EURECOM
Sophia Antipolis

Enrichir des vidéos d'actualités par la création d'instantanés sémantiques et contextualisés

Raphael Troncy <raphael.troncy@eurecom.fr>

Multimedia Semantics, EURECOM

@rtroncy

@peputo

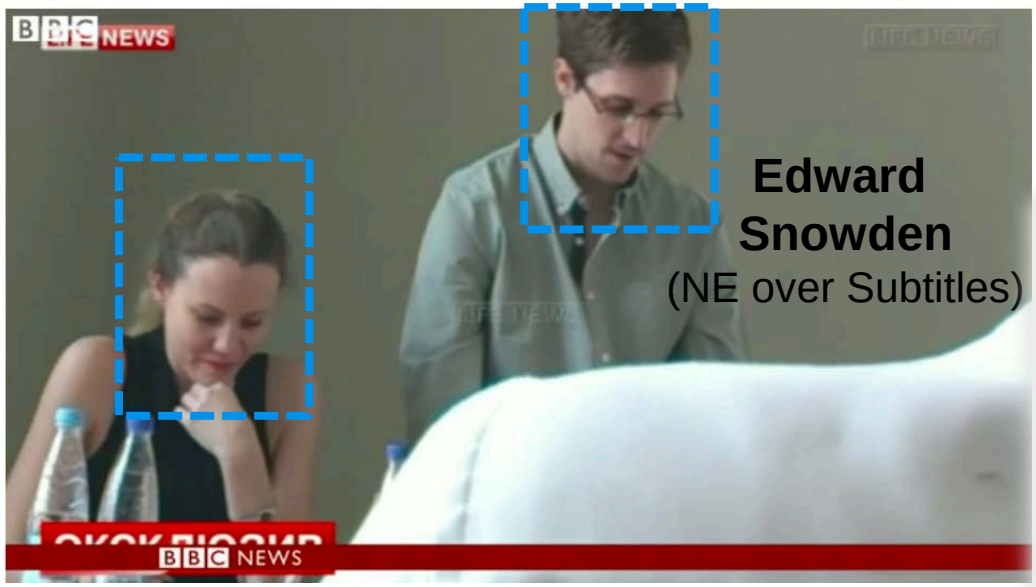
The Use Case: Contextualizing News

BBC | Sign in | News | Sport | Weather | Shop | Earth | Travel | More | Search

NEWS

Home | Video | World | UK | Business | Tech | Science | Magazine | Entertainment & Arts | Health | In Pictures | More

World | Africa | Asia | Australia | Europe | Latin America | Middle East | US & Canada



Edward Snowden
(NE over Subtitles)

Fugitive Edward Snowden applies for asylum in Russia

Sarah Harrison

WikiLeaks Editor

Sheremetyevo

Международный аэропорт Шереметьево

Airport in Moscow

http://www.bbc.com/news/world-europe-23339199#t=34.1,39.8
(Media Fragment URI 1.0)

EURECOM

The News Semantic Snapshot (NSS)

What is on top:

Entities explicitly appearing in the documents

Edward Snowden

Anatoly Kucherena

Going deep down…

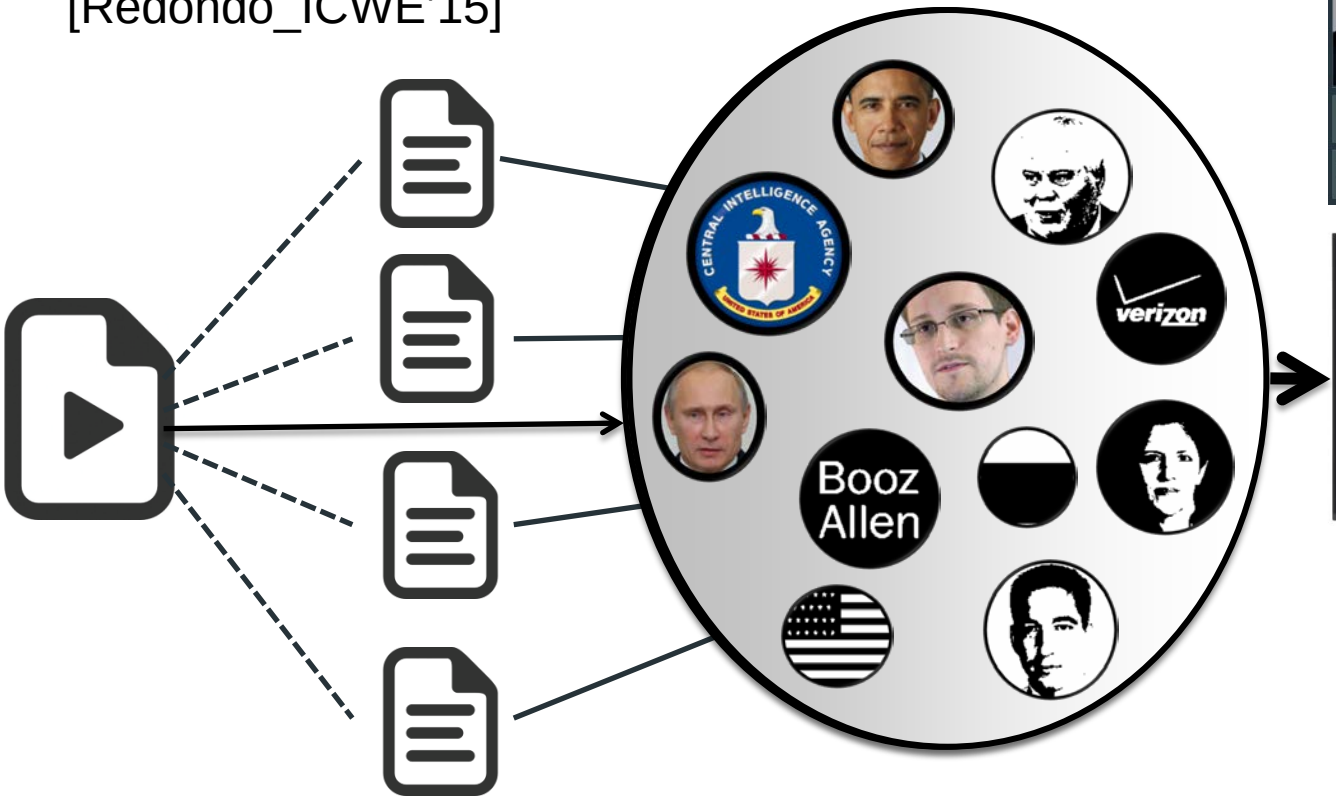
It is always challenging

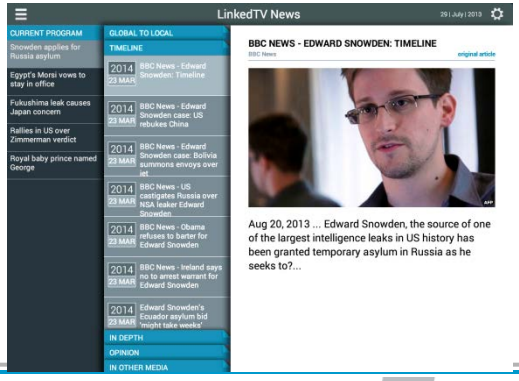
Laura Poitras

NSS for Feeding Second Screen Applications

News Semantic Snapshot
(NSS)

[Redondo_ICWE'15]





EURECOM

The News Semantic Snapshot: Gold Standard

- High Level of detail, significant human Intervention:
  (Experts in the news domain + users)
- Entities in 5 Dimensions: (Visual & Text)

(4) Suggestions of an expert

(2) Image in the video

(3) Text in the video image

(1) Video Subtitles

(5) Related articles

The New York Times
ALJAZEERA
theguardian

"We don't have any extradition treaty with Russia. (1) Broadly speaking our policy remains the same: that we'd like him returned

[Romero_TVX'14]

USER SURVEY

The News Semantic Snapshot: Gold Standard

Newscast Title	Person	Organisation	Location	Total
Fugitive Edward Snowden applies for asylum in Russia	11	7	10	28
Egypt's Morsi Vows to Stay in Power	4	5	4	17
Fukushima leak causes Japan concern	7	5	5	13
Rallies in US after Zimmerman Verdict	9	2	8	19
Royal Baby Prince Named George	15	1	6	22
Total	46	20	33	99

Table 1: Breakdown entity figures per type and per newscast.

25

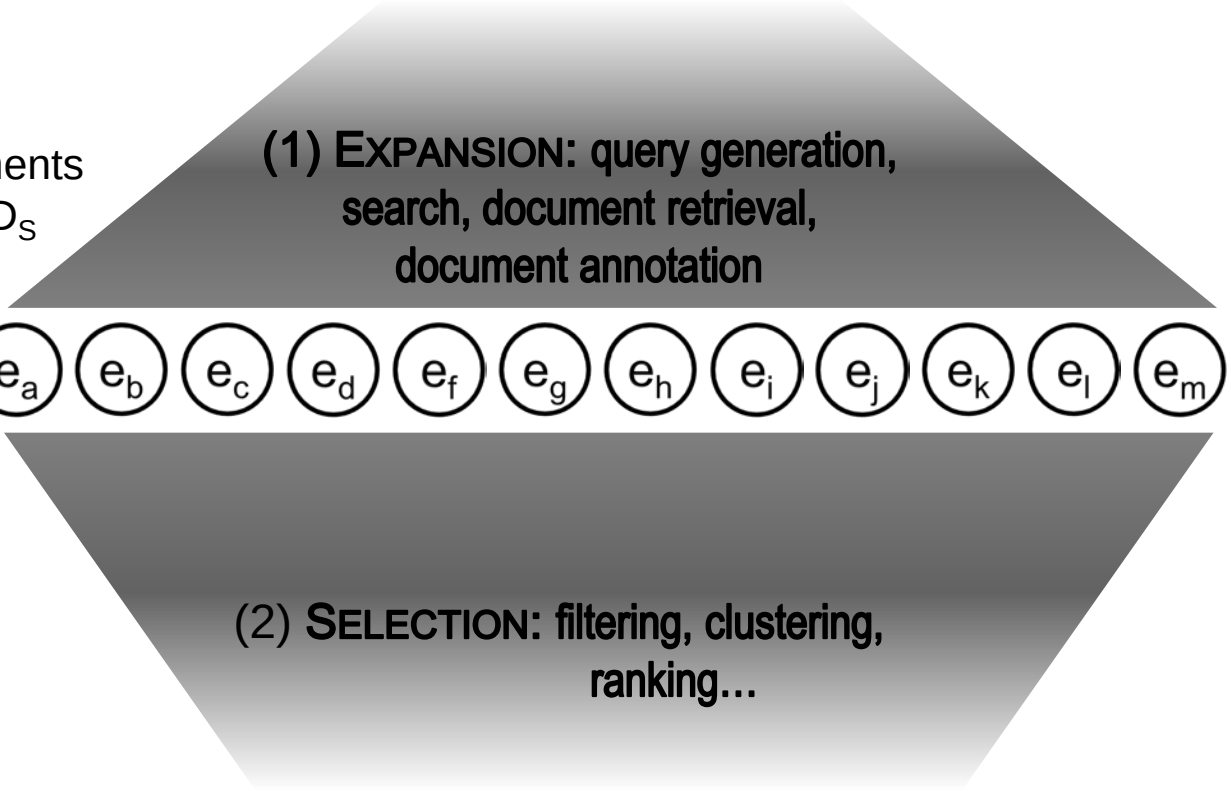
Play with the data and help us to extend it at:
https://github.com/jluisred/NewsConceptExpansion/wiki/Golden-Standard-Creation

Generating the NSS: General Method

a) Entities from Seed Document D_S e_a e_b e_c e_d

[Redondo_SNOW'14]

