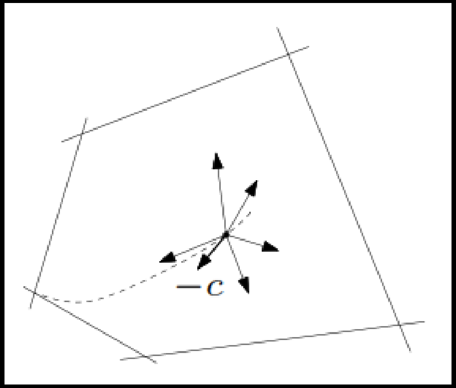
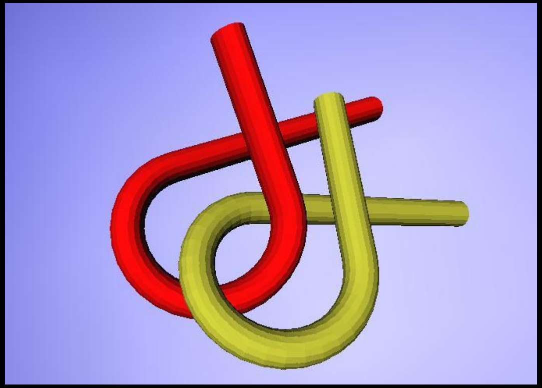
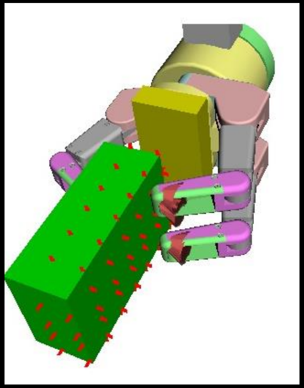
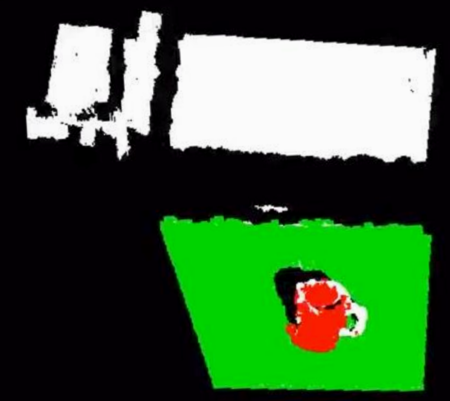
Probability and Bayes Nets

Using material from probabilistic-robotics.org and Russell+Norvig AI lectures

Previously…

- Everything was known and deterministic

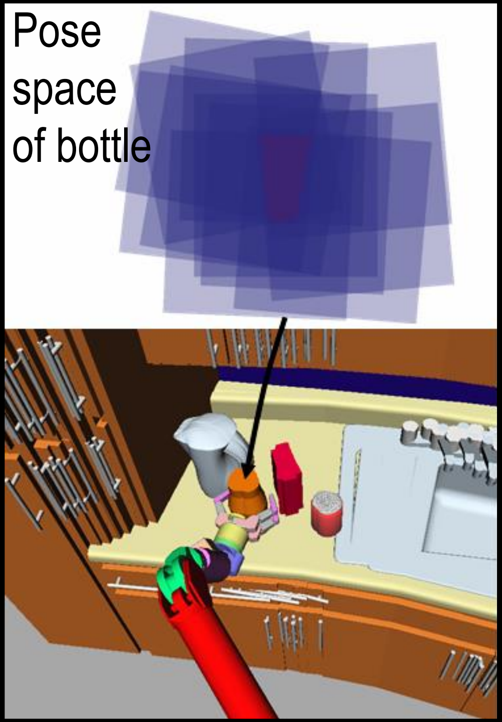


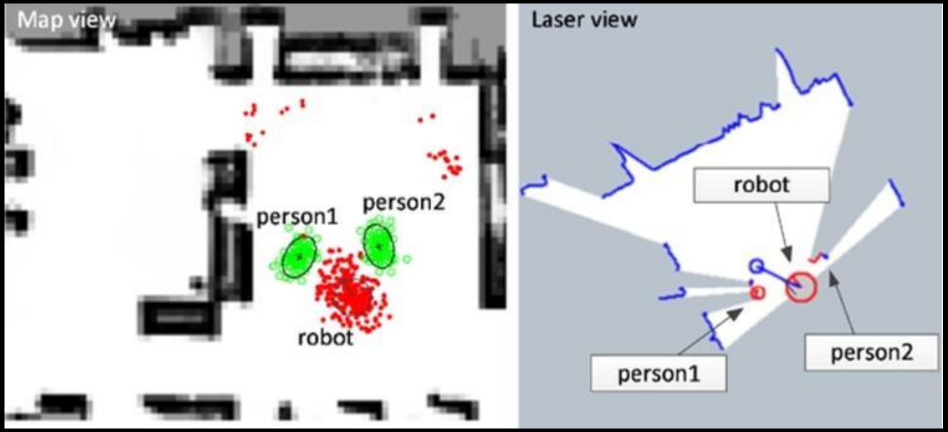
- But the real world is uncertain!
  - Or at least there is some information we don't know
- How do we quantify and reason about this uncertainty?

Uncertainty in Robotics

- (Almost) everything is uncertain in robotics!
  - Sensor data is never perfectly accurate
- Reasoning about uncertainty is essential for processing sensor data
- Actions can have uncertain results
  - E.g. for a car "drive forward" will have different results depending on friction between tire and ground (which is unknown)



Uncertainty in object pose in manipulation



Localization uncertainty for mobile robot and people

Outline

- Probability basics
- Bayes Rule
- Independence and Conditional Independence
- Bayes nets

Probability basics

Begin with a set Ω—the sample space

e.g., 6 possible rolls of a die.

$\omega \in \Omega$ is a sample point/possible world/atomic event

A probability space or probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.

$0 \leq P(\omega) \leq 1$

$\Sigma_\omega P(\omega) = 1$

e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.

