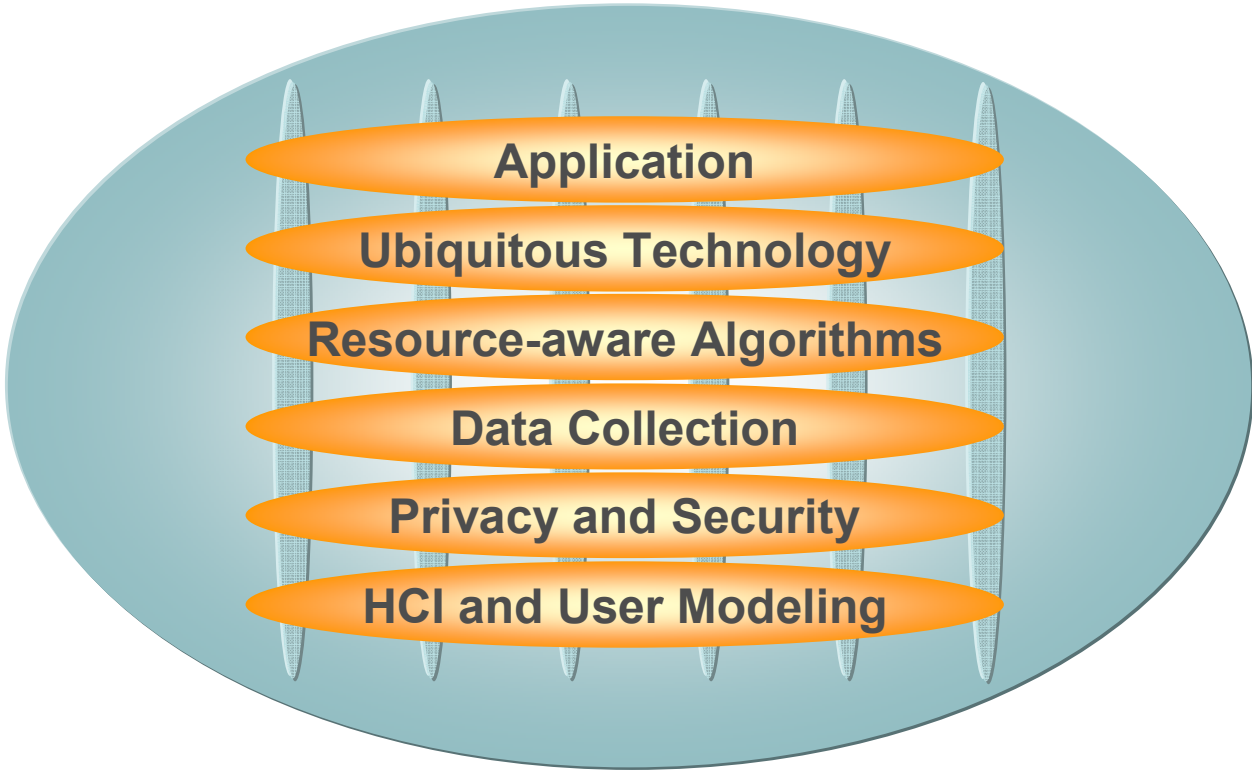
Research Challenges in Ubiquitous Knowledge Discovery

KDubiq network

Michael May
Fraunhofer Institute for
Intelligent Analysis and
Information Systems

NGDM'07, 12.10.07
Baltimore

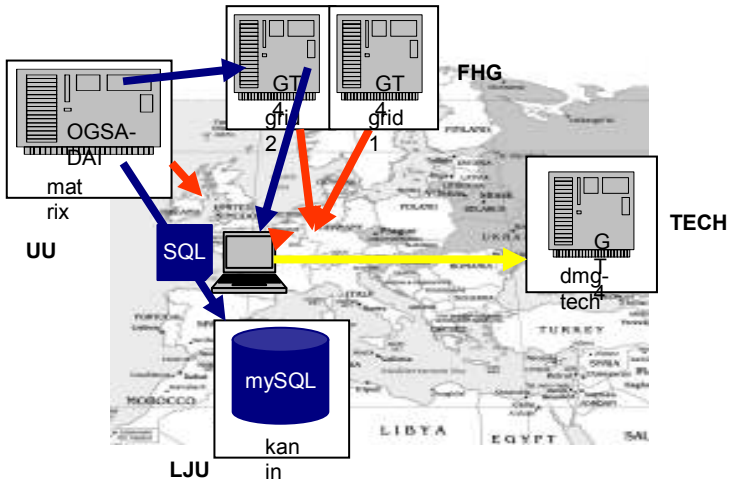


Next Generation Data Mining –

a collection of hot topics or an integrated perspective?

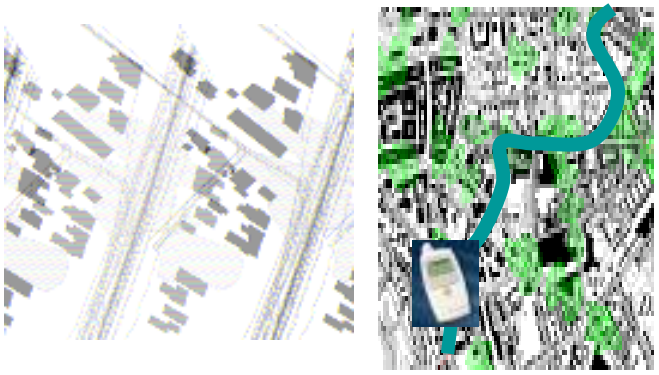
How can we get people to collaborate with each other that work in different fields?

Some Challenges …



Grid-based Data Mining & Data Mining Based Grid Monitoring (Technion, Fraunhofer, Daimler)

Data Stream Mining (Univ. Porto)

