Big Data and Location-Based Services: An Introduction

Yunjun Gao (高云君)

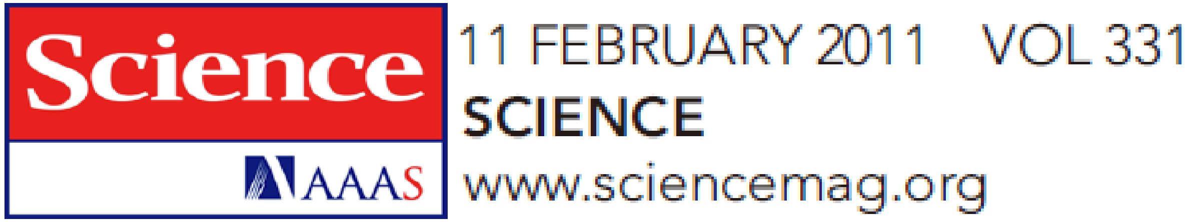
College of Computer Science
Zhejiang University
gaoyj@zju.edu.cn
13957167510

Information Explosion

- 988EB (1EB = 1024PB) data will be produced in 2010 (IDC) ⇔ 18 million times of all info in books
- IT
  - 850 million photos & 8 million videos every day (Facebook)
  - 50PB web pages, 500PB log (Baidu)
- Public Utilities
  - Health care (medical images - photos)
  - Public traffic (surveillance - videos)
- …

Research Frontier and Hot

- 《Science》: Special Online Collection: Dealing with Data
  - In this, *Science* joins with colleagues from *Science Signaling*, *Science Translational Medicine*, and *Science Careers* to provide a broad look at the issues surrounding the increasingly huge influx of research data. This collection of articles highlights both the challenges posed by the data deluge and the opportunities that can be realized if we can better organize and access the data.



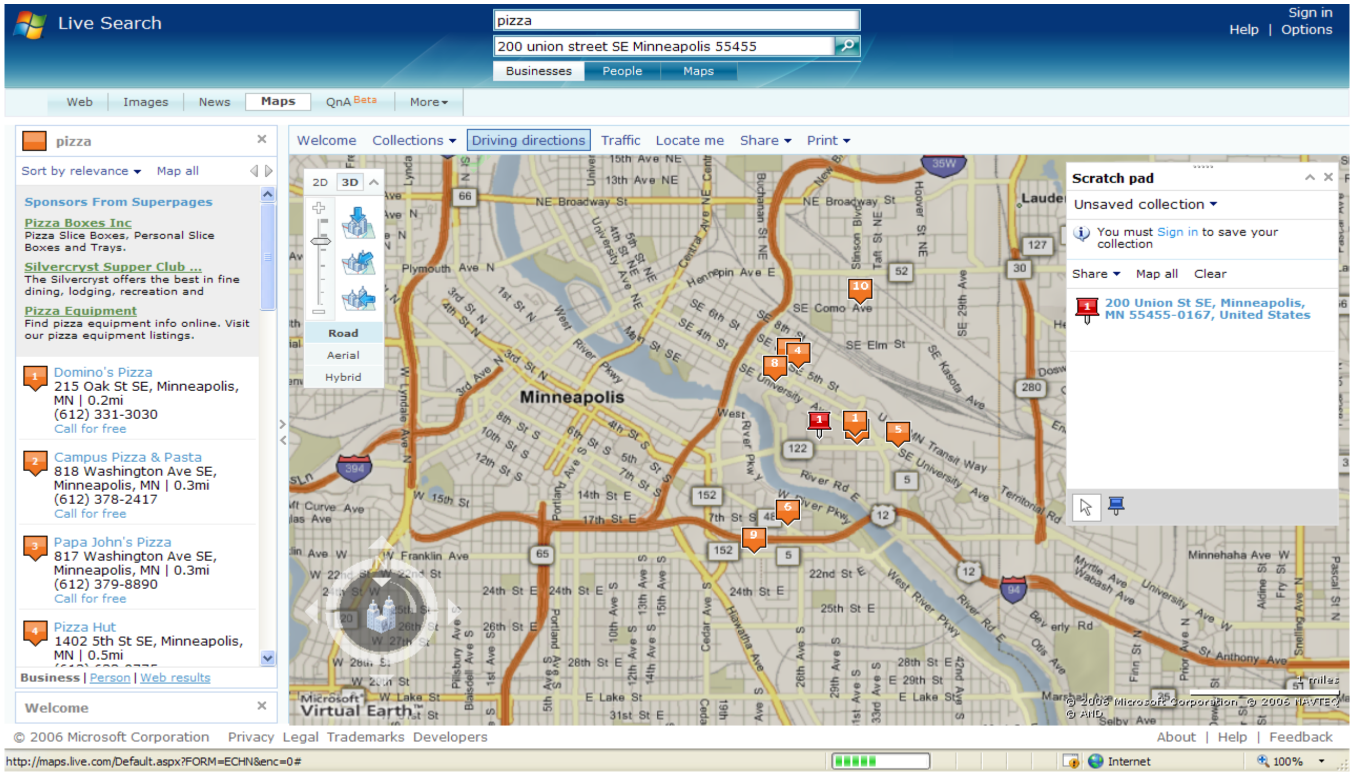
- 《Nature》:

Big Data Use Cases

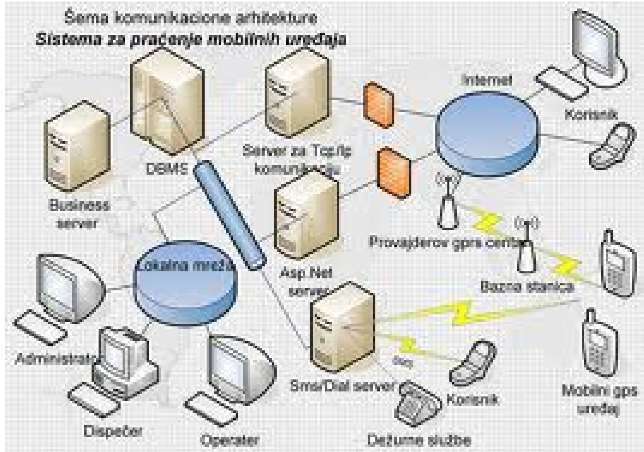
Today's Challenge	New Data	What's Possible
Healthcare Expensive office visits	Remote patient monitoring	Preventive care, reduced hospitalization
Manufacturing In-person support	Product sensors	Automated diagnosis, support
Location-Based Services Based on home zip code	Real time location data	Geo-advertising, traffic, local search
Public Sector Standardized services	Citizen surveys	Tailored services, cost reductions
Retail One size fits all marketing	Social media	Sentiment analysis segmentation

Location-Based Services

- Location-based services (LBS) provide the ability to find the geographical location of a mobile device and then provide services based on that location.
  - E.g., Yahoo/Google Maps, MapPoint, MapQuest, …



