

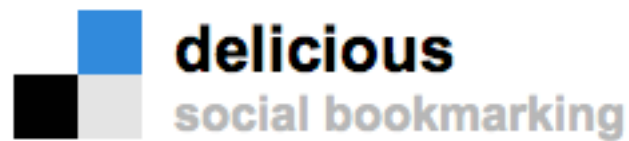
Extracting Time and Location Concepts Related to Tags

INSEMTIVE2008

October 26, 2008

Yukino Baba, Fuyuki Ishikawa
and Shinichi Honiden
University of Tokyo
National Institute of Informatics

Folksonomy



[The Future of Web 3.0 According to Yahoo! - ReadWriteWeb](#) SAVE

164

First saved by: stevechou

web3.0 < yahoo < web < trends < social

flickr LOVES YOU TM

You Tube

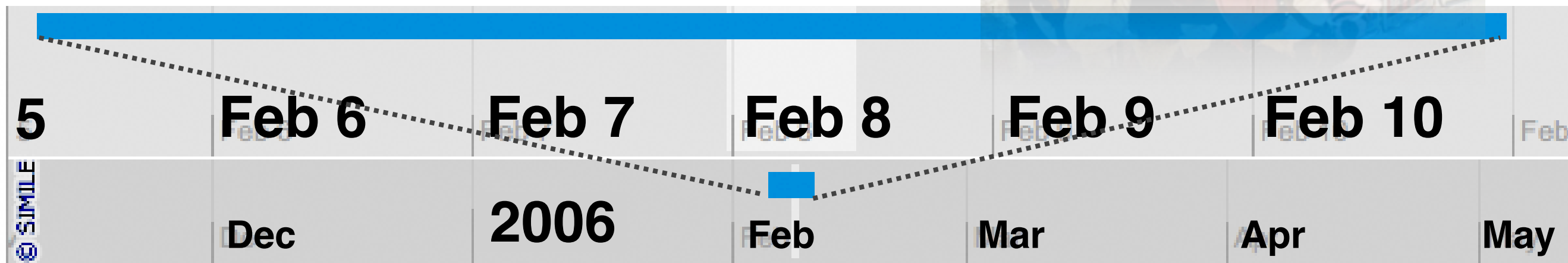
Broadcast Yourself™

- Folksonomy is widely used in web services
- There are a large amount of **[tag, content]** data generated **by many users**
- Our goal is **to extract the concept of a tag by focusing on the features of the content**
- In this work, we target the **temporal and spatial concepts related to tags**

Example of outputs (temporal concept)

Tag: sapporosnowfestival

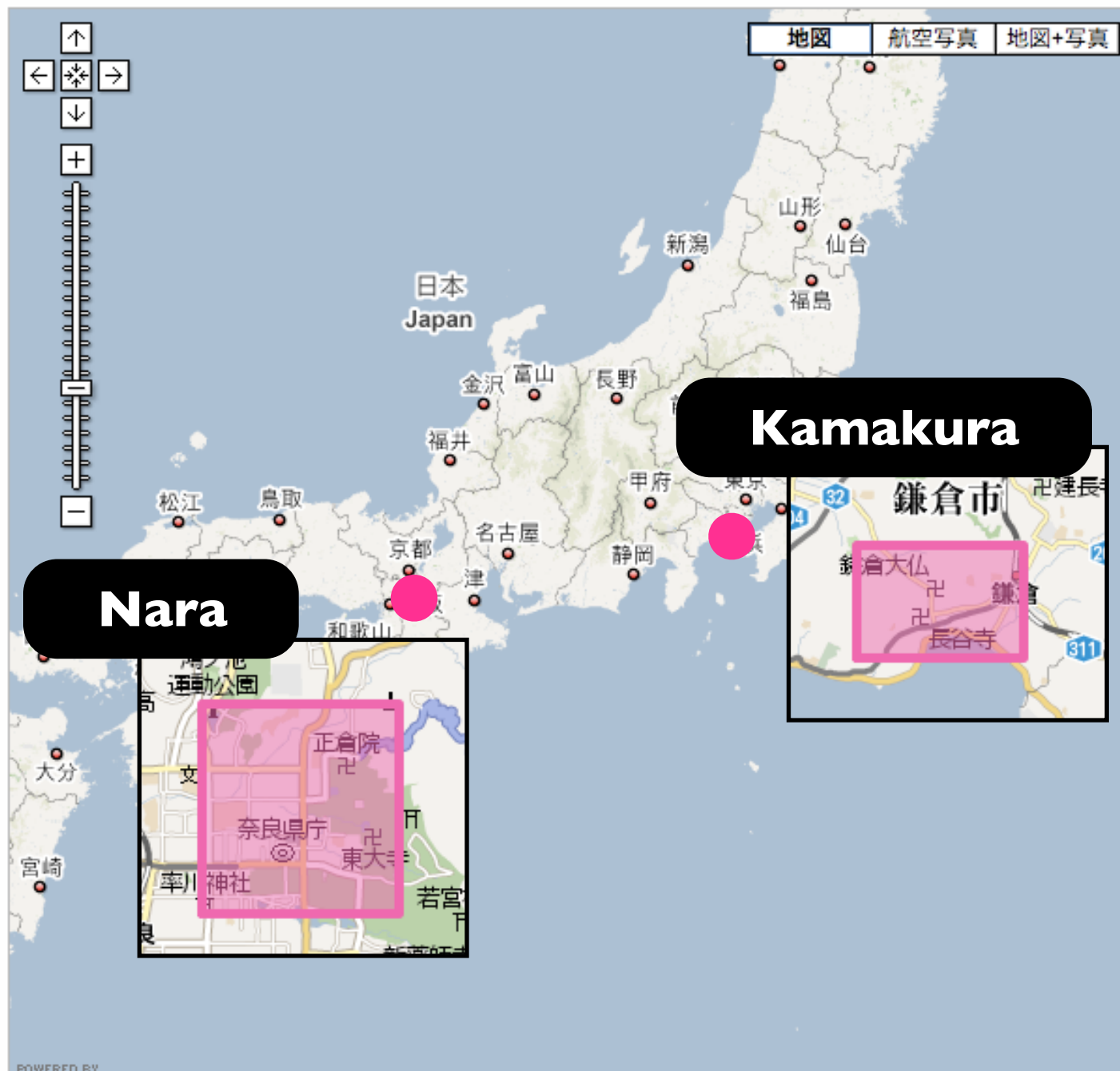
Time usage distribution: 524288



Example of outputs (spatial concept)

Tag: **daibutsu**

Location usage distribution: 16384



The Concept of a tag(I)

A tag is a label attached to a concept

Content



Concept



Sagrada Familia

Barcelona

Gaudi

Tags

The Concept of a tag(2)

