Advanced Multimodal Machine Learning

Lecture 10.1: Probabilistic Graphical Models

Louis-Philippe Morency
Tadas Baltrusaitis

Lecture Objectives

- Probabilistic Graphical Models
- Markov Random Fields
  - Boltzmann/Gibbs distribution
  - Factor graphs
- Conditional Random Fields
  - Multi-View Conditional Random Fields
- CRFs and Deep Learning
  - DeepConditional Neural Fields
  - CRF and Bilinear LSTM
- Continuous and Fully-Connected CRFs

Administrative Stuff

Lecture Schedule

Classes	Lectures
Spring break 3/13 – 3/17	
Week 9 3/21 & 3/23	**Multimodal alignment** • Attention models and multi-instance learning • Multimodal synchrony and prediction
Week 10 3/28 & 3/30	**Probabilistic Graphical Models** • Markov random field and Boltzmann machines • Latent, continuous and fully-connected CRFs
Week 11 4/4 & 4/6	***Mid-term project assignment - Presentations***

Lecture Schedule

Classes	Lectures
Week 12 4/11 & 4/13	**Multimodal fusion** • Multi-kernel learning and fusion • Sample-based late fusion
Week 13 4/18 & 4/20	**Advanced multimodal representations** • Image and video description • Guest lecture: Ruslan Salakhutdinov
Week 14 4/25 & 4/27	**Multilingual computational models** • Neural Machine Translation • Guest lecture: Graham Neubig
Week 15 5/2 & 5/4	***Final project assignment - Presentations***

Upcoming Schedule

- Pre-proposal (tomorrow Wednesday 2/5 at 9am)
- First project assignment:
  - Proposal presentation (2/21 and 2/23)
  - First project report (Sunday 3/5)
- Second project assignment
  - Midterm presentations (4/4 and 4/6)
  - Midterm report (Sunday 4/9)
- Final project assignment
  - Final presentation (5/2 & 5/4)
  - Final report (Sunday 5/7)

Midterm Presentation Instructions

- 8-9 minute presentations (12-18 slides)
  - +2-3 minutes for written feedback and notes
- All team members should be involved during the presentation.
- The ordering of the presentations (Tuesday vs. Thursday) will be inverted based on proposal presentations.
- The presentations will be from 4:30pm – 6:15 to give everyone more time
  - Let us know if you need to leave before then

Midterm Project Report Instructions

- PART 1
  - **Research problem:** describe and motivate the research problem you are planning to work on. Explain why this problem is important for the research community and, if possible, the society in general. Define in generic terms the main computational challenges involved in this research problem.
  - **Related Work:** Present an overview of the work happening in this research area. This section should include about 12-15 citations of prior work, grouped in similar topics. Also, you should present in more details the 3-4 research papers most related to your proposed work. The related work section should end emphasizing how your proposed approach differ from previous work.
  - **Dataset and Input Modalities:** Describe the dataset(s) you are planning to use for this project. If many options exist, please motivate your choice of dataset for this research problem. Describe the input modalities and annotations available in this dataset. Specify which subset of these modalities and annotations you are planning to use.

Read carefully all the comments we gave you in the proposal reports. We will be stricter for the midterm reports.

Midterm Project Report Instructions

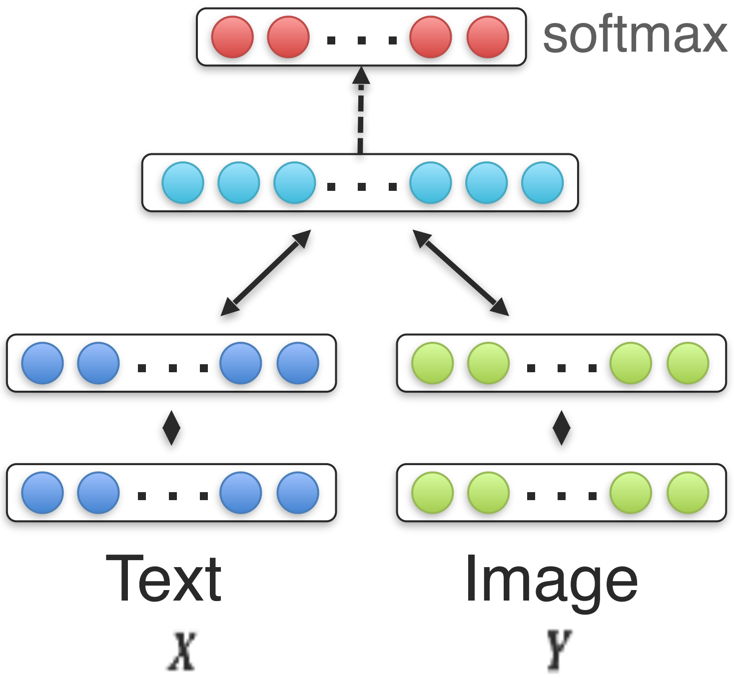
- PART 2
  - **Problem statement:** formalize mathematically your research problem. This should include the mathematical definition of the variables involved in your problem.
  - **Multimodal baseline models:** Describe mathematically at least one multimodal baseline model for your research problem.
  - **Experimental methodology:** Describe your experimental methodology for evaluating the multimodal baseline model(s).
  - **Results and Discussion:** Present in tables and/or figures your experimental results. This section should include more than re-running existing baseline models.
  - **Proposed approach:** Describe what models you are planning to test for the final report experiments. Whenever possible, you should write down the loss function of these models, following the same mathematical formulation previously used.

Quick Recap

Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

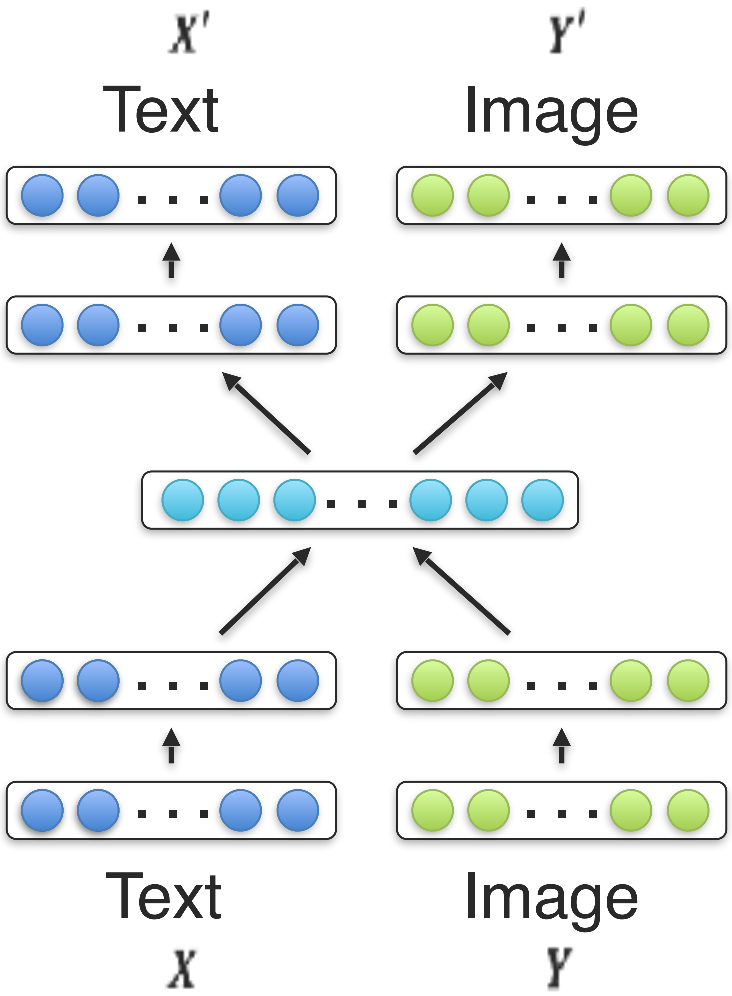
- ❑ Deep Multimodal Boltzmann machines



Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

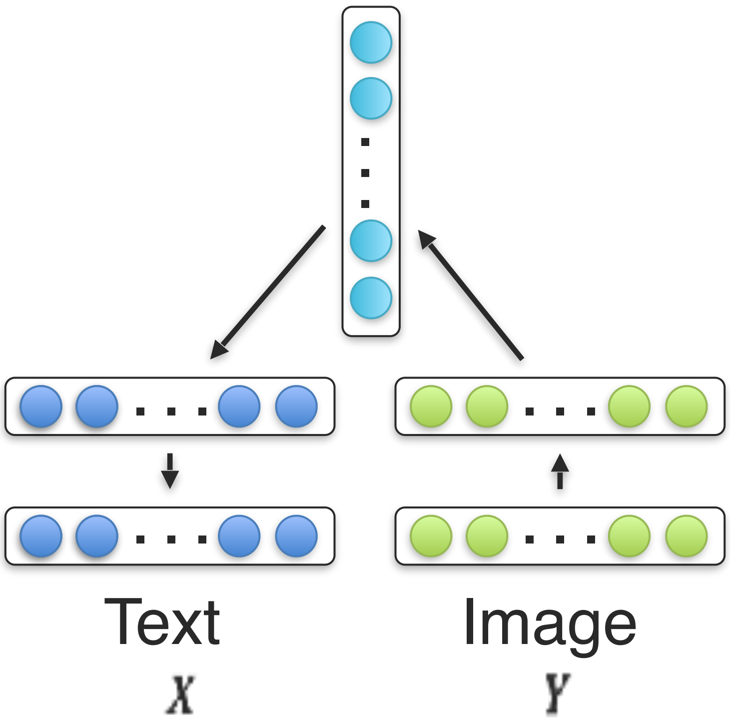
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder



Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

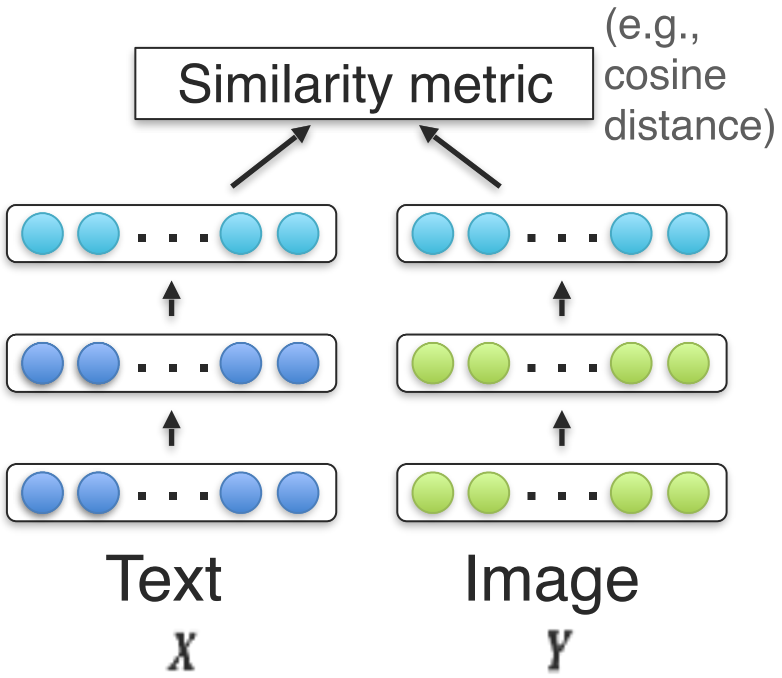
- Deep Multimodal Boltzmann machines
- Stacked Autoencoder
- Encoder-Decoder



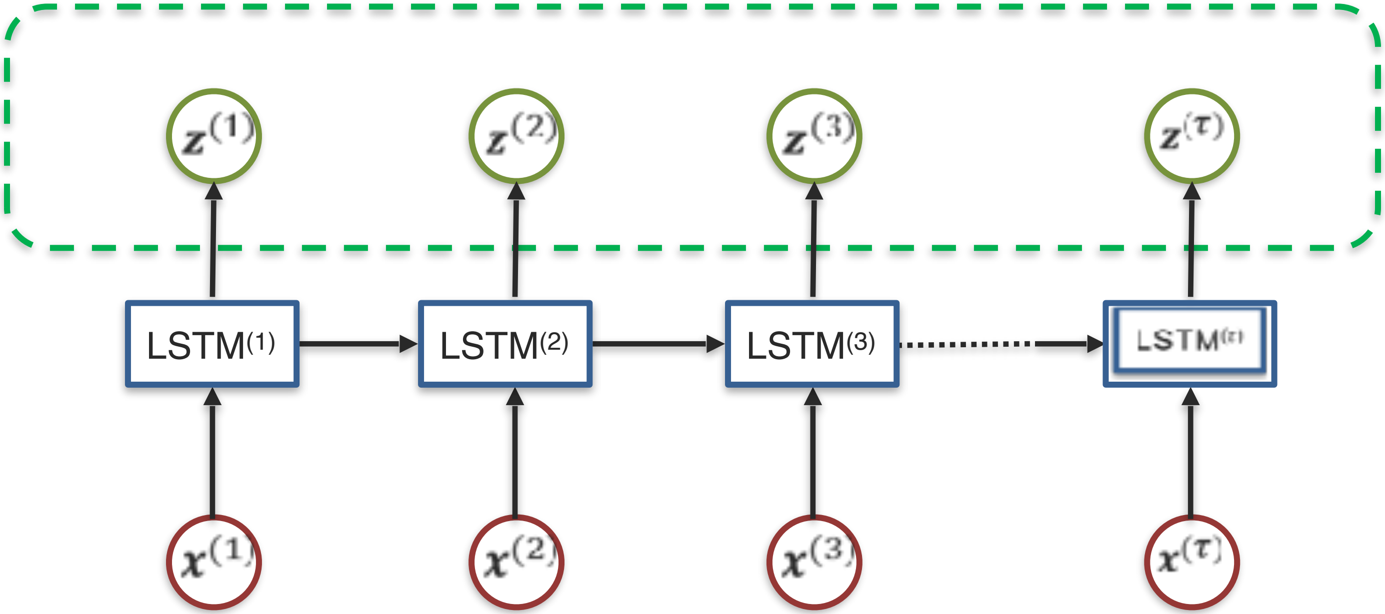
Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

- Deep Multimodal Boltzmann machines
- Stacked Autoencoder
- Encoder-Decoder
- “Minimum-distance” Multimodal Embedding



Recurrent Neural Network using LSTM Units



How can we improve reasoning by including prior domain knowledge?

