

# Probabilistic Models for KG Construction

---

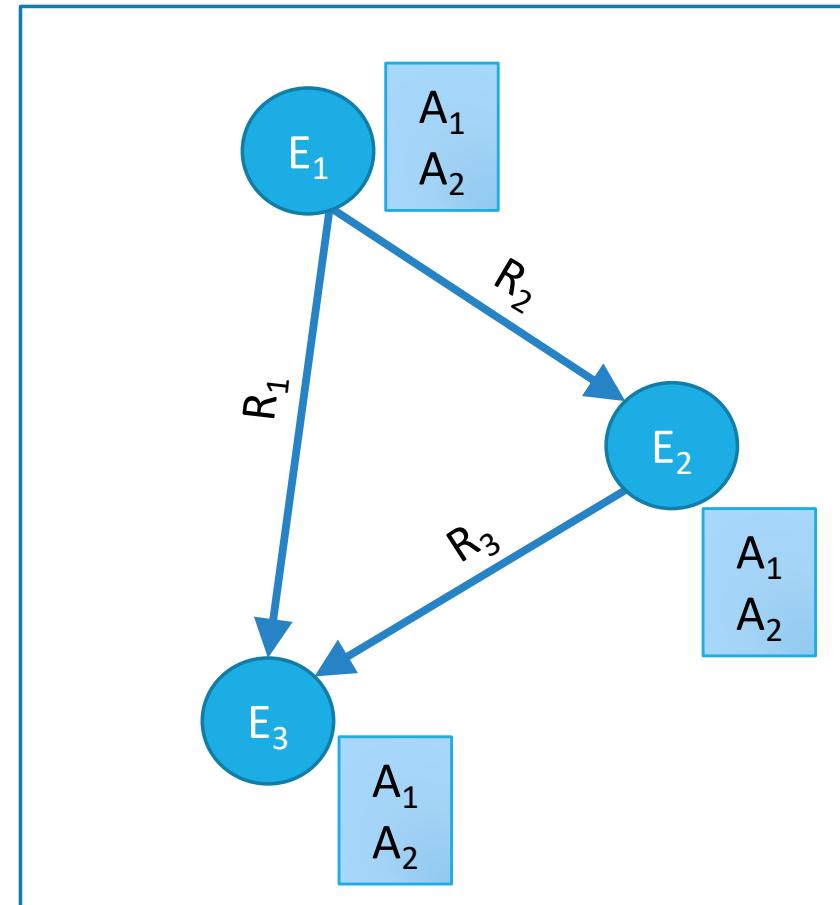
DSCI 558 – 3/3/21

JAY PUJARA

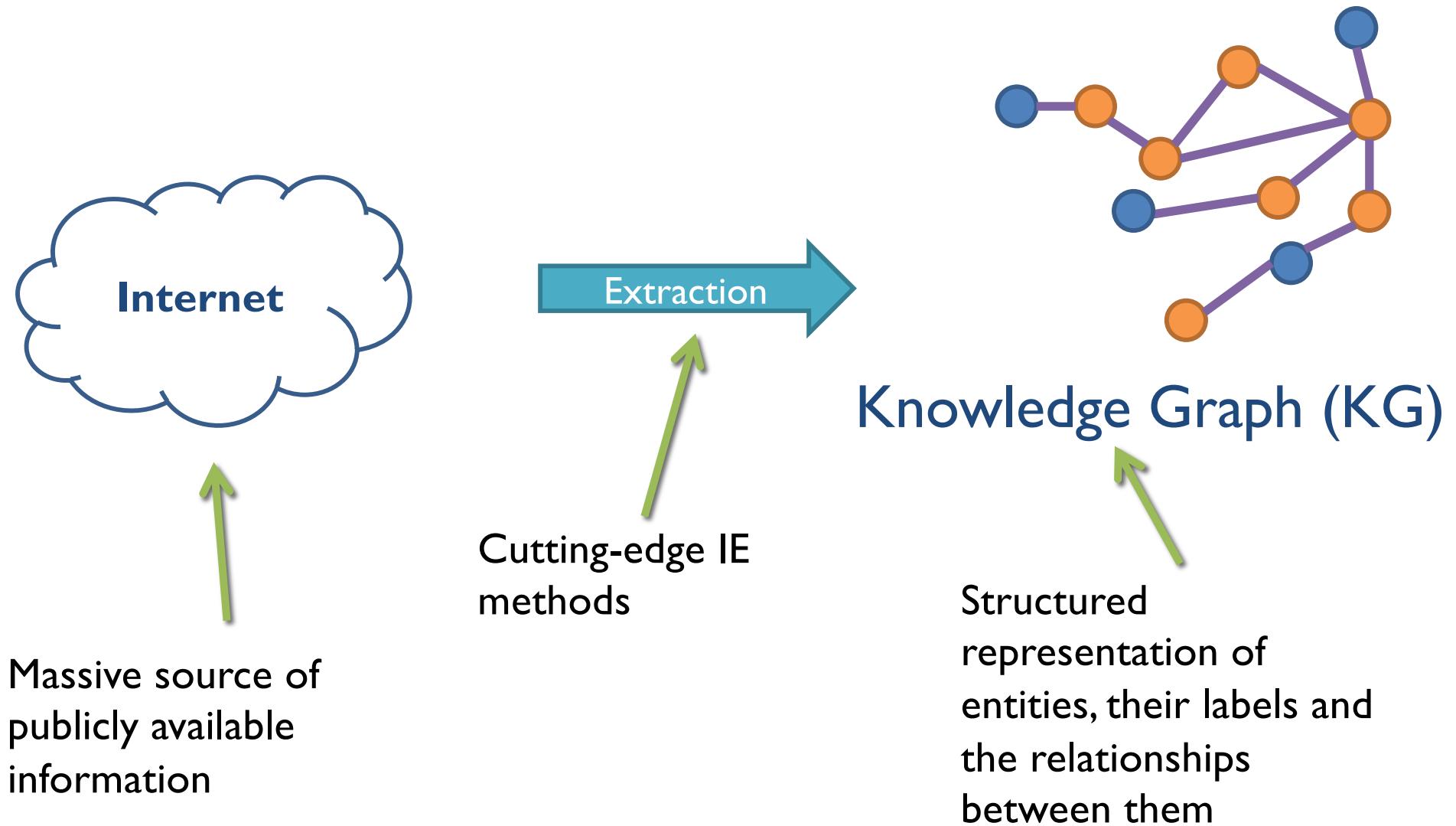
# Reminder: Basic problems

---

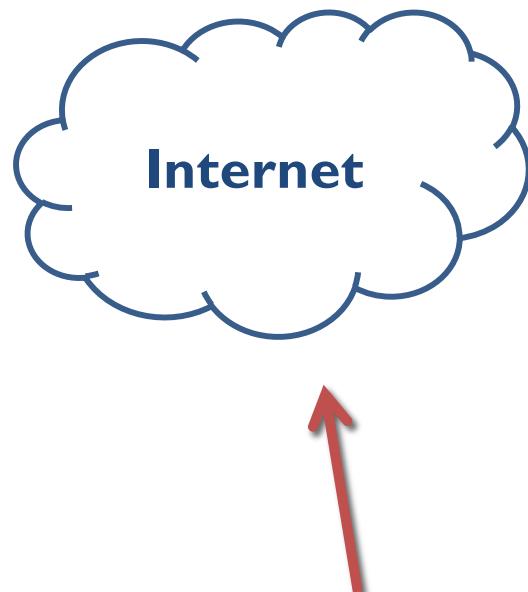
- **Who** are the entities (nodes) in the graph?
- **What** are their attributes and types (labels)?
- **How** are they related (edges)?



# Motivating Problem: New Opportunities



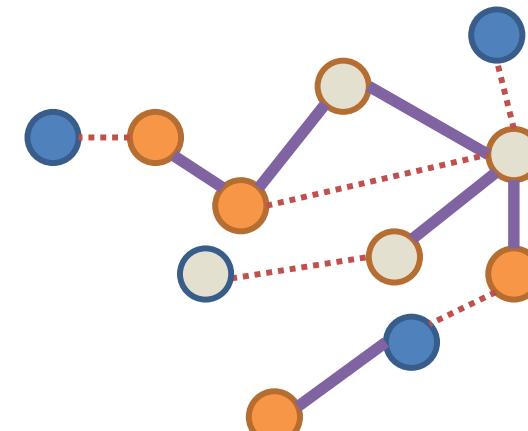
# Motivating Problem: Real Challenges



Noisy!



Difficult!



Knowledge Graph

Contains many errors  
and inconsistencies

# Graph Construction Issues

---

Extracted knowledge is:

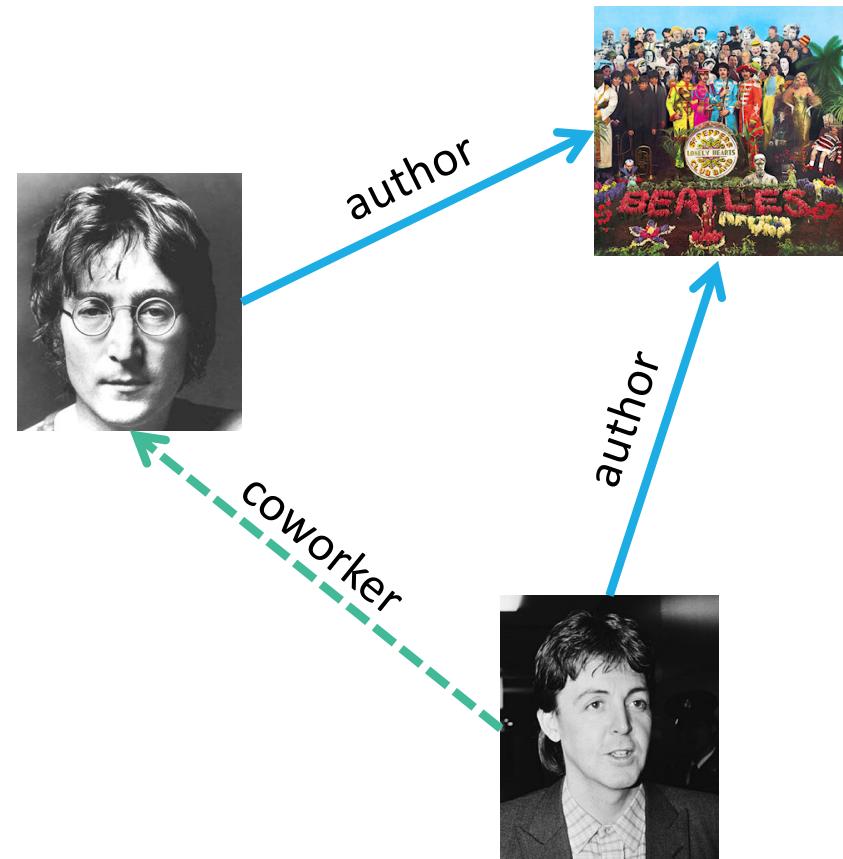
- ambiguous:
  - Ex: Beetles, beetles, Beatles
  - Ex: citizenOf, livedIn, bornIn



# Graph Construction Issues

Extracted knowledge is:

- ambiguous
- incomplete
  - Ex: missing relationships
  - Ex: missing labels
  - Ex: missing entities



# Graph Construction Issues

Extracted knowledge is:

- ambiguous
- incomplete
- inconsistent
  - Ex: Cynthia Lennon, Yoko Ono
  - Ex: exclusive labels (alive, dead)
  - Ex: domain-range constraints



spouse



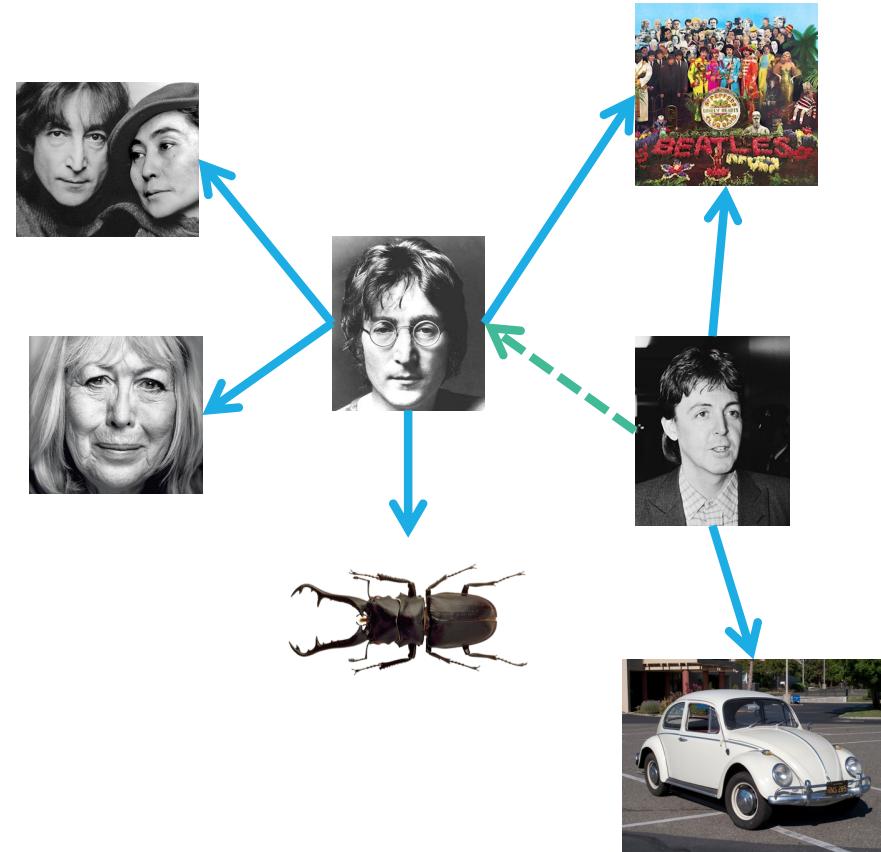
spouse



# Graph Construction Issues

Extracted knowledge is:

- ambiguous
- incomplete
- inconsistent



# NELL: The Never-Ending Language Learner

NELL @cmunell  
True or False? "kevn tv" is a #TVStation (bit.ly/18JQ8gs)  
Expand

1 Oct

NELL @cmunell  
True or False? "metro-Atlanta" is a #County (bit.ly/1hhsefI)  
Expand

1 Oct

NELL @cmunell  
True or False? "exclusive right" is an #Artery (bit.ly/1bZq2LA)  
Expand

1 Oct

NELL @cmunell  
True or False? "Fireplace" is #SomethingFoundInOrOnBuildings  
(bit.ly/17E1JhW)  
Expand

1 Oct

Reply Retweet Favorite More

NELL @cmunell  
True or False? "will\_whalen" is an #AustralianPerson (bit.ly/1fUzRdT)  
Expand

1 Oct

NELL @cmunell  
True or False? "iron\_chair" is a #HouseholdItem (bit.ly/14ZsCNk)  
Expand

30 Sep

NELL @cmunell  
True or False? "jerry gordon" is a #Chef (bit.ly/19Ry4QN)  
Expand

30 Sep

- Large-scale IE project  
(Carlson et al., 2010)

- Lifelong learning: aims to  
“read the web”

- Ontology of known  
labels and relations

- Knowledge base  
contains millions of facts

- person
  - monarch
  - astronaut
- personbylocation
  - personnorthamerica
    - personcanada
    - personus
    - politicianus
    - personmexico
  - personeurope
  - personaaustralia
  - personafrica
  - personsouthamerica
  - personasia
  - personantarctica
- visualartist
- model
- scientist
- journalist
- female
- actor
- professor
- director
- architect
- politician
  - politicianus
- musician
- athlete
- chef
- male
- writer
- ceo
- judge
- mlauthor
- coach
- celebrity
- comedian
- criminal

# Examples of NELL errors

# Entity co-reference errors

Kyrgyzstan has many variants:

- Kyrgystan
- Kyrgistan
- Kyrghyzstan
- Kyrgzstan
- Kyrgyz Republic

Saudi Cultural Days in the **Kyrgyz Republic** has concluded its activities in the capital Bishkek in the weekend in a special ceremony held on this occasion. The event was attended by Deputy Minister of Culture and Tourism of the **Kyrgyz Republic** Koulev Mirza; Kyrgyzstan's Ambassador to Saudi Arabia Jusupbek Sharipov; the Saudi Embassy Acting Chargé d'affaires to Kyrgyzstan, Mari bin Barakah Al-Derbas and members of the embassy staff, in the presence of a heavy turnout of Kyrgyz citizens.

