Modeling Sequence Data: HMMs and Viterbi

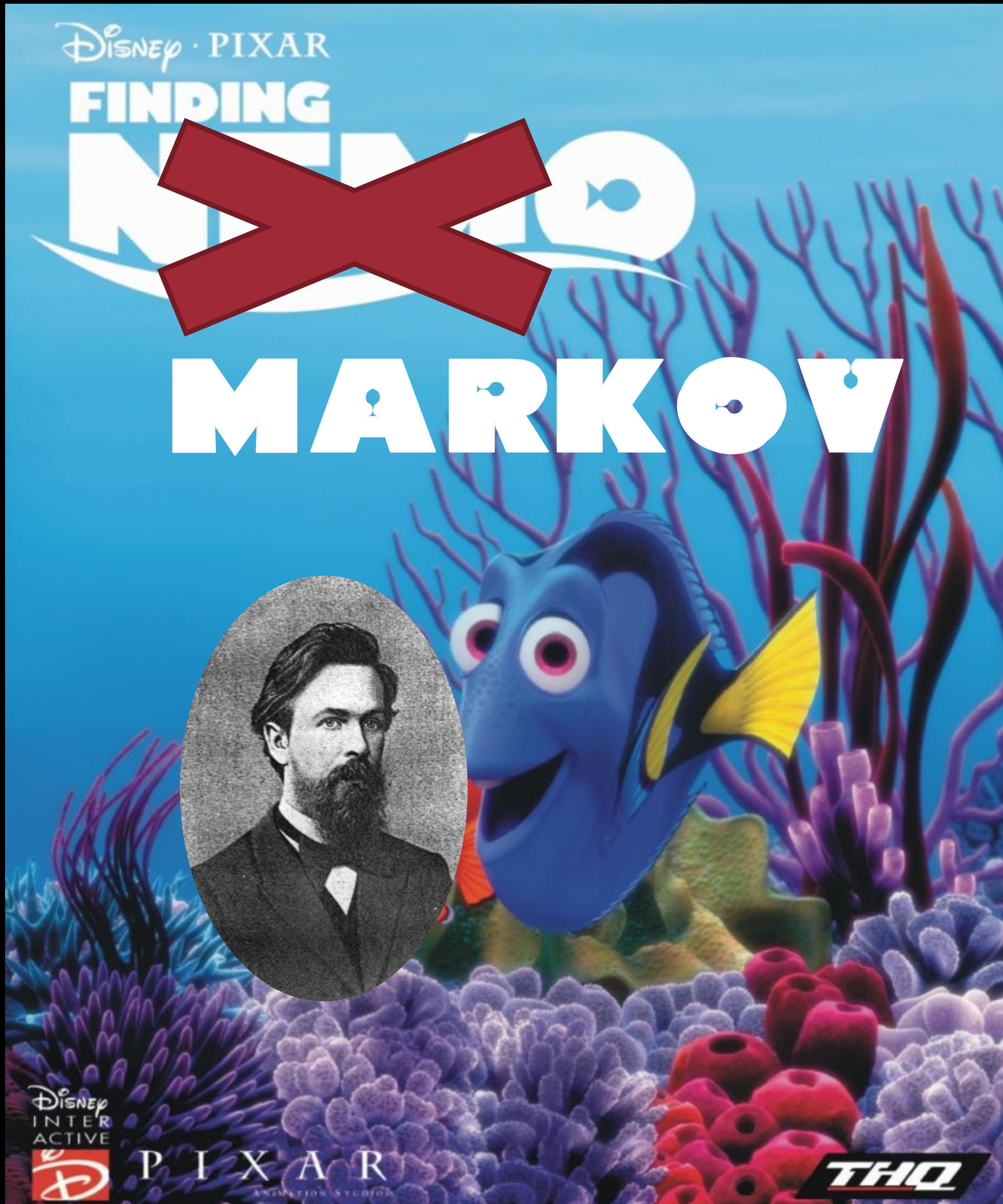
CS4780/5780 – Machine Learning
Fall 2014

