Soft Computing – Probabilistic Reasoning

- Dynamic Bayesian Networks-

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

Course Syllabus (Tentative)

Probability basics (fast and furious)

- Frequentists vs Bayesians
- Joint and Naive Distributions

Probabilistic graphical models

- Directed graphical models (Bayesian Networks)
- Conditional independence and d-separation
- Inference in directed graphical models

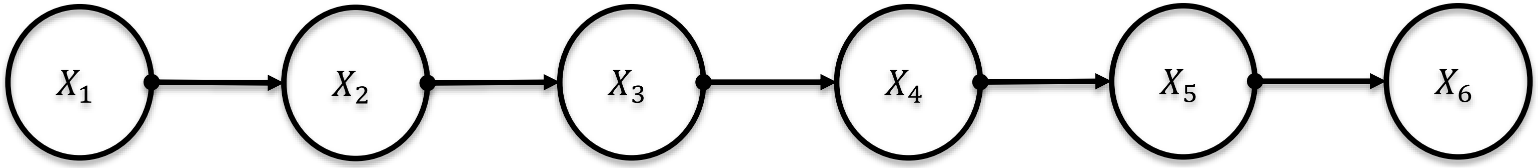
Dynamical graphical models

- Markov chains
- Hidden Markov models

Learning directed graphical models …

Probabilistic Reasoning for (Time) Series

To describe an ever changing world we can use a series of random variables describing the world state at any time instant!



- It represents a sequence of states $X_1, X_2, X_3, \ldots$ where the number represents the position in the sequence (often time)
- We assume the transition from $X_{t-1} = x_i$ to $X_t = x_j$ depends only on X_{t-1}

$$P(X_t | X_{t-1}, X_{t-2}, \ldots, X_0) = P(X_t | X_{t-1})$$

Markov Property

- In a <u>*Stationary Process*</u> transition probabilities are the same a any t
- This is just a <u>*Bayesian Network*</u> that forms a chain!

Soft Computing – Probabilistic Reasoning

- Markov Chains-

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

Stochastic Processes and Markov Chains

Given X_t the value of a (state) random variable at time t :

- *Discrete Stochastic Process* describes the relationship between the stochastic description of a system $(X_0, X_1, X_2, \ldots)$ at some discrete time steps.
- *Continuous Stochastic Process* is a stochastic process where the state can be observed at any time.

A Discrete Stochastic Process is a (first order) *Markov Chain* when we have that $\forall t = 1,2,3, \ldots$ and for all N states it holds:

$$P(X_t | X_{t-1}, X_{t-2}, .., X_0) = P(X_t | X_{t-1})$$

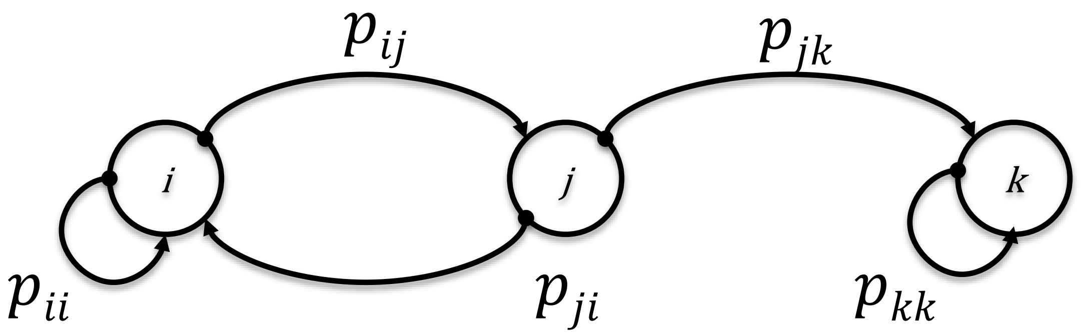
Whenever the probability of an event is independent from time the Markov Chain is *Stationary*: $P(X_{t+1} = j | X_t = i) = p_{ij}$

Markov Chain Description

A Markov Chain can be described using a <u>*Transition Matrix*</u> where $\boldsymbol{p_{ij}}$ describes the probability of getting into state $\boldsymbol{j}$ starting from state $\boldsymbol{i}$:

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & \ddots & \vdots \\ p_{N1} & \cdots & p_{NN} \end{bmatrix}, \qquad \sum_{j=1}^{N} p_{ij} = 1$$

This transition matrix can be described also using a directed graph



<u>Note:</u> different values of the same random variable!

Computing Probabilities

Given a Markov Chain in state i at time m, states probability after n steps:

$$P(X_{m+n} = j | X_m = i) = P(X_n = j | X_0 = i) = P_{ij}(n)$$

If we take $n = 2$ we have

$$P_{ij}(n) = \sum_k p_{ik} \cdot p_{kj}$$

Scalar product of row i and column j

In general $P_{ij}(n) \; = \; ij^{th}$ element of P^n

Probability of being in a given state j at time n without knowing the exact state of Markov Chain at time $\mathbf{0}$ is:

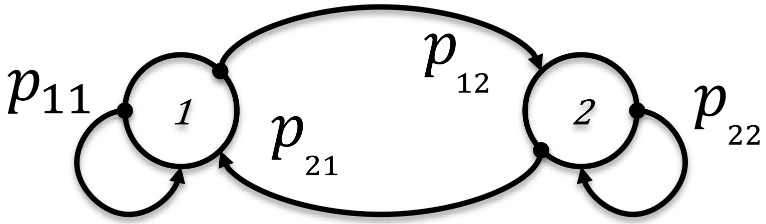
$$\sum_i q_i \cdot P_{ij}(n) = q \cdot (column\, j\; of\;\; P^n)$$

q_i is the state probability at time 0

The Cola Example (1)

We have just two brands of Cola on the market (i.e., $\boldsymbol{Cola_1}$, and $\boldsymbol{Cola_2}$). A person buying $\boldsymbol{Cola_1}$ will buy $\boldsymbol{Cola_1}$ again with probability 0.9. A persona buying $\boldsymbol{Cola_2}$will buy $\boldsymbol{Cola_2}$ again with probability 0.8.

$$P = \begin{array}{c c} & \begin{array}{cc} Cola_1 & Cola_2 \end{array} \\ \begin{array}{c} Cola_1 \\ Cola_2 \end{array} & \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \end{array}$$



- Someone has bought $\boldsymbol{Cola_2}$, how likely she'll buy $\boldsymbol{Cola_1}$after 2 times?
- Someone has bought $\boldsymbol{Cola_1}$, how likely she'll buy $\boldsymbol{Cola_1}$ again after 3 times?
- At some time 60% of clients bought $\boldsymbol{Cola_1}$ and 40% $\boldsymbol{Cola_2}$. After three purchases what's the percentage of people buying $\boldsymbol{Cola_1}$?

The Cola Exmple (2)

Someone has bought $\boldsymbol{Cola_2}$, how likely she'll buy $\boldsymbol{Cola_1}$ after 2 times?

$$P(X_2 = 1 | X_0 = 2) = P_{21}\ (2)$$

$$P(2) = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

Someone has bought $\boldsymbol{Cola_1}$, how likely she'll buy $\boldsymbol{Cola_1}$ again after 3 times?

$$P(X_3 = 1 | X_0 = 1) = P_{11}\ (3)$$

$$P(1) = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

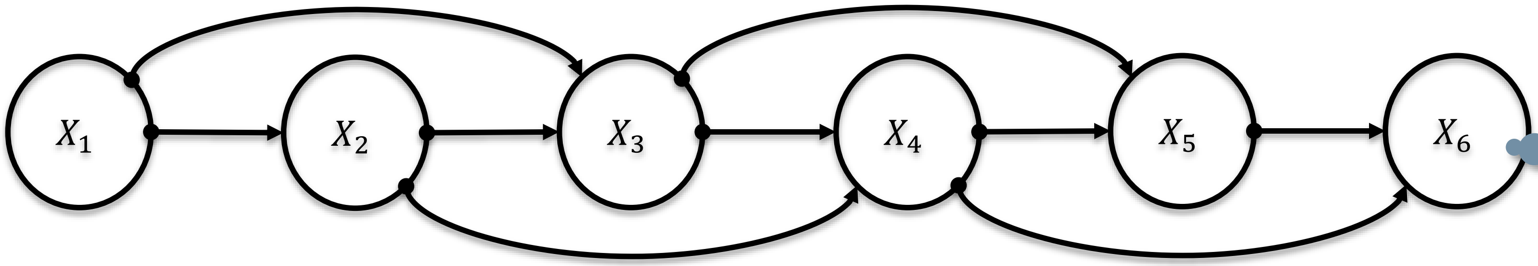
The Cola Example (3)

Suppose at some time 60% of clients bought $\boldsymbol{Cola_1}$ and 40% $\boldsymbol{Cola_2}$. After three purchases what's the percentage of people buying $\boldsymbol{Cola_1}$?

$$\sum_i q_i \cdot P_{ij}(3) = q \cdot (column\ 1\ of\ P^3)$$

$$p = [0.60 \quad 0.40] \begin{bmatrix} 0.781 \\ 0.438 \end{bmatrix} = 0.6438$$

<u>Note:</u> we have see so far first-order Markov Chain. More generally, in $\boldsymbol{k^{th}}$ Markov Chain, each state transition depends on previous $\boldsymbol{k}$ states.



What's the size of the transition matrix?

