Semantically Annotating and Contextualising Big Collections of Human-Readable Documents

José Luis Redondo García

Jluisred.github.io @peputo

Outline

1. Me:
   Research Interests, Projects, Publications.

2. My Research:
   Contextualize news stories.

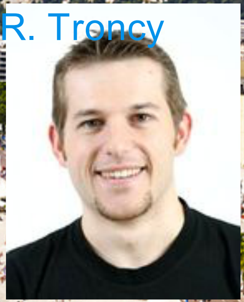
3. Future:
   Upcoming Challenges

About me…

1

PhD @ Nice
Sophia Antipolis

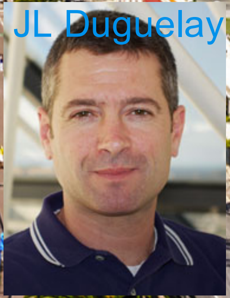
EURECOM
Sophia Antipolis



Semantic Multimedia

Natural Language Processing

Audio and Video Analysis



20+

Postdoc @ Madrid

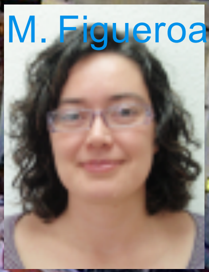
Politécnica de Madrid

Ontology Engineering Group

Ontology Engineering

Knowledge Representation

Linguistics and NPL



30+

Projects:

open SEM

MEDIA MIXER

LINKEDTV

WAVE

DRINVENTOR

3cixty
an EIT Digital initiative

NEXGEN TV

Publications

Journals (2), Conferences (6), Workshops(5), Demo/Poster(7) TOTAL (+30)

Journals

- **Redondo Garcia J. L** and Adolfo Lozano-Tello: OntoTV: an Ontology Based System for the Management of Information about Television Content. International Journal of Semantic Computing, 6(01), 111-130, 2012.

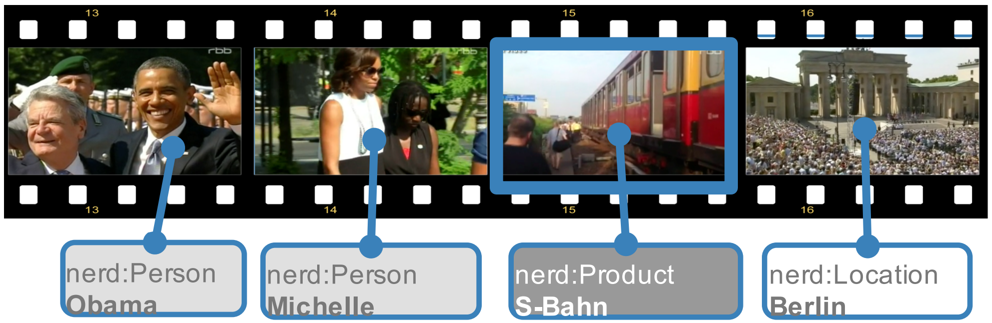
Conferences

- **Redondo Garcia J. L**., Rizzo G., Troncy R. (2015) Capturing News Stories Once, Retelling a Thousand Ways. In: 8th International Conference on Knowledge Capture (K-CAP'15), Palisades, NY, USA.
- **Redondo Garcia J. L.**, Rizzo G., Troncy R. (2015) The Concentric Nature of News Semantic Snapshots: Knowledge Extraction for Semantic Annotation of News Items. In: 8th International Conference on Knowledge Capture (K-CAP'15), Palisades, NY, USA.

  Best Paper Award
- **Redondo Garcia J. L**., Rizzo G., Romero L. P., Hildebrand M., Troncy R. (2015) Generating Semantic Snapshots of Newscasts using Entity Expansion. In: 15th International Conference on Web Engineering (ICWE'15), Rotterdam, the Netherlands.
- Rizzo G., Steiner T., Troncy R., Verborgh R., **Redondo Garcia J. L**. and Van de Walle R. (2012), What Fresh Media Are You Looking For? Extracting Media Items from Multiple Social Networks. In (ACM Multimedia) International Workshop on Socially-Aware Multimedia (SAM'12), Nara, Japan

Timeline

1. Semantic Multimedia **(2012-2014)**

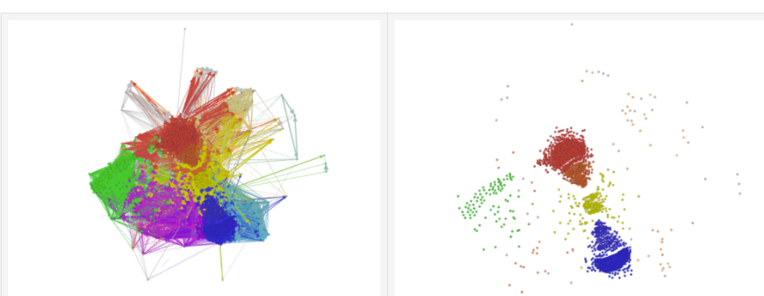


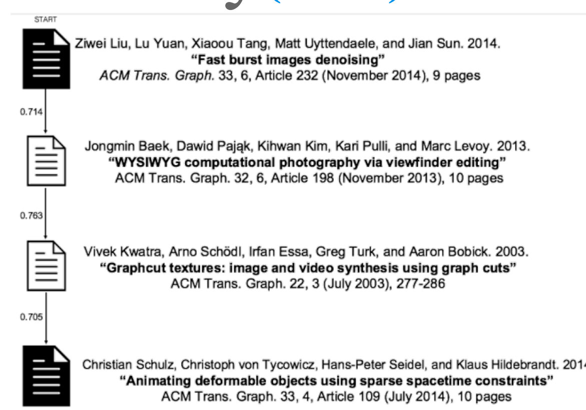
Media Fragment URI's
Multimodal Semantic Annotation

http://data.linkedtv.eu/media/e2899e7f#t=840,900

2. **Context of News Stories** **(2014-2016)**
3. Scientific Knowledge Discovery **(2016)**

- http://drinventor.dia.fi.upm.es/
- **Topic Modelling, Ontology Learning**



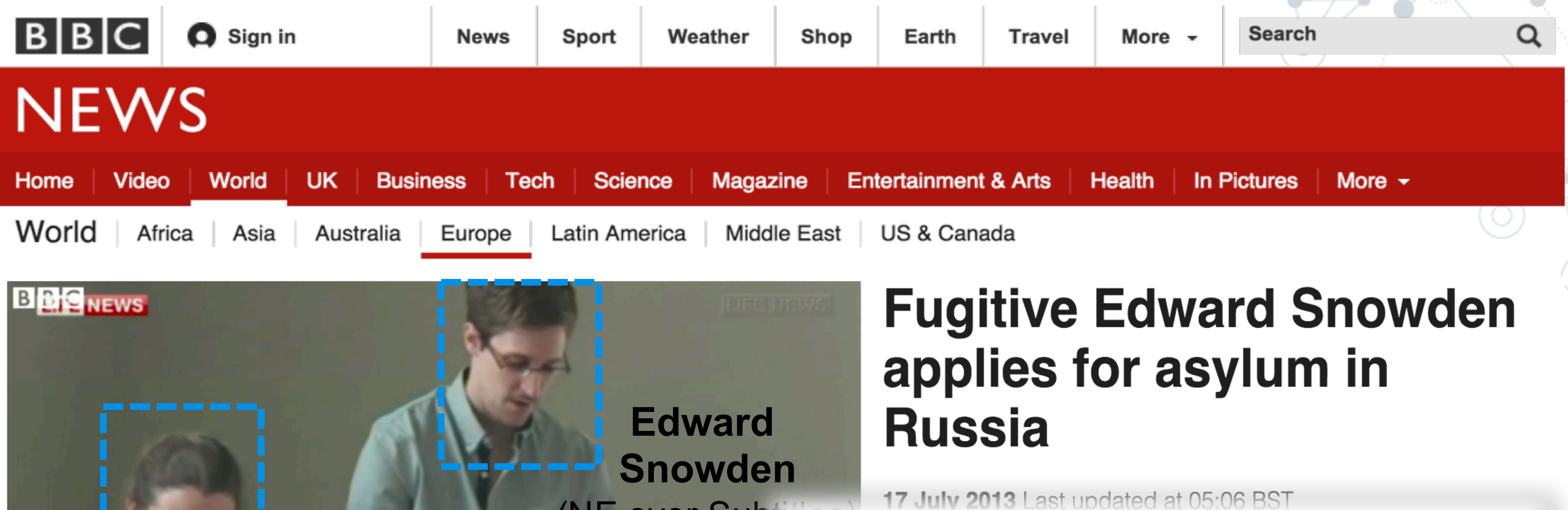


My Research

Semantically Contextualizing News Stories

2

The Use Case: Contextualizing News



Edward Snowden

(NE over Subt

http://www.bbc.com/news/world-europe-2333919

(Media Fragment

Sarah Harrison

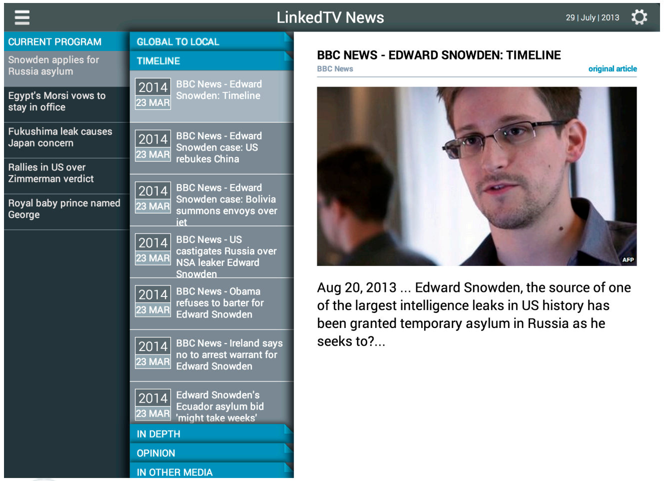
WikiLeaks Editor

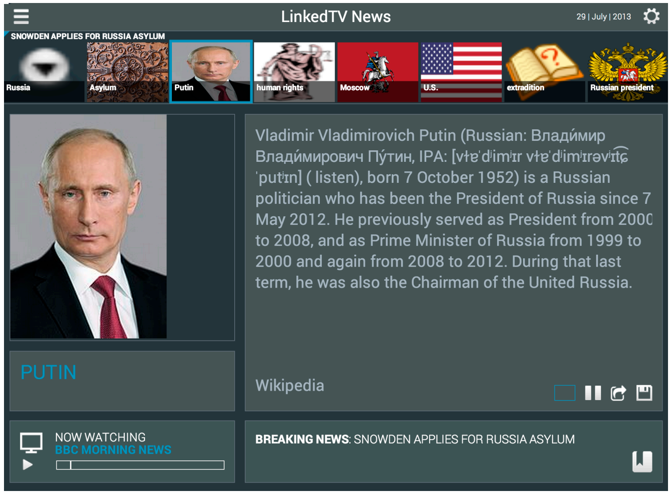
Sheremetyevo

Airport in Moscow

The Use Case: Contextualizing News

МЕЖДУНАРОДНЫЙ АЭРОПОРТ ШЕРЕМЕТЬЕВО





Semantic Snapshot of News (NSS)

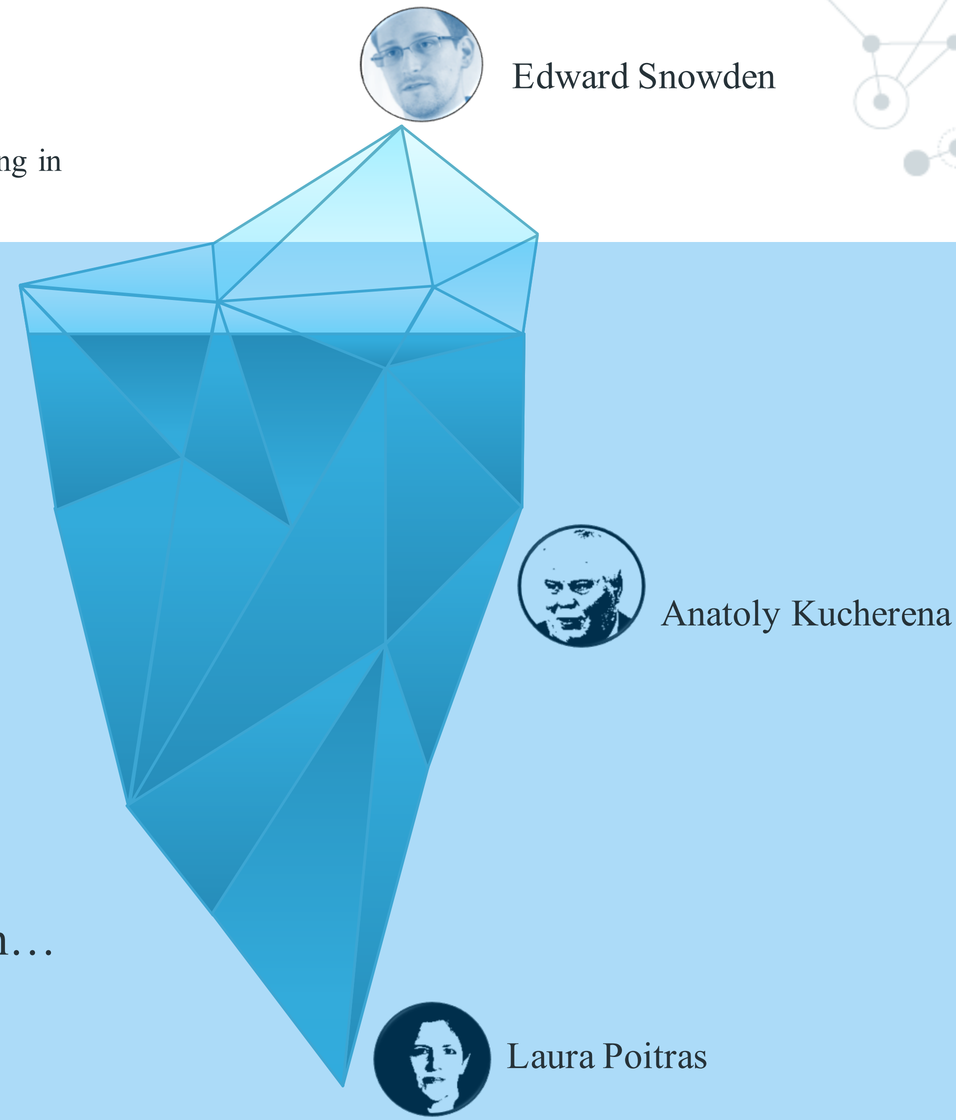
- Definition and Motivation
- A Gold Standard of News Entities

The News Semantic Snapshot (NSS)

Part 2.a

What is on top:

Entities explicitly appearing in the documents



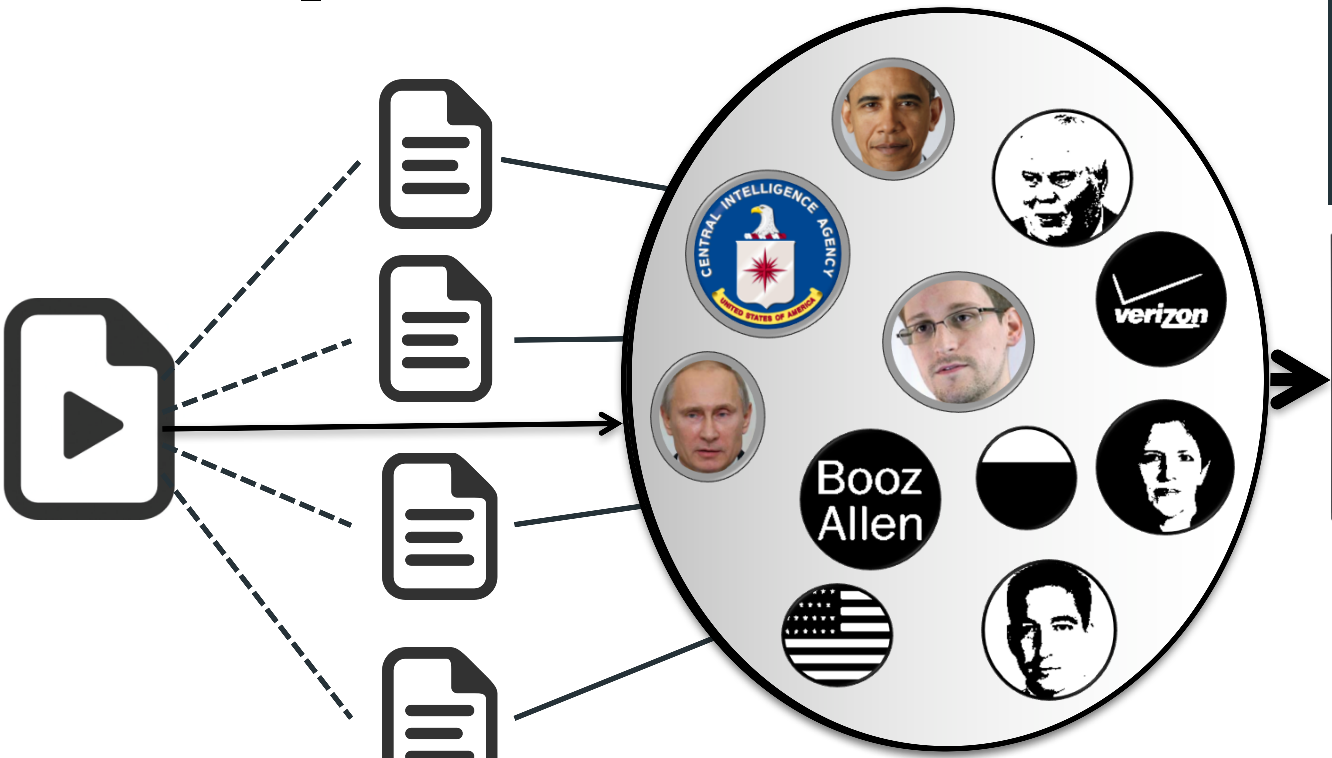
Going deep down…

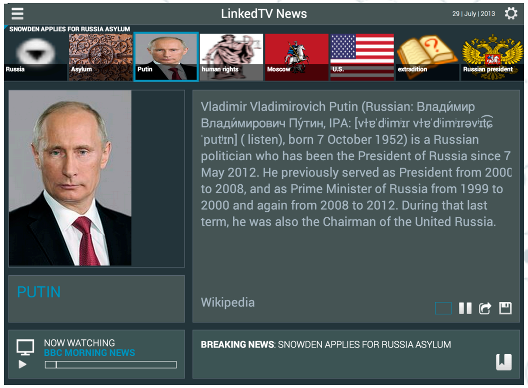
It is always challenging

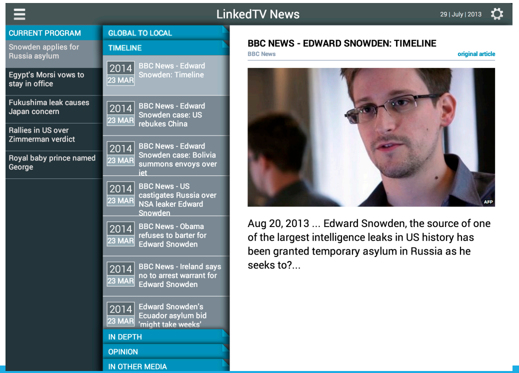
The News Semantic Snapshot (NSS)

[Redondo_ICWE'15]

News Semantic Snapshot
(NSS)







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

(5) Related articles

(1)

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

(1) Video Subtitles

[Romero_TVX’14]

USER SURVEY

Part 2.a

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

Automatically Generating the NSS

2.b

- The Selection problem
- Approaches: frequency-based, multidimensional, concentric
- Experiments and Results

Generating the NSS: General Method

Part 2.b

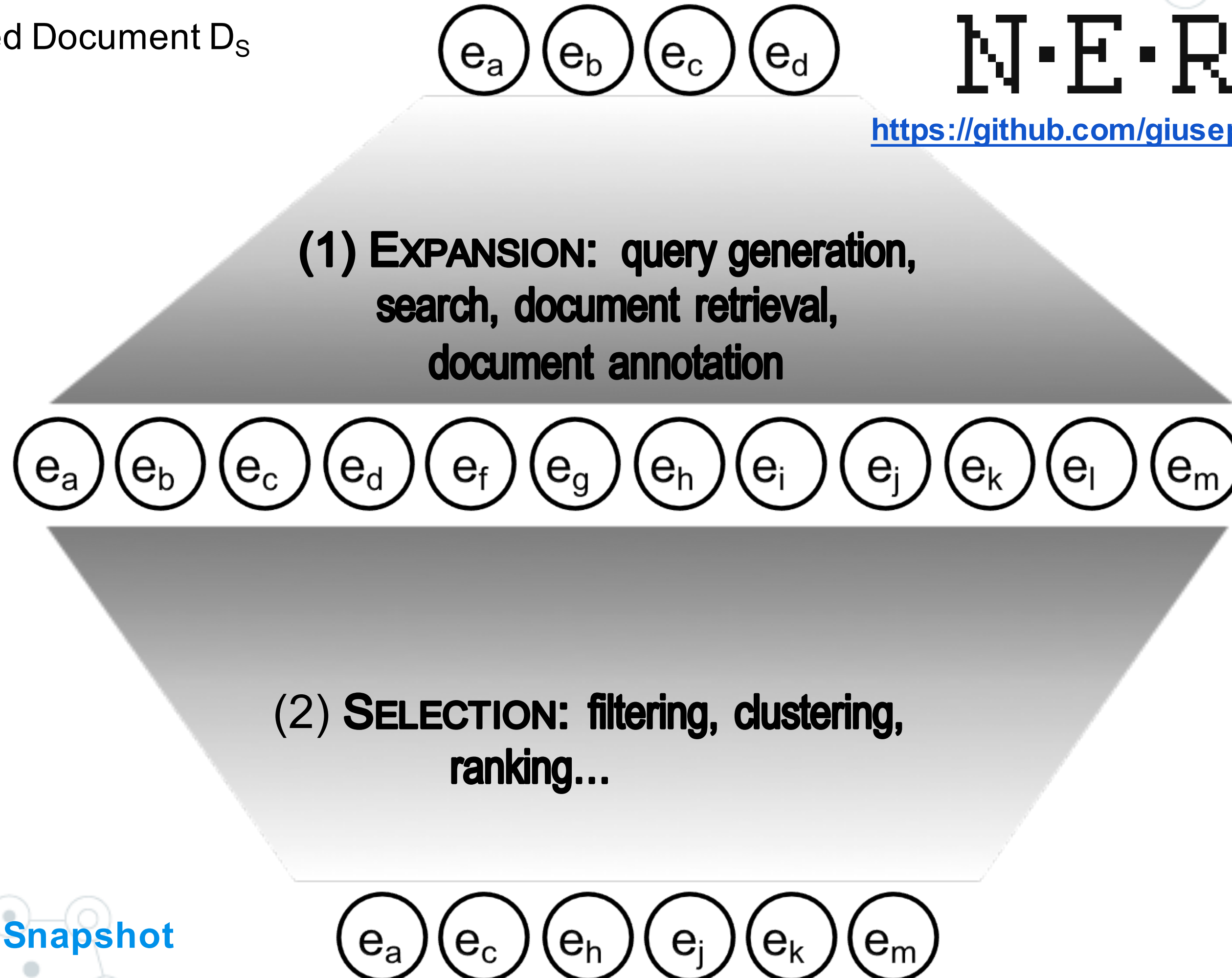
[Redondo_SNOW'14]

a) Entities from Seed Document D_S

https://github.com/giusepperizzo/nerdml

b) **Expanded Entities**

c) **News Semantic Snapshot**

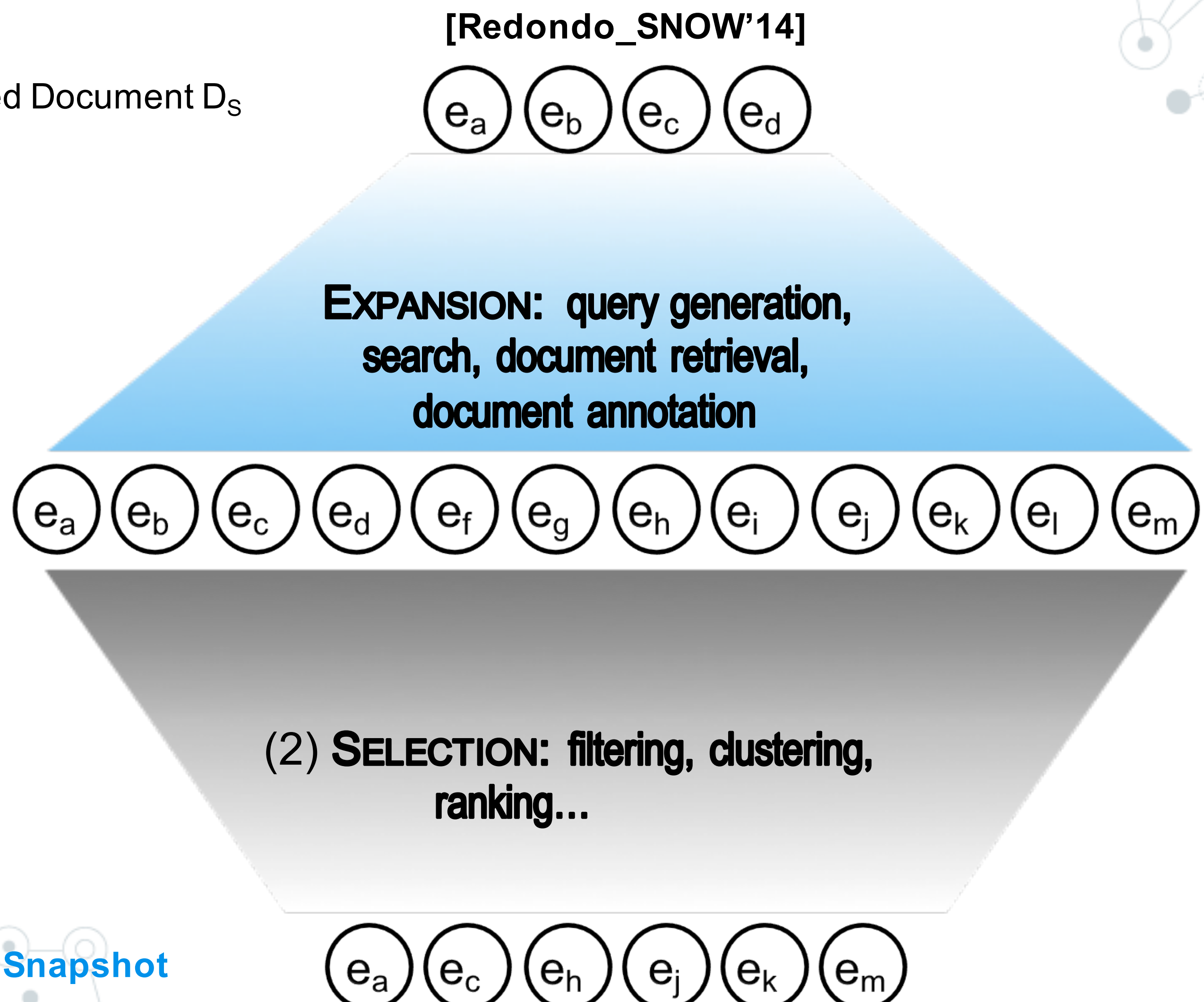


Generating the NSS: Entity Expansion

Part 2.b

a) Entities from Seed Document D_S

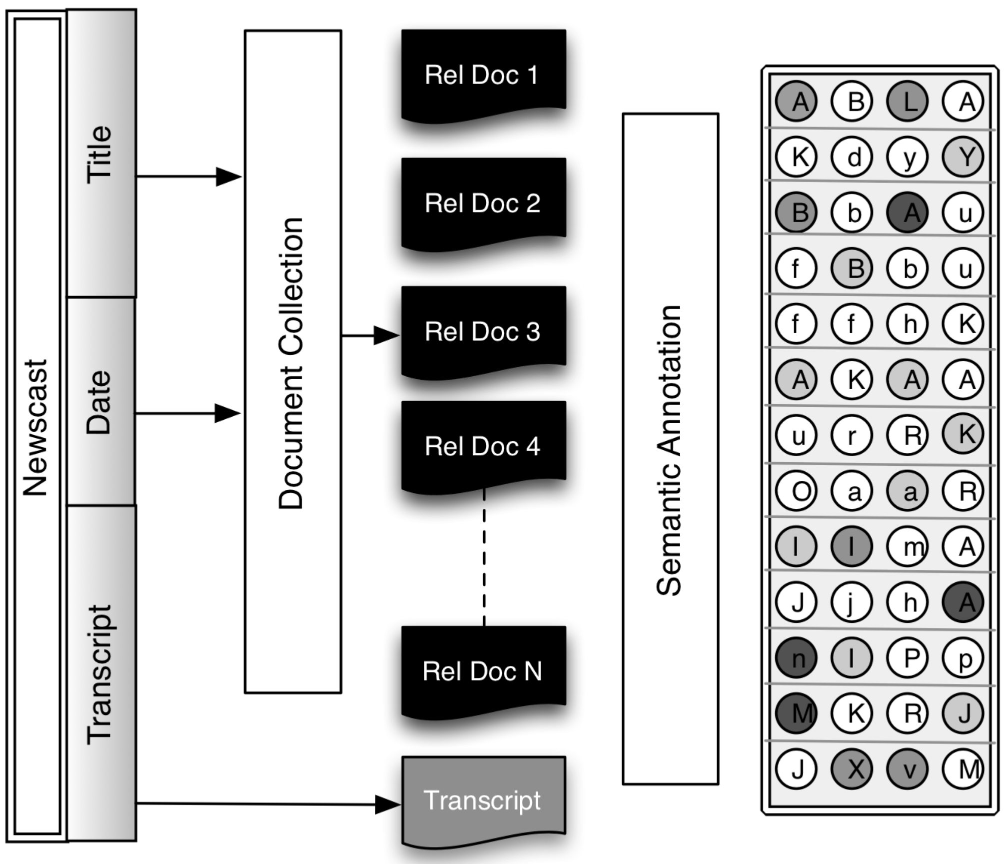
b) **Expanded Entities**



c) News Semantic Snapshot

Generating the NSS: Expansion's Settings

[Redondo_ICWE'15]



Google Custom Search Engine (CSE)

Parameters:

Query

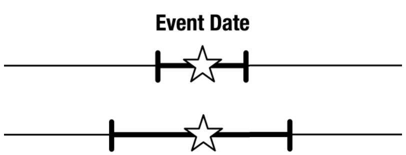
- Title
- 5 W's over Subtitles Entities

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**

schema.org

Generating the NSS: Expansion's Settings

[Redondo_SNOW'14]

a) Entities D_S

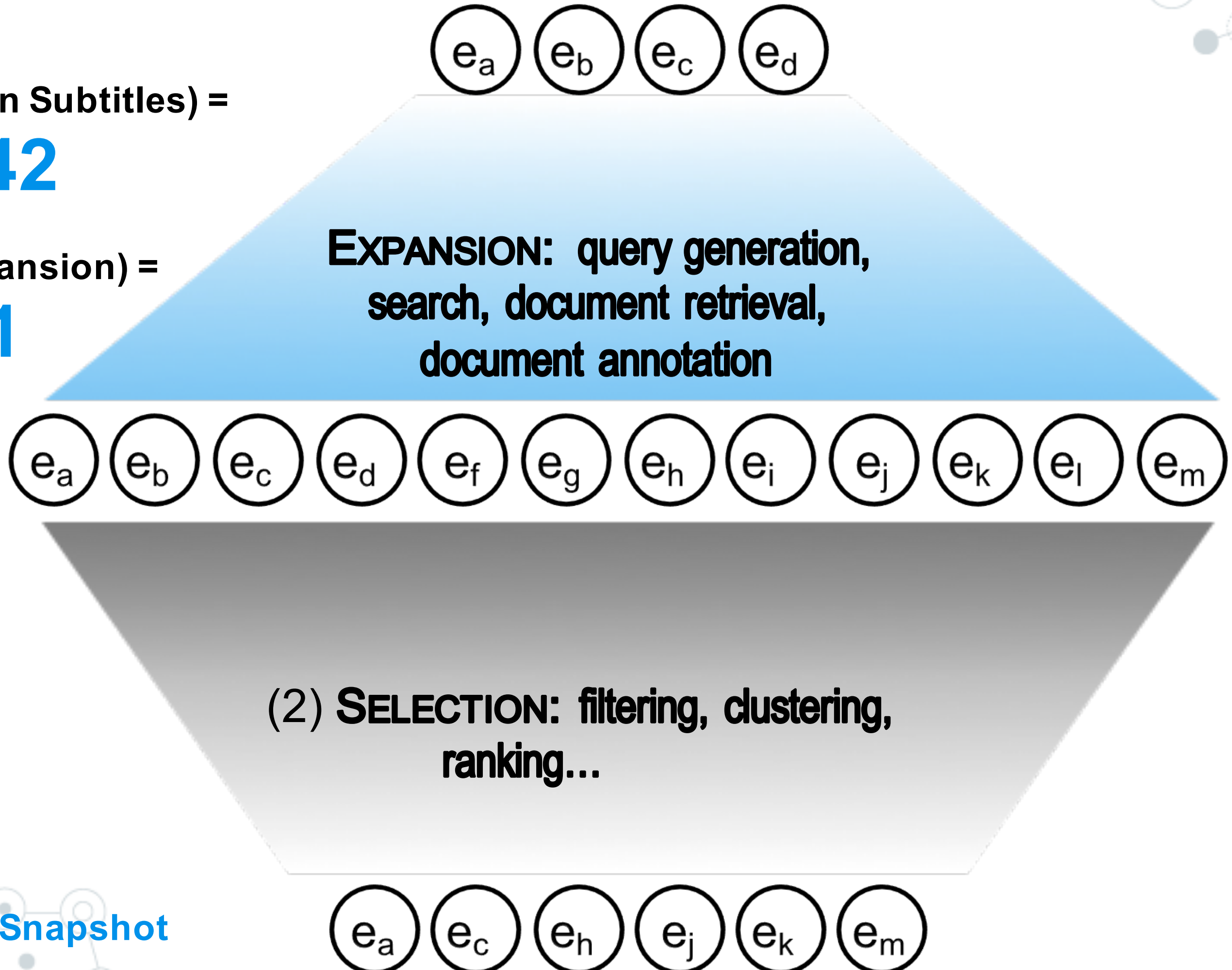
Recall (NER on Subtitles) =

0.42

Recall (E. Expansion) =

0.91

b) Expanded Entities



c) News Semantic Snapshot

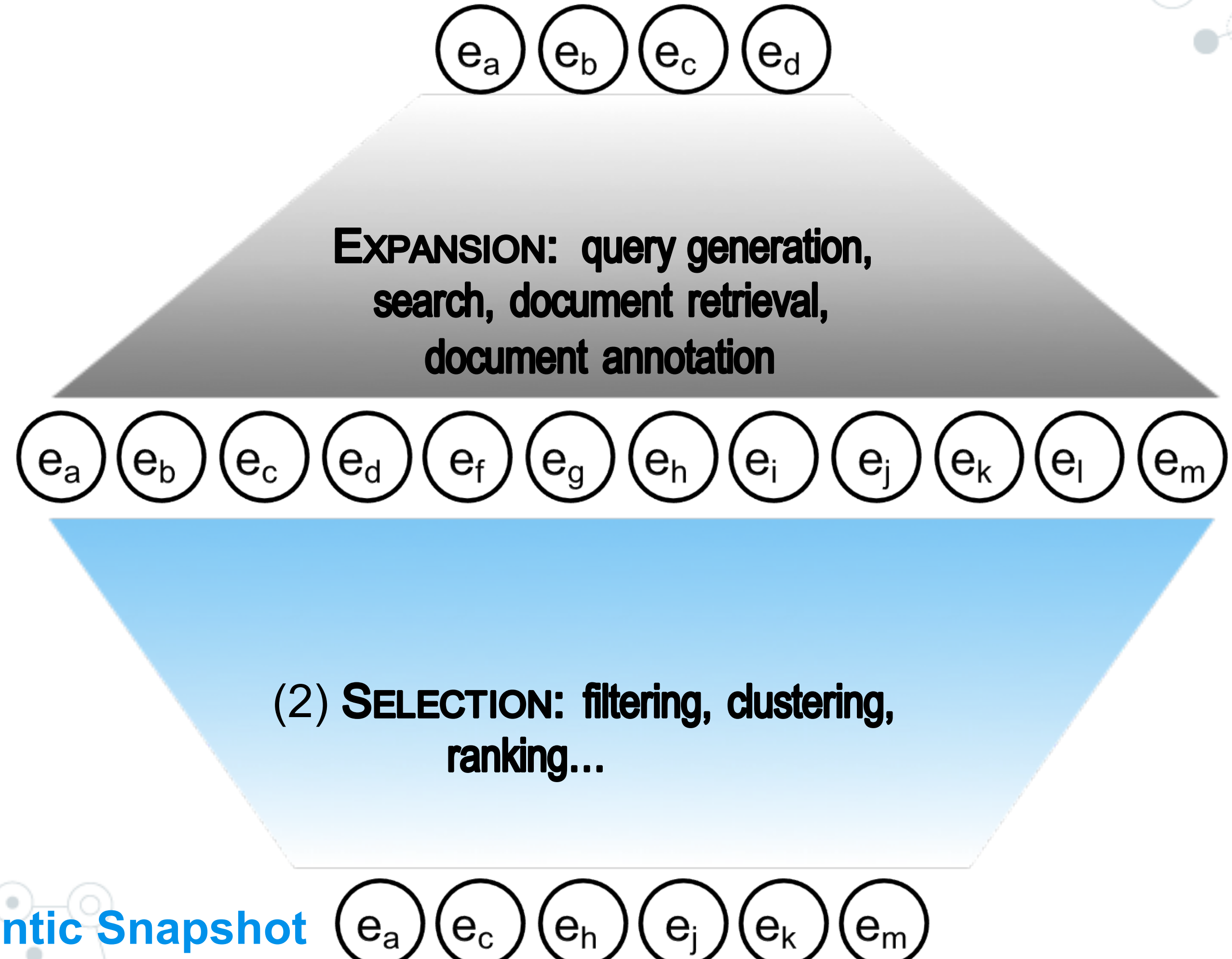
Generating the NSS: Selection

[Redondo_SNOW'14]

a) Entities D_S

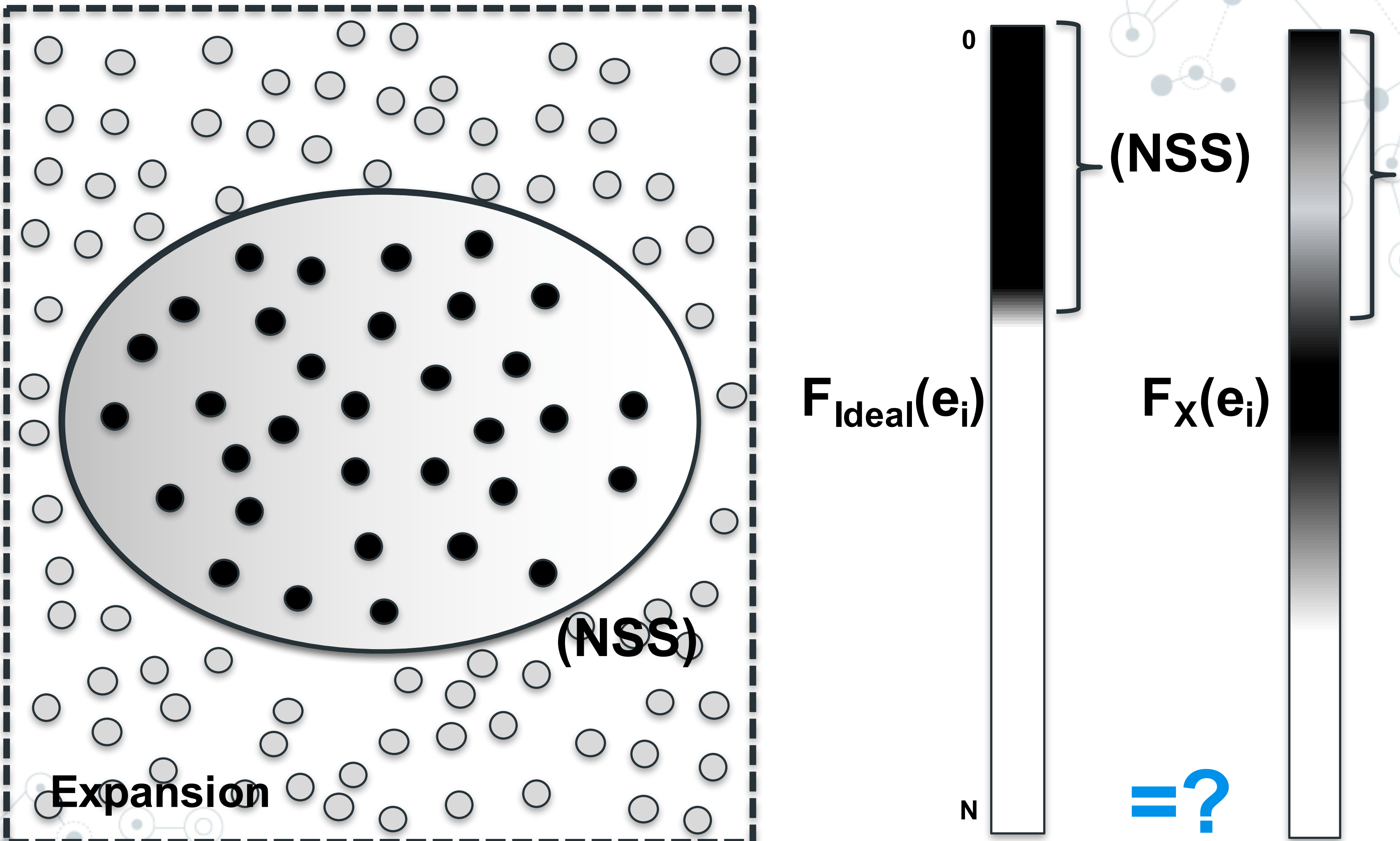
b) Expanded Entities

c) News Semantic Snapshot



Generating the NSS: The Selection problem

Part 2.b



Generating the NSS: Measures

Part 2.b

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:

- Compromise between: Recall and NSS size

$$Com\,(R, f, v) = |min(NSS \in Res)| \mid f(NSS) \geq v$$

Generating the NSS: Compactness Example

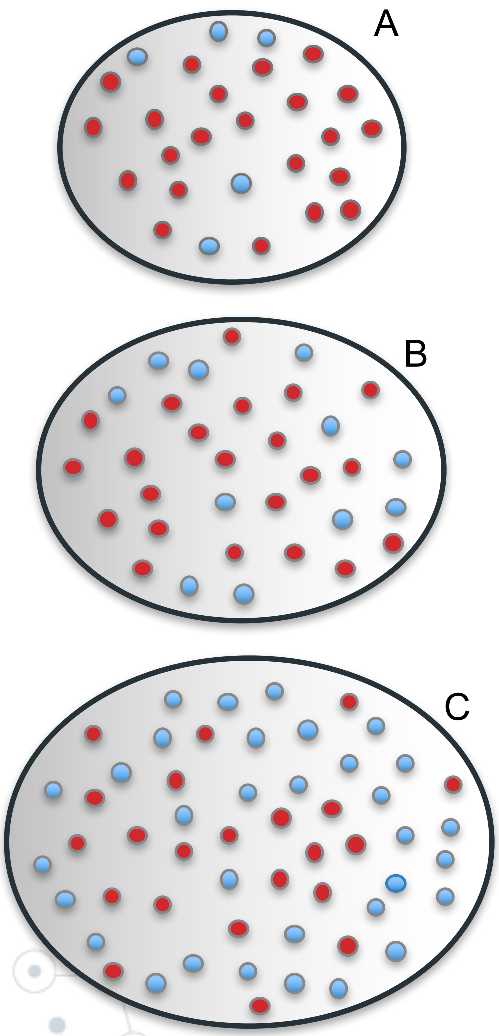
Part 2.b

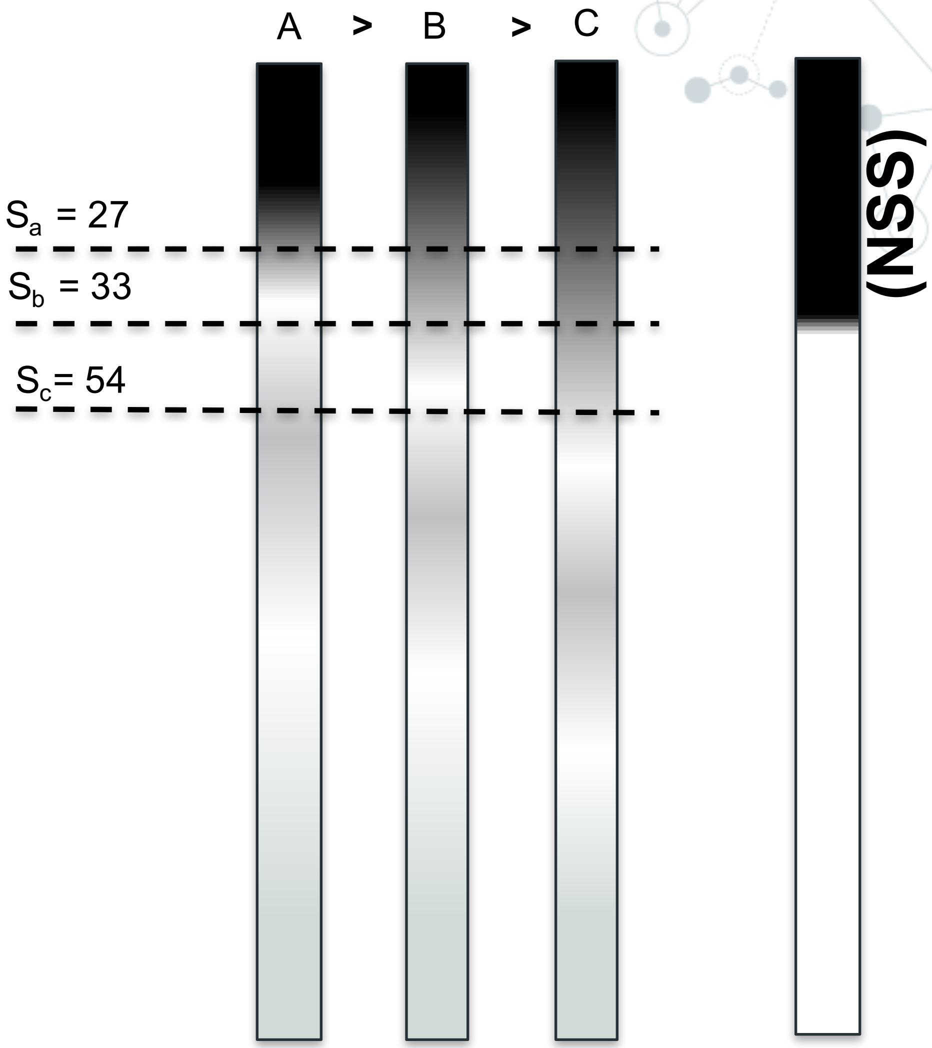
Recall: **22/33 = 0.66**

S_a = 27

S_b = 33

S_c = 54





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

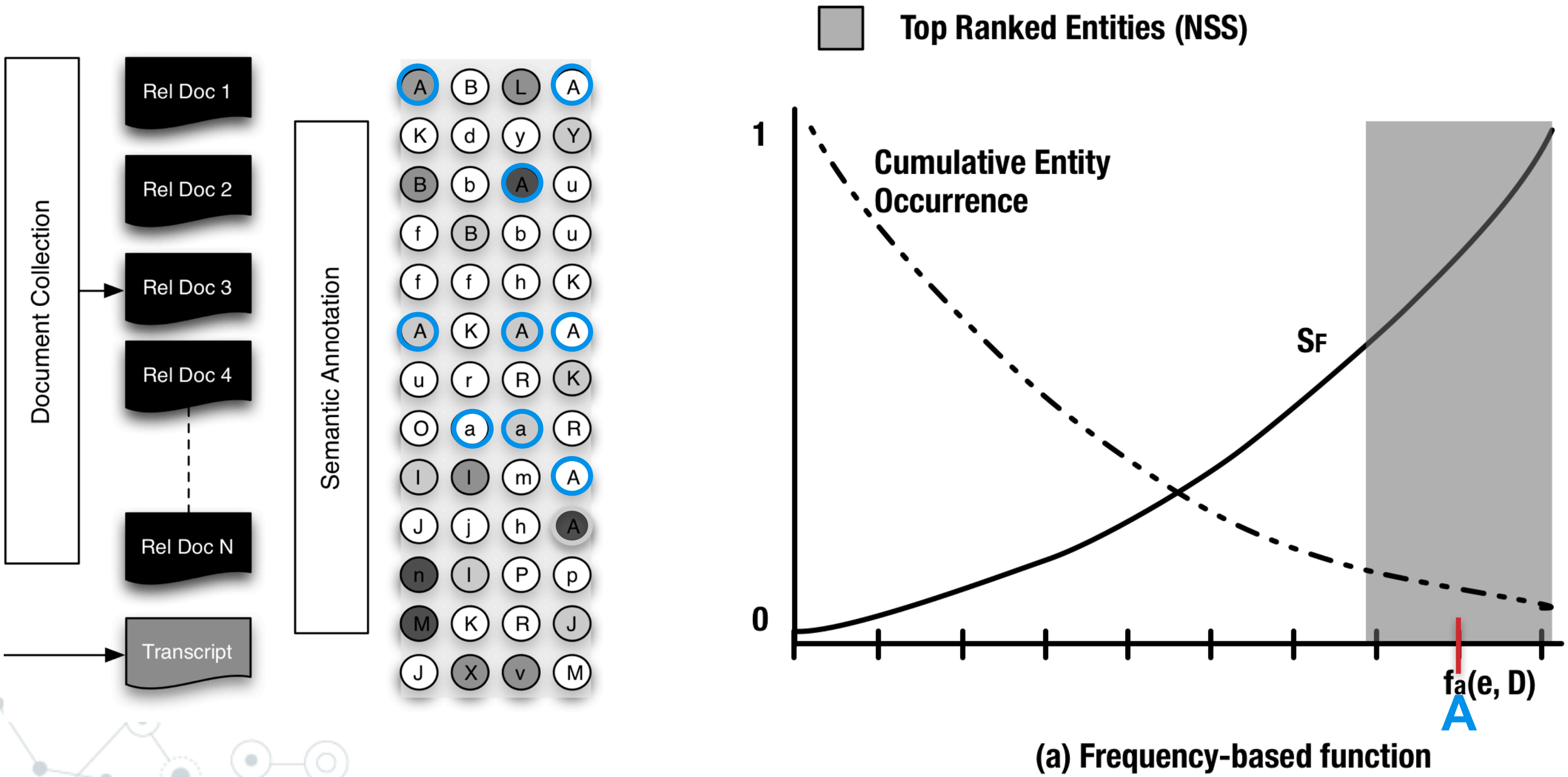
3 Concentric Based Approach [Redondo_KCAP'15A]

