Graphical Models

Lecture 1:

Motivation and Foundations

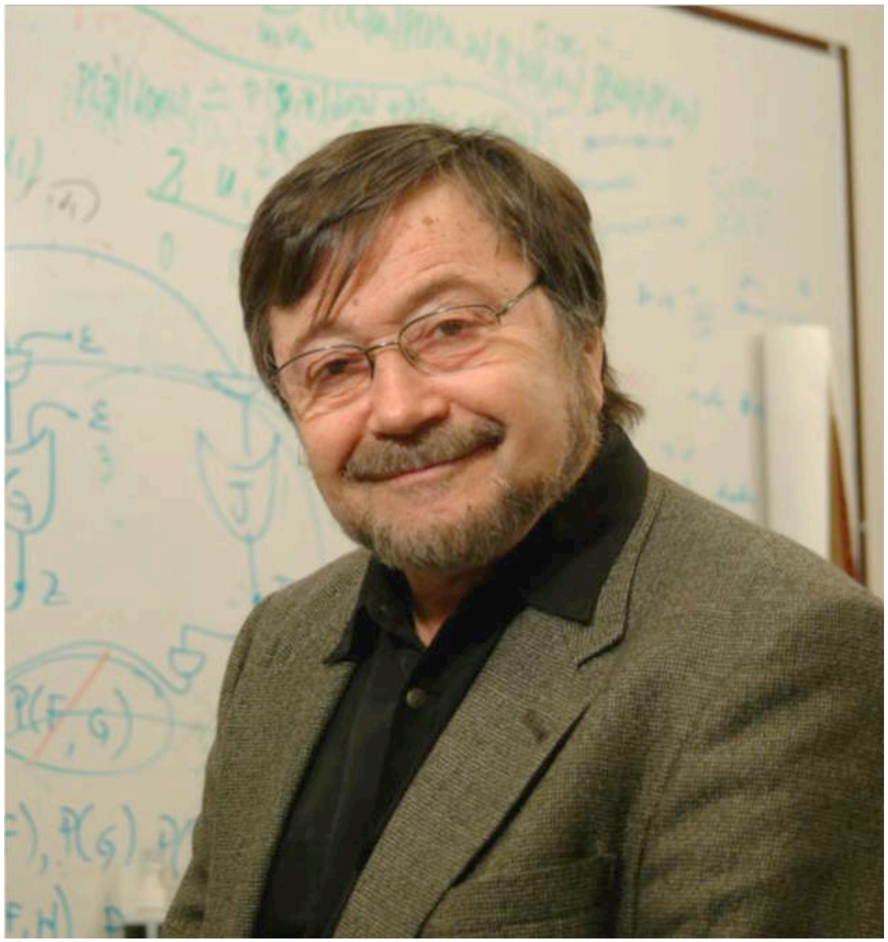
Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.

Board work

- Expert systems
  - the desire for probability and dependencies.
- Joint probability tables
- Exponential blow-up in size
- MYCIN & certainty factors

Judea Pearl (1936-)

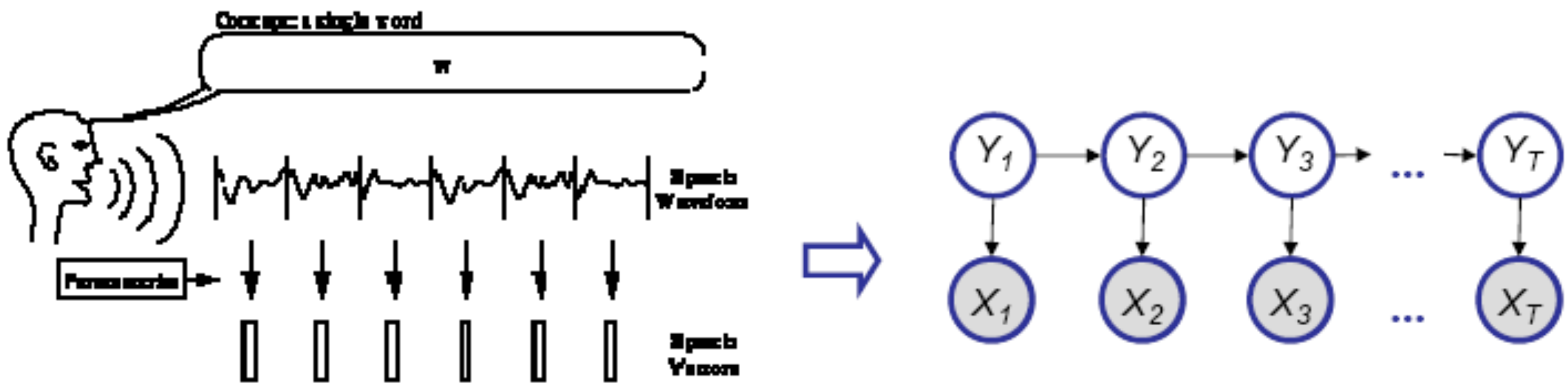


- First proposed Bayesian Networks, qualitative structure for encoding independence relations. c.1988
- *Probabilistic Reasoning in Intelligent Systems*
- Now working on causality.

