

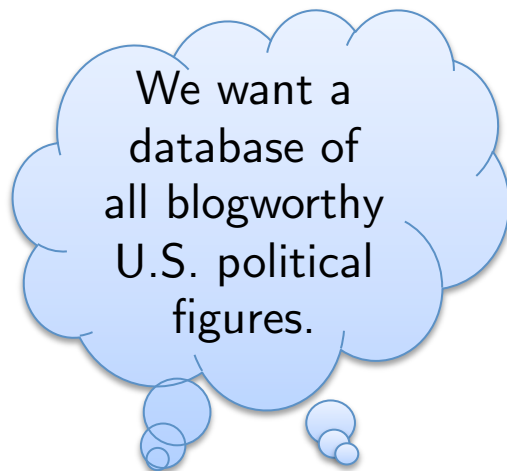
Structured Databases of Named Entities from Bayesian Nonparametrics

Dr.	Jacob		Eisenstein	Machine	Learning		Department	Carnegie	Mellon	University
Ms.	Tae		Yano		Language	Technologies	Institute	Carnegie	Mellon	University
Prof.	William	W.	Cohen	Machine	Learning		Department	Carnegie	Mellon	University
Prof.	Noah	A.	Smith		Language	Technologies	Institute	Carnegie	Mellon	University
Prof.	Eric	P.	Xing	Computer		Science	Department	Carnegie	Mellon	University

In a Nutshell

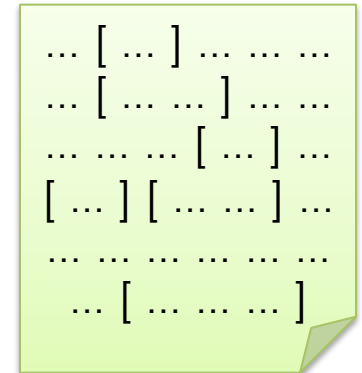
- A joint model over
 - a collection of named entity mentions from text and
 - a structured database table (entities \times name-fields) with data-defined dimensions
- Model aims to solve three problems:
 1. canonicalize the entities
 2. infer a schema for the names
 3. match mentions to entities (i.e., coreference resolution)
- Preliminary experiments on political blog data, only task 1 in this paper.

An Imagined Information Extraction Scenario

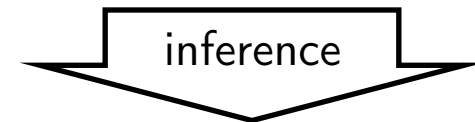


John	McCain	Sen.			Mr.
George	Bush		W.		Mr.
Hillary	Clinton			Rodham	Mrs.
Barack	Obama	Sen.			
Sarah	Palin				

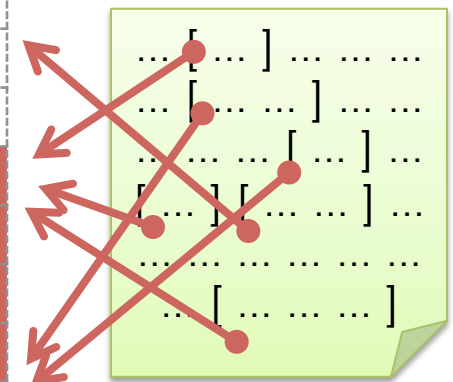
initial table



NER-tagged text: systematic variation in mentions



John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.



Caveat



- Sen. Tom Coburn, M.D. (Rep., Oklahoma), a.k.a. “Dr. No,” does *not* approve of this research.

Prior Work

<i>Research problem</i>	<i>Related papers</i>	<i>Diff</i>
Information extraction	Haghighi and Klein, 2010	Predefined schema (columns/fields).
Name structure models	Charniak, 2001; Elsnar et al., 2009	No resolution to entities.
Record linkage	Felligi and Sunter, 1969; Cohen et al., 2000; Pasula et al., 2002; Bhattacharya and Getoor, 2007	Often on bibliographies (not raw text); predefined schema.
Multi-document coreference resolution	Li et al., 2004; Haghighi and Klein, 2007; Poon and Domingos, 2008; Singh et al., 2011	No canonicalization of entity names.
Morphological paradigm learning	Dreyer and Eisner, 2011	Fixed schema, linguistic analysis problem.

Goal

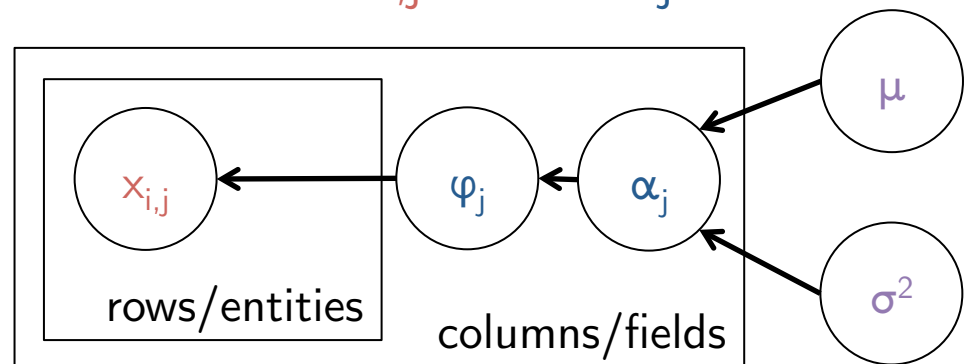
We want a model that solves three problems:

1. canonicalize mentioned entities
2. infer a schema for their names
3. match mentions to entities (i.e., coreference resolution)

Generative Story: Types

First, generate the table.