Recalled concept from the same content
varies with the individual

Content

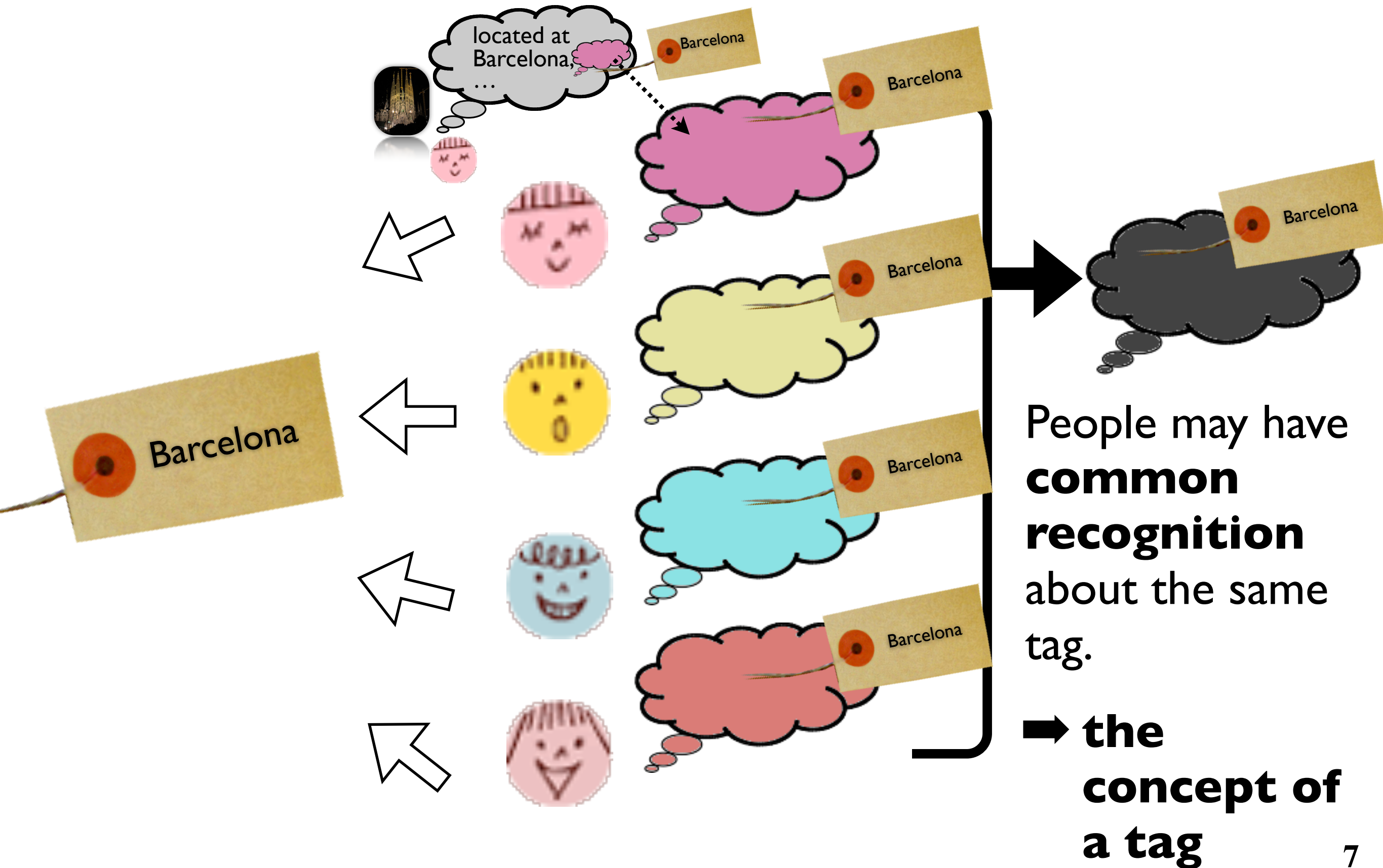


Concept

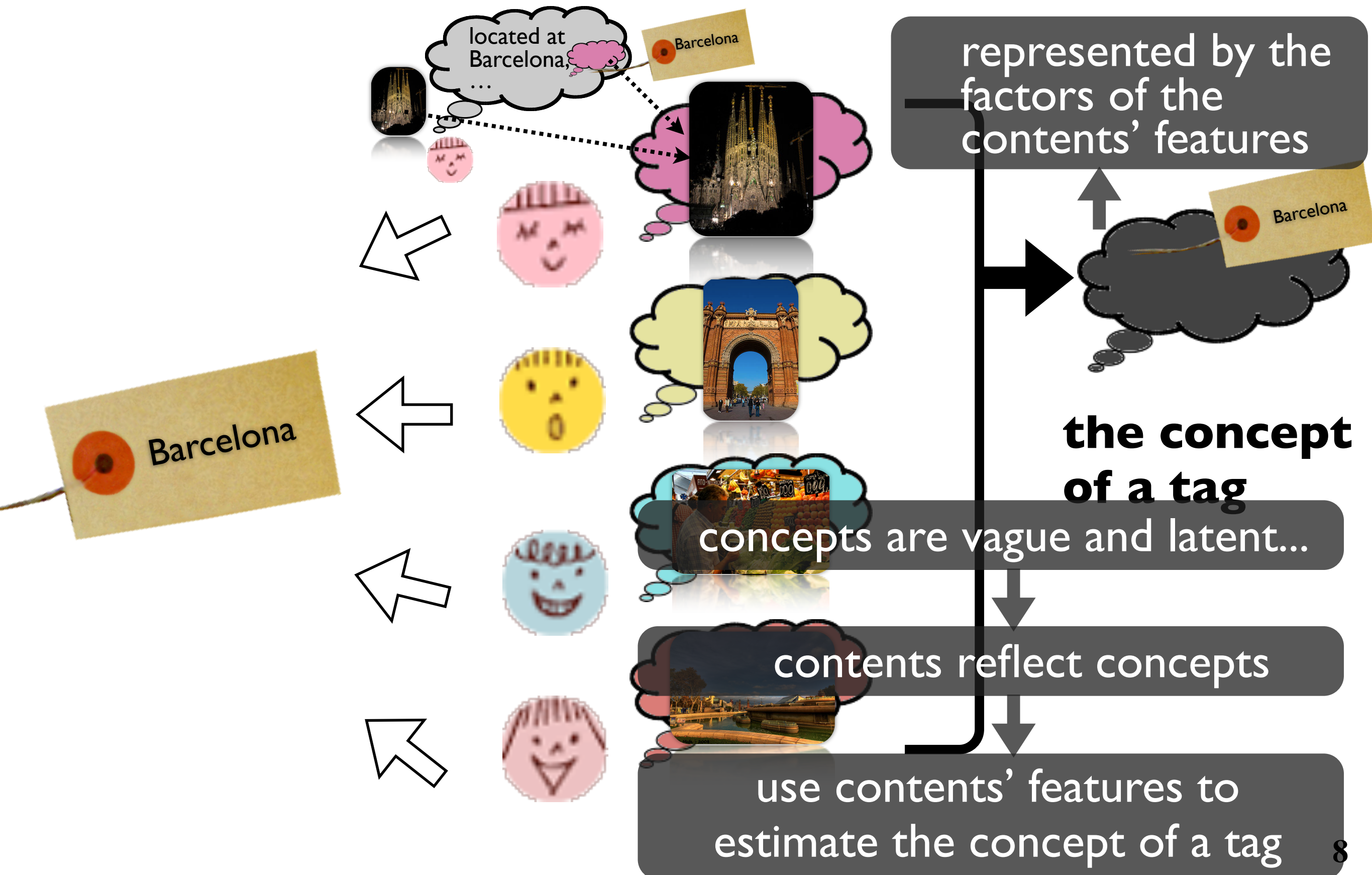


Tags

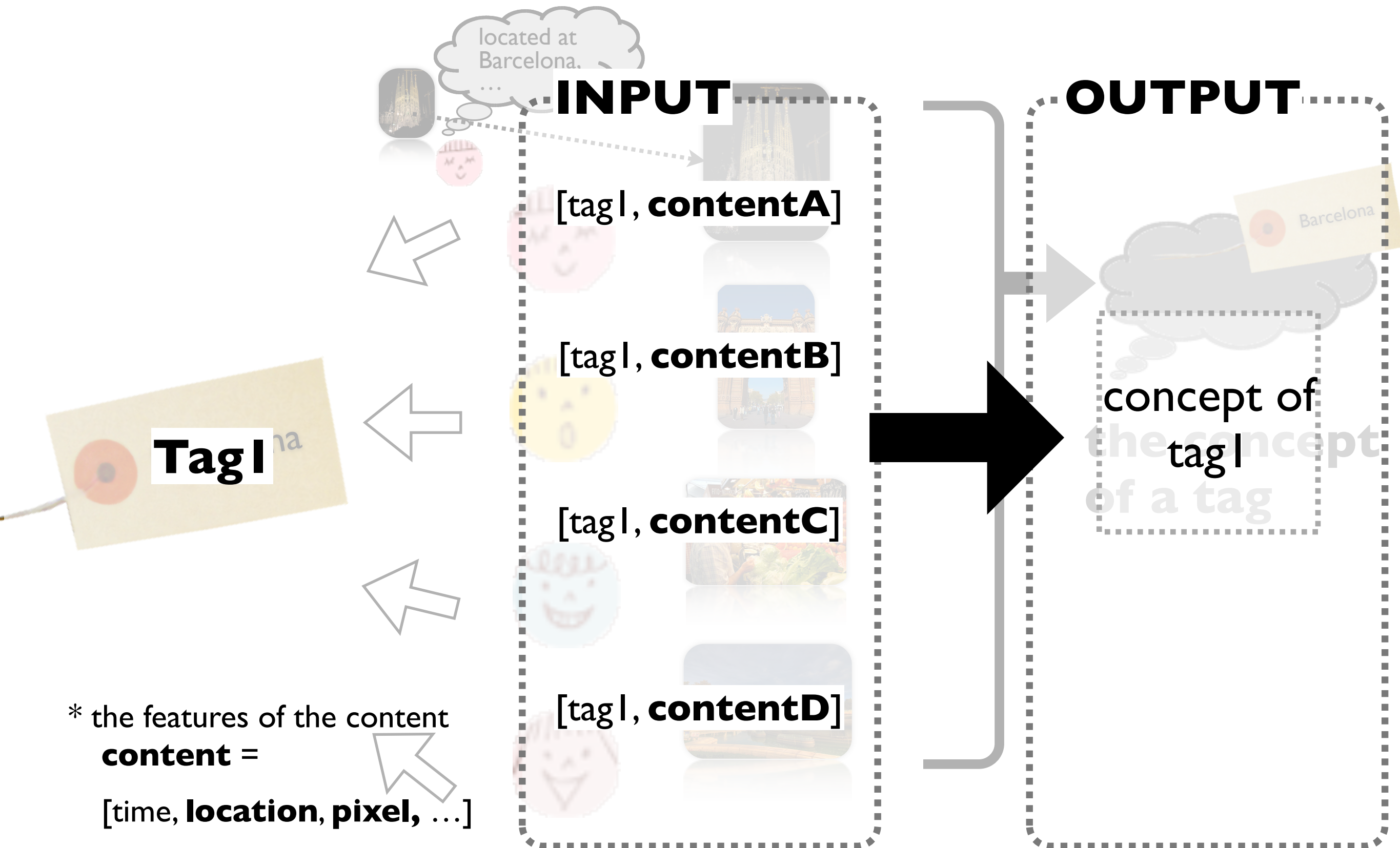
The Concept of a tag(3)



