

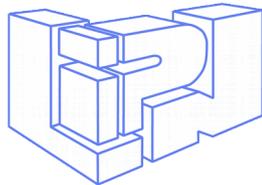
# Cartografía de la diáspora altamente calificada aplicando técnicas de minería semántica a resultado de búsquedas en la web

**Jorge García Flores**  
**Laboratorio de Informática de París Norte**  
**CNRS – Universidad de París 13**

# ECOS NORD 2017-2021

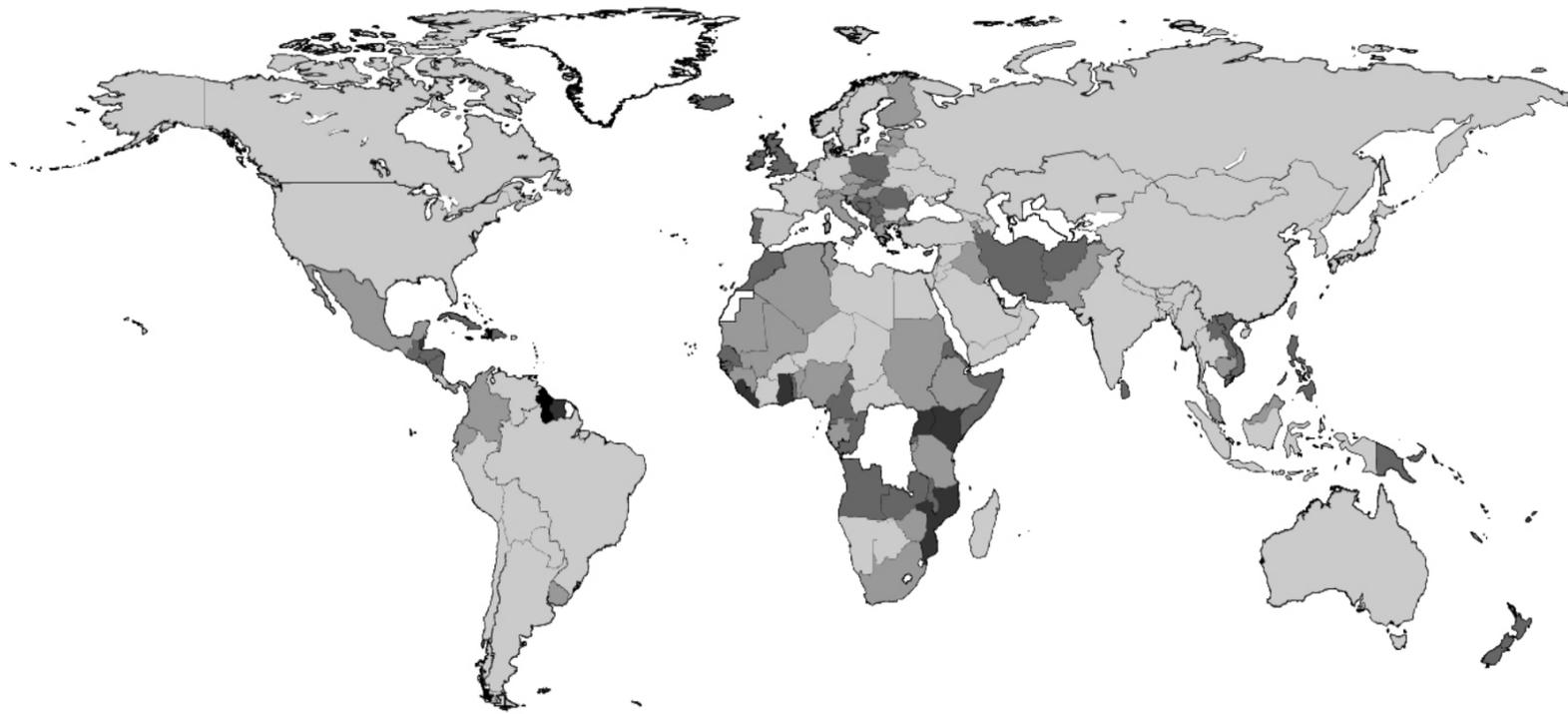


- Aldo Gangemi (LIPN)
- Thierry Charnois (LIPN)
- Dana Diminescu (Télécom-Paristech)
- Jean-Baptiste Meyer (IRD)
- Pegah Alizedeh (LIPN)

