COMP90051 **Statistical Machine Learning**

Semester 2, 2015

Lecturer: Ben Rubinstein

7. Example PGMs

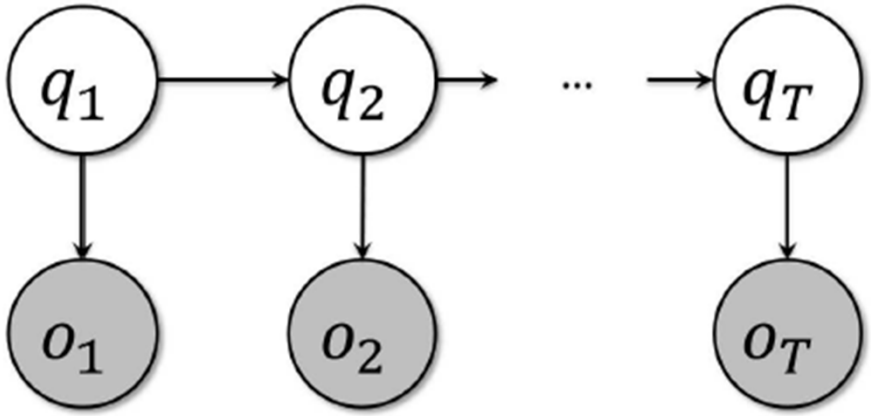
THE UNIVERSITY OF MELBOURNE

Sequence Analysis

The hidden Markov model (HMM);
and related Kalman Filter

The HMM (and Kalman Filter)

- Sequential observed outputs from hidden state



$A = \{a_{ij}\}$ transition probability matrix; $\forall i : \sum_j a_{ij} = 1$

$B = \{b_i(o_k)\}$ output probability matrix; $\forall i : \sum_k b_i(o_k) = 1$

$\Pi = \{\pi_i\}$ the initial state distribution; $\sum_i \pi_i = 1$

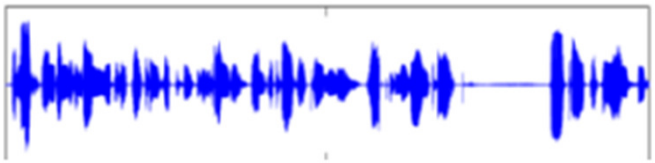
- The Kalman filter same with continuous Gaussian r.v.'s

HMM Applications

- NLP – part of speech tagging: given words in sentence, infer hidden parts of speech

  "I love Machine Learning" → noun, verb, object

- Speech recognition: given waveform, determine phonemes



- Biological sequences: classification, search, alignment
- Computer vision: identify who's walking in video, tracking

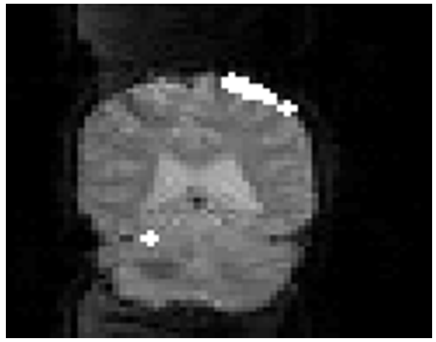
Fundamental HMM Tasks

HMM Task	PGM Task
Evaluation. Given an HMM μ and observation sequence O, determine likelihood $\Pr(O\|\mu)$	Probabilistic inference
Decoding. Given an HMM μ and observation sequence O, determine most probable hidden state sequence Q	MAP point estimate
Learning. Given an observation sequence O and set of states, learn parameters A, B, Π	Statistical inference

Computer Vision

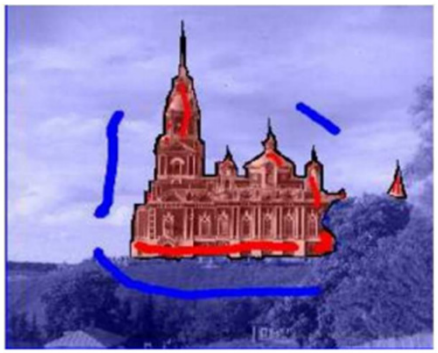
Hidden square-lattice Markov random fields

Pixel labelling tasks in Computer Vision



fMRI analysis (Kim et al. 2000)

