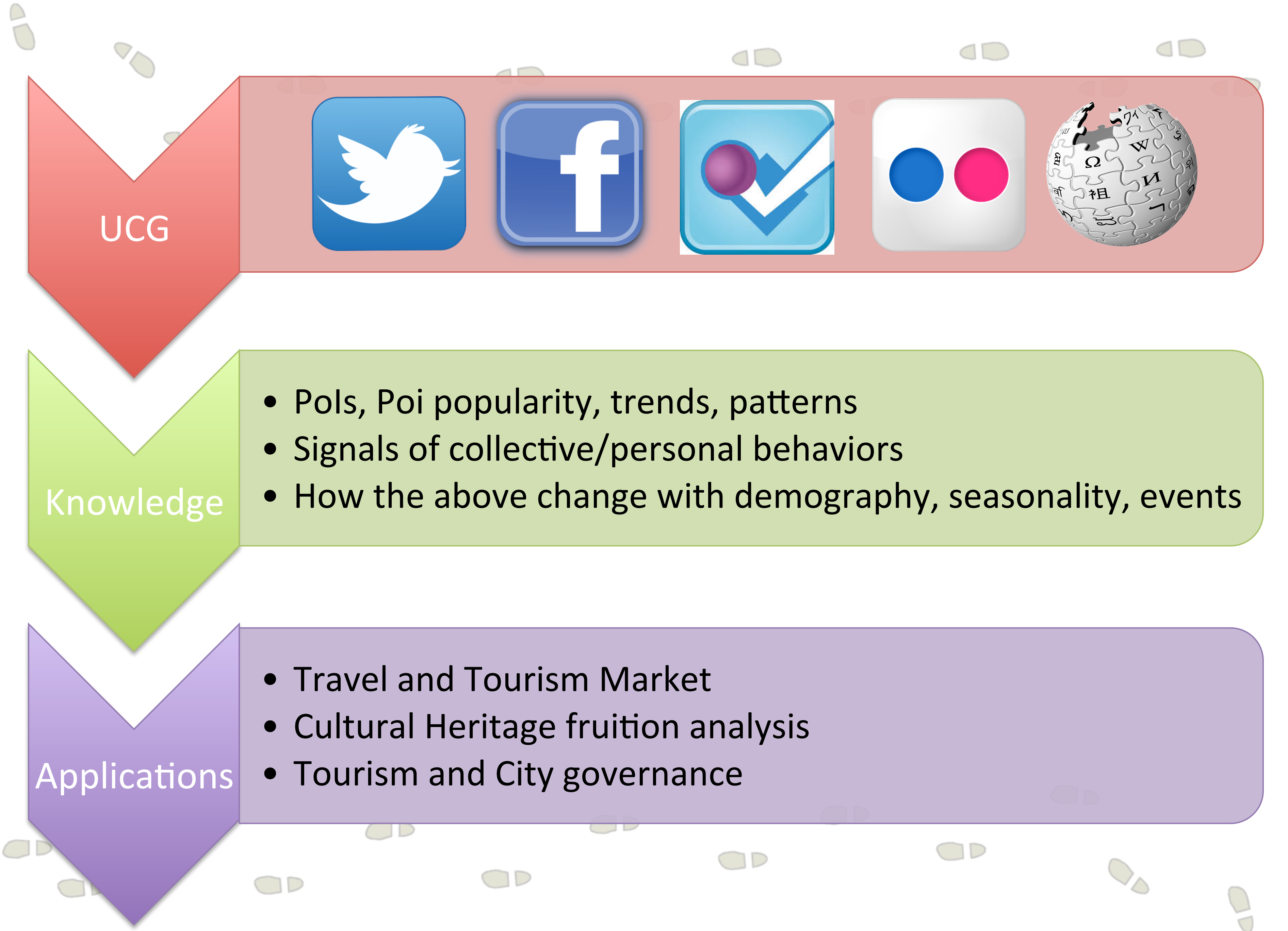
i-ASC 2014 Workshop in conjunction with ECIR 2014, Amsterdam, the Netherlands, 13th April 2014

Mining digital footprints for smart tourism

Raffaele Perego, ISTI-CNR, Italy

I. Brilhante, C. Lucchese, J.F. de Macedo, F.M. Nardini, C. Renso, C. Muntean, F. Silvestri, U. Vespier



UCG

Knowledge

- PoIs, Poi popularity, trends, patterns
- Signals of collective/personal behaviors
- How the above change with demography, seasonality, events

Applications

- Travel and Tourism Market
- Cultural Heritage fruition analysis
- Tourism and City governance

Twitter

- ~1% of tweets are geotagged
- ~25% of geotagged tweets are LBS checkins
- Easy to crawl via the public API
  - 1% but can access more if you ask for a specific topic or place

Facebook

- Huge repository of user personal information
- Friendship constraints on posts visibility
  - Large crawl not possible
- FB Apps can collect data

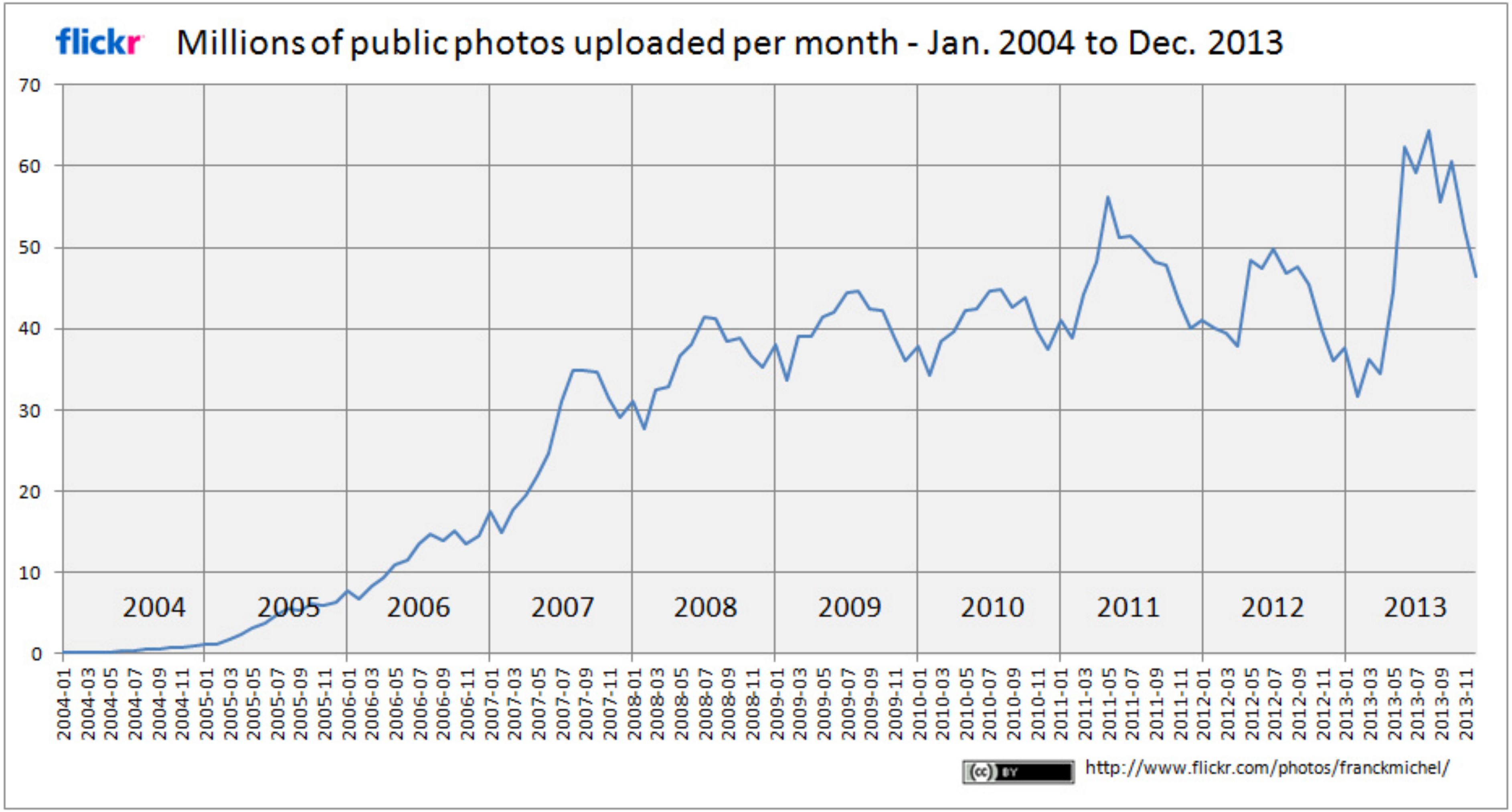
Location-Based Services

- *I'm at Stazione Roma Termini - @fsnews_it (Roma, RM) w/ 12 others http://4sq.com/1l4mk81*
- Users can check-in at a pre-defined place, or create a new place
  - Data biased toward sponsored listings
- Not available for public use:
  - Crawl Twitter as people can choose to publish their Foursquare check-ins on Twitter.