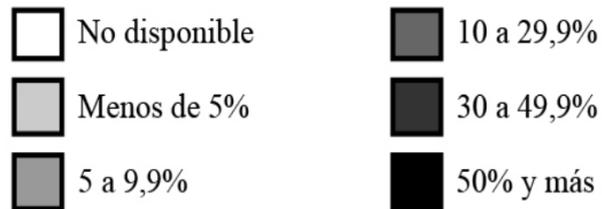


- Luis Pineda (IIMAS)
- Iván Vladimir Meza (IIMAS)
- Gibrán Fuentes (IIMAS)
- Fernando Lozano (CRIM)
- Josué Urbina (FI-IIMAS)
- Karl Posthuma (FI-IIMAS)





Tasa de emigración calificada



*Fuente:* elaboración propia con datos de *Brain Brain Data Set*. Beine, M. y otros (2006)

Fuente:



[Home](#)

[Contact](#)

[Description](#)

[Research](#)

[Publications](#)

**[Members](#)**

[Projects](#)

[Seminars](#)

[News](#)

[Job Offers](#)

[Demos](#)

[Software](#)

## Diana MATEUS



**I'm PhD Candidate, working at** the Perception Team - INRIA Rhone Alpes (Grenoble 2004-2008) under the supervision of Radu Horaud. My current research interests concern the registration and segmentation of Articulated Shapes trough spectral embedding methods. I've previously work on a project to estimate the Sparse

Scene-Flow of a scene (tracking of 3-D points and surfels) from image sequences obtained with a multiple-camera set-up. I've also studied the segmentation of the resultant scene-flow trajectories in groups of coherent motion.

Before coming to France, I got an Electronic Engineer BsC, (Bogota-Colombia 1996-2002). Then I completed a research master (DEA) in Automation Systems (Toulouse 2003-2004). During this time, I made an internship at the LAAS laboratory -  
PIA team, under the direction of Michel Doy, and in

## Contact

- >>> INRIA Rhône-Alpes  
655, avenue de  
l'Europe  
38330 Montbonnot,  
France
- >>> [diana.mateus@inrialpes.fr](mailto:diana.mateus@inrialpes.fr)
- >>> Tel: +33 476 61 55 85

## Links

- >>> [All publications](#)
- >>> [CV](#)
- >>> [Website](#)
- >>> [Research projects](#)

¿Hasta qué punto la realidad del web permite responder preguntas acerca de la realidad fuera del web?

[Google.org home](#)

Flu Trends

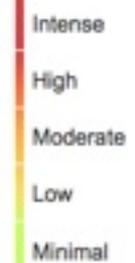
Select country/region ▾

[Home](#)

[How does this work?](#)

[FAQ](#)