Probabilistic Graphical Models

Probabilistic Graphical Model

Definition: A probabilistic graphical model (PGM) is a graph formalism for compactly modeling joint probability distributions and dependence structures over a set of random variables.

- Random variables: $X_1,\ldots,X_n$
- P is a joint distribution over $X_1,\ldots,X_n$

Can we represent P more compactly?

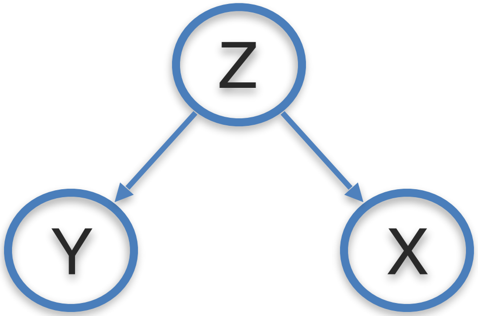
- Key: Exploit independence properties

Independent Random Variables

- Two variables X and Y are independent if
  - $P(X=x|Y=y) = P(X=x)$ for all values x,y
  - Equivalently, knowing Y does not change predictions of X
- If X and Y are independent then:
  - $P(X, Y) = P(X|Y)P(Y) = P(X)P(Y)$
- If $X_1,\ldots,X_n$ are independent then:
  - $P(X_1,\ldots,X_n) = P(X_1)\ldots P(X_n)$

Conditional Independence

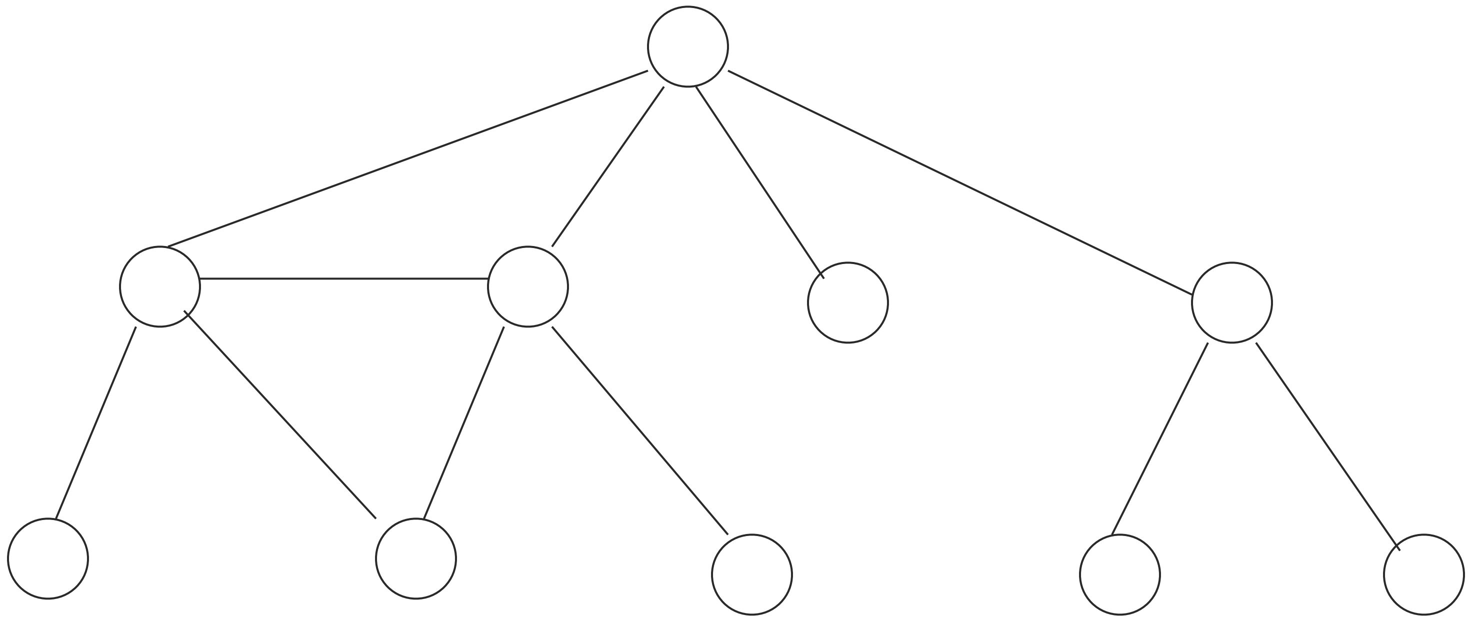
- X and Y are conditionally independent given Z if
  - $P(X=x|Y=y, Z=z) = P(X=x|Z=z)$ for all values x, y, z
  - Equivalently, if we know Z, then knowing Y does not change predictions of X



Graphical Model

- A tool that visually illustrate conditional dependence among variables in a given problem.
- Consisting of nodes (Random variables or States) and edges (Connecting two nodes, directed or undirected).
- The lack of edge represents conditional independence between variables.

Graphical Model



- Chain, Path, Cycle, Directed Acyclic Graph (DAG), Parents and Children

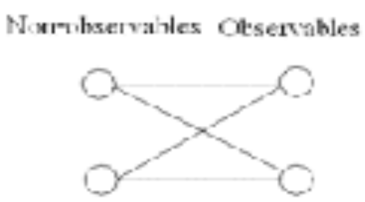
Reasoning

- The activity of guessing the state of the domain from prior knowledge and observations.

Uncertain Reasoning (Guessing)

- Some aspects of the domain are often unobservable and must be estimated indirectly through other observations.

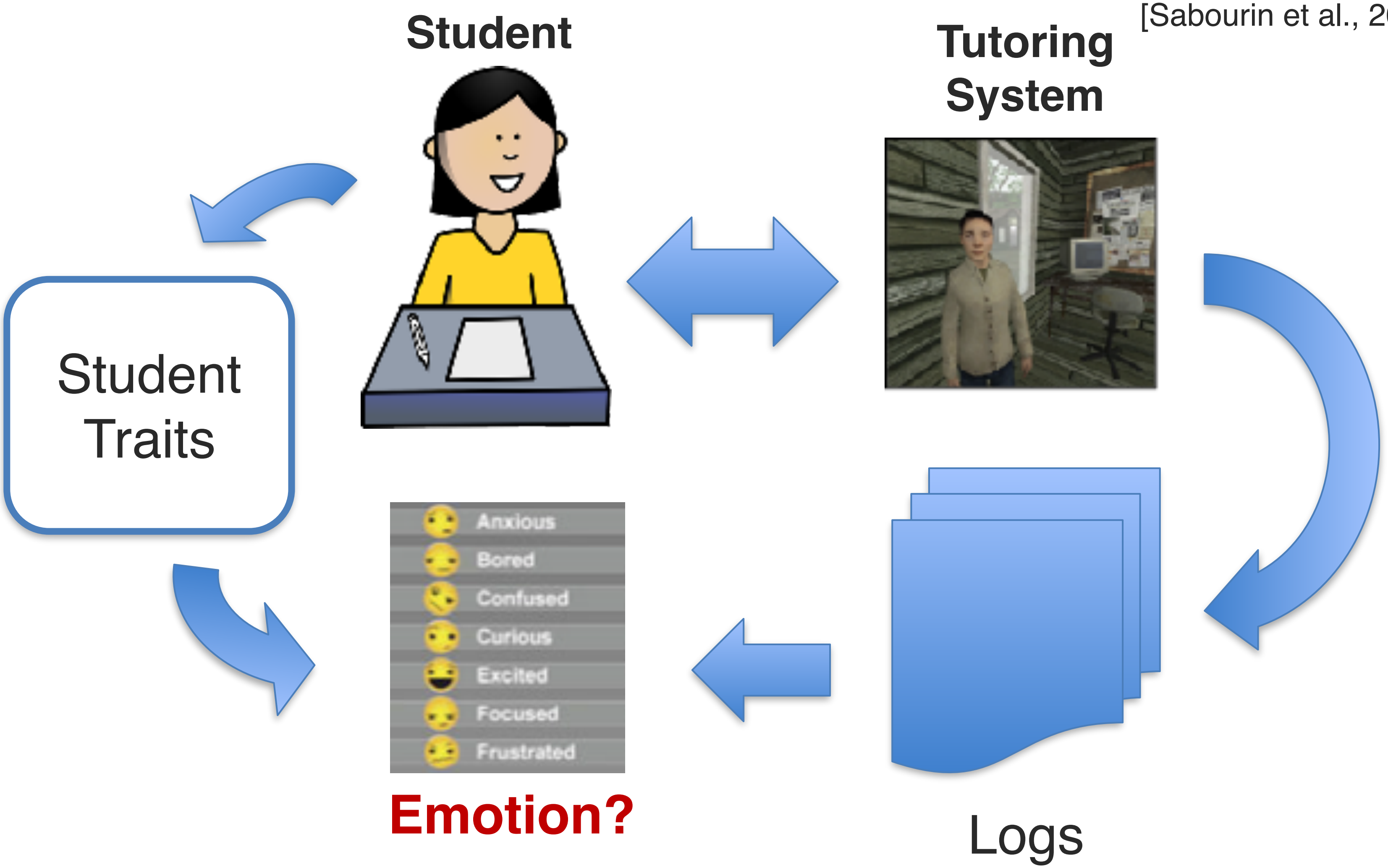
- The relationships among domain events are often uncertain, particularly the relationship between the observables and non-observables.



Developing a Graphical Model

Example: Inferring Emotion from Interaction Logs

[Sabourin et al., 2011]

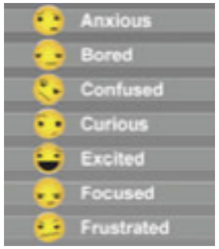


Example: Graphical Model Representation

[Sabourin et al., 2011]

Hypothesis (non-observable)

Emotion



Evidences (observable)

book views

correct ans.

notes taken

incorrect ans.

poster views

Total goals

Observable environment variables

Openness

Agreeableness

Conscientious

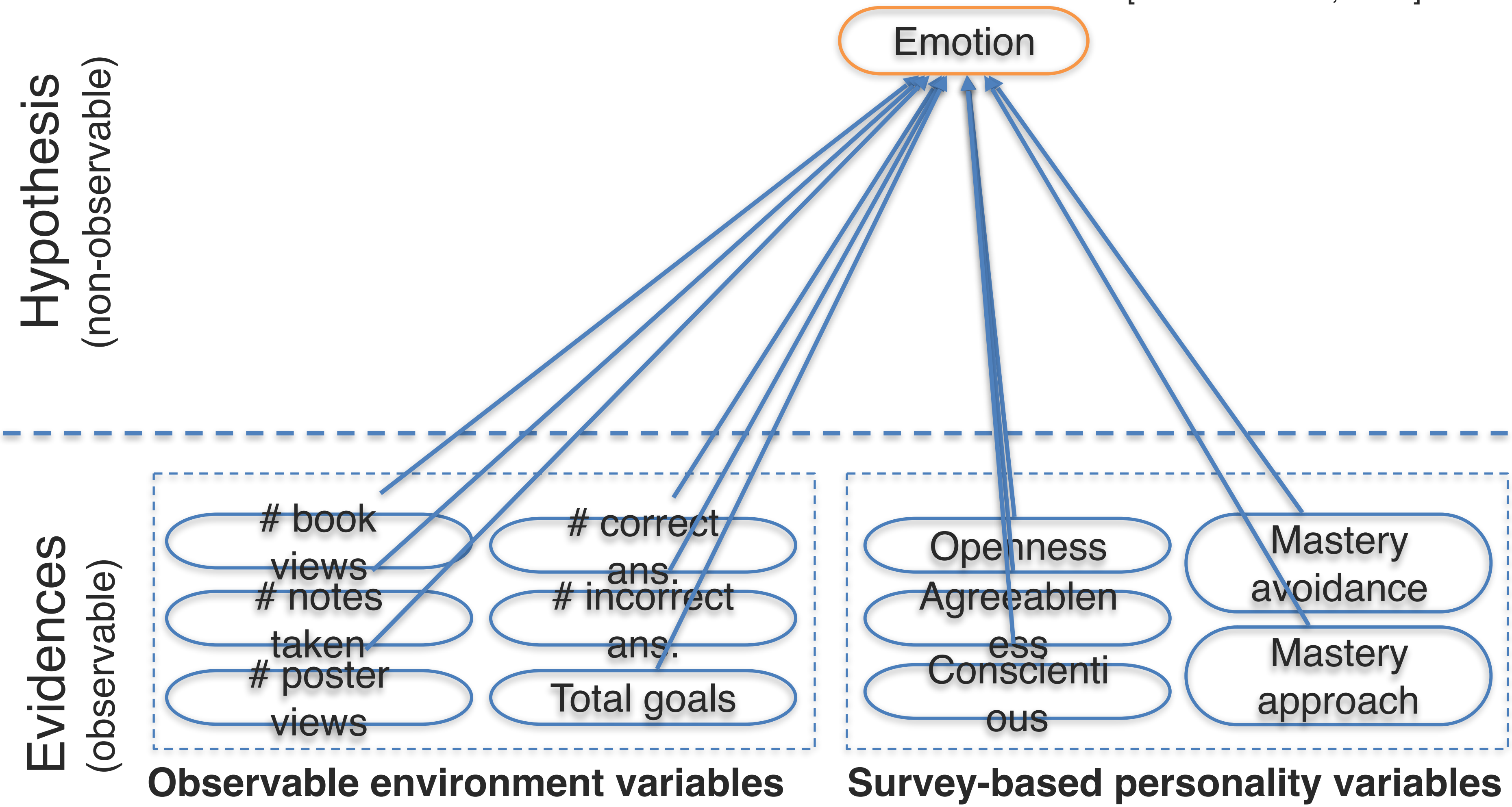
Mastery avoidance

Mastery approach

Survey-based personality variables

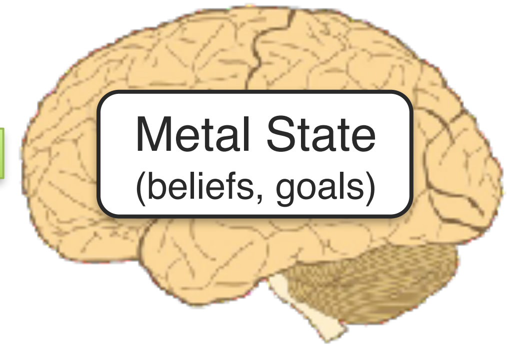
Example: Direct Prediction Approach

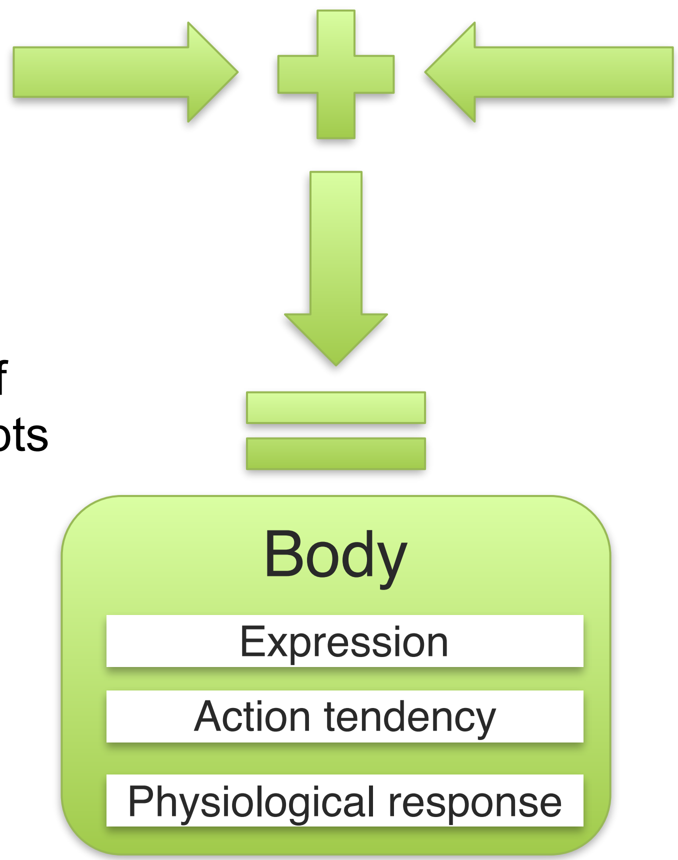
[Sabourin et al., 2011]



