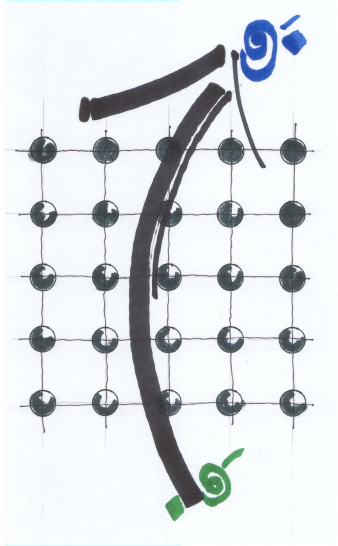
PRIN 2004 Project
GeoPKDD

Geographic Privacy-aware Knowledge Discovery and Delivery



MID-TERM MEETING

Venezia, 17-18 ottobre 2005

Agenda of the meeting

Monday, October 17, Ca' Dolfin (Dorsoduro)

14:00-15:00

Introduction to the project scientific context and objectives (for our special guests); Communications from the coordinator; Relation with the starting European-level GeoPKDD

15:00-16:30

Alignment Reports: discussion

(I responsabili dei 6 Alignment Reports illustrano le scelte fatte, la struttura del report, lo stato attuale, etc.)

16:30-16:45 Pausa Caffe

Agenda of the meeting

Monday, October 17, Ca' Dolfin (Dorsoduro)

16:45-17:45 WP1 (Trajectory Warehouse)

Maria Damiani: *Spatial data warehousing & security/privacy in LBS*

Andrea Mazzoni: *CENTRE - un generatore di dati di posizionamento per reti cellulari*

Alessandra Raffaetà: *Aggregati per traiettorie* ***

17:45-19:45 WP2 (ST Data Mining):

Giuseppe Manco: *Model based clustering of trajectories*

Mirco Nanni: *Time-Focuesd Density Based Clustering of Trajectories*

Mirco Nanni: *Mining Sequences with Temporal Annotations*

Salvo Rinzivillo: *Spatial Clustering*

CENA

Agenda of the meeting

Tuesday, October 18, Ca' Dolfin (Dorsoduro)

09:00-10:00

Invited Talk by Carlo Zaniolo (UCLA): *Temporal Queries in Decision Support and Business Intelligence: Should we Use SQL or XML?*

10:00-11:00 WP2 (ST Data Mining):

Domenico Talia: *WEKA4WS: enabling distributed data mining on grids*

Claudio Silvestri : *Approximate mining di frequent itemsets from distributed and streamed data sources*

11:00-11:15 Pausa Caffè

Agenda of the meeting

Tuesday, October 18, Ca' Dolfin (Dorsoduro)

11:15-12:00 : WP2 (Privacy-aware Data Mining)
Maurizio Atzori: *Anonymity-aware Data Mining*
Mimmo Saccà: *Visione sul tema Privacy*

12:00-13:00: Discussione finale: Pianificazione attività secondo anno (parte 1)

13:00-14:00 Pausa pranzo

14:00-16:00: Discussione finale: Pianificazione attività secondo (parte 2)

Alignment reports

[TR1.1] Report di allineamento sul warehousing per flussi continui di dati di oggetti in movimento e relative problematiche di privacy e security, eventuale specifica preliminare dei requisiti.

[TR1.2] Report di allineamento su tecniche di data mining spaziale e spazio-temporale.

[TR1.3] Report di allineamento su tecniche di data mining con rispetto della privacy.

[TR1.4] Report di allineamento su tecniche e sistemi per data mining distribuito.

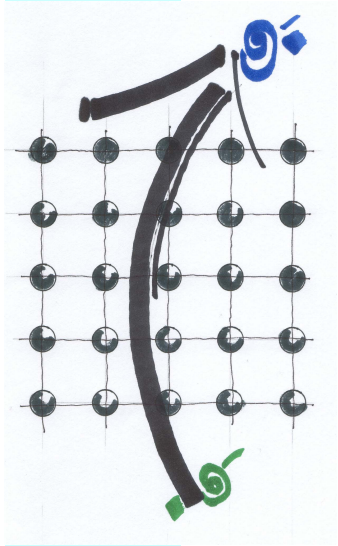
[TR1.5] Report di allineamento su tecniche di ragionamento su dati spazio-temporali.

[TR1.6] Report su caratterizzazione delle applicazioni GeoPKDD e considerazioni di fattibilita' preliminari.

FET Project IST-014915: **GeoPKDD**

Geographic Privacy-aware Knowledge Discovery and Delivery

November 2005-October 2008



Dino Pedreschi and Fosca Giannotti

MSTD @ ECML/PKDD 2005

Porto, October 3rd

The consortium

ID	Acronym	Partner	Country
1	**KDDLAB**	Knowledge Discovery and Delivery Laboratory, ISTI-CNR, Istituto di Scienza e Tecnologie dell'Informazione, Pisa. http://www.isti.cnr.it/ - jointly with Univ. Pisa, Dept. of Computer Science http://www.di.unipi.it	I
2	**LUC**	Univ. Limburg, Theoretical Computer Science Group. http://www.luc.ac.be/theocomp	B
3	**EPFL**	EPFL, Lab. DB, Lausanne. http://lbdwww.epfl.ch/e/	CH
4	**FAIS**	Fraunhofer Institute for Autonomous Intelligent Systems, Sankt Augustin. http://www.ais.fraunhofer.de/	D
5	**WUR**	Wageningen UR, Centre for GeoInformation. http://cgi.girs.wageningen-ur.nl/	NL
6	**CTI**	Research Academic Computer Technology Institute, Research and Development Division. http://www.cti.gr/ - jointly with Univ. Piraeus, Dept. of Informatics http://www.unipi.gr	GR
7	**UNISAB**	Sabanci University, Faculty of Engineering and Natural Sciences. http://www.sabanciuniv.edu/	TK
8	**WIND**	WIND Telecomunicazioni SpA, Direzione Reti Wind Progetti Finanziati & Technology Scouting.	I

Plan of the talk

- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced

The Wireless Explosion

- *Mobile devices, linking the real and virtual worlds could change your perception of your surroundings.* (The Economist, may 2003)
- Mobile devices and sensor network have a potential of changing how we work and how we use personal technology
- Mobile devices and sensor network are in their infancy: in 1990 the roads were almost the same, 1990 GSM phones did not exist

Location Awareness

- Managing location information implies introducing **context awareness**, **time** and **identity**
- **Location and sensor services** are merging: from macro to micro Geography
- Location awareness has a vast range of benefits and threats. **Privacy and control** are the most glaring examples

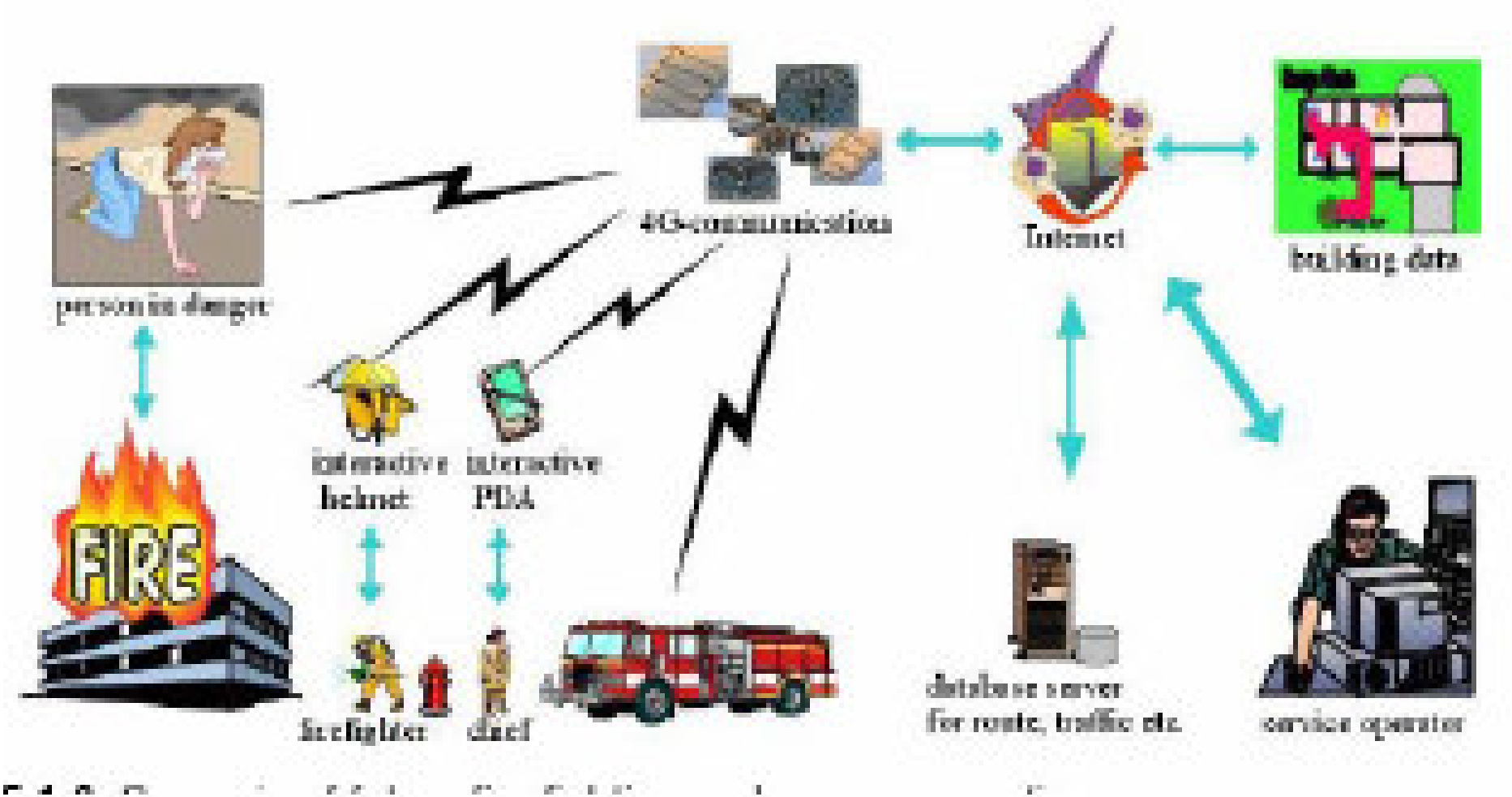
Context-aware demands: **where are you now?**

- Where is the 112 call coming from?
- I cannot find the device that need maintenance!
- Where is patient Brown?
- Is area 2B clear of staff?
- The convoy has deviated from the route!
- How do I alert people in this area?

