Incorporating World Knowledge to Document Clustering via Heterogeneous Information Networks

KDD'15 Sydney, Australia

Chenguang Wang, Yangqiu Song, Ahmed El-Kishky, Dan Roth, Ming Zhang, Jiawei Han

Text Categorization

- A classical machine learning problem that impacts many applications!
  - Social network analysis, health care, machine reading …
- Traditional approach:

Text Categorization

- A classical machine learning problem that impacts many applications!
  - Social network analysis, health care, machine reading …
- Traditional approach:

Label data → Train a classifier → Make prediction

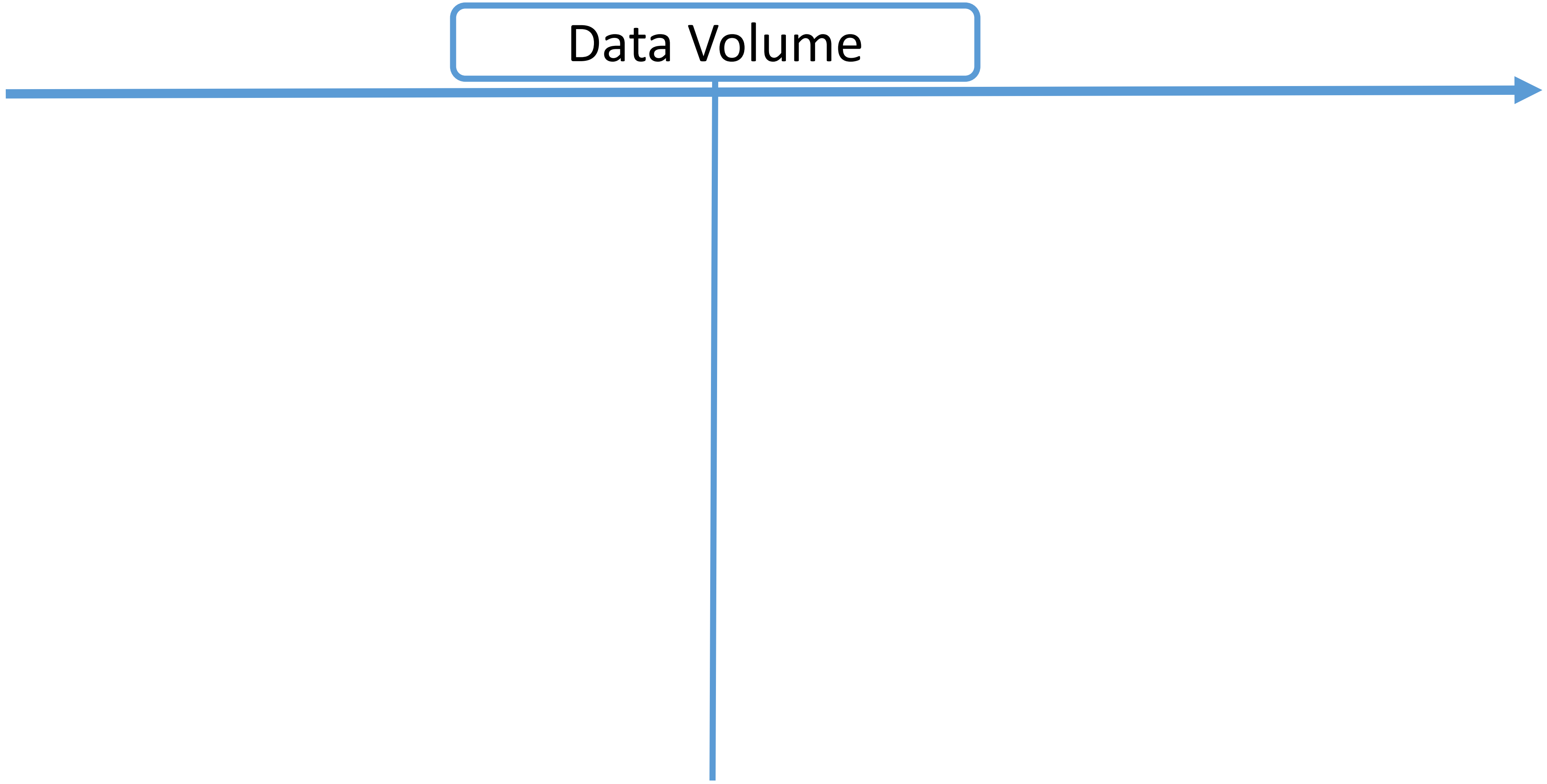
Text Categorization

- A classical machine learning problem that impacts many applications!
  - Social network analysis, health care, machine reading …
- Traditional approach:

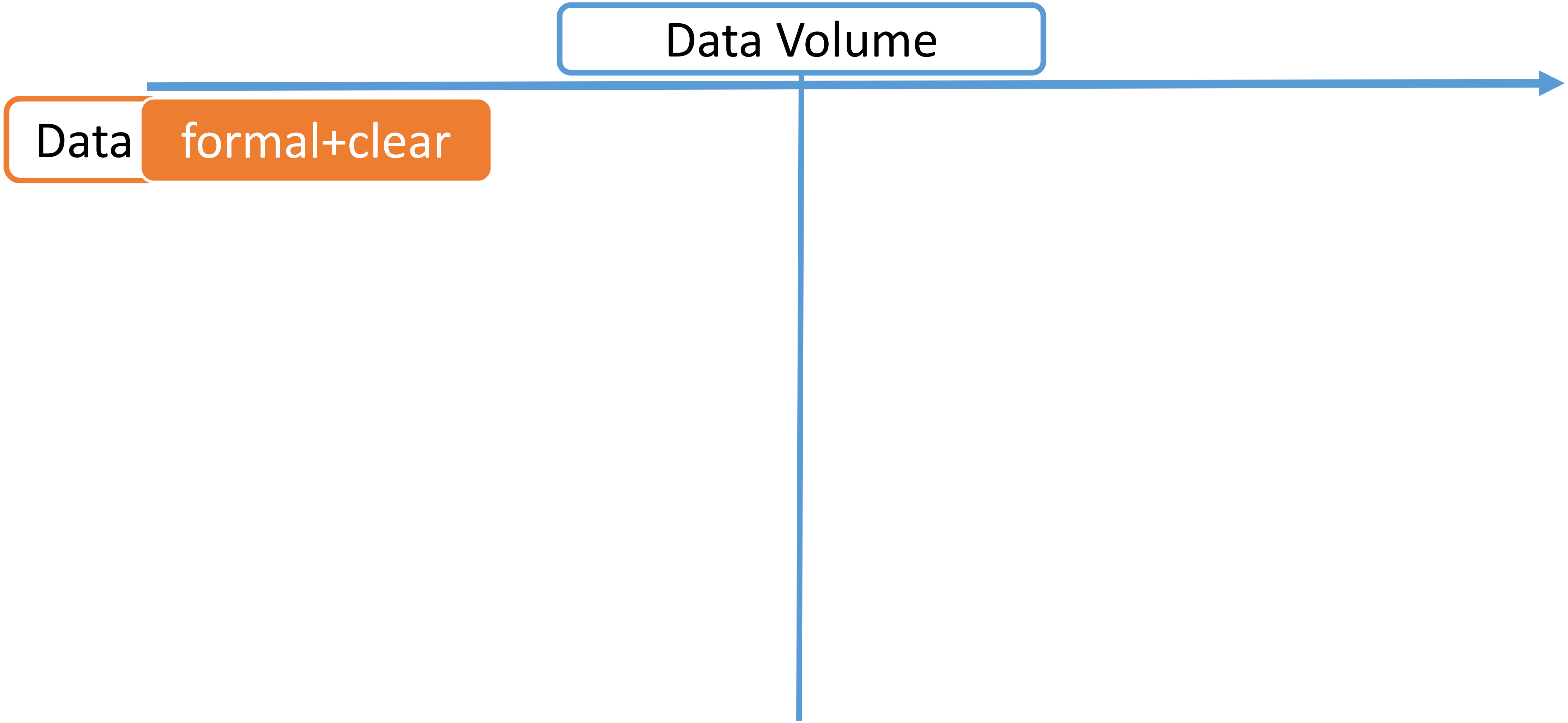
Label data → Train a classifier → Make prediction

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?



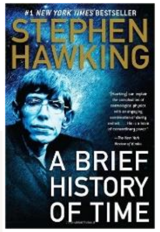
Challenges in Big Data Era:
How to Make Machine Learning Still Effective?



Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data | formal+clear | e.g., articles



the beginning of time would have been a point of infinite density and infinite curvature of space-time. All the known laws of science would break down at such a point. One might suppose that there were new laws that held at singularities, but it would be very difficult even to formulate such laws at such badly behaved points, and we would have no guide from observations as to what those laws might be. However, what the singularity theorems really indicate is that the gravitational field becomes so strong that quantum gravitational effects become important: classical theory is no longer a good description of the universe. So one has to use a quantum theory of gravity to discuss the very early stages of the universe. As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere, including at the beginning of time: it is not necessary to postulate new laws for singularities, because there need not be any singularities in the quantum theory.

We don't yet have a complete and consistent theory that combines quantum mechanics and gravity. However, we are fairly certain of some features that such a unified theory should have. One is that it should incorporate Feynman's proposal to formulate quantum theory in terms of a sum over histories. In this approach, a particle does not have just a single history, as it would in a classical theory. Instead, it is supposed to follow every possible path in

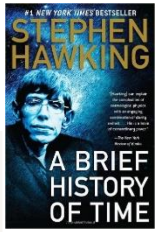
Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Data Volume

Data | formal+clear

e.g.,
articles

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere



Challenges in Big Data Era: How to Make Machine Learning Still Effective?

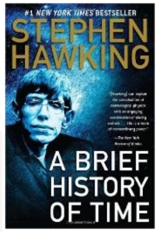
Data Volume

Data

formal+clear

e.g., articles

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

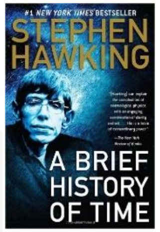


informal+noisy

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

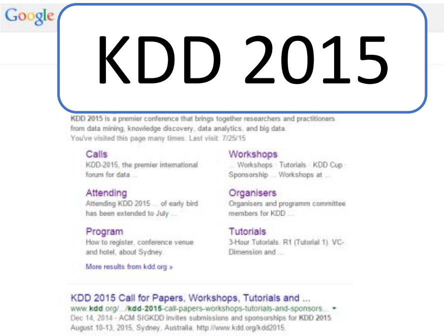
Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

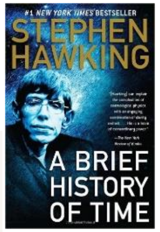
informal+noisy | e.g., Tweets, Queries



Challenges in Big Data Era: How to Make Machine Learning Still Effective?

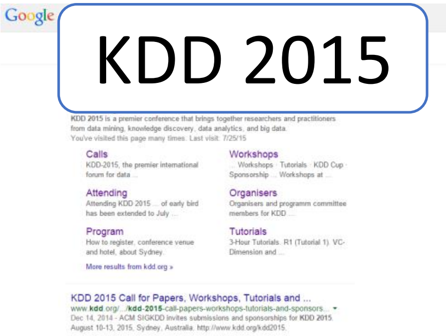
Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

informal+noisy | e.g., Tweets, Queries

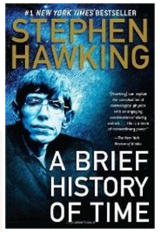


Label | coarse-grained

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

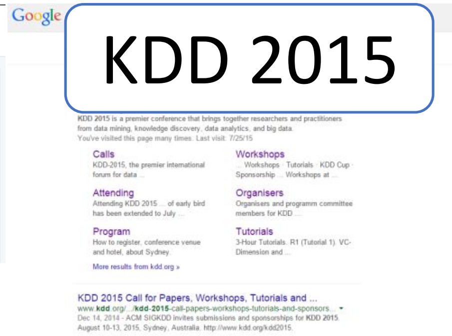
Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

informal+noisy | e.g., Tweets, Queries



KDD 2015

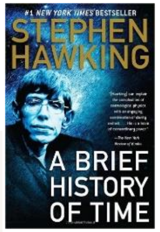
Label | coarse-grained

Insurance

Challenges in Big Data Era: How to Make Machine Learning Still Effective?

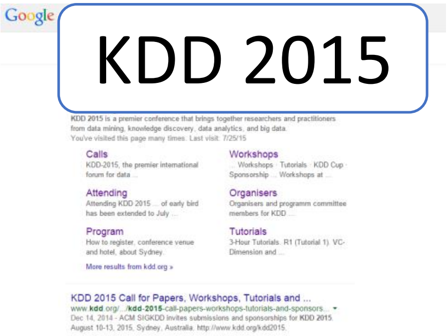
Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

informal+noisy | e.g., Tweets, Queries



Label | coarse-grained

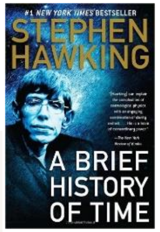
fine-grained

Insurance

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

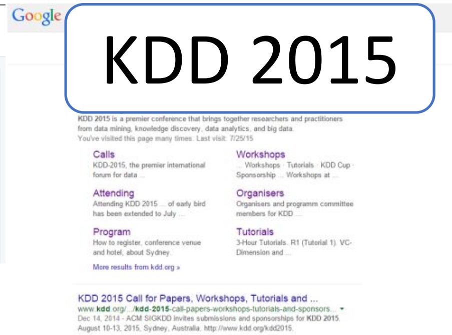
Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

informal+noisy | e.g., Tweets, Queries



...

Label | coarse-grained | fine-grained

Insurance → Health insurance, Auto insurance, ...

Health insurance → Vision insurance, Dental insurance, Disability insurance

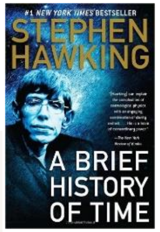
Auto insurance → ...

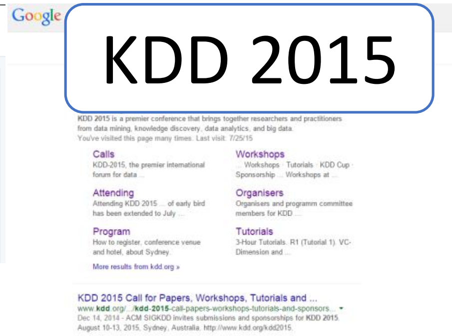
Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Data Volume

Data	formal+clear (e.g., articles)	informal+noisy (e.g., Tweets, Queries)
Label	coarse-grained	fine-grained

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere





...

Insurance

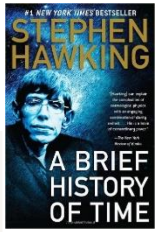
- Health insurance
  - Vision insurance
  - Dental insurance
  - Disability insurance
- Auto insurance
  - ...
- ...

Challenges in Big Data Era: How to Make Machine Learning Still Effective?

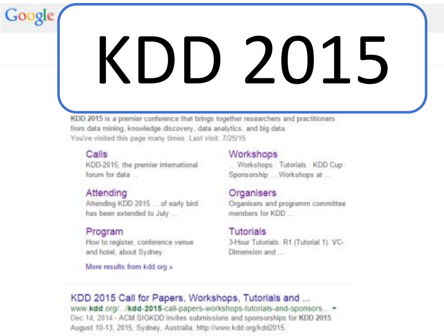
Data Volume

Data: formal+clear (e.g., articles) + informal+noisy (e.g., Tweets, Queries)

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere



KDD 2015



...

Label: coarse-grained + fine-grained

Insurance → Health insurance, Auto insurance, ...

Health insurance → Vision insurance, Dental insurance, Disability insurance

Auto insurance → ...

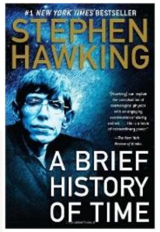
Machine Learning Ability

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Data Volume

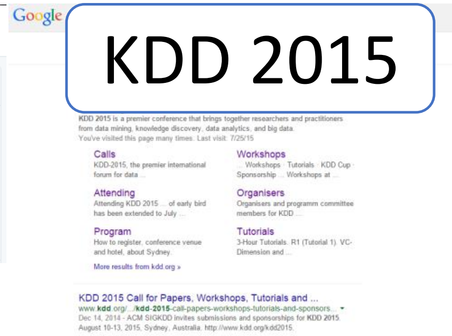
Data | formal+clear | e.g., articles

As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere



informal+noisy | e.g., Tweets, Queries

KDD 2015



...

Label | coarse-grained

NOKIA, SONY, DELL, intel, Apple, IBM, hp, Microsoft, ORACLE

fine-grained

Health insurance → Vision insurance, Dental insurance, Disability insurance

Auto insurance → ...

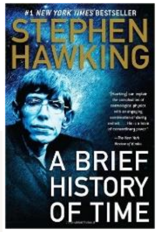
...

Machine Learning Ability

Challenges in Big Data Era:
How to Make Machine Learning Still Effective?

Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

Label | coarse-grained

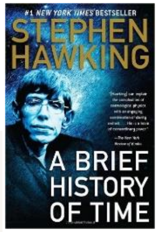
...

Machine Learning Ability

Challenges in Big Data Era: How to Make Machine Learning Still Effective?

Data Volume

Data | formal+clear | e.g., articles



As we shall see, it is possible in the quantum theory for the ordinary laws of science to hold everywhere

Label | coarse-grained

More applications!

...

Machine Learning Ability

Acquire Labeled Data

Acquire Labeled Data

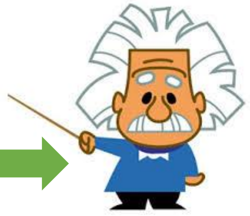
Expert Annotation

Costly

Acquire Labeled Data

Expert Annotation

Costly

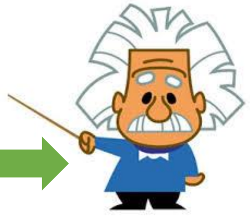


Only big companies can hire a lot of experts

Acquire Labeled Data

Expert Annotation

Costly



Only big companies can hire a lot of experts

Crowdsourcing

Simple tasks

Low quality

Still costly

Acquire Labeled Data

Expert Annotation

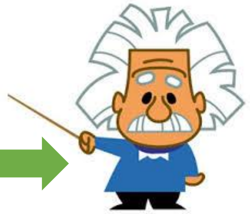
Costly

Facebook LinkedIn IBM

Microsoft

Google YAHOO!

amazon



Only big companies can hire a lot of experts

Crowdsourcing

Simple tasks

Low quality

Still costly

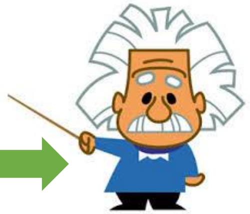
amazon mechanical turk

freelancer

Acquire Labeled Data

Expert Annotation

Costly



Only big companies can hire a lot of experts

Crowdsourcing

Simple tasks

Low quality

Still costly

Semi-supervised /transfer learning

Not generalizable

Acquire Labeled Data

Expert Annotation

Costly

Crowdsourcing

Simple tasks

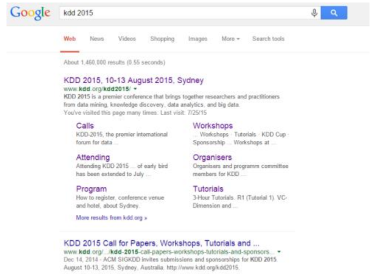
Low quality

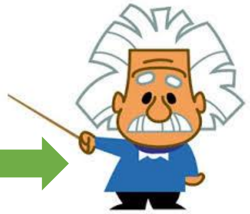
Still costly

Semi-supervised /transfer learning

Not generalizable

Fast changing domains



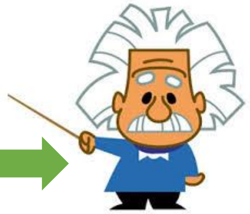


Only big companies can hire a lot of experts

Acquire Labeled Data

Expert Annotation

Costly



Only big companies can hire a lot of experts

Crowdsourcing

Simple tasks

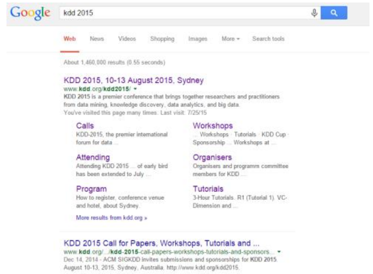
Low quality

Still costly

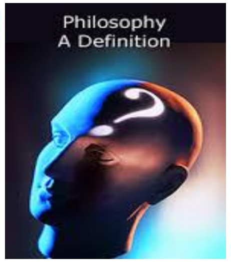
Semi-supervised /transfer learning

Not generalizable

Fast changing domains



Many diverse domains



Knowledge Enabled Learning: use knowledge as indirect supervision

Knowledge Enabled Learning:
use knowledge as indirect supervision

World knowledge bases

Knowledge Enabled Learning: use knowledge as indirect supervision

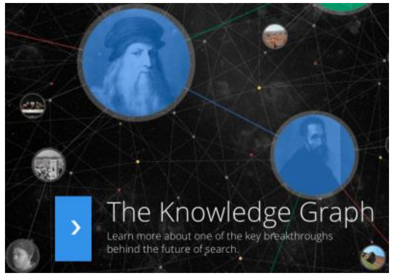
World knowledge bases

WIKIPEDIA
The Free Encyclopedia

DBpedia

Freebase Labs

yago
select knowledge



Knowledge Enabled Learning: use knowledge as indirect supervision

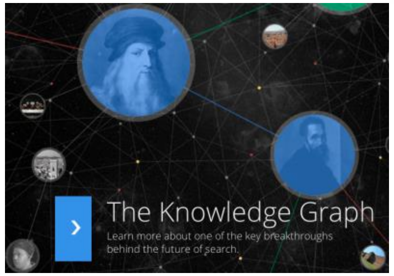
World knowledge bases | *Indirect* label

Wikipedia
The Free Encyclopedia

DBpedia

Freebase
Labs

yago
select knowledge



Knowledge Enabled Learning: use knowledge as indirect supervision

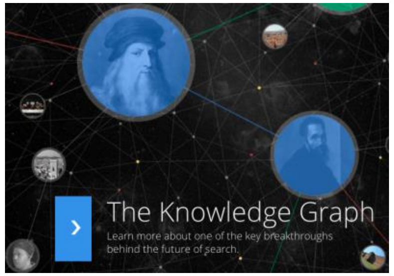
World knowledge bases | *Indirect* label

