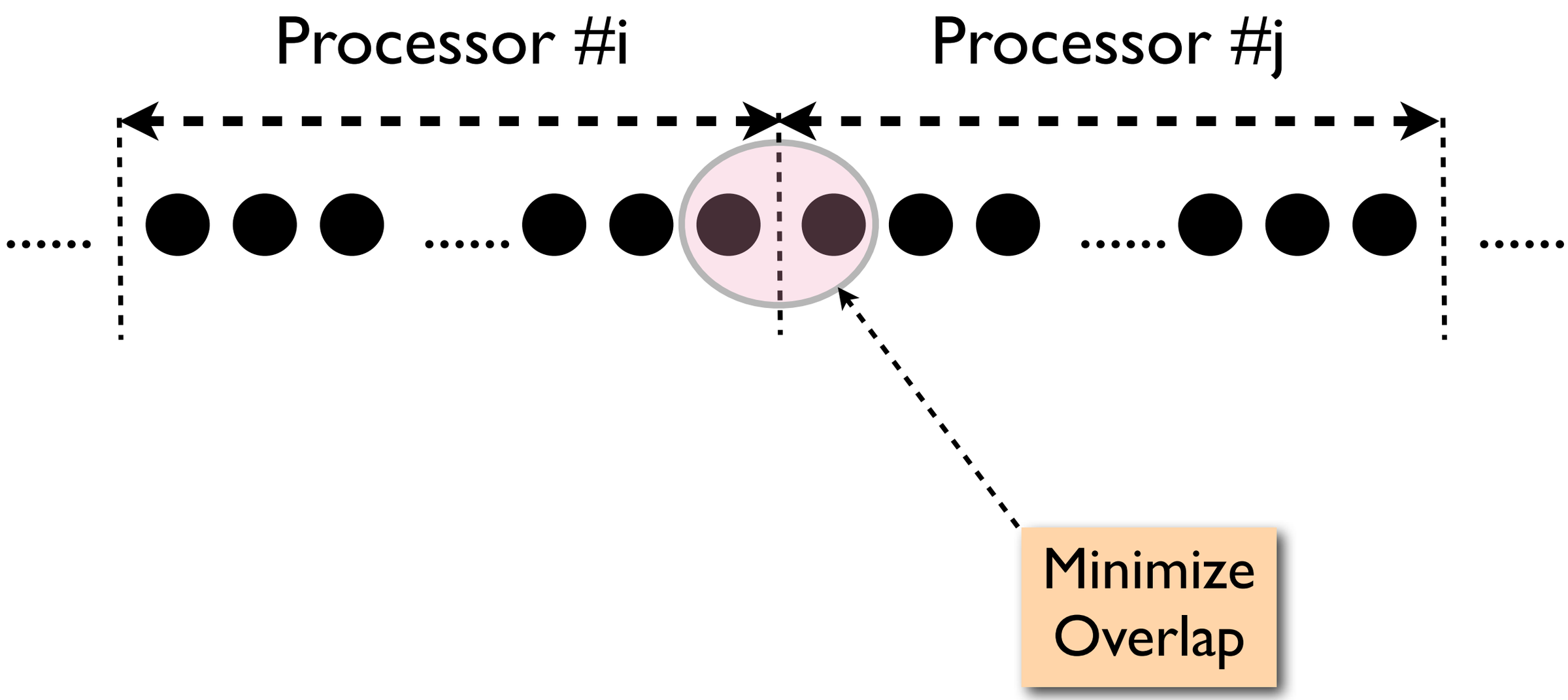


[Bilmes & Subramanya, 2011]

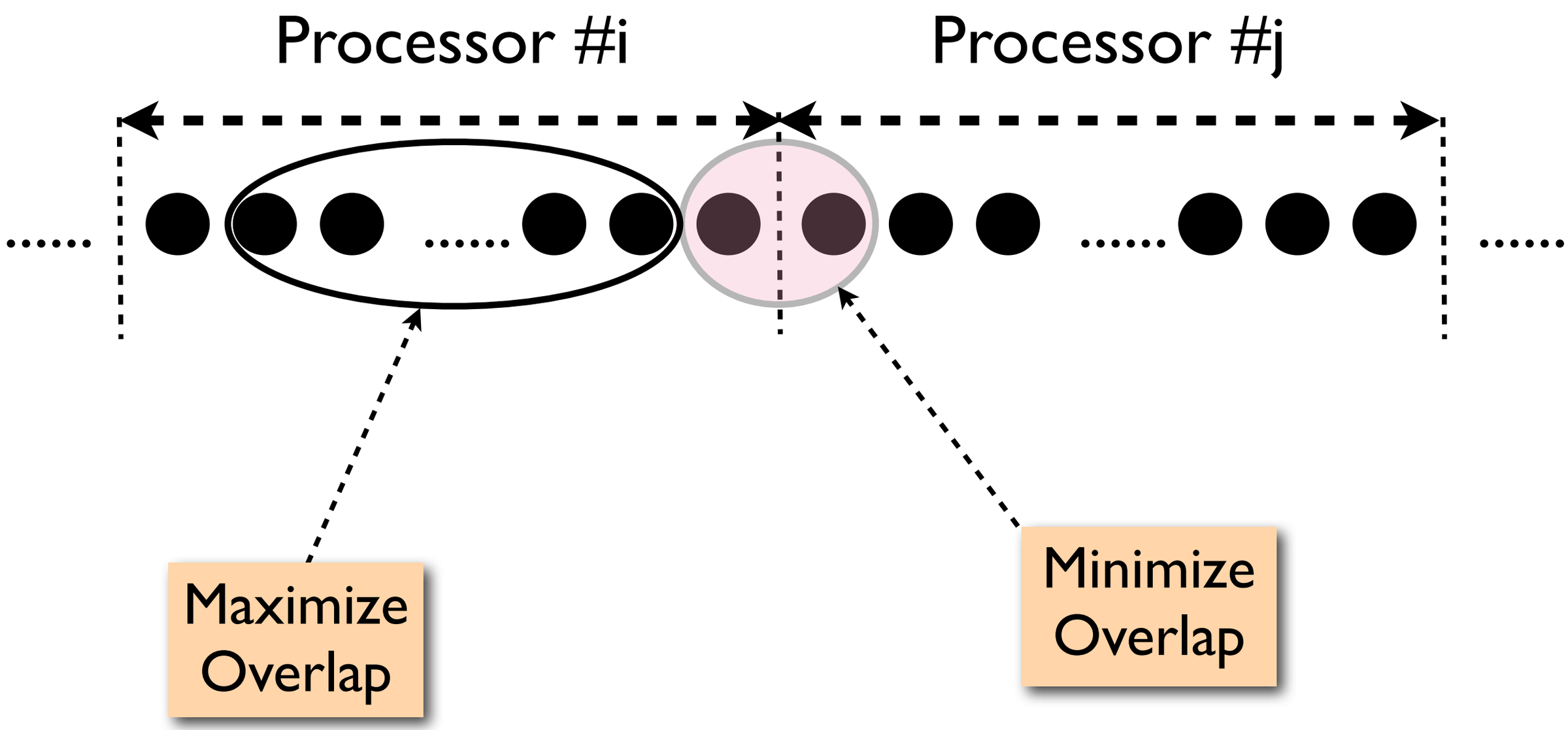
Node reordering for Distributed Computer



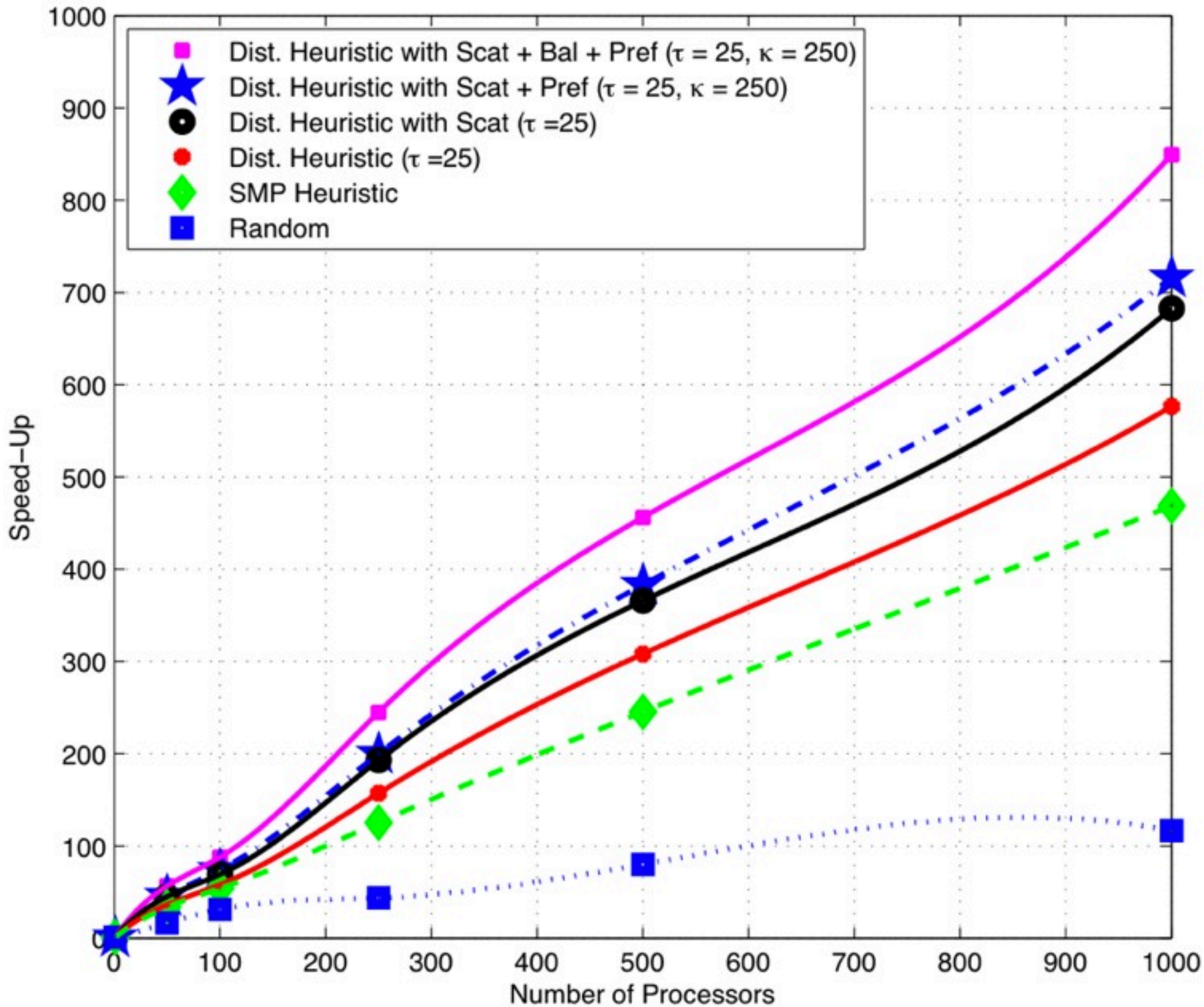
Node reordering for Distributed Computer



Node reordering for Distributed Computer



Distributed Processing Results

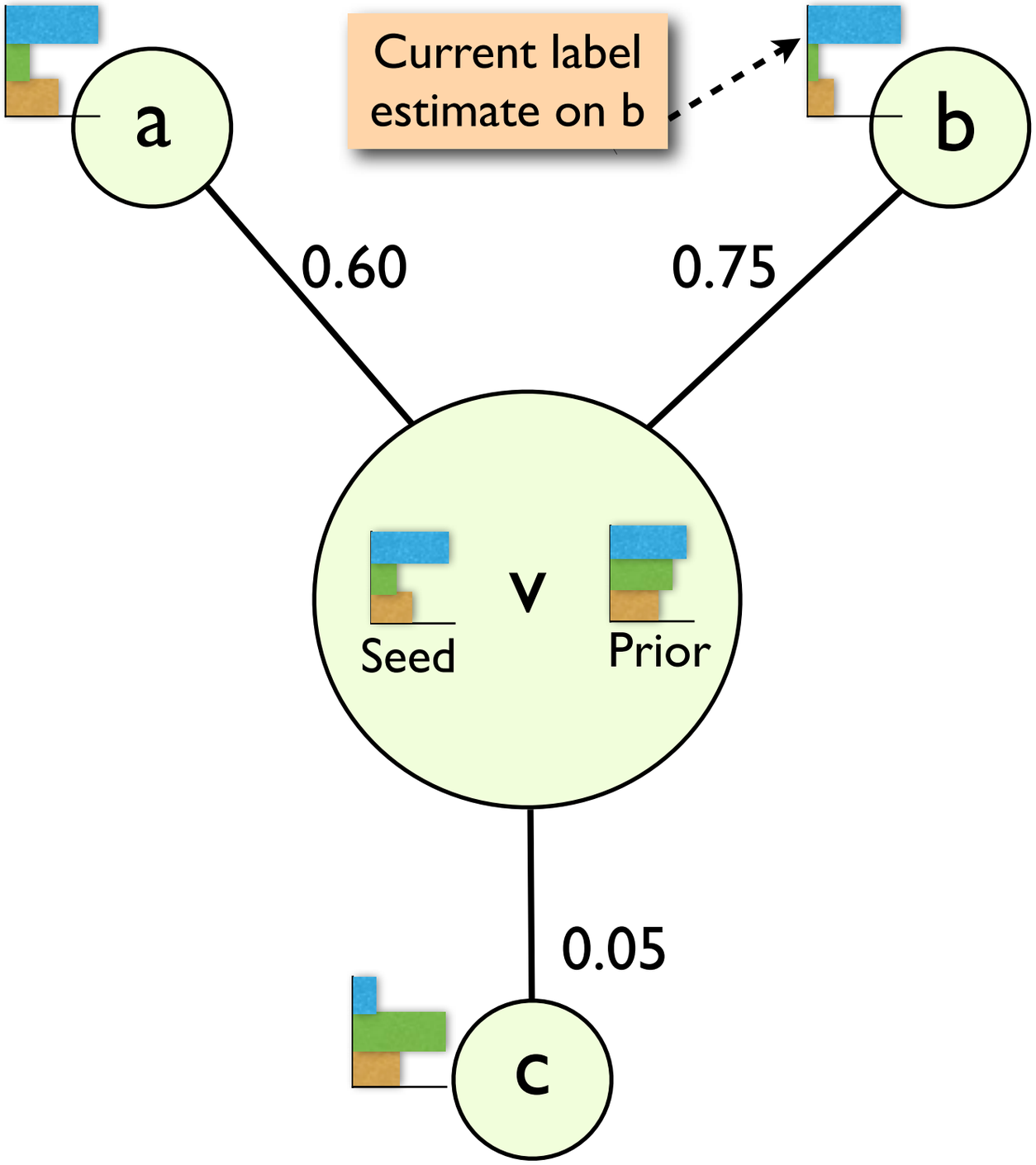


[Bilmes & Subramanya, 2011]

Outline

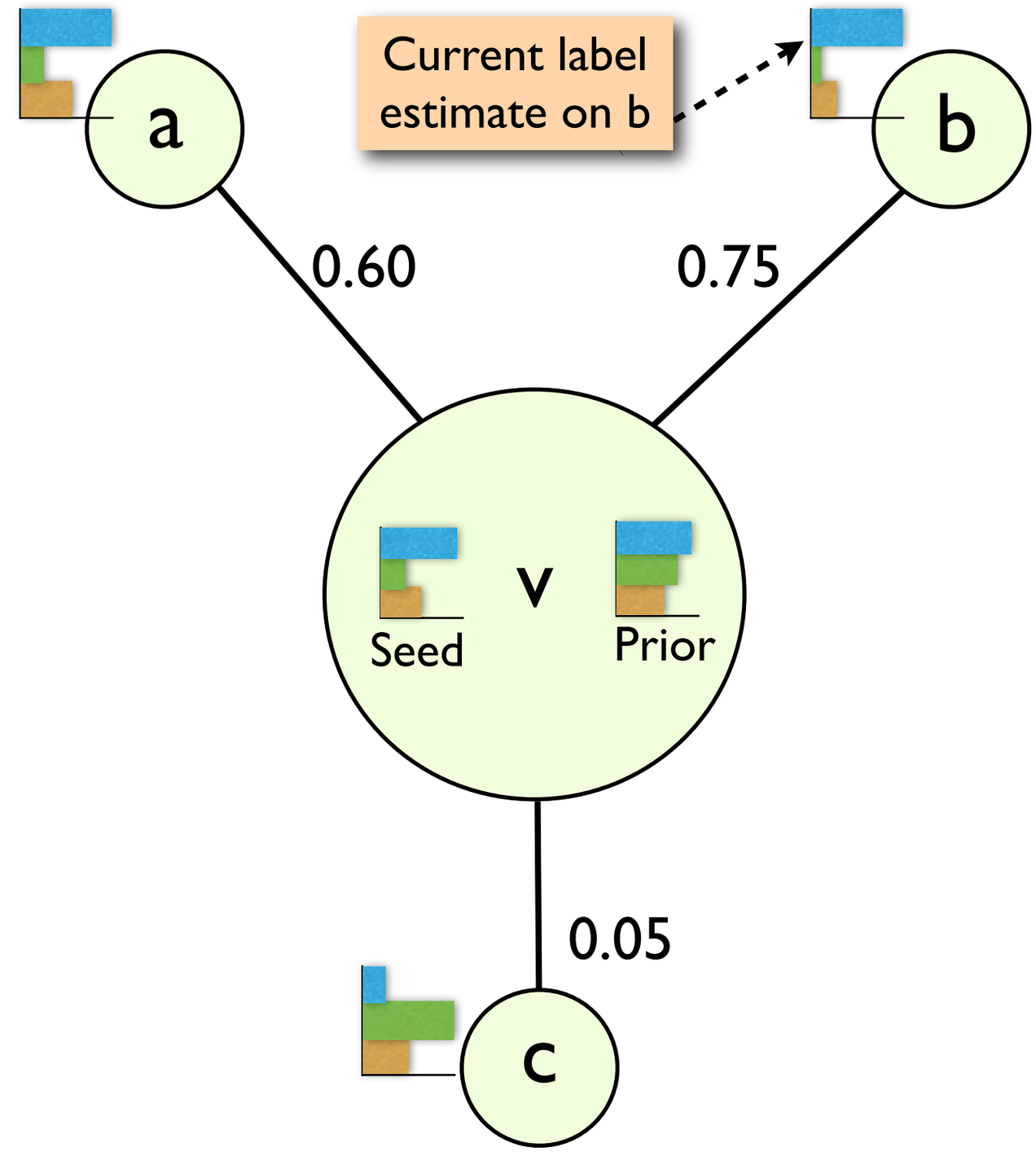
- Motivation
- Graph Construction
- Inference Methods
- Scalability
  - Scalability Issues
  - Node reordering
  - MapReduce Parallelization
- Applications
- Conclusion & Future Work

MapReduce Implementation of MAD



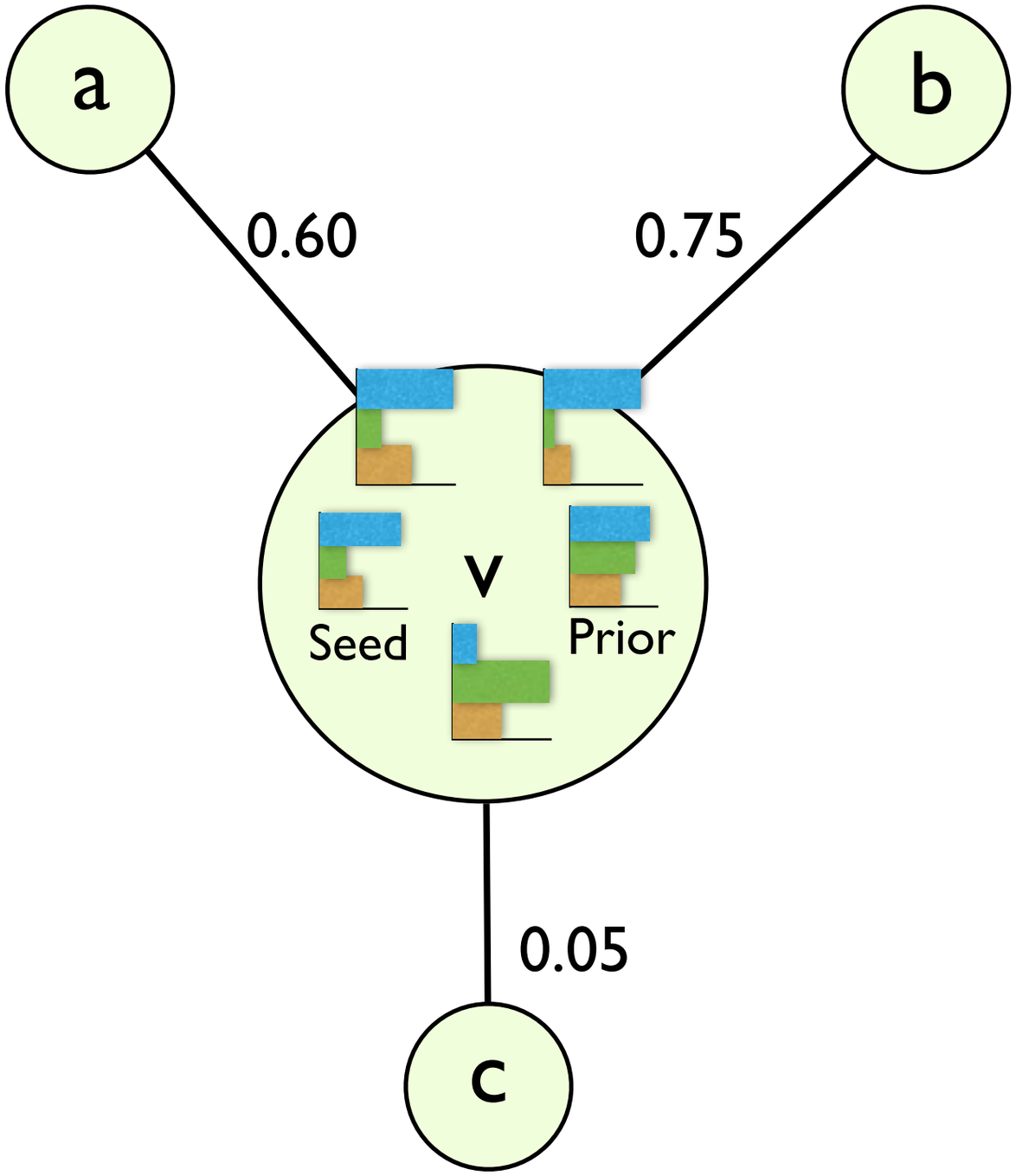
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors



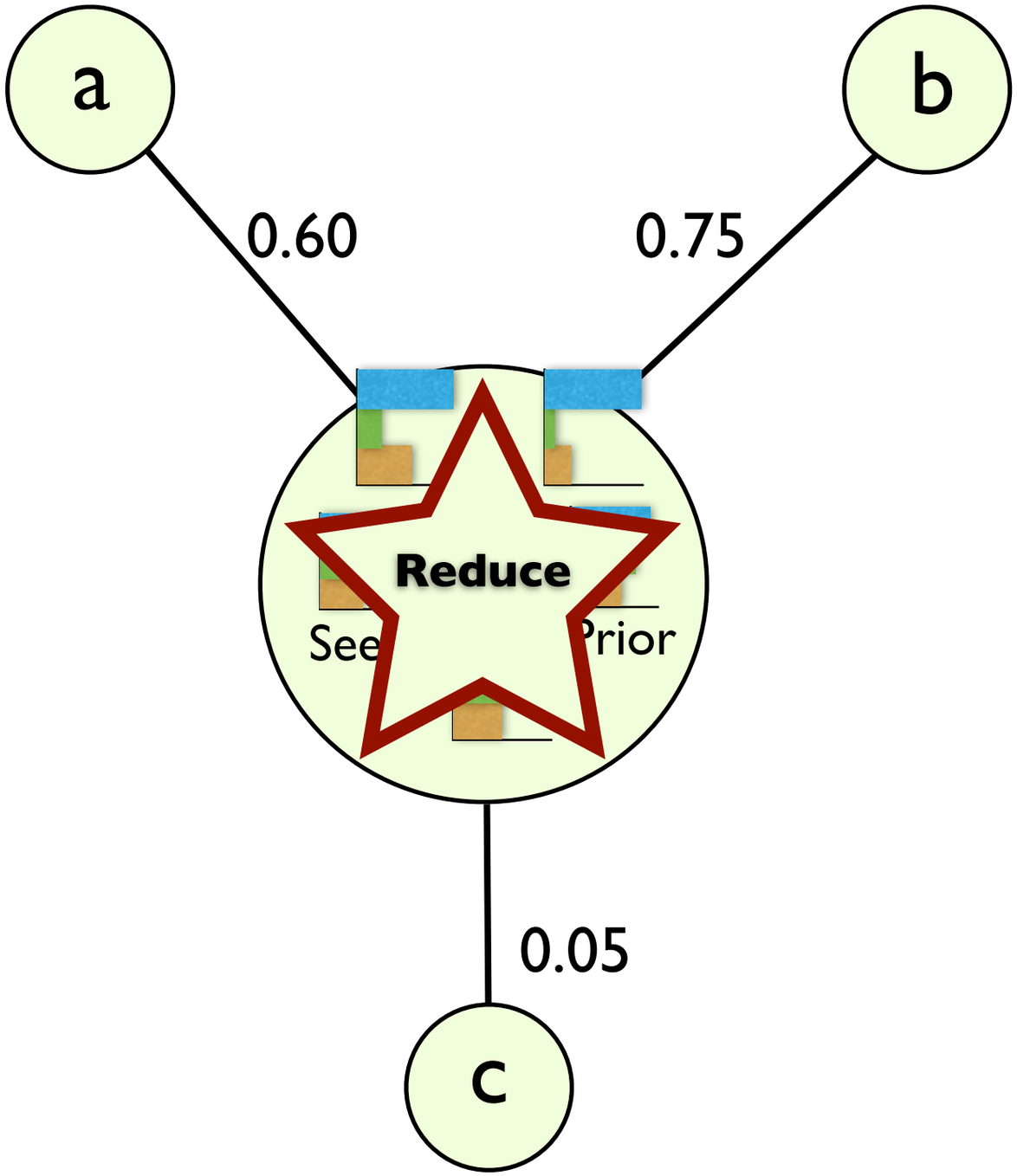
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors



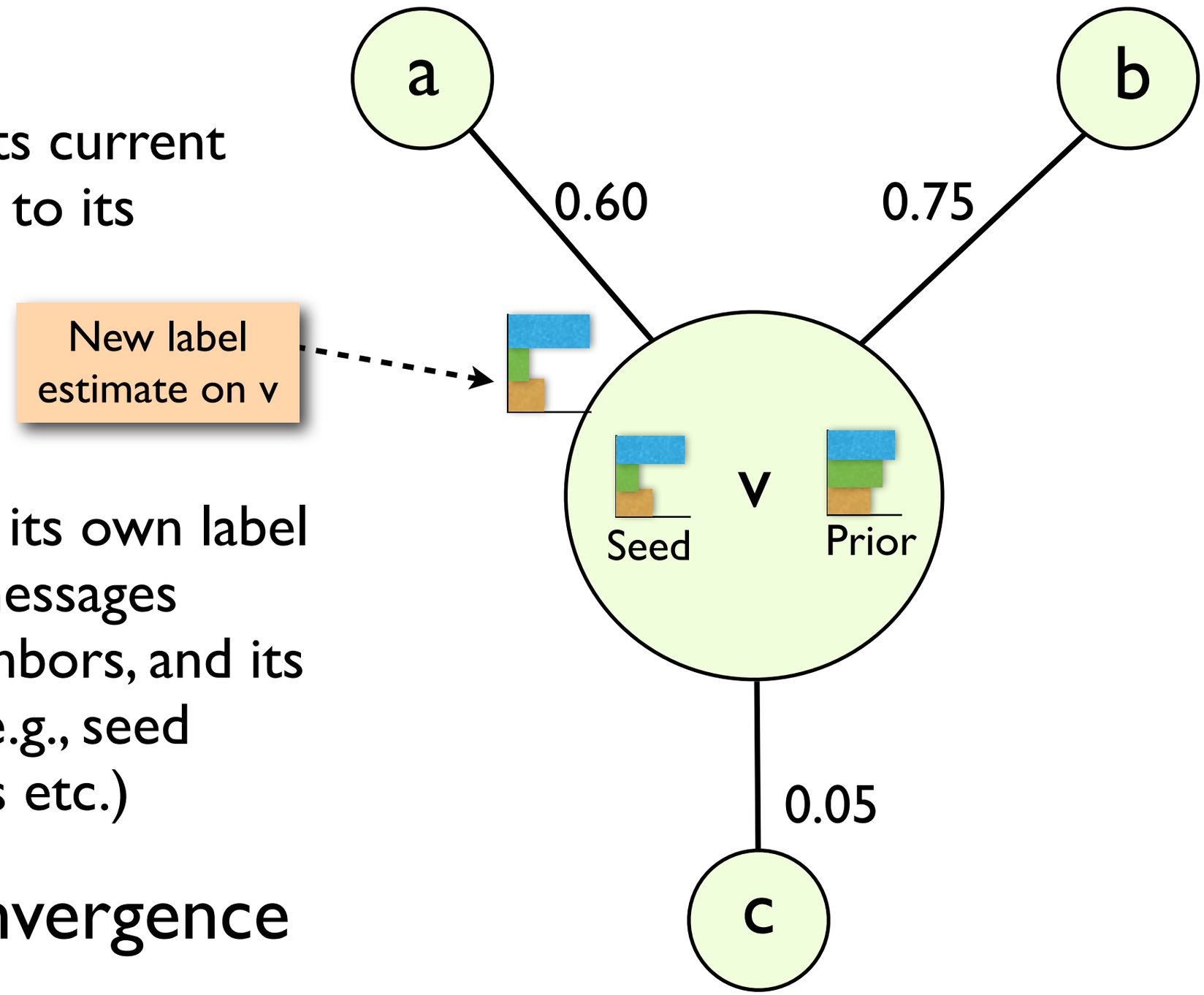
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)
- Repeat until convergence



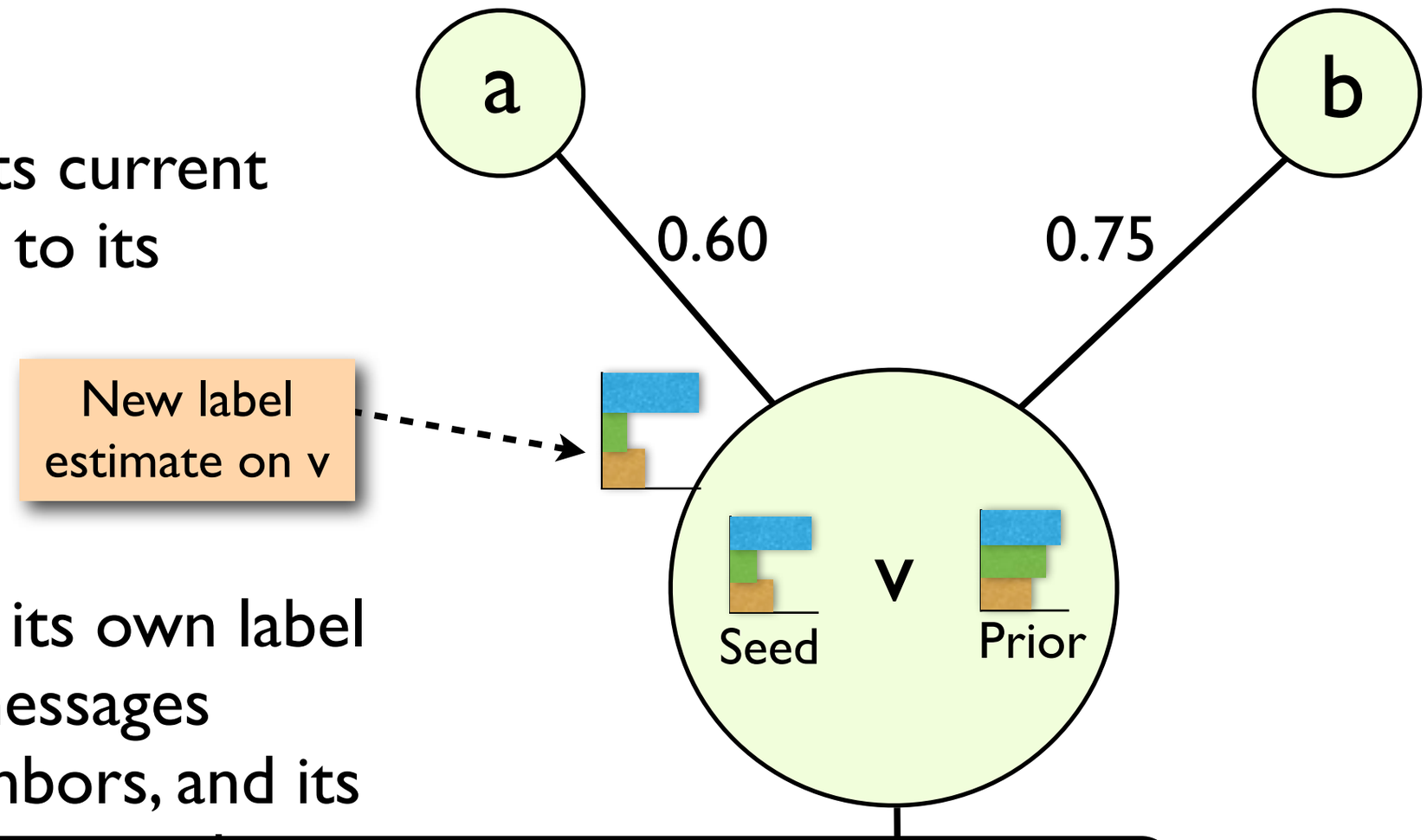
MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)
- Repeat until convergence



MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node updates its own label assignment using messages received from neighbors, and its own
    label
- Repe



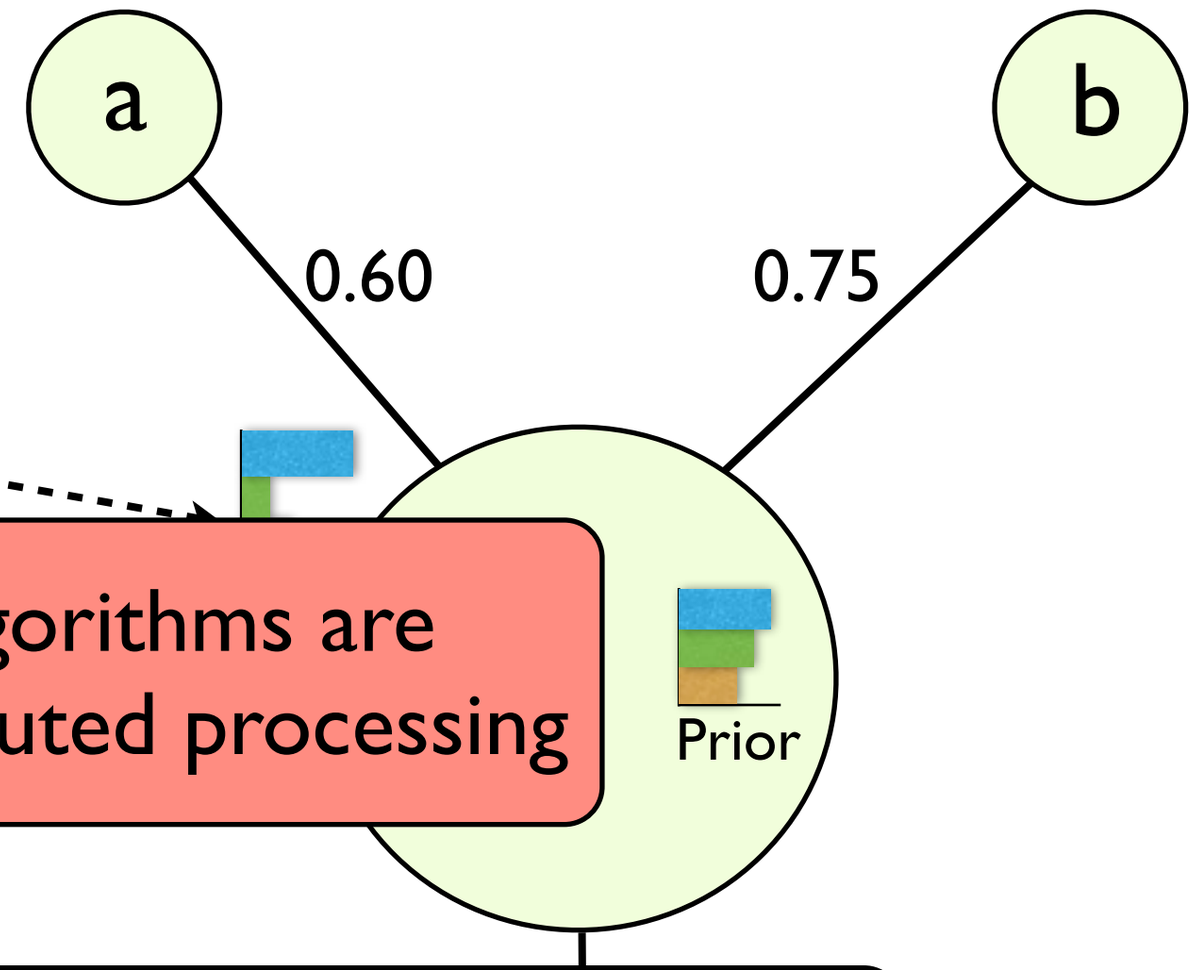
Code in Junto Label Propagation Toolkit

(includes Hadoop-based implementation)

http://code.google.com/p/junto/

MapReduce Implementation of MAD

- Map
  - Each node send its current label assignments to its neighbors
- Reduce
  - Each node
    assignment
    received from neighbors, and its
    own
    label
- Repe



New label

Graph-based algorithms are amenable to distributed processing

Code in Junto Label Propagation Toolkit

(includes Hadoop-based implementation)

http://code.google.com/p/junto/

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Phone Classification
  - Text Categorization
  - Dialog Act Tagging
  - Statistical Machine Translation
  - POS Tagging
  - MultiLingual POS Tagging
- Conclusion & Future Work

Problem Description & Motivation

Problem Description & Motivation

- Given a “frame” of speech classify it into one of n phones

Problem Description & Motivation

- Given a “frame” of speech classify it into one of n phones
- Training supervised models requires large amounts of labeled data (phone classification in resource-scarce languages)

TIMIT

TIMIT

- Corpus of read speech

TIMIT

- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English

TIMIT

- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English
- Each speaker has read 10 sentences

TIMIT

- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English
- Each speaker has read 10 sentences
- Includes time-aligned phonetic transcriptions

TIMIT

- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English
- Each speaker has read 10 sentences
- Includes time-aligned phonetic transcriptions
- Phone set has 61 phones [Lee & Hon, 89]

TIMIT

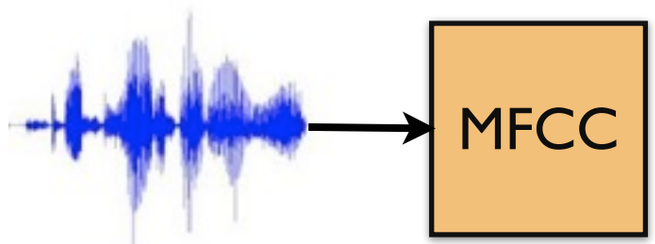
- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English
- Each speaker has read 10 sentences
- Includes time-aligned phonetic transcriptions
- Phone set has 61 phones [Lee & Hon, 89]
  - mapped down to 48 phones for modeling

TIMIT

- Corpus of read speech
- Broadband recordings of 630 speakers of 8 major dialects of American English
- Each speaker has read 10 sentences
- Includes time-aligned phonetic transcriptions
- Phone set has 61 phones [Lee & Hon, 89]
  - mapped down to 48 phones for modeling
  - further mapped down to 39 phones for scoring

