

Genome 540 Discussion

February 22th, 2024

Clifford Rostomily

Assignment 7 Questions?

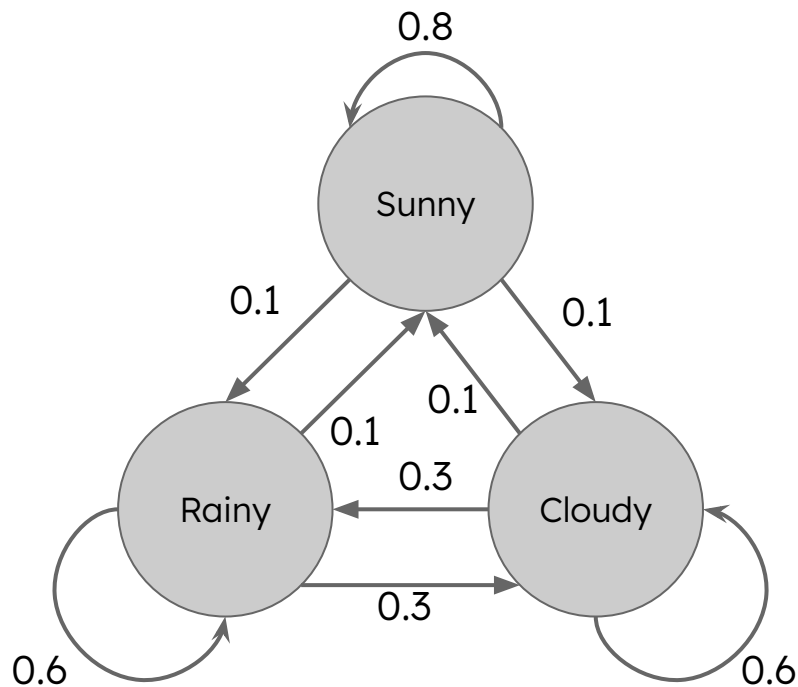
- Part 1: Use your predicted D-segments from hw6 to
 - Generate a new scoring scheme
 - Simulate background sequence
- Part 2: Run your D-segment program on the background and compare to the real data
- Part 3: Answer some questions



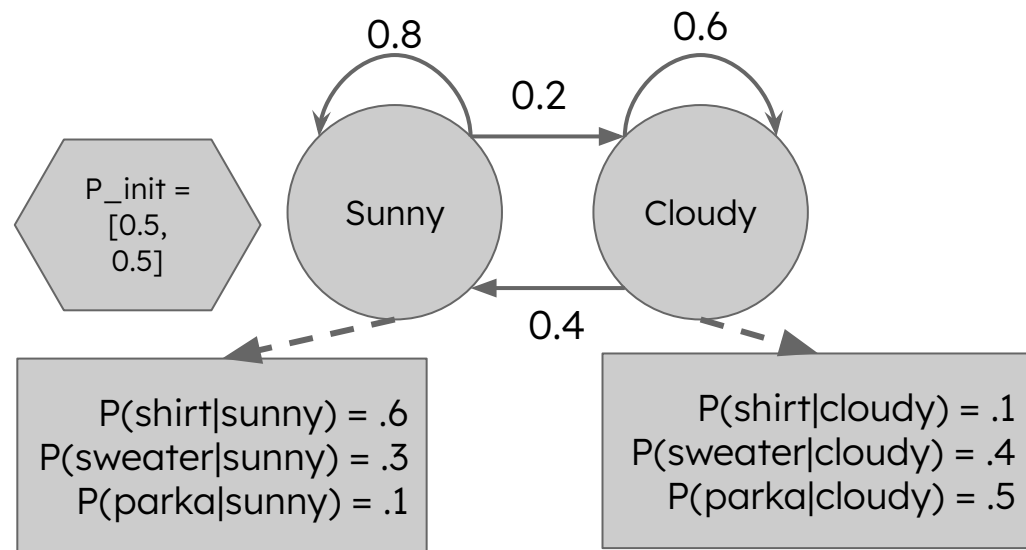
Assignment 8

Markov Chain vs. HMM

Markov Chain



HMM



Markov Chain vs. HMM

Markov Chain

What is the probability of observing this sequence of states?

HMM

What are the most probable (unobserved) states given my observations?

e.g. I observed the sequence ATG, am I in a gene?

HMM Tasks

Rabiner 1989:

Likelihood: Given an HMM $\lambda = (A, B)$ and an observation sequence O , determine the *likelihood* $P(O|\lambda)$.

Decoding: Given an observation sequence O and an HMM $\lambda = (A, B)$, discover the *best hidden state* sequence Q .

Learning: Given an observation sequence O and the set of states in the HMM, learn the HMM *parameters* A and B .

Example

Your dog is very moody and you want to know when they **like** or **hate** you so you start recording what they are doing when you get home everyday...