The Days of Culture of Saudi Arabia in **Kyrgyzstan** will be held from 6 to 9 May.

[Home](#) > [Holiday Destinations](#) > [Kyrgyzstan](#) > [Bishkek](#) > Climate Profile



**Fast Forecast**

**Holiday Weather**

Refugees are often from areas where conflict is historically embedded and marked in ideology and injustice. The Tsarnaev family emigrated from the Chechen diaspora in **Kyrgyzstan**, a region Stalin deported the Chechens to in 1943. After the fall of the Berlin Wall in 1991, Chechens engaged in a battle for independence from Russia that led to the Tsarnaevs' petition for refugee status in the early

# Missing and spurious labels

[Anssi Kullberg](#) has sent along some great trip reports to unusual places, including [Kyrgyzstan](#), [Pakistan](#), [Egypt/Jordan](#), and [Afghanistan](#). I had to create a whole new country page for [Afghanistan](#) to hold that last one! Thanks so much, Anssi!

[Erik Kleyheeg](#) has just returned from Lesvos with some new bird images. Included here are: [Common Scops-Owl](#), [Wood Warbler](#), [Spanish Sparrow](#), [Red-throated Pipit](#), [Eurasian Chiff-chaff](#), and [Cretzschmar's Bunting](#).

**Kyrgyzstan** (/kɜrgɪ'sta:n/ *kur-gi-STAHN*;<sup>[5]</sup> Kyrgyz: Кыргызстан (IPA: [қырғызстан]); Russian: Киргизия), officially the **Kyrgyz Republic** (Kyrgyz: Кыргыз Республикасы; Russian: Кыргызская Республика), is a country located in Central Asia.<sup>[6]</sup> Landlocked and mountainous, Kyrgyzstan is bordered by Kazakhstan to the north, Uzbekistan to the west, Tajikistan to the southwest and China to the east. Its capital and largest city is Bishkek.

Kyrgyzstan is labeled a bird and a country

# Missing and spurious relations

Guidance

## Kazakhstan / Kyrgyzstan – Consular Fees

Organisation: Foreign & Commonwealth Office  
Page history: Published 4 April 2013

Kyrgyzstan's location is ambiguous – Kazakhstan, Russia and US are included in possible locations

## Kyrgyzstan U.S. Air Base Future Unclear

A Central Asian country of incredible natural beauty and proud nomadic traditions, most of Kyrgyzstan was formally annexed to Russia in 1876. The Kyrgyz staged a major revolt against the Tsarist Empire in 1916 in which almost one-sixth of the Kyrgyz population was killed. Kyrgyzstan became a Soviet republic in 1936 and

# Violations of ontological knowledge

- Equivalence of co-referent entities (`sameAs`)
  - `SameEntity(Kyrgyzstan, Kyrgyz Republic)`
- Mutual exclusion (`disjointWith`) of labels
  - `MUT(bird, country)`
- Selectional preferences (domain/range) of relations
  - `RNG(countryLocation, continent)`

Enforcing these constraints require **jointly** considering multiple extractions

# Graph Construction approach

---

- Graph construction **cleans** and **completes** extraction graph
- Incorporate ontological constraints and relational patterns
- Discover statistical relationships within knowledge graph

# Graph Construction Probabilistic Models

---

TOPICS:

OVERVIEW

GRAPHICAL MODELS

RANDOM WALK METHODS

# Graph Construction Probabilistic Models

---

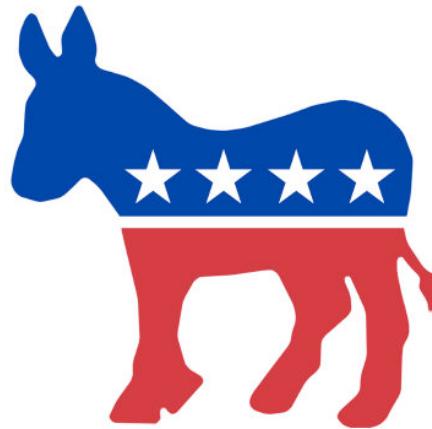
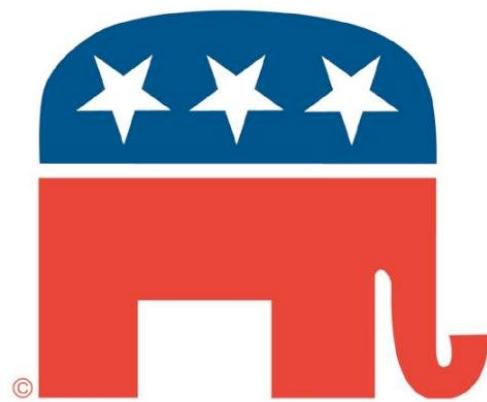
TOPICS:

OVERVIEW

GRAPHICAL MODELS

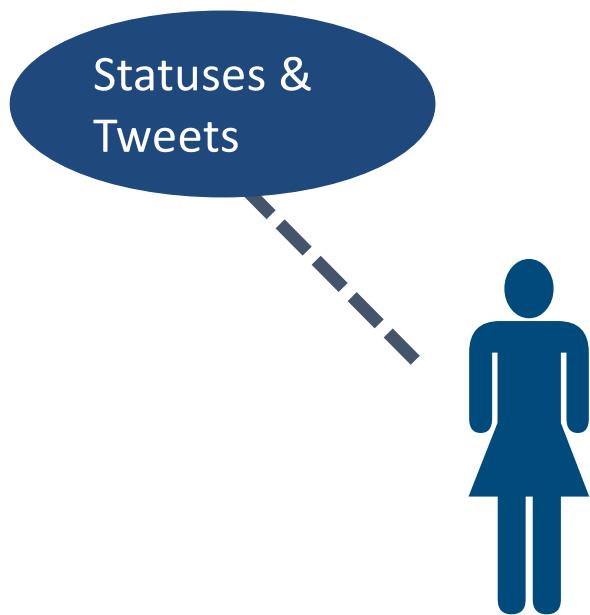
RANDOM WALK METHODS

# Voter Party Classification



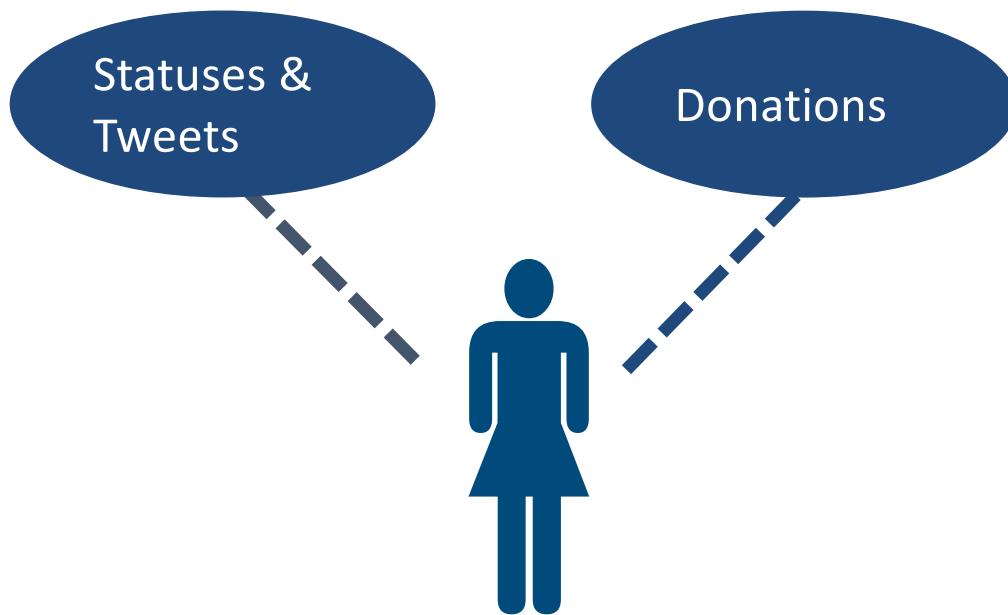
# Voter Party Classification

## Multiple Sources of Information



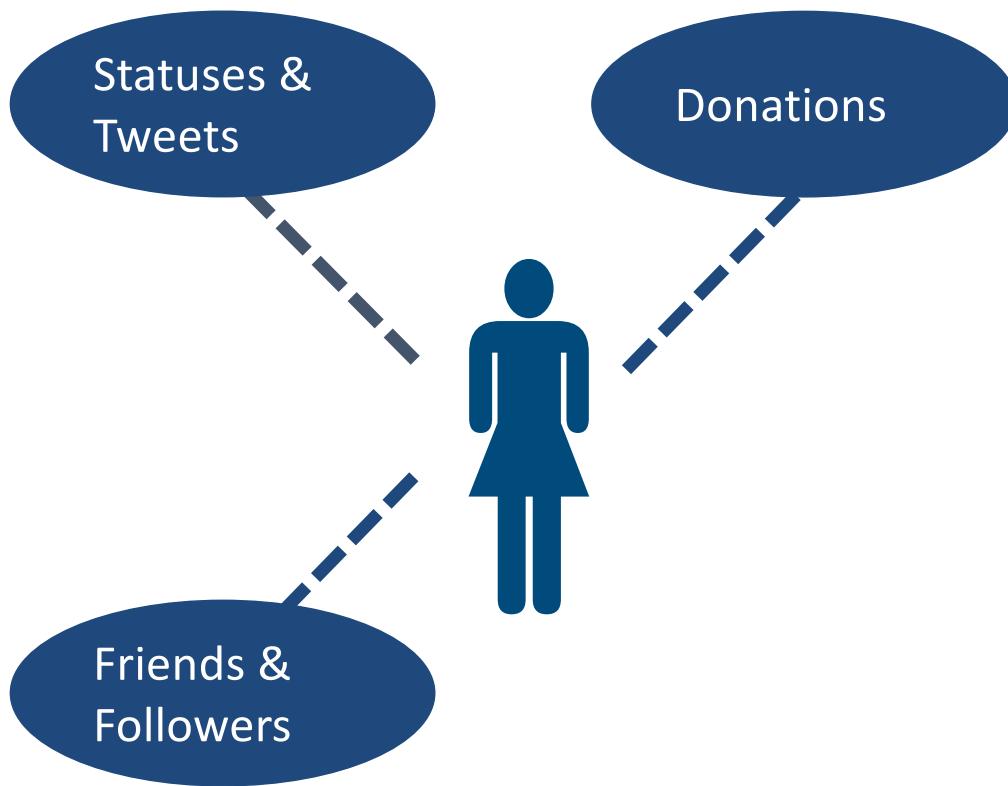
# Voter Party Classification

**Multiple Sources of Information**



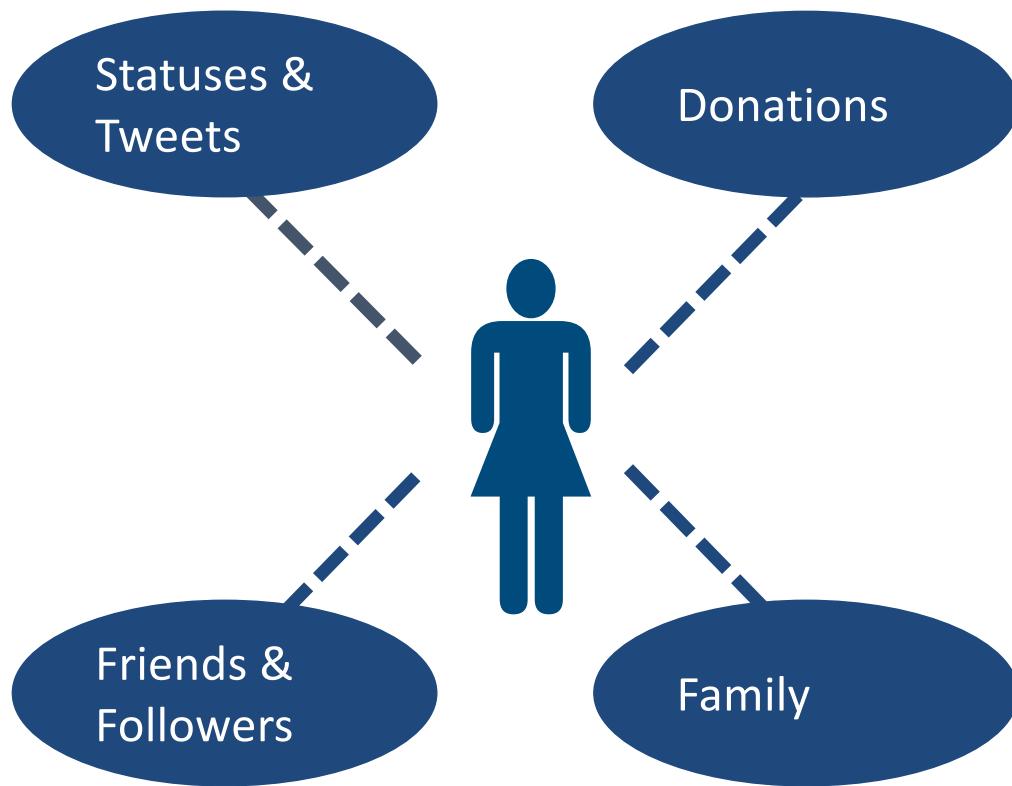
# Voter Party Classification

## Multiple Sources of Information



# Voter Party Classification

## Multiple Sources of Information



# Voter Party Classification



**Tim Long**   
@mrtimlong



 Follow

Forget Trump & his 100's of accusers — this new Hillary REVELATION will turn the race UPSIDE DOWN!!!

**Eyewitness News** @ABC7NY

Clinton Team Ran Highly Scripted Campaign, WikiLeaks Emails Indicate  
[7ny.tv/2deCcrL](http://7ny.tv/2deCcrL)

RETWEETS

8

LIKES

44



11:16 AM - 15 Oct 2016



 8

 44

...

# Voter Party Classification



Reply to @mrtimlong



**sliderwave** @sliderwave · Oct 15

@mrtimlong @ABC7NY OMG! Someone in the campaign was ready and prepared because it's the most important job on Earth? Why would they do that?



2

...



**Emily Fun Buns** @MsEffieLou · Oct 15

@mrtimlong almost as bad as her sympathizing with Bernie supporters and not calling them Basement Dwellers



1

...



**((((Voter))) Eitan** @AnotherEitan · Oct 15

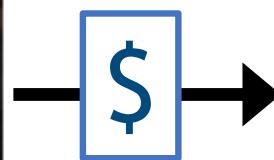
@mrtimlong @ABC7NY Holy shit. I'm starting to think that she WANTS to be President.



