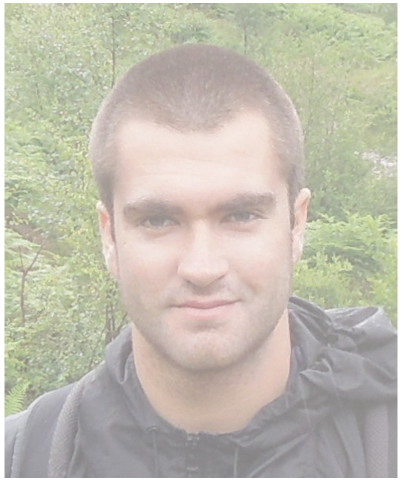
Light-Supervision of Structured Prediction Energy Networks

Andrew McCallum

Pedram Rooshenas
Oregon PhD→UMass Postdoc

Aishwarya Kamath
UMass MS

SPENs
[2016]
David Belanger
UMass PhD→Google Brain



Generalized Expectation
[Mann; Druck 2010-12]
Greg Druck
UMass PhD→Yummly

Light-Supervision

Prior Knowledge as *Generalized Expectation*

…induces extra structural dependencies…

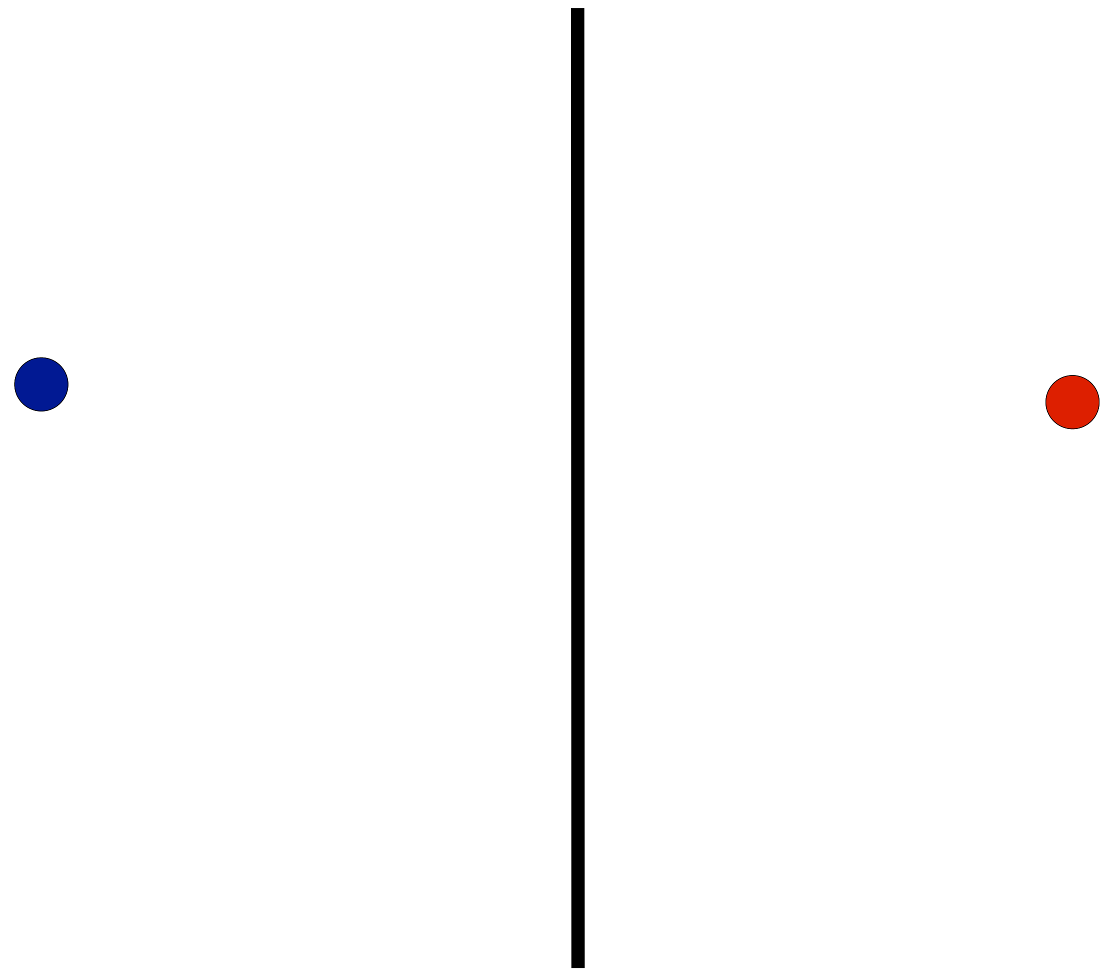
Structured Prediction

Complex dependencies with *SPENs*

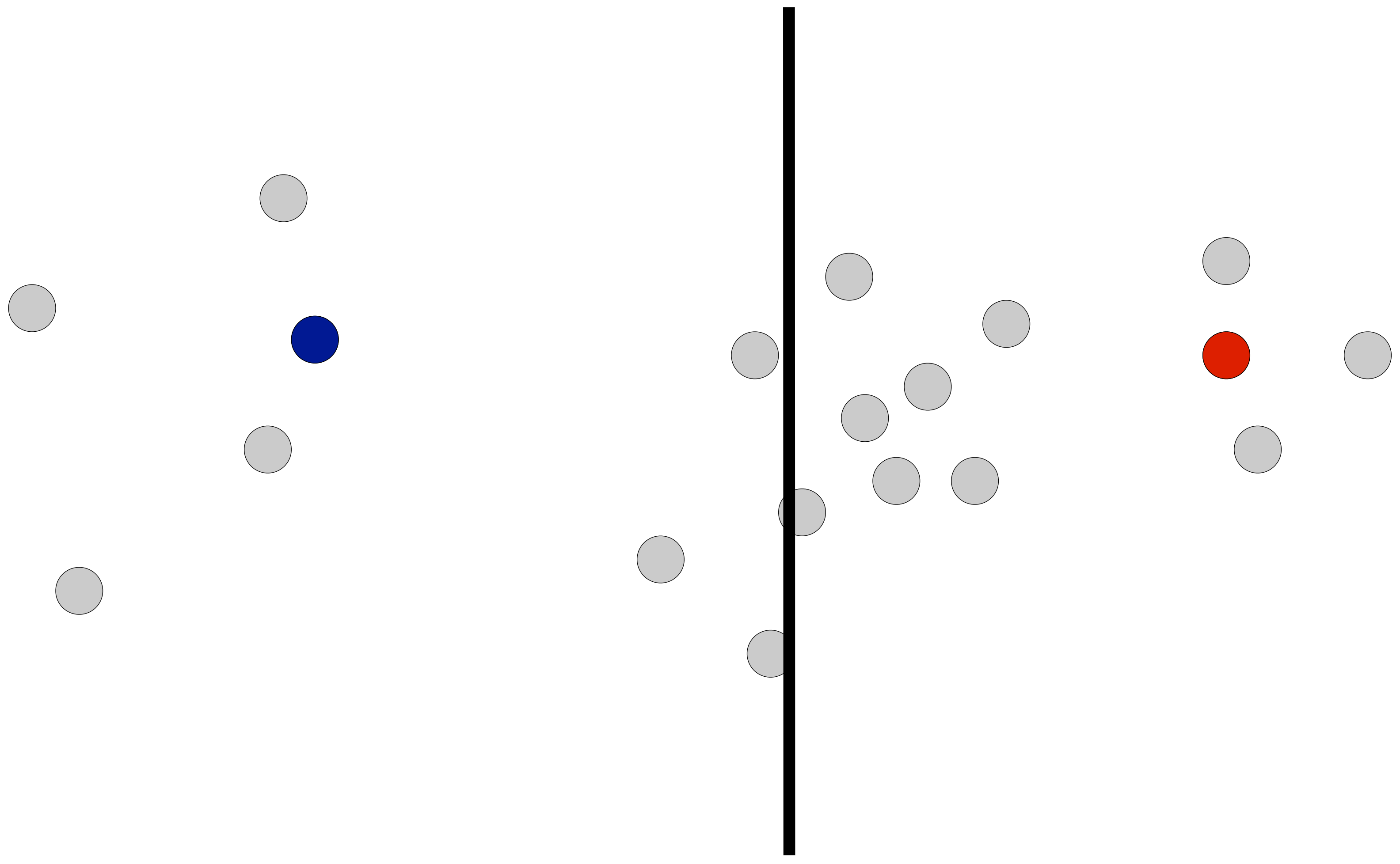
Chapter 1

Generalized Expectation

Learning from small labeled data

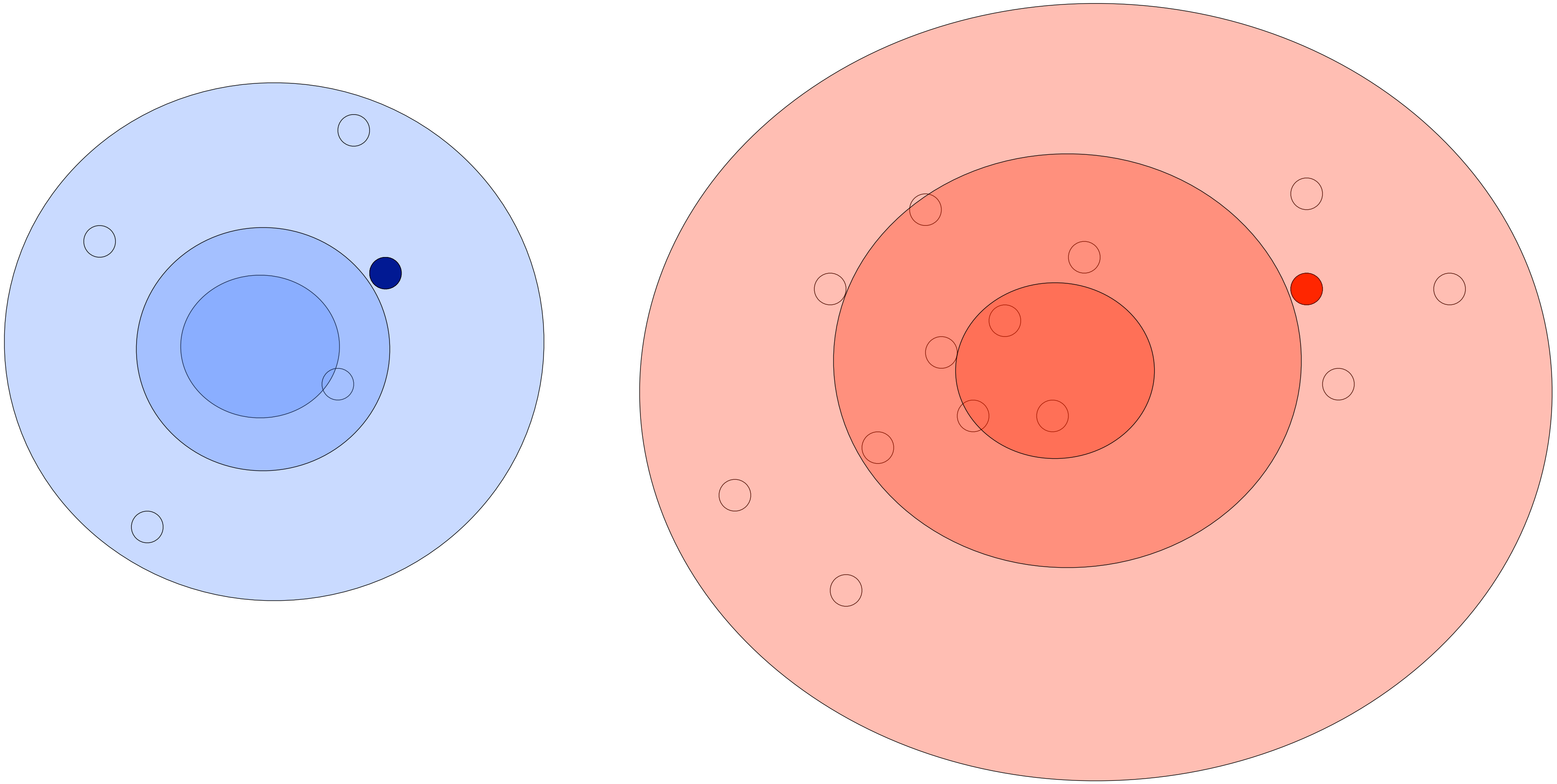


Leverage unlabeled data



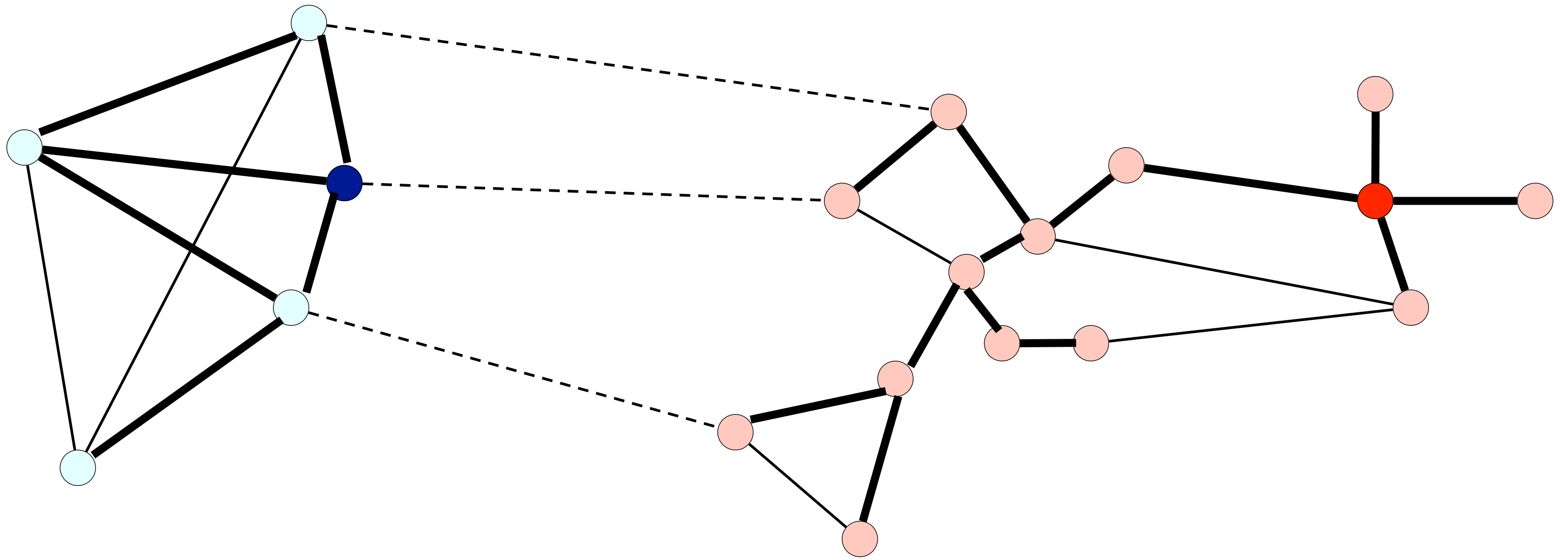
Family 1: Expectation Maximization

[Dempster, Laird, Rubin, 1977]



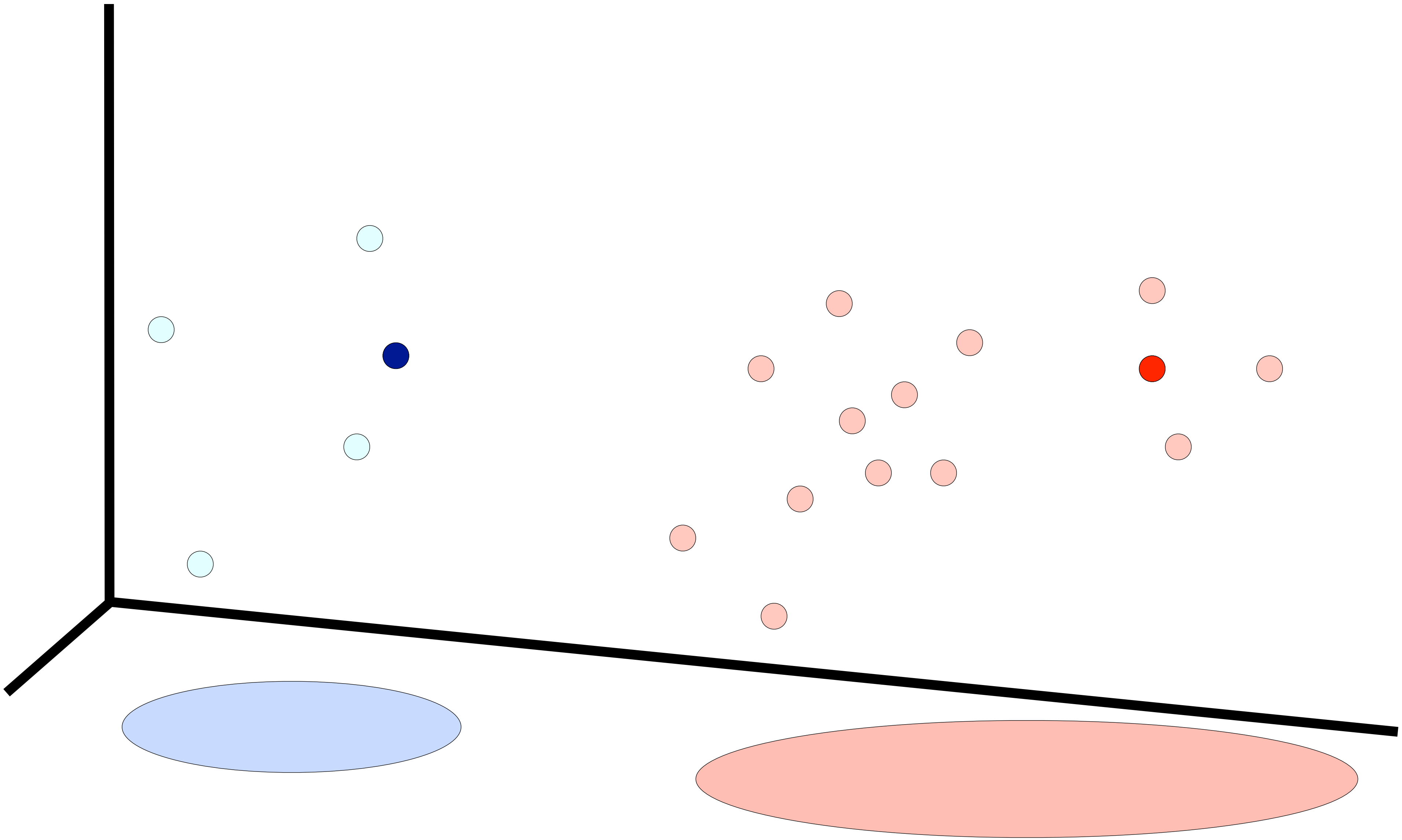
Family 2: Graph-Based Methods

[Szummer, Jaakkola, 2002]   [Zhu, Ghahramani, 2002]



Family 3: Auxiliary-Task Methods

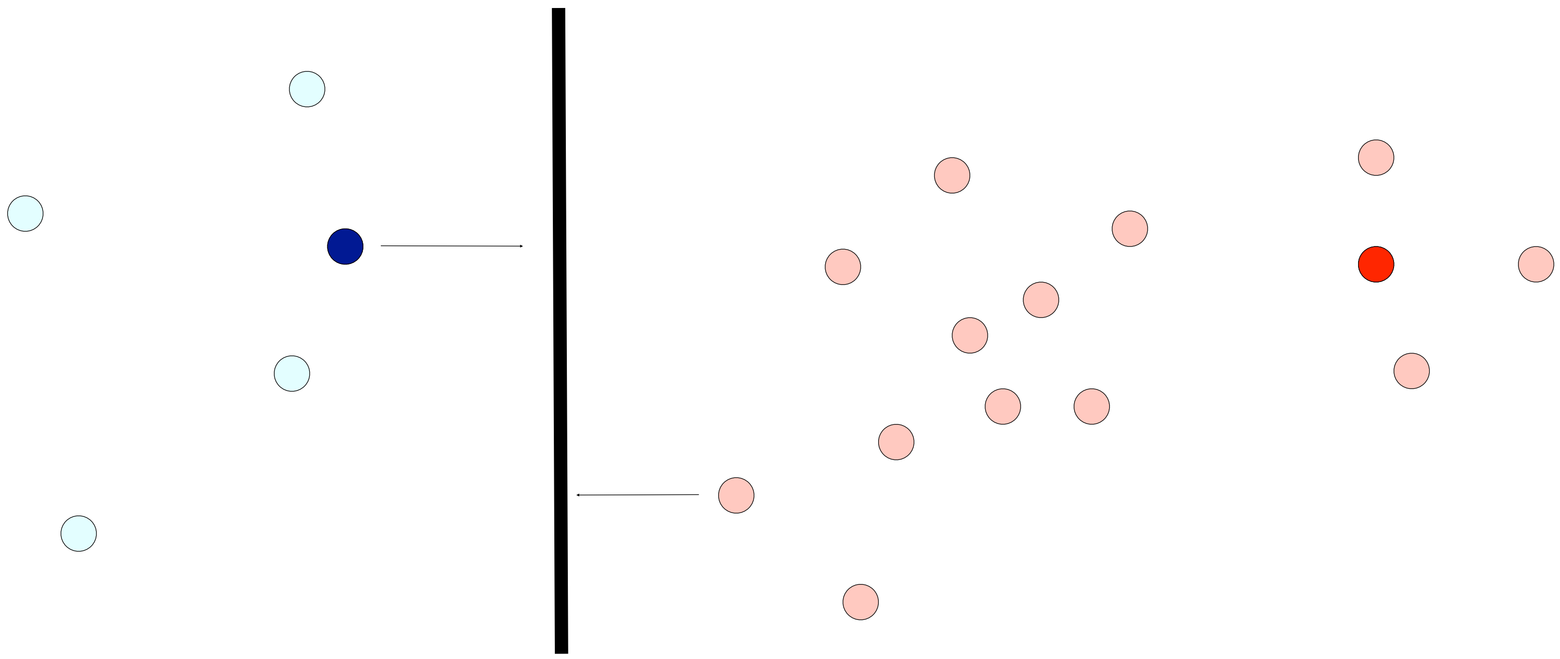
[Ando and Zhang, 2005]



Family 4: Boundary in Sparse Region

Transductive SVMs [Joachims, 1999]: Sparsity measured by margin

Entropy Regularization [Grandvalet & Bengio, 2005]: minimize label entropy



Family 4: Boundary in Sparse Region

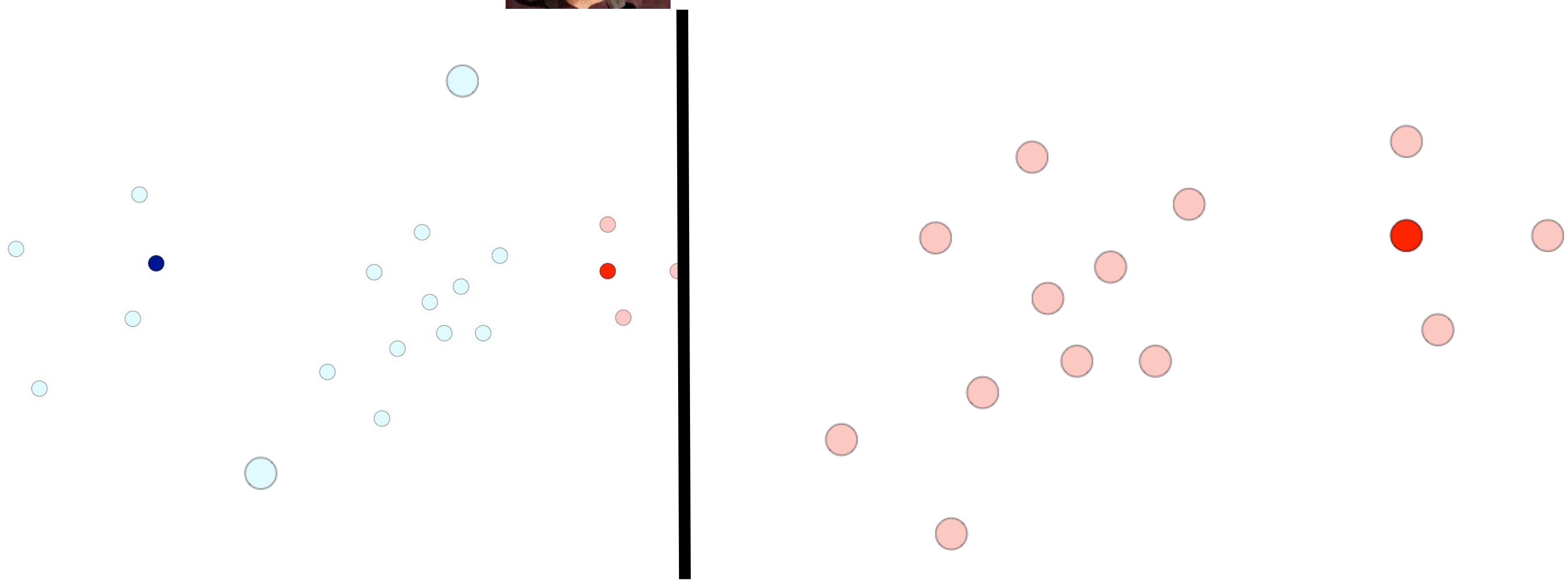
Family 5: Generalized Expectation Criteria

Transductive SVMs [Joachims, 1999]: Sparsity measured by margin

[Mann, McCallum 2010; Druck, Mann, McCallum 2011; Druck McCallum 2012]

Entropy Regularization  lvalet & Bengio, 2005]: minimize label entropy

best solution?