A Bunch of Definitions

Given a Markov Chain we define:

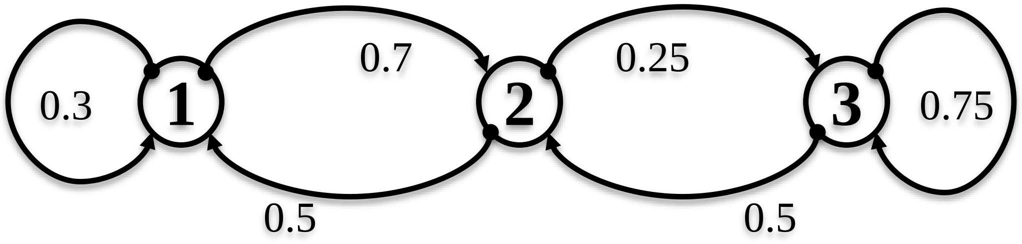
- State j is reachable from i if it exist a path from i to j
- States i and j communicate if i is reachable from j and viceversa
- A set of states S is closed if no state outside S is reachable from a state in S
- A state i is an absorbing state if $p_{ii} = 1$
- A state i is transient if exists j reachable from i, but i is not reachable from j
- A state that is not transient is defined as recurrent
- A state i is periodic with period $k > 1$ if k is the biggest number that divides the length of all path from i to i, a state that is not periodic is said a-periodic

If all states in a Markov Chain are recurrent, a-periodic, and communicate with each other, it is said to be ***<u>Ergothic</u>***

Examples of Ergothic Markov Chains

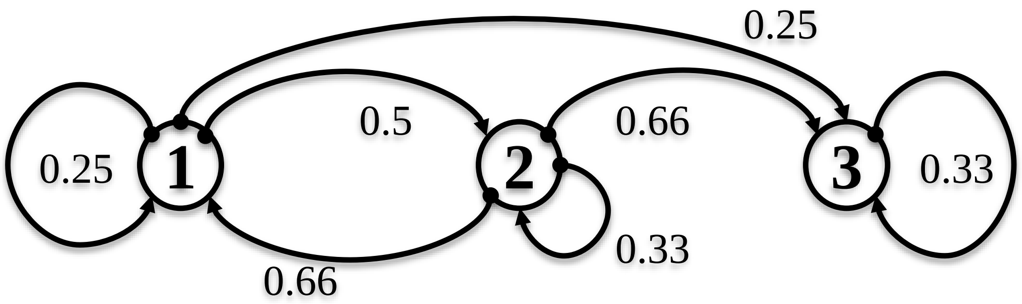
A simple example of Ergothic Markov Chain is the following:

$$P = \begin{bmatrix} 0.3 & 0.7 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 0.25 & 0.75 \end{bmatrix}$$

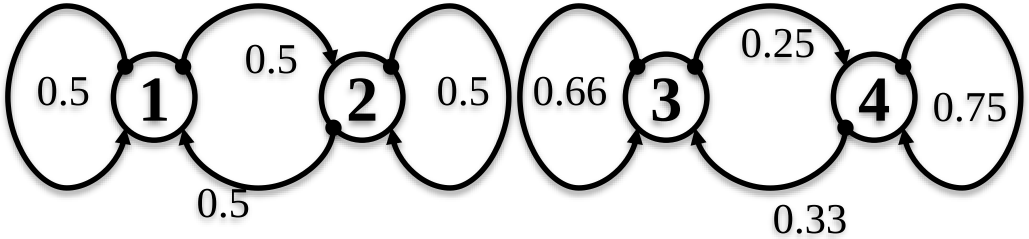


Do the following transitions represent Ergothic Markov Chains?

$$P = \begin{bmatrix} 1/4 & 1/2 & 1/4 \\ 2/3 & 1/3 & 0 \\ 0 & 2/3 & 1/3 \end{bmatrix}$$



$$P = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 2/3 & 1/3 \\ 0 & 0 & 1/4 & 3/4 \end{bmatrix}$$



Steady State Distribution

Being P the transition matrix of an Ergothic Markov Chain with N states:

$$\lim_{n\to\infty} P_{ij}(n) = \pi_j$$

with $\pi = [\pi_1, \pi_2, \dots, \pi_N]$ being the Steady State Distribution

The Cola Example:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$\lim_{n\to\infty} P(n) = \pi = \begin{bmatrix} 0.67 & 0.33 \\ 0.67 & 0.33 \end{bmatrix}$$

n	$p_{11}(n)$	$p_{12}(n)$	$p_{21}(n)$	$p_{22}(n)$
1	.90	.10	.20	.80
2	.83	.17	.34	.66
3	.78	.22	.44	.56
5	.72	.28	.56	.44
10	.68	.32	.65	.35
20	.67	.33	.67	.33
30	.67	.33	.67	.33
40	.67	.33	.67	.33

Transitory Behavior

The behavior of a Markov Chain before getting to the Steady State is defined transitory; we can compute the expected number of transition to reach state $\boldsymbol{j}$ being in state $\boldsymbol{i}$ for an Ergothic Markov Chain as:

$$m_{ij} = p_{ij}(1) + \sum_{k \neq j} p_{ik} \cdot (1 + m_{kj}) = 1 + \sum_{k \neq j} p_{ik} m_{kj}$$

The Cola Example:

- How many bottle on average $\boldsymbol{Cola_1}$ buyer will have before switching to $\boldsymbol{Cola_2}$?

$$m_{12} = 1 + \sum_{k \neq j} p_{1k} m_{k2} = 1 + p_{11} m_{12} = 1 + 0.9 * m_{12} = \frac{1}{1 - 0.9} = 10$$

- What about viceversa?

$$m_{21} = 1 + \sum_{k \neq j} p_{2k} m_{k1} = 1 + p_{22} m_{21} = 1 + 0.8 * m_{21} = \frac{1}{1 - 0.8} = 5$$

Why Should I Care All This Crazy Math?

"Nice, but unless I want to gamble why should I care? I'm a computer engineer what this has to do with practical intelligent systems?"

Assume a link from page A to page B is a recommendation of page B by the author of A (we say B is successor of A).

- The quality of a page is related to its in-degree.
- The quality of a page is related to the quality of pages linking to it

This recursively defines the PageRank of a page [Brin & Page '98]

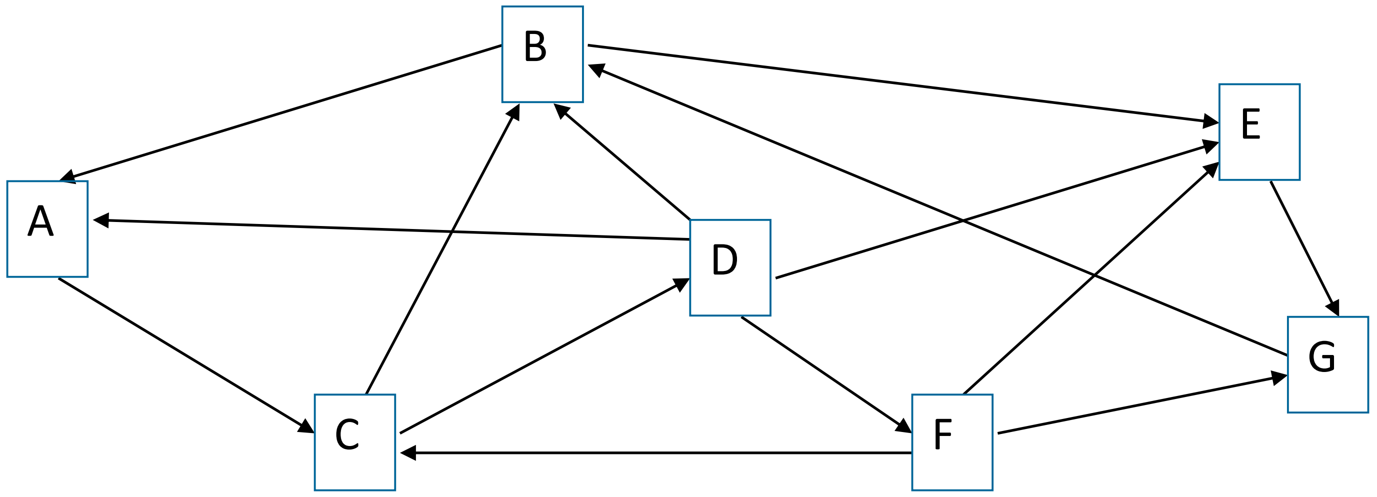
For a (better) detailed description feel free to read:
http://www-db.stanford.edu/~backrub/google.html
http://www.iprcom.com/papers/pagerank/

Google’s PageRank

Suppose the web is an Ergothic Markov Chain, and browsing is an infinite random walk (surfing):

- Initially the surfer is at a random page
- At each step, the surfer proceeds
  - to a randomly chosen web page with probability d
  - to a randomly chosen successor of the current page with probability 1-d

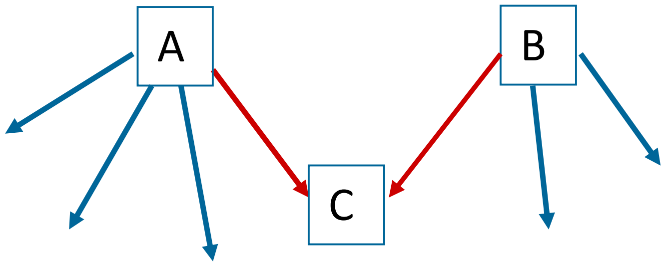
The PageRank of a page is the fraction of steps the surfer spends on it in the limit.