- Let μ and σ^2 be hyperparameters.
- For each column j :
 - Sample α_j from $\text{LogNormal}(\mu, \sigma^2)$
 - Sample multinomial φ_j from $\text{DP}(G_0, \alpha_j)$, where G_0 is uniform up to a fixed string length.
 - For each row i , draw cell value $x_{i,j}$ from φ_j

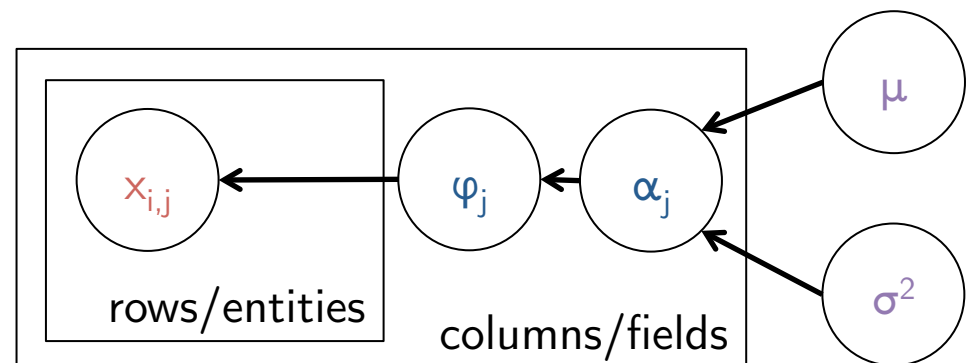


Field-wise Dirichlet Process Priors

very high diversity (high α_j)

very high repetition (low α_j)

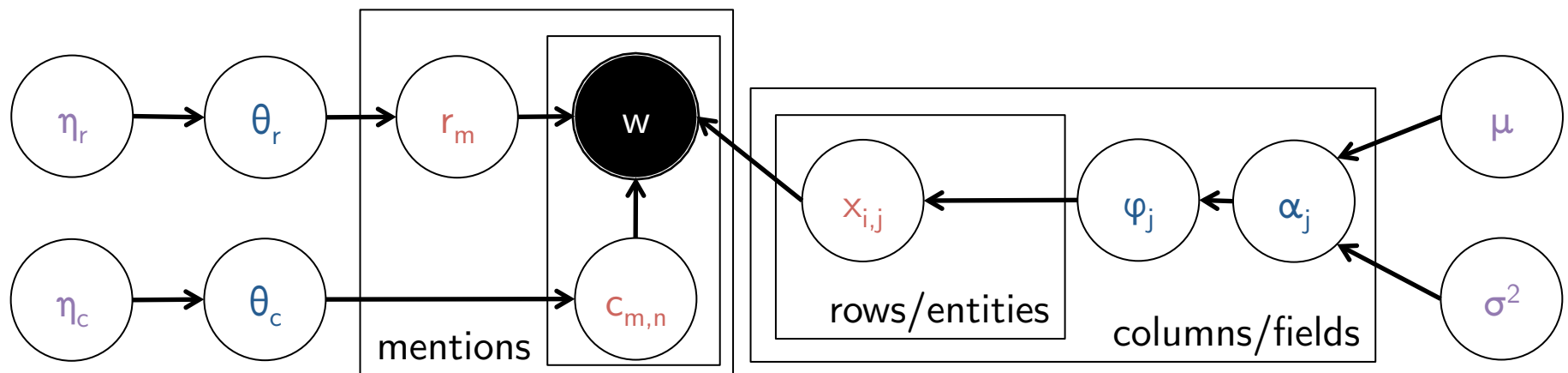
John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.



Generative Story: Tokens

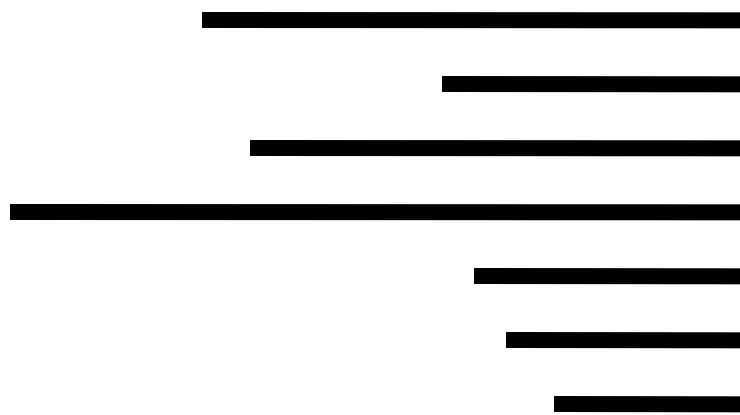
Next, generate the mention tokens.

- Draw the distribution over rows/entities to be mentioned, θ_r , from $\text{Stick}(\eta_r)$.
- Draw the distribution over columns/fields to be used in mentions, θ_c , from $\text{Stick}(\eta_c)$.
- For each mention m , sample its row r_m from θ_r .
 - For each word in the mention, sample its column $c_{m,n}$ from θ_c .
 - Fill in the word to be $x_{r_m, c_{m,n}}$.

