Computer Science

CSC535: Probabilistic Graphical Models

Course Wrap-Up

Prof. Jason Pacheco

Some material from Prof. Erik Sudderth

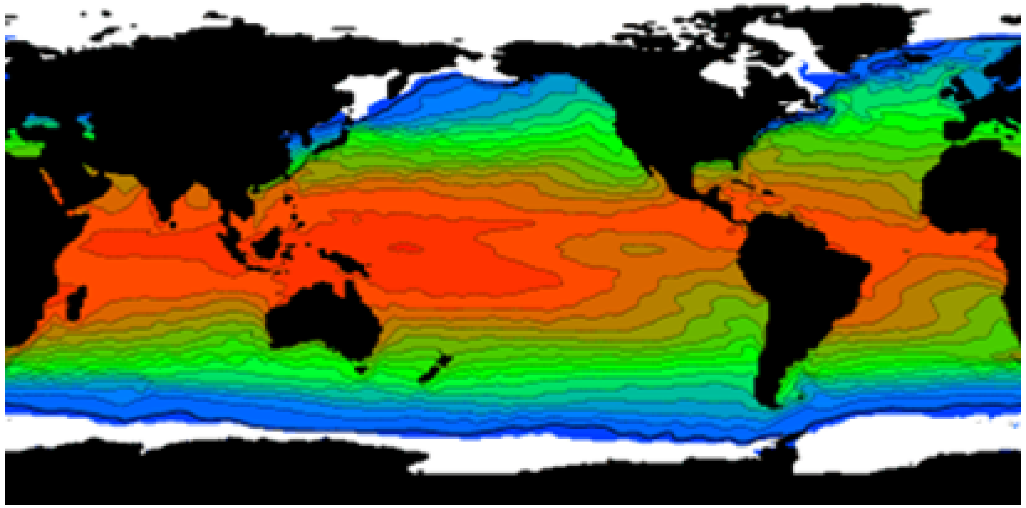
Final Exam

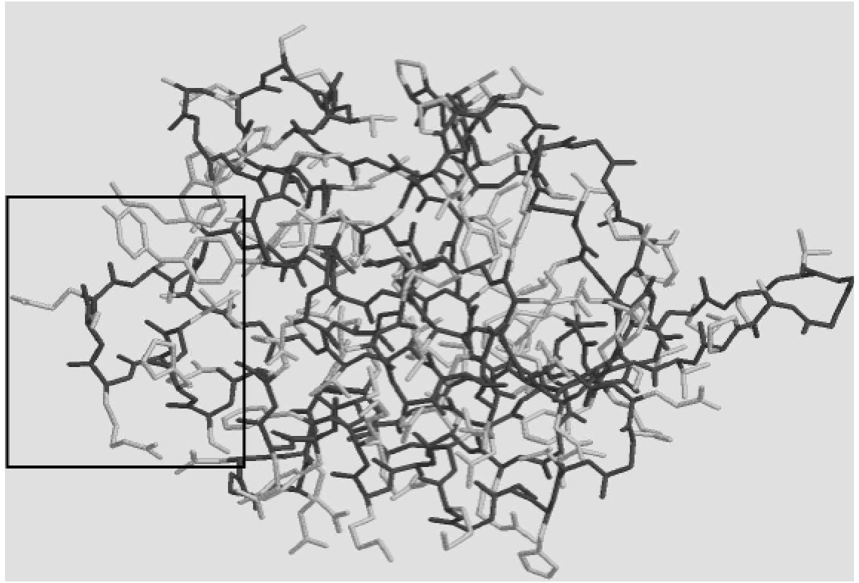
- Out by Monday morning
- Due 11:59pm Wednesday (6/11)
- 4 Questions (5 points each) + 1 Extra Credit

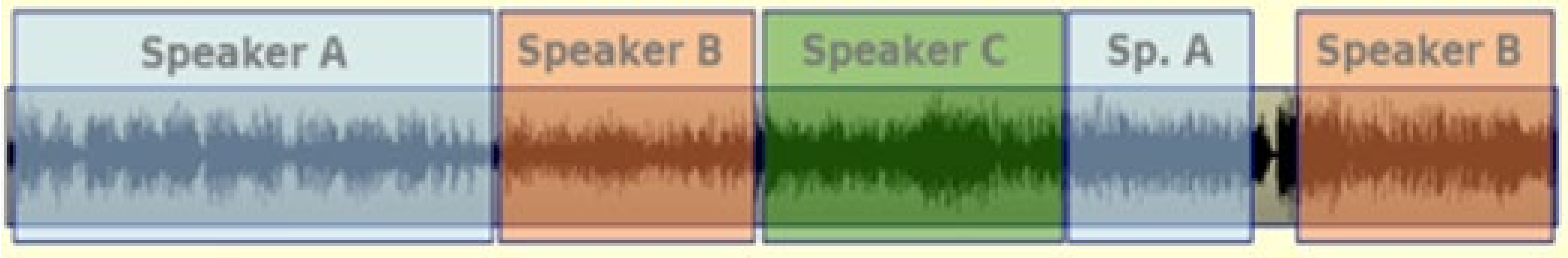
Topics

- PGM models, probability
- Gibbs sampling (compute complete conditionals)
- Expectation Maximization
- Mean Field (compute update, extra credit)

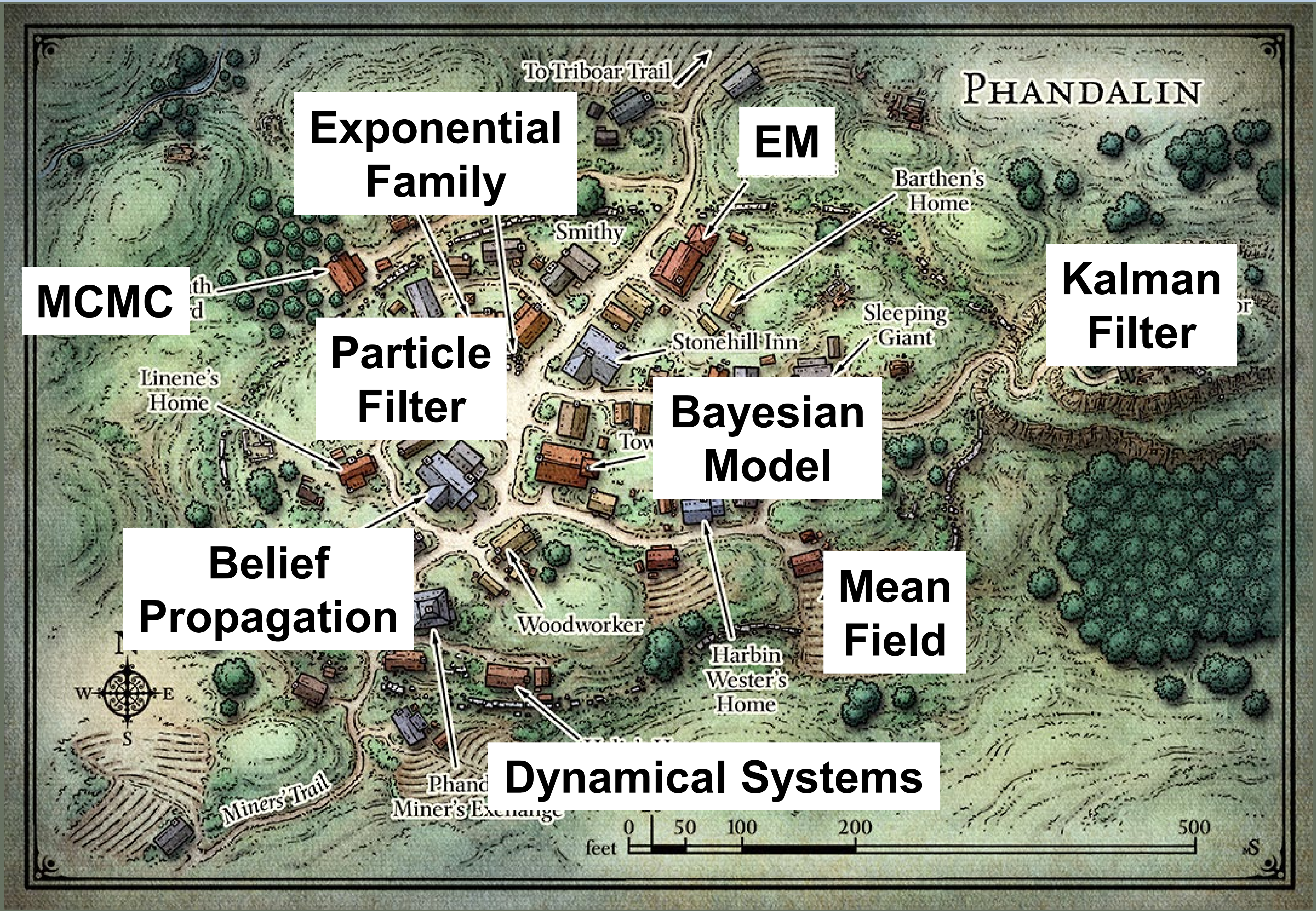
Learning from Structured Data







Roadmap for ML Practice & Research



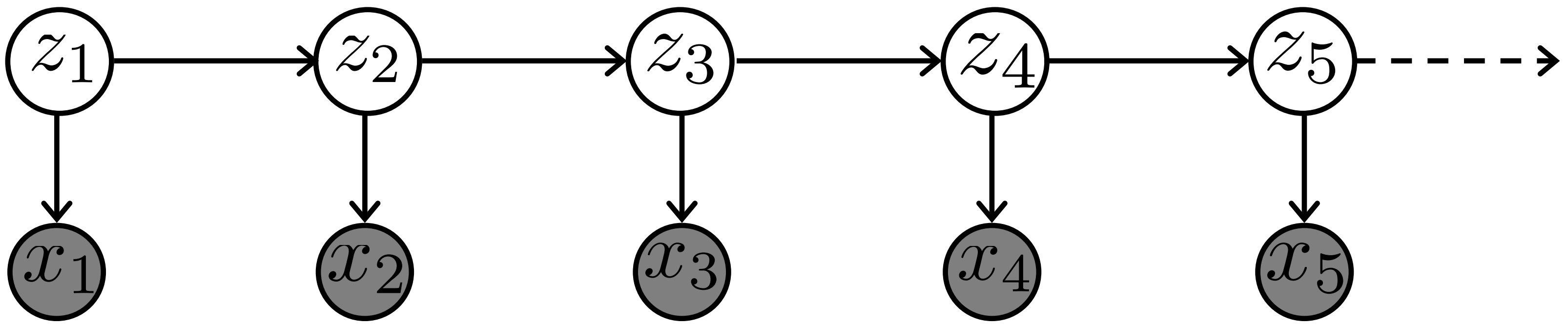
What we covered…

Probability and Statistics	Message Passing Algorithms	Parameter Learning	Monte Carlo Methods	Dynamical Systems	Variational Inference
Probability primer, Bayesian statistics, PGMs, Exponential families	Elimination, Junction tree, Sum-product / max-product, Belief propagation, Viterbi decoding	Maximum likelihood, Maximum a posteriori, Expectation Maximization (EM)	Rejection sampling, Importance sampling, Metropolis-Hastings, Gibbs	Linear and switching state-space models, Kalman filter, Particle filter	Mean field, Stochastic variational, Bethe energy methods

There's so much more to cover…

Models & Applications	Bayesian Deep Learning	Representation Learning	Bayesian Nonparametrics	Advanced MCMC	Still more…
Course was mostly focused on algorithms, limited attention to modelling	Probabilistic uncertainty models for deep learning	Unsupervised representation learning from structured data	A class of probability models where model complexity is inferred from the data	Avoiding random walk dynamics and allowing parallel computatoin	

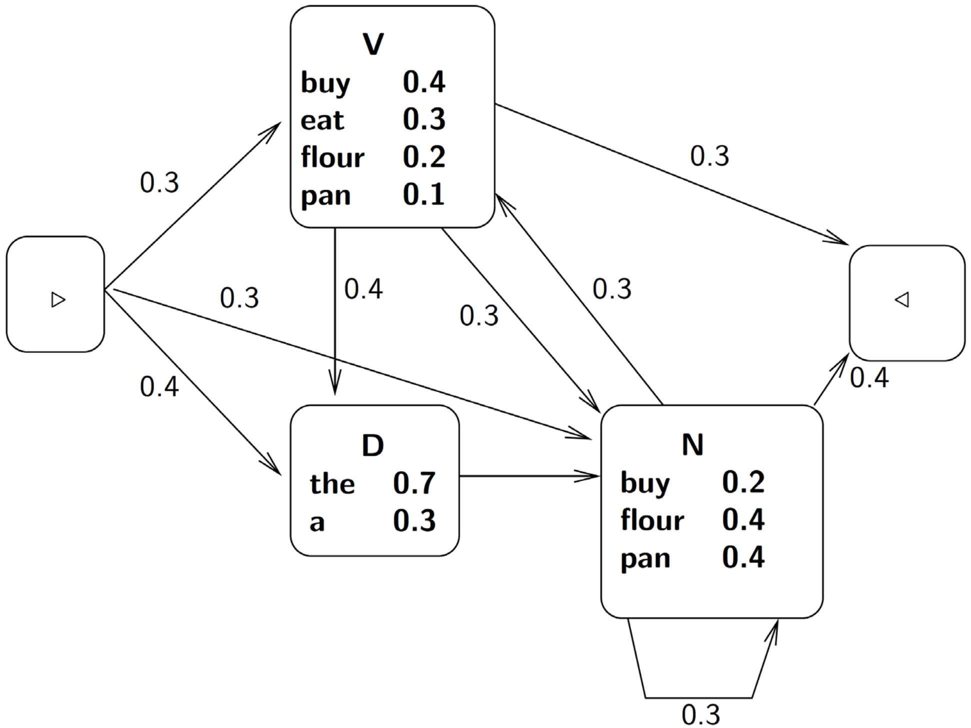
Hidden Markov Models (HMMs)

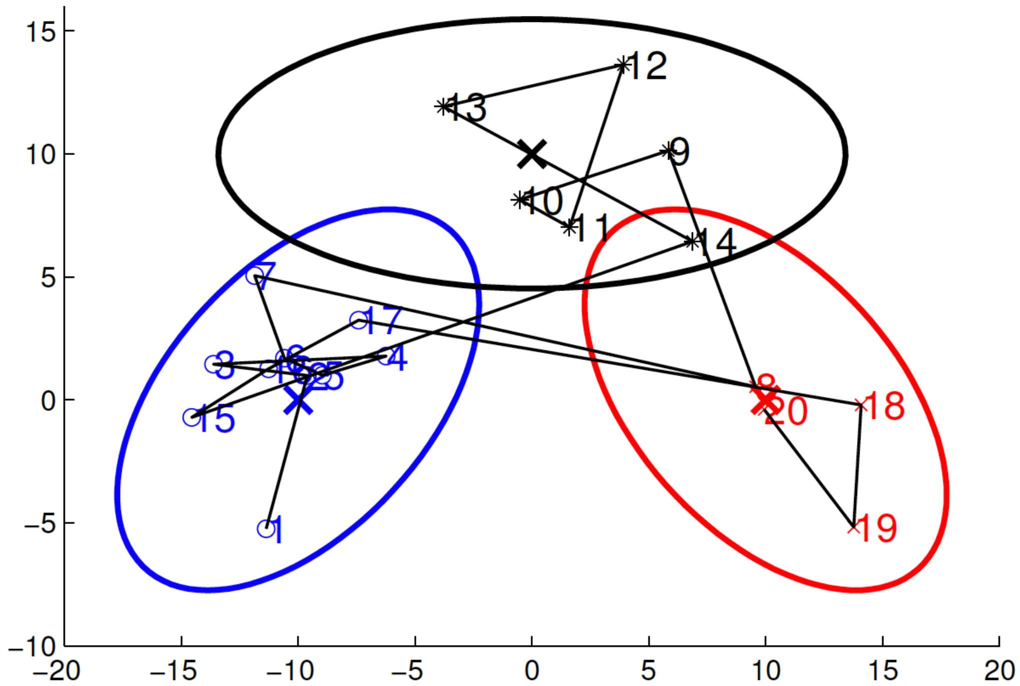


$z_t \rightarrow$ Hidden states taking 1 of *K* discrete values.

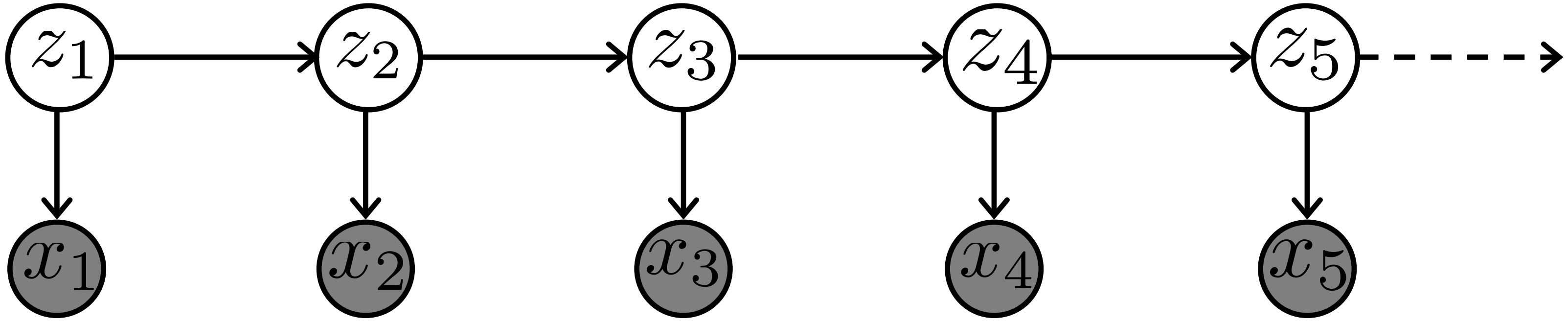
$x_t \rightarrow$ Observations taking values in any space.

$$p(z,x) = p(z)p(x \mid z) = \left[p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \right] \cdot \left[\prod_{t=1}^{T} p(x_t \mid z_t) \right]$$





Example: Sequence Labeling in NLP



Part of speech (POS) tagging:

$\mathbf{z}$:	DT	JJ	NN	VBD	NNP	.
$\boldsymbol{x}$:	the	big	cat	bit	Sam	.