1

...

# Voter Party Classification



→ CarlyFiorinaforVicePresident.com



**Tim Long** 

@mrtimlong

Simpsons Writer/ Man of Peace

📍 Los Angeles

📅 Joined December 2010

# Voter Party Classification



**Tim Long**

@mrtimlong

Simpsons Writer/ Man of Peace

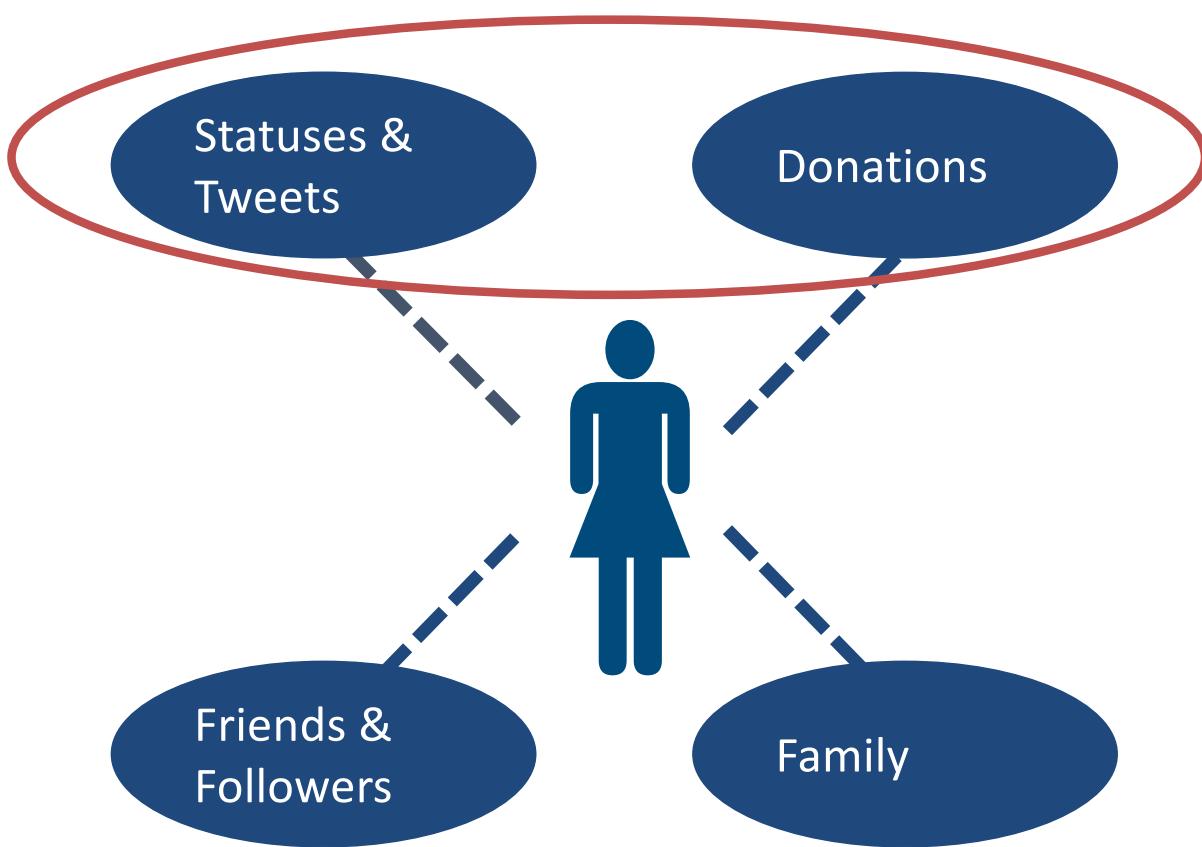
Los Angeles

Joined December 2010



# Voter Party Classification

## Multiple Sources of Information



# Standard Classification



**Tim Long**

@mrtimlong

Simpsons Writer/ Man of Peace

Los Angeles

Joined December 2010

Forget Trump & his 100's of accusers — this new Hillary REVELATION will turn the race UPSIDE DOWN!!!

**Eyewitness News @ABC7NY**

Clinton Team Ran Highly Scripted Campaign, WikiLeaks Emails Indicate  
[7ny.tv/2deCcrL](https://7ny.tv/2deCcrL)

CarlyFiorinaforVicePresident.com



Bag-of-words features

# Standard Classification



**Tim Long**

@mrtimlong

Simpsons Writer/ Man of Peace

📍 Los Angeles

📅 Joined December 2010

Forget Trump & his 100's of accusers — this new Hillary REVELATION will turn the race UPSIDE DOWN!!!

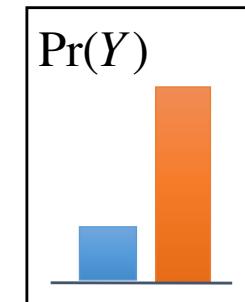
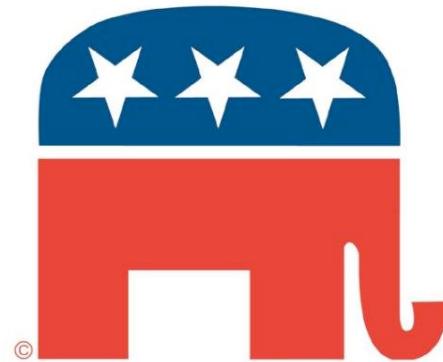
Eyewitness News @ABC7NY

Clinton Team Ran Highly Scripted Campaign, WikiLeaks Emails Indicate  
[7ny.tv/2deCcrL](https://7ny.tv/2deCcrL)

CarlyFiorinaforVicePresident.com



Bag-of-words features



# Standard Classification



**Tim Long**

@mrtimlong

Simpsons Writer/ Man of Peace

📍 Los Angeles

📅 Joined December 2010

Forget Trump & his 100's of accusers — this new Hillary REVELATION will turn the race UPSIDE DOWN!!!

Eyewitness News @ABC7NY

Clinton Team Ran Highly Scripted Campaign, WikiLeaks Emails Indicate  
[7ny.tv/2deCcrL](http://7ny.tv/2deCcrL)

CarlyFiorinaforVicePresident.com

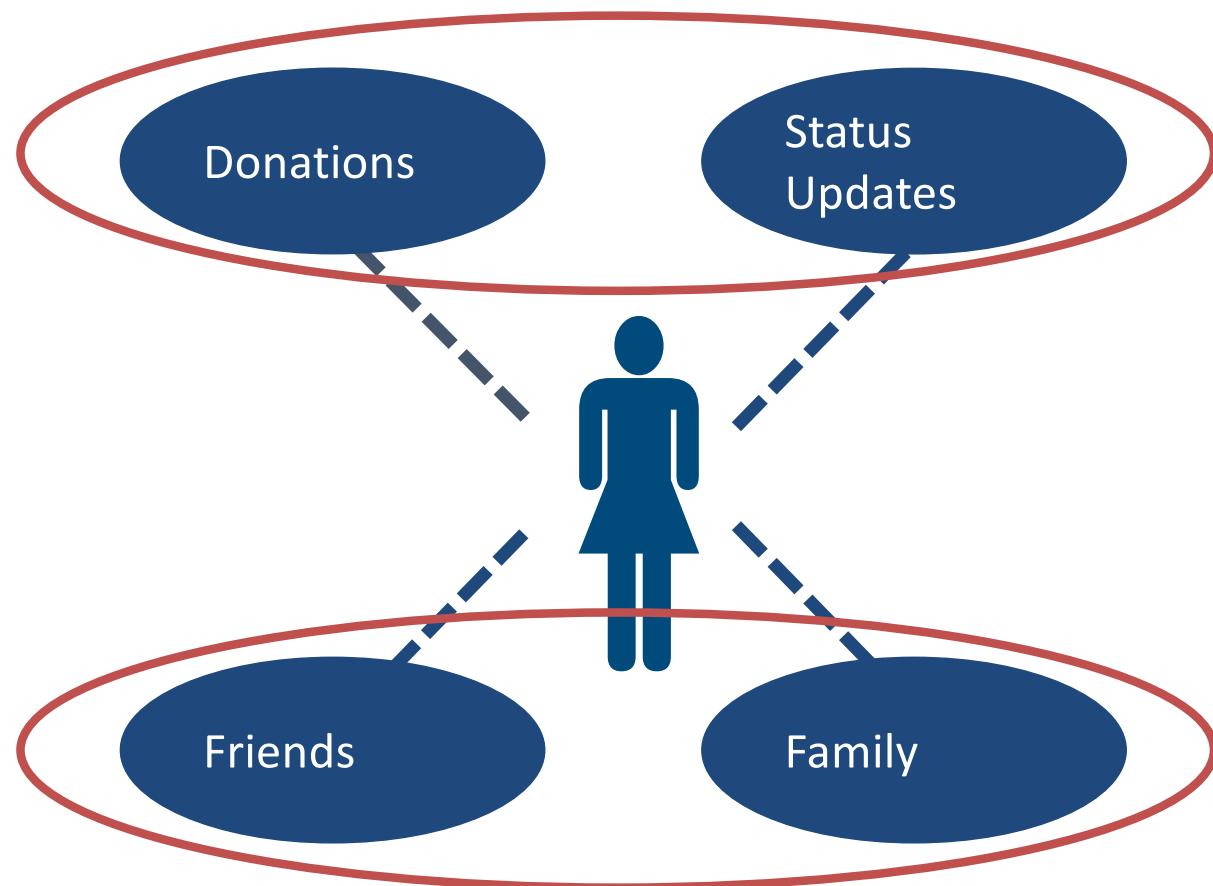


Bag-of-words features



# Voter Party Classification

## Multiple Sources of Information



# Collective Classification



**Tim Long** 

@mrtimlong

Simpsons Writer/ Man of Peace

 Los Angeles

 Joined December 2010

Follows



**Terri Lee Dee**

@terrileedee

# Collective Classification



Follows



Terri Lee Dee Retweeted



RhysHRCTays @iRhysTay · May 21

The most qualified, optimistic, skilled, and competent candidate for POTUS. She will never quit! #ImWithHer

Tin  
@mrt

Simpsons Writer/ Man of Peace

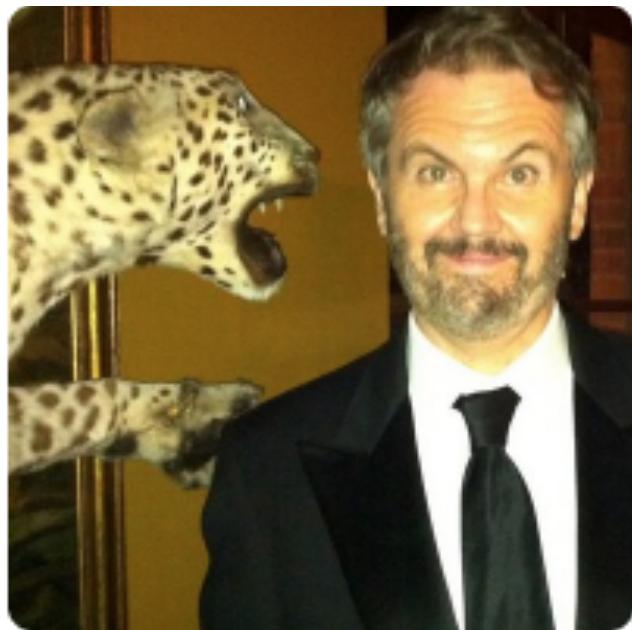
Los Angeles

Joined December 2010

Terri Lee Dee

@terrileedee

# Collective Classification



**Tim Long**

@mrtimlong

Simpsons Writer/ Man of Peace

📍 Los Angeles

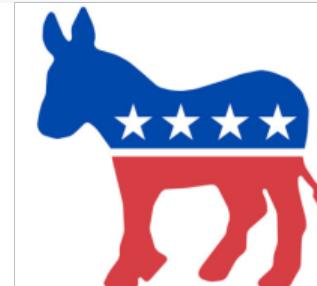
📅 Joined December 2010

Follows

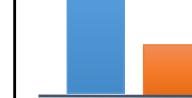


**Terri Lee Dee**

@terrileedee



$\Pr(Y)$



# Collective Classification



← Follows



My label is likely to match that of my follower

**Tim Long**

@mrtimlong

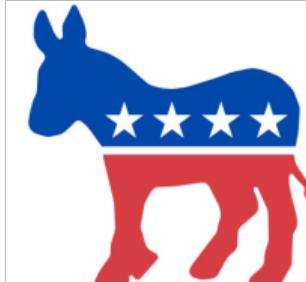
Simpsons Writer/ Man of Peace

📍 Los Angeles

📅 Joined December 2010

**Terri Lee Dee**

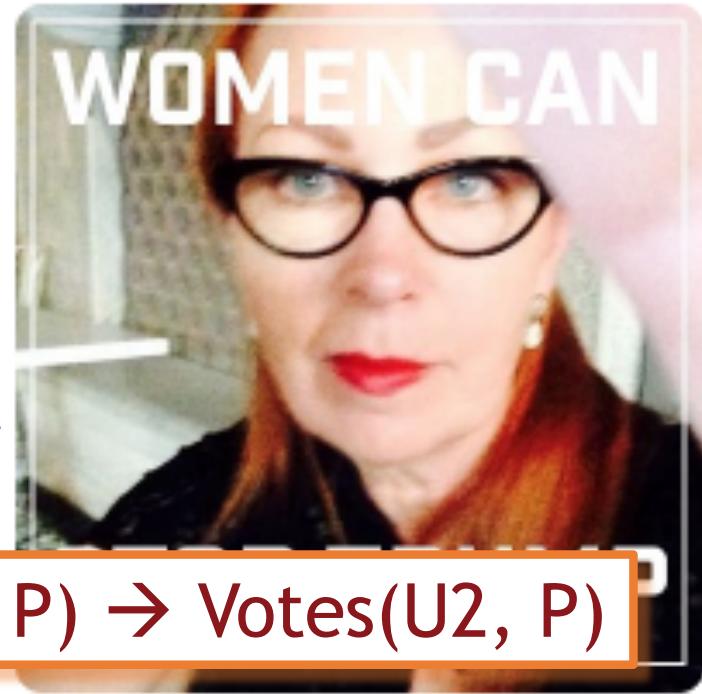
@terrileedee



# Collective Classification



Follows



Follows(U<sub>1</sub>, U<sub>2</sub>) & Votes(U<sub>1</sub>, P) → Votes(U<sub>2</sub>, P)