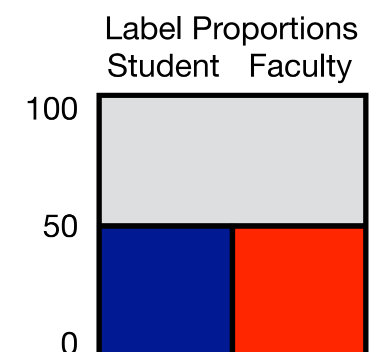


Label | Feature
Expectations

$$E[p(y|f(x))]$$

Label Prior
Expectations

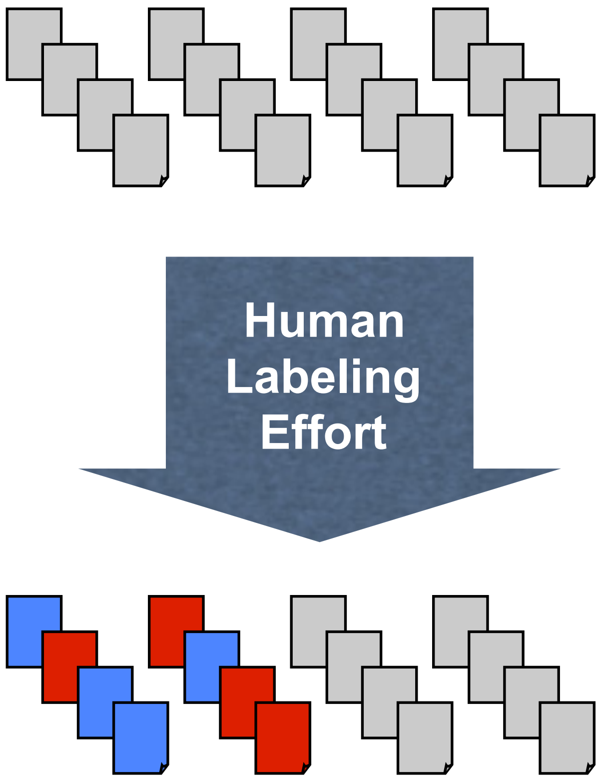
$$E[p(y)]$$



Expectations on Labels | Features

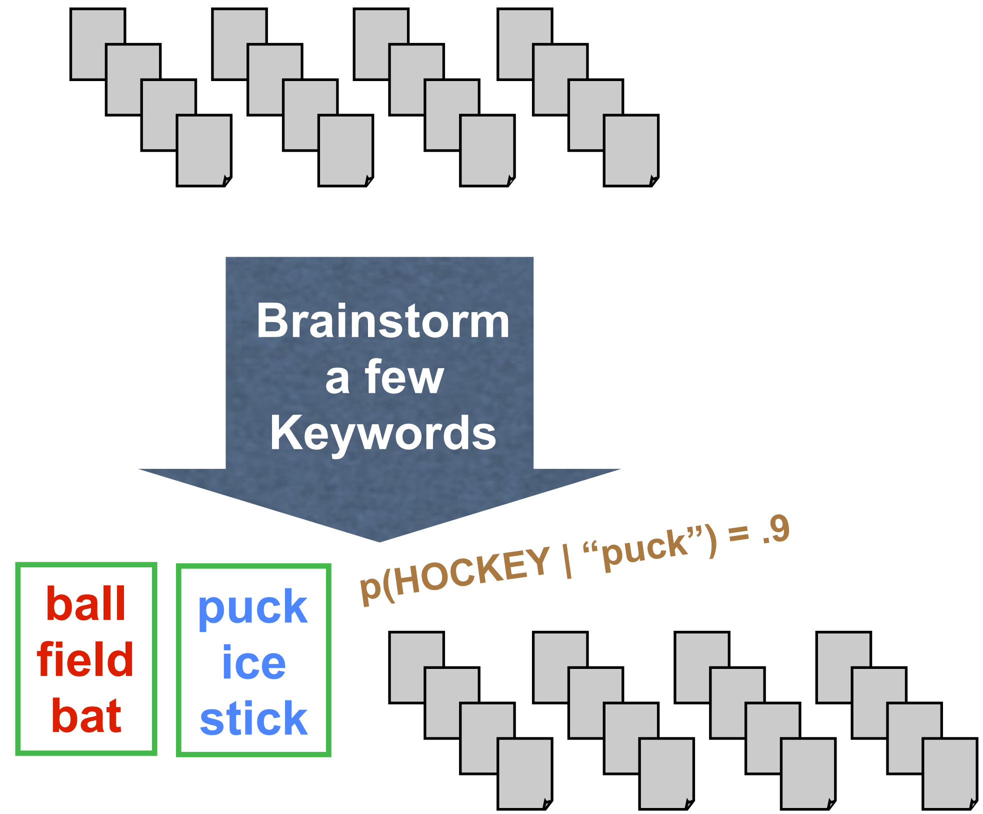
Classifying *Baseball* versus *Hockey*

Traditional



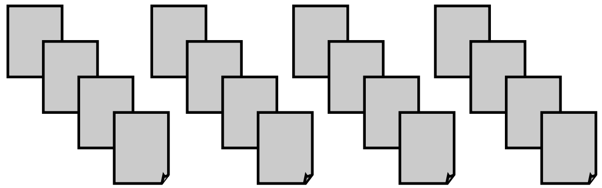
(Semi-)Supervised Training via Maximum Likelihood

Generalized Expectation



Semi-Supervised Training via *Generalized Expectation*

Labeling Features



~1000 unlabeled examples

features labeled . . .

hockey baseball HR Mets	goal Buffalo Leafs puck Lemieux	ball Oilers Sox Pens runs	batting base NHL Bruins Penguins

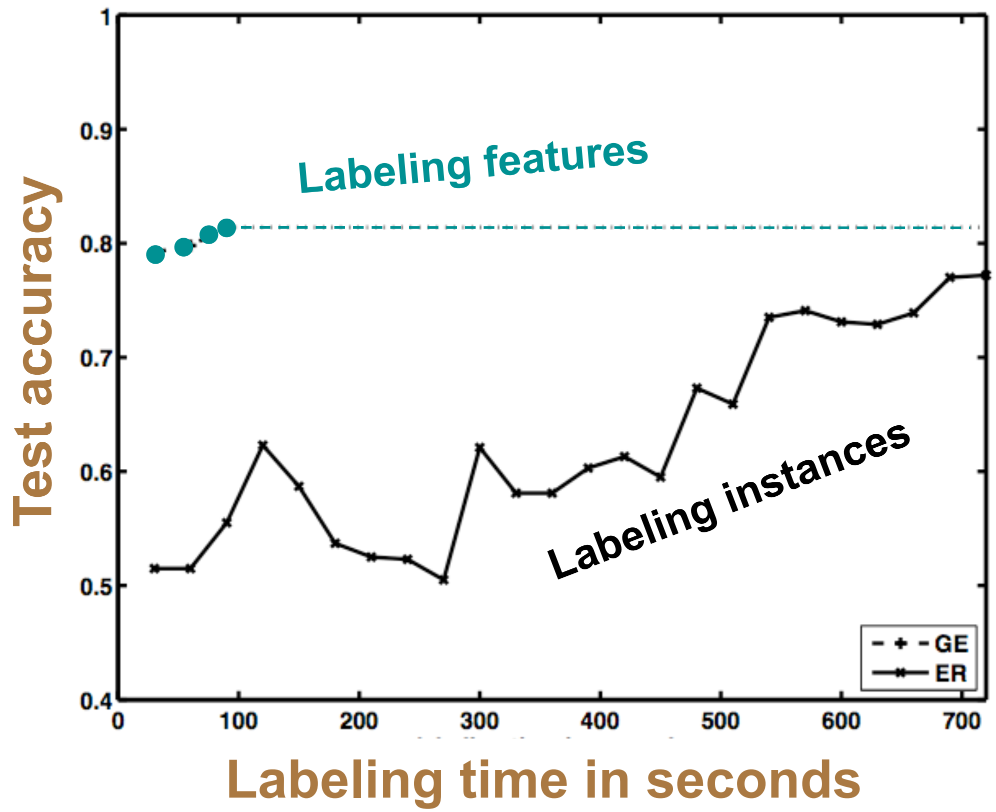
Toronto Maple Leafs

Edmonton Oilers

Pittsburgh Penguins

Accuracy	85%	92%	94.5%	96%

Accuracy per Human Effort



Prior Knowledge

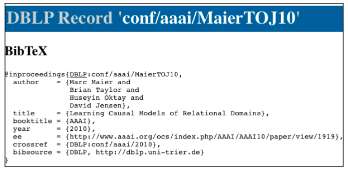
Feature labels from humans

baseball/***hockey*** classification

baseball	**hockey**
hit	puck
braves	goal
runs	nhl

many other sources

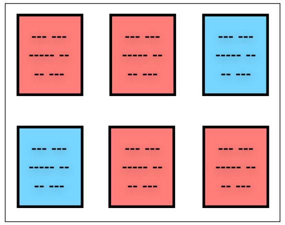
resources on the web


DBLP Record 'conf/aaai/MaierTOJ10'

BibTeX

```
@inproceedings{DBLP:conf/aaai/MaierTOJ10,
  author    = {Marc Maier and
               Brian Taylor and
               Huseyin Oktay and
               David Jensen},
  title     = {Learning Causal Models of Relational Domains},
  booktitle = {AAAI},
  year      = {2010},
  ee        = {http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1919},
  crossref  = {DBLP:conf/aaai/2010},
  bibsource = {DBLP, http://dblp.uni-trier.de}
}
```

WIKIPEDIA
The Free Encyclopedia



data from related tasks

W. H. Enright. Improving the efficiency of matrix operations in the numerical solution of stiff ordinary differential equations. ACM Trans. Math. Softw., 4(2), 127-136, June 1978.

Generalized Expectation (GE)

input variables

output variables

$$\mathbf{g}(\mathbf{x}, \mathbf{y})$$

constraint features

returns 1 if **x** contains “hit” and **y** is **baseball**

Generalized Expectation (GE)

assume general CRF [Lafferty et al. 01]

$$p(\mathbf{y}|\mathbf{x};\theta) = \frac{1}{Z_{\theta,\mathbf{x}}} \exp\left(\theta^\top \mathbf{f}(\mathbf{x},\mathbf{y})\right)$$

$$\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]$$

model distribution

model features

model probability of **baseball** if **x** contains "hit"

Generalized Expectation (GE)

$$\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]]$$

↑

empirical distribution

(can be defined as)
model's probability that
documents that contain "hit" are labeled **baseball**

Generalized Expectation (GE)

(soft) expectation constraint

$$S(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]])$$

score function

larger score if model expectation matches prior knowledge

Generalized Expectation (GE)

Objective Function

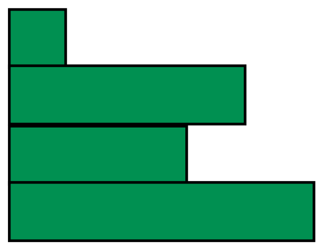
$$\mathcal{O}(\theta) = S(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x}, \mathbf{y})]]) + r(\theta)$$

regularization

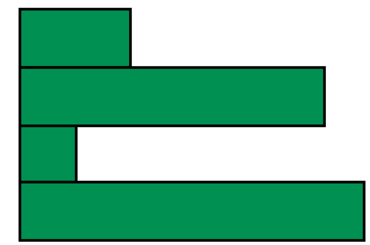
GE Score Functions

$$\mathcal{O}(\theta) = S(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]]) + r(\theta)$$

target expectations

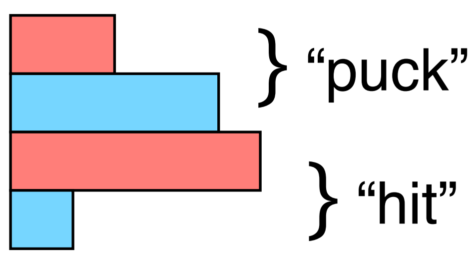
$\hat{\mathbf{g}} =$ 

model expectations

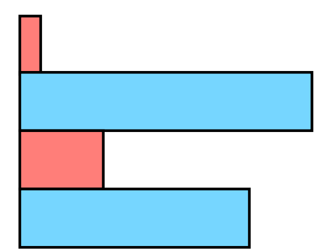
$\mathbf{g}_\theta =$ 

squared error: $S_{l_2^2}(\theta) = -\|\ \hat{\mathbf{g}} - \mathbf{g}_\theta\ \|_2^2$

target expectations

$\hat{\mathbf{g}} =$ 

model expectations

$\mathbf{g}_\theta =$ 

KL divergence: $S_{KL}(\theta) = -\sum_q \hat{\mathbf{g}}_q \log \frac{\hat{\mathbf{g}}_q}{\mathbf{g}_{\theta,q}}$

Estimating Parameters with GE

$$\mathcal{O}(\theta) = S(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]]) + r(\theta)$$

violation term:

KL: $v_i = \frac{\hat{g}_i}{g_{\theta i}}$

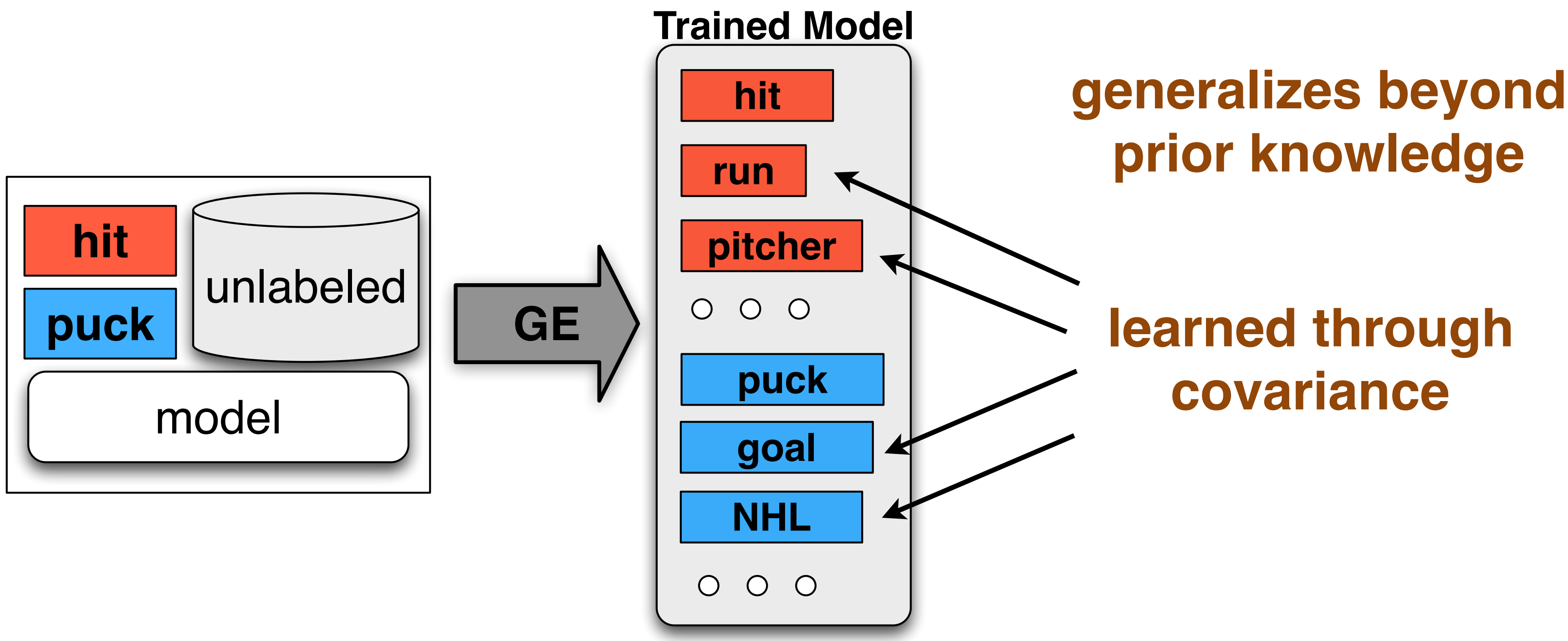
sq. error: $v_i = -2(\hat{g}_i - g_{\theta i})$

$$\nabla_\theta \mathcal{O}(\theta) = \mathbf{v}^\top \Big(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})\mathbf{f}(\mathbf{x},\mathbf{y})^\top] - \mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{f}(\mathbf{x},\mathbf{y})^\top]] \Big) + \nabla_\theta r(\theta)$$

violation

estimated covariance between model and constraint features

Learning About Unconstrained Features



Generalized Expectation criteria

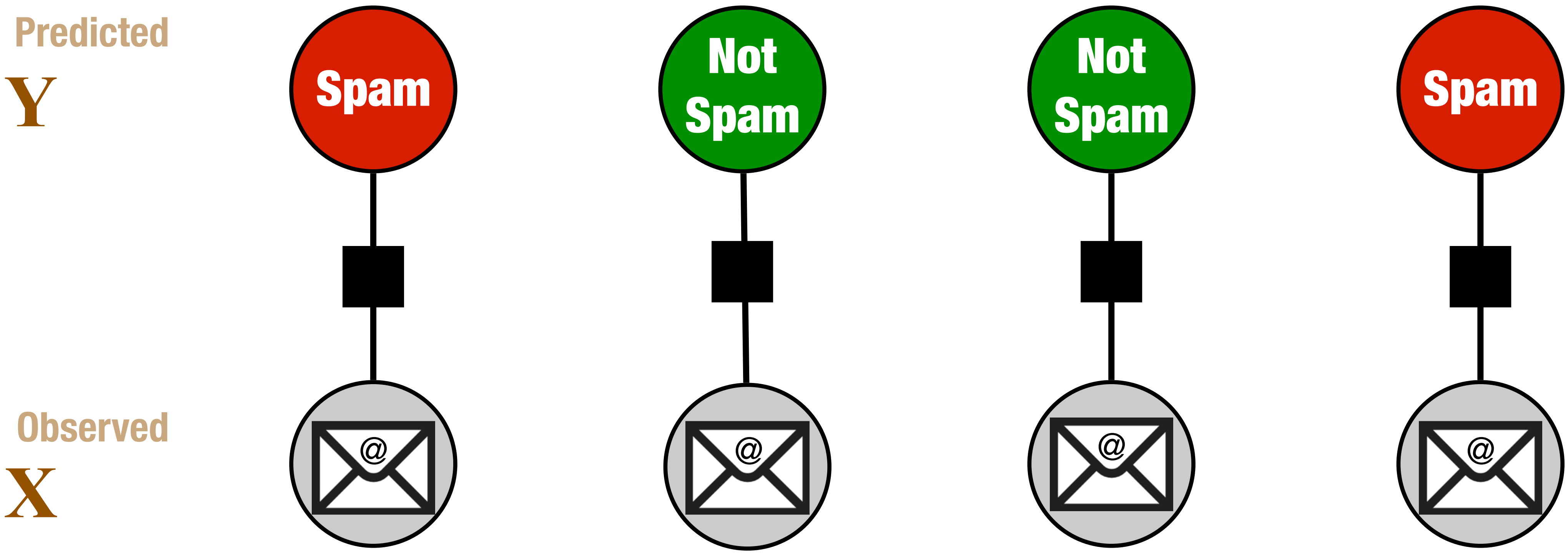
Easy communication with domain experts

- Inject domain knowledge into parameter estimation
- Like “informative prior”...
- ...but rather than the “language of parameters”
(difficult for humans to understand)
- ...use the “language of expectations”
(natural for humans)

