City Data Fusion

http://citydatafusion.org

DeRiVE 2013 Workshop
21.10.2013, ISWC 2013, Sydney, Australia

Listening to the pulse of our cities during City Scale Events

Emanuele Della Valle
emanuele.dellavalle@polimi.it
http://emanueledellavalle.org

Share, Remix, Reuse — Legally

Agenda

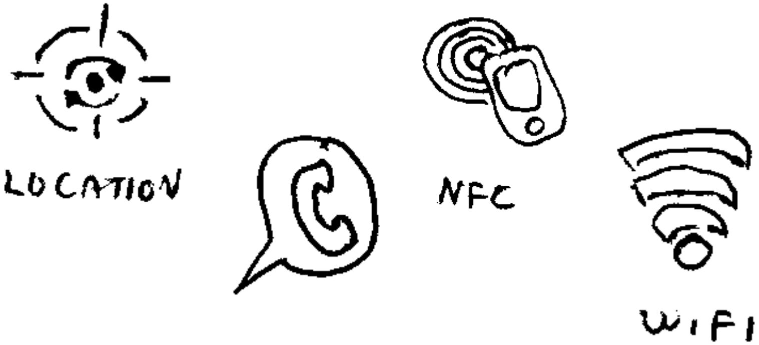
- Context
- Ingredients and challenges
- Research hypothesis
- Streaming Linked Data Framework
  - design principles
  - Architecture
  - Components
- Testing the research hypothesis
  - London Olympic Games 2013
  - Milano Design Week 2013
- Conclusions

The digital reflection of reality is sharpening

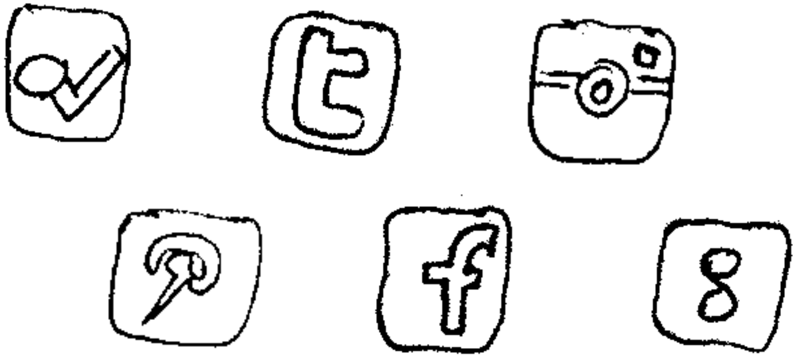
Streams of information flows through our cities thanks to:

the pervasive deployment of sensors in our cities

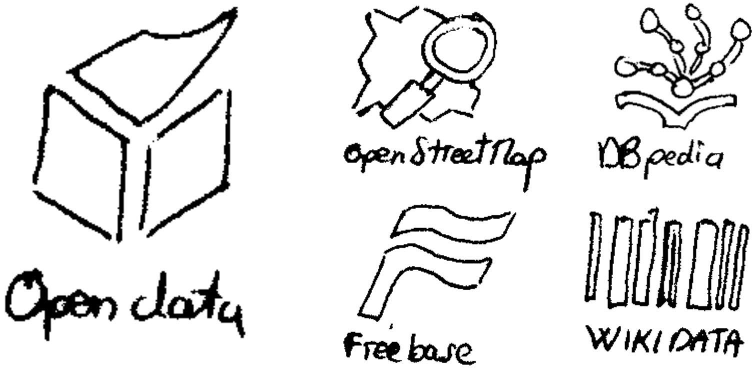
the wide adoption of smart phones (equipped with sensors)



the usage of (location-based) social networks

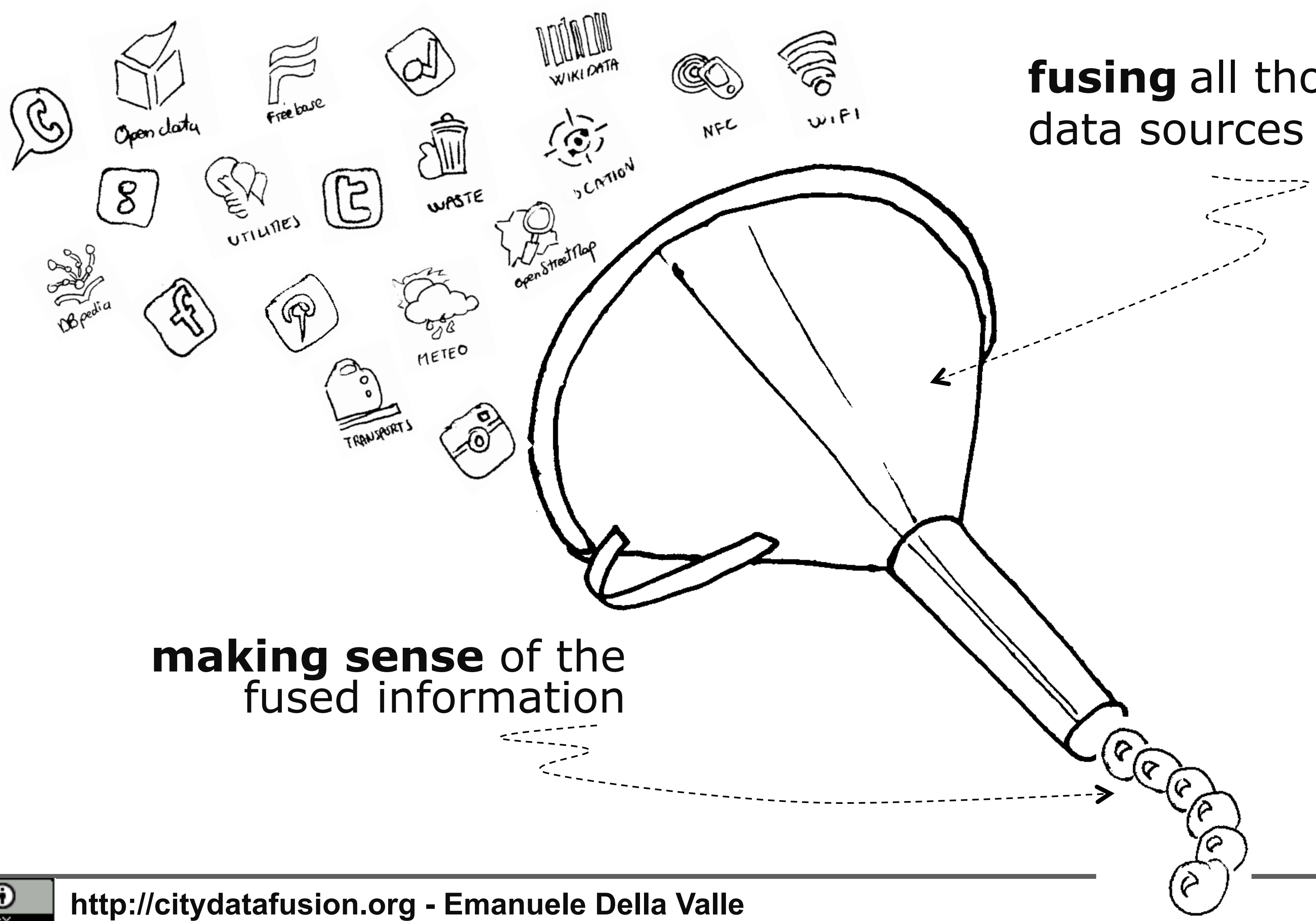


the availability of datasets about urban environment



For the first time in human history

- **We can feel the pulse of our cities by**



fusing all those data sources

making sense of the fused information

City Scale Events as test beds

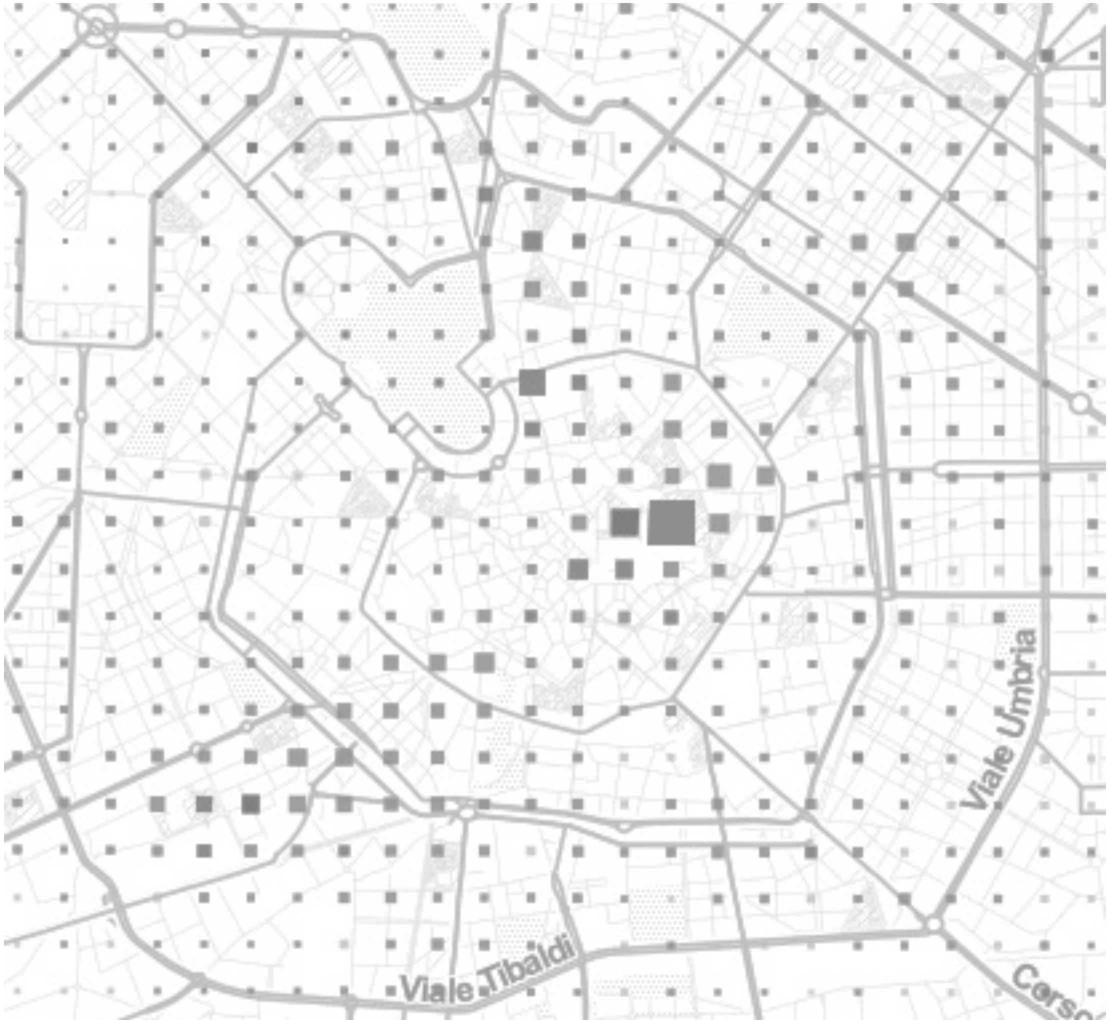
- Characteristics
  - Lasting days
  - Hundreds of venues
  - Thousands of events
  - Hundreds of thousands of visitors
- Questions
  - Which are the most attractive events?
  - What do visitors think about the events they join?
  - What is their mood before, during and after the event they join?
- Ground truth
  - The program of the event
  - News about the event

Example of City Scale Event
The Milano Design Week 2013 (MDW'13)

- Problem
  - Sponsor and organizer of a city scale event needs to quantify the return on investment
- Existing solutions
  - Spread people around the hundreds of event to asses the success of the various events is effective, but expensive
- Challenge
  - Obtaining comparable results by analysing public social streams

E.g., is Milano Design Week perceivable?

Step 1: associate mobile traffic to urban areas

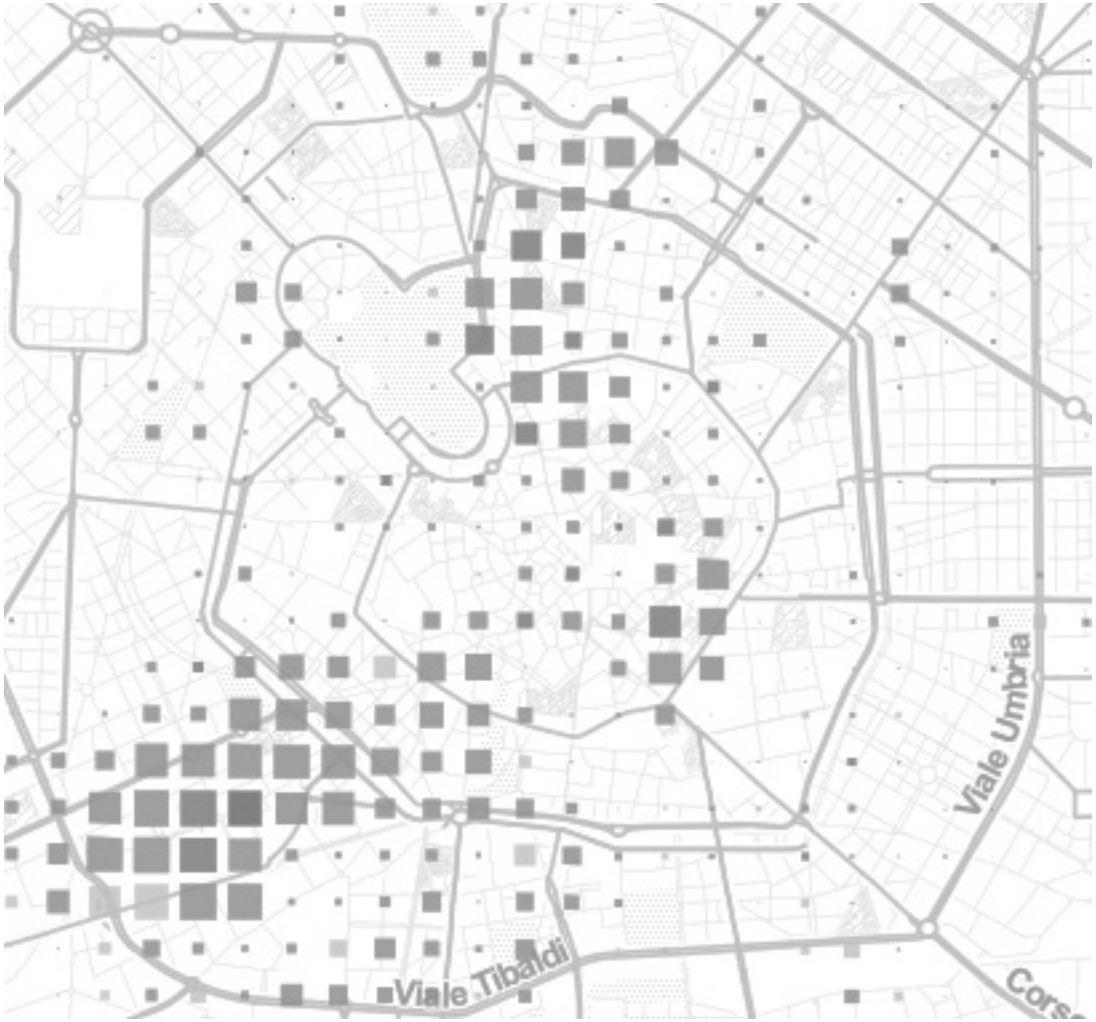


Real data recorded on 13 April 2013 between 13:00 and 00:00

E.g., is Milano Design Week perceivable?