**Tim Long**

@mrtimlong

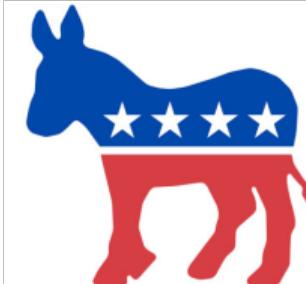
Simpsons Writer/ Man of Peace

Los Angeles

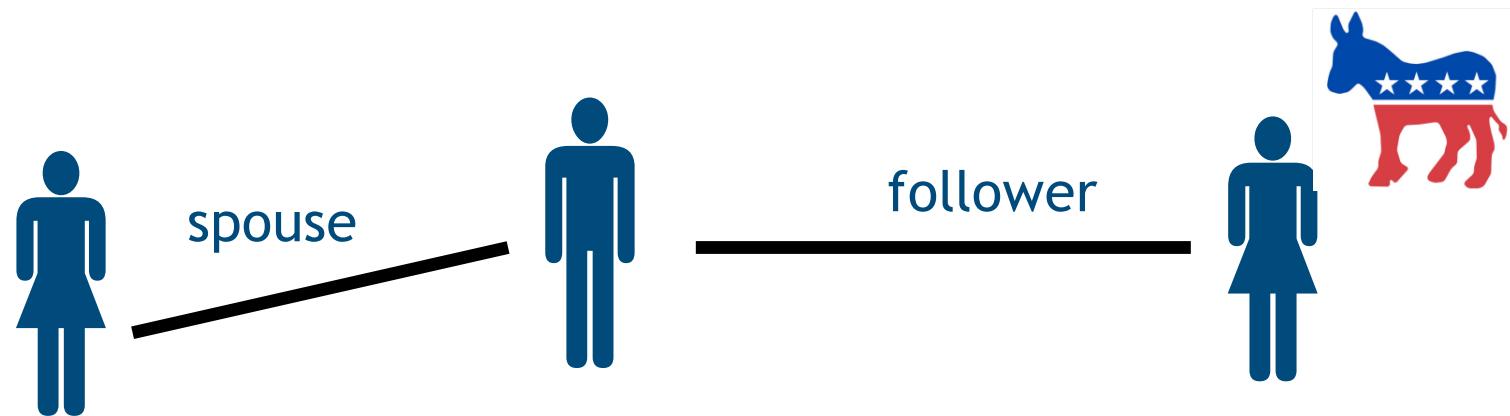
Joined December 2010

**Terri Lee Dee**

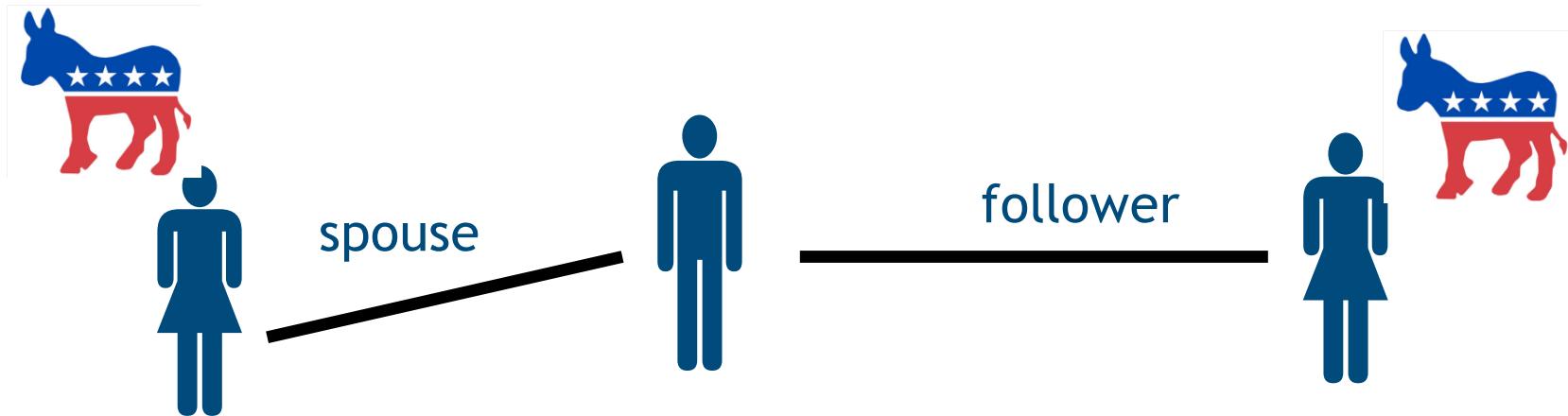
@terrileedee



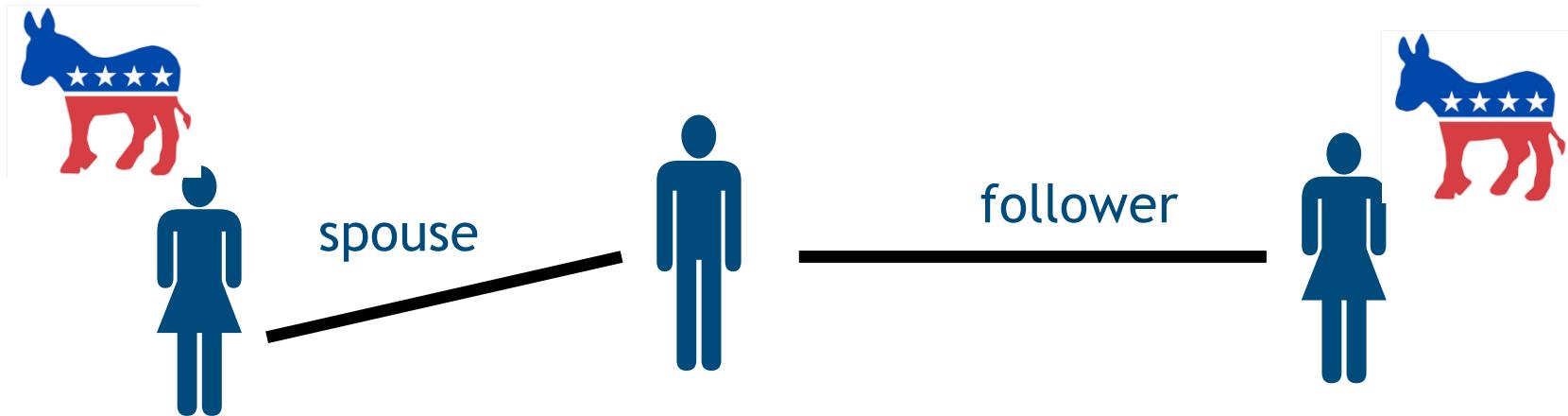
# Collective Classification



# Collective Classification



# Collective Classification



$\text{Spouse}(U_1, U_2) \& \text{Votes}(U_1, P) \rightarrow \text{Votes}(U_2, P)$

$\text{Follows}(U_1, U_2) \& \text{Votes}(U_1, P) \rightarrow \text{Votes}(U_2, P)$

# Collective Classification



**Tim Long**   
@mrtimlong



 Follow

Wait don't forget that Hillary caught a bug & stayed home for THREE DAYS



Reply to @mrtimlong



**Werewalker** @warwalker · Oct 7



@mrtimlong Don't blame me, I voted for Kodos.



1

...



**Jake Schroeder** @bluejake14star · Oct 7

@mrtimlong They both have problems.



...

# Collective Classification



Tim Long   
@mrtimlong



 Follow

Wait don't forget that Hillary caught a bug & stayed home for THREE DAYS



Reply to @mrtimlong



Werewalker @warwalker · Oct 7

@mrtimlong Don't blame me, I voted for Kodos.



1

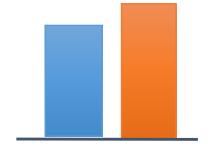


Jake Schroeder @bluejake14star · Oct 7

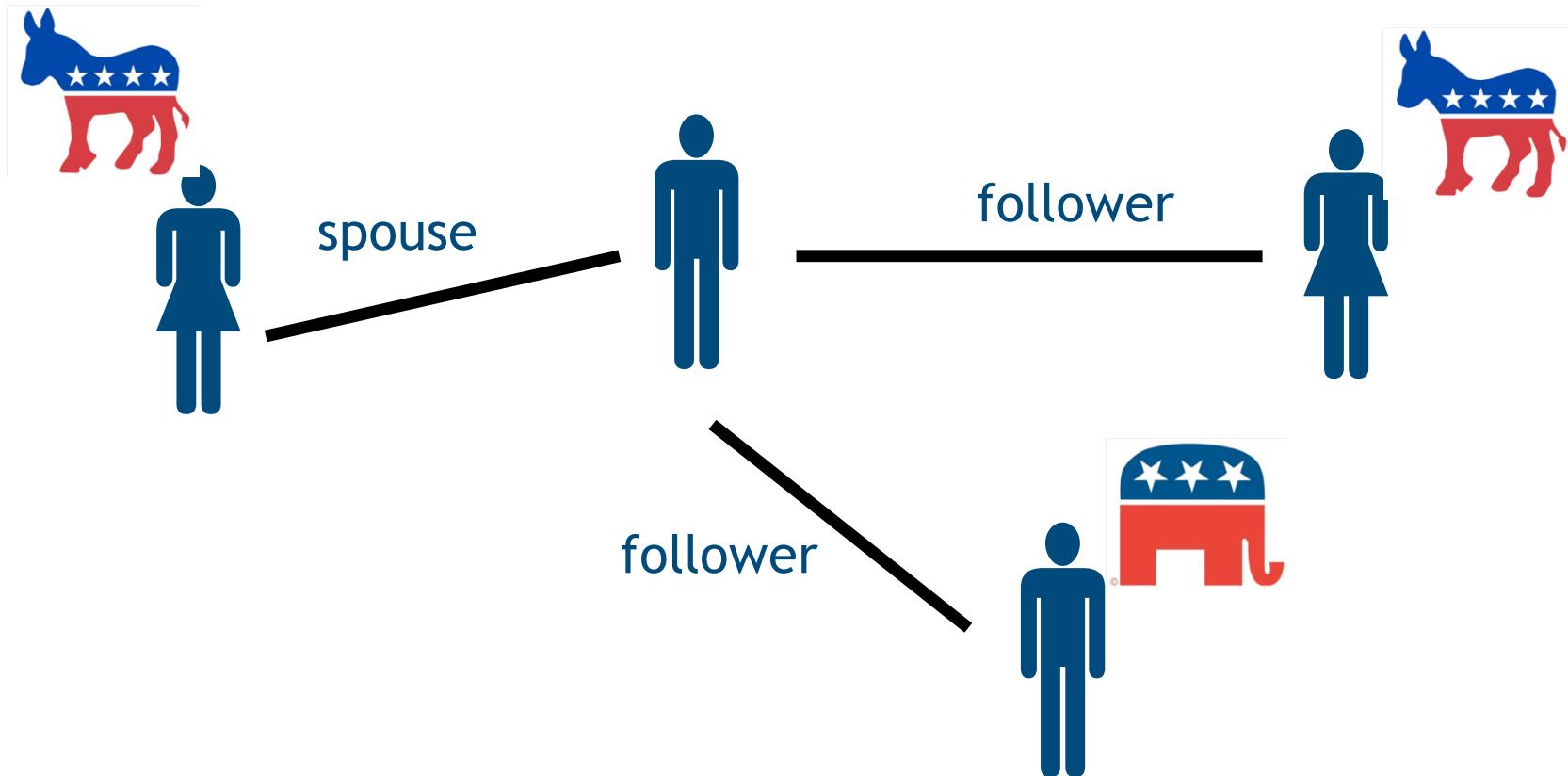
@mrtimlong They both have problems.