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

Generating the NSS: (1) Frequency-Based

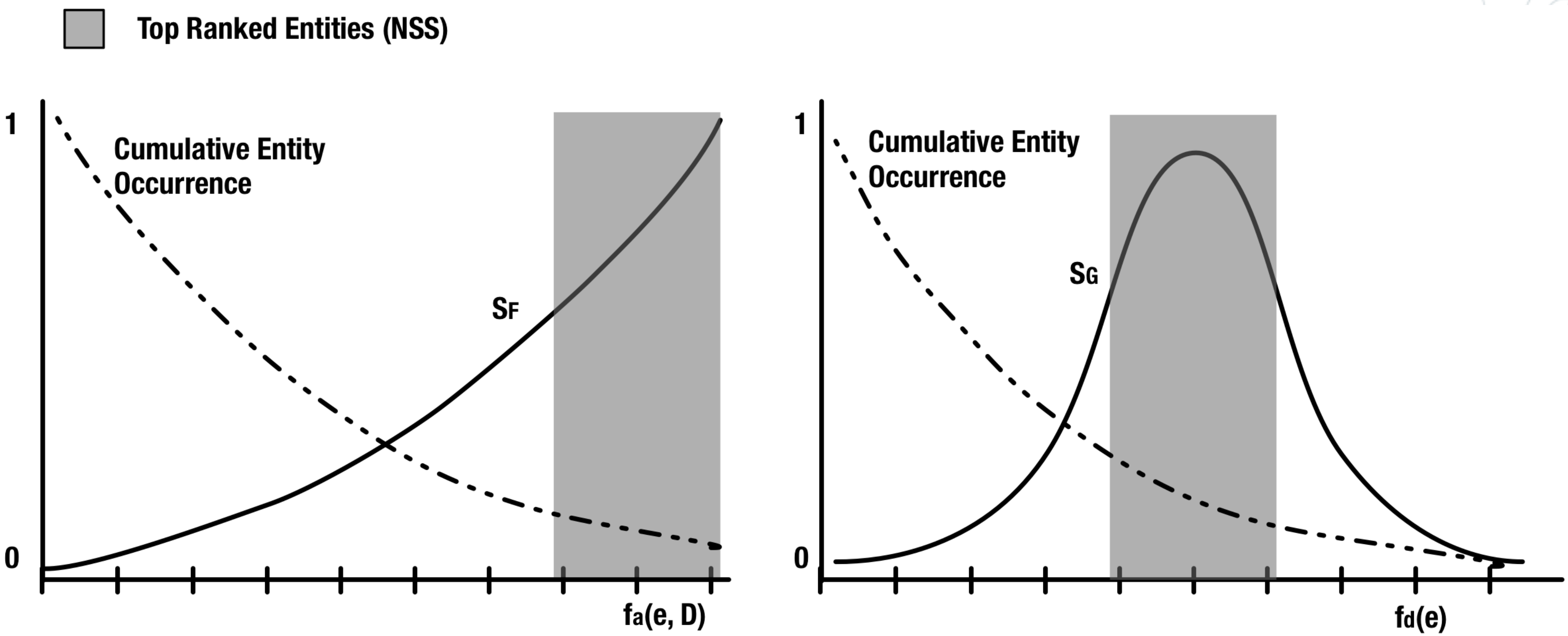
Part 2.b

[Redondo_SNOW'14]



Generating the NSS: (2) **Multidimensional**

[Redondo_ICWE2015]



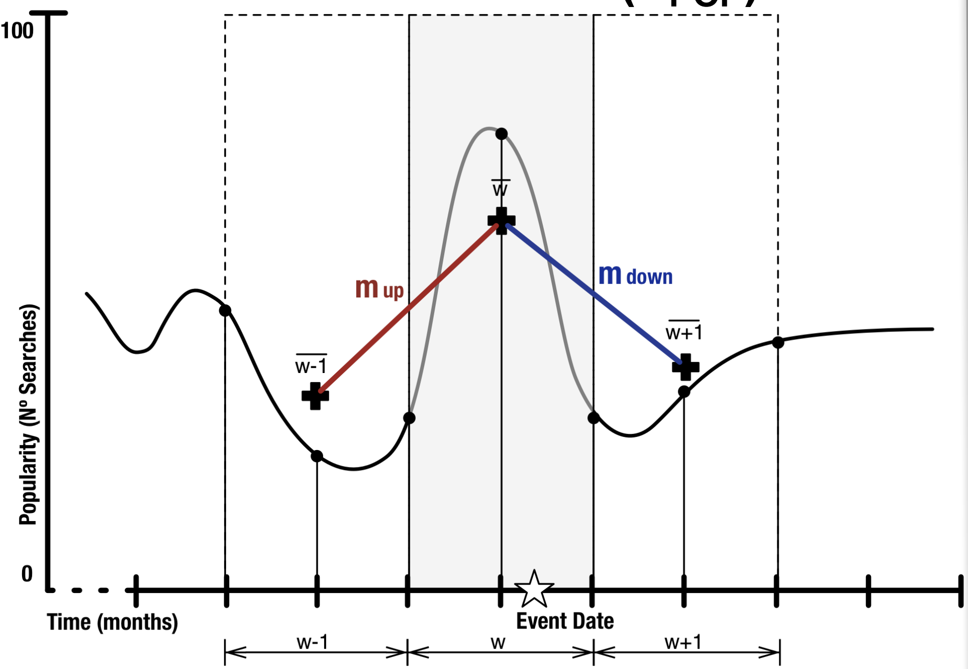
(a) Frequency-based function

(b) Gaussian-based function

Generating the NSS: (2) Multidimensional

Part 2.b

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:

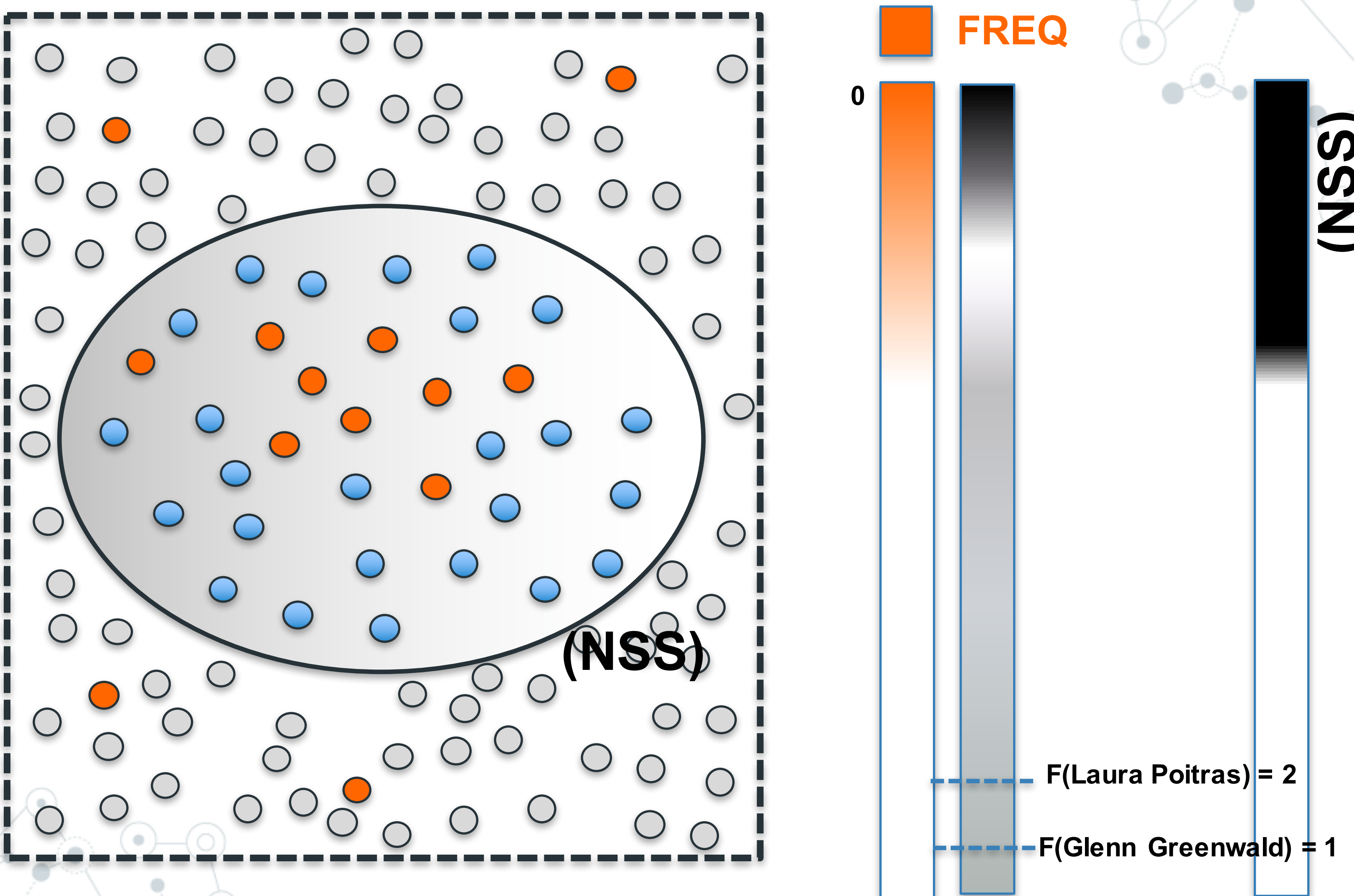
- [Location, = 0.43]
- [Person, = 0.78]
- [Organization, = 0.95]
- [$f_{doc}(e_i)$< 2 , = 0.0]