Challenges of LBS

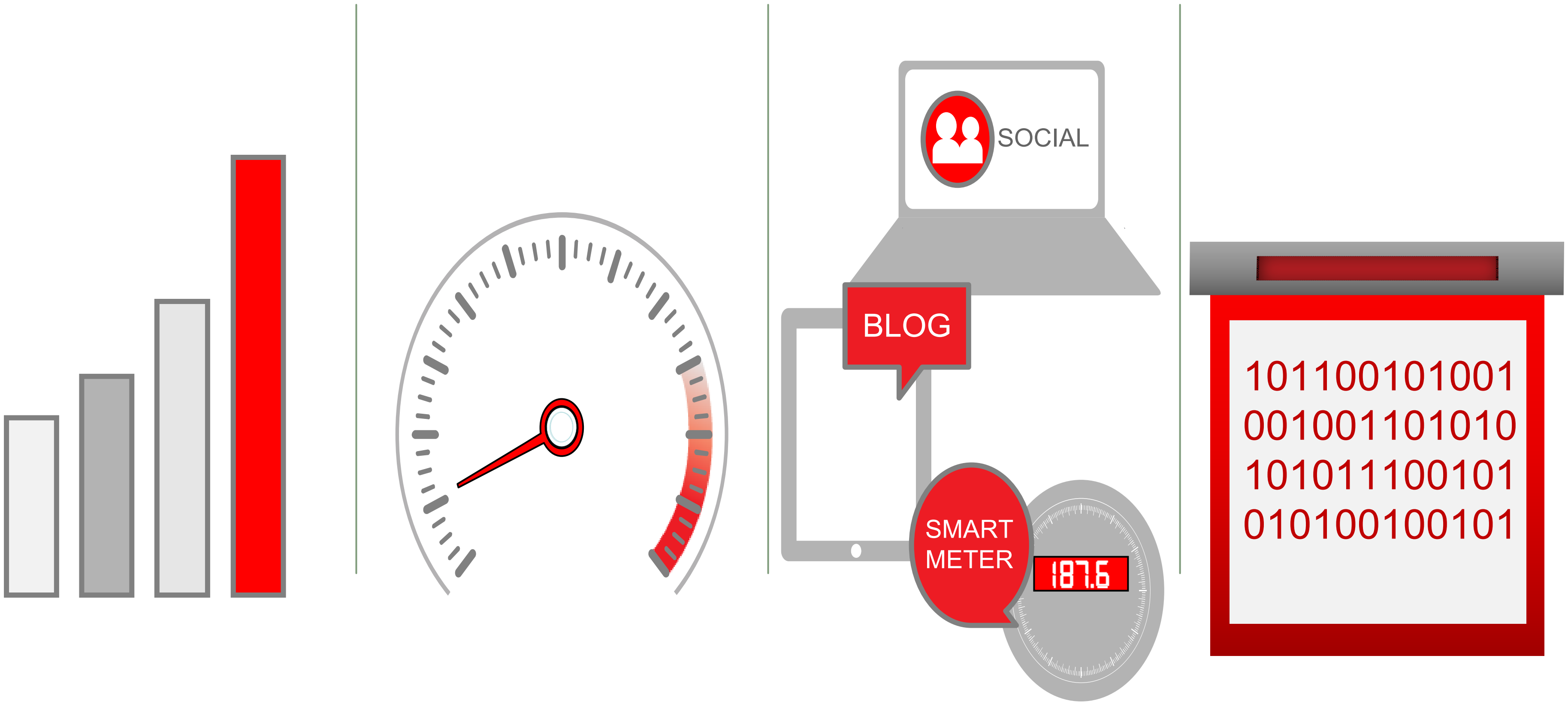


- ❑ Scalability
- ❑ Performance
  - ➢ Sustain high insertion rates
  - ➢ Query processing
  - ➢ Real-time query support
- ❑ High-precision positioning
- ❑ Privacy preservation
- ❑ Load Balance, i.e., overcome spatial and/or temporal data skew distribution

Outline

- Big Data
  - Definition
  - Properties
  - Applications
  - Framework
  - Challenges
  - Principles
  - Research Status
- Location-Based Services
  - Introduction
  - Research Status
  - Potential Research Contents
- Conclusions

