# Selective Information Dissemination on the Real-time Web - A Database Perspective

Bernd AMANN, UPMC – LIP6

Keynote WEBIST 2015

Lisbon, Portugal

May 20, 2015

# LIP6 Database Research Group

- Web Information Streams
  - continuous top-k queries
  - multi-query optimization
- Web Crawling and Archiving
  - crawling dynamic contents
  - archive construction and querying
- Cloud Data Processing
  - large-scale transactions
  - data distribution and replication
- Data provenance and quality
  - provenance generation and quality of data-centric workflows
- Text document representation
  - computational linguistics
  - machine learning

# My personal « Web » timeline

Hypertext Query Languages (PHD)

XML repositories and Active views

Semantic XML data integration

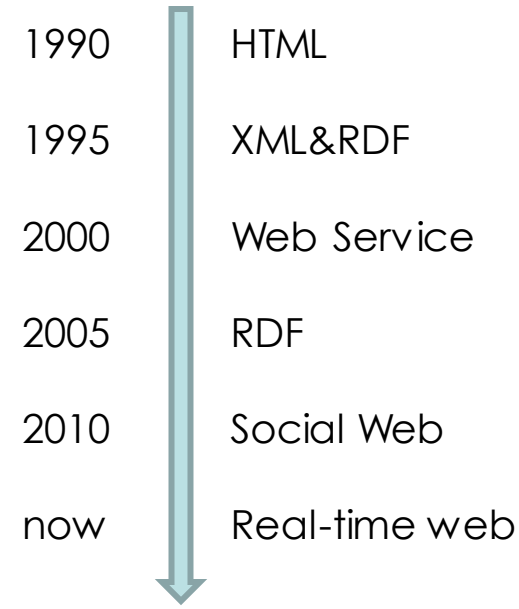
Intensional XML data

Web Service Ranking

RSS data acquisition and aggregation

XML Workflow provenance

Continuous Top-k Query Processing



# The Web Revolution

Web of Knowledge

reason

## Web of Contents

publish, search and explore contents

## Web of Services

acces and update data

interact with data

## Web of People

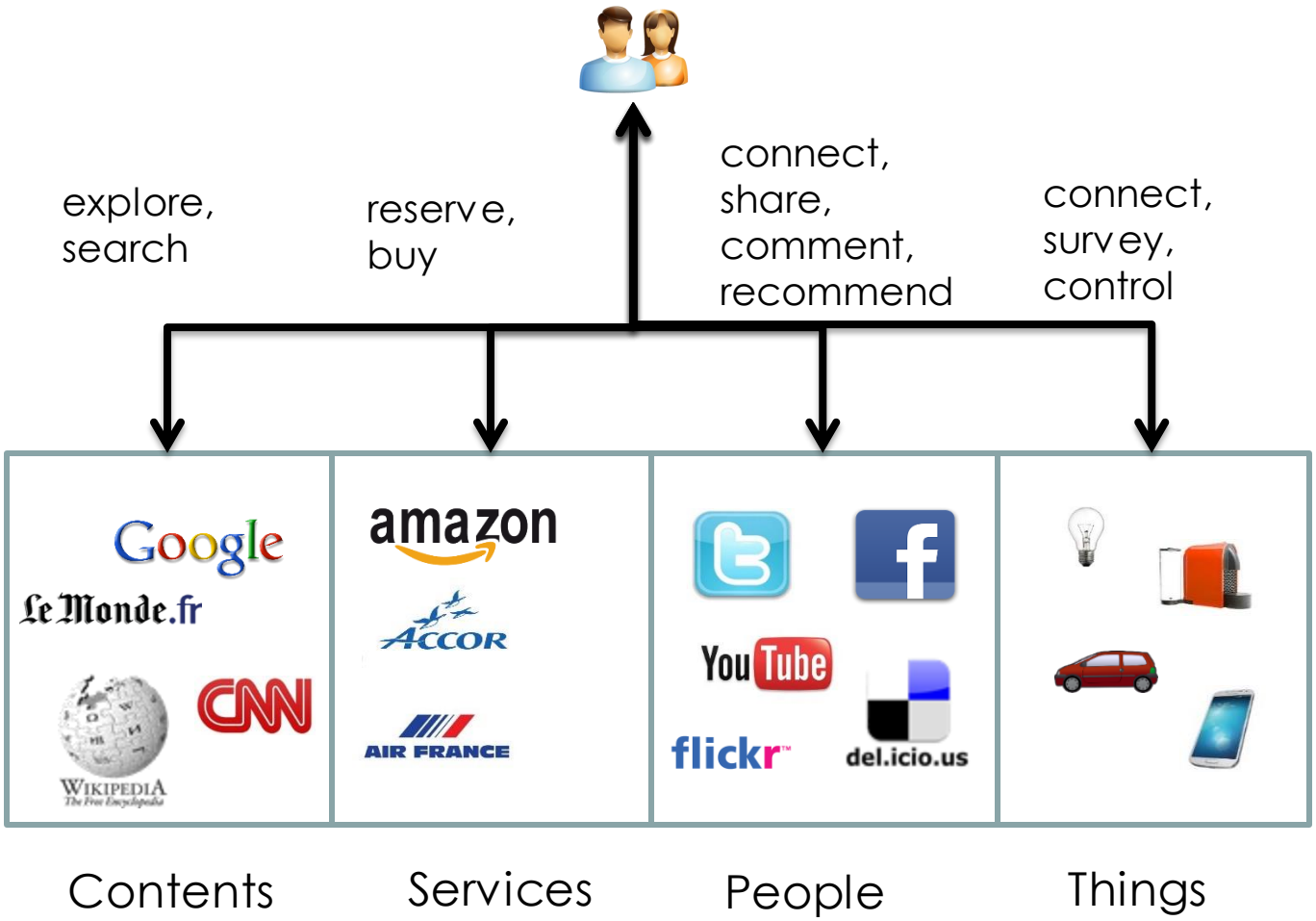
publish and share

interact with people

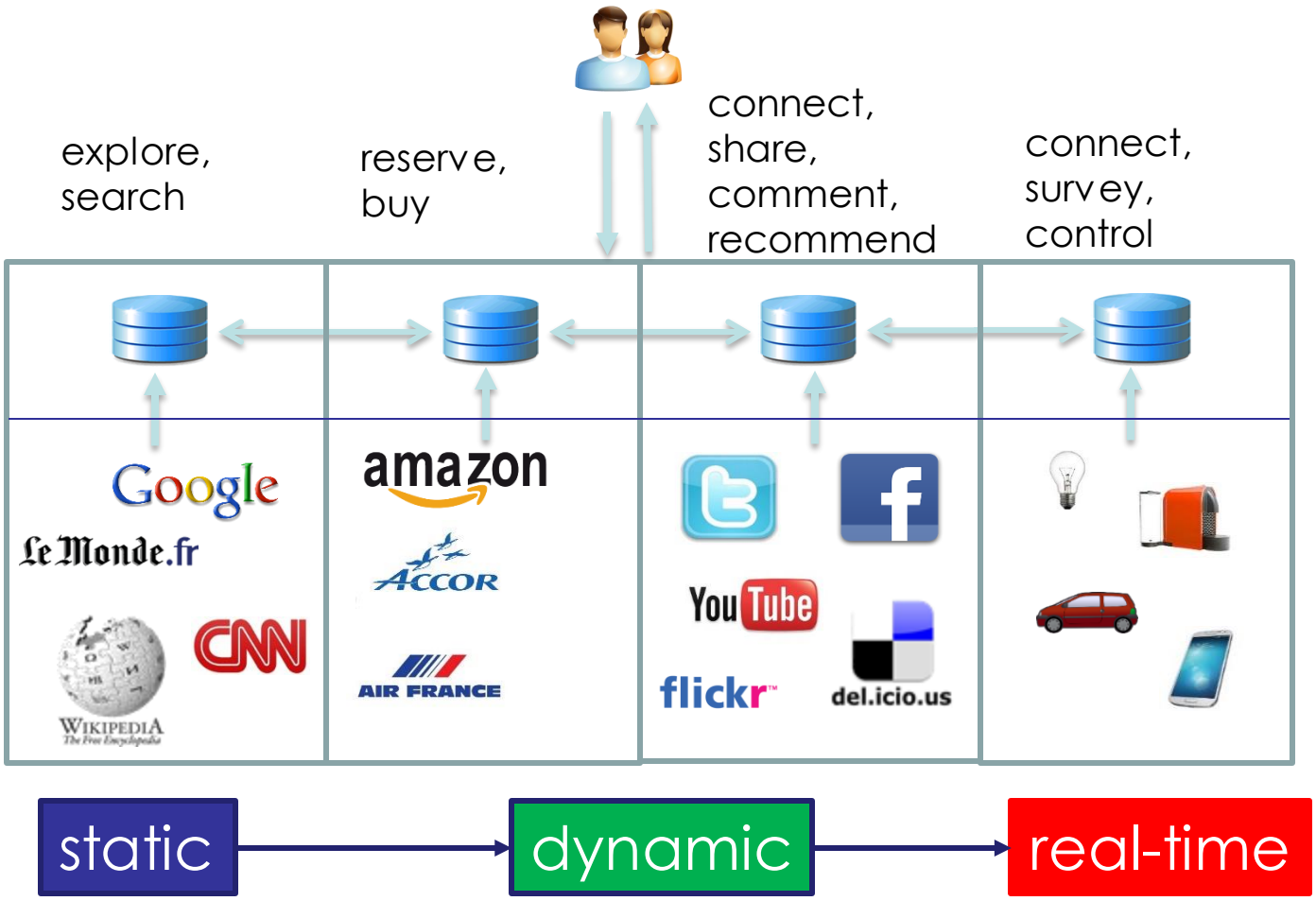
## Web of Things

interact with things

# "Interactive" Web



# Interaction = dynamic data



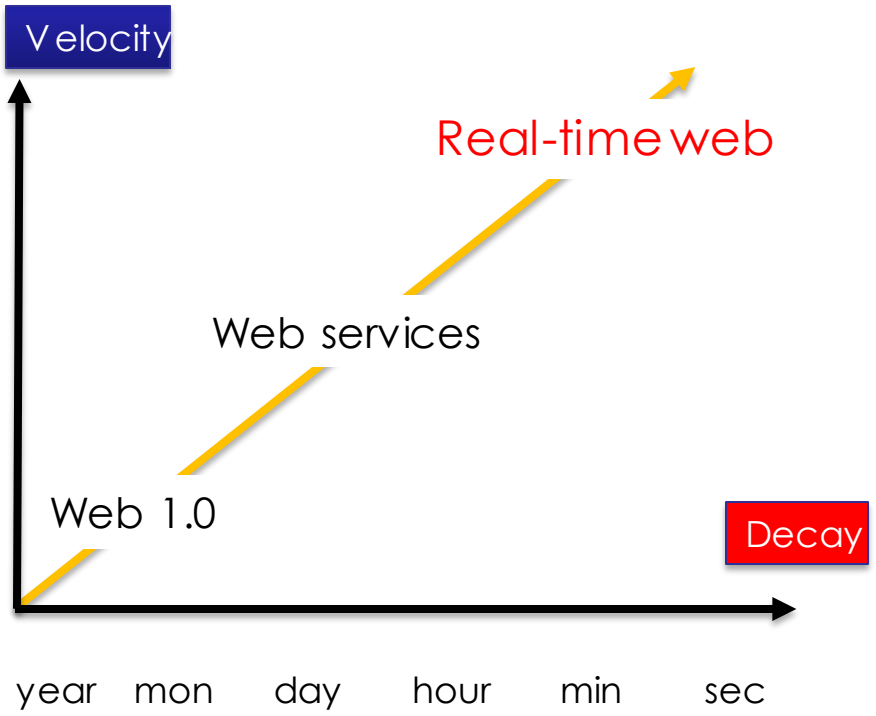
# Real-time web

Static web

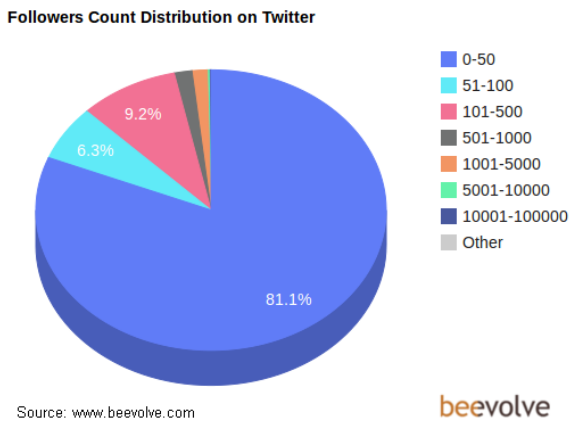
- Large, complex documents
- Low information decay
- Ad-hoc search
- On demand processing

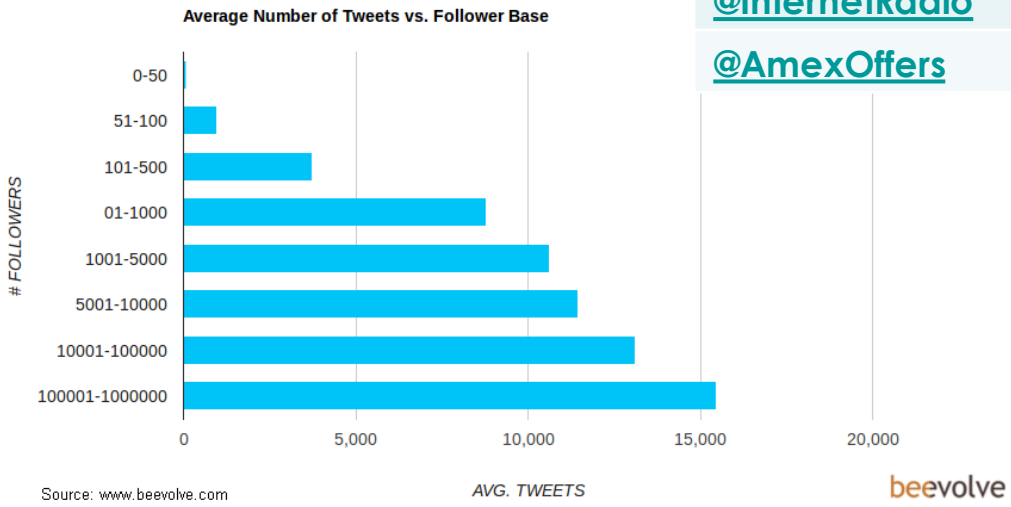
Real-time web

- Simple, short messages
- High information decay
- Publish-subscribe dissemination
- Continuous processing