IID Prediction

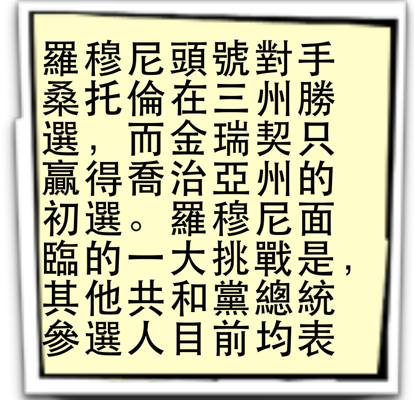
“classification” *e.g. logistic regression*

Example: Spam Filtering



Structured Prediction

e.g. "sequence labeling" Chinese Word Segmentation



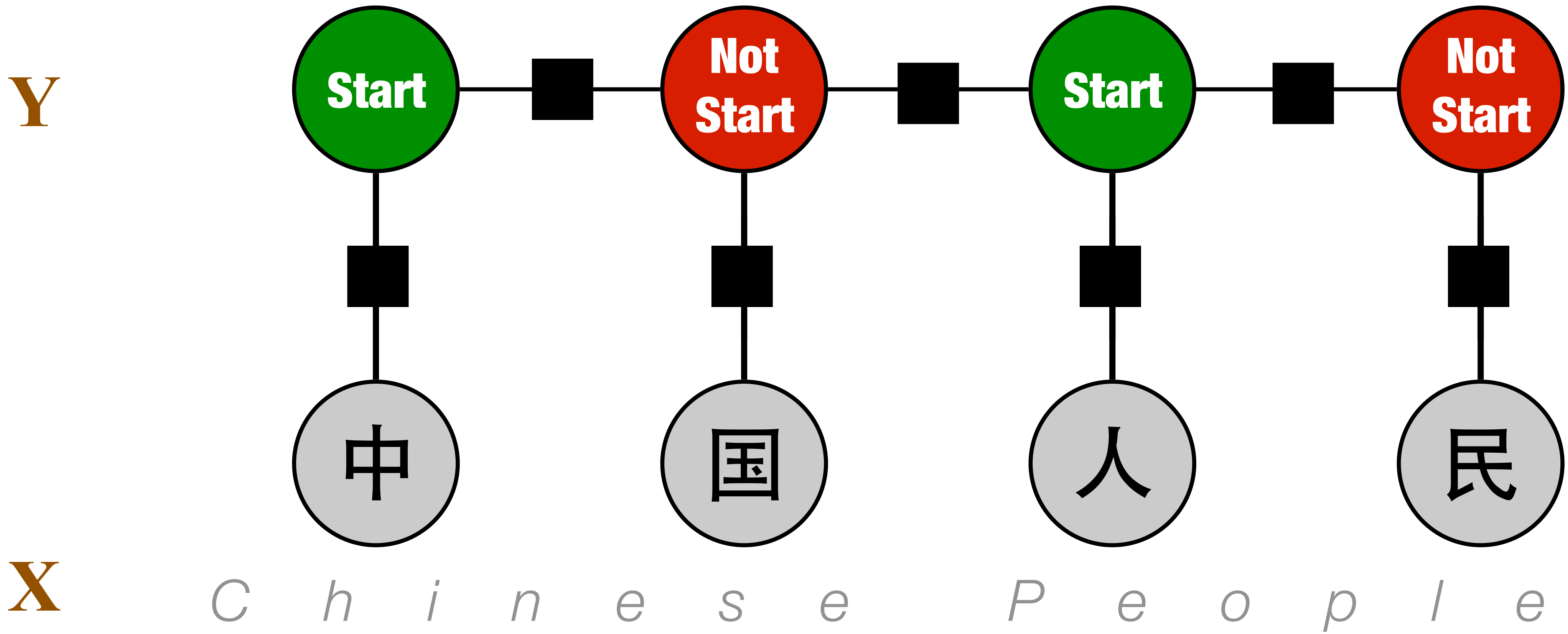
$$\mathcal{O}(\theta) = S(\mathrm{E}_{\tilde{p}(\mathbf{x})}[\mathrm{E}_{p(\mathbf{y}|\mathbf{x};\theta)}[\mathbf{g}(\mathbf{x},\mathbf{y})]]) + r(\theta)$$

Linear-chain CRF

GE Gradient

$$\mathbf{v}^\top \sum_{\mathbf{y}} \sum_i \sum_j p(y_{i-1}, y_i, y_j|\mathbf{x};\theta)\mathbf{g}(\mathbf{x}, y_j, j)\mathbf{f}(\mathbf{x}, y_{i-1}, y_i, i)^\top$$

marginal over three, non-consecutive positions



Natural Expectations lead to Difficult Training Inference

"AUTHOR field should be contiguous, only appearing once."

AUTHOR AUTHOR
Anna Popescu (2004), "Interactive Clustering,"

EDITOR EDITOR
Wei Li (Ed.), Learning Handbook, Athos Press,

LOCATION
Souroti.

$p(y_{i-1}, y_i, y_j, y_k)$ **The downfall of GE.**

Chapter 2

A framework providing easier inference for complex dependencies?

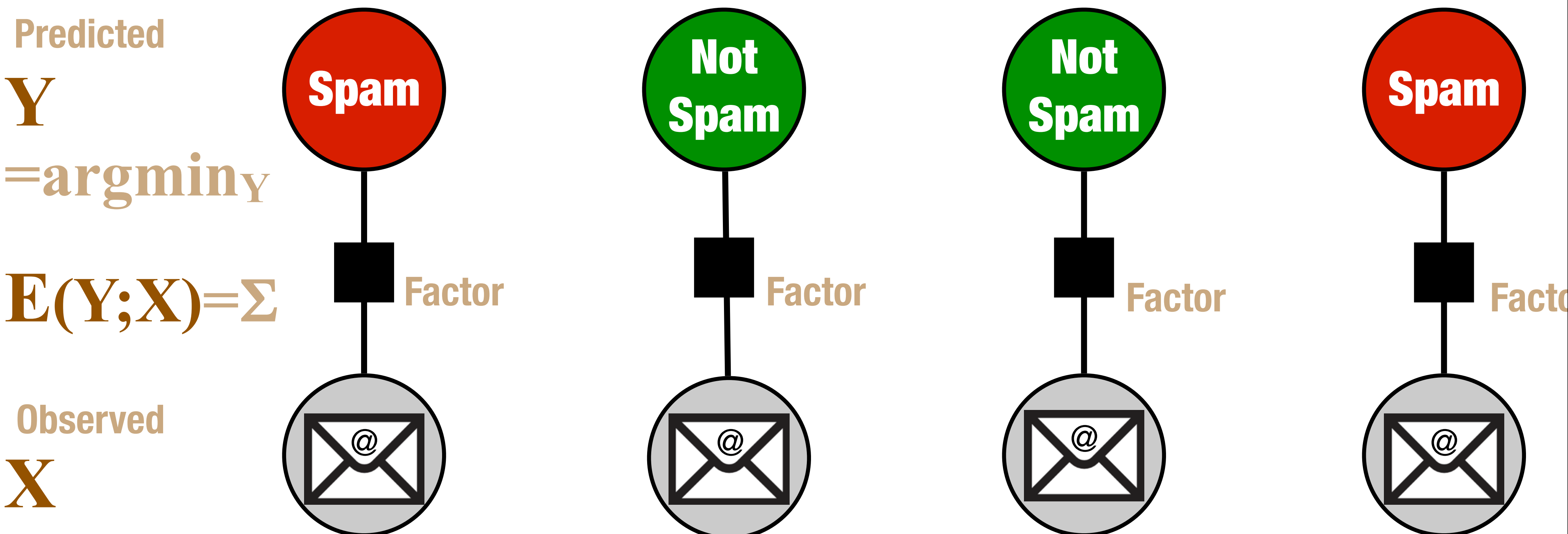
Structured Prediction Energy Networks

Deep Learning
+
Structured Prediction

Structured **Prediction**

"classification" *e.g. logistic regression*

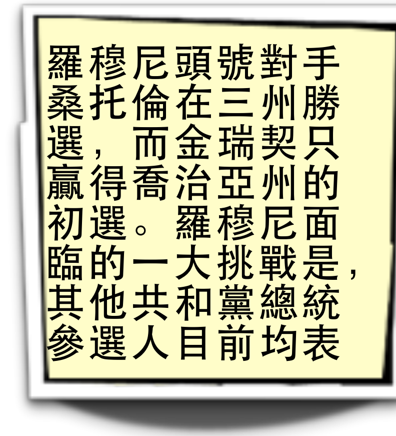
Example: Spam Filtering

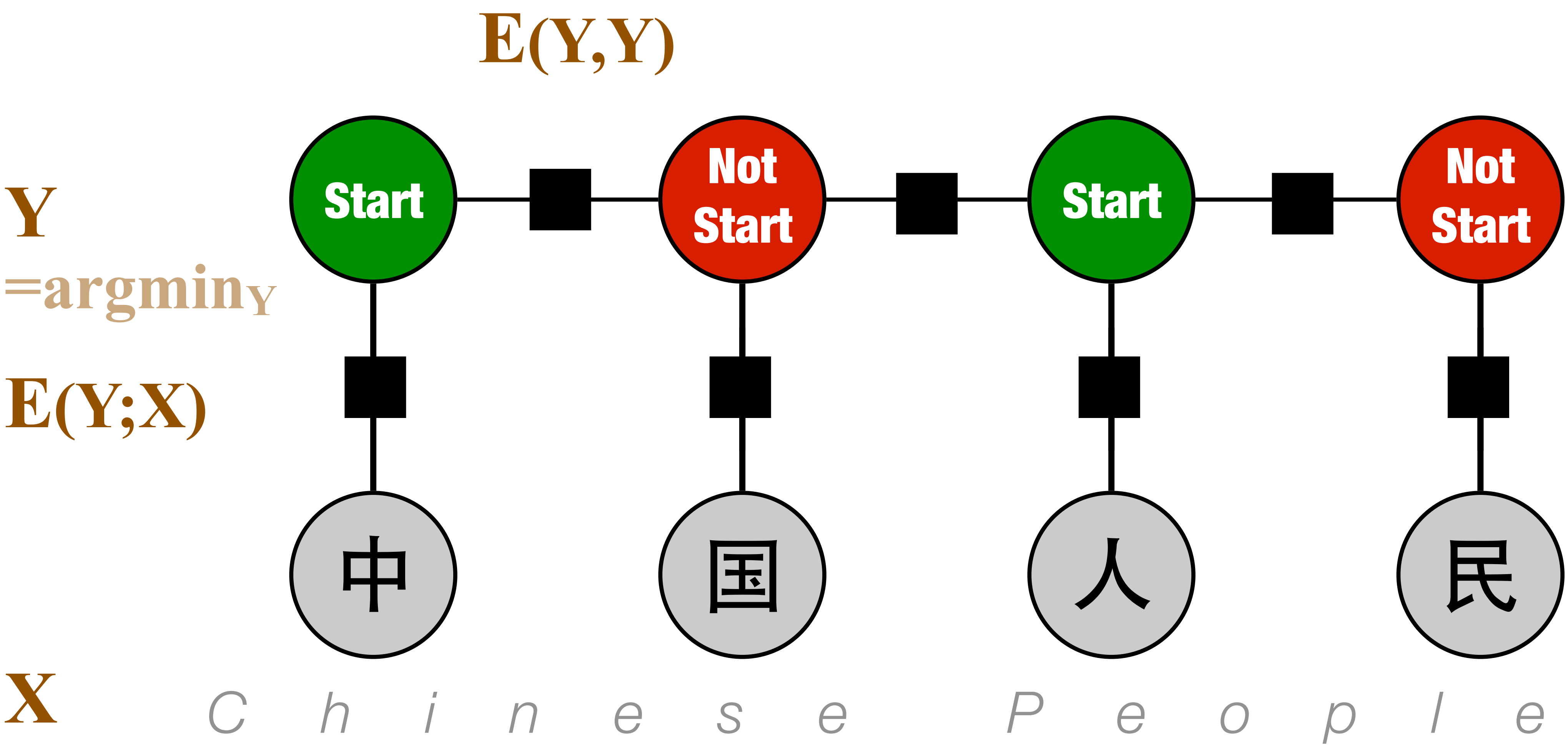


Structured **Prediction**

e.g. "sequence labeling"

Example: Chinese Word Segmentation


羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表

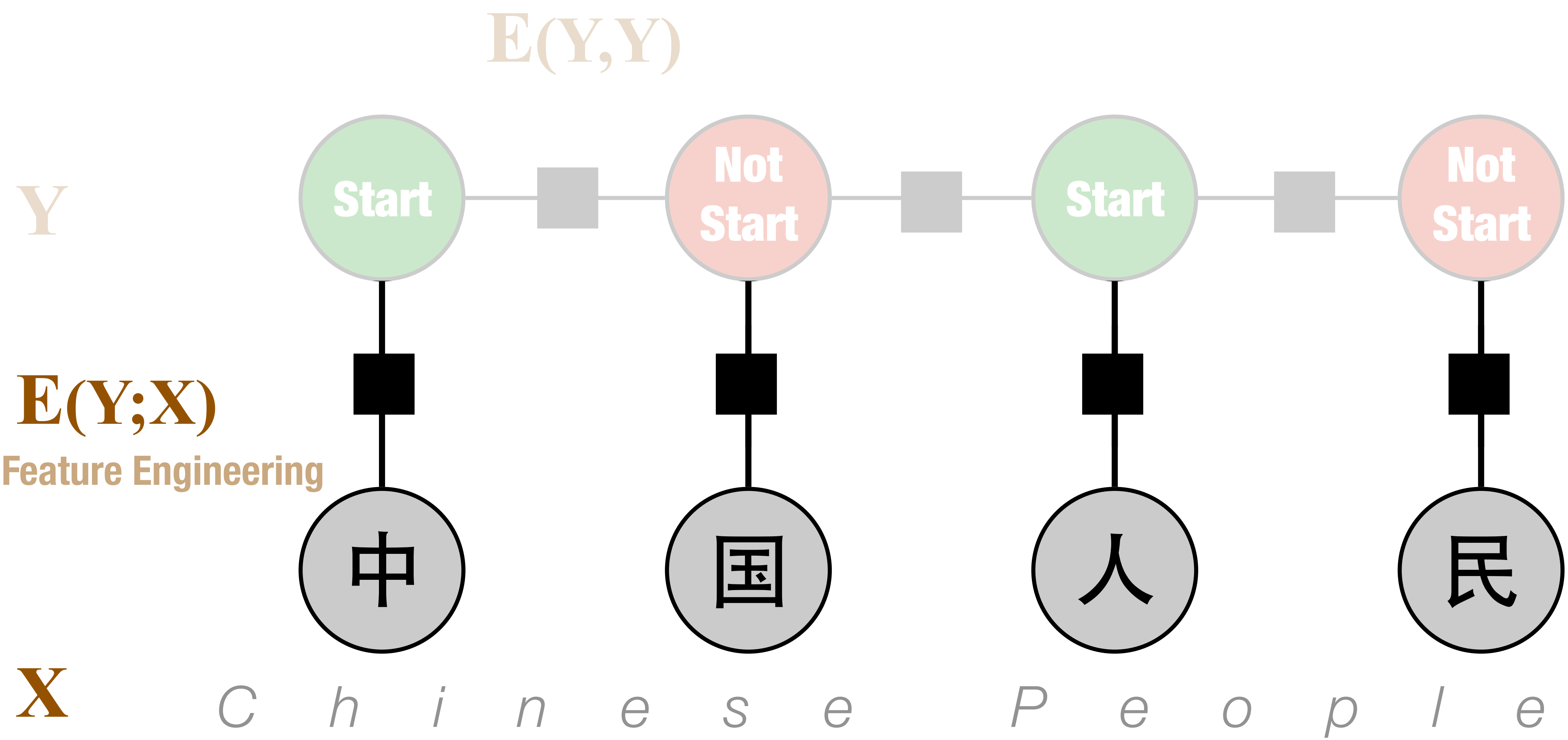


Structured Prediction

e.g. "sequence labeling"

Example: Chinese Word Segmentation

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表

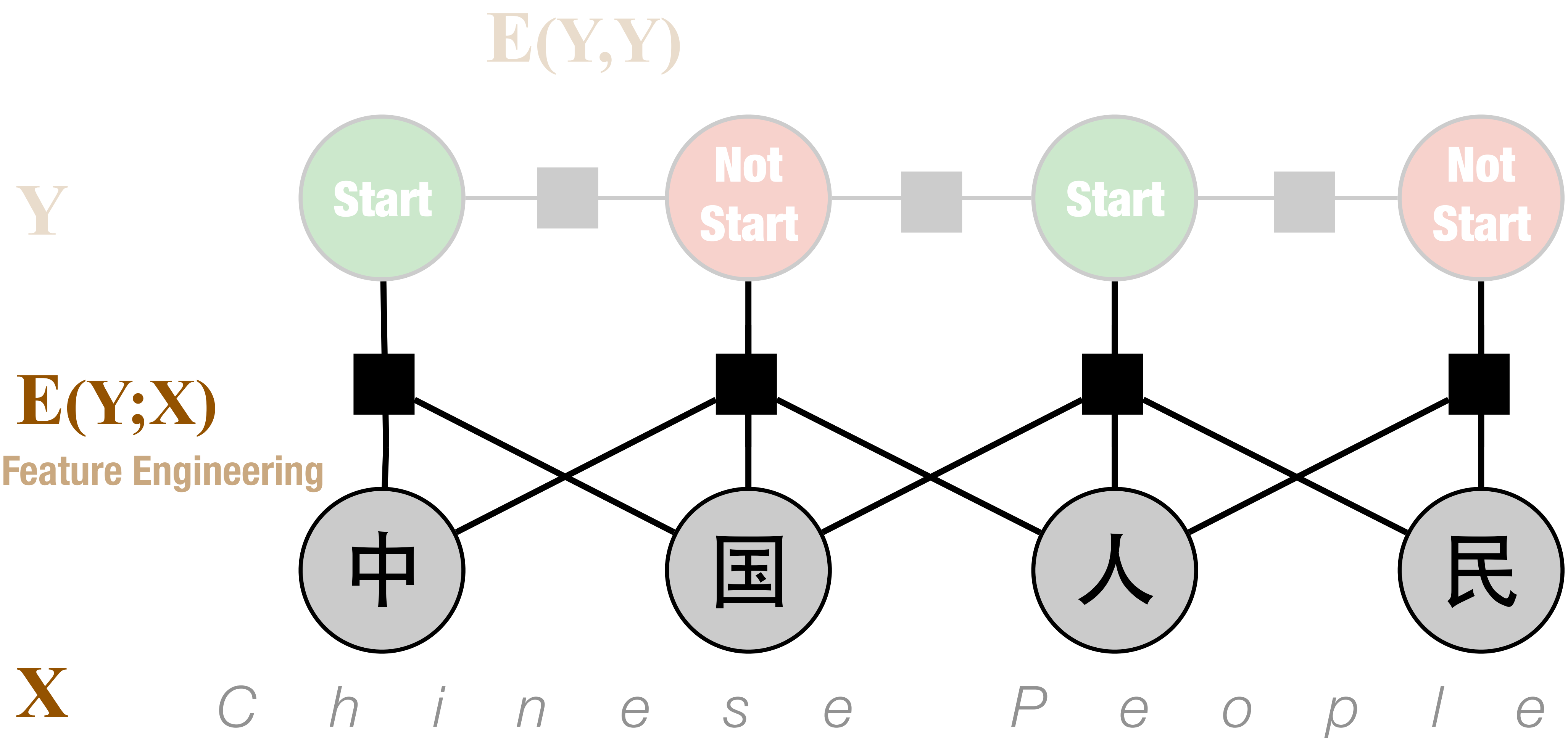


Structured Prediction

e.g. “sequence labeling”

Example: Chinese Word Segmentation

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表



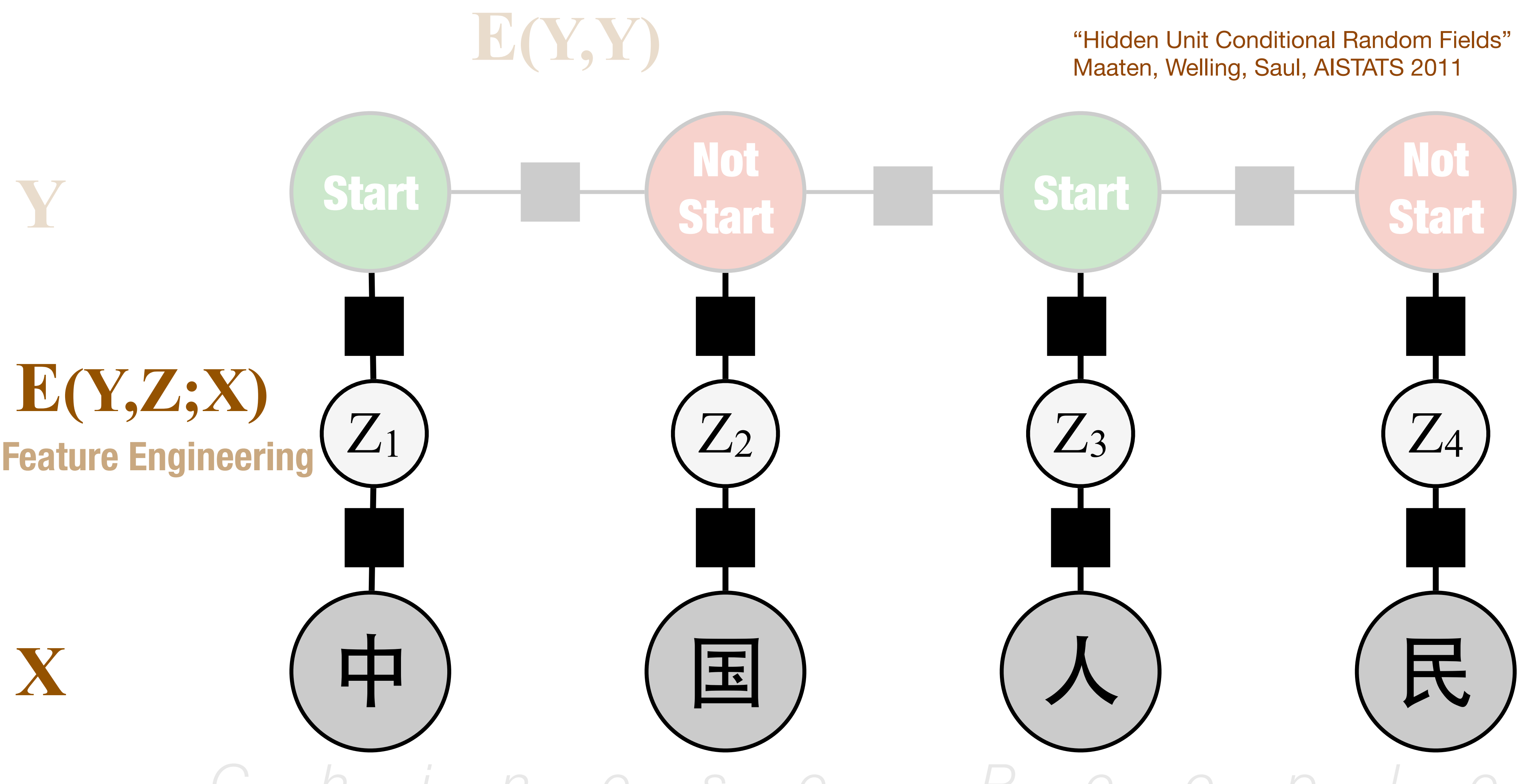
Structured Prediction

e.g. "sequence labeling"