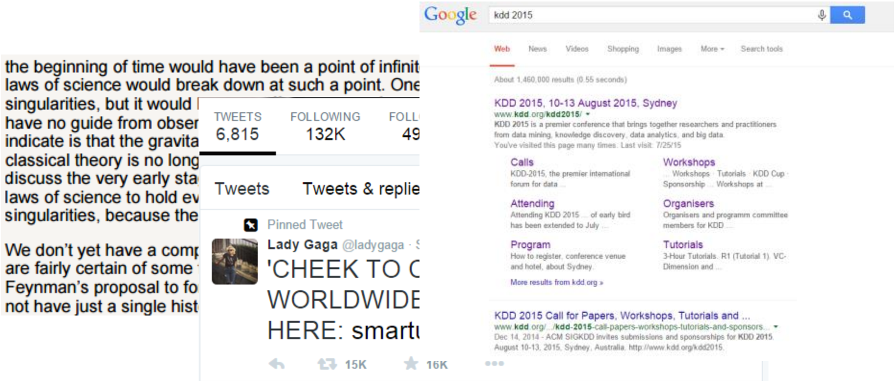
Wikipedia
The Free Encyclopedia

DBpedia

Freebase Labs

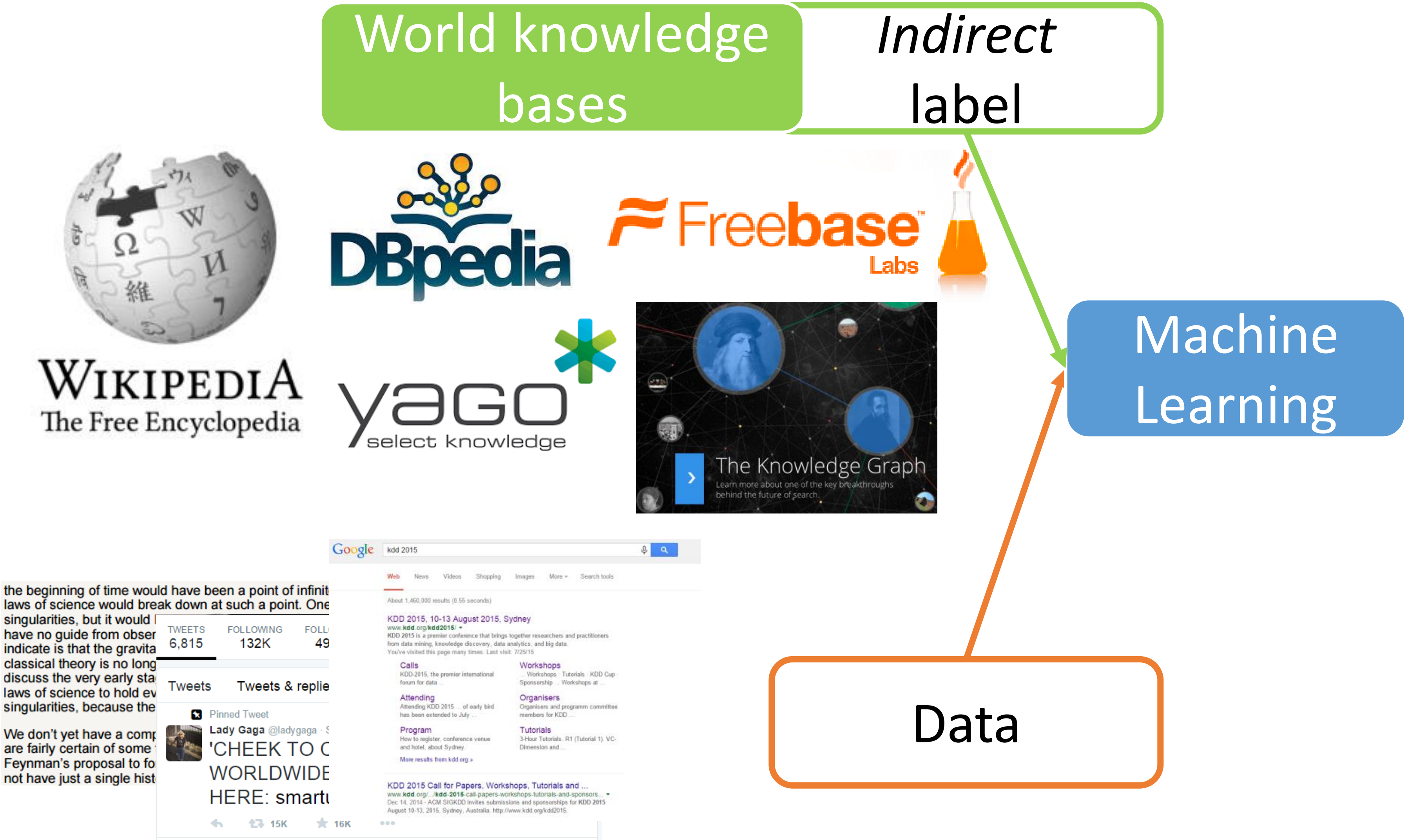
yago select knowledge



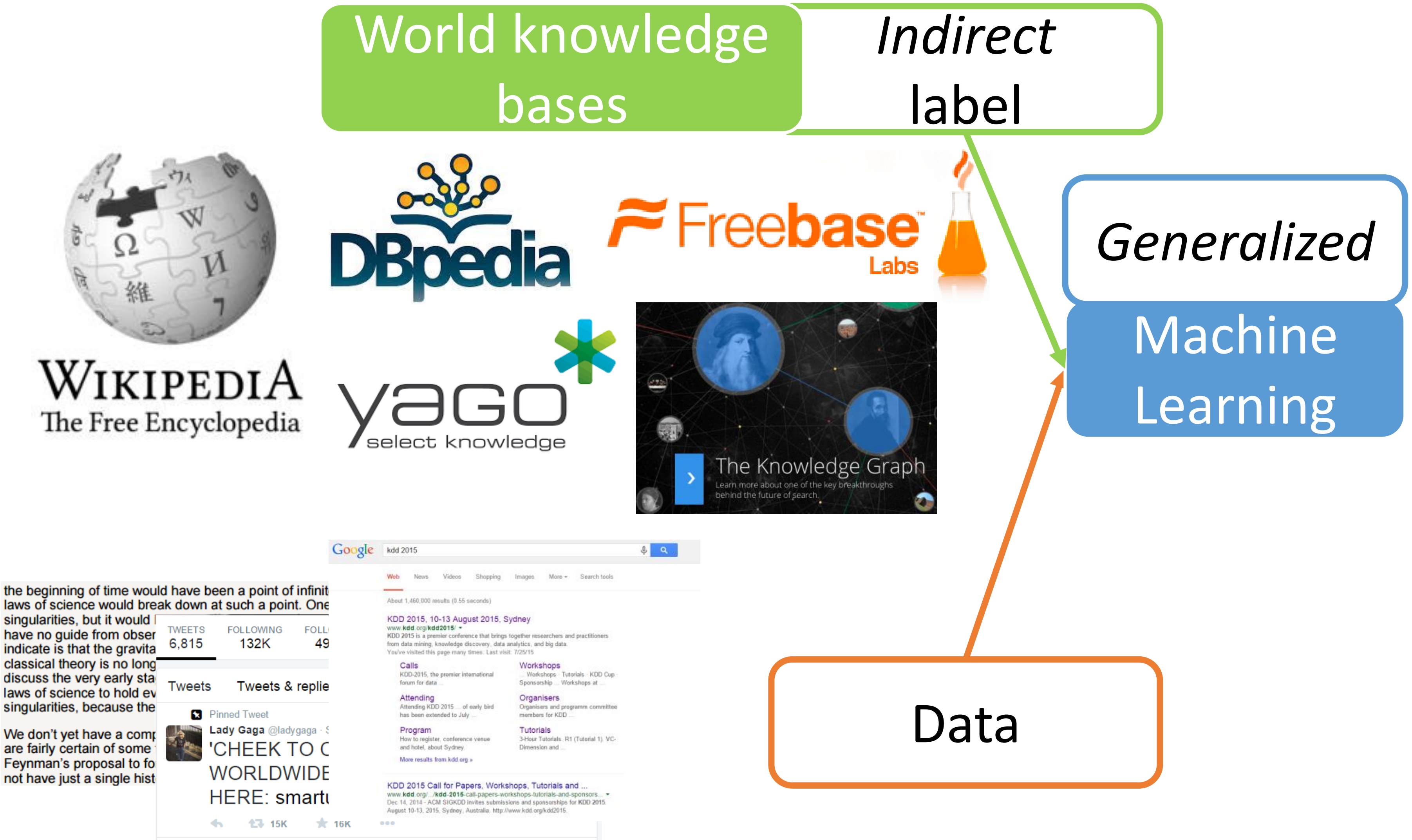


Data

Knowledge Enabled Learning: use knowledge as indirect supervision



Knowledge Enabled Learning: use knowledge as indirect supervision



Knowledge Enabled Learning: use knowledge as indirect supervision

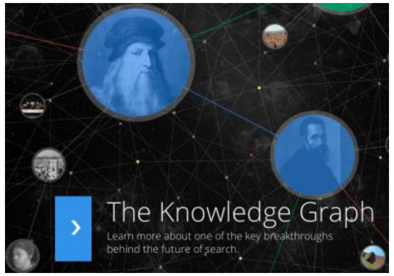
World knowledge bases

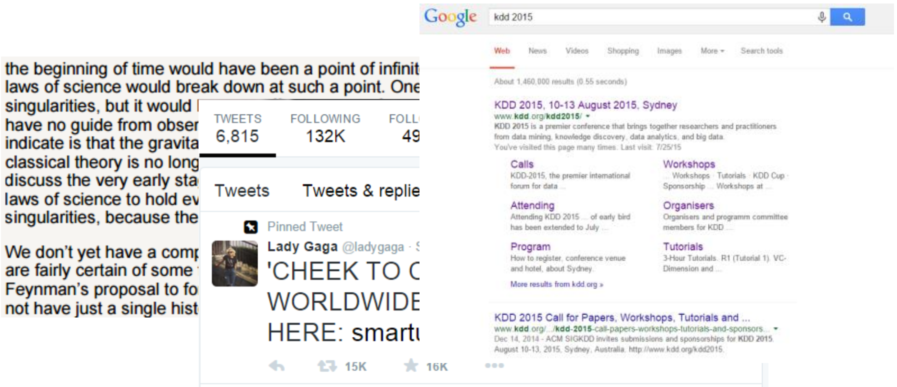
Indirect label

Generalized

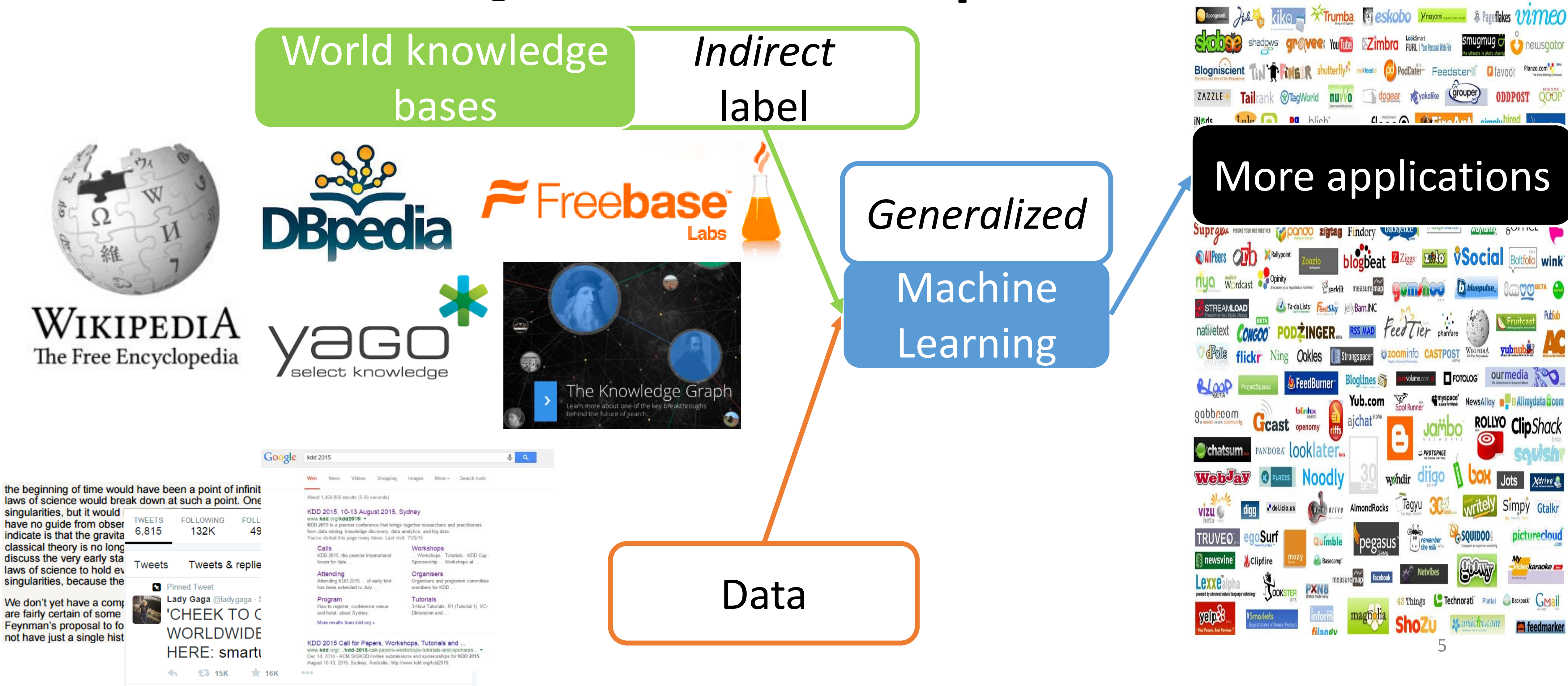
Machine Learning

Data

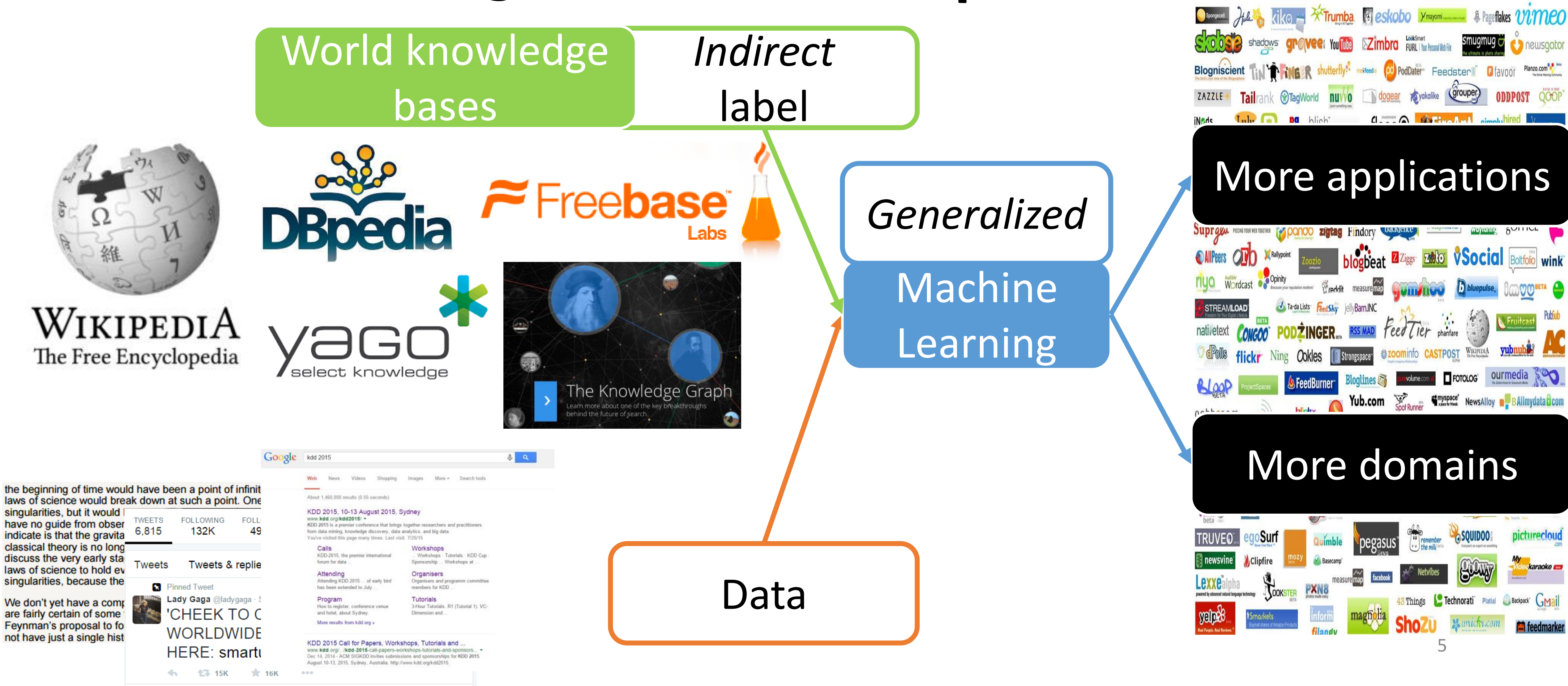




Knowledge Enabled Learning: use knowledge as indirect supervision



Knowledge Enabled Learning: use knowledge as indirect supervision



Example: Knowledge Enabled Text Clustering

Example: Knowledge Enabled Text Clustering

On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

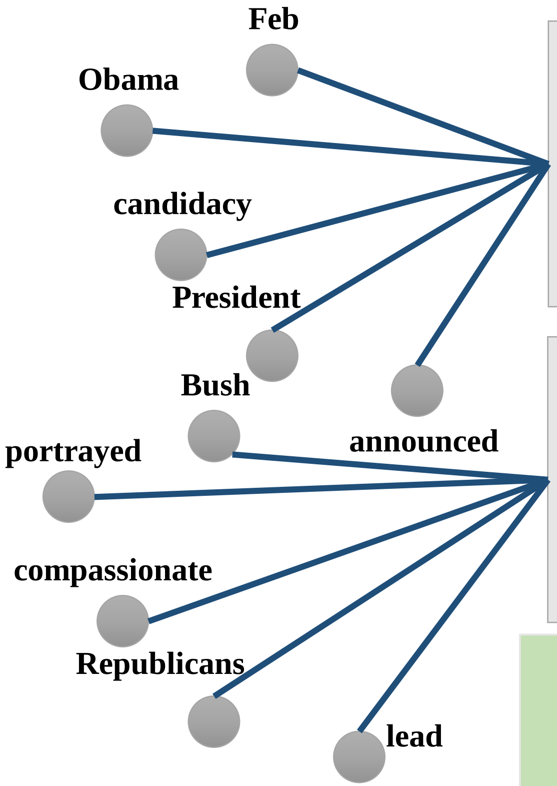
Example: Knowledge Enabled Text Clustering

On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

Are the two documents belong to the same cluster?

Example: Knowledge Enabled Text Clustering

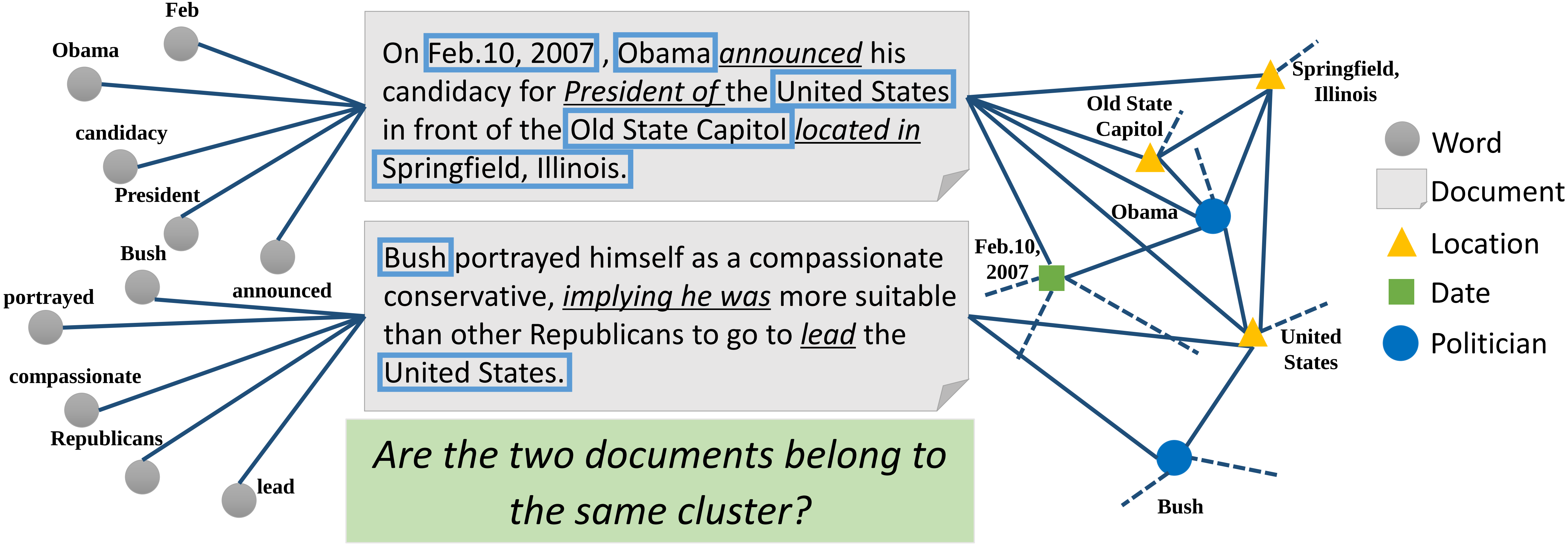


On Feb.10, 2007 , Obama *announced* his candidacy for *President of* the United States in front of the Old State Capitol *located in* Springfield, Illinois.

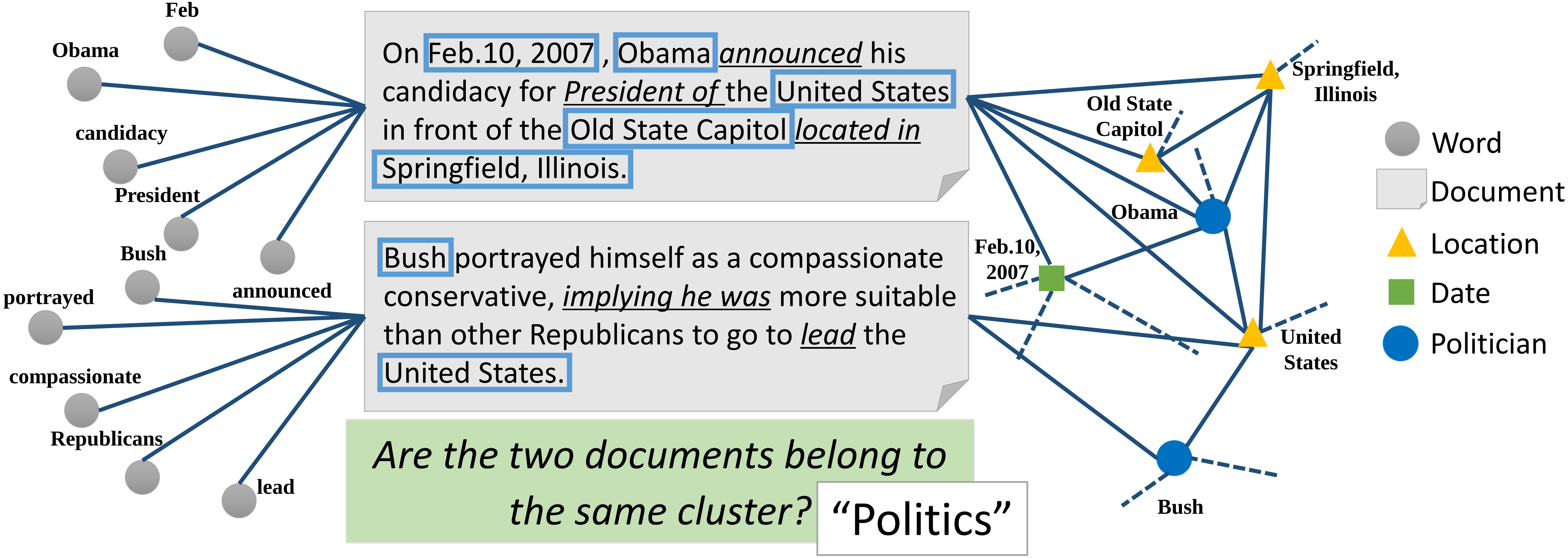
Bush portrayed himself as a compassionate conservative, *implying he was* more suitable than other Republicans to go to *lead* the United States.

Are the two documents belong to the same cluster?

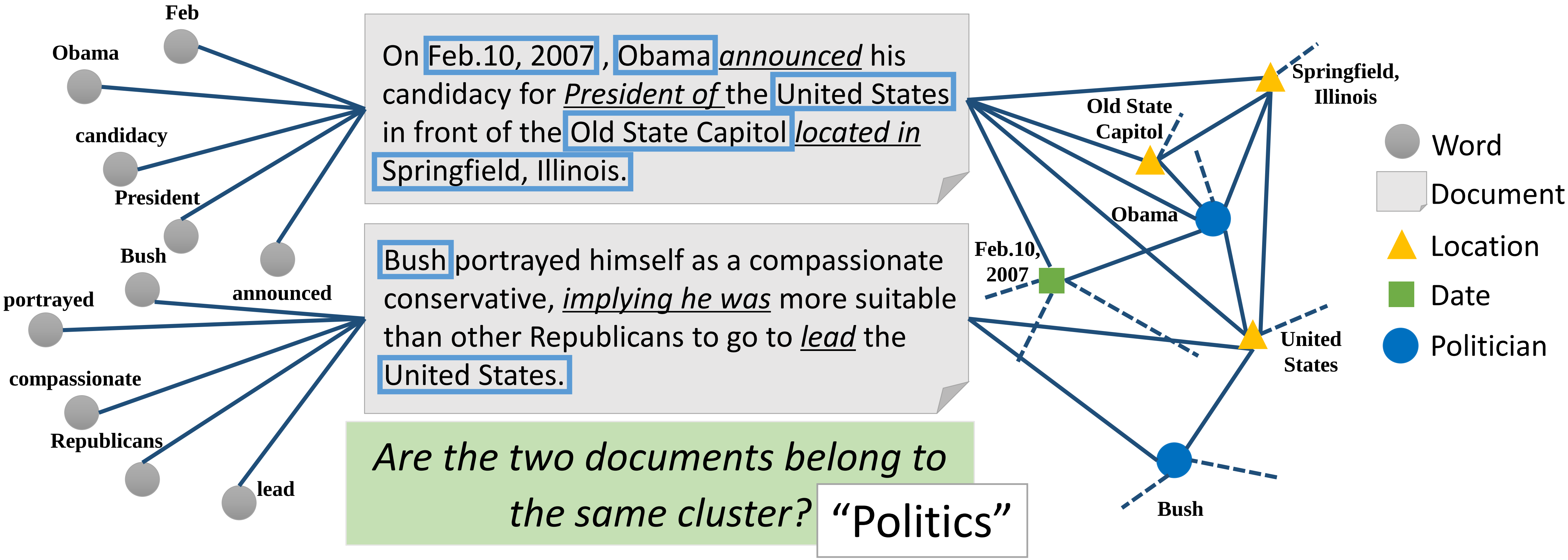
Example: Knowledge Enabled Text Clustering



Example: Knowledge Enabled Text Clustering

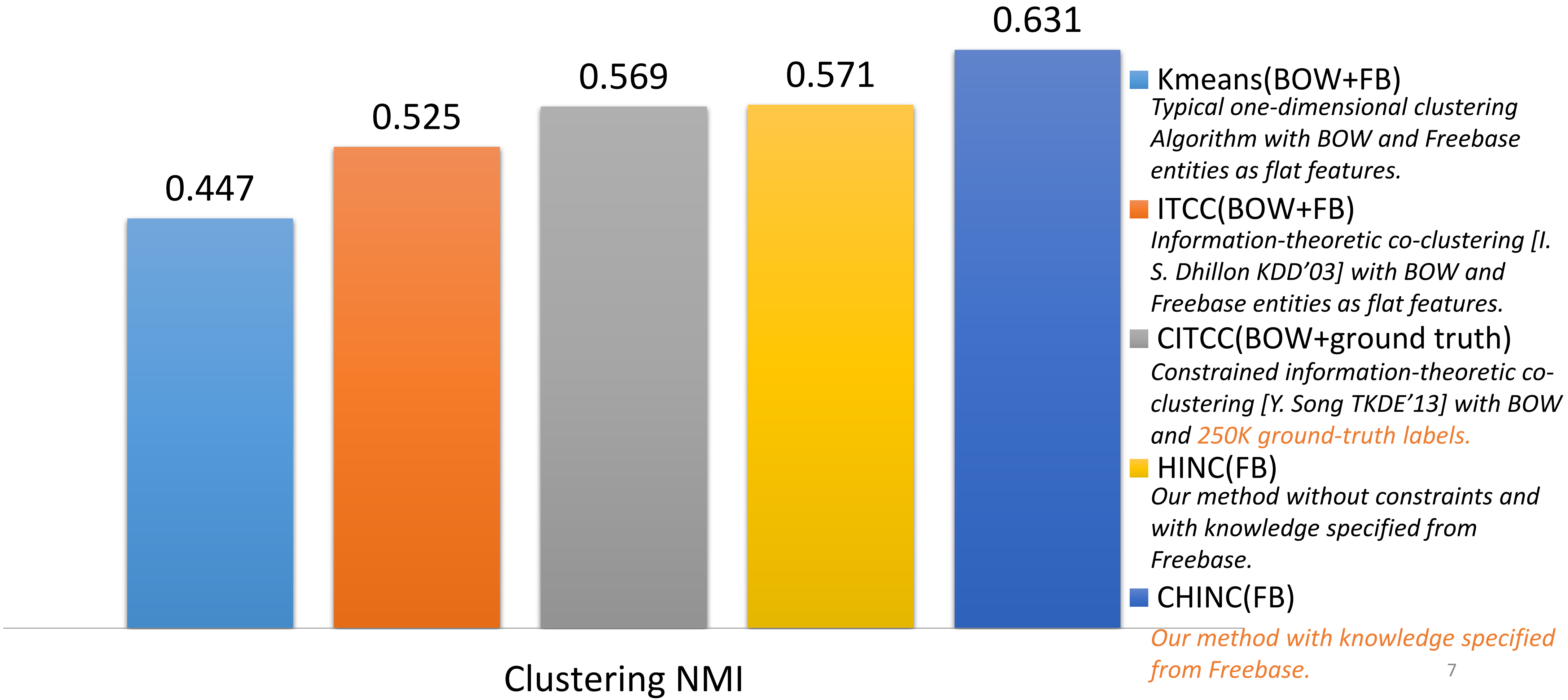


Example: Knowledge Enabled Text Clustering



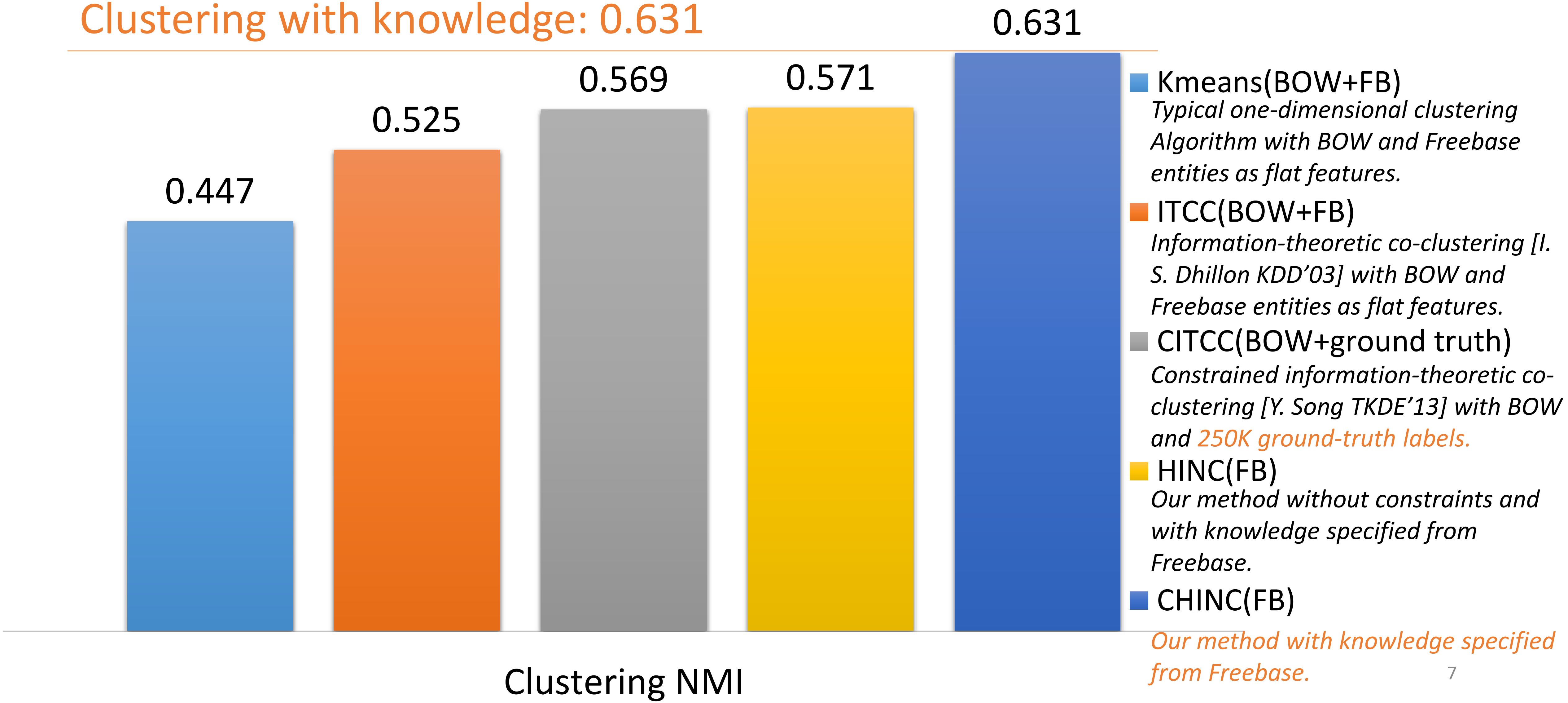
- Links and types carry a lot of information!
- But traditional approaches are not using them

Clustering of 20 Newsgroups Documents



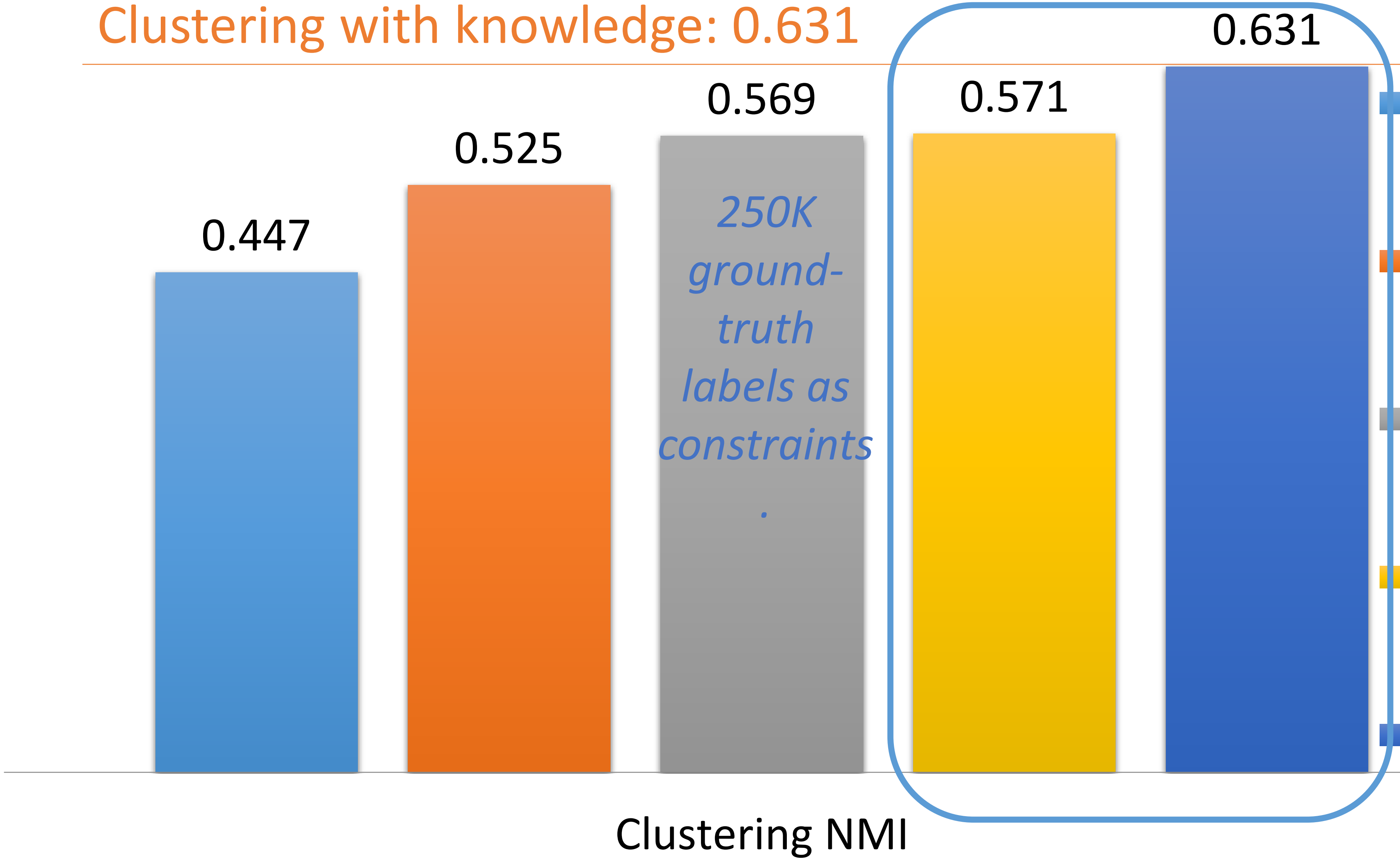
Clustering of 20 Newsgroups Documents

Clustering with knowledge: 0.631



Clustering of 20 Newsgroups Documents

Clustering with knowledge: 0.631



Kmeans(BOW+FB)
Typical one-dimensional clustering Algorithm with BOW and Freebase entities as flat features.