An event A is any subset of Ω

$$P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

<u>Discrete</u> Random Variables

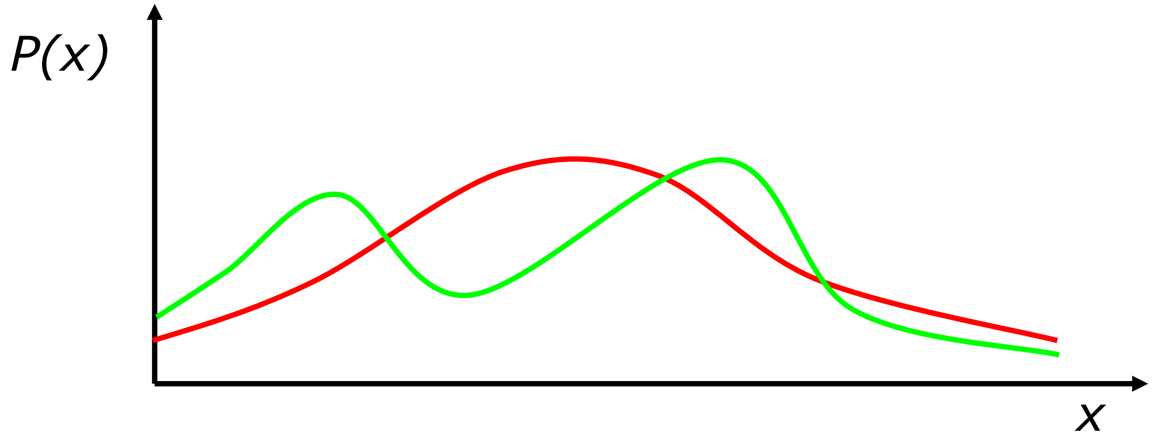
- X denotes a random variable.
- X can take on a countable number of values in $\{x_1, x_2, \ldots, x_n\}$.
- $P(X=x_i)$, or $P(x_i)$, is the probability that the random variable X takes on value x_i.
- $P(\ldots)$ is called the probability **mass** function.
- E.g. $$P(Room) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

Continuous Random Variables

- X takes on values in the continuum.
- $P(X=x)$, or $P(x)$, is a probability **density** function.

$$P(x \in (a,b)) = \int_a^b P(x)dx$$

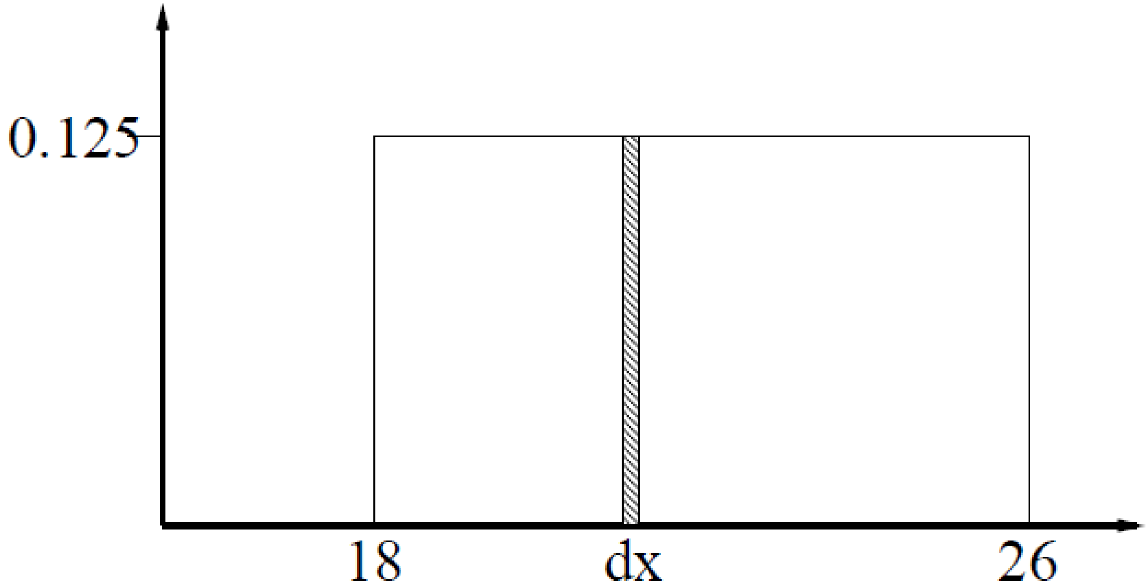
- E.g.



Probability for continuous variables

Express distribution as a parameterized function of value:

$P(X = x) = U[18, 26](x) =$ uniform density between 18 and 26



Here P is a density; integrates to 1.
$P(X = 20.5) = 0.125$ really means

$$\lim_{dx \to 0} P(20.5 \leq X \leq 20.5 + dx)/dx = 0.125$$

Propositions

Think of a proposition as the event (set of sample points) where the proposition is true

{true, false}

Given Boolean random variables A and B:

event a = set of sample points where $A(\omega) = true$

event $\neg a$ = set of sample points where $A(\omega) = false$

event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$

i.e. "or"

Proposition = disjunction of atomic events in which it is true

e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$

$\Rightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Logic symbol reminder: $\wedge$ means "and"
$\vee$ means "or"
$\neg$ means "not"

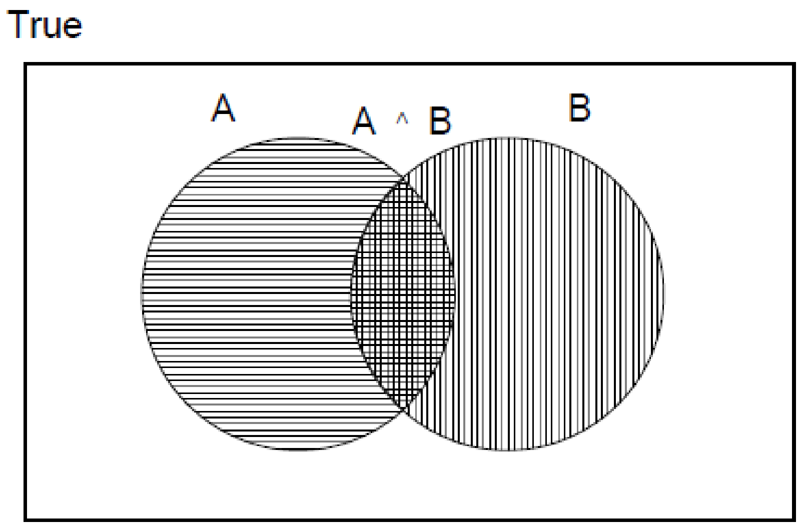
Axioms of Probability Theory

- $0 \le P(a) \le 1$
- $P(true) = 1 \qquad P(false) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Why use probability?

The definitions imply that certain logically related events must have related probabilities

E.g., $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$



de Finetti (1931): an agent who bets according to probabilities that violate these axioms can be forced to bet so as to lose money regardless of outcome.

