

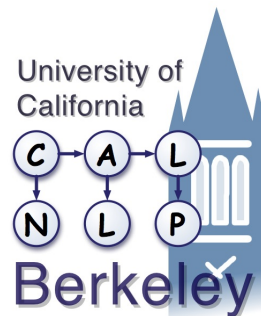
# Measurements: A Unifying Framework for Cost-Effective Learning

IBM SMiLe Workshop

October 8–9, 2009

Percy Liang

joint work with Michael Jordan and Dan Klein



# Supervised learning

Motivating example: information extraction from Craigslist ads

*y:* FEAT FEAT FEAT FEAT FEAT ...

*x:* View of Los Gatos Foothills ...

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

Available July 1 ... 2 bedroom 1 bath ...

# Supervised learning

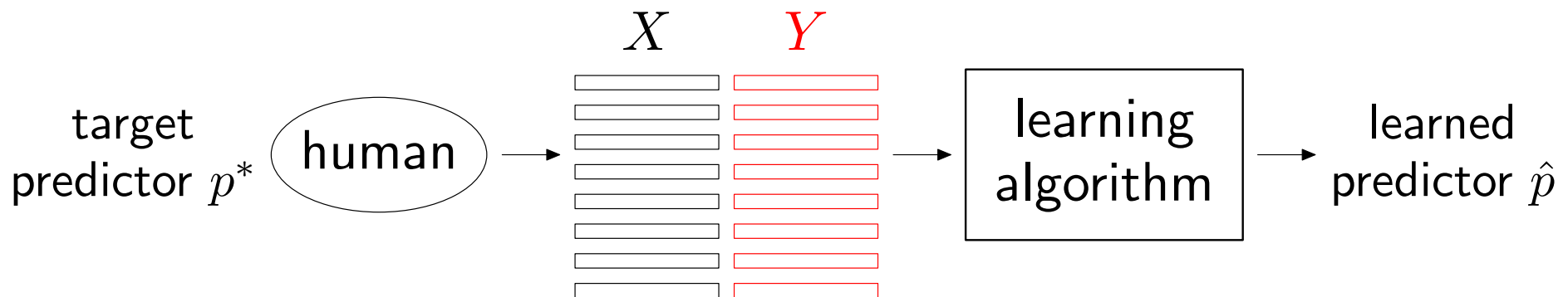
Motivating example: information extraction from Craigslist ads

$y$ : FEAT FEAT FEAT FEAT FEAT ...

$x$ : *View of Los Gatos Foothills ...*

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

*Available July 1 ... 2 bedroom 1 bath ...*



# Semi-supervised learning

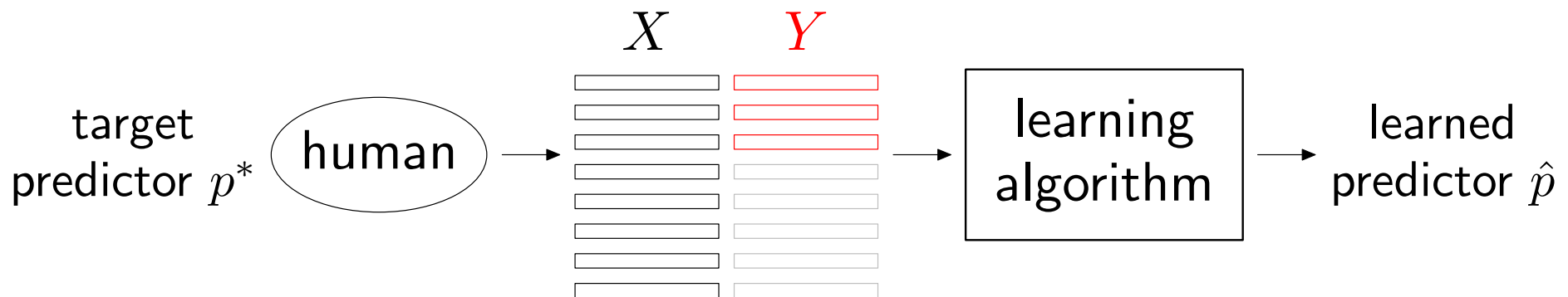
Motivating example: information extraction from Craigslist ads

$y$ : FEAT FEAT FEAT FEAT FEAT ...

$x$ : *View of Los Gatos Foothills ...*

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

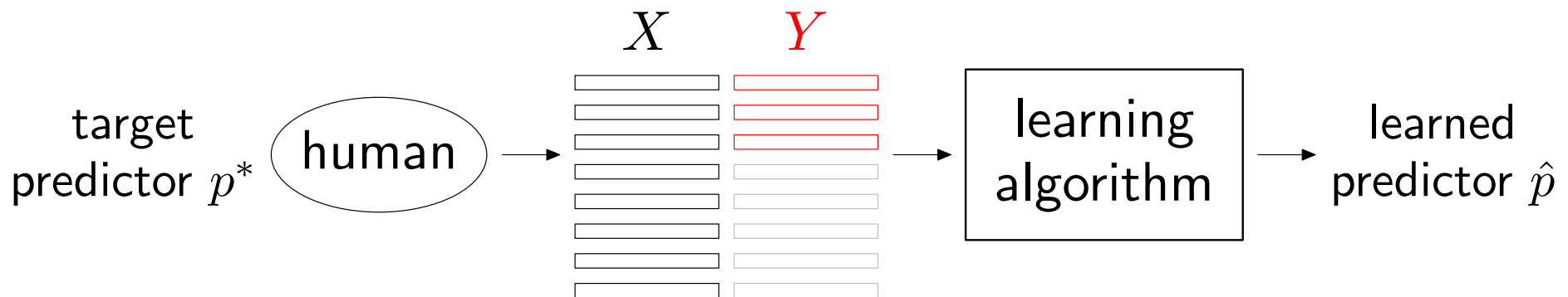
*Available July 1 ... 2 bedroom 1 bath ...*



# Semi-supervised learning

Motivating example: information extraction from Craigslist ads

$y$ :	FEAT	FEAT	FEAT	FEAT	FEAT	...			
$x$ :	View	of	Los	Gatos	Foothills	...			
	AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
	Available	July	1	...	2	bedroom	1	bath	...



Examples:

transductive SVMs, entropy regularization, etc.

# Learning from general constraints

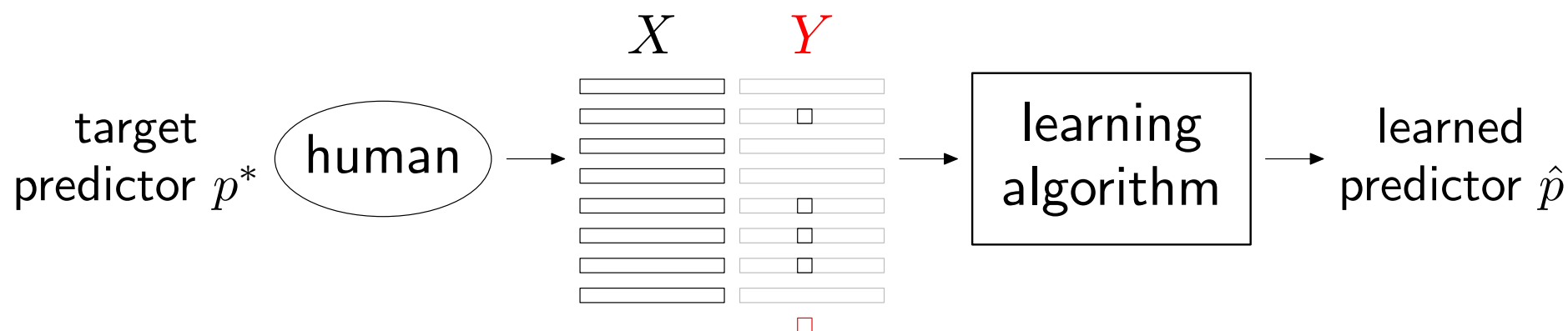
Motivating example: information extraction from Craigslist ads

$y$ : FEAT FEAT FEAT FEAT FEAT ...