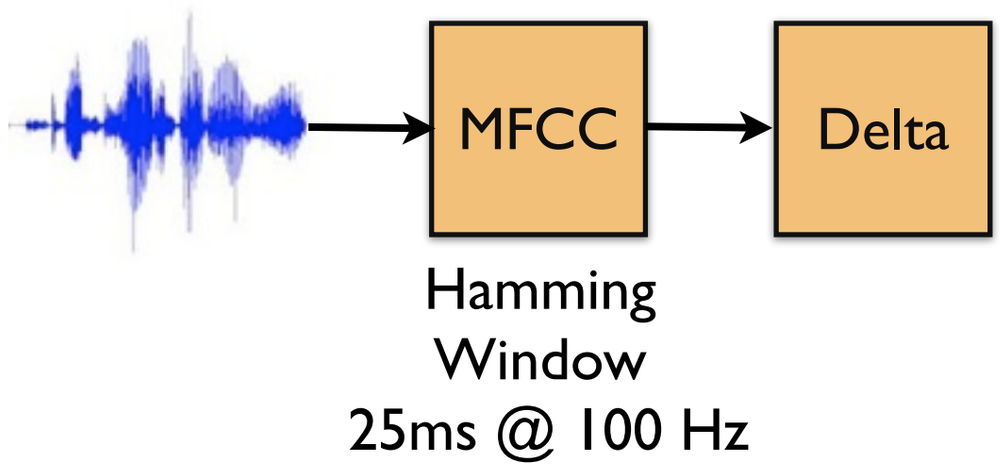
Feature Extraction

Feature Extraction

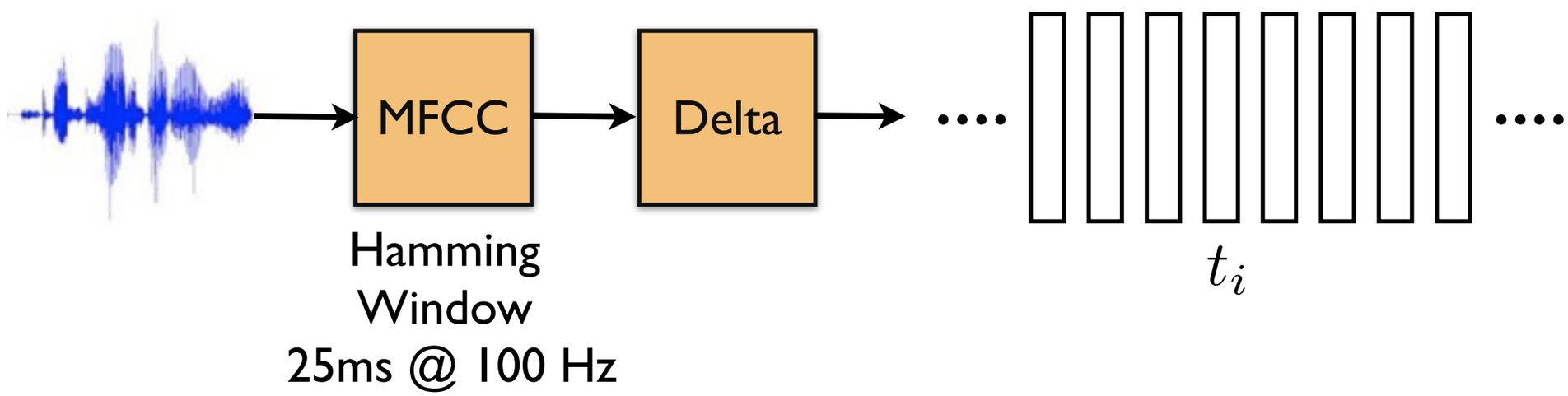


Hamming
Window
25ms @ 100 Hz

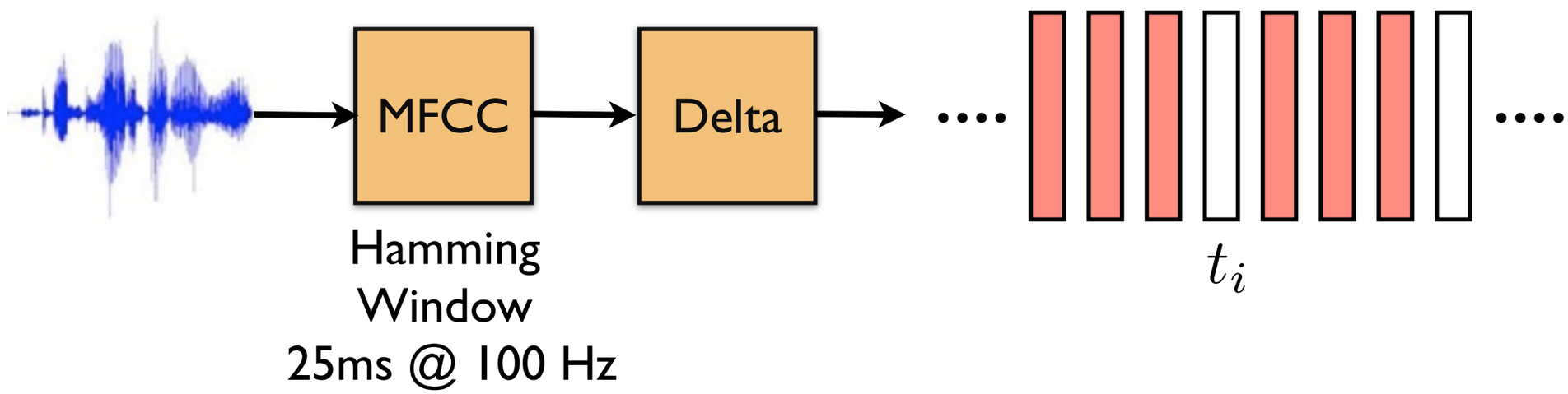
Feature Extraction



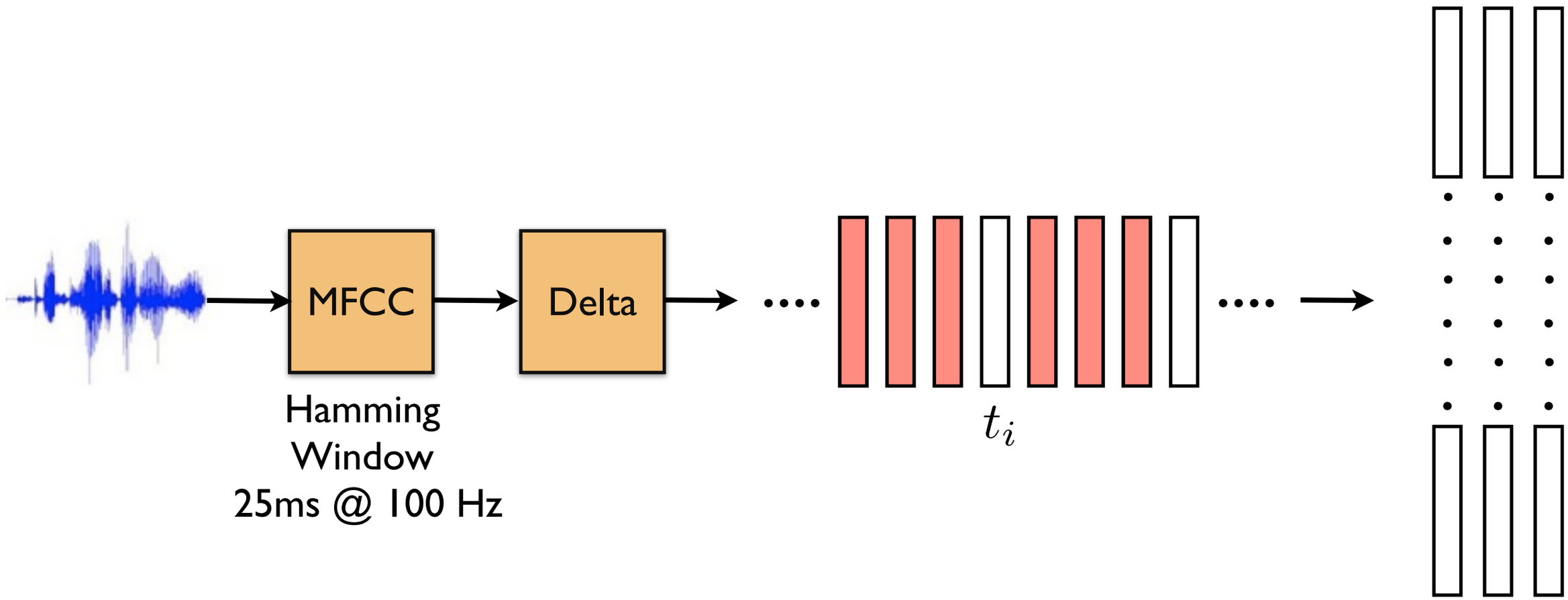
Feature Extraction



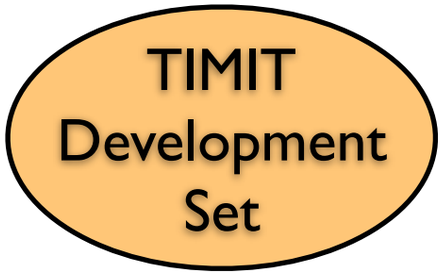
Feature Extraction



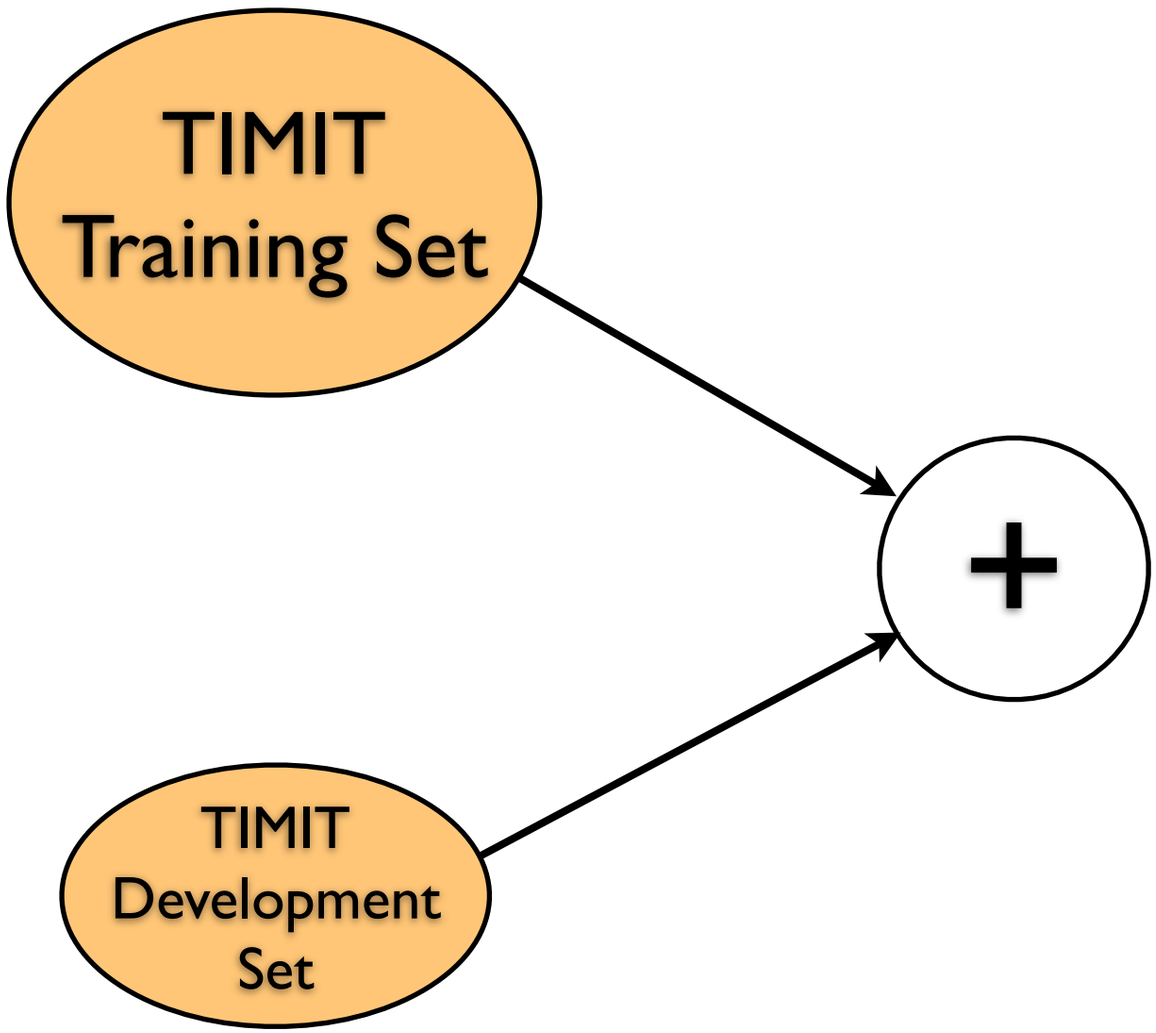
Feature Extraction



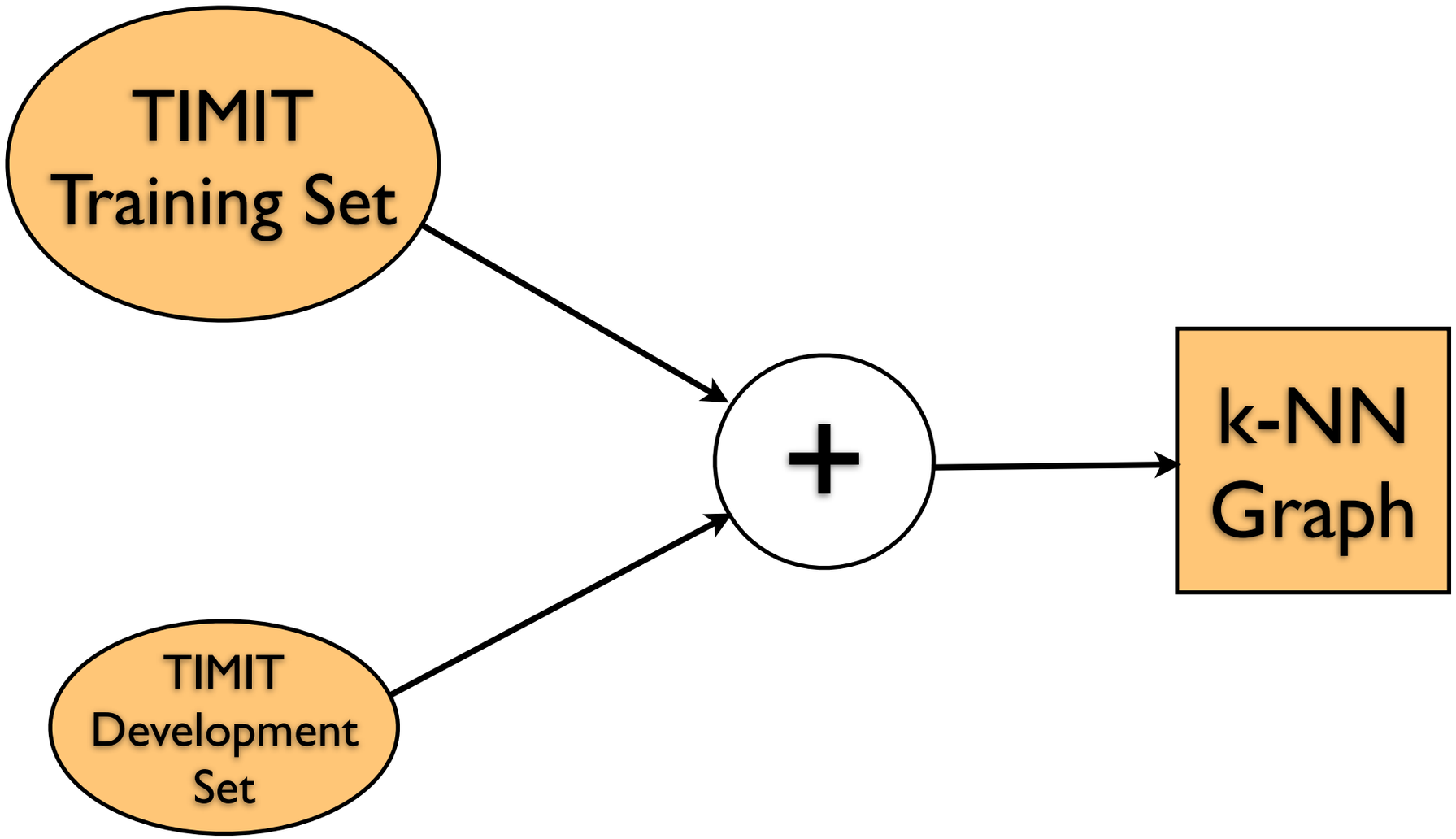
Graph Construction



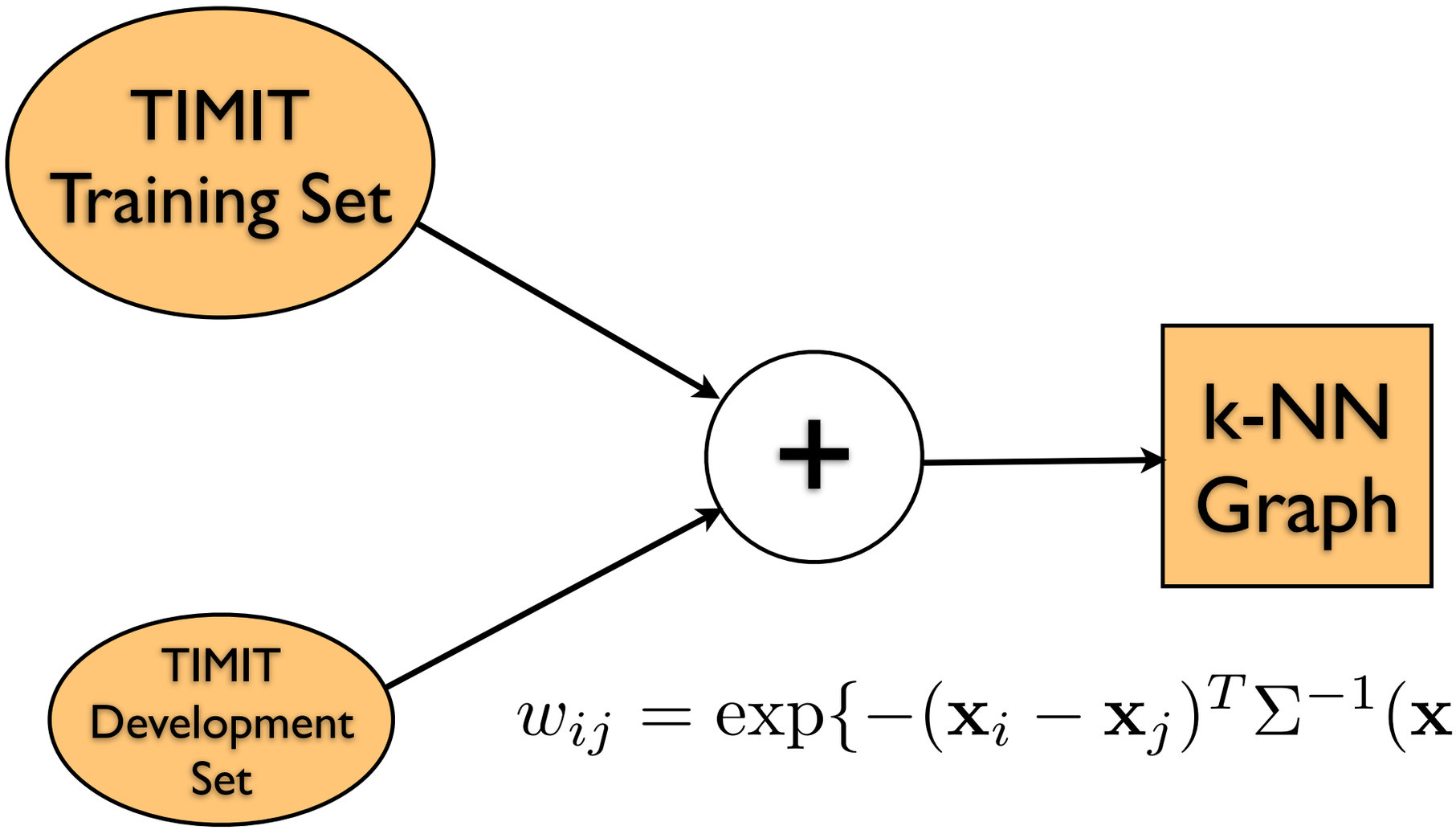
Graph Construction



Graph Construction

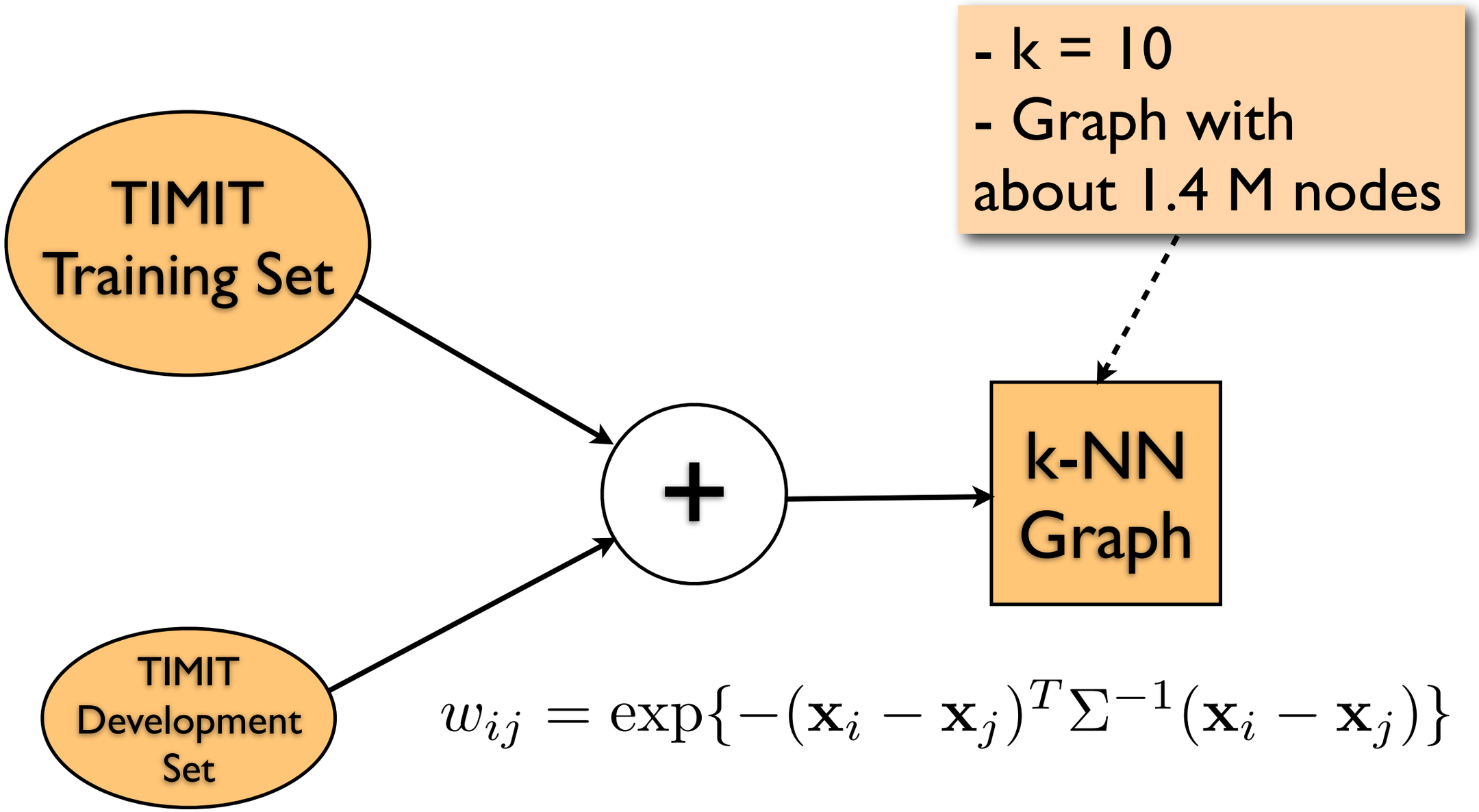


Graph Construction



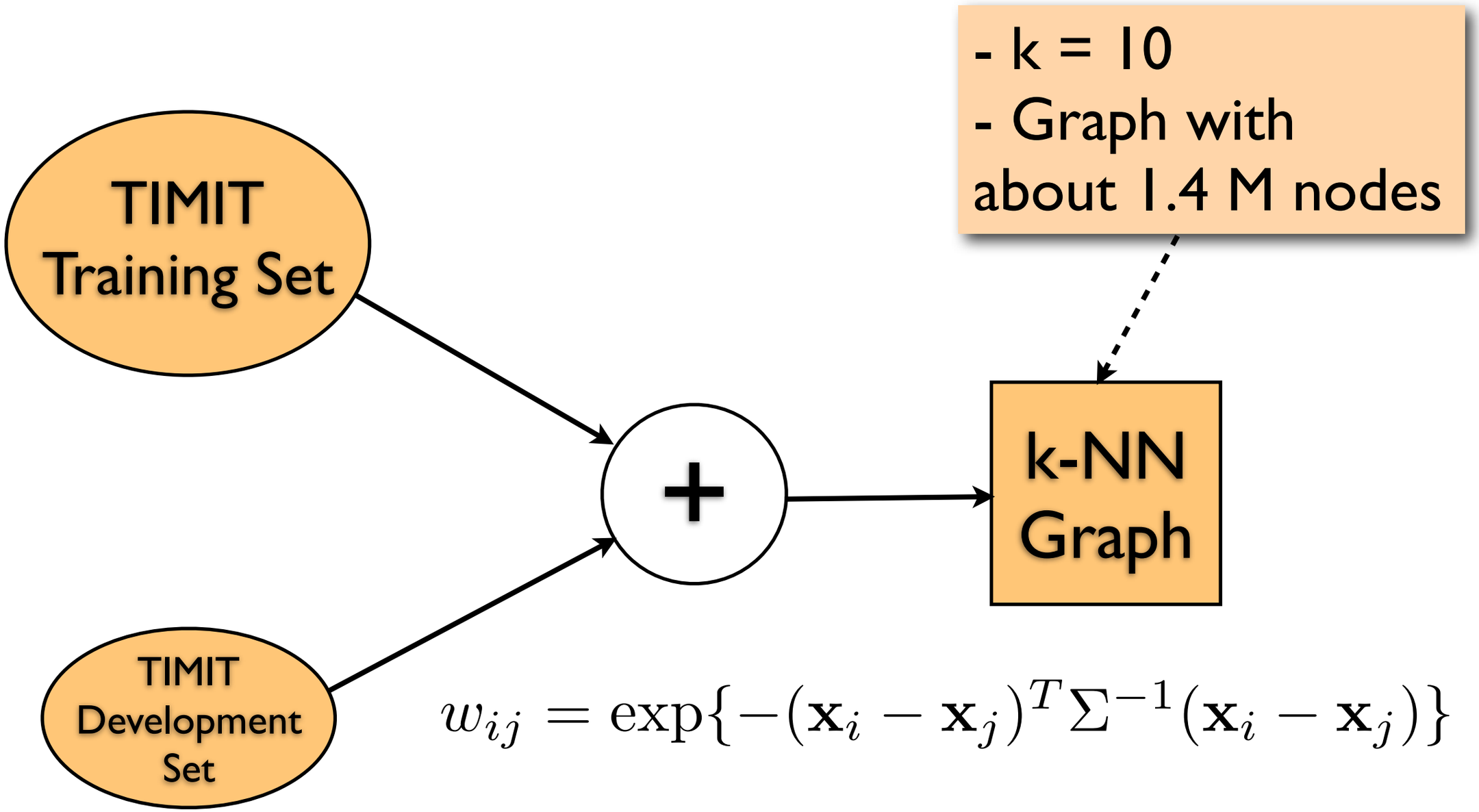
$$w_{ij} = \exp\{-(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)\}$$

Graph Construction



$$w_{ij} = \exp\{-(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)\}$$

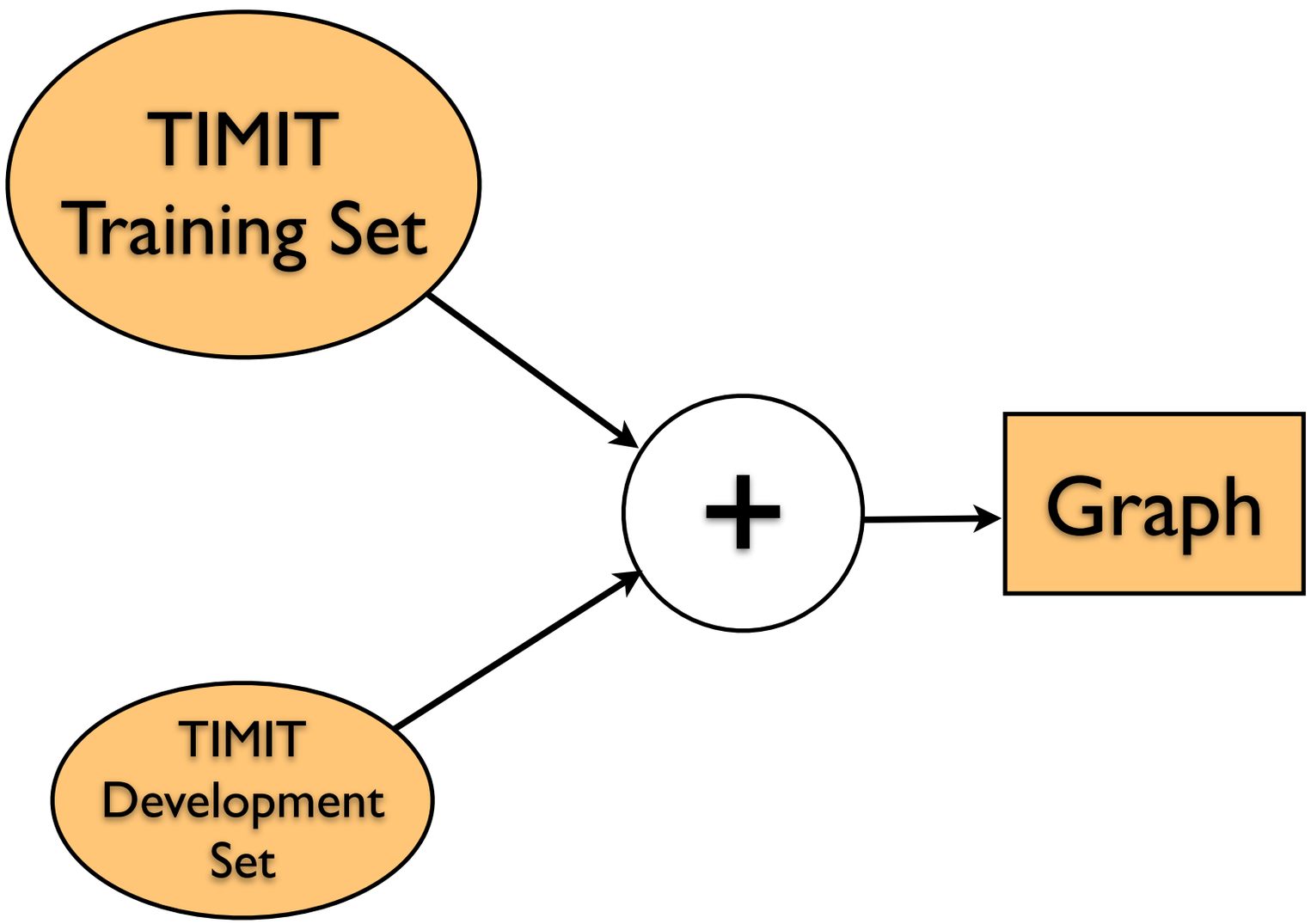
Graph Construction



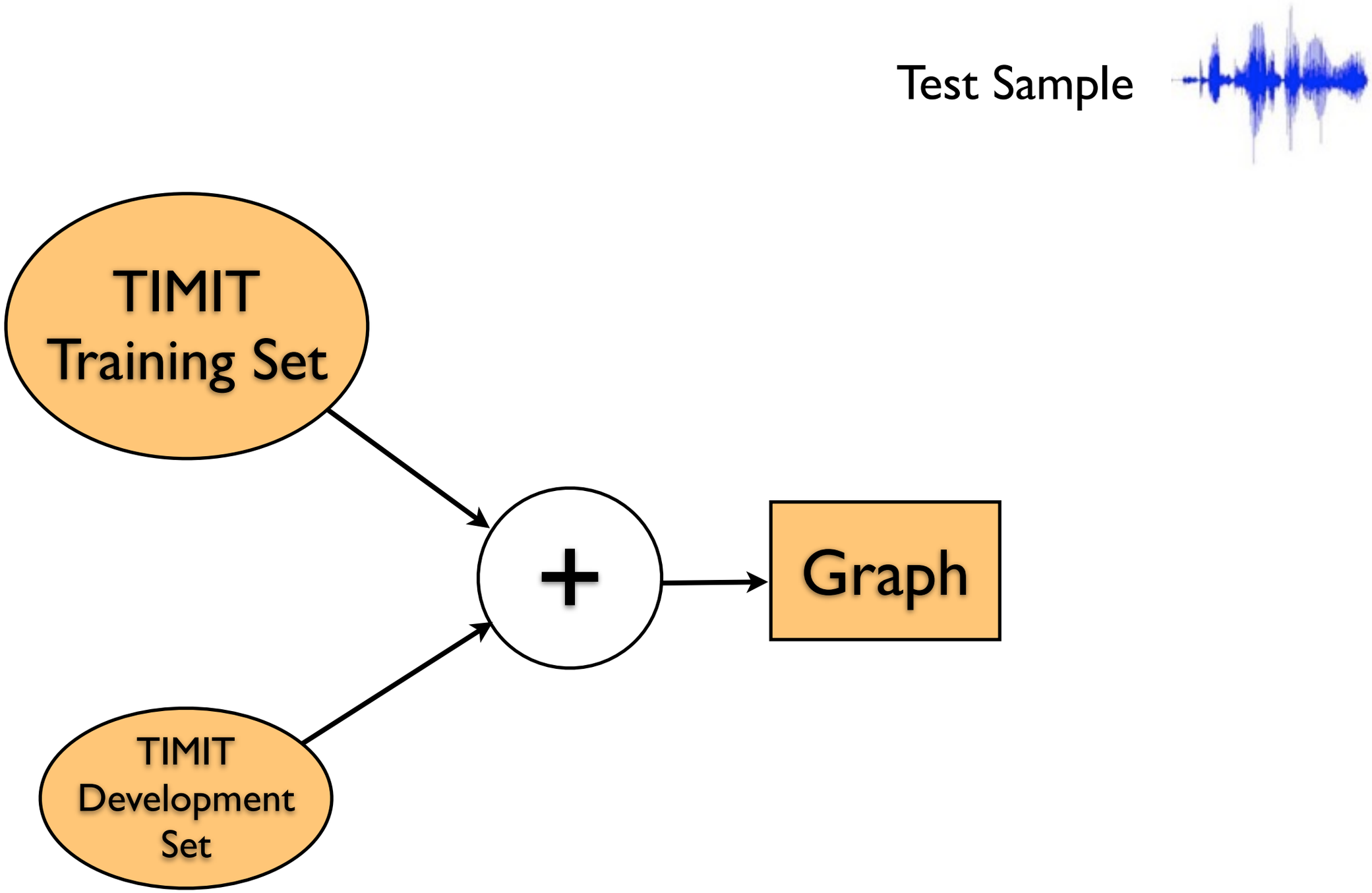
$$w_{ij} = \exp\{-(\mathbf{x}_i - \mathbf{x}_j)^T \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)\}$$

Labels not used during graph construction

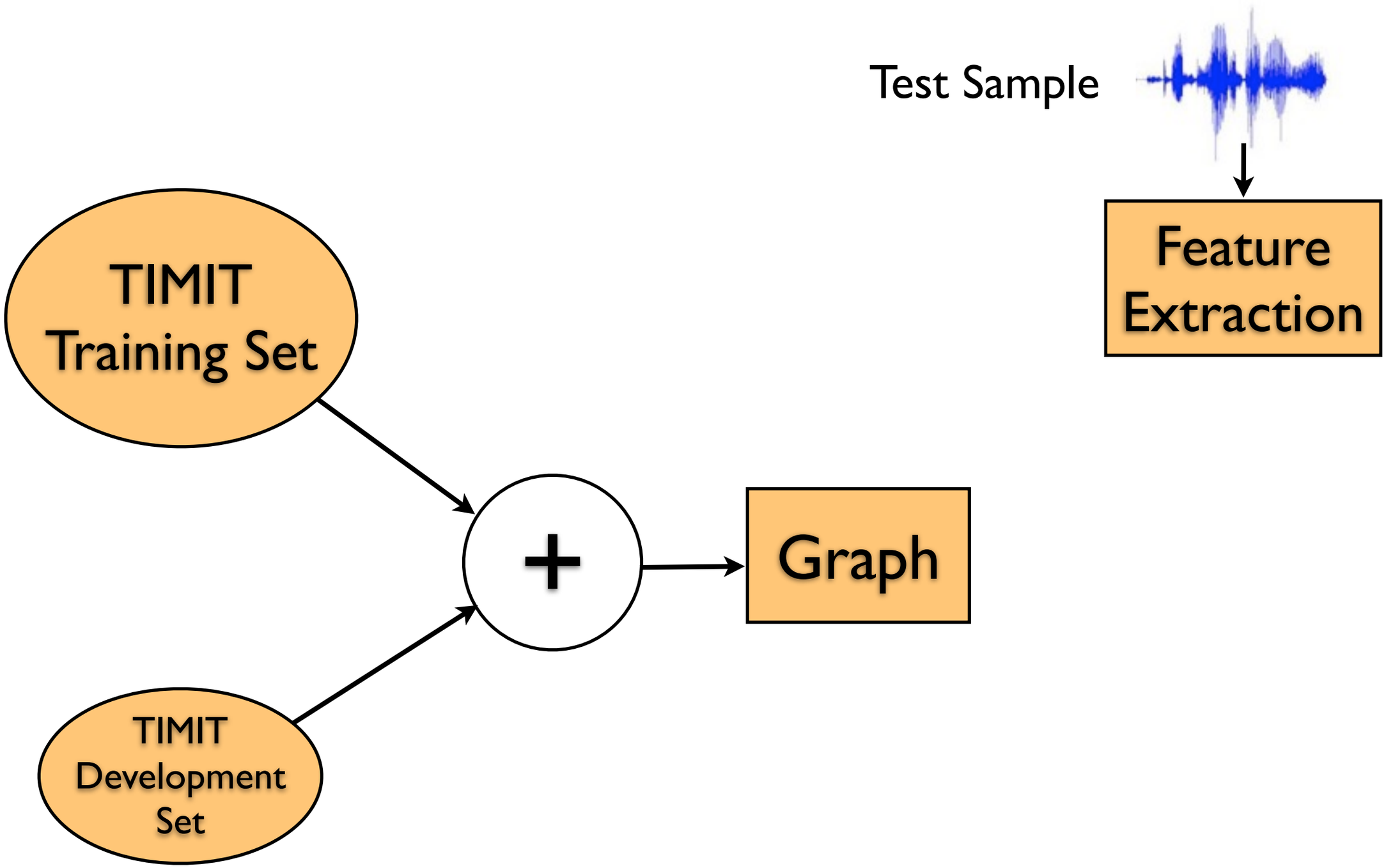
Inductive Extension



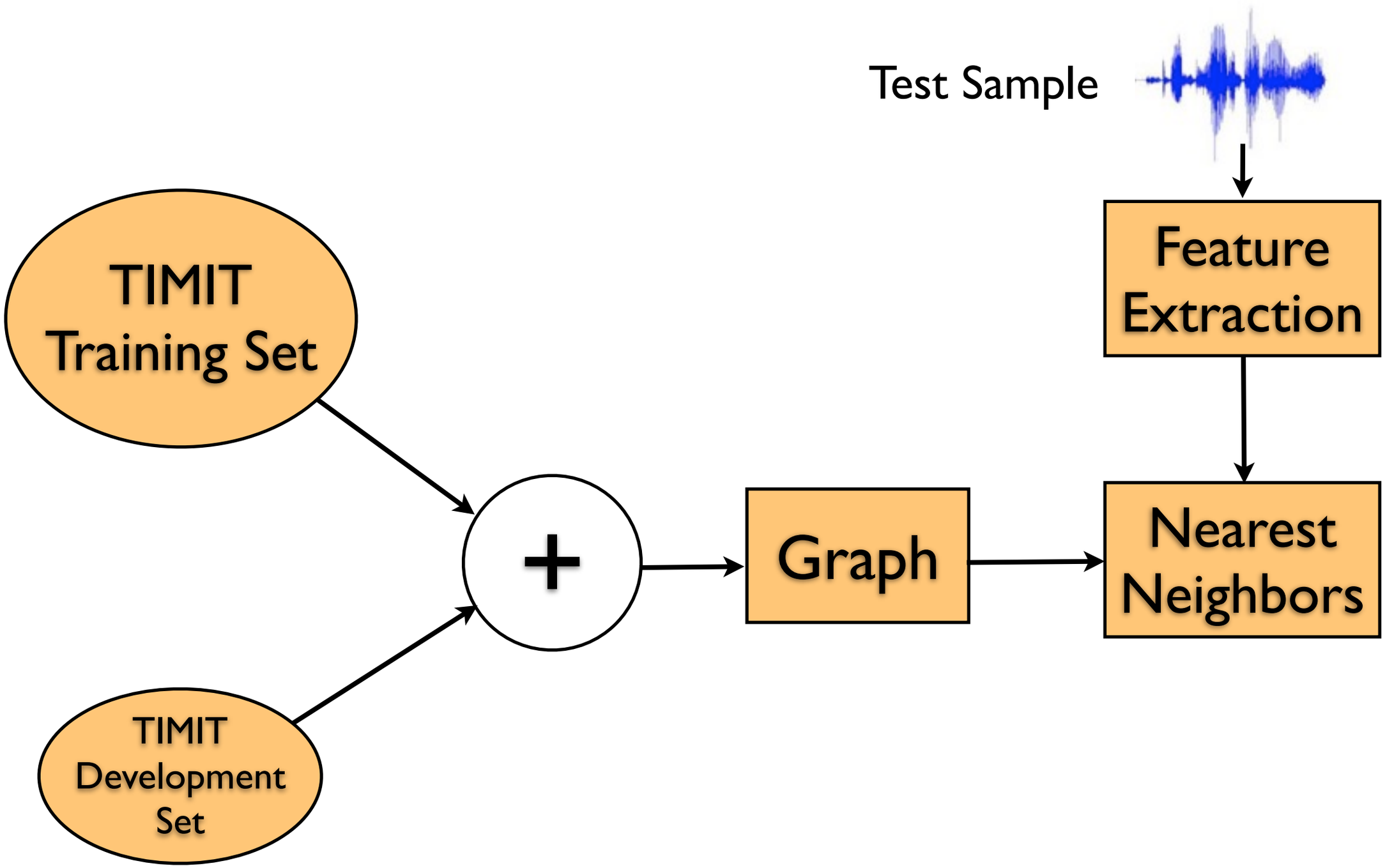
Inductive Extension



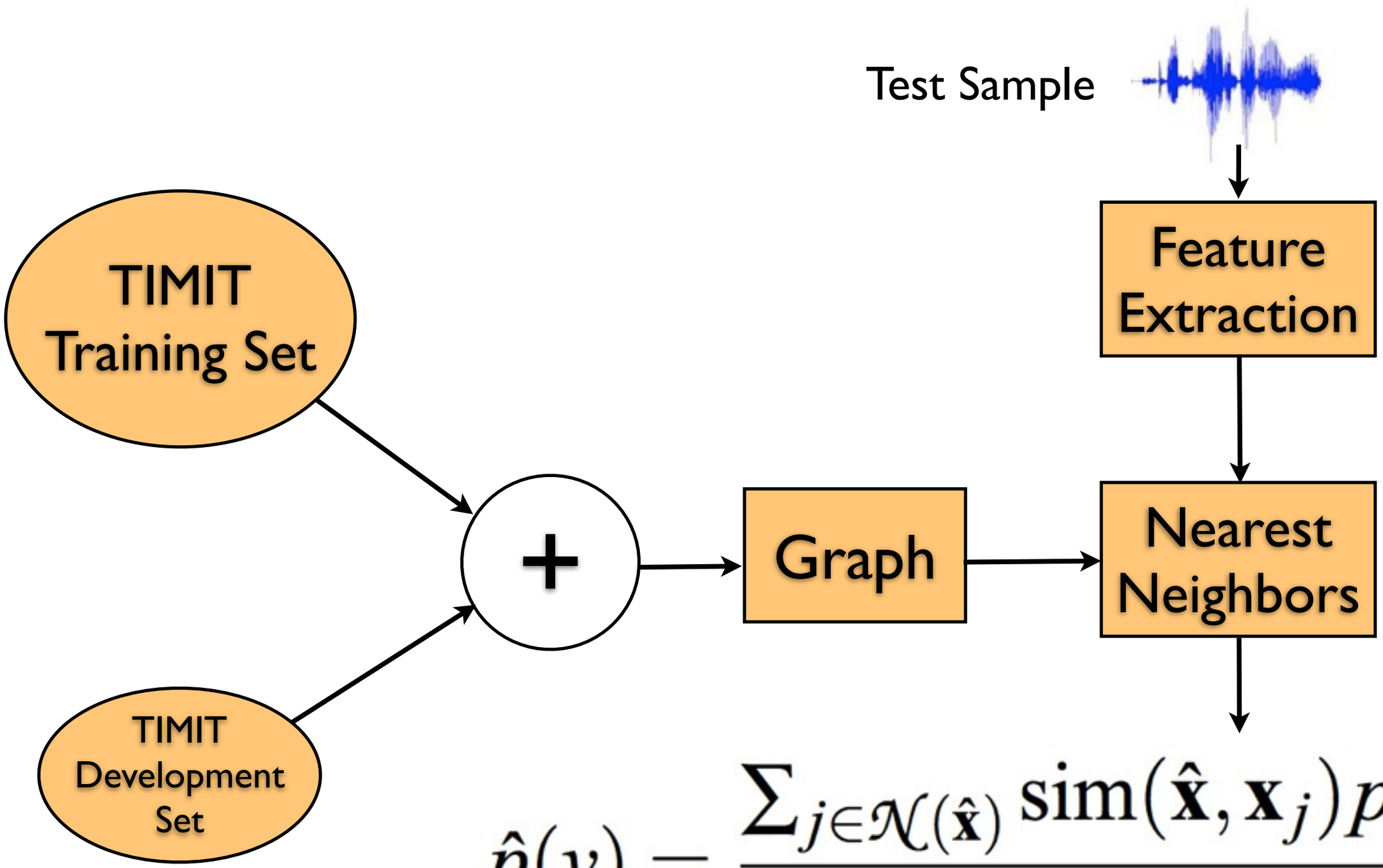
Inductive Extension



Inductive Extension

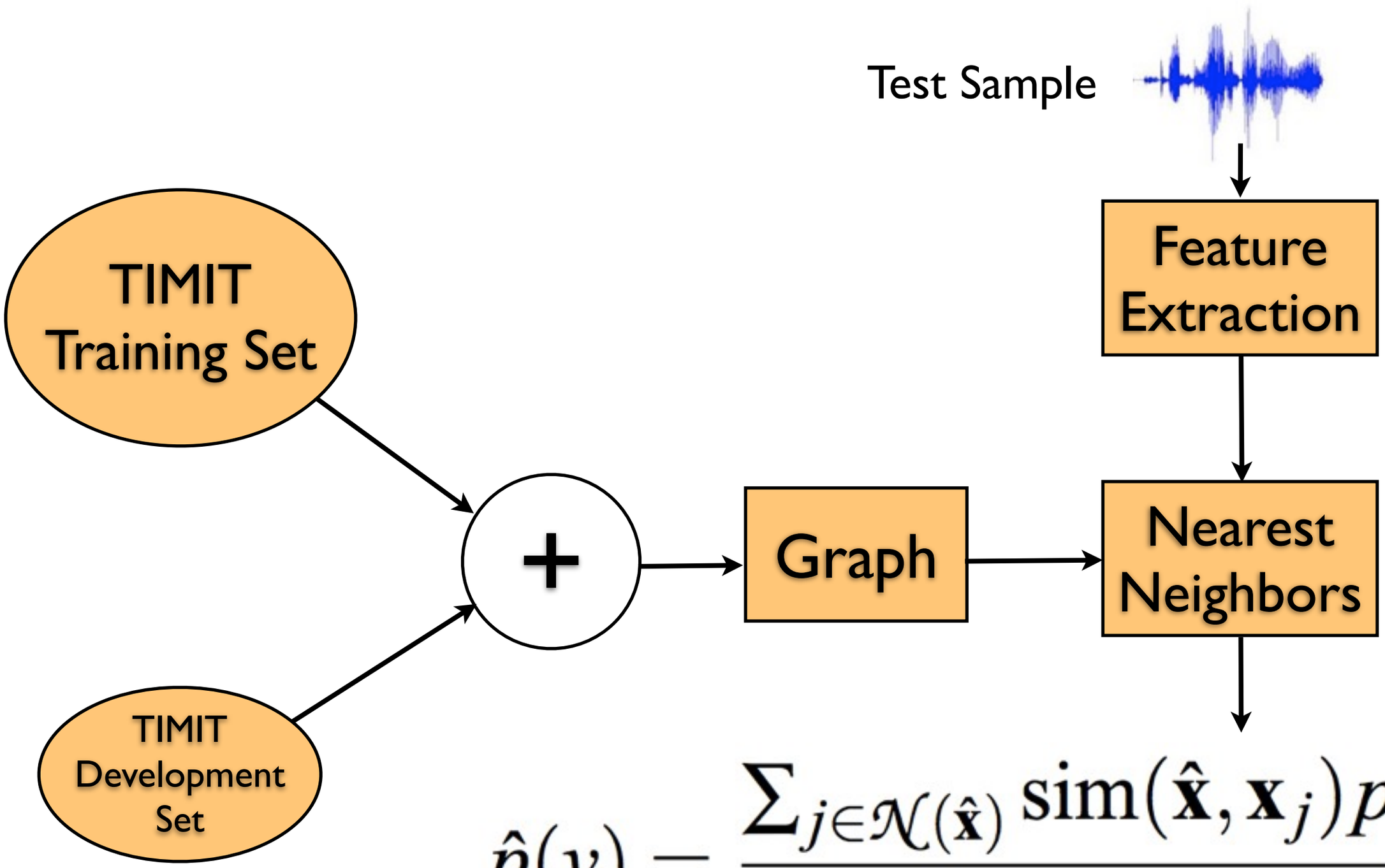


Inductive Extension



$$\hat{p}(y) = \frac{\sum_{j \in \mathcal{N}(\hat{\mathbf{x}})} \mathrm{sim}(\hat{\mathbf{x}}, \mathbf{x}_j) p_j^*(y)}{\sum_{j \in \mathcal{N}(\hat{\mathbf{x}})} \mathrm{sim}(\hat{\mathbf{x}}, \mathbf{x}_j)}$$

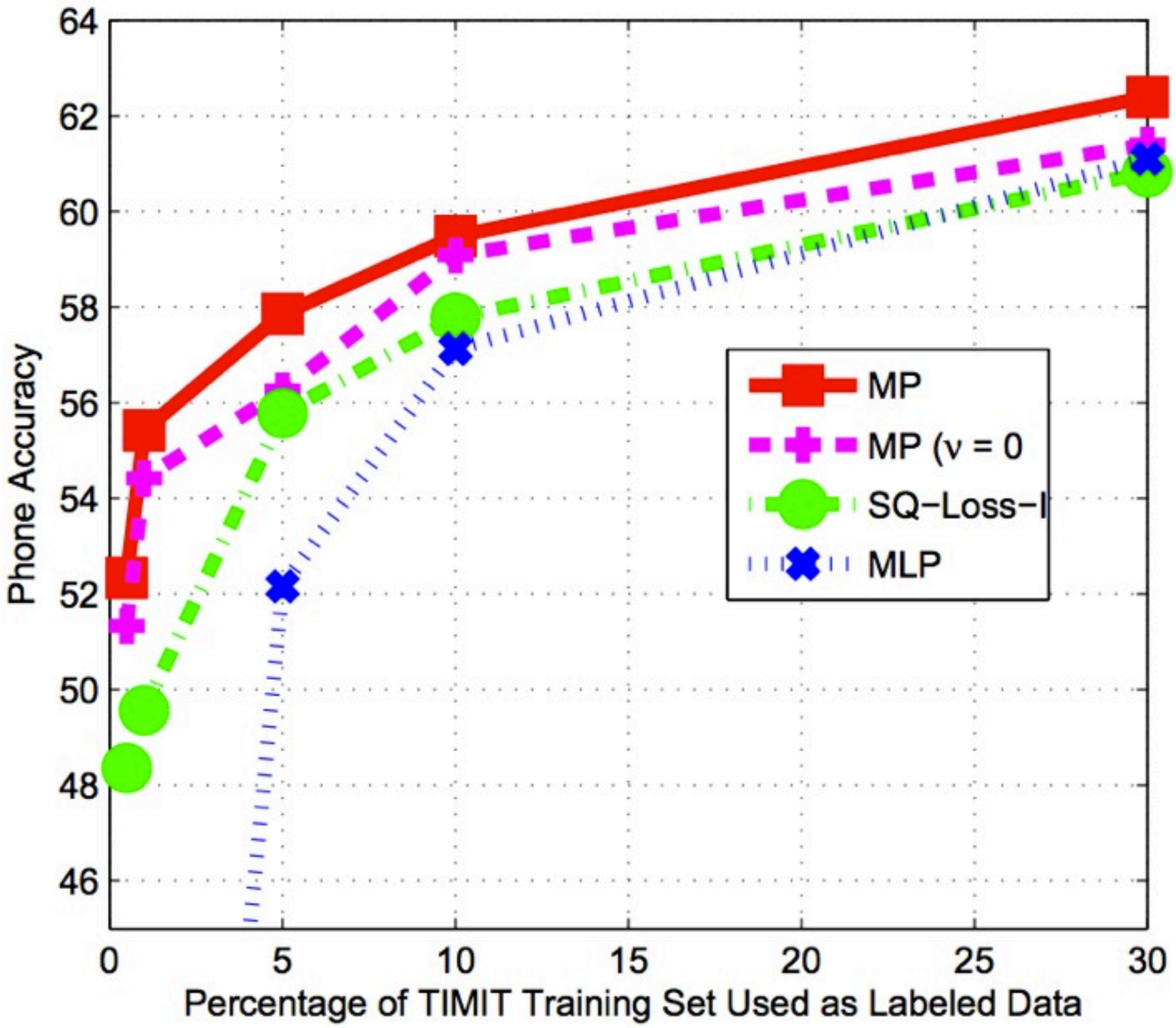
Inductive Extension



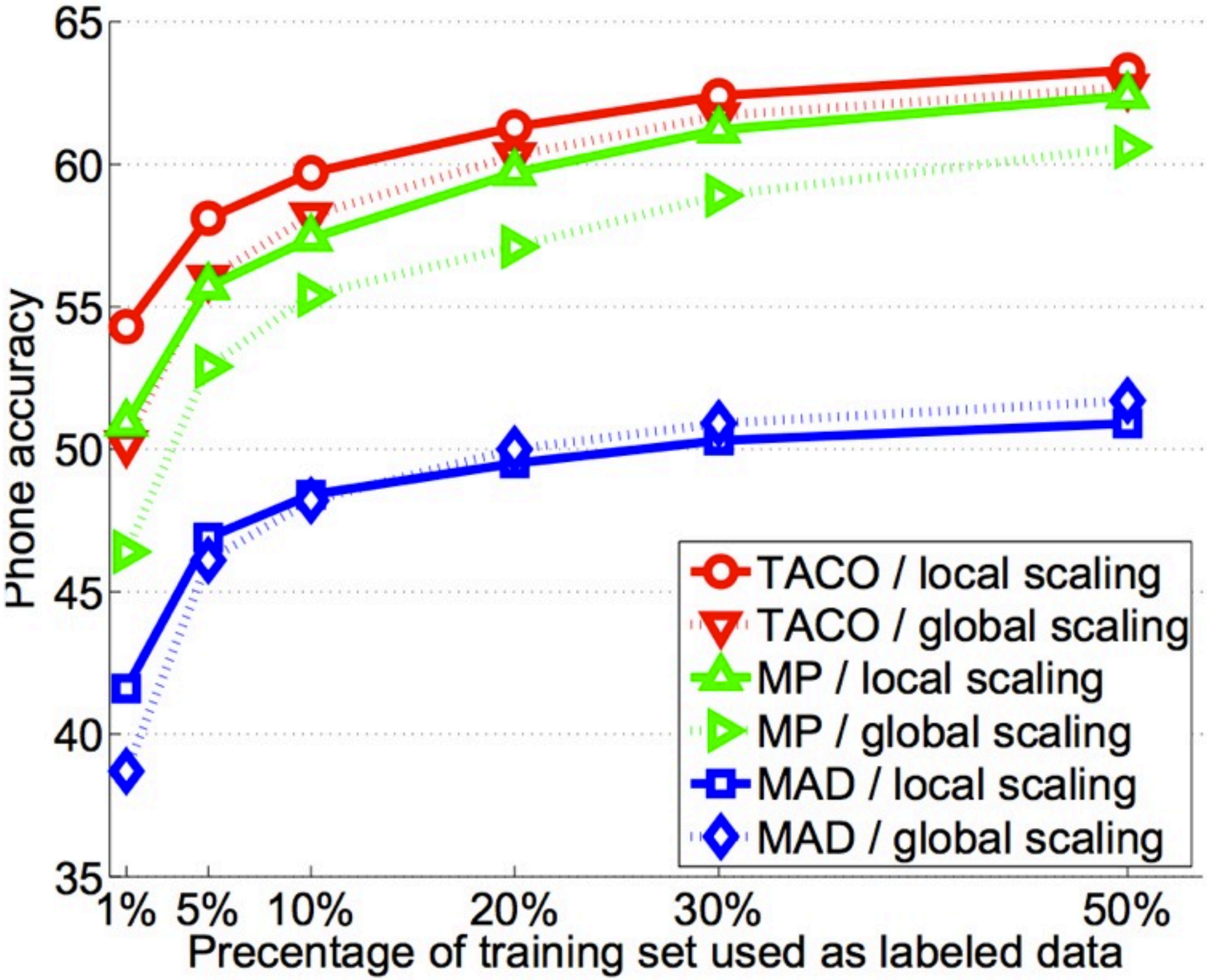
$$\hat{p}(y) = \frac{\sum_{j \in \mathcal{N}(\hat{\mathbf{x}})} \mathrm{sim}(\hat{\mathbf{x}}, \mathbf{x}_j) p_j^*(y)}{\sum_{j \in \mathcal{N}(\hat{\mathbf{x}})} \mathrm{sim}(\hat{\mathbf{x}}, \mathbf{x}_j)}$$

Nadaraya-Watson Estimator

Results (I)


64
62
60
58
56
54
52
50
48
46
Phone Accuracy
0
5
10
15
20
25
30
Percentage of TIMIT Training Set Used as Labeled Data
MP
MP (ν = 0
SQ−Loss−I
MLP

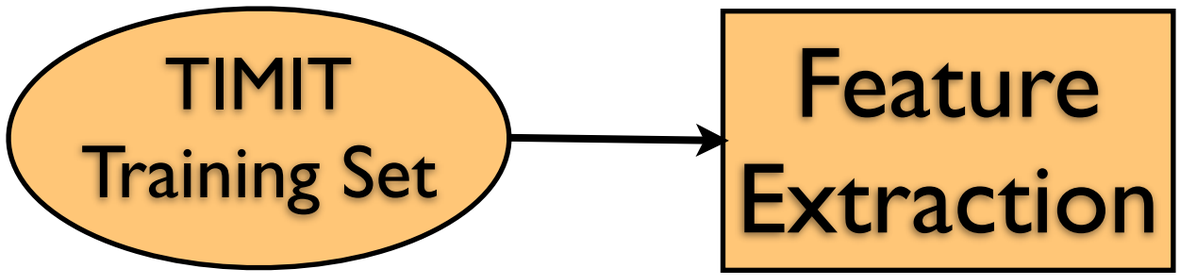
Results (II)


65
60
55
50
45
40
35
Phone accuracy
1%
5%
10%
20%
30%
50%
Precentage of training set used as labeled data
TACO / local scaling
TACO / global scaling
MP / local scaling
MP / global scaling
MAD / local scaling
MAD / global scaling

“Labeled” Graph Construction

[Alexandrescu & Kirchoff, ASRU 2007]

“Labeled” Graph Construction

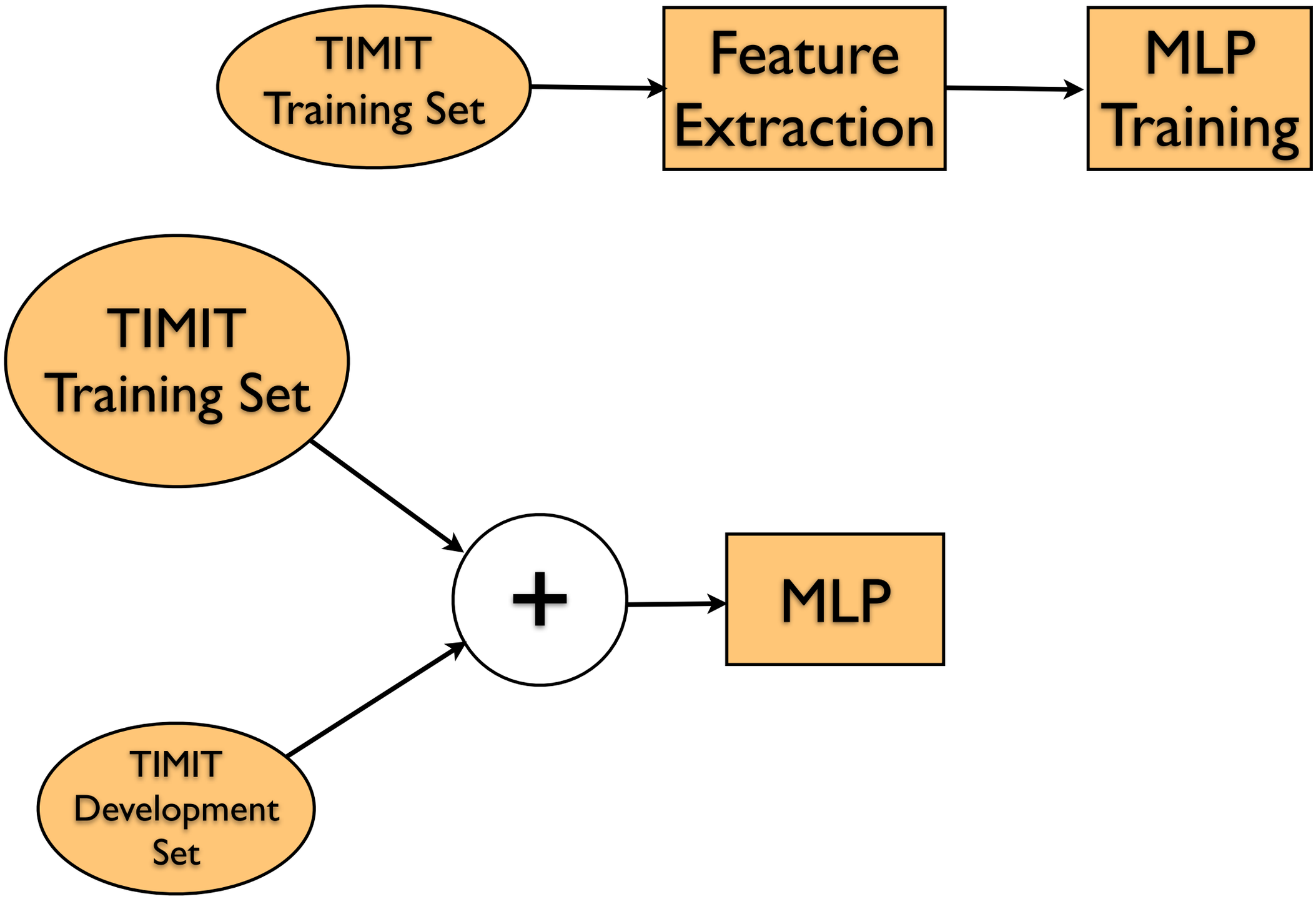


[Alexandrescu & Kirchoff, ASRU 2007]

“Labeled” Graph Construction

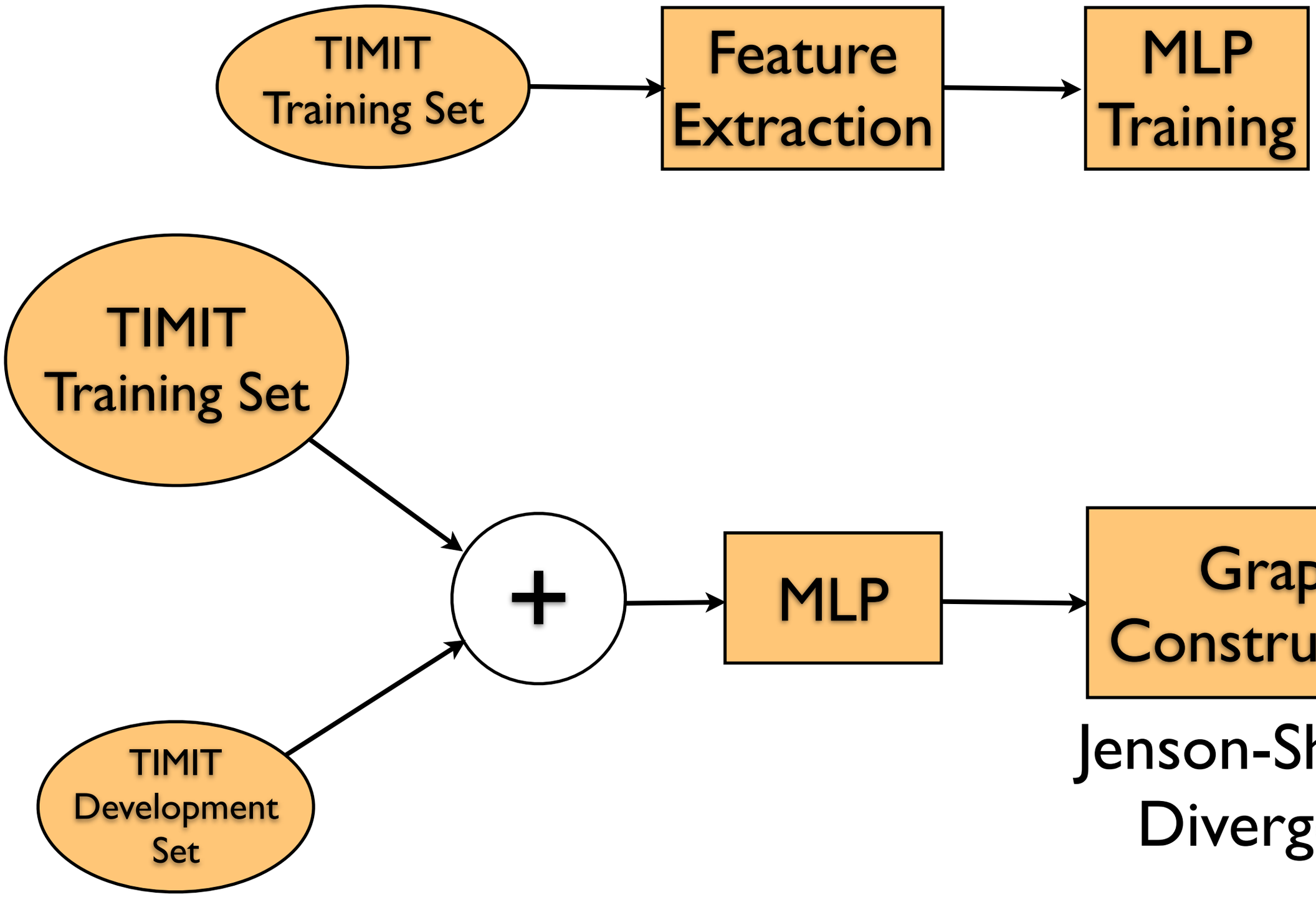
[Alexandrescu & Kirchoff, ASRU 2007]

"Labeled" Graph Construction



[Alexandrescu & Kirchoff, ASRU 2007]