Joint and Conditional Probability

- $P(X=x \text{ and } Y=y) = P(x \wedge y) = P(x,y)$ (joint probability)
- If X and Y are independent then

$$P(x,y) = P(x)P(y)$$

- $P(x \mid y)$ is the probability of x given y (conditional probability)

$$P(x \mid y) = P(x,y) \,/\, P(y)$$

$$P(x,y) = P(x \mid y)\, P(y)$$

- If X and Y are independent then

$$P(x \mid y) = P(x)$$

Law of Total Probability

Discrete variables

$$\sum_x P(x) = 1$$

$$P(x) = \sum_y P(x, y)$$

$$P(x) = \sum_y P(x \mid y) P(y)$$

Continuous variables

$$\int P(x)\, dx = 1$$

$$P(x) = \int P(x, y)\, dy$$

$$P(x) = \int P(x \mid y) P(y)\, dy$$

Conditioning

- Law of total probability:

$$P(x) = \int P(x, z)dz$$

$$P(x) = \int P(x \mid z)P(z)dz$$

$$P(x \mid y) = \int P(x \mid y, z)\, P(z \mid y)\, dz$$

Bayes Rule

Product rule:

$$P(x,y) = P(x|y)P(y)$$

$$\|\qquad\qquad\qquad\|$$

$$P(y,x) = P(y|x)P(x)$$

Bayes rule:

$$\boxed{P(x|y) = \frac{P(y|x)\ P(x)}{P(y)} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}}$$

Bayes Rule with more conditioning

$$P(x \mid y, z) = \frac{P(y \mid x, z)\ P(x \mid z)}{P(y \mid z)}$$

Normalization

- Back to Bayes Rule (for distributions):

$$P(X \mid Y) = \frac{P(Y \mid X)\ P(X)}{\boxed{P(Y)}} = \frac{\text{likelihood} \cdot \text{prior}}{\text{evidence}}$$

- Often *P(Y=y)* (the probability of the given evidence) is difficult to compute
- Normalization trick:
  - Compute $P(y|x_i)P(x_i)$ for every $X = x_i$, put results in a vector and normalize the vector
  - This is often written as:

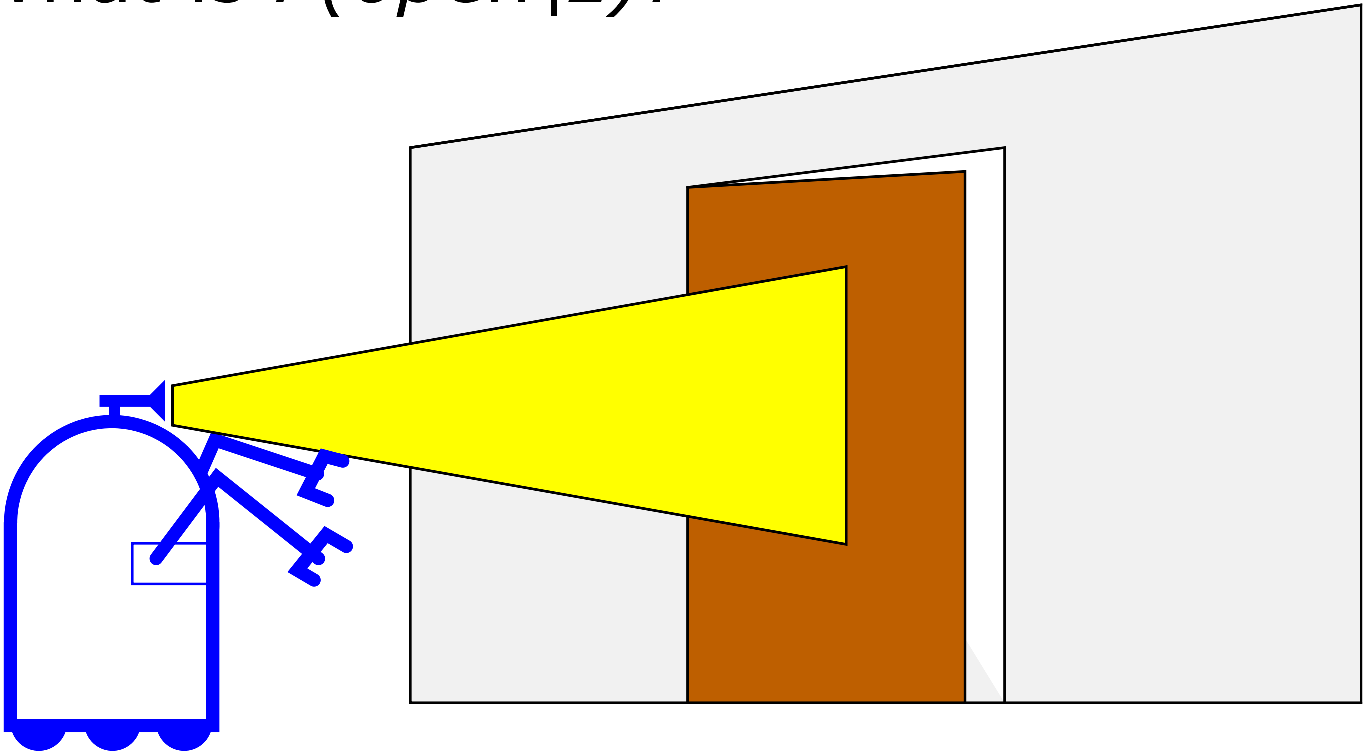
$$P(X \mid y) = \frac{P(y \mid X)\ P(X)}{P(y)} = \eta\, P(y \mid X)\, P(X)$$

$$\eta = P(y)^{-1} = \frac{1}{\sum_{x \in X} P(y \mid x) P(x)}$$

You don't actually compute η in practice, it just means normalize the results!

Bayes Rule State Estimation Example

- A robot wants to determine the probability of a door being open
- Suppose the robot obtains measurement z
- What is $P(open|z)$?



Causal vs. Diagnostic Reasoning

- *P(open|z)* is diagnostic.
- *P(z|open)* is causal.
- Often causal knowledge is easier to obtain and reason about.
- Bayes rule allows us to use causal knowledge:

$$P(open \mid z) = \frac{P(z \mid open) P(open)}{P(z)}$$

State Estimation Example

- Prior probability: $P(open) = P(\neg open) = 0.5$
- Assume sensor model says: $P(z|open) = 0.6 \;\; P(z|\neg open) = 0.3$

$$P(open|z) = \frac{P(z|open)P(open)}{P(z|open)P(open) + P(z|\neg open)P(\neg open)}$$

$$P(open|z) = \frac{0.6 \cdot 0.5}{0.6 \cdot 0.5 + 0.3 \cdot 0.5} = \frac{2}{3} = 0.67$$

- *z* raises the probability that the door is open (was 0.5, now is 0.67)

Combining Evidence

- Suppose our robot obtains another observation z_2.
- How can we integrate this new information?
- More generally, how can we estimate $P(x | z_1 ... z_n)$?