Flu activity

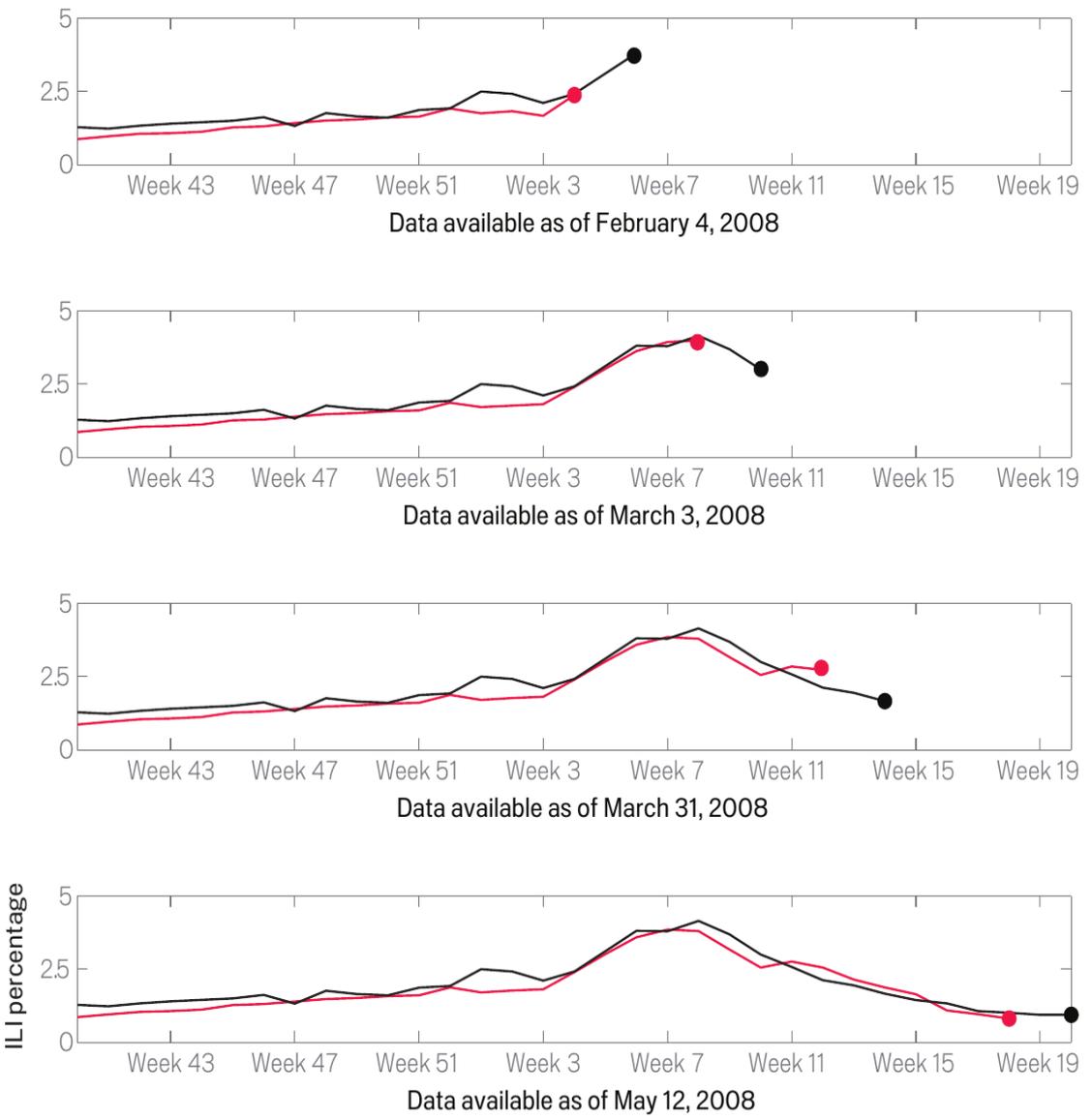


## Explore flu trends around the world

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »](#)



[Download world flu activity data](#)



¿Hay entre la emigración calificada algún mexicano experto en frenado por corrientes de Foucault ?

# Estado del conocimiento:

[Garcia Flores et al. 2012]