$x$ : *View of Los Gatos Foothills ...*

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

*Available July 1 ... 2 bedroom 1 bath ...*



# Learning from general constraints

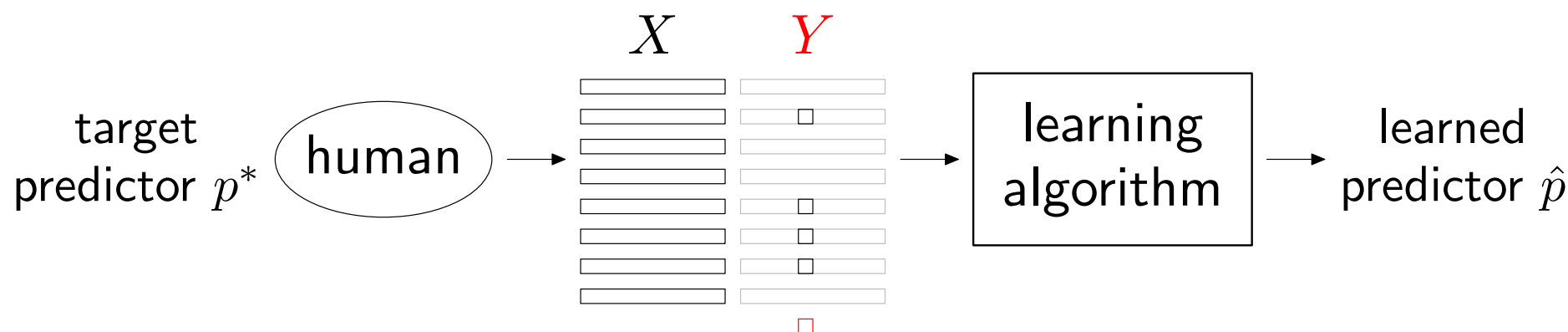
Motivating example: information extraction from Craigslist ads

*y*: FEAT FEAT FEAT FEAT FEAT ...

*x*: View of Los Gatos Foothills ...

AVAIL AVAIL AVAIL ... SIZE SIZE SIZE SIZE ...

Available July 1 ... 2 bedroom 1 bath ...



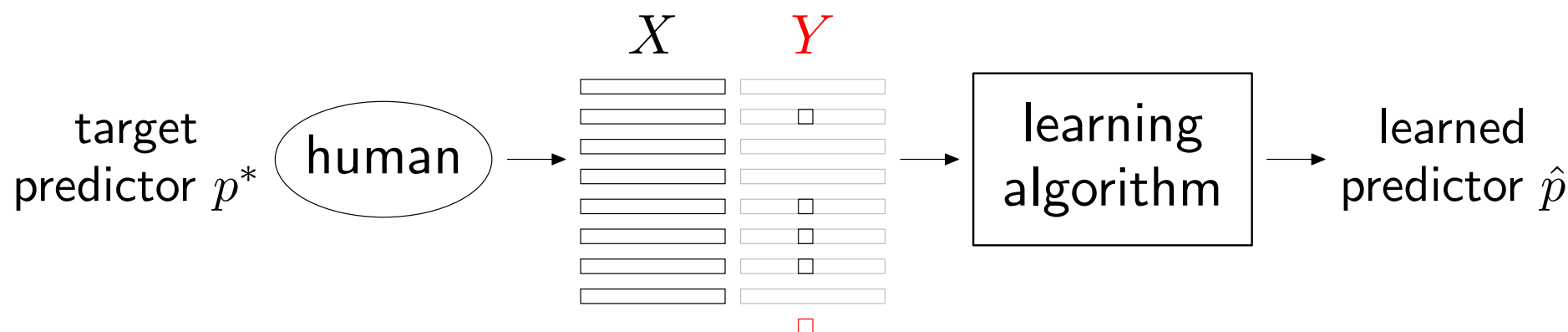
General constraints are declarative:

Example: word “bedroom” is tagged as **SIZE** 75% of the time

# Learning from general constraints

Motivating example: information extraction from Craigslist ads

<i>y</i> :	FEAT	FEAT	FEAT	FEAT	FEAT	...			
<i>x</i> :	View	of	Los	Gatos	Foothills	...			
	AVAIL	AVAIL	AVAIL	...	SIZE	SIZE	SIZE	SIZE	...
	Available	July	1	...	2	bedroom	1	bath	...



General constraints are declarative:

Example: word “bedroom” is tagged as **SIZE** 75% of the time  
[Chang, et al., 2007; Druck, et al., 2008; Graça, et al., 2008]



# The general picture



# The general picture



Many types of information:

- Labels (very specific)

- Partial labels

- Constraints

- Preferences (very general)

# The general picture



Many types of **information**:

- Labels (very specific)

- Partial labels

- Constraints

- Preferences (very general)

Questions:

1. How to incorporate diverse information?

# The general picture



Many types of **information**:

- Labels (very specific)

- Partial labels

- Constraints

- Preferences (very general)

Questions:

1. How to incorporate diverse information?
2. What is the most cost-effective one to acquire?

# The general picture



Many types of **information**:

- Labels (very specific)
- Partial labels
- Constraints
- Preferences (very general)

Questions:

1. How to incorporate diverse information?
2. What is the most cost-effective one to acquire?

**Measurements:** unifying framework for specifying information about the desired predictor  $p^*$

# The general picture



Many types of **information**:

Labels (very specific)	}	<b>measurements</b>
Partial labels		
Constraints		
Preferences (very general)		

Questions:

1. How to incorporate diverse information?
2. What is the most cost-effective one to acquire?

**Measurements**: unifying framework for specifying information about the desired predictor  $p^*$

# Measurements: definition/notation

$X_1$  ,  $Y_1$

$X_2$  ,  $Y_2$

$X_3$  ,  $Y_3$

... ..

$X_i$  ,  $Y_i$

... ..

$X_n$  ,  $Y_n$

# Measurements: definition/notation

Measurement features:  $\sigma(x, y) \in \mathbb{R}^k$

$$\sigma( X_1 , Y_1 )$$

$$\sigma( X_2 , Y_2 )$$

$$\sigma( X_3 , Y_3 )$$

$$\begin{matrix} \dots & \dots \\ \sigma( X_i , Y_i ) \end{matrix}$$

$$\begin{matrix} \dots & \dots \\ \sigma( X_n , Y_n ) \end{matrix}$$



# Measurements: definition/notation

Measurement features:  $\sigma(x, y) \in \mathbb{R}^k$

Measurement values:  $\tau \in \mathbb{R}^k$

$$\begin{array}{l} \sigma( X_1 , Y_1 ) \\ \sigma( X_2 , Y_2 ) \\ \sigma( X_3 , Y_3 ) \\ \dots \quad \dots \\ \sigma( X_i , Y_i ) \\ \dots \quad \dots \\ \sigma( X_n , Y_n ) \\ + \text{ noise} \\ \hline \tau \end{array}$$

$$\tau = \sum_{i=1}^n \sigma(X_i, Y_i) + \text{noise}$$

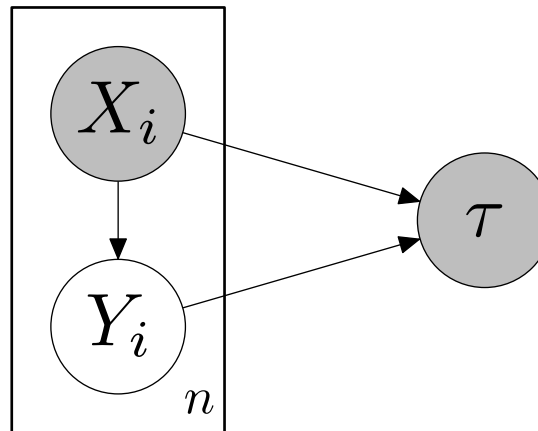
# Measurements: definition/notation

Measurement features:  $\sigma(x, y) \in \mathbb{R}^k$

Measurement values:  $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^n \sigma(X_i, Y_i) + \text{noise}$$

$$\begin{array}{c} \sigma(X_1, Y_1) \\ \sigma(X_2, Y_2) \\ \sigma(X_3, Y_3) \\ \dots \\ \sigma(X_i, Y_i) \\ \dots \\ \sigma(X_n, Y_n) \\ + \text{noise} \\ \hline \tau \end{array}$$

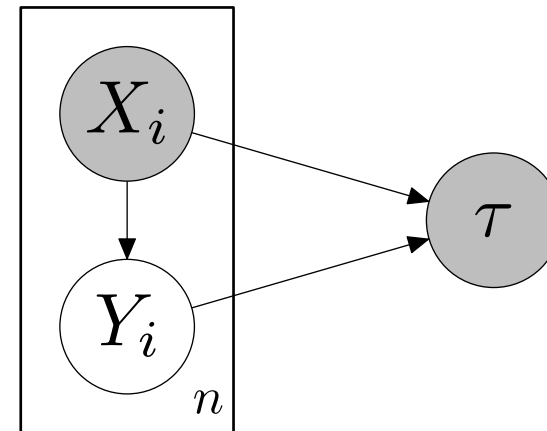


# Measurements: definition/notation

Measurement features:  $\sigma(x, y) \in \mathbb{R}^k$

Measurement values:  $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^n \sigma(X_i, Y_i) + \text{noise}$$



$$\begin{array}{l} \sigma(X_1, Y_1) \\ \sigma(X_2, Y_2) \\ \sigma(X_3, Y_3) \\ \dots \\ \sigma(X_i, Y_i) \\ \dots \\ \sigma(X_n, Y_n) \\ + \text{noise} \\ \hline \tau \end{array}$$

How it works:

Human sets  $\sigma$  to reveal desired information about  $Y$

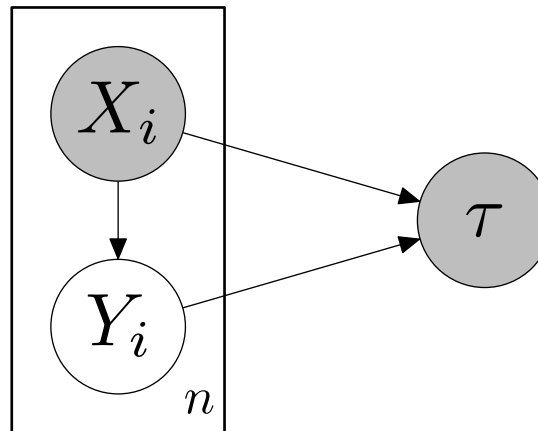
# Measurements: definition/notation

Measurement features:  $\sigma(x, y) \in \mathbb{R}^k$

Measurement values:  $\tau \in \mathbb{R}^k$

$$\tau = \sum_{i=1}^n \sigma(X_i, Y_i) + \text{noise}$$

$$\begin{array}{l} \sigma(X_1, Y_1) \\ \sigma(X_2, Y_2) \\ \sigma(X_3, Y_3) \\ \dots \\ \sigma(X_i, Y_i) \\ \dots \\ \sigma(X_n, Y_n) \\ + \text{noise} \\ \hline \tau \end{array}$$



How it works:

Human sets  $\sigma$  to reveal desired information about  $Y$

Human “measures” and observes  $\tau$

# Examples of measurements

# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{AVAIL FEAT ...}] \Rightarrow \tau = 0$$

# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{AVAIL FEAT ...}] \Rightarrow \tau = 0$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT AVAIL ...}] \Rightarrow \tau = 0$$

...



# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{AVAIL FEAT ...}] \Rightarrow \tau = 0$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT AVAIL ...}] \Rightarrow \tau = 0$$

...

Labeled predicate:

$$\sigma_j(x, y) = \sum_{i=1}^{\text{len}(x)} \mathbb{I}[x_i = \textit{View}, y_i = \text{FEAT}] \Rightarrow \tau = 17$$

# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{AVAIL FEAT ...}] \Rightarrow \tau = 0$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT AVAIL ...}] \Rightarrow \tau = 0$$

...

Labeled predicate:

$$\sigma_j(x, y) = \sum_{i=1}^{\text{len}(x)} \mathbb{I}[x_i = \textit{View}, y_i = \text{FEAT}] \Rightarrow \tau = 17$$

$$\sigma_j(x, y) = \sum_{i=1}^{\text{len}(x)} \mathbb{I}[x_i = \textit{View}, y_i = \text{AVAIL}] \Rightarrow \tau = 2$$

...

# Examples of measurements

Fully-labeled example:

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT FEAT ...}] \Rightarrow \tau = 1$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{AVAIL FEAT ...}] \Rightarrow \tau = 0$$

$$\sigma_j(x, y) = \mathbb{I}[x = \textit{View of ...}, y = \text{FEAT AVAIL ...}] \Rightarrow \tau = 0$$

...

Labeled predicate:

$$\sigma_j(x, y) = \sum_{i=1}^{\text{len}(x)} \mathbb{I}[x_i = \textit{View}, y_i = \text{FEAT}] \Rightarrow \tau = 17$$

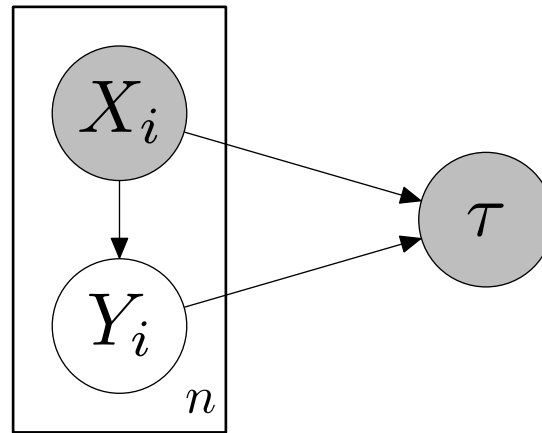
$$\sigma_j(x, y) = \sum_{i=1}^{\text{len}(x)} \mathbb{I}[x_i = \textit{View}, y_i = \text{AVAIL}] \Rightarrow \tau = 2$$

...

**Note:** To get a measurement value  $\tau$ ,  
need to look at only small subset of examples

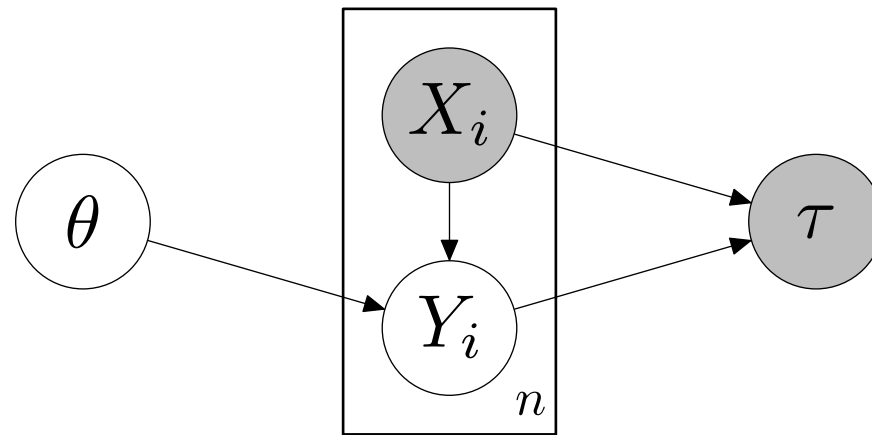
# Prediction model

Bayesian framework:



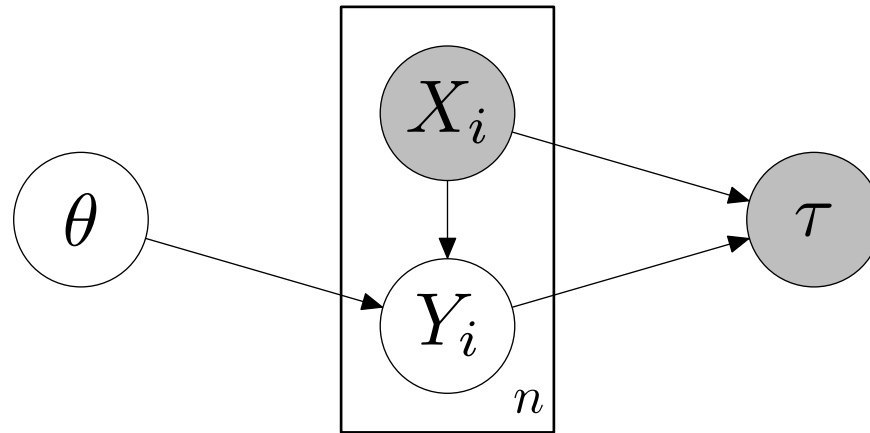
# Prediction model

Bayesian framework:



# Prediction model

Bayesian framework:

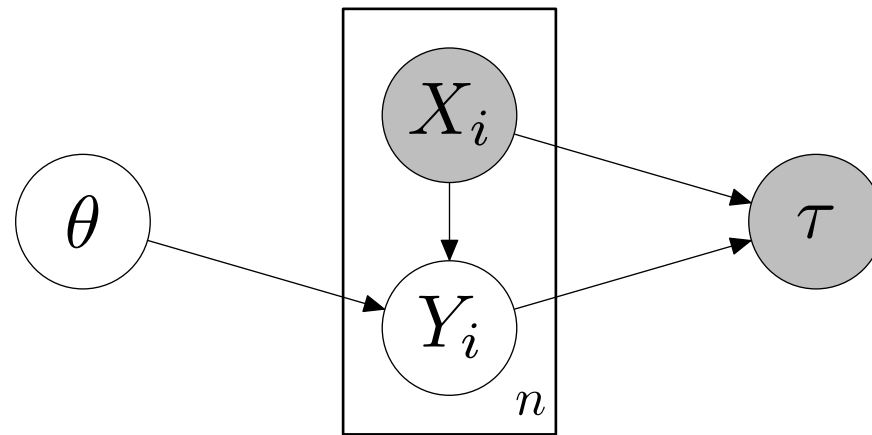


Exponential families:

$$p_{\theta}(y \mid x) \propto \exp\{\langle \phi(x, y), \theta \rangle\}$$

# Prediction model

Bayesian framework:



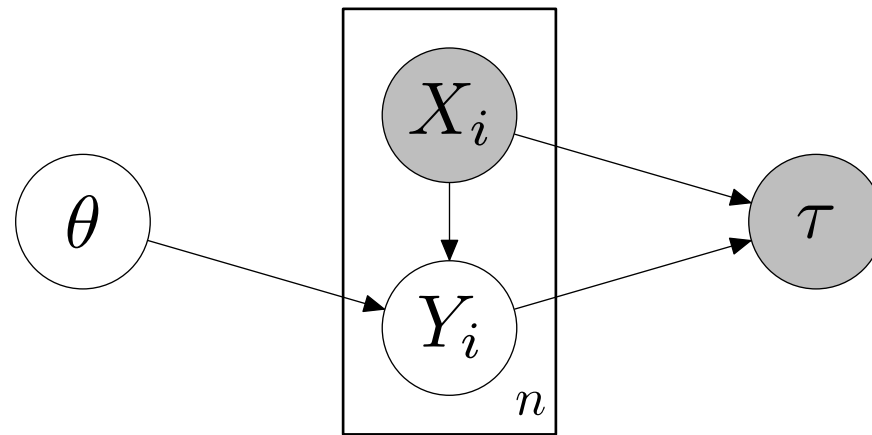
Exponential families:

$$p_{\theta}(y \mid x) \propto \exp\{\langle \phi(x, y), \theta \rangle\}$$

$\phi(x, y) \in \mathbb{R}^d$ : model features (not measurement features)

# Prediction model

Bayesian framework:



Exponential families:

$$p_{\theta}(y \mid x) \propto \exp\{\langle \phi(x, y), \theta \rangle\}$$

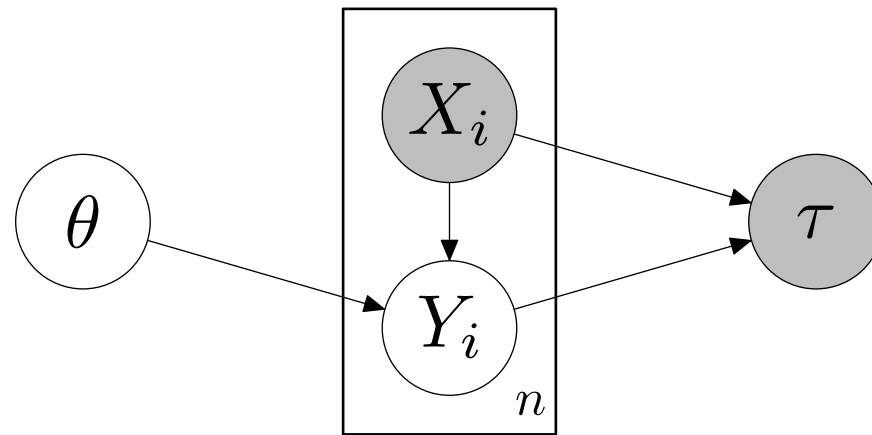
$\phi(x, y) \in \mathbb{R}^d$ : model features (not measurement features)

$\theta \in \mathbb{R}^d$ : model parameters



# Prediction model

Bayesian framework:



Exponential families:

$$p_{\theta}(y \mid x) \propto \exp\{\langle \phi(x, y), \theta \rangle\}$$

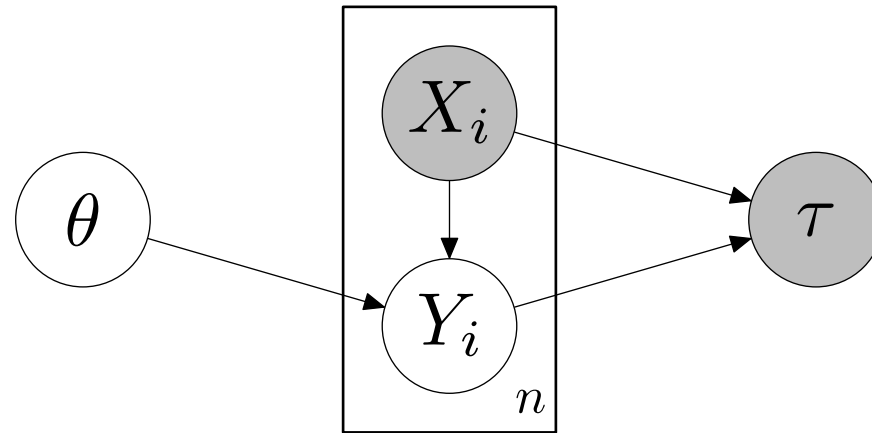
$\phi(x, y) \in \mathbb{R}^d$ : model features (not measurement features)

$\theta \in \mathbb{R}^d$ : model parameters

Examples: logistic regression, conditional random fields

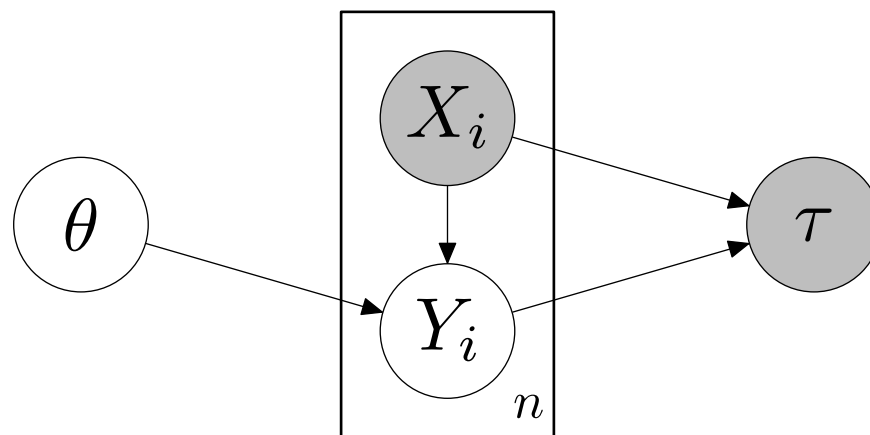
# Learning via Bayesian inference

Bayesian framework:



# Learning via Bayesian inference

Bayesian framework:

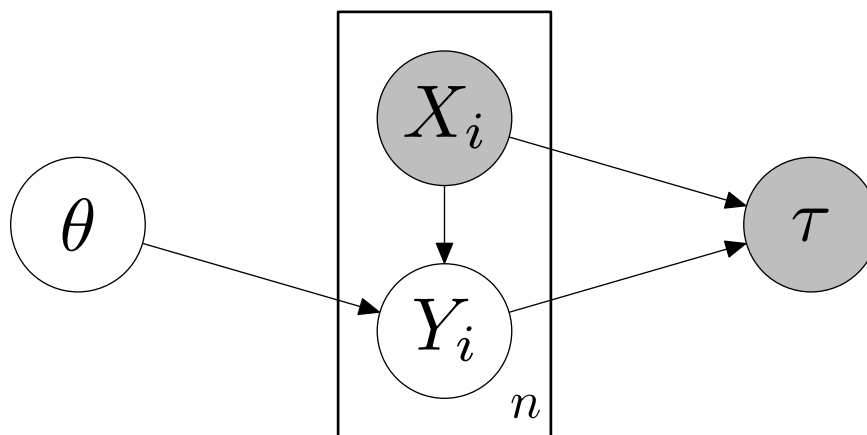


Goal:

Compute  $p(\theta, Y_1, \dots, Y_n \mid \tau, X_1, \dots, X_n)$

# Learning via Bayesian inference

Bayesian framework:



Goal:

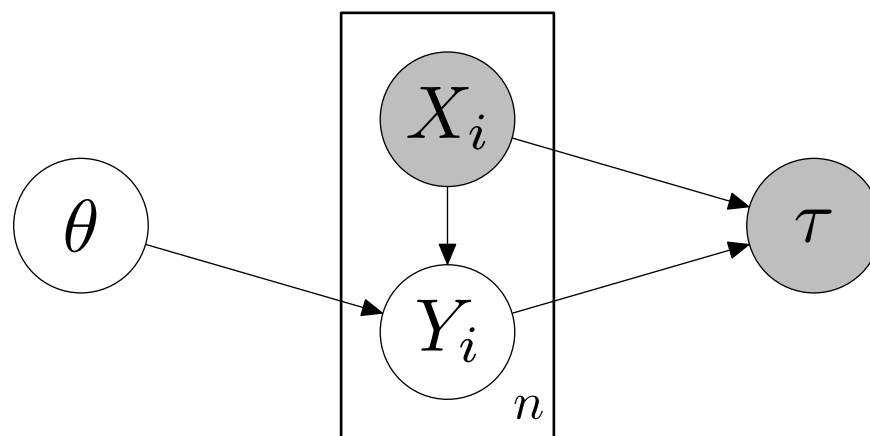
Compute  $p(\theta, Y_1, \dots, Y_n \mid \tau, X_1, \dots, X_n)$

Solution:

Based on variational approximation

# Learning via Bayesian inference

Bayesian framework:



Goal:

Compute  $p(\theta, Y_1, \dots, Y_n \mid \tau, X_1, \dots, X_n)$

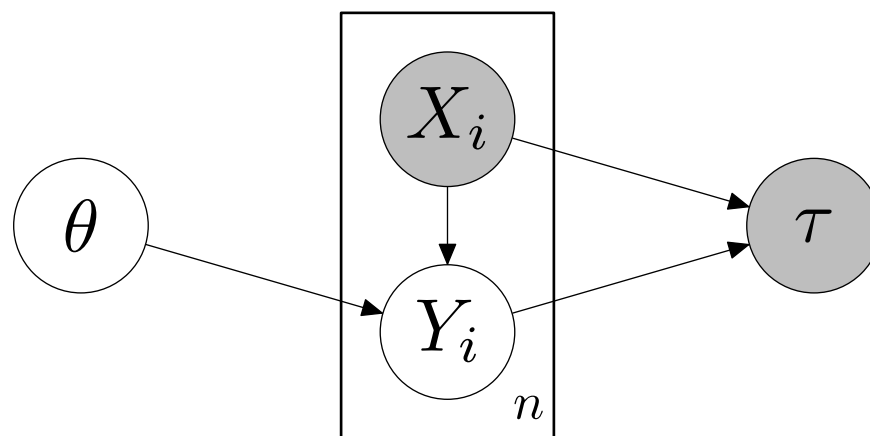
Solution:

Based on variational approximation

Algorithm just requires computing model expectations

# Learning via Bayesian inference

Bayesian framework:



Goal:

Compute  $p(\theta, Y_1, \dots, Y_n \mid \tau, X_1, \dots, X_n)$

Solution:

Based on variational approximation

Algorithm just requires computing model expectations

Output:  $q_{\beta}(y \mid x)$  and  $p_{\theta}(y \mid x)$

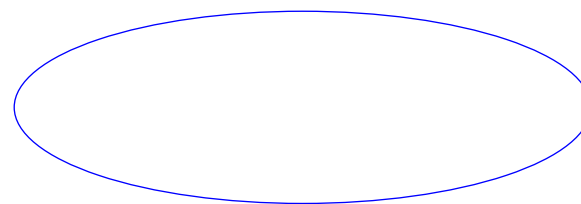
# Information geometry viewpoint

(assume zero measurement noise)

# Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{P} \stackrel{\text{def}}{=} \{p_{\theta}(y \mid x) : \theta \in \mathbb{R}^d\}$$



$\mathcal{P}$ : all predictors we want to consider



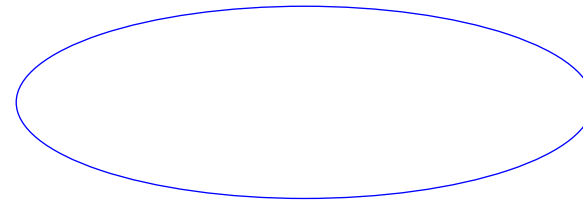
# Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\}$$



$$\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



$\mathcal{P}$ : all predictors we want to consider

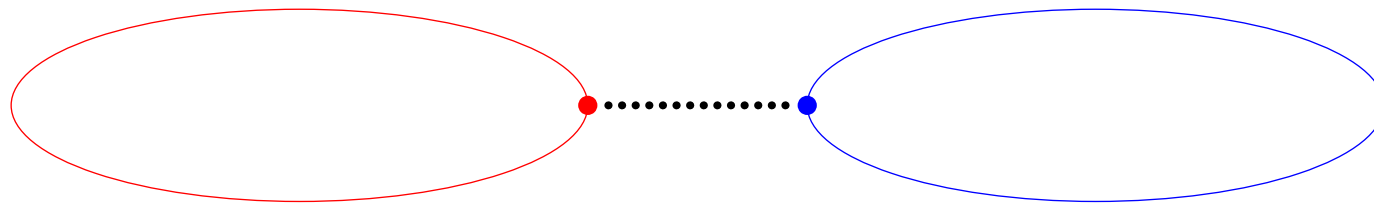
$\mathcal{Q}$ : all predictors consistent with our measurements

# Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\}$$

$$\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \text{KL} (q \parallel p)$$

$\mathcal{P}$ : all predictors we want to consider

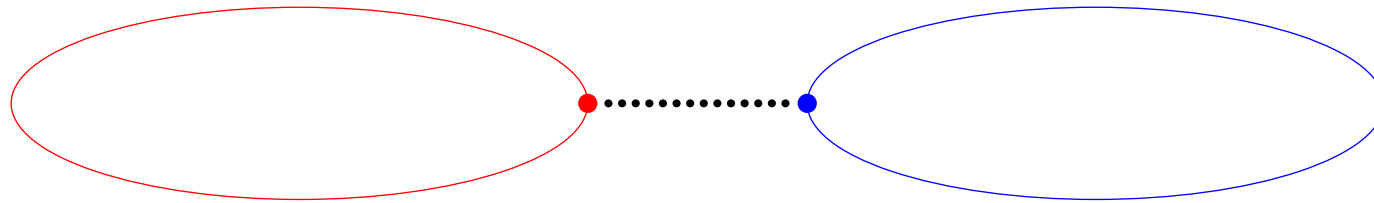
$\mathcal{Q}$ : all predictors consistent with our measurements

# Information geometry viewpoint

(assume zero measurement noise)

$$\mathcal{Q} \stackrel{\text{def}}{=} \{q(y \mid x) : \mathbb{E}_q[\sigma] = \tau\}$$

$$\mathcal{P} \stackrel{\text{def}}{=} \{p_\theta(y \mid x) : \theta \in \mathbb{R}^d\}$$



$$\min_{q \in \mathcal{Q}, p \in \mathcal{P}} \text{KL}(\textcolor{red}{q} \parallel \textcolor{blue}{p})$$

$\mathcal{P}$ : all predictors we want to consider

$\mathcal{Q}$ : all predictors consistent with our measurements

Interpretation:

Measurements shape  $\mathcal{Q}$       Find model in  $\mathcal{P}$  with best fit

# Results on the Craigslist task

$n = 1000$  total examples (ads), 11 possible labels

Model:

Conditional random field with standard NLP features

# Results on the Craigslist task

$n = 1000$  total examples (ads), 11 possible labels

## Model:

Conditional random field with standard NLP features

## Measurements:

- Fully-labeled examples
- 33 labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = \textit{View}, y_i = \textit{FEAT}]$ )

# Results on the Craigslist task

$n = 1000$  total examples (ads), 11 possible labels

## Model:

Conditional random field with standard NLP features

## Measurements:

- Fully-labeled examples
- 33 labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = \textit{View}, y_i = \textit{FEAT}]$ )

## Per-position test accuracy (on 100 examples):

# labeled examples	10	25	100
General Expectation Criteria	74.6	77.2	80.5
Constraint-Driven Learning	<b>74.7</b>	<b>78.5</b>	81.7
Measurements	71.4	76.5	<b>82.5</b>

# Results on the Craigslist task

$n = 1000$  total examples (ads), 11 possible labels

## Model:

Conditional random field with standard NLP features

## Measurements:

- Fully-labeled examples
- 33 labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = \textit{View}, y_i = \textit{FEAT}]$ )

## Per-position test accuracy (on 100 examples):

# labeled examples	10	25	100
General Expectation Criteria	74.6	77.2	80.5
Constraint-Driven Learning	<b>74.7</b>	<b>78.5</b>	81.7
Measurements	71.4	76.5	<b>82.5</b>

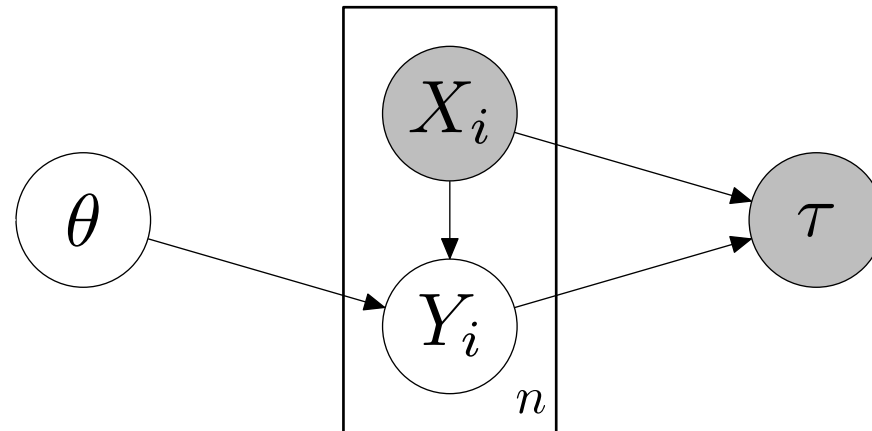
Able to integrate labeled examples and predicates gracefully

