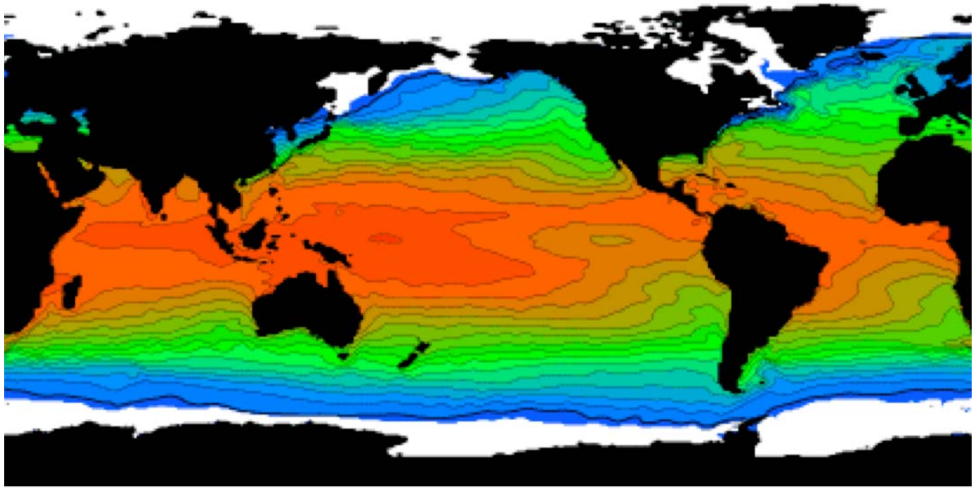
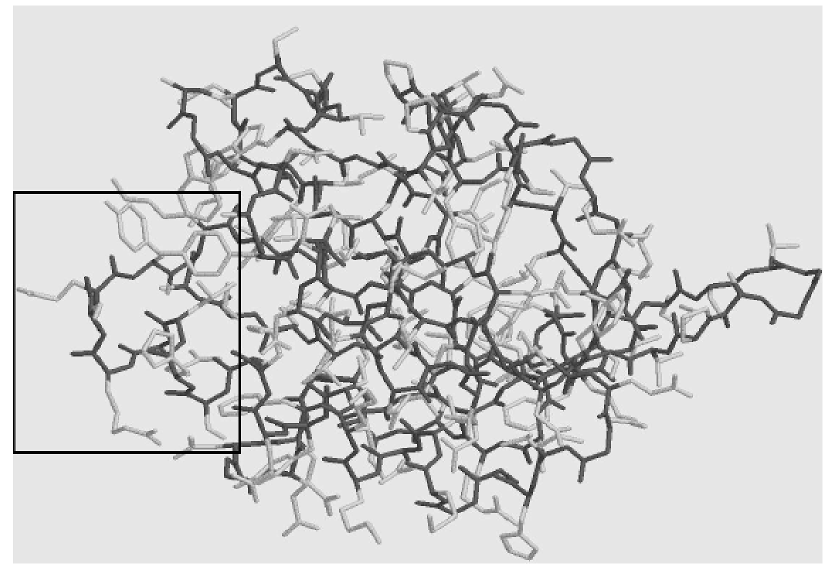
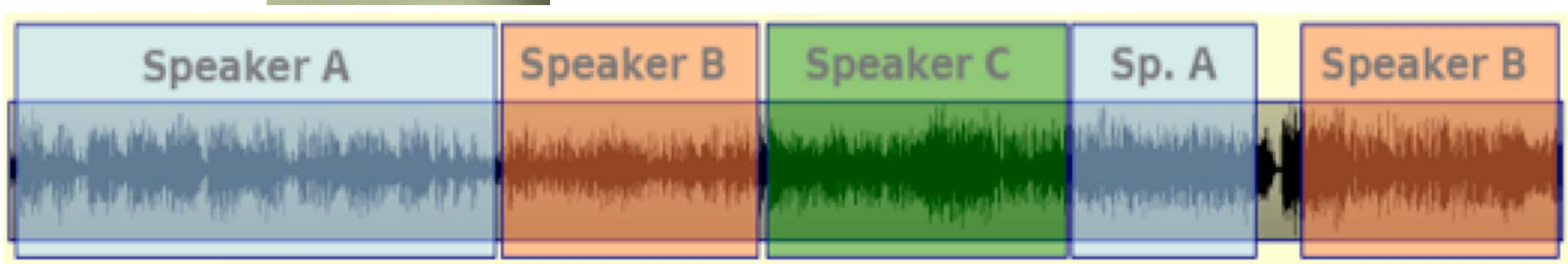
Learning and Inference in Probabilistic Graphical Models

CSCI 2950-P: Special Topics in Machine Learning
Spring 2010
Prof. Erik Sudderth

Learning from Structured Data

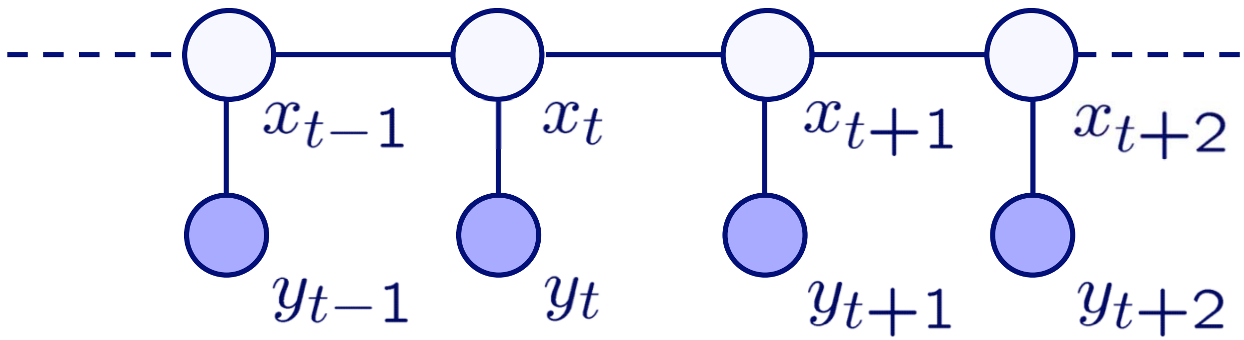






Hidden Markov Models (HMMs)

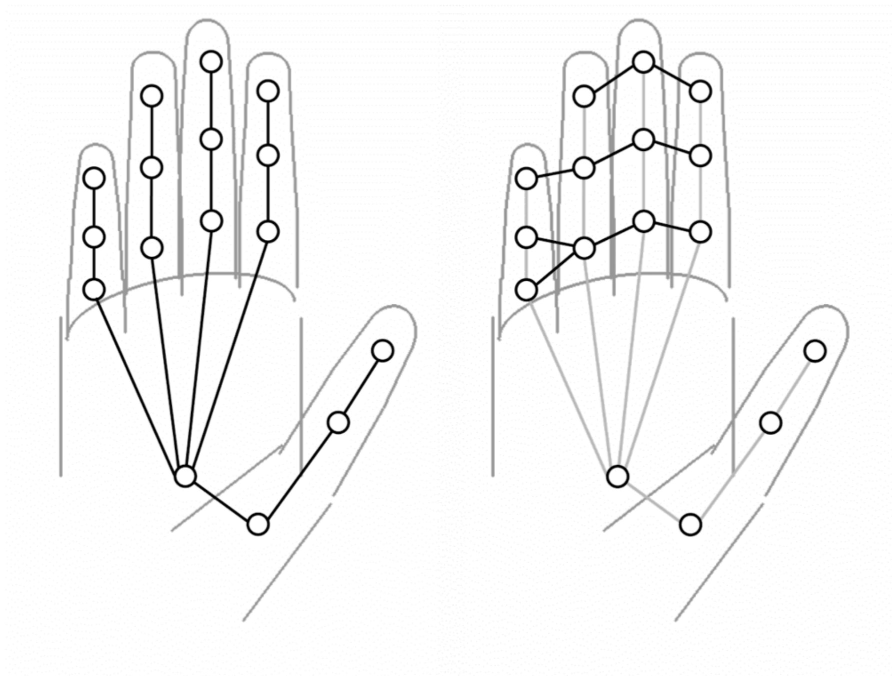
Visual Tracking

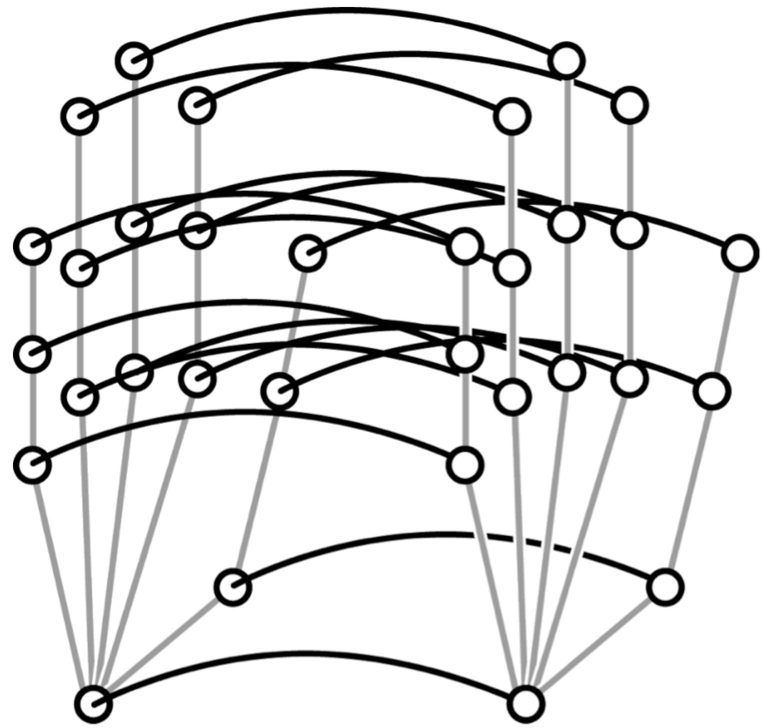


$$p(x, y) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1}) p(y_t \mid x_t)$$

"Conditioned on the present, the past and future are statistically independent"