Experiment 1: Frequency VS Multidimensional

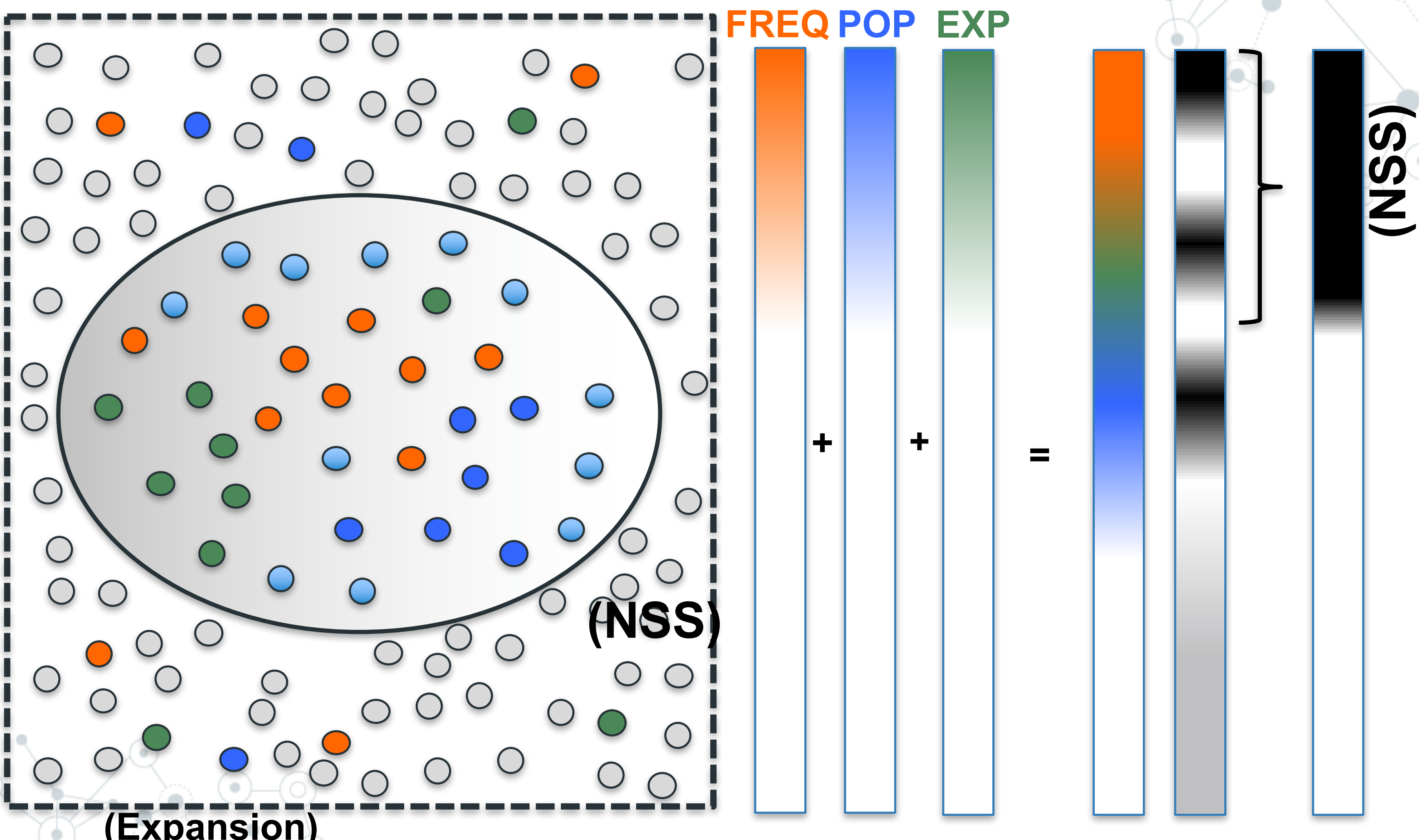
20 x 4 x 4 =
320 formulas

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



Experiment 1: Frequency VS Multidimensional



Experiment 2: Multidimensional ++

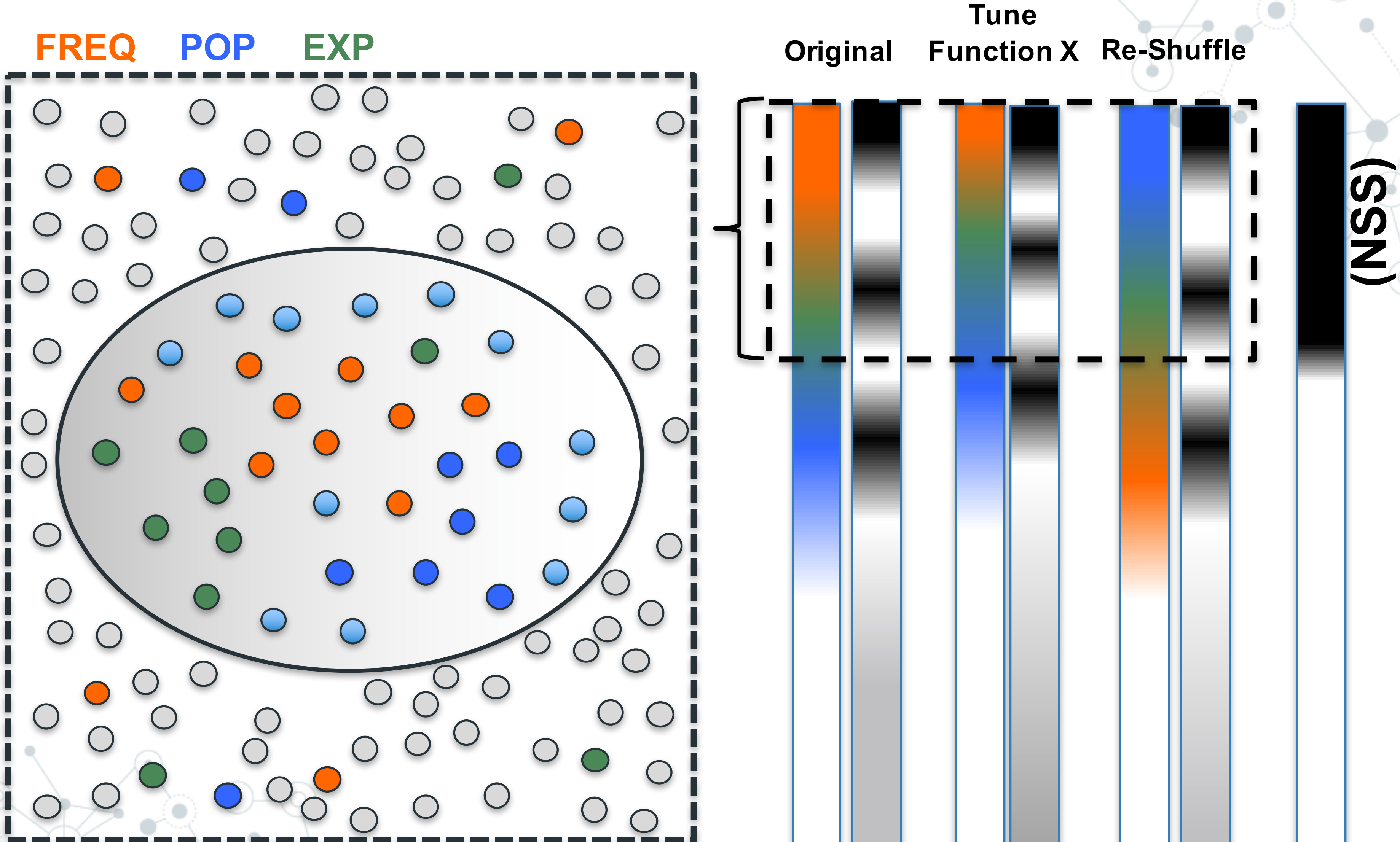
NMDCG @ 10:

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 -

Part 2.b

Experiment 2: Multidimensional ++



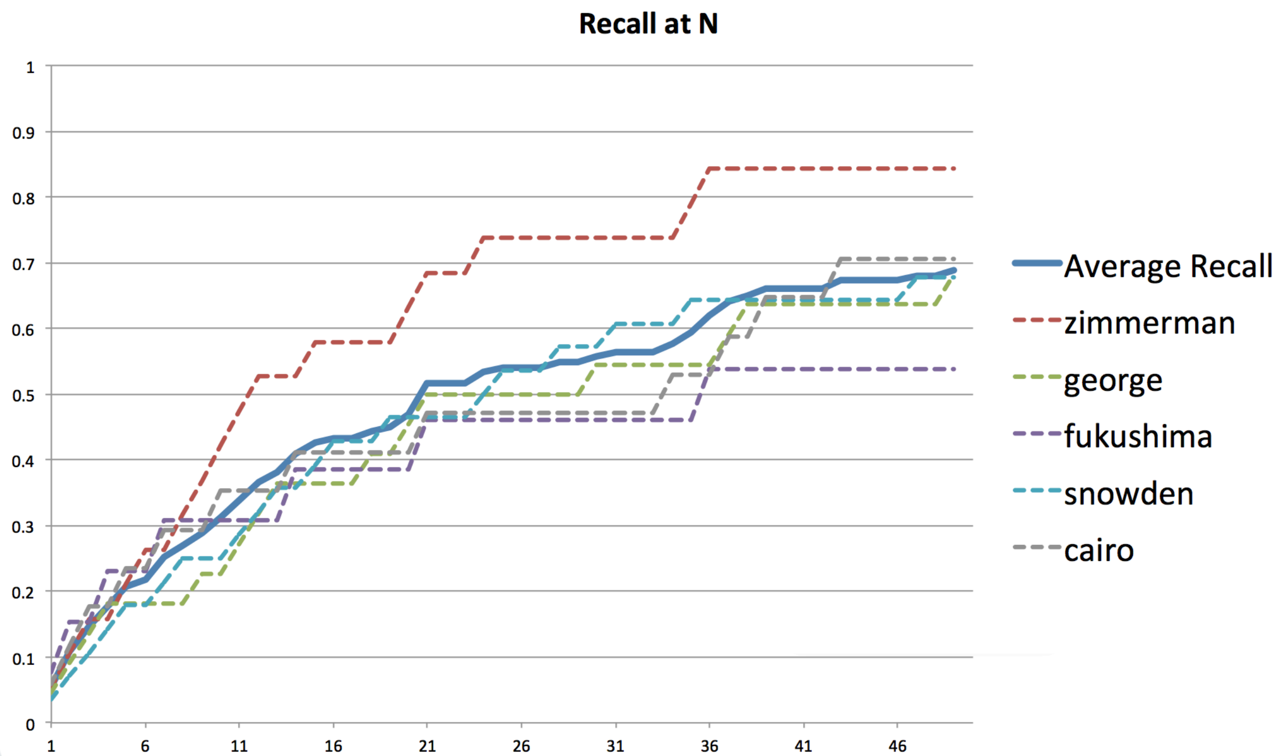
Re-thinking the problem: measures

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



Part 2.b

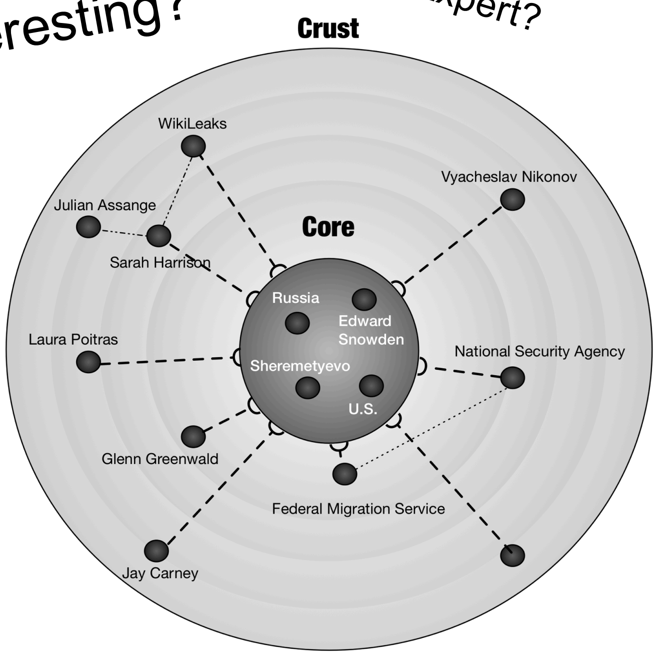
Re-thinking the problem: dimensions

Highly Popular? Explicative?

Informative? Unexpected?

Suggested by Expert?

Interesting?



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

Generating the NSS: (3) Concentric Approach

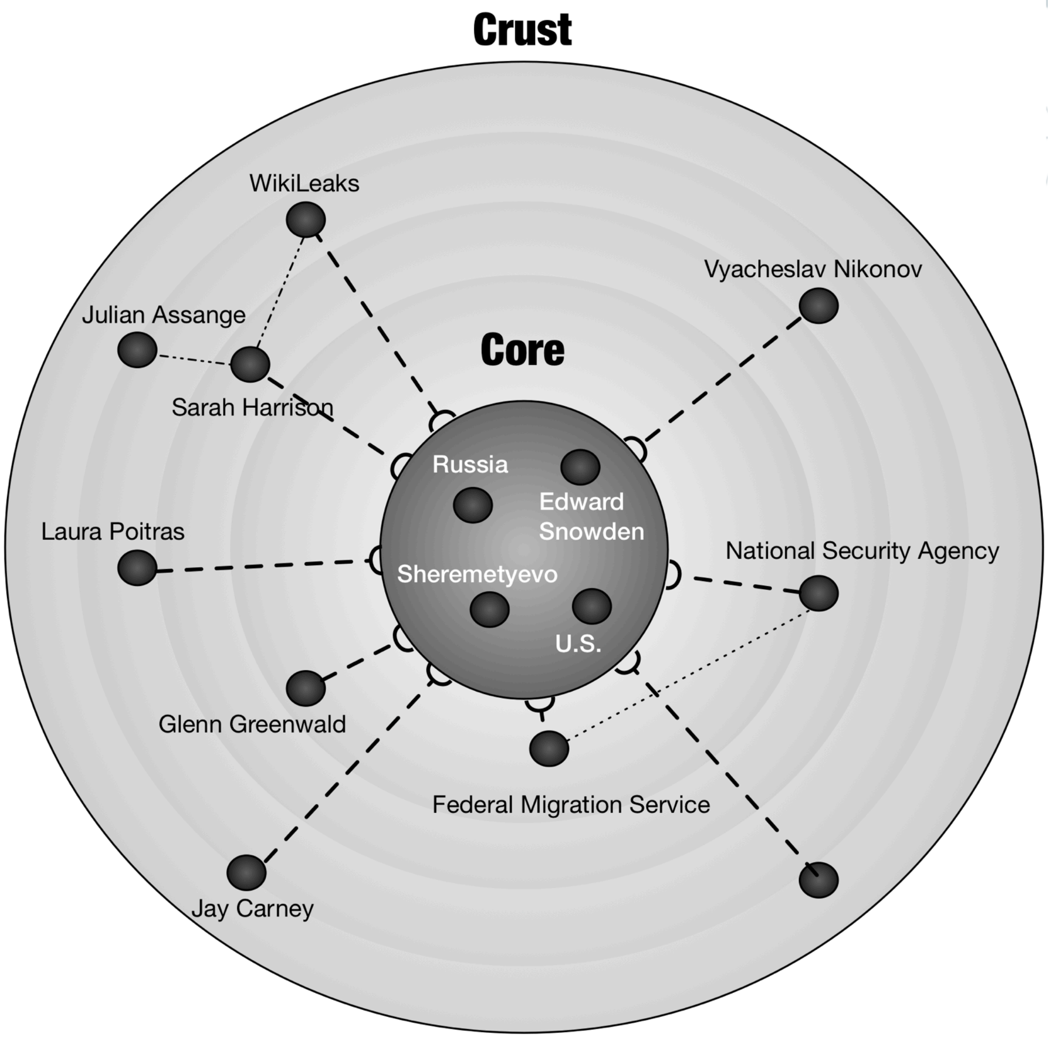
[Redondo_KCAP2015A]

Core

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

Crust

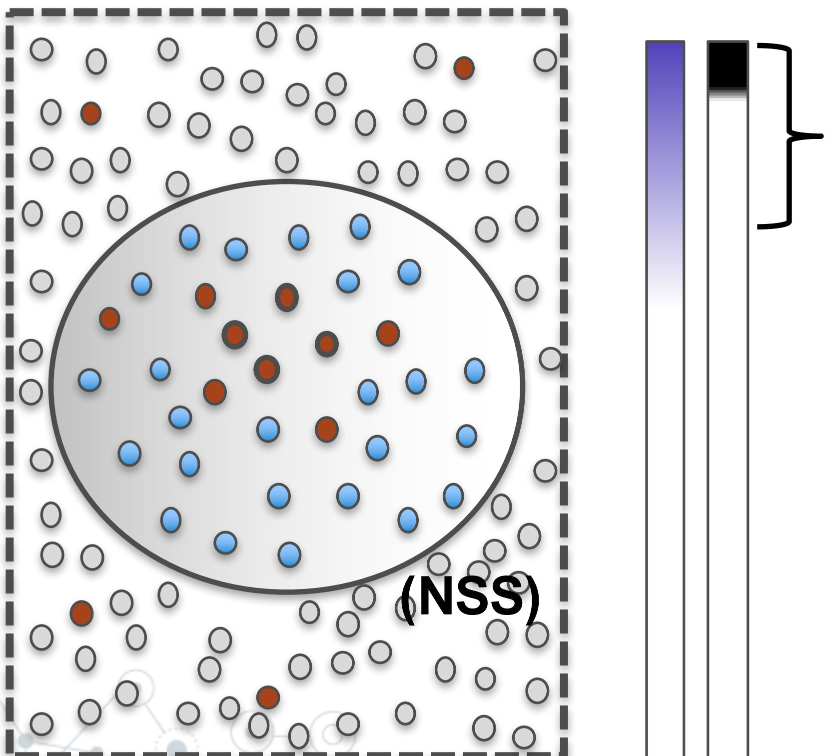
- Attached to the Core via semantic relations
- Agnostic to relevancy nature: informativeness, interestingness, etc.



Generating the NSS: (3) Core Creation

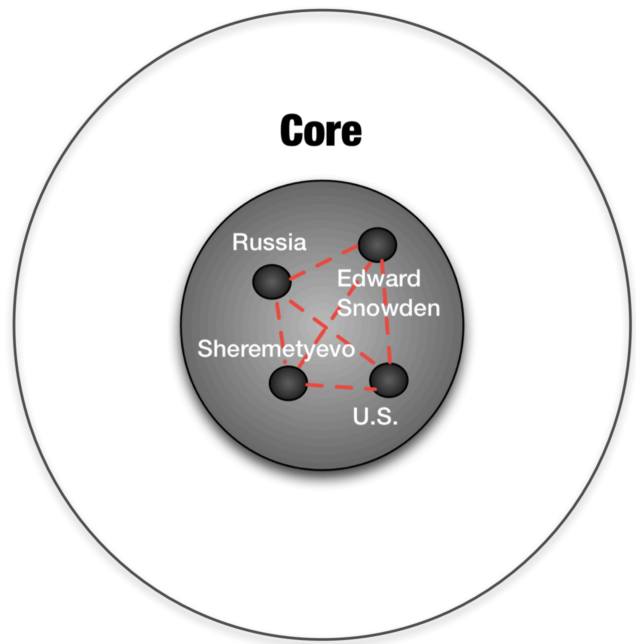
a) Spot representative entities: Frequency Dimension

