Inference and Representation

David Sontag

New York University

Lecture 1, September 8, 2015

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

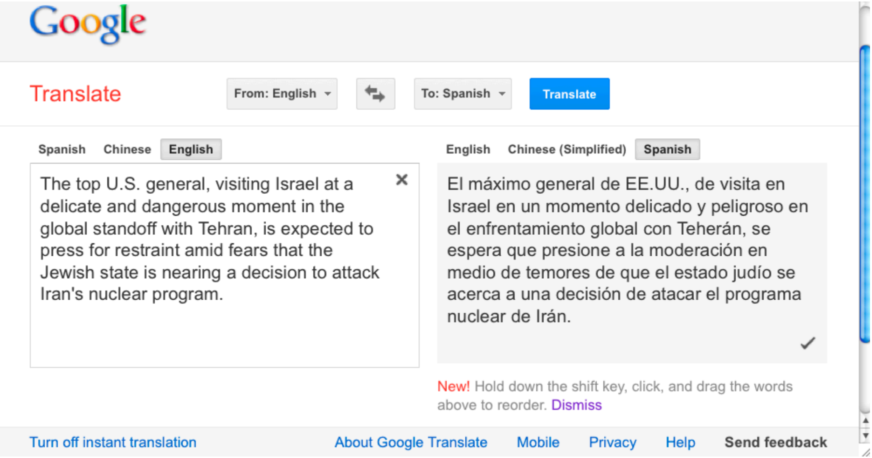
1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
2. **Learn** the distribution from data
3. Perform "**inference**" (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)

Reasoning under uncertainty

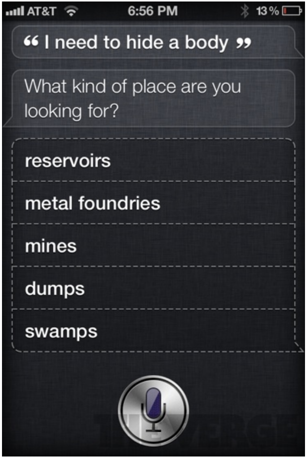
- As humans, we are continuously making predictions under uncertainty
- Classical AI and ML research ignored this phenomena
- Many of the most recent advances in technology are possible because of this new, *probabilistic*, approach

Applications: Deep question answering

Applications: Machine translation

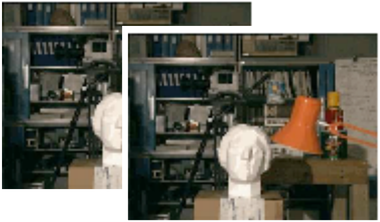


Applications: Speech recognition



Applications: Stereo vision

input: two images



output: disparity

Key challenges

1. **Represent** the world as a collection of random variables $X_1, \ldots, X_n$ with joint distribution $p(X_1, \ldots, X_n)$
   - How does one *compactly describe* this joint distribution?
   - Directed graphical models (Bayesian networks)
   - Undirected graphical models (Markov random fields, factor graphs)
2. **Learn** the distribution from data
   - Maximum likelihood estimation. Other estimation methods?
   - How much data do we need?
   - How much computation does it take?
3. Perform "**inference**" (compute conditional distributions $p(X_i \mid X_1 = x_1, \ldots, X_m = x_m)$)

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 variables
  - Partially observed data during learning (hidden variables)
  - *Approximate* inference & learning
- Learn about algorithms, theory & applications

Logistics: class

- **Class webpage:**
  - `http://cs.nyu.edu/~dsontag/courses/inference15/`
  - Sign up for Piazza!
- **Book:** *Machine Learning: a Probabilistic Perspective* by Kevin Murphy, MIT Press (2012)
  - Required readings for each lecture posted to course website.
  - A good optional reference is *Probabilistic Graphical Models: Principles and Techniques* by Daphne Koller and Nir Friedman, MIT Press (2009)
- **Office hours:** Thurs 3:30-4:30pm. 715 Broadway, 12th floor, #1204 (**except for** 9/22, 11/3, 11/10, 11/24, 12/8: Tues 10:30-11:30am)
- **Lab:** Wednesdays, 7:10-8:00pm in WWH 102 (same as class)
  - Instructor: Rachel Hodos (hodos@cims.nyu.edu)
  - Required attendance; no exceptions.
- **Grader:** Prasoon Goyal (pg1338@nyu.edu)

Logistics: prerequisites & grading

- **Prerequisite:**
  - DS-GA-1003/CSCI-GA.2567 (Machine Learning and Computational Statistics)
- **Grading:** *not finalized* problem sets (55%) + in class midterm exam (20%) + in class final exam (20%) + participation (5%)
  - I would love to see an active Piazza with students asking & responding to each other's questions. Will contribute to your participation grade.
  - 6 assignments (every 1–2 weeks). Both theory and programming.
  - First homework out **tomorrow**, due Friday Sept. 18 at 5pm (via email)
  - **Important:** See collaboration policy on class webpage
- Solutions to the theoretical questions must be rigorous.
- For the programming assignments, I recommend Python.