Recursive Bayesian Updating

$$P(x \mid z_1,\ldots,z_n) = \frac{P(z_n \mid x, z_1,\ldots,z_{n-1})\, P(x \mid z_1,\ldots,z_{n-1})}{P(z_n \mid z_1,\ldots,z_{n-1})}$$

Markov assumption: z_n is independent of $z_1,\ldots,z_{n-1}$ if we know x.

$$P(x \mid z_1,\ldots,z_n) = \frac{P(z_n \mid x)\, P(x \mid z_1,\ldots,z_{n-1})}{P(z_n \mid z_1,\ldots,z_{n-1})}$$

$$= \eta\, P(z_n \mid x)\, P(x \mid z_1,\ldots,z_{n-1})$$

$$= \eta_{1\ldots n} \prod_{i=1\ldots n} P(z_i \mid x)\, P(x)$$

Example: Second Measurement

- New prior (from last measurement): $P(open|z_1)$=2/3
- Assume sensor model says: $P(z_2|open) = 0.5$ $P(z_2|\neg open) = 0.6$

$$P(open \mid z_2, z_1) = \frac{P(z_2 \mid open)\, P(open \mid z_1)}{P(z_2 \mid open)\, P(open \mid z_1) + P(z_2 \mid \neg open)\, P(\neg open \mid z_1)}$$

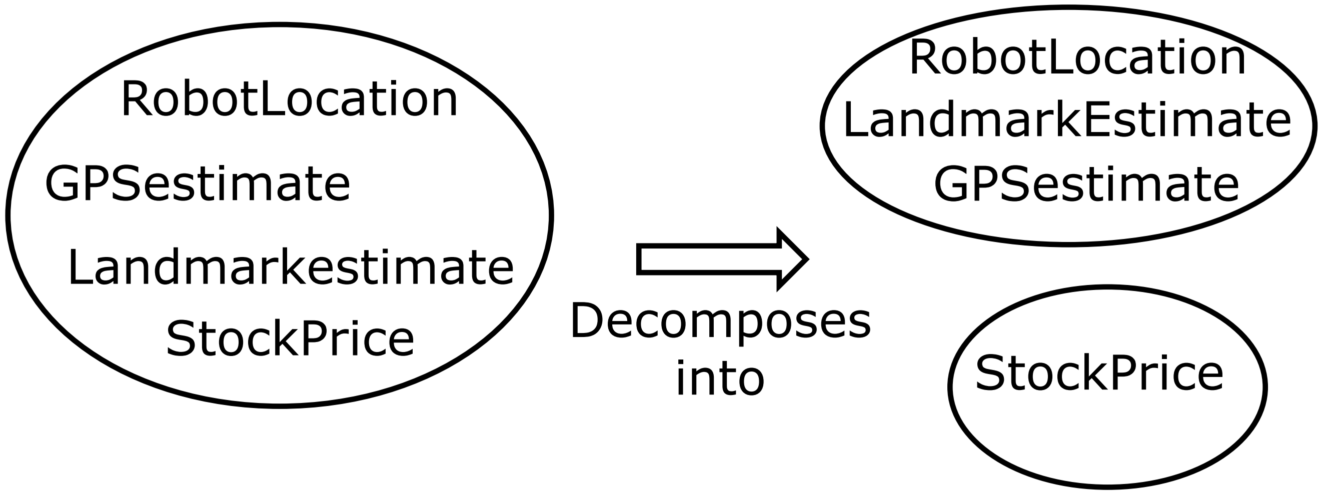
$$= \frac{\frac{1}{2}\cdot\frac{2}{3}}{\frac{1}{2}\cdot\frac{2}{3}+\frac{3}{5}\cdot\frac{1}{3}} = \frac{5}{8} = 0.625$$

- z_2 lowers the probability that the door is open (was 0.67 , now 0.625)

Independence

- A and B are independent iff

$$\mathbf{P}(A|B) = \mathbf{P}(A) \quad \text{or} \quad \mathbf{P}(B|A) = \mathbf{P}(B) \quad \text{or} \quad \mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$$



I.e. $P(RobotLocation, GPSestimate, LandmarkEstimate, StockPrice)$
$= P(RobotLocation, GPSestimate, LandmarkEstimate)P(StockPrice)$

- This makes distributions easier to compute, but true independence is rare
  - Often everything in a domain depends on everything else