Definition of PageRank

PageRank = the steady state probability for this Markov Chain

$$PageRank(u) = d + (1-d) \sum_{(v,u) \in E} PageRank(v)/outdegree(v)$$



- n is the total number of nodes in the graph
- d is the probability of a random jump

$$PageRank(C) = \frac{d}{n} + (1-d)\left(\frac{1}{4} PageRank(A) + \frac{1}{3} PageRank(B)\right)$$

Summarizes the "web opinion" about the page importance

- Query-independent
- It can be faked … read the provided links if you are curious!

Dealing with Absoring States

We have and absorbing Markov Chain if there exist one or more absorbing states and all the other are transient; its transition matrix is:

$$P = \left[\begin{array}{c:c} Q & R \\ \hdashline 0 & 1 \end{array}\right]$$

What kind of inference with such a model?

where

- Q is the transition matrix for transient states
- R is the transition matrix from transient to absorbing states

Inference in Absorbing Markov Chains

How long do I remain in a transient state starting from a transient one?

- Being in a transient state $\boldsymbol{i}$ the average time spent in a transient state $\boldsymbol{j}$ is the $\boldsymbol{ij^{th}}$ element of $(\boldsymbol{I}-\boldsymbol{Q})^{-1}$

Starting from a transient state, how long does it takes to get to an absorbing one?

- Being in transient state $\boldsymbol{i}$ the probability to get into an absorbing state $\boldsymbol{j}$ is the $\boldsymbol{ij^{th}}$ element of $(\boldsymbol{I}-\boldsymbol{Q})^{-1}\cdot\boldsymbol{R}$

<u>*Example*</u>: in a company there are 3 levels (J, S, P):

- How long does a junior remains in the company?
- What's the probability for a junior to leave the company as partner?

$$
P = \begin{array}{c} \begin{array}{ccccc} J & S & P & LN & LP \end{array} \\ \left[\begin{array}{ccc|cc} 0.80 & 0.15 & 0 & 0.05 & 0 \\ 0 & 0.70 & 0.20 & 0.10 & 0 \\ 0 & 0 & 0.95 & 0 & 0.05 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right] \end{array}
$$

The Company Example

$$(I-Q)^{-1} = \begin{bmatrix} 5 & 2.5 & 10 \\ 0 & 3.3 & 13.3 \\ 0 & 0 & 20 \end{bmatrix} \qquad (I-Q)^{-1} \cdot R = \begin{bmatrix} 0.5 & 0.5 \\ 0.3 & 0.7 \\ 0 & 1 \end{bmatrix}$$

How long does a junior remains in the company?

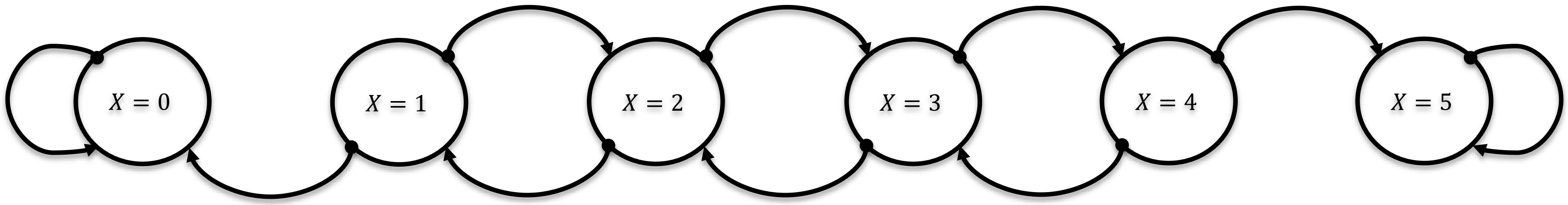
$$\left.\begin{array}{l} \bullet\ \text{He/she will stay as Junior: } m_{11} = 5 \\ \bullet\ \text{He/she will stay as Senior: } m_{12} = 2.5 \\ \bullet\ \text{He/She will stay as Partner: } m_{13} = 10 \end{array}\right\} 17.5\ years$$

What's the probability for a junior to leave the company as partner?

- He/She will end up in state LP: $m_{12} = 0.5$

Exercise: Gambler's Ruin

Suppose you start from a 3$ capital. With probability $\boldsymbol{p = 1/3}$ you can win 1$ and with $\boldsymbol{1 - p = 2/3}$ you loose 1$. You succeed if capital gets 5.



- Possible states: 0, 1, 2, 3, 4, 5
- Transition probability: p(Xt+1=Xt+1)=1/3, p(Xt+1=Xt-1)=2/3

What kind of reasoning can we apply to this model?

- What's the probability of sequence 3, 4, 3, 2, 3, 2, 1, 0?
- What's the probability of success for the gambler?
- What's the average number of bets the gambler will make?

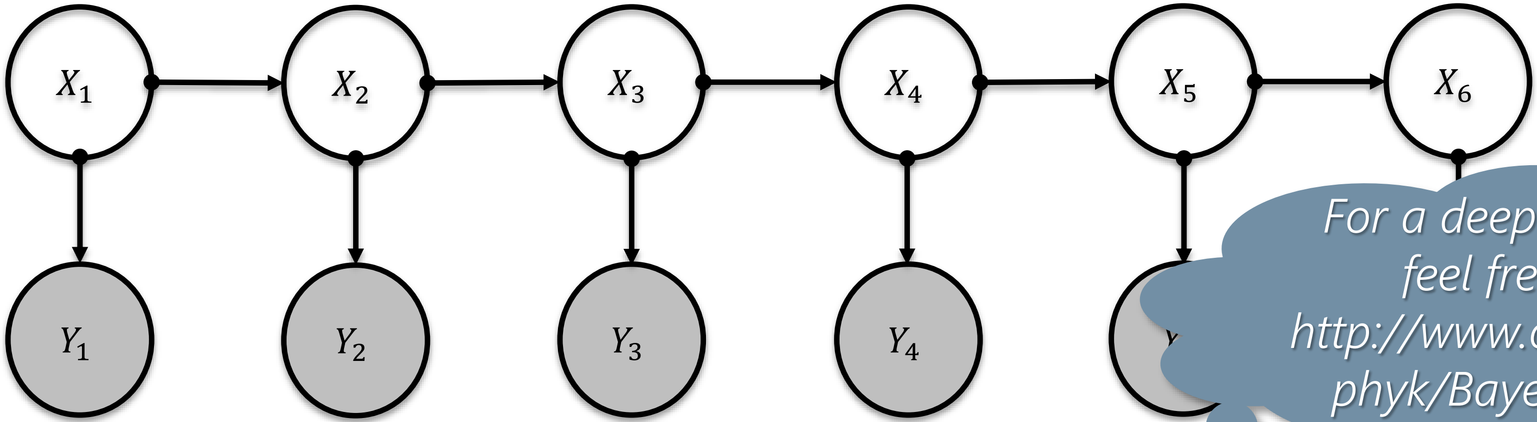
Soft Computing – Probabilistic Reasoning

- Hidden Markov Models -

Prof. Matteo Matteucci – *matteo.matteucci@polimi.it*

Hidden Markov Models (HMM)

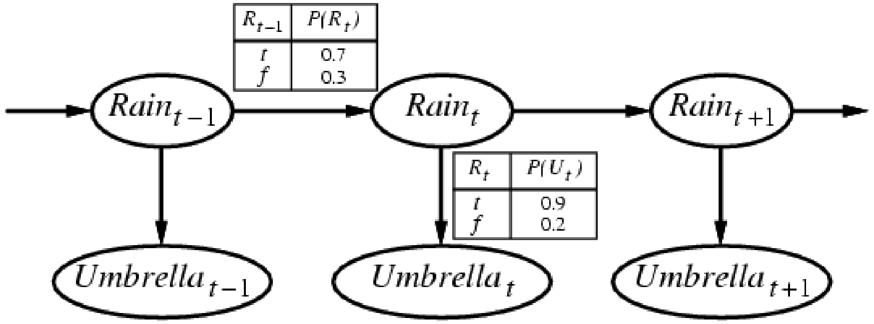
If we may not observe directly the states, we get another Bayesian Network named as Hidden Markov Model (HMM).



For a deeper description feel free to read: http://www.cs.ubc.ca/~murphyk/Bayes/rabiner.pdf

An HMM is described by a quintuple <S,E,P,A,B>

- S : $\{S_1, \dots, S_N\}$ are the values for the hidden states
- E : $\{e_1, \dots, e_T\}$ are the values for the observations
- P: probability distribution of the initial state
- A: transition probability matrix
- B: emission probability matrix

