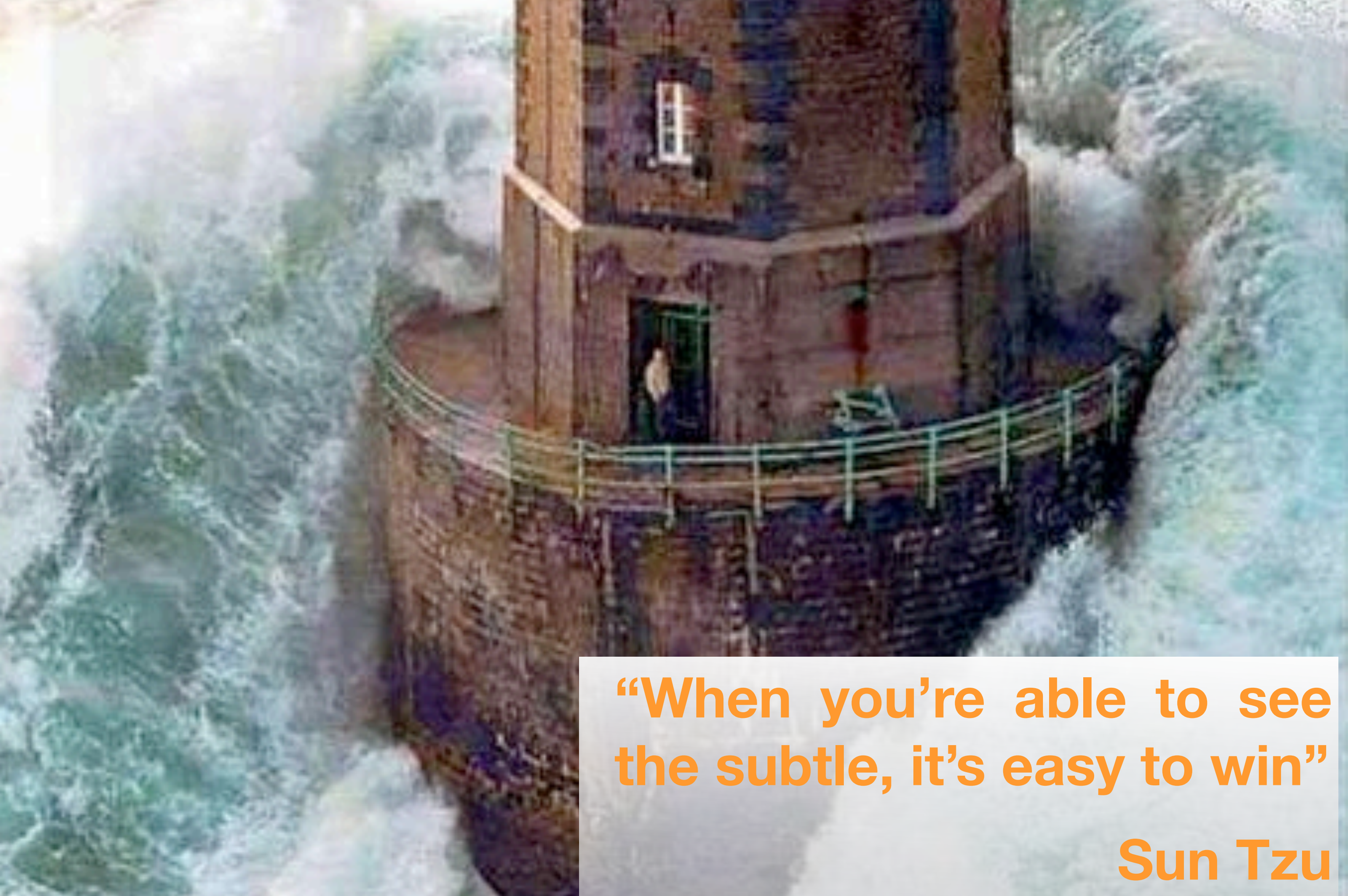
AUTORITAS

MUIInf
Máster Universitario en **Ingeniería Informática**

autoritas®
nuevas ideas, nuevas soluciones

Classification of Spanish by Regions
Universitat Politècnica de València, Apr 24th



“When you’re able to see the subtle, it’s easy to win”

Sun Tzu

Smart Listening

DATA -> INFORMATION -> KNOWLEDGE -> INTELLIGENCE

Tools and Methods

Human Interpretation

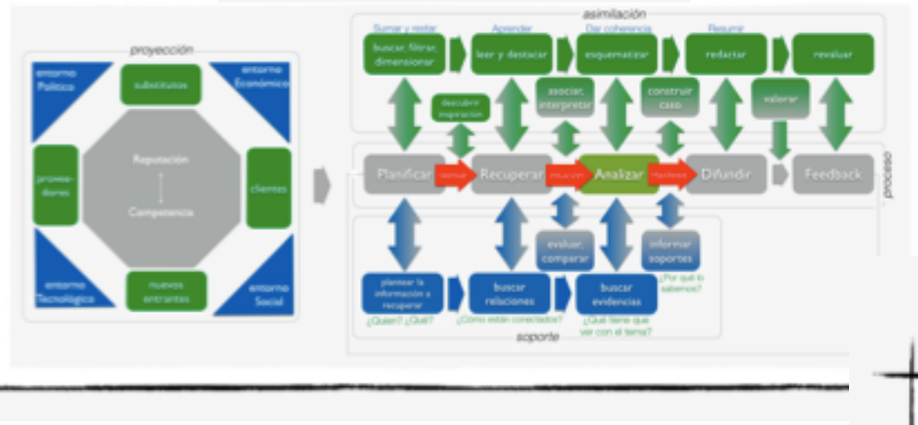
Strategic Application

http://hdwall.co/facebook-wallpaper-digg/

How does Autoritas do it?

Technology

Method



Team

The Smart Cycle

THINK, THINK, THINK...

...since thought precedes action

What do you expect from Internet / Social Networks for you ***Business Objectives****?*

Where is your ROI?

Police Investigation Bureau

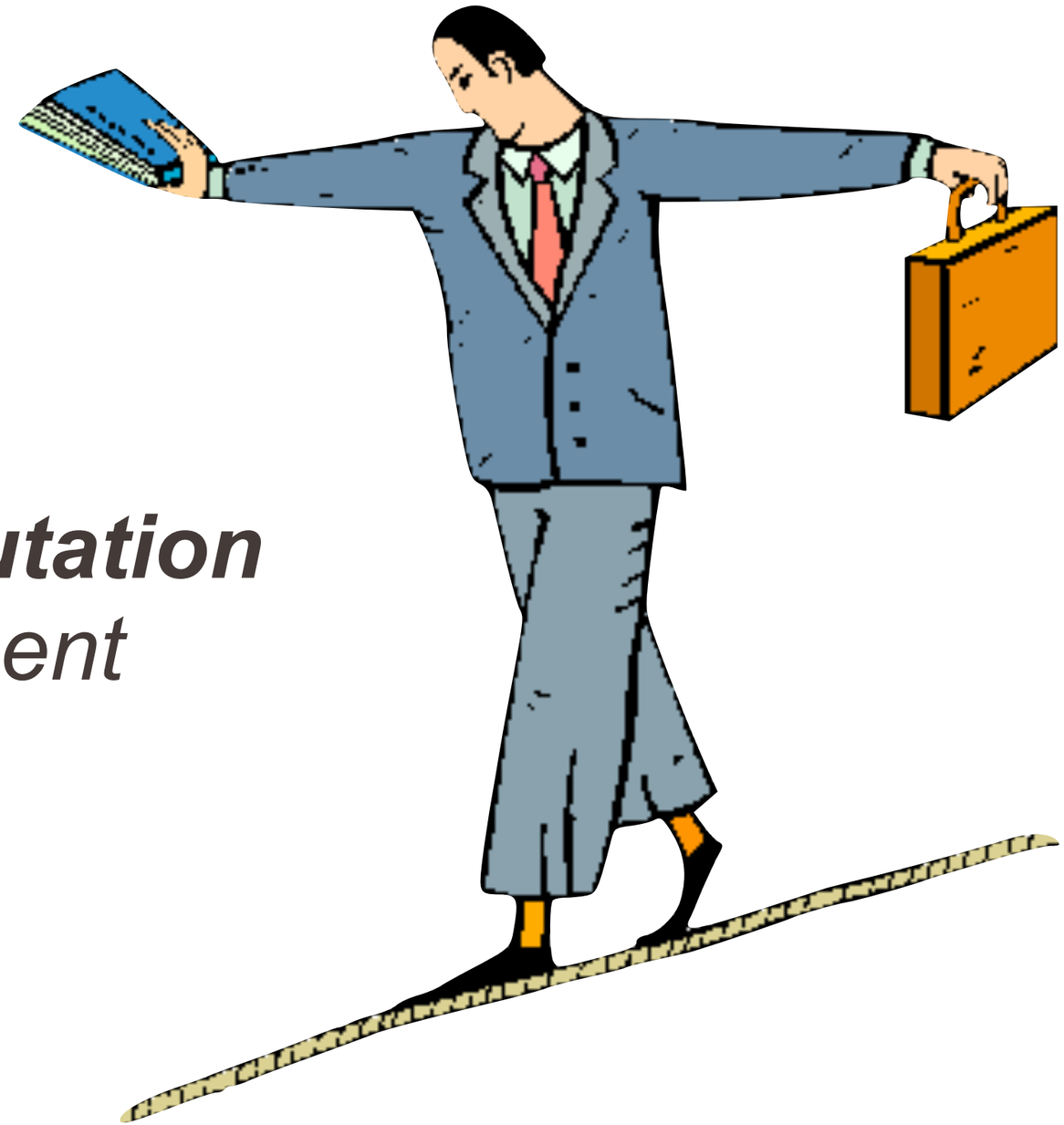
¡They have it very clear!

