10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Hidden Markov Models

Matt Gormley
Lecture 19
Mar. 27, 2019

Reminders

- Homework 6: Learning Theory / Generative Models
  - Out: Fri, Mar 22
  - Due: Fri, Mar 29 at 11:59pm (1 week)
- Midterm Exam 2
  - Thu, Apr 4 – evening exam, details announced on Piazza
- Homework 7: HMMs
  - Out: Fri, Mar 29
  - Due: Wed, Apr 10 at 11:59pm

- Today's In-Class Poll
  - http://p19.mlcourse.org

Reminders

- **Schedule Change:**
  - Fri (3/29) - Lecture 20: HMMs (Part II) / Midterm Exam Review
  - Mon (4/1) - Recitation 7: HW7

HIDDEN MARKOV MODEL (HMM)

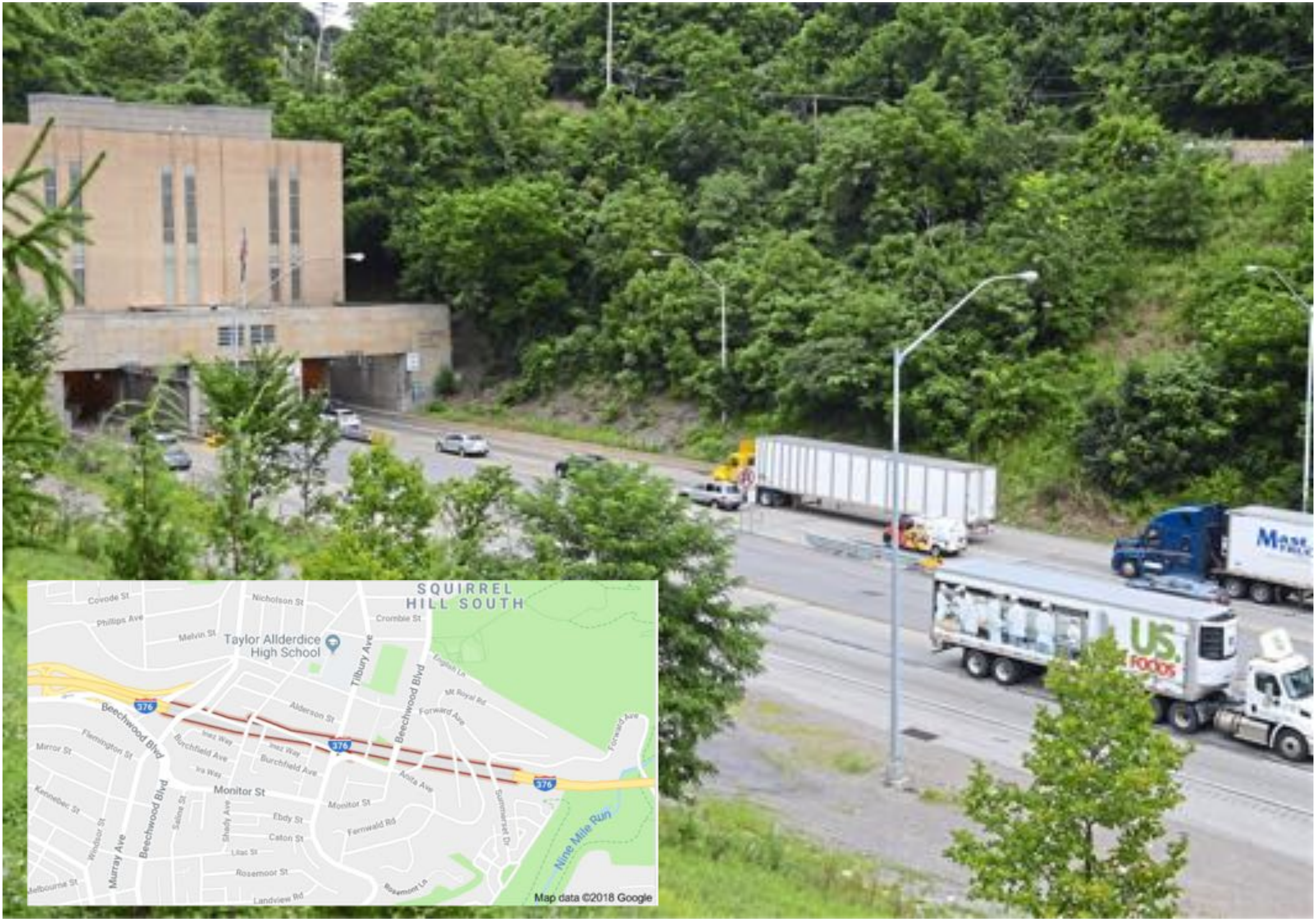
HMM Outline

- **Motivation**
  - Time Series Data
- **Hidden Markov Model (HMM)**
  - Example: Squirrel Hill Tunnel Closures
    [courtesy of Roni Rosenfeld]
  - Background: Markov Models
  - From Mixture Model to HMM
  - History of HMMs
  - Higher-order HMMs
- **Training HMMs**
  - (Supervised) Likelihood for HMM
  - Maximum Likelihood Estimation (MLE) for HMM
  - EM for HMM (aka. Baum-Welch algorithm)
- **Forward-Backward Algorithm**
  - Three Inference Problems for HMM
  - Great Ideas in ML: Message Passing
  - Example: Forward-Backward on 3-word Sentence
  - Derivation of Forward Algorithm
  - Forward-Backward Algorithm
  - Viterbi algorithm

Markov Models

Whiteboard

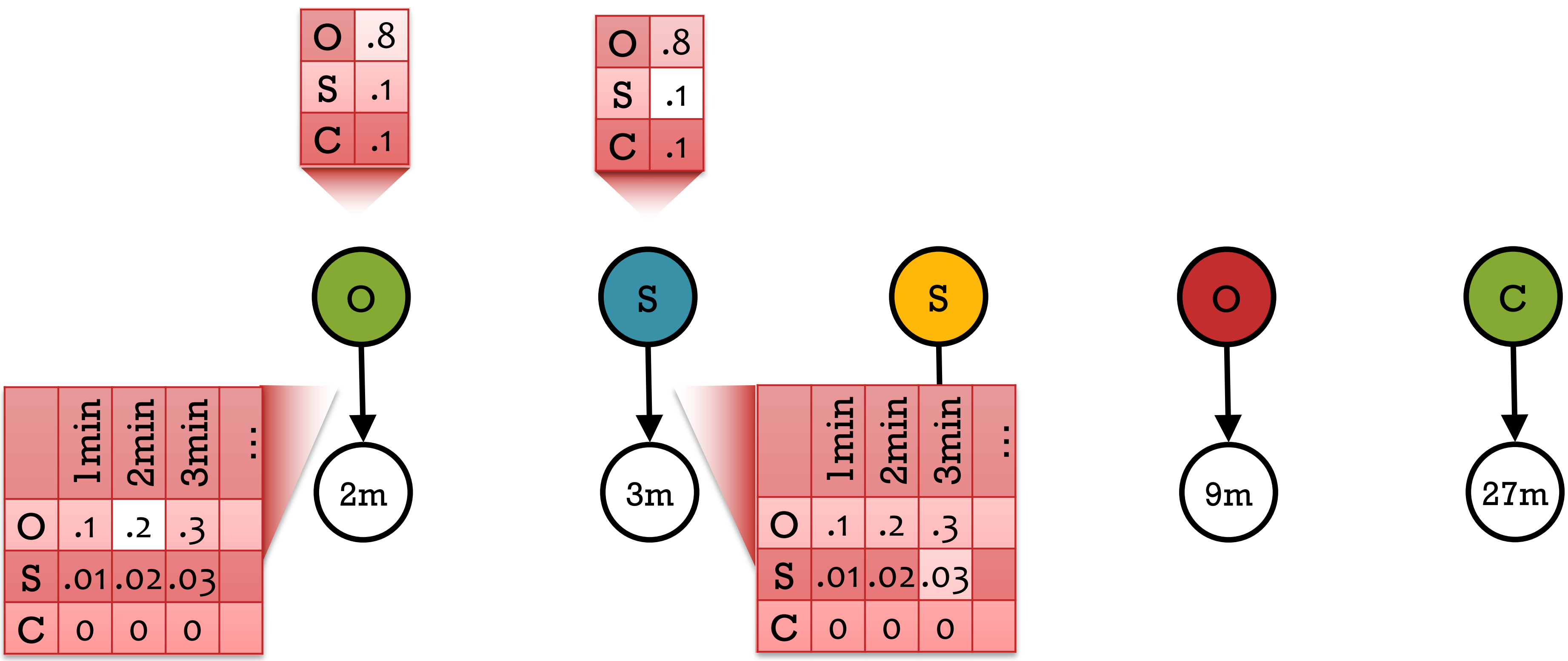
- Example: Tunnel Closures
  [courtesy of Roni Rosenfeld]
- First-order Markov assumption
- Conditional independence assumptions


SQUIRREL HILL SOUTH
Taylor Allderdice High School
Covode St
Phillips Ave
Nicholson St
Melvin St
Crombie St
Tilbury Ave
Beechwood Blvd
English Ln
Mt Royal Rd
Forward Ave
Alderson St
376
Burchfield Ave
Inez Way
Ira Way
Flemington St
Mirror St
Kennebec St
Monitor St
Anita Ave
Summerset Dr
Ebdy St
Caton St
Fernwald Rd
Lilac St
Rosemoor St
Windsor St
Murray Ave
Landview Rd
Rosemont Ln
Nine Mile Run
Map data ©2018 Google
US. FOODS

Mixture Model for Time Series Data

We could treat each (tunnel state, travel time) pair as independent. This corresponds to a Naïve Bayes model with a single feature (travel time).

