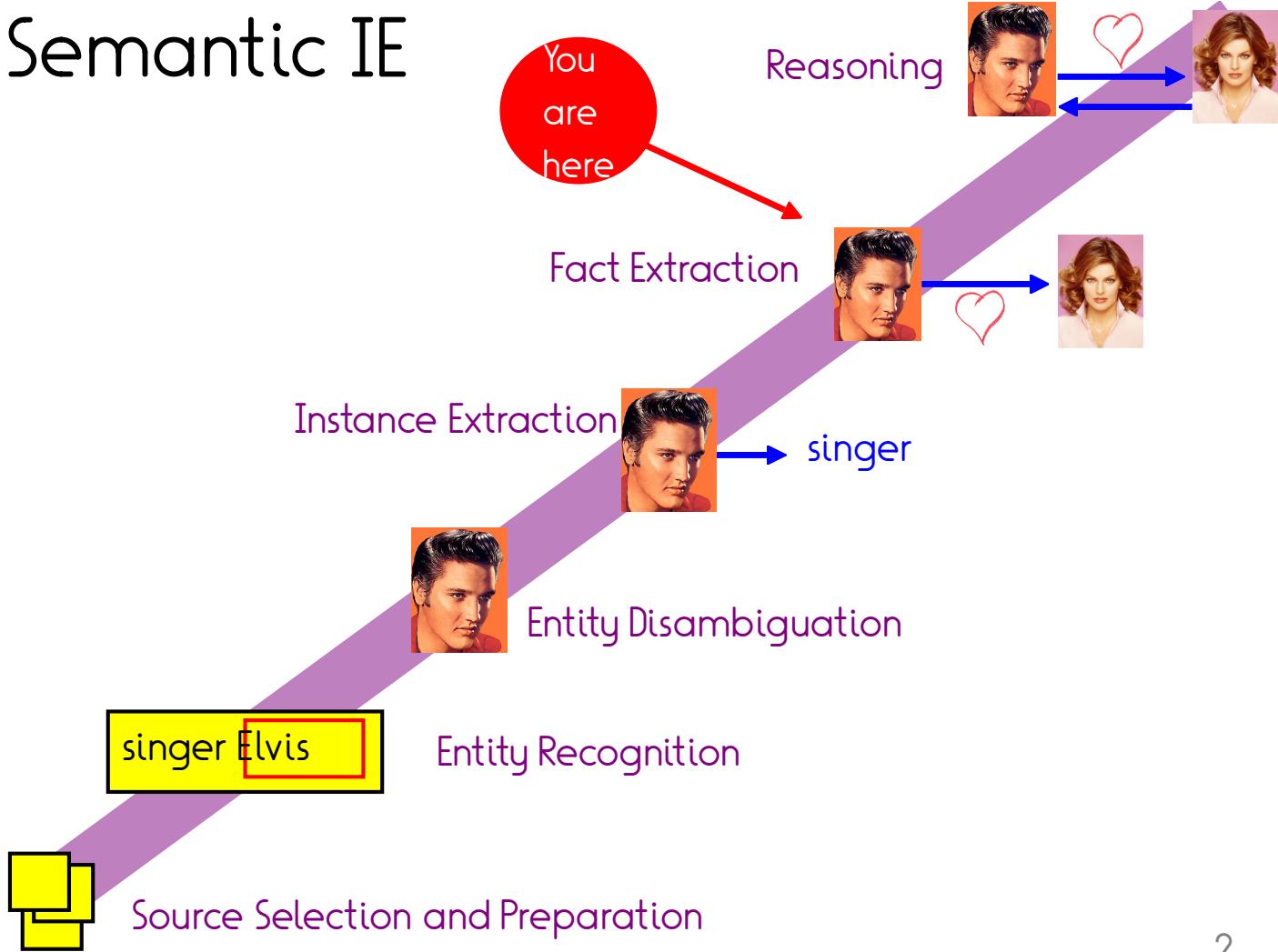


# The DIPRE Algorithm

Nada Mimouni

Based on slides by:  
Fabian M. Suchanek

# Semantic IE

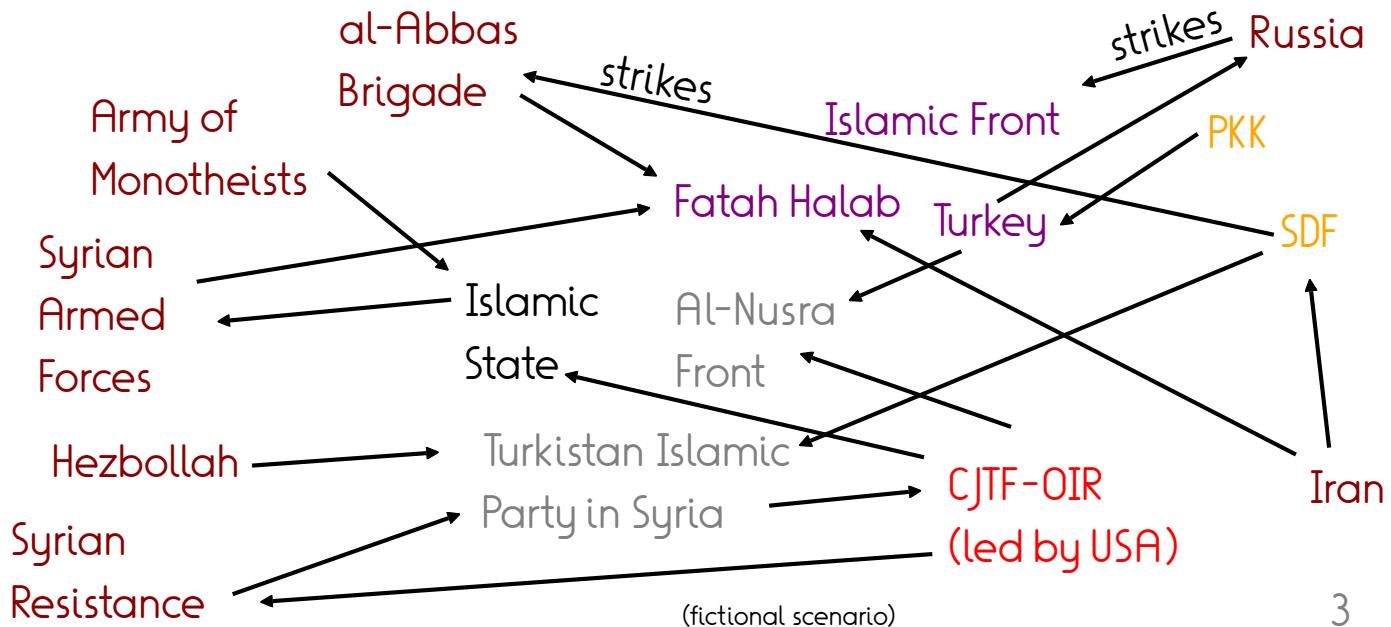


# Def: Fact Extraction

Fact extraction is the extraction of facts about entities from a corpus.

For now, we concentrate on facts with a single relation.

في أوائل نوفمبر، جرت اشتباكات بين الجيش السوري الحر و قوات الأمن في حمص مما ساهم في توسيع الحصار الشمالي طويلاً في العديد من الأحياء، كانت المقاومة في حمص أكبر بكثير من البلديات والمدن الأخرى، وحمص، فقد فشلت العمليات في حمص حتى الآن في قمع الاضطرابات. في نوفمبر تشرين الثاني ديسمبر 2011،

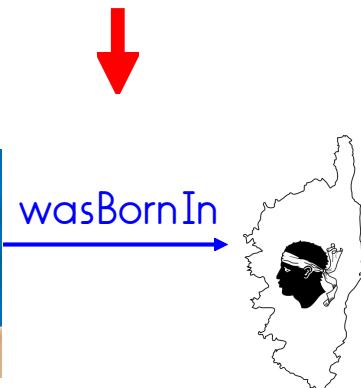


# Fact Extraction, happier example

Fact extraction is the extraction of facts about entities from a corpus.

