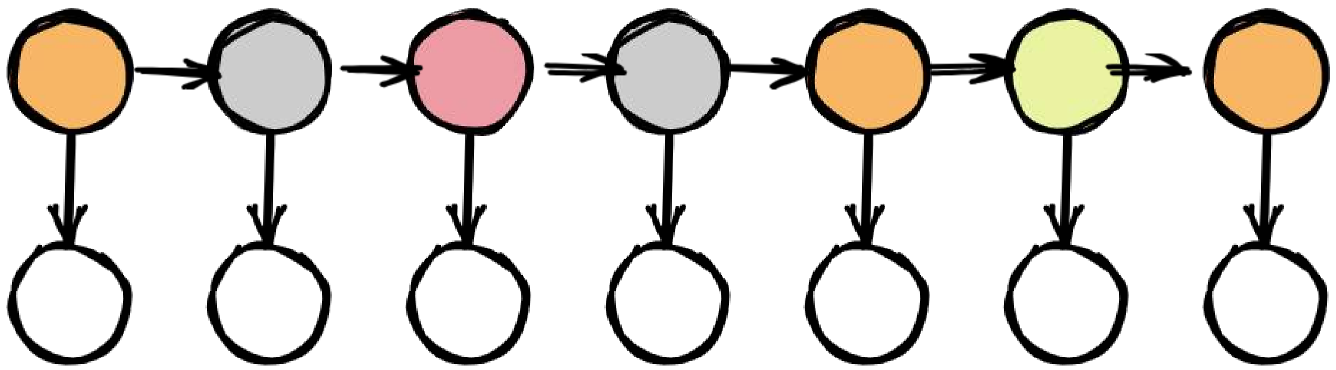
Hidden Markov Models and Sequence Data



Week 18

Middlesex University Dubai; CST4050 Fall21;
Instructor: Dr. Ivan Reznikov

Plan

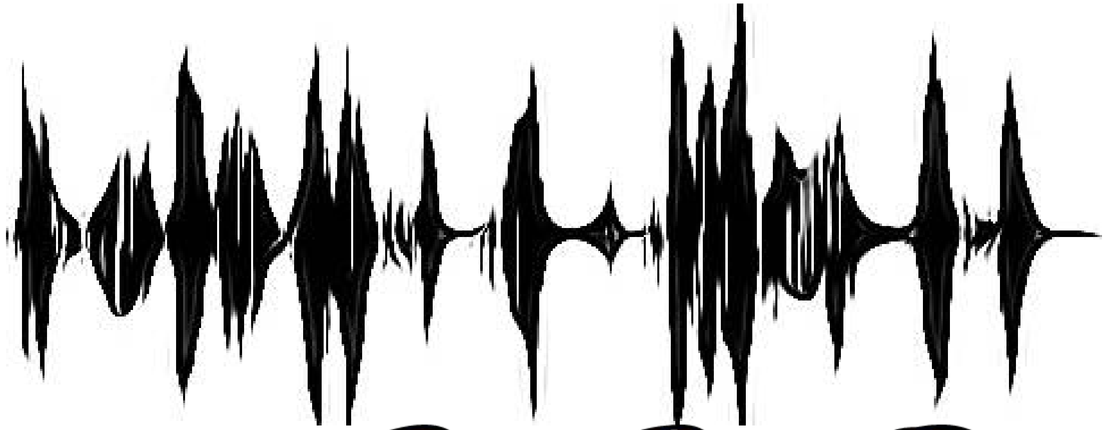
- Sequential Data
- Sequential Labeling
- Bayesian Networks
- Mixture Models
- Markov assumption
- Hidden Markov Model

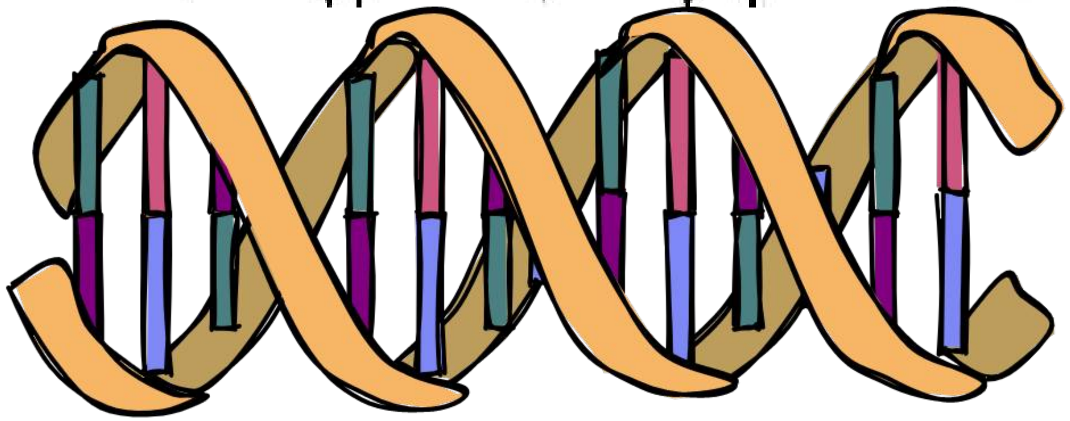
What is sequence data?

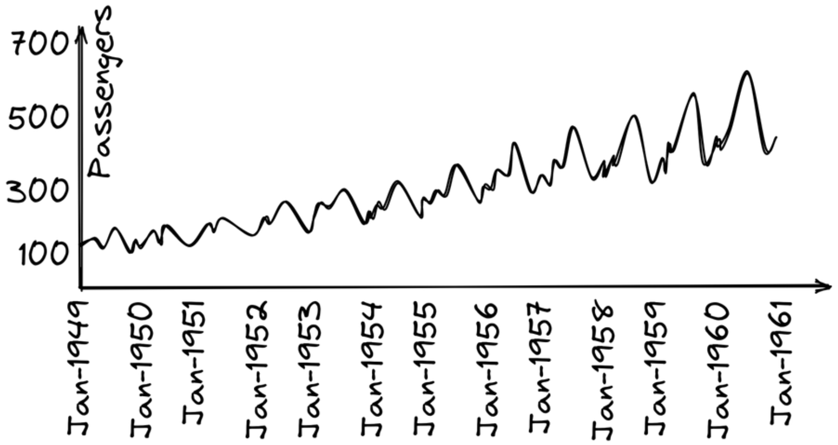
- Ordered set of elements: $x = x_1, x_2, \ldots x_N$
- Order determined by time or position and could be regular or irregular
- Each element x_i could be
  - Numerical (sales, stock price, etc.)
  - Categorical (weather, part-of-speech)
  - Multiple attributes
- The length N of a sequence isn't fixed