Catching the bad

Tourist Agency

Demand
prediction

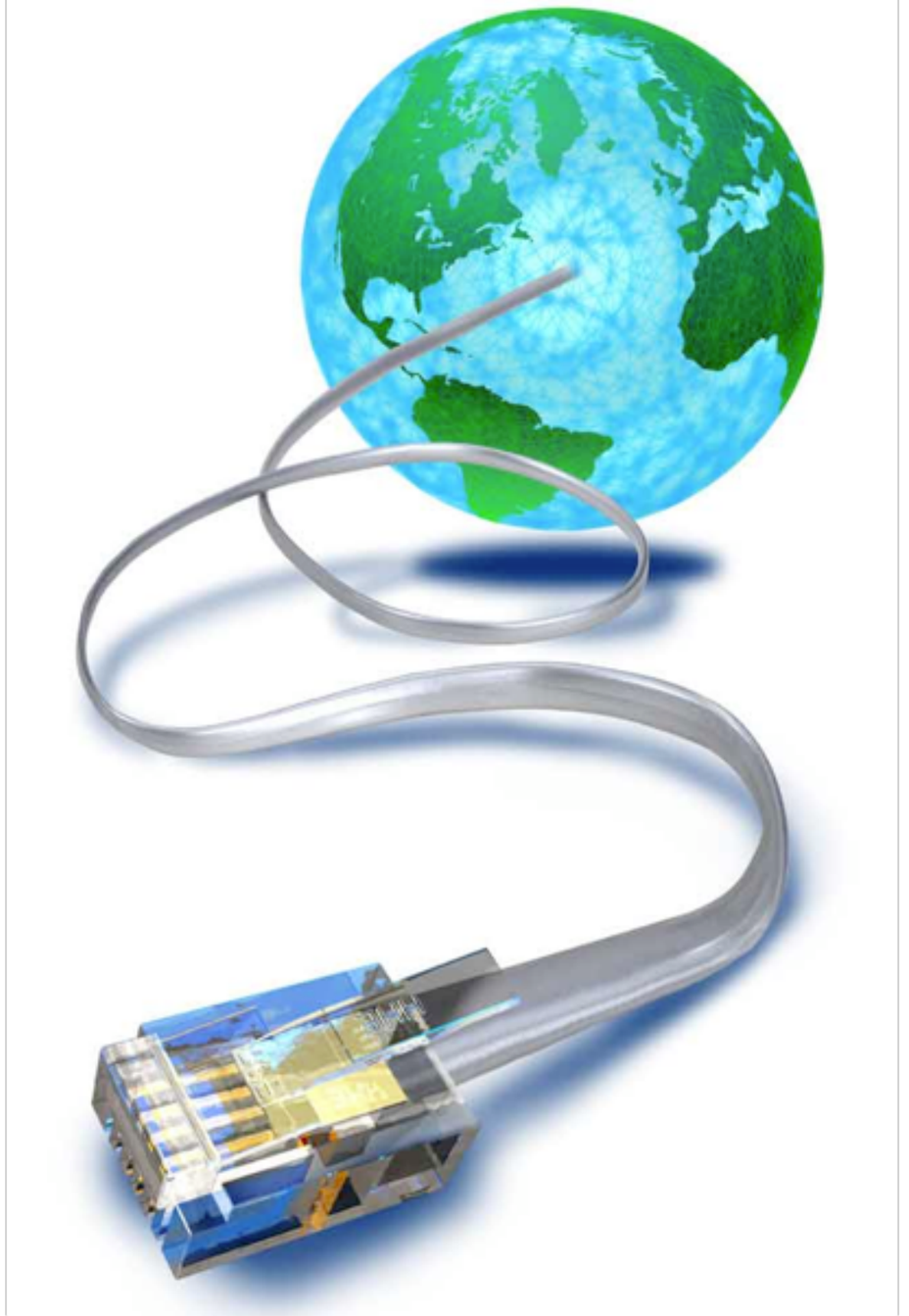
Insurance Companies



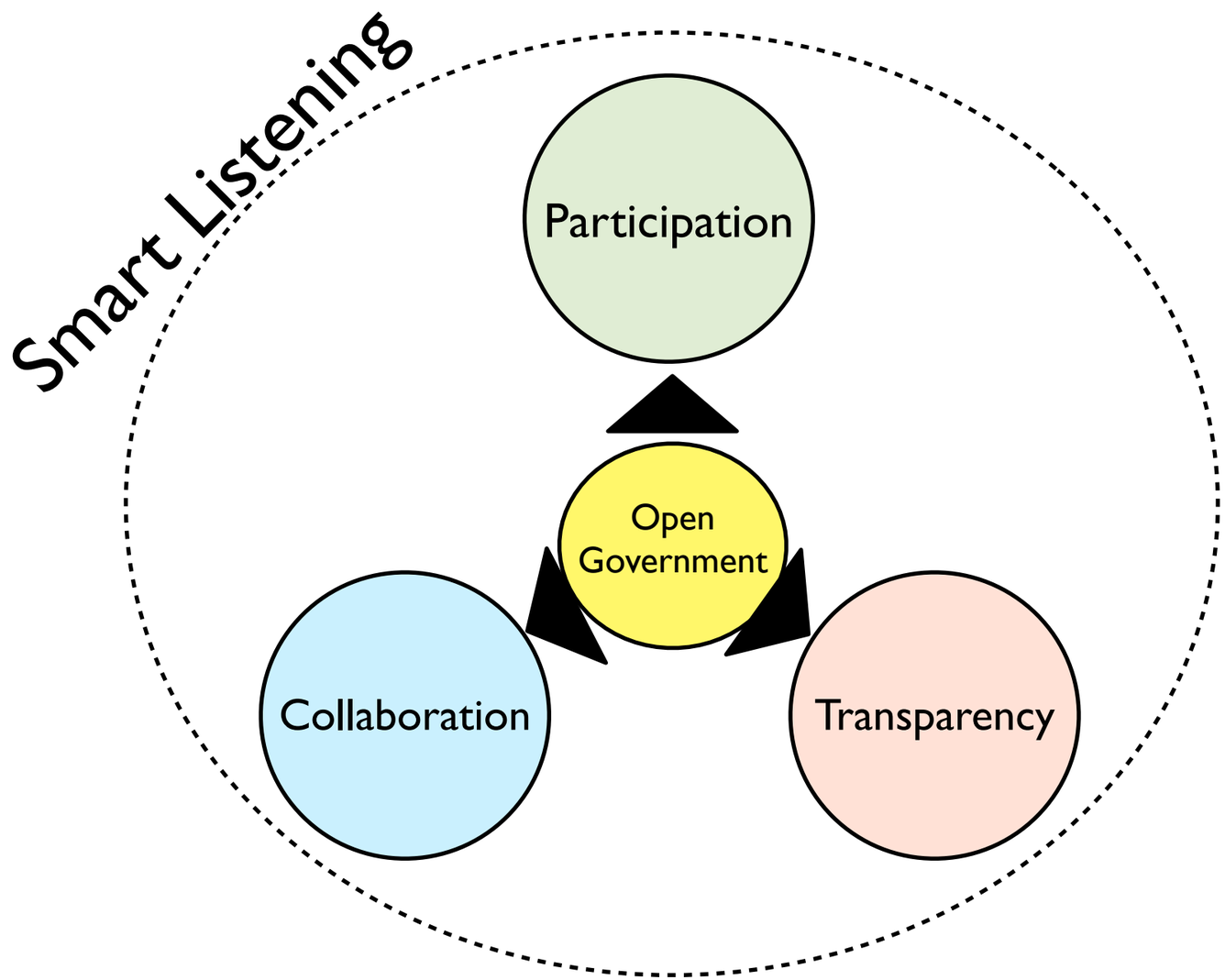
Measuring Reputation
*in **risk** management*

Marketing Agencies

Internet as
Information *source*



Governments



Telcos

To improve ***customer*** *service*

Mass media

Social content ***generation***

Audience *meassurement*

And yours?

Think, think, think...

“Truth is out there”
X-Files

$$\lim_{x \to \infty}$$

OBJECTIVE:

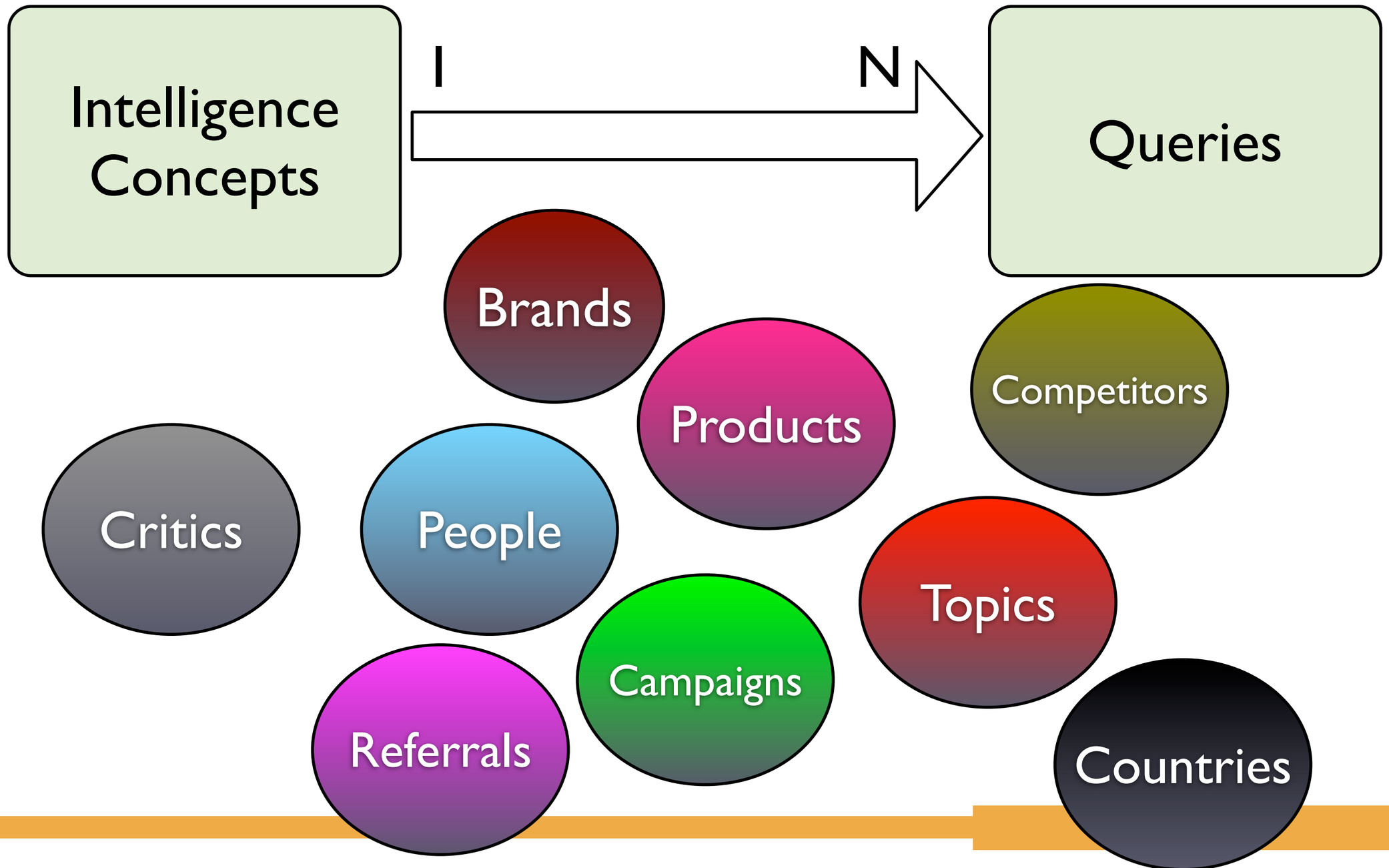
To retrieve all that we should retrieve and/but without retrieving nothing that we should not retrieve