Context-aware services

- Aimed at
  - delivering personalized, timely, location-aware information services to the mobile visitors
  - E.g. WebPark or Fire Alert System
- Depending on the **CURRENT** user position
- **ON LINE** services
- Privacy trivializes, it is more security and secrecy

FIRE ALERT SCENARIO



Changing the focus:
Movement awareness

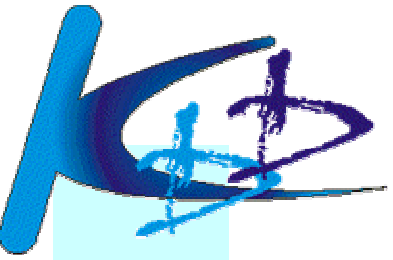
- Managing location information also gives the possibility to access space-time trajectories of the personal devices.
- Trajectories are the traces left behind by moving objects and individuals
- Trajectories contain detailed information on mobile behaviour and therefore offer opportunity to mine behavioral patterns

Movement-aware demand: **where people have been?**

- How people move around in the town
  - During the day, during the week, etc.
- Are there typical movement behaviours?
- How frequently people access the network?
- How are people movement habits changing in this area in last decade-year-month-day?
- ….

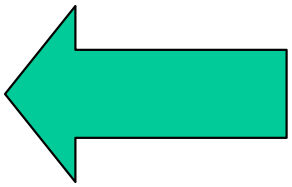
Movement aware services

- Aimed at modeling the **movement behaviours**
- Depending on the **traces** (the logs) left behind during the mobile activity
- Depending on the **HISTORY** of traces
- **OFF-LINE** services
- **Privacy** is a big issue

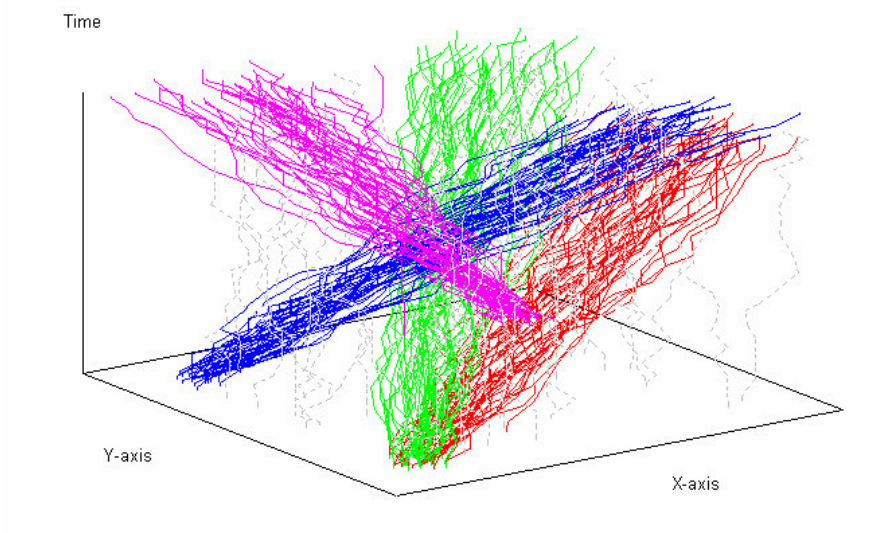


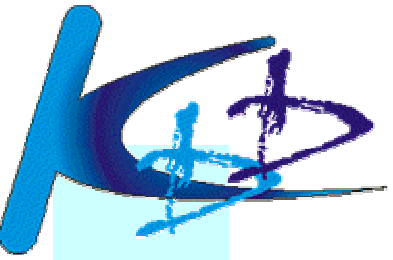
Plan of the talk

- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced



From movement data to movement patterns





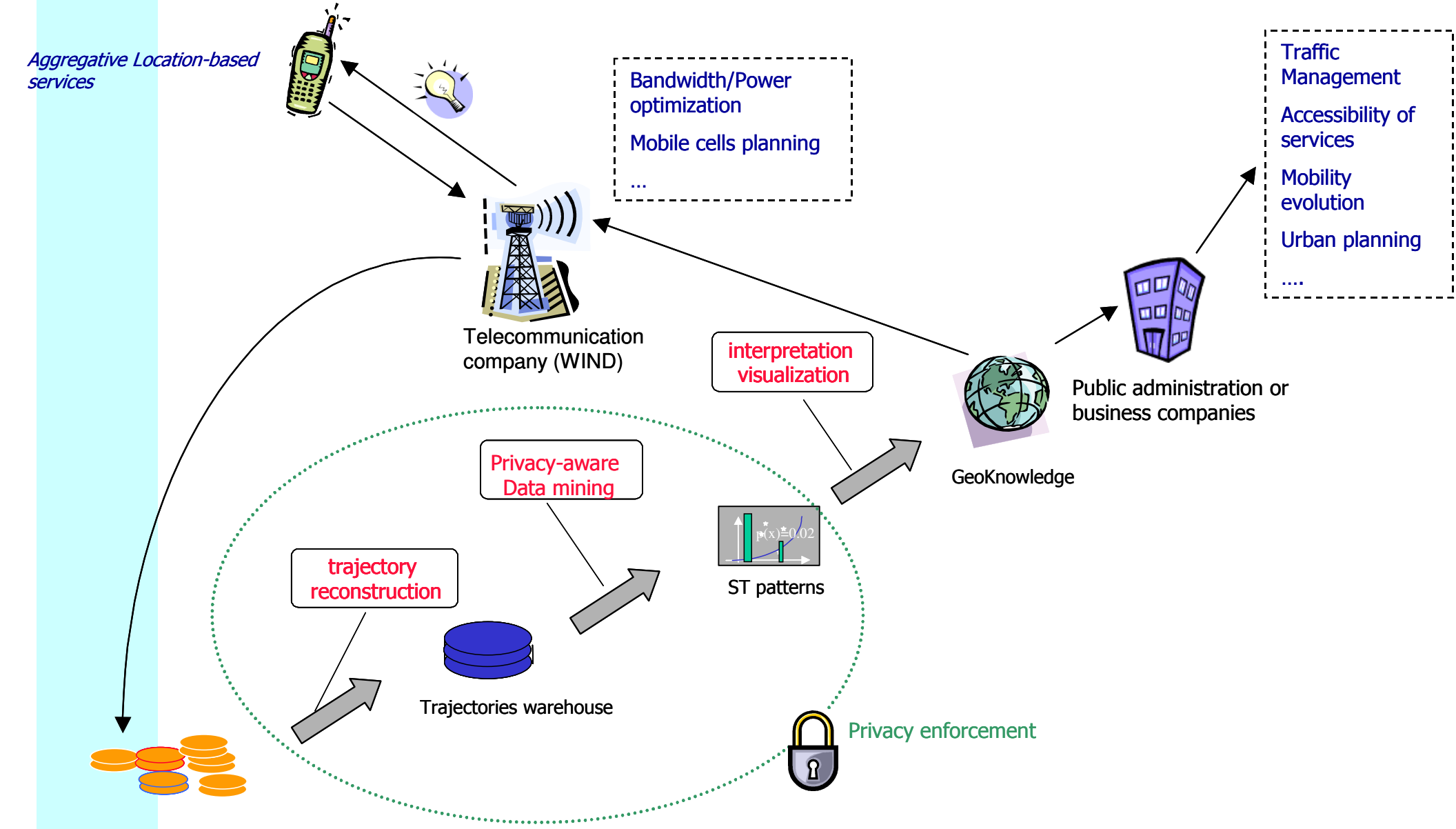
GeoPKDD applications

- enabled by *movement patterns*
  - extracted from positioning data
  - at the server level
  - in a safe, privacy-preserving way,
- delivered in the appropriate form to various end users

Exploitation scenarios

1. Towards the society: dynamic traffic monitoring and management for sustainable mobility, urban planning
2. Towards the network: network optimization, e.g. adaptive band allocation to cells,
3. Towards the individual: personalization of location-based services, car traffic reports, traffic information and predictions

Geographic privacy-aware Knowledge Discovery process



GeoPKDD – general project idea

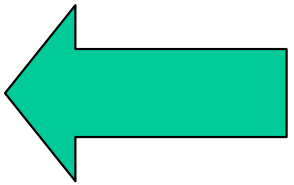
- extracting user-consumable forms of knowledge from large amounts of raw geographic data referenced in **space** and in **time**.
- knowledge discovery and analysis methods for **trajectories of moving objects**, which change their position in time, and possibly also their shape or other significant features
- devising **privacy-preserving** methods for data mining from sources that typically contain personal sensitive data.

GeoPKDD – specific goals

- models for moving objects, and data warehouse methods to store their trajectories,
- knowledge discovery and analysis methods for moving objects and trajectories,
- techniques to make such methods privacy-preserving,
- techniques for reasoning on spatio-temporal knowledge and on background knowledge
- techniques for delivering the extracted knowledge within the geographic framework

Plan of the talk

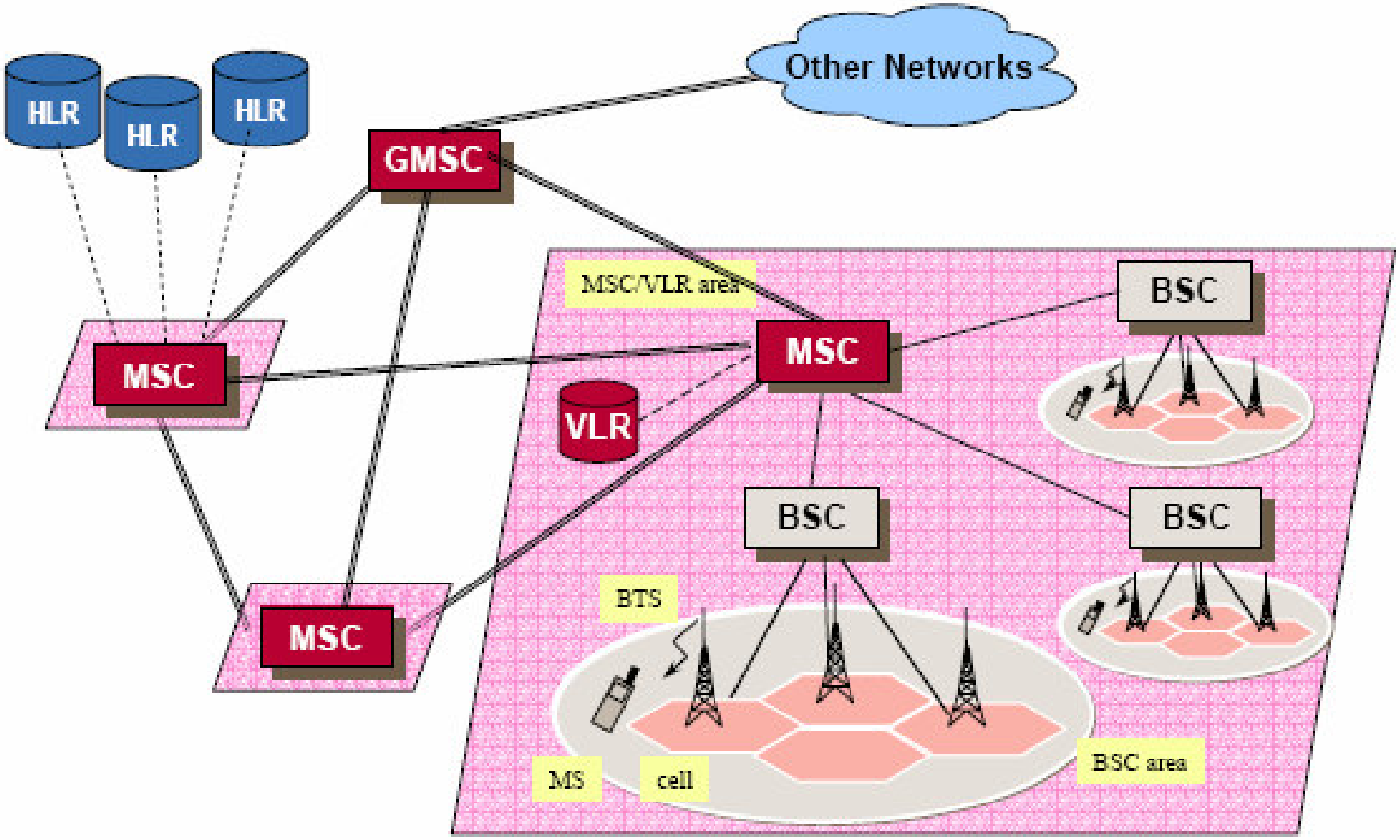
- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced

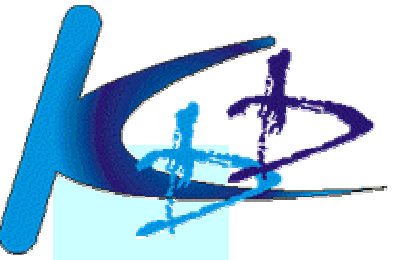


From traces to trajectories: the source data

- Streams of log data of mobile phones, i.e. sampling a trajectory by means of a set of of localization points (e.g., cells in the GSM/UMTS network).
  - Entering the cell –
    - e.g. (UserID, time, IDcell, in)
  - Exiting the cell –
    - e.g. (UserID, time, IDcell, out)
  - Movements inside the cell?
    - Eg (UserID, time, X,Y, Idcell

GSM network


Other Networks
HLR
HLR
HLR
GMSC
MSC/VLR area
MSC
MSC
VLR
BSC
BSC
BSC
MSC
BTS
MS
cell
BSC area



From trajectories to logs and backwards

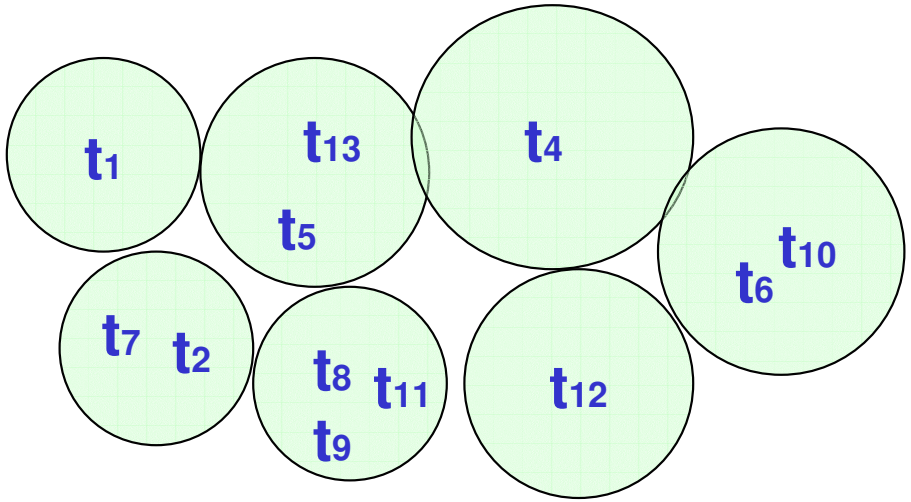
- Real trajectories are continuous functions
- Logs are discrete *sampling* of real trajectories, dependent on the wireless network technology
  - unregular granularity in time and space
  - possible imperfection/imprecision
- An approximated reconstruction of the real trajector from its log traces is needed

Reconstructing trajectories

Scene 1

In the log entries we have no ID

➔ Log entries become time-stamped events



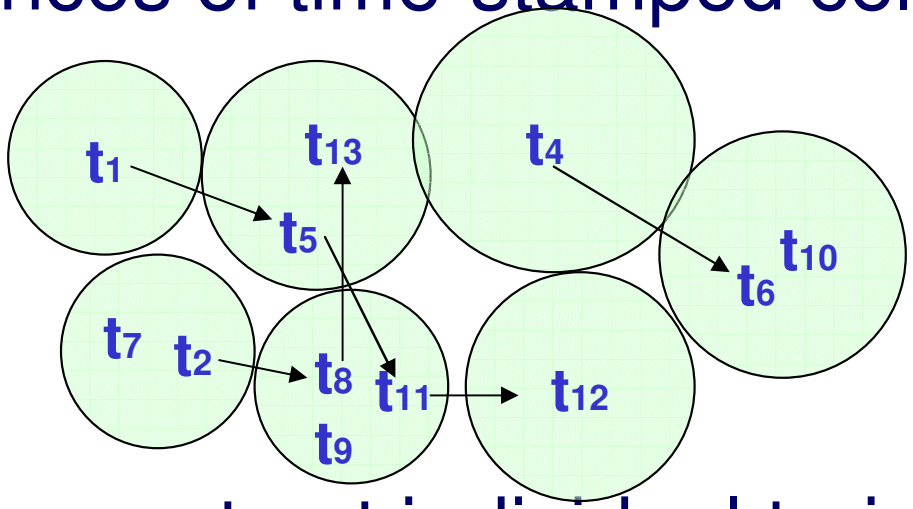
- We can only compute aggregated info on *traffic flow*, but not reconstruct individual trajectories

Reconstructing trajectories

Scene 2

In the log entries we have (encrypted) IDs

➔ Log entries can be grouped by ID to obtain sequences of time-stamped cells

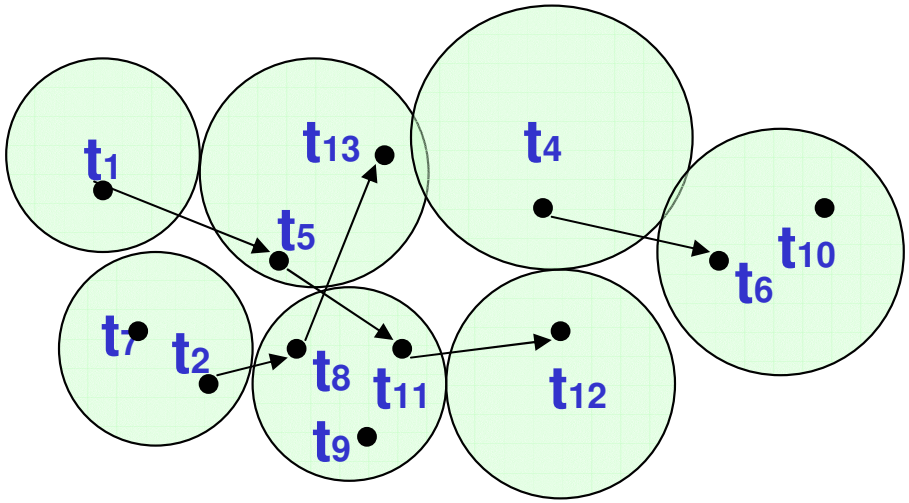


- We can reconstruct individual trajectories, with the *spatial granularity* of a cell:
  - positions of t_5 and t_8 can be distinguished, but not t_5 and t_{13}

Reconstructing trajectories

Scene 3

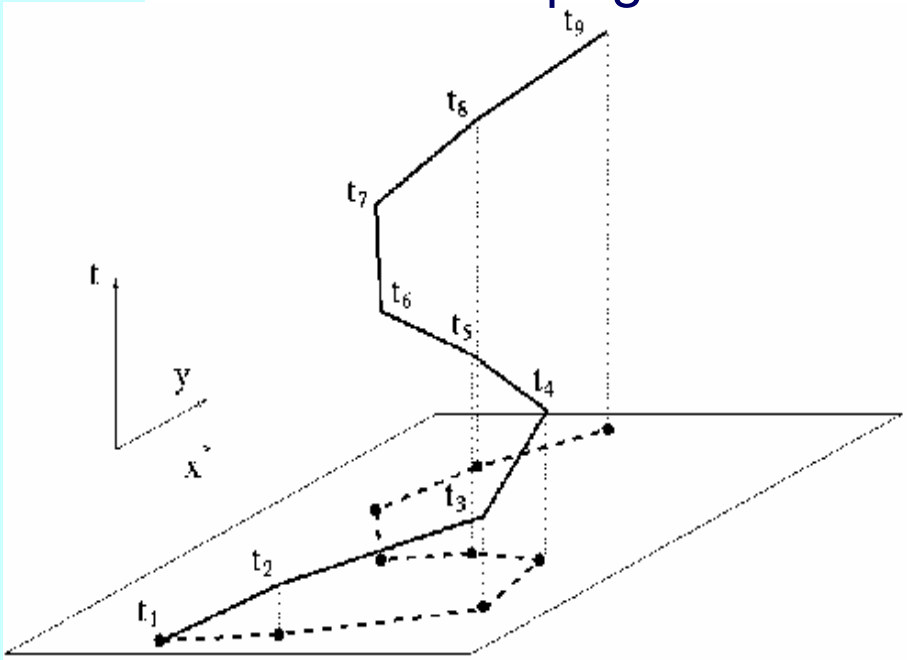
In the log entries we IDs and (approximated) position in the cell

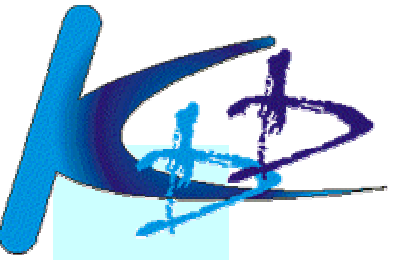


- We can reconstruct individual trajectories, with a finer *spatial granularity*: now, positions of t_5 and t_{13} can be distinguished.