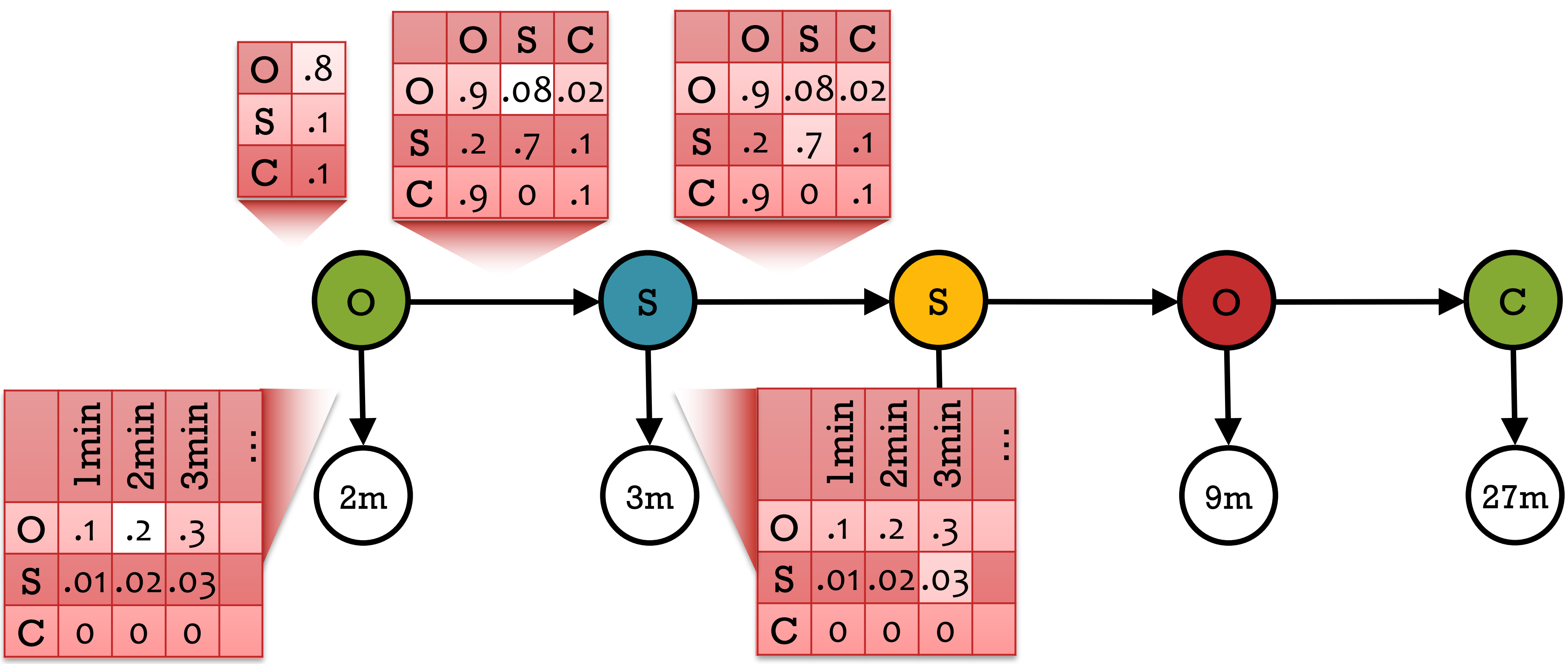
$$p(\text{O}, \text{S}, \text{S}, \text{O}, \text{C}, 2\text{m}, 3\text{m}, 18\text{m}, 9\text{m}, 27\text{m}) = (.8 * .2 * .1 * .03 * \ldots)$$



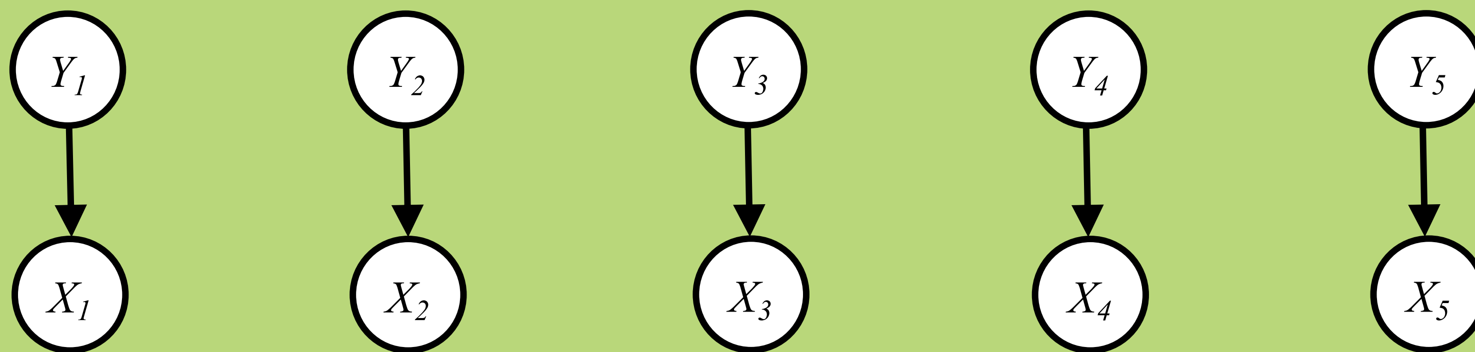
Hidden Markov Model

A Hidden Markov Model (HMM) provides a joint distribution over the the tunnel states / travel times with an assumption of dependence between adjacent tunnel states.

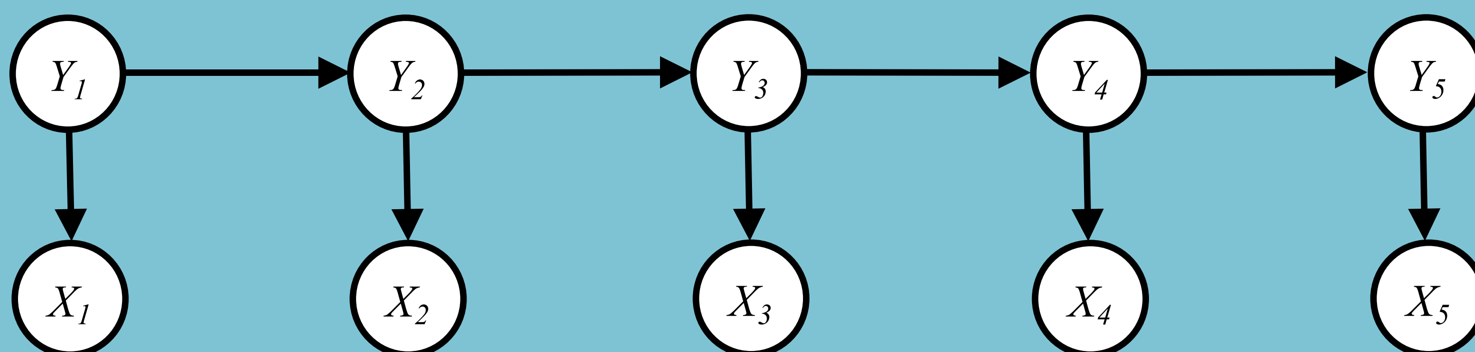
$$p(O, S, S, O, C, 2m, 3m, 18m, 9m, 27m) = (.8 * .08 * .2 * .7 * .03 * \ldots)$$



From Mixture Model to HMM

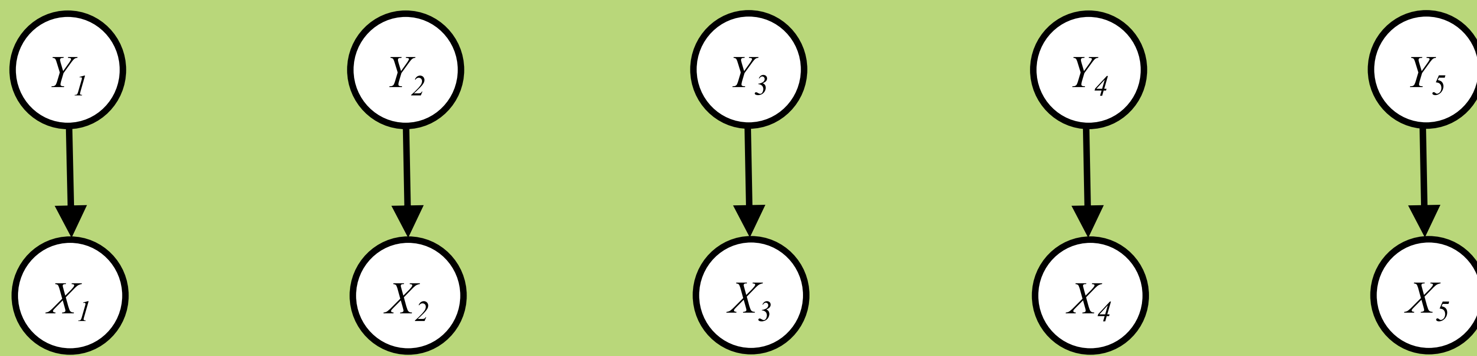


"Naïve Bayes": $$P(\mathbf{X},\mathbf{Y}) = \prod_{t=1}^{T} P(X_t|Y_t)p(Y_t)$$

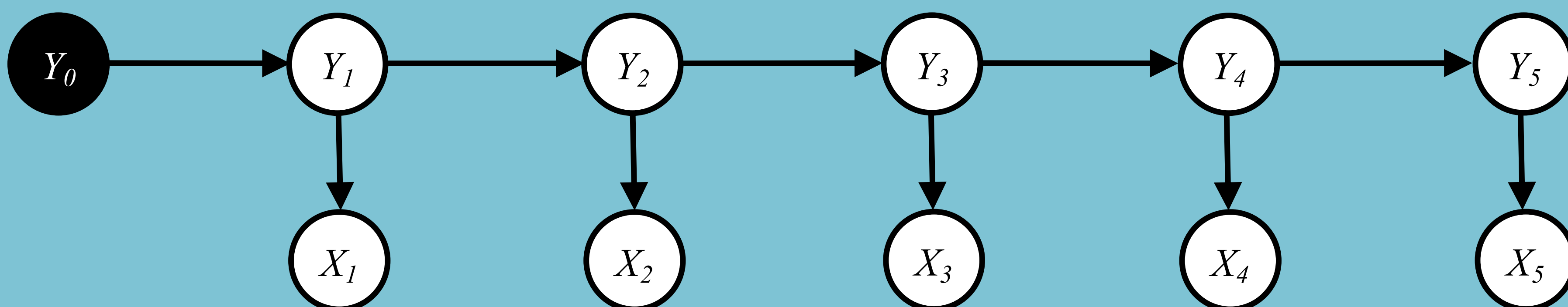


HMM: $$P(\mathbf{X},\mathbf{Y}) = P(Y_1)\left(\prod_{t=1}^{T} P(X_t|Y_t)\right)\left(\prod_{t=2}^{T} p(Y_t|Y_{t-1})\right)$$