# Twitter : real-time information dissemination




| | followers | following | tweets/day |
|---|---|---|---|
| @HuffingtonPost | **5 700 000** | 5 500 | 200-500 |
| @justinbieber | 53 000 000 | 210 000 | 5-15 |
| @BarackObama | 69 000 000 | 642 000 | 5-20 |
| @NBA | 14 000 000 | 1 300 | 130-230 |
| @MTV | 12 000 000 | 31 000 | 100-250 |
| @CNN | 17 000 000 | 1 000 | 40-70 |
| @InternetRadio | 1 300 | 32 | 800 – 1000 |
| @AmexOffers | 56 000 | 0 | 100 – 40 000 |

http://twittercounter.com/

500 000 000 tweets/day

# Rest of this talk

Online information aggregation, ranking and filtering

- News
- Web syndication (RSS)
- Micro-blogging (Twitter)
- Social media

Outline :

- RoSeS : Content-based RSS aggregation
- Meows : Continuous top-k queries over the real-time web

- Perspectives


# Really Open, Simple and Efficient Syndication

**Jordi Creus[1], Roxana Horincar[1]**

Bernd Amann[1], Nicolas Travers[2], Dan Vodislav[3]

[1] LIP6 – Université Pierre & Marie Curie

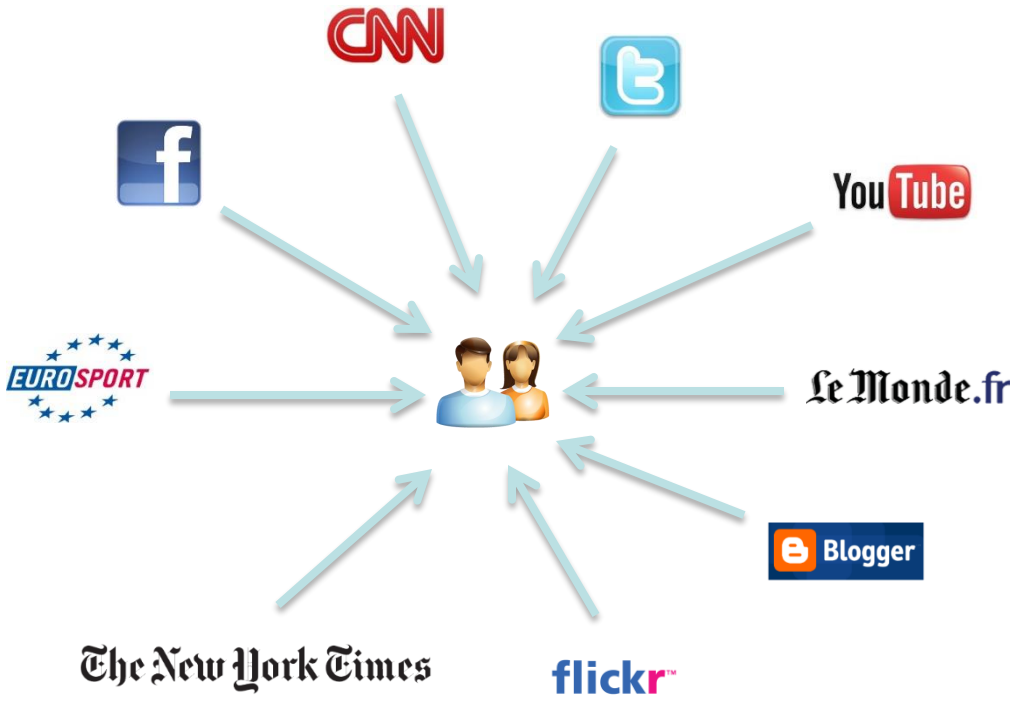
[2] CEDRIC – Conservatoire National des Arts et Métiers

[3] ETIS – Université de Cergy-Pontoise

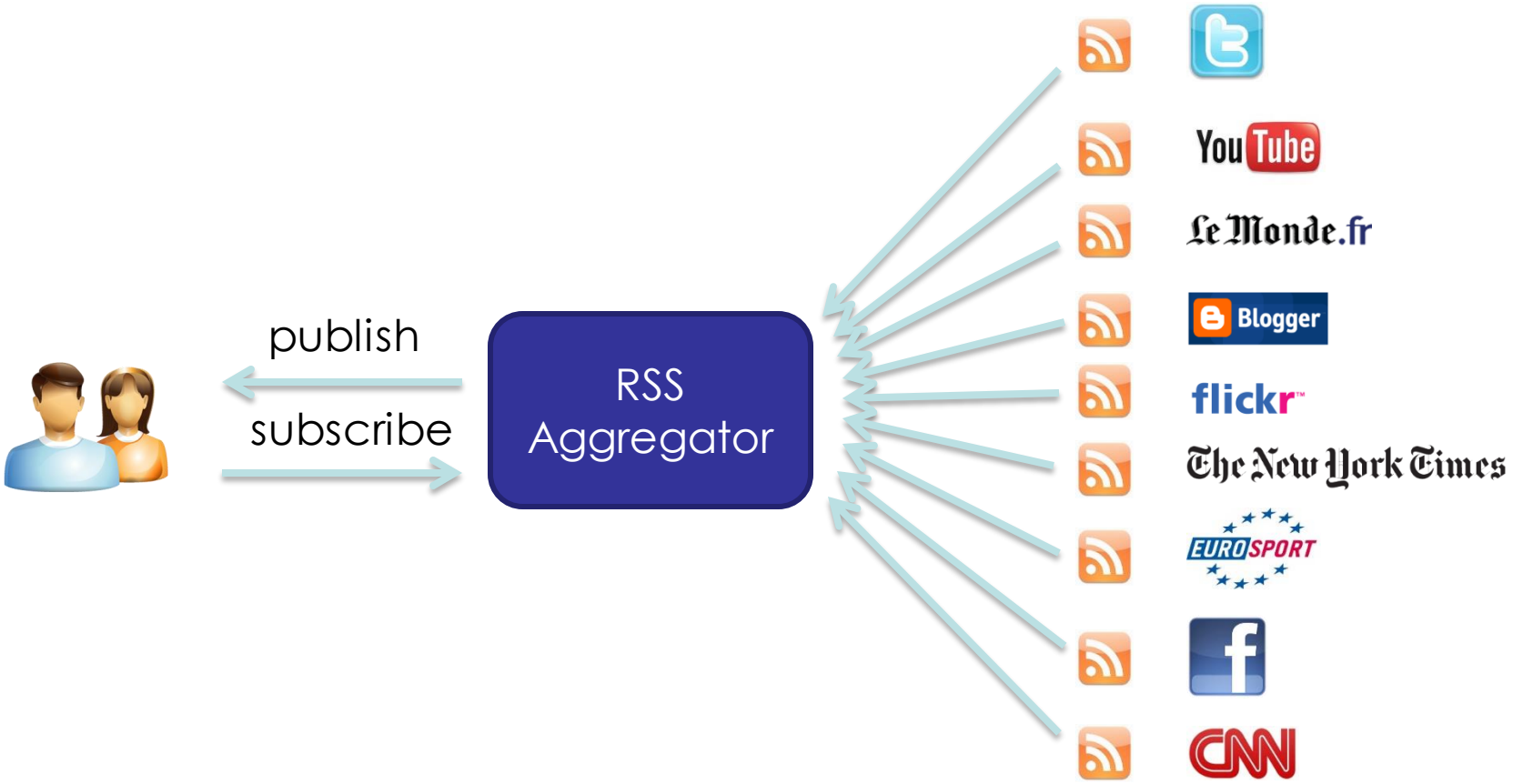





# Content-based RSS Aggregation



# RSS Aggregation



# RSS web syndication

Really Simple Syndication

- standard web feed format to publish frequently updated information

RSS feed

- XML document including full or summarized text and metadata with links

RSS aggregators

- Feedly : 40 M feeds / 15 M users
- Google Reader (< June 1013)
- Yahoo Pipes
  - https://pipes.yahoo.com/

RSS usage :

- top 10k sites : 21 %
- top 100k sites : 22 % top
- 1M sites : 30 %
- entire Internet : 6 % (20M)

Site types:

- Business
- News
- Social
- Technology



source http://trends.builtwith.com

# RSS Feeds

http://www.wikicfp.com/cfp/rss?cat=web

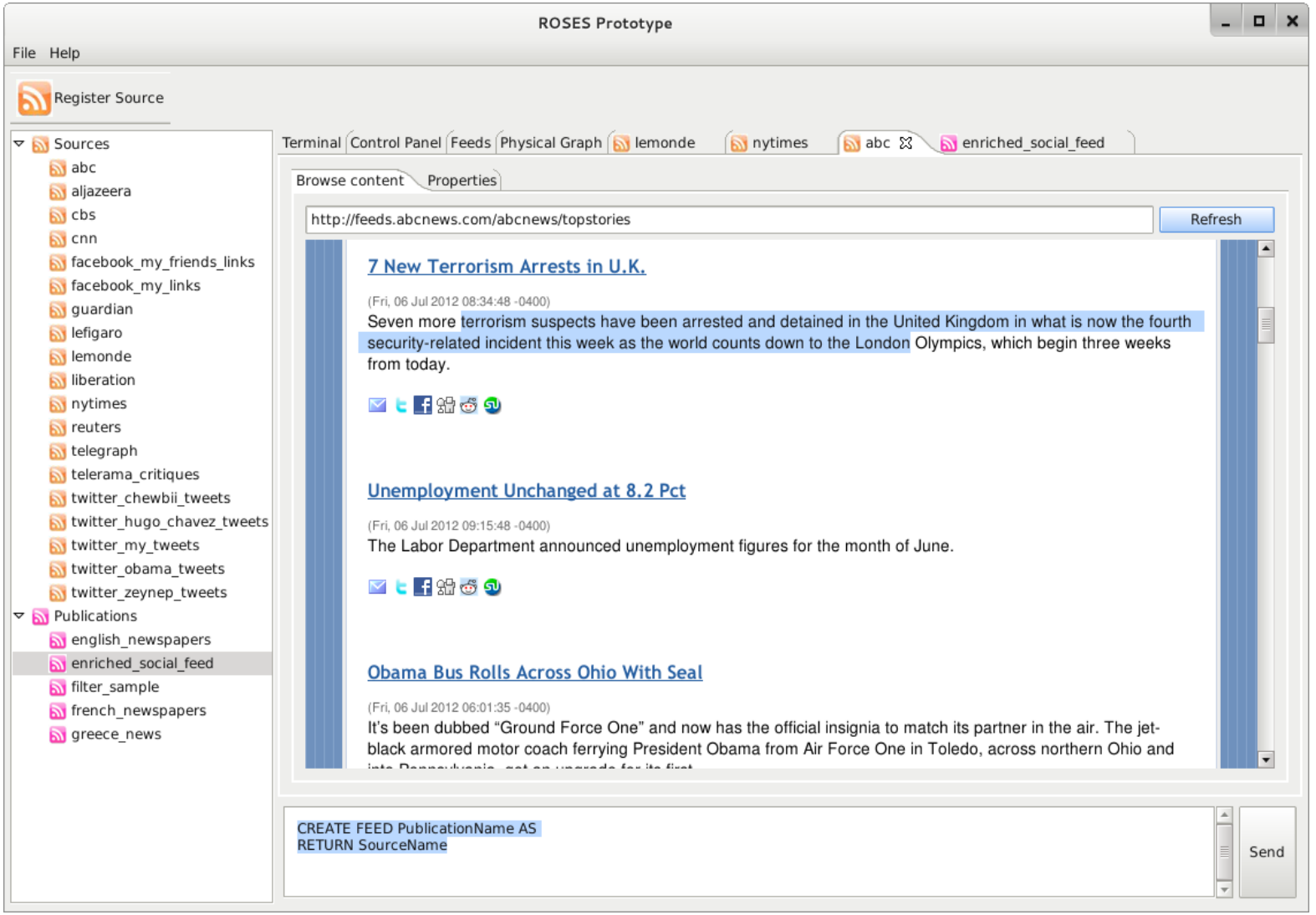
```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0">
 <channel>
  <title>CFPs on Web : WikiCFP</title>
  <link>http://www.wikicfp.com/cfp/call?conference=web</link>
  <description>A Wiki to Organize and Share Calls For Papers</description>
  <item>
    <title>SaW 2015 : International Workshop on Semantic and Web</title>
    <link>http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=46102</link>
    <description>International Workshop on Semantic and Web [Porto - Portugal]
                 [Oct 21, 2015 - Oct 23, 2015]
    </description>
   <guid isPermaLink="false">cfp-528064-S@wikicfp.com</guid>
  </item>
   ....
</rss>
```


# Really Open, Simple and Efficient Syndication

[WISE'2010]
[DEXA'2011]
[CIKM'2011 demo]
[ICWE'2012]
[WWWJ'2014]



ROSES


ROSES Prototype
File Help
Register Source
Sources
abc
aljazeera
cbs
cnn
facebook_my_friends_links
facebook_my_links
guardian
lefigaro
lemonde
liberation
nytimes
reuters
telegraph
telerama_critiques
twitter_chewbii_tweets
twitter_hugo_chavez_tweets
twitter_my_tweets
twitter_obama_tweets
twitter_zeynep_tweets
Publications
english_newspapers
enriched_social_feed
filter_sample
french_newspapers
greece_news
Terminal
Control Panel
Feeds
Physical Graph
lemonde
nytimes
abc
enriched_social_feed
Browse content
Properties
http://feeds.abcnews.com/abcnews/topstories
Refresh
7 New Terrorism Arrests in U.K.
(Fri, 06 Jul 2012 08:34:48 -0400)
Seven more terrorism suspects have been arrested and detained in the United Kingdom in what is now the fourth security-related incident this week as the world counts down to the London Olympics, which begin three weeks from today.
Unemployment Unchanged at 8.2 Pct
(Fri, 06 Jul 2012 09:15:48 -0400)
The Labor Department announced unemployment figures for the month of June.
Obama Bus Rolls Across Ohio With Seal
(Fri, 06 Jul 2012 06:01:35 -0400)
It's been dubbed "Ground Force One" and now has the official insignia to match its partner in the air. The jet-black armored motor coach ferrying President Obama from Air Force One in Toledo, across northern Ohio and
CREATE FEED PublicationName AS
RETURN SourceName
Send