So far: given measurements, how to learn

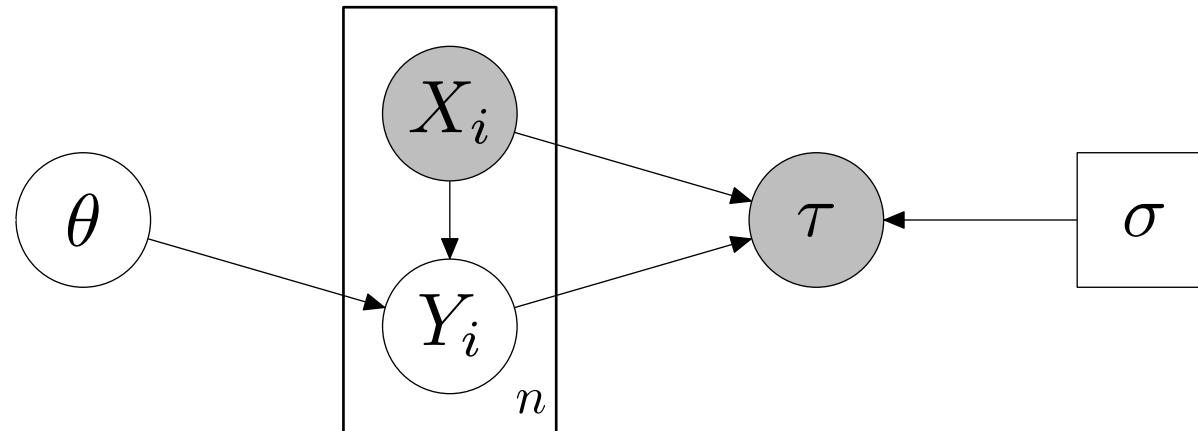
Next: how to choose measurements?



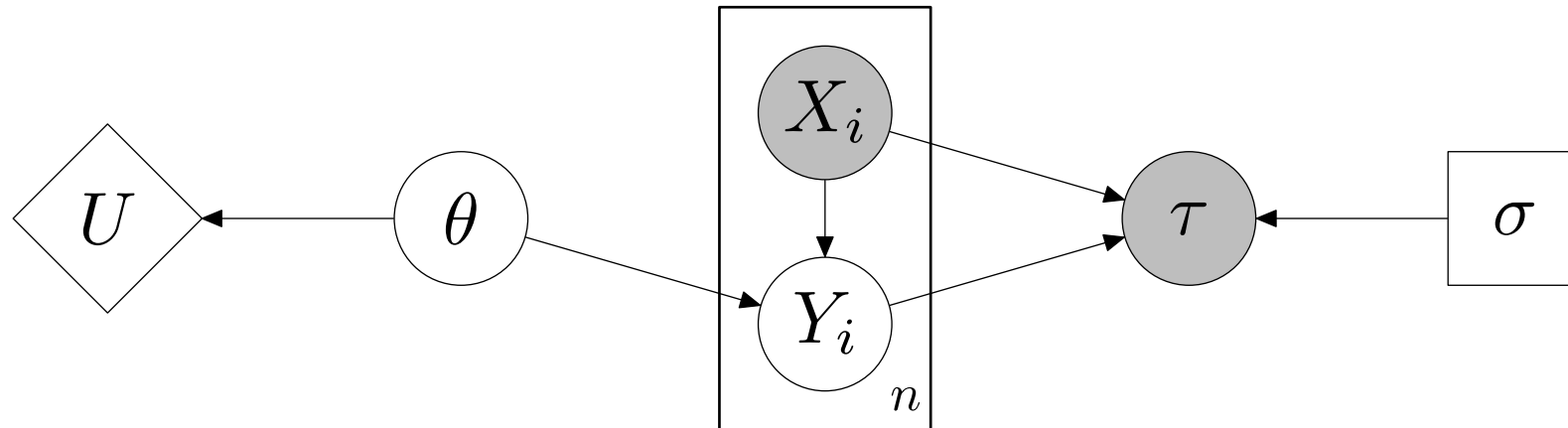
# Bayesian experimental design (active learning)



# Bayesian experimental design (active learning)



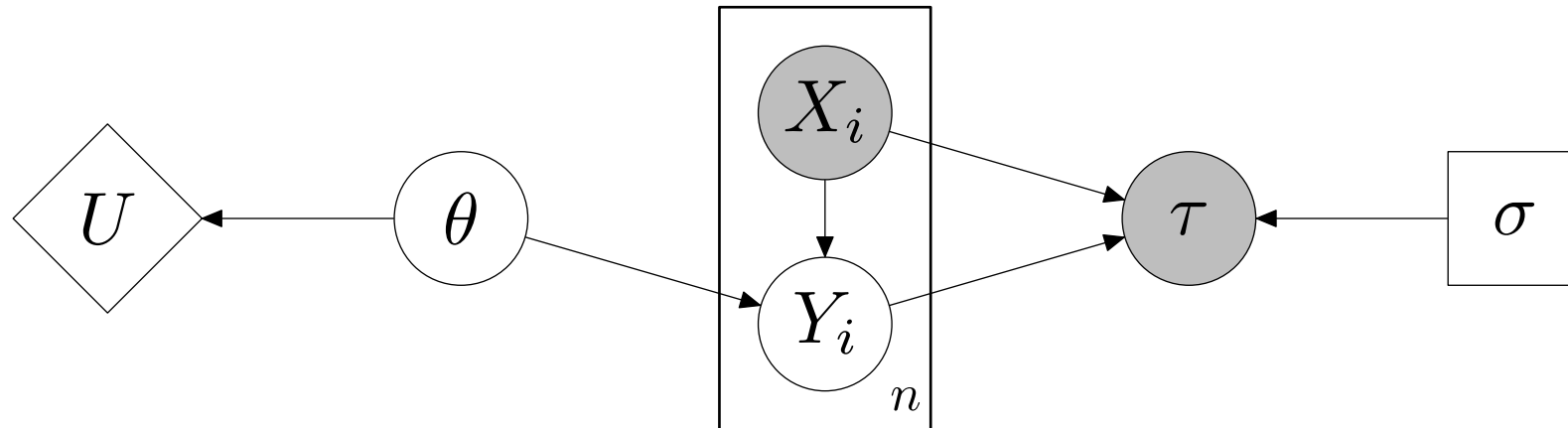
# Bayesian experimental design (active learning)



Utility of measurement  $(\sigma, \tau)$ :

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

# Bayesian experimental design (active learning)



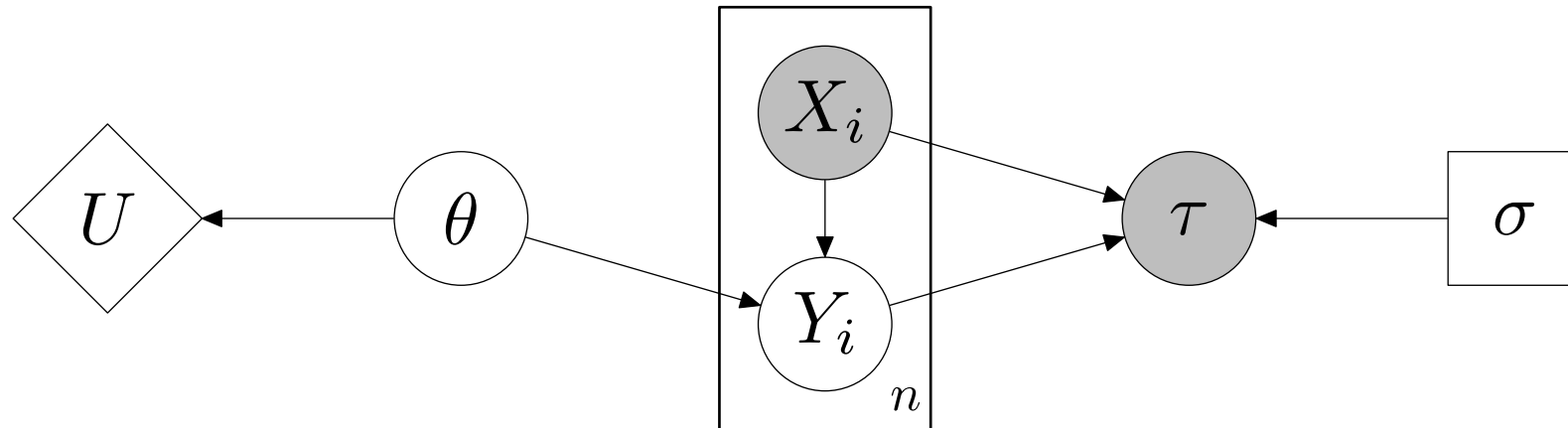
Utility of measurement  $(\sigma, \tau)$ :

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

Reward:

Prediction loss (e.g., Hamming) on a heldout set

# Bayesian experimental design (active learning)



Utility of measurement  $(\sigma, \tau)$ :

$$U(\sigma, \tau) = \underbrace{R(\sigma, \tau)}_{\text{reward}} - \underbrace{C(\sigma)}_{\text{cost}}$$