From Mixture Model to HMM



"Naïve Bayes": $$P(\mathbf{X}, \mathbf{Y}) = \prod_{t=1}^{T} P(X_t|Y_t)p(Y_t)$$



HMM: $$P(\mathbf{X}, \mathbf{Y}|Y_0) = \prod_{t=1}^{T} P(X_t|Y_t)p(Y_t|Y_{t-1})$$

SUPERVISED LEARNING FOR HMMS

Recipe for Closed-form MLE

1. Assume data was generated i.i.d. from some model (i.e. write the generative story)

   $x^{(i)} \sim p(x|\boldsymbol{\theta})$
2. Write log-likelihood

   $\ell(\boldsymbol{\theta}) = \log p(x^{(1)}|\boldsymbol{\theta}) + \ldots + \log p(x^{(N)}|\boldsymbol{\theta})$
3. Compute partial derivatives (i.e. gradient)

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_1 = \ldots$

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_2 = \ldots$

   …

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_M = \ldots$
4. Set derivatives to zero and solve for $\boldsymbol{\theta}$

   $\partial\ell(\boldsymbol{\theta})/\partial\theta_m = 0$ for all $m \in \{1, \ldots, M\}$

   $\boldsymbol{\theta}^{MLE}$ = solution to system of M equations and M variables
5. Compute the second derivative and check that $\ell(\boldsymbol{\theta})$ is concave down at $\boldsymbol{\theta}^{MLE}$

MLE of Categorical Distribution

1. Suppose we have a **dataset** obtained by repeatedly rolling a M-sided (weighted) die N times. That is, we have data

$$\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$$

where $x^{(i)} \in \{1, \ldots, M\}$ and $x^{(i)} \sim \text{Categorical}(\phi)$.

2. A random variable is **Categorical** written $X \sim \text{Categorical}(\phi)$ iff

$$P(X = x) = p(x; \phi) = \phi_x$$

where $x \in \{1, \ldots, M\}$ and $\sum_{m=1}^{M} \phi_m = 1$. The **log-likelihood** of the data becomes:

$$\ell(\phi) = \sum_{i=1}^{N} \log \phi_{x^{(i)}} \text{ s.t. } \sum_{m=1}^{M} \phi_m = 1$$

3. Solving this *constrained* optimization problem yields the **maximum likelihood estimator** (MLE):

$$\phi_m^{MLE} = \frac{N_{x=m}}{N} = \frac{\sum_{i=1}^{N} \mathbb{I}(x^{(i)} = m)}{N}$$

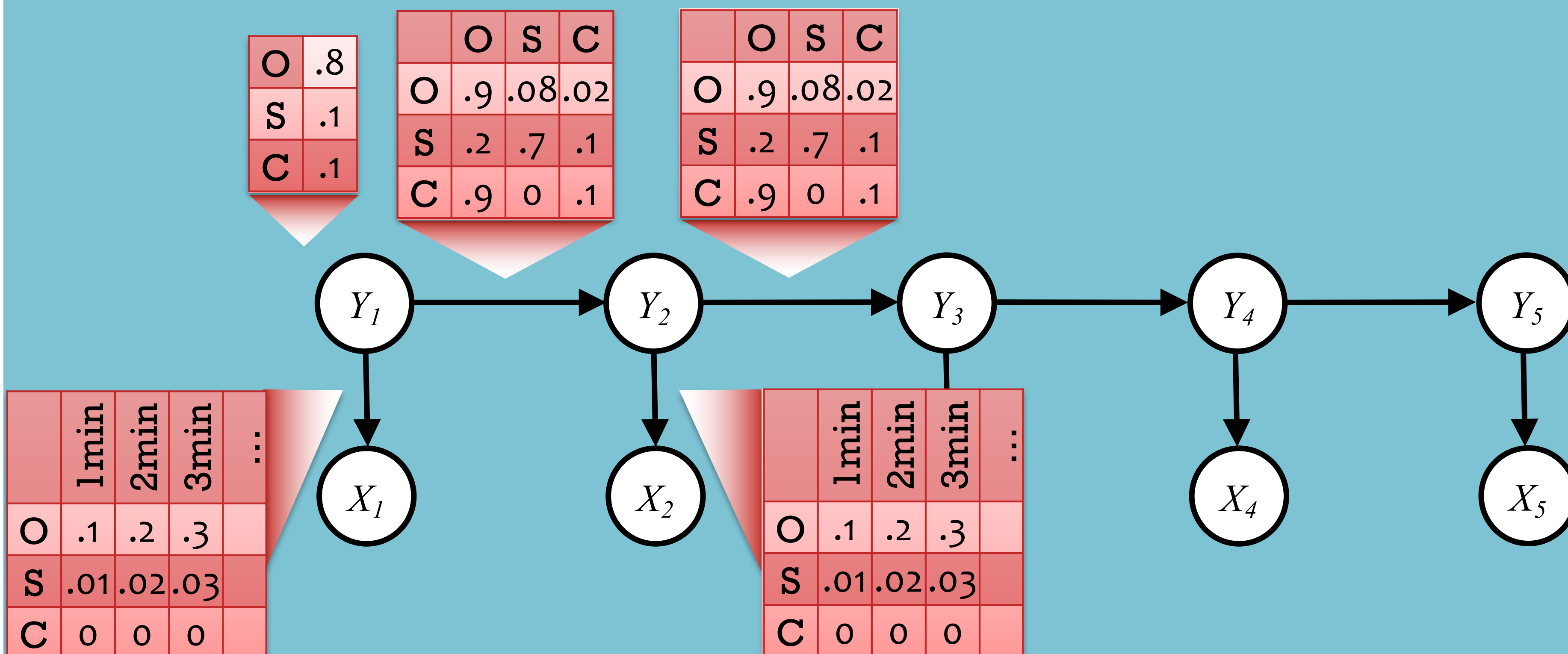
Hidden Markov Model

HMM Parameters:

Emission matrix, $\mathbf{A}$, where $P(X_t = k | Y_t = j) = A_{j,k}, \forall t, k$

Transition matrix, $\mathbf{B}$, where $P(Y_t = k | Y_{t-1} = j) = B_{j,k}, \forall t, k$

Initial probs, $\mathbf{C}$, where $P(Y_1 = k) = C_k, \forall k$



O	.8
S	.1
C	.1

	O	S	C
O	.9	.08	.02
S	.2	.7	.1
C	.9	0	.1

	O	S	C
O	.9	.08	.02
S	.2	.7	.1
C	.9	0	.1

	1min	2min	3min	…
O	.1	.2	.3	
S	.01	.02	.03	
C	0	0	0	

	1min	2min	3min	…
O	.1	.2	.3	
S	.01	.02	.03	
C	0	0	0	

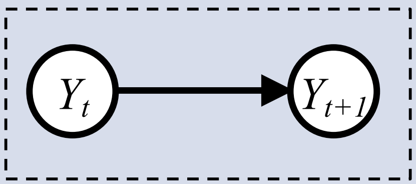
Training HMMs

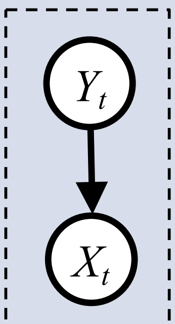
Whiteboard

- (Supervised) Likelihood for an HMM
- Maximum Likelihood Estimation (MLE) for HMM

Supervised Learning for HMMs

Learning an HMM decomposes into solving two (independent) Mixture Models





Data: $\mathcal{D} = \{(\vec{x}^{(i)}, \vec{y}^{(i)})\}_{i=1}^{N}$ $\quad \vec{x} = [x_1, \ldots, x_T]^T$, $\vec{y} = [y_1, \ldots, y_T]^T$

Likelihood:

$$\ell(A,B,C) = \sum_{i=1}^{N} \log p(\vec{x}^{(i)}, y^{(i)} \mid A,B,C)$$

$$= \sum_{i=1}^{N} \Big[\underbrace{\log p(y_1^{(i)} \mid C)}_{\text{initial}} + \underbrace{\Big(\sum_{t=2}^{T} \log p(y_t^{(i)} \mid y_{t-1}^{(i)}, B)\Big)}_{\text{transition}} + \underbrace{\Big(\sum_{t=1}^{T} \log p(x_t^{(i)} \mid y_t^{(i)}, A)\Big)}_{\text{emission}} \Big]$$

MLE:

$$\hat{A}, \hat{B}, \hat{C} = \underset{A,B,C}{\operatorname{argmax}}\ \ell(A,B,C)$$

$$\Rightarrow \hat{C} = \underset{C}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(y_1^{(i)} \mid C)$$

$$\hat{B} = \underset{B}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=2}^{T} \log p(y_t^{(i)} \mid y_{t-1}^{(i)}, B)$$

$$\hat{A} = \underset{A}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{t=1}^{T} \log p(x_t^{(i)} \mid y_t^{(i)}, A)$$

Can solve in closed form, which yields...

$$\hat{C}_k = \frac{\#(y_1^{(i)} = k)}{N} \quad \forall i, k$$

$$\hat{B}_{jk} = \frac{\#(y_t^{(i)} = k \text{ and } y_{t-1}^{(i)} = j)}{\#(y_{t-1}^{(i)} = j)} \quad \forall i, t>1, j, k$$

$$\hat{A}_{jk} = \frac{\#(x_t^{(i)} = k \text{ and } y_t^{(i)} = j)}{\#(y_t^{(i)} = j)} \quad \forall i, t, j, k$$

Hidden Markov Model

HMM Parameters:

Emission matrix, $\mathbf{A}$, where $P(X_t = k | Y_t = j) = A_{j,k}, \forall t, k$

