Lecture 12
Undirected Graphical Models

Lei Li and Yu-xiang Wang
UCSB

Part of slides borrowed from Alex Smola

Recap

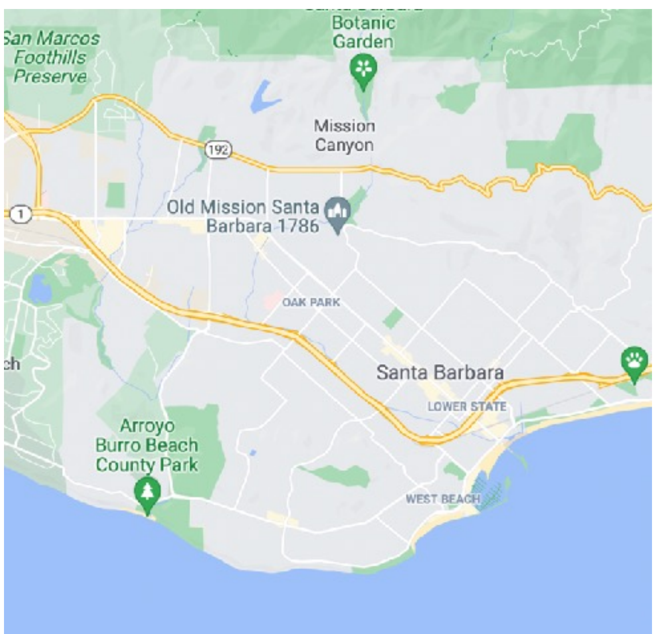
- Gaussian Mixture Models
- Expectation-Maximization
- Linear Dynamical Systems
  - Forward-backward algorithm (Kalman filter, Kalman smoothing)

Understanding Query Intent

Building a conversational assistant for Google Map?

Noodle house [Keyword] near Santa Barbara [Location]

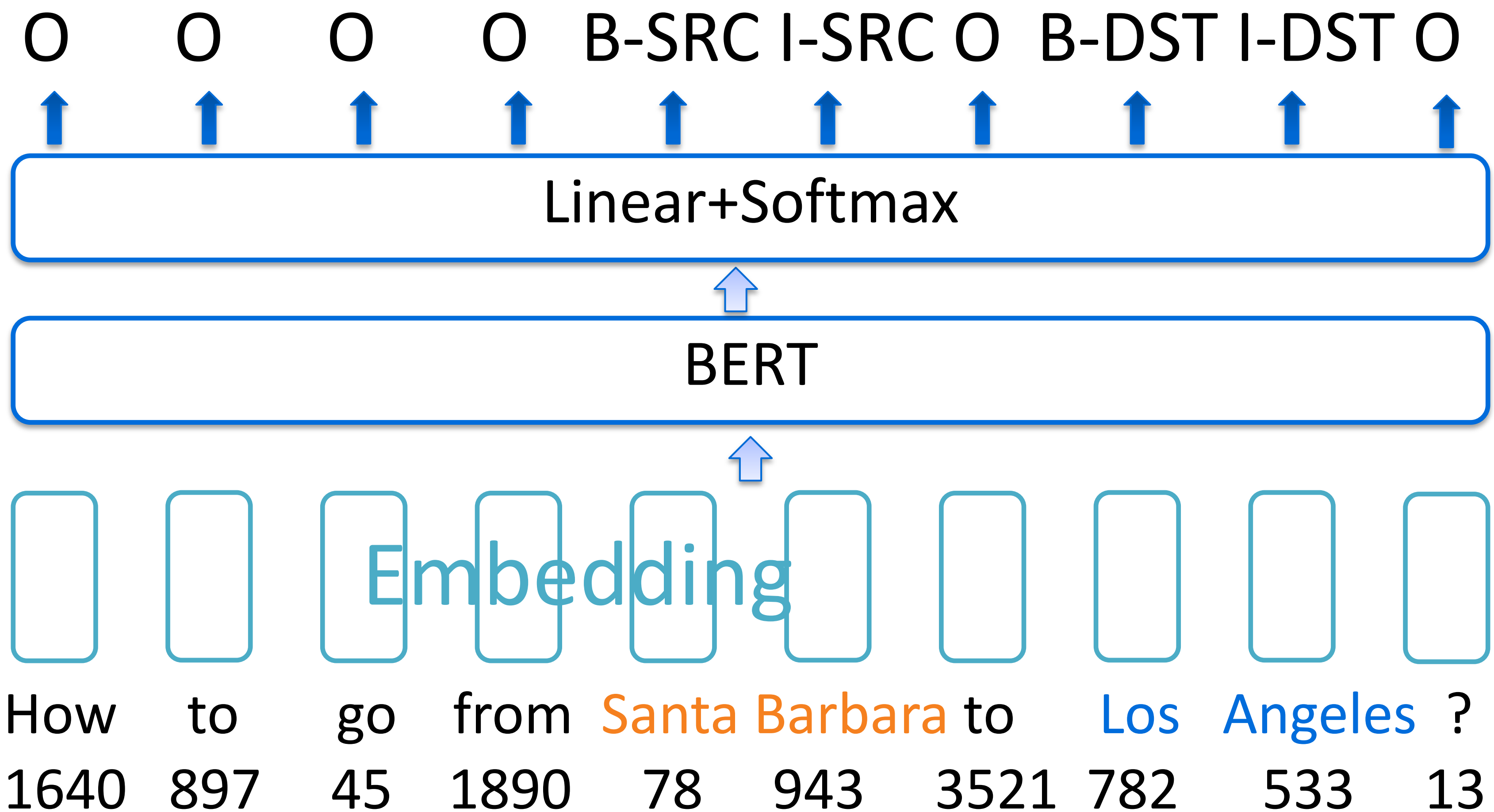
How to go from Santa Barbara [Origin] to Log Angeles [Destination] ?