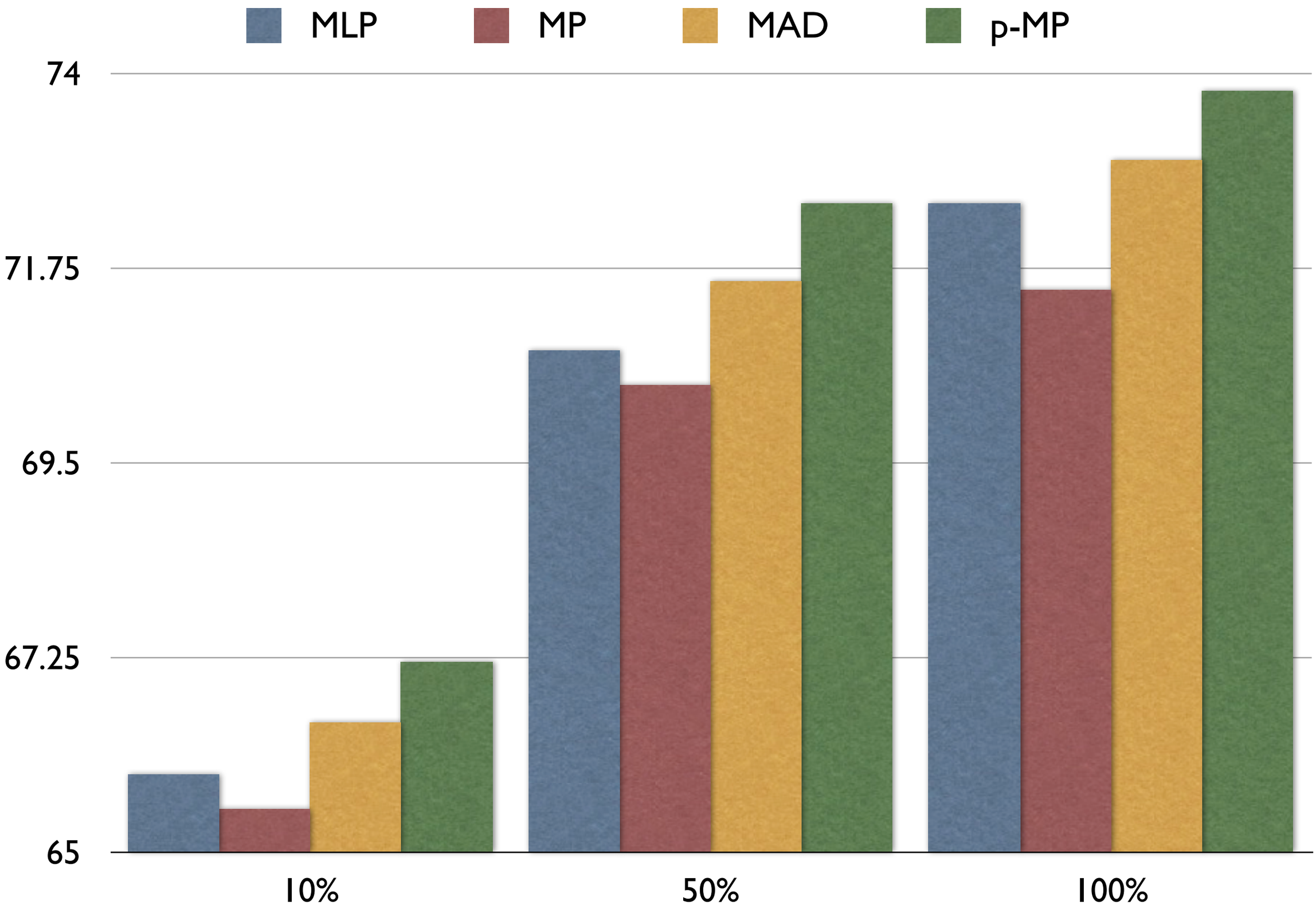
"Labeled" Graph Construction



Jenson-Shannon
Divergence

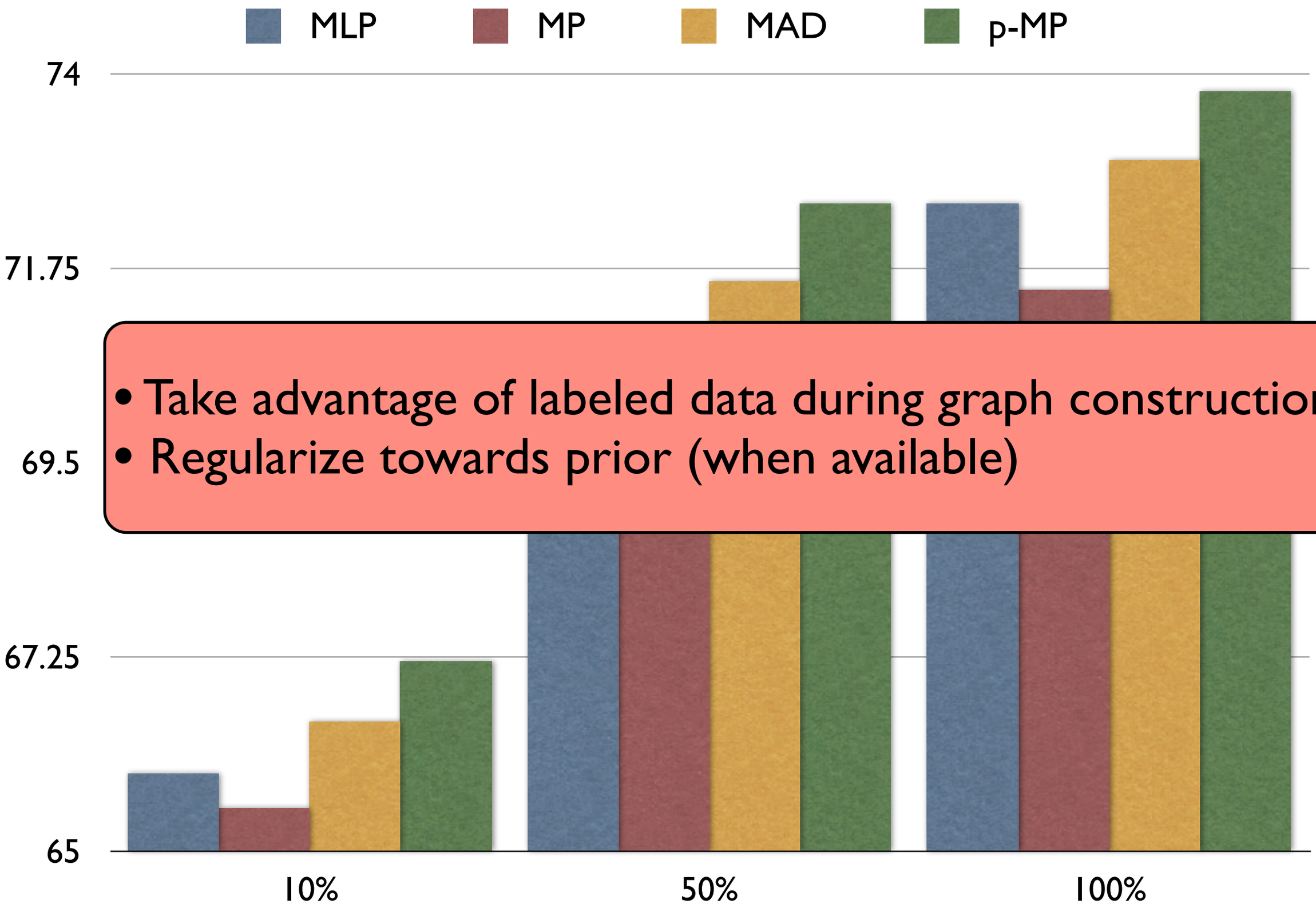
[Alexandrescu & Kirchoff, ASRU 2007]

Results (III)



[Liu & Kirchoff, UW-EE TR 2012]

Results (III)



- Take advantage of labeled data during graph construction
- Regularize towards prior (when available)

[Liu & Kirchoff, UW-EE TR 2012]

Switchboard Phonetic Annotation

Switchboard Phonetic Annotation

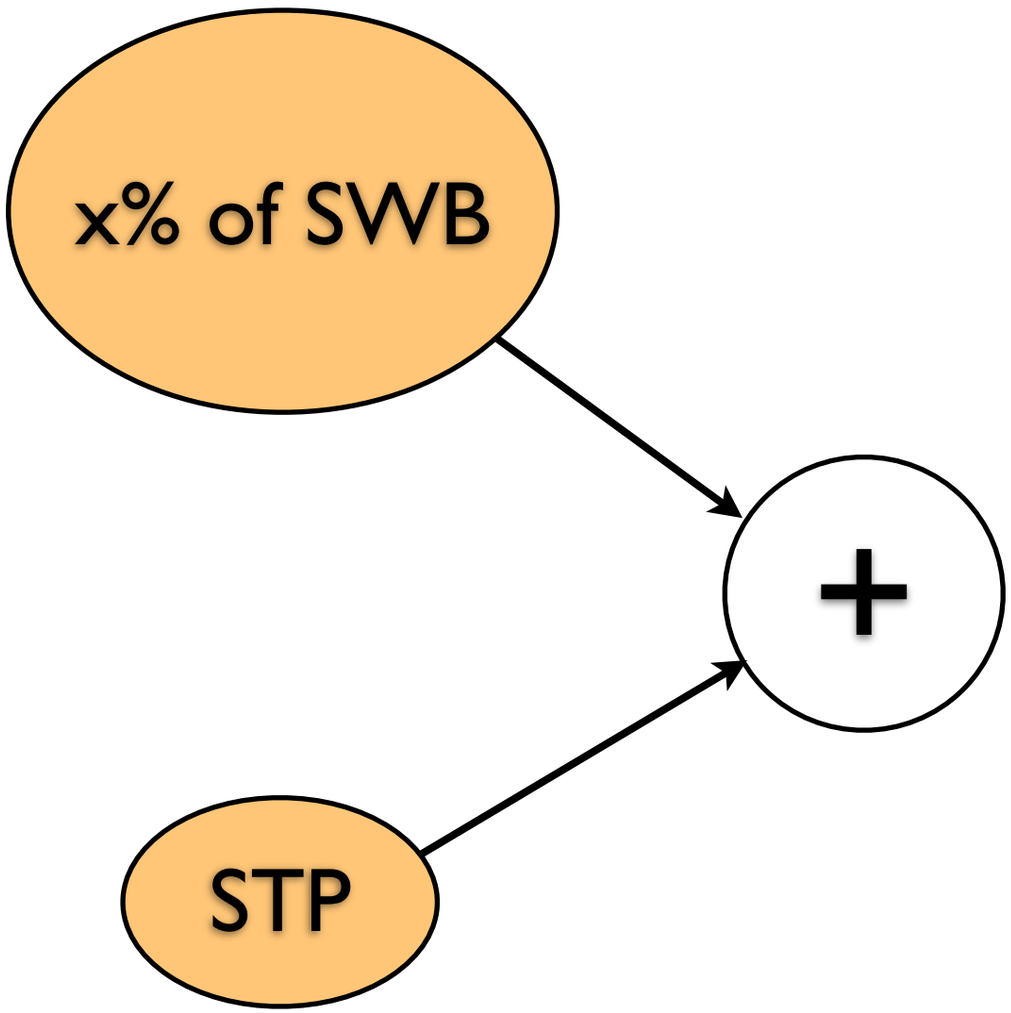
- Switchboard corpus consists of about 300 hours of conversational speech.
  - Less reliable automatically generated phone-level annotations [Deshmukh et al., 1998]

Switchboard Phonetic Annotation

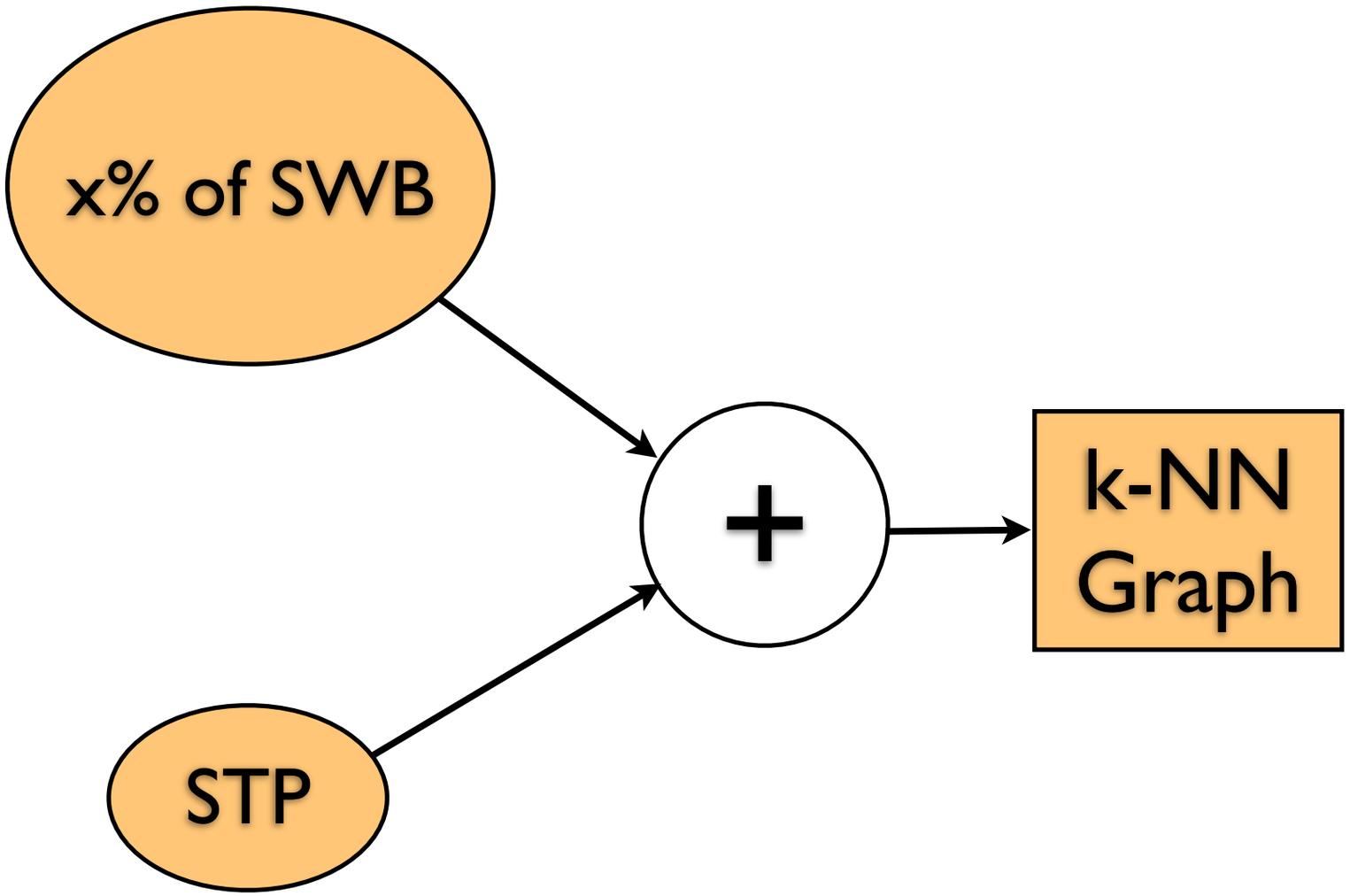
- Switchboard corpus consists of about 300 hours of conversational speech.
  - Less reliable automatically generated phone-level annotations [Deshmukh et al., 1998]
- **Switchboard transcription project** (STP) [Greenberg, 1995]
  - Manual phonetic annotation
  - Only about 75 minutes of data annotated

Graph Construction

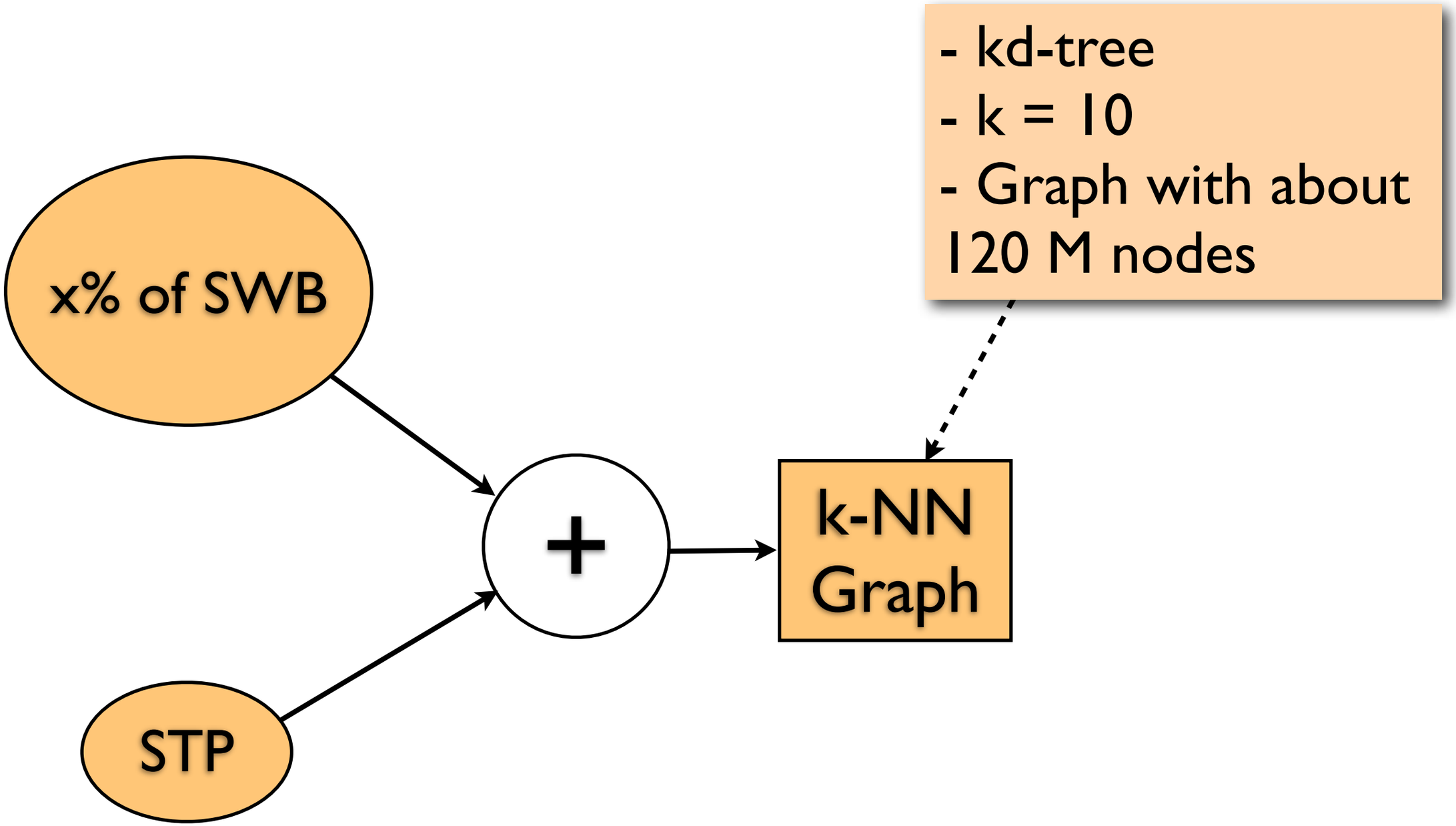
Graph Construction



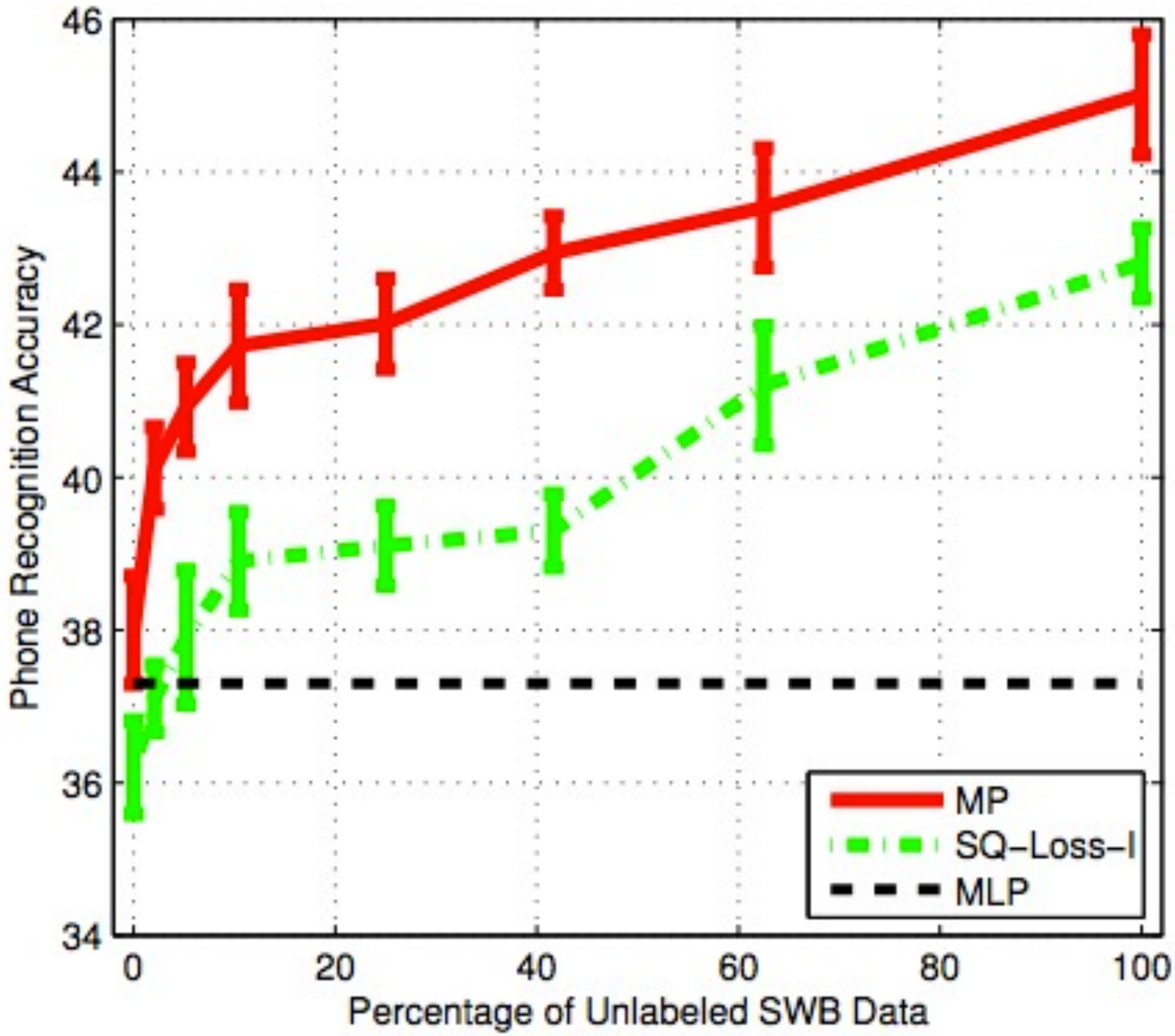
Graph Construction



Graph Construction

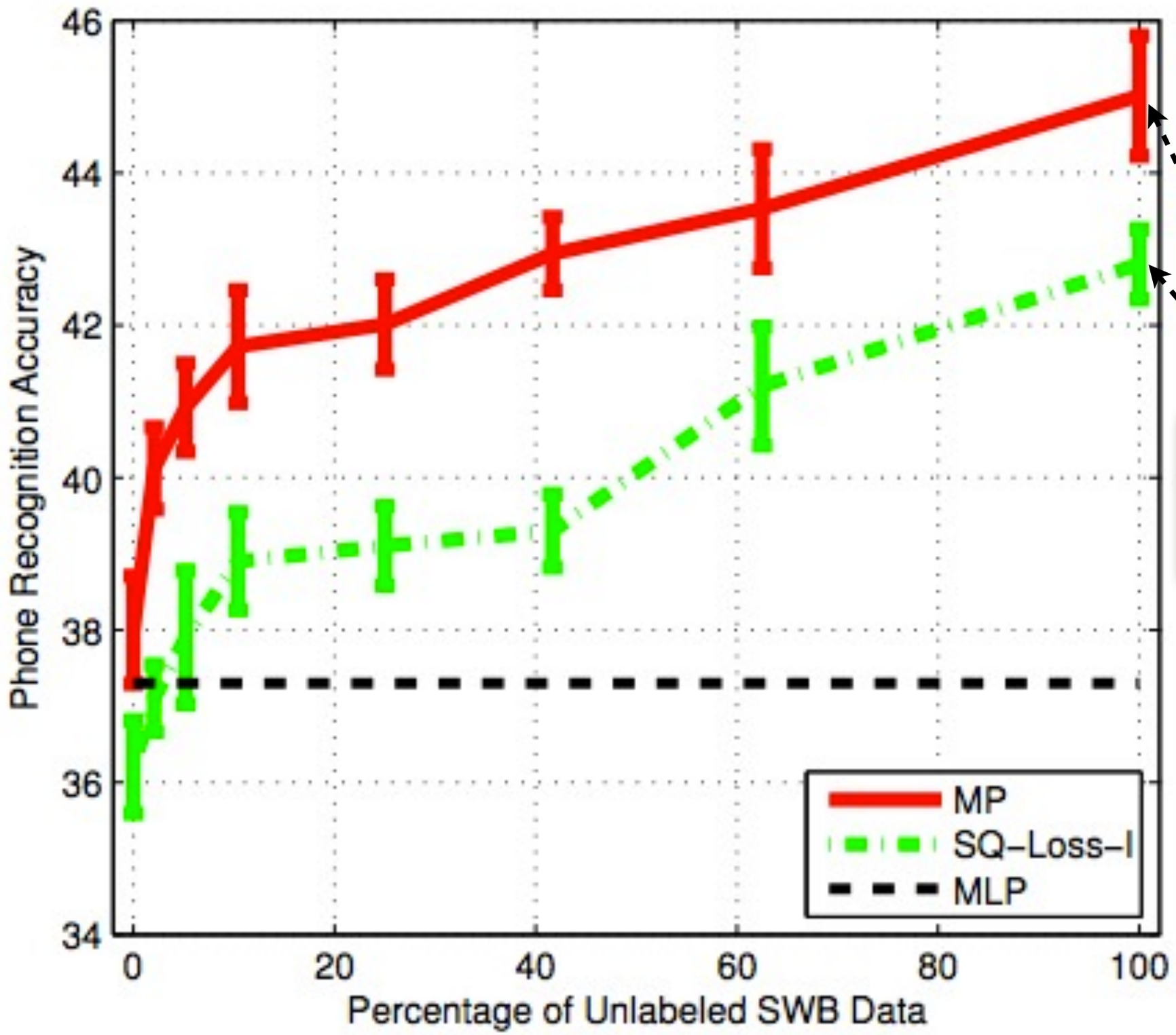


Results



[Subramanya & Bilmes, JMLR 2011]

Results



Graph with **120m** vertices

[Subramanya & Bilmes, JMLR 2011]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Phone Classification
  - Text Categorization
  - Dialog Act Tagging
  - Statistical Machine Translation
  - POS Tagging
  - MultiLingual POS Tagging
- Conclusion & Future Work

Problem Description & Motivation

Problem Description & Motivation

- Given a document (e.g., web page, news article), assign it to a fixed number of semantic categories (e.g., sports, politics, entertainment)

Problem Description & Motivation

- Given a document (e.g., web page, news article), assign it to a fixed number of semantic categories (e.g., sports, politics, entertainment)
- <u>Multi-label</u> problem

Problem Description & Motivation

- Given a document (e.g., web page, news article), assign it to a fixed number of semantic categories (e.g., sports, politics, entertainment)
- Multi-label problem
- Training supervised models requires large amounts of labeled data [Dumais et al., 1998]

Corpora

- **Reuters** [Lewis, et al., 1978]
  - Newswire
  - About 20K document with 135 categories. Use top 10 categories (e.g., “earnings”, “acquistions”, “wheat”, “interest”) and label the remaining as “other”

Corpora

- **Reuters** [Lewis, et al., 1978]
  - Newswire
  - About 20K document with 135 categories. Use top 10 categories (e.g., “earnings”, “acquistions”, “wheat”, “interest”) and label the remaining as “other”
- **WebKB** [Bekkerman, et al., 2003]
  - 8K webpages from 4 academic domains
  - Categories include “course”, “department”, “faculty” and “project”

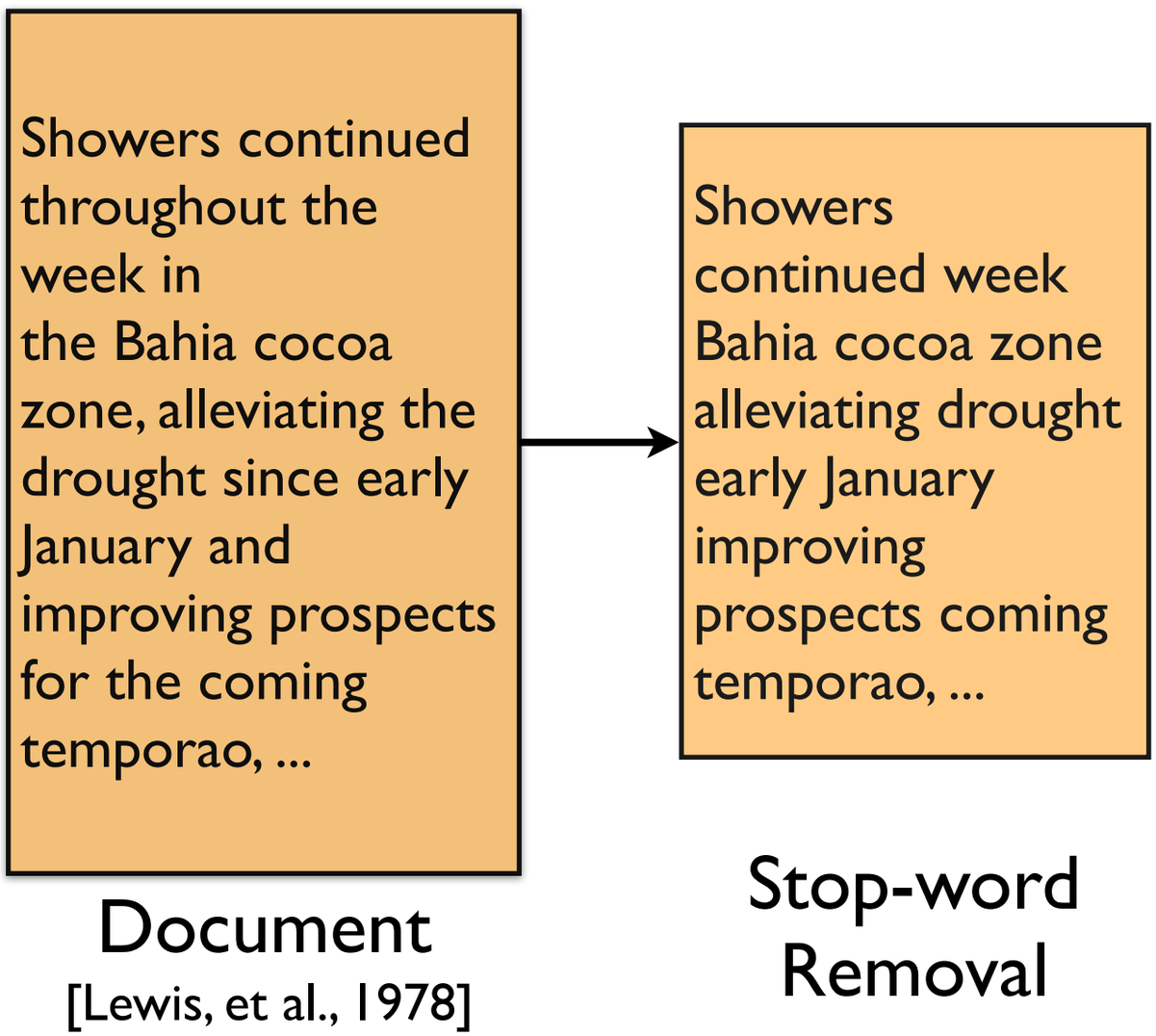
Feature Extraction

Showers continued throughout the week in
the Bahia cocoa zone, alleviating the drought since early January and improving prospects for the coming temporao, ...

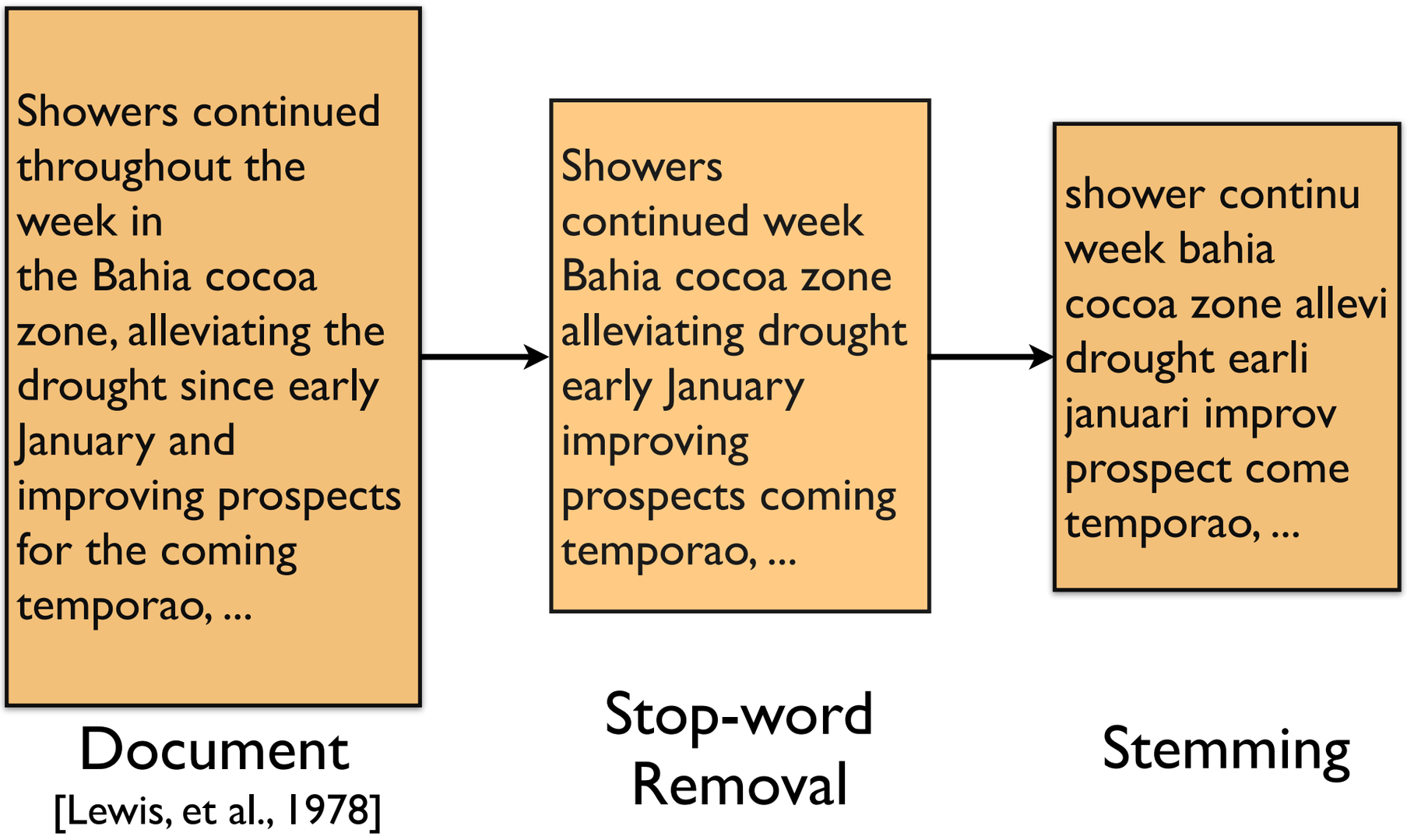
Document

[Lewis, et al., 1978]

Feature Extraction



Feature Extraction



Feature Extraction

Showers continued throughout the week in the Bahia cocoa zone, alleviating the drought since early January and improving prospects for the coming temporao, ...

Document
[Lewis, et al., 1978]

Showers continued week Bahia cocoa zone alleviating drought early January improving prospects coming temporao, ...

Stop-word Removal

shower continu week bahia cocoa zone allevi drought earli januari improv prospect come temporao, ...

Stemming

shower	0.01
bahia	0.34
cocoa	0.33
.	.
.	.
.	.
.	.

TFIDF weighted Bag-of-words

Results

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Support Vector Machine (Supervised)

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Transductive SVM [Joachims 1999]

Support Vector Machine (Supervised)

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

Precision-recall break even point (PRBEP)

Results

Support Vector Machine (Supervised)

Transductive SVM [Joachims 1999]

Spectral Graph Transduction (SGT) [Joachims 2003]

Label Propagation [Zhu & Ghahramani 2002]

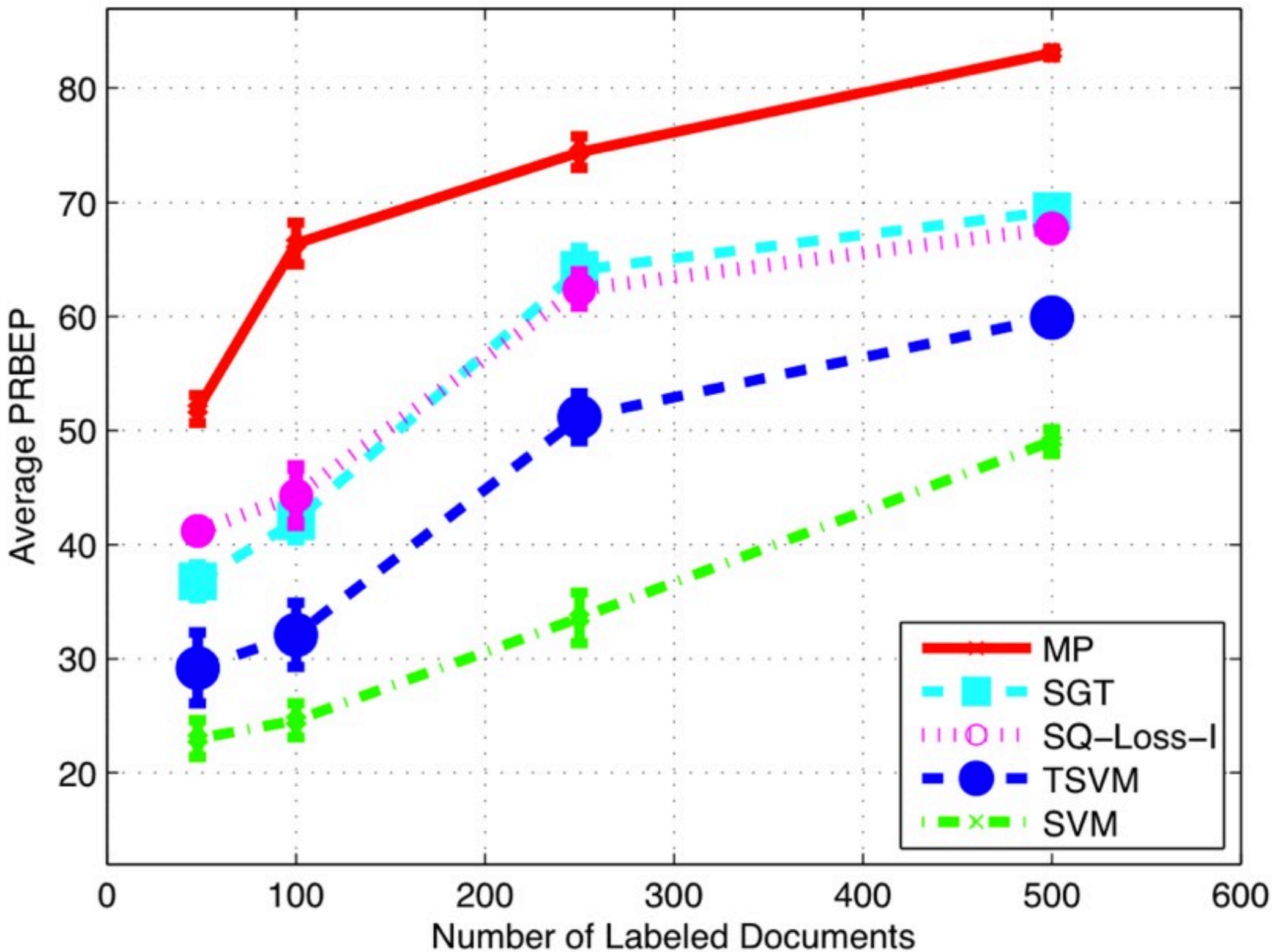
Measure Propagation [Subramanya & Bilmes 2008]

Modified Adsorption [Talukdar & Crammer 2009]

Average PRBEP	SVM	TSVM	SGT	LP	MP	MAD
Reuters	48.9	59.3	60.3	59.7	**66.3**	-
WebKB	23.0	29.2	36.8	41.2	51.9	**53.7**

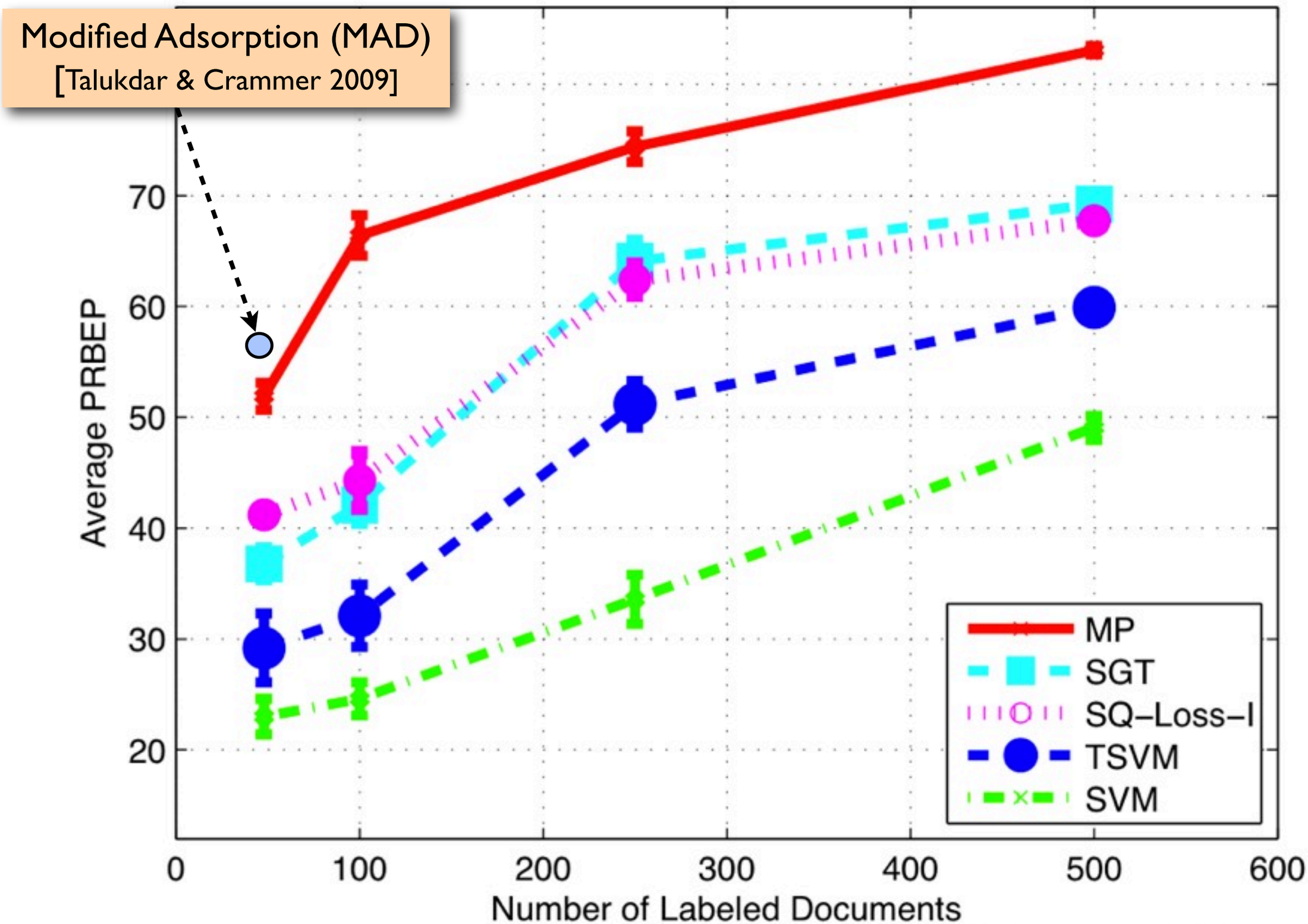
Precision-recall break even point (PRBEP)

Results on WebKB



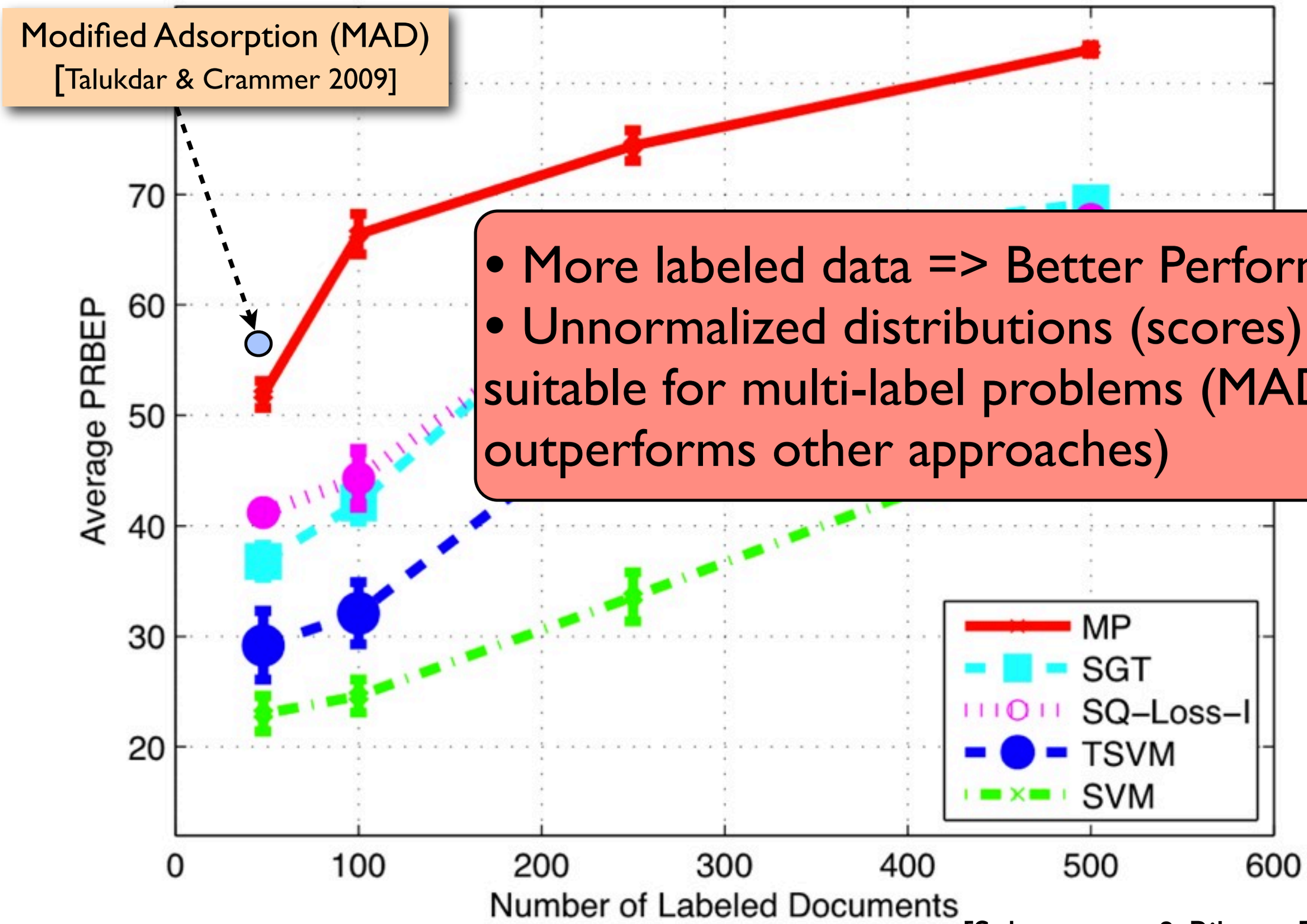
[Subramanya & Bilmes, EMNLP 2008]

Results on WebKB



[Subramanya & Bilmes, EMNLP 2008]

Results on WebKB



- More labeled data => Better Performance
- Unnormalized distributions (scores) more suitable for multi-label problems (MAD outperforms other approaches)

[Subramanya & Bilmes, EMNLP 2008]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

Applications:
- Phone Classification
- Text Categorization
- Dialog Act Tagging [Subramanya & Bilmes, JMLR 2011]
- Statistical Machine Translation
- POS Tagging
- MultiLingual POS Tagging

Problem description

- Dialog acts (DA) reflect the function that utterances serve in discourse
- Applications in automatic speech recognition (ASR), machine translation (MT) & natural language processing (NLP)

Switchboard DA Corpus

Switchboard DA Corpus

- Switchboard dialog act (DA) tagging project [Jurafsky, et al., 1997]

Switchboard DA Corpus

- Switchboard dialog act (DA) tagging project [Jurafsky, et al., 1997]
  - Manually labeled 1155 conversations (about 200k sentences)

Switchboard DA Corpus

- Switchboard dialog act (DA) tagging project [Jurafsky, et al., 1997]
  - Manually labeled 1155 conversations (about 200k sentences)
  - Example labels: question, answer, backchannel, agreement... (total of 42 different DAs)

Switchboard DA Corpus

- Switchboard dialog act (DA) tagging project [Jurafsky, et al., 1997]
  - Manually labeled 1155 conversations (about 200k sentences)
  - Example labels: question, answer, backchannel, agreement... (total of 42 different DAs)
  - Use top 18 DAs (about 185k sentences)

Graph Construction

Graph Construction

- Bigram & trigram TFIDF

Graph Construction

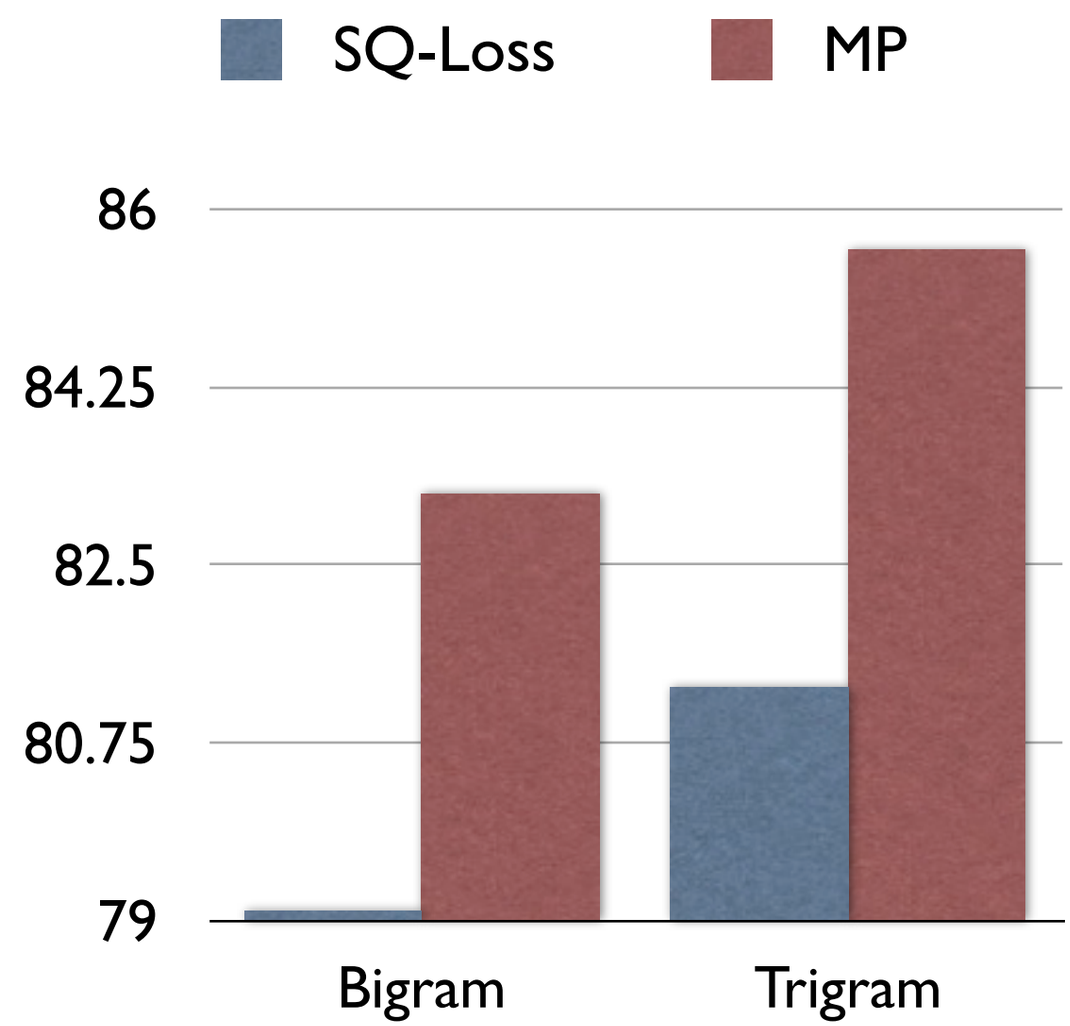
- Bigram & trigram TFIDF
- Cosine similarity

Graph Construction

- Bigram & trigram TFIDF
- Cosine similarity
- k-NN graph

SWB DA Tagging Results

- Baseline performance: 84.2% [Ji & Bilmes, 2005]



[Subramanya & Bilmes, JMLR 2011]

Dihana Corpus

Dihana Corpus

- Computer-human dialogues
  - answering telephone queries about train services in Spain
  - about 900 dialogues
  - topics include timetables, fares and services

Dihana Corpus

- Computer-human dialogues
  - answering telephone queries about train services in Spain
  - about 900 dialogues
  - topics include timetables, fares and services
- 225 speakers

Dihana Corpus

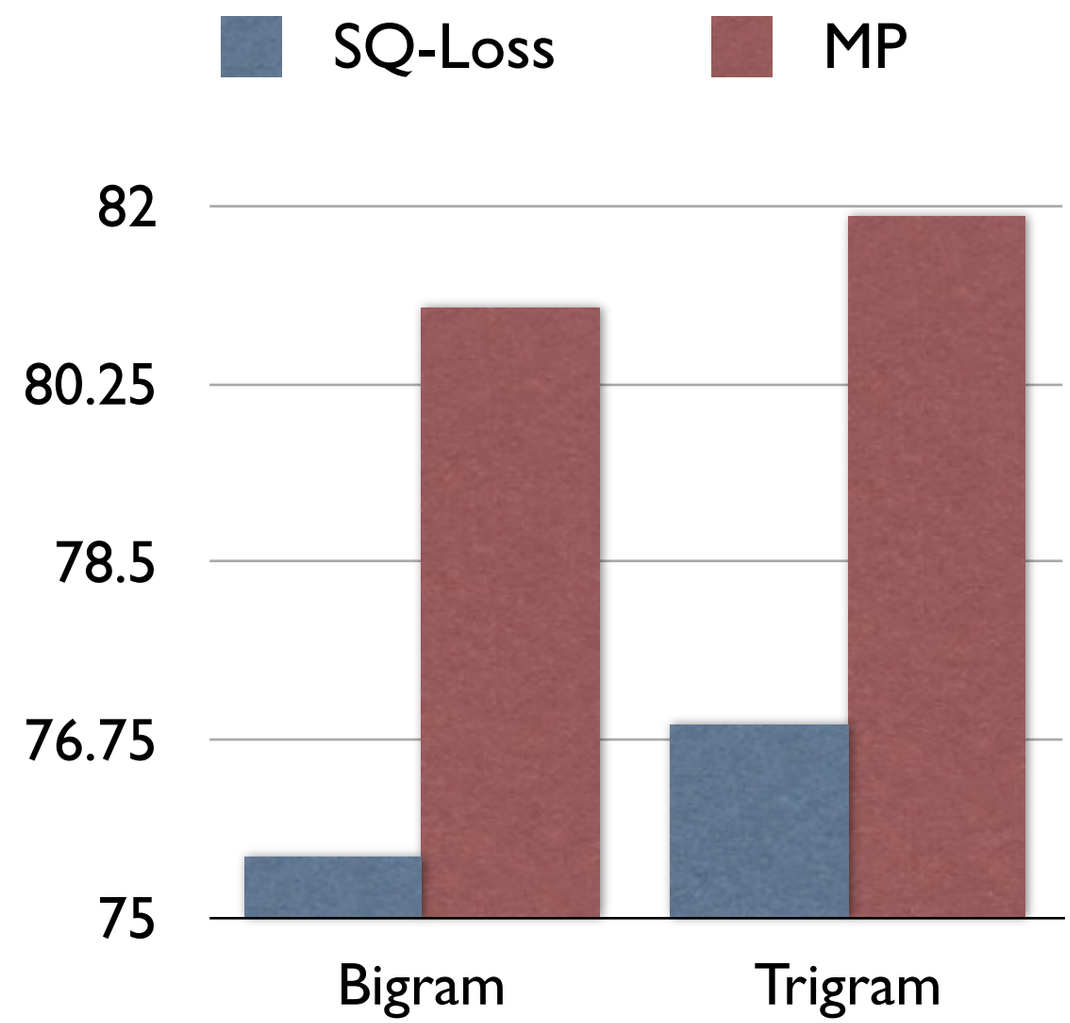
- Computer-human dialogues
  - answering telephone queries about train services in Spain
  - about 900 dialogues
  - topics include timetables, fares and services
- 225 speakers
- Standard train/test set (16k/7.5k sentences)

Dihana Corpus

- Computer-human dialogues
  - answering telephone queries about train services in Spain
  - about 900 dialogues
  - topics include timetables, fares and services
- 225 speakers
- Standard train/test set (16k/7.5k sentences)
- Number of labels = 72

Dihana Results

- Results for classifying user turns
- Baseline Performance: 76.4%



[Subramanya & Bilmes, JMLR 2011]

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Phone Classification
  - Text Categorization
  - Dialog Act Tagging
  - Statistical Machine Translation [Alexandrescu & Kirchoff, NAACL 2009]
  - POS Tagging
  - MultiLingual POS Tagging
- Conclusion & Future Work

Problem Description

- Phrase-based statistical machine translation (SMT)
- Sentences are translated in isolation

Motivation

Motivation

Source Asf lA ymknk *lk hnAk .....