Appraisal Theory of Emotion

[Scherer et al., 2001]





Argues for importance of three interrelated concepts

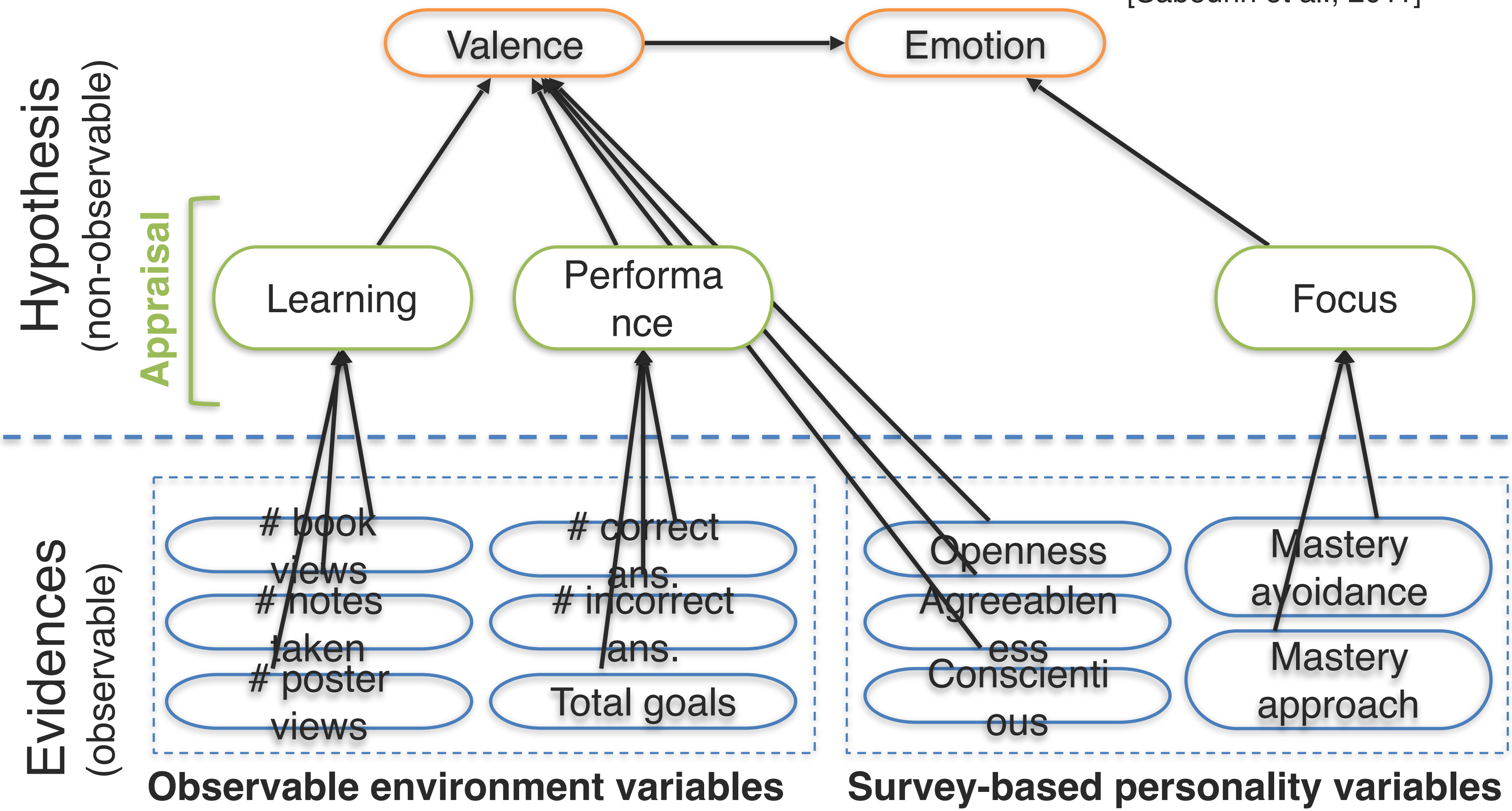
- World events
- Mental state
- Emotional Response

If we know two of these variables, we can make predictions about the third

Response= f(Env., Mind)

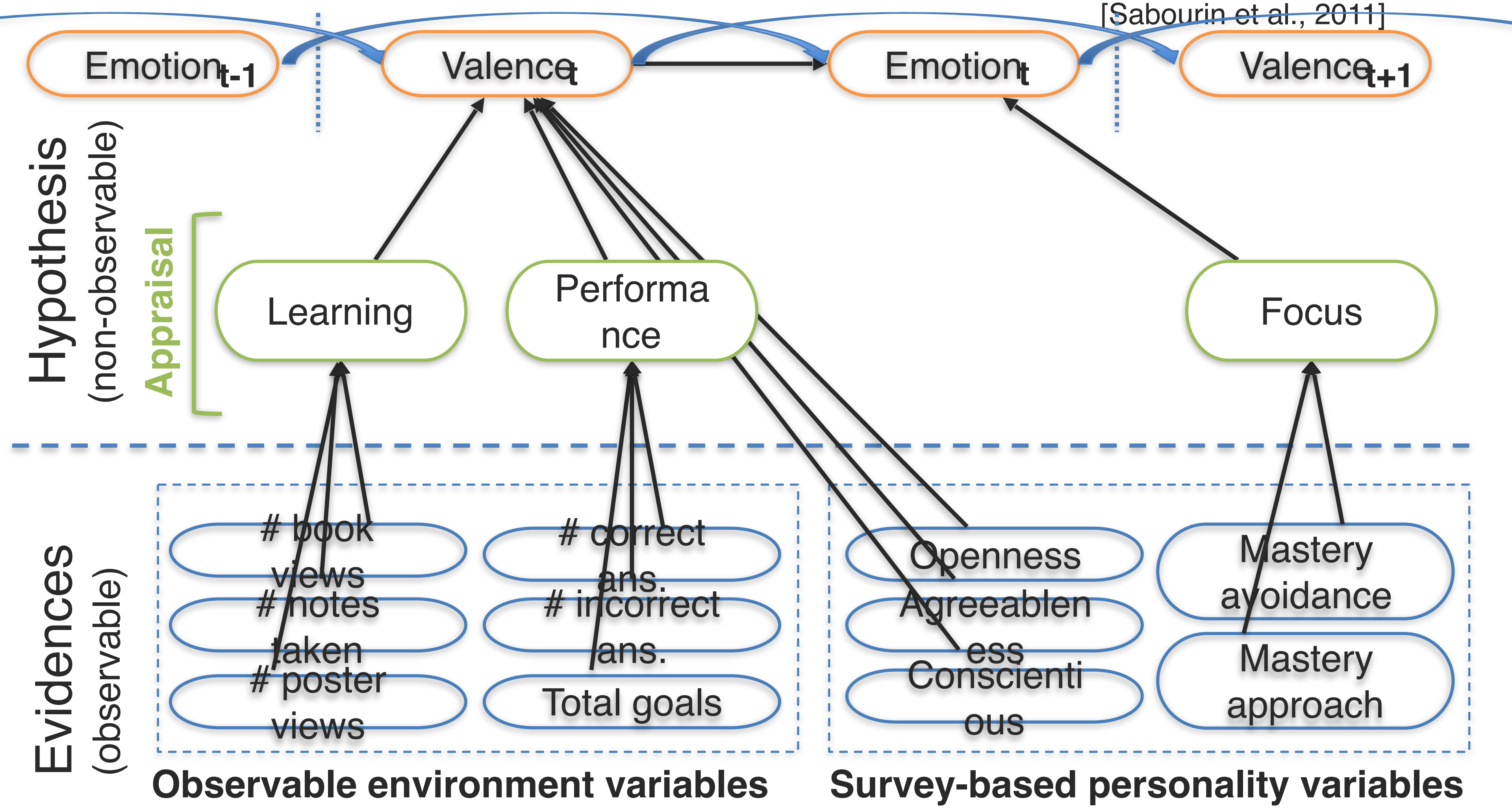
Example: Graphical Model Approach

[Sabourin et al., 2011]



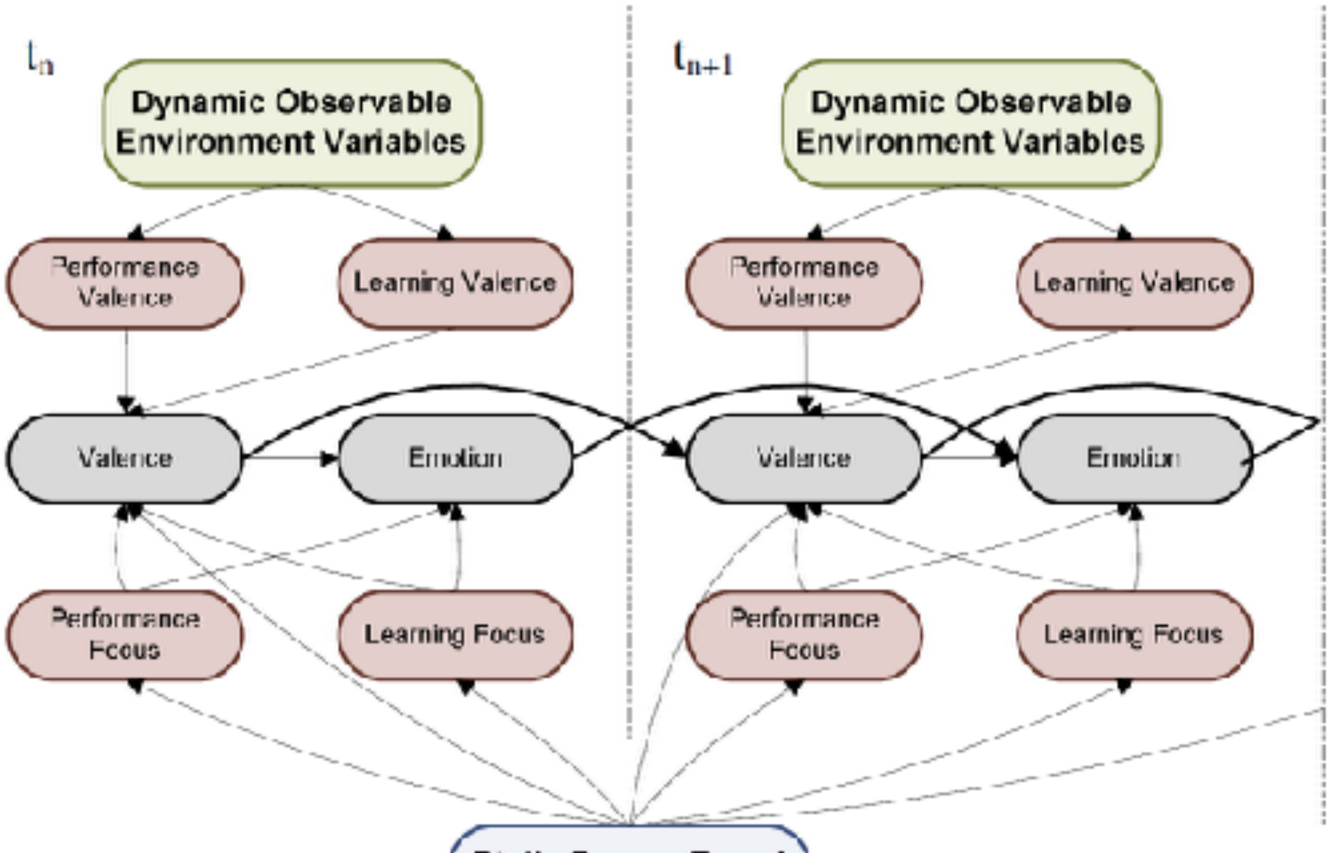
Example: Dynamic Graphical Model Approach

[Sabourin et al., 2011]



Example: Dynamic Bayesian Network Approach

[Sabourin et al., 2011]

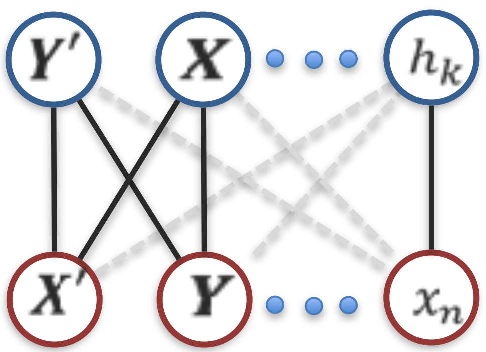


What if the “evidences” require neural network architectures to perform automatic perception?