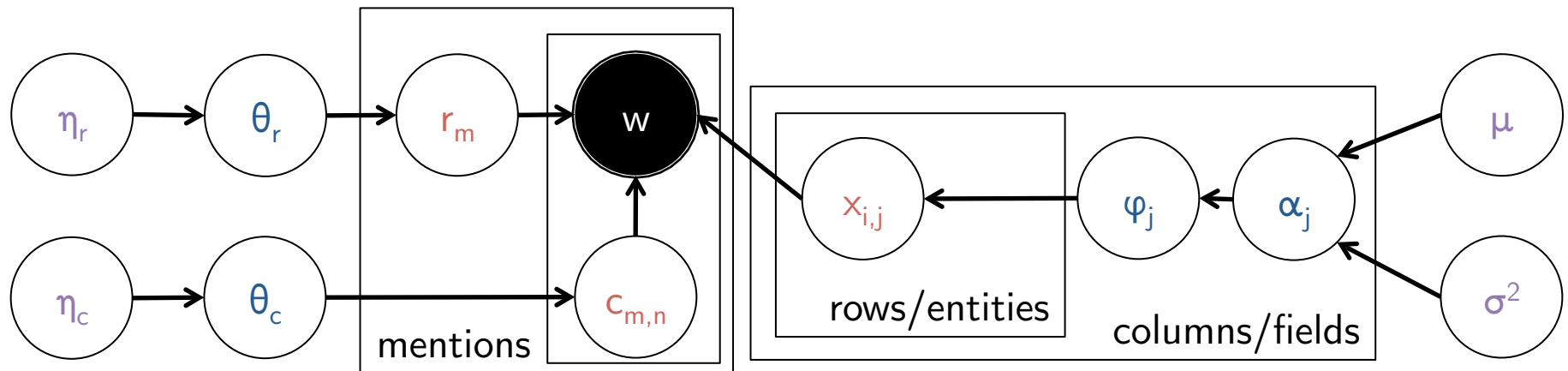


Entity-wise Dirichlet Process Priors

entities receive different amounts of attention (fictitious)

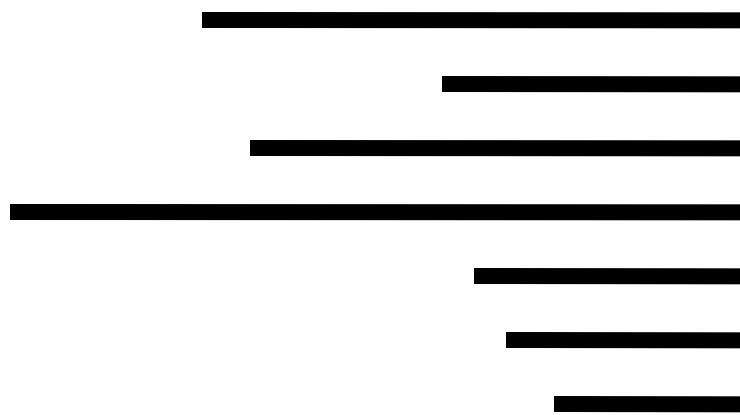


John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.

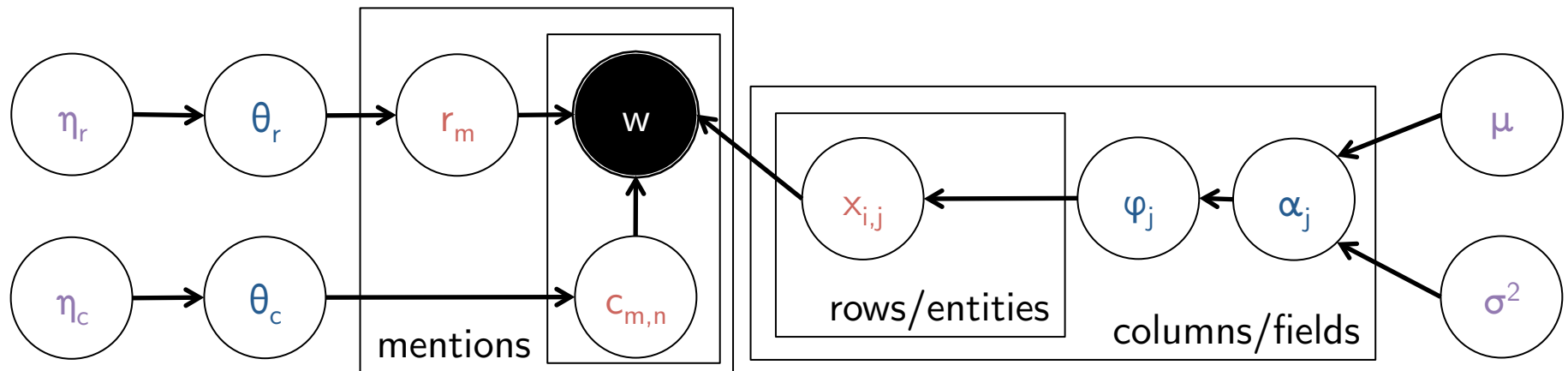


Entity-wise Dirichlet Process Priors

entities receive different amounts of attention (fictitious)

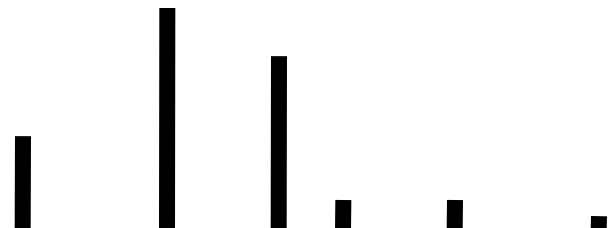


John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.

