

In district of county Precinct 10, Plumb Name of institution X  
Jailed city, town, or village, within the above-named district. X Ward of city X  
Discovered by me on the 1st day of June, 1900. John F. Stick Commissioner 1170

# Using a Hidden-Markov Model in Semi-Automatic Indexing of Historical Handwritten Records

Thomas Packer, Oliver Nina, Ilya Raykhel

Computer Science

Brigham Young University

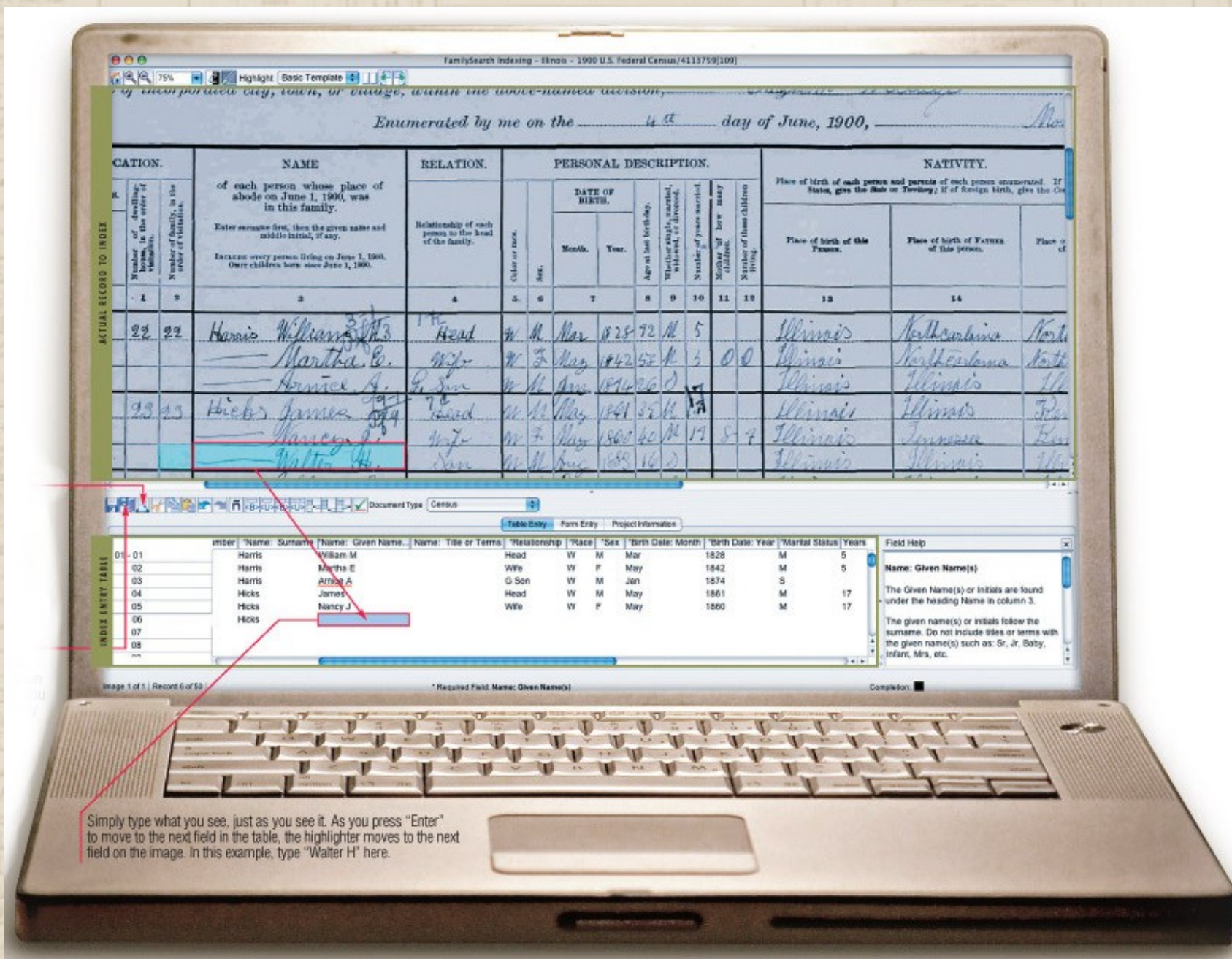
# The Challenge: Indexing Handwriting

- Millions of historical documents.
- Many hours of manual indexing.
- Years to complete using hundreds of thousands of volunteers.
- Previous transcriptions not fully leveraged.





