CS 4260 and CS 5260
Vanderbilt University

Lecture on Uncertainty (Sequential Models)

This lecture assumes that you have

- Read Section 8.1 through 8.3, watched lecture on belief network inference, and read section 8.5 of ArtInt

ArtInt: Poole and Mackworth, Artificial Intelligence 2E
at http://artint.info/2e/html/ArtInt2e.html
to include slides at http://artint.info/2e/slides/ch08/lect5.pdf

Douglas H. Fisher

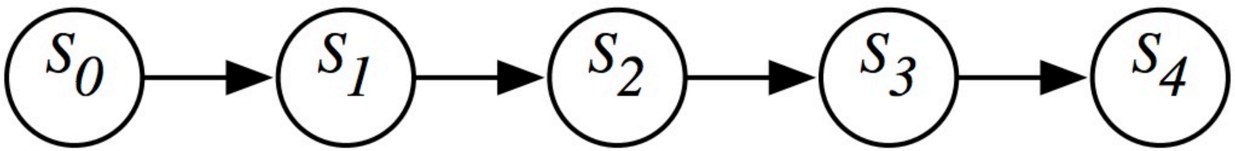
Project ideas

- Increased functionality -- Filling in courses to make 12 credit minimums
  - Randomly?
  - Heuristically?
  - Interactively?
  - Prior knowledge? (semantic web)
  - Machine Learning

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- A Markov chain is a special sort of belief network:



What probabilities need to be specified?

- $P(S_0)$ specifies initial conditions
- $P(S_{i+1}|S_i)$ specifies the dynamics

What independence assumptions are made?

- $P(S_{i+1}|S_0, \ldots, S_i) = P(S_{i+1}|S_i)$.
- Often S_t represents the state at time t. Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- A **Markov chain** is a special sort of belief network:

What probabilities need to be specified?

- $P(S_0)$ specifies initial conditions
- $P(S_{i+1}|S_i)$ specifies the dynamics

As with previous lectures, capital S_i represents an outcome space (i.e., as set of possible states, s_{ik}, and $P(S_i)$ is a probability distribution over s_{ik}s, so

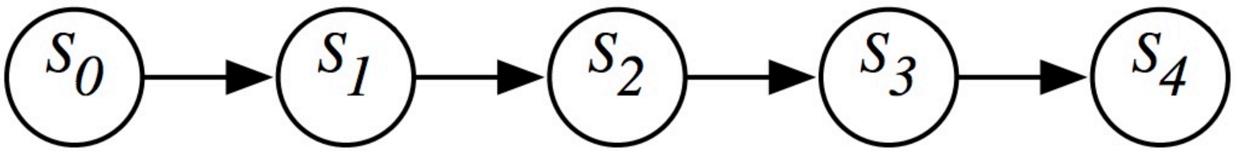
$P(S_0)$ is $P(s_{01})$, $P(s_{02})$, …, $P(s_{0n_0})$

$P(S_1 | S_0)$ is $P(s_{11} | s_{01})$, …, $P(s_{11} | s_{0n_0})$, $P(s_{12} | s_{01})$, …, $P(s_{12} | s_{0n_0})$, …, $P(s_{1n_1} | s_{0n_0})$

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- A Markov chain is a special sort of belief network:



What probabilities need to be specified?

- $P(S_0)$ specifies initial conditions
- $P(S_{i+1}|S_i)$ specifies the dynamics

S_i can represent a primitive state (or outcome) that is not decomposable, like the node in a graph, or more typically, each S_i will be a joint outcome space, as in the state of a plan from chapter 6 or of a training (or test) datum from chapter 7.

$P(s_{ik})$ = P(< lab, ~rhc, swc, mw, rhm>)

For example,

s_{ik} might be < lab, ~rhc, swc, mw, rhm>, or

s_{ik} might be [SciFi = -1, Suspense = 1, Romance = -1, Ebert = 1, Siskel = 1, …, Watch-it = 1]

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

What independence assumptions are made?

- $P(S_{i+1}|S_0, \ldots, S_i) = P(S_{i+1}|S_i)$.
- Often S_t represents the state at time t. Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

Strips-style operators from chapter 6 make this assumption trivially
e.g., puc: Precondition {cs, ~rhc}; Effect {rhc},
where P({cs, rhc,…} | {cs,~rhc,…})=1.0 if puc applied, regardless of path,
and 0.0 otherwise
or perhaps 0.3 overall (just made this up!!!)

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

What independence assumptions are made?

- $P(S_{i+1}|S_0, \ldots, S_i) = P(S_{i+1}|S_i)$.
- Often S_t represents the state at time t. Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

- But what if future does depend on past, as well as the present (where "past" corresponds to the path to the present state)?
- For example, if I am in downtown Nashville, I might be down there for different reasons, and my next step may be dependent of more than the state (e.g., intersection) I am at.
- Years of historic warfare and other grievance may be the classic example of a non-Markov process

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

What independence assumptions are made?

- $P(S_{i+1} | S_0, \ldots, S_i) = P(S_{i+1} | S_i)$.
- Often S_t represents the **state** at time t. Intuitively S_t conveys all of the information about the history that can affect the future states.
- "The future is independent of the past given the present."

- But what if future does depend on past, as well as the present (where "past" corresponds to the past to the present state)? Consider state s_{ik} and path to it as $p\_s_0\_s_{ik}$
- We can still represent as Markov process by representing a state as $\langle s_{ik} , p\_s_0\_s_{ik} \rangle$ That is, embed the path (i.e., "the past") to a state, into a new state description.
- What if next action also depends on a goal, g_m, that agent is pursuing? Then state is $\langle s_{ik} , g_m, p\_s_0\_s_{ik} \rangle$

Markov chain

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- A stationary Markov chain is when for all $i > 0$, $i' > 0$, $P(S_{i+1}|S_i) = P(S_{i'+1}|S_{i'})$. *i.e., transition probabilities never change*
- We specify $P(S_0)$ and $P(S_{\not{i+1}\,1}|S_{\not{i}\,0})$.
  - Simple model, easy to specify
  - Often the natural model
  - The network can extend indefinitely

For example, P(< lab, rhc, swc, mw, rhm>|< lab, ~rhc, swc, mw, rhm>)=0.95 at step 2 , …, at step 23, …, at step 10037, …. then stationary dynamics (or model)

But what if robot is learning? So P(< lab, rhc, swc, mw, rhm>|< lab, ~rhc, swc, mw, rhm>)=0.25 at step 2 , …, 0.86 at step 23, …, 0.995 at step 10037, …. then NON-stationary dynamics

Markov chain

- A distribution over states, P is a stationary distribution if for each state s, $P(S_{i+1}{=}s) = P(S_i{=}s)$.

  i.e., a given state s, is equally likely at each step

- A Markov chain is ergodic if, for any two states s_1 and s_2, there is a non-zero probability of eventually reaching s_2 from s_1. *i.e., s2 is reachable from s1*

- A Markov chain is periodic if there is a strict temporal regularity in visiting states. A state is only visited divisible at time t if $t \mod n = m$ for some n, m.