Kinematic Hand Tracking



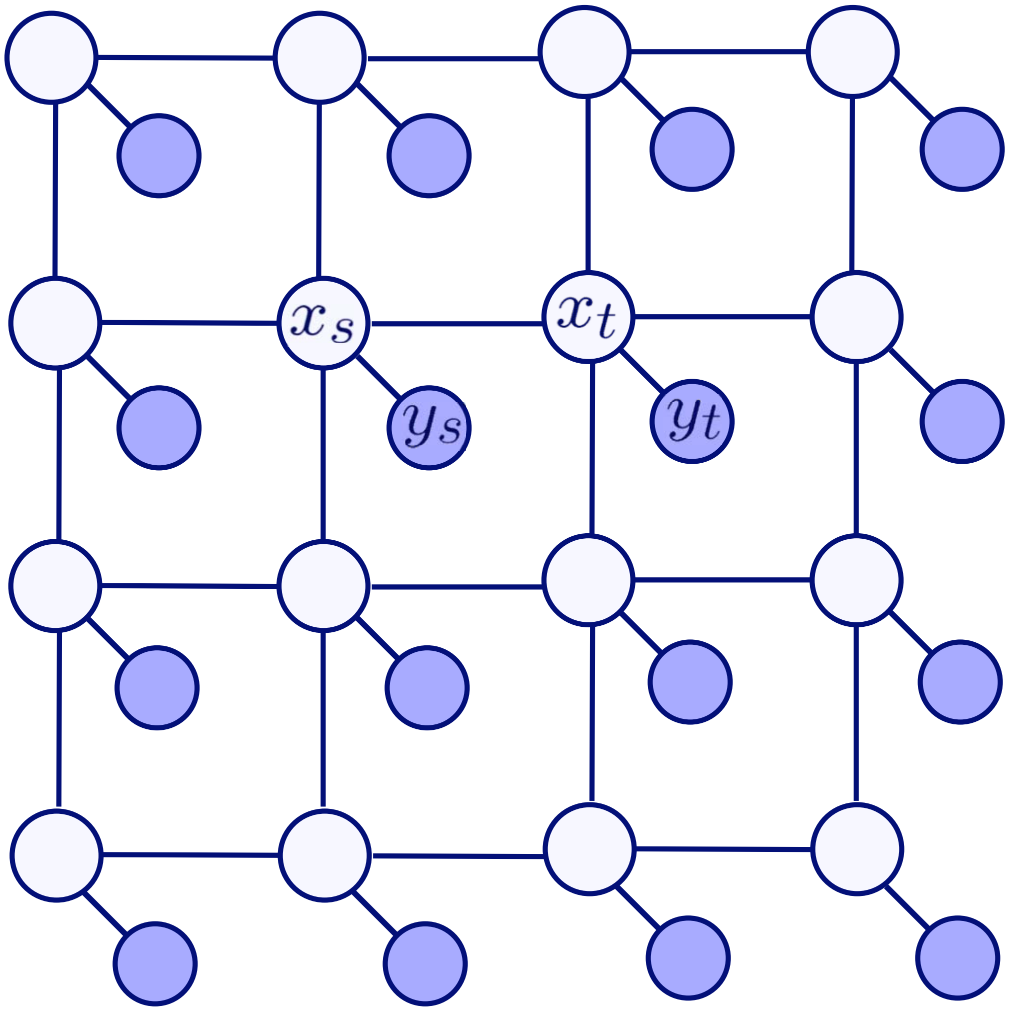


Kinematic Prior

Structural Prior

Dynamic Prior

Nearest-Neighbor Grids



Low Level Vision

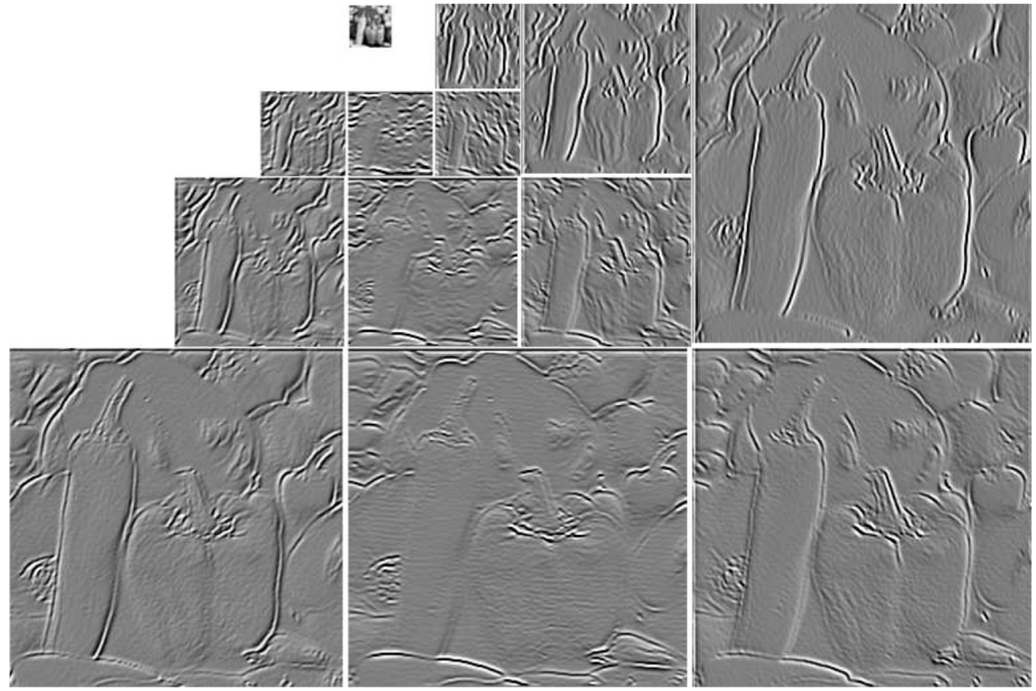
- Image denoising
- Stereo
- Optical flow
- Shape from shading
- Superresolution
- Segmentation

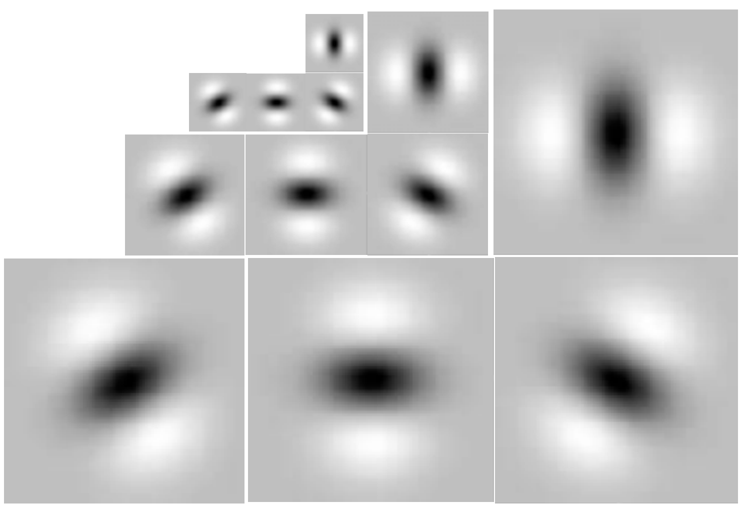
x_s → unobserved or hidden variable

y_s → local observation of x_s

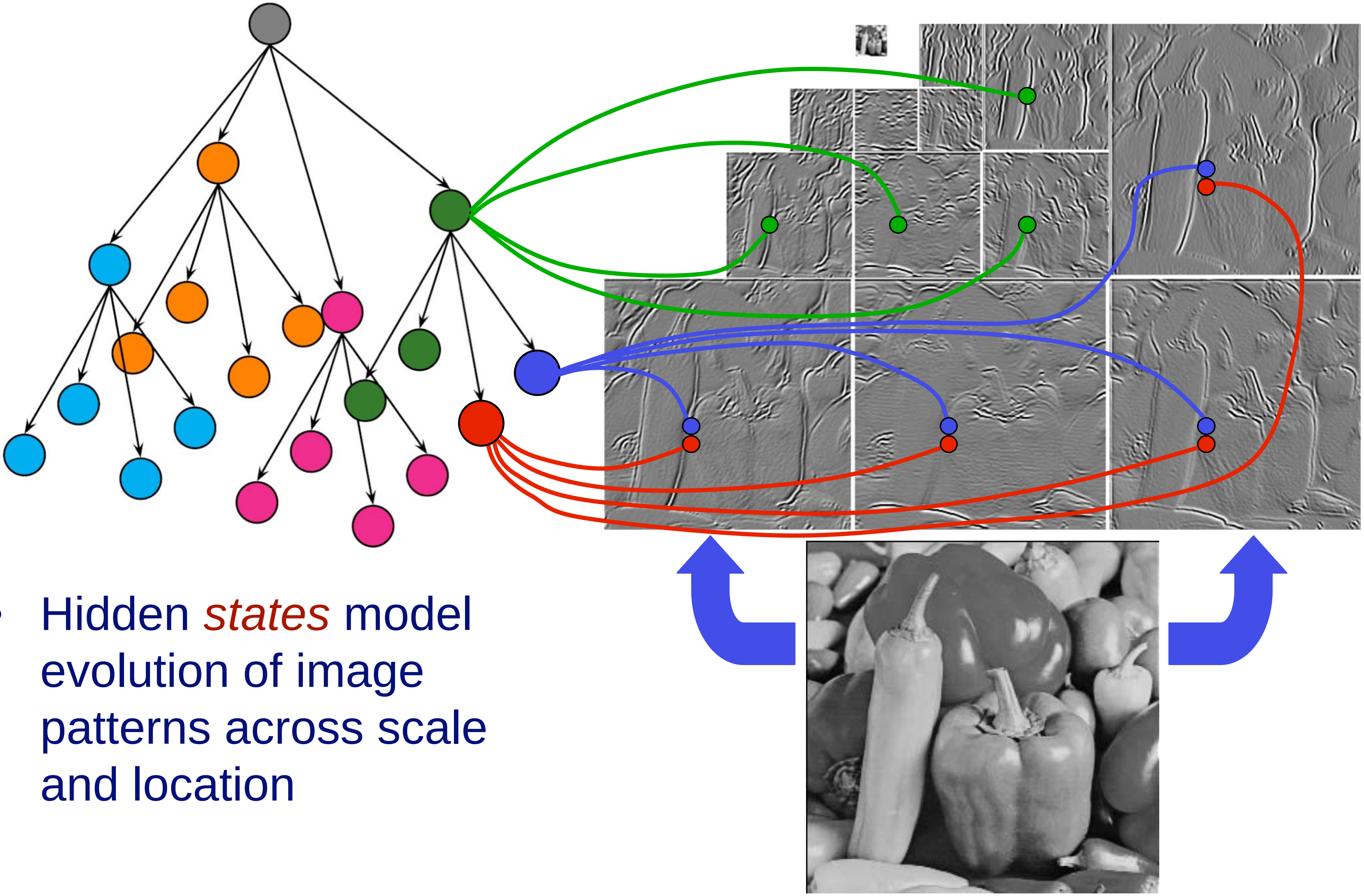
Wavelet Decompositions

- Bandpass decomposition of images into multiple *scales* & *orientations*
- Dense features which *simplify* statistics of natural images