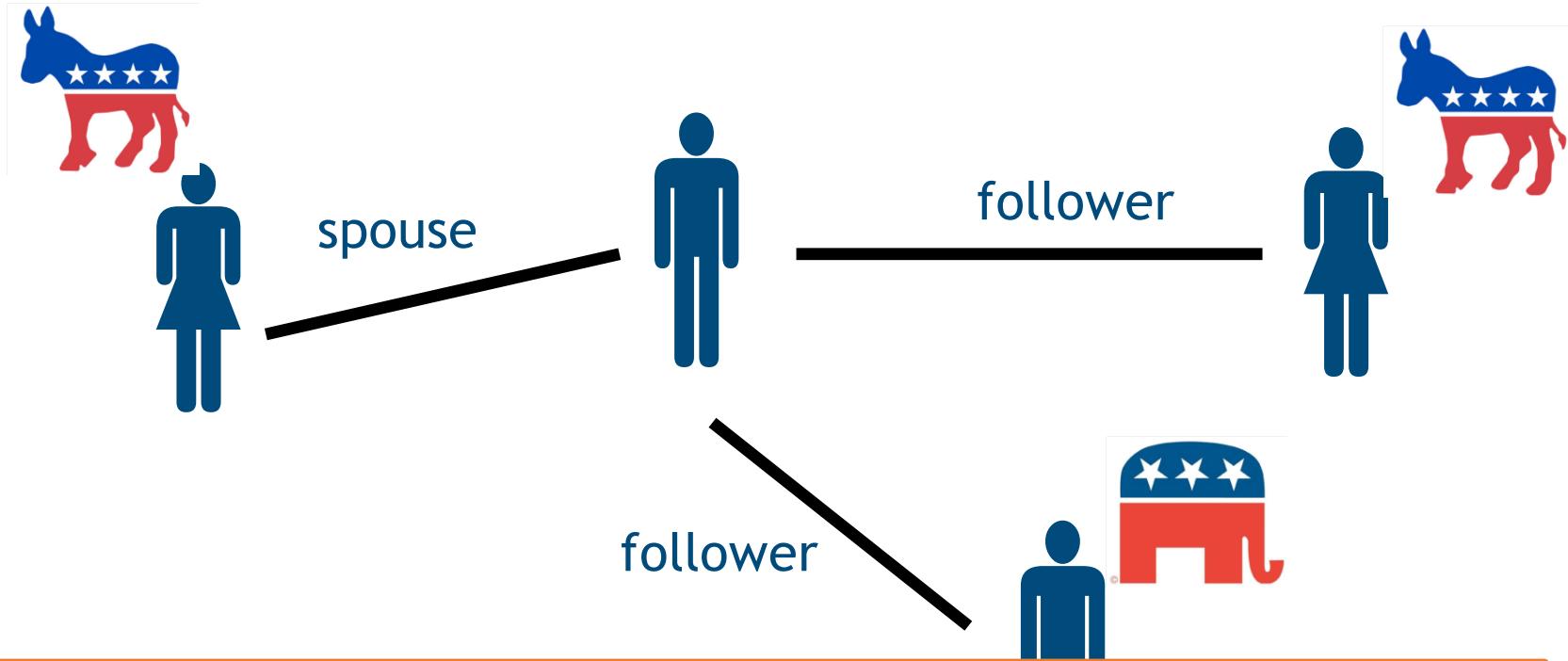
$\Pr(Y)$



# Collective Classification



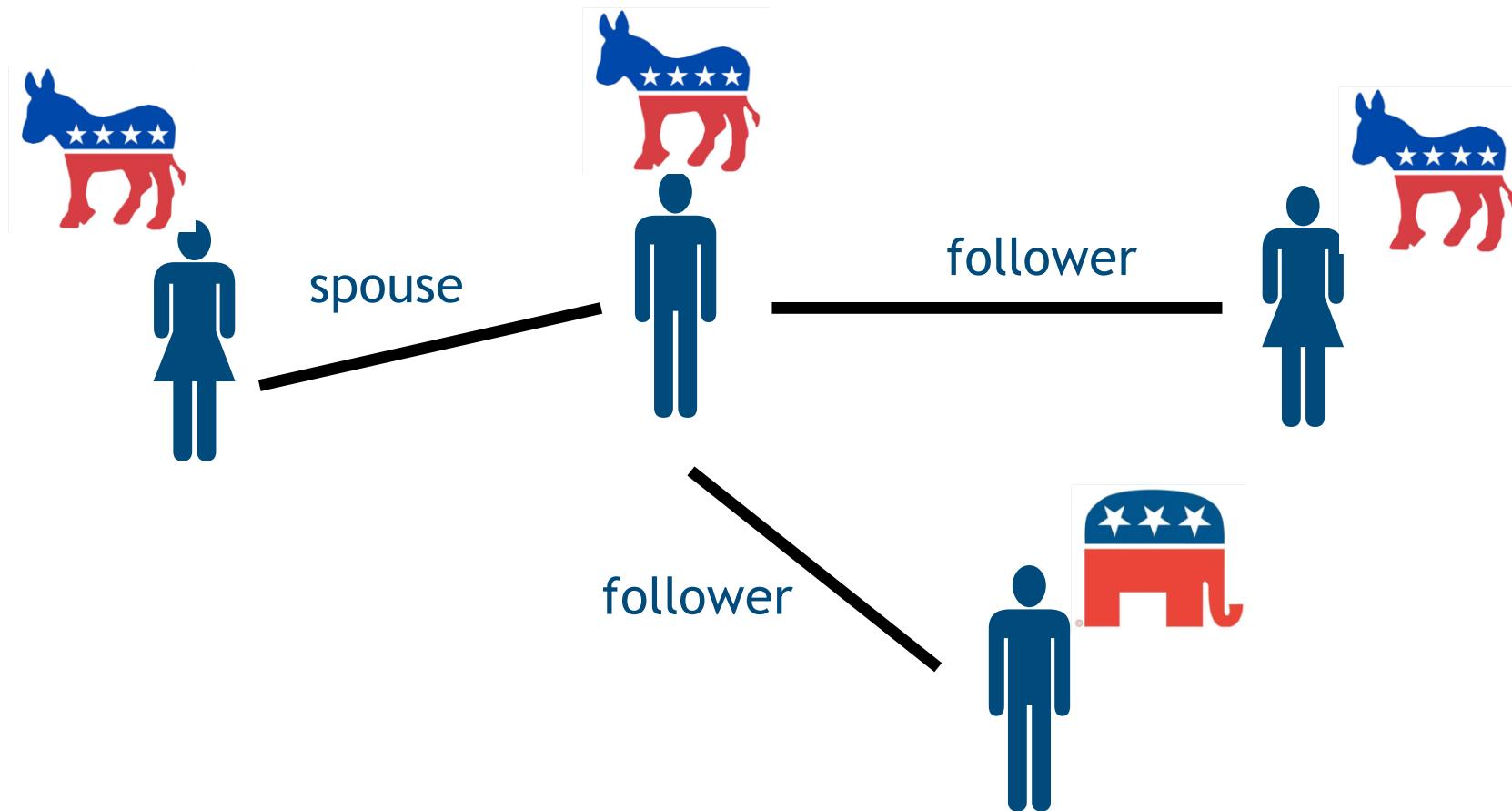
# Collective Classification



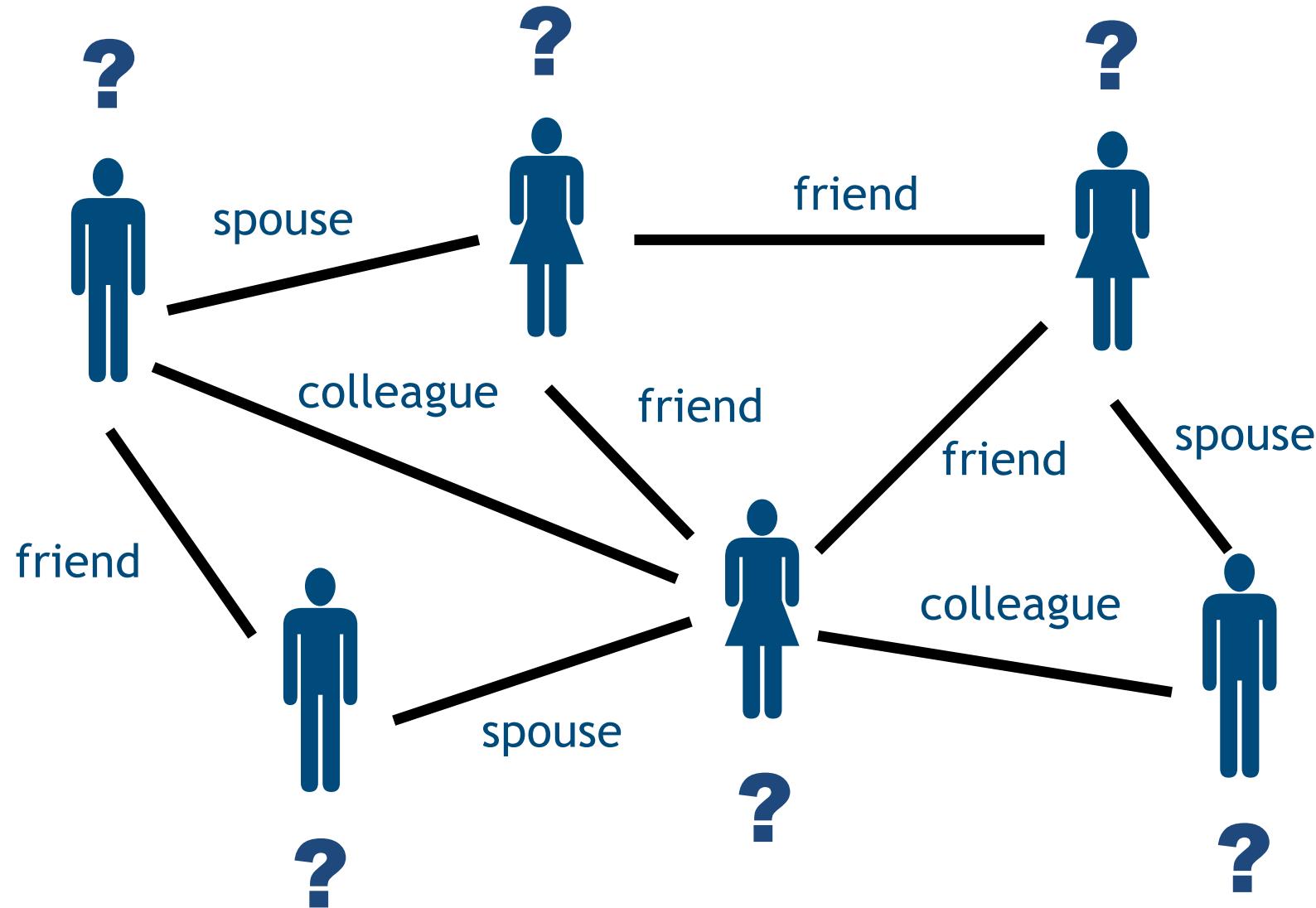
5.0: Spouse(U1, U2) ^& Votes(U1, P) → Votes(U2, P)

2.0: Follows(U1, U2) & Votes(U1, P) → Votes(U2, P)

# Collective Classification



# Collective Classification



# Collective Classification with PSL

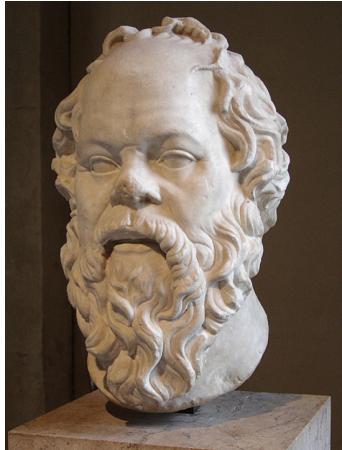
```
/* Local rules */
5.0: Donates(A, P) -> Votes(A, P)
0.3: Mentions(A, "Affordable Health") -> Votes(A, "Democrat")
0.3: Mentions(A, "Tax Cuts") -> Votes(A, "Republican")

/* Relational rules */
1.0: Votes(A,P) & Spouse(B,A) -> Votes(B,P)
0.3: Votes(A,P) & Friend(B,A) -> Votes(B,P)
0.1: Votes(A,P) & Colleague(B,A) -> Votes(B,P)

/* Range constraint */
Votes(A, "Republican") + Votes(A, "Democrat") = 1.0 .
```

# Beyond Pure Reasoning

---

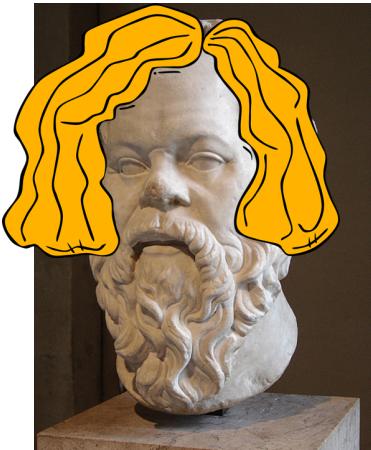


- Classical AI approach to knowledge: reasoning

$\text{Lbl}(\text{Socrates}, \text{Man}) \& \text{Sub}(\text{Man}, \text{Mortal}) \rightarrow \text{Lbl}(\text{Socrates}, \text{Mortal})$

# Beyond Pure Reasoning

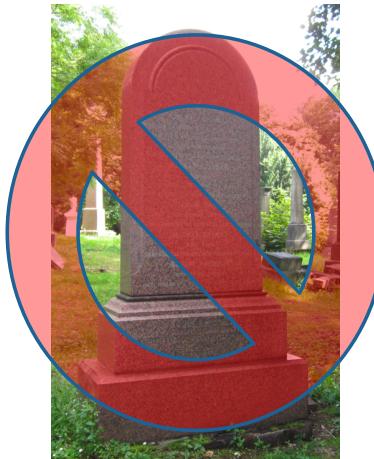
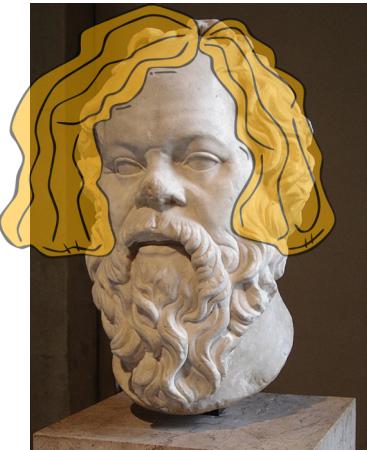
---



- Classical AI approach to knowledge: reasoning  
 $\text{Lbl}(\text{Socrates}, \text{Man}) \ \& \ \text{Sub}(\text{Man}, \text{Mortal}) \rightarrow \text{Lbl}(\text{Socrates}, \text{Mortal})$
- Reasoning difficult when extracted knowledge has errors

# Beyond Pure Reasoning

---



- Classical AI approach to knowledge: reasoning  
 $\text{Lbl}(\text{Socrates}, \text{Man}) \& \text{Sub}(\text{Man}, \text{Mortal}) \rightarrow \text{Lbl}(\text{Socrates}, \text{Mortal})$
- Reasoning difficult when extracted knowledge has errors
- Solution: probabilistic models  
 $P(\text{Lbl}(\text{Socrates}, \text{Mortal}) | \text{Lbl}(\text{Socrates}, \text{Man})) = 0.9$

# Logic Refresher: Satisfaction

```
/* Model Snippet */  
Mentions(A, "Affordable Health") -> Votes(A, "Democrat")
```

Affordable Health	Democrat	Logical Satisfaction
TRUE	TRUE	
TRUE	FALSE	
FALSE	TRUE	
FALSE	FALSE	

# Logic and Noisy Data

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Logical Satisfaction	[2] Logical Satisfaction
TRUE	TRUE	TRUE		
TRUE	TRUE	FALSE		

# Logic and Noisy Data

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Logical Satisfaction	[2] Logical Satisfaction
TRUE	TRUE	TRUE	😊	😢
TRUE	TRUE	FALSE	😢	😊

In logic, much as in politics, it is hard to satisfy everyone

# Soft Logic to the Rescue!

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Logical Satisfaction	[2] Logical Satisfaction
TRUE	TRUE	0.5	😐	😐
TRUE	TRUE	0.5	😐	😐

What does 0.5 MEAN?

# What does 0.5 mean?

- Rounding probability:
  - Flip a coin with bias 0.5
  - Heads = **TRUE**
  - Tails = **FALSE**
- **Using this method is a  $\frac{3}{4}$  optimal solution to the NP hard weighted MAX SAT problem**  
[Goemans&Williams, 94]

What does 😐 MEAN?

# What does 😐 mean?

P -> Q

- /\* Soft Logic Penalty \*/
- if P < Q  
    return 😊
- else:
- return P-Q

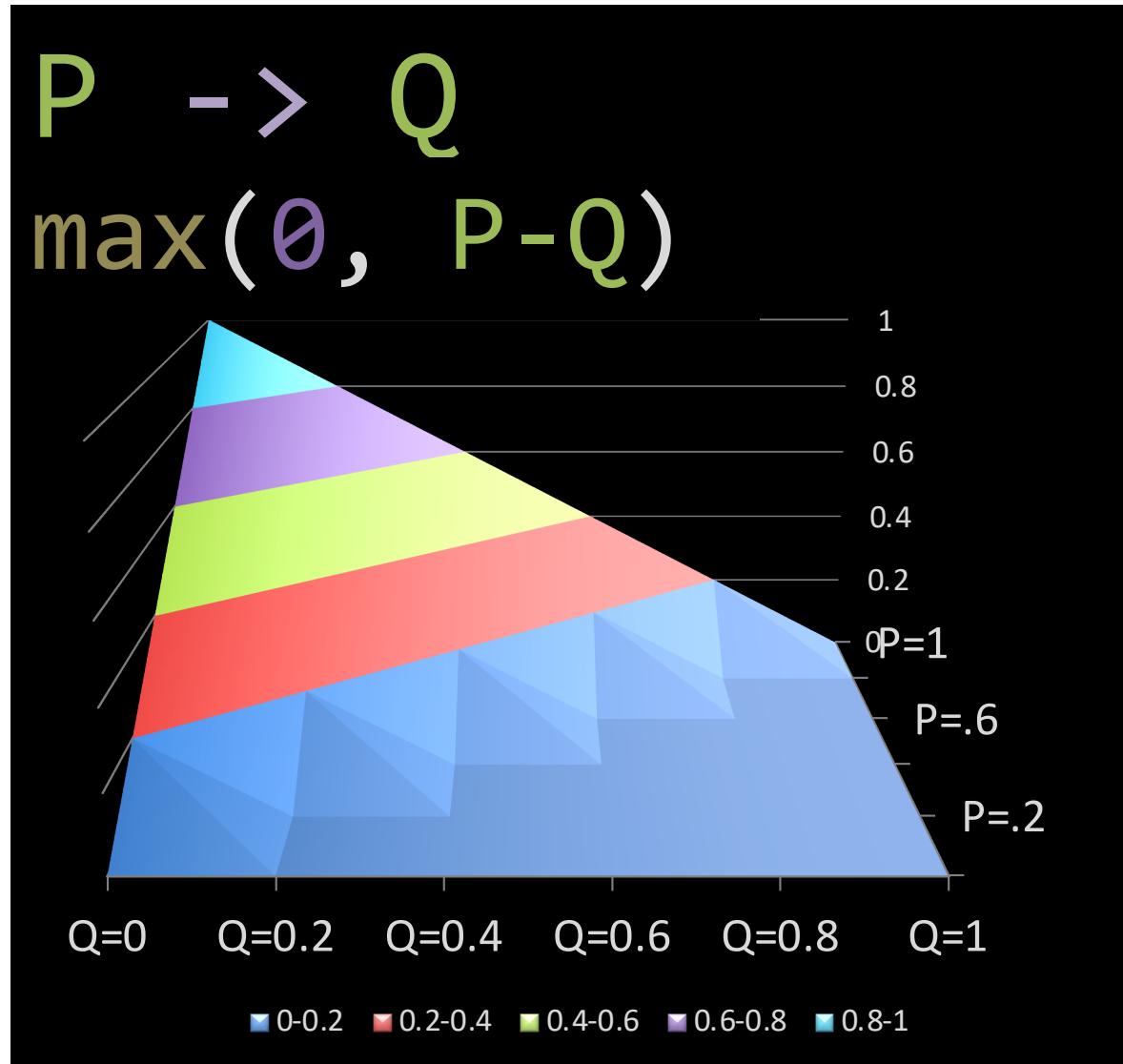


: Closed Form

P -> Q

• max(θ, P-Q)