Example: Chinese Word Segmentation

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表

"Hidden Unit Conditional Random Fields"
Maaten, Welling, Saul, AISTATS 2011

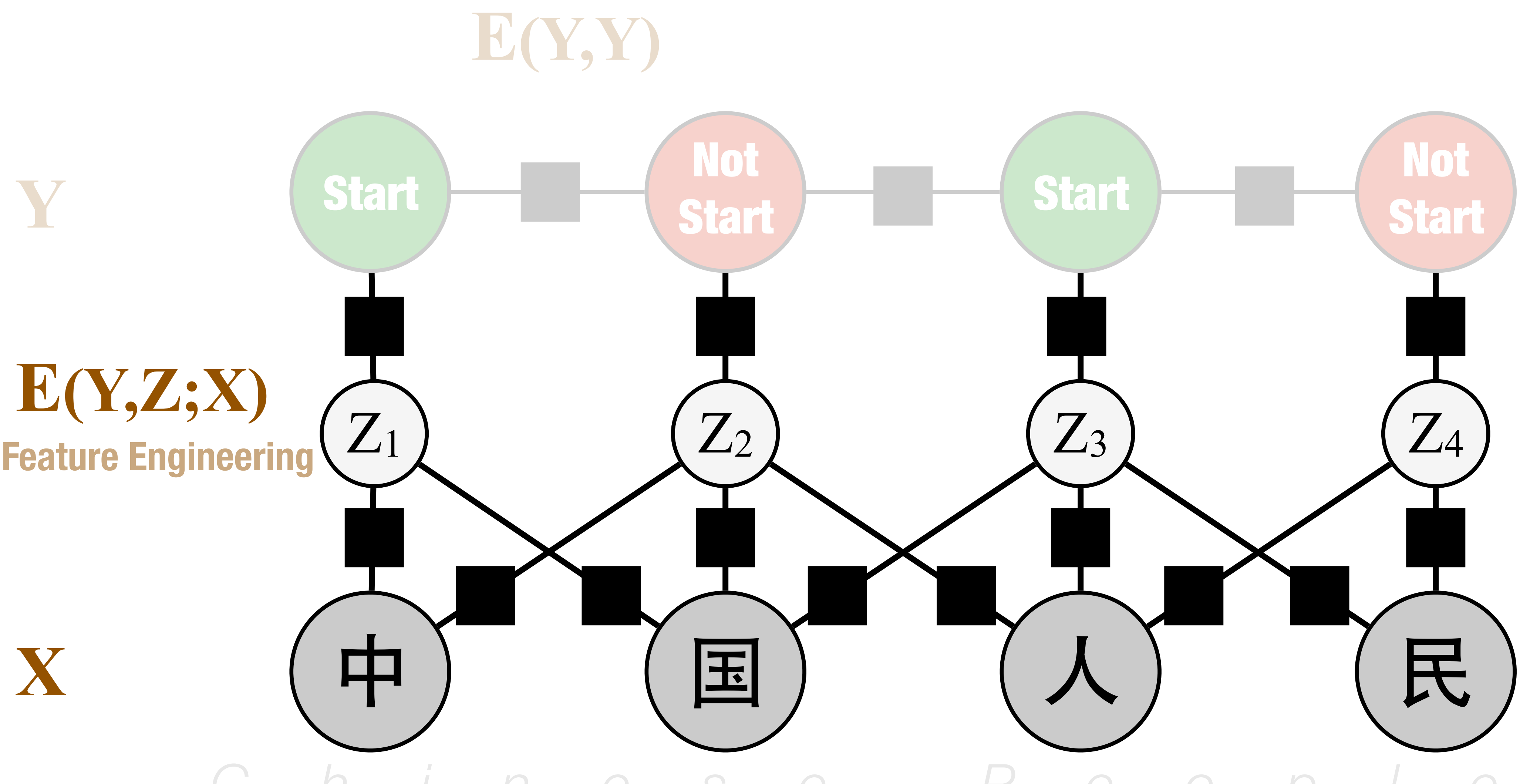


Structured Prediction

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表

e.g. "sequence labeling"

Example: Chinese Word Segmentation

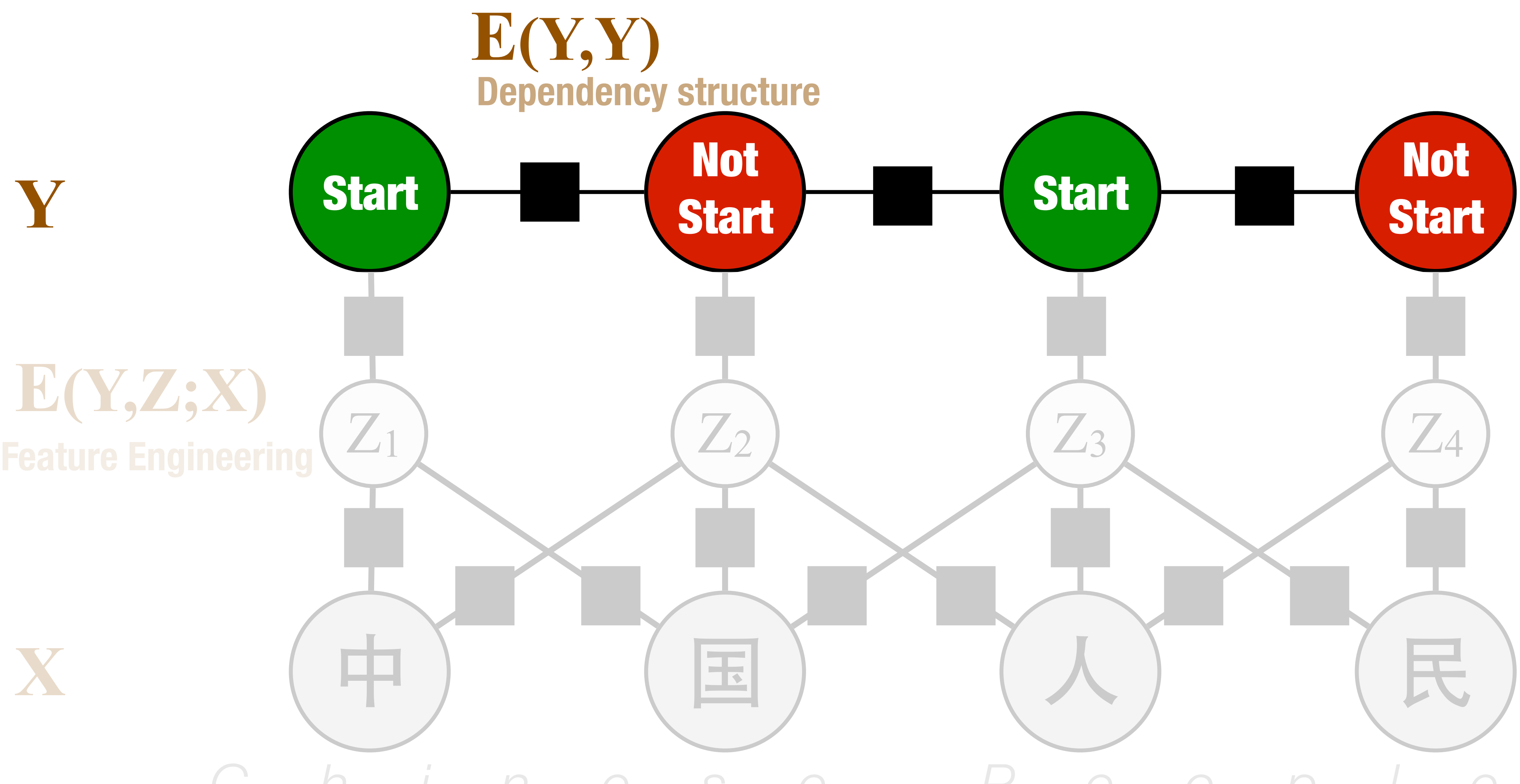


Structured Prediction

e.g. "sequence labeling"

Example: Chinese Word Segmentation

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表

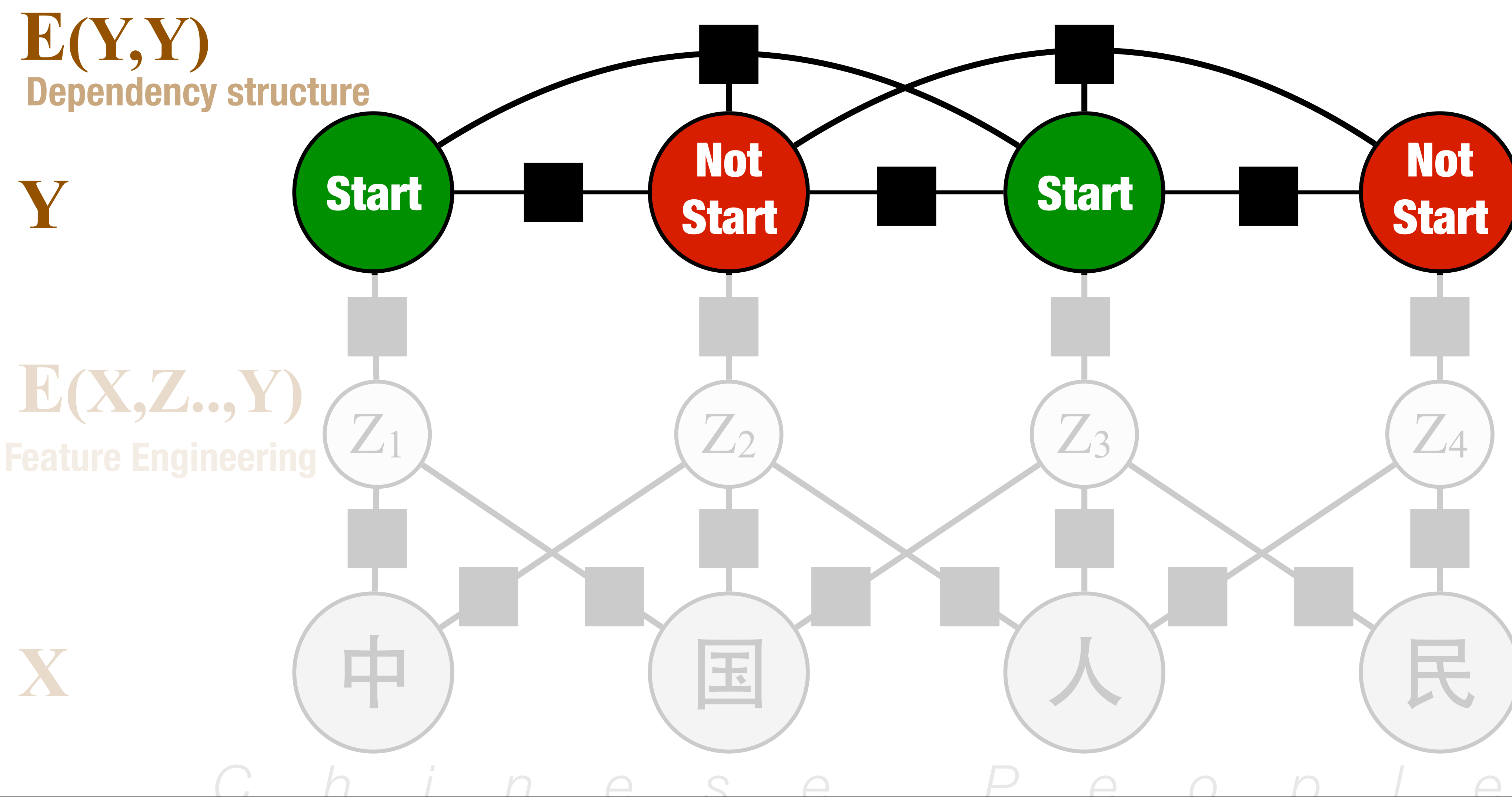


Structured Prediction

e.g. "sequence labeling"

Example: Chinese Word Segmentation

羅穆尼頭號對手
桑托倫在三州勝
選，而金瑞契只
贏得喬治亞州的
初選。羅穆尼面
臨的一大挑戰是，
其他共和黨總統
參選人目前均表



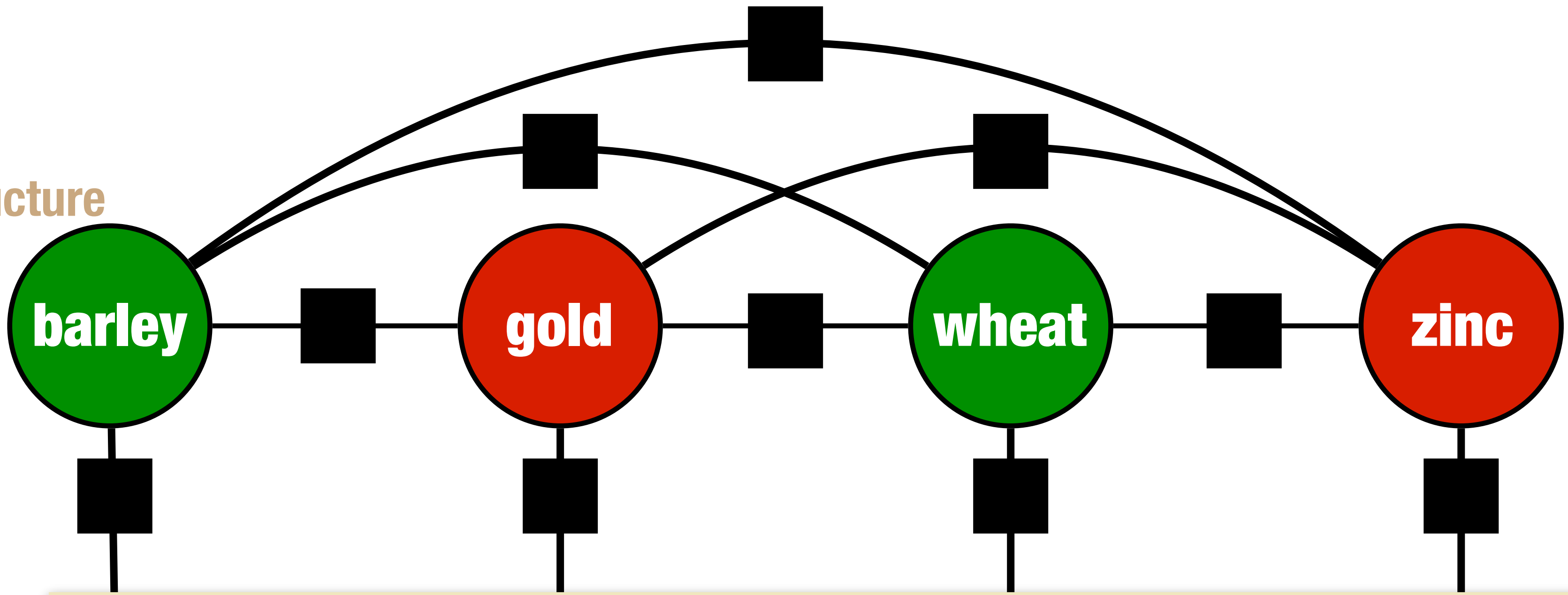
Structured Prediction

e.g. "multi-label classification"

Example: Multi-label Document Classification

E(Y,Y)
Dependency structure

Y



E(X,Y)
Feature Engineering

X

LONDON, March 3 - The U.K. Exported 535,460 tonnes of wheat and 336,750 tonnes of barley in January, the Home Grown Cereals Authority (HGCA) said, quoting adjusted Customs and Excise figures. Based on the previous January figures issued on February 9, wheat exports increased by nearly 64,000 tonnes and barley by about 7,000 tonnes. The new figures bring cumulative wheat exports for the period July 1/February 13 to 2.99 mln tonnes, and barley to 2.96

Structured Prediction

e.g. "multi-label classification"

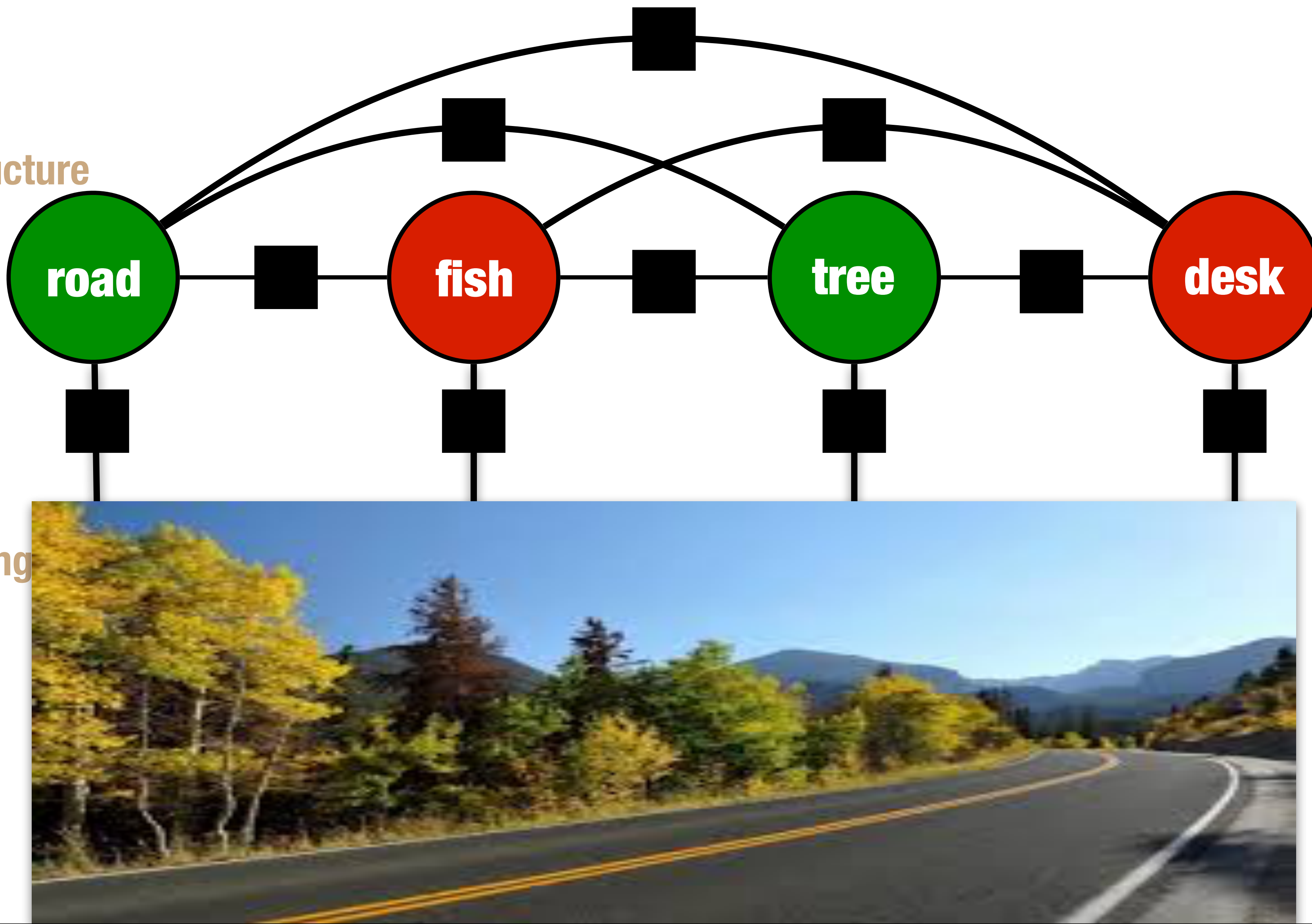
Example: Multi-label Image Classification

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

X



Structured Prediction

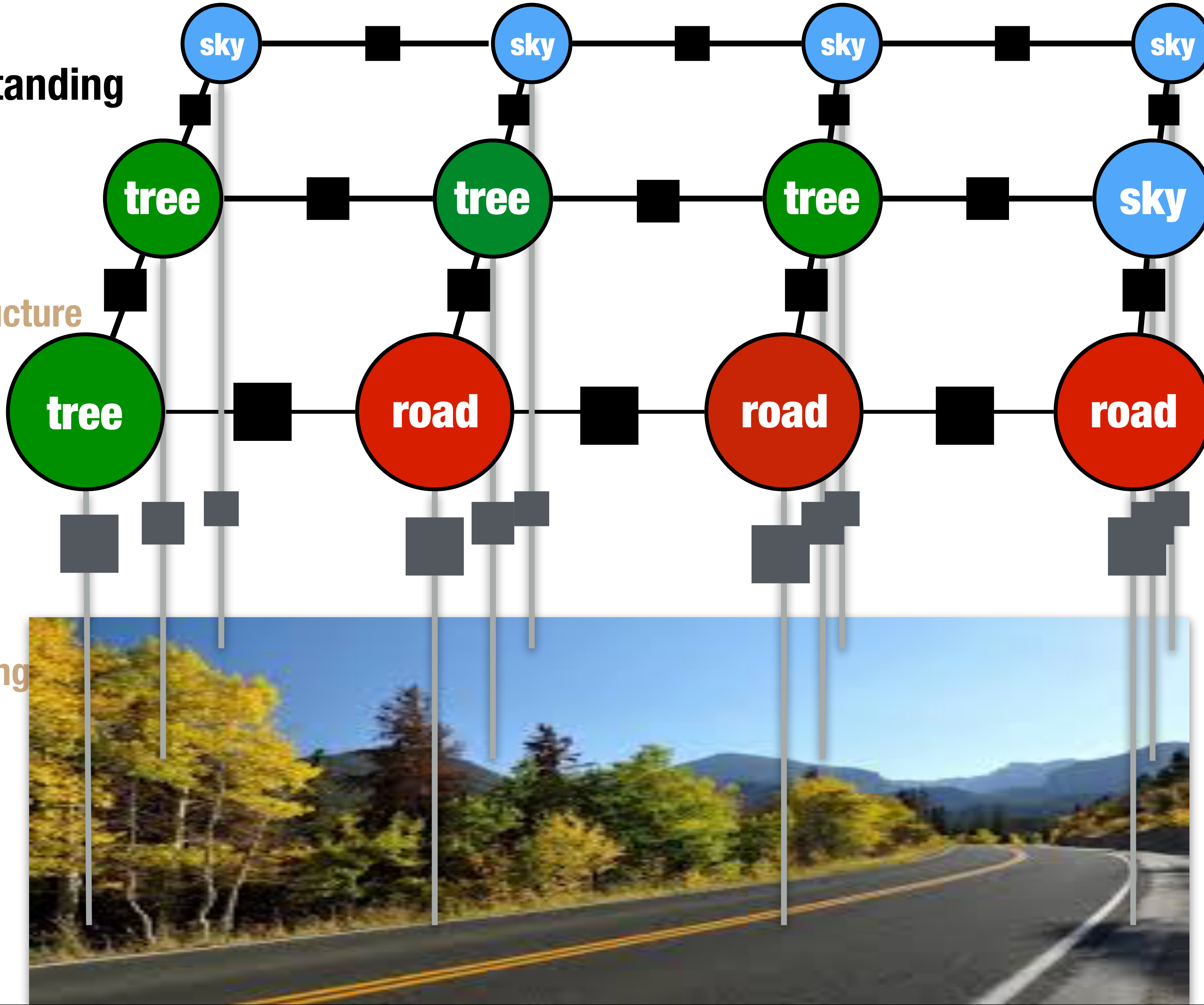
Example:
Scene Understanding

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

X



Structured Prediction

Example:
Scene Understanding

E(Y,Y)
Dependency structure

- **Expressivity** of dependencies
- **Parsimony** of parameterization
- **Tractability** of inference