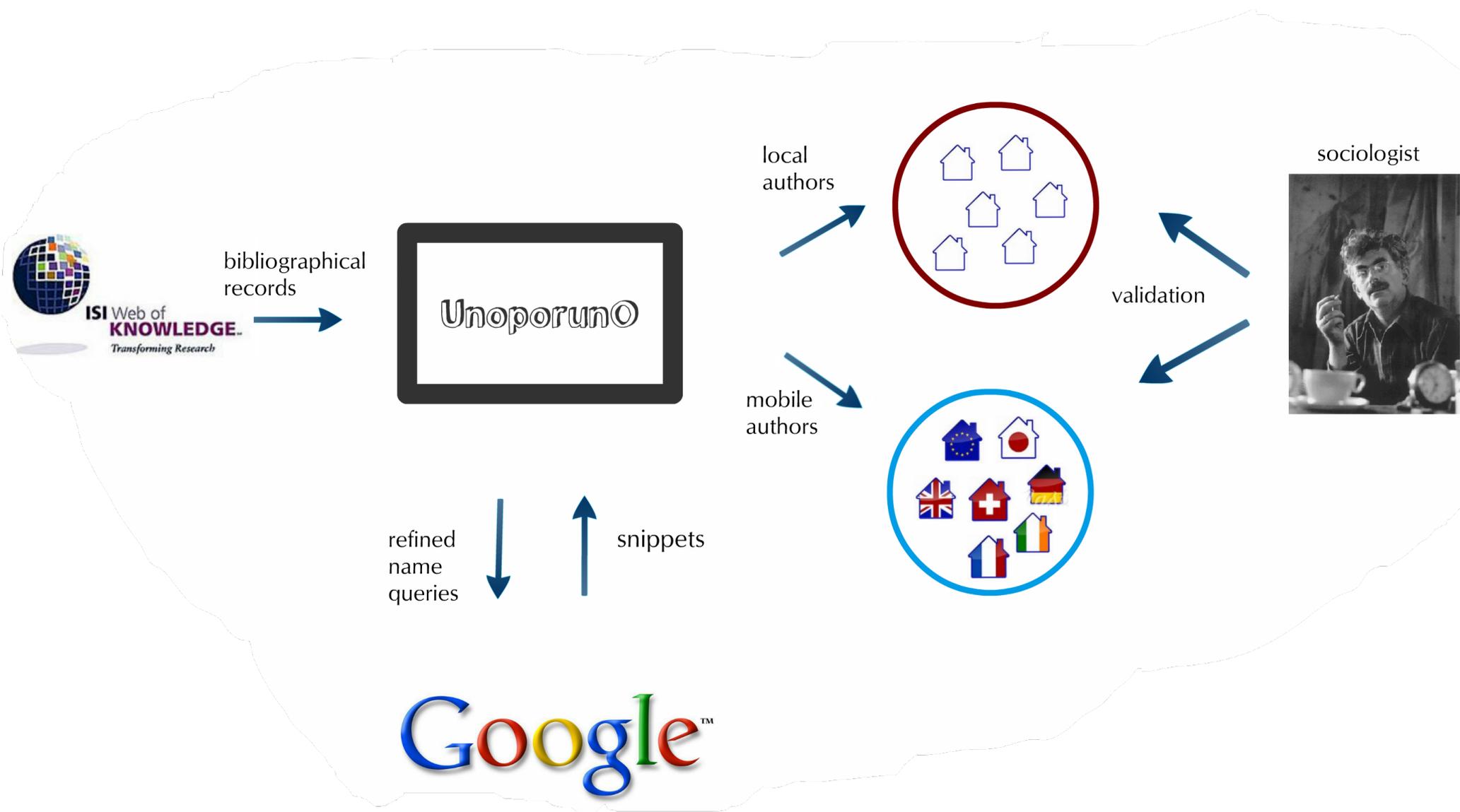
[Bordea & Buitelaar 2012]

[Diminescu, 2008]

[Artiles et al. 2010]

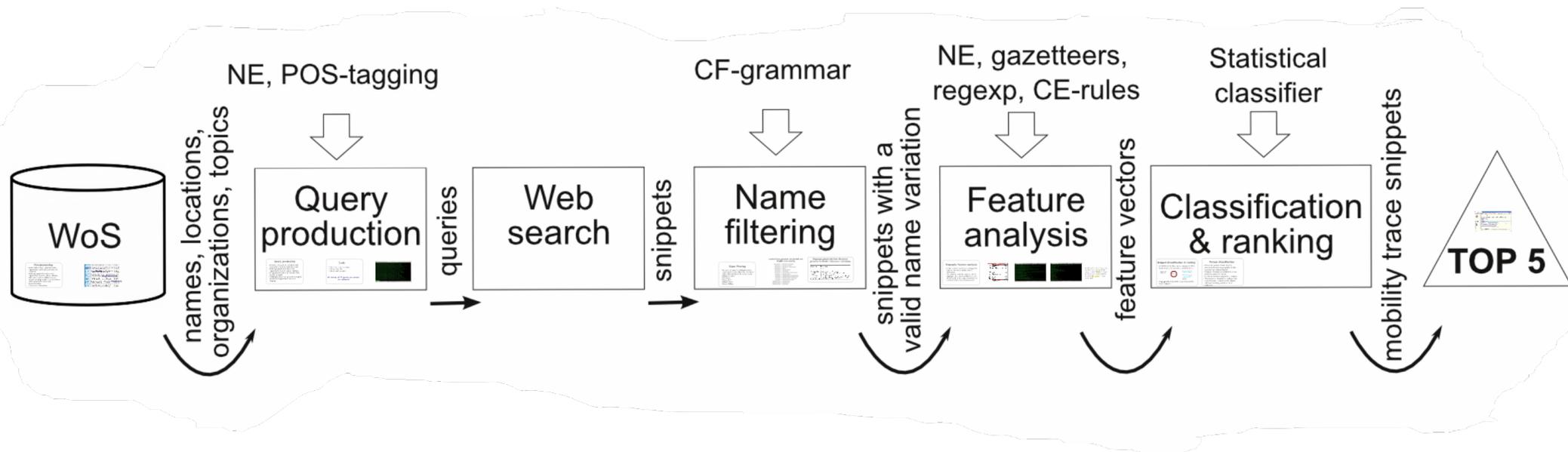
[Di Marco & Navigli,  
2013]