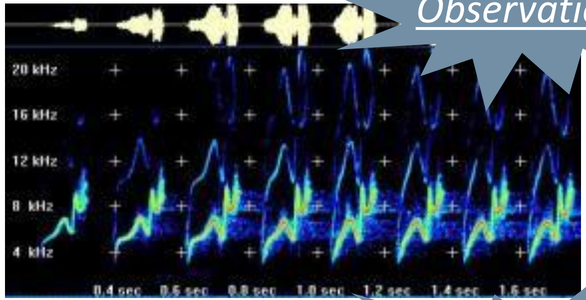


An Example: The Audio Spectrum

Audio Spectrum of the song for the Prothonotary Warbler

State

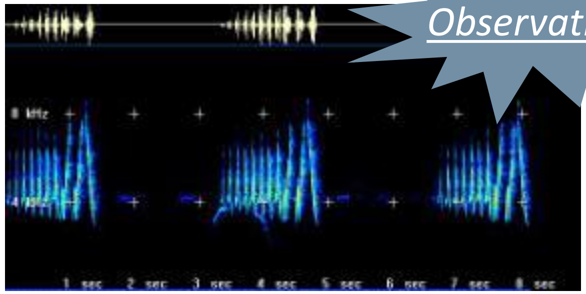
Observation



Audio Spectrum of the song for the Chestnut-sided Warbler

State

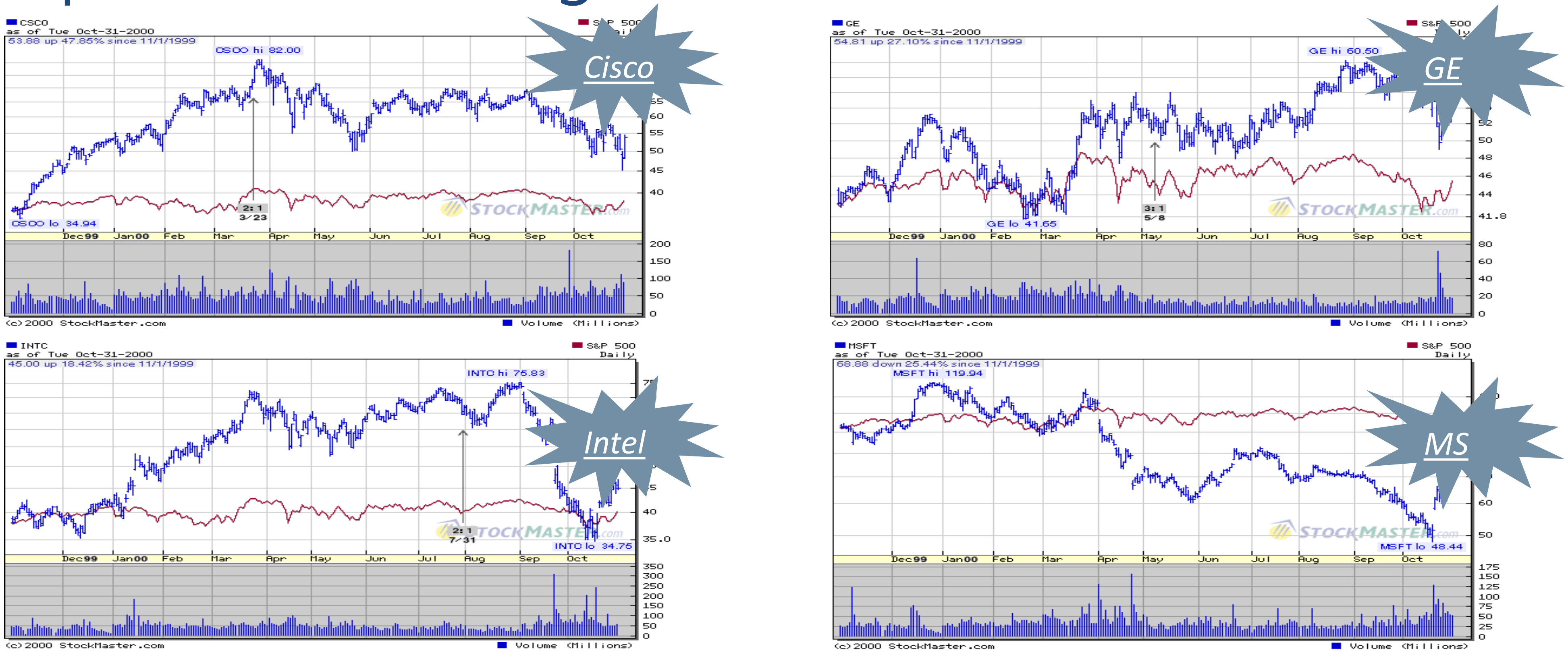
Observation



What can we ask to an HMM?

- What bird is this? → Time Series Classification
- How will the song continue? → Time Series Prediction
- Is this bird sick? → Outlier Detection
- What phases does this song have? → Time Series Segmentation

An Example: Stock Exchange



What can we ask to an HMM?

- What kind of stock is this (e.g., risky)? → Time Series Classification
- Will the stock go up or down? → Time Series Prediction
- Is the behavior abnormal (e.g., bank fraud)? → Outlier Detection

An Example: Music Analysis

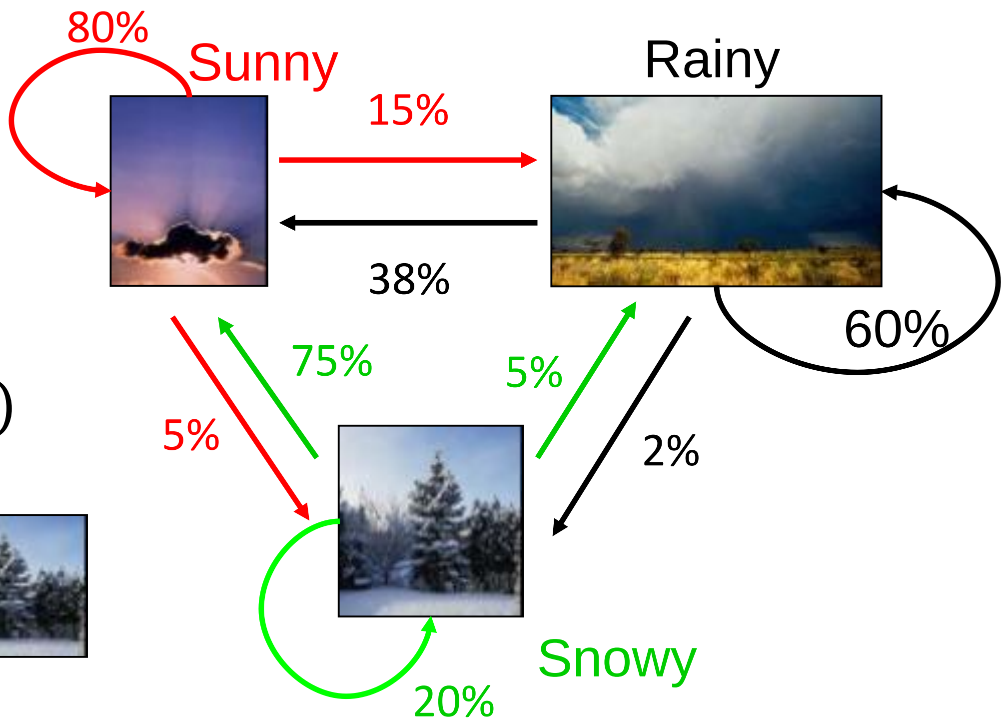
What can we ask to an HMM?

- Is this Beethoven or Bach? → Time Series Classification
- Can we compose more of that? → Time Series Prediction
- Can we segment it into themes? → Time Series Segmentation

Weather: A Markov Chain Model

States: $\{S_{sunny}, S_{rainy}, S_{snowy}\}$

State transitions : $P = \begin{bmatrix} 0.80 & 0.15 & 0.05 \\ 0.38 & 0.60 & 0.02 \\ 0.75 & 0.05 & 0.20 \end{bmatrix}$



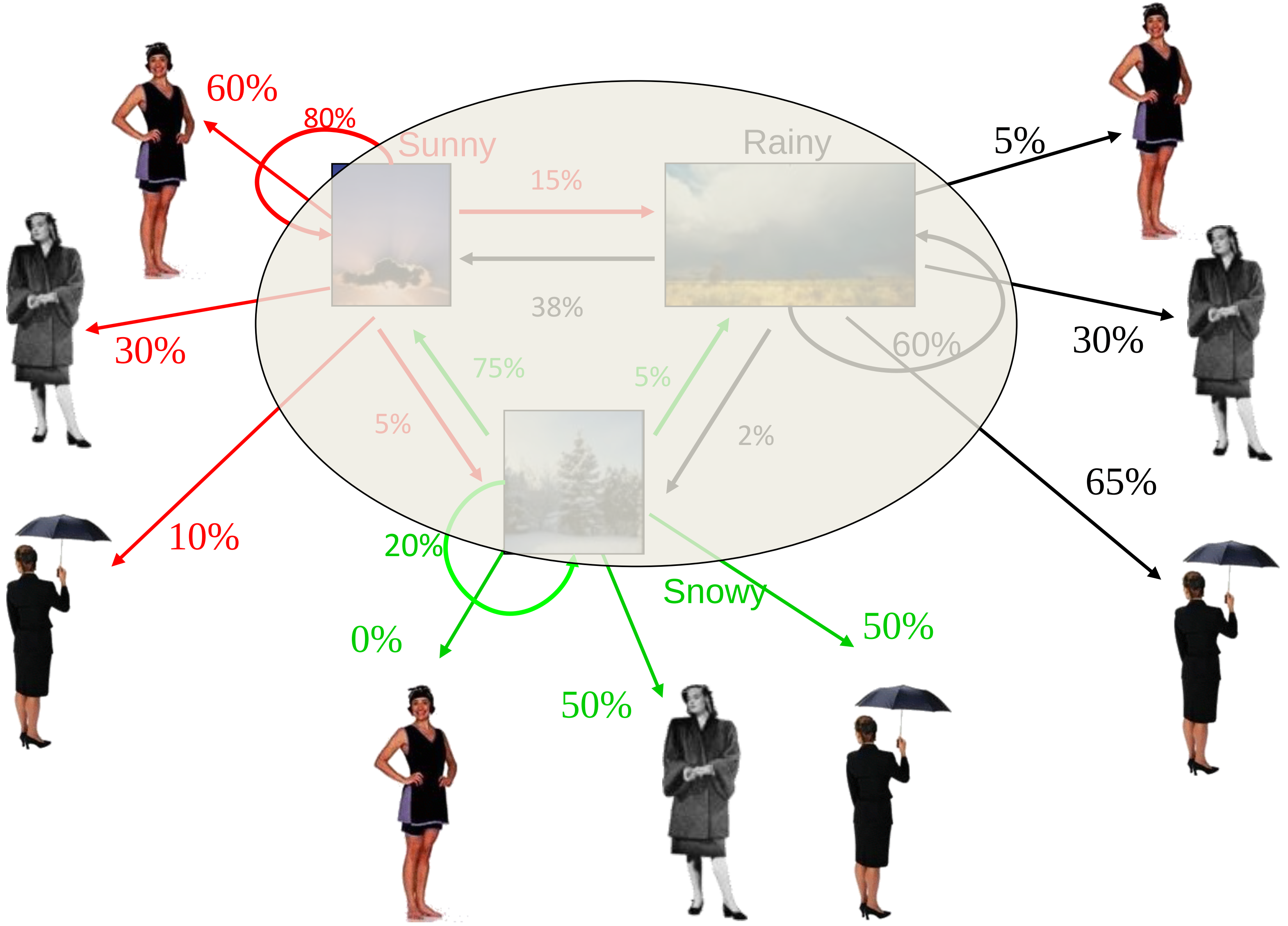
Initial state distribution: $q = (0.7 \;\; 0.25 \;\; 0.05)$