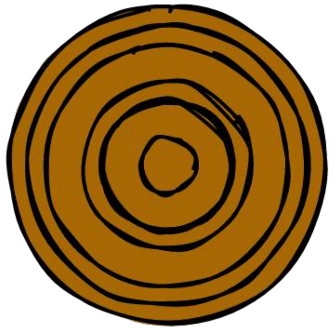
Examples of sequence data

- Speech (sequence of phonemes)
- Language-related (sequence of words)
- Bioinformatics (genes – sequence of 4 possible nucleotides and proteins – sequence of 20 possible amino-acids)
- Telecommunications (sequence of data packets)
- Time series (sequence of events per time)
- ...









Sequence labeling

Address:

221B Baker Street, London, UK

House number Street
City Country

Citation:

Pauling, L. (1931). The nature of the chemical bond. II. The one-electron bond and the three-electron bond. Journal of the American Chemical Society, 53(9), 3225-3237.

Author Year
Article title Journal
Journal number
Volume Pages

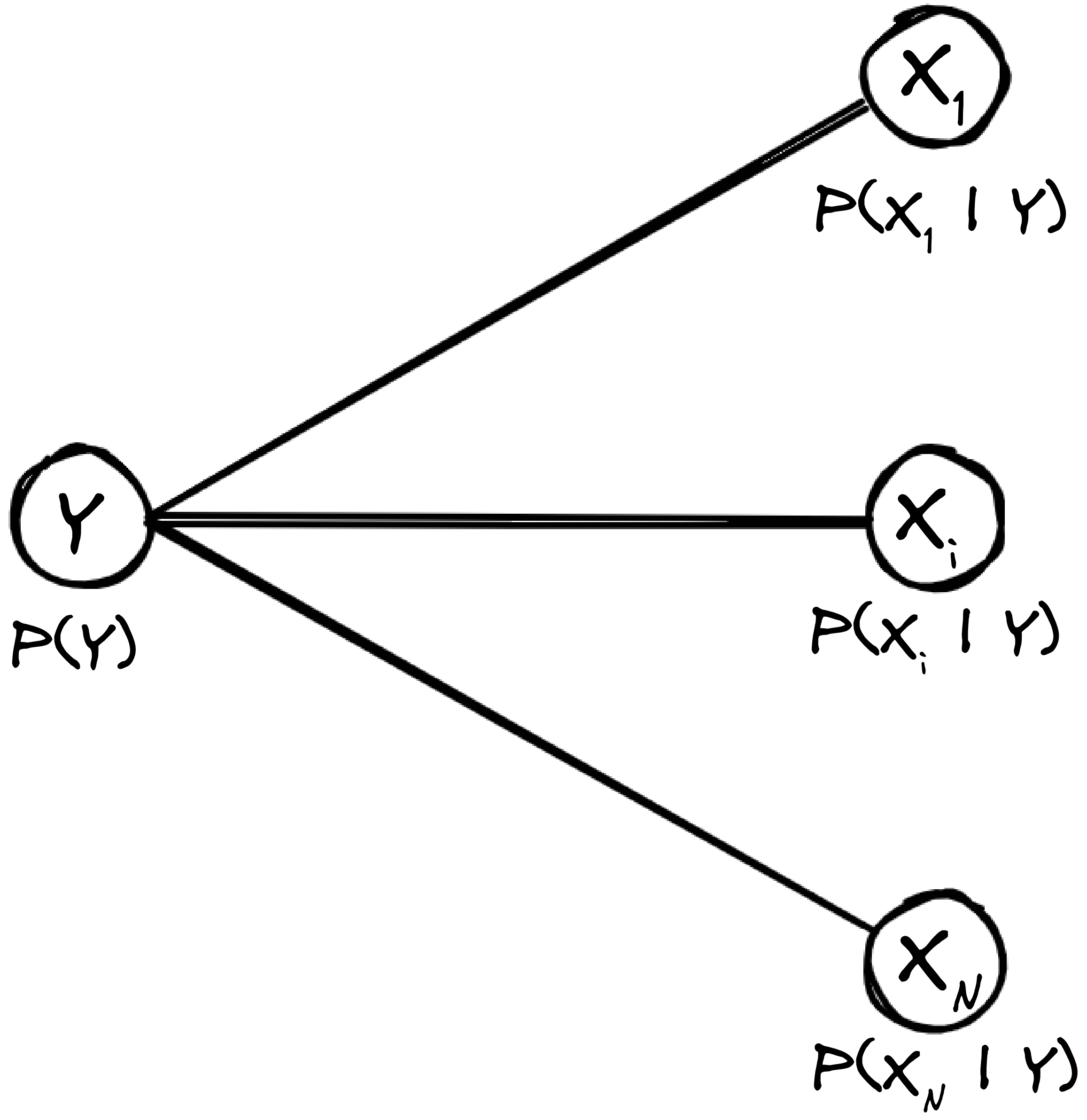
Input: a sequence $x = (x_1, \ldots x_n)$

Output: a sequence $y = (y_1, \ldots y_n)$, where y_i is a label for x_i

Graphical model

Let's assume we have a condition Y. There are several X, that can occur with Y happening. We can draw represent our graph as a probability tree:

- Edges showing dependencies
- Each node has associated conditional
- Probability distribution, conditioned on its parent nodes
- Nodes are independent



$$P(X_1, X_2, \ldots X_N, Y) = P(X_1 | Y) \times P(X_2 | Y) \ldots P(X_N | Y) \times P(Y)$$