WEB:
Other documents similar to D_S



b) **Expanded Entities** e_a e_b e_c e_d e_f e_g e_h e_i e_j e_k e_l e_m

c) News Semantic Snapshot e_a e_c e_h e_j e_k e_m

EURECOM

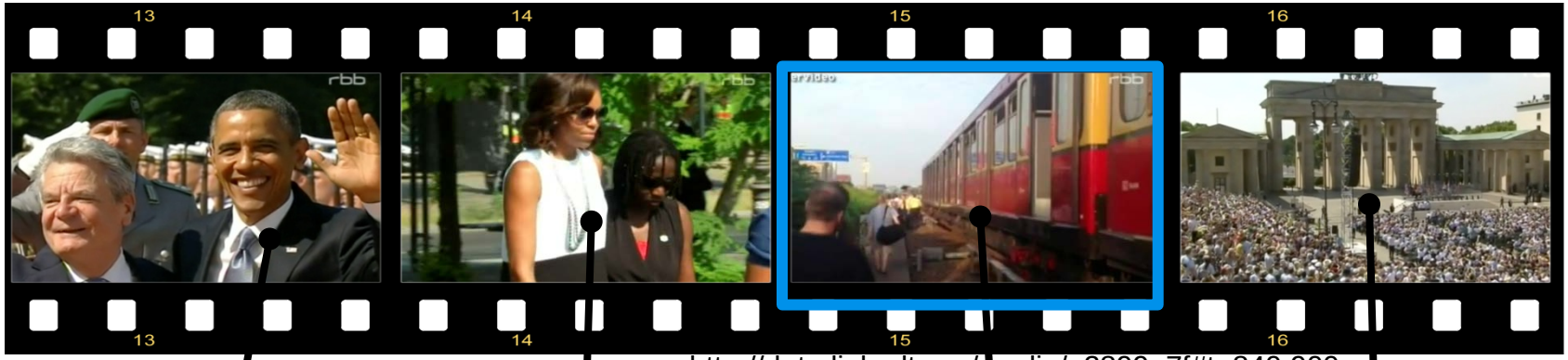
Named Entity Recognition

N·E·R·D ML

https://github.com/giusepperizzo/nerdml

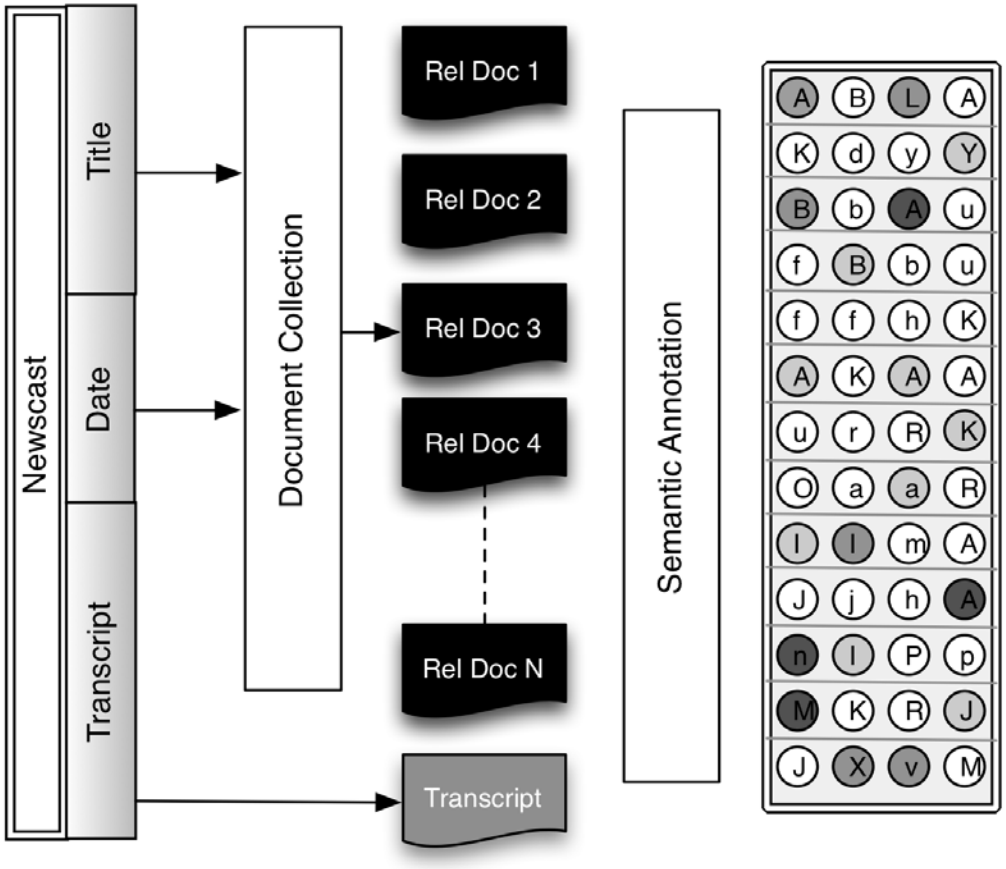
[Rizzo_LREC'14]

[1] ontology http://nerd.eurecom.fr/ontology
[2] API http://nerd.eurecom.fr/api/application.wadl
[3] UI http://nerd.eurecom.fr



Generating the NSS: Expansion's Settings

[Redondo_ICWE'15]



Parameters:

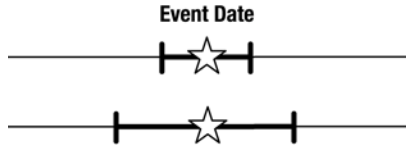
Query:

- Title
- 5 W's over Subtitles Entities

Web sites to be crawled:

- **Google**
- **L1** : A set of 10 internationals English speaking newspapers
- **L2** : A set of 3 international newspapers used in GS

Temporal Window:



- **1W:**
- **2W:**

Annotation filtering

- **Schema.org**

Available @ http://linkedtv.eurecom.fr/entitycontext/api/

Generating the NSS: Expansion Results

a) Entities from Seed Document D_S e_a e_b e_c e_d

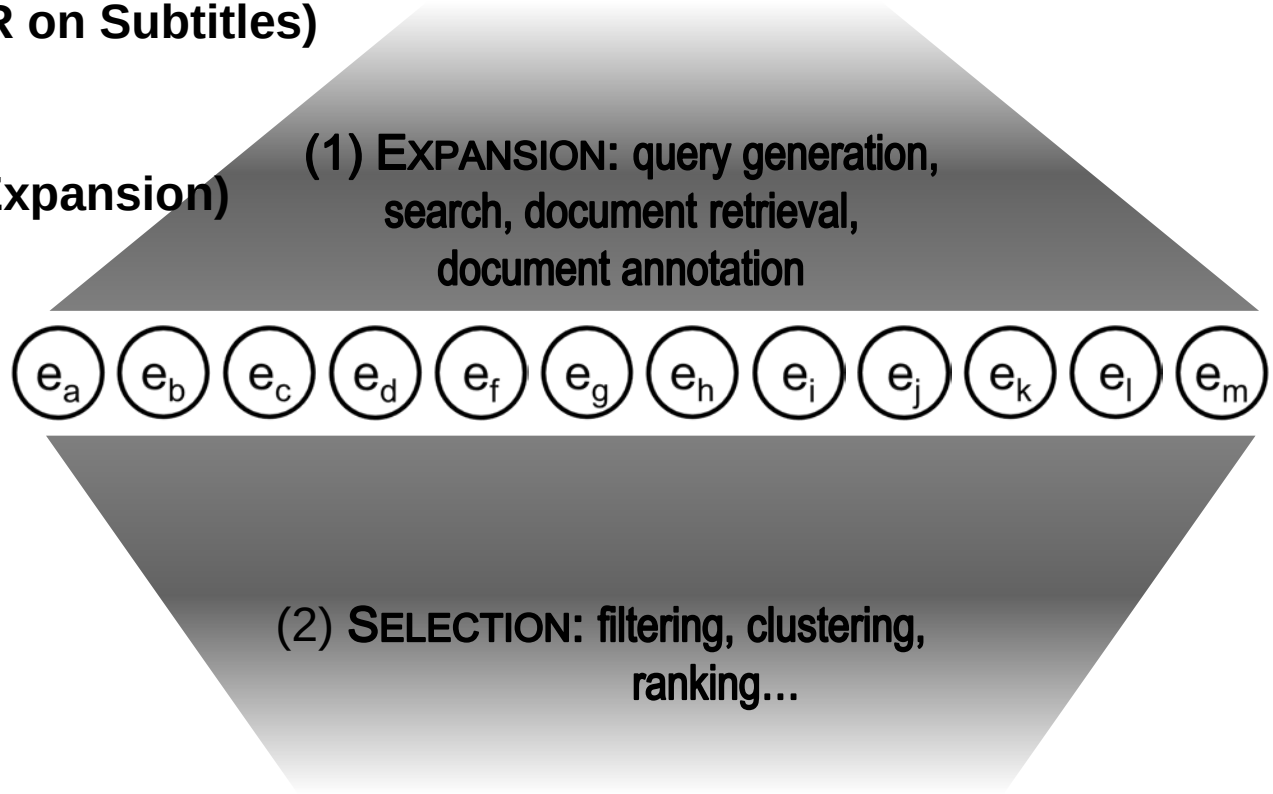
[Redondo_SNOW'14]

Recall (NER on Subtitles)

= 0.42

Recall (E. Expansion)