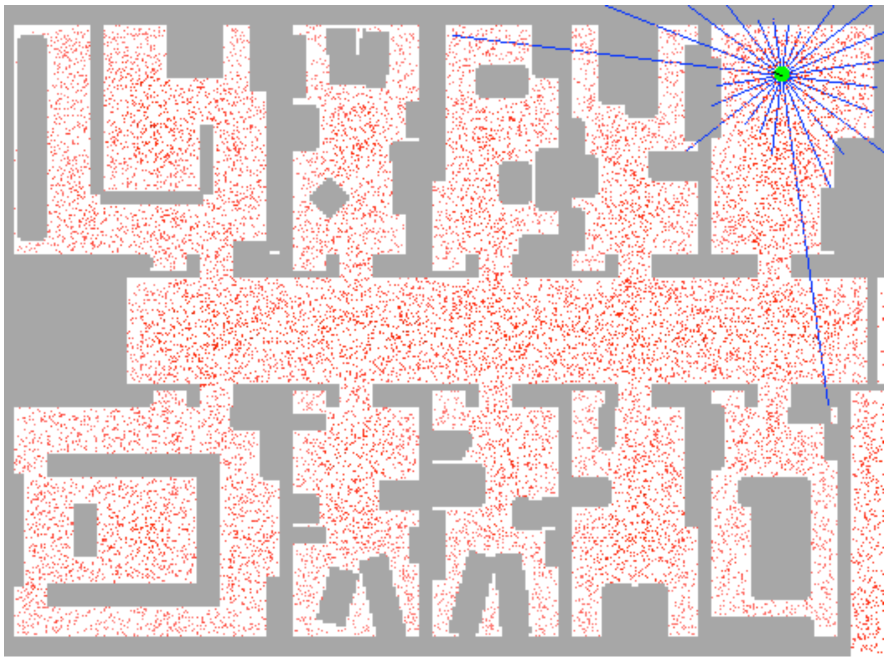
Probabilistic Graphical Models

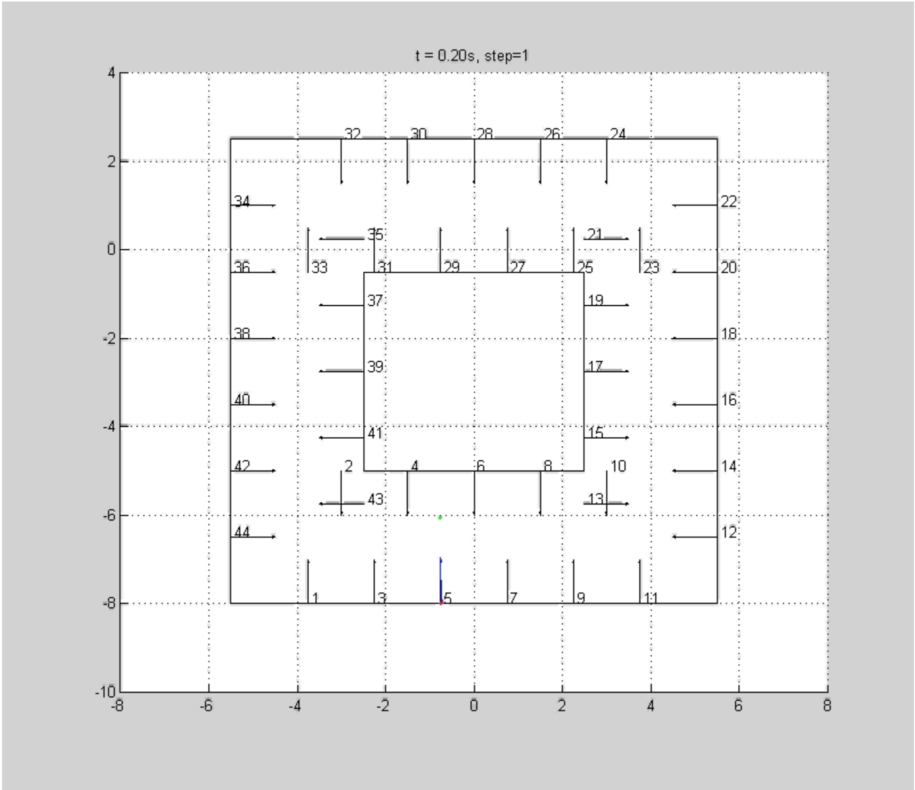
- Framework for obtaining, representing, querying large probability distributions.
- A beautiful formalism that *generalizes* many ideas from CS and from Statistics.
- Carlos Guestrin: "one of the most exciting developments in machine learning (knowledge representation, AI, EE, Stats, …) in the last two (or three, or more) decades…"