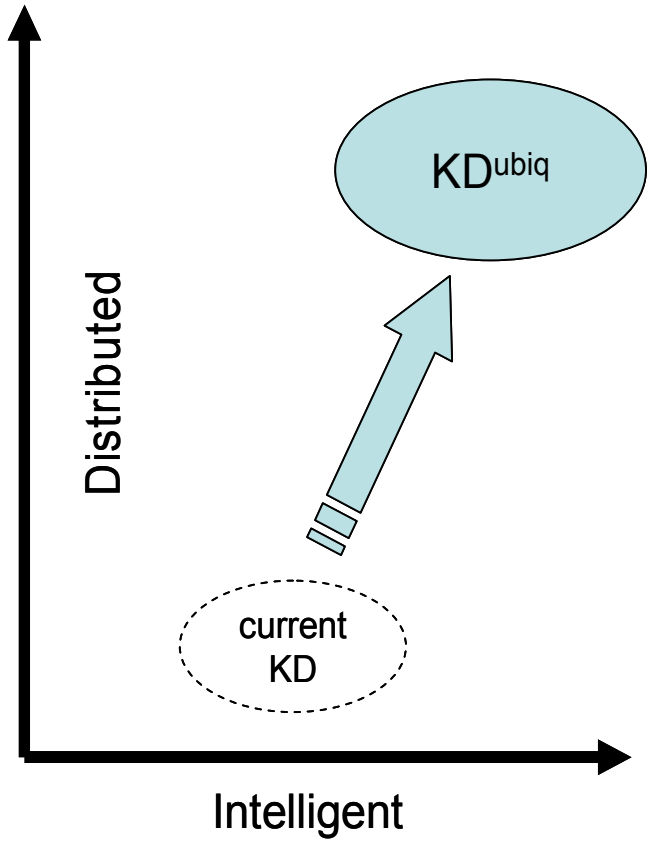


Mobility Mining from GPS-Tracks (Fraunhofer, Univ. Pisa, Univ. Sabanci)

P2P Music Mining

(Univ. Dortmund)

Ubiquitous Knowledge Discovery



.

Knowledge discovery process in mobile, distributed, dynamic environments, in presence of massive amounts of data

= **Ubiquitous Knowledge Discovery**

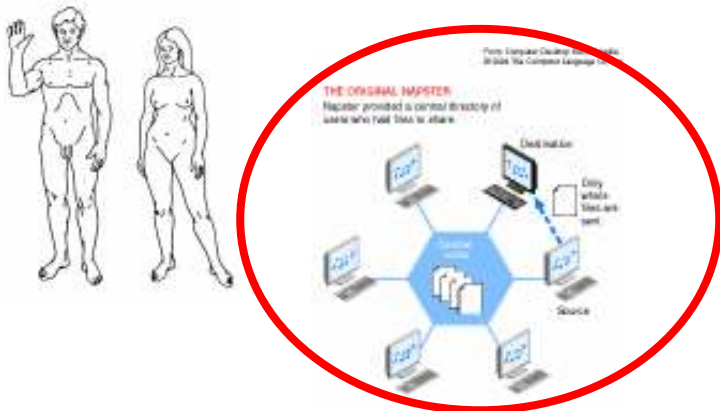
Key characteristics

1. **Time and space**. The objects of analysis exist in time and space. Often they are able to move.
2. **Dynamic environment**. These objects might not be stable over the life-time of an application. Instead they might appear or disappear.They exist in a dynamic and unstable environment, evolving incrementally over time.
3. **Information processing capability**. The objects are endowed with information processing capabilities
4. **Locality**. The objects never see the global picture - they know only their local spatio-temporal environment.
5. **Real-Time**. They often have to take decisions or act upon their environment - analysis and inference has to be done in real-time. The models have to evolve incrementally in correspondence with the evolving environment.
6. **Distributed**. In many cases the object will be able to exchange information with other objects, thus forming a truly distributed environment

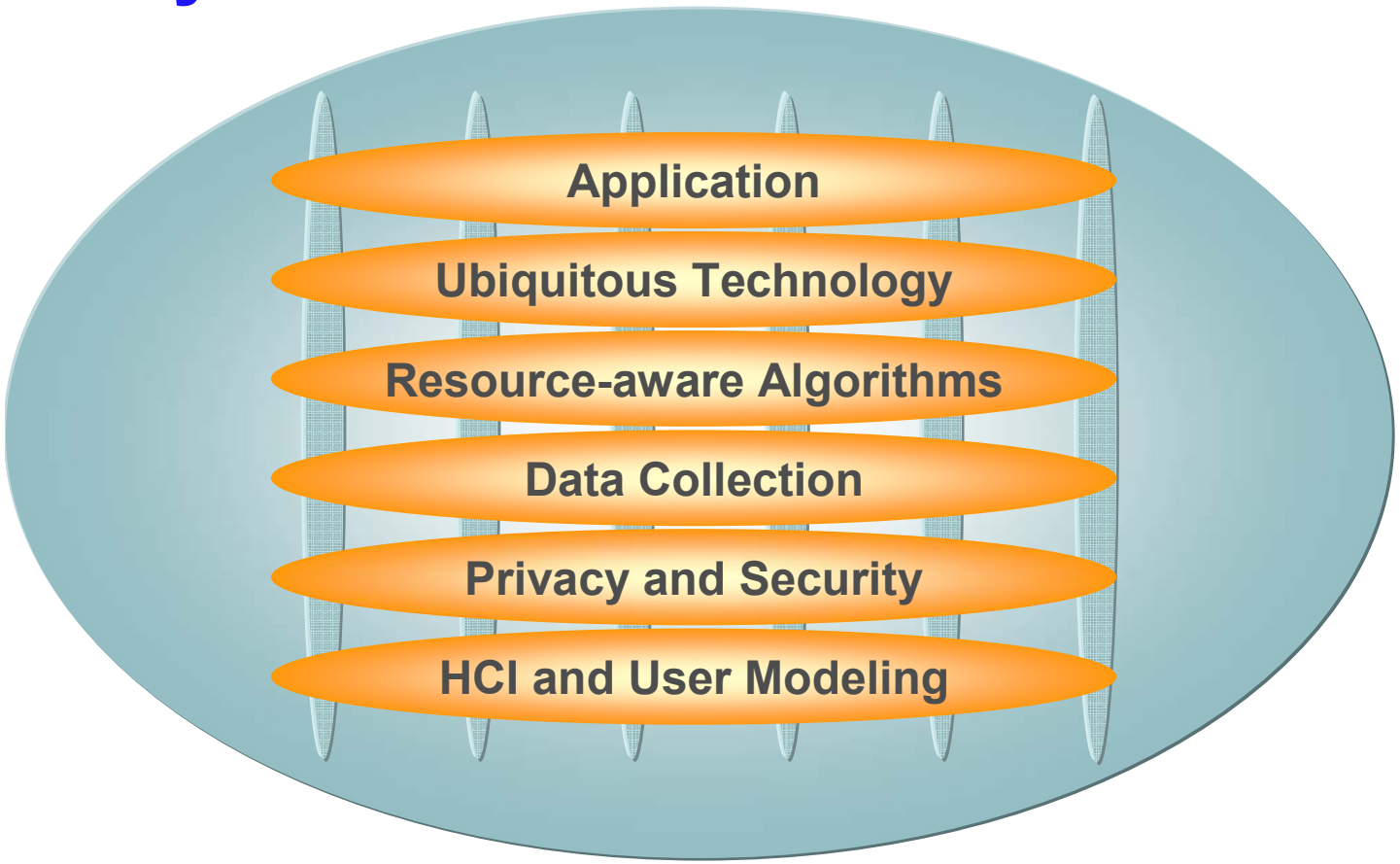
Objects of Study

- Systems that have these properties are humans, animals, and increasingly, computing devices
- KDUbiq investigates artificial systems
  - The machine learning or data mining is not applied to data *about* the system,
  - it is rather *part* of the information processing capabilities of the system

- *This is a large departure from the current mainstream in machine learning and data mining!*



Dimensions of Ubiquitous Knowledge Discovery Systems



When designing a ubiquitous knowledge discovery system, major design decisions in each of these six dimensions have to be taken.