- unoporunO
- expert mining
- e-diaspora Atlas
- web people search
- clustering de resultados de búsquedas en el web



**Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis.** J. J. Garcia Flores, Pierre Zwigenbaum, Zhao Yue and W.A. Turner (2012) T In H. Ishara & K. Kanzaki, *Advances in Natural Language Processing (JapTAL 2012)* Springer Lecture Notes in Computer Sciences, vol. 7614, , pp. 180-191.

# Arquitectura de UnoporunO:



# Generación de consultas Google

```
INFO:root:busca::querying Jorge B Aquino last_query_time:: Fri Oct 7 03:03:49 2011
INFO:root:busca::querying Jorge B Aquino Buenos Aires last_query_time:: Fri Oct 7 03:04:50 2011
INFO:root:busca::querying Jorge B Aquino Buenos Aires last_query_time:: Fri Oct 7 03:06:52 2011
INFO:root:busca::querying Jorge B Aquino Argentina last_query_time:: Fri Oct 7 03:07:22 2011
INFO:root:busca::querying Jorge B Aquino Argentina last_query_time:: Fri Oct 7 03:08:23 2011
INFO:root:busca::querying Jorge B Aquino novel synergistic combination last_query_time:: Fri Oct 7 03:09:23 2011
INFO:root:busca::querying Jorge B Aquino cyclophosphamide last_query_time:: Fri Oct 7 03:11:25 2011
INFO:root:busca::querying Jorge B Aquino gene transfer last_query_time:: Fri Oct 7 03:12:26 2011
INFO:root:busca::querying Jorge B Aquino colorectal carcinoma last_query_time:: Fri Oct 7 03:14:27 2011
INFO:root:busca::querying Jorge B Aquino mice last_query_time:: Fri Oct 7 03:16:29 2011
INFO:root:busca::querying Jorge B Aquino Gene Therapy Laboratory last_query_time:: Fri Oct 7 03:17:30 2011
INFO:root:busca::querying Jorge B Aquino Gene Therapy Laboratory last_query_time:: Fri Oct 7 03:18:31 2011
INFO:root:busca::querying Jorge B Aquino Liver Unit last_query_time:: Fri Oct 7 03:19:32 2011
INFO:root:busca::querying Jorge B Aquino Liver Unit last_query_time:: Fri Oct 7 03:20:32 2011
INFO:root:busca::querying Jorge B Aquino School of Medicine last_query_time:: Fri Oct 7 03:22:34 2011
INFO:root:busca::querying Jorge B Aquino School of Medicine last_query_time:: Fri Oct 7 03:24:36 2011
INFO:root:busca::querying Jorge B Aquino Austral University last_query_time:: Fri Oct 7 03:26:07 2011
INFO:root:busca::querying Jorge B Aquino Austral University last_query_time:: Fri Oct 7 03:27:38 2011
INFO:root:busca::querying Jorge Reinheimer last_query_time:: Fri Oct 7 03:28:41 2011
INFO:root:busca::querying Jorge Reinheimer last_query_time:: Fri Oct 7 03:30:12 2011
INFO:root:busca::querying Jorge Reinheimer Santa Fe last_query_time:: Fri Oct 7 03:31:12 2011
INFO:root:busca::querying Jorge Reinheimer Santa Fe last_query_time:: Fri Oct 7 03:32:44 2011
INFO:root:busca::querying Jorge Reinheimer Argentina last_query_time:: Fri Oct 7 03:34:15 2011
INFO:root:busca::querying Jorge Reinheimer Argentina last_query_time:: Fri Oct 7 03:35:46 2011
INFO:root:busca::querying Jorge Reinheimer Characterization last_query_time:: Fri Oct 7 03:36:46 2011
INFO:root:busca::querying Jorge Reinheimer new virulent phage last_query_time:: Fri Oct 7 03:38:18 2011
INFO:root:busca::querying Jorge Reinheimer MLC-A last_query_time:: Fri Oct 7 03:39:18 2011
INFO:root:busca::querying Jorge Reinheimer Lactobacillus paracasei last_query_time:: Fri Oct 7 03:41:20 2011
INFO:root:busca::querying Jorge Reinheimer Instituto de Lactologia Industrial (INLAIN) last_query_time:: Fri Oct 7 03:42:21
2011
```