Example: Medical diagnosis

- Variable for each **symptom** (e.g. "fever", "cough", "fast breathing", "shaking", "nausea", "vomiting")
- Variable for each **disease** (e.g. "pneumonia", "flu", "common cold", "bronchitis", "tuberculosis")
- Diagnosis is performed by **inference** in the model:

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

- One famous model, Quick Medical Reference (QMR-DT), has 600 diseases and 4000 findings

Representing the distribution

- Naively, could represent multivariate distributions with table of probabilities for each outcome (assignment)
- How many outcomes are there in QMR-DT? 2^{4600}
- **Estimation** of joint distribution would require a huge amount of data
- **Inference** of conditional probabilities, e.g.

$$p(\text{pneumonia} = 1 \mid \text{cough} = 1, \text{fever} = 1, \text{vomiting} = 0)$$

  would require summing over exponentially many variables' values
- Moreover, defeats the purpose of probabilistic modeling, which is to make predictions with *previously unseen observations*

Structure through independence

- If $X_1, \ldots, X_n$ are independent, then

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2)\cdots p(x_n)$$

- 2^n entries can be described by just n numbers (if $|\mathrm{Val}(X_i)| = 2$)!
- However, this is not a very *useful* model – observing a variable X_i cannot influence our predictions of X_j
- If $X_1, \ldots, X_n$ are *conditionally independent* given Y, denoted as $X_i \perp \mathbf{X}_{-i} \mid Y$, then

$$\begin{aligned} p(y, x_1, \ldots, x_n) &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid x_1, \ldots, x_{i-1}, y) \\ &= p(y)p(x_1 \mid y)\prod_{i=2}^{n} p(x_i \mid y). \end{aligned}$$

- This is a simple, yet *powerful*, model

Example: naive Bayes for classification

- Classify e-mails as spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $1 : n$ index the words in our vocabulary (e.g., English)
  - $X_i = 1$ if word i appears in an e-mail, and 0 otherwise
  - E-mails are drawn according to some distribution $p(Y, X_1, \ldots, X_n)$
- Suppose that the words are conditionally independent given Y. Then,

$$p(y, x_1, \ldots x_n) = p(y) \prod_{i=1}^{n} p(x_i \mid y)$$

Estimate the model with maximum likelihood. **Predict** with:

$$p(Y = 1 \mid x_1, \ldots x_n) = \frac{p(Y = 1) \prod_{i=1}^{n} p(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} p(Y = y) \prod_{i=1}^{n} p(x_i \mid Y = y)}$$

- Are the independence assumptions made here reasonable?
- Philosophy: Nearly all probabilistic models are “wrong”, but many are nonetheless useful

Bayesian networks

Reference: Chapter 10

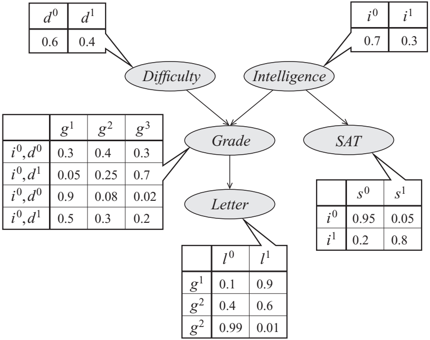
- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

- Powerful framework for designing *algorithms* to perform probability computations
- Enables use of *prior knowledge* to specify (part of) model structure

Example

- Consider the following Bayesian network:



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^0, d^0	0.9	0.08	0.02
i^0, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^2	0.99	0.01

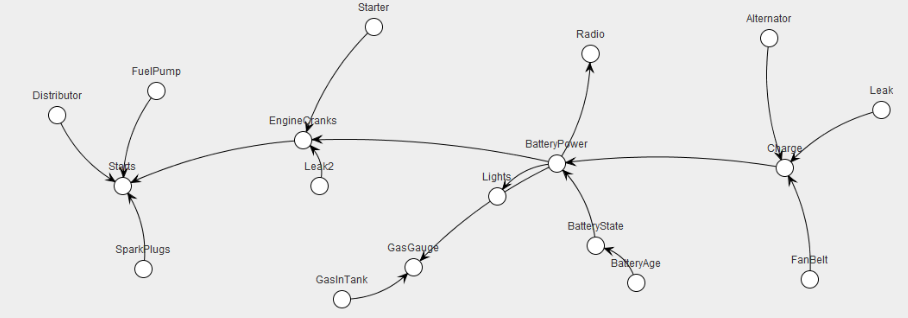
- What is its joint distribution?