Sequence Labelling problem

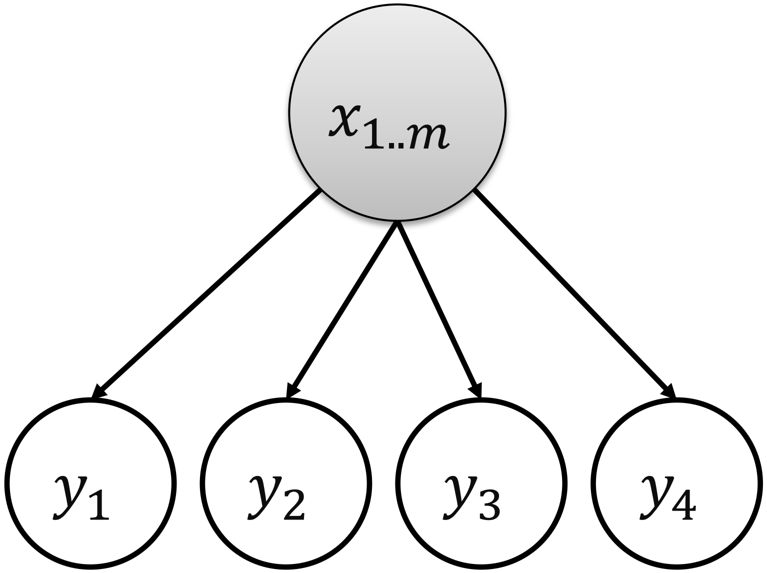
Parsing Query Intent

$$p(y_1, \ldots, y_m | x_1, \ldots, x_m)$$



Label Independent?

- Are y_i y_j independent given $x_1, \dots, x_m$?
- But neighboring labels are correlated



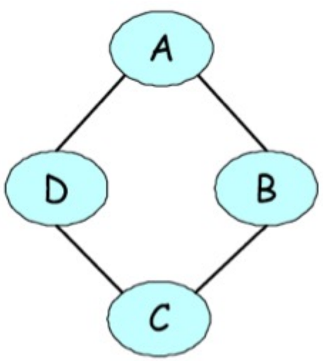
How to go from Santa Barbara to Log Angeles ?

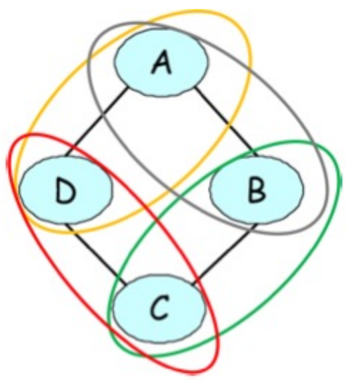
Markov Random Fields

- Using undirected graphs to represent probability distributions of random variables

$$\tilde{p}(A,B,C,D) = \phi(A,B)\phi(B,C)\phi(C,D)\phi(D,A)$$

$$\phi(X,Y) = \begin{cases} 10 \; if \; X = Y = 1 \\ 5 \; if \; X = Y = 0 \\ 1 \; otherwise \end{cases}$$





$$p(A,B,C,D) = \frac{1}{Z}\tilde{p}(A,B,C,D)$$

Normalizing term, also partition function

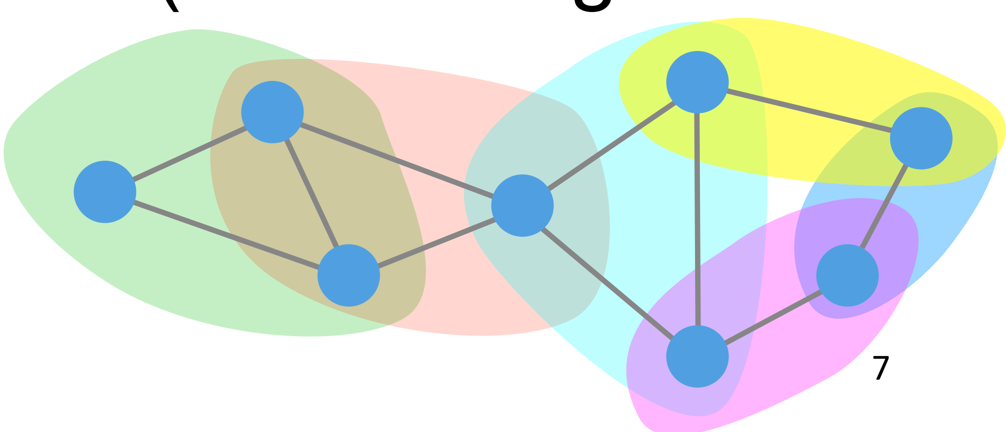
MRF

- A Markov Random Field is a probability distribution p over variables $x_1 .. x_n$ defined by an undirected graph G, s.t.