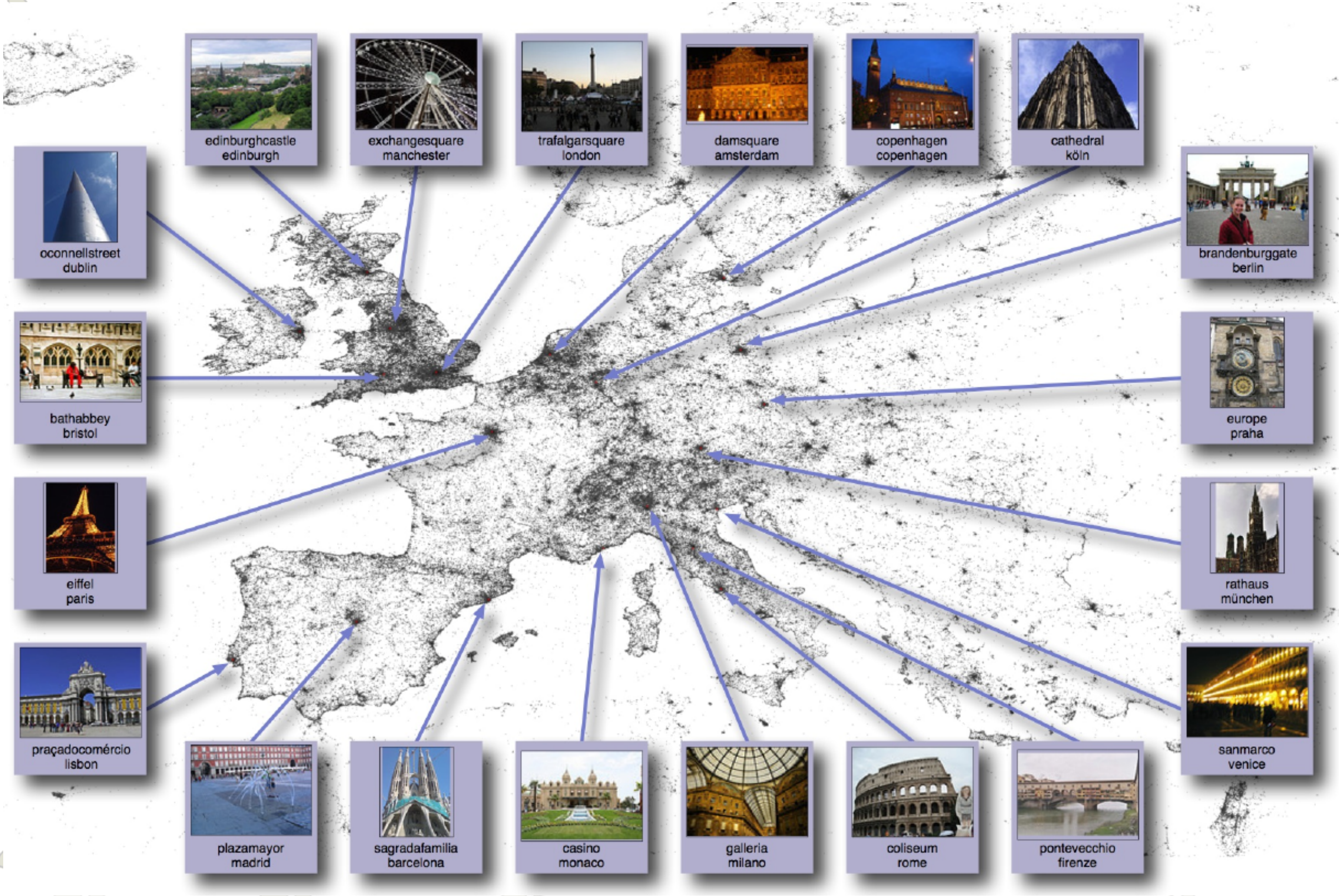
Flickr

- Vast amount of rich data
  - 586 M public Photos uploaded in 2013
  - (Geo-)Tags, Titles, likes, Descriptions, Comments, Social profiles
- Easy to crawl
- Existing large public crawls:
  - CoPhIR: http://cophir.isti.cnr.it/
- Bulk uploading very common

How many photos are uploaded to Flickr every day, month, year?



http://www.flickr.com/photos/franckmichel/

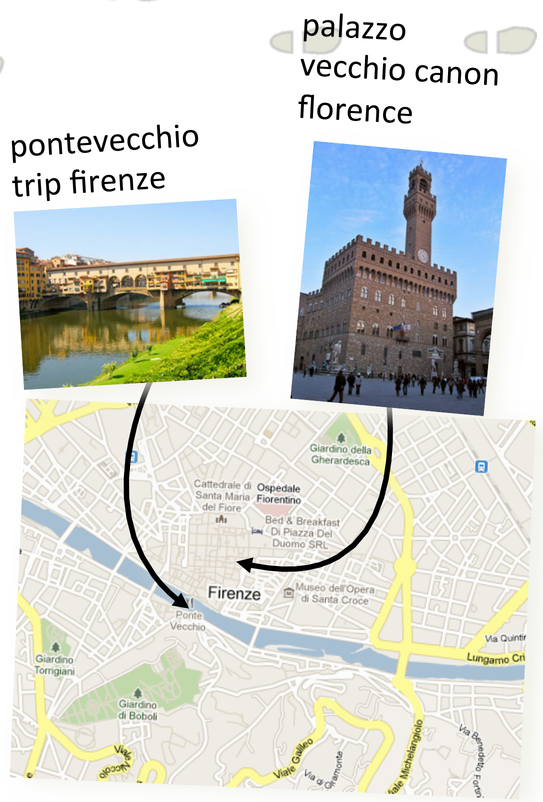


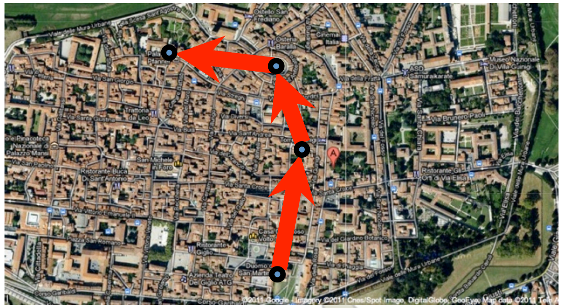
(credits to David Crandall et al., Cornell University)

Our Challenges

Given a large and noisy collection of photo albums taken in a given city:

1. Clean and organize the collection in semantically coherent clusters
2. Associate relevant PoIs with these clusters
3. Devise routes of tourists through these PoIs and characterize as precisely as possible their behaviors
4. Exploit such knowledge to provide personalized recommendations

