KnowledgeStore

— Scalable Framework for Interlinking Text and Knowledge —

Marco Rospocher, Bernardo Magnini, Luciano Serafini
Fondazione Bruno Kessler (FBK)

FONDAZIONE BRUNO KESSLER

NewsReader
POST HOC ERGO PROPTER HOC

SEVENTH FRAMEWORK PROGRAMME

INTRODUCTION

- Information is typically available both in unstructured and structured form

- Deep and Large scale NLP now enables to bridge the two "world"

- Development of frameworks for integrating unstructured and structured content only partially investigated

KnowledgeStore

- A scalable, fault-tolerant, and Semantic Web grounded storage system to jointly store, manage, retrieve, and query, both structured and unstructured data

KnowledgeStore

Motivating scenario

- Among a collection of news articles, a user is interested in retrieving all 2014 articles reporting statements of a 20th century US president where he is positively mentioned as "commander-in-chief".

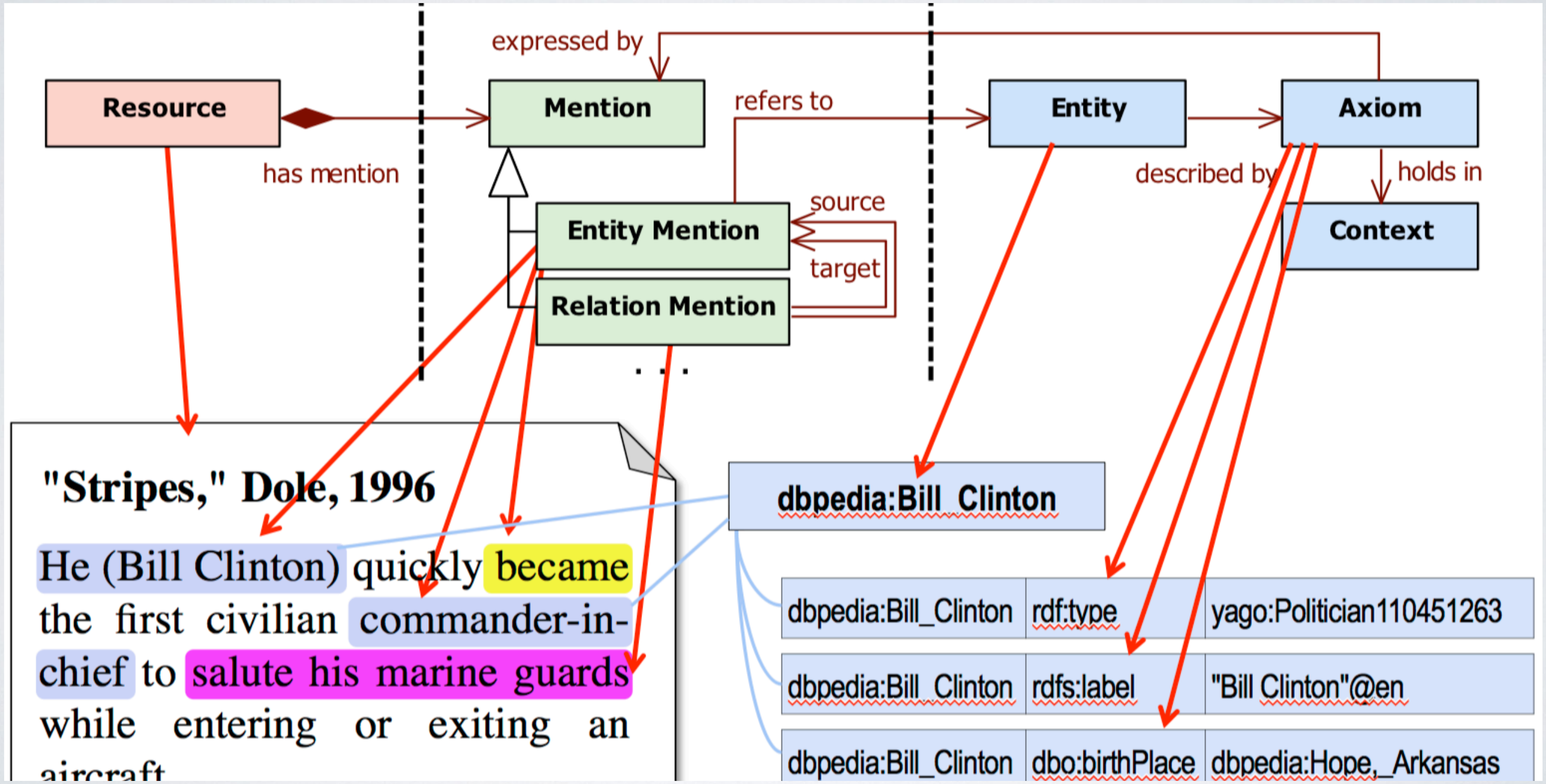
KnowledgeStore

Motivating scenario

- Among a collection of news articles, a user is interested in retrieving all 2014 articles reporting statements of a 20th century US president where he is positively mentioned as "commander-in-chief".

KnowledgeStore

In a nutshell



KnowledgeStore

Exploitation

- Enhanced applications (e.g., decision support systems)

- Developing, debugging, training, and evaluating NLP and knowledge processing tasks

- Reasoning on Extracted Information (e.g., on Events)

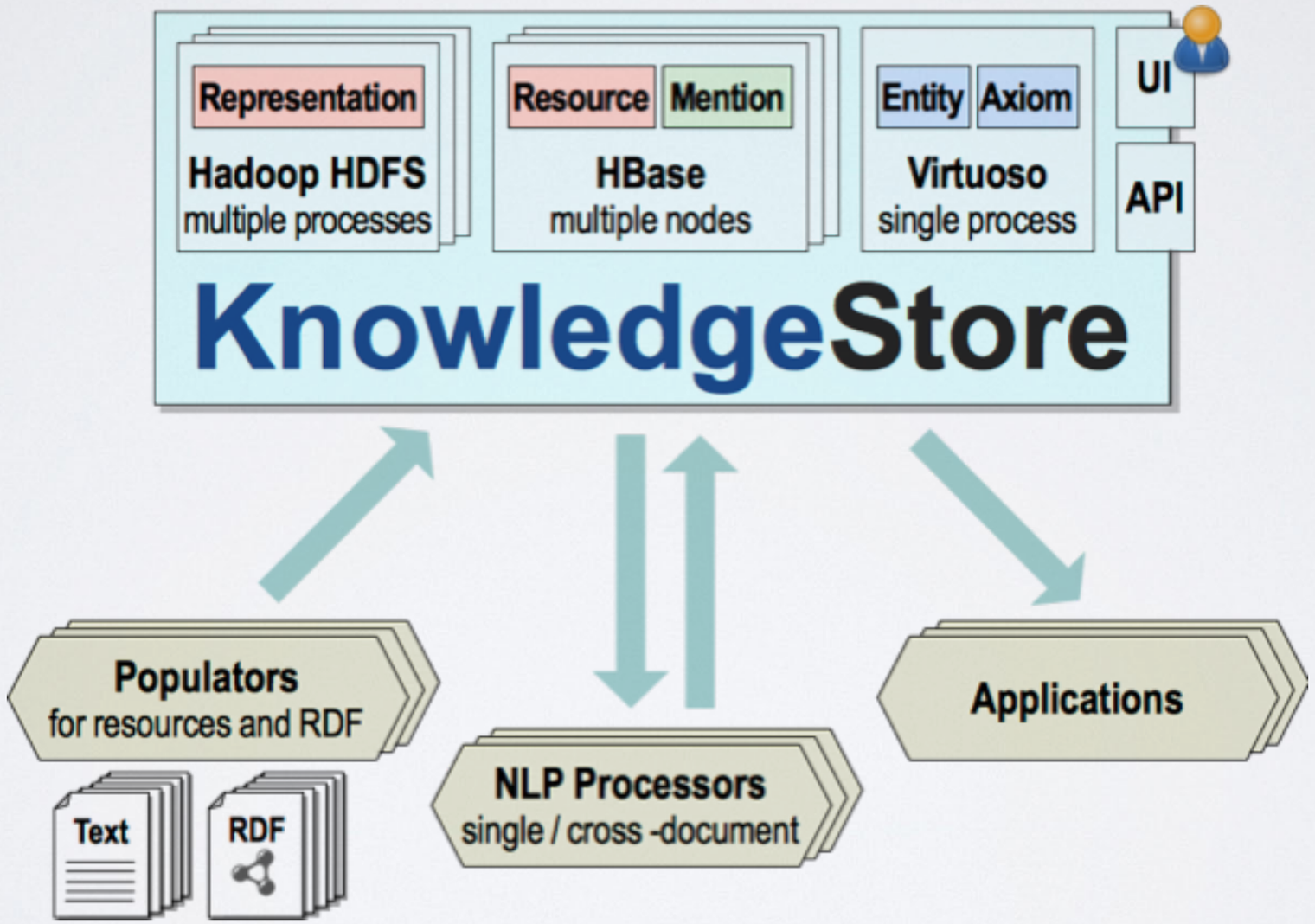
- Text Exploration

OUTLINE

- A walk through the KnowledgeStore
- The KnowledgeStore “live”
- The KnowledgeStore in NewsReader
- Next Challenges

KnowledgeStore

Functional View



KnowledgeStore

Data Model

Resources

- arbitrary text documents, e.g., news and associated annotation files
- both digital representation (file) and associated metadata attributes