😐: Closed Form



# What does 😐 mean?

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

```
/* Soft Logic Penalty */  
  
if Mentions(A, "Tax Cuts") < !Votes(A, "Democrat"):  
    return 0  
else:  
    return Mentions(A, "Tax Cuts") - !Votes(A, "Democrat")
```

# Computing 😐

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Penalty	[2] Penalty
1	1	0.7		
1	1	0.2		

$$!Q = 1 - Q$$

$$P \rightarrow Q = \max(\theta, P - Q)$$

# Computing 😐

```
/* Model Snippet */  
[1] Mentions(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Mentions(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Penalty	[2] Penalty
1	1	0.7	0.3	0.7
1	1	0.2	0.8	0.2

$$!Q = 1 - Q$$

$$P \rightarrow Q = \max(\theta, P - Q)$$

# Computing 😐 with soft evidence

```
/* Model Snippet */  
[1] Supports(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Supports(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Penalty	[2] Penalty
0.4	0.1	0.65		
0.4	0.1	0.2		
0.4	0.1	0.9		

$$!Q = 1 - Q$$

$$P \rightarrow Q = \max(\theta, P - Q)$$

# Computing 😐 with soft evidence

```
/* Model Snippet */  
[1] Supports(A, "Affordable Health") -> Votes(A, "Democrat")  
[2] Supports(A, "Tax Cuts")           -> !Votes(A, "Democrat")
```

Affordable Health	Tax Cuts	Democrat	[1] Penalty	[2] Penalty
0.4	0.1	0.65	0	0
0.4	0.1	0.2	0.2	0
0.4	0.1	0.9	0	0

$$!Q = 1 - Q$$

$$P \rightarrow Q = \max(0, P - Q)$$

# From penalties to truth values

```
/* Soft Logic Penalty */  
P -> Q = max(θ, P-Q)
```

```
/* Soft Logic Truth Value */  
P -> Q = 1-max(θ, P-Q)  
= min(1-θ, 1-(P-Q))  
= min(1, 1-P+Q)
```

# Computing truth values for arbitrary formulas

$$\text{!}Q = \text{1-Q}$$

$$P \rightarrow Q = \min(1, \text{1-P+Q})$$

$$P \& Q = \max(0, P+Q-1)$$

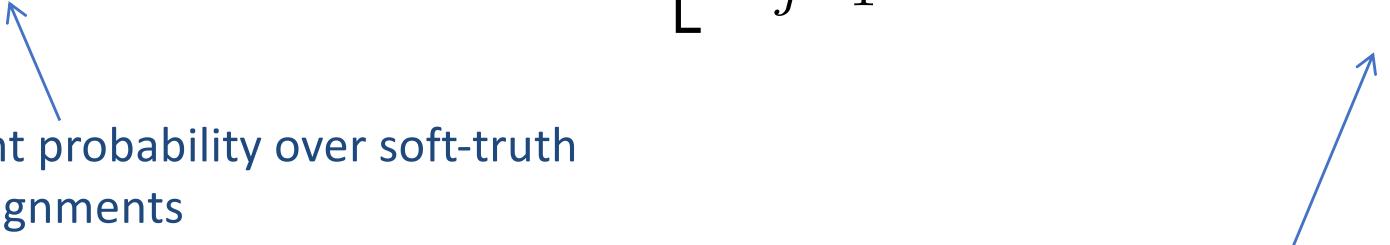
$$P \mid Q = \min(1, P+Q)$$

# Underlying Probability Distribution

$$p(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(w, \mathbf{X})} \exp \left[ - \sum_{j=1}^m w_j \left[ \max \{\ell_j(\mathbf{Y}, \mathbf{X}), 0\} \right]^{\{1,2\}} \right]$$

Joint probability over soft-truth assignments

Sum over rule penalties



# Optimizing PSL models

- PSL finds optimal assignment for all unknowns
- Optimal = minimizes the soft-logic penalty
- **Fast**, joint convex optimization using ADMM
- Supports learning rule weights and latent variables

# Graph Construction Probabilistic Models

---

TOPICS:

OVERVIEW

GRAPHICAL MODELS

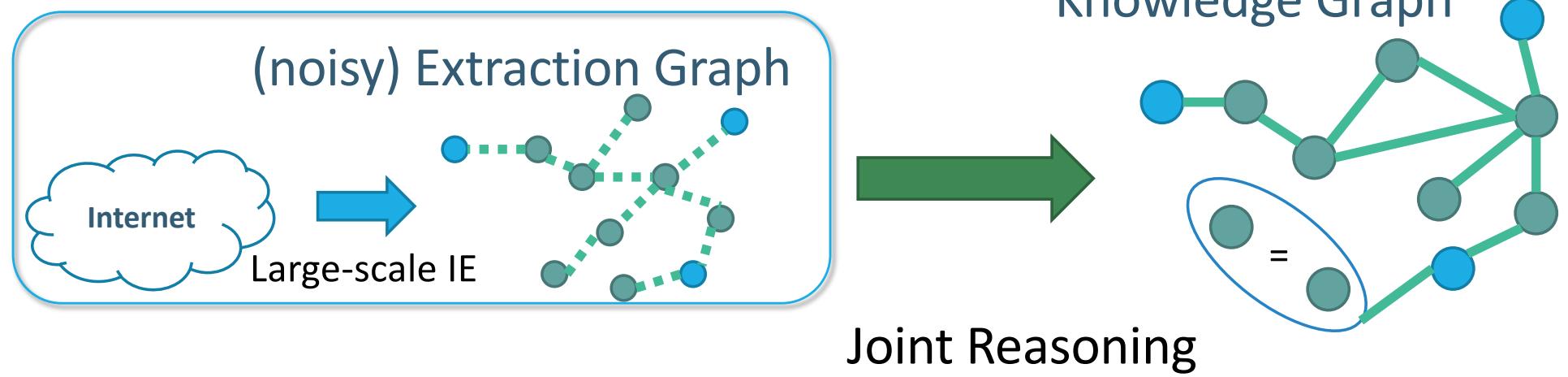
RANDOM WALK METHODS

# Graphical Models: Overview

---

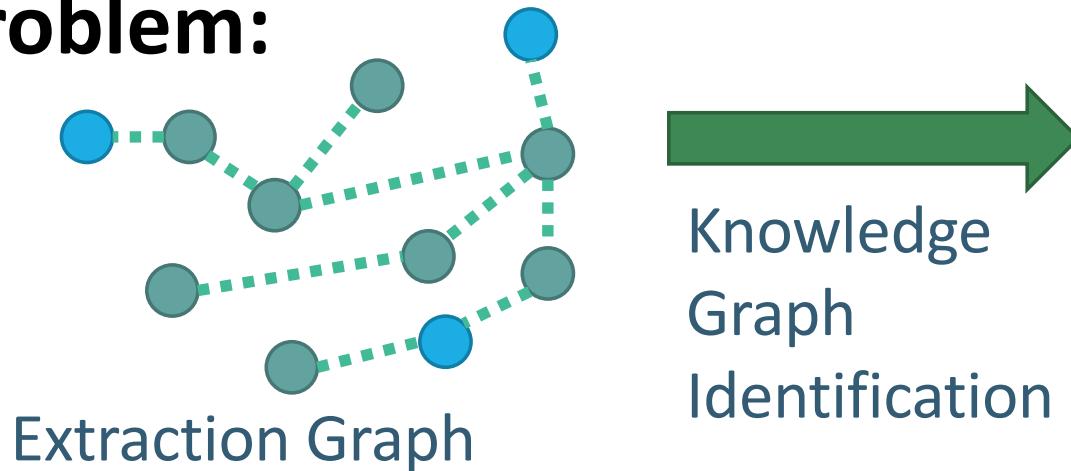
- Define **joint probability distribution** on knowledge graphs
- Each candidate fact in the knowledge graph is a **variable**
- Statistical signals, ontological knowledge and rules parameterize the **dependencies** between variables
- Find most likely knowledge graph by **optimization/sampling**

# Motivating Problem (revised)

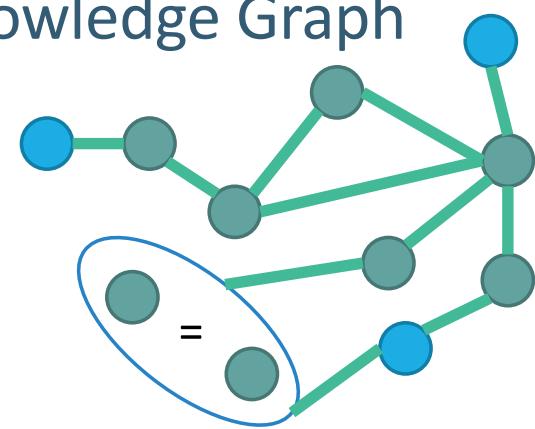


# Knowledge Graph Identification

## Problem:



## Knowledge Graph



## Solution: *Knowledge Graph Identification (KGI)*

Performs *graph identification*:

- entity resolution
- collective classification
- link prediction

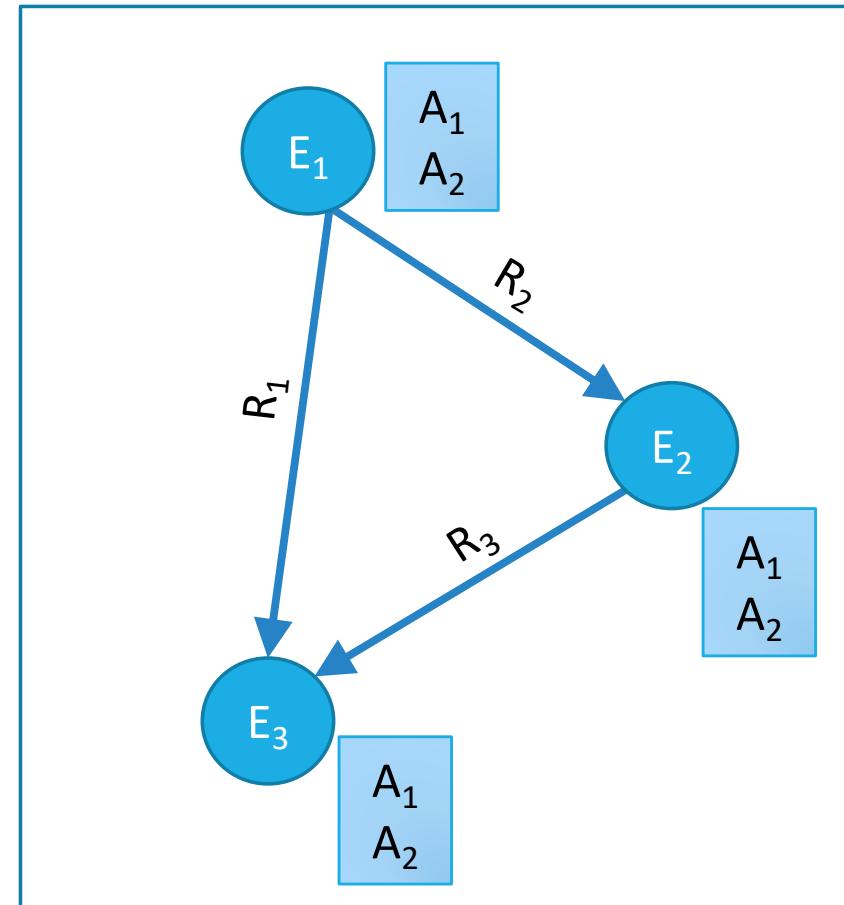
Enforces *ontological constraints*

Incorporates *multiple uncertain sources*

# Knowledge Graph Identification

Define a graphical model to perform all three of these tasks simultaneously!

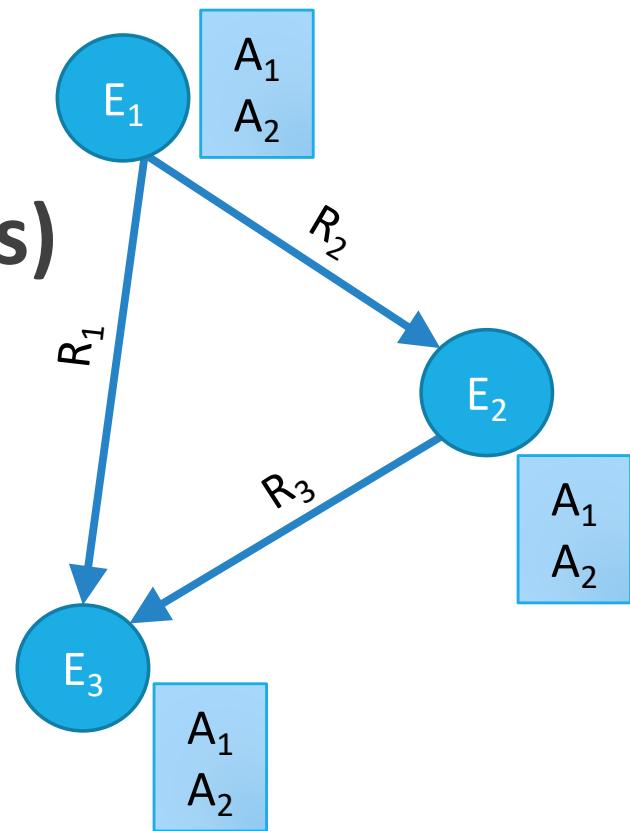
- **Who** are the entities (nodes) in the graph?
- **What** are their attributes and types (labels)?
- **How** are they related (edges)?