$$
\begin{aligned}
p(x_1, \ldots x_n) &= \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)}) \\
p(d, i, g, s, l) &= p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)
\end{aligned}
$$

More examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

Will my car start this morning?



Heckerman *et al.*, Decision-Theoretic Troubleshooting, 1995

More examples

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

What is the differential diagnosis?

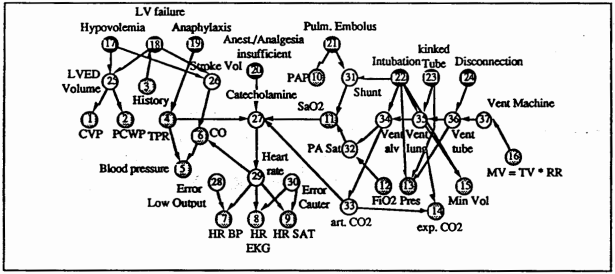
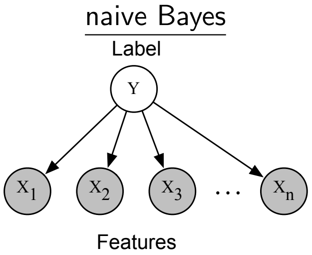


Fig. 1 The ALARM network representing causal relationships is shown with diagnostic (●), intermediate (○) and measurement (◎) nodes. CO: cardiac output, CVP: central venous pressure, LVED volume: left ventricular end-diastolic volume, LV failure: left ventricular failure, MV: minute ventilation, PA Sat: pulmonary artery oxygen saturation, PAP: pulmonary artery pressure, PCWP: pulmonary capillary wedge pressure, Pres: breathing pressure, RR: respiratory rate, TPR: total peripheral resistance, TV: tidal volume

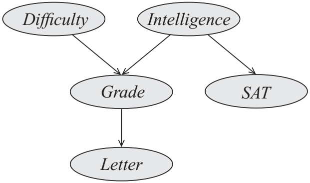
Beinlich *et al.*, The ALARM Monitoring System, 1989

Bayesian networks are *generative models*



- Evidence is denoted by shading in a node
- Can interpret Bayesian network as a **generative process**. For example, to *generate* an e-mail, we
  1. Decide whether it is spam or not spam, by samping $y \sim p(Y)$
  2. For each word $i = 1$ to n, sample $x_i \sim p(X_i \mid Y = y)$

Bayesian network structure implies conditional independencies!



- The joint distribution corresponding to the above BN factors as

$$p(d, i, g, s, l) = p(d)p(i)p(g \mid i, d)p(s \mid i)p(l \mid g)$$

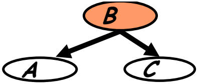
- However, by the chain rule, *any* distribution can be written as

$$p(d, i, g, s, l) = p(d)p(i \mid d)p(g \mid i, d)p(s \mid i, d, g)p(l \mid g, d, i, g, s)$$

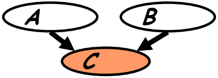
- Thus, we are assuming the following additional independencies:
  $D \perp I, \qquad S \perp \{D, G\} \mid I, \qquad L \perp \{I, D, S\} \mid G.$ What else?

Bayesian network structure implies conditional independencies!

- Generalizing the above arguments, we obtain that a variable is independent from its non-descendants given its parents

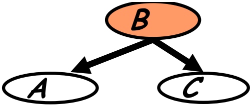


- **Common parent** – fixing B *decouples* A and C
- **Cascade** – knowing B *decouples* A and C



- **V-structure** – Knowing C *couples* A and B
  - This important phenomona is called **explaining away** and is what makes Bayesian networks so powerful

A simple justification (for common parent)



We'll show that $p(A, C \mid B) = p(A \mid B)p(C \mid B)$ for *any* distribution $p(A, B, C)$ that factors according to this graph structure, i.e.

$$p(A, B, C) = p(B)p(A \mid B)p(C \mid B)$$

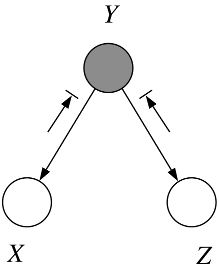
Proof.