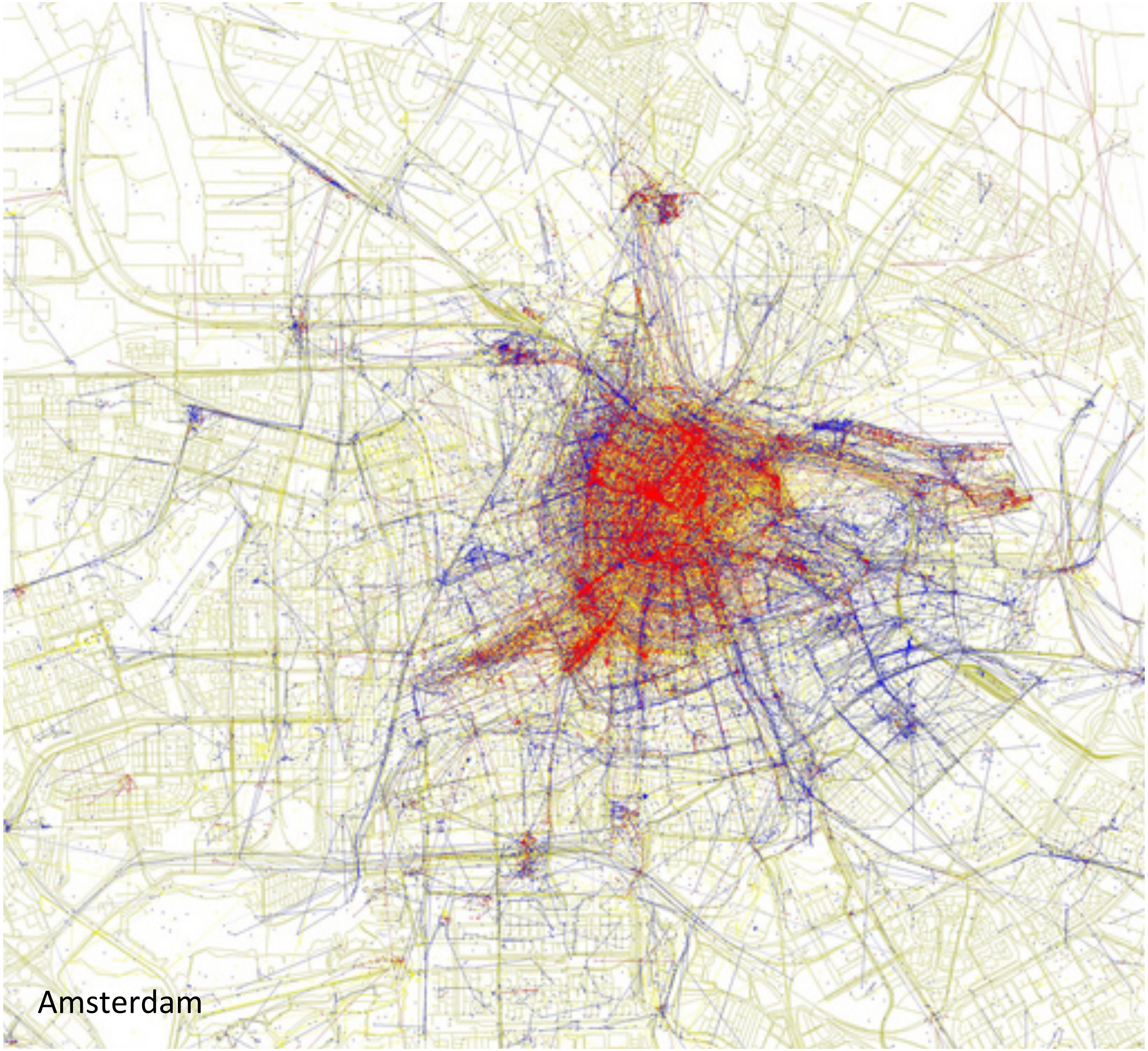


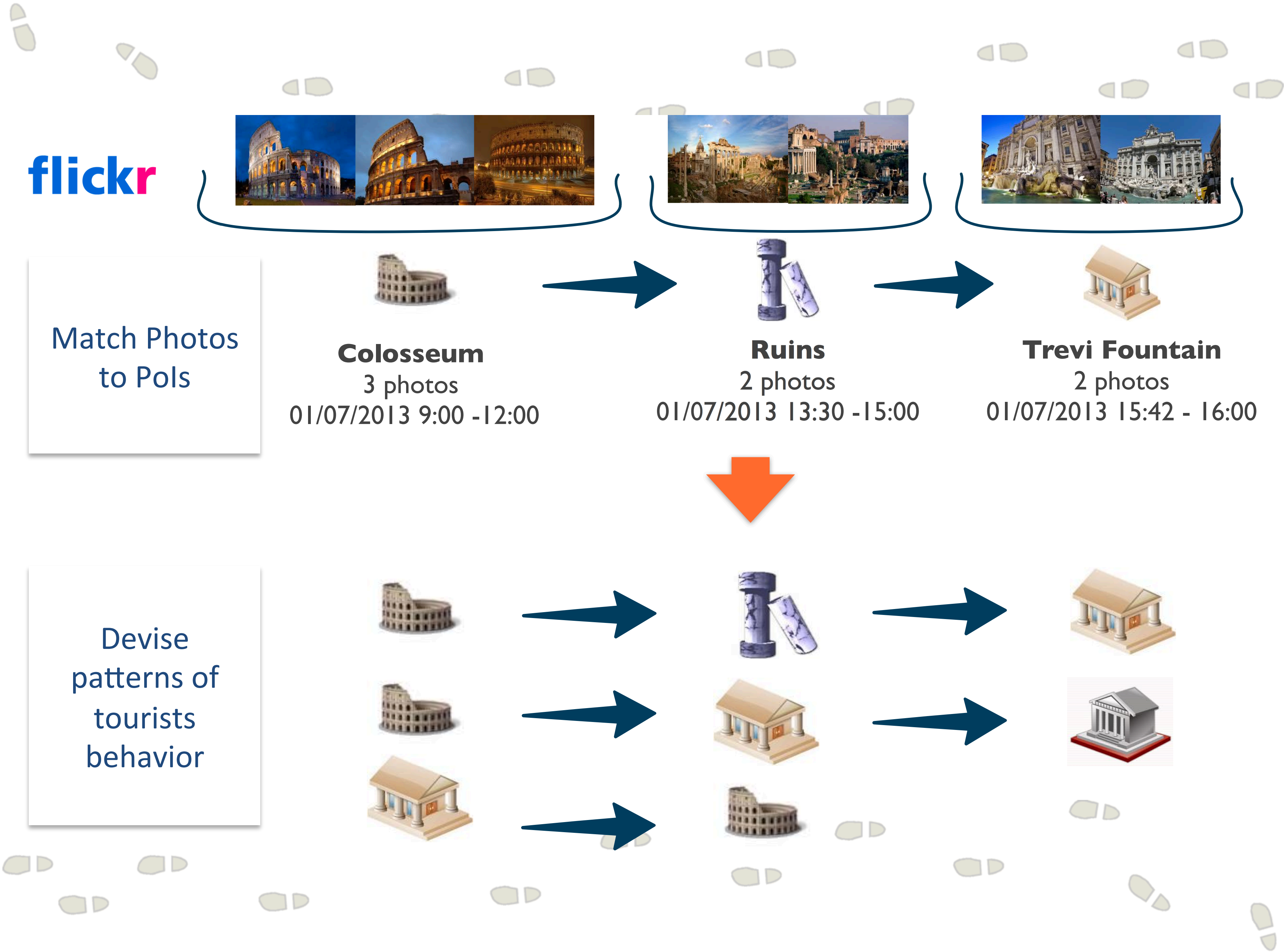


- Where shall we go today?: planning touristic tours with Tripbuilder. CIKM 2013.
- TripBuilder: A Tool for Recommending Sightseeing Tours. Demo. ECIR 2014.
- LearNext: learning to predict tourists movements. CIKM 2013

Locals and tourists

(credits to Erik Fisher)





Match Photos to PoIs

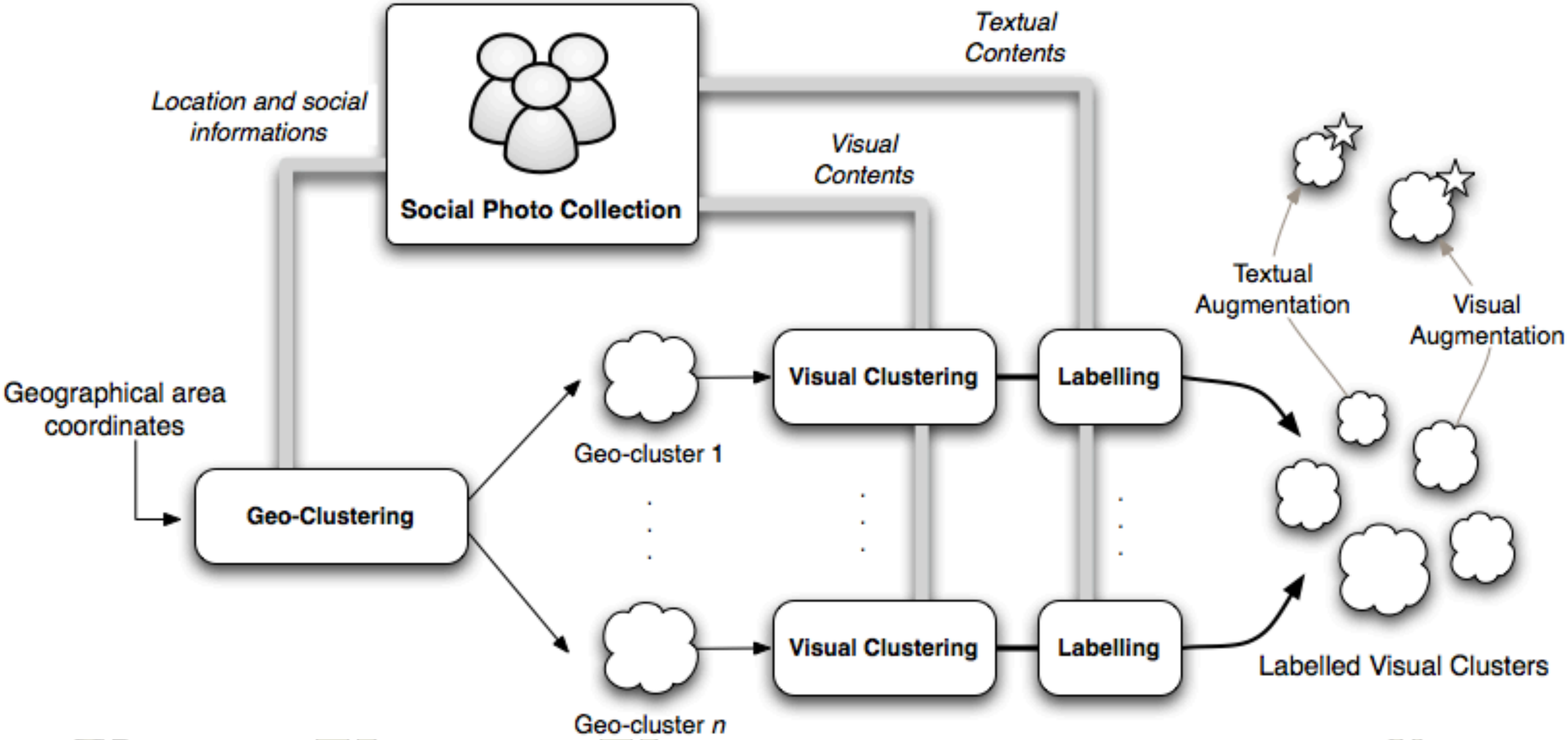
Colosseum
3 photos
01/07/2013 9:00 -12:00

Ruins
2 photos
01/07/2013 13:30 -15:00

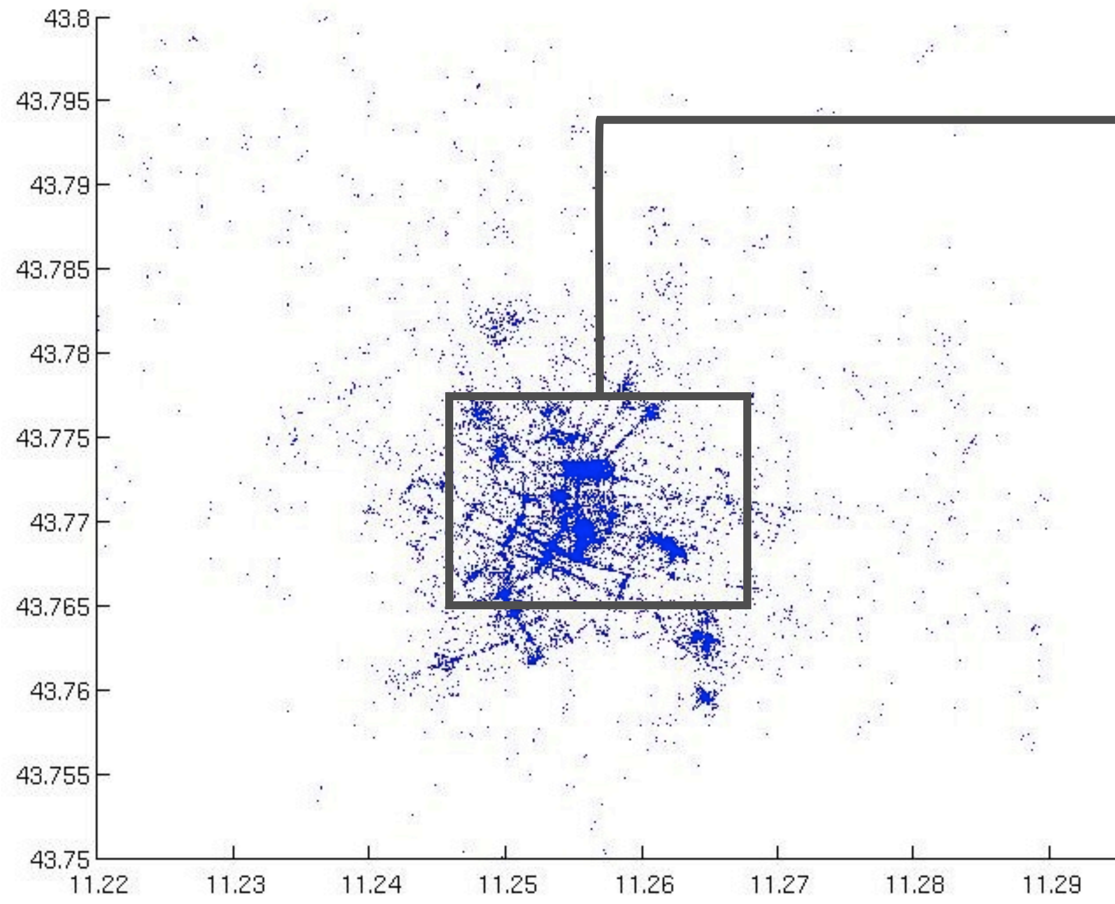
Trevi Fountain
2 photos
01/07/2013 15:42 - 16:00