Waiting



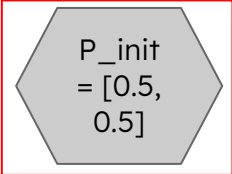
Lounging



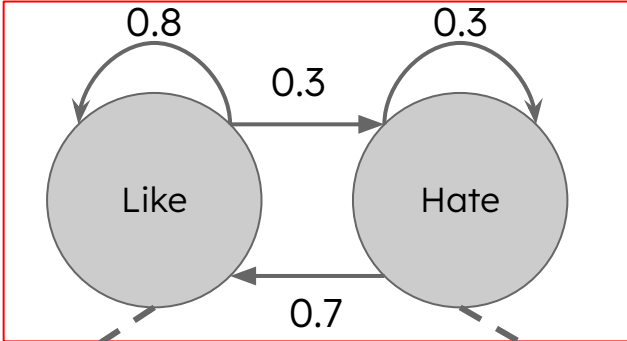
Sleeping



Model



Initiation probabilities (p_k)

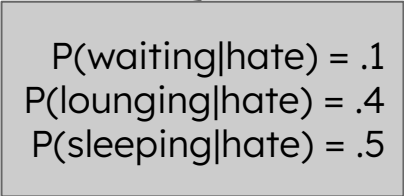
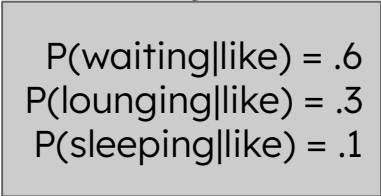


Transition Matrix (A)

	L	H
L	0.8	0.3
H	0.7	0.2

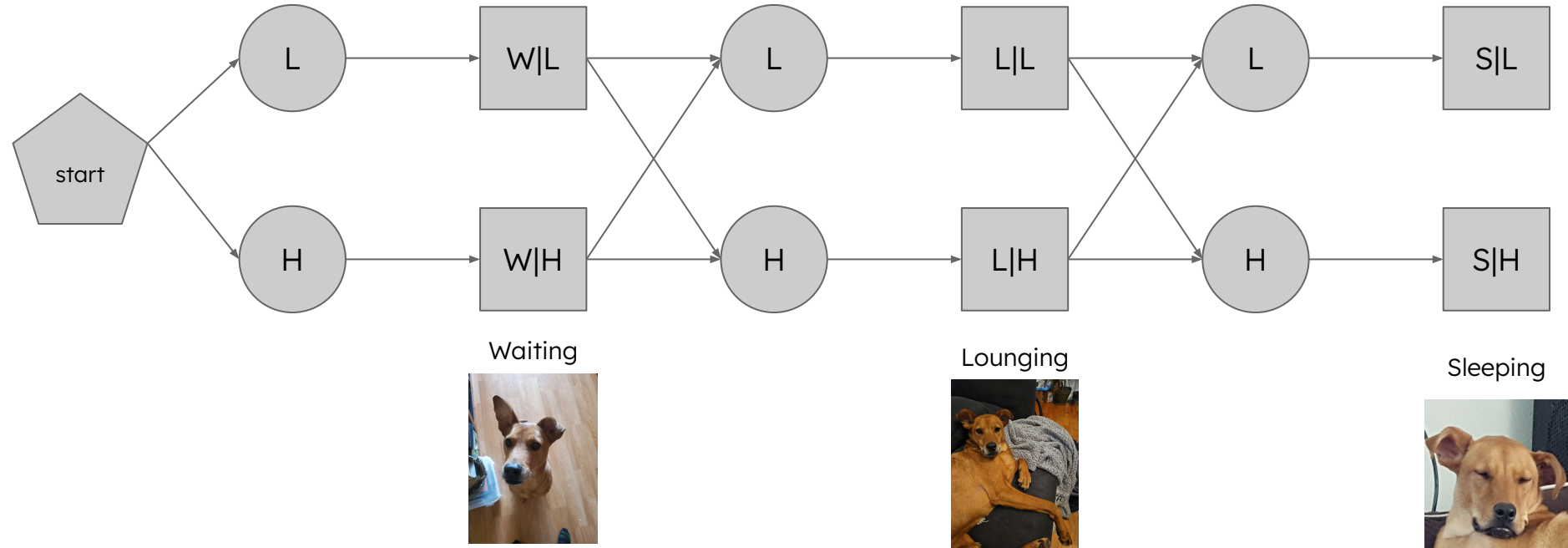
a_{LH}

Set of Hidden States ($S = \{k\}$)