Markov Random Fields

Restricted Boltzmann Machine (RBM)

- Undirected Graphical Model
- A generative rather than discriminative model
- Connections from every hidden unit to every visible one
- No connections across units (hence Restricted), makes it easier to train and do inference



Restricted Boltzmann Machine (RBM)

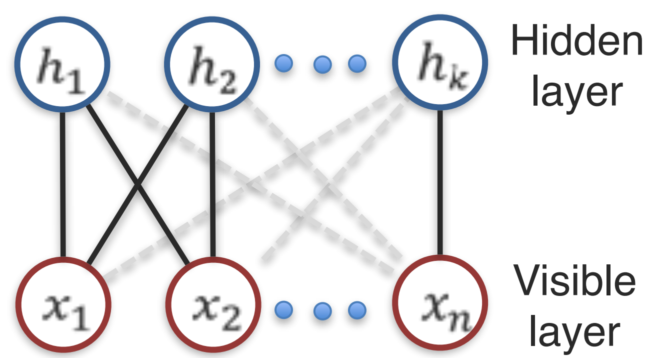
$$p(\boldsymbol{x}, \boldsymbol{h}; \theta) = \frac{\exp(-\mathrm{E}(\boldsymbol{x}, \boldsymbol{h}; \theta))}{\underbrace{\sum_{\boldsymbol{x}'} \sum_{\boldsymbol{h}'} \exp(-\mathrm{E}(\boldsymbol{x}', \boldsymbol{h}'; \theta))}_{\text{Partition function } Z}}$$

- Hidden and visible layers are binary (e.g. $\boldsymbol{x} = \{0, \dots, 1, 0, 1\}$)

- Model parameters $\theta = \{W, \boldsymbol{b}, \boldsymbol{a}\}$

$$\mathrm{E} = -\boldsymbol{x}W\boldsymbol{h} - \boldsymbol{b}\boldsymbol{x} - \boldsymbol{a}\boldsymbol{h}$$

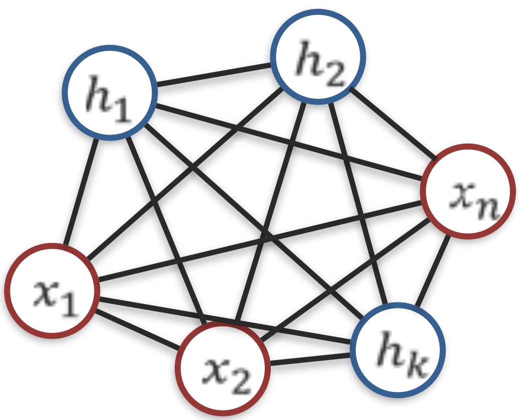
$$\mathrm{E} = \underbrace{-\sum_i \sum_j w_{i,j} x_i h_j}_{\text{Interaction term}} \underbrace{- \sum_i b_i x_i - \sum_j a_j h_j}_{\text{Bias terms}}$$

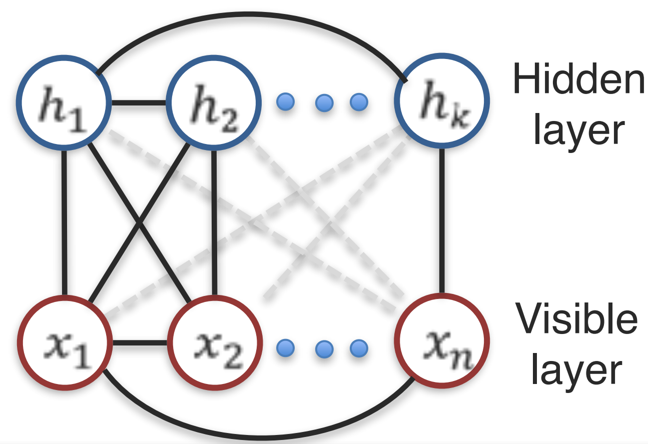


Boltzmann Machine

$$p(\boldsymbol{x}, \boldsymbol{h}; \theta) = \frac{\exp(-\mathrm{E}(\boldsymbol{x}, \boldsymbol{h}; \theta))}{\sum_{\boldsymbol{x}'} \sum_{\boldsymbol{h}'} \exp(-\mathrm{E}(\boldsymbol{x}', \boldsymbol{h}'; \theta))}$$

- Hidden and visible layers are binary (e.g. $\boldsymbol{x} = \{0, \ldots, 1, 0, 1\}$)



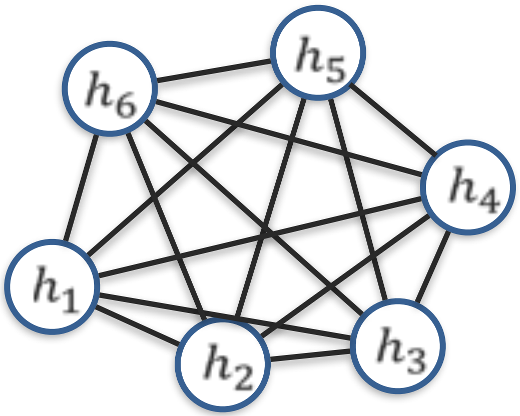


Statistical Mechanics: Boltzmann Distribution

[also called Gibbs measure]

$$p(\boldsymbol{h};\theta) = \frac{\exp(-\mathrm{E}(\boldsymbol{h};\theta)/kT)}{\sum_{\boldsymbol{h}'}\exp(-\mathrm{E}(\boldsymbol{h}';\theta)/kT)}$$

- probability distribution that gives the probability that a system will be in a certain state $\boldsymbol{h}$



$\mathrm{E}(\boldsymbol{h};\theta)$: Energy of state $\boldsymbol{h}$

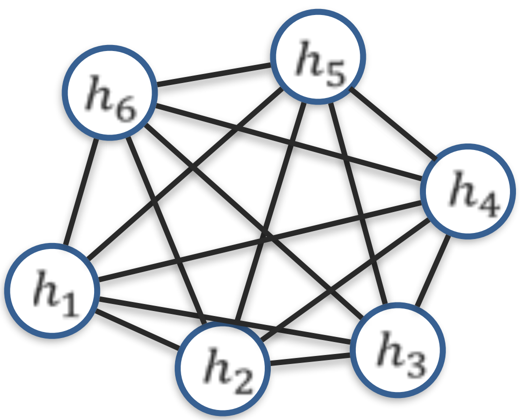
k: Boltzmann constant

T: Thermodynamic temperature

Markov Random Fields

$$p(H = \boldsymbol{h}; \theta) = \frac{\exp(-\mathrm{E}(\boldsymbol{h}; \theta))}{\sum_{\boldsymbol{h}'} \exp(-\mathrm{E}(\boldsymbol{h}'; \theta))} = \frac{\Phi(\boldsymbol{h}; \theta)}{\sum_{\boldsymbol{h}'} \Phi(\boldsymbol{h}'; \theta)}$$

- Set of random variables H having a Markov property described by undirected graph



$$\Phi(\boldsymbol{h}; \theta) = \prod_k \phi_k(\boldsymbol{h}; \theta_k)$$

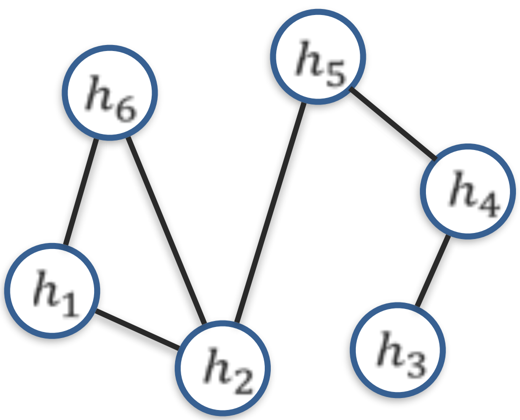
$$= \exp\left(-\sum_k \mathrm{E}_k(\boldsymbol{h}; \theta_k)\right)$$

Potential functions
$\phi_k(\boldsymbol{h}; \theta) > 0$

Markov Random Fields

$$p(H = \boldsymbol{h}; \theta) = \frac{\Phi(\boldsymbol{h}; \theta)}{\sum_{\boldsymbol{h}'} \Phi(\boldsymbol{h}'; \theta)} = \frac{\sum_k \phi_k(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\sum_{\boldsymbol{y}'} \sum_k \phi_k(\boldsymbol{y}', \boldsymbol{x}; \theta)}$$

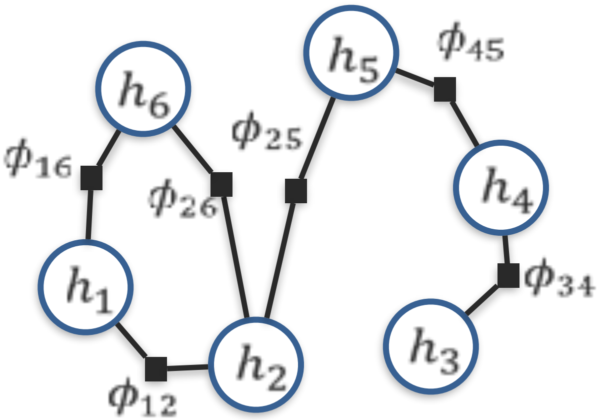
$$\begin{aligned}\Phi(\boldsymbol{h}; \theta) = {} & \phi_{12}(h_1, h_2; \theta_{12}) \times \\ & \phi_{16}(h_1, h_6; \theta_{16}) \times \\ & \phi_{26}(h_2, h_6; \theta_{26}) \times \\ & \phi_{25}(h_2, h_5; \theta_{25}) \times \\ & \phi_{45}(h_4, h_5; \theta_{45}) \times \\ & \phi_{34}(h_3, h_4; \theta_{34})\end{aligned}$$



Markov Random Fields: Factor Graphs

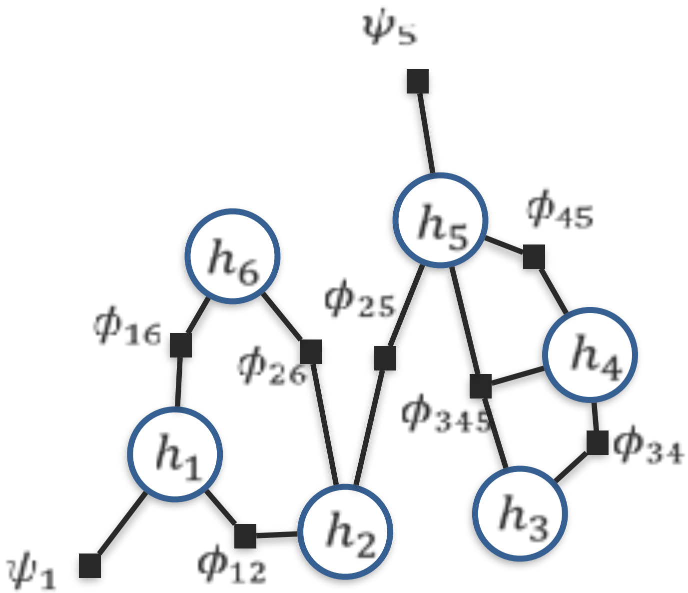
$$p(H = \boldsymbol{h}; \theta) = \frac{\Phi(\boldsymbol{h}; \theta)}{\sum_{\boldsymbol{h}'} \Phi(\boldsymbol{h}'; \theta)} = \frac{\sum_k \phi_k(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\sum_{\boldsymbol{y}'} \sum_k \phi_k(\boldsymbol{y}', \boldsymbol{x}; \theta)}$$

$$\begin{aligned}\Phi(\boldsymbol{h}; \theta) = \; & \phi_{12}(h_1, h_2; \theta_{12}) \times \\ & \phi_{16}(h_1, h_6; \theta_{16}) \times \\ & \phi_{26}(h_2, h_6; \theta_{26}) \times \\ & \phi_{25}(h_2, h_5; \theta_{25}) \times \\ & \phi_{45}(h_4, h_5; \theta_{45}) \times \\ & \phi_{34}(h_3, h_4; \theta_{34})\end{aligned}$$



Markov Random Fields (Factor Graphs)

$$p(H = \boldsymbol{h}; \theta) = \frac{\Phi(\boldsymbol{h}; \theta)}{\sum_{\boldsymbol{h}'} \Phi(\boldsymbol{h}'; \theta)} = \frac{\sum_k \phi_k(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\sum_{\boldsymbol{y}'} \sum_k \phi_k(\boldsymbol{y}', \boldsymbol{x}; \theta)}$$



$$\begin{aligned}\Phi(\boldsymbol{h}; \theta) = {} & \phi_{12}(h_1, h_2; \theta_{12}) \times \\ & \phi_{16}(h_1, h_6; \theta_{16}) \times \\ & \phi_{26}(h_2, h_6; \theta_{26}) \times \\ & \phi_{25}(h_2, h_5; \theta_{25}) \times \\ & \phi_{45}(h_4, h_5; \theta_{45}) \times \\ & \phi_{34}(h_3, h_4; \theta_{34}) \times \\ & \psi_1(h_1; \theta_1) \times \psi_5(h_5; \theta_5) \\ & \times \phi_{345}(h_3, h_4, h_5; \theta_{345})\end{aligned}$$

pairwise potentials

Unary potentials

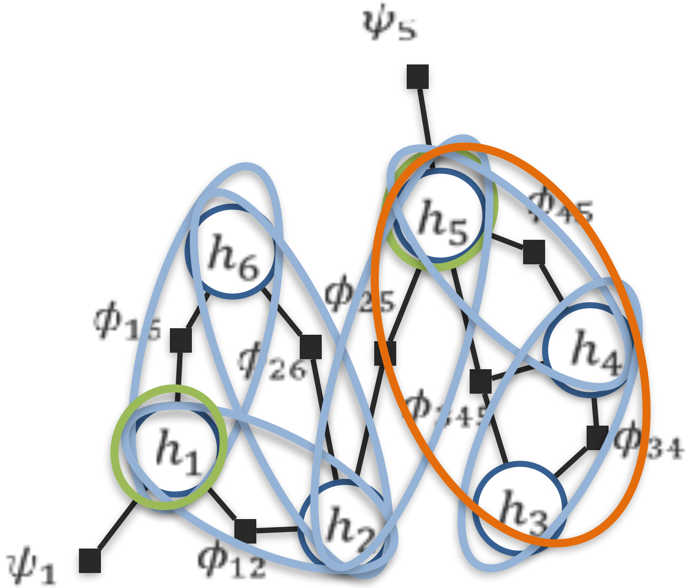
Markov Random Fields – Clique Factorization

$$p(H = \boldsymbol{h}; \theta) = \frac{\Phi(\boldsymbol{h}; \theta)}{\sum_{\boldsymbol{h}'} \Phi(\boldsymbol{h}'; \theta)} = \frac{\sum_k \phi_k(\boldsymbol{y}, \boldsymbol{x}; \theta)}{\sum_{\boldsymbol{y}'} \sum_k \phi_k(\boldsymbol{y}', \boldsymbol{x}; \theta)}$$