Semantic labelling (Gould et al. 09)

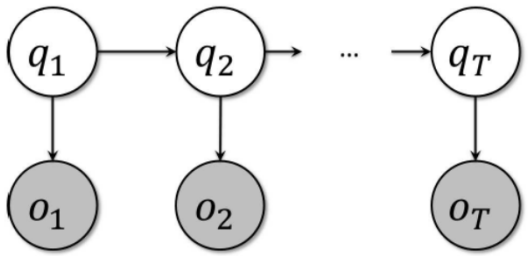


Interactive figure-ground segmentation (Boykov & Jolly 2011)

Denoising (Felzenszwalb & Huttenlocher 04)

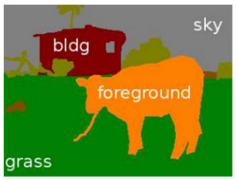
What these tasks have in common

- Hidden state representing semantics of image
  - fMRI: Tumour vs benign tissue
  - Semantic labelling: Cow vs. tree vs. grass vs. sky vs. house
  - Fore-back segment: Figure vs. ground
  - Denoising: Clean pixels
- Pixels of image
  - What we observe of hidden state
- Remind you of HMMs?

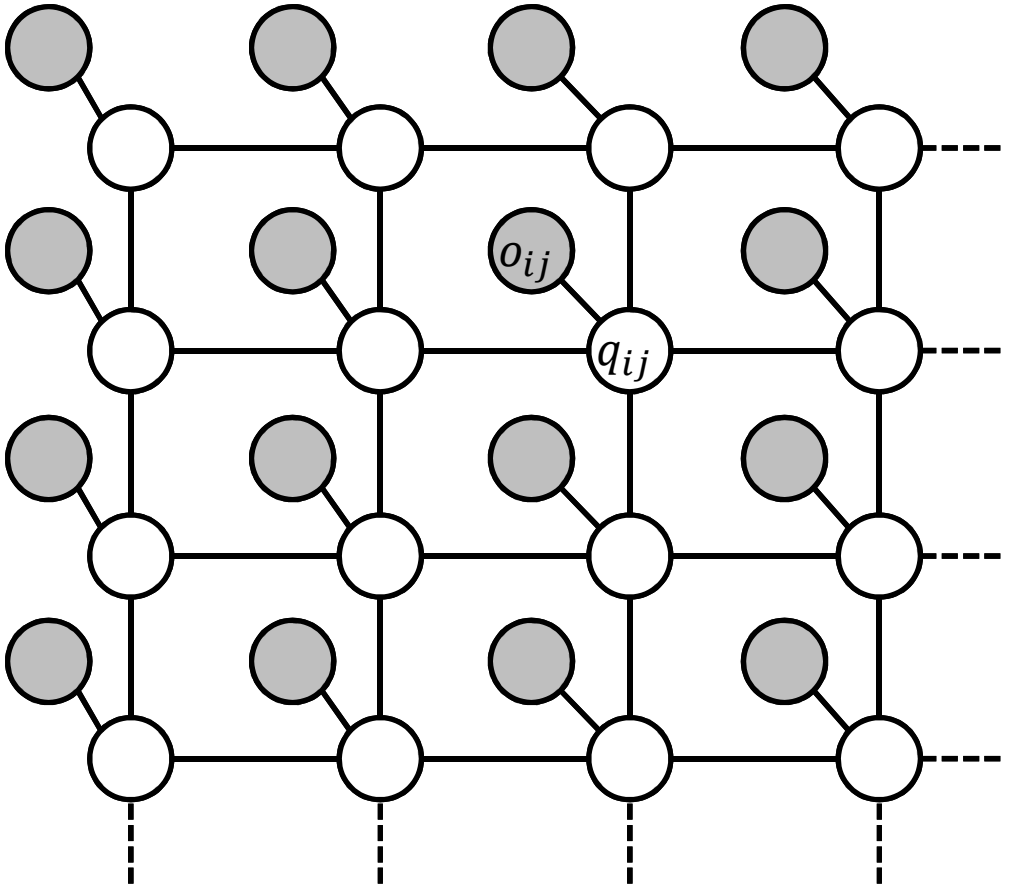


A hidden square-lattice MRF

- Hidden states: square-lattice model
  - Boolean for two-class states
  - Discrete for multi-class
  - Continuous for denoising