Transition matrix, $\mathbf{B}$, where $P(Y_t = k | Y_{t-1} = j) = B_{j,k}, \forall t, k$

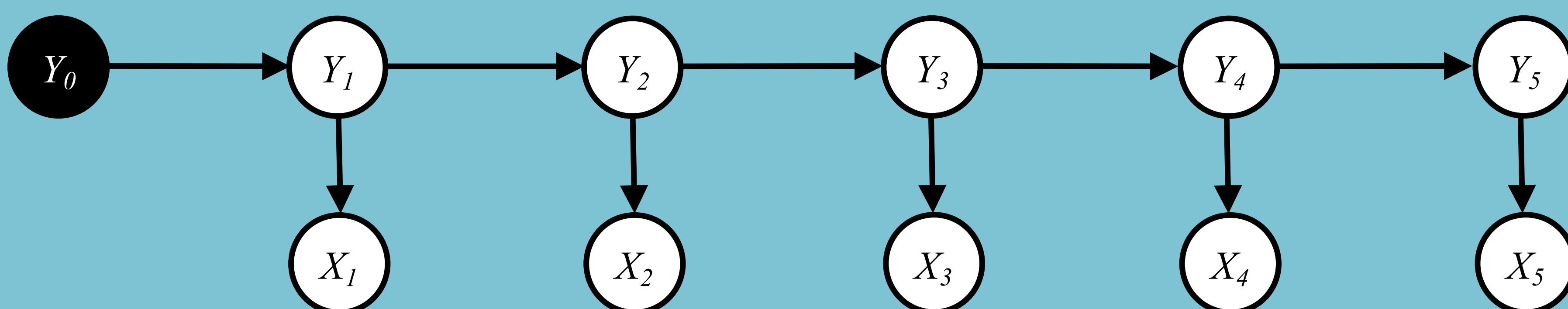
Assumption: $y_0 =$ START

Generative Story:

$$Y_t \sim \text{Multinomial}(\mathbf{B}_{Y_{t-1}}) \ \forall t$$

$$X_t \sim \text{Multinomial}(\mathbf{A}_{Y_t}) \ \forall t$$

For notational convenience, we fold the *initial probabilities* **C** into the *transition matrix* **B** by our assumption.



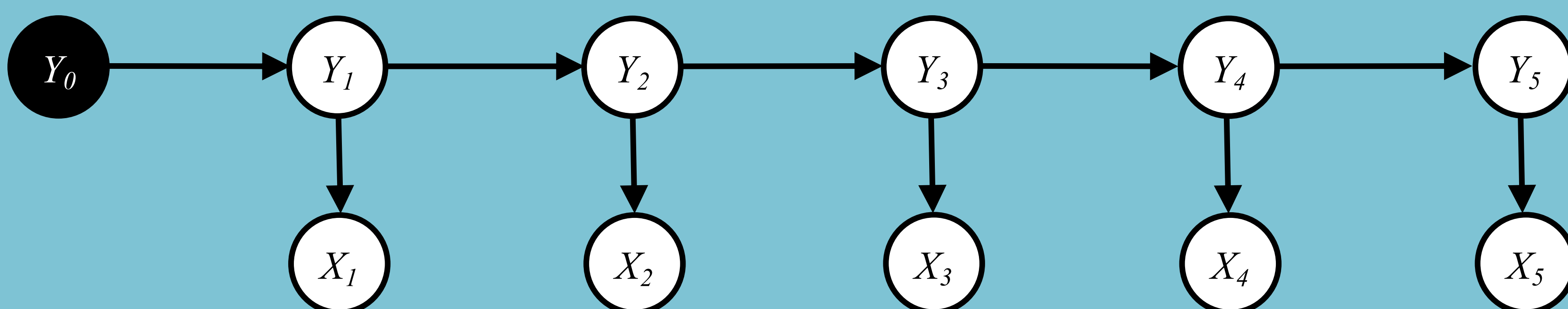
Hidden Markov Model

Joint Distribution:

$$y_0 = \text{START}$$

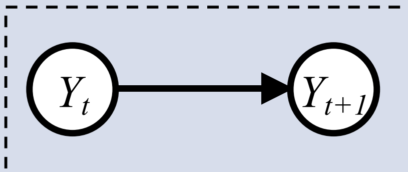
$$p(\mathbf{x}, \mathbf{y} | y_0) = \prod_{t=1}^{T} p(x_t | y_t) p(y_t | y_{t-1})$$

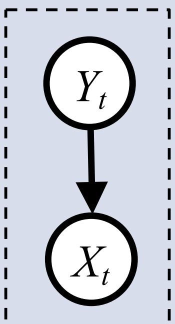
$$= \prod_{t=1}^{T} A_{y_t, x_t} B_{y_{t-1}, y_t}$$



Supervised Learning for HMMs

Learning an HMM decomposes into solving two (independent) Mixture Models





$$\mathcal{D} = \{(\vec{x}^{(i)}, \vec{y}^{(i)})\}_{i=1}^{N}$$

<u>Likelihood</u>:

$$\ell(A,B) = \sum_{i=1}^{N} \log p(\vec{x}^{(i)}, \vec{y}^{(i)})$$

$$= \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \log p(y_t^{(i)} \mid y_{t-1}^{(i)}, B) + \log p(x_t^{(i)} \mid y_t^{(i)}, A) \right]$$

<u>MLE</u>:

$$\hat{A}, \hat{B} = \operatorname{argmax} \ \ell(A,B)$$

$$\hat{A} = \operatorname{argmax} \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \log p(x_t^{(i)} \mid y_t^{(i)}, A) \right]$$

$$\hat{B} = \operatorname{argmax} \sum_{i=1}^{N} \left[\sum_{t=1}^{T} \log p(y_t^{(i)} \mid y_{t-1}^{(i)}, B) \right]$$

↑ can solve in closed form to get...

$$\hat{B}_{jk} = \frac{\#(y_t^{(i)} = k \text{ and } y_{t-1}^{(i)} = j)}{\#(y_{t-1}^{(i)} = j)}$$

$$\hat{A}_{jk} = \frac{\#(x_t^{(i)} = k \text{ and } y_t^{(i)} = j)}{\#(y_t^{(i)} = j)}$$

Unsupervised Learning for HMMs

- Unlike **discriminative** models p(y|x), **generative** models p(x,y) can maximize the likelihood of the data D = {$x^{(1)}$, $x^{(2)}$, …, $x^{(N)}$} where we don't observe any y's.
- This **unsupervised learning** setting can be achieved by finding parameters that maximize the **marginal likelihood**
- We optimize using the **Expectation-Maximization** algorithm

Since we don't observe $\mathbf{y}$, we define the marginal probability:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}) \quad (1)$$

The log-likelihood of the data is thus:

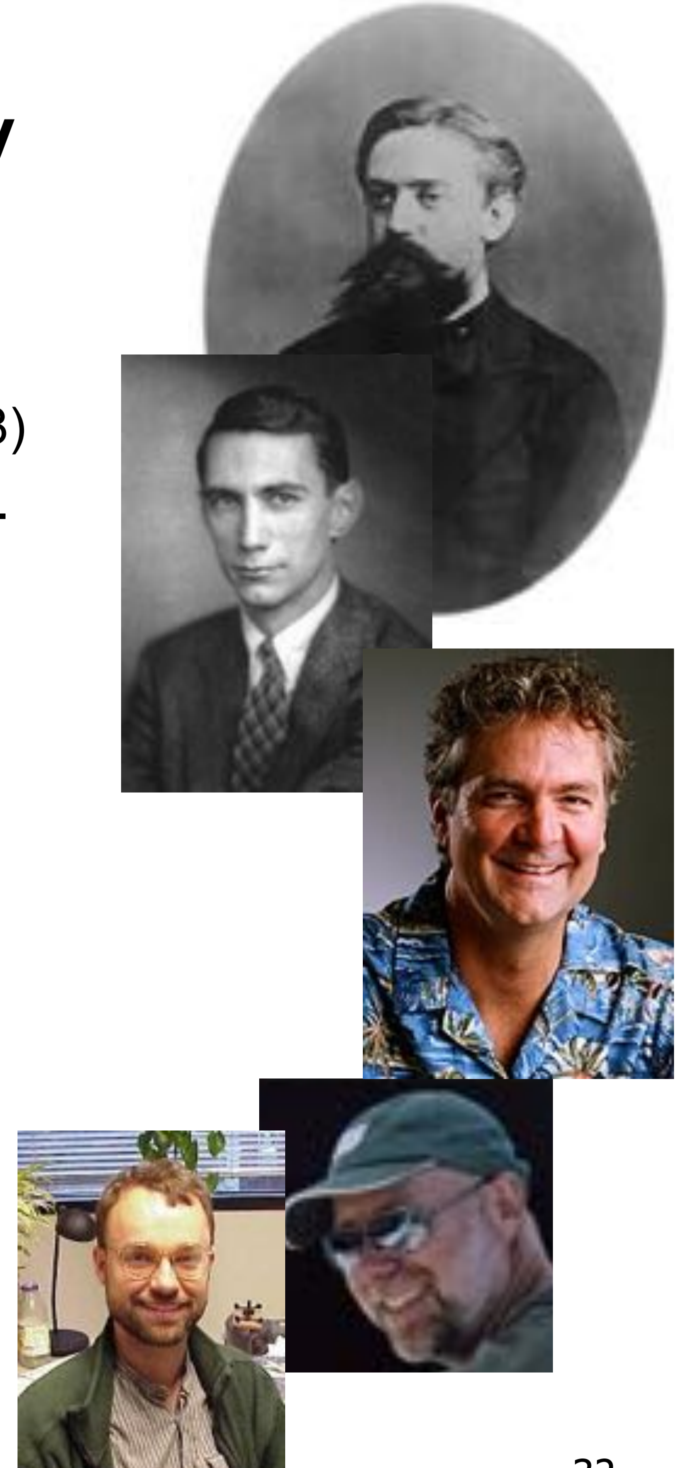
$$\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$$

$$= \sum_{i=1}^{N} \log \sum_{\mathbf{y} \in \mathcal{Y}} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{y}) \quad (3)$$

Beyond the scope of today's lecture!