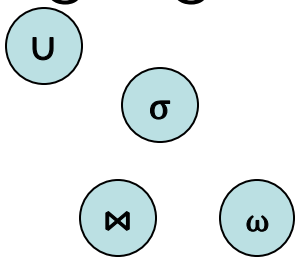
# Publication language

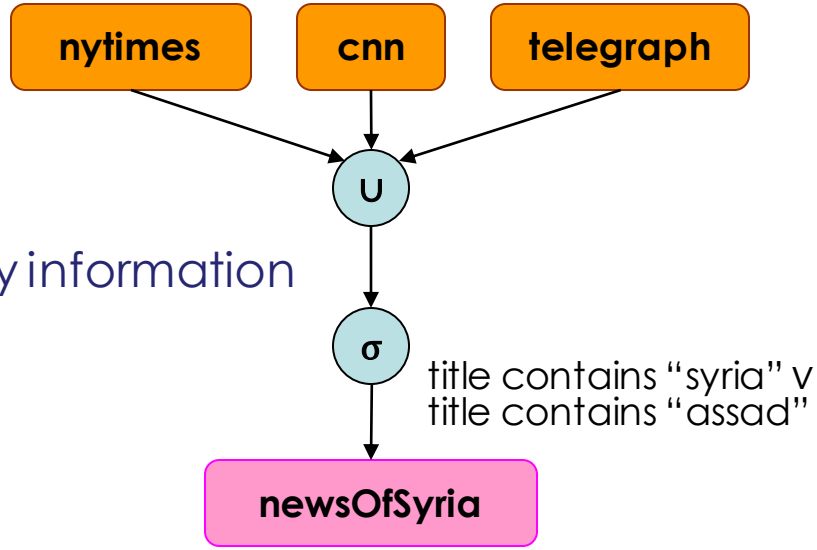
## Requirements

- Query collections of feeds
- Apply text filters
- Annotate items with complementary information

## Publication language :

- union
- selection
- join & window



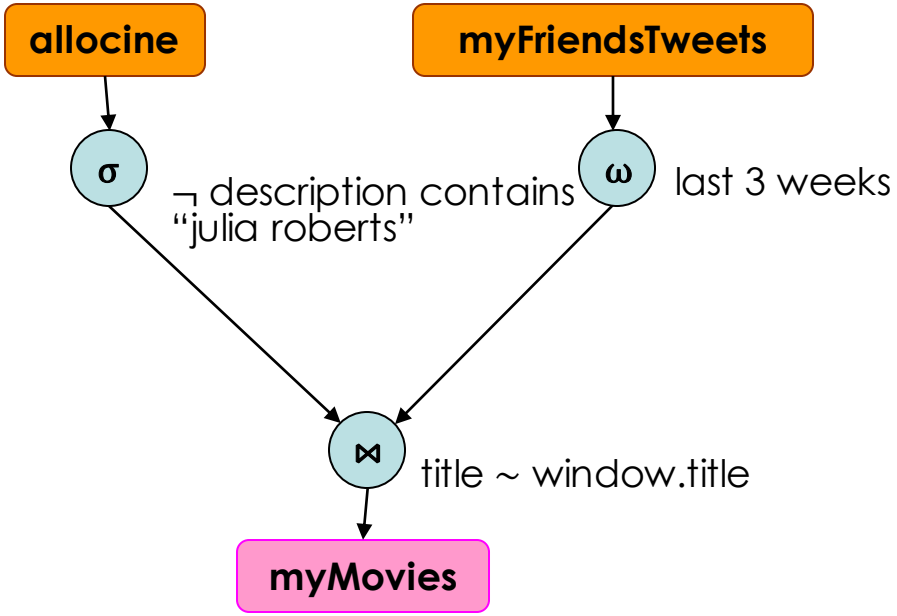




## Example

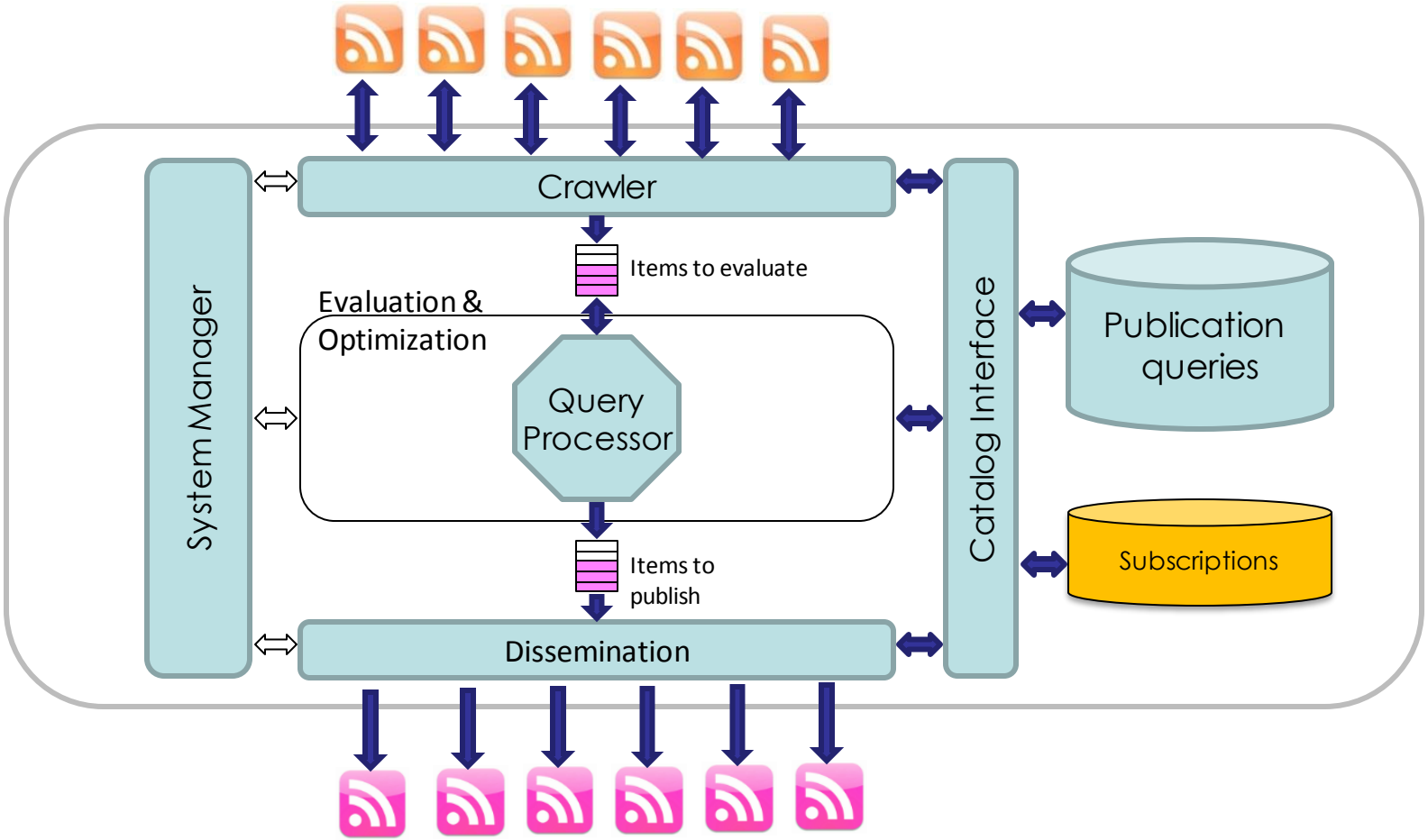
create feed newsOfSyria
from nytimes | cnn | telegraph
where title contains "syria" or title contains "assad"

# Publication language

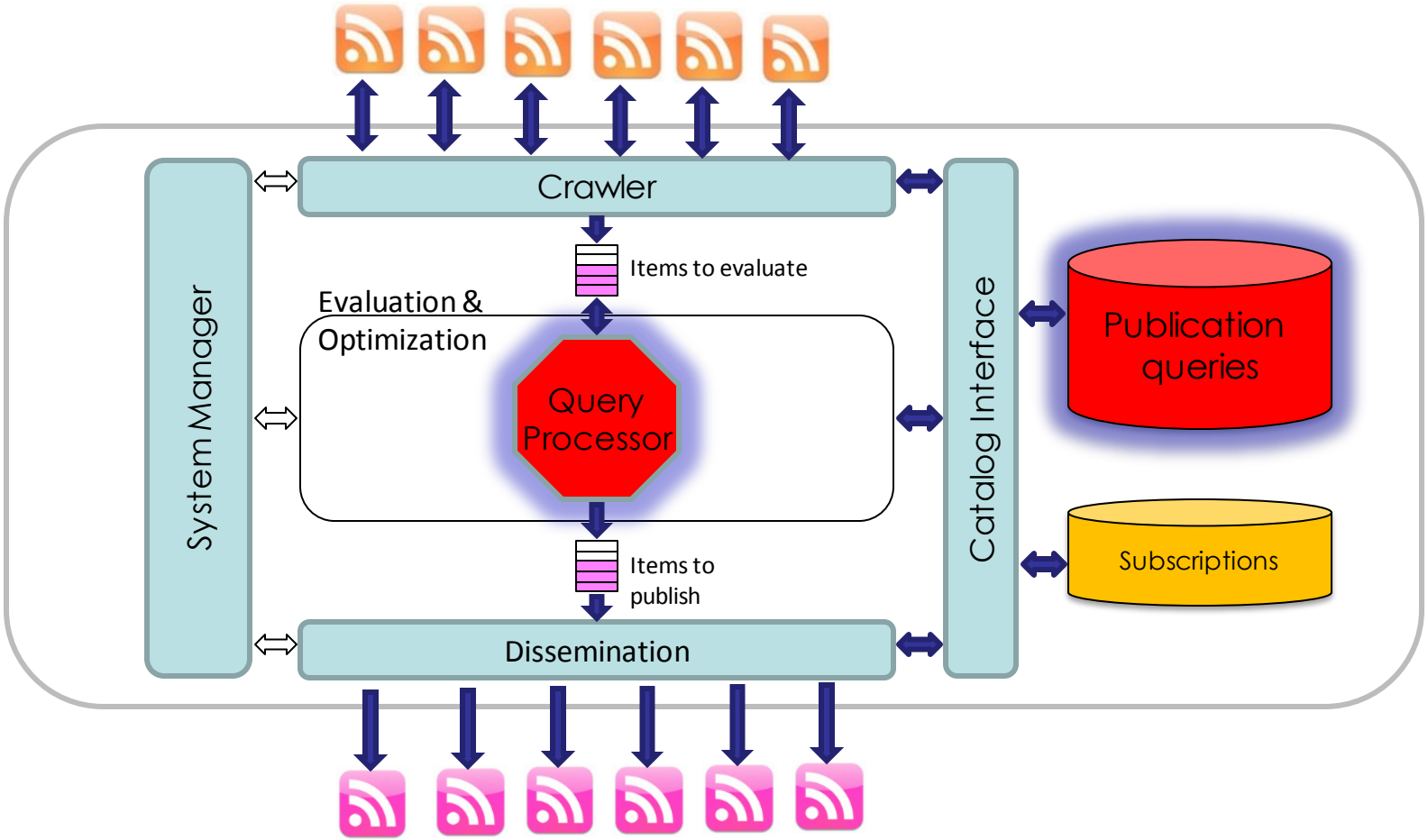
create feed myMovies from allocine as $a
join last 3 weeks on myFriendsTweets with $a[title similar window.title]
where $a[description not contains "julia roberts"]