$$\lim_{x \to 0}$$

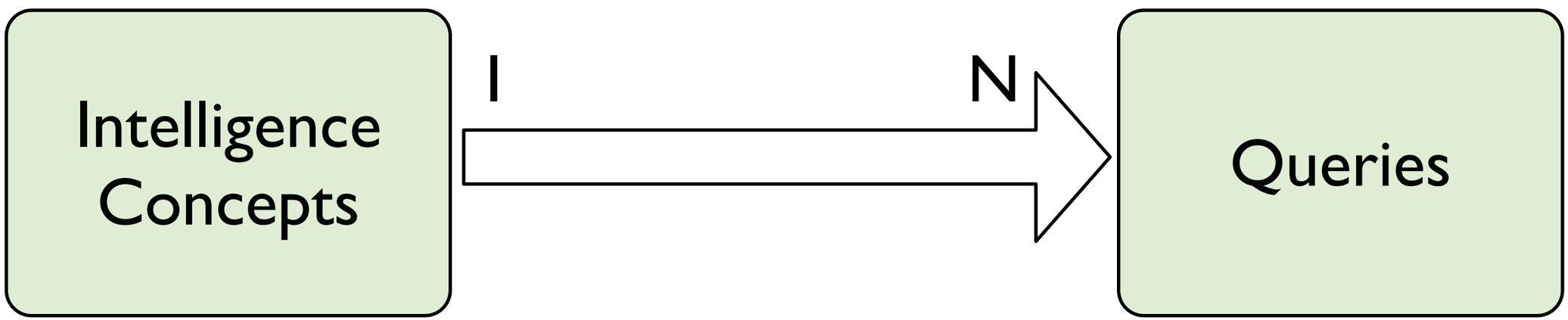
Information Sources (channels)

