Probabilistic Graphical Models

Lecture 1 – Introduction

CS/CNS/EE 155
Andreas Krause

One of the **most exciting advances** in machine learning (AI, signal processing, coding, control, …) in the last decades

How can we gain **global insight** based on **local observations**?

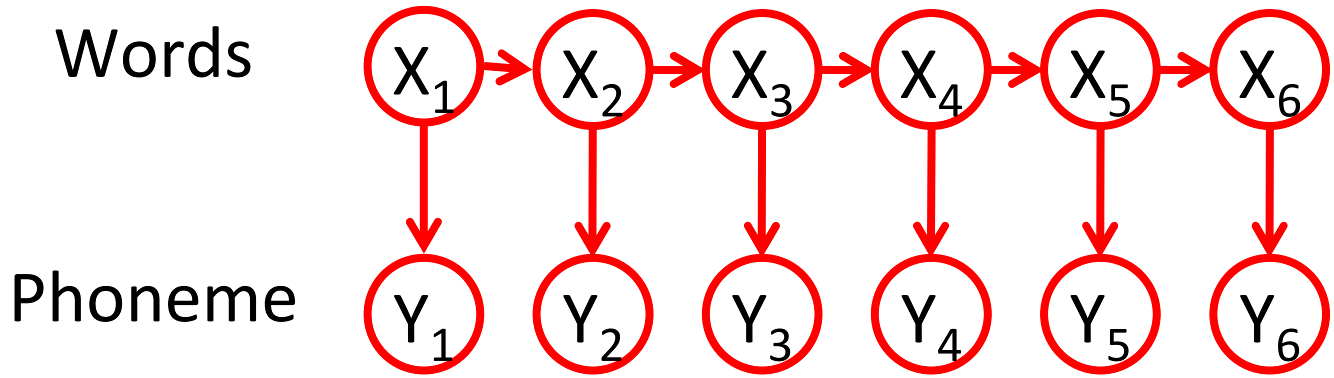
Key idea:

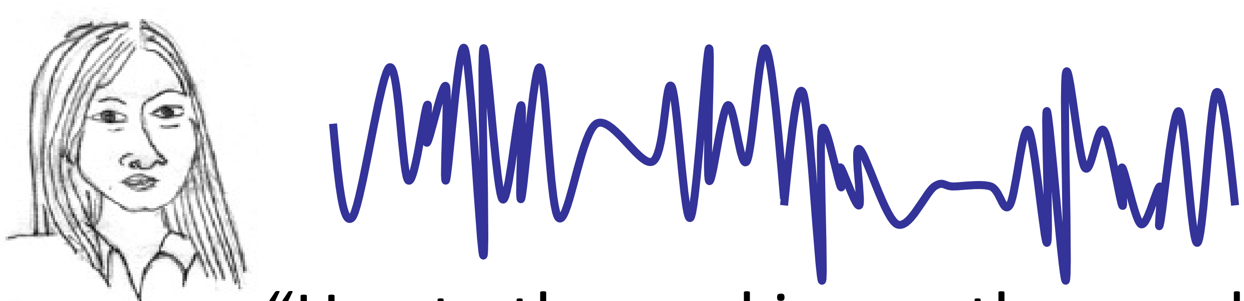
- **Represent** the world as a collection of random variables $X_1, \ldots X_n$ with joint distribution $P(X_1,\ldots,X_n)$
- **Learn** the distribution from data
- Perform "**inference**" (compute conditional distributions $P(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$

Applications

Natural Language Processing

Speech recognition

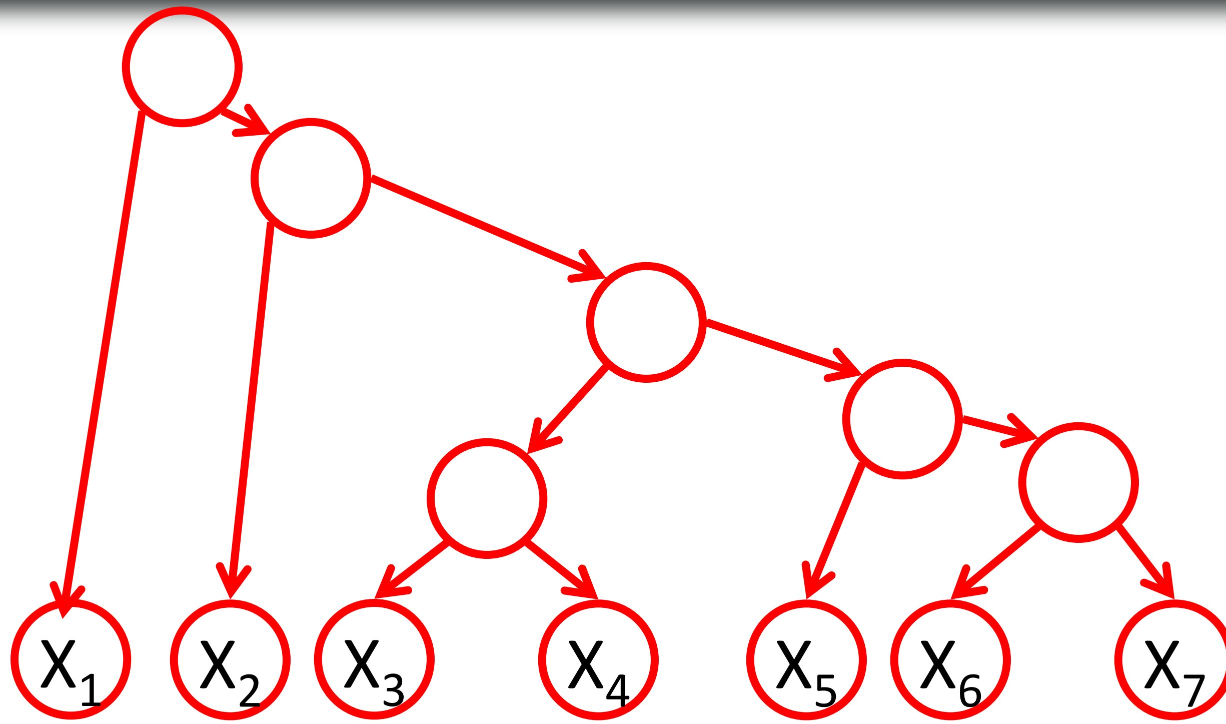




"He ate the cookies on the couch"