William A. Turner

WA Turner

Turner, W.A.

William Turner

Turner, William

# Filtro onomástico

PersonName → FirstName LastName  
FirstName → FirstName Initial  
FirstName → FirstName CommonName  
FirstName → Initial  
FirstName → CommonName  
FirstName → TypoName  
FirstName → CommonName ParticleName  
LastName → MainLastName  
LastName → MainLastName CommonName  
LastName → MainLastName ParticleName  
LastName → MainLastName TypoName  
LastName → MainLastName Initial  
MainLastName → CommonName  
MainLastName → ParticleName  
MainLastName → TypoName

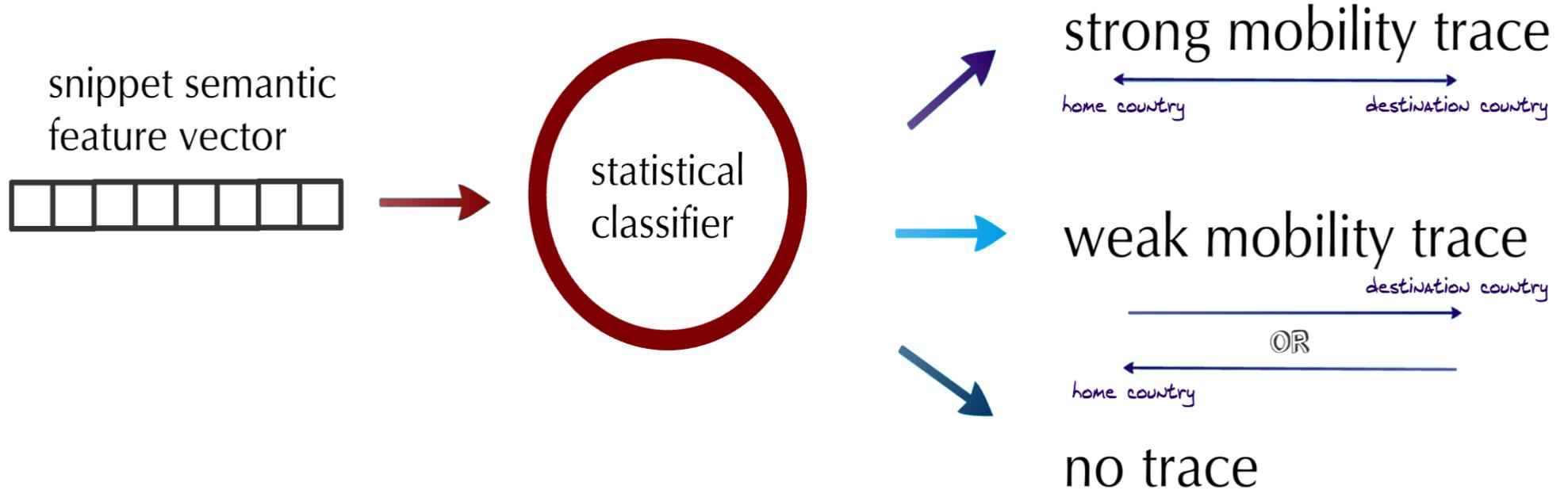
# Filtro onomástico

Operation	Description
$n$ : name	$n \rightarrow N$
$a$ : surname	$a \rightarrow A$
$C$ : compression	$CnLa(\text{Noe Lopez}) \rightarrow N \setminus .? \text{Lopez}$
$E$ : expansion	$EnLa(\text{Eva M Perez}) \rightarrow \text{Eva M}[a - z]^+ \text{Perez}$
$L$ : literal	$LnLa(\text{Noe Lopez}) \rightarrow \text{Noe Lopez}$
$X$ : extra element	$LnXLa(\text{Eva M Perez}) \rightarrow \text{Eva M}[A - Z][a - z]^+ \text{Perez}$
$V$ : inversion	$VCnLa(\text{Eva M Perez}) \rightarrow \text{Perez},? +E[\setminus.]?[-]?M \setminus .?$
$SI$ : suppress initial	$SInSIa(\text{Noe J Lopez F}) \rightarrow \text{Noe Lopez}$

# Atributos semánticos (*features*)

	<b>Feature</b>	<b>Type</b>
<b>biographical features</b>	<b>Biographical sentence</b>	regex
	<b>Academic degree</b>	regex
	<b>Profession</b>	gazet
	<b>nationality (Latin America)</b>	gazet
	<b>nationality (not Latin America)</b>	gazet
	<b>e-mail</b>	regex
<b>geographical features</b>	<b>City</b>	regex
	<b>Country</b>	regex