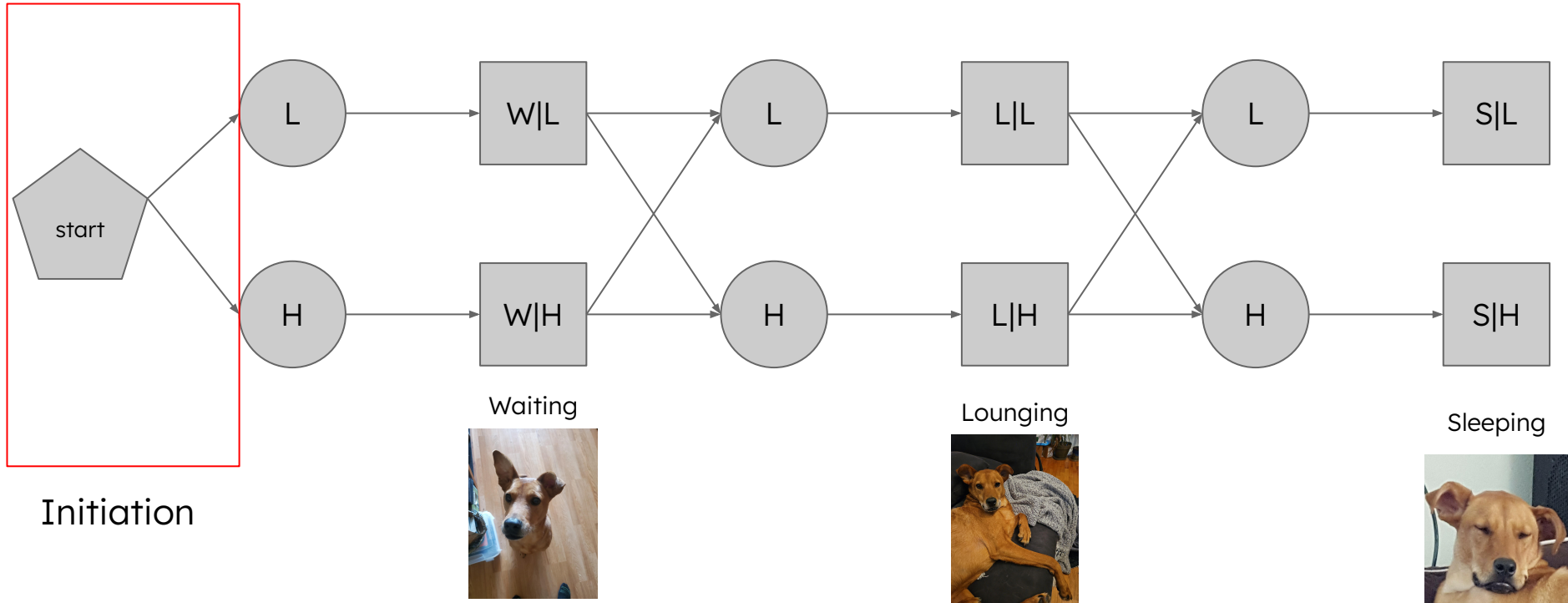


Emission Probabilities of the observations ($B = \{b\}$) given the state

Graphical representation with data

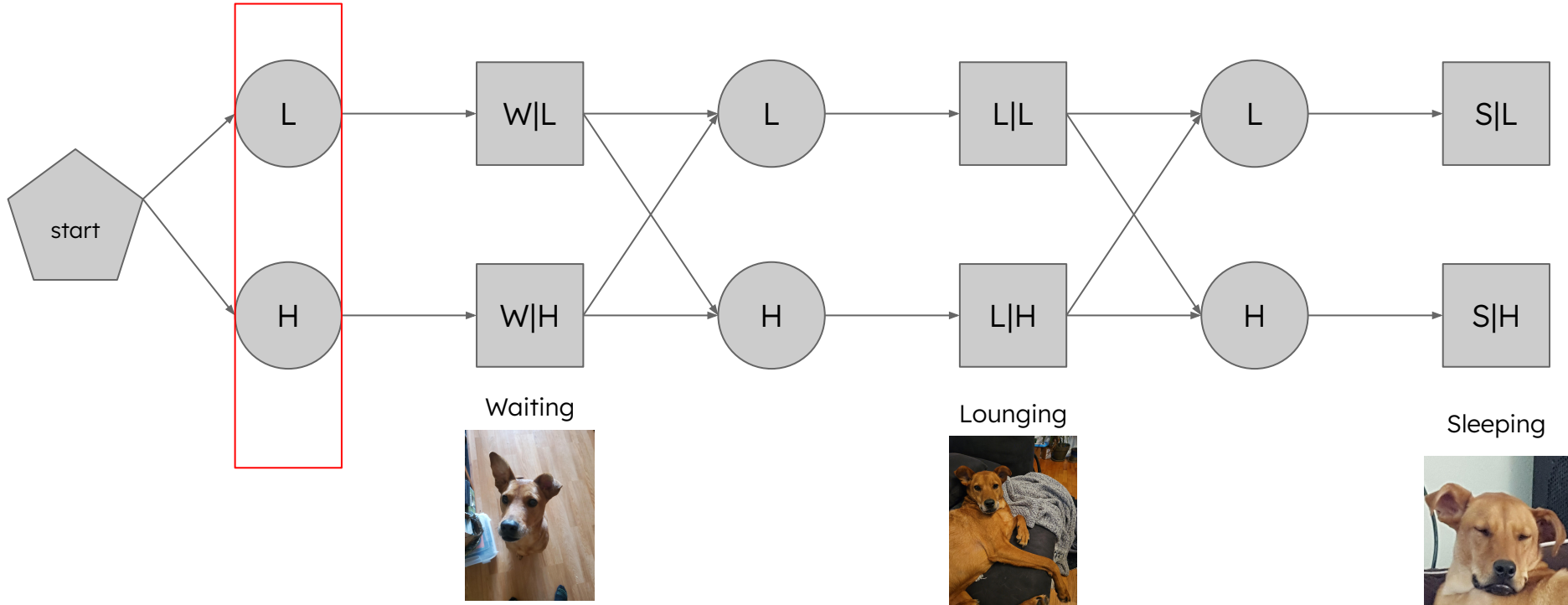


Graphical representation with data



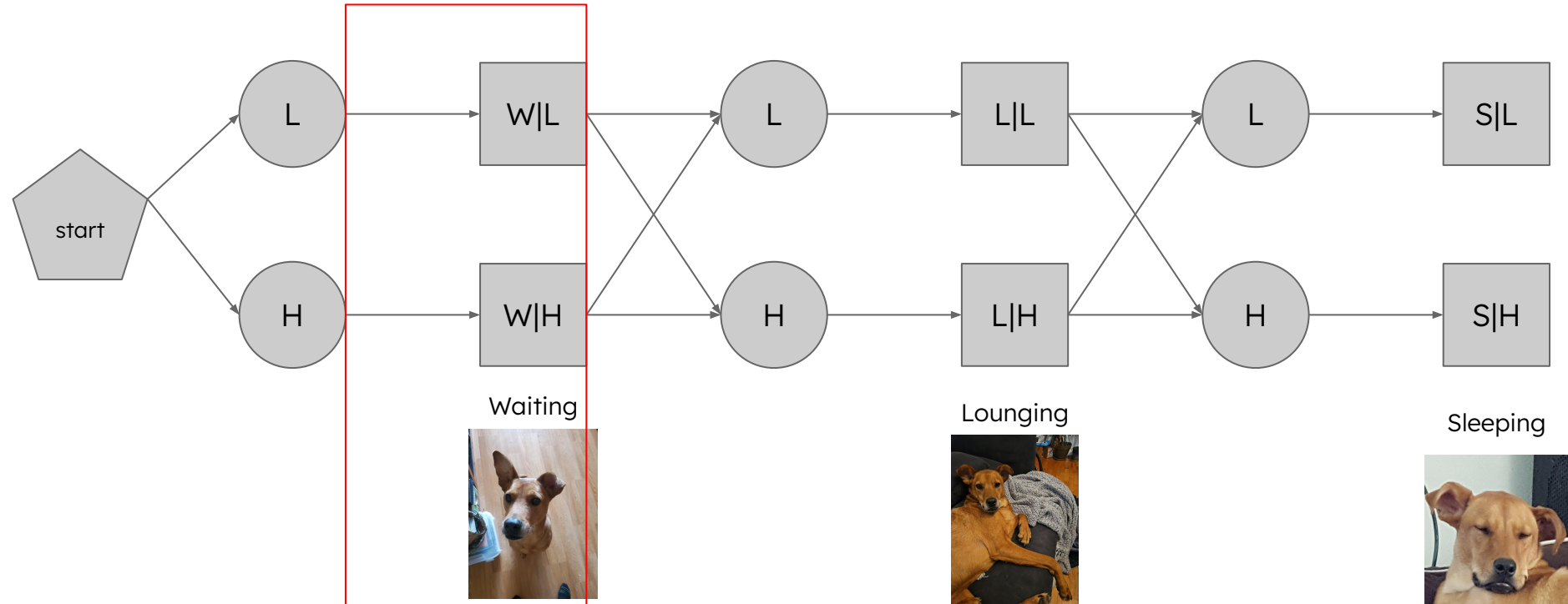
Graphical representation with data

State 1



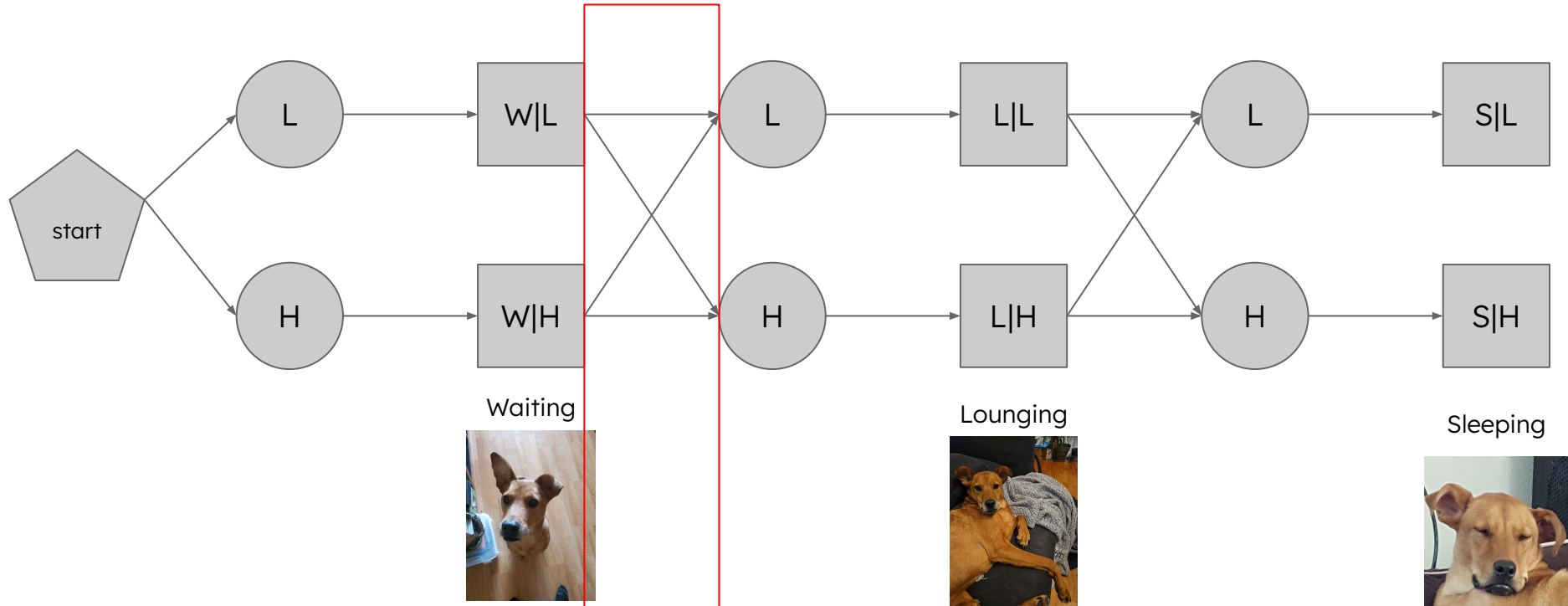