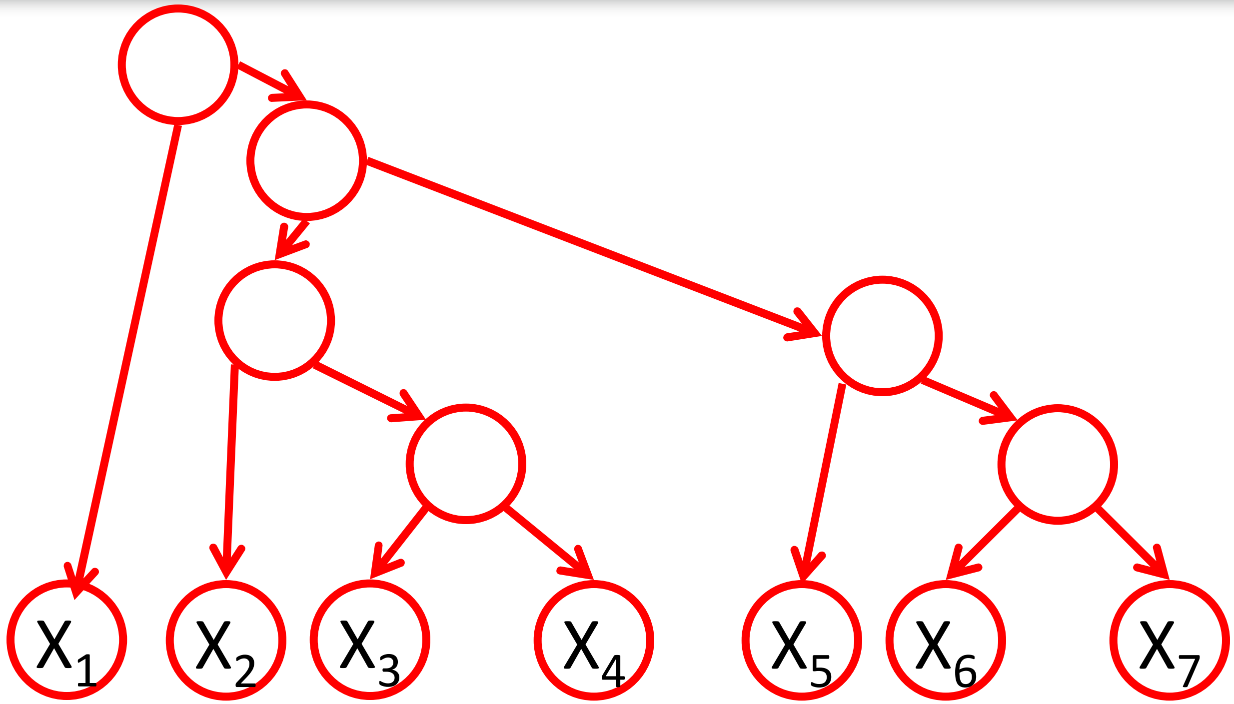
- Infer spoken words from audio signals
- "Hidden Markov Models"

Natural language processing



"He ate the cookies on the couch"

Natural language processing

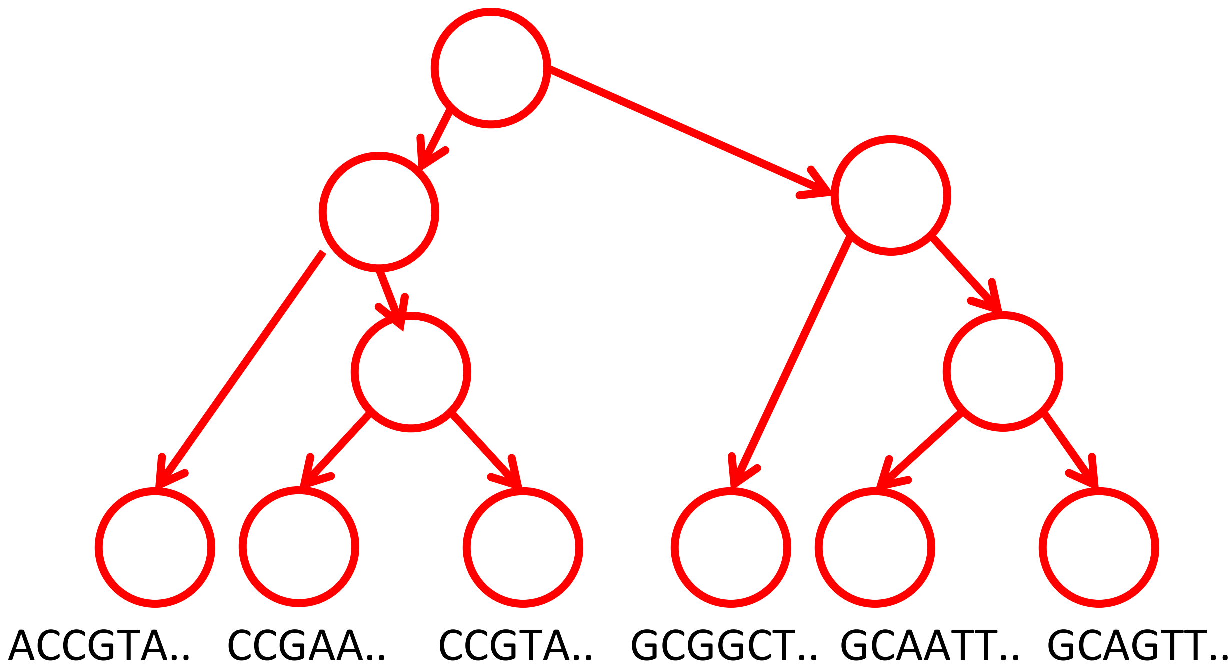


"He ate the cookies on the couch"

- Need to deal with ambiguity!
- Infer grammatical function from sentence structure
- "Probabilistic Grammars"

Evolutionary biology

[Friedman et al.]



