Co-Training and **Estimating Accuracy from Unlabeled Data**

Anthony Platanios

Semi-Supervised Learning

Unlabeled Data

↓

Often easier to obtain

Labeled Data

Program that classifies a person's salary as high/low, given the city they live in and they job that they do

Labeled data are expensive

Demographics → **Unlabeled data are cheap**

Semi-Supervised Learning

Unlabeled Data

↓

Often easier to obtain

Labeled Data

Unlabeled data can give us information about the underlying distribution of the input data

Co-Training

Program that classifies a person’s salary as high/low, given the **city** they live in and they **job** that they do

Two **views** of the input data

Can **co-train** two classifiers

Main Assumption / Intuition

Agreement / consistency of the two classifiers

Co-Training

Program that classifies a person's salary as high/low, given the **city** they live in and they **job** that they do

Two **views** of the input data

Can **co-train** two classifiers

Algorithm

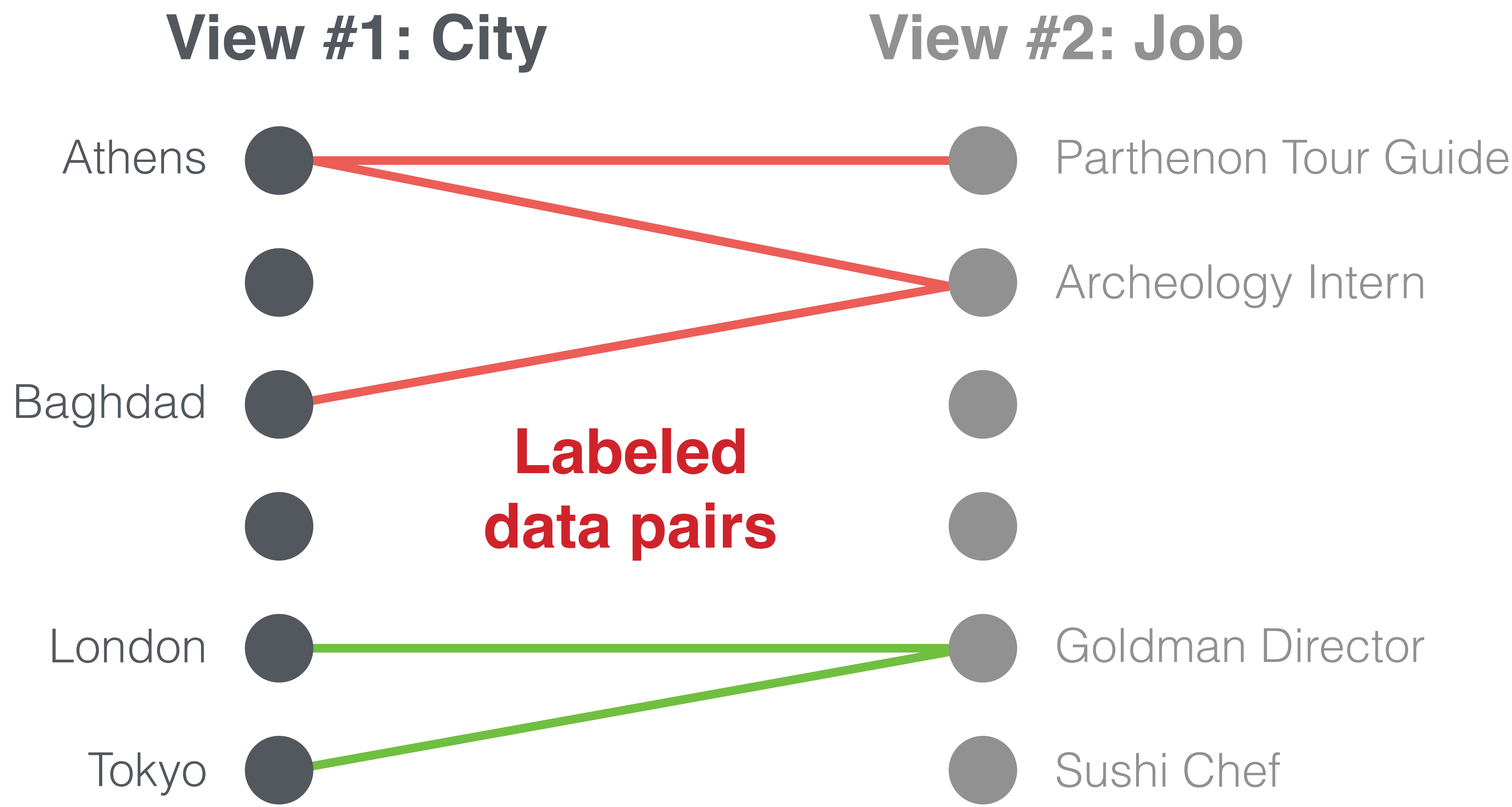
1. Train weak predictors using small set of labeled data
2. Make predictions with both over unlabeled data
3. Add most confident predictions to training data
4. Repeat

Co-Training **Intuition**

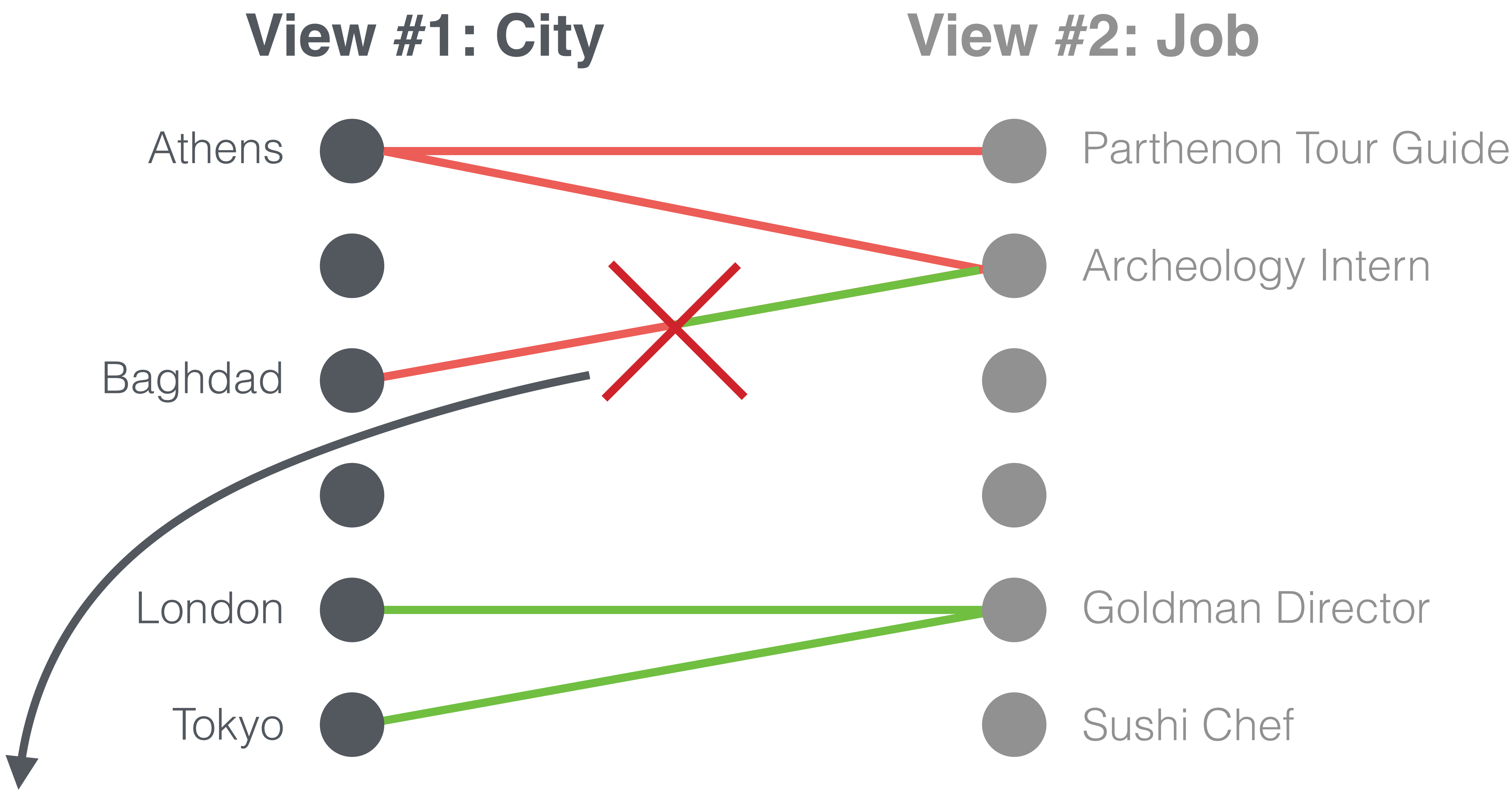
View #1: City	**View #2: Job**
Athens	Parthenon Tour Guide
	Archeology Intern
Baghdad	
London	Goldman Director
Tokyo	Sushi Chef