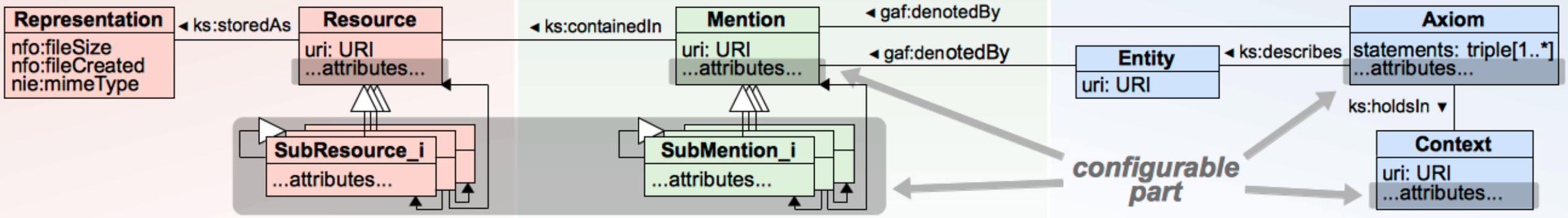
Mentions

[Tim]$_{PER}$ [W3C]$_{ORG}$ [Trento]$_{LOC}$ [ISWC]$_{EVENT}$

- snippets of resources denoting something of interest - entity / axiom
- sign vs referent distinction
- described by linguistic metadata attributes

Entities

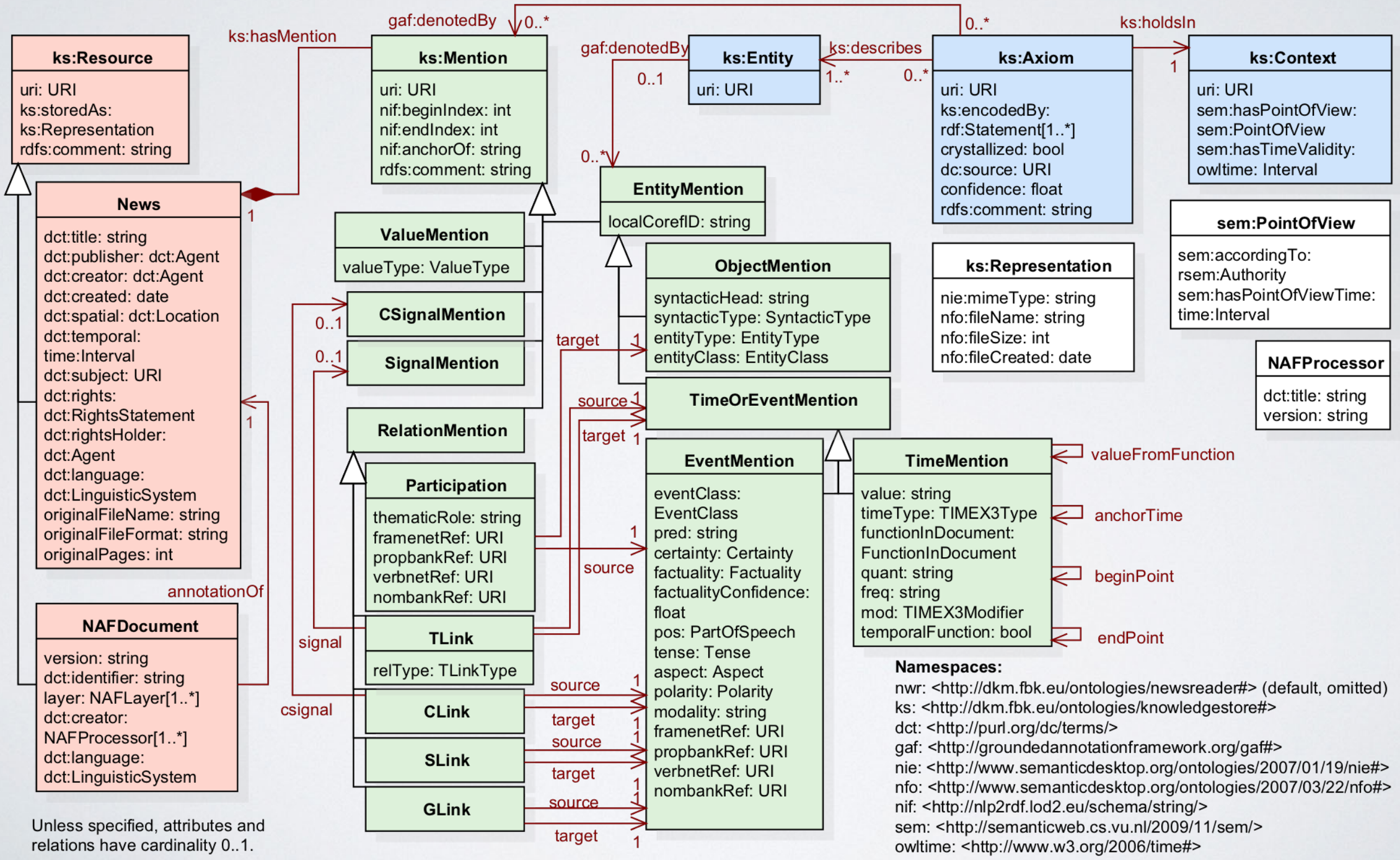
- persons, organizations, events, ...
- described by logical axioms (~triples)
- axiom validity restricted to a specific context (e.g., time span, attribution pair)



- It's Flexible
- It's an OWL 2 Ontology

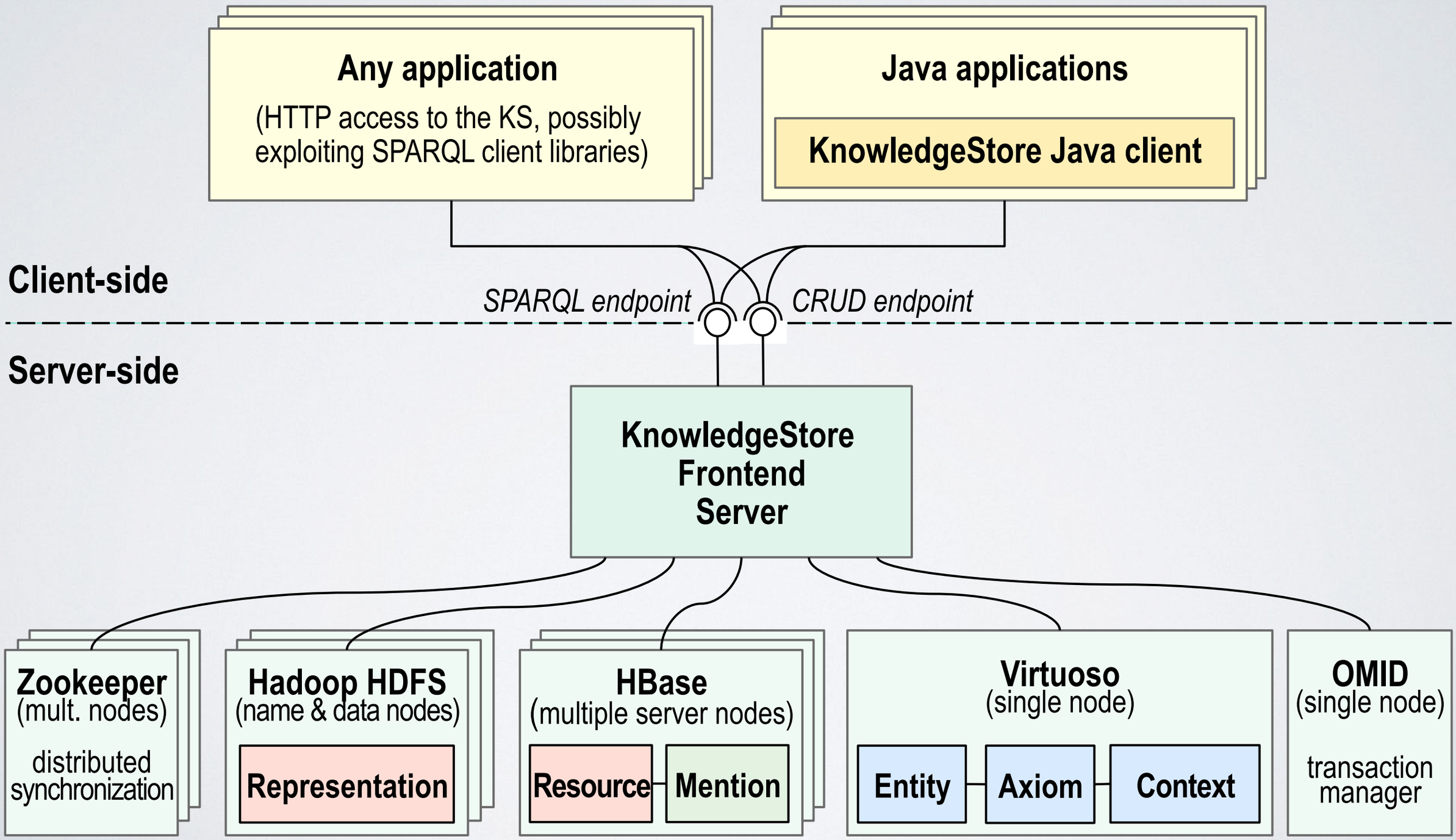
KnowledgeStore

Example: Data Model For Event Extraction from News



KnowledgeStore

Architectural View



KnowledgeStore

Looking through the glass box

- The KnowledgeStore **User Interface**
  - lookup: given the URI of an object (i.e., resource, mention, entity), retrieves all the content about it
  - SPARQL: run arbitrary queries against the SPARQL endpoint