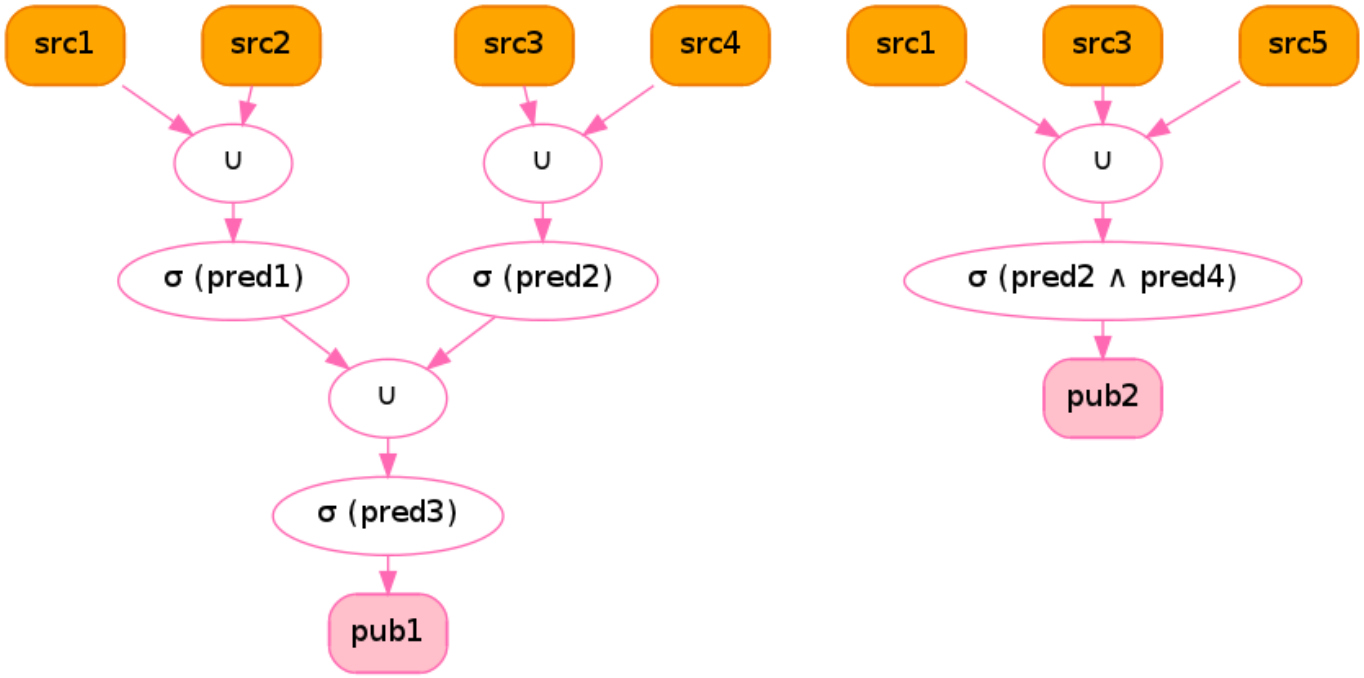
# System Overview



# Query Processor



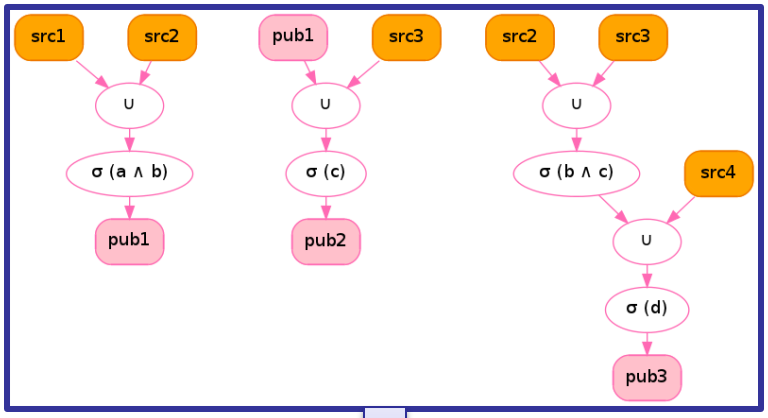
# Multi-Query Optimization

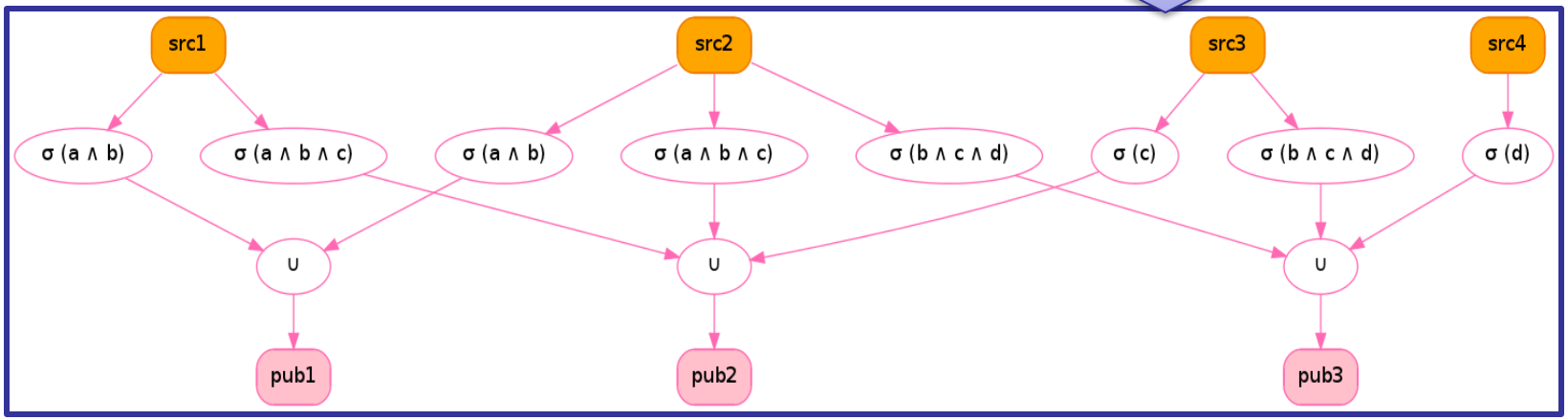


# Build Shared Filter Plans

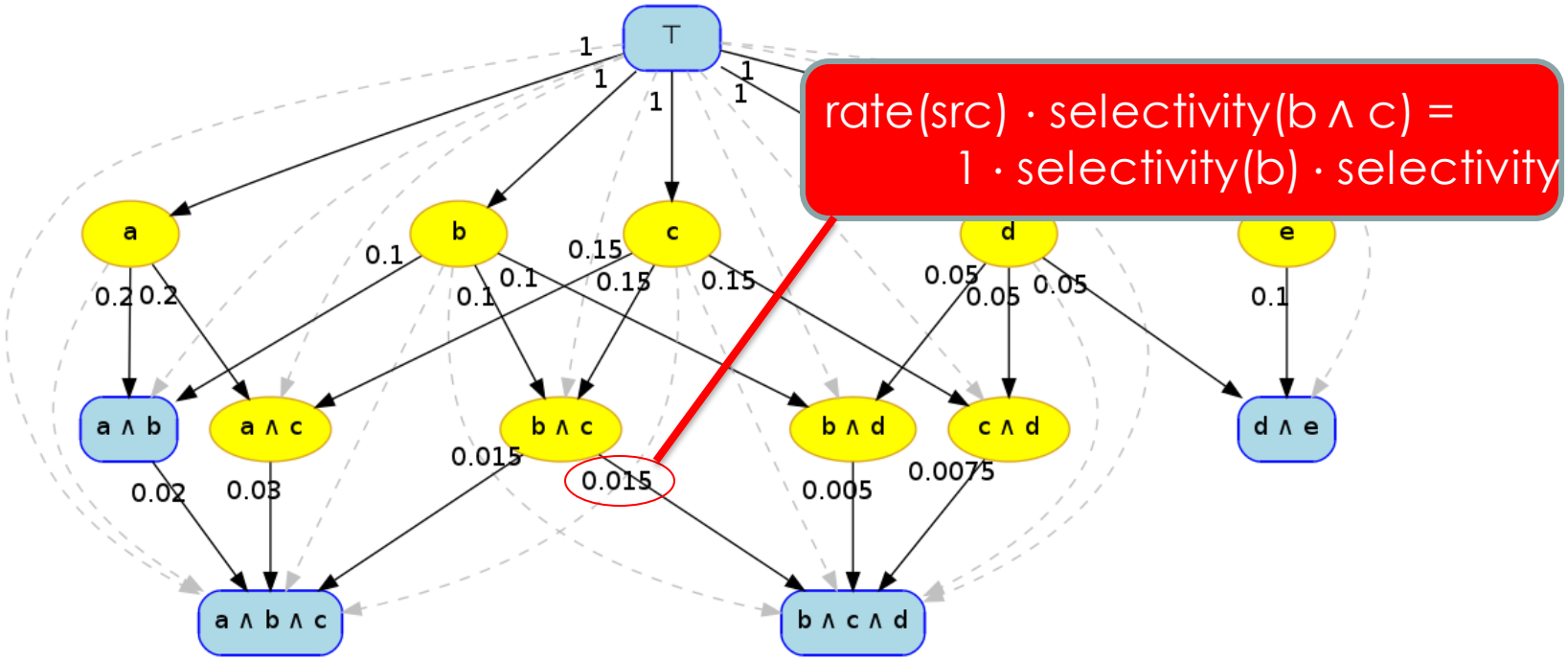
Rewriting rules

- Distribute selection over unions
- Flatten cascading selections
- View decomposition
- Commute join and selection, …

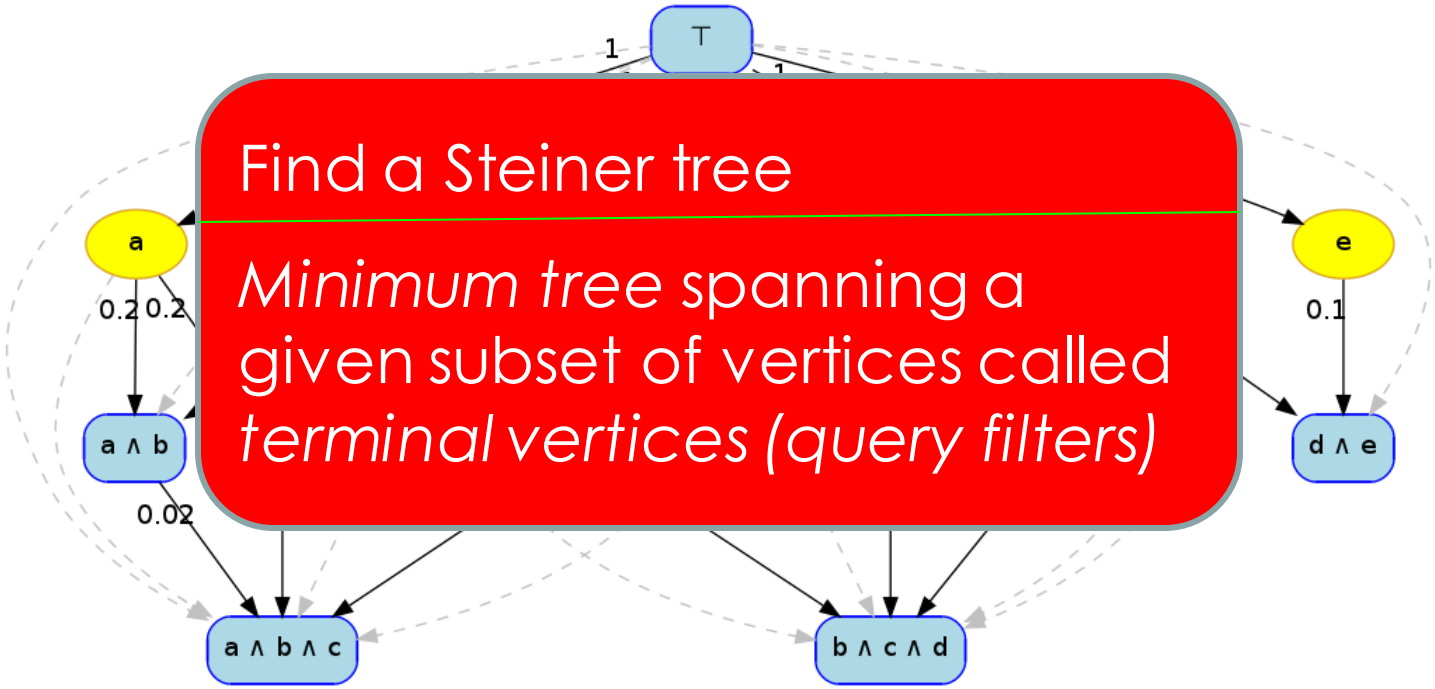


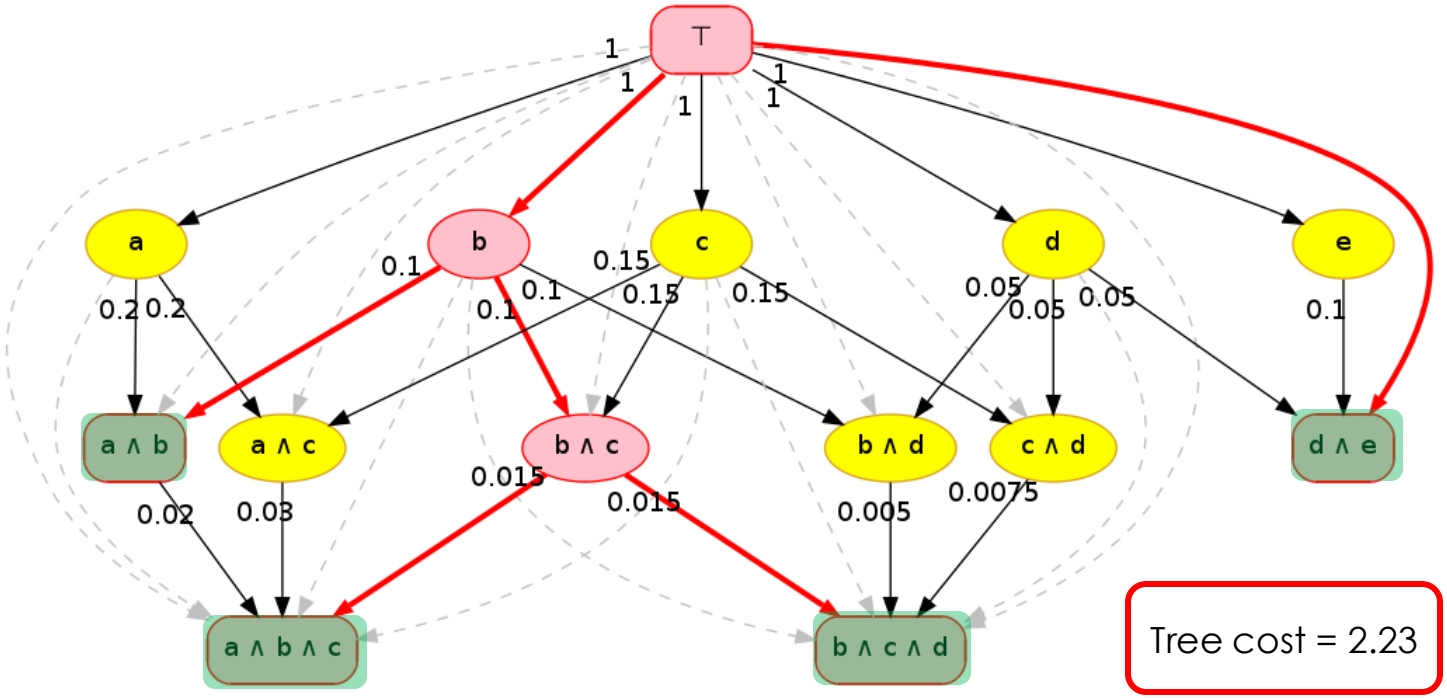
# Extended Filter Plans (with cost estimation)