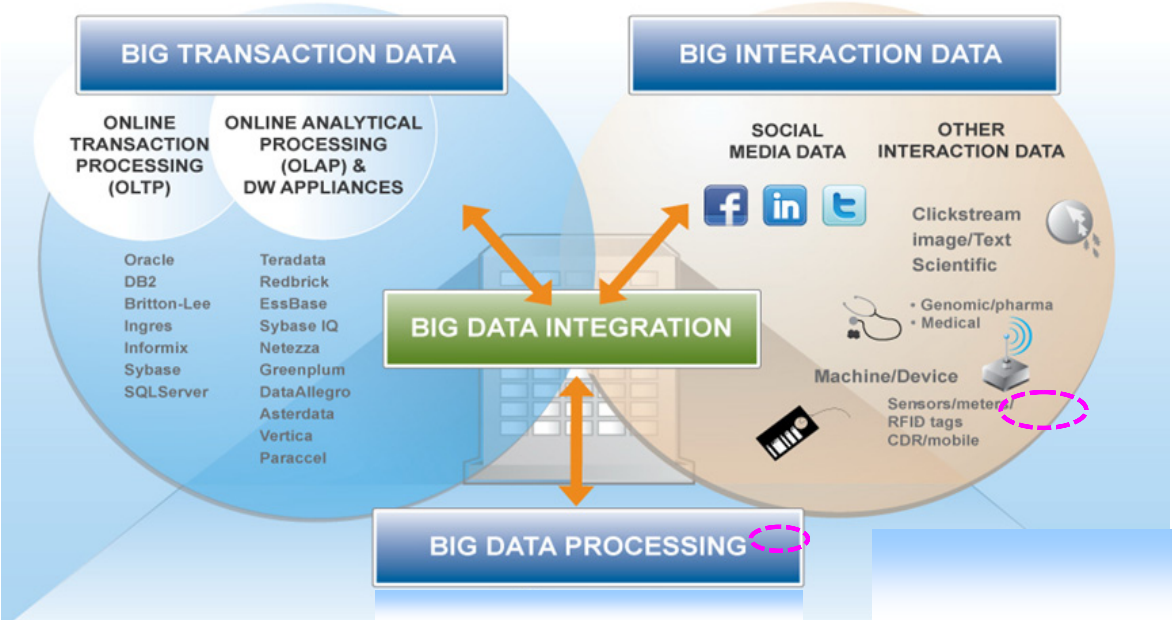
What Makes it Big Data?



VOLUME VELOCITY VARIETY VALUE

What is Big Data?

- **Definition:** Big Data refers to datasets that grow so large that it is difficult to capture, store, manage, share, analyze and visualize using the typical database software tools.

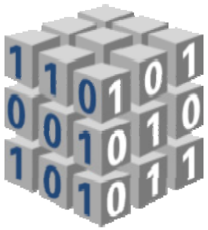


......
Unstructured data
Interaction Data

Structural and
Semi-Structural
Transaction Data

- **Questions:** Big Data = Large-Scale Data (Massive Data) 

Where Do We See Big Data?



Data Warehouses | OLTP | Social Networks | Scientific Devices

Everywhere

Diverse Data Sets

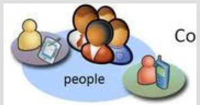
Big Data:

Decisions based on all your data

Video and Images

Documents

Social Data



Machine-Generated Data

Information Architectures Today:

Decisions based on database data

Transactions

IBM DB2

Microsoft SQL Server

ORACLE SIEBEL

SAP

Why Is Big Data Important?

US HEALTH CARE	MANUFACTURING	GLOBAL PERSONAL LOCATION DATA	EUROPE PUBLIC SECTOR ADMIN	US RETAIL
Increase industry value per year by	Decrease dev., assembly costs by	Increase service provider revenue by	Increase industry value per year by	Increase net margin by
\$300 B	**–50%**	**\$100 B**	**€250 B**	**60+%**

The Properties of Big Data

- Huge
- Distributed
  - Dispersed over many servers
- Dynamic
  - Items add/deleted/modified continuously
- Heterogeneous
  - Many agents access/update data
- Noisy
  - Inherent
  - Unintentional/Malicious
- Unstructured/semi-structured
  - No database schema
- Complex interrelationships