KnowledgeStore

NewsReader
POST HOC ERGO PROPTER HOC

recording history
by processing massive streams of daily news

ICT 316404
FP7-ICT-2011-8
www.newsreader-project.eu
Jan 2013 - Dec 2015

VU UNIVERSITY AMSTERDAM

Universidad del País Vasco · Euskal Herriko Unibertsitatea

FONDAZIONE BRUNO KESSLER

ScraperWiki

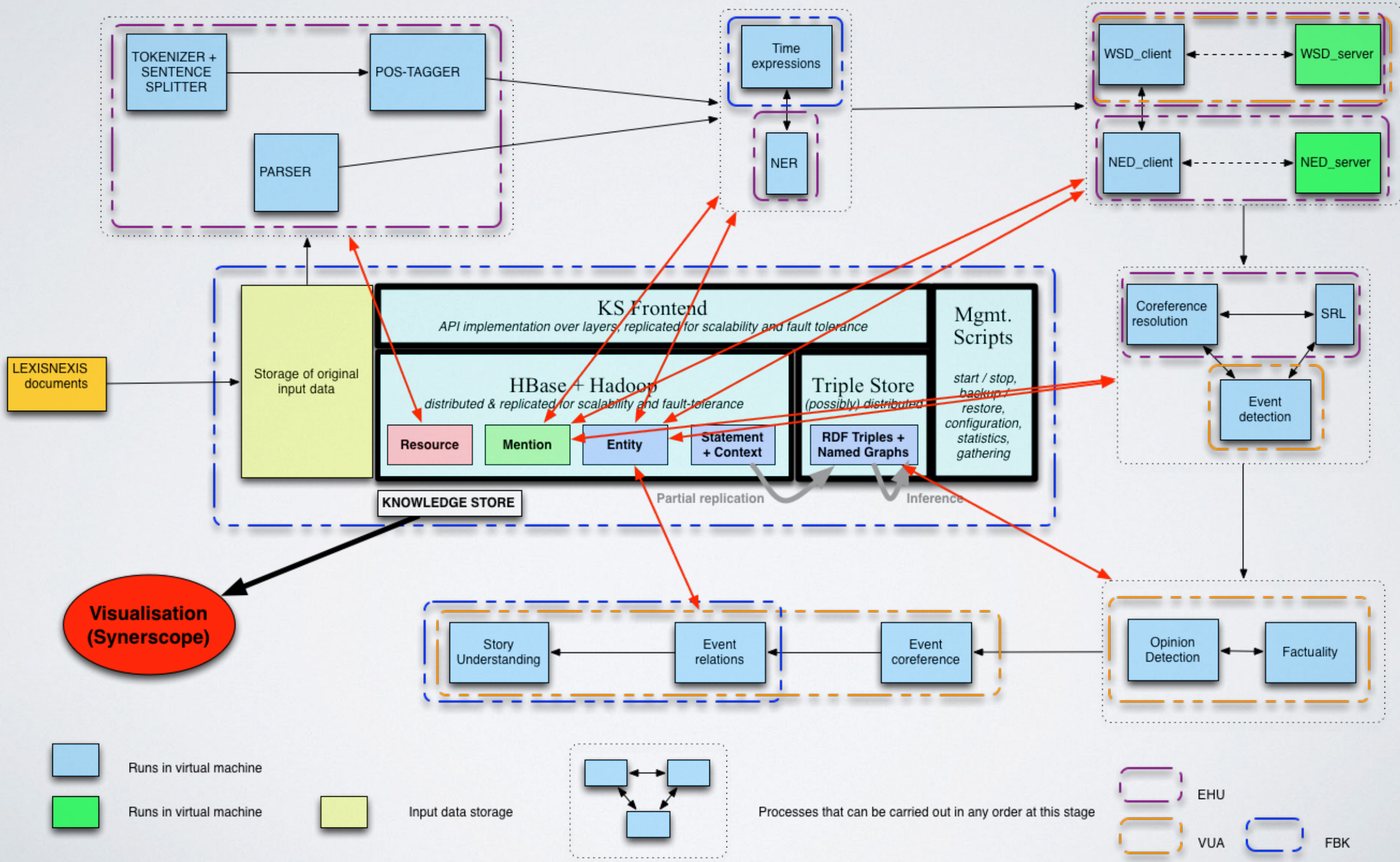
LexisNexis

COOPERATION

synerscope
connecting the dots

KnowledgeStore in NewsReader

POST HOC ERGO PROPTER HOC



KnowledgeStore

LIVE DEMO

Dataset used in the Demo (NewsReader Project):

- Domain: Global Automotive Industry
- **1.3M news documents** (2003-2013), provided by LexisNexis (www.lexisnexis.nl)
- 1.3M NLP annotation files (NAF format) obtained processing the news (NewsReader NLP Pipeline)
- **205M mentions** of **events**, **persons**, **organisations**, **locations**, **time expressions**...
- **535M of RDF triples** about **events**, **persons**, **organisations**, **locations**, **time expressions**...:
  - 439M extracted from text
  - 96M coming from selected background knowledge (DBpedia)

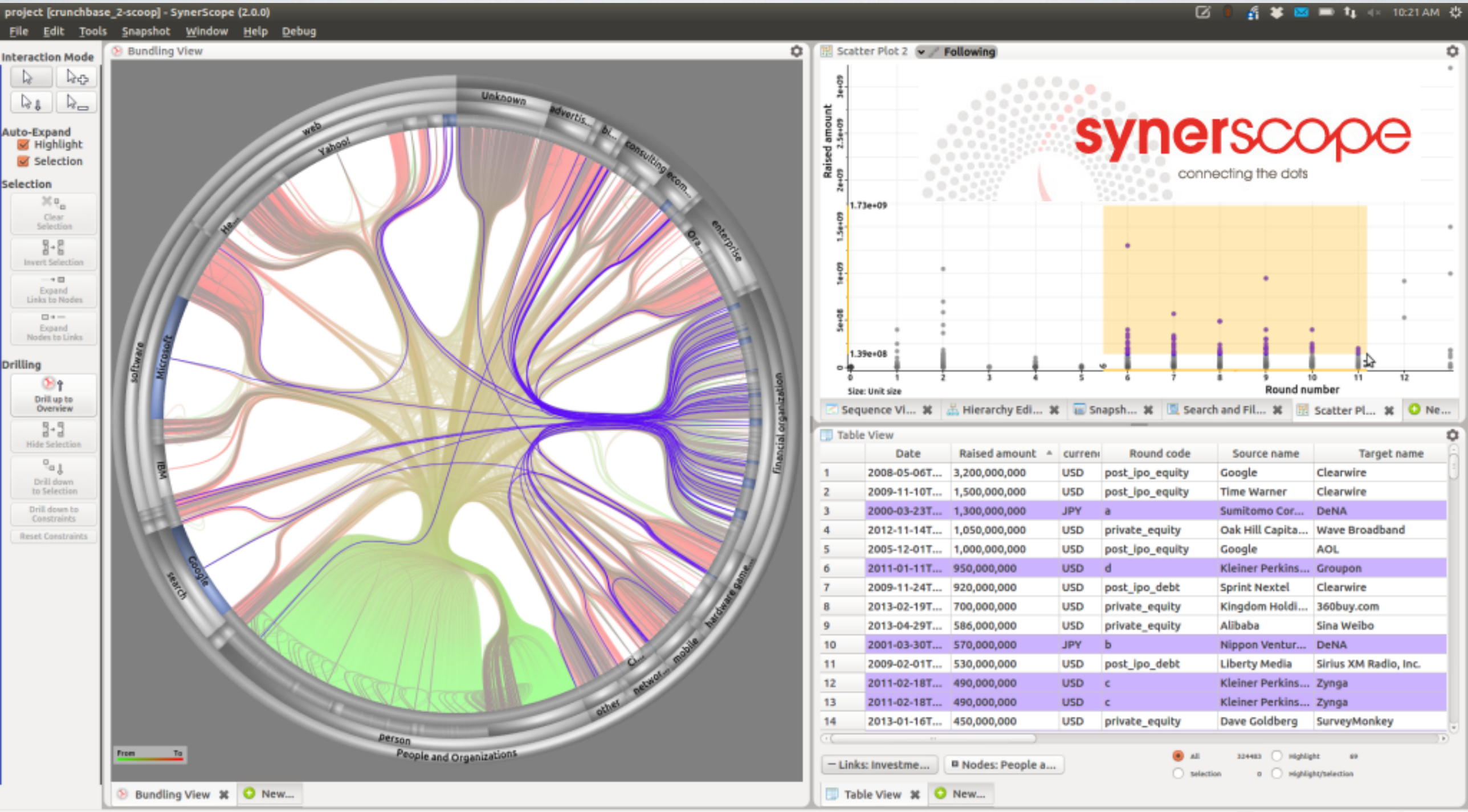
KnowledgeStore in NewsReader

POST HOC ERGO PROPTER HOC

	WikiNews	**Cars (Ver. 1)**	**FIFA WorldCup**	**Cars (Ver. 2)**
Domain	General News	Automotive Industry	Sport, Football	Automotive Industry
Resource Providers	en.wikinews.org	LexisNexis	LexisNexis, BBC, The Guardian	LexisNexis
Populated in	February 2015	January 2014	May 2014	December 2014
Resources	18,510	63,635	212,258	1,259,748
Mentions	2,629,176	9,110,683	76,165,114	205,114,711
Entities	951,879	2,212,691	10,246,338	27,123,724
Events	*624,439*	*1,783,991*	*9,387,356*	*25,156,574*
Persons	*94,731*	*199,999*	*403,021*	*729,797*
Organizations	*101,754*	*187,842*	*431,232*	*947,262*
Locations	*130,955*	*40,859*	*24,729*	*290,091*
Axioms (Triples)	105,610,963	316,034,616	240,731,408	535,011,673
from Mentions	*9,700,585*	*46,359,300*	*136,135,841*	*439,101,295*
from Background Knowledge	*95,910,378*	*269,675,316*	*104,595,567*	*95,910,378*
distilled from:	DBpedia 2014 (EN)	DBpedia 3.9 (EN,NL,ES,IT)	DBpedia 3.9 (EN)	DBpedia 2014 (EN)
Total Disk Space (GB)	17.64	30.67	82.48	260.20
Resource Layer	*1.25*	*3.10*	*16.55*	*108.27*
Mention Layer	*1.49*	*4.77*	*41.72*	*112.00*
Entity Layer	*14.90*	*22.80*	*24.21*	*39.93*
Approx. Population Total Time (hrs)	2	30	56	160
Approx. Rate (resources/hour)	*9,300*	*2,250*	*4,000*	*7,800*