Given:

What is the probability of this series?

$$P(S) = p(S_{sunny}) \cdot p(S_{rainy}|S_{sunny}) \cdot p(S_{rainy}|S_{rainy}) \cdot p(S_{rainy}|S_{rainy})$$
$$\cdot\, \mathrm{p}(S_{snowy}|S_{rainy}) \cdot p(S_{snowy}|S_{snowy}) = 0.7 \cdot 0.15 \cdot 0.6 \cdot 0.6 \cdot 0.02 \cdot 0.2 = 0.0001512$$

Weather: An Hidden Markov Models



HMM Ingredients and Fundamental Questions

States: $\{S_{sunny}, S_{rainy}, S_{snowy}\}$

Observations: $\{O_{shorts}, O_{coat}, O_{umbrella}\}$

State transition probabilities: $A = \begin{bmatrix} 0.80 & 0.15 & 0.05 \\ 0.38 & 0.60 & 0.02 \\ 0.75 & 0.05 & 0.20 \end{bmatrix}$

Observation probabilities: $B = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.05 & 0.30 & 0.65 \\ 0.00 & 0.50 & 0.50 \end{bmatrix}$

Initial state distribution: $q = (0.7 \;\; 0.25 \;\; 0.05)$

Given:

...

How can I learn the structure of this HMM?

How can I learn the HMM parameters?

What is the underlying sequence of states?

What is the probability of this series?

Computing Forward Probability

Forward Probability is the join probability of actual state and observations

$$P(X_t = s_i, e_{1:t})$$

Why are we interested in forward probability?

- Probability of observations: $P(e_{1:t})$
- Prediction: $P(X_{t+1} = s_i \mid e_{1:t}) = ?$

No panic! It is just message passing after all …

Same form, use recursion!

$$\overbrace{P(X_t = s_i, e_{1:t})}^{\alpha_i(t)} = P(X_t = s_i, e_{1:t-1}, e_t) = \sum_j P(X_{t-1} = s_j, X_t = s_i, e_{1:t-1}, e_t) =$$

$$= \sum_j P(e_t | X_t = s_i, X_{t-1} = s_j, e_{1:t-1}) P(X_t = s_i, X_{t-1} = s_j, e_{1:t-1}) =$$

$$= \sum_j P(e_t | X_t = s_i) P\big(X_t = s_i \big| X_{t-1} = s_j, e_{1:t-1}\big) P\big(X_{t-1} = s_j, e_{1:t-1}\big) =$$

$$= \sum_j P(e_t | X_t = s_i) P\big(X_t = s_i \big| X_{t-1} = s_j\big) P\big(X_{t-1} = s_j, e_{1:t-1}\big) = \sum_j A_{ij} B_{je_t} \overbrace{P\big(X_{t-1} = s_j, e_{1:t-1}\big)}^{\alpha_j(t-1)}$$

The Viterbi Algorithm (1)

From observations, compute the most likely hidden state sequence:

$$argmax\ P(X_{1:t}|e_{1:t}) = argmax\ P(X_{1:t}, e_{1:t})/P(e_{1:t}) = argmax\ P(X_{1:t}, e_{1:t})$$

By applying the Bayesian Network factorization

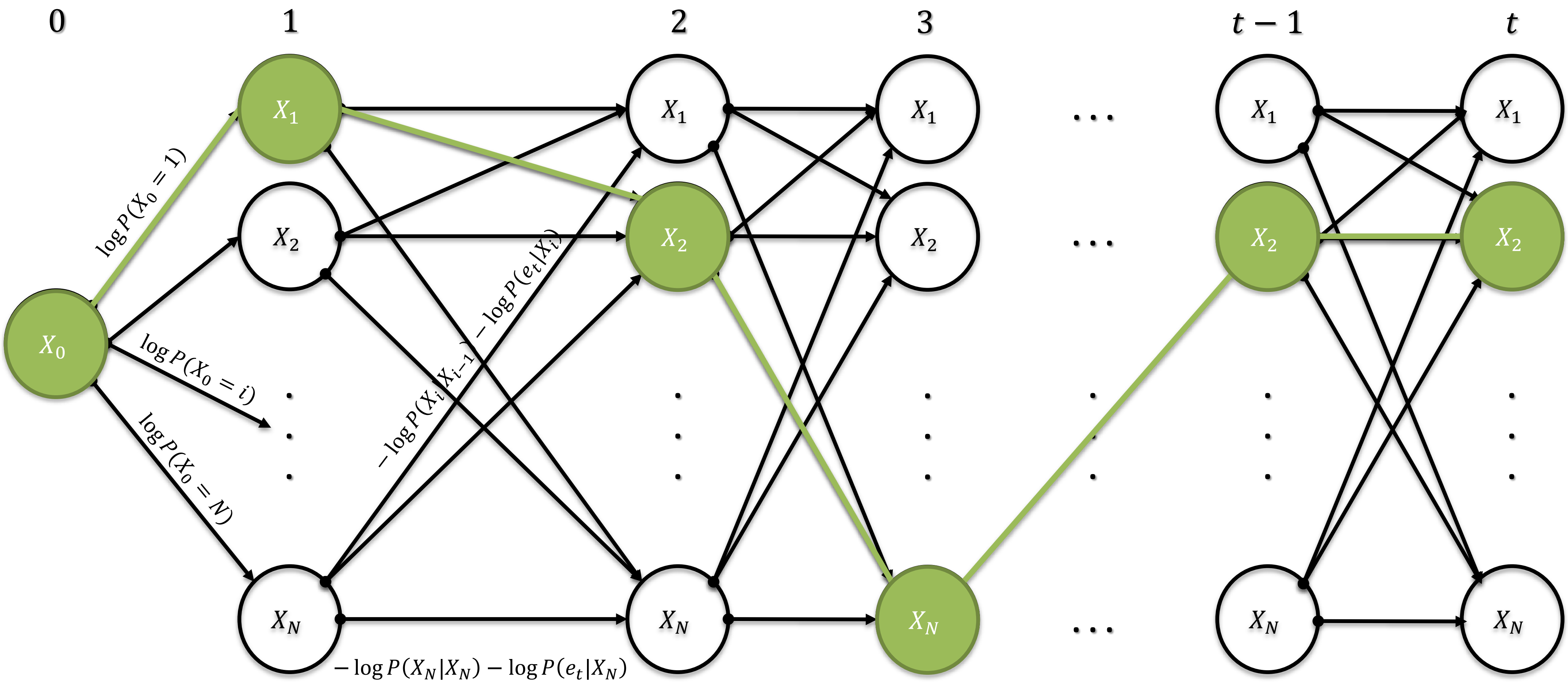
$$P(X_{1:t}, e_{1:t}) = P(X_0) \prod_{i=1:t} P(X_i|X_{i-1})P(e_t|X_i)$$

The solution we are looking for is the one that minimizes

$$-\log P(X_{1:t}, e_{1:t}) = -\log P(X_0) + \sum_{i=1:t} (-\log P(X_i|X_{i-1}) - \log P(e_i|X_i))$$

Construct a graph that consists $1 + t \cdot N$ nodes, one initial node and N node at time i where j^{th} represents $X_i = s_j$.

