Active Learning for Sparse Bayesian Multilabel Classification

Deepak Vasisht, MIT & IIT Delhi
Andreas Domianou, University of Sheffield
Manik Varma, MSR, India
Ashish Kapoor, MSR, Redmond

Multilabel Classification

Given a set of datapoints, the goal is to annotate them with a set of labels.

Multilabel Classification

Given a set of datapoints, the goal is to annotate them with a set of labels.

Multilabel Classification

Given a set of datapoints, the goal is to annotate them with a set of labels.

$$x_i \in \mathcal{R}^d$$

Feature vector, d: dimension of the feature space

Multilabel Classification

Given a set of datapoints, the goal is to annotate them with a set of labels.

$$x_i \in \mathcal{R}^d$$

Feature vector, d: dimension of the feature space

Iraq	Flowers	Human	Brick	Sea	Sun	Sky

Multilabel Classification

Given a set of datapoints, the goal is to annotate them with a set of labels.

$$x_i \in \mathcal{R}^d$$

Feature vector, d: dimension of the feature space

Iraq	Flowers	Human	Brick	Sea	Sun	Sky
✓		✓				✓

Training


LIVE Central London
JEREMY COOK
Chief Economist, World First
HURRICANE WARNING
26.01.09 Frodo geboren
news-tv
19:28
Onlinewahn.de übernimmt Microsoft
PITCH FX
INSIDE
IRAQ

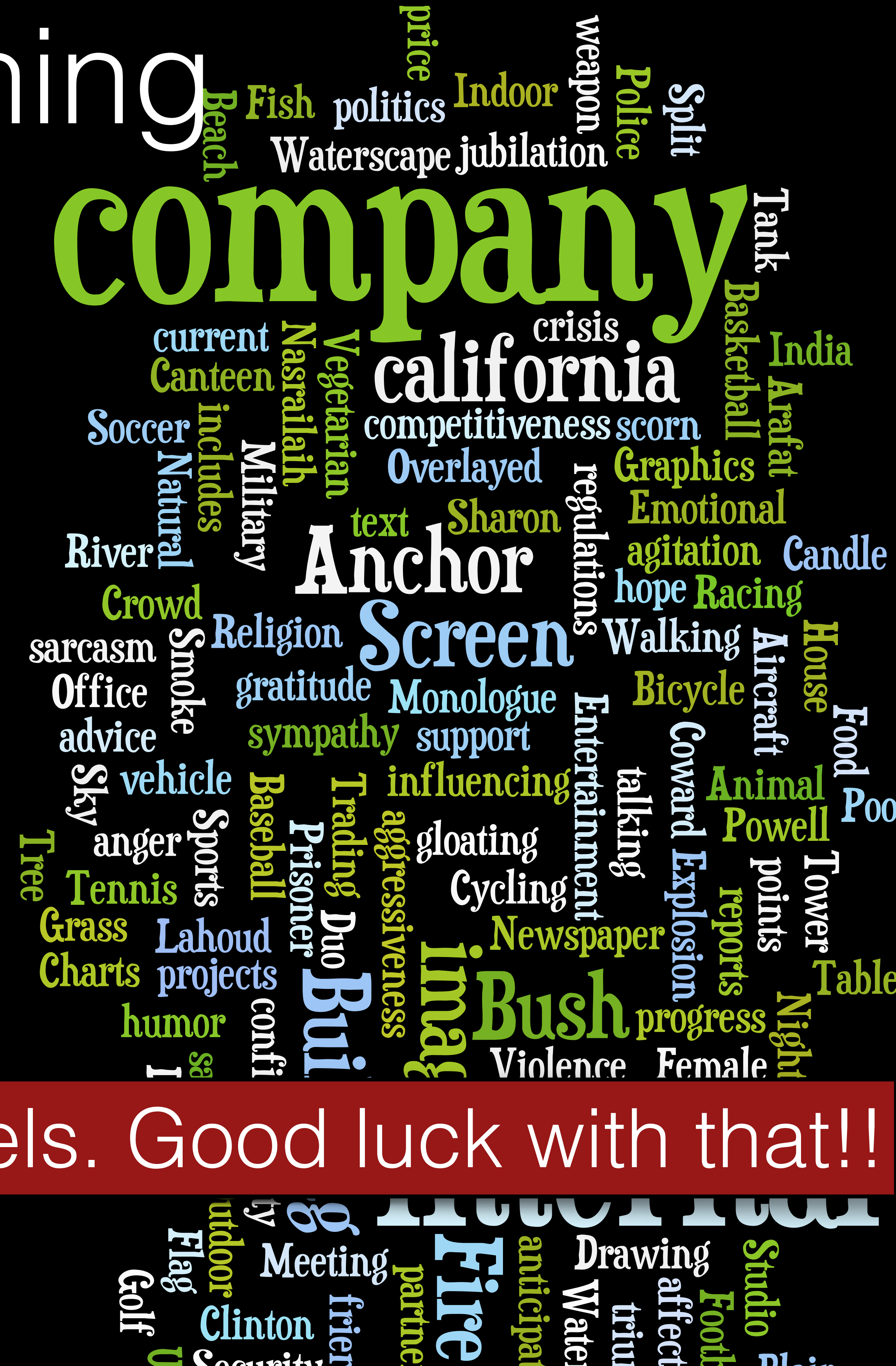

Desert
Ship
Beach
Urban
Truck
Baseball
Police
Waterscape
Blair
Meeting
Prisoner
Sports
Walking
Bus
Crowd
Building
Newspaper
Arafat
Office
Airplane
Vegetation
Entertainment
Computer
Studio
Military
Road
TV-screen
Flag-US
Maps
Face
Explosion
Court
Natural-Disaster
Boat
Outdoor
Person
Car
Bicycle
Weather
Animal
Allawi
Mountain
Government
Basketball
Sky
Charts
Running
Waterfront
Snow
Corporate-Leader

Training


LIVE Central London
JEREMY COOK
Chief Economist, World First
TV8 NEWS
HURRICANE WARNING
26.01.09 Frodo geboren
news-tv
19:28
Onlinewahn.de übernimmt Microsoft
PITCH FX
INSIDE
IRAQ


Building
Anchor
Screen
Fire
Bush
Corporation
Vegetarian
Disaster
Outdoor
Cloud
text
Dog
Table
Female
Grass
Chair
Split
Urban
Entertainment
Military
Boat
Tank
Monologue
Overlayed
Night
Allawi
USA
Swimming
Drawing
Sports
Road
House
Flag
Police
Powell
Sky
Bus
Beach
Walking
Waterfall
Baseball
Smoke
Trading
Cartoon
Weapon
Studio
Map
Duo
Fish
Leader
Crafts
People
Bicycle
Religion
Marching
Blair
Violence
Basketball
Candle
Clinton
Canteen
Truck
Golf
Food
Car
Government
Animal
India
Tower
Arafat
Coward
Aircraft
Pool
Charts
Explosion
Security
Racing
Meeting
Prisoner
Natural
Jintao
Weather
Horse
River
Crowd
Sharon
Indoor
Tree
Office
Lahoud
Football
Cycling
Motorbike
Waterscape
Newspaper
Mountain
Soccer
Tennis
Snow
Court
Kerry
vehicle
Nasrailaih
Face
Dessert
Graphics
Bird

Training



WikiLSHTC has 325k labels. Good luck with that!!

Training Is Expensive

- Training data can also be very expensive, like genomic data, chemical data
- Getting each label incurs additional cost

Training Is Expensive

- Training data can also be very expensive, like genomic data, chemical data
- Getting each label incurs additional cost

Need to reduce the required training data.

Active Learning

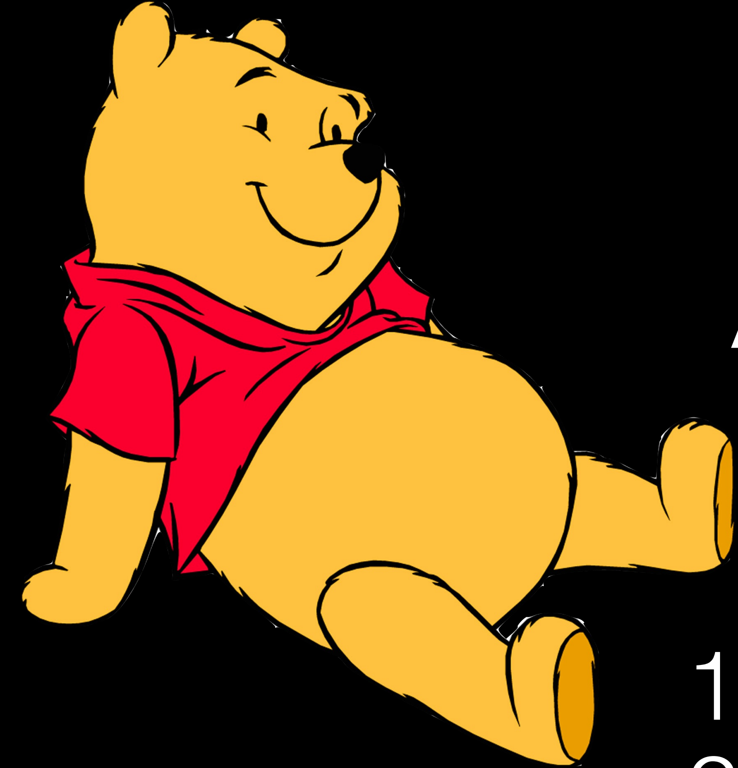
Datapoints

Active Learning

Labels

	Iraq 1	Flowers 2	3	…	…	…	Sun	Sky L
1								
2		1					0	
3								
⋮								
⋮								
⋮								
⋮								
⋮								
⋮								
⋮								
⋮								
⋮								
⋮								
N								

