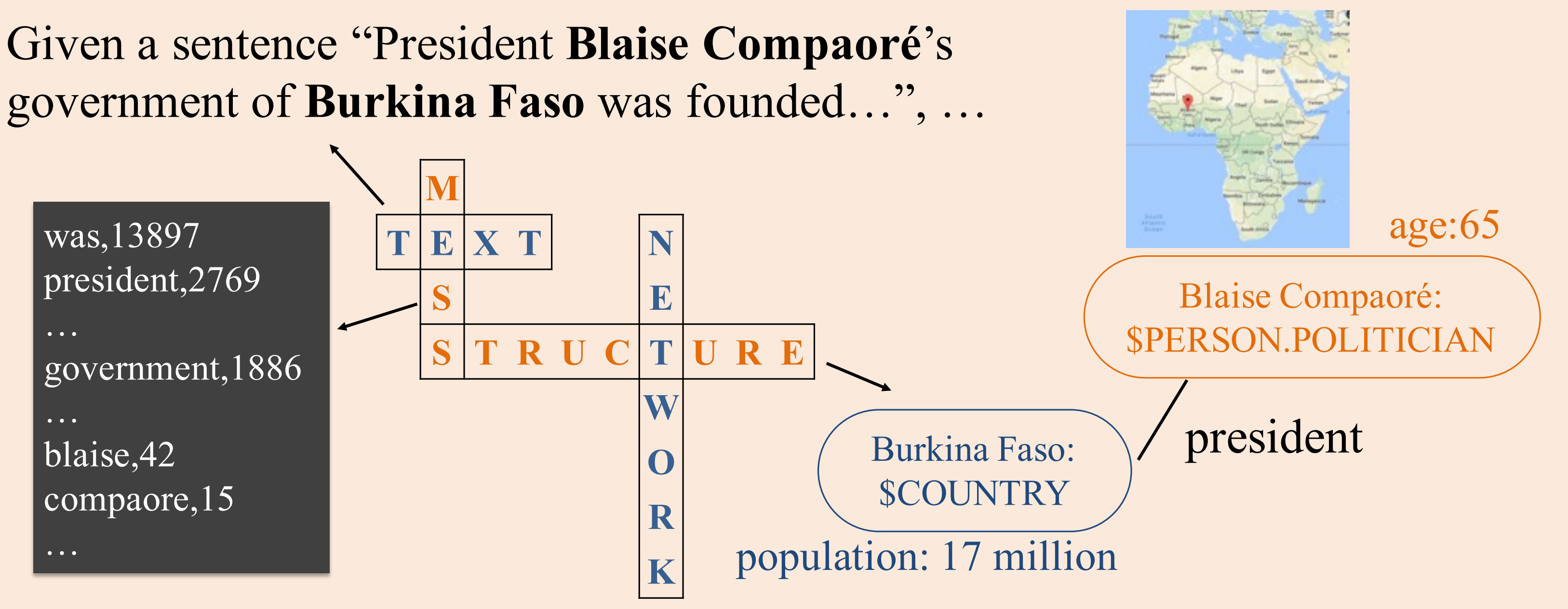
MetaPAD: Meta Pattern Discovery from Massive Text Corpora

Meng Jiang[1], Jingbo Shang[1], Talyor Cassidy[2], Xiang Ren[1],
Lance M. Kaplan[2], Timothy P. Hanratty[2], Jiawei Han[1]
[1] Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
[2] Computational & Information Sciences Directorate, Army Research Laboratory, Adelphi, MD, USA

The 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Halifax, Canada, 2017.

Motivation

Task: *Fact extraction* from massive corpora (e.g., news, tweets, papers) to facilitate heterogeneous information network construction



Given a sentence "President **Blaise Compaoré**'s government of **Burkina Faso** was founded…", …

Task 1: *(entity, attribute name, attribute value)-tuple extraction*

(Burkina Faso, president, Blaise Compaoré)
(Burkina Faso, population, 17 million)
(Blaise Compaoré, age, 65)