Each bullet is a feature under each view

Co-Training **Intuition**



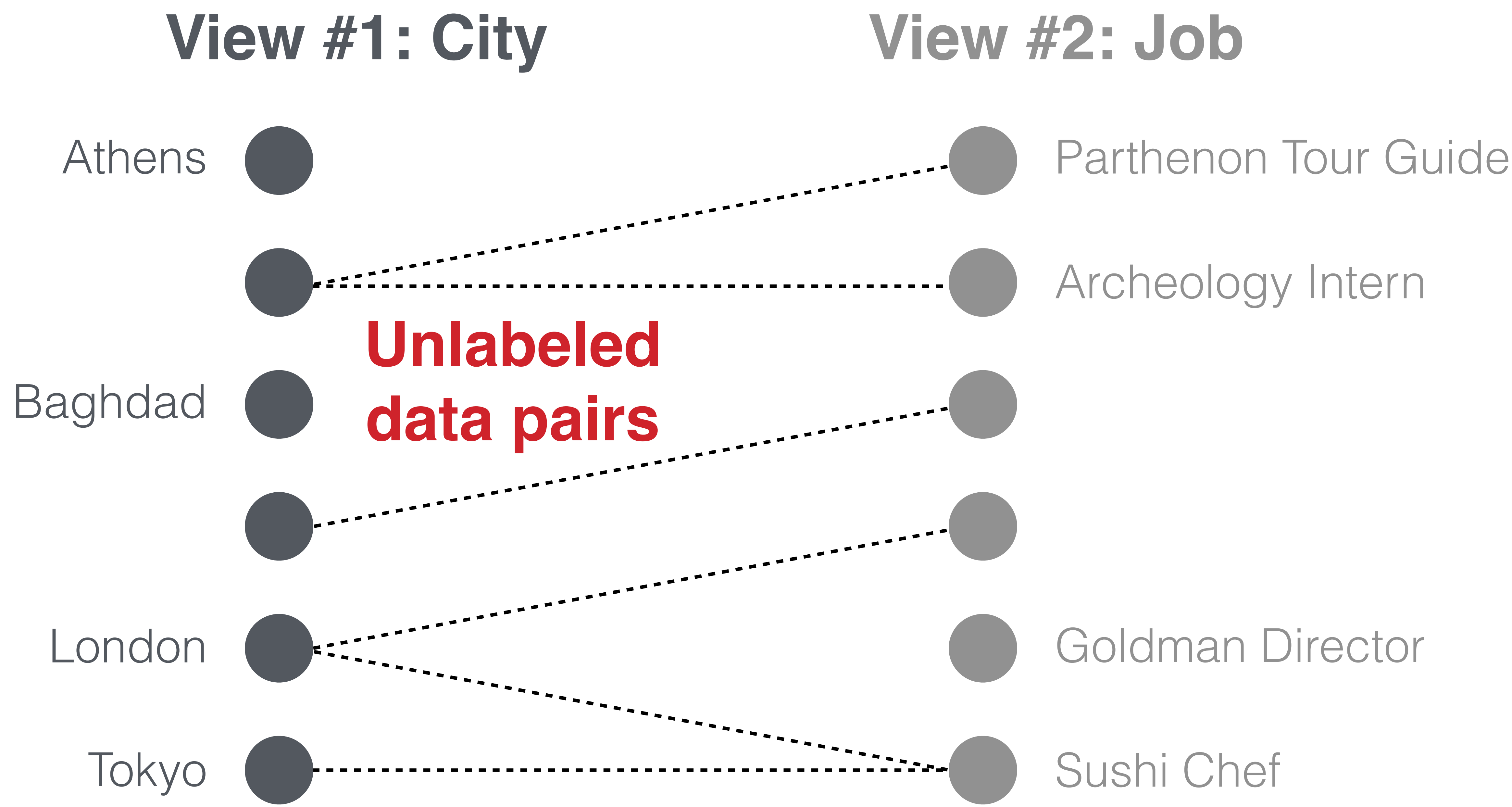
Co-Training **Intuition**



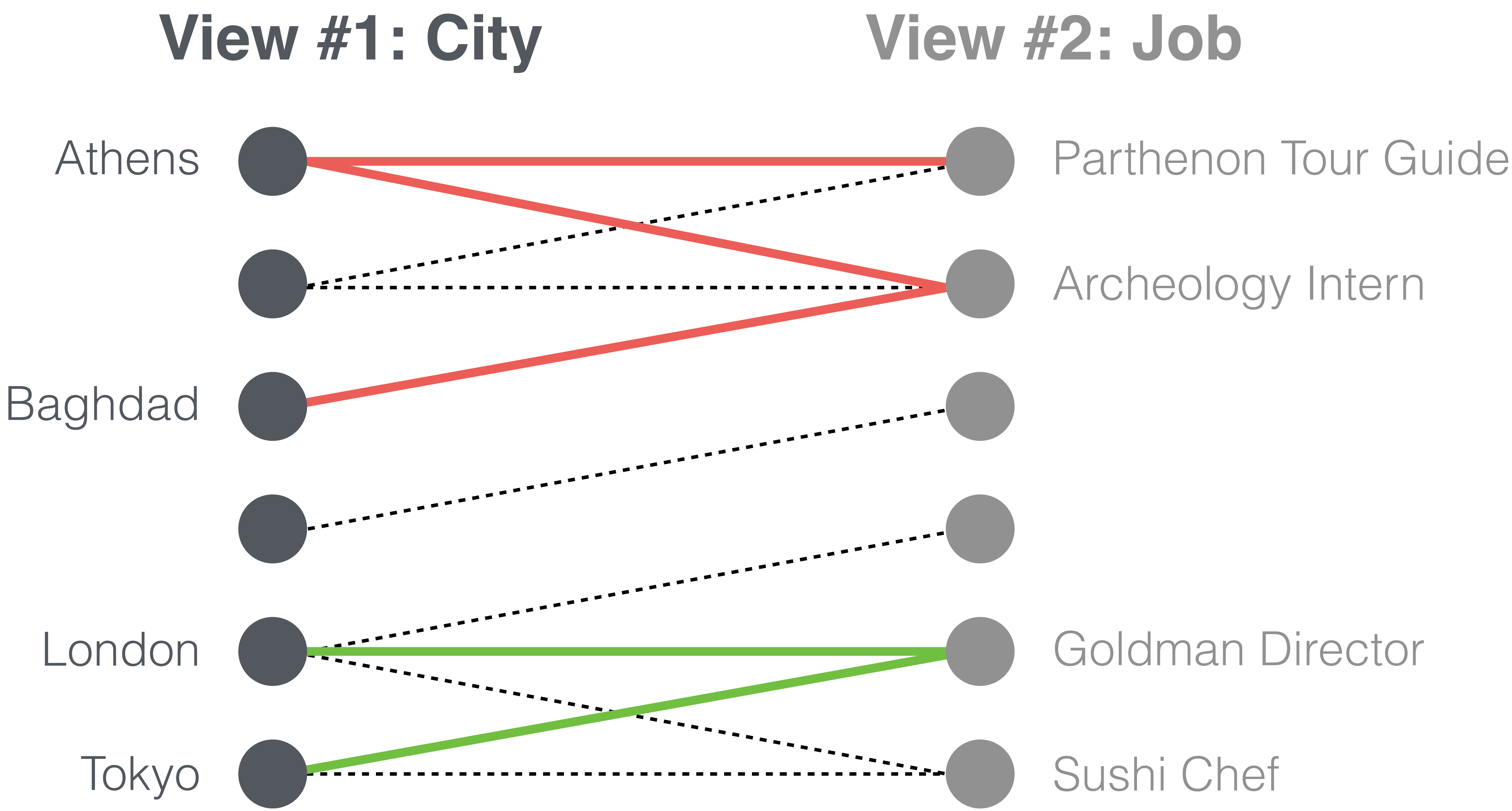
Main Assumption

Agreement / consistency of the two classifiers

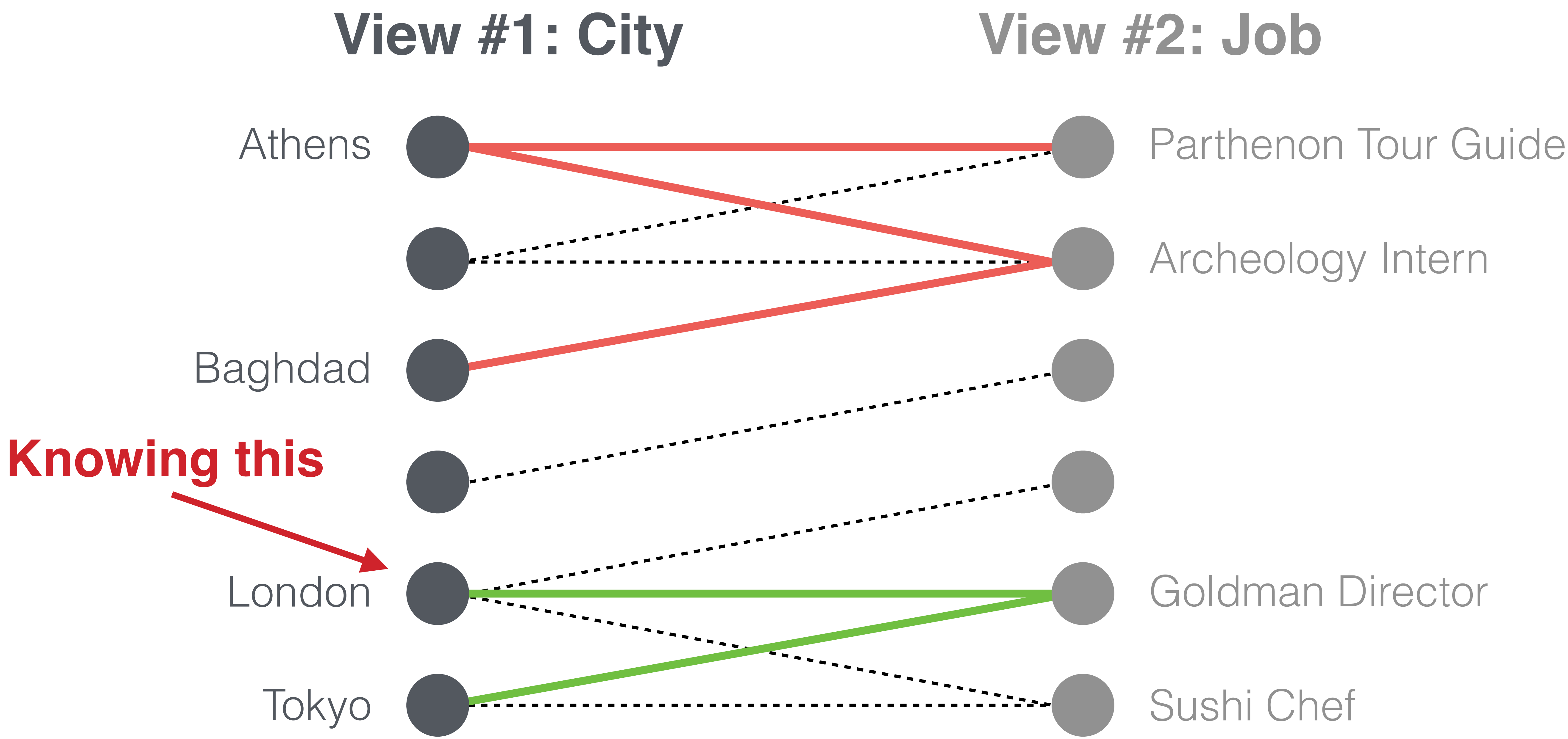
Co-Training **Intuition**



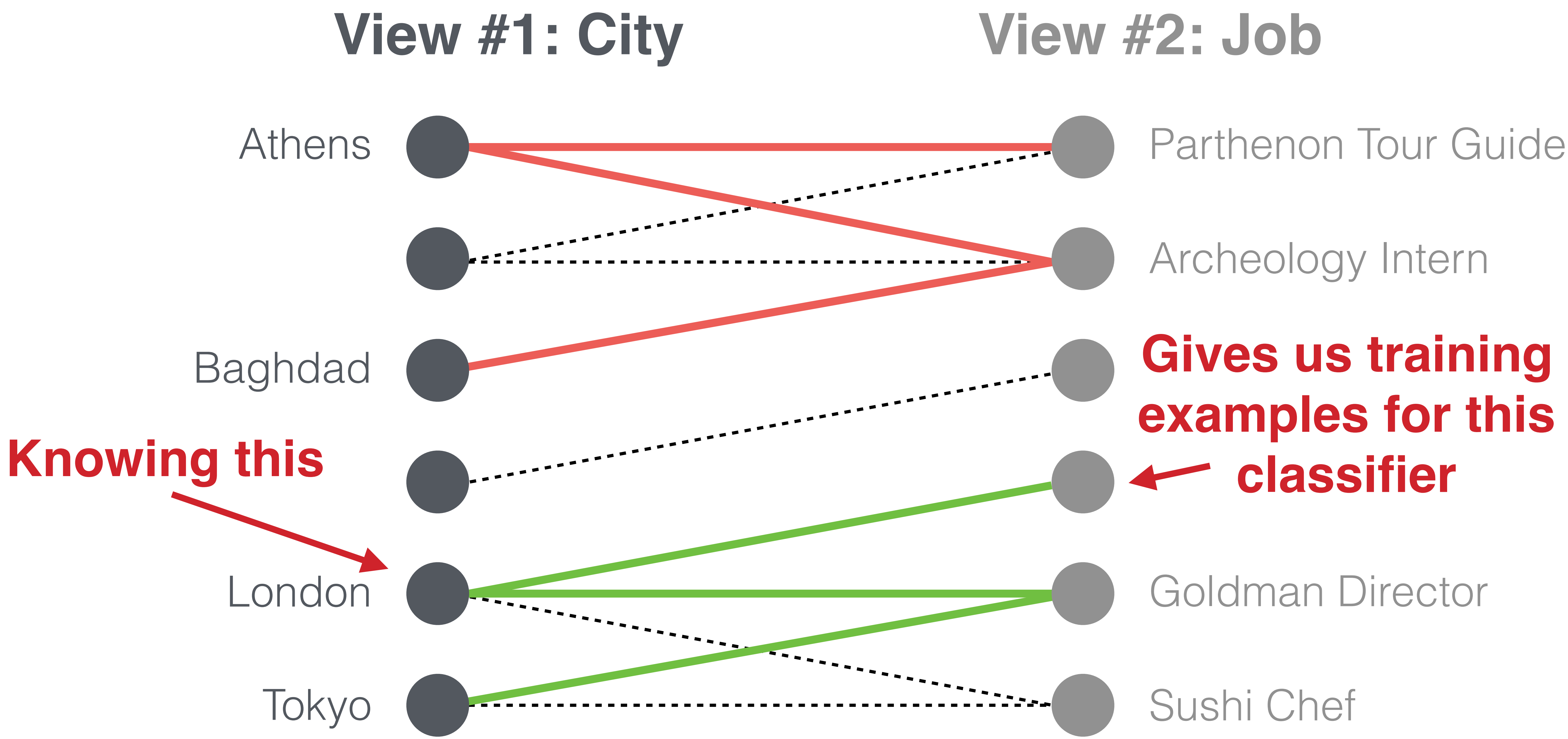
Co-Training **Intuition**



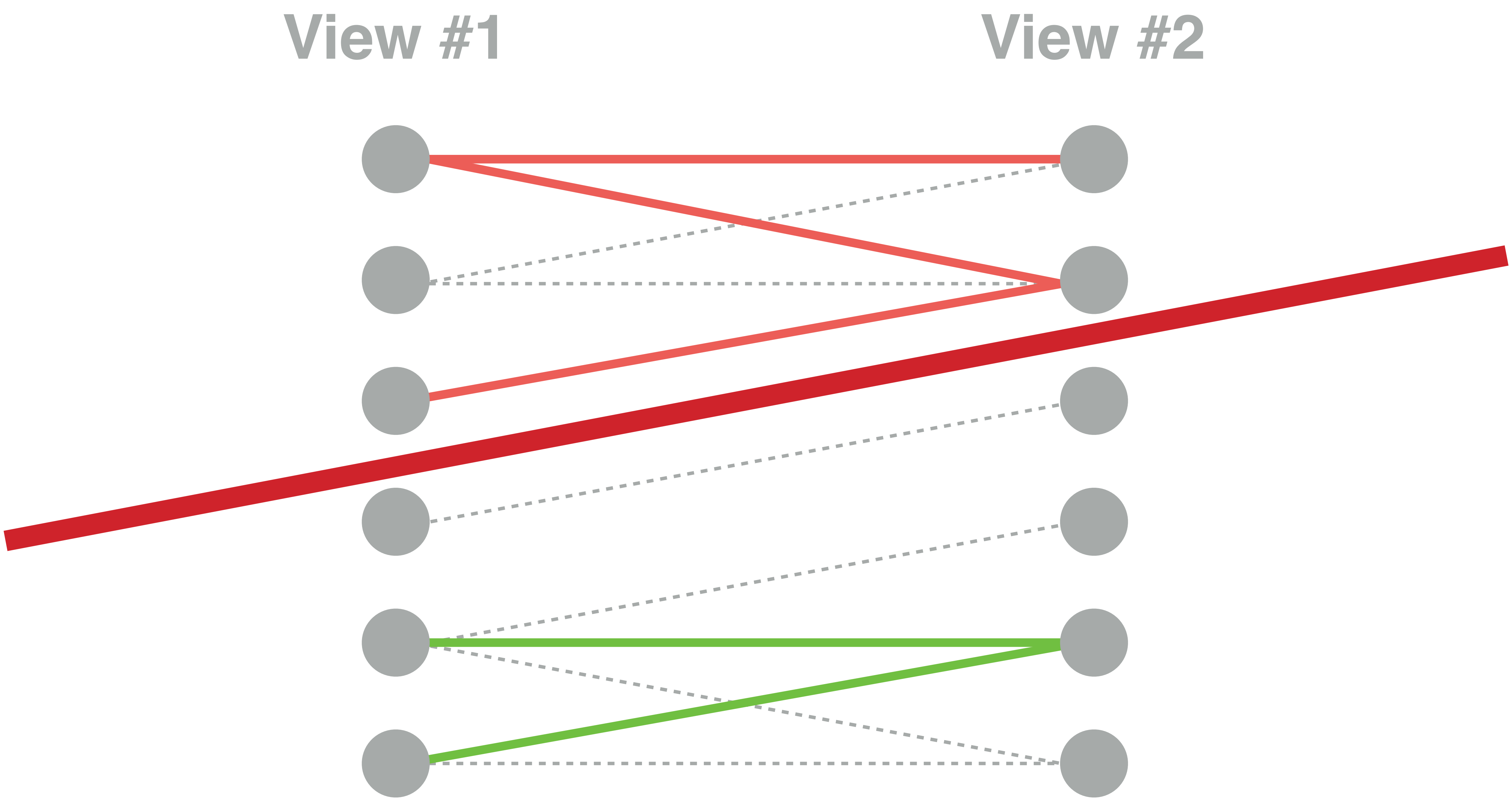
Co-Training **Intuition**



Co-Training **Intuition**



Co-Training **Assumptions**



Different class connected components are separated

