# Grounding Topic Models with Knowledge Bases


Zhiting Hu[1*], Gang Luo[2], Mrinmaya Sachan[1] , Eric Xing[1], Zaiqing Nie[3]

[1]Carnegie Mellon University

[2]Microsoft, California, US

[3]Microsoft Research, Beijing, China

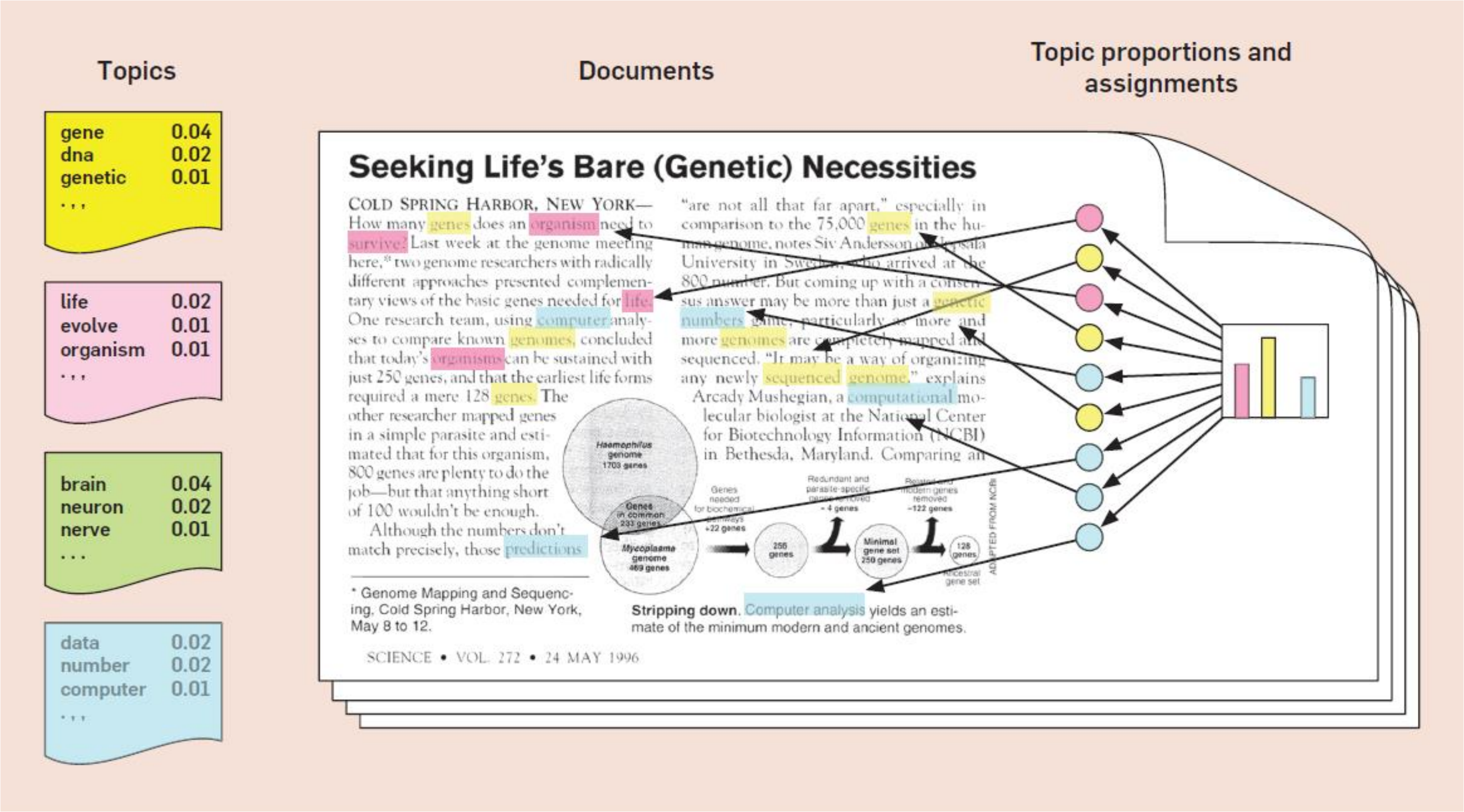

*This work was done when the first two authors were at Microsoft Research, Beijing

# Topic Modeling

- Represents latent topics as probability distributions over words

# Topic Modeling

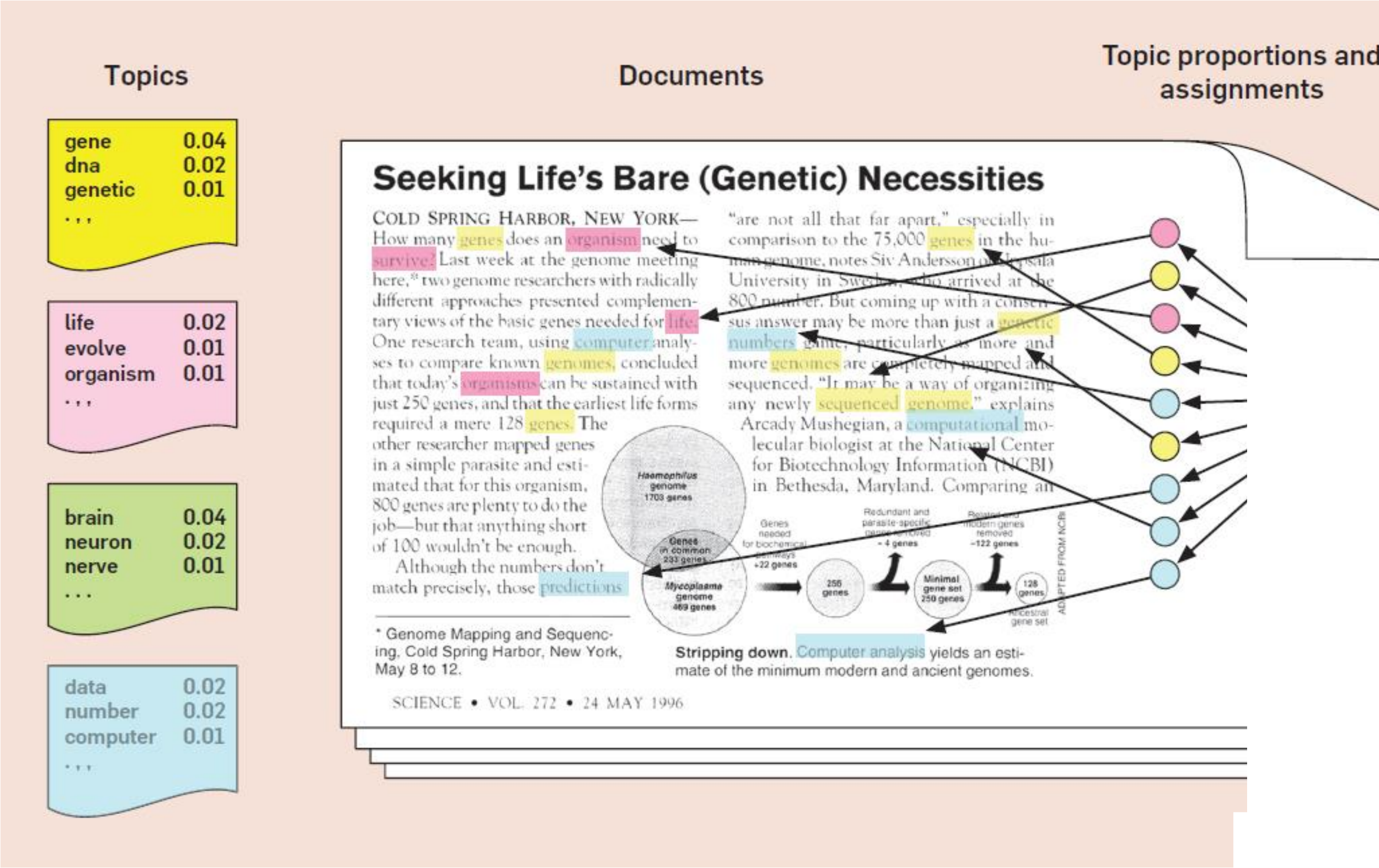
- Represents latent topics as probability distributions over words