Hidden Markov Trees



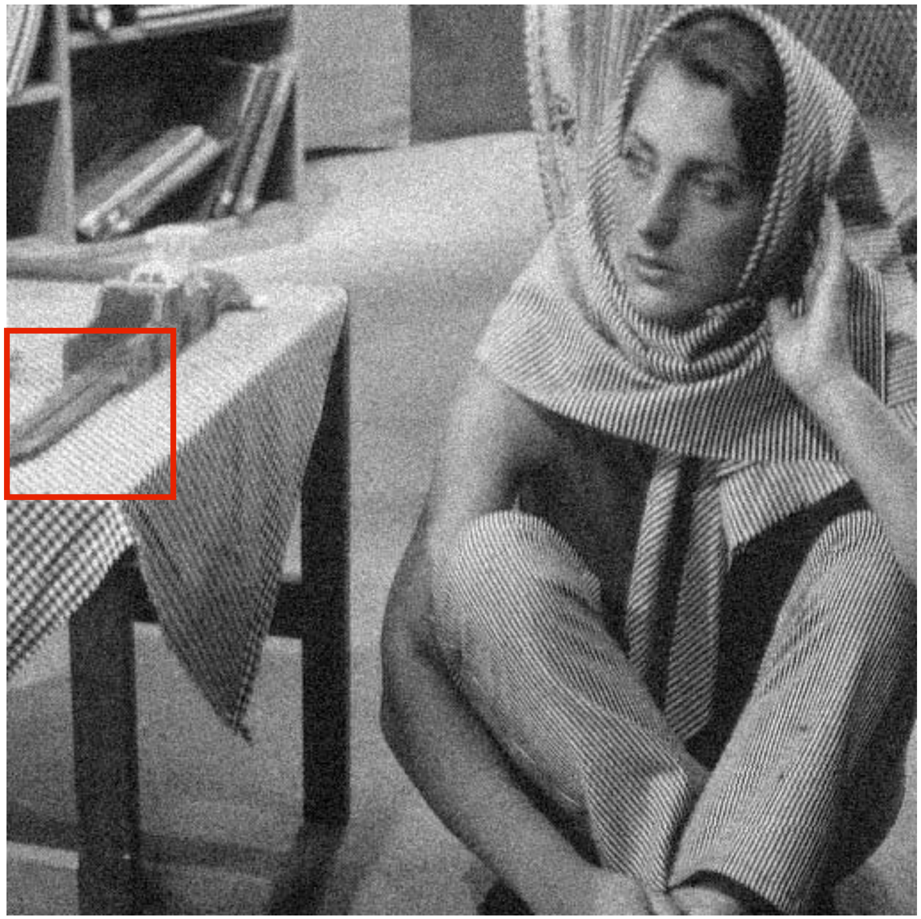
- Hidden *states* model evolution of image patterns across scale and location

Validation: Image Denoising

Original Image: *Barbara*

Corrupted by Additive White Gaussian Noise
(PSNR = 24.61 dB)

Denoising Results: Barbara

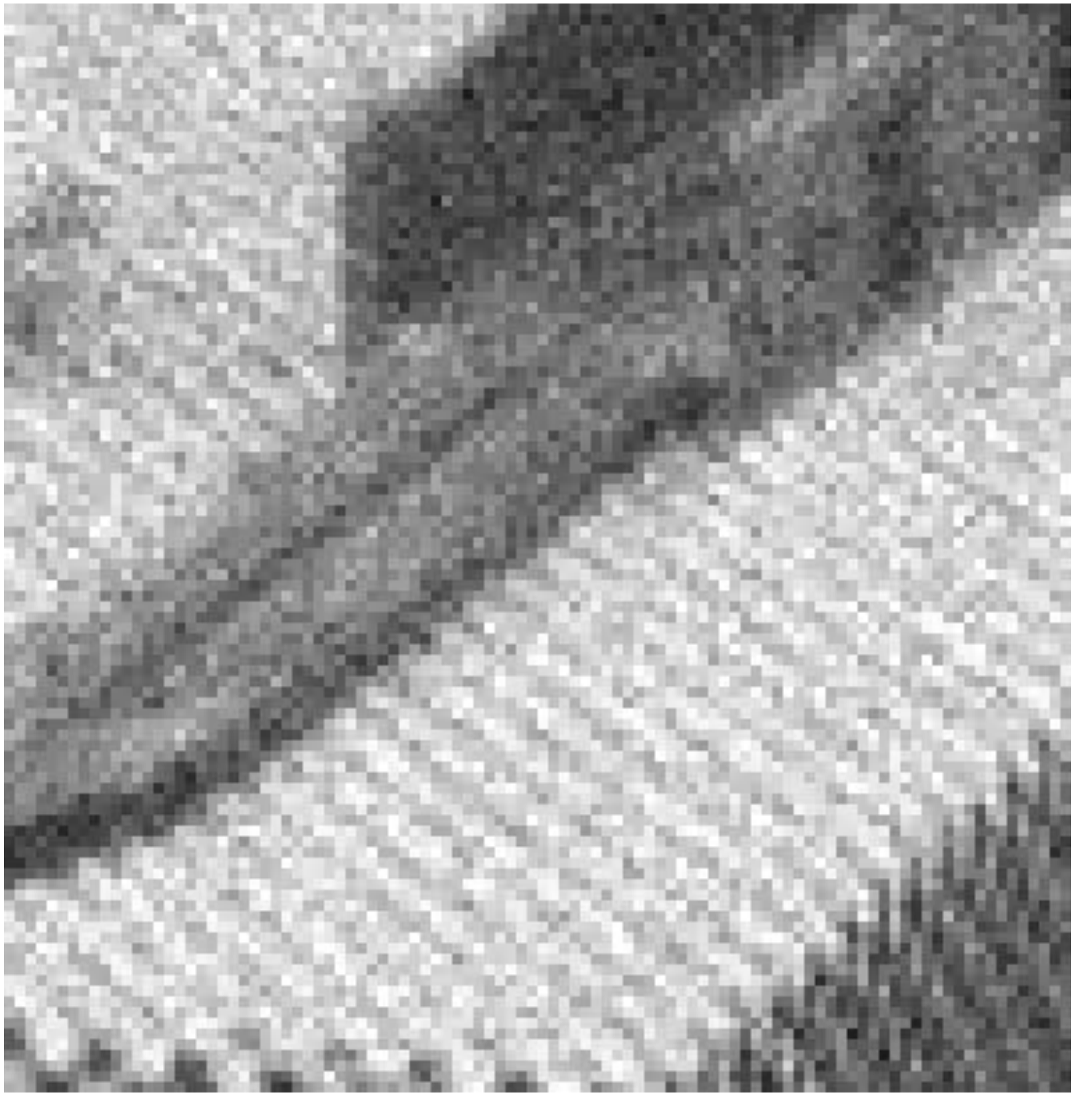


Noisy Input *(24.61 dB)*

HDP-HMT *(32.10 dB)*

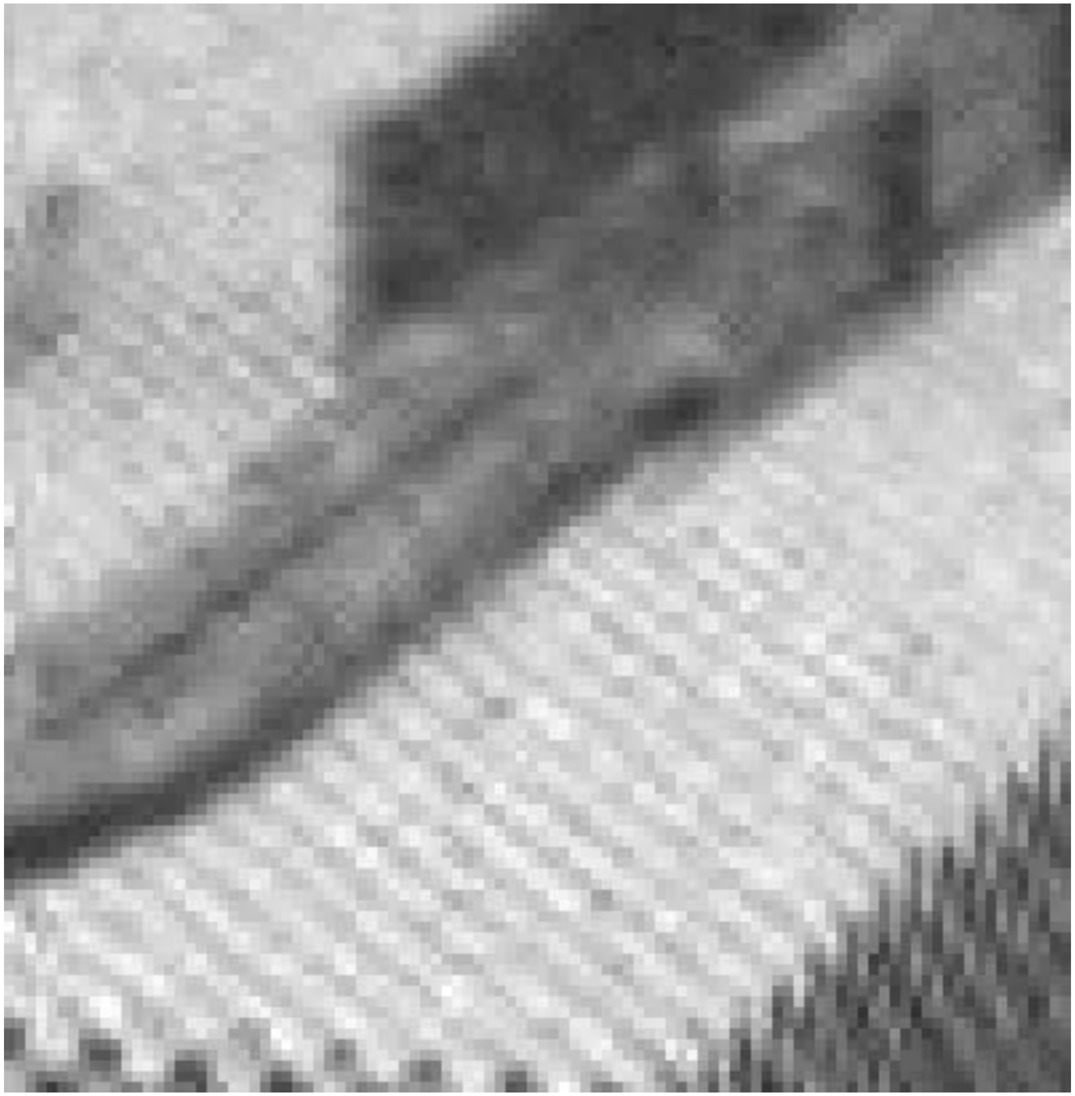
- Posterior mean of wavelet coefficients averages samples with varying numbers of states (model *averaging*)