LOCATION

Step 2: subtract what is systematic

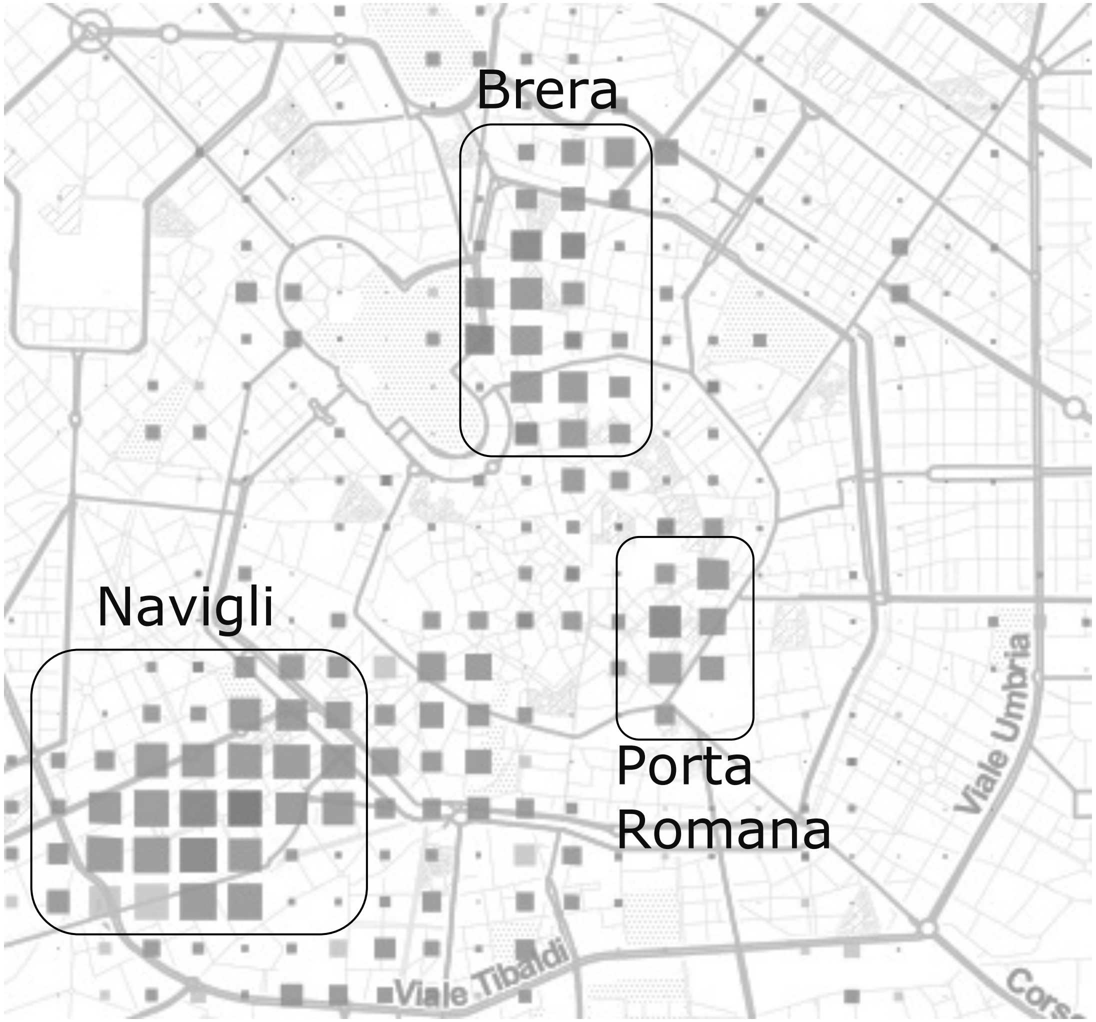


Real data recorded on 13 April 2013 between 13:00 and 00:00

OpenStreetMap

E.g., is Milano Design Week perceivable?

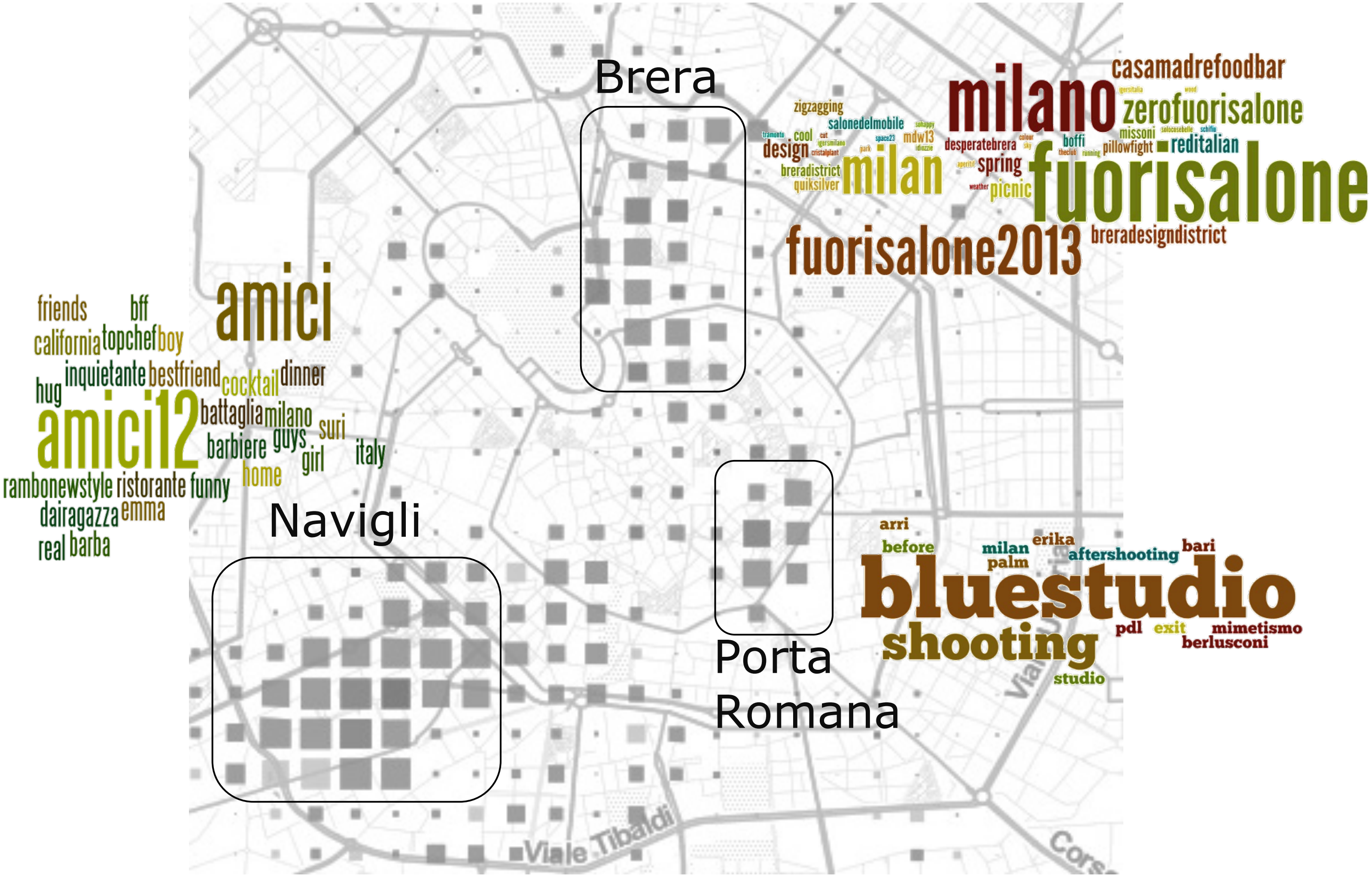
Step 3: Identify interesting areas



Real data recorded on 13 April 2013 between 13:00 and 00:00

E.g., is Milano Design Week perceivable?

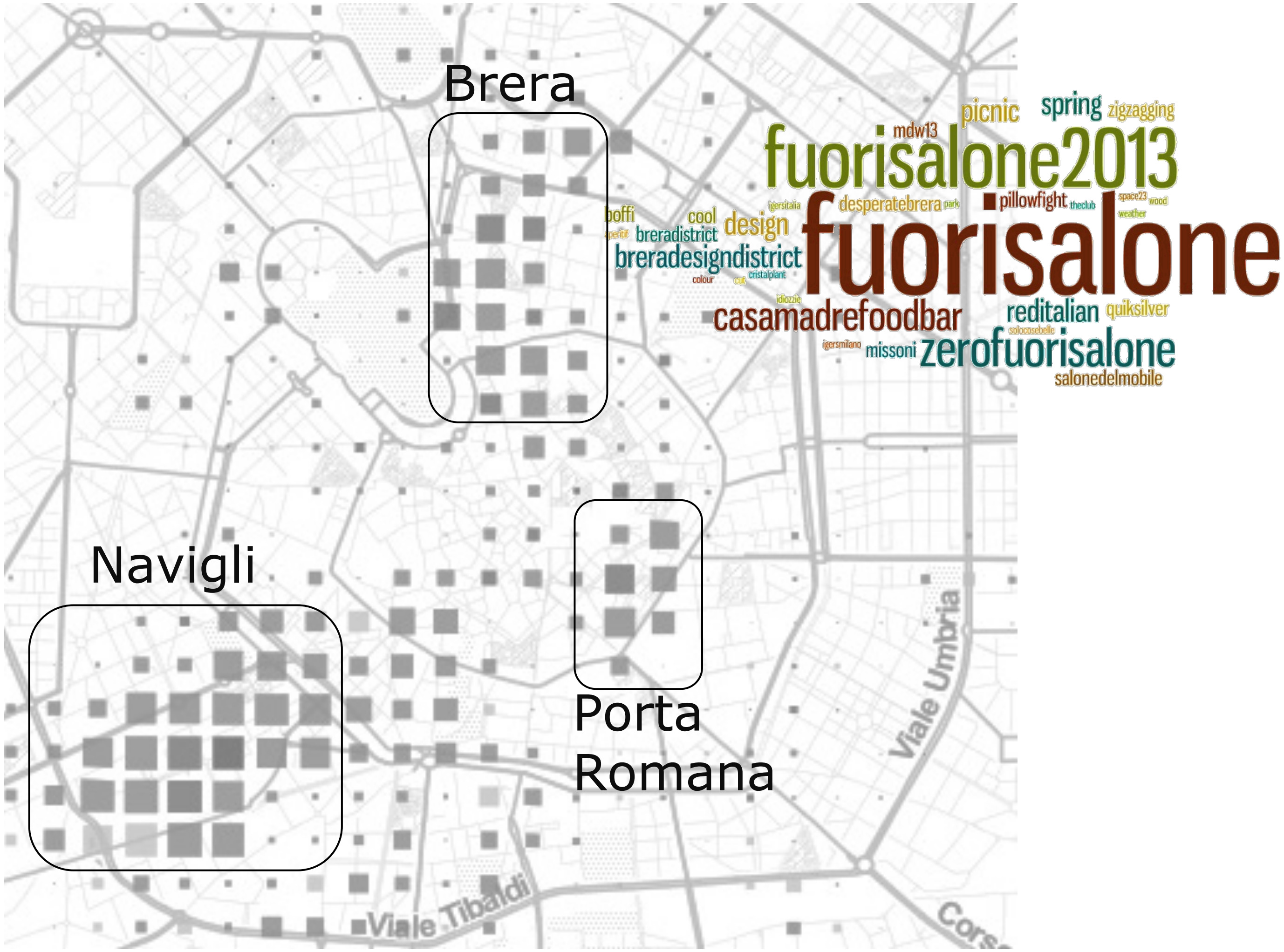
Step 4: retrieve the top hashtags



Real data recorded on 13 April 2013 between 13:00 and 00:00

E.g., is Milano Design Week perceivable?

Step 5: exclude what is systematic



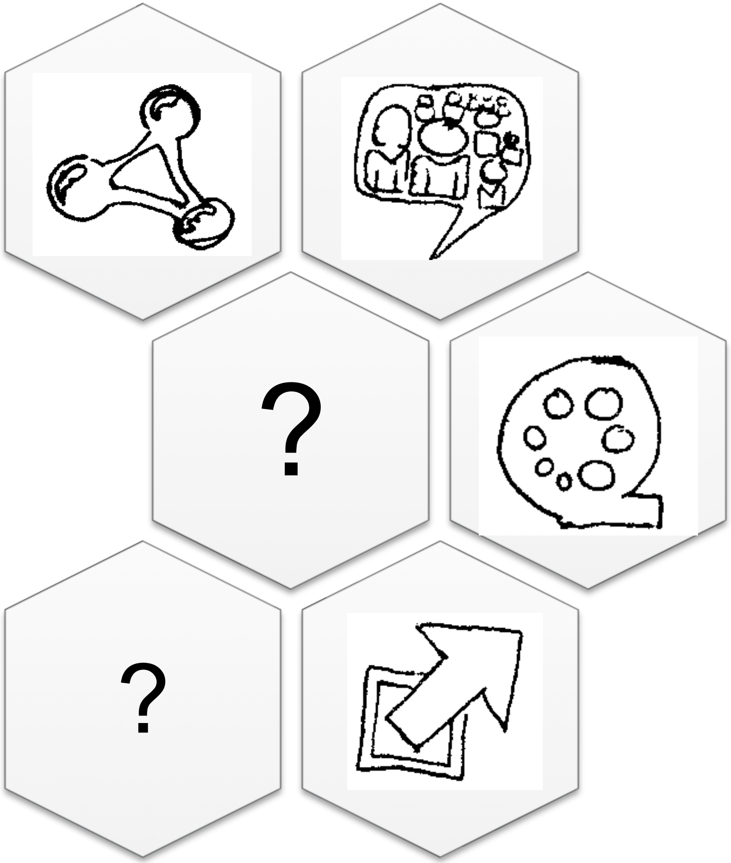
Real data recorded on 13 April 2013 between 13:00 and 00:00

Ingredients to be combined

- **semantic technologies**
  - Address "*variety*" using Ontology Based Data Access
  - Named Entity recognition and linkage
  - Knowledge discovery (e.g., detecting systematicy)
- **streaming algorithms**
  - Address "*velocity*" of data stream
  - Address "*volume*" by being able to process data that do not fit in main memory
- **crowd-sourcing techniques**
  - Address "*veracity*" by cleansing and enriching data
- **Visual analytics**
  - Allow no-expert access to data
  - Tell stories out of data

Limitation of current systems

- Insufficient methods for making sense in **real-time** of **heterogeneous** data and social streams w.r.t. the vast collections of (open) data
- Lack of crowd-sourcing techniques whose **incentives** leverage needs of people in the urban environment
- Lack of visualization techniques tailored to **non-experts**

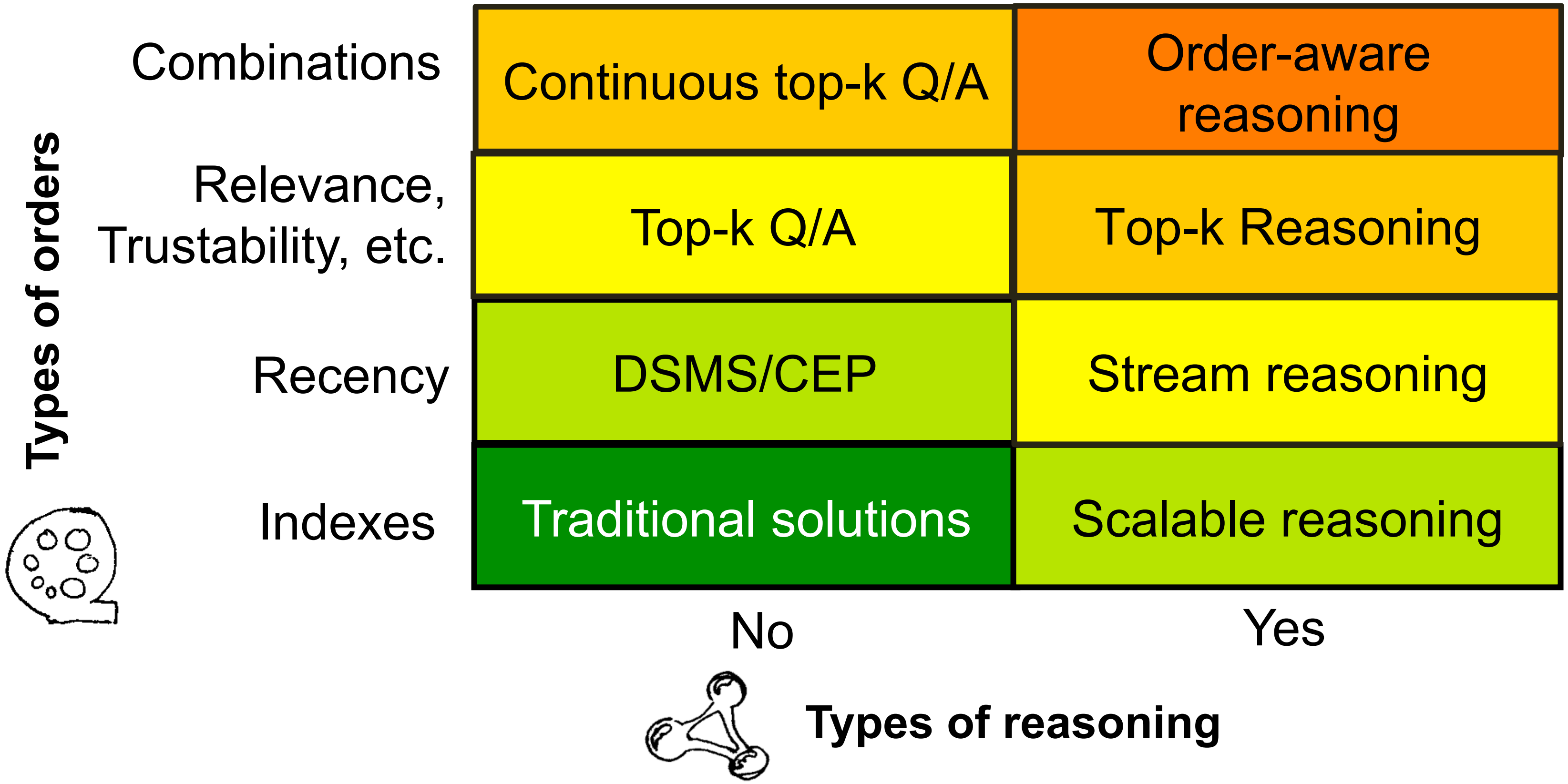


Research hypothesis

1. To scale order matters
2. Crowdsourcing needs the urban-centric incentives
3. Visualization must tell stories

Research hypothesis: order matters!