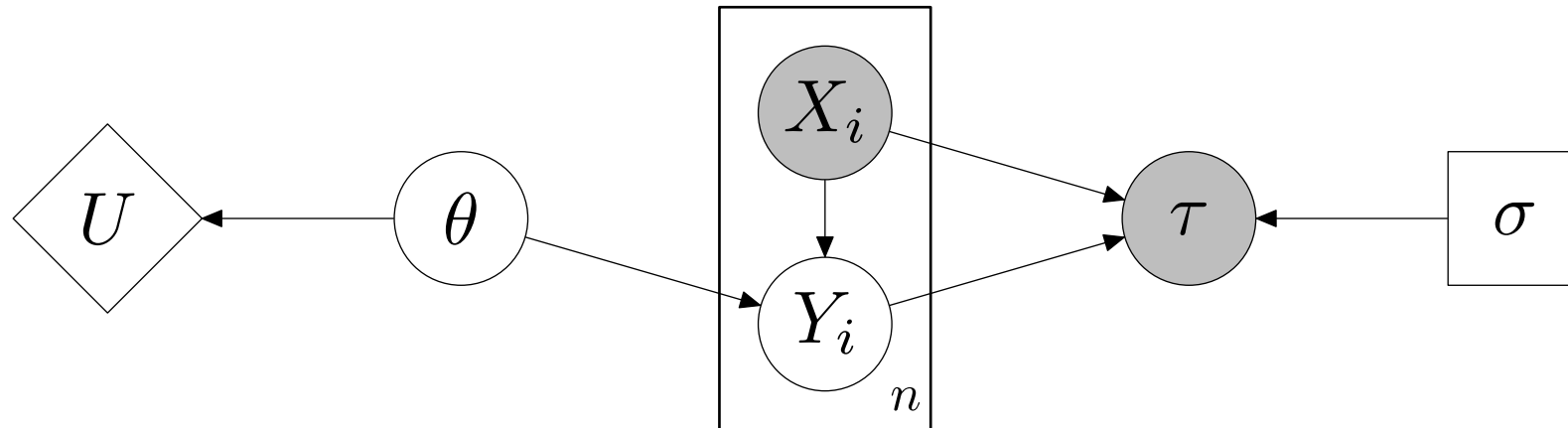
Reward:

Prediction loss (e.g., Hamming) on a heldout set

Cost:

Estimated time for human to produce measurement value

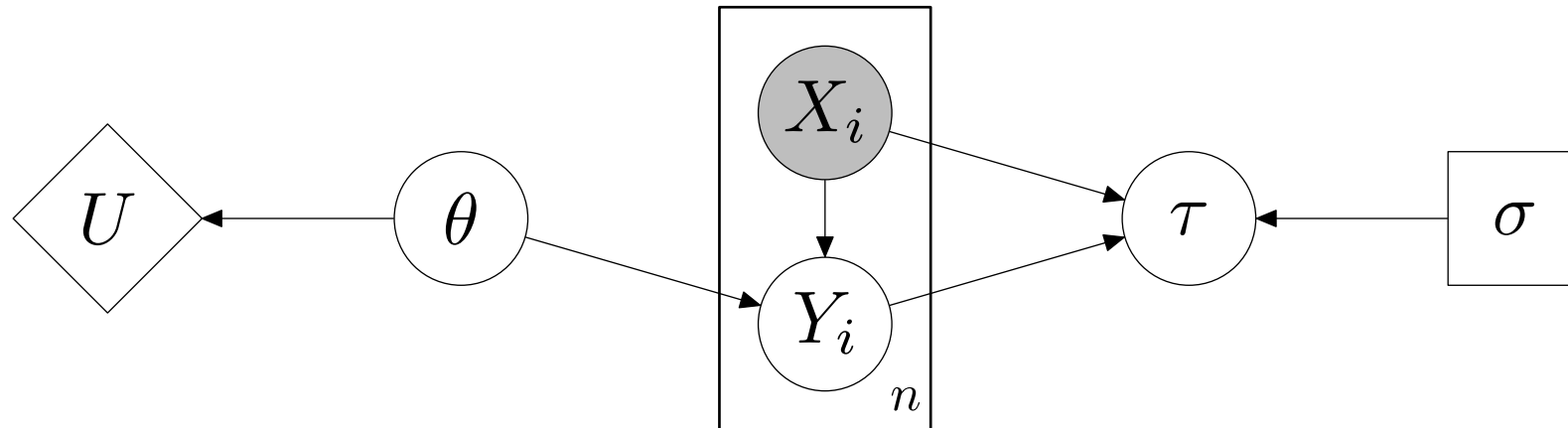
# Bayesian experimental design (active learning)



Don't know  $\tau$ , so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

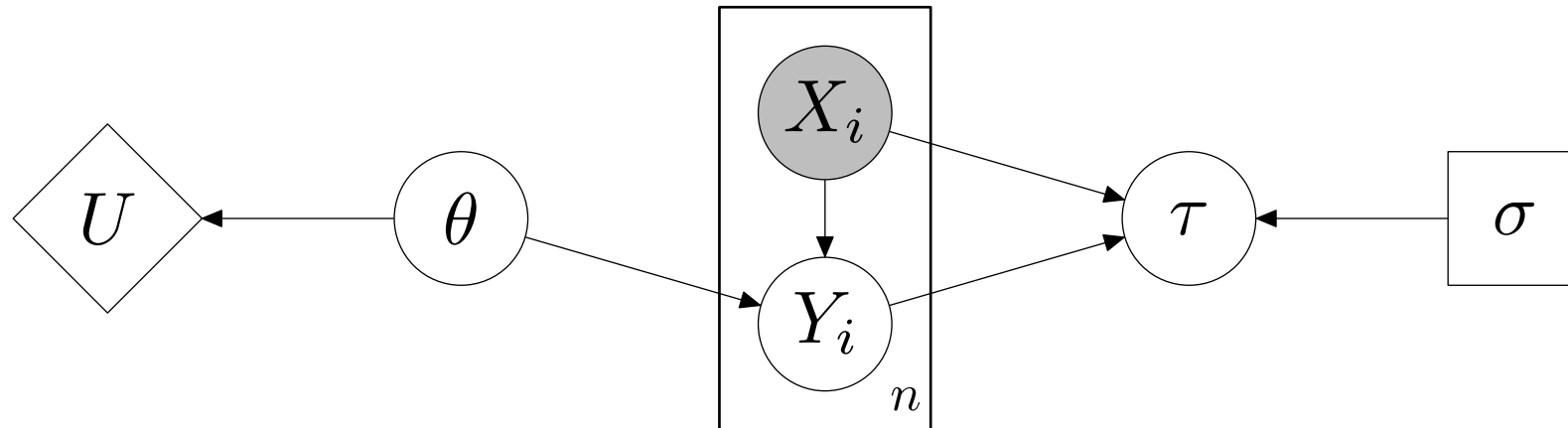
# Bayesian experimental design (active learning)



Don't know  $\tau$ , so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

# Bayesian experimental design (active learning)



Don't know  $\tau$ , so integrate out:

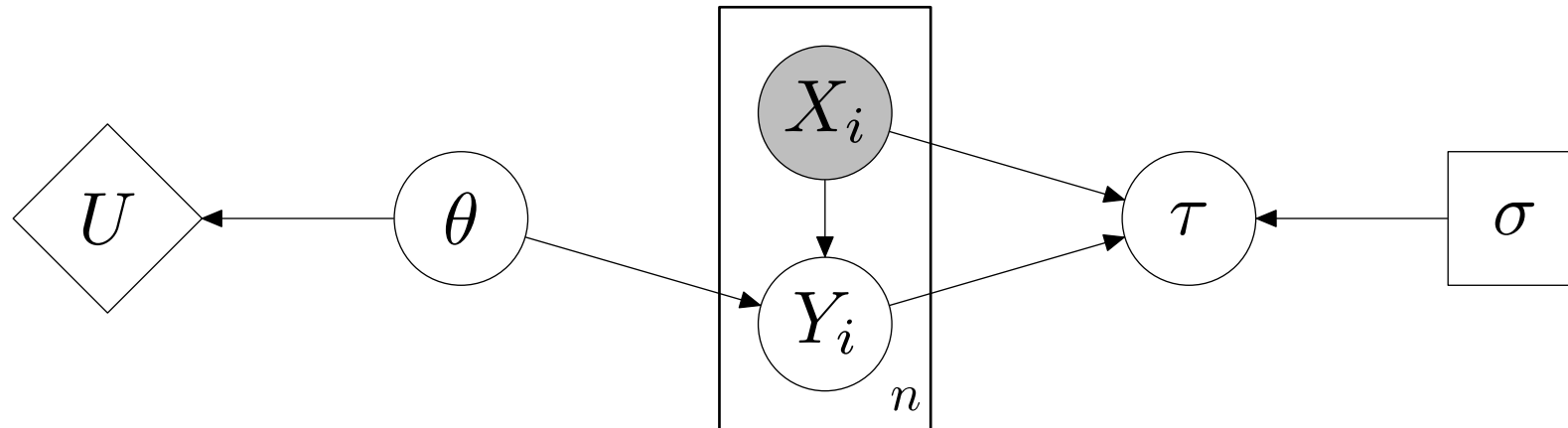
$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