Task 2: *(entity type, attribute name, value type)-tuple extraction*

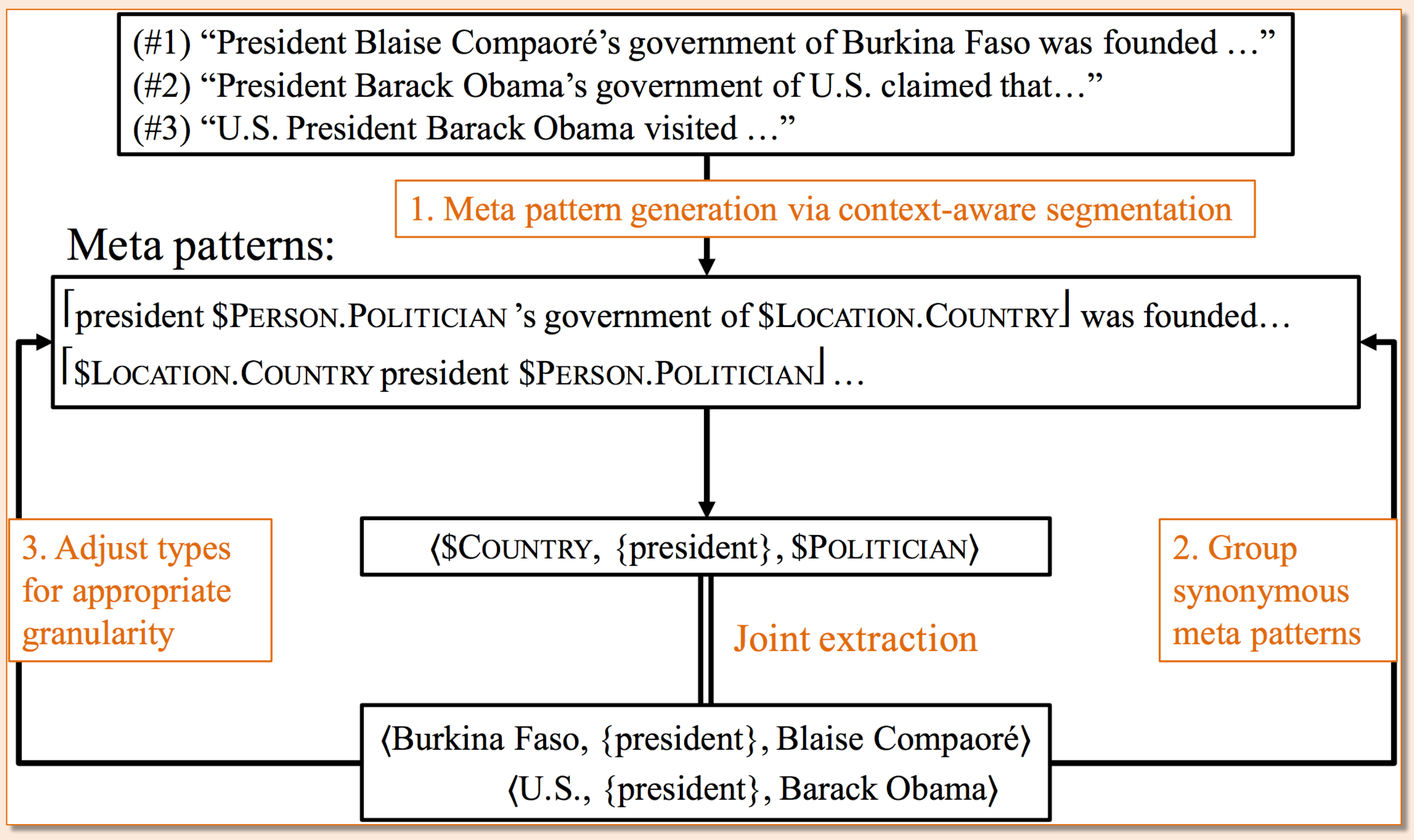
($Location.Country, president, $Person.Politician)
($Location, population, $Digit $DigitUnit)
($Person, age, $Digit)

Idea: Discovering a group of **synonymous "meta patterns"** to find facts.

president $Politician's government of $Country
(e.g., President **Blaise Compaoré**'s government of **Burkina Faso**)
$Country president $Politician
$Country 's president $Politician
president $Politician of $Country
…