ITCC(BOW+FB)
Information-theoretic co-clustering [I. S. Dhillon KDD'03] with BOW and Freebase entities as flat features.

CITCC(BOW+ground truth)
Constrained information-theoretic co-clustering [Y. Song TKDE'13] with BOW and 250K ground-truth labels.

HINC(FB)
Our method without constraints and with knowledge specified from Freebase.

CHINC(FB)
Our method with knowledge specified from Freebase.

Knowledge Enabled Learning

Challenges

Future

Knowledge Enabled Learning

Challenges

Future

World Knowledge

Data

Knowledge Enabled Learning

Challenges

Future

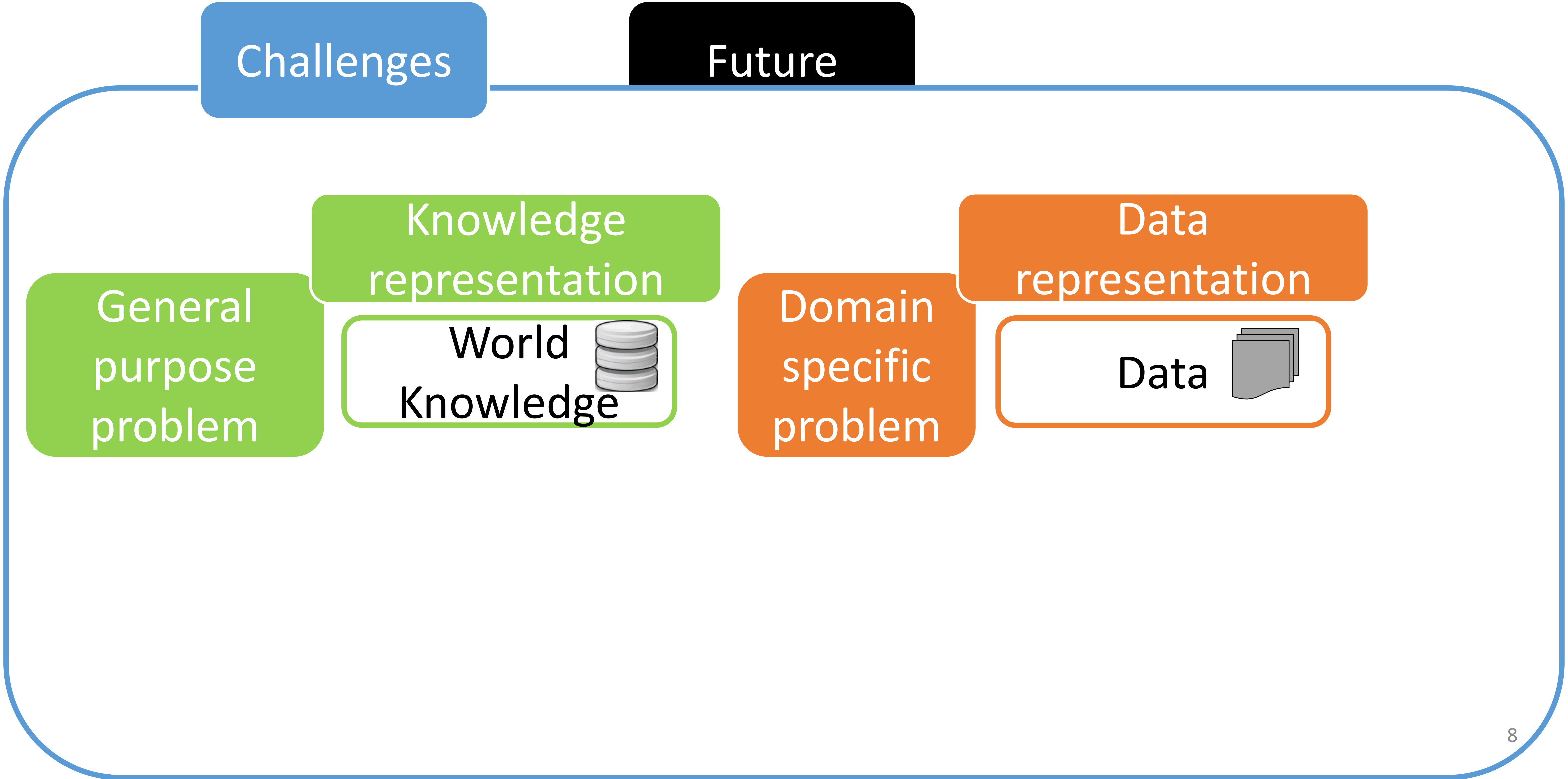
General purpose problem

World Knowledge

Domain specific problem

Data

Knowledge Enabled Learning



Knowledge Enabled Learning

Challenges

Future

General purpose problem

Knowledge representation

World Knowledge

Large scale inference

Domain specific problem

Data representation

Data

Small scale inference

Knowledge Enabled Learning

Challenges

Future

Knowledge Enabled Learning

Challenges

Future

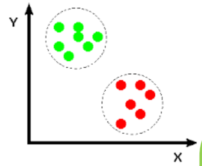
More general and effective machine learning

Knowledge Enabled Learning

Challenges

Future

More applications



e.g., clustering, classification, recommendation

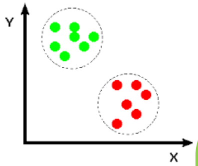
More general and effective machine learning

Knowledge Enabled Learning

Challenges

Future

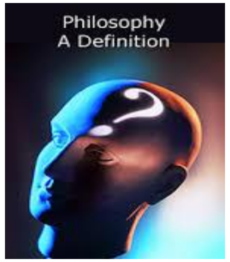
More applications



e.g., clustering, classification, recommendation

More general and effective machine learning

More domains



e.g., tweets, blogs, websites, medical, psychology

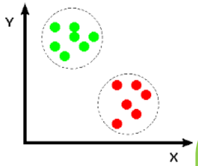
Knowledge Enabled Learning

Challenges

Future

Next generation of machine learning

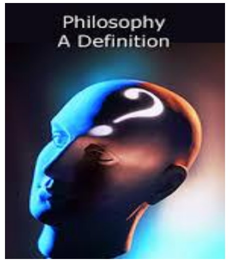
More applications



e.g., clustering, classification, recommendation

Machine learning algorithms + World knowledge bases = **More general and effective machine learning**

More domains



e.g., tweets, blogs, websites, medical, psychology

Knowledge Enabled Learning

Challenges

Future

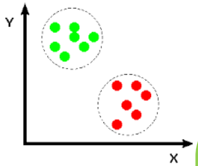
Next generation of machine learning

Machine learning algorithms + World knowledge bases = **More general and effective machine learning**

Data

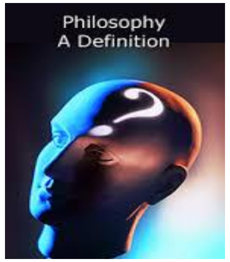
Big data enabled machine learning

More applications



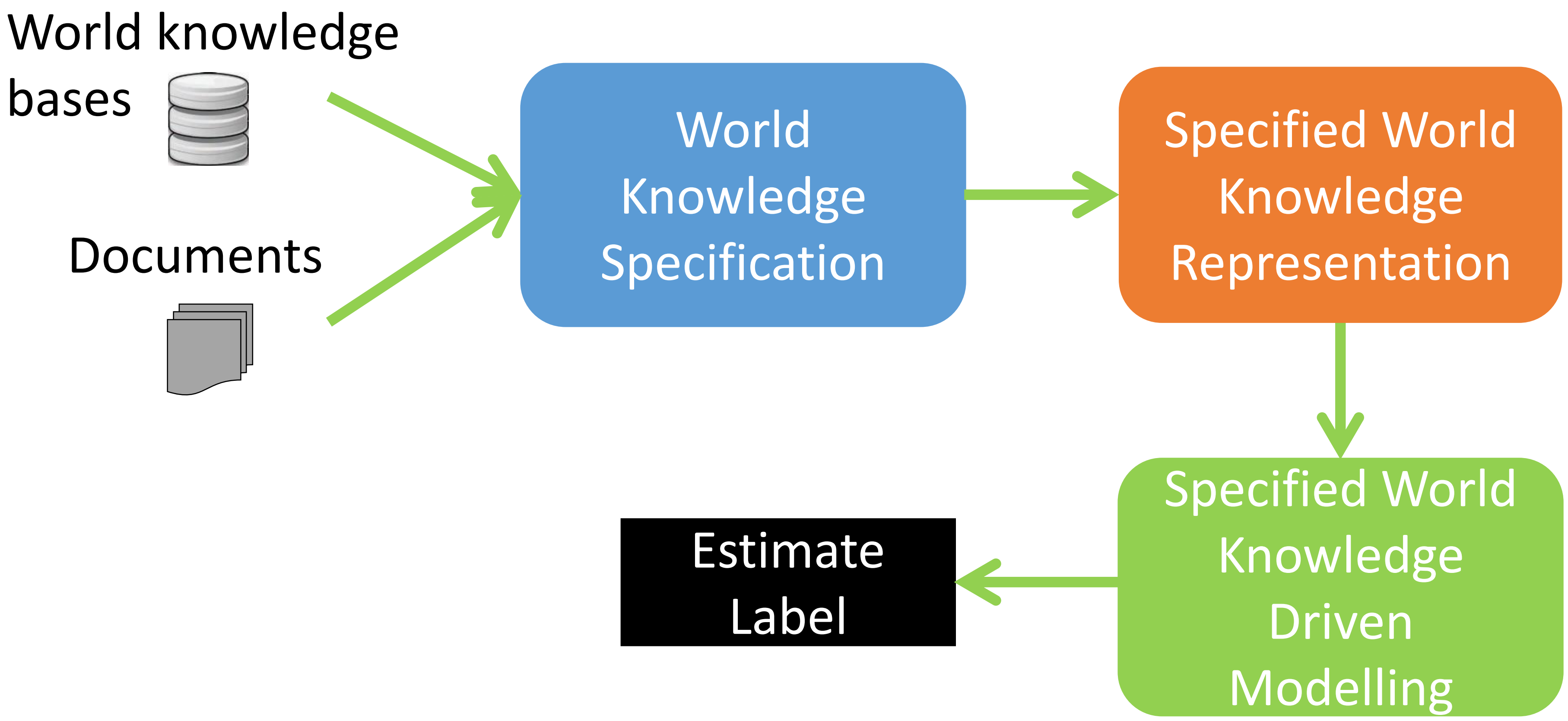
e.g., clustering, classification, recommendation

More domains

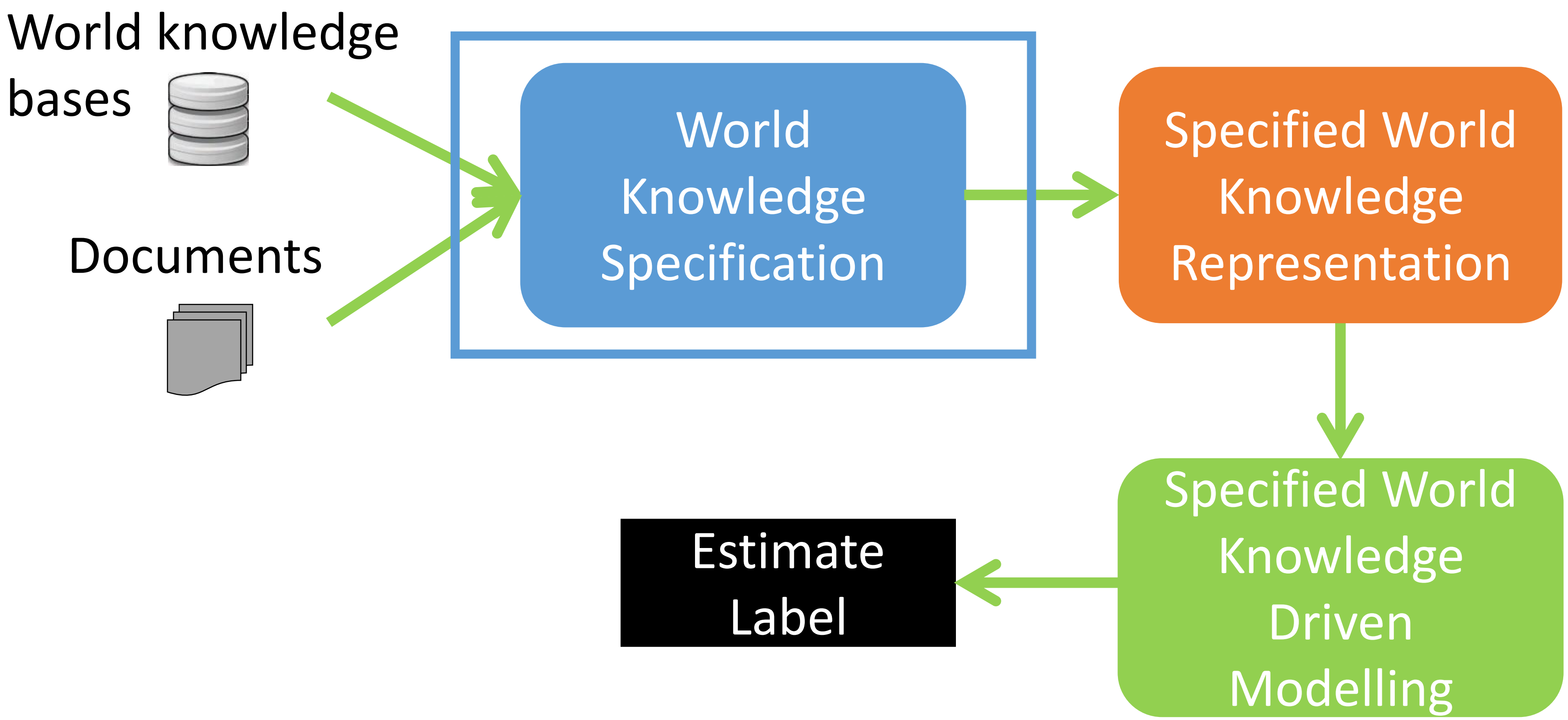


e.g., tweets, blogs, websites, medical, psychology

Text Clustering with World Knowledge



Text Clustering with World Knowledge



Unsupervised Semantic Parsing for Documents

Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America

Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Logic form *People.BarackObama* $\sqcap$ PresidentofCountry.*Country.USA*

Unsupervised Semantic Parsing for Documents

Document Obama is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

Logic form *People.BarackObama* $\sqcap$ PresidentofCountry.*Country.USA*

- Motivation: [J. Berant et al. EMNLP'13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase
  - Existing semantic parsing approaches, that require expert annotation
  - Scales to large scale knowledge-bases, supervised by the QA pairs

Unsupervised Semantic Parsing for Documents

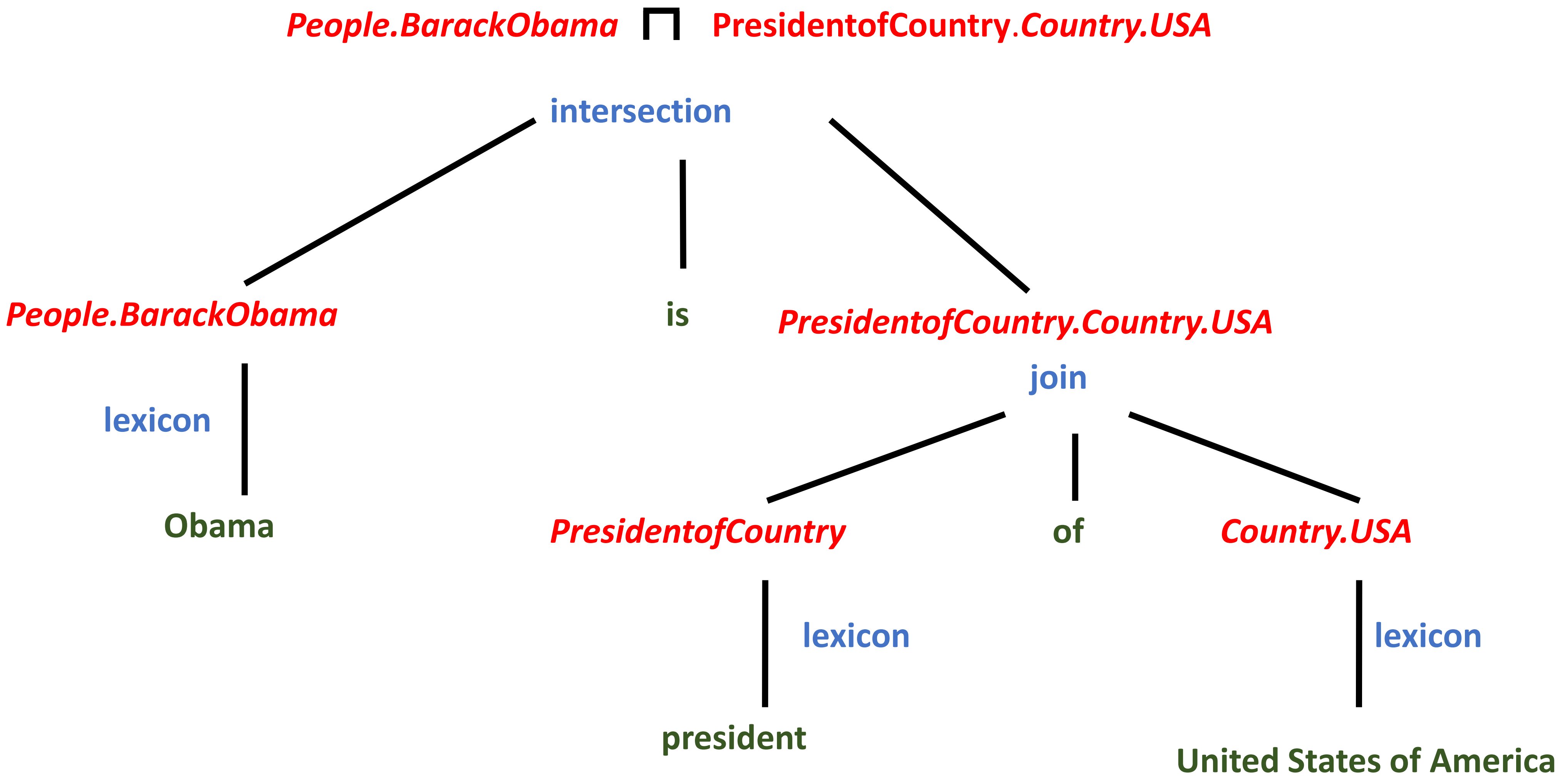
Document Obama is the president of the United States of America

Semantic parsing is the task of mapping a piece of natural language text to a formal meaning representation.

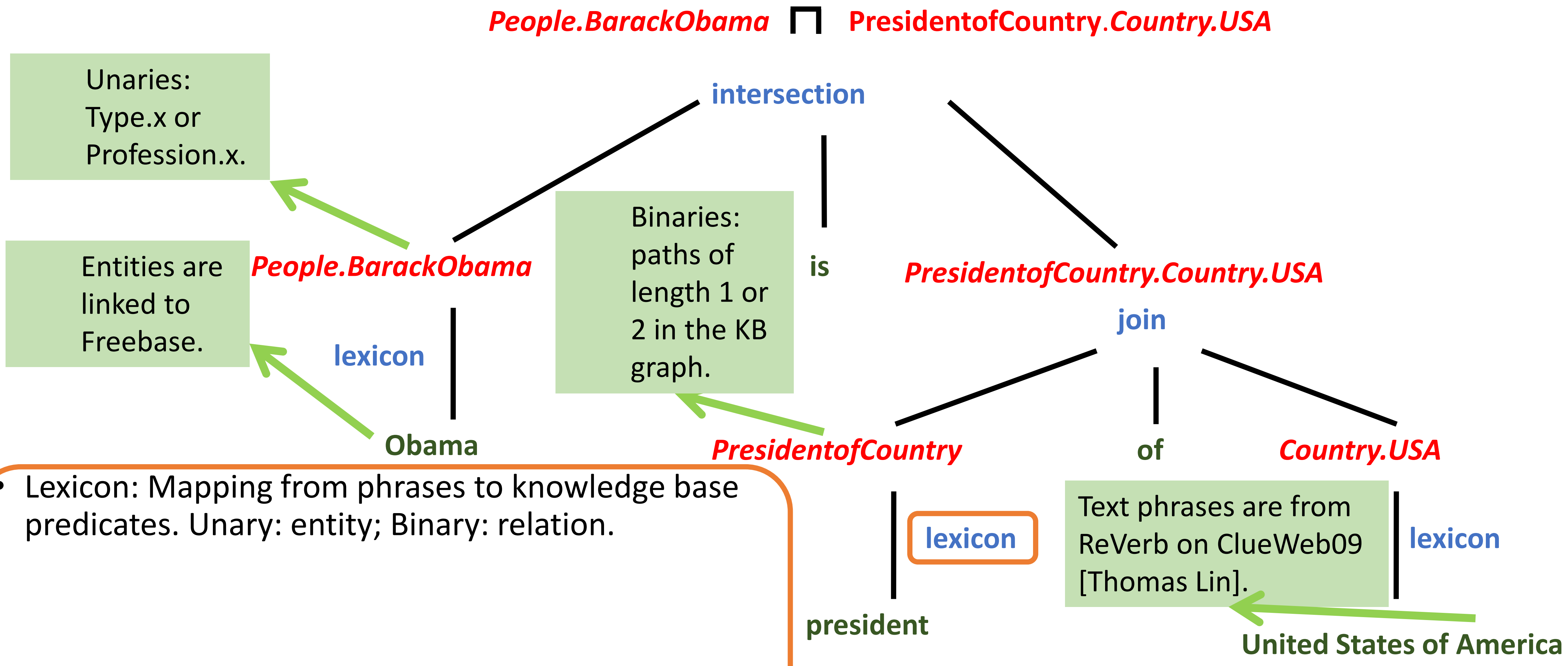
Logic form *People.BarackObama* $\sqcap$ PresidentofCountry.*Country.USA*

- Motivation: [J. Berant et al. EMNLP'13] aim to train a parser from question/answer pairs on a large knowledge-base Freebase
  - Existing semantic parsing approaches, that require expert annotation
  - Scales to large scale knowledge-bases, supervised by the QA pairs
- No such training data for the document dataset.

Unsupervised Semantic Parsing for Documents

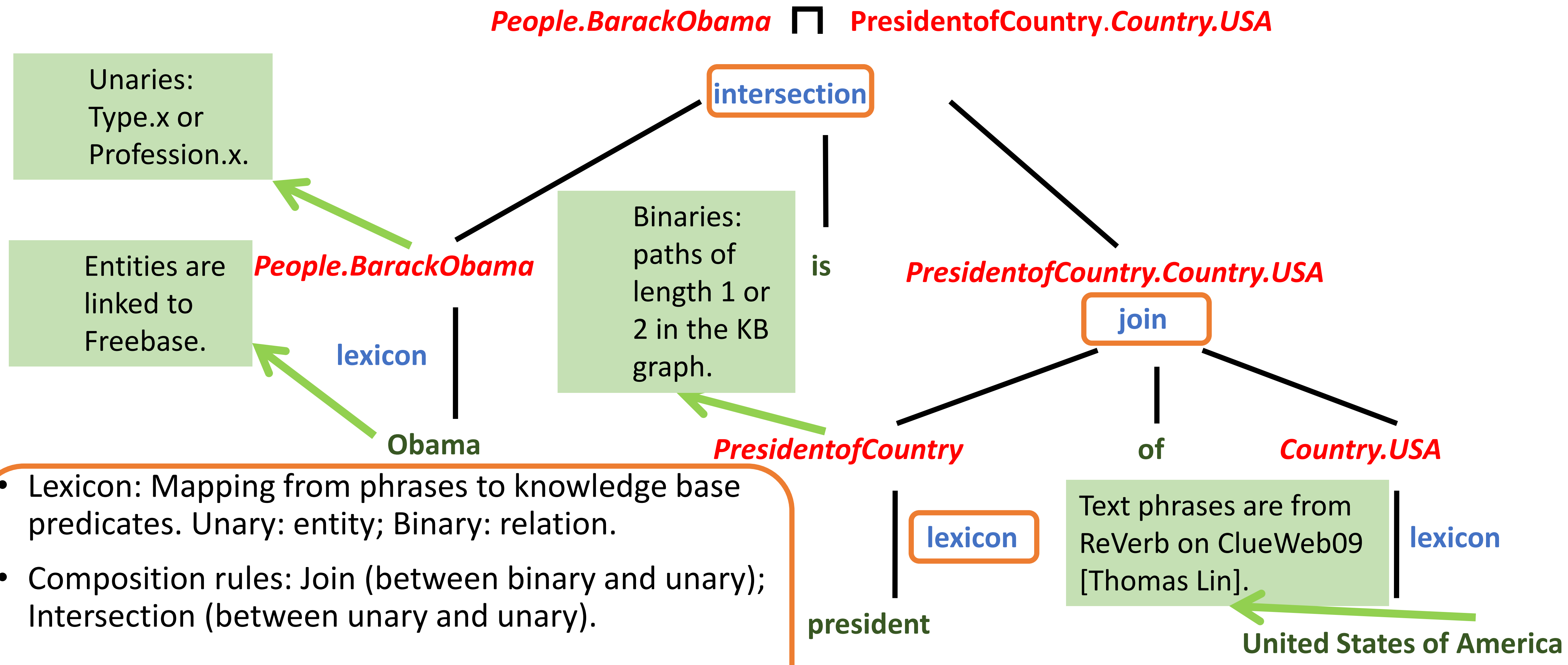


Unsupervised Semantic Parsing for Documents



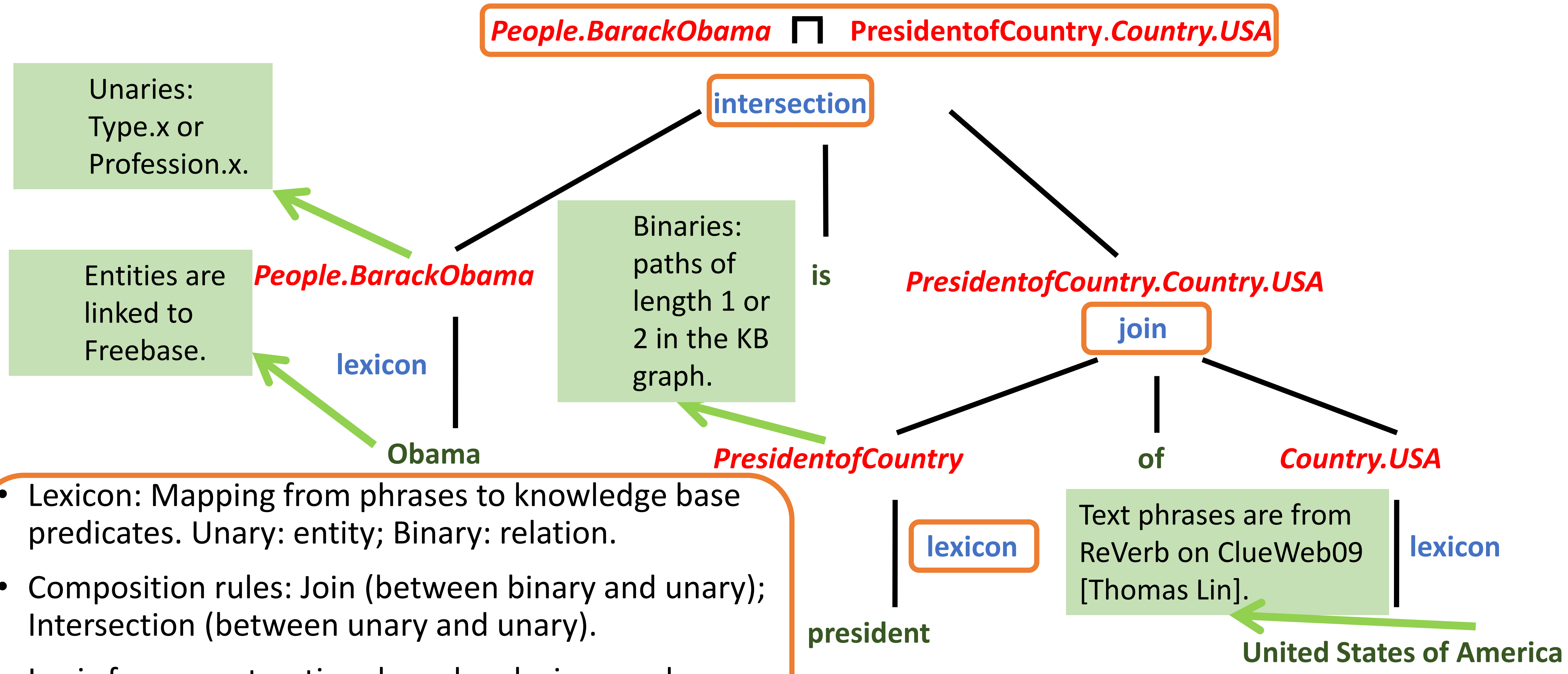
- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.

