Heterogeneous Supervision for Relation Extraction: A Representation Learning Approach

Liyuan Liu[❖], Xiang Ren[❖], Qi Zhu[❖], Huan Gui[❖], Shi Zhi[❖], Heng Ji[◇] and Jiawei Han[❖]

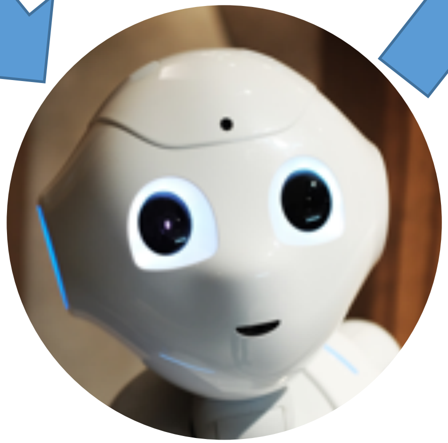
[❖]University of Illinois at Urbana-Champaign, Urbana, IL, USA

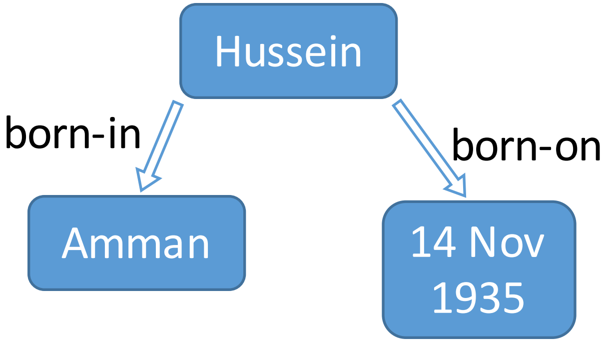
[◇]Computer Science Department, Rensselaer Polytechnic Institute, USA

Relation Extraction

- Goal: acquire structured knowledge from unstructured text

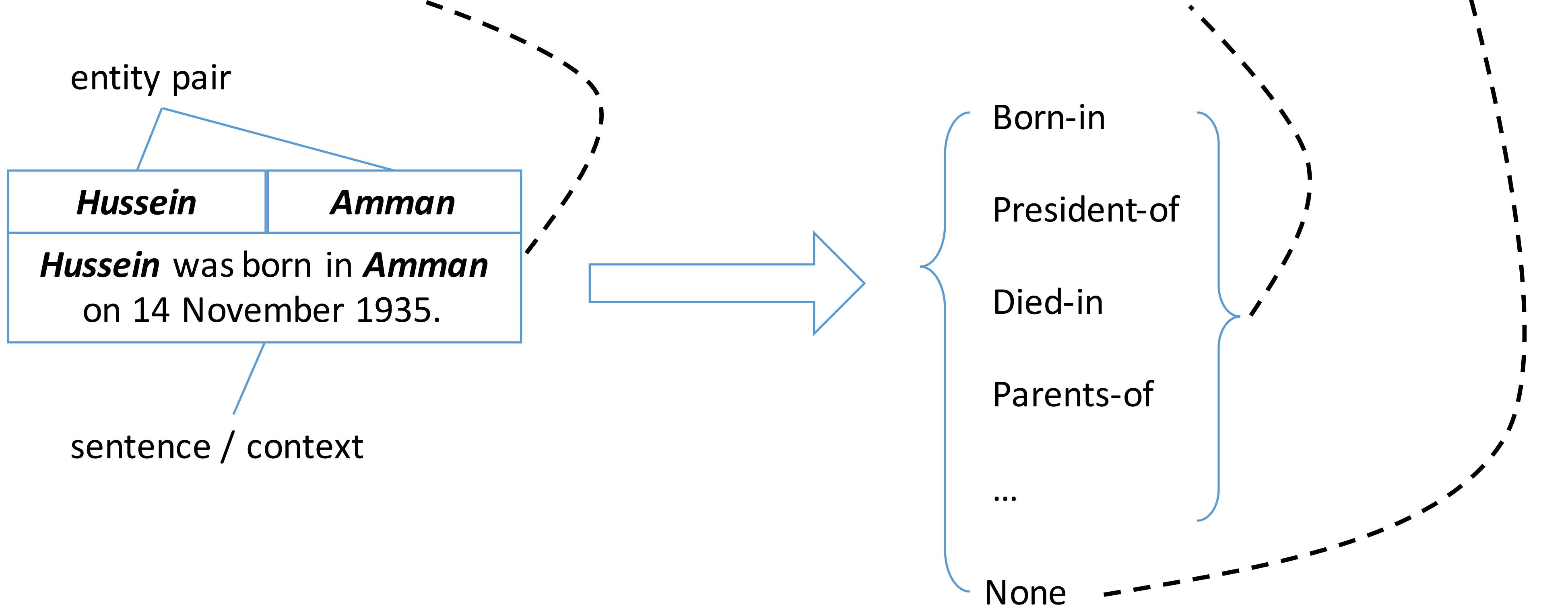
“Hussein was born in Amman on 14 November 1935.”





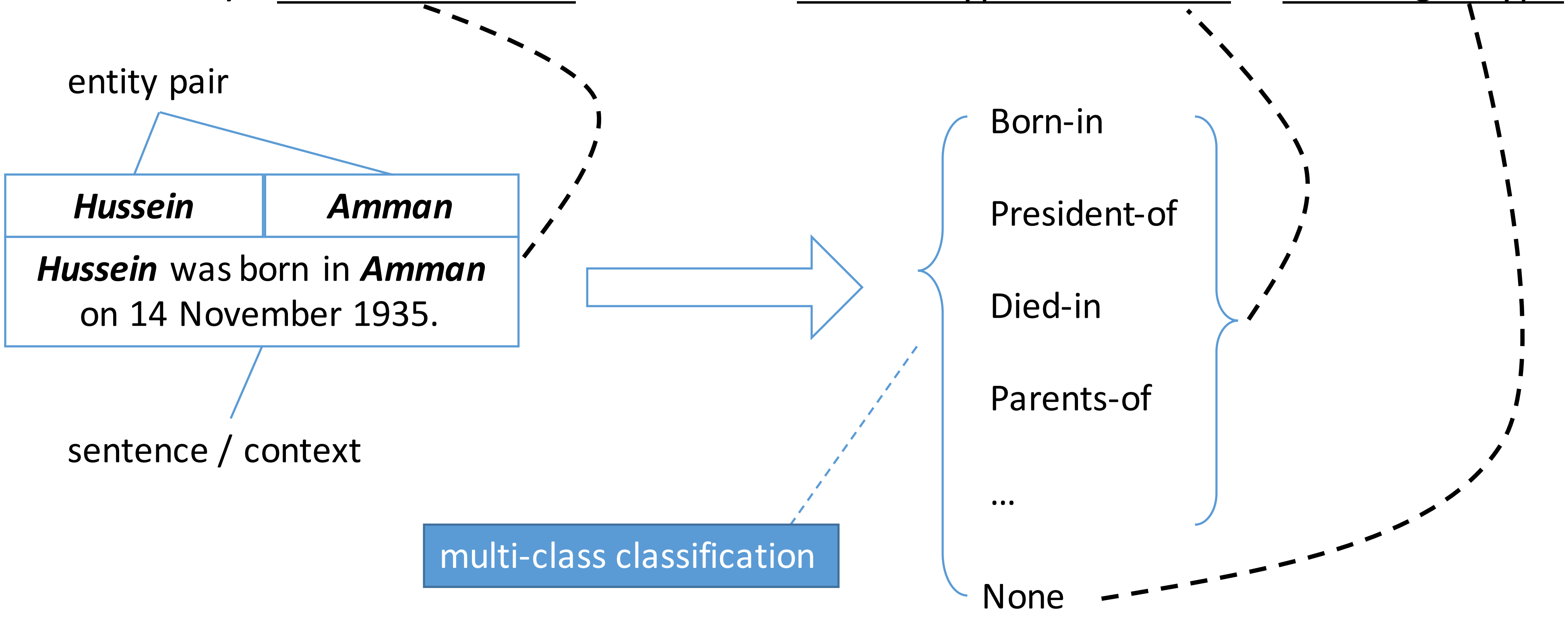
Relation Extraction

- Formal Definition:
  - Sentence-level relation extraction:
    - Classify a <u>relation mention</u> into a set of <u>relation types of interest</u> or <u>Not-Target-Type</u> (None)



Relation Extraction

- Formal Definition:
  - Sentence-level relation extraction:
    - Classify a relation mention into a set of relation types of interest or Not-Target-Type (None)



Related Work

- Supervised Learning:
  - Multi-class classification

Related Work

- Supervised Learning:
  - Multi-class classification

Dataset with human annotation is the bottleneck

Limited, might even not existed for many domains

Hard to get, and costly

Slow, and sometimes outdated

……

Related Work

- Bootstrap learning:
  - Start with a set of seed patterns / annotations, iteratively generate more
  - Suffers from semantic shift

Mintz et al. “Distant supervision for relation extraction without labeled data”, ACL 2009

Related Work

- Distant Supervision:
  - Automatically generate annotations by Knowledge Base

Mintz et al. "Distant supervision for relation extraction without labeled data", ACL 2009

Related Work

- Distant Supervision:
  - Automatically generate annotations by Knowledge Base
    - ("Obama", "USA", Obama was born in Honolulu, Hawaii, USA as he has always said)
      - *Born-in* (correct)
      - *President-of* (wrong).