Markov chain (Pagerank)

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Consider the Markov chain:

- Domain of S_i is the set of all web pages
- $P(S_0)$ is uniform; $P(S_0 = p_j) = 1/N$

See https://en.wikipedia.org/wiki/PageRank for more details

$$P(S_{i+1} = p_j \mid S_i = p_k)$$

$$= (1 - d)/N + d * \begin{cases} 1/n_k & \text{if } p_k \text{ links to } p_j \\ 1/N & \text{if } p_k \text{ has no links} \\ 0 & \text{otherwise} \end{cases}$$

- if p_k links to p_j: *equally likely that each link will be taken*
- if p_k has no links: *uniform random jump to p_j*
- otherwise: *If p_k has links, but p_j is not one of them*

$(1-d)$: *Probability of mental break*; d: *Probability surfing continues*

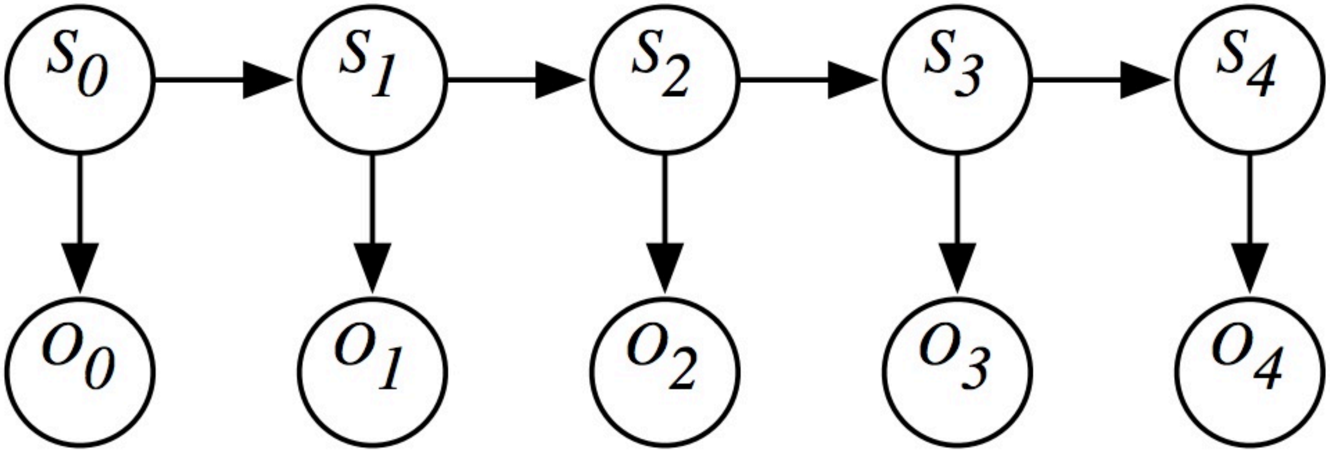
where there are N web pages and n_k links from page p_k

- $d \approx 0.85$ is the probability someone keeps surfing web

Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- A Hidden Markov Model (HMM) is a belief network:



The probabilities that need to be specified:

- $P(S_0)$ specifies initial conditions
- $P(S_{i+1}|S_i)$ specifies the dynamics
- $P(O_i|S_i)$ specifies the sensor model

Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Filtering:

$$P(S_i | o_1, \ldots, o_i)$$ *Probability distribution of each state conditioned on all prior observations*

What is the current belief state based on the observation history?

s_{ik}

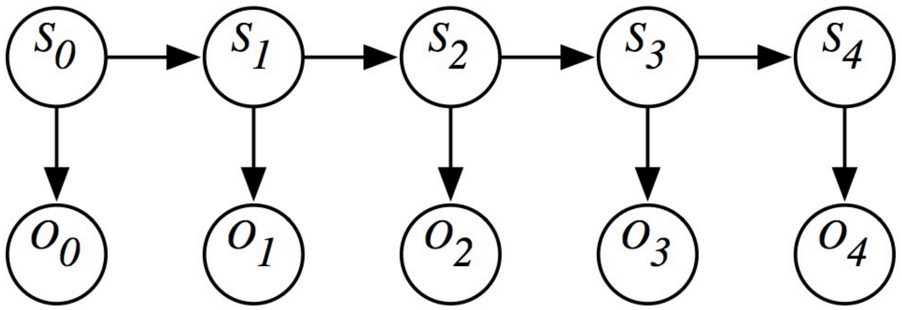
$$P(\not{S}_i | o_1, \ldots, o_i) = P(s_{ik}, o_1, \ldots, o_i) / P(o_1, \ldots, o_i)$$

propto $P(s_{ik}, o_1, \ldots, o_i)$

$P(S_0=x|O_0=a)?$
$= P(O_0=a|S_0=x)P(S_0=x)/P(O_0=a)$

$P(S_0=y|O_0=a)?$
$= P(O_0=a|S_0=y)P(S_0=y)/P(O_0=a)$

- Observe O_0, query S_0.
- then observe O_1, query S_1.
- then observe O_2, query S_2.
- ...



Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Filtering:

$$P(S_i | o_1, \ldots, o_i)$$ *Probability distribution of each state conditioned on all prior observations*

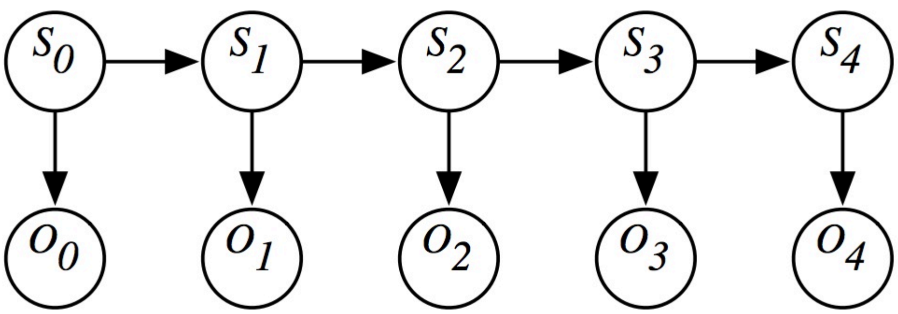
What is the current belief state based on the observation history?

$$P(s_{ik} | o_1, \ldots, o_i) = P(s_{ik}, o_1, \ldots, o_i) / P(o_1, \ldots, o_i)$$

propto $P(s_{ik}, o_1, \ldots, o_i)$

Using what you learned about inference with belief networks, give $P(S_1{=}x \mid O_1{=}a, O_2{=}b)$ only in terms of probabilities found in (or trivially computed from) the probability tables of the belief network below, where the domain of each step S are the states x and y. The observation variables at each step are the same, with the same domains (a, b).

- Observe O_0, query S_0.
- then observe O_1, query S_1.
- then observe O_2, query S_2.
- ...



Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Using what you learned about inference with belief networks, give $P(S_1=x \mid O_1=a, O_2=b)$ only in terms of probabilities found in (or trivially computed from) the probability tables of the belief network below, where the domain of each step S are the states x and y. The observation variables at each step are the same, with the same domains (a, b).

$P(S_1=x \mid O_0=a, O_1=b)$

$= P(S_1=x, O_0=a, O_1=b) / P(O_0=a, O_1=b)$

propto $= P(S_1=x, O_0=a, O_1=b)$

$= P(O_1=b \mid S_1=x, O_0=a)P(S_1=x, O_0=a)$

$= P(O_1=b \mid S_1=x)[P(S_1=x, O_0=a, S_0=x) + P(S_1=x, O_0=a, S_0=y)]$

= ????

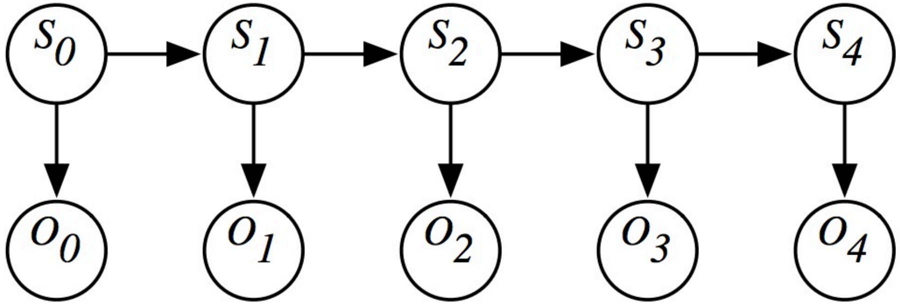
$P(S_1=y \mid O_0=a, O_1=b)$

$= P(S_1=y, O_0=a, O_1=b) / P(O_0=a, O_1=b)$

propto $= P(S_1=y, O_0=a, O_1=b)$

= ????

- Observe O_0, query S_0.
- then observe O_1, query S_1.
- then observe O_2, query S_2.
- ...

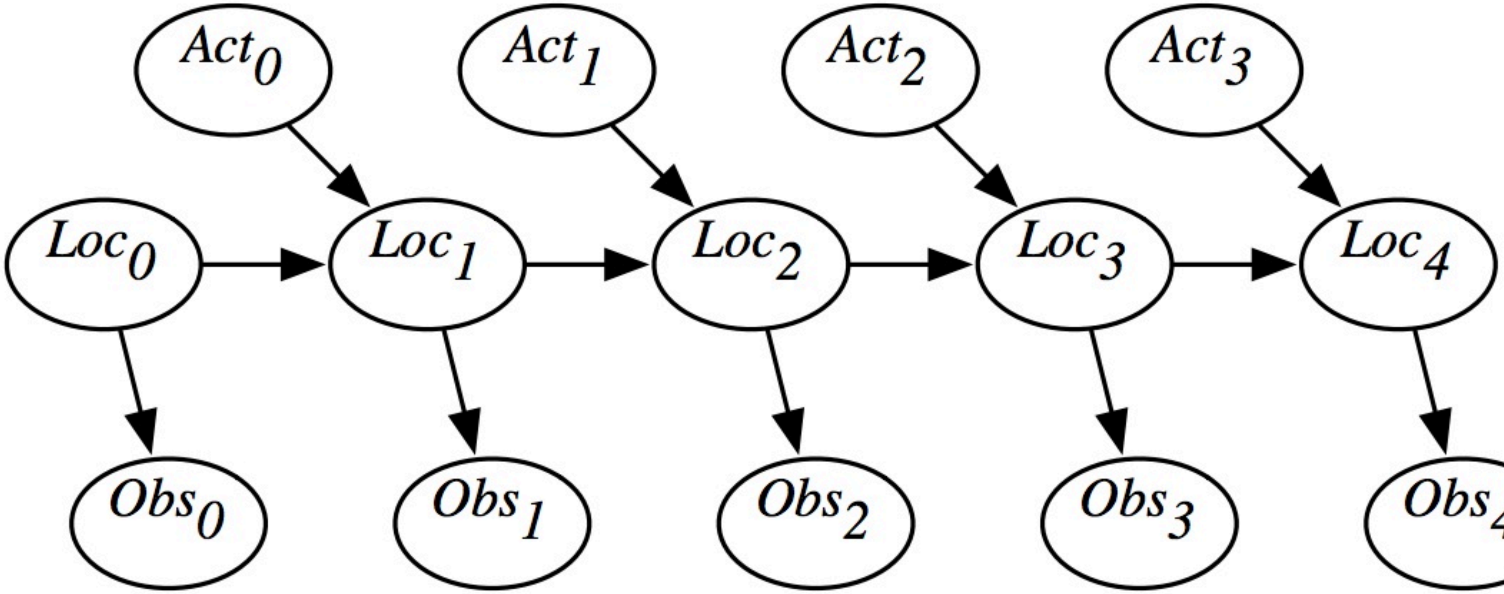


Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

HMMs augmented with actions, like STRIPS operators, though with probabilistically qualified effects

