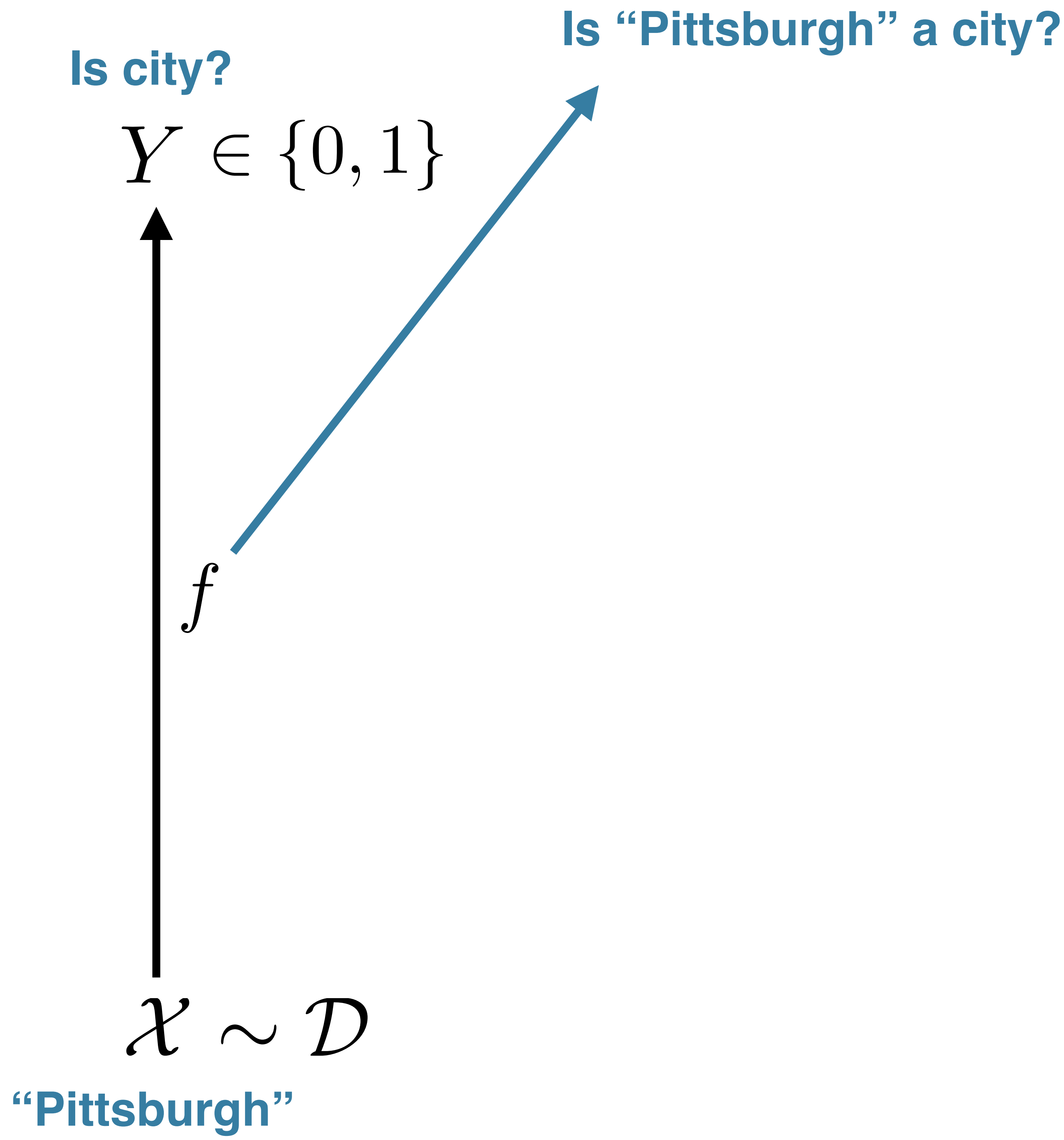
Estimating Accuracy from Unlabeled Data

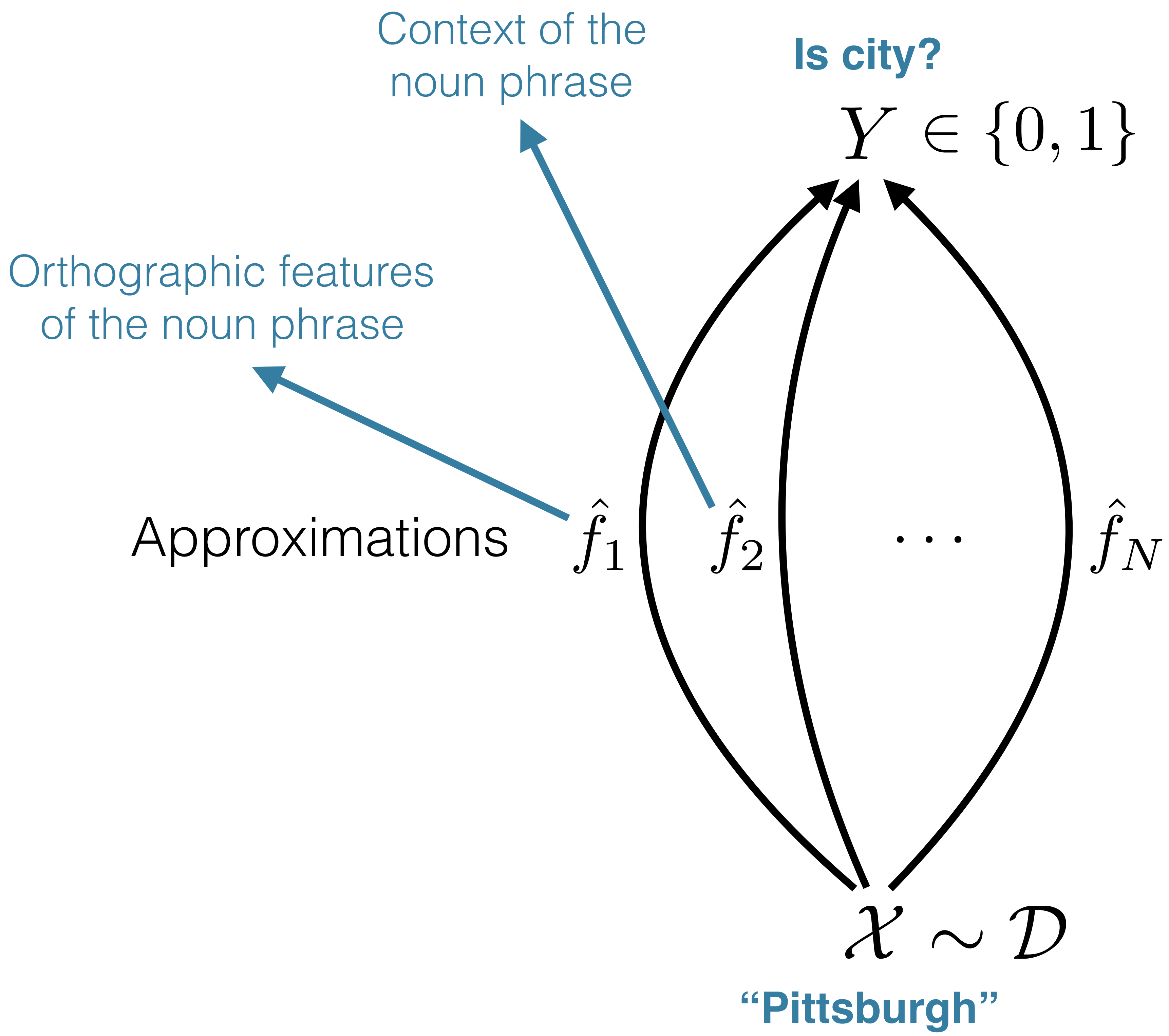
A Bayesian Approach

Anthony Platanios, Avinava Dubey, and Tom Mitchell

Machine Learning Department
Carnegie Mellon University

Presented by Anthony Platanios


Is city?
Is “Pittsburgh” a city?
$Y \in \{0, 1\}$
f
$\mathcal{X} \sim \mathcal{D}$
“Pittsburgh”


Context of the noun phrase
Is city?
$Y \in \{0, 1\}$
Orthographic features of the noun phrase
Approximations