Conditional Independence

- x and y are *conditionally independent* given z iff

$$P(x, y \mid z) = P(x \mid z) P(y \mid z)$$

- Equivalent to

$$P(x \mid z) = P(x \mid z, y)$$

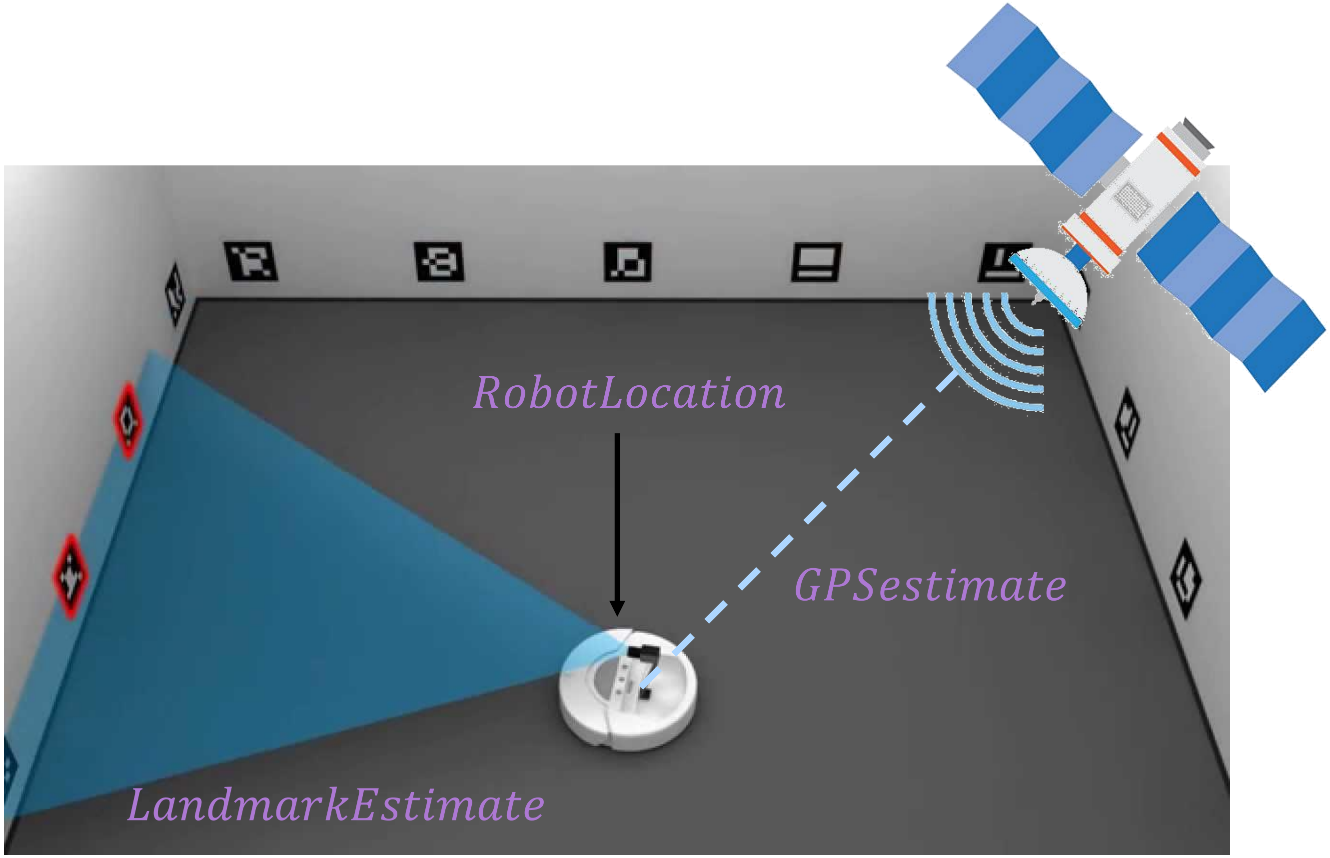
and

$$P(y \mid z) = P(y \mid z, x)$$

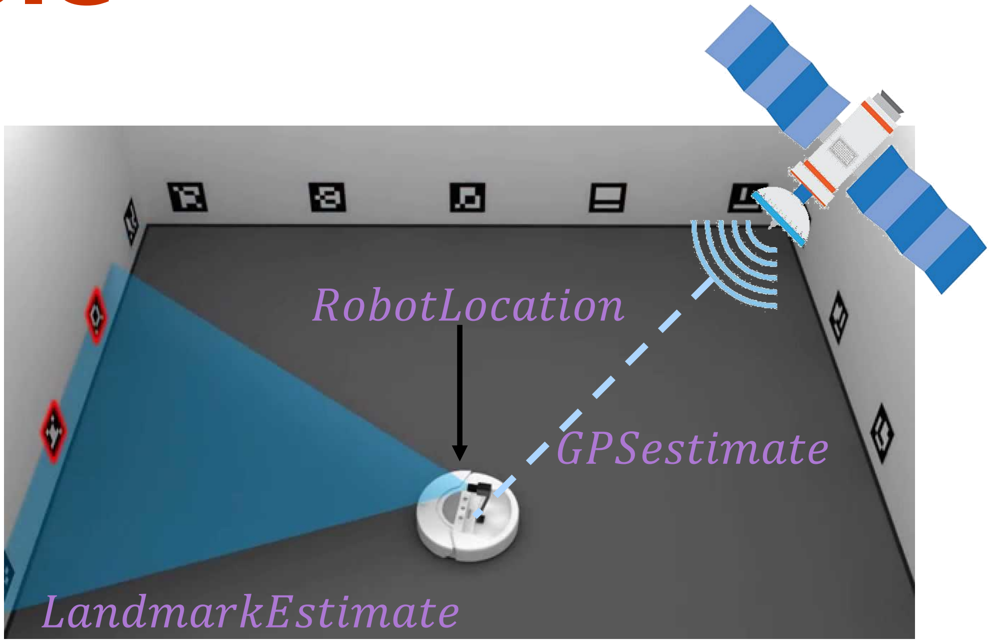
i.e. “If I know z, I don’t need to know x to compute the probability of y”

Conditional Independence Example

- Consider the variables $RobotLocation, GPSestimate, LandmarkEstimate$

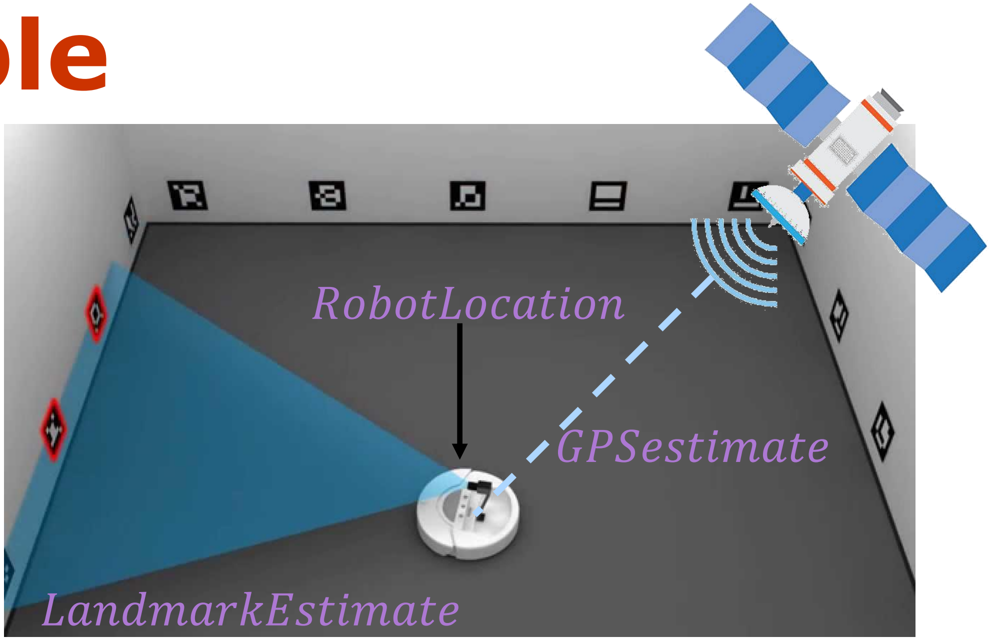


Conditional Independence Example



- $GPSestimate$ and $LandmarkEstimate$ are NOT independent:
  $P(GPSestimate|LandmarkEstimate) \neq P(GPSestimate)$
  - Because knowing $LandmarkEstimate$ gives us information about $GPSestimate$