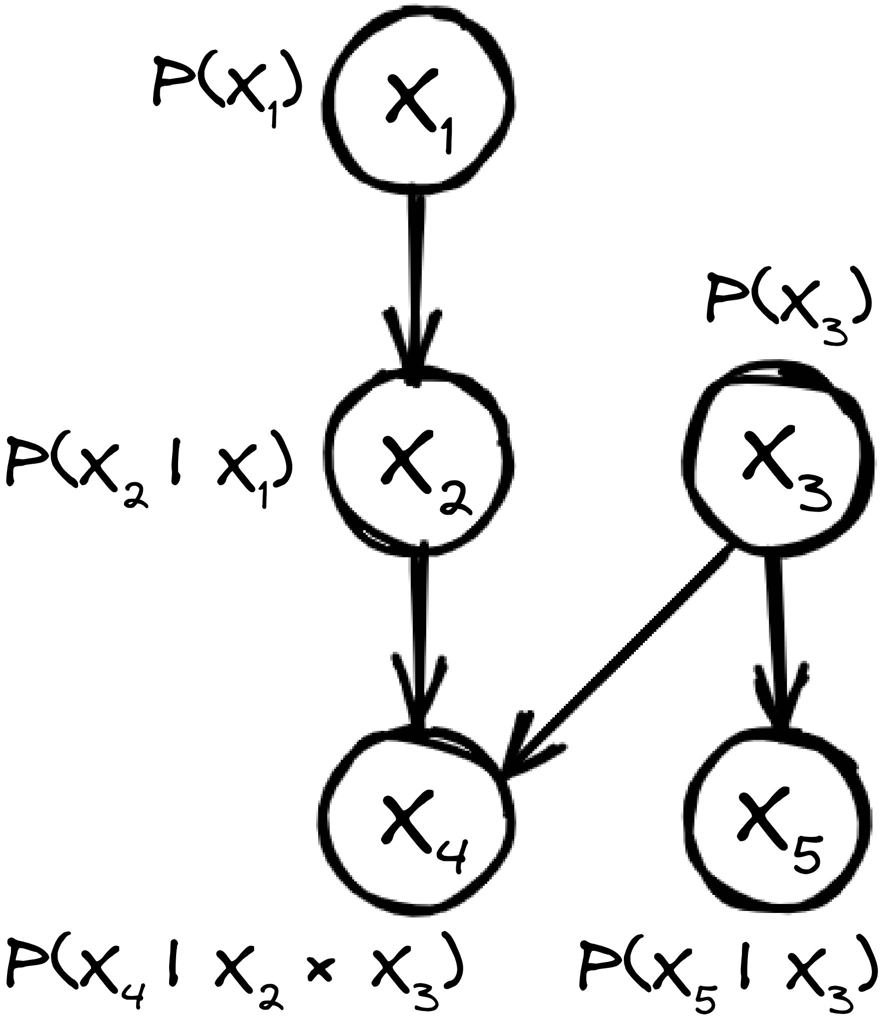
Graphical model

Let's now draw a directed graph out of 5 nodes.

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_5 \mid X_3) \times P(X_4 \mid X_2 \times X_3) \times P(X_2 \mid X_1) \times P(X_3) \times P(X_1)$$