LDA (latent Dirichlet process)

# Topic Modeling

- Represents latent topics as probability distributions over words




LDA (latent Dirichlet process)

| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

[Blei et al., 2003]

# Topic Modeling

- Represents latent topics as probability distributions over words
  - hard to interpret due to incoherence
  - lack of background context
  - no grounded semantics

| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

[Blei et al., 2003]

# Topic Modeling

- Represents latent topics as probability distributions over words
  - hard to interpret due to incoherence
  - lack of background context
  - no grounded semantics
- Previous work combines external knowledge
  - improves coherence, but topics = word distributions
  - imposes one-to-one binding of topics to pre-defined knowledge base (KB) entities
    - Sacrifices flexibility

| "Arts" | "Budgets" | "Children" | "Education" |
|---|---|---|---|
| NEW | MILLION | CHILDREN | SCHOOL |
| FILM | TAX | WOMEN | STUDENTS |
| SHOW | PROGRAM | PEOPLE | SCHOOLS |
| MUSIC | BUDGET | CHILD | EDUCATION |
| MOVIE | BILLION | YEARS | TEACHERS |
| PLAY | FEDERAL | FAMILIES | HIGH |
| MUSICAL | YEAR | WORK | PUBLIC |
| BEST | SPENDING | PARENTS | TEACHER |
| ACTOR | NEW | SAYS | BENNETT |
| FIRST | STATE | FAMILY | MANIGAT |
| YORK | PLAN | WELFARE | NAMPHY |
| OPERA | MONEY | MEN | STATE |
| THEATER | PROGRAMS | PERCENT | PRESIDENT |
| ACTRESS | GOVERNMENT | CARE | ELEMENTARY |
| LOVE | CONGRESS | LIFE | HAITI |

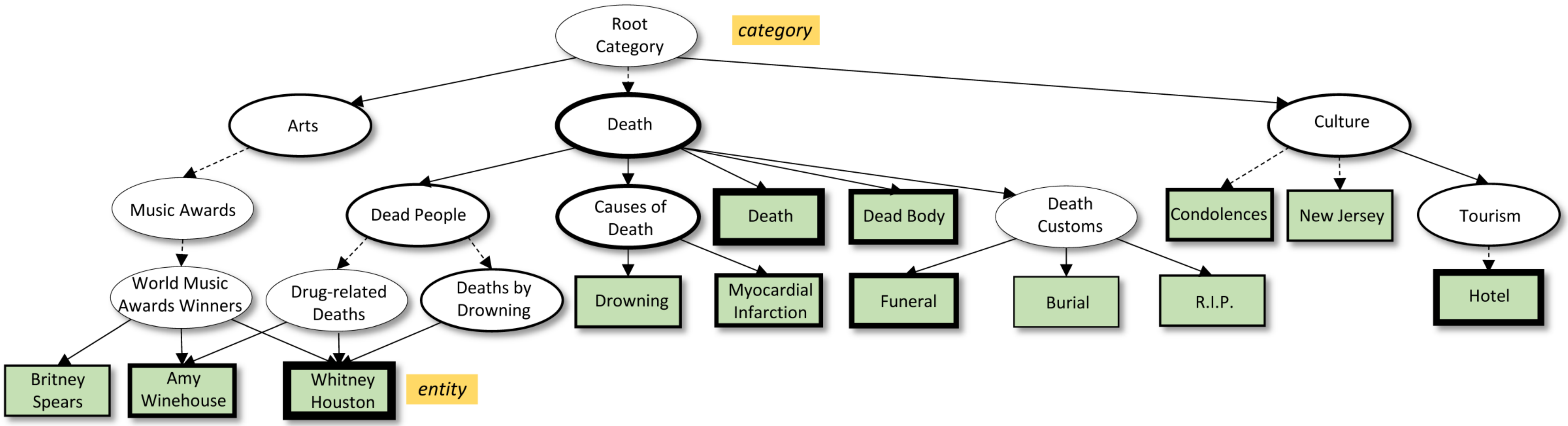
[Blei et al., 2003]

# This work

- A structured topic representation based on *entity taxonomy* from KBs

# This work

- A structured topic representation based on *entity taxonomy* from KBs




Topic ``Death of Whitney Houston''

# This work

- A structured topic representation based on *entity taxonomy* from KBs
  - grounded semantics
  - improved coherenceness: captures entity correlations encoded in the taxonomy