$\hat{f}_1$ $\hat{f}_2$ $\cdots$ $\hat{f}_N$
$\mathcal{X} \sim \mathcal{D}$
"Pittsburgh"

Using **only unlabeled data** we can measure

consistency

but not

correctness

consistency

↓

Does this implication hold?

correctness

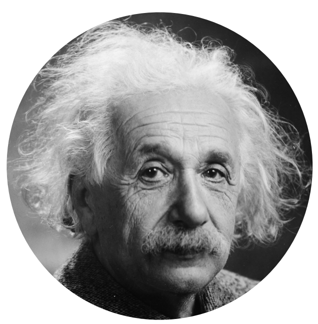
If yes, under what conditions?

Is quantum physics probabilistic?

Independent Groups of Scientists

Schrödinger

Planck



Einstein

Cramer

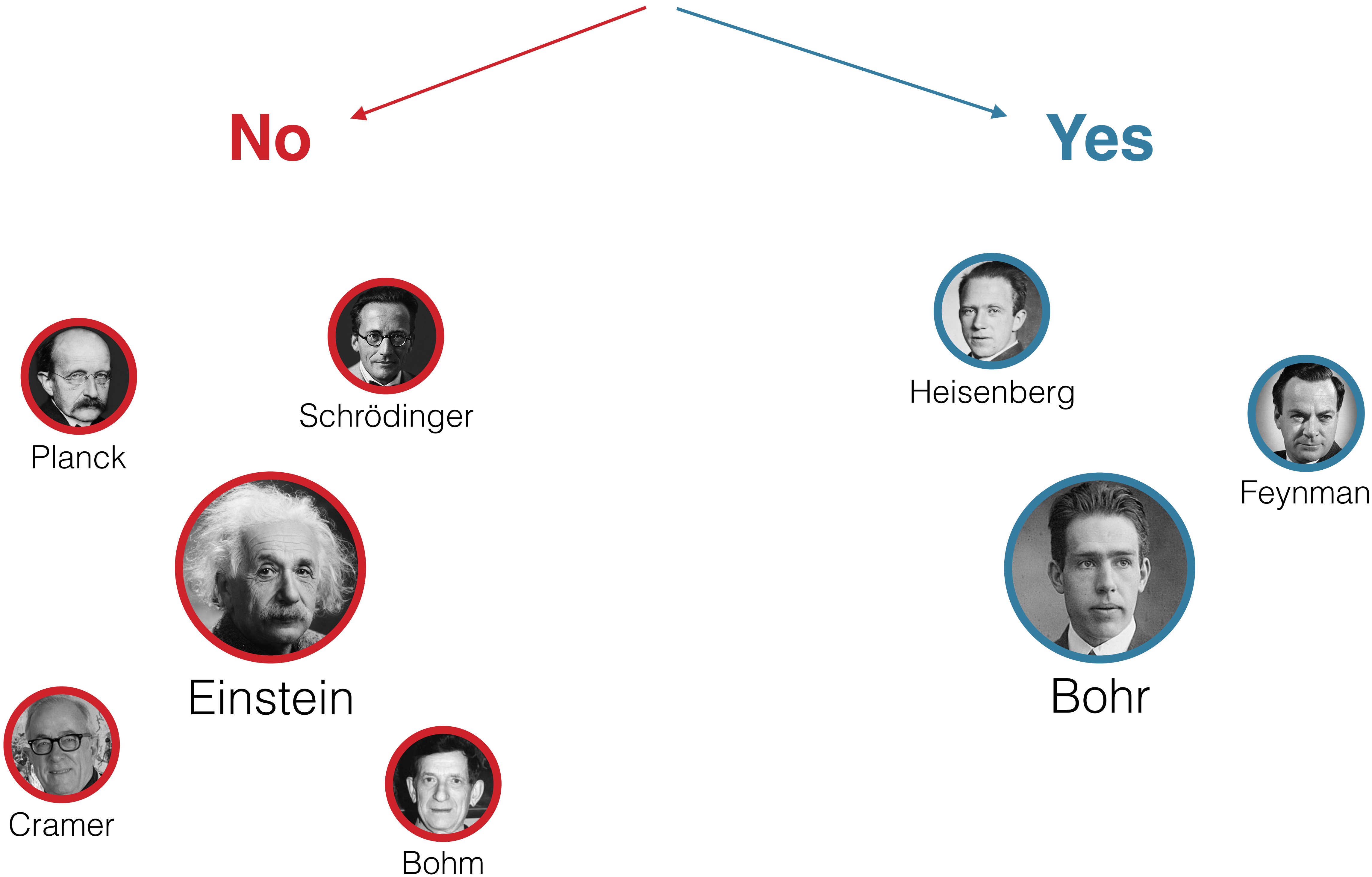
Bohm

Heisenberg

Feynman

Bohr

Is quantum physics probabilistic?