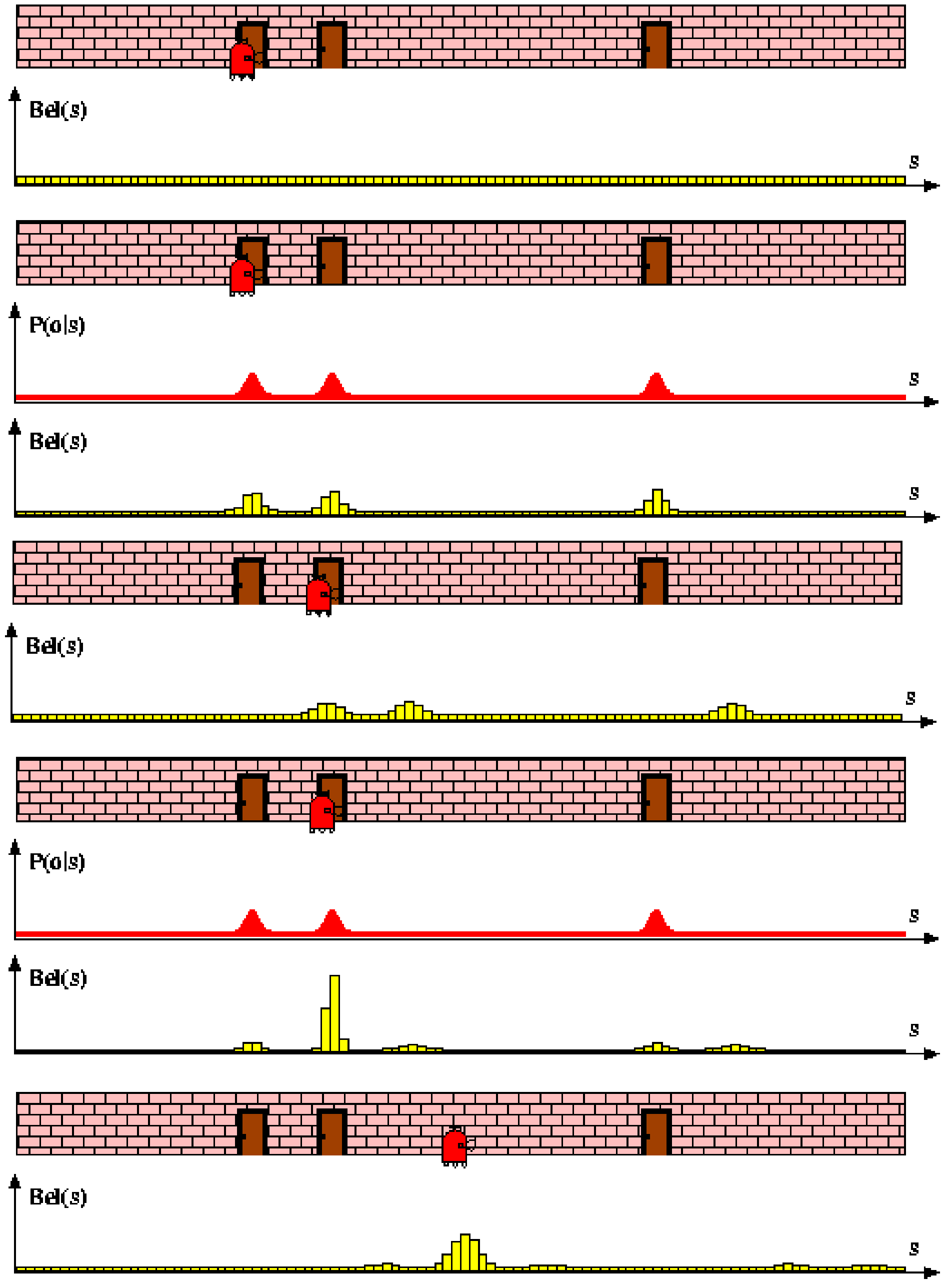
Named entity detection:

$\mathbf{z}$:	[CO	CO]	_	[LOC]	_	[PER]	_
$\boldsymbol{x}$:	XYZ	Corp.	of	Boston	announced	Spade's	resignation

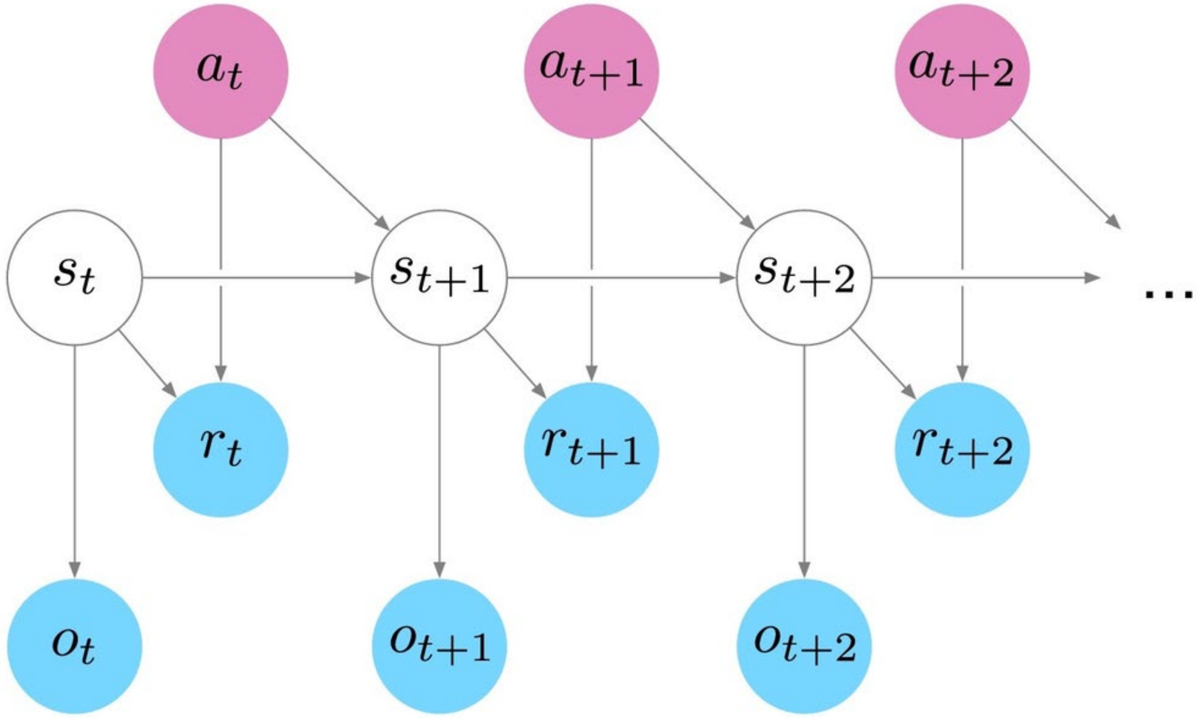
Speech recognition: The $\boldsymbol{x}$ are 100 msec. time slices of acoustic input, and the $\mathbf{z}$ are the corresponding phonemes (i.e., $\mathbf{z}_i$ is the phoneme being uttered in time slice x_i)

M. Johnson, 2009

HMM Localization for Mobile Robots

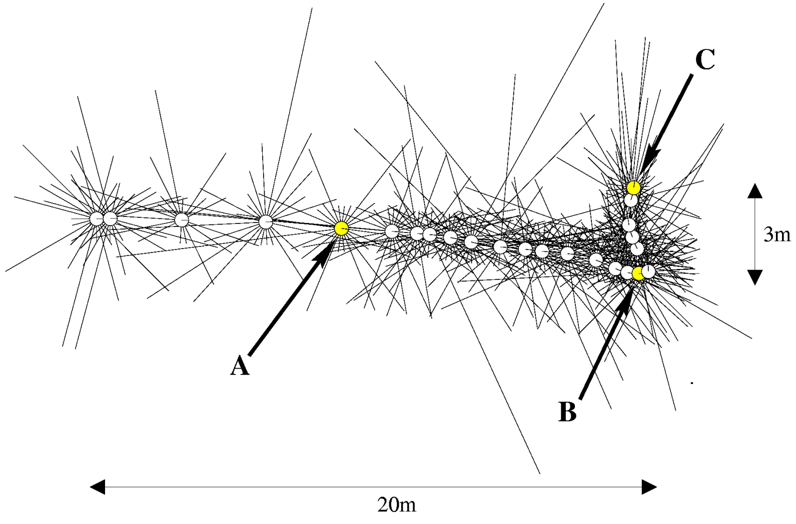


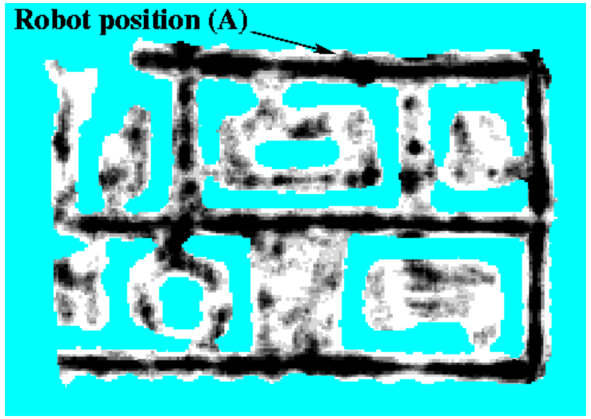
(a) Partially observable Markov decision process (POMDP)

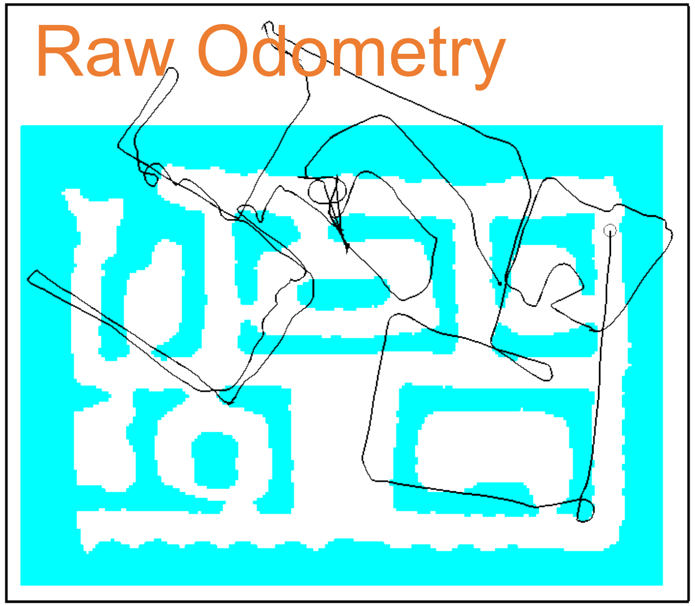


Fox, Burgard, & Thrun, JAIR 1999
Probabilistic Robotics, 2006

HMM Localization for Mobile Robots

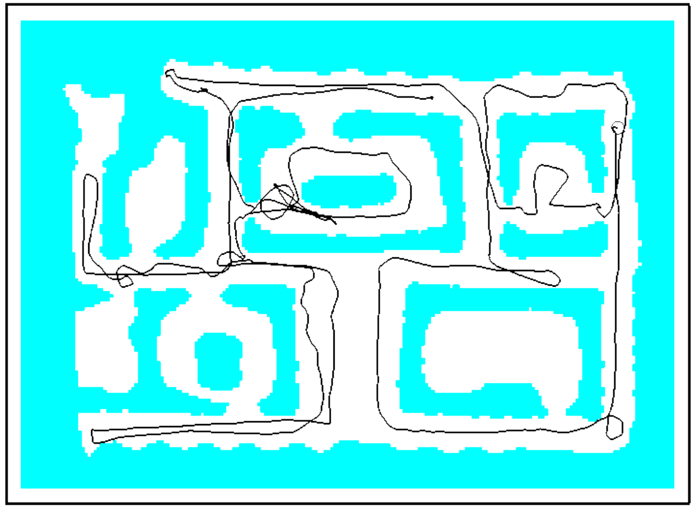




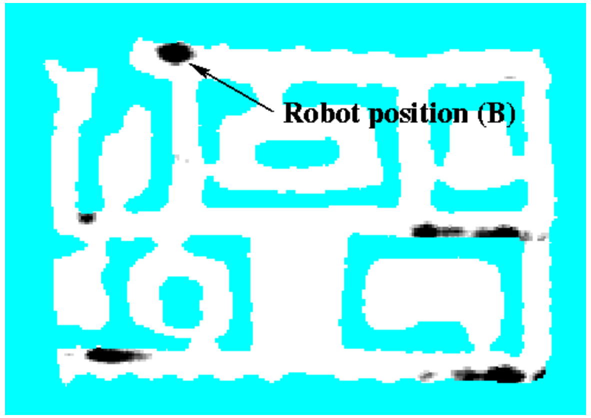


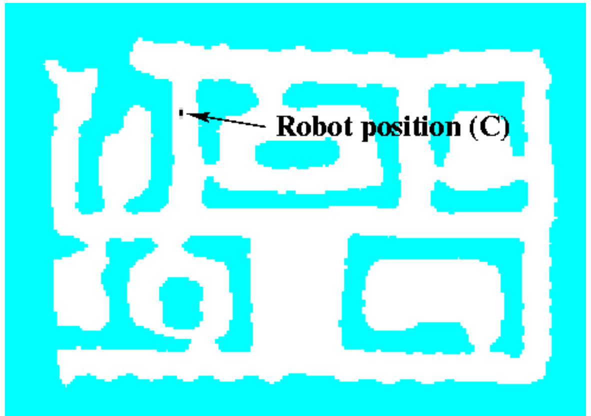
(a)

HMM Estimate

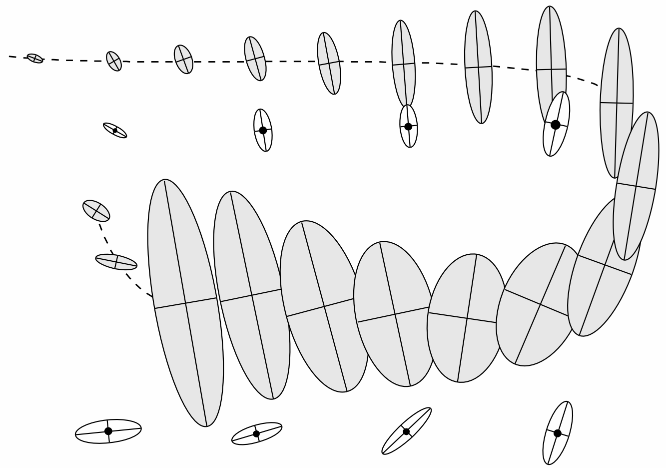


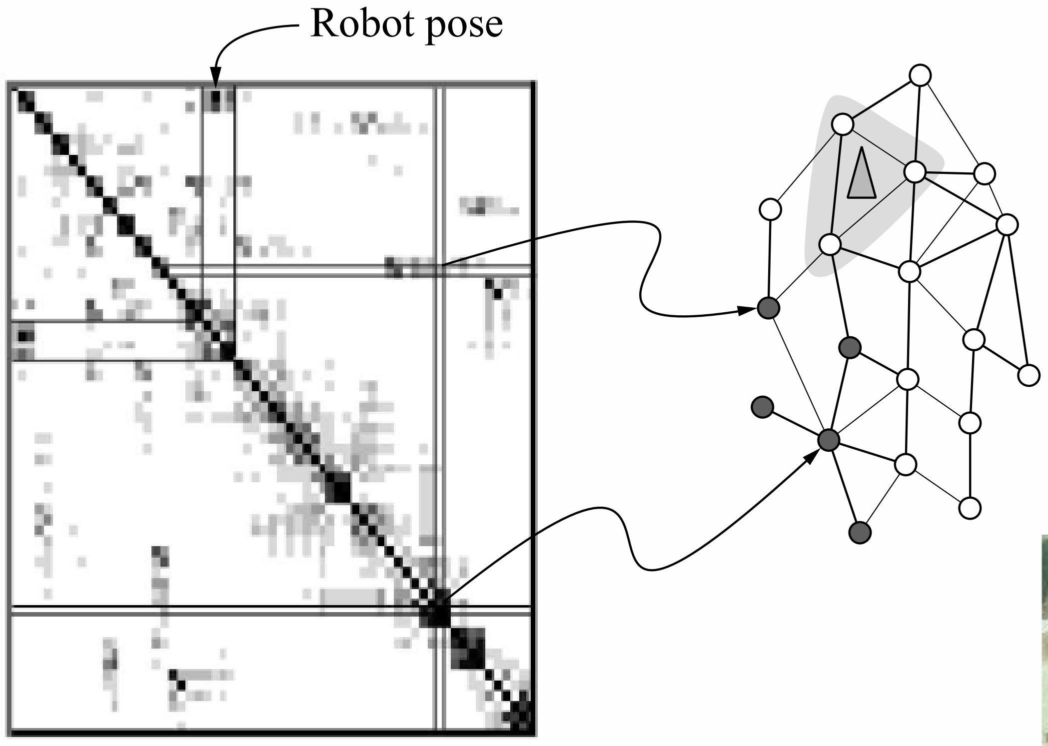
(b)

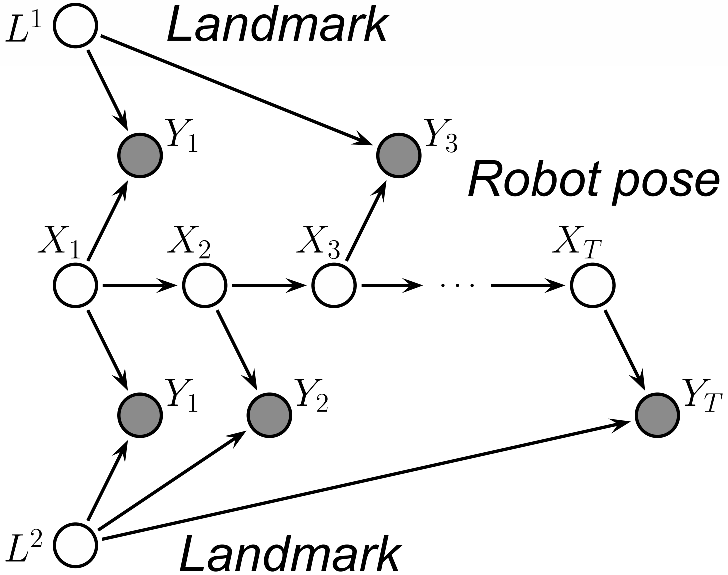




Simultaneous Localization & Mapping

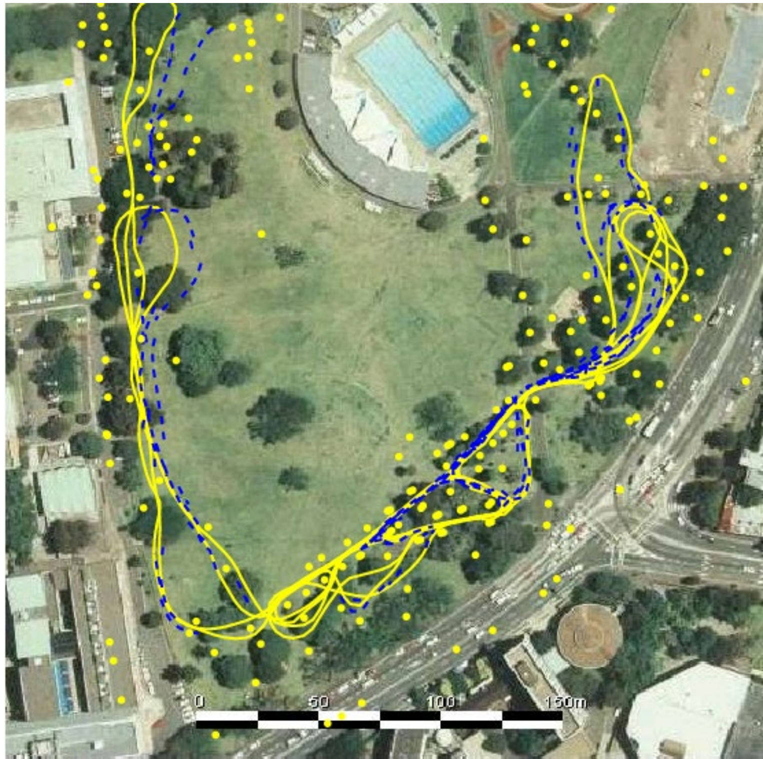






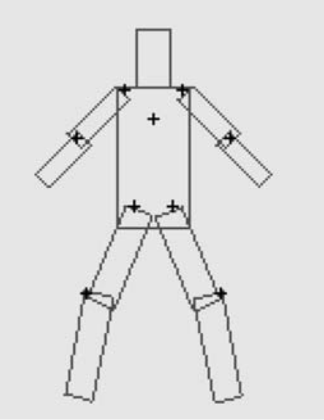
Landmark SLAM

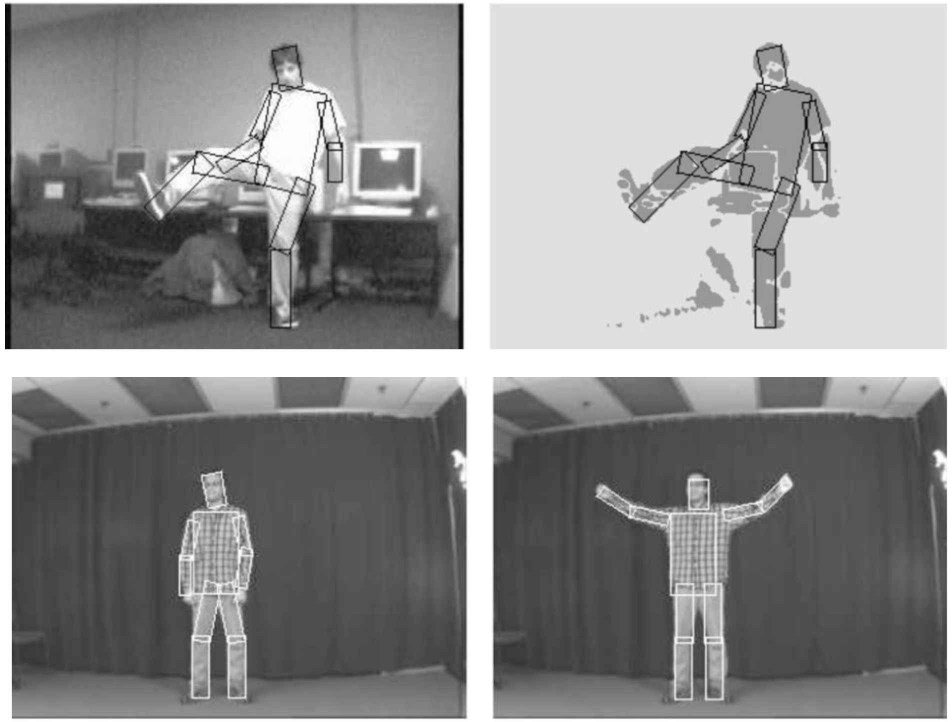
(E. Nebot, Victoria Park)