$$\lim_{x \to \infty}$$

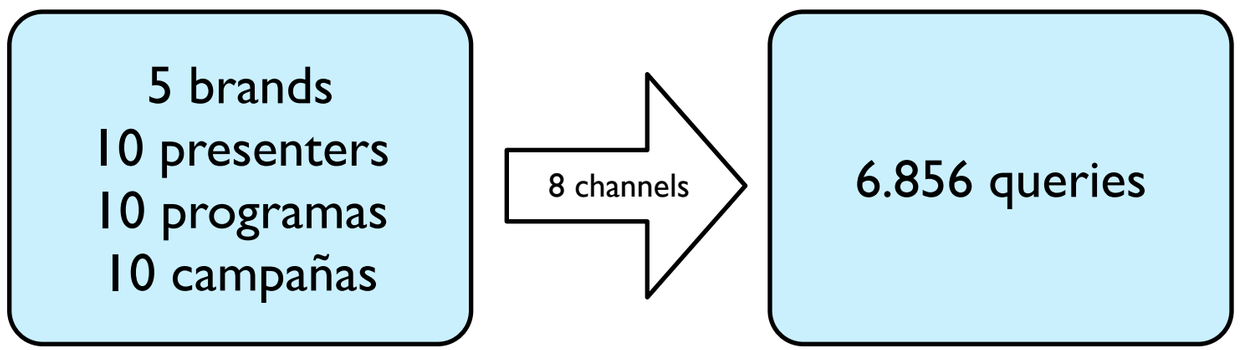
Concept-query mapping



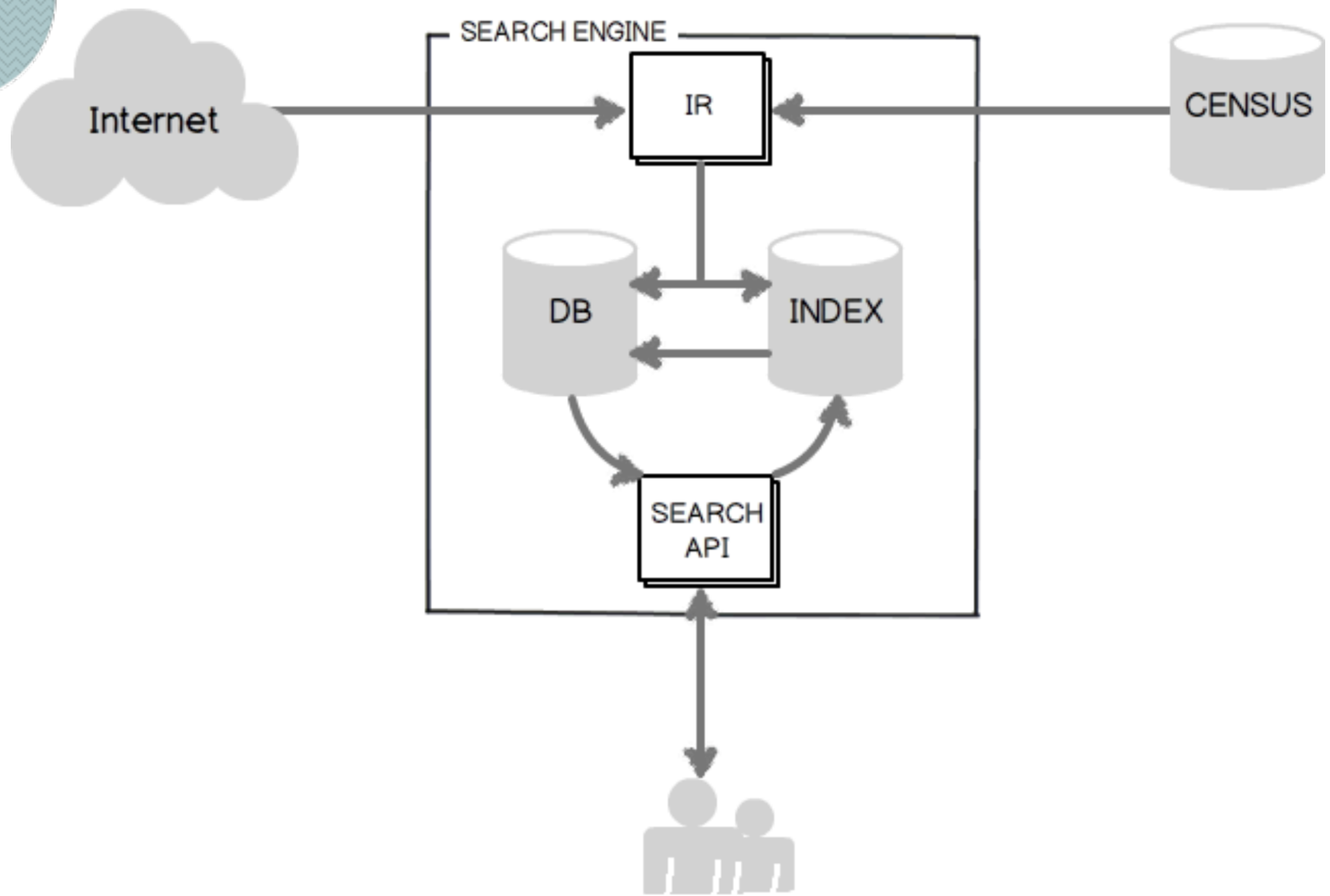
Concept-query mapping



A real example: Spanish public television



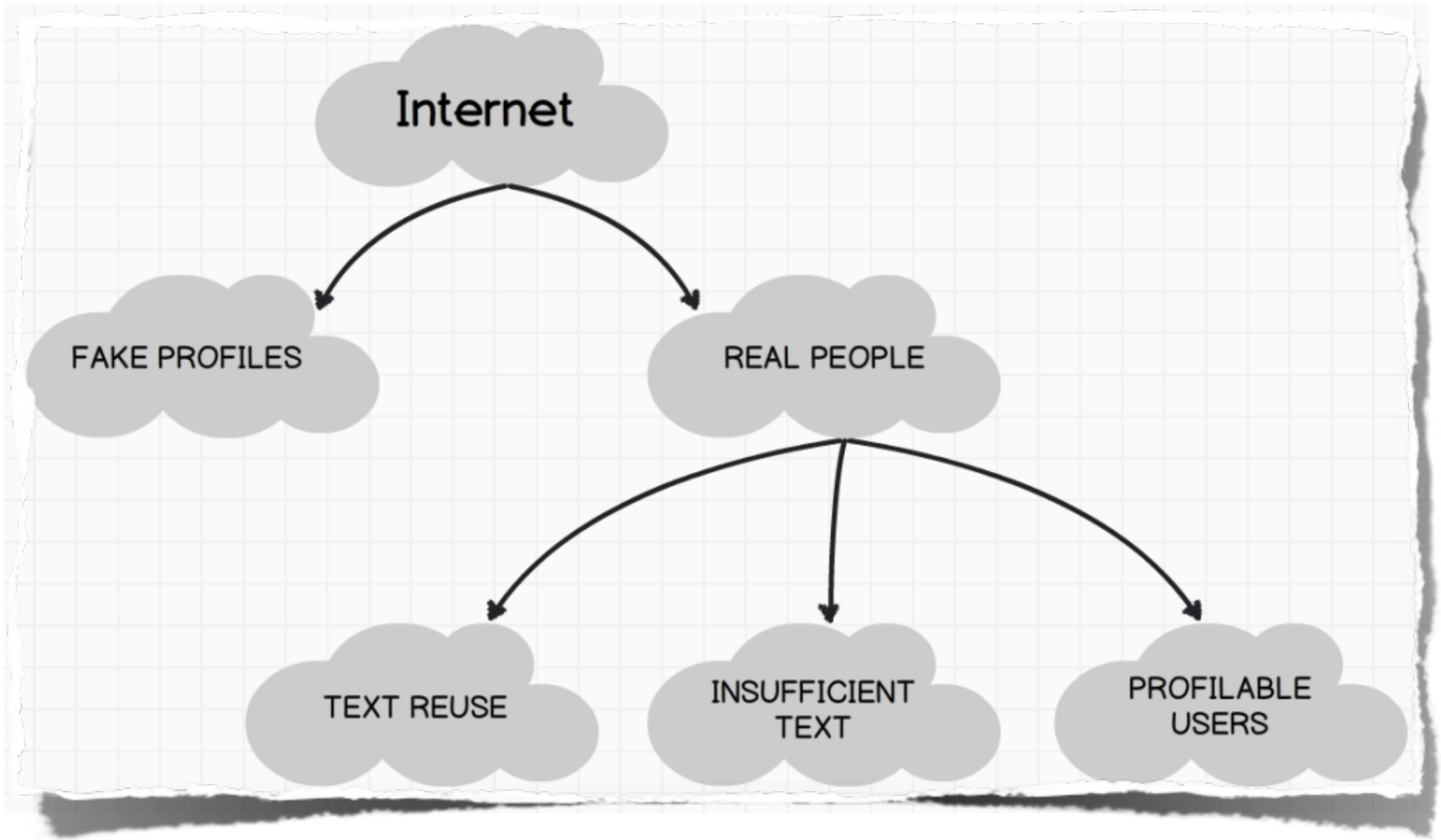
API vs. Crawling



information = data - noise

$$\lim_{x \to 0}$$

Why is there noise?



Users lie, plagiarize or say foolishness...

For example...

findability $\neq$ ***relevance***

Usable content retrieval

TRAGEDIA EN BUENOS AIRES

Casi medio centenar de muertos en un accidente ferroviario en Argentina

- El convoy, que transporta diariamente a más de [illegible] personas, no frenó al entrar en una de las principales estaciones de la capital argentina
- Vídeo: Momento en que el tren choca al llegar a la estación [illegible]
- Sigue en directo la transmisión de A24 sobre la tragedia

FRANCISCO PEREGIL | Buenos Aires | 22 FEB 2012 - 21:41 CET 255

Archivado en: Buenos Aires Latinoamérica Argentina Accidentes ferrocarril Sudamérica América Accidentes Sucesos

Rescatistas trasladan a un herido. / JULIO SANDERS (REUTERS)

Recomendar 376 Twittear 264 Enviar Compartir Enviar Imprimir

El tren de cercanías Sarmiento iba con casi todos los viajeros apiñados de pie, como siempre en hora punta. Salió a las [illegible] y media de la estación de Moreno para recorrer 14 estaciones hasta Buenos Aires. Sobre la estación cuarta, en la de Castelar, cambió de conductor. El nuevo maquinista, de 28 años, iba a emprender su primer trayecto de la mañana. Y el tren siguió frenando y arrancando en cada una de las paradas. Parecía un viaje normal, tal vez un poco más incómodo que otros para sus más de 1.200 viajeros, porque era la primera jornada laborable tras un largo puente de carnaval. A mil metros de su destino redujo la velocidad de 47 a 39 kilómetros por hora. En el andén entró a 26 kilómetros por hora, según el ministro de Transporte de Argentina, Juan Pablo Schiavi. Eran las velocidades normales de entrada en la estación. A 40 metros del final ya había frenado hasta los 20 por hora. Pero ya no volvió a frenar más. De pronto, el tren impactó contra el muro de contención y el segundo vagón se incrustó más de cinco metros en el primero. Eran las 8.32 (las 12.32 hora peninsular española). Murieron al menos 49 personas y 600 resultaron heridas. Uno de los que resultaron con vida fue el propio maquinista de 28 años, quien anoche se encontraba en una unidad de cuidados intensivos. "No sabemos qué ocurrió en los últimos 40 metros", reconoció el ministro.

Advertisements without relevance for the contents

Usable contents

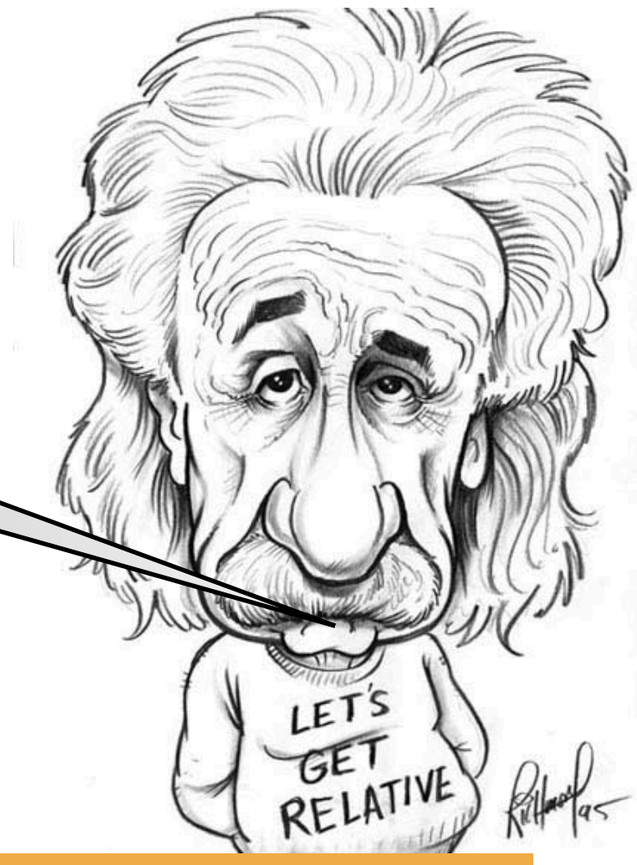
Latest news section that distorts the semantics of the page

$$\lim_{x \to 0}$$

The importance (and difficulty) of obtaining the right date

If the url includes the date, it's easy to know it

That's relative. Is this url from july or january: http://xxx/07/01/2010/crawler-403-forbidden.html



Language filtering

English

estoy sin internet ¬¨¬¨ fuuuuck!!!

Finnish

... euskocaja, como euskolabel, euskotren, euskomueble... XDDD

German

Vierrrrrrrrrrrrrnes, egunon!!

Portuguese

Flowah Powah!

Language Models vs. n-Gramms vs. Machine Learning

Geography filtering

Koldo M Martin
@iTitanMiller Mi casa, bilbao!
Ufff!!!

nerea miguel andrada
@Nereabskt bilbooo
i lovee baskett!!! BBB & ATHLETIC:)

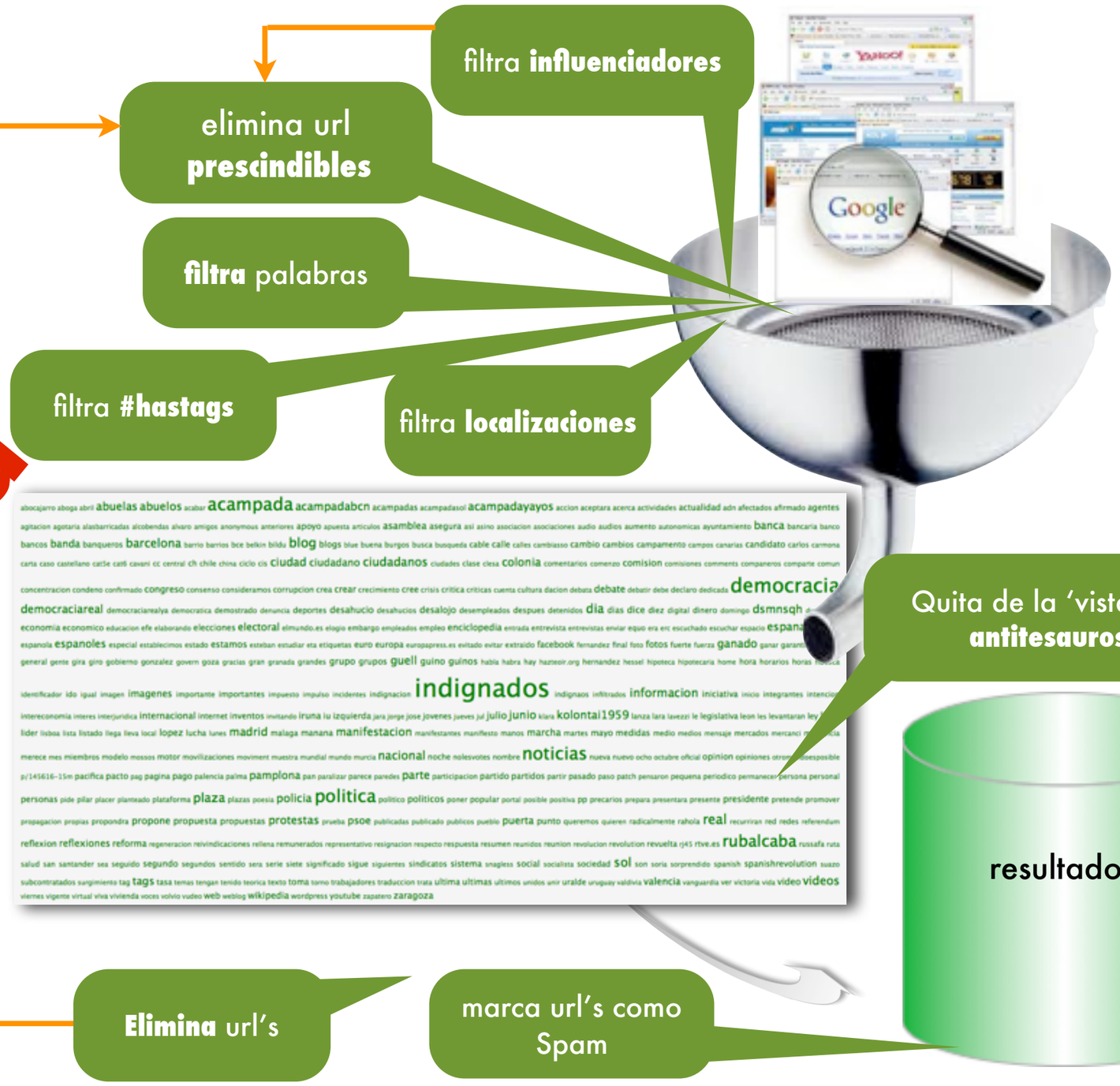
Jeremy Hagger
@mac_english M.A.R.S.
The Love Jones
http://twitter.com/lovejones

$\lim_{x \to 0}$

Fernanda
@FernandNavarro Narnia
Exactly who we are is just enough.