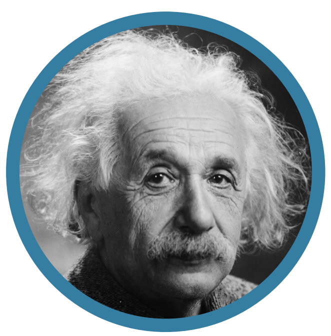
Is quantum physics probabilistic?

No

Yes

Planck

Schrödinger



Einstein

Cramer

Bohm

What if he switched sides?

Heisenberg

Feynman

Bohr

It becomes more likely that the correct answer is "Yes"

Is quantum physics probabilistic?

No

Yes

Planck
Schrödinger
Heisenberg
Feynman
Einstein
Bohr
Cramer
Bohm
Dirac
Hilbert

consistency

Does this implication hold?

correctness

If yes, under what conditions?

dependency

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

Definition

consistency

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

$$a_{\{i,j\}} = \mathbb{P}_{\mathcal{D}}\left(\hat{f}_i\left(X\right) = \hat{f}_j\left(X\right)\right)$$

Definition

consistency

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

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

Definition

correctness

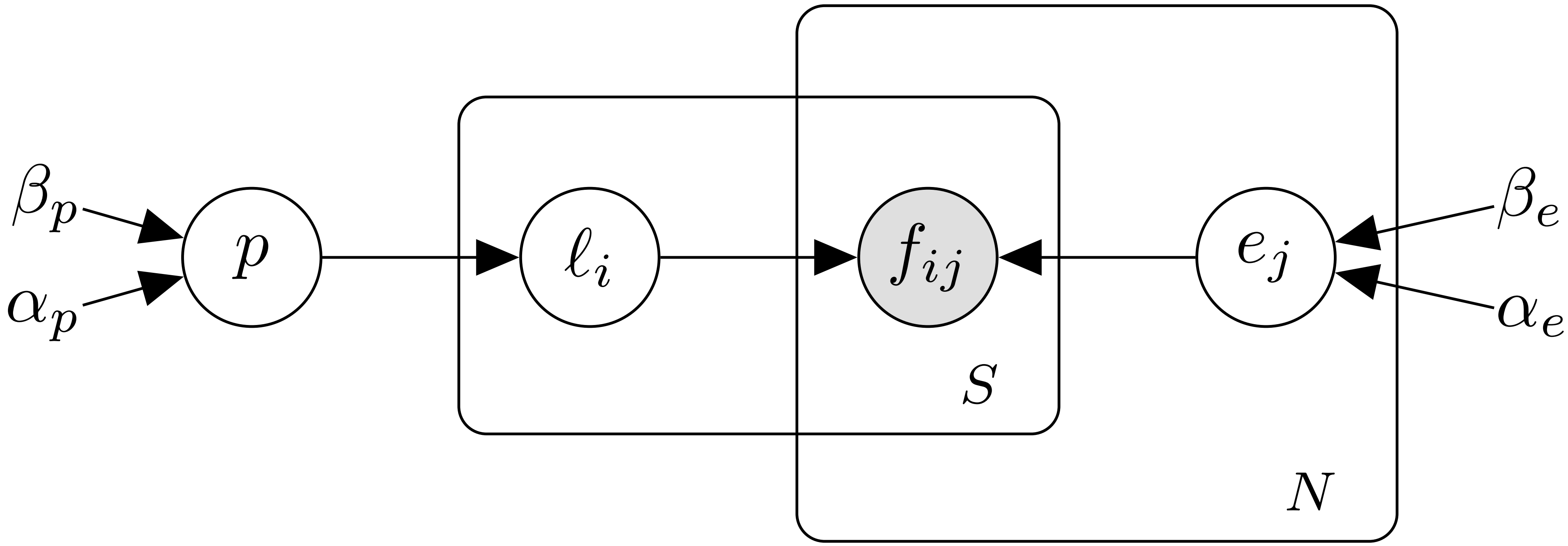
Error Rate: The probability over $\mathbb{P}\left(\mathcal{X}\right) = \mathcal{D}$ of disagreeing with the correct output label.

$$e_i = \mathbb{P}_{\mathcal{D}}\left(f_i(X) \neq Y\right)$$

Error Estimation

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

Error Estimation



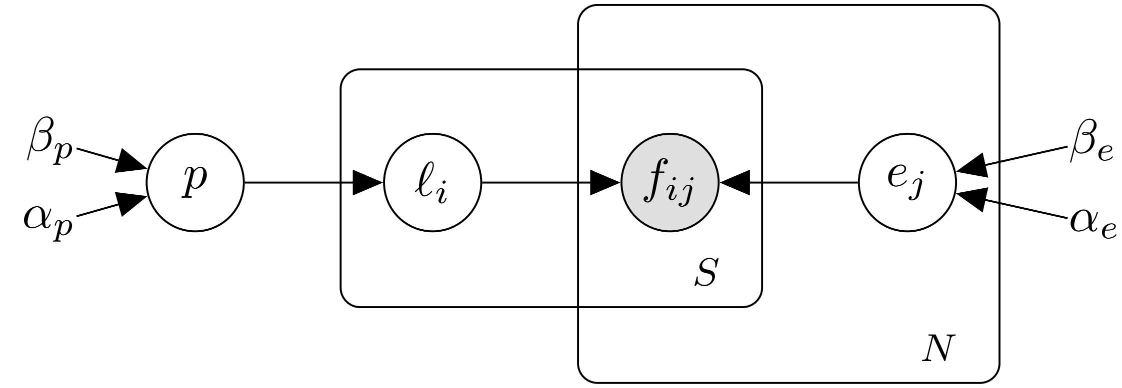
Label Prior ← $p \sim \text{Beta}(\alpha_p, \beta_p),$

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:

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

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

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

where:

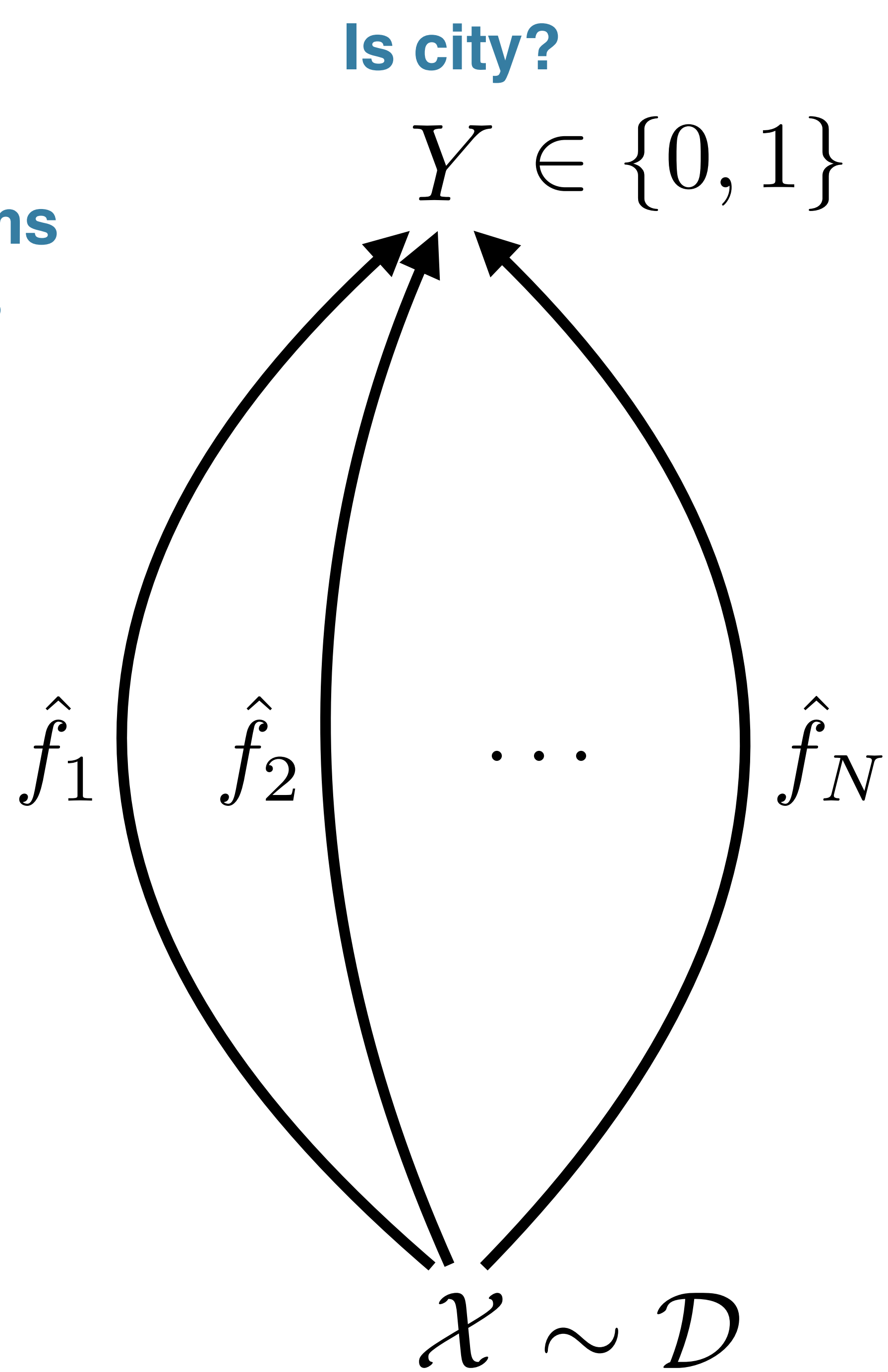
$$\sigma_{\boldsymbol{\ell}} = \sum_{i=1}^{S} \ell_i, \quad \sigma_j = \sum_{i=1}^{S} \mathbb{1}_{\{\hat{f}_{ij} \neq \ell_i\}},$$

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

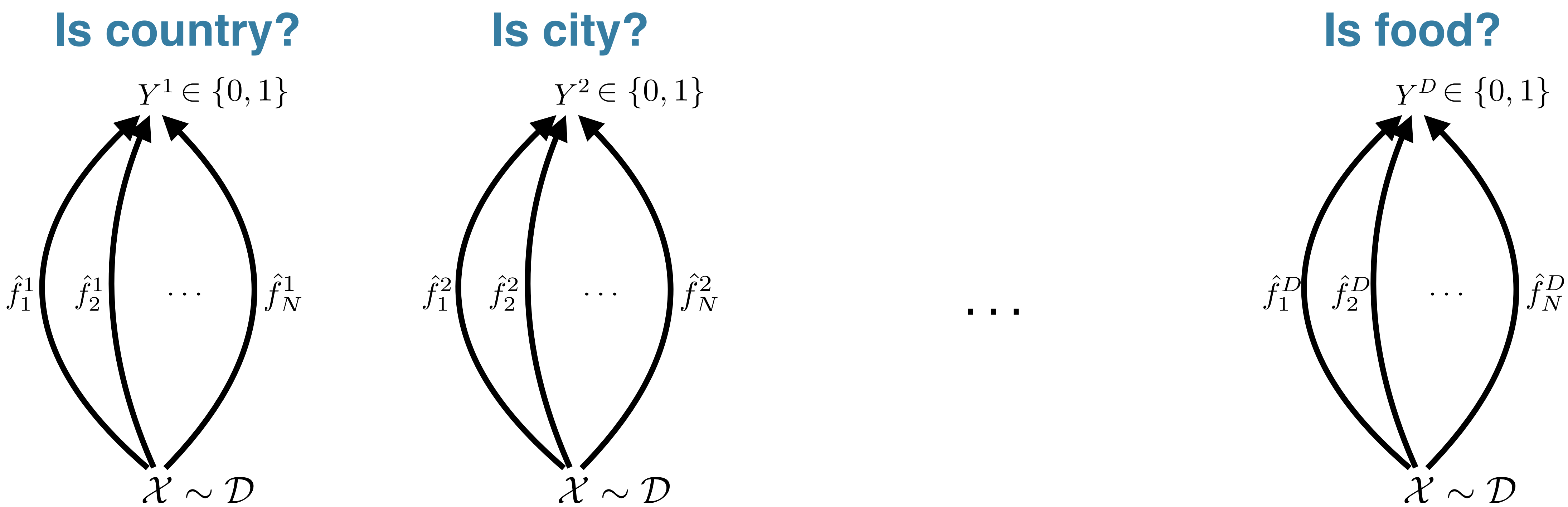
Disagreement Rate

Single Domain Settings So Far

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



What About **Multiple Domains**?