The Applications of Big Data

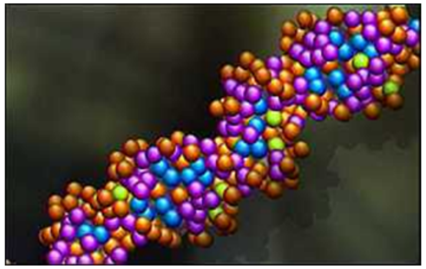
Celestial body
Exobiology
……

Astronomy

- Ast…
- …
- …

Data Mining
Consuming habit
……

Genomics



- 25,000 ge…
- 3 bil…
- 3 G…

Inheritance
Sequence of cancer
……

Changing router
……

Phone call billing records

AT&T

- 250M calls/day
- 60G calls/year
- 40 bytes/call
- 2.5 Terabytes/year

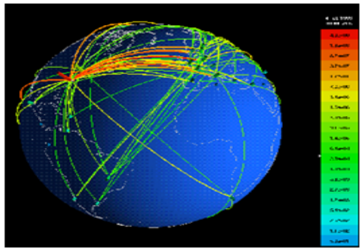
Advertisement
Finding communities
……

SNA
Finding communities
……

Credit card transactions

- 47.5 billion transactions in 2005 worldwide
- 115 Terabytes of data transmitted to VisaNet data processing center in 2004

Internet traffic



Traffic in a typical router:

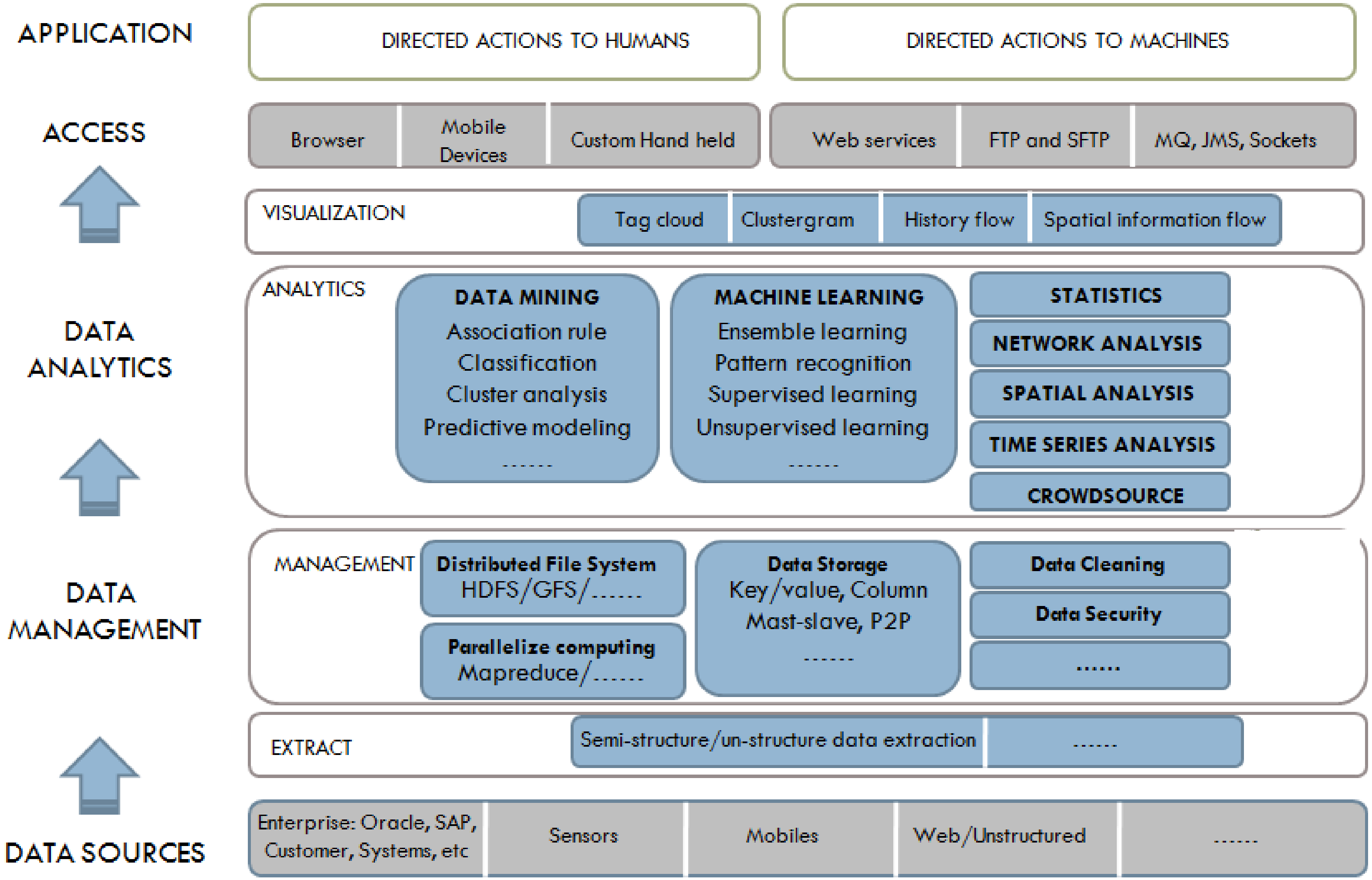
- 42 kB/second
- 3.5 Gigabytes/day
- 1.3 Terabytes/year