KnowledgeStore in NewsReader

Decision Making on top of the KnowledgeStore



	Date	Raised amount	currenc	Round code	Source name	Target name
1	2008-05-06T...	3,200,000,000	USD	post_ipo_equity	Google	Clearwire
2	2009-11-10T...	1,500,000,000	USD	post_ipo_equity	Time Warner	Clearwire
3	2000-03-23T...	1,300,000,000	JPY	a	Sumitomo Cor...	DeNA
4	2012-11-14T...	1,050,000,000	USD	private_equity	Oak Hill Capita...	Wave Broadband
5	2005-12-01T...	1,000,000,000	USD	post_ipo_equity	Google	AOL
6	2011-01-11T...	950,000,000	USD	d	Kleiner Perkins...	Groupon
7	2009-11-24T...	920,000,000	USD	post_ipo_debt	Sprint Nextel	Clearwire
8	2013-02-19T...	700,000,000	USD	private_equity	Kingdom Holdi...	360buy.com
9	2013-04-29T...	586,000,000	USD	private_equity	Alibaba	Sina Weibo
10	2001-03-30T...	570,000,000	JPY	b	Nippon Ventur...	DeNA
11	2009-02-01T...	530,000,000	USD	post_ipo_debt	Liberty Media	Sirius XM Radio, Inc.
12	2011-02-18T...	490,000,000	USD	c	Kleiner Perkins...	Zynga
13	2011-02-18T...	490,000,000	USD	c	Kleiner Perkins...	Zynga
14	2013-01-16T...	450,000,000	USD	private_equity	Dave Goldberg	SurveyMonkey

KnowledgeStore in NewsReader

Exploited in Three Hack Day Events

KnowledgeStore in NewsReader

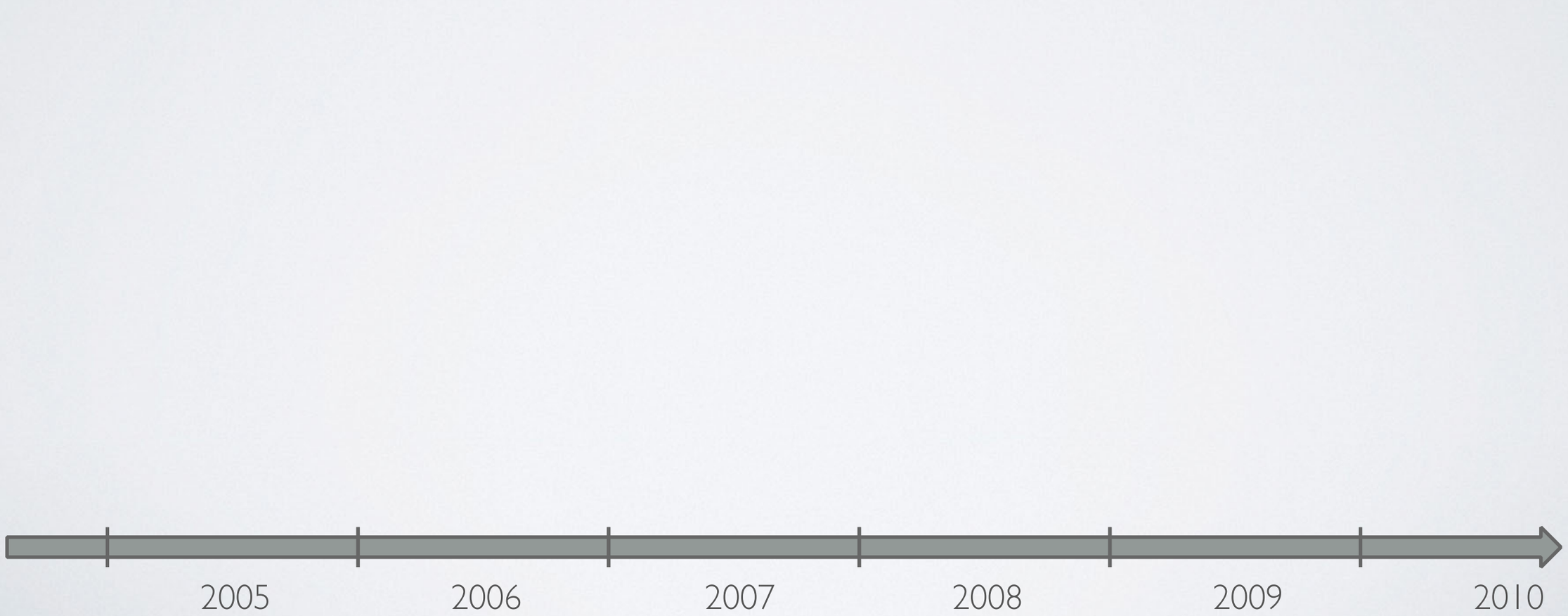
Exploited in Three Hack Day Events

- Capable of handling:
  - large number of requests (>110K)
  - multiple concurrent requests (40 requests/sec.)
  - low response time (30-214ms)

KnowledgeStore

Reasoning on Events

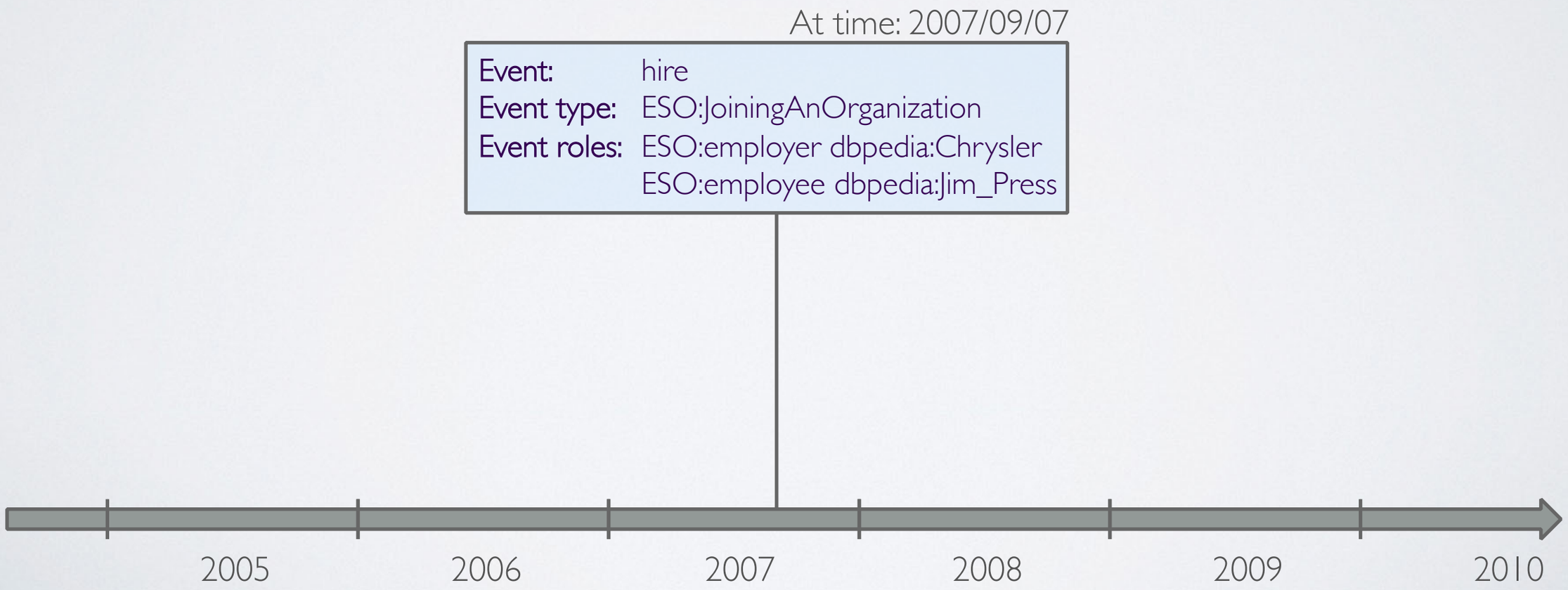
- Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)
  - Example: "Yesterday, Chrysler hired Jim Press to lead its sales and marketing"



KnowledgeStore

Reasoning on Events

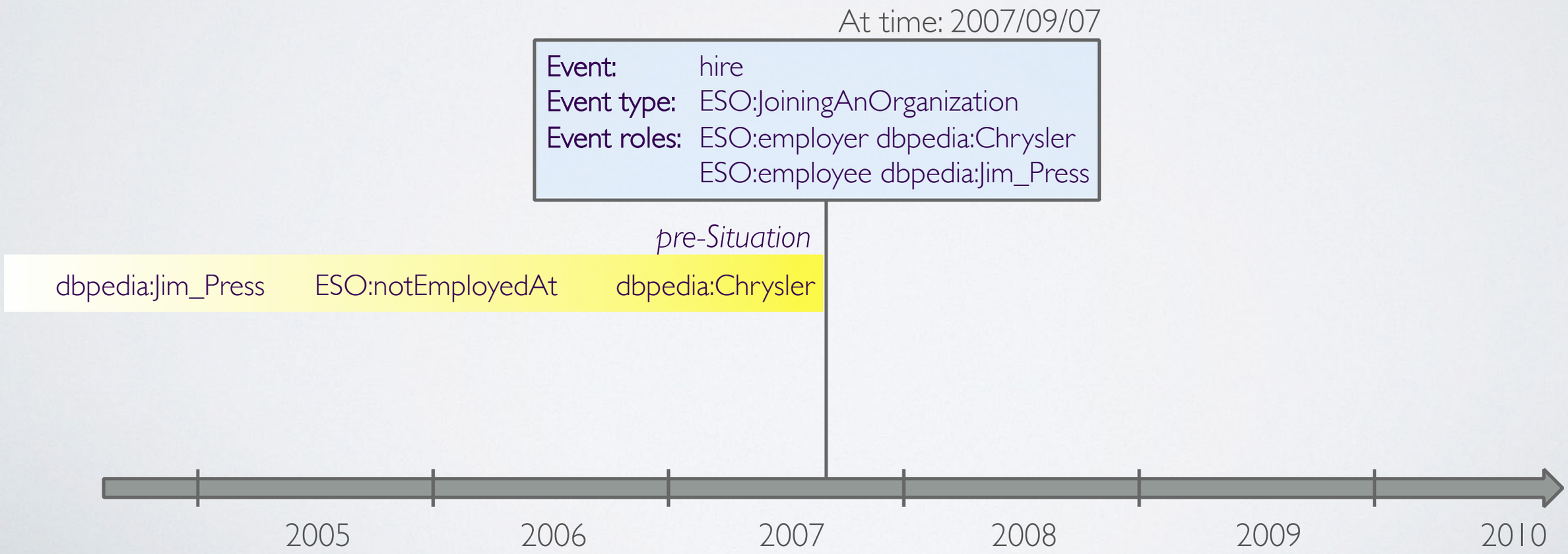
- Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)
  - Example: "Yesterday, Chrysler hired Jim Press to lead its sales and marketing"