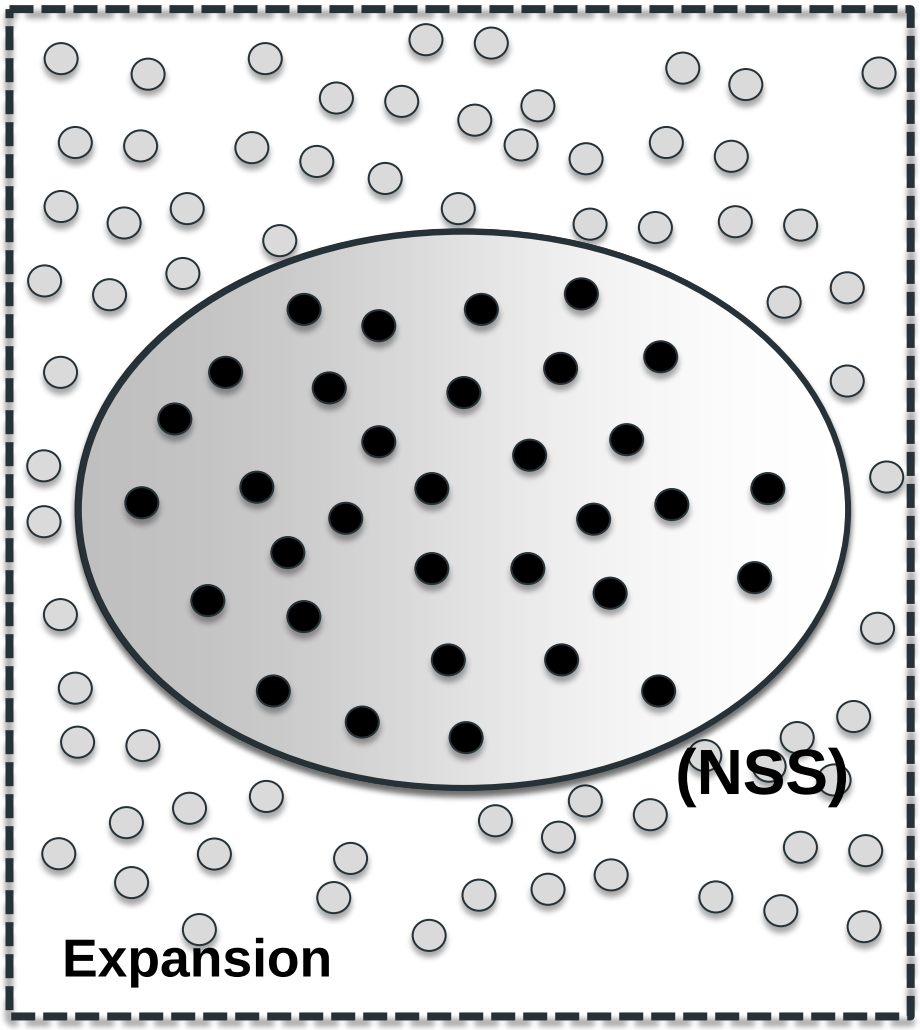
= 0.91

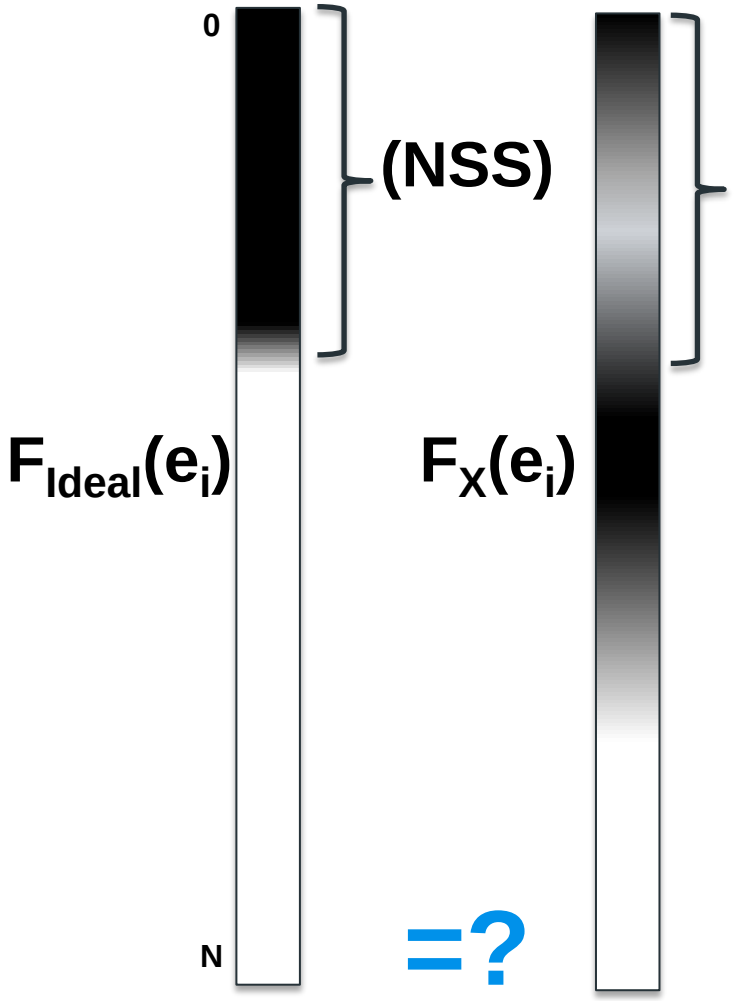


b) **Expanded Entities**

c) News Semantic Snapshot e_a e_c e_h e_j e_k e_m

Generating the NSS: The Selection problem





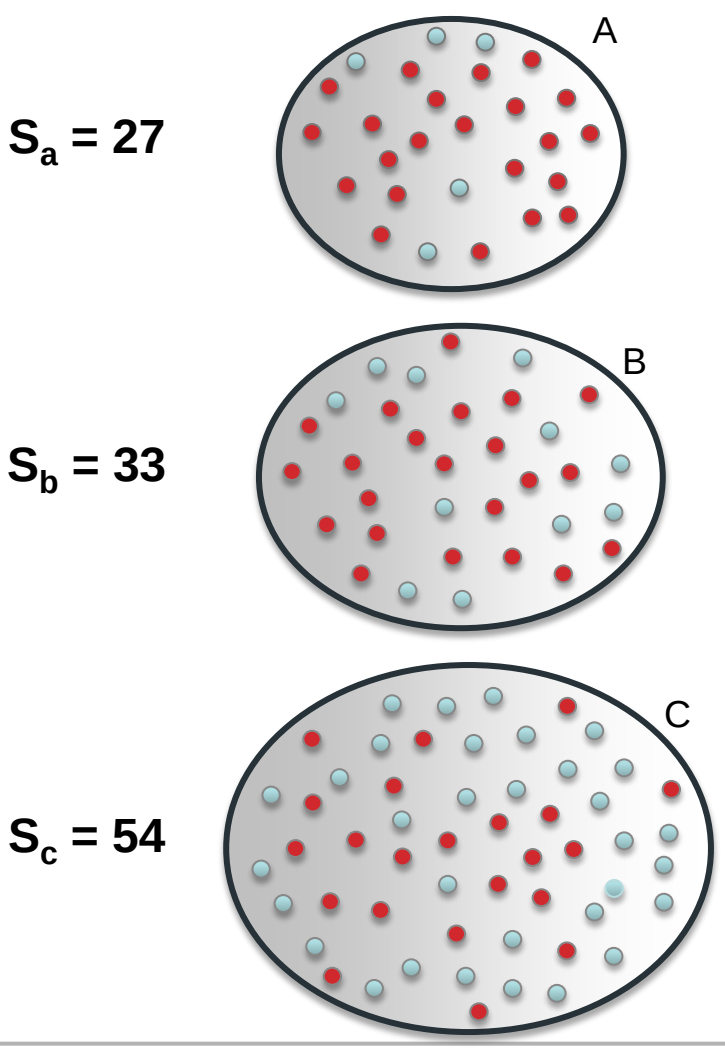
EURECOM

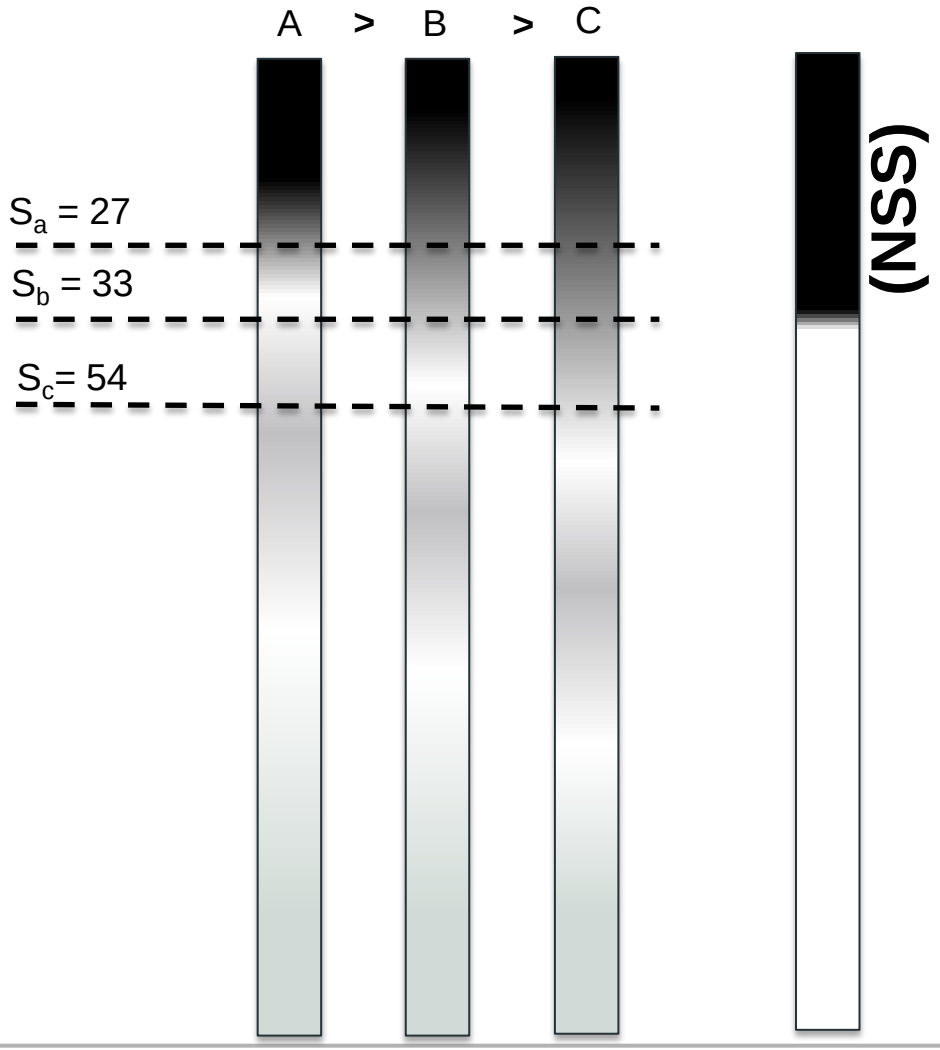
Generating the NSS: Measures

1. Precision / Recall @ N
   - Popular
   - Easy to interpret
2. Mean Normalized Discounted Cumulative Gain (MNDCG) @ N:
   - Considers ranking
   - Relevant documents at the top positions
3. Compactness for Recall R:
   - $Com\,(R, f, v) = |min(NSS \in Res)| \mid f(NSS) \geq v$

Generating the NSS: Compactness Example

Recall: **22/33 = 0.66**





EURECOM

Generating the NSS: The Approaches

1 Frequency-Based Ranking [Redondo_SNOW'14]

- Leverages on biggest sample provided by expansion
- Prioritizes representativeness

2 Multidimensional Entity Relevance Ranking [Redondo_ICWE'15]

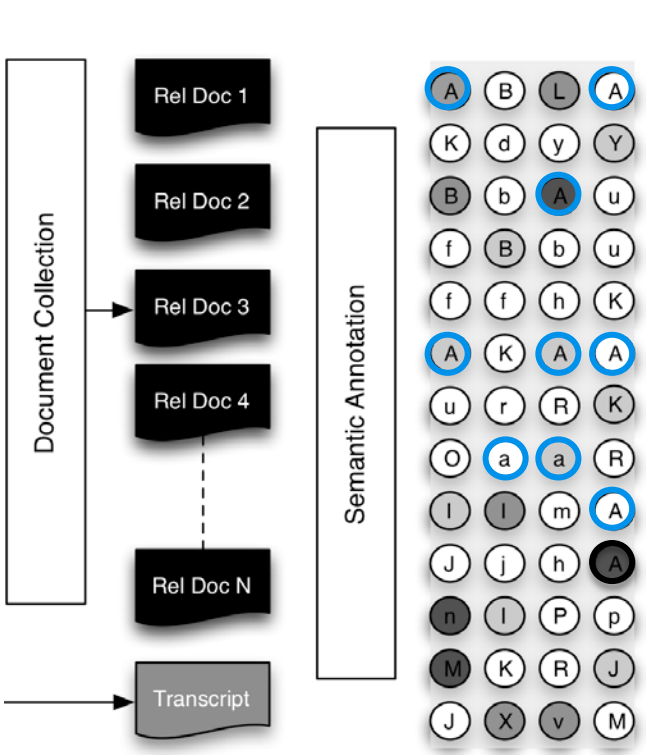
- Relevancy of entities is ground on different dimensions

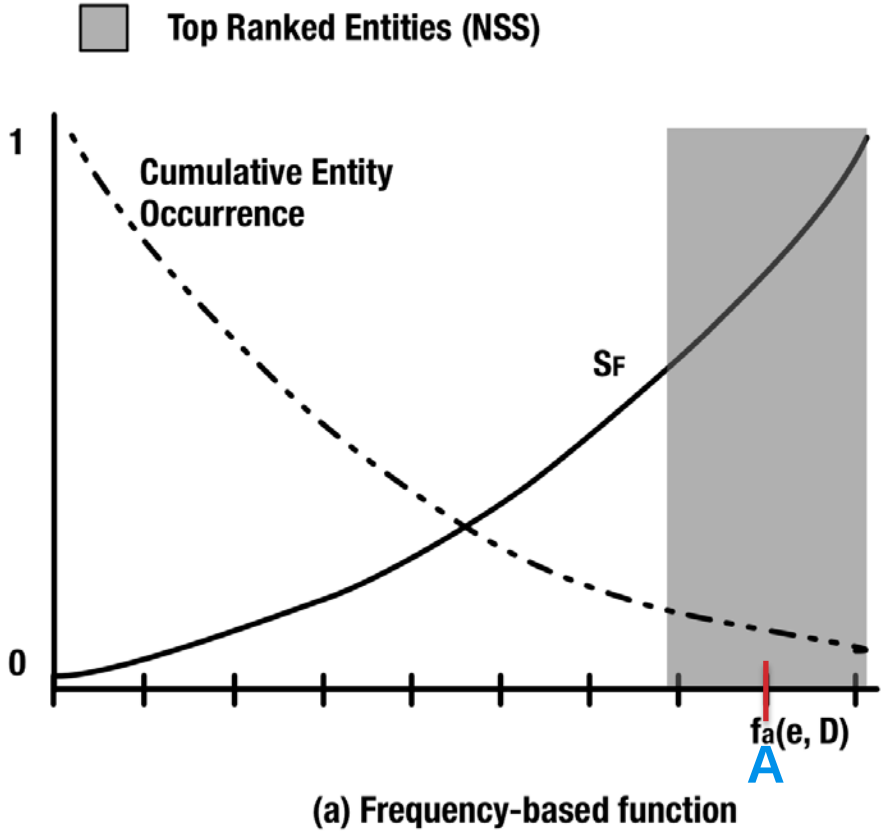
3 Concentric Based Approach [Redondo_KCAP'15A]

- Core / Crust model
- Alleviates the problem of dealing with many dimensions

Generating the NSS: (1) Frequency-Based

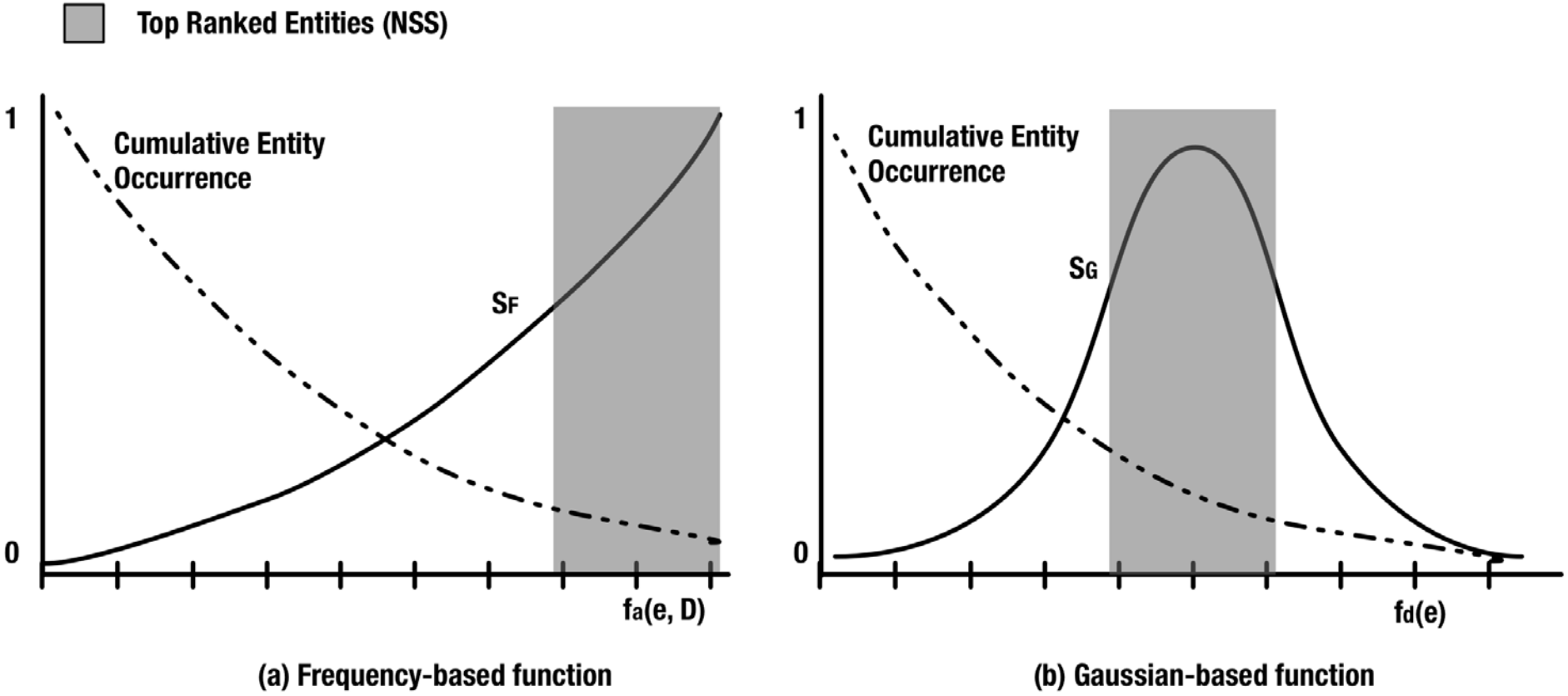
[Redondo_SNOW'14]