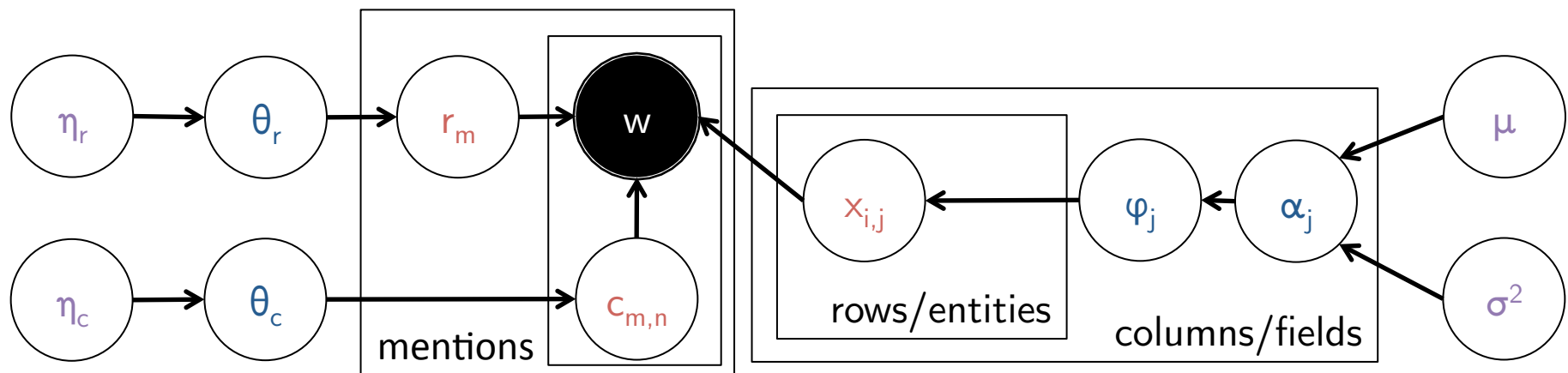


Field-wise Dirichlet Process Priors

fields are used with different frequencies (fictitious)

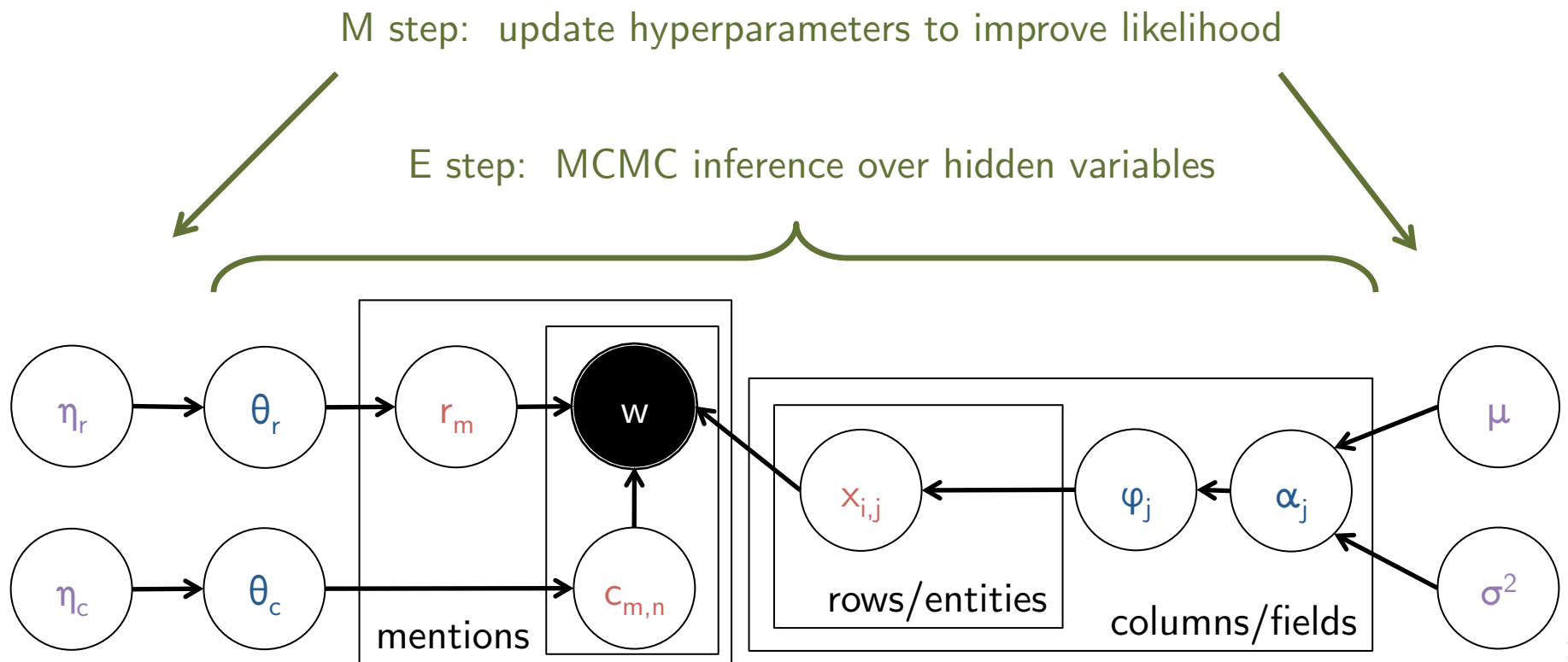


John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.