	<b>Organization</b>	regex
	<b>Organization acronym</b>	gazet
<b>source document features</b>	<b>CV</b>	regex
	<b>PhD tesis</b>	regex
	<b>LinkedIn</b>	regex
	<b>Publication</b>	regex
	<b>Person name in URL</b>	regex
	<b><i>Feature count</i></b>	<i>count</i>

# Atributos semánticos (*features*)



# Corpus de evaluación

Gold standard	Training set	Testing set
Researchers	748	50
Home country	Argentina, Uruguay, Colombia	Uruguay
Queries	11253	609
Filtered snippets	58220	3476
Mobility traces	<b>611</b>	<b>134</b>
Home/destination country traces	<b>1691</b>	<b>1091 (h)</b> <b>252 (d)</b>
No trace	55918	1999

# Evaluation 1/2



50 researchers (Arg, Uru, Col)  
25 mobiles / 25 locals



1 sociologist

88% of persons found



78% of persons found

# Evaluation 2/2



30 researchers  
15 mobiles/15 locals



6 evaluators

avg. of persons found = 77%  
inter-annotator agreement = 60%

## Dos experimentos en curso:

- 1: ¿A qué se dedican hoy en día 10 000 exbecarios de Conacyt (1995-2006)
- 2: ¿A qué se dedican hoy en día 3000 miembros de grupos Facebook de mexicanos en Francia?

## Dos experimentos en curso:

1: ¿A qué se dedican hoy en día 10 000  
exbecarios de Conacyt (1995-2006)

\_\_regresaron

\_\_se quedaron allá

\_\_regresaron y se quedaron allá ;)

Conacyt calcula que 30% de sus exbecarios al extranjero no regresan...

¿qué cifra arrojará la web?

Experimento con aprendizaje reforzado:  
completar la base de datos *Conacyt\_95-06*:

1:

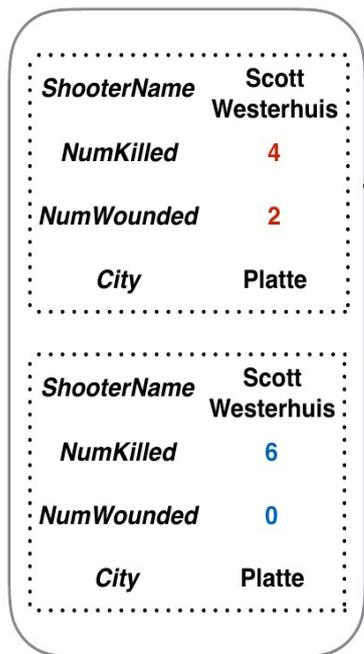
\_\_consultar en Google sólo el nombre  
\_\_encontrar institución y país

2:

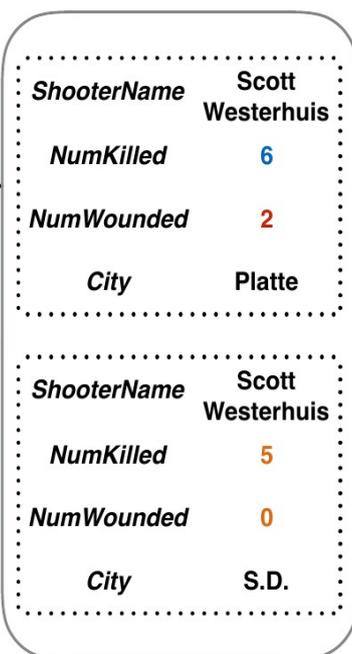
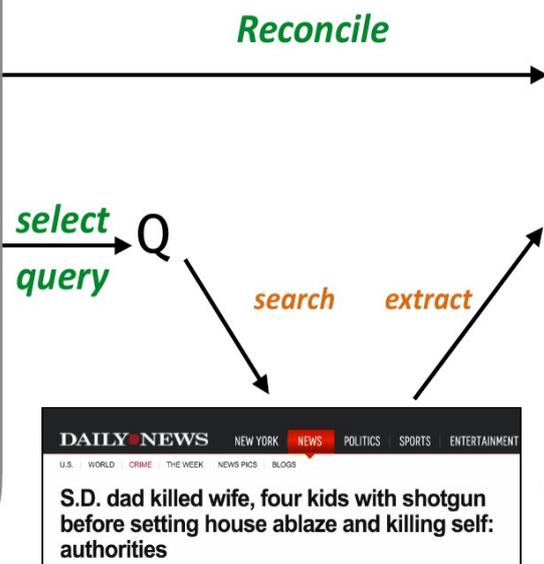
\_\_consultar nombre, institución y país  
\_\_encontrar el año

3:

\_\_consultar nombre, institución país y año  
\_\_encontrar otras instituciones en otros  
años



State 1



State 2

**Current Values:**  
*ShooterName* → Scott Westerhuis  
*NumKilled* → 4  
*NumWounded* → 2  
*City* → Platte

**New Values:**  
*ShooterName* → Scott Westerhuis  
*NumKilled* → 6  
*NumWounded* → 0  
*City* → Platte

**State:**  
 ⟨0.3, 0.2, 0.5, 0.1, ← *currentConf*  
 0.4, 0.6, 0.2, 0.4, ← *newConf*  
 1, 0, 0, 1, 0, 1, 1, 0, ← *matches*  
 0.2, 0.3, ..., 0.1, 0.5, ← *contextWords*  
 0.65⟩ ← *document tf-idf similarity*

Iván Vladimir Meza Ruiz:

1996-2001 : UNAM – Facultad de Ingeniería  
2003-2009 : MSc y PhD en Edimburgo  
2010-2017 : UNAM - IIMAS



I'm a researcher at [IIMAS](#), [UNAM](#) working in the [Department of Computer Science](#).

My main interest is the intersection between (human) language and computing, I focus on [Machine Learning](#), [Artificial Intelligence](#) and [Natural Language Processing](#).

Recently I have been working on the following topics:

- Deep learning for language, speech and sound (mainly bird songs)
- Topic mining and discovery
- Narrative generation for robots

# Statistical Semantic Processing using Markov Logic

*Ivan Vladimir Meza-Ruiz*

*University of Edinburgh*



Doctor of Philosophy

Búsqueda de información:

— ¿cómo llegar a las páginas que contienen los datos que buscamos?

Extracción de información:

—¿cómo extraer las fechas, los nombres propios, los topónimos y las instituciones que buscamos?

Ética informática:

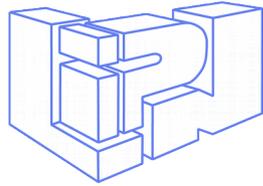
—¿cómo transparentar el manejo de información personal proveniente de Google en nuestra plataforma?

## Sociología de la migración:

—¿por qué retorna (o no) la emigración mexicana altamente calificada?

## Humanidades digitales:

— ¿cuál es el grado de correlación entre los datos extraídos de las fuentes tradicionales (Conacyt, institutos migratorios, OCDE) y los datos extraídos del web?



¡Gracias!

[garciaflores@lipn.univ-paris13.fr](mailto:garciaflores@lipn.univ-paris13.fr)