Devise patterns of tourists behavior

Solution 1: exploit visual content

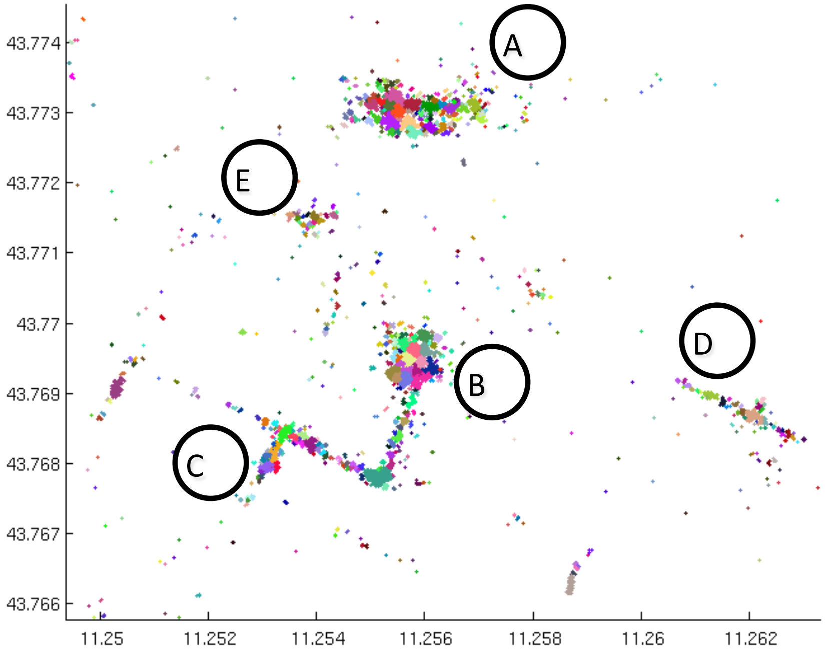


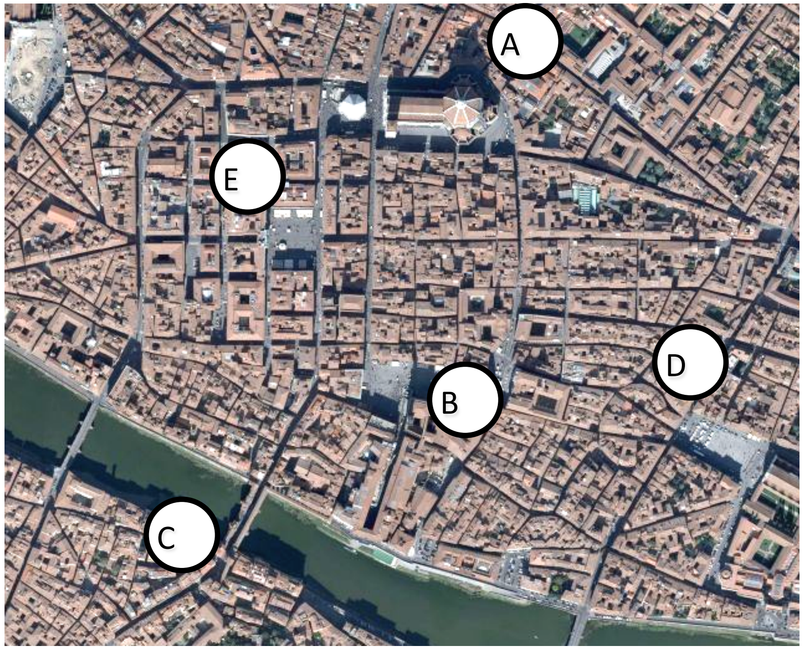
Flickr Geo-Tags in Florence



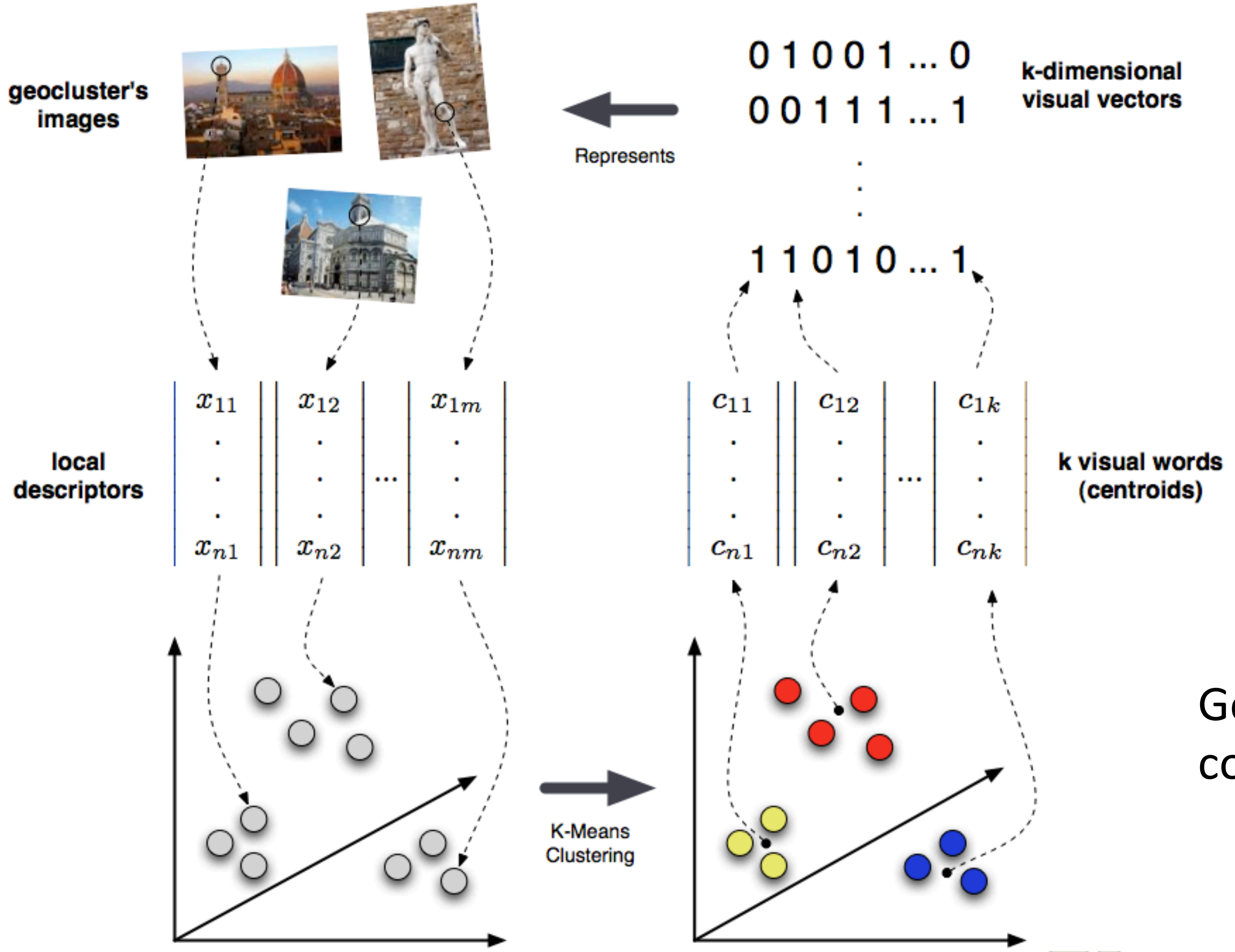
How do we spatially group the photos?

Geo-clustering with DBSCAN





Visual Clustering



Goal: to reduce the cost of computing similarity

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. Computer Vision–ECCV 2006
G Csurka, C. Dance, L Fan, J Willamowski, and C. Bray. Visual categorization with bags of keypoints. ECCV, 2004.

Labeling with tags

florence firenze dmclx2 torres lumix florenz towers italien palazzovecchio tuscany leica tours architecture digital italy landscapes palazzodellasignoria torri

Two key ideas:

- Using the *spatial relevance* of tags
  - Measure: ratio between the tag area and the overall geographical area analyzed
- Using the *social relevance* of tags
  - Measure: number of different users using a given tag

$$\textsc{GeoRelevance}(tag_k) < t_{geo} \wedge \textsc{SocialRelevance}(tag_k) > t_{social}$$