Datapoints

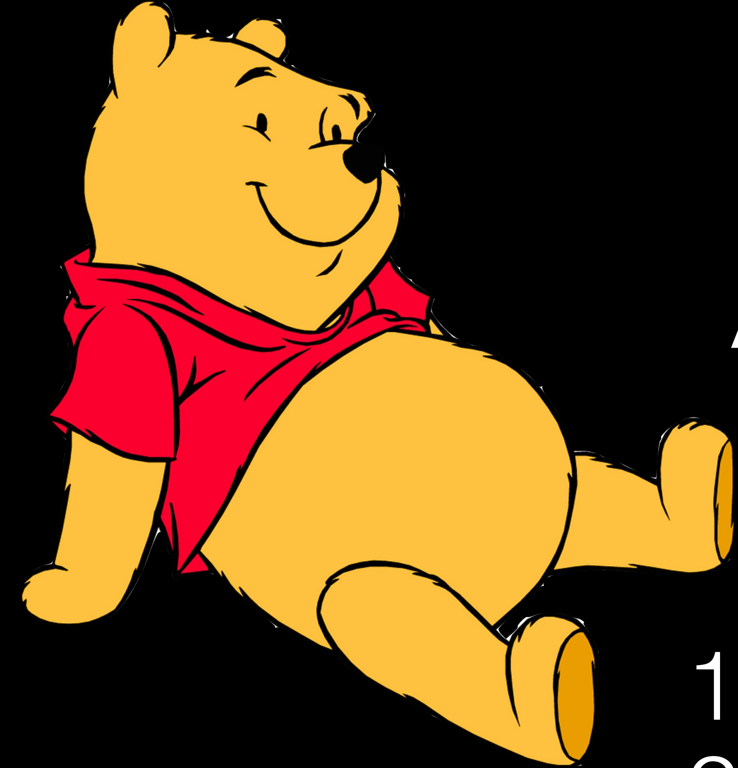


Active Learning

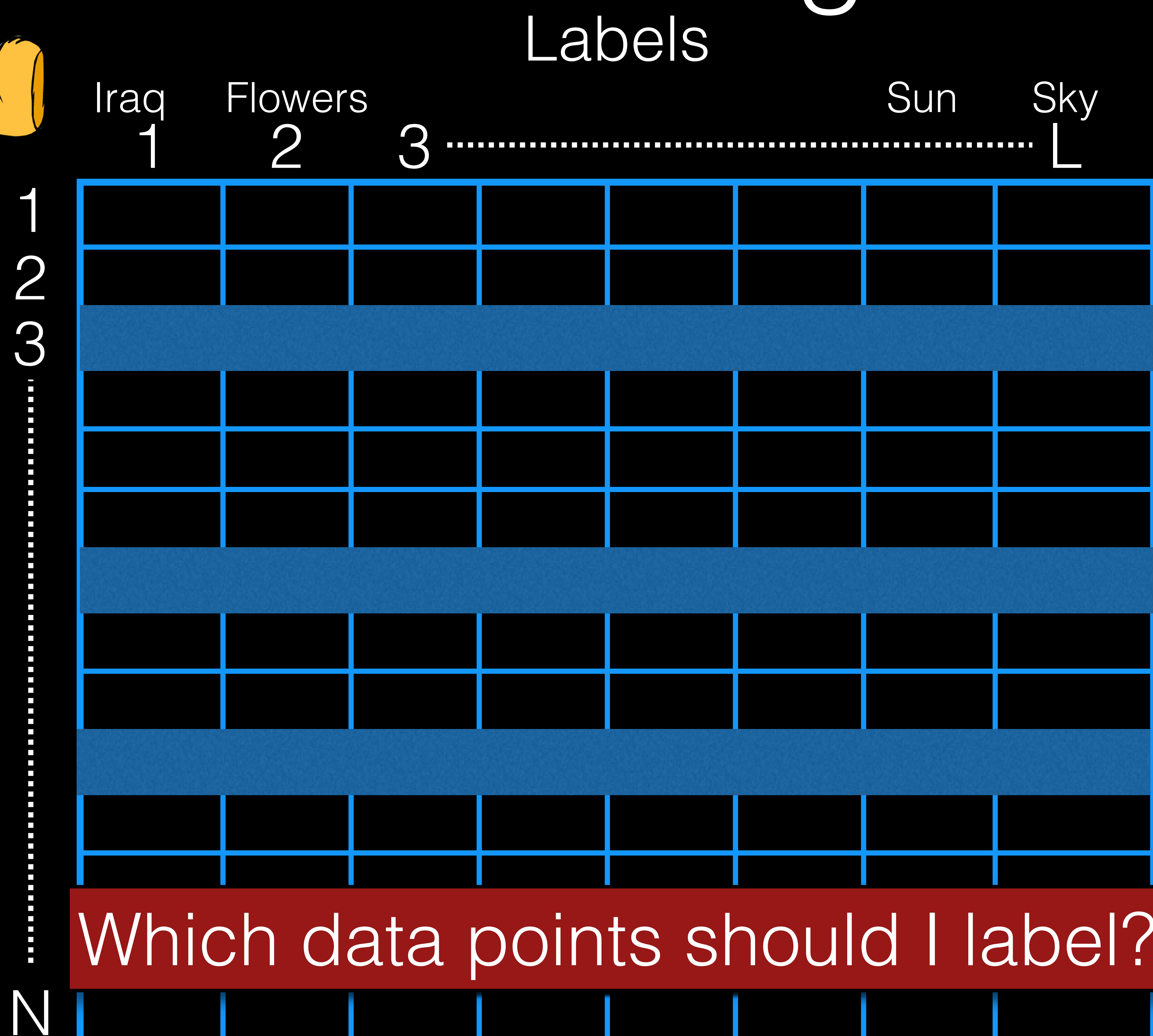
Labels

	Iraq 1	Flowers 2	3				Sun	Sky L
1								
2		1					0	
3								
N								

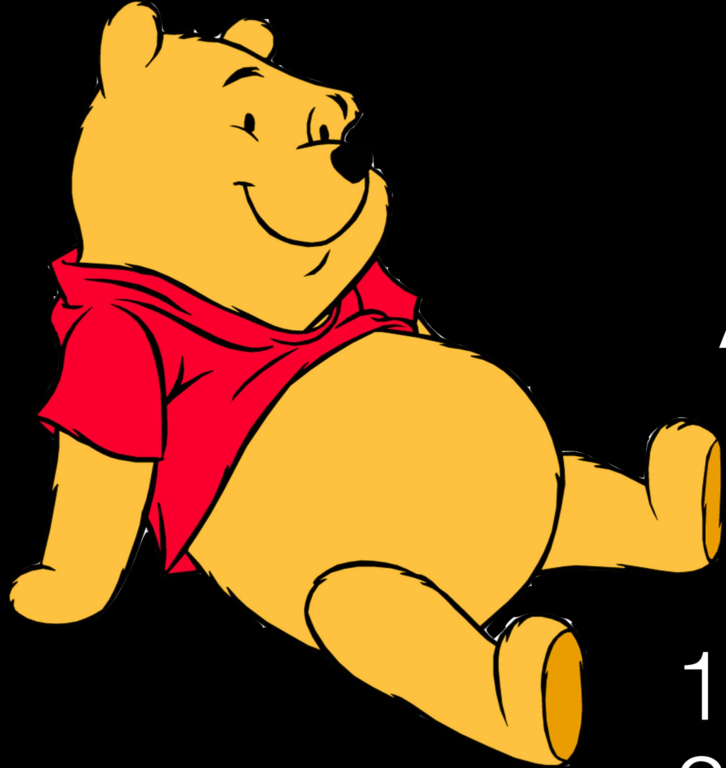
Datapoints



Active Learning

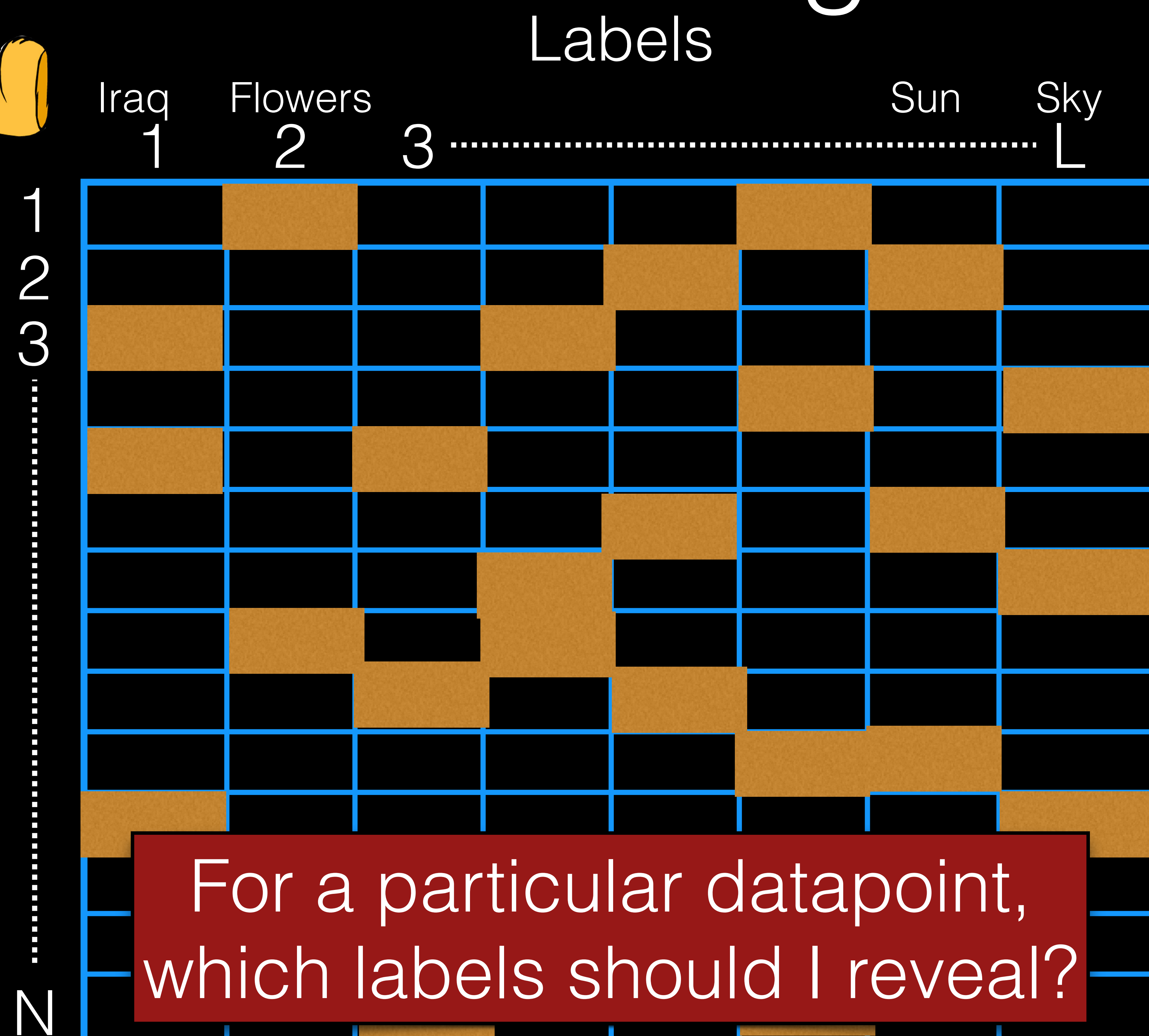


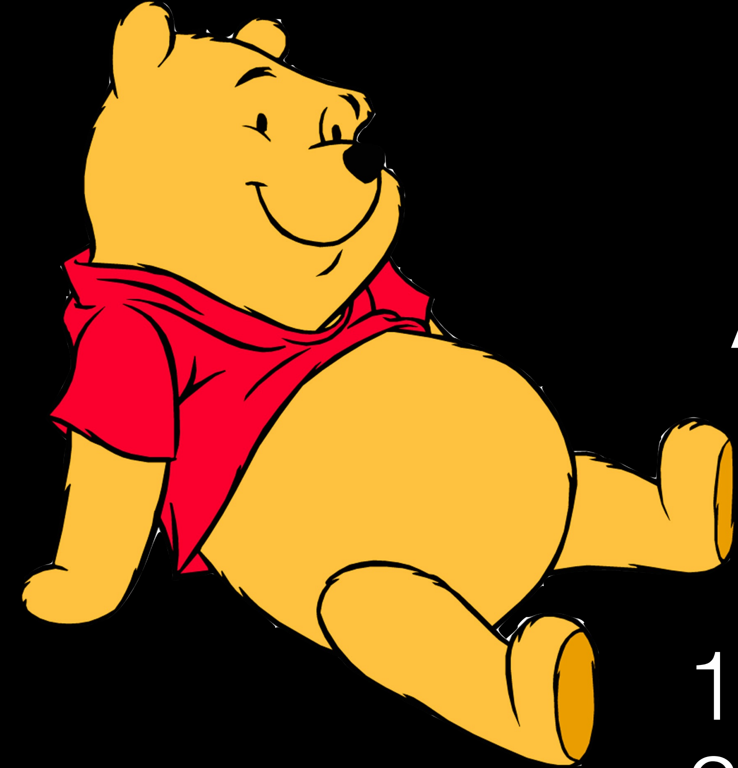
Datapoints



Active Learning

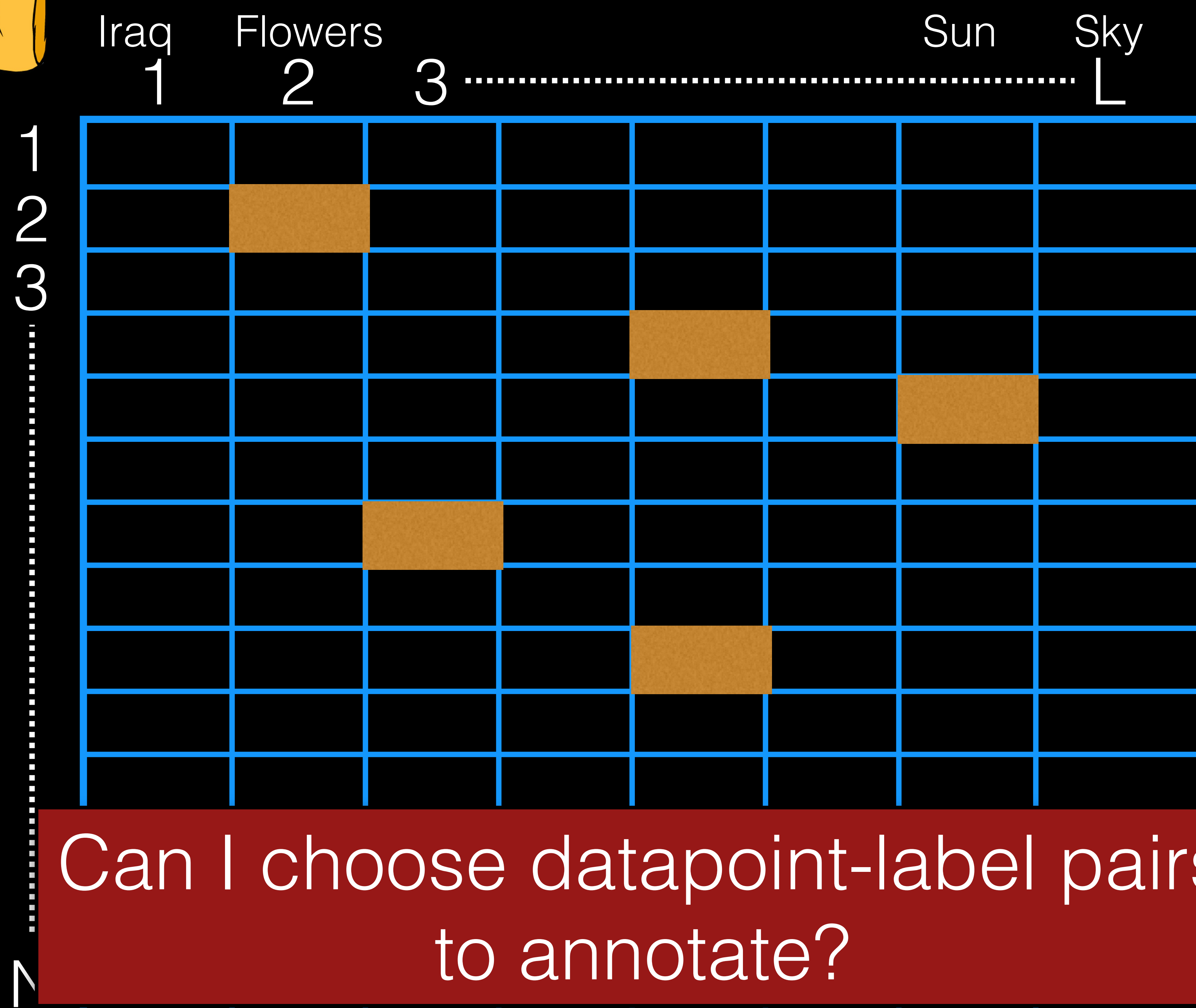
Datapoints





Active Learning

Datapoints



Can I choose datapoint-label pairs to annotate?

In this talk

- An active learner for Multi-label classification that:
  - Answers all your questions
  - Is Computationally Cheap
  - Is Non myopic and near-optimal
  - Incorporates label sparsity
  - Achieves higher accuracy than state-of-the-art