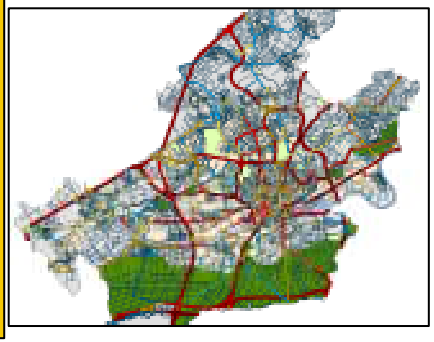
Choices are mutually constraining each other

The best way to understand a system is to build one!

KDUbiq thus adopts a *systems* view on how to build next generation knowledge discovery systems

The path to ubiquity

?



Ubiquitous

Knowledge Discovery

Activity Mining inside Mobile Devices

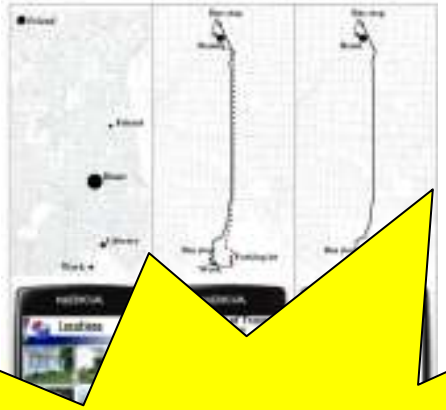
GPS Track Mining

Traffic Prediction using video measurements

KD ubiq

Some more challenges …

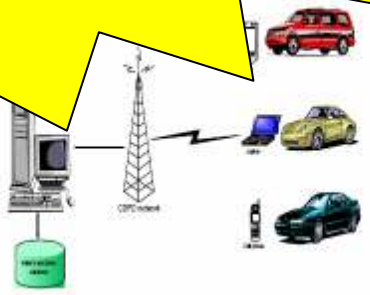
Probabilistic Robotics - Self Driving Car



…hine Learning –
Collaborative P2P Mining

Machine Learning makes the difference!

Intelligent Agents – Smart Homes

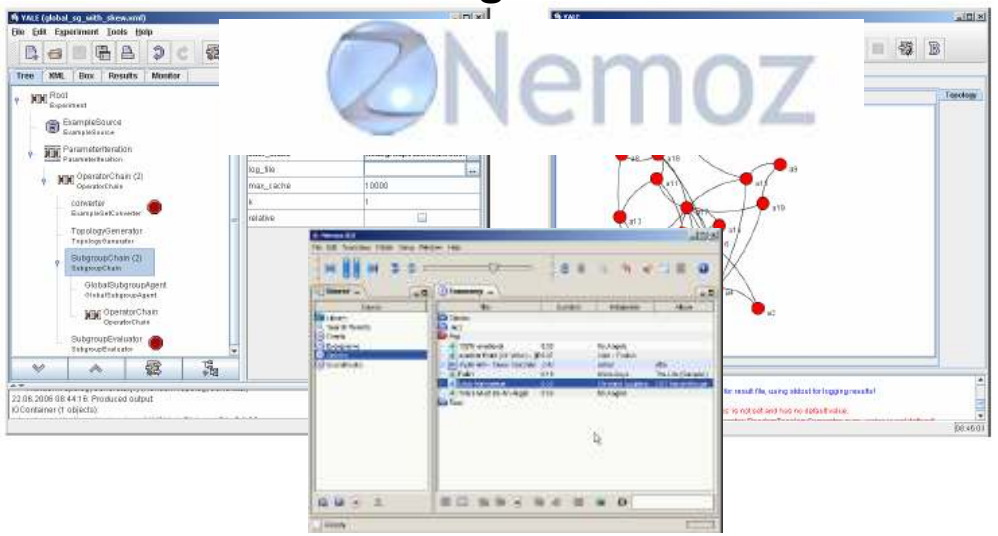


Data Mining - Vehicle Monitoring

A self-driving car

What do these two systems have in common?