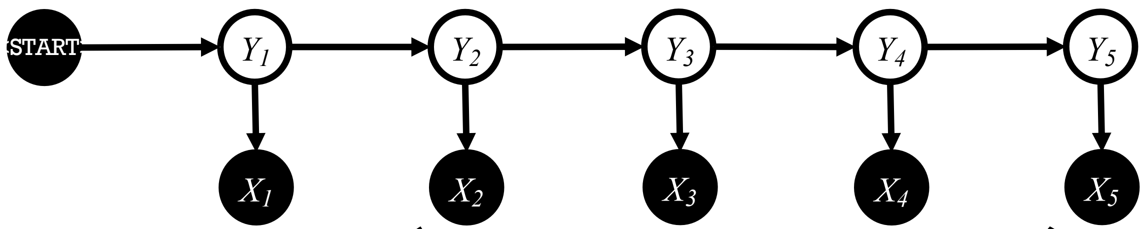
HMMs: History

- Markov chains: Andrey Markov (1906)
  - Random walks and Brownian motion
- Used in Shannon’s work on information theory (1948)
- Baum-Welsh learning algorithm: late 60’s, early 70’s.
  - Used mainly for speech in 60s-70s.
- Late 80’s and 90’s: David Haussler (major player in learning theory in 80’s) began to use HMMs for modeling biological sequences
- Mid-late 1990’s: Dayne Freitag/Andrew McCallum
  - Freitag thesis with Tom Mitchell on IE from Web using logic programs, grammar induction, etc.
  - McCallum: multinomial Naïve Bayes for text
  - With McCallum, IE using HMMs on CORA
- …

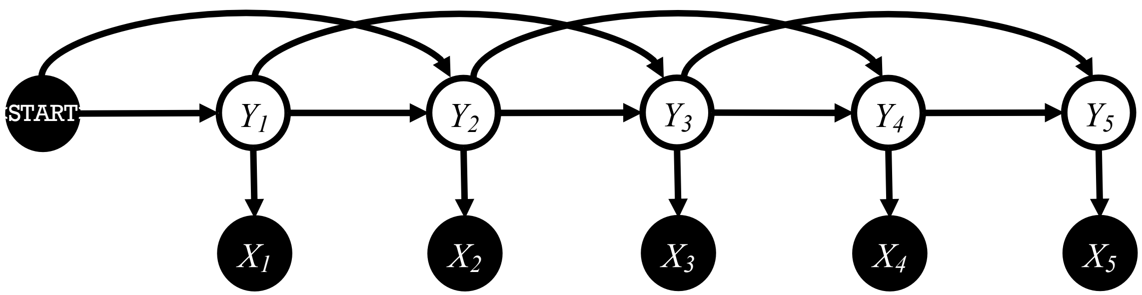


Higher-order HMMs

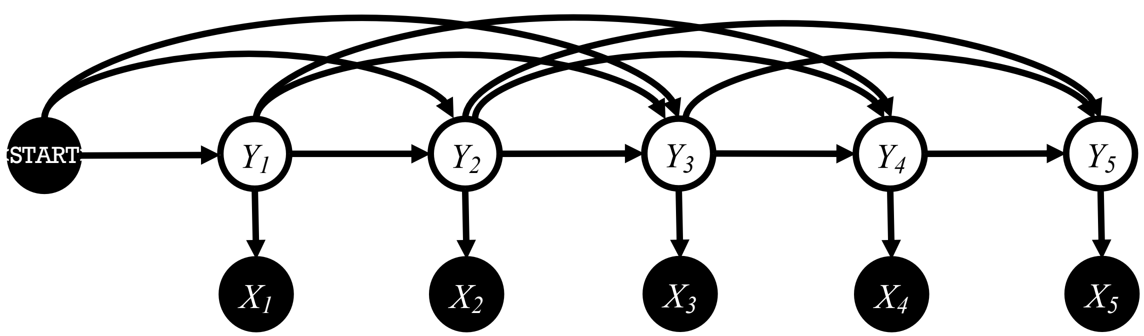
- 1^{st}-order HMM (i.e. bigram HMM)



- 2^{nd}-order HMM (i.e. trigram HMM)

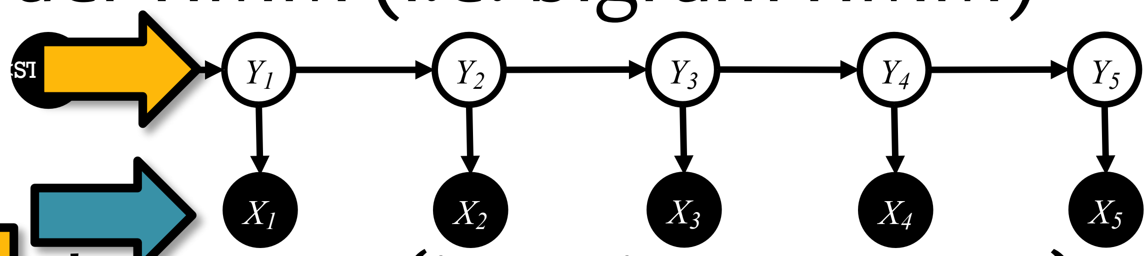


- 3^{rd}-order HMM



Higher-order HMMs

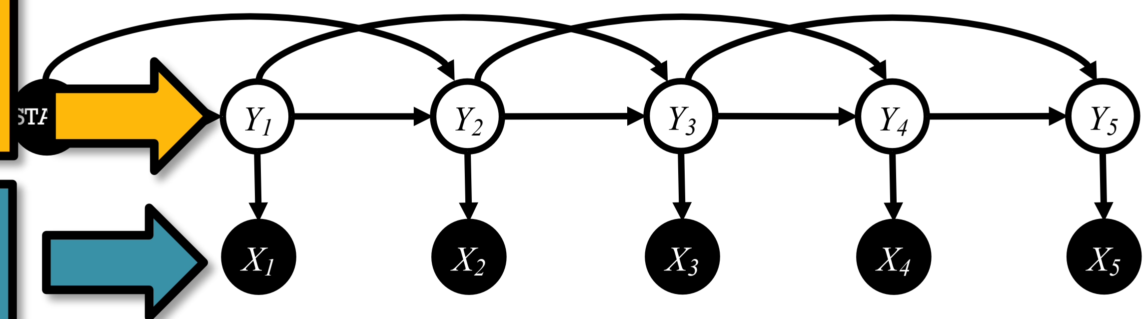
- 1st-order HMM (i.e. bigram HMM)



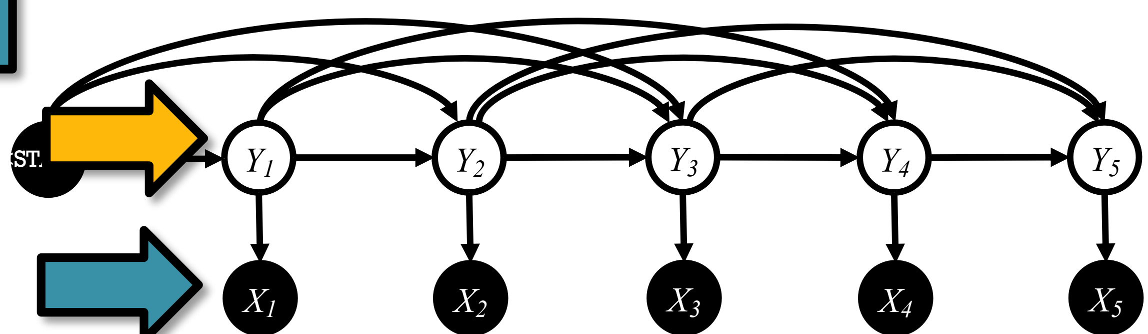
Hidden States, **y**

Observa -tions, **x**

- rder HMM (i.e. trigram HMM)