- Observation: order reflects recency, relevance, trustability …



Types of orders	Types of reasoning: No	Types of reasoning: Yes
Combinations	Continuous top-k Q/A	Order-aware reasoning
Relevance, Trustability, etc.	Top-k Q/A	Top-k Reasoning
Recency	DSMS/CEP	Stream reasoning
Indexes	Traditional solutions	Scalable reasoning

- harnessing orders is key to make sense in real-time of heterogeneous, massive and volatile data

Research hypothesis: urban-centric incentives!

[source: http://www.behance.net/gallery/Maslows-Hierarchy-of-Needs-Infographic/4376921]

Maslow's hierarchy of needs

SELF-ACTUALIZATION:
Creativity, Spontaneity, Problem Solving, Lack of Prejudice, Acceptance of Facts
ESTEEM:
Self Esteem, Confidence, Achievement, Respect of Others
#1
BEST EVER
LOVE/BELONGING:
Friendship, Family, Sexual Intimacy
SAFETY:
Security of body: Employment, Resources, Morality, the Family, Health, Property
zzZ
XXX
PHYSIOLIGOCAL:
Breathing, Food, Water, Sex, Sleep, Homeostasis, Excretion

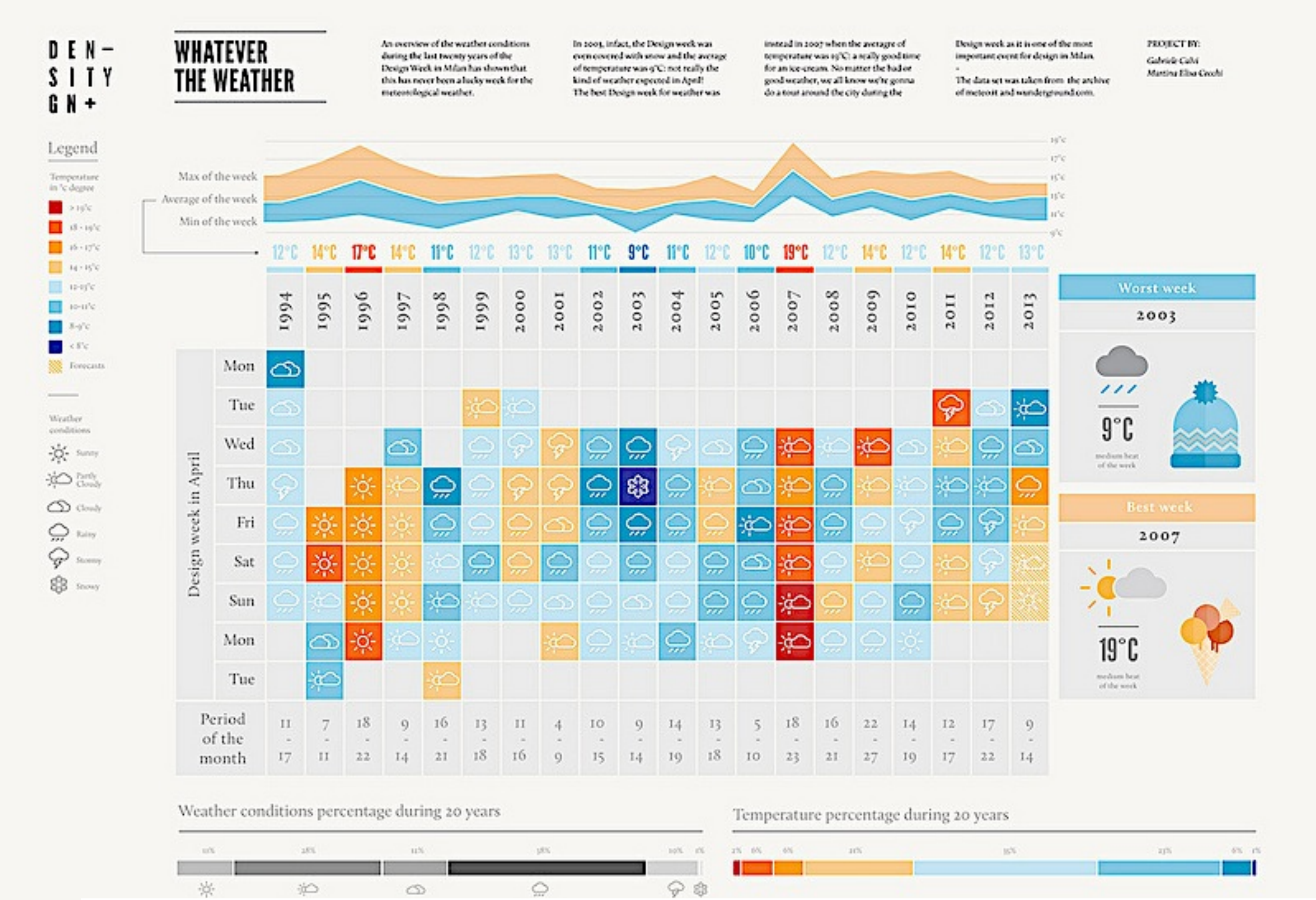
INNOCENTIVE®

amazon® mechanical turk

ReCAPTCHA™

- **incentives** designed **for urban environment and life styles** are key

Research hypothesis: visualizations must tell stories



An overview of the weather conditions during the last twenty years of the Design Week in Milan has shown that this has never been a lucky week for the meteorological weather.

In 2003, infact, the Design week was even covered with snow and the average of temperature was 9°C: not really the kind of weather expected in April! The best Design week for weather was instead in 2007 when the average of temperature was 19°C: a really good time for an ice-cream. No matter the bad or good weather, we all know we're gonna do a tour around the city during the Design week as it is one of the most important event for design in Milan.

-

The data set was taken from the archive of meteoit and wunderground.com.

PROJECT BY:

Gabriele Calvi
Martina Elisa Cecchi

[Source: http://www.densitydesign.org/2013/04/whatever-the-weather/]

Testing the research hypothesis

1. To scale order matters
   - Stream Reasoning
     - RDF Stream
     - Continuous SPARQL
     - Incremental Materialization for RDF Streams (IMaRS)
     - C-SPARQL Engine
     - RESTFul Services for RDF Stream Processors
     - Streaming Linked Data Framework
   - SPARQL Rank
     - Rank aware SPARQL algebra
     - ARQ-Rank
2. Crowdsourcing needs the urban-centric incentives
   - Urban Games With A Purpose
     - UrbanMatch
     - Urbanopoly
3. Visualization must tell stories
   - On going work

Focus of this key note

1. To scale order matters
   - Stream Reasoning
     - RDF Stream
     - Continuous SPARQL
     - Incremental Materialization for RDF Streams (IMaRS)
     - C-SPARQL Engine
     - RESTFul Services for RDF Stream Processors
     - Streaming Linked Data Framework
   - SPARQL Rank
     - Rank aware SPARQL algebra
     - ARQ-Rank
2. Crowdsourcing needs the urban-centric incentives
   - Urban Games With A Purpose
     - UrbanMatch
     - Urbanopoly
3. Visualization must tell stories
   - On going work

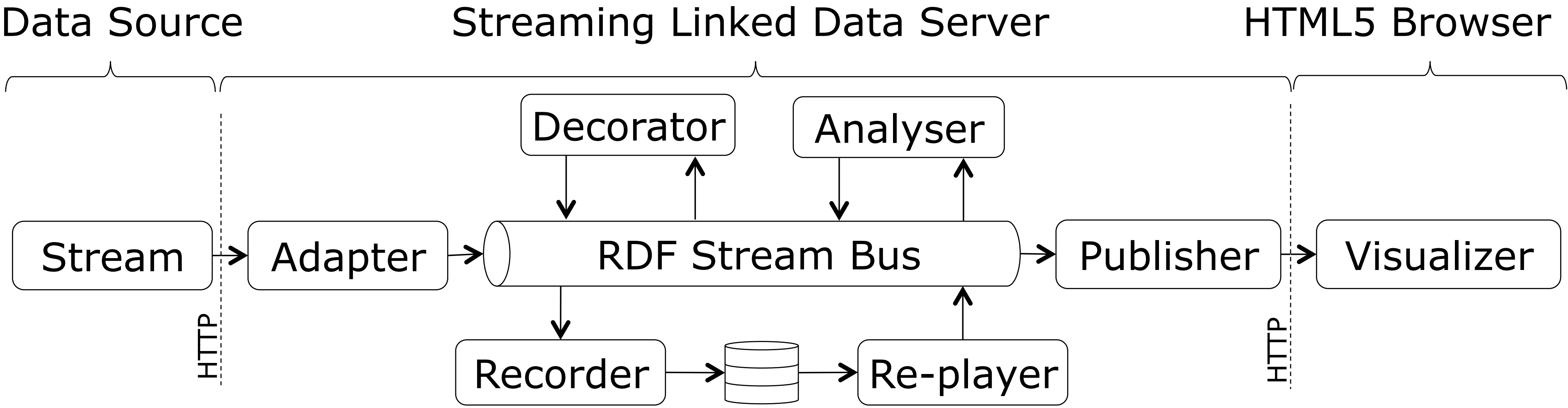
Streaming Linked Data Framework

- Input Data Formats
  - Streaming information: RDF streams
  - Background Information: RDF graphs
- Query Language
  - Continuous SPARQL
- Features
  - Adapters to access the social streams, e.g., twitter
  - Ability to record and replay portions of the social stream
  - Possibility to decorate the social stream with sentiment information
  - Possibility to continuously analyzing the social stream
  - Possibility to built complex application composing complex networks of decorators and analyzers
  - Possibility to publishing and visualizing results of continuous analysis

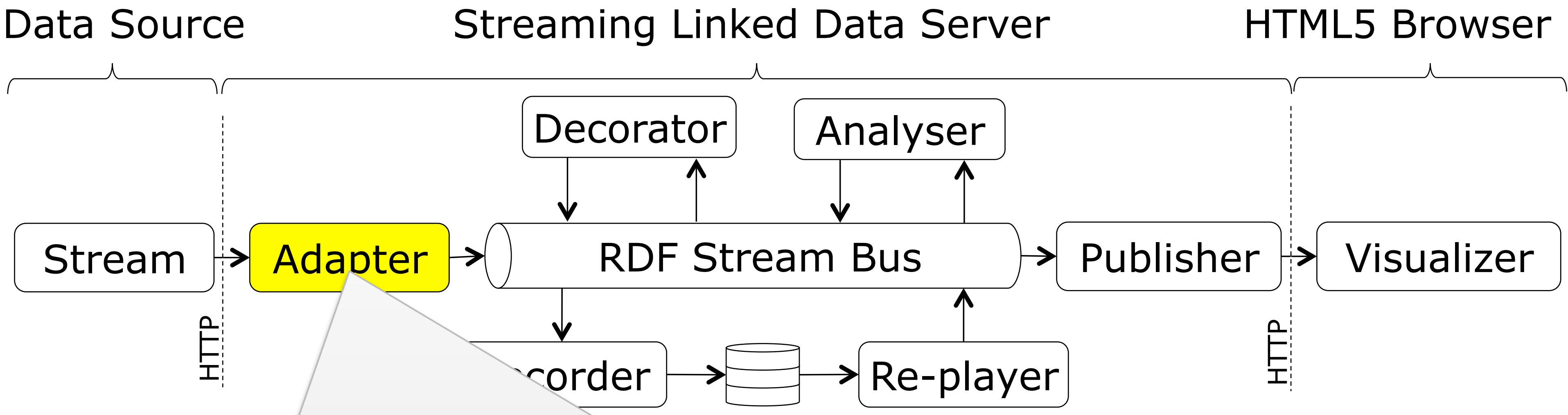
Design Principles

- Follow publish/subscribe patter
  - Publisher and subscribers do not have to know each other
  - Subscribers can join and leave in any moment
- Adopt a reliable message-passing
  - Guarantees delivery order
- Minimise latency by using main memory
  - Avoiding disk I/O bottleneck

Architecture



Architecture - Adapters

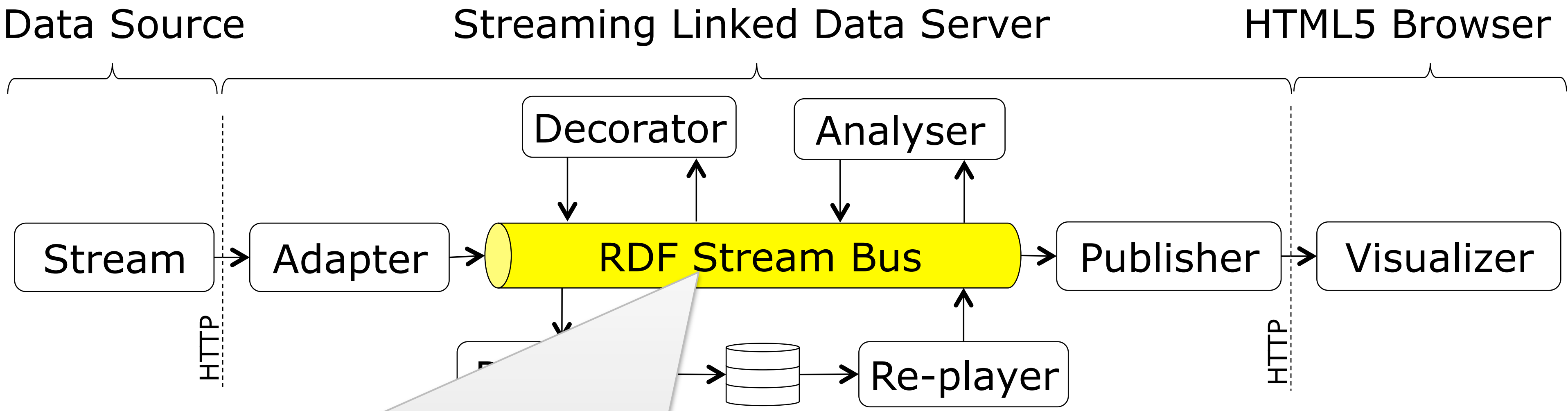


An adapter streams time-stamped RDF graphs

```
[] sioc:content "This is for everyone #london2012 #oneweb #openingceremony";
    sioc:has_creator :timberners_lee;
    sioc:topic :london2012, :oneweb, :openingceremony .
```

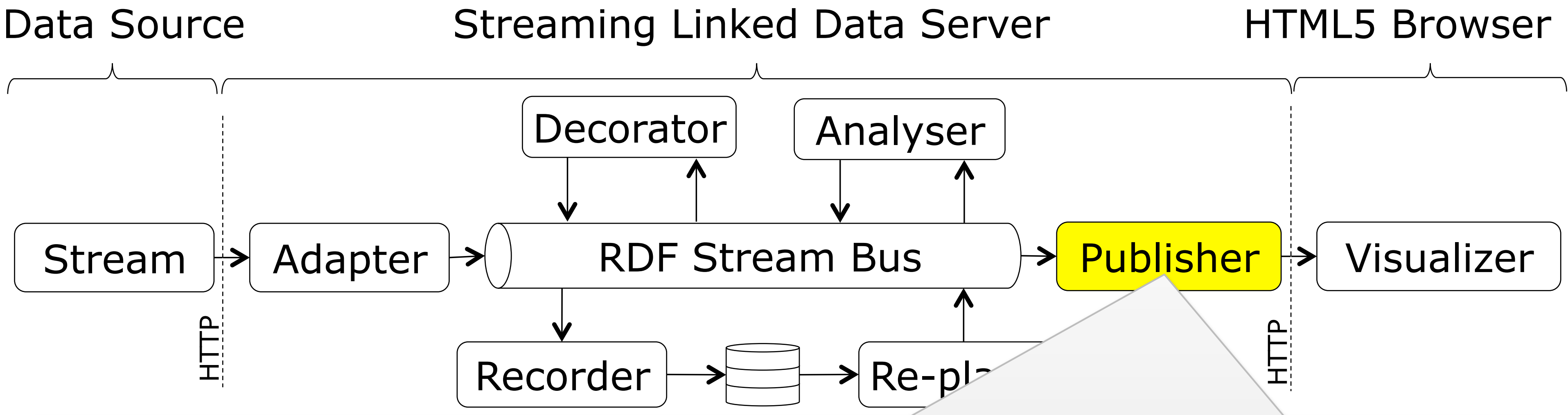
Available adapters include: twitter, instagram, pachube, and linked sensor data