Y

E(X,Y)
Feature Engineering

X

Structured Prediction

Sampling
Inference

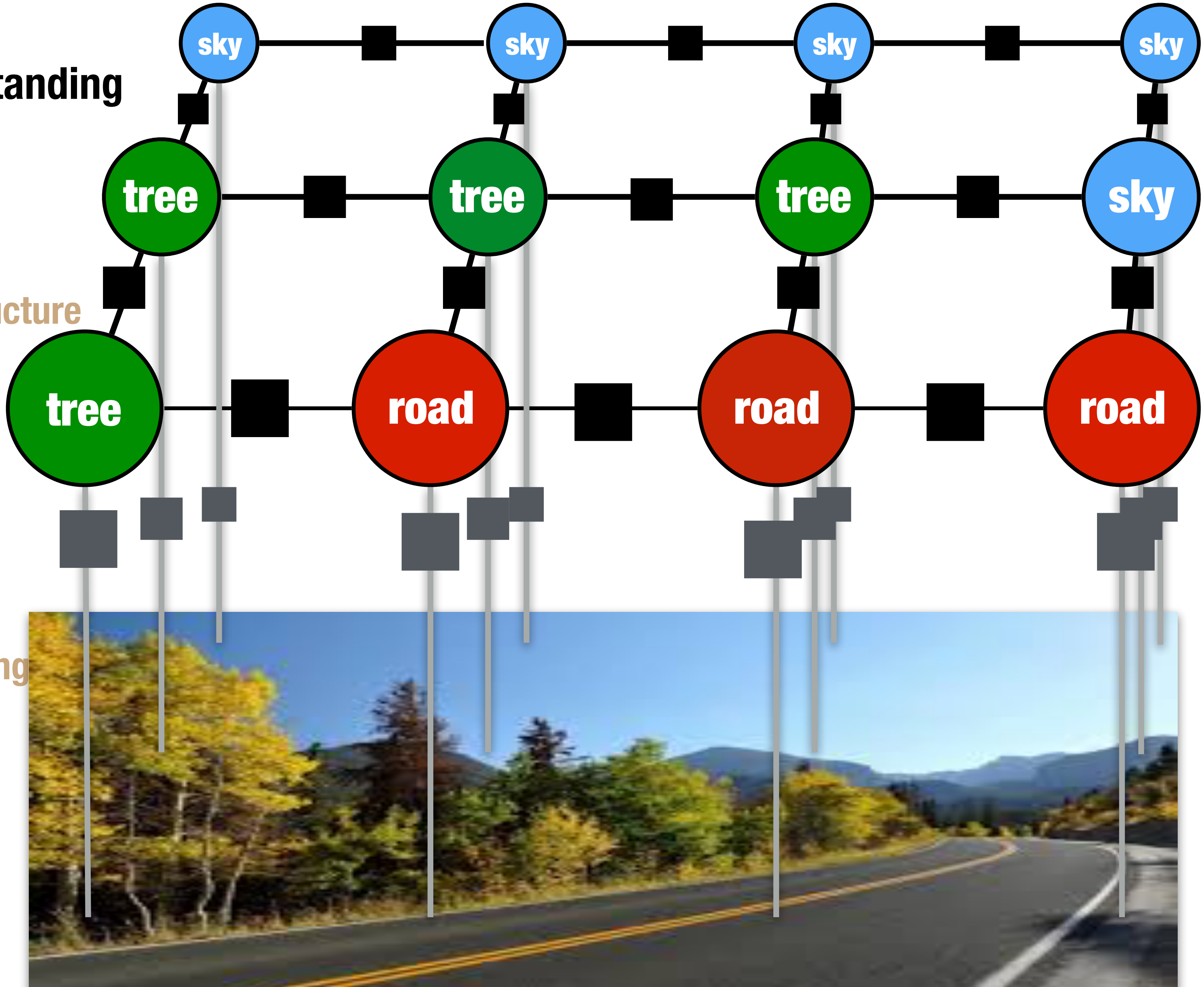
Example:
Scene Understanding

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

X



Structured Prediction

Sampling
Inference

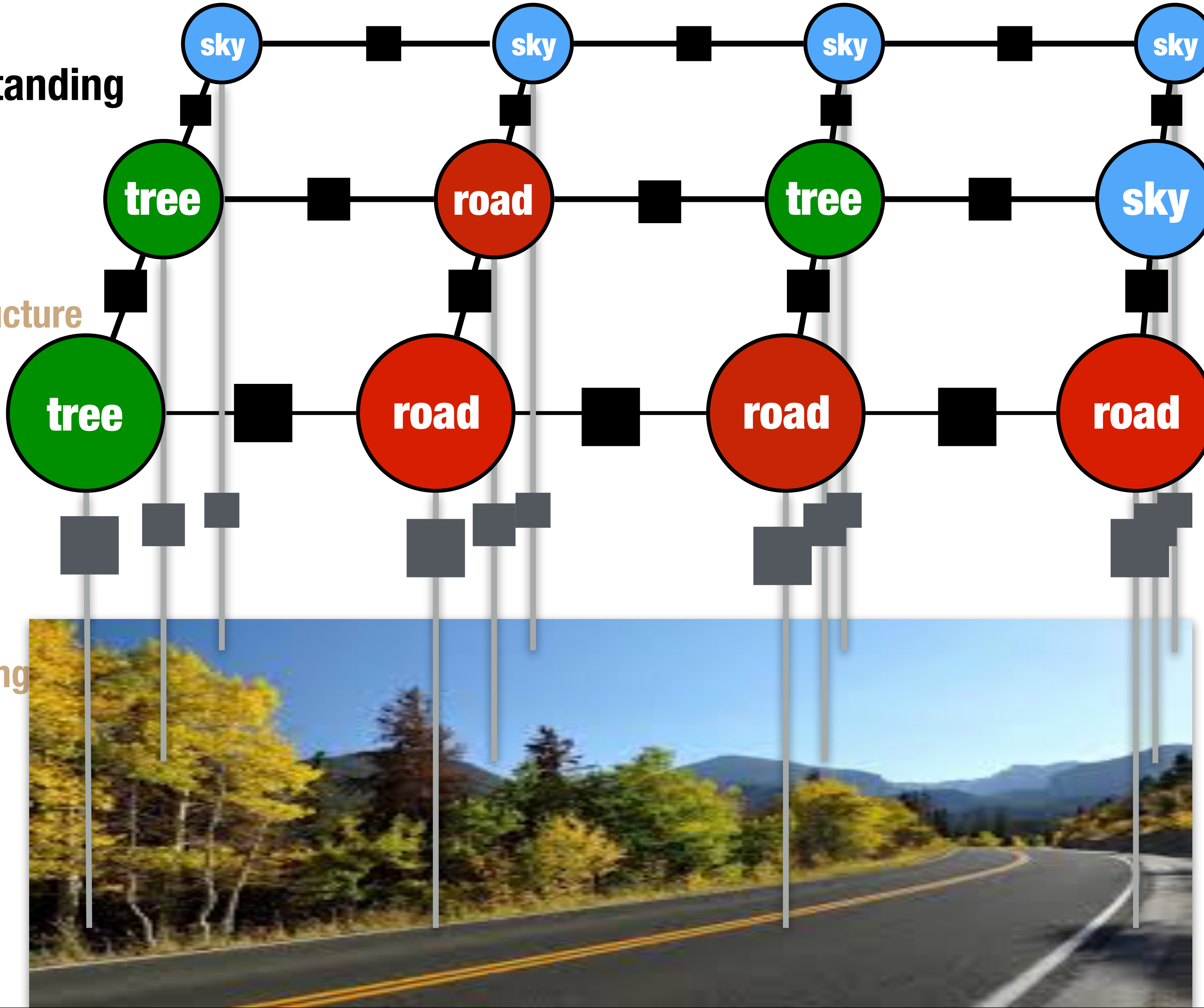
Example:
Scene Understanding

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

X



Structured Prediction

Sampling
Inference

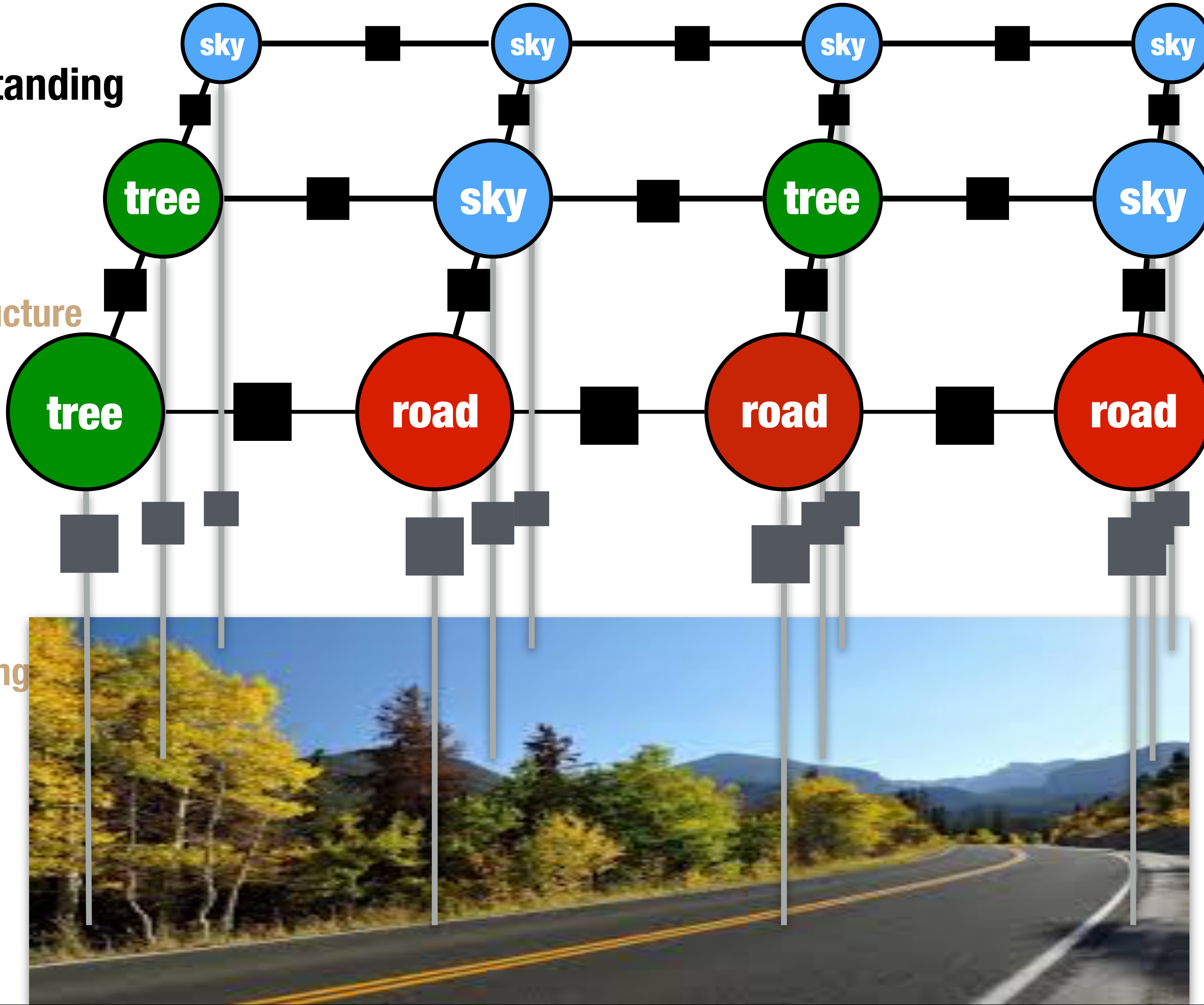
Example:
Scene Understanding

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

X



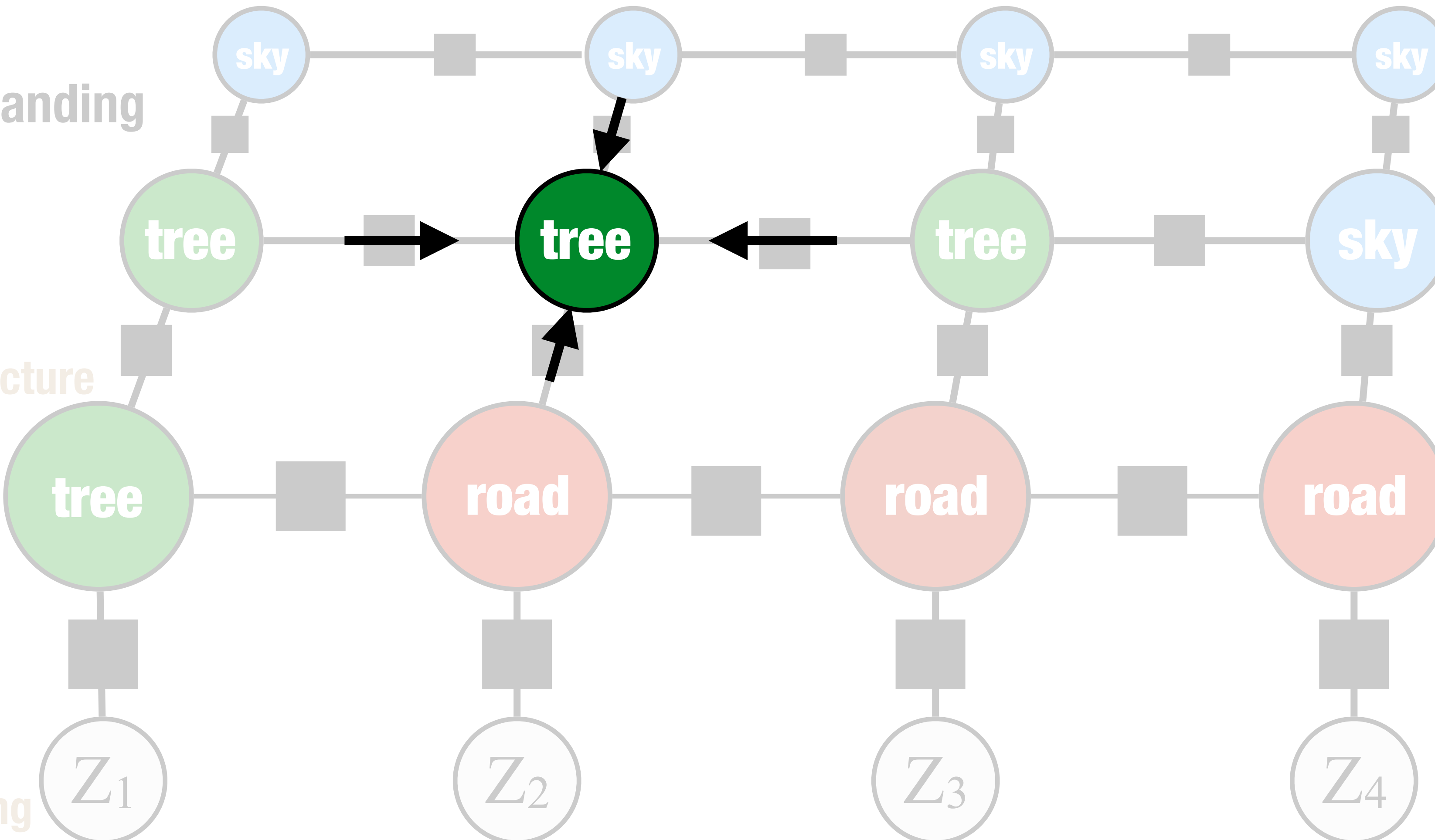
Structured Prediction *Variational Inference*

Example:
Scene Understanding

E(Y,Y)
Dependency structure

Y

E(X,Y)
Feature Engineering

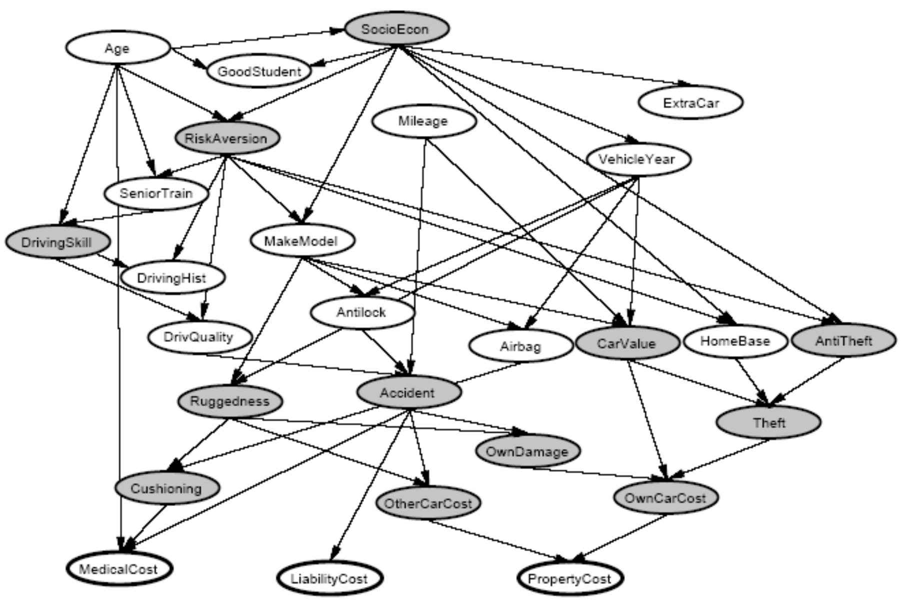


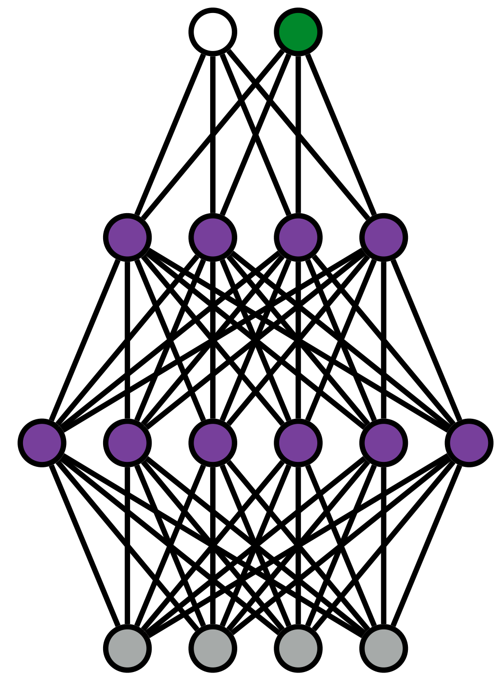
$$m_{i\to j}^{(t+1)}(x_j) = \sum_{x_i} \Phi_{ij}(x_i, x_j)\Phi_i(x_i) \prod_{k\in N(i)\ j} m_{k\to i}^{(t)}(x_i)$$

X

Bayesian Network

Deep Learning





Sparsely connected
Hand-designed representations
Loopy/iterated inference (typically)
Cautious about capacity
“Statistically conscientious”

Densely connected (learn connectivity)
Learned, distributed representations
Feed-forward inference (typically)
Wild about high capacity
“Wild West” 😃

Deep Learning

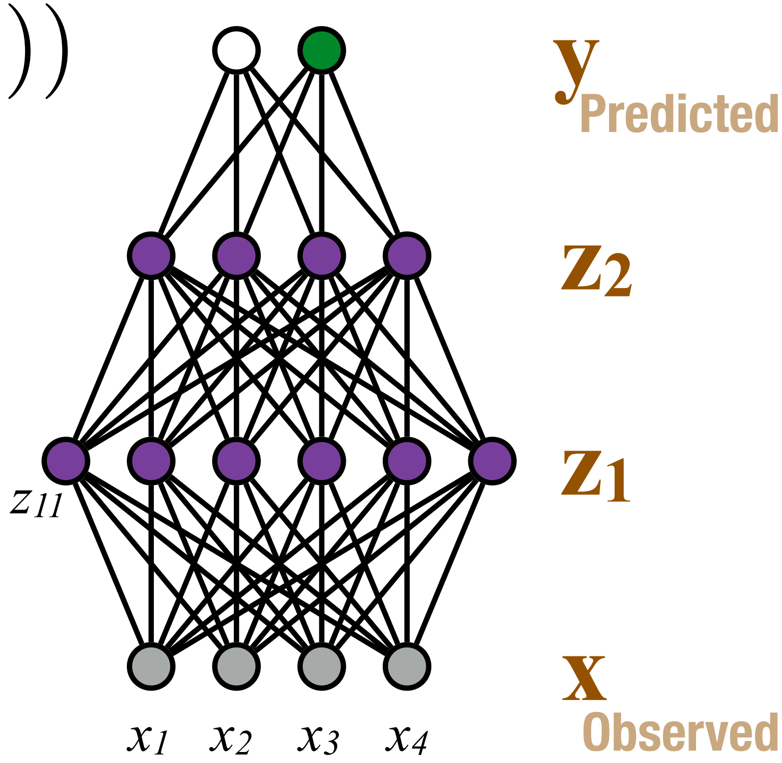
$$\mathbf{y} = \sigma\left(W_3 \mathbf{z_2}\right)$$