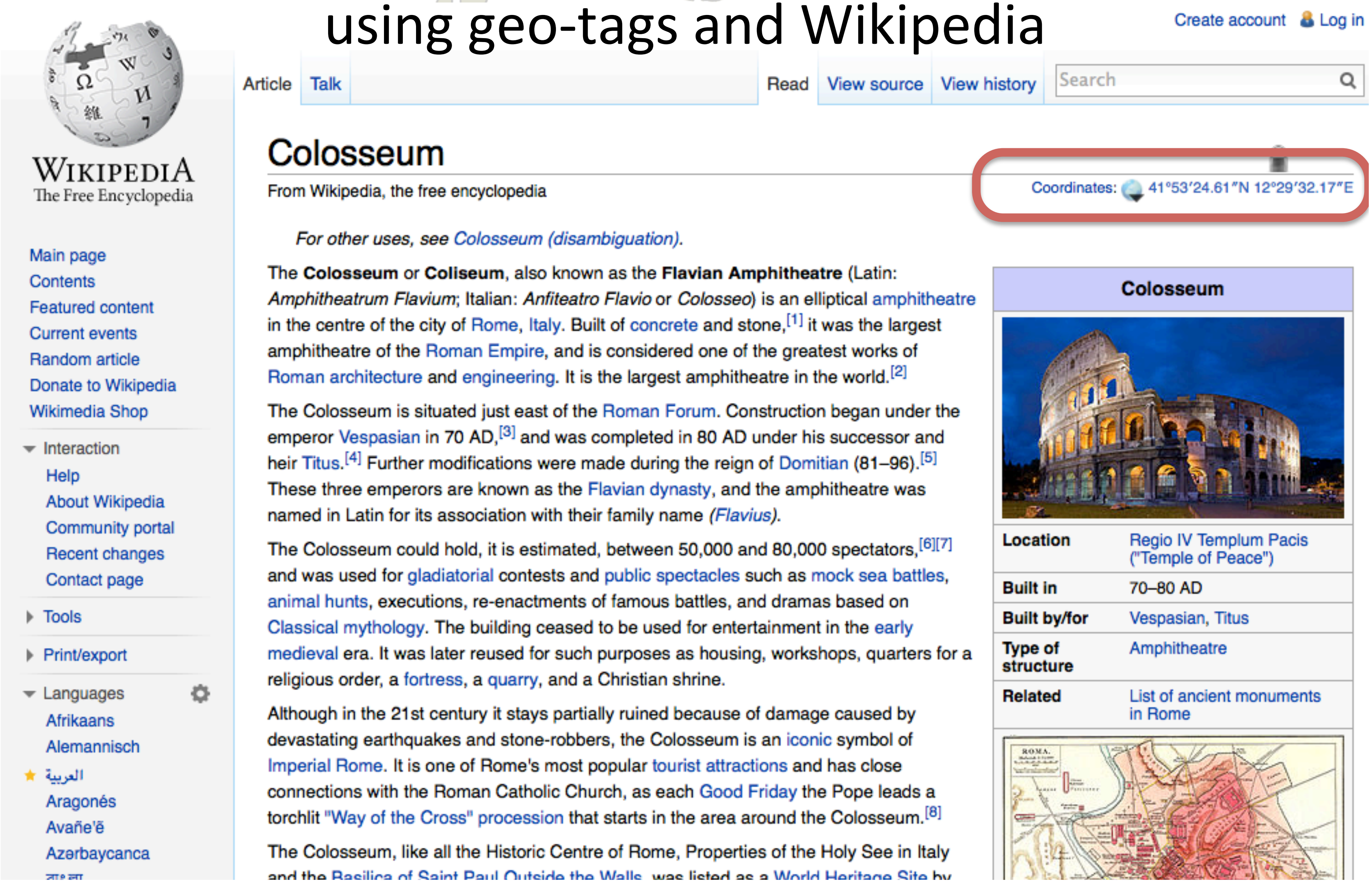
Simple Demo

http://hpc.isti.cnr.it:8000

Flickr photos collection of Florence

Images crawled	53563
Geo-clusters	112
Geo-clustered images	37187
Visual Clusters	743
Visual clustered images	4235

Solution 2:
using geo-tags and Wikipedia



Colosseum

From Wikipedia, the free encyclopedia

Coordinates: 41°53′24.61″N 12°29′32.17″E

For other uses, see Colosseum (disambiguation).

The **Colosseum** or **Coliseum**, also known as the **Flavian Amphitheatre** (Latin: *Amphitheatrum Flavium*; Italian: *Anfiteatro Flavio* or *Colosseo*) is an elliptical amphitheatre in the centre of the city of Rome, Italy. Built of concrete and stone,[1] it was the largest amphitheatre of the Roman Empire, and is considered one of the greatest works of Roman architecture and engineering. It is the largest amphitheatre in the world.[2]

The Colosseum is situated just east of the Roman Forum. Construction began under the emperor Vespasian in 70 AD,[3] and was completed in 80 AD under his successor and heir Titus.[4] Further modifications were made during the reign of Domitian (81–96).[5] These three emperors are known as the Flavian dynasty, and the amphitheatre was named in Latin for its association with their family name *(Flavius)*.

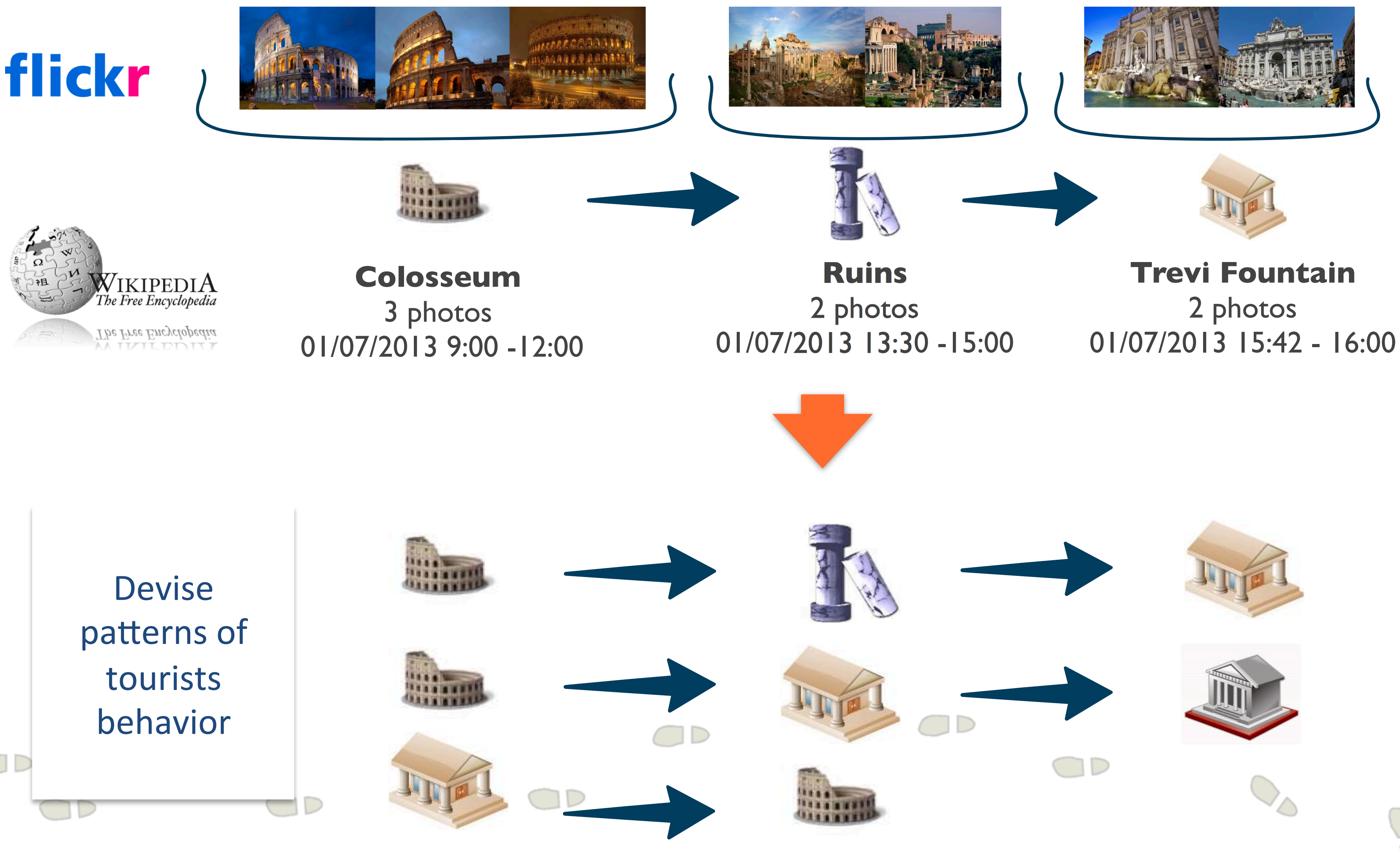
The Colosseum could hold, it is estimated, between 50,000 and 80,000 spectators,[6][7] and was used for gladiatorial contests and public spectacles such as mock sea battles, animal hunts, executions, re-enactments of famous battles, and dramas based on Classical mythology. The building ceased to be used for entertainment in the early medieval era. It was later reused for such purposes as housing, workshops, quarters for a religious order, a fortress, a quarry, and a Christian shrine.

Although in the 21st century it stays partially ruined because of damage caused by devastating earthquakes and stone-robbers, the Colosseum is an iconic symbol of Imperial Rome. It is one of Rome's most popular tourist attractions and has close connections with the Roman Catholic Church, as each Good Friday the Pope leads a torchlit "Way of the Cross" procession that starts in the area around the Colosseum.[8]

The Colosseum, like all the Historic Centre of Rome, Properties of the Holy See in Italy and the Basilica of Saint Paul Outside the Walls, was listed as a World Heritage Site by

Colosseum	
Location	Regio IV Templum Pacis ("Temple of Peace")
Built in	70–80 AD
Built by/for	Vespasian, Titus
Type of structure	Amphitheatre
Related	List of ancient monuments in Rome

Trajectories from Flickr & Wikipedia



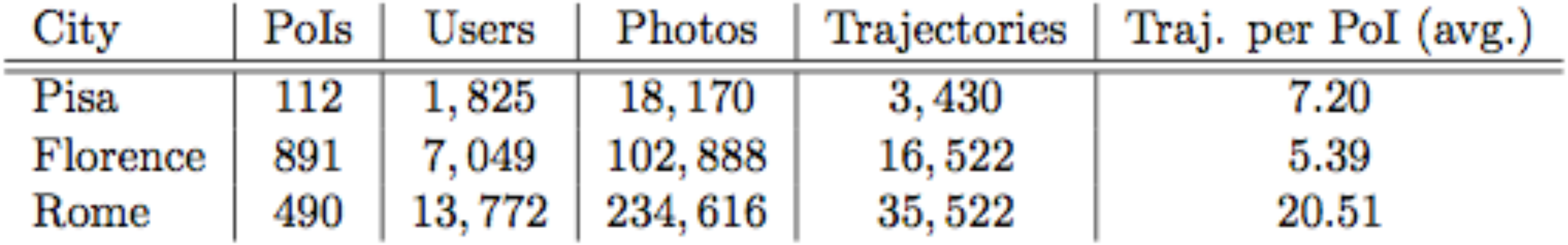
Colosseum
3 photos
01/07/2013 9:00 -12:00