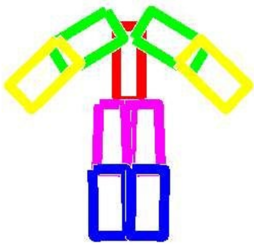
Pose Estimation & Tree-Structured Graphs

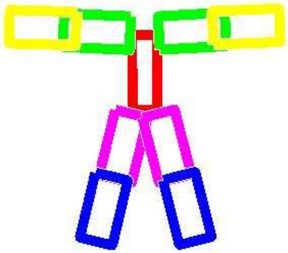
Felzenszwalb & Huttenlocher, 2005

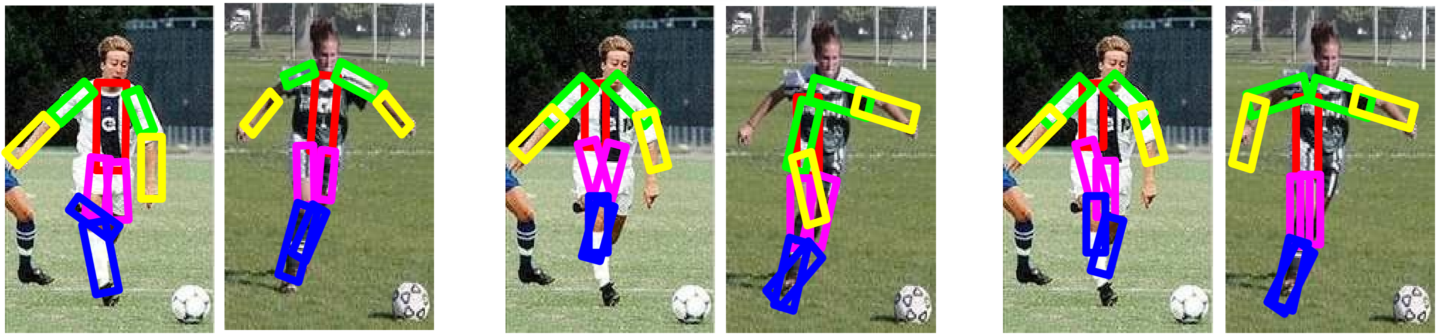




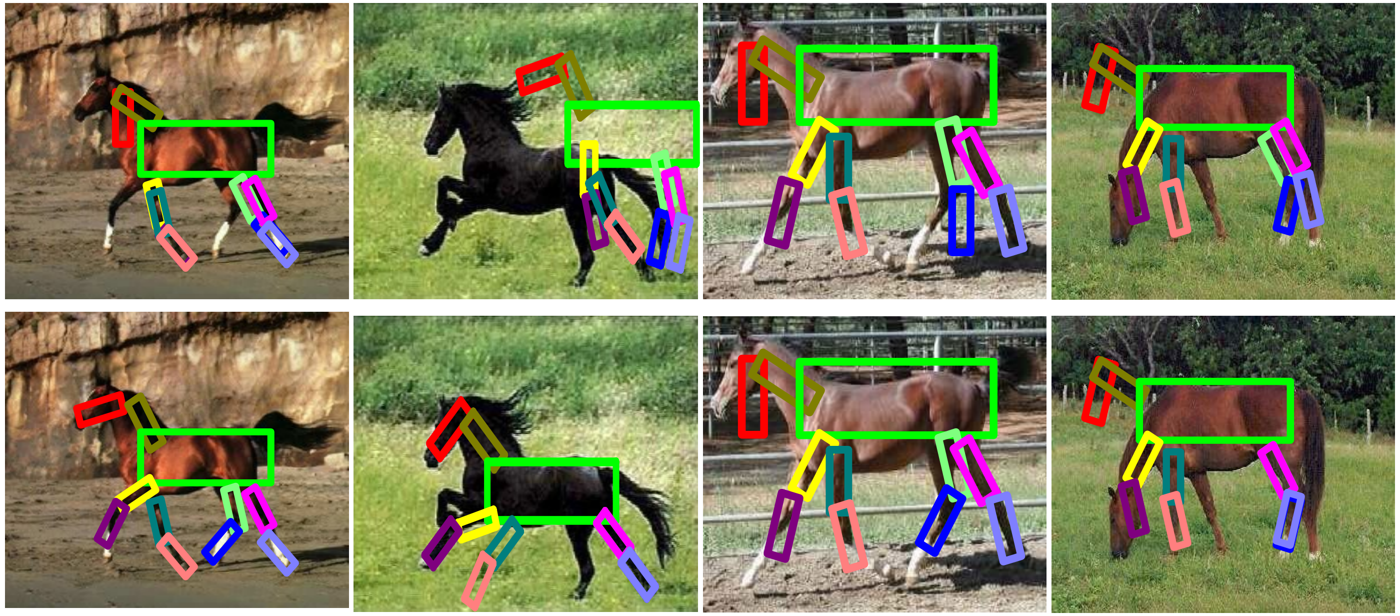
Ramanan & Sminchisescu, 2006



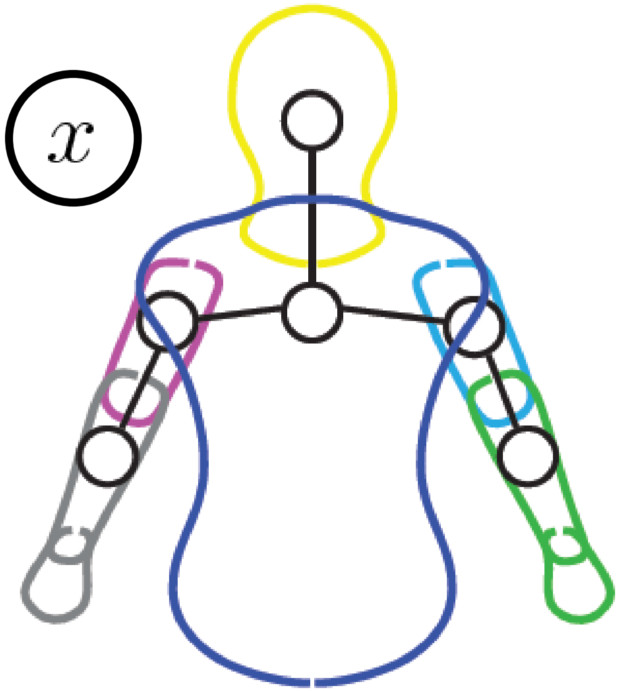




Training Data | Maximum Likelihood Model | Conditional Likelihood Model



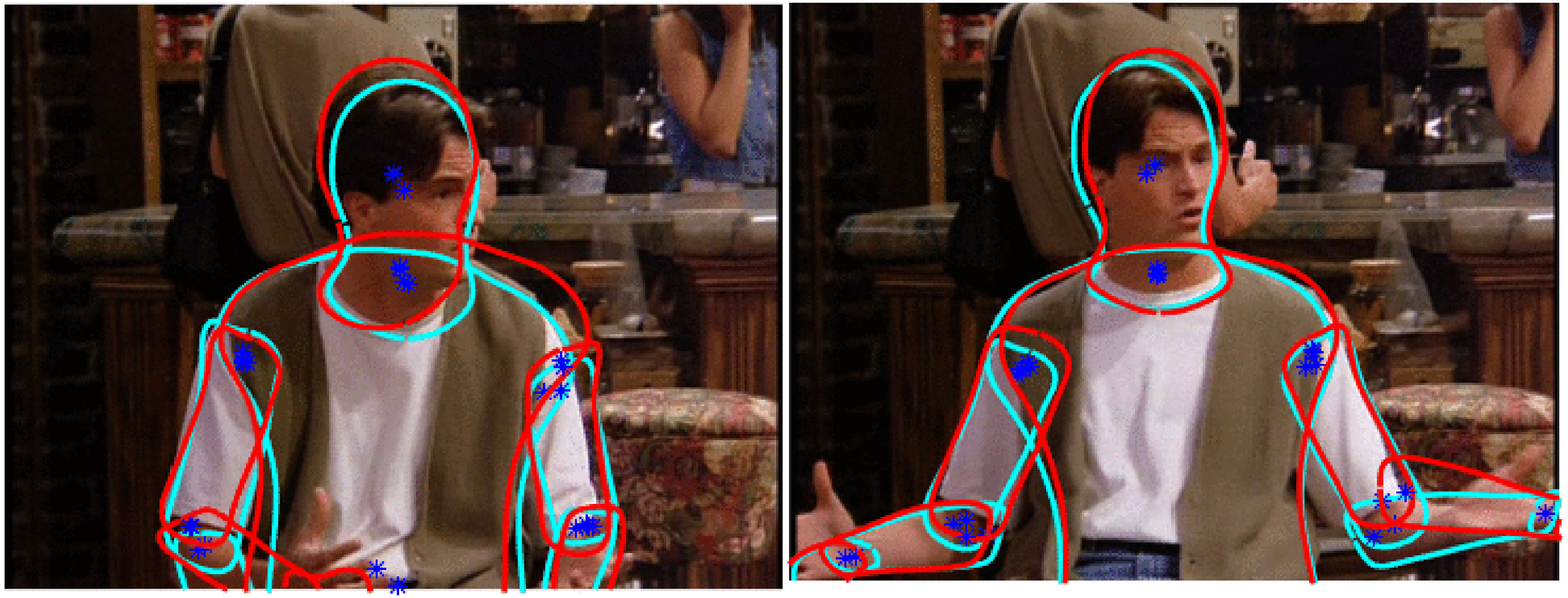
Pose and Shape Estimation



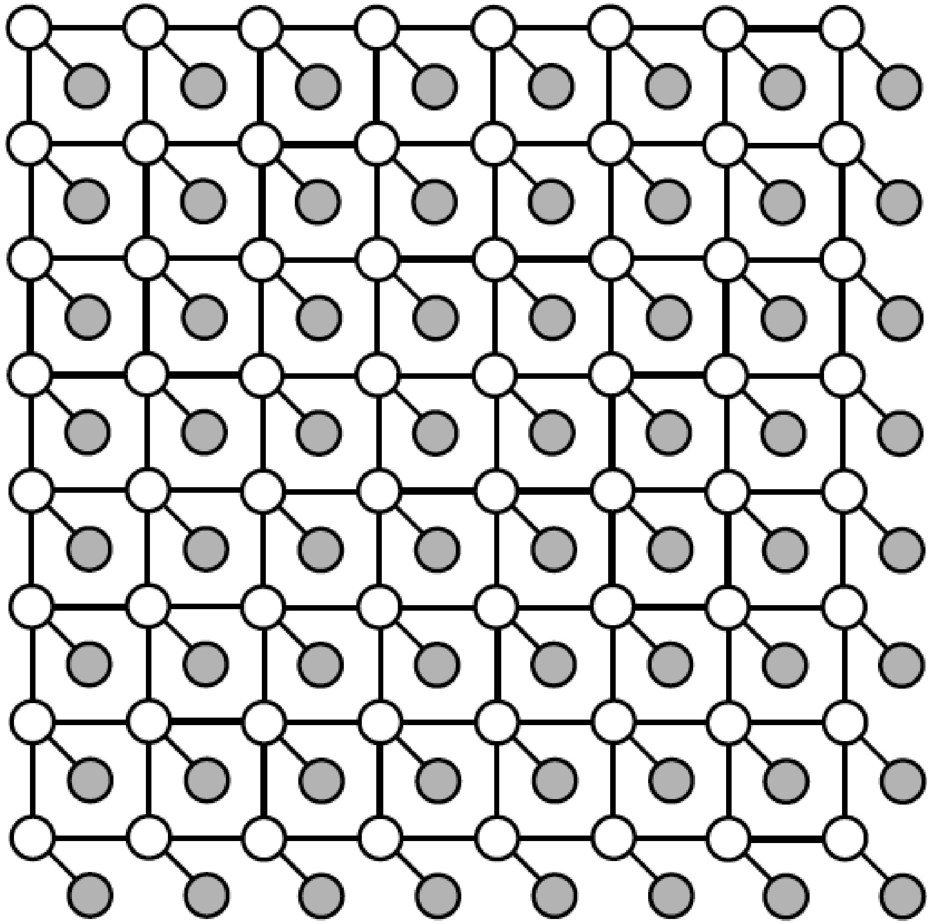
$$p(x,y) \propto \prod_{s \in \mathcal{V}} \psi_s(x_s, y) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t)$$

Complicated Likelihood

Non-Gaussian Prior

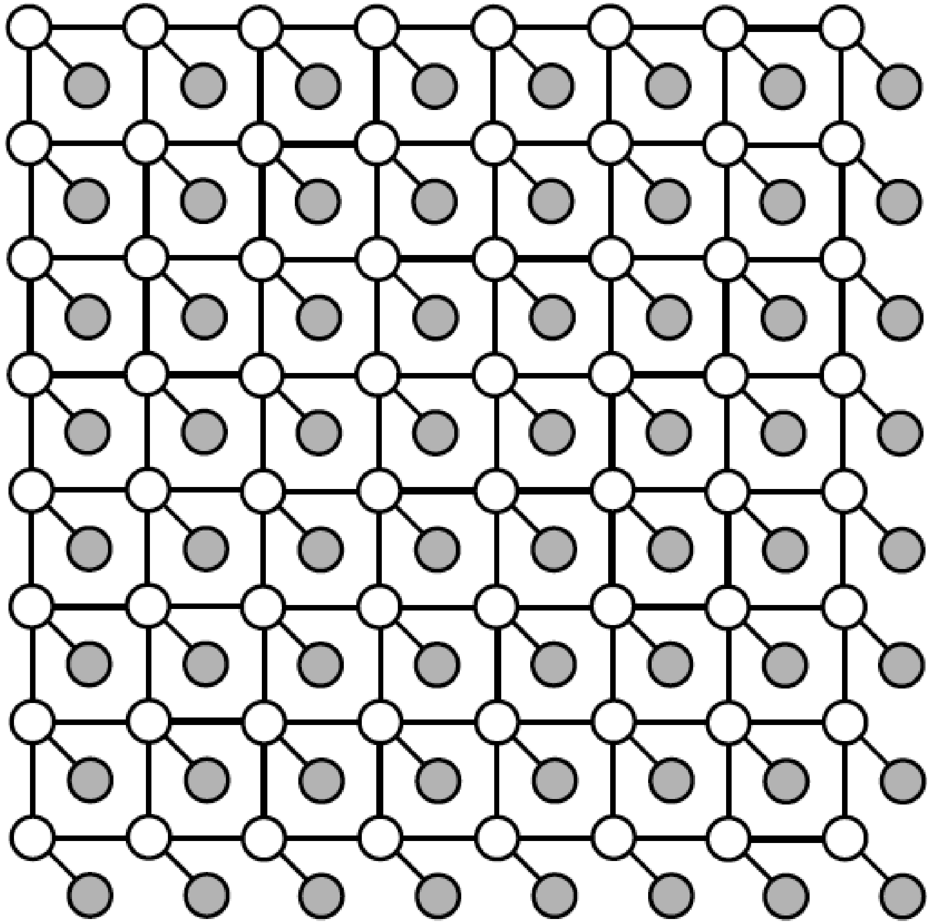


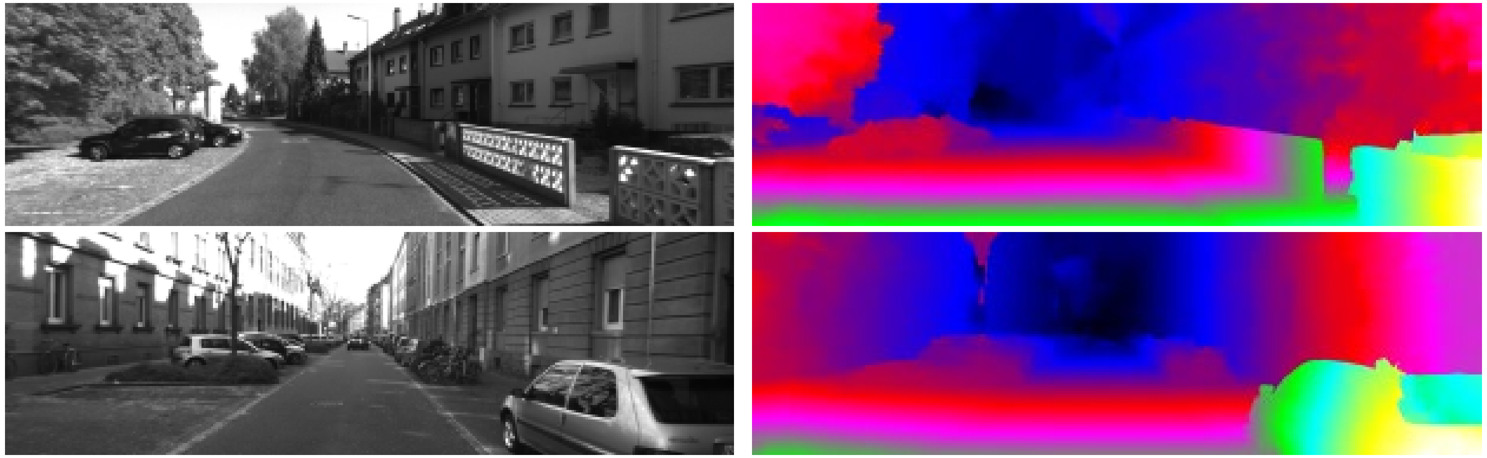
Spatial Markov Random Fields (MRFs)



- Observed nodes: Features of 2D image (intensity, color, texture, …)
- Hidden nodes: Property of 3D world (depth, motion, object category, …)