The Concept of a tag(4)



INPUT/OUTPUT



Time and Location

- Use the **time** and **location** data as features of the content
 - Both data are given in some services
 - e.g., photo sharing services give the time and location data as at which the photo was taken
- **Time tag**: A tag related to time
 - e.g., August, cherry blossoms, Sapporo Snow Festival
- **Location tag**: A tag related to location
 - e.g., Tokyo, Daibutsu, Sapporo Snow Festival

INPUT/OUTPUT

(Time&Location)

INPUT

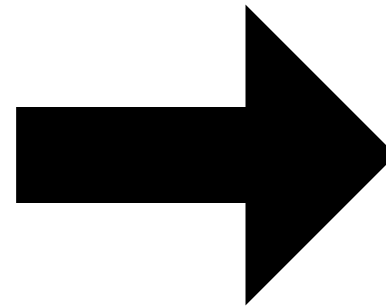
[Time tag1, **TimeA**]

[Time tag1, **TimeB**]

[Time tag1, **TimeC**]

...

Time = YYYY-MM-DD hh:mm:ss



OUTPUT

concept of
Time tag1

INPUT

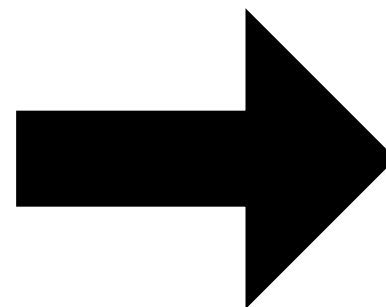
[Location tag2, **LocationD**]

[Location tag2, **LocationE**]

[Location tag2, **LocationF**]

...

Location = [latitude, longitude]



OUTPUT

concept of
Location
tag2

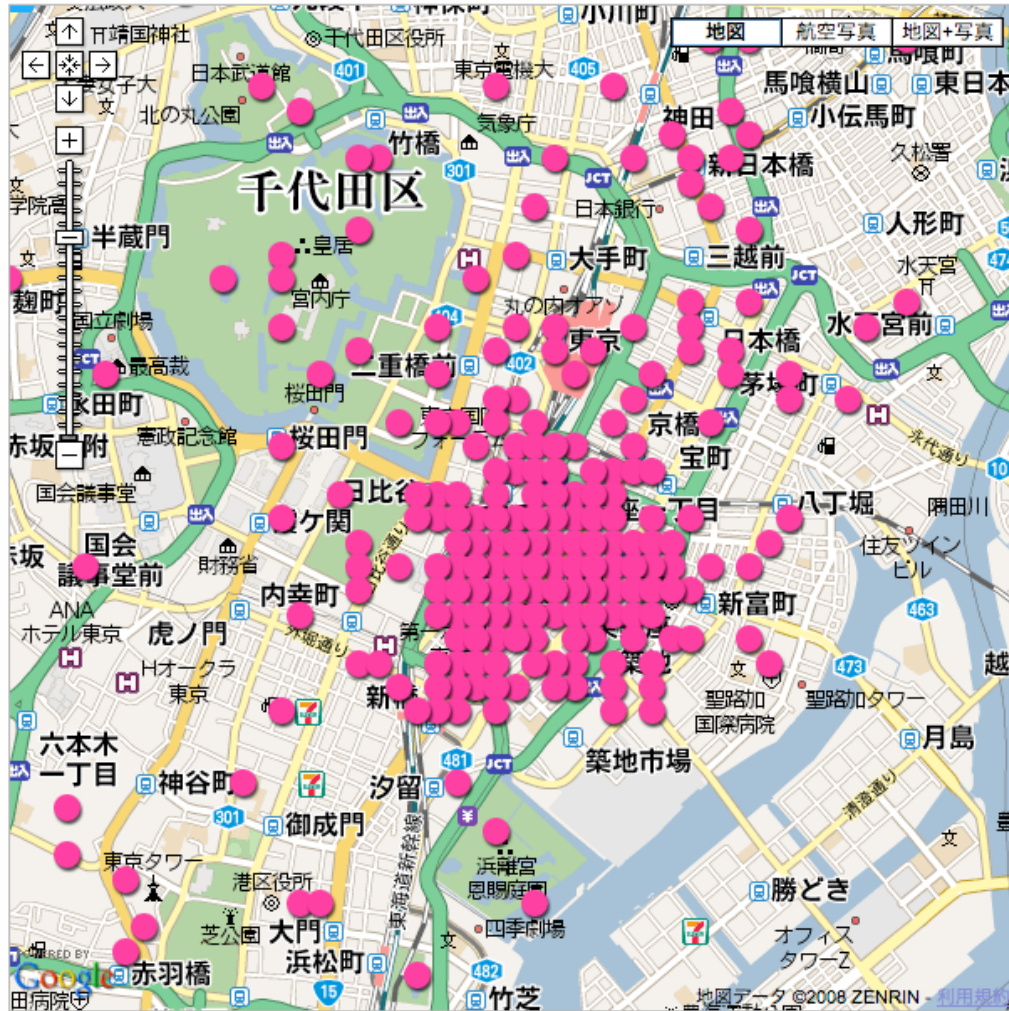
The Concept of the Time/ Location Tag

- The ranges related to the tags
 - e.g., 2007-02-05 12:21:58 to 2007-02-16 23:48:24
 - e.g., (lat: 35.616279, lng: 139.650307) to (lat: 35.772702, lng: 139.859047))
- ➡ The dense region in the tag's temporal/spatial distribution

Approach

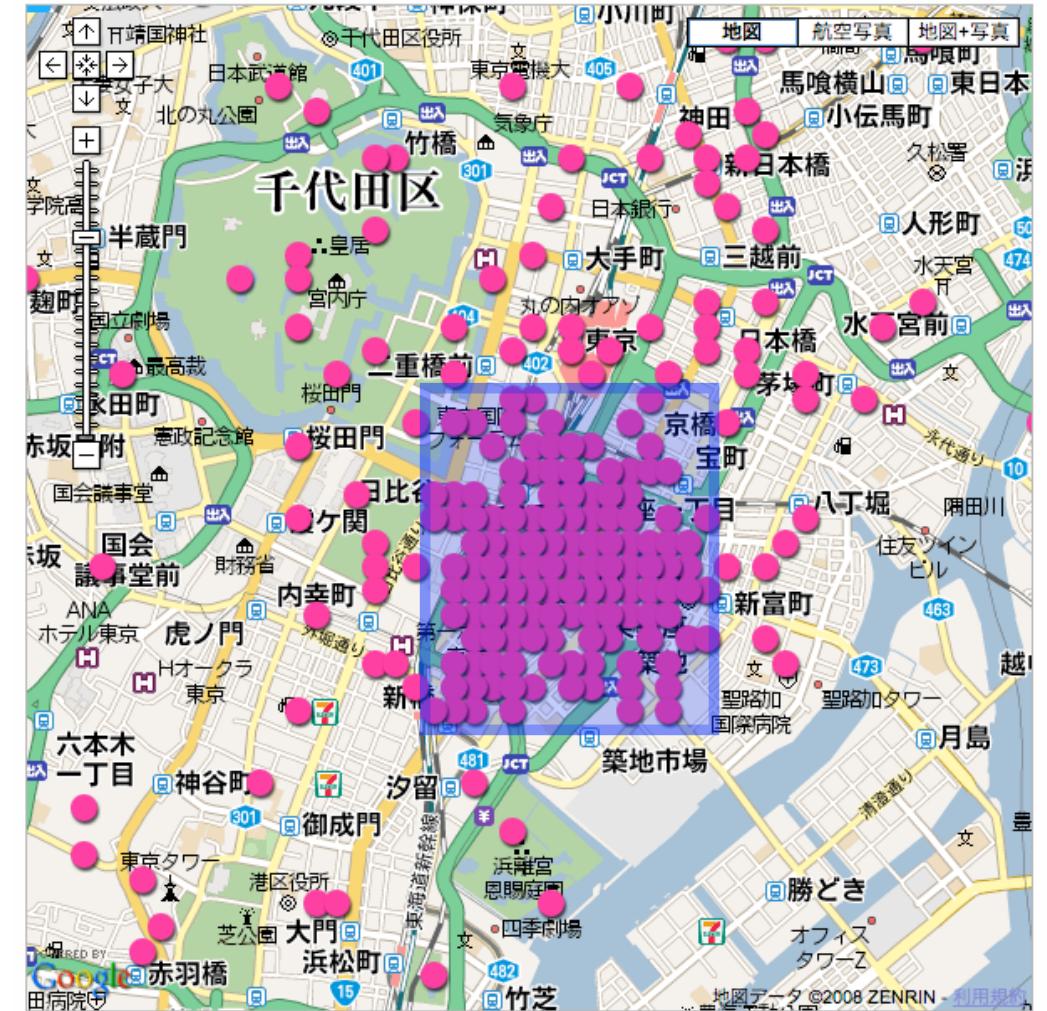
Tag: **ginza**

Location usage distribution



Tag: **ginza**

Location usage distribution



- Extract the Concept of the Time/Location Tag from the tag's usage distribution
- ➡ Determine the dense region in the tag's temporal/spatial usage distribution