Tobias Schnabel and Igor Labutov
Cornell University

Reading:
Manning/Schütze, Sections 9.1-9.3 (except 9.3.1)
Leeds Online HMM Tutorial (except Forward and Forward/Backward Algorithm)
(http://www.comp.leeds.ac.uk/roger/HiddenMarkovModels/html_dev/main.html)

Outline

- Hidden Markov Models
  - Viterbi Algorithm
  - Estimation with fully observed training data
  - Applications: Part-of-speech tagging


Disney · PIXAR
FINDING
~~NEMO~~
MARKOV
Disney
INTERACTIVE
PIXAR
THQ

Hidden Markov Model

- States: $y \in \{s_1, \dots, s_k\}$
- Outputs symbols: $x \in \{o_1, \dots, o_m\}$

Parameter	
Starting probability	$P(Y_1 = y_1)$
Transition probability	$P(Y_i = y_i \mid Y_{i-1} = y_{i-1})$
Output/Emission probability	$P(X_i = x_i \mid Y_i = y_i)$

Hidden Markov Model

- Every output/state sequence has a probability

$$P(x,y) = P(x_1, \dots, x_l, y_1, \dots, y_l)$$
$$= P(y_1)P(x_1|y_1)\prod_{i=2}^{l} P(x_i|y_i)P(y_i|y_{i-1})$$

- Different visualizations

Estimating the Probabilities

- Fully observed data:

$$P(Y_i = a \mid Y_{i-1} = b) = \frac{\text{\# of times state } a \text{ follows state } b}{\text{\# of times state } b \text{ occurs}}$$

$$P(X_i = a \mid Y_i = b) = \frac{\text{\# of times output } a \text{ is observed in state } b}{\text{\# of times state } b \text{ occurs}}$$

  - Smoothing the estimates:
    - See Naïve Bayes for text classification
- Partially observed data (Y_i unknown):
  - Expectation-Maximization (EM)

HMM Decoding: Viterbi Algorithm

- Question: What is the most likely state sequence given an output sequence
  - Find $y^* = \underset{y \in \{y_1,\ldots,y_l\}}{\operatorname{argmax}} P(x_1, \ldots, x_l, y_1, \ldots, y_l)$

$$= \underset{y \in \{y_1,\ldots,y_l\}}{\operatorname{argmax}} \left\{ P(y_1) P(x_1|y_1) \prod_{i=2}^{l} P(x_i|y_i) P(y_i|y_{i-1}) \right\}$$

Going on a trip

- Deal: 3 trips to cities 3 different countries:

Going on a trip

- Deal: 3 trips to cities 3 different countries:

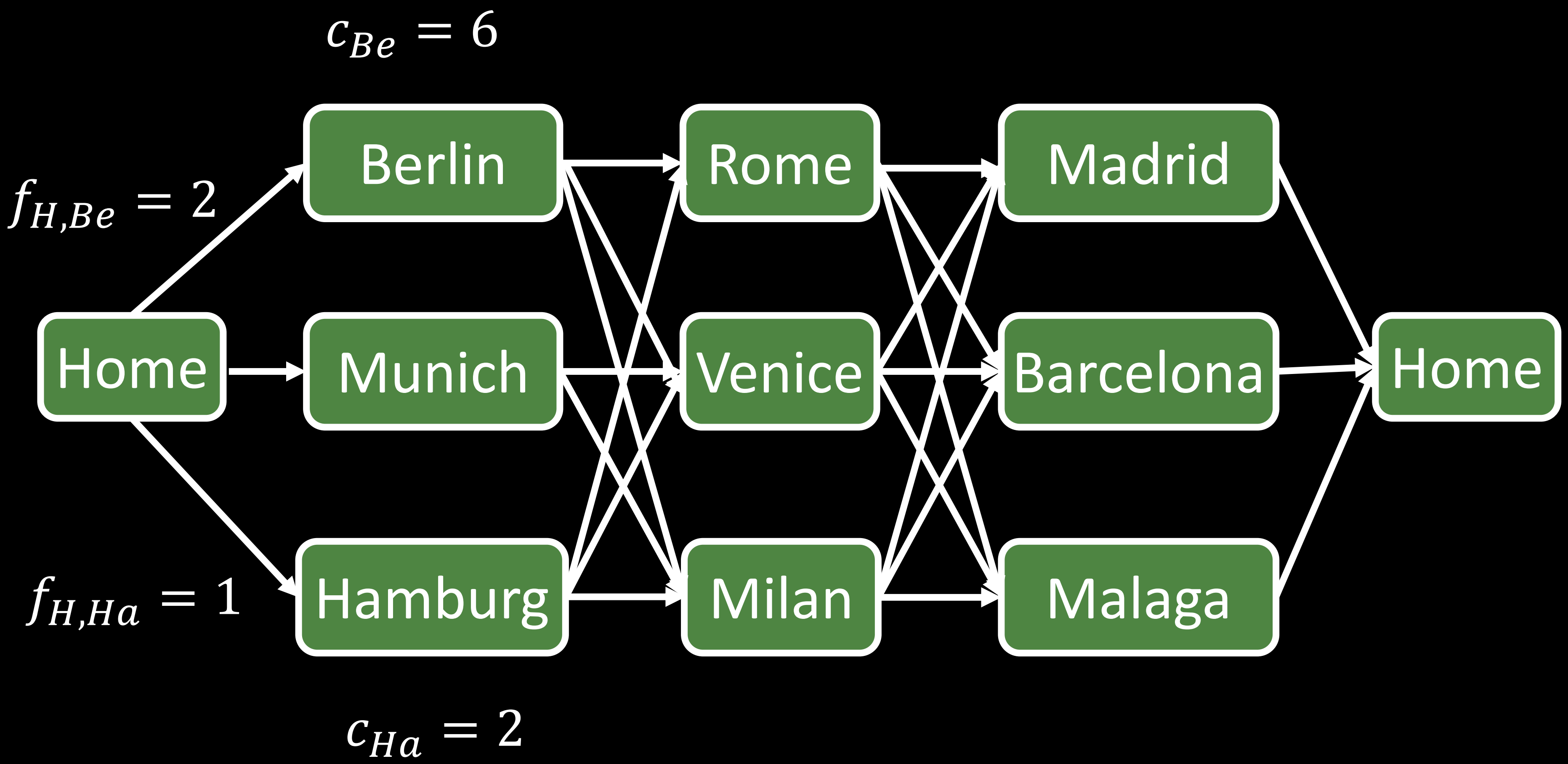
Country	City options
Germany	Berlin/Munich/Hamburg
Italy	Rome/Venice/Milan
Spain	Madrid/Barcelona/Malaga