MRFs for Stereo Vision

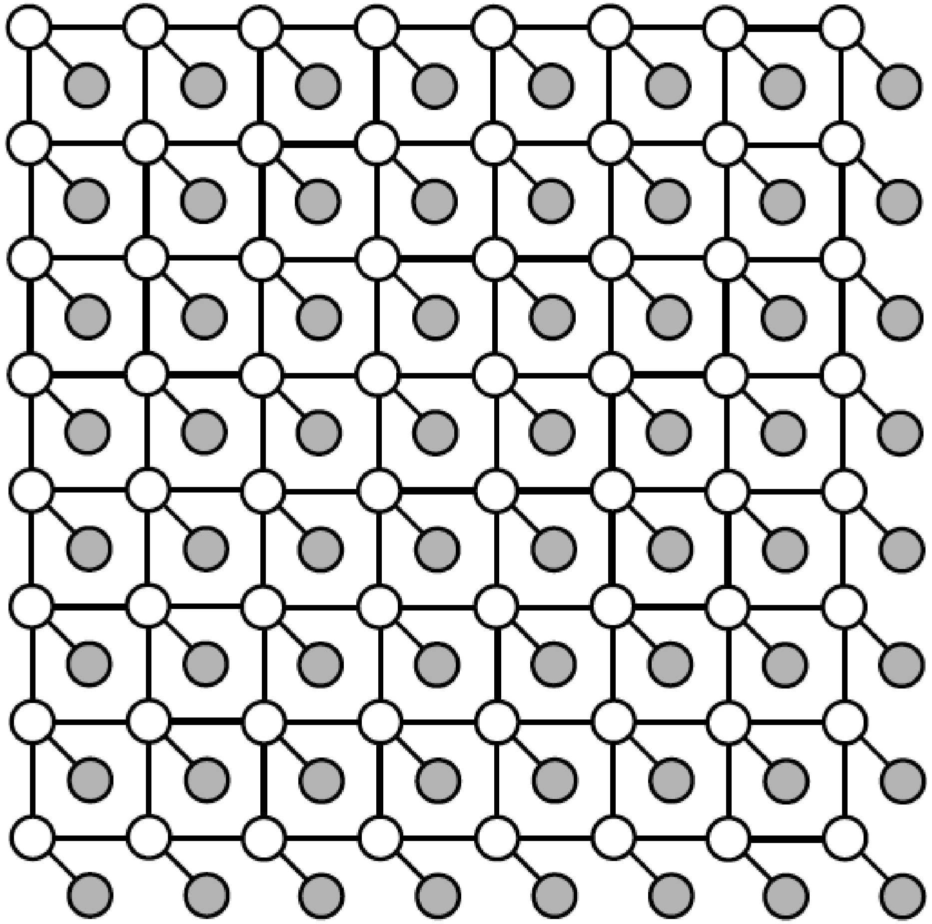


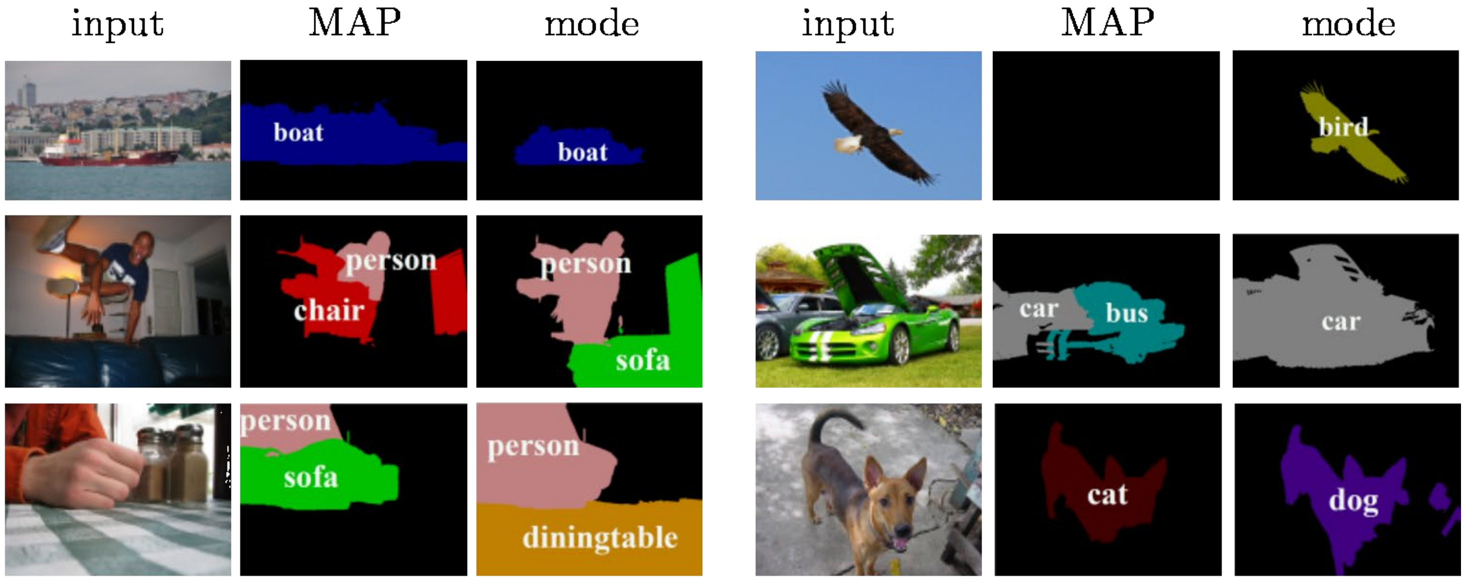


Yamaguchi et al., ECCV 2012

- Observed nodes: Features of 2D image (intensity, color, texture, …)
- Hidden nodes: Property of 3D world (depth, motion, object category, …)

MRFs for Object Segmentation

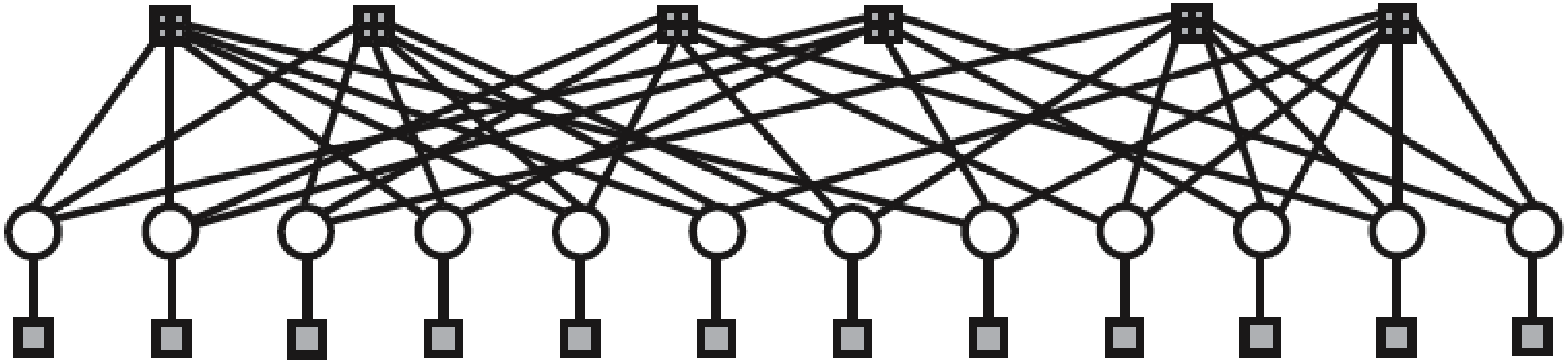


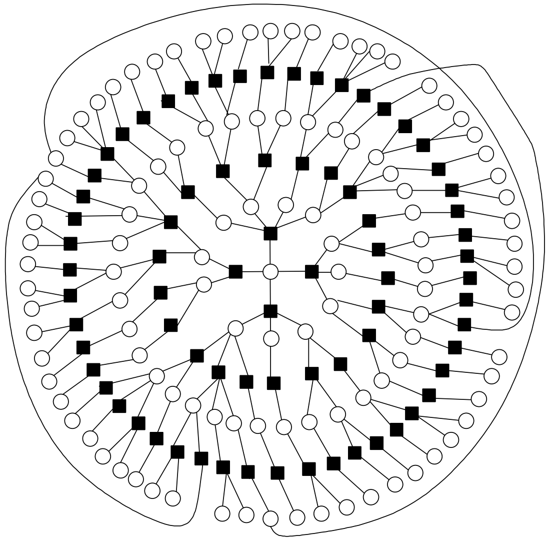


Batra et al., ECCV 2012

- Observed nodes: Features of 2D image (intensity, color, texture, …)
- Hidden nodes: Property of 3D world (depth, motion, object category, …)

Low Density Parity Check (LDPC) Codes

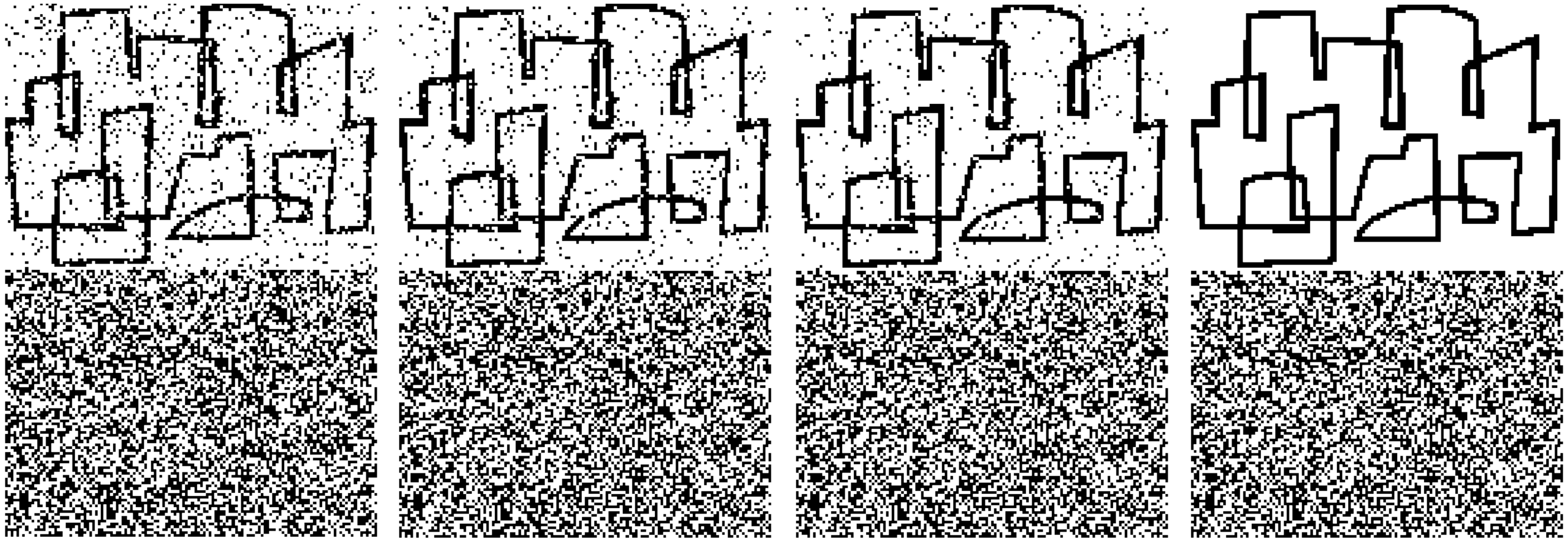




Related Graphical Model on HW2!

Data Bits

Parity Bits

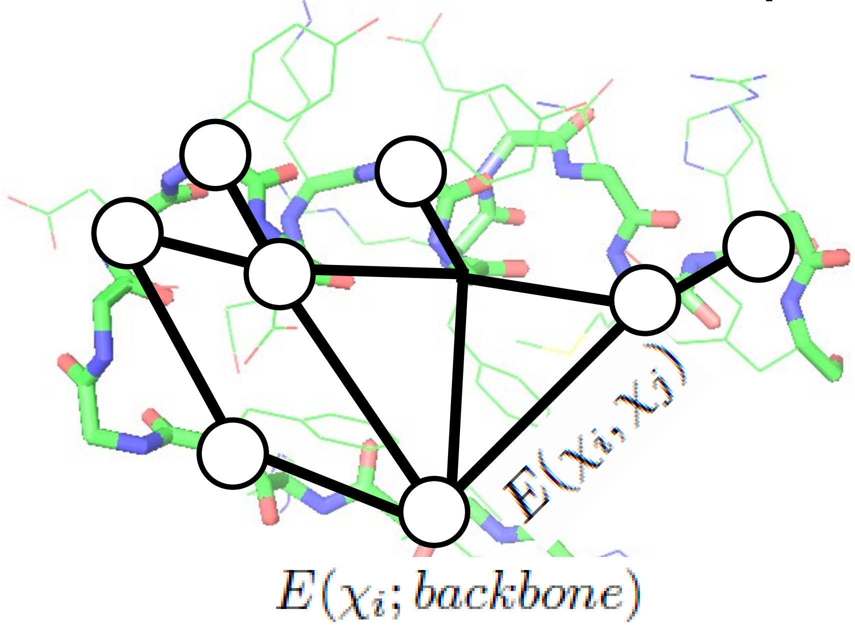


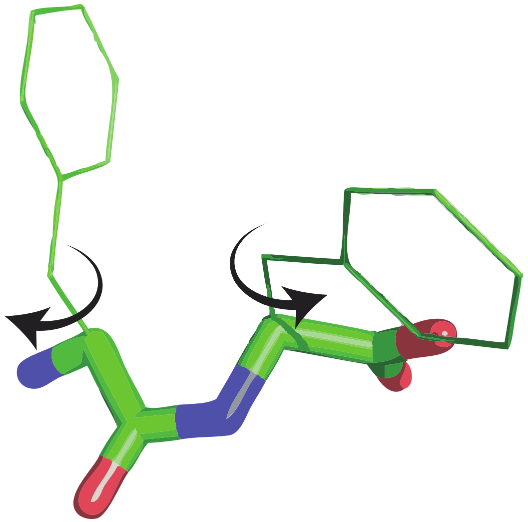
Inference: 0 5 15 \

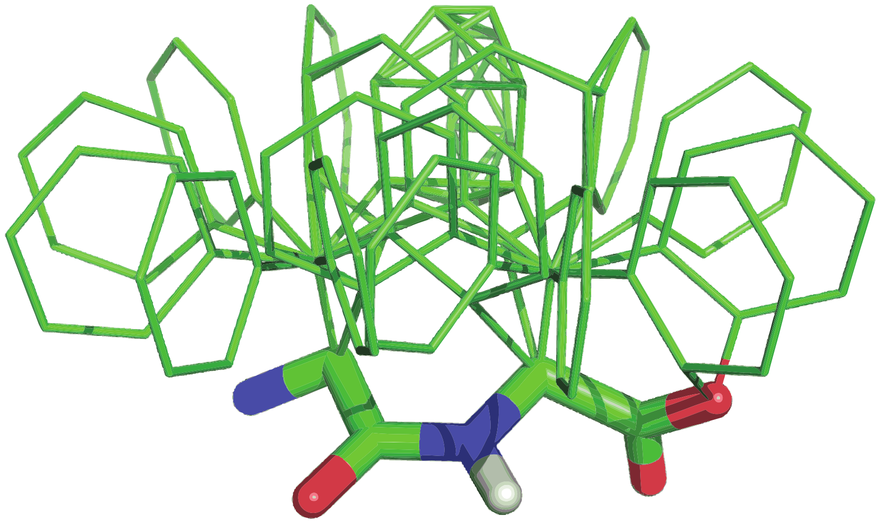
Protein Side-Chain Structure Prediction

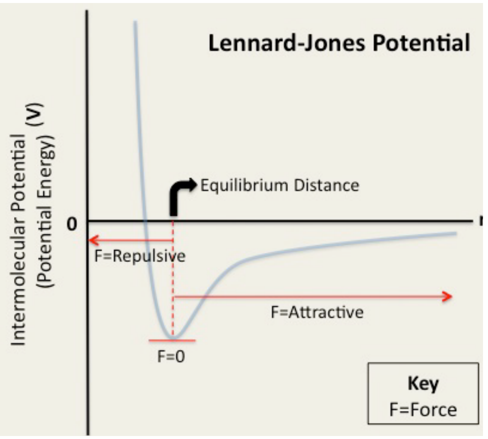
- A protein is a sequence of *amino acids*, each with a *side-chain*
- Side-chain structure prediction is MAP in pairwise MRF:

Pacheco et al., ICML 2015



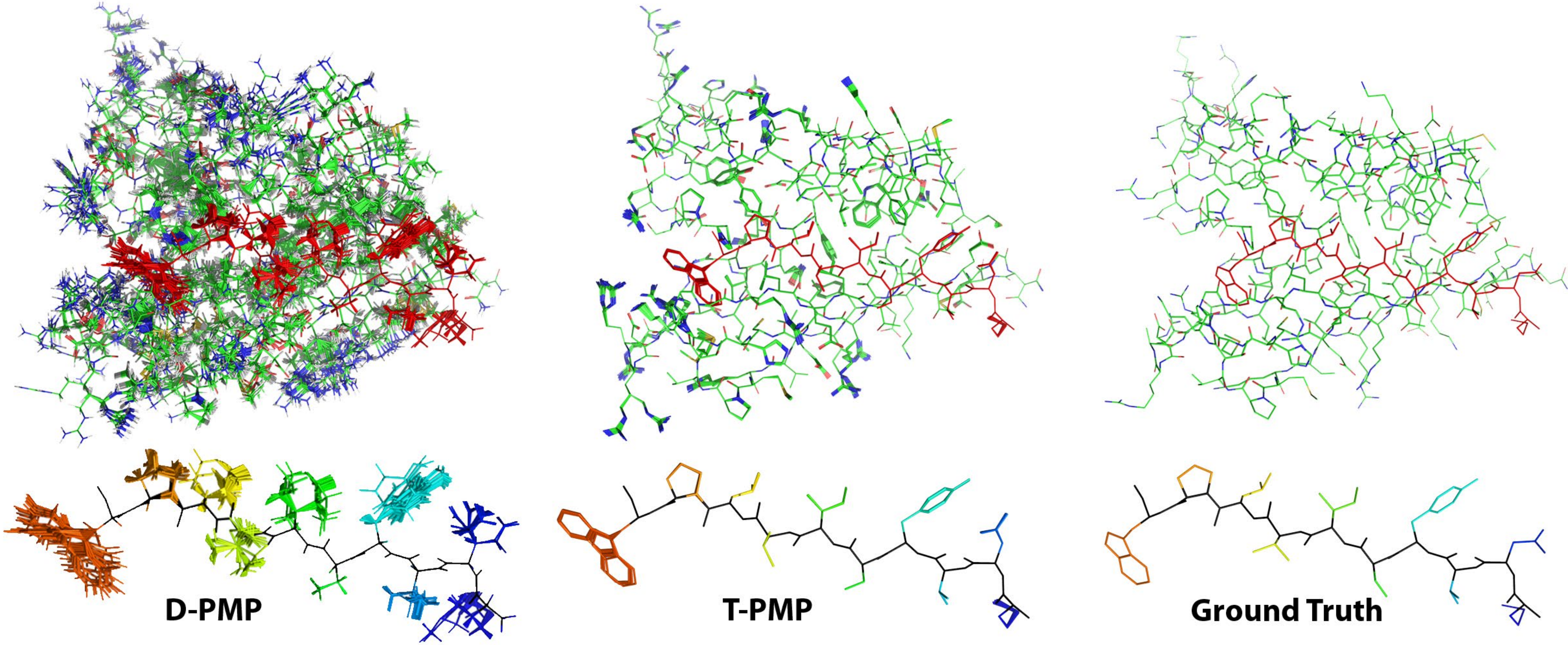






- Pairwise potentials describe repulsive (Pauli exclusion) and attractive (van der Waals force) energetic interactions
- Predicting structure lets biochemists better understand and predict function

Protein Side-Chain Structure Prediction



- Qualitative example of side-chain predictions for one protein.
- Energy evaluated via state-of-the-art Rosetta package.