- Suppose a robot wants to determine its location based on its actions and its sensor readings: Localization
- This can be represented by the augmented HMM:



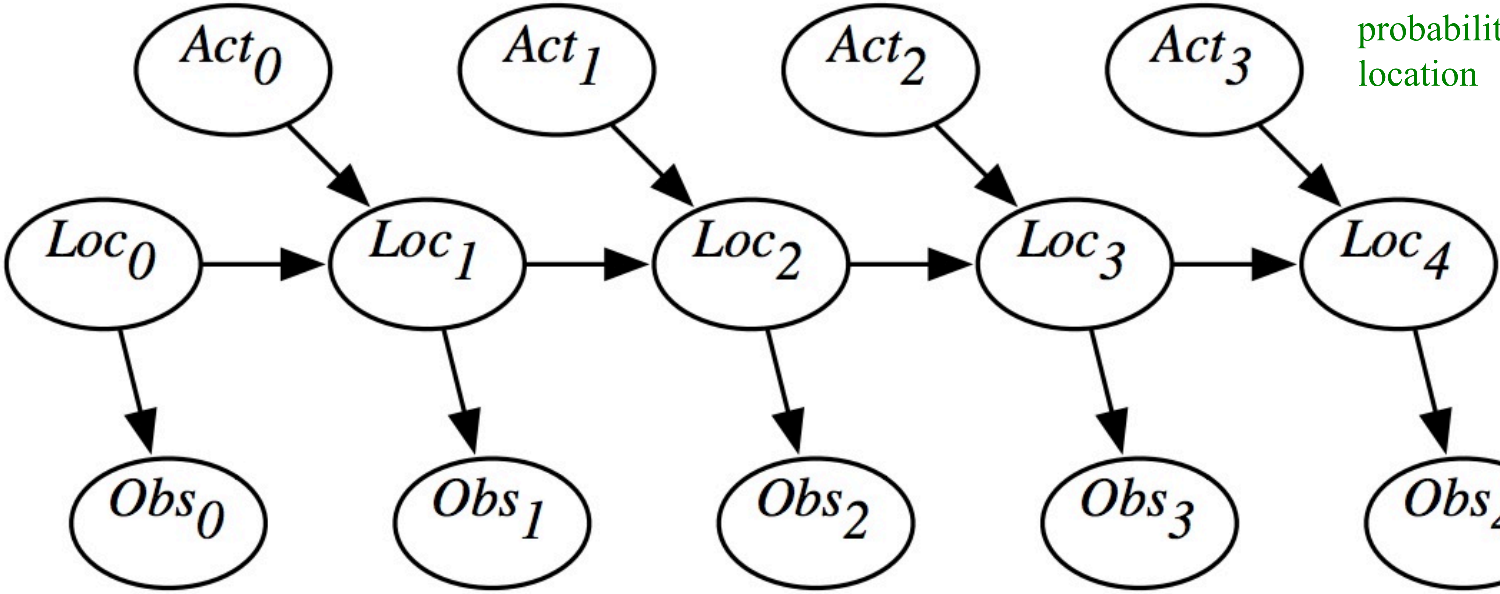
Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

HMMs augmented with actions, like STRIPS operators, though with probabilistically qualified effects

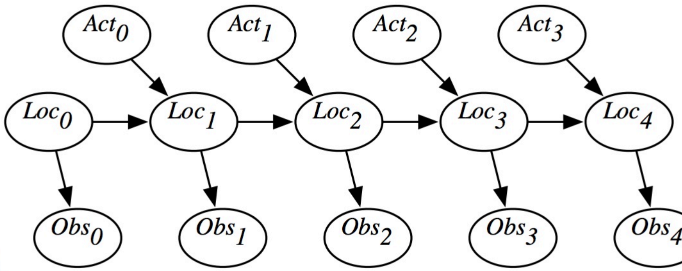
- Suppose a robot wants to determine its location based on its actions and its sensor readings: Localization
- This can be represented by the augmented H

This example is a bit misleading, because the example assumes that an observer is computing probabilities of the robot's location



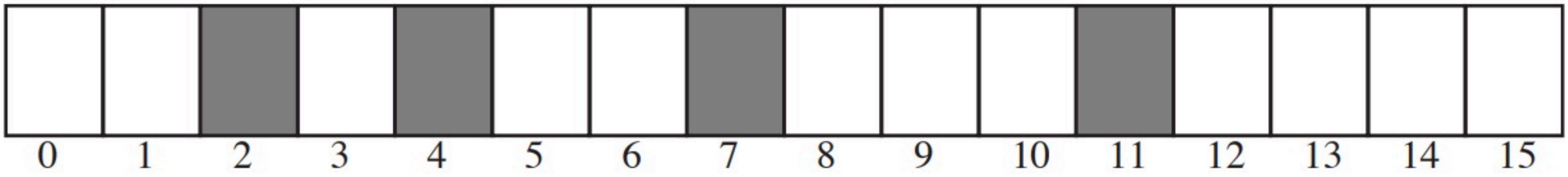
Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf



Example of localization

- Circular corridor, with 16 locations:

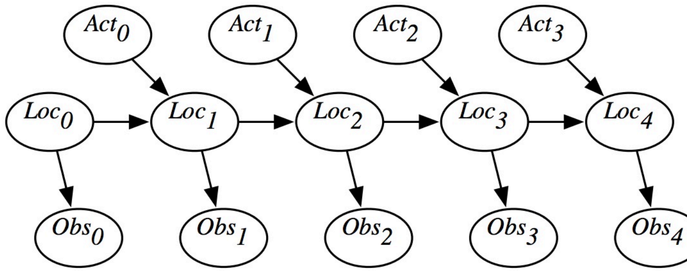


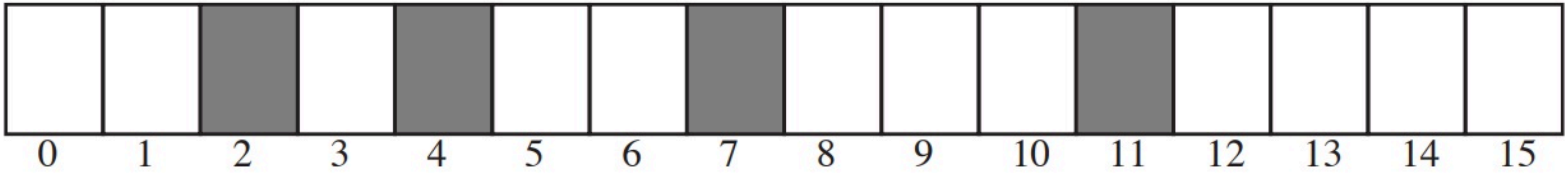
- Doors at positions: 2, 4, 7, 11.
- Noisy Sensors *to sense whether in front of a door*
- Stochastic Dynamics *transition probabilities*
- Robot starts at an unknown location and must determine where it is.

Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

$P(Loc_2{=}7 \mid O_0{=}{\sim}od, A_0{=}gR, O_1{=}od, A_1{=}gR, O_2{=}od)$???





Sensor Model

- $P(Observe\ Door \mid At\ Door) = 0.8$ $P({\sim}od \mid ad) = 0.2$
- $P(Observe\ Door \mid Not\ At\ Door) = 0.1$ $P({\sim}od \mid {\sim}ad) = 0.9$

Dynamics Model

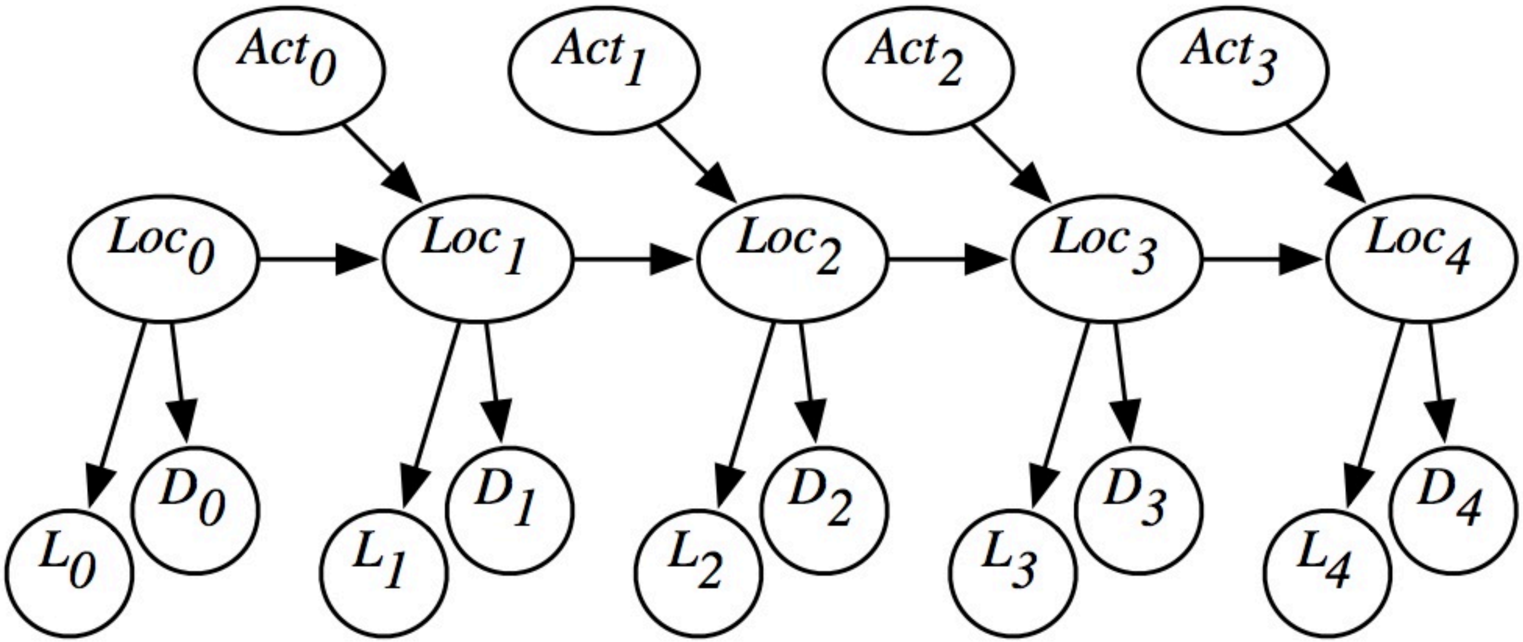
- $P(loc_{t+1} = L \mid action_t = goRight \wedge loc_t = L) = 0.1$
- $P(loc_{t+1} = L + 1 \mid action_t = goRight \wedge loc_t = L) = 0.8$
- $P(loc_{t+1} = L + 2 \mid action_t = goRight \wedge loc_t = L) = 0.074$
- $P(loc_{t+1} = L' \mid action_t = goRight \wedge loc_t = L) = 0.002$ for any other location L'.
  - All location arithmetic is modulo 16.
  - The action *goLeft* works the same but to the left.

Hidden Markov Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

- **Example:** we can combine information from a light sensor and the door sensor **Sensor Fusion**

Specify probability tables, and Perform localization



S_t robot location at time t

D_t door sensor value at time t

L_t light sensor value at time t

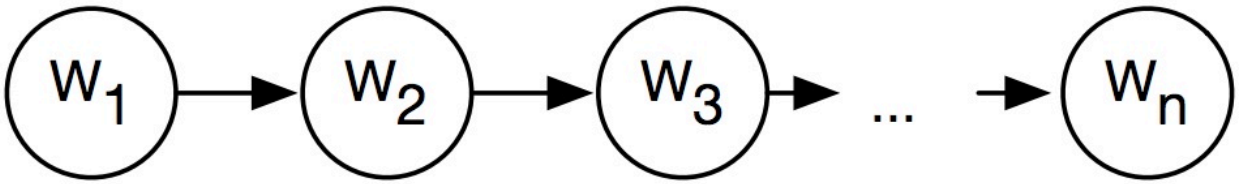
Location induces conditional dependence between prior location and action