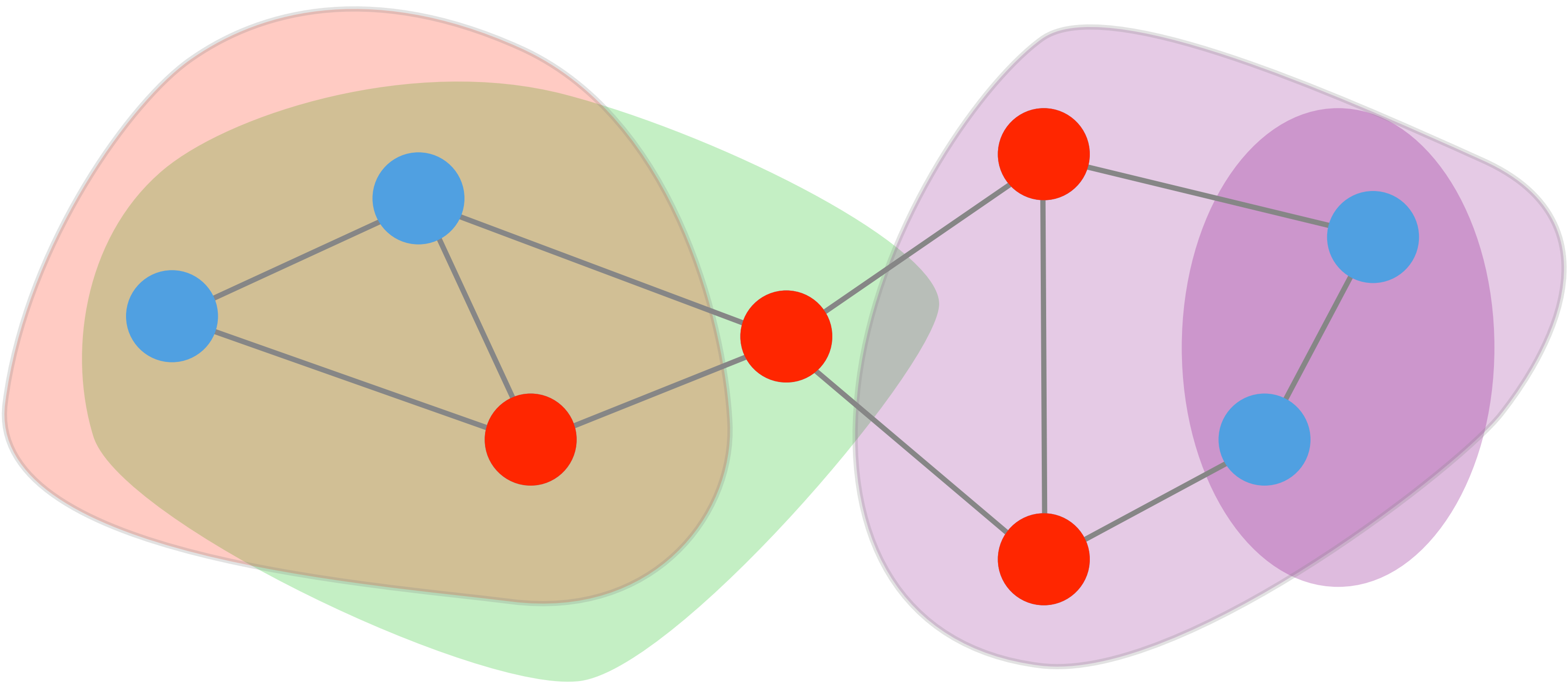
$$p(x_1, \ldots x_n) = \frac{1}{Z} \prod_{c \in C(G)} \phi_c(x_c)$$

Z is the partition function (normalizing constant)