Related Work

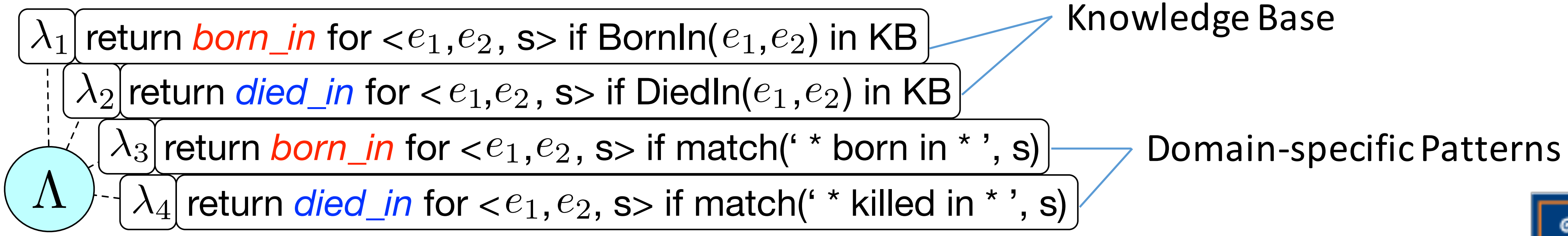
- Distant Supervision:
    - Automatically generate annotations by Knowledge Base

Distant supervision only encodes KB, while we have more than KB

Heterogeneous Supervision

- Provide a general framework to encode knowledge for supervision:
  - Knowledge Base, domain-specific patterns, ……

- Labelling functions:



Ratner, Alexander J., et al. "Data programming: Creating large training sets, quickly." NIPS. 2016.

Heterogeneous Supervision & Distant Supervision:

- Heterogeneous Supervision is an extension of Distant Supervision:
  - Both encode external information and provide supervision,
  - Heterogeneous Supervision can encode more.

Information type	KBP		NYT	
	# of Relation Types	# of Relation Mentions	# of Relation Types	# of Relation Mentions
Knowledge Base	7	133955	25	530767
Domain-specific Patterns	13	225977	16	43820

Table1. Statistic of Heterogeneous Supervision

Heterogeneous Supervision & Distant Supervision:

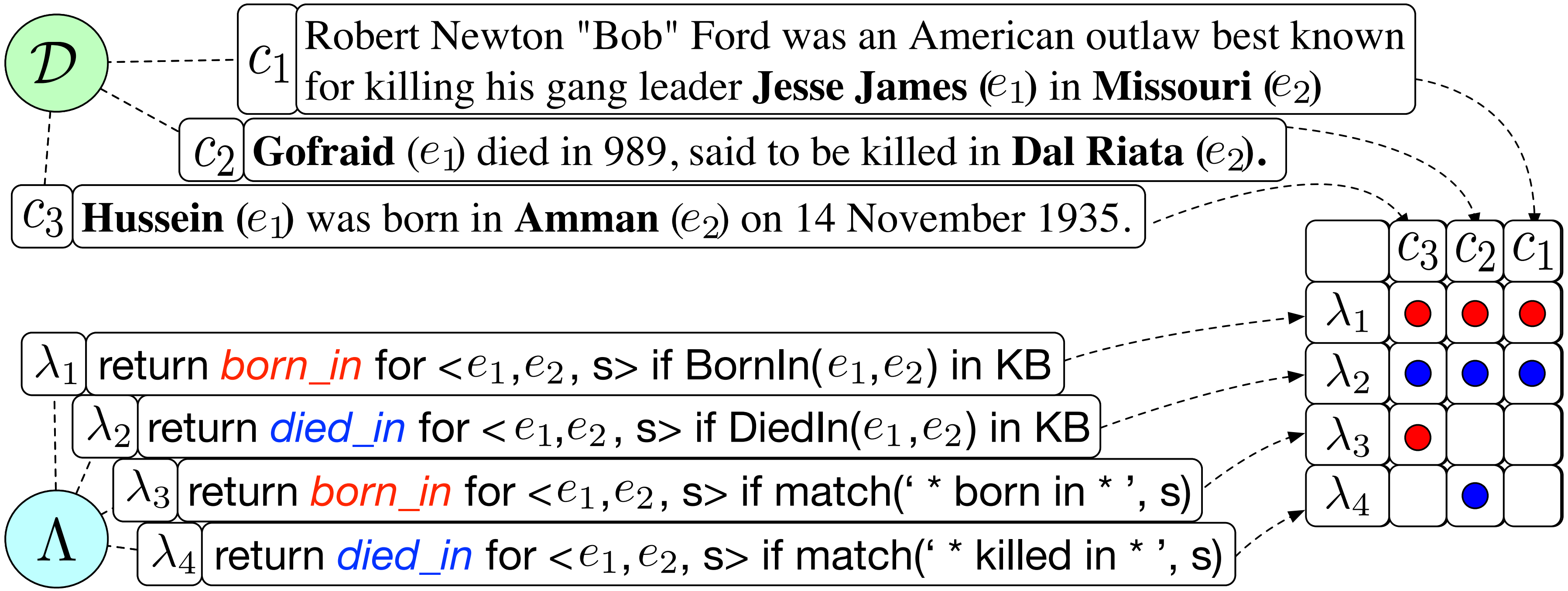
- Heterogeneous Supervision is an extension of Distant Supervision:
  - Both encode external information and provide supervision,
  - Heterogeneous Supervision can encode more.

Information type	KBP		NYT	
	# of Relation Types	# of Relation Mentions	# of Relation Types	# of Relation Mentions
Knowledge Base	7	133955	25	530767
Domain-specific Patterns	13	225977	16	43820