Graphical representation with data


Emission



Graphical representation with data

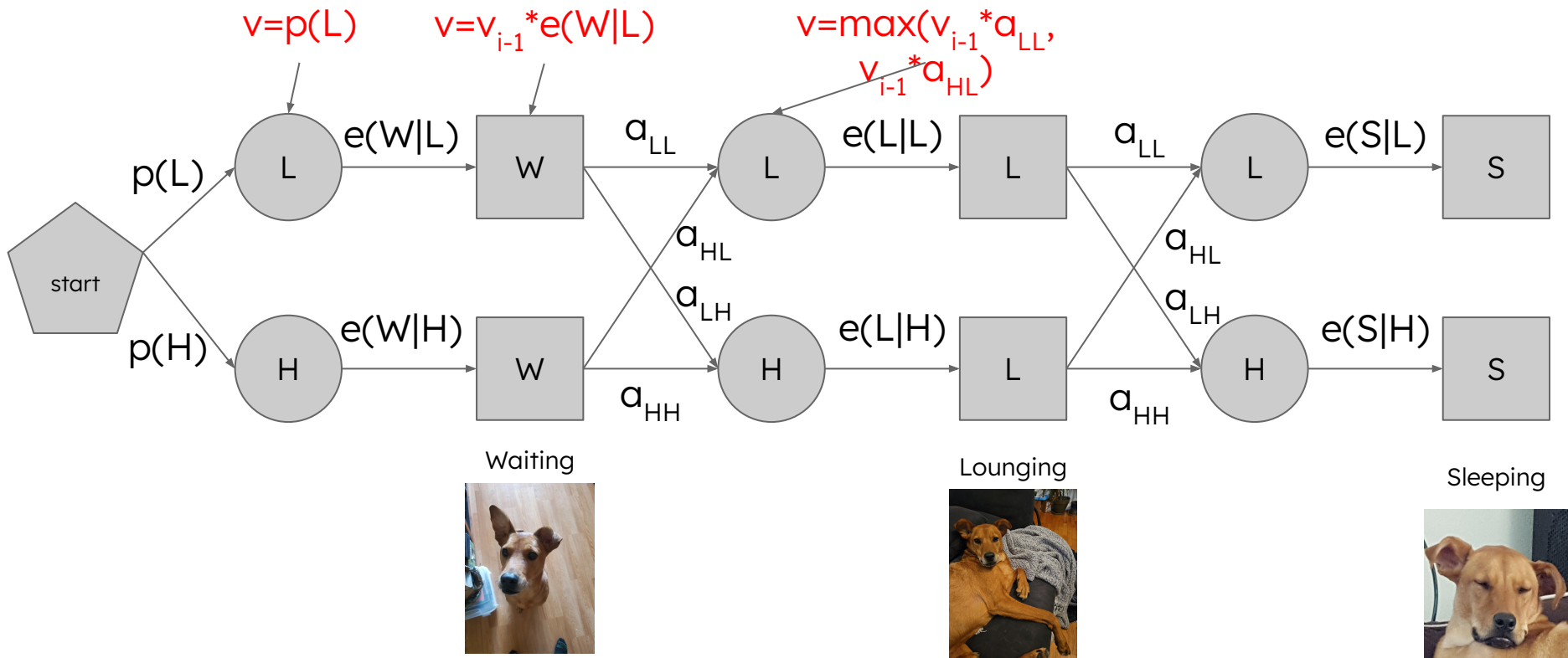
Transition





What is the most optimal state sequence given our model?

Viterbi - Most probable sequence of states





How probable is our model given the data?

Forward Algorithm - **Likelihood** of an observed sequence