Speech Recognition

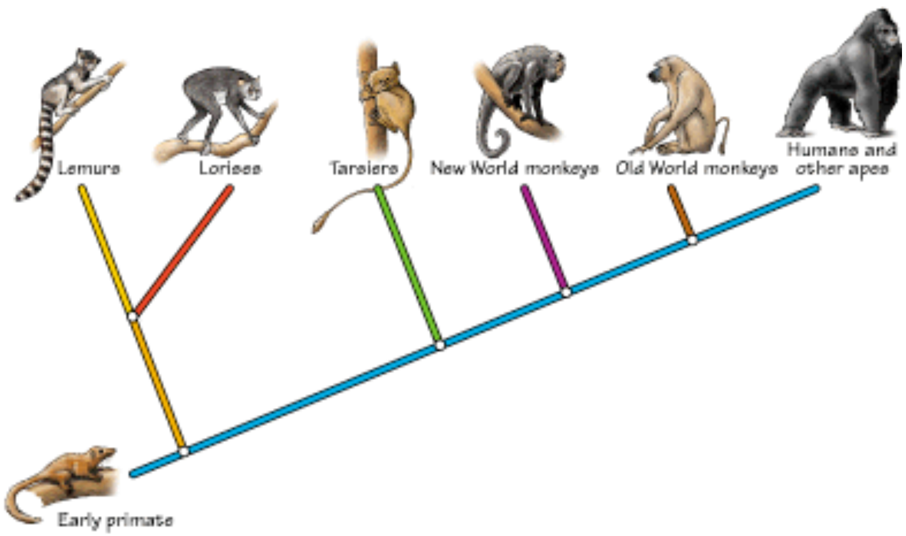


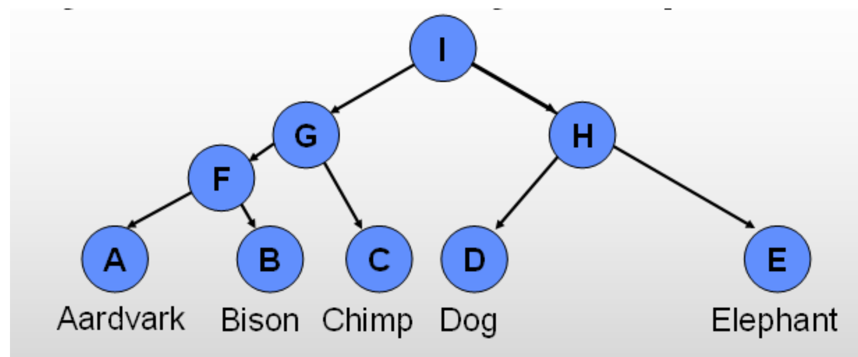
Robot Navigation



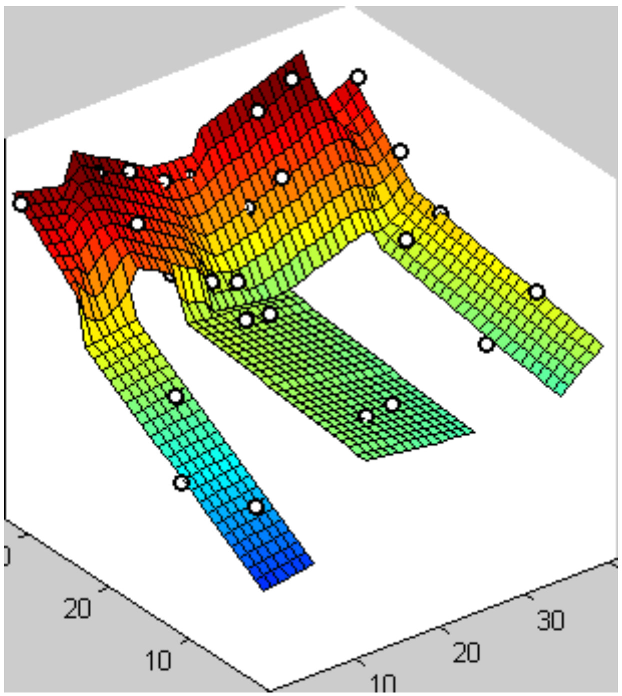


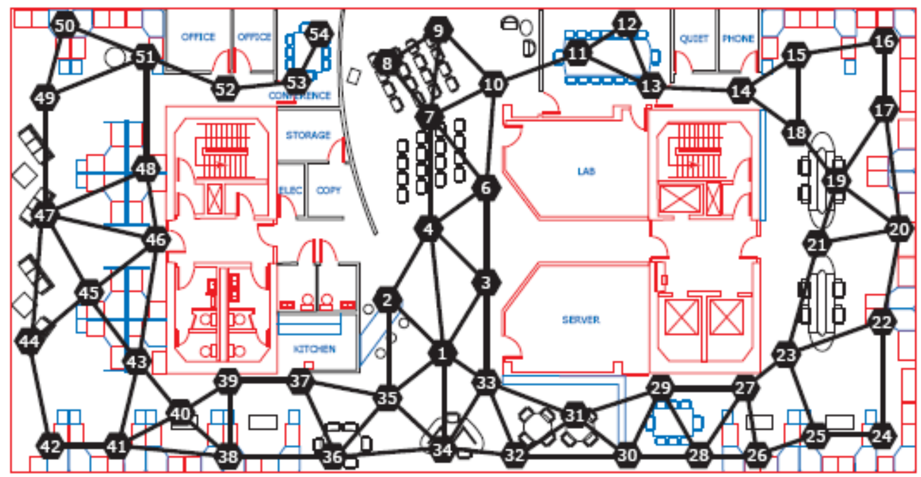
Evolutionary Biology

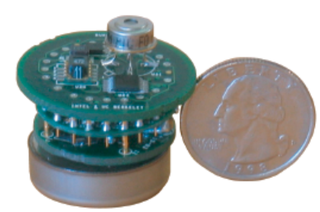




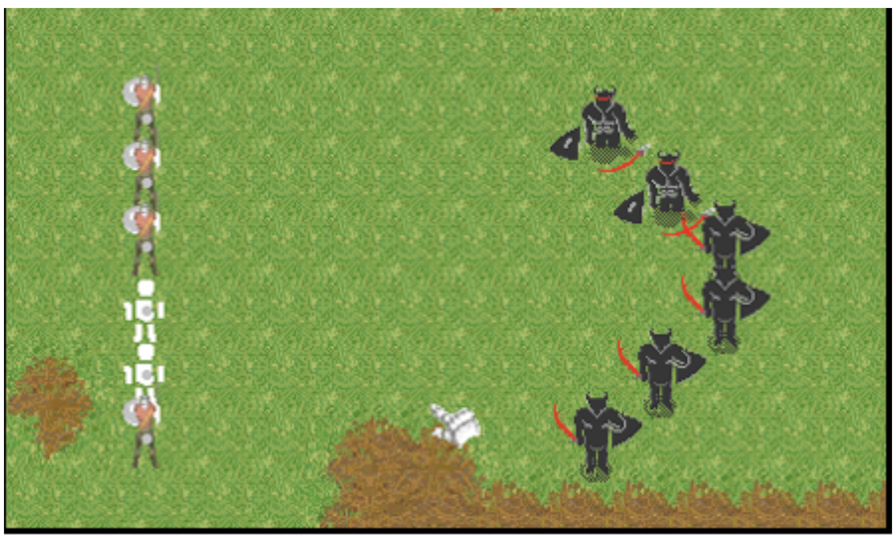
Sensor Data

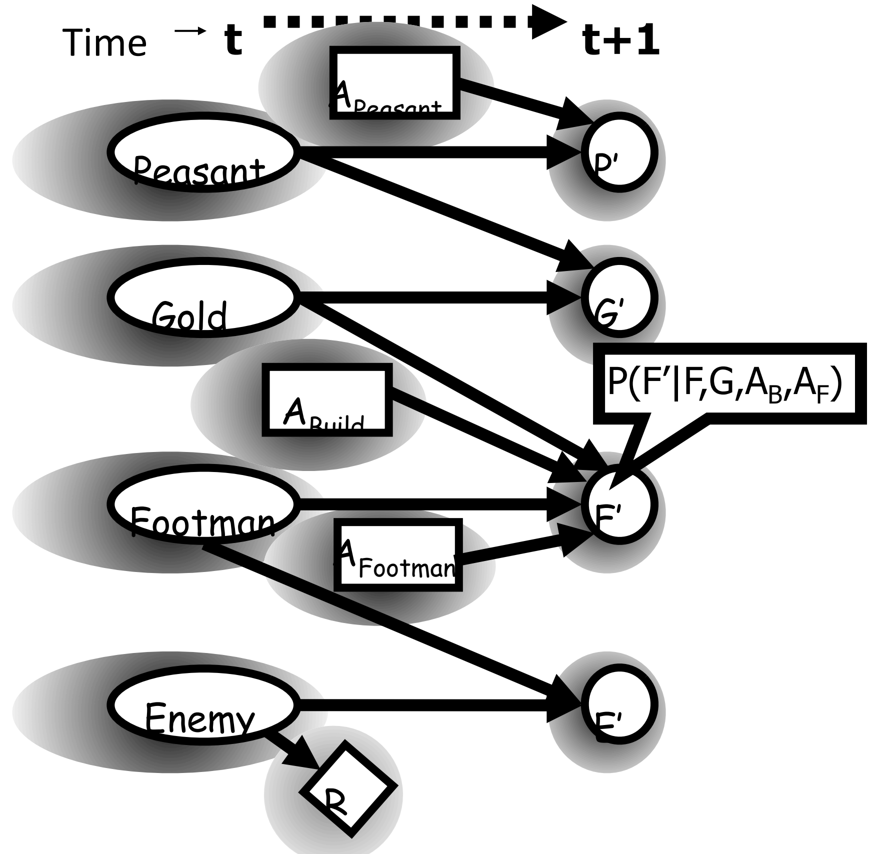

20
10
30
20
10


OFFICE
OFFICE
CONFERENCE
STORAGE
ELEC
COPY
KITCHEN
LAB
SERVER
QUIET
PHONE

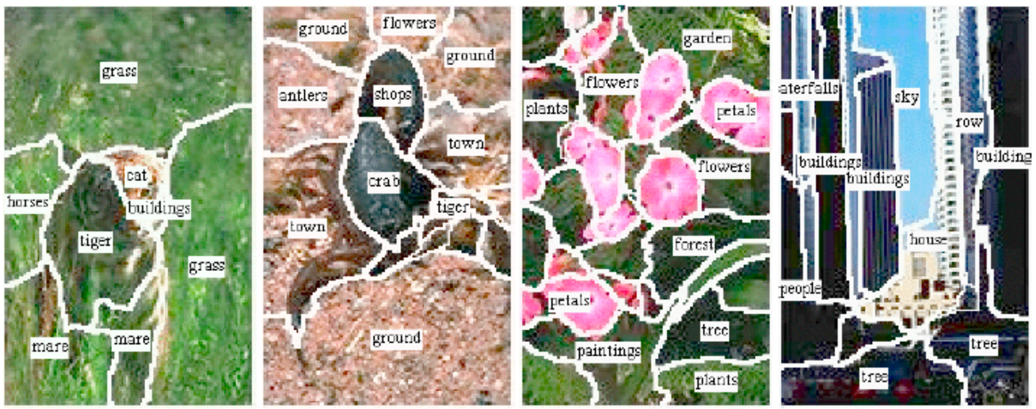


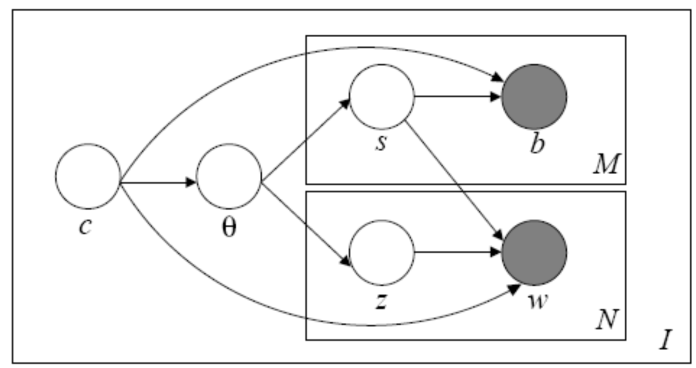
Planning



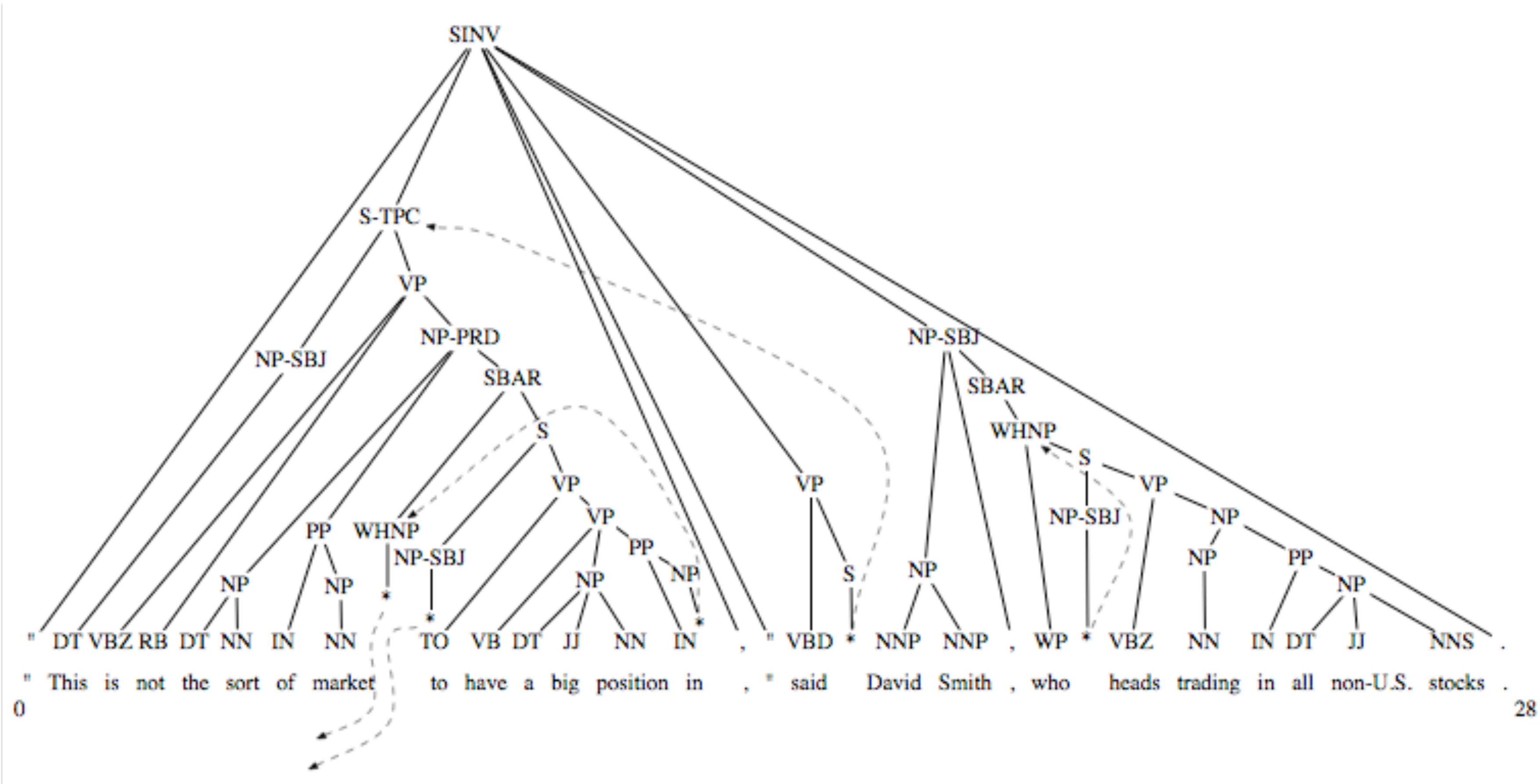


Images

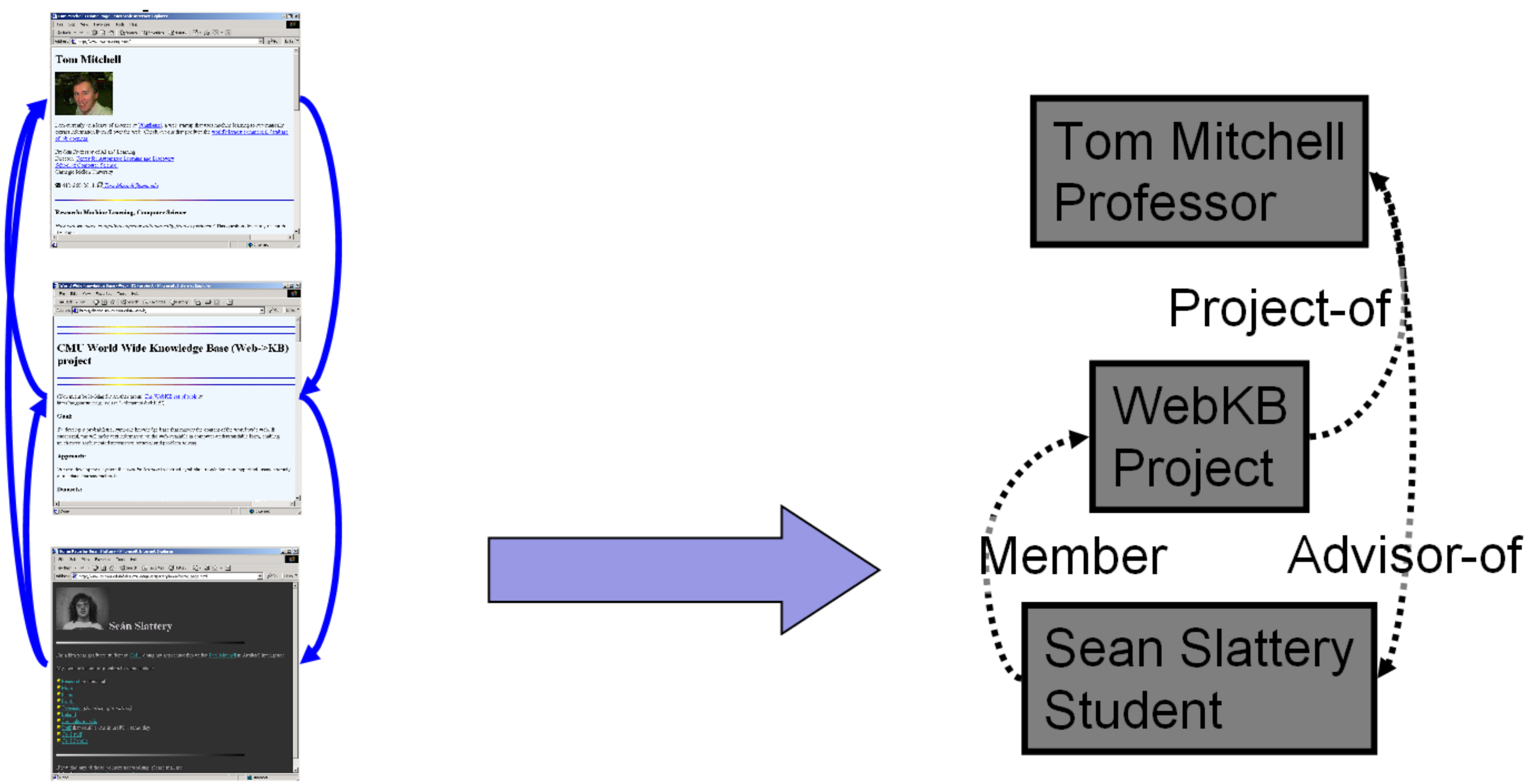

grass
horses
cat
buildings
tiger
grass
mare
mare
ground
flowers
ground
antlers
shops
town
crab
tiger
town
ground
garden
flowers
plants
petals
flowers
forest
petals
tree
paintings
plants
aterfalls
sky