Motivation

Source	Asf lA ymknk *lk hnAk .....
1-best	i'm sorry you can't there is a cost

Motivation

Source	Asf lA ymknk *lk hnAk .....
1-best	i'm sorry you can't there is a cost

Source	E*rA lA ymknk t$gyl .....
1-best	excuse me i you turn tv ....

Motivation

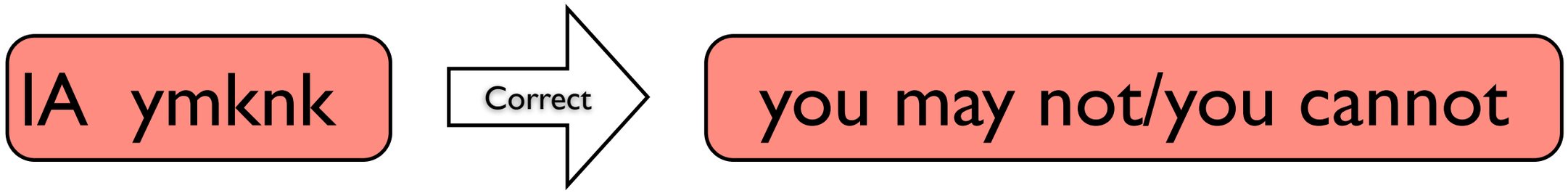
Source Asf lA ymknk *lk hnAk .....

1-best i'm sorry you can't there is a cost

Source E*rA lA ymknk t$gyl .....

1-best excuse me i you turn tv ....

Motivation



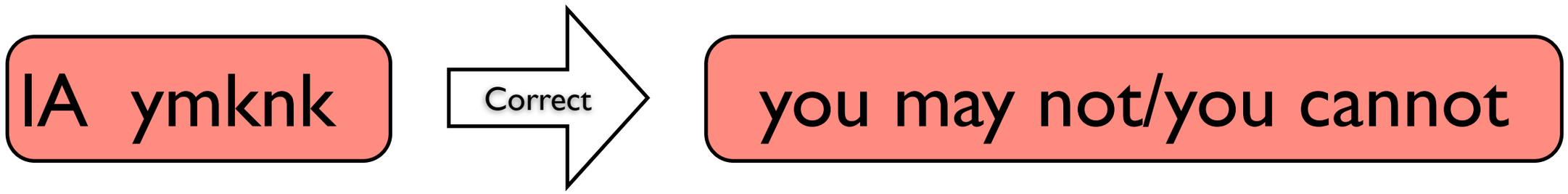
Source Asf lA ymknk *lk hnAk .....

1-best i'm sorry you can't there is a cost

Source E*rA lA ymknk t$gyl .....

1-best excuse me i you turn tv ....

Motivation



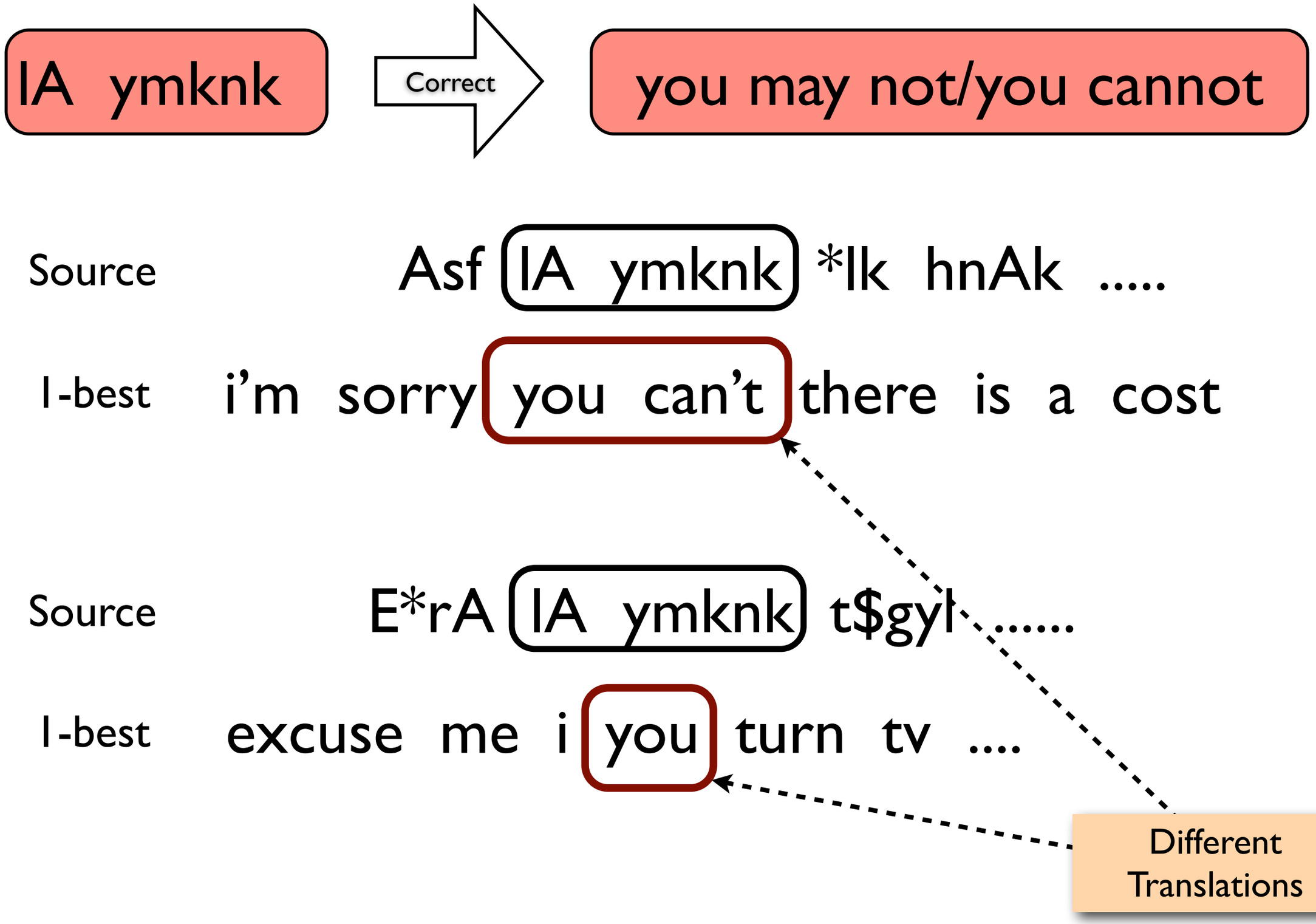
Source Asf lA ymknk *lk hnAk .....

1-best i'm sorry you can't there is a cost

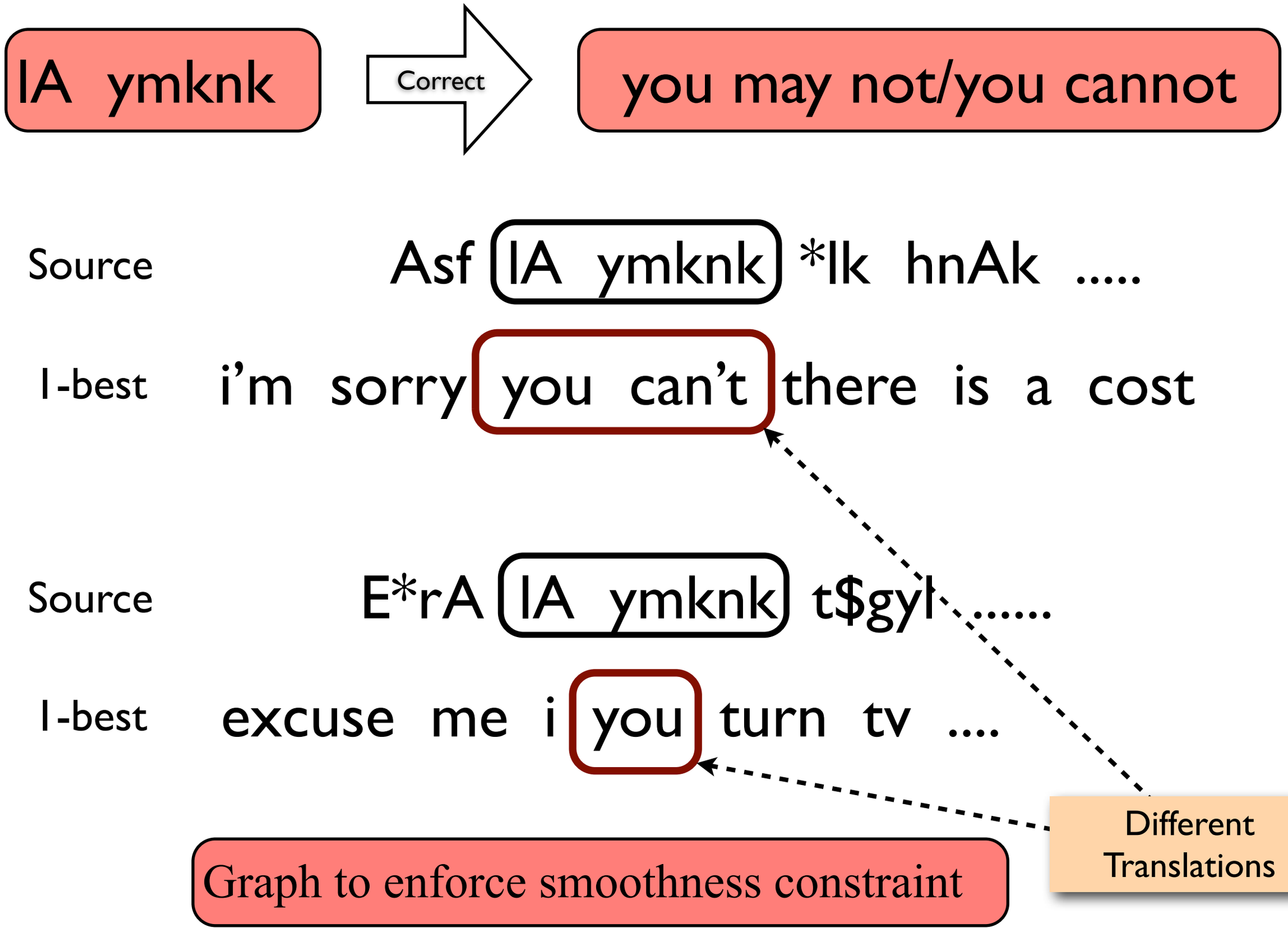
Source E*rA lA ymknk t$gyl .....

1-best excuse me i you turn tv ....

Motivation


lA ymknk
Correct
you may not/you cannot
Source
Asf lA ymknk *lk hnAk .....
1-best
i'm sorry you can't there is a cost
Source
E*rA lA ymknk t$gyl .....
1-best
excuse me i you turn tv ....
Different Translations

Motivation


lA ymknk
Correct
you may not/you cannot
Source
Asf lA ymknk *lk hnAk .....
1-best
i'm sorry you can't there is a cost
Source
E*rA lA ymknk t$gyl .....
1-best
excuse me i you turn tv ....
Different Translations
Graph to enforce smoothness constraint

Issues

Issues

- What we want to do -
  - exploit similarity between sentences

Issues

- What we want to do -
  - exploit similarity between sentences
- Input consists of variable-length word strings

Issues

- What we want to do -
  - exploit similarity between sentences
- Input consists of variable-length word strings
- Output space is structured (number of possible "labels" is very large)

Graph Construction (I)

Graph Construction (I)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Graph Construction (I)

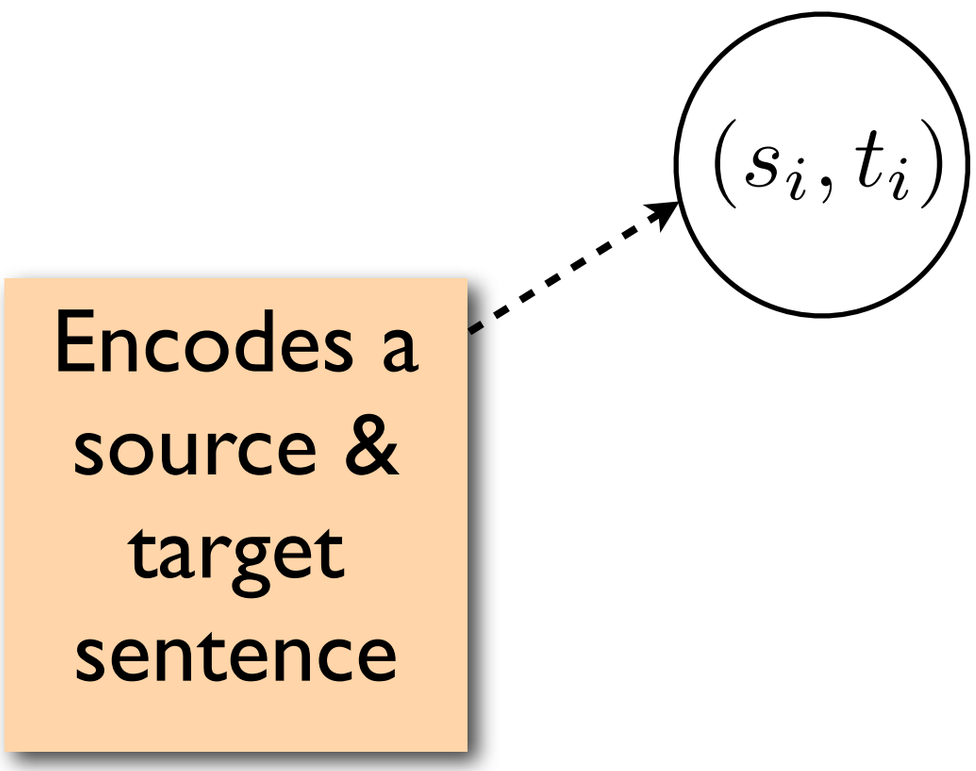
Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$

Graph Construction (I)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

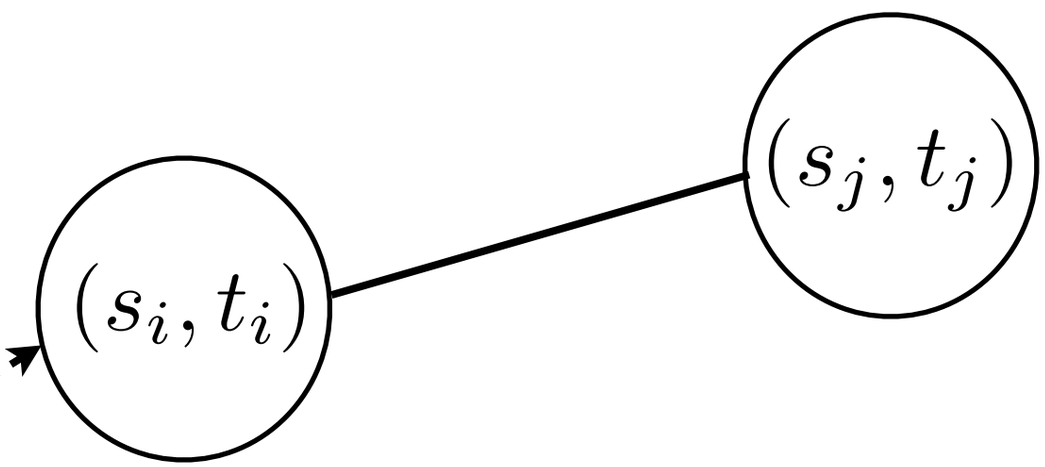
Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$



Graph Construction (I)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$

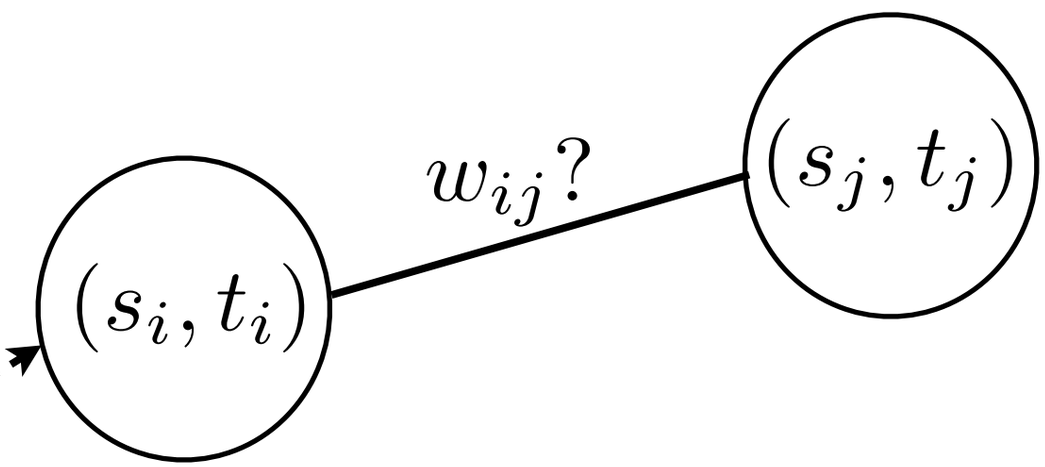


Encodes a source & target sentence

Graph Construction (I)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$

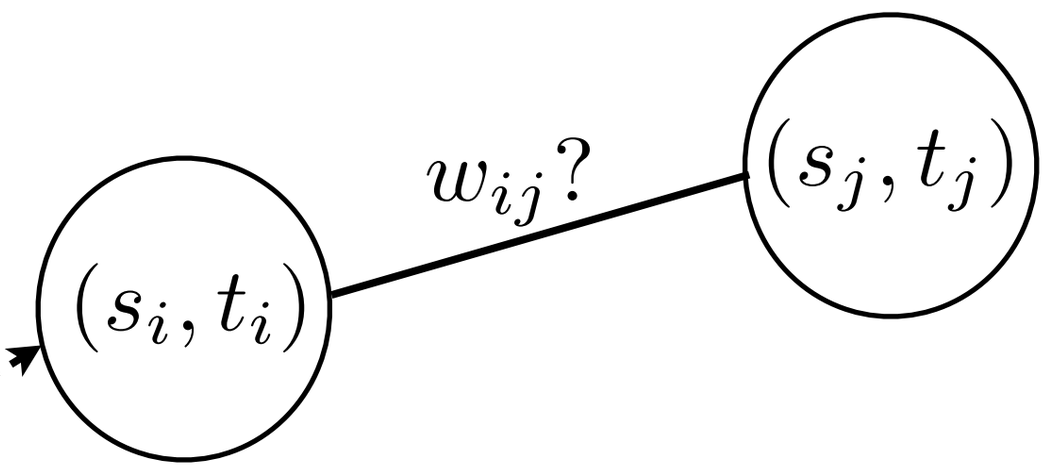


Encodes a source & target sentence

Graph Construction (I)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$



Encodes a source & target sentence

- How do we compute similarity
- What about the test set?

Graph Construction (II)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$

Graph Construction (II)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$

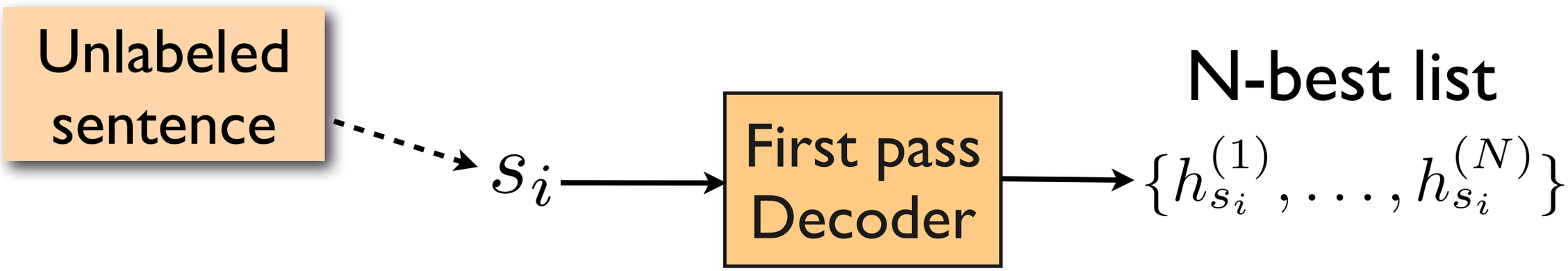
Unlabeled sentence

s_i

Graph Construction (II)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

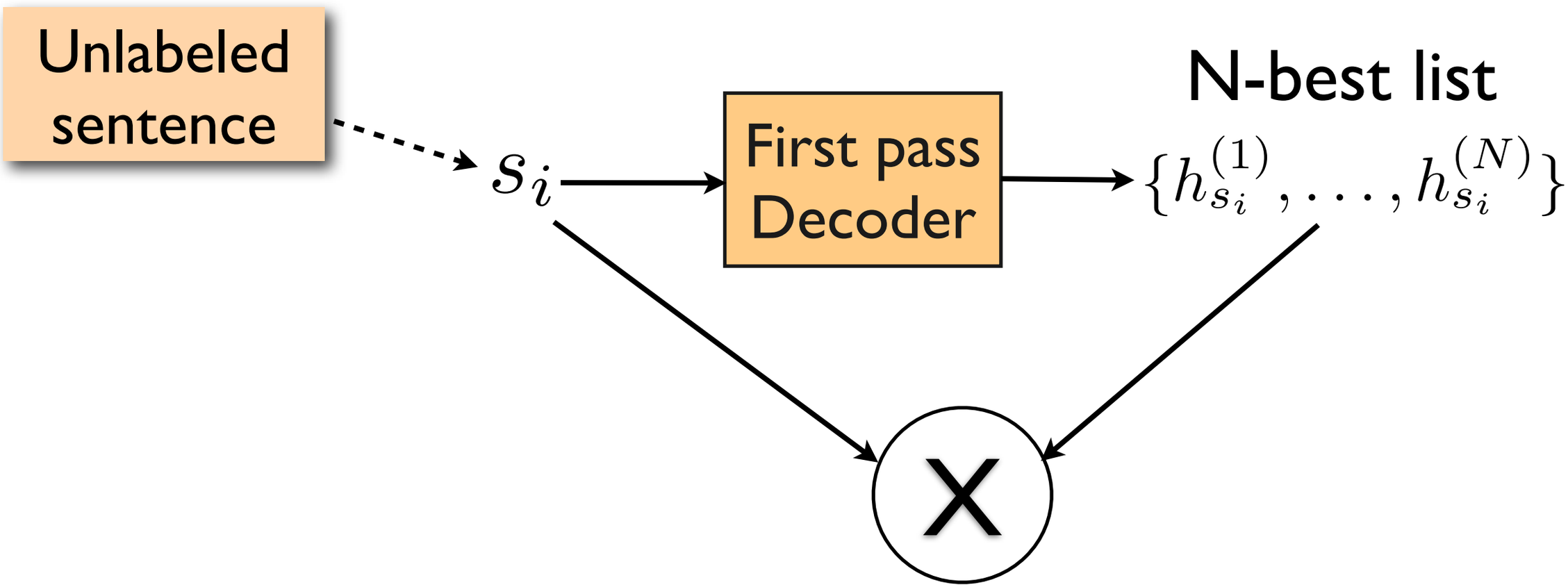
Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$



Graph Construction (II)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

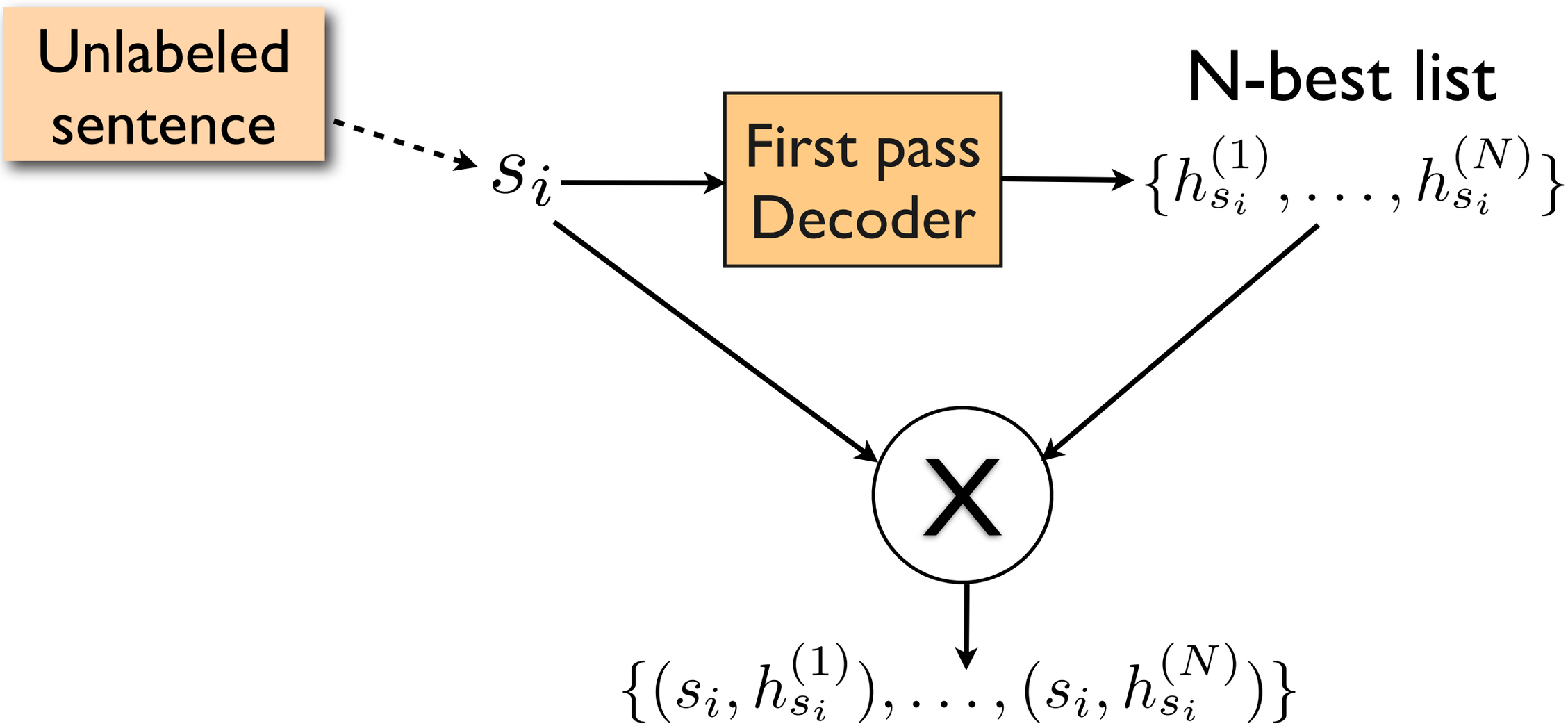
Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$



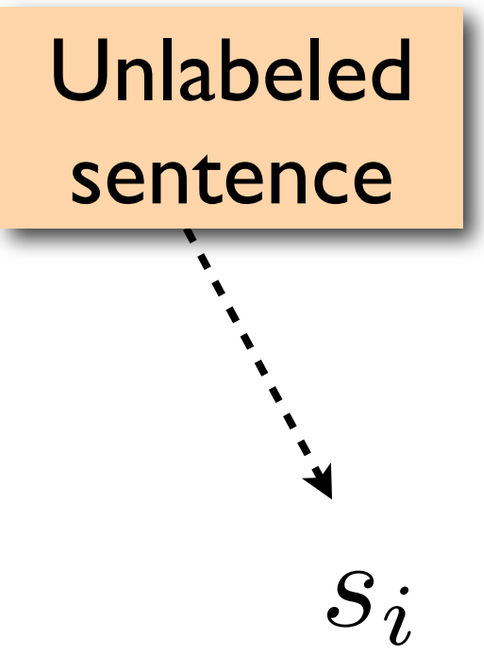
Graph Construction (II)

Labeled data: $\{(s_1, t_1), \ldots, (s_l, t_l)\}$

Unlabeled data (test set): $\{s_{l+1}, \ldots, s_n\}$



Graph Construction (III)



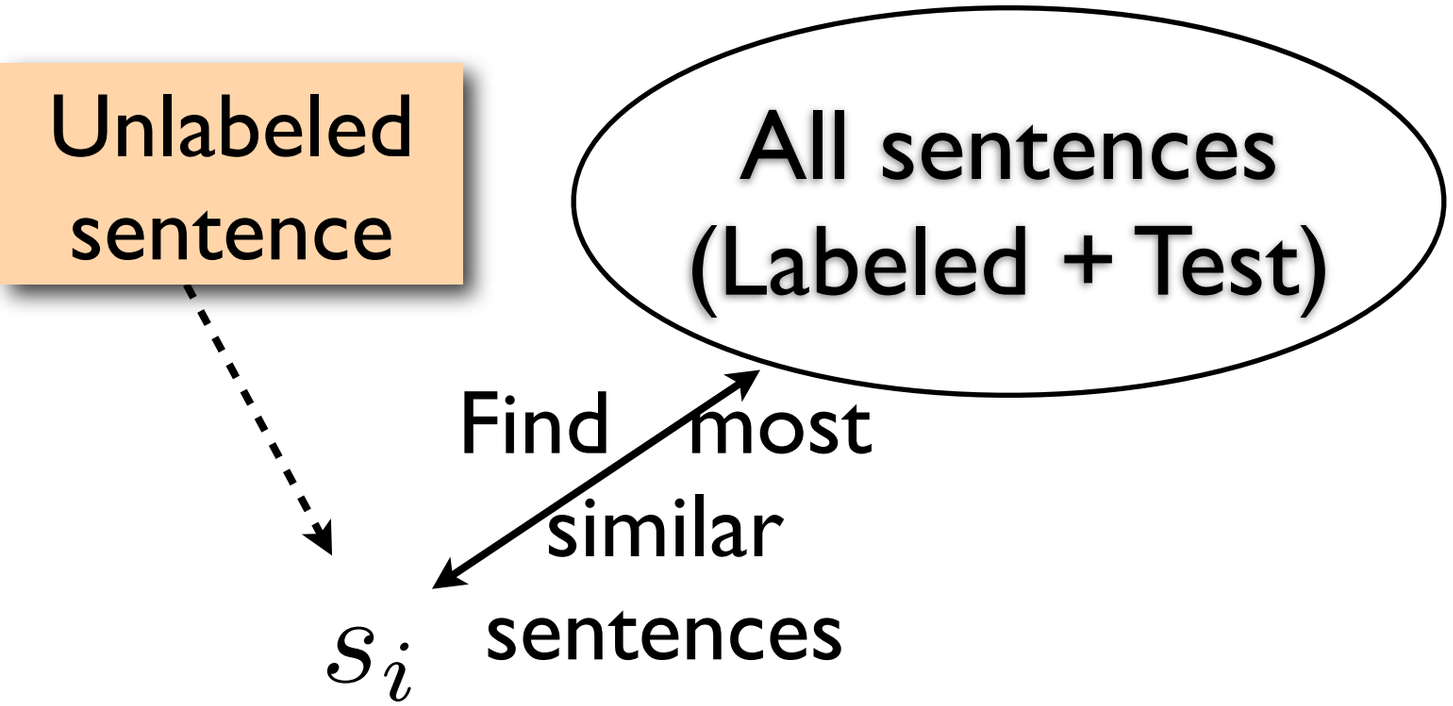
Graph Construction (III)

Unlabeled sentence

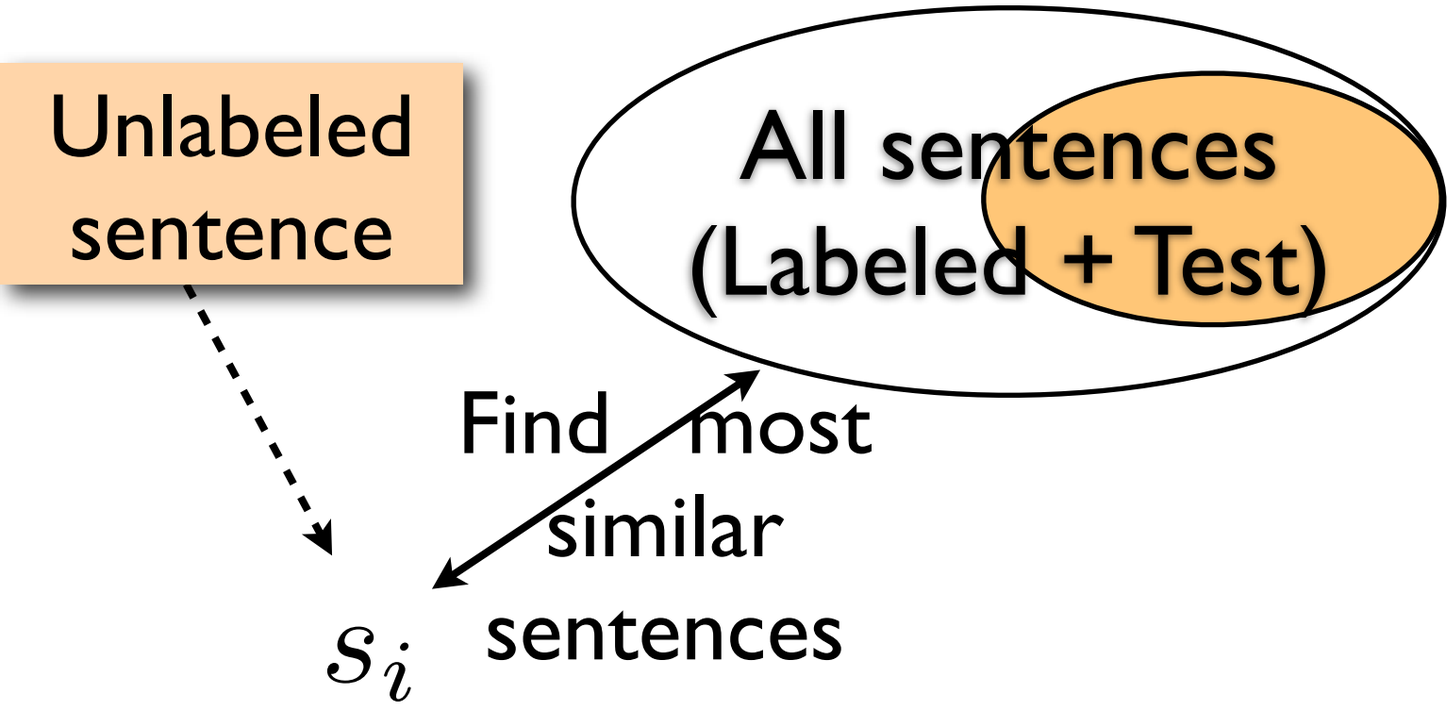
All sentences (Labeled + Test)

s_i

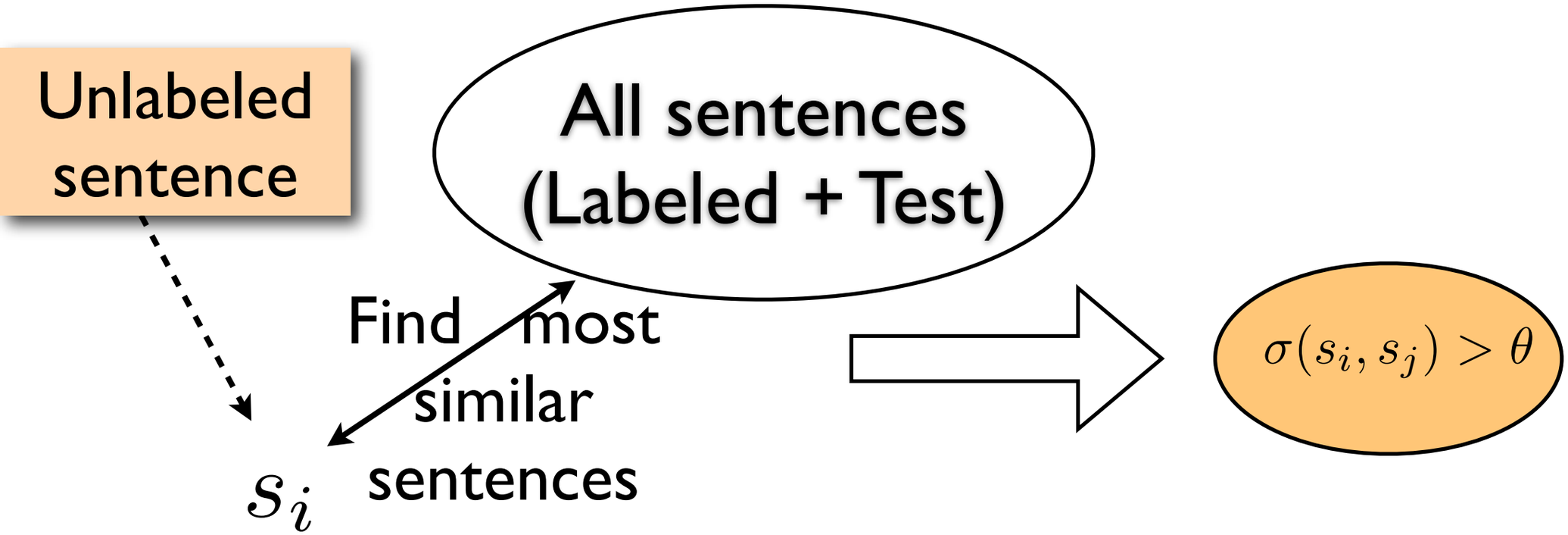
Graph Construction (III)



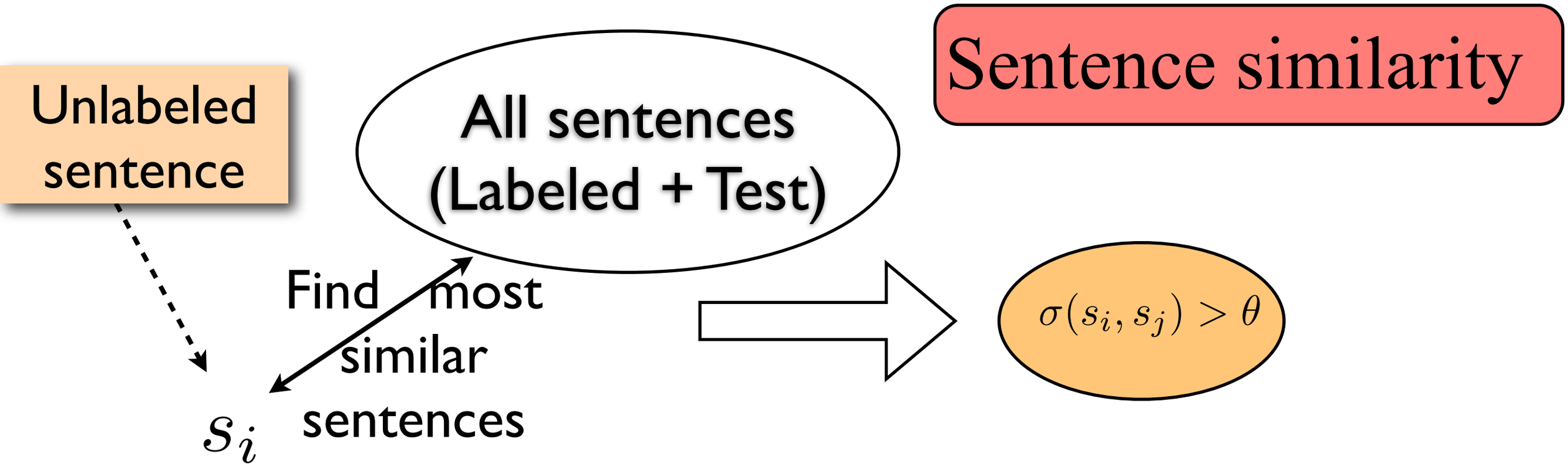
Graph Construction (III)



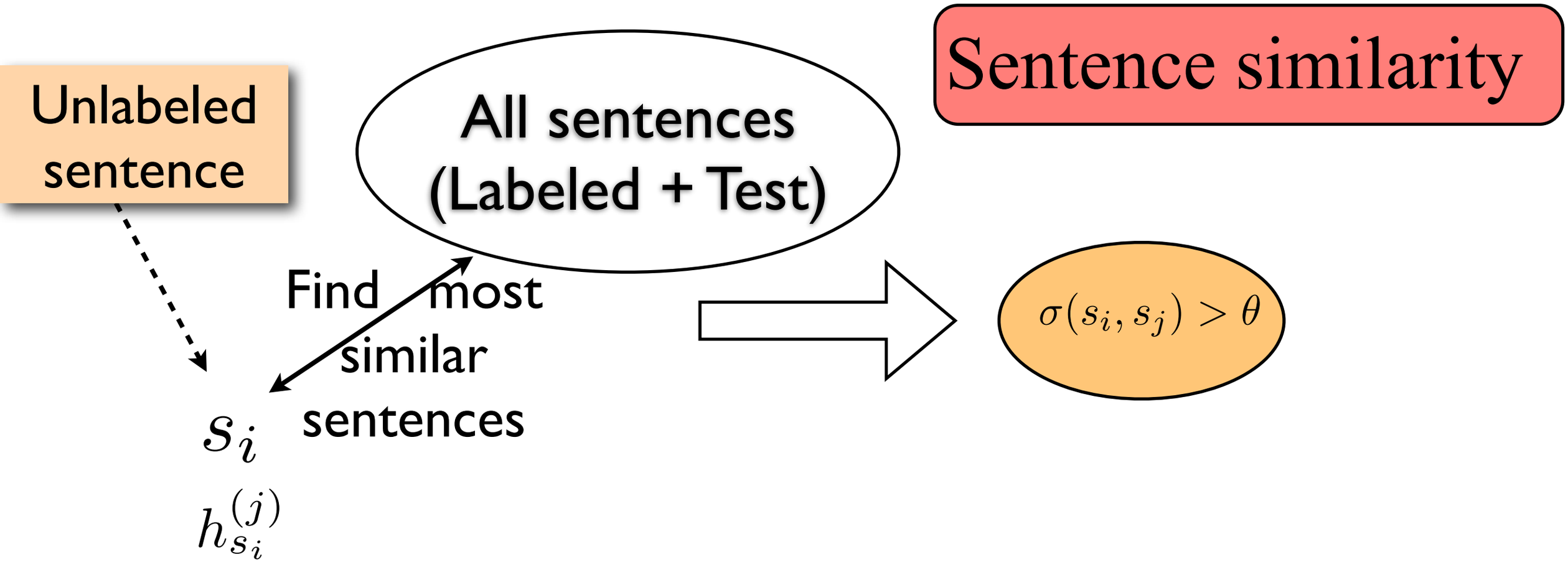
Graph Construction (III)



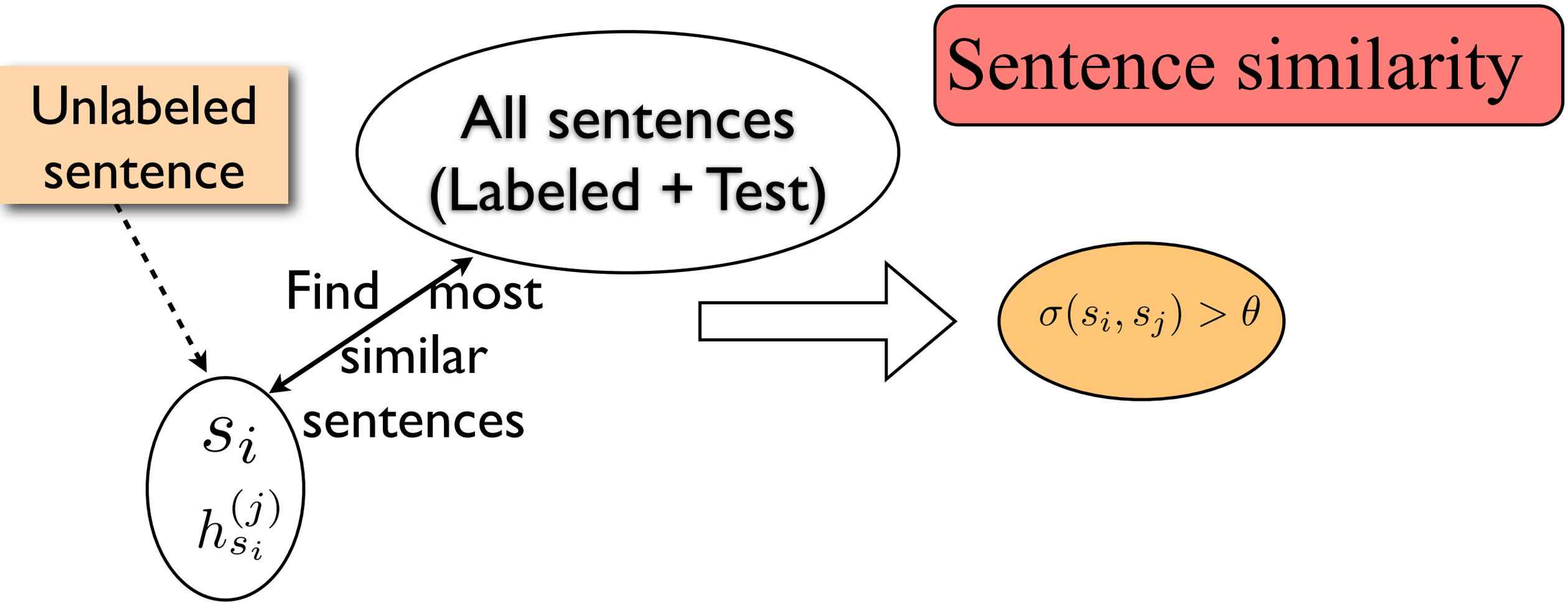
Graph Construction (III)



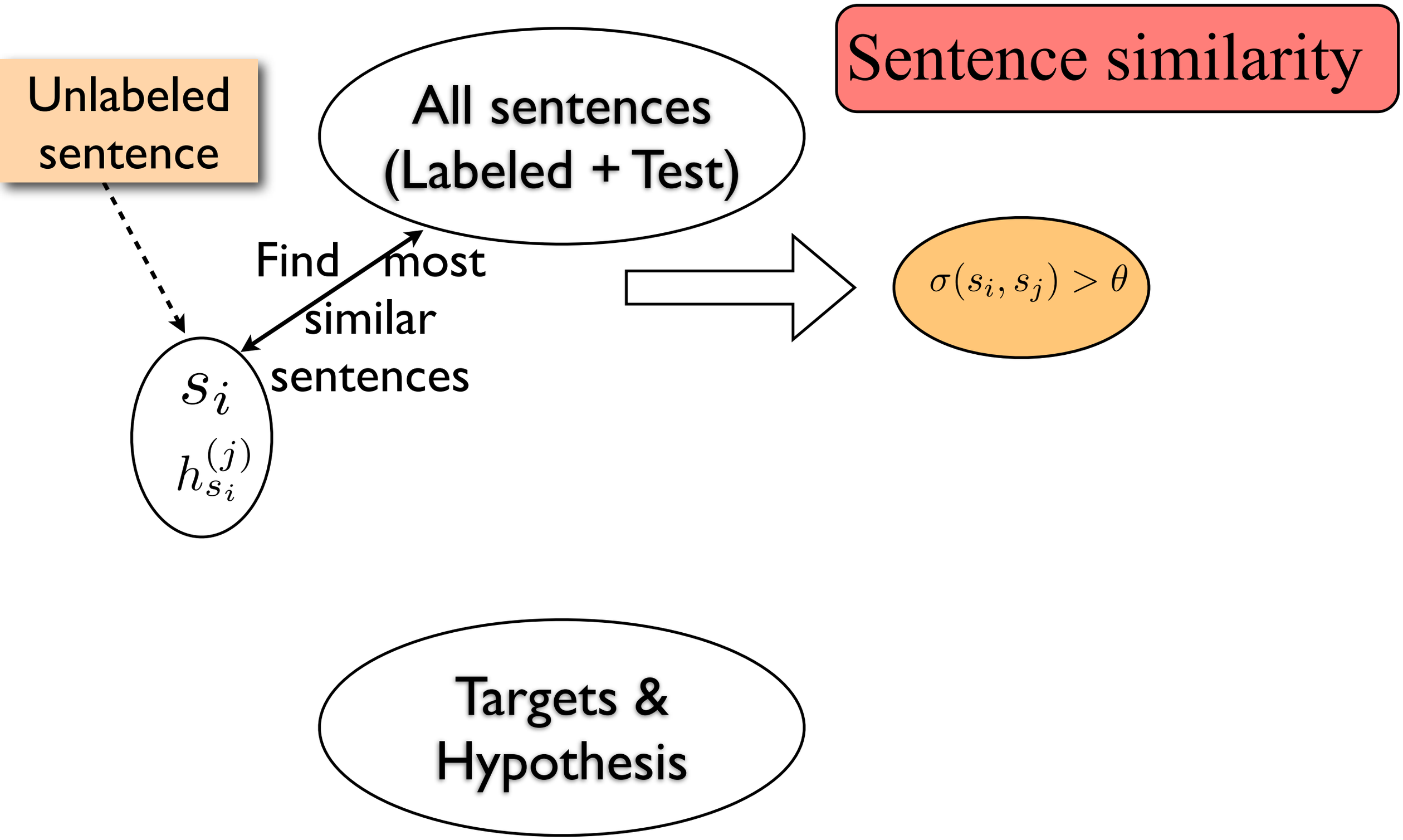
Graph Construction (III)



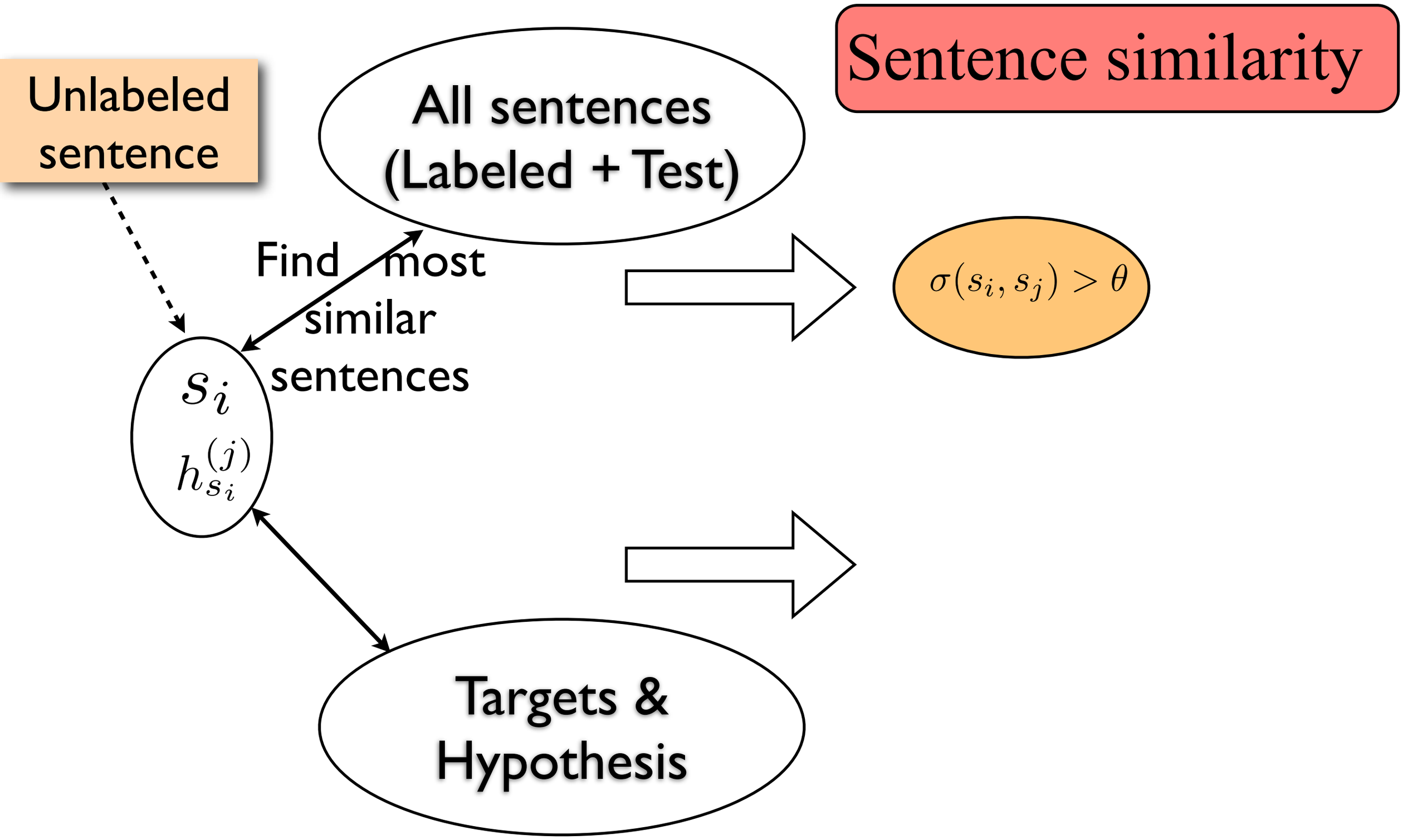
Graph Construction (III)