Trajectory data models

- Discrete data model
  - Trajectory is represented as a set of time-stamped coordinates
  - $T=(t_1,x_1,y_1), \ldots, (t_n, x_n, y_n)$ => position at time t_i was (x_i,y_i)
- Continuous data model
  - Trajectory is represented as a function of space and time
  - Parametric-spaghetti: linear interpolation of consecutive points



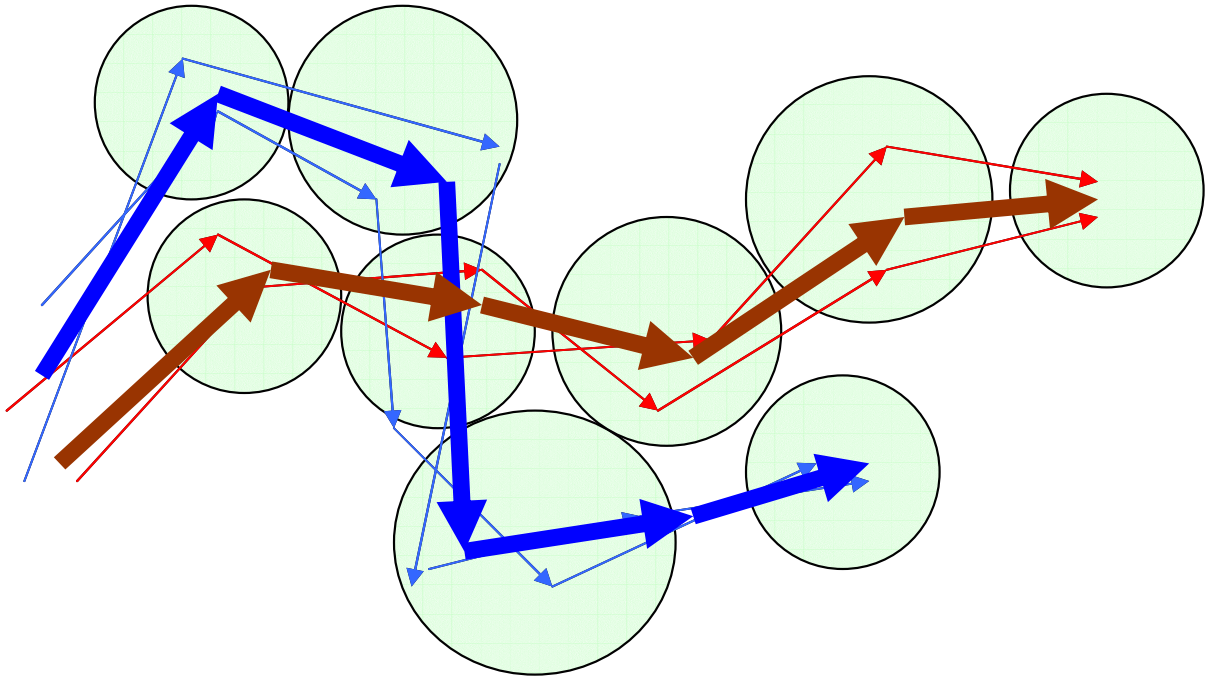


Plan of the talk

- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced

Movement patterns: Clustering

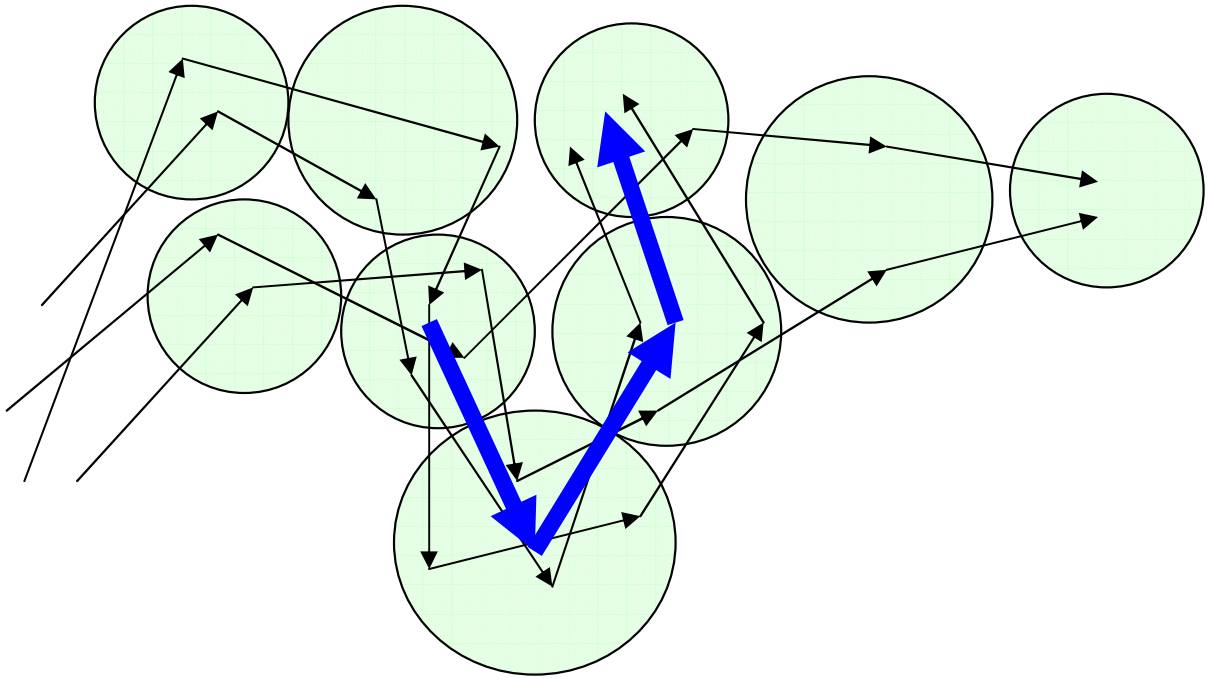
- Group together similar trajectories
- For each group produce a summary



= cell

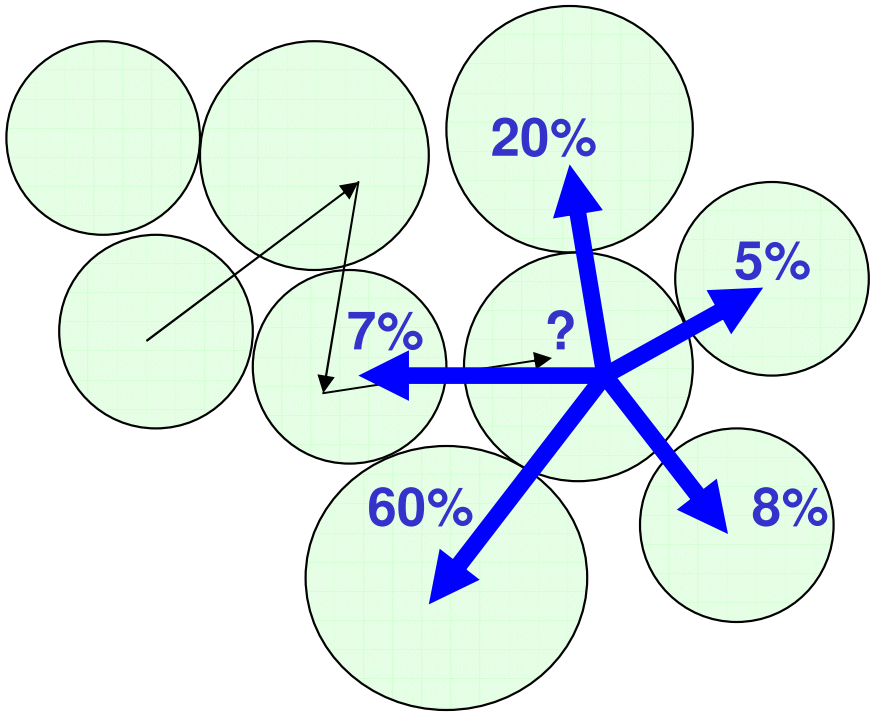
Movement patterns: Frequent patterns

- Discover frequently followed (sub)paths



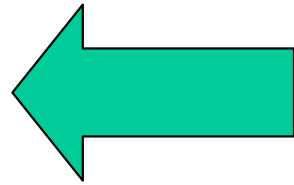
Movement patterns: classification models

- Extract behaviour rules from history
- Use them to predict behaviour of future users



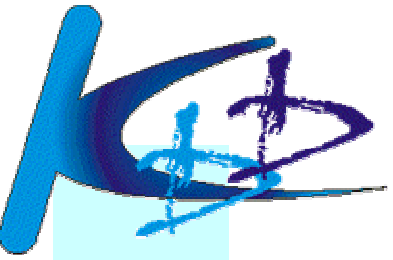
Plan of the talk

- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced



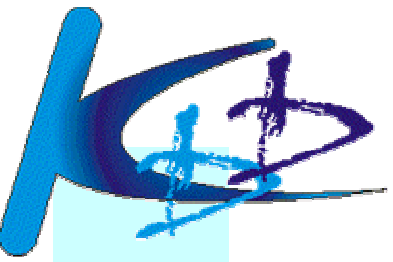
Why emphasis on privacy?

- More, better data are gathered, more vulnerability from correlation
- On the other hand, more and new data bring new opportunities
  - Public utility, new markets/paradigms, new services
- Need to maintain privacy without giving up opportunities
- Need to obtain social acceptance through demonstrably trustworthy solutions



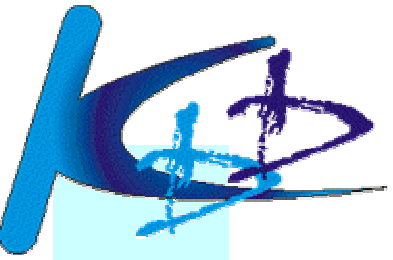
Privacy in GeoPKDD

- to develop trustable data mining technology,
- capable of using logs to produce ***provably*** privacy-preserving patterns,
- which may be safely distributed
  - Patterns, not data!