Going on a trip

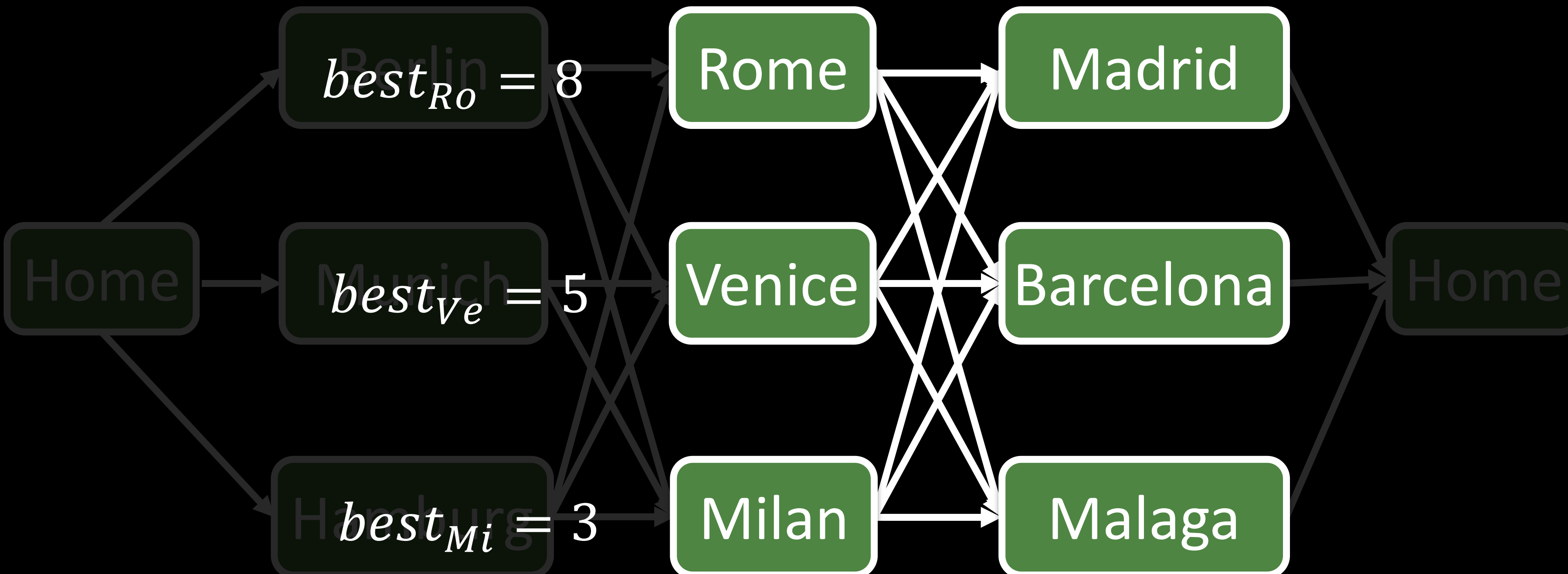
- Deal: 3 trips to cities 3 different countries:
  - Each city i has an attractiveness score $c_i \in [0, 10]$
  - Each flight has an comfort score $f_{i,j} \in [0, 10]$

- Find the best trip!