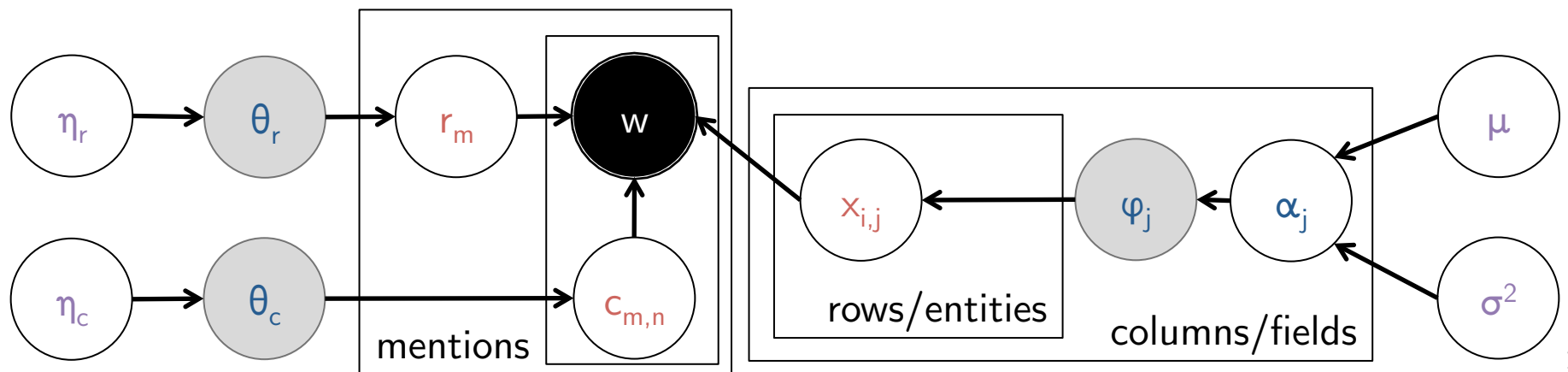
Inference

At a high level, we are doing Monte Carlo EM.



Gibbs Sampling

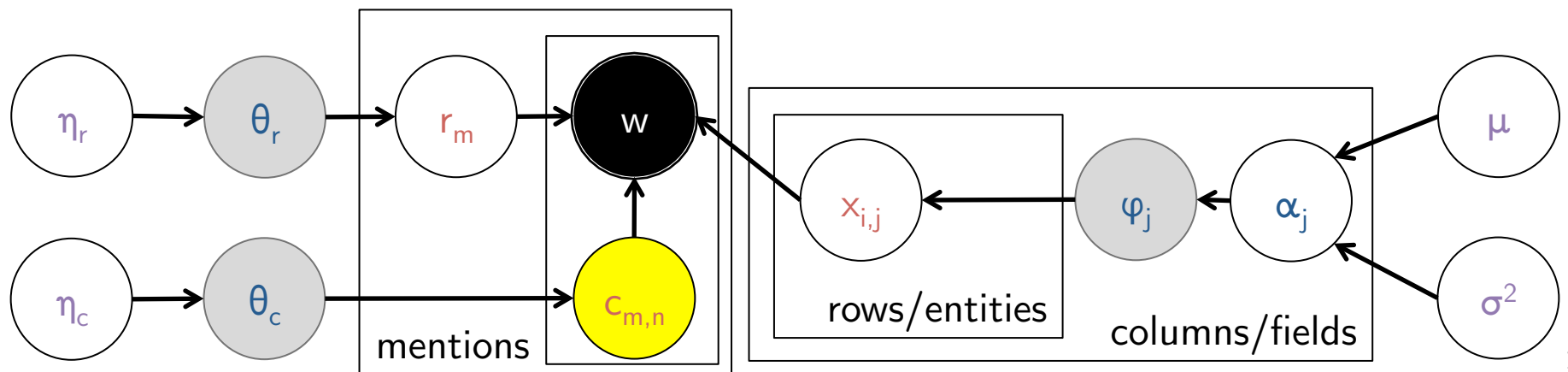
- Collapse out θ_r , θ_c , and φ_j (standard collapsed Gibbs sampler for Dirichlet process).
- Given rows, columns, and words, some of x is determined, and we marginalize the rest.
- I'll describe how we sample columns, rows, and concentrations α_j .