C: max-cliques

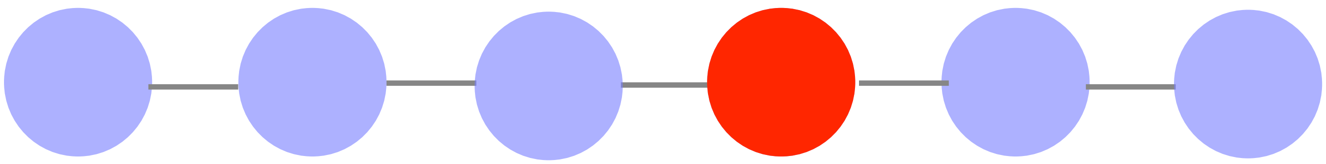


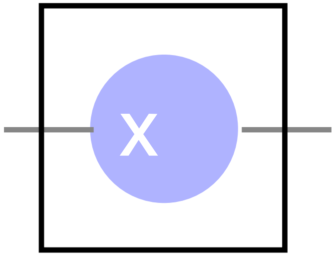
Independence in MRF



Key Concept
Observing nodes makes remainder conditionally independent

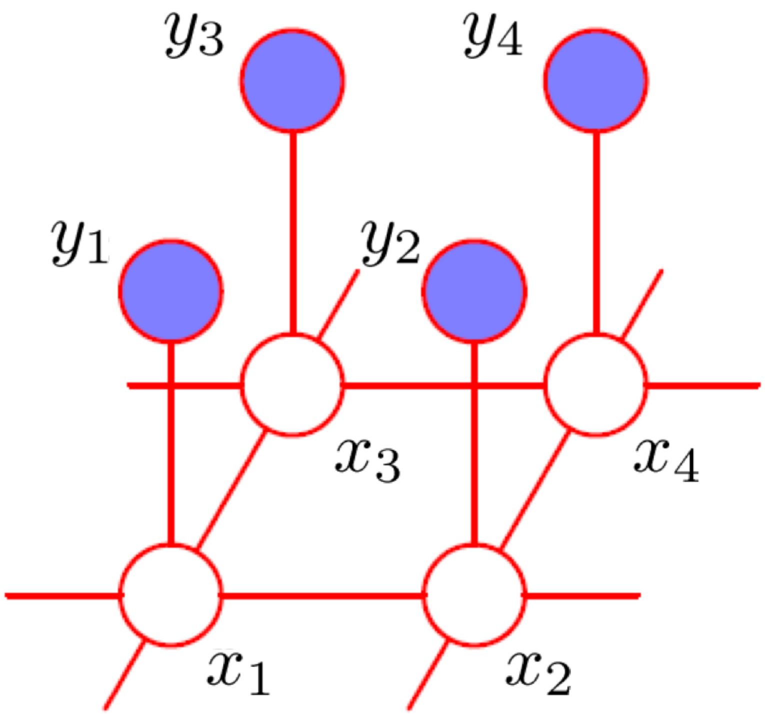
Example