Denoising: Input



24.61 dB

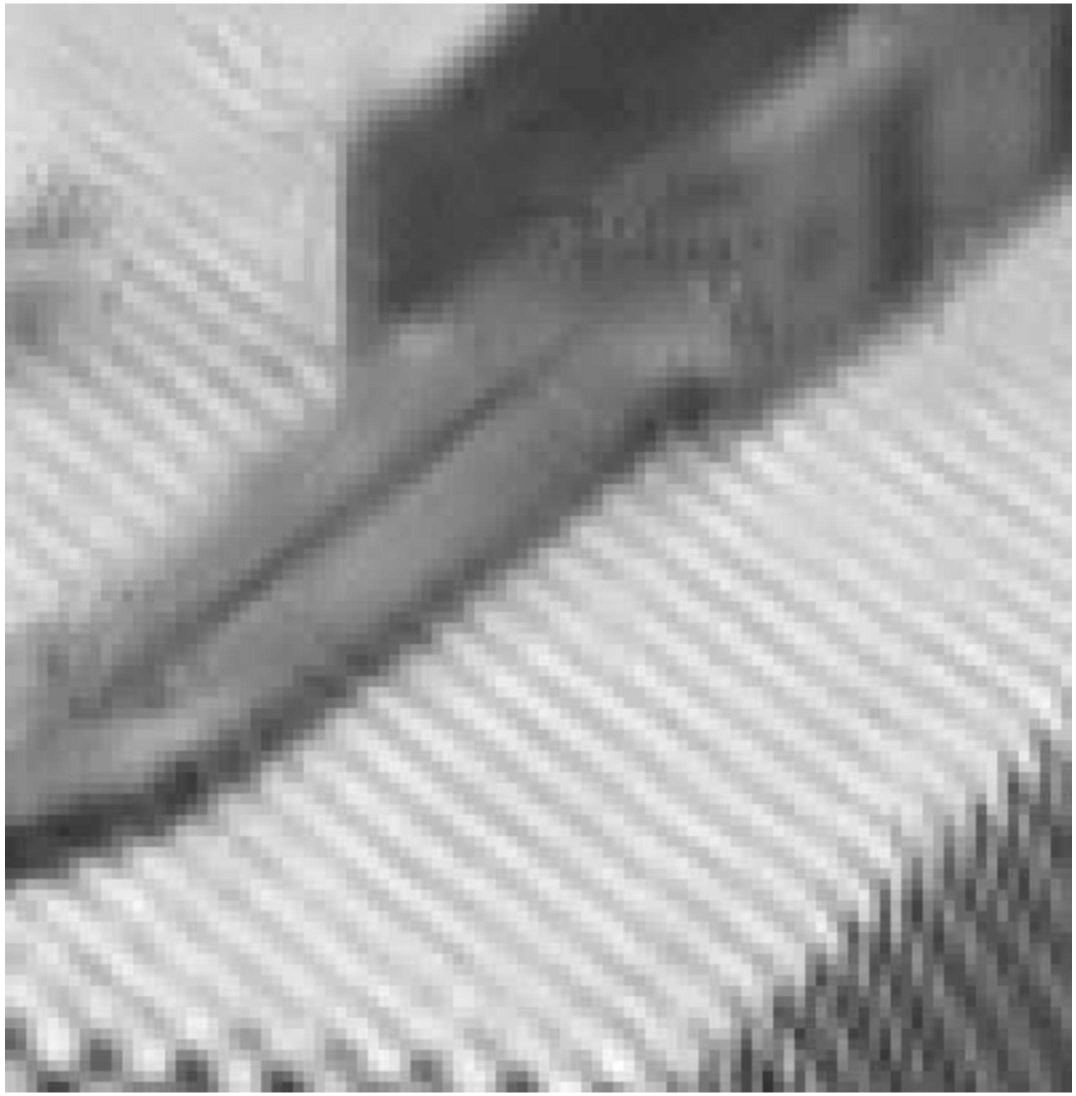
Denoising: Binary HMT



29.35 dB

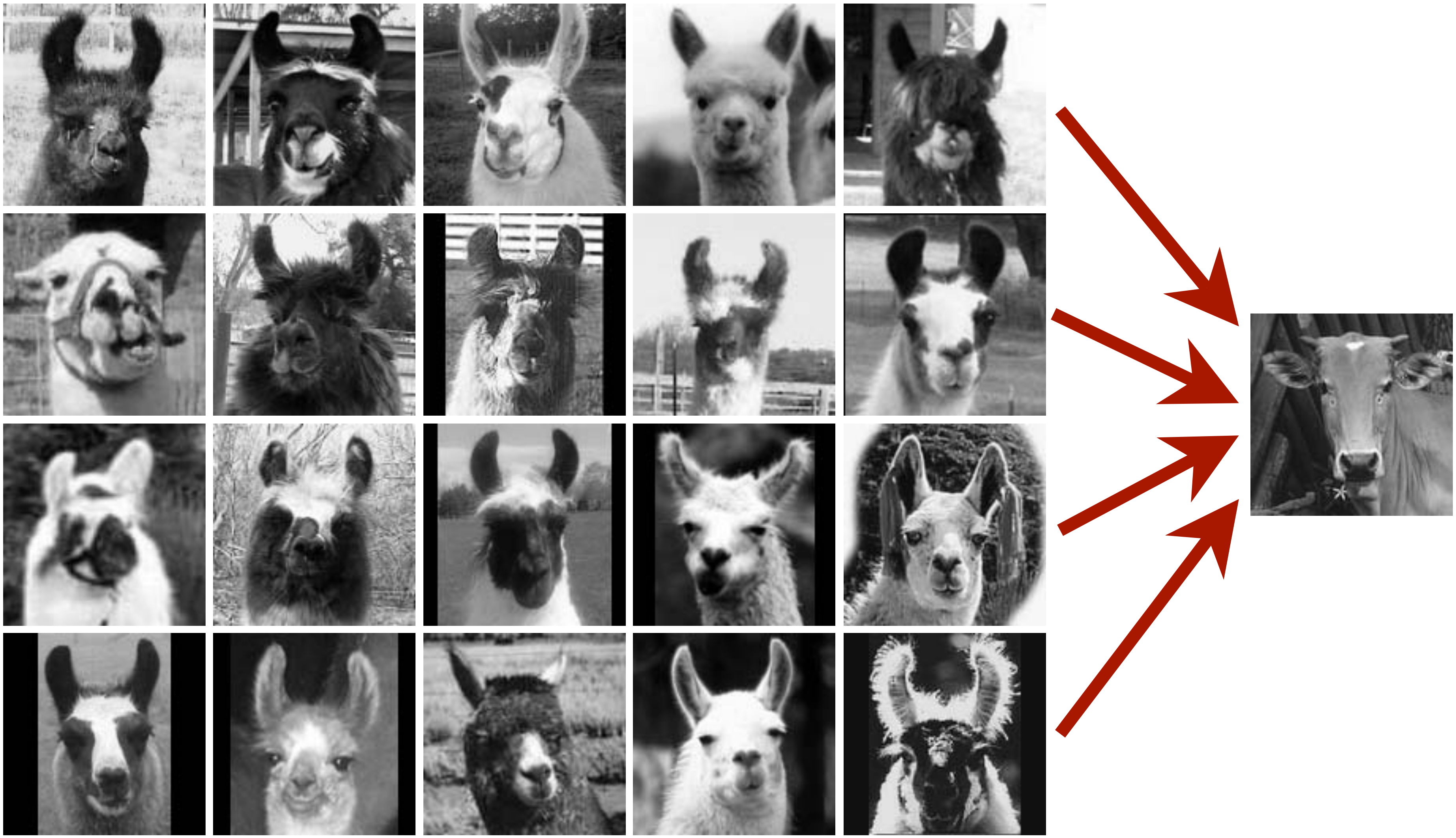
Crouse, Nowak, & Baraniuk, 1998

Denoising: HDP-HMT



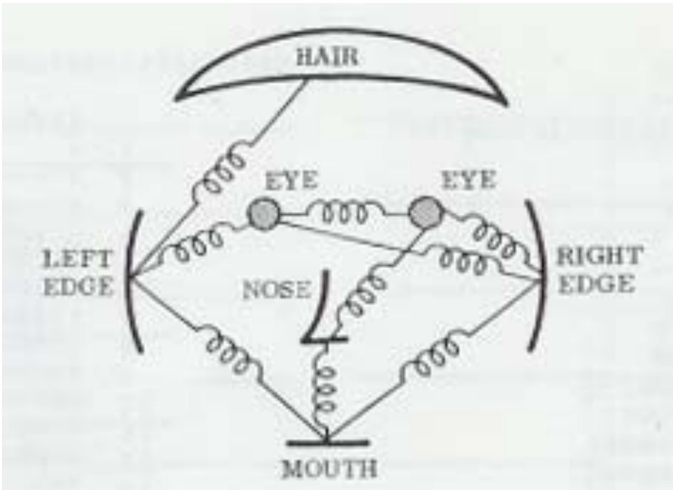
32.10 dB

Visual Object Recognition

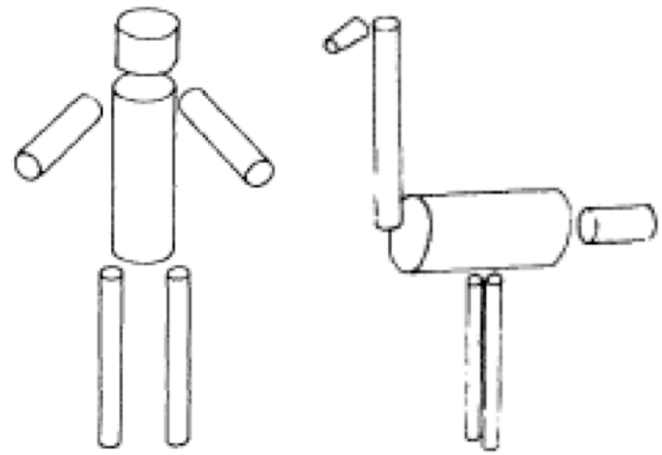


Can we transfer knowledge from one object category to another?