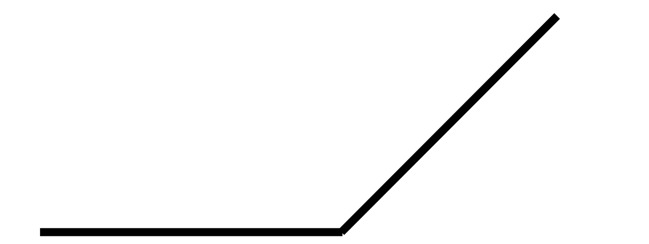
$$\mathbf{y} = \sigma\left(W_3 \sigma\left(W_2 \sigma\left(W_1 \mathbf{x}\right)\right)\right)$$

$$\mathbf{z}_2 = \sigma\left(W_2 \mathbf{z_1}\right)$$

$$z_{11} = \sigma\left(\sum_i w_{1i} x_i\right)$$

$$\mathbf{z}_1 = \sigma\left(W_1 \mathbf{x}\right)$$





$$\sigma(\cdot) = \max(\cdot, 0)$$

Deep Learning

$$\mathbf{y} = F\left(\mathbf{x}; W\right)$$

Training Data $\left\{\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right\}_{i=1}^{N}$

Loss

$$\mathcal{L} = \sum_i L\left(F(\mathbf{x}^{(\mathbf{i})}; W), \mathbf{y}^{(i)}\right)$$

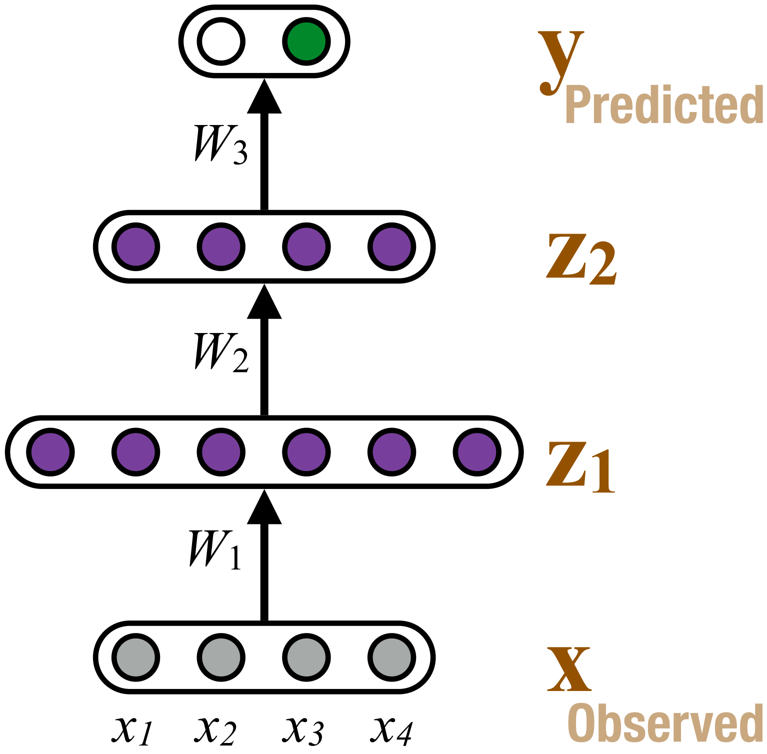
e.g. Squared error, Cross-entropy,…

Training

$$\arg\min_W \mathcal{L}$$

Gradient descent

$$W_{\text{new}} = W_{\text{old}} - \alpha \frac{\partial \mathcal{L}(W)}{\partial W}$$



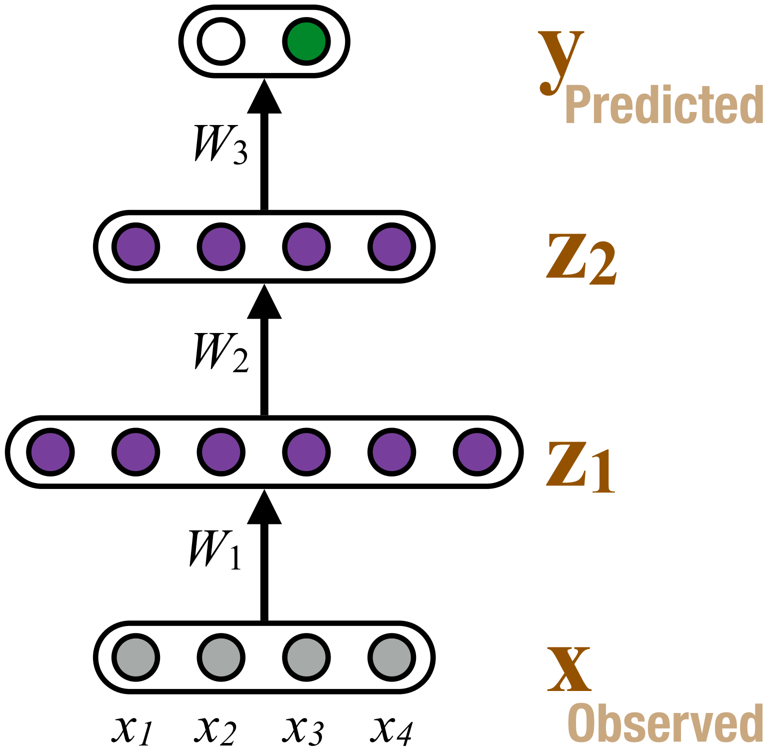
Key tools:

(1) Back-propagation

(2) Stochastic gradient descent

Deep Learning

$$\mathbf{y} = F\left(\mathbf{x}; W\right)$$



Back-propagation

Deep Learning

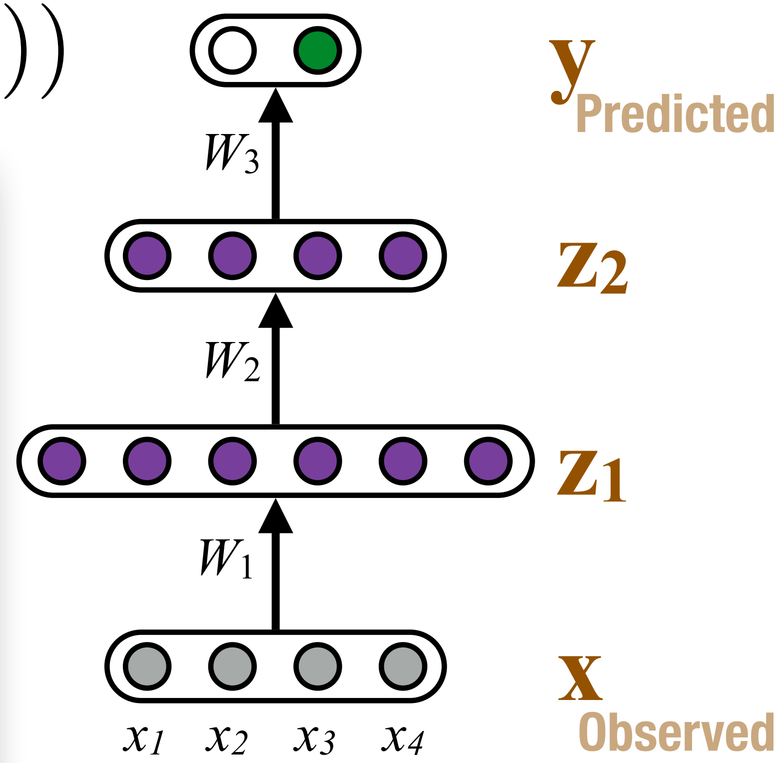
$$\mathbf{y} = \sigma\left(W_3\sigma\left(W_2\sigma\left(W_1\mathbf{x}\right)\right)\right)$$

The "chain rule"

$$g(f(x))' = g'(f(x)) \cdot f'(x)$$

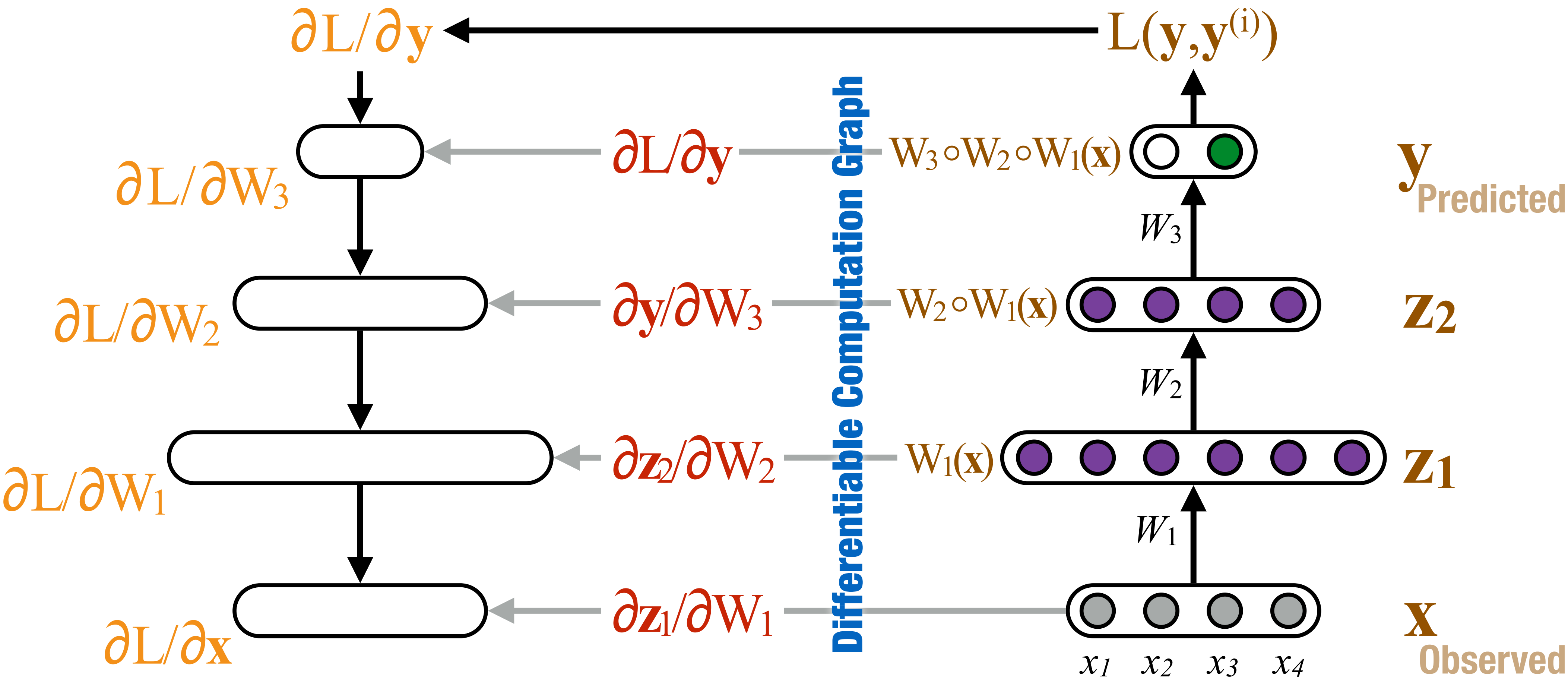
$$\frac{\partial g \circ f}{\partial x} = \frac{\partial g}{\partial f} \cdot \frac{\partial f}{\partial x}$$

$$\frac{\partial j \circ i \circ h \circ g \circ f}{\partial x} = \frac{\partial j}{\partial i}\frac{\partial i}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial f}\frac{\partial f}{\partial x}$$



Back-propagation

Deep Learning



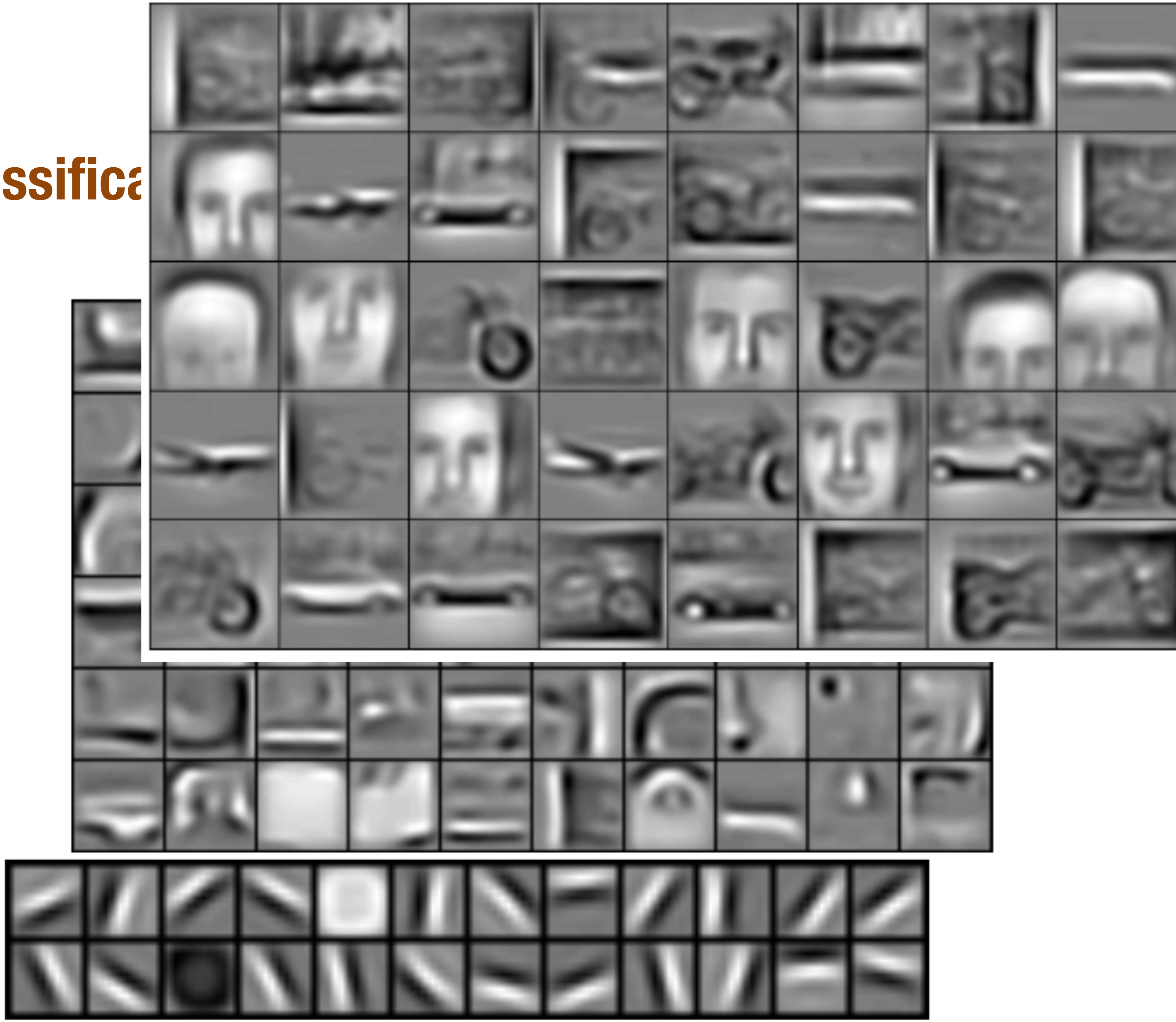
Can get gradient of *Loss* wrt parameters at any depth from

(1) local partial derivative functions

(2) numeric gradient from above

Example:
CNNs for
Object Classifica
in Images

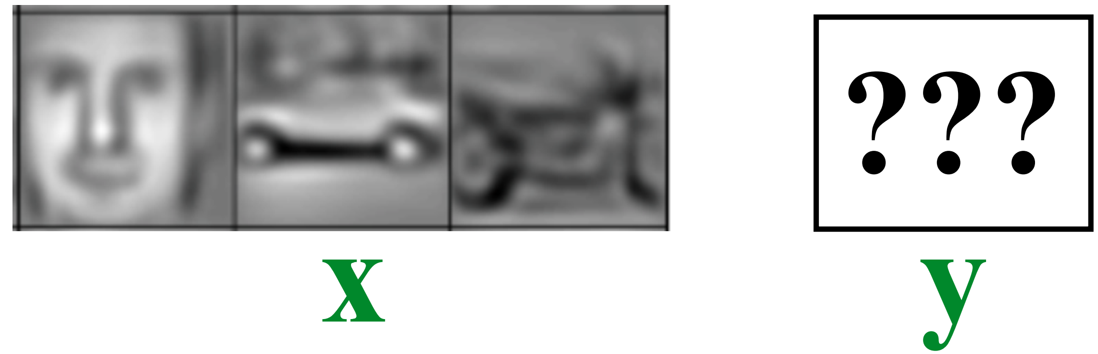
Representation
Learning



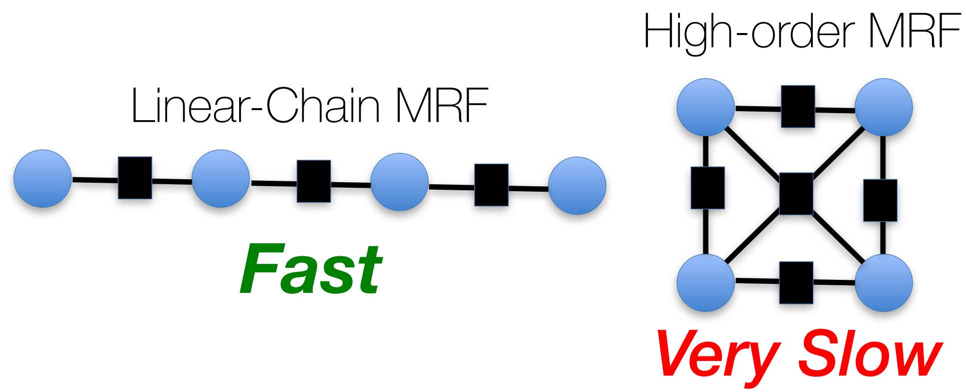
Lee, Grosse, Ranganath, Ng. "Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations."

Motivation for SPENs

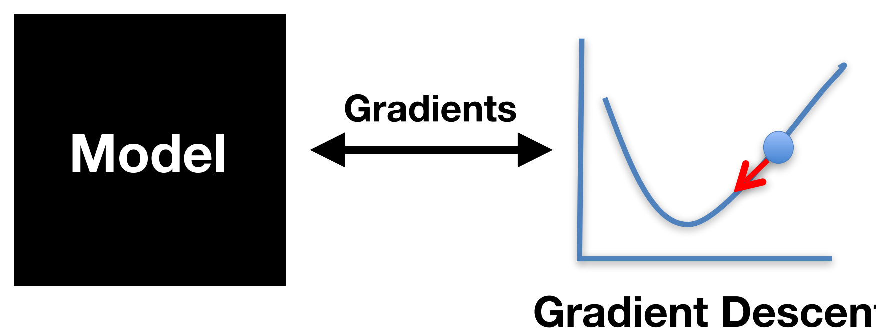
1. Use power of deep learning for *structure learning*



2. Provide an alternative to graphical models.

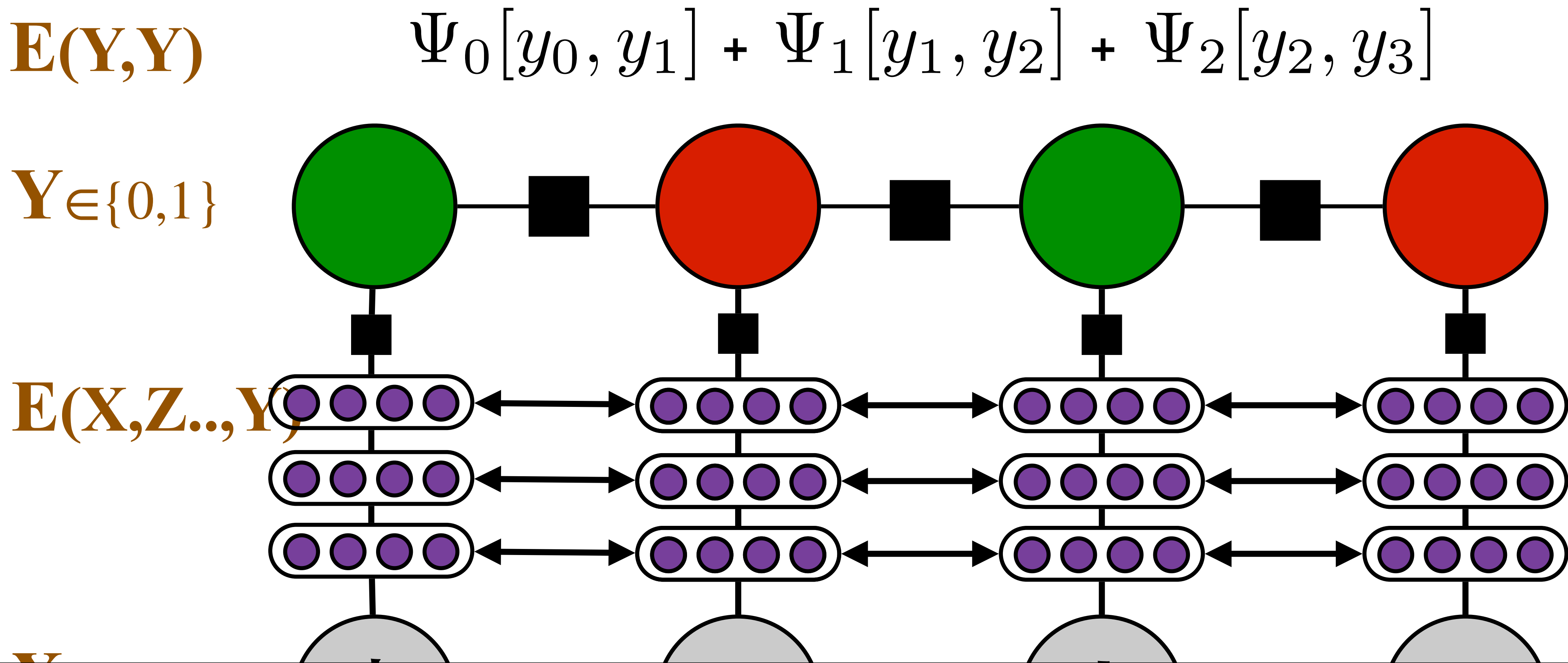


3. Black-box interaction with model.



Structured Prediction Energy Networks

[Belanger, McCallum, ICML 2016]