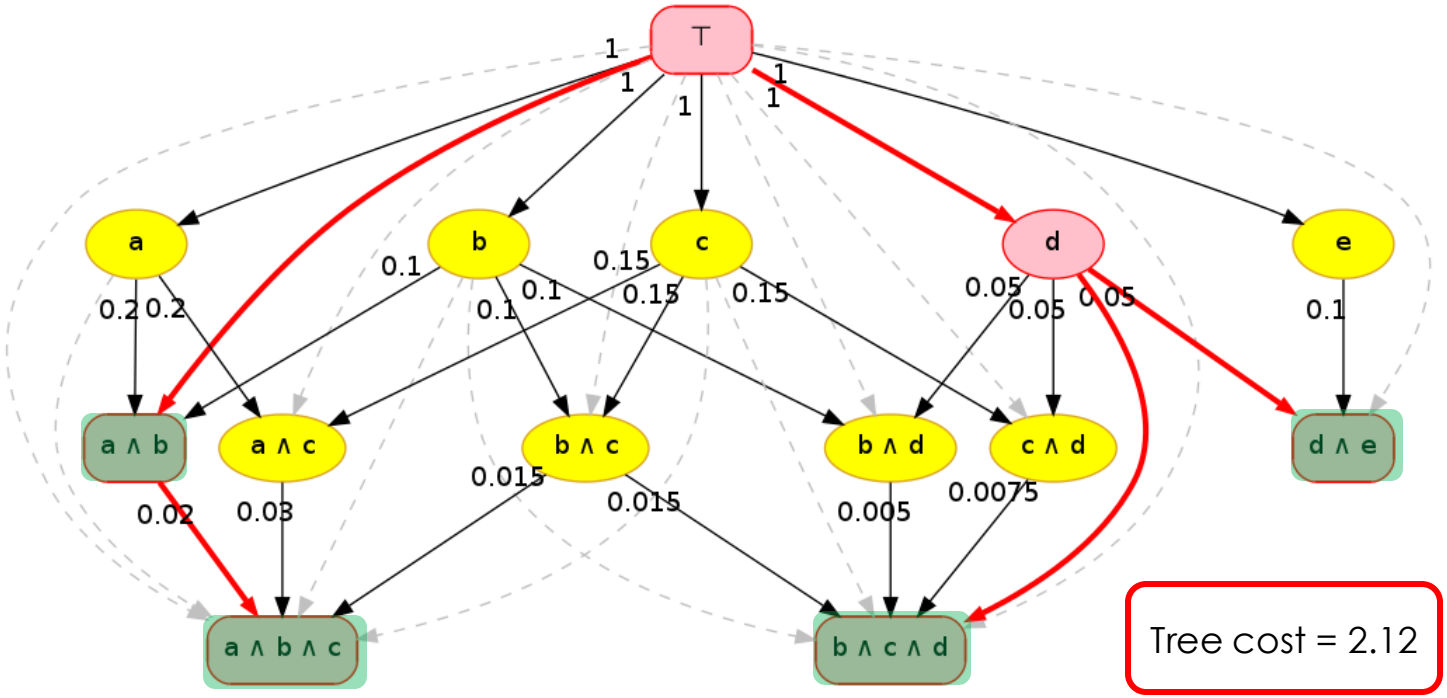
# Extended Filter Plan Optimization



# Extended Filter Plan Optimization



# Extended Filter Plan Optimization



# VCA Algorithm

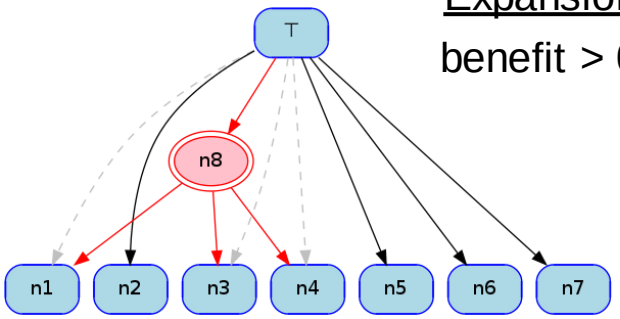
Steiner tree problem is NP-complete

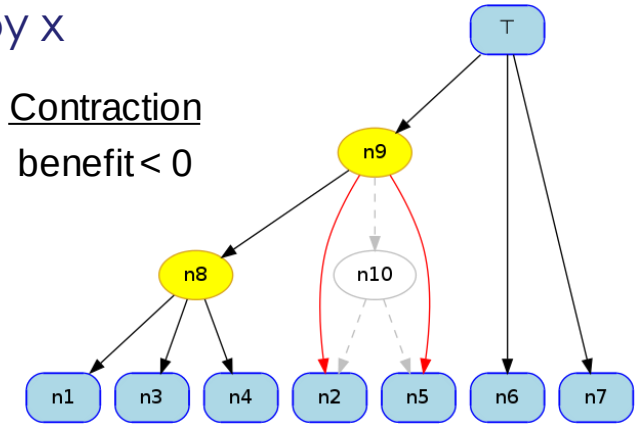
STA Algorithm

- greedy state-of-the art algorithm with approximation guarantees

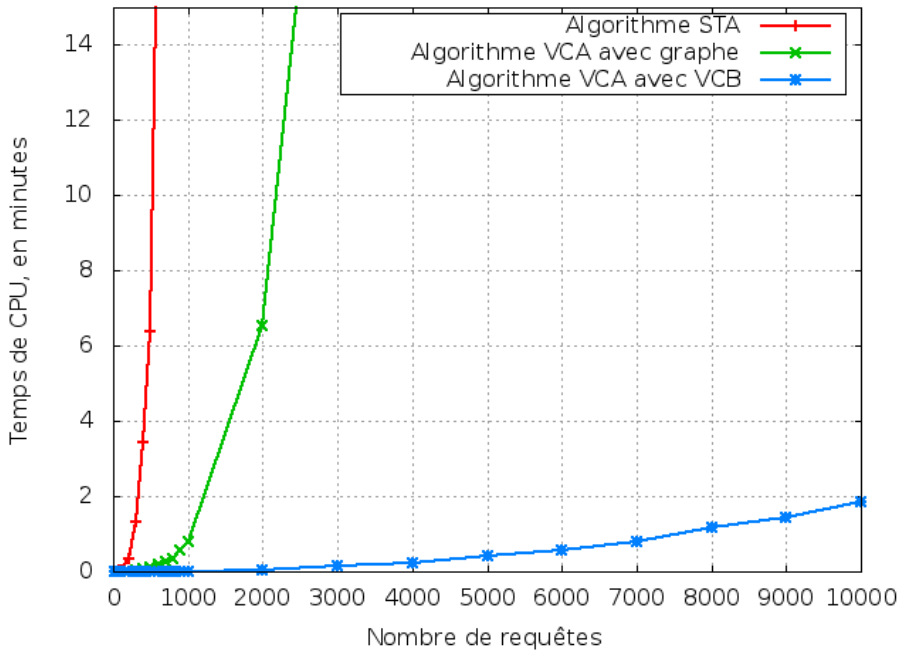
VCA – Algorithm

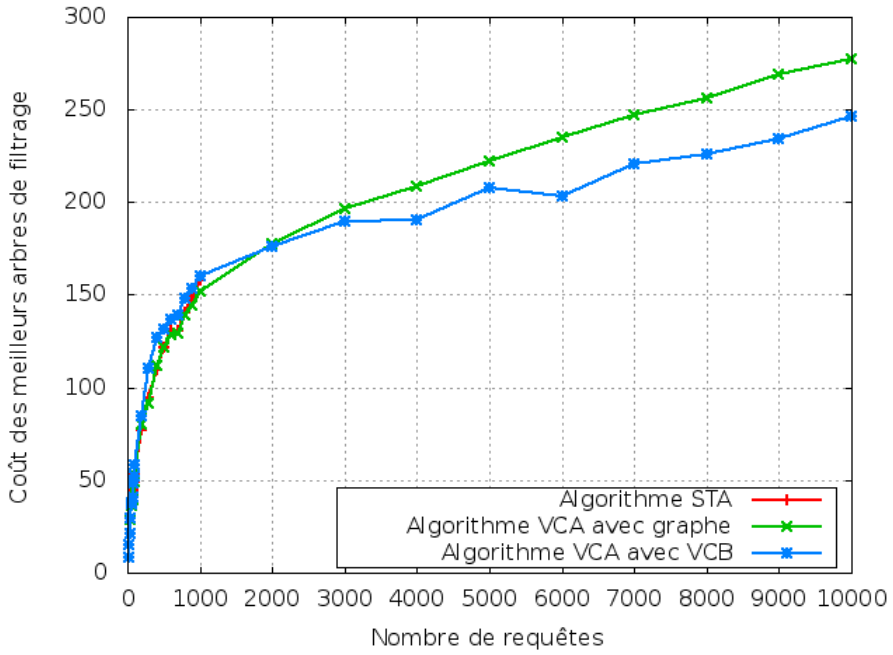
- expansion/reduction step at each iteration
- benefit(x, y) = (n – 1) * selectivity(y) – n * selectivity(x)
- n : number of sons of y also subsumed by x



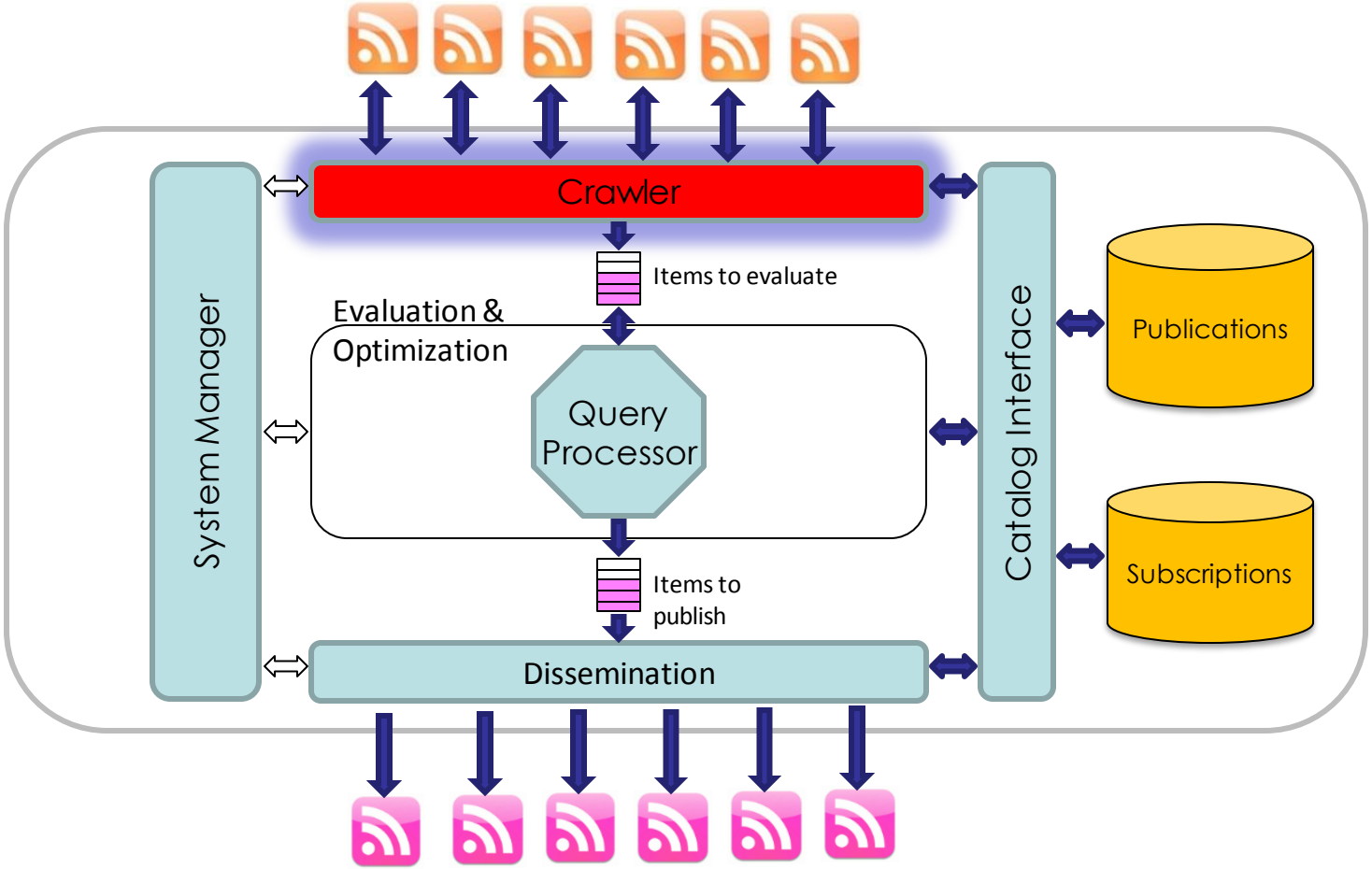



# Experiments : number of queries



# RSS Crawler


Crawler
Items to evaluate
Evaluation &
Optimization
Query
Processor
Items to
publish
Dissemination
System Manager
Catalog Interface
Publications
Subscriptions

# RSS Crawler - Challenges

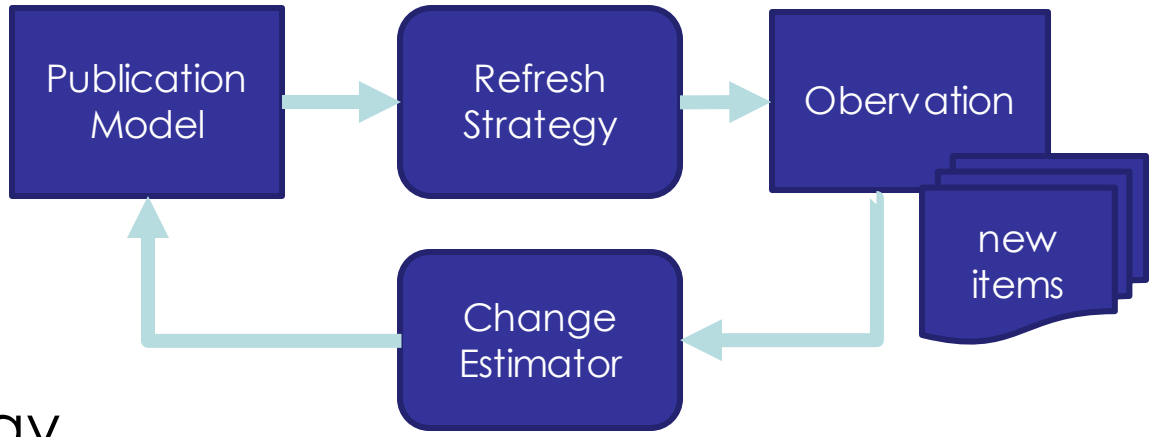
Maximize aggregation quality
- stream completeness (long-term information loss)
- window freshness (short-term outdated information)

Limited resources
- bandwidth
- storage
- memory or computing capacity

Highly dynamic content
- publication model
- change estimation

Refresh strategies

# Crawler architecture



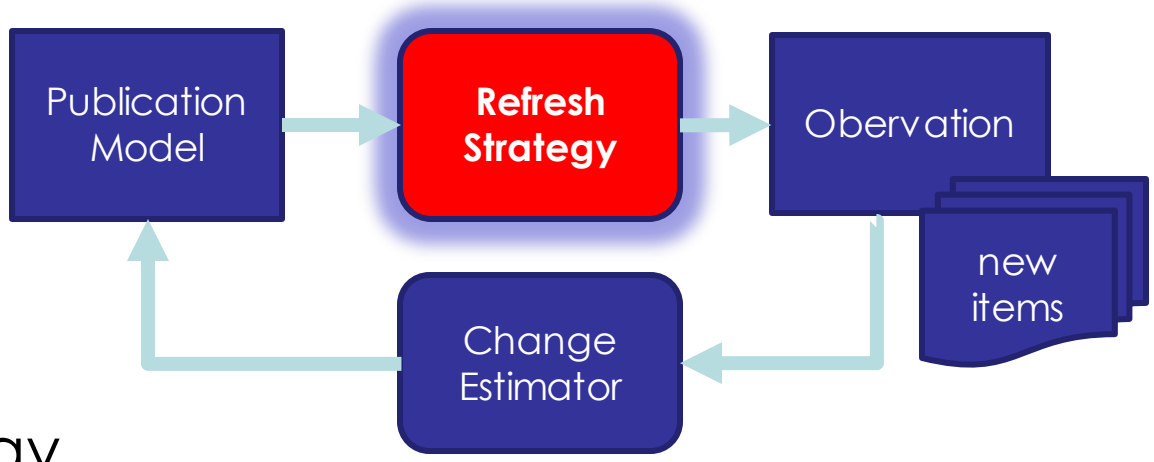

## Refresh strategy

uses source publication models in order to take refresh decisions

## Online change estimator

updates the source publication models based on the observations of the real source publication behaviors