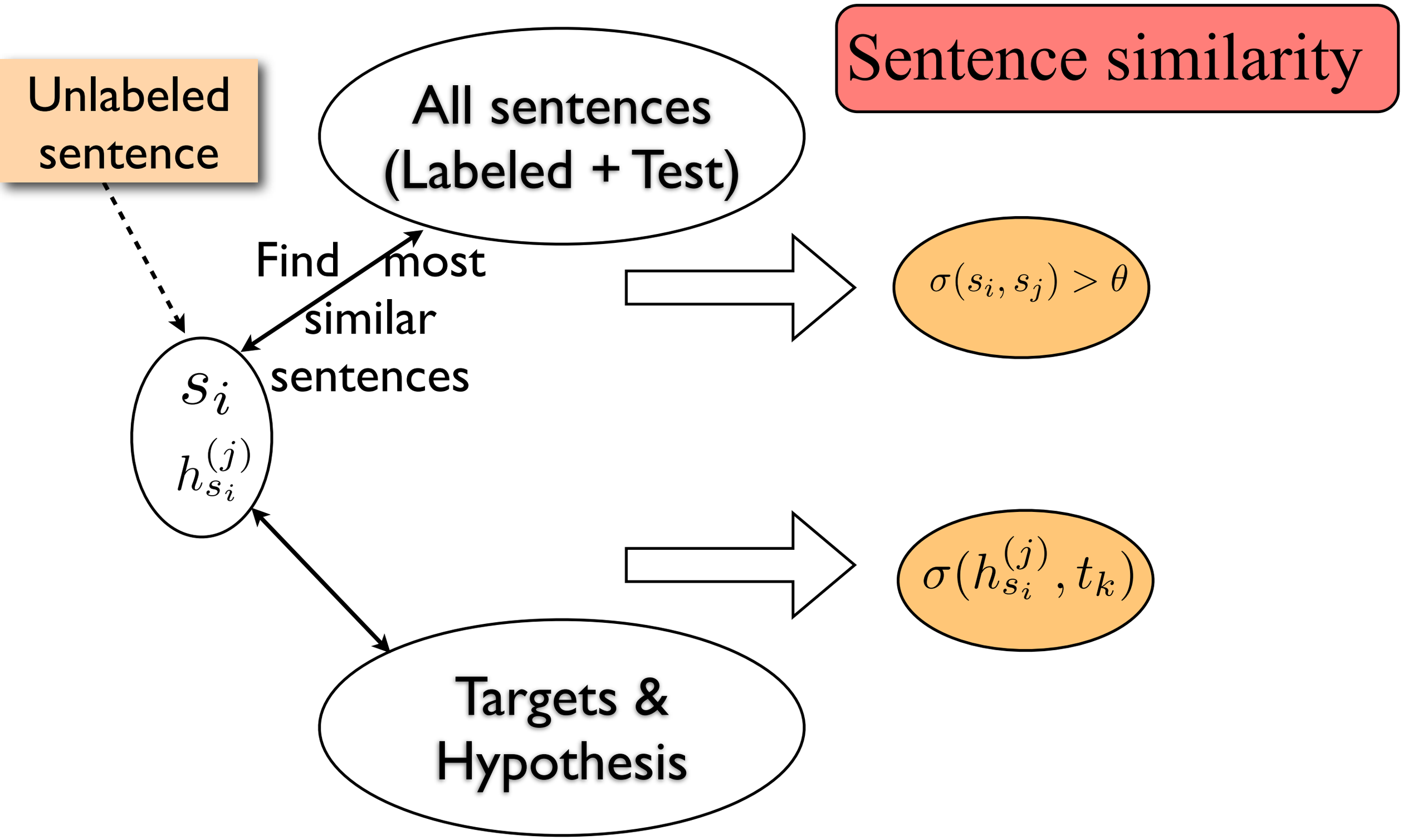
Graph Construction (III)



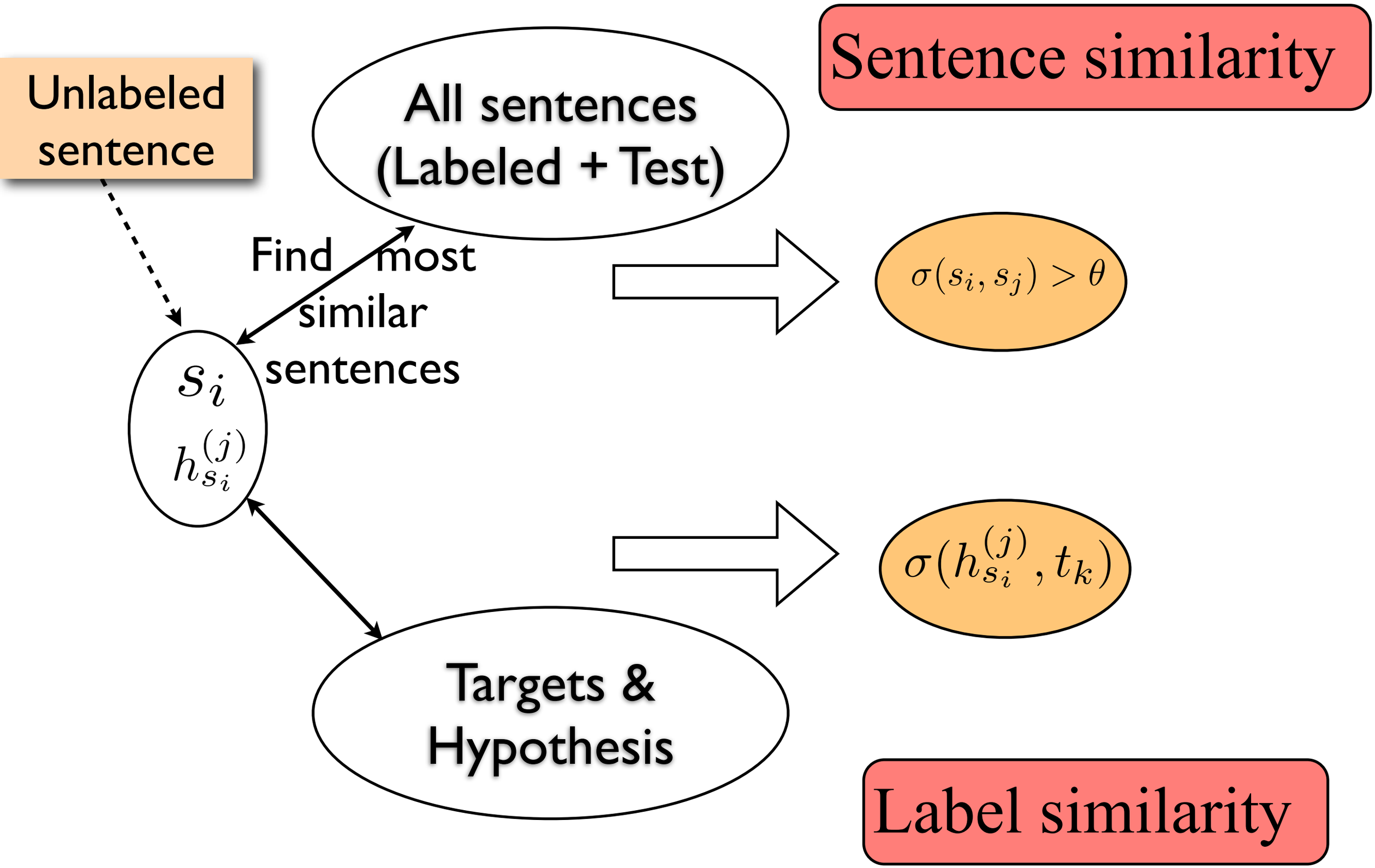
Graph Construction (III)



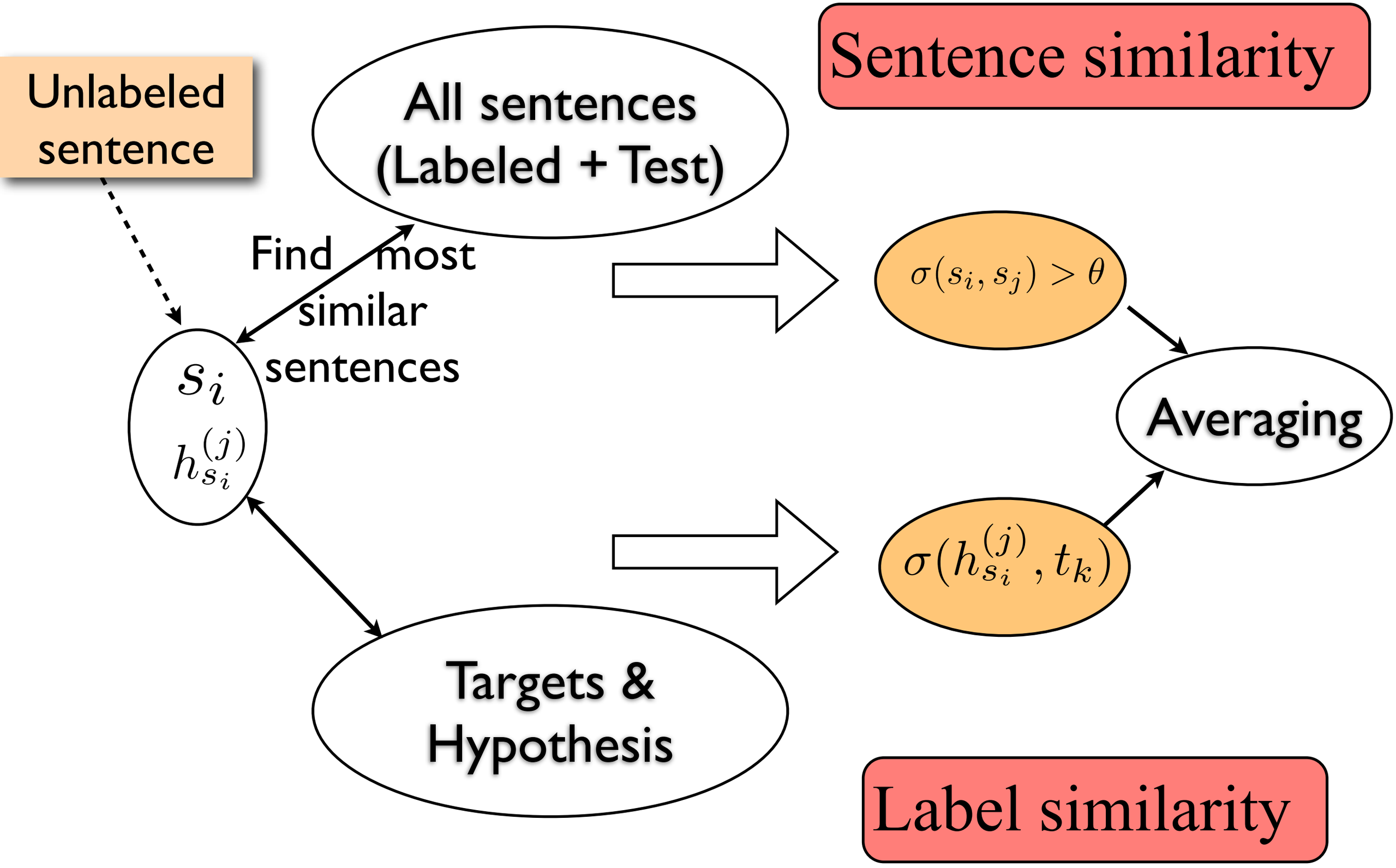
Graph Construction (III)



Graph Construction (III)



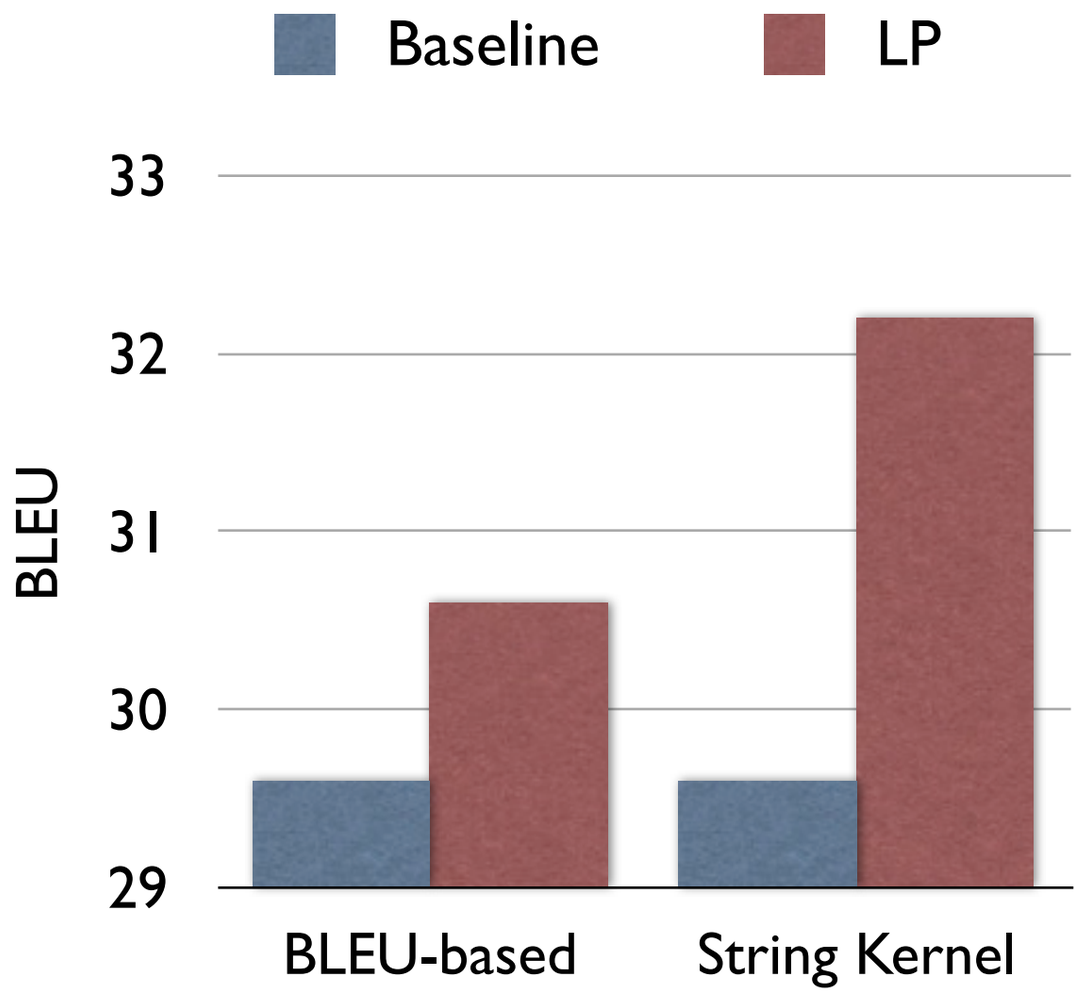
Graph Construction (III)



Corpora

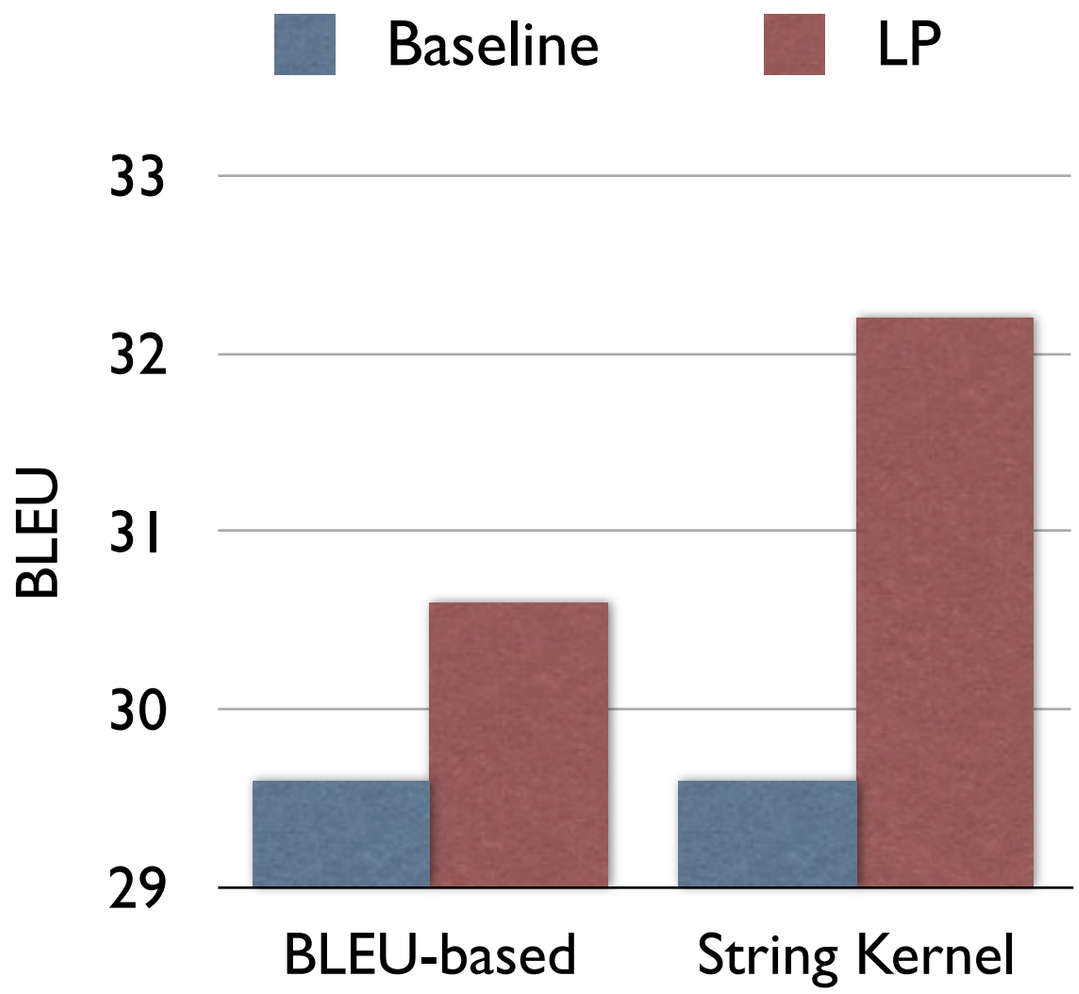
- IWSLT 2007 task
  - Italian-to-English (IE) & Arabic-to-English (AE) travel tasks
- Each task has train/dev/eval sets
- Baseline: Standard phrase-based SMT based on a log-linear model. Yields state-of-the-art performance.
- Results are measured using BLEU score and Phrase error rate (PER) [Papineni et al., ACL 2002]

Results (I)



IE Results on eval set

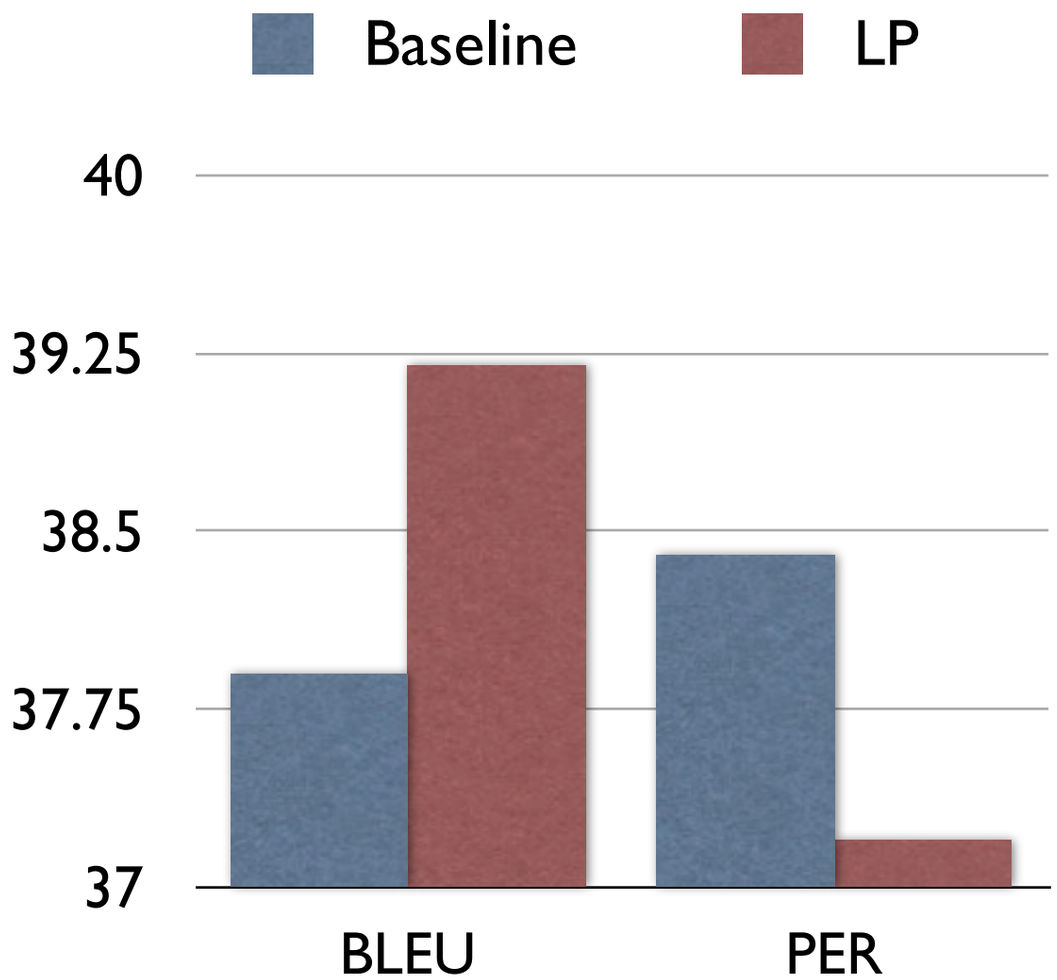
Results (I)



IE Results on eval set

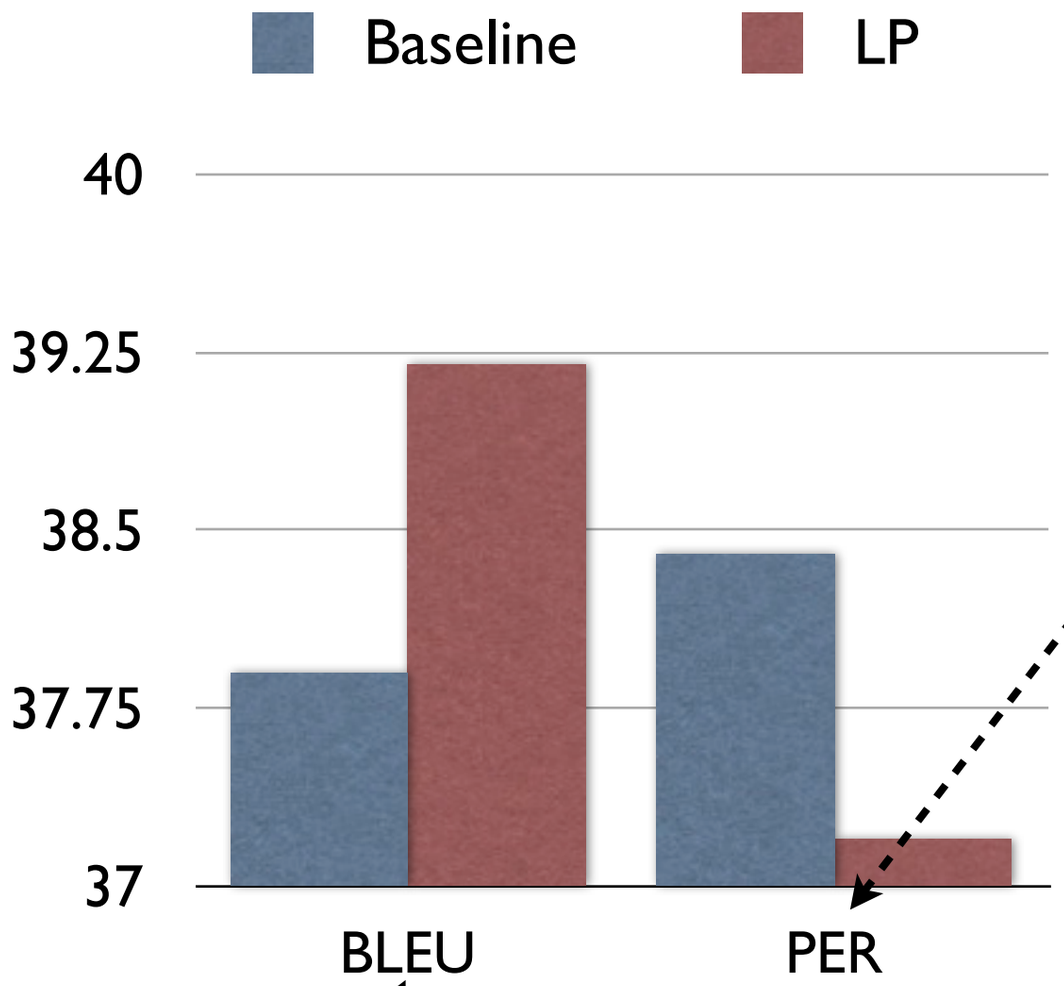
Geometric mean based averaging worked the best

Results (II)



IE Results on eval set

Results (II)

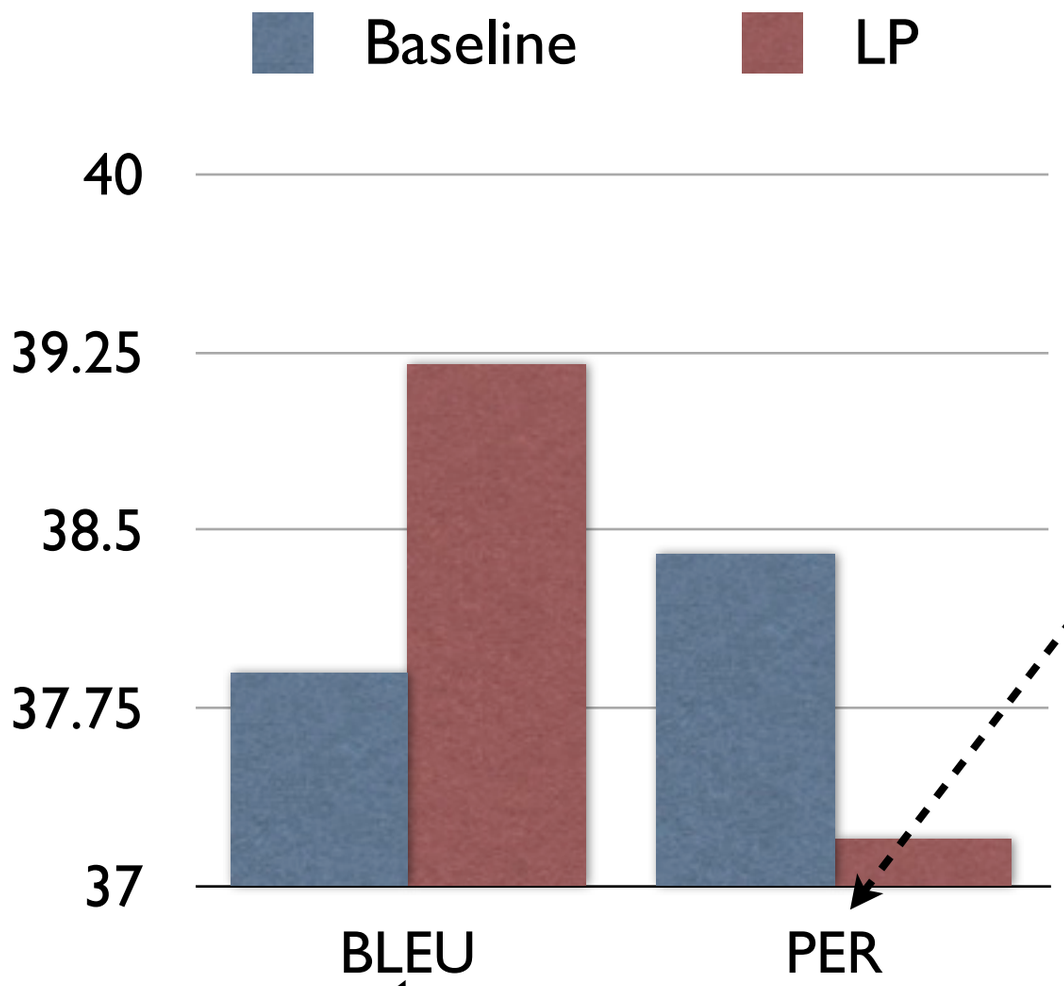


Lower is better

IE Results on eval set

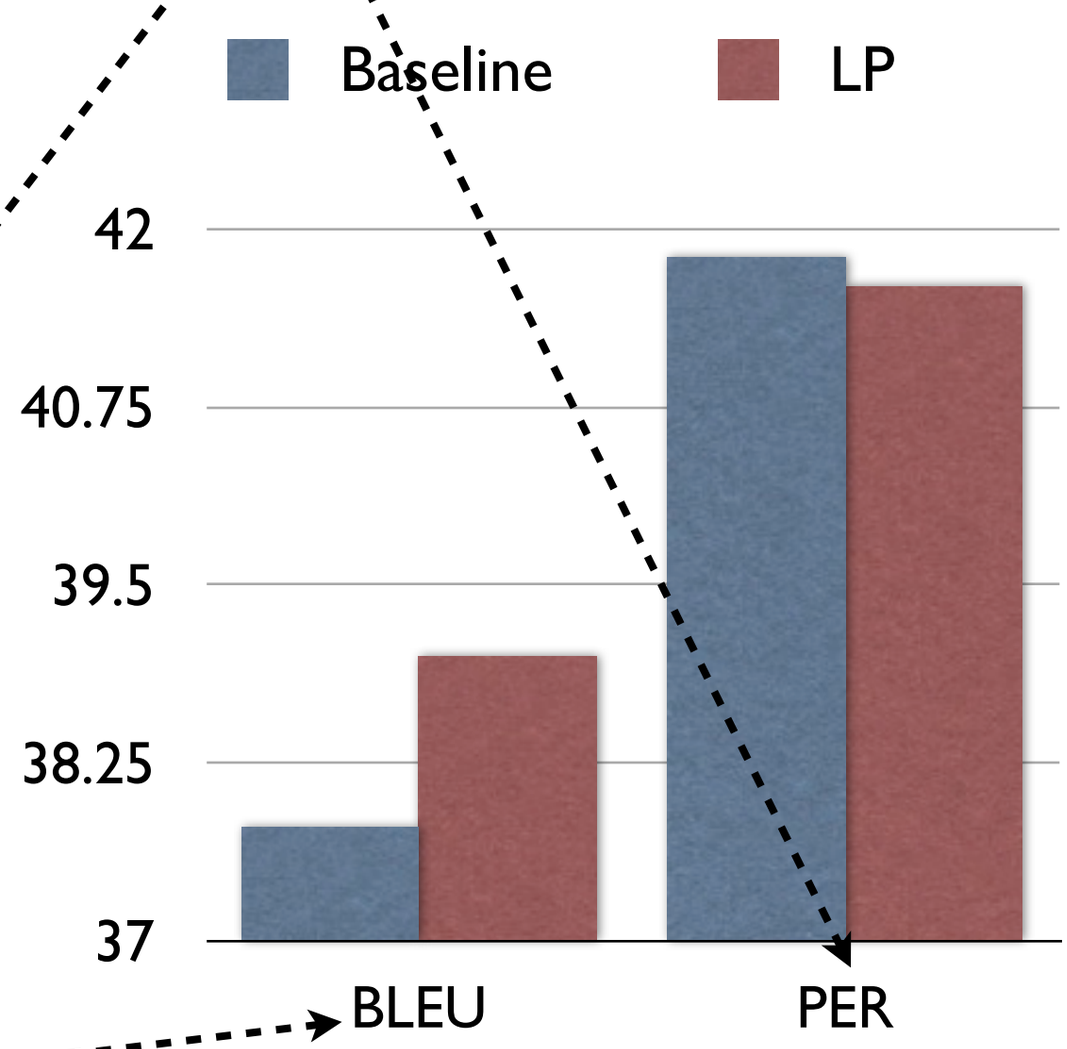
Higher is better

Results (II)



IE Results on eval set

Lower is better



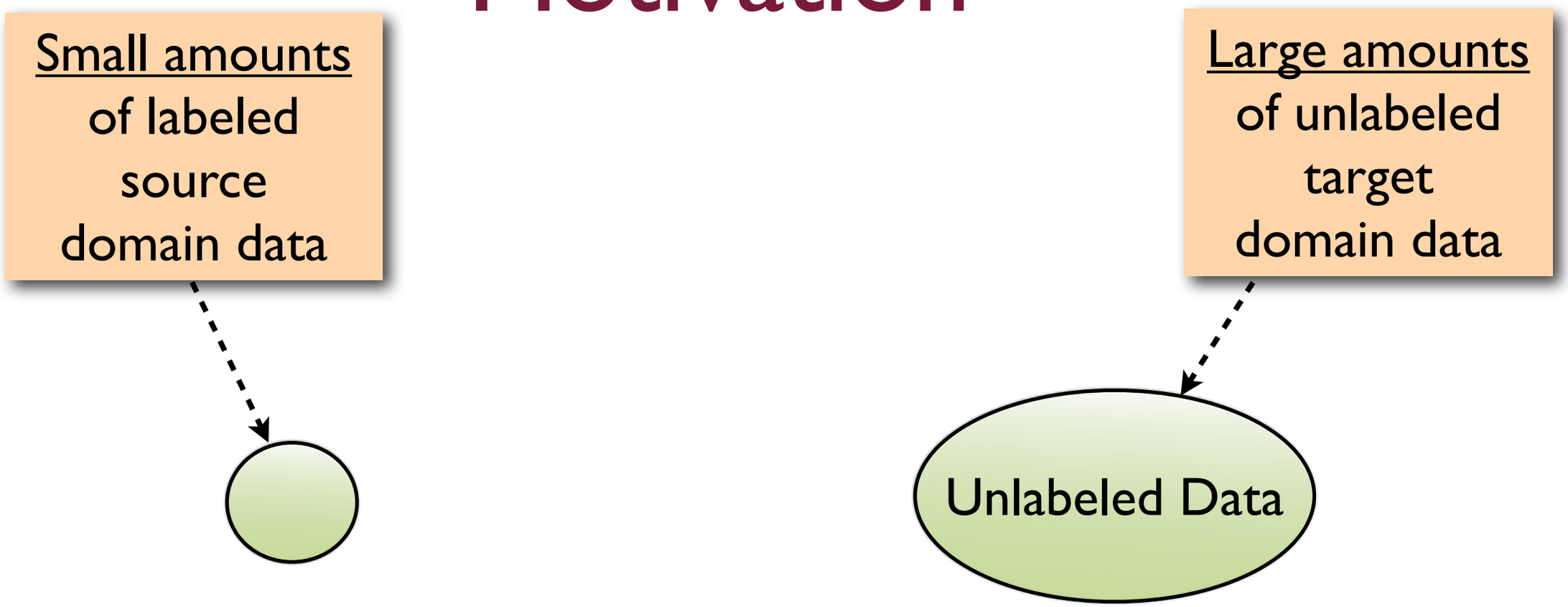
AE Results on eval set

Higher is better

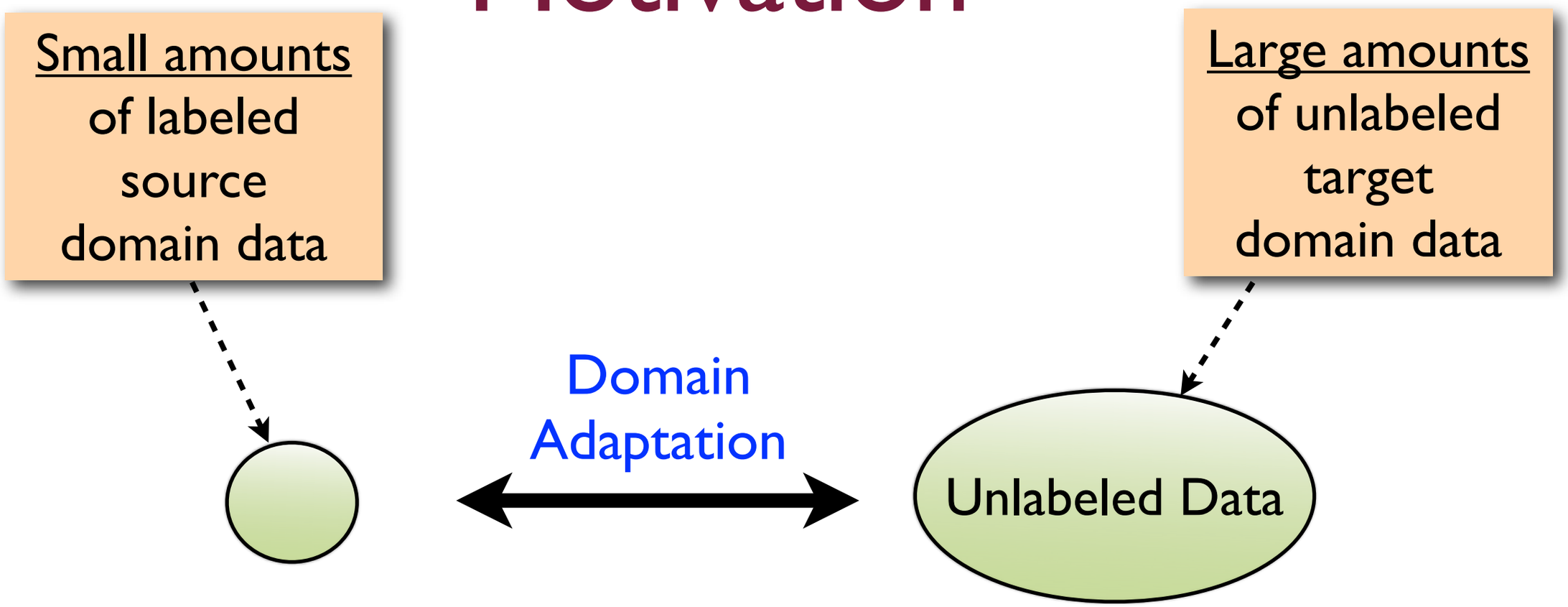
Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Phone Classification
  - Text Categorization
  - Dialog Act Tagging
  - Statistical Machine Translation
  - POS Tagging
    [Subramanya et. al., EMNLP 2008]
  - MultiLingual POS Tagging
- Conclusion & Future Work

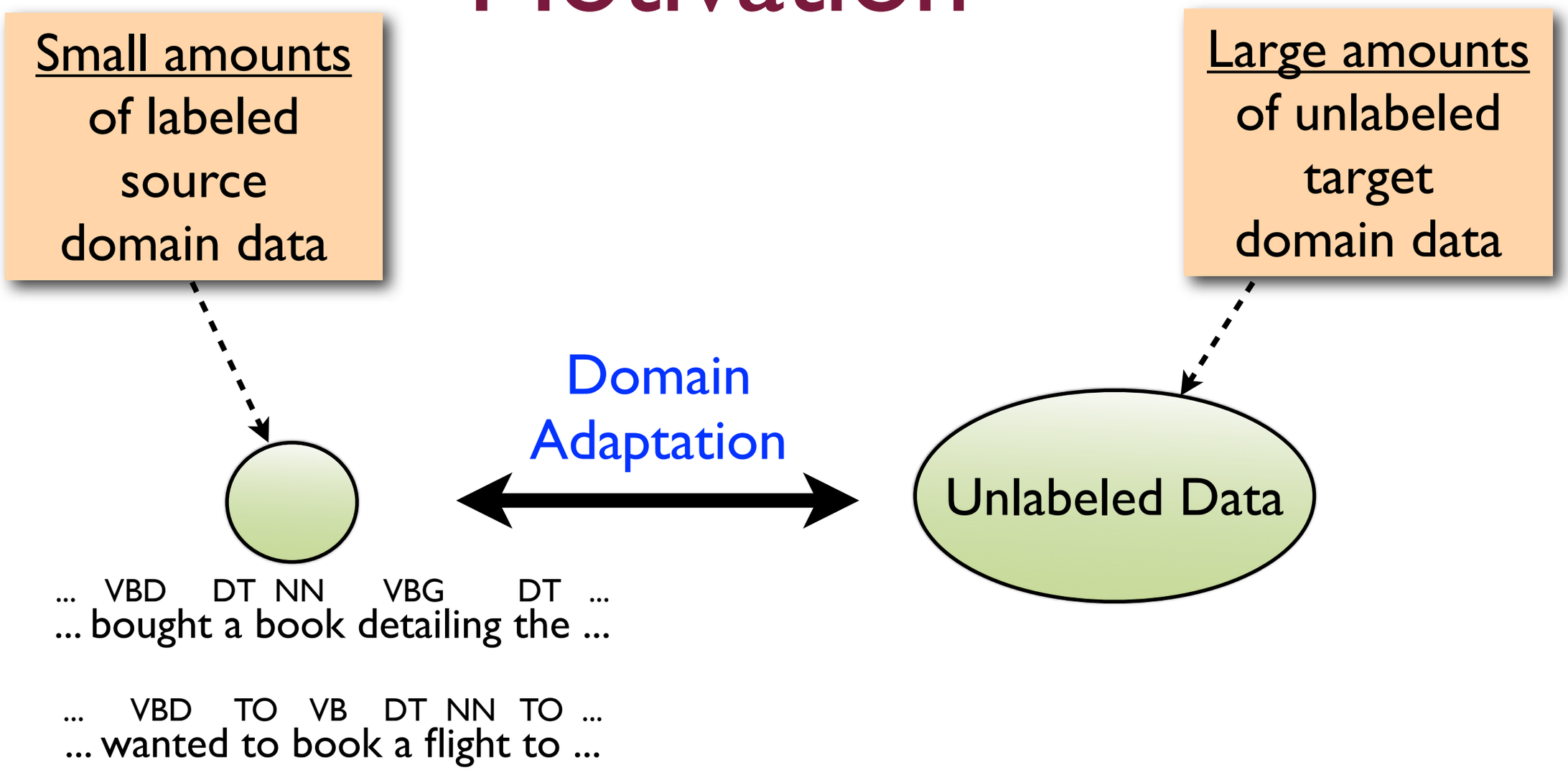
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Unlabeled Data

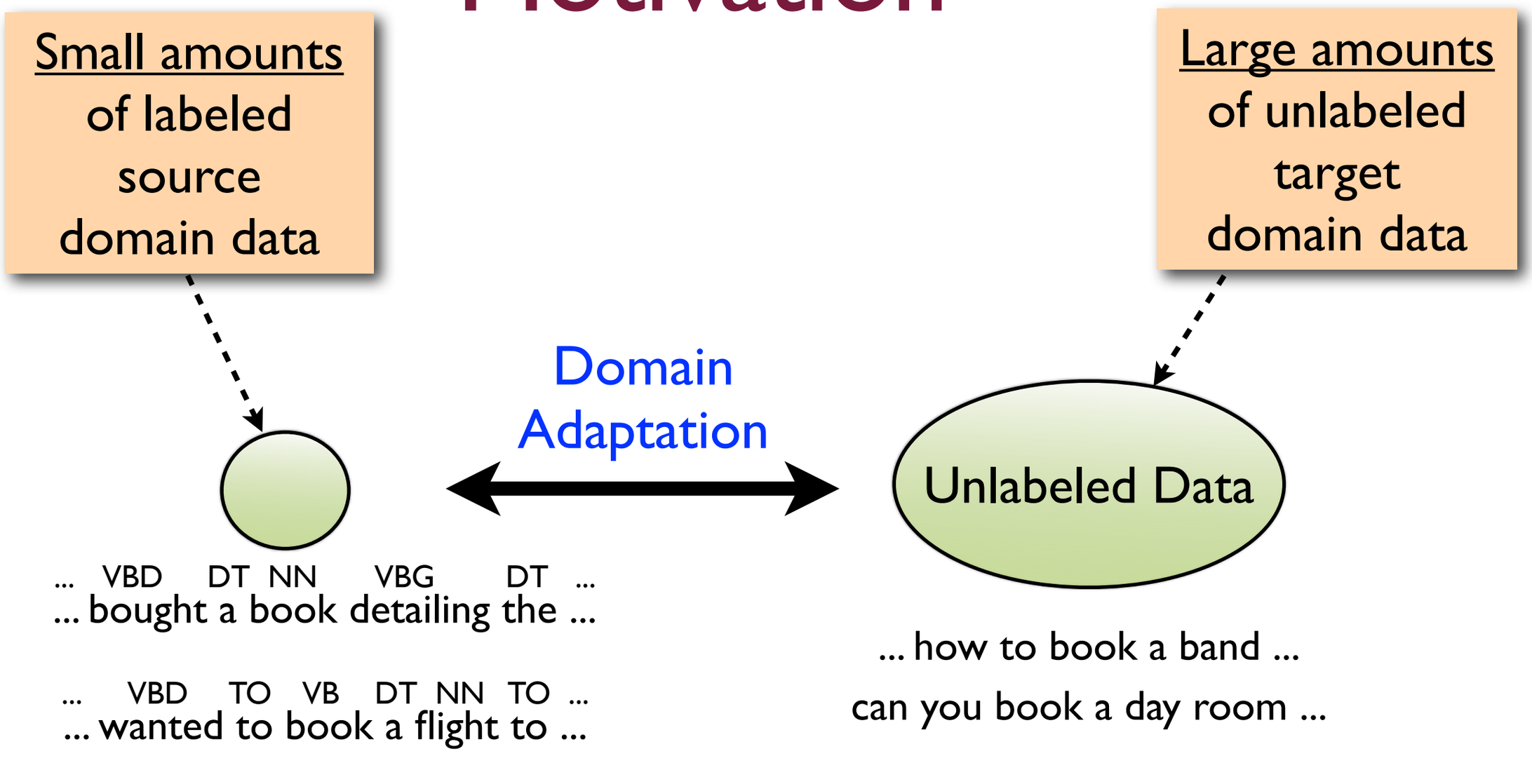
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data

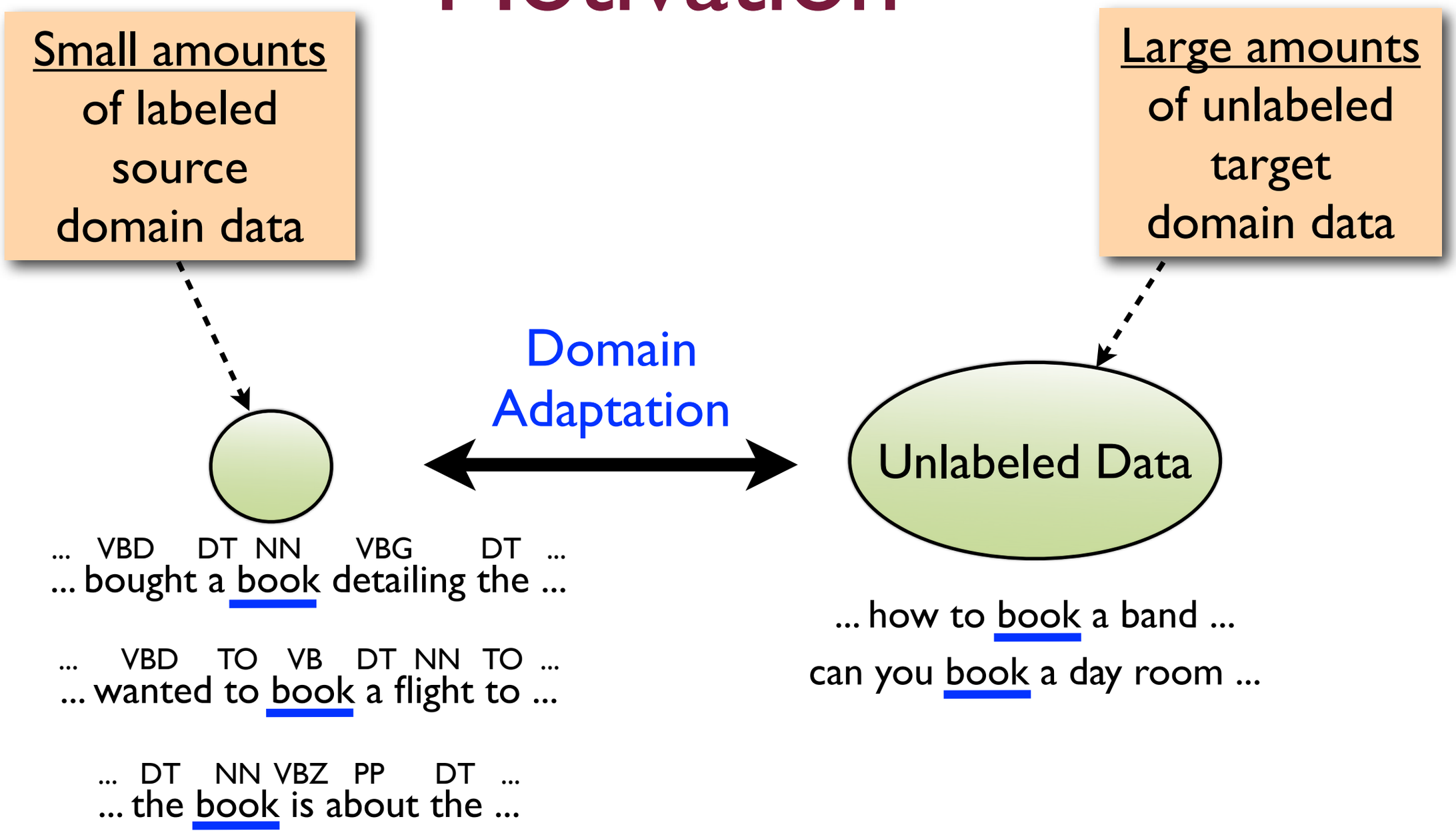
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...

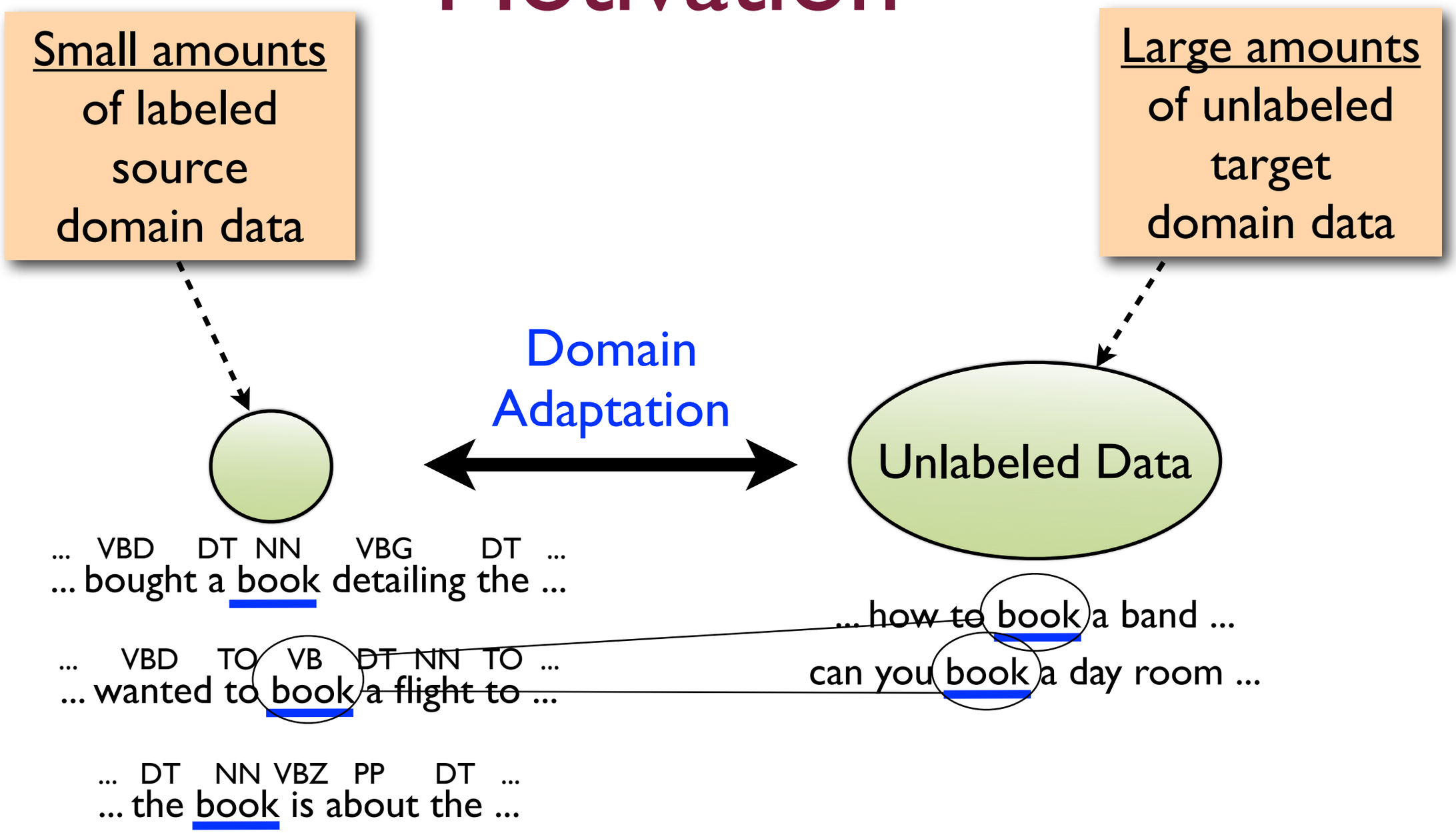
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...