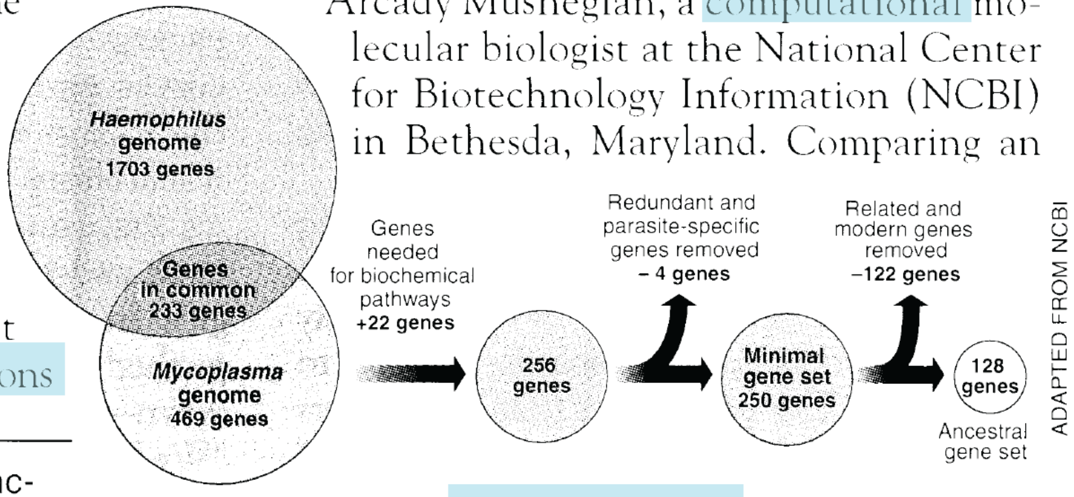
Pacheco et al., ICML 2015

Latent Dirichlet Allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

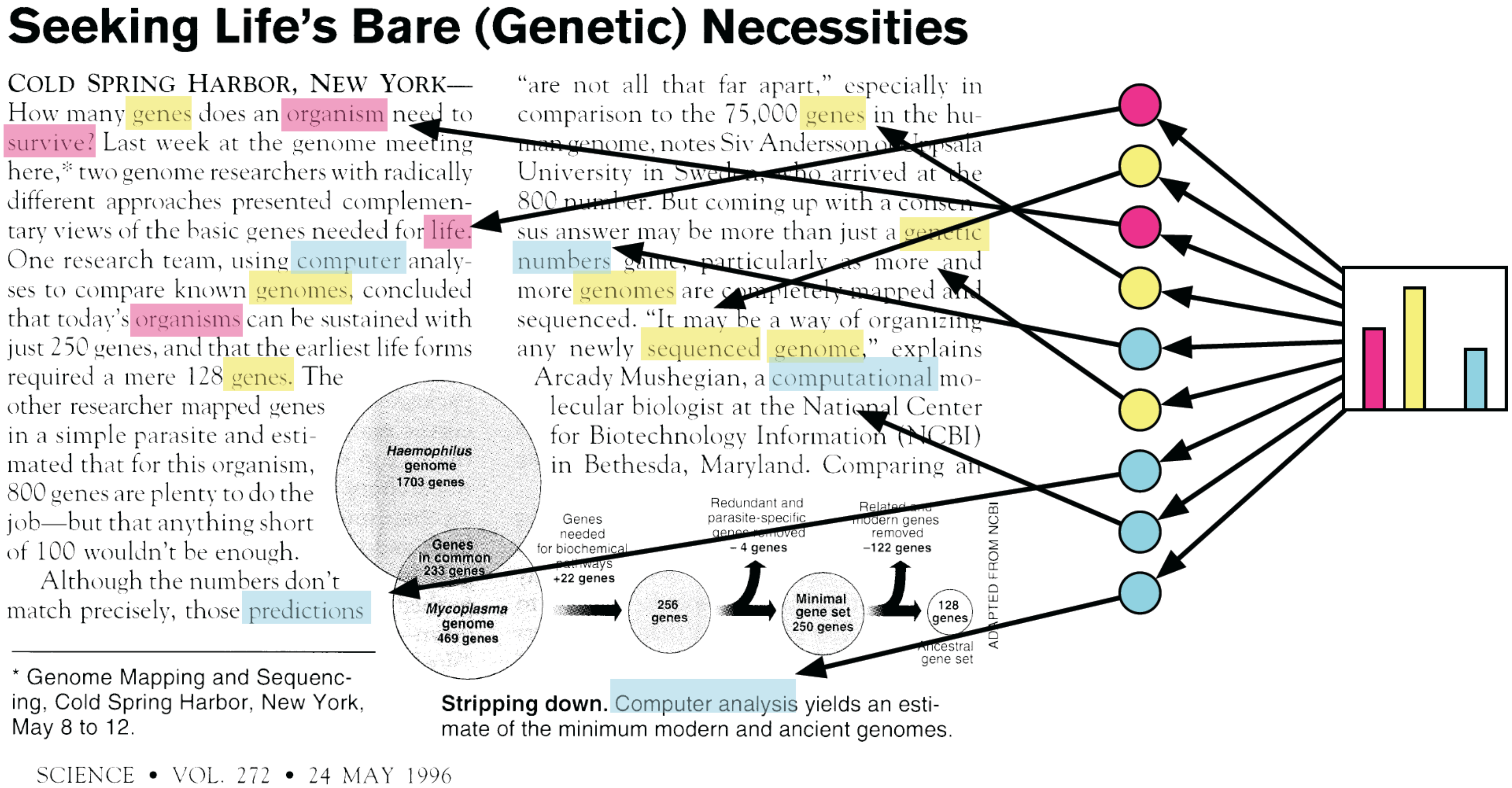
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Every text document discusses a mixture of multiple topics.

D. Blei, 2008

LDA: Generative Model


Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

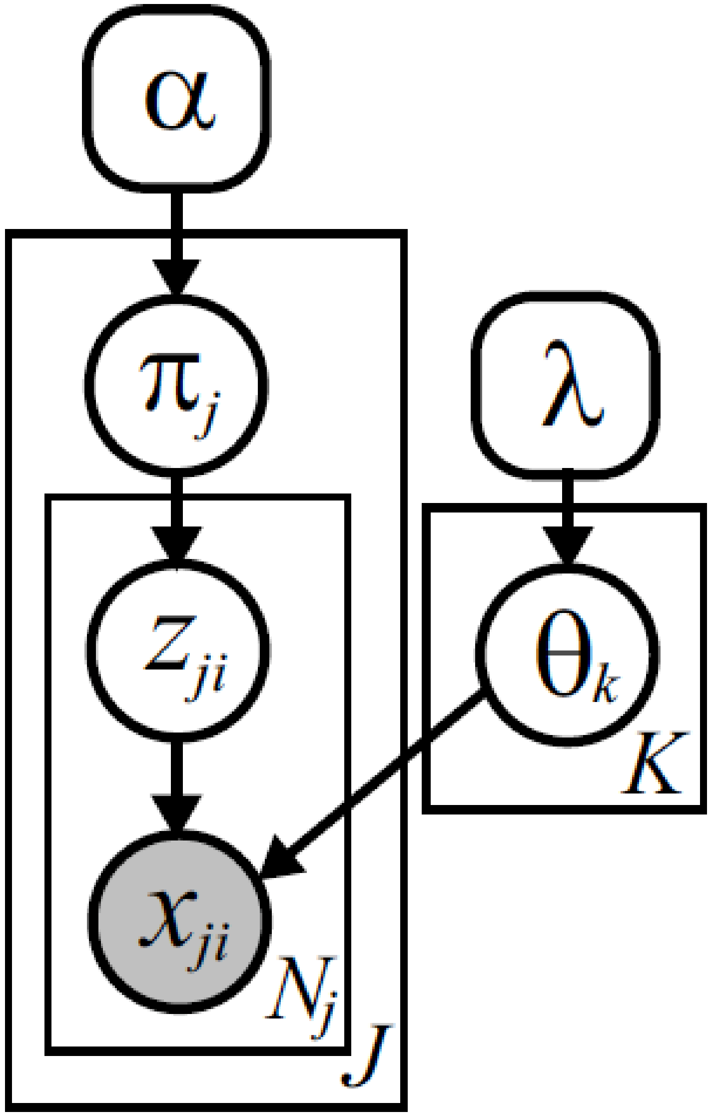
Generative Probabilistic Model:

- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

D. Blei, 2008

LDA as a Graphical Model

Given J documents, with N_j words (observations) in document j:

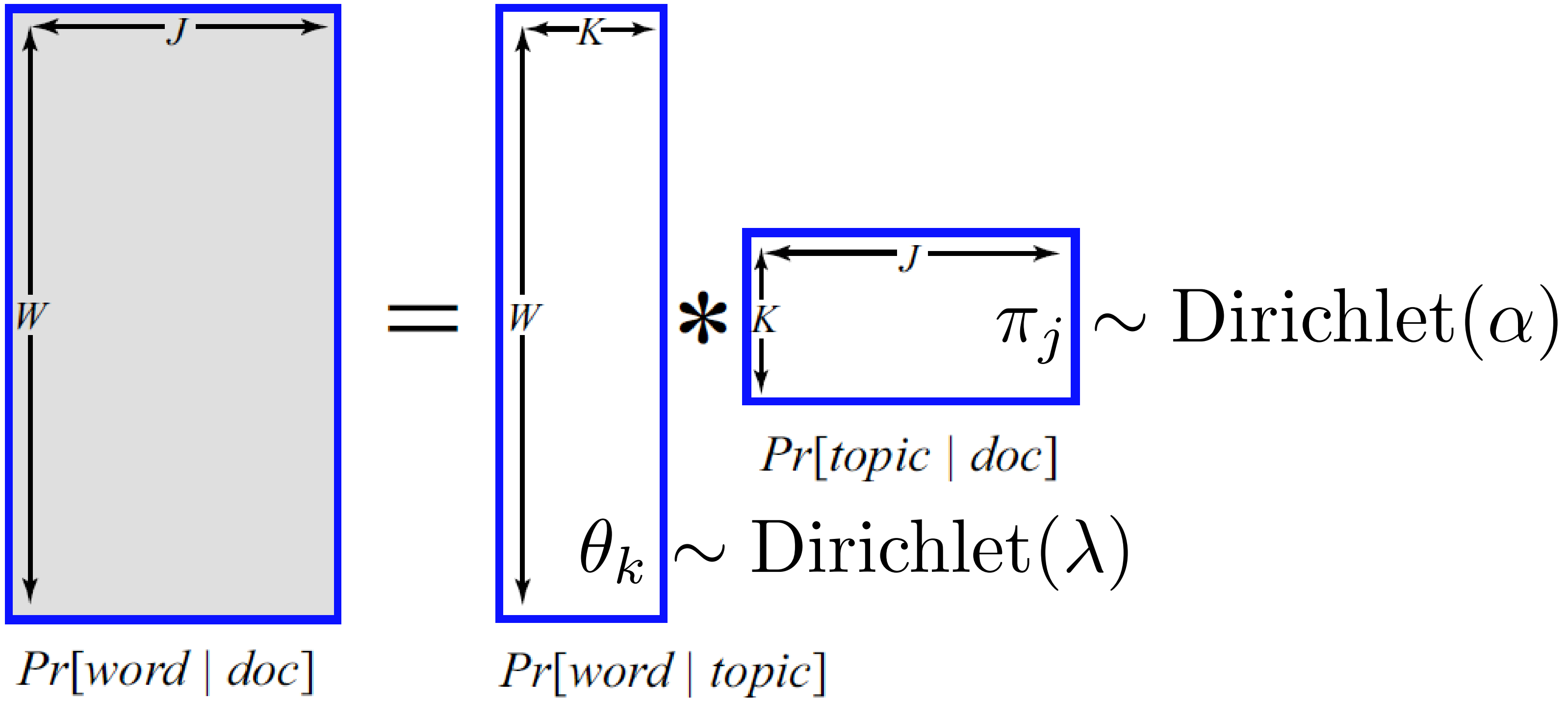


$x_{ji} \longrightarrow$ *word i in document j*

$z_{ji} \longrightarrow$ *cluster (topic) for word i in document j*

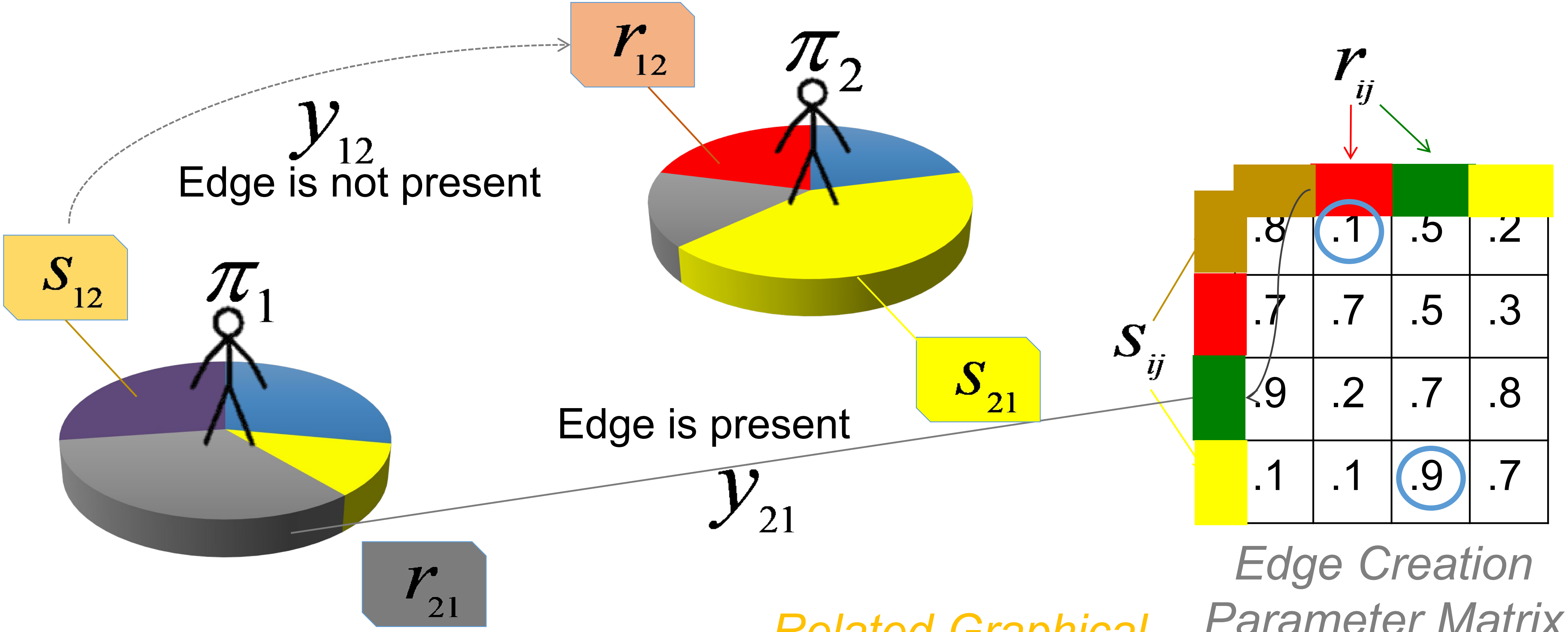
$\pi_{jk} \longrightarrow$ *expected fraction of document j about topic k*

$\theta_k \longrightarrow$ *word usage frequencies for topic k*



Community Models of Social Networks

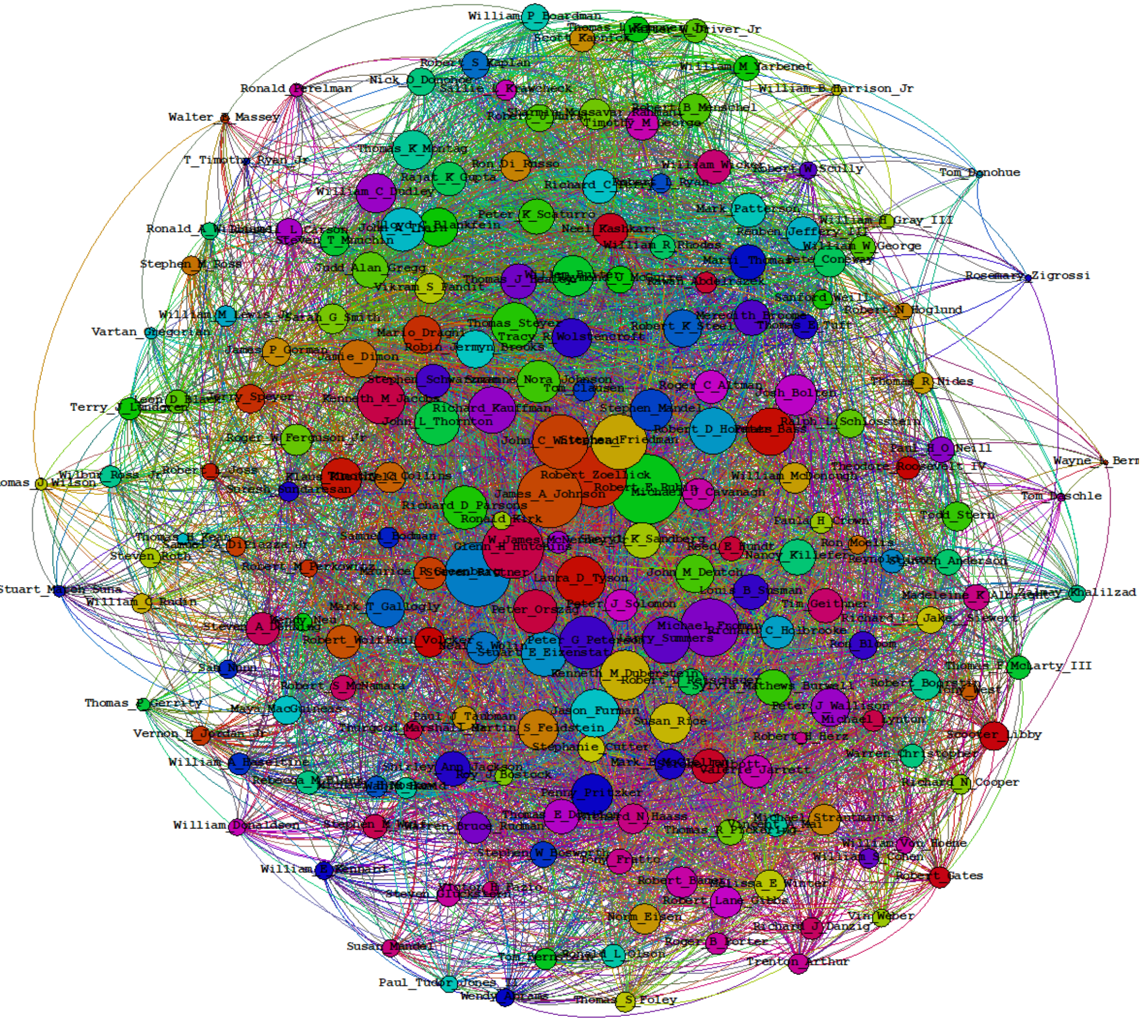
Parametric mixed membership stochastic blockmodel, Airoldi et al. JMLR 2008

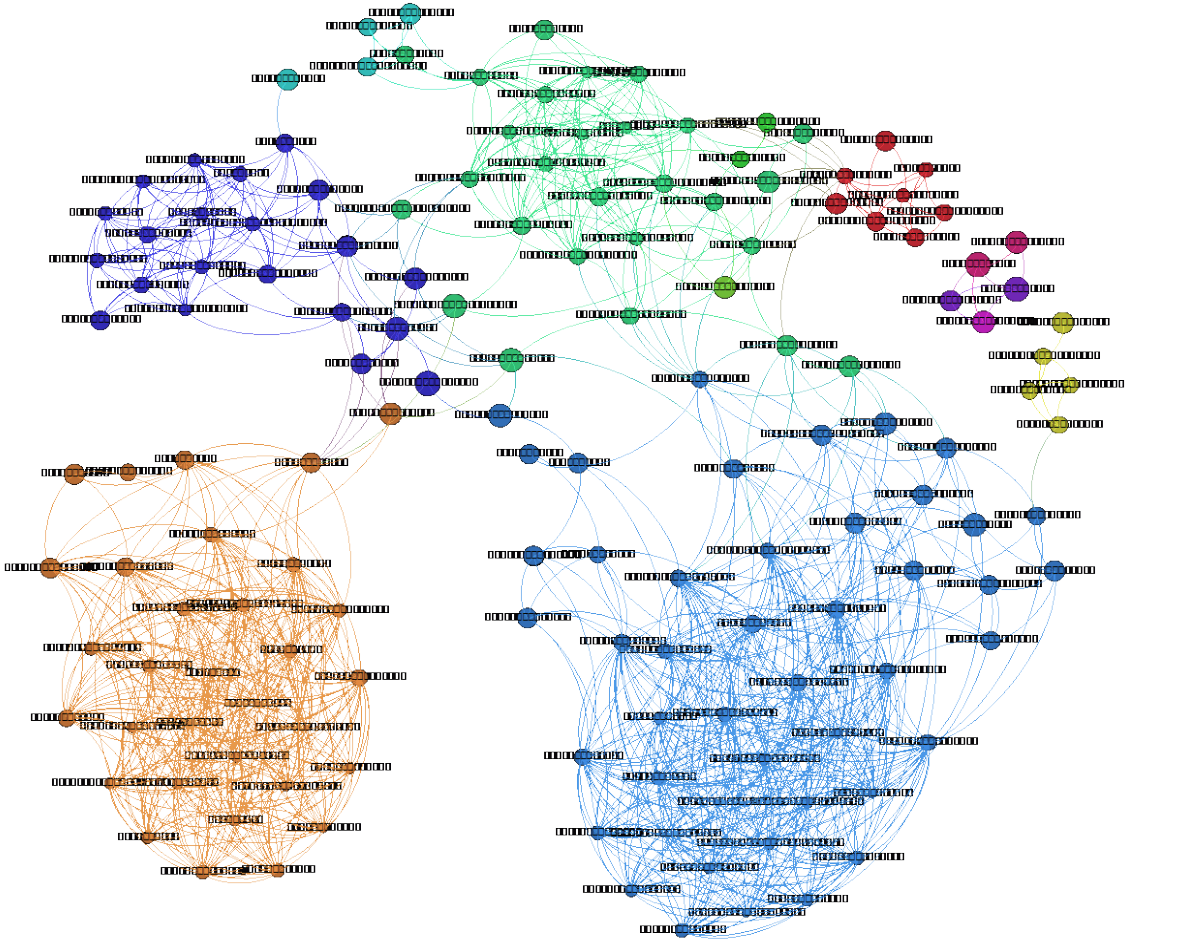


Edge Creation Parameter Matrix

Related Graphical Model on HW4!