Classification

Classification Model*

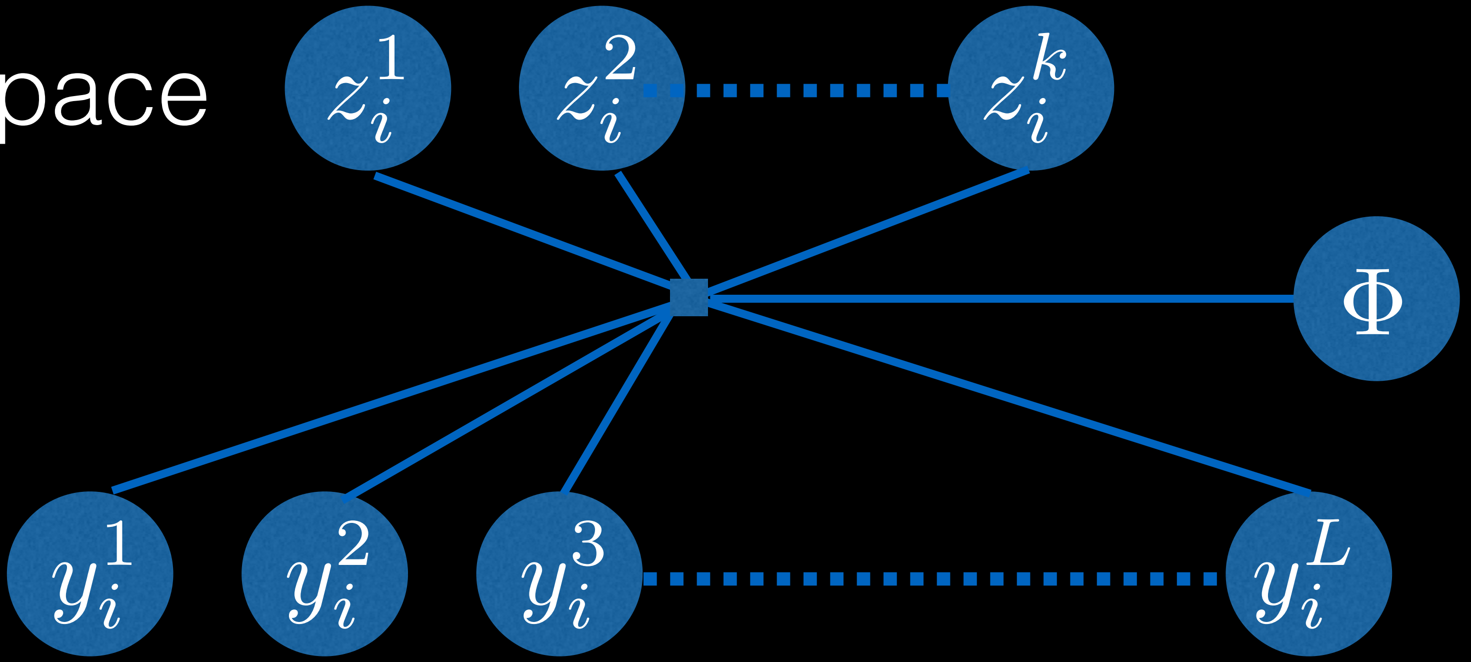
x_i

Labels y_i^1 y_i^2 y_i^3 $\cdots$ y_i^L

*Kapoor et al, NIPS 2012

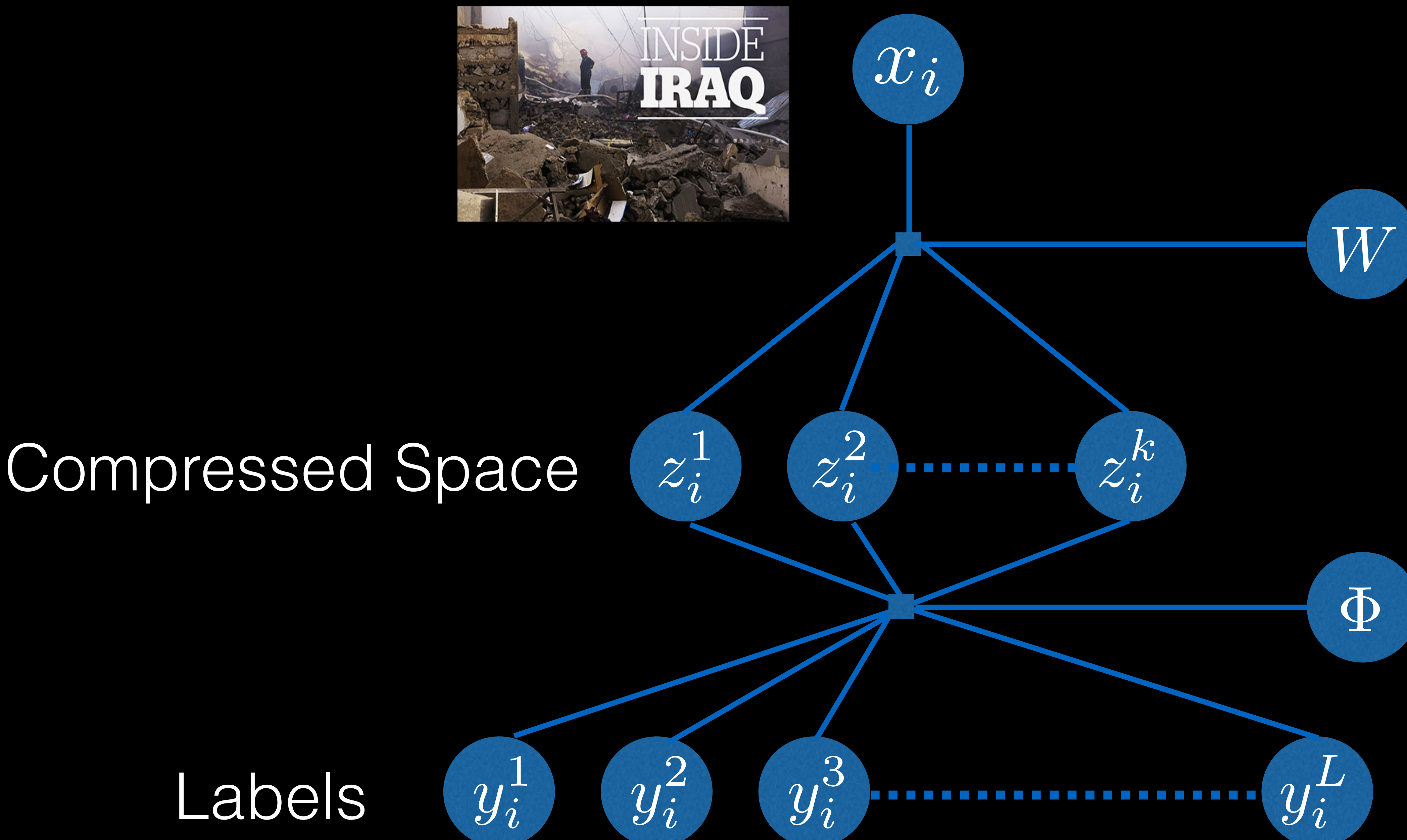
Classification Model*

x_i



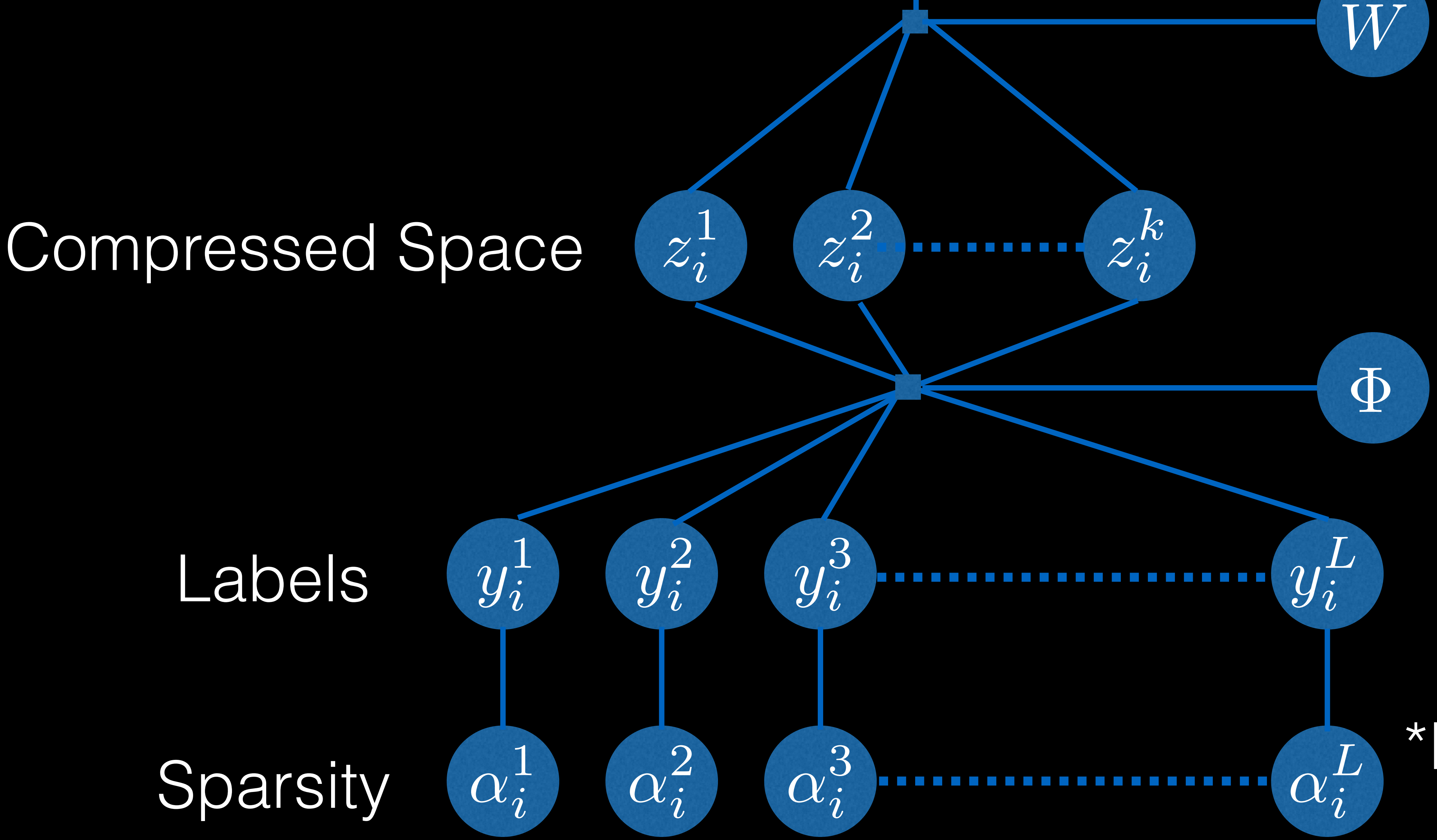
*Kapoor et al, NIPS 2012

Classification Model*



*Kapoor et al, NIPS 2012

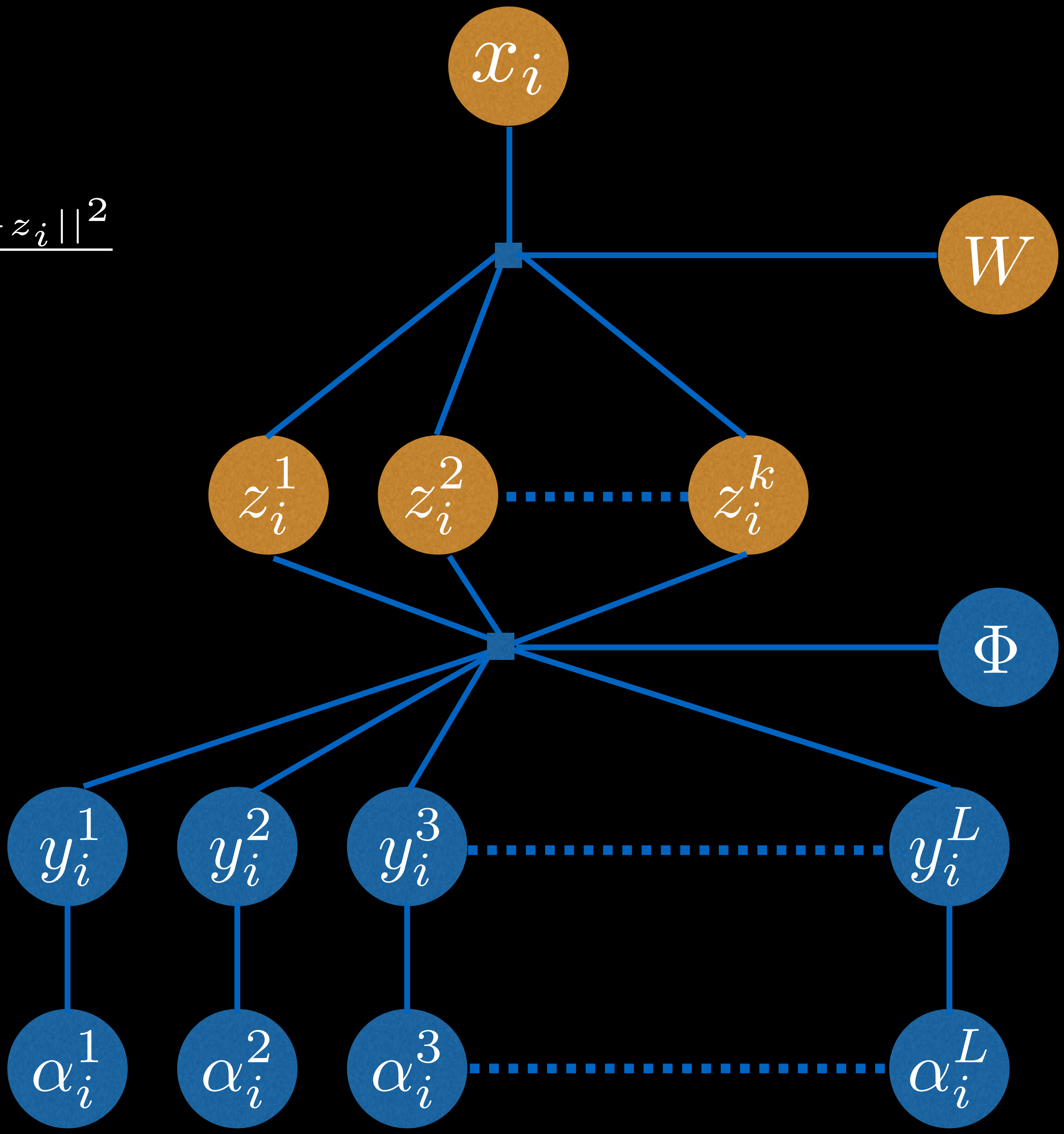
Classification Model*



*Kapoor et al, NIPS 2012

Classification Model: Potentials

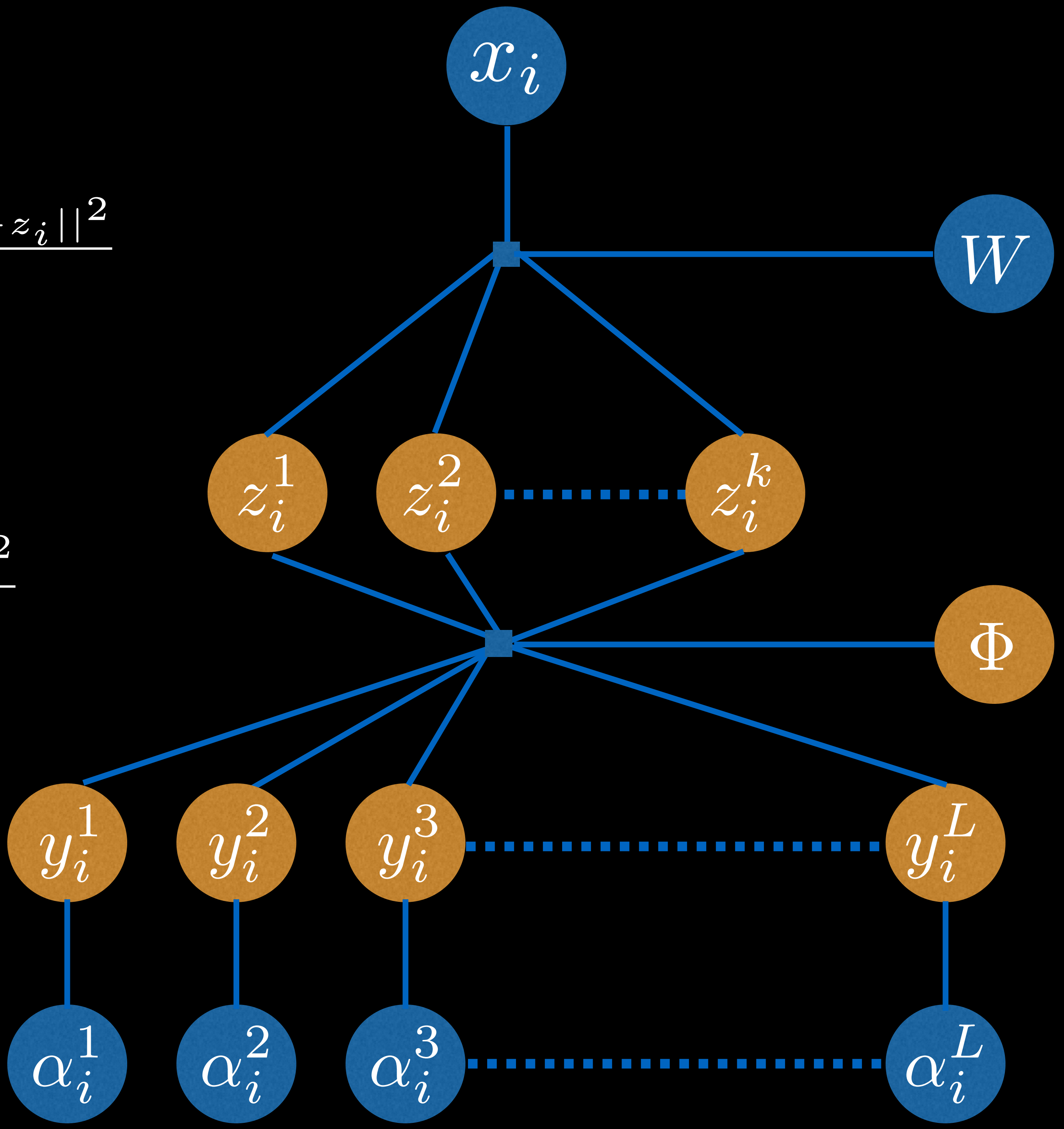
$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$



Classification Model: Potentials

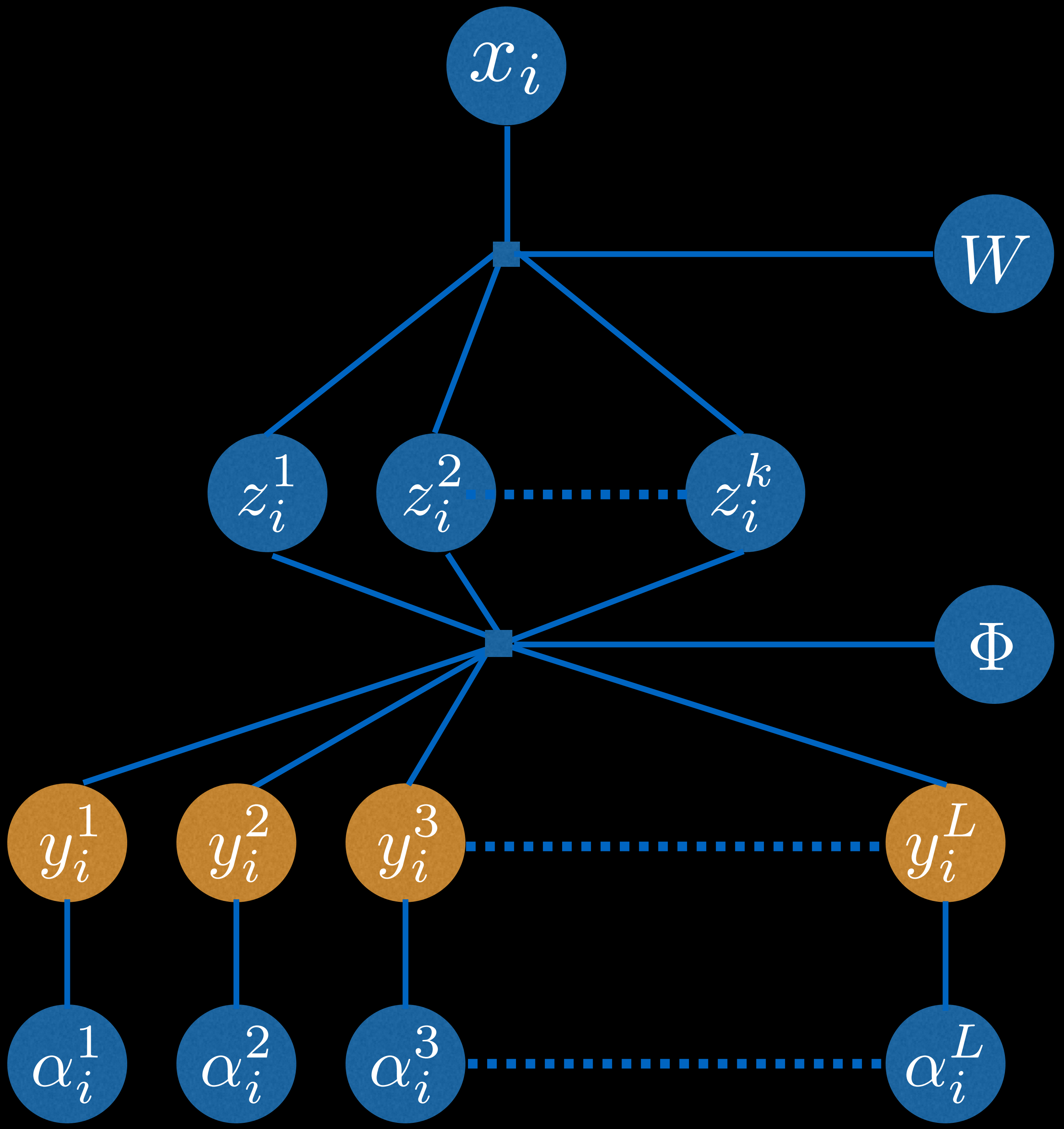
$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