Unsupervised Semantic Parsing for Documents



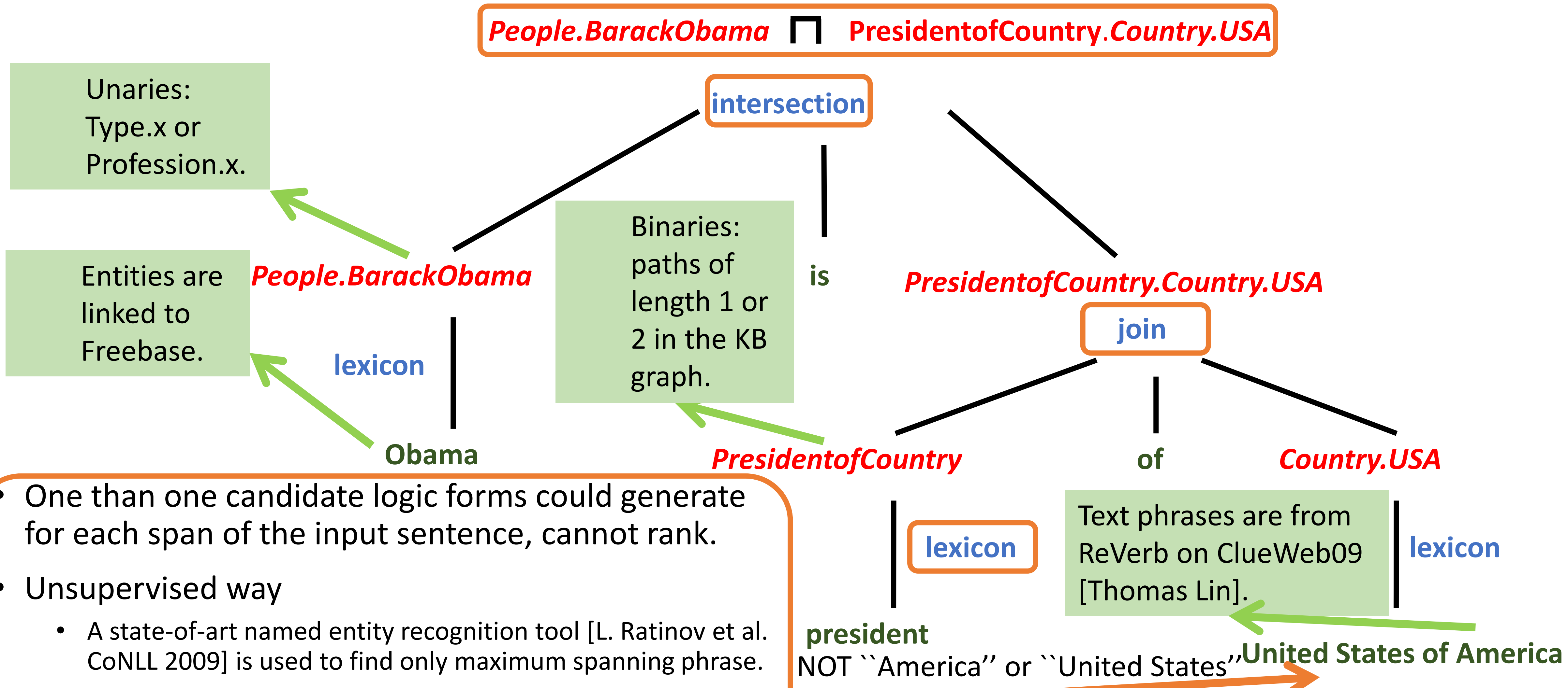
- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).

Unsupervised Semantic Parsing for Documents



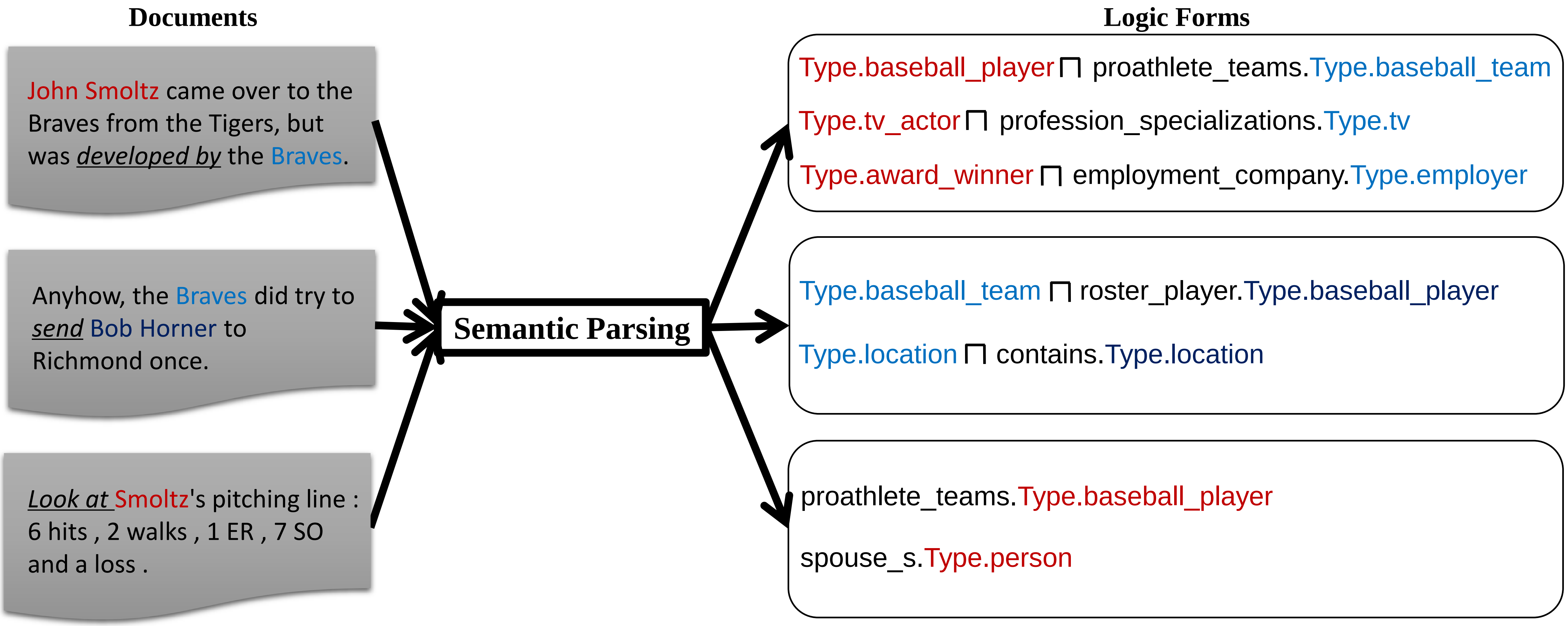
- Lexicon: Mapping from phrases to knowledge base predicates. Unary: entity; Binary: relation.
- Composition rules: Join (between binary and unary); Intersection (between unary and unary).
- Logic form construction: based on lexicon and composition rules recursively.

Unsupervised Semantic Parsing for Documents



- One than one candidate logic forms could generate for each span of the input sentence, cannot rank.
- Unsupervised way
  - A state-of-art named entity recognition tool [L. Ratinov et al. CoNLL 2009] is used to find only maximum spanning phrase.
  - Only generate partial immediate logic form based on the maximum spanning phrase.

Examples of Semantic Parsing on 20NG



Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

  Assumption: correct semantic meaning can best fit the context.
  Different entities can be used to disambiguate each other.

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple adobe

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple

software company, brand, fruit

adobe

brand, software company

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple — software company, brand, fruit

adobe — brand, software company

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple adobe

software company, brand, fruit brand, software company

software company, brand

P(type | related entities)

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple — software company, brand, fruit

adobe — brand, software company

software company, brand

P(type | related entities)

A cluster of entities of type features

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI'11.

Semantic Filtering

- Conceptualization based semantic filter (CBSF).

Assumption: correct semantic meaning can best fit the context.
Different entities can be used to disambiguate each other.

apple — software company, brand, fruit

adobe — brand, software company

software company, brand

largest probability ones are selected — P(type | related entities) — A cluster of entities of type features

Song et al. Short text conceptualization using a probabilistic knowledgebase. IJCAI’11.

Examples of Semantic Filtering on 20NG

Logic Forms

Type.baseball_player $\sqcap$ proathlete_teams.Type.baseball_team

Type.tv_actor $\sqcap$ profession_specializations.Type.tv

Type.award_winner $\sqcap$ employment_company.Type.employer

Type.baseball_team $\sqcap$ roster_player.Type.baseball_player

Type.location $\sqcap$ contains.Type.location

proathlete_teams.Type.baseball_player

spouse_s.Type.person

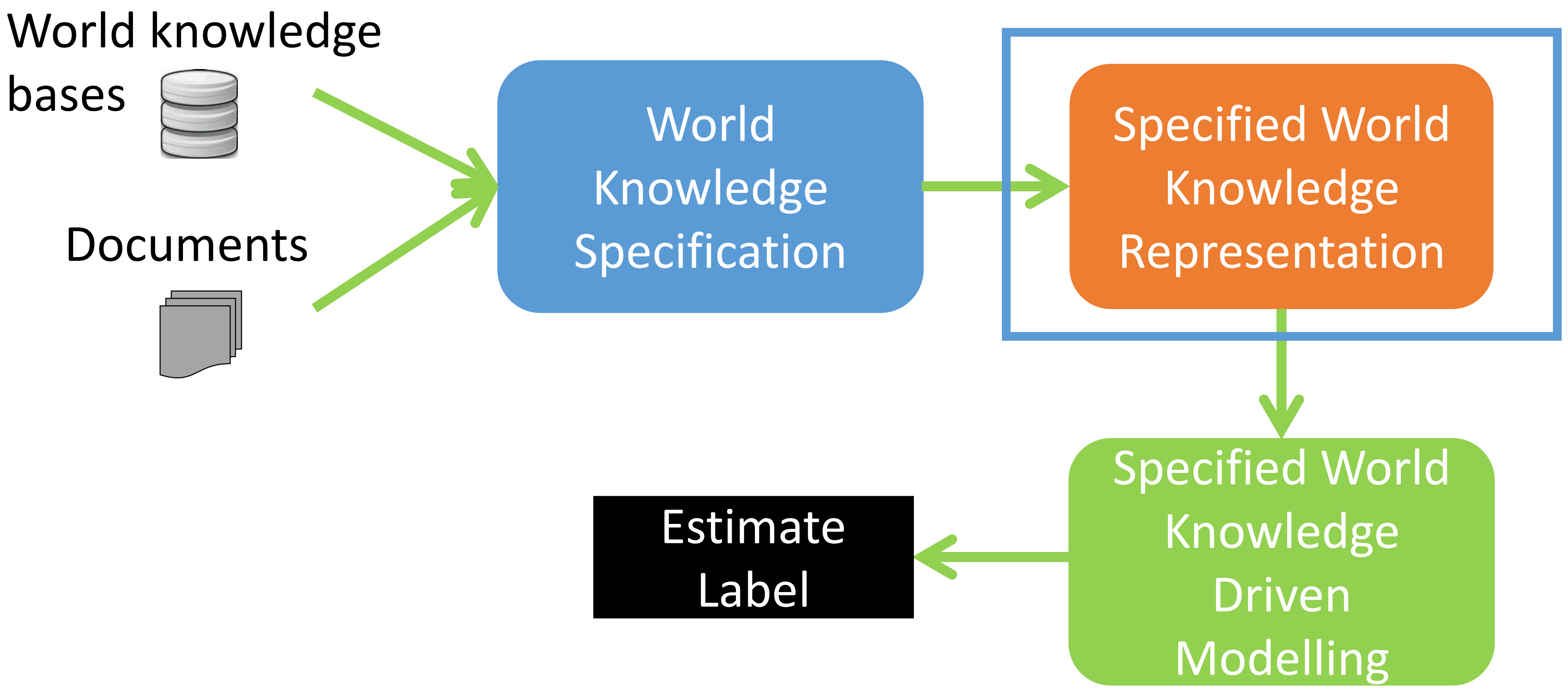
Semantic Filtering

Filtered Semantics

John Smoltz:Type.baseball_player

Braves:Type.baseball_team

Text Clustering with World Knowledge



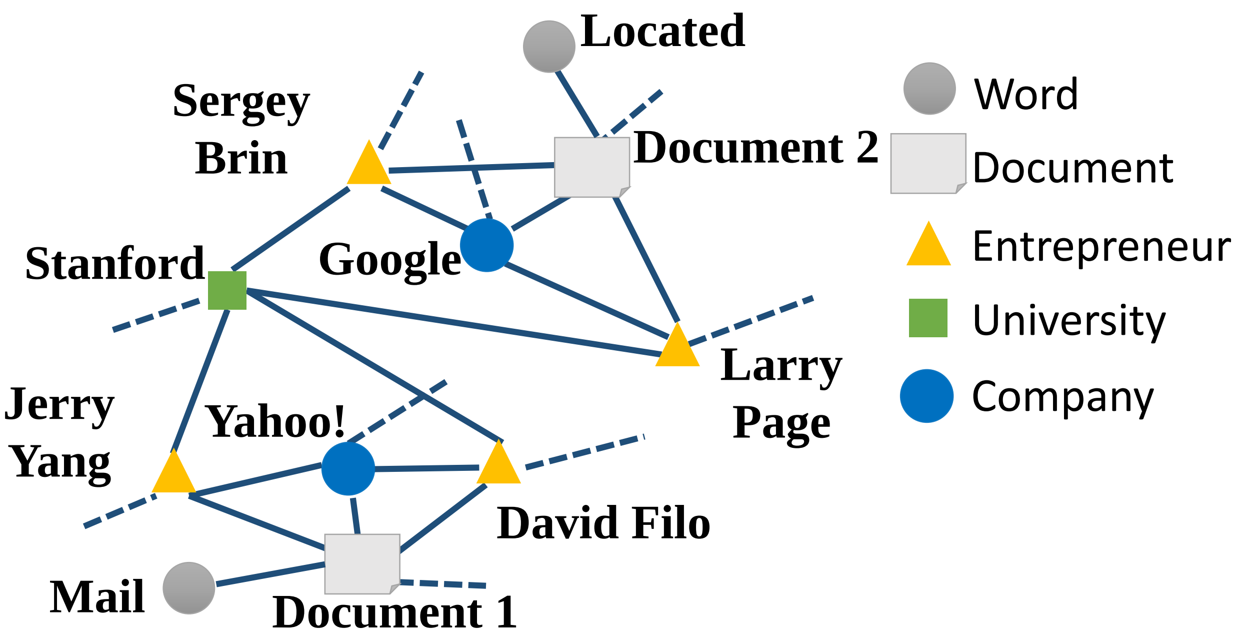
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

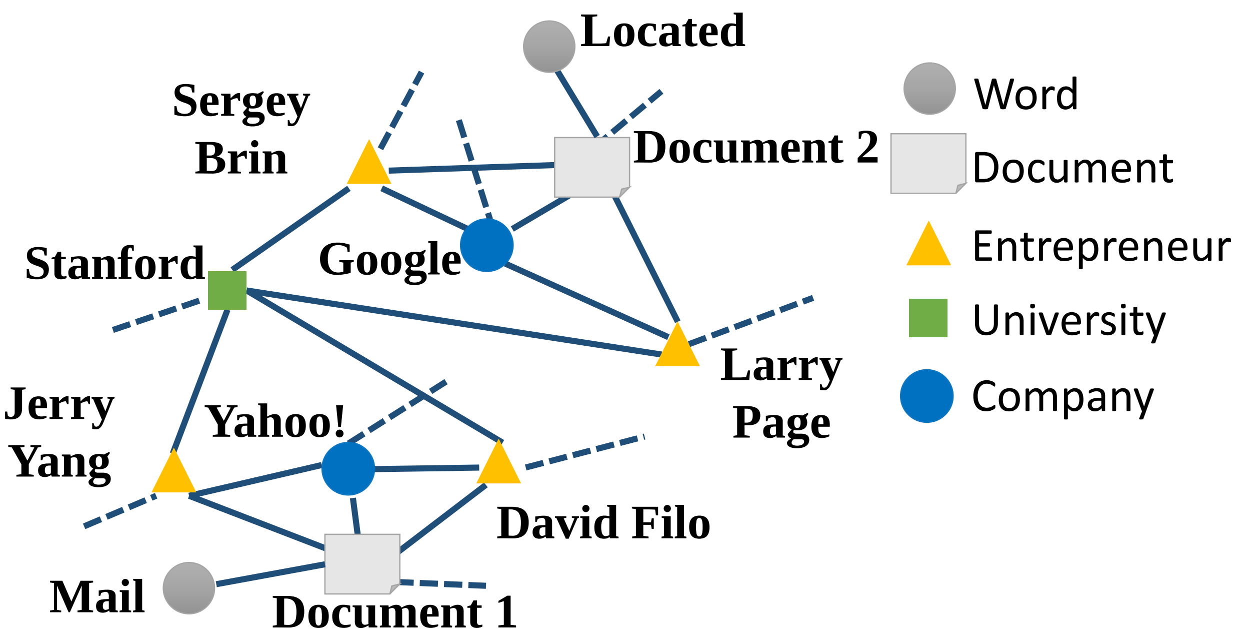


Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!

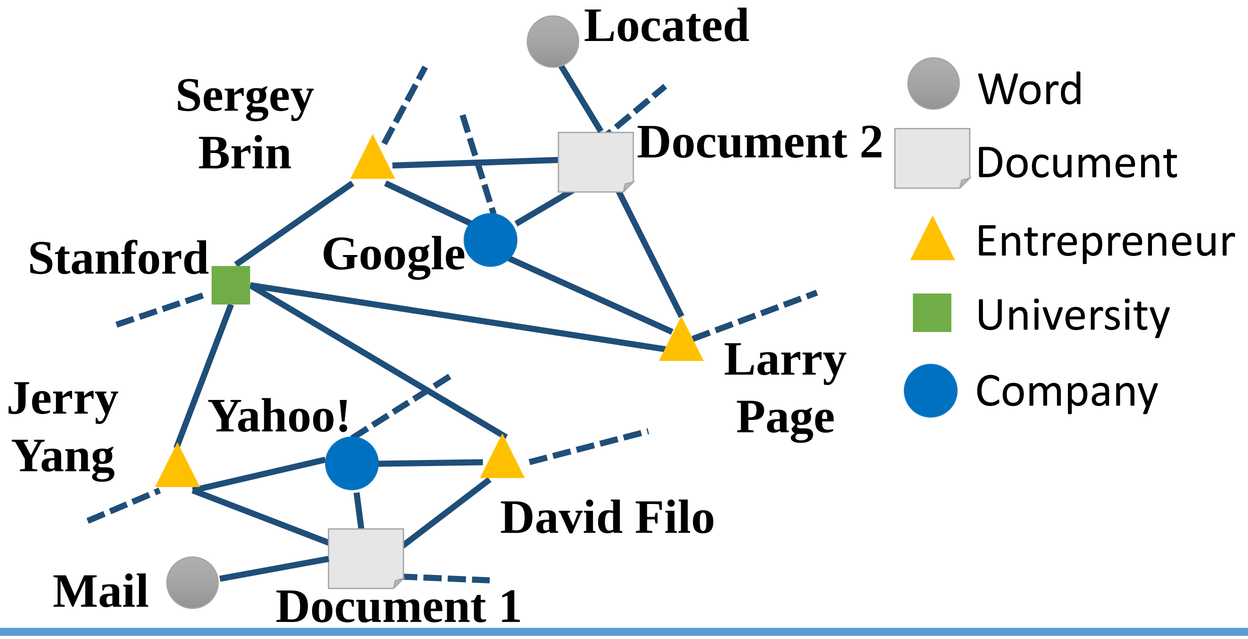


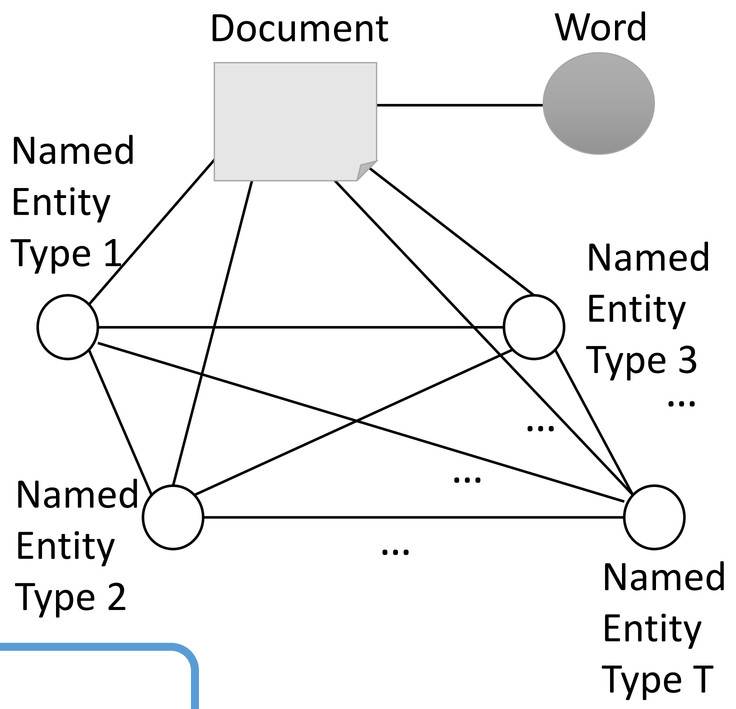
Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!





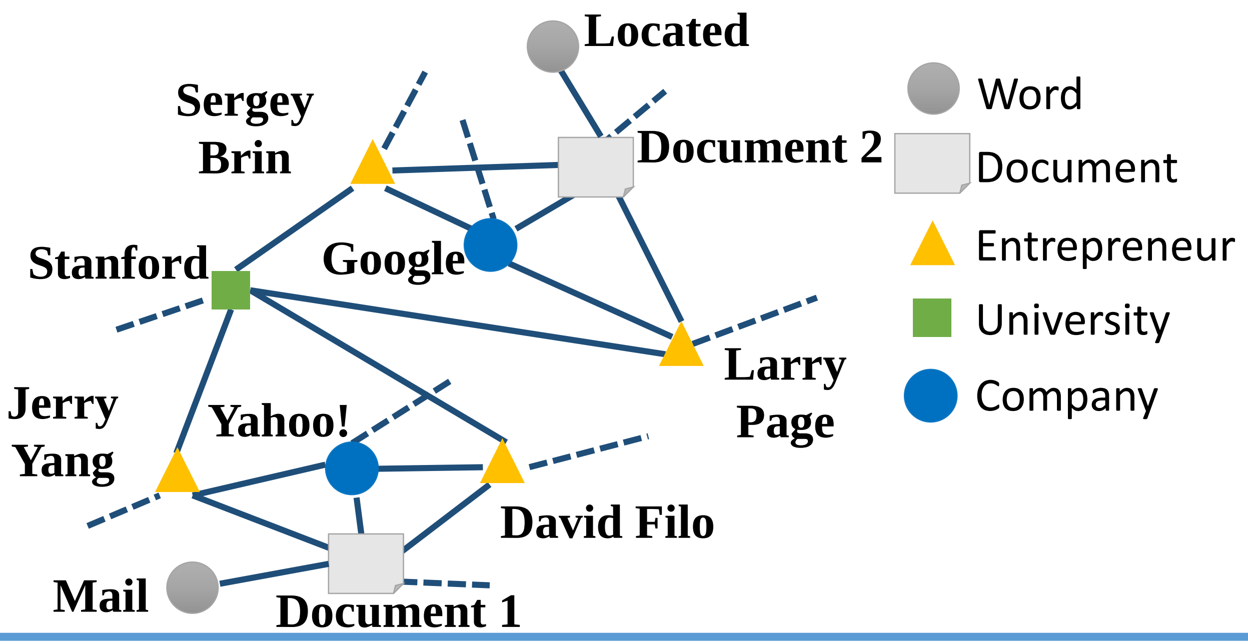
Network schema: High-level description of a network.

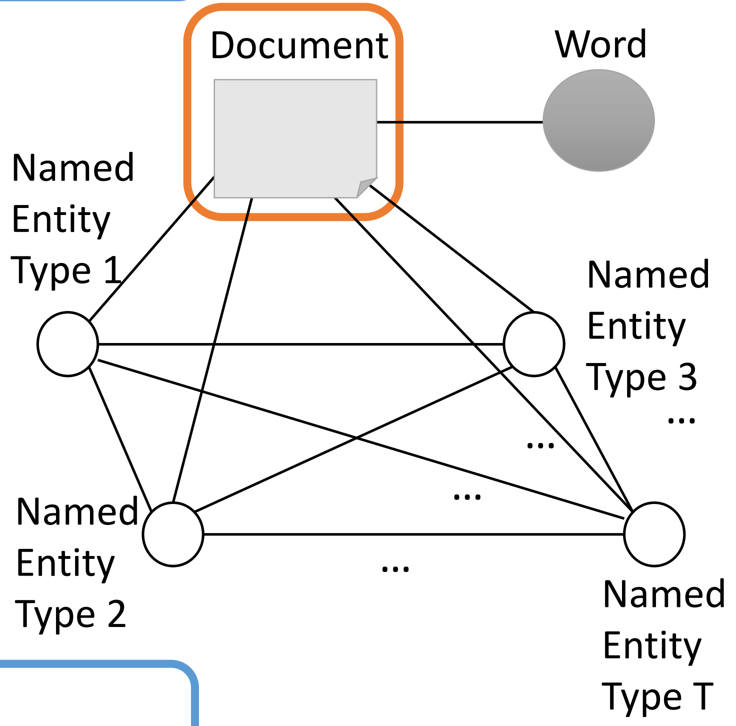
Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!





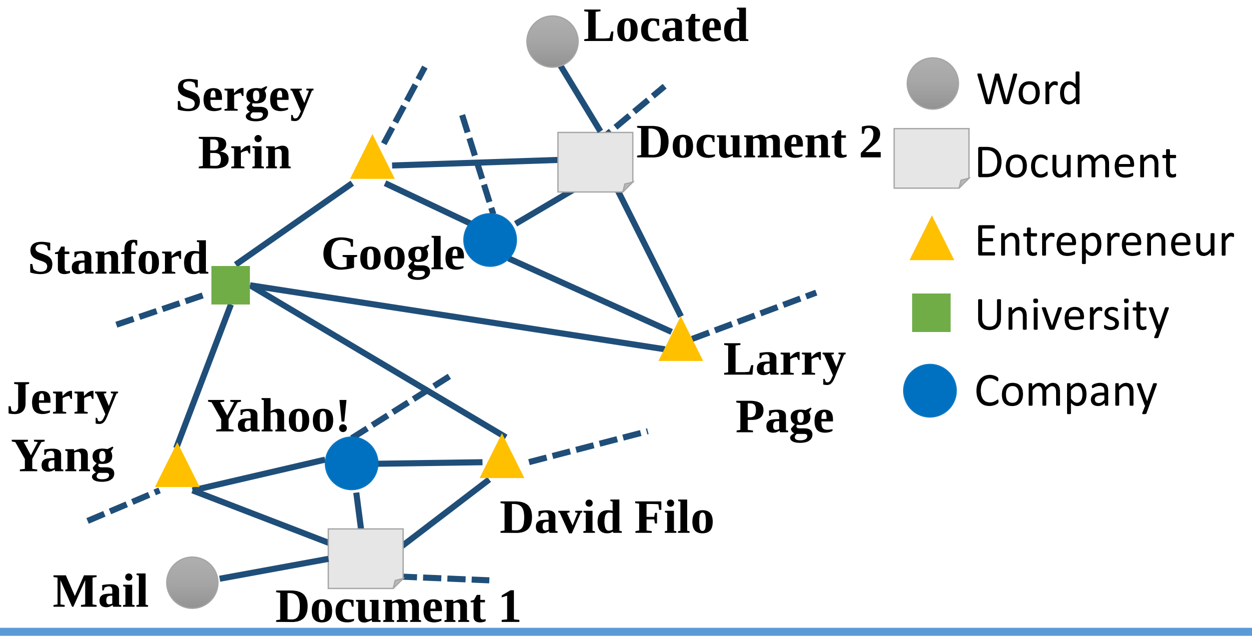
Network schema: High-level description of a network.