Table1. Statistic of Heterogeneous Supervision

Challenges

- Relation Extraction
- Resolve Conflicts among Heterogeneous Supervision

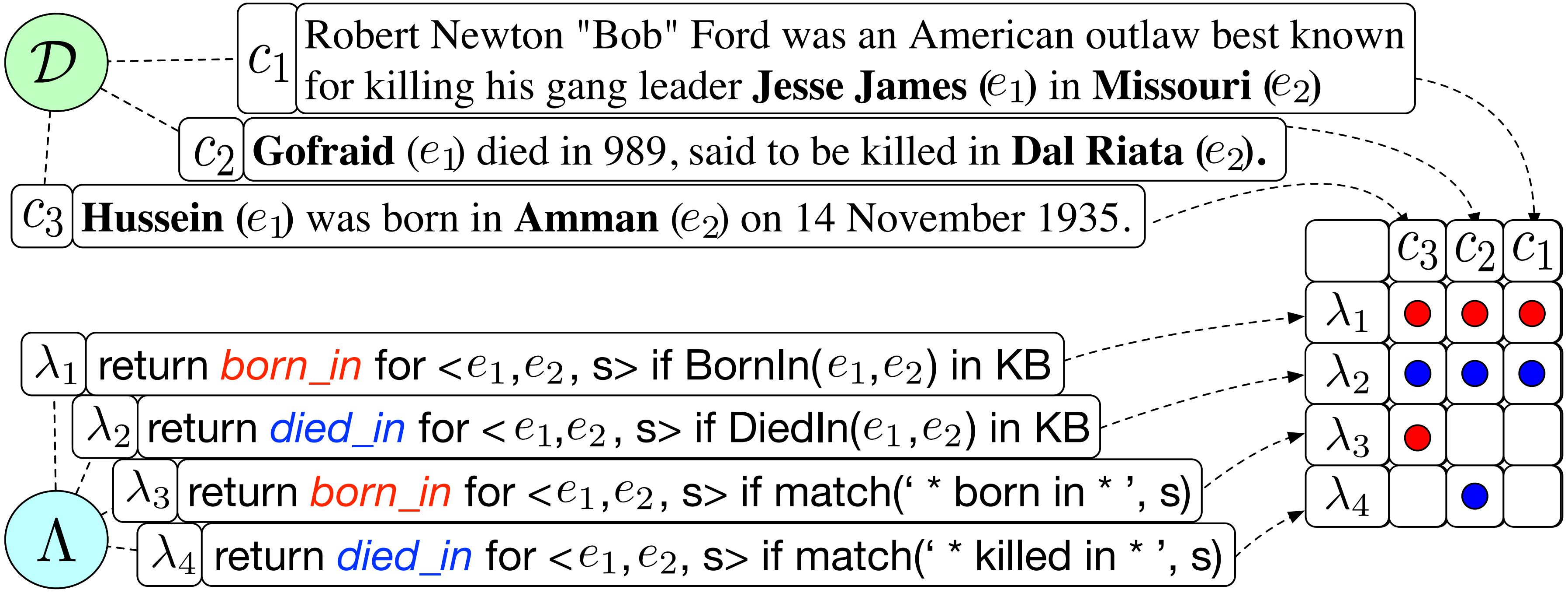


ReHession

- Heterogeneous Supervision
- Our Solution: A Representation Learning Approach
  - Relation Mention Representation
  - True Label Discovery component
  - Relation Extraction component
- Experiments

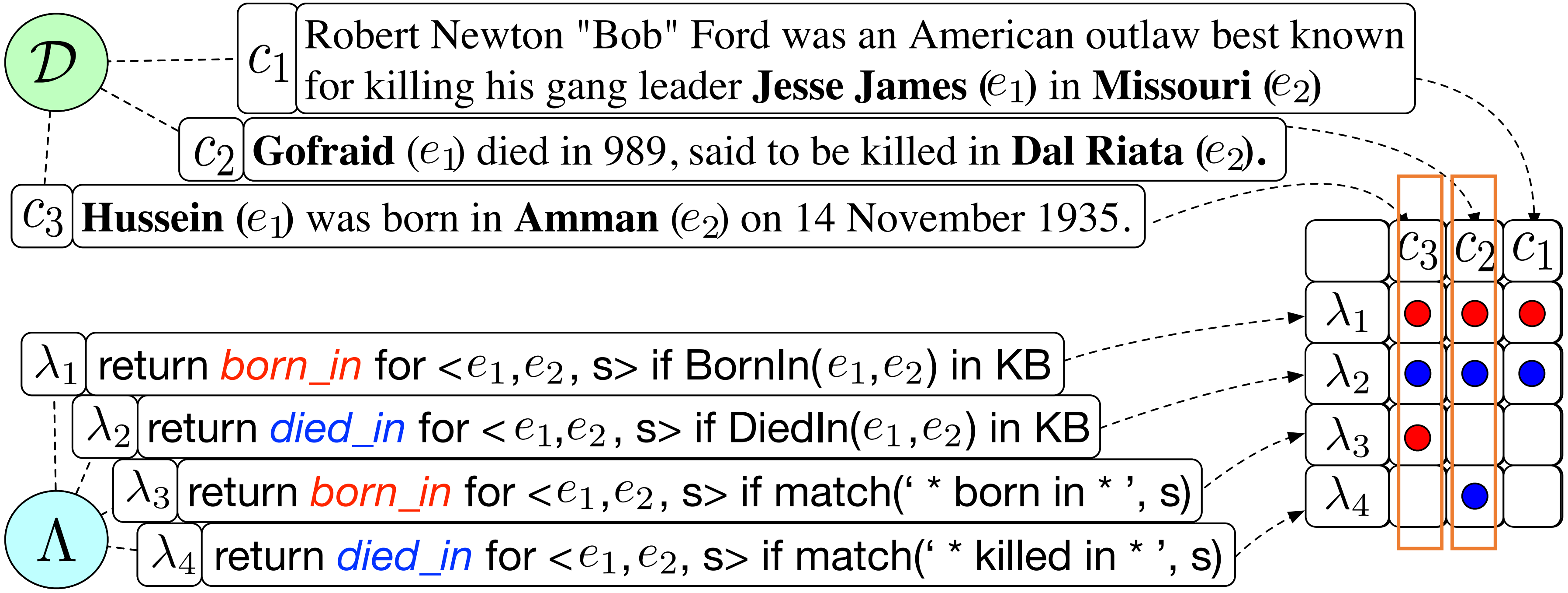
Conflicts among Heterogeneous Supervision

- Most simple way: majority voting



Conflicts among Heterogeneous Supervision

- How to resolve conflicts among Heterogeneous Supervision?
  - Works for C3 and C2, but not work for C1



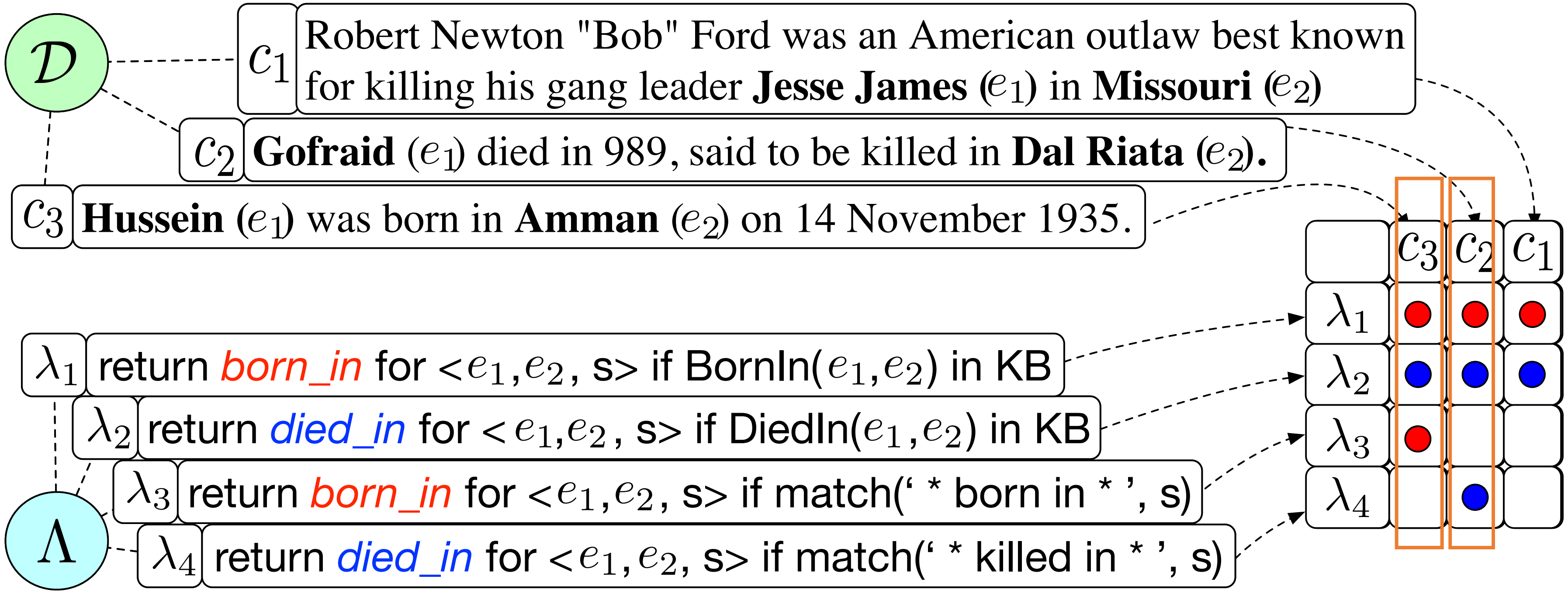
Conflicts among Heterogeneous Supervision

- For more complicated models, several principles have been proposed:
  - Truth Discovery:
    - Some sources (labeling functions) would be more reliable than others
    - Refer the reliability of different sources and the true label at the same time
    - Source Consistency Assumption: a source is likely to provide true information with the same probability for all instances.