# Crawler architecture




## Refresh strategy

uses source publication models in order to take refresh decisions

## Online change estimator

updates the source publication models based on the observations of the real source publication behaviors

# "Best effort" Refresh strategy

"Best effort" strategy (Lagrange Multipliers):

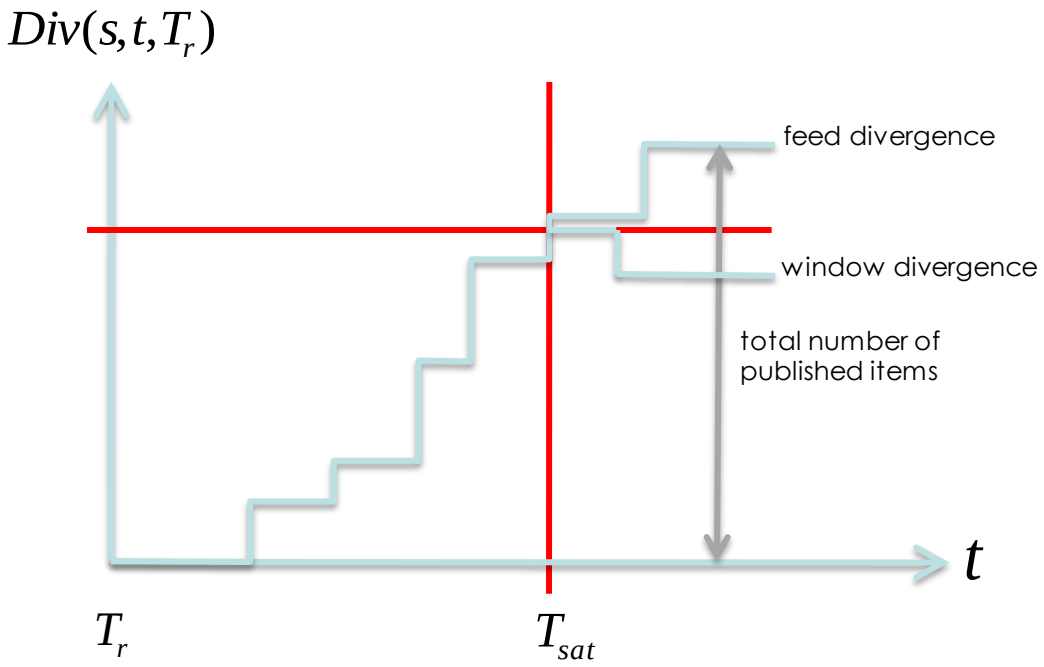
Let $Uti(s, q, t, T_r)$ be a *monotonic utility function* and $\tau$ a positive constant. At each time instant $t \geq Tr$, refresh all sources $s$ where $Uti(s, q, t, T_r) \geq \tau$.

→ obtains maximum quality compared with all other strategies that use an equal cost (number of refreshes).

Applications

web pages, cache synchronization, RSS (retrieval delay)

# Saturation & Monotonicity




## Saturation

feed divergence ≥ publication window size

## Divergence

stream – monotonic

window – non monotonic after saturation

# Refresh strategy for saturated and non saturated sources

Saturated sources : top-k divergence

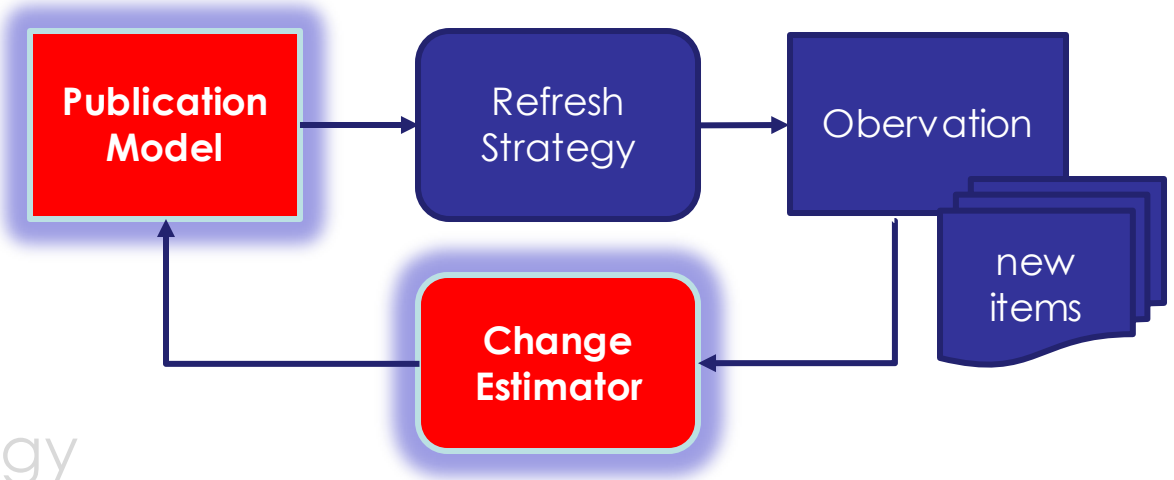
refresh sources with maximal divergence score

Non saturated sources : best effort

refresh sources with $Uti(s, q, t, T_r) \geq \tau$

→ obtains maximum utility compared with all other strategies that use an equal cost (average number of refreshes).

# Crawler architecture




Refresh strategy

uses source publication models in order to take refresh decisions

## Online change estimator

updates the source publication models based on the observations of the real source publication behaviors

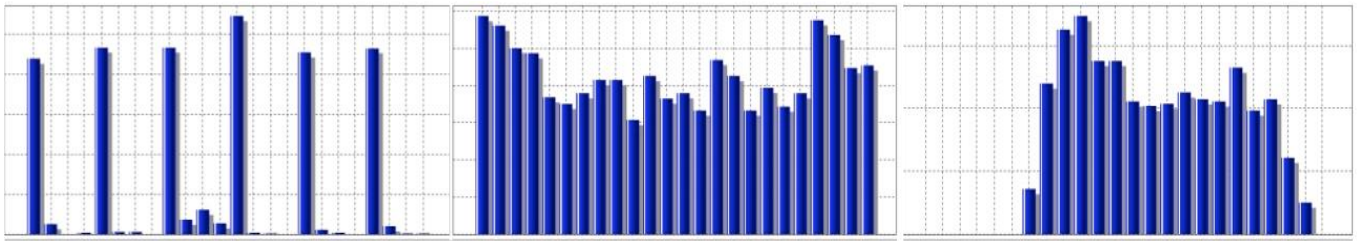
# Source publication activity analysis

RSS source feeds:

- 1658 random feeds over 4 weeks
- 963 feeds over 4 weeks : online French and international newspapers

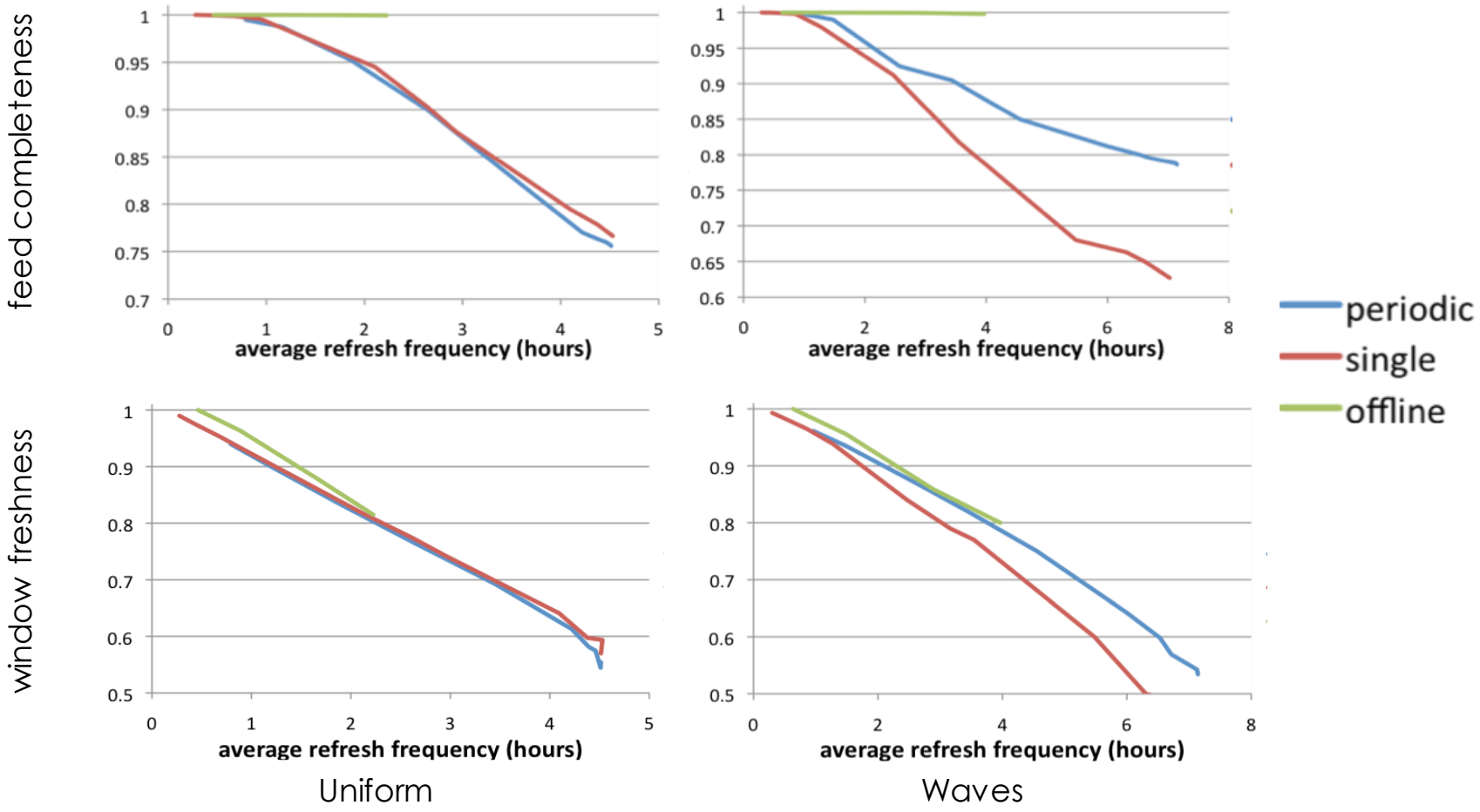
Shape discovery algorithm

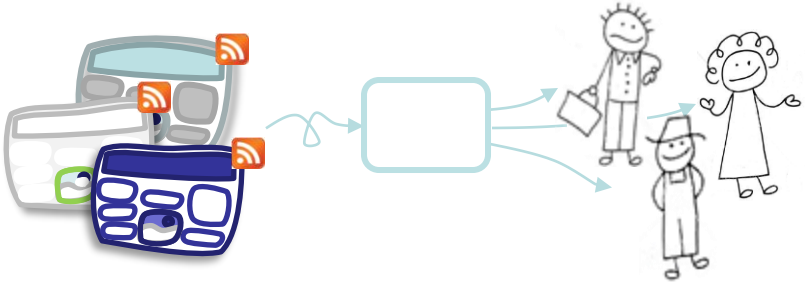
Publication shapes:



Peaks Uniform Waves

# Experiments
# "2 steps" Refresh strategy with Online estimation





# Meows: Continuous Top-k Filtering over the Real-time Web

**Nelly Vouzoukidou**[1], Bernd Amann[1], Vassilis Christophides[2],

[1] LIP6 – Université Pierre & Marie Curie

[2] University of Crete, Heraklion, Greece

[CIKM'2012]
[CIKM'2014 demo]


Meows Reader
Daily News
All my Meows
Malaysia Airlines(1)
breast cancer
android
premier league(2)
barack obama
Politics
Sports
Technology
International
MeowsReader > All my Meows > Malaysia Airlines
International
All my Meows
Another Dead End, But Malaysia Airlines Searchers Not Giving Up
The first objects fished out of the Indian Ocean that searchers hoped would give them a clue about the missing Malaysia Airlines jet turned out to be another dead end, but organizers of the international effort said they were "well, well short"
ABC News | Mon, 31 March 2013, 2:20PM
MH370: Hopes dashed as orange objects turn out to be fishing equipment
Kuala Lumpur, Malaysia (CNN) -- Potential leads on the missing Malaysian jetliner keep coming. So do the setbacks and frustrations. Monday's search ended without finding anything significant, Australian officials said...
CNN | Mon, 31 March 2013, 2:10PM
Race on to find missing plane's black box before it stops pinging
Royal Australian Air Force crew members read navigation maps during Thursday's search effort. Photo: Pool/Getty Images. MORE ON: Flight MH370 · Black box detecting
Trends:
Health care
Empire State Building
Crimea
"The Walking Dead"
Russia
Pervez Musharraf
Oculus
Flight 370
FedEx Lawsuit
Ice Storm Study
Software Patents