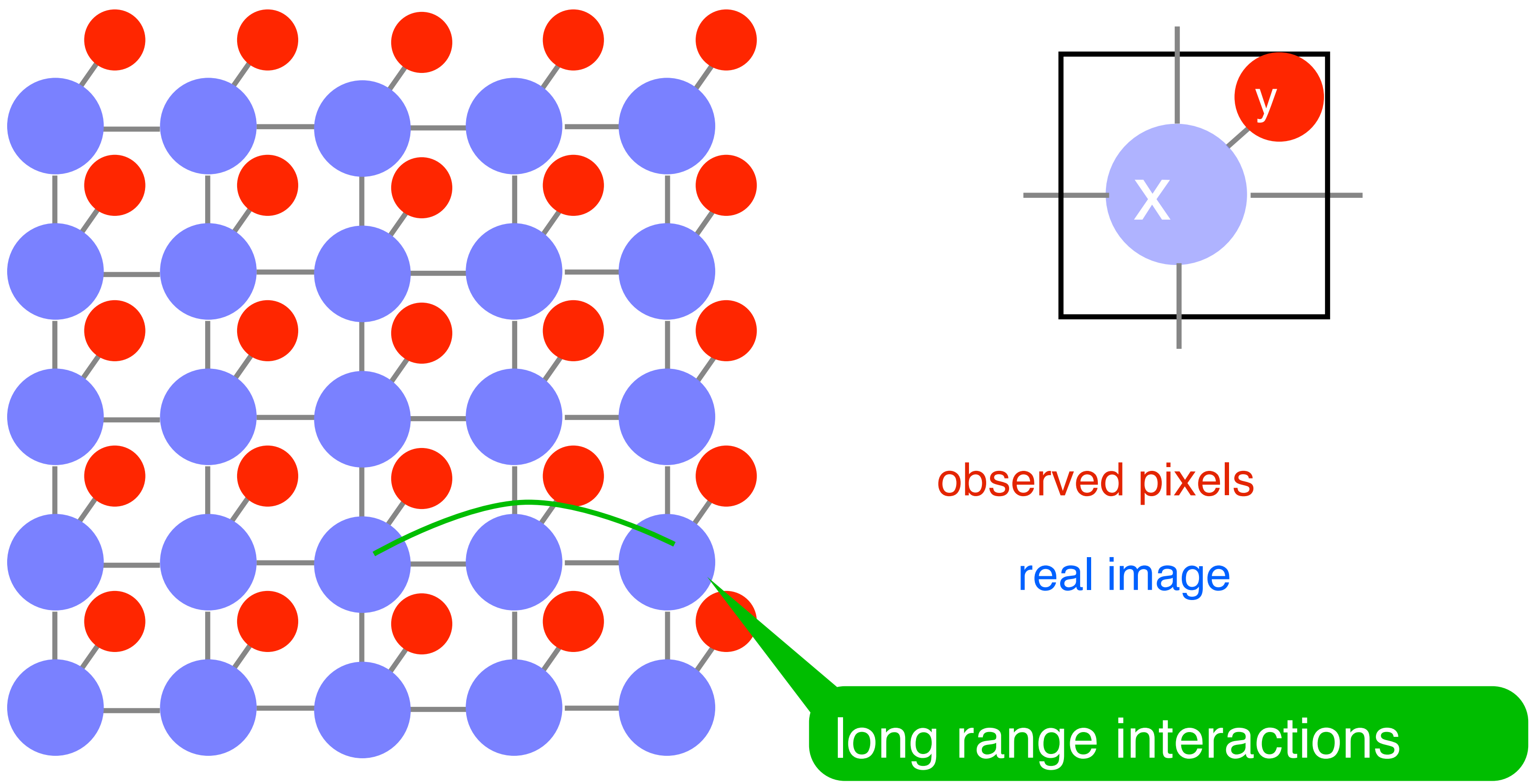


$$p(x) = \prod_i \psi_i(x_i, x_{i+1})$$

Example

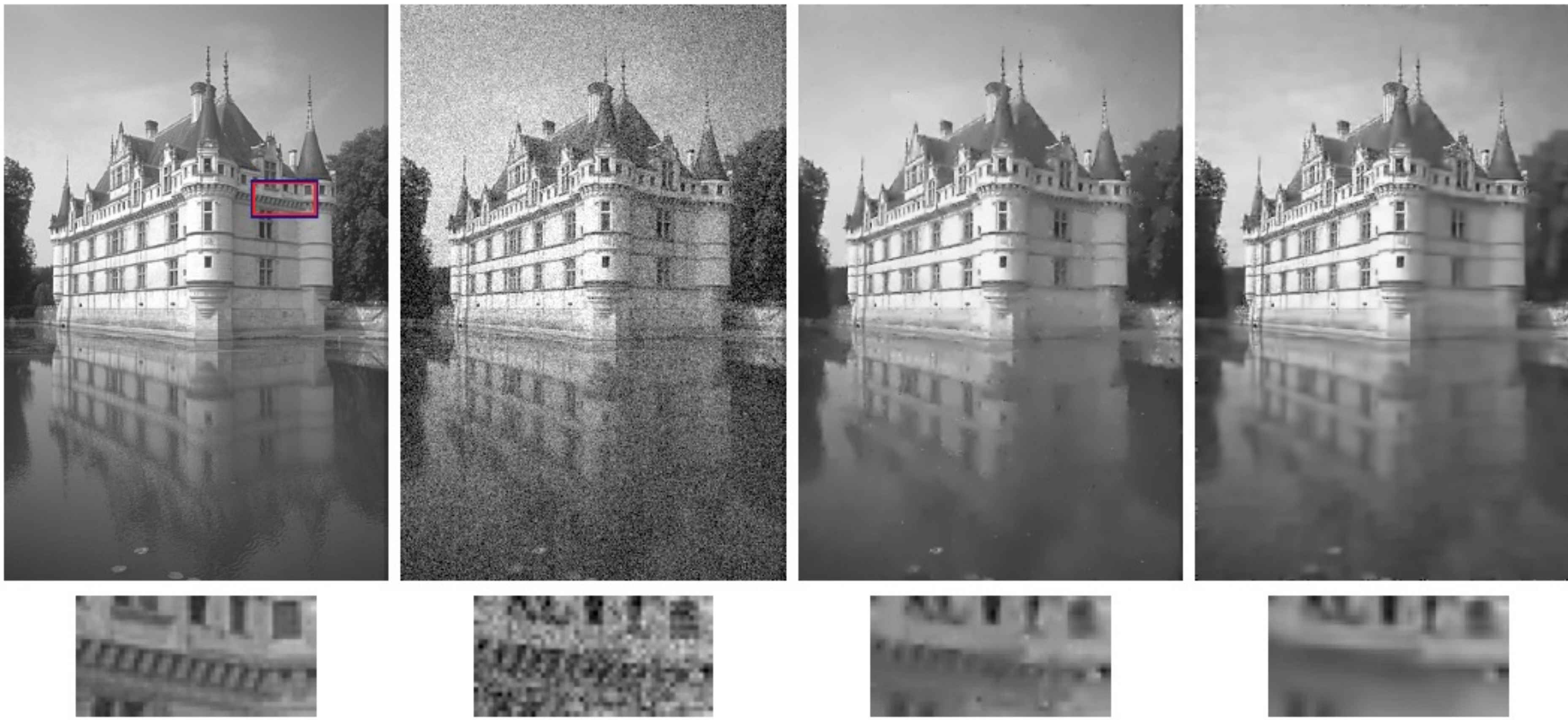


Spin Glasses + Images



$$p(x|y) = \prod_{ij} \psi^{\text{right}}(x_{ij}, x_{i+1,j}) \psi^{\text{up}}(x_{ij}, x_{i,j+1}) \psi^{xy}(x_{ij}, y_{ij})$$