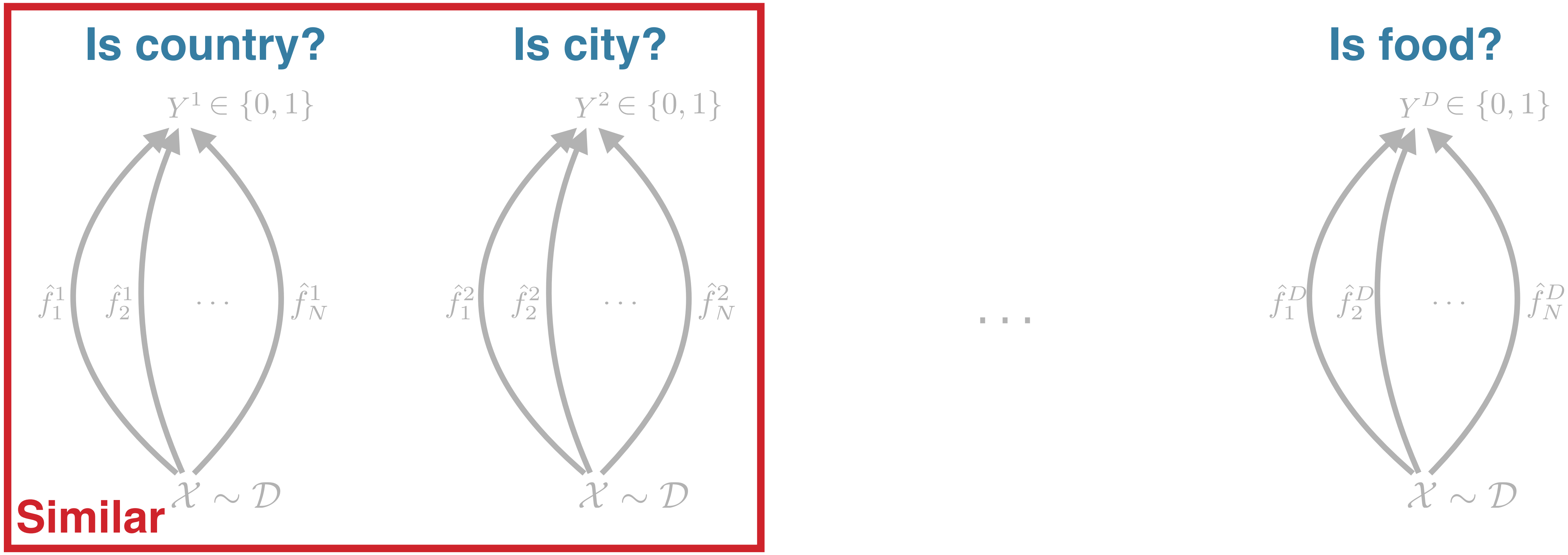


We have functions of the **same parametric form** using the same input data and features, answering **different questions**!

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

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

What About **Multiple Domains**?