(Burkina Faso, president, Blaise Compaoré)
(U.S., president, Barack Obama)
…

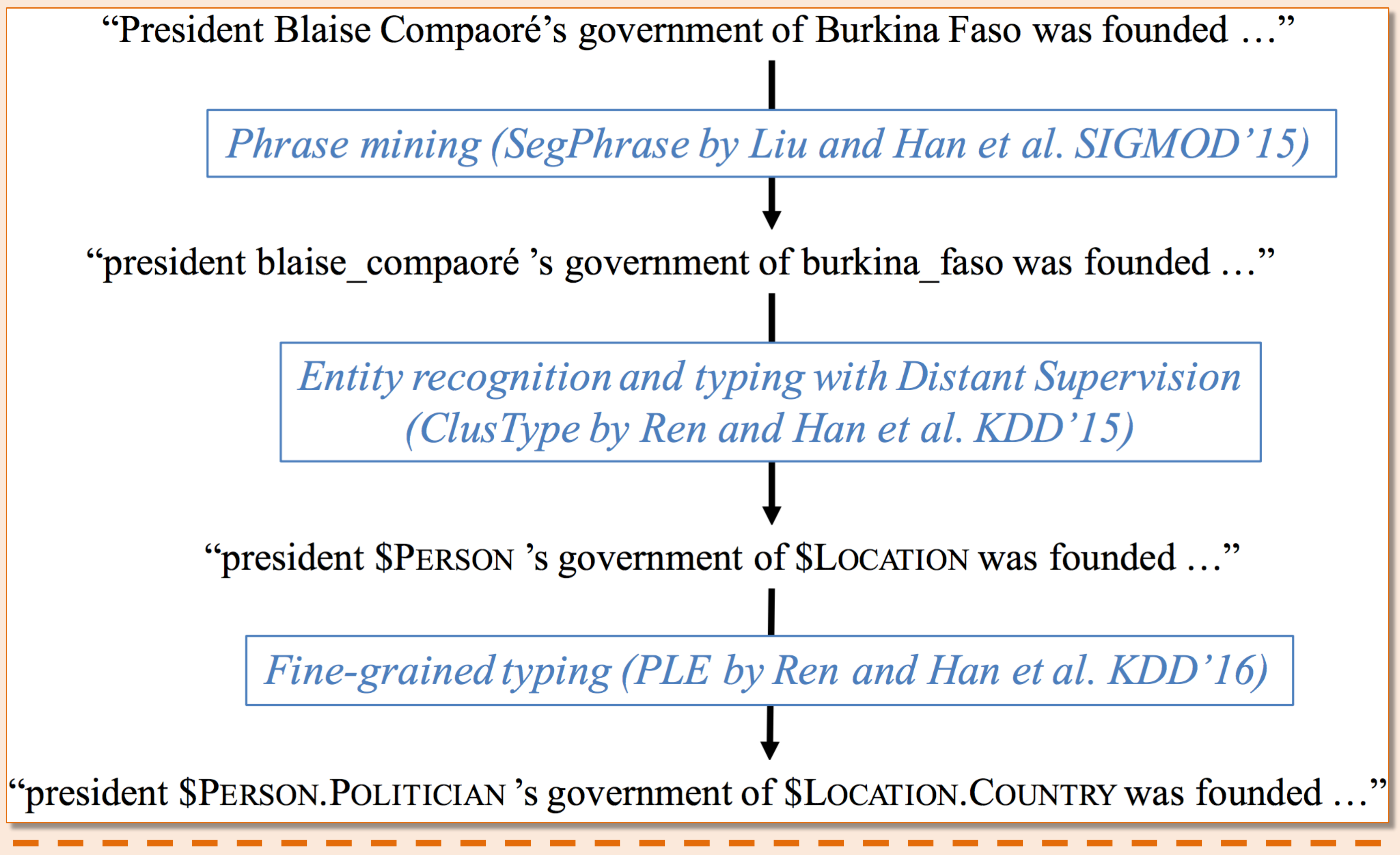
The MetaPAD Framework



(#1) "President Blaise Compaoré's government of Burkina Faso was founded …"
(#2) "President Barack Obama's government of U.S. claimed that…"
(#3) "U.S. President Barack Obama visited …"
1. Meta pattern generation via context-aware segmentation
Meta patterns:
⌈president $Person.Politician 's government of $Location.Country⌋ was founded…
⌈$Location.Country president $Person.Politician⌋ …
3. Adjust types for appropriate granularity
⟨$Country, {president}, $Politician⟩
2. Group synonymous meta patterns
Joint extraction
⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩

mutual enhancement: Generating meta patterns: **c(pattern) >> c(tuple)**
Grouping synonymous meta patterns: **#(patterns) ↑ → #(tuples)↑**