Co-Training **Assumptions**

We need cases when **one classifier makes a confident prediction** and the other does not. Different approaches make different assumptions in order to derive **theoretical guarantees**:

[Blum and Mitchell, 1998]

- Independence of views given the true label
- Existence of an algorithm for learning from noise

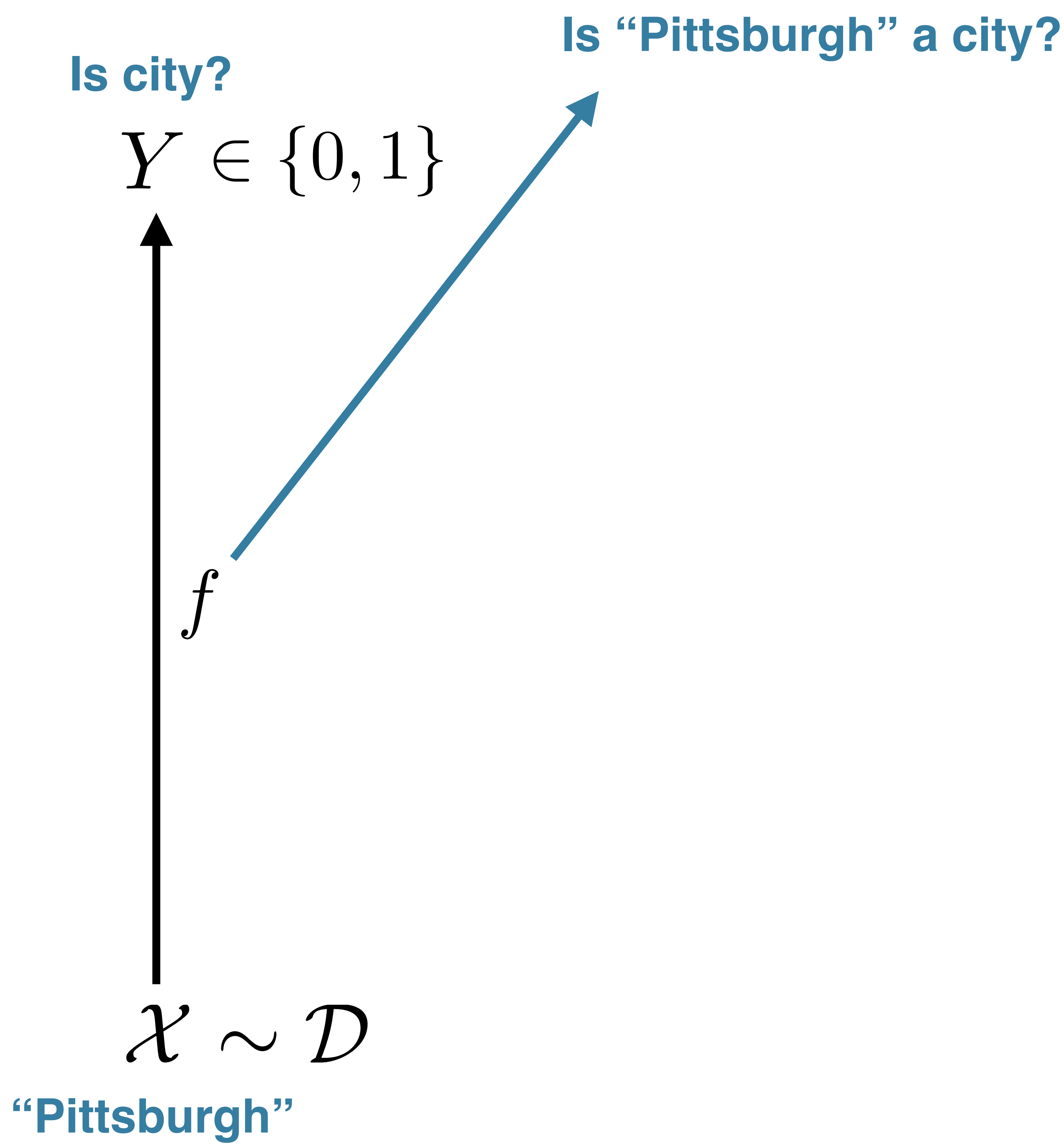
[Balcan, Blum and Yang, 2004]

- Distribution expansion (**weaker assumption**)
- Existence of an algorithm for learning from positive examples only