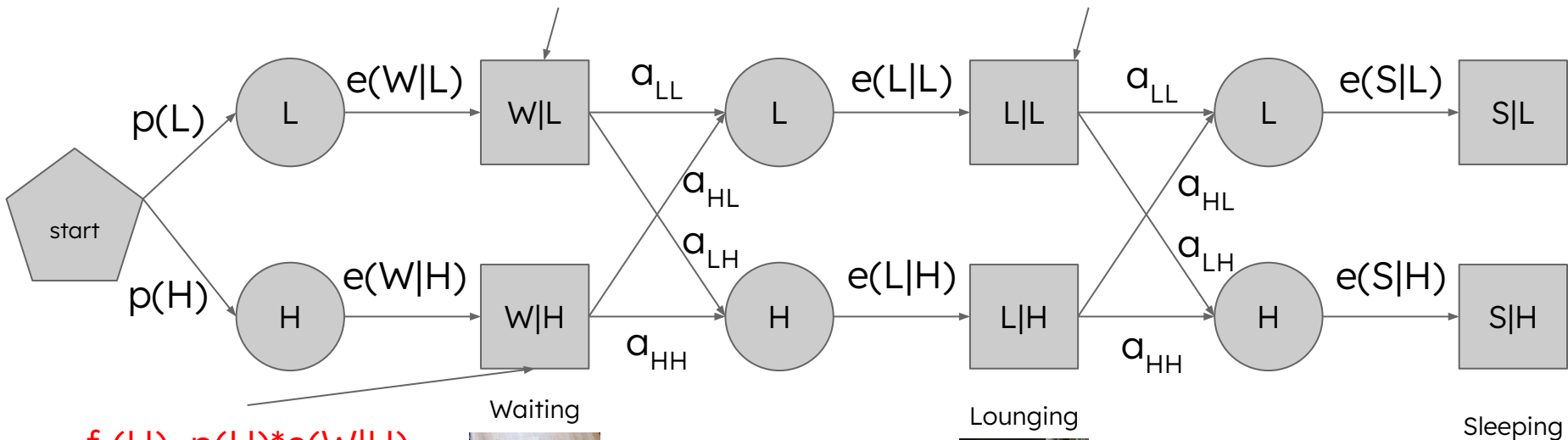
3 steps:

1. Initialization
2. Recursion
3. Termination

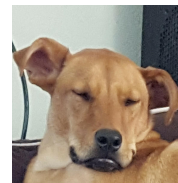
Forward Algorithm - Likelihood of an observed sequence

$$f_1(L) = p(L) * e(W|L)$$

$$f_2(L) = (f_1(L) * a_{LL} + f_1(H) * a_{HL}) * e(L|L)$$



$$f_1(H) = p(H) * e(W|H)$$




Forward Algorithm - **Likelihood** of an observed sequence

Finally...