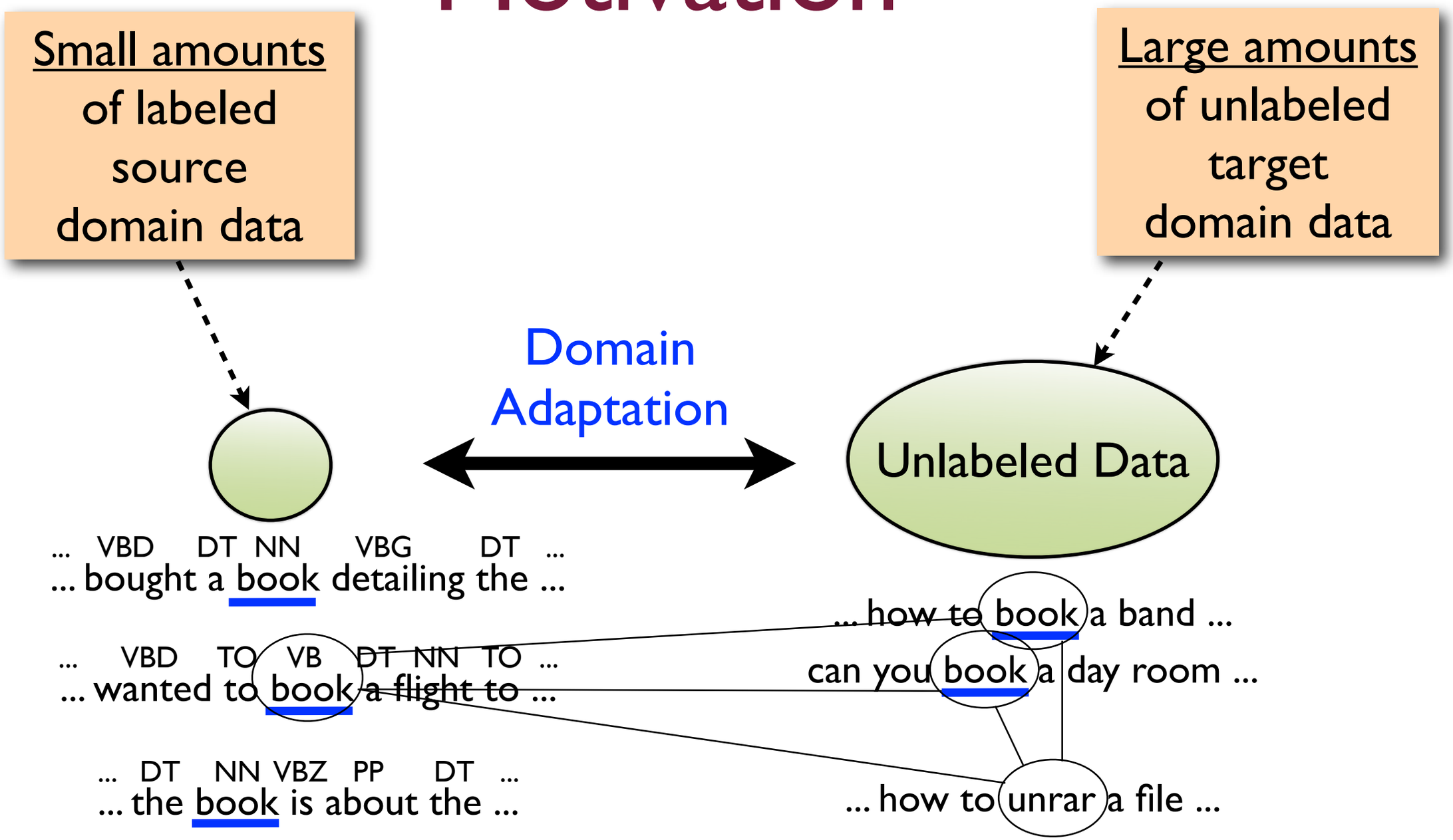
Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...

Motivation


Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...

Motivation

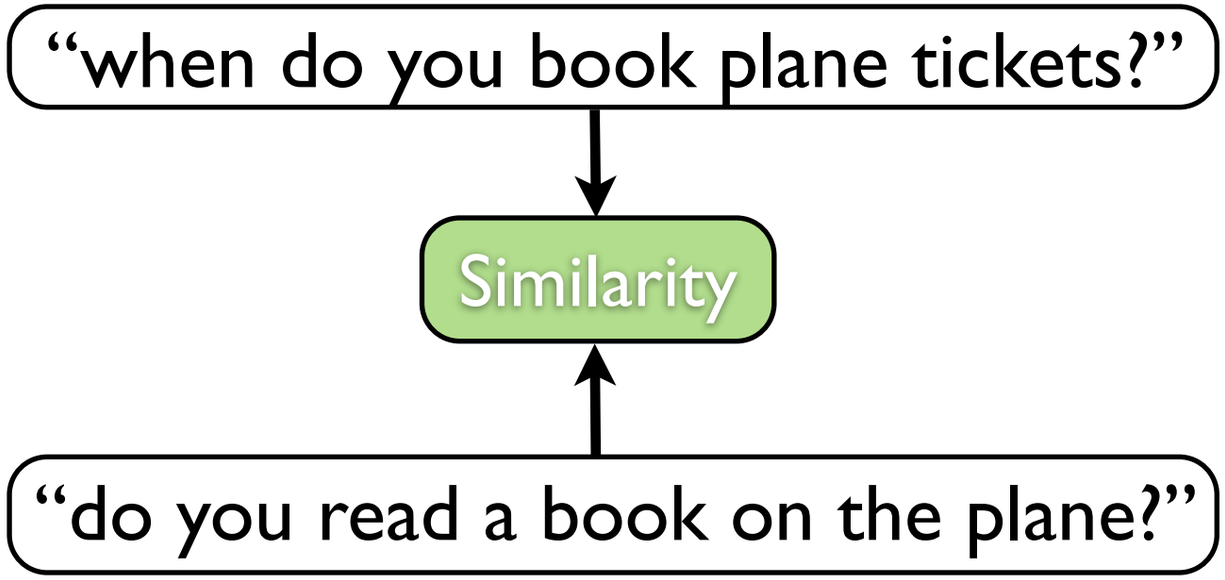

Small amounts of labeled source domain data
Large amounts of unlabeled target domain data
Domain Adaptation
Unlabeled Data
... VBD DT NN VBG DT ...
... bought a book detailing the ...
... VBD TO VB DT NN TO ...
... wanted to book a flight to ...
... DT NN VBZ PP DT ...
... the book is about the ...
... how to book a band ...
can you book a day room ...
... how to unrar a file ...

Graph Construction (I)

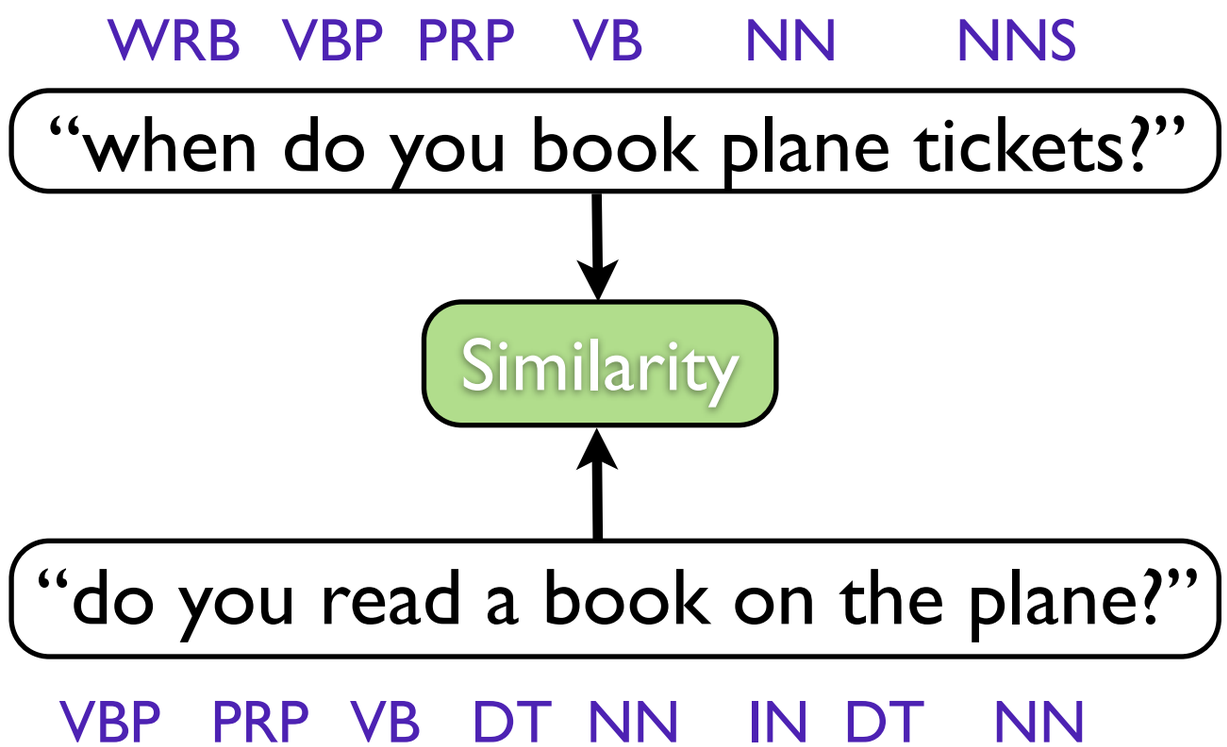
"when do you book plane tickets?"

"do you read a book on the plane?"

Graph Construction (I)



Graph Construction (I)



Graph Construction (II)

can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

how much does it cost to book a band ?

how to get a book agent ?

Graph Construction (II)

can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

how much does it cost to book a band ?

how to get a book agent ?

Graph Construction (II)

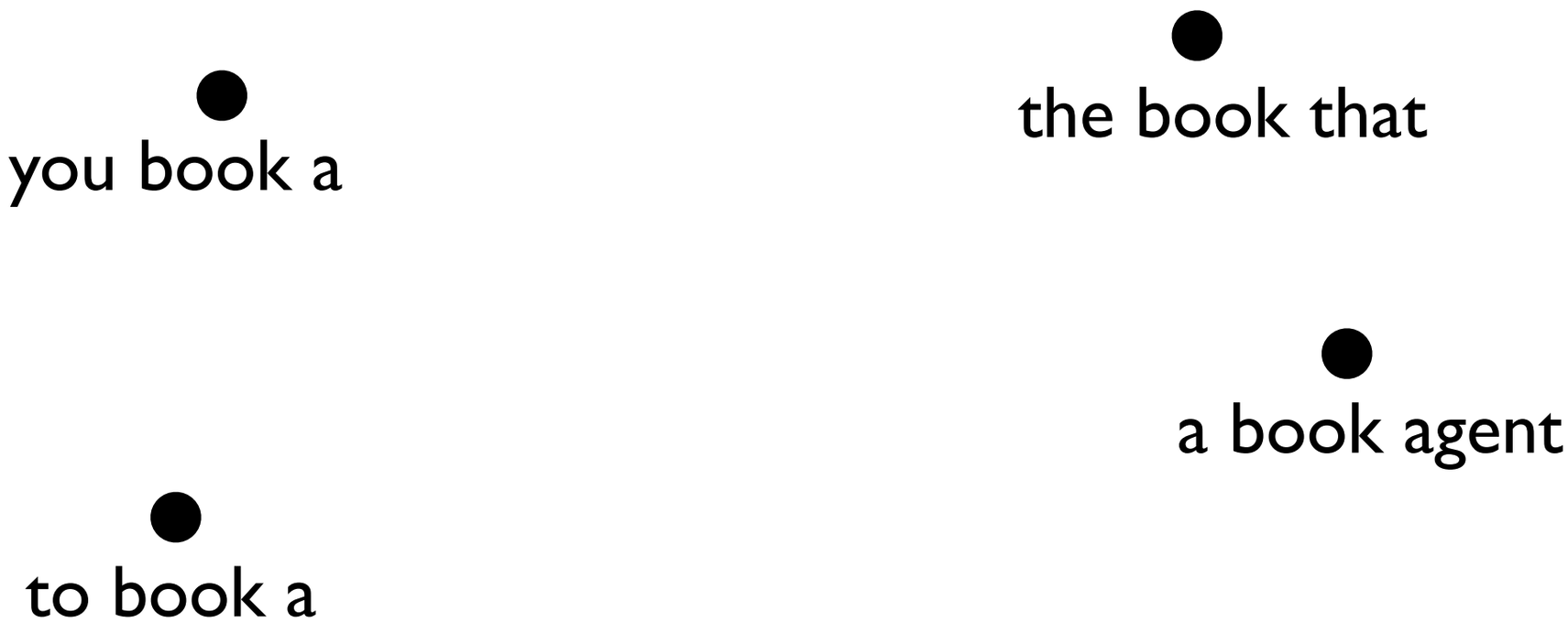
can you book a day room at hilton hawaiian village ?

what was the book that has no letter e ?

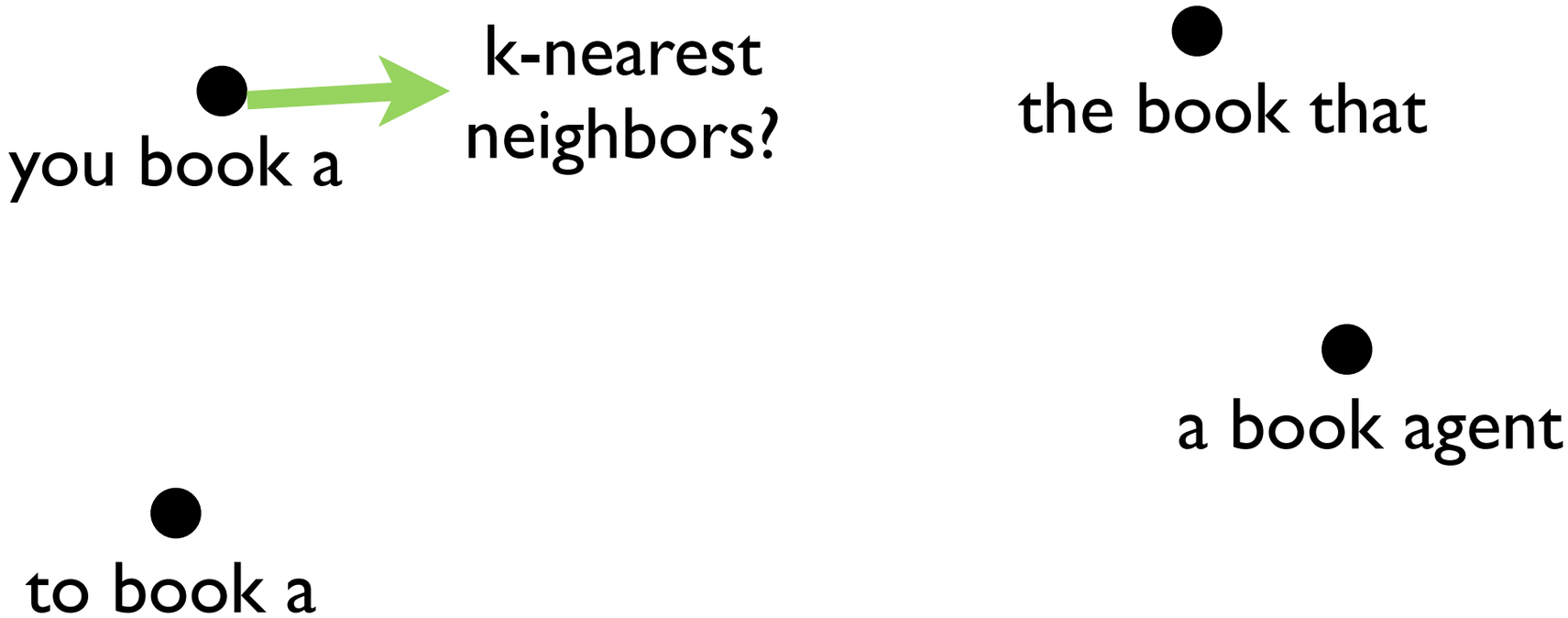
how much does it cost to book a band ?

how to get a book agent ?

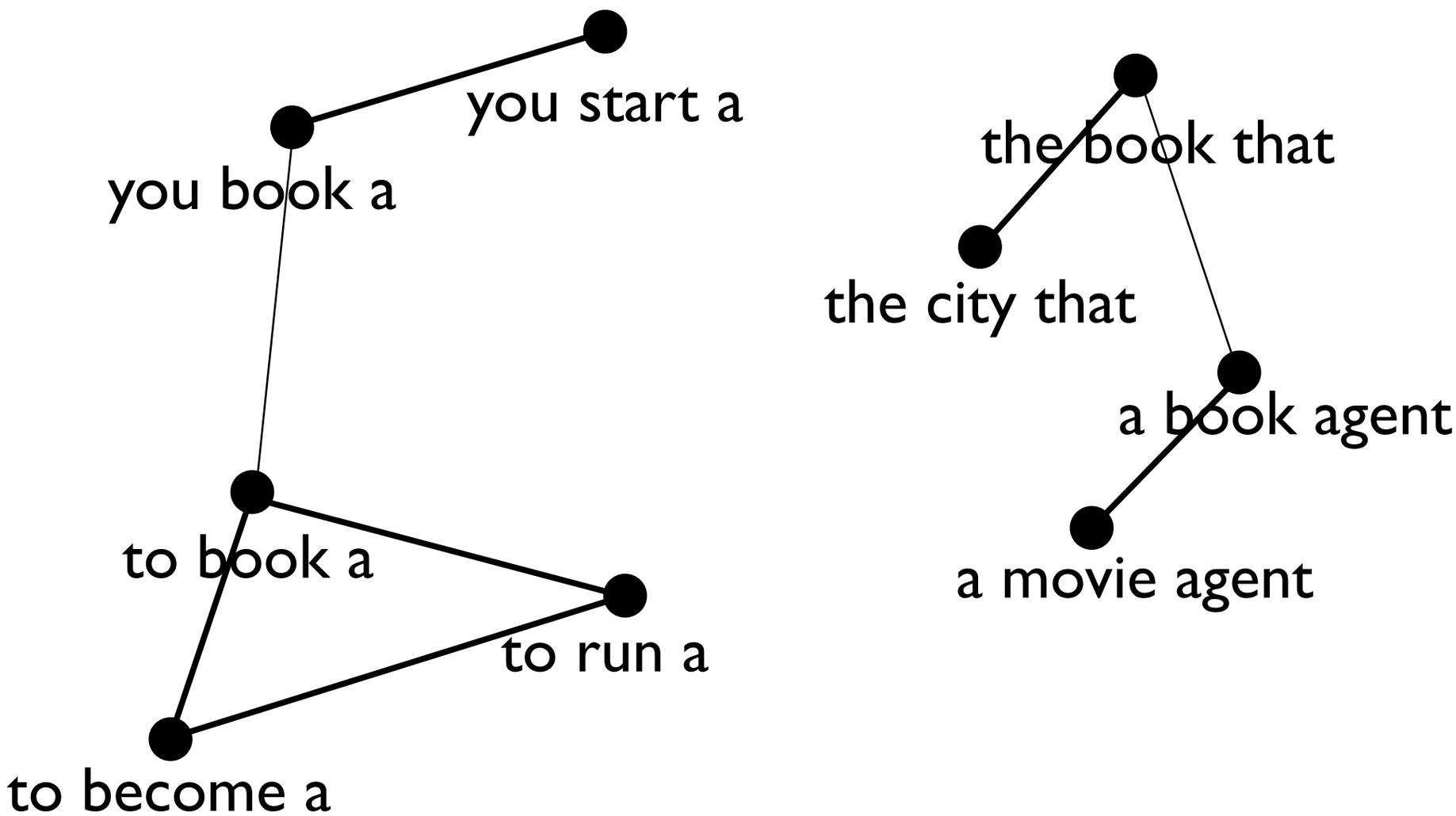
Graph Construction (II)



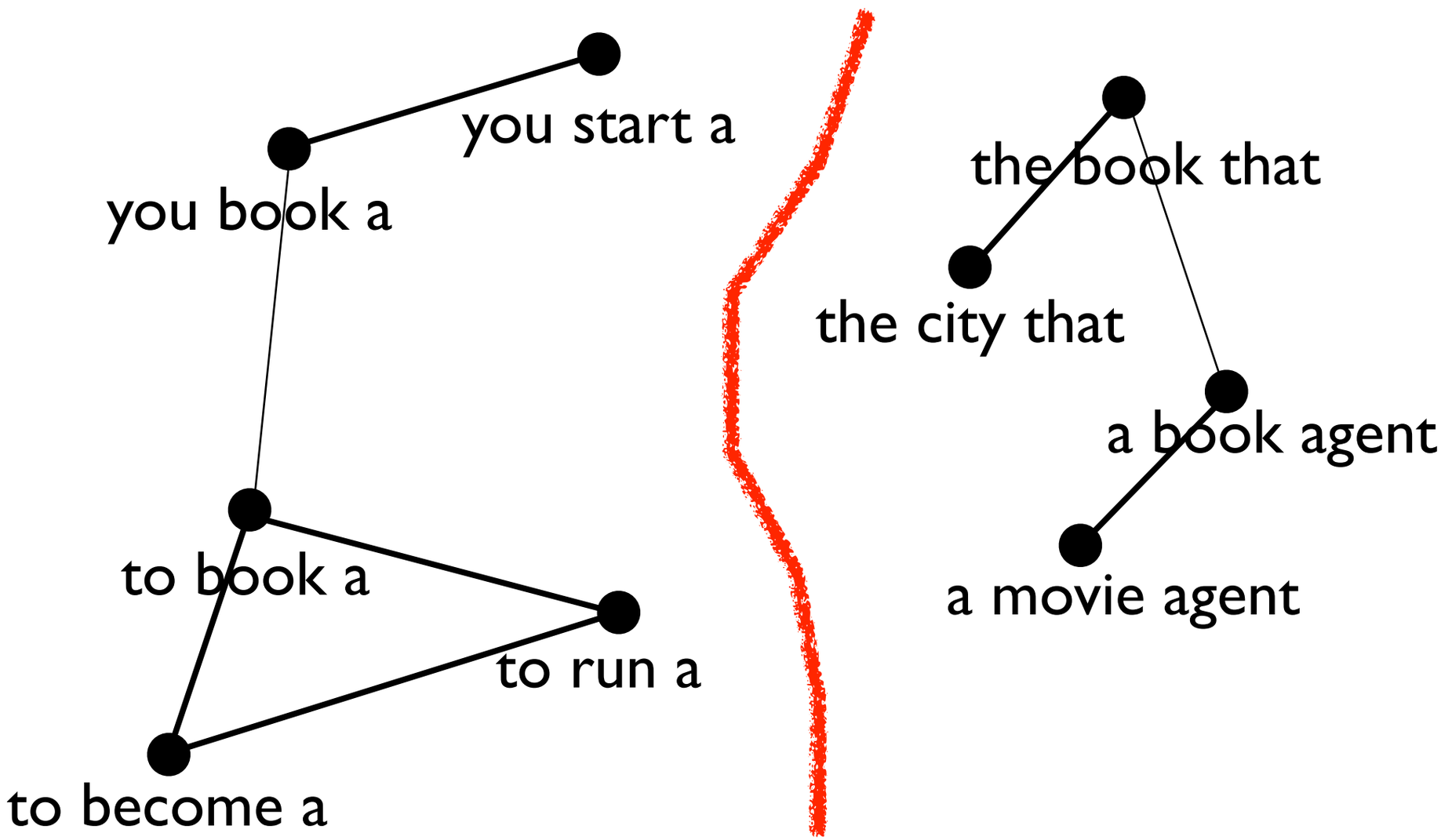
Graph Construction (II)



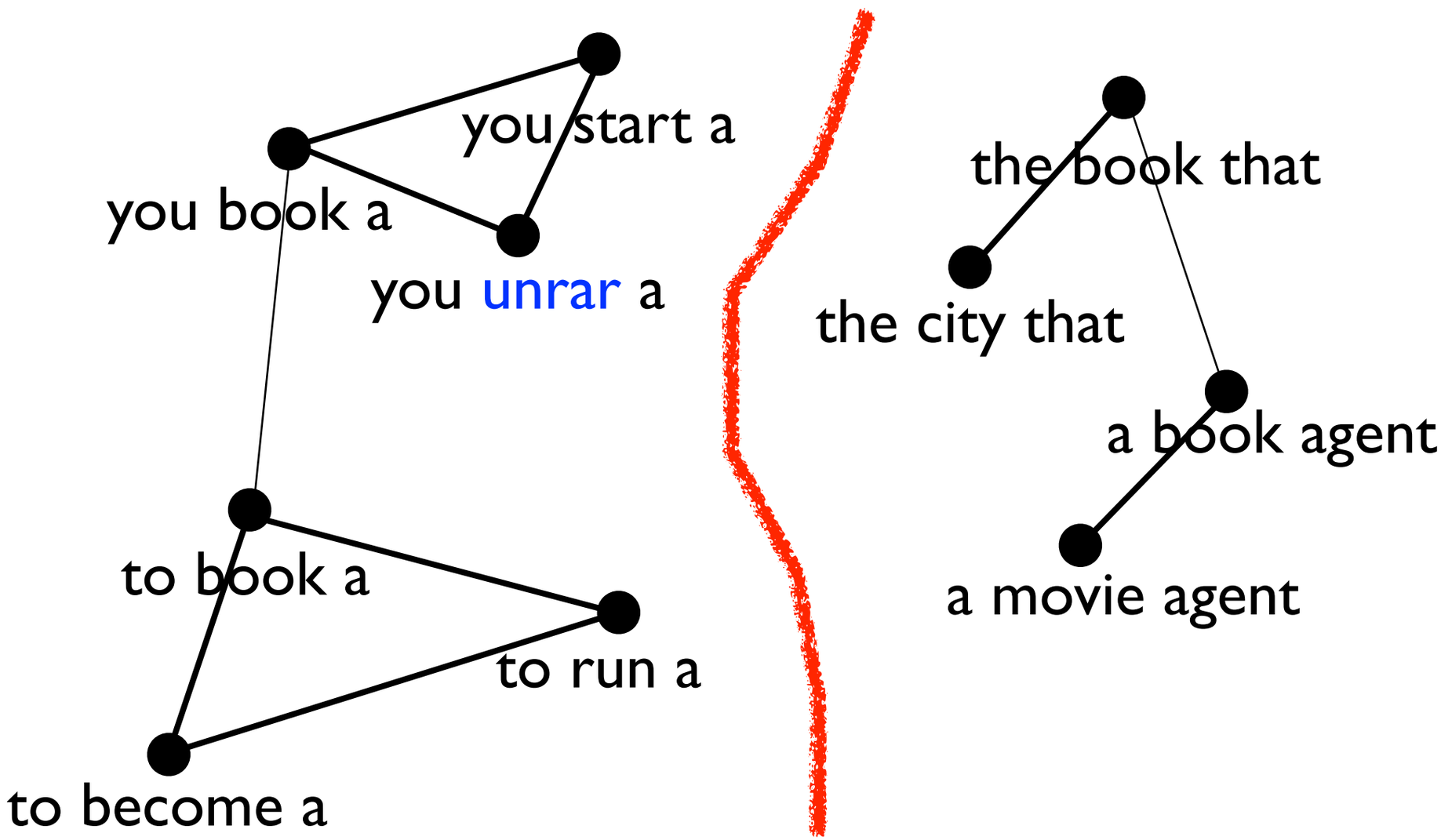
Graph Construction (III)



Graph Construction (III)



Graph Construction (III)



Graph Construction - Features

how much does it cost to book a band ?

Graph Construction - Features

how much does it cost to book a band ?

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band
Center Word	book

Graph Construction - Features

how much does it cost to book a band ?

Trigram + Context	cost to book a band
Left Context	cost to
Right Context	a band
Center Word	book
Trigram - Center Word	to _____ a
Left Word + Right Context	to _____ a band
Left Context + Right Word	cost to _____ a
Suffix	none

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

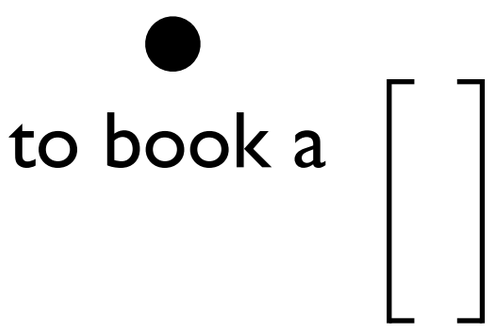
Graph Construction - Features

how much to book a flight to paris?

how much does it cost to book a band ?

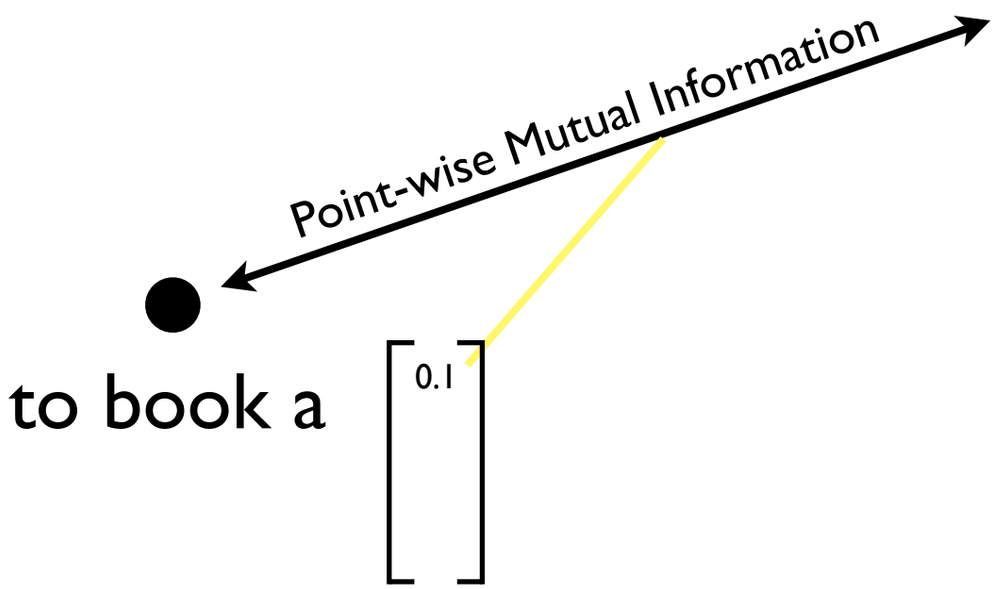
Graph Construction - Features

Graph Construction - Features



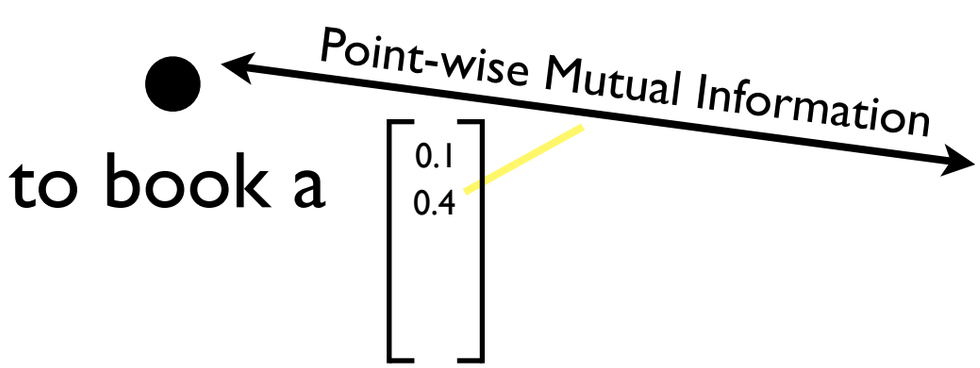
Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features



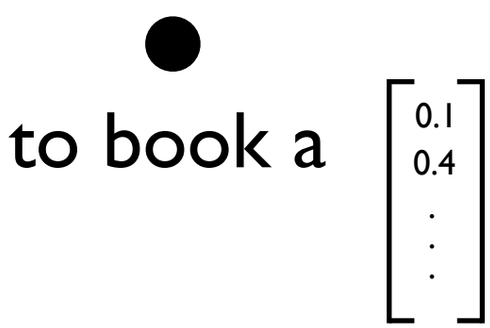
Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features



Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Graph Construction - Features



Trigram + Context
Left Context
Right Context
Center Word
Trigram - Center Word
Left Word + Right Context
Left Context + Right Word
Suffix

Similarity Function

to book a $\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}$

Cosine Similarity (,) = 0.56

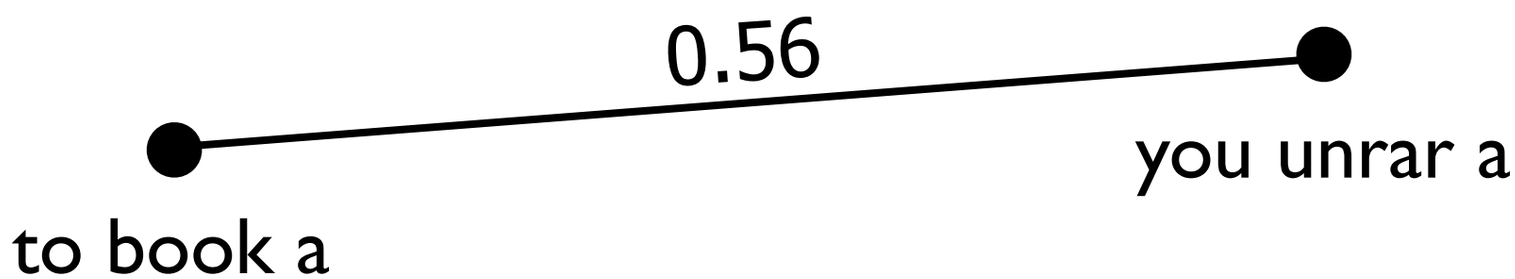
Similarity Function

to book a $\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}$

you unrar a

Cosine Similarity (,) = 0.56

Similarity Function



$$\text{Cosine Similarity}\left(\begin{bmatrix} 0.1 \\ 0.4 \\ \vdots \end{bmatrix}, \begin{bmatrix} 0.2 \\ 0.3 \\ \vdots \end{bmatrix} \right) = 0.56$$

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

can you book a day room at hilton hawaiian village ?

how to unrar a zipped file ?

how to get a book agent ?

how do you book a flight to multiple cities ?

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

CRF

can you book a day room at hilton hawaiian village ?

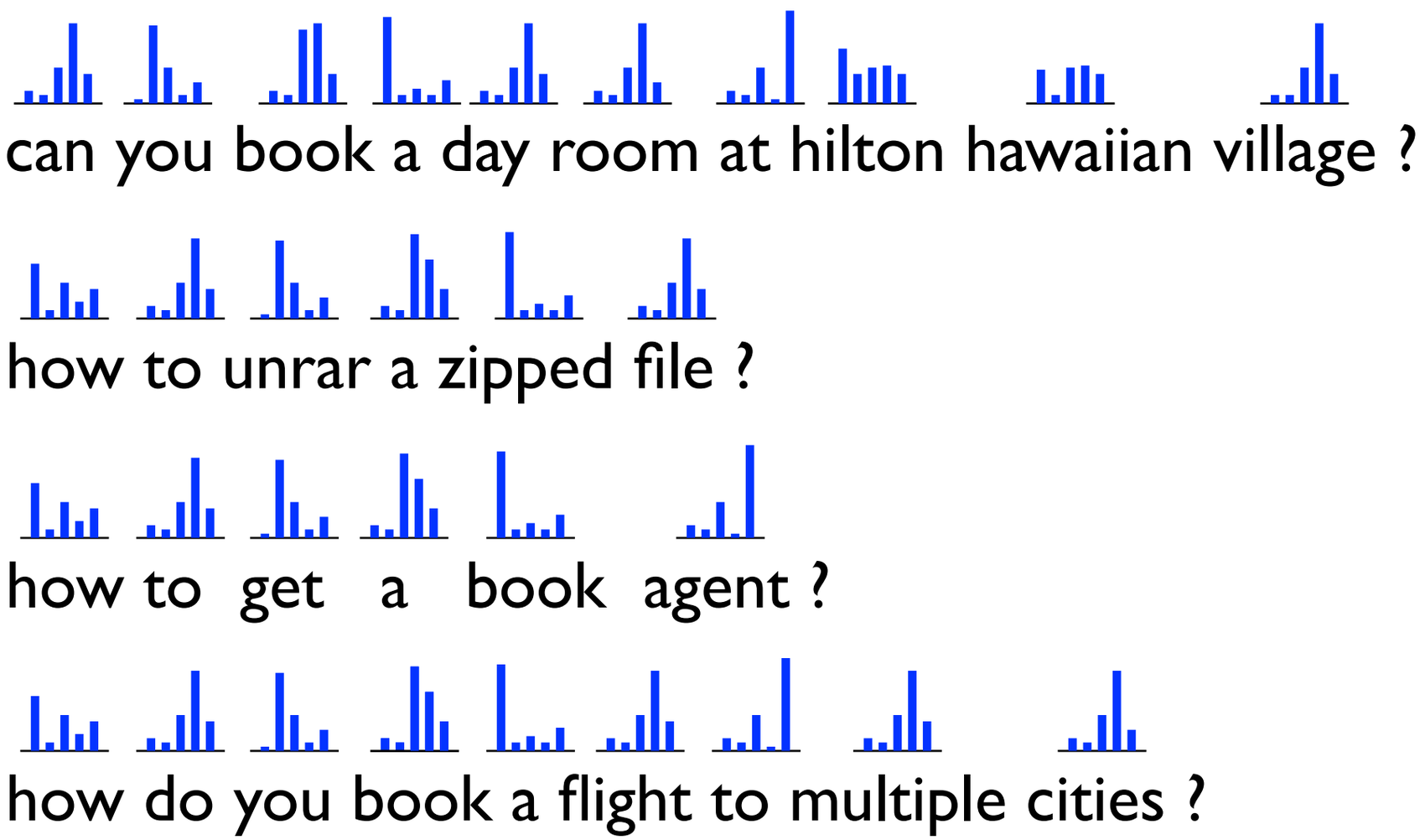
how to unrar a zipped file ?

how to get a book agent ?

how do you book a flight to multiple cities ?

Approach (I)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF

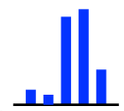


Approach (II)

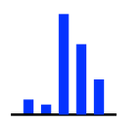
1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)

Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)



can you book a day room at hilton hawaiian village ?



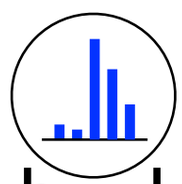
how do you book a flight to multiple cities ?

Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)

can you book a day room at hilton hawaiian village ?

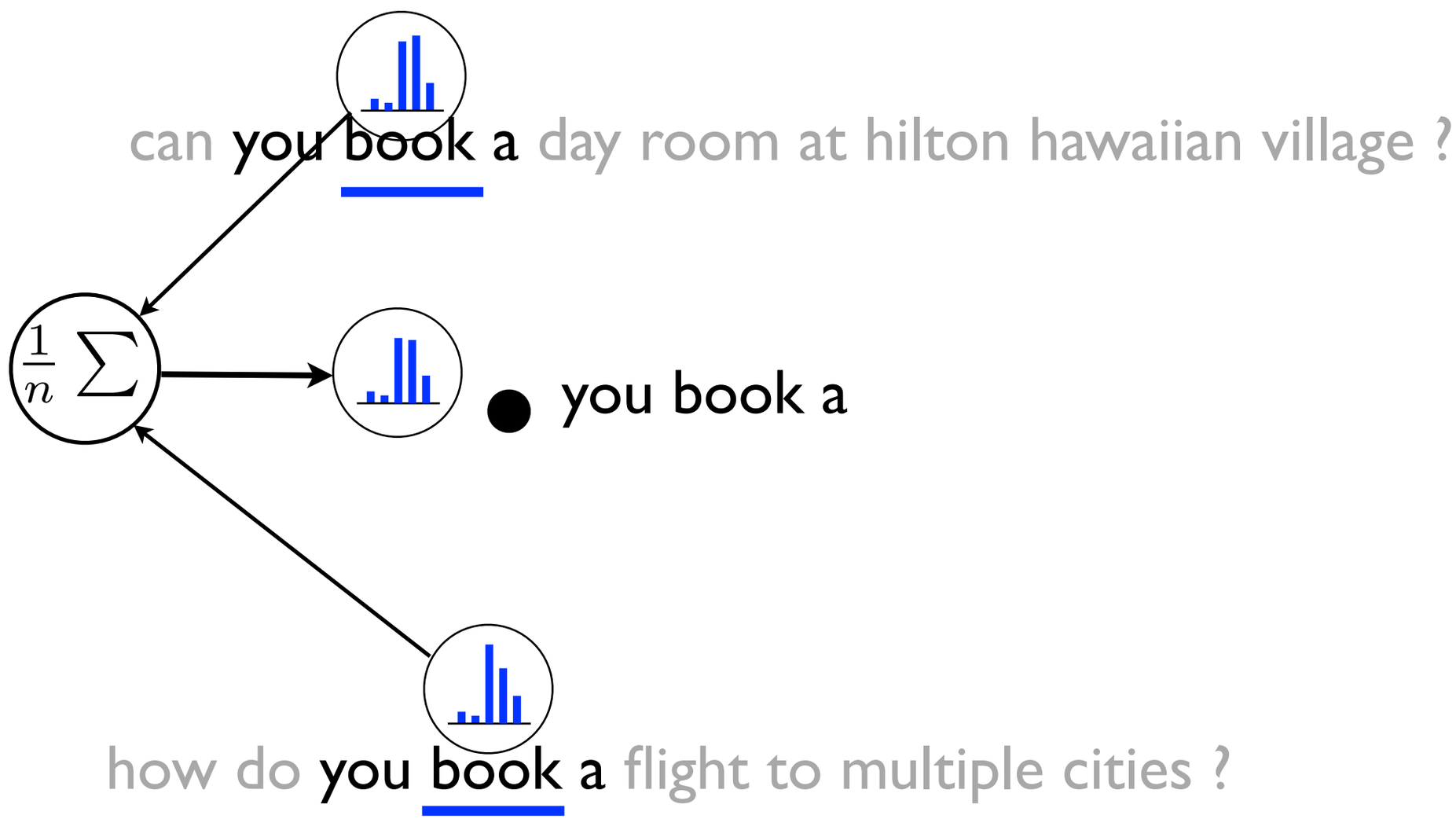
- you book a



how do you book a flight to multiple cities ?

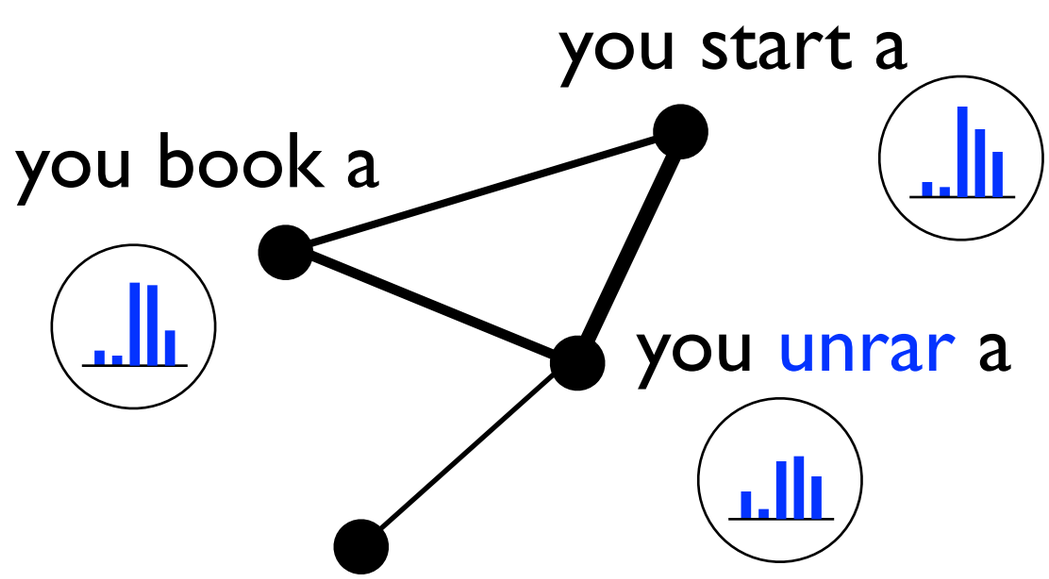
Approach (II)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)



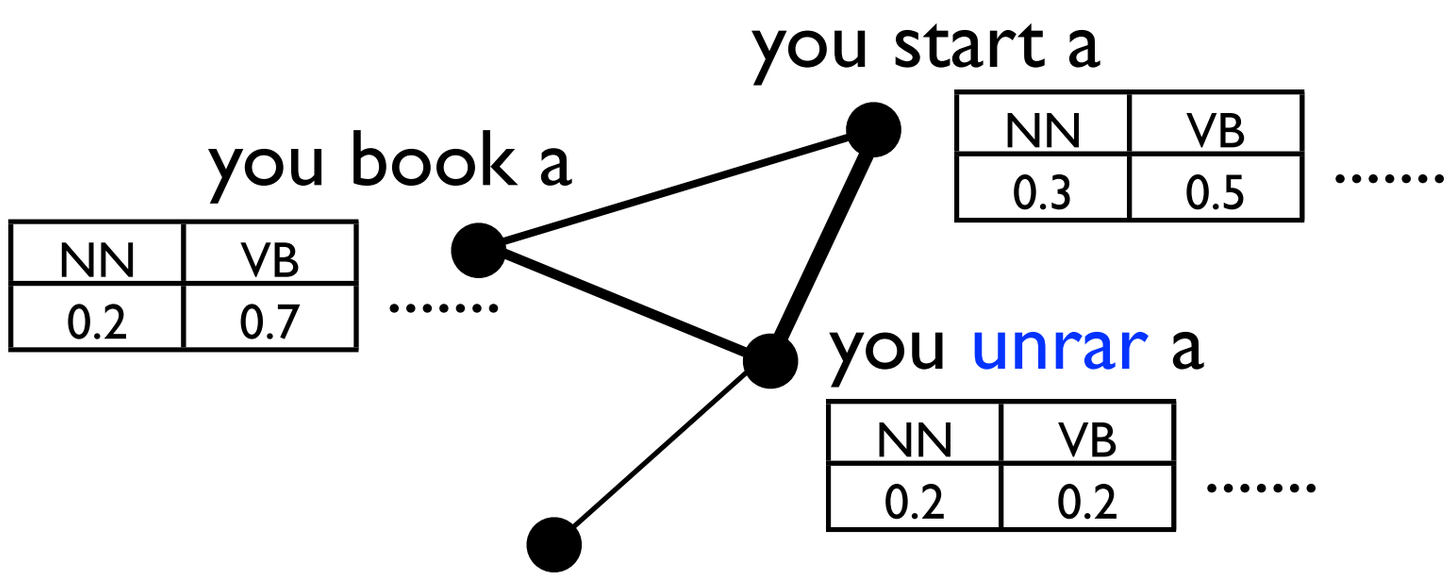
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



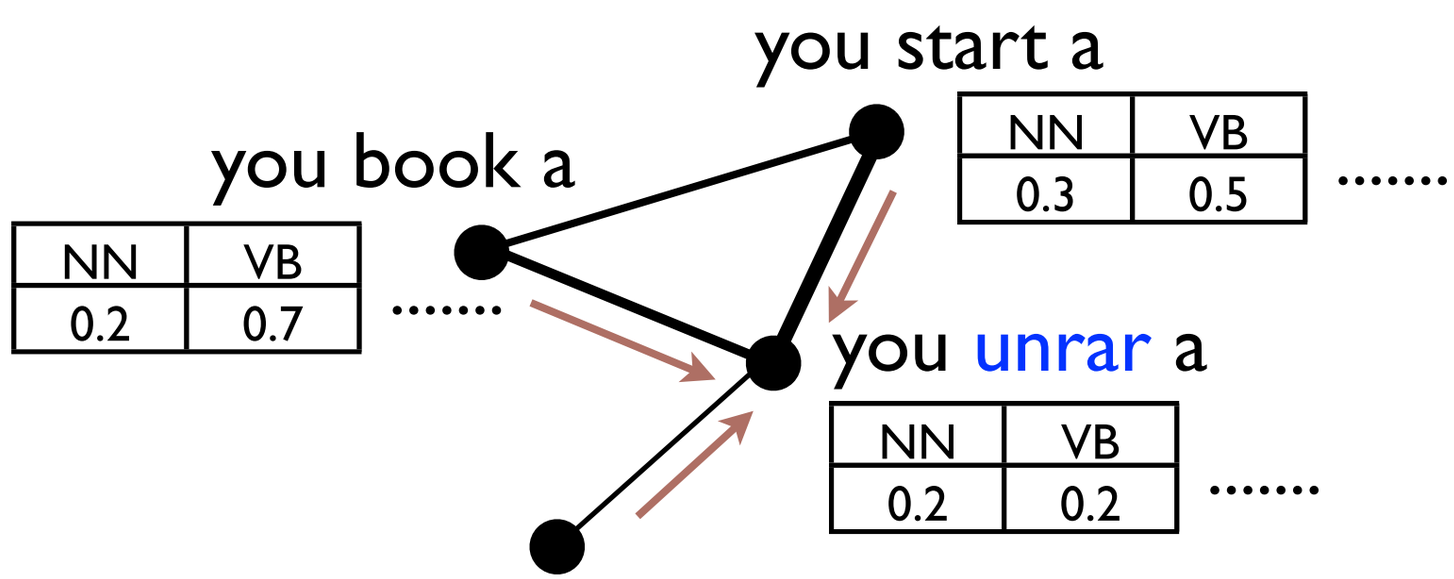
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



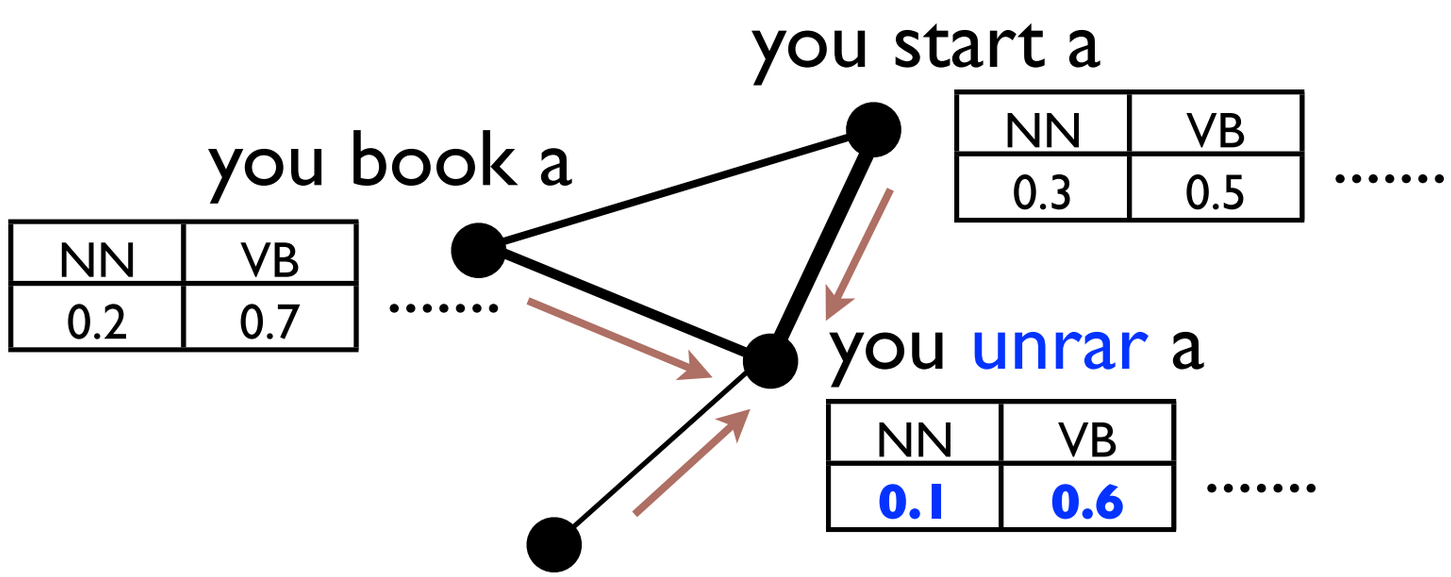
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



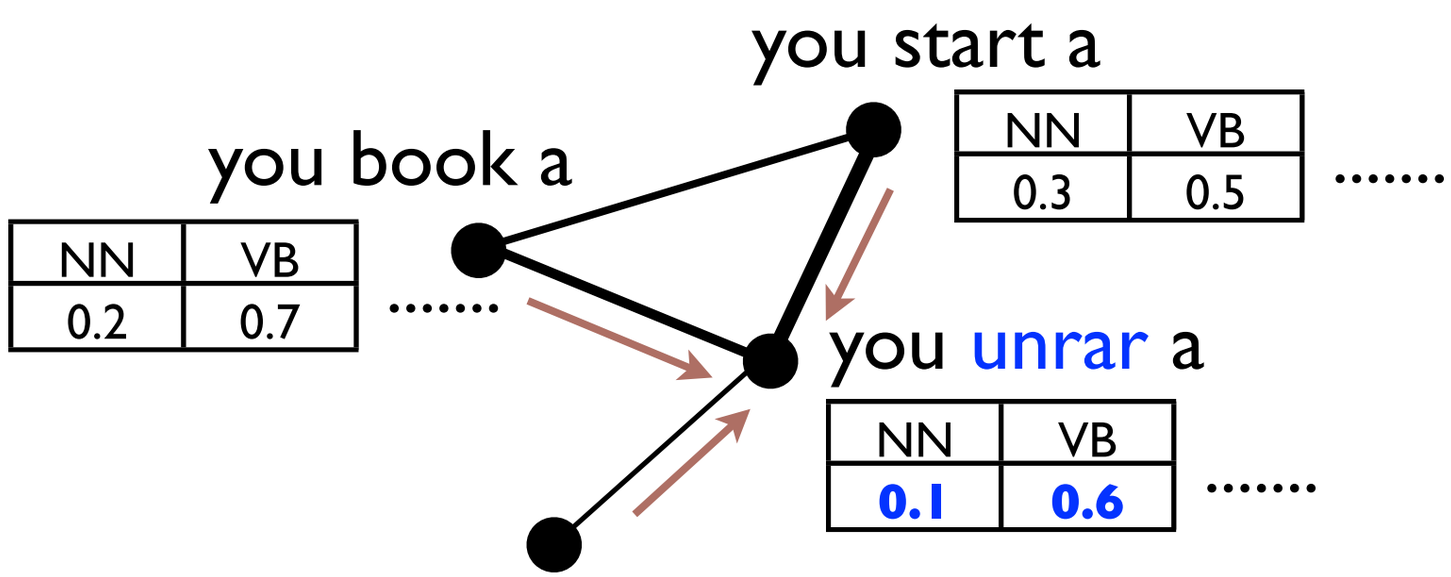
Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



Approach (III)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation



If two n-grams are similar according to the graph then their output distributions should be similar

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Can you unrar a zipped file?

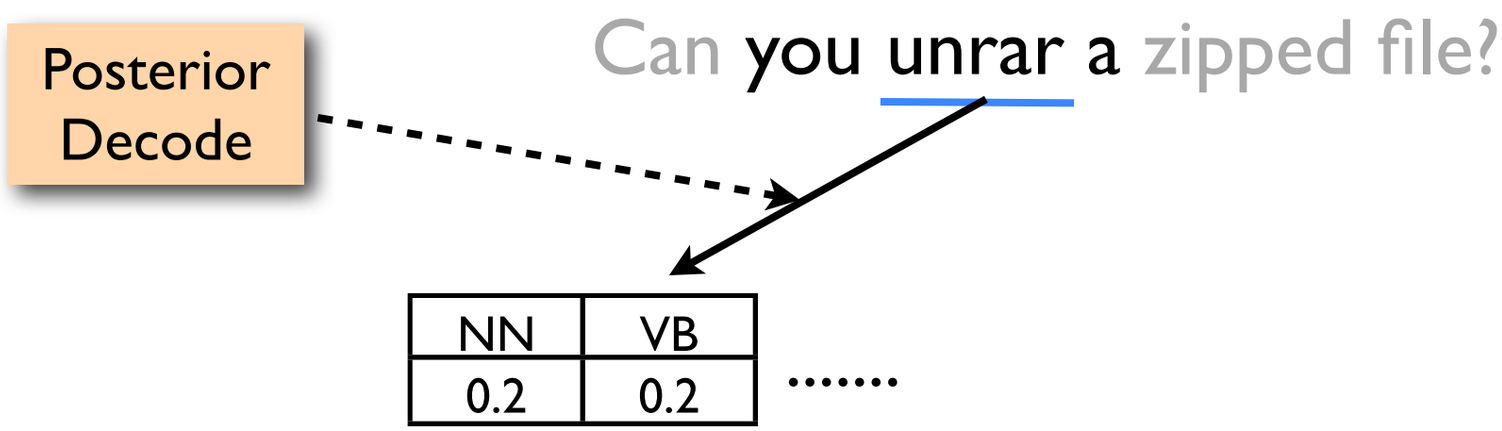
Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode

Can you unrar a zipped file?