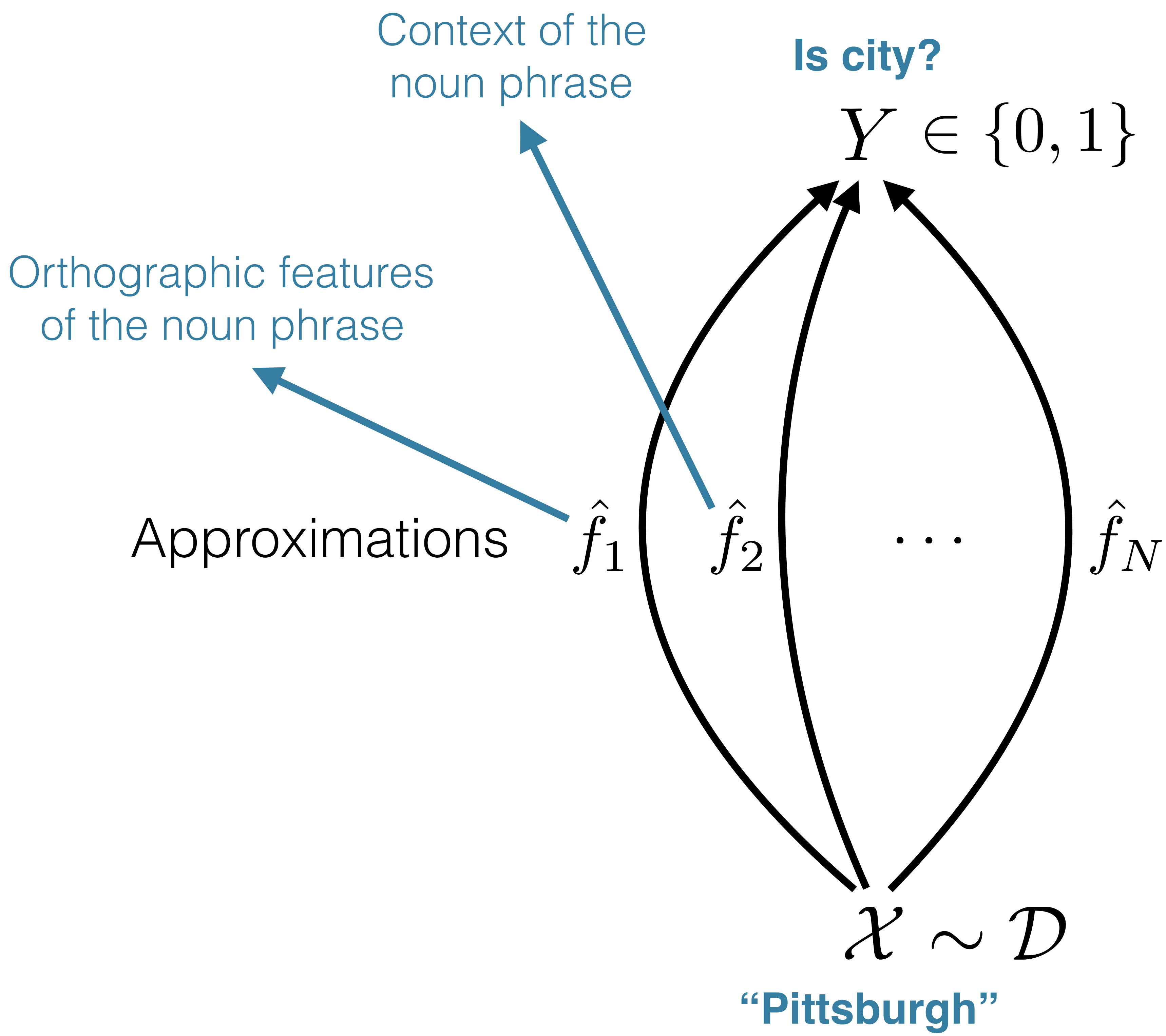
Never Ending Language Learning (NELL)

$$\begin{tikzcd} Y \in \{0,1\} \\ \mathcal{X} \sim \mathcal{D} \arrow[u, "f"'] \end{tikzcd}$$

Never Ending Language Learning (NELL)



Never Ending Language Learning (NELL)



Co-Training is core to NELL

Co-Training **Issue**

In a practical real-life setting, such as NELL, **our classifiers can make errors**. If they are confident in wrong predictions these predictions are treated as training data and these **errors can propagate** and **worsen the performance of the classifiers**.

It would be great if we could **estimate the accuracy of these classifiers** using very few labeled data, or, even better, **using only unlabeled data**.

Using **only unlabeled data** we can measure

consistency

but not

correctness

consistency

Does this implication hold?

correctness

If yes, under what conditions?

Why only unlabeled data?

It is often **impossible** to have
enough labeled data!

Never Ending Language Learning (NELL):

1. Huge knowledge-base with **thousands of functions**
2. Refined **daily** over **several years**
3. Constantly creating **new functions** automatically

Outline

1. Useful Definitions
2. Agreement Rates Method
3. Graphical Model Approaches
   1. Error Estimation
   2. Coupled Error Estimation
   3. Hierarchical Coupled Error Estimation
4. Experiments
   1. NELL Data Set
   2. Brain Data Set
5. Summary

Definition

consistency

Agreement Rate: The probability over $\mathbb{P}\left(\mathcal{X}\right) = \mathcal{D}$ of two function outputs agreeing.

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\left(\bigcap_{\substack{i,j \in \mathcal{A} \\ i \neq j}} [\hat{f}_i(X) = \hat{f}_j(X)] \right)$$

Definition

consistency

Given **unlabeled input data**, $X_1, \ldots, X_S$, we observe the **sample agreement rates**:

$$\hat{a}_{\mathcal{A}} = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left\{ \hat{f}_i(X_s) = \hat{f}_j(X_s), \forall i, j \in \mathcal{A} : i \neq j \right\}$$

Definition

correctness

Error Rate: The probability over $\mathbb{P}(\mathcal{X}) = \mathcal{D}$ of disagreeing with the correct output label.

Definition

correctness

Error Rate ⟵ $$e_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\underbrace{\left(\bigcap_{i \in \mathcal{A}} [\hat{f}_i(X) \neq Y] \right)}_{E_{\mathcal{A}}}$$

$E_{\mathcal{A}}$ ⟶ **Error Event**

Definition

correctness

$$\textbf{Error Rate} \longleftarrow e_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\Bigg(\underbrace{\bigcap_{i\in\mathcal{A}} [\hat{f}_i(X) \neq Y]}_{E_{\mathcal{A}} \longrightarrow \textbf{ Error Event}}\Bigg)$$

$$e_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i\in\mathcal{A}} [\hat{f}_i(X) = f(X)]\right)$$

Outline

1. Useful Definitions
2. **Agreement Rates Method**
3. Graphical Model Approaches
    i. Error Estimation
    ii. Coupled Error Estimation
    iii. Hierarchical Coupled Error Estimation
4. Experiments
    i. NELL Data Set
    ii. Brain Data Set
5. Summary

Agreement Rates Method

Agreement rate between $\hat{f}_i$ and $\hat{f}_j$:

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\left(E_{\{i\}} \cap E_{\{j\}}\right) + \mathbb{P}_{\mathcal{D}}\left(\bar{E}_{\{i\}} \cap \bar{E}_{\{j\}}\right)$$

Agreement Rates Method

Agreement rate between $\hat{f}_i$ and $\hat{f}_j$:

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\underbrace{\left(E_{\{i\}} \cap E_{\{j\}}\right)}^{\text{both are wrong}} + \mathbb{P}_{\mathcal{D}}\left(\bar{E}_{\{i\}} \cap \bar{E}_{\{j\}}\right)$$

Agreement Rates Method

Agreement rate between $\hat{f}_i$ and $\hat{f}_j$:

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\left(E_{\{i\}} \cap E_{\{j\}}\right) + \mathbb{P}_{\mathcal{D}}\overbrace{\left(\bar{E}_{\{i\}} \cap \bar{E}_{\{j\}}\right)}^{\text{both are right}}$$

Agreement Rates Method

Agreement rate between $\hat{f}_i$ and $\hat{f}_j$:

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\left(E_{\{i\}} \cap E_{\{j\}}\right) + \mathbb{P}_{\mathcal{D}}\left(\bar{E}_{\{i\}} \cap \bar{E}_{\{j\}}\right)$$

$$\downarrow$$

$$a_{\{i,j\}} = 1 - e_{\{i\}} - e_{\{j\}} + 2e_{\{i,j\}}$$

Probability that $\hat{f}_i$ makes an error

Probability that $\hat{f}_j$ makes an error

Probability that both make an error

Agreement Rates Method

Agreement rates and error rates are related!

$$a_{\{i,j\}} = 1 - e_{\{i\}} - e_{\{j\}} + 2e_{\{i,j\}}$$

Agreement Rates Method

Agreement rates and error rates are related!

$$a_{\{i,j\}} = 1 - e_{\{i\}} - e_{\{j\}} + 2e_{\{i,j\}}$$

Independent errors

$$e_{\{i\}}e_{\{j\}}$$

Agreement Rates Method

Agreement rates and error rates are related!

$$a_{\{i,j\}} = 1 - e_{\{i\}} - e_{\{j\}} + 2\underbrace{e_{\{i,j\}}}_{e_{\{i\}}e_{\{j\}}}$$

Independent errors

3 functions that make independent errors:

$\binom{3}{2} = 3$ equations

3 unknowns

$$e_{\{i\}} = \frac{c \pm \left(1 - 2\hat{a}_{\{j,k\}}\right)}{\pm 2\left(1 - 2\hat{a}_{\{j,k\}}\right)}$$

where $i \in \{1,2,3\}$, $j,k \in \{1,2,3\} \backslash i$ with $j < k$ and:

$$c = \sqrt{\left(2\hat{a}_{\{1,2\}} - 1\right)\left(2\hat{a}_{\{1,3\}} - 1\right)\left(2\hat{a}_{\{2,3\}} - 1\right)}$$

Agreement Rates Method

Independent errors → **Too strong** assumption → We do not make it

But we end up with **more unknowns than equations**

Agreement Rates Equations

Objective Function

Constraints

Constrained Optimization Problem

Agreement Rates Method

The objective function tries to **minimize the dependence** between the error rates:

$$c(\boldsymbol{e}) = \sum_{i,j \in \{1,\ldots,N\}} \left(e_{\{i,j\}} - e_{\{i\}} e_{\{j\}} \right)^2$$

Relaxes the independence assumption

More **constraints:**

$$e_{\{i,j\}} \leq \min \left\{ e_{\{i\}}, e_{\{j\}} \right\}$$

Agreement Rates Method

Agreement rate for a bigger set of functions:

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} E_i \right) + \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} \bar{E}_i \right)$$

Agreement Rates Method

Agreement rate for a bigger set of functions:

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\overbrace{\left(\bigcap_{i \in \mathcal{A}} E_i\right)}^{\text{all are wrong}} + \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} \bar{E}_i\right)$$

Agreement Rates Method

Agreement rate for a bigger set of functions:

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} E_i\right) + \mathbb{P}_{\mathcal{D}}\left(\overbrace{\bigcap_{i \in \mathcal{A}} \bar{E}_i}^{\text{all are right}}\right)$$

Agreement Rates Method

Agreement rate for a bigger set of functions:

$$a_{\mathcal{A}} = \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} E_i \right) + \mathbb{P}_{\mathcal{D}}\left(\bigcap_{i \in \mathcal{A}} \bar{E}_i \right)$$

$$\Downarrow$$

$$a_{\mathcal{A}} = e_{\mathcal{A}} + 1 + \sum_{k=1}^{|\mathcal{A}|} \left[(-1)^k \sum_{\substack{I \subseteq \mathcal{A} \\ |I| = k}} e_I \right]$$

Agreement Rates Method

Objective function:

$$c(\boldsymbol{e}) = \sum_{\mathcal{A}:|\mathcal{A}|\geq 2} \left(e_{\mathcal{A}} - \prod_{i \in \mathcal{A}} e_i \right)^2$$