KnowledgeStore

Reasoning on Events

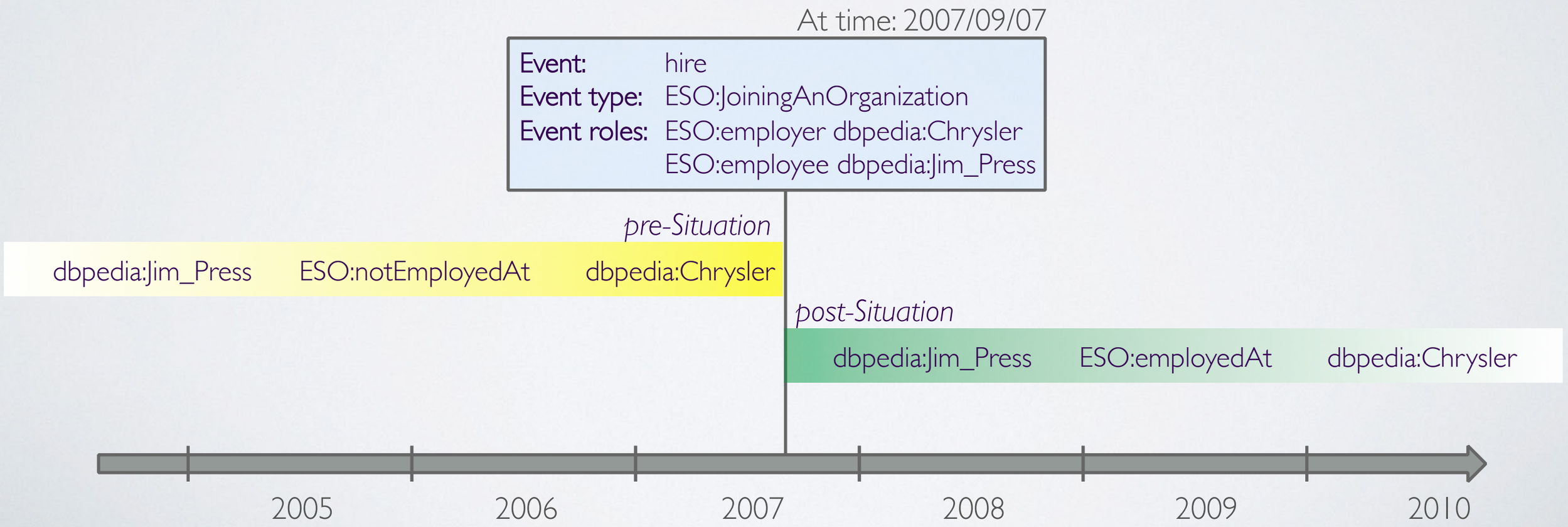
- Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)
  - Example: "Yesterday, Chrysler hired Jim Press to lead its sales and marketing"



KnowledgeStore

Reasoning on Events

- Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)
  - Example: "Yesterday, Chrysler hired Jim Press to lead its sales and marketing"



KnowledgeStore

Reasoning on Events

- Applied on the **1.3M** global automotive industry **news**
- Extremely fast: 1,333s (**~22m**) to process **~500M triples**
- Triggered **2M new triples** (i.e., **not in the text**), organised in 397,885 situations
- 255,470 events have at least a pre/post/during situation:
  - 71.2% of the events having at least two distinct roles

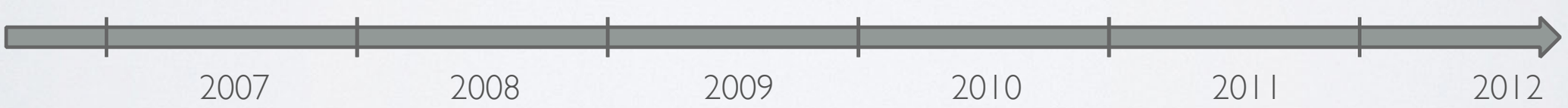
KnowledgeStore

Looking ahead to the future…

KnowledgeStore

Beyond "Asserted" Knowledge…

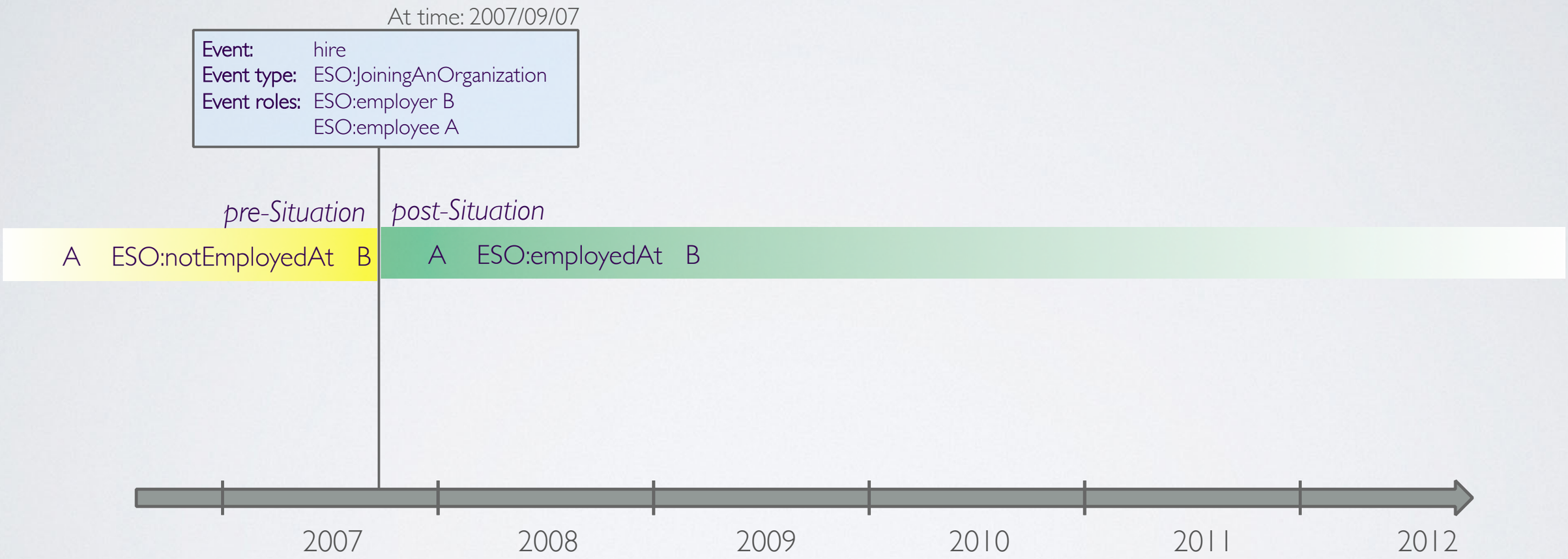
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

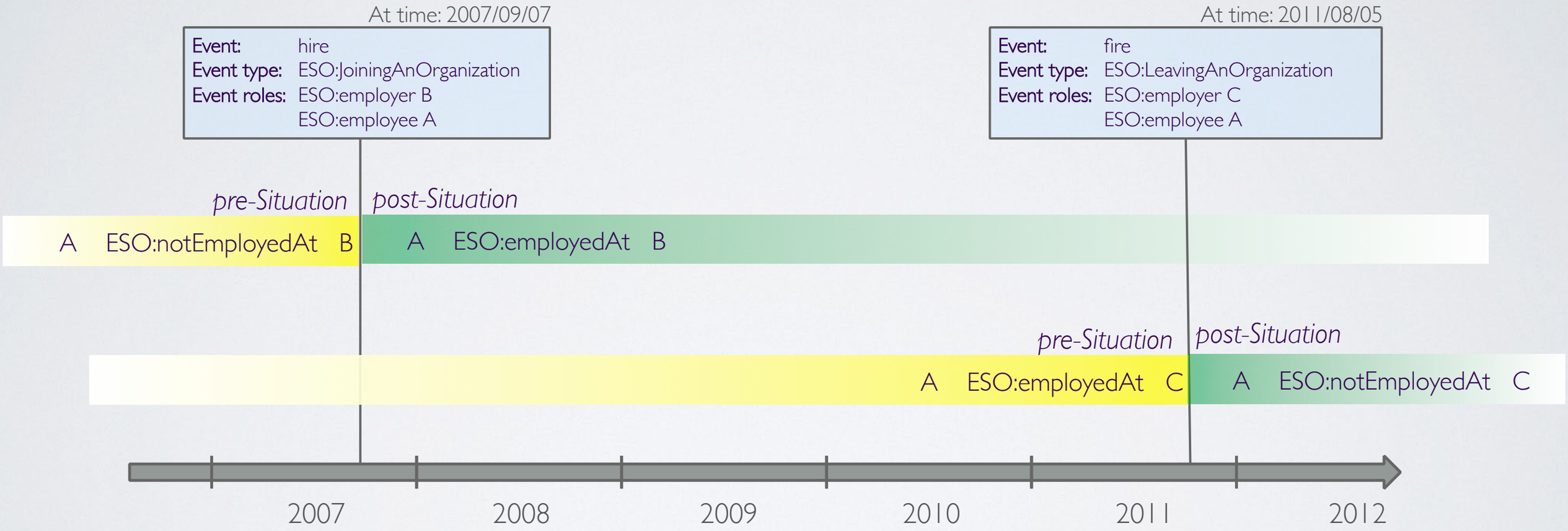
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