Conditional Independence Example



- $GPSestimate$ and $LandmarkEstimate$ **are conditionally independent given** $RobotLocation$, which means…

$$P(GPSestimate|RobotLocation, LandmarkEstimate)$$
$$= P(GPSestimate|RobotLocation)$$

and

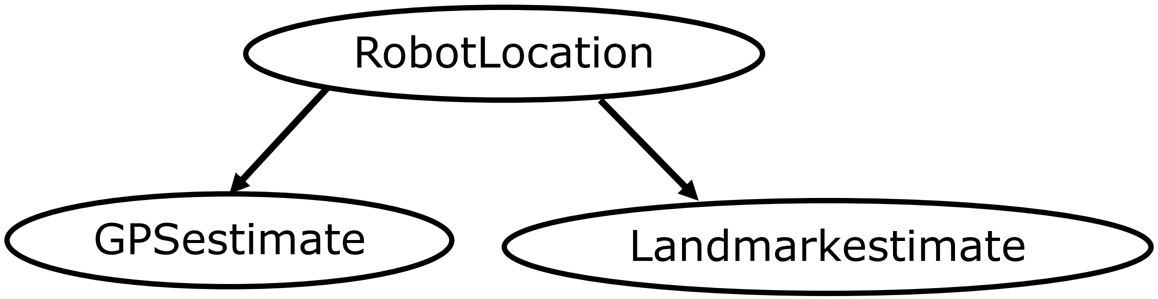
$$P(LandmarkEstimate|RobotLocation, GPSestimate)$$
$$= P(LandmarkEstimate|RobotLocation)$$

i.e. "if I know the robot's location, I can compute the landmark estimate without knowing the GPS estimate"

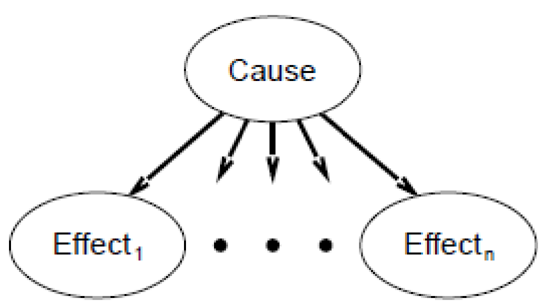
BREAK

Bayes nets

- Convenient to encode conditional independence relationships in a *Bayes Net*



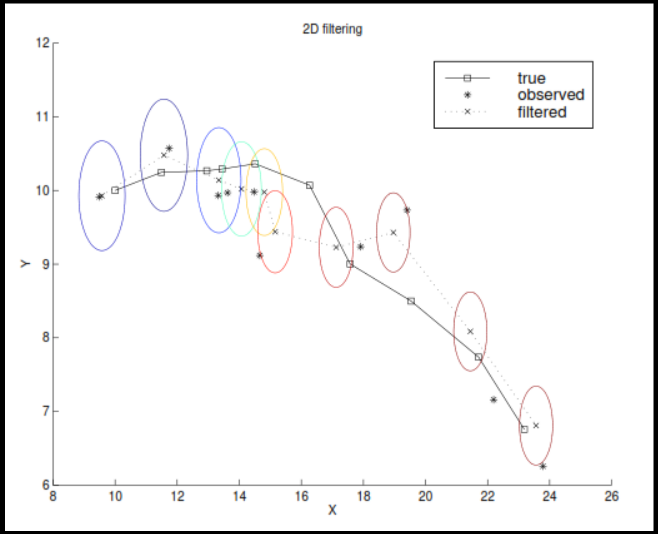
- Bayes nets are often used to describe cause-effect relationships:

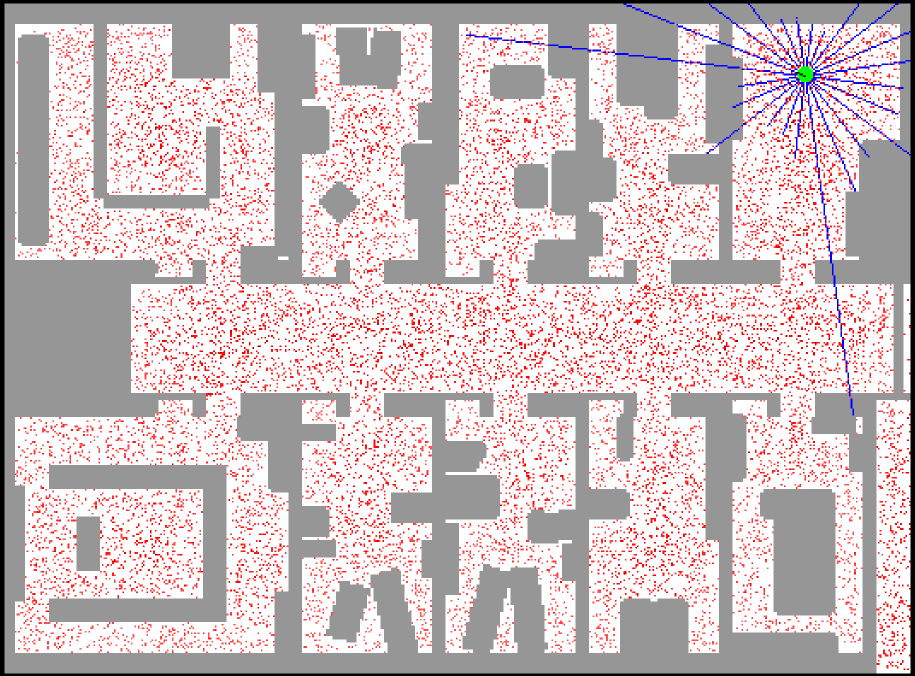


$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\prod_i \mathbf{P}(Effect_i|Cause)$$

Bayes nets in Robotics

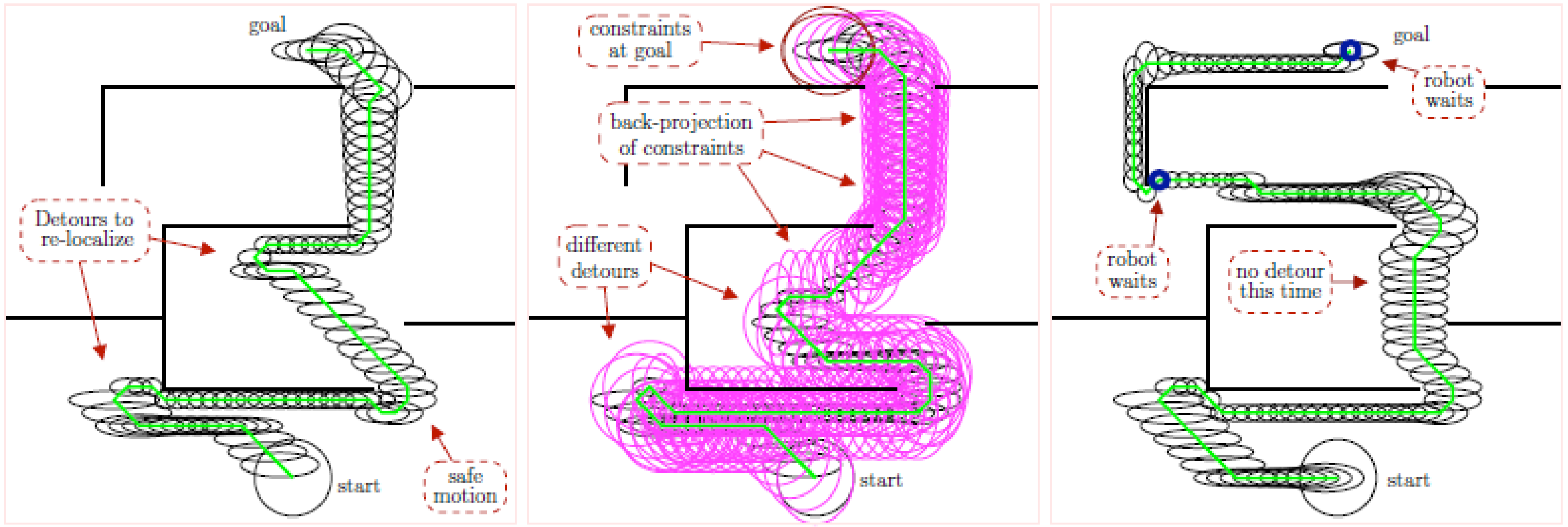
- State estimation
  - Kalman Filters
  - Particle Filters
  - We will cover these later!
- SLAM