# Family Search Indexing Tool



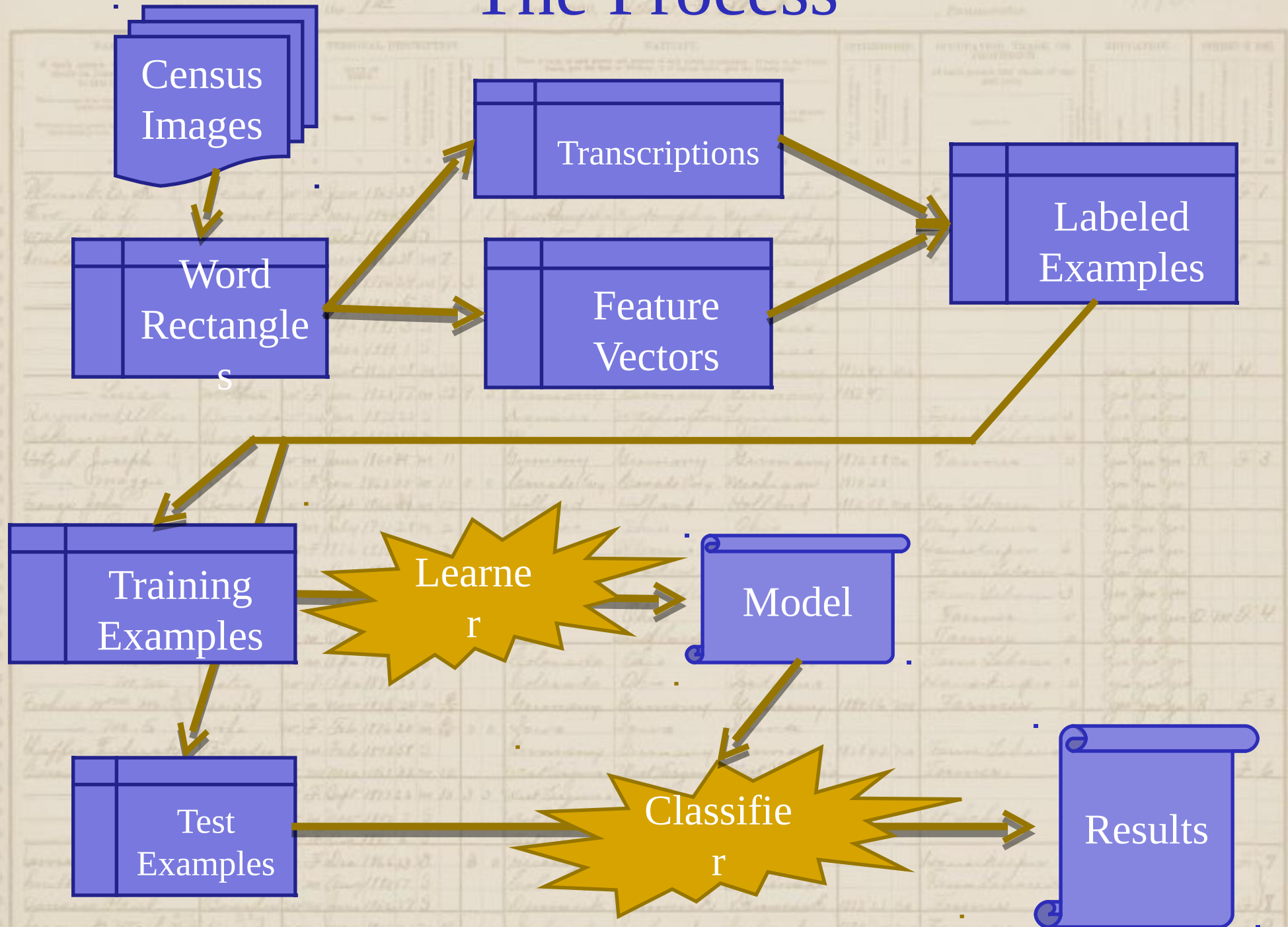
Simply type what you see, just as you see it. As you press "Enter" to move to the next field in the table, the highlighter moves to the next field on the image. In this example, type "Walter H" here.