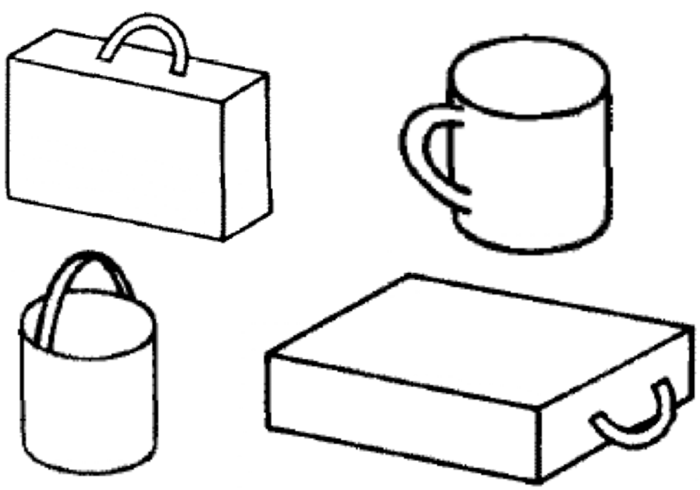
Describing Objects with Parts



Pictorial Structures
Fischler & Elschlager, 1973



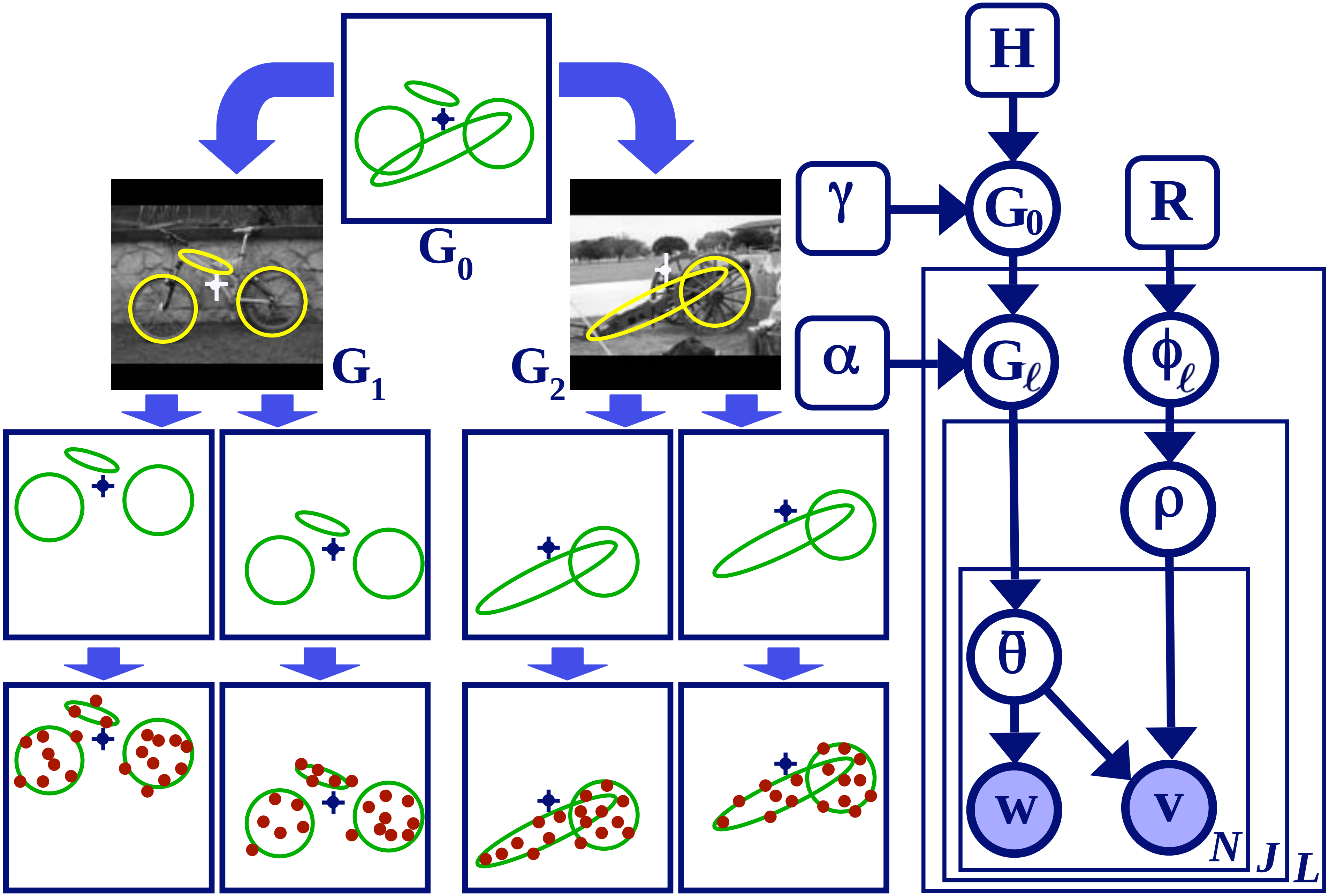
Generalized Cylinders
Marr & Nishihara, 1978



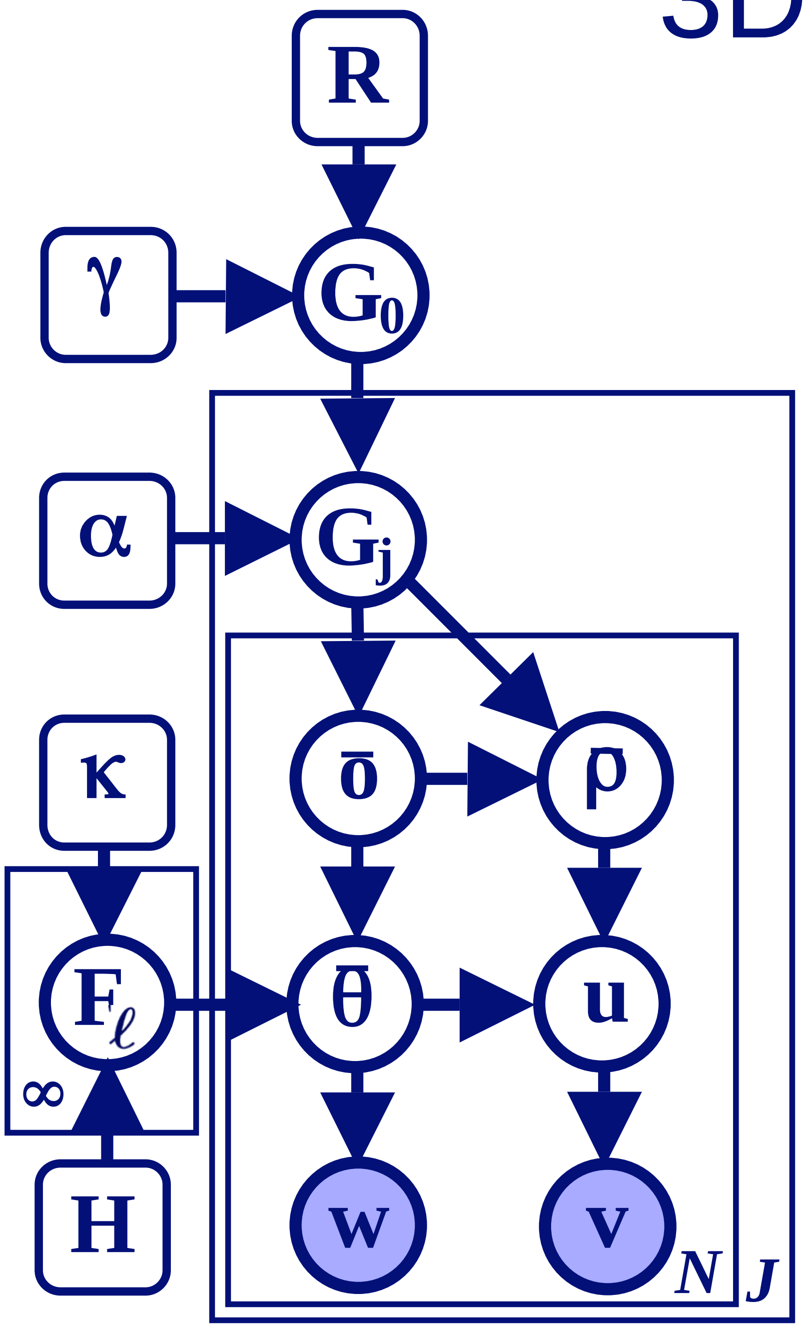
Recognition by Components
Biederman, 1987