Step 0: Preprocessing



"President Blaise Compaoré's government of Burkina Faso was founded …"

Phrase mining (SegPhrase by Liu and Han et al. SIGMOD'15)

"president blaise_compaoré 's government of burkina_faso was founded …"

Entity recognition and typing with Distant Supervision (ClusType by Ren and Han et al. KDD'15)

"president $Person 's government of $Location was founded …"

Fine-grained typing (PLE by Ren and Han et al. KDD'16)

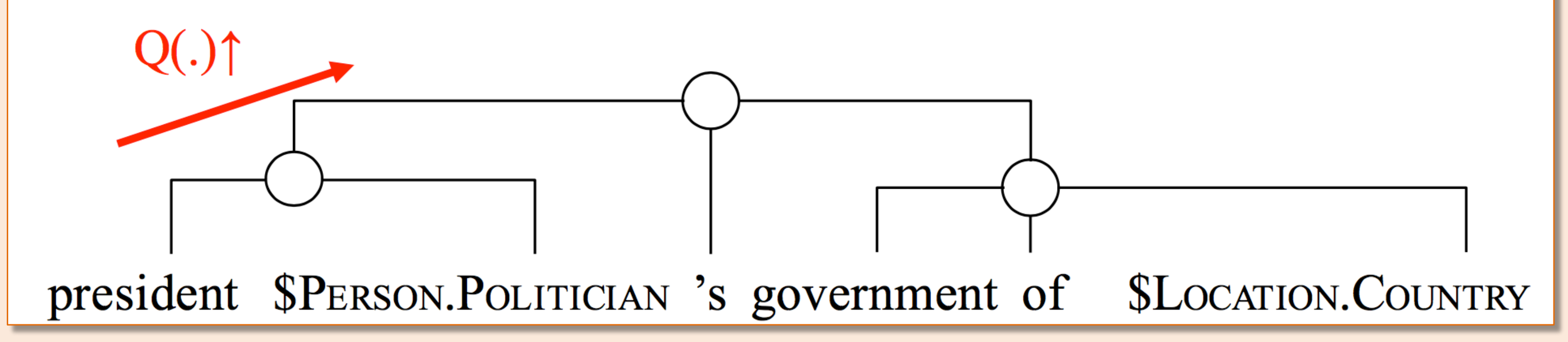
"president $Person.Politician 's government of $Location.Country was founded …"

Step 1: Meta pattern quality assessment and segmentation

A rich set of features:

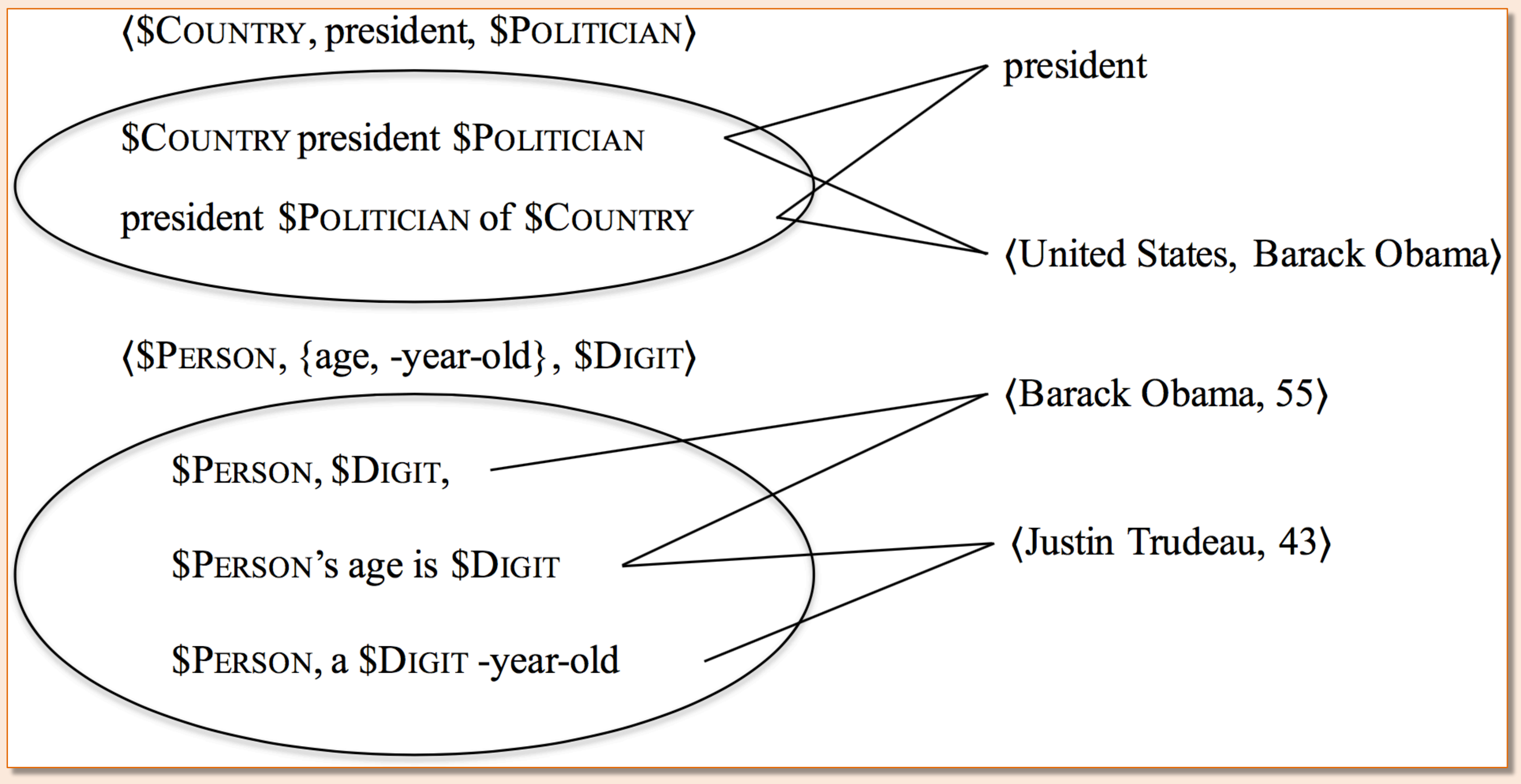
✓ Frequency
✓ Concordance: "$Person 's wife"
✓ Completeness: "$Country president" vs "$Country president $Politician"
✓ Informativeness: "$Person and $Person" vs "$Person 's wife, $Person"