- Reconstruct phylogenetic tree from current species (and their DNA samples)

Applications

Computer Vision

Image denoising

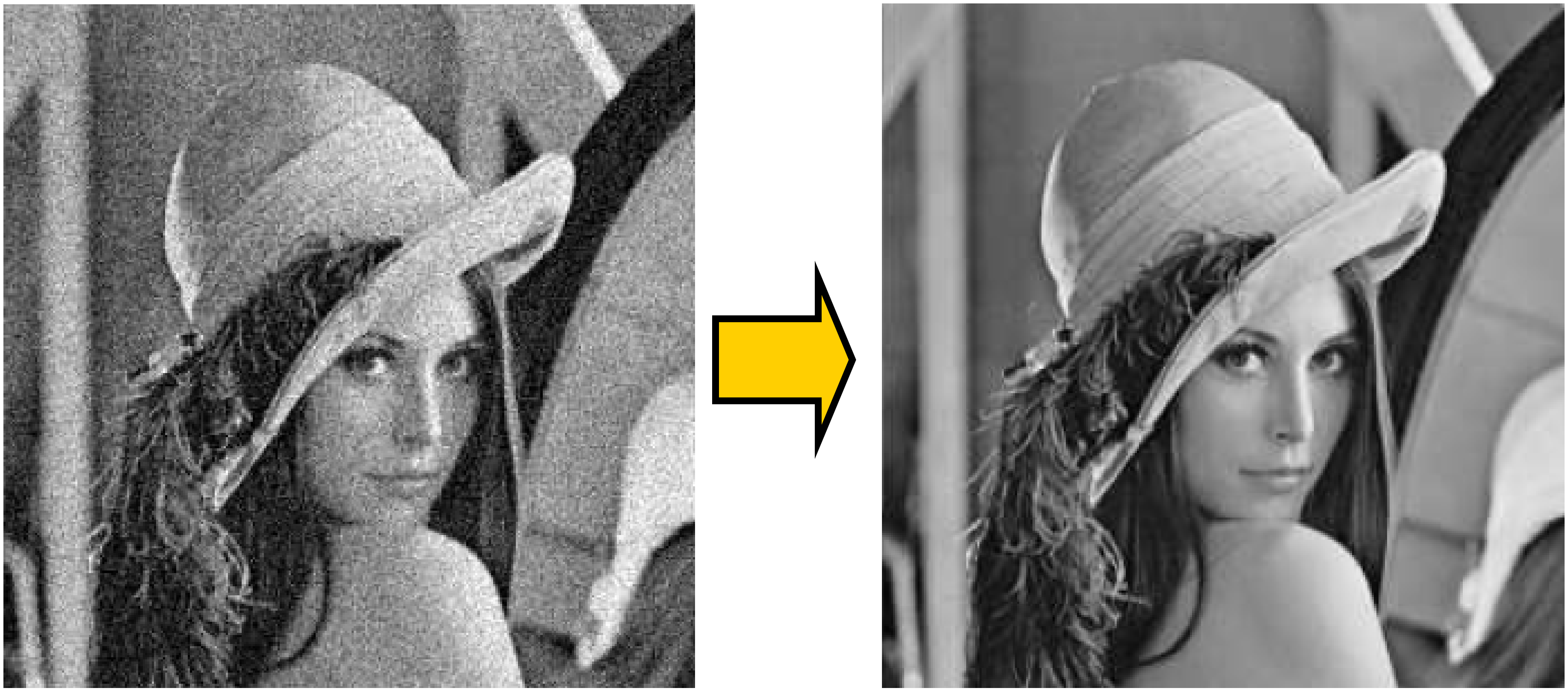
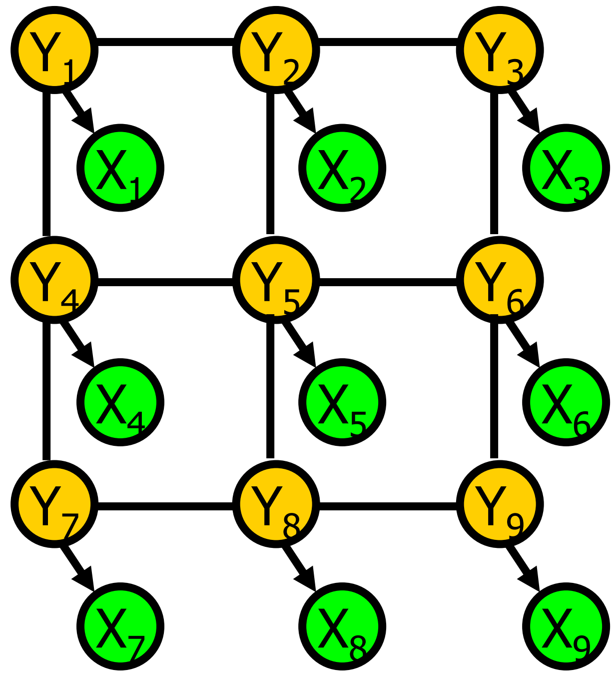