Li, Yaliang, et al. "A survey on truth discovery." Acm Sigkdd Explorations Newsletter 17.2 (2016): 1-16.

Conflicts among Heterogeneous Supervision

- For more complicated models, several principles have been proposed:
  - Truth Discovery:
    - May not fit our scenario very well



Conflicts among Heterogeneous Supervision

- For more complicated models, several principles have been proposed:
  - Truth Discovery:
    - These models are context-agnostic, while context is important for Relation Extraction

Conflicts among Heterogeneous Supervision

- For more complicated models, several principles have been proposed:
  - Truth Discovery:
  - Distant Supervision:
    - Partial-label association has been proposed to resolve conflicts among Distant Supervision, and proved to be effective.

$$l(z, O_z) = \max\{0, 1 - [\underbrace{\max_{r \in O_z} \phi(z, r)}_{\text{Most likely positive relation type}} - \underbrace{\max_{r' \notin O_z} \phi(z, r')}_{\text{Most likely negative relation type}}]\}$$

Xiang Ren, et al. "CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases", WWW, 2017.

Conflicts among Heterogeneous Supervision

- For more complicated models, several principles have been proposed:
  - Truth Discovery:
  - Distant Supervision:
    - Partial-label association has been proposed to resolve conflicts among Distant Supervision, and proved to be effective.

$$l(z, O_z) = \max\{0, 1 - \left[\max_{r \in O_z} \phi(z, r) - \max_{r' \notin O_z} \phi(z, r')\right]\}$$

    - For Distant Supervision, all annotations come from Knowledge Base.
    - For Heterogeneous Supervision, annotations are from different sources, and some could be more reliable than others.

Conflicts among Heterogeneous Supervision

- To fit our problem, we introduce context awareness to truth discovery, and modified the assumption:

    - A source is likely to provide true information with the same probability for instances *with similar context*.

Heterogeneous Supervision

- To fit our assumption, we add one constraint to labeling functions:
  - each labeling function can annotate ***only one*** relation type based on ***one source*** of information

- Reasons:
  - Different information sources often have different reliabilities
  - Some sources annotate different relation types without consistency
    - KB-based labeling function may have higher recall on ‘president-of’ than ‘born-in’

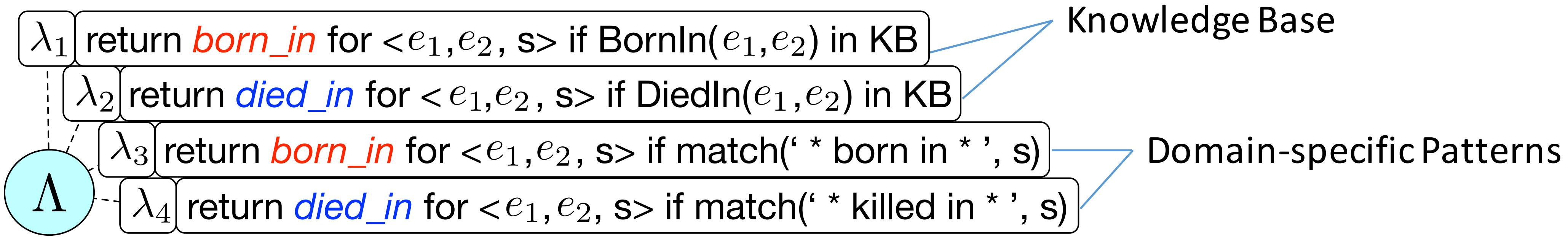
Heterogeneous Supervision

- To fit our assumption, we add one constraint to labeling functions:
  - each labeling function can annotate ***only one*** relation type based on ***one source*** of information

return ***r*** for <e1, e2, s> if ***r*** (e1, e2) in KB

Heterogeneous Supervision

- To fit our assumption, we add one constraint to labeling functions:
  - each labeling function can annotate ***only one*** relation type based on ***one source*** of information



ReHession

- Heterogeneous Supervision
- Our Solution: A Representation Learning Approach
  - Relation Mention Representation
  - True Label Discovery component
  - Relation Extraction component
- Experiments

Heterogeneous Supervision for Relation Extraction

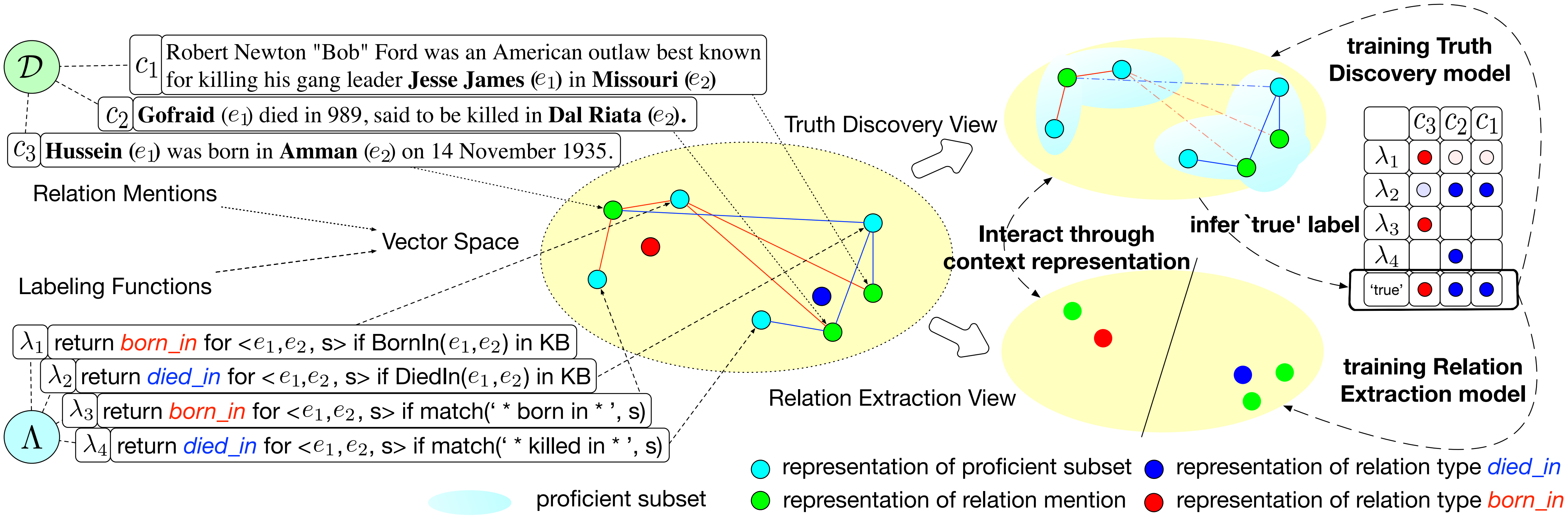
- Relation Extraction:
  - Matching context with proper relation type
- Heterogeneous Supervision:
  - Refer true labels in a context-aware manner