The Viterbi Algorithm (2)



Harmonising Chorales by Probabilstic Inference

Moray Allan & Chris Williams (NIPS 2004) used an HMM

- Observed sequence $Y_{0:T}$ Soprano melody
- Latent sequence $X_{0:T}$ chord & and harmony

Figure 2: Most likely harmonisation under our model of chorale K4, BWV 48

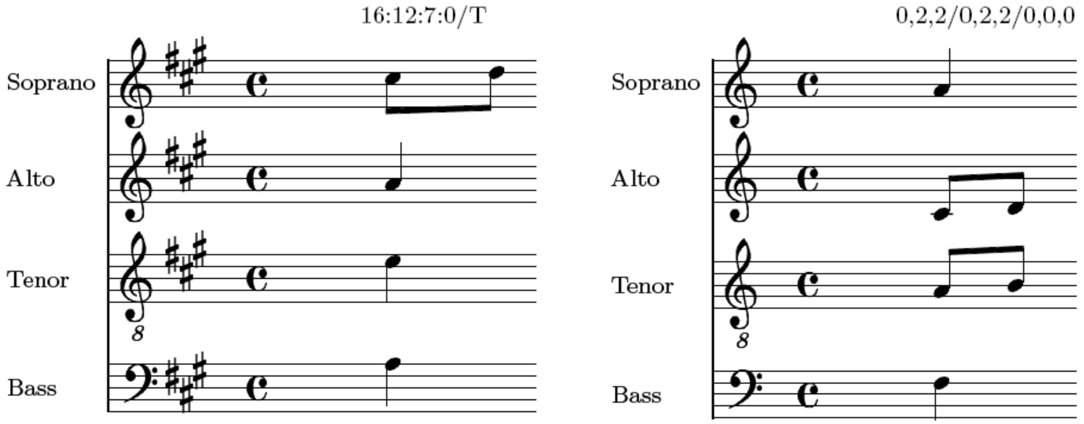
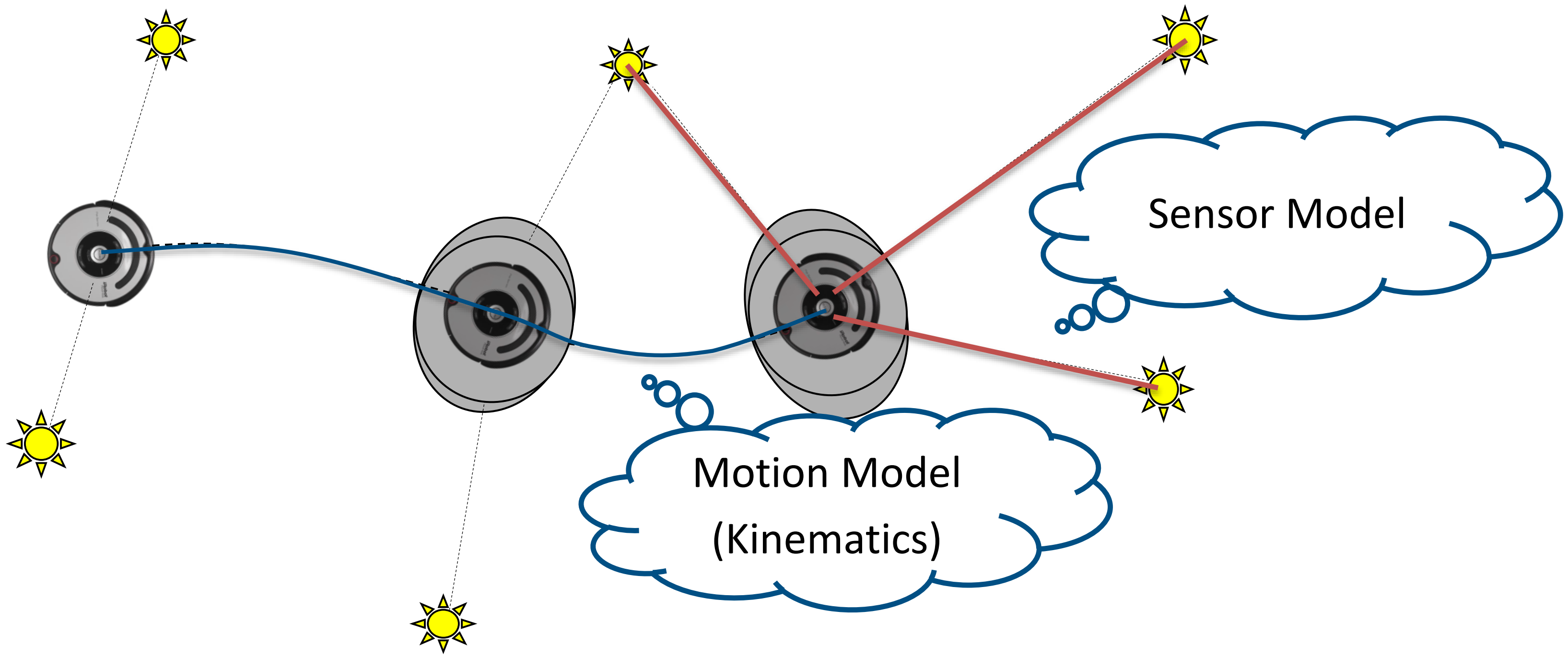
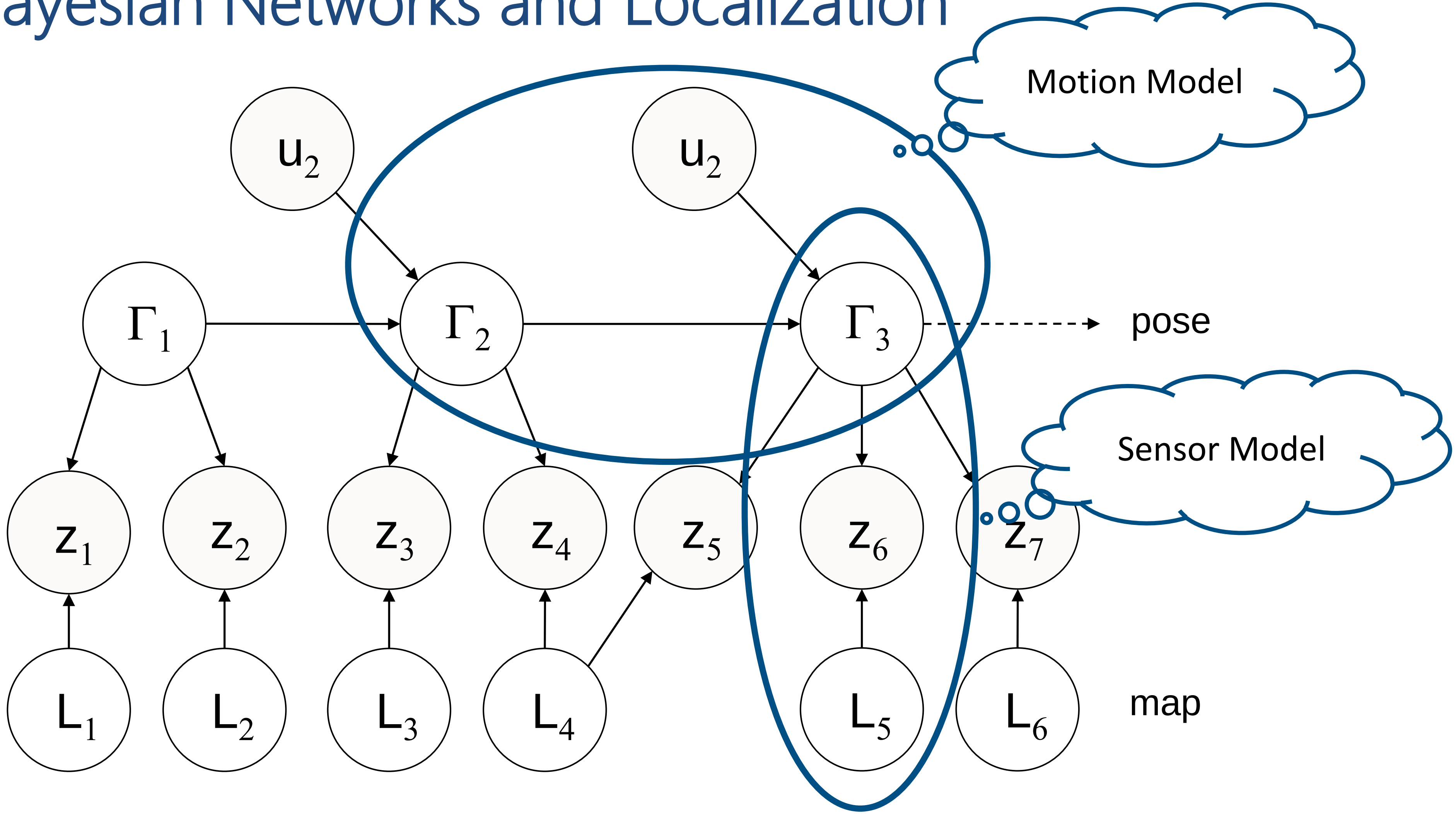


Figure 1: Hidden state representations (a) for harmonisation, (b) for ornamentation.