For now, we concentrate on facts with a single relation.

Alizée kommt aus Corsica.



For the computer,  
the corpus is  
completely  
incomprehensible –  
as if it were written  
in a foreign language!

The extracted facts,  
on the other hand,  
use well-defined  
relations.

# Def: Extraction Pattern

An extraction pattern for a binary relation  $r$  is a string that contains two place-holders  $X$  and  $Y$ , indicating that two entities stand in relation  $r$ .

$X$  kommt aus  $Y$ .

$X$  wurde geboren in  $Y$ .

$X$  stammt aus  $Y$ .

$X$  ist gebürtig aus  $Y$ .

Extraction  
patterns



# Where do we get the patterns?

- Option 1: Manually compile patterns.



Public Domain

- Option 2: Manually find the patterns in texts

Angela Merkel stammt aus Hamburg. Sie ist  
seit 2005 Kanzlerin von Deutschland und...

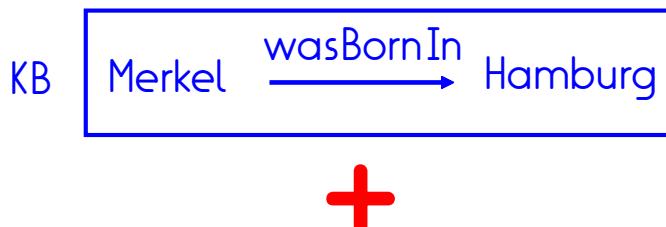


" $X$  stammt aus  $Y$ " is a pattern for  $\text{bornIn}(X, Y)$

- Option 3: Pattern deduction

# Def: Pattern Deduction

Given a corpus, and given a KB, pattern deduction is the process of finding extraction patterns that produce facts of the KB when applied to the corpus.

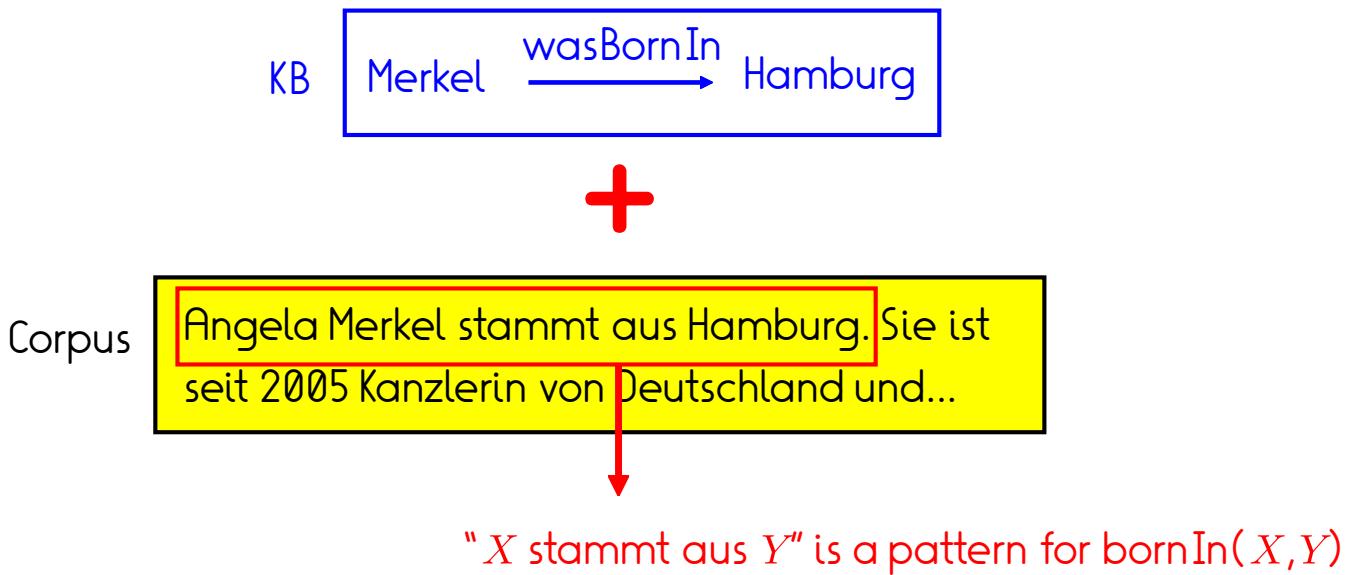


Corpus

Angela Merkel stammt aus Hamburg. Sie ist seit 2005 Kanzlerin von Deutschland und...