# A Solution: On-Line Machine Learning

- Holistic handwritten word recognition using a Hidden Markov Model (HMM), based on Lavrenko et al. (2004).
- HMM selects words to maximize joint probability:
  - Word-feature probability model
  - Word-transition probability model
- Word-feature model predicts a word from its visual features.
- Word-transition model predicts a word from its neighboring word.



# The Process



# Census Images

- 3 US Census images
- Same census taker
- Preprocessing: Kittler's algorithm to threshold images

RELATION.	PERSONAL DESCRIPTION.						
	Color or race.	Sex.	DATE OF BIRTH.		Age at last birthday.	Whether single, married, widowed, or divorced.	Number of years married.
			Month.	Year.			
4	5	6	7		8	9	10
1913 Head	w m		Jan	1865	33	S	
Servant.	w f		May	1844	56	D.	
Boarder	w m		Oct	1874	23	S.	

RELATION.	PERSONAL DESCRIPTION.						
	Color or race.	Sex.	DATE OF BIRTH.		Age at last birthday.	Whether single, married, widowed, or divorced.	Number of years married.
			Month.	Year.			
4	5	6	7		8	9	10
1913 Head	w m		Jan	1865	33	S	
Servant.	w f		May	1844	56	D.	
Boarder	w m		Oct	1874	23	S.	

# Extracted Fields

- Manually copied bounding rectangles
- 3 columns:
  1. Relationship to Head (14)
  2. Sex (2)
  3. Marital Status (4)
- 123 rows total
- N-fold cross validation
- N = 24 (5 rows to test)

Head	m	S
Servant	F	D
Boarder	m	S
Head	m	M
wife	F	M
Son	m	S
Daughter	F	S
Daughter	F	S



# Examples to Feature Vectors

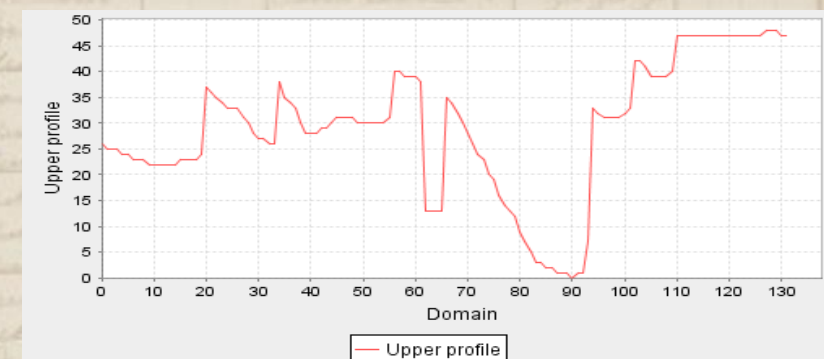
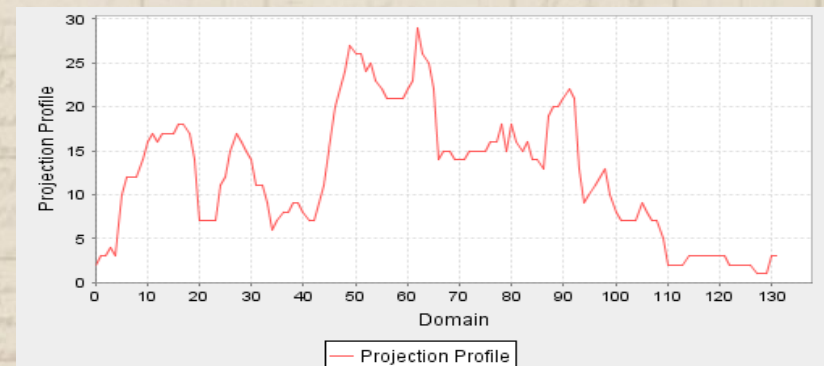
## 25 Numeric Features Extracted:

### o Scalar Features:

- height ( $h$ )
- width ( $w$ )
- aspect ratio ( $w / h$ )
- area ( $w * h$ )

### o Profile Features:

- projection profile
- upper/lower word profile
- 7 lowest scalar values from DFT