Clique factorization

$$\Phi(\boldsymbol{h}; \theta) = \underbrace{\phi_{12}(h_1, h_2; \theta_{12}) \times \phi_{16}(h_1, h_6; \theta_{16}) \times \phi_{26}(h_2, h_6; \theta_{26}) \times \phi_{25}(h_2, h_5; \theta_{25}) \times \phi_{45}(h_4, h_5; \theta_{45}) \times \phi_{34}(h_3, h_4; \theta_{34})}_{\text{pairwise potentials}} \times$$

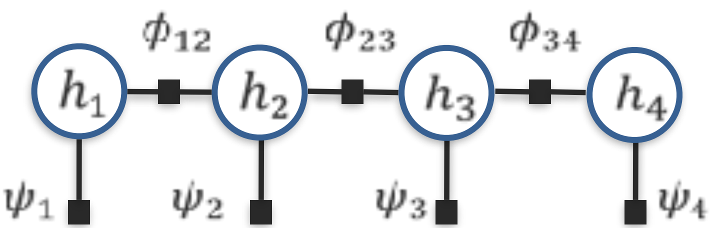
$$\underbrace{\psi_1(h_1; \theta_1) \times \psi_5(h_5; \theta_5)}_{\text{Unary potentials}} \times \phi_{345}(h_3, h_4, h_5; \theta_{345})$$



Chain Markov Random Fields (Factor Graphs)

$$p(H=\boldsymbol{h};\theta)=\frac{\Phi(\boldsymbol{h};\theta)}{\sum_{\boldsymbol{h}'}\Phi(\boldsymbol{h}';\theta)}=\frac{\sum_k \phi_k(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\sum_k \phi_k(\boldsymbol{y}',\boldsymbol{x};\theta)}$$

$$\begin{aligned}\Phi(\boldsymbol{h};\theta) = & \left.\begin{aligned}&\phi_{12}(h_1,h_2;\theta_{12})\times\\ &\phi_{23}(h_2,h_3;\theta_{23})\times\\ &\phi_{34}(h_3,h_4;\theta_{34})\times\end{aligned}\right\}\text{pairwise potentials}\\ & \left.\begin{aligned}&\psi_1(h_1;\theta_1)\times\\ &\psi_2(h_2;\theta_2)\times\\ &\psi_3(h_3;\theta_3)\times\\ &\psi_4(h_4;\theta_4)\end{aligned}\right\}\text{Unary potentials}\end{aligned}$$



Conditional Random Fields

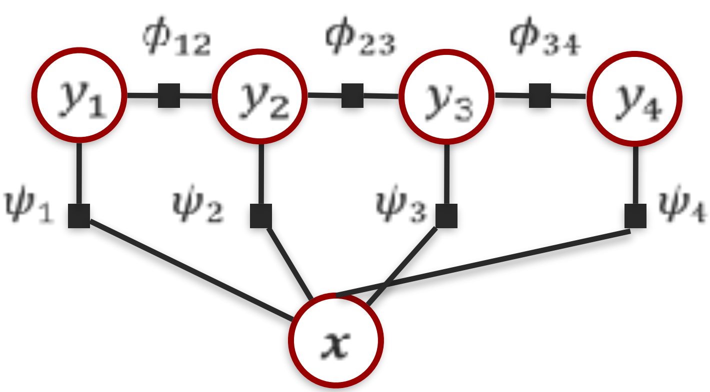
Conditional Random Fields (Factor Graphs)

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{\Phi(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\Phi(\boldsymbol{y}',\boldsymbol{x};\theta)} = \frac{\sum_k \phi_k(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\sum_k \phi_k(\boldsymbol{y}',\boldsymbol{x};\theta)}$$

$$\begin{aligned}\Phi(y,\boldsymbol{x};\theta) = \; & \phi_{12}(y_1,y_2,\boldsymbol{x};\theta_{12}) \times \\ & \phi_{23}(y_2,y_3,\boldsymbol{x};\theta_{23}) \times \\ & \phi_{34}(y_3,y_4,\boldsymbol{x};\theta_{34}) \times \\ & \psi_1(y_1,\boldsymbol{x};\theta_1) \times \\ & \psi_2(y_2,\boldsymbol{x};\theta_2) \times \\ & \psi_3(y_3,\boldsymbol{x};\theta_3) \times \\ & \psi_4(y_4,\boldsymbol{x};\theta_4)\end{aligned}$$

pairwise potentials

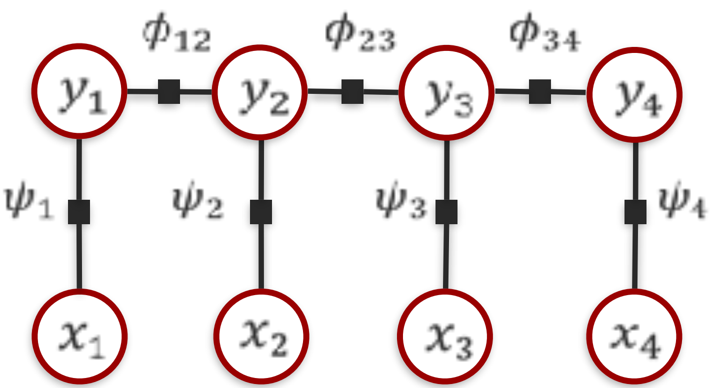
Unary potentials



Conditional Random Fields (Factor Graphs)

$$p(\mathbf{y}|\mathbf{x};\theta) = \frac{\Phi(\mathbf{y},\mathbf{x};\theta)}{\sum_{\mathbf{y}'}\Phi(\mathbf{y}',\mathbf{x};\theta)} = \frac{\sum_k \phi_k(\mathbf{y},\mathbf{x};\theta)}{\sum_{\mathbf{y}'}\sum_k \phi_k(\mathbf{y}',\mathbf{x};\theta)}$$

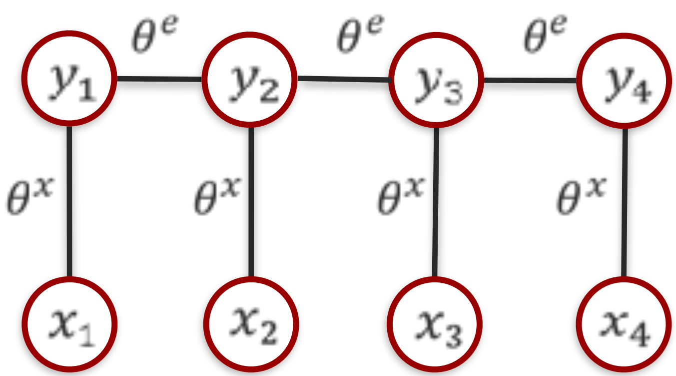
$$\begin{aligned}\Phi(\mathbf{y},\mathbf{x};\theta) = {} & \left.\begin{aligned}&\phi_{12}(y_1,y_2,\mathbf{x};\theta_{12}) \times \\ &\phi_{23}(y_2,y_3,\mathbf{x};\theta_{23}) \times \\ &\phi_{34}(y_3,y_4,\mathbf{x};\theta_{34}) \times\end{aligned}\right\} \text{pairwise potentials} \\ & \left.\begin{aligned}&\psi_1(y_1,x_1;\theta_1) \times \\ &\psi_2(y_2,x_2;\theta_2) \times \\ &\psi_3(y_3,x_3;\theta_3) \times \\ &\psi_4(y_4,x_4;\theta_4)\end{aligned}\right\} \text{Unary potentials}\end{aligned}$$



Conditional Random Fields (Log-linear Model)

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{\Phi(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\Phi(\boldsymbol{y}',\boldsymbol{x};\theta)} = \frac{\sum_k \phi_k(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\sum_k \phi_k(\boldsymbol{y}',\boldsymbol{x};\theta)}$$

$$= \frac{\exp(\sum_k \theta_k f_k(\boldsymbol{y},\boldsymbol{x}))}{\sum_{\boldsymbol{y}'}\exp\left(\sum_k \theta_k f_k(\boldsymbol{y}',\boldsymbol{x})\right)}$$



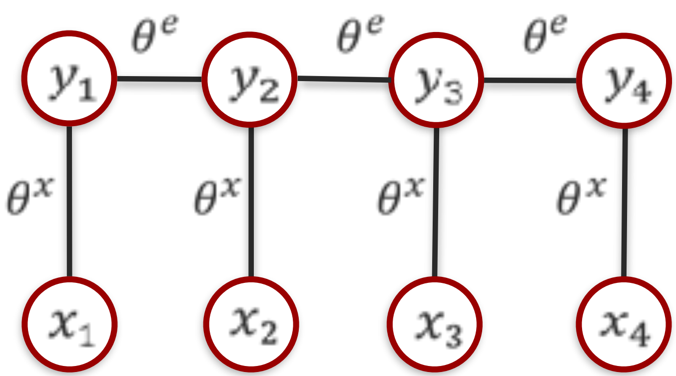
$f_k(\boldsymbol{y},\boldsymbol{x})$: feature function

- Pairwise feature function
  $f_k(y_i, y_j, \boldsymbol{x}; \theta^e)$
- Unary feature function
  $f_k(y_i, \boldsymbol{x}; \theta^x)$

Learning Parameters of a CRF Model

$$\underset{\hat{y}}{\operatorname{argmax}} \log\big(p(\boldsymbol{y}|\boldsymbol{x};\theta)\big)$$

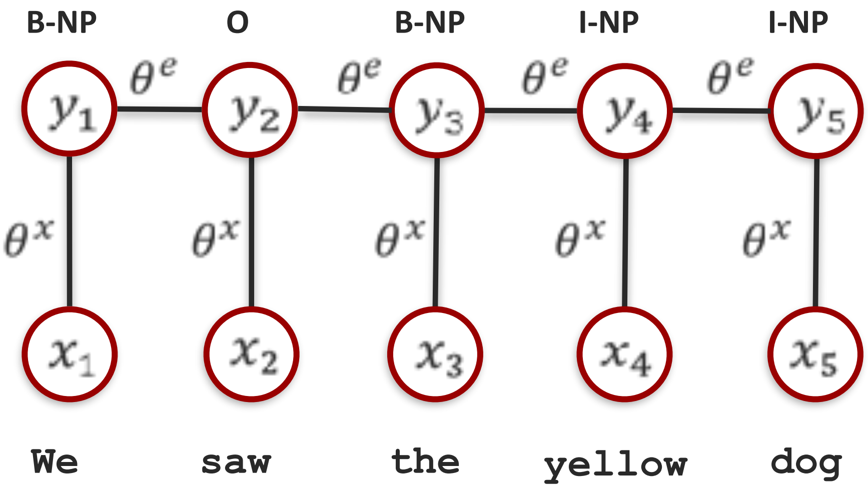
- Gradient can be computed analytically
  - Inference of marginal probabilities using belief propagation (or loopy belief propagation for cyclic graphs)
- Optimized with stochastic or batch approaches

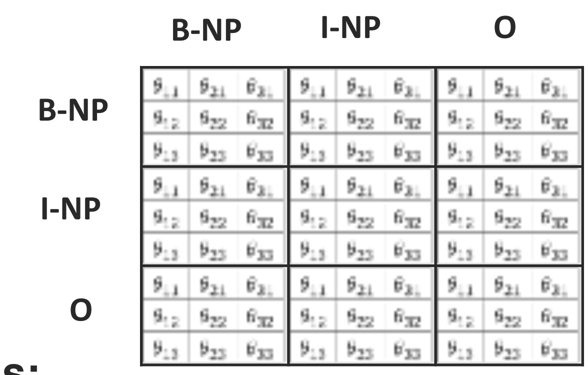


CRFs for Shallow Parsing

$$p(\boldsymbol{y}|\boldsymbol{x};\theta) = \frac{\Phi(\boldsymbol{y},\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}'}\Phi(\boldsymbol{y}',\boldsymbol{x};\theta)} = \frac{\exp(\sum_k \theta_k f_k(\boldsymbol{y},\boldsymbol{x}))}{\sum_{\boldsymbol{y}'}\exp\left(\sum_k \theta_k f_k(\boldsymbol{y}',\boldsymbol{x})\right)}$$

- How many θ^x parameters?
- What did θ^x learn?
- What did θ^e learn?