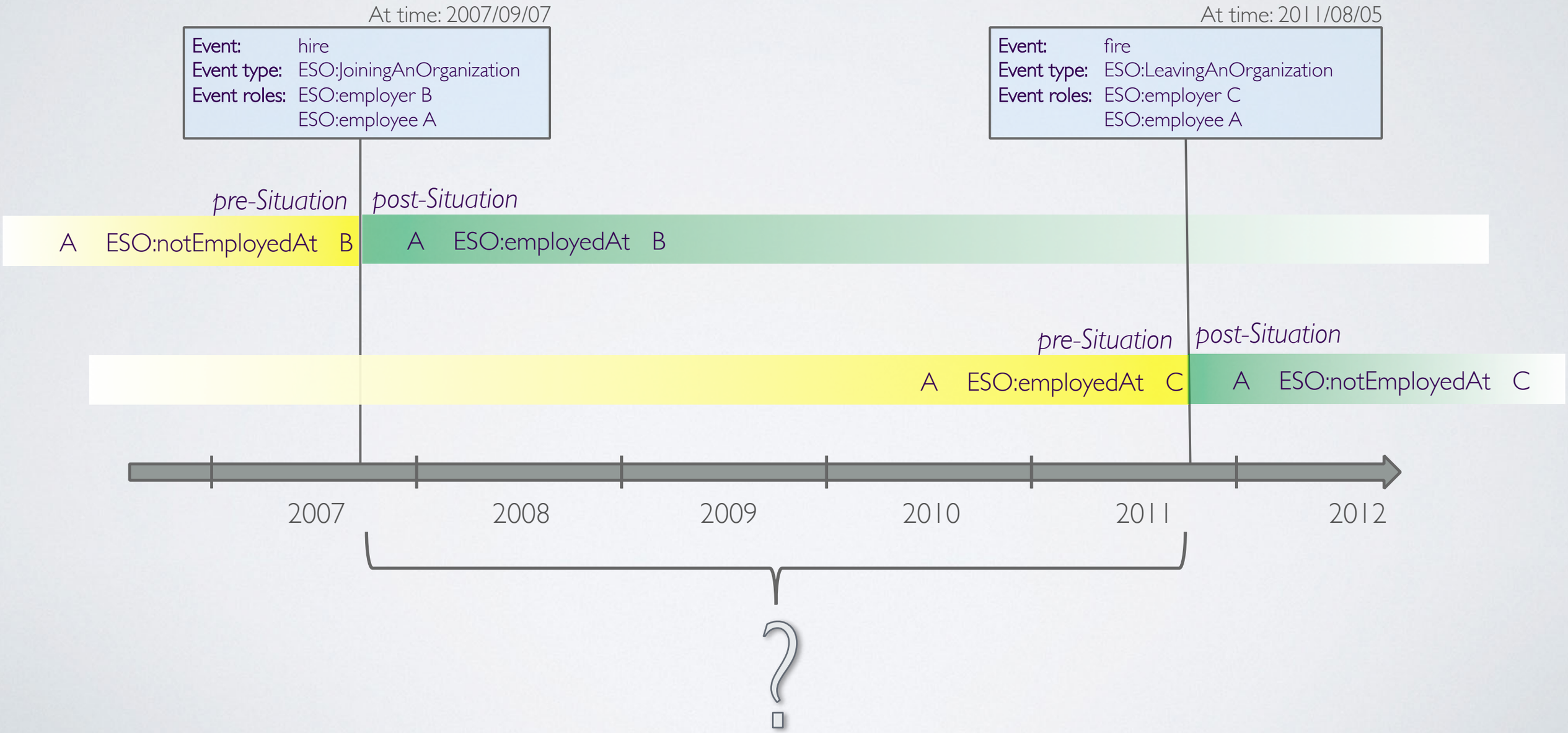
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

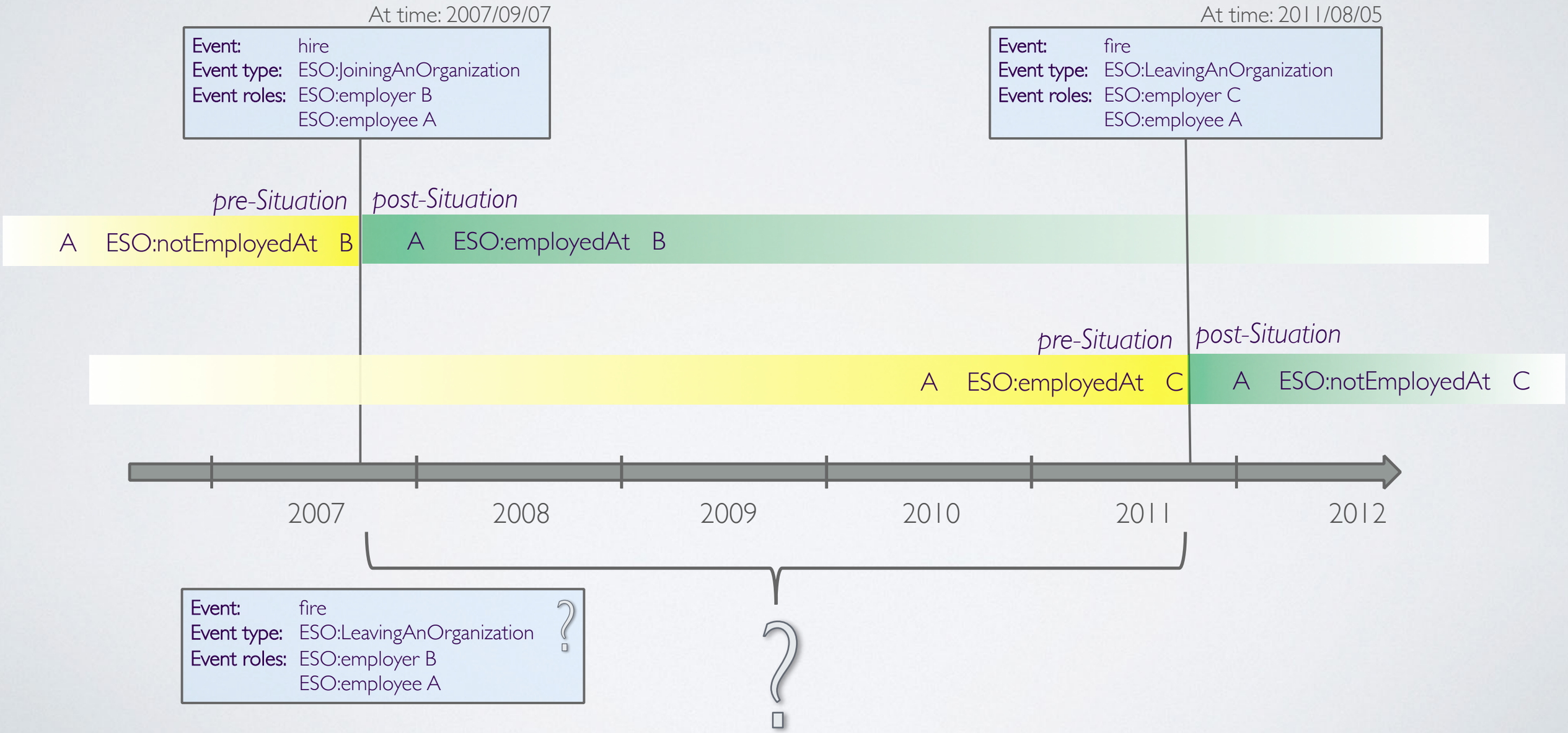
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