Inequality constraints:

$$e_{\mathcal{A}} \leq \min_{i \in \mathcal{A}} e_{\mathcal{A} \setminus i}$$

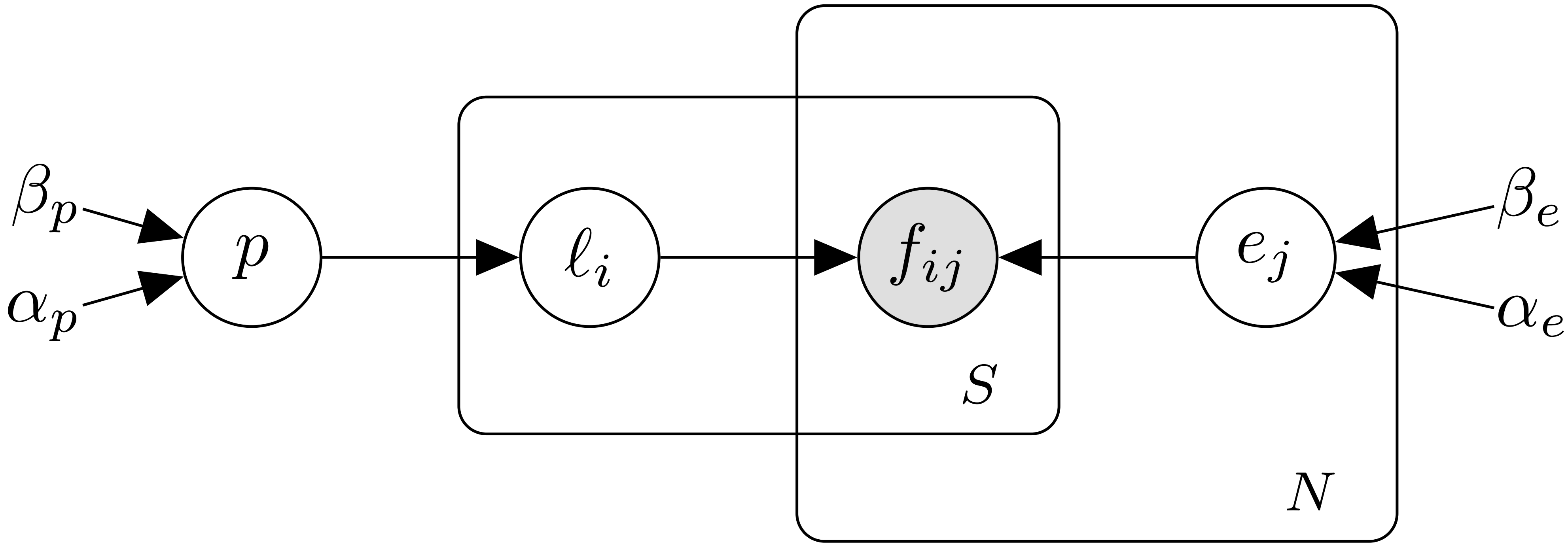
Outline

1. Useful Definitions
2. Agreement Rates Method
3. **Graphical Model Approaches**
    i. **Error Estimation**
    ii. **Coupled Error Estimation**
    iii. **Hierarchical Coupled Error Estimation**
4. Experiments
    i. NELL Data Set
    ii. Brain Data Set
5. Summary

Error Estimation

We designed a **generative process** describing how our observations are generated.

Error Estimation



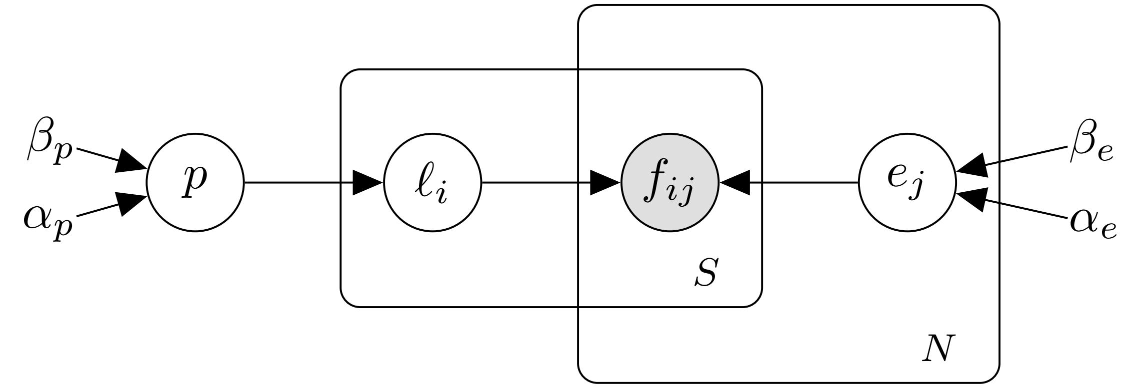
Label Prior ← $p \sim \text{Beta}(\alpha_p, \beta_p),$

True Labels ← $\ell_i \sim \text{Bernoulli}(p), \text{ for } i = 1, \ldots, S,$

Error Rates ← $e_j \sim \text{Beta}(\alpha_e, \beta_e), \text{ for } j = 1, \ldots, N,$

Actual Outputs ← $\hat{f}_{ij} = \begin{cases} \ell_i & , \text{ with probability } 1 - e_j, \\ 1 - \ell_i & , \text{ otherwise.} \end{cases}$

Error Estimation



We use **Gibbs sampling** to perform inference:

$$P(p \mid \cdot) = \mathrm{Beta}(\alpha_p + \sigma_{\boldsymbol{\ell}}, \beta_p + S - \sigma_{\boldsymbol{\ell}}),$$

$$P(\ell_i \mid \cdot) \propto p^{\ell_i}(1-p)^{1-\ell_i}\pi_i,$$

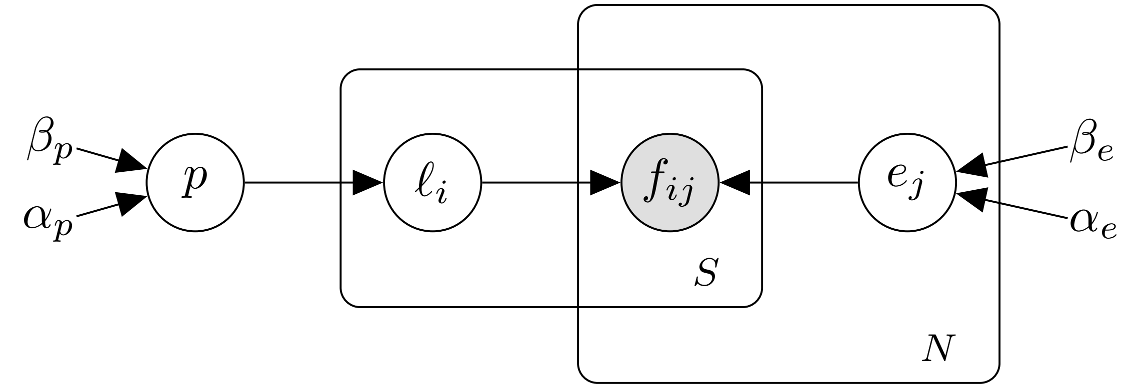
$$P(e_j \mid \cdot) = \mathrm{Beta}(\alpha_e + \sigma_j, \beta_e + S - \sigma_j),$$

where:

$$\sigma_{\boldsymbol{\ell}} = \sum_{i=1}^{S} \ell_i, \quad \sigma_j = \sum_{i=1}^{S} \mathbb{1}_{\{\hat{f}_{ij} \neq \ell_i\}},$$

$$\pi_i = \prod_{j=1}^{N} e_j^{\mathbb{1}_{\{\hat{f}_{ij} \neq \ell_i\}}} (1 - e_j)^{\mathbb{1}_{\{\hat{f}_{ij} = \ell_i\}}}.$$

Error Estimation



We use **Gibbs sampling** to perform inference:

$$P(p \mid \cdot) = \mathrm{Beta}(\alpha_p + \sigma_{\boldsymbol{\ell}}, \beta_p + S - \sigma_{\boldsymbol{\ell}}),$$
$$P(\ell_i \mid \cdot) \propto p^{\ell_i}(1-p)^{1-\ell_i}\pi_i,$$
$$P(e_j \mid \cdot) = \mathrm{Beta}(\alpha_e + \sigma_j, \beta_e + S - \sigma_j),$$

where:

$$\sigma_{\boldsymbol{\ell}} = \sum_{i=1}^{S} \ell_i, \quad \sigma_j = \sum_{i=1}^{S} \mathbb{1}_{\{\hat{f}_{ij} \neq \ell_i\}},$$

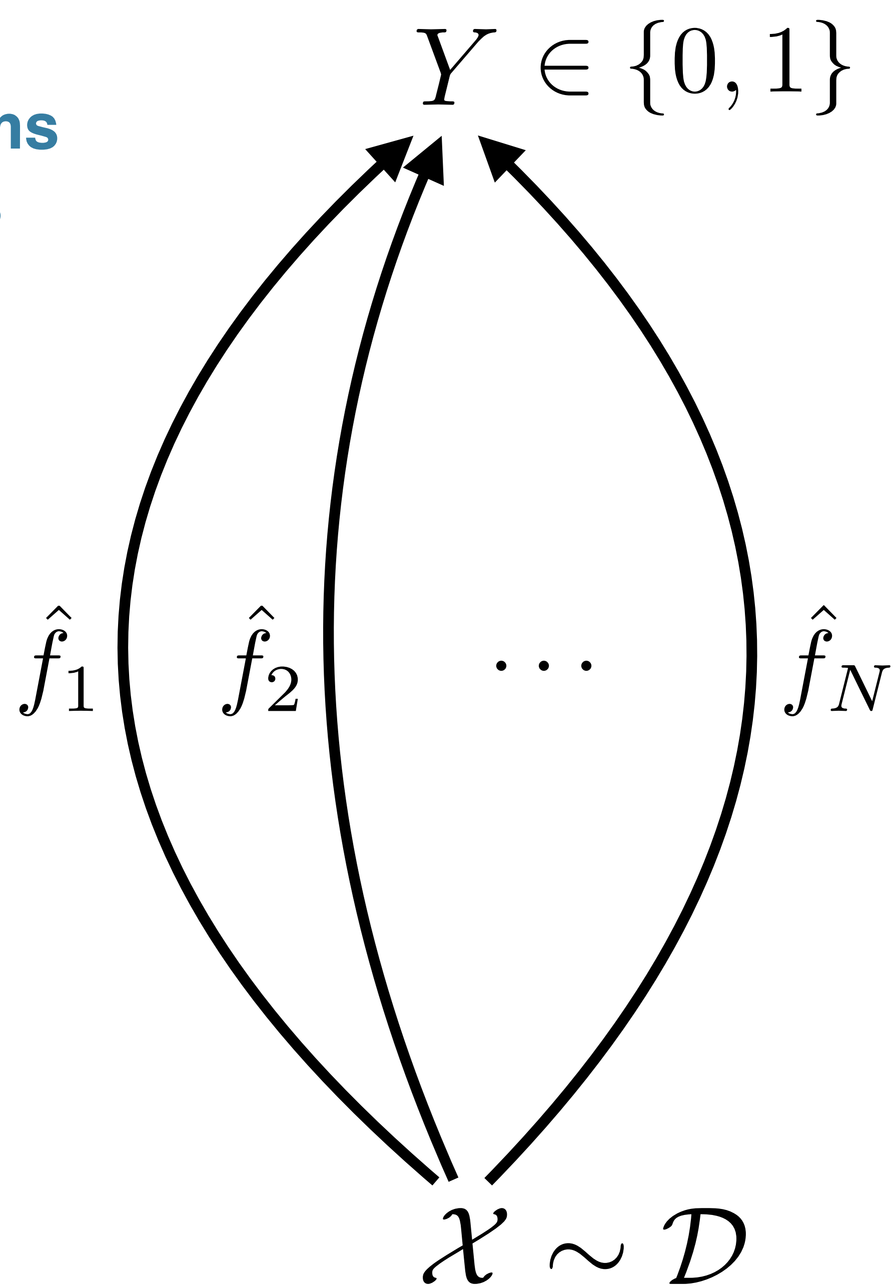
$$\pi_i = \prod_{j=1}^{N} e_j^{\mathbb{1}_{\{\hat{f}_{ij} \neq \ell_i\}}} (1-e_j)^{\mathbb{1}_{\{\hat{f}_{ij} = \ell_i\}}}.$$