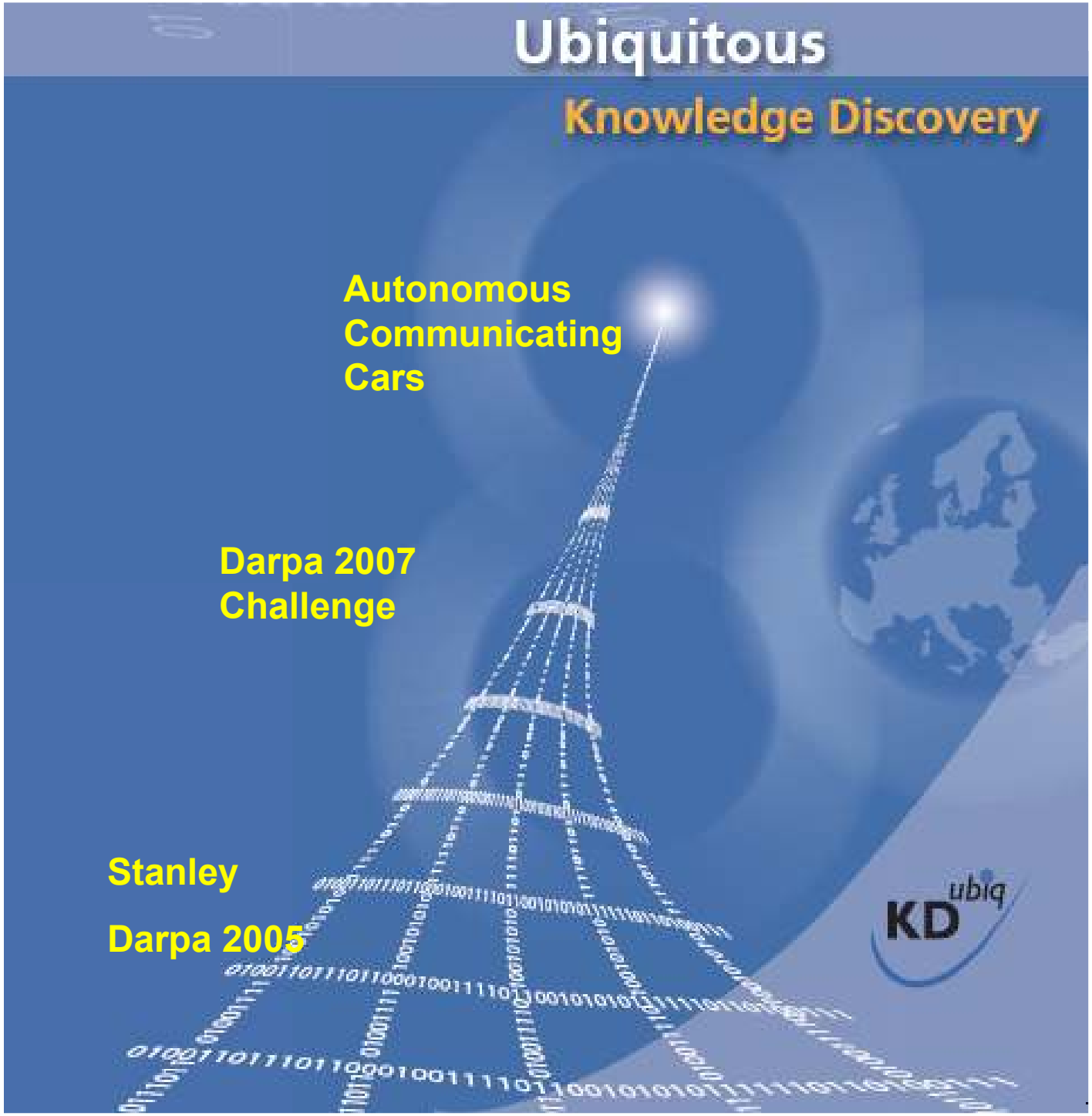
A Web 2.0 application for distributed, collaborative clustering of music collections



The path to ubiquity

? Communicating cars, respecting privacy

? Operating under city traffic conditions



What are the biggest research challenges?

- Coming up with algorithms and a theoretical framework for problems that violate the iid assumption

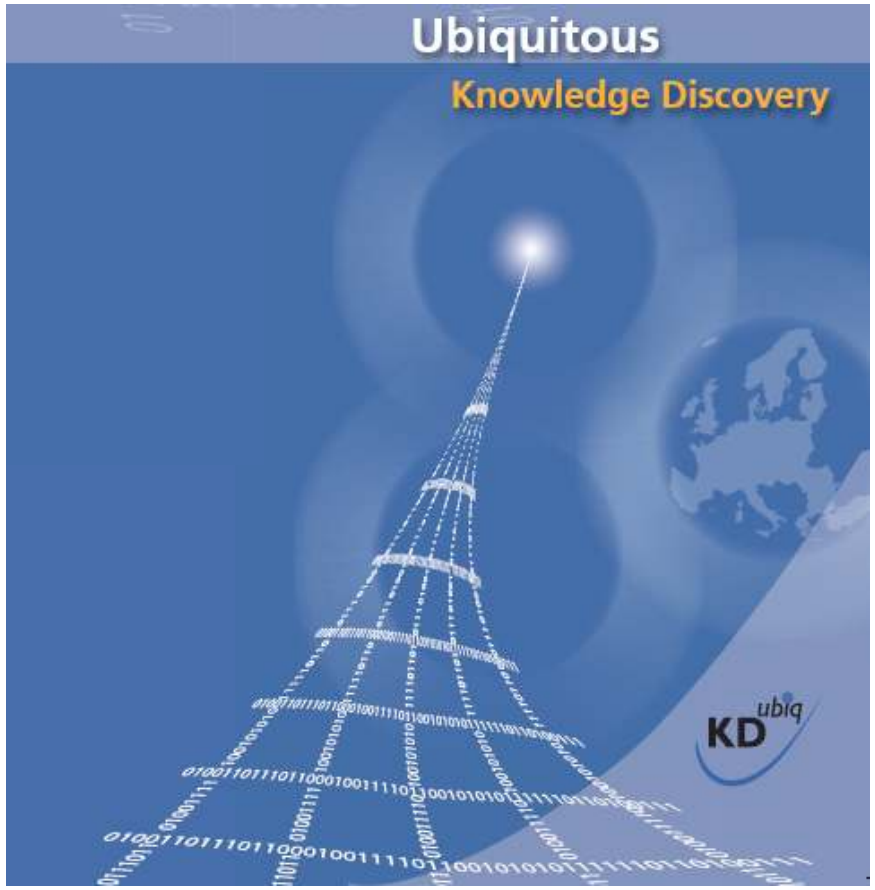
- Integrating results from
  - distributed data mining
  - privacy preserving data mining
  - spatio-temporal data mining
  - data stream mining
  - collaborative data mining