Sum over all state probabilities to get

$$P(\text{observations}|\text{model}) = \sum_{\mathbf{i}} f(\mathbf{i})$$



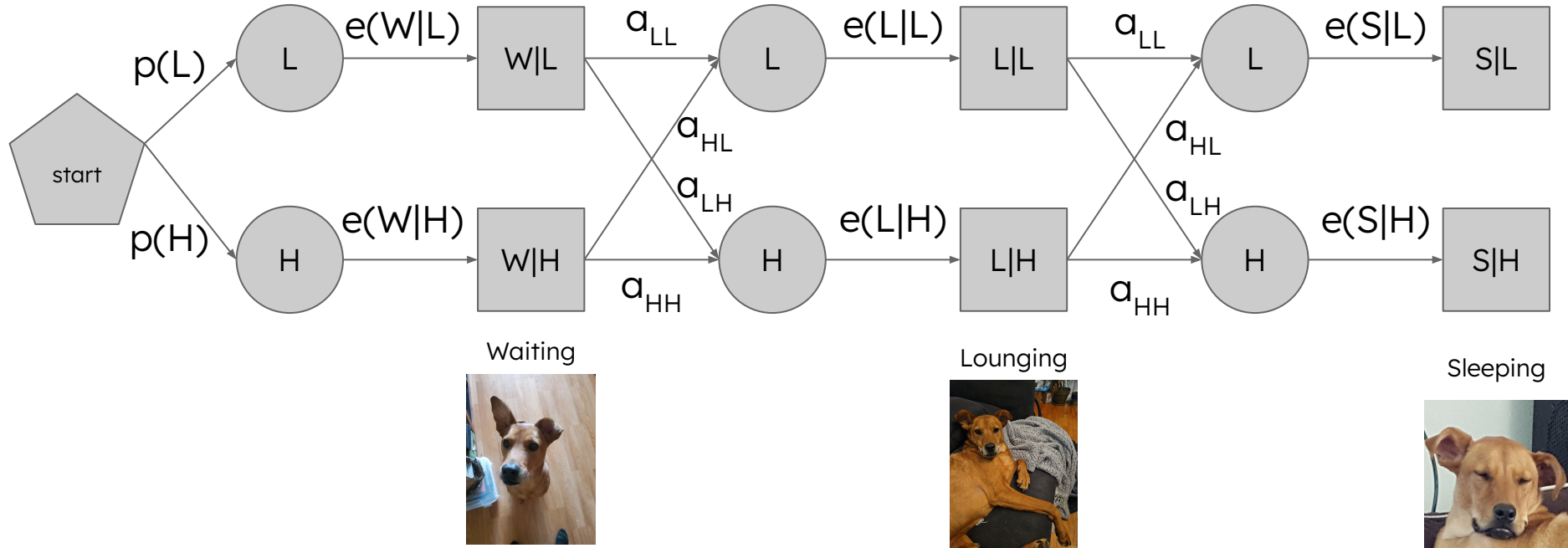
How do we find the transition and emission probabilities given the data?

Baum Welch (Forward/Backward) - “Training” an HMM

1. Step 1: Expectation
 - a. Compute the forward probabilities
 - b. Compute the backward probabilities
2. Step 3: Maximization
 - a. Update the transition and emission probabilities

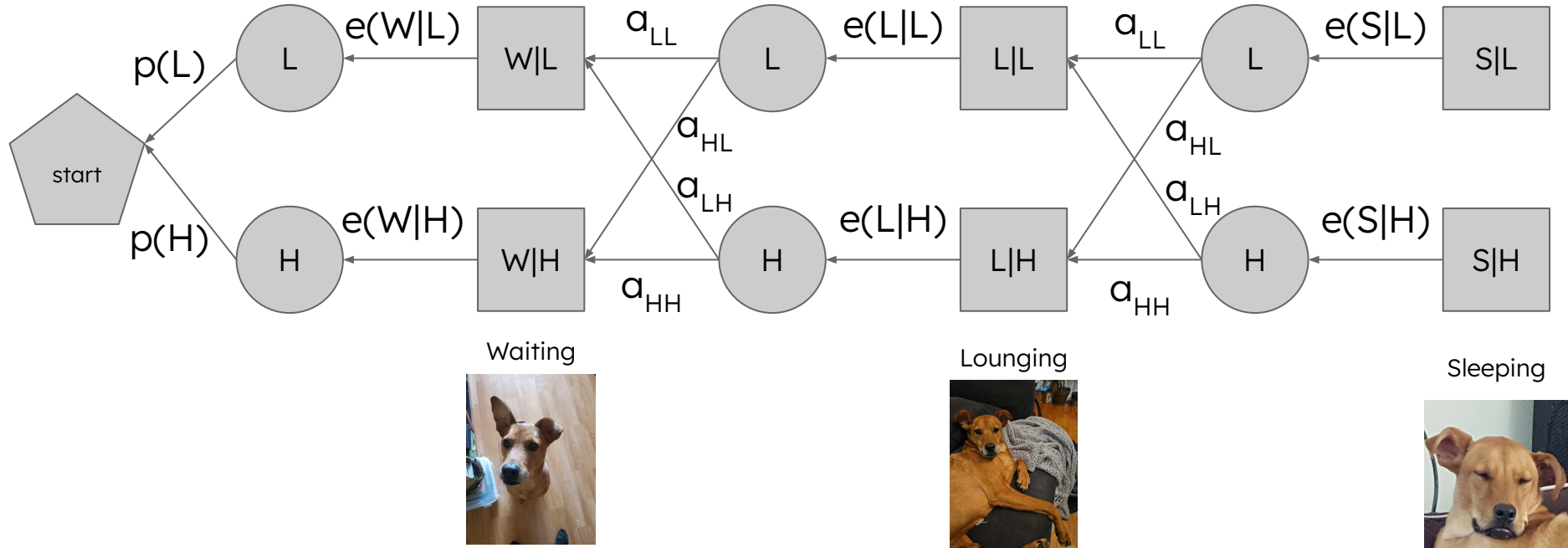
Computing the backward probabilities

Backward probabilities: probability of seeing the observations from time $t + 1$ to the end