Benefits

- Our method might be able to extract the concept that people often recognize
- ➡ Estimate concepts which are difficult to determine using the top-down definition approach

- Place name: recognized different as from the definition on the map

- ▶ e.g., “tokyo”
the place really called as “Tokyo”
may be only a part of the Tokyo Metropolitan Area



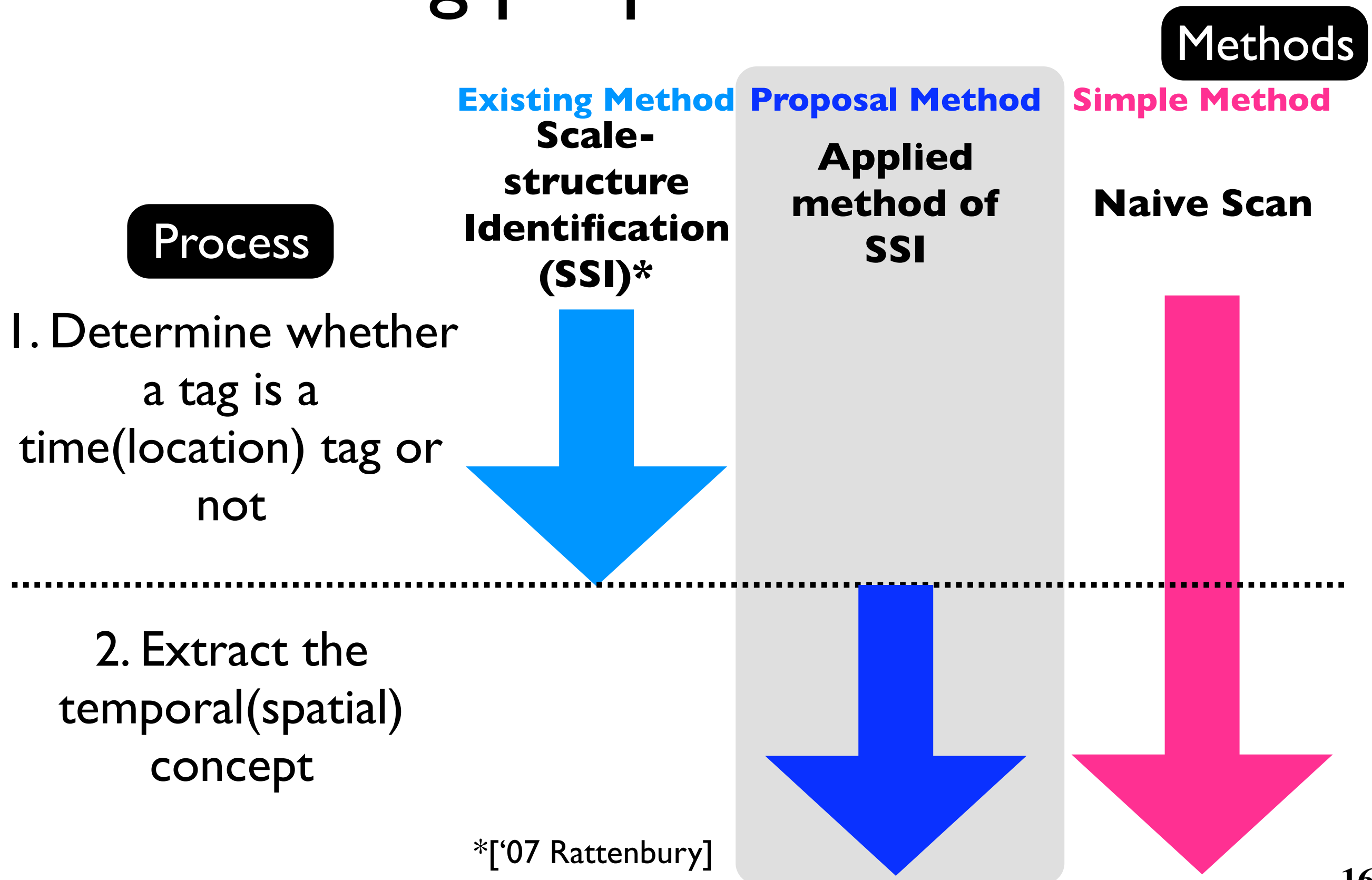
- Tags that are implicitly related to time or location

- ▶ e.g., “cherry blossom”

Applications

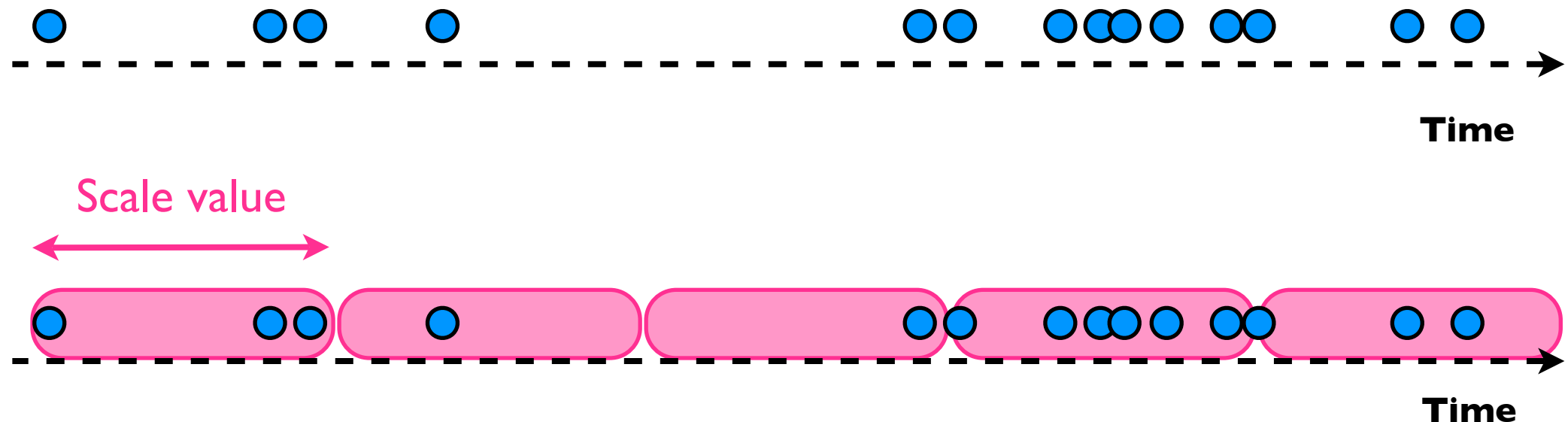
- Metadata estimation
 - Estimate the contents' time or location metadata from the tags
- Solution for Synonyms
 - Tags are synonyms if they have similar concepts
- Tag Suggestion
 - Suggest a tag related to the content

Existing/proposal methods



Naive Scan

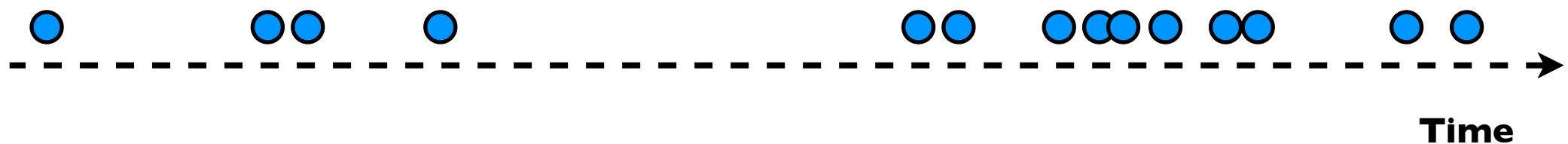
- Standard algorithm used in signal processing to detect a burst



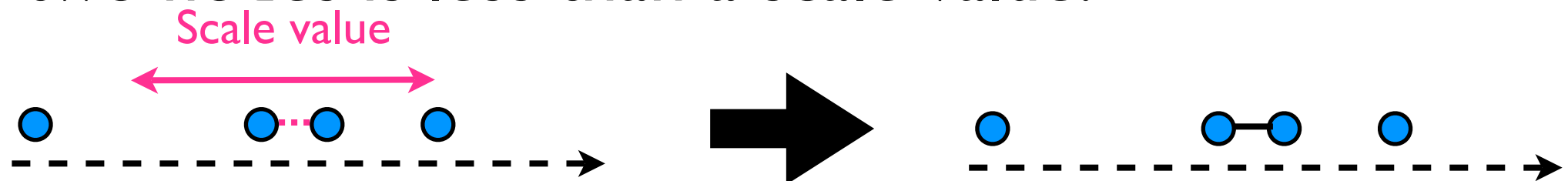
The result of Naive Scan method depends on the scale value and segment separation