Image denoising

Markov Random Field

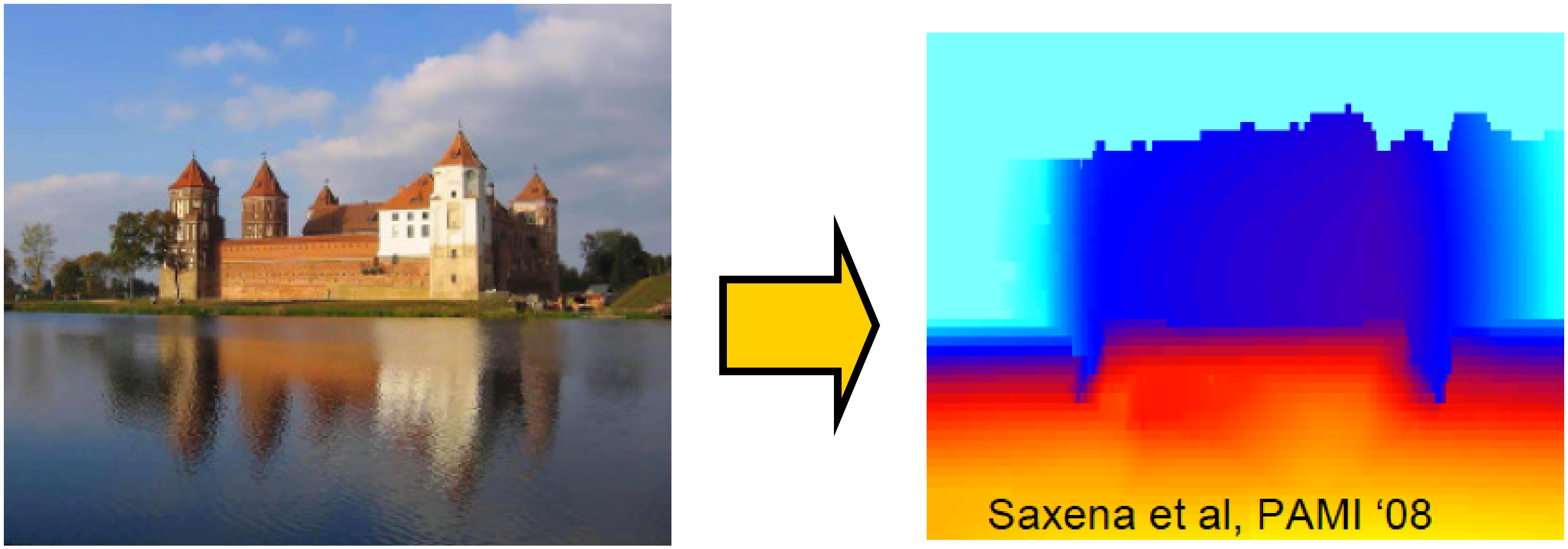


X_i: noisy pixels

Y_i: “true” pixels

12

Make3D

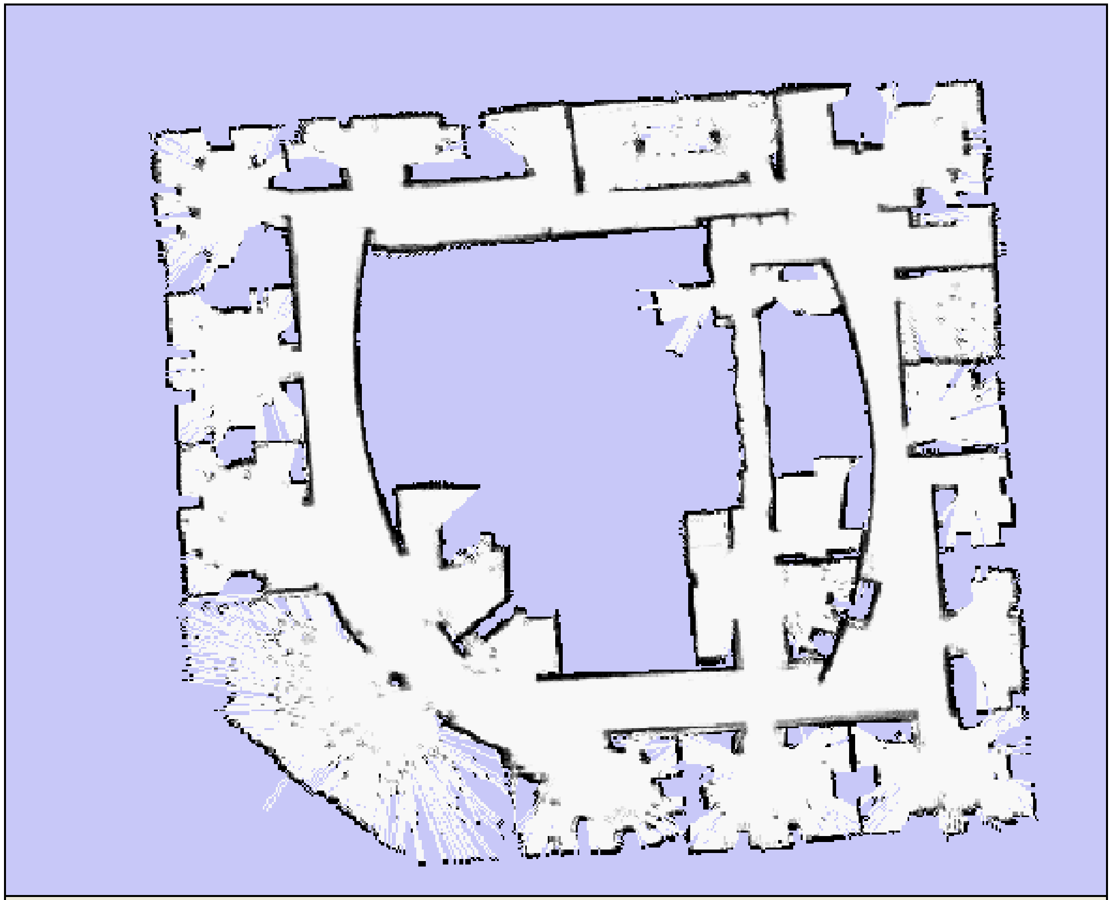


- Infer depth from 2D images
- “Conditional random fields”