Heterogeneous Supervision for Relation Extraction

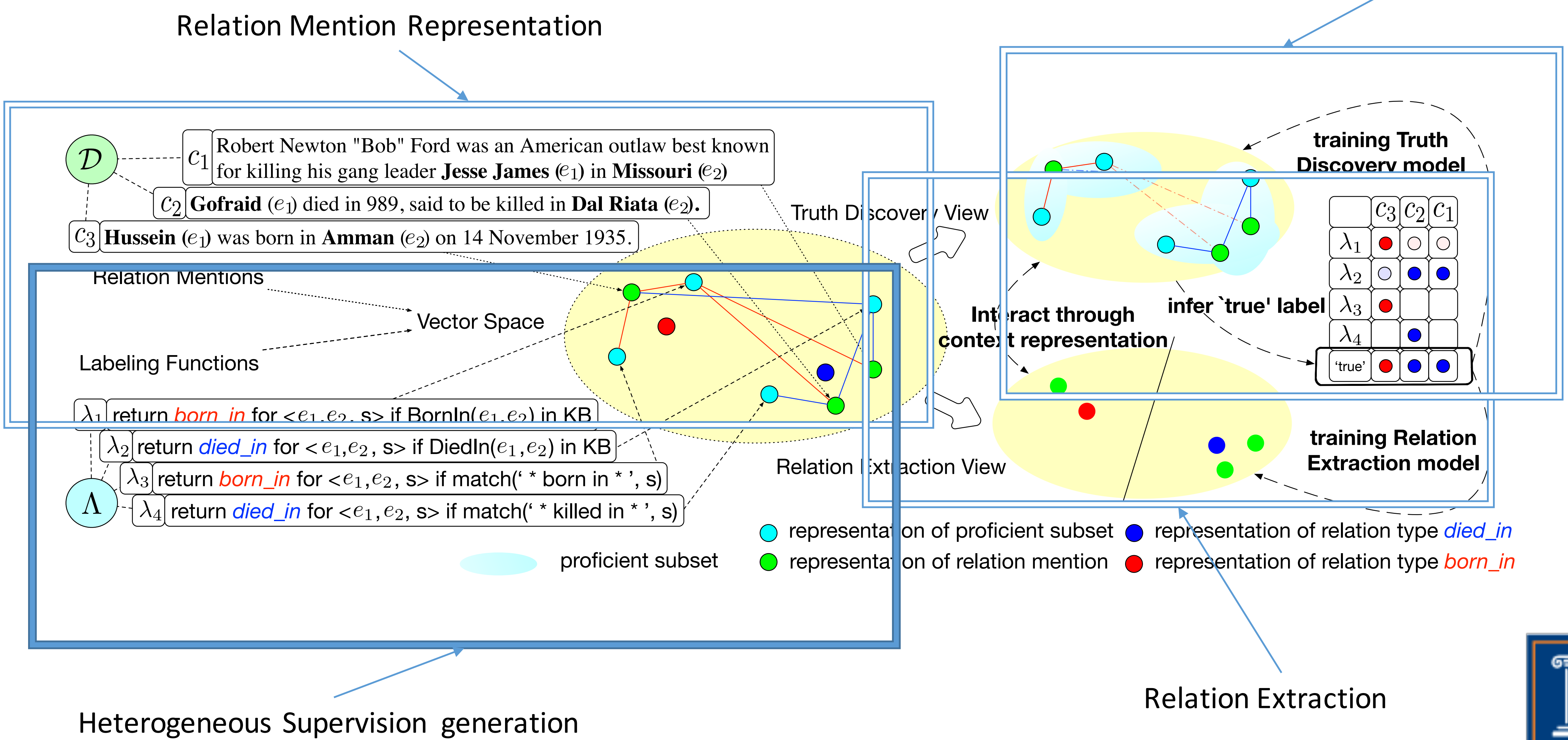
- Relation Extraction:
  - Matching ***context*** with proper relation type
- Heterogeneous Supervision:
  - Refer true labels in a ***context***-aware manner