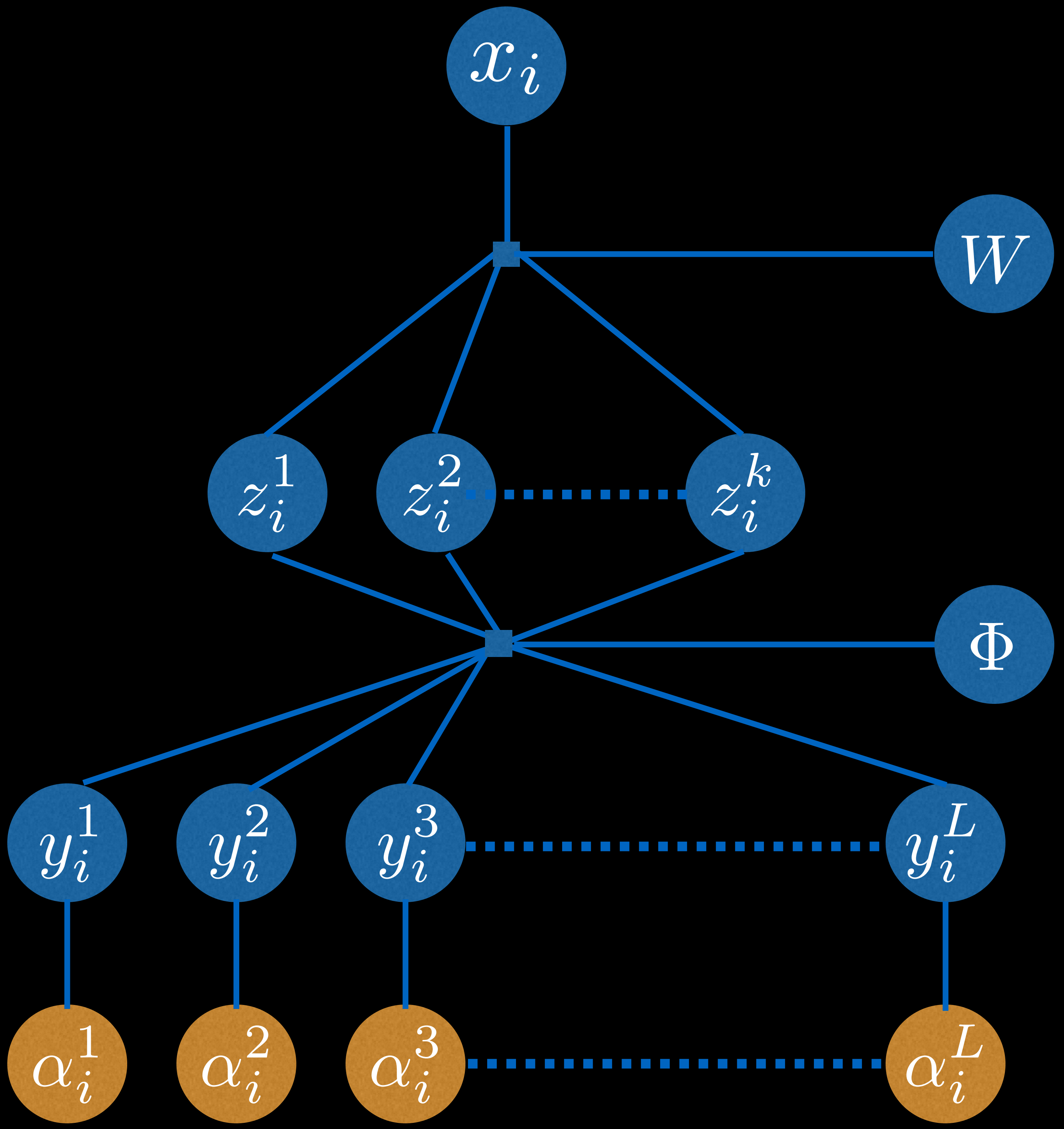


Classification Model: Priors

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$



Classification Model: Priors

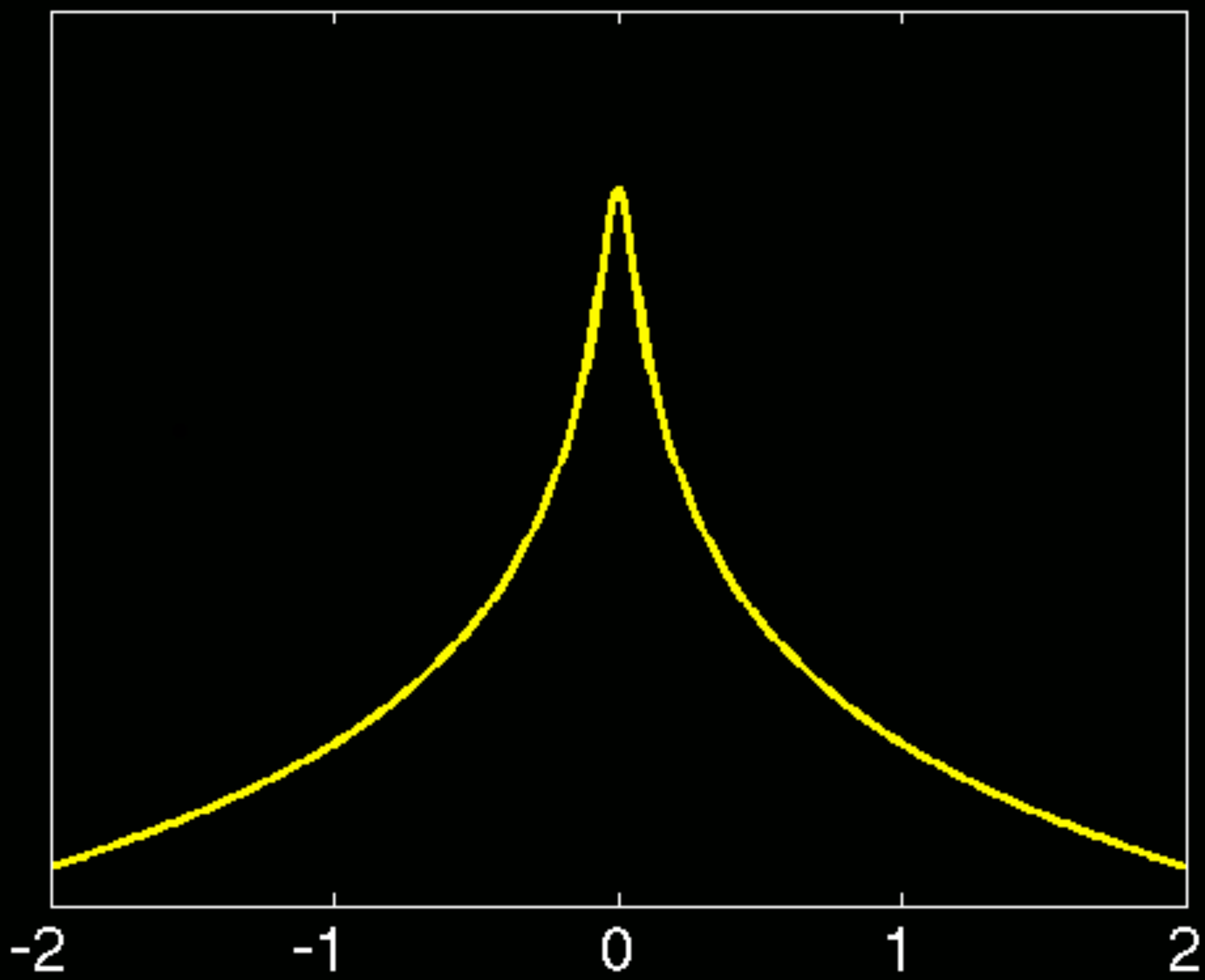


$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$

Sparsity Priors

$$a_0 = 10^{-6}, b_0 = 10^{-6}$$



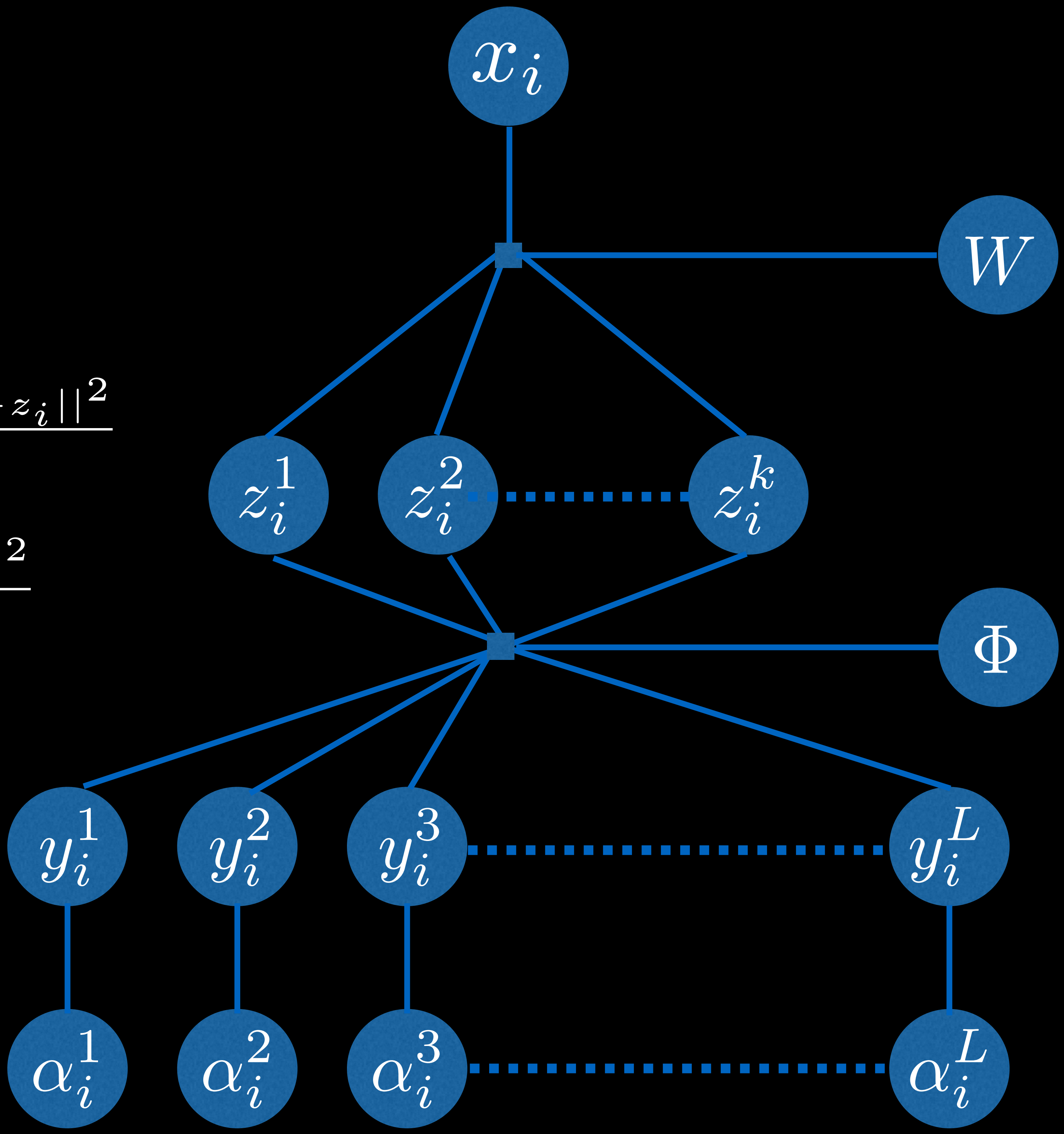
Classification Model

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



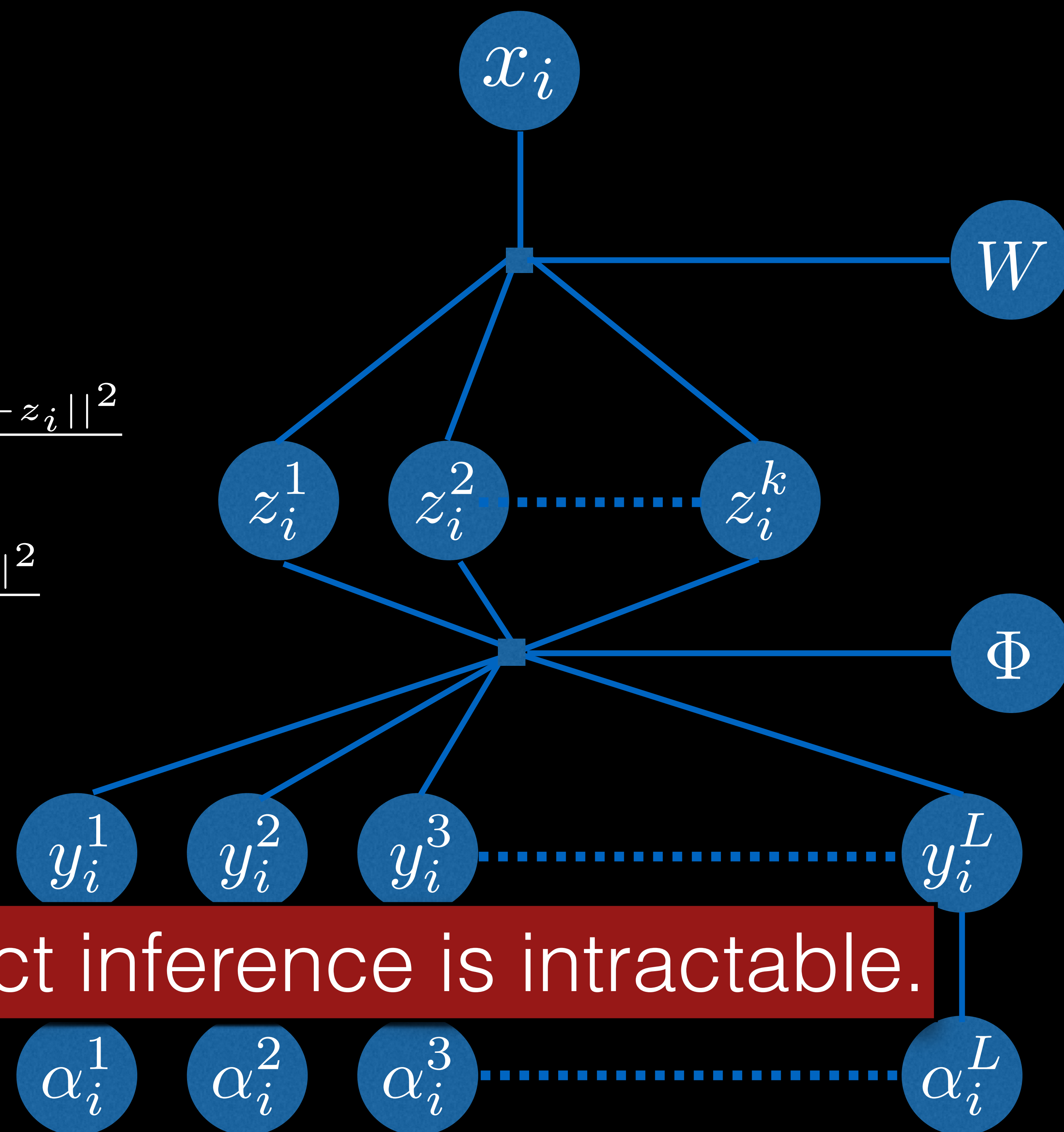
Classification Model

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

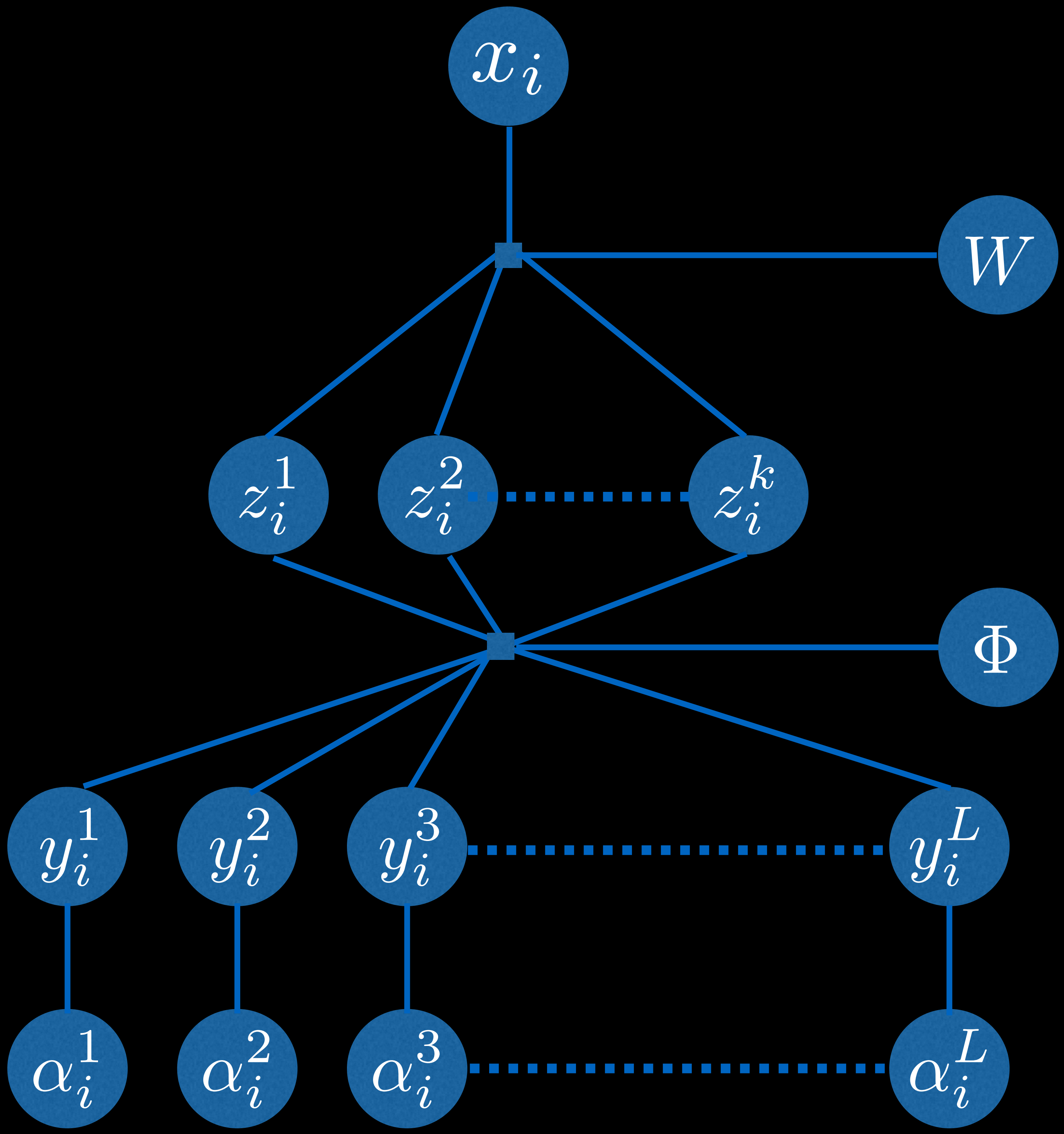
$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



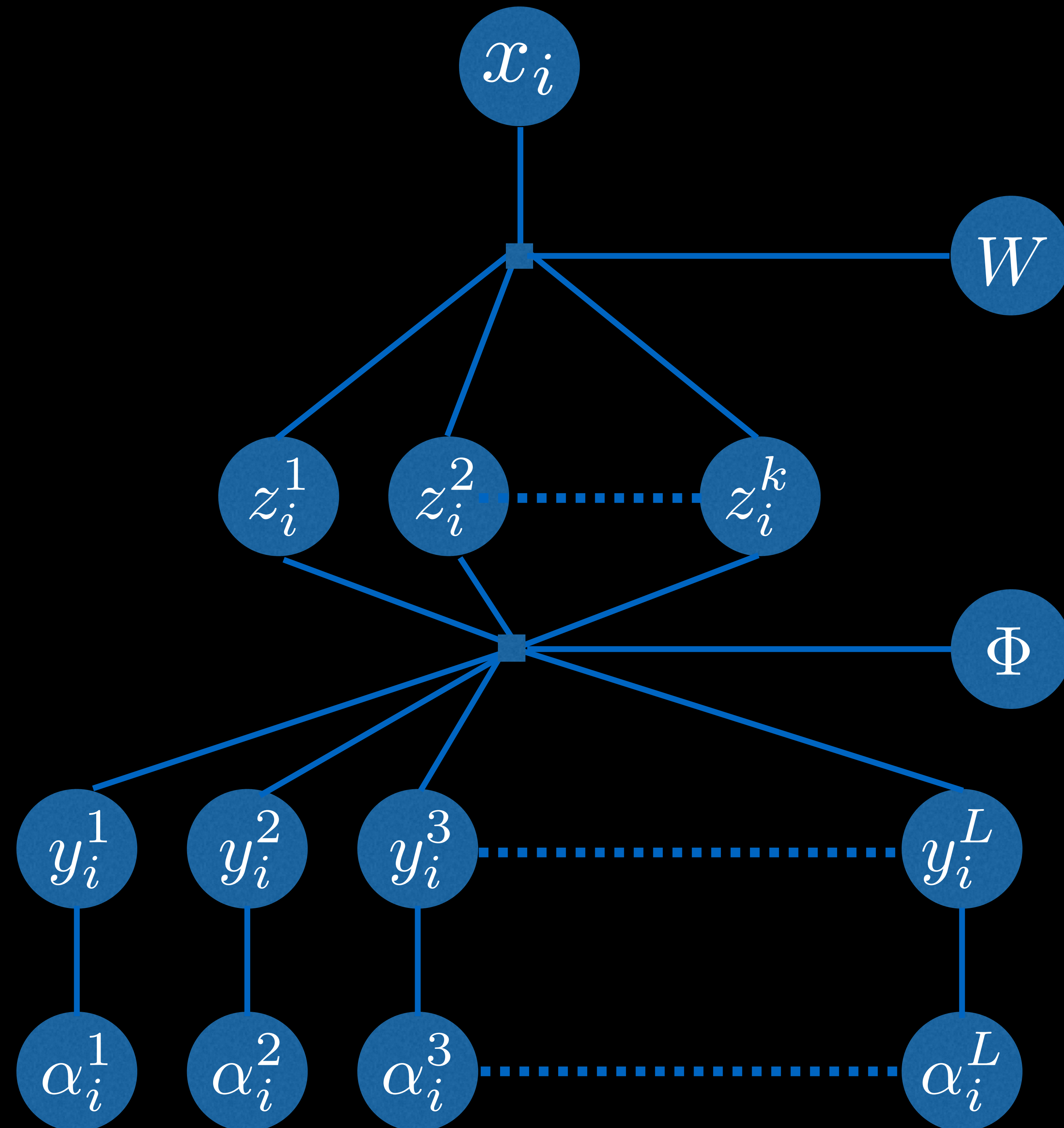
Problem: Exact inference is intractable.