Choose best measurement feature  $\sigma$  from candidate set  $\Sigma$ :

$$\sigma^* = \operatorname{argmax}_{\sigma \in \Sigma} U(\sigma)$$



# Bayesian experimental design (active learning)



Don't know  $\tau$ , so integrate out:

$$U(\sigma) = E_{p(\tau|X)}[U(\sigma, \tau)]$$

Choose best measurement feature  $\sigma$  from candidate set  $\Sigma$ :

$$\sigma^* = \operatorname{argmax}_{\sigma \in \Sigma} U(\sigma)$$

In general, iterate to add more measurements

# Part-of-speech tagging results

$n = 1000$  total examples (sentences), 45 possible labels

**Model:** Indep. logistic regression with standard NLP features

# Part-of-speech tagging results

$n = 1000$  total examples (sentences), 45 possible labels

**Model:** Indep. logistic regression with standard NLP features

**Measurements:**

- Fully-labeled examples
- Labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = \textit{the}, y_i = \text{DT}]$ )

Use label entropy as surrogate for assessing measurements

# Part-of-speech tagging results

$n = 1000$  total examples (sentences), 45 possible labels

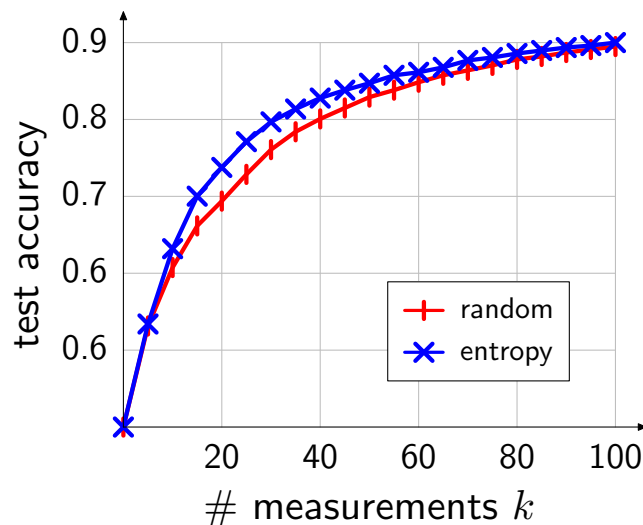
**Model:** Indep. logistic regression with standard NLP features

**Measurements:**

- Fully-labeled examples
- Labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = \textit{the}, y_i = \text{DT}]$ )

Use label entropy as surrogate for assessing measurements

**Test accuracy (on 100 examples):**



(a) Labeling examples

# Part-of-speech tagging results

$n = 1000$  total examples (sentences), 45 possible labels

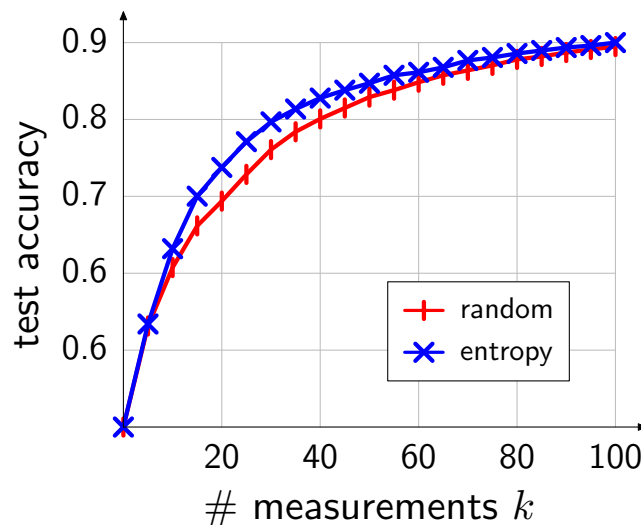
**Model:** Indep. logistic regression with standard NLP features

**Measurements:**

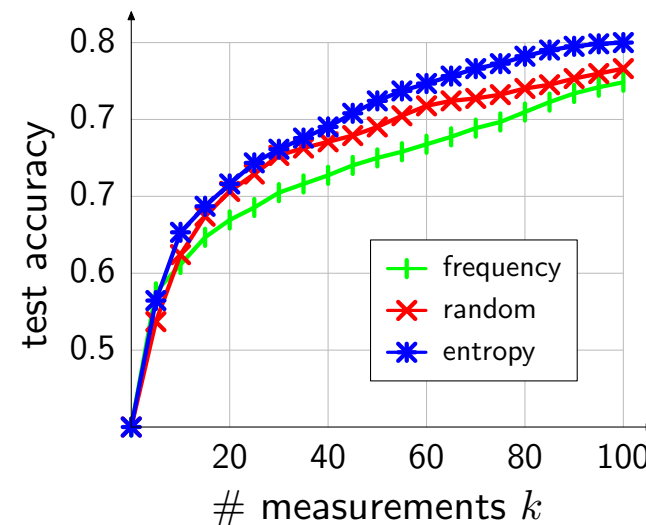
- Fully-labeled examples
- Labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = the, y_i = \text{DT}]$ )

Use label entropy as surrogate for assessing measurements

**Test accuracy (on 100 examples):**



(a) Labeling examples



(b) Labeling word types

# Part-of-speech tagging results

$n = 1000$  total examples (sentences), 45 possible labels

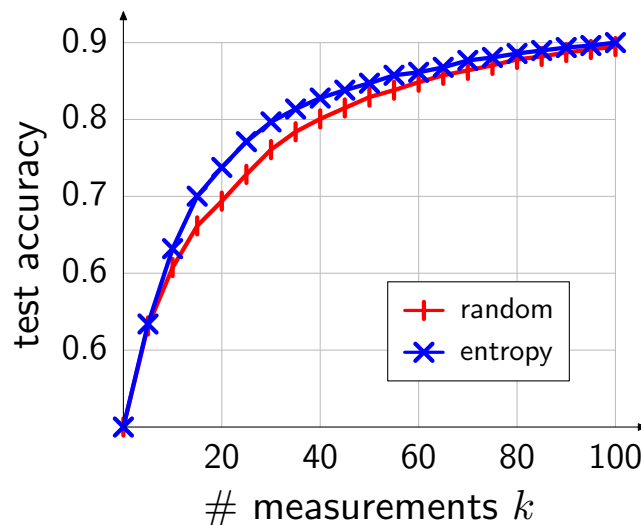
**Model:** Indep. logistic regression with standard NLP features

**Measurements:**

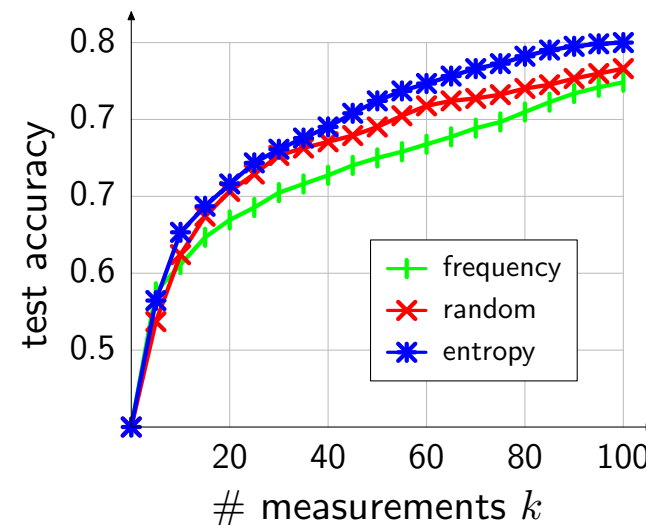
- Fully-labeled examples
- Labeled predicates (e.g.,  $\sum_i \mathbb{I}[x_i = the, y_i = \text{DT}]$ )

Use label entropy as surrogate for assessing measurements

**Test accuracy (on 100 examples):**



(a) Labeling examples



(b) Labeling word types

# Summary

- Think beyond standard supervised/semi-supervised learning

# Summary

- Think beyond standard supervised/semi-supervised learning
- Goal: learn best predictor in a cost-effective way





# Summary

- Think beyond standard supervised/semi-supervised learning
- Goal: learn best predictor in a cost-effective way



# Summary

- Think beyond standard supervised/semi-supervised learning
- Goal: learn best predictor in a cost-effective way



- Use a Bayesian decision-theoretic framework