Image Denoising



Li&Huttenlocher, ECCV'08

Hammersley-Clifford Theorem

- Set of distributions that factorize according to the graph – F
- Set of distributions that respect conditional independencies implied by graph-separation – I
- F $\Leftrightarrow$ I

Directed vs. Undirected

- Causal description
- Normalization automatic
- Intuitive
- Requires knowledge of dependencies
- Conditional independence tricky (d-separation)

- Noncausal description (correlation only)
- Intuitive
- Easy modeling
- Normalization difficult
- Conditional independence easy to read off (graph connectivity)

Exponential Family

- Density function

$$p(x;\theta) = \exp\left(\langle\phi(x),\theta\rangle - g(\theta)\right)$$
$$\text{where } g(\theta) = \log\sum_{x'}\exp\left(\langle\phi(x'),\theta\rangle\right)$$

- Log partition function generates cumulants

$$\partial_\theta g(\theta) = \mathbf{E}\left[\phi(x)\right]$$
$$\partial^2_\theta g(\theta) = \mathrm{Var}\left[\phi(x)\right]$$

- g is convex (second derivative is p.s.d.)

Log Partition Function

$$p(x|\theta) = e^{\langle \phi(x), \theta \rangle - g(\theta)}$$

$$g(\theta) = \log \sum_x e^{\langle \phi(x), \theta \rangle}$$

Unconditional model

$$\partial_\theta g(\theta) = \frac{\sum_x \phi(x) e^{\langle \phi(x), \theta \rangle}}{\sum_x e^{\langle \phi(x), \theta \rangle}} = \sum_x \phi(x) e^{\langle \phi(x), \theta \rangle - g(\theta)}$$

$$p(y|\theta, x) = e^{\langle \phi(x,y), \theta \rangle - g(\theta|x)}$$

Conditional model

$$g(\theta|x) = \log \sum_y e^{\langle \phi(x,y), \theta \rangle}$$

$$\partial_\theta g(\theta|x) = \frac{\sum_y \phi(x,y) e^{\langle \phi(x,y), \theta \rangle}}{\sum_y e^{\langle \phi(x,y), \theta \rangle}} = \sum_y \phi(x,y) e^{\langle \phi(x,y), \theta \rangle - g(\theta|x)}$$

Estimation

- Conditional log-likelihood

$$\log p(y|x;\theta) = \langle \phi(x,y), \theta \rangle - g(\theta|x)$$

- Log-posterior (Gaussian Prior)

$$\log p(\theta|X,Y) = \sum_i \log(y_i|x_i;\theta) + \log p(\theta) + \text{const.}$$

$$= \left\langle \sum_i \phi(x_i,y_i), \theta \right\rangle - \sum_i g(\theta|x_i) - \frac{1}{2\sigma^2} \|\theta\|^2 + \text{const.}$$

- First order optimality conditions

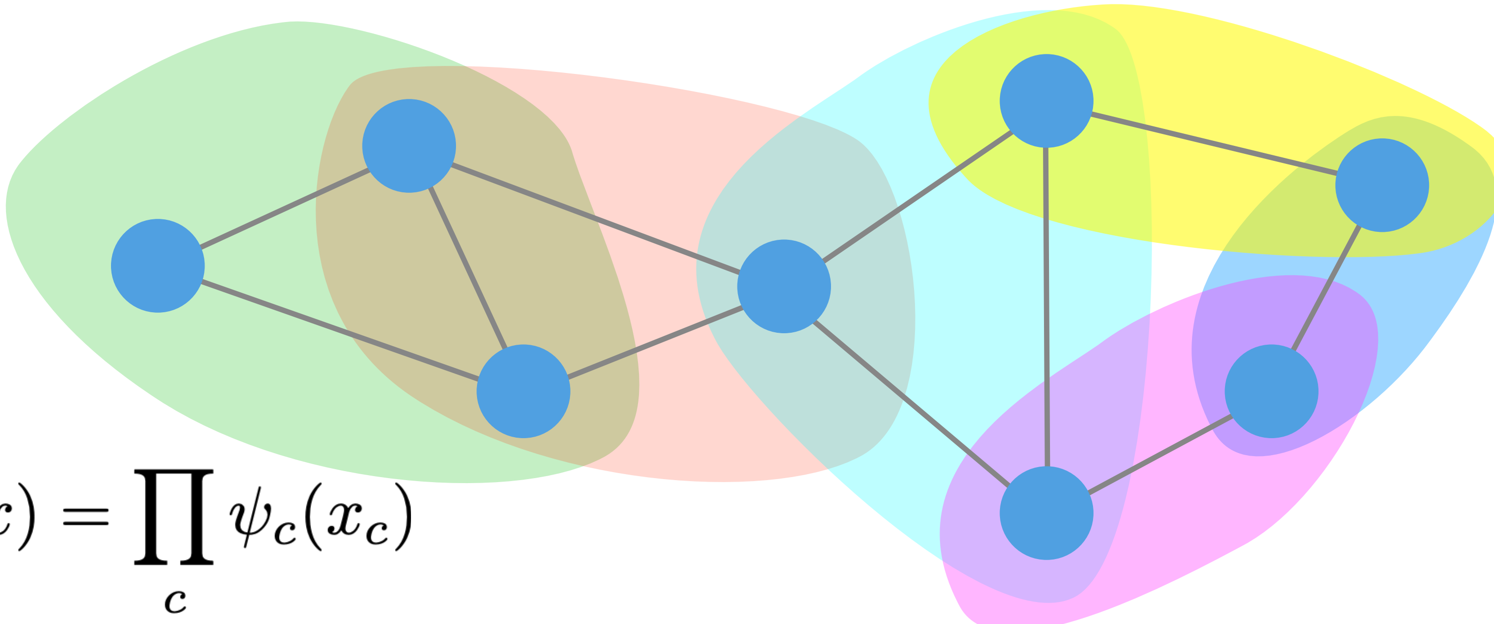
$$\sum_i \phi(x_i,y_i) = \sum_i \mathbf{E}_{y|x_i} [\phi(x_i,y)] + \frac{1}{\sigma^2}\theta$$