Scale-structure Identification(SSI)

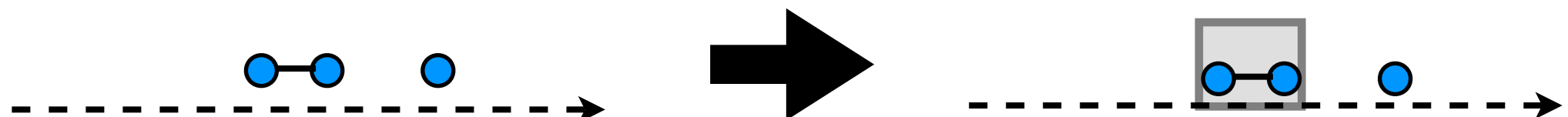
- Analyze the results on multiple scales and determine whether a tag is time (location) tag or not



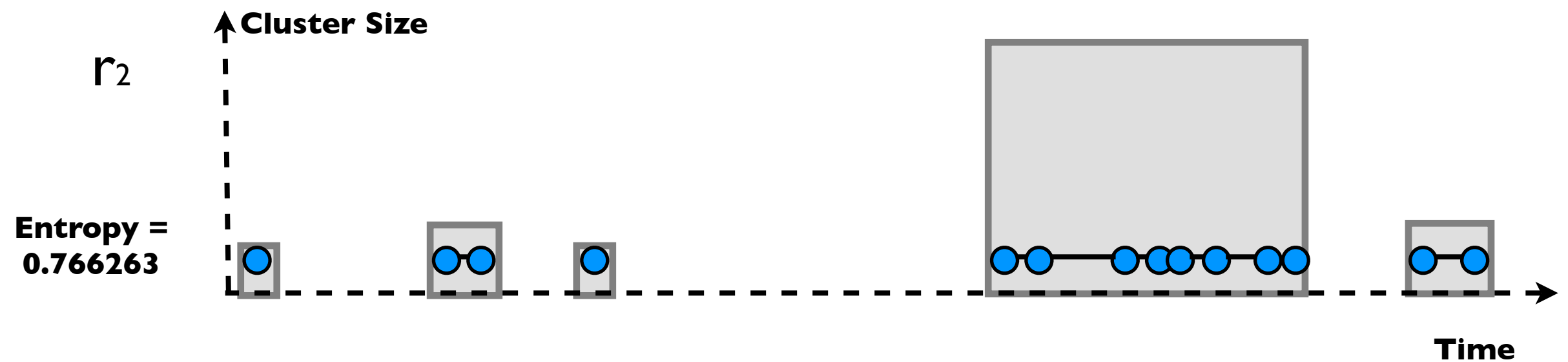
- The edges of a node pair exist if the distance between the two nodes is less than a scale value.



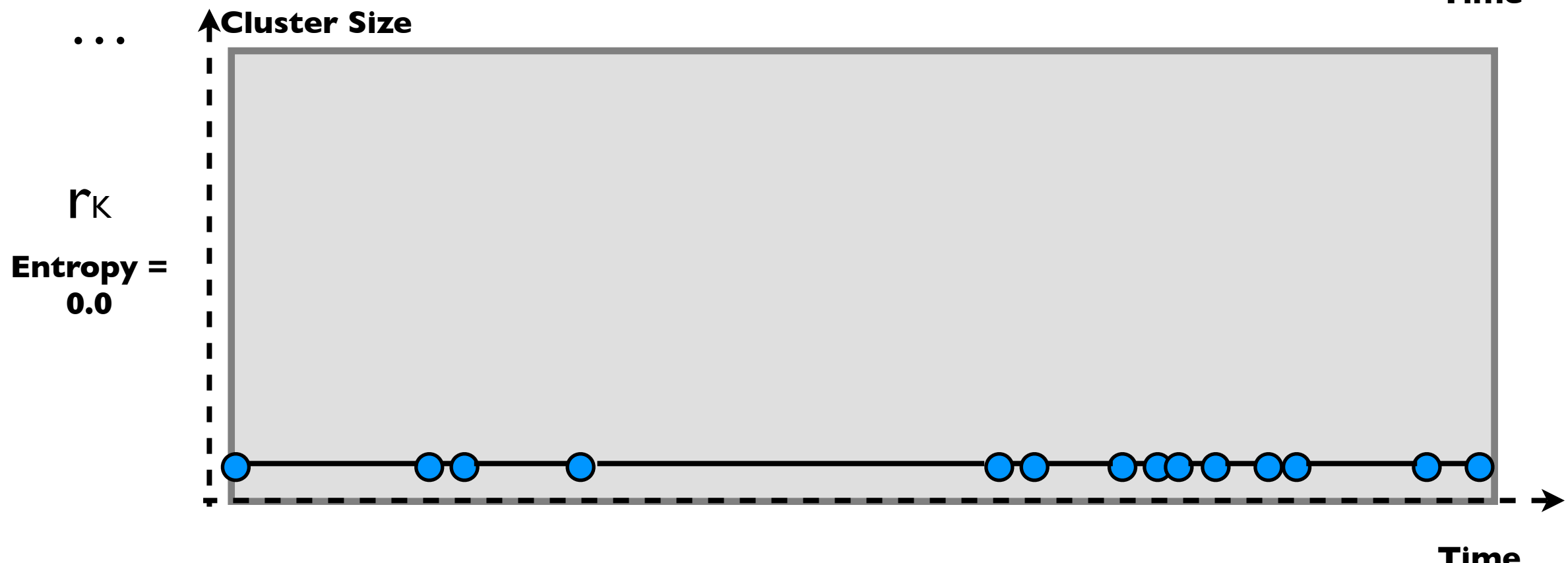
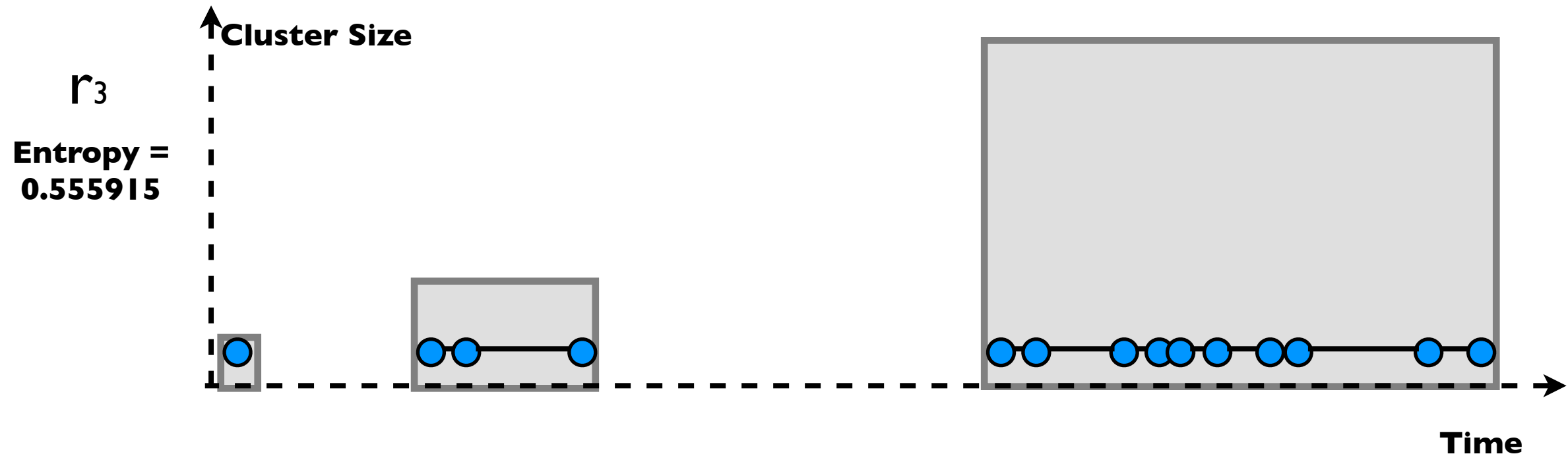
- Consider the set of the connected graphs as a cluster.