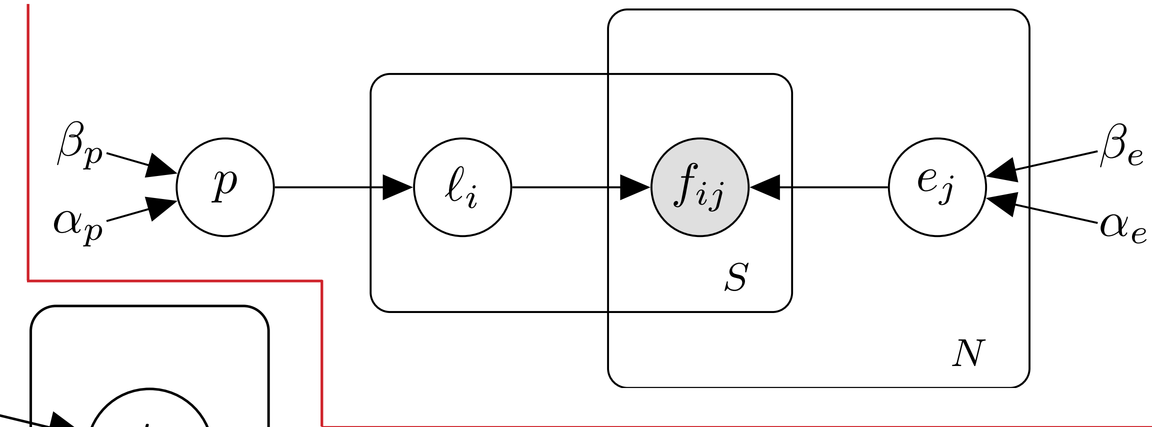


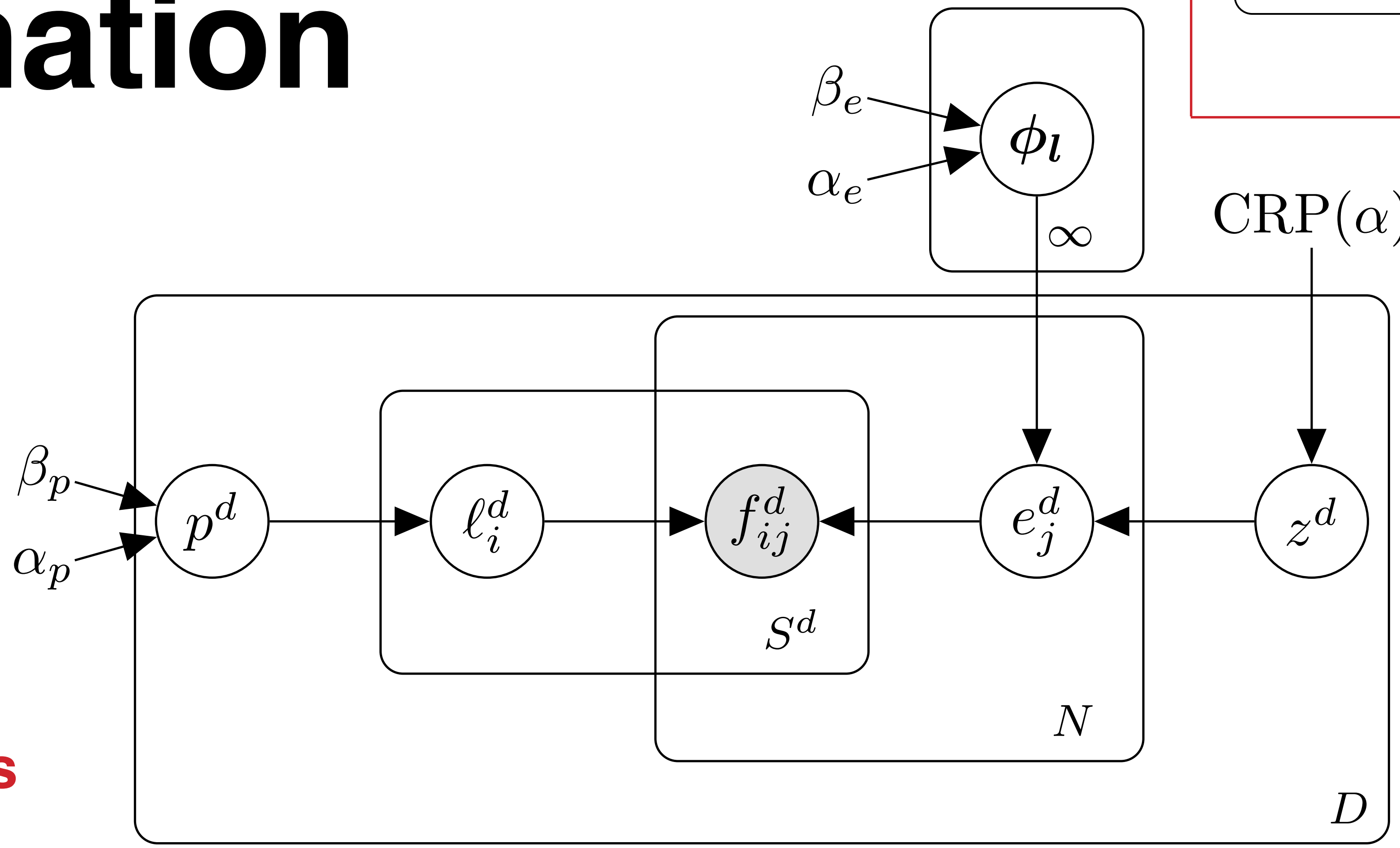
Coupled Error Estimation

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

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

Coupled Error Estimation

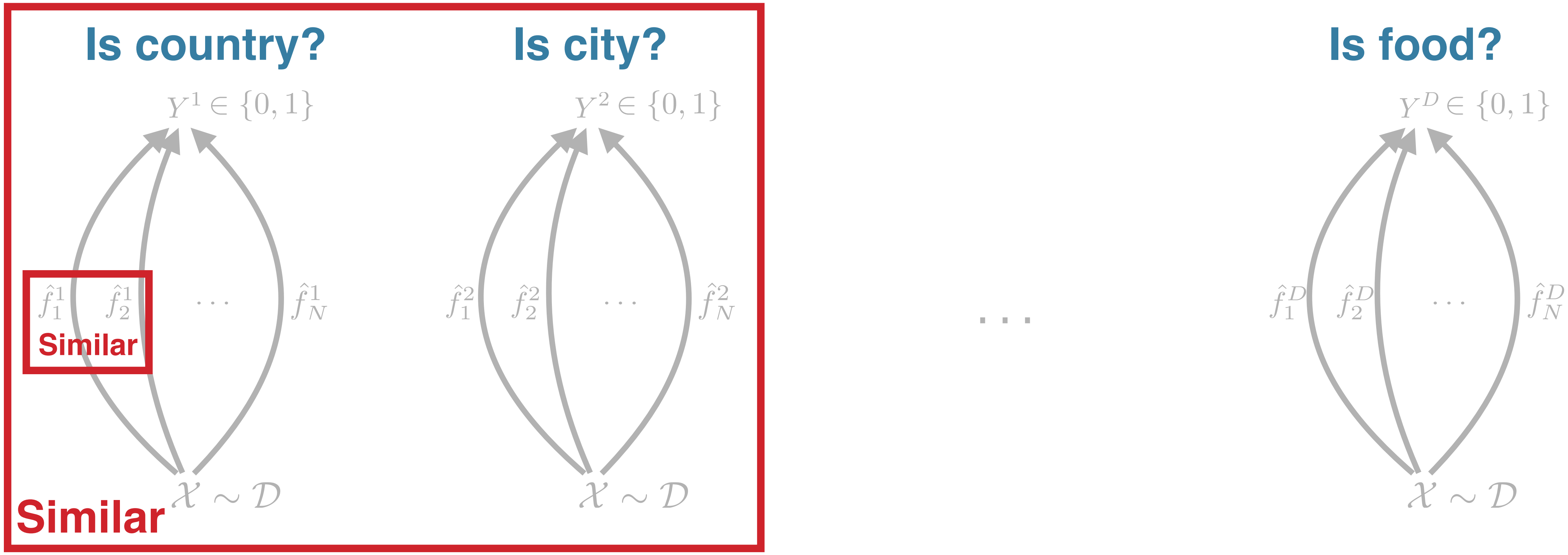




Dirichlet process clusters function error rates across domains

$$
\begin{aligned}
p^d &\sim \text{Beta}(\alpha_p, \beta_p), \;\text{ for } d = 1, \ldots, D, \\
\ell_i^d &\sim \text{Bernoulli}(p^d), \;\text{ for } i = 1, \ldots, S^d, \;\text{ and } d = 1, \ldots, D, \\
[\phi_l]_j &\sim \text{Beta}(\alpha_e, \beta_e), \;\text{ for } j = 1, \ldots, N, \;\text{ and } l = 1, \ldots, \infty, \\
z^d &\sim \text{CRP}(\alpha), \;\text{ for } d = 1, \ldots, D, \\
e_j^d &= [\phi_{z^d}]_j, \;\text{ for } j = 1, \ldots, N, \;\text{ and } d = 1, \ldots, D, \\
\hat{f}_{ij}^d &= \begin{cases} \ell_i^d & \text{, with probability } 1 - e_j^d, \\ 1 - \ell_i^d & \text{, otherwise.} \end{cases}
\end{aligned}
$$

What About **Multiple Domains**?