Labels:

B-NP: Beginning of a noun phrase

I-NP: Continuation of a noun phrase

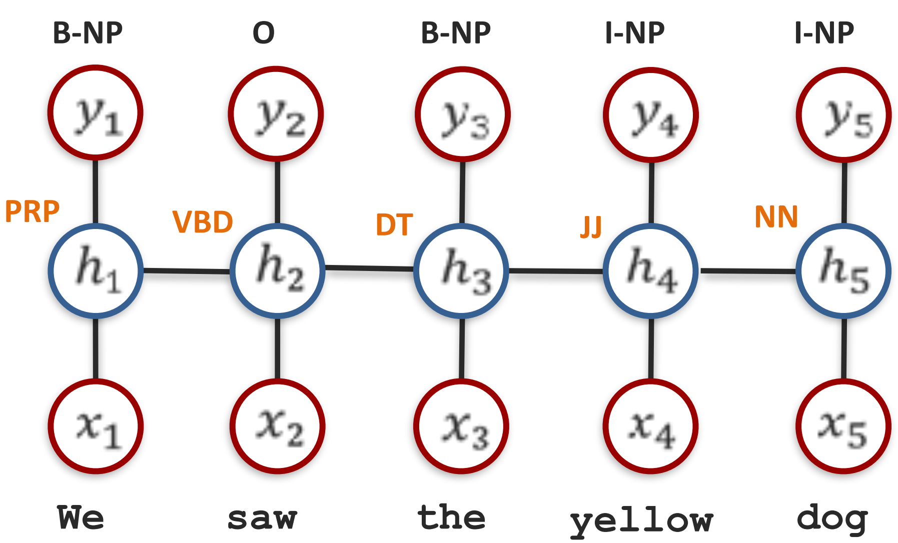
O: Outside a noun phrase

Dictionary size: 10,000 words

Latent-Dynamic CRF

$$p(\mathbf{y}|\mathbf{x};\theta) = \sum_{\mathbf{h}} p(\mathbf{y}|\mathbf{h};\boldsymbol{\theta})p(\mathbf{h}|\mathbf{x};\boldsymbol{\theta})$$

$$= \sum_{\mathbf{h}:\forall h_t \in \mathcal{H}_{y_t}} p(\mathbf{h}|\mathbf{x};\theta) \qquad = \sum_{\mathbf{h}:\forall h_t \in \mathcal{H}_{y_t}} \frac{\Phi(\mathbf{h},\mathbf{x};\theta)}{\sum_{\mathbf{h}'}\Phi(\mathbf{h}',\mathbf{x};\theta)}$$



Latent variables (e.g., POS tags)

$\mathbf{h} = \{h_1, h_2, h_3, \ldots, h_t\}$ where $h_t \in \{\mathcal{H}_{y_t}\}$

For example:

$\mathcal{H} = \{\mathcal{H}_{B-NP}, \mathcal{H}_{I-NP}, \mathcal{H}_{O}\}$

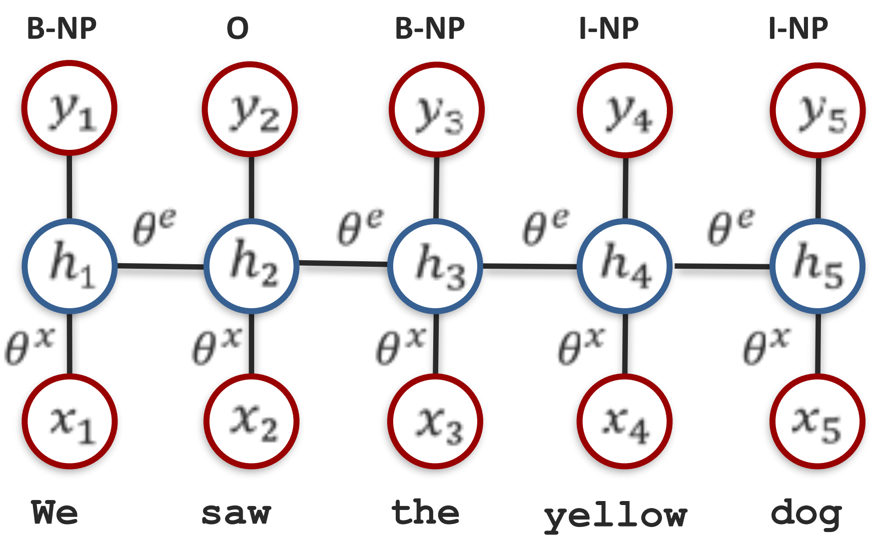
$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$

Dictionary size: 10,000 words

Latent-Dynamic CRF

$$p(y|x;\theta) = \sum_{h: \forall h_j \in \mathcal{H}_{y_j}} \frac{\exp(\sum_k \theta_k f_k(h,x))}{\sum_{h'} \exp(\sum_k \theta_k f_k(h',x))}$$

- How many θ^x parameters?
- What did θ^x learn?
- How many θ^e parameters?
- What did θ^e learn?



- Intrinsic dynamics
- Extrinsic dynamics

Latent variables (e.g., POS tags)

$h = \{h_1, h_2, h_3, \ldots, h_\tau\}$ where $h_\tau \in \{\mathcal{H}_{y_\tau}\}$

For example:

$\mathcal{H} = \{\mathcal{H}_{B-NP}, \mathcal{H}_{I-NP}, \mathcal{H}_{O}\}$

$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$

Dictionary size: 10,000 words

Latent-Dynamic CRF for Shallow Parsing

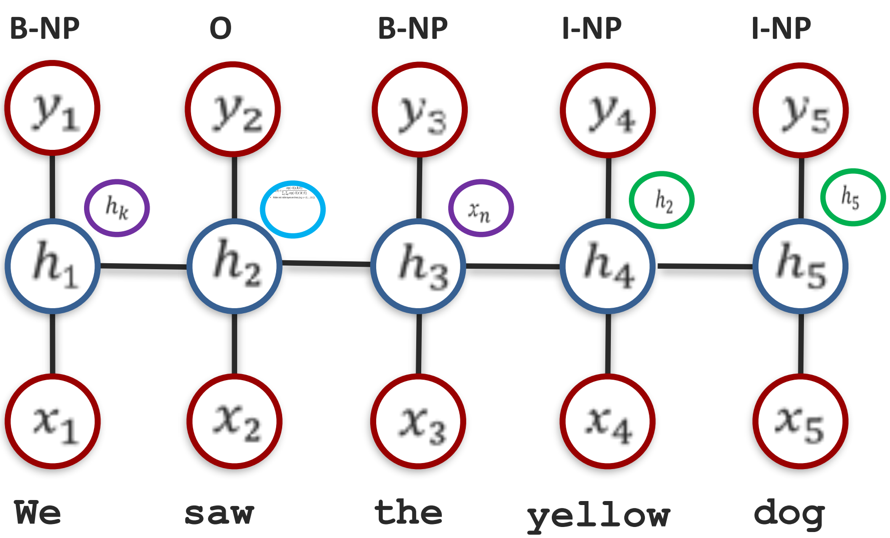
Experiment – Analyzing latent variables

- **Task:** Shallow parsing with CoNLL 2000 dataset
- **Input features:** word feature only
- **Output labels:** Noun phrase labels

1) Select hidden state a^* with highest marginal:

$$a^* = \arg\max_{a} p(h_t = a | x; \theta)$$

2) Compute relative frequency for each word



Label	State	Words	POS	Freq.
		That	WDT	0.85
		who	WP	0.49
		Who	WP	0.33
		any	DT	1.00
		an	DT	1.00
		a	DT	0.98
		They	PRP	1.00
		we	PRP	1.00
		he	PRP	1.00
		Nasdaq	NNP	1.00
		Florida	NNP	0.99
		cities	NNS	0.99

Label	State	Words	POS	Freq.
		but	CC	0.88
		by	IN	0.73
		or	IN	0.67
		4.6	CD	1.00
		1	CD	1.00
		1 1	CD	0.62
		were	VBD	0.94
		rose	VBD	0.93
		have	VBP	0.92
		been	VBN	0.97
		be	VB	0.94
		to	TO	0.92

Latent variables (e.g., POS tags)

$$h = \{h_1, h_2, h_3, \ldots, h_t\} \qquad \text{where } h_t \in \{\mathcal{H}_{y_t}\}$$

For example:

$$\mathcal{H} = \{\mathcal{H}_{B-NP}, \mathcal{H}_{I-NP}, \mathcal{H}_{O}\}$$

$$\mathcal{H} = \{B_1, B_2, B_3, B_4, I_1, I_2, I_3, I_4, O_1, O_2, O_3, O_4\}$$

Dictionary size: 10,000 words

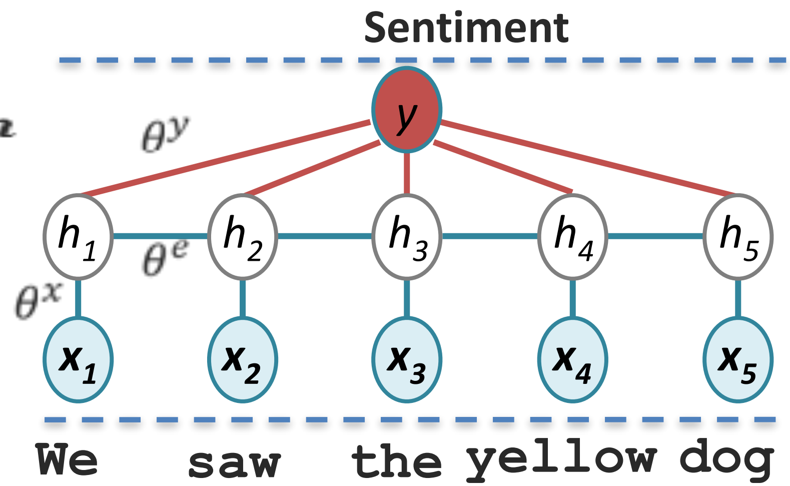
Hidden Conditional Random Field

Sequence label:

k: Boltzmann constant $\mathrm{E}(\boldsymbol{h};\theta)$: Energy of state $\boldsymbol{h}$

Latent variables with shared hidden states:

$\boldsymbol{h} = \{h_1, h_2, h_3, \ldots, h_t\}$ $\Phi(\boldsymbol{h};\theta) = \prod_k \phi_k(\boldsymbol{h};\theta_k)$



$$p(H = \boldsymbol{h};\theta) = \frac{\exp(-\mathrm{E}(\boldsymbol{h};\theta))}{\sum_{\boldsymbol{h}'}\exp(-\mathrm{E}(\boldsymbol{h}';\theta))}$$

T: Thermodynamic te

- Inference is tractable: $O(YH^2T)$
  - Linear in sequence length T !
- Parameter learning $(\theta^x, \theta^e, \theta^y)$:
  - Gradient descent or L-BFGS

Shared hidden states

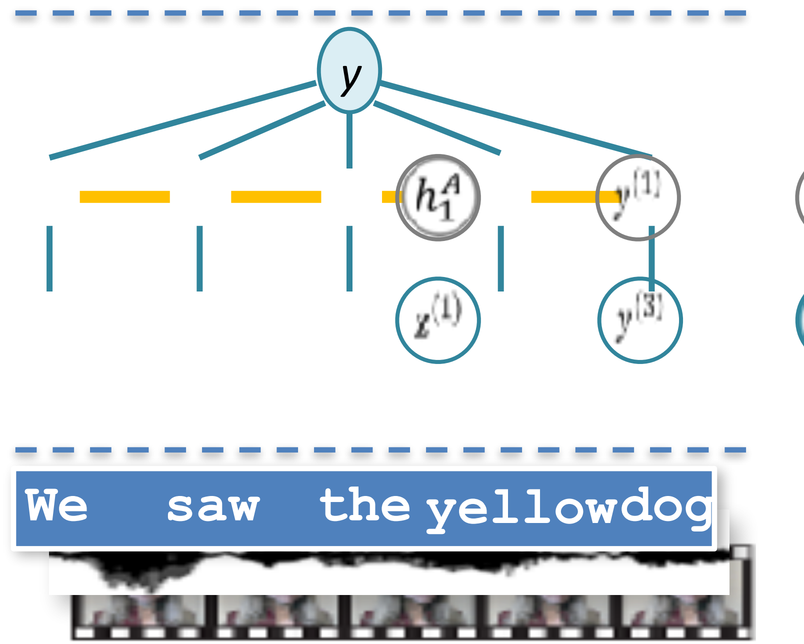
Learning Multimodal Structure

Modality-*private* structure

- Internal grouping of observations

Modality-*shared* structure

- Interaction and synchrony

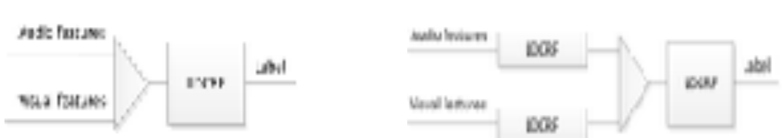


Early / Late fusion is inappropriate

- Early – strong modality can dominate
- Late – cross-modality dependency is discarded

Early fusion

Late fusion



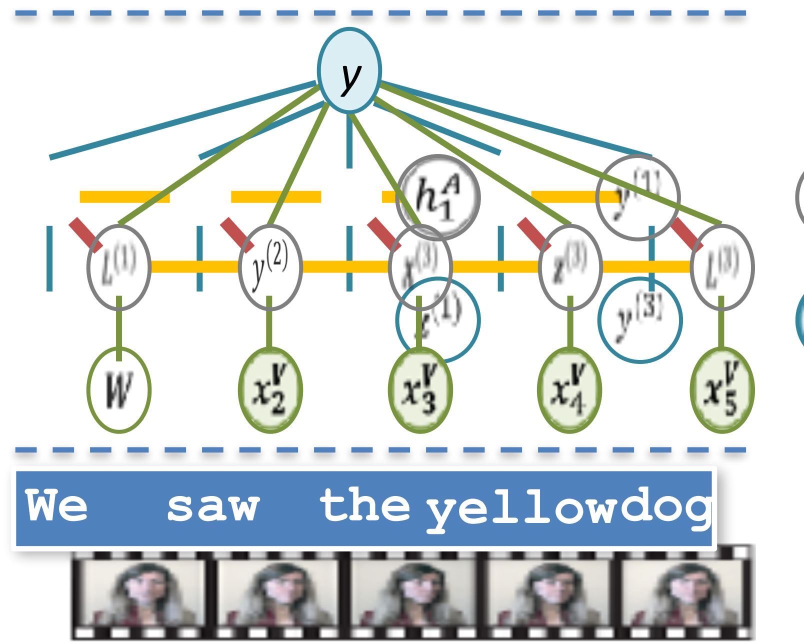
Multi-view Latent Variable Discriminative Models

Modality-*private* structure

- Internal grouping of observations

Modality-*shared* structure

- Interaction and synchrony



Early / Late fusion is inappropriate

- Early – strong modality can dominate
- Late – cross-modality dependency is discarded

$$p(y|\,\boldsymbol{x}^A, \boldsymbol{x}^V; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}^A, \boldsymbol{h}^V} p\left(y, \boldsymbol{h}^A, \boldsymbol{h}^V \middle| \boldsymbol{x}^A, \boldsymbol{x}^V; \boldsymbol{\theta}\right)$$