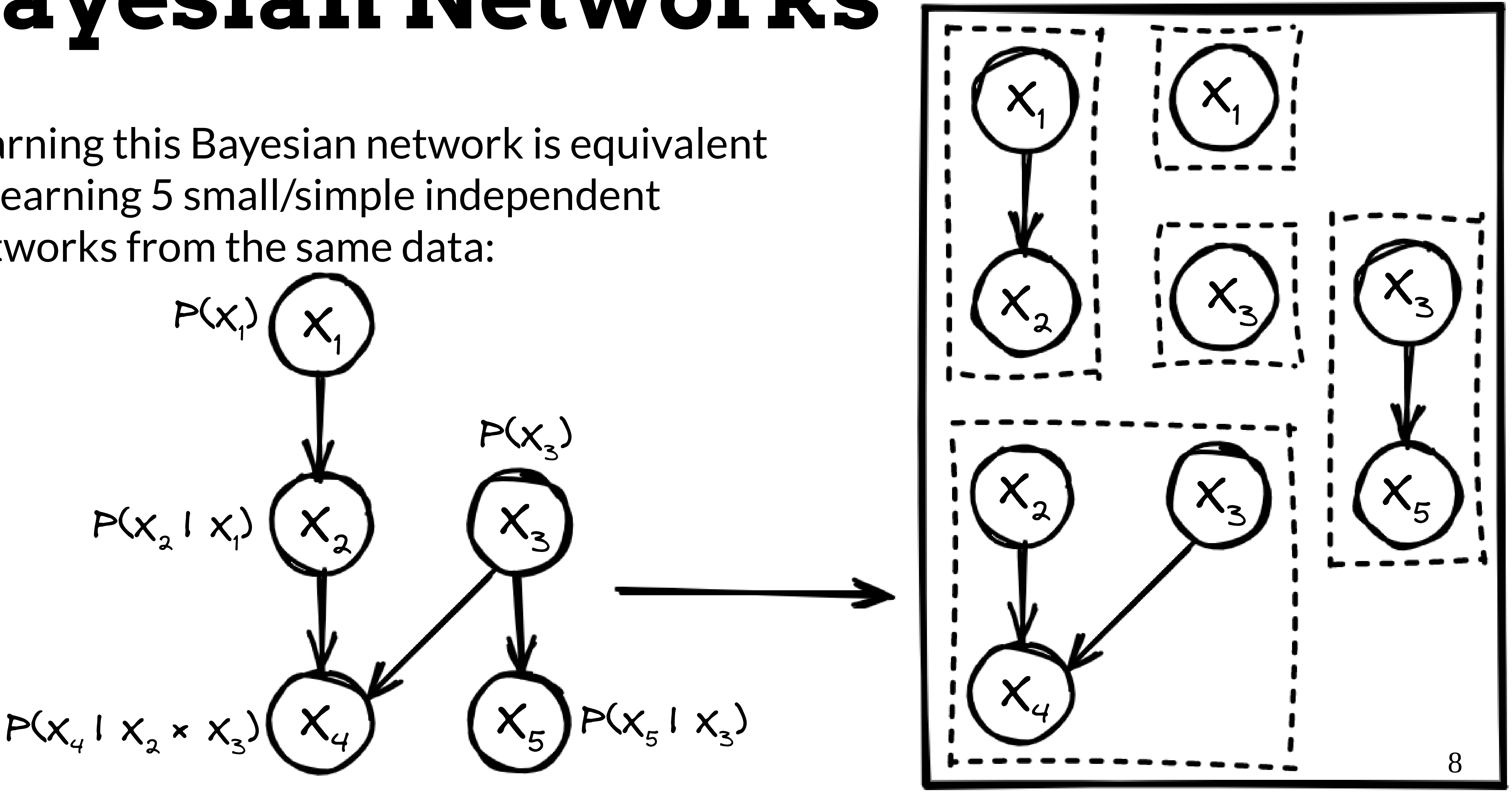
conditional distributions

marginal distributions



Bayesian Networks

Learning this Bayesian network is equivalent to learning 5 small/simple independent networks from the same data:

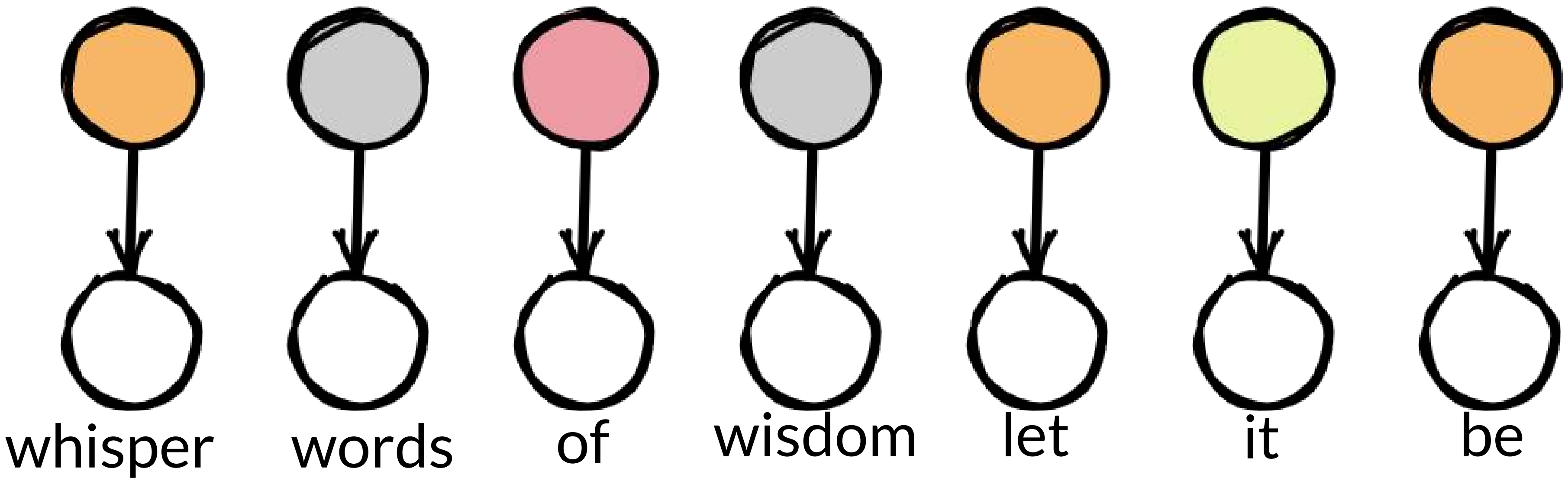


Mixture model

whisper words of wisdom let it be

verb noun
preposition
pronoun

$$
\begin{aligned}
P(y, x) &= P(\text{verb}, \text{noun}, \text{preposition}, \text{noun}, \text{verb}, \text{pronoun}, \text{verb}, \text{whisper}, \text{words}, \text{of}, \text{wisdom}, \text{let}, \text{it}, \text{be}) \\
&= P(\text{verb}, \text{whisper}) \times P(\text{noun}, \text{words}) \times \\
&\quad \times \ldots \\
&= P(\text{whisper} \mid \text{verb}) \times P(\text{verb}) \times \\
&\quad \times P(\text{words} \mid \text{noun}) \times P(\text{noun}) \times \\
&\quad \times \ldots
\end{aligned}
$$