Inference: Variational Bayes



Inference: Variational Bayes

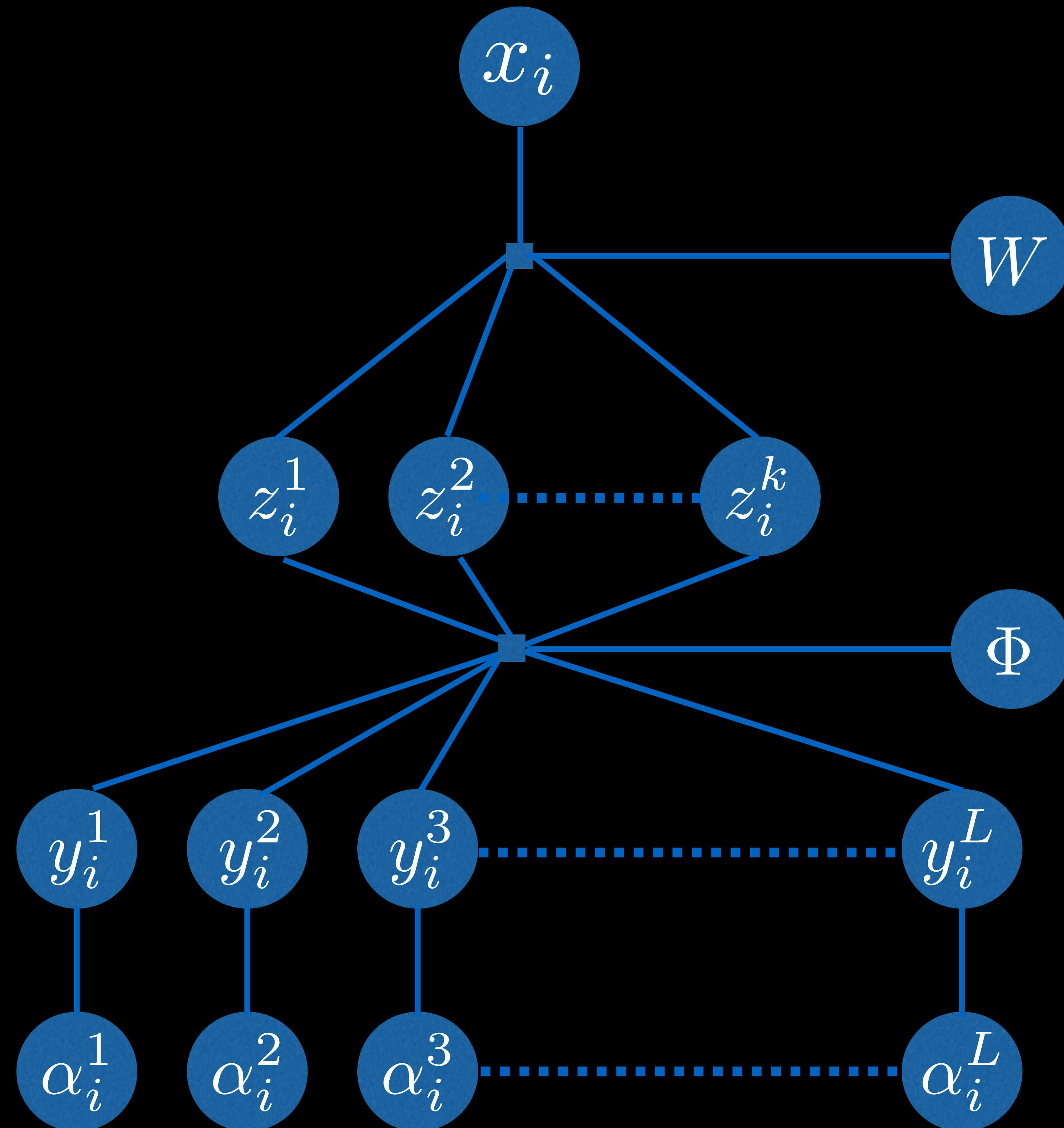
Approximate Gaussian



Inference: Variational Bayes

Approximate Gaussian

Approximate Gaussian

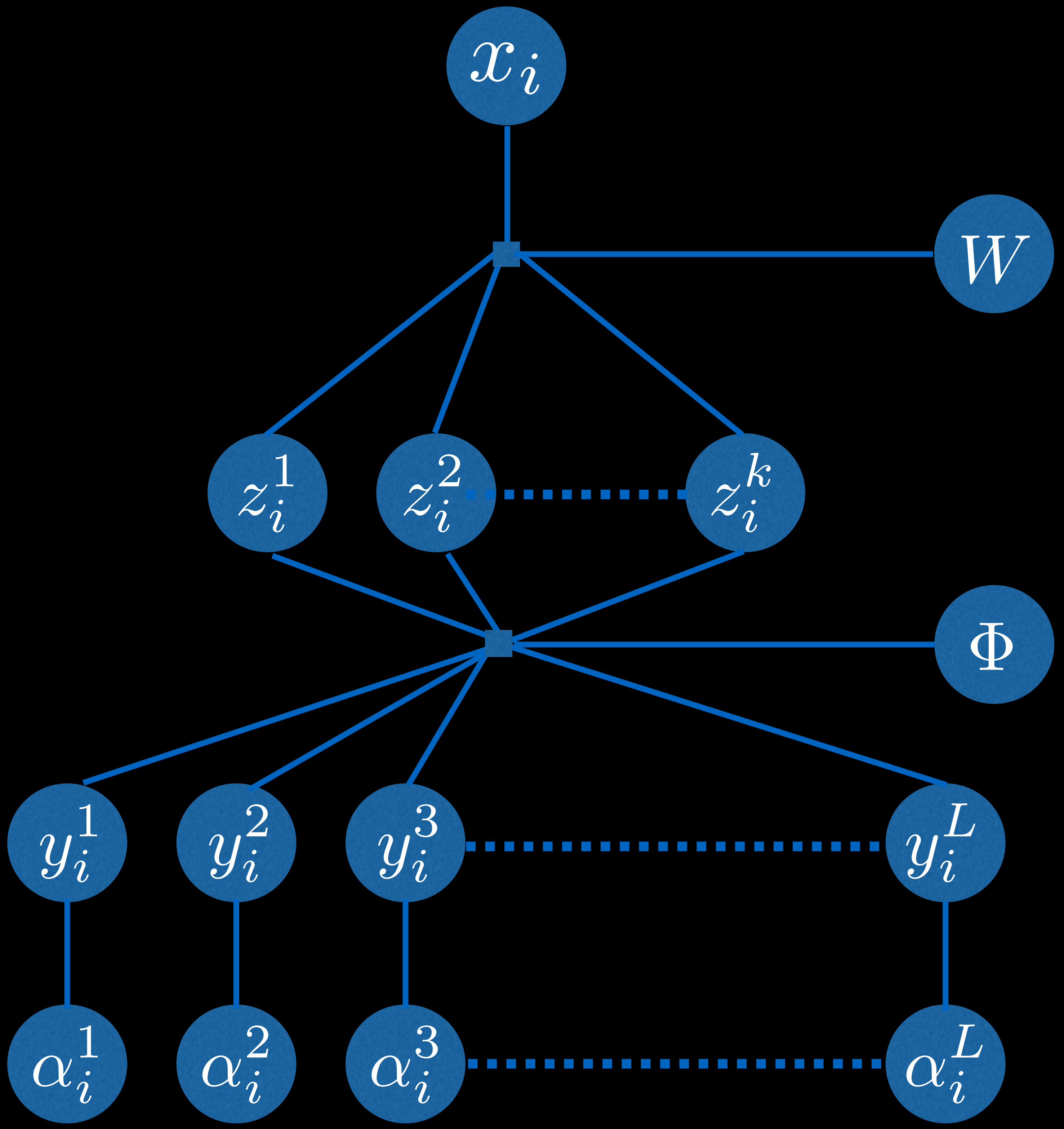


Inference: Variational Bayes

Approximate Gaussian

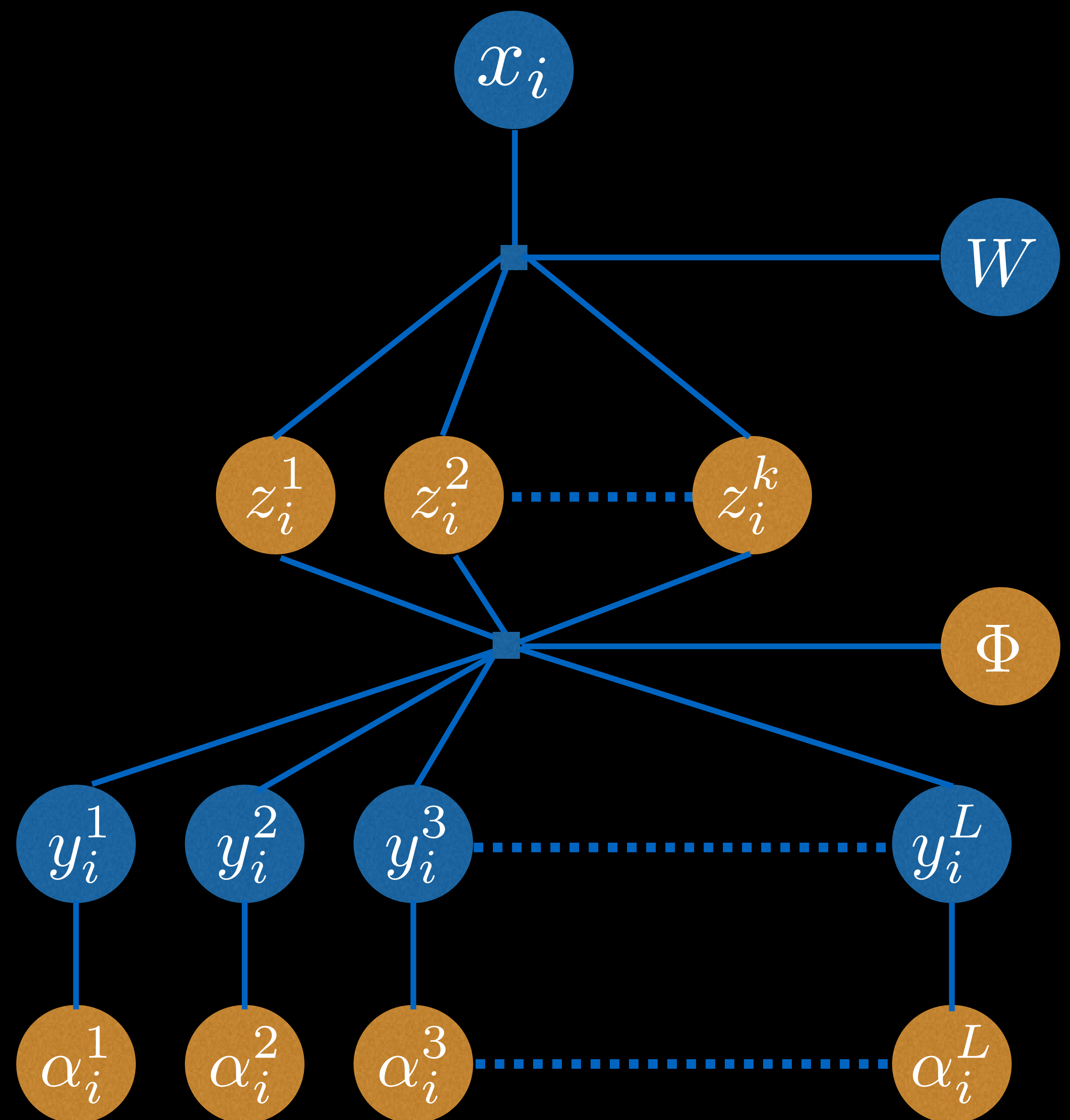
Approximate Gaussian

Approximate Gamma



Inference: Variational Bayes

Approximate Gaussian



Active Learning Criteria

- Entropy: Is a measure of uncertainty. For a random variable X, the entropy H is given as:

$$H(X) = -\sum_i P(x_i)\log(P(x_i))$$

- Picks points far apart from each other
- For a Gaussian process, $H = \frac{1}{2}\log(|\Sigma|) + const$

Active Learning Criteria

- Mutual Information: Measures reduction in uncertainty over unlabeled space

$$MI(A, B) = H(A) - H(A|B)$$

- Used in past work successfully for regression

Active Learning: Mutual Information

- We have already modeled the distribution over labels, Y as a Gaussian process

- The goal is to select a subset of labels that offers the maximum reduction in entropy over the remaining space

$$\mathcal{A}^* = \arg_{\mathcal{A} \subseteq \mathcal{U}} \max H(Y_{\mathcal{U} \setminus \mathcal{A}}) - H(Y_{\mathcal{U} \setminus \mathcal{A}} | \mathcal{A})$$

Active Learning: Mutual Information

- We have already modeled the distribution over labels, Y as a Gaussian process

- The goal is to select a subset of labels that offers the maximum reduction in entropy over the remaining space