row
buildings
buildings
building
house
people
tree
tree


c
θ
s
b
M
z
w
N
I

Linguistic Structure



Relational, Structured Data



Applications of GMs

- Speech recognition (hidden Markov models)
- Tracking and robot localization (Kalman filters)
- Evolutionary biology (Bayesian networks)
- Modeling sensor data (undirected GMs)
- Planning under uncertainty (dynamic BNs, factored Markov decision problems)
- Images (hierarchical BNs)
- Natural language processing (probabilistic grammars)
- Structured data like social networks and linked documents (probabilistic relational models)
- ... (your additions?)

GMs as a *Lingua Franca*

- A major barrier to research is communication: people from different technical backgrounds speak different "languages."
- PGMs are a language with a diverse following; they enable cross-fertilization that wasn't possible before.
- Cases in point:
  - missiles and speech
  - chemistry and CS
  - computational social science

What We're Going to Cover

1. Representation
   – Bayesian networks (directed GMs)
   – Markov networks (undirected GMs)
2. Inference
   – exact, approximate
   – variational, sampling
3. Learning
   – parameters, structure
4. Research Topics

Keyword Soup

D-separation

Variable Elimination

Expectation Maximization

Bayes Ball algorithm

Junction Tree

Maximum Likelihood

Mean Field Inference

Tree-width

Gradient Optimization

Factor Graphs

Variational Inference

Conditional Random Fields

Context-specific Independence

Generalized BP

Loopy BP

Boltzmann Machines

Partition Function

Lagrange Multipliers

Free Energy

Gaussian Random Field

Deep Belief Networks

Expectation Propagation

Exponential Family

Collapsed Gibbs Sampling

Chow-Liu Algorithm

Markov Blanket

Markov-chain Monte Carlo

Topic Models

Metropolis-Hastings

Moralization

Blocked Gibbs Sampling

Variational Bayes

Probabilistic Relational Models

Particle filtering

Non-parametric Models

MAP Inference

Dirichlet Process

Linear-Programming

Foundations

Events & Random Variables

- **Space** of possible **outcomes**
- An **event** is a subset of the outcomes.
- Events are complicated!
  - We tend to *group* events by **attributes**
  - Person → Age, Grade, HairColor
- **Random variables** formalize attributes:
  - "Grade = A" is shorthand for the set of events: $\{\omega \in \Omega : f_{\text{Grade}}(\omega) = A\}$
- Properties of random variable X:
  - Val(X) = possible values of X
  - For discrete (categorical): $\sum_{x \in \text{Val}(X)} P(X = x) = 1$
  - For continuous: $\int P(X = x) dx = 1$