Mixture model

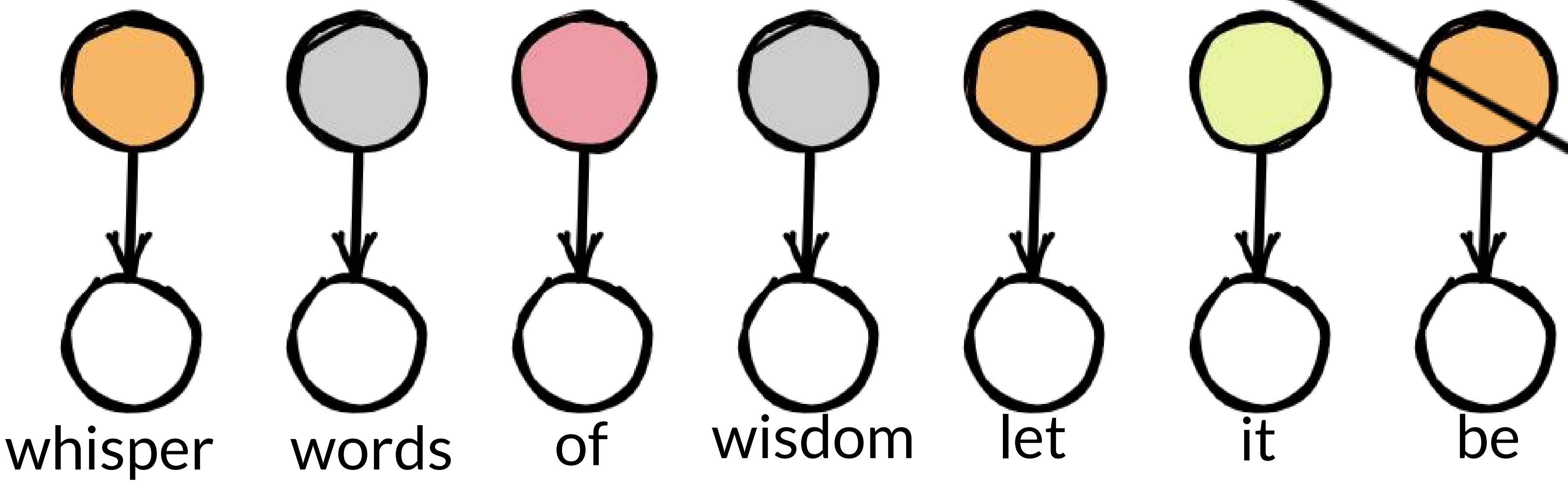
whisper words of wisdom let it be

	whispers	words	of	wisdom	let	it	be
verb (0.35)	0.7	0.2	0.1	0.05	0.6	0.0	0.9
noun (0.4)	0.3	0.7	0.1	0.85	0.3	0.15	0.0
prep (0.15)	0.0	0.0	0.7	0.0	0.05	0.1	0.1
pronoun (0.1)	0.0	0.1	0.1	0.1	0.05	0.65	0.0

$$P(y, x) = P(\text{verb}, \text{noun}, \text{preposition}, \text{noun}, \text{verb}, \text{pronoun}, \text{verb}, \text{whisper}, \text{words}, \text{of}, \text{wisdom}, \text{let}, \text{it}, \text{be}) = P(\text{verb}, \text{whisper}) \times P(\text{noun}, \text{words}) \times \ldots$$

$$= P(\text{whisper} \mid \text{verb}) \times P(\text{verb}) \times P(\text{words} \mid \text{noun}) \times P(\text{noun}) \times \ldots$$

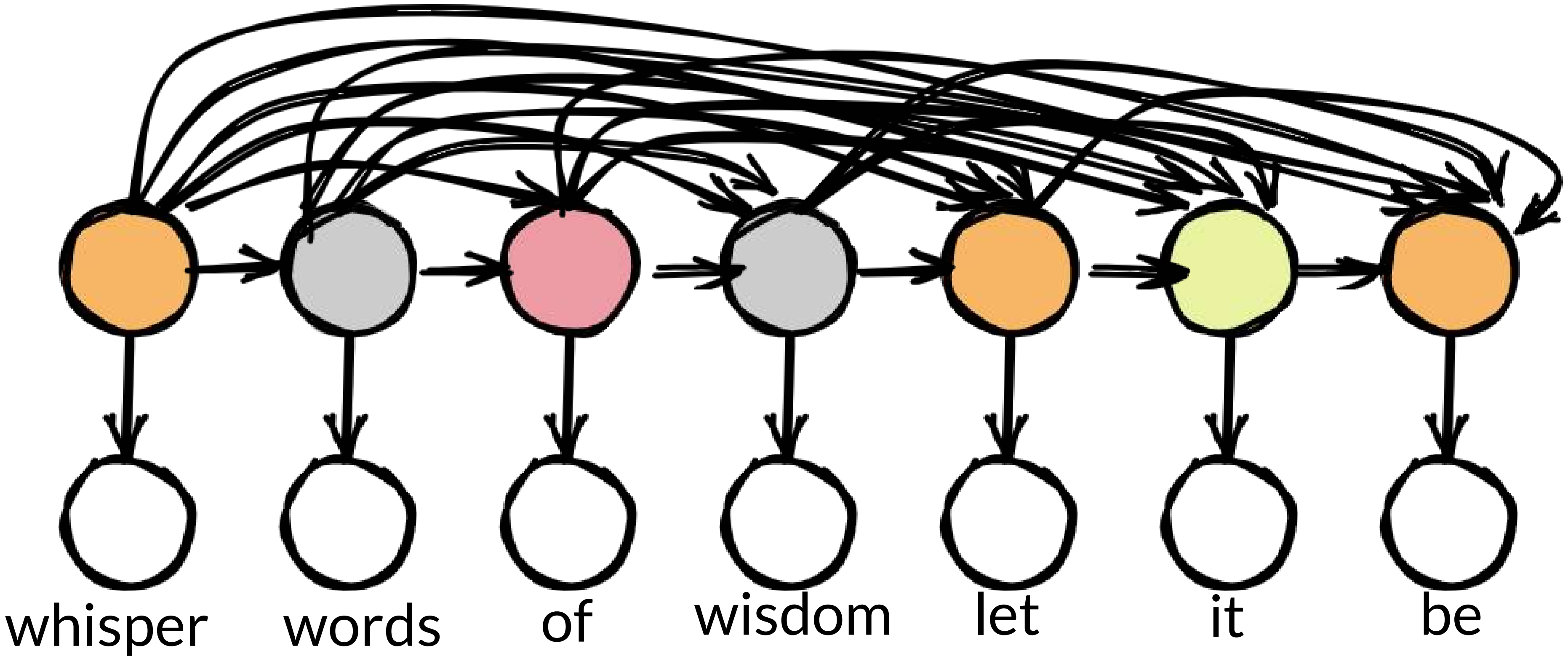
$$= (0.7 \times 0.35) \times (0.7 \times 0.4) \times (0.7 \times 0.15) \times \ldots$$



Emission model

Better model

The previous probabilistic model is too simple. Context (adjacent words and labels) is essential. We'll add dependencies between labels (<u>not between words</u>)