Problem: Variance is not preserved across layers

Idea: Collapsed Variational Bayes

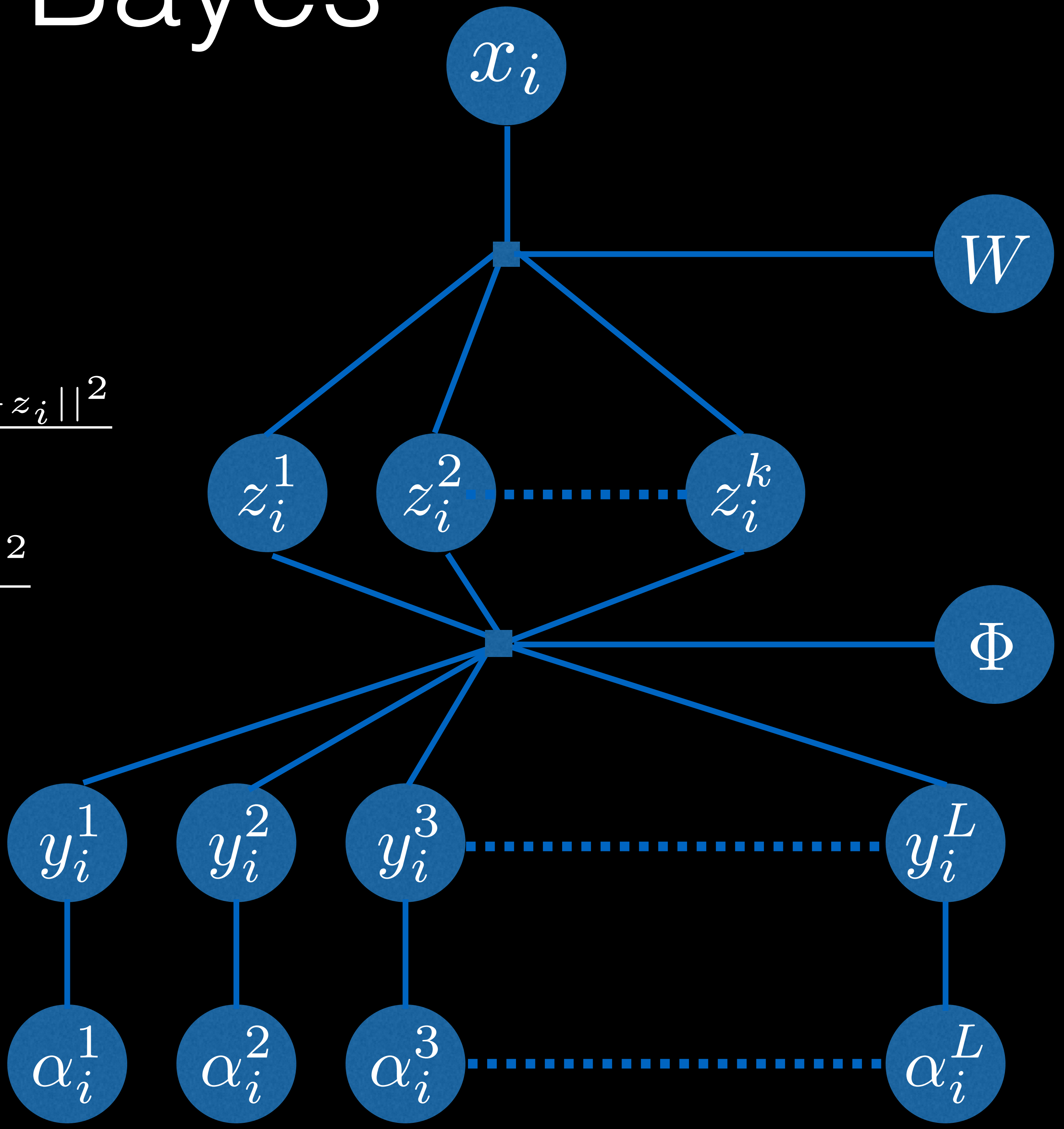
Idea: Collapsed Variational Bayes

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



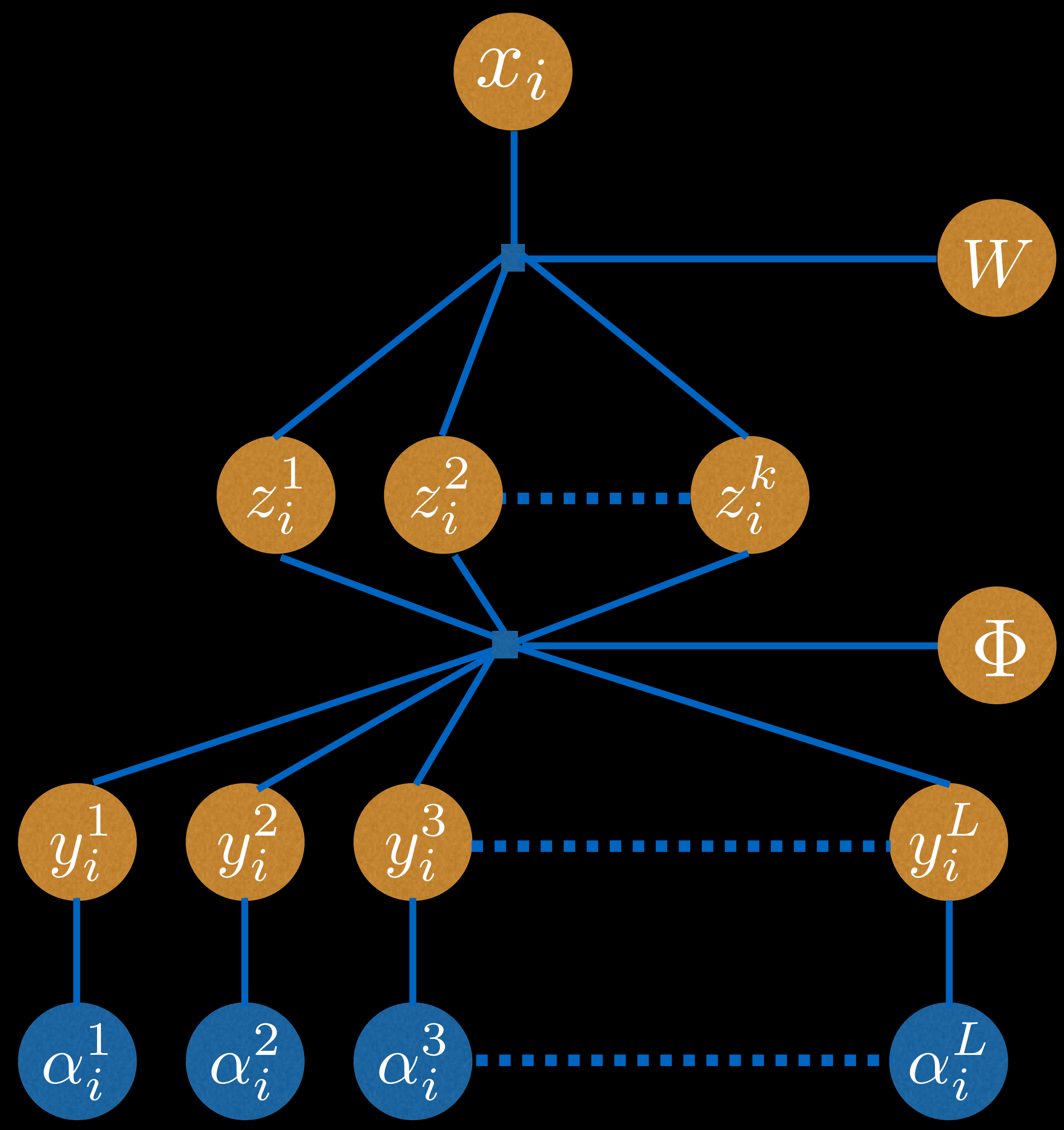
Idea: Collapsed Variational Bayes

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



Idea: Collapsed Variational Bayes

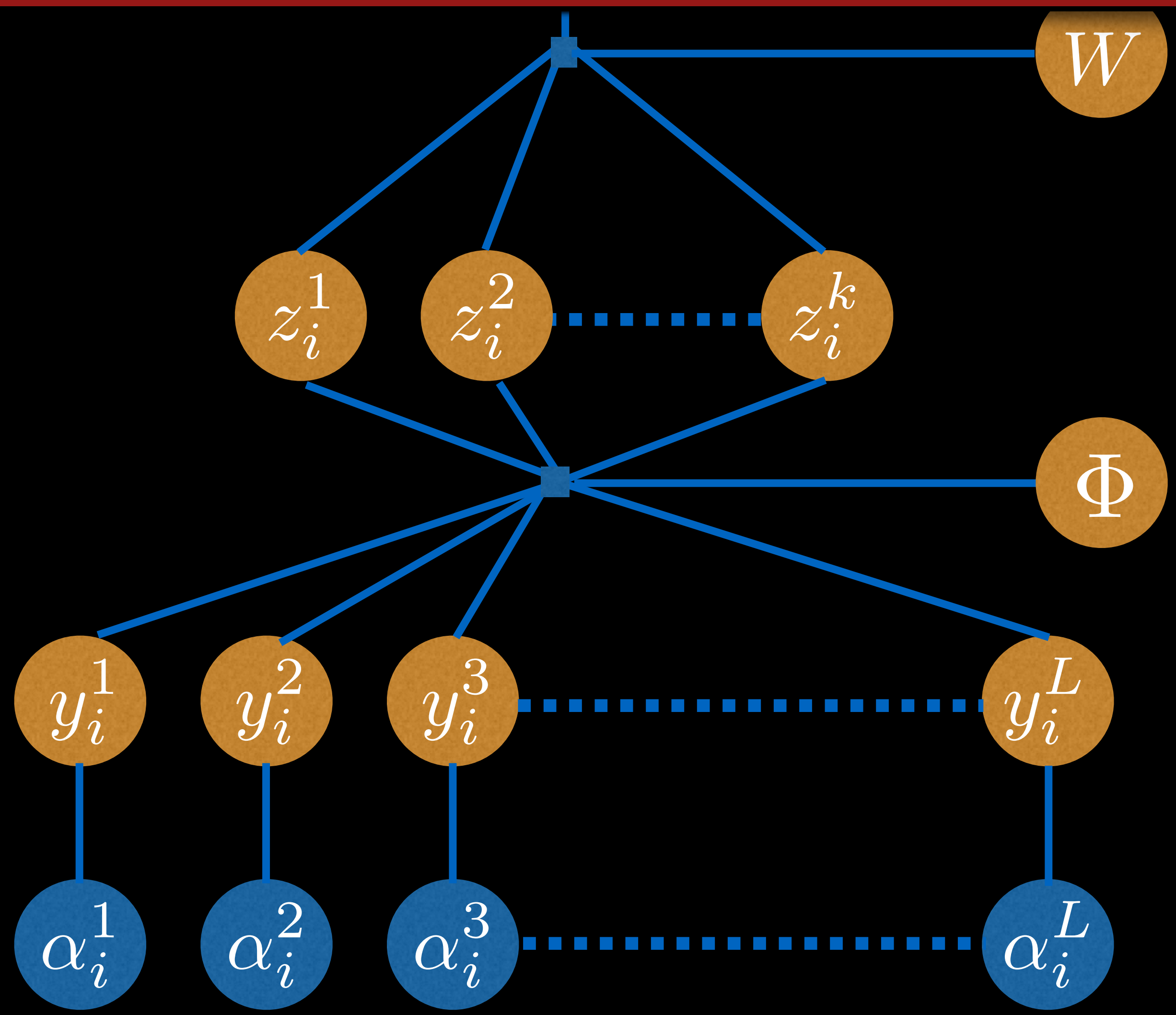
Integrate to get a Gaussian distribution over Y

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



Idea: Collapsed Variational Bayes

Integrate to get a Gaussian distribution over Y

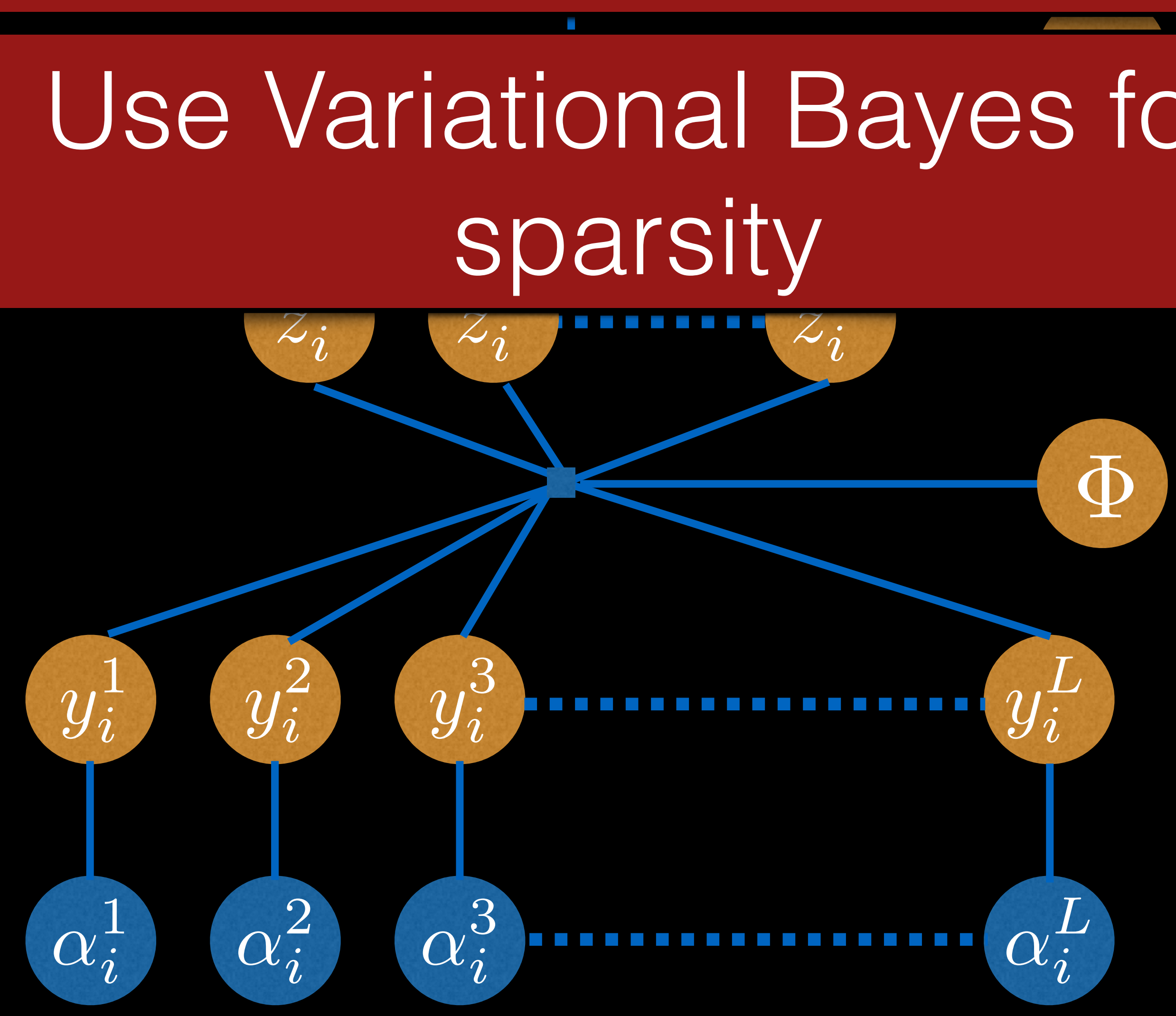
Use Variational Bayes for sparsity

$$f_{x_i}(W, z_i) = e^{-\frac{||W^T x_i - z_i||^2}{2\sigma^2}}$$

$$g_\phi(y_i, z_i) = e^{-\frac{||\Phi y_i - z_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$



Active Learning: Mutual Information

- We have already modeled the distribution over labels, Y as a Gaussian process
- The goal is to select a subset of labels that offers the maximum reduction in entropy over the remaining space

$$\mathcal{A}^* = \arg_{\mathcal{A} \subseteq \mathcal{U}} \max H(Y_{\mathcal{U} \setminus \mathcal{A}}) - H(Y_{\mathcal{U} \setminus \mathcal{A}} | \mathcal{A})$$

Active Learning: Mutual Information

- We have already modeled the distribution over labels, Y as a Gaussian process
- The goal is to select a subset of labels that offers the maximum reduction in entropy over the remaining space

Problem: Computing Mutual Information still needs exponential time

Solution: Approximate Mutual Information

- Approximate the final distribution over Y by a Gaussian
- Use the Gaussian to estimate the mutual information
- Theorem 1: $\lim_{a_0 \to 0, b_0 \to 0} \hat{M}I \to MI$

Active Learning: Mutual Information

- We have already modeled the distribution over labels, Y as a Gaussian process
- The goal is to select a subset of labels that offers the maximum reduction in entropy over the remaining space

Problem: Subset selection problem is NP complete

Solution: Use Submodularity

- Under some weak conditions, the objective is sub-modular
- Sub-modularity ensures that the greedy solution is a constant times the optimal solution

Algorithm

- Input: Feature vectors for a set of unlabeled instance, U and a budget n

- Iteratively, add a datapoint x to labeled set A, such that x leads to maximum increase in MI

$$x \leftarrow \arg \max_{x \in \mathcal{U} \setminus A} \hat{M}I(A \cup x) - \hat{M}I(A)$$

Performance Evaluation

Datasets

Dataset	Type	Instances	Features	Labels
Yeast	Biology	2417	103	14
MSRC	Image	591	1024	23
Medical	Text	978	1449	45
Enron	Text	1702	1001	53
Mediamill	Video	43907	120	101
RCV1	Text	6000	47236	101
Bookmarks	Text	87856	2150	208
Delicious	Text	16105	500	983

Setup

- Unlabeled pool size: 4000 points, test size: 2000 points
- For smaller datasets, the entire data was in unlabeled pool. Testing on all unlabeled data
- Initial seed size: 500 points

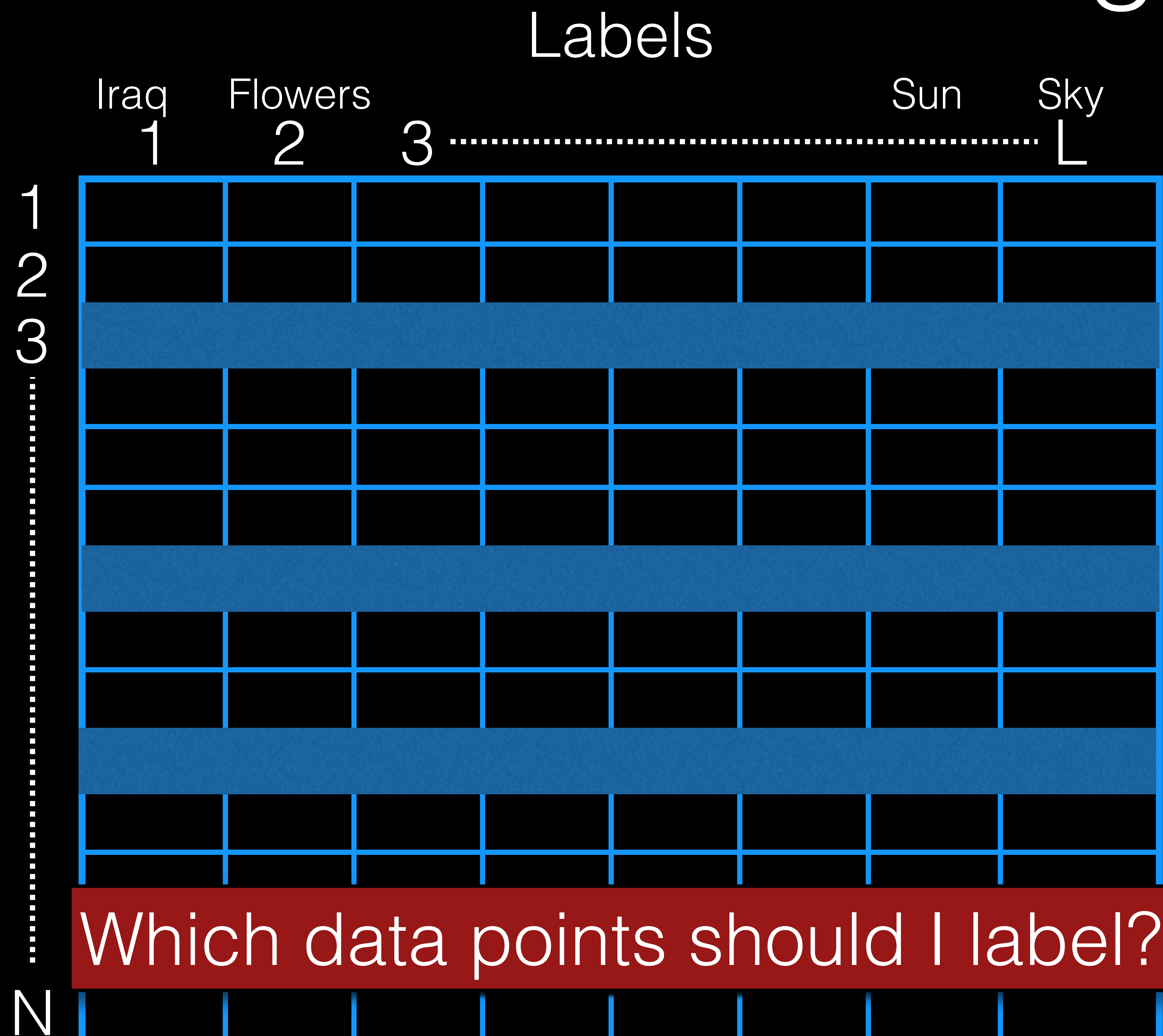
Compared Algorithms

- **MIML:** Mutual Information for Multilabel Classification (proposed method).
- **Uncert:** Uncertainty sampling (Entropy based)
- **Rand:** Random sampling
- **Li-Adaptive*:** SVM based adaptive active learning