Architecture – RDF stream Bus

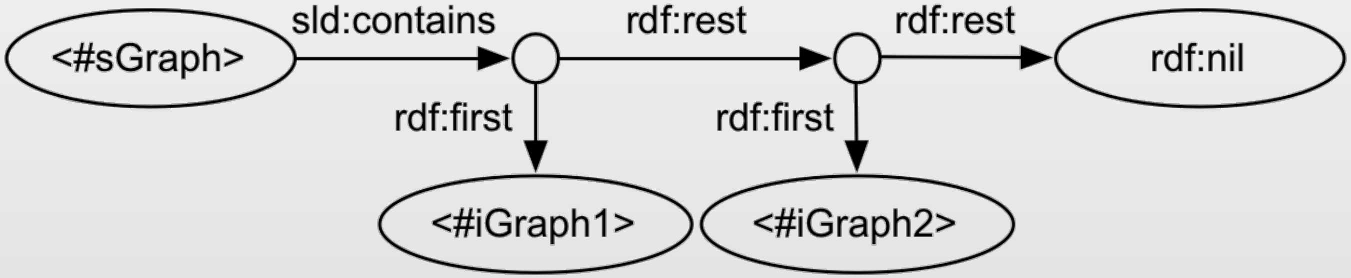


The RDF Stream Bus supports the publish/subscribe communication

Architecture - Publishers



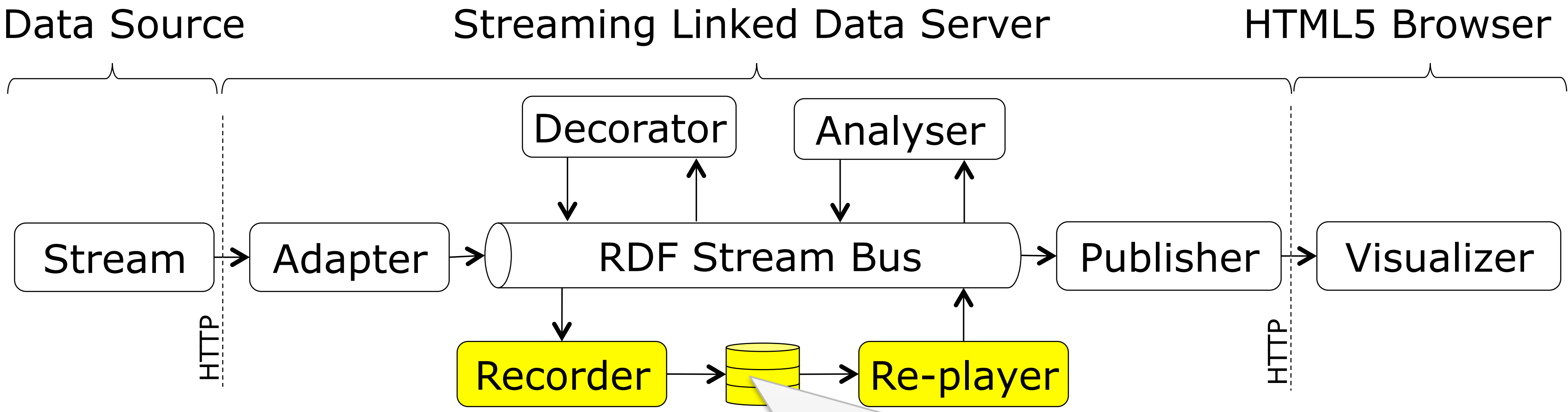
A publisher make available on the Web the last RDF graphs of an RDF stream following a variation of Streaming Linked Data format*:



* Barbieri, D.F., Della Valle, E.: A proposal for publishing data streams as linked data - a position paper. In: LDOW. (2010)

Streaming Linked Data Framework

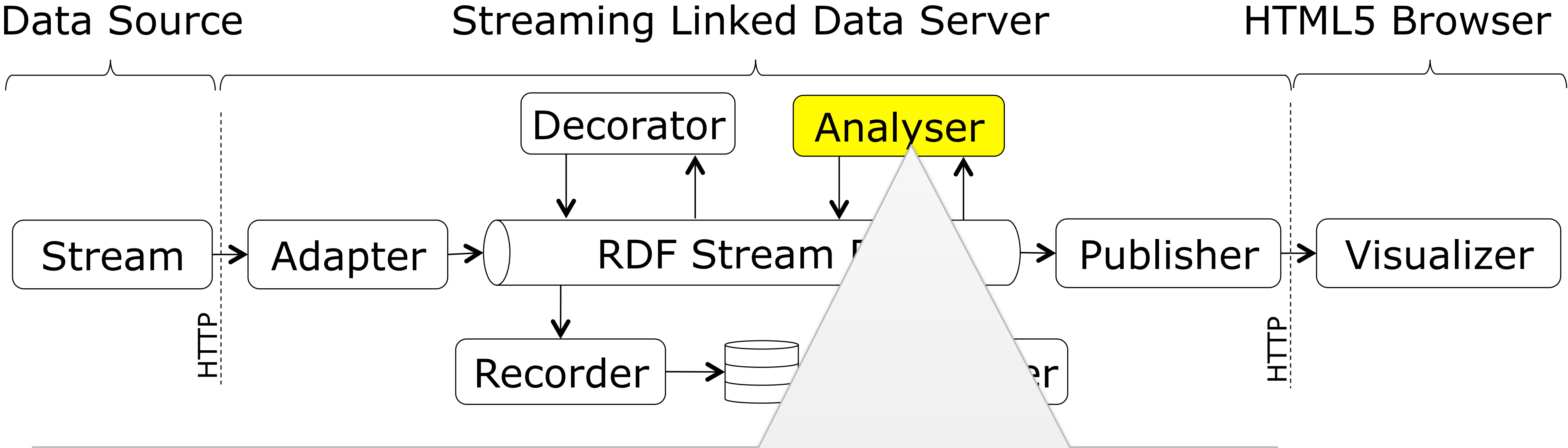
Architecture – Recorders and Re-players



A recorder records the content of an RDF stream following a variation of Streaming Linked Data format.
A re-player re-plays a recorded stream. It can vary the speed.

Streaming Linked Data Framework

Architecture - Analysers

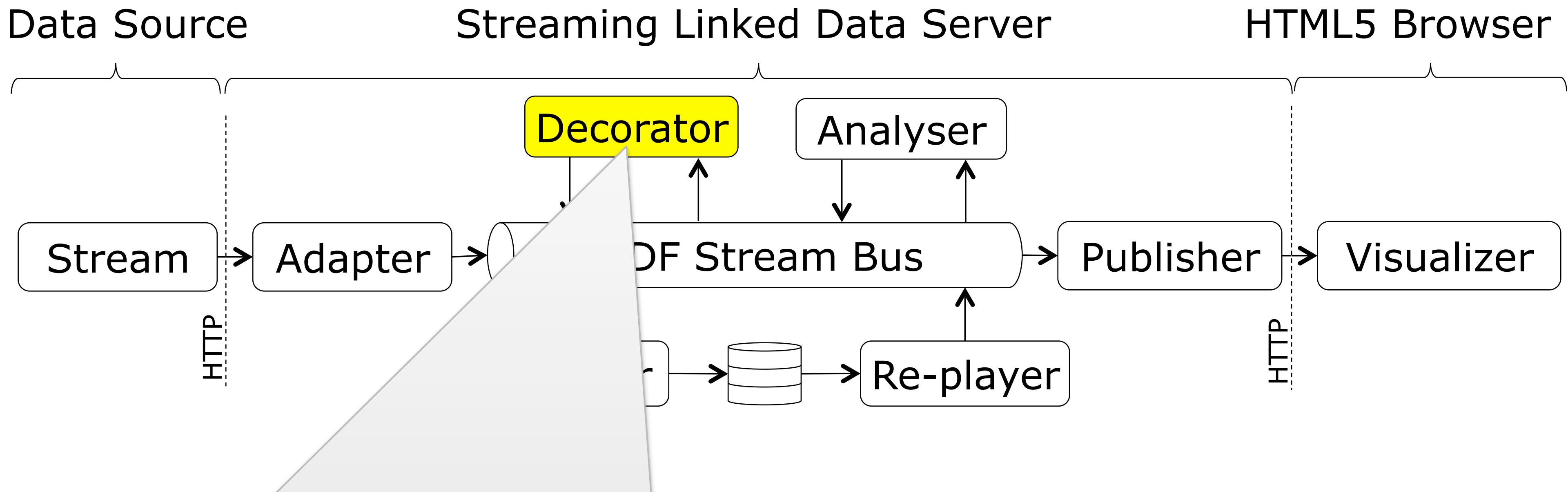


An analysers continuously execute C-SPARQL queries.
E.g., *count the number of times each hashtag is used in the last 15 minutes updating the counting every minute*.

```
REGISTER STREAM HashtagAnalysis AS
CONSTRUCT { [] sld:about ?tag ; sld:count ?n . }
FROM STREAM <http://.../London2012> [RANGE 15m STEP 1m]
WHERE { { SELECT ?tag (COUNT(?tweet) AS ?n)
          WHERE { ?tweet sioc:topic ?tag . } GROUP BY ?tag } }
```

Streaming Linked Data Framework

Architecture - Decorators

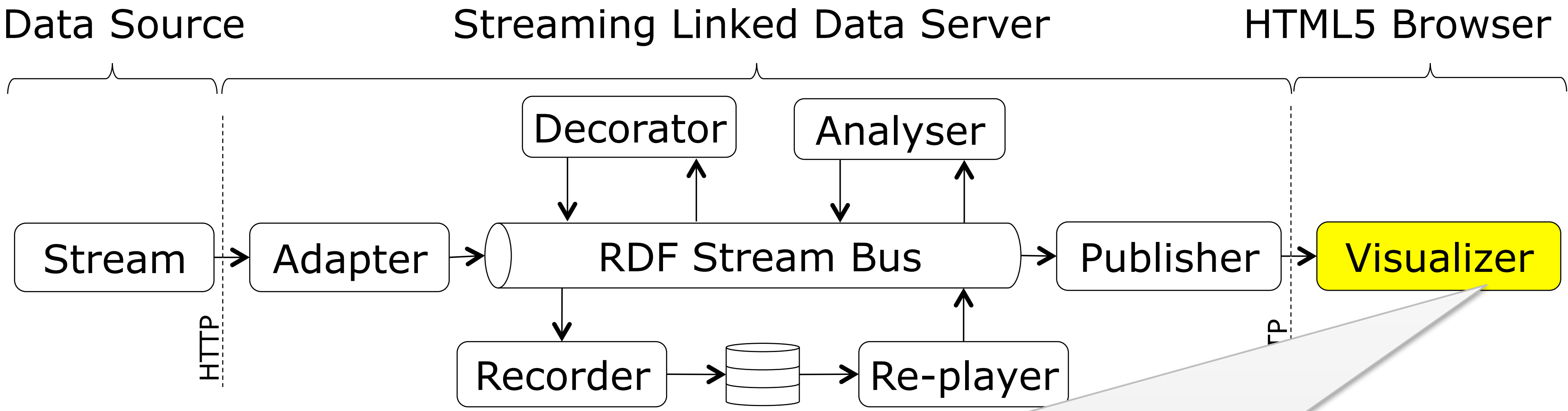


A decorator adds information to streamed RDF graphs that match a given patter.
E.g., Dictionary-based sentiment classifiers* (known to be efficient for short texts concentrating on a single topic, such as tweets) was used in this work to decorate each tweet.

* Tsytsarau, M., Palpanas, T., Denecke, K.: Scalable Detection of Sentiment-Based Contradictions. In: DiversiWeb workshop, WWW, Hyberabad, India (2011)

Streaming Linked Data Framework

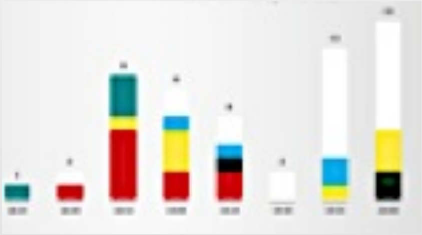
Architecture - Visualizers



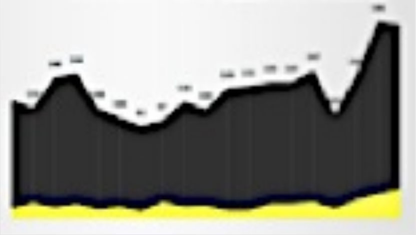
A visualizer displays the published linked data.
Available visualizers include:

Heatmaps

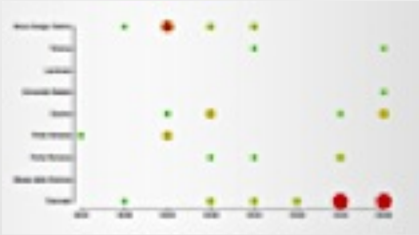
Bar charts



Area charts



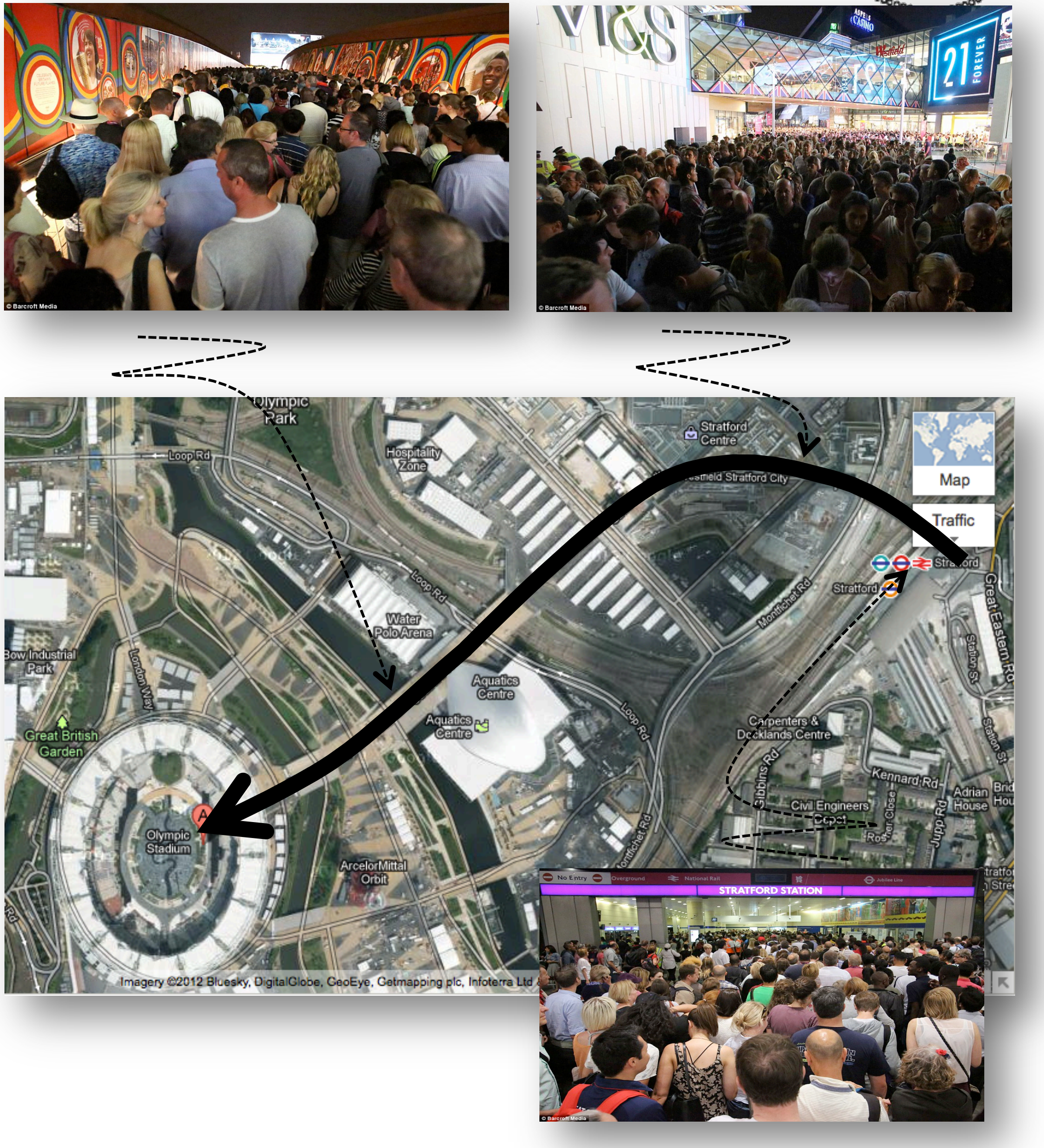
Dot charts



Testing the research hypothesis

The London Olympic Games 2012 (LOG'12)

- Problem
  - To manage a big event requires tracking in real time the movement of crowds.
- Existing solutions
  - CCTV, and mobile network data analysis are effective, but expensive
- Challenge
  - Obtaining comparable results by analysing public social streams