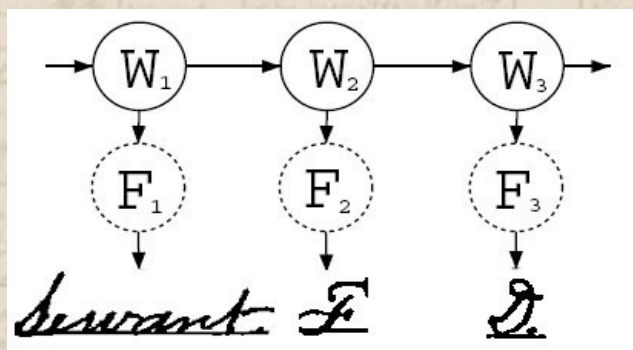




# HMM and Transition Probability Model

- **Probability Model:**

- Hidden Markov Model



$$P(w_1 \dots w_n, I_1 \dots I_n) = \prod_{j=1}^n P(w_j | w_{j-1}) p(f_j | w_j)$$

- State Transition Probabilities

$$P_T(w | w_0) = \left( \frac{\text{number of times } w \text{ occurs in } T}{\text{total number of words in } T} \right)$$

$$P_T(w | v) = \left( \frac{\text{number of times } v, w \text{ occurs in } T}{\text{number of times } v \text{ occurs in } T} \right)$$

# Observation Probability Model

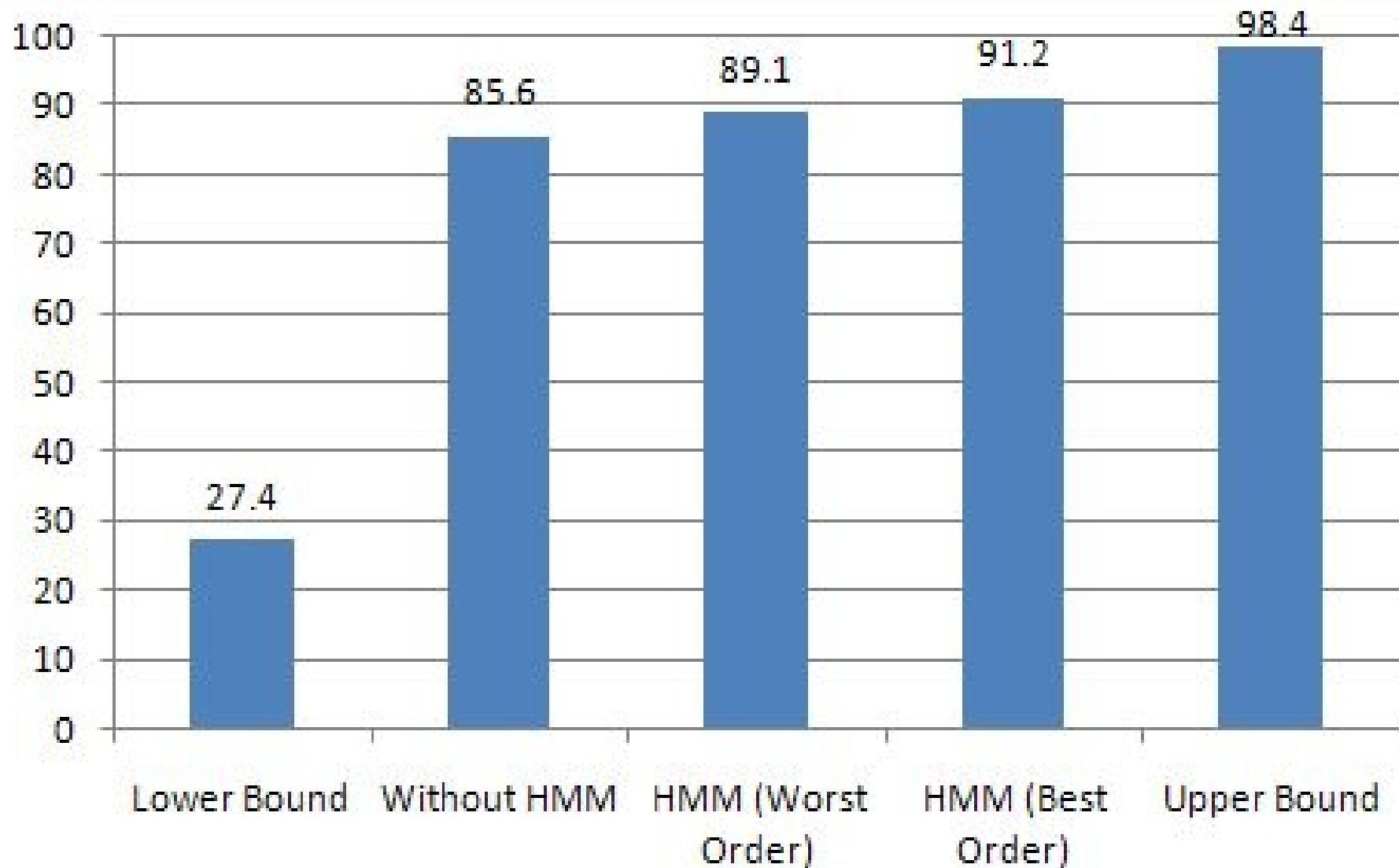
o Multi-variate normal distribution:

$$p(f|w) = \frac{\exp \left\{ (f - \mu_w)^\top \Sigma_w^{-1} (f - \mu_w) \right\}}{\sqrt{2^D \pi^D |\Sigma_w|}}$$

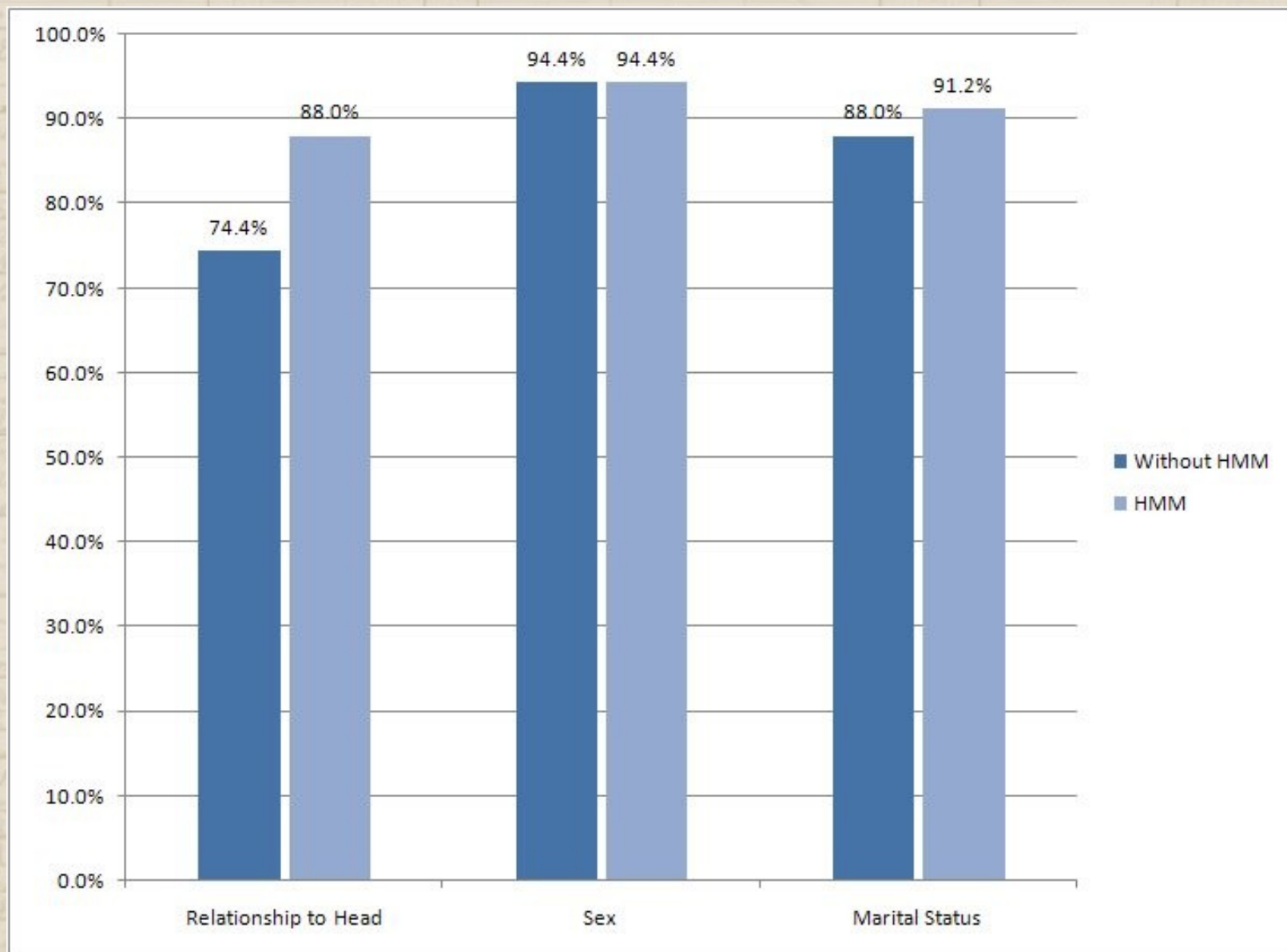
$$p(\mathbf{x}|\omega_i) = (2\pi)^{-d/2} |\mathbf{C}_i|^{-1/2} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^\top \mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)}$$