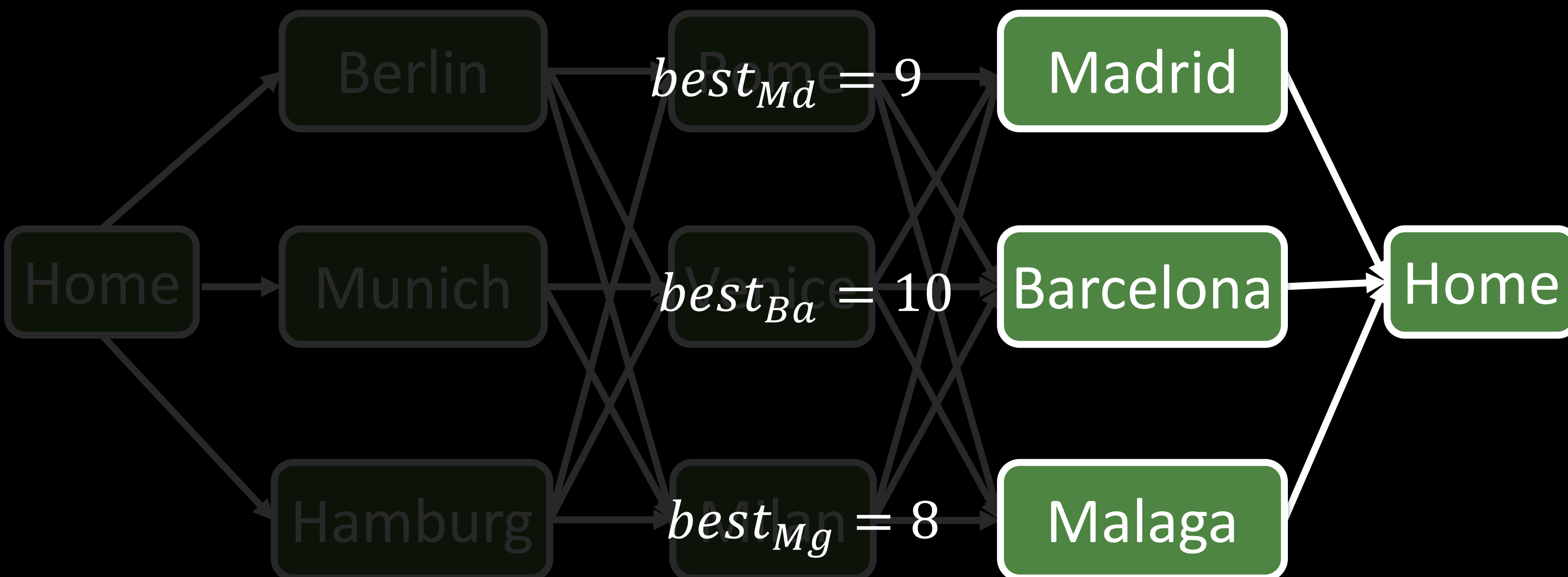
Going on a trip



Going on a trip



Going on a trip



HMM Decoding: Viterbi Algorithm

- Question: What is the most likely state sequence given an output sequence
  - Find $y^* = \underset{y \in \{y_1,\dots,y_l\}}{\operatorname{argmax}} P(x_1, \dots, x_l, y_1, \dots, y_l)$
    $= \underset{y \in \{y_1,\dots,y_l\}}{\operatorname{argmax}} \left\{ P(y_1)P(x_1|y_1) \prod_{i=2}^{l} P(x_i|y_i)P(y_i|y_{i-1}) \right\}$
  - Viterbi algorithm has runtime linear in length of sequence

Viterbi Example

$P(X_i \mid Y_i)$	A+	B	C
happy	0.6	0.3	0.1
grumpy	0.1	0.4	0.5

$P(Y_1)$	
happy	0.7
grumpy	0.3

$P(Y_i \mid Y_{i-1})$	happy	grumpy
happy	0.8	0.2
grumpy	0.3	0.7

- What the most likely mood sequence for x = (C, A+, A+)?

HMM's for POS Tagging

- Design HMM structure (vanilla)
  - States: one state per POS tag
  - Transitions: fully connected
  - Emissions: all words observed in training corpus
- Estimate probabilities
  - Use corpus, e.g. Treebank
  - Smoothing
  - Unseen words?
- Tagging new sentences
  - Use Viterbi to find most likely tag sequence

Experimental Results

Tagger	Accuracy	Training time	Prediction time
HMM	96.80%	20 sec	18.000 words/s
TBL Rules	96.47%	9 days	750 words/s

- Experiment setup
  - WSJ Corpus
  - Trigram HMM model
  - from [Pla and Molina, 2001]

Discriminative vs. Generative

- Bayes Rule: $$\mathrm{h}_{\mathrm{bayes}}(\mathrm{x}) = \underset{y \in \mathrm{Y}}{\operatorname{argmax}} \left[P(Y = y | X = x) \right]$$ $$= \underset{y \in \mathrm{Y}}{\operatorname{argmax}} \left[P(X = x | Y = y) P(Y = y) \right]$$
- Generative:
  - Model $P(X = x | Y = y)$ and $P(Y = y)$
- Discriminative:
  - Find h in H that best approximates the classifications made by $$\mathrm{h}_{\mathrm{bayes}}(\mathrm{x}) = \underset{y \in \mathrm{Y}}{\operatorname{argmax}} \left[P(Y = y | X = x) \right]$$
- Question: Can we train HMM's discriminately?