Generating the NSS: (2) Multidimensional

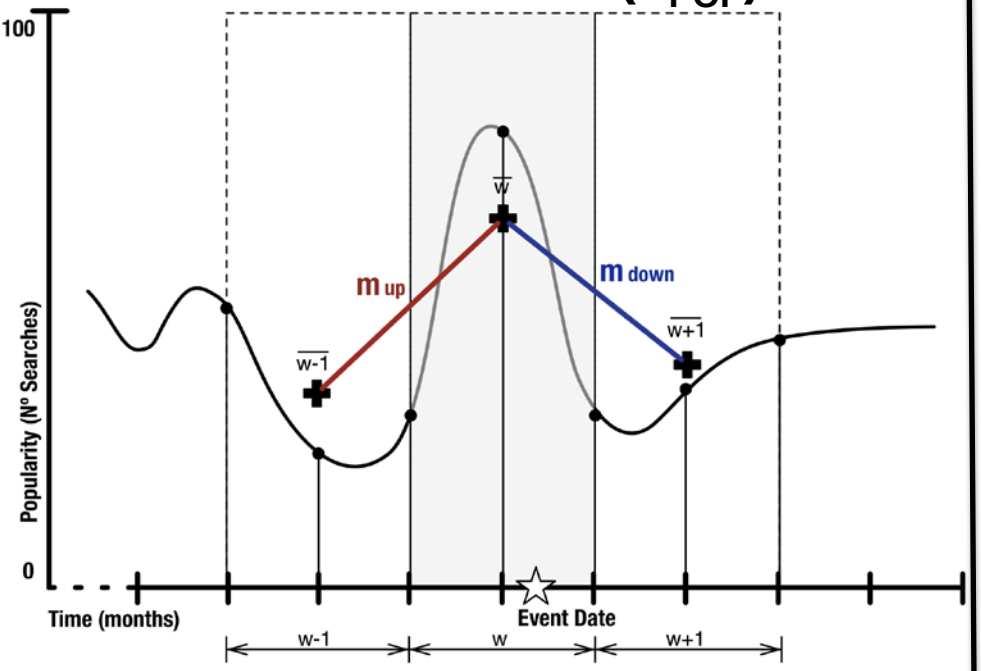
[Redondo_ICWE2015]



EURECOM

Generating the NSS: (2) Multidimensional

POPULARITY (F_{POP})



- Based on **Google Trends**
  - w = 2 months
  - μ + 2*σ (2.5%)

EXPERT RULES (F_{EXP})

$$S_{expert}(e) = S_{F-1}(e) * Op_{expert}$$

Example:

- [Location, = 0.43]
- [Person, = 0.78]
- [Organization, = 0.95]
- [$f_{doc}(e_i) < 2$, = 0.0]

EURECOM

Experiment 1: Frequency *vs* Multidimensional

20 x 4 x 4 =

320 formulas

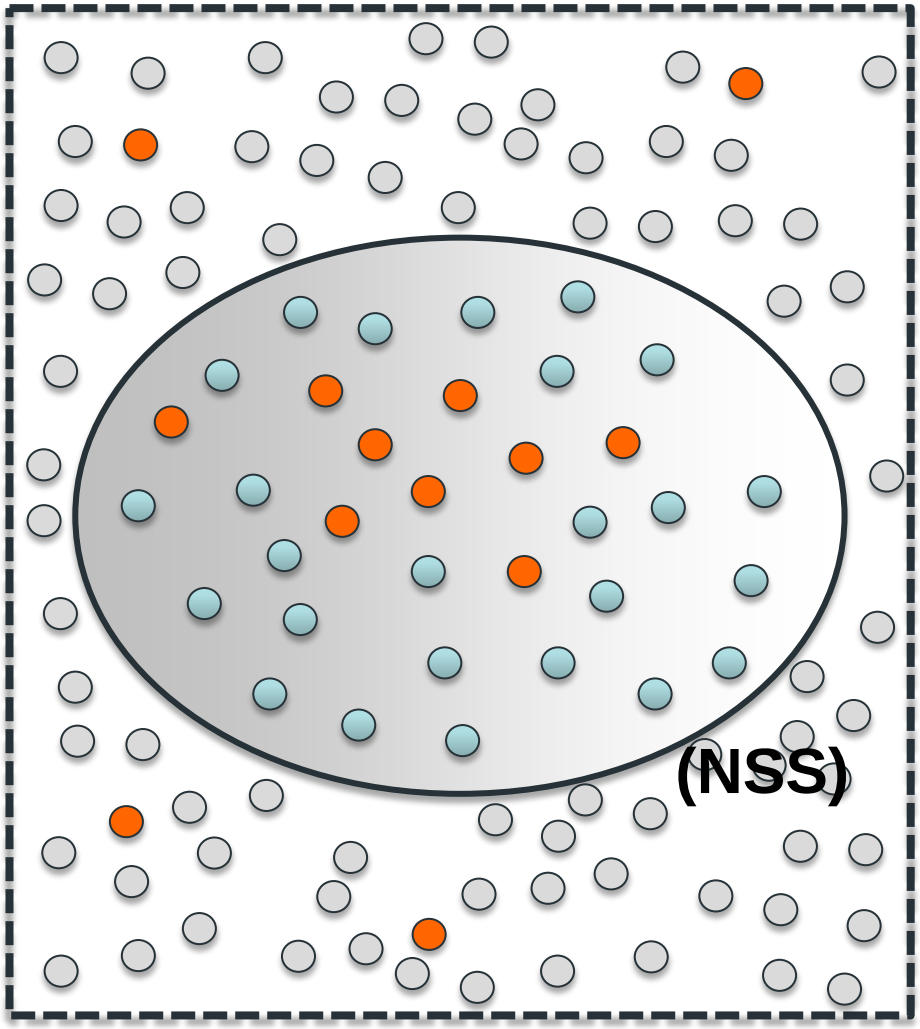
Run	Collection			Filtering	Functions			Result			
	Sources	T_{Window}	Schema.org		Freq	Pop	Exp	$MNDCG_{10}$	MAP_{10}	MP_{10}	MR_{10}
Ex0	L1+Google	2W		F3	Freq		✓	0.698	0.93	0.68	0.35
Ex1	L2+Google	2W		F3	Freq		✓	0.695	0.93	0.68	0.35
Ex2	L1+Google	2W	✓	F1+F3	Freq		✓	0.689	0.93	0.62	0.31
Ex3	L1	2W	✓	F3	Freq		✓	0.681	0.9	0.64	0.35
Ex4	L2+Google	2W		F1+F3	Freq		✓	0.679	0.92	0.7	0.36
Ex5	L1+Google	2W	✓	F1+F3	Freq		✓	0.67	0.91	0.62	0.31
Ex6	L1	2W	✓	F3	Freq	✓	✓	0.668	0.86	0.6	0.32
Ex7	L2+Google	2W		F3	Freq	✓	✓	0.659	0.85	0.56	0.29
Ex8	Google	2W		F3	Freq		✓	0.654	0.88	0.66	0.34
Ex9	L1	2W		F3	Freq		✓	0.654	0.88	0.66	0.35
Ex10	Google	2W	✓	F1+F3	Freq		✓	0.653	0.9	0.62	0.31
Ex11	Google	2W		F3	Freq	✓	✓	0.653	0.81	0.56	0.29
Ex12	L1+Google	2W	✓	F1+F3	Freq			0.652	0.93	0.64	0.32
Ex13	L2	2W	✓	F3	Freq		✓	0.651	0.89	0.64	0.34
Ex14	Google	2W		F1+F3	Freq		✓	0.649	0.88	0.64	0.33
Ex15	L2+Google	2W		F1+F3	Freq			0.649	0.94	0.72	0.37
Ex16	L1+Google	2W		F3	Freq			0.649	0.9	0.68	0.35
Ex17	Google	2W		F1+F3	Freq			0.648	0.93	0.72	0.37
Ex18	L1	2W		F1+F3	Freq		✓	0.646	0.89	0.66	0.34
Ex19	L1+Google	2W		F1+F3	Freq			0.646	0.94	0.7	0.37
Ex20	L1+Google	2W		F1+F3	Freq		✓	0.646	0.89	0.66	0.34
...	...	...	...	...	...	...	...	...	...	...	...
Ex78	Google	2W	✓	F1+F3	Gaussian		✓	0.552	0.66	0.66	0.34
Ex80	L2+Google	2W	✓	F1+F3	Gaussian		✓	0.55	0.69	0.7	0.36
Ex82	L1	2W	✓	F3	Gaussian		✓	0.549	0.68	0.64	0.33
...	...	...	...	...	...	...	...	...	...	...	...
BS2	Google	2W			Freq			0.473	0.53	0.42	0.22
...	...	...	...	...	...	...	...	...	...	...	...
BS1	Google	2W			TFIDF			0.063	0.08	0.06	0.03

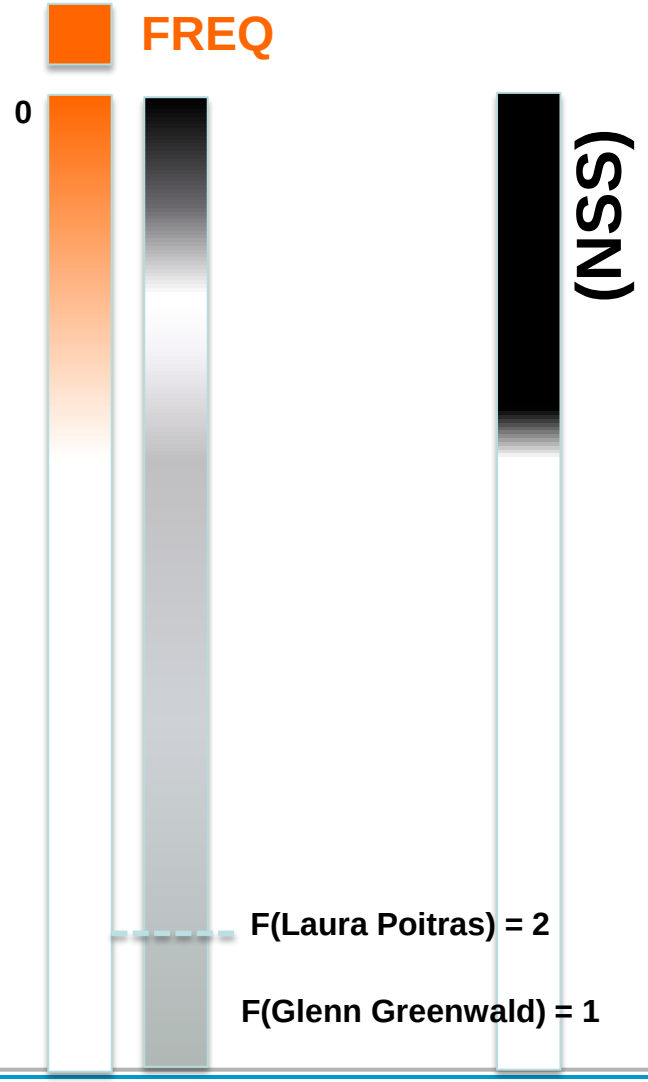
Experiment 1: Frequency *vs* Multidimensional

- News Entity Expansion & Dimensions ➔ Generate NSS
- Frequency-based score: 0.473 **MNDCG @ 10**
- Best score: **0.698 MNDCG @ 10**
  - Collection:
    - CSE (Google + 2W + ~~Schema.org~~)
  - Ranking:
    - Expert Rules
    - Popularity