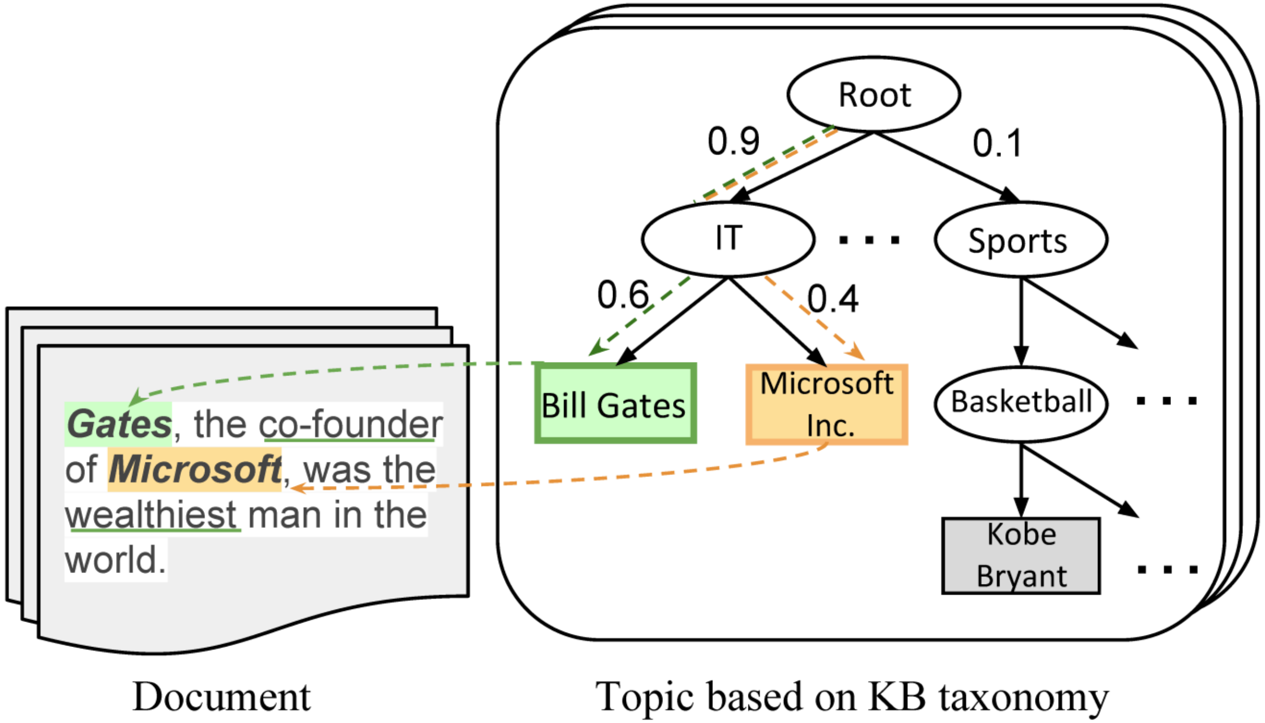
# This work

- A structured topic representation based on *entity taxonomy* from KBs
  - grounded semantics
  - improved coherenceness: captures entity correlations encoded in the taxonomy
- A probabilistic model to infer both hidden *topics* and *entities* from text corpora

# Document Modeling

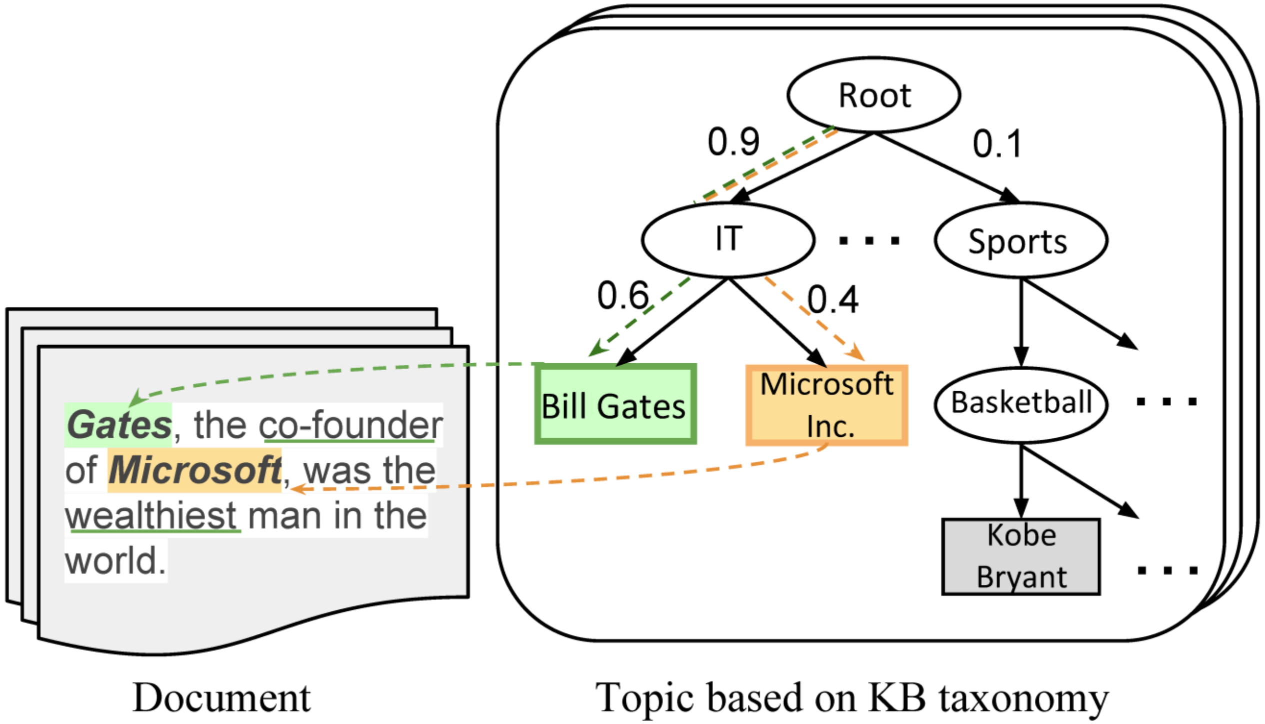
- Augments bag-of-word documents with *entity mentions*
  - mentions carry salient semantics of a document

- *{co-founder, wealthiest, man, …}*
- *{Gates, Microsoft, …}*



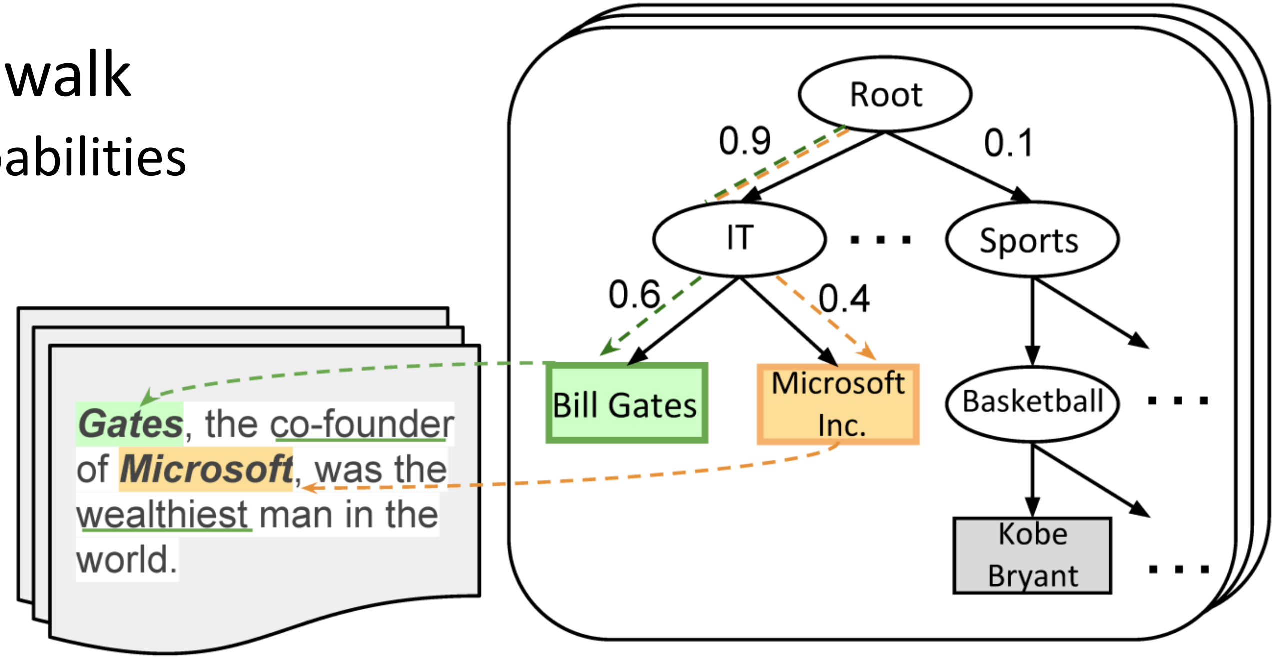
# Document Modeling

- Generative process:
  - each mention <- an entity and a topic
  - each word <- an index indicating which mention to describe