Constellation Model
Perona et. al., 2000 to present

A Graphical Model for Object Parts

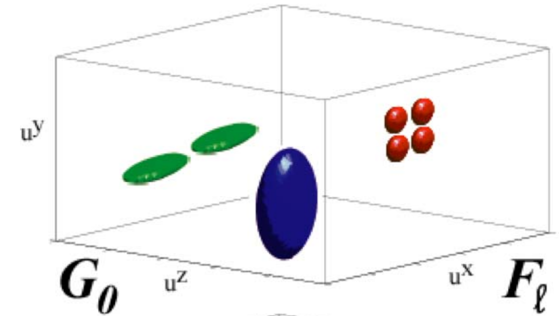


3D Scenes



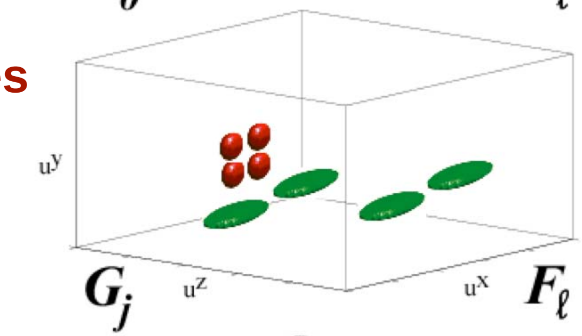
Global Density

Object category
Part size & shape
Transformation prior



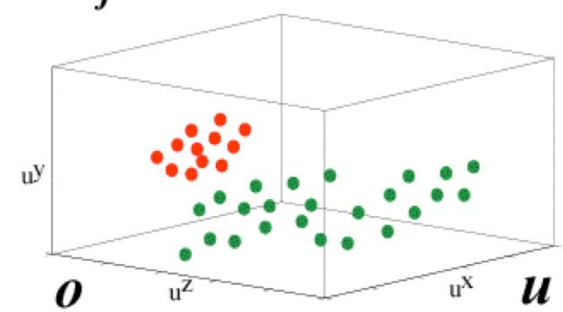
Transformed Densities

Object category
Part size & shape
Transformed locations



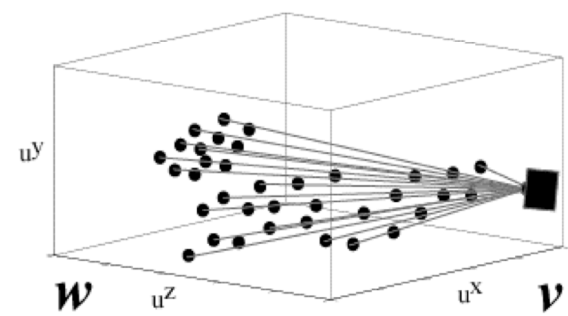
3D Scene Features

Object category
3D Location

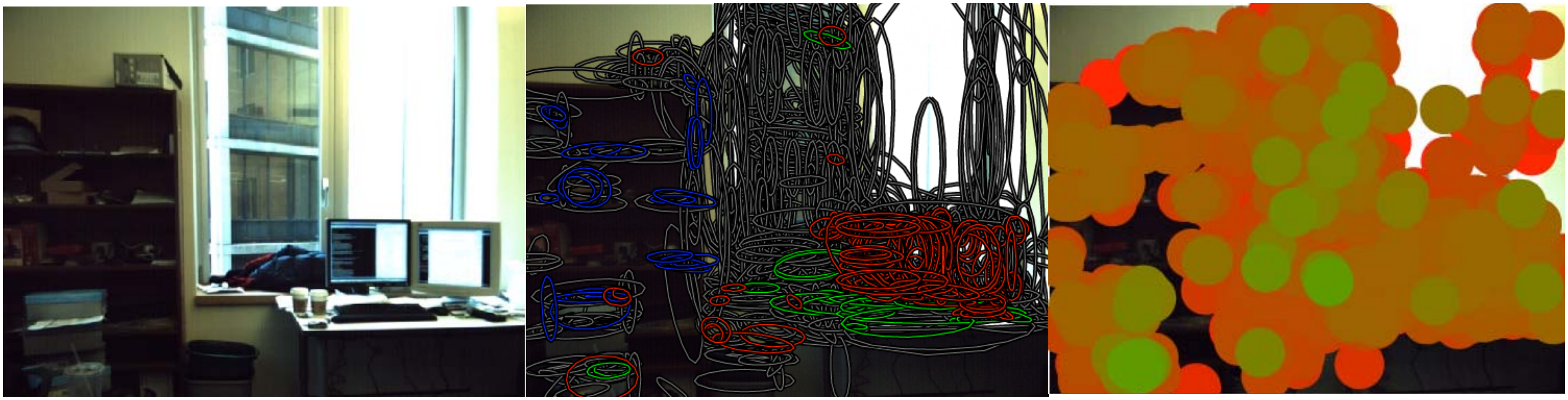


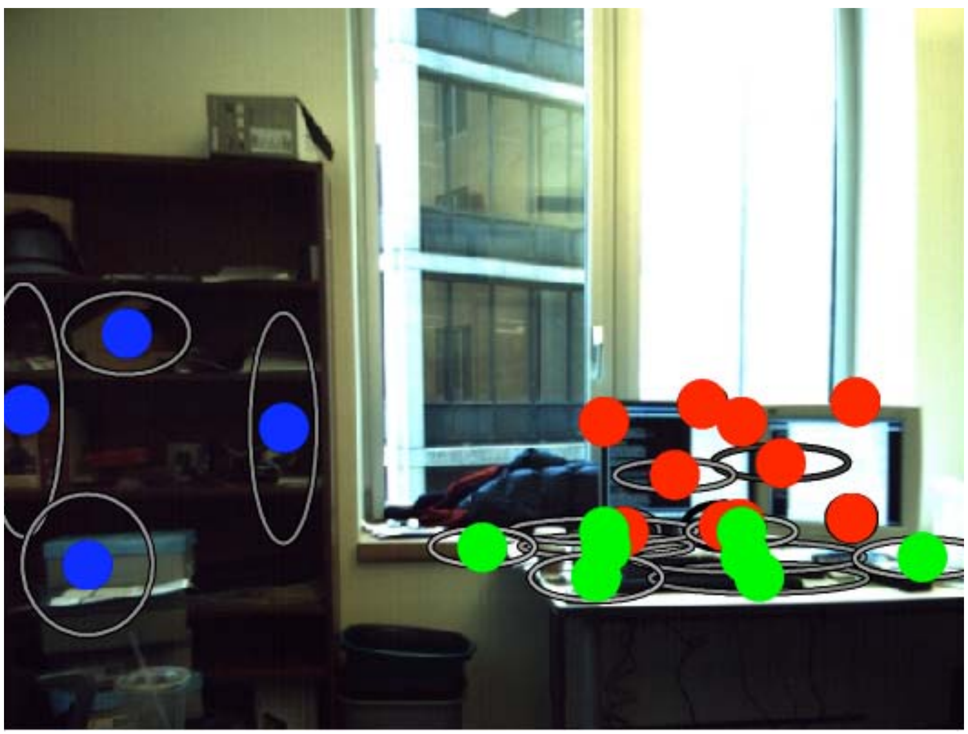
2D Image Features

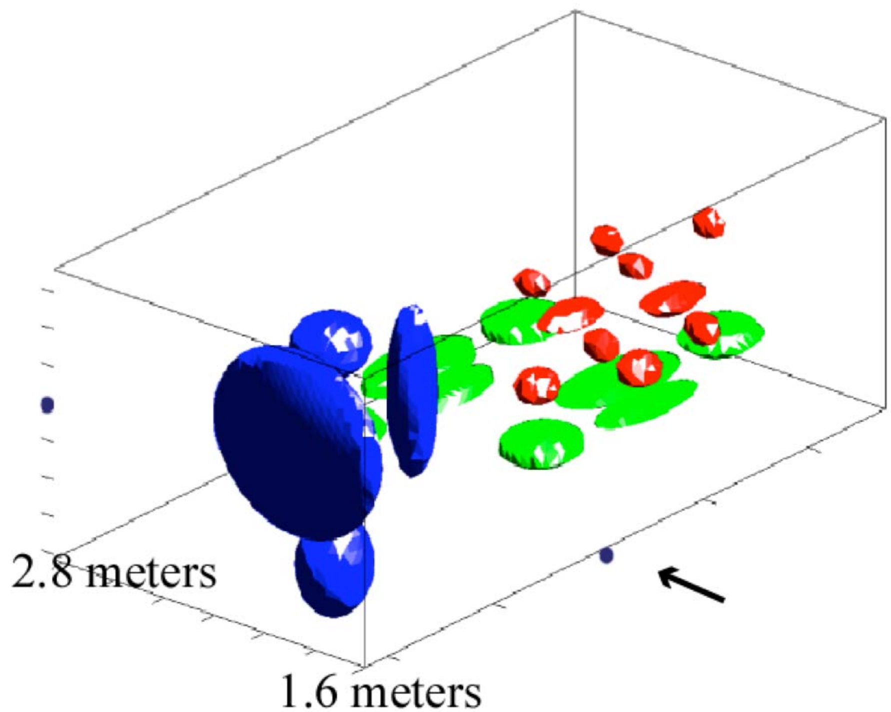
Appearance Descriptors
2D Pixel Coordinates



Stereo Test Image







Many Other Applications

- Speech recognition & speaker diarization
- Natural language processing: parsing, topic models, …
- Robotics: mapping, navigation & control, …
- Error correcting codes & wireless communications
- Bioinformatics
- Nuclear test monitoring
- ………

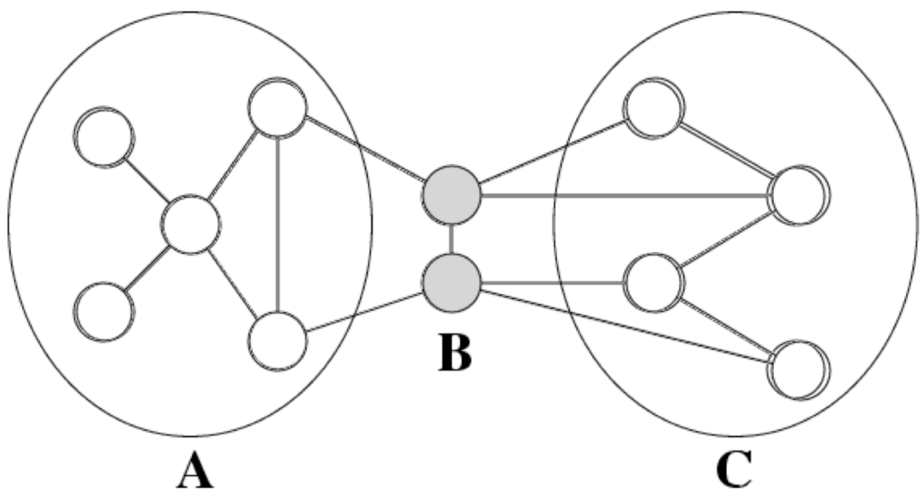
Undirected Graphical Models

An undirected graph $\mathcal{G}$ is defined by

$\mathcal{V} \longrightarrow$ set of N nodes $\{1, 2, \ldots, N\}$

$\mathcal{E} \longrightarrow$ set of edges (s, t) connecting nodes $s, t \in \mathcal{V}$

Nodes $s \in \mathcal{V}$ are associated with random variables x_s



Graph Separation

$\Updownarrow$

Conditional Independence

$$p(x_A, x_C | x_B) = p(x_A | x_B) p(x_C | x_B)$$

Inference in Graphical Models

$$p(x \mid y) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t)$$

y → observations (implicitly encoded via compatibilities)

Maximum a Posteriori (MAP) Estimates

$$\widehat{x} = \arg\max_x \; p(x \mid y)$$

Posterior Marginal Densities

$$p_t(x_t \mid y) = \sum_{x_{\mathcal{V} \setminus t}} p(x \mid y)$$

- Provide both estimators and confidence measures
- Sufficient statistics for iterative *parameter estimation*

Why the Partition Function?

$$Z = \sum_x \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t)$$

Statistical Physics

- Sensitivity of physical systems to external stimuli

Hierarchical Bayesian Models

- Marginal likelihood of observed data
- Fundamental in hypothesis testing & model selection

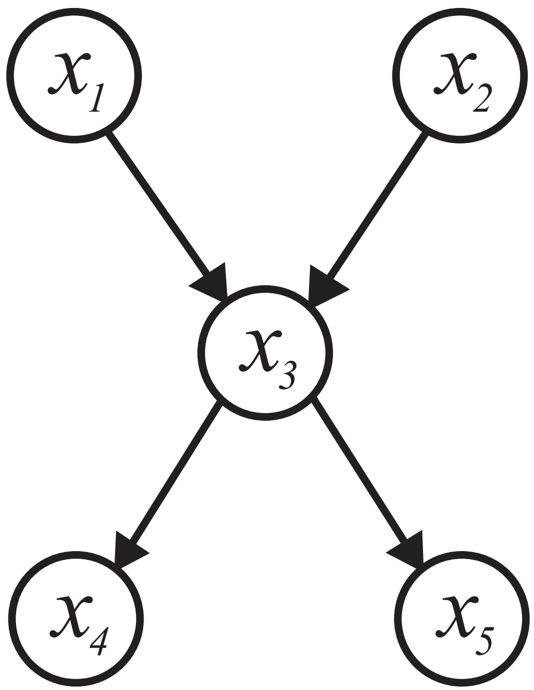
Cumulant Generating Function

- For exponential families, derivatives with respect to parameters provide marginal statistics

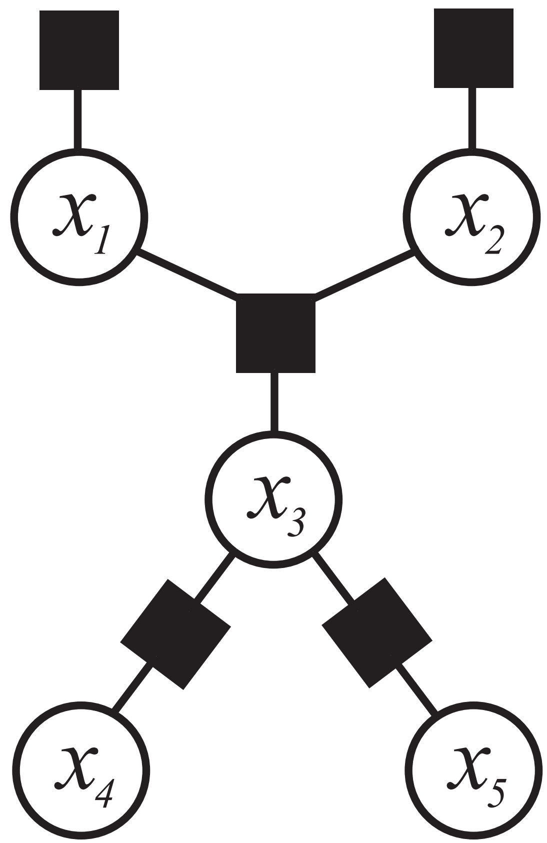
PROBLEM: Computing Z in general graphs is NP-complete

What do you want to learn about?