Applications

State estimation

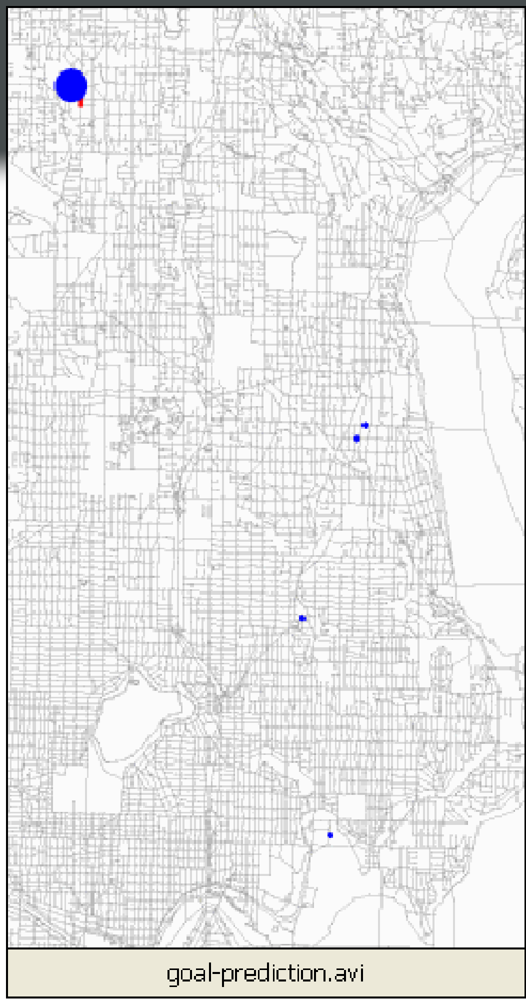
Robot localization & mapping



fastslam.avi

D. Haehnel,
W. Burgard,
D. Fox, and
S. Thrun.
IROS-03.

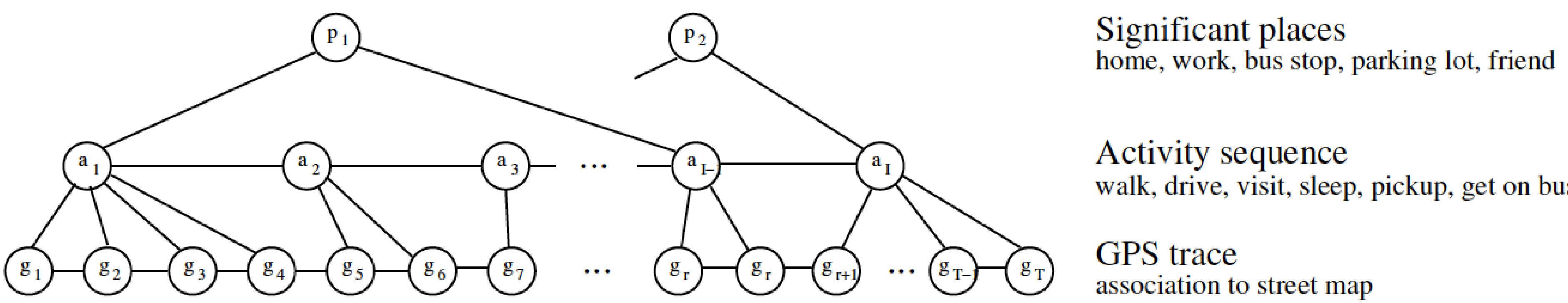
- Infer both location and map from noisy sensor data
- Particle filters



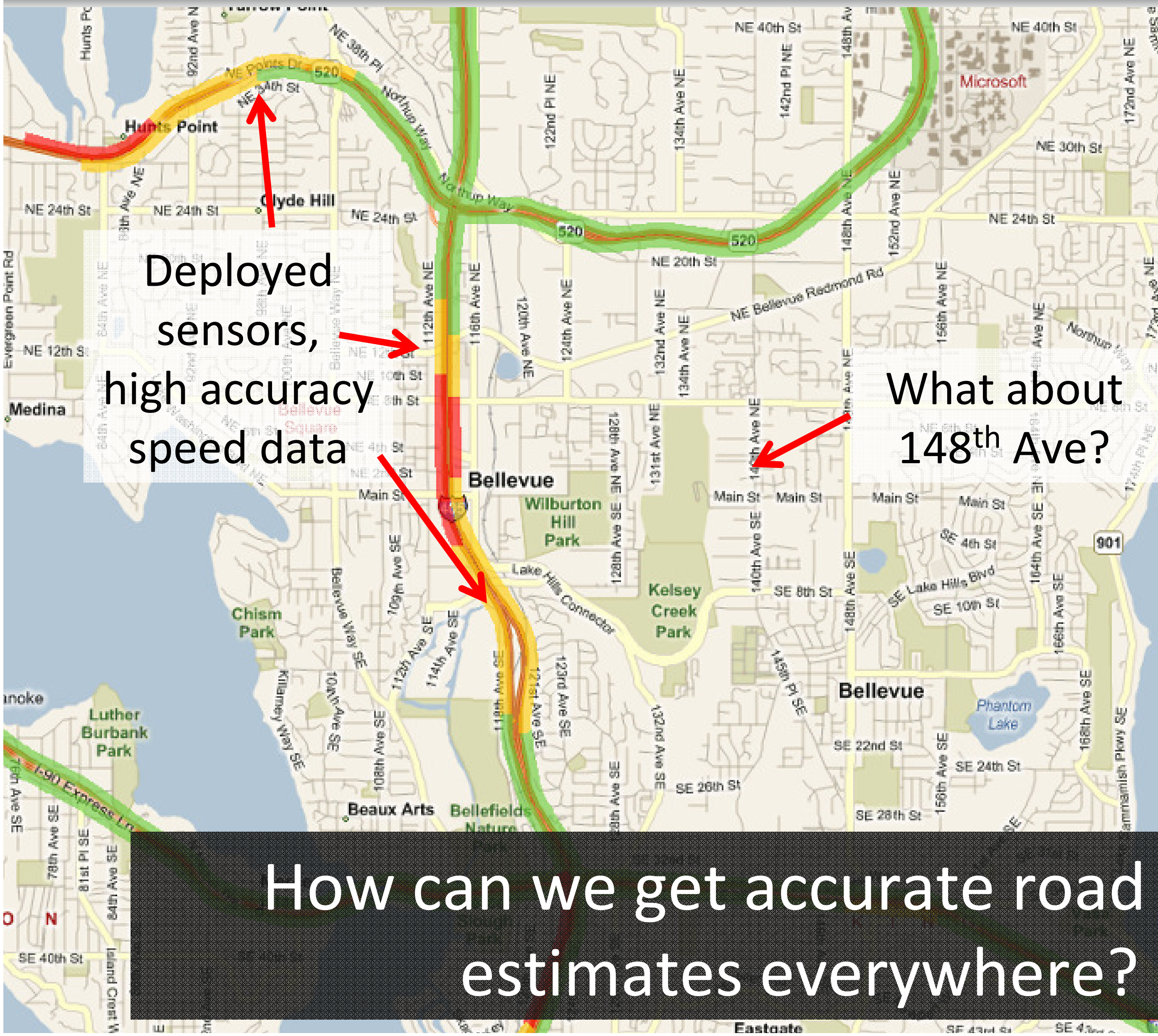
Activity recognition

L. Liao, D. Fox, and H. Kautz. *AAAI-04*

Predict “goals” from raw GPS data
“Hierarchical Dynamical Bayesian networks”