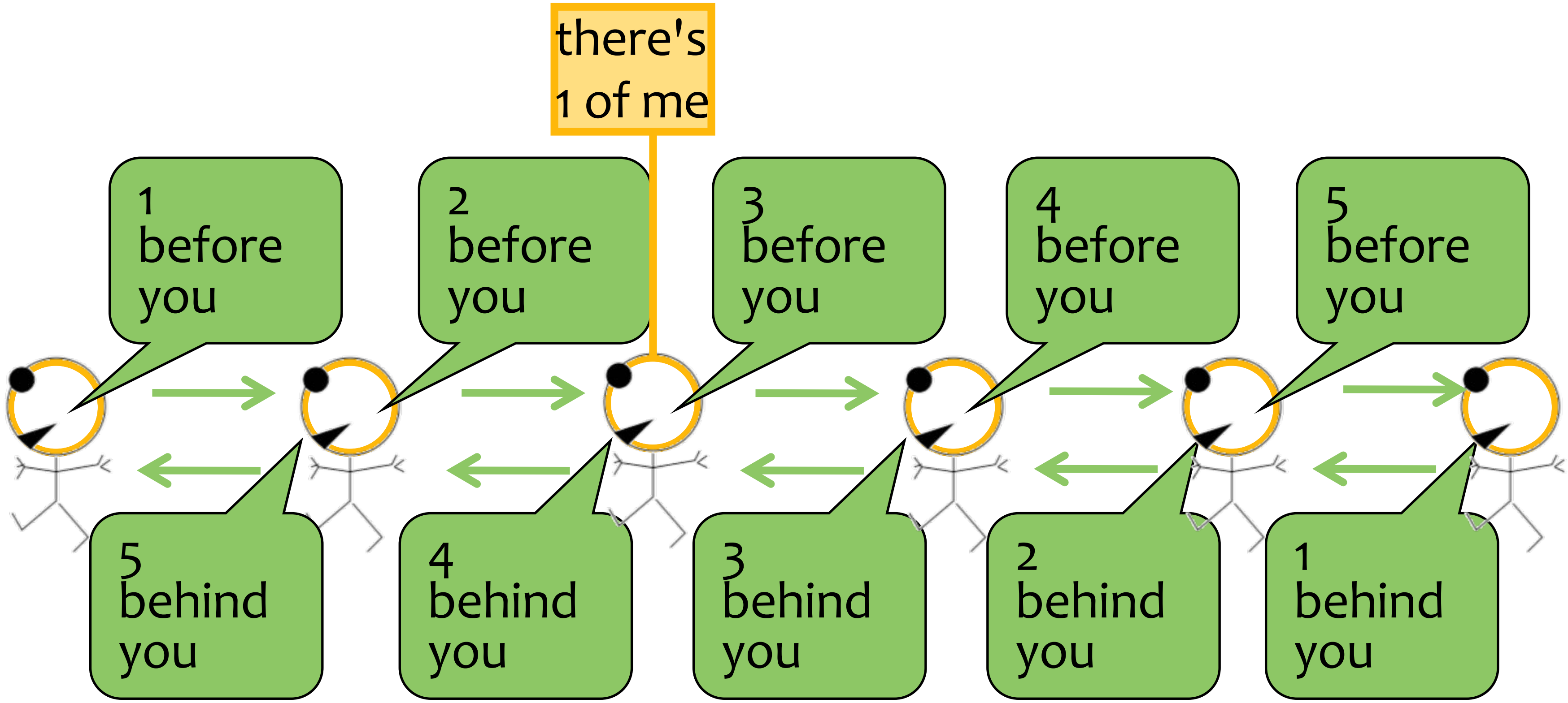
- rder HMM



BACKGROUND: MESSAGE PASSING

Great Ideas in ML: Message Passing

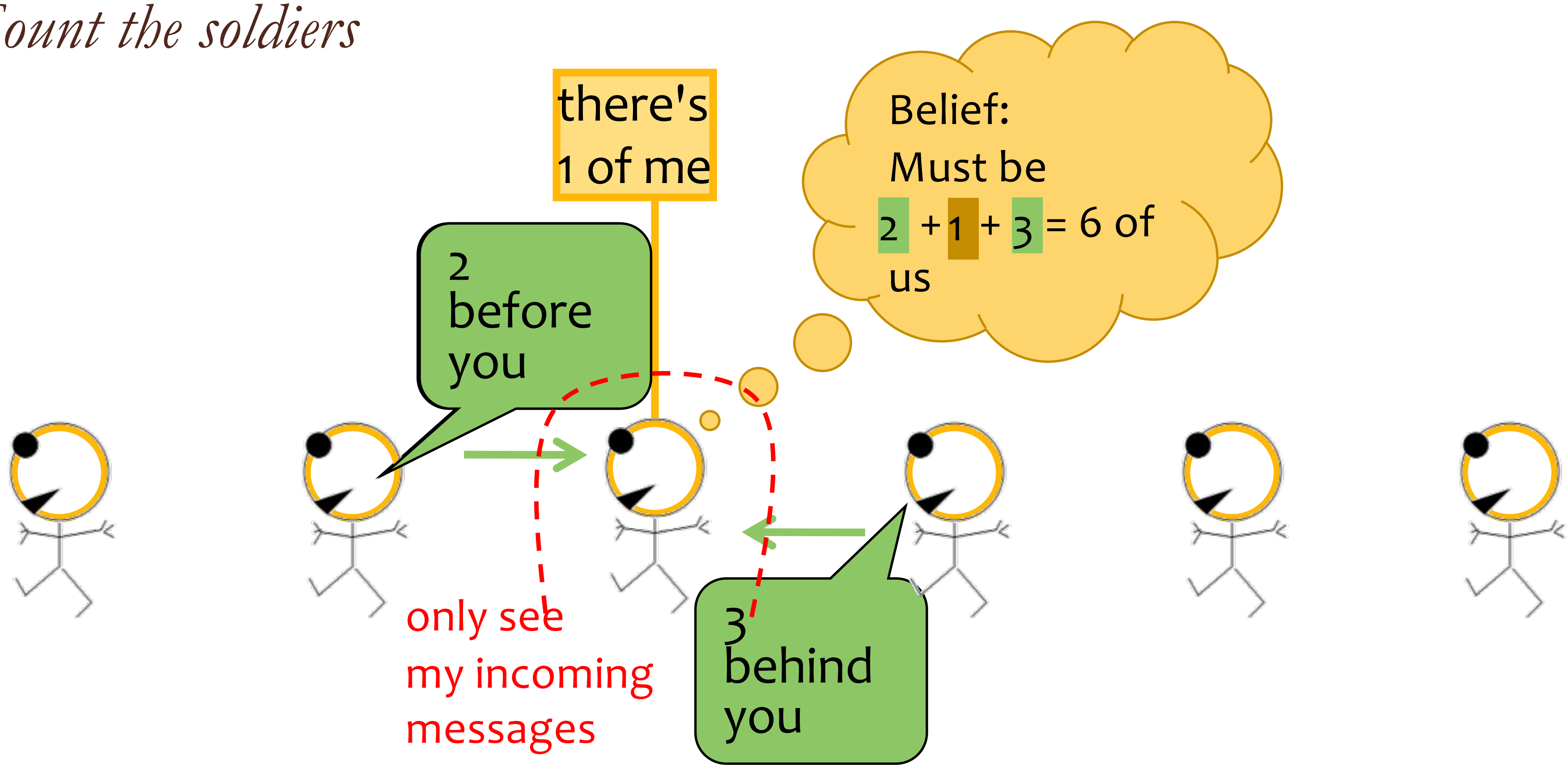
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

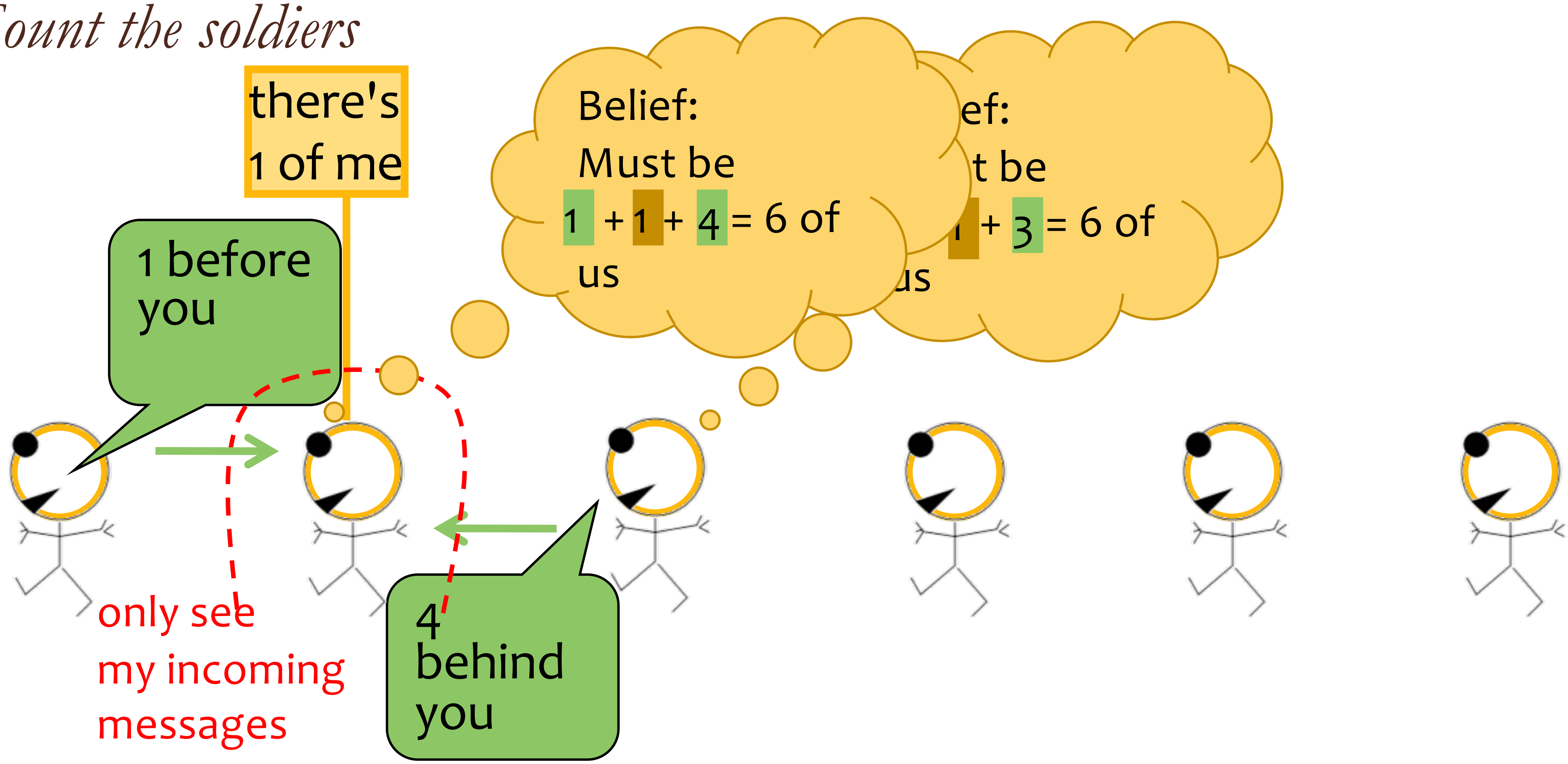
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