Community Models of Social Networks





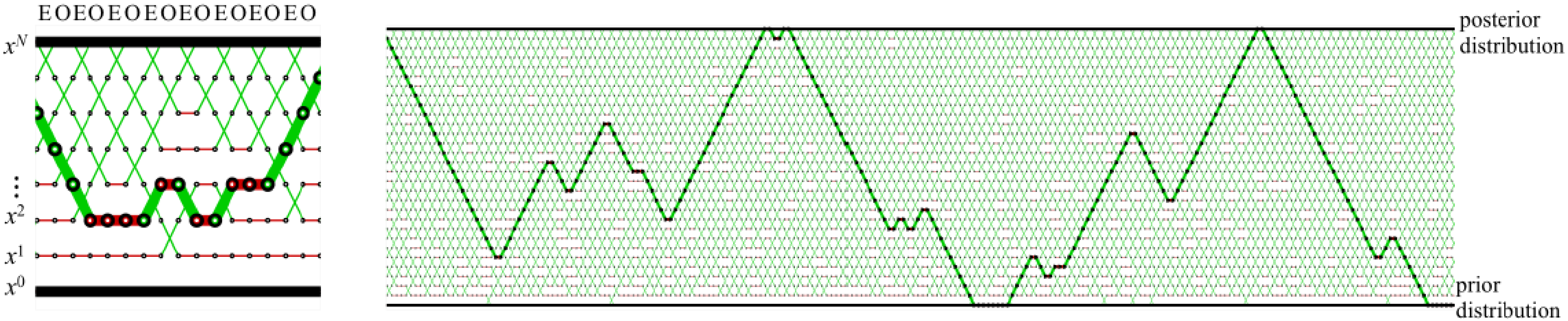
Top 200 degree nodes

Full network has N=18,831

Advanced Markov Chain Monte Carlo

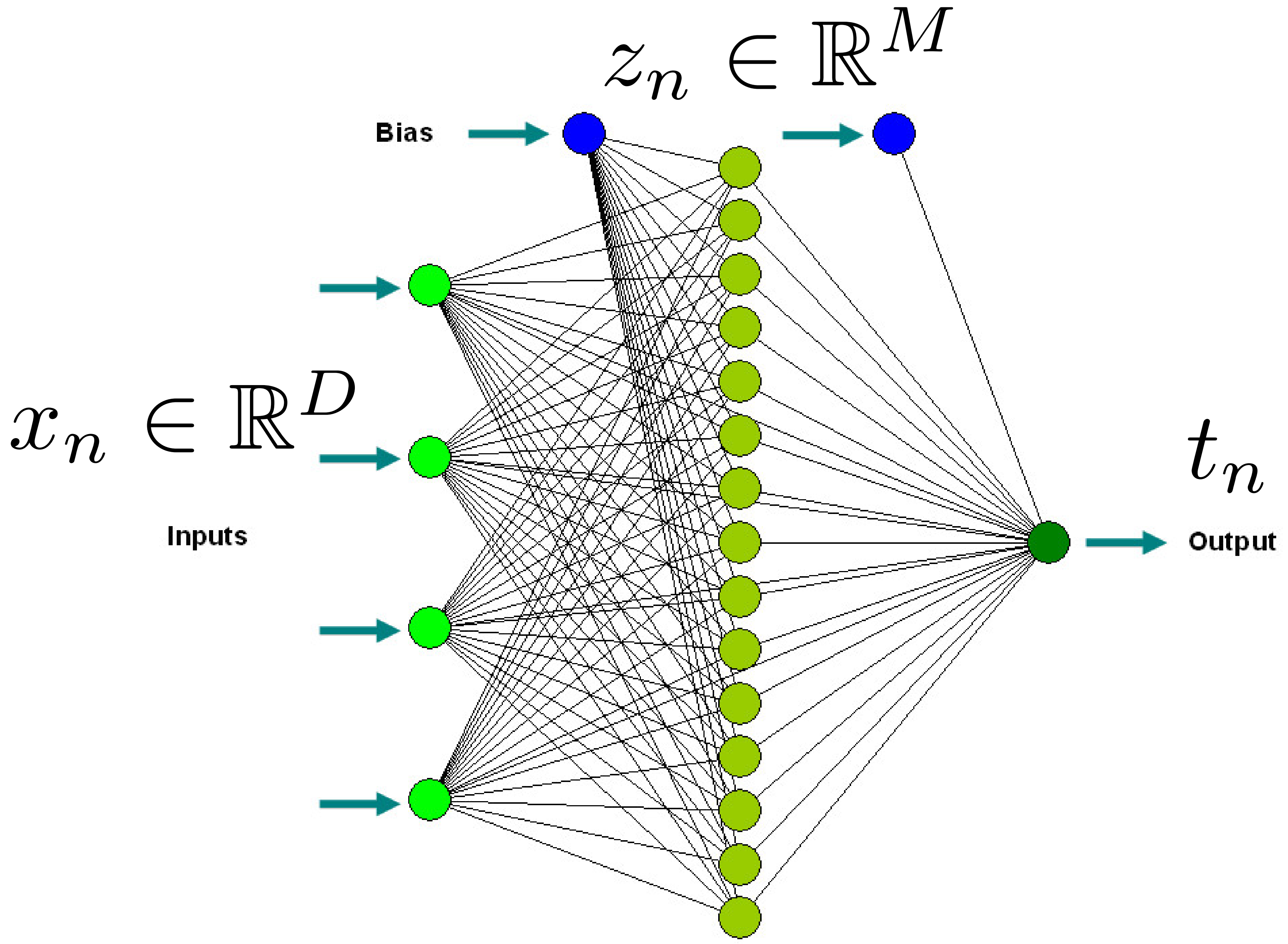
Advanced MCMC techniques reduce sample complexity and avoid getting stuck in local energy minima

[Source: Syed et al, 2019]



Example: Parallel tempering exchange replicates across multiple MCMC chains running in (embarrassingly) parallel

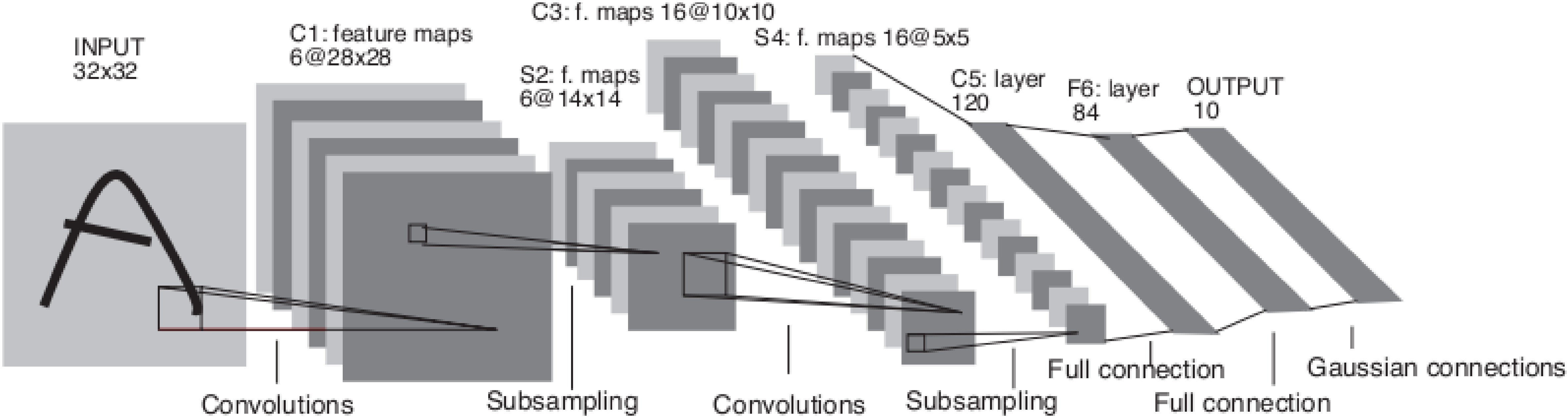
Neural Networks



- Feed-forward
- Transform input into ***hidden***, ***non-linear, tunable*** feature representation
- Use this hidden representation to produce output
- Size of hidden layer M and weights can all be optimized.

Multiple Layers and Deep Networks

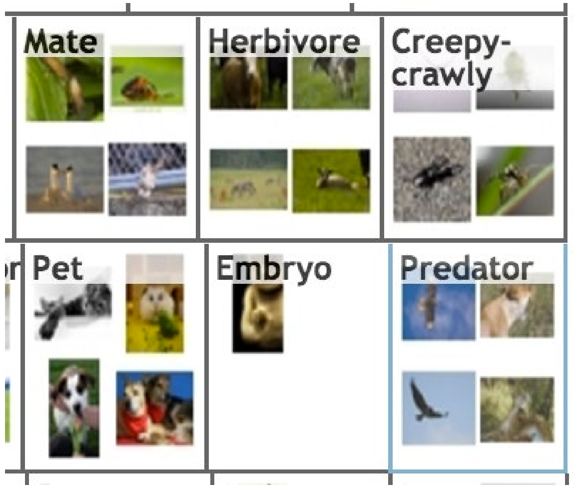
LeNet5: *Convolutional Neural Net for Digit Classification (LeCun et al., 1998)*



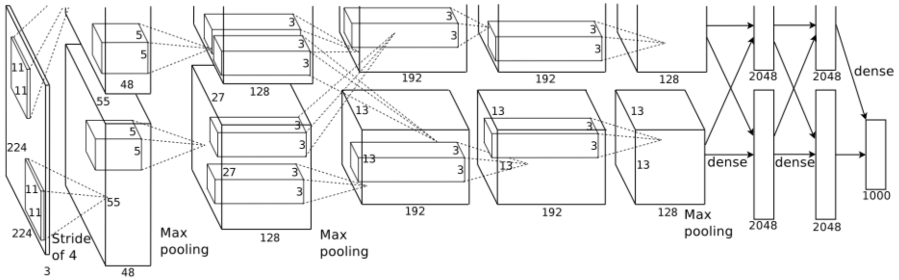
Deep Learning for Object Recognition

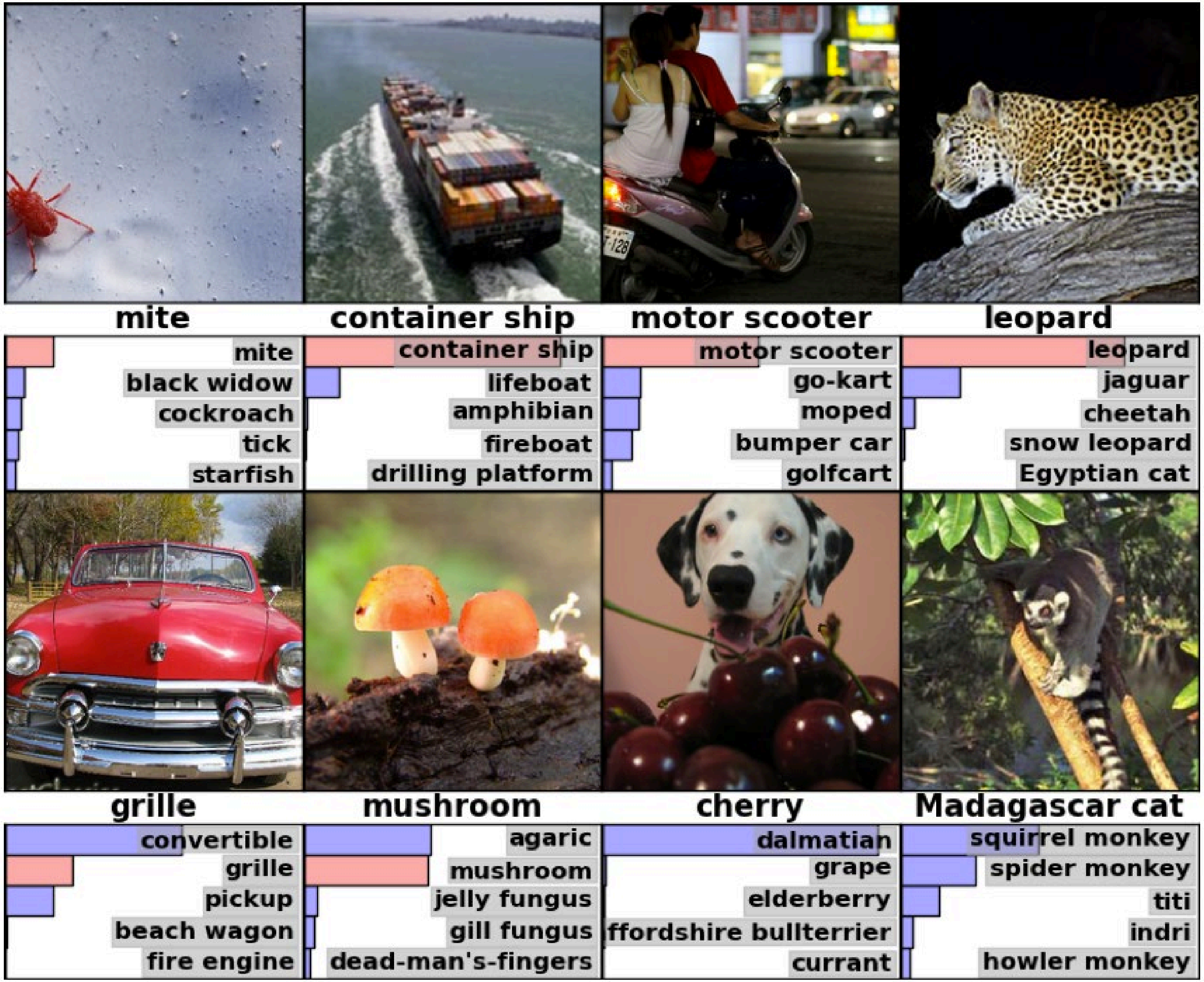
ImageNet dataset: 15 million images

22,000 categorie



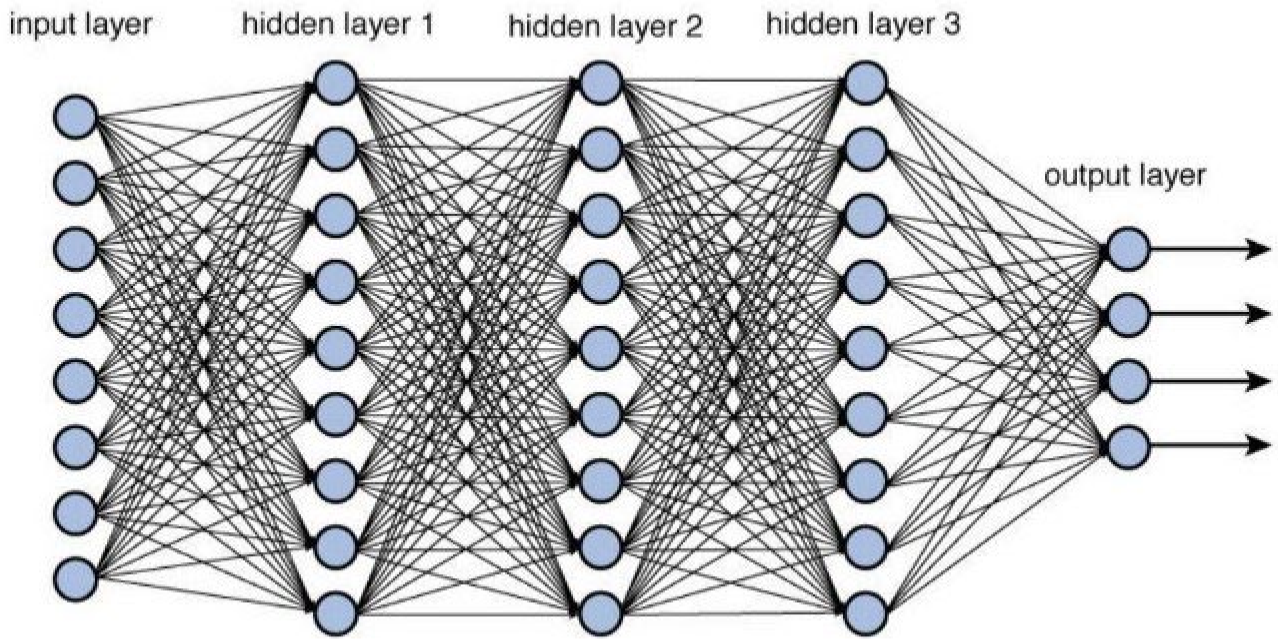
AlexNet *(Alex Krishevsky et al, NIPS 2012)*
Deep convolutional neural network,
trained via backprop on multiple GPUs.