➢ Approximate inference using loopy-belief

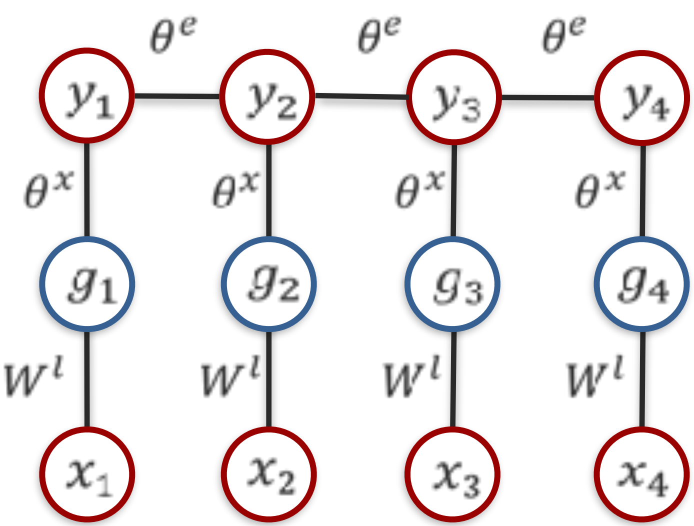
CRFs and Deep Learning

Conditional Neural Fields

$$\mathcal{G}^l(x_i, W^l) = [g_1^l(x_i \cdot W_1^l), g_2^l(x_i \cdot W_i^l), \ldots, g_n^l(x_i \cdot W_n^l)]$$

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \propto \exp\left\{\sum_i \boldsymbol{\theta}^x \cdot \boldsymbol{f}^x(y_i, \mathbf{x}_i) + \sum_i \boldsymbol{\theta}^e \cdot \boldsymbol{f}^e(y_i, y_{i-1})\right\}$$

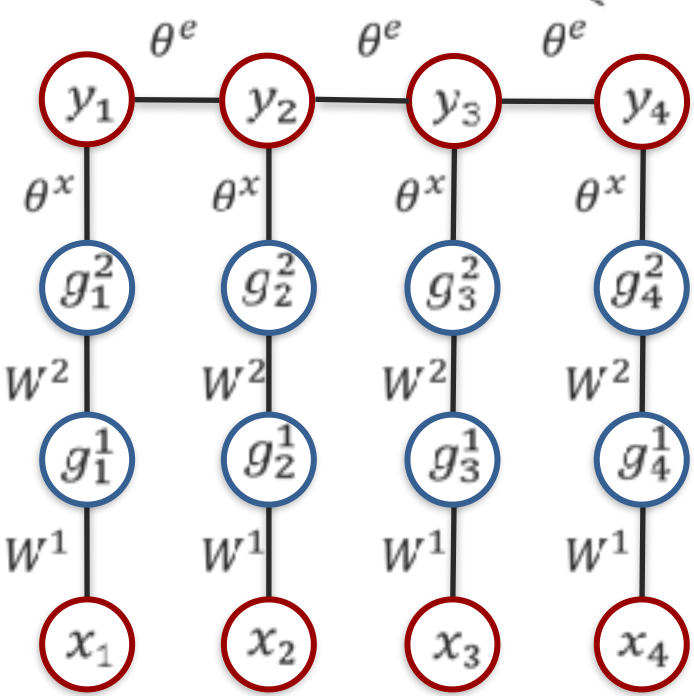
$$f^x(y_i, \mathbf{x}_i) = \mathbb{I}[y_i = y'] \cdot \mathcal{G}(\mathbf{x}_i, \boldsymbol{W}^l)$$



Deep Conditional Neural Fields

$$\mathcal{G}^l(x_i, W^l) = [g_1^l(x_i \cdot W_1^l), g_2^l(x_i \cdot W_i^l), \ldots, g_n^l(x_i \cdot W_n^l)]$$

$$p(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) \propto \exp\left\{\sum_i \boldsymbol{\theta}^x \cdot \boldsymbol{f}^x(y_i, \mathbf{x}_i) + \sum_i \boldsymbol{\theta}^e \cdot \boldsymbol{f}^e(y_i, y_{i-1})\right\}$$



$$f^x(y_i, \mathbf{x}_i) = \mathbb{I}[y_i = y'] \cdot \mathcal{G}(\boldsymbol{a}_i^{m-1}, \boldsymbol{W}^l)$$

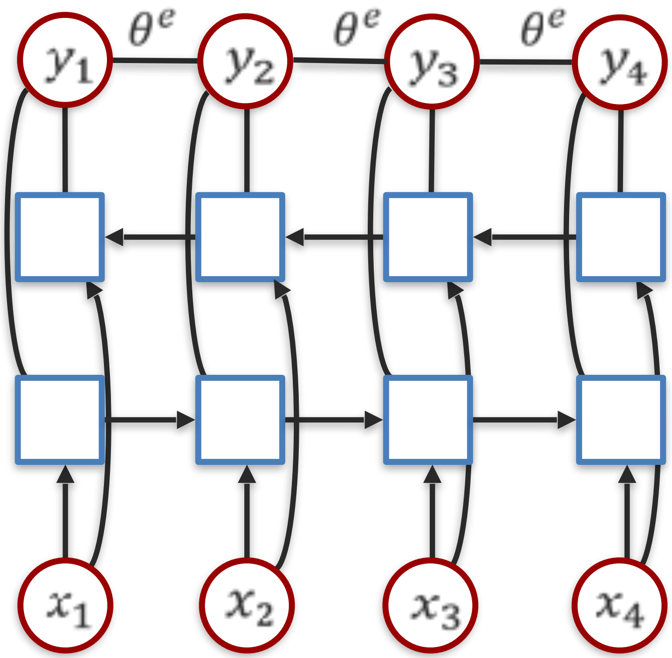
$$a^l = \mathcal{G}(\boldsymbol{a}_i^{l-1}, \boldsymbol{\theta}^g) \quad \text{for } l = 2 \ldots m-1$$

Iterate

CRF and Bilinear LSTM [Dyer, 2016]

Learning:

1. Feedforward
2. Gradient
   a) Belief propagation
3. Backpropagation



Output labels:
- Name entities

Input features:
- Word embedding

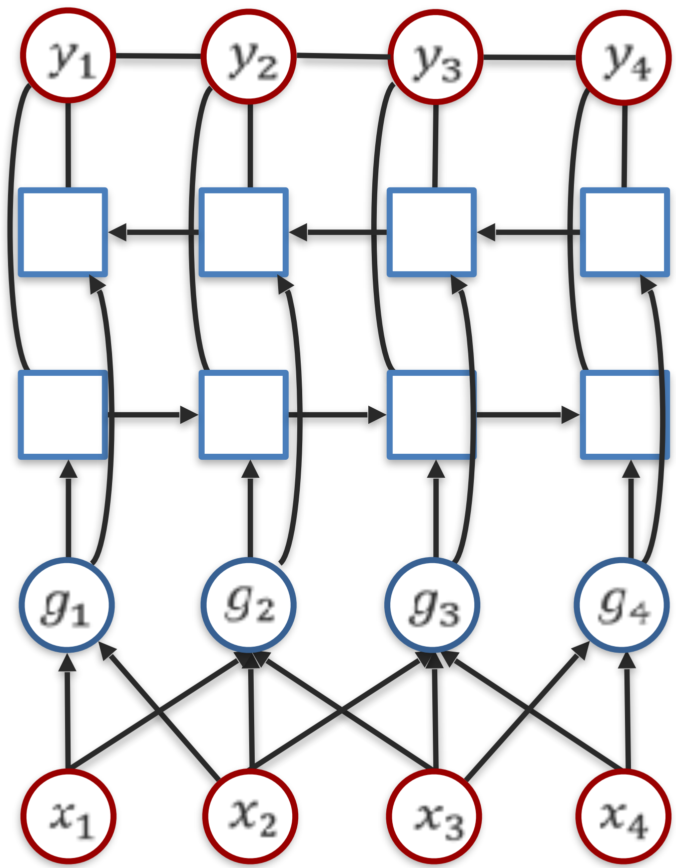
➢ What did θ^e paramters learn?

➢ What does LSTM parameters learns?

CNN and CRF and Bilinear LSTM [Hovy, 2016]

Learning:

1. Feedforward
2. Gradient
    a) Belief propagation
3. Backpropagation



Output labels:

- Name entities

Input features:

- Character embedding

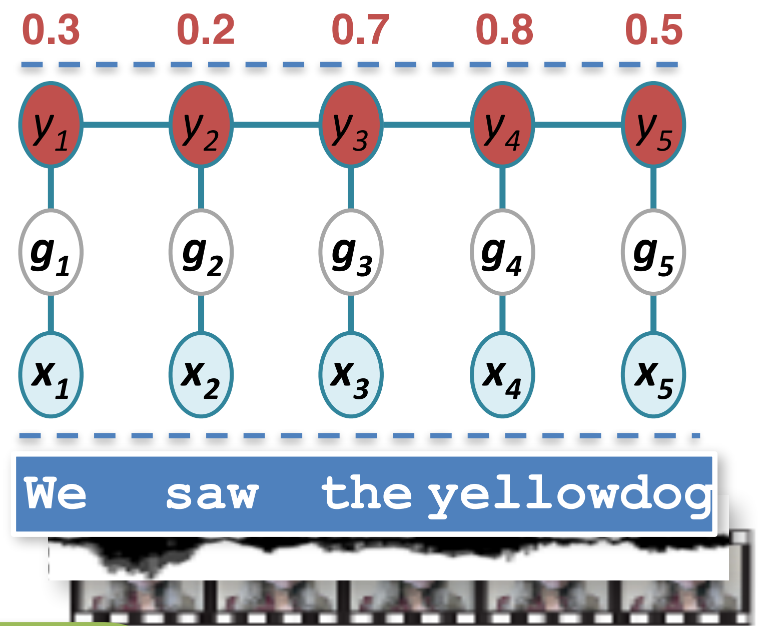
Continuous and Fully-Connected CRFs

Continuous Conditional Neural Field [Baltrusaitis 2014]

Continuous output variables: **(e.g., continuous emotional label)**

$$\psi_1(y_1, x; \theta_1) \propto \quad y_1$$

$$p(\mathbf{y} | \mathbf{x}; \boldsymbol{\theta}) = \underbrace{\frac{1}{Z(\mathbf{x}; \boldsymbol{\theta})} \exp\left\{ \sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \boldsymbol{\theta}^g) \right\}}_{y_3}$$



- How to solve

Multivariate Gaussian integral:

$$\int_{-\infty}^{\infty} \exp\{\tfrac{1}{2} \boldsymbol{y}^T \Sigma^{-1} \boldsymbol{y} + \boldsymbol{y} \Sigma^{-1} \boldsymbol{\mu}\} d\boldsymbol{y}$$

$$= \frac{(2\pi)^{n/2}}{|\Sigma^{-1}|^{1/2}} \exp(\tfrac{1}{2} \boldsymbol{\mu} \, \Sigma^{-1} \boldsymbol{\mu})$$

[Radosavljevic et al., 2010]

Continuous Conditional Neural Field

Continuous output variables: **(e.g., continuous emotional label)**

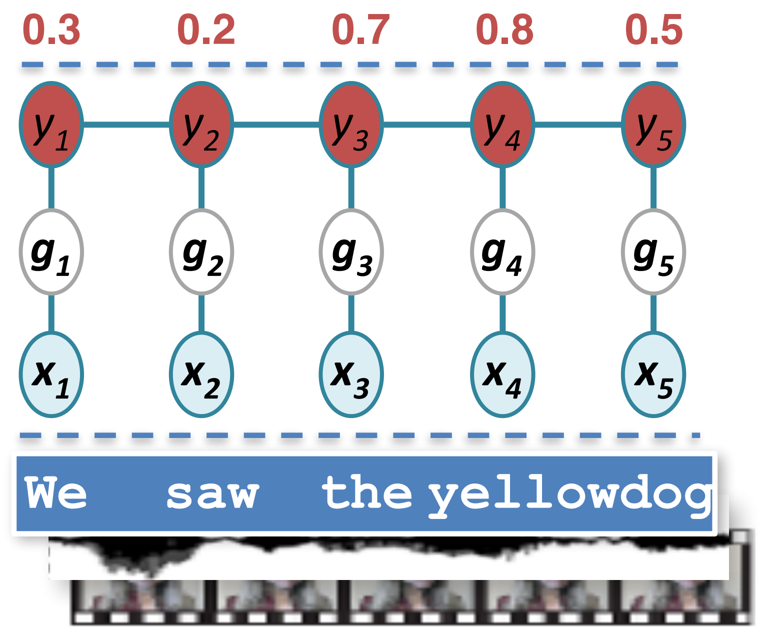
$\psi_1(y_1, x; \theta_1) \times \quad y_1$

$$p(\mathbf{y}|\,\mathbf{x};\boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x};\boldsymbol{\theta})} \exp\left\{\sum_t \boldsymbol{\theta} \cdot F(y_t, y_{t-1}, \mathbf{x}_t, \theta^g)\right\}$$

y_3

y_4

$\psi_2(y_2, x; \theta_2) \times$



Continuous Conditional Neural Field

Continuous output variables: **(e.g., continuous emotional label)**

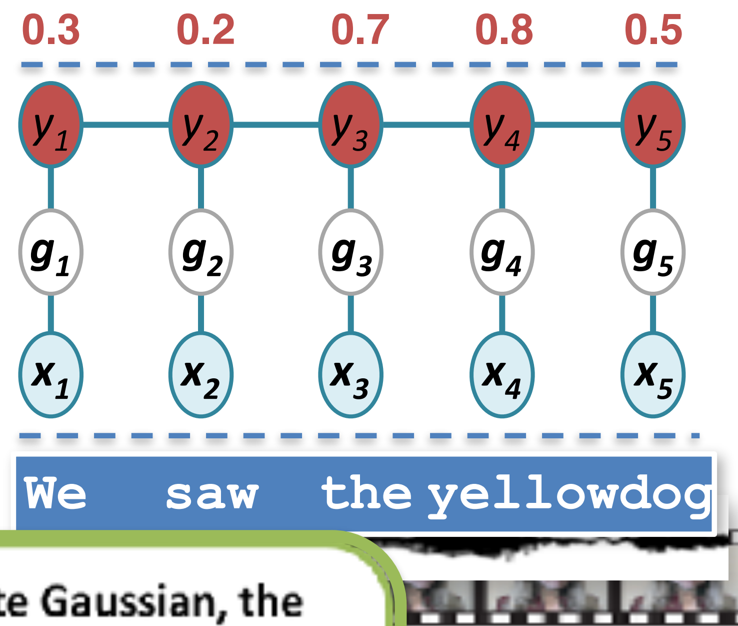
$\psi_1(y_1, x; \theta_1) \times \quad y_1$

Multivariate Gaussian distribution:

$$\psi_3(y_3, x; \theta_3) \times$$

$\psi_4($

$\phi_{23}(y_2,$



Since CCNF can be viewed as a multivariate Gaussian, the prediction of y' is simply the mean value of distribution:

$$y' = \arg\max_{y}\left(P(y|x)\right) = \mu$$

➢ Optimized using gradient ascent or BFGS.