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

$$P(y) = P(y_1) \times P(y_2 \mid y_1) \times$$
$$\times P(y_3 \mid y_1, y_2) \times \ldots \times$$
$$\times P(y_N \mid y_1, y_2, \ldots y_{N-1}) =$$

$$= P(y_1) \times \prod_{N=2} P(y_i \mid y_1, \ldots y_{i-1})$$

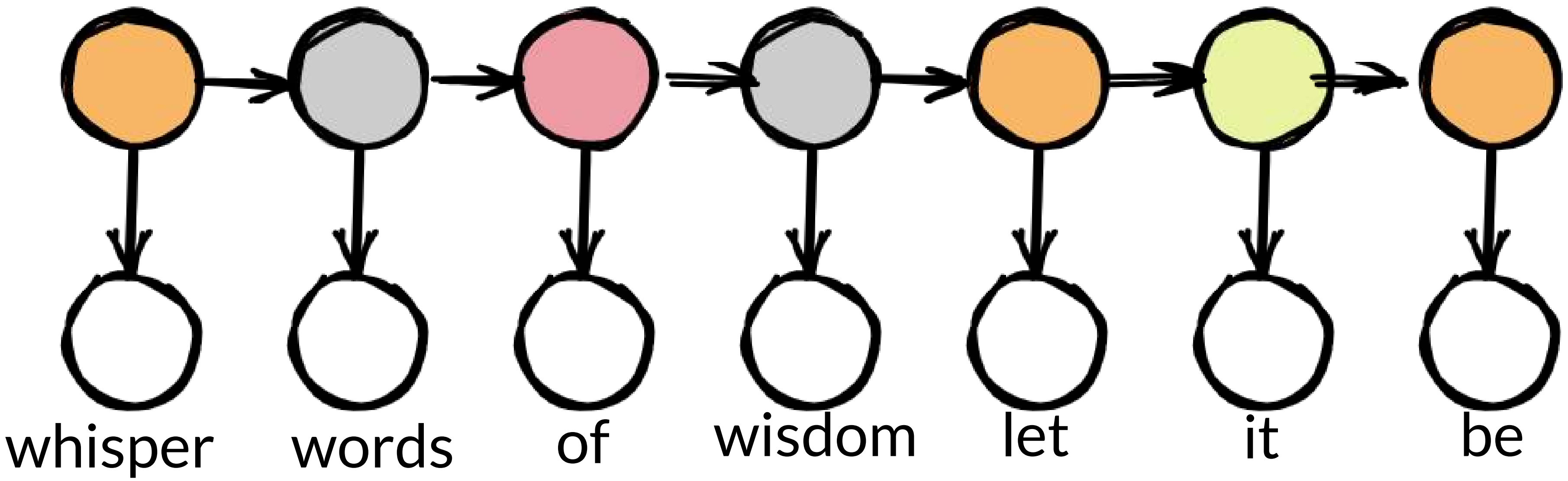
thus each y_i depends on all previous $i-1$ states

Markov assumption

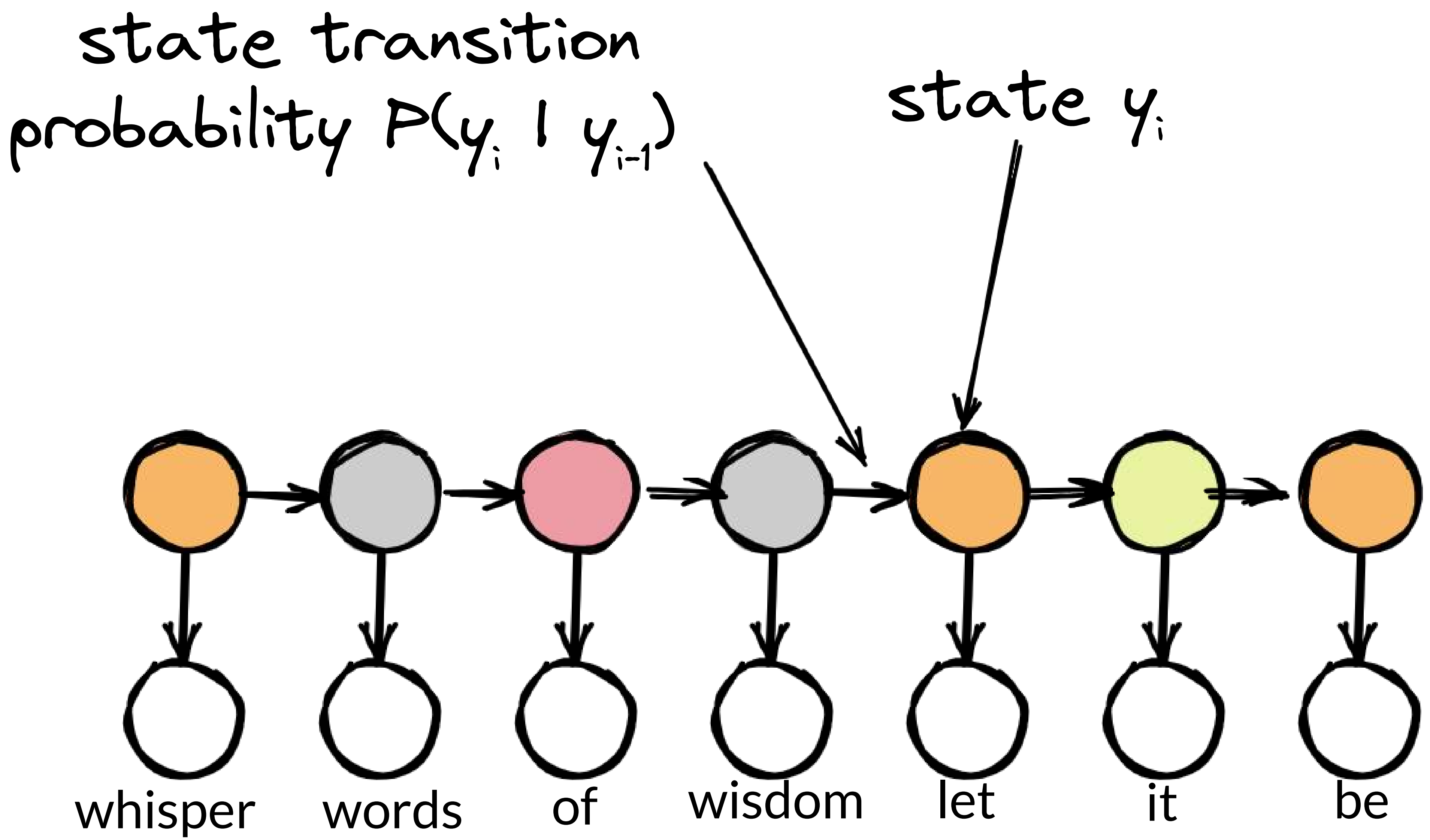
Hard to believe the y_i element depends on all, including the first one.
Markov assumption allows us to consider y_i being dependent on **only** the last element (y_{i-1})