# Def: Pattern Deduction

Given a corpus, and given a KB, pattern deduction is the process of finding extraction patterns that produce facts of the KB when applied to the corpus.



# Def: Pattern Application

Given a corpus, and given a pattern, **pattern application** is the process of finding the facts produced by the pattern.



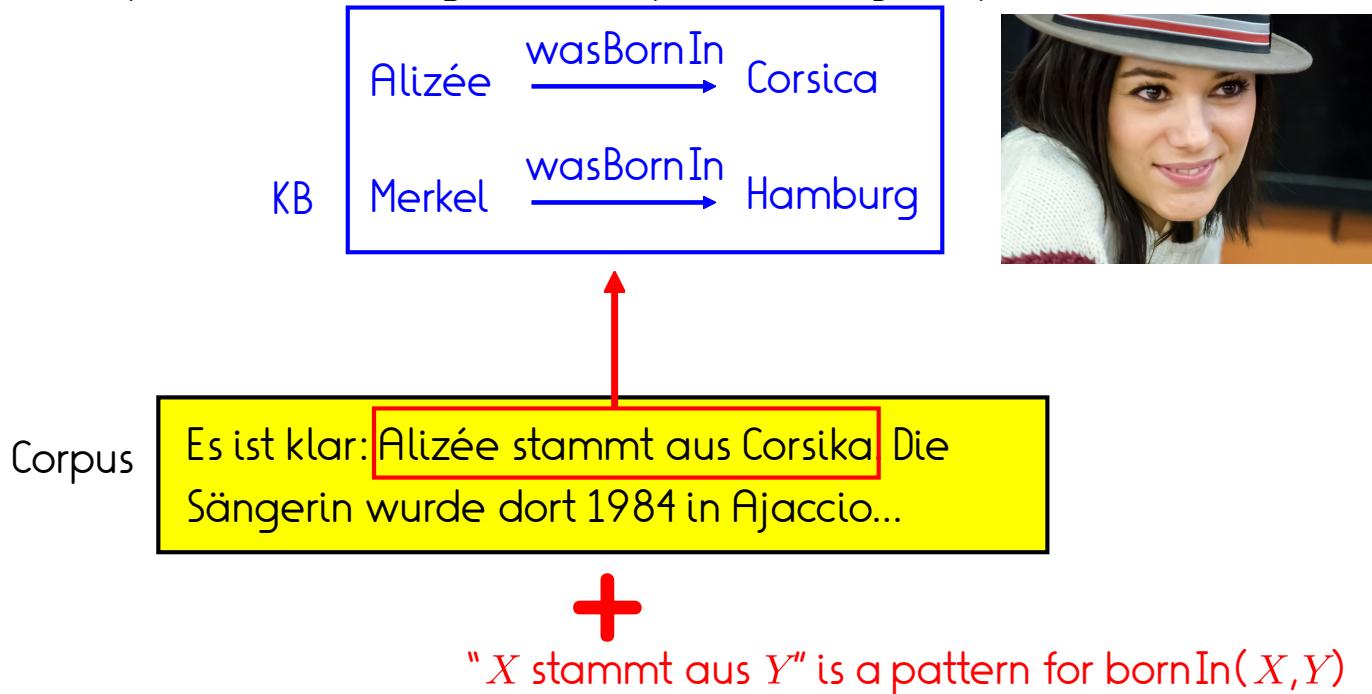
Corpus

Es ist klar: Alizée stammt aus Corsika Die Sängerin wurde dort 1984 in Ajaccio...