# Continuous Top-k Query Processing

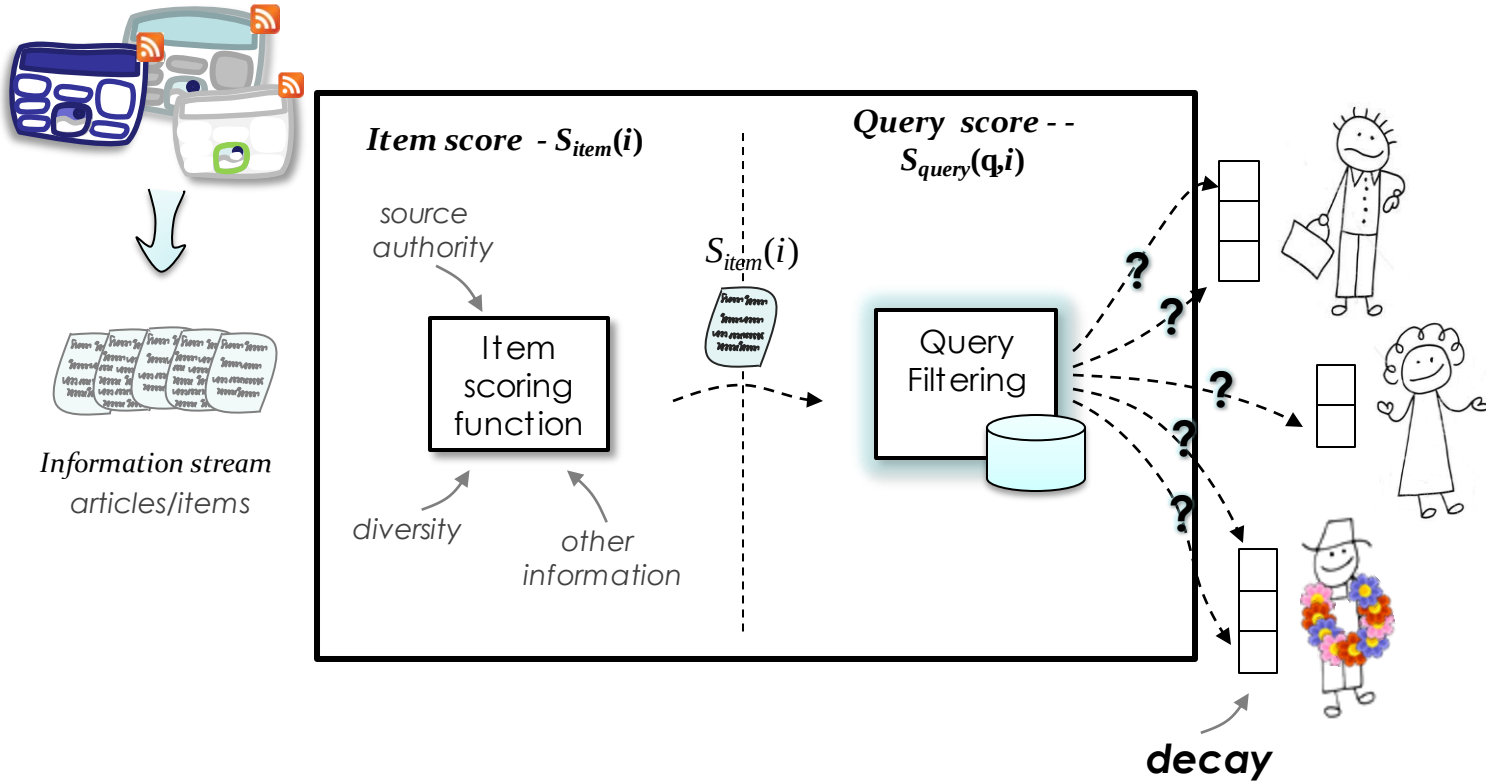
## Context :

- Continuous information retrieval in text streams (news, tweets)
- Top-k queries : produce top-k result wrt. some ranking function

## Contribution :

- Formal top-k continuous query filtering model
- Inhomogeneous ranking functions combining query-dependent (similarity) and query-independent (importance) item scores
- Efficient index structures for continuous top-k text queries with inhomogeneous ranking functions

# Inhomogenuous ranking function

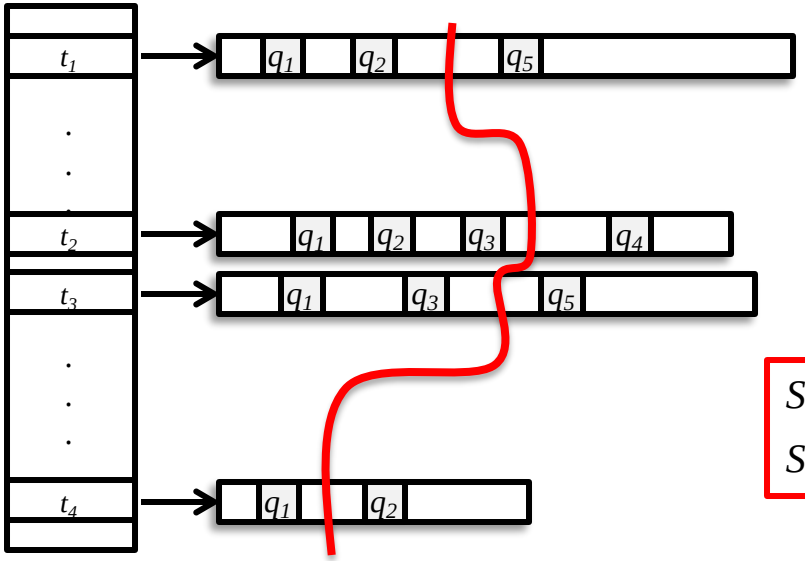



$$S_{total}(q, i) = \alpha \cdot S_{item}(i) + \beta \cdot S_{query}(q, i)$$

# State of the Art: COL-Filter Using Threshold Algorithm

State of art system (for homogenous score functions) :
COL-Filter [HMA2010]:

TA-like algorithm [Fagin01]

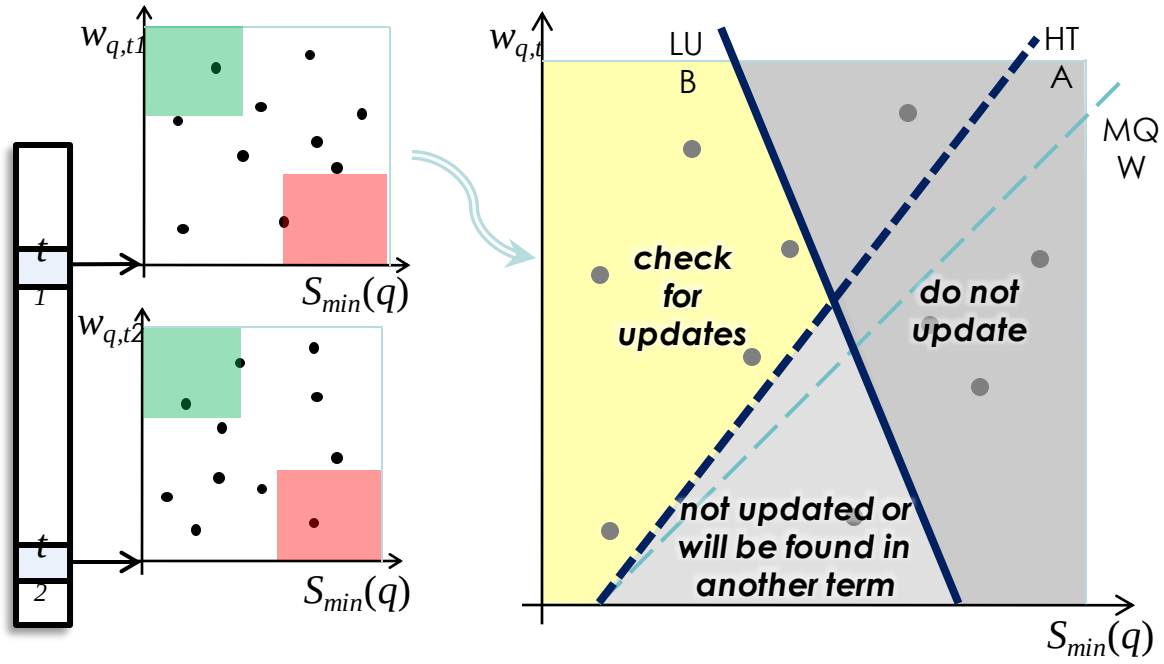



$q_1 = \{t_1, t_2, t_3, t_4\}$
$q_2 = \{t_1, t_2, t_4\}$
$q_3 = \{t_2, t_3\}$
$q_4 = \{t_2\}$
$q_5 = \{t_1, t_3\}$

$S_{total}(q, i) > S_{min}(q)$
$S_{total}(q, i) = \sum w_{q,t}\, w_{i,t}$

# Term Weights/Query Score Constraints

Filtering constraints depending on minimal query score and term weights



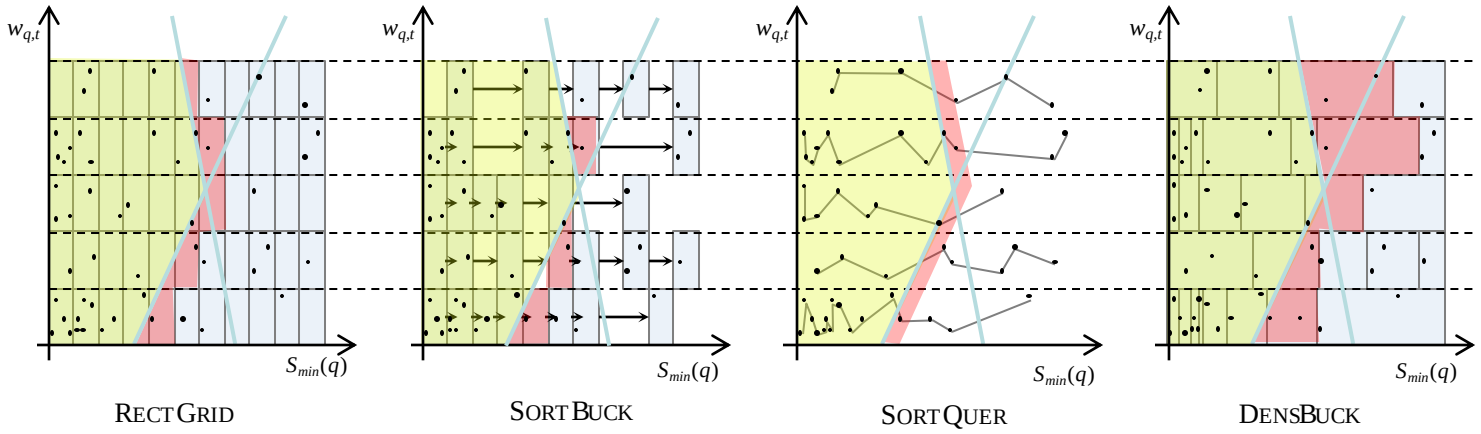
# Query Indexes

Spatial indexing problem

We consider variations of:

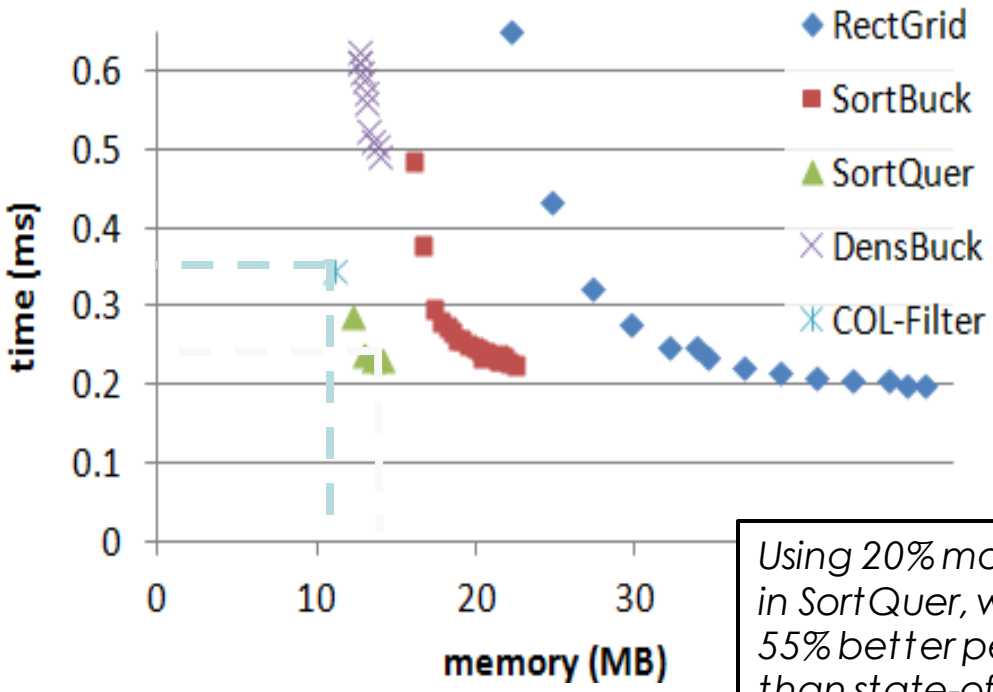
- Rectangular grids
- R-Trees

Different space, filtering and update time costs



# Experiments : Time/Space trade-off

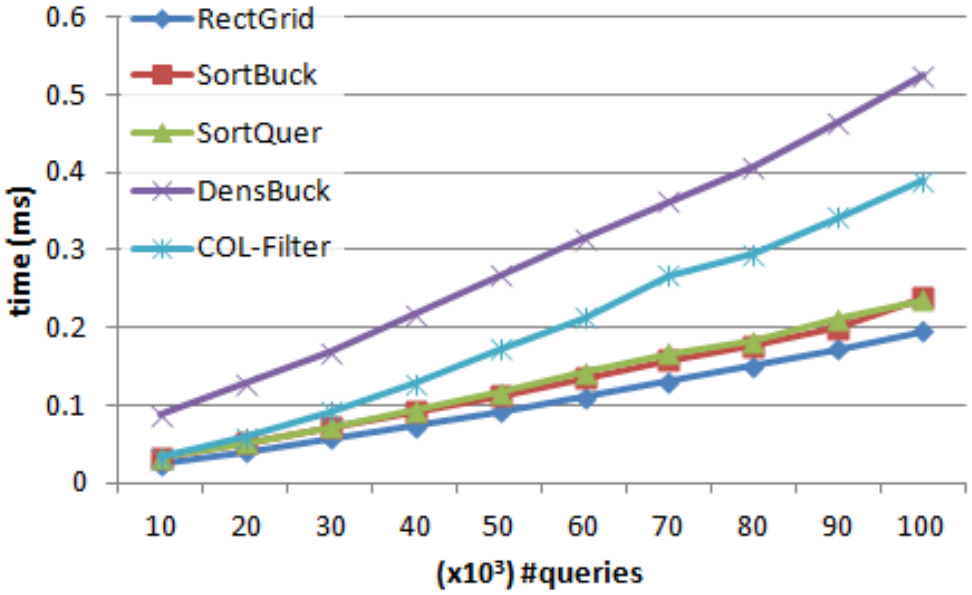
Average per item query detection time over number of stored queries