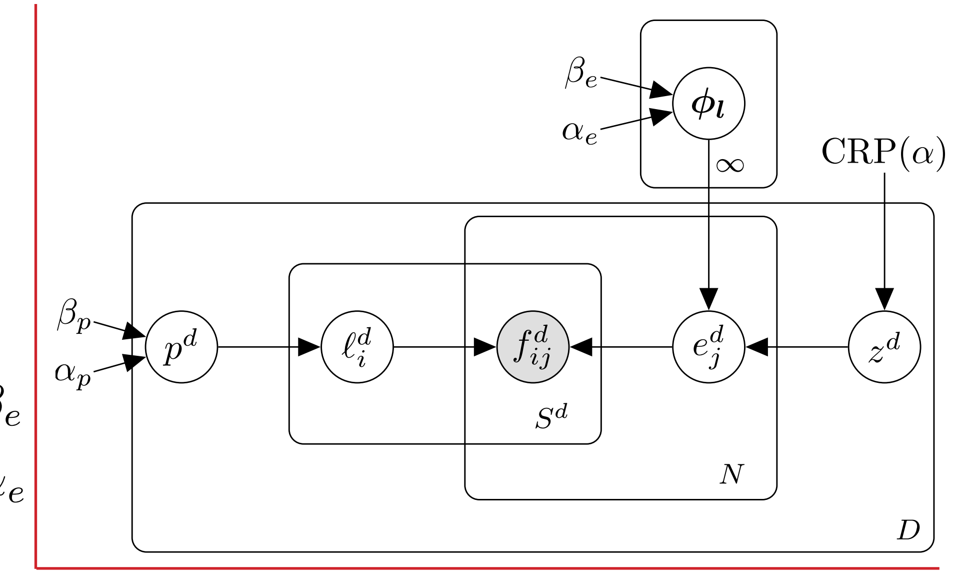


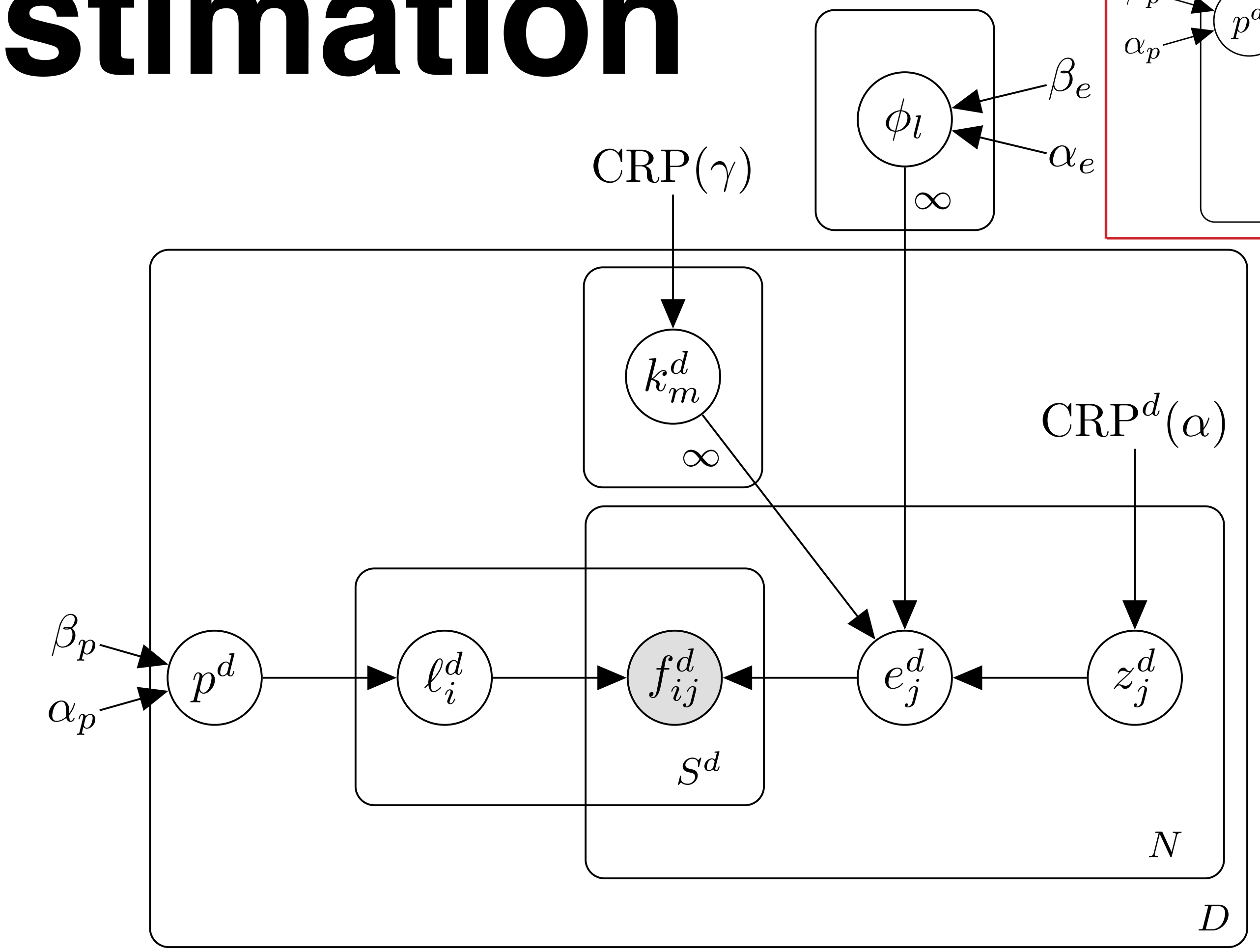
Hierarchical Coupled Error Estimation

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**.

Hierarchical Coupled Error Estimation





Hierarchical Dirichlet process further clusters function error rates across classifiers

$$
\begin{aligned}
p^d &\sim \text{Beta}(\alpha_p, \beta_p), \text{ for } d = 1, \dots, D, \\
\ell_i^d &\sim \text{Bernoulli}(p^d), \text{ for } i = 1, \dots, S^d, \text{ and } d = 1, \dots, D, \\
\phi_l &\sim \text{Beta}(\alpha_e, \beta_e), \text{ for } l = 1, \dots, \infty, \\
k_m^d &\sim \text{CRP}(\gamma), \text{ for } d = 1, \dots, D, \text{ and } m = 1, \dots, \infty, \\
z_j^d &\sim \text{CRP}^d(\alpha), \text{ for } d = 1, \dots, D, \text{ and } j = 1, \dots, N, \\
e_j^d &= \phi_{k^d_{z_j^d}}, \text{ for } j = 1, \dots, N, \text{ and } d = 1, \dots, D,
\end{aligned}
$$

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

Experiments

We report the **error mean squared deviation (MSE_{error})** between:

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

and the **target label mean absolute deviation (MAD_{label})**.