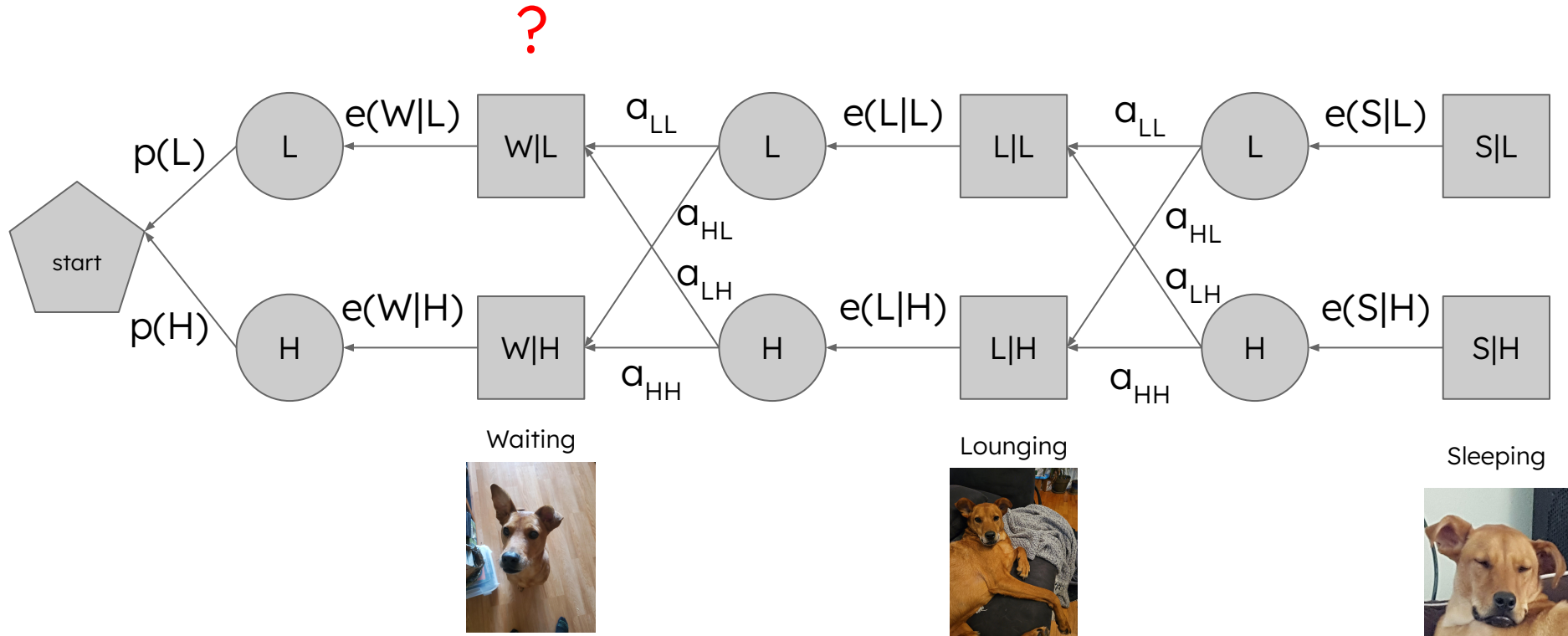
Computing the backward probabilities

Backward probabilities: probability of seeing the observations from time $t + 1$ to the end