context

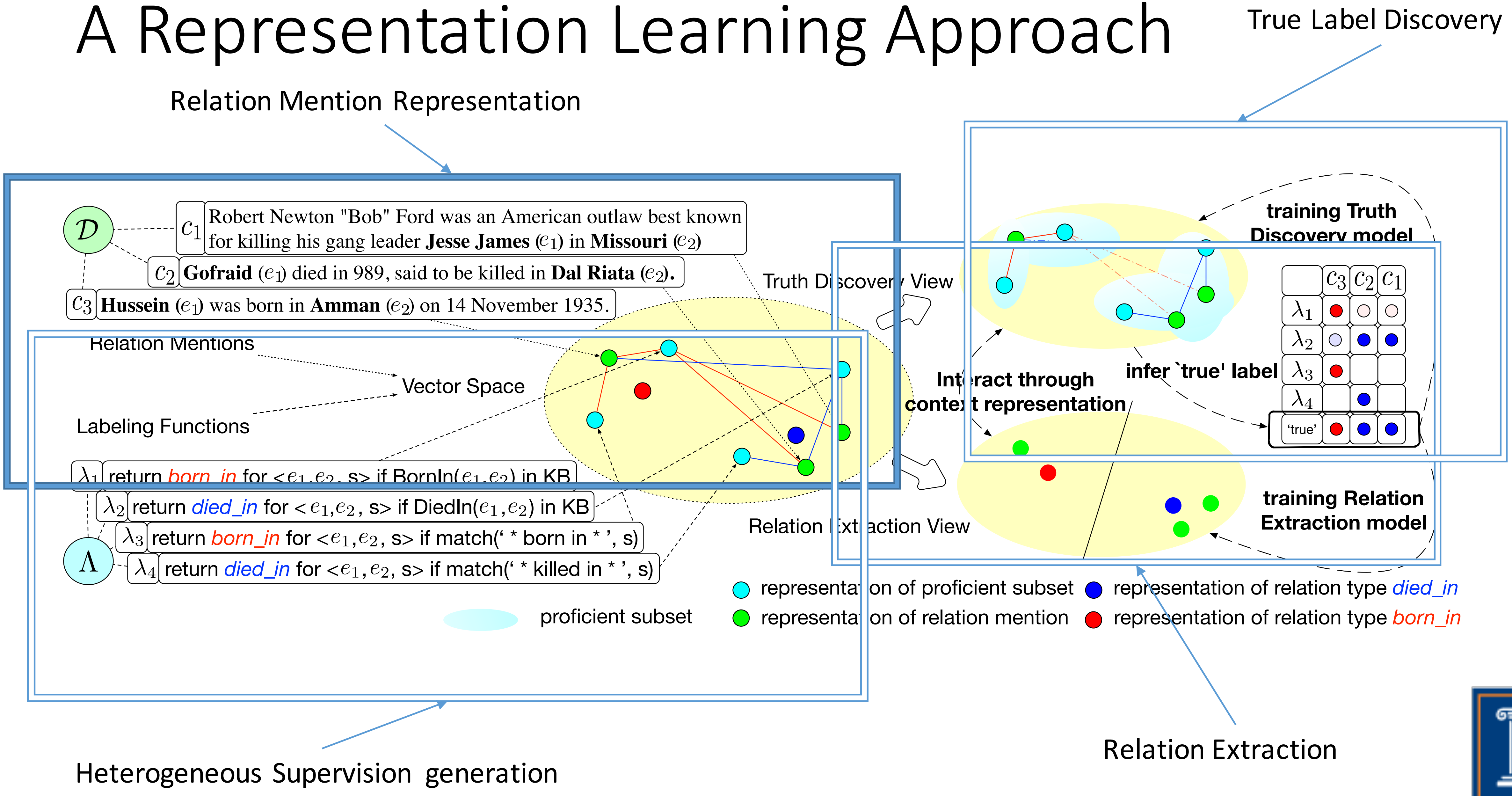
A Representation Learning Approach



A Representation Learning Approach

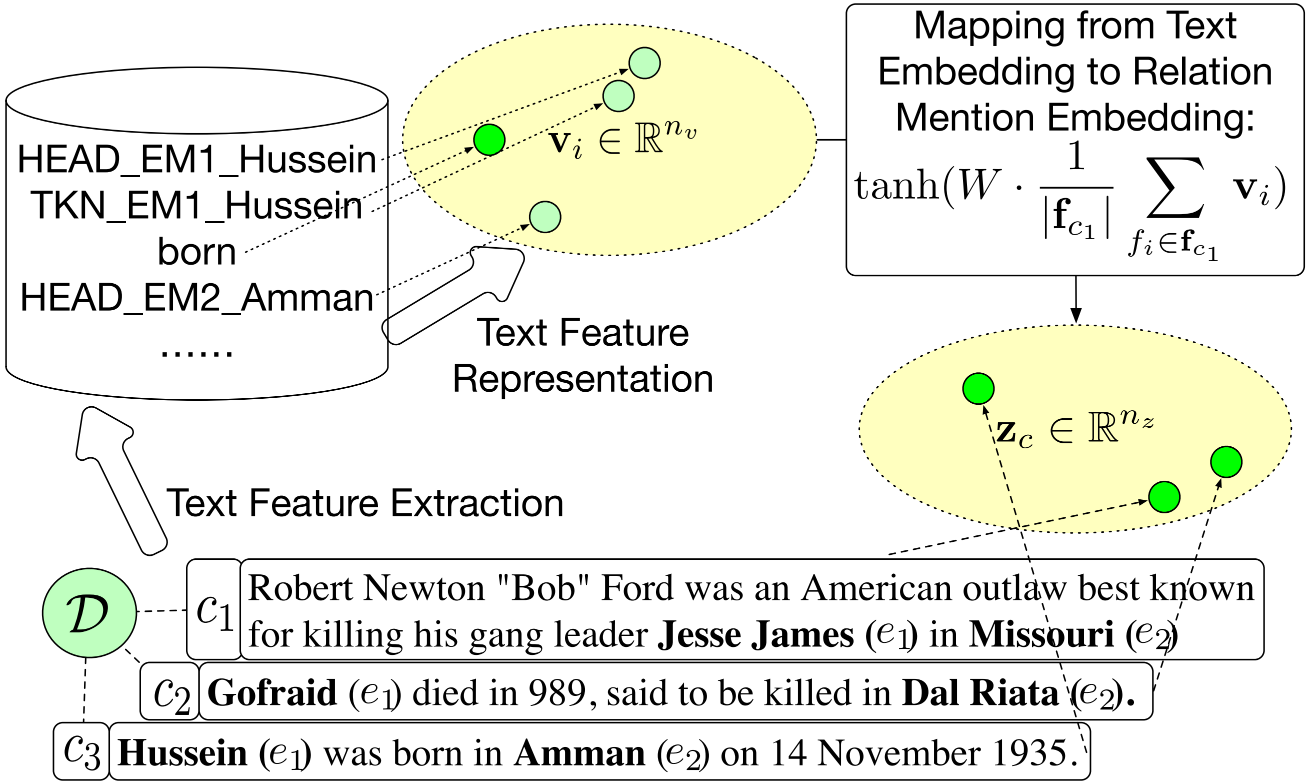


A Representation Learning Approach



Relation Mention Representation

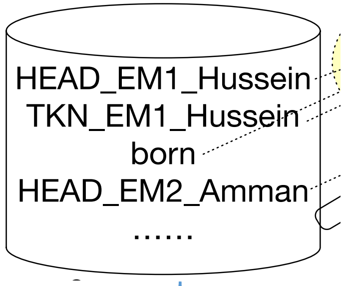
- Text Feature Extraction
- Text Feature Representation
- Relation Mention Representation



Text Feature Extraction

We adopted texture features, POS-tagging and **brown clustering** to extract features

C3: ***Hussein*** was born in ***Amman*** on 14 November 1935



Feature	Description	Example
Entity mention (EM) head	Syntactic head token of each entity mention	"*HEAD_EM1_Hussein*", ...
Entity Mention Token	Tokens in each entity mention	"*TKN_EM1_Hussein*", ...
Tokens between two EMs	Tokens between two EMs	"*was*", "*born*", "*in*"
Part-of-speech (POS) tag	POS tags of tokens between two EMs	"*VBD*", "*VBN*", "*IN*"
Collocations	Bigrams in left/right 3-word window of each EM	"*Hussein was*", "*in Amman*"
Entity mention order	Whether EM 1 is before EM 2	"*EM1_BEFORE_EM2*"
Entity mention distance	Number of tokens between the two EMs	"*EM_DISTANCE_3*"
Body entity mentions numbers	Number of EMs between the two EMs	"*EM_NUMBER_0*"
Entity mention context	Unigrams before and after each EM	"*EM_AFTER_was*", ...
Brown cluster (learned on $\mathcal{D}$)	Brown cluster ID for each token	"*BROWN_010011001*", ...

Text Feature Representation

- Leverage features' co-occurrence information to learn the representation , and help the model generalize better.

- Loss function of this part:

$$\mathcal{J}_E = \sum_{\substack{c\in\mathcal{C}_l \\ f_i, f_j \in \mathbf{f}_c}} \left(\log\sigma(\mathbf{v}_i^T \mathbf{v}_j^*) - \sum_{k=1}^{V} \mathbb{E}_{f_{k'}\sim\hat{P}}[\log\sigma(-\mathbf{v}_i^T \mathbf{v}_{k'}^*)]\right)$$

Feature set of c

Feature embedding for feature f_i

Negative sampling

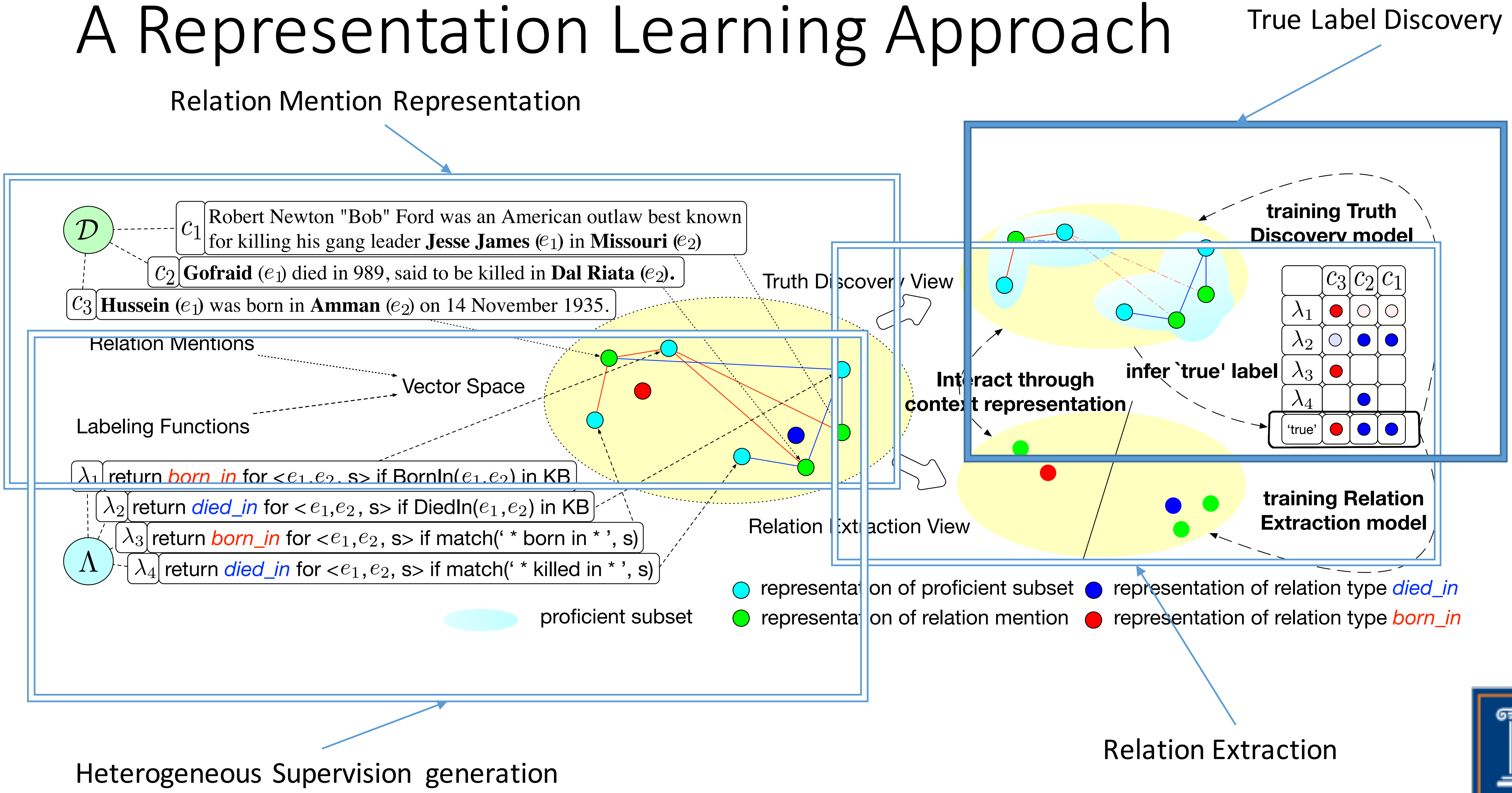
co-occurrence here refers to features occur in the same relation mention instead of the same shifting window

Relation Mention Representation

- Here, we adopted the bag-of-features assumption, and add transformation weights to allow representation of relation mention and features to be in different semantic space.

$$\mathbf{z}_c = g(\mathbf{f}_c) = \tanh(W \cdot \frac{1}{|\mathbf{f}_c|} \sum_{f_i \in \mathbf{f}_c} \mathbf{v}_i)$$

A Representation Learning Approach



True label discovery

- Assume:
  - A labeling function would annotate similar instances with the same reliability

Context Information: $\mathbf{z}$

True label discovery

- Assume:
  - A labeling function would annotate similar instances with the same reliability

Context Information: $\mathbf{z}$

for each labeling function, there exists an proficient subset, containing instances that it can precisely annotate.