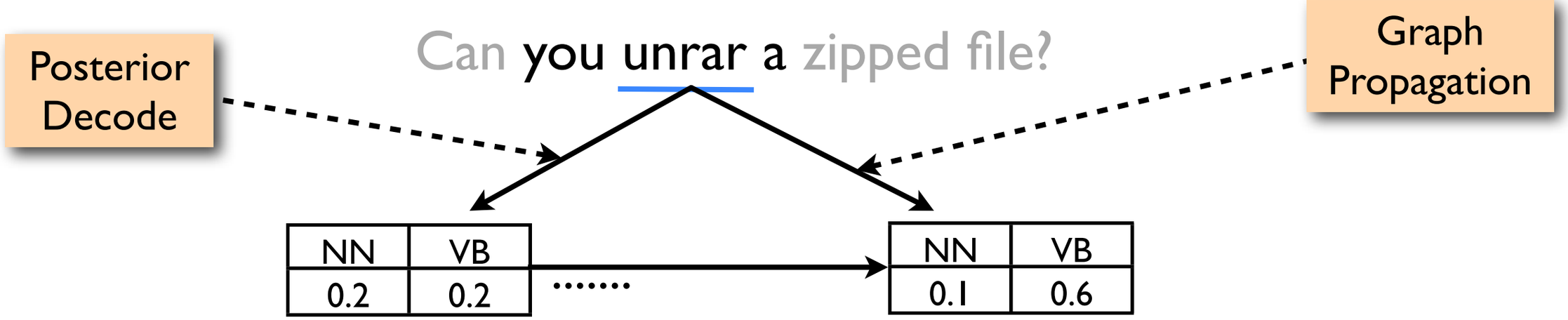
Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode



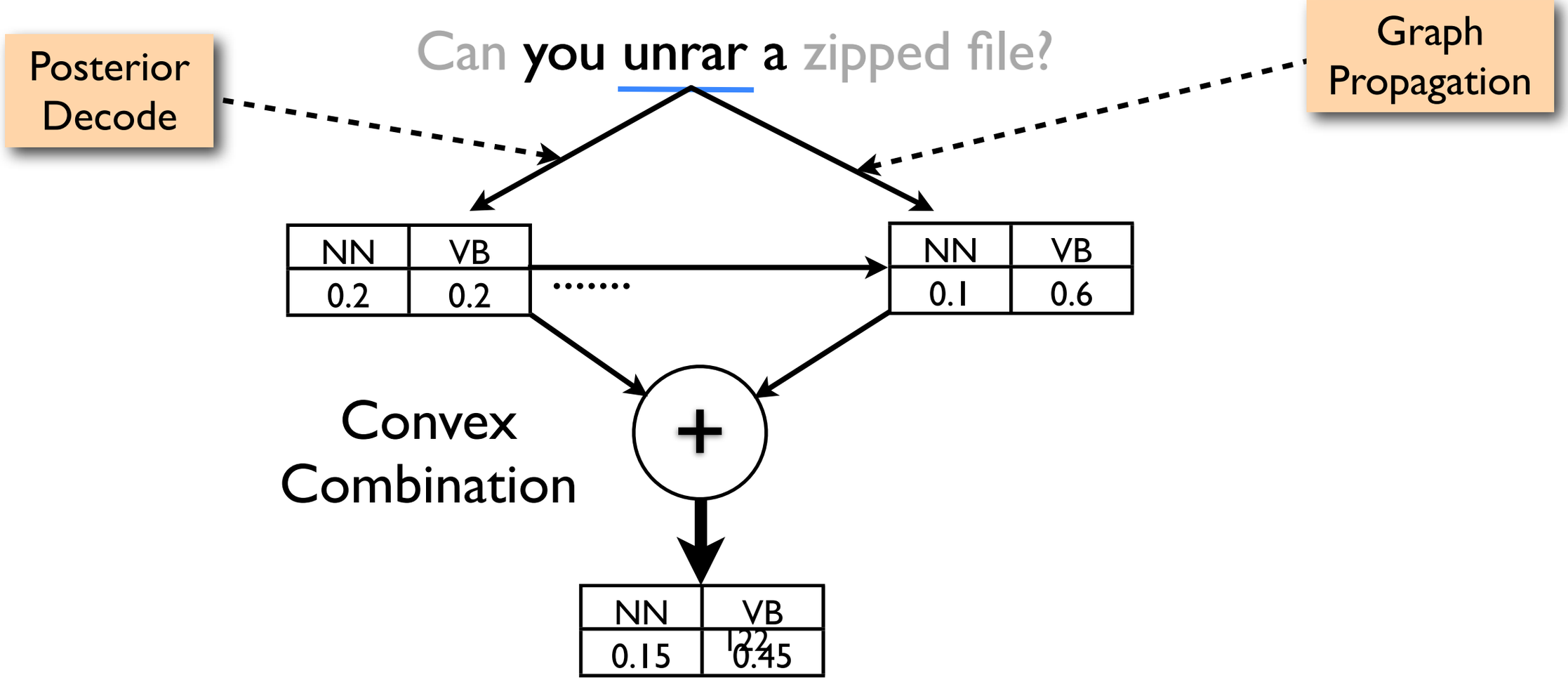
Approach (IV)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode



Approach (IV)

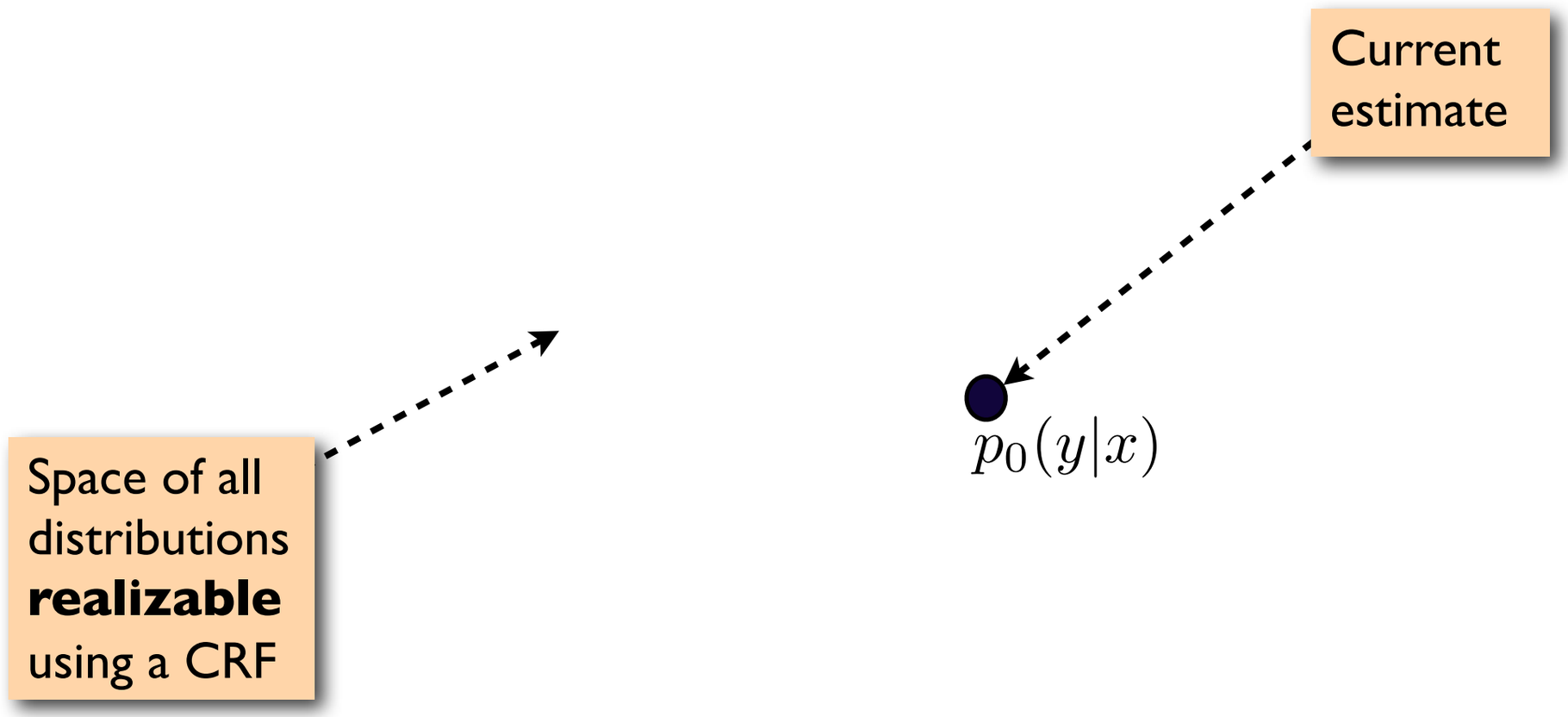
1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode



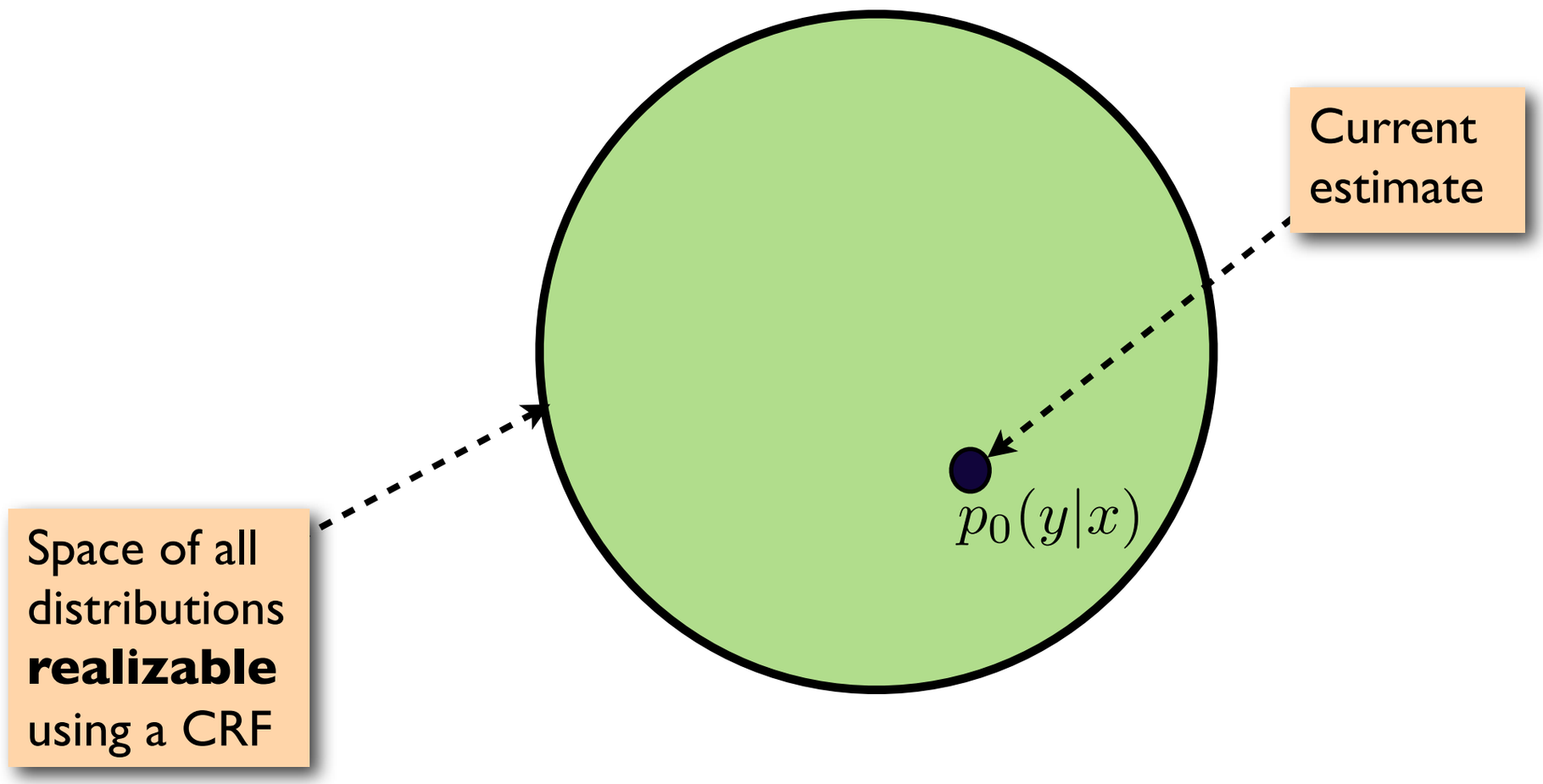
Approach (V)

1. Train a CRF on labeled data
2. While not converged do:
   2.1. Posterior decode **unlabeled data** using CRF
   2.2. Aggregate posteriors (token-to-type mapping)'
   2.3. Graph propagation
   2.4. Viterbi Decode
   2.5. Retrain CRF on labeled & **automatically labeled** unlabeled data

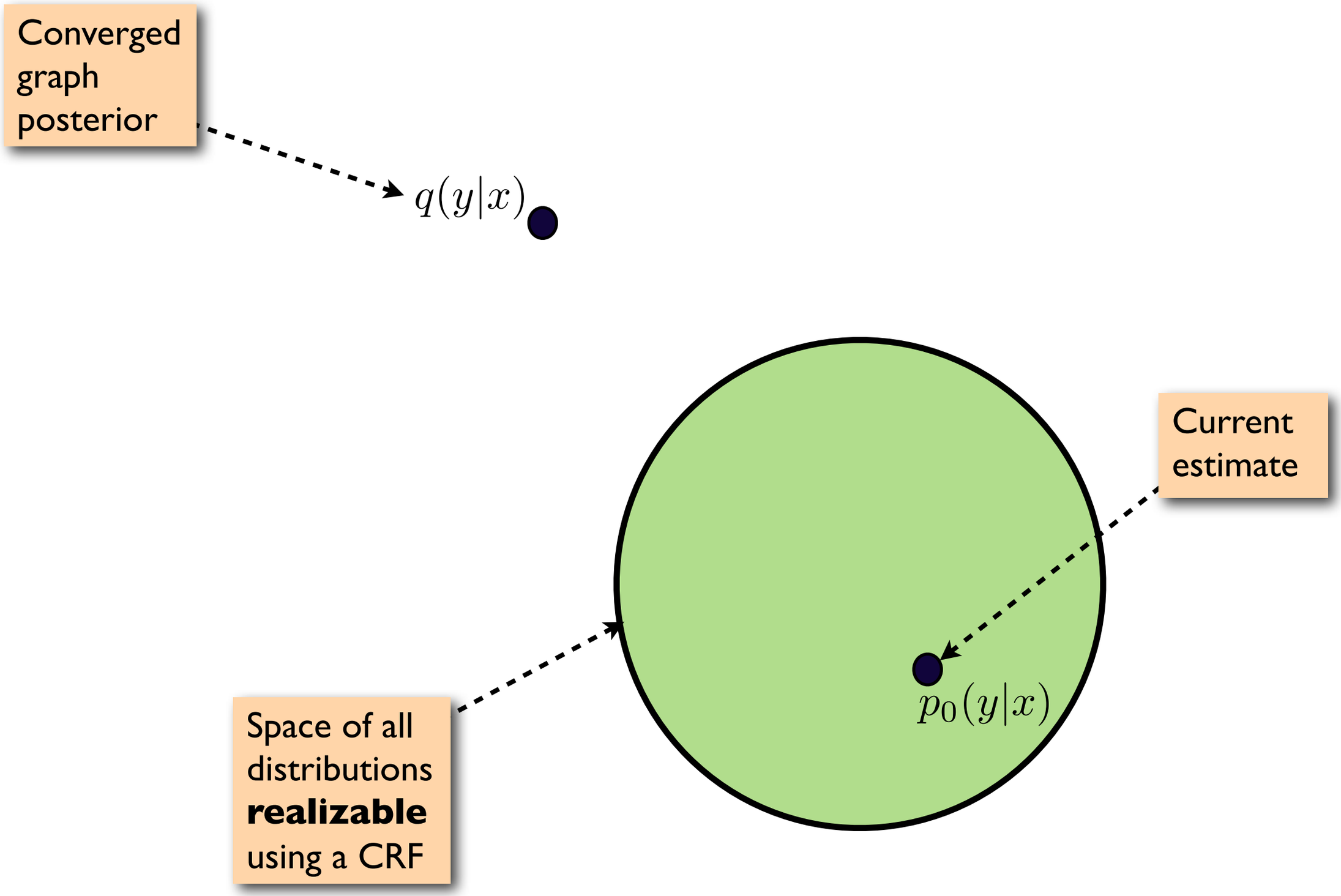
Viterbi Decoding : Intuition



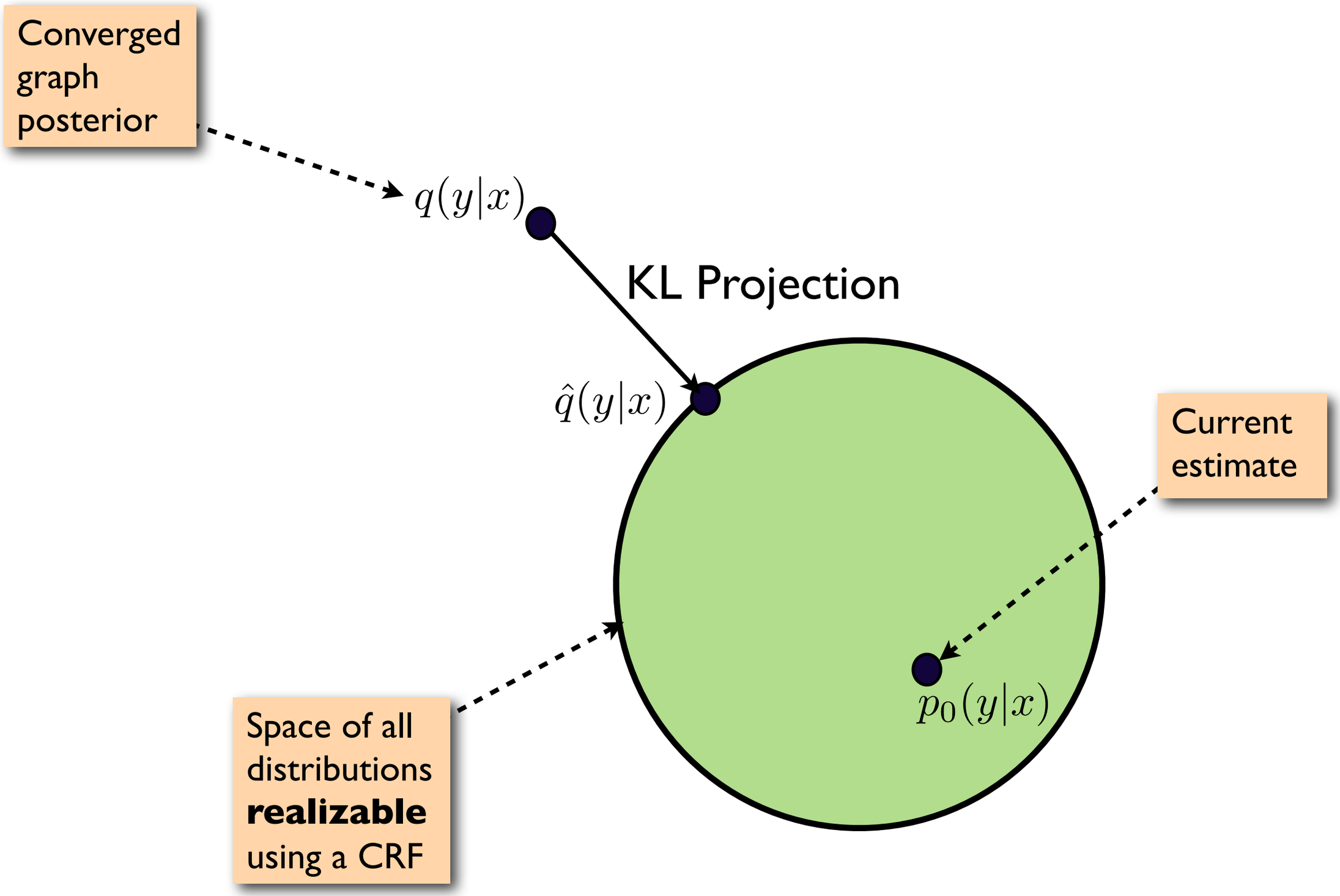
Viterbi Decoding : Intuition



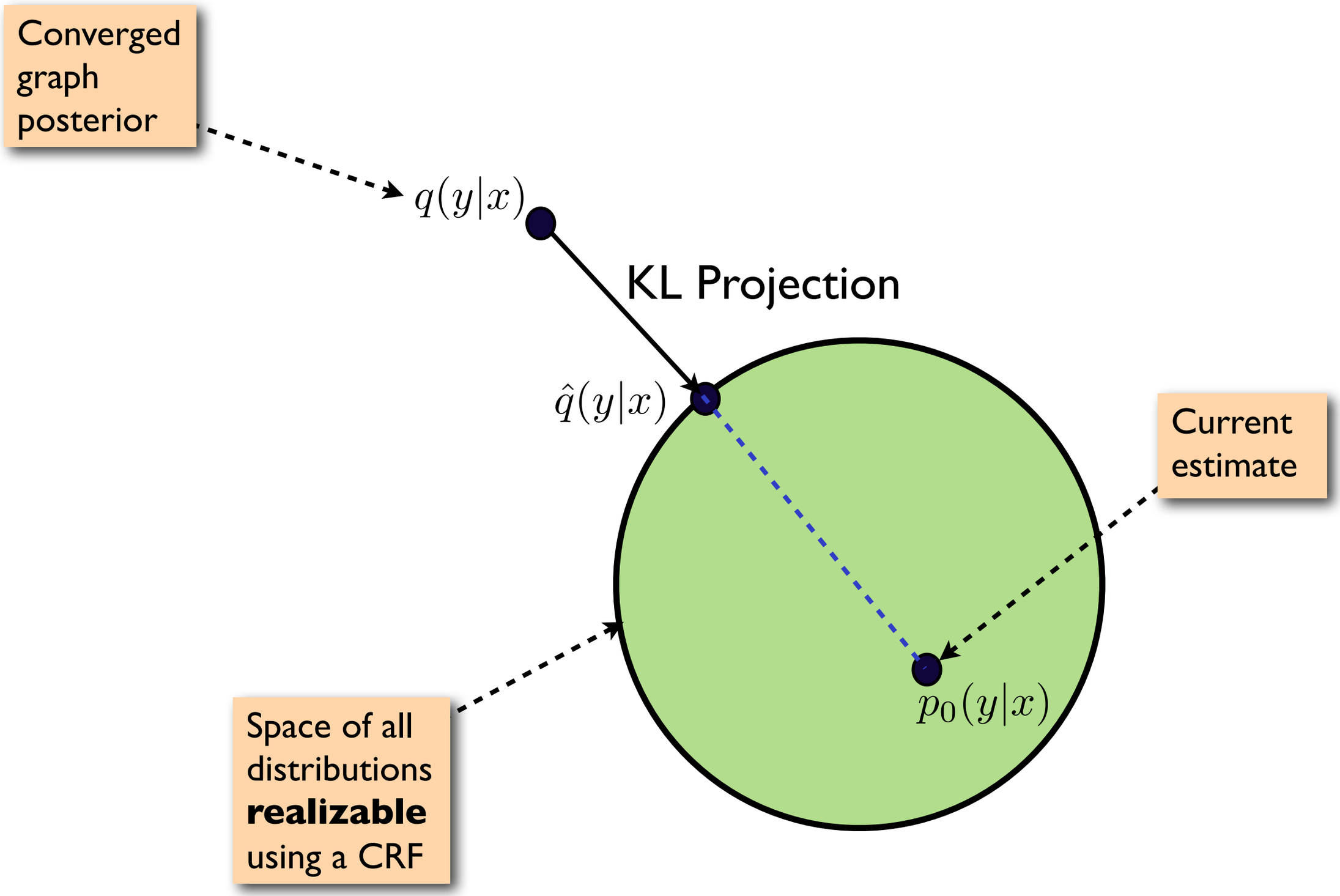
Viterbi Decoding : Intuition



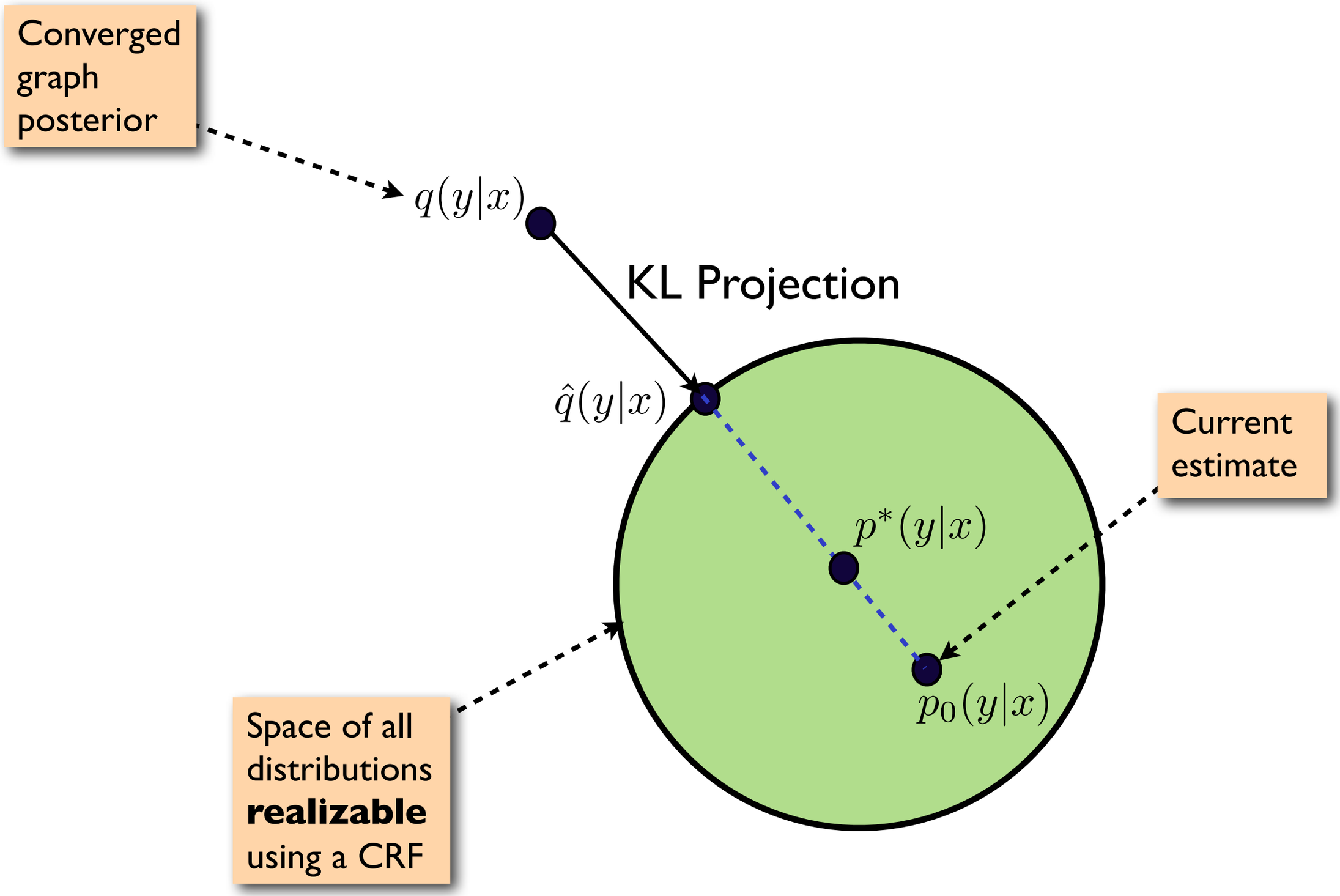
Viterbi Decoding : Intuition



Viterbi Decoding : Intuition



Viterbi Decoding : Intuition

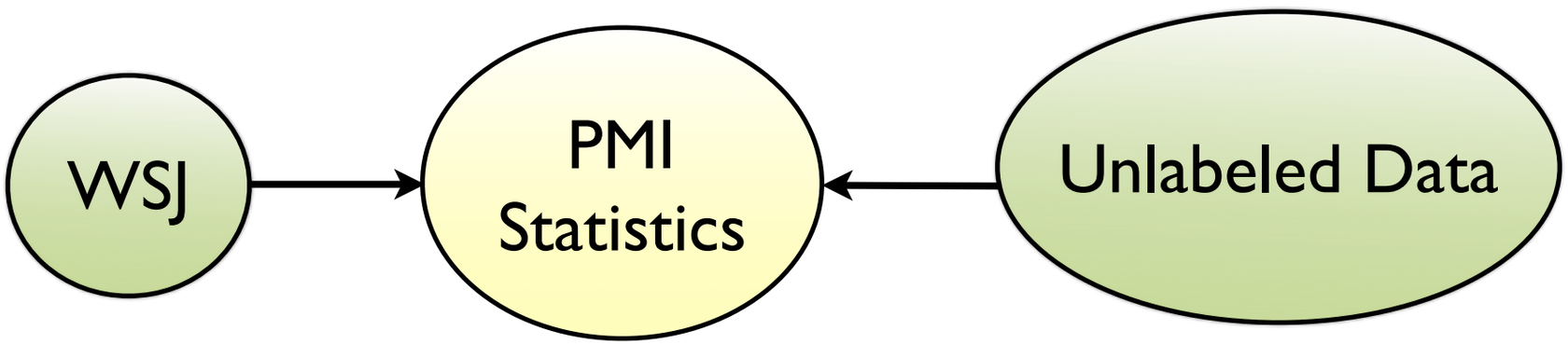


Corpora

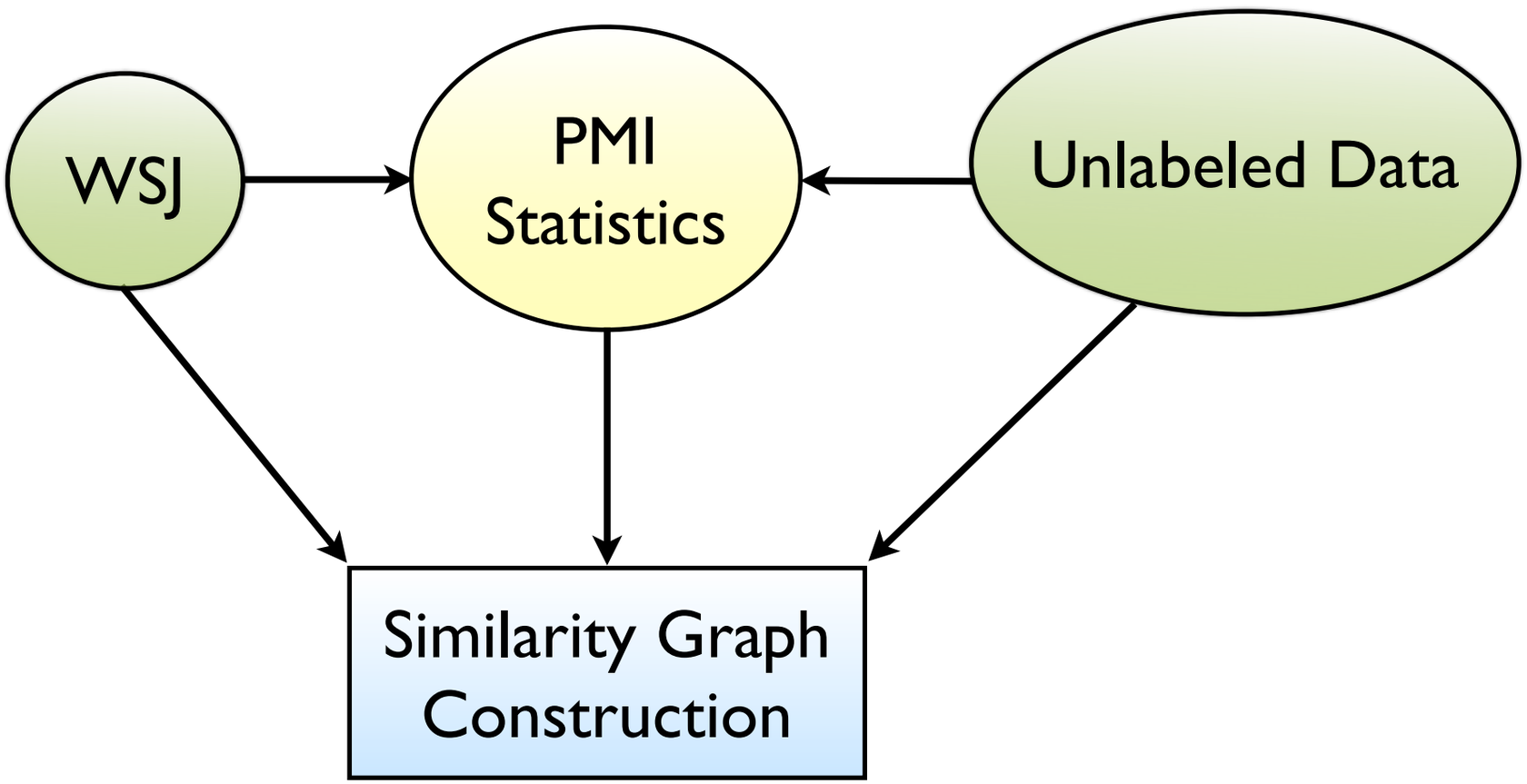
- Source Domain (labeled): Wall Street Journal (WSJ) section of the Penn Treebank.
- Target Domain:
  - QuestionBank: 4000 labeled sentences
  - Penn BioTreebank: 1061 labeled sentences

Graph Construction: Question Bank

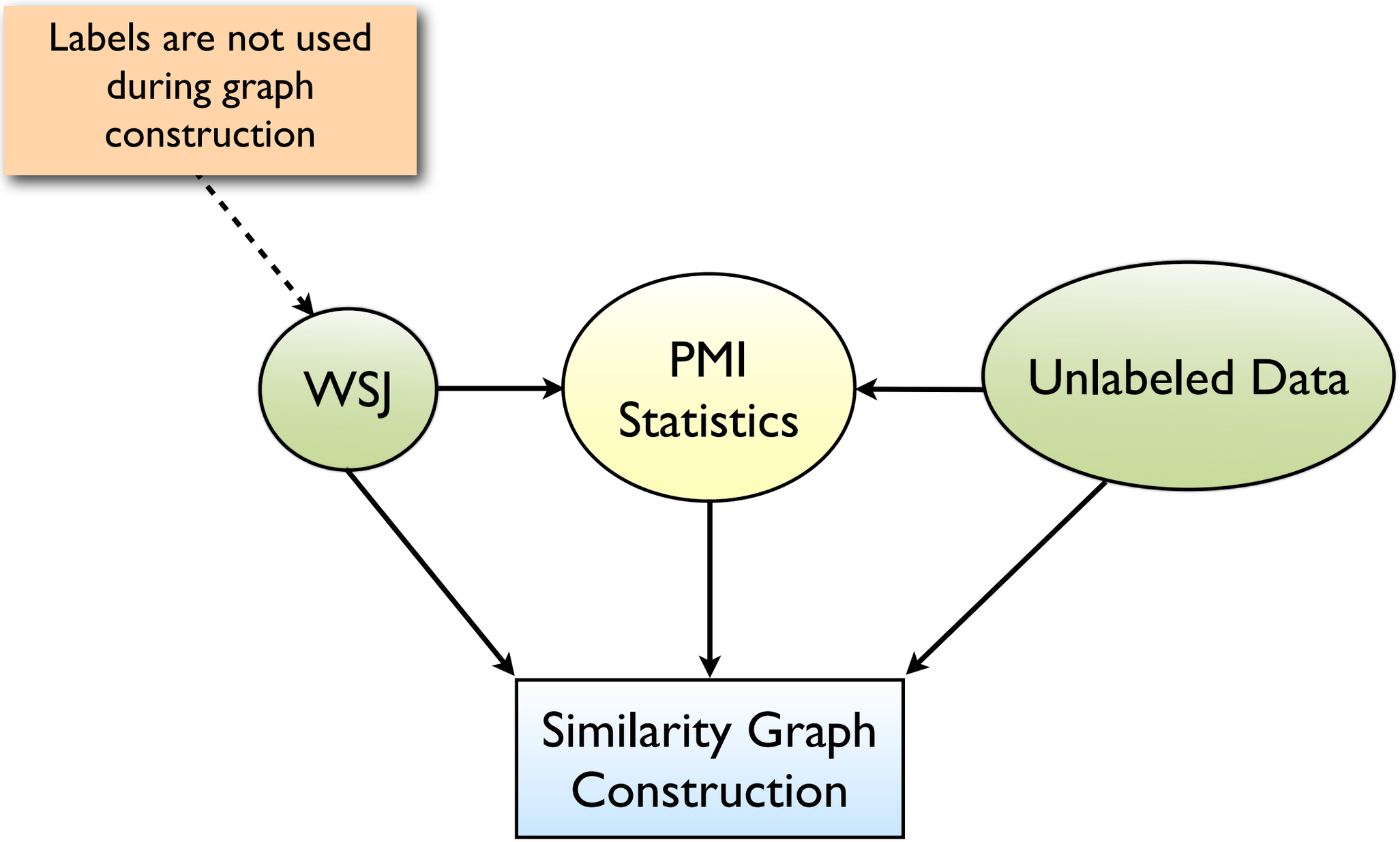
Graph Construction: Question Bank



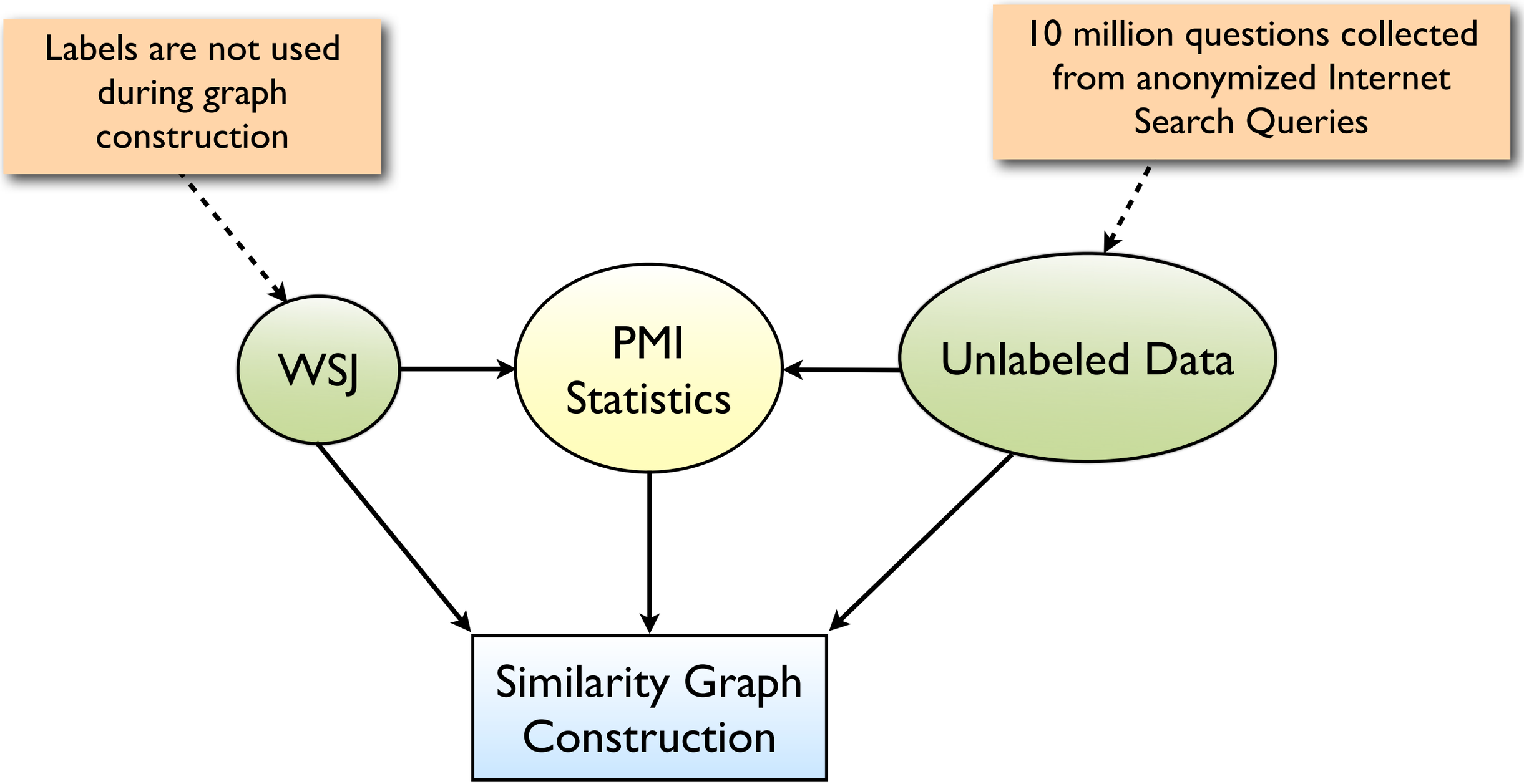
Graph Construction: Question Bank



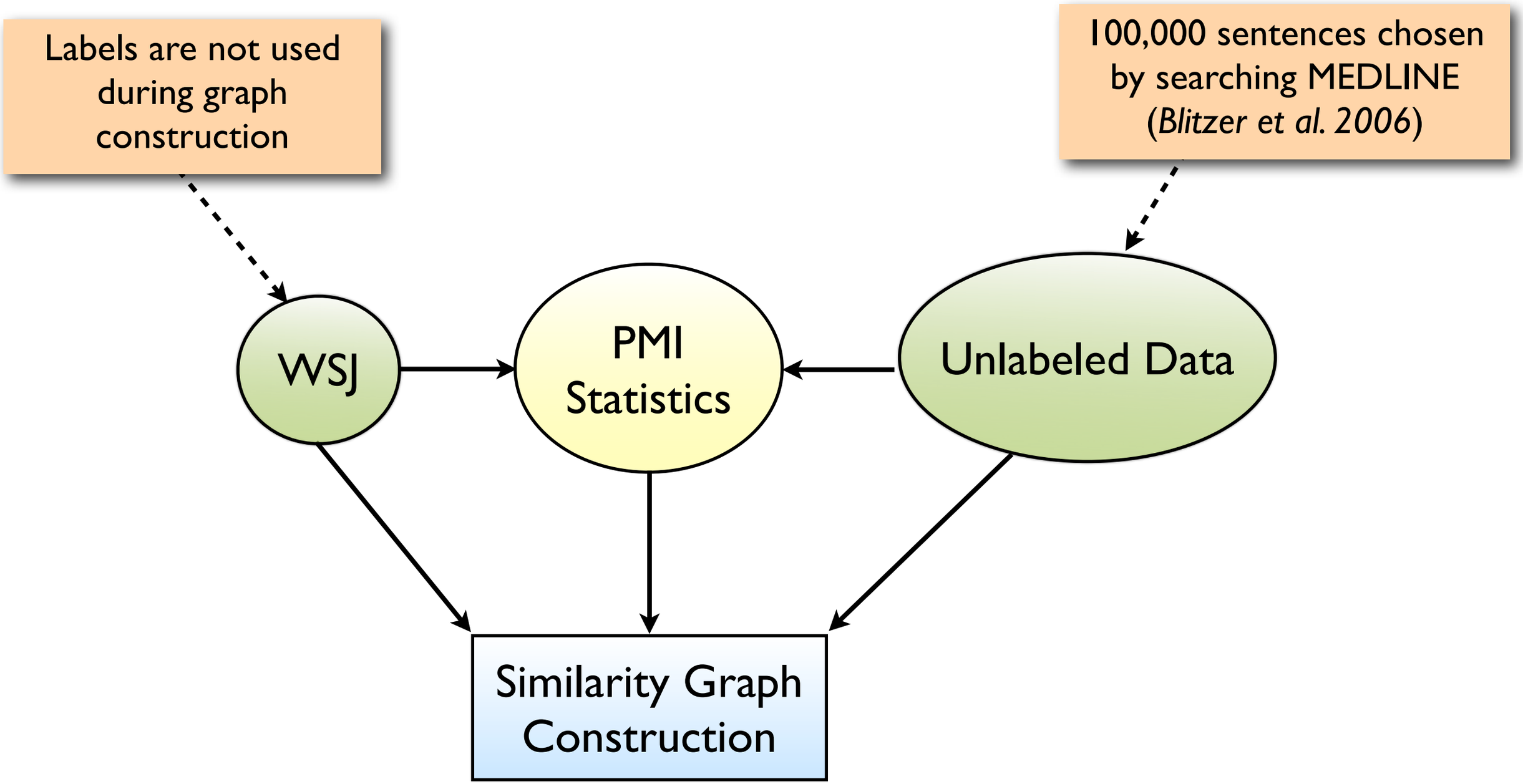
Graph Construction: Question Bank



Graph Construction: Question Bank



Graph Construction: Bio



Baseline (Supervised)

Not the same as features used using graph construction

- Features: word identity, suffixes, prefixes & special character detectors (dashes, digits, etc.).
- Achieves 97.17% accuracy on WSJ development set.

Results

	Questions	Bio
Baseline	83.8	86.2
Self-training	84.0	87.1
Semi-supervised CRF	**86.8**	**87.6**

Analysis

	Questions	Bio
percentage of unlabeled trigrams not connected to and any labeled trigram	12.4	46.8
average path length between an unlabeled trigram and its nearest labeled trigram	9.4	22.4

Analysis

Sparse Graph

	Questions	Bio
percentage of unlabeled trigrams not connected to and any labeled trigram	12.4	46.8
average path length between an unlabeled trigram and its nearest labeled trigram	9.4	22.4

Analysis

- Pros
  - Inductive
  - Produces a CRF (standard CRF inference infrastructure may be used)
- Issues
  - Graph construction
  - Graph is not integrated with CRF training

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
  - Phone Classification
  - Text Categorization
  - Dialog Act Tagging
  - Statistical Machine Translation
  - POS Tagging
  - MultiLingual POS Tagging [Das & Petrov, ACL 2011]
- Conclusion & Future Work

Motivation

Motivation

- Supervised POS taggers for English have accuracies in the high 90's for most domains

Motivation

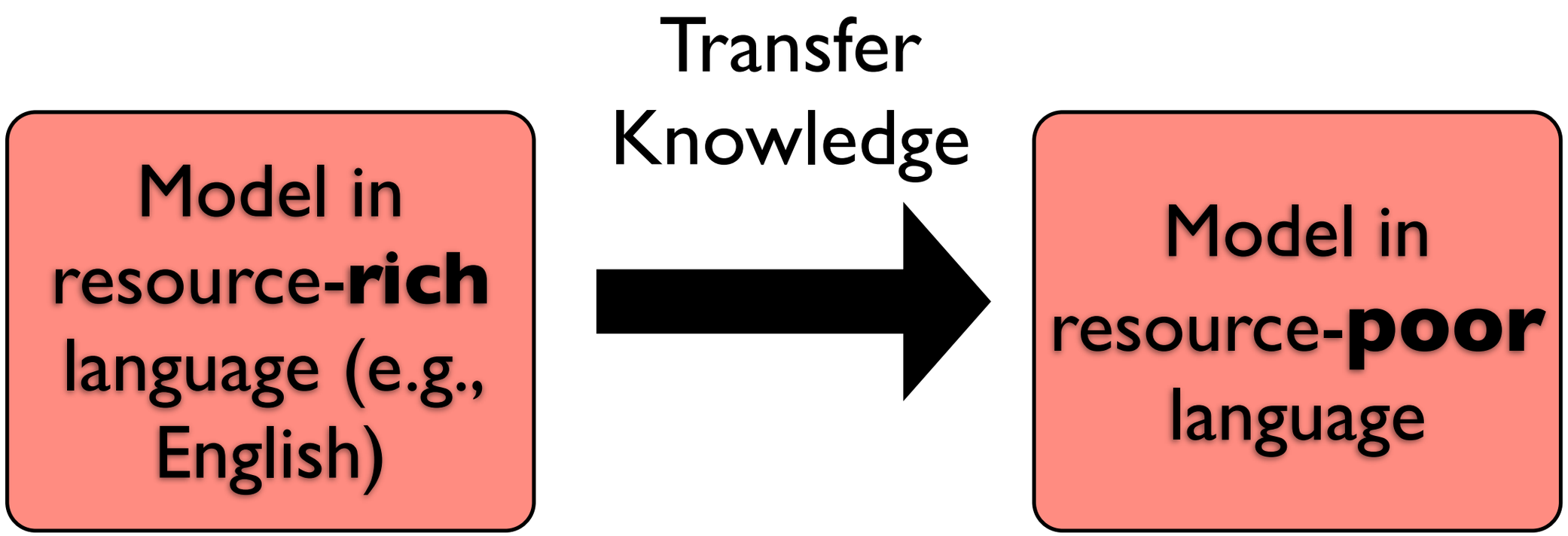
- Supervised POS taggers for English have accuracies in the high 90's for most domains
- By comparison taggers in other languages are not as accurate

Motivation

- Supervised POS taggers for English have accuracies in the high 90's for most domains
- By comparison taggers in other languages are not as accurate
  - Performance ranges from between 60 - 80%

Motivation

- Supervised POS taggers for English have accuracies in the high 90's for most domains
- By comparison taggers in other languages are not as accurate
  - Performance ranges from between 60 - 80%



Cross-Lingual Projection

The food at Google is good .

Cross-Lingual Projection

96% Accuracy

DET	**NOUN**	**ADP**	**NOUN**	**VERB**	**ADJ**	**.**
The	food	at	Google	is	good	.

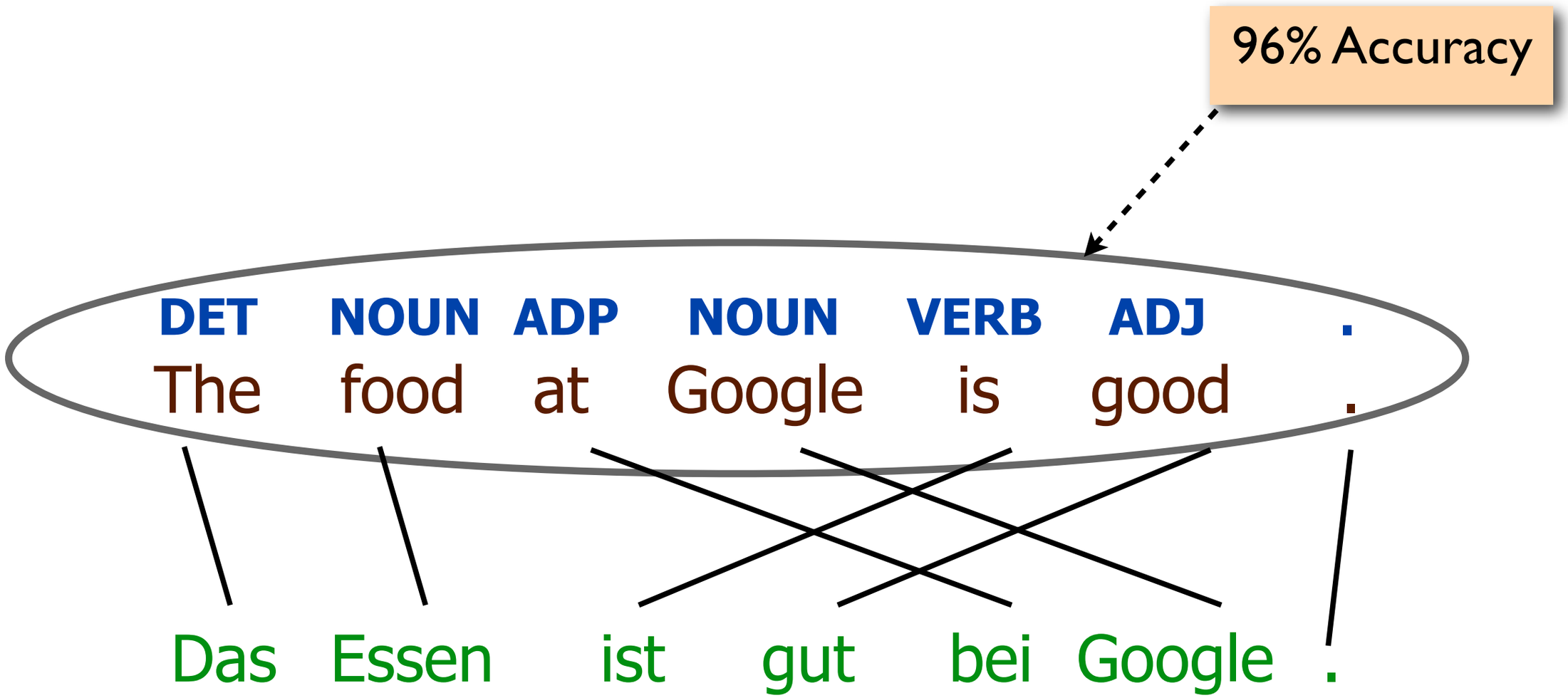
Cross-Lingual Projection

96% Accuracy

DET	**NOUN**	**ADP**	**NOUN**	**VERB**	**ADJ**	**.**
The	food	at	Google	is	good	.

Das Essen ist gut bei Google .