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



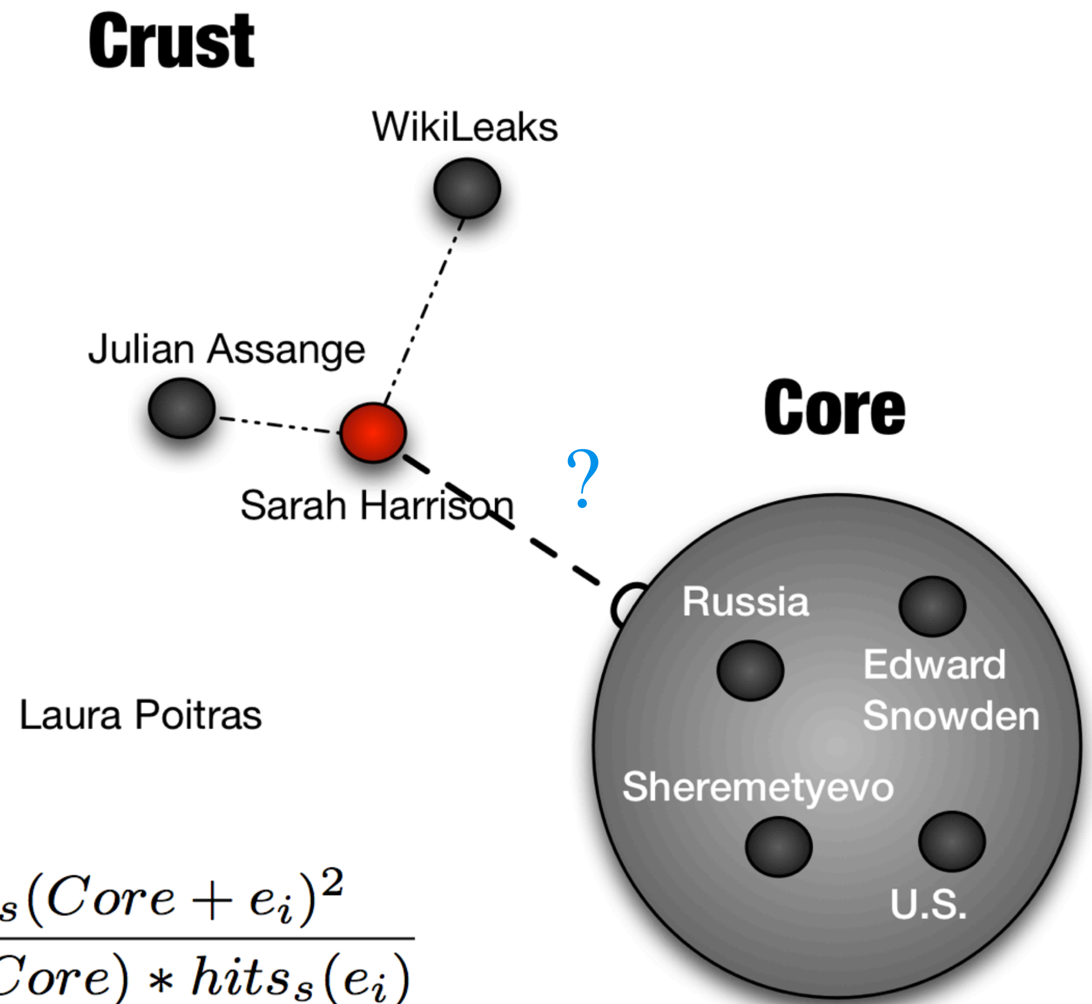
b) Cohesiveness (DBpedia)

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



Generating the NSS: (3) Crust Creation

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



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

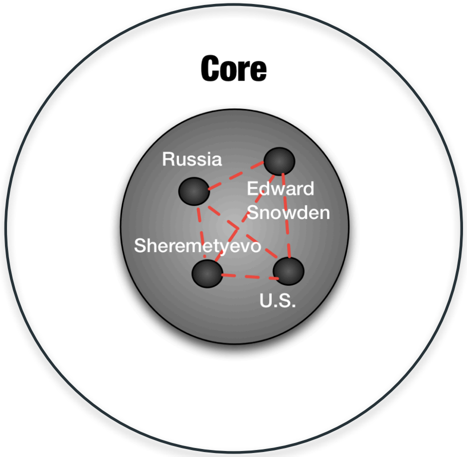
Part 2.b

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

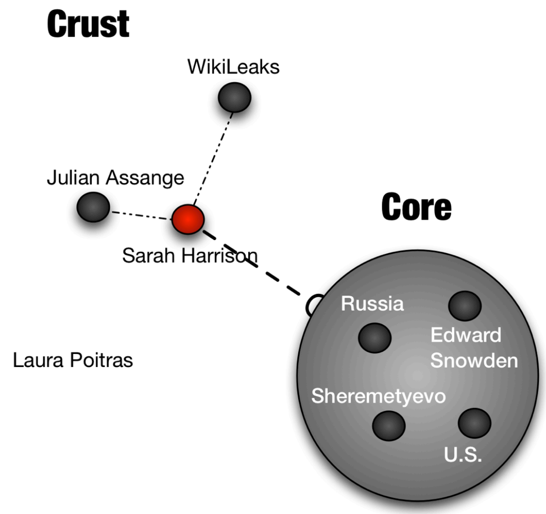
Experiment 3: Multidimensional VS Concentric

Part 2.b

Concentric Crust:

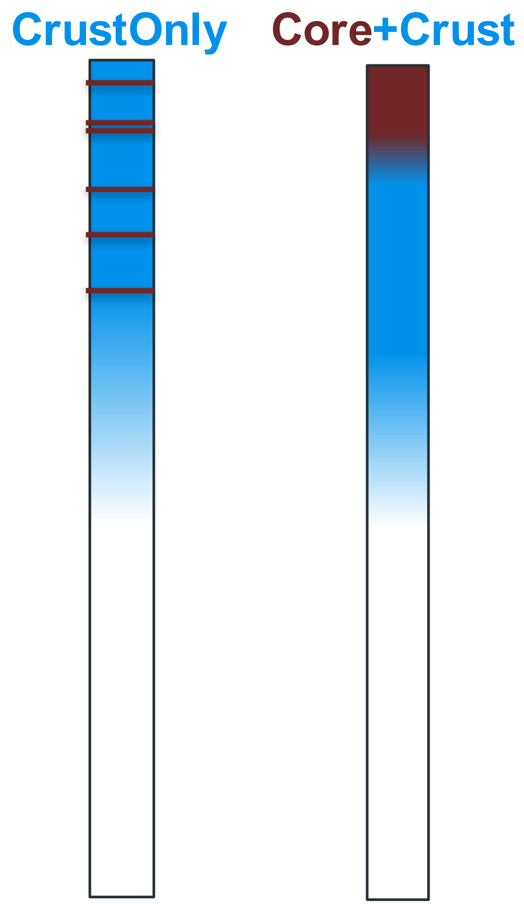
1. Candidates for CRUST generation:
   - Ex1: 1° ICWE2015 by R*(50): L2+Google, F3 1W, Gauss+ POP
   - Ex2: 2° ICWE 2015 by R*(50): L2+Google, F3 1W, Freq + POP

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



Experiment 3: Multidimensional VS Concentric

Combining CORE and CRUST:



Part 2.b

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)

Part 2.b

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

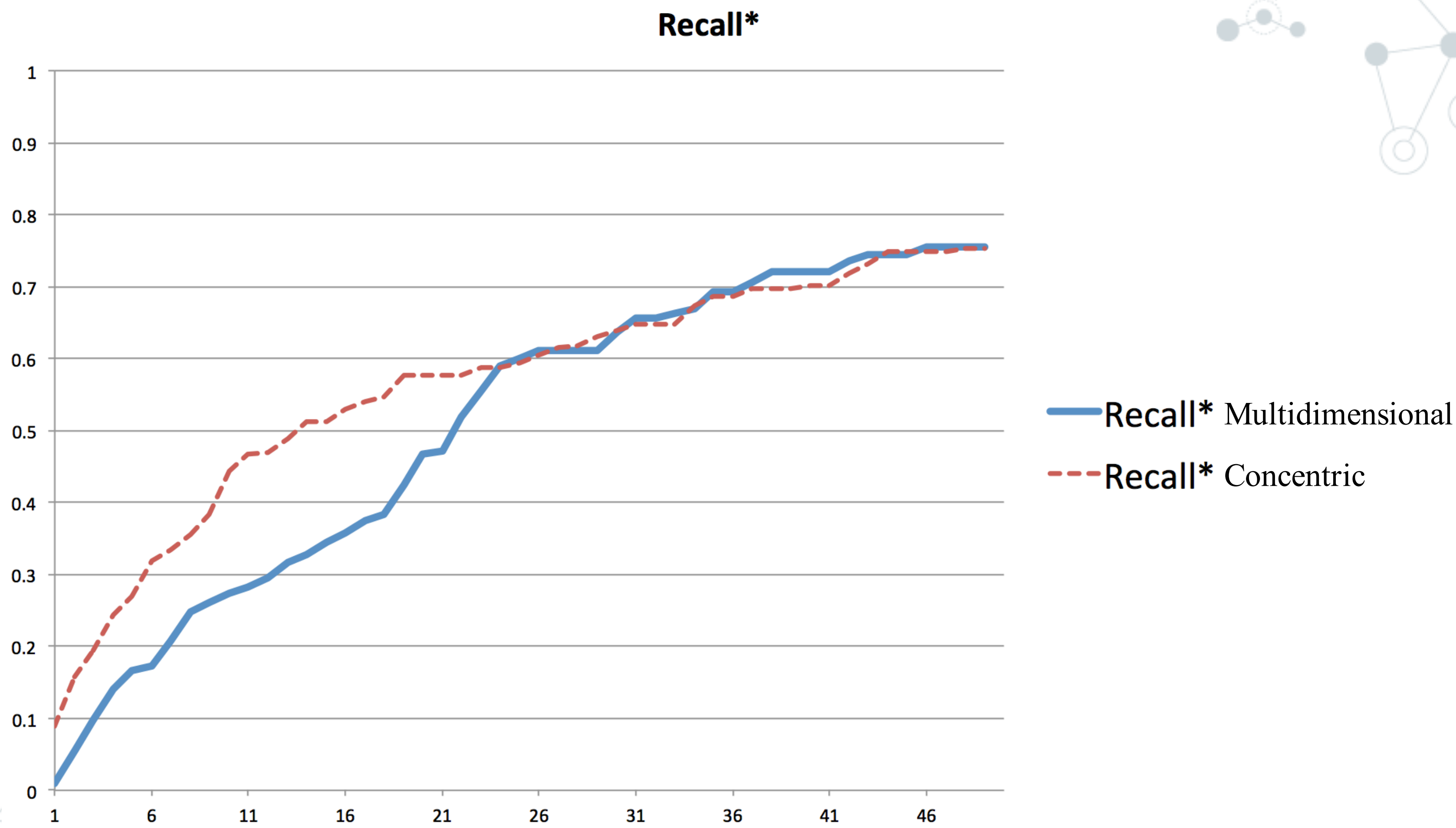
n=22

NSS
Gold Standard

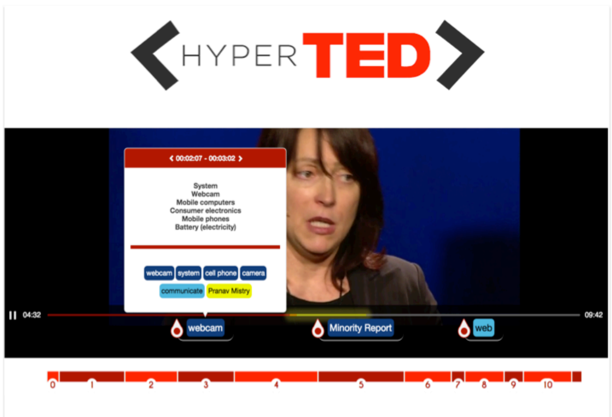
Fukushima Disaster 2013

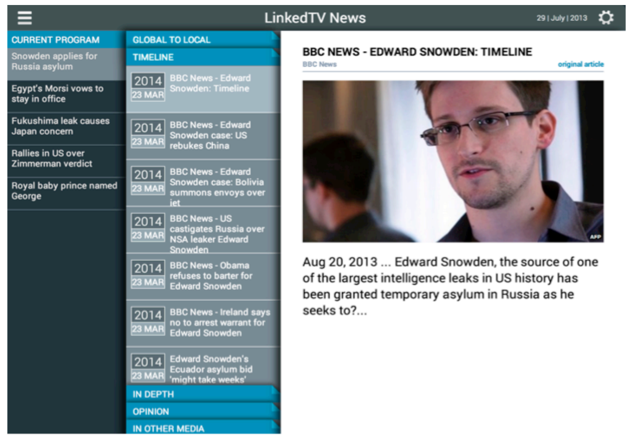
Experiment 3: Multidimensional VS Concentric

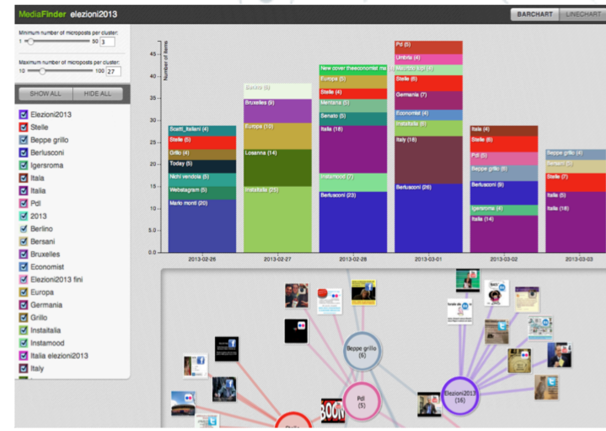
Part 2.b



NSS Consumption: News Prototypes







… short summaries, previews, hotspots …

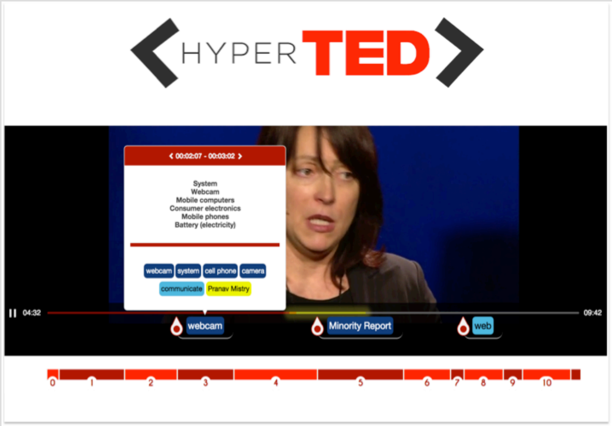
… second screen apps, slideshows, info-boxes …