Regression Q(.): random forest with only 300 labels



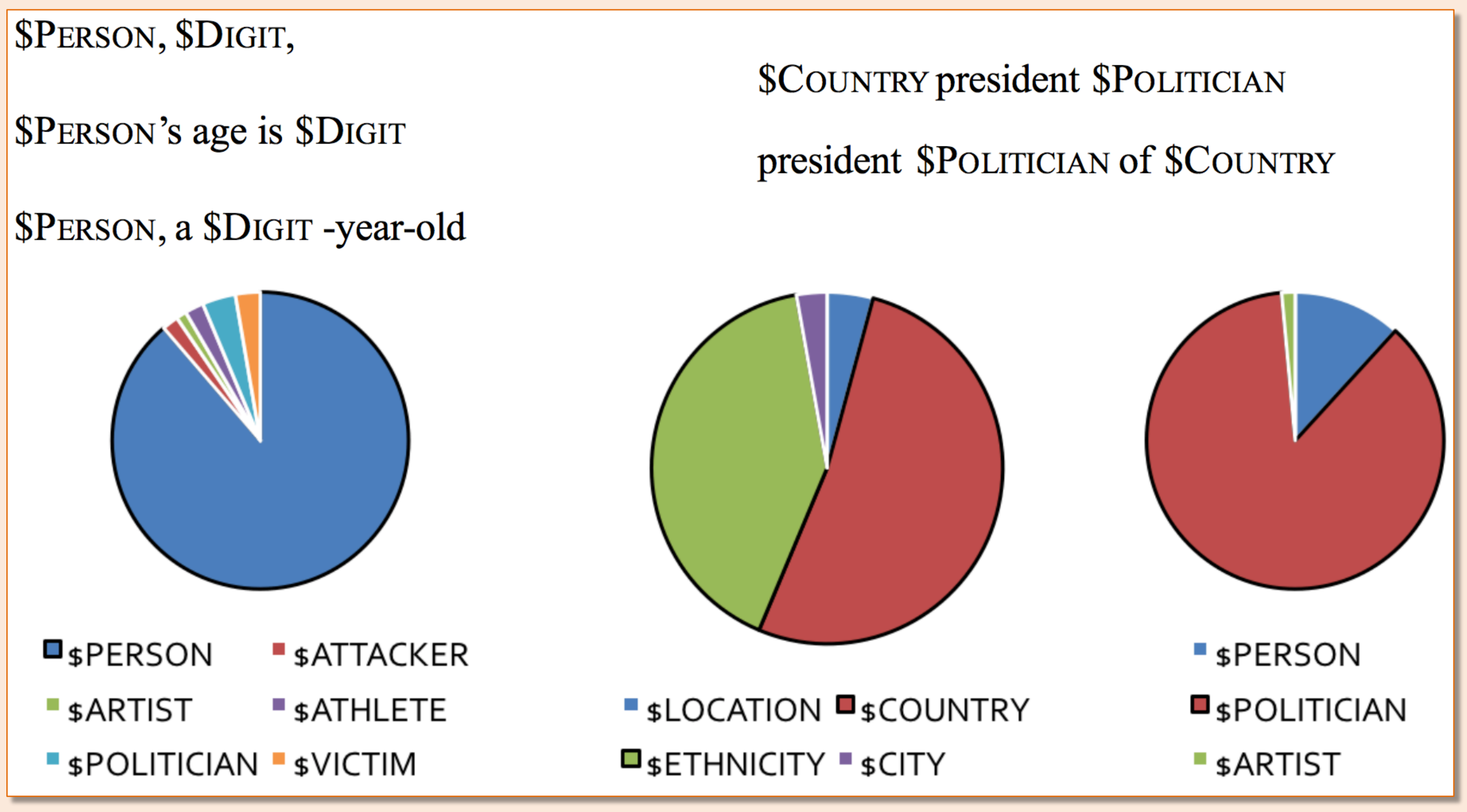
Q(.)↑
president $Person.Politician 's government of $Location.Country

Step 2: Grouping synonymous meta patterns



⟨$Country, president, $Politician⟩
$Country president $Politician
president $Politician of $Country
president
⟨United States, Barack Obama⟩
⟨$Person, {age, -year-old}, $Digit⟩
$Person, $Digit,
$Person's age is $Digit
$Person, a $Digit -year-old
⟨Barack Obama, 55⟩
⟨Justin Trudeau, 43⟩

Step 3: Adjusting types for appropriate granularity



Meta patterns	Entity	Attribute value
$Country president $Politician $Country's president $Politician president $Politician of $Country … president $Politician's government of $Country	United States	Barack Obama
	Russia	Vladimir Putin
	France	Francois Hollande
	…	…
	Burkina Faso	Blaise Compaoré

Meta patterns	Entity	Attribute value
$Company CEO $Person $Company chief executive $Person $Person, the $Company CEO, … $Company former CEO $Person $Person, the $Company former CEO,	Apple	Tim Cook
	Facebook	Mark Zuckerberg
	Hewlett-Packard	Carly Fiorina
	…	…
	Infor	Charles Phillips
	Afghan Citadel	Roya Mahboob

Meta patterns	Entity	Attribute value
$Bacteria was resistant to $Antibiotics $Bacteria are resistant to $Antibiotics $Bacteria is the most resistant to $Antibiotics … $Bacteria, particularly those resistant to $Antibiotics	corynebacterium striatum BM4687	gentamicin
	methicillin-susceptible S aureus	vancomycin
	multidrug-resistant enterobacteriaceae	gentamicin

Meta patterns	Entity	Attribute value
$Treatment was used to treat $Disease $Disease using the $Treatment $Treatment has been used to treat $Disease $Treatment of patients with $Disease …	zoledronic acid therapy	Paget's disease of bone
	bisphosphonates	osteoporosis
	calcitonin	Paget's disease of bone