The World-Wide Web

Google™

- 25 billion pages indexed
- 10kB/Page
- 250 Terabytes of indexed text data
- "Deep web" is supposedly 100 times as large

The Framework of Big Data



The Challenges of Big Data

- Efficiency requirements for Algorithm
  - Traditionally, “efficient” algorithms
    - Run in (small) polynomial time: O(nlogn)
    - Use linear space: O(n)
  - For large data sets, efficient algorithms
    - Must run in linear or even sub-linear time: o(n)
    - Must use up to poly-logarithmic space: (logn)2

- Mining Big Data
  - Association Rule and Frequent Patterns
    - Two parameters: support, confidence
  - Clustering
    - Distance measure (L1, L2, L∞, Edit Distance, etc,.)
  - Graph structure
    - Social Networks, Degree distribution (heavy trail)

The Challenges of Big Data (Cont.)

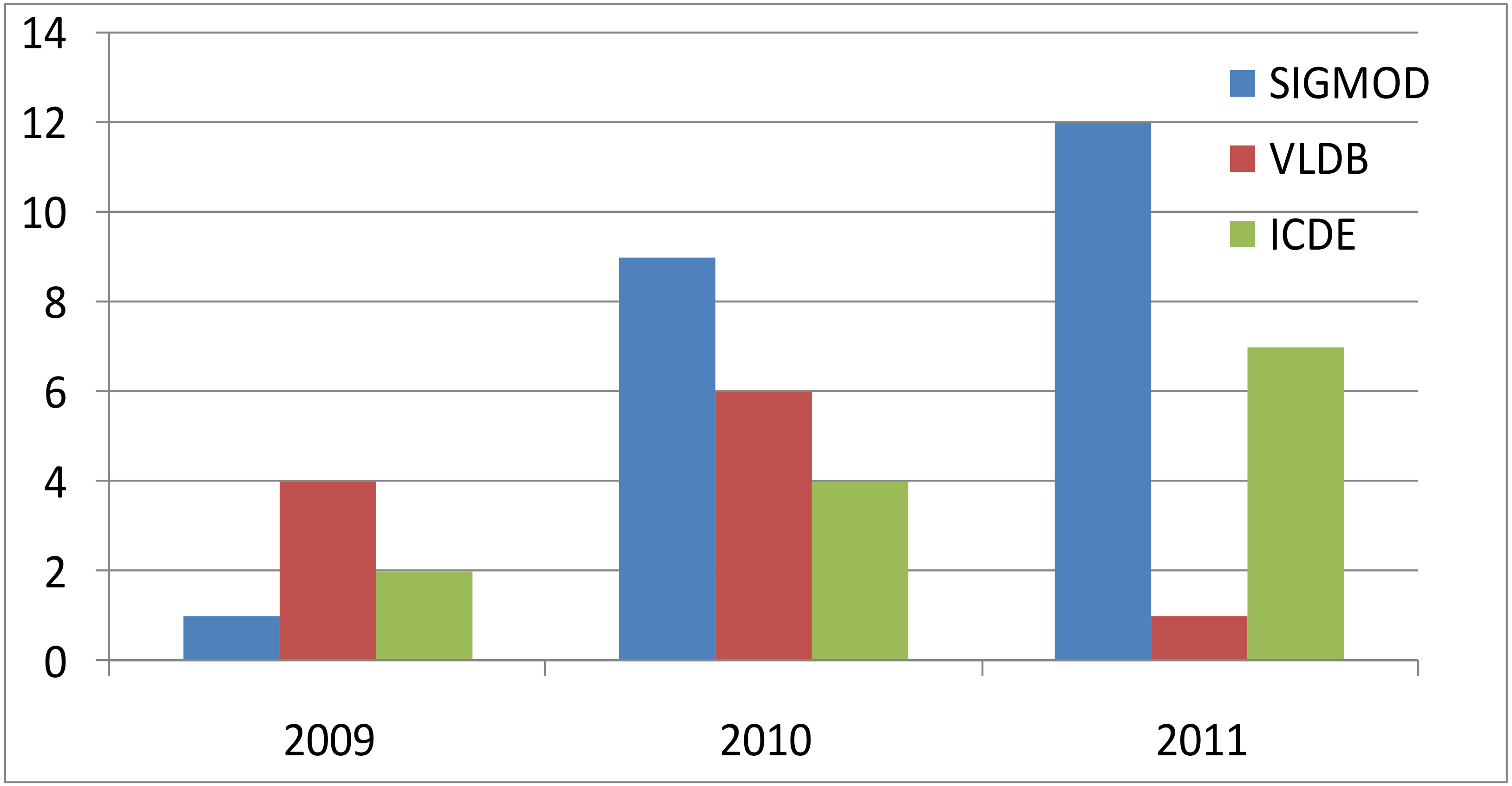
- Clean Big Data
  - Noise in data distorts
    - Computation results
    - Search results
  - Need automatic methods for “cleaning” the data
    - Duplicate elimination
    - Quality evaluation

- Computing Model
  - Accuracy and Approximation
  - Efficiency

The Principles of Big Data

- Partition Everything and key-value storage
  - 1st normal form cannot be satisfied
- Embrace Inconsistency
  - ACID properties are not satisfied
- Backup everything
  - Guarantee 99.999999% safety