  in Ubiquitous Knowledge Discovery systems

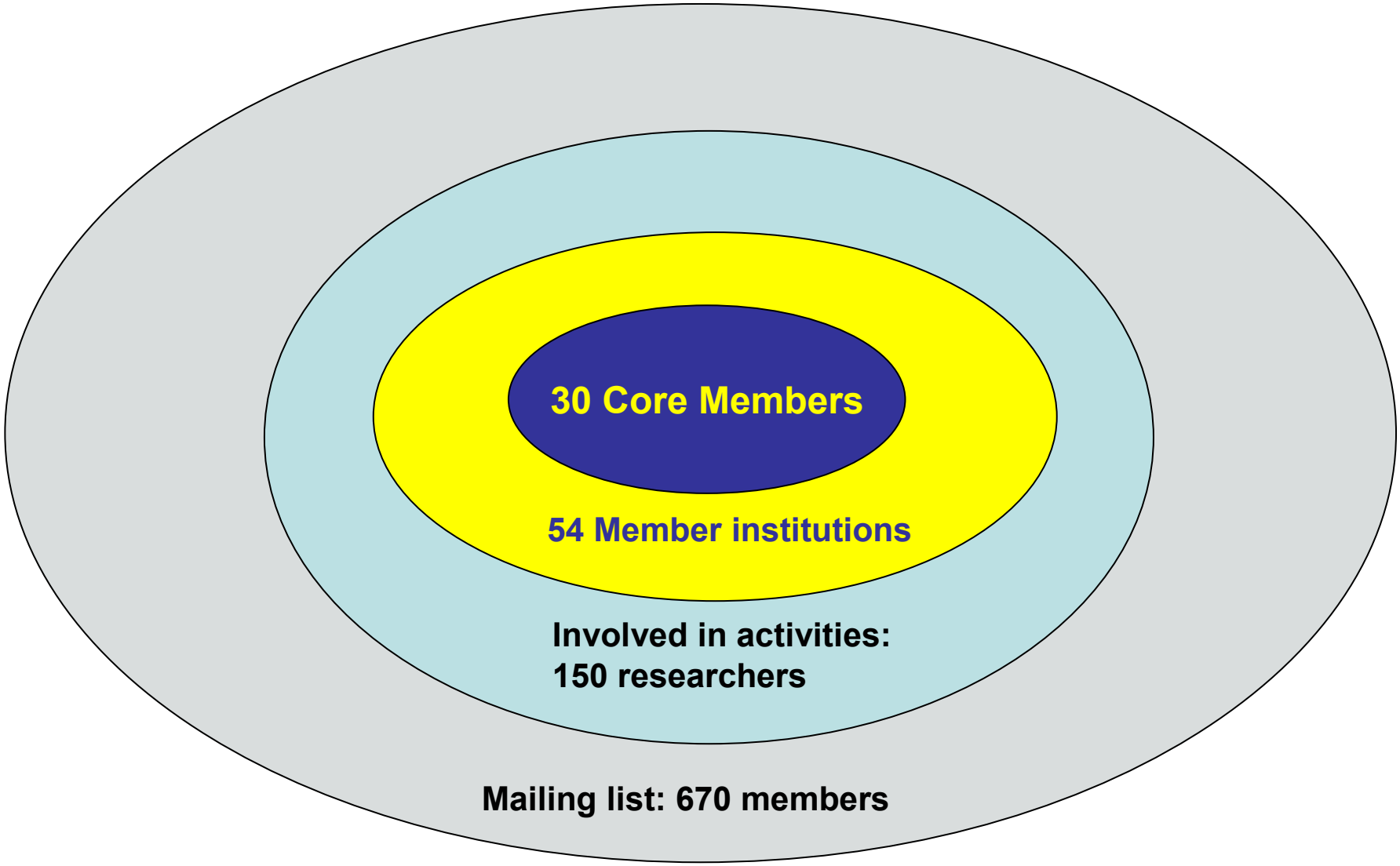
Towards KDUbiq

KD^{ubiq} Coordination Action

- To stimulate research, to define the field, and to shape the community in Europe, the KDUbiq research network was launched in 2006.
- It is funded by the European Commission
- Currently it has more than 50 members from research and industry
- Not a research project, it's about shaping a community
- Buget 1.2 Mio $, 2006-2008
- www.kdubiq.org



KDubiq Community



Objectives

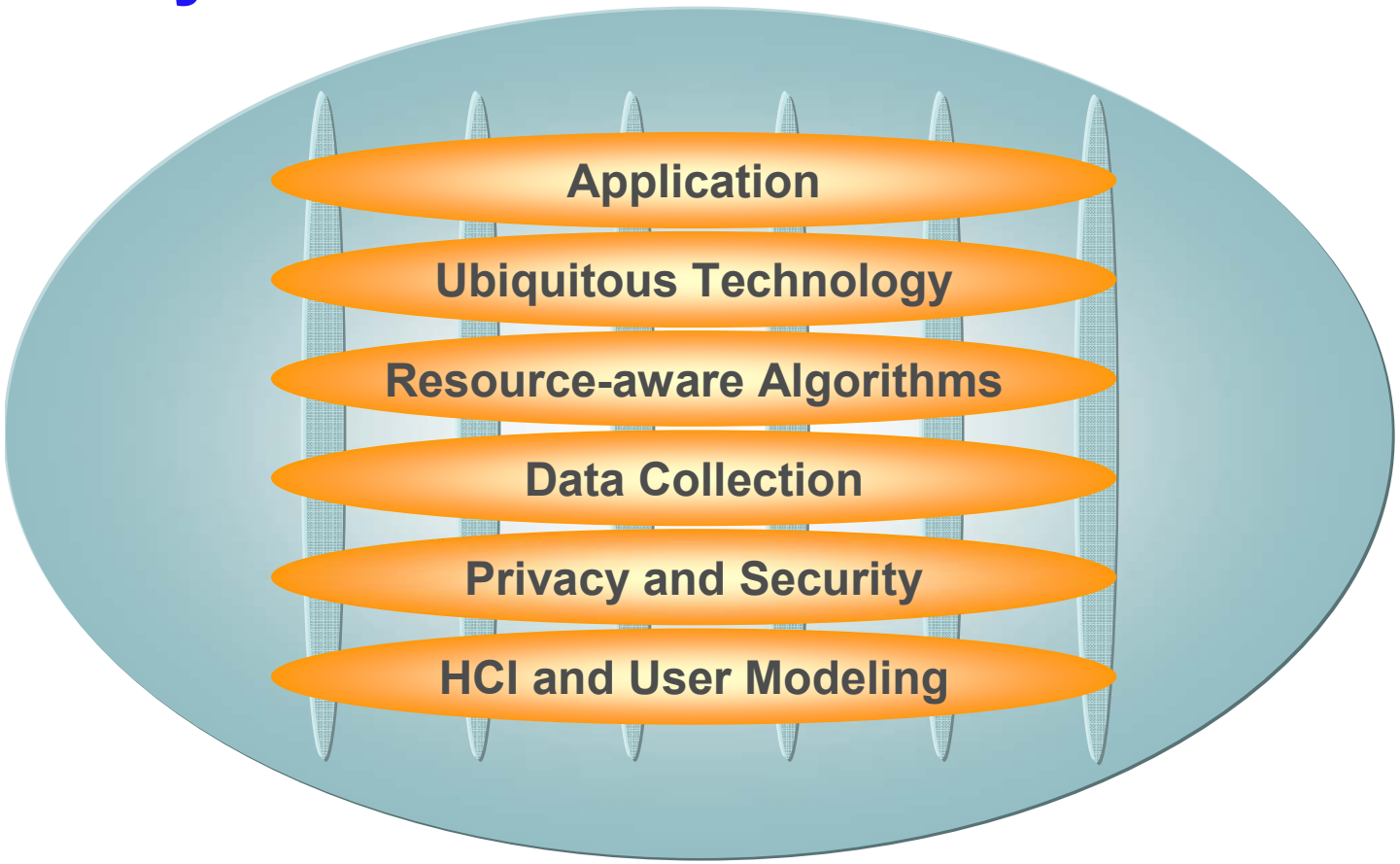
Today

- Relevant research is scatterd around several communities
- Highly interdisciplinary skills are needed
- Projects usually fall short of realizing the full potential of ubiquitous knowledge discovery

Goal

- bring researchers from relevant fields together
- form critical mass
- establish community
- define a joint research agenda

Dimensions of Ubiquitous Knowledge Discovery Systems



When designing a ubiquitous knowledge discovery system, major design decisions in each of these six dimensions have to be taken.