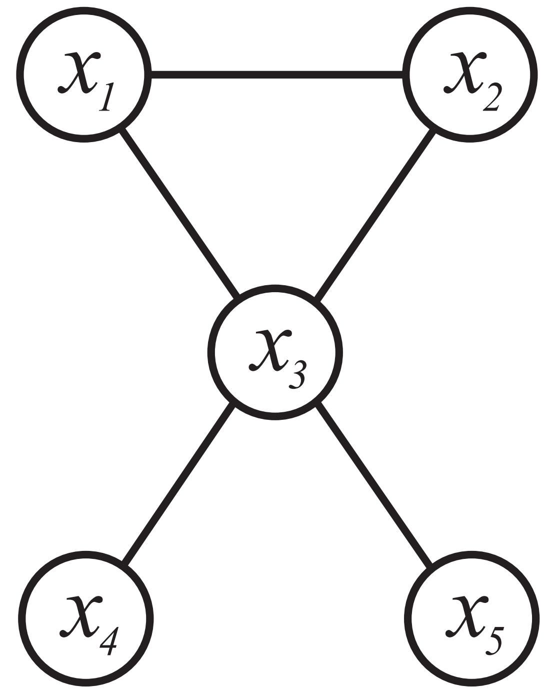
Graphical Models



Directed Bayesian Network



Factor Graph

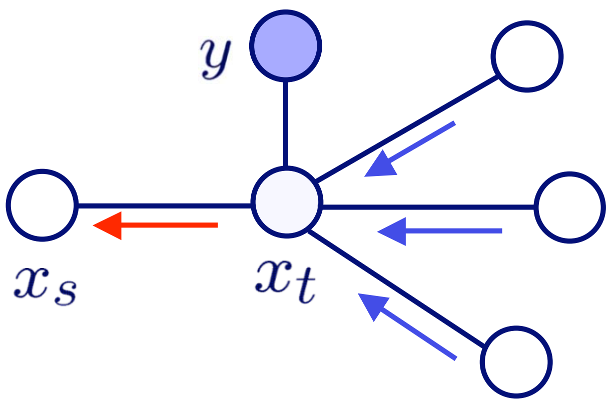


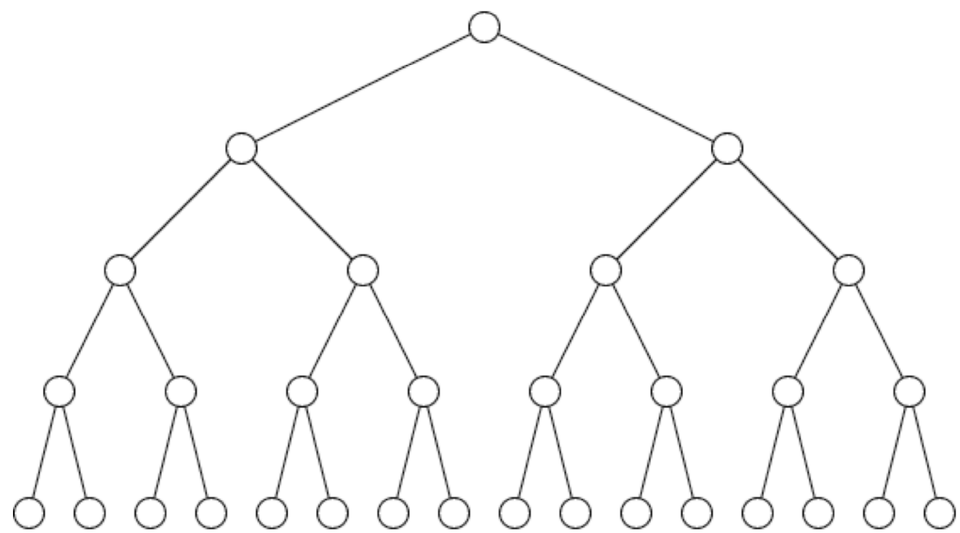
Undirected Graphical Model

Exact Inference

MESSAGES: Sum-product or belief propagation algorithm

$$m_{ts}(x_s) = \alpha \sum_{x_t} \psi_{st}(x_s, x_t)\psi_t(x_t, y) \prod_{u \in \Gamma(t) \backslash s} m_{ut}(x_t)$$





Computational cost:

N → *number of nodes*

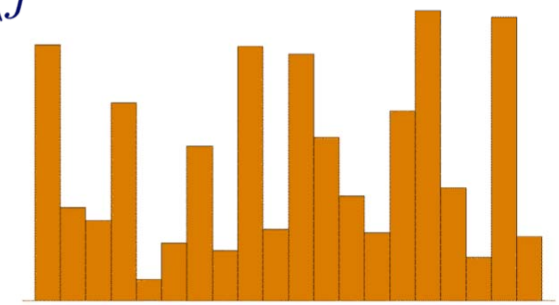
M → *discrete states for each node*

Belief Prop: $\mathcal{O}(NM^2)$

Brute Force: $\mathcal{O}(M^N)$

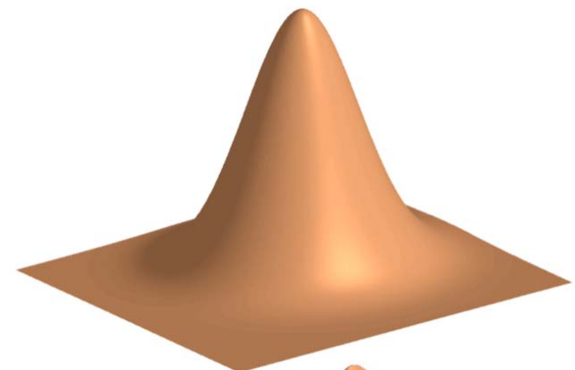
Continuous Variables

$$m_{ij}(x_j) \propto \int_{x_i} \psi_{j,i}(x_j, x_i)\psi_i(x_i, y) \prod_{k \in \Gamma(i) \setminus j} m_{ki}(x_i)\, dx_i$$



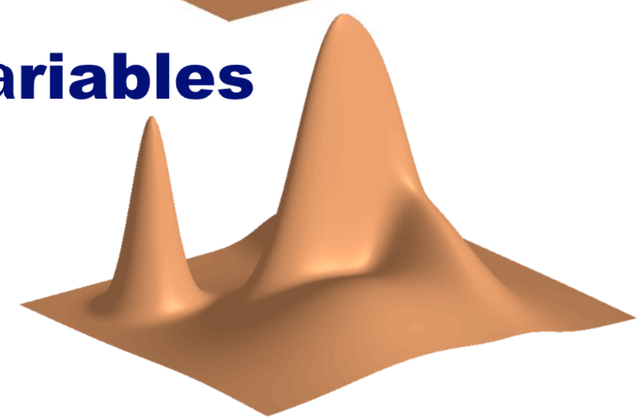
Discrete State Variables

- Messages are *finite vectors*
- Updated via matrix-vector products



Gaussian State Variables

- Messages are *mean & covariance*
- Updated via information Kalman filter



Continuous Non-Gaussian State Variables

- Closed parametric forms unavailable
- Discretization can be *intractable* even with 2 or 3 dimensional states

Variational Inference: An Example

$$p(x \mid y) = \frac{1}{Z} \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s\in\mathcal{V}} \psi_s(x_s, y)$$

- Choose a family of approximating distributions which is tractable. The simplest example:

$$q(x) = \prod_{s\in\mathcal{V}} q_s(x_s)$$

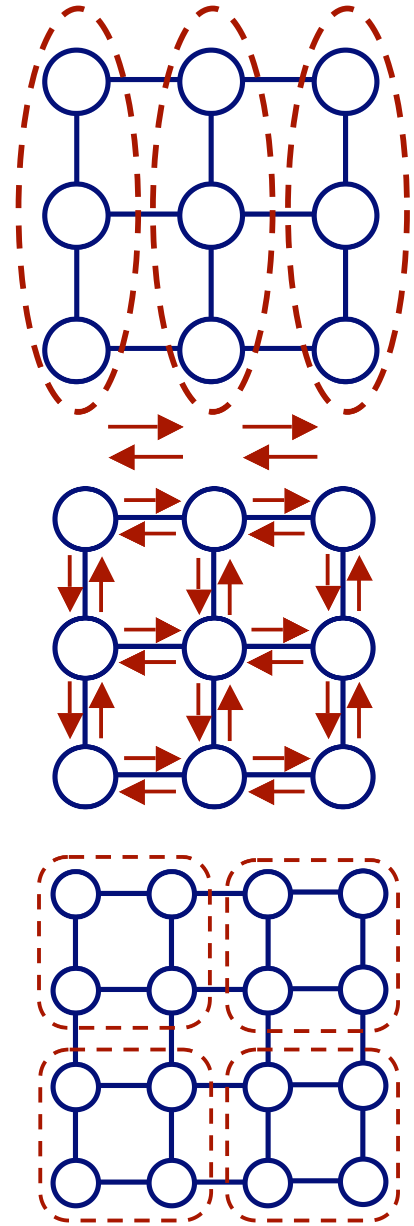
- Define a distance to measure the quality of different approximations. One possibility:

$$D(q \,||\, p) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)}$$

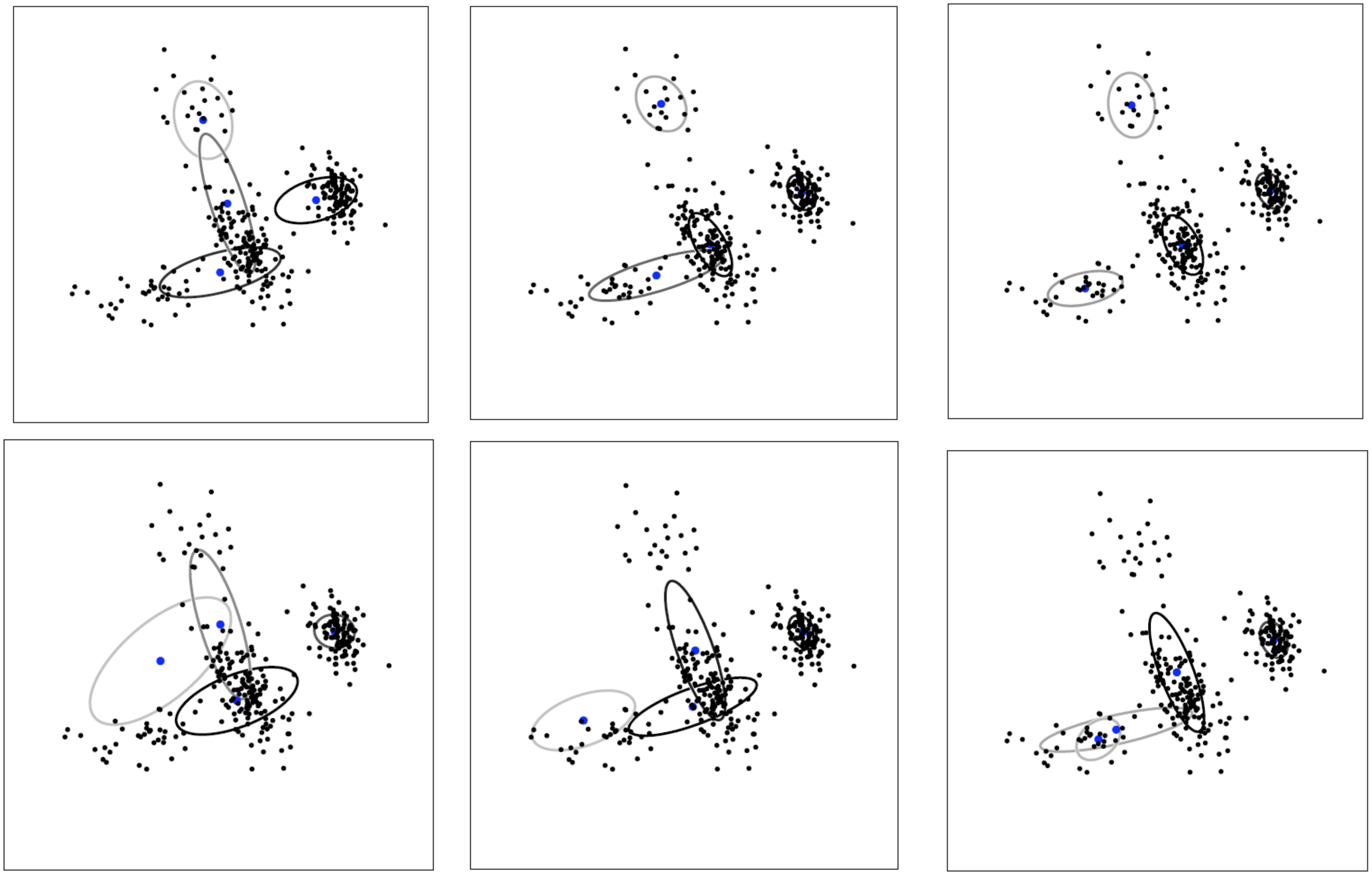
- Find the approximation minimizing this distance

Advanced Variational Methods

- Exponential families
- Mean field methods: naïve and structured
- Variational EM for parameter estimation
- Loopy belief propagation (BP)
- Bethe and Kikuchi entropies
- Generalized BP, fractional BP
- Convex relaxations and bounds
- MAP estimation and linear programming
- ………



Markov Chain Monte Carlo



...