$$P(y_1, x) = P(x \mid y) \times P(y)\ P(y)$$
$$= P(y_1) \times P(y_2 \mid y_1) \times$$
$$\times P(y_3 \mid y_2) \times \ldots \times P(y_N \mid y_{N-1}) =$$

$$= P(y_1) \times \prod_{N=2} P(y_i \mid y_{i-1})$$



Markov chain

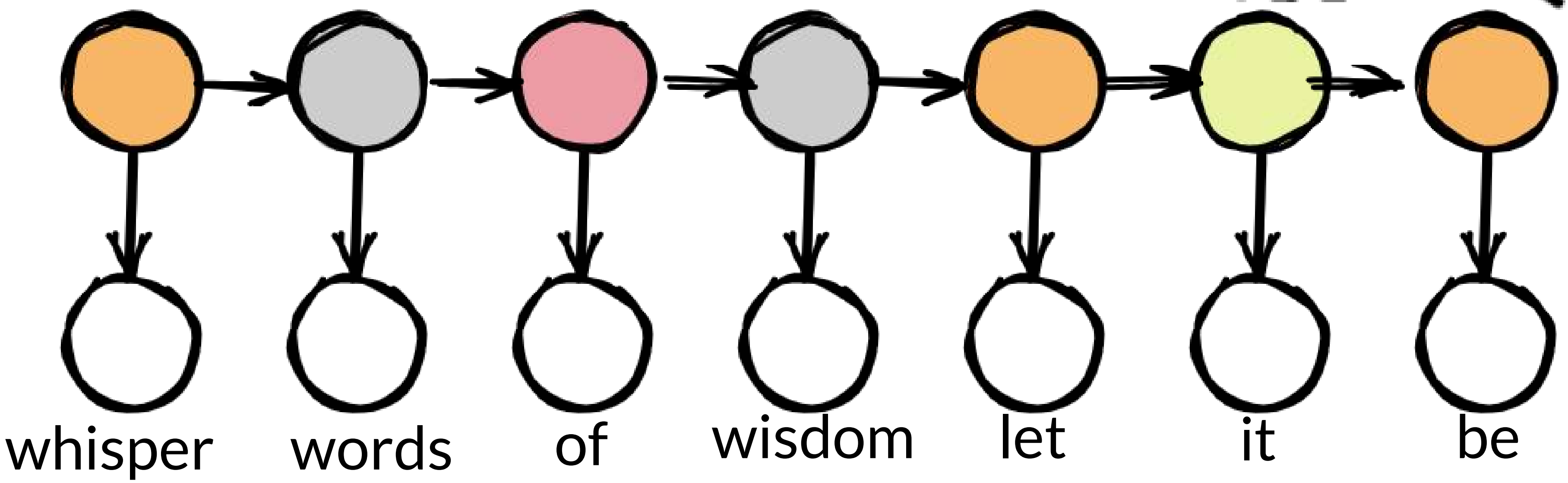


Markov model

whisper words of wisdom let it be

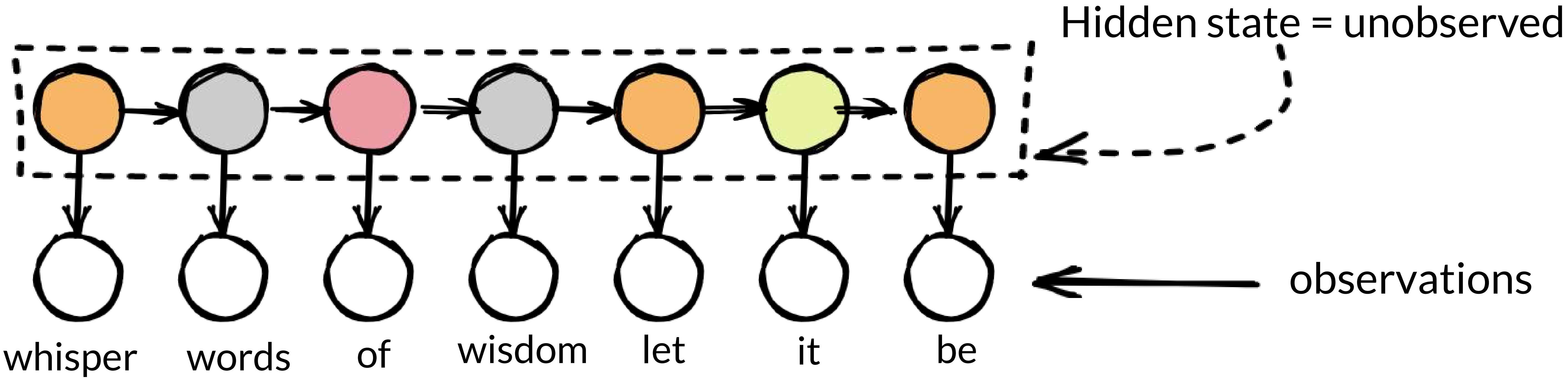
								i-1			
i	whispers	words	of	wisdom	let	it	be	v	n	p	p
verb (0.35)	0.7	0.2	0.1	0.05	0.6	0.0	0.9	0.1	0.4	0.2	0.3
noun (0.4)	0.3	0.7	0.1	0.85	0.3	0.15	0.0	0.8	0.1	0.1	0.0
prep (0.15)	0.0	0.0	0.7	0.0	0.05	0.1	0.1	0.2	0.3	0.2	0.3
pronoun (0.1)	0.0	0.1	0.1	0.1	0.05	0.65	0.0	0.2	0.8	0.0	0.0

$$
\begin{aligned}
&P(y, x) = \\
&= P(\text{whisper} \mid \text{verb}) \times P(\text{verb}) \times \\
&\times P(\text{words} \mid \text{noun}) \times P(\text{noun} \mid \text{verb}) \times \\
&\times P(\text{of} \mid \text{prep}) \times P(\text{prep} \mid \text{noun}) \times \\
&\times \ldots = \\
&= (0.7 \times 0.35) \times (0.7 \times 0.8) \times \\
&\times (0.7 \times 0.3) \times \ldots
\end{aligned}
$$