$$p(A, C \mid B) = \frac{p(A, B, C)}{p(B)} = p(A \mid B)p(C \mid B)$$

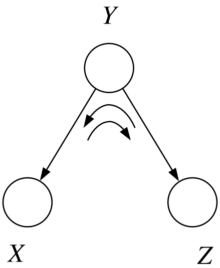
□

D-separation ("directed separated") in Bayesian networks

- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



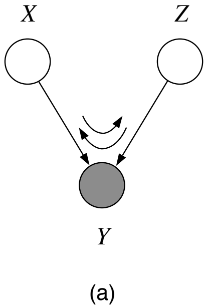
(a)

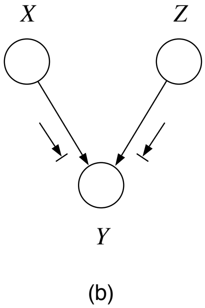


(b)

D-separation ("directed separated") in Bayesian networks

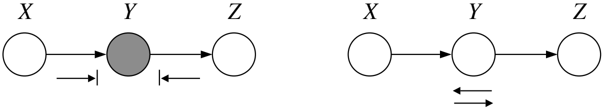
- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:





D-separation ("directed separated") in Bayesian networks

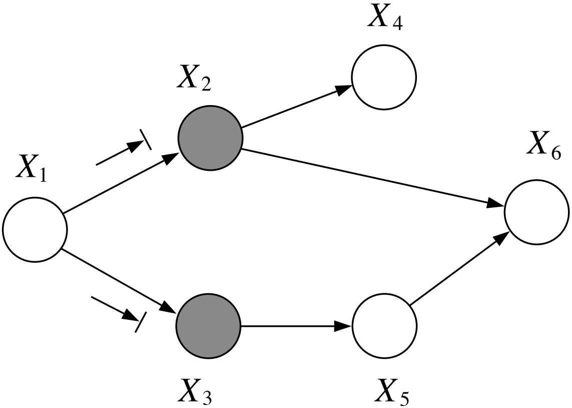
- Algorithm to calculate whether $X \perp Z \mid \mathbf{Y}$ by looking at graph separation
- Look to see if there is **active path** between X and Z when variables $\mathbf{Y}$ are observed:



(a) (b)

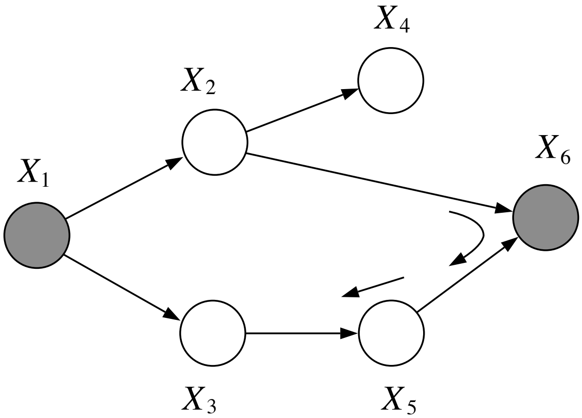
- If no such path, then X and Z are **d-separated** with respect to $\mathbf{Y}$
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy)
- Important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query

D-separation example 1



Is $X_6 \perp X_5 \mid X_2, X_3$? Is $X_4 \perp X_5 \mid X_2, X_3$?

D-separation example 2



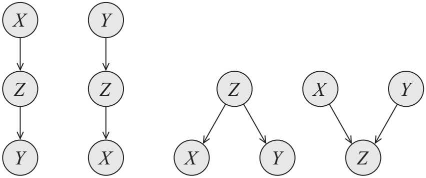
Is $X_4 \perp X_5 \mid X_1, X_6$? What about if X_6 is not observed? I.e., is $X_4 \perp X_5 \mid X_1$?

Independence maps

- Let $I(G)$ be the set of all conditional independencies implied by the directed acyclic graph (DAG) G
- Let $I(p)$ denote the set of all conditional independencies that hold for the joint distribution p.
- A DAG G is an **I-map** (independence map) of a distribution p if $I(G) \subseteq I(p)$
  - A fully connected DAG G is an I-map for *any* distribution, since $I(G) = \emptyset \subseteq I(p)$ for all p
- G is a **minimal I-map** for p if the removal of even a single edge makes it not an I-map
  - A distribution may have several minimal I-maps
  - Each corresponds to a specific node-ordering
- G is a **perfect map** (P-map) for distribution p if $I(G) = I(p)$

Equivalent structures

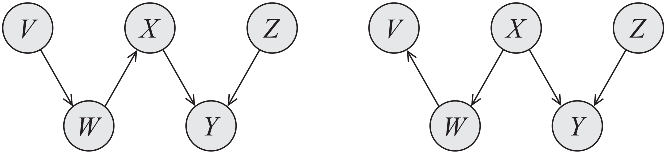
- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Which of these are equivalent?