Sampling $c_{m,n}$

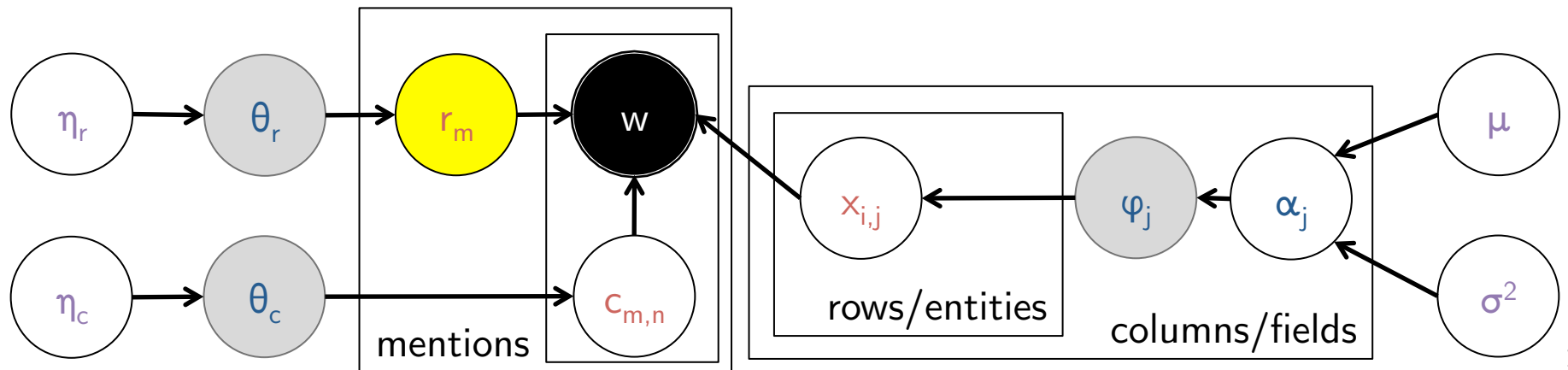
Hinges on $p(w \mid \dots)$ factors:

$$p(c_{m,n} \mid \dots) \propto p(w_{m,n} \mid r_m, c_{m,n}, x_{\text{obs}}, \dots) \times \frac{1}{N(c_{-(m,n)}) + \eta_c} \begin{cases} N(c_{-(m,n)} = j) & \text{if } N(c_{-(m,n)} = j) > 0 \\ \eta_c & \text{otherwise} \end{cases}$$



Sampling r_m

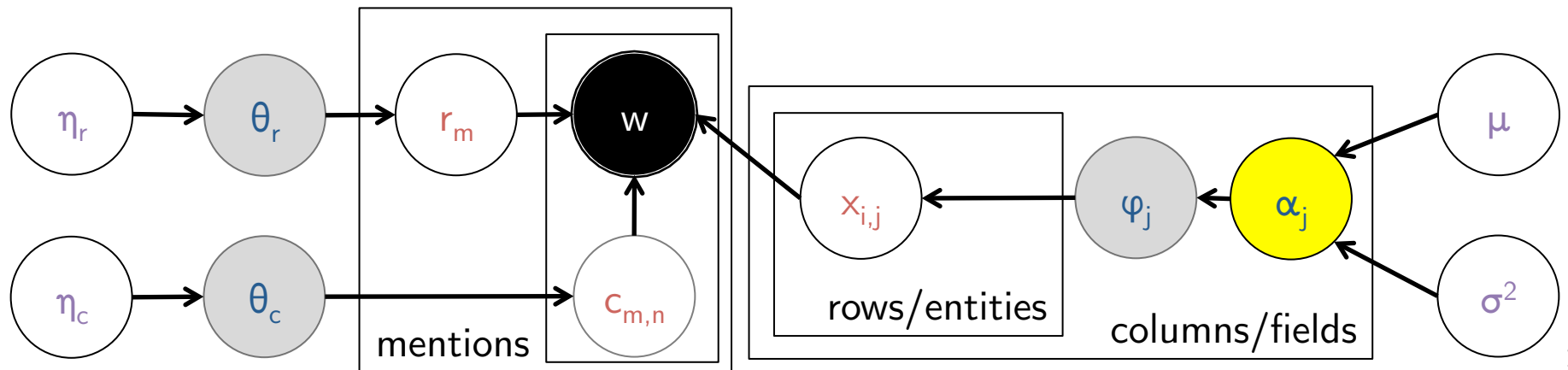
- Need to multiply together $p(w \mid \dots)$ quantities (see paper) for all words in the mention.
- We speed things up by marginalizing out $c_{m,*}$.
- This calculation exploits conditional independence of tokens given the row.



Sampling α_j

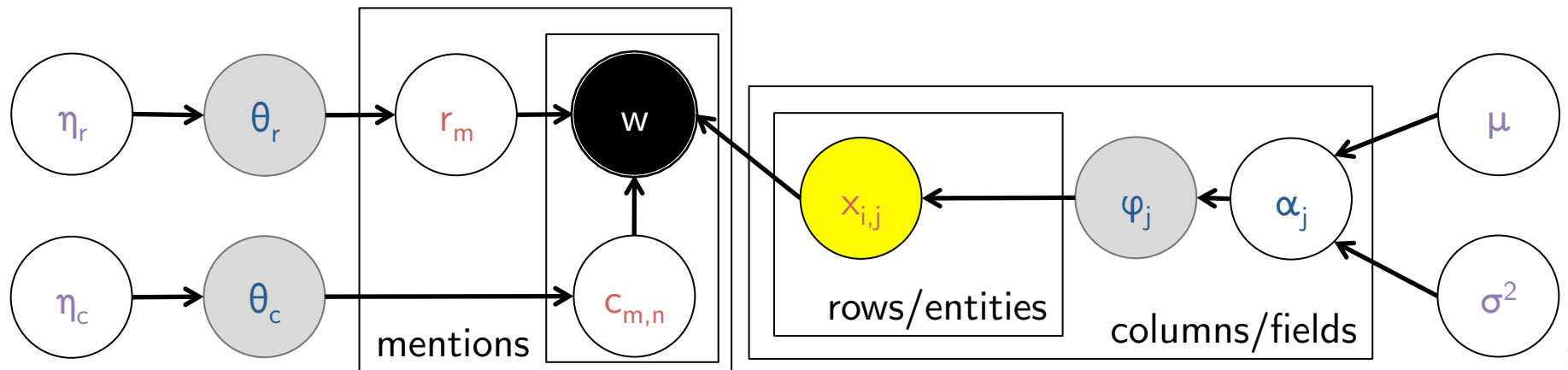
- Given number of specified entries in $x_{*,j}$ (n_j) and number of unique entries in $x_{*,j}$ (k_j):