Localization with Knowm Map

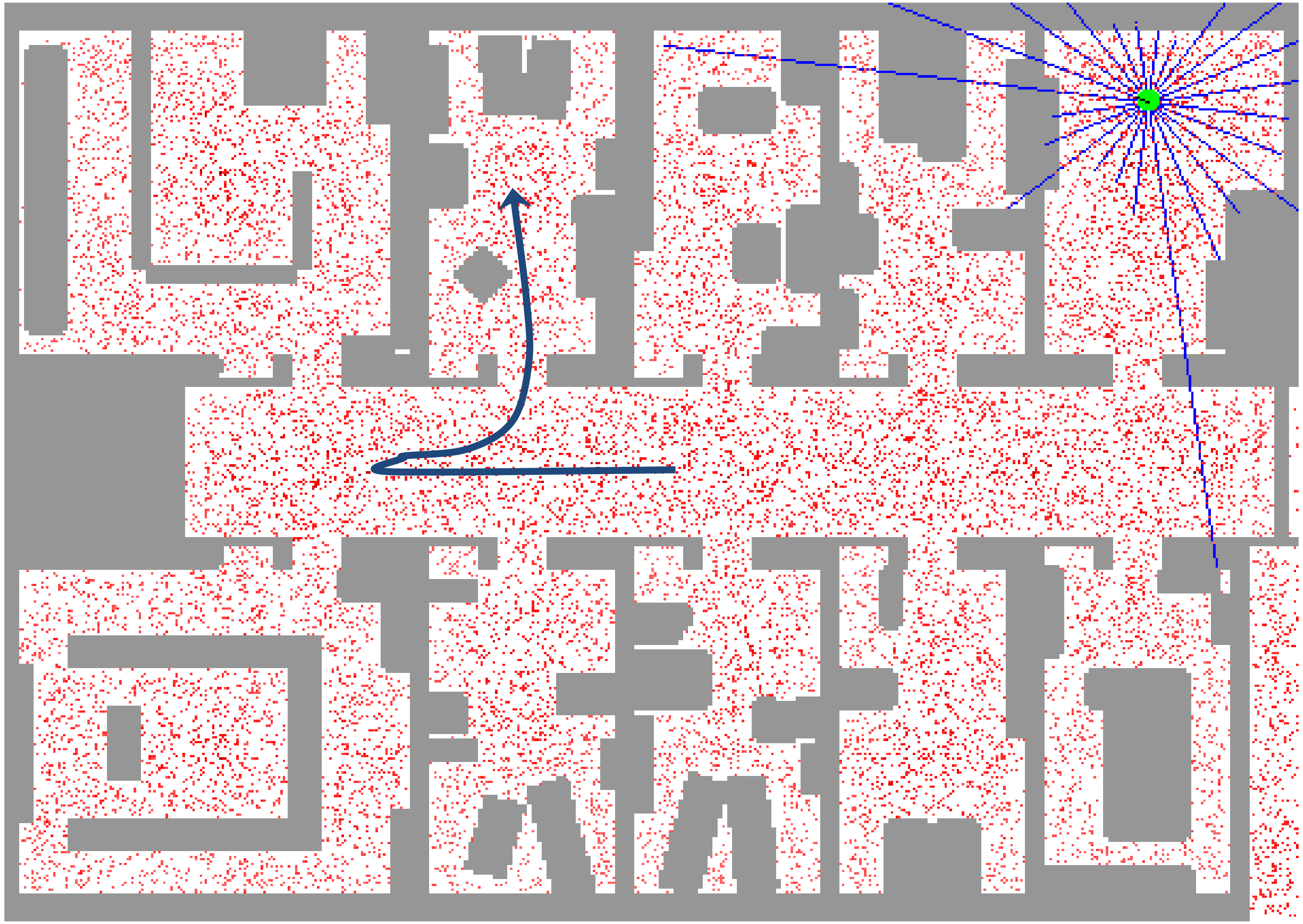


Dynamic Bayesian Networks and Localization

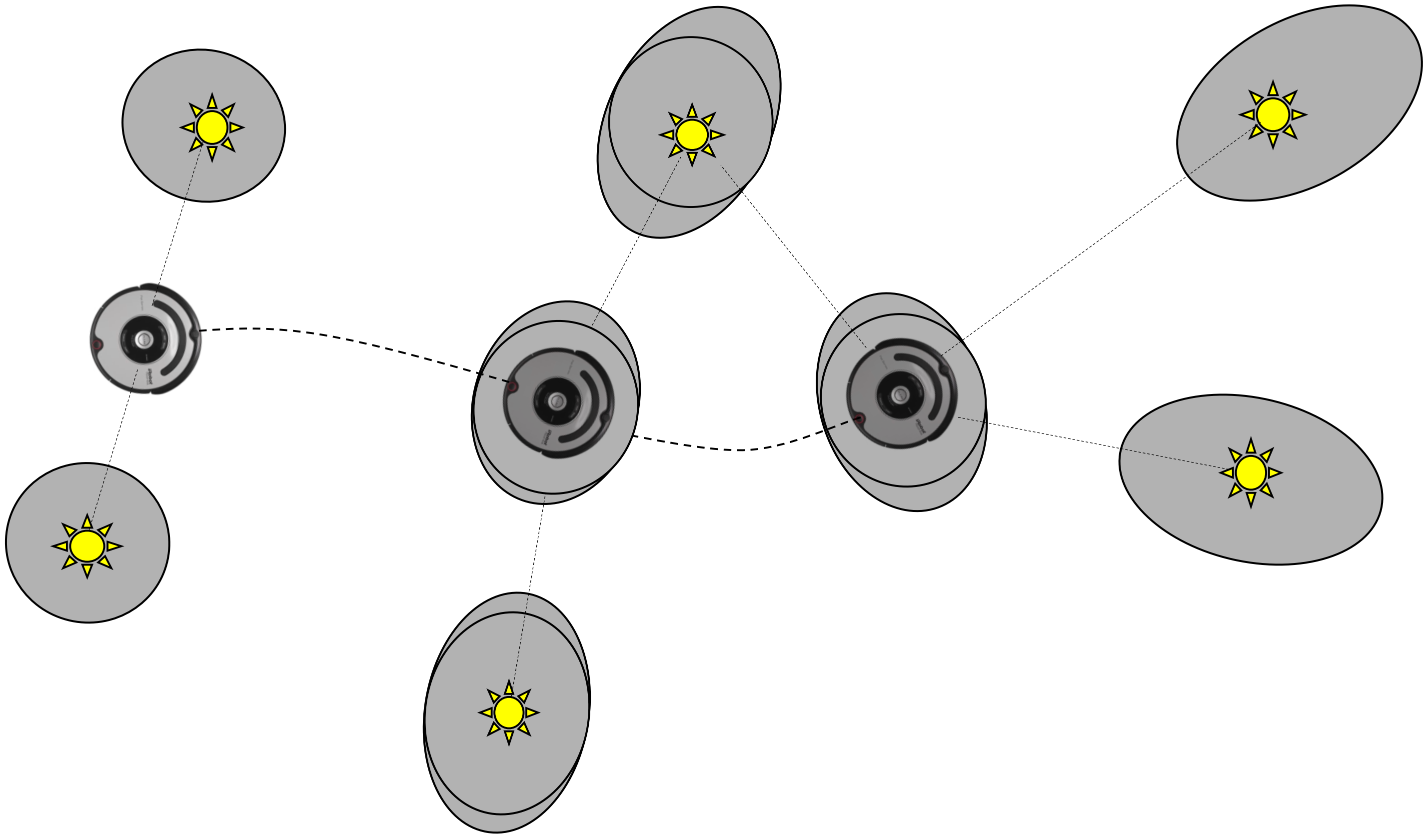


$$\text{Filtering:} \quad p(\Gamma_t \mid Z_{1:t}, U_{1:t}, l_1, \ldots, l_N) = \iiint_{1:t-1} p(\Gamma_{1:t} \mid Z_{1:t}, U_{1:t}, l_1, \ldots, l_N)$$

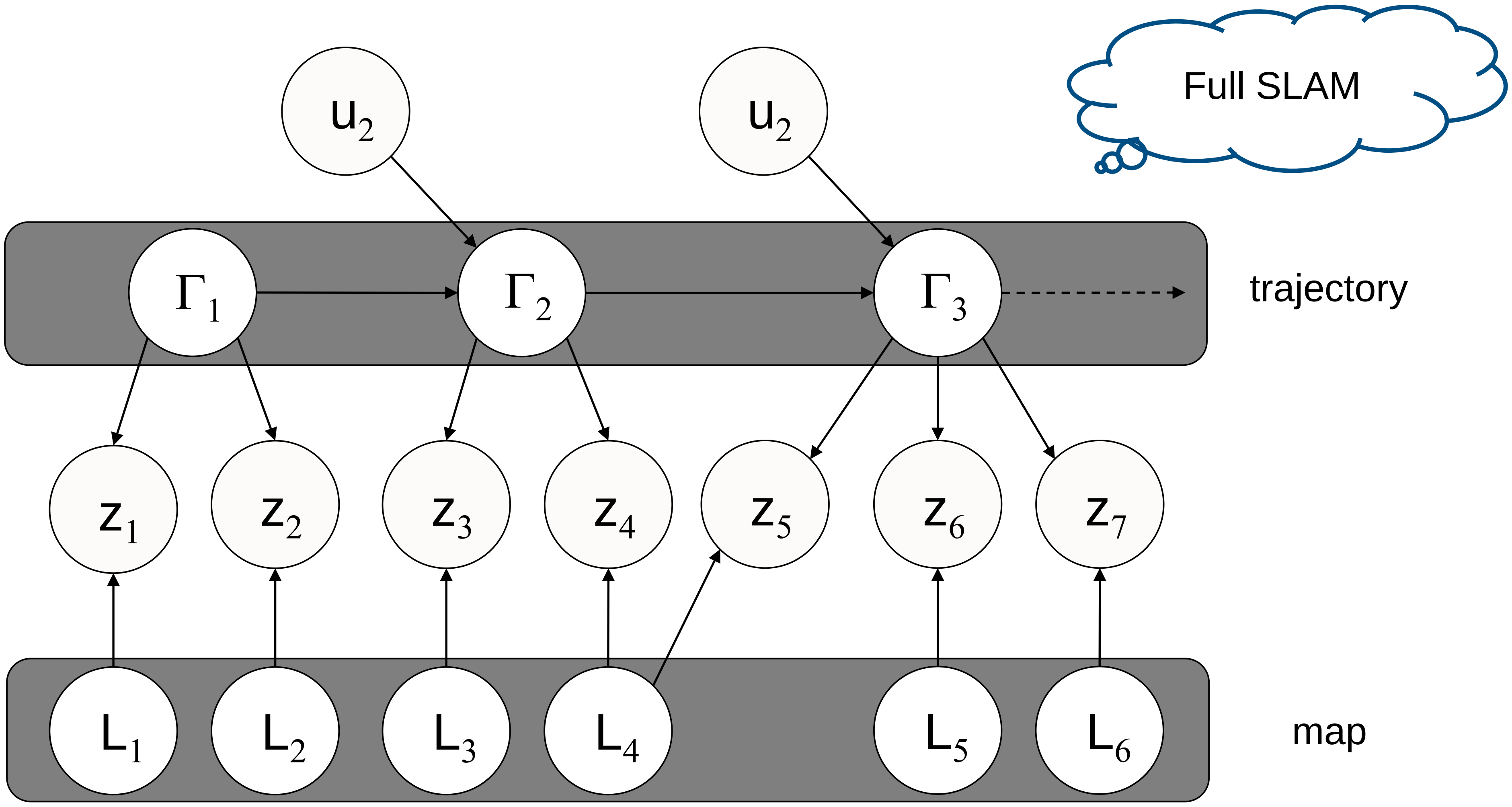
Sample-based Localization (sonar)