Simple Language Models: bigram

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

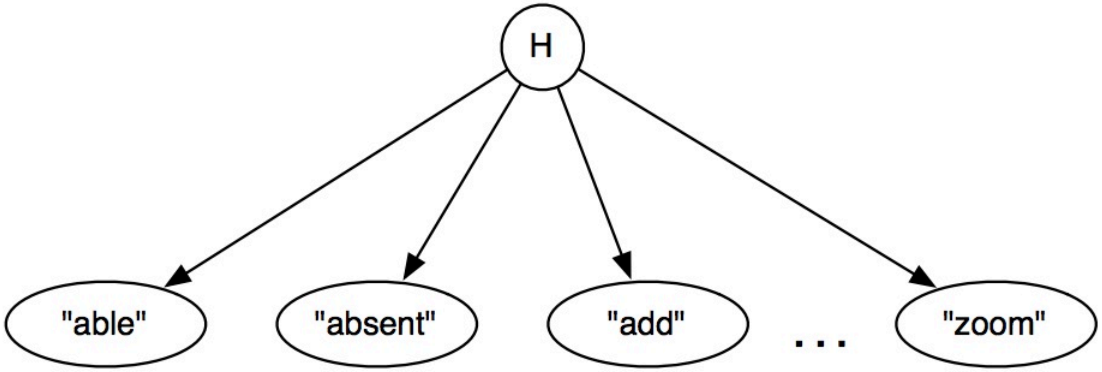
Sentence: $w_1, w_2, w_3, \ldots, w_n$.

bigram:



- Domain of each variable is the set of all words.
- What probabilities are provided?
  - $P(w_i | w_{i-1})$ is a distribution over words for each position given the previous word
- How do we condition on the question "how can I phone my phone"?

Naive Bayes Classifier: User's request for help



H is the help page the user is interested in.
What probabilities are required?

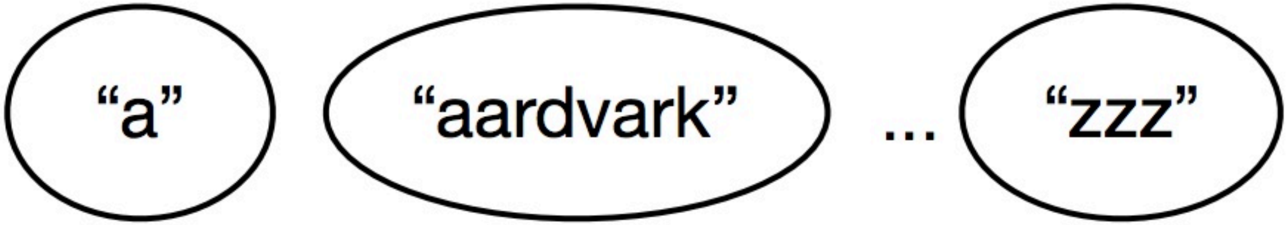
- $P(h_i)$ for each help page h_i. The user is interested in one best web page, so $\sum_i P(h_i) = 1$.
- $P(w_j \mid h_i)$ for each word w_j given page h_i. There can be multiple words used in a query.
- Given a help query: condition on the words in the query and display the most likely help page.

Simple Language Models: set-of-words

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Sentence: $w_1, w_2, w_3, \ldots$.

Set-of-words model:



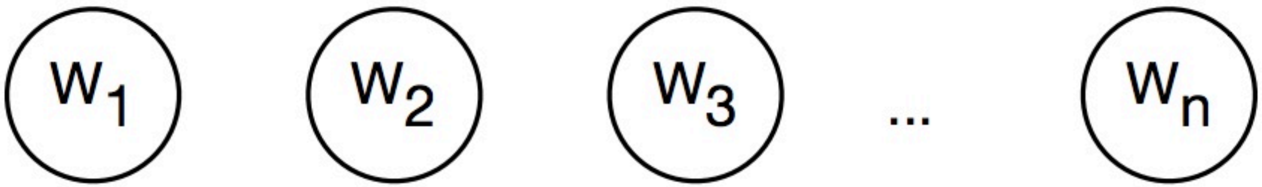
- Each variable is Boolean: *true* when word is in the sentence and *false* otherwise.
- What probabilities are provided?
  - $P(\text{"a"}), P(\text{"aardvark"}), \ldots, P(\text{"zzz"})$
- How do we condition on the question "how can I phone my phone"?

Simple Language Models: bag-of-words

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Sentence: $w_1, w_2, w_3, \ldots, w_n$.

Bag-of-words or unigram:



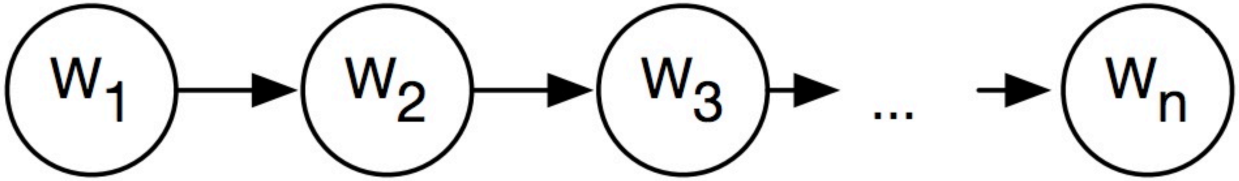
- Domain of each variable is the set of all words.
- What probabilities are provided?
  - $P(w_i)$ is a distribution over words for each position
- How do we condition on the question "how can I phone my phone"?

Simple Language Models: bigram

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Sentence: $w_1, w_2, w_3, \ldots, w_n$.

bigram:



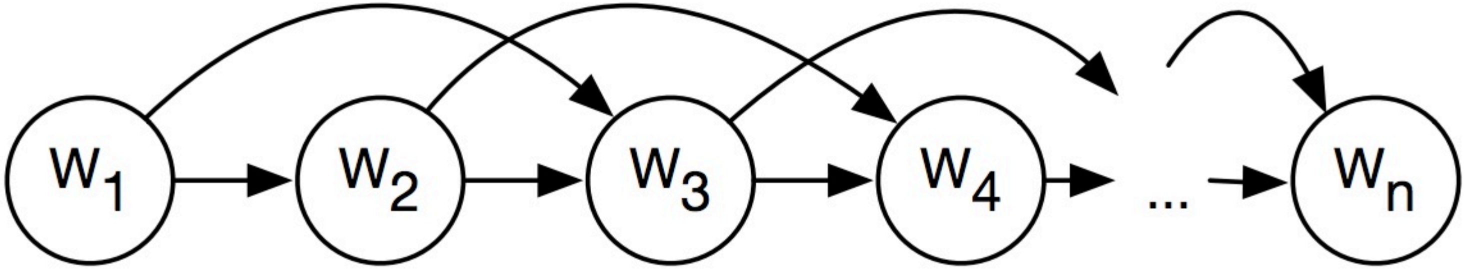
- Domain of each variable is the set of all words.
- What probabilities are provided?
  - $P(w_i|w_{i-1})$ is a distribution over words for each position given the previous word
- How do we condition on the question "how can I phone my phone"?