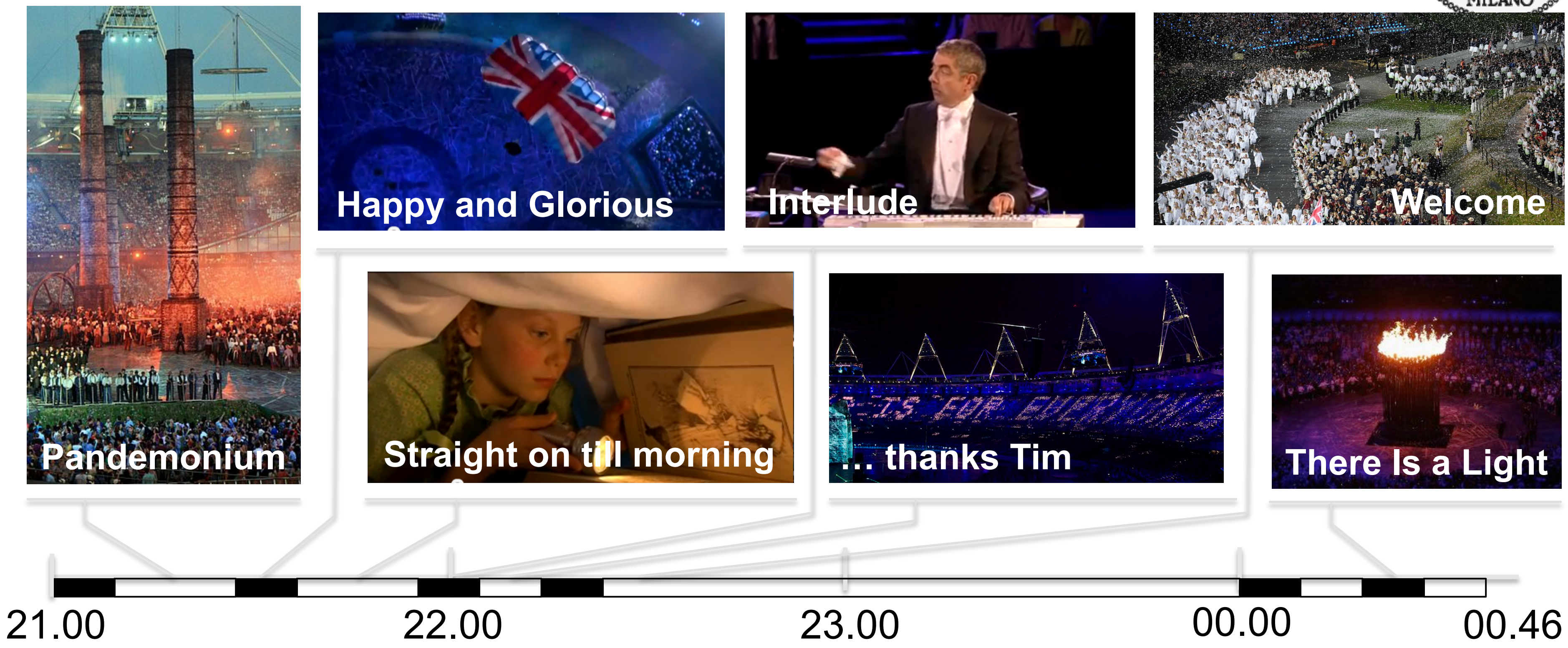
Test bed LOG'12

Case study #1 tracking attention of the crowds

- The problem
  - Managers of Big event want to track in real time if the event is capturing the attention of the audience.
- Input
  - 20 thousand of tweets streamed by Twitter between 9pm July 27th and 1am July 28th
- Ground truth
  - The Opening Ceremony Broadcast
    - http://www.youtube.com/watch?v=4As0e4de-rI
  - The wikipedia page describing the Opening Ceremony
    - http://en.wikipedia.org/wiki/2012_Summer_Olympics_opening_ceremony

Test bed LOG'12 – study case 1

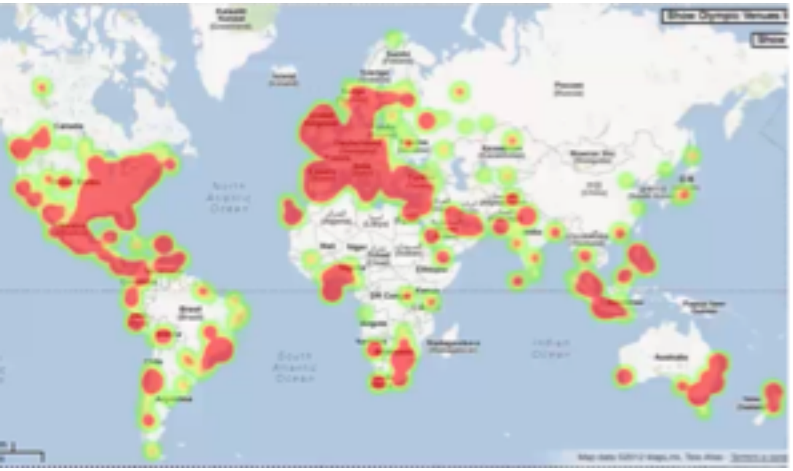
The Opening Ceremony



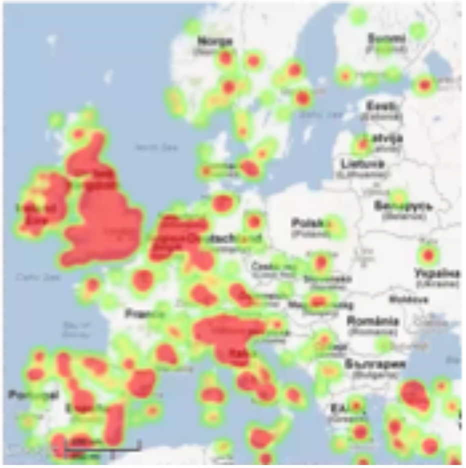
Test bed LOG'12 – study case 1

Findings

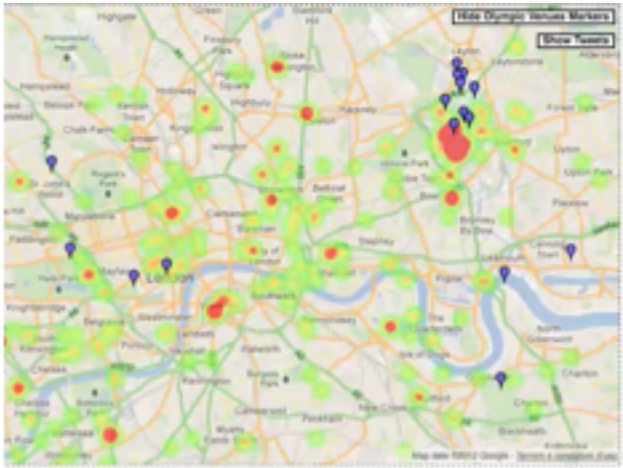
- Interesting phenomena are visible at different scales



World



Continent



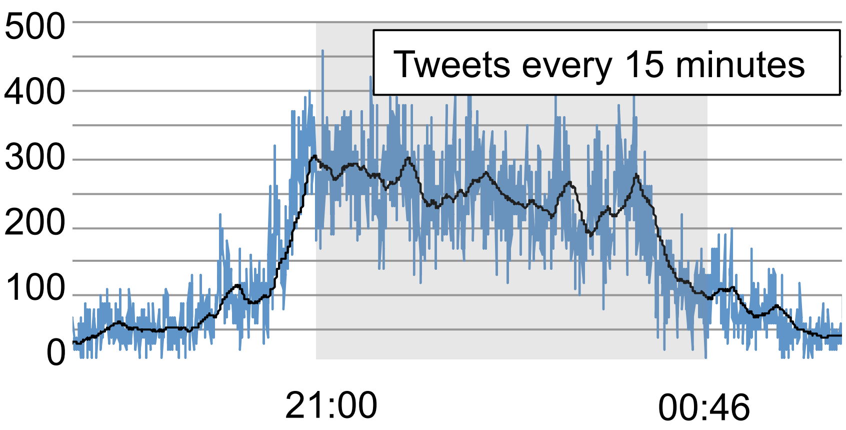
City

☺

Test bed LOG'12 – study case 1

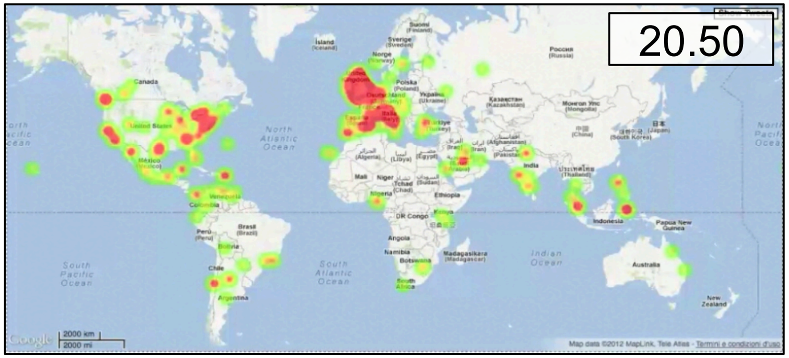
Findings

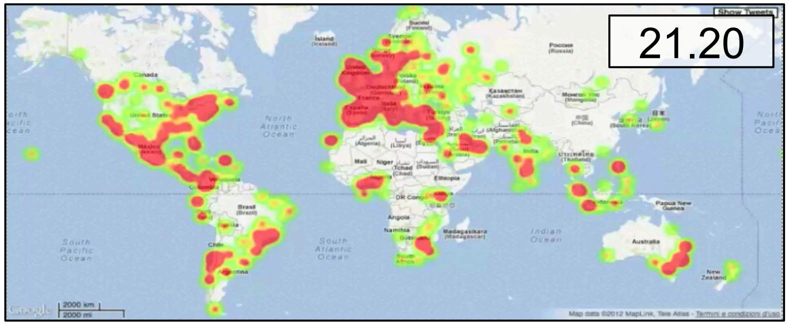
- The Opening Ceremony is clearly visible in the volume of tweets containing LOG'12 related hashtags

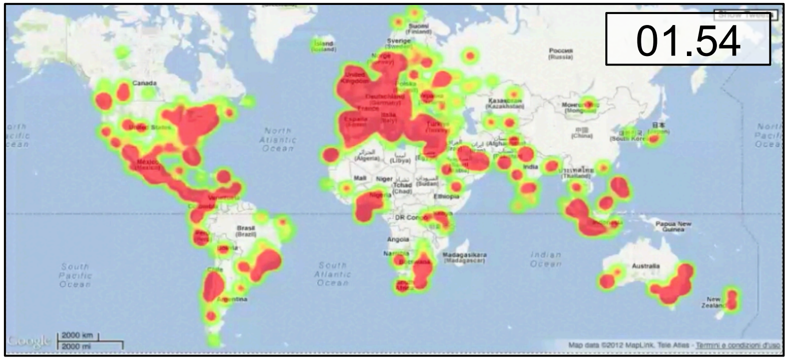


☺

- Global scope

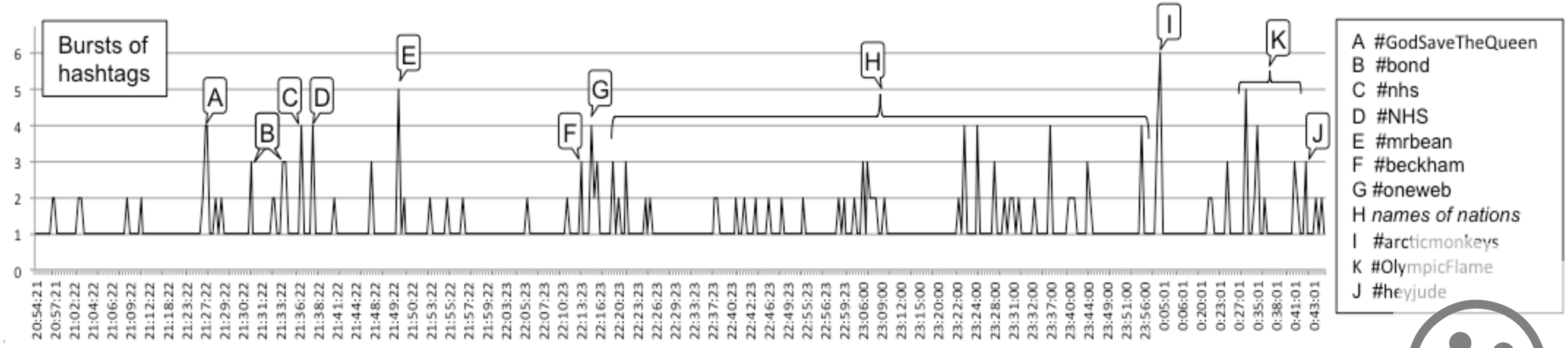






Findings

- Bursts of hashtags usage capture what attracts the attention of those watching the ceremony world wide



- Detailed Analysis

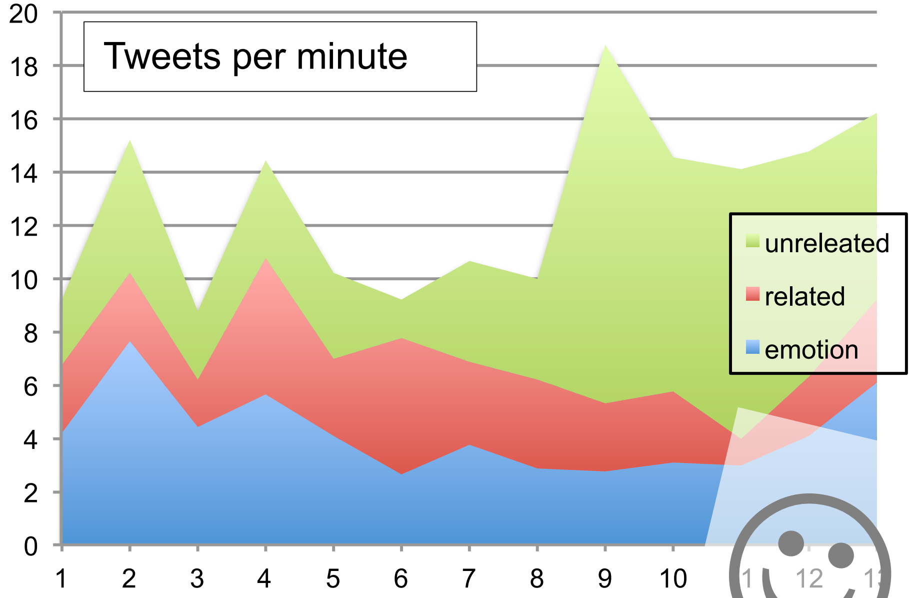
Moments of Ceremony	# of hashtags	Fraction
Total	189	100.00%
Hashtagged with an emotional state	34	17.99%
Correctly hashtagged	72	38.10%
- Right on time (1 min tolerance)	50	26.46%
- After the event (15 min tolerance)	13	6.88%
- Before the event (15 min tolerance)	9	4.76%

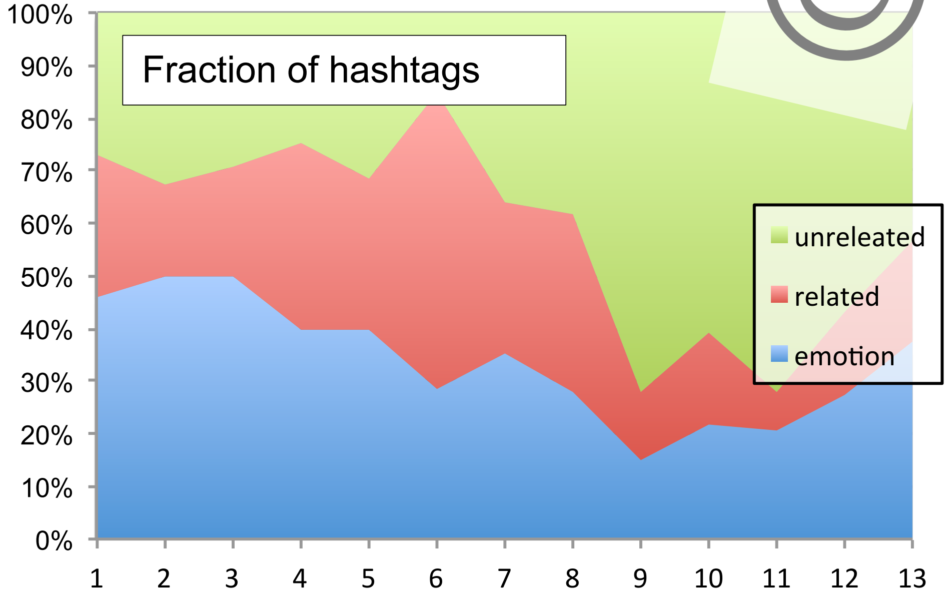
Test bed LOG'12 – study case 1

Findings

- Audience loosed attention while the ceremony was progressing
- Audience emotions where more evident in the first and the last part of the ceremony

1. Countdown	(21:00-21:04)
2. Green and Pleasant Land	(21:04–21:09)
3. Pandemonium	(21:09–21:25)
4. Happy and Glorious	(21:25–21:35)
5. Second to the right	(21:35–21:47)
6. Interlude	(21:47–21:52)
7. Frankie and June say…	(21:52–22:09)
8. Abide with Me	(22:09–22:20)
9. Welcome	(22:20–00:00)
10. Bike a.m.	(00:00–00:07)
11. Let the Games Begin	(00:07–00:24)
12. There Is a Light	(00:24–00:38)
13. And in the end	(00:38–00:46)