# Topic: Random Walk on Taxonomy

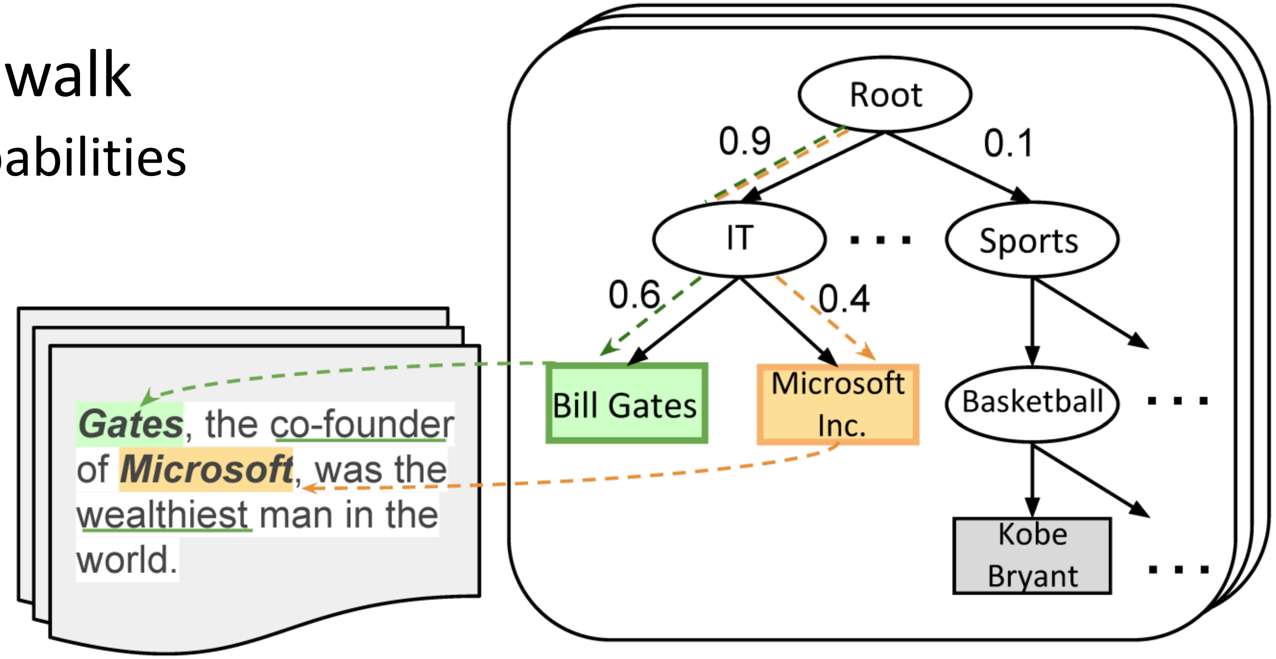
- Entity taxonomy
  - leaf: entity
  - internal nodes: category
- Each topic as a root-to-leaf random walk
  - a set of parent-to-child transition probabilities
  - -> entity/category weights




Document | Topic based on KB taxonomy

# Topic: Random Walk on Taxonomy

- Entity taxonomy
  - leaf: entity
  - internal nodes: category
- Each topic as a root-to-leaf random walk
  - a set of parent-to-child transition probabilities
  - -> entity/category weights
- Path-sharing:
  - encourages clustering correlated entities into the same topic