$$p(\alpha_j \mid \dots) \propto \frac{\exp(-(\log \alpha_j - \mu)^2) \alpha_j^{k_j} \Gamma(\alpha_j)}{2\sigma^2 \Gamma(n_j + \alpha_j)}$$



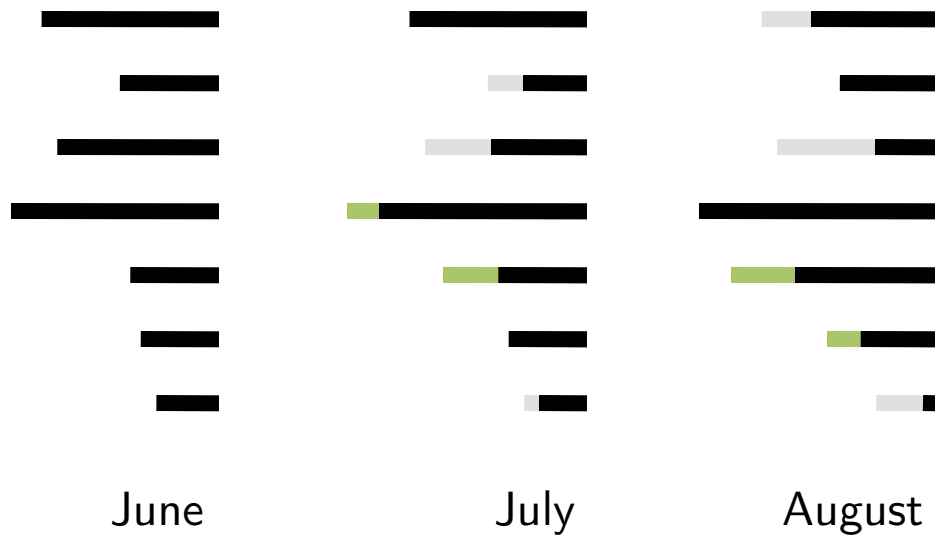
Column Swaps

- One additional move: in a single row, swap entries in two columns of \mathbf{x} .
- The swap also implies changing some \mathbf{c} variables.
- See the paper for details on this Metropolis-Hastings step.



Temporal Dynamics

entities receive different amounts of attention **at different times**



John	McCain	Sen.			Mr.
George	Bush	Pres.	W.		Mr.
Hillary	Clinton	Sen.		Rodham	Mrs.
Barack	Obama	Sen.	H.		Mr.
Sarah	Palin	Gov.			Mrs.
Joe	Biden	Sen.			Mr.
Ron	Paul	Rep.			Mr.

Recurrent Chinese Restaurant Process (Ahmed and Xing, 2008)

- Data are divided into discrete epochs.
- Row Dirichlet process includes pseudocounts from previous epoch.
- Entities come and go; reappearing after disappearance is vanishingly improbable.

In Chinese restaurant view:

$$p(r_m^{(t)} = i \mid r_{1,\dots,m-1}^{(t)}, r^{(t-1)}, \eta_r) \propto \begin{cases} \eta_r & \text{if positive} \\ N(r_{1,\dots,m-1}^{(t)} = i) + N(r^{(t-1)} = i) & \text{otherwise} \end{cases}$$

This affects updates to η_r and sampling of r .