Choices are mutually constraining each other

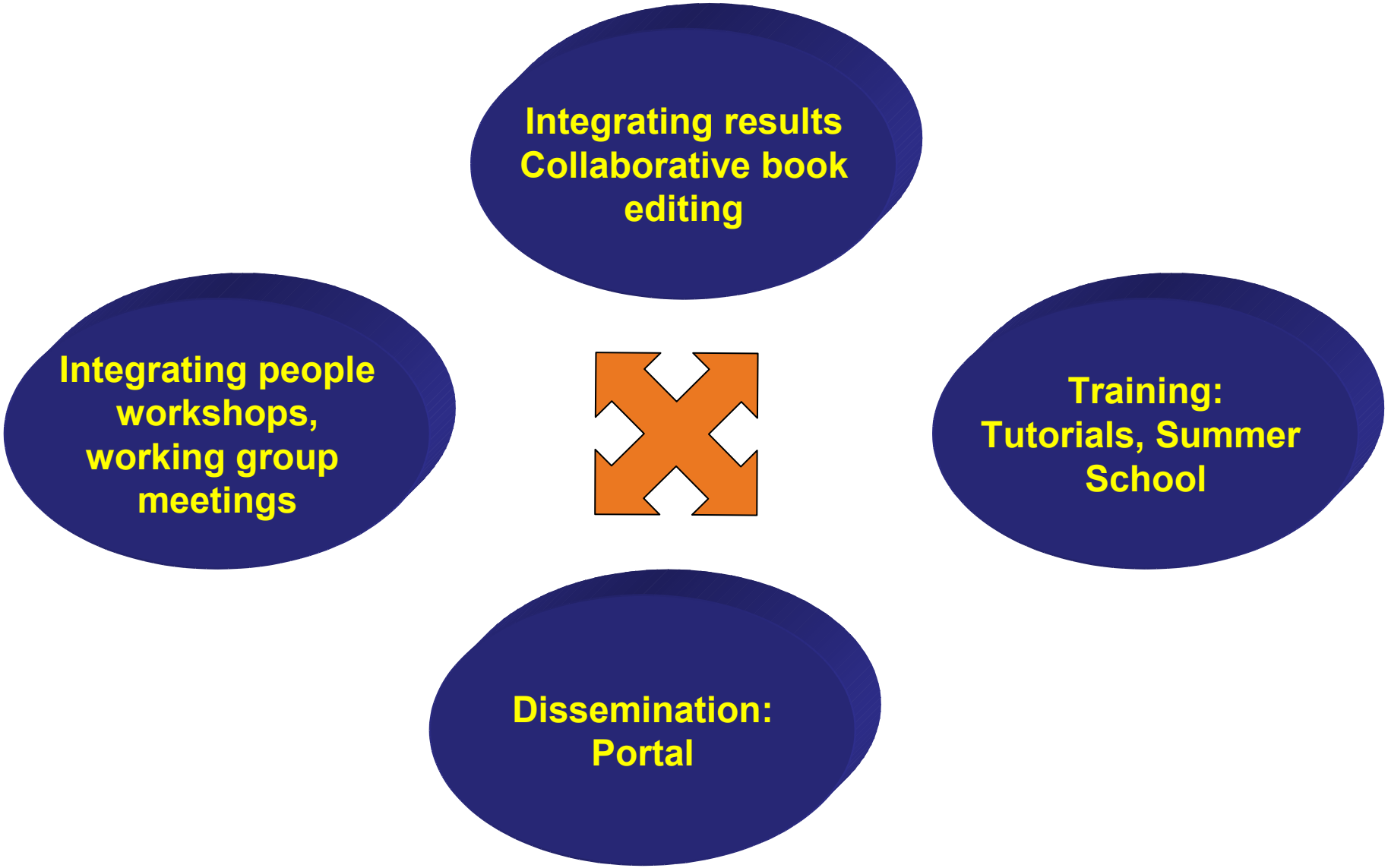
The best way to understand a system is to build one!

KDUbiq thus adopts a *systems* view on how to build next generation knowledge discovery systems

Working Group Chairs

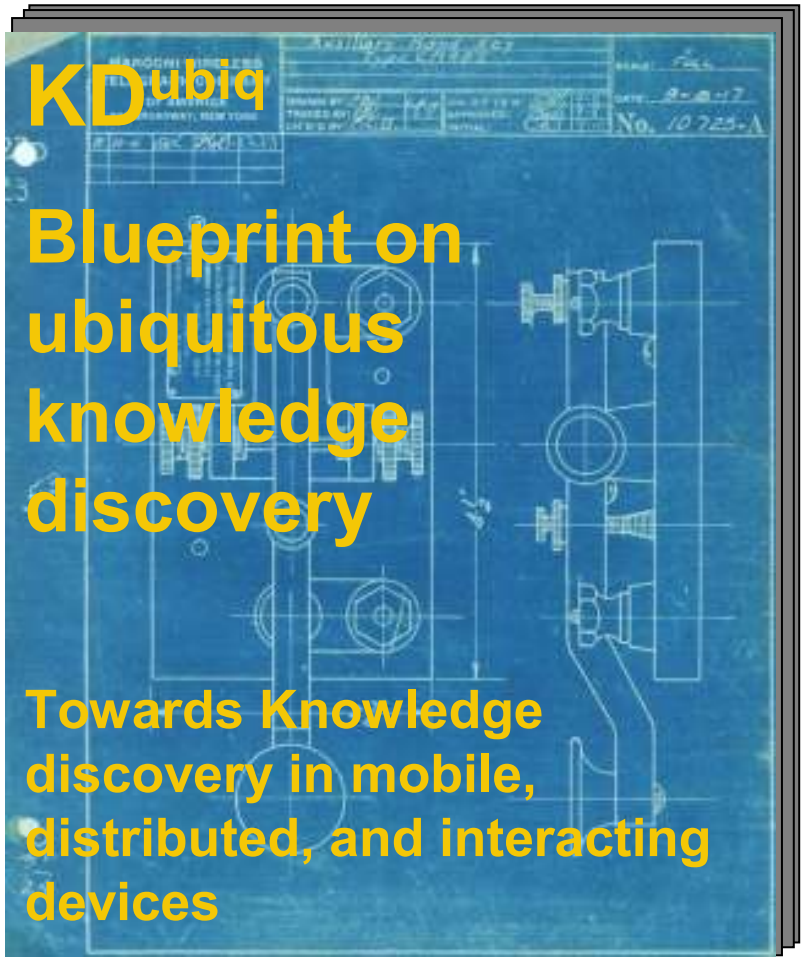
Working Group	WG Chairs	Country
Application Environments	**Koon Vanhoof, Hasselt Univ.**	Belgium
Ubiquitous Technologies	**Assaf Schuster, Technion** Ran Wolff, Univ. Haifa	Israel
Resource aware, distributed algorithms	**Joao Gama, Univ. Porto** Antoine Cornuéjols, Univ. Paris Sud	France Portugal
Ubiq. Interaction and Data Collection	**Rasmus Pedersen, Univ Copenhagen** Andreas Hotho, Univ. Kassel	Germany Denmark
Security & Privacy	**Fosca Giannotti, Univ. Pisa** Yücel Saygin, Sabanci Univ.	Italy Turkey
HCI & Cognitive Modelling	**Bettina Berendt, Humboldt Univ, Berlin** Ernestina Menasalvas, Univ. Madrid	Germany Spain