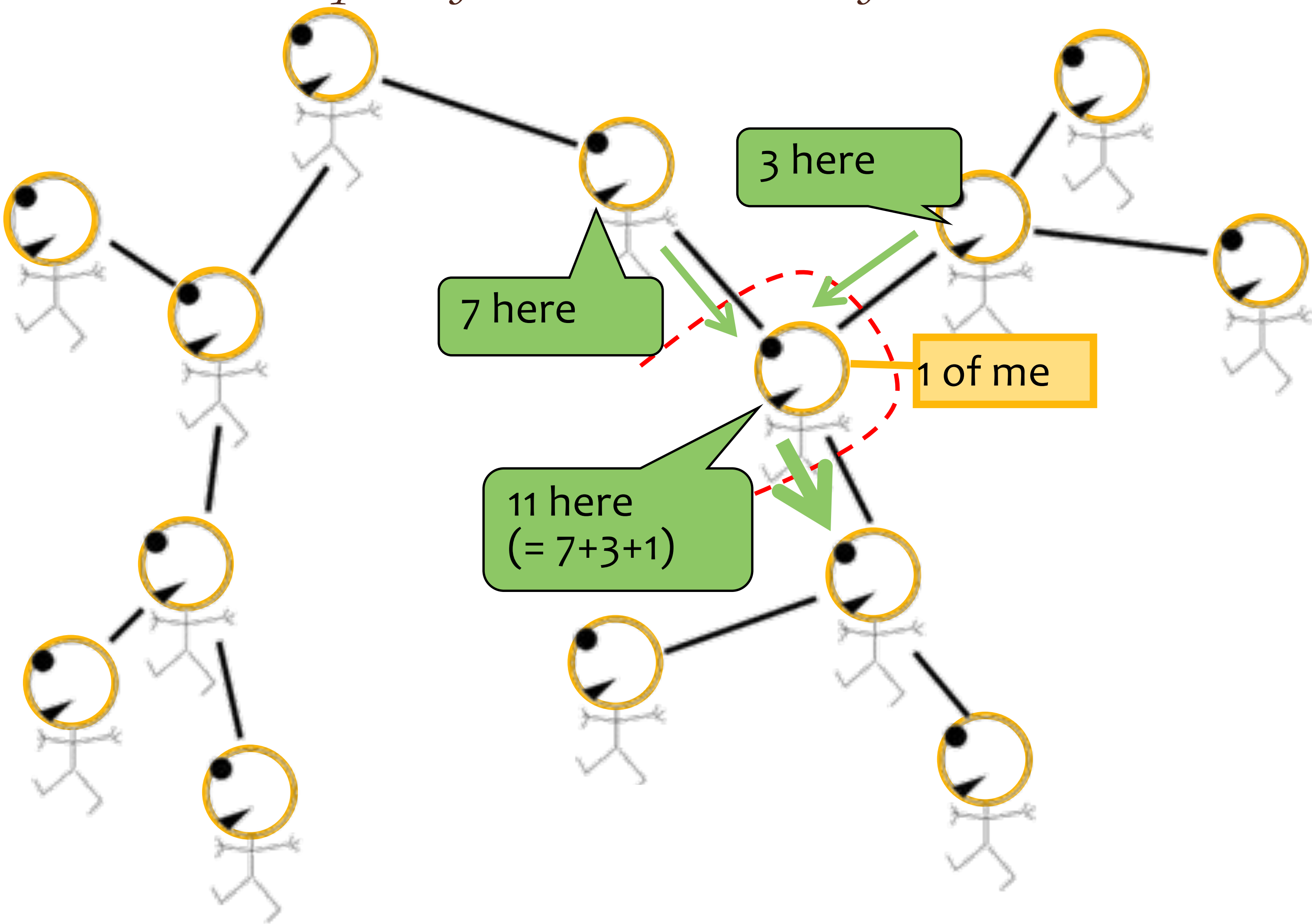
Count the soldiers



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

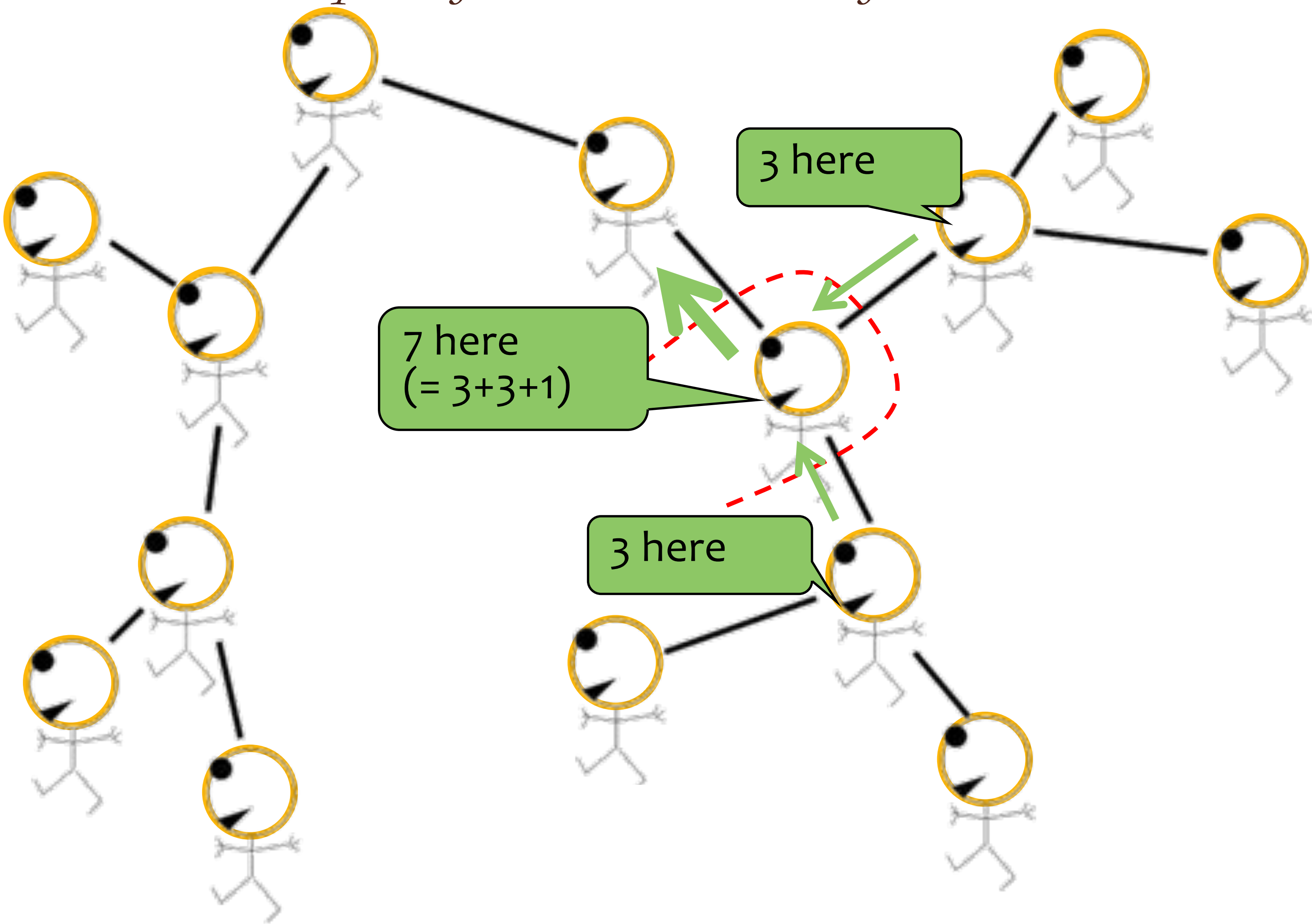
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

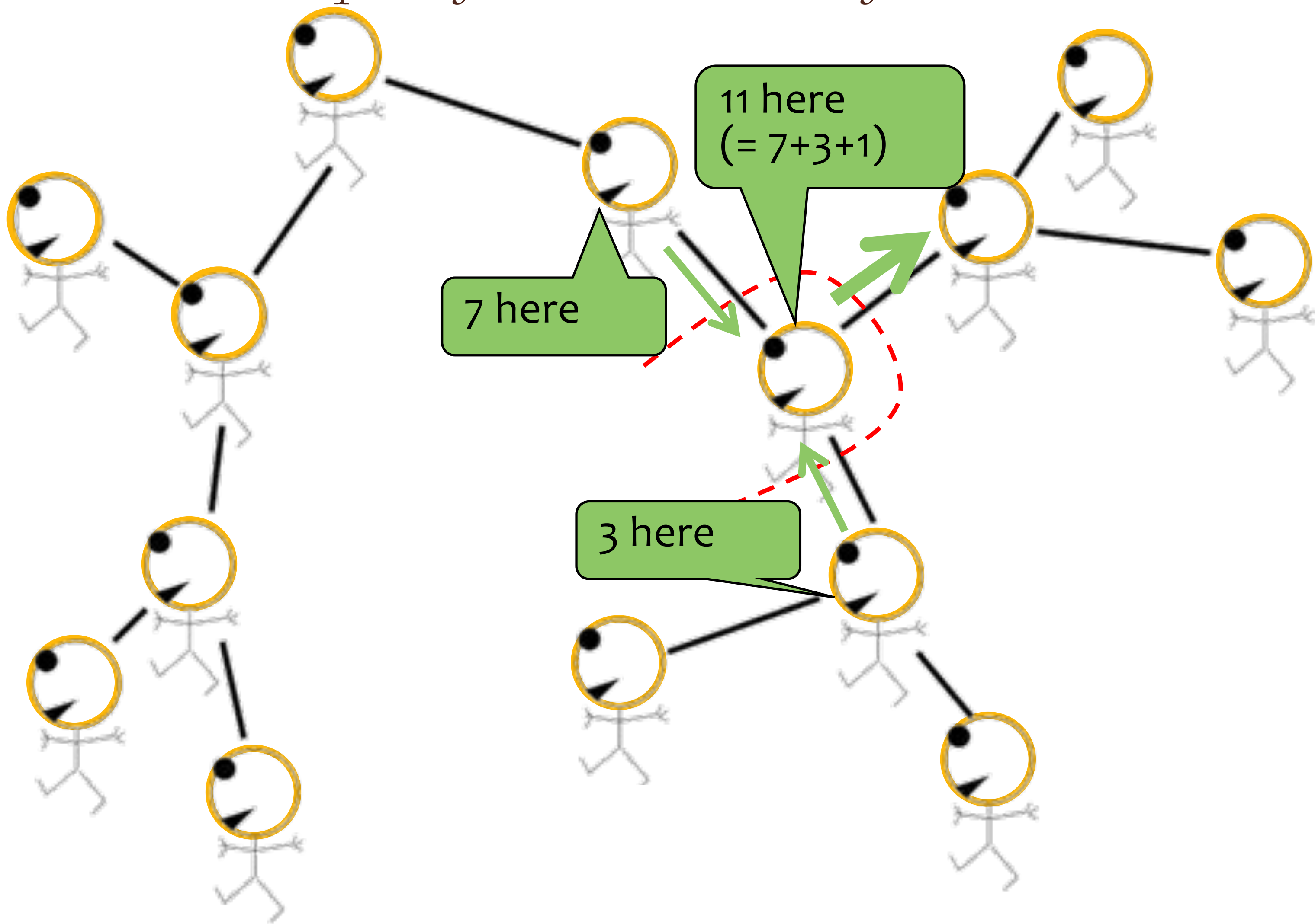
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