Multidimensional Nature of the NSS

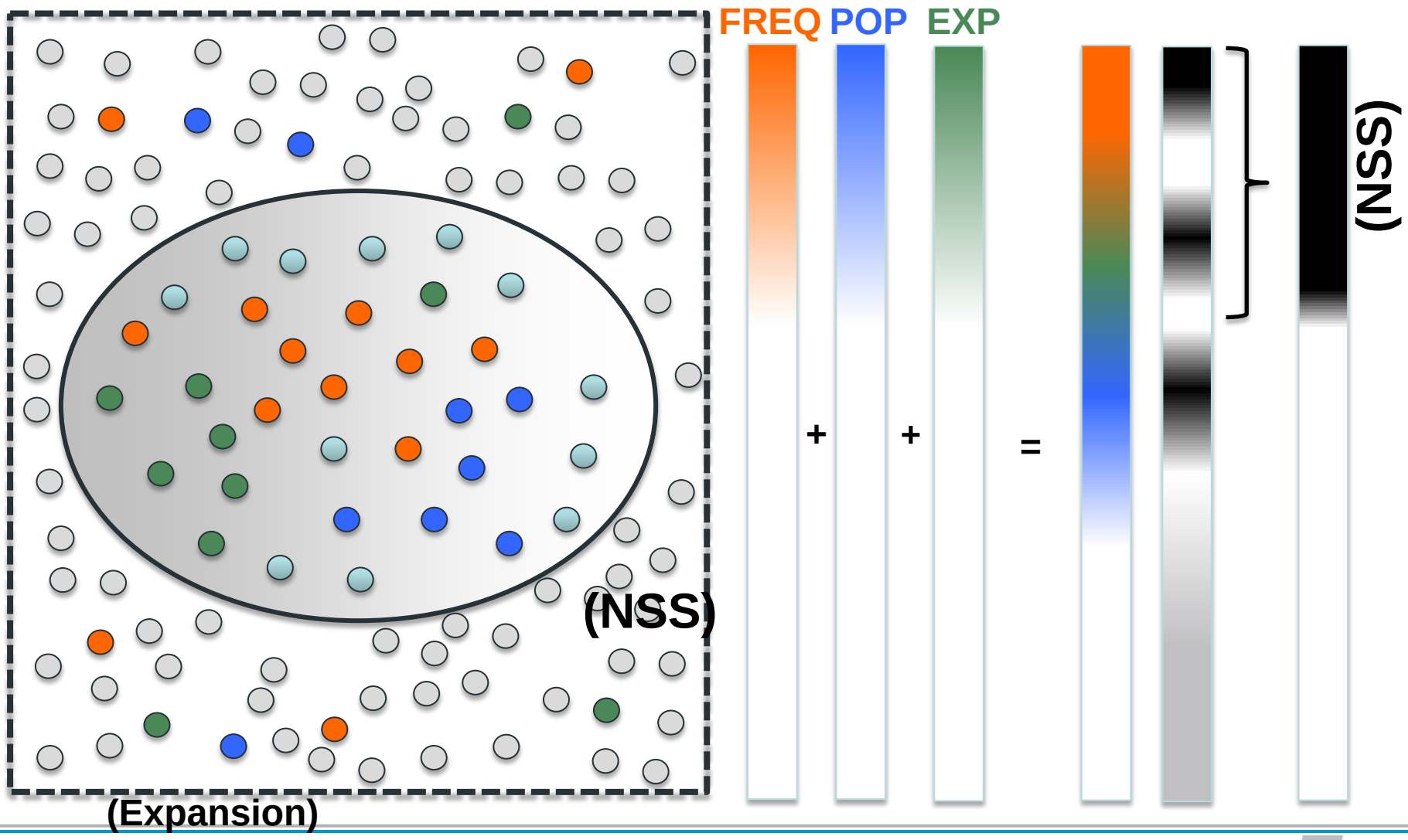
Experiment 1: Frequency *vs* Multidimensional





EURECOM

Experiment 1: Frequency *vs* Multidimensional



EURECOM

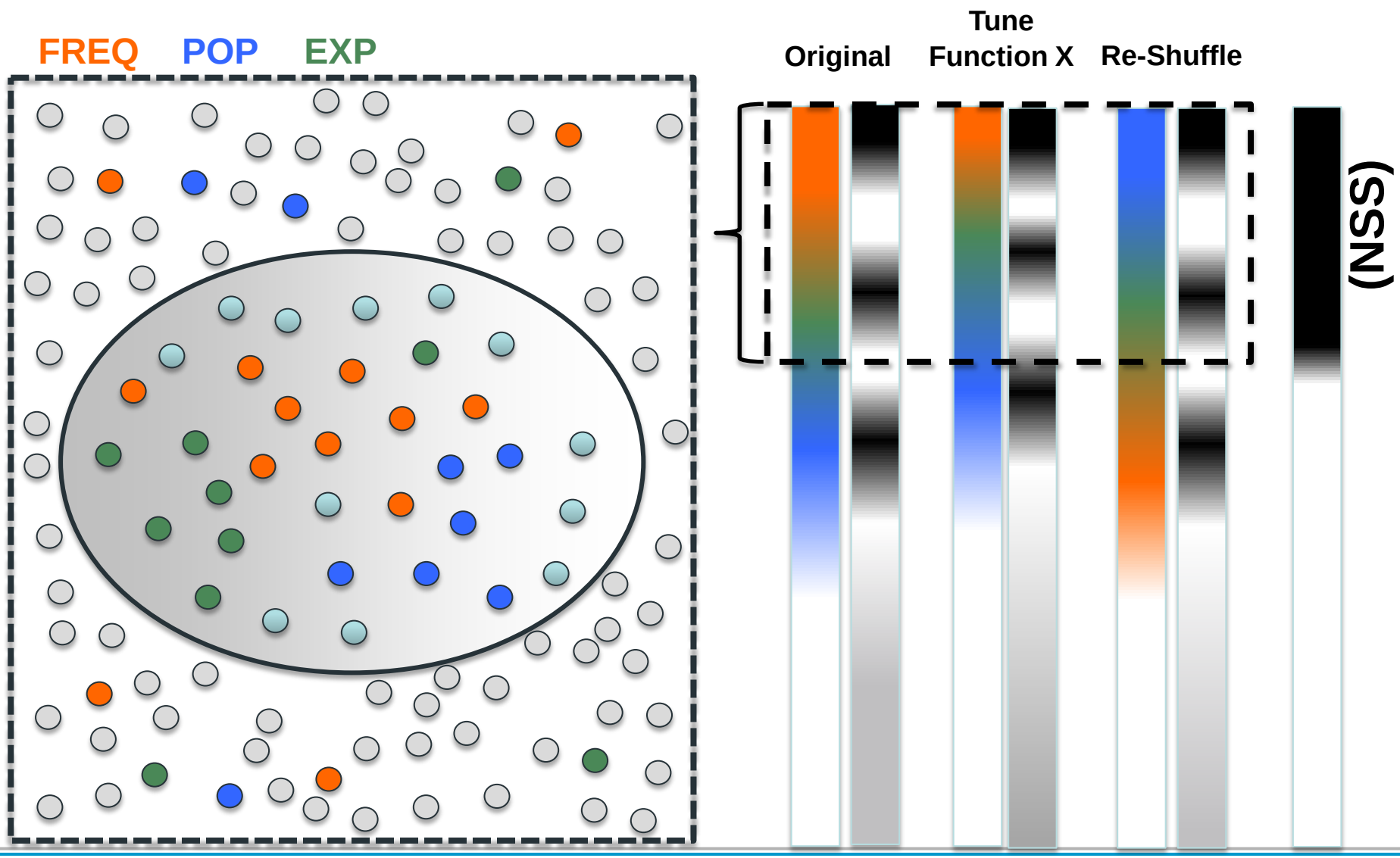
Experiment 2: Multidimensional ++

NMDCG @ 10:

1. Exploit Google relevance (+1.80%)
2. Promote subtitle entities (+2.50%)
3. Exploit named entity extractor's confidence (+0.20%)
4. Interpret popularity dimension (+1.40%)
5. Performing clustering before filtering (-0.60%)

- NO SIGNIFICANT IMPROVEMENT -

Experiment 2: Multidimensional ++



EURECOM

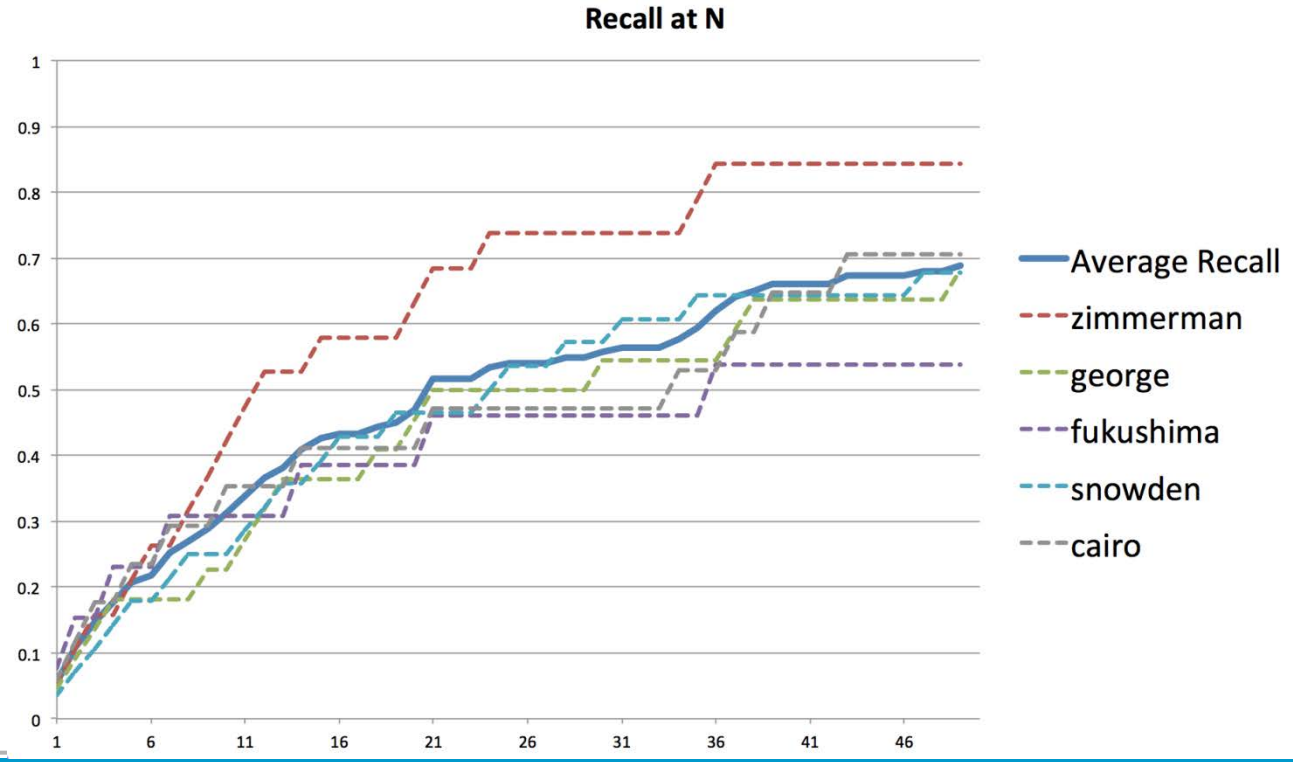
Re-thinking the problem: measures

MNDCG:

- Too focused on success at first positions (decay Function)
- NSS intends to be flexible, ranking is application-dependent

COMPACTNESS:

- Prioritizes coverage over ranking while minimizing NSS size



EURECOM

Re-thinking the problem: dimensions

Duality in news entity spectrum:

- Representative entities:
  - Driving the **plot** of the story
- Relevant entities
  - Related to former via specific reasons
- Exploit the entity semantic relations