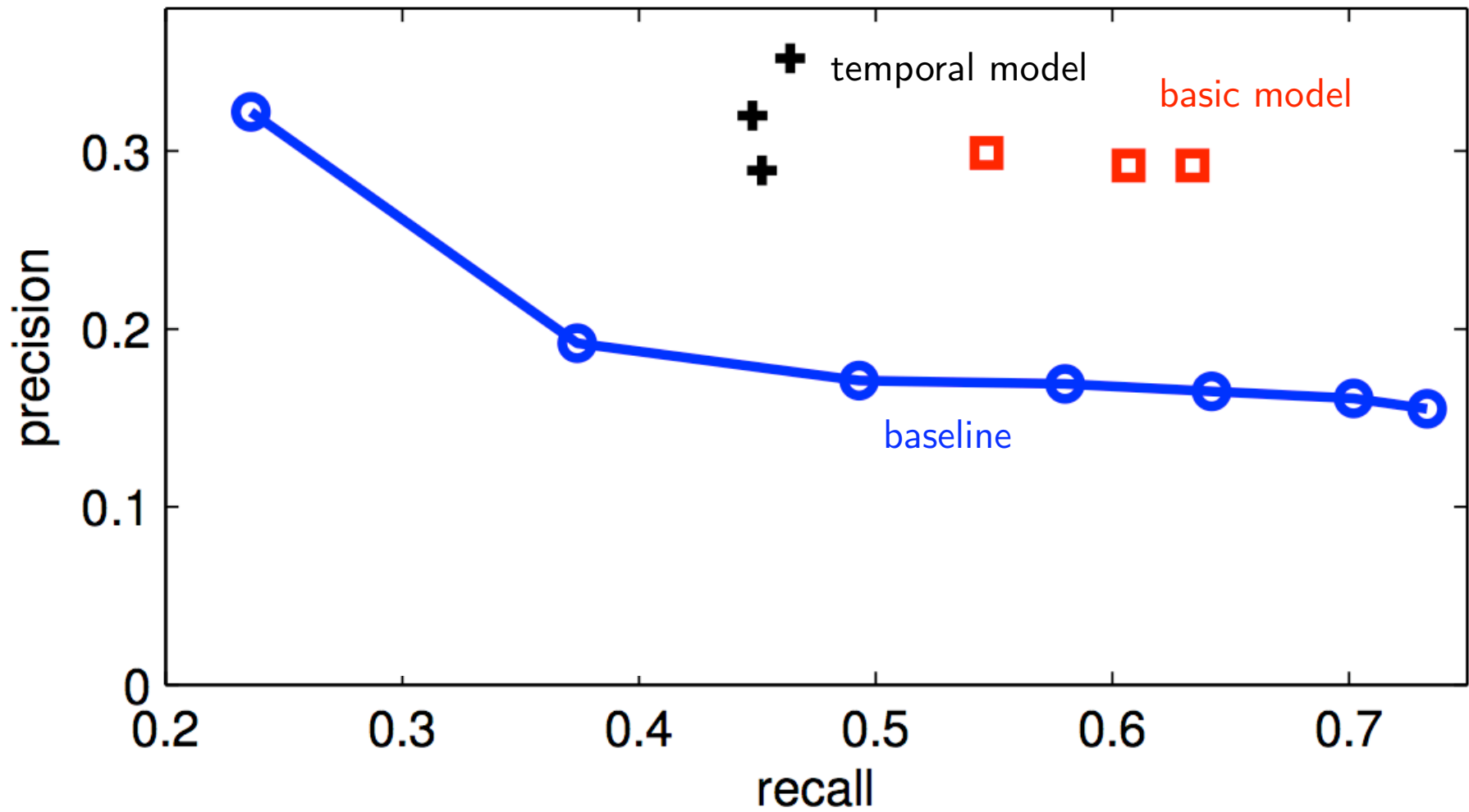
Data for Evaluation

- Data: blogs on U.S. politics from 2008 (Eisenstein and Xing, 2008)
 - Stanford NER → 25,000 mentions
 - Eliminate those with frequency less than 4 and more than 7 tokens
 - 19,247 mentions (45,466 tokens), 813 unique
- Annotation: 100 reference entities
 - Constructed by merging sets of most frequent mentions, discarding errors
 - Example: { Barack, Obama, Mr., Sen. }

Evaluation

- Bipartite matching between reference entities and rows of **x**.
- Measure precision and recall.
 - Precision is very harsh (only 100 entities in reference set, and finding anything else incurs a penalty!) – same problem is present in earlier work.
- Baseline: agglomerative clustering based on string edit distance (Elmacioglu et al., 2007); different stopping points define a P-R curve.
 - No database!

Results



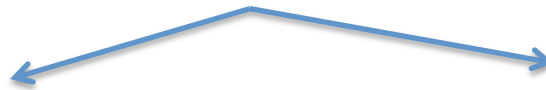
Examples



Bill	Clinton	Benazir		Bhutto
Nancy	Pelosi	Speaker		
John	Kerry	Sen.		Roberts
Martin	King	Dr.	Jr.	Luther
Bill	Nelson			

☺ Bill Clinton is not Bill Nelson

Examples



Bill	Clinton	Benazir		Bhutto
Nancy	Pelosi	Speaker		
John	Kerry	Sen.		Roberts
Martin	King	Dr.	Jr.	Luther
Bill	Nelson			

- ☺ Bill Clinton is not Bill Nelson
- ☹ Bill Clinton *is* Benazir Bhutto
- ☹ John Kerry is John Roberts
 - Hard to create a new row once we're "stuck"
 - Common names are garbage collectors

Examples



Bill	Clinton	Benazir		Bhutto
Nancy	Pelosi	Speaker		
John	Kerry	Sen.		Roberts
Martin	King	Dr.	Jr.	Luther
Bill	Nelson			

- ☺ Bill Clinton is not Bill Nelson
- ☹ Bill Clinton *is* Benazir Bhutto
- ☹ John Kerry is John Roberts
- ☺ Rare “Speaker” title for Pelosi; fields generally good

Future Extensions

- Structured model over name structure
- Optionality within a cell?
- Changes in the database over time
- Joint inference with named entity recognition
- “Topics” (some entities are likely to coocur)
- Lexical context of mentions to aid disambiguation
- Burstiness within a document
- Events (cf., Chambers and Jurafsky, 2011)
- Information used in coreference resolution: linguistic cues (Bengtson and Roth, 2008) and external knowledge (Haghighi and Klein, 2010)

Conclusions

- A joint model over
 - a collection of named entity mentions from text and
 - a structured database table (entities \times name-fields) with data-defined dimensions
- Model aims to solve three problems:
 1. canonicalize the entities
 2. infer a schema for the names
 3. match mentions to entities (i.e., coreference resolution)

Thanks!