maxent model

expensive

prior

Exponential Clique Decomposition



$$p(x) = \prod_c \psi_c(x_c)$$

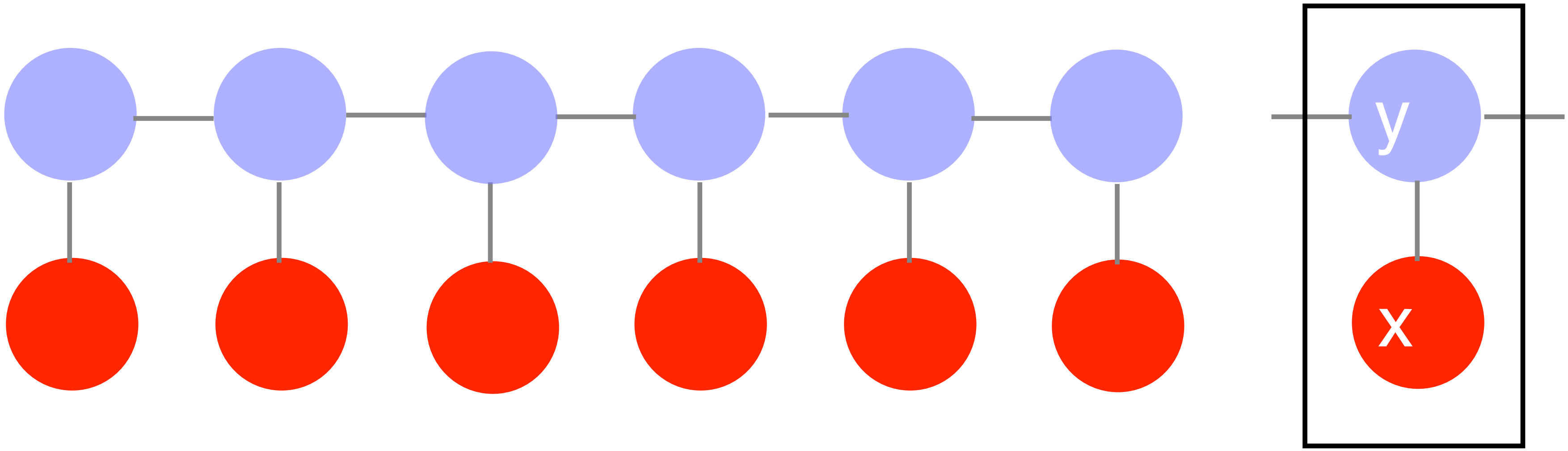
Theorem: Clique decomposition holds in sufficient statistics

$$\phi(x) = (\ldots, \phi_c(x_c), \ldots) \text{ and } \langle \phi(x), \theta \rangle = \sum_c \langle \phi_c(x_c), \theta_c \rangle$$

Corollary: we only need expectations on cliques

$$\mathbf{E}_x[\phi(x)] = (\ldots, \mathbf{E}_{x_c}[\phi_c(x_c)], \ldots)$$

Conditional Random Fields



$$\phi(x) = (y_1\phi_x(x_1), \ldots, y_n\phi_x(x_n), \phi_y(y_1, y_2), \ldots, \phi_y(y_{n-1}, y_n))$$

$$\langle\phi(x), \theta\rangle = \sum_i \langle\phi_x(x_i, y_i), \theta_x\rangle + \sum_i \langle\phi_y(y_i, y_{i+1}), \theta_y\rangle$$

$$g(\theta|x) = \sum_y \prod_i f_i(y_i, y_{i+1}) \text{ where}$$

dynamic Programming (examples later)

$$f_i(y_i, y_{i+1}) = e^{\langle\phi_x(x_i,y_i),\theta_x\rangle + \langle\phi_y(y_i,y_{i+1}),\theta_y\rangle}$$