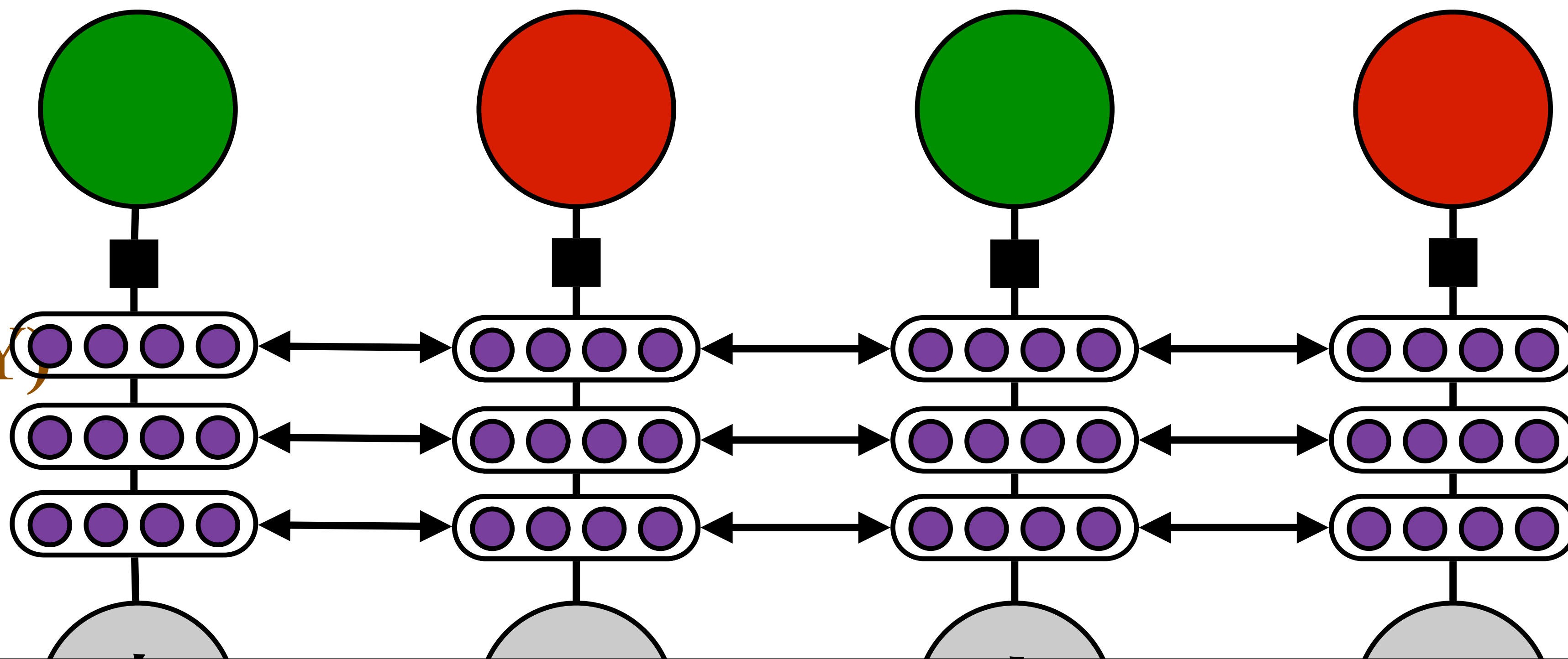
Structured Prediction Energy Networks

[Belanger, McCallum, ICML 2016]

E(Y,Y)

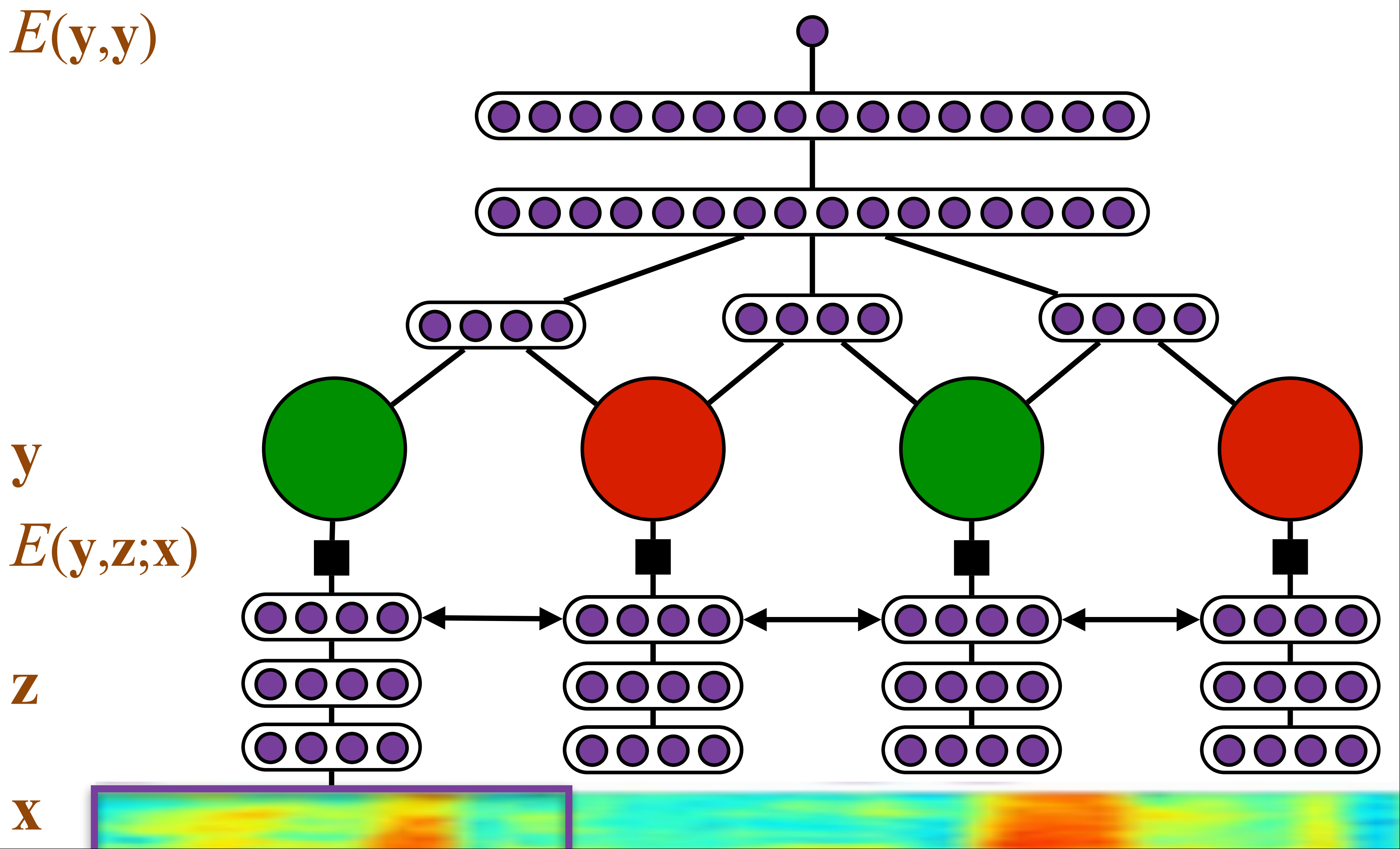
Y$\in\{0,1\}$

E(X,Z..,Y)



Structured Prediction Energy Networks

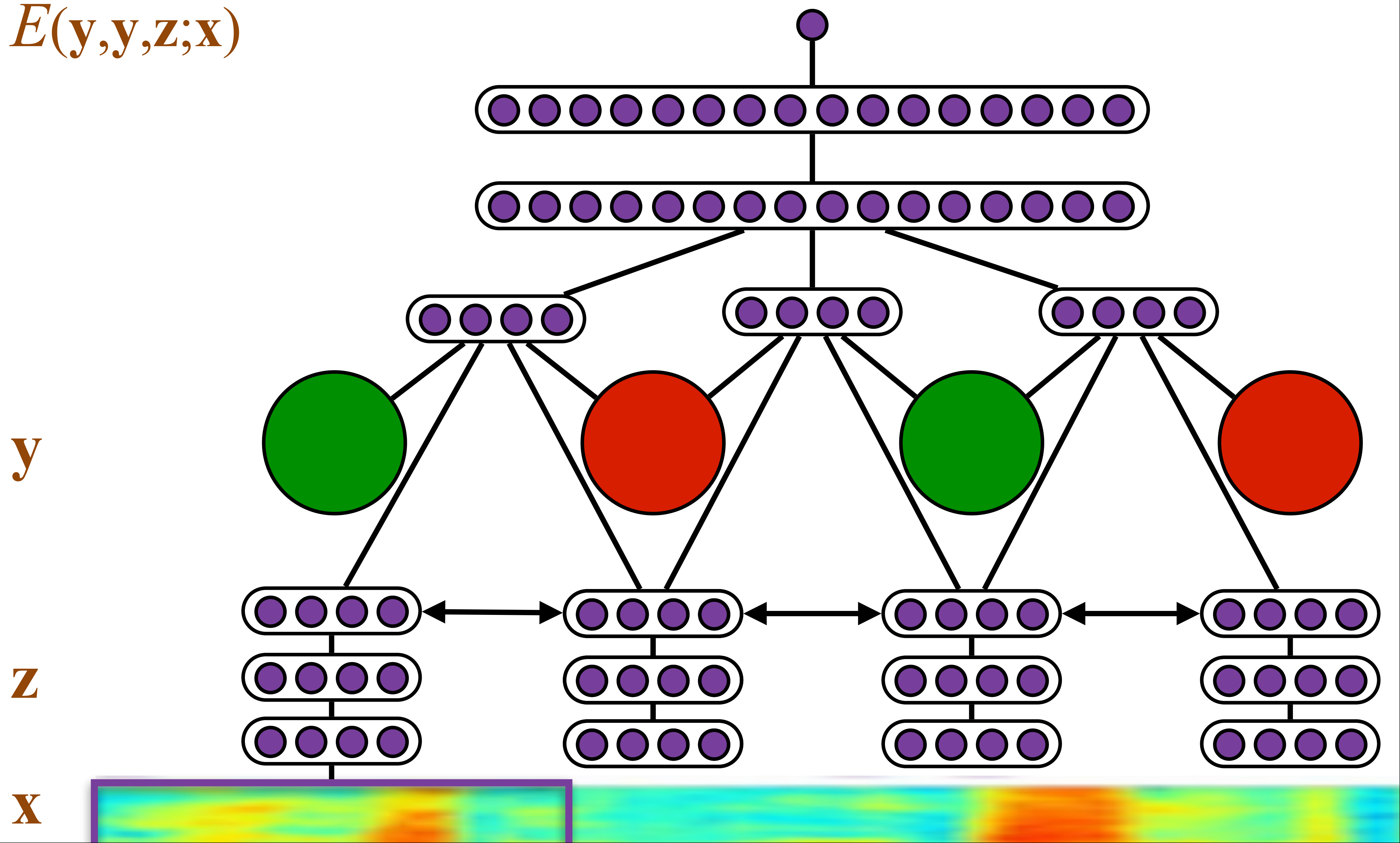
[Belanger, McCallum, ICML 2016]



Structured Prediction Energy Networks

[Belanger, McCallum, ICML 2016]

$E(\mathbf{y},\mathbf{y},\mathbf{z};\mathbf{x})$



Structured Prediction Energy Networks

[Belanger, McCallum, ICML 2016]

Energy network $E(\bar{\mathbf{y}}; F(\mathbf{x}))$

Soft prediction… **found by gradient descent**

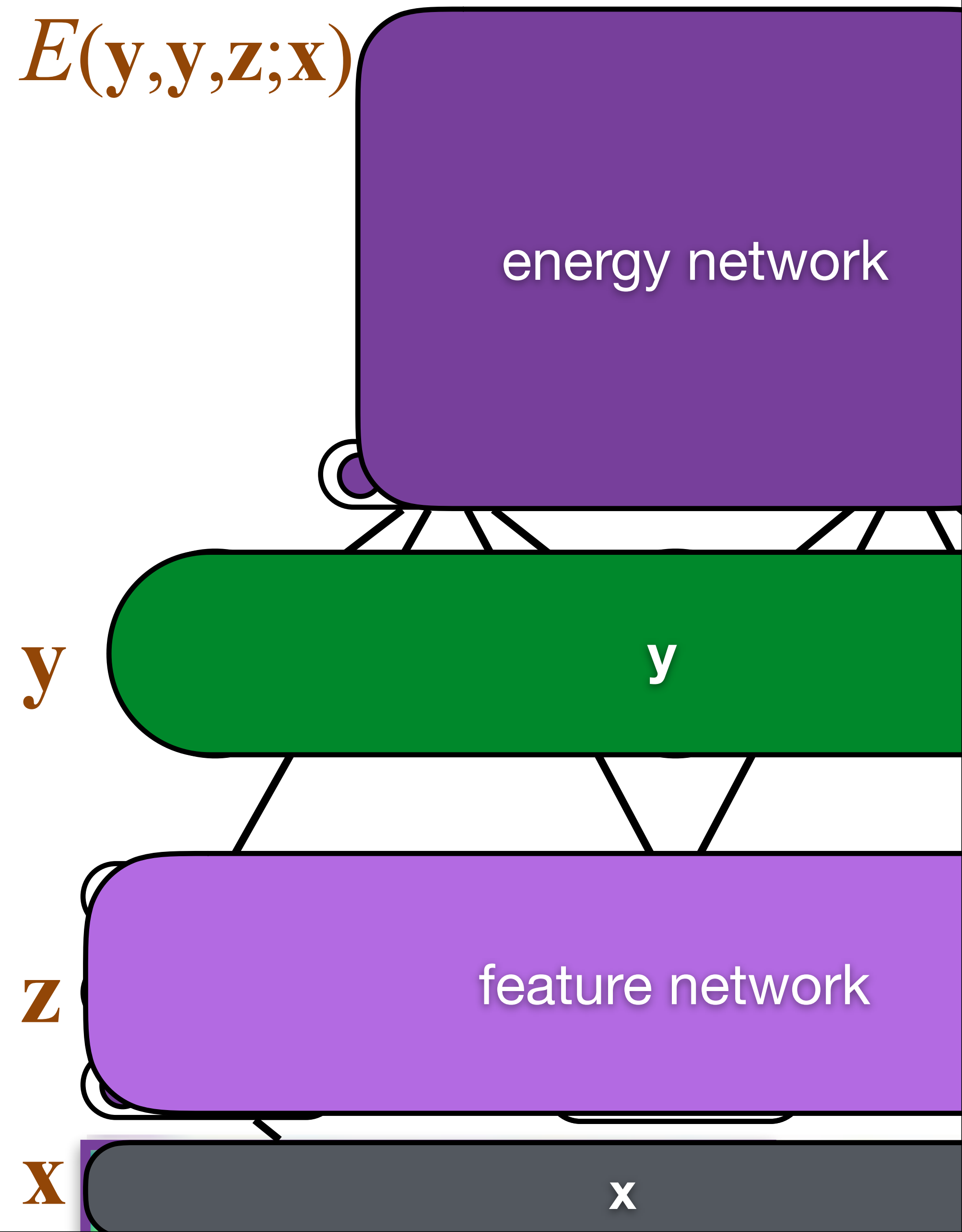
$$\bar{\mathbf{y}}^* = \arg\min_{\bar{\mathbf{y}} \in [0,1]^L} E(\bar{\mathbf{y}}; F(\mathbf{x}))$$

$$\frac{\partial E(\bar{\mathbf{y}}; F(\mathbf{x}))}{\partial \bar{\mathbf{y}}}$$

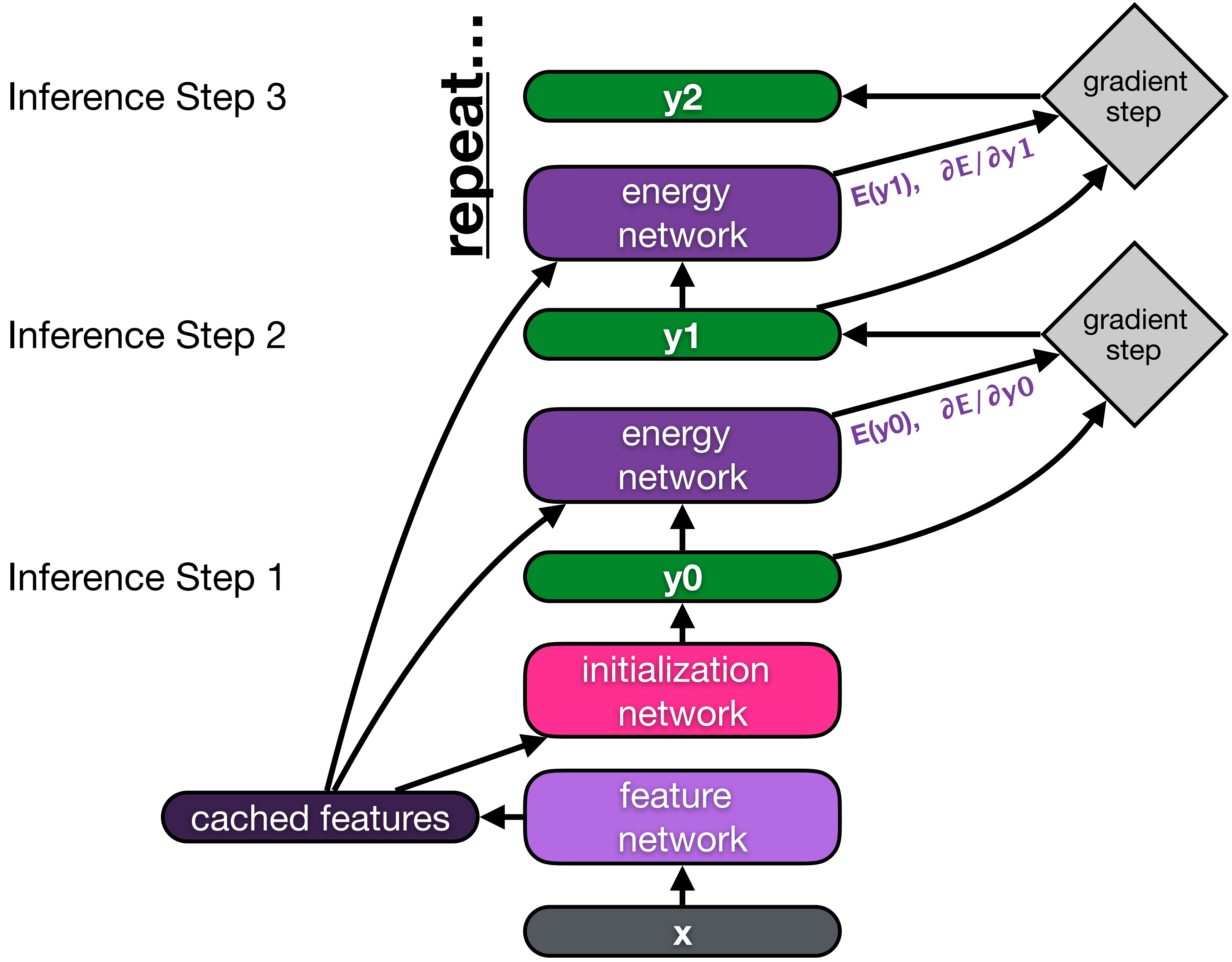
Relax y, to be continuous

$$\mathbf{y} \in \{0,1\}^L \rightarrow \bar{\mathbf{y}} \in [0,1]^L$$

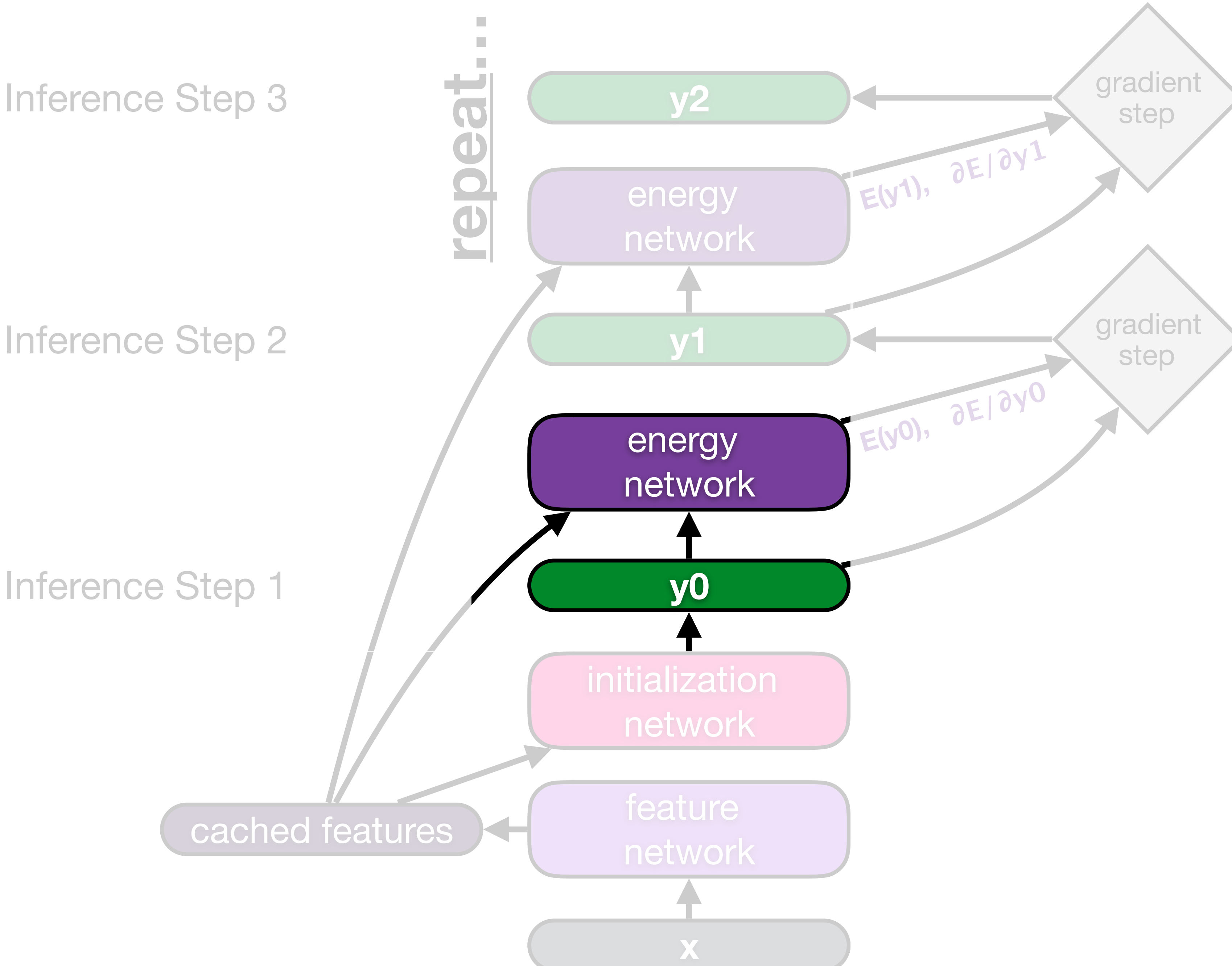
Feature Network $F(\mathbf{x})$



SPEN Inference Graph



SPEN Inference Graph



Gradient used to Modify Inputs

"A Neural Algorithm for Artistic Style"
[Gatys et al. 2015]

SPENs use similar idea:
Optimize energy using backprop all the way down to the raw pixels.