" $X$  stammt aus  $Y$ " is a pattern for  $\text{bornIn}(X, Y)$

# Def: Pattern Application

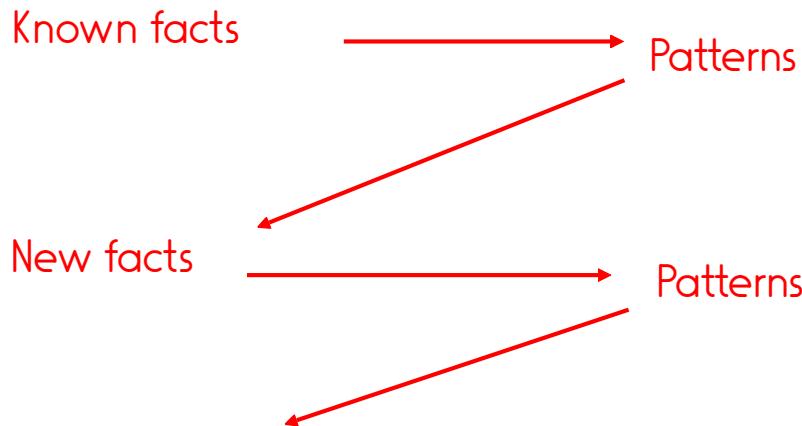
Given a corpus, and given a pattern, **pattern application** is the process of finding the facts produced by the pattern.



# Def: Pattern iteration/DIPRE

Pattern iteration (also: DIPRE) is the process of repeatedly

- applying pattern deduction
- using the patterns to find new facts
- ... thus continuously augmenting the KB.



# Example: DIPRE

KB

Obama            chases      Osama

Obama hetzt Osama. Tom jagt Jerry. Tom hetzt Jerry.

# Example: DIPRE

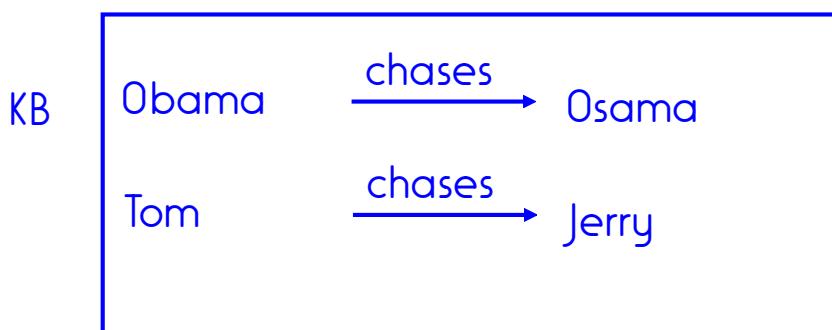
KB

Obama       $\xrightarrow{\text{chases}}$       Osama

Obama hetzt Osama. Tom jagt Jerry. Tom hetzt Jerry.

=> "X hetzt Y" is a pattern for  $\text{chases}(X, Y)$

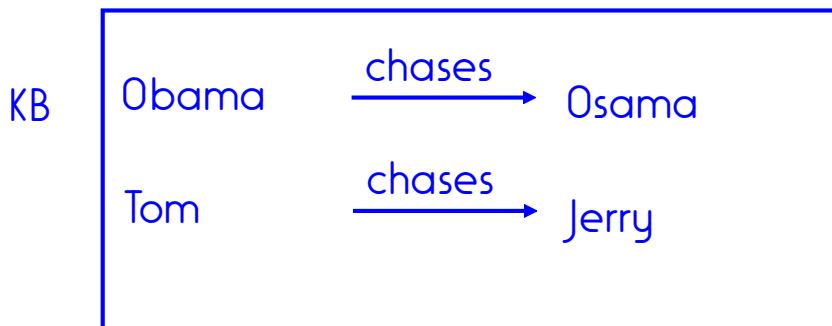
# Example: DIPRE



Obama hetzt Osama. Tom jagt Jerry. Tom hetzt Jerry.

=> "X hetzt Y" is a pattern for  $\text{chases}(X, Y)$

# Example: DIPRE



Obama hetzt Osama. Tom jagt Jerry. Tom hetzt Jerry.