Simultaneous Localization and Mapping

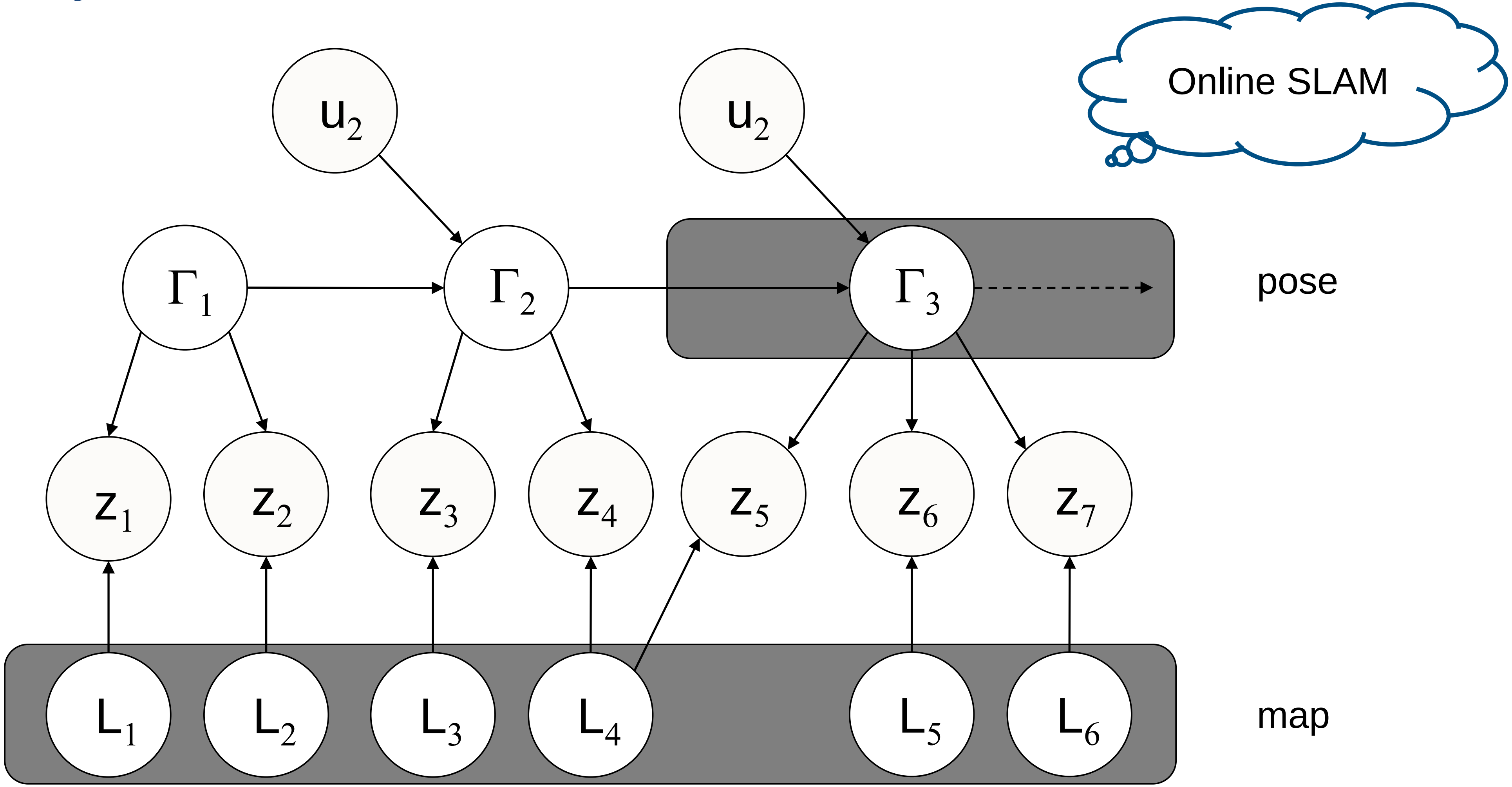


Dynamic Bayesian Networks and (Full) SLAM



Smoothing : $p(\Gamma_{1:t}, l_1, \ldots, l_N \mid Z_{1:t}, U_{1:t})$

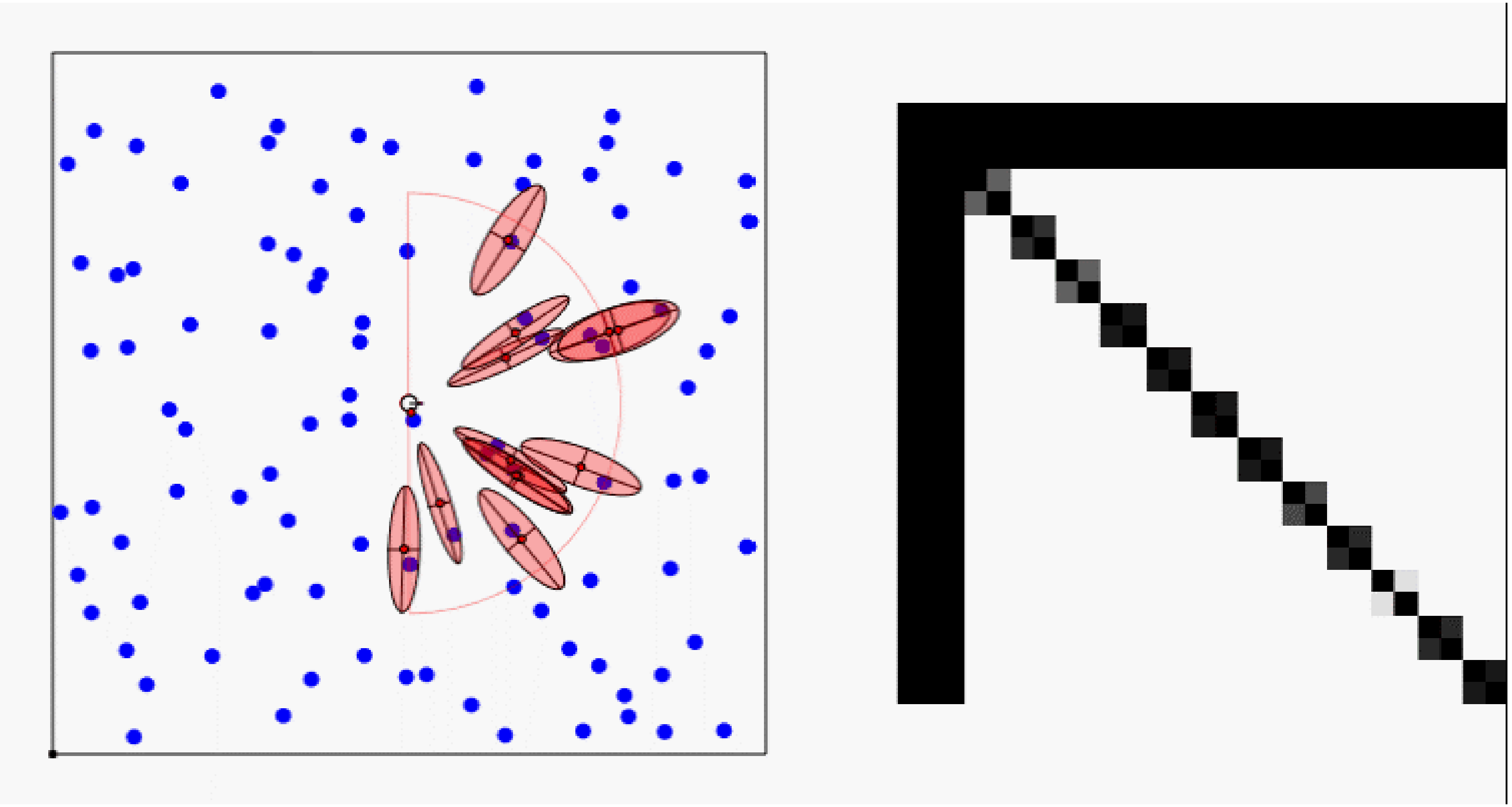
Dynamic Bayesian Networks and (Online) SLAM



$$\text{Filtering} \quad : \quad p(\Gamma_t, l_1, \ldots, l_N \mid Z_{1:t}, U_{1:t}) = \iiint_{1:t-1} p(\Gamma_{1:t}, l_1, \ldots, l_N \mid Z_{1:t}, U_{1:t})$$

Classical Solution – The Extended Kalman Filter

Approximate the SLAM posterior with a high-dimensional Gaussian



Blue path = true path **Red path** = estimated path **Black path** = odometry

Monte Carlo (Fast-SLAM) Example