# Knowledge Graph Identification

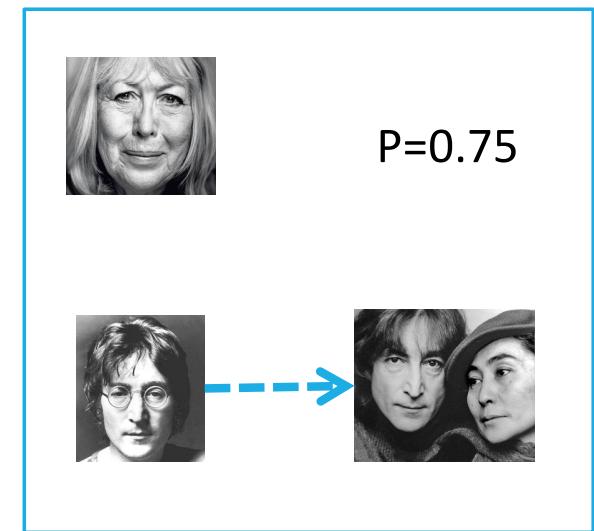
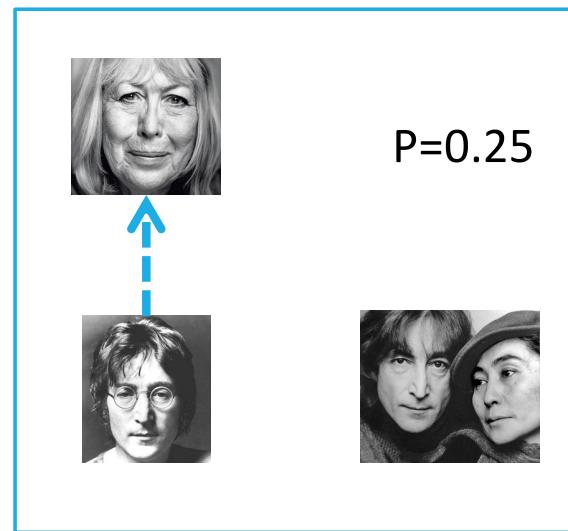
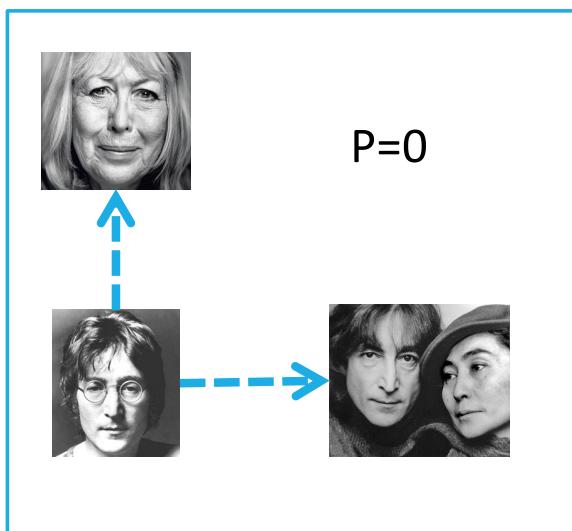
---

$P(\text{Who, What, How} \mid \text{Extractions})$



# Probabilistic Models

- Use dependencies between facts in KG
- Probability defined *jointly* over facts



# What determines probability?

---

- Statistical signals from text extractors and classifiers

# What determines probability?

---

- **Statistical signals from text extractors and classifiers**
  - $P(R(\text{John}, \text{Spouse}, \text{Yoko}))=0.75$ ;  $P(R(\text{John}, \text{Spouse}, \text{Cynthia}))=0.25$
  - LevenshteinSimilarity(Beatles, Beetles) = 0.9

# What determines probability?

---

- Statistical signals from text extractors and classifiers
- Ontological knowledge about domain

# What determines probability?

---

- Statistical signals from text extractors and classifiers
- **Ontological knowledge about domain**
  - Functional(Spouse) & R(A,Spouse,B) -> !R(A,Spouse,C)
  - Range(Spouse, Person) & R(A,Spouse,B) -> Type(B, Person)

# What determines probability?

---

- Statistical signals from text extractors and classifiers
- Ontological knowledge about domain
- Rules and patterns mined from data

# What determines probability?

---

- Statistical signals from text extractors and classifiers
- Ontological knowledge about domain
- Rules and patterns mined from data
  - $R(A, \text{Spouse}, B) \& R(A, \text{Lives}, L) \rightarrow R(B, \text{Lives}, L)$
  - $R(A, \text{Spouse}, B) \& R(A, \text{Child}, C) \rightarrow R(B, \text{Child}, C)$

# What determines probability?

---

- **Statistical signals from text extractors and classifiers**
  - $P(R(\text{John}, \text{Spouse}, \text{Yoko}))=0.75$ ;  $P(R(\text{John}, \text{Spouse}, \text{Cynthia}))=0.25$
  - LevenshteinSimilarity(Beatles, Beetles) = 0.9
- **Ontological knowledge about domain**
  - Functional(Spouse) &  $R(A, \text{Spouse}, B) \rightarrow !R(A, \text{Spouse}, C)$
  - Range(Spouse, Person) &  $R(A, \text{Spouse}, B) \rightarrow \text{Type}(B, \text{Person})$
- **Rules and patterns mined from data**
  - $R(A, \text{Spouse}, B) \& R(A, \text{Lives}, L) \rightarrow R(B, \text{Lives}, L)$
  - $R(A, \text{Spouse}, B) \& R(A, \text{Child}, C) \rightarrow R(B, \text{Child}, C)$

# Example: The Fab Four

THE  
**BEATLES**



# Illustration of KG Identification

---

## Uncertain Extractions:

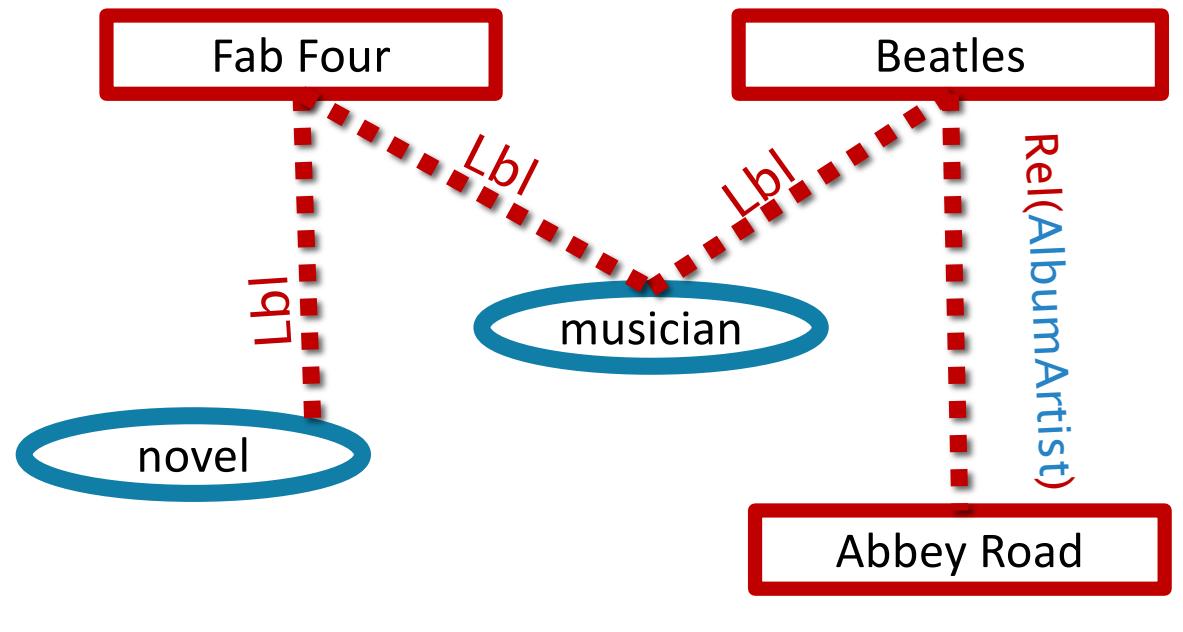
- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist,  
Abbey Road)

# Illustration of KG Identification

## Uncertain Extractions:

- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist, Abbey Road)

(Annotated) Extraction Graph



# Illustration of KG Identification

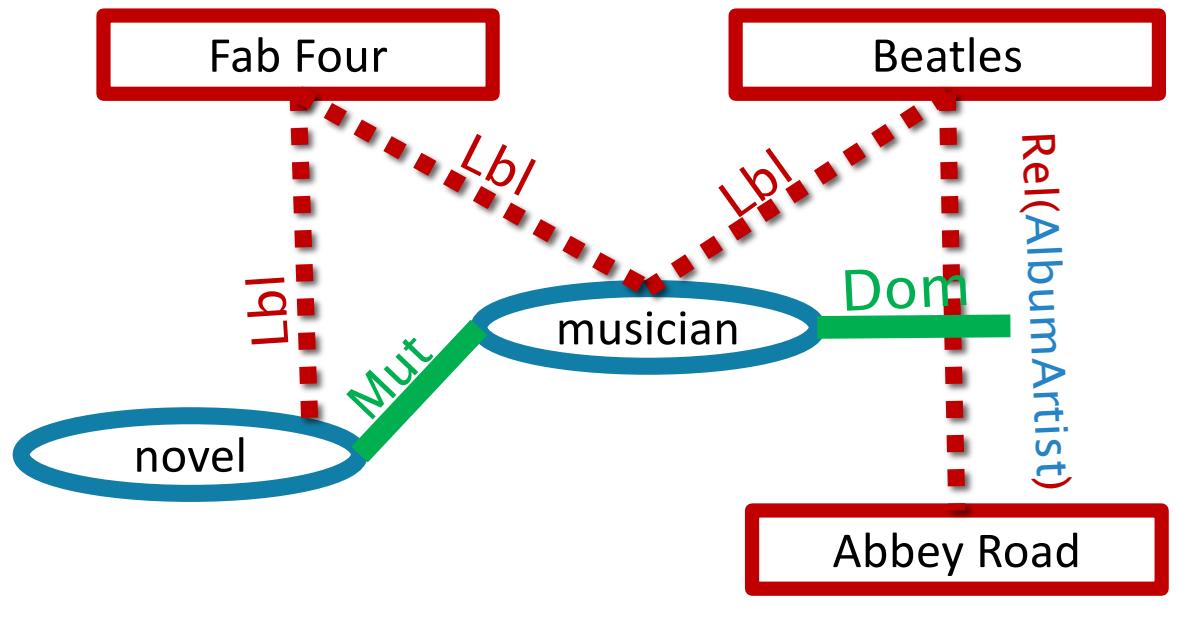
## Uncertain Extractions:

- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist,  
Abbey Road)

## Ontology:

- Dom(albumArtist, musician)
- Mut(novel, musician)

## Extraction Graph



# Illustration of KG Identification

## Uncertain Extractions:

- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist, Abbey Road)

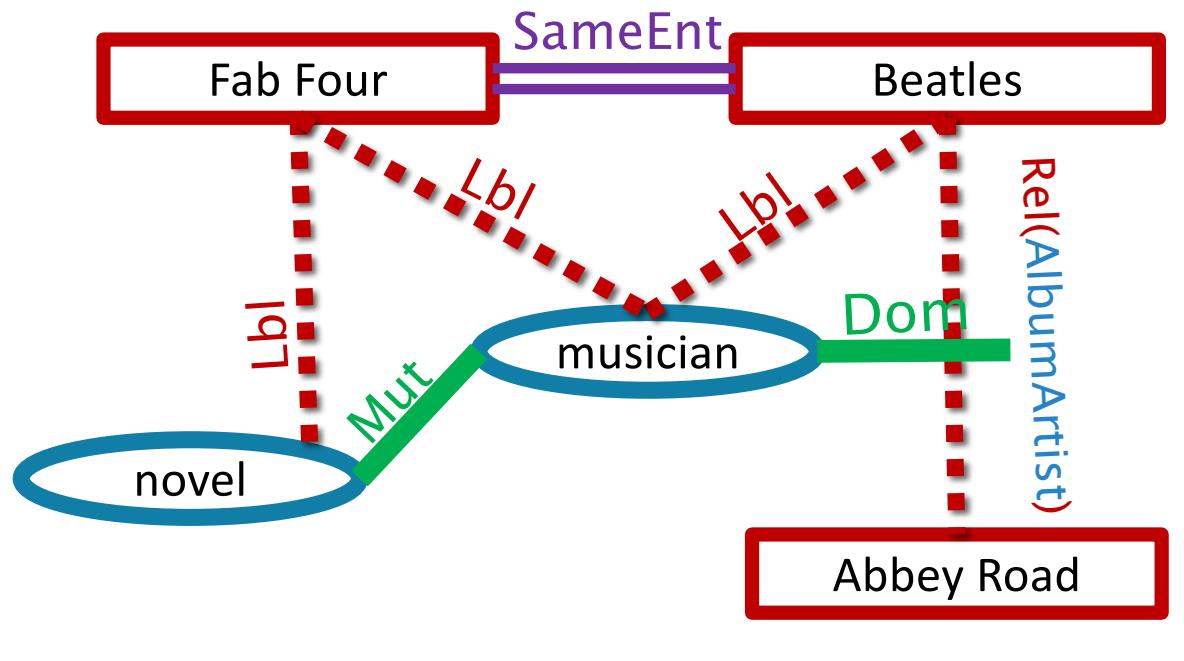
## Ontology:

- Dom(albumArtist, musician)
- Mut(novel, musician)

## Entity Resolution:

SameEnt(Fab Four, Beatles)

(Annotated) Extraction Graph



# Illustration of KG Identification

## Uncertain Extractions:

- .5: Lbl(Fab Four, novel)
- .7: Lbl(Fab Four, musician)
- .9: Lbl(Beatles, musician)
- .8: Rel(Beatles, AlbumArtist, Abbey Road)

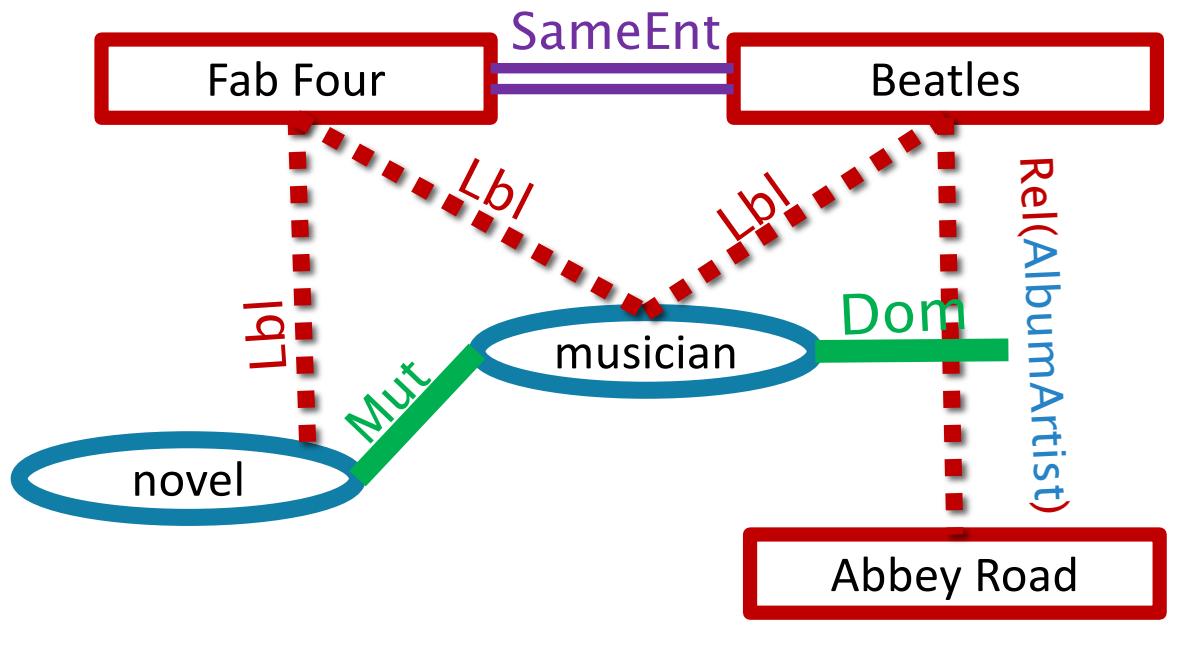
## Ontology:

- Dom(albumArtist, musician)
- Mut(novel, musician)

## Entity Resolution:

SameEnt(Fab Four, Beatles)