EURECOM

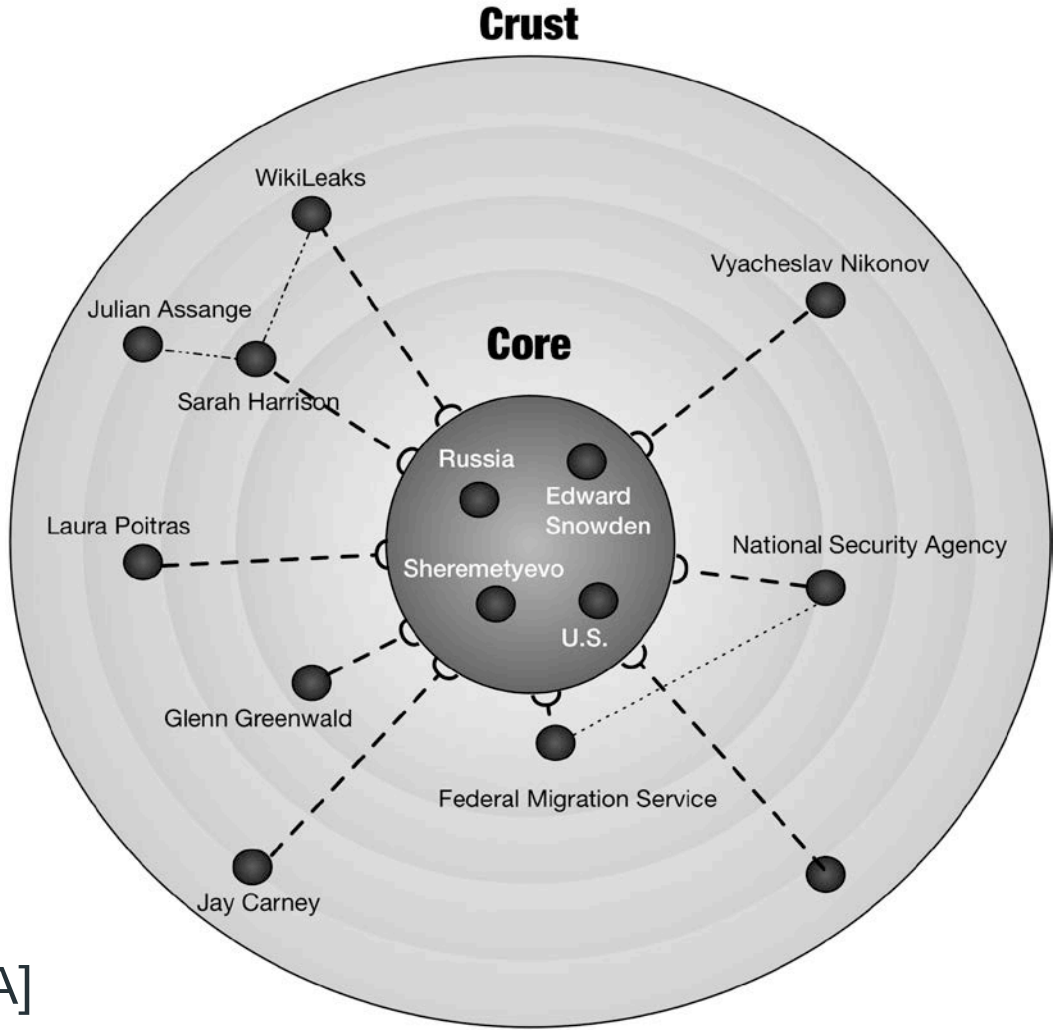
Generating the NSS: (3) Concentric Approach

Core

- Representative entities
- Spottable via frequency dimensions
- High degree of cohesiveness

Crust

- Attached to the Core via semantic relations
- Agnostic to relevancy nature: informativeness, interestingness, etc.



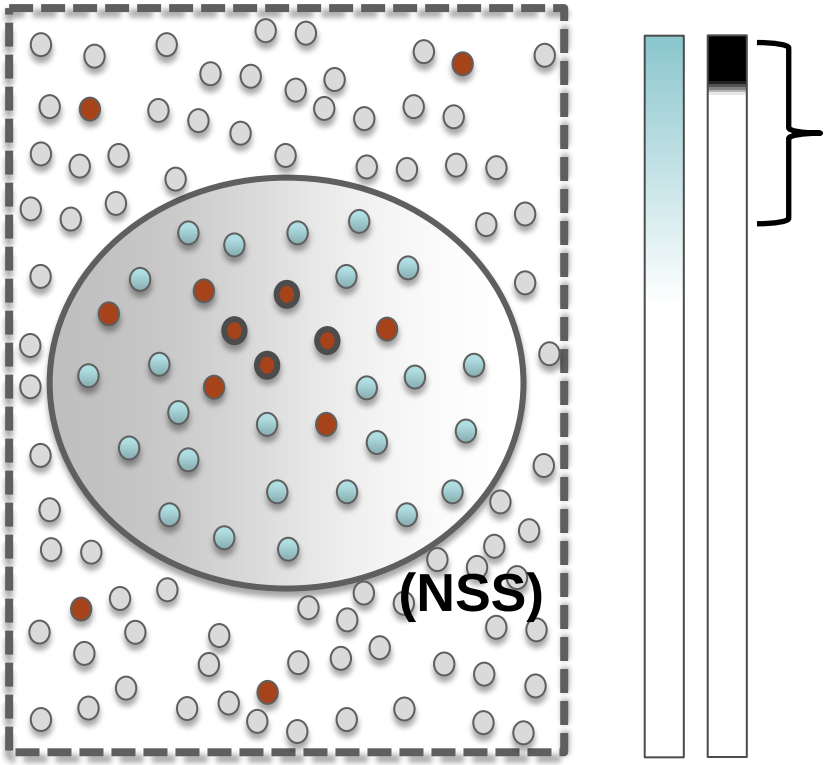
[Redondo_KCAP2015A]

EURECOM

Generating the NSS: (3) Core Creation

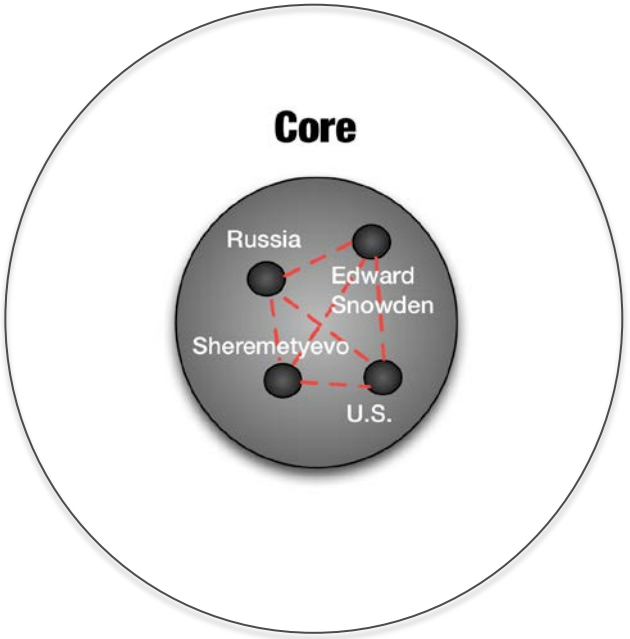
a) Spot representative entities: Frequency Dimension

$$f_{Core}(e, D) = f_{doc}(e_i, D) + \frac{f_a(e_i, D)}{f_{doc}(e_i, D)}$$



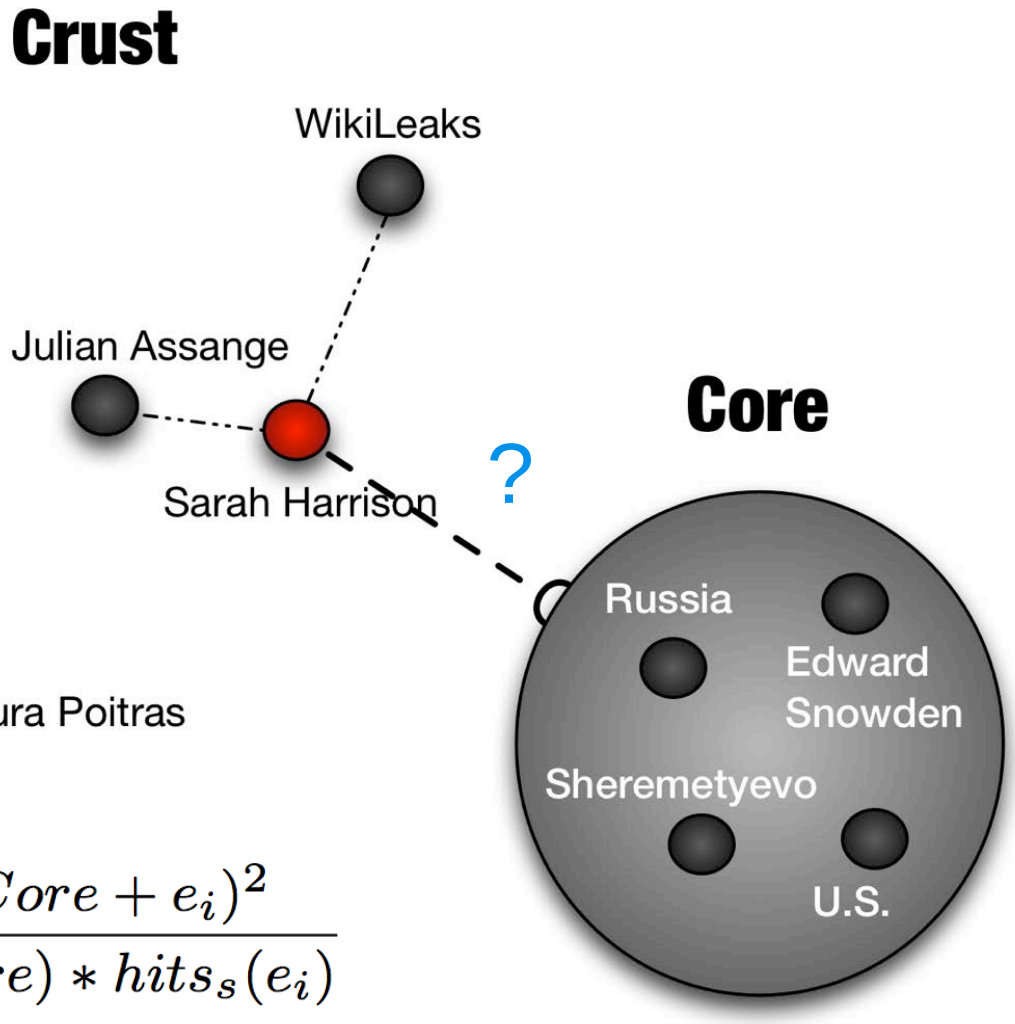
b) Cohesiveness (DBpedia)

$$S_{KB}(e_i, e_j) = \sum_{1}^{p} \frac{1}{|path_{i,j}|}$$



Generating the NSS: (3) Crust Creation

The number of Web documents talking simultaneously about a particular entity e and the Core:



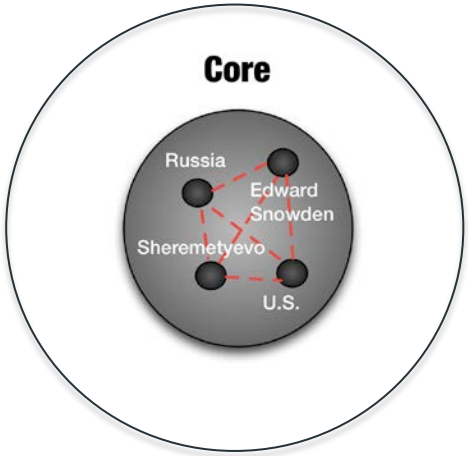
$$S_{Web}(e_i, Core) = \frac{hits_s(Core + e_i)^2}{hits_s(Core) * hits_s(e_i)}$$

Experiment 3: Multidimensional *vs* Concentric

Concentric Core:

1. Entity Frequency
   - Core1: Jaro-Winkler > 0.9
   - Core2: Frequency based on Exact String matching

2. Cohesiveness:
   - Everything is Connected Engine, $S_{kb}(e1, e2) > 0.125$



Everything is Connected Engine:

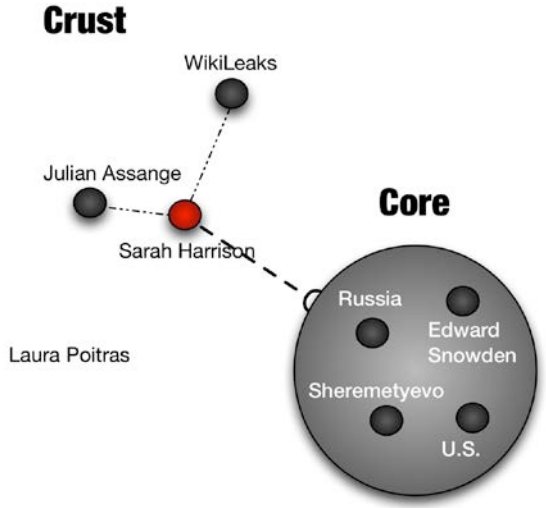
https://github.com/mmlab/eice

Experiment 3: Multidimensional *vs* Concentric

Concentric Crust:

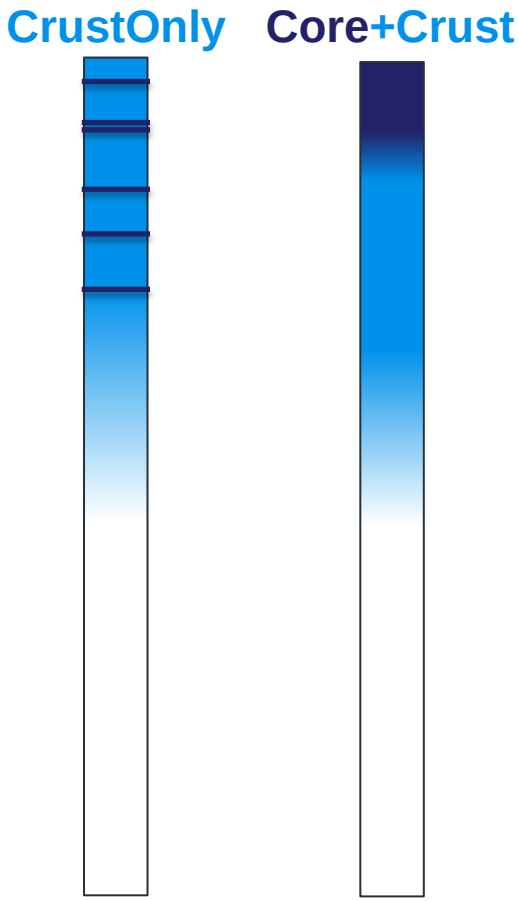
1. Candidates for CRUST generation:
   - Ex1: 1° ICWE2015 by R*(50): L2+Google, F3 1W, Gauss+ POP
   - Ex2: 2° ICWE 2015 by R*(50): L2+Google, F3 1W, Freq + POP