=> "X hetzt Y" is a pattern for  $\text{chases}(X, Y)$

=> "X jagt Y" is a pattern for  $\text{chases}(X, Y)$

# Task: DIPRE



Michelle ist verheiratet mit Barack.

Merkel ist die Frau von Sauer.

Michelle ist die Frau von Barack.

Priscilla ist verheiratet mit Elvis.

# Example: Patterns in NELL

NELL (Never Ending Language Learner) is an information extraction project at Carnegie Mellon University.

Apple produced MacBook

- CPL @851 (100.0%) on 28-jun-2014 [ "arg1 claims the new arg2" "arg1 were to release arg2" "arg2 are trademarks of arg1" "arg1 Store to get arg2" "arg1 AppleCare Protection Plan for arg2" "arg1 will announce a new arg2" "arg1 would release a new arg2" "arg2 Pro now includes arg1" "arg2 nano at arg1" "arg1 will release a new arg2" "arg1 announced their new arg2" "arg1 releases a new version of arg2" "arg1 already sells arg2" "arg1 announced that the new arg2" "arg1 recently switched their arg2" "arg2 and iPod are trademarks of arg1" "arg1 TV and arg2" "arg2 Pro from arg1" "arg1 says the new arg2" "arg1 unveils new arg2" "arg1 iMac and arg2" "arg1 has now released arg2" ] using (apple, macbook)

>skip summary

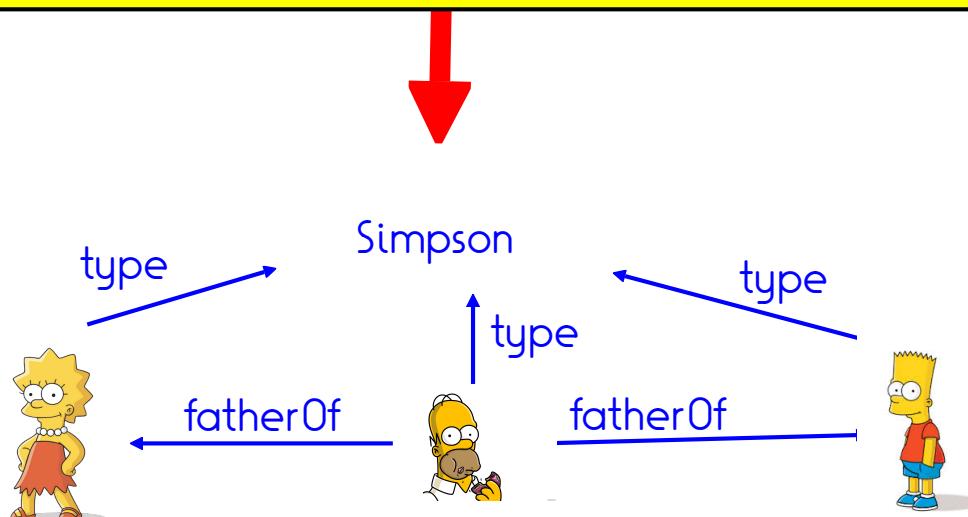
# Summary: Information Extraction

Congratulations, you can now transform (parts of) natural language text into structured information!

I love Simpsons such as Bart, Lisa, and Homer.

Homer is the father of Bart.

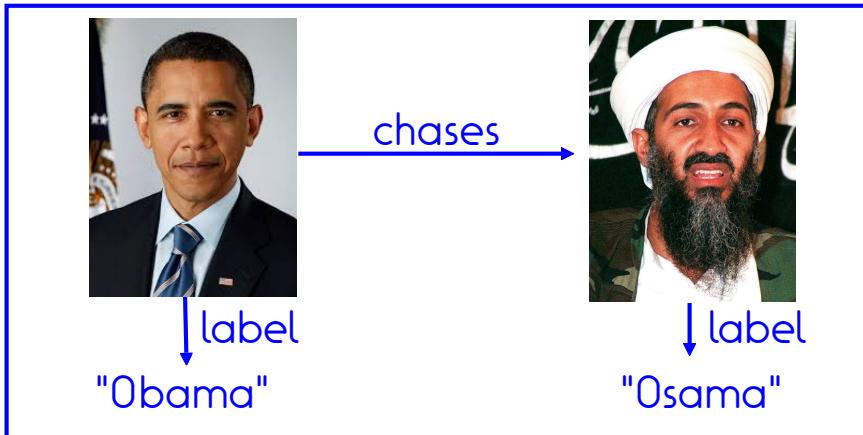
Homer is the father of Lisa.