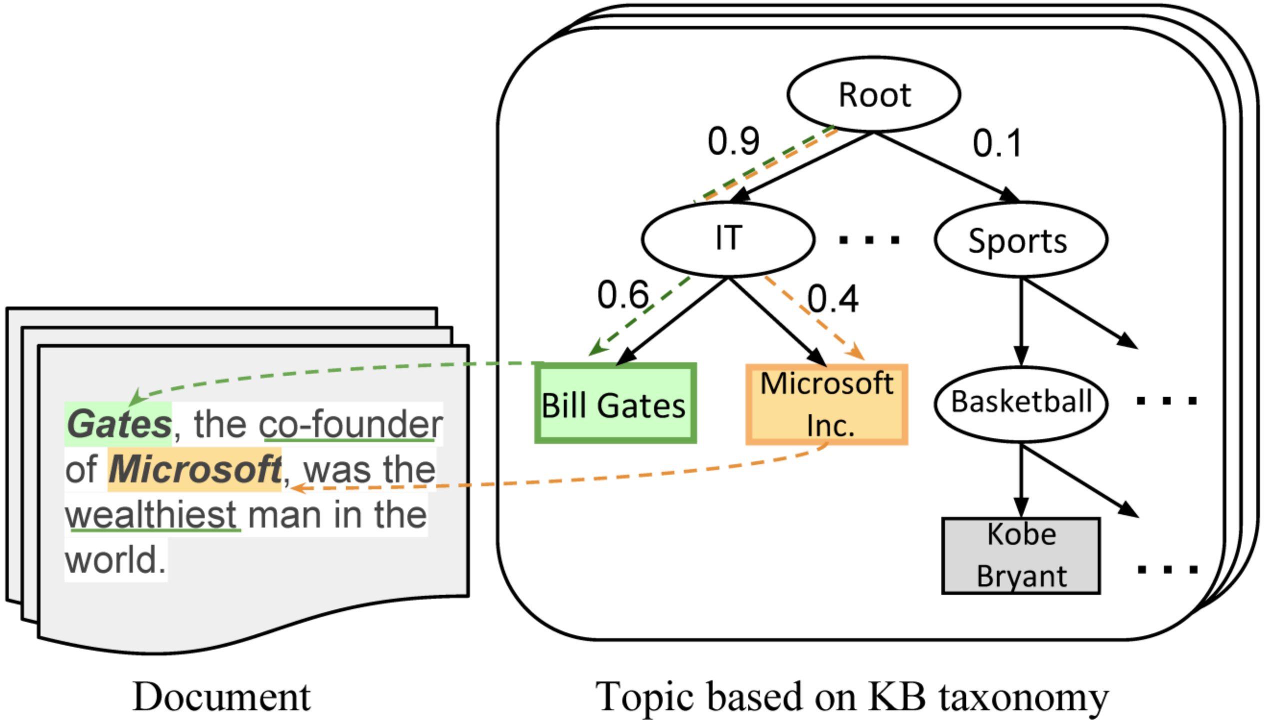



Document

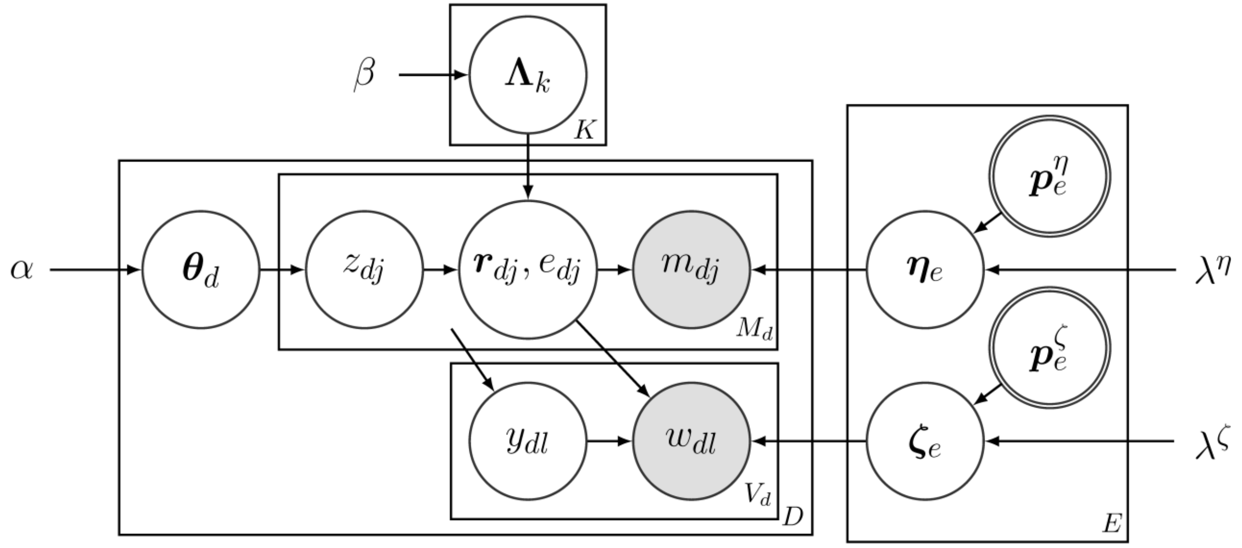
Topic based on KB taxonomy

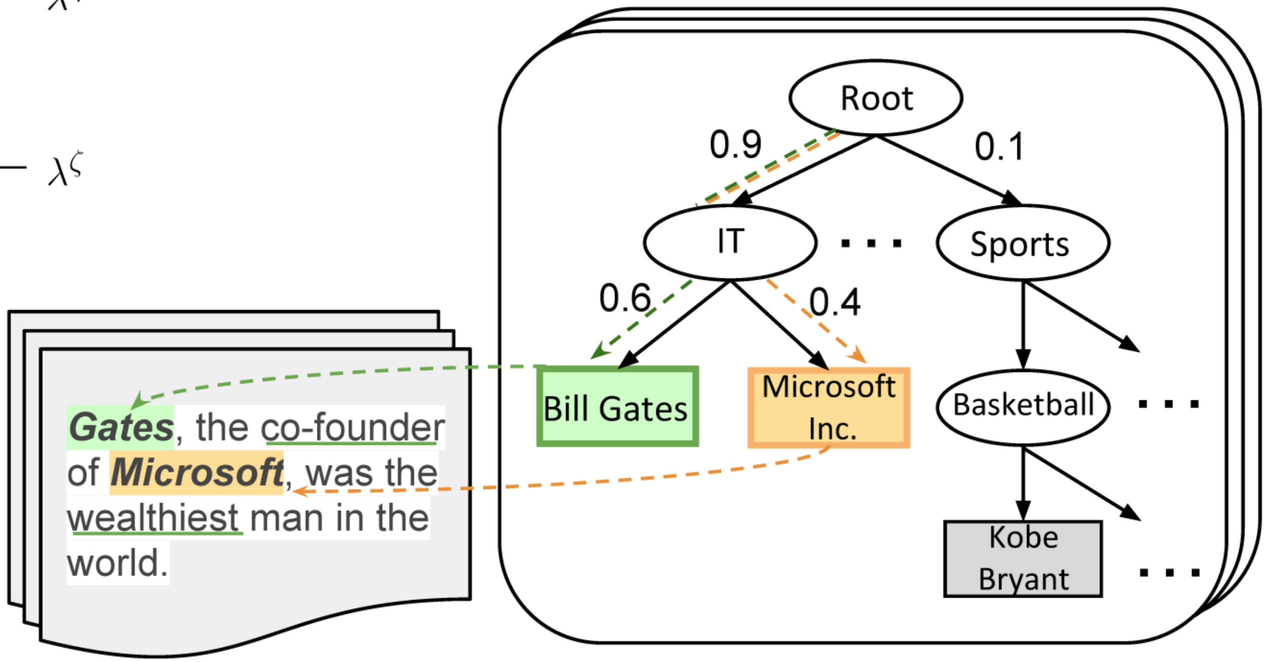
# Entity Modeling

- A distribution over mentions
  - captures relatedness between the entity and mentions
  - *Microsoft Inc.* – MS, Gates
- A distribution over words
  - characterizes the entity attributes
  - *Bill Gates* - wealthiest