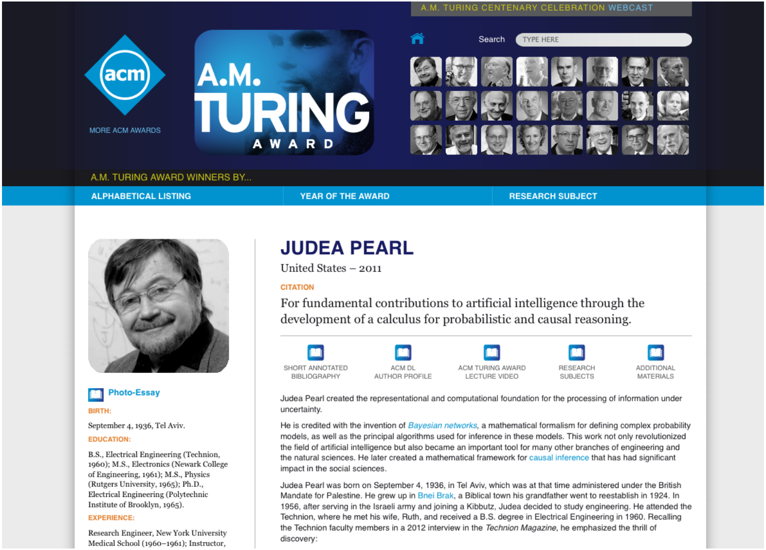
(a) (b) (c) (d)

Equivalent structures

- Different Bayesian network structures can be **equivalent** in that they encode precisely the same conditional independence assertions (and thus the same distributions)
- Are these equivalent?



2011 Turing Award was for Bayesian networks



A.M. TURING CENTENARY CELEBRATION WEBCAST

MORE ACM AWARDS

A.M. TURING AWARD

Search TYPE HERE

A.M. TURING AWARD WINNERS BY...

ALPHABETICAL LISTING | YEAR OF THE AWARD | RESEARCH SUBJECT

Photo-Essay

BIRTH:

September 4, 1936, Tel Aviv.

EDUCATION:

B.S., Electrical Engineering (Technion, 1960); M.S., Electronics (Newark College of Engineering, 1961); M.S., Physics (Rutgers University, 1965); Ph.D., Electrical Engineering (Polytechnic Institute of Brooklyn, 1965).

EXPERIENCE:

Research Engineer, New York University Medical School (1960–1961); Instructor,

JUDEA PEARL

United States – 2011

CITATION

For fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

SHORT ANNOTATED BIBLIOGRAPHY | ACM DL AUTHOR PROFILE | ACM TURING AWARD LECTURE VIDEO | RESEARCH SUBJECTS | ADDITIONAL MATERIALS

Judea Pearl created the representational and computational foundation for the processing of information under uncertainty.

He is credited with the invention of *Bayesian networks*, a mathematical formalism for defining complex probability models, as well as the principal algorithms used for inference in these models. This work not only revolutionized the field of artificial intelligence but also became an important tool for many other branches of engineering and the natural sciences. He later created a mathematical framework for causal inference that has had significant impact in the social sciences.

Judea Pearl was born on September 4, 1936, in Tel Aviv, which was at that time administered under the British Mandate for Palestine. He grew up in Bnei Brak, a Biblical town his grandfather went to reestablish in 1924. In 1956, after serving in the Israeli army and joining a Kibbutz, Judea decided to study engineering. He attended the Technion, where he met his wife, Ruth, and received a B.S. degree in Electrical Engineering in 1960. Recalling the Technion faculty members in a 2012 interview in the *Technion Magazine*, he emphasized the thrill of discovery:

Summary

- **Bayesian networks** given by (G, P) where P is specified as a set of local **conditional probability distributions** associated with G's nodes
- One interpretation of a BN is as a **generative model**, where variables are sampled in topological order
- Local and global independence properties identifiable via **d-separation** criteria
- Computing the probability of any assignment is obtained by multiplying CPDs
  - **Bayes' rule** is used to compute conditional probabilities
  - Marginalization or **inference** is often computationally difficult