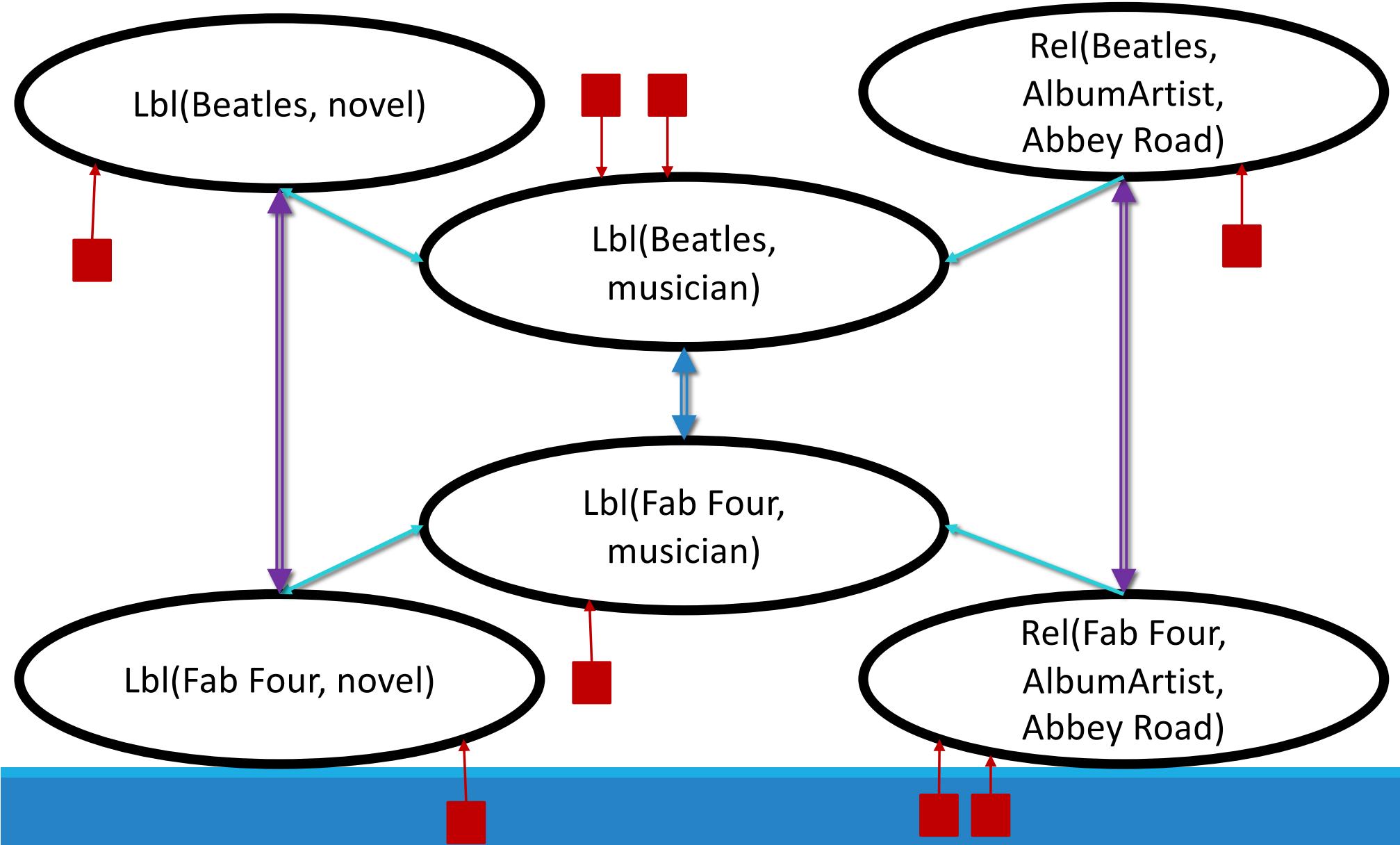
## (Annotated) Extraction Graph



## After Knowledge Graph Identification



# Probabilistic graphical model for KG



# Defining graphical models

---

- Many options for defining a graphical model
- We focus on two approaches, MLNs and PSL, that use **rules**
- **MLNs** treat facts as Boolean, use sampling for satisfaction
- **PSL** infers a “truth value” for each fact via optimization



# Rules for KG Model

100:	Subsumes(L1,L2) & Label(E,L1)	-> Label(E,L2)
100:	Exclusive(L1,L2) & Label(E,L1)	-> !Label(E,L2)
100:	Inverse(R1,R2) & Relation(R1,E,O)	-> Relation(R2,O,E)
100:	Subsumes(R1,R2) & Relation(R1,E,O)	-> Relation(R2,E,O)
100:	Exclusive(R1,R2) & Relation(R1,E,O)	-> !Relation(R2,E,O)
100:	Domain(R,L) & Relation(R,E,O)	-> Label(E,L)
100:	Range(R,L) & Relation(R,E,O)	-> Label(O,L)
10:	SameEntity(E1,E2) & Label(E1,L)	-> Label(E2,L)
10:	SameEntity(E1,E2) & Relation(R,E1,O)	-> Relation(R,E2,O)
1:	Label_OBIE(E,L)	-> Label(E,L)
1:	Label_OpenIE(E,L)	-> Label(E,L)
1:	Relation_Pattern(R,E,O)	-> Relation(R,E,O)
1:		-> !Relation(R,E,O)
1:		-> !Label(E,L)

# Rules to Distributions

- Rules are *grounded* by substituting literals into formulas

$w_r : \text{SAMEENT}(\text{Fab Four}, \text{Beatles}) \wedge$

$\text{LBL}(\text{Beatles}, \text{musician}) \Rightarrow \text{LBL}(\text{Fab Four}, \text{musician})$

- Each ground rule has a weighted satisfaction derived from the formula's truth value

$$P(G|E) = \frac{1}{Z} \exp \left[ \sum_{r \in R} w_r \phi_r(G, E) \right]$$

- Together, the ground rules provide a joint probability distribution over knowledge graph facts, conditioned on the extractions

# Probability Distribution over KGs

$$P(G | E) = \frac{1}{Z} \exp \left[ - \sum_{r \in R} w_r \varphi_r(G) \right]$$

CANDLBL<sub>T</sub>(FabFour, novel)

$\Rightarrow$  LBL(FabFour, novel)

MUT(novel, musician)

$\wedge$  LBL(Beatles, novel)

$\Rightarrow$   $\neg$ LBL(Beatles, musician)

SAMEENT(Beatles, FabFour)

$\wedge$  LBL(Beatles, musician)

$\Rightarrow$  LBL(FabFour, musician)

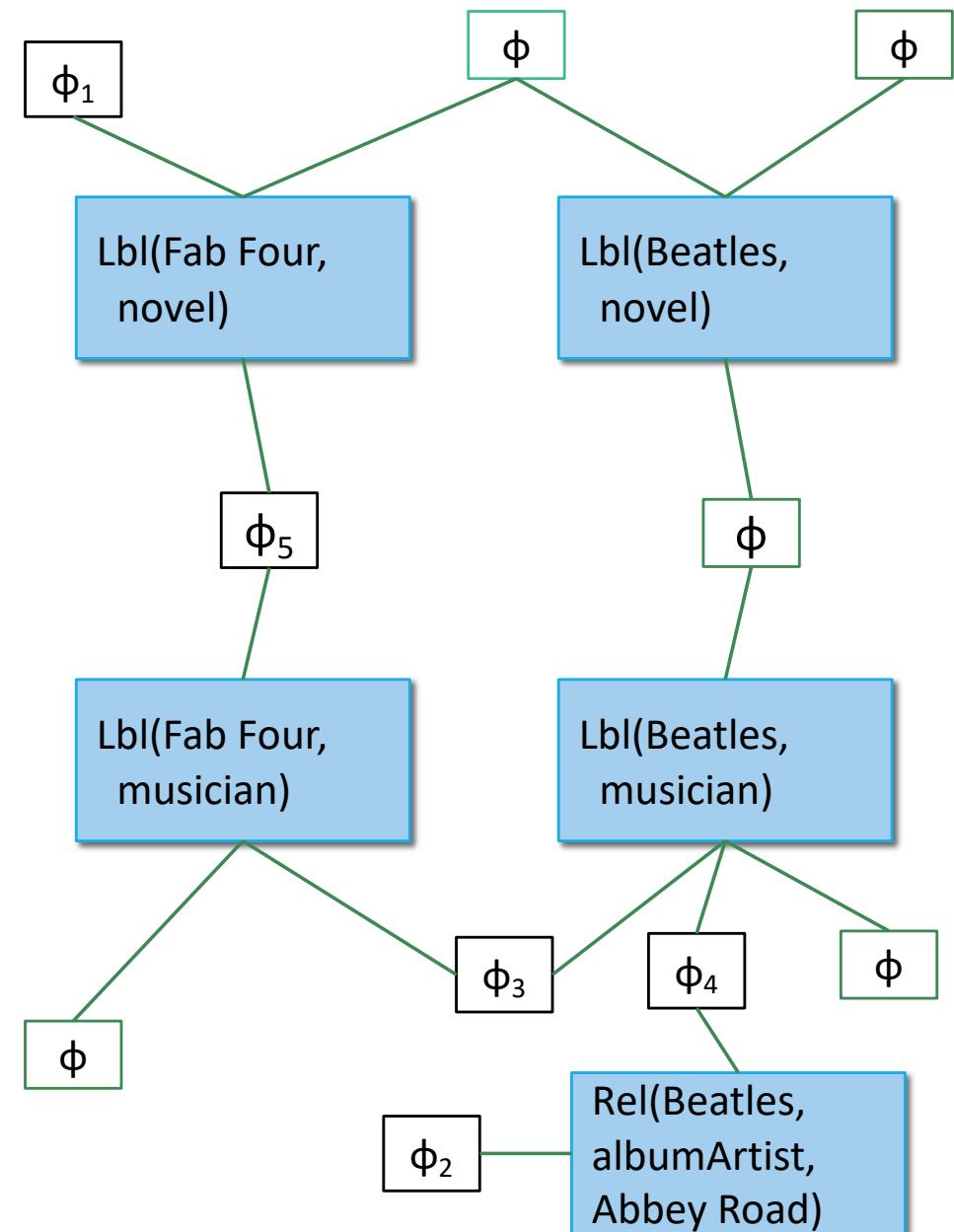
$[\phi_1] \text{ CANDLBL}_{\text{struct}}(\text{FabFour}, \text{novel})$   
 $\Rightarrow \text{LBL}(\text{FabFour}, \text{novel})$

$[\phi_2] \text{ CANDREL}_{\text{pat}}(\text{Beatles}, \text{AlbumArtist}, \text{AbbeyRoad})$   
 $\Rightarrow \text{REL}(\text{Beatles}, \text{AlbumArtist}, \text{AbbeyRoad})$

$[\phi_3] \text{ SAMEENT}(\text{Beatles}, \text{FabFour})$   
 $\wedge \text{LBL}(\text{Beatles}, \text{musician})$   
 $\Rightarrow \text{LBL}(\text{FabFour}, \text{musician})$

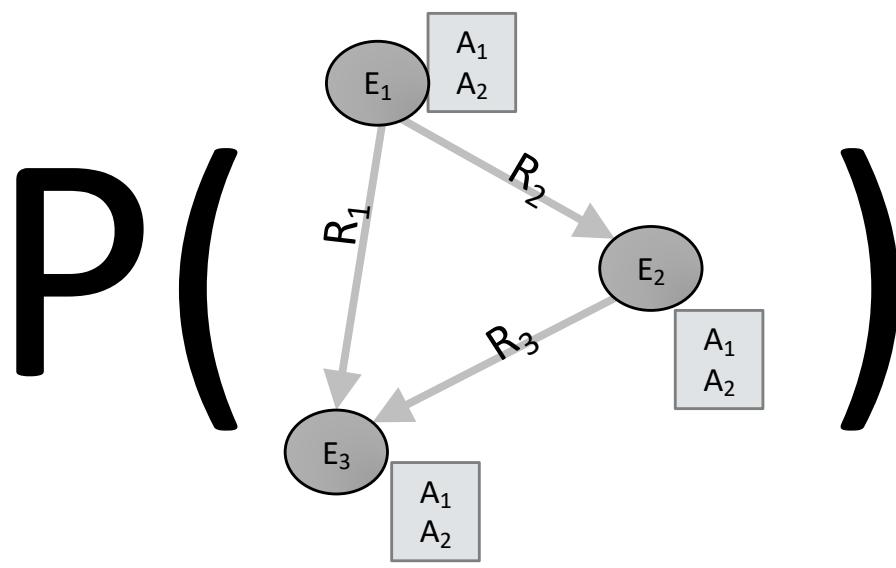
$[\phi_4] \text{ DOM}(\text{AlbumArtist}, \text{musician})$   
 $\wedge \text{REL}(\text{Beatles}, \text{AlbumArtist}, \text{AbbeyRoad})$   
 $\Rightarrow \text{LBL}(\text{Beatles}, \text{musician})$

$[\phi_5] \text{ MUT}(\text{musician}, \text{novel})$   
 $\wedge \text{LBL}(\text{FabFour}, \text{musician})$   
 $\Rightarrow \neg \text{LBL}(\text{FabFour}, \text{novel})$

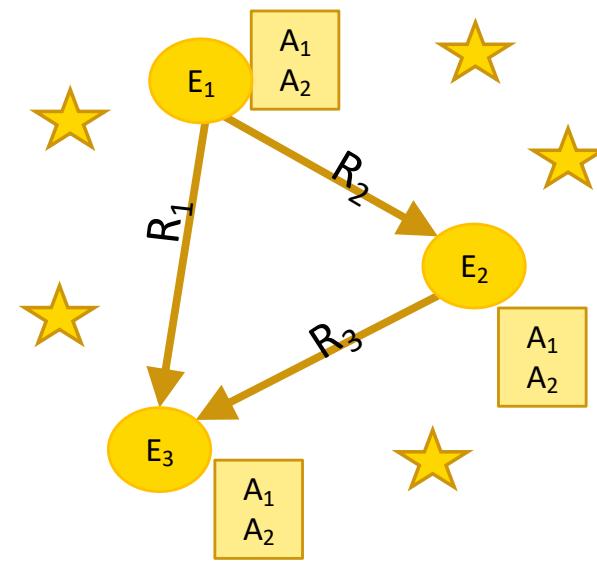


# How do we get a knowledge graph?

Have:  $P(KG)$  for all KGs



Need: best KG



MAP inference: optimizing over distribution to find the best knowledge graph

# Inference and KG optimization

---

- Finding the best KG satisfying weighed rules: NP Hard
- MLNs [discrete]: Monte Carlo sampling methods
  - Solution quality dependent on burn-in time, iterations, etc.
- PSL [continuous]: optimize convex linear surrogate
  - Fast optimization,  $\frac{3}{4}$ -optimal MAX SAT lower bound

# Graphical Models Experiments

---

**Data:** ~1.5M extractions, ~70K ontological relations, ~500 relation/label types

**Task:** Collectively construct a KG and evaluate on 25K target facts

## Comparisons:

Extract	Average confidences of extractors for each fact in the NELL candidates
Rules	Default, rule-based heuristic strategy used by the NELL project
MLN	Jiang+, ICDM12 – estimates marginal probabilities with MC-SAT
PSL	Pujara+, ISWC13 – convex optimization of continuous truth values with ADMM

**Running Time:** Inference completes in 10 seconds, values for 25K facts

	AUC	F1
Extract	.873	.828
Rules	.765	.673
MLN (Jiang, 12)	.899	.836
PSL (Pujara, 13)	.904	.853

# Graphical Models: Pros/Cons

---

## BENEFITS

- Define probability distribution over KGs
- Easily specified via rules
- Fuse knowledge from many different sources

## DRAWBACKS

- Requires optimization over all KG facts - overkill
- Dependent on rules from ontology/expert
- Require probabilistic semantics - unavailable

# Two classes of Probabilistic Models

---

## GRAPHICAL MODELS

- Possible facts in KG are variables
- Logical rules relate facts
- Probability  $\propto$  satisfied rules
- Universally-quantified

## RANDOM WALK METHODS

- Possible facts posed as queries
- Random walks of the KG constitute “proofs”
- Probability  $\propto$  path lengths/transitions
- Locally grounded