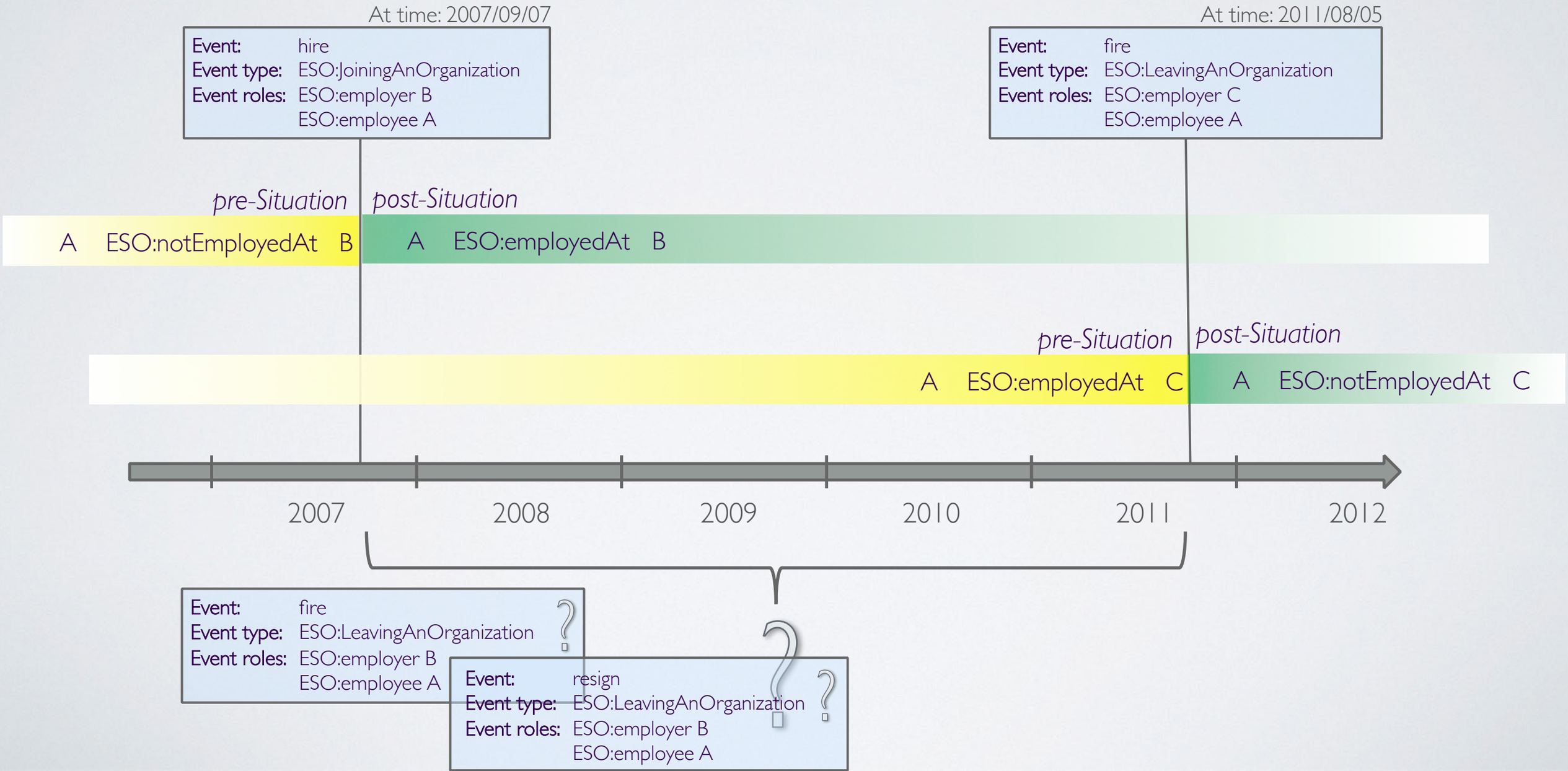
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

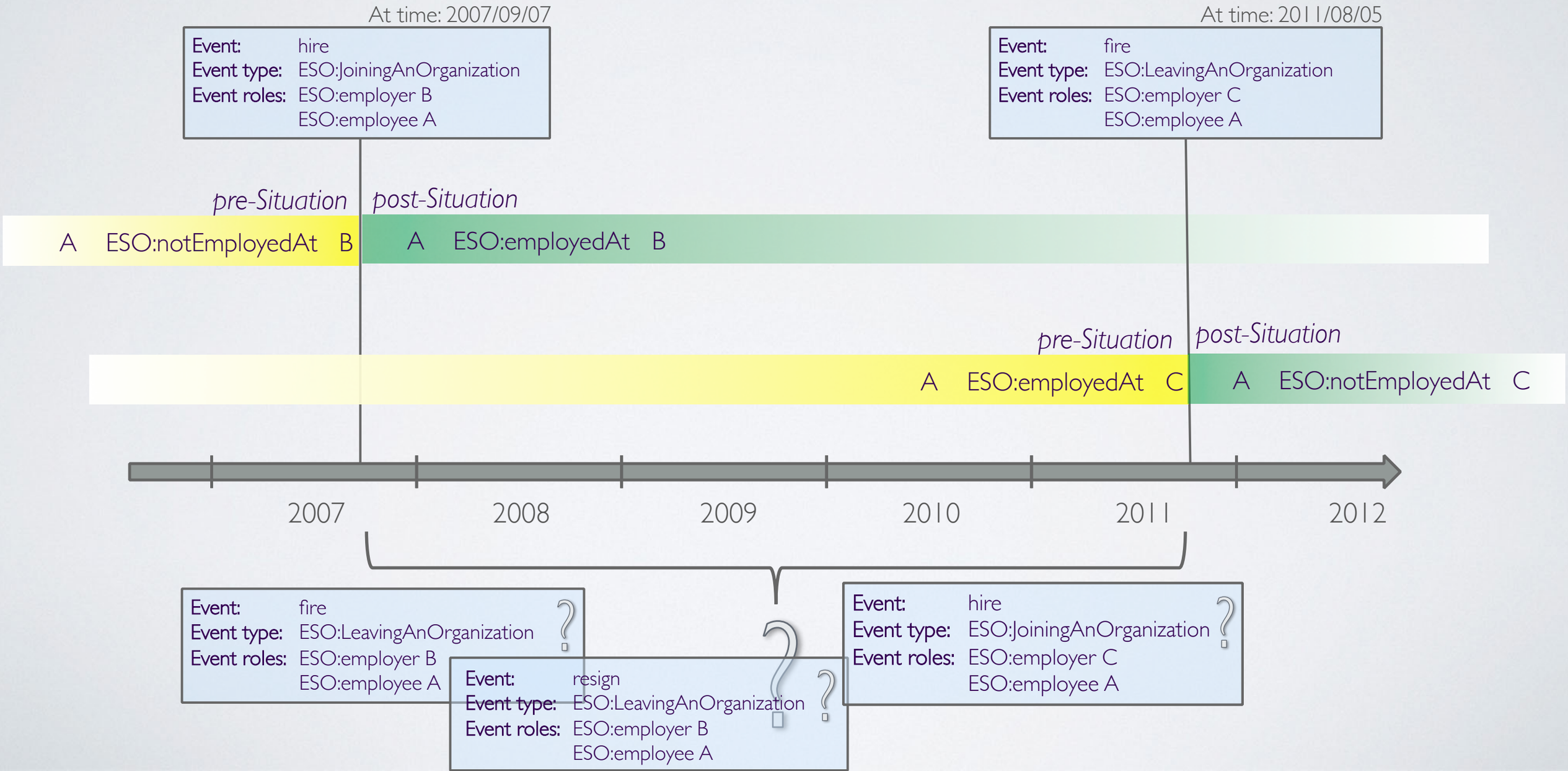
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

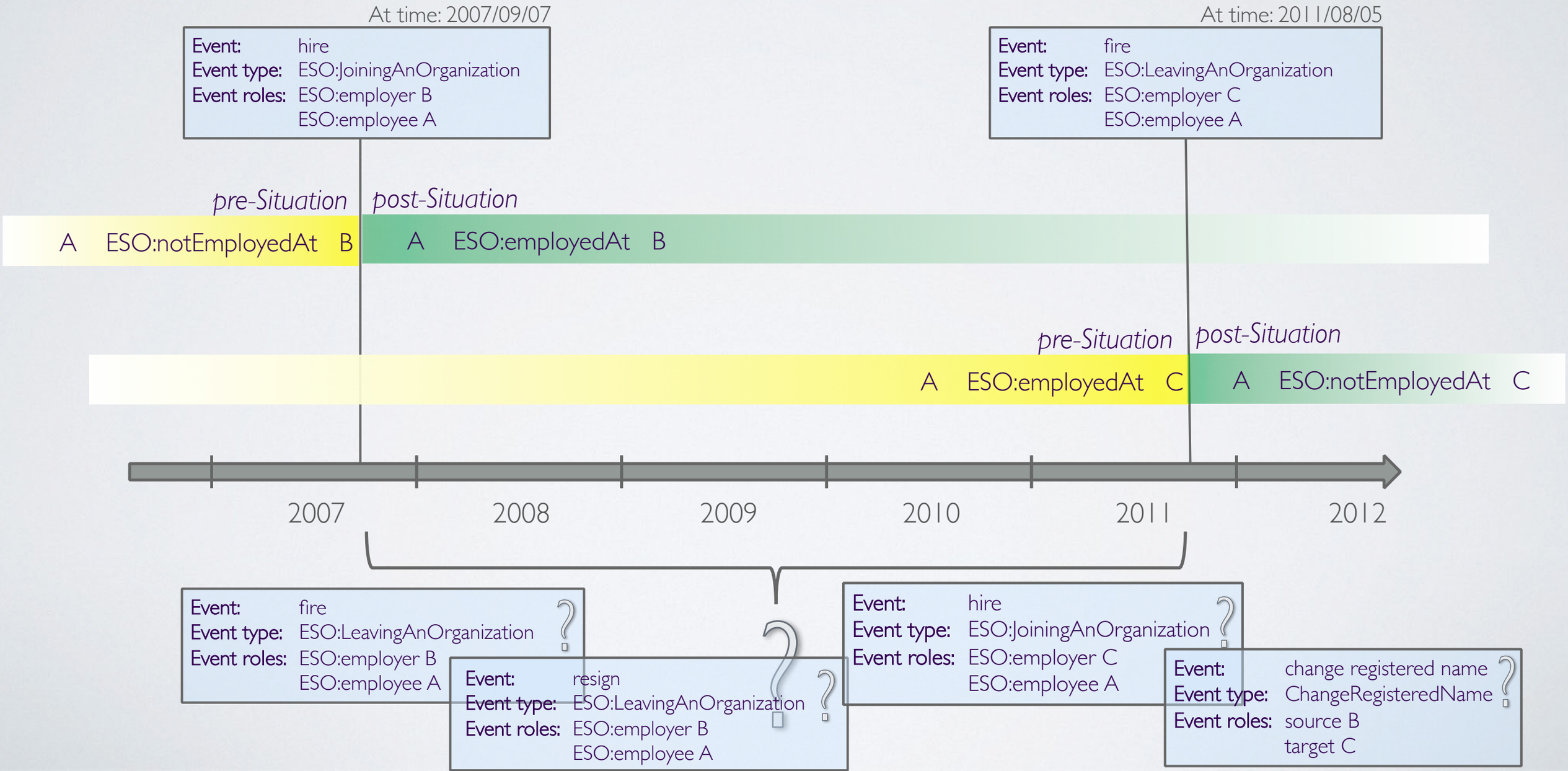
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