2. Function for attaching entities to CORE:
   - $S_{WEB}(e_i, \text{Core})$ over Google CSE, default configuration



Experiment 3: Multidimensional *vs* Concentric

Combining CORE and CRUST:



EURECOM

Experiment 3: Multidimensional *vs* Concentric

(2*2*2 + 2) Runs

IdealGT: size of SSN according to Gold Standard

Run	Expansion				$Com\,(R, f, v)$					
	Collection	Core	Crust	Fusion	v_1	v_2	v_3	v_4	v_5	Avg
IdealGT		-	-	-	16	11	22	27	19	**19**
Cm4	Ex2	CoreA	S_{Google}	Core_Crust	21	9	41	44	45	**32**
Cm5	Ex2	CoreA	S_{Google}	CrustBased	20	14	41	44	45	**32.8**
Cm6	Ex2	CoreB	S_{Google}	Core_Crust	27	10	43	44	42	**33.2**
Cm0	Ex1	CoreA	S_{Google}	Core_Crust	22	13	42	43	47	**33.4**
Cm1	Ex1	CoreA	S_{Google}	CrustBased	21	16	42	43	47	**33.8**
Cm7	Ex2	CoreB	S_{Google}	CrustBased	27	13	43	44	42	**33.8**
Cm2	Ex1	CoreB	S_{Google}	Core_Crust	28	13	43	43	44	**34.2**
Cm3	Ex1	CoreB	S_{Google}	CrustBased	28	16	43	43	44	**34.8**
BAS01	L2+AllGoogle, 1W F3 Gaussian + EXP + POP	-	-	-	41	45	34	41	37	**39.6**
BAS02	L2+AllGoogle, 1W F3 Freq + EXP + POP	-	-	-	24	39	49	48	39	**39.8**

36.9% more compact than **Multidimensional**
(NSS's size decrease)

Experiment 3: Multidimensional *vs* Concentric

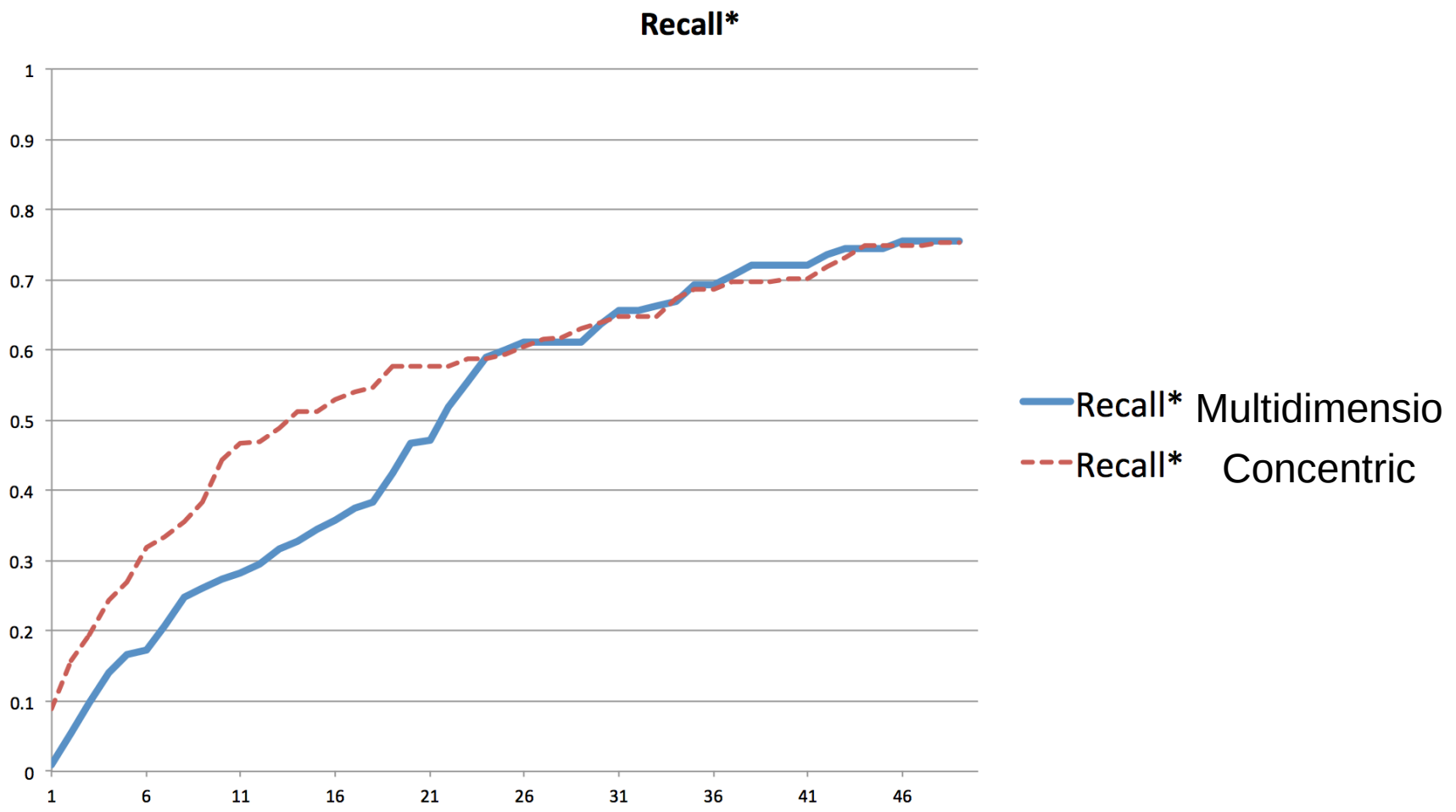
MULTIDIMENSIONAL	CONCENTRIC
Tokyo Electric Power Company	Fukushima Daiichi nuclear disaster
Fukushima Daiichi nuclear disaster	Tokyo Electric Power Company
Japan	Fukushima Daiichi nuclear power plant
BBC News	Nuclear Regulation Authority
Barack Obama	Japanese government
Pacific Ocean	Nuclear Regulatory Commission
Nuclear Regulatory Commission	Pacific Ocean
concern	Chernobyl disaster
plant	TOKYO
Radiation	Radiation
Nuclear Reactor	Groundwater
Liberal Democratic Party	Liberal Democratic Party
United States	Edward Snowden
No. 1	NHK
Officials	Japan
Canada	workers
23 Jul	San Onofre Nuclear Generating Station
Tokyo	Officials
Steam	NRA
Groundwater	NSA
Nuclear Regulation Authority	Barack Obama
Nuclear & Energy	Tokyo University of Marine Science and Technology

NSS Gold Standard

n=22

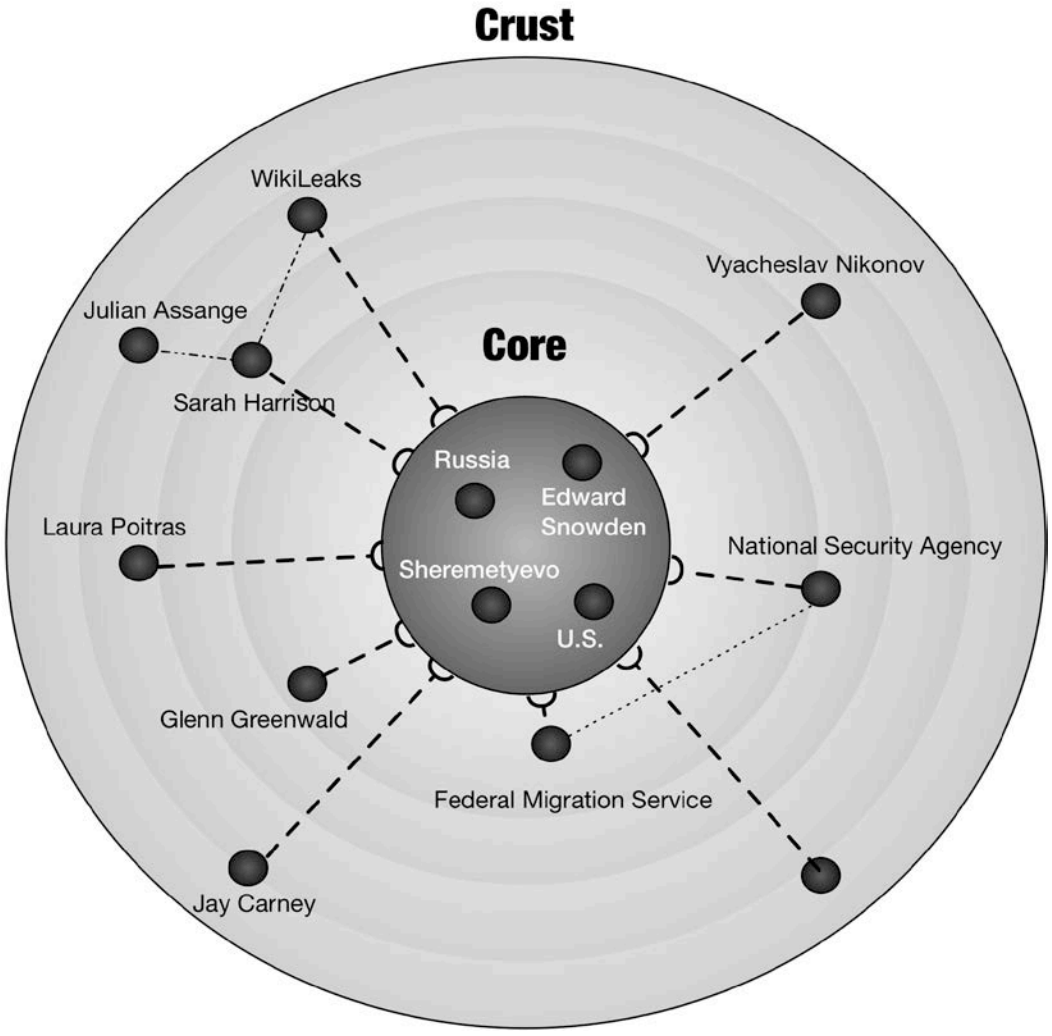
Fukushima Disaster 2013

Experiment 3: Multidimensional *vs* Concentric



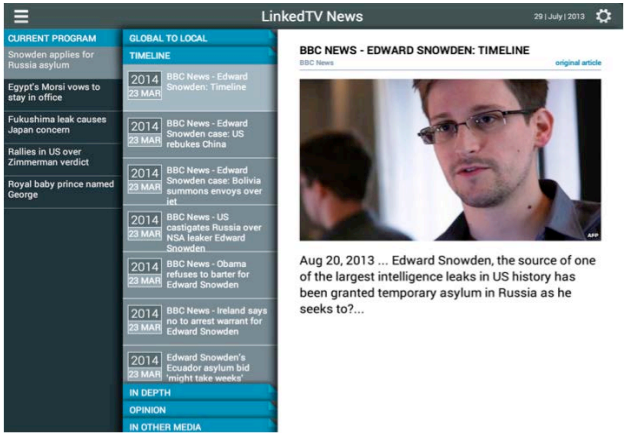
EURECOM

NSS: Suitable model for news applications ?