Traffic monitoring



Deployed sensors, high accuracy speed data

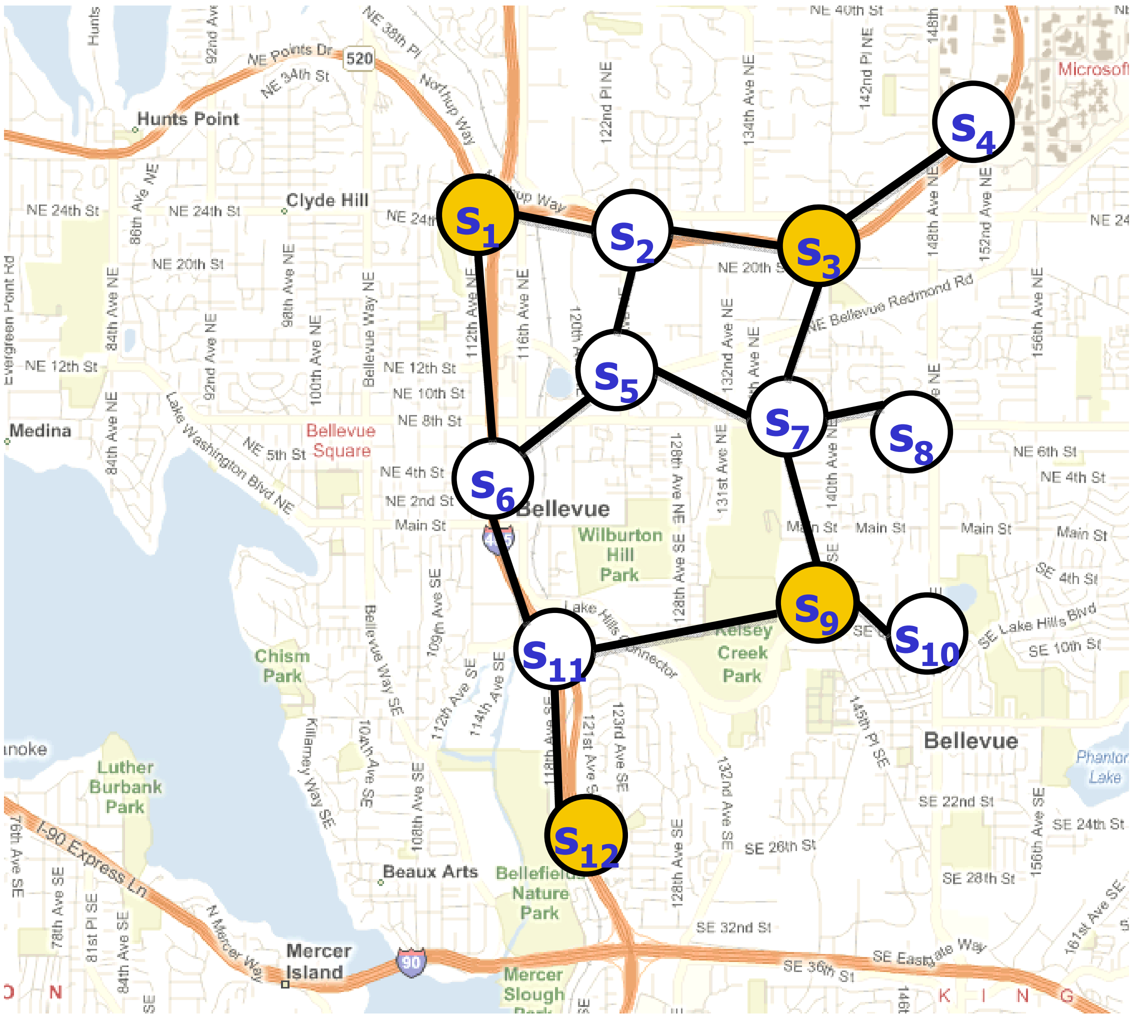
What about 148th Ave?

Detector loops

Traffic cameras

How can we get accurate road speed estimates everywhere?

Cars as a sensor network
[Krause, Horvitz et al.]

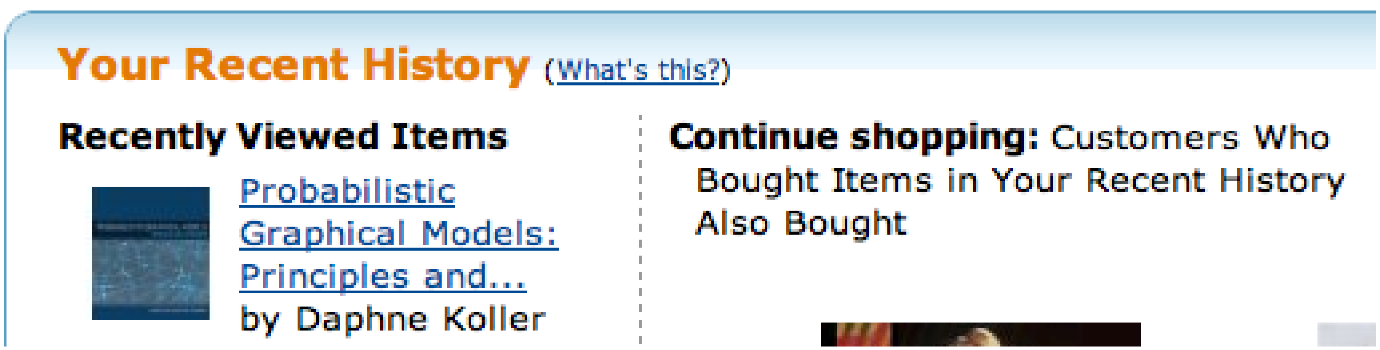


- (Normalized) speeds as random variables
- Joint distribution allows modeling correlations
- Can **predict unmonitored** speeds from monitored speeds using $\mathbf{P(S_5 \mid S_1, S_9)}$

Applications

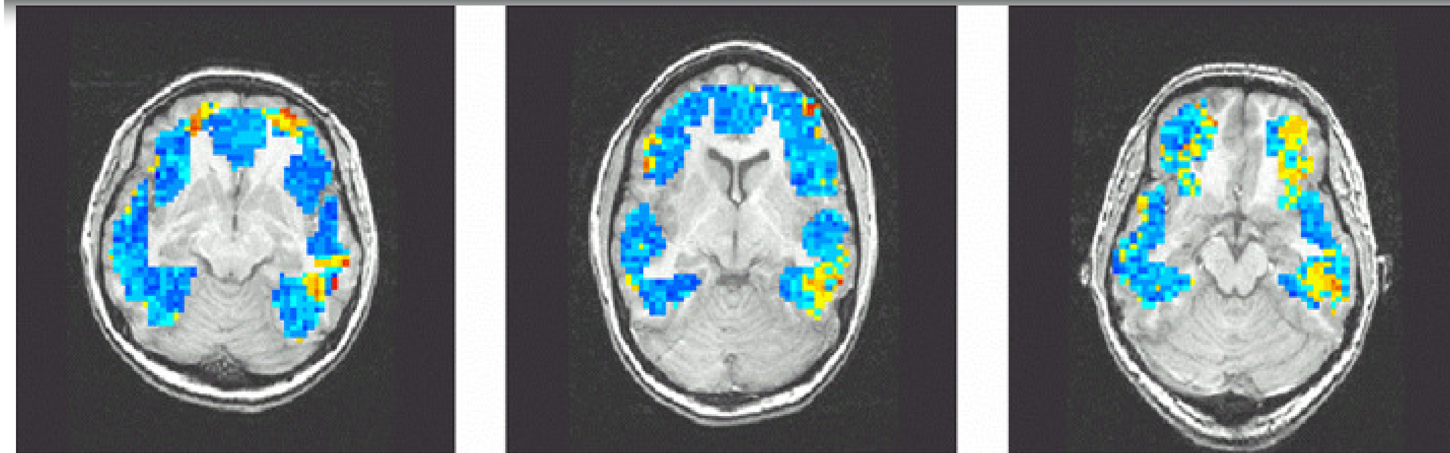
Structure Prediction

Collaborative Filtering and Link Prediction



- Predict "missing links", ratings...
- "Collective matrix factorization", Relational models