Learning Algorithm 1: Structured SVM

Belanger, McCallum, ICML 2016

(Taskar et al., 2004; Tsochantaridis et al., 2004)

Training Loss= $\mathcal{L} =$

sum / worst violation / Penalty / Model's energy difference

$$\sum_{\{\mathbf{x}^{(i)},\ \mathbf{y}^{(i)}\}} \max_{\bar{\mathbf{y}}} \left[\Delta(\mathbf{y}^{(i)}, \bar{\mathbf{y}}) - \left(E_W(\bar{\mathbf{y}}; \mathbf{x}^{(i)}) - E_W(\mathbf{y}^{(i)}; \mathbf{x}^{(i)}) \right) \right]_+$$

training data / true / predicted / predicted / true

search requires **Loss-Augmented Inference**

$$\arg\min_{\bar{\mathbf{y}}} \left(-\Delta(\mathbf{y}^{(i)}, \bar{\mathbf{y}}) + E_W(\bar{\mathbf{y}}; \mathbf{x}^{(i)}) \right)$$

Penalty must be differentiable

Stochastic Gradient

$$\frac{\partial \mathcal{L}}{\partial W}$$

Belanger, McCallum, ICML 2017

Learning Algorithm 2: End-to-end "backprop through inference"

Training Loss= $\mathcal{L} =$

Direct Risk Minimization

sum

$$\sum_i L\left(\mathbf{y}^{(i)}, \text{Algorithm}_W(\mathbf{x}^{(i)})\right)$$

$$\sum_i L\left(\mathbf{y}^{(i)}, \arg\min_{\mathbf{y}} E_W(\mathbf{y}, \mathbf{x}^{(i)})\right)$$

training data

Direct application of:
Justin Domke, AISTATS, 2012.
"Generic Methods for Optimization-Based Modeling"

Algorithm for inference

$$\bar{\mathbf{y}}^* = \bar{\mathbf{y}}^{[0]} + \sum_{t=1}^{T} \alpha_t \frac{\partial}{\partial \bar{\mathbf{y}}} E_W\left(\mathbf{x}, \bar{\mathbf{y}}^{[t-1]}\right)$$

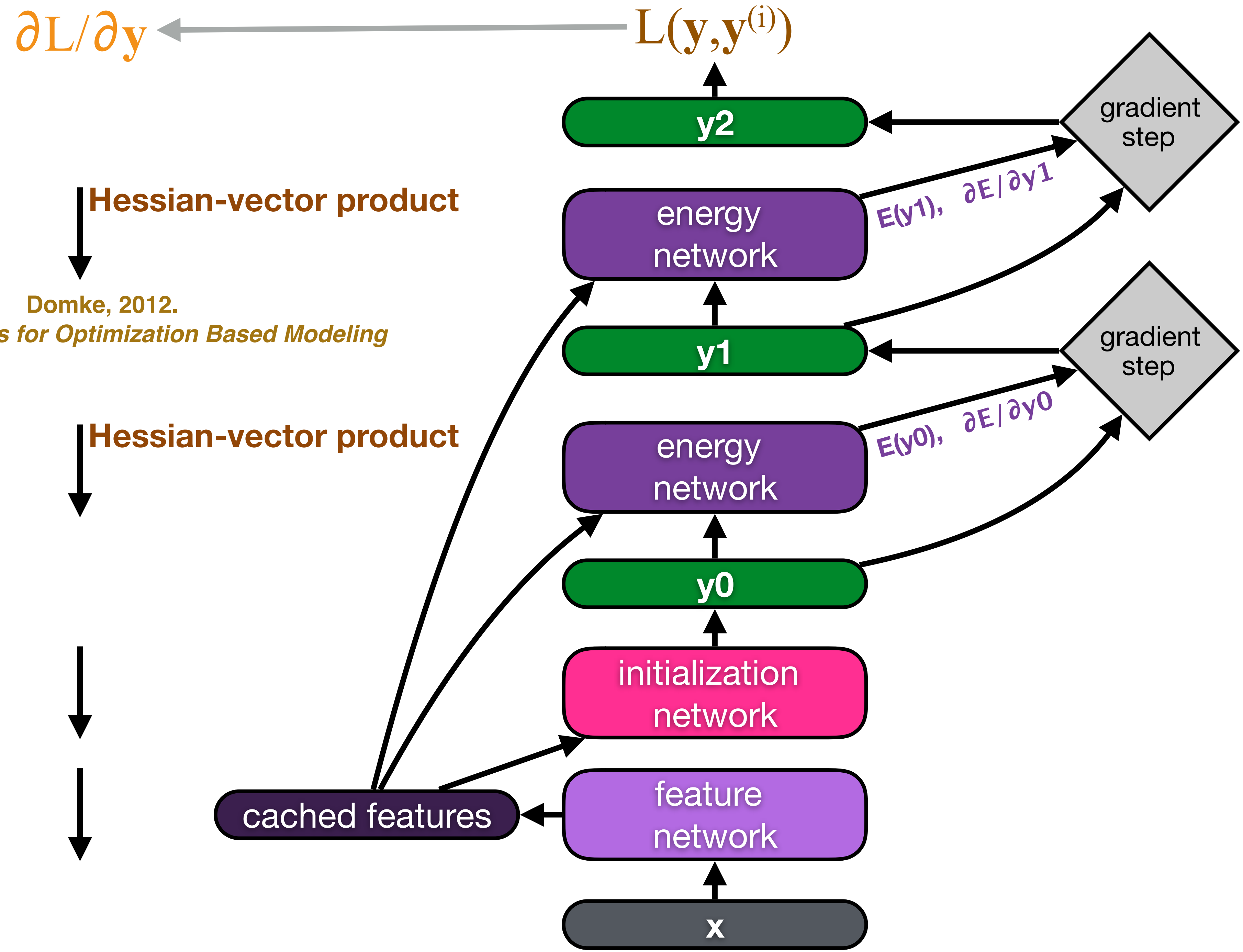
sum over "time steps" of inference

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial L}{\partial \bar{\mathbf{y}}^*} \frac{\partial \bar{\mathbf{y}}^*}{\partial W} = \sum_{t=1}^{T} \alpha_t \frac{\partial L}{\partial \bar{\mathbf{y}}^*} \left(\frac{\partial}{\partial W} \frac{\partial}{\partial \mathbf{y}} E_W\left(\mathbf{x}, \bar{\mathbf{y}}^{[t-1]}\right)\right)$$

Hessian-Vector product can be approximated using one-dimensional finite differences

sum over "time steps" of inference

Learning Algorithm 2 Graph



Domke, 2012.
Generic Methods for Optimization Based Modeling

Chapter 3

Light Supervision training of Structured Prediction Energy Networks

(Turing complete!)

1. Human writes arbitrary prior knowledge

(SPEN)

2. Learn model with arbitrary dependencies.

3. Efficient inference by gradient descent.

AUTHOR AUTHOR
Anna Popescu (2004), "Interactive Clustering,"
EDITOR EDITOR
Wei Li (Ed.), Learning Handbook, Athos Press,
LOCATION
Souroti.

Human writes arbitrary prior knowledge…

"AUTHOR field should be contiguous, only appearing once."

…as a scoring function $V(x$=citation, y=labeling$)$

```
score = 0
score -= 1 foreach AUTHOR non-contiguous
score -= 1 if has both JOURNAL & BOOKTITLE
score -= 1 foreach "using" not in TITLE
score -= 1 foreach [A-Z]\. not AUTHOR|EDITOR
score -= 1 if PUBLISHER before JOURNAL ...
```

(like rule-based AI before ML was popular)

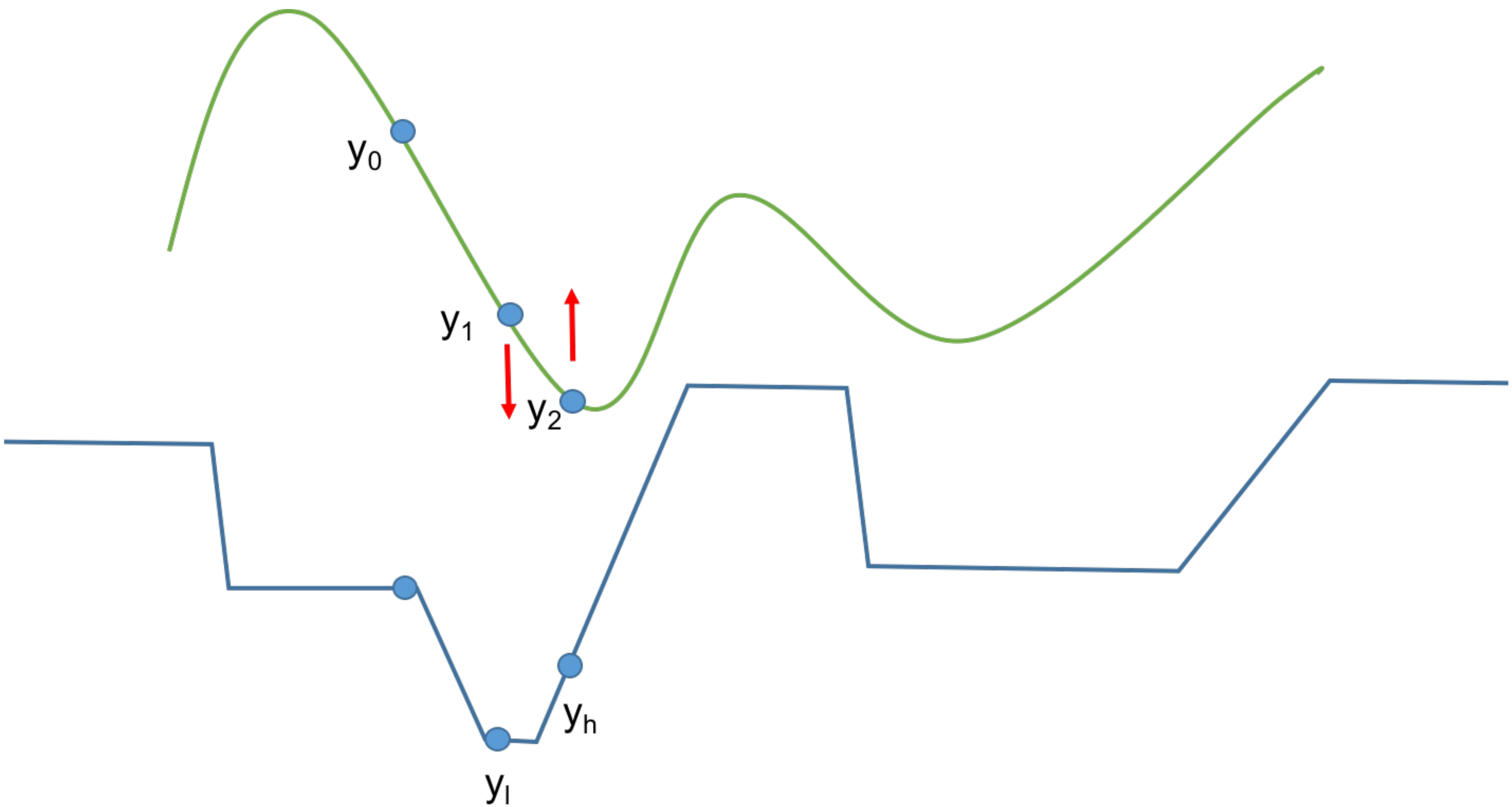
Why use ML if we get a ruled-based scoring function?

- **Doesn't generalize**
  - examines just a few features
  - SPENs will learn correlated features, labels.
- **No inference procedure** just scores for given (x,y)
  - stochastic optimization is slow
  - SPENs provide gradient-descent inference

Learning Algorithm 3: "ranking successive gradient steps"

Rooshenas, McCallum,...
forthcoming

Training Loss=
$$\sum_{\mathbf{x}\in\mathcal{D}}[\alpha(V(\mathbf{y}_h,\mathbf{x})-V(\mathbf{y}_l,\mathbf{x}))-E_{\mathbf{w}}(\mathbf{y}_h,\mathbf{x})+E_{\mathbf{w}}(\mathbf{y}_l,\mathbf{x})]_+$$



Preliminary Experiments

(…much more work and comparisons in future…)

Weak-Sup SPEN: simple test
Multi-label Document Classification

x = Medical bag-of-words

[amount, cystourethrogram, diagnosed, episode, evaluate, exam, fever, grade, growth, hematuria, infection, interval, kidney, left, lower, occurred, patient, pole, previously, purpose, reflux, renal, scar, scarring, small, study, tract, urinary, vesicoureteral, voiding, year]

y = multiple ICM-9-CD codes

[593-70, 599-00]

x = Human background knowledge

Keyword descriptions of ICM-9-CD codes. (Not gathering any labeled correlation knowledge.)

593-70: vesicoureteral, reflux, unspecified, nephropathy
V79-99: viral, chlamydial, infection, conditions, unspecified
753-00: renal, agenesis, dysgenesis

Scoring function gives +1 for each label:keyword cooccurrence.

$$V(y^i, x^i) = \sum_j I(l_j \in y^i) I(|x^i \cap w_j| > 0) - \gamma \max(|y^i| - 1, 0)$$

Label, Keyword matches

Sparsity constraint

Does the SPEN generalize over the human scoring function?

ICM-9-CD code data set, evaluate **F1 of label set**

Human Scoring Function, Exhaustive Search						SPEN
N≦1	N≦2	N≦3	N≦4	N≦5	N≦6	
15.5	**18.3**	**19.6**	**20.5**	**21.1**	**20.3**	**22.6**

(~10x faster)

Weak-Sup SPEN: better test
Citation Field Extraction

x = Citation Token Sequence

Anna Popescu (2004), “Interactive Clustering,”
Wei Li (Ed.), Learning Handbook,
Athos Press, Souroti.

y = Seq. of Labels ∈ |14|

AUTHOR AUTHOR YEAR TITLE TITLE
EDITOR, EDITOR EDITOR BOOKTITLE, BOOKTITLE
PUBLISHER PUBLISHER LOCATION

x = Human background knowledge

Human-written scoring function. 50 lines of code. Written in ~1 hour.

```
score -= 1 foreach AUTHOR non-contiguous
score -= 1 if has both JOURNAL & BOOKTITLE
score -= 1 foreach “using” not in TITLE
...
```

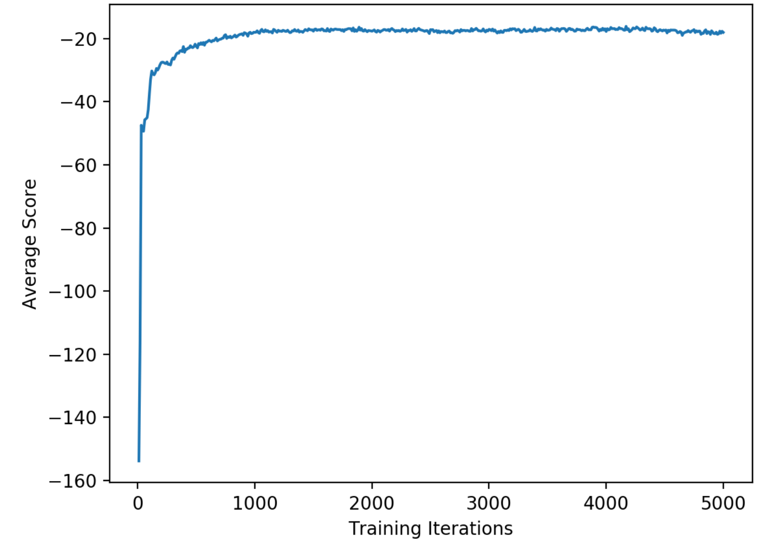
~4000 unlabeled examples, 0 labeled.

Scoring function advice:

- Penalties only, so 0 = best.
- Can use varying magnitudes, -1, -5, -10.
- Debug with some stochastic optimization.

Citation Field Extraction Accuracy

Method (no labeled data)	Token **accuracy**	Time sec/citation	Ave. V() score
GE [Mann & McCallum '10]	37%	?	N/A
V search 10	34%	14	-1.86
V search 100	39%	170	-0.98
V search 1000	42%	1240	-0.62
SPEN	**52%**	0.0008	~ -20



Example text

Wright, A. K. Simple imperative polymorphism. Lisp and Symbolic Computation 8, 4 (Dec. 1995), 343-356.

V search 100 output

AUTHOR TITLE AUTHOR AUTHOR AUTHOR AUTHOR NOTE NOTE NOTE NOTE NOTE NOTE DATE DATE PUB PUB PUB

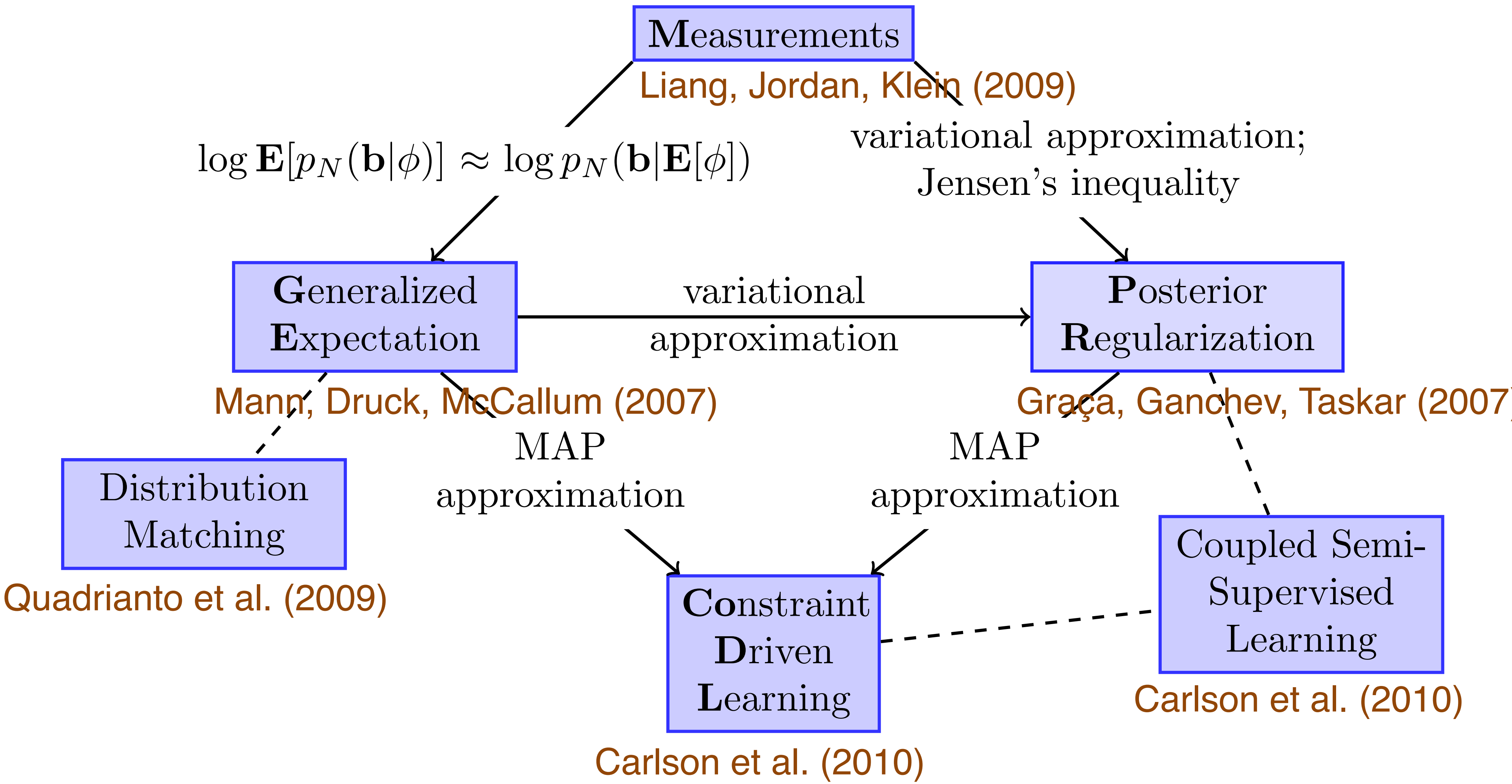
SPEN output

AUTHOR TITLE TITLE TITLE TITLE TITLE TITLE TITLE TITLE TITLE TITLE DATE DATE DATE PAGES PAGES PAGES

Related Work

- **Deep Value Networks**…
  [Gygli, Norouzi, Angelova 2017 ICML]
  - Matching magnitude (rather than just ranking).
  - Hurts accuracy? 5% vs SPEN's 52%
- **Constraint-Driven Learning**
  [Chang, Ratinov, Roth 2007 ACL]
  - Supervised training ➡ Pseudo-label data w/ constraints ↩
- **Snorkel: Rapid Training Data Creation with Weak Supervision**
  [Ratner, Bach, Ehrenberg, Fries, Wu, Ré 2017 VLDB]
  - Rules ➡ Pseudo-labeled data ➡ Supervised (self) training
- **Label-Free Supervision of NNs w/ … Domain Knowledge**
  [Stewart, Ermon 2017 AAAI]
  - Constraints ➡ Loss function ➡ Train feed-forward NN.

GE Related Work



Summary

- ***Generalized Expectation***
  - Learning from unlabeled data + “labeled features”
  - Hard to do inference
- ***Structured Prediction Energy Networks***
  - Representation learning for *output* variables
  - Test-time inference by gradient descent
  - New SPEN training method: Ranking
- **Experiments**
  - Multi-label Classification: ICM-9
  - Sequence labeling: Citation field extraction
- **Next**
  - Training on corpus-wide expectations.
  - Interactive tools for score function development.