>Problems

# We use labels to find patterns

KB



Corpus

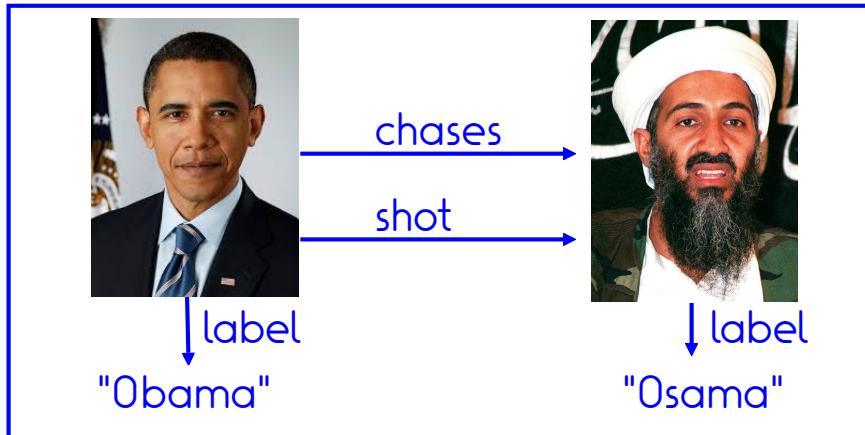
Obama verfolgt Osama.

=> "X verfolgt Y" is a pattern for  $\text{chases}(X, Y)$

>Problems

# Different Relations

KB



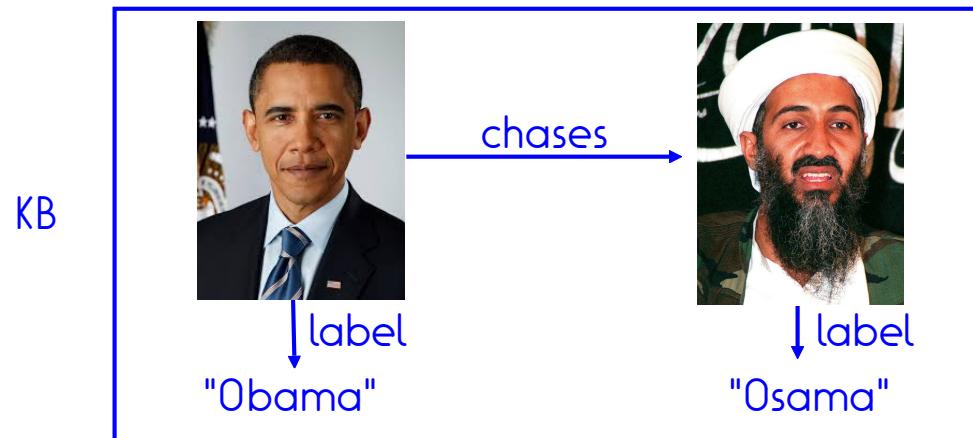
Corpus

Obama verfolgt Osama.

=> "X verfolgt Y" is a pattern for  $\text{chases}(X, Y)$  for  $\text{shot}(X, Y)$ ?