Analyzing fMRI data



Mitchell et al., *Science,* 2008

- Predict activation patterns for nouns
- Predict connectivity (Pittsburgh Brain Competition)

Other applications

- Coding (LDPC codes, …)
- Medical diagnosis
- Identifying gene regulatory networks
- Distributed control
- Computer music
- Probabilistic logic
- Graphical games
- ….

MANY MORE!!

Key challenges:

How do we

… **represent** such probabilistic models?

(distributions over vectors, maps, shapes, trees, graphs, functions…)

… perform **inference** in such models?

… **learn** such models from data?

Syllabus overview

- We will study Representation, Inference & Learning
- First in the simplest case
  - Only discrete variables
  - Fully observed models
  - Exact inference & learning
- Then generalize
  - Continuous distributions
  - Partially observed models (hidden variables)
  - Approximate inference & learning
- Learn about algorithms, theory & applications

Overview

- Course webpage
  - **http://www.cs.caltech.edu/courses/cs155/**
- Teaching assistant: Pete Trautman (trautman@cds.caltech.edu)
- Administrative assistant: Sheri Garcia (sheri@cs.caltech.edu)

Background & Prerequisites

- Basic probability and statistics
- Algorithms
- CS 156a or permission by instructor

- Please fill out the questionnaire about background (not graded ☺)

- Programming assignments in MATLAB.
- Do we need a MATLAB review recitation? No.

Coursework

- Grading based on
  - 4 homework assignments (one per topic) (40%)
  - Course project (40%)
  - Final take home exam (20%)
- 3 late days
- Discussing assignments allowed, but everybody must turn in their own solutions
- Start early! ☺ !!!

Course project

- “Get your hands dirty” with the course material
- Implement an algorithm from the course or a paper you read and apply it to some data set
- Ideas on the course website (soon)
- Application of techniques you learnt to your own research is encouraged
- Must be something new (e.g., not work done last term)

Project: Timeline and grading

- Small groups (2-3 students)
- October 19: Project proposals due (1-2 pages); feedback by instructor and TA
- November 9: Project milestone
- December 4: Project report due; poster session

- Grading based on quality of poster (20%), milestone report (20%) and final report (60%)

Review: Probability

- This should be familiar to you…

- Probability Space (Ω, F, P)
  - Ω: set of "atomic events"
  - F $\subseteq 2^{\Omega}$: set of all (non-atomic) events

    F is a σ-Algebra

    (closed under complements and countable unions)
  - P: F$\rightarrow$ [0,1] probability measure

    For $\omega \in$ F, P(ω) is the probability that event ω happens

$\alpha \in F$

$\Omega \setminus \alpha \in F$

Interpretation of probabilities

- Philosophical debate..
- Frequentist interpretation
  - P(α) is relative frequency of α in repeated experiments
  - Often difficult to assess with limited data
- Bayesian interpretation
  - P(α) is “degree of belief” that α will occur
  - Where does this belief come from?
  - Many different flavors (subjective, pragmatic, …)
- Most techniques in this class can be interpreted either way.

Independence of events

- Two events $\alpha, \beta \in$ F are independent if

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

- A collection S of events is independent, if for any subset $\alpha_1, \ldots, \alpha_n \in$ S it holds that

$$P(\alpha_1, \ldots \alpha_n) = P(\alpha_1) P(\alpha_2) \cdot \ldots \cdot P(\alpha_n)$$

$$\cancel{Q: \forall i, j : P(\alpha_i \cap \alpha_j) = P(\alpha_i) P(\alpha_j) \Rightarrow \alpha_1, \ldots \alpha_n \text{ independent?}}$$

No.

Conditional probability

- Let α, β be events, P(β)>0
- Then:

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

Most important rule #1:

- Let $\alpha_1,\ldots,\alpha_n$ be events, P(α_i)>0
- Then

$$P(\alpha_1 \cap \ldots \cap \alpha_n) = P(\alpha_1) \cdot P(\alpha_2 \mid \alpha_1) \cdot \ldots \cdot P(\alpha_n \mid \alpha_1 \ldots \alpha_{n-1})$$

Chain rule

Most important rule #2:

- Let α, β be events with prob. P(α) > 0, P(β) > 0
- Then

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

$$P(\beta) = P(\beta \mid \alpha) \cdot P(\alpha) + P(\beta \mid \neg\alpha) \cdot P(\neg\alpha)$$

Bayes 'rule

Random variables

- Events are cumbersome to work with.

- Let D be some set (e.g., the integers)
- A **random variable** X is a mapping $X: \Omega \rightarrow D$
- For some $x \in D$, we say
  $P(X = x) = P(\{\omega \in \Omega: X(\omega) = x\})$

"probability that variable X assumes state x"

- Notation: Val(X) = set D of all values assumed by X.

Examples

- Bernoulli distribution: "(biased) coin flips"

  $D = \{H,T\}$

  Specify $P(X = H) = p$. Then $P(X = T) = 1-p$.

  Write: $X \sim Ber(p)$;

- Multinomial distribution: "(biased) m-sided dice"

  $D = \{1,\ldots,m\}$

  Specify $P(X = i) = p_i$, s.t. $\sum_\iota p_i = 1$

  Write: $X \sim Mult(p_1,\ldots,p_m)$

Multivariate distributions

- Instead of random variable, have random vector

  $\mathbf{X}(\omega) = [X_1(\omega),\dots,X_n(\omega)]$

- Specify $P(X_1=x_1,\dots,X_n=x_n)$

- Suppose all X_i are Bernoulli variables.
- How many parameters do we need to specify?