Ruins
2 photos
01/07/2013 13:30 -15:00

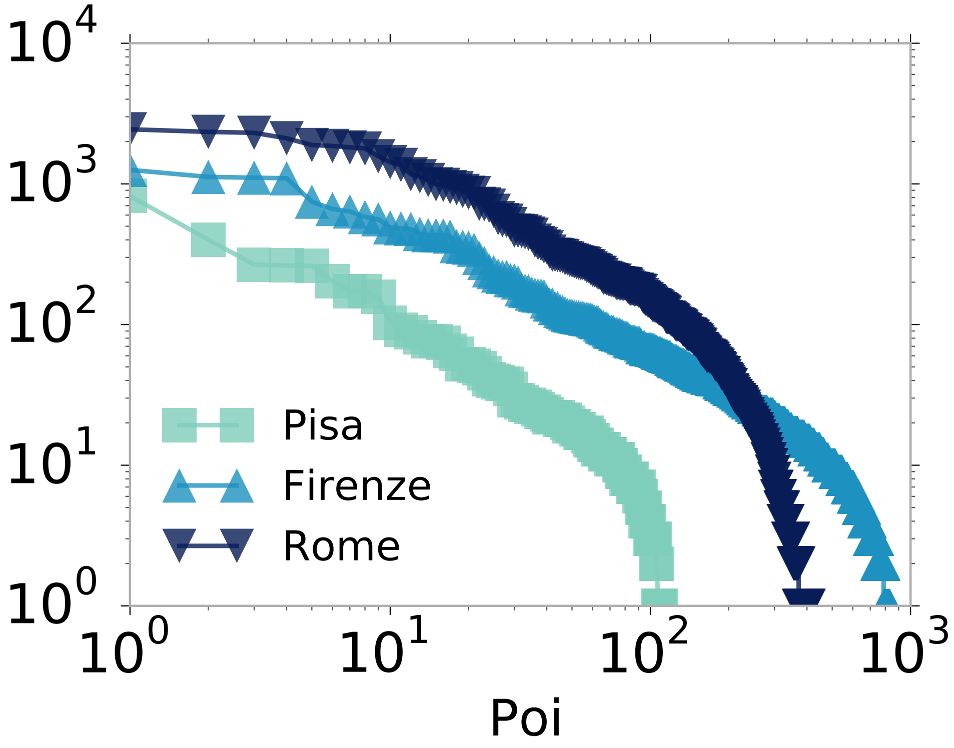
Trevi Fountain
2 photos
01/07/2013 15:42 - 16:00

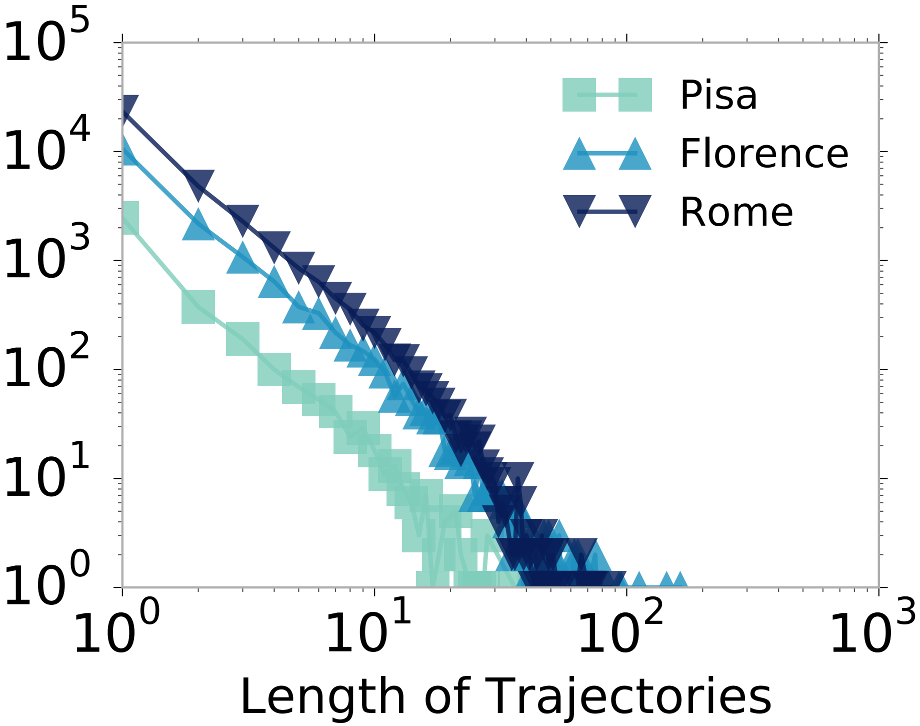
Devise patterns of tourists behavior

Datasets



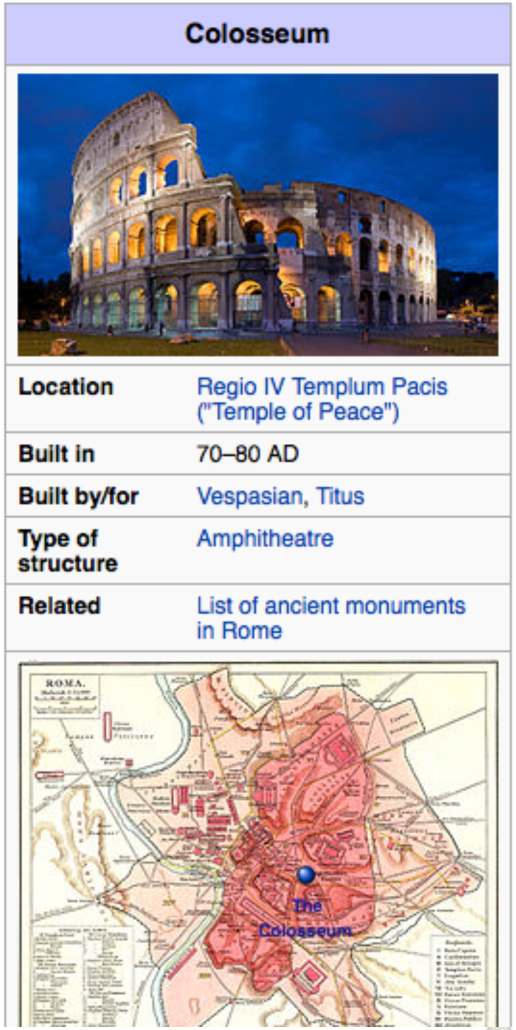
City	PoIs	Users	Photos	Trajectories	Traj. per PoI (avg.)
Pisa	112	1, 825	18, 170	3, 430	7.20
Florence	891	7, 049	102, 888	16, 522	5.39
Rome	490	13, 772	234, 616	35, 522	20.51





Wikipedia Categories

- Categories for Colosseum:
  - 1st-century architecture
  - Amphitheatres in Rome
  - Ancient Roman architecture
  - Building projects of the Flavian dynasty
  - Roman archaeology
  - Ruins in Italy



Colosseum	
Location	Regio IV Templum Pacis ("Temple of Peace")
Built in	70–80 AD
Built by/for	Vespasian, Titus
Type of structure	Amphitheatre
Related	List of ancient monuments in Rome

User and Trajectory profiling

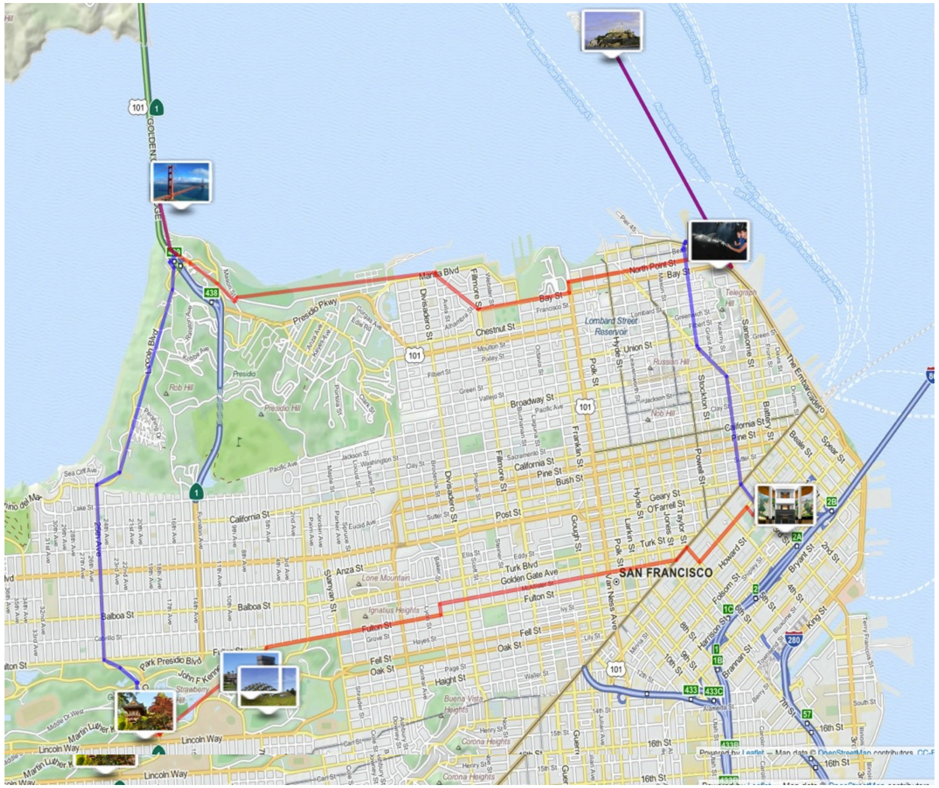
Categories: Amphitheatres in Rome, Ancient Roman architecture, Roman archaeology, Ruins in Italy

- A profile for a user can be built from the categories of the PoIs visited
  - We can measure the “interest” of a user for a (set of) PoI, e.g.:

$$\Gamma(p, u) = \alpha \cdot sim(\vec{v_p}, \vec{v_u}) + (1 - \alpha) \cdot pop(p)$$

- Each trajectory can be labeled with the set of categories of the constituent PoIs

Planning Sightseeing Tours with TripBuilder



What should I visit in San Francisco?

Given:

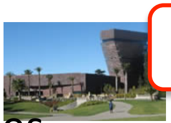
- Time: 2 days;
- My preferences.

Golden Gate Bridge

Golden Gate Park

4 h

California Academy of Sciences


de Young Museum

4 h

San Francisco Museum of Modern Art

Aquarium of the Bay

Alcatraz

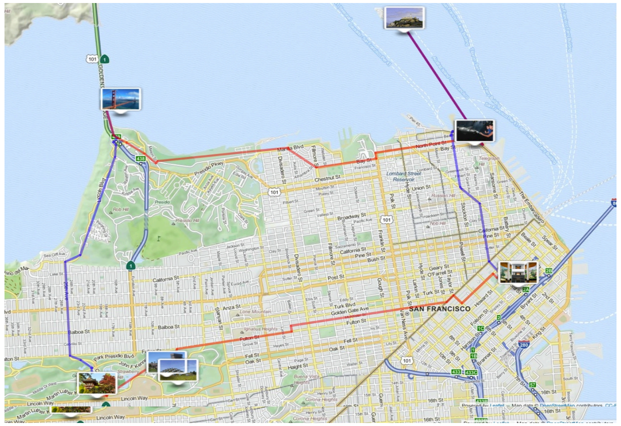
8 h

How many of these “trajectories” can I enjoy?

How do these tourists visit such places?