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

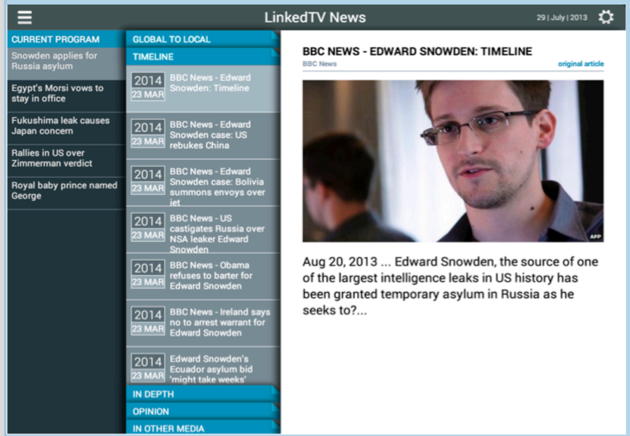
NSS Consumption: Consumptions Phases

Part 2.c

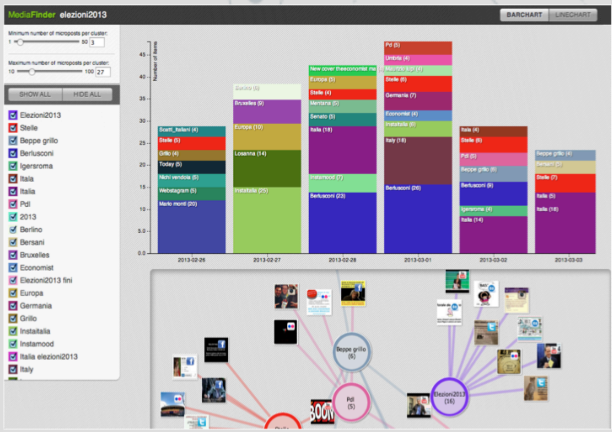
The Before

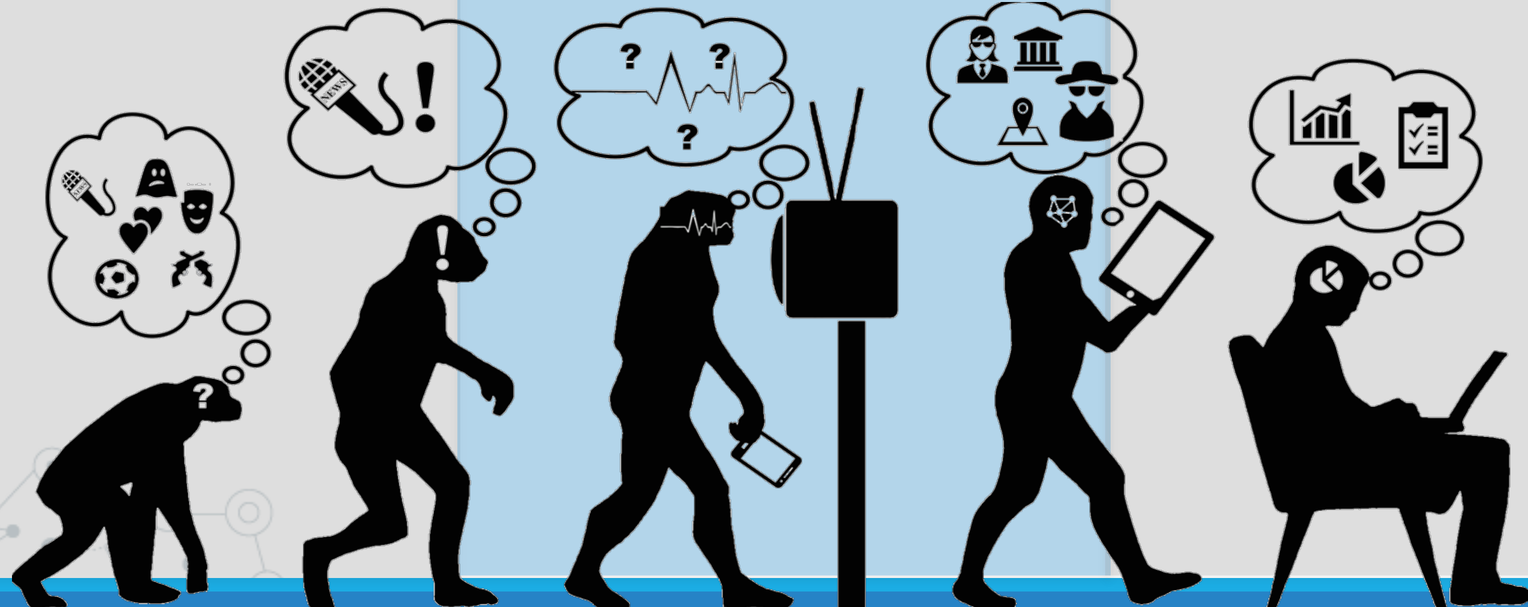


The During



The After

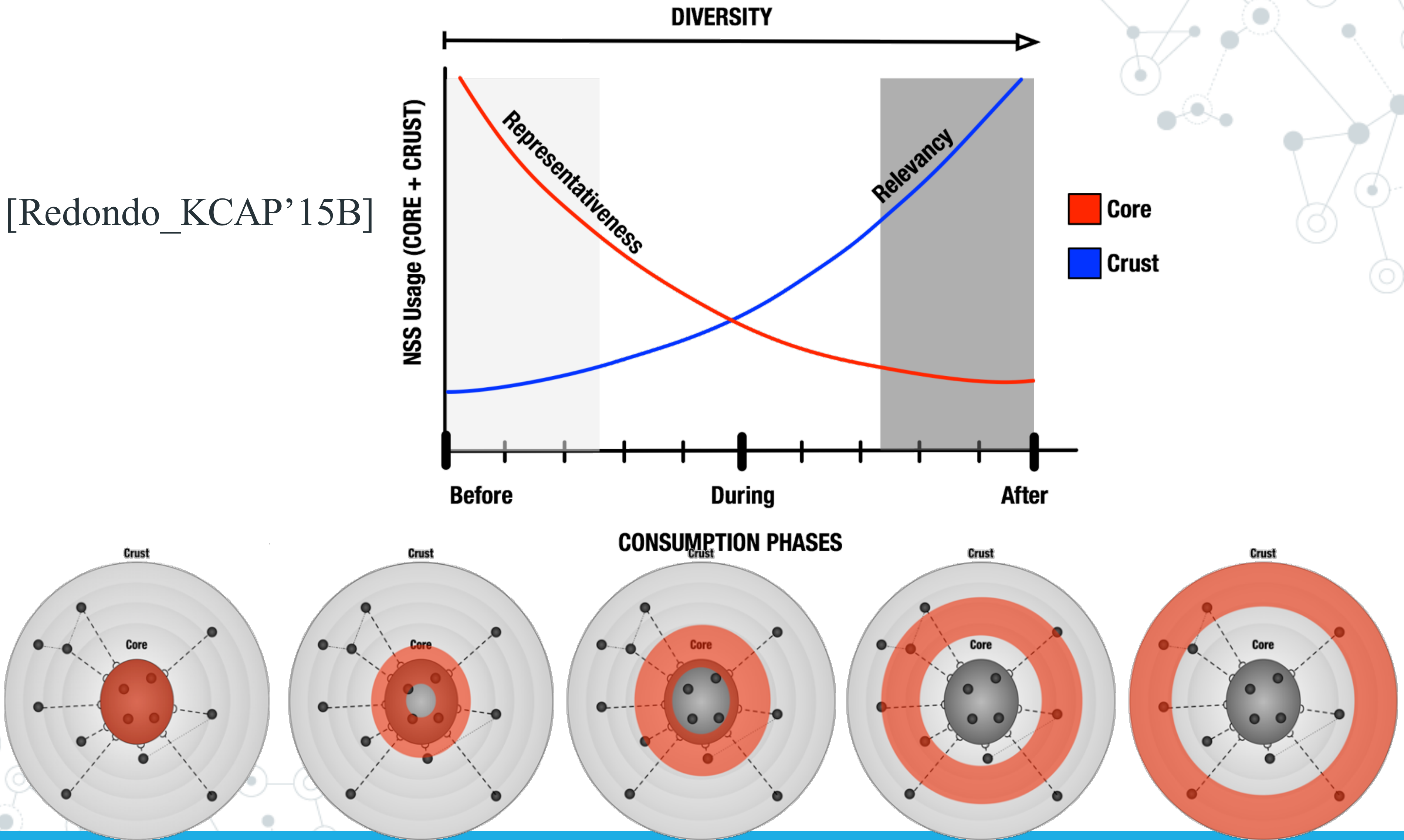




NSS Consumption: Phases VS Layers

Part 2.c

[Redondo_KCAP'15B]



What's Next?

- Future Challenges

Conclusions

Q3: Is it possible to automatically contextualize news stories with background information so they can be effectively interpreted by humans and machines?

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:

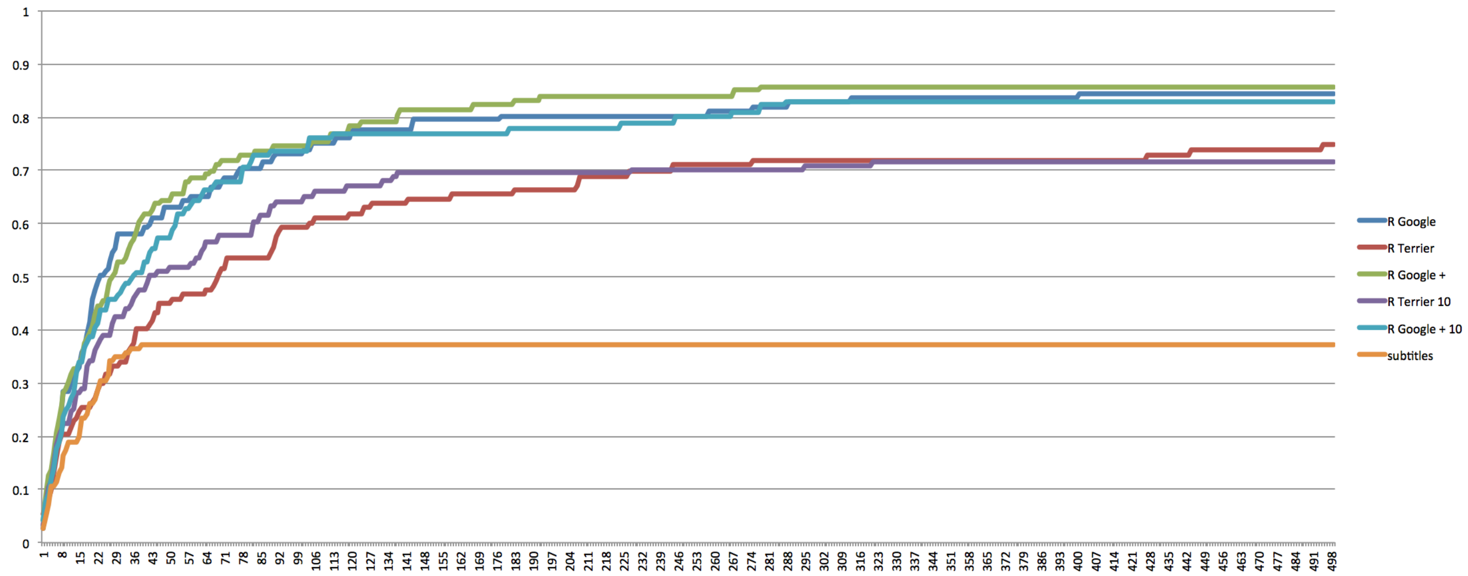
- Better Compactness: 36.9% over BAS01 (similar recall, smaller size)
- Core/Crust brings up relevant entities without having to deal with fuzzy dimensions

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

Future Work

- Applying those those IR and KR techniques to :
  - Bigger Corpora: **Big Data**
  - Different Domains: **Scientific Papers, Libraries, etc.**

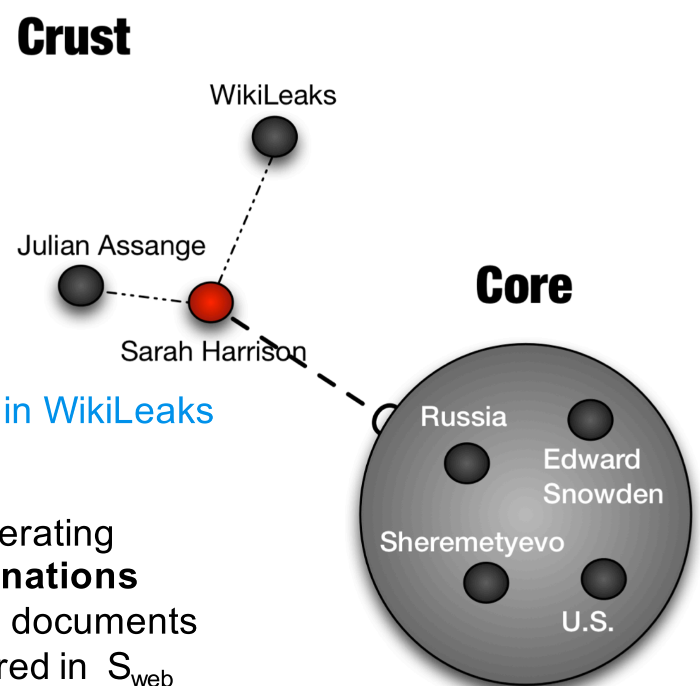
In parallel, stop depending on “big players” for retrieving knowledge during the expansion phase (Terrier VS Google experiments)



Future Work:

Explanations :

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



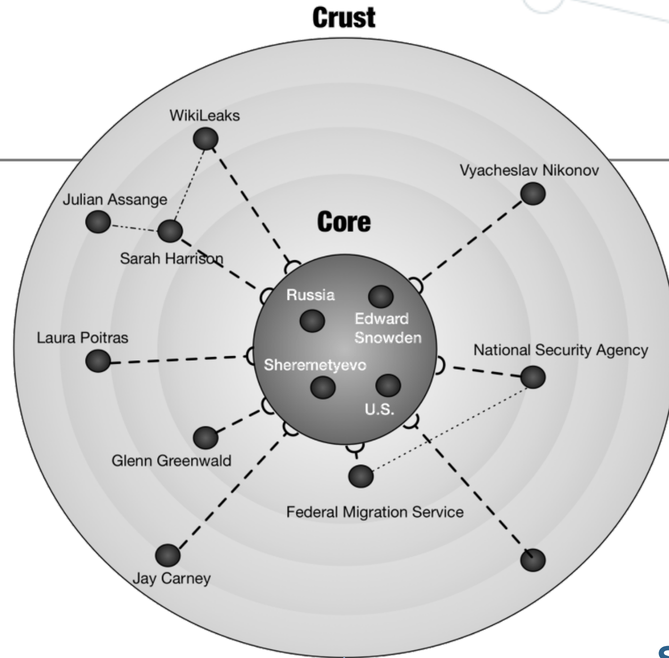
Editor in WikiLeaks

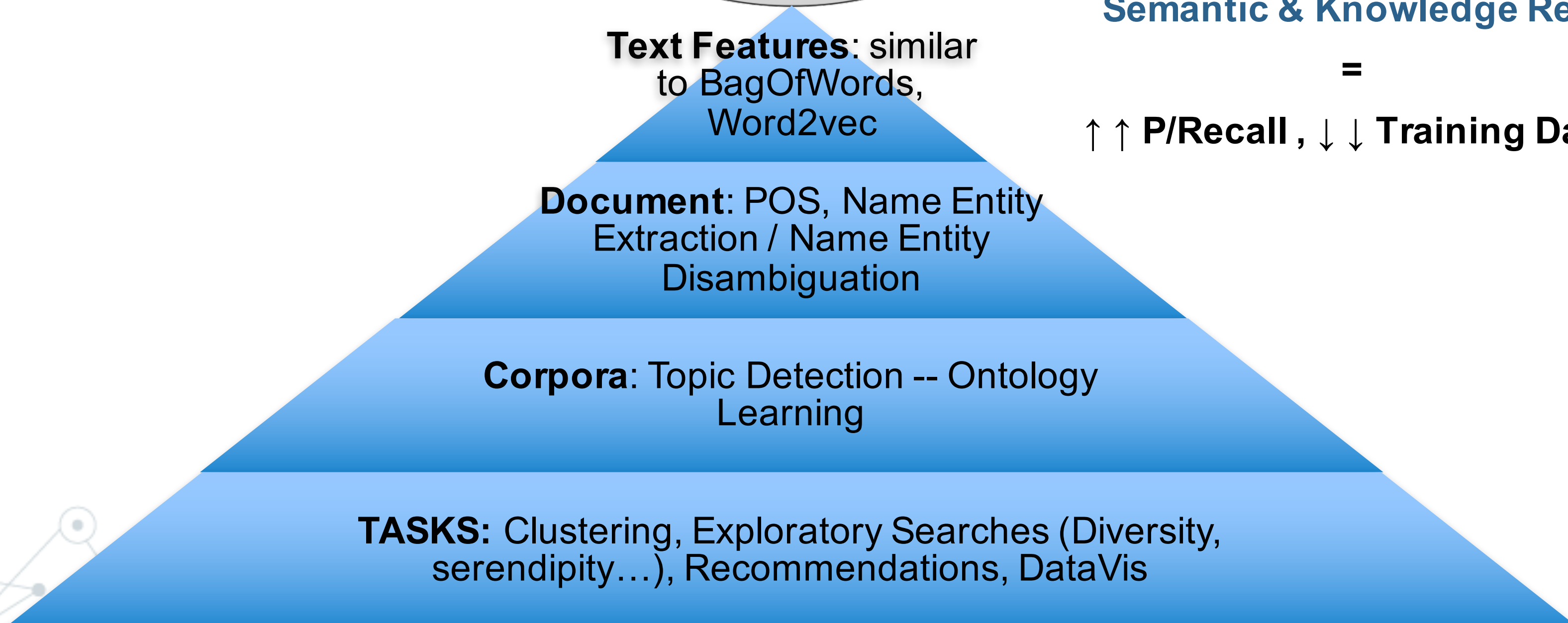
Generating **Explanations** analyzing documents considered in S_{web}

Question - Answering

Future Work:

Reusing concentric model in IA, and context-related algorithms:





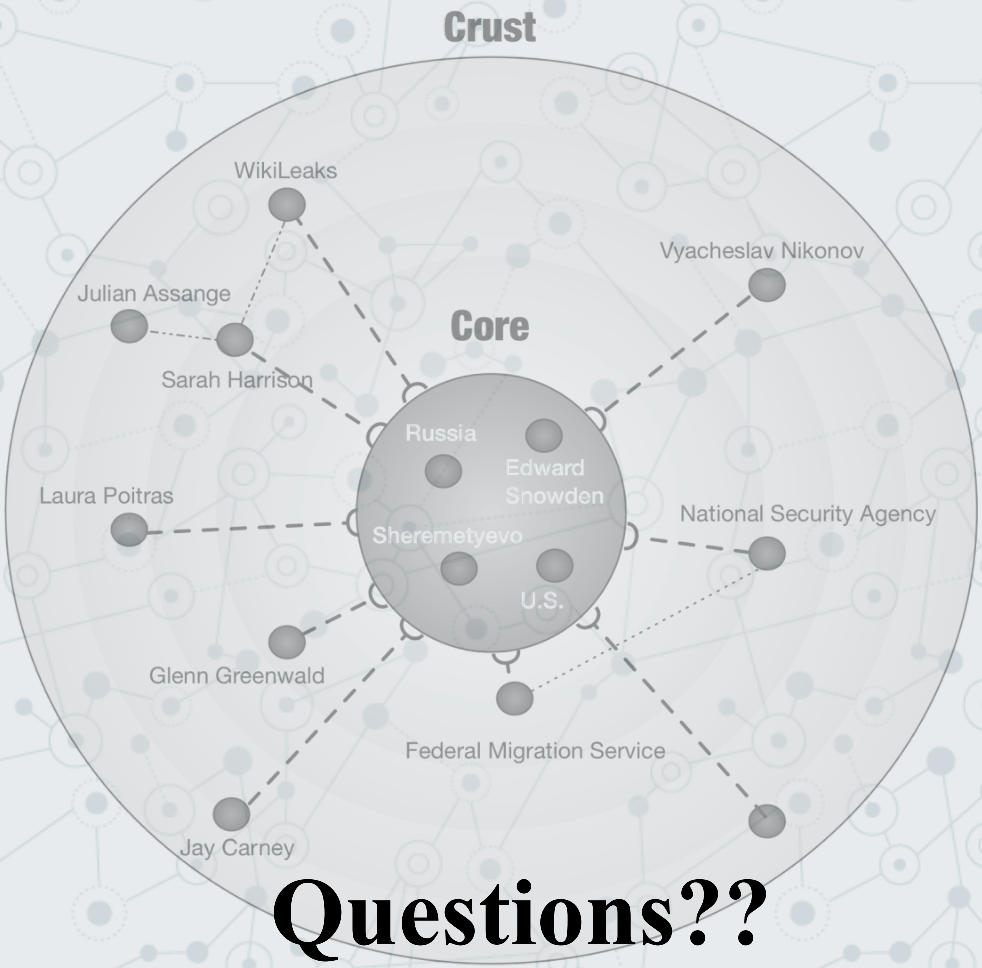
MACHINE LEARNING

+

Semantic & Knowledge Rep

=

↑↑ P/Recall , ↓↓ Training Data



Questions??