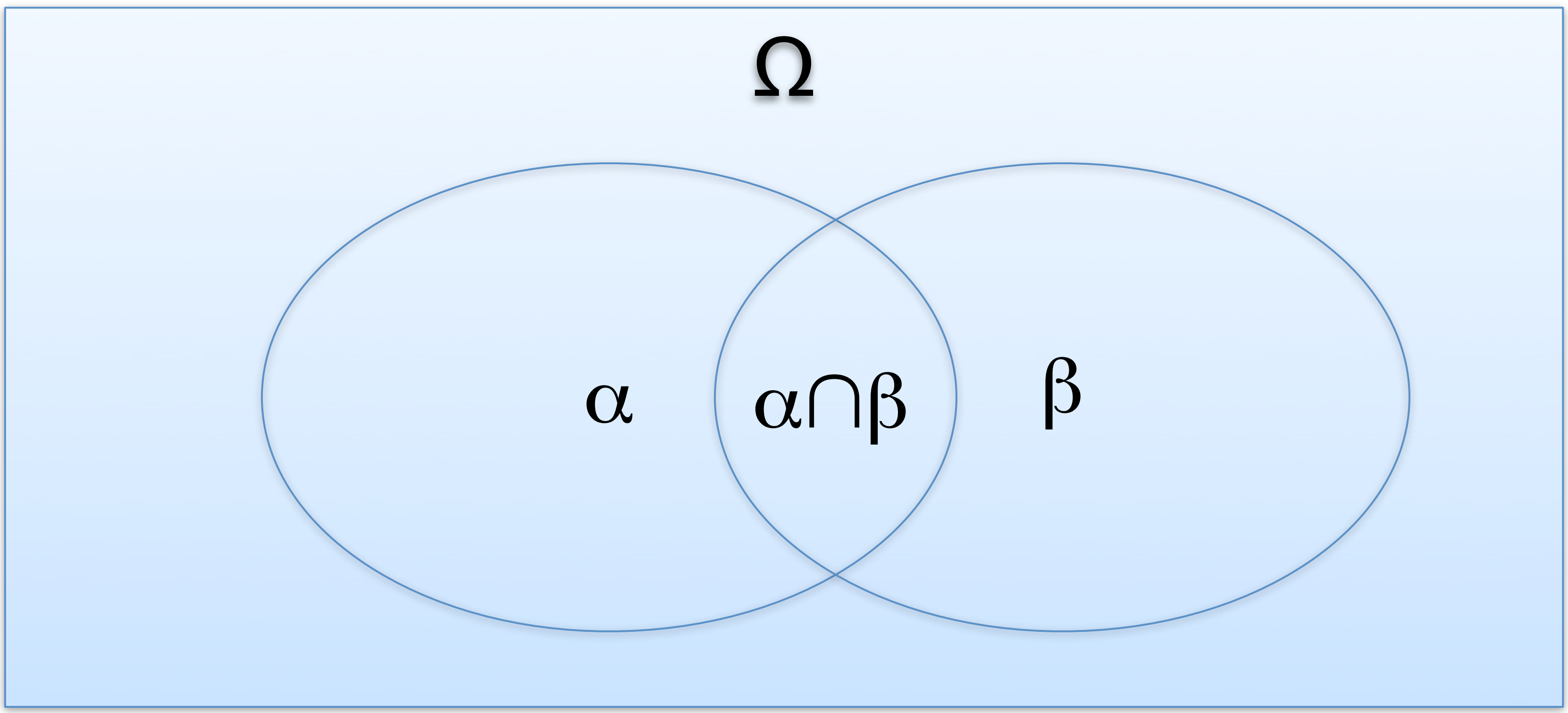
$$\forall x \in \text{Val}(X), P(X = x) \geq 0$$

Two Interpretations of Probability

- Frequentists
  - $P(\alpha)$ is the frequency of α in the limit
  - Many arguments against this interpretation
    - What is the frequency of the event "it will rain tomorrow"?
- Subjective (Bayesian) interpretation
  - $P(\alpha)$ is my degree of belief that α will happen
  - What does "degree of belief" mean? Ground in betting.
  - Vaguely expecting a horse, catching a glimpse of a donkey, and strongly believing you've seen a mule.
- For this class, we (mostly) don't care which camp you are in.

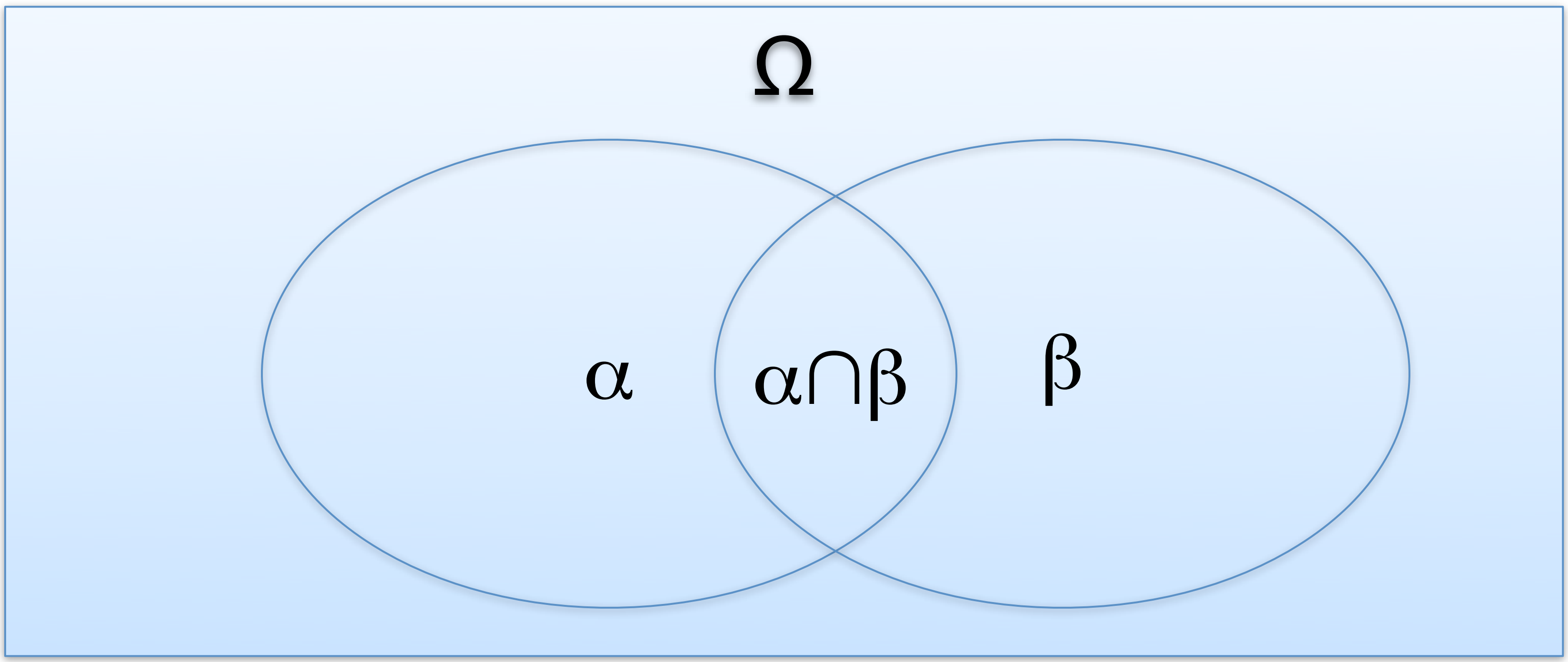
Conditional Probabilities

- After learning that α is true, how do we feel about β? $P(\beta \mid \alpha)$



Chain Rule

$$P(\beta|\alpha) = \frac{P(\alpha \cap \beta)}{P(\alpha)} \qquad P(\alpha \cap \beta) = P(\alpha)P(\beta \mid \alpha)$$



$$P(\alpha_1 \cap \cdots \cap \alpha_k) = P(\alpha_1)P(\alpha_2 \mid \alpha_1) \cdots P(\alpha_k \mid \alpha_1 \cap \ldots \cap \alpha_{k-1})$$