The TripCover Problem

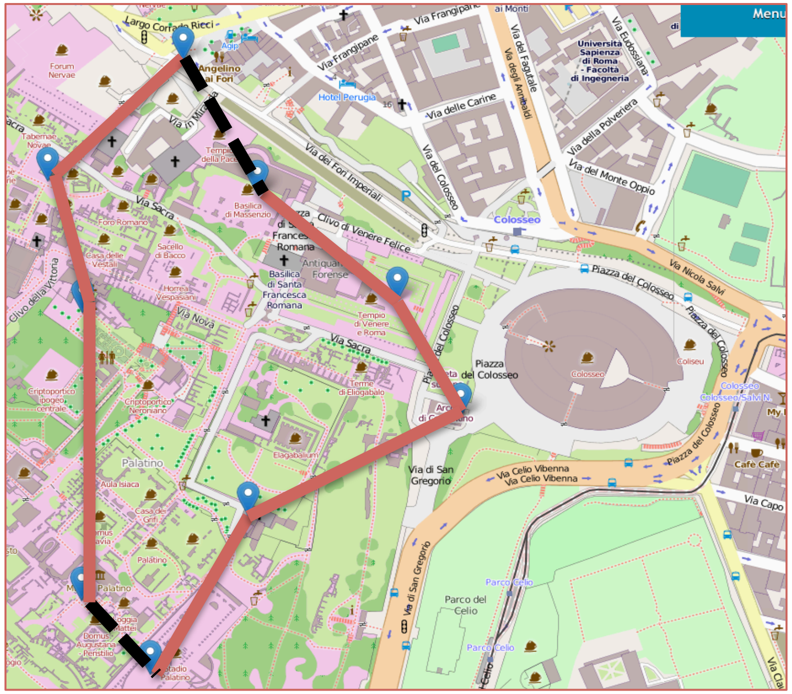
- **Given:**
  - A set of popular trajectories crossing a set of PoIs and their time cost
  - The relevance of the trajectories w.r.t. the category set
  - The **Time Budget** and **Preferences** of a user
  - A measure of **PoI-User interest**
- **Find:**
  - the subset of trajectories that maximizes user interest and fits in the time budget

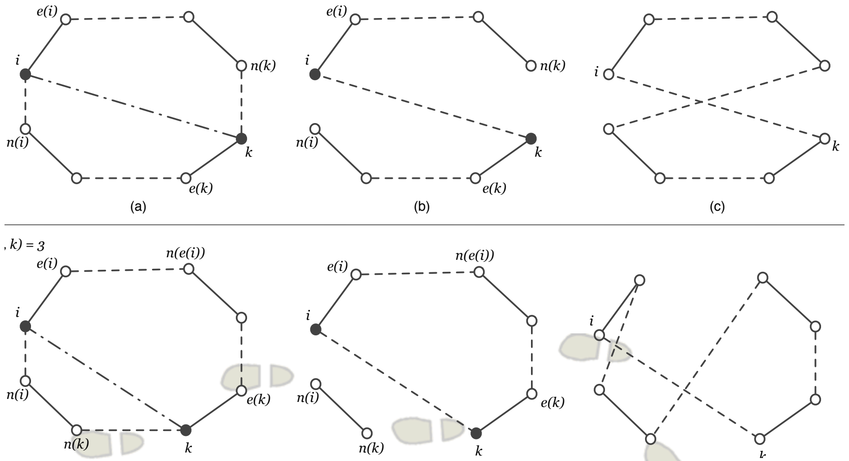


TripCover is an instance of the **Generalized Maximum Coverage** (GMC) problem. NP-Hard with a (e/(e-1))-approximation algorithm.

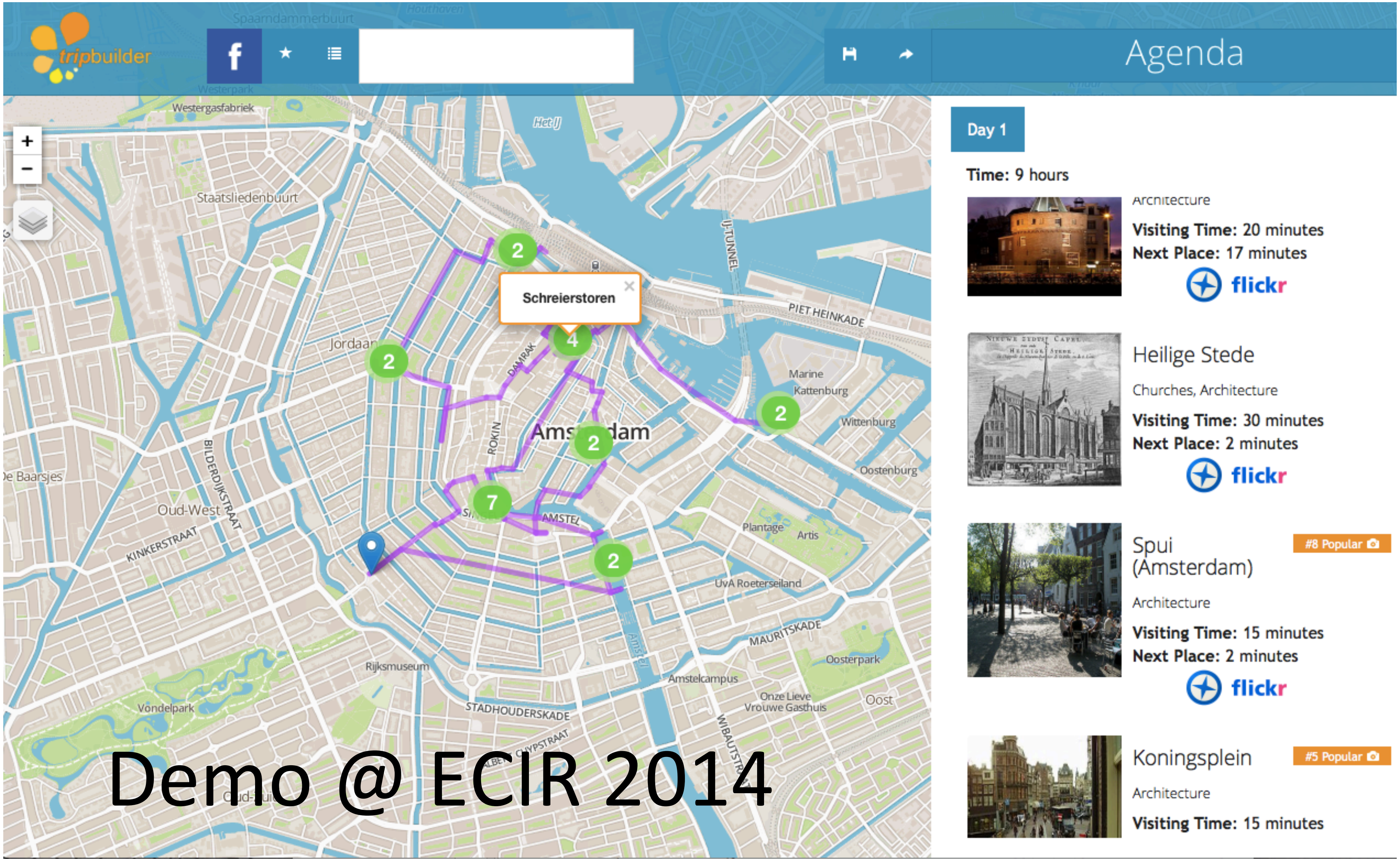
TrajSP: joining the trajectories

- TripCover solution is a set of trajectories fitting user interest and time budget
  - Local search heuristics based on 2-opt and 3-opt moves for connecting the solution in a single sightseeing tour





http://tripbuilder.isti.cnr.it



Agenda

Day 1

Time: 9 hours

Architecture

Visiting Time: 20 minutes

Next Place: 17 minutes

flickr

Heilige Stede

Churches, Architecture

Visiting Time: 30 minutes

Next Place: 2 minutes

flickr

Spui (Amsterdam) #8 Popular

Architecture

Visiting Time: 15 minutes

Next Place: 2 minutes

flickr

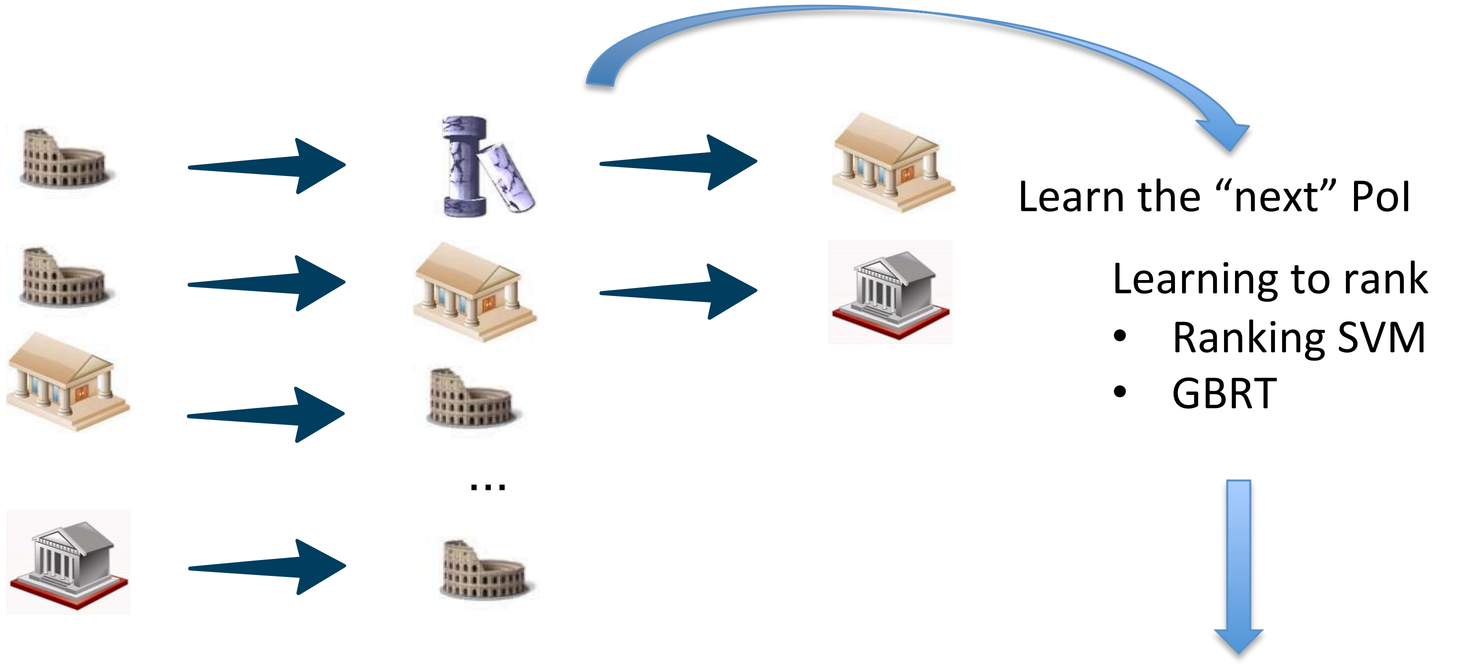
Koningsplein #5 Popular

Architecture

Visiting Time: 15 minutes

Demo @ ECIR 2014

Next PoI prediction with LearNext



Learn the “next” PoI

Learning to rank

- Ranking SVM
- GBRT

Feature engineering

Features

🟥 Popularity	🟦 Frequent seq. of PoIs	🟩 User preference	🟪 Distance and time	🟫 Poi Characteristics	▣ Session Characteristics