Bayesian Deep Learning

Neural networks are graphical models too…

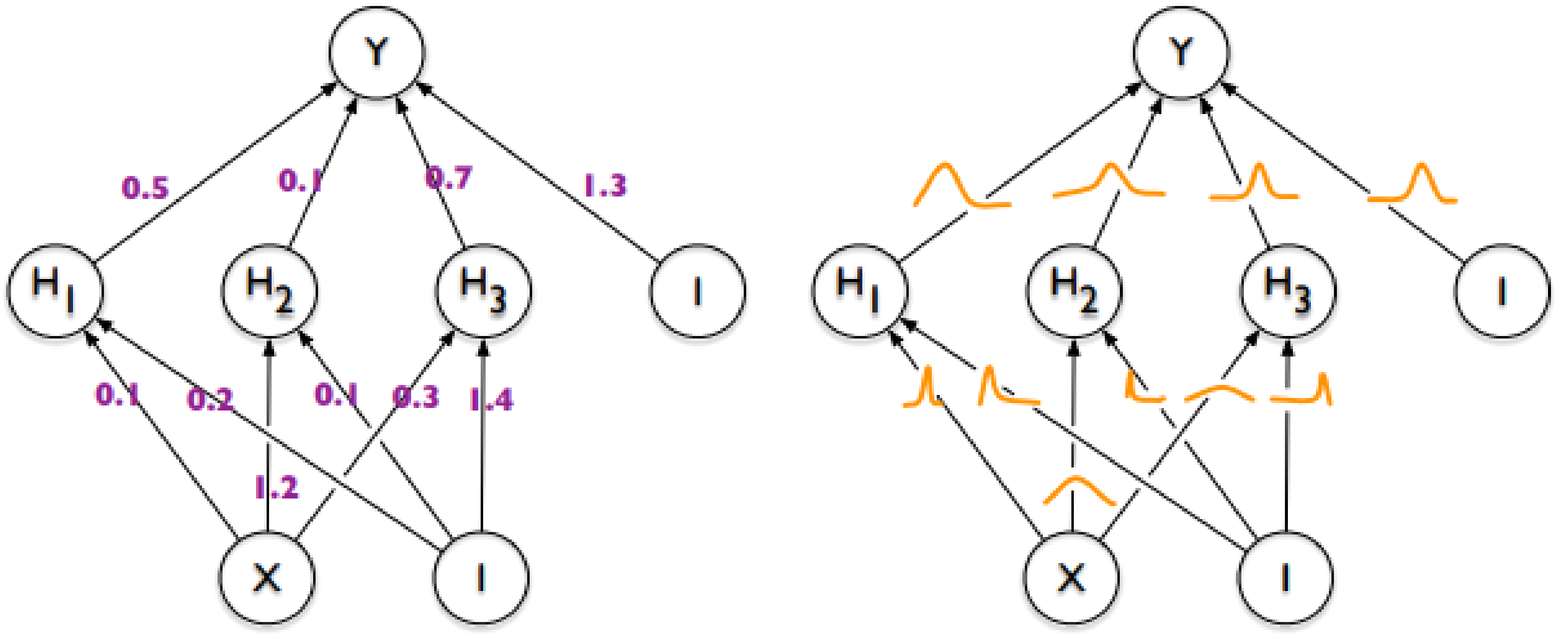


…but they are *typically* not <u>probabilistic</u>

Idea Combine representation flexibility of DNNs with uncertainty modeling of PGMs

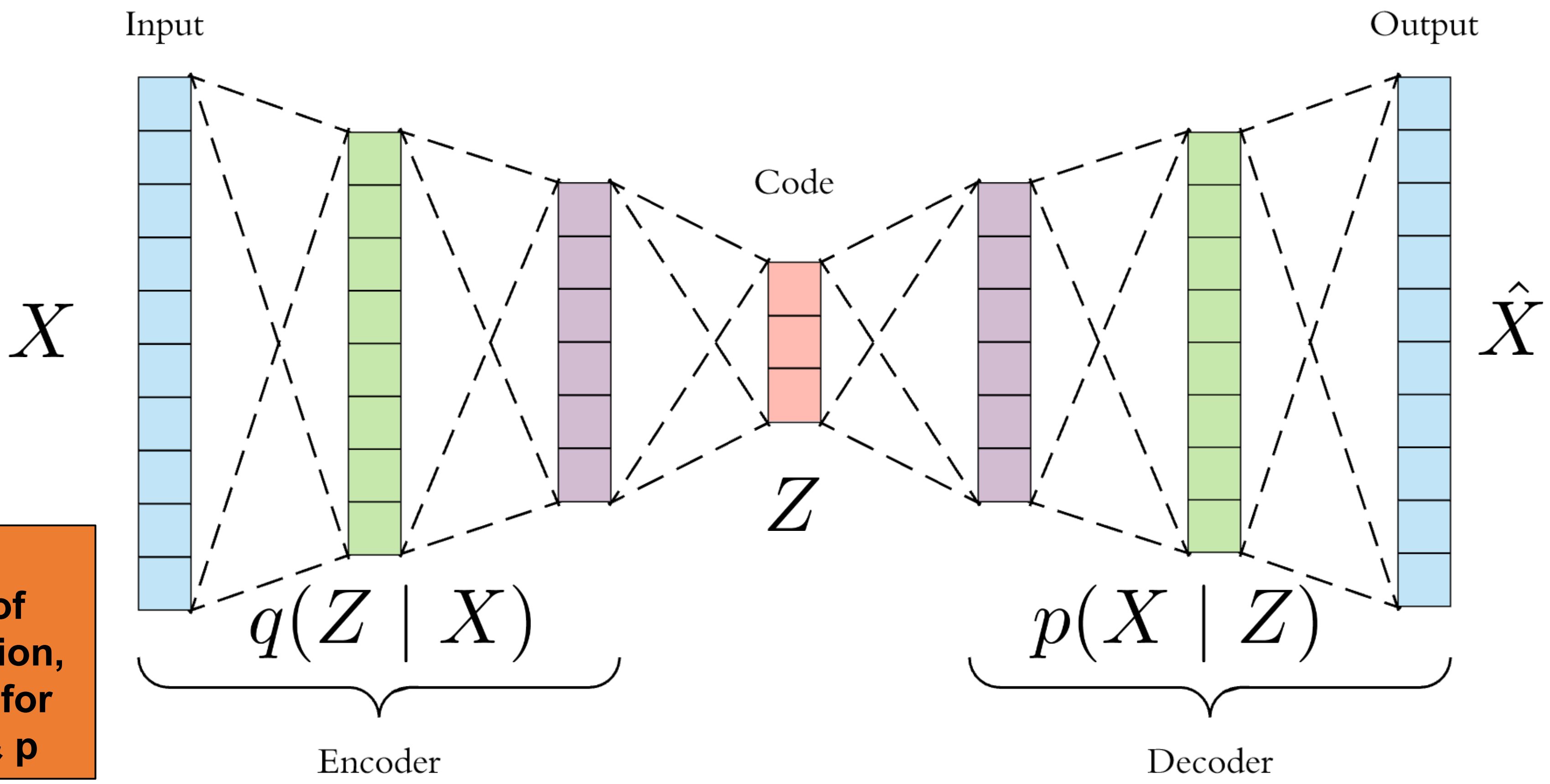
Bayesian Neural Network

Standard DNNs learn *point estimate* of weights from training…



- Predictions can be brittle / sensitive to adversarial attack
- Robustness requires training data include all possible realities
- Bayesian approach treats weights as random quantities to be inferred
- Assigns posterior probabilities to all network parameters / predictions

Variational Autoencoder



NN learns parameters of some distribution, e.g. mean/var for Gaussian q & p

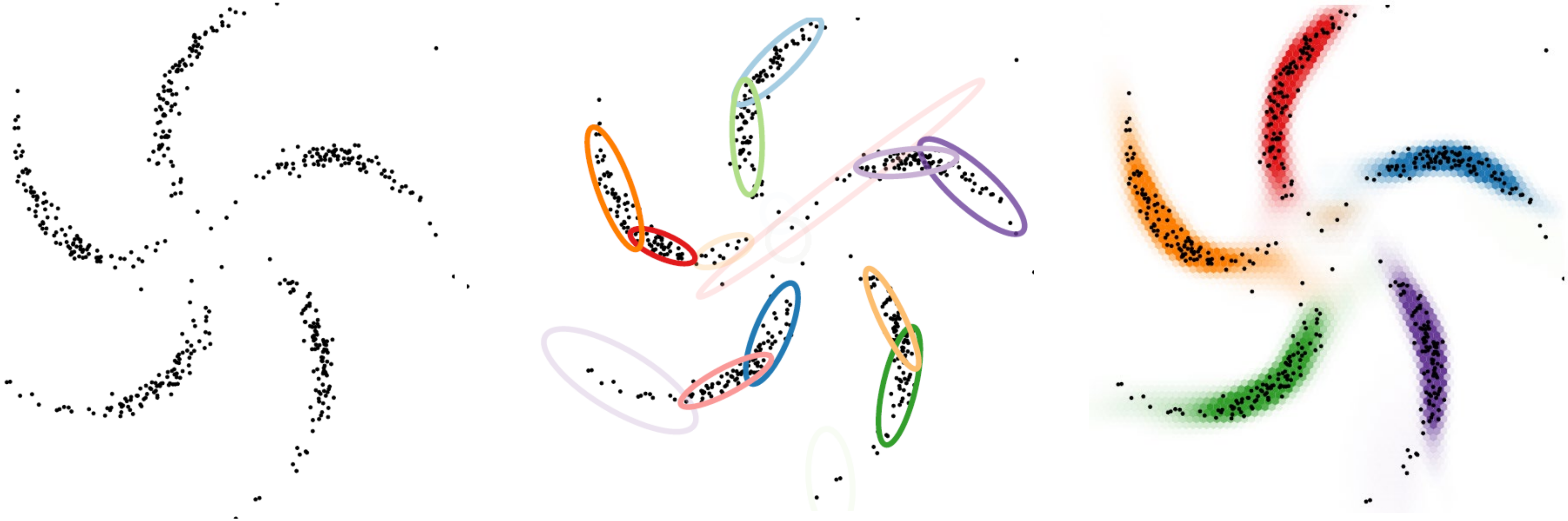
NOTE: This is completely *unsupervised*

Train by minimizing reconstruction loss and fit to marginal:

$$\min \mathcal{L}(x, \hat{x}) + KL(q(z \mid x) \| p(z))$$

Structured VAE

Combines VAEs with structured models (mixtures, dynamical systems, …)

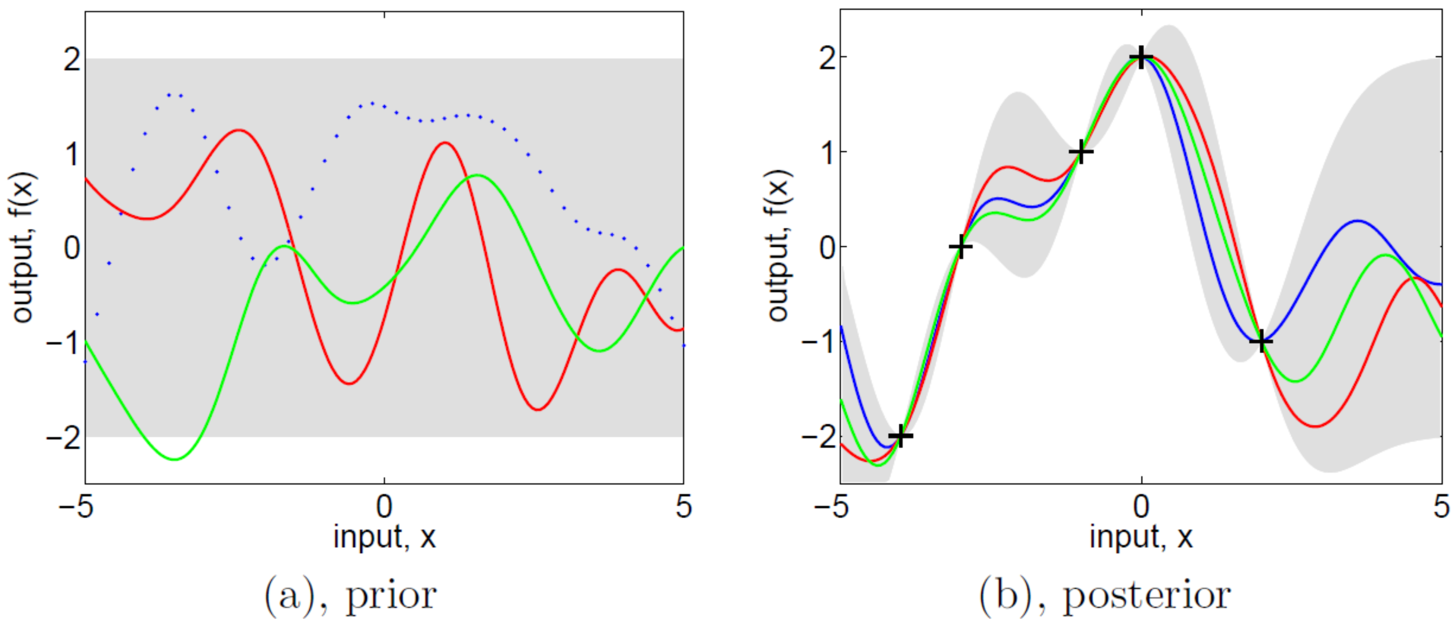


Data **Gaussian Mixture Model (GMM)** **GMM Structured Variational Autoencoder**

[Source: Johnson et al., NIPS 2016]

Gaussian Processes (GPs)

Distribution over random continuous functions…



$$\mathbf{f}_* \sim \mathcal{N}(\mathbf{0}, \underbrace{K(X_*, X_*)})$$

Kernel function encodes correlation between evaluation points in the domain

GPs are generative models…

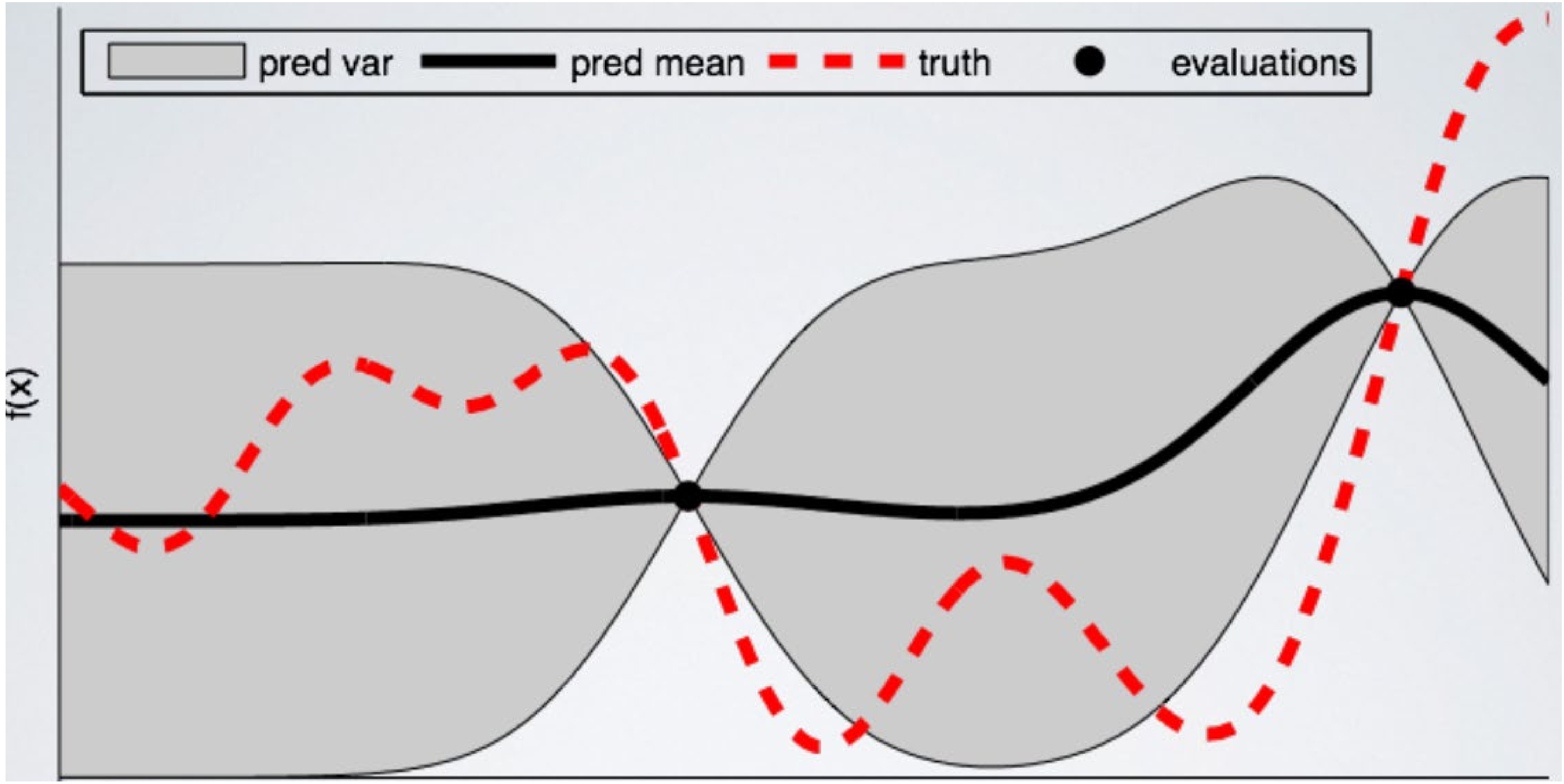
- Can sample function from prior
- Tractable posterior
- Posterior predictive

…equivalent to Bayesian linear regression in function space

[Source: C. Rassmussen]

Bayesian Optimization

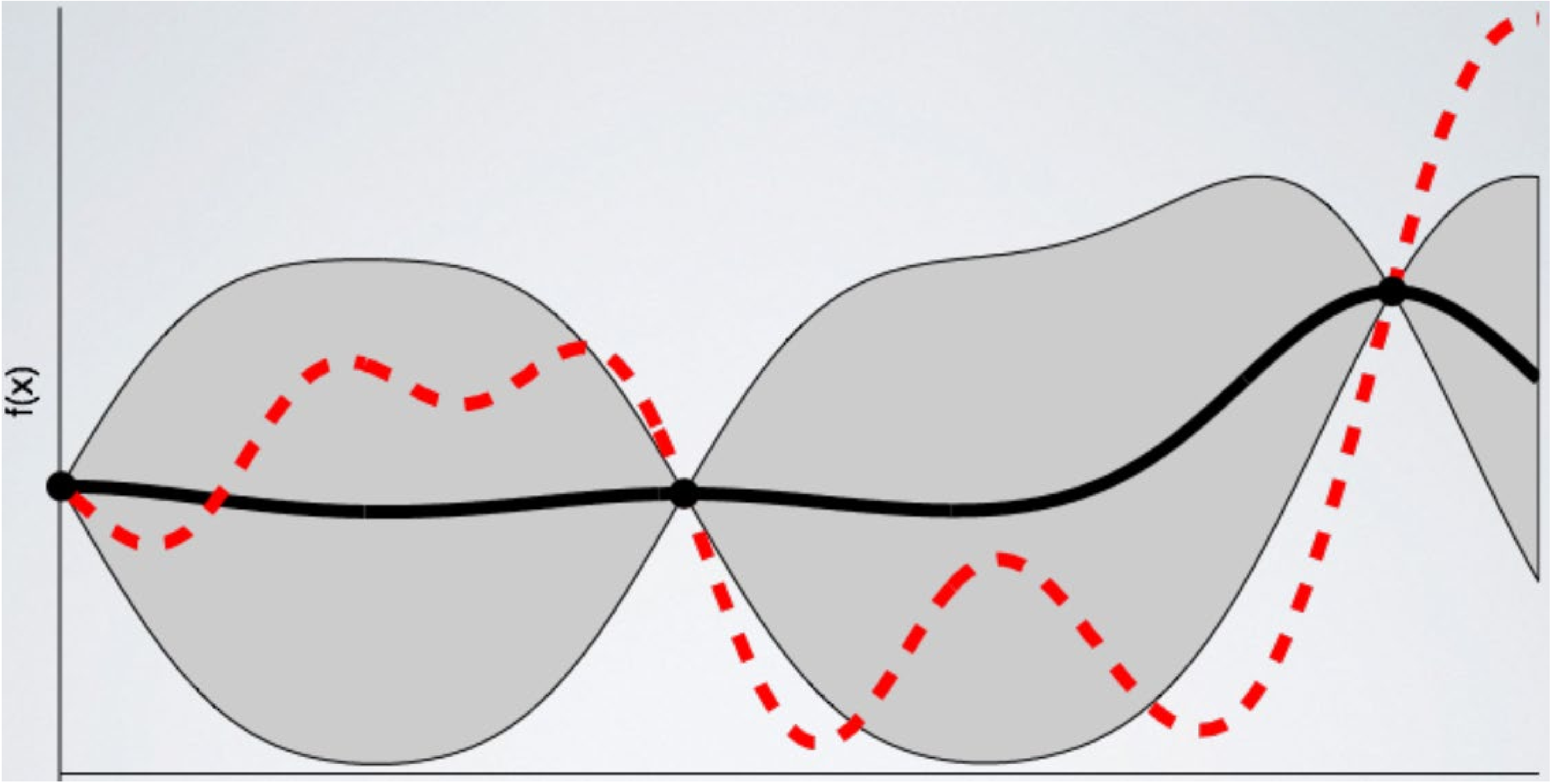
Global optimization of <u>random functions</u>: $\min_x f(x)$



[Source: Ryan Adams]