Our code and data are available at http://www.platanios.org/code

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)*

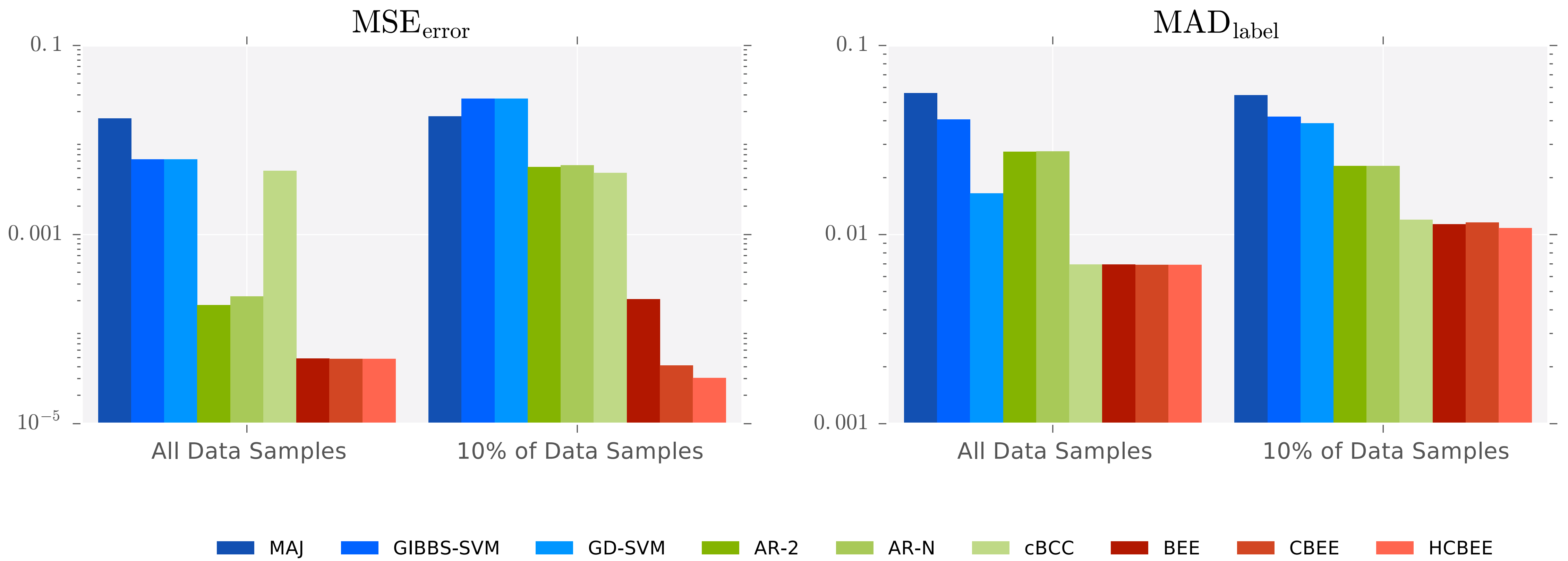
4 logistic regression classifiers using different features:

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

Domain	# 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



Experiments

2 Brain Data Set

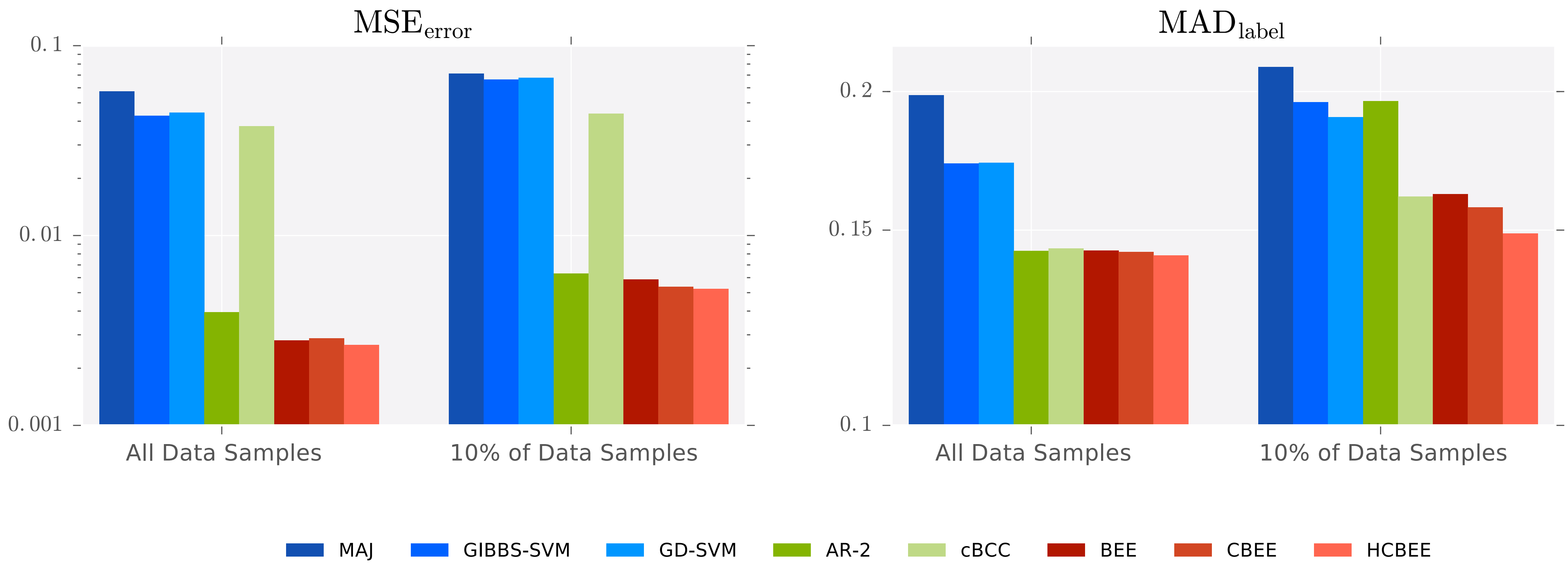
Task: *Find which of two 40 second long story passages corresponds to an unlabeled 40 second time series of fMRI neural activity*

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.

Experiments

2 Brain Data Set



Conclusion

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

3 Approaches presented

Highly accurate error rates estimates

↓

on two very different data sets

consistency

↓

correctness

Use **logical constraints** for error estimation

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

Thank You