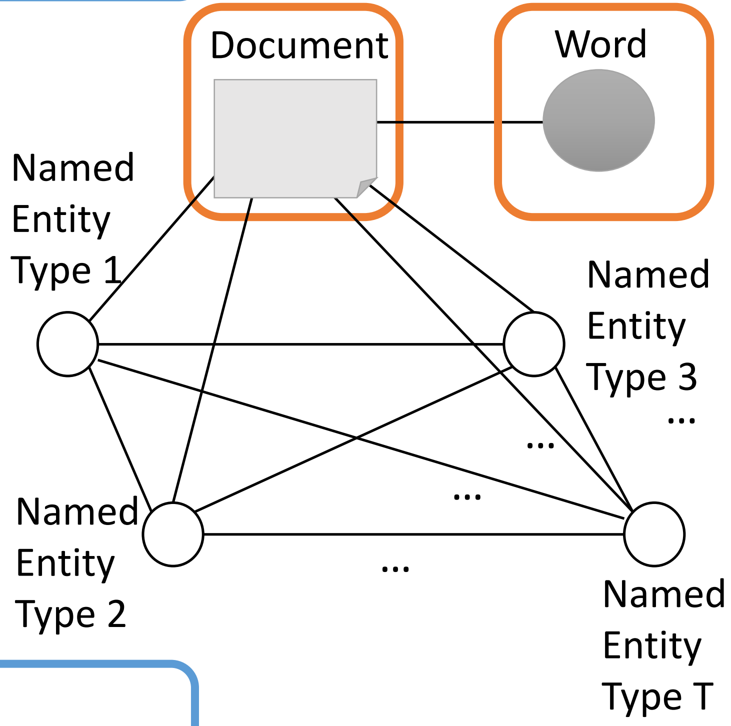
Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!





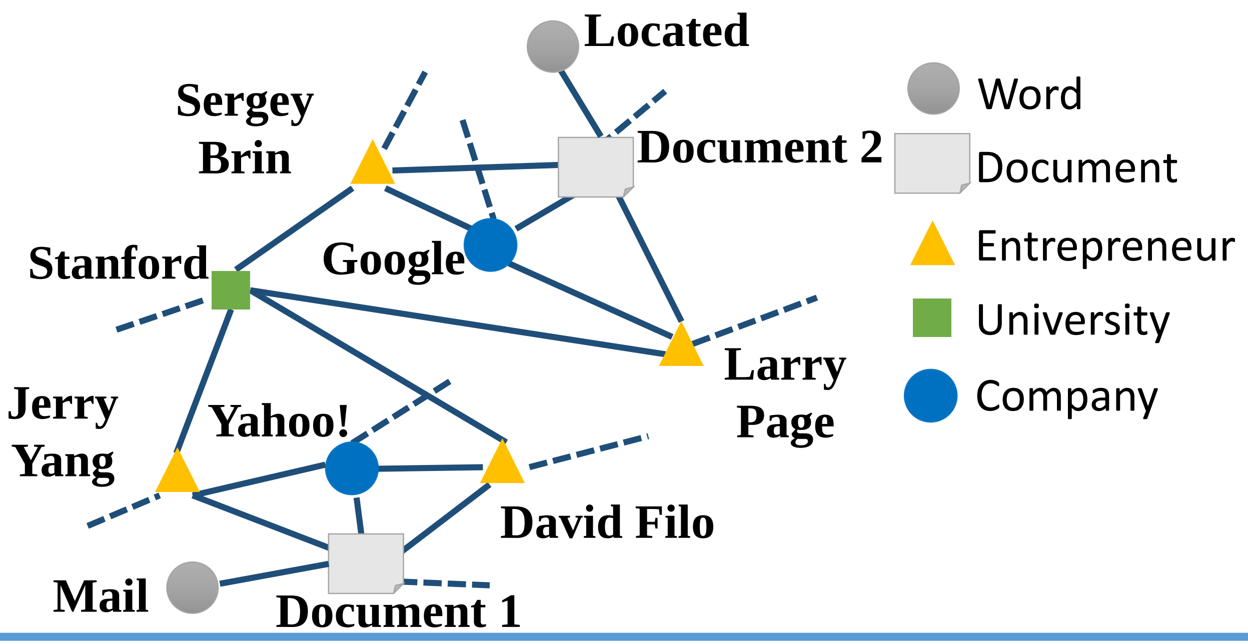
Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

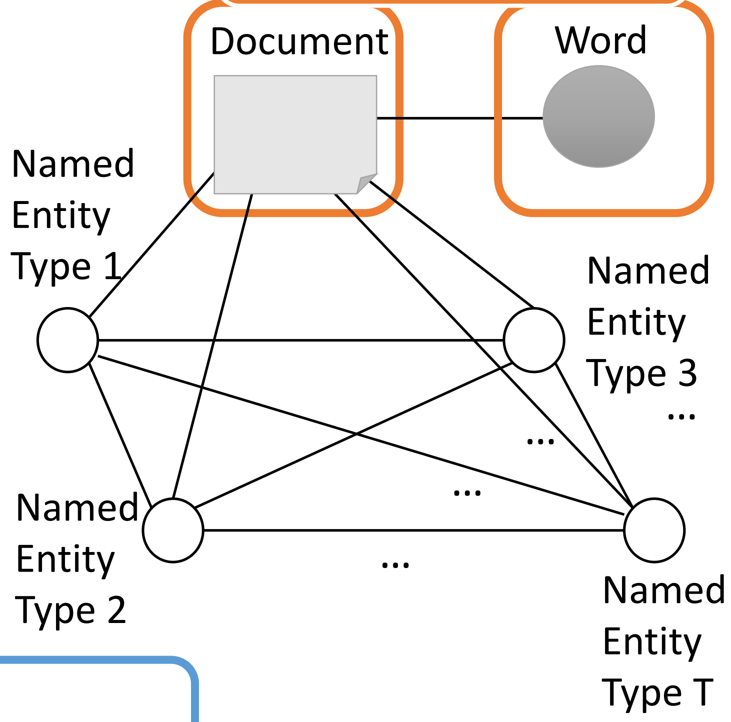
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!



Two entity types in document-based HIN.



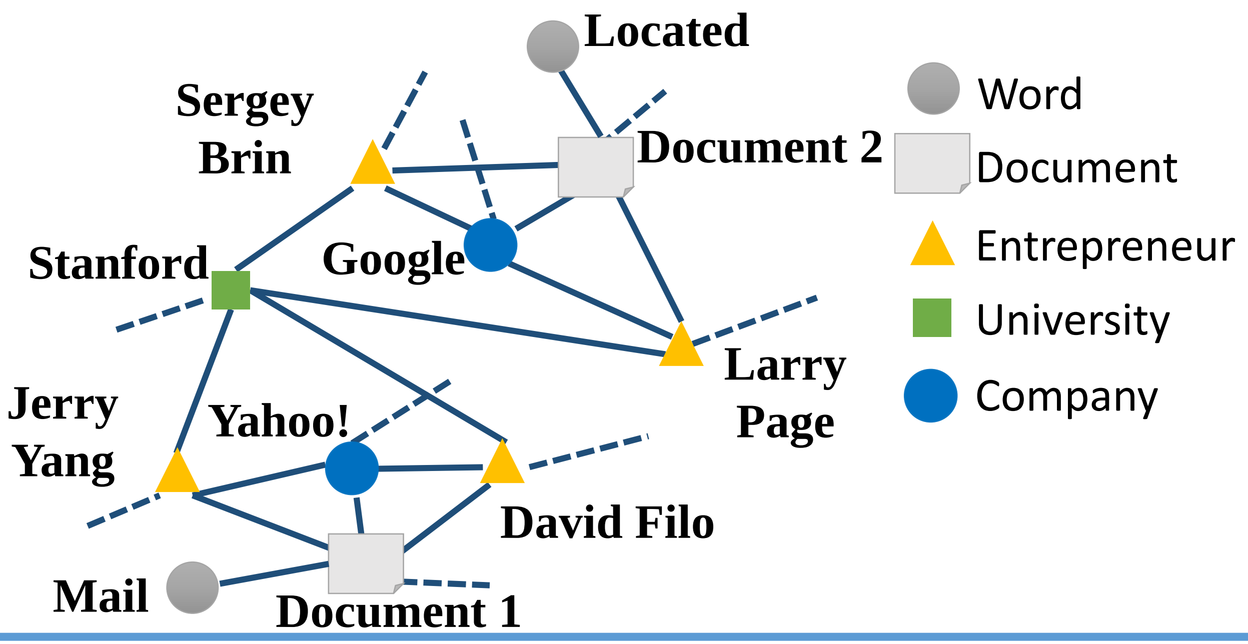
Network schema: High-level description of a network.

Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

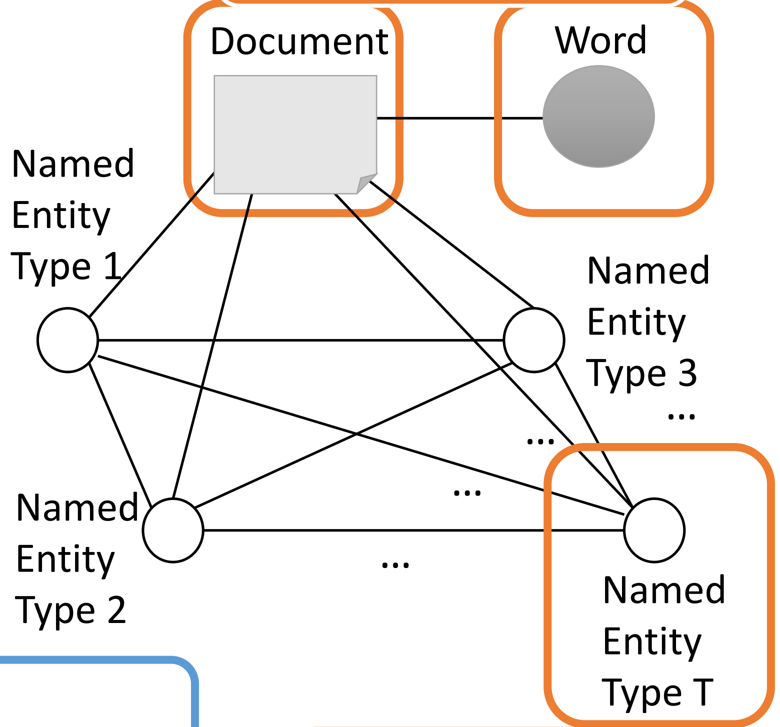
Document-based Heterogeneous Information Network (HIN)

HIN: Network with multiple object types and/or multiple link types.

A good way to model real world data!



Two entity types in document-based HIN.

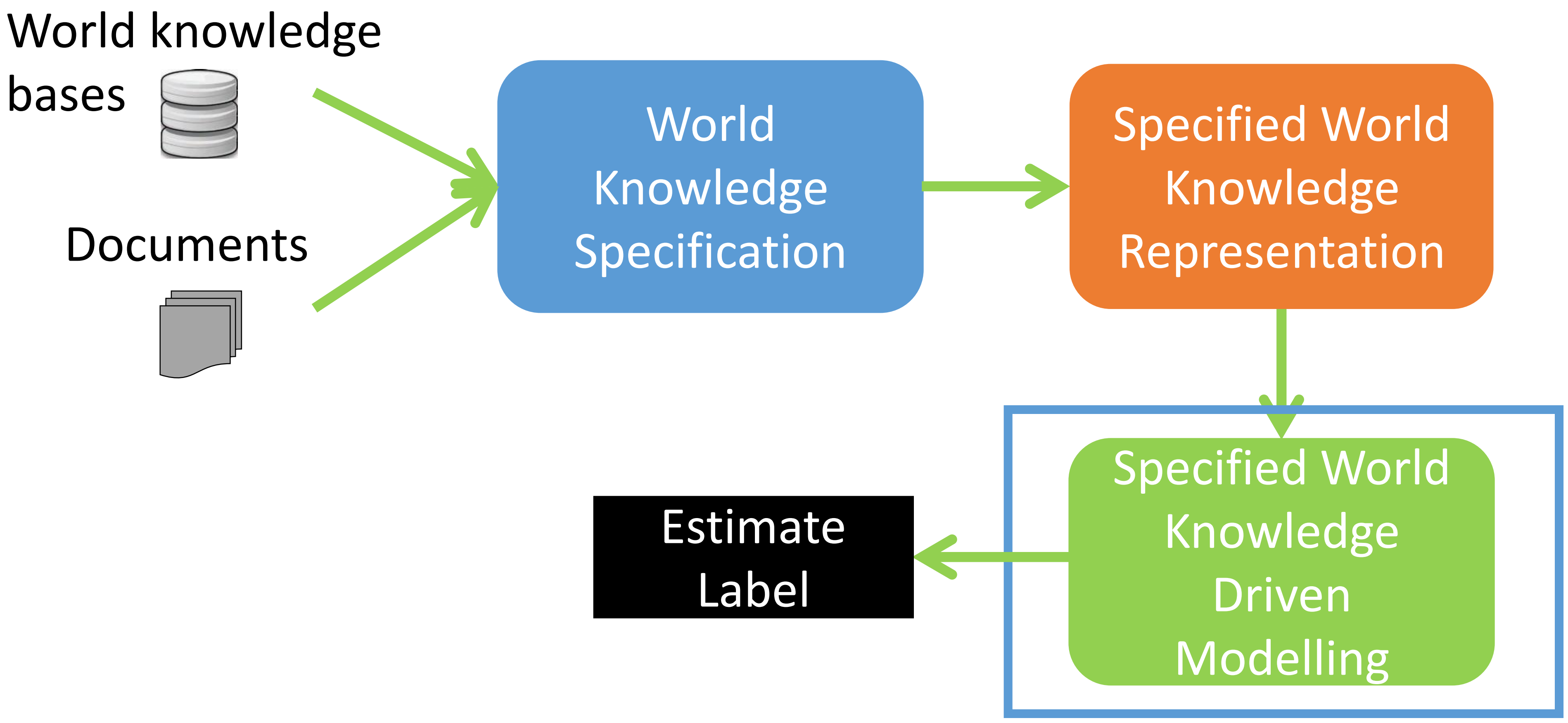


Network schema: High-level description of a network.

Represent the type of the name in text, e.g, person name.
NOT entity type (node type in HIN).

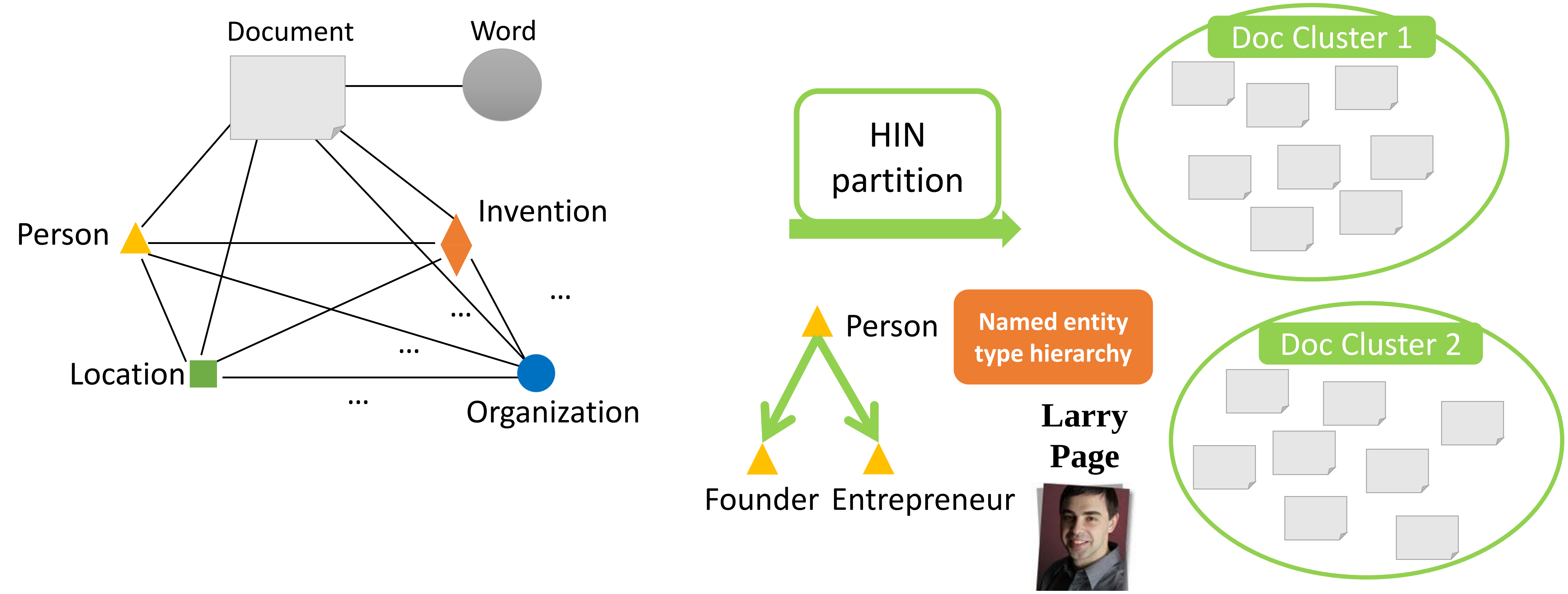
Y. Sun et al. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. PVLDB'11.

Text Clustering with World Knowledge

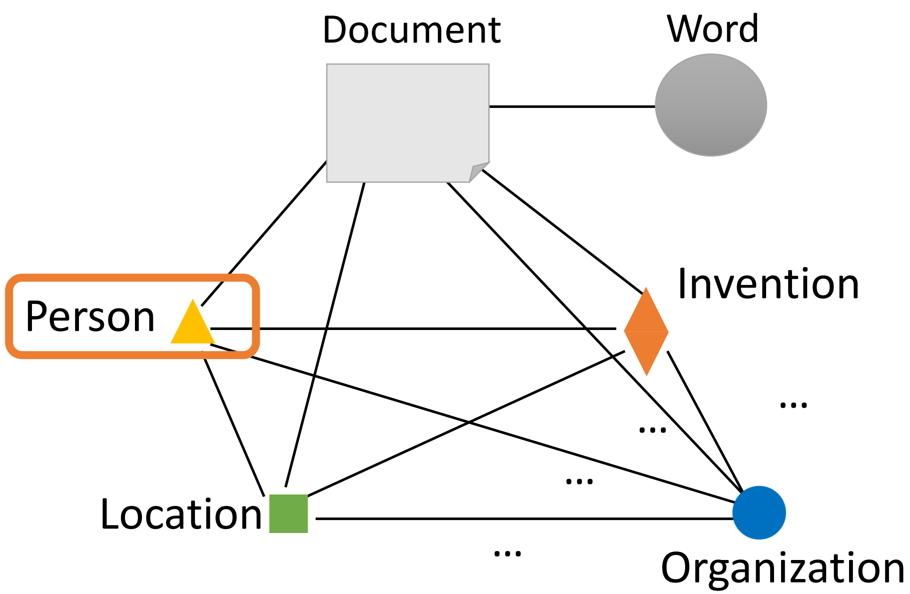


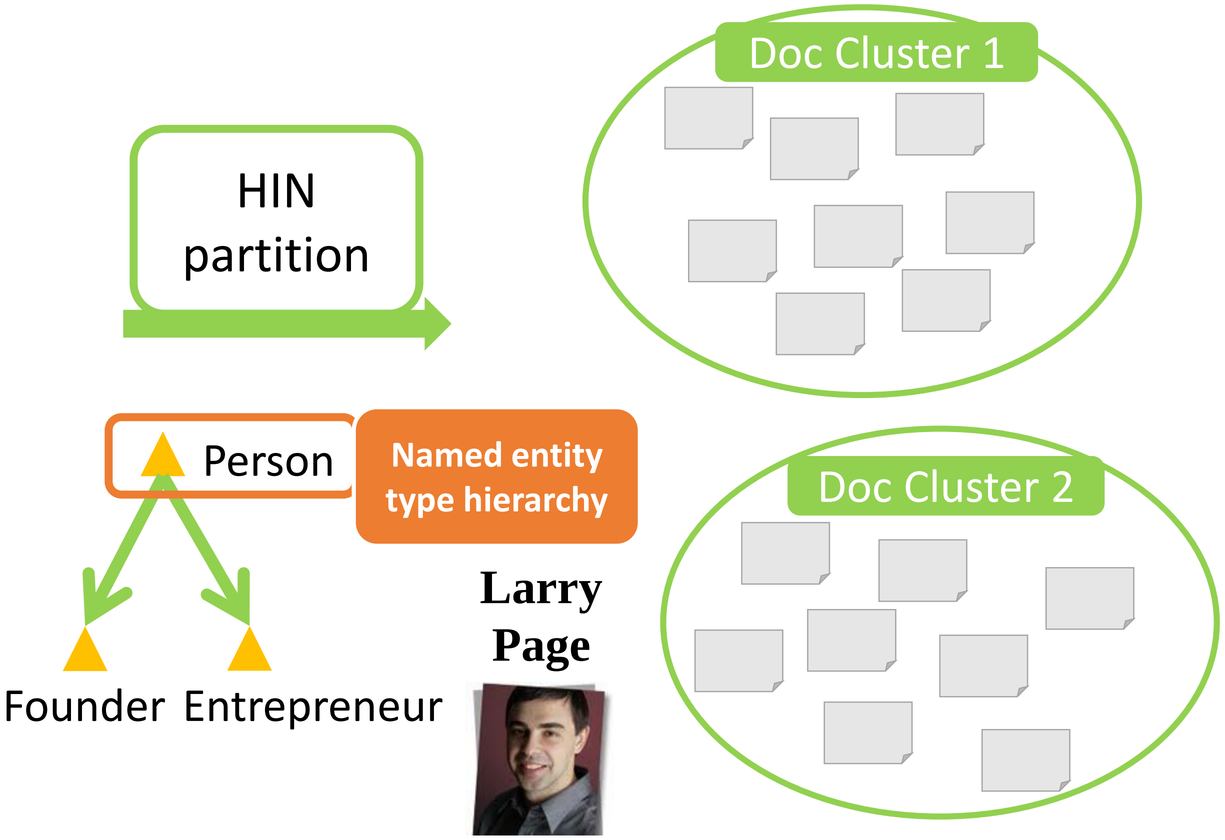
Constrained Clustering Modeling

Constrained Clustering Modeling



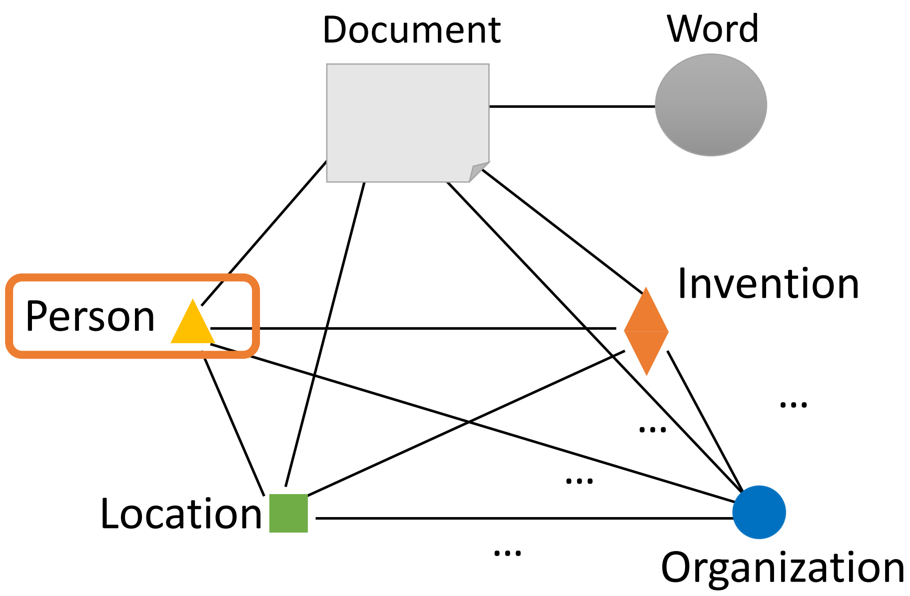
Constrained Clustering Modeling

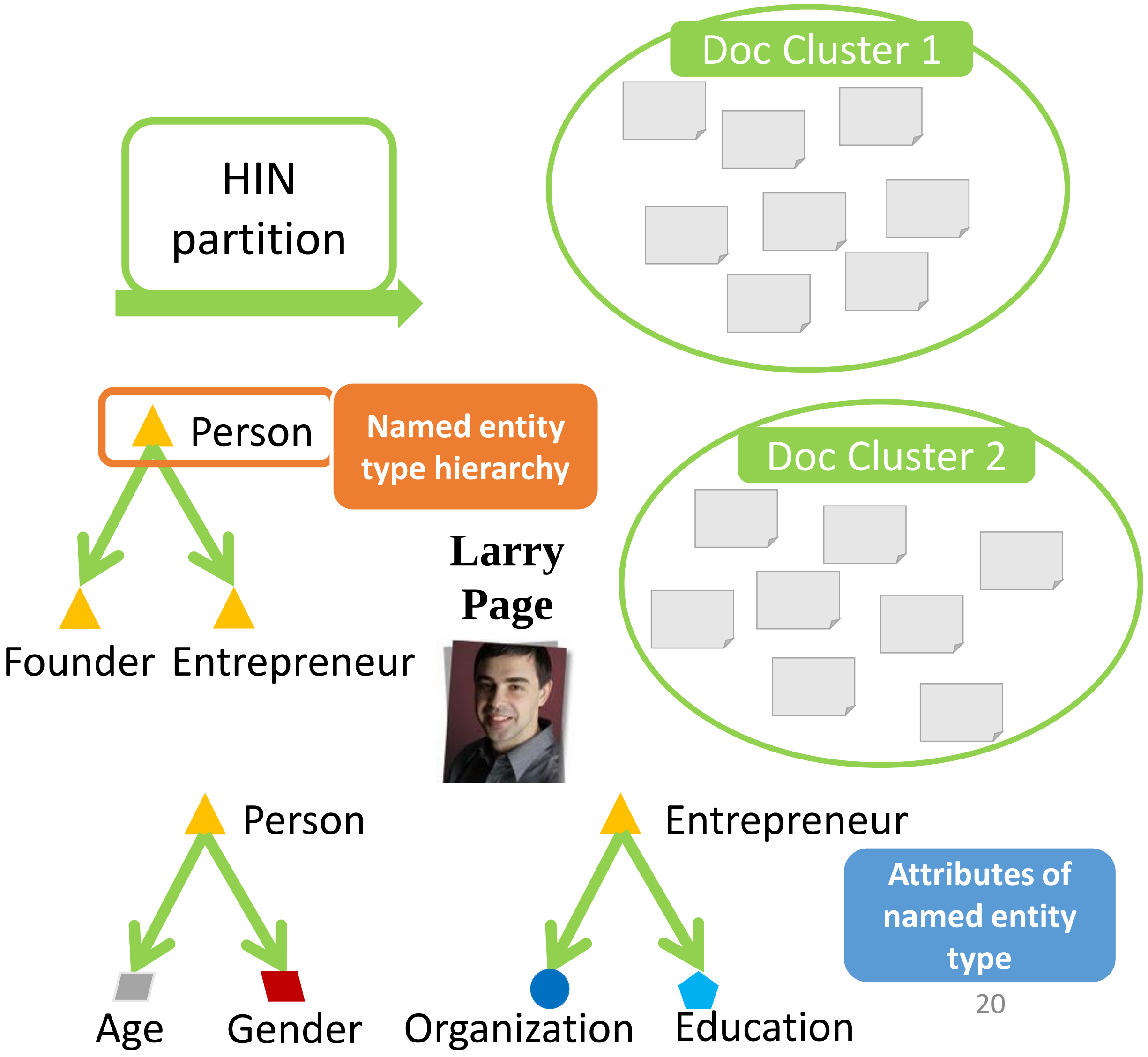




- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.