>Problems

# Phrase Structure can be a Problem



Corpus

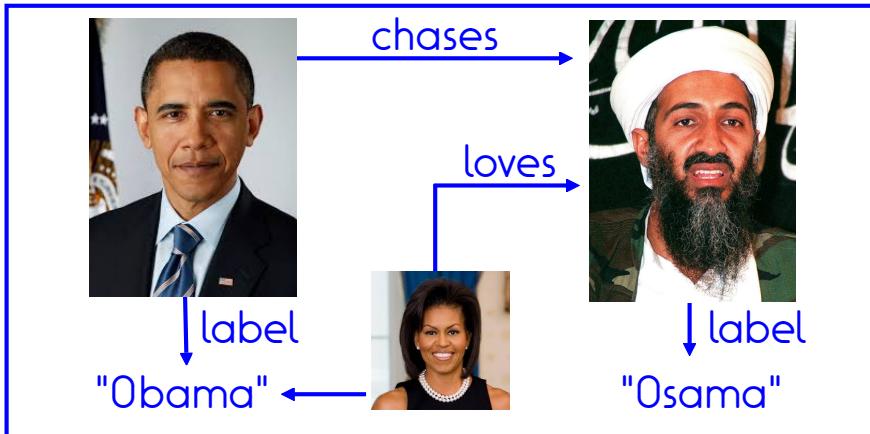
Obama hat Osama verfolgt.

=> "X hat Y" is a pattern for  $\text{chases}(X, Y)$ ?

>Problems

# Ambiguity is a Problem

KB



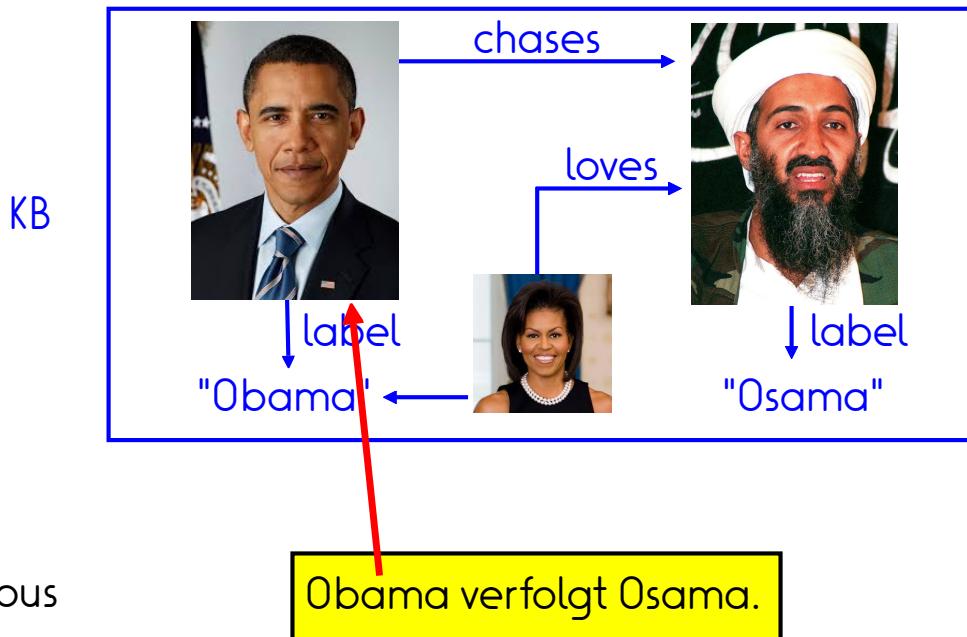
Corpus

Obama verfolgt Osama.

=> "X verfolgt Y" is a pattern for  $\text{chases}(X,Y)$  for  $\text{loves}(X,Y)$ ?

>Problems

# Disambiguation helps



=> "X verfolgt Y" is a pattern for  $\text{chases}(X, Y)$

>Problems

# Confidence of a pattern

The confidence of an extraction pattern is the number of matches that produce known facts divided by the total number of matches.

Pattern produces mostly new facts  
=> risky

Pattern produces mostly known facts  
=> safe

# Simple word match is not enough

Coyote invents a wonderful machine.

+

"X invents a Y"



=

invents(Coyote, wonderful)

>Problems

# Patterns may be too specific

Coyote invents a wonderful machine.

+

"X invents a gorgeous Y"

=

~~-invents(Coyote,machine)-~~



# References

Brin: Extracting Patterns and Relations from the WWW

Agichtein: Snowball

- >ie-by-reasoning
- >pos-tagging
- >semantic-web
- >security