# Graphical Model Representation

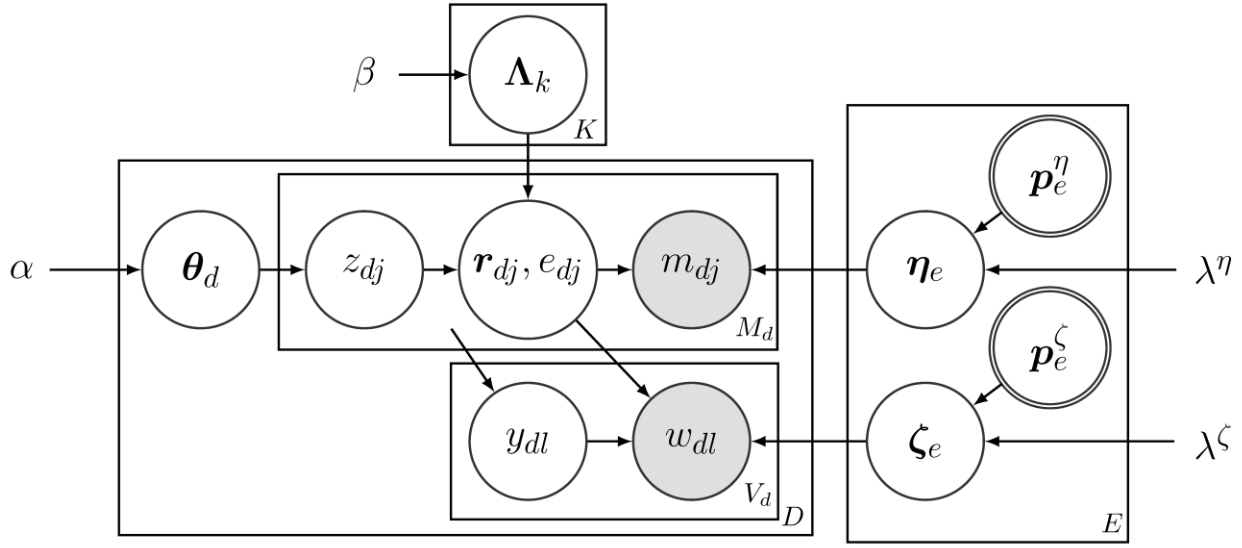


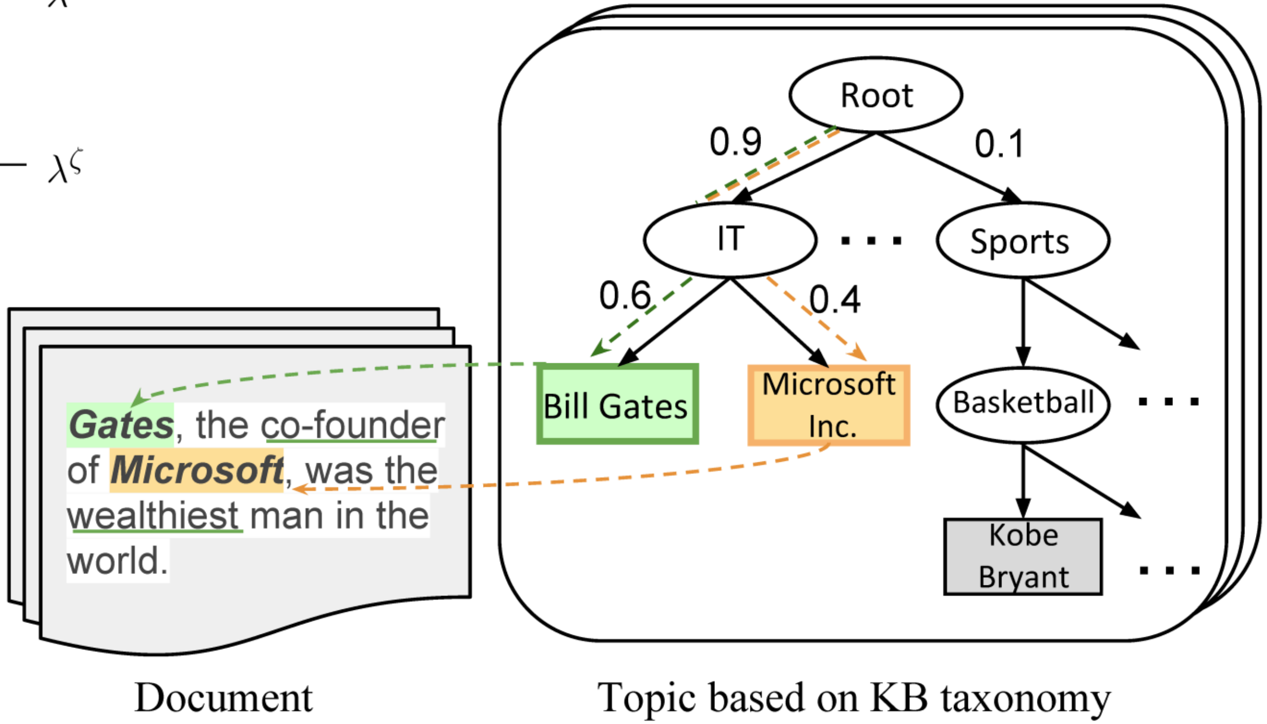

Document

Topic based on KB taxonomy

# Graphical Model Representation

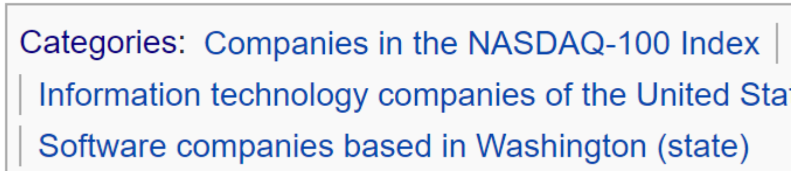



Latent Grounded Semantic Analysis (LGSA)



# Experiments

https://en.wikipedia.org/wiki/Microsoft

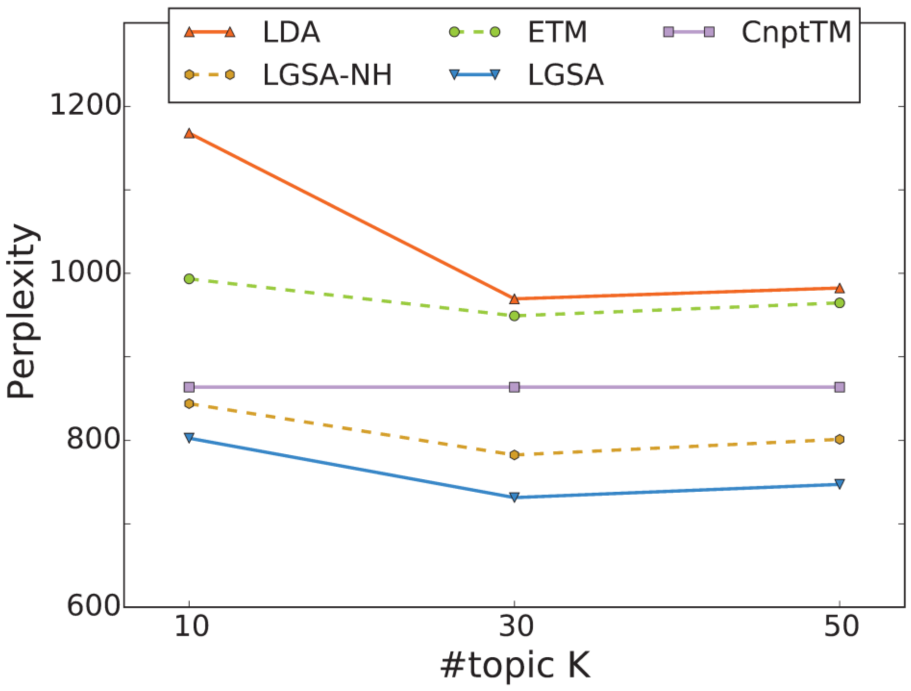




- Knowledge Base: Wikipedia
  - Entity Wikipedia pages
  - Entity category hierarchy
- Datasets
  - TMZ (tmz.com): celebrity gossip news
    - celebrity labels
    - #doc ~= 30K
  - New York Times news (LDC)
    - #doc ~= 330K
- Baselines