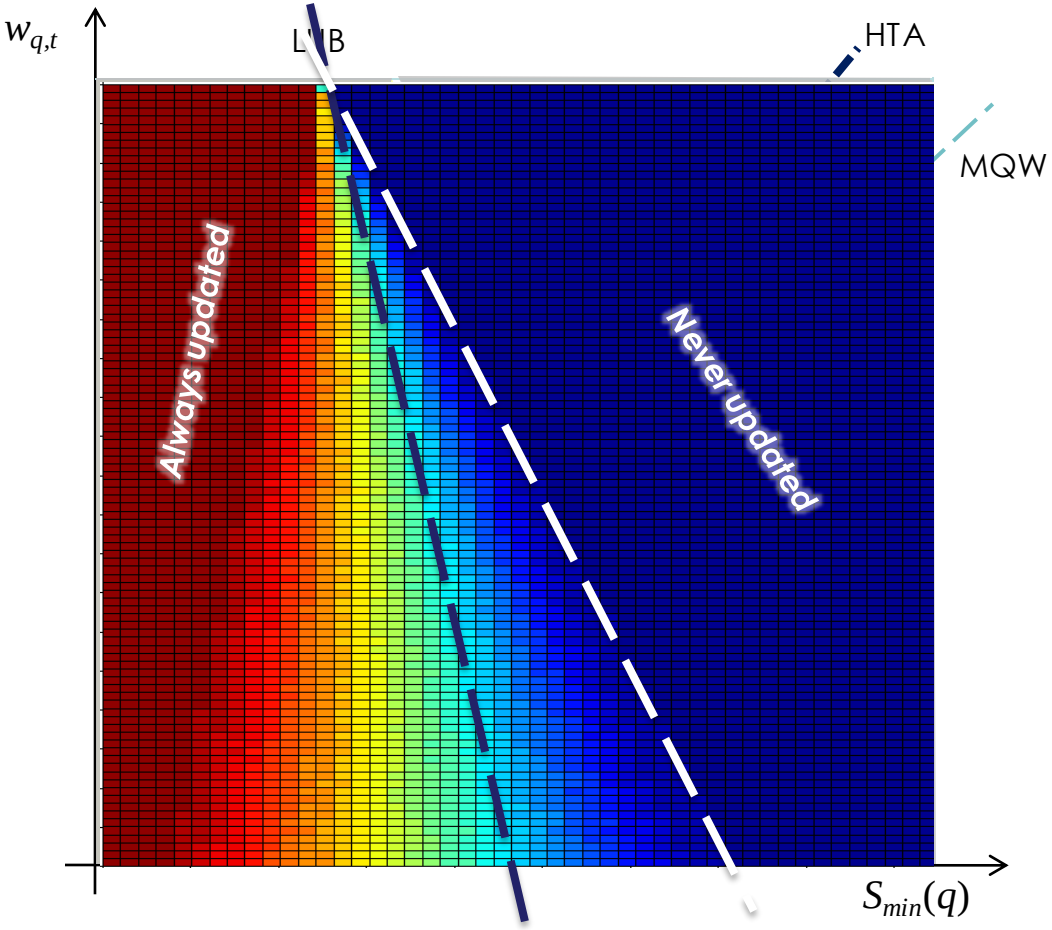

*Using 20% more memory in SortQuer, we achieve 55% better performance than state-of-art COL-Filter*

# Experiments : Query Scale

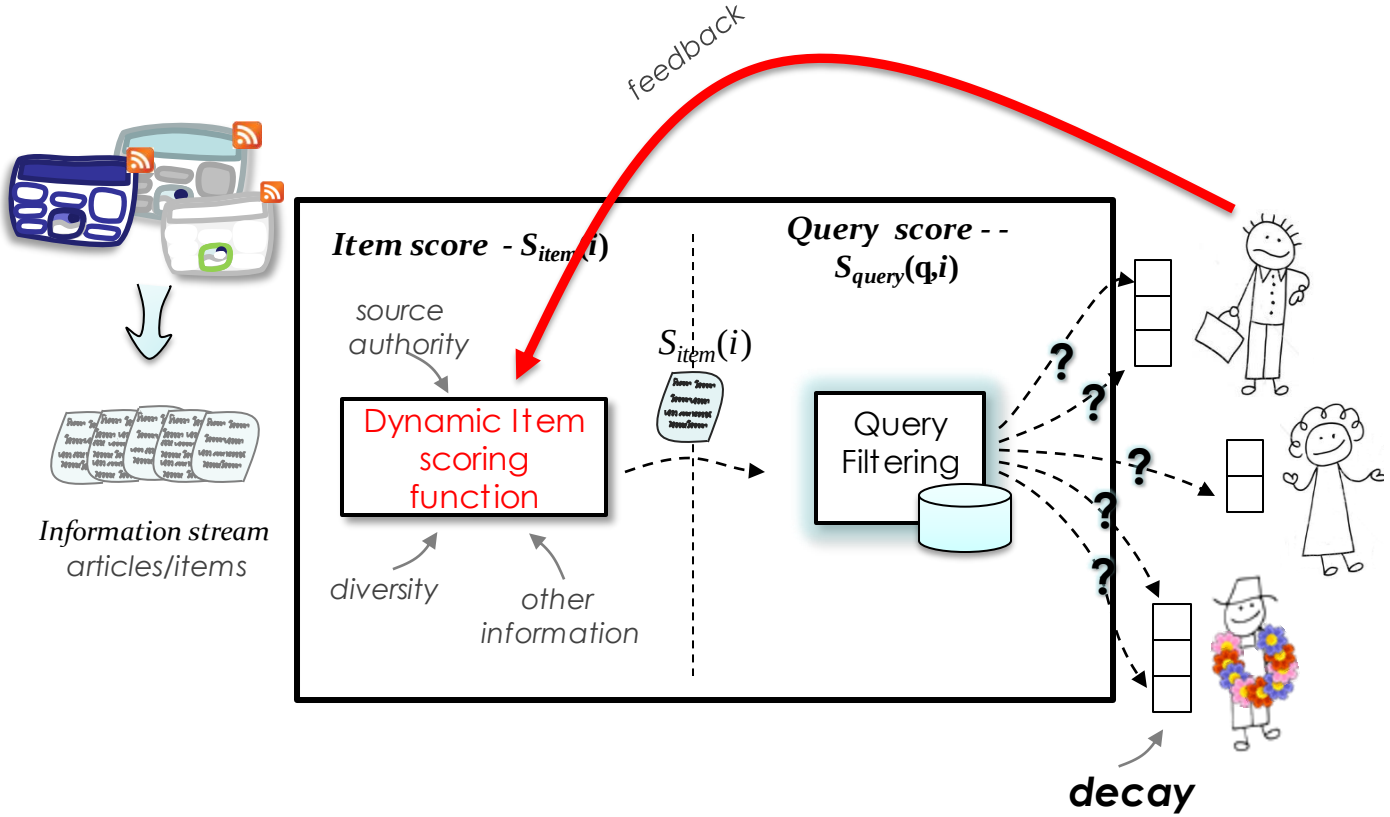
Average per item query detection time over number of stored queries



# Approximate Query Filtering



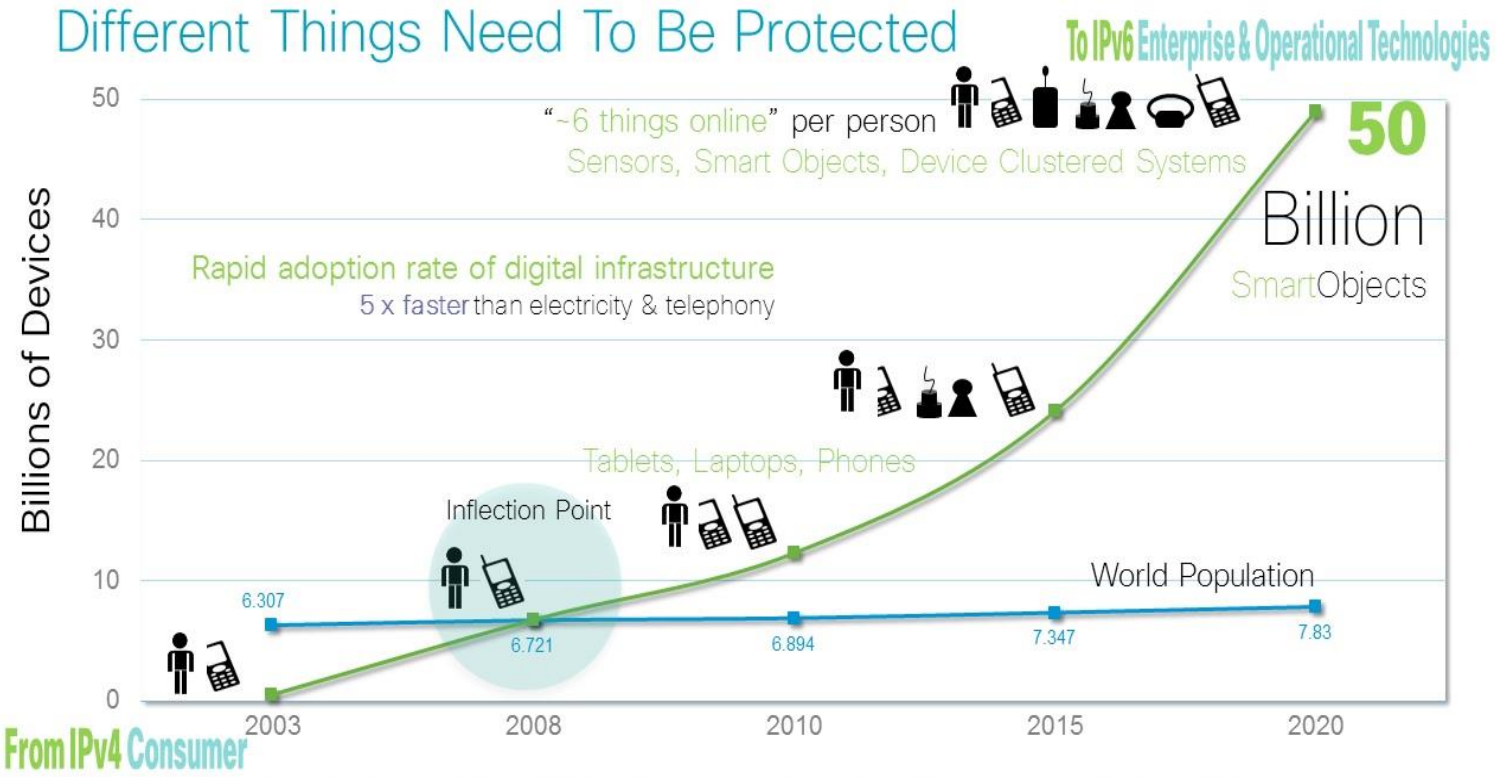
# Inhomogeneous dynamic ranking function




$$S_{total}(q, i) = \alpha \cdot S_{item}(i) + \beta \cdot S_{query}(q, i) + \gamma\, S_{feedback}(i)$$

# Perspectives and Conclusion

# Web of Things



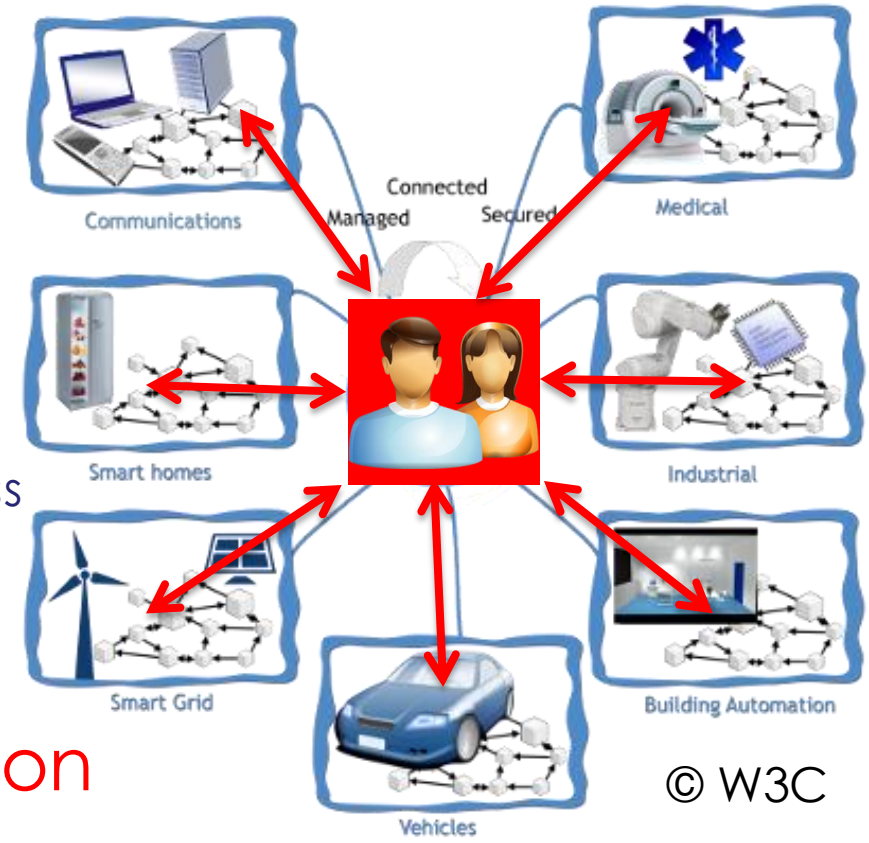

Source: Cisco IBSG projections, UN Economic & Social Affairs http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf

© CISCO

# What's next : "Ubiquitous" Web

Data and service Ubiquity
- universal connectivity
- communication
- pervasive computing
- data cloud

Real-time/real-world interaction
- "smart" car/home/city/industry
- personal/ambient/social awareness

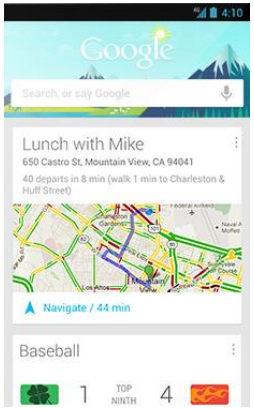



© W3C

Data and Service Explosion

# Data Dissemination in the "Ubiquitous" Web

- Data complexity
  - heterogeneity (contents, structure, semantics)
  - metadata : source, annotation
  - user context : social, spatial, …
- Data quality
  - correctness
  - completeness
  - timeliness
- Data control
  - privacy
  - security




- High-dimensional ranking and recommendation
  - contents, space, time, semantics, feedback, …
- Data & stream processing
  - new hybrid architectures
  - new algorithms
- Provenance
  - why/where/how provenance
- Logics and semantics
  - reasoning and verification
- User interfaces
  - interaction
  - publication & subscription
  - administration

Thank you for your attention!