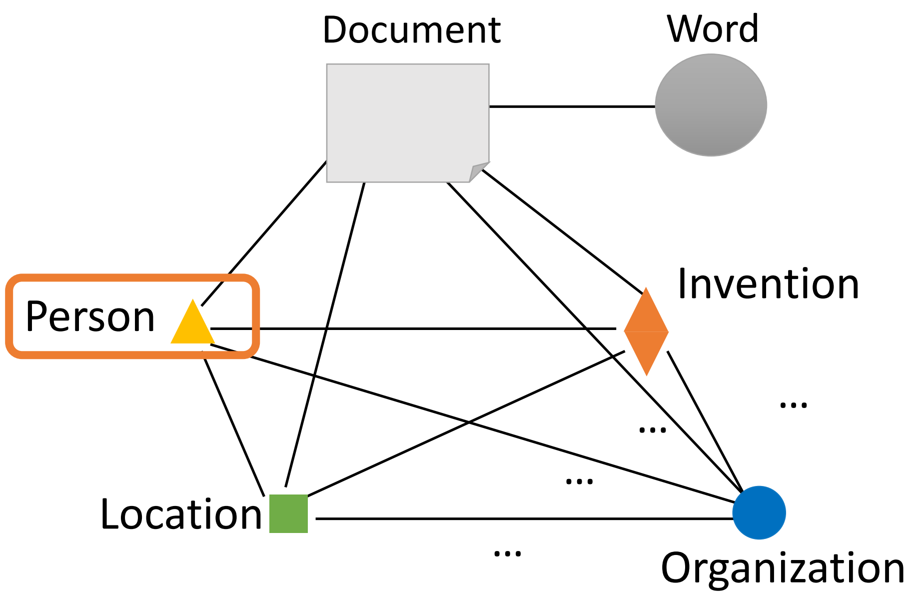
Constrained Clustering Modeling

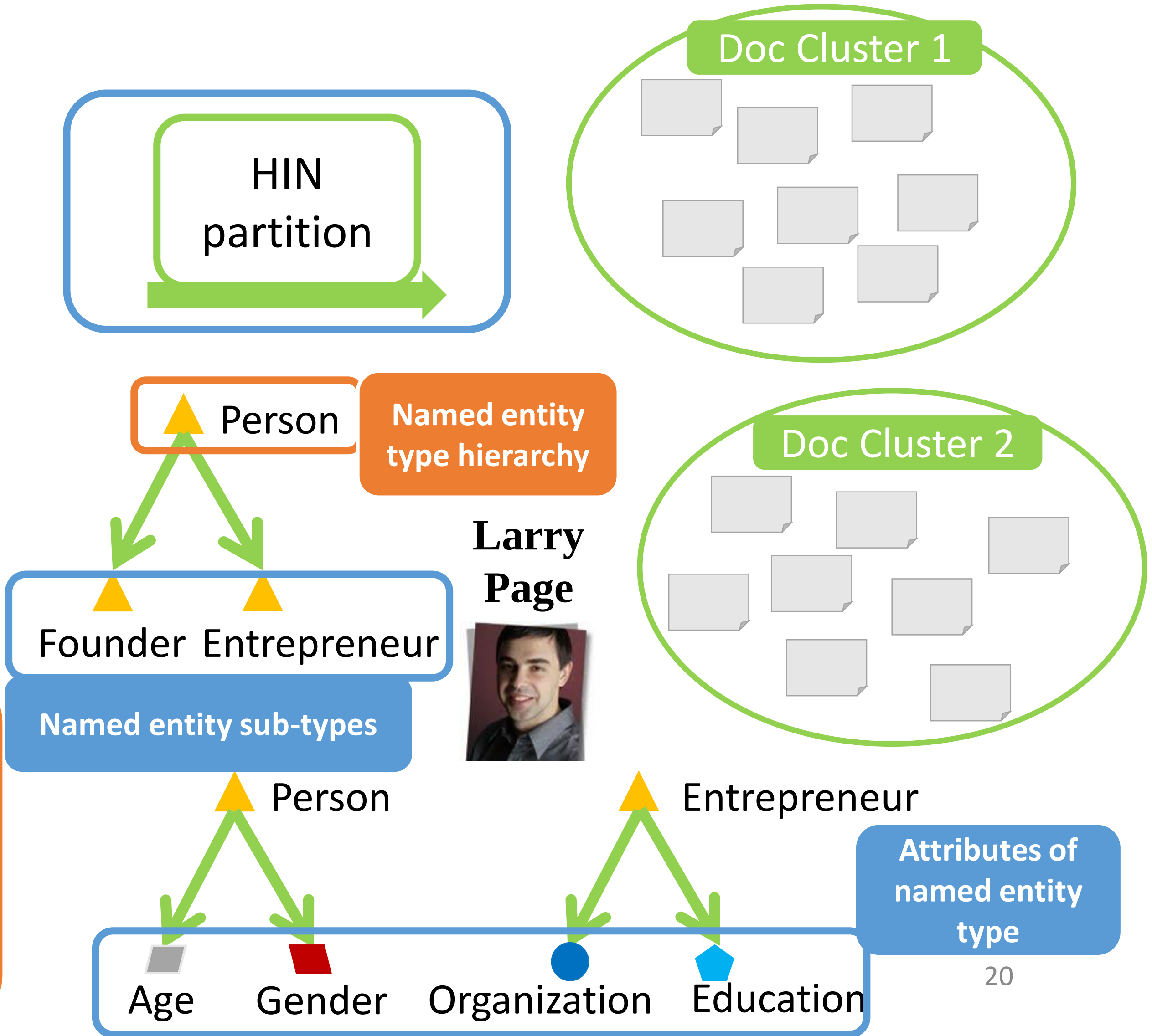




- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

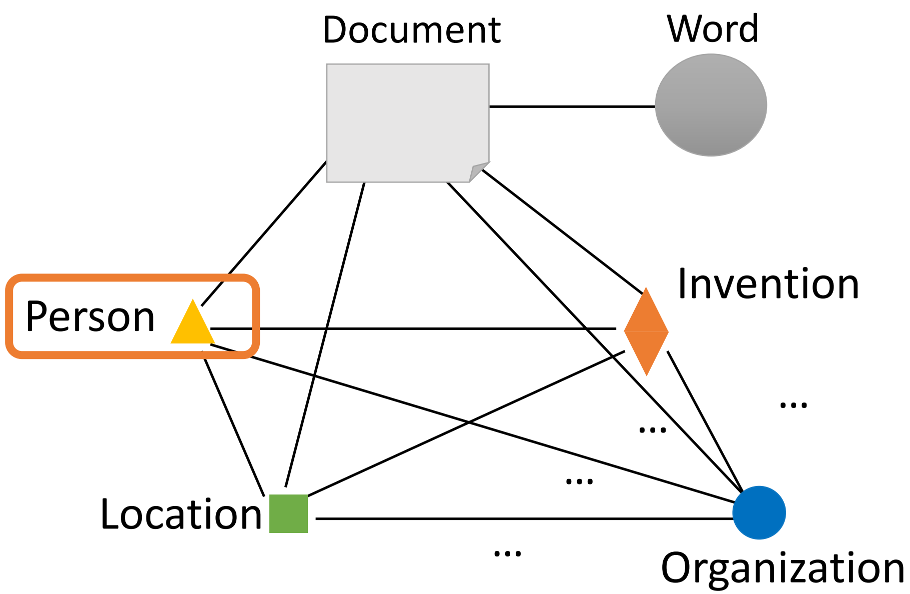
Constrained Clustering Modeling

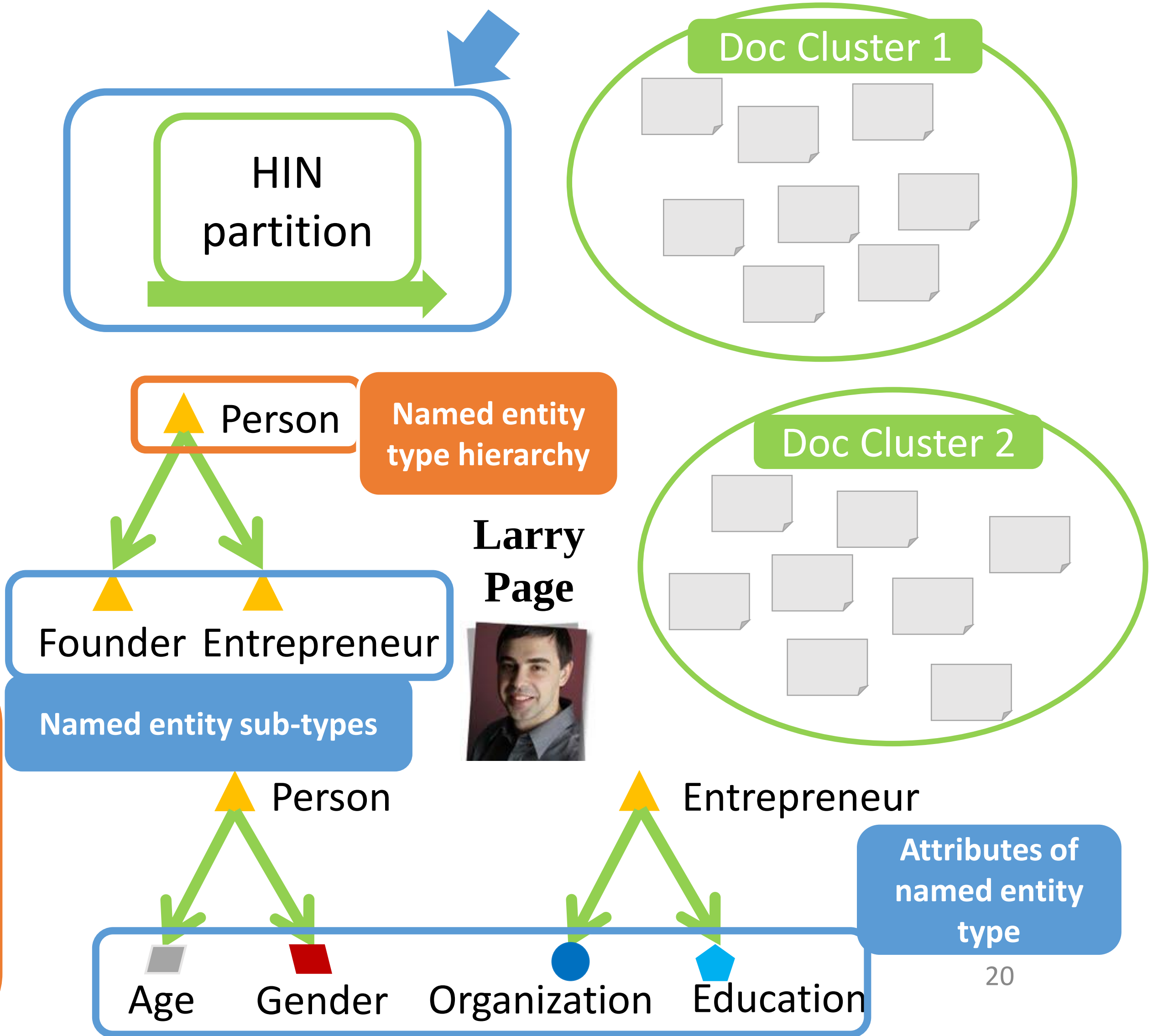




- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

Constrained Clustering Modeling





- Use the top level named entity types as the entity types in HIN.
  - have a relatively dense graph.
- Use named entity sub-types and attributes in HIN clustering model.
  - Useful to identify the topics or clusters of the documents.

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w})p(d_m|\hat{d}_{k_d})p(w_i|\hat{w}_{k_w})$$

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices

Constrained Clustering Modeling

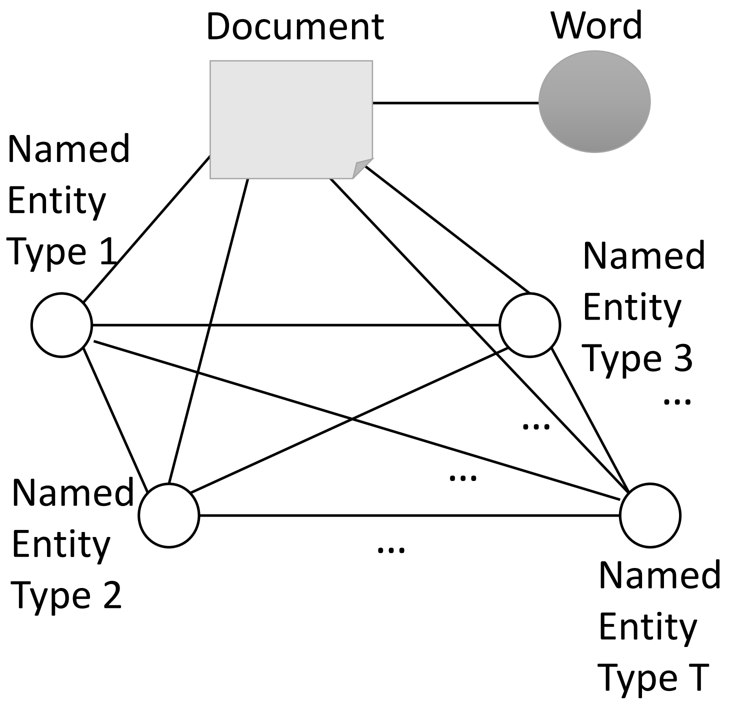
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



$$\begin{aligned}
J_{CHINC} &= D_{KL}\big(p(D,W) || q(D,W)\big) \\
&+ \sum_{t=1}^{T} D_{KL}\big(p(D,E^t) || q(D,E^t)\big) \\
&+ \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s) || q(E^t,E^s)\big) \\
&+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) \\
&+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)
\end{aligned}$$

Constrained Clustering Modeling

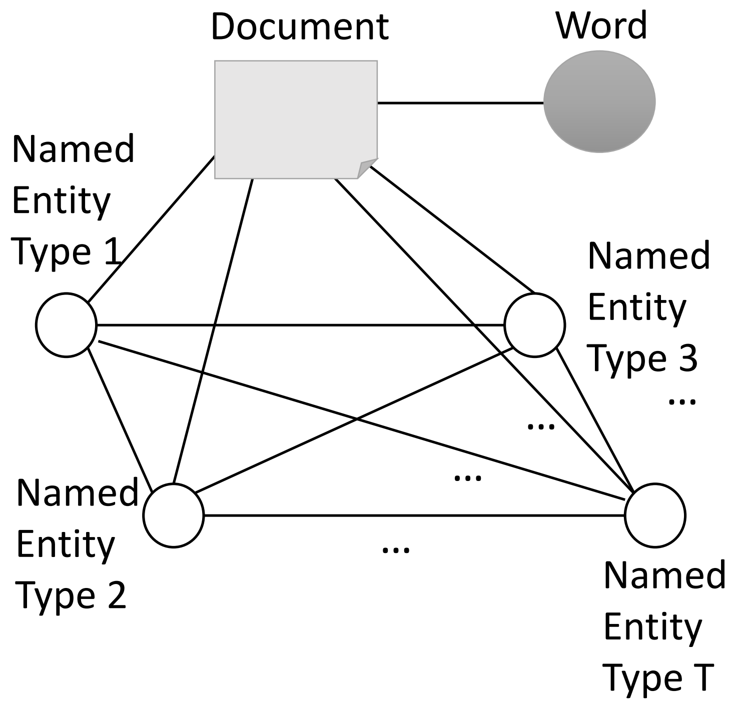
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{k_w})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$\begin{aligned} J_{CHINC} = & \, D_{KL}\big(p(D,W) \| q(D,W)\big) \\ & + \sum_{t=1}^{T} D_{KL}\big(p(D,E^t) \| q(D,E^t)\big) \\ & + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s) \| q(E^t,E^s)\big) \\ & + \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) \\ & + \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right) \end{aligned}$$

Constrained Clustering Modeling

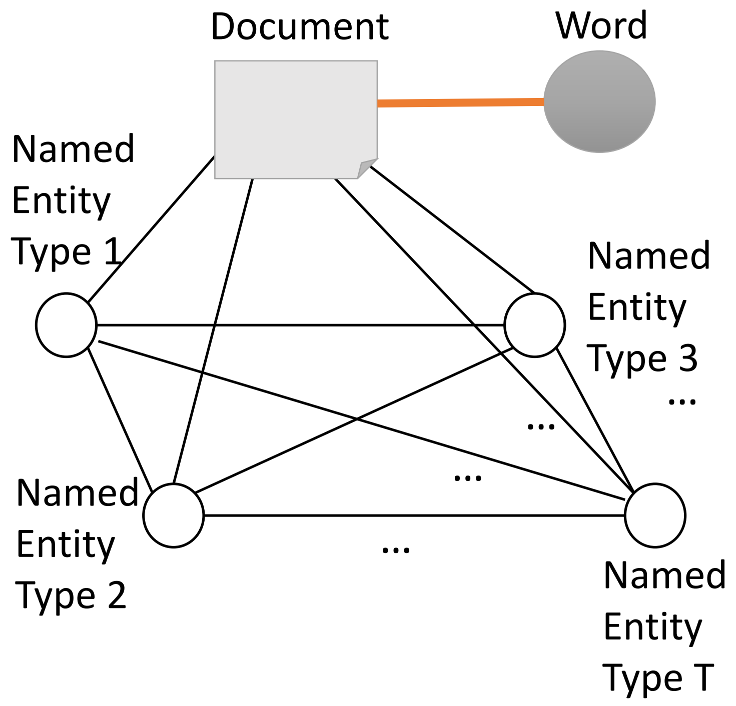
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$\boxed{J_{CHINC}} = \boxed{D_{KL}\big(p(D,W) || q(D,W)\big)}$$
$$+ \sum_{t=1}^{T} D_{KL}\big(p(D,E^t) || q(D,E^t)\big)$$
$$+ \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s) || q(E^t,E^s)\big)$$
$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$
$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Constrained Clustering Modeling

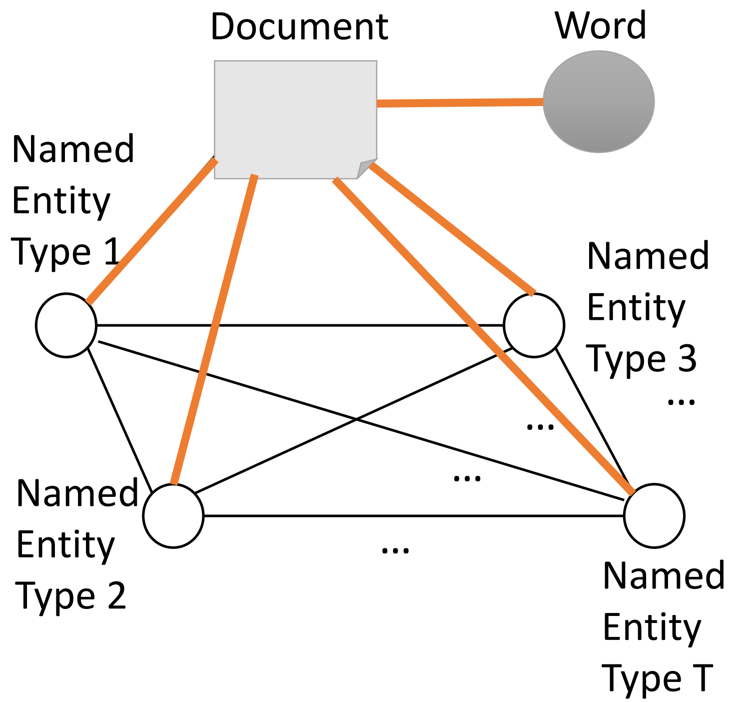
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w})\, p(d_m|\hat{d}_{k_d})\, p(w_i|\hat{w}_{k_w})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$\begin{aligned} J_{CHINC} = {} & D_{KL}(p(D,W)||q(D,W)) \\ & + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) \\ & + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s)) \\ & + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) \\ & + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right) \end{aligned}$$

Constrained Clustering Modeling

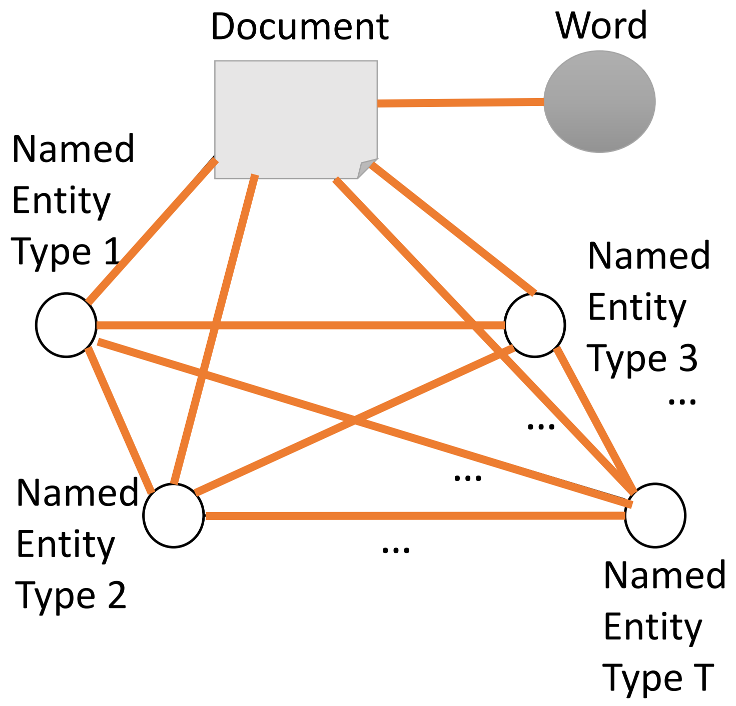
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



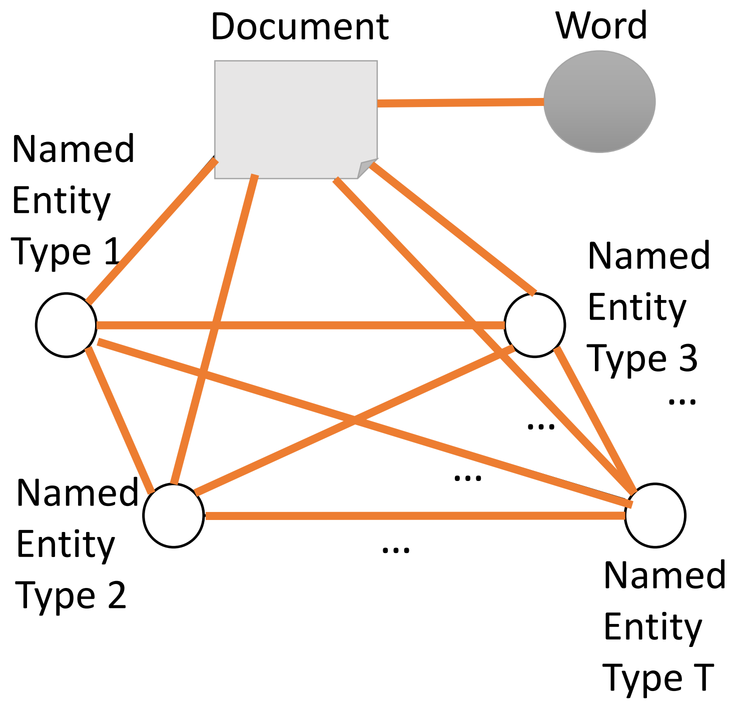
Minimize

$$\boxed{J_{CHINC}} = \boxed{D_{KL}\big(p(D,W) \| q(D,W)\big)}$$
$$+ \boxed{\sum_{t=1}^{T} D_{KL}\big(p(D,E^t) \| q(D,E^t)\big)}$$
$$+ \boxed{\sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s) \| q(E^t,E^s)\big)}$$
$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$
$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\underbrace{q(d_m, w_i)}_{\text{Joint probability } p(d_m, w_i) \text{ approximation}} = p(\underbrace{\hat{d}_{k_d}}_{\text{Cluster indicators}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\underbrace{k_w}_{\text{Cluster indices}}})$$



Minimize

$$J_{CHINC} = D_{KL}\big(p(D,W) || q(D,W)\big) + \sum_{t=1}^{T} D_{KL}\big(p(D,E^t) || q(D,E^t)\big) + \sum_{t=1}^{T} \sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s) || q(E^t,E^s)\big)$$

$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

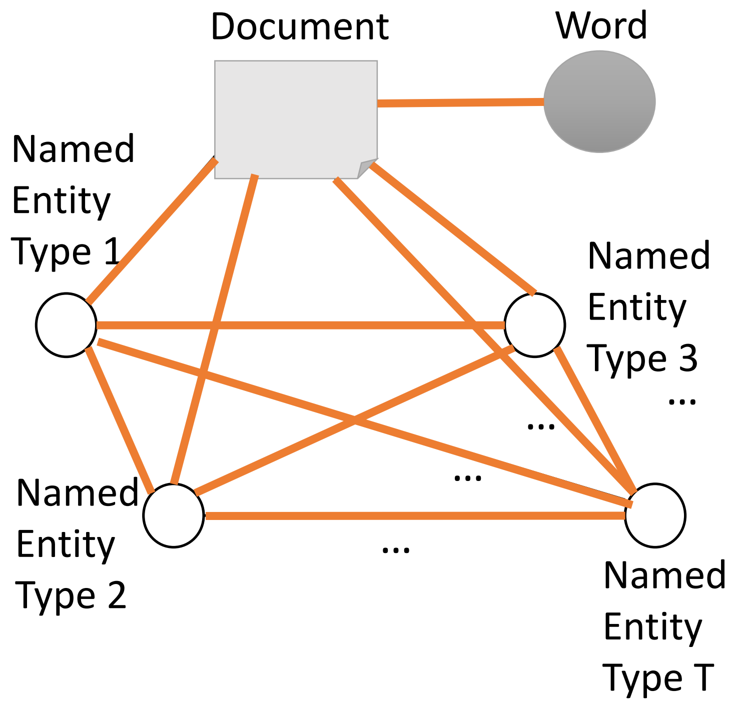
$$+ \sum_{t=1}^{T} \sum_{e_{i_1}^t=1}^{V_t} \sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Minimizing KL means **approximation** q should be **similar to original** p.

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\underbrace{q(d_m, w_i)}_{\text{Joint probability } p(d_m, w_i) \text{ approximation}} = p(\underbrace{\hat{d}_{k_d}}_{\text{Cluster indicators}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\underbrace{k_w}_{\text{Cluster indices}}})$$



Minimize

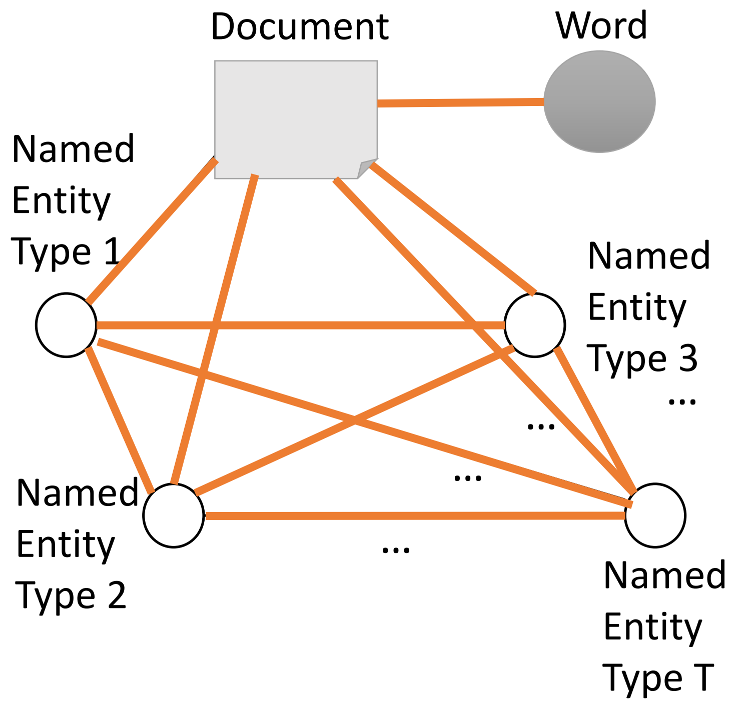
$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W)) + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$
$$w_M D_{KL}\left(p(D|e_{i_1}^t)||p(D|e_{i_2}^t)\right) \cdot I_{l_{e_{i_1}^t} \neq l_{e_{i_2}^t}}$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Minimizing KL means **approximation** q should be **similar to original** p.

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\underbrace{q(d_m, w_i)}_{\text{Joint probability } p(d_m, w_i) \text{ approximation}} = p(\underbrace{\hat{d}_{k_d}}_{\text{Cluster indicators}}, \hat{w}_{k_w})\, p(d_m|\hat{d}_{k_d})\, p(w_i|\hat{w}_{\underbrace{k_w}_{\text{Cluster indices}}})$$



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W)) + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$

$$+ \sum_{t=1}^{T}\sum_{e^t_{i_1}=1}^{V_t}\sum_{e^t_{i_2}\in M_{e^t_{i_1}}} \underbrace{V_M\left(e^t_{i_1}, e^t_{i_2} \in M_{e^t_{i_1}}\right)}_{w_M D_{KL}\left(p(D|e^t_{i_1})||p(D|e^t_{i_2})\right)\cdot I_{l_{e^t_{i_1}} \neq l_{e^t_{i_2}}}}$$

$$+ \sum_{t=1}^{T}\sum_{e^t_{i_1}=1}^{V_t}\sum_{e^t_{i_2}\in C_{e^t_{i_1}}} \underbrace{V_C\left(e^t_{i_1}, e^t_{i_2} \in C_{e^t_{i_1}}\right)}_{w_C\left(D^t_{max} - D_{KL}\left(p(D|e^t_{i_1})||p(D|e^t_{i_2})\right)\right)\cdot I_{l_{e^t_{i_1}} \neq l_{e^t_{i_2}}}}$$

Minimizing KL means **approximation** q should be **similar to original** p.

Constrained Clustering Modeling

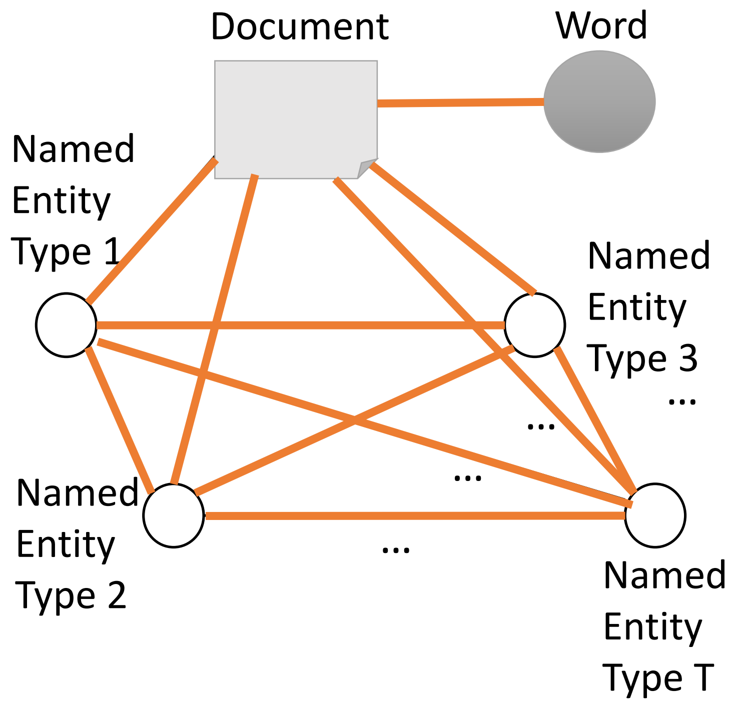
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w})\, p(d_m|\hat{d}_{k_d})\, p(w_i|\hat{w}_{k_w})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$\begin{aligned} J_{CHINC} = & D_{KL}(p(D,W)||q(D,W)) \\ & + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) \\ & + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s)) \\ & + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) \\ & + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right) \end{aligned}$$

Minimizing KL means **approximation** q should be **similar to original** p.