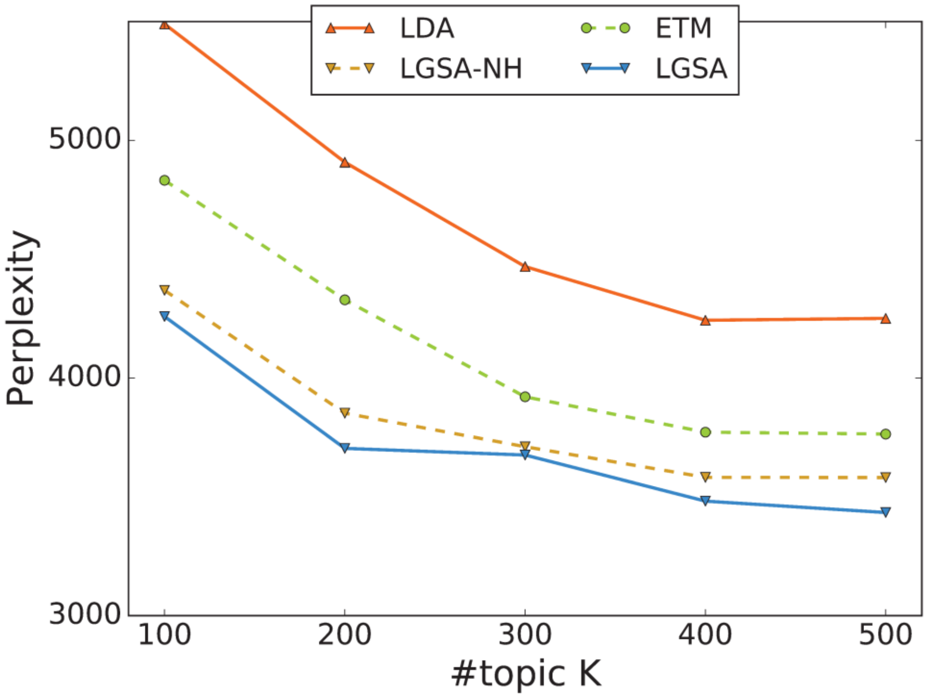
| Method | Features | | | Tasks | |
|---|---|---|---|---|---|
| | word | mention | structured knowledge | topic extraction | key entity identification |
| CnptTM | √ | | √ | √ | √ |
| ETM | √ | √ | | √ | |
| LDA | √ | | | √ | √ |
| ESA | √ | | √ | | √ |
| MA-C | √ | √ | √ | | √ |
| LGSA-NH | √ | √ | | √ | √ |
| LGSA | √ | √ | √ | √ | √ |