Privacy in GeoPKDD

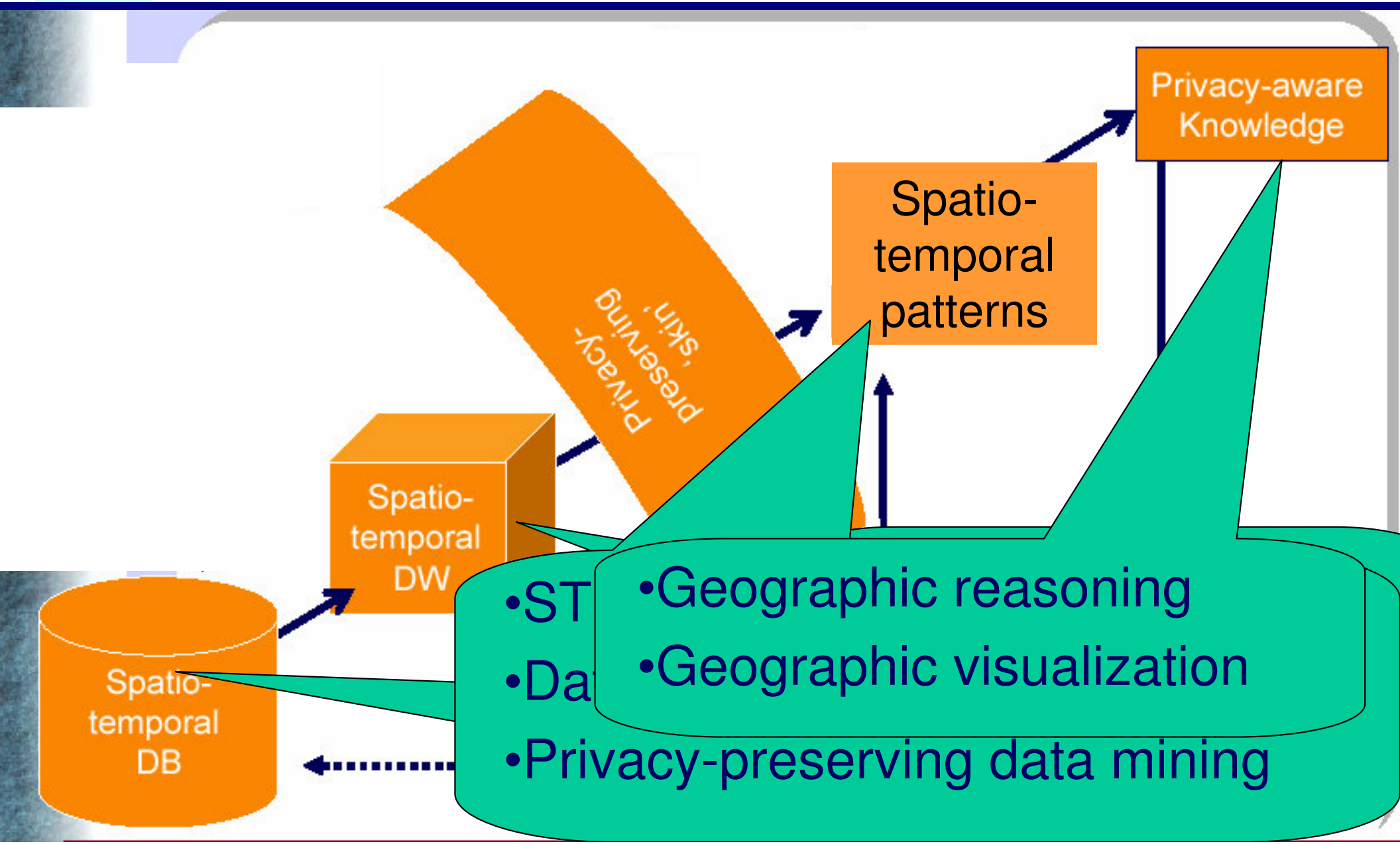
- ... is a technical issue, besides ethical, social and legal, in the specific context of ST data
- How to formalize privacy constraints over ST data and ST patterns?
  - E.g., anonimity threshold on clusters of individual trajectories
- How to design DM algorithms that, by construction, only yield patterns that meet the privacy constraints?
- How to perform multidimensional analysis of ST data the meet the the privacy constraints?



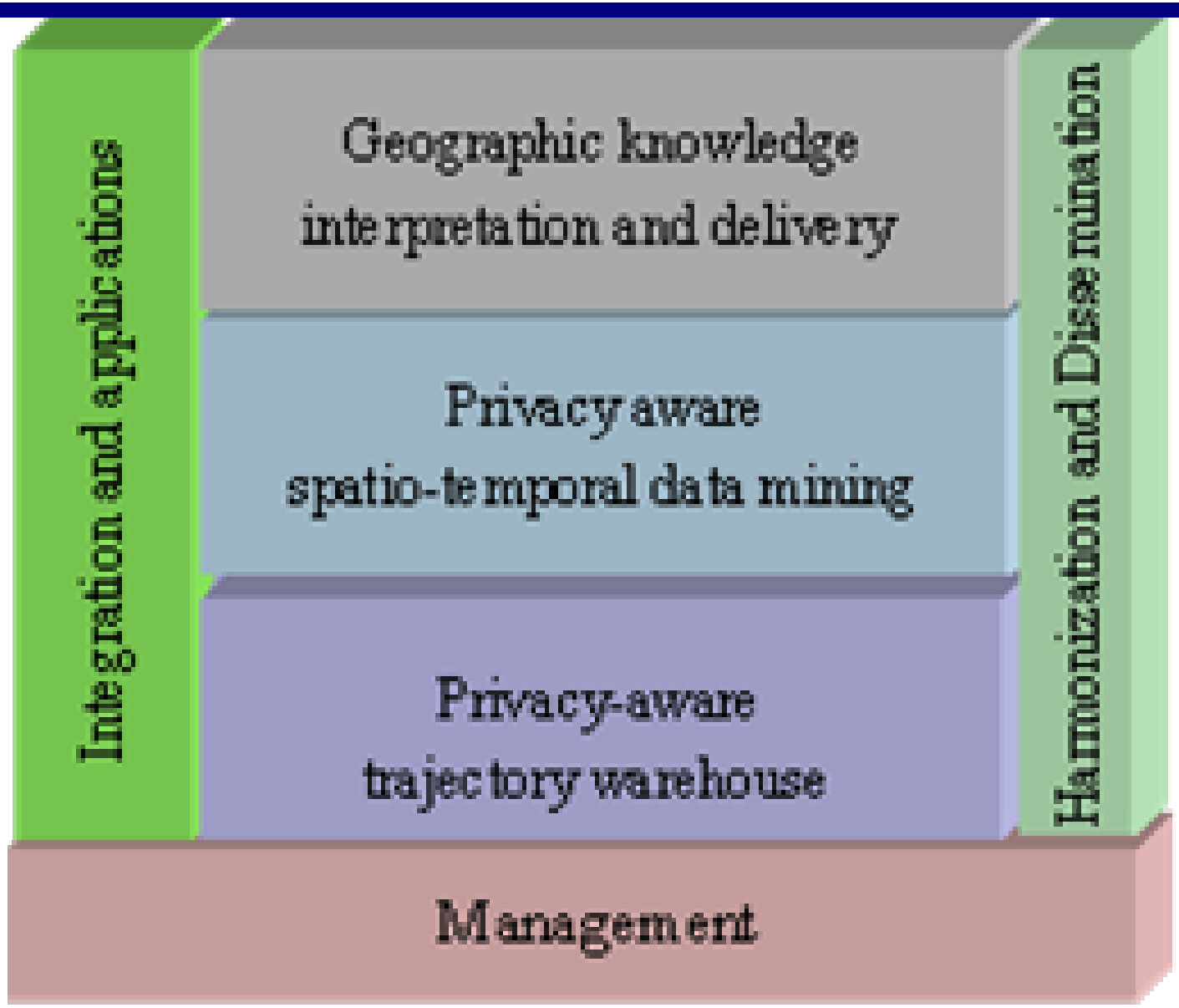
Plan of the talk

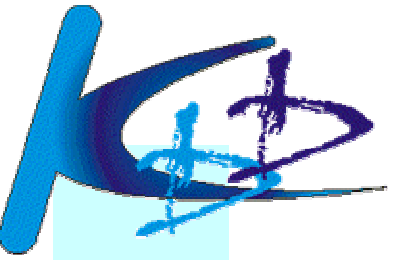
- The wireless explosion:
  - Location- vs Movement-aware services
- GeoPKDD vision and goals
- The source data:
  - From logs to trajectories
- The movement patterns
  - Spatio-temporal models of mobility behaviour
- The privacy challenge
- The building blocks:
  - Methods and technologies to be invented/enhanced
- First investigations at Pisa KDD Lab

GeoPKDD research issues



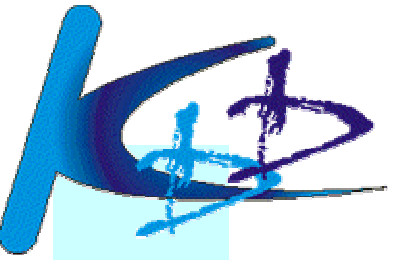
GeoPKDD workpackages





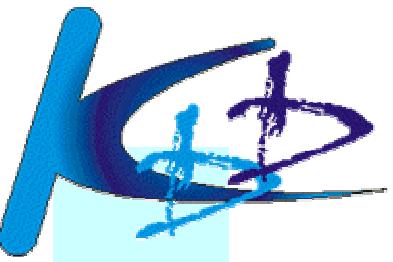
GeoPKDD basic workpackages

- (WP1) Privacy-aware trajectory warehouse
- (WP2) Privacy-aware spatio-temporal data mining methods
- (WP3) Geographic knowledge interpretation and delivery
- (WP4) Harmonization, integration and applications



Privacy-aware trajectory warehouse

- Tasks:
  1. a **trajectory model** able to represent moving objects, and to support multiple representations, multiple granularities both in space and in time, and uncertainty;
  2. a **trajectory data warehouse** and associated OLAP mechanisms, able to deal with multi-dimensional trajectory data;



Privacy-aware spatio-temporal data mining

- Task: algorithms for **spatio-temporal data mining**, specifically meant to extract spatio-temporal **patterns** from trajectories of moving objects, equipped with:
  1. methods for provably and measurably protecting **privacy** in the extracted patterns;
  2. mechanisms to express constraints and queries into a **data mining query language**, in which the data mining tasks can be formulated

Geographic knowledge interpretation and delivery

- Task: interpretation of the extracted spatio-temporal patterns, by means of **ST reasoning** mechanisms
- Issues
  - **uncertainty**
  - georeferenced visualization methods for trajectories and spatio-temporal patterns

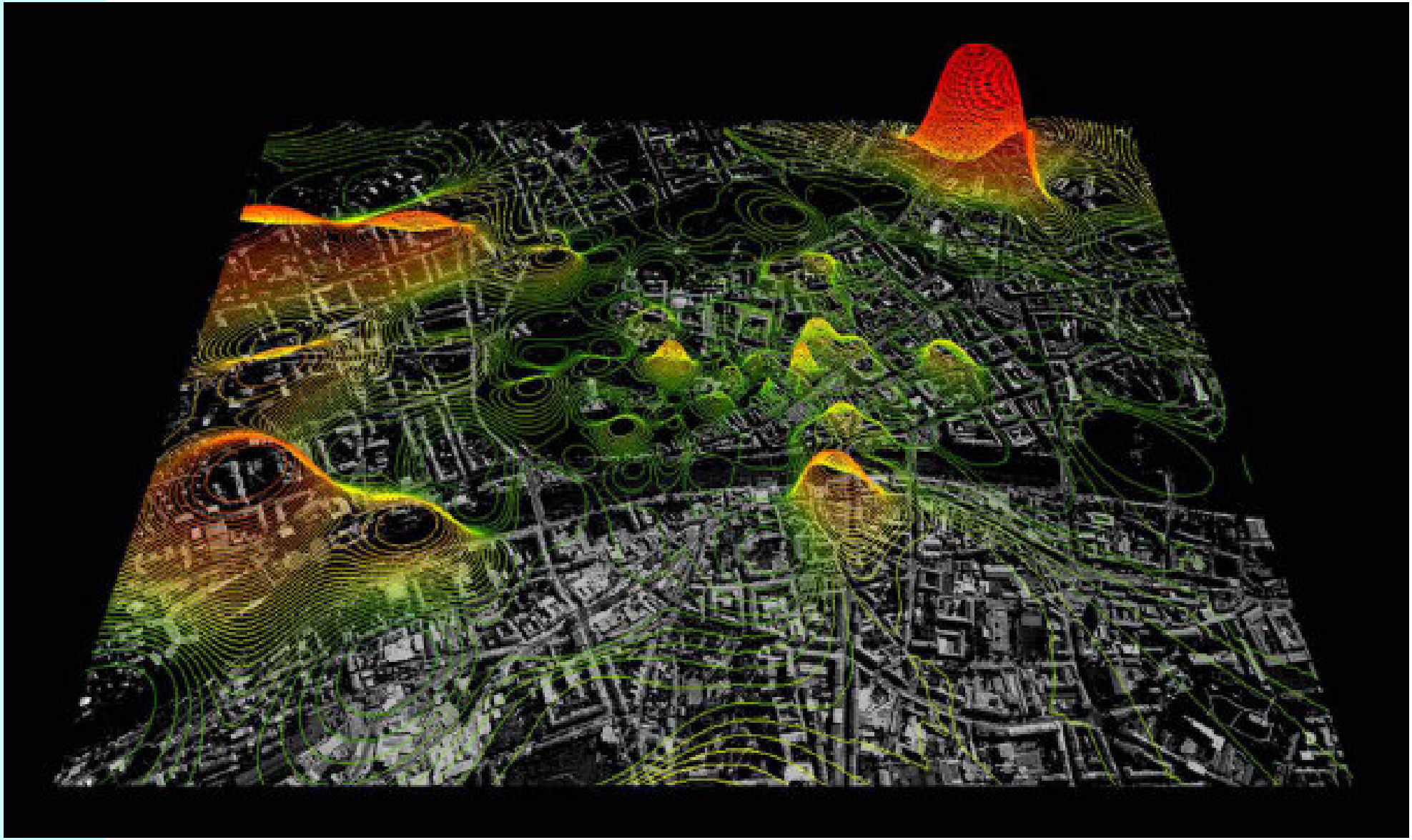
Harmonization, Integration and Applications