Session features:

- 🟪 actualTransferTime
- 🟪 actualVisitTime
- 🟪 distLat_Avg/Max/Min/Tot
- 🟪 distLen_Avg/Max/Min/Tot
- 🟪 euclideanDist_Avg/Max/Min/Tot
- 🟪 sessTime
- ▣ phPoISess_Avg/Max/Min/Tot
- ▣ categsPerSess
- ▣ uniqueCategsPerSess
- ▣ sessLen
- 🟩 userSessLen_Avg/Max/Min/Tot
- 🟩 userSessRatio

Candidate PoI features:

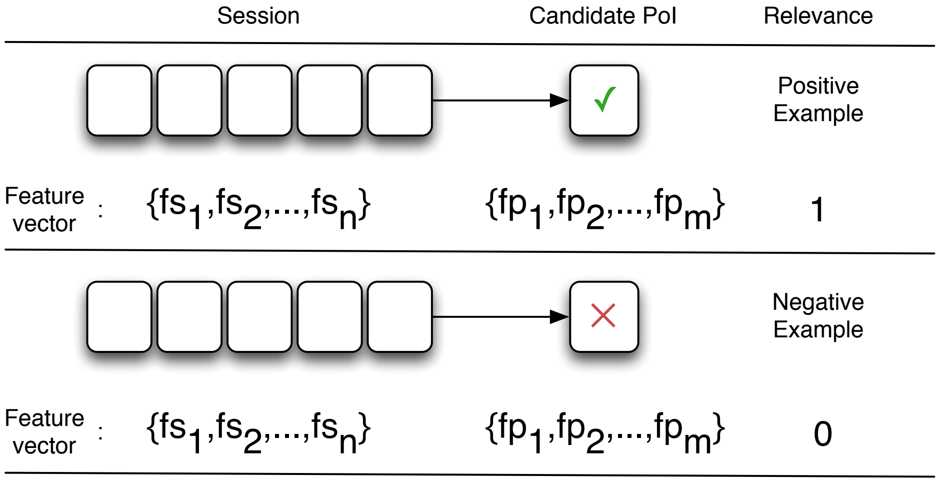
- 🟪 distFromFirstPoI_Eucl/Lat/Len
- 🟪 distFromLastPoI_Eucl/Lat/Len
- 🟪 visitTimePoI_User
- 🟪 visitTime_Avg/Max/Min/StDev/Tot
- 🟦 freqBigrams
- 🟦 freqTrigrams
- 🟫 start/stop/middleProbab
- 🟫 cat1, cat2, ..., cat10
- 🟫 entropy
- 🟫 numCategs
- 🟥 numPhotos_Avg/Max/Min/Tot
- 🟥 noOfVisists
- 🟥 photosPoIUserId_Avg/Total
- 🟥 ratioPhotosPoI
- 🟥 ratioSessWithPoI
- 🟥 ratioUsersVisitingPoI
- 🟩 photosPerUser
- 🟩 ratioPoIInUserPhotos

Learning to Rank approach

- Building a model that ranks highest the PoI most likely to be visited as next by the tourist.
  - A trail is represented by a 68-dimension feature space.
  - Each example is represented by the feature vector and its label indicates the PoI’s degree of relevance to the user {0,1}.
- The learning algorithm is trained to predict the relevance from the feature vector. Approaches:
  - Ranking SVM
  - GBRT

Experiments

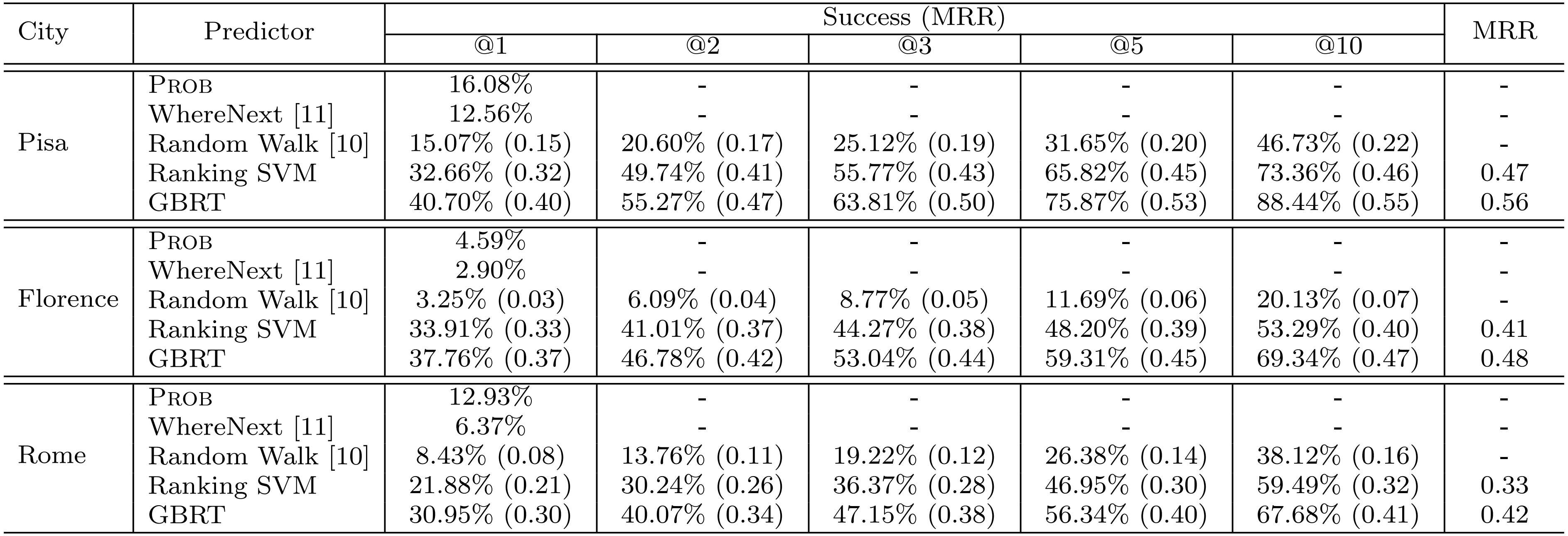
- Baselines:
  - Probability: suggesting the most probable next PoI from the current one
  - WhereNext[1]: a trajectory pattern mining approach
  - Random Walk[2]: a graph based approach "Itinerary Graph" exploiting RWR
- Training/test methodology
  - Training: 1 positive, 3 negative
  - Test: all unseen PoIs for the trail
- Metrics:
  - Success@k
  - MRR@k
  - MRR



	Session	Candidate PoI	Relevance
		✓	Positive Example
Feature vector :	$\{fs_1, fs_2, \ldots, fs_n\}$	$\{fp_1, fp_2, \ldots, fp_m\}$	1
		×	Negative Example
Feature vector :	$\{fs_1, fs_2, \ldots, fs_n\}$	$\{fp_1, fp_2, \ldots, fp_m\}$	0

1. A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. "WhereNext: a location predictor on trajectory pattern mining". In Proc. SIGKDD. ACM, 2009.
2. C. Lucchese, R. Perego, F. Silvestri, H. Vahabi, and R. Venturini. "How random walks can help tourism". In Proc. ECIR. LNCS, 2012.

Results



City	Predictor	Success (MRR)					MRR
		@1	@2	@3	@5	@10	
Pisa	Prob	16.08%	-	-	-	-	-
	WhereNext [11]	12.56%	-	-	-	-	-
	Random Walk [10]	15.07% (0.15)	20.60% (0.17)	25.12% (0.19)	31.65% (0.20)	46.73% (0.22)	-
	Ranking SVM	32.66% (0.32)	49.74% (0.41)	55.77% (0.43)	65.82% (0.45)	73.36% (0.46)	0.47
	GBRT	40.70% (0.40)	55.27% (0.47)	63.81% (0.50)	75.87% (0.53)	88.44% (0.55)	0.56
Florence	Prob	4.59%	-	-	-	-	-
	WhereNext [11]	2.90%	-	-	-	-	-
	Random Walk [10]	3.25% (0.03)	6.09% (0.04)	8.77% (0.05)	11.69% (0.06)	20.13% (0.07)	-
	Ranking SVM	33.91% (0.33)	41.01% (0.37)	44.27% (0.38)	48.20% (0.39)	53.29% (0.40)	0.41
	GBRT	37.76% (0.37)	46.78% (0.42)	53.04% (0.44)	59.31% (0.45)	69.34% (0.47)	0.48
Rome	Prob	12.93%	-	-	-	-	-
	WhereNext [11]	6.37%	-	-	-	-	-
	Random Walk [10]	8.43% (0.08)	13.76% (0.11)	19.22% (0.12)	26.38% (0.14)	38.12% (0.16)	-
	Ranking SVM	21.88% (0.21)	30.24% (0.26)	36.37% (0.28)	46.95% (0.30)	59.49% (0.32)	0.33
	GBRT	30.95% (0.30)	40.07% (0.34)	47.15% (0.38)	56.34% (0.40)	67.68% (0.41)	0.42

Print-outs of all the images uploaded to Flickr in a day
(installation by Erik Kessels, Amsterdam)

THANKS!