EURECOM

NSS Consumption: News Prototypes



… short summaries, previews, hotspots …

… second screen apps, slideshows, info-boxes …

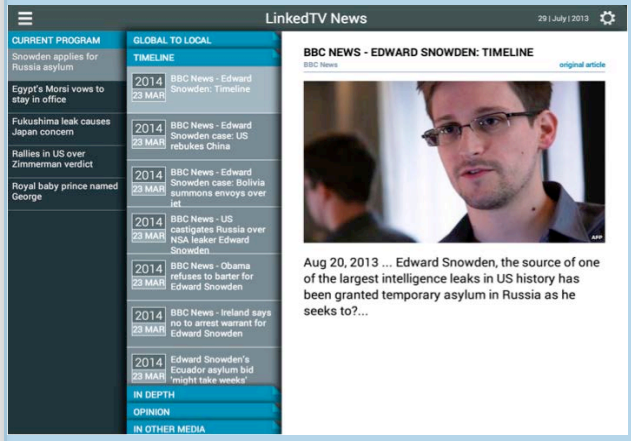
… advanced graphs and diagrams, timelines, in-depth summaries …

EURECOM

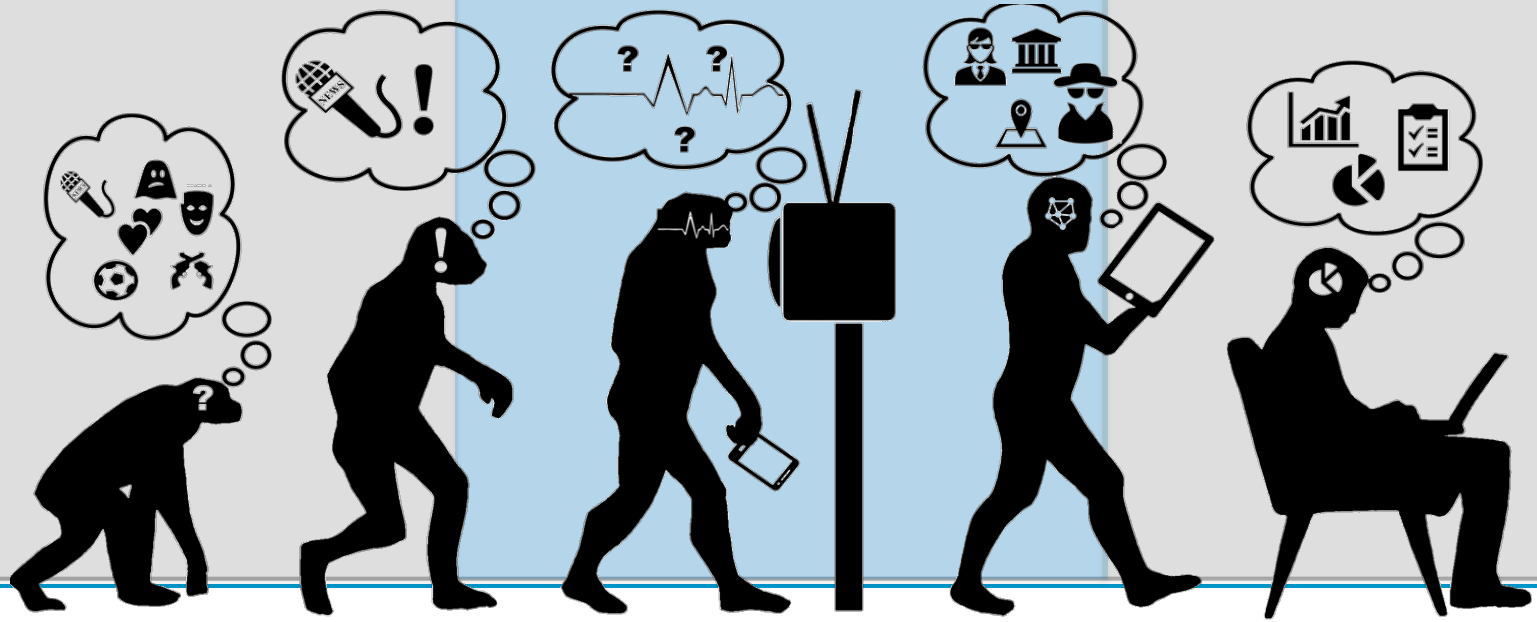
NSS Consumption: Consumptions Phases

The Before

The During



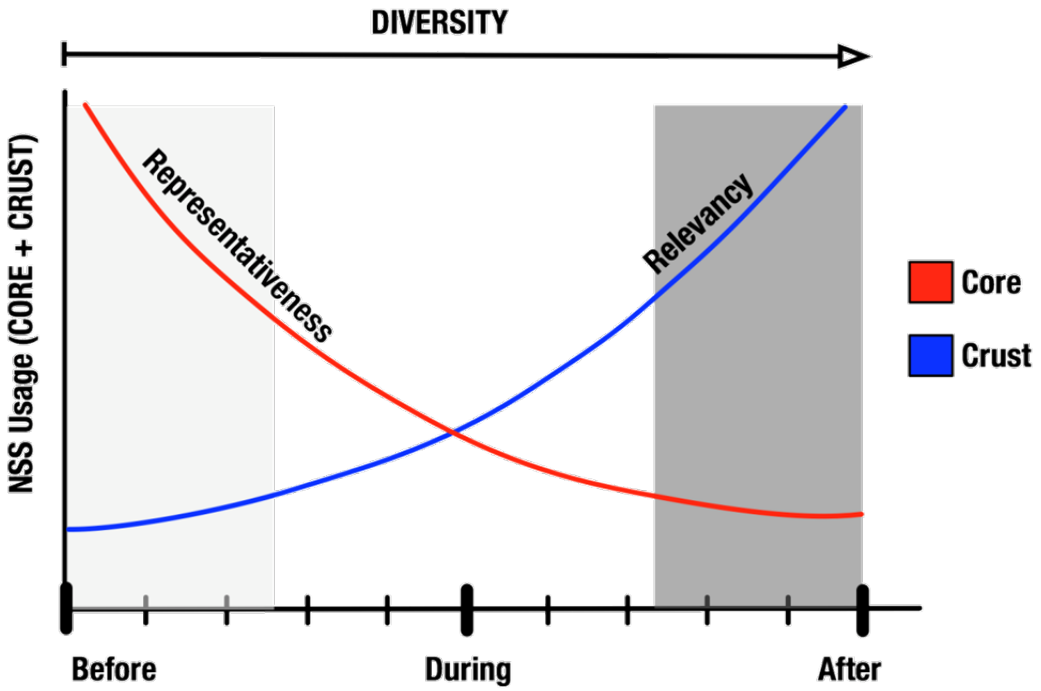
The After

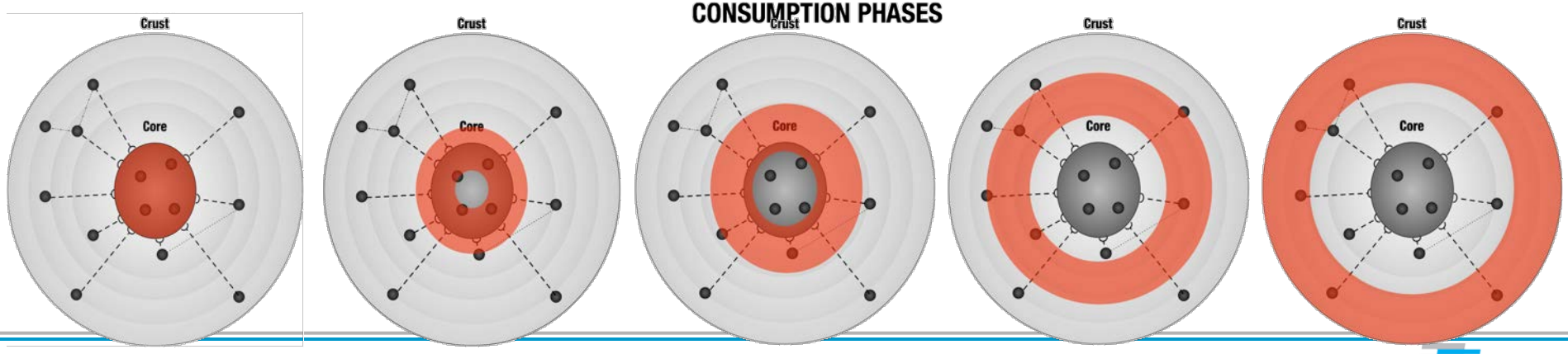


EURECOM

NSS Consumption: Phases VS Layers

[Redondo_KCAP'15B]





EURECOM

Conclusions

a. Proposed the NSS model and a Gold Standard

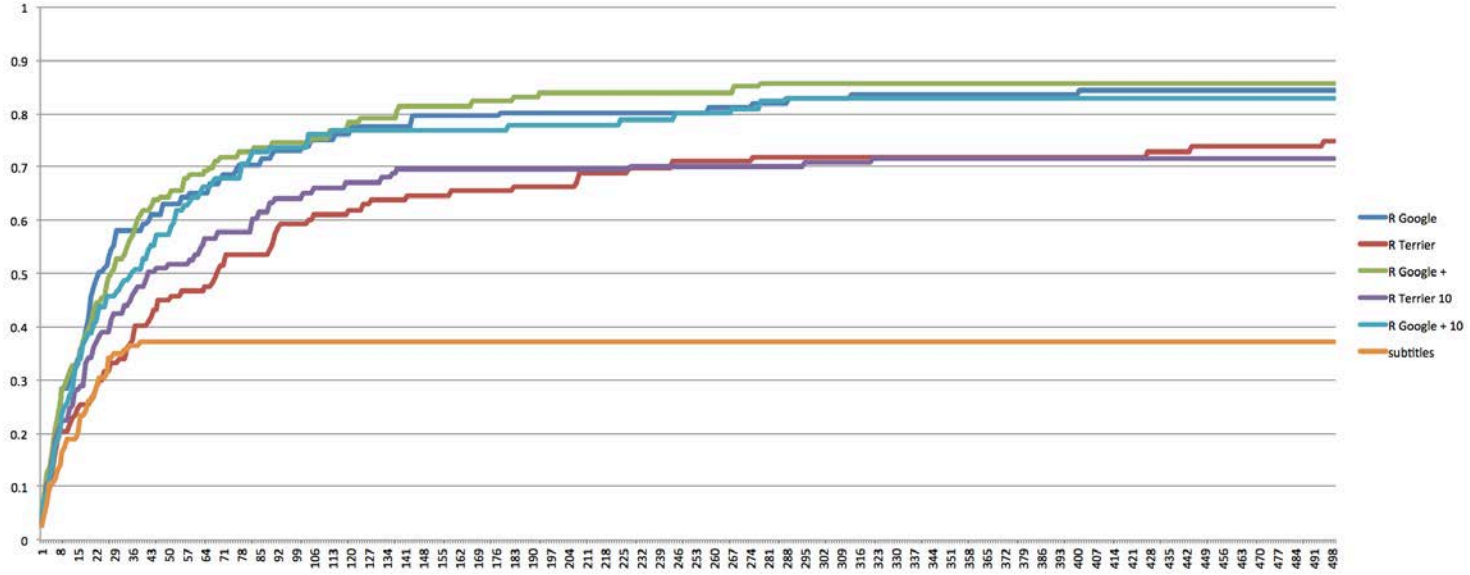
b. The multidimensional nature of the entity relevance
   - Gaussian function, popularity, experts rules…

c. Concentric model better reproduces the NSS:
   - Better Compactness: 36.9% over BAS01 (similar recall, smaller size)
   - Core/Crust brings up relevant entities without having to deal with fuzzy dimensions

d. NSS better supports the news consumption phases: (Before, During, After)

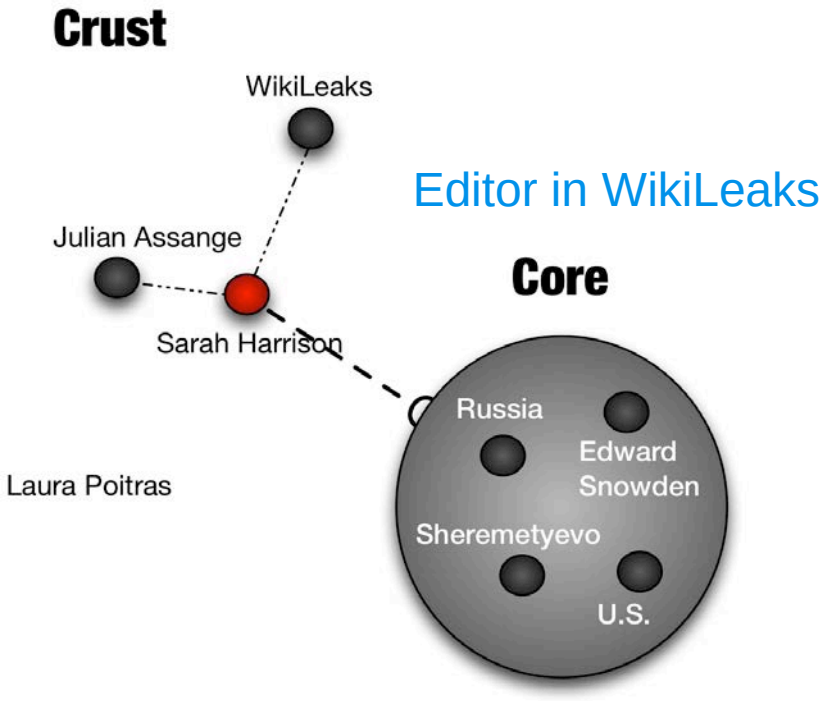
Future Work

- [S] Publish generated NSS on the Web (Linked Data)
- [S] Extend the Gold Standard:
  - From 5 to 23 videos, concentric based model for candidate selection
  - Submission to TOIS
- [S] Not depending on "big players" for retrieving knowledge during the expansion phase
(Terrier VS Google experiments)



Future Work

- [L] Spot not only the strength of the relationships between **Crust** and the **Core,** but also the **predicates**



Editor in WikiLeaks

Generating **Explanations** analyzing documents considered in S_{web}

Credits

http://jluisred.github.io/

http://giusepperizzo.github.io/