Bayes nets in Robotics

- Bayesian reasoning in motion planning:

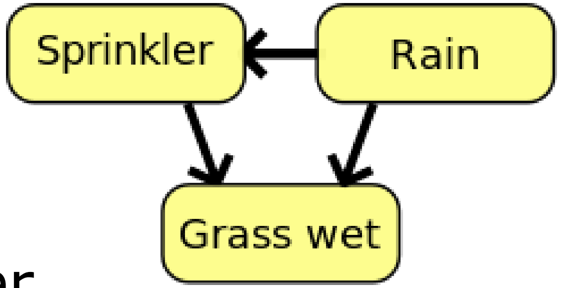


(e) Minimum-time solution (forward algorithm) (f) Minimum-time solution (backward algorithm) (g) Minimum-final-covariance solution

A Bayesian framework for optimal motion planning with uncertainty,
Andrea Censi, Daniele Calisi, Alessandro De Luca, Giuseppe Oriolo, ICRA, 2008

Bayesian Networks Structure

- Directed and acyclic graph
- Nodes represent random variables
- Edges represent conditional dependencies
  - nodes which are not connected represent variables which are *conditionally independent* of each other
- Each node is associated with a probability function $P(X_i|Parents(X_i))$ that
  - takes as input a particular set of values for the node's parent variables
  - gives the probability of the variable represented by the node



Rain influences whether the sprinkler is activated, and both rain and the sprinkler influence whether the grass is wet.

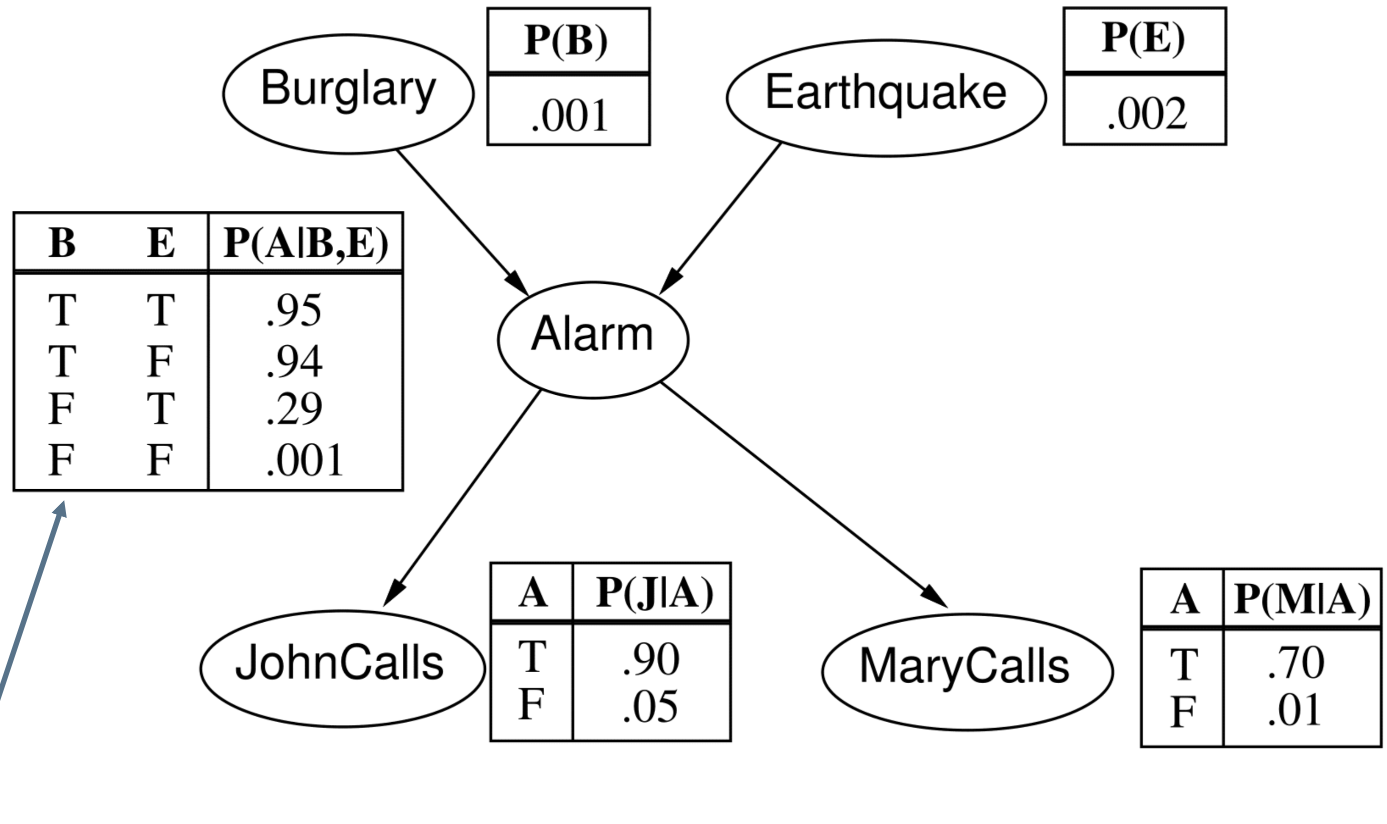
Example

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Conditional Probability Table (CPT) defines $P(X_i|Parents(X_i))$ for discrete variables