Source geography vs. contents geography vs. profile geography

“Limpia, fija y da explendor” Lema de la RAE

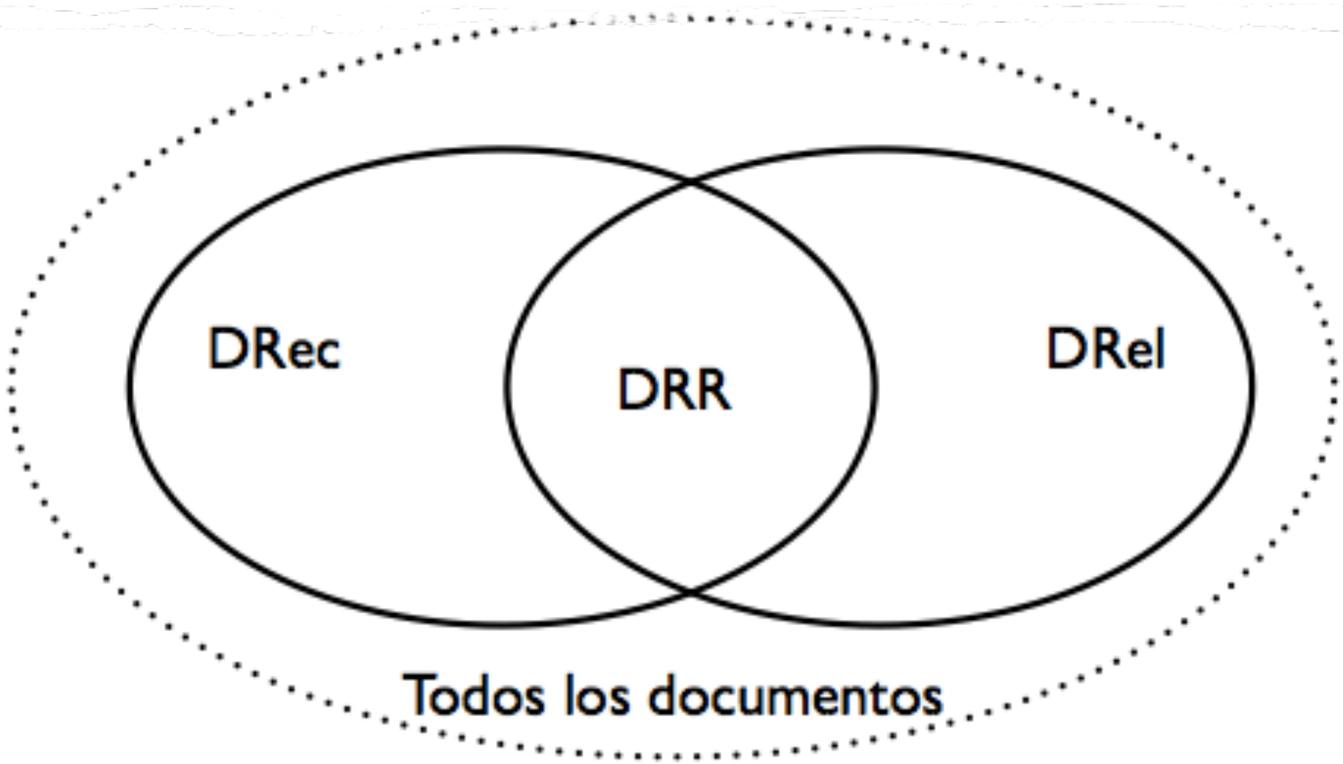


80% of the workload

I believe I solved the equation:

$$\lim_{x \to \infty} \quad \lim_{x \to 0}$$

Information Retrieval Evaluation...



DRel: Documentos Relevantes
DRec: Documentos Recuperados
DRR: Documentos Relevantes Recuperados

Precisión = DRR / DRec
Alcance = DRR / DRel

*- Do you want that all retrieved content is good? -> **Precision***

$$\lim_{x \to 0}$$

*- Don't you want to leave anything without retrieving? -> **Recall***

$$\lim_{x \to \infty}$$

...in Science

Information Retrieval Evaluation...

7.000 retrieved
54 wrong
99.23% precision

3.000 retrieved
50 not retrieved
98.36% recall

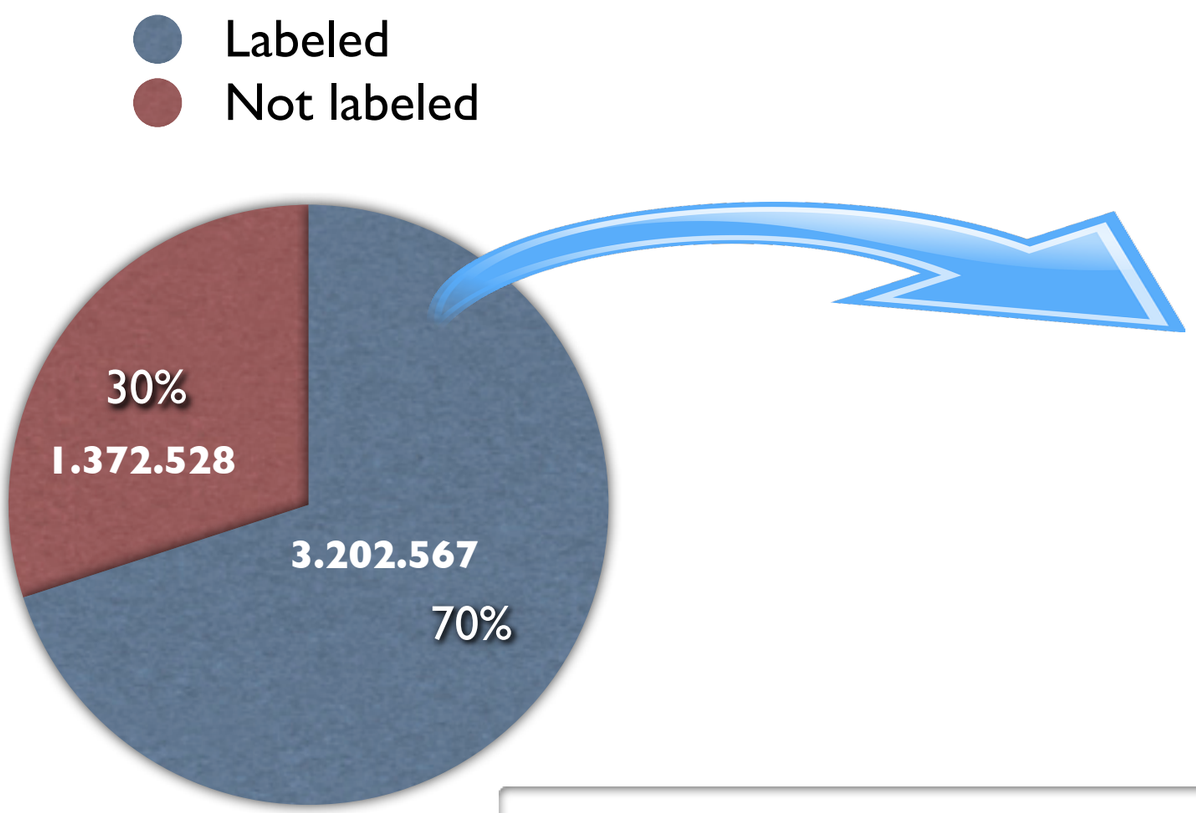
...in the Industry

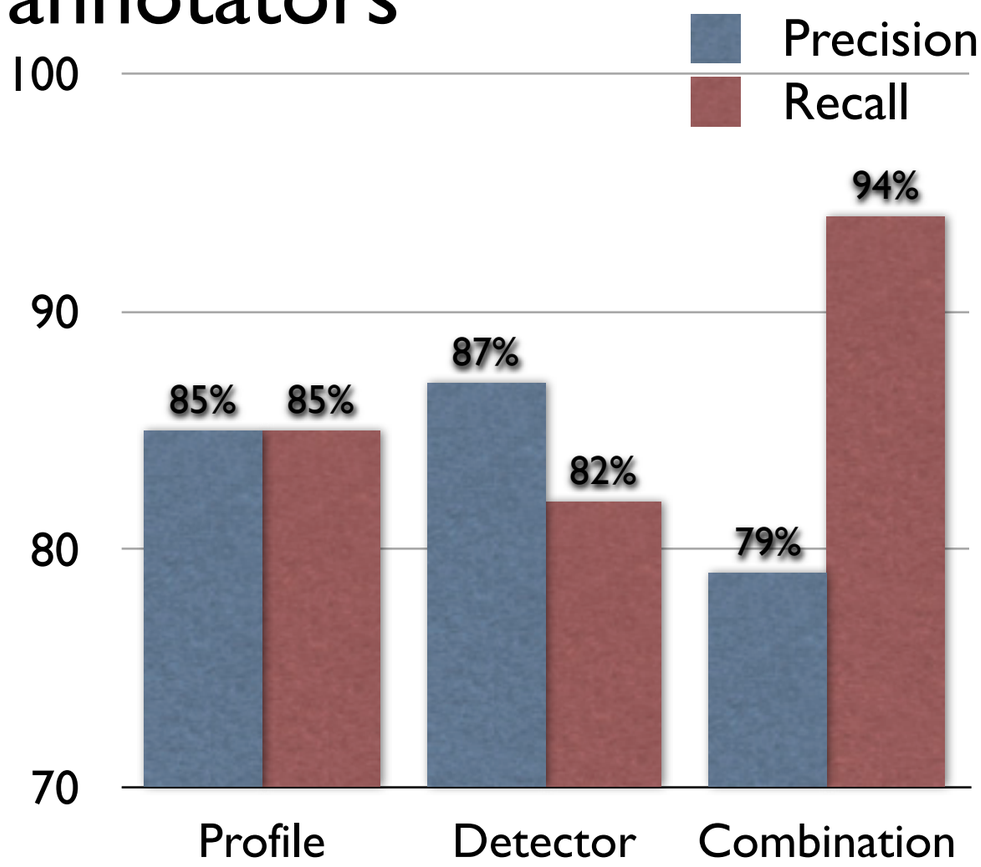
Example

Touristic project: 4.575.096 tweets

QUERIES: Mallorca, Menorca, Ibiza, Formentera, Baleares, ...