Tweets per minute


Fraction of hashtags

Test bed LOG'12
Case study #2 detecting events

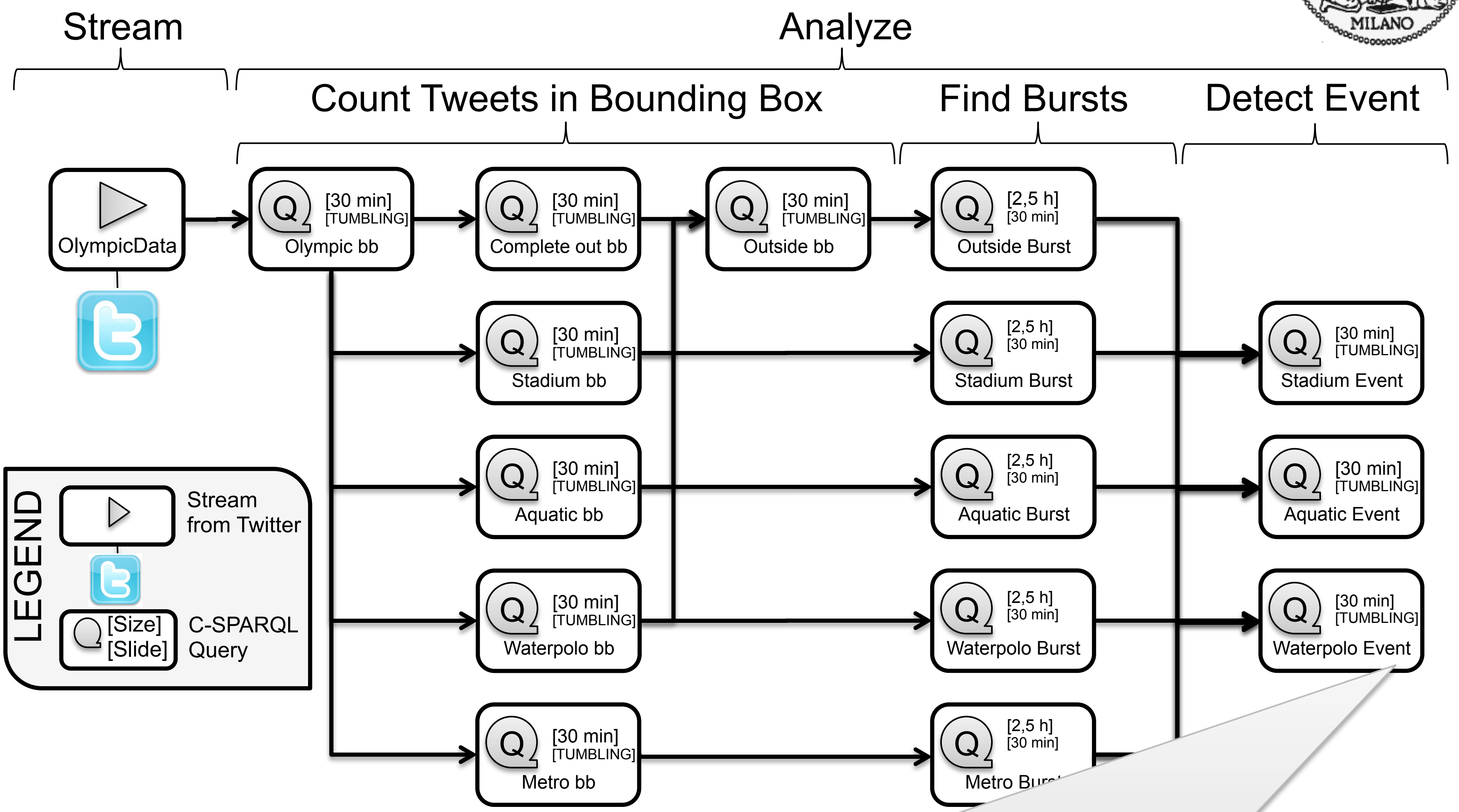
- The problem
  - detect the events given the position of a set of venues and socially listening their surroundings
- Input
  - three million tweets streamed by Twitter between July 25th and August 13th 2012
- Conditions

Type	Venue	Events	Capacity (seats)
Large	Olympic stadium	athletic games	80,000
Medium	Aquatic centre	swimming, diving and synchronized swimming	17,500
Small	Water polo arena	Water polo	5,000

- Ground truth
  - calendar of Olympic Games

Test bed LOG'12 – case study 2

SLD application

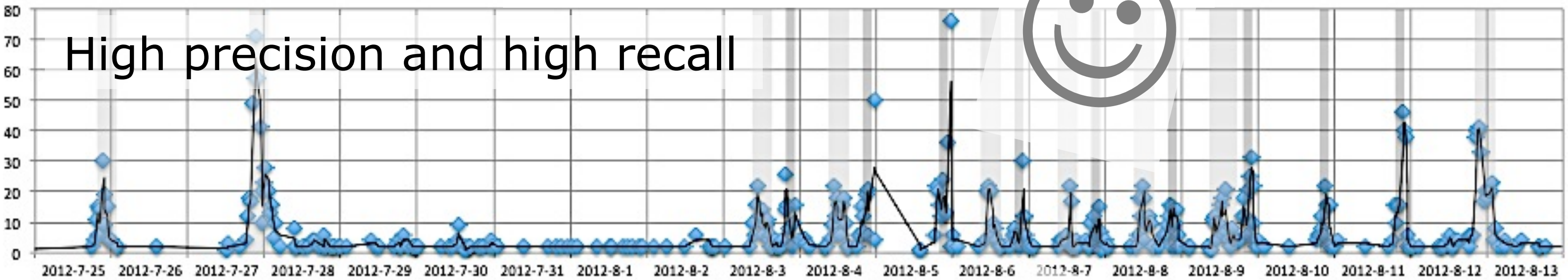

Stream
Analyze
Count Tweets in Bounding Box
Find Bursts
Detect Event
OlympicData
[30 min] [TUMBLING] Olympic bb
[30 min] [TUMBLING] Complete out bb
[30 min] [TUMBLING] Outside bb
[2,5 h] [30 min] Outside Burst
[30 min] [TUMBLING] Stadium bb
[2,5 h] [30 min] Stadium Burst
[30 min] [TUMBLING] Stadium Event
[30 min] [TUMBLING] Aquatic bb
[2,5 h] [30 min] Aquatic Burst
[30 min] [TUMBLING] Aquatic Event
[30 min] [TUMBLING] Waterpolo bb
[2,5 h] [30 min] Waterpolo Burst
[30 min] [TUMBLING] Waterpolo Event
[30 min] [TUMBLING] Metro bb
[2,5 h] [30 min] Metro Burst
LEGEND
Stream from Twitter
[Size] [Slide] C-SPARQL Query

(Metro Burst -> Outside Burst -> In ?venue Burst) within 30 min => event in ?venue

Test bed LOG'12 – case study 2

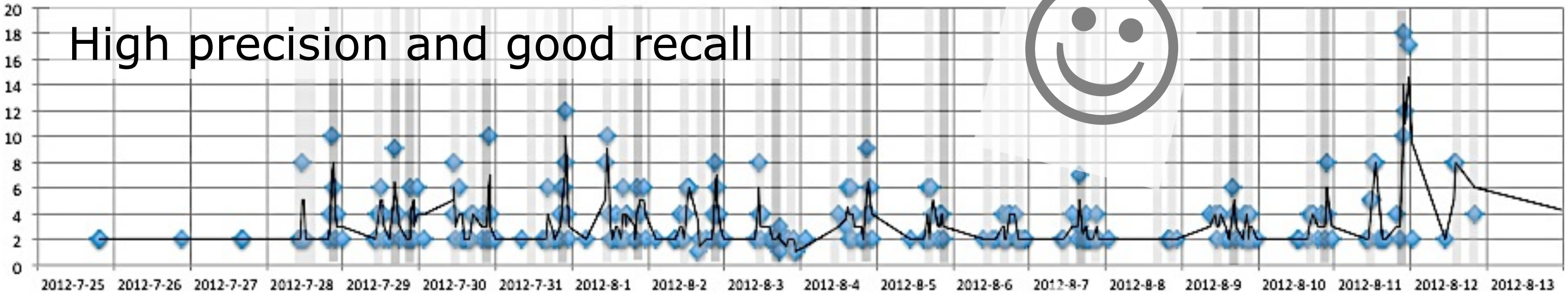
Evaluation

Stadium



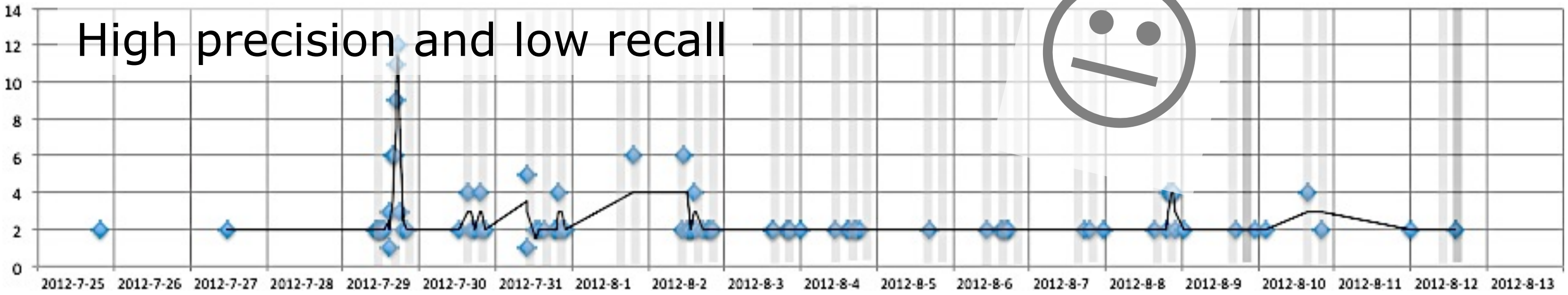
High precision and high recall

Water Centre



High precision and good recall

Water-polo Arena



High precision and low recall

Capacity (seats)

Case study #3 visualizing crowds movements

- The problem
  - Visualize the movement of the crowds
- Input
  - three million tweets streamed by Twitter between July 25th and August 13th 2012
- Conditions

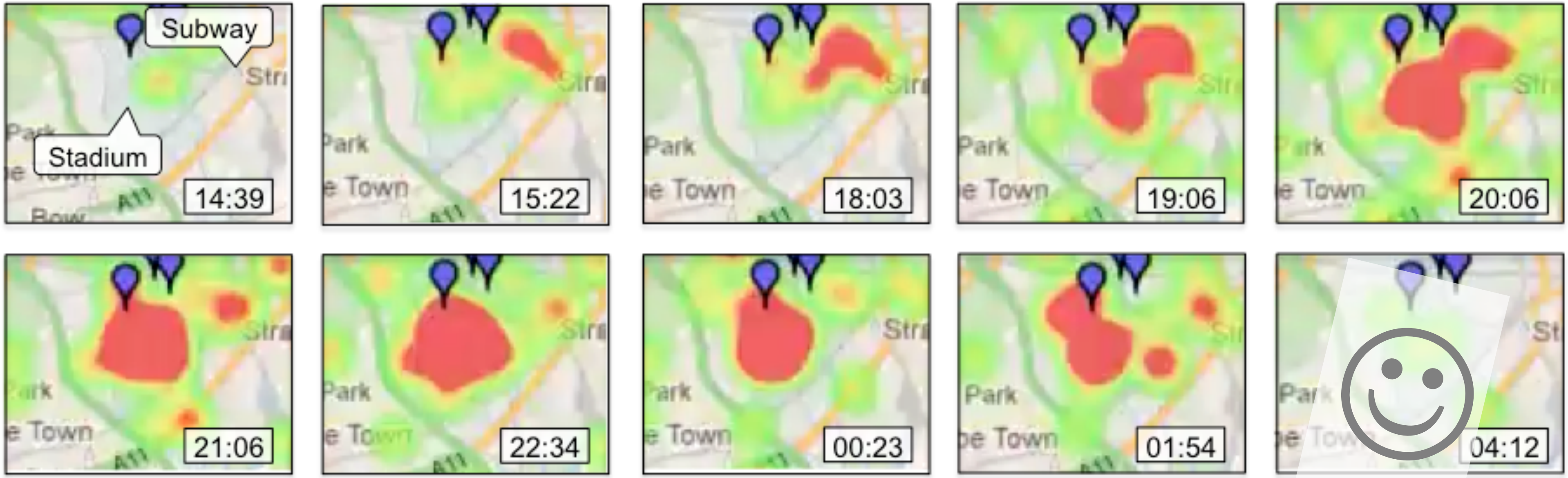
Type	Venue	Events	Capacity (seats)
Large	Olympic stadium	athletic games	80,000
Medium	Aquatic centre	swimming, diving and synchronized swimming	17,500
Small	Water polo arena	Water polo	5,000

- Ground truth
  - Expert judgement ;-)

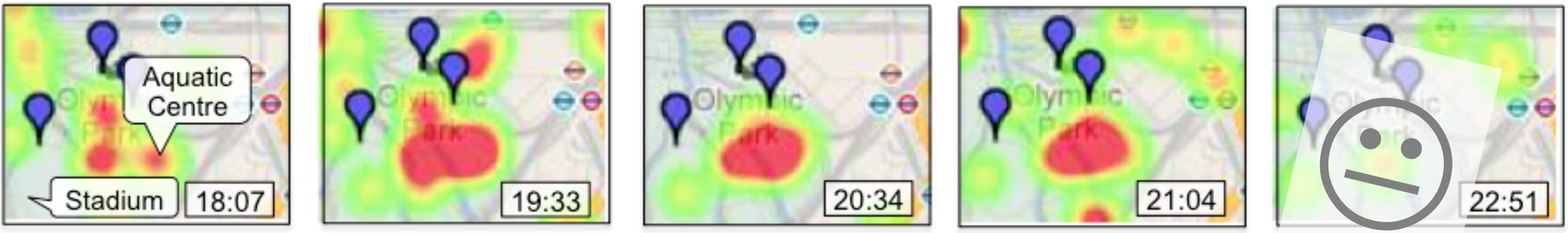
Test bed LOG'12 – case study 3

Evaluation

Opening Ceremony at the Olympic Stadium



A crowded event at the Aquatic Centre (July 31st, 2012)



The Milano Design Week 2013 (MDW'13)

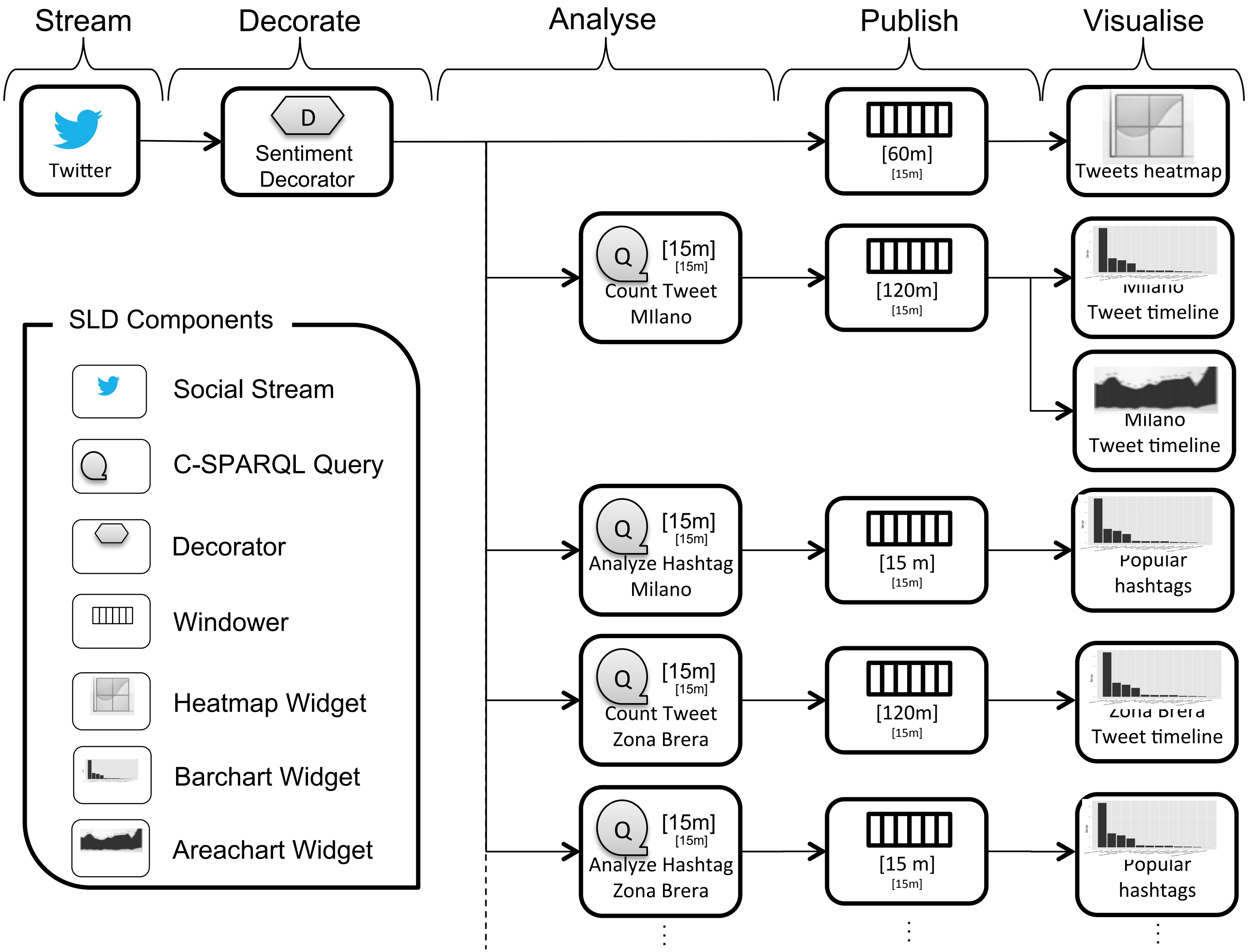
- Problem
  - Sponsor and organizer of a city scale event needs to quantify the return on investment
- Existing solutions
  - Spread people around the hundreds of event to asses the success of the various events is effective, but expensive
- Challenge
  - Obtaining comparable results by analysing public social streams