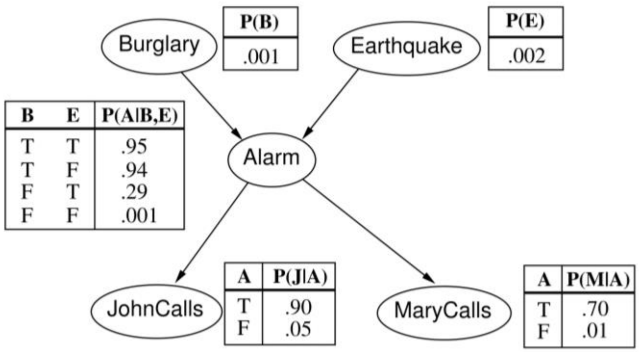
Global semantics

Global semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

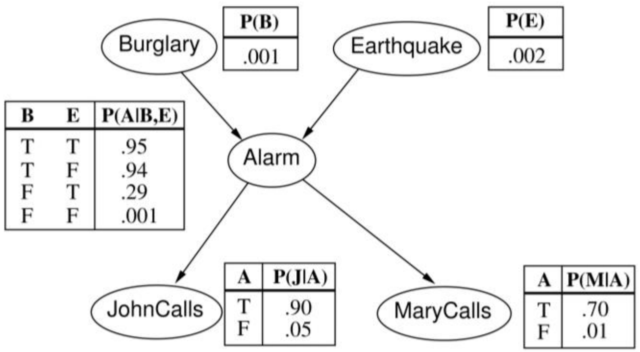
e.g., $P(j \land m \land a \land \neg b \land \neg e)$

$$=$$



Global semantics

"Global" semantics defines the full joint distribution as the product of the local conditional distributions:



$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

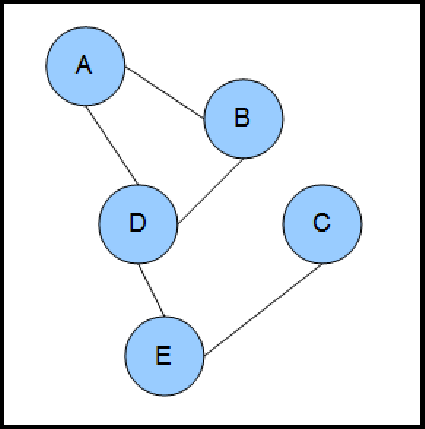
$$\begin{aligned} &= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ &= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ &\approx 0.00063 \end{aligned}$$

- Calculating the probability of a given set of variable values (like we did above) is called *inference*
- Processing every variable (like above) can be very inefficient in large Bayes nets
- The AI book has several algorithms that give efficient approximations (Ch. 14.4-14.5)

Going Further: Markov Random Fields

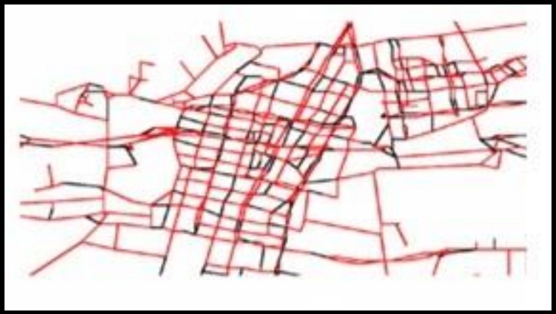
- Same idea as Bayesian networks but…
- Graph is undirected and may be cyclic

- Useful for many applications
  - E.g. reconstructing missing traffic data



A depends on B and D. B depends on A and D. D depends on A, B, and E. E depends on D and C. C depends on E.

True traffic density



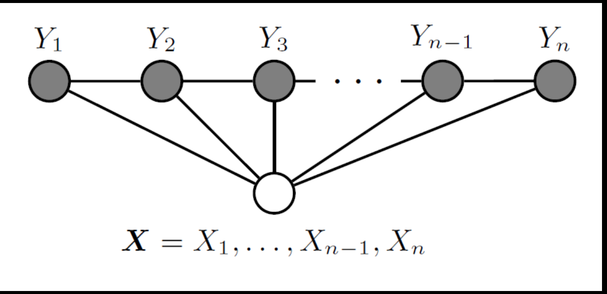
Unobserved roads (red)

Reconstructed traffic data

Traffic data reconstruction based on Markov random field modeling, S. Kataoka, M. Yasuda, C. Furtlehner, and K. Tanaka, Inverse Problems, Volume 30, Number 2, January 2014

Going Further: Conditional Random Fields

- A CRF is an undirected graphical model whose nodes can be divided into exactly two disjoint sets:
  - X, the input variables
  - Y, the observed and output variables
- The conditional distribution P(Y | X) is what is modeled
- In computer vision, CRFs are can be used for object recognition and image segmentation



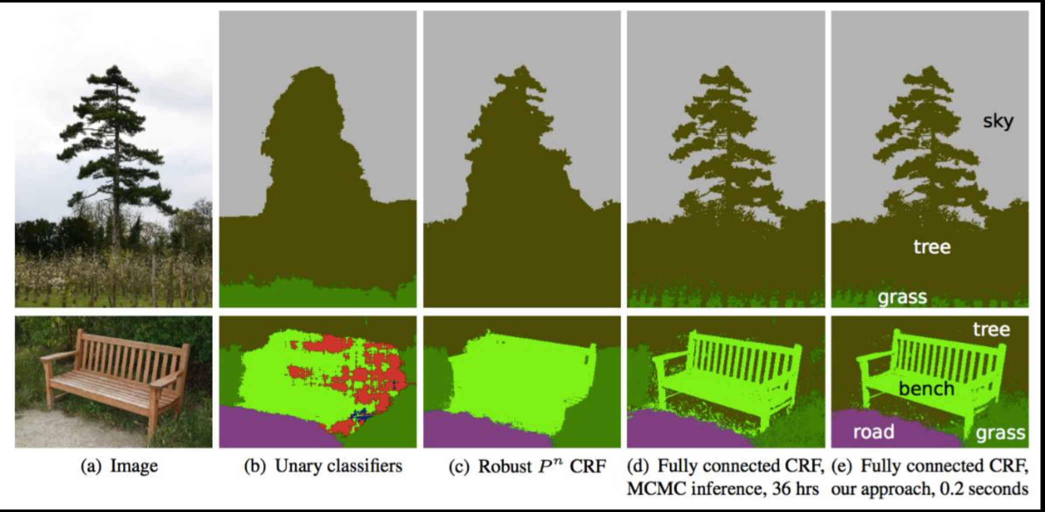


Image from Qure.ai blog

Summary

- Probability is a way to quantify uncertainty
- Bayes rule allows us to reason about causality
- Independent variables make reasoning easier
  - But true independence is rare ☹
- Conditional independence can be described with Bayes nets
- Bayes nets allow us to infer the probability of a complex event happening
  - Like in the earthquake/burglary example

Homework

- AI book Ch. 20
- Remember to start HW 4!