X_1	X_2	$\dots$	X_n	$P(x_1 \dots x_n)$
0	0		0	?
0		0	1	?
0	.	0 1	0	

$2^n - 1$

Rules for random variables

- Chain rule

$$P(X_1 \ldots X_n) = P(X_1)\, P(X_2 \mid X_1) \cdot \ldots P(X_n \mid X_1 \ldots X_{n-1})$$

- Bayes' rule

$$P(X \mid Y) = \frac{P(Y \mid X)\, P(X)}{P(Y)}$$

How do we get $P(Y)$?

Marginal distributions

- Suppose, X and Y are RVs with distribution P(X,Y)

X : Intelligence
Y : Grade

Y \ X	VH	H
A	0.7	0.15
B	0.1	0.05

P(Grade = A) = .85

Marginal distributions

- Suppose we have joint distribution $P(X_1,\ldots,X_n)$
- Then

$$P(X_i = x_i) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} P(x_1 \ldots x_n)$$

- If all X_i binary: How many terms? 2^{n-1}

Independent RVs

- What if RVs are independent?

  RVs $X_1,\ldots,X_n$ are independent, if for any assignment

  $$P(X_1=x_1,\ldots,X_n=x_n) = P(x_1)\,P(x_2)\,\ldots\,P(x_n)$$

$\Leftrightarrow$ $\{\omega : X_i(\omega) = x_i\}$ $\forall i,\ x_i \in Val(X_i)$ indep.

- How many parameters are needed in this case? $\underline{\underline{n}}$

$n \ll 2^n$

X_i, X_j indep $\Rightarrow$ $P(X_i \mid X_j) = P(X_i)$

- Independence too strong assumption… Is there something weaker?

Key concept: Conditional independence

- Events α, β conditionally independent given γ if

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

- Random variables X and Y cond. indep. given Z if for all x$\in$ Val(X), y$\in$ Val(Y), Z$\in$ Val(Z)

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)\, P(Y = y \mid Z = z)$$

- If P(Y=y |Z=z)>0, that's equivalent to

$$P(X = x \mid Z = z, Y = y) = P(X = x \mid Z = z)$$

Similarly for sets of random variables **X**, **Y**, **Z**

We write: P $\models$ **X** $\perp$ **Y** | **Z**

Why is conditional independence useful?

- $P(X_1,\ldots,X_n) = P(X_1)\, P(X_2 \mid X_1) \ldots \underbrace{P(X_n \mid X_1,\ldots,X_{n-1})}_{2^{n-1}}$

  How many parameters?

  $2^0 + 2^1 + 2^2 + \ldots + 2^{n-1} = 2^n - 1$

- Now suppose $X_1 \ldots X_{i-1} \perp X_{i+1} \ldots X_n \mid X_i$ for all i

  Then

  $P(X_1,\ldots,X_n) = \underbrace{P(X_1)}_{1} \cdot \underbrace{P(X_2 \mid X_1)}_{2} \cdot \underbrace{P(X_3 \mid X_2)}_{2} \cdot \ldots \cdot \underbrace{P(X_n \mid X_{n-1})}_{2}$

  $2n-1 \ll 2^n$

  How many parameters? Exponential reduction in # params

- Can we compute $P(X_n)$ more efficiently? Yes (often)

Properties of Conditional Independence

- **Symmetry**
  - $X \perp Y \mid Z \Rightarrow Y \perp X \mid Z$
- **Decomposition**
  - $X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z$
- **Contraction** "Inverse Decomposition"
  - $(X \perp Y \mid Z) \wedge (X \perp W \mid Y,Z) \Rightarrow X \perp Y,W \mid Z$
- **Weak union**
  - $X \perp Y,W \mid Z \Rightarrow X \perp Y \mid Z,W$
- **Intersection**
  - $(X \perp Y \mid Z,W) \wedge (X \perp W \mid Y,Z) \Rightarrow X \perp Y,W \mid Z$
  - Holds only if distribution is positive, i.e., P>0

Key questions

- How do we specify distributions that satisfy particular independence properties?

  ➔ **Representation**

- How can we exploit independence properties for efficient computation?

  ➔ **Inference**

- How can we identify independence properties present in data?

  ➔ **Learning**

Will now see examples: Bayesian Networks

Bayesian networks

- A powerful class of probabilistic graphical models
- Compact parametrization of high-dimensional distributions
- In many cases, efficient exact inference possible
- Many applications
  - Natural language processing
  - State estimation
  - Link prediction
  - …
- Demo..

Key idea

- Conditional parametrization
  (instead of joint parametrization)
- For each RV, specify $P(X_i \mid \mathbf{X_A})$ for set $\mathbf{X_A}$ of RVs
- Then use chain rule to get joint parametrization

- Have to be careful to guarantee legal distribution…

Example: 2 variables

$$\begin{tikzcd} I \arrow[d] \\ G \end{tikzcd}$$

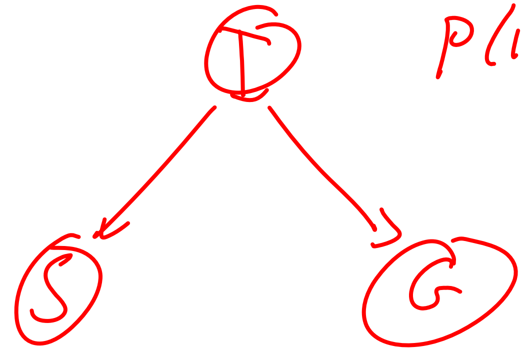
$P(I = VH) = 0.8$

$P(G | I)$

I \ G	A	B	
VH	0.8	0.2	$\Sigma = 1$
H	0.6	0.4	$\Sigma = 1$

Example: 3 variables

$P(I)$



$P(S|I)$ $P(G|I)$

$$P(I,S,G) = P(I)\,P(G|I)\,P(S|I)$$

Example: Naïve Bayes models

- Class variable Y
- Evidence variables $X_1,\ldots,X_n$
- Assume that $X_A \perp X_B \mid Y$
  for all subsets X_A, X_B of $\{X_1,\ldots,X_n\}$

- Conditional parametrization:
  - Specify P(Y)
  - Specify $P(X_i \mid Y)$
- Joint distribution

$$P(X_1 \ldots X_n, Y) = P(Y) \prod_i P(X_i \mid Y)$$

What you need to know

- Basic probability
- Independence and conditional independence
- Chain rule & Bayes' rule
- Naïve Bayes models

Tasks

- By tomorrow (October 1, 4pm): hand in questionnaire about background to Sheri Garcia
- Read Chapter 2 in Koller & Friedman
- Start thinking about project teams and ideas (proposals due October 19)