# Accuracies with and without HMM

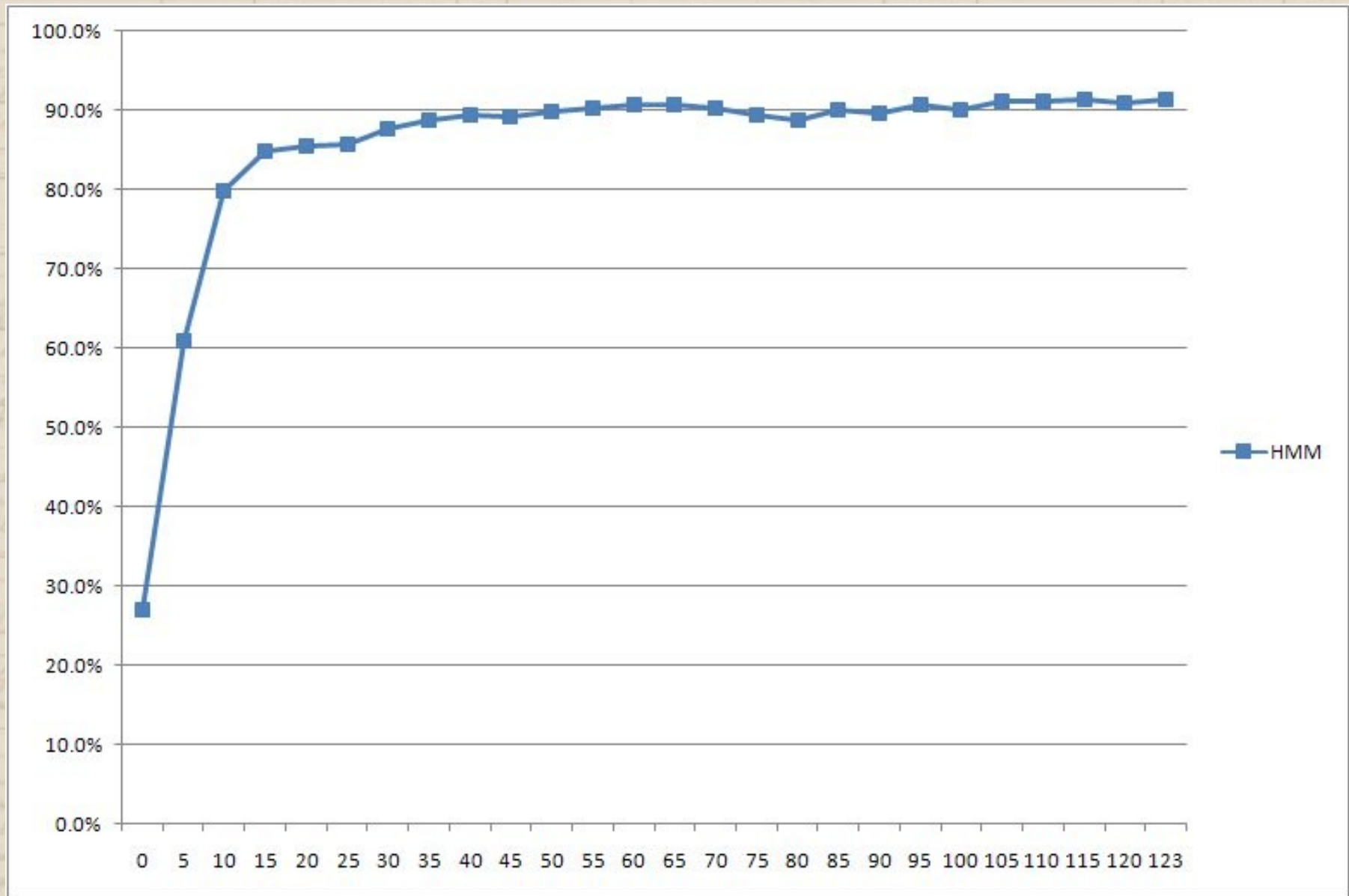


# Accuracies for Separate Columns with and without HMM

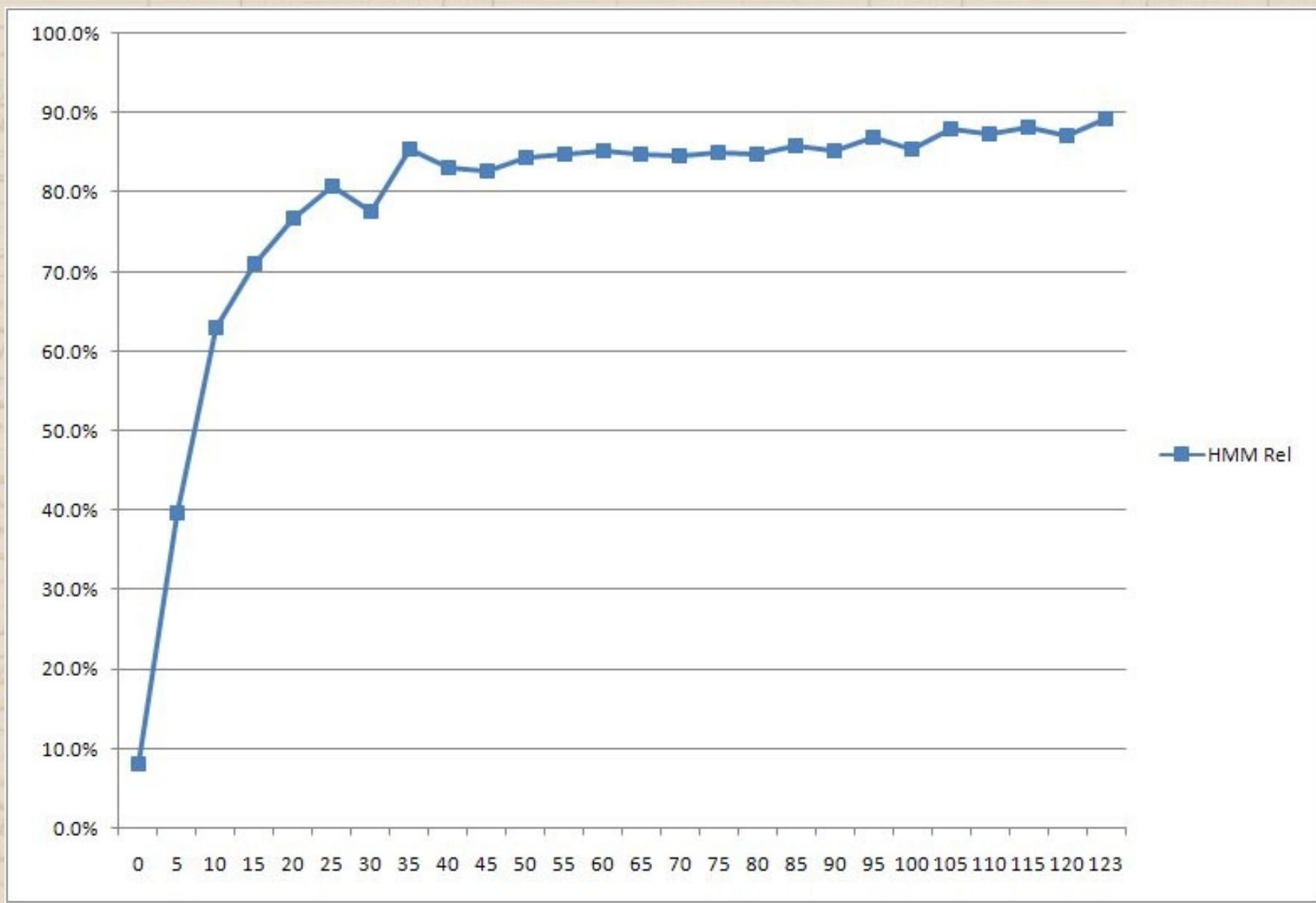




# Accuracies of HMM for Varying Numbers of Training Examples



# Accuracies of “Relationship to Head” for Varying Numbers of Examples





# Conclusions and Future Work

- 10% correction rate for chosen columns after one page.
- Measure indexing time.
- Update models in real-time.
- Columns with larger vocabularies.
- More image preprocessing.
- More visual features.
- More dependencies among words (in different rows).
- More training data.