State transition matrix.
If constant => homogeneous Markov model

Hidden Markov Model



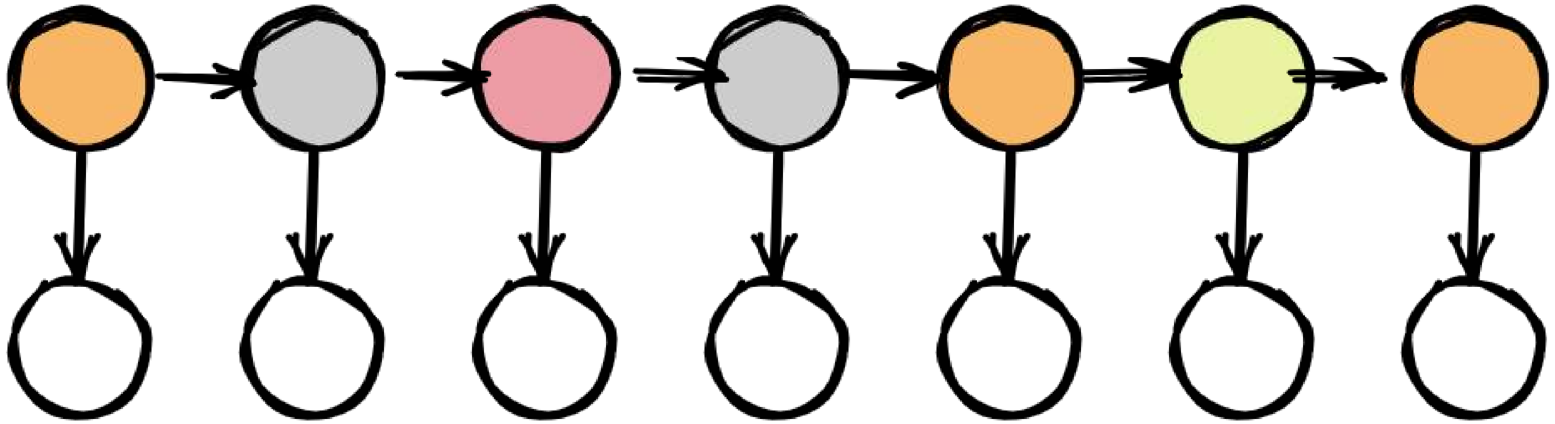
For a hidden Markov Model:

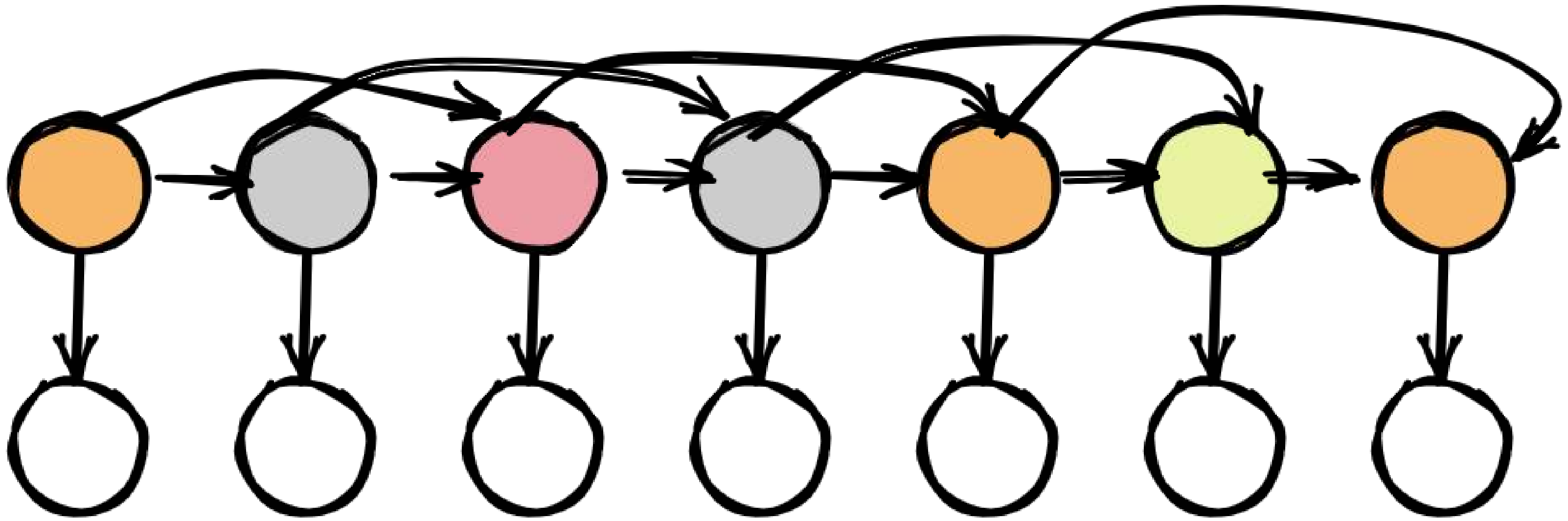
$P(y, x) = P(x \mid y) \times P(y)$

$P(y)$ is the probability of the hidden sequence. For a Markov chain y_i depends only on previous state. $P(y) = i(y_1) \times \prod_{N=2} s(y_i, y_{i-1})$

$P(x \mid y)$ is the emission model of the HMM => $P(y, x)$ = Markov chain × emission model

Higher-order HMMs





Inference problems for HMMs

Given an observation sequence $\boldsymbol{x}$ and an HMM model λ, how do we efficiently compute $P(\boldsymbol{x}|\lambda)$, i.e., the probability of the observation sequence given the model

Evaluation

forward algorithm

Given an observation sequence $\boldsymbol{x}$ and an HMM model λ, how do we choose a corresponding state sequence $\boldsymbol{y}$ which is optimal in some sense, i.e., best explains the observations

Decoding
(Recognition)

Viterbi algorithm

Given an observation sequence $\boldsymbol{x}$, how do we adjust (learn) the model parameters λ, to maximise $P(\boldsymbol{x}|\lambda)$

Training

Baum-Welch algorithm