*Li et al, IJCAI 2013

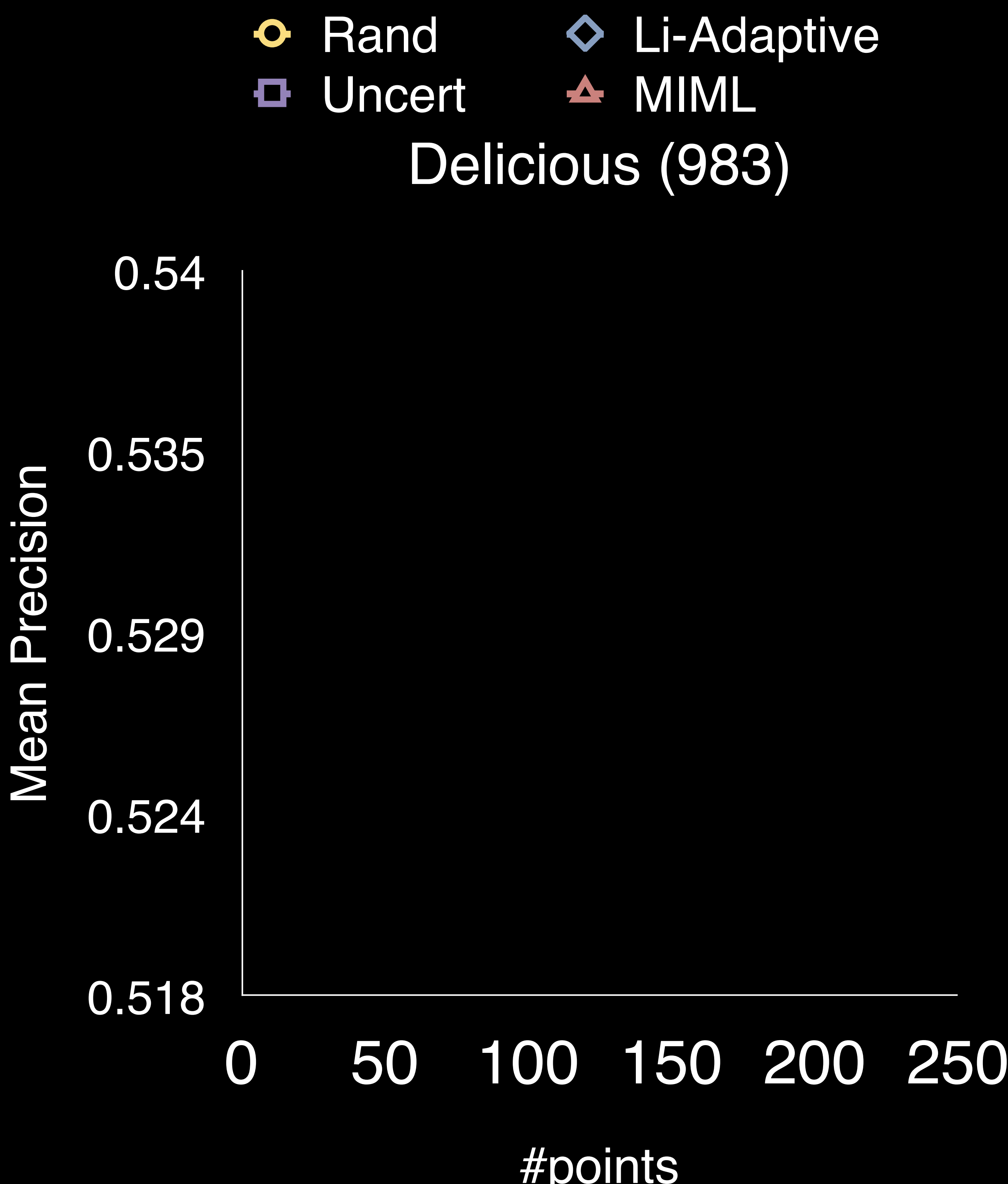
Traditional Active Learning

Datapoints

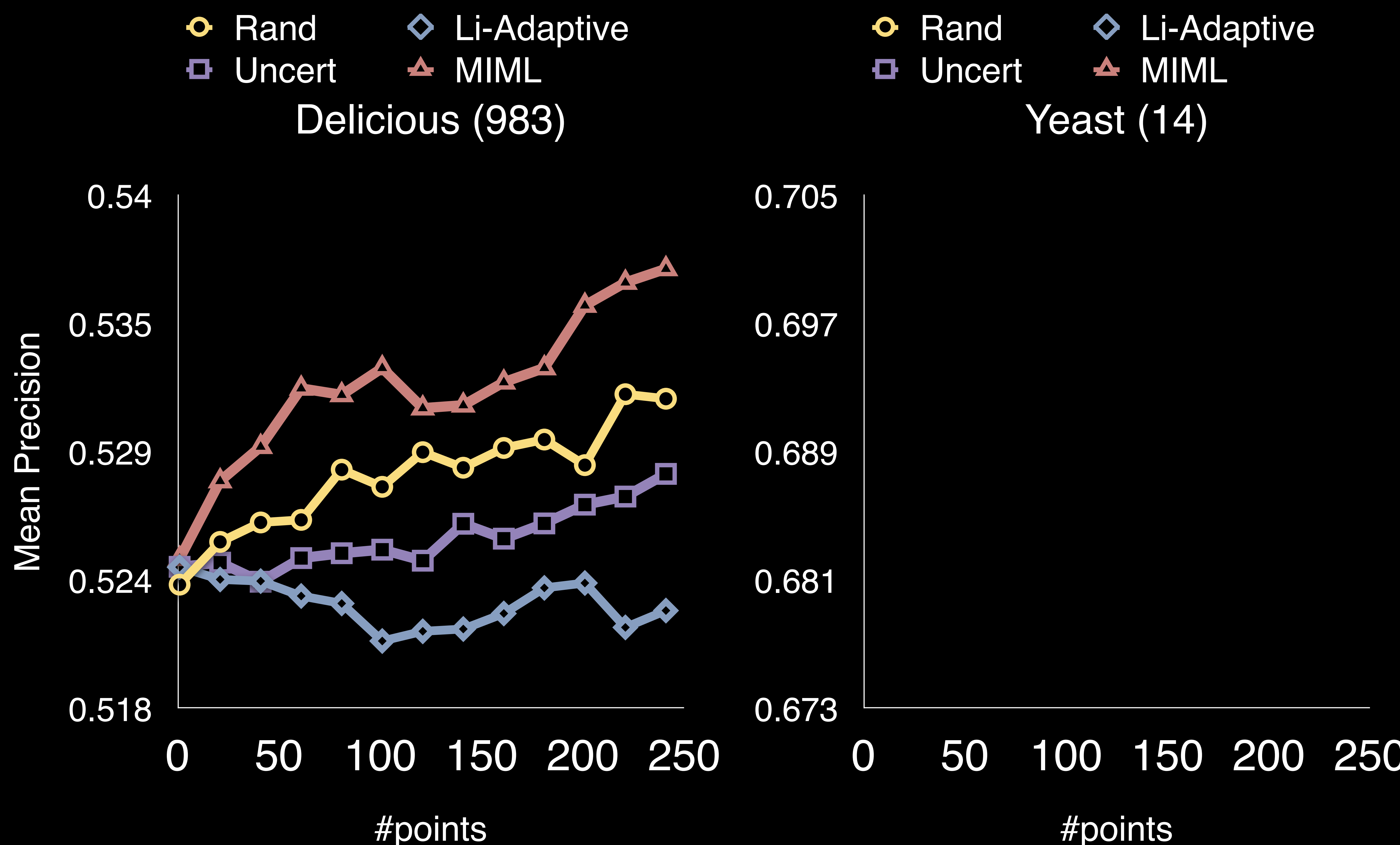


Traditional Active Learning

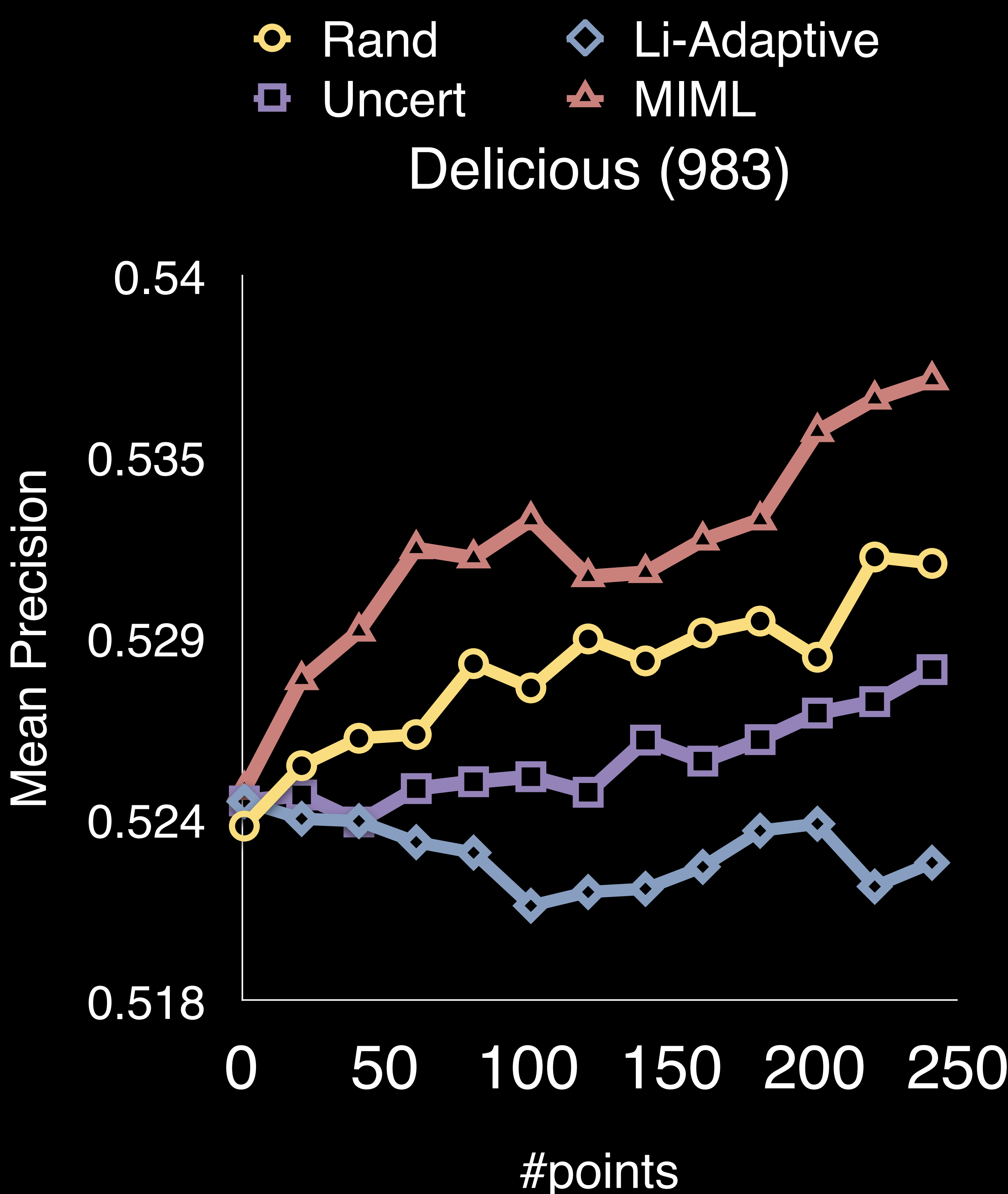
Traditional Active Learning

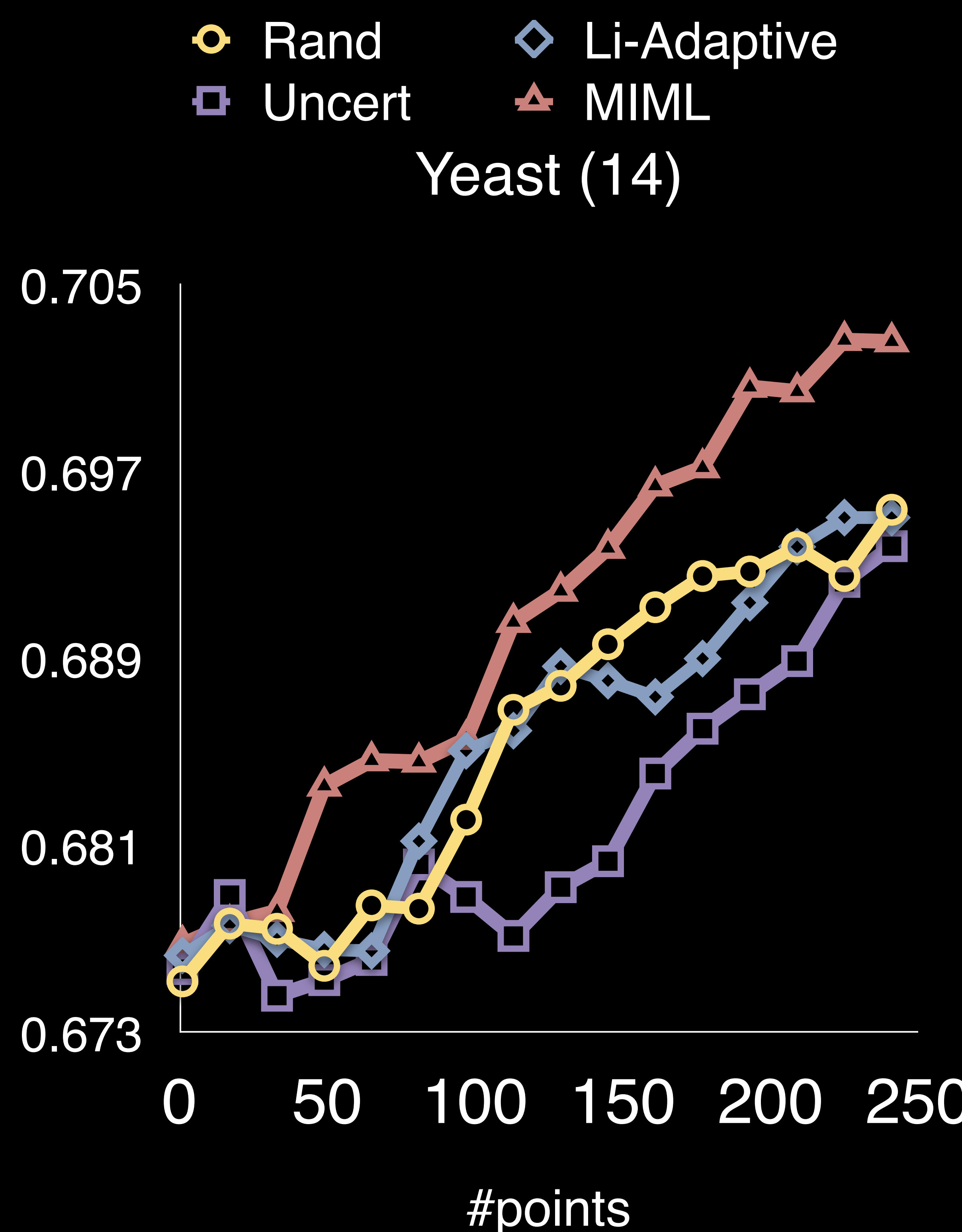


Traditional Active Learning



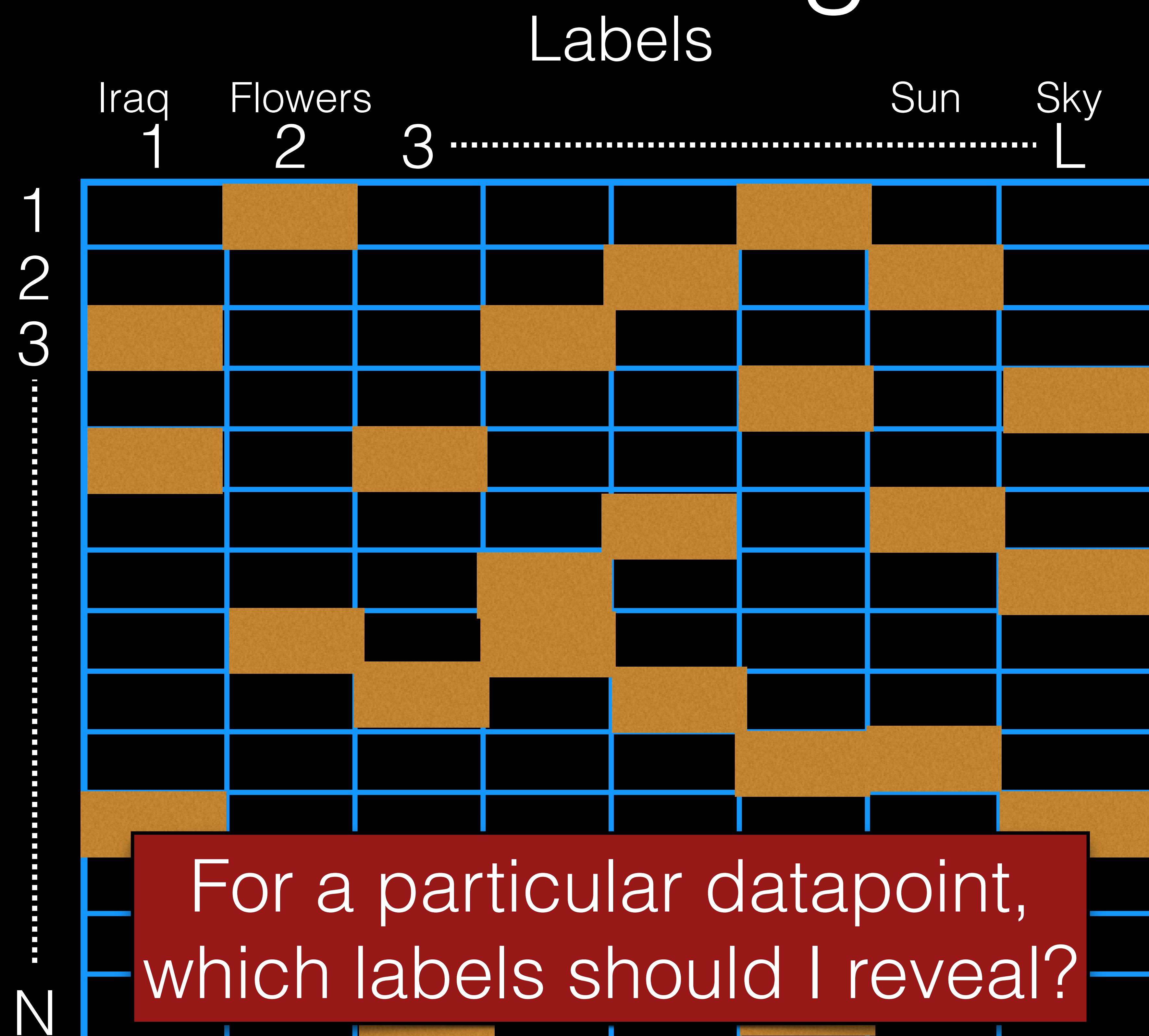
Traditional Active Learning





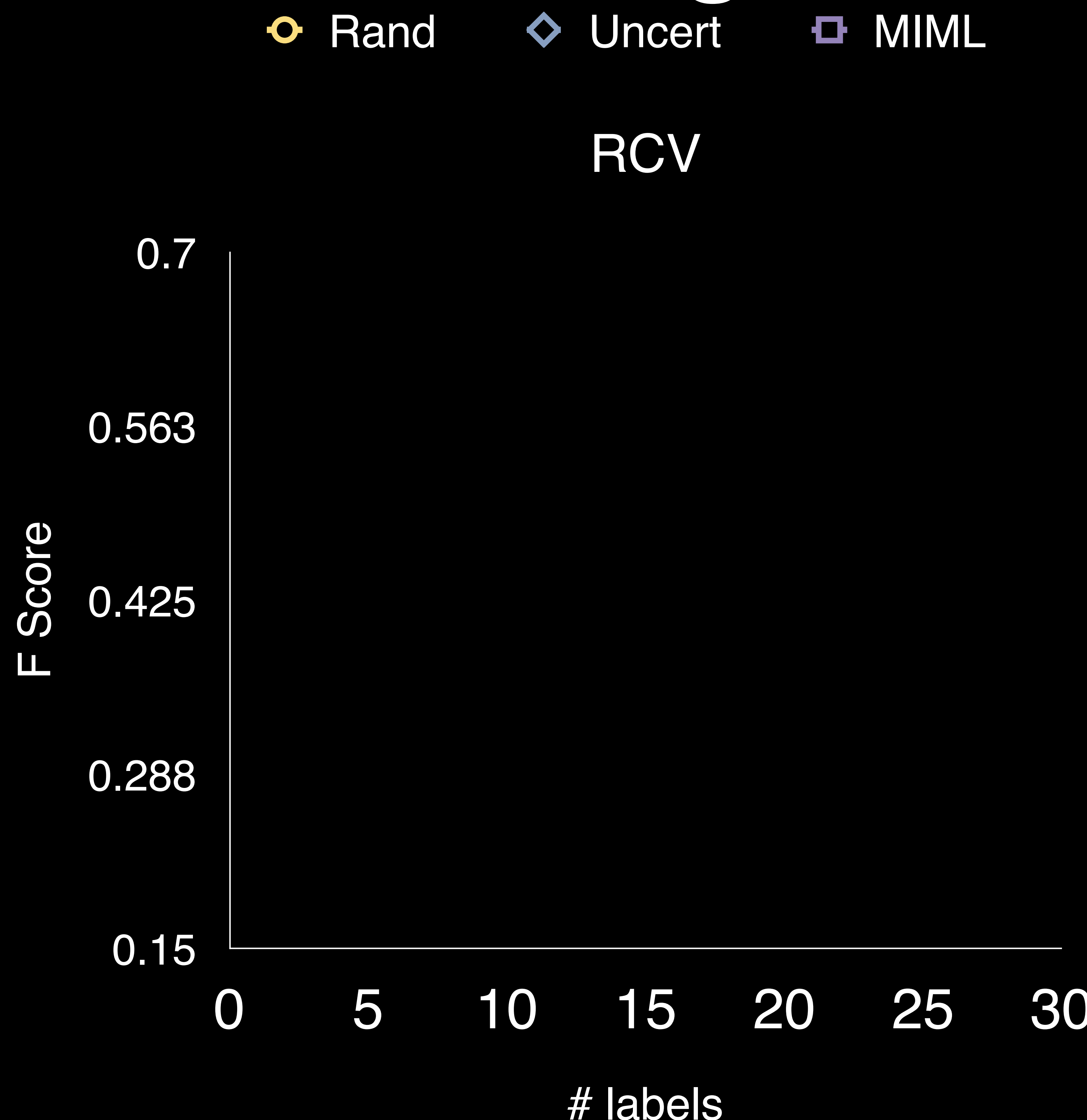
Active Learning

Datapoints

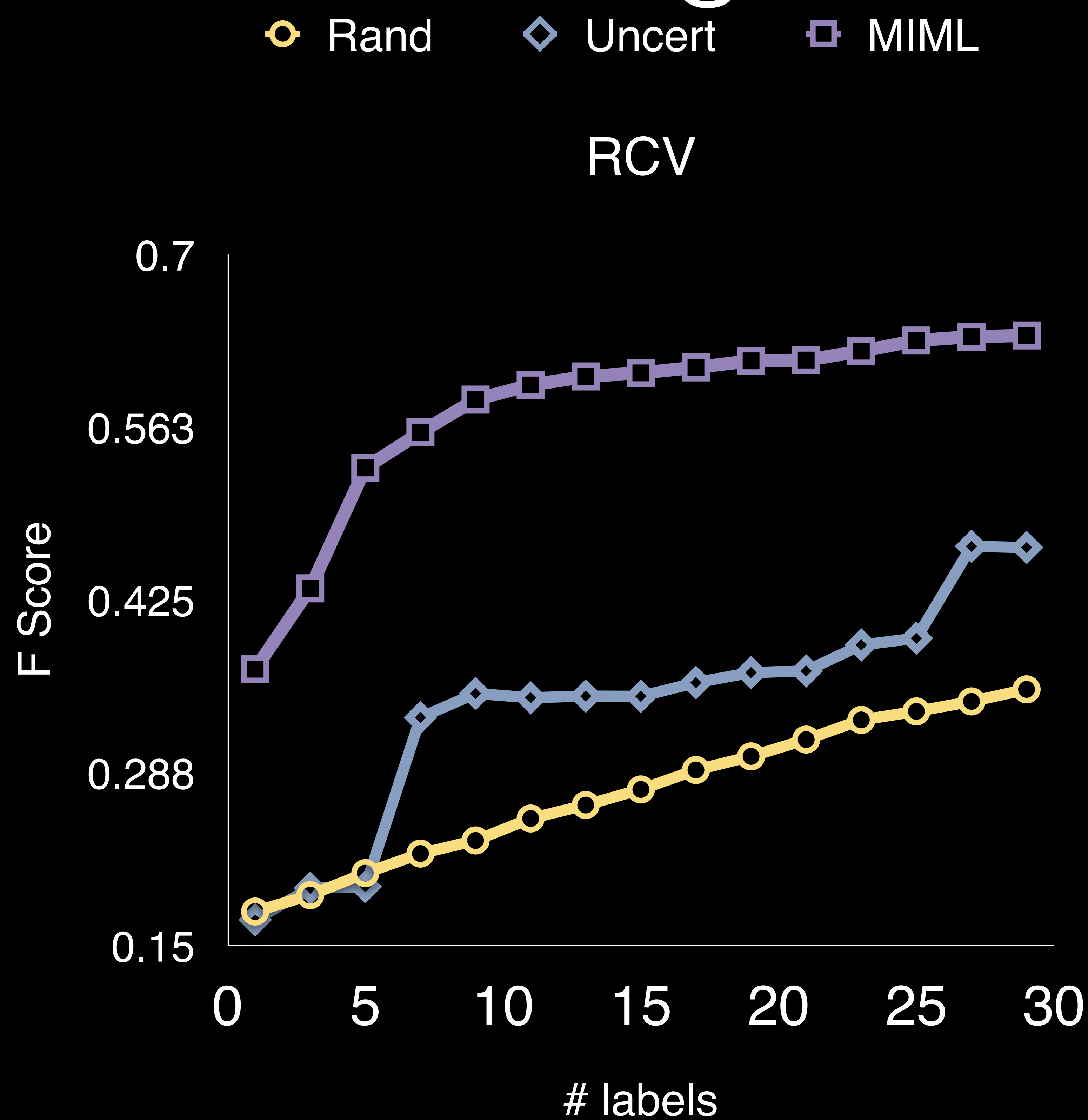


Active Diagnosis

Active Diagnosis

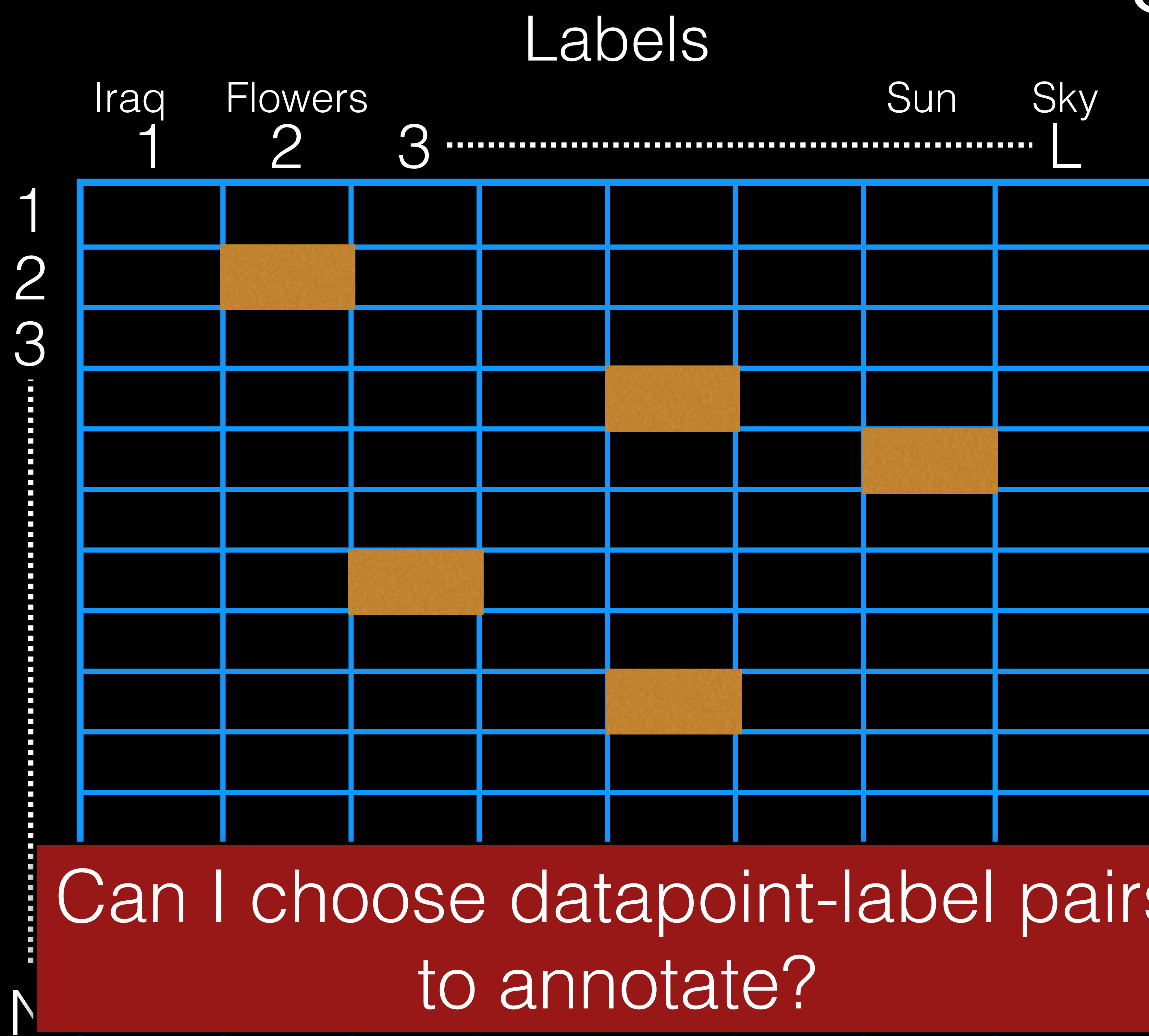


Active Diagnosis



Generalized Active Learning

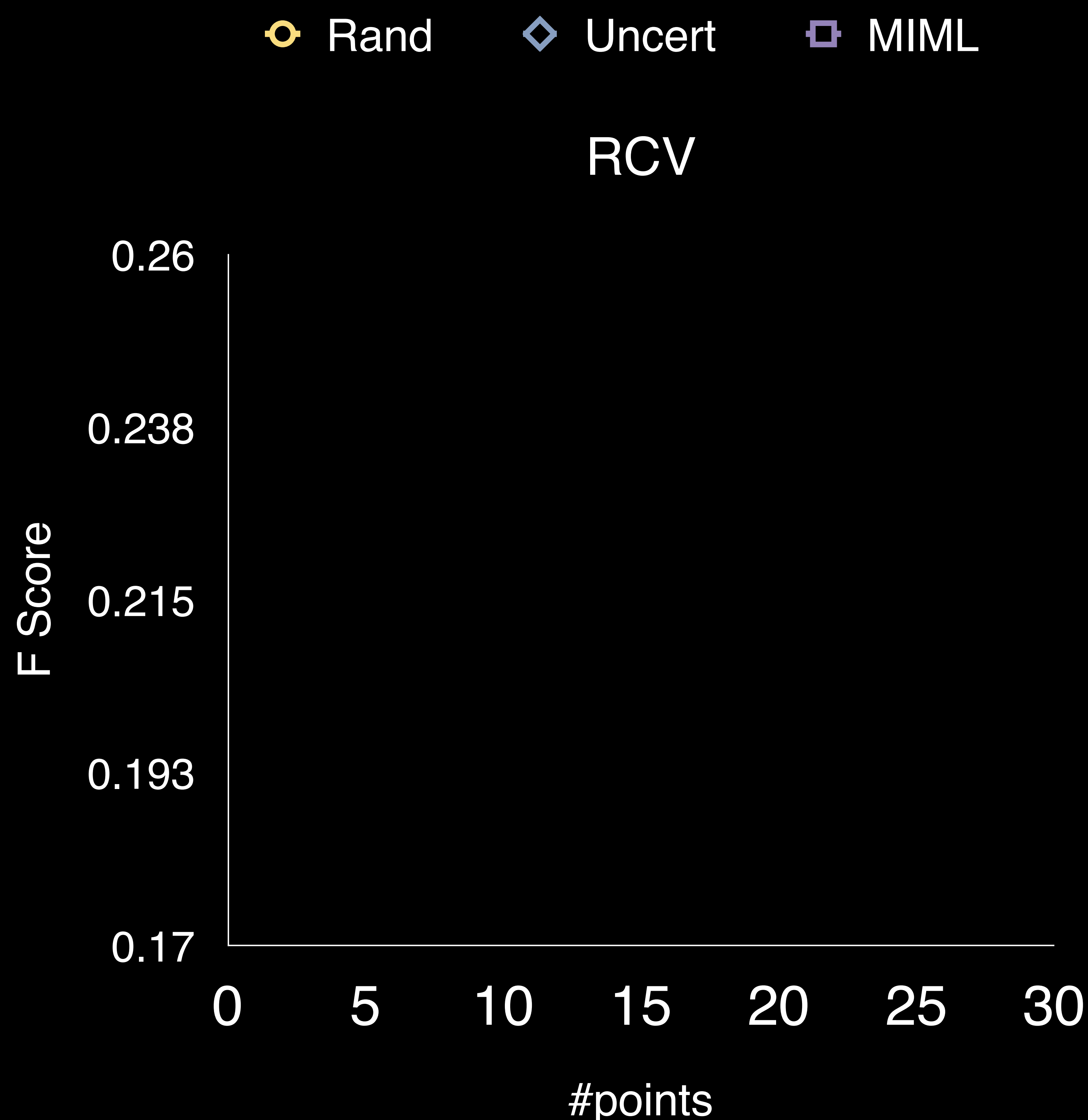
Datapoints



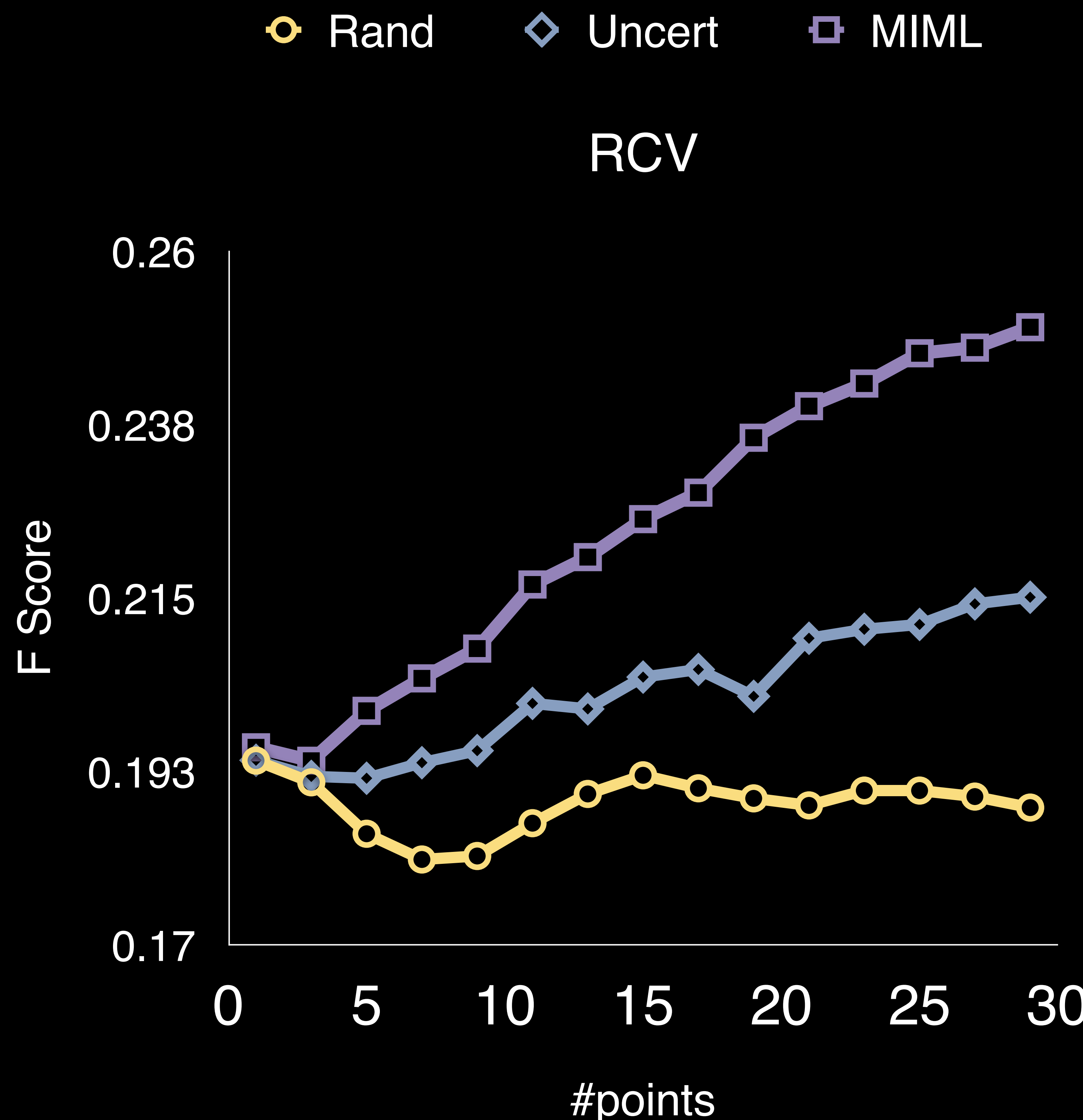
Can I choose datapoint-label pairs to annotate?

Generalized Active Learning

Generalized Active Learning



Generalized Active Learning



Time Complexity

Dataset	Labels	MIML	Li-Adaptive
Yeast	14	3m 25s	1m 54s
Mediamill	101	41m 29s	54m 35s
RCV1	101	30m 45s	37m 35s
Bookmarks	208	48m 58s	3h 57m
Delicious	983	1h 11m	20h 15m

Related Work

- SVM based Active Learning: Li et al [IJCAI, 2013], Yang et al [KDD 2009], Esuli et al [ECIR 2009], Li et al [ICIP 2004], …

- Mutual Information: Krause et al [UAI 2005], Krause et al [JMLR 2008], Singh et al [JAIR 2009], …

Conclusion

- Proposed mutual information based active learning for multi-label classification
- Collapsed Variational Bayes to infer variances
- Theoretical analysis of mutual information approximation showing that it is near-optimal
- Showed significant empirical improvements over the state-of-the-art