Disagreement Rate

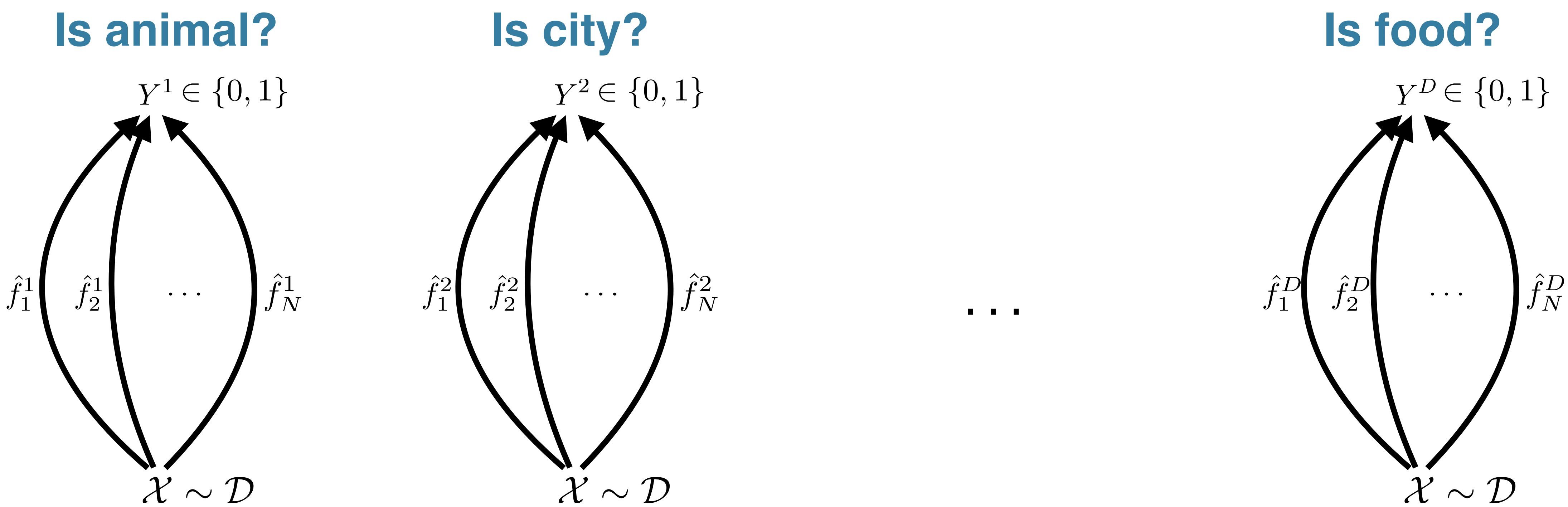
Single Domain Settings So Far

Is city?

We refer to different **classification problems** as different **domains**

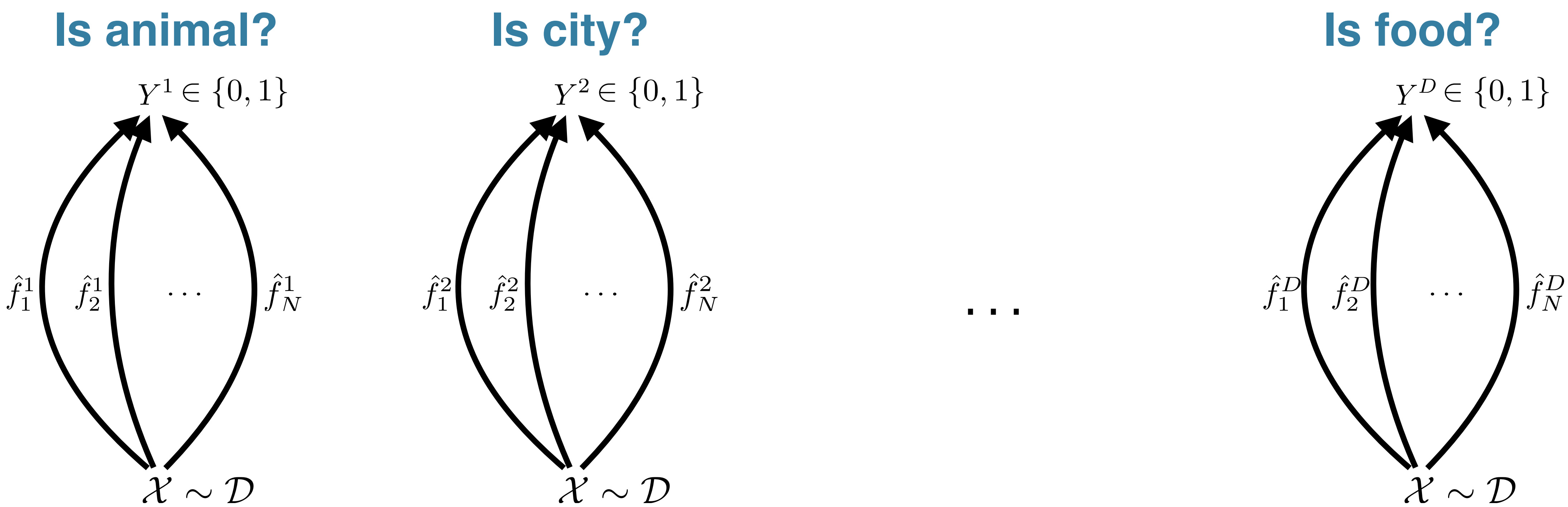


What About **Multiple Domains**?



We have functions of the **same parametric form** using the same input data and features, answering **different questions**!

What About **Multiple Domains**?

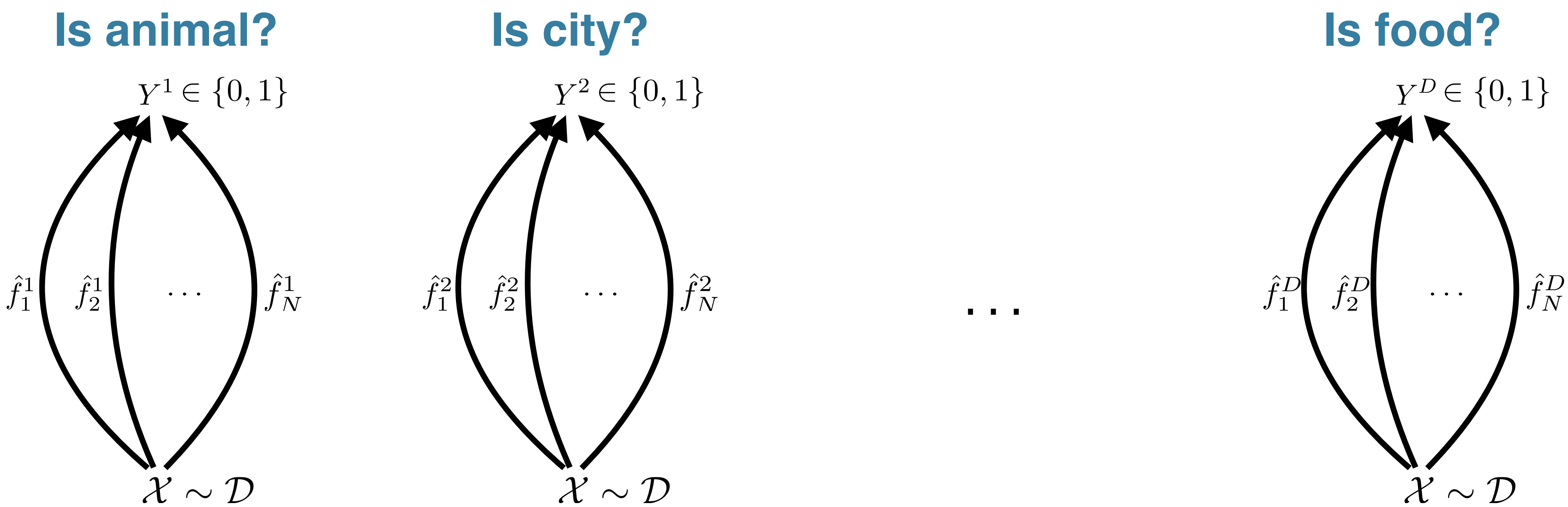


We have functions of the **same parametric form** using the same input data and features, answering **different questions**!

We could potentially gain by **sharing information** across those accuracy estimation problems.

We can **cluster the functions across domains**.

What About **Multiple Domains**?

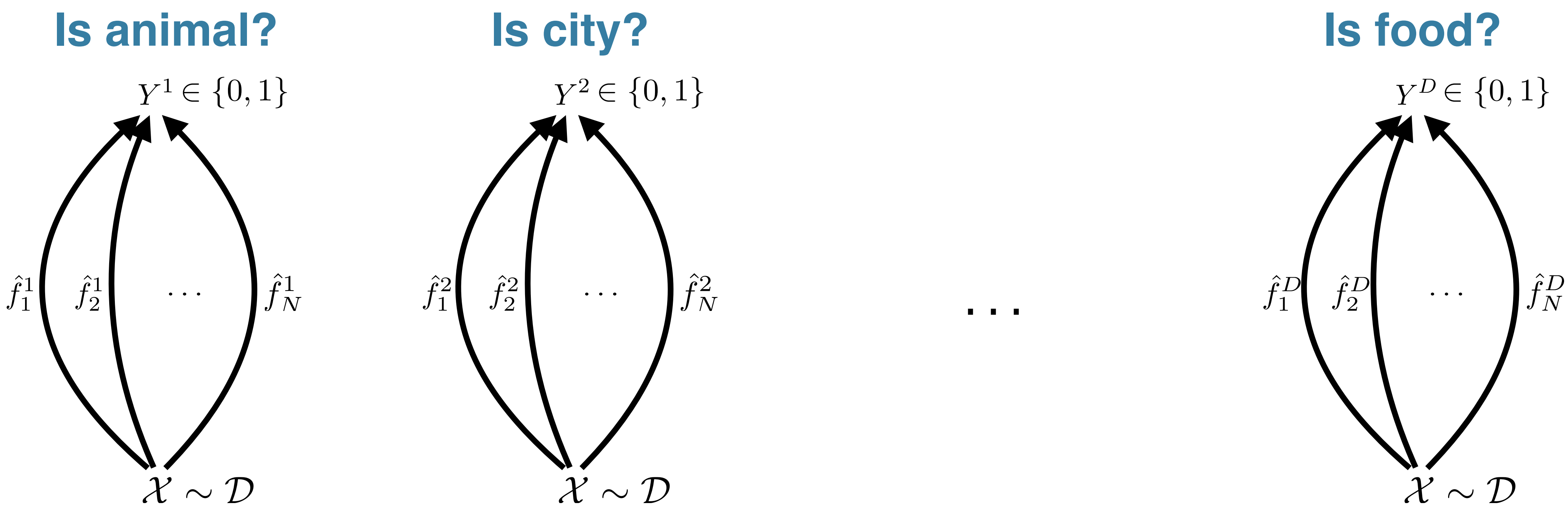


Coupled Error Estimation

We could potentially gain by **sharing information** across those accuracy estimation problems.

We can **cluster the functions across domains**.

What About **Multiple Domains**?



Hierarchical Coupled Error Estimation

We can **further cluster error rates across functions** to share even more information in a structured manner.

Note that this sharing of information can in general be very **useful** in the case of **limited data**.

Outline

1. Useful Definitions
2. Agreement Rates Method
3. Graphical Model Approaches
   i. Error Estimation
   ii. Coupled Error Estimation
   iii. Hierarchical Coupled Error Estimation
4. **Experiments**
   i. NELL Data Set
   ii. Brain Data Set
5. Summary

Experiments

We report the **error mean absolute deviation (MAD_{error})** between:

- True error rates (estimated from labeled data)
- Error rates estimates from unlabeled data

and the **label mean absolute deviation (MAD_{label})** between:

- True labels
- Predicted labels

Note that this is simply the **label accuracy**.