Must–link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are

Knowledge indirect supervision: fine-grained named entity sub-types or the attributes

Constrained Clustering Modeling

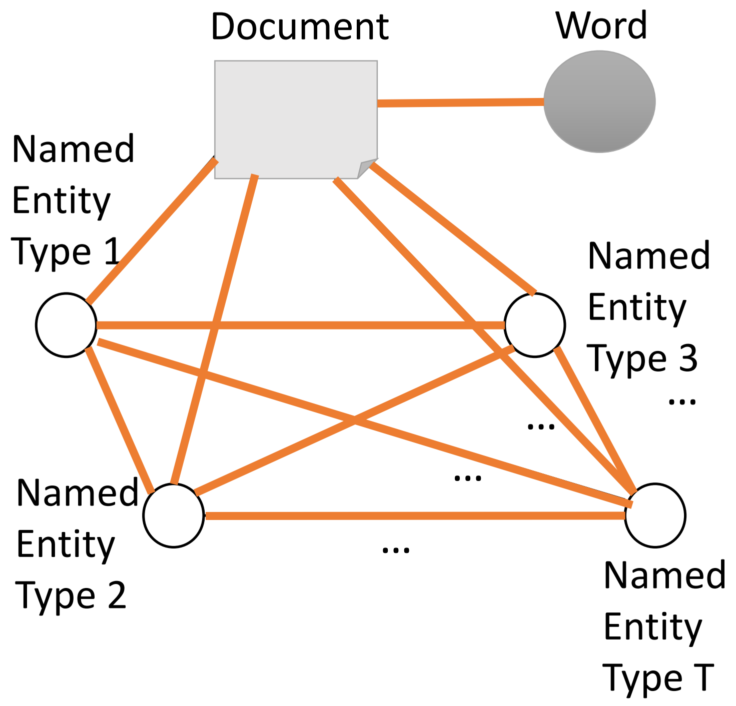
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD’03] and constrained ITCC [Y. Song et al. TKDE’13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w})p(d_m|\hat{d}_{k_d})p(w_i|\hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$\boxed{J_{CHINC}} = \boxed{D_{KL}(p(D,W)||q(D,W))} + \boxed{\sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t))} + \boxed{\sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))}$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} \boxed{V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)}$$

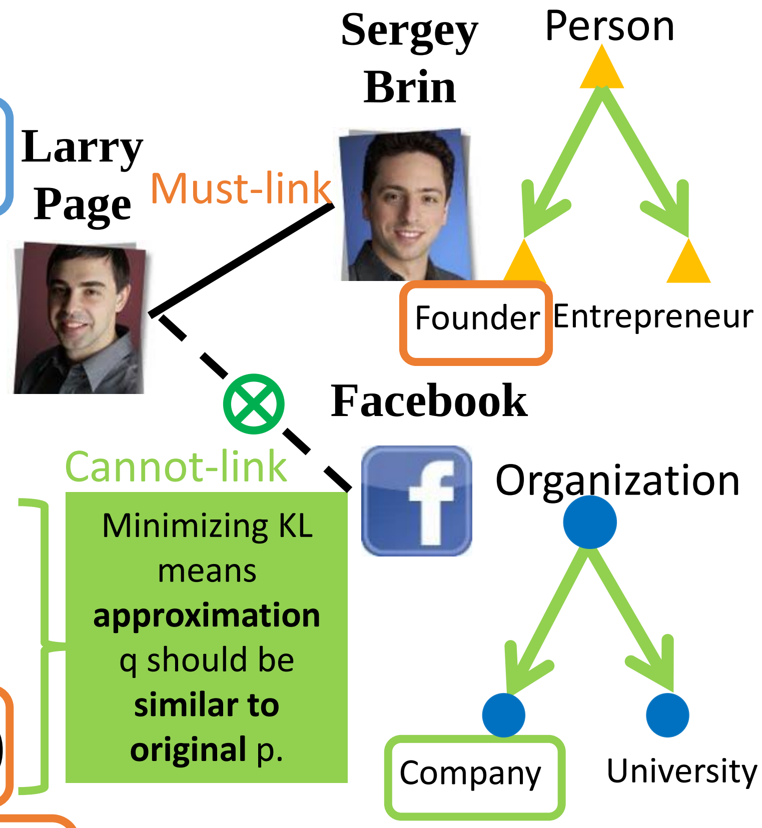
Must–link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} \boxed{V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)}$$

Cannot-link: if two labels are equal, consider how similar they are

Minimizing KL means **approximation** q should be **similar to original** p.

Knowledge indirect supervision: fine-grained named entity sub-types or the attributes



Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

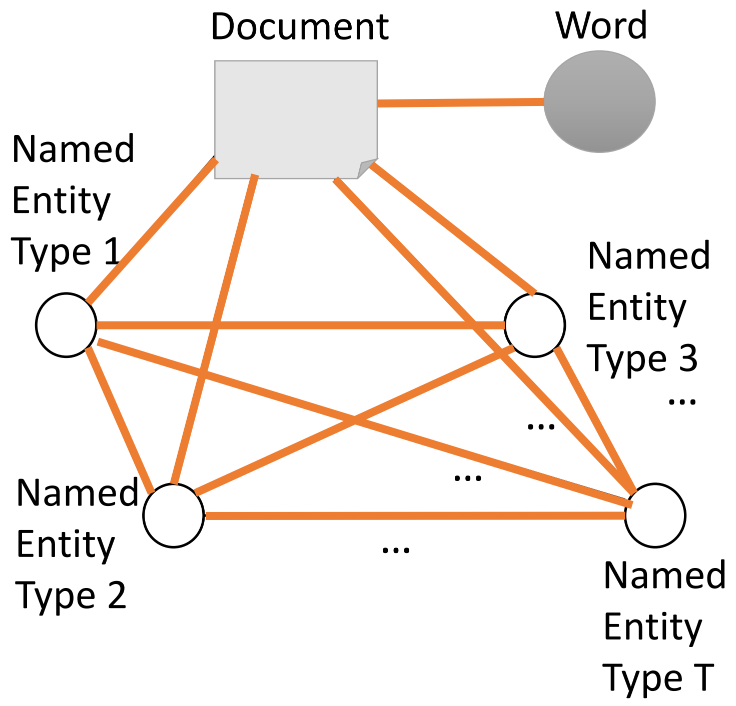
Globally optimizing the latent labels and the approximating function is intractable

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w})p(d_m|\hat{d}_{k_d})p(w_i|\hat{w}_{k_w})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W)) + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

Must-link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Cannot-link: if two labels are equal, consider how similar they are

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \dots, T$, Set maxIter and maxδ.

while iter $<$ maxIter and $\delta >$ maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1,…,T **do**

E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_j^s, e_i^t)$.

end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update $q(d_m, w_i)$.

Compute cost change δ.

end while

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

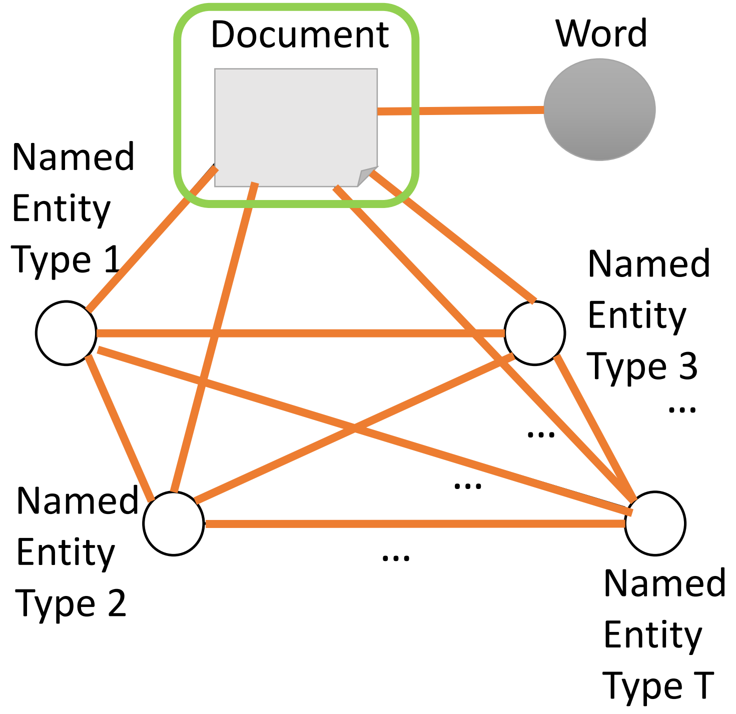
Globally optimizing the latent labels and the approximating function is intractable

$$q(d_m, w_i) = p(\hat{d}_{k_d}, \hat{w}_{k_w})\, p(d_m|\hat{d}_{k_d})\, p(w_i|\hat{w}_{k_w})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$J_{CHINC} = D_{KL}\big(p(D,W)||q(D,W)\big) + \sum_{t=1}^{T} D_{KL}\big(p(D,E^t)||q(D,E^t)\big) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}\big(p(E^t,E^s)||q(E^t,E^s)\big) + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right) + \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Must-link: if two labels are not equal, consider how dissimilar they are

Cannot-link: if two labels are equal, consider how similar they are

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \dots, T$, Set maxIter and maxδ.

while iter < maxIter and δ > maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1,…,T **do**
E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_j^s, e_i^t)$.
end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update $q(d_m, w_i)$.

Compute cost change δ.
end while

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

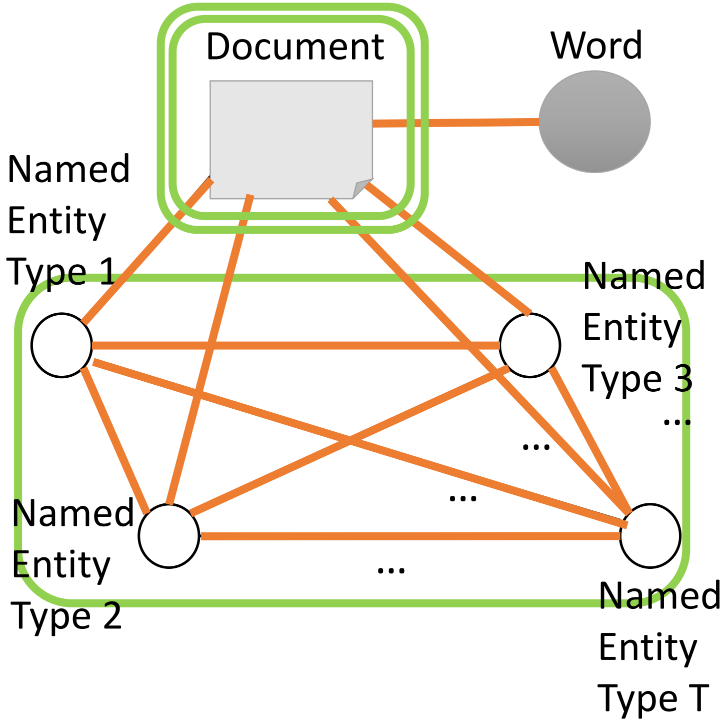
Globally optimizing the latent labels and the approximating function is intractable

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W)) + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

Must–link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Cannot-link: if two labels are equal, consider how similar they are

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \ldots, T$, Set maxIter and maxδ.

while iter < maxIter and δ > maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1,...,T **do**

E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_i^s, e_i^t)$.

end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update $q(d_m, w_i)$.

Compute cost change δ.

end while

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

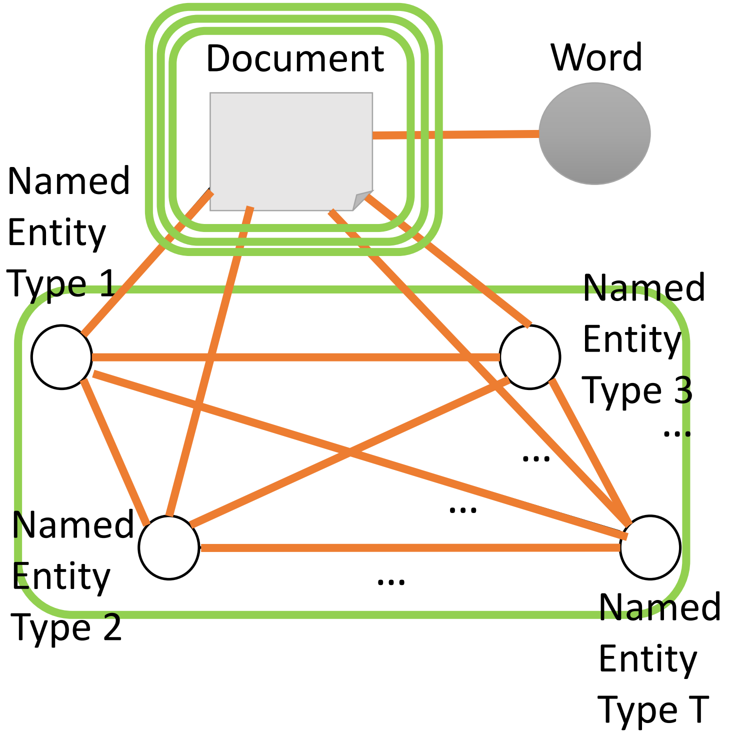
Globally optimizing the latent labels and the approximating function is intractable

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w})p(d_m|\hat{d}_{k_d})p(w_i|\hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W))$$
$$+ \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t))$$
$$+ \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

Must–link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Cannot-link: if two labels are equal, consider how similar they are

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \dots, T$, Set maxIter and maxδ.

while iter < maxIter and δ > maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1,...,T **do**

E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_i^s, e_i^t)$.

end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update$q(d_m, w_i)$.

Compute cost change δ.

end while

Constrained Clustering Modeling

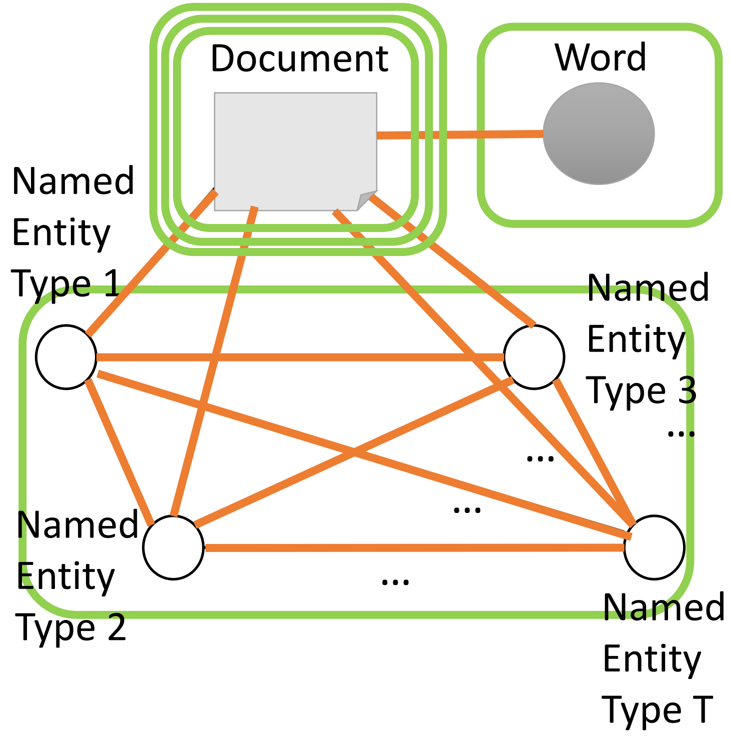
Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w})p(d_m|\hat{d}_{k_d})p(w_i|\hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W))$$
$$+ \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t))$$
$$+ \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

Must-link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Cannot-link: if two labels are equal, consider how similar they are

Globally optimizing the latent labels and the approximating function is intractable

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \ldots, T$, Set maxIter and maxδ.

while iter < maxIter and δ > maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1.....T **do**

E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_i^s, e_i^t)$.

end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update$q(d_m, w_i)$.

Compute cost change δ.

end while

Constrained Clustering Modeling

Motivation: The framework of information-theoretic co-clustering (ITCC) [I. S. Dhillon et al. KDD'03] and constrained ITCC [Y. Song et al. TKDE'13].

Globally optimizing the latent labels and the approximating function is intractable

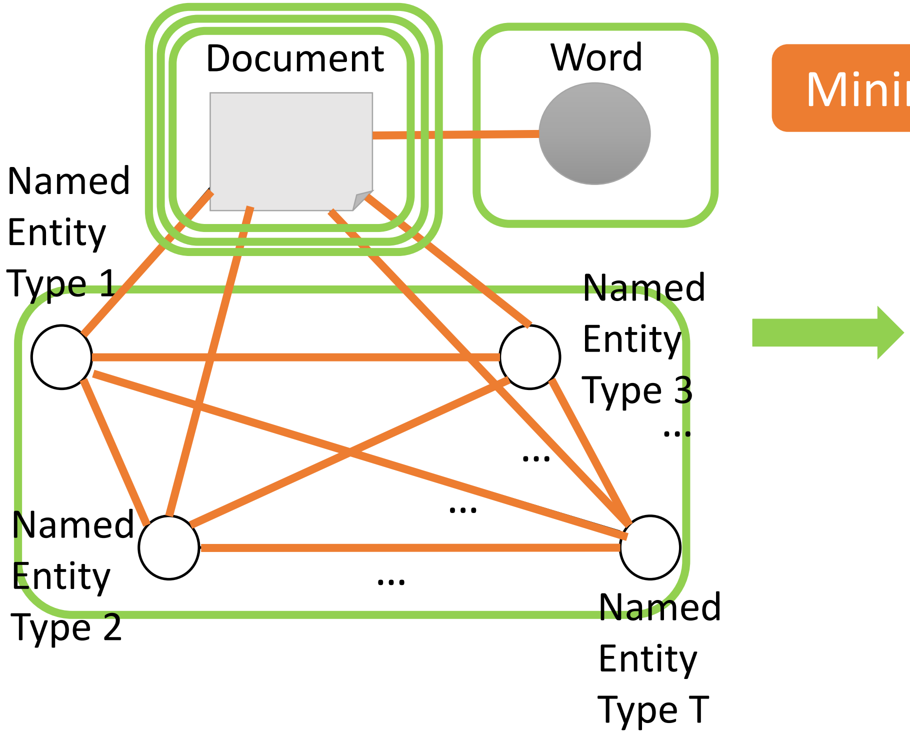
$$\boxed{q(d_m, w_i)} = p(\boxed{\hat{d}_{k_d}}, \hat{w}_{k_w}) p(d_m | \hat{d}_{k_d}) p(w_i | \hat{w}_{\boxed{k_w}})$$

Joint probability $p(d_m, w_i)$ approximation

Cluster indicators

Cluster indices

Knowledge indirect supervision: sub-types or attributes *cannot directly affect the document labels*. Constraints affect entity labels, entity labels will be transferred to affect the document labels.



Minimize

$$J_{CHINC} = D_{KL}(p(D,W)||q(D,W)) + \sum_{t=1}^{T} D_{KL}(p(D,E^t)||q(D,E^t)) + \sum_{t=1}^{T}\sum_{s=1}^{T} D_{KL}(p(E^t,E^s)||q(E^t,E^s))$$
$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in M_{e_{i_1}^t}} V_M\left(e_{i_1}^t, e_{i_2}^t \in M_{e_{i_1}^t}\right)$$

Must–link: if two labels are not equal, consider how dissimilar they are

$$+ \sum_{t=1}^{T}\sum_{e_{i_1}^t=1}^{V_t}\sum_{e_{i_2}^t \in C_{e_{i_1}^t}} V_C\left(e_{i_1}^t, e_{i_2}^t \in C_{e_{i_1}^t}\right)$$

Cannot-link: if two labels are equal, consider how similar they are

Algorithm: Alternating Optimization

Input: HIN defined on documents D, words W, entities $E^t, t = 1, \ldots, T$, Set maxIter and maxδ.

while iter < maxIter and δ > maxδ **do**

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

for t = 1.....T **do**

E^t **Label Update**: minimize J_{CHINC} w.r.t. L_{e^t}.
E^t **Model Update**: update $q(d_m, e_i^t)$ and $q(e_i^s, e_i^t)$.

end for

D **Label Update**: minimize J_{CHINC} w.r.t. L_d.
D **Model Update**: update $q(d_m, w_i)$ and $q(d_m, e_i^t)$.

W **Label Update**: minimize J_{CHINC} w.r.t. L_w.
W **Model Update**: update$q(d_m, w_i)$.

Compute cost change δ.

end while

Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

The number is reported in [X. Dong et al. KDD'14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.

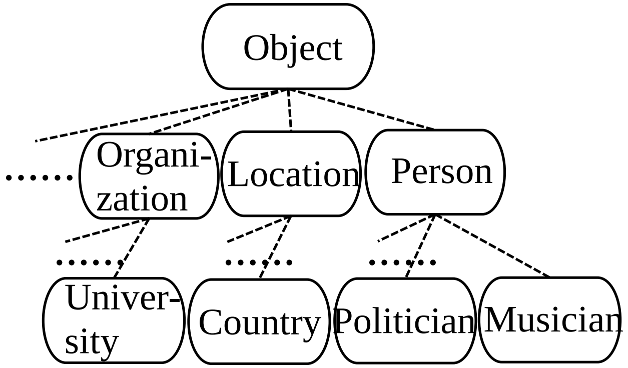
Experiments

Document datasets			
Name	#(Categories)	#(Leaf Categories)	#(Documents)
20Newsgroups (20NG)	6	20	20,000
MCAT (Markets)	9	7	44,033
CCAT (Corporate/Industrial)	31	26	47,494
ECAT (Economics)	23	18	19,813

MCAT, CCAT, ECAT are top categories in RCV1 dataset containing manually labeled newswire stories from Reuter Ltd.

World knowledge bases				
Name	#(Entity Types)	#(Entity Instances)	#(Relation Types)	#(Relation Instances)
Freebase	1,500	40 millions	35,000	2 billions
publicly available knowledge base with entities and relations collaboratively collected by its community members.				
YAGO2	350,000	10 millions	100	120 millions
a semantic knowledge base, derived from Wikipedia, WordNet and GeoNames.				

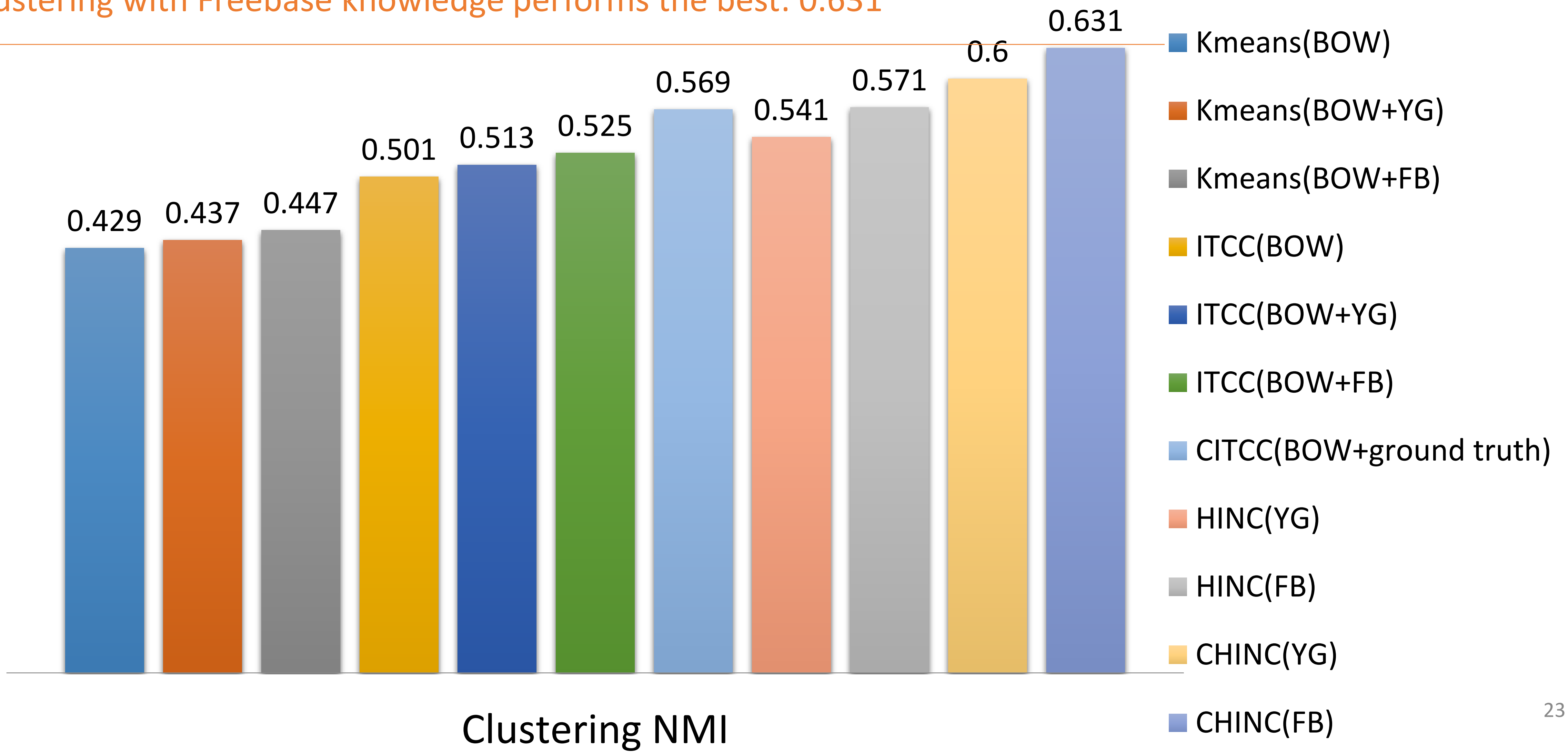
The number is reported in [X. Dong et al. KDD'14], In our downloaded dump of Freebase, we found 79 domains, 2,232 types, and 6,635 properties.