METHOD: 1.200 random tweets / 3 annotators





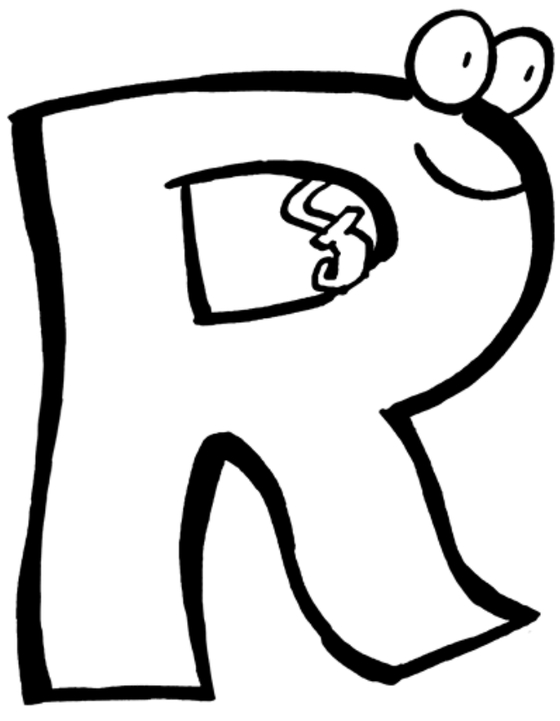
672.539 wrongly retrieved
204.419 good but not retrieved
1.372.528 that we don't know...

Bring order to the information...

...to analyse it and get value

...to ask questions...

...and to know which new questions to ask



WHAT are people talking ab, the consonant, the pro ing language or the ga telecommunic company?

Semantic ambiguity in millions of documents!!!

WHEN? -> Crisis management

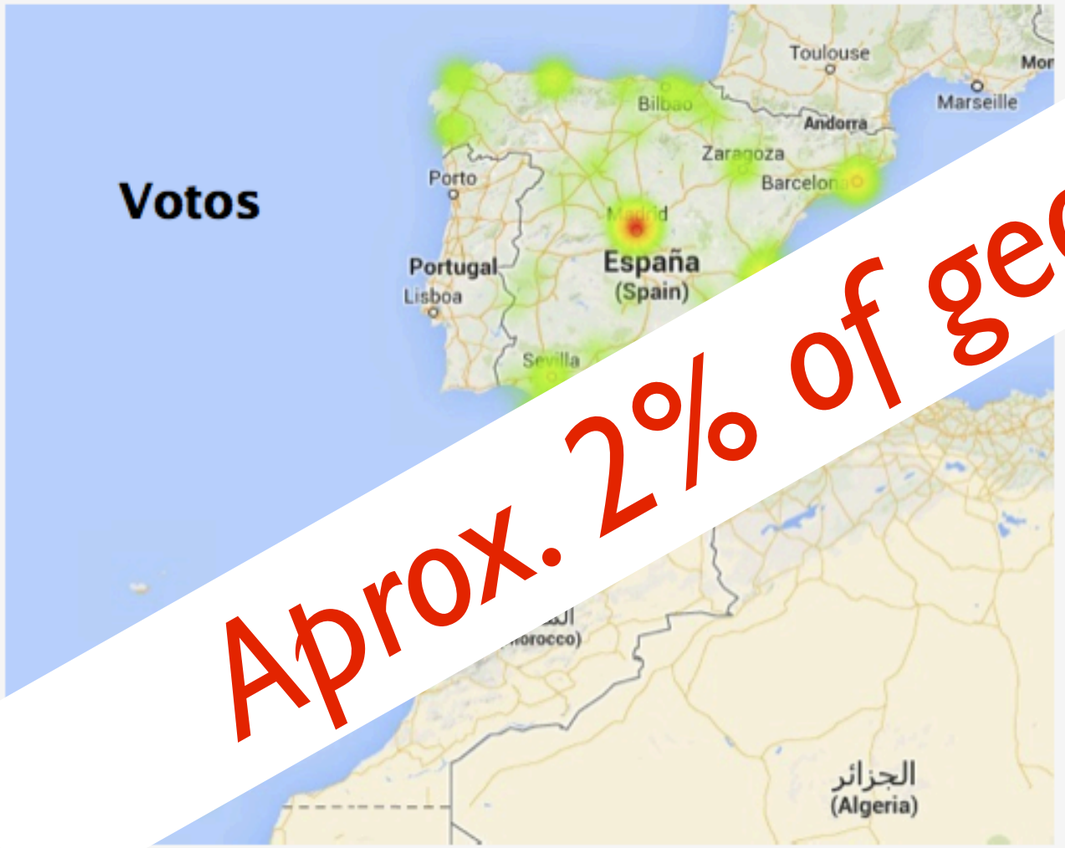
In real time!!

And let remember the issue of dates identification...

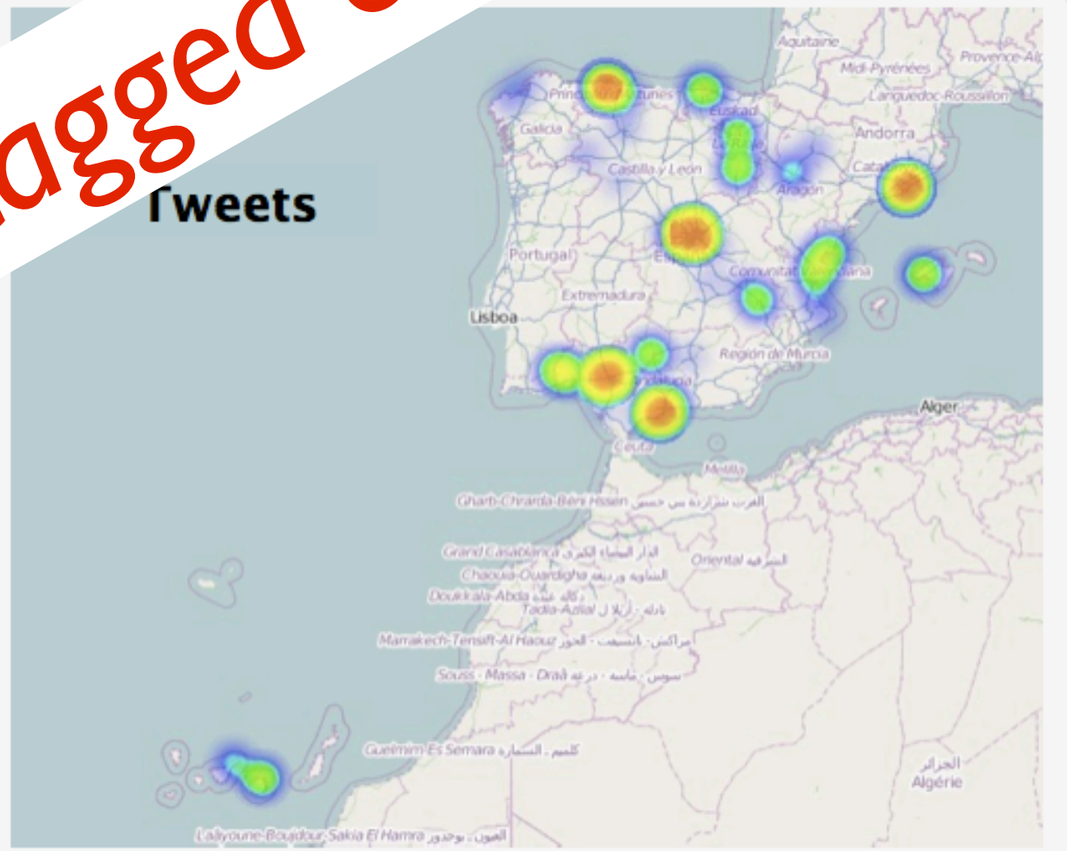
When are things happening?

WHERE/FROM WHERE is people talking ab t?

Distribución geográfica



Mapa de calor en función del reparto de votos por mesas electorales



Mapa de calor de los tweets emitidos, que han podido geolocalizarse

Aprox. 2% of geotagged contents!!

HOW? -> Not only sentiment analysis

Polarity is onlye one dimension:
- Emotions
- Motivations
- Values
- SWOT
All of them answer the question “how?”

An example: "The risk premium in Spain is 235"

Positive, *negative*, *neutral* o *none*?