- Tasks:
  - Harmonization with national privacy regulations and authorities – **privacy observatory**
  - **Integration** of the achieved results into a coherent framework to support the GeoPKDD process
  - **Demonstrators** for some selected applications: for public authorities, network operators and/or marketing operators, e.g., in sustainable mobility, network optimization, geomarketing.

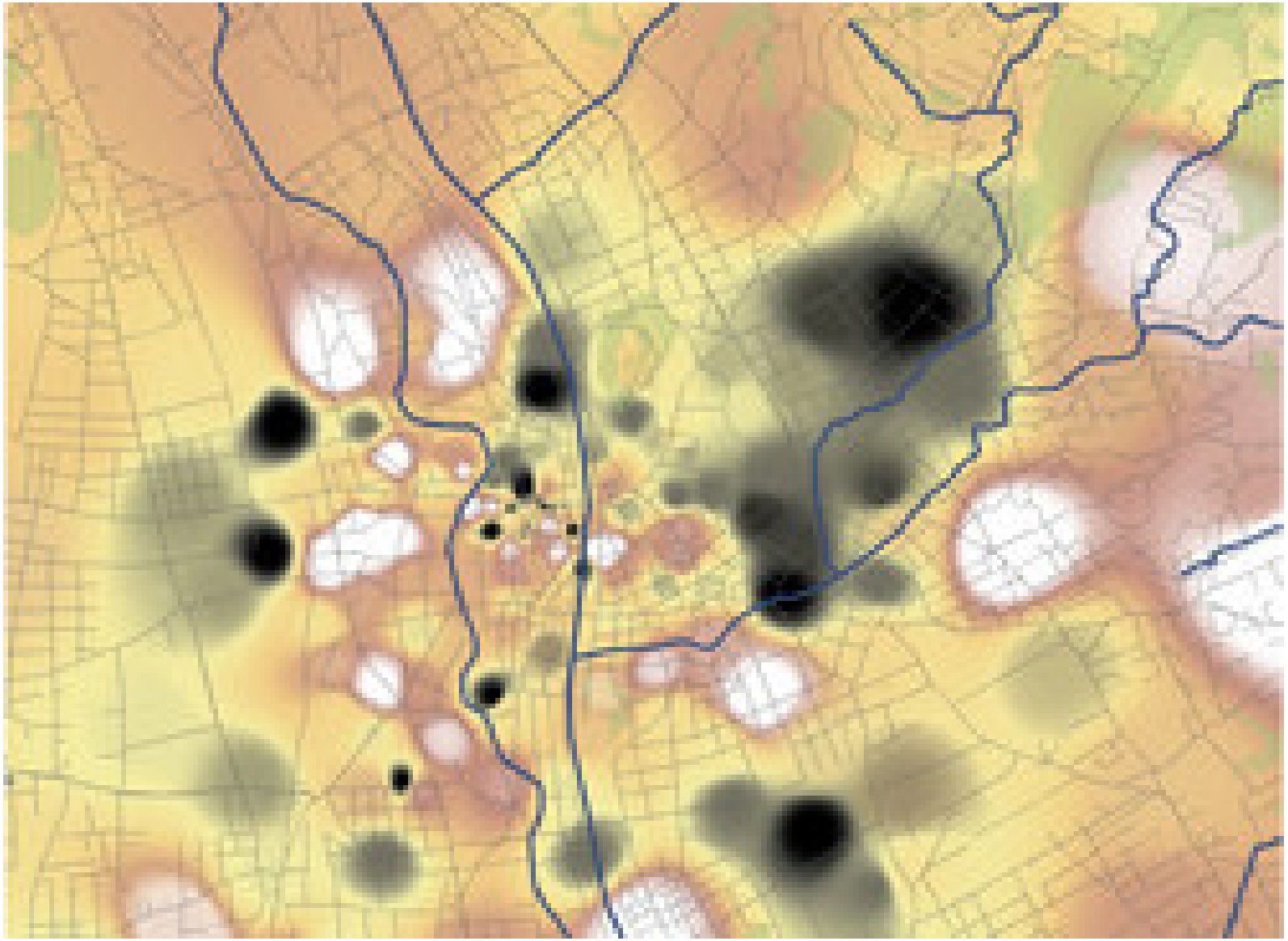
Summarizing….

- GeoPKDD is
  - –Strong pull from emerging applications
  - –Strong push for fundamental research
  - –Scientifically exciting
  - –Timely….

The senseable project: http://senseable.mit.edu/grazrealtime/



cell-phone traffic intensity in real-time

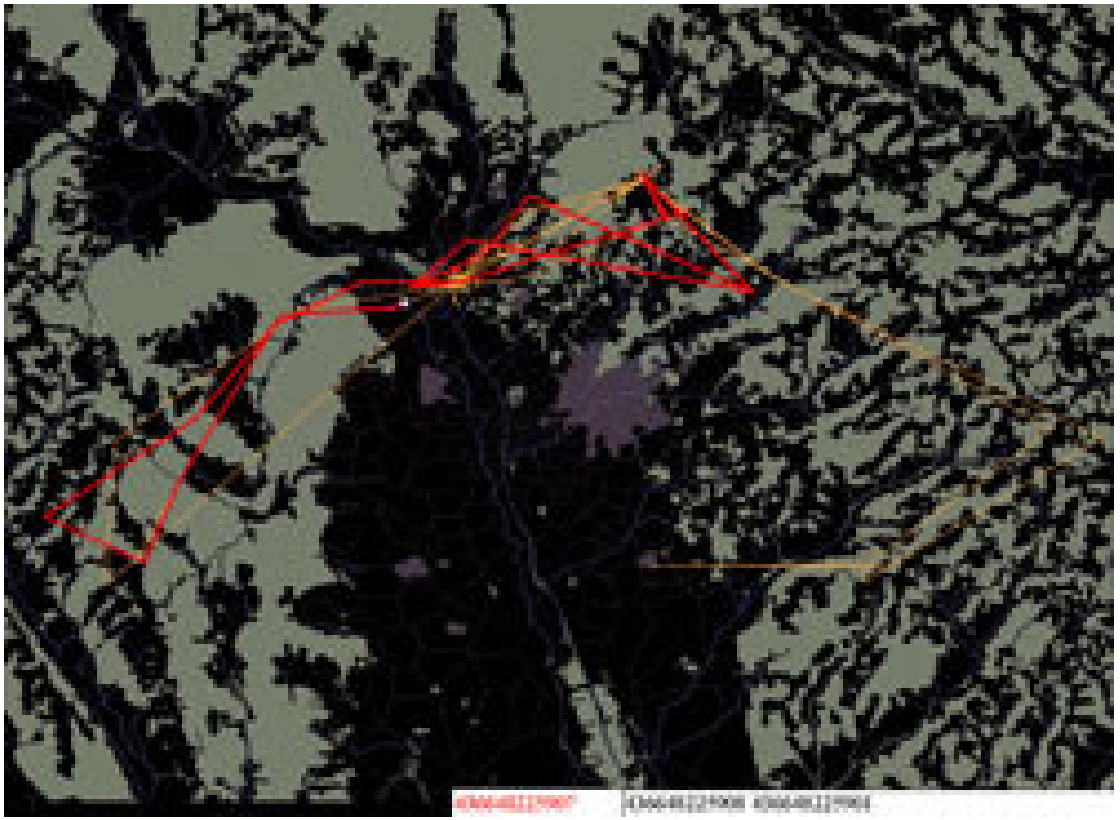


Call handovers between cells

- This map computes origins and destinations of cell-phone calls passing through the city of Graz.