- Pixels: observed outputs
  - Continous e.g. Normal



Topic Modelling

Latent Dirichlet Allocation

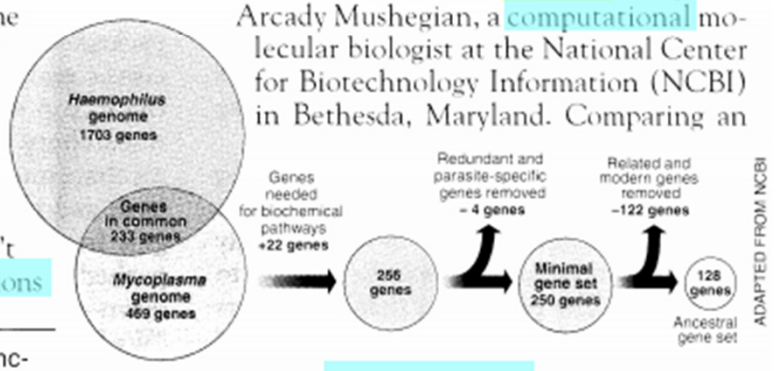
Based on David Blei's 2008 Science slides

Documents exhibit multiple topics

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

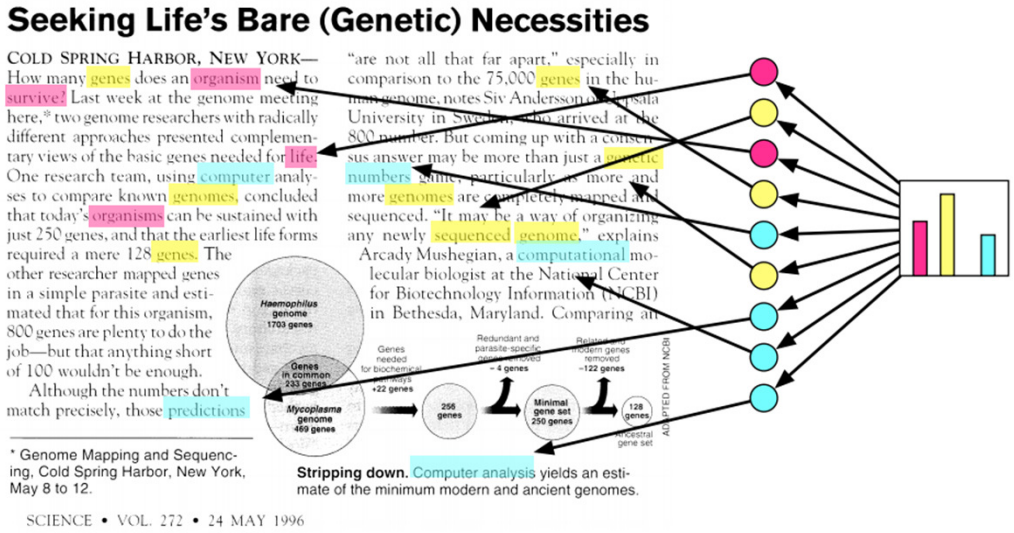


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Generative process



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

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

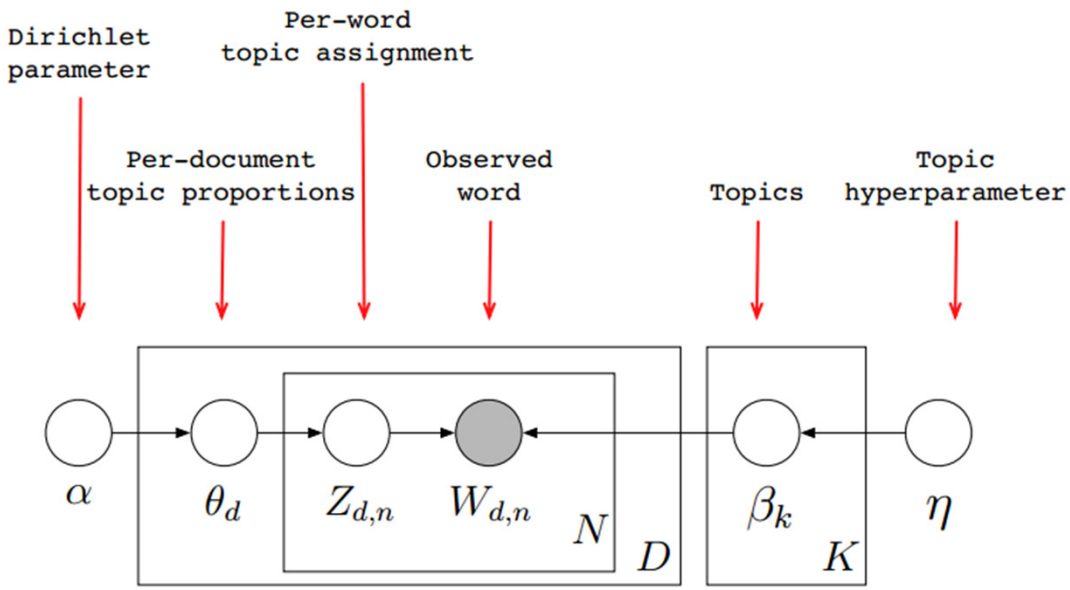
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

- "Corpus" of documents
- Corpus-wide topics
- Generative model
  - Each document a random mixture of topics