KDubiq Activities



Blueprint – collaborative book editing

- A collaborative effort to define the research challenges
- Six working groups corresponding to six main chapters
- 30 partners actively contributing
- Will result in a joint book in 2008



Workshops ´07

KD^{ubiq} Summer School

- Covers various areas relevant for ubiquitous knowledge discovery
  - sensor networks
  - embedded systems
  - distributed data mining
  - distributed computing and grid
  - collaborative environments
  - privacy and security
- Addresses both foundations and state of the art
- puts emphasis on exercises, will allow participants to get hands on experience in interdisciplinary work.

Second European Summer School on Knowledge Discovery for Ubiquitous Computing

Prof. Dr. Katharina Morik
Dr. João Gama

March 3 – 9, 2008
University of Porto, Portugal
www.kdubiq.org

Project Clustering

- KDubiq Project Presentation Day at ECML/PKDD 2006 and 2007 program
- The goal is networking and exchanging of results/experience.
- In total 24 projects presented
- each time ~ 350 partcipants

Rotes Rathaus, Berlin 2006

University Warsaw, 2007

KD ubiq ubiquitous knowledge discovery

SIMDAT

Anti Phish

DATAMINING grid DATA MINING TOOLS AND SERVICES FOR GRID COMPUTING ENVIRONMENTS

GeoPKDD

NeOn

PASCAL

ACGT

THESEUS

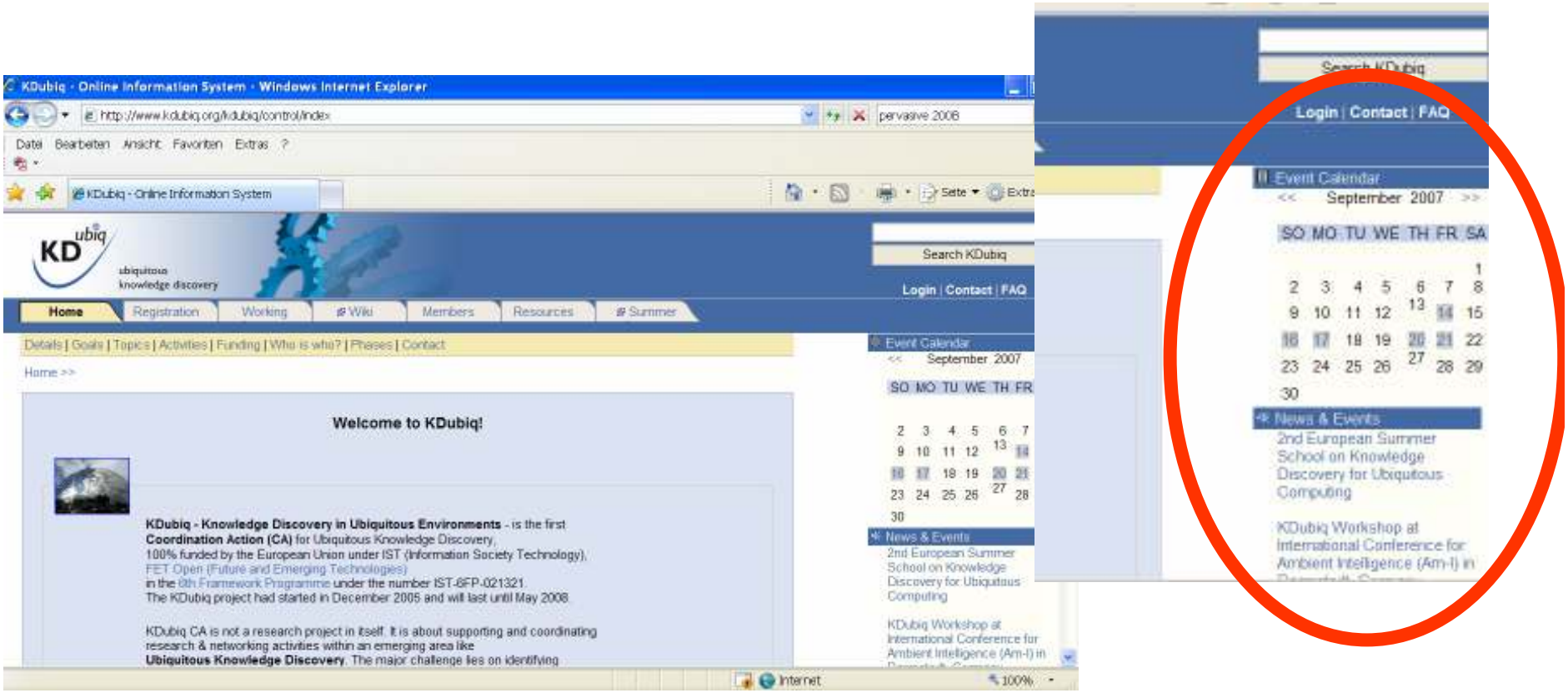
iWebCare

ASSIST The ASSIST E.C. Project

SWING

RAC WeB Risk Assessment for Customs in Western Balkans

KDubiq portal

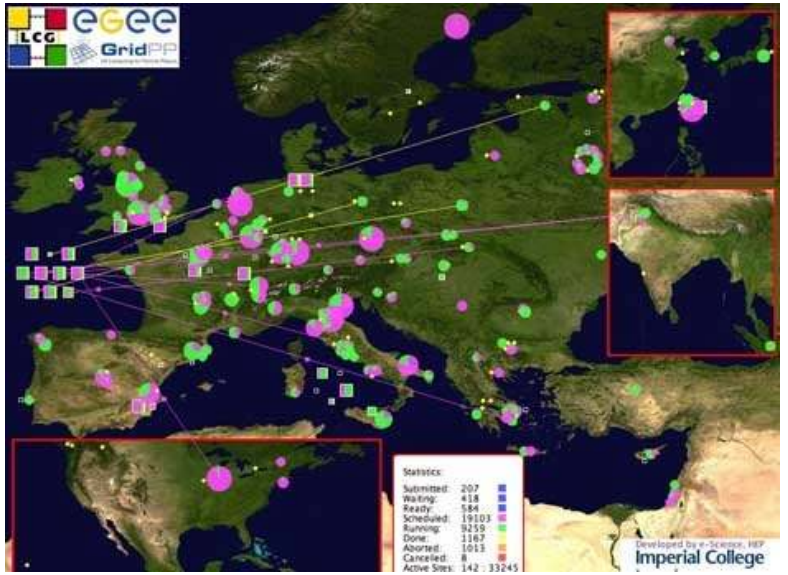


www.kdubiq.org

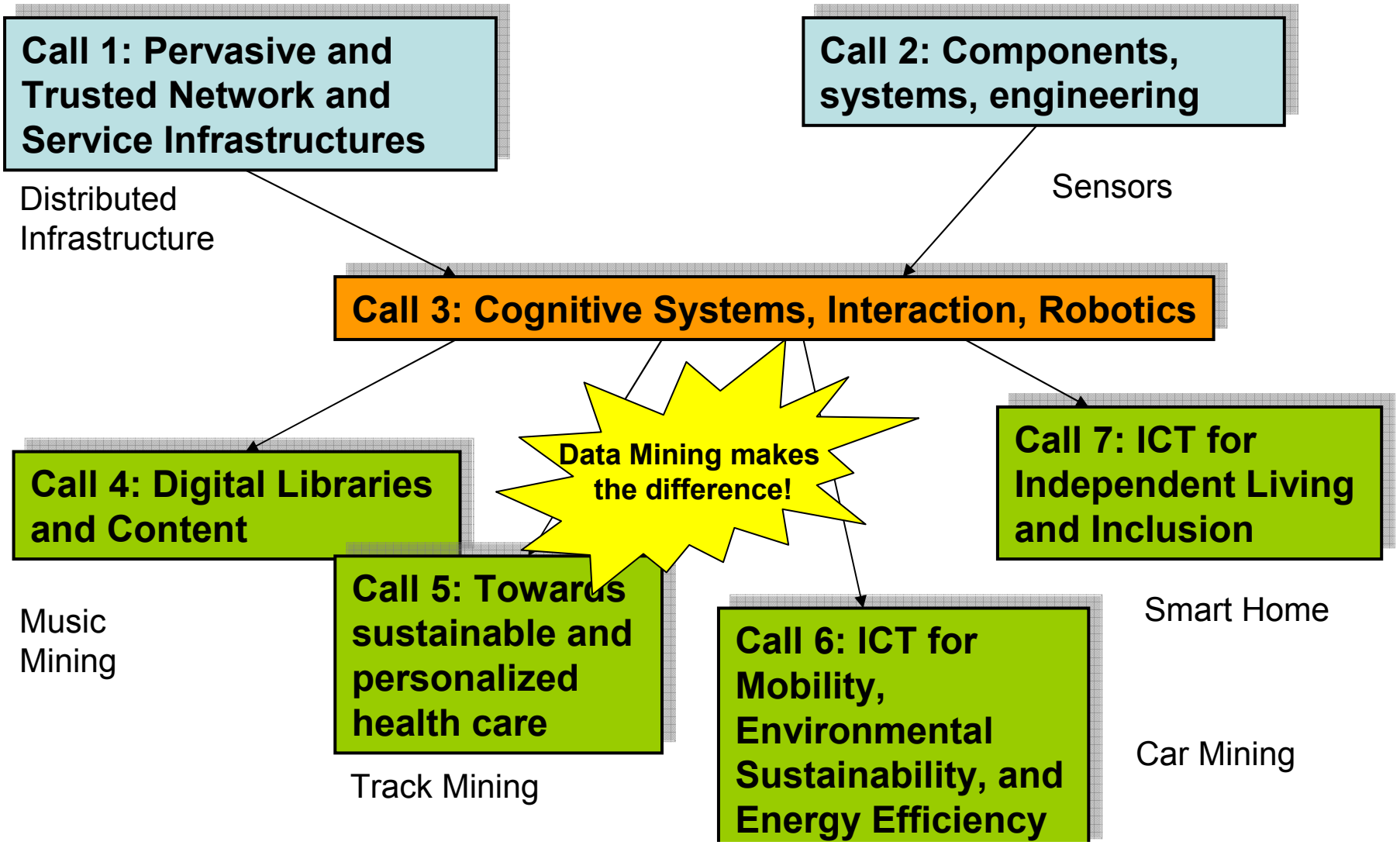
News Ticker
& KDNet &
KDUbiq
mailing lists

Benchmarking

- The availability of benchmarking data sets can have big impact on shaping a field
- The idea is to have some relevant and large-scale application in order to benchmark KDubiq algorithms,
- There has been work on identifying benchmarking data sets.
- a first data set from the grid computing area from the EGEE project.
- the data set will be used during the Summer School in Porto 2008.



A KDUbiq perspective on the European ICT Research Programme



Summary 1

- For many disciplines, further progress crucially depends on advances in data mining and machine learning
- Ubiquitous Knowledge Discovery provides an integrated perspective for studying data mining challenges in
  - distributed data mining
  - privacy preserving data mining
  - spatio-temporal data mining
  - data stream mining
  - collaborative data mining
- Taking a systems point view demonstrates the need for for work that *combines* several (or all) of these topics
- Biggest algorithmic challenge is to study learning settings that violate the iid assumption

Summary 2

Results on KDubiq networking activites

- Data Mining is a fragmented discipline
- people that never cooperated before and did not pay attention to the other‘s research work started collaborating
- Everyone in the network has a much broader view on ermerging issues and interrelations
- integrating research from data mining sub-communitites in research projects

Resources

- CDI should
  - allow for integrating work across data mining sub-communities
  - provide benchmarking data sets
  - provide support for networking
  - establish international links