Bayesian Optimization

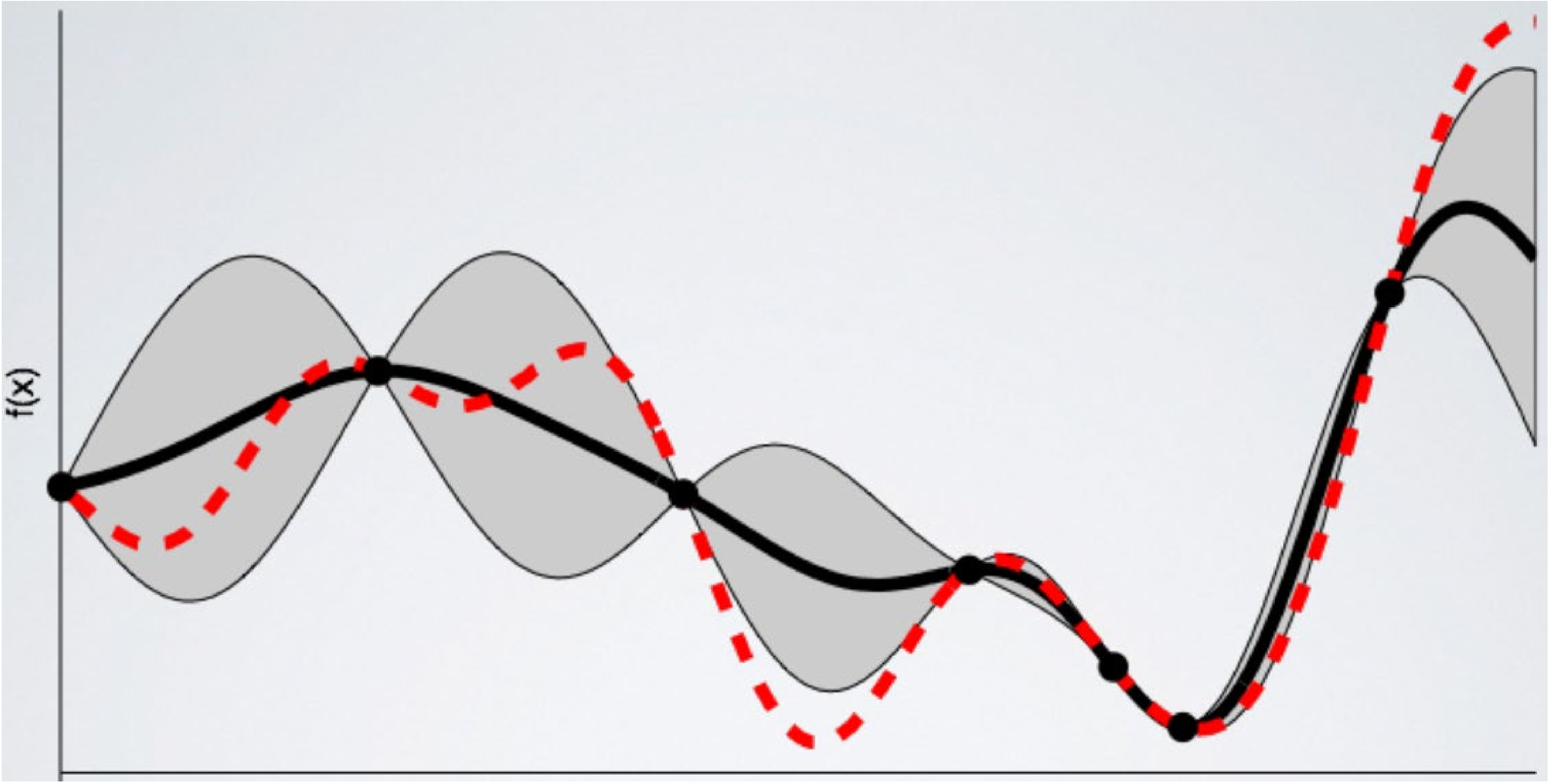
Iteratively updates distribution over function value (regression)



[Source: Ryan Adams]

Bayesian Optimization

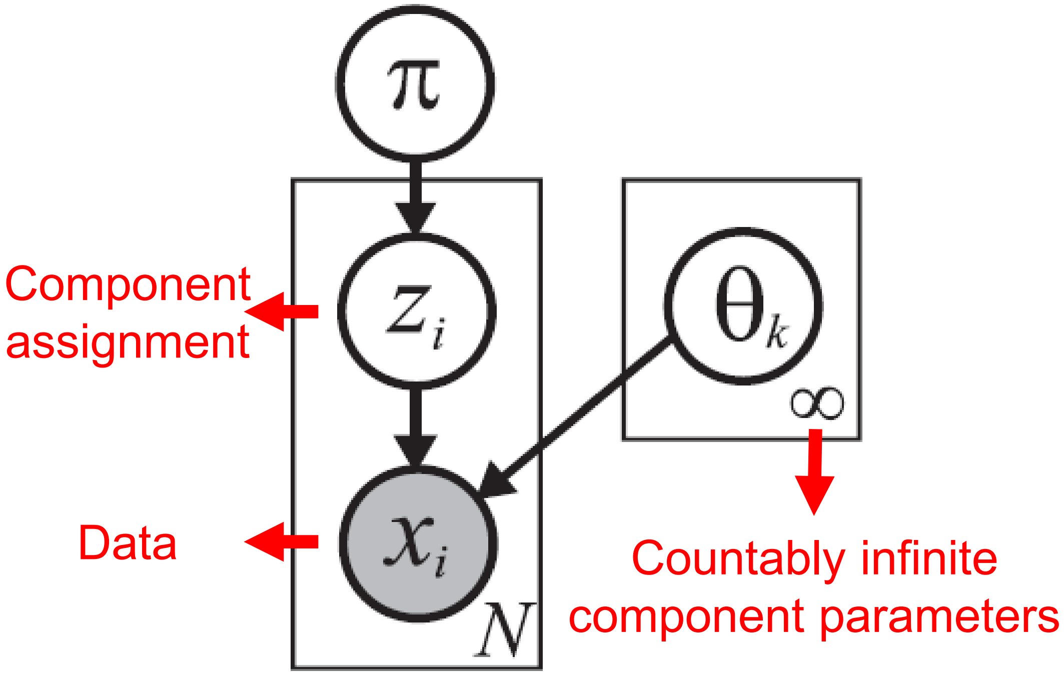
The function is well-approximated around the minimizer

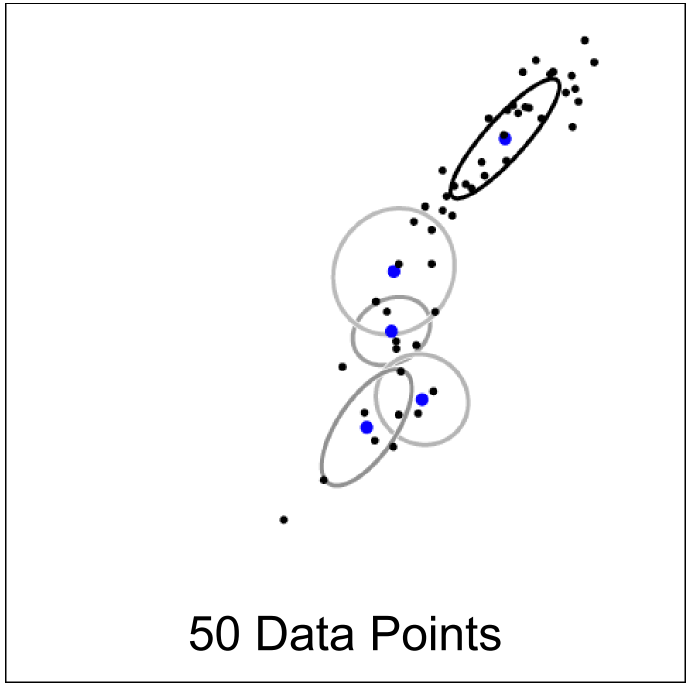


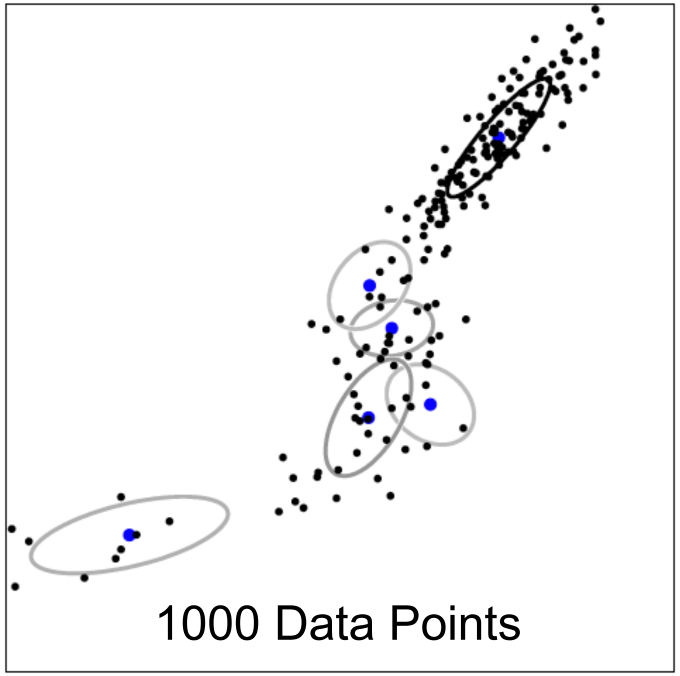
[Source: Ryan Adams]

Bayesian Nonparametrics

Amount and nature of data drive model complexity

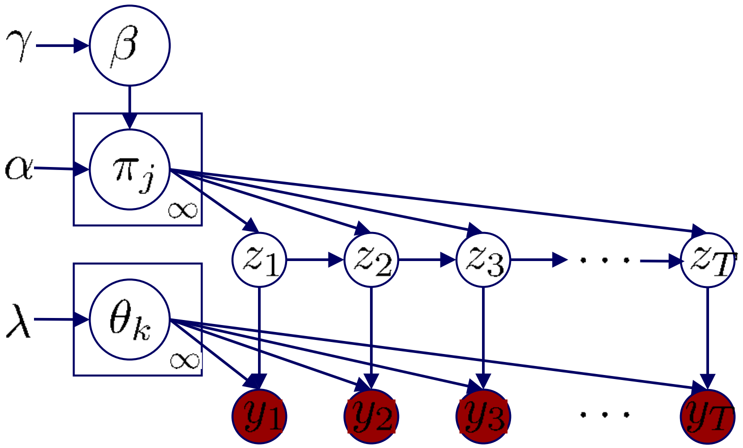


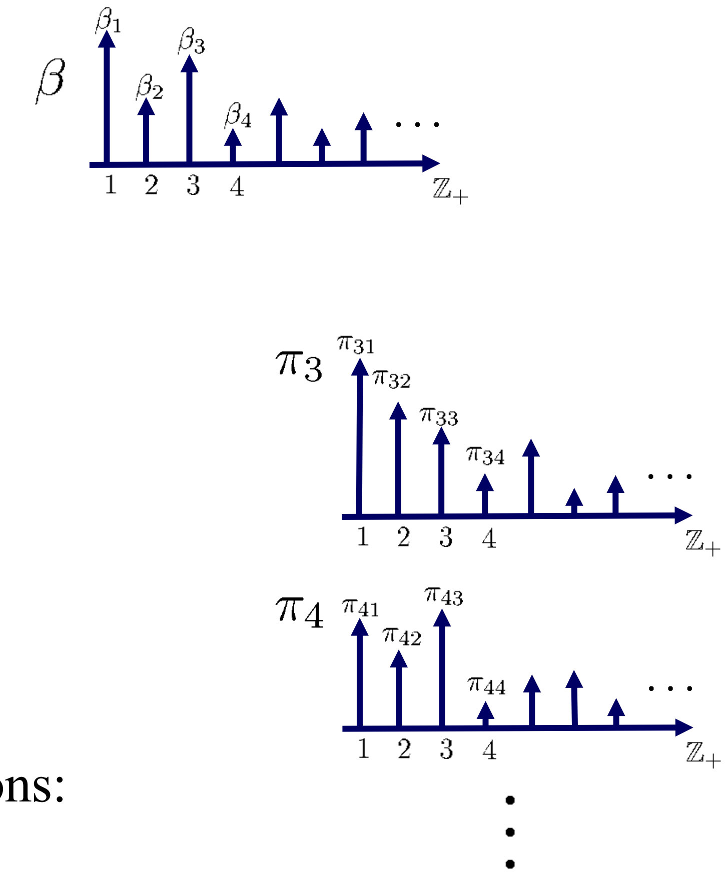




Example: Dirichlet process mixture models a distribution over an <u>infinite</u> number of mixture components

HDP-HMM





Hierarchical Dirichlet Process HMM

- Global transition distribution:

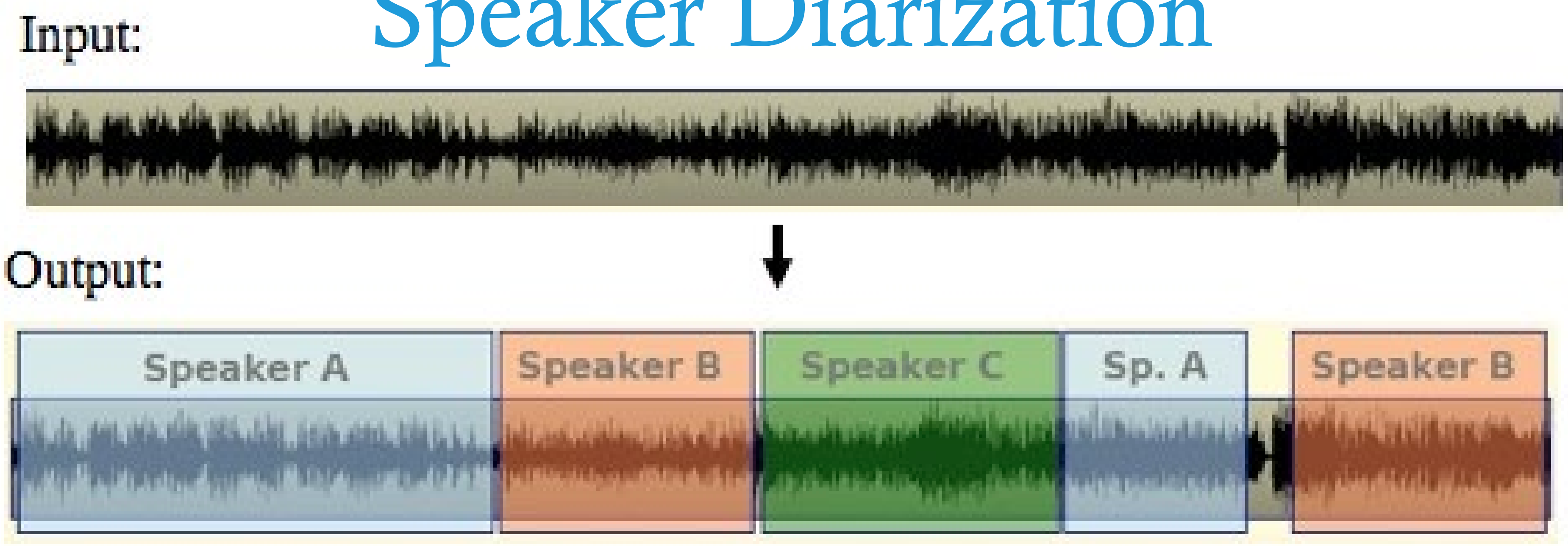
$$\beta \sim \text{Stick}(\gamma)$$

- Mode-specific transition distributions:

$$\pi_j \sim \text{DP}(\alpha\beta) \quad j = 1, 2, 3, \ldots$$

sparsity of β is shared $\longrightarrow$ $\boxed{E[\pi_{jk}] = \beta_k}$

Speaker Diarization

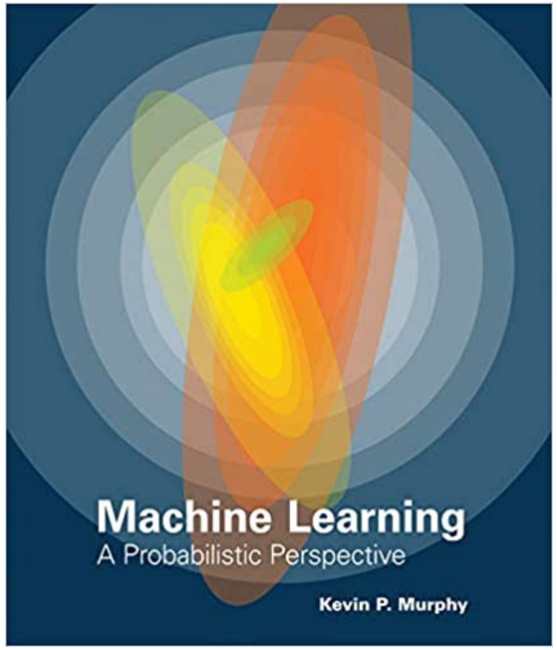


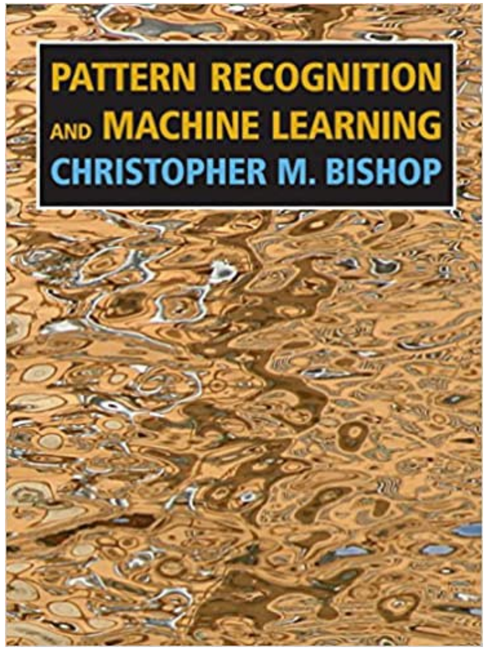
- ➢ Sticky HDP-HMM comparable to a state-of-the-art, heavily engineered speaker diarization system (Berkeley ICSI)

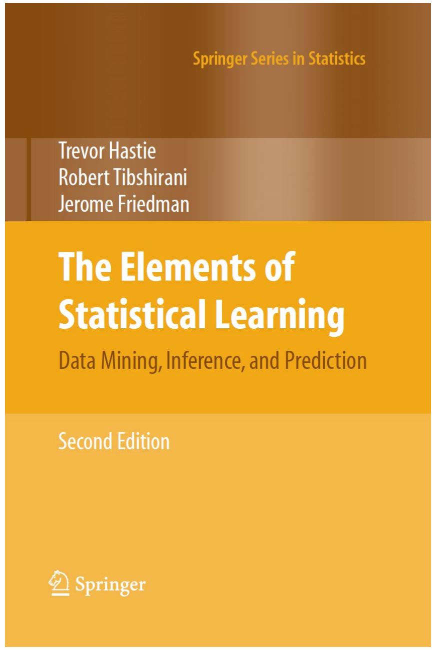
	Overall DER	Best DER	Worst DER
Sticky HDP-HMM	**17.84%**	1.26%	34.29%
Non-Sticky HDP-HMM	23.91%	6.26%	46.95%
ICSI	**18.37%**	4.39%	32.23%

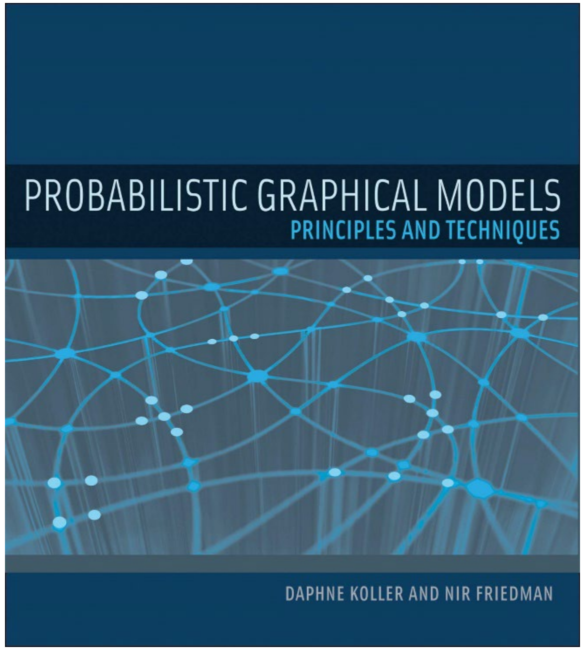
Summary

We covered a lot of ground…but there is a lot more to cover!









Important conferences to follow…

- **NeurIPS**
- **ICML**
- **AISTATS**
- AAAI / UAI

- ICRA
- IROS
- COLT

- IJCAI
- ICLR