  - Each word in document drawn from one of these topics

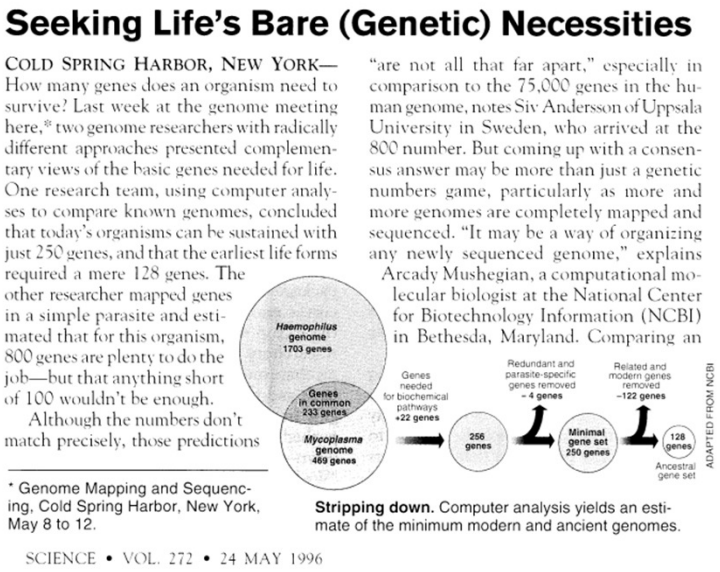
Latent Dirichlet allocation

- For all K topics
  - Draw topic's word distribution β_i from Dirichlet
- For all D documents
  - Draw topic proportions θ_d from Dirichlet
  - For all N words
    - Draw word's topic $Z_{d,n}$ from Multinomial in per-document topics
    - Draw word itself $W_{d,n}$ from Multinomial in topic's word distrib



LDA Example

- Data: journal *Science* 1990-2000
  - 17k documents, 11m words, 20k unique terms
- *K*=100 topic LDA



Seeking Life's Bare (Genetic) Necessities

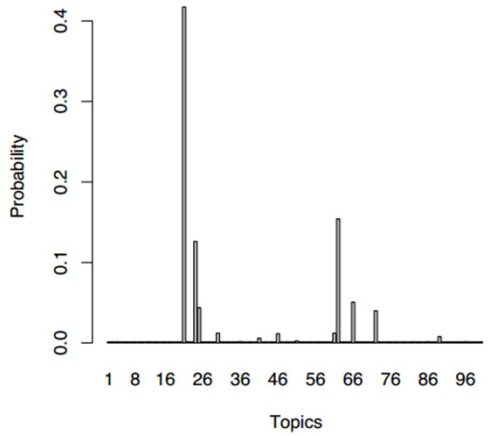
COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known 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