Simple Language Models: trigram

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

Sentence: $w_1, w_2, w_3, \ldots, w_n$.

trigram:



Domain of each variable is the set of all words.

What probabilities are provided?

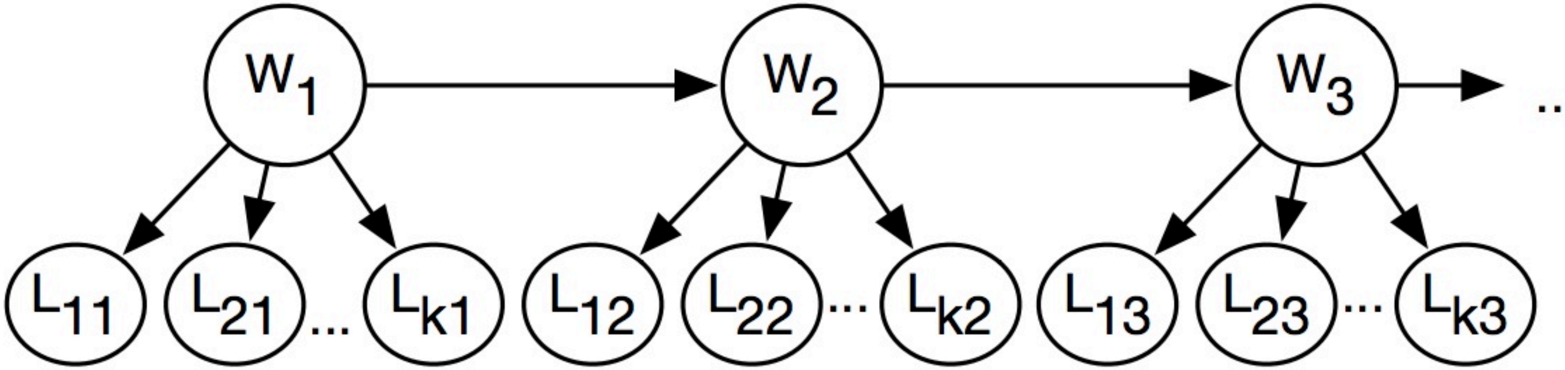
- $P(w_i | w_{i-1}, w_{i-2})$

N-gram

- $P(w_i | w_{i-1}, \ldots w_{i-n+1})$ is a distribution over words given the previous $n-1$ words

Predictive Typing and Error Correction

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf



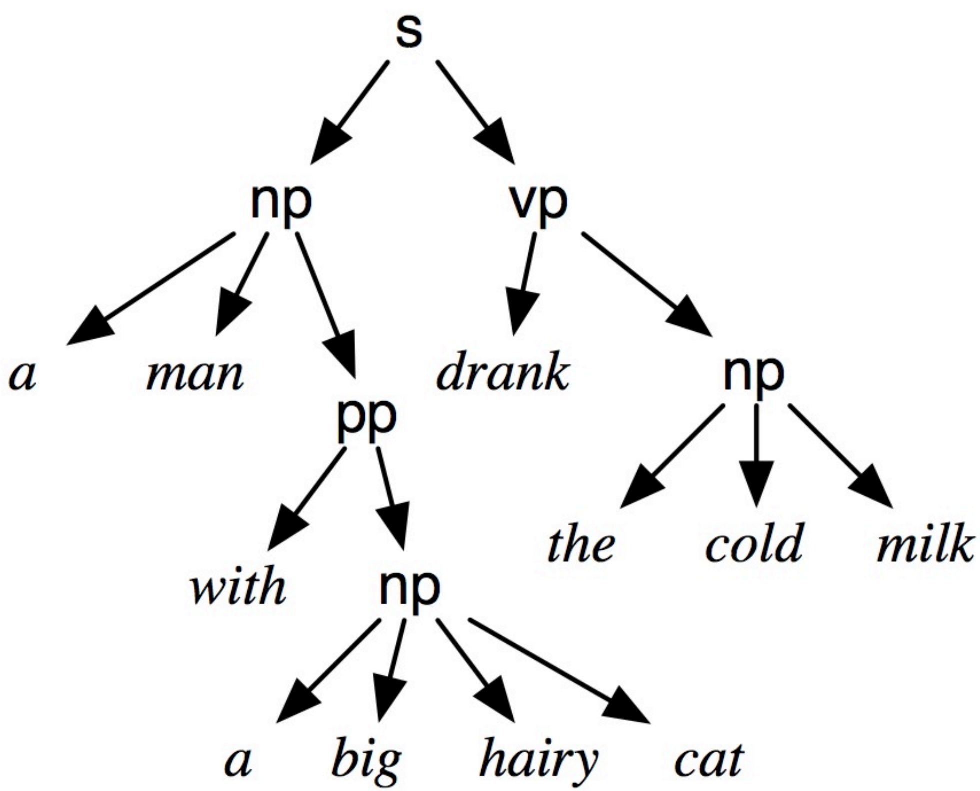
$$domain(W_i) = \{"a", "aarvark", \ldots, "zzz", "\bot", "?"\}$$
$$domain(L_{ji}) = \{"a", "b", "c", \ldots, "z", "1", "2", \ldots\}$$