Scale-structure Identification(SSI)



Scale-structure Identification(SSI)



Scale-structure Identification(SSI)

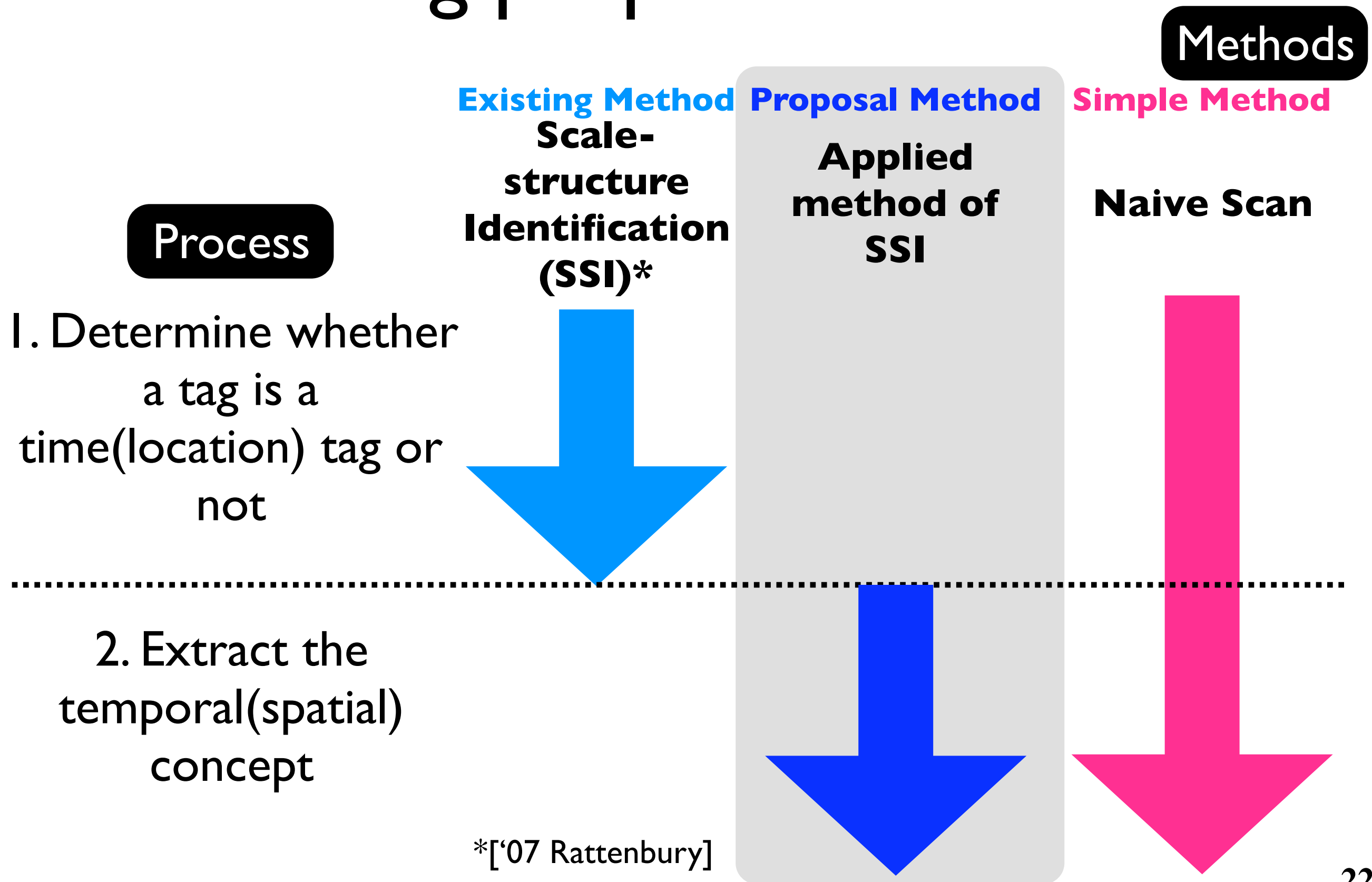
- A tag is a Time tag if the number of clusters is low for all scales.
 - ─ Because the distribution may have a dense region

➡ A tag is a Time tag when the sum of entropies E_{r_k} for all scales is less than a certain threshold.

$$E_{r_k} = \sum_{Y \in Y_{r_k}} \frac{|Y|}{|T_x|} \log_2 \frac{|T_x|}{|Y|} \quad \sum_{k=1}^K E_{r_k}$$

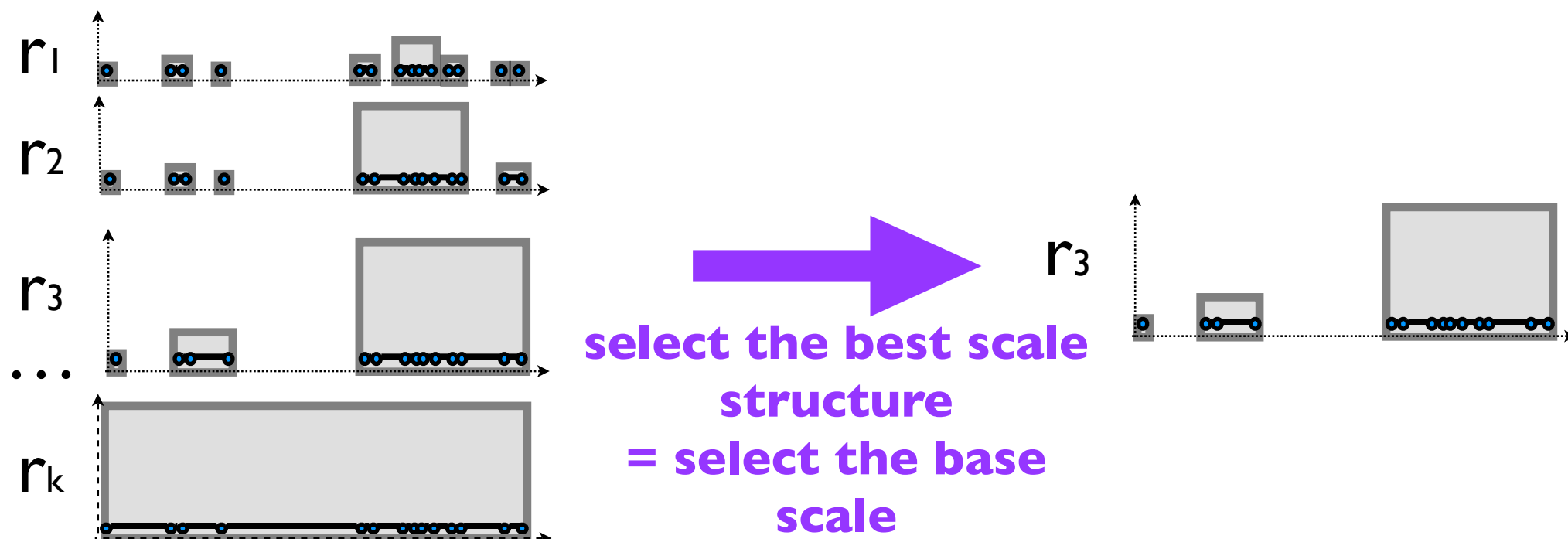
- ─ E_{r_k} indicates the degree to which the whole distribution is similar to one cluster (When $|Y_{r_k}| = 1$, $E_{r_k} = 0$)

Existing/proposal methods



Applied method of Scale-structure Identification

- Which clustering structure of scale values best characterize the time distribution?
- Requirements for better clustering structure:
 - There is more than one cluster having a lot of nodes **Compactness**
 - Clusters are distant from one another **Isolation**



Evaluation formula['00 Leung]

Evaluation for
each cluster

$$compactness(Y_i) = \frac{\sum_{t \in Y_i} \exp^{-\|t - t_i\|^2 / 2r^2}}{\sum_{t \in Y_i} \sum_{Y_j \in Y} \exp^{-\|t - t_j\|^2 / 2r^2}}$$

$$isolation(Y_i) = \frac{\sum_{t \in Y_i} \exp^{-\|t - t_i\|^2 / 2r^2}}{\sum_t \exp^{-\|t - t_i\|^2 / 2r^2}}$$