For the agreement rates method we used the **IpOpt 3.11.9** interior point optimization solver and all the methods were implemented in **Java**.

Experiments

1. NELL Data Set
2. Brain Data Set

Experiments

1 NELL Data Set

Task: *Predict whether a noun phrase (NP) belongs to a category (e.g. city)*

4 logistic regression classifiers using different features:

ADJ: Adjectives that occur with the NP
CMC: Orthographic features of the NP
CPL: Phrases that occur with the NP
VERB: Verbs that appear with the NP

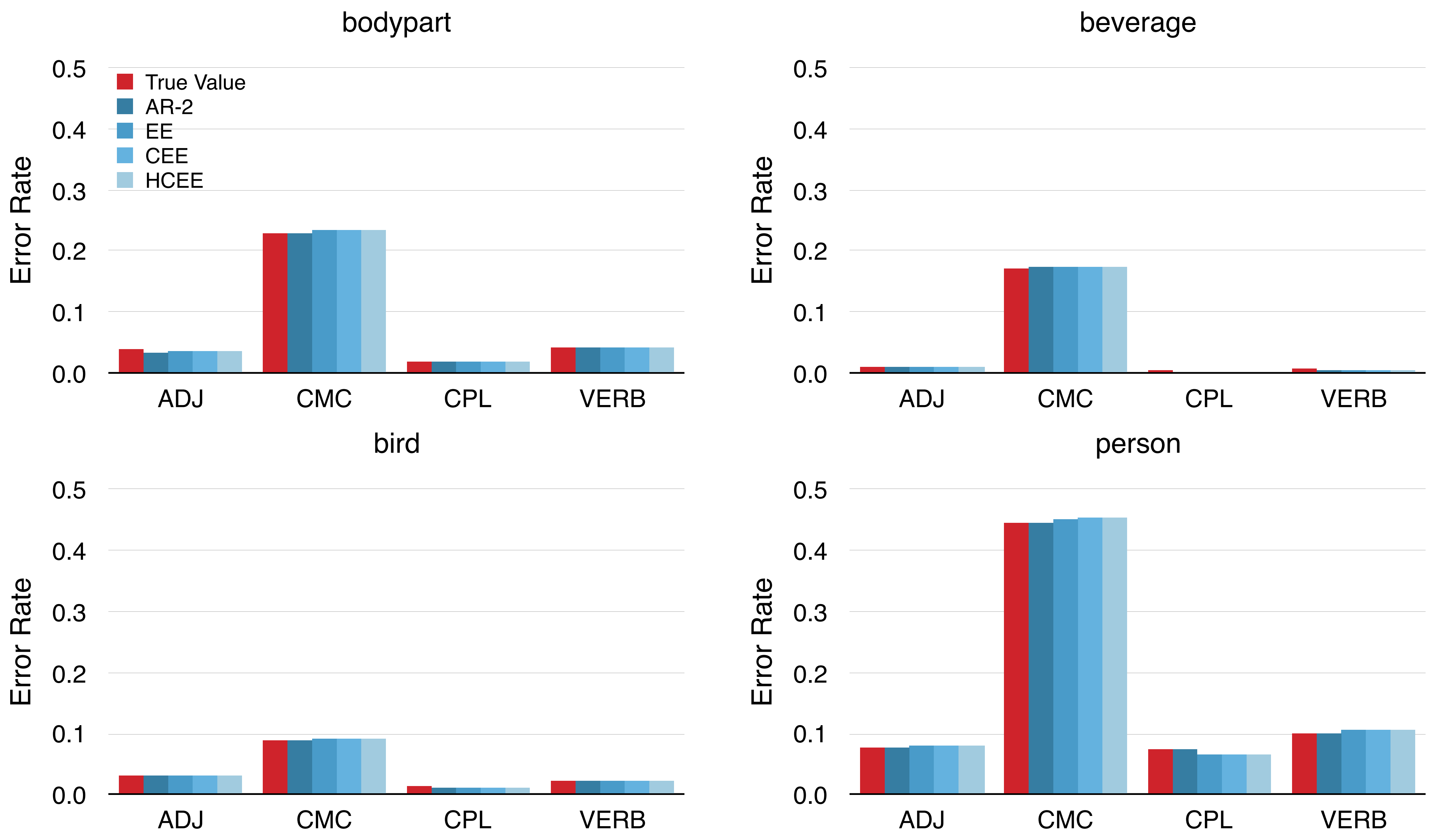
Category	# Examples
animal	20,733
beverage	18,932
bird	19,263
bodypart	21,840
city	21,778
disease	21,827
drug	20,452
fish	19,162
food	19,566
fruit	18,911
muscle	21,606
person	21,700
protein	21,811
river	21,723
vegetable	18,826

Experiments

1 NELL Data Set

True error rates (estimated from labeled data)

Error rates estimated from unlabeled data



Experiments

1 NELL Data Set

4 functions without the independence assumption:

$\times 10^{-2}$	All Data Samples		10% of Data Samples	
	MAD_{error}	MAD_{label}	MAD_{error}	MAD_{label}
MAJ	-	5.60	-	5.47
AR-2	0.59	2.21	1.00	2.36
AR	0.66	2.20	0.70	2.36
EE	**0.29**	0.96	0.65	1.32
CEE	0.31	**0.94**	0.58	0.96
HCEE	0.31	0.96	**0.31**	**0.95**

3 functions under independence assumption: 2.82×10^{-2}.

Experiments

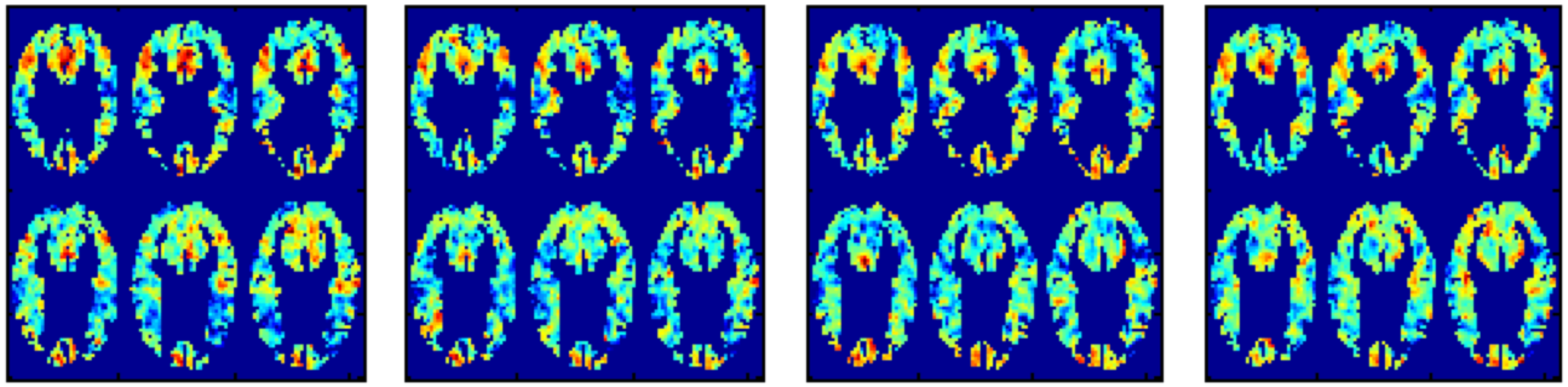
2 Brain Data Set

Task: *Find which of two 40 second long story passages corresponds to an unlabeled 40 second time series of fMRI neural activity*

11 logistic regression classifiers using a different representation of the text passage. For example:

- Number of letters in each word
- Part of speech tag of each word
- Emotions experienced by characters in the story
- etc.

1,000 labeled samples for 11 brain regions

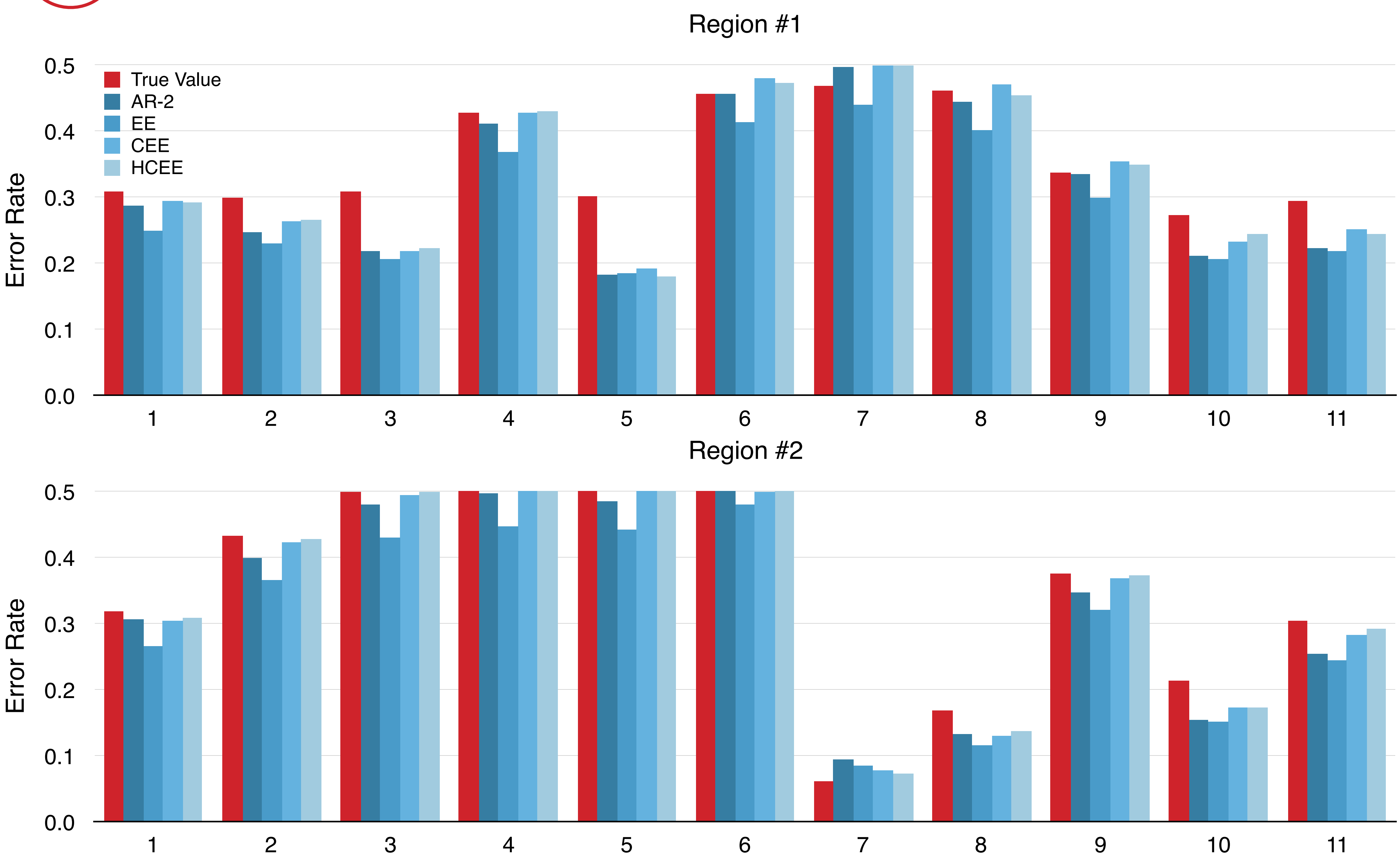


Experiments

True error rates (estimated from labeled data)