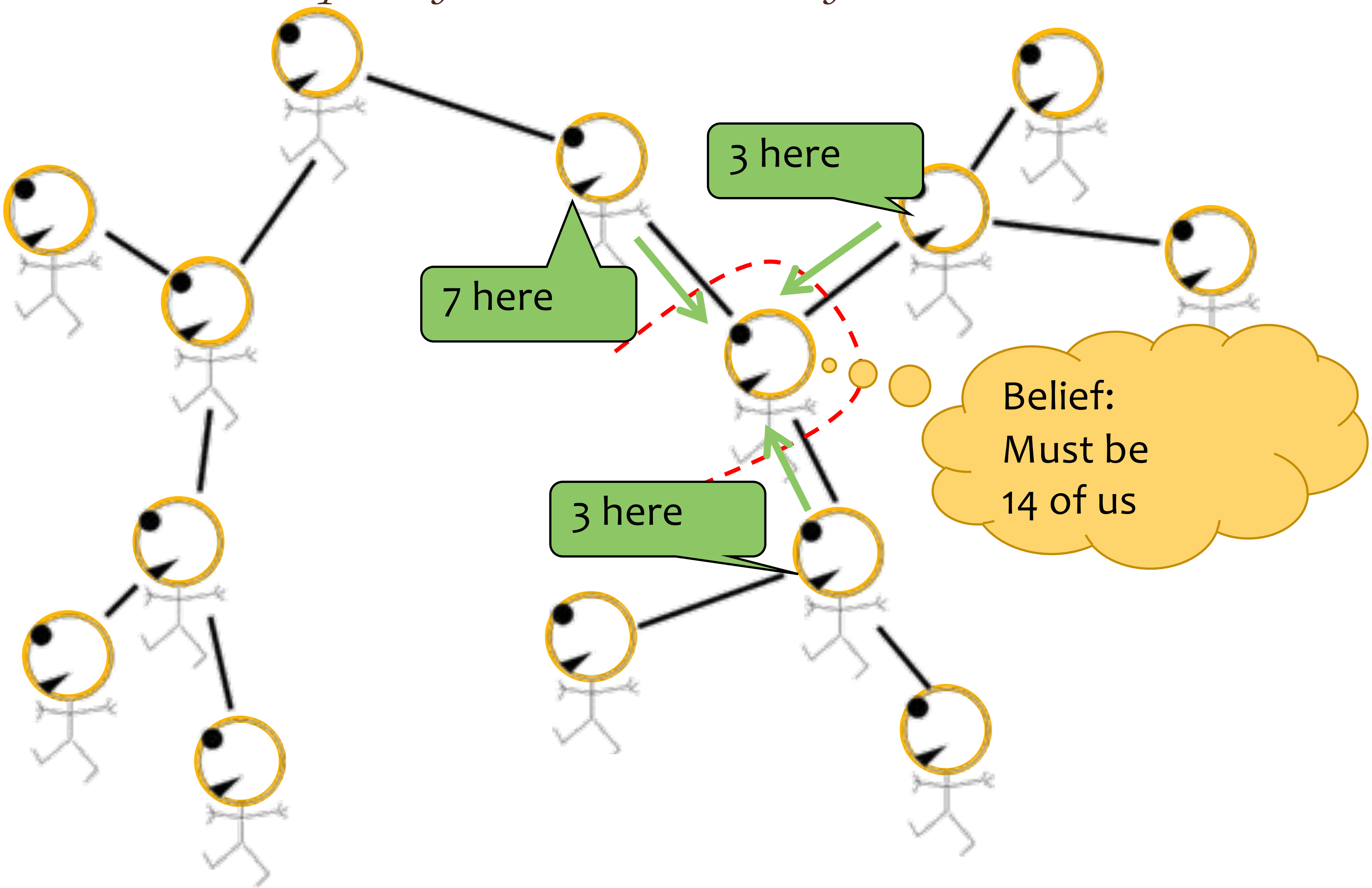
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

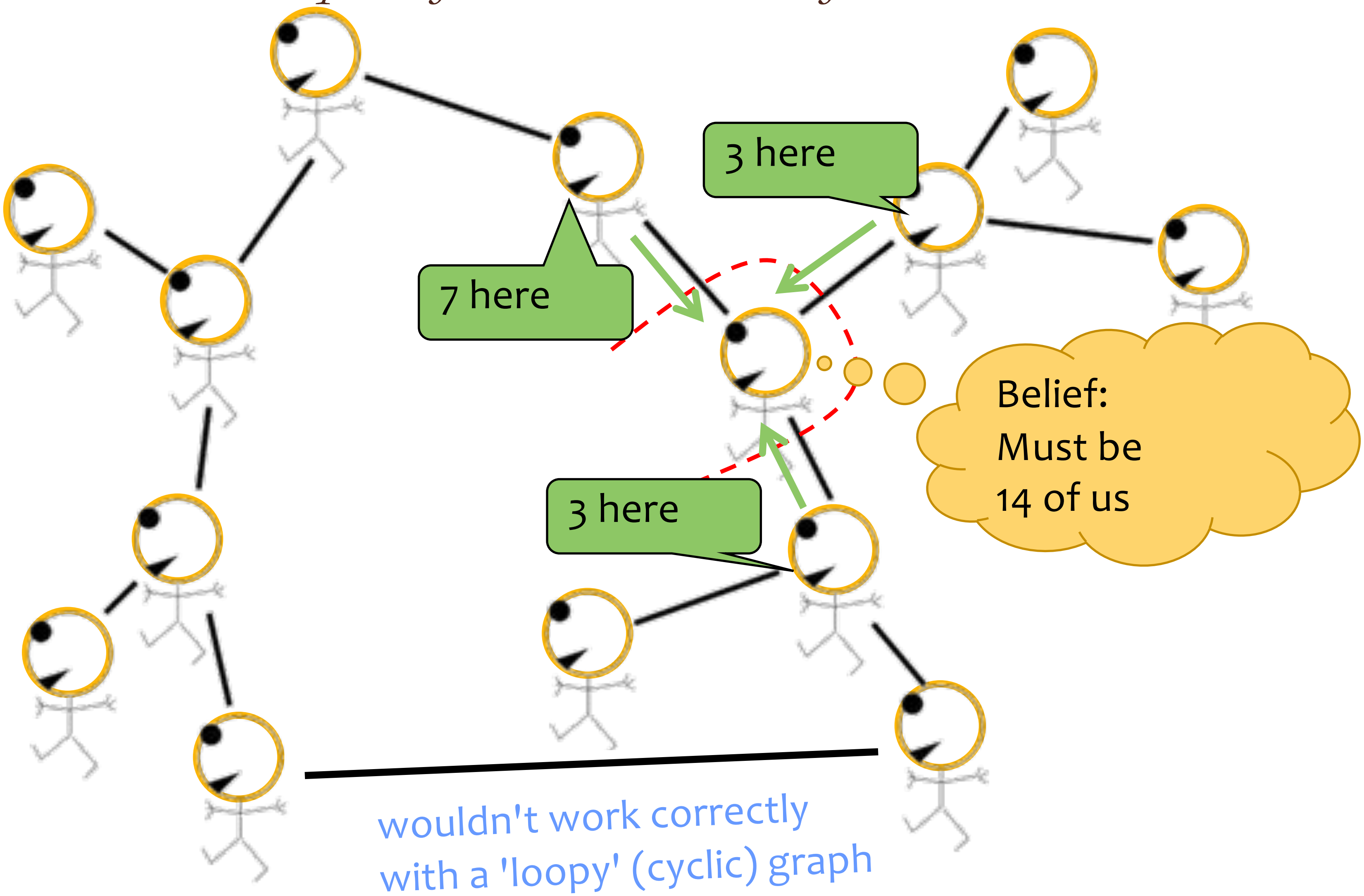
Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

Great Ideas in ML: Message Passing

Each soldier receives reports from all branches of tree



adapted from MacKay (2003) textbook

THE FORWARD-BACKWARD ALGORITHM

Inference

Question:

True or False: The **joint probability of the observations and the hidden states** in an HMM is given by:

$$P(\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}) = C_{y_1} \left[\prod_{t=1}^{T} A_{y_t, x_t} \right] \left[\prod_{t=1}^{T-1} B_{y_{t+1}, y_t} \right]$$

Recall:

Emission matrix, $\mathbf{A}$, where $P(X_t = k | Y_t = j) = A_{j,k}, \forall t, k$

Transition matrix, $\mathbf{B}$, where $P(Y_t = k | Y_{t-1} = j) = B_{j,k}, \forall t, k$

Initial probs, $\mathbf{C}$, where $P(Y_1 = k) = C_k, \forall k$

Inference

Question:

True or False: The **probability of the observations** in an HMM is given by:

$$P(\mathbf{X} = \mathbf{x}) = \prod_{t=1}^{T} A_{x_t, x_{t-1}}$$

Recall:

Emission matrix, $\mathbf{A}$, where $P(X_t = k | Y_t = j) = A_{j,k}, \forall t, k$

Transition matrix, $\mathbf{B}$, where $P(Y_t = k | Y_{t-1} = j) = B_{j,k}, \forall t, k$

Initial probs, $\mathbf{C}$, where $P(Y_1 = k) = C_k, \forall k$

Inference for HMMs

Whiteboard

– Three Inference Problems for an HMM

1. Evaluation: Compute the probability of a given sequence of observations
2. Viterbi Decoding: Find the most-likely sequence of hidden states, given a sequence of observations
3. Marginals: Compute the marginal distribution for a hidden state, given a sequence of observations