Beyond N-grams

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf

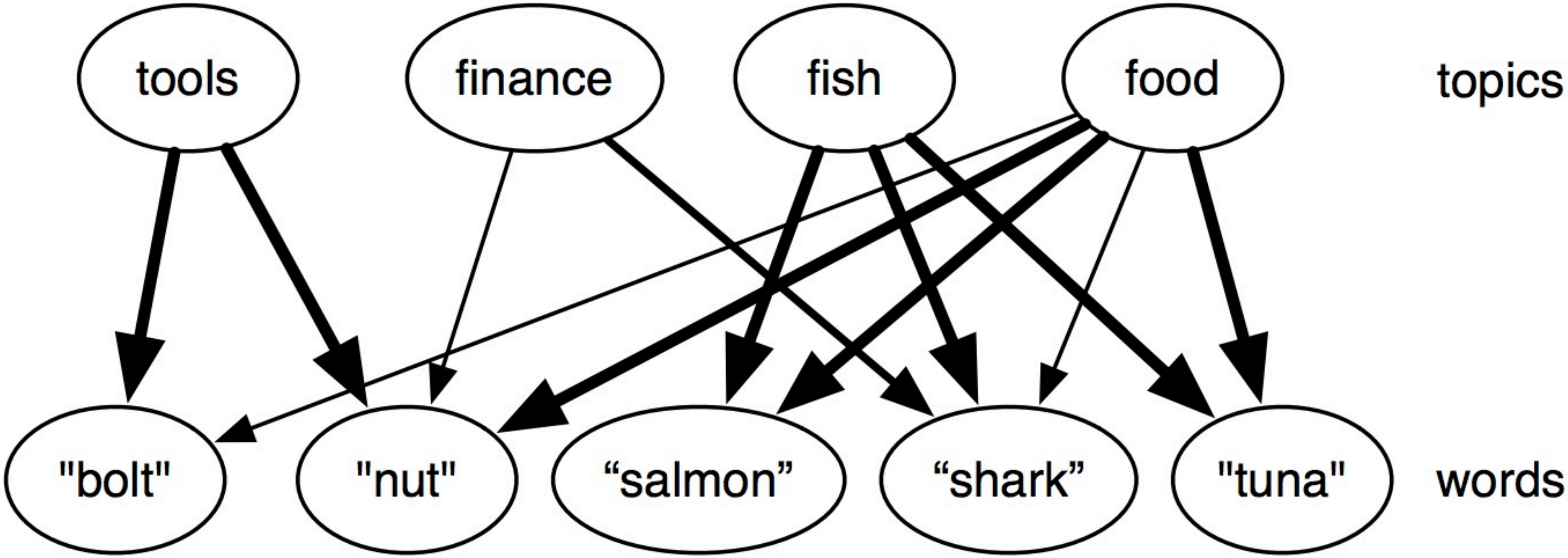
- *A man with a big hairy cat drank the cold milk.*
- Who or what drank the milk?

Simple syntax diagram:



Topic Model

Adapted from Poole and Mackworth, Artificial Intelligence 2E slides at http://artint.info/2e/slides/ch08/lect5.pdf



An example of topic modeling

Incorporating Sustainability into Computing Education

Douglas H. Fisher, Zimei Bian, Selina Chen

IEEE Intelligent Systems, Vol. 31, No. 5 (2016)

- Sustainability and Assistive Computing (Bryn Mawr College, Fall 2010);
- Computing and the Environment (Vanderbilt University, Spring 2011);
- Topics in Computational Sustainability (Cornell University, Spring 2011);
- Computational Sustainability (University of British Columbia, Winter 2013–2014);
- Computational Sustainability (Georgia Tech, Spring 2014);
- Seminar on Computational Sustainability: Algorithms for Ecology and Conservation (University of Massachusetts Amherst, Spring 2014)

Incorporating Sustainability into Computing Education

Douglas H. Fisher, Zimei Bian, Selina Chen

IEEE Intelligent Systems, Vol. 31, No. 5 (2016)

TOPICS GENERATED			
Topic #	*Weight*	*Keywords*	*Topic Name*
0	0.15074	energy power data consumption time carbon electricity environmental system	GreenIT/Energy
1	0.18246	problem algorithm set time sensor greedy network number optimal	Optimization/Sensor
2	0.16311	data environmental urban energy services development science land government	Urban/Policy
3	0.09139	problem cost solution budget corridor connectivity habitat connected conservation	Optimization/Land
4	0.08485	waste electronic media hazardous equipment social nigeria computer countries	GreenIT/Materials
5	0.27841	model data models species distribution set maxent detection modeling	Modeling/Species
6	0.11874	energy building cost design optimization model optimisation objective buildings	Optimization/Built
7	0.09318	model capture data survival time models rates parameters recapture	Modeling/Method
8	0.12163	food network species webs web time information data networks	Ecology Webs
9	0.09067	climate change global water ocean sea earth fish system	Earth Systems

Incorporating Sustainability into Computing Education

Douglas H. Fisher, Zimei Bian, Selina Chen

IEEE Intelligent Systems, Vol. 31, No. 5 (2016)

COURSE TOPIC WEIGHTS

School	*Topic 0*	*Topic 1*	*Topic 2*	*Topic 3*	*Topic 4*
Bryn Mawr	0.090943549	**0.127644406**	**0.20480037**	2.10E-05	**0.265664737**
Cornell	7.22E-05	0.085409982	**0.174295598**	0.009161242	0.005980967
Georgia Tech	0.081458989	**0.136824135**	**0.100419814**	**0.125061275**	0.061678773
UBC	**0.200559536**	0.018010526	**0.172902203**	0.044725581	0.052835175
UMass Amherst	1.87E-05	**0.177675797**	6.20E-04	**0.217023506**	2.66E-06
Vanderbilt	**0.354199272**	0.033780717	0.02020729	**0.253033232**	0.072572848

School	*Topic 5*	*Topic 6*	*Topic 7*	*Topic 8*	*Topic 9*
Bryn Mawr	**0.29306572**	0.001092996	0.002332577	0.005188805	0.009245879
Cornell	0.054950987	0.056984767	0.089727397	**0.474219654**	0.04919718
Georgia Tech	**0.193939583**	**0.14640088**	0.028616956	0.038639172	0.086960423
UBC	**0.102387938**	**0.100914674**	5.24E-05	0.010594252	**0.297017732**
UMass Amherst	**0.284061303**	0.030038263	**0.283903305**	0.006486598	1.70E-04
Vanderbilt	0.048782513	0.020952409	2.51E-04	**0.137485102**	0.058735835