Error rates estimated from unlabeled data

2 Brain Data Set

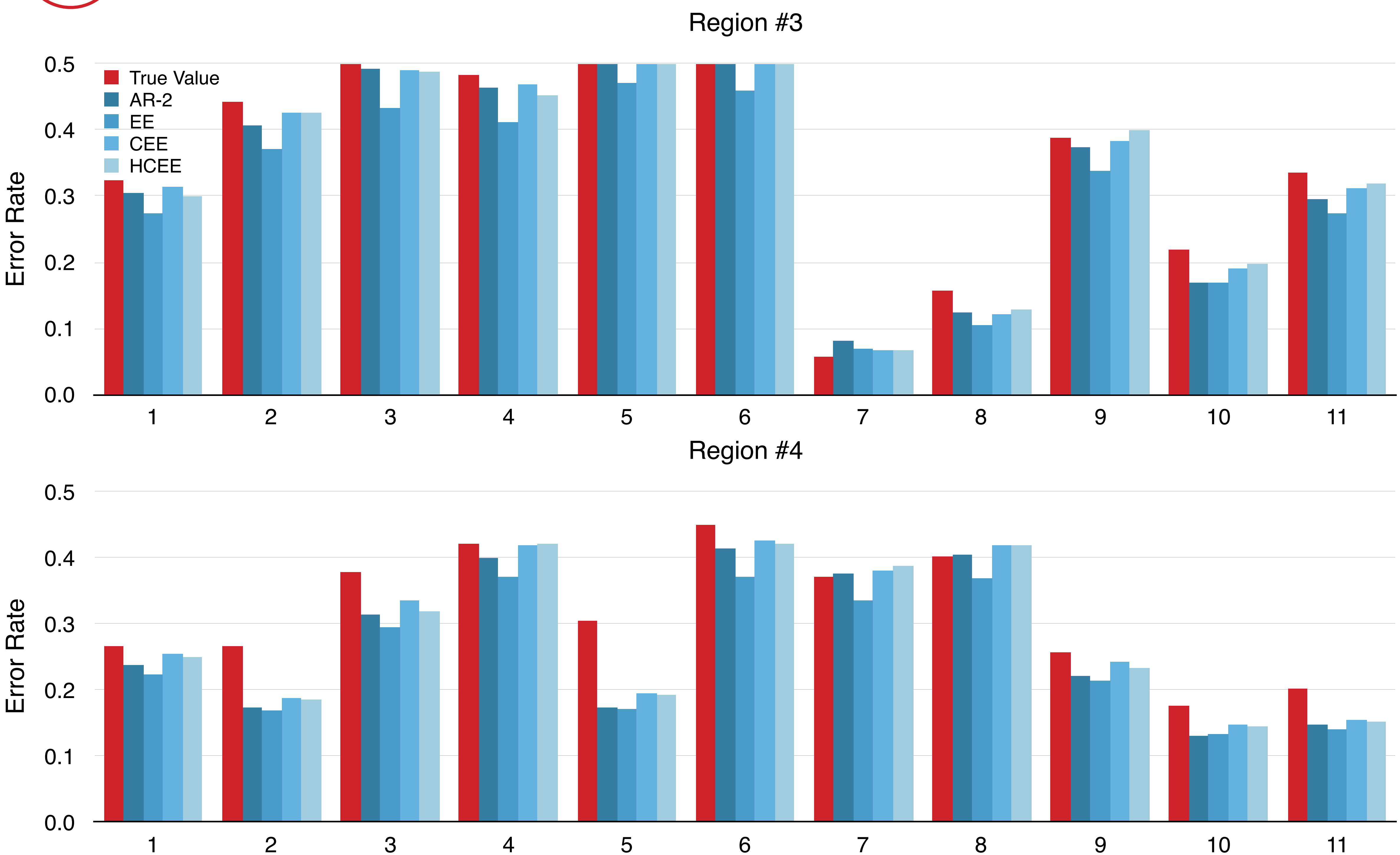


Experiments

True error rates (estimated from labeled data)

Error rates estimated from unlabeled data

2 Brain Data Set



Experiments

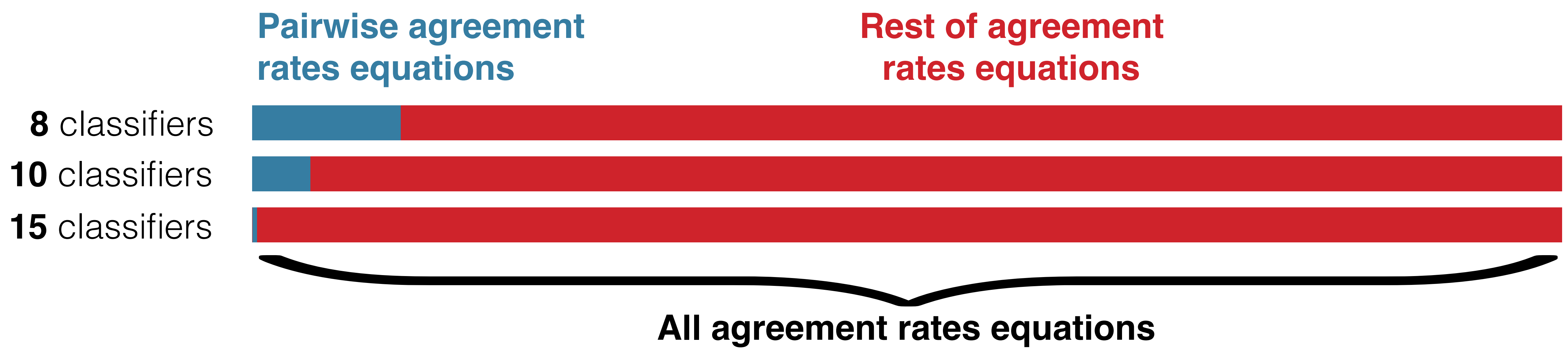
2 Brain Data Set

$\times 10^{-2}$	All Data Samples		10% of Data Samples	
	MAD_{error}	MAD_{label}	MAD_{error}	MAD_{label}
MAJ	-	19.82	-	20.82
AR-2	5.14	18.67	5.84	20.14
AR	15.29	19.82	14.96	19.86
EE	6.77	**17.23**	20.20	20.03
CEE	4.07	17.51	**4.69**	**17.42**
HCEE	**4.04**	17.34	5.74	18.51

Experiments

High order sample agreement rates are often bad estimates of the actual agreement rates

2 Brain Data Set



$\times 10^{-2}$	Pairwise Agreement Rates		All Agreement Rates	
	NELL	NELL 10%	NELL	NELL 10%
MAD_{error}	**0.59**	1.00	0.66	**0.70**
MAD_{label}	2.21	**2.36**	**2.20**	**2.36**

Runs 4 times faster and performs as well on average!

Accuracy Estimation Summary

Estimating binary functions' **error rates** using **unlabeled data**

4 Methods presented

↓

1 formulated as an optimization problem and 3 graphical models

Highly accurate error rates estimates

↓

on two very different data sets

Much higher than when making the independence assumption

Accuracy Estimation Summary

Estimating binary functions' **error rates** using **unlabeled data**

4 Methods presented

1 formulated as an optimization problem and 3 graphical models

Highly accurate error rates estimates

on two very different data sets

Much higher than when making the independence assumption

consistency

correctness

Accuracy Estimation Summary

Estimating binary functions' **error rates** using **unlabeled data**

Extend to non-boolean, discrete-valued and even **real-valued functions**

Use those error rates in the context of **self-reflection**

4 Methods presented

1 formulated as an optimization problem and 3 graphical models

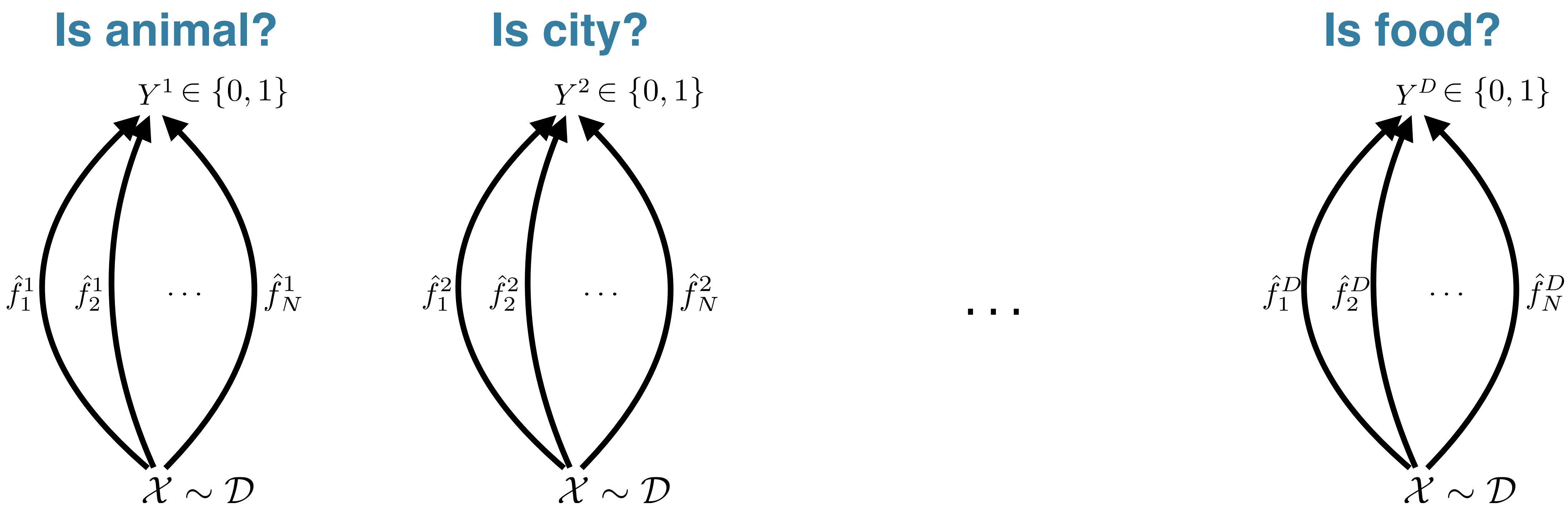
Try using **different objective functions** for AR

Highly accurate error rates estimates

on two very different data sets

Much higher than when making the independence assumption

What About **Multiple Domains**?



Logic Error Estimation

What if there are **constraints** between the domains? What if "city" and "animal" are **mutually exclusive**, for example?

If two classifiers say that a NP is both a city and animal at the same time, then at least one of them has to be making a mistake

Related Work

Disagreement rate as **distance metric for model selection and regularization** [Schuurmans et al., 2006; Bengio and Chapados, 2003].

Use of disagreement along with an ontology to **estimate the error of the prediction vector for multi-class prediction**, from unlabeled data, under an **assumption of independence of the input features given the labeling** [Balcan et. al., 2013].

Work at developing **more robust semi-supervised learning algorithms** by using the concept of agreement rates [Collins and Singer, 1999] or some task specific constraints [Chang et al., 2007].

Bounding error rates using the pairwise agreement rates only, under the **assumption that the functions make independent errors** [Dasgupta et. al., 2011].

Estimation of average error rate of two predictors using their disagreement rate [Madani et. al., 2004].

Estimation of per-function prediction risk, under the **assumption that the true probability distribution of the output labels is known** [Donmez et. al., 2010].

Questions?