José Luis Redondo García

http://jluisred.github.io

@peputo

http://github.com/jluisred

References

[Redondo_KCAP'15B]	Capturing News Stories Once, Retelling a Thousand Ways
[Redondo_KCAP'15A]	The Concentric Nature of News Semantic Snapshots
[Redondo_ICWE'15]	Generating Semantic Snapshots of Newscasts using Entity Expansion
[Redondo_ISWC'14]	Finding and sharing hot spots in Web Videos
[Redondo_ESWC'14]	Augmenting TV Newscasts via Entity Expansion
[Redondo_SNOW'14]	Describing and Contextualizing Events in TV News Show
[LinkedTV_D2.6'14]	LinkedTV Framework for Generating Video Enrichments with Annotations
[Romero_TVX'14]	LinkedTV News: A dual mode second screen companion for web-enriched news broadcasts
[Hoang_MediaEval'14]	LinkedTV at MediaEval 2014 Search and Hyperlinking Task
[Rizzo_LREC'14]	Benchmarking the Extraction and Disambiguation of Named Entities on the Semantic Web
[Li_LIMe'13]	Enriching Media Fragments with Named Entities for Video Classification
[Milicic_WWW'13]	Live Topic Generation from Event Streams
[Milicic_ESWC'13]	Tracking and Analyzing The 2013 Italian Election
[Sahuguet_MediaEval'13]	LinkedTV at MediaEval 2013 Search and Hyperlinking Task
[Rizzo_SAM'12]	What Fresh Media Are You Looking For? Extracting Media Items from Multiple Social Networks

Experiment 3: Multimodal VS Concentric (NMDCG)

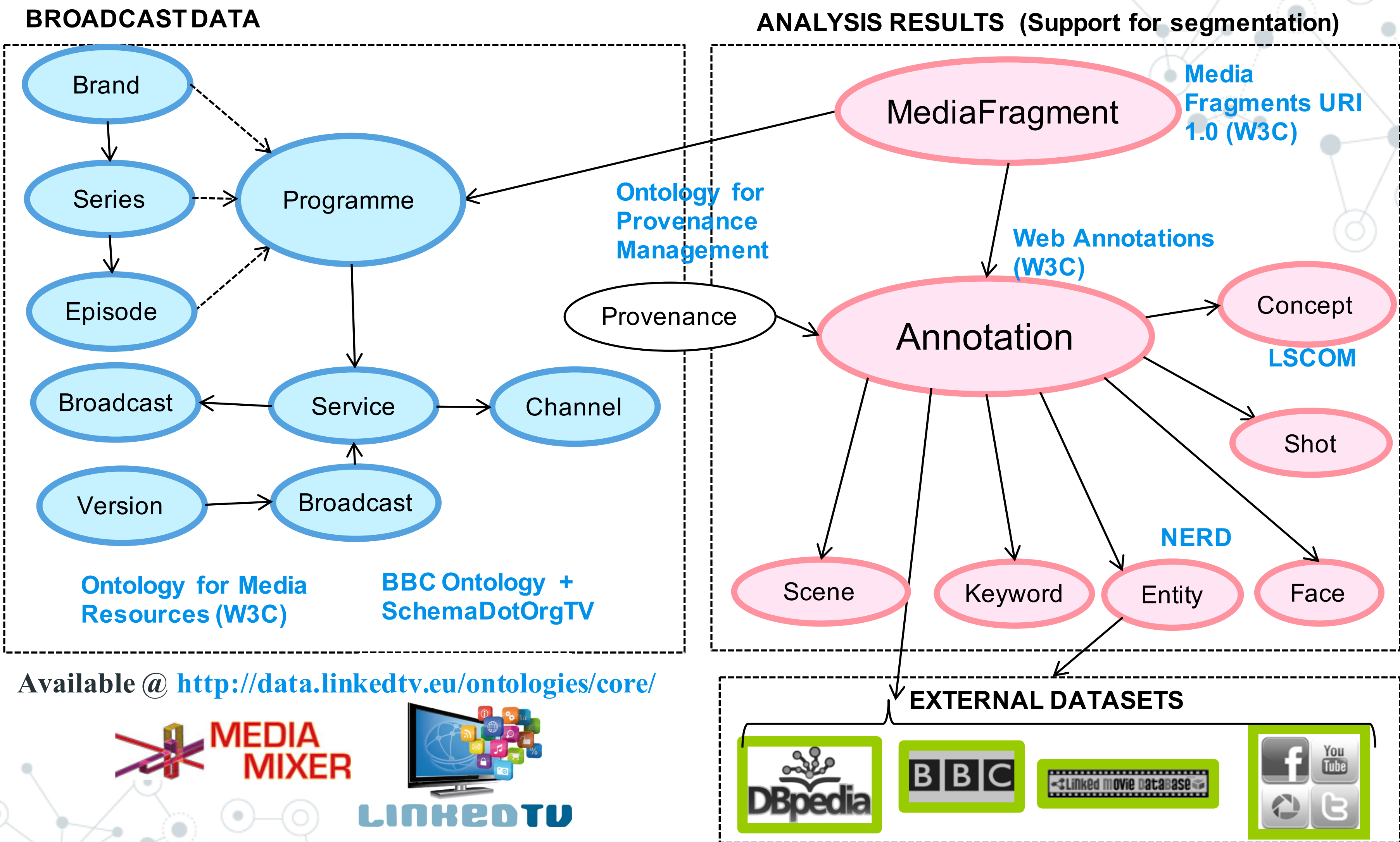
- Best Multimodal Run, NMDCG @ 10: 0.698

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

- Best Concentric Approach, NMDCG @ 10: 0.645
  - Best Run: [Ex1, Core A, Core+Crust]
  - Slightly lower NMDCG @ 10:
    - Ranking is too much application dependent
  - Need of reconsider evaluation:
    - Recall is higher (0.35 VS 0.37)
    - Why @ 10?

Multimedia Model: LinkedTV Model

Part 1.b

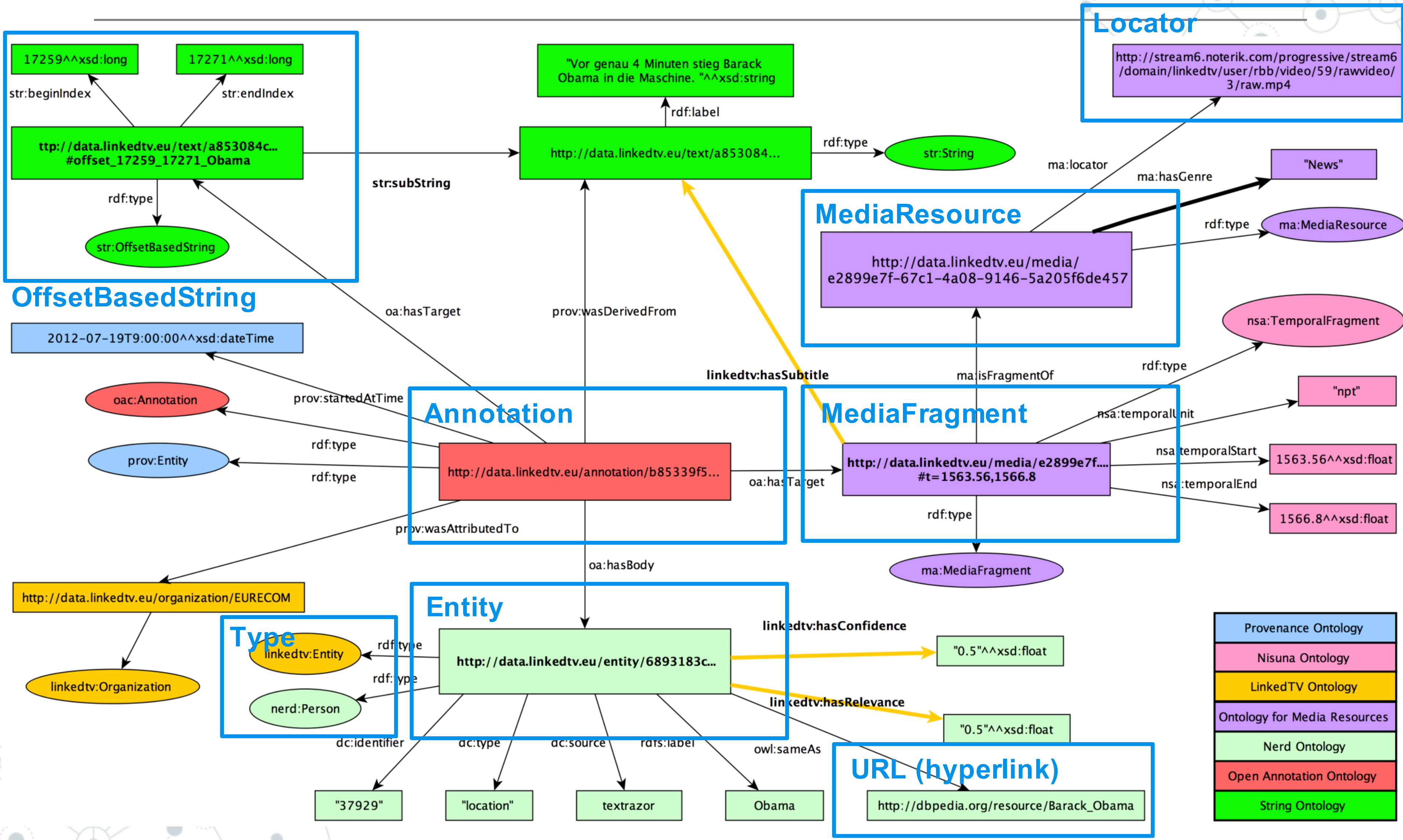


Available @ http://data.linkedtv.eu/ontologies/core/

MEDIA MIXER

LINKEDTV

Multimedia Model: LinkedTV Model

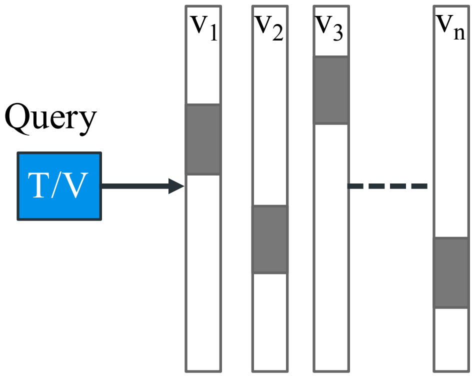


Evaluation: Multimodal @ Mediaeval 2013

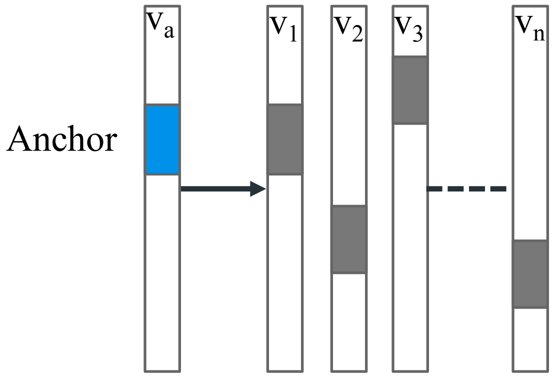
Part 1.c

MediaEval Benchmark

Search Task



Hyperlinking Task



~ 1697h of BBC video data, 2323 videos

- Different TV shows (news, sports, politics…) from 2012
- Subtitles and ASR (English)
- Output of some visual algorithms: shot and face detection

Evaluation: Multimodal @ Mediaeval 2013

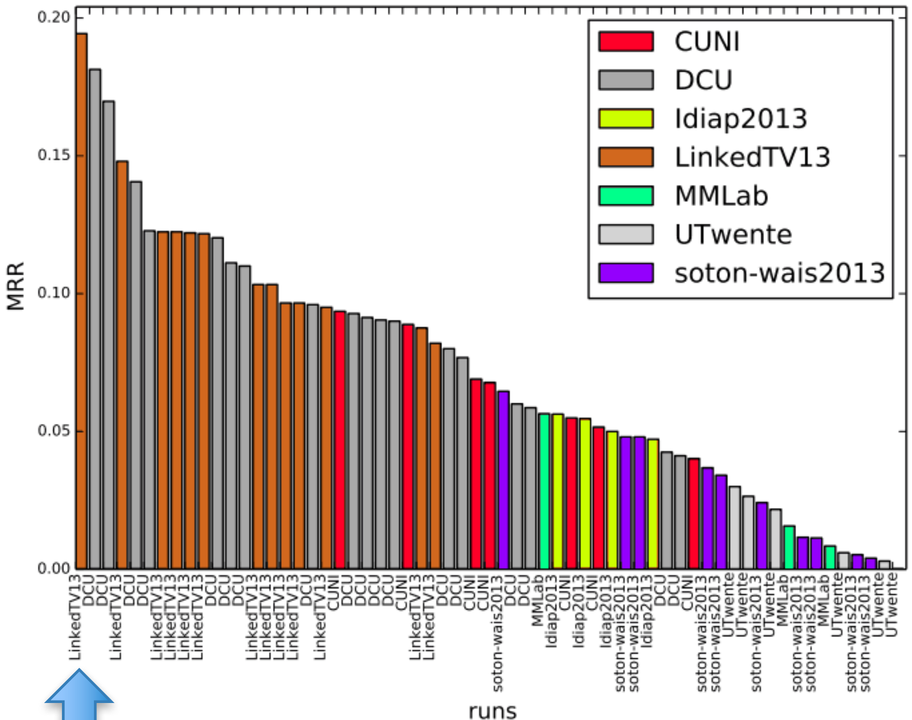
Annotations	Processing Time	Type
Visual Concept Detection (151)	**20 days on 100 cores**	**Visual** **
Scene Segmentation	2 days on 6 cores	**Visual**
OCR	1 day on 10 cores	**Visual**
Keywords Extraction	5 hours	**Textual** **
Named Entities Extraction	4 days	**Textual**
Face detection and Tracking	4 days on 160 cores	**Visual**

Approach:

- Data Indexing:
  - Lucene & Solr
  - Granularities: Shot, Scenes, Sliding Windows…
  - Multimodality
- Query Formulation:
  - Search: Text + Visual Cues + Visual Concept Mapping, LSCOM
  - Hyperlink: Subtitles, Keywords, LSCOM concepts (MoreLikeThis)

Evaluation: Mediaeval 2013 Results

Part 1.c



Search Task

0.19 MRR

(Mean R. Rank)

[Sahuguet_MediaEval'13]

0,72 P_{10}

Hyperlinking Task