High-Order Continuous Conditional Neural Field

Continuous output variables: **(e.g., continuous emotional label)**

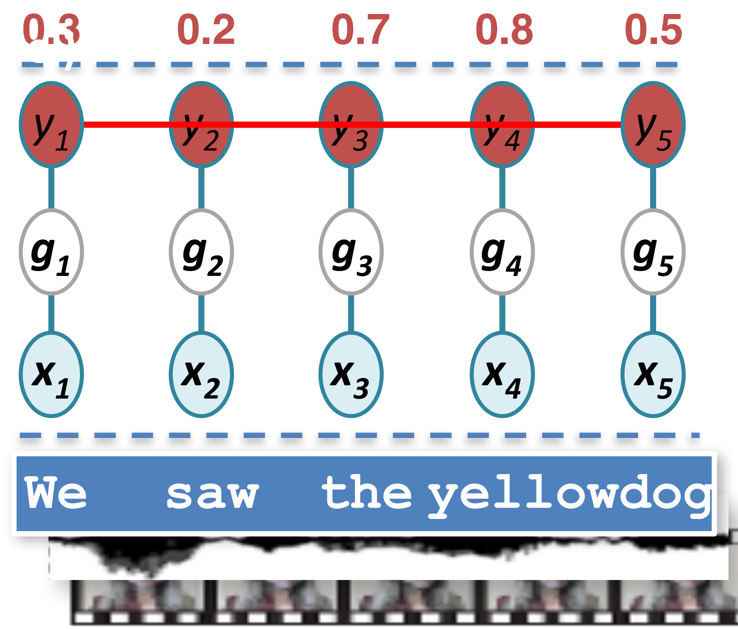
$\psi_1(y_1, x; \theta_1) \times \quad y_1$

Multivariate Gaussian distribution:

$$\psi_3(y_3, x; \theta_3) \times$$

k-order potential functions:

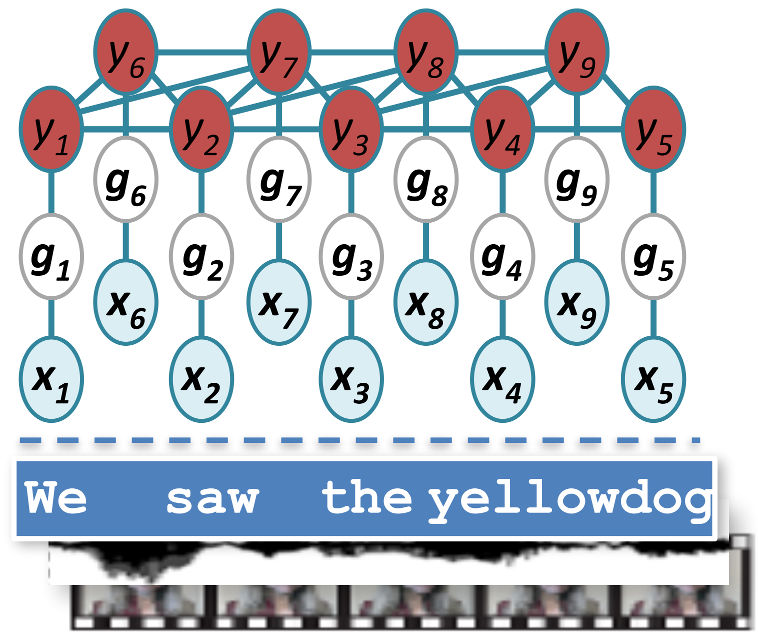
$$\psi_1(y_1, x_1; \theta_1) \times$$



Fully-Connected Continuous Conditional Neural Field

Continuous output variables: **(e.g., continuous emotional label)**

$$\psi_1(y_1, x; \theta_1) \times \qquad y_1$$



Multivariate Gaussian distribution:

$$\psi_3(y_3, x; \theta_3) \times$$

k-order potential functions:

$$\psi_1(y_1, x_1; \theta_1) \times$$

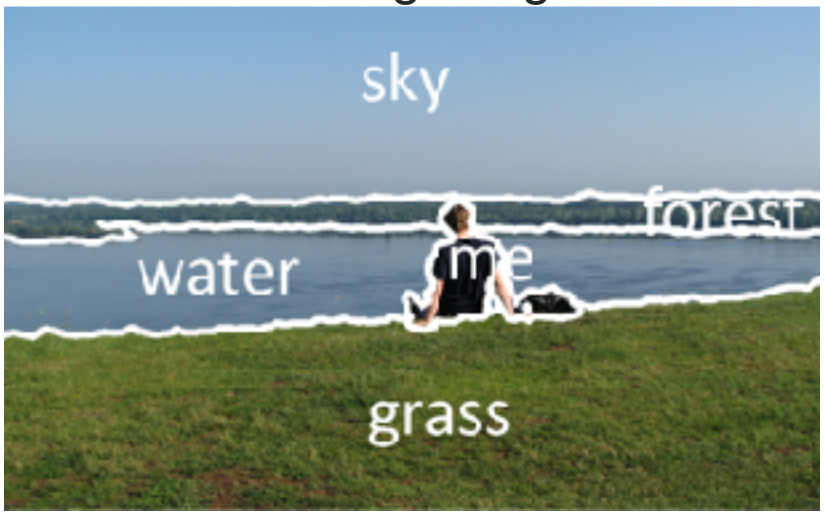
Grid potential functions:

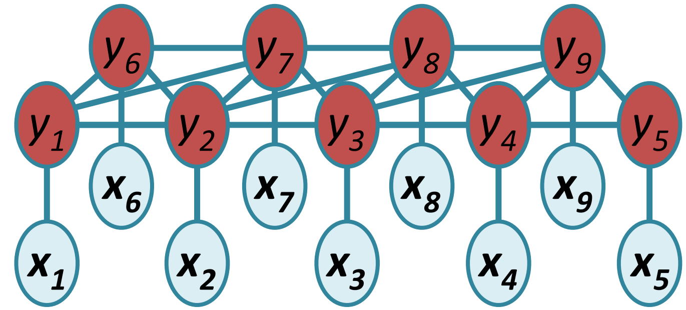
$$f^{2D}(y_i, y_j) = -\frac{1}{2} S_{ij}(y_i - y_j)^2$$

where $S_{i,j}$ specifies which nodes are connected.

Fully-Connected CRF [Krahenbuhl and Koltun, 2013]

“Semantic” image segmentation





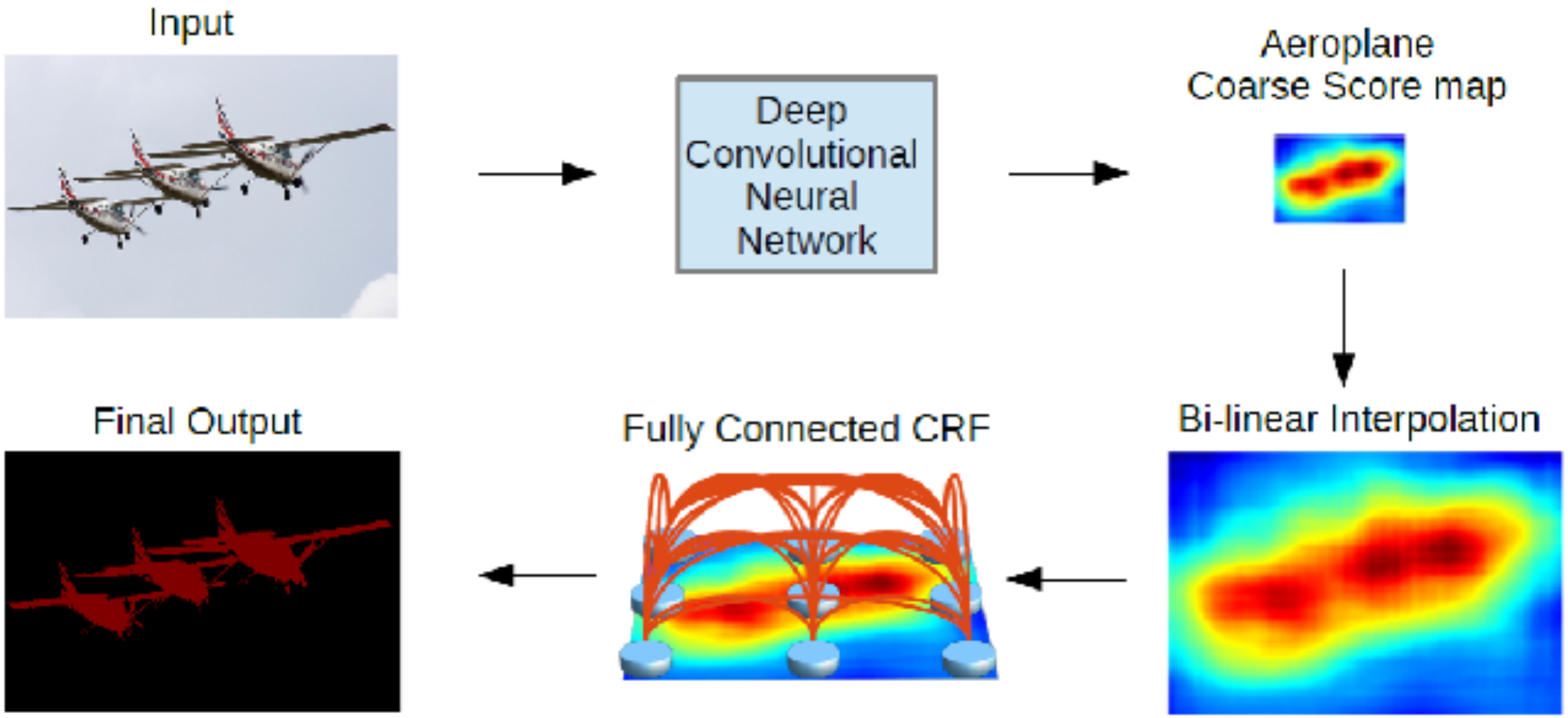
y_i: object class label

$\boldsymbol{x}_i$: local pixel features

$$p(\boldsymbol{y}|x;\theta) = \frac{\Phi(\boldsymbol{y},;\theta)}{\sum_{\boldsymbol{y}'}\Phi(\boldsymbol{y}',\boldsymbol{x};\theta)}$$

where $$\Phi_{ij}(y_i,y_j;\boldsymbol{\theta}) = \sum_{m=1}^{C} \overbrace{u^{(m)}(y_i,y_j|\boldsymbol{\theta})k^{(m)}(\boldsymbol{x}_i,\boldsymbol{x}_j)}^{\text{Mixture of kernels}}$$

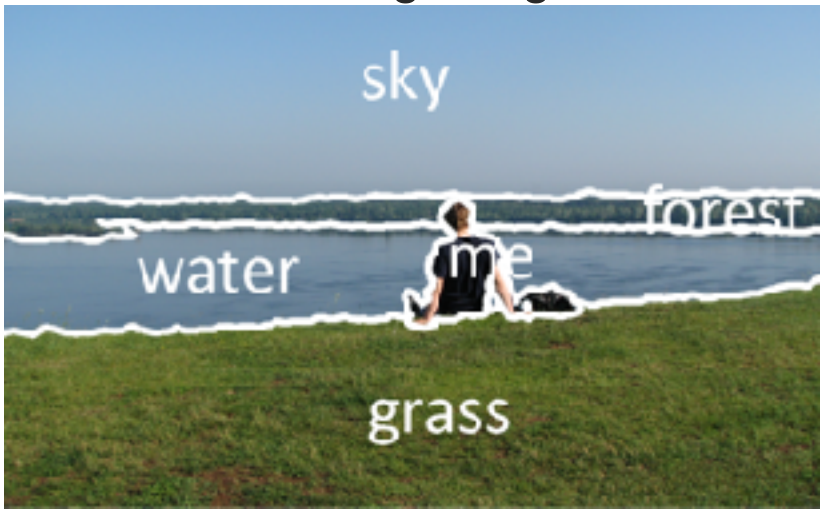
CNN and Fully-Connected CRF [Chen et al., 2014]

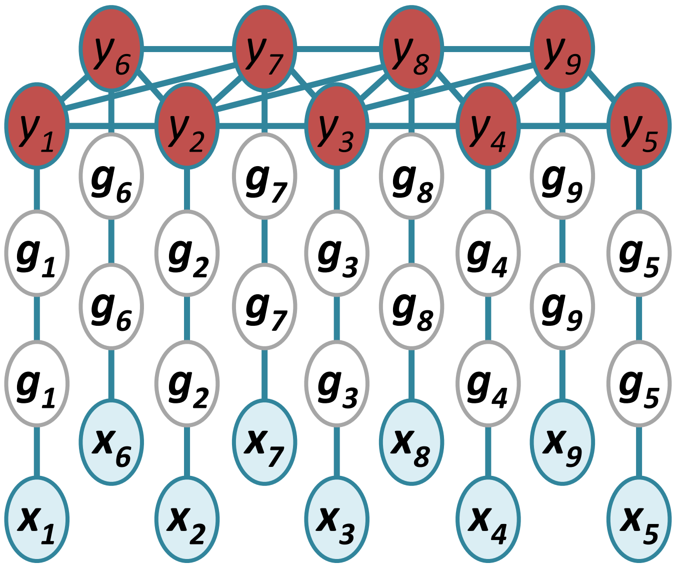


Fully Connected Deep Structured Networks

[Zheng et al., 2015; Schwing and Urtasun, 2015]

“Semantic” image segmentation





Algorithm: Learning Fully Connected Deep Structured Models
Repeat until stopping criteria

1. Forward pass to compute $f_r(x, \hat{y}_r; w)\ \forall r \in \mathcal{R}, y_r \in \mathcal{Y}_r$
2. Computation of marginals $q^t_{(x,y),i}(\hat{y}_i)$ via filtering for $t \in \{1, \ldots, T\}$
3. Backtracking through the marginals $q^t_{(x,y),i}(\hat{y}_i)$ from $t = T - 1$ down to $t = 1$
4. Backward pass through definition of function via chain rule
5. Parameter update

Using mean field approximation