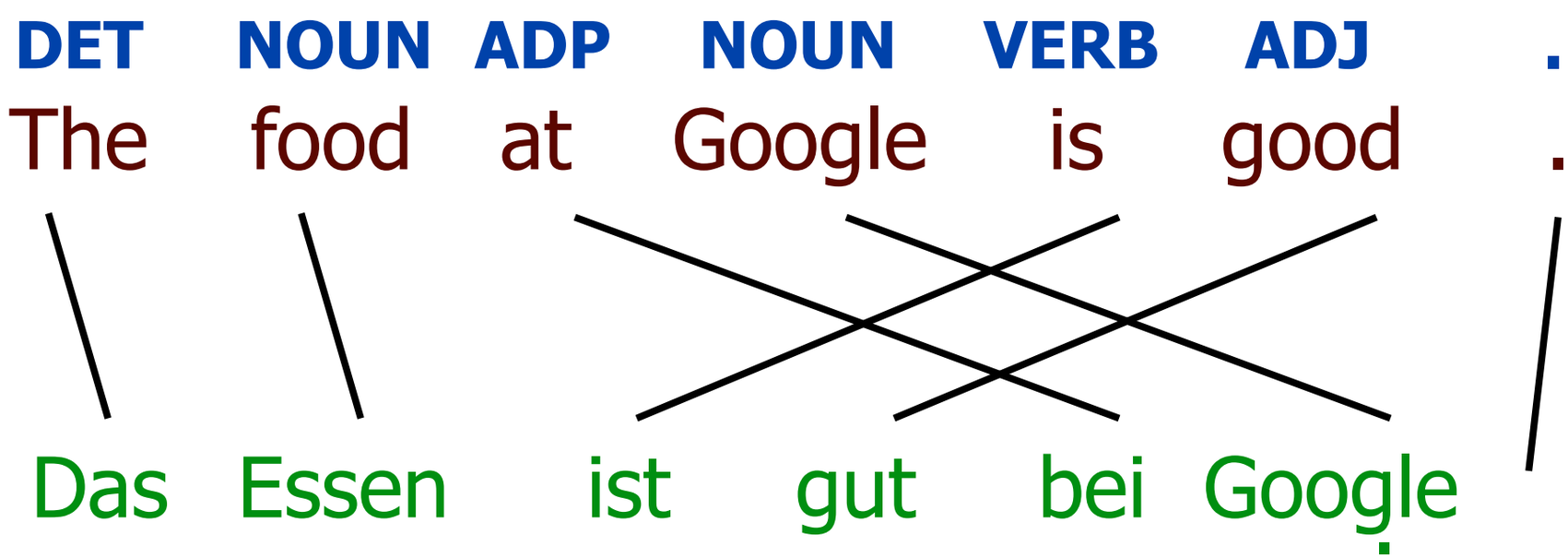
Cross-Lingual Projection



Automatic alignments from translation data

(available for more than 50 languages)

Cross-Lingual Projection



Cross-Lingual Projection

NOUN
food

DET
The

Essen

Das

VERB
is

ist

ADJ
good

gut

.

bei

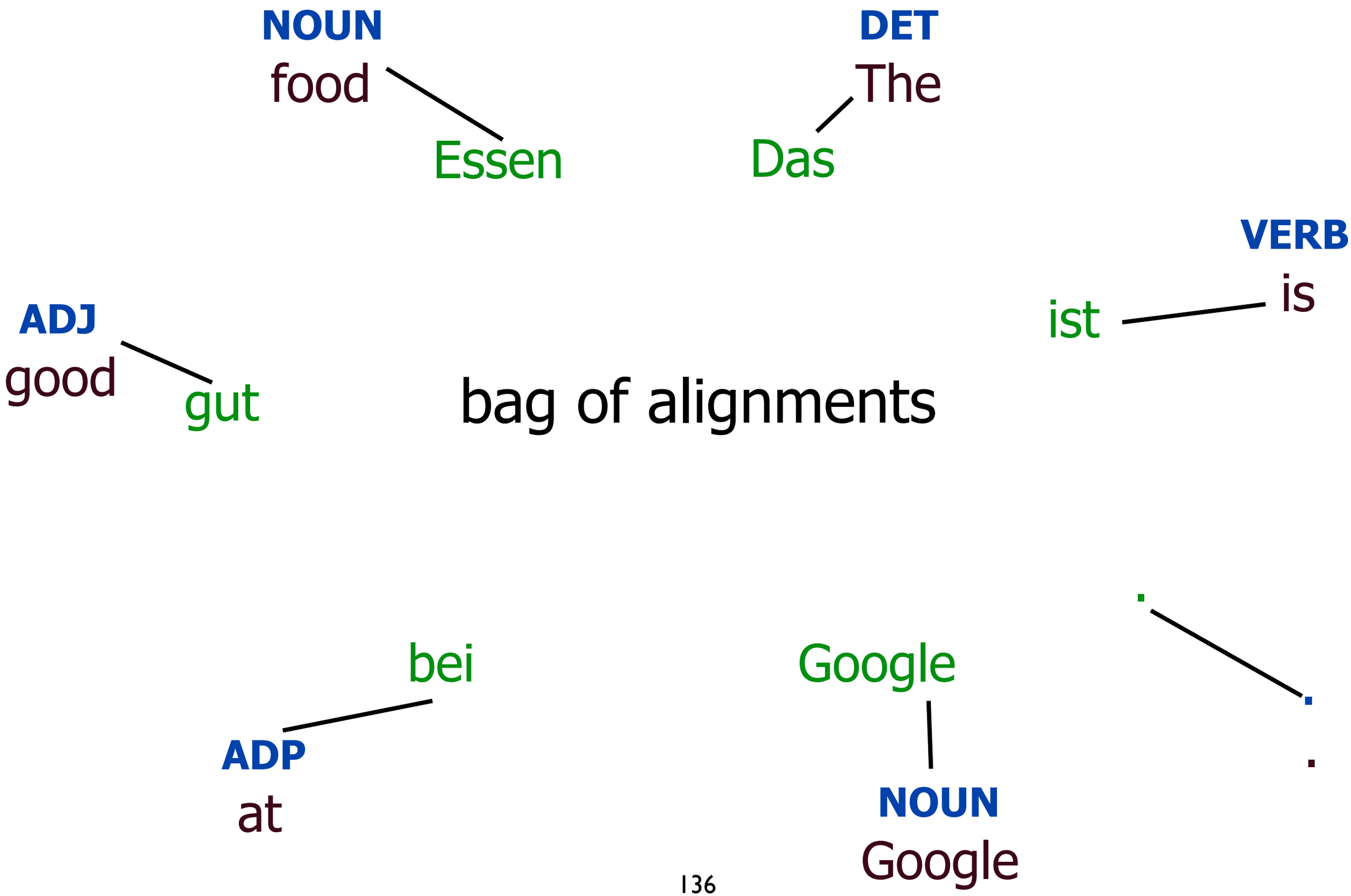
Google

.

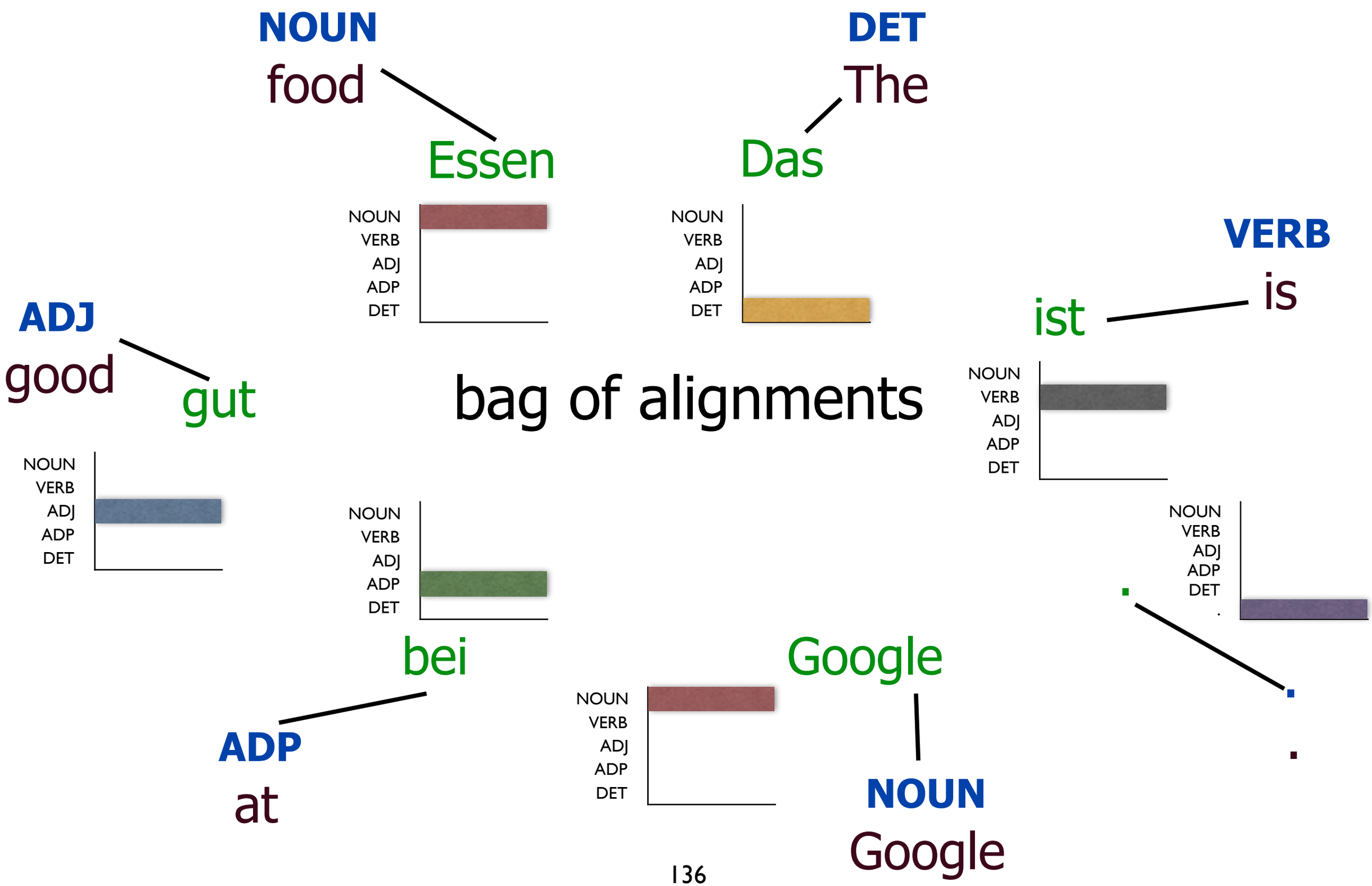
ADP
at

NOUN
Google

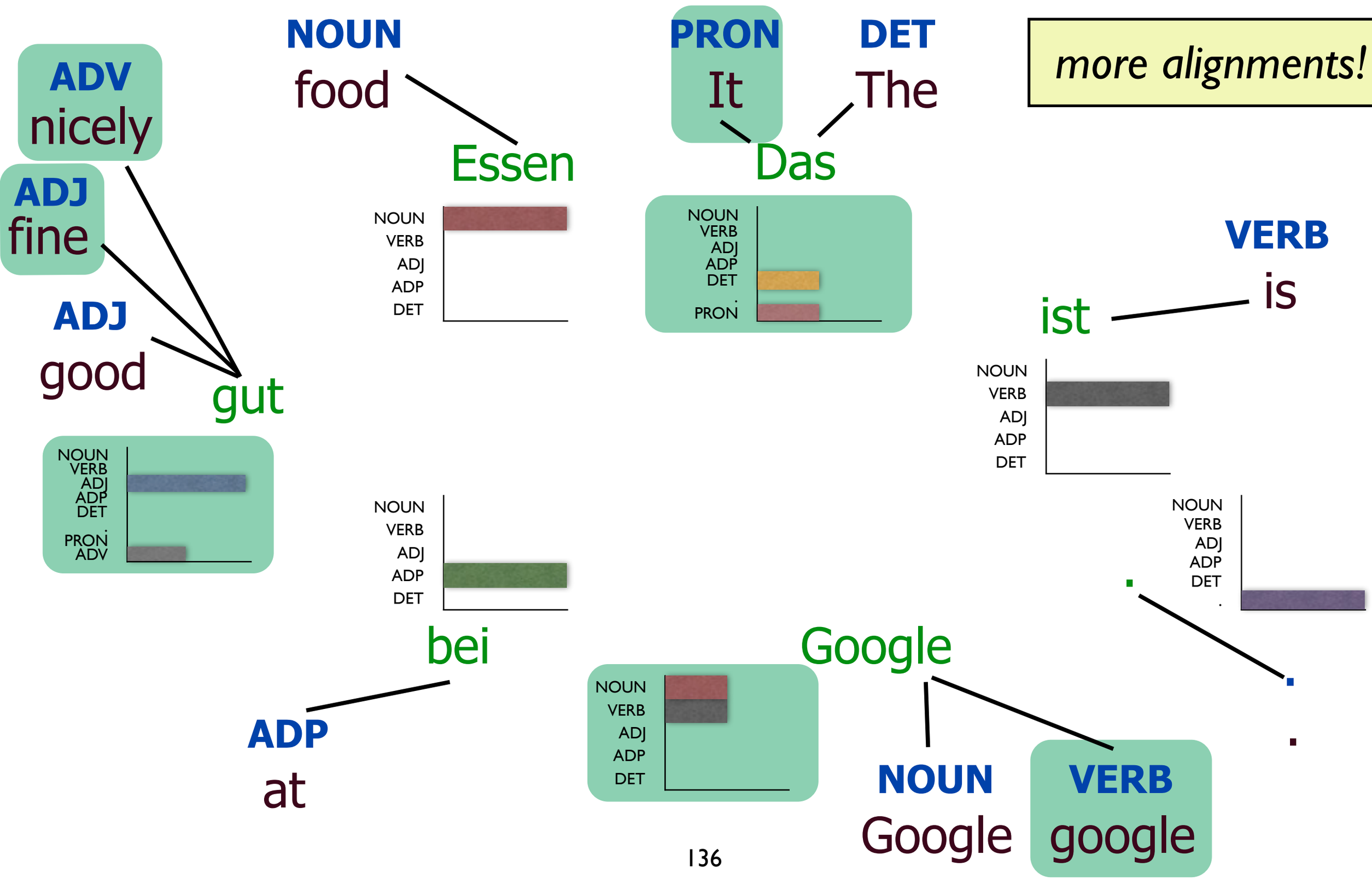
Cross-Lingual Projection



Cross-Lingual Projection



Cross-Lingual Projection



Cross-Lingual Projection Results

	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0

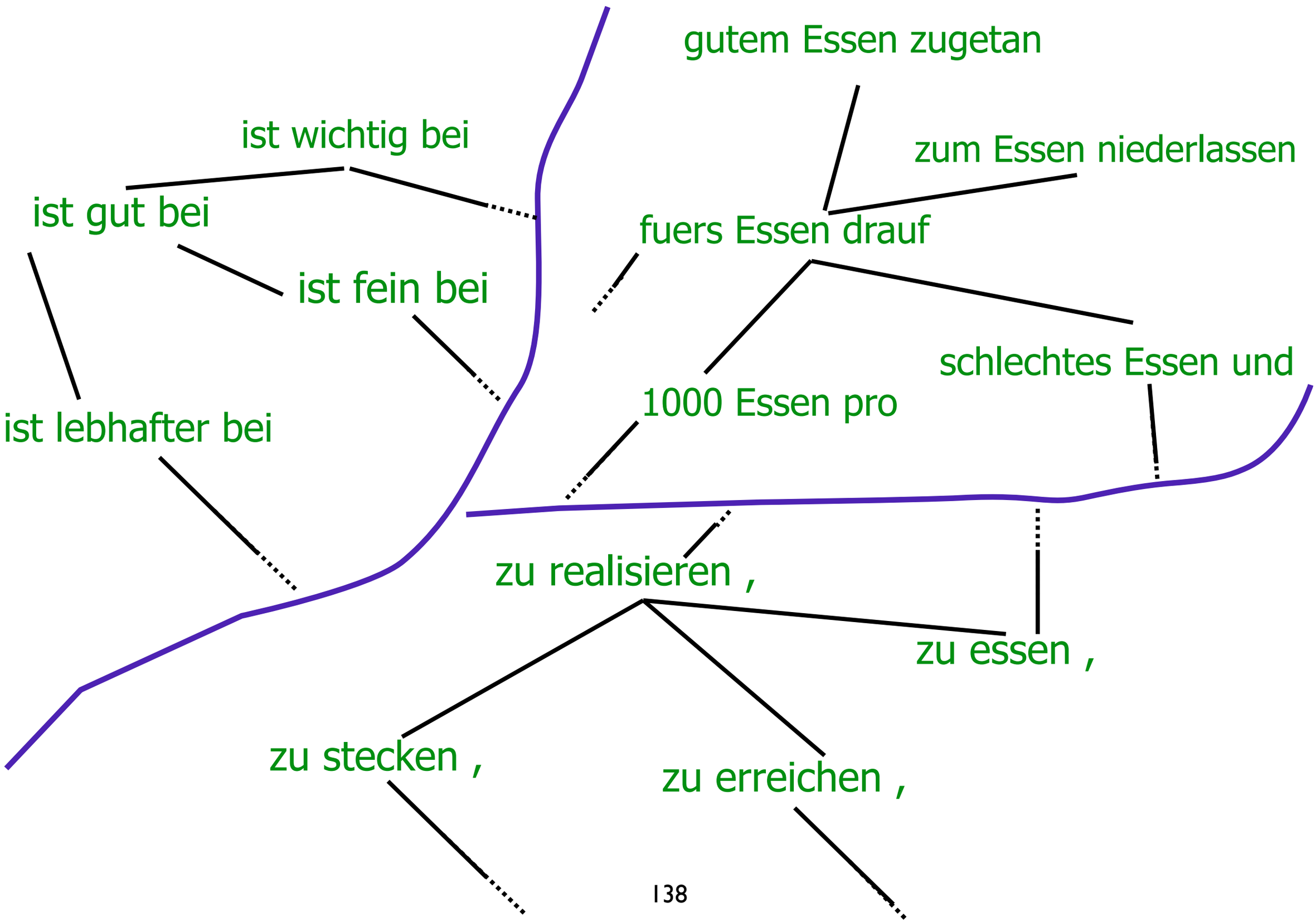
Feature-HMM [Berg-Kirkpatrick, NAACL 2010]

Cross-Lingual Projection Results

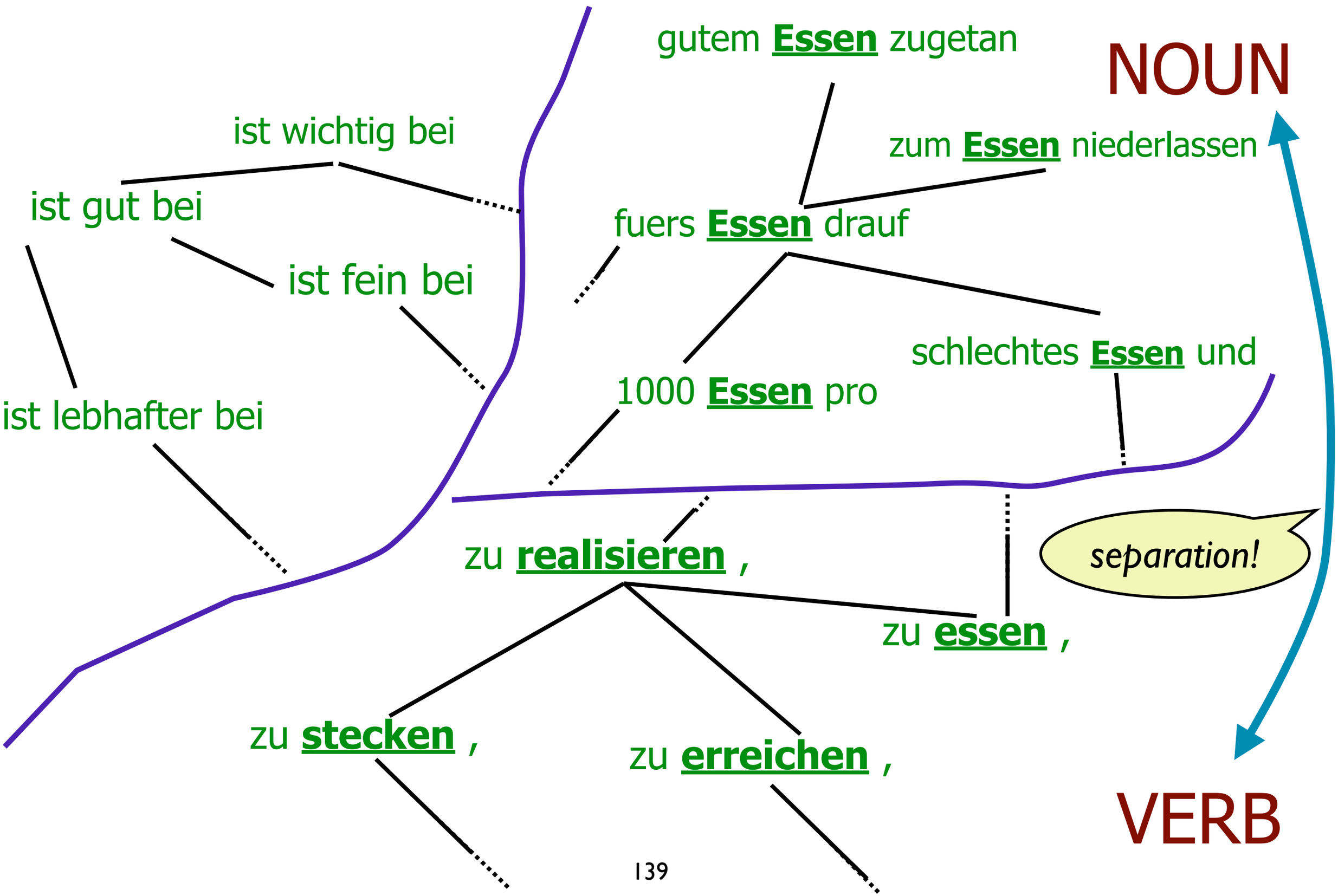
	Danish	Dutch	German	Greek	Italian	Portuguese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	**73.6**	**77.0**	**83.2**	**79.3**	**79.7**	**82.6**	**80.1**	**74.7**	**78.8**

Feature-HMM [Berg-Kirkpatrick, NAACL 2010]

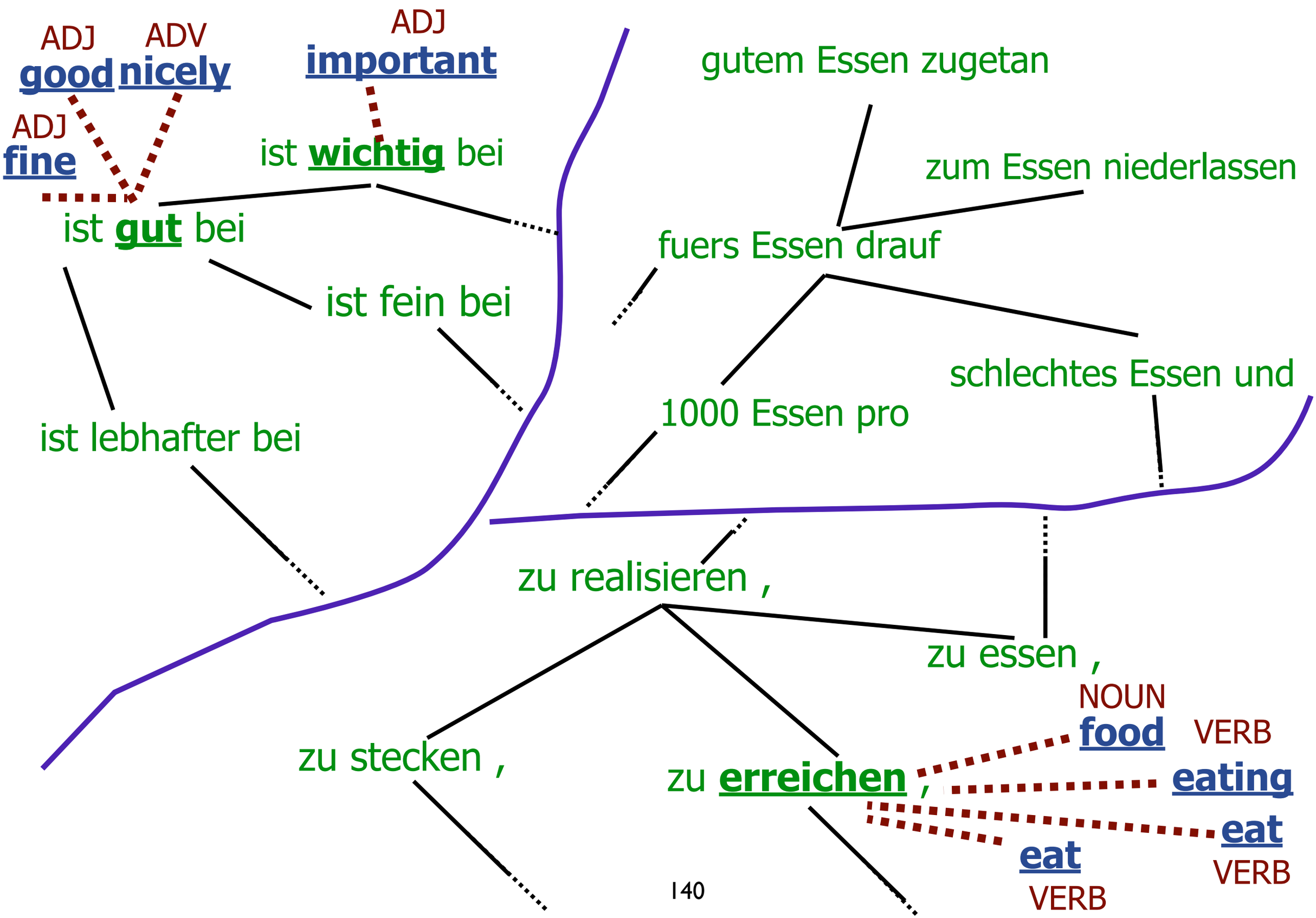
Graph Regularization



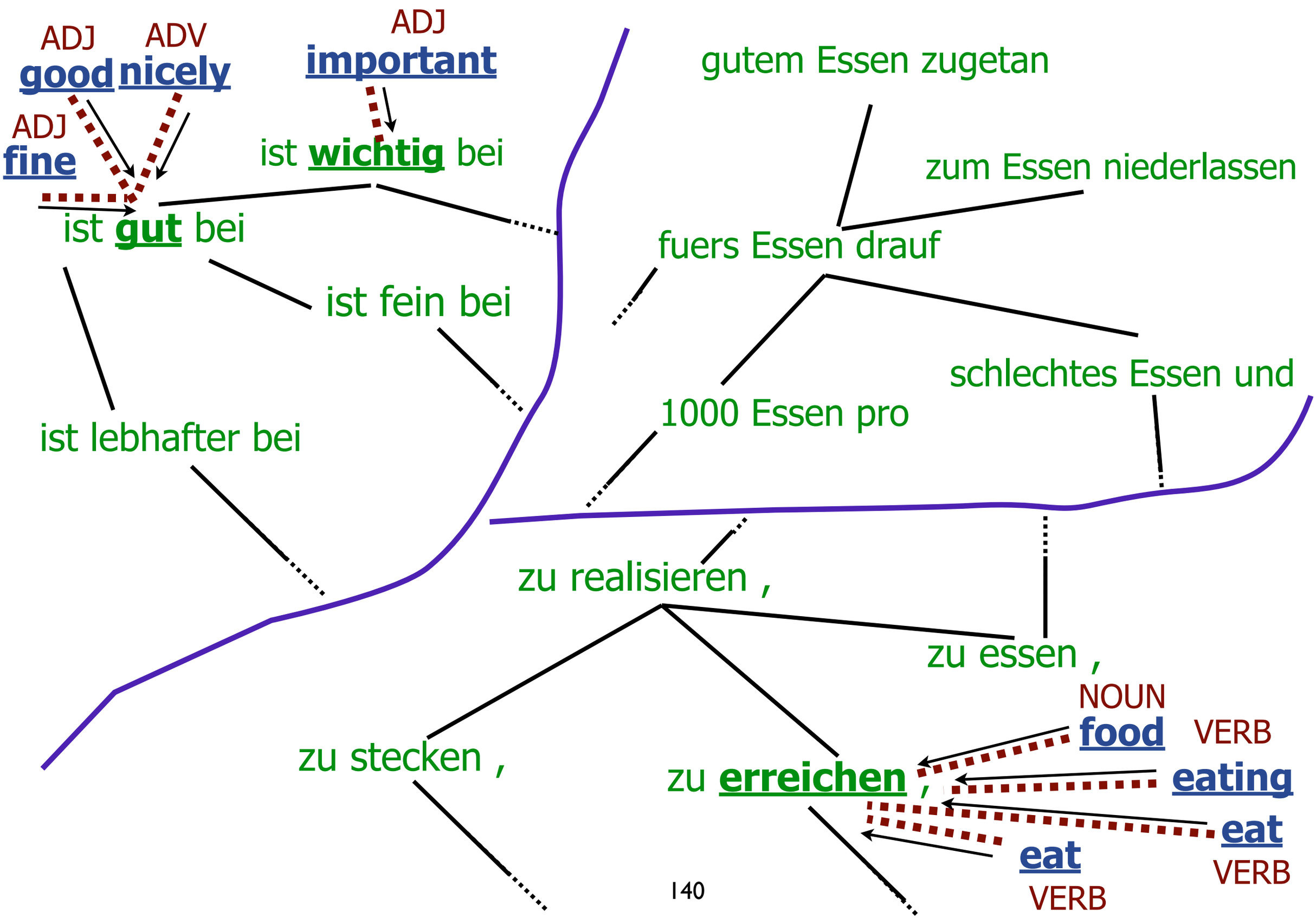
Graph Regularization



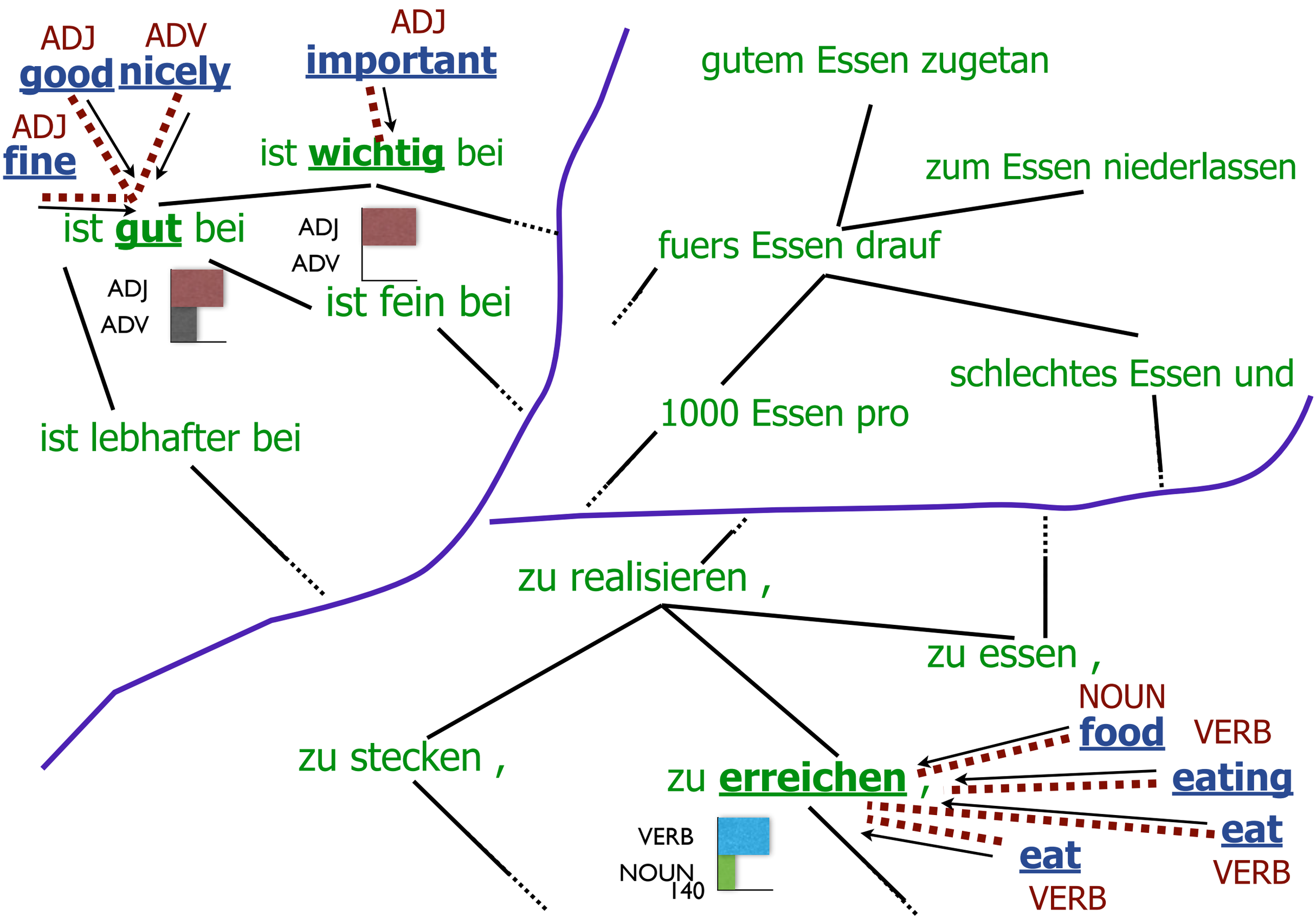
Graph Regularization



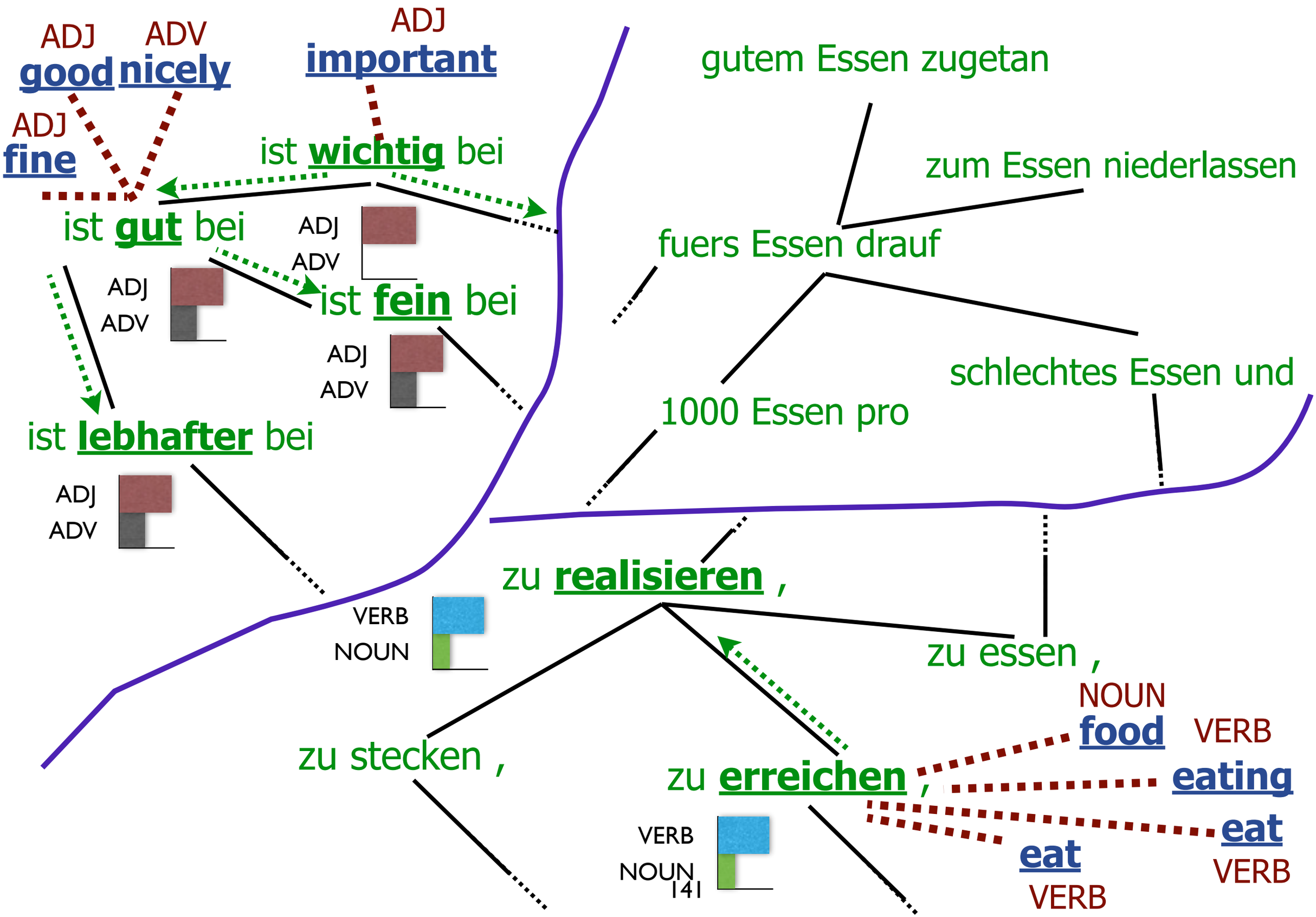
Graph Regularization



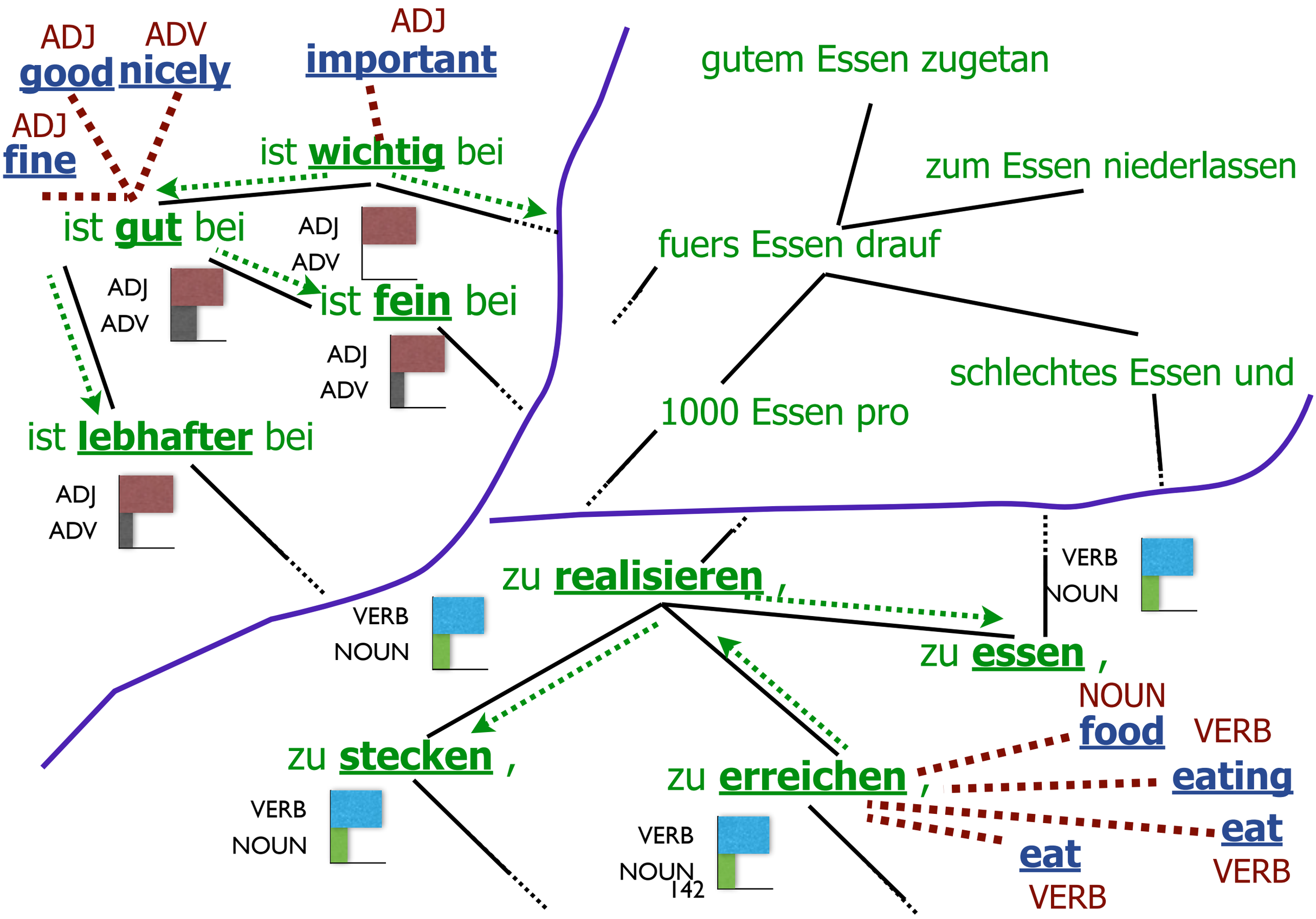
Graph Regularization



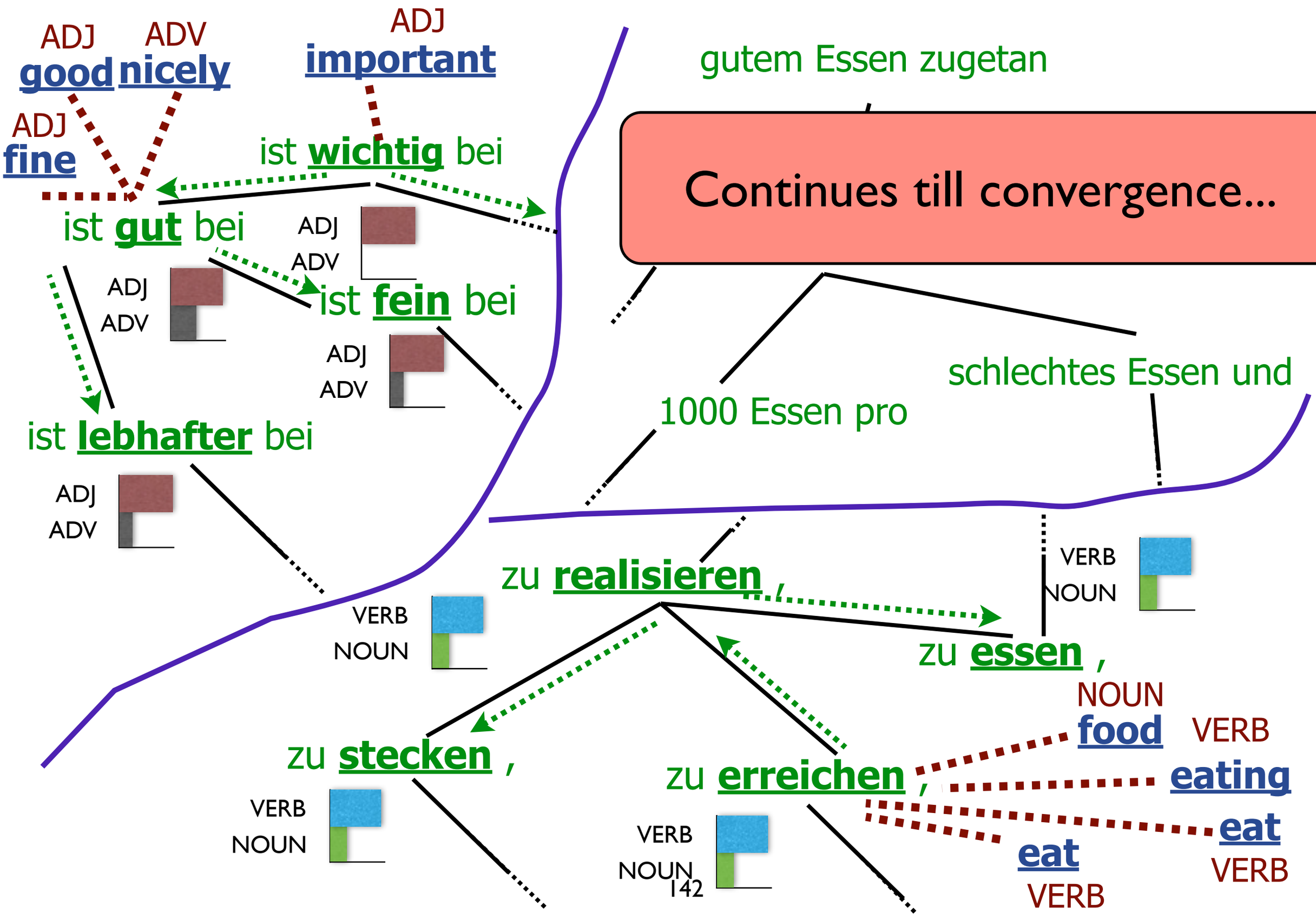
Graph Regularization



Graph Regularization



Graph Regularization



Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8

Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Graph-based Projection	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**

Results

	Danish	Dutch	German	Greek	Italian	Portugese	Spanish	Swedish	Average
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	73.0
Direct Projection	73.6	77.0	**83.2**	79.3	79.7	82.6	80.1	74.7	78.8
Graph-based Projection	**83.2**	**79.5**	82.8	**82.5**	**86.8**	**87.9**	**84.2**	**80.5**	**83.4**
Oracle (Supervised)	96.9	94.9	98.2	97.8	95.8	97.2	96.8	94.8	96.6

Outline

- Motivation
- Graph Construction
- Inference Methods
- Scalability
- Applications
- Conclusion & Future Work

When to use Graph-based SSL and which method?

When to use Graph-based SSL and which method?

- When input data itself is a graph (relational data)
  - or, when the data is expected to lie on a manifold

When to use Graph-based SSL and which method?

- When input data itself is a graph (relational data)
  - or, when the data is expected to lie on a manifold
- MAD, Quadratic Criteria (QC)
  - when labels are not mutually exclusive
  - MADDL: when label similarities are known

When to use Graph-based SSL and which method?

- When input data itself is a graph (relational data)
  - or, when the data is expected to lie on a manifold
- MAD, Quadratic Criteria (QC)
  - when labels are not mutually exclusive
  - MADDL: when label similarities are known
- Measure Propagation (MP)
  - for probabilistic interpretation

When to use Graph-based SSL and which method?

- When input data itself is a graph (relational data)
  - or, when the data is expected to lie on a manifold
- MAD, Quadratic Criteria (QC)
  - when labels are not mutually exclusive
  - MADDL: when label similarities are known
- Measure Propagation (MP)
  - for probabilistic interpretation
- Manifold Regularization
  - for generalization to unseen data (induction)

Graph-based SSL: Summary

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key
- Scalable: Node Reordering and MapReduce

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key
- Scalable: Node Reordering and MapReduce
- Can handle labeled as well as unlabeled data

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key
- Scalable: Node Reordering and MapReduce
- Can handle labeled as well as unlabeled data
- Can handle multi class, multi label settings

Graph-based SSL: Summary

- Provide flexible representation
  - for both IID and relational data
- Graph construction can be key
- Scalable: Node Reordering and MapReduce
- Can handle labeled as well as unlabeled data
- Can handle multi class, multi label settings
- Effective in practice

Open Challenges

Open Challenges

- Graph-based SSL for Structured Prediction
  - Algorithms: Combining Inductive and graph-based methods
  - Applications: Constituency and dependency parsing, Coreference

Open Challenges

- Graph-based SSL for Structured Prediction
  - Algorithms: Combining Inductive and graph-based methods
  - Applications: Constituency and dependency parsing, Coreference
- Scalable graph construction, especially with multi-modal data

Open Challenges

- Graph-based SSL for Structured Prediction
  - Algorithms: Combining Inductive and graph-based methods
  - Applications: Constituency and dependency parsing, Coreference
- Scalable graph construction, especially with multi-modal data
- Extensions with other loss functions, sparsity, etc.

Open Challenges

- Graph-based SSL for Structured Prediction
  - Algorithms: Combining Inductive and graph-based methods
  - Applications: Constituency and dependency parsing, Coreference
- Scalable graph construction, especially with multi-modal data
- Extensions with other loss functions, sparsity, etc.
- Using side information

Acknowledgments

- National Science Foundation (NSF) IIS-0447972
- DARPA HRO1107-1-0029, FA8750-09-C-0179
- Google Research Award
- Dipanjan Das (Google), Fernando Pereira (Google), Matan Orbach (Technion), Noah Smith (CMU)

References

1. A. Alexandrescu and K. Kirchhoff. Data-driven graph construction for semi-supervised graph-based learning in nlp. In NAACL HLT, 2007.
2. Y. Liu and K. Kirchhoff. A comparison of graph construction and learning algorithms for graph-based phonetic classification. In UWEETR-2012-0005.
3. S. Greenberg. The Switchboard Transcription Project. The Johns Hopkins University (CLSP) Summer Research Workshop 1995.
4. N. Deshmukh and A. Ganapathiraju and A. Gleeson and J. Hamaker and J. Picone. Resegmentation of Switchboard. ICSLP 1998.
5. D. Jurafsky and C.V. Ess-Dykema. Switchboard Discourse Language Modeling Project. Johns Hopkins Summer Workshop. 1997.
6. G. Ji and J. Bilmes. Dialog Act Tagging using Graphical Models. ICASSP 2005.
7. A. Alexandrescu and K. Kirchhoff. Graph-based Learning for Statistical Machine Translation. NAACL-HLT 2009.
8. A. Alexandrescu and K. Kirchhoff. Graph-Based Learning for Phonetic Classification. ASRU 2007.
9. K. Papineni and S. Roukos and T. Ward and W. Zhu. BLEU: a method for automatic evaluation of machine translation. ACL 2002.
10. Y. Altun, D. McAllester, and M. Belkin. Maximum margin semi-supervised learn- ing for structured variables. NIPS, 2006.
11. B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, M. A. T. Figueiredo. On Semi-Supervised ClassiÔ¨Åcation. In NIPS 2004.
12. S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In WWW, 2008.
13. R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. J. Mach. Learn. Res., 3:1183,Äì1208, 2003.
14. M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 7:2399,Äì2434, 2006.
15. Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. Semi-supervised learning, 2006.
16. T. Berg-Kirkpatrick, A. Bouchard-CoÀÜt ÃÅe, J. DeNero, and D. Klein. Painless unsupervised learning with features. In HLT-NAACL, 2010.
17. J. Bilmes and A. Subramanya. Scaling up Machine Learning: Parallel and Distributed Approaches, chapter Parallel Graph-Based Semi-Supervised Learning. 2011.
18. S. Blair-goldensohn, T. Neylon, K. Hannan, G. A. Reis, R. Mcdonald, and J. Reynar. Building a sentiment summarizer for local service reviews. In In NLP in the Information Explosion Era, 2008.
19. M. Cafarella, A. Halevy, D. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. VLDB, 2008.
20. O. Chapelle, B. Scho Ãàlkopf, A. Zien, et al. Semi-supervised learning. MIT press Cambridge, MA:, 2006.

References

21. Y. Choi and C. Cardie. Adapting a polarity lexicon using integer linear program- ming for domain specific sentiment classification. In EMNLP, 2009.
22. S. Daitch, J. Kelner, and D. Spielman. Fitting a graph to vector data. In ICML, 2009.
23. D. Das and S. Petrov. Unsupervised part-of-speech tagging with bilingual graph- based projections. In ACL, 2011.
24. D. Das, N. Schneider, D. Chen, and N. A. Smith. Probabilistic frame-semantic parsing. In NAACL-HLT, 2010.
25. D. Das and N. Smith. Graph-based lexicon expansion with sparsity-inducing penalties. NAACL-HLT, 2012.
26. D. Das and N. A. Smith. Semi-supervised frame-semantic parsing for unknown predicates. In ACL, 2011.
27. J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In ICML, 2007.
28. O. Delalleau, Y. Bengio, and N. L. Roux. Efficient non-parametric function induction in semi-supervised learning. In AISTATS, 2005.
29. P. Dhillon, P. Talukdar, and K. Crammer. Inference-driven metric learning for graph construction. Technical report, MS-CIS-10-18, University of Pennsylvania, 2010.
30. S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In CIKM, 1998.
31. J. Friedman, J. Bentley, and R. Finkel. An algorithm for finding best matches in logarithmic expected time. ACM Transaction on Mathematical Software, 3, 1977.
32. J. Garcke and M. Griebel. Data mining with sparse grids using simplicial basis functions. In KDD, 2001.
33. A. Goldberg and X. Zhu. Seeing stars when there aren,Äôt many stars: graph-based semi-supervised learning for sentiment categorization. In Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, 2006.
34. A. Goldberg, X. Zhu, and S. Wright. Dissimilarity in graph-based semi-supervised classification. AISTATS, 2007.
35. M. Hu and B. Liu. Mining and summarizing customer reviews. In KDD, 2004.
36. T. Jebara, J. Wang, and S. Chang. Graph construction and b-matching for semi-supervised learning. In ICML, 2009.
37. T. Joachims. Transductive inference for text classification using support vector machines. In ICML, 1999.
38. T. Joachims. Transductive learning via spectral graph partitioning. In ICML, 2003.
39. M. Karlen, J. Weston, A. Erkan, and R. Collobert. Large scale manifold transduction. In ICML, 2008.
40. S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In Proceedings of the 20th International conference on Computational Linguistics, 2004.

References

41. F. Kschischang, B. Frey, and H. Loeliger. Factor graphs and the sum-product algorithm. Information Theory, IEEE Transactions on, 47 (2):498,Äì519, 2001.
42. K. Lerman, S. Blair-Goldensohn, and R. McDonald. Sentiment summarization: evaluating and learning user preferences. In EACL, 2009.
43. D.Lewisetal.Reuters-21578.http://www.daviddlewis.com/resources/testcollections/reuters21578, 1987.
44. J. Malkin, A. Subramanya, and J. Bilmes. On the semi-supervised learning of multi-layered perceptrons. In InterSpeech, 2009.
45. B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP, 2002.
45b. D. Rao and D. Ravichandran. Semi-supervised polarity lexicon induction. In EACL, 2009.
46. A. Subramanya and J. Bilmes. Soft-supervised learning for text classification. In EMNLP, 2008.
47. A. Subramanya and J. Bilmes. Entropic graph regularization in non-parametric semi-supervised classification. NIPS, 2009.
48. A. Subramanya and J. Bilmes. Semi-supervised learning with measure propagation. JMLR, 2011.
49. A. Subramanya, S. Petrov, and F. Pereira. Efficient graph-based semi-supervised learning of structured tagging models. In EMNLP, 2010.
50. P. Talukdar. Topics in graph construction for semi-supervised learning. Technical report, MS-CIS-09-13, University of Pennsylvania, 2009.
51. P. Talukdar and K. Crammer. New regularized algorithms for transductive learn- ing. ECML, 2009.
52. P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In ACL, 2010.
53. P. Talukdar, J. Reisinger, M. Pa Ãßsca, D. Ravichandran, R. Bhagat, and F. Pereira. Weakly-supervised acquisition of labeled class instances using graph random walks. In EMNLP, 2008.
53b. P. Talukdar, D. Wijaya, T, Mitchell. Acquiring Temporal Constraints between Relations. In CIKM 2012.
54. B. Van Durme and M. Pasca. Finding cars, goddesses and enzymes: Parametriz- able acquisition of labeled instances for open-domain information extraction. In AAAI, 2008.
55. L. Velikovich, S. Blair-Goldensohn, K. Hannan, and R. McDonald. The viability of web-derived polarity lexicons. In HLT-NAACL, 2010.
56. F. Wang and C. Zhang. Label propagation through linear neighborhoods. In ICML, 2006.
57. J. Wang, T. Jebara, and S. Chang. Graph transduction via alternating minimiza- tion. In ICML, 2008.
58. R. Wang and W. Cohen. Language-independent set expansion of named entities using the web. In ICDM, 2007.
59. K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. The Journal of Machine Learning Research, 10:207,Äì244, 2009.
60. T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase- level sentiment analysis. In HLT-EMNLP, 2005.

References

61. D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scho Ãàlkopf. Learning with local and global consistency. NIPS, 2004.
62. D. Zhou, J. Huang, and B. Scho Ãàlkopf. Learning from labeled and un- labeled data on a directed graph. In ICML, 2005.
63. D. Zhou, B. Scho Ãàlkopf, and T. Hofmann. Semi-supervised learning on directed graphs. In NIPS, 2005.
64. X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University, 2002.
65. X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, 2002.
66. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In ICML, 2003.
67. X. Zhu and J. Lafferty. Harmonic mixtures: combining mixture models and graph- based methods for inductive and scalable semi-supervised learning. In ICML, 2005.

Thank You!

Web: http://graph-ssl.wikidot.com/