Lafferty, McCallum, Pereira, ICML 2001

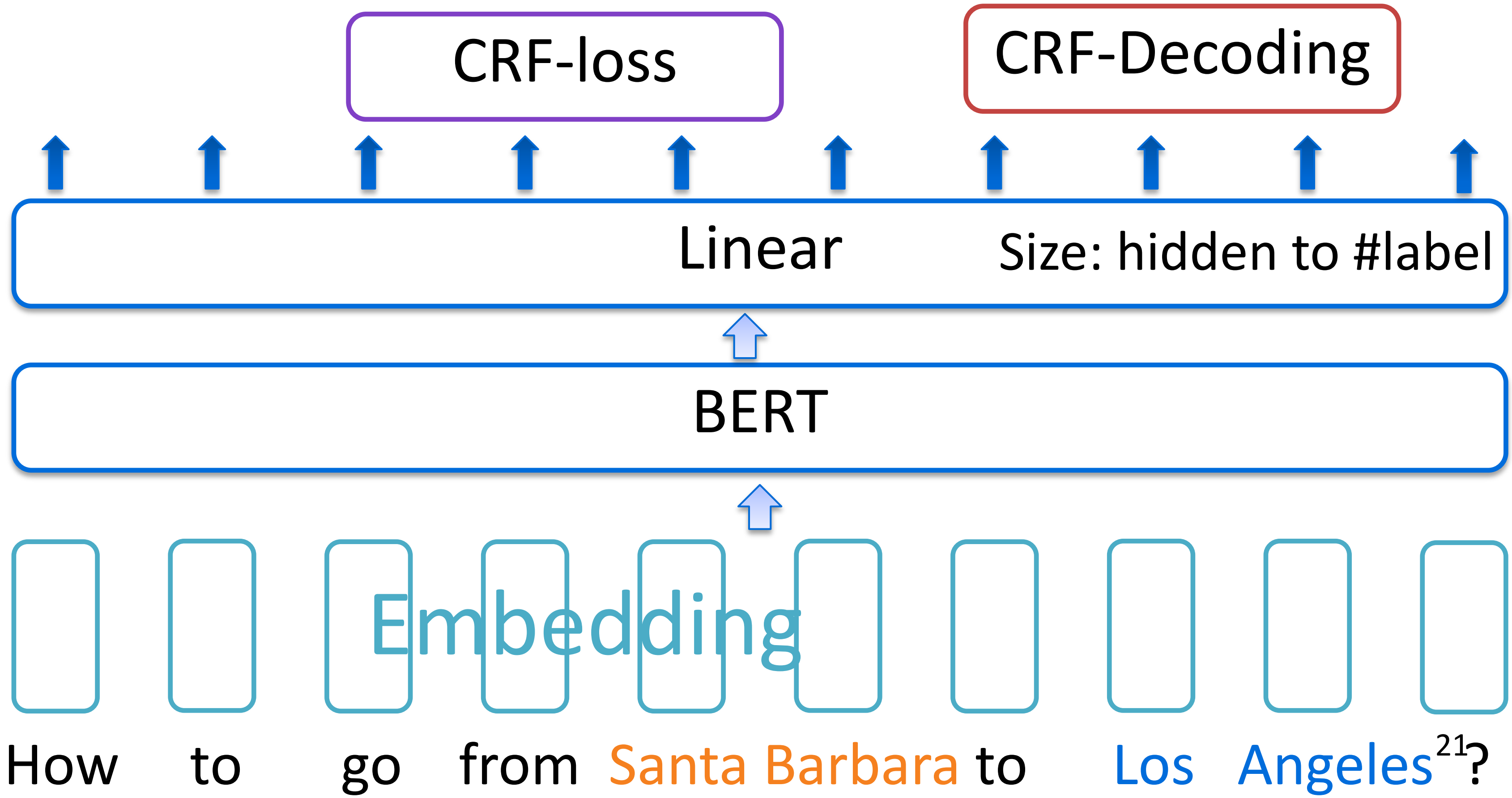
Conditional Random Fields

- Compute distribution over marginal and adjacent labels
- Take conditional expectations
- Take update step (batch or online)

- More general techniques for computing normalization via message passing ...

Combining NN and CRF

- BiLSTM+CRF
- BERT+CRF



Training BERT-CRF

- Labels: K
- A: transition matrix (K x K)
- Using dynamic programming to compute the log-partition function

Decoding

- Forward pass to compute last hidden layer from BERT
- Using Viterbi algorithm to compute max prob. label seq

Case study

- Building query intent parsing for Baidu Map

Next up

- Approximate Inference
  - Variational Inference
  - Sampling