An example: "The risk premium in Sp

Positive, negative, neutr

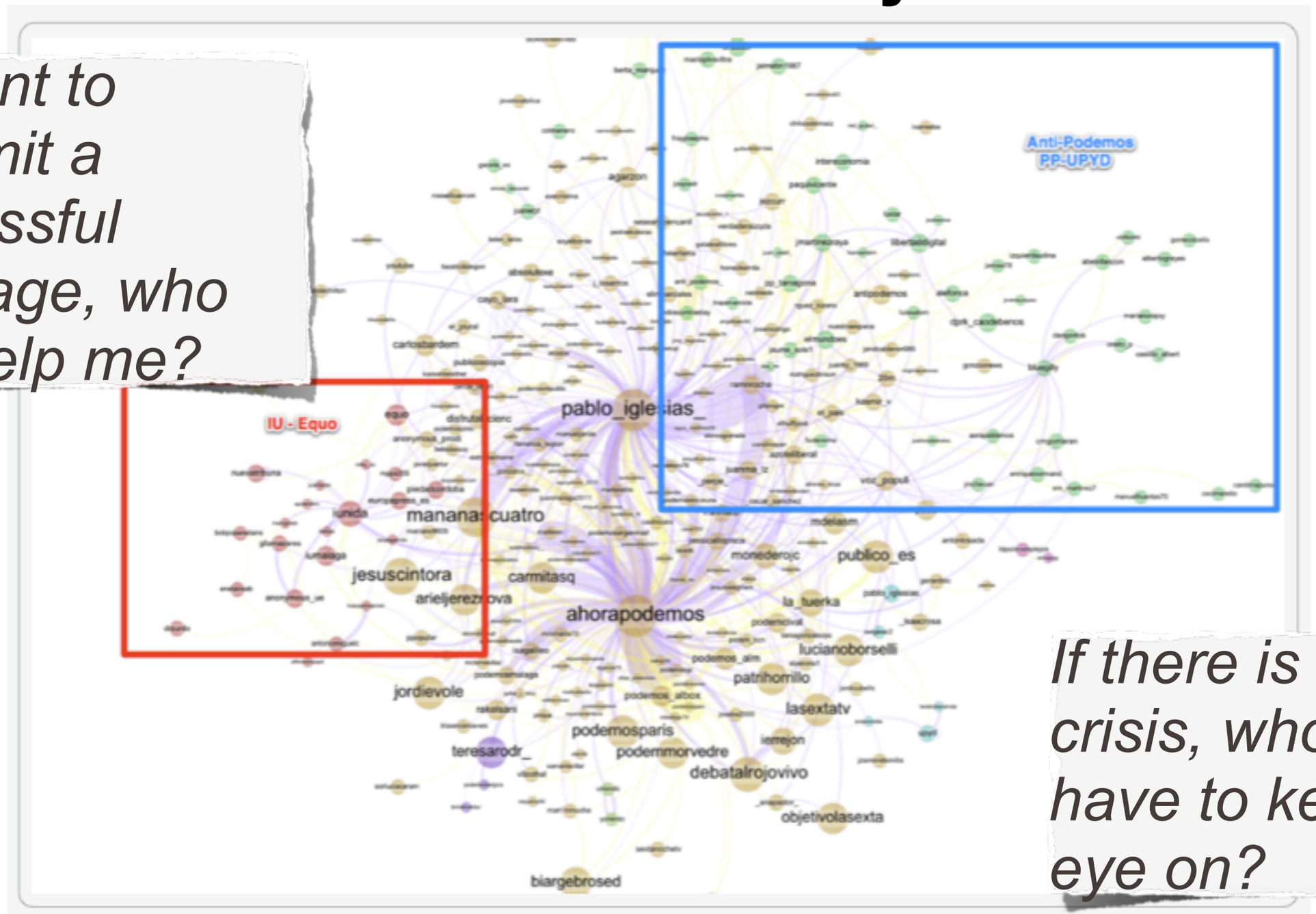
*My question: **For***

country?
e opposition?
. of the Spanish Bank?
eigner investor?
e national capitalist?
or whom has a morgage?

Subjectivity in transmitter, and in receiver!!

WHO? -> Social Network Analysis

If I want to transmit a successful message, who can help me?

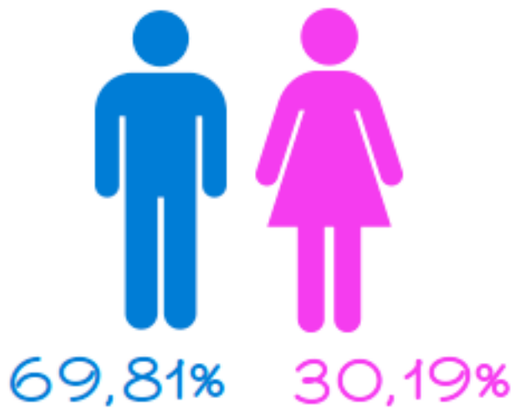


Usuarios de twitter más influyentes en la conversación sobre PODEMOS

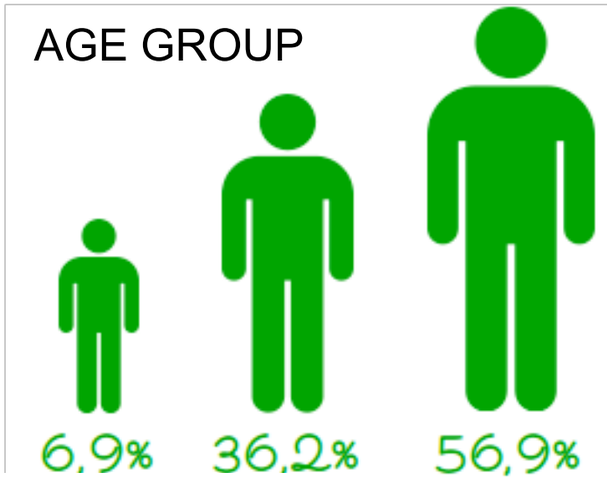
If there is a crisis, who do I have to keep an eye on?

WHY -> Author Profiling

GENDER

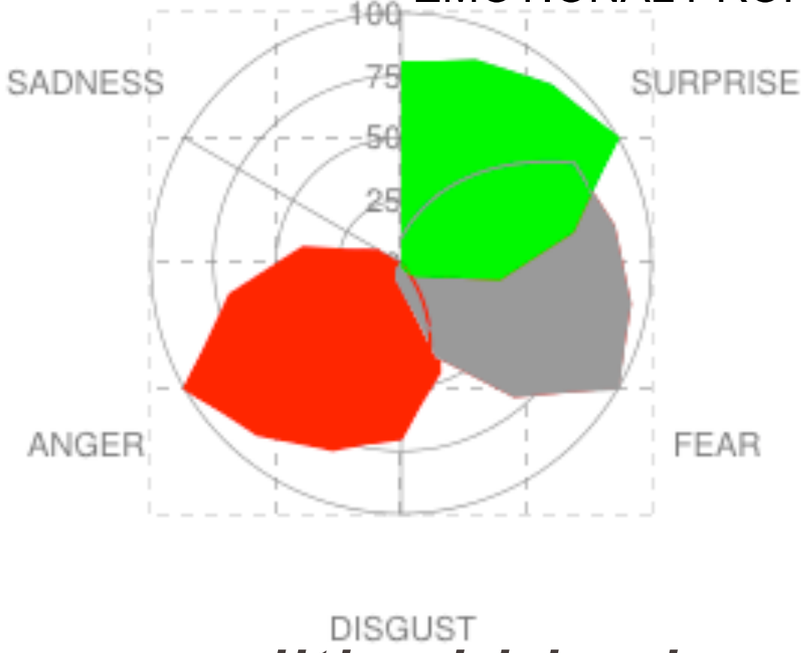


AGE GROUP



NATIVE LANGUAGE

EMOTIONAL PROFILE



PERSONALITY TRAITS

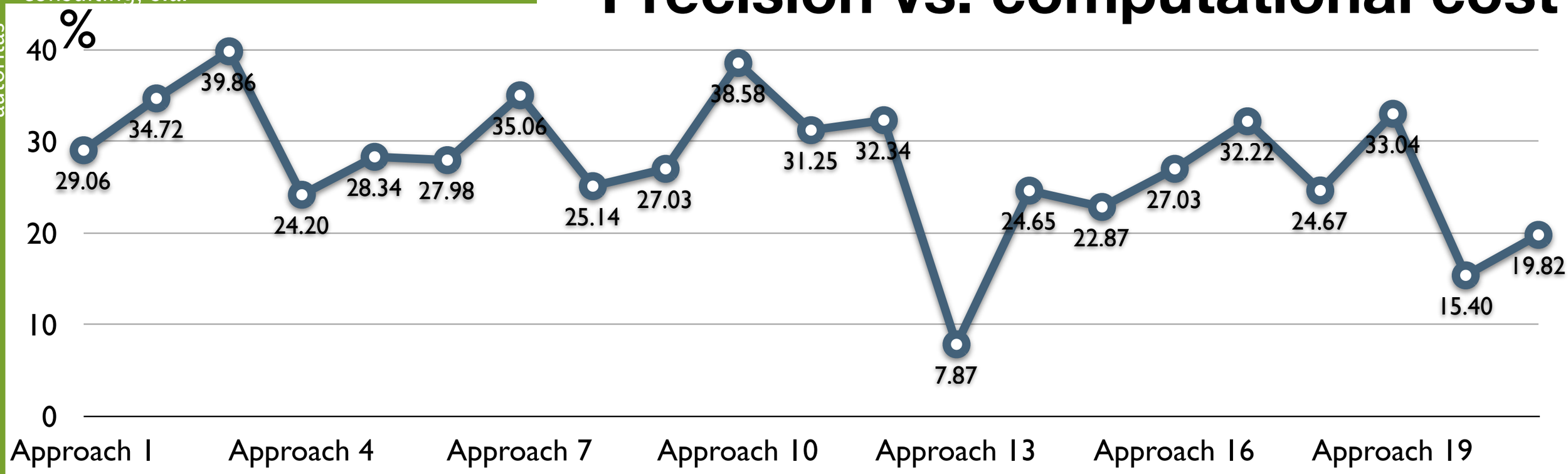
... political ideology, religious beliefs, and much more!

Big Data is both the solution and the problem...
...but mainly the opportunity

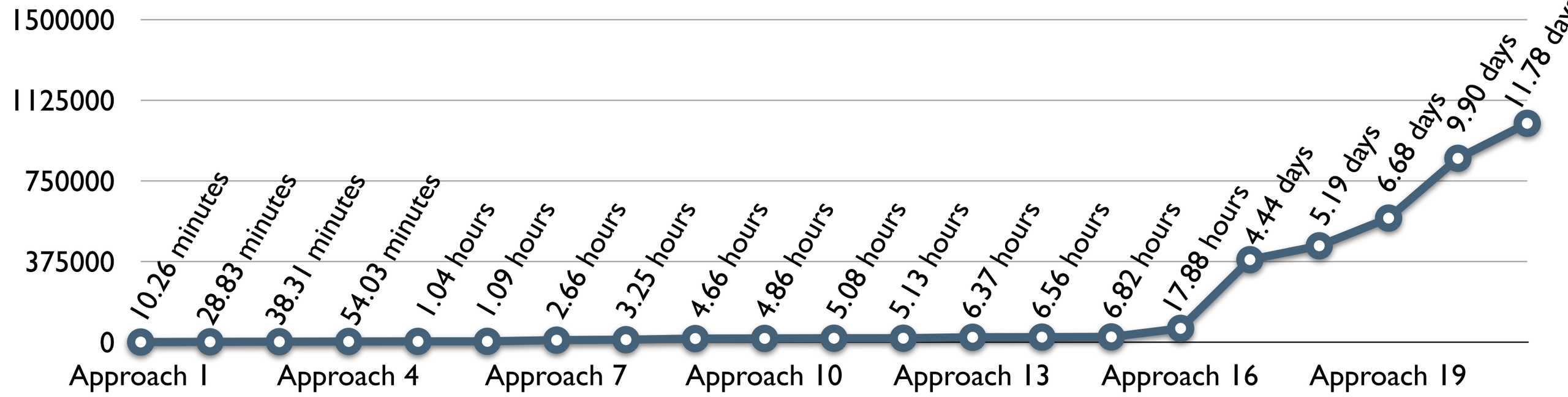
- Retrieval & Storing
- Evolution
- Words & Topics
- Tagging
- Hashtags
- People
- Locations
- Brands
- Sentiment & Emotion
- Users & Relationships
- Influence
- Gender & Age
- Language Variety
- ...