True label discovery

- How to decide which label is correct?
  - Probability model and maximum likelihood estimate

Corresponding to our
assumption and setting

Identify the true label

True label discovery

- Probability Model:
  - Describing the generation of Heterogeneous Supervision?
  - Different from crowdsourcing. E.g., ONE worker may annotate:
    - ("Obama", "USA", Obama was born in Honolulu, Hawaii, USA as he has always said)
      - *Born-in* Cautious Worker
      - *President-of*
      - *Citizen-of*
      - …

Careless Worker

Exists some randomness

True label discovery

- Probability Model:
  - Describing the generation of Heterogeneous Supervision?
  - Different from crowdsourcing. E.g., <u>ONE</u> worker may annotate:
    - ("Obama", "USA", Obama was born in Honolulu, Hawaii, USA as he has always said)
      - *Born-in*
      - *President-of*
      - *Citizen-of*
      - …

  - But <u>One</u> labeling function can only annotate <u>One</u> relation type:
    - Randomness exists in the correctness, not in the choice of relation type

True label discovery

- Describing the correctness of Heterogeneous Supervision

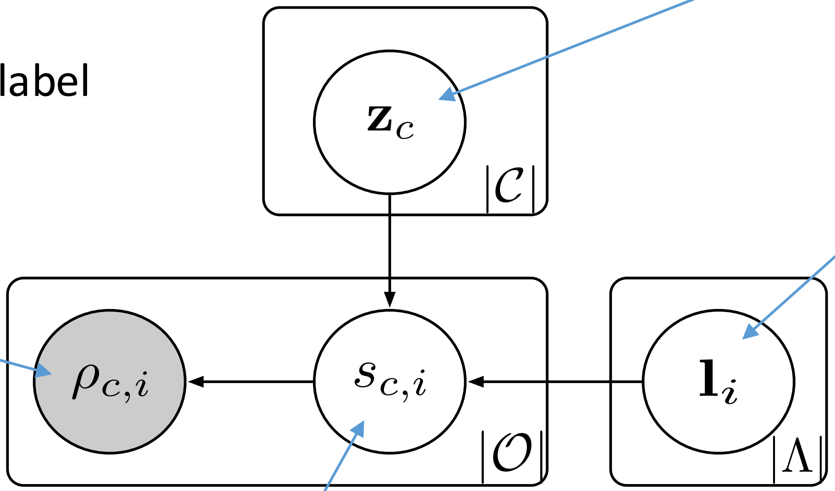
Representation of relation mention

observed annotation

underlying true label

$$\rho_{c,i} = \delta(o_{c,i} == o_c^*)$$

Representation of labeling function

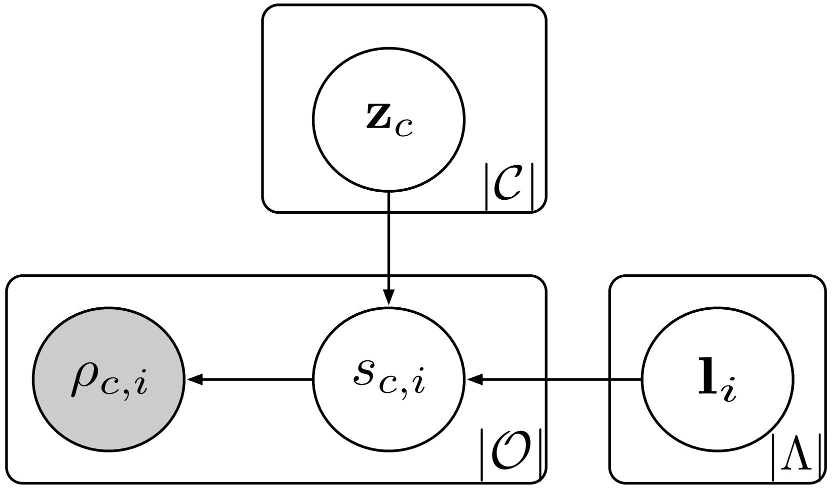


correctness of annotation $o_{c,i}$

whether c belongs to the proficient subset of $\mathbf{l_i}$

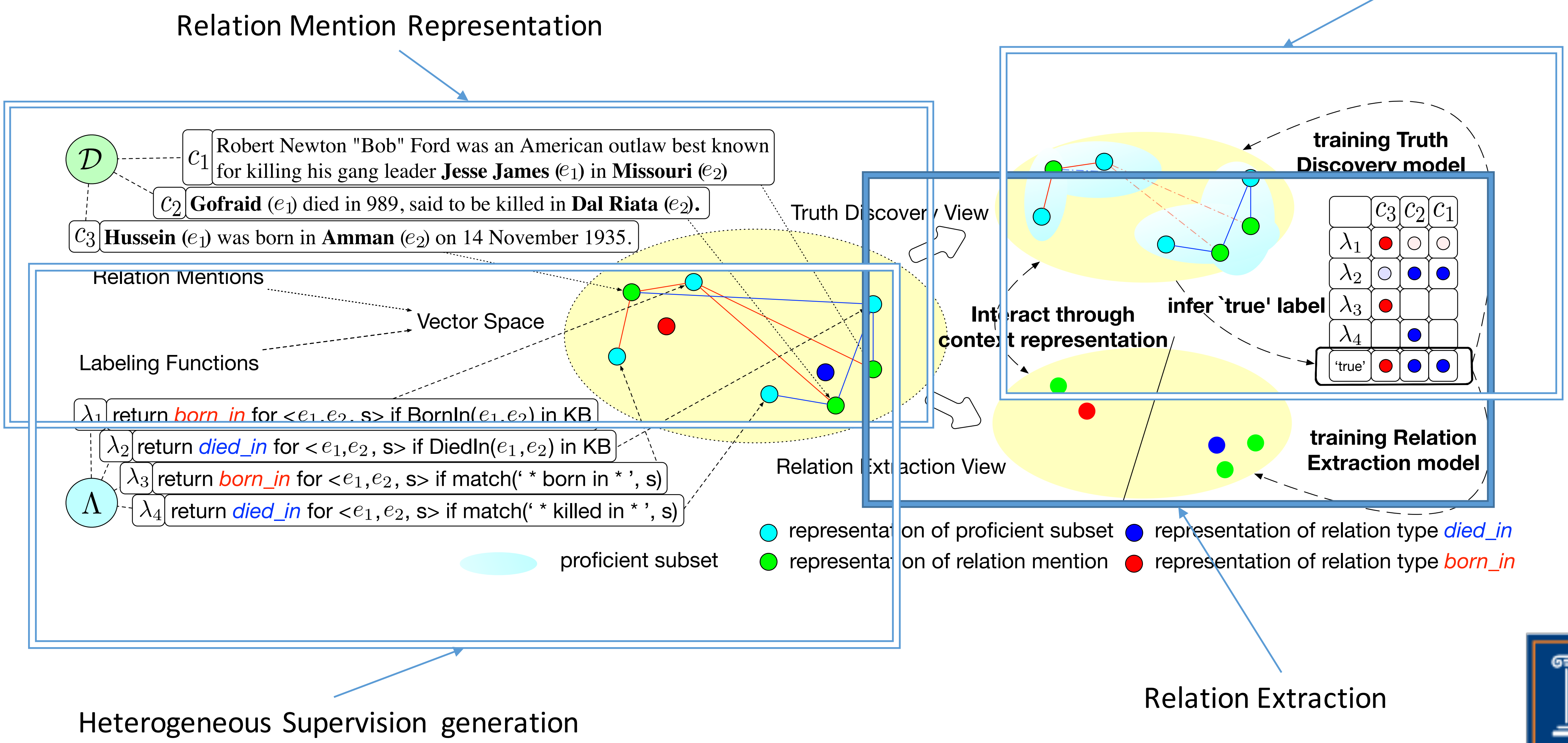
True label discovery

- Describing the correctness of Heterogeneous Supervision
  - $p(\rho_{c,i} = 1) = p(\rho_{c,i} = 1 | s_{c,i} = 1) * p(s_{c,i} = 1) + p(\rho_{c,i} = 1 | s_{c,i} = 0) * p(s_{c,i} = 0)$
  - $p(s_{c,i} = 1) = \sigma(\mathbf{l}_i^t * \mathbf{z}_c)$
  - $$\mathcal{J}_T = \sum_{o_{c,i} \in \mathcal{O}} \log(\sigma(\mathbf{z}_c^T \mathbf{l}_i) \phi_1^{\delta(o_{c,i} = o_c^*)} (1 - \phi_1)^{\delta(o_{c,i} \neq o_c^*)} + (1 - \sigma(\mathbf{z}_c^T \mathbf{l}_i)) \phi_0^{\delta(o_{c,i} = o_c^*)} (1 - \phi_0)^{\delta(o_{c,i} \neq o_c^*)})$$