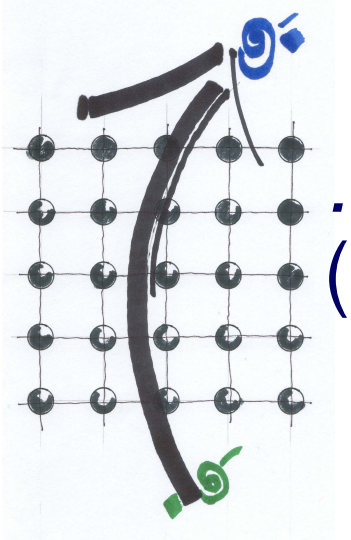
Traces: cell phone tracking in GRAZ

- The orange lines show the physical location of the M-Graz exhibition visitors who voluntarily registered and allowed their cell phones to be tracked as they move through the city.
- The red lines retrace individual paths of movement, indicating the person's code number at the bottom of the page



Minard's depiction of Napoleon's 1812 March on Moscow

… defied the pen of the historian in its brutal eloquence (Marey, 1887)



… is the best statistical graphic ever drawn (Tufte, 1983)

The Terrible Fate of Napoleon's Grand Army in Russia: 1812

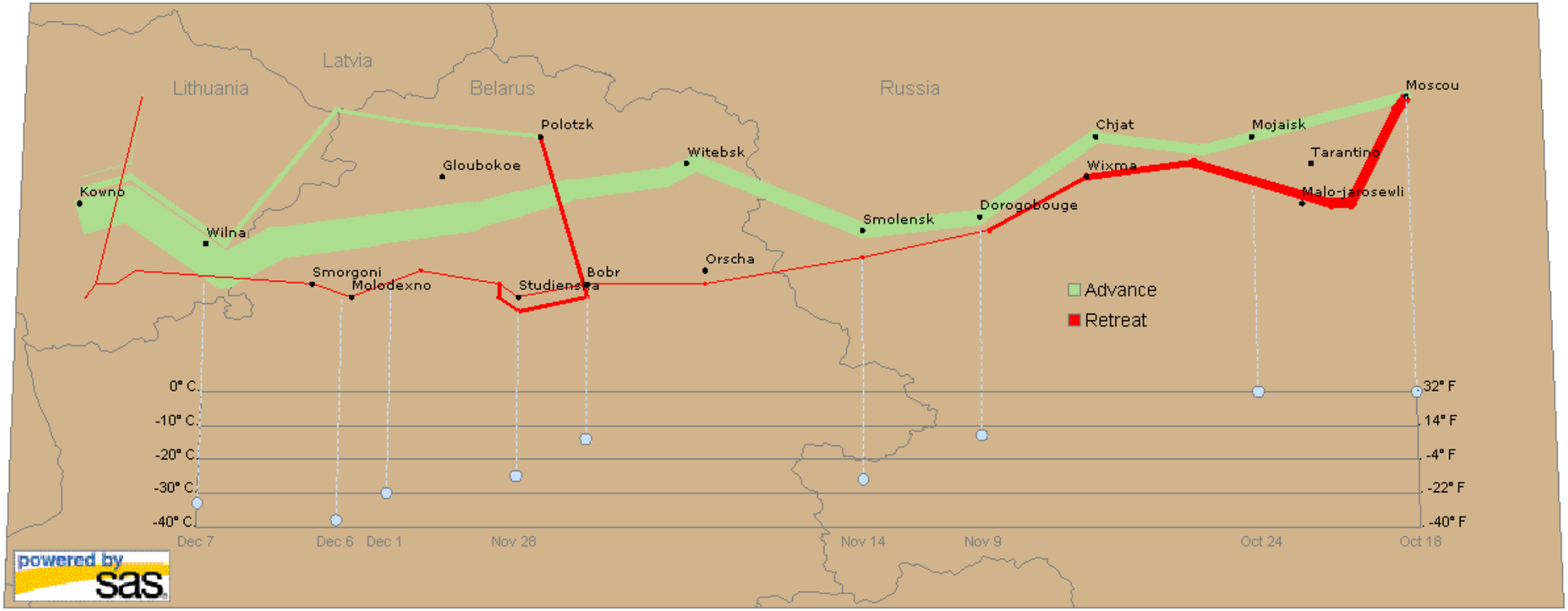
Adapted from the Statistical Graph of Charles Minard (1861),
by Aaron Walburg and Stephen Hartzog (1996)