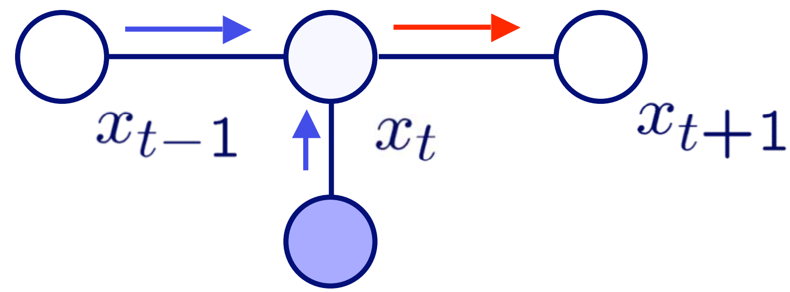
...

Metropolis-Hastings, Gibbs sampling, Rao-Blackwellization, …

Sequential Monte Carlo

Particle Filters, Condensation, Survival of the Fittest,…

- Nonparametric approximation to optimal BP estimates
- Represent messages and posteriors using a set of samples, found by simulation



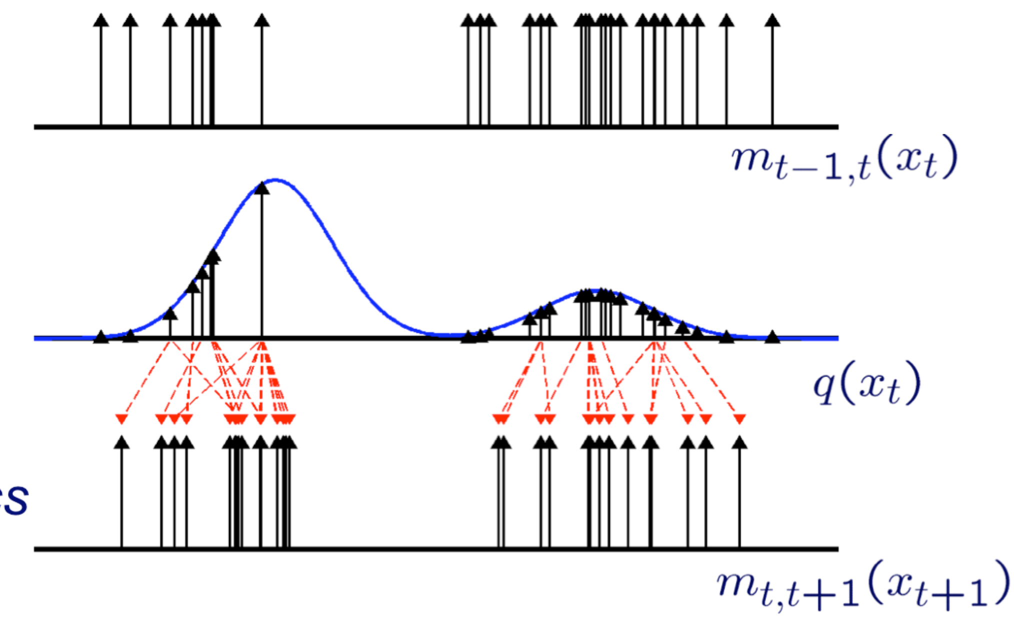
Sample-based density estimate

↓

Weight by observation likelihood

↓

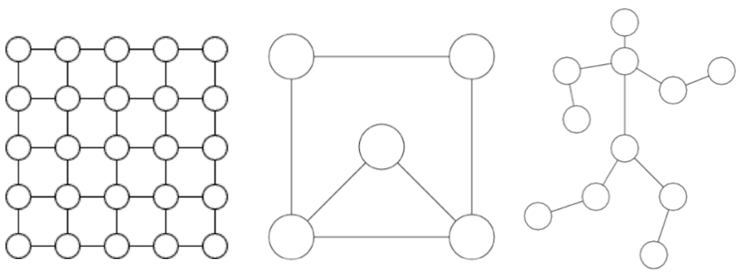
Resample & propagate by dynamics

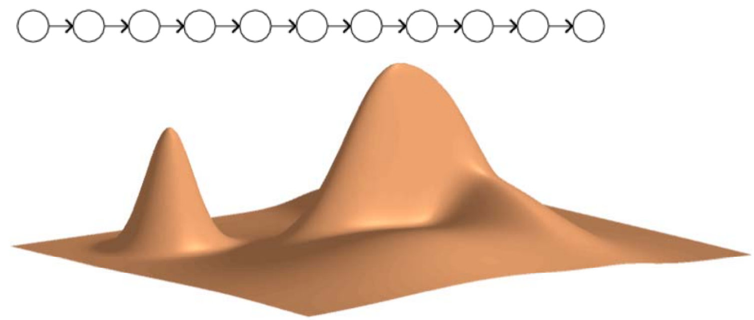


Nonparametric Belief Propagation

Belief Propagation
- General graphs
- Discrete or Gaussian

Particle Filters
- Markov chains
- General potentials



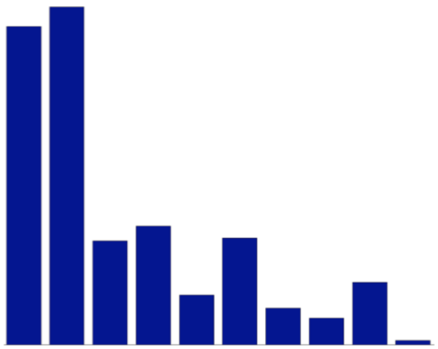


Nonparametric BP
- General graphs
- General potentials

Nonparametric Bayes

$$p(x) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x \mid 0, \Lambda_k)$$

Dirichlet process mixture model



Nonparametric $\neq$ No Parameters

- Model complexity grows as data observed:
  - Small training sets give *simple, robust* predictions
  - Reduced sensitivity to prior assumptions

Flexible but Tractable

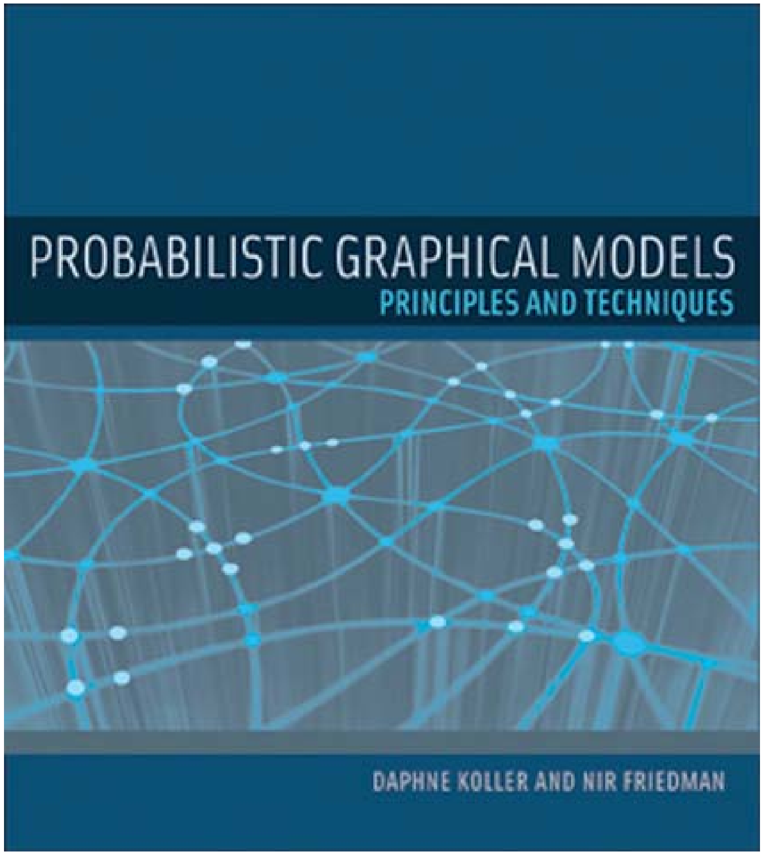
- Literature showing attractive *asymptotic properties*
- Leads to simple, effective *computational methods*
  - Avoids challenging model selection issues

Prereq: Intro Machine Learning

	Supervised Learning	***Unsupervised Learning***
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- Bayesian and frequentist estimation
- Model selection, cross-validation, overfitting
- Expectation-Maximization (EM) algorithm

Textbook & Readings



- Variational tutorial by Wainwright and Jordan (2008)
- Background chapter of Prof. Sudderth's thesis
- Many classic and contemporary research articles…

Grading

Class Participation: 30%

- Attend class and participate in discussions
- Prepare summary overview presentation, and lead class discussion, for ~2 papers
  - Prof. Sudderth will lecture 50% of the time
- Upload comments about the assigned reading before each lecture (due at 9am)

Final Project: 70%

- Proposal: 1-2 pages, due in March (10%)
- Presentation: ~10 minutes, during finals week (10%)
- Conference-style technical report (50%)

Reading Comments

The Good: 1-2 sentences

- What is the most exciting or interesting model, idea, or technique described here? Why is it important?
- Don't just copy the abstract - what do *you* think?

The Bad: 1-2 sentences

- No method is perfect, and many are far from it!
- What is the biggest weakness of this model or approach?
- Problems could be a lack of empirical validation, missing theory, unacknowledged assumptions, …

The Ugly: 1-2 sentences

- Poorly written or unclear sections of the paper: terse explanations, steps you didn't follow, etc.
- What would you like to have explained in class?

Final Projects

Best case: Application of course material to your own area of research

Key Requirements: Novelty, use of graphical models

- Propose a new family of graphical models suitable for a particular application, try baseline learning algorithms
- Propose, develop, and experimentally test an extension of some existing learning or inference algorithm
- Experimentally compare different models or algorithms on an interesting, novel dataset
- Survey the latest advances in a particular application area, or for a particular type of learning algorithm
- …

Administration

Mailing List: E-mail sudderth@cs.brown.edu with

- Your name
- Your CS account username
- Your department, major, and year
- Your experience in machine learning
  - If you took CS195-F in Fall 2009, just say so
  - Otherwise, 1-2 sentences about previous exposure

Readings for Monday:

- Introductory chapters of Koller & Friedman; specific sections announced via e-mail
- No comments required for Monday’s lecture