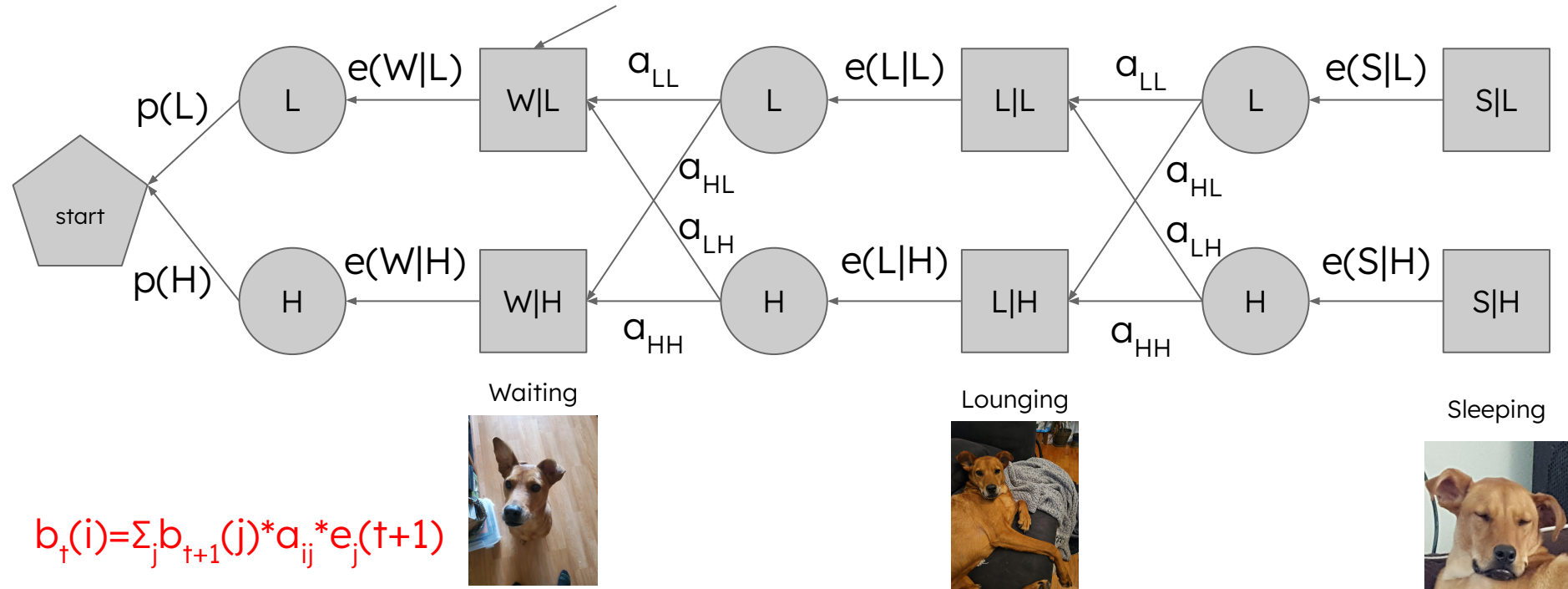
Computing the backward probabilities

?



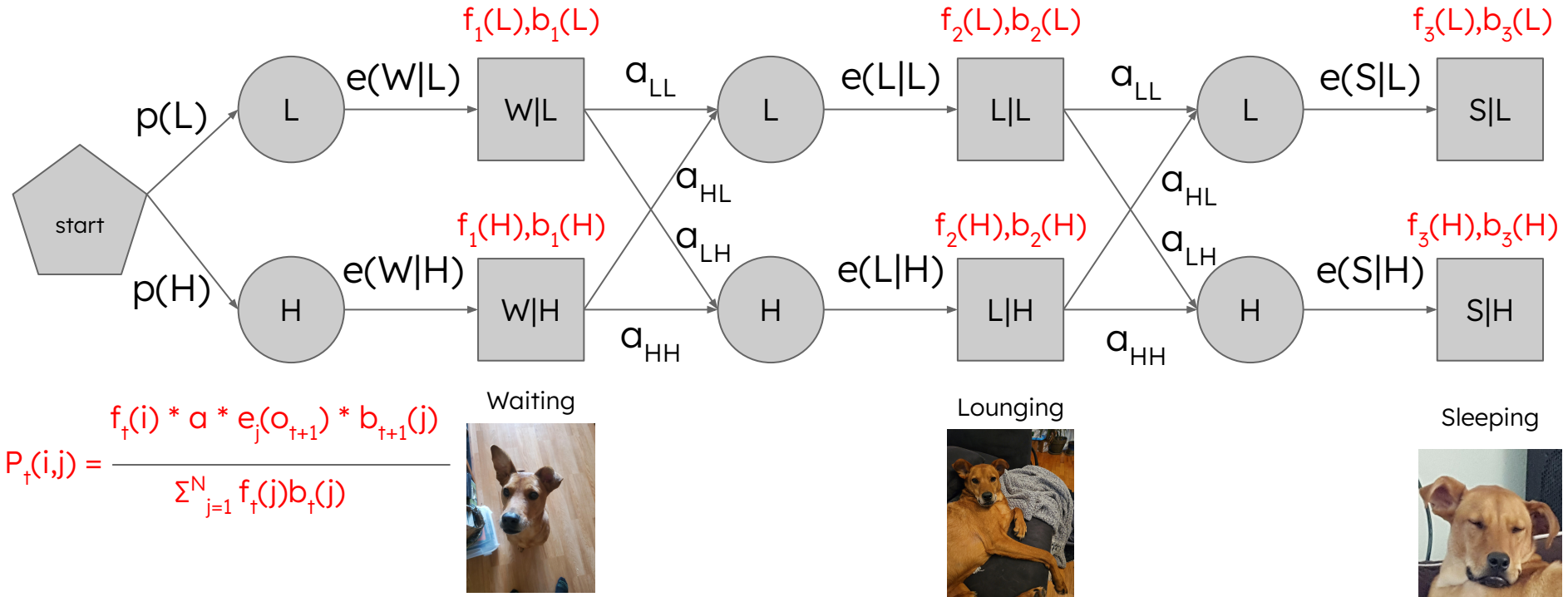
Computing the backward probabilities

$$b_t(i) = b_{t+1}(L) * a_{LL} * e(L|L) + b_{t+1}(H) * a_{LH} * e(L|H)$$



$$b_t(i) = \sum_j b_{t+1}(j) * a_{ij} * e_j(t+1)$$

Calculating the transition probabilities



Calculating the transition probabilities

$$P_t(i,j) = \frac{f_t(i) * a * e_j(o_{t+1}) * b_{t+1}(j)}{\sum_{j=1}^N f_t(j)b_t(j)}$$

$$\underline{q}(i,j) = \frac{\sum_{t=1}^{T-1} P_t(i,j)}{\sum_{t=1}^{T-1} \sum_{k=1}^N P_t(i,k)}$$

Calculating the emission probabilities

Next time



To be continued...



Avoiding vanishing probabilities

- Work in log space
- Scaling

Reminders

- HW7 due this Sunday, 11:59pm
- Please have your name in the filename of your homework assignment and match the template