Evaluation for
clustering
structure

$$F_c(r) = \sum_i^m compactness(Y_i) - m$$

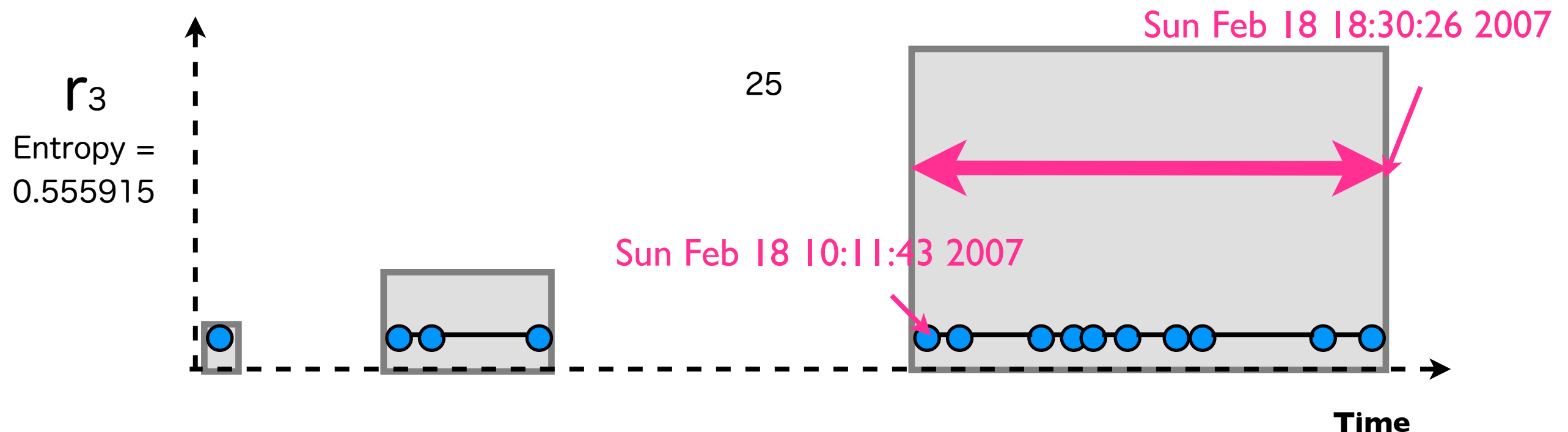
$$F_i(r) = \sum_i^m isolation(Y_i) - m$$

24

The clustering structure is better if both $F_c(r)$
and $F_i(r)$ are larger

Determine the region

- After selecting the best clustering structure, select the best cluster(s)
- If the number of nodes in a cluster is larger than a certain percentage of all nodes, the cluster is characteristic



Experiment - Dataset

- Used photo data on Flickr
 - Photos taken between 2004/1/1 and 2007/12/31 and annotated with location data in Japan
 - Tags annotating more than 100 photos and that were used by more than three users
 - No. of Photos: 3,826,253, No. of Tags: 2,453
- Chose 100 Time tags and 100 Location tags

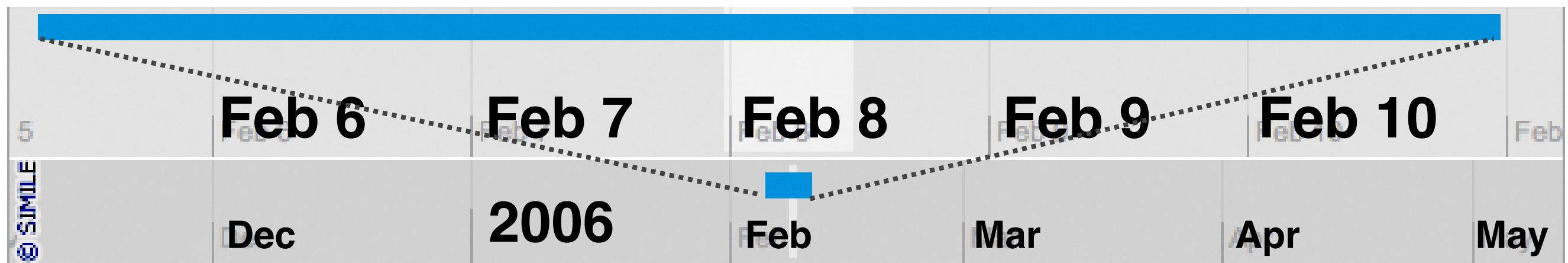
Example of Time and Location Tags

- Time Tags
august2007, carnival, comiket, cosmos, firework, fujirock, gameshow, ginkgo, gionmatsuri, jidai, june, newyear, obon, rama, september, snowboarding, tgs2006, ume
- Location Tags
beppu, chatan, daibutsu, enoshima, fl, hakodate, himeji, kanazawa, matsumoto, nara, otaru, roppongi, shinagawa, takayama, tokyodisneysea

Example of outputs(Time Tag)

Tag: sapporosnowfestival

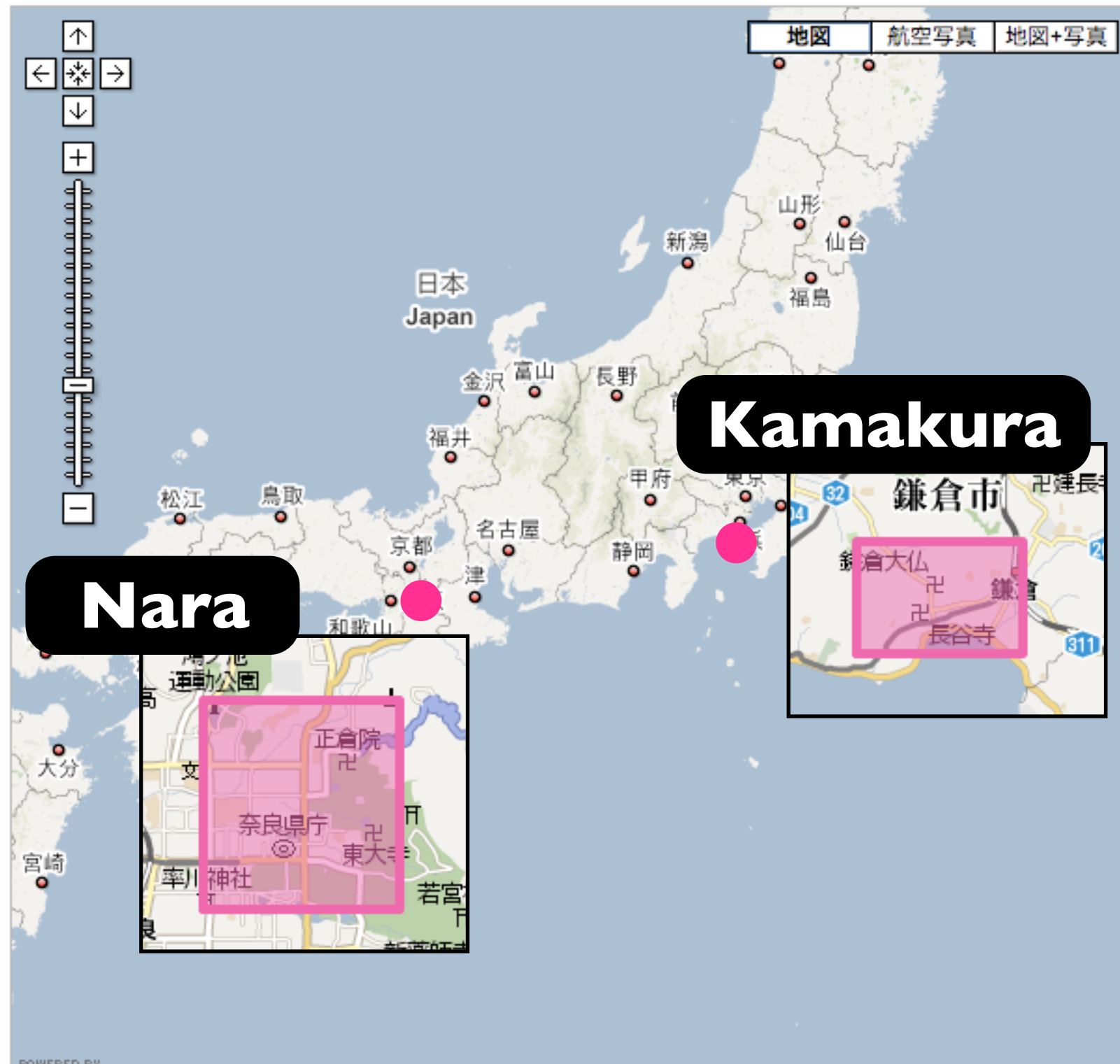
Time usage distribution: 524288



Example of outputs(Location Tag)

Tag: **daibutsu**

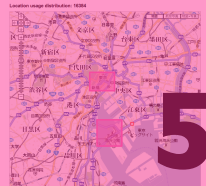
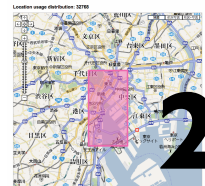
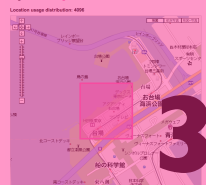
Location usage distribution: 16384



Evaluation

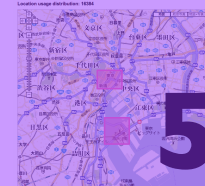
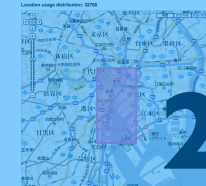
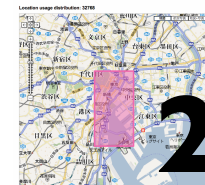
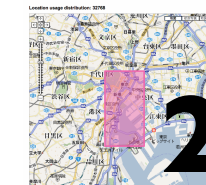
- The outputs for each of the 100 Time and Location tags were manually ranked from 1 (low) to 5 (high).