Test bed MDW'13 – case study 1

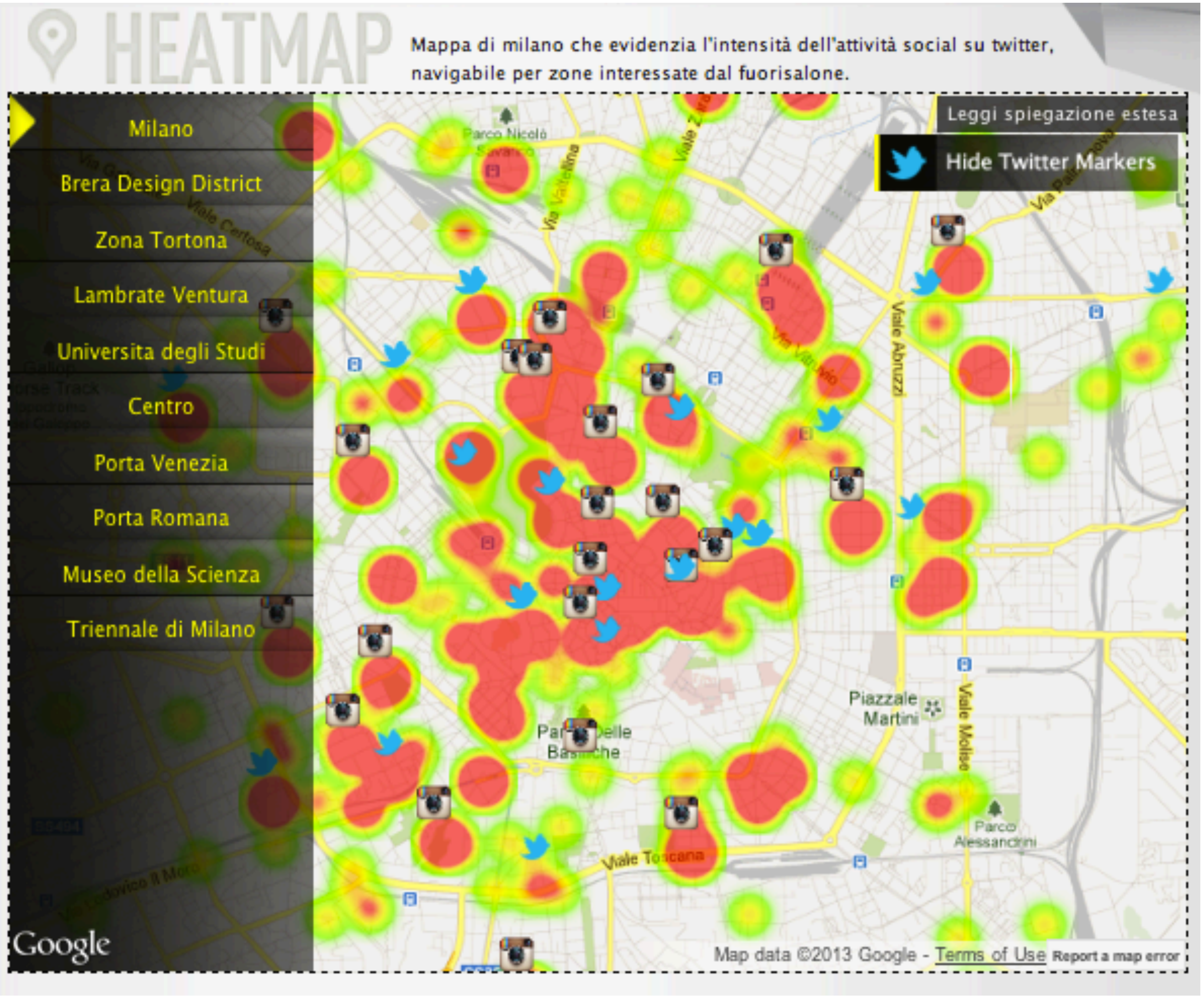
- The problem
  - Is MDW visible in the social streams posted by people in Milano area? If yes in real-time,
    1. What are the districts from which MDW visitors post the most?
    2. What are the most frequently used hashtags?
    3. How do people feel before, during and after the event they join?
  - Can these question answered at a cost a SME can afford?
- Input
  - 106,770 tweets streamed by Twitter between April 9th and April 14th 2013

Test bed MDW'13 – case study 1

SLD application



Test bed MDW'13 – case study 1
http://twindex.fuorisalone.it


HEATMAP

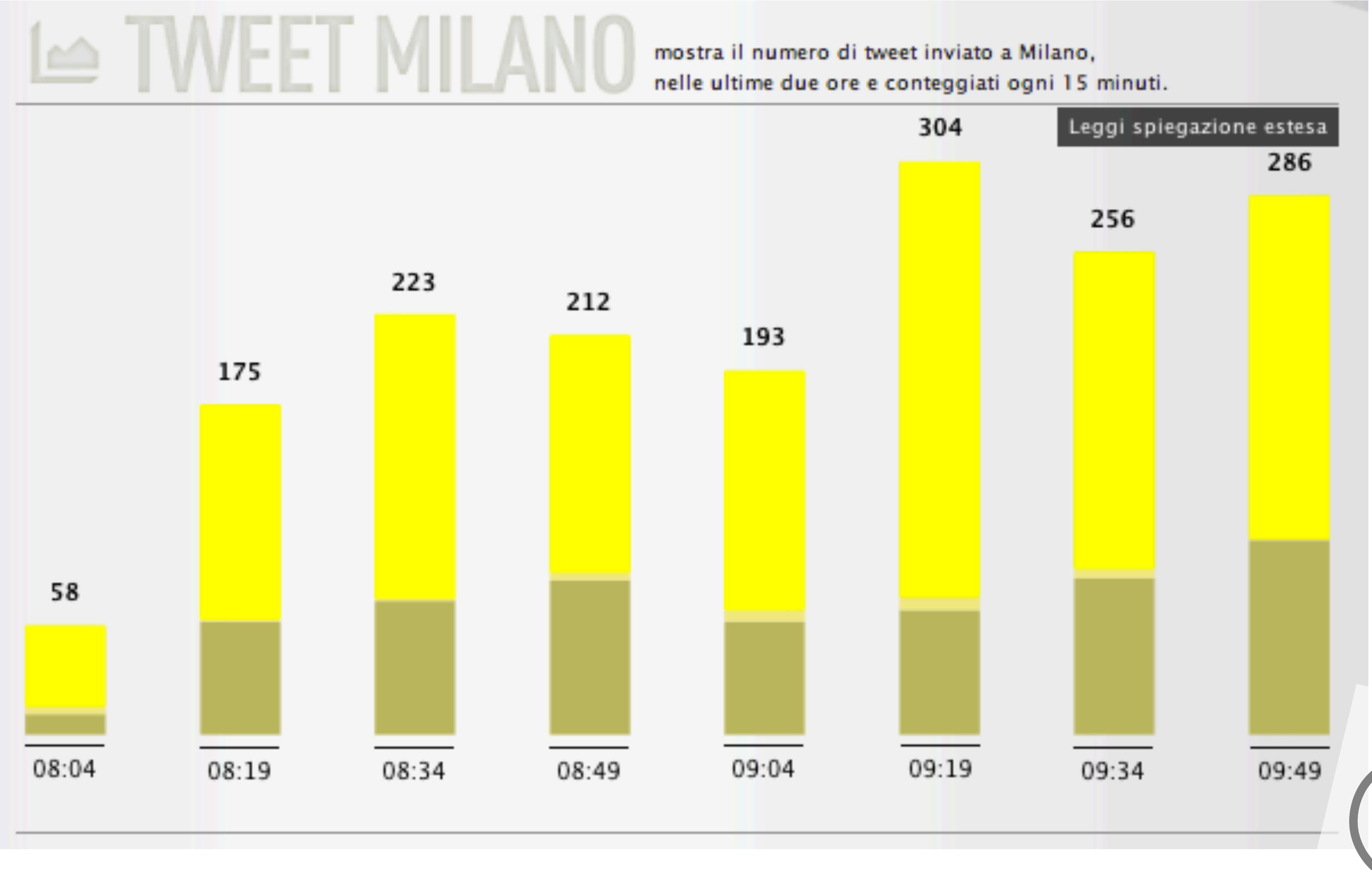
Mappa di milano che evidenzia l'intensità dell'attività social su twitter, navigabile per zone interessate dal fuorisalone.

April 9-14, 2013

- Distinct users
  - 12.031
- Invocation of Linked Data Publisher
  - 1,136,052
- Cost
  - 25 €/month
  - 2 cores, 2 GB

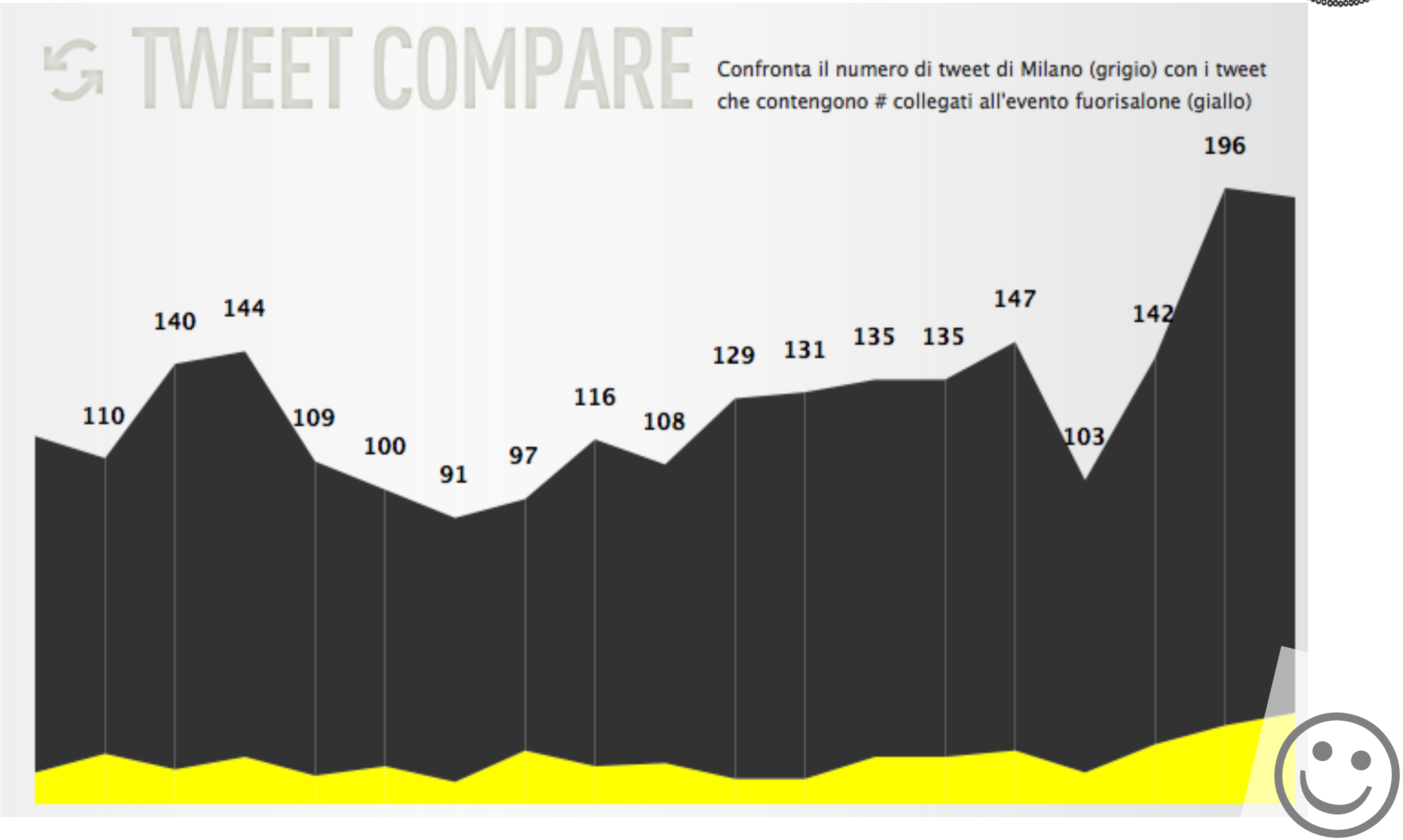
☺

Test bed MDW'13 – case study 1
http://twindex.fuorisalone.it



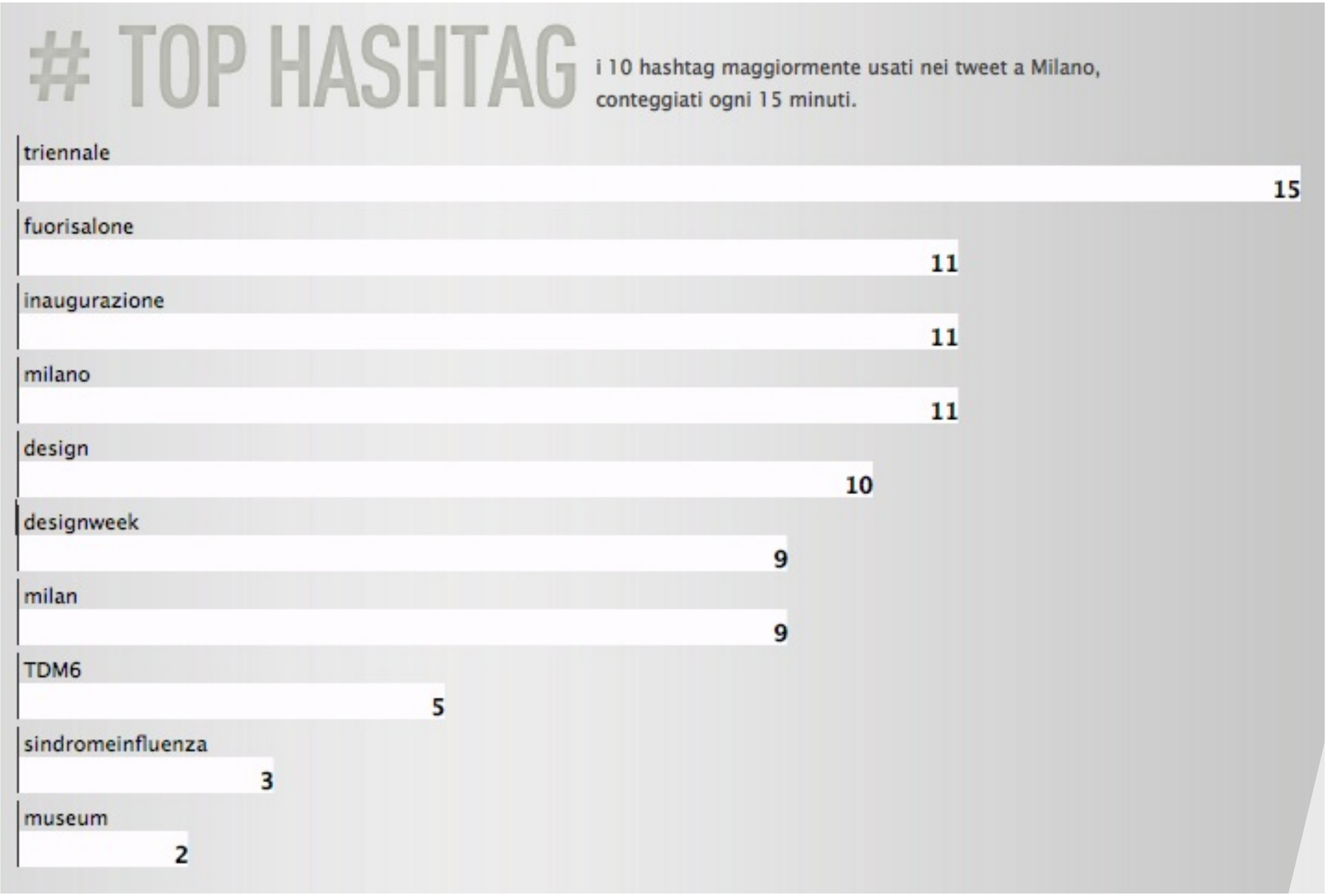
☺

Test bed MDW'13 – case study 1
http://twindex.fuorisalone.it


TWEET COMPARE

Confronta il numero di tweet di Milano (grigio) con i tweet che contengono # collegati all'evento fuorisalone (giallo)