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

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

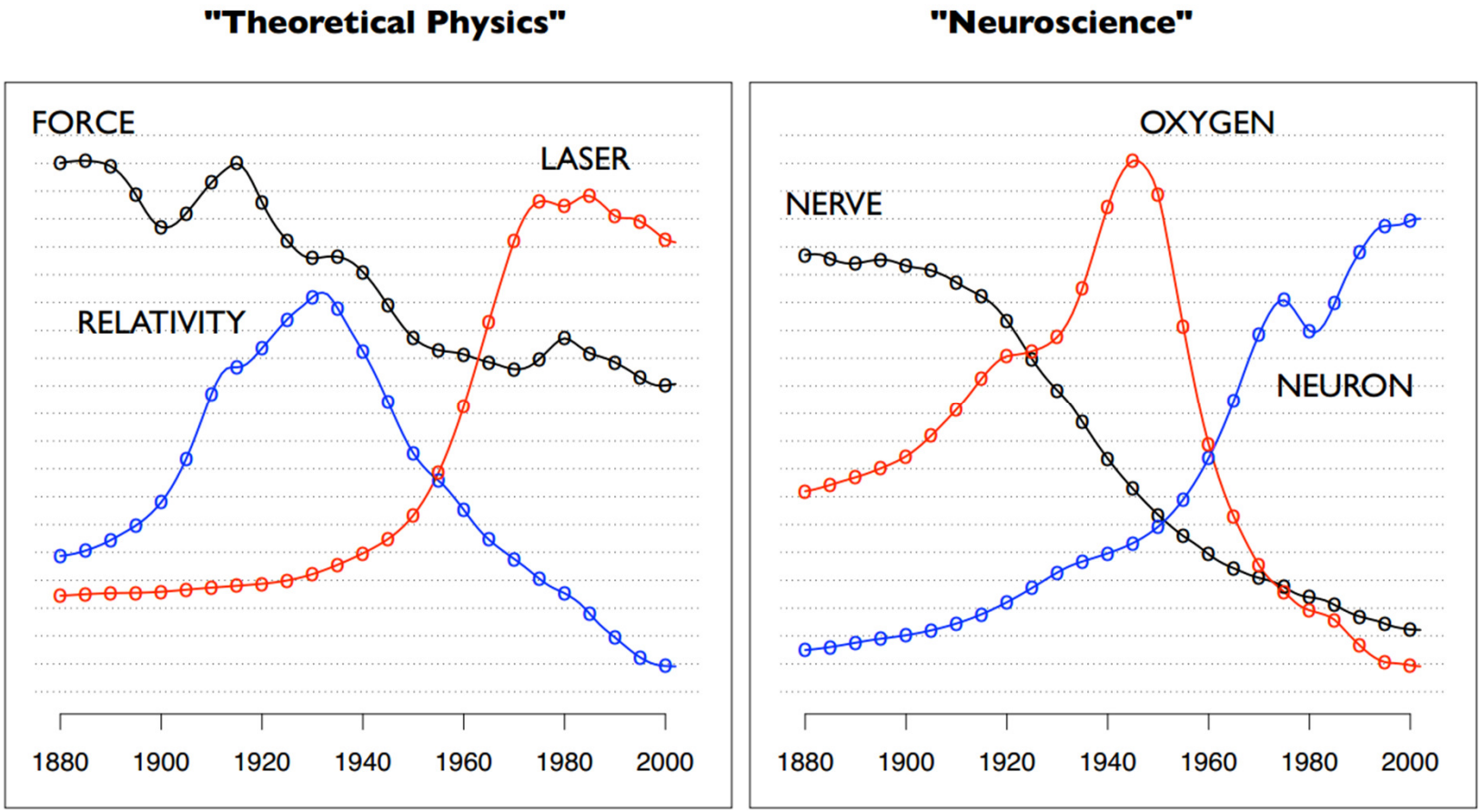
SCIENCE • VOL. 272 • 24 MAY 1996



LDA Example: Topics

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

LDA Example: Dynamic model



Summary

- Diverse examples of PGMs
  - HMM/Kalman filter → speech reco, bioinformatics
  - Square-lattice MRFs → pixel labelling, computer vision
  - Latent Dirichlet allocation → topic modelling in IR/NLP
- Share common training, prediction algorithms
  - Elimination-based or sampling for probabilistic inference
  - MLE or EM for training