80~120 tw/sec =4.800~7.200 tw/min =288.000~432.000 tw/hour =6.912.000~10.368.000 tw/day

Precision vs. computational cost



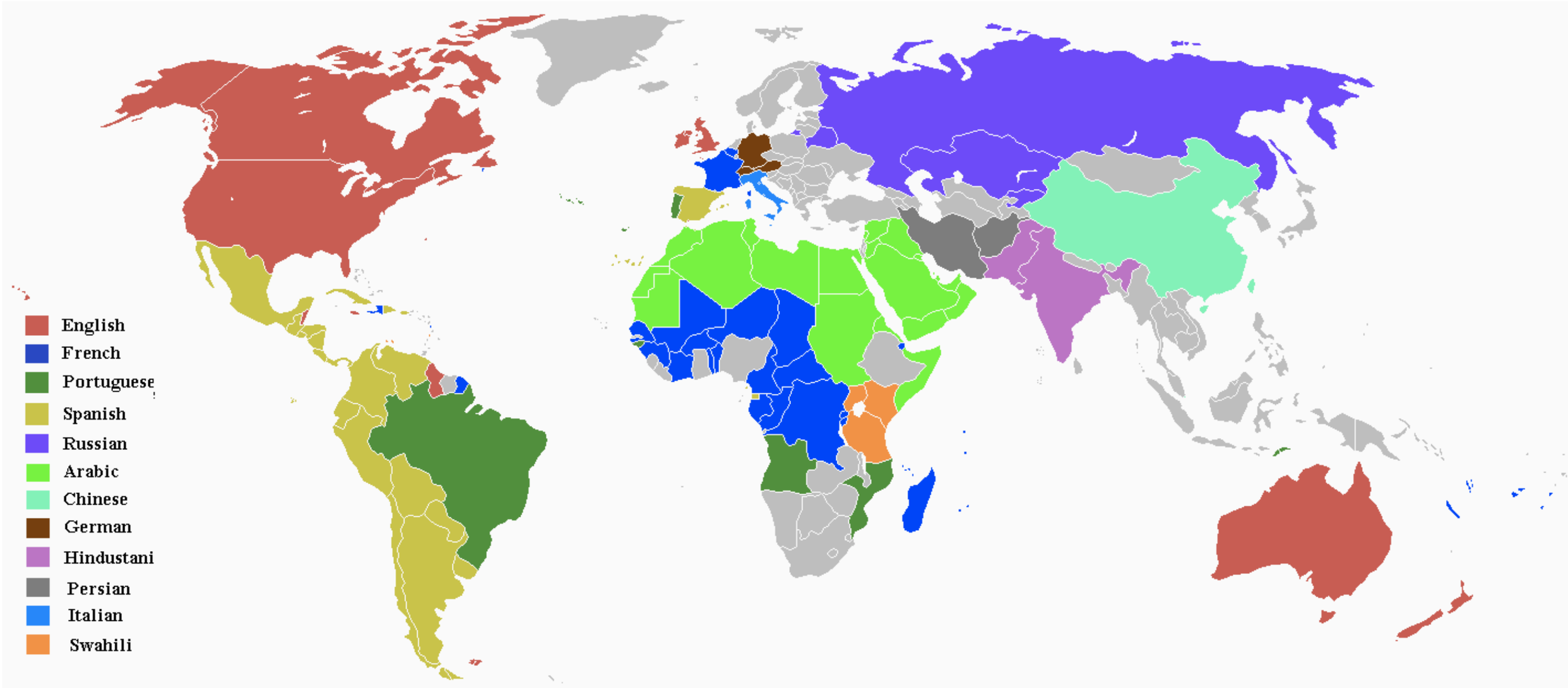
AGE AND GENDER IDENTIFICATION ~240K AUTHORS



Classification of Spanish by Regions

Autoritas Case Study

On the Internet there are no boundaries...



...except the language

Show me how you talk...

HispaBlogs

TRAINING	TEST
450	200
x 5 varieties	

https://github.com/autoritas/RD-Lab/tree/master/data/HispaBlogs

...and I tell you where you are

State-of-the-art

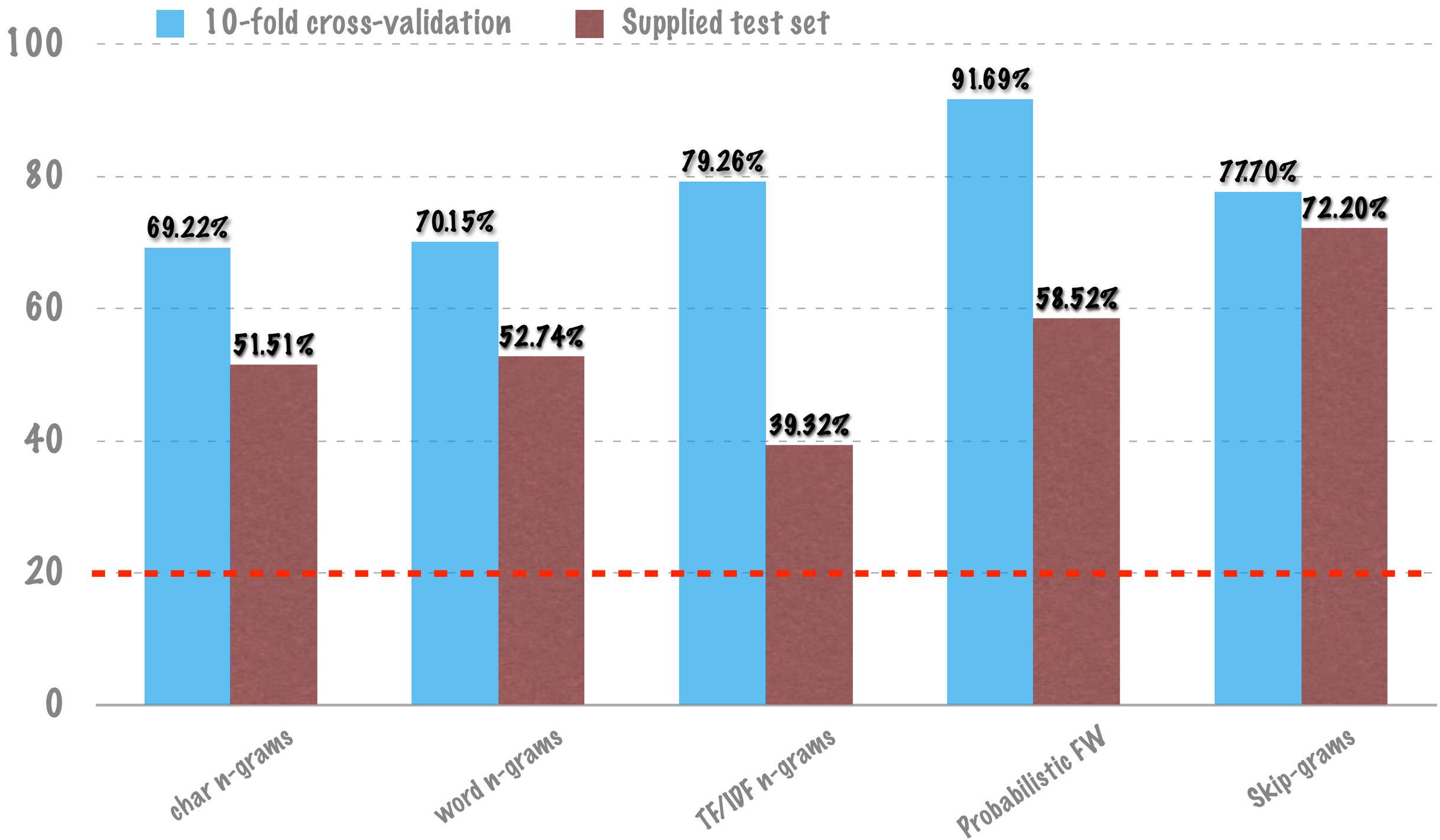
Automatic Identification of Language Varieties: The Case of Portuguese.
Zampieri, M., Gebrekidan-Gebre, B.
In Proceedings of the Conference on Natural Language Processing 2012

* Corpus (1000 documents from newsletters): 2 regional variations
* Features: word and character n-grams
* ML Algorithm: Language probability distributions with log-likelihood function for probability estimation
* Evaluation method: 50/50 split
* Accuracy:
  * Word uni-grams: 99.6%
  * Word bi-grams: 91.2%
  * Character 4-grams: 99.8%

Automatic Identification of Arabic Language Varieties and Dialects in Social Media.
Sadat, F., Kazemi, F., Farzindar, A.
In Proceeding of the 1st. International Workshop on Social Media Retrieval and Analysis SoMeRa 2014

* Corpus (blogs and forum documents): 6 regional variations
* Features: character n-grams
* ML Algorithm: Markov language model vs. Naïve Bayes
* Evaluation method: 50/50 split
* Accuracy: 98% (78% F-measure)

Our approach: compared results



Your turn...

- *Course work*

- *Master project*

- *Scholarship, job positions, collaborations*

- *...*

@kicorangel
francisco.rangel@autoritas.es

tocamos**internet**

autoritas®
nuevas ideas, nuevas soluciones

ARGENTINA - BRASIL - CHILE - ESPAÑA - MÉXICO - PANAMÁ - UK