A Representation Learning Approach

True Label Discovery



Relation Extraction

- Adopts soft-max as the relation extractor:

$$p(r_i|\mathbf{z}_c) = \frac{\exp(\mathbf{z}_c^T \mathbf{t}_i)}{\sum_{r_j \in \mathcal{R} \cup \{\texttt{None}\}} \exp(\mathbf{z}_c^T \mathbf{t}_j)}$$

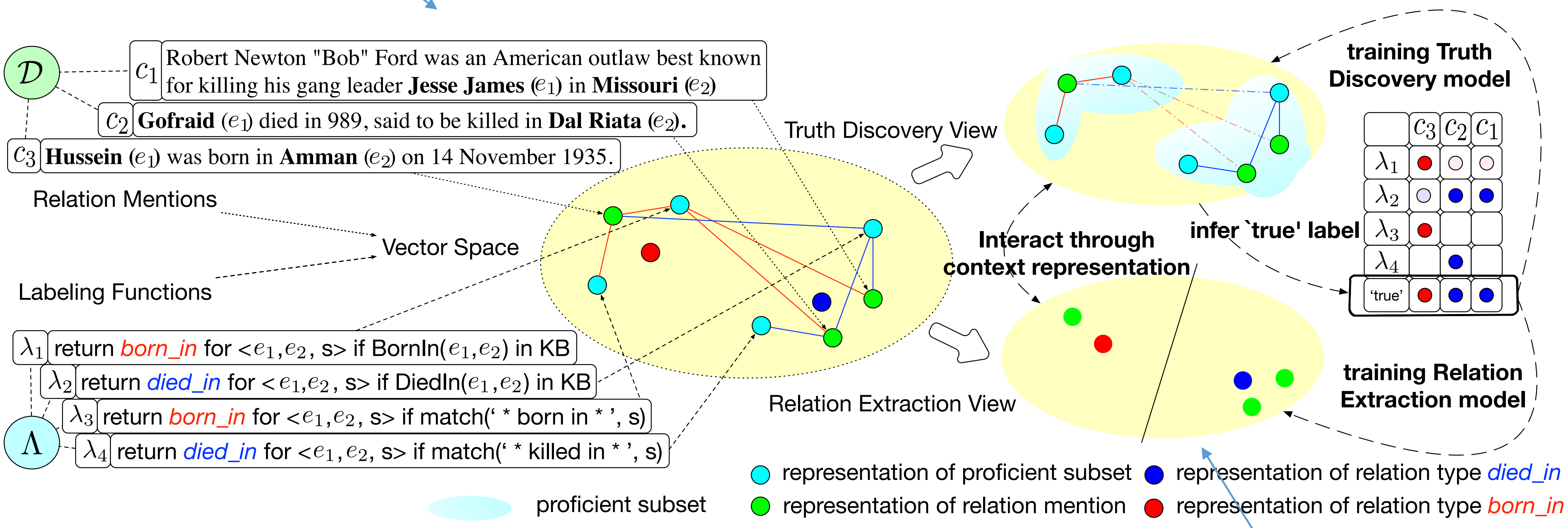
- Loss function: KL-Divergence:

$$\mathcal{J}_R = -\sum_{c \in \mathcal{C}_l} KL(p(.|\mathbf{z}_c)||p(.|o_c^*))$$

A Representation Learning Approach

True Label Discovery

Relation Mention Representation



Heterogeneous Supervision generation

Relation Extraction

Model Learning

- Joint optimize three components

$$\min_{W,\mathbf{v},\mathbf{v}^*,\mathbf{l},\mathbf{t},o^*} \mathcal{J} = -\mathcal{J}_R - \lambda_1 \mathcal{J}_E - \lambda_2 \mathcal{J}_T$$

$$\text{s.t. } \forall c \in \mathcal{C}_l, o_c^* = \operatorname*{argmax}_{o_c^*} \mathcal{J}_T, \mathbf{z}_c = g(\mathbf{f}_c)$$

ReHession

- Heterogeneous Supervision
- Our Solution: A Representation Learning Approach
  - Relation Mention Representation
  - True Label Discovery component
  - Relation Extraction component
- Experiments

Experiments

- 1. Relation extraction (with None) and Relation classification (without None):
  - NL: train relation extractor with all annotations
  - TD: train relation extractor with `true’ label inferred by Investment (compared true label discovery model)

Experiments

Method	Relation Extraction						Relation Classification	
	NYT			Wiki-KBP			NYT	Wiki-KBP
	Prec	Rec	F1	Prec	Rec	F1	Accuracy	Accuracy
NL+FIGER	0.2364	0.2914	0.2606	0.2048	0.4489	0.2810	0.6598	0.6226
NL+BFK	0.1520	0.0508	0.0749	0.1504	0.3543	0.2101	0.6905	0.5000
NL+DSL	0.4150	0.5414	0.4690	0.3301	0.5446	0.4067	0.7954	0.6355
NL+MultiR	0.5196	0.2755	0.3594	0.3012	0.5296	0.3804	0.7059	0.6484
NL+FCM	0.4170	0.2890	0.3414	0.2523	0.5258	0.3410	0.7033	0.5419
NL+CoType-RM	0.3967	0.4049	0.3977	**0.3701**	0.4767	0.4122	0.6485	0.6935
TD+FIGER	0.3664	0.3350	0.3495	0.2650	**0.5666**	0.3582	0.7059	0.6355
TD+BFK	0.1011	0.0504	0.0670	0.1432	0.1935	0.1646	0.6292	0.5032
TD+DSL	0.3704	0.5025	0.4257	0.2950	0.5757	0.3849	0.7570	0.6452
TD+MultiR	**0.5232**	0.2736	0.3586	0.3045	0.5277	0.3810	0.6061	0.6613
TD+FCM	0.3394	0.3325	0.3360	0.1964	0.5645	0.2914	0.6803	0.5645
TD+CoType-RM	0.4516	0.3499	0.3923	0.3107	0.5368	0.3879	0.6409	0.6890
REHESSION	0.4122	**0.5726**	**0.4792**	0.3677	0.4933	**0.4208**	**0.8381**	**0.7277**

Table 6: Performance comparison of relation extraction and relation classification

Experiments

- 2. Effectiveness of proposed true label discovery component:
  - Ori: with proposed context-aware true label discovery component
  - LD: with Investment (compared true label discovery model)

Dataset & Method		Prec	Rec	F1	Acc
Wiki-KBP	Ori	**0.3677**	0.4933	**0.4208**	**0.7277**
	TD	0.3032	**0.5279**	0.3850	0.7271
NYT	Ori	**0.4122**	**0.5726**	**0.4792**	**0.8381**
	TD	0.3758	0.4887	0.4239	0.7387

Table 7: Comparison between REHESSION (Ori) and REHESSION-TD (TD) on relation extraction and relation classification

Case Study

Relation Mention	**REHESSION**	**Investment**
Ann Demeulemeester (**born** 1959 , Waregem , *Belgium*) is a ...	born-in	None
Raila Odinga was **born** at ..., in *Maseno*, Kisumu District, ...	born-in	None
Ann Demeulemeester (**elected** 1959 , Waregem , *Belgium*) is a ...	None	None
Raila Odinga was **examined** at ..., in *Maseno*, Kisumu District, ...	None	None

Table 8: Example output of true label discovery. The first two relation mentions come from Wiki-KBP, and their annotations are {born-in, None}. The last two are created by replacing key words of the first two. Key words are marked as bold and entity mentions are marked as Italics.

Thank You

Q & A