Table 3: Feature and task comparison of different methods

# Topic Perplexity




On the TMZ dataset



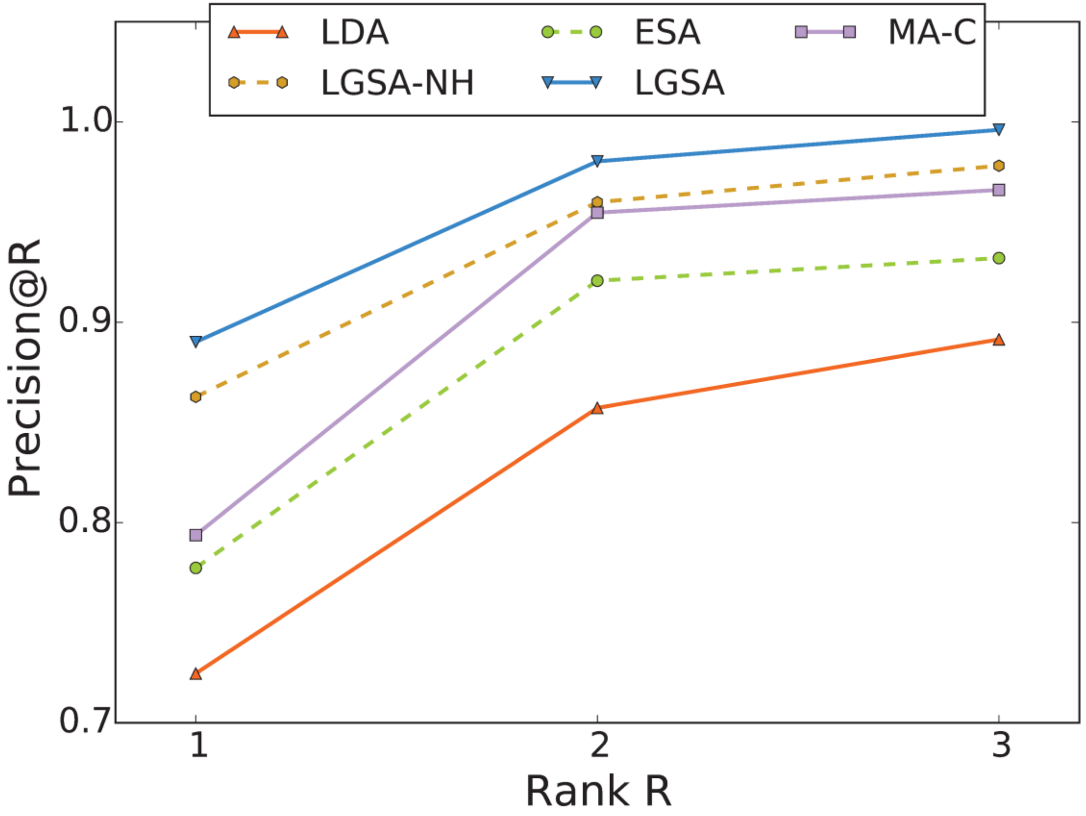

On the NYT dataset

# Key Entity Identification

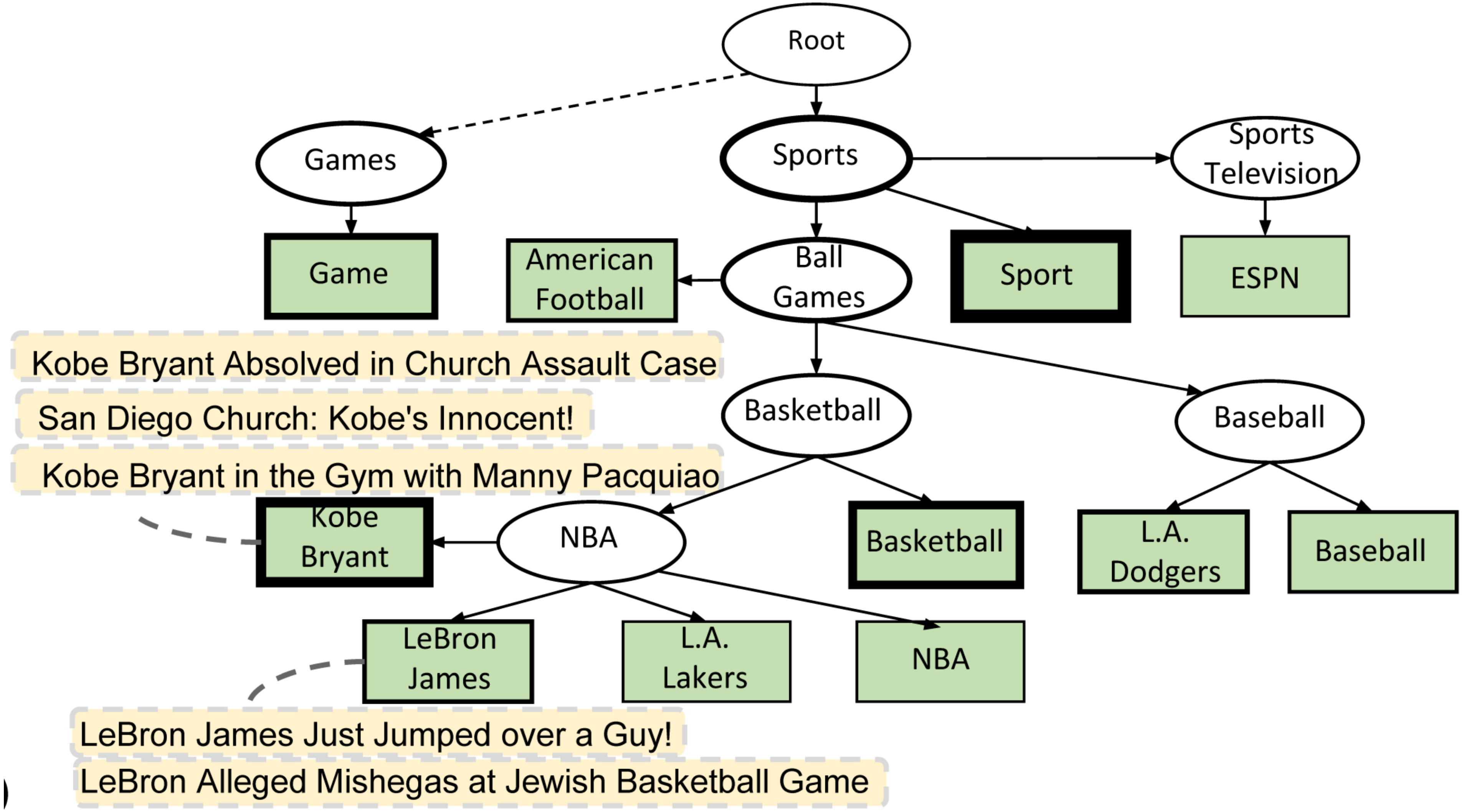
- Key entity of a document
  - E.g., the persons a news article is mainly about
- TMZ dataset: ground truth (celebrity label) available
- LGSA: $\theta'_d$ - distribution over entities for document $d$

# Key Entity Identification

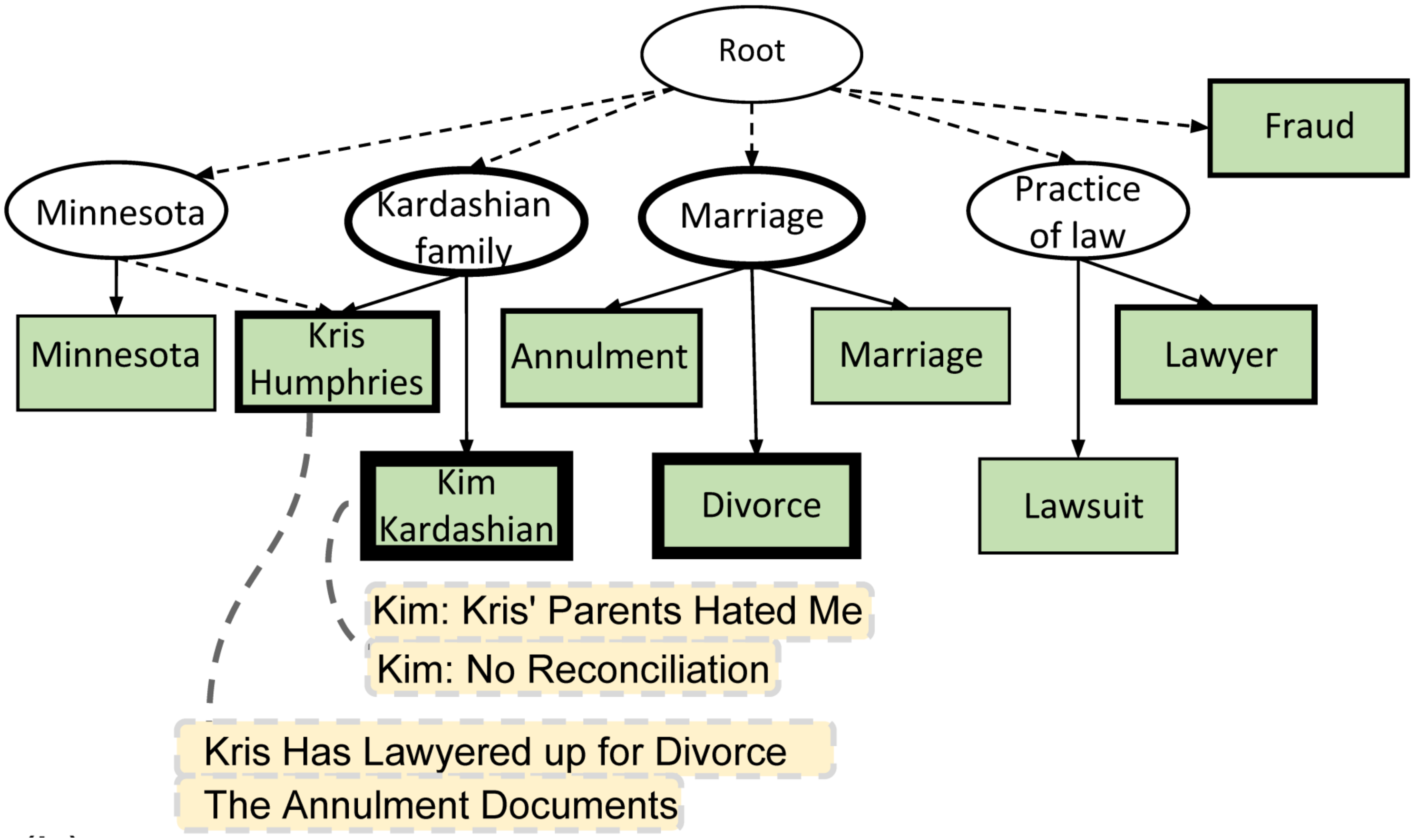
- Key entity of a document
  - E.g., the persons a news article is mainly about
- TMZ dataset: ground truth
- LGSA: $\theta'_d$ - distribution over



# Example Topics: Sports



# Example Topics: Kardashian and Humphries' Divorce



# Conclusion

- Traditional word-based topic representation lacks interpretability and grounded semantics
- A structured topic representation based on entity taxonomy from KBs
- A probabilistic model (LGSA) to infer latent grounded topics
- Improved performance on topic perplexity and key entity identification

# Thanks..