Entity type hierarchy in Freebase and YAGO2

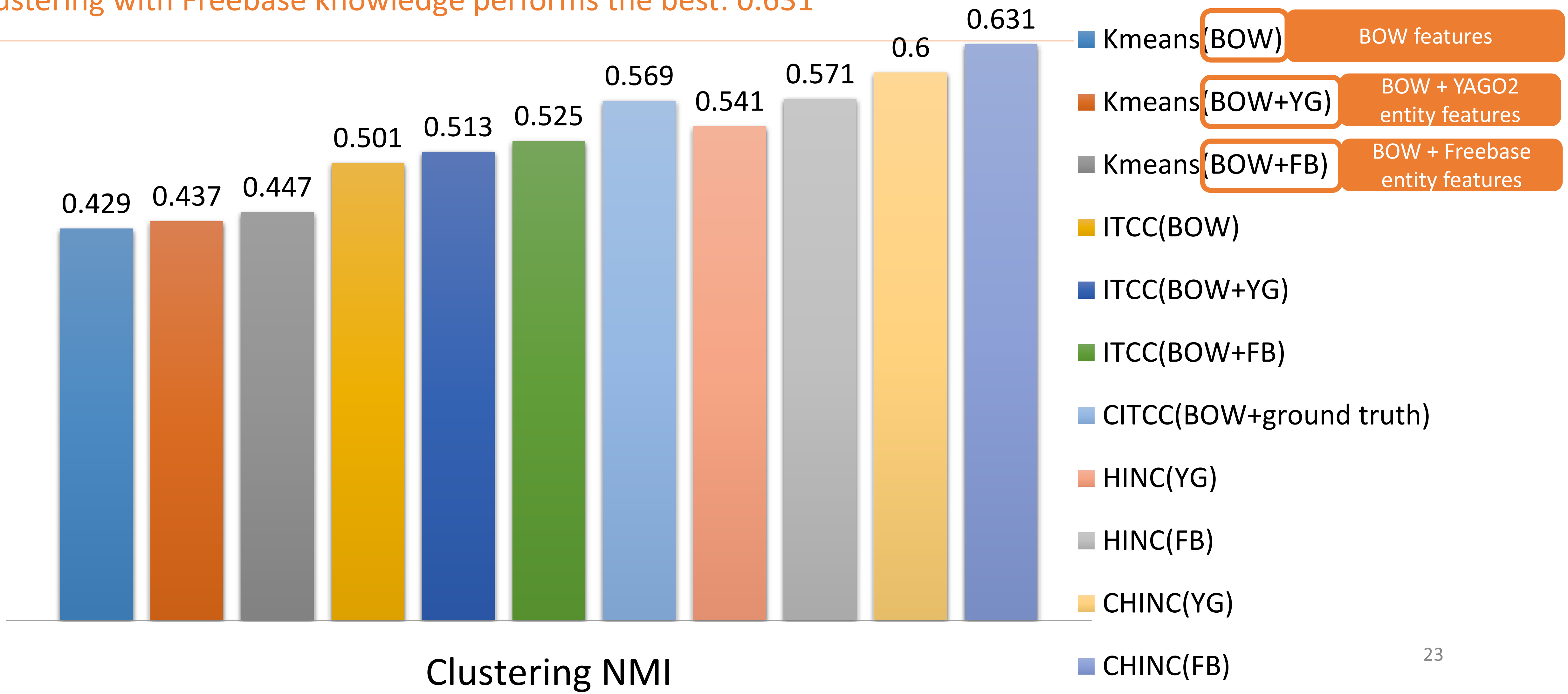
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



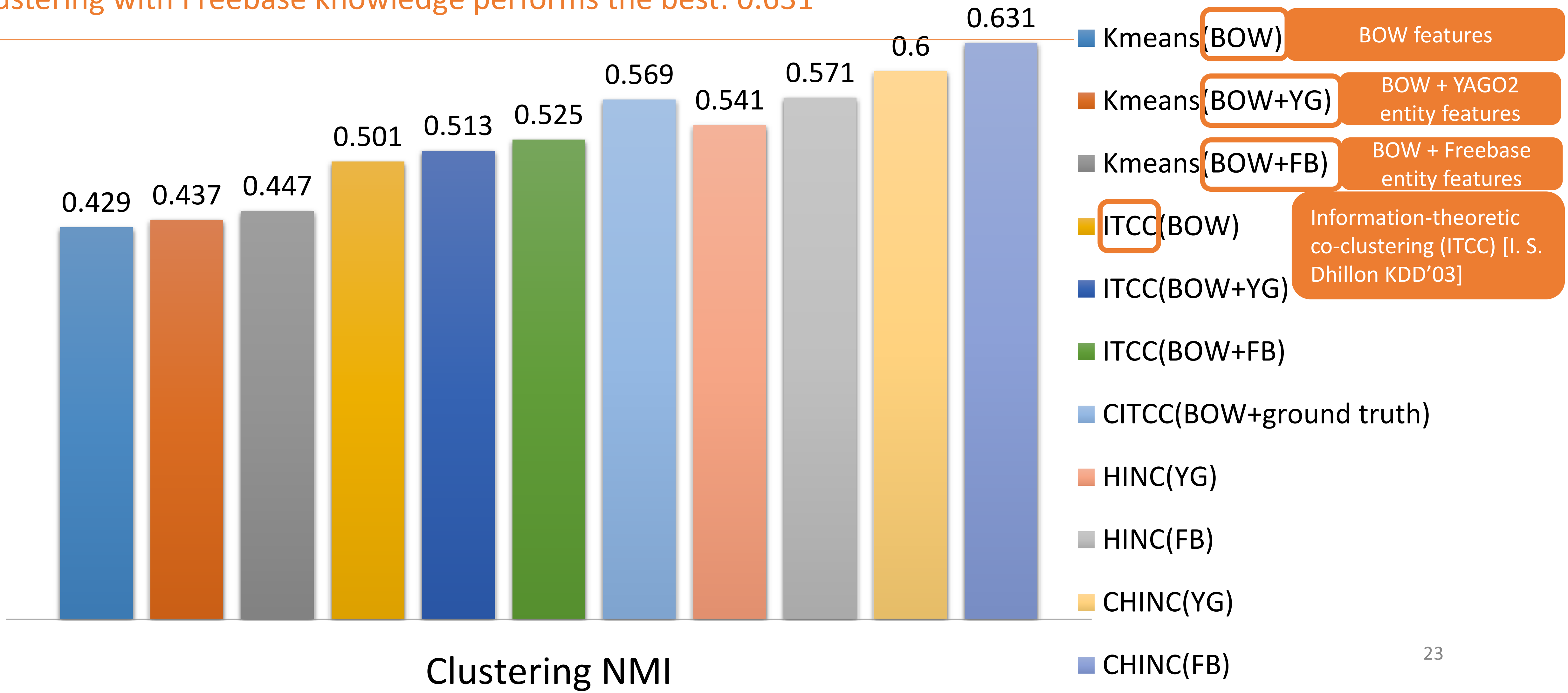
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



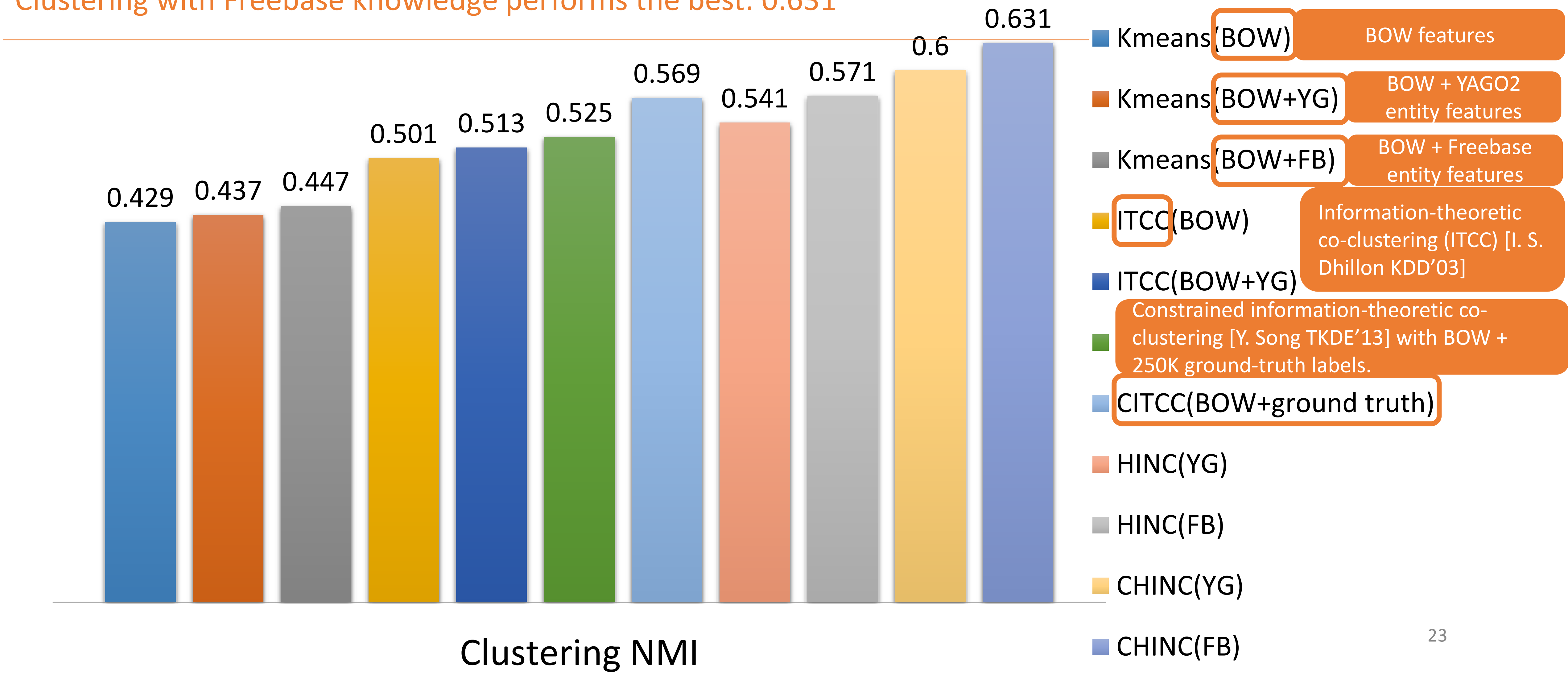
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



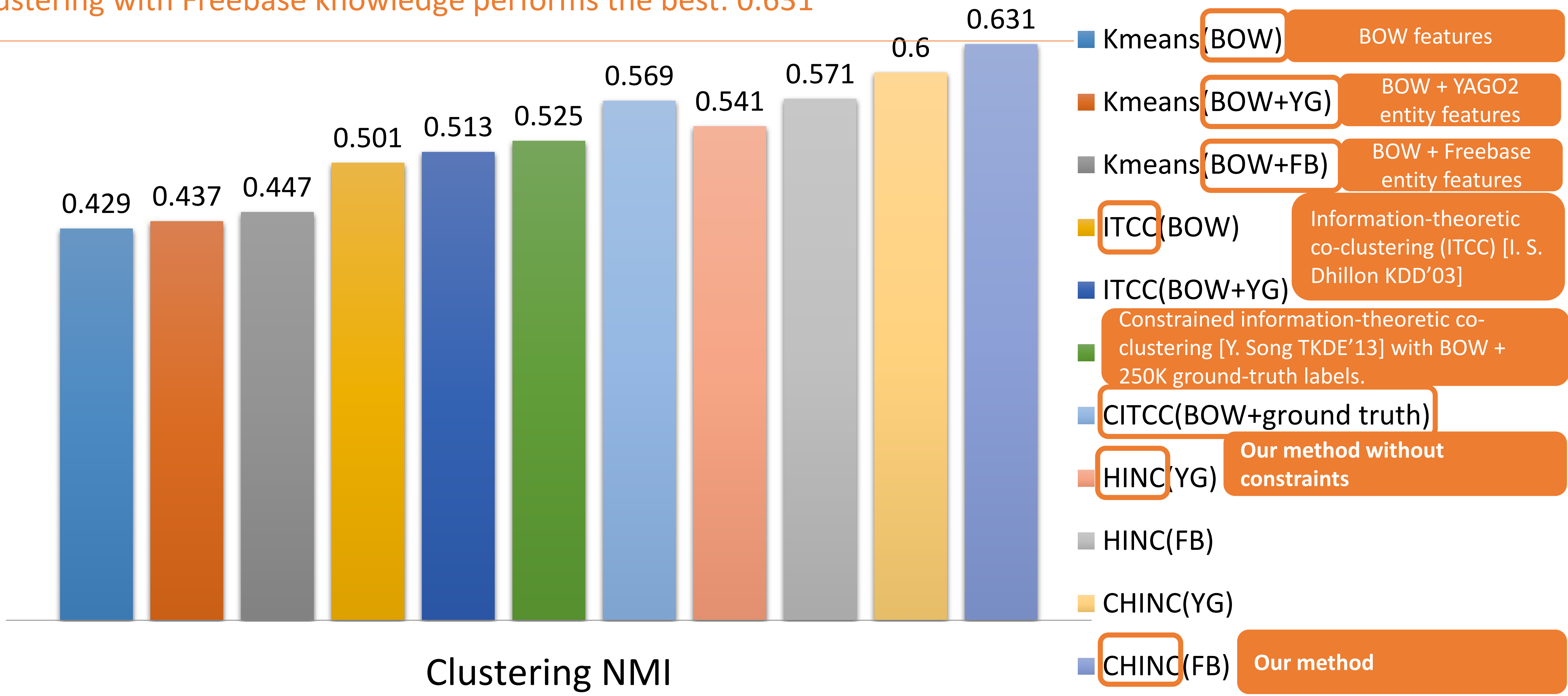
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



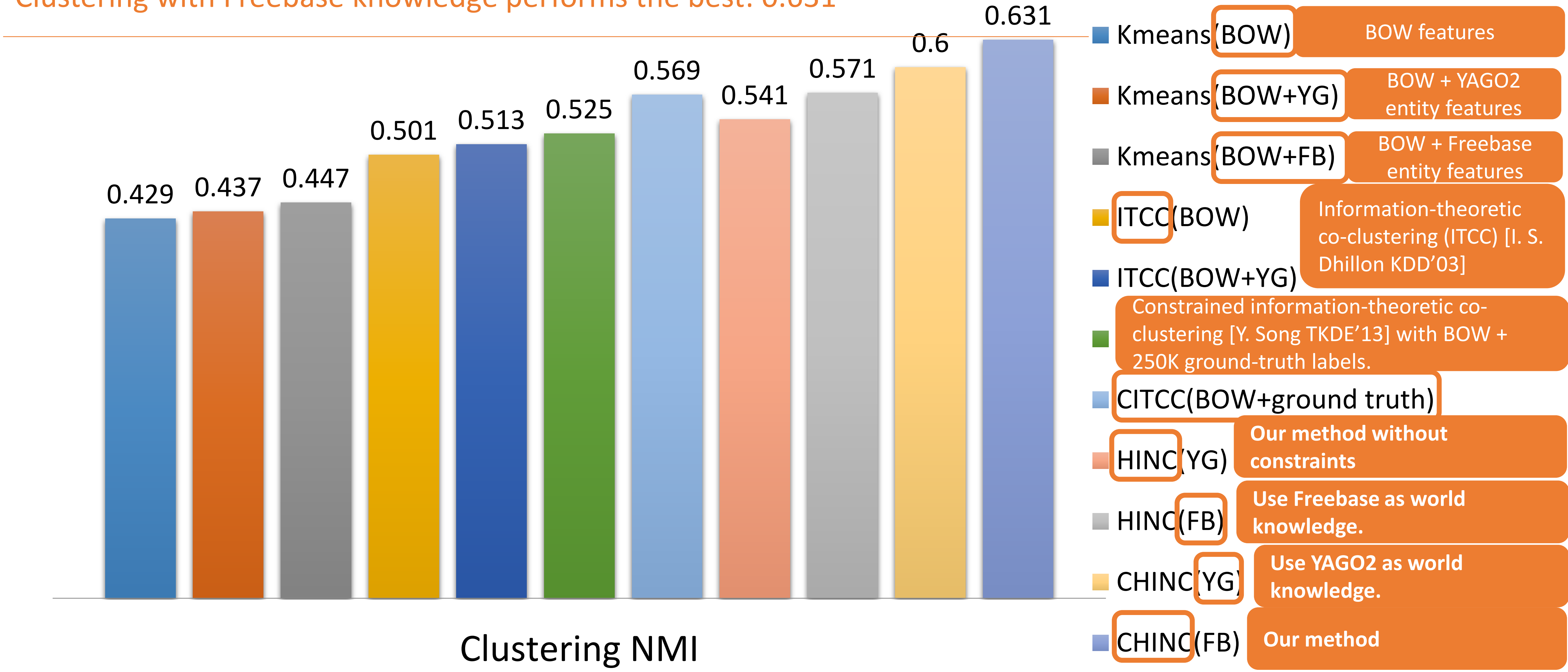
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



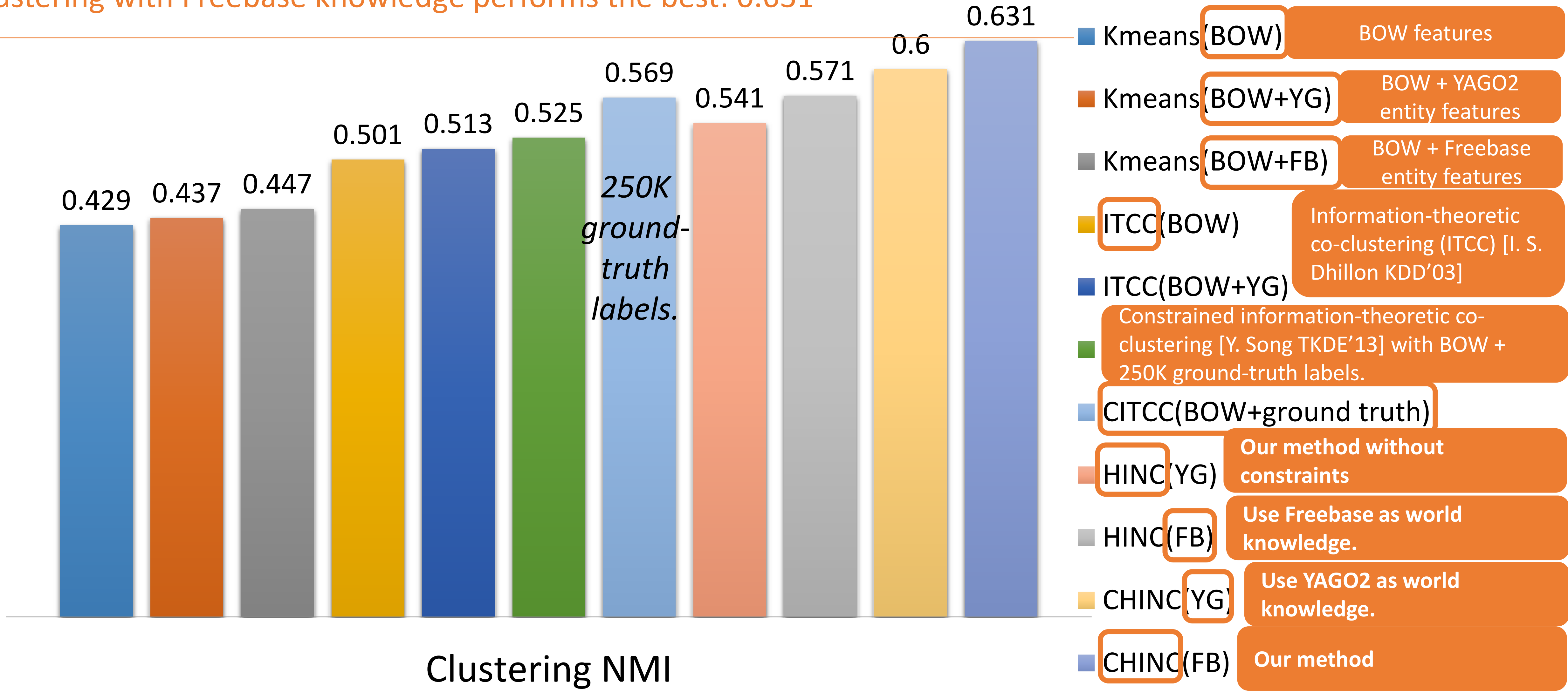
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



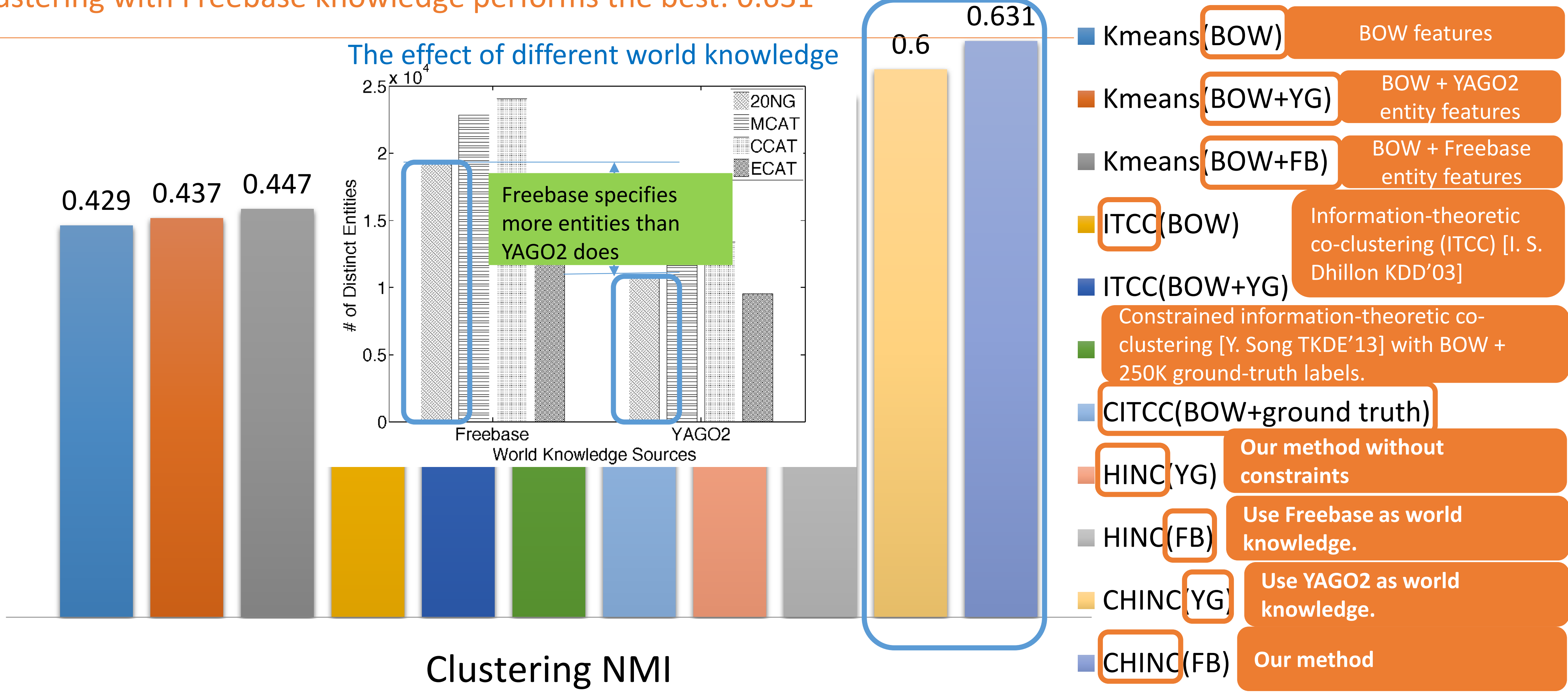
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



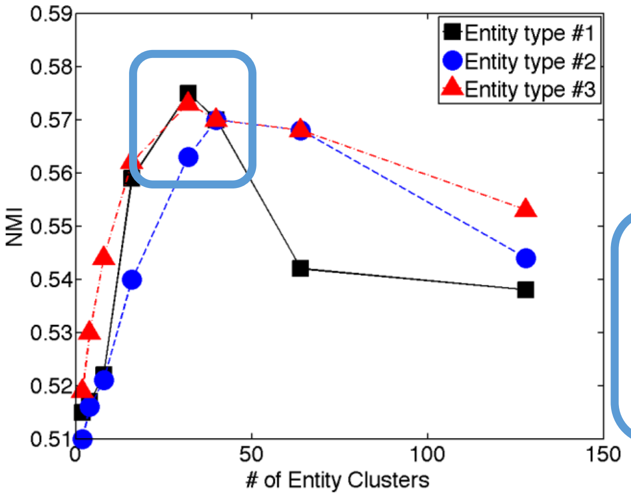
Clustering Results on 20 Newsgroups

Clustering with Freebase knowledge performs the best: 0.631



Parameter Study

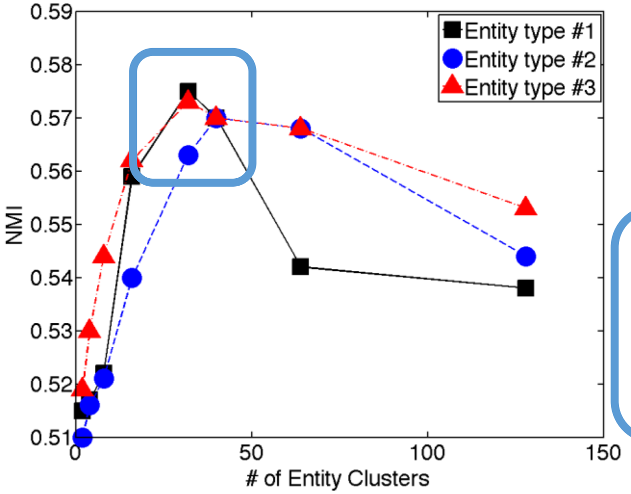
Clustering with different numbers of entity clusters of each entity type



Finding #1: <u>certain values</u> of the number of entity clusters leading to the best clustering performance.

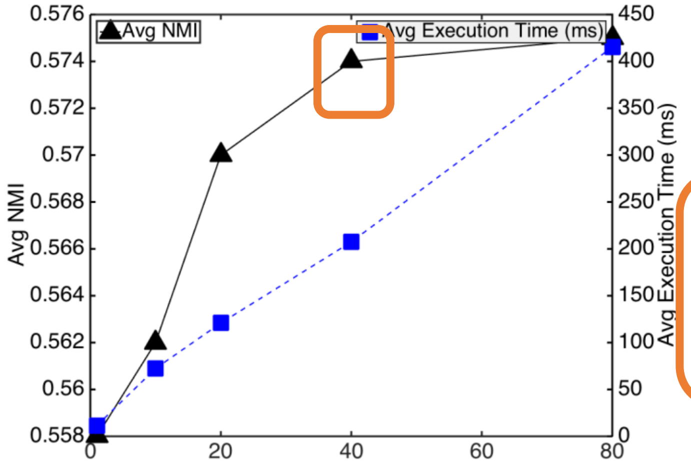
Parameter Study

Clustering with different numbers of entity clusters of each entity type



Finding #1: certain values of the number of entity clusters leading to the best clustering performance.

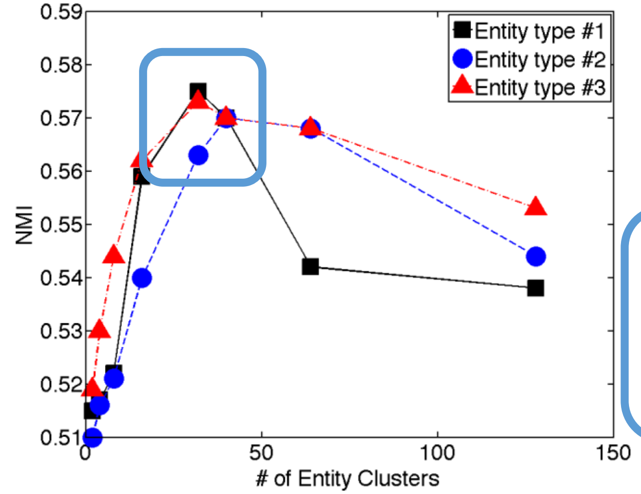
Optimization algorithm with different numbers of iterations



Finding #2: larger number of iterations, the clustering improves more, and become stable.
Because it comes to convergence.

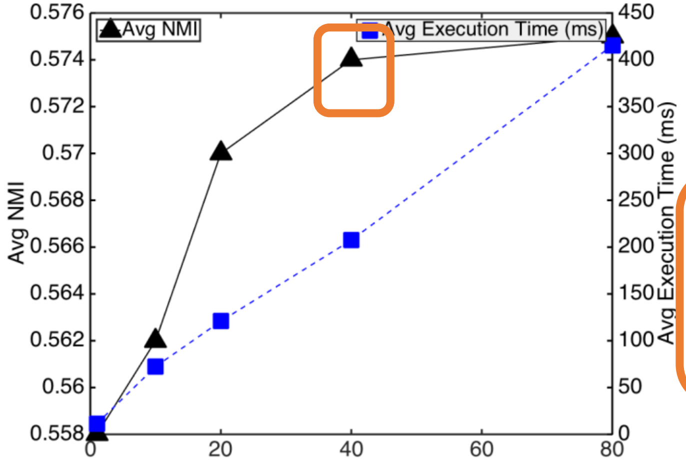
Parameter Study

Clustering with different numbers of entity clusters of each entity type



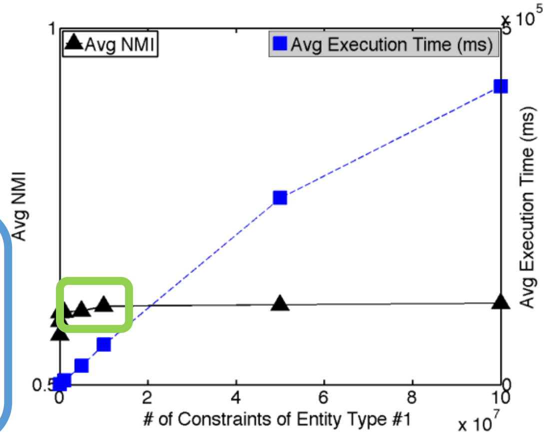
Finding #1: <u>certain values</u> of the number of entity clusters leading to the best clustering performance.

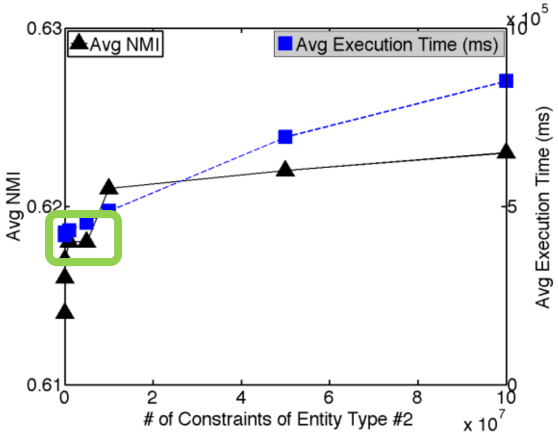
Optimization algorithm with different numbers of iterations



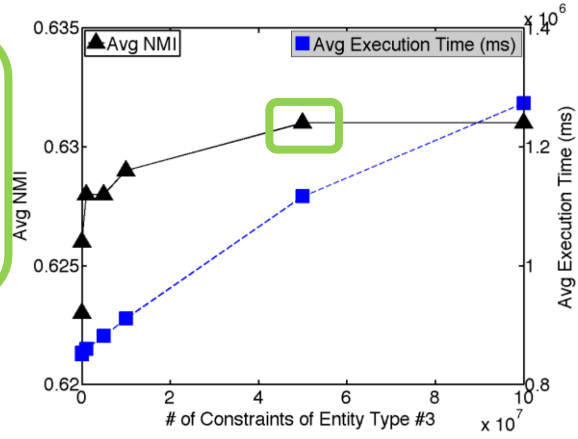
Finding #2: larger number of iterations, the clustering improves more, and <u>become stable.</u>
Because it comes to convergence.

Clustering with world knowledge constraints





Finding #3: adding more and more constraints leading to better performance. <u>Then become stable.</u>
The entity sub-type information is transferred to the document side.



Recall

Problem	Document clustering with world knowledge as indirect supervision.
Framework	World knowledge specification: unsupervised semantic parsing and conceptualization based semantic filtering.
Model	Constrained clustering model with the specified world knowledge represented in heterogeneous information network.

Recall

Problem	Document clustering with world knowledge as indirect supervision.
Framework	World knowledge specification: unsupervised semantic parsing and conceptualization based semantic filtering.
Model	Constrained clustering model with the specified world knowledge represented in heterogeneous information network.

Thank You! ☺