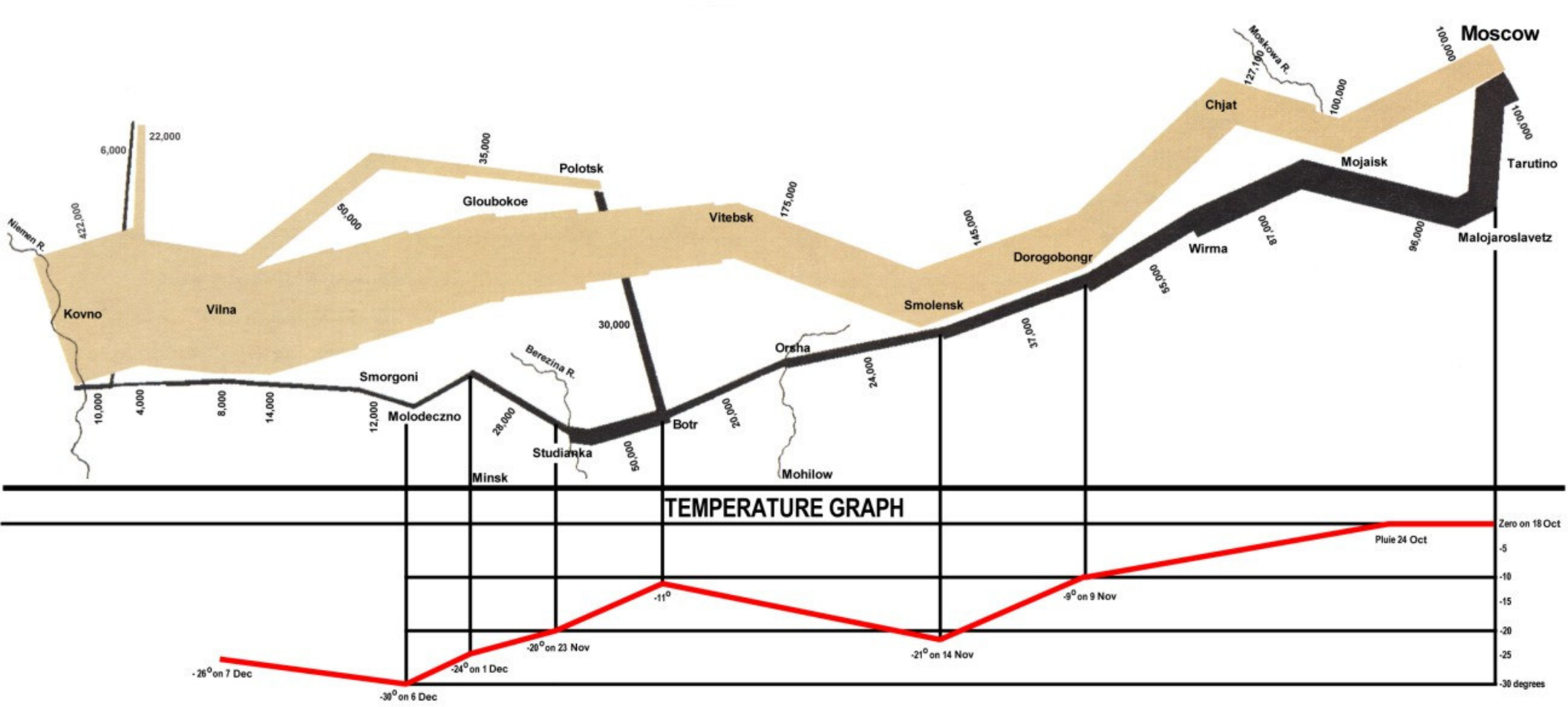
Napoleon's Russian Campaign, 1812

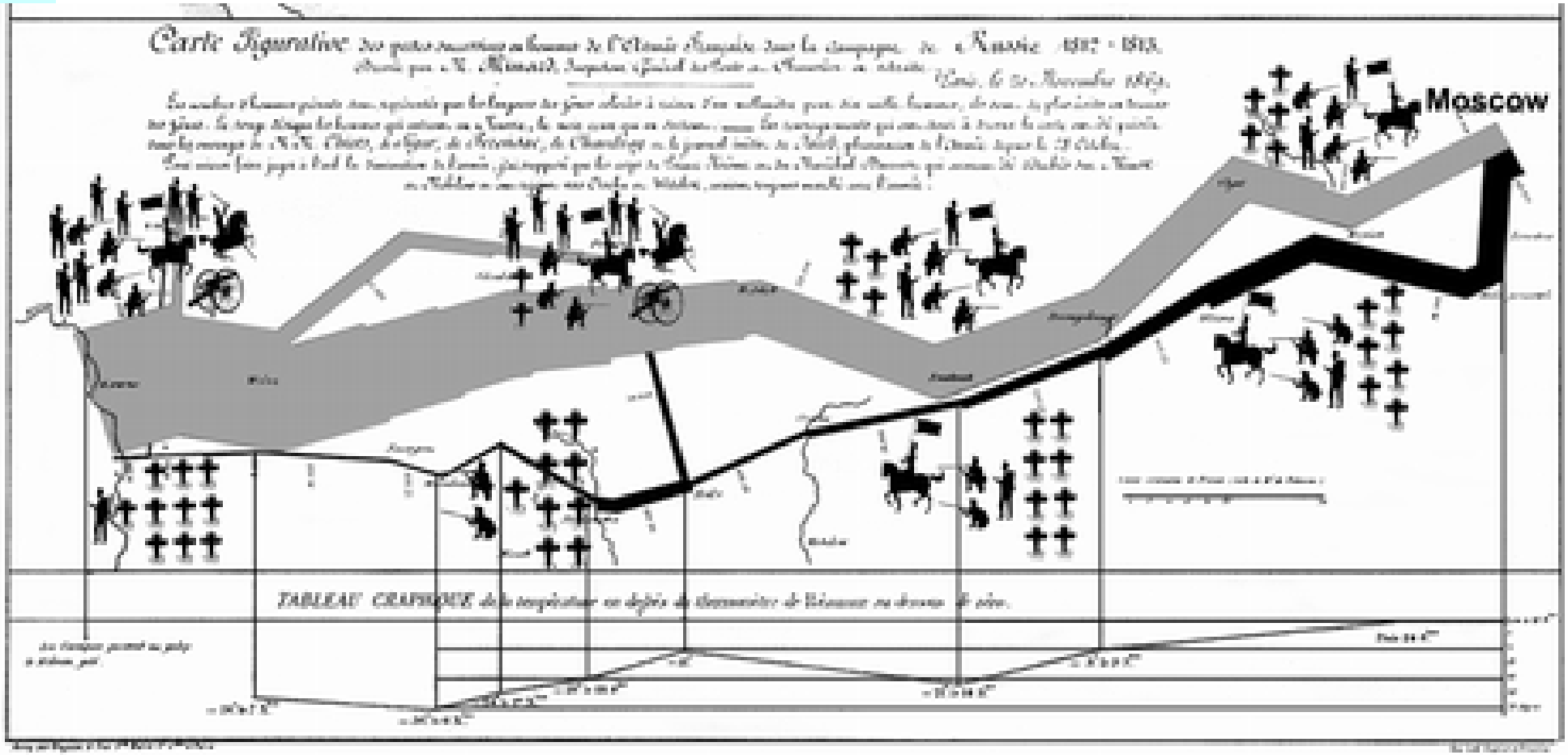
Plotted on modern map

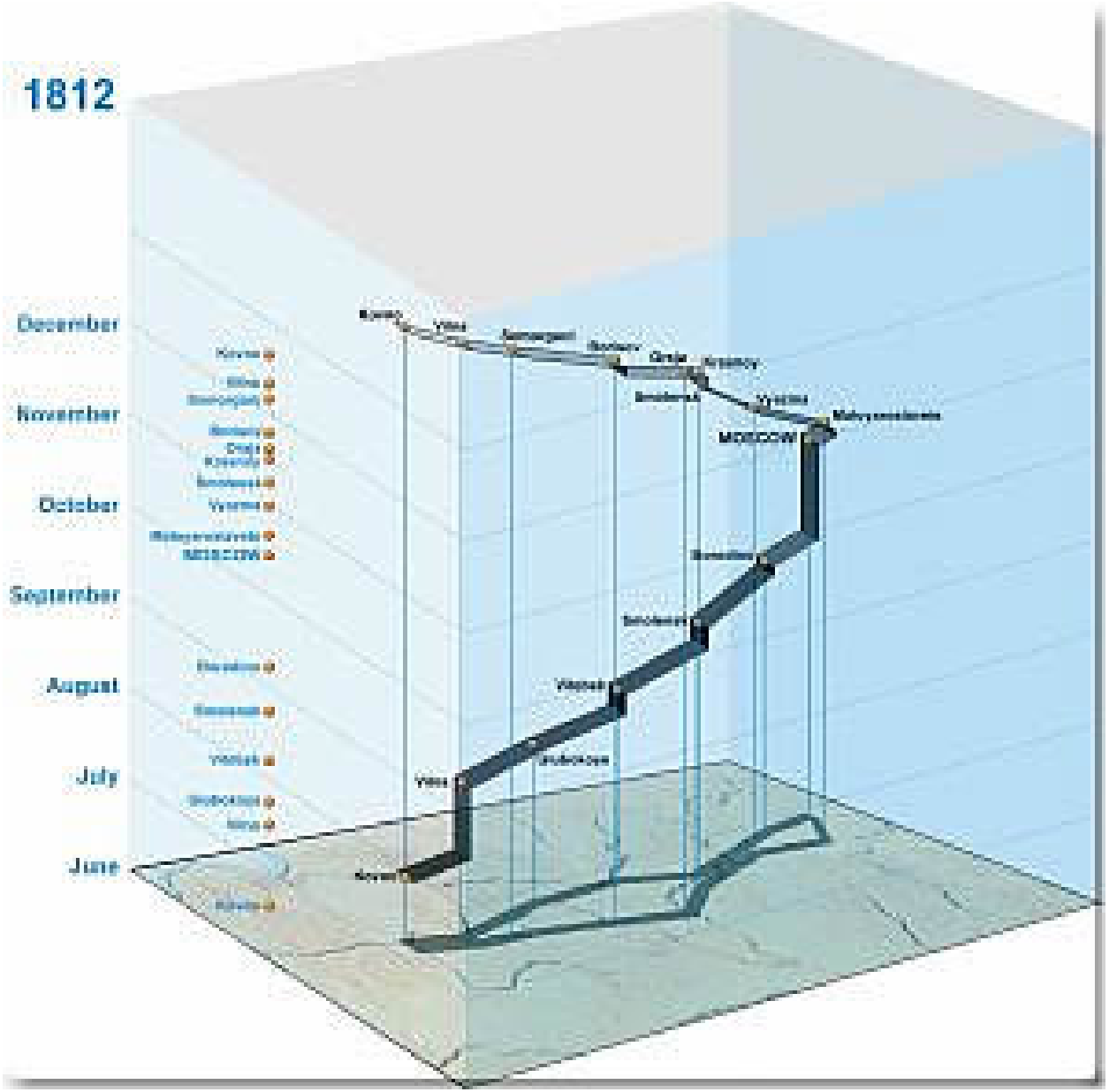
Napoleon's Russian Campaign, 1812

Plotted on modern map

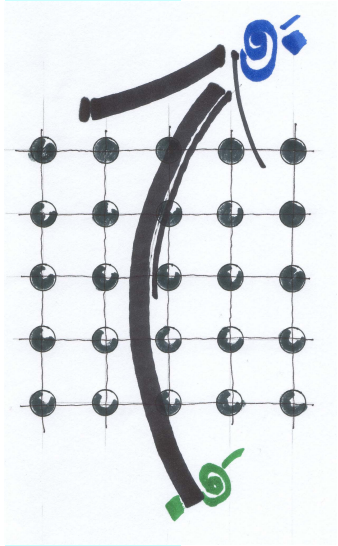



Moscow
Moskowa R.
127,000
100,000
100,000
100,000
Chjat
Mojaisk
Tarutino
Malojaroslavetz
96,000
87,000
Wirma
Dorogobongr
65,000
Smolensk
37,000
145,000
175,000
Vitebsk
Gloubokoe
Polotsk
35,000
50,000
22,000
6,000
423,000
Niemen R.
Kovno
Vilna
30,000
Orsha
24,000
20,000
Botr
50,000
Studianka
28,000
Berezina R.
Smorgoni
Molodeczno
12,000
14,000
8,000
4,000
10,000
Minsk
Mohilow
TEMPERATURE GRAPH
Zero on 18 Oct
Pluie 24 Oct
-5
-10
-15
-20
-25
-30 degrees
-9° on 9 Nov
-11°
-21° on 14 Nov
-20° on 23 Nov
-24° on 1 Dec
-26° on 7 Dec
-30° on 6 Dec


Moscow


1812
December
November
October
September
August
July
June
MOSCOW

Goals of this mid-term meeting



Goals of mid-term meeting

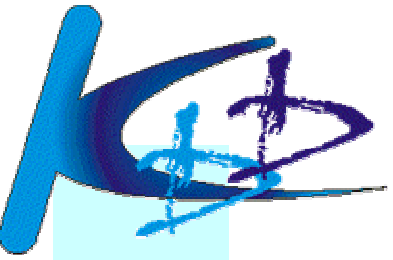
- What have we done in the first year?
  - Finalization plan of alignment reports
- What do we plan do to in the second year? Any refocusing needed? Coordination with European GeoPKDD.
  - Plan of activity for the second year – each partner should consolidate its plan w.r.t. to WP goals and deliverables
  - Tuesday morning 12:00 thru 13:00 the three units present briefly their (remodulated) aims and goals
  - Afternoon: concrete planning – who does what?

Pisa: objectives

- spatial and spatio-temporal privacy-preserving data mining, with particular focus on
  - clustering,
  - constraint-based frequent pattern mining
  - spatial classification;
- spatio-temporal logical formalisms to reason on extracted patterns and background knowledge.

Venezia (+ Milano): objectives

- trajectory model and privacy-preserving data warehouse, within a streamed and distributed context
- methods to mine sequential and non sequential frequent patterns from trajectories, within a streamed and distributed context
- postprocessing and interpretation of the extracted spatio-temporal patterns



Cosenza: objectives

- Trajectory mining
  - Clustering
- Privacy-preserving data mining
  - Probabilistic approach
- Distributed data mining

Deliverables of Phase 1 (months 1-5)

- WP1: Privacy-aware trajectory warehouse
  - [TR1.1] Alignment report and preliminary specification of requirements.
- WP2: Privacy-aware spatio-temporal data mining
  - [TR1.2] Alignment report on ST data mining techniques.
  - [TR1.3] Alignment report on privacy-preserving data mining techniques.
  - [TR1.4] Alignment report on distributed data mining.
- WP3: Geographic knowledge interpretation and delivery
  - [TR1.5] Alignment report on ST reasoning techniques.
- WP4: Harmonization, Integration and Applications
  - [TR1.6] Report on characterization of GeoPKDD applications and preliminary feasibility study.
  - [A1.7] Implantation of the Privacy Regulation Observatory.

Deliverables of Phase 2 (months 6-17)

- WP1: Privacy-aware trajectory warehouse
  - [TR2.1] TR on design of the trajectory warehouse.
  - [P2.2] Prototype of the trajectory warehouse.
- WP2: Privacy-aware spatio-temporal data mining
  - [TR2.3] TR on new techniques for ST and trajectory Data Mining.
  - [TR2.4] TR on new privacy-preserving ST Data Mining.
  - [TR2.5] TR on distributed data mining
  - [P2.6] Prototype(s) of privacy-aware ST data mining methods.
- WP3: Geographic knowledge interpretation and delivery
  - [TR2.7] TR on ST reasoning techniques and DMQL for geographic knowledge interpretation and delivery.
  - [P2.8] Prototype(s) of the ST reasoning formalism and DMQL
- WP4: Harmonization, Integration and Applications
  - [TR2.9] Requirements of the application demonstrator(s).

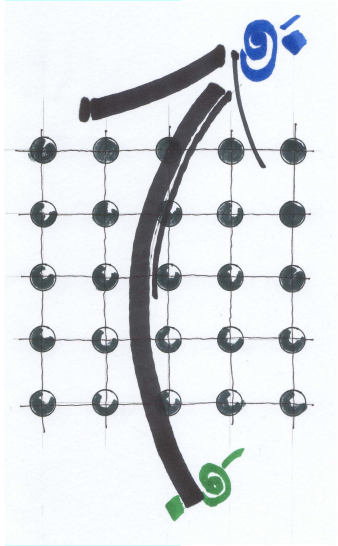
Deliverables of Phase 3 (months 18-24)

- WP4: Harmonization, Integration and Applications
  - [TR3.1] TR on the design of a system prototype allowing the application of privacy-preserving data mining tools to spatio-temporal and trajectory data.
  - [P3.2] Prototype implementing the system described in the technical report [TR3.1].
  - [P3.3] Prototype extending the system prototype [P3.2] to work on a distributed system.
  - [TR3.4] TR on the description of the prototypes developed and the results of the experimentation.
  - [TR3.5] Final report on harmonisation actions and mutual impact between privacy regulations and project results.

Transversal activities

- experiments,
- application demonstrators,
- harmonization with privacy regulations and authorities,
- dissemination of results.

Decision making and action plan



Alignment reports

- Drafts on-line on the project web site by November 20
- Prospective consolidation and review process for integration with ARs of European GeoPKDD and publication in a dedicated book (Springer?)
- Send material (including slides of this meeting) to Salvo Rinzivillo at rinziv@di.unipi.it

Research in year 2

- Goals of units confirmed? YES
- Which is the specificity of Italian level GeoPKDD?
  - Streaming
  - Distributed
  - Neighboring application fields: logistics, workflows, RF IDs and sensor networks
- Task force on Spatio-temporal DMQL
  - Franco, Fosca, Alessandra, Giuseppe, Chiara, Miriam
- Next meeting: 9-10 February 2006
  - Ischia, giovedi 9 pomeriggio, venerdi 10 tutto il giorno
  - Tutorial su aspetti specifici: (circa 1,5 h ciascuno, incluso discussione)
    - Fosca, Dino: privacy;
    - Carlo: RFIDs;
    - Sergio: logistica;
    - VE: stream DM;
    - Malerba: multi-relational ST DM
    - Task force DMQL: DMQL
  - Goal: Definizione proposta PRIN 2006

Research in year 2

- European GeoPKDD:
  - Kick-off a Pisa, 1-2-3 dicembre 2006, verrà diffusa agenda appena disponibile
  - Un rappresentante di GeoPKDD.it presenta il progetto italiano al kick-off
- Acronimo del progetto italiano: GeoPKDD.it
- What on the web site?
  - Sezione riservata
  - Sezione sui prodotti (deliverables)
  - Formato standard dei TR (latex e word)
  - Ispirarsi al sito del progetto PRIN D2I: http://www.dis.uniroma1.it/~lembo/D2I/
- Which relation between the European and Italian web sites: mutual reference