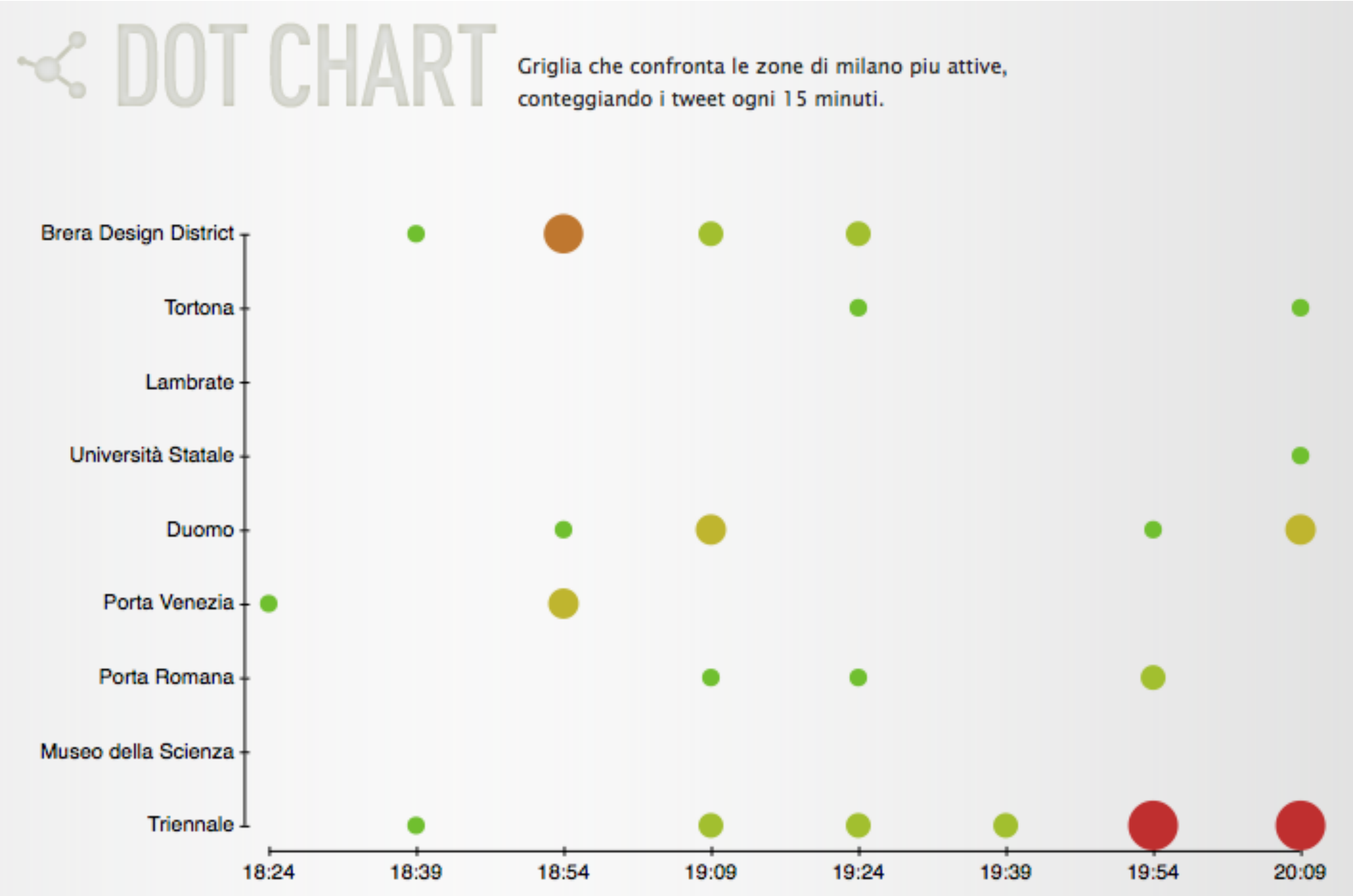
Test bed MDW'13 – case study 1
http://twindex.fuorisalone.it



TOP HASHTAG

i 10 hashtag maggiormente usati nei tweet a Milano, conteggiati ogni 15 minuti.

Test bed MDW'13 – case study 1
http://twindex.fuorisalone.it

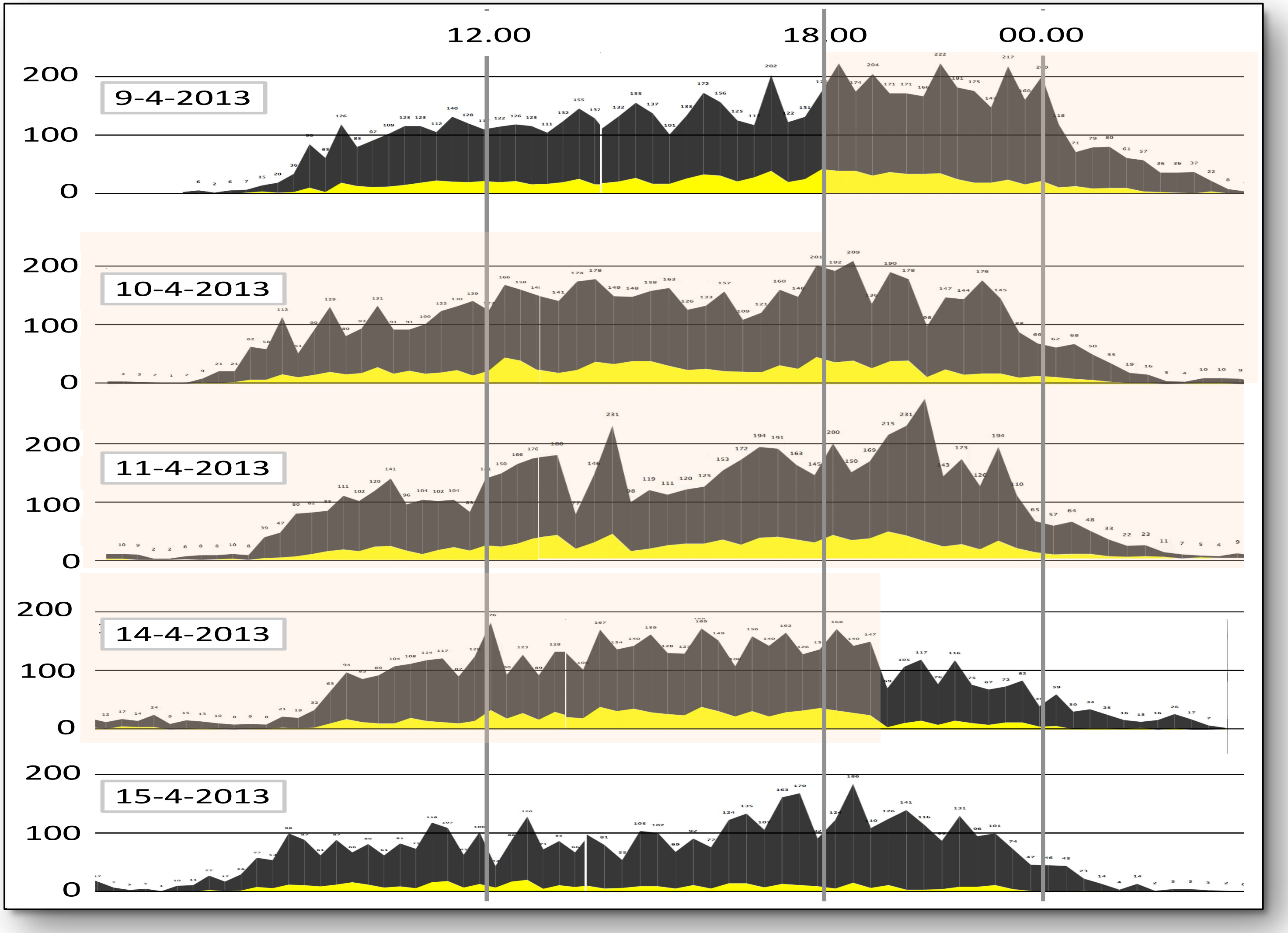


DOT CHART

Griglia che confronta le zone di milano piu attive, conteggiando i tweet ogni 15 minuti.

Test bed MDW'13 – case study 1

The results question 1



MDW 2013 is visible in the volume of micro-posts

Test bed MDW'13 – case study 1

The results question 1

April 9th, 2013 at 18.00	posts
fuorisalone	30
designweek	28
nabasalone	20
milano	9
design	6

April 11th, 2013 at 18.00	posts
milano	25
fuorisalone	22
design	10
designweek	6
32giornata	6

April 13th, 2013 at 18.00	posts
fuorisalone	28
designweek	21
nabasalone	17
milano	10
inter	8

April 15th, 2013 at 18.00	posts
inter	20
diretta	11
cagliarii	6
milan	4
seriea	3

MDW 2013 is visible in the top-5 hashtags used in the micro-posts

☺

Test bed MDW'13 – case study 1

The results question 2

Venue	posts
cesati antiques & works of art	16653
Porta nuova 46/b	13416
Circolo Filologico	9891
Adele Svettini Antichità	7366
ALTAI	5592
Bigli19	5175
Dudalina	4875
Galleria DadaEast	3550
borronichemicals	1078
Antonio Lupi Showroom Milano	995
Instituto Cervantes Milano	752
GALLERIA D'ARTE CONTEMPORANEA CINESE	560

The most attractive venues are found

☺

Test bed MDW'13 – case study 1

The results question 3

- The number of tweets posted by the same user where not enough to answer question 3 (How do people feel before, during and after the event they join?)

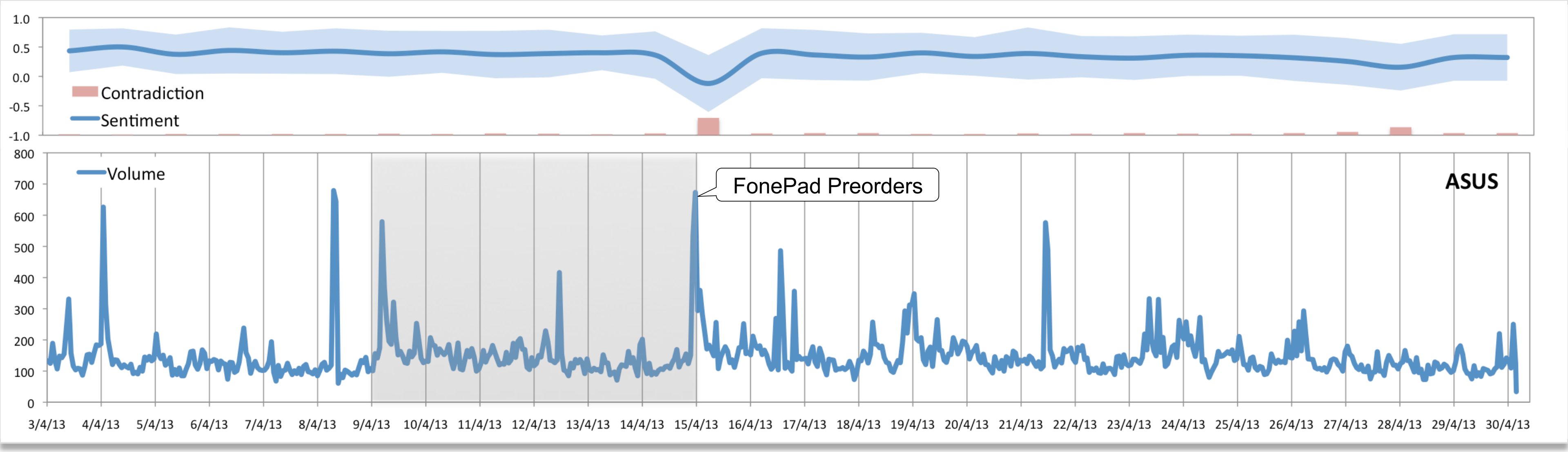
☹

Test bed MDW'13 – case study 2

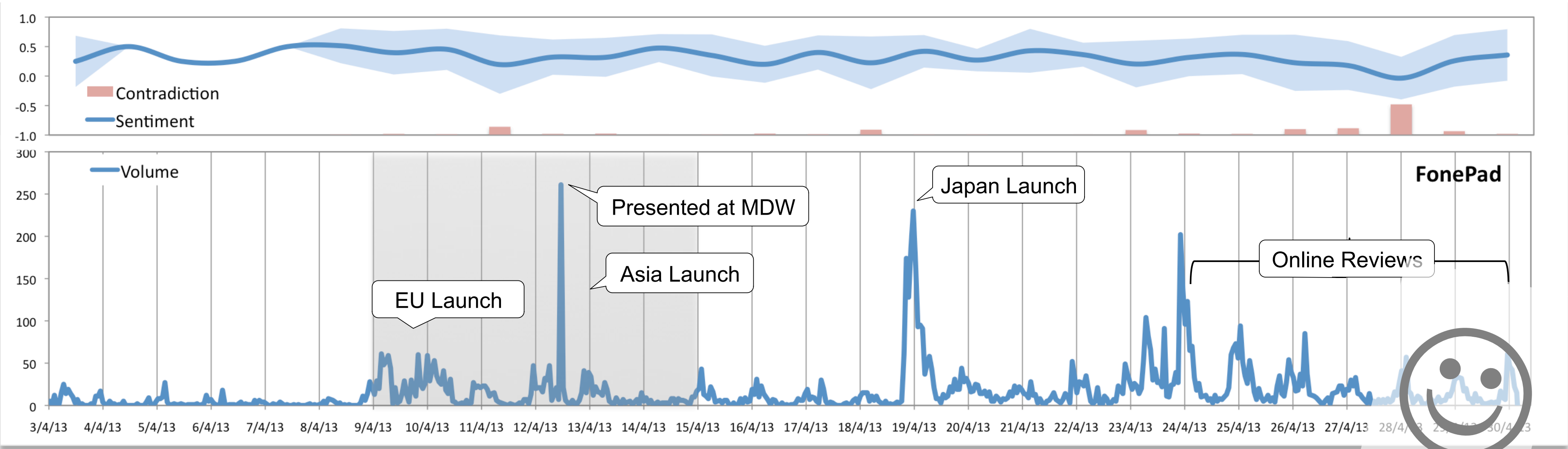
- The problem
  - Is the launch of ASUS products during MDW visible in the social streams posted by people around the world?
  - If yes, not necessarily in real-time,
    1. What are the products that attract more attention?
    2. What is the global sentiment before, during and after the launch?
- Input
  - 107,044,487 tweets registered with SLD between April 3rd and April 30th, 2013 asking Twitter to send to SLD tweets containing 300 words related to MDW, ASUS and its products
- Ground Truth
  - News, movies, and other media published on the Web in the same time period
  - Tool: google advance search

Test bed MDW'13 – case study 2

The results



FonePad Preorders



EU Launch

Presented at MDW

Asia Launch

Japan Launch

Online Reviews

Test bed MDW'13 – case study 2

The results

- As expected, the method did not handle sarcasm in a satisfactory manner
- some tweets about FonePad contained sarcastic sentences
  - e.g., *“wanna buy it so bad!”*
  - It was classified as negative
  - it was expressing positive sentiment.

☹

Wrapping up

- Listening to social streams and proper visual analytics methods can **unveil interesting city scale phenomena**
  - where people are gathering
  - what people are interested in
  - where people interested in a given topic are
  - if an event is running
  - what people feels about some topic
  - if the people are perceiving the message an event organizer want to transmit
  - how the feeling of the people evolves over time
  - …

Problematic aspects

- **lack of volume** in social streams prevents to perform meaningful analysis
- **Sarcasm and idioms** hinder the possibility to capture the opinion of people from highly volatile social streams
- **limited a priori knowledge** about the event hampers the ability to link social content to background data

Next steps

- Basic research
  - continuous top-k query answering
  - crowdsource data cleansing and linking
  - determining what is systematic is difficult
- Applied research
  - profile a crowd
  - find opinion makers
  - predict social trends

On-going collaborations

Who	Semantic techs	Streaming algorithms	Crowd-sourcing	Visual analytics
CEFRIEL	■		■	
Density Design Lab – PoliMi				■
DISI – University of Trento	■	■		
KDD Lab – ISTI, CNR, Pisa	■		■	
ML Group – SIEMENS	■			
Ontology Eng. Group – UPM	■	■		
Saltlux – Korea	■			■
SKIL Lab - Telecom Italia	■	■		
Studio Labo				■
Web IS - TU Delft	■		■	

City Data Fusion

http://citydatafusion.org

DeRiVE 2013 Workshop
21.10.2013, ISWC 2013, Sydney, Australia

Thank you! Any question?

Emanuele Della Valle
emanuele.dellavalle@polimi.it
http://emanueledellavalle.org