Naive Scan

	scale1	scale2	scale3
tag1	 4	 5	 2
tag2	 3	 3	 3

Scores for Naive Scan

Proposal method

	scale1	scale2	scale3
tag1	 4	 5	 2
tag2	 5	 2	 2

Scores for Proposal
Ideal scores method

Average Score and Precision

		Average Score	Precision
Time	Ideal	3.74	0.80
	Proposal method	2.70	0.56
	Naive Scan	2.64	0.53
Location	Ideal	3.86	0.84
	Proposal method	2.67	0.57
	Naive Scan	2.62	0.49

- Average score and precision were higher for Proposal method than for Naive Scan
 - Average ideal score is near four and precision is higher than 80%
- ➡ If we can select a suitable scale, we can extract the time or location concept of tags much better corresponding to human recognition

Distribution of difference between each method's scores and ideal scores

Time Tags

Proposal method
- ideal

Bad

-4
-3
-2
-1
0
+1
+2
+3
+4

Good

Mode value: 0

NS - ideal

Bad

-4
-3
-2
-1
0
+1
+2
+3
+4

Good

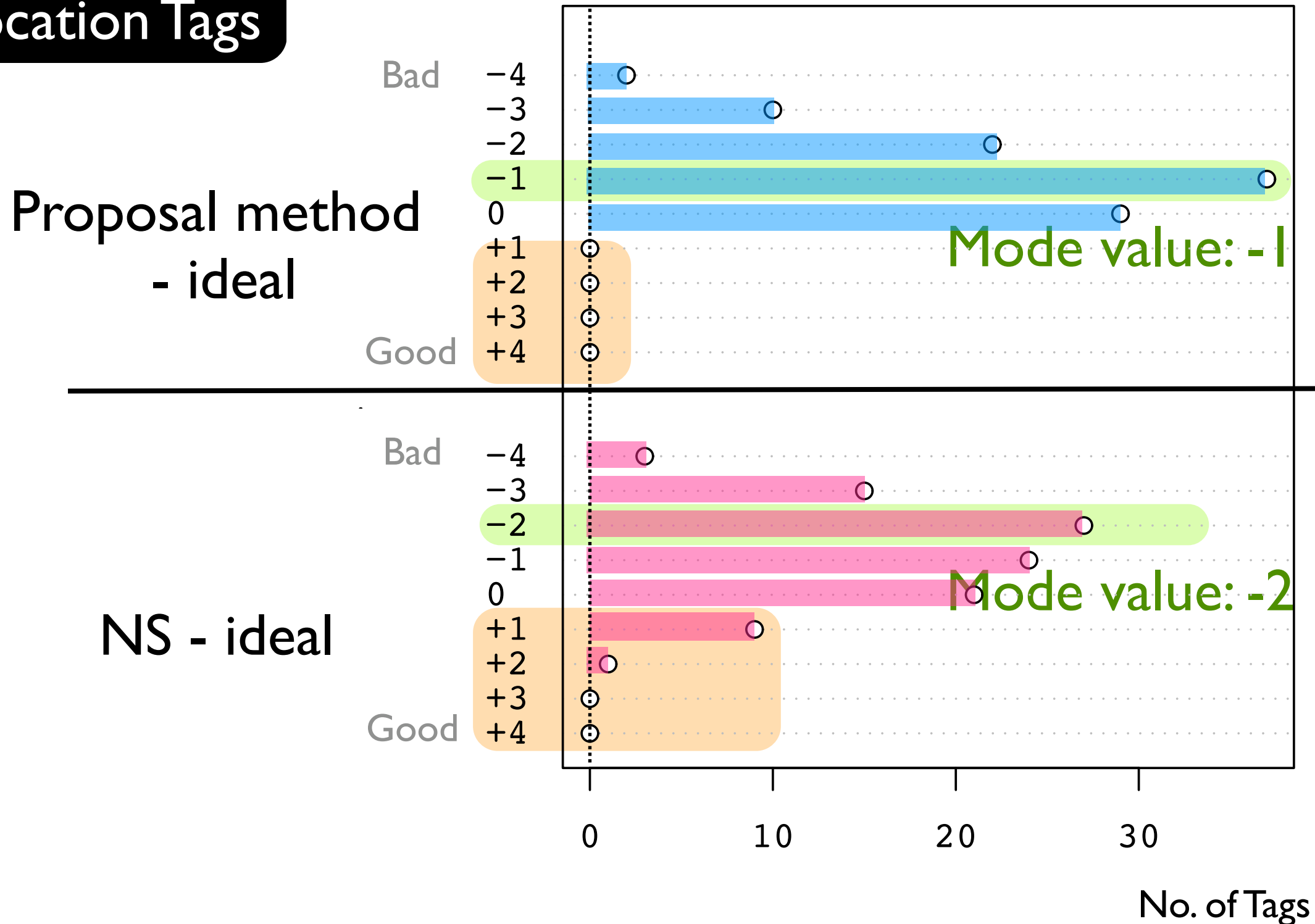
Mode value: -1

Naive Scan's results
vary in quality

No. of Tags

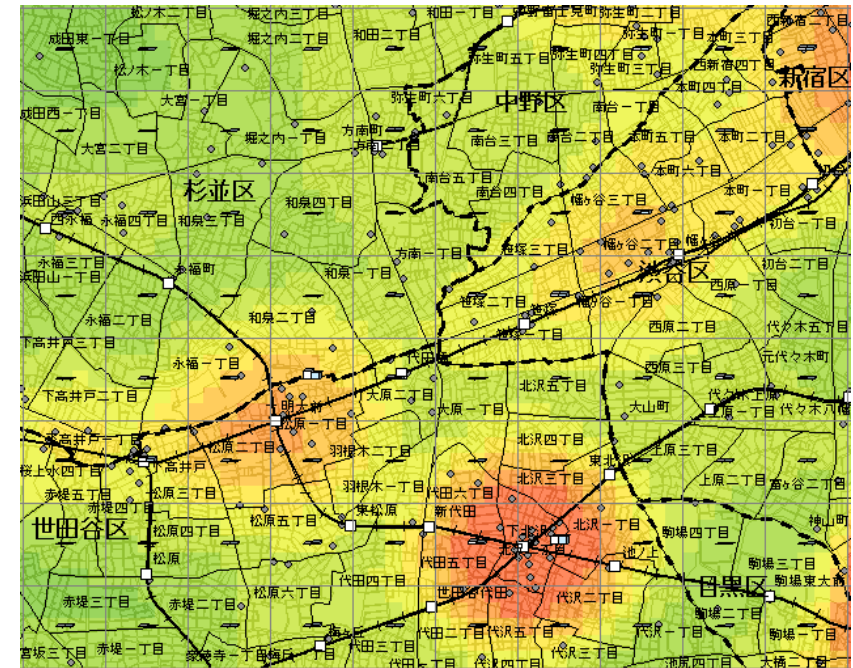
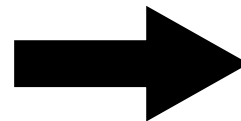
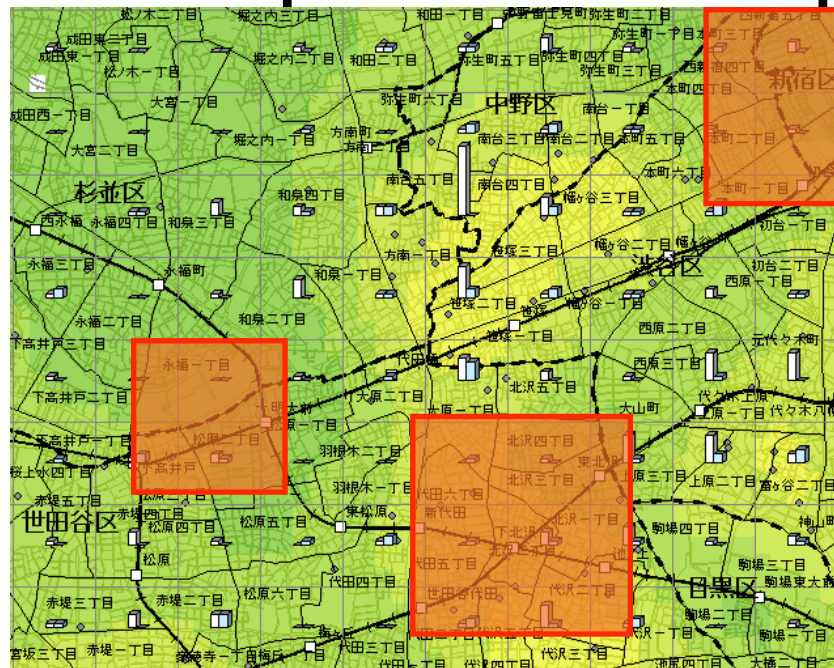
Distribution of difference between each method's scores and Best SSI scores

Location Tags



Conclusion

- We propose a method to extract the temporal/spatial concept related to a tag
- Ongoing work
 - Human recognition of a tag may not involve a well defined region. It might be more fuzzy.
 - I'm trying to extract the tags' concept by probabilistic approach



**Thank you for your
attention :-)**