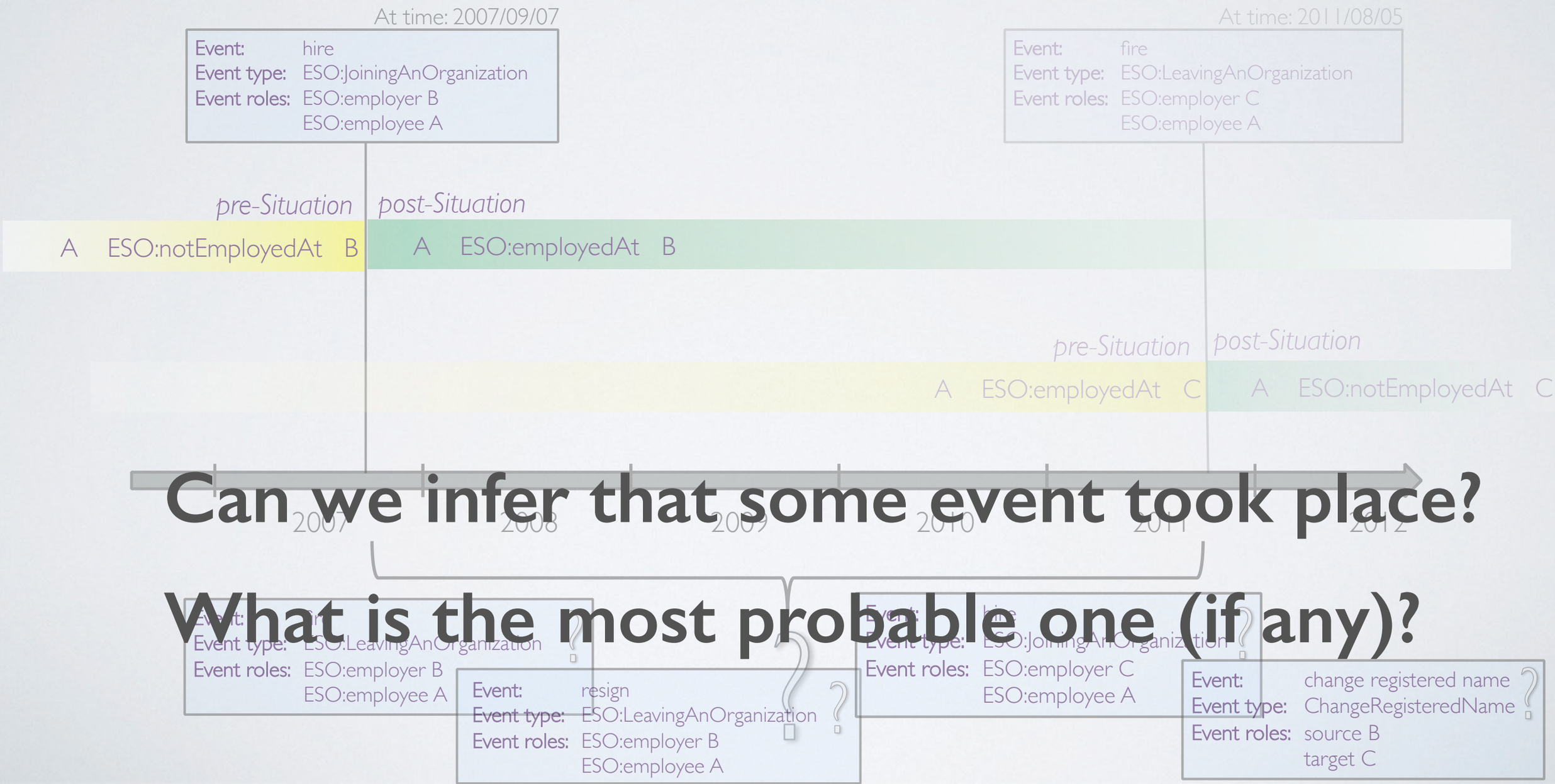
- What knowledge can be additionally inferred from what mentioned in text?



KnowledgeStore

Beyond "Asserted" Knowledge…

- What knowledge can be additionally inferred from what mentioned in text?



Can we infer that some event took place?

What is the most probable one (if any)?

KnowledgeStore

Crystallising Extracted Knowledge

- **Knowledge Crystallisation**
  - When can a fact **automatically extracted** from information extraction tools be considered as "**Background Knowledge**"?
- Aspects to be considered:
  - "**cleaning**" of the data
  - **consistency** of event
    - "John" has been fired today from company "Alpha AGf", but the KnowledgeStore contains the fact that "Alpha AGf" was closed 4 years ago
  - event **matching** and **integration**
    - augment background knowledge entities with facts extracted from the pipeline
- How To Tackle This? Combining **Statistical** and **Crisp** Reasoning?

KnowledgeStore

Beyond Text…

- Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…
  - E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

KnowledgeStore

Beyond Text…

- Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…
  - E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.

KnowledgeStore

Beyond Text…

- Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…
  - E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

"Stripes," Dole, 1996

He (Bill Clinton) quickly
the first civilian comm
chief to salute his marin
while entering or exiting an
aircraft.

KnowledgeStore

Beyond Text…

- Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…
  - E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

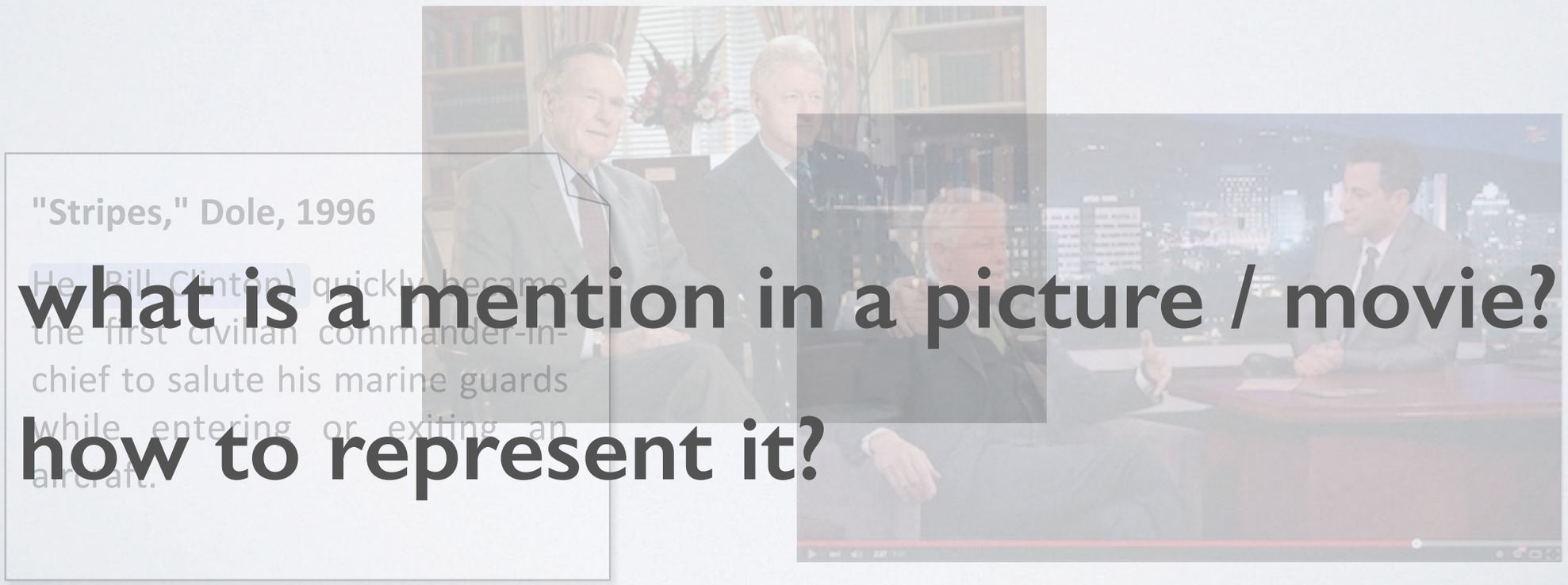
"Stripes," Dole, 1996

He (Bill Clinton) quickly
the first civilian comm
chief to salute his marin
while entering or exiting an
aircraft.

KnowledgeStore

Beyond Text…

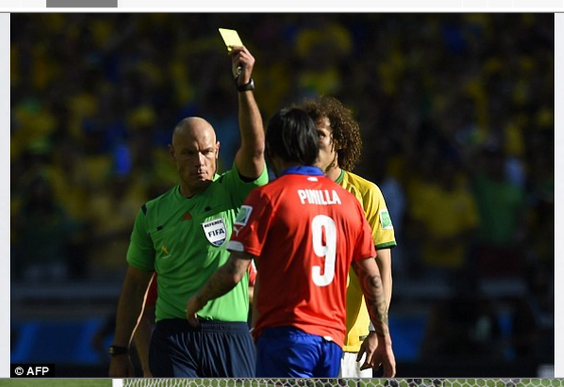
- Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…
  - E.g., retrieve all documents and media where Bill Clinton makes statements about the Army



KnowledgeStore

Beyond Text… and even more…

- Interpreting / extracting / aligning knowledge from different media (e.g., video, commentary, text, …)

Frame	Commentary	Knowledge
	"Sanchez, Sanchez,. . . goal. Sanchez equalizes for Chile"	dbpedia:Alexis_Sanchez scorestAt 32min
	"Yellow card for the Chilean defender"	dbpedia:Mauricio_Pinilla yellowCardAt 102min
	"Now is Marcelo turn, to kick the fourth penalty" "Marcelo. . . Goal"	dbpedia:Marcelo_Vieira kicks SuppPenalty4 SuppPenalty4 leadsTo goal

Thank You! Questions?

The 知識ストア Team

Roldano Cattoni, Francesco Corcoglioniti, Bernardo Magnini, Alessio Palmero Aprosio, Mohammed Qwaider, Marco Rospocher, Luciano Serafini

— http://knowledgestore.fbk.eu —