"Factorization"

Bayes Rule

$$P(\alpha|\beta)P(\beta) = P(\alpha \cap \beta) = P(\beta|\alpha)P(\alpha)$$

likelihood

prior

$$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)P(\alpha)}{P(\beta)}$$

posterior

normalization constant

$$P(\alpha \mid \beta \cap \gamma) = \frac{P(\beta \mid \alpha \cap \gamma)P(\alpha \mid \gamma)}{P(\beta \mid \gamma)}$$

where γ is an “external event”

Marginalization and Conditioning

- (Shown in tables on board)

Back to our Problem

Large number of variables.

Massive joint probability table.

Independence

- α and β are **independent** if
  $P(\beta|\alpha) = P(\beta)$
- Independence implied by joint table $P(\alpha,\beta)$
  $P : (\alpha \perp \beta)$
- α and β are independent, **written**: $\alpha \perp \beta$
- **Proposition:** α and β are **independent**
  if and only if
  $P(\alpha \cap \beta) = P(\alpha)\, P(\beta)$

Conditional Independence

- Independence is rarely true.

- α and β are **conditionally independent** given γ if
  $P(\beta \mid \alpha \cap \gamma) = P(\beta \mid \gamma)$

$$P : (\alpha \perp \beta \mid \gamma)$$

Proposition: $P : (\alpha \perp \beta \mid \gamma)$ if and only if

$P(\alpha \cap \beta \mid \gamma) = P(\alpha \mid \gamma)\, P(\beta \mid \gamma)$

Board Work

- Conditional independence and compression

Conditional Independence of Random Variables

- **Sets** of variables **X**, **Y**, **Z**
- **X** is independent of **Y** given **Z** if
  - $P : (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} | \mathbf{Z}=\mathbf{z})$,
    $\forall\ \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$

- Shorthand:
  - **Conditional independence:** $P : (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P : (\mathbf{X} \perp \mathbf{Y} \mid \emptyset)$, write $P : (\mathbf{X} \perp \mathbf{Y})$

- **Proposition:** P satisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\ P(\mathbf{Y}|\mathbf{Z})$

Properties of Independence

- Symmetry:
  - $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{Y} \perp \mathbf{X} \mid \mathbf{Z})$
- Decomposition:
  - $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
- Weak union:
  - $(\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}, \mathbf{W})$
- Contraction:
  - $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y}, \mathbf{Z}) \wedge (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$
- Intersection:
  - $(\mathbf{X} \perp \mathbf{W} \mid \mathbf{Y}, \mathbf{Z}) \wedge (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{W}, \mathbf{Z}) \Rightarrow (\mathbf{X} \perp \mathbf{Y}, \mathbf{W} \mid \mathbf{Z})$
  - Only for positive distributions: $P(\alpha) > 0, \forall\alpha, \alpha \neq \emptyset$

Course Administration

Learning Philosophy

- Learn by doing!
  - Don’t really understand belief-propagation until you have implemented it yourself.
  - Therefore, 7 programming assignments.
- How to make this OK.
  - Simple, yet open-ended; you define.
  - Use whatever programming language you like. MatLab, Java, F#.
  - Write a ~2 page report about your experiences.
    Informal format.
  - We will give you data, but not code.
  - No long written HW assignments
  - No midterm. No final project.
  - Mini-quizzes. “pen & pencil quizzes”. (Sometimes take home)
  - Drop lowest programming assignment grade.

Grading

- 50% homework (programming assignments)
  - We will read your reports lovingly
  - Coarse-grained grading: check, ++, +, -, --.
- 10% quizzes
- 25% final exam
- 15% class participation
  - Might include brief presentation of your HW

Prerequisites

- Helpful to have some working knowledge of...
  - Probability (distributions, densities, marginalization)
  - Basic statistics (moments, typical distributions, regression)
  - Algorithms (dynamic programming, basic data structures, a little complexity)

- Programming
  - Facility with some programming language of your choice
  - Helpful to have some experience programming for machine learning (e.g. NB, HMM)

- Ability to deal with "abstract mathematical concepts"

About the Instructor

- Main research focus:
  - Information extraction ⇔ Databases ⇔ Data mining
  - Natural language processing, Machine Learning
  - Joint Inference
  - Structured prediction
  - Semi-supervised learning
- Why I'm teaching this course
  - promote the topic
  - perceived need & interest in the department

About the TAs

- Michael Wick
  - parameter estimation in undirected graphical models
  - probabilistic databases

- Sameer Singh
  - joint inference
  - parallel & distributed graphical models

More Info & Contacting Us

- Course web site:
  http://www.cs.umass.edu/~mccallum/courses/gm2011

- Mailing Lists (coming soon):
  `691gm-staff@cs.umass.edu`
  `691gm-all@cs.umass.edu`

Summary So Far

- Graphical model motivation.
- Basic definitions of probabilities, independence, conditional independence, chain rule, bayes rule
- Basic idea: exploiting conditional independence to compress joint prob table.

- Next time: semantics of Bayesian networks