- Scalable and high performance

Research Status



Research Status (Cont.)

- Indexes on Big Data ~ 4 papers
- Transactions on Big Data 4~5 papers
- Processing Architecture on Big Data 6~7 papers
- Applications in MapReduce Parallel Processing 6~7 papers
- Benchmark of Big Data Management System 3~4 papers

Outline

- Big Data
  - Definition
  - Properties
  - Framework
  - Applications
  - Challenges
  - Principles
  - Research Status
- Location-Based Services
  - Introduction
  - Research Status
  - Potential Research Contents
- Conclusions

Mobile Devices and Services

- Large diffusion of mobile devices, mobile services, and location-based services.

Which Location Data?

- Location data from mobile phones (e.g., iPhone, GPhone, etc.)
  - Cell positions in the GSM/UMTS network
- Location data from GPS-equipped devices
  - Humans (pedestrians, drivers) with GPS-equipped smart-phones
  - Vessels with AIS transmitters (due to maritime regulations)
- Location data from intelligent transportation environments
  - Vehicular ad-hoc networks (VANET)
- Location data from indoor positioning systems
  - RFIDs (radio-frequency ids)
  - Wi-Fi access points

Examples of Location Data

- Vehicles (private cars) moving in Milan
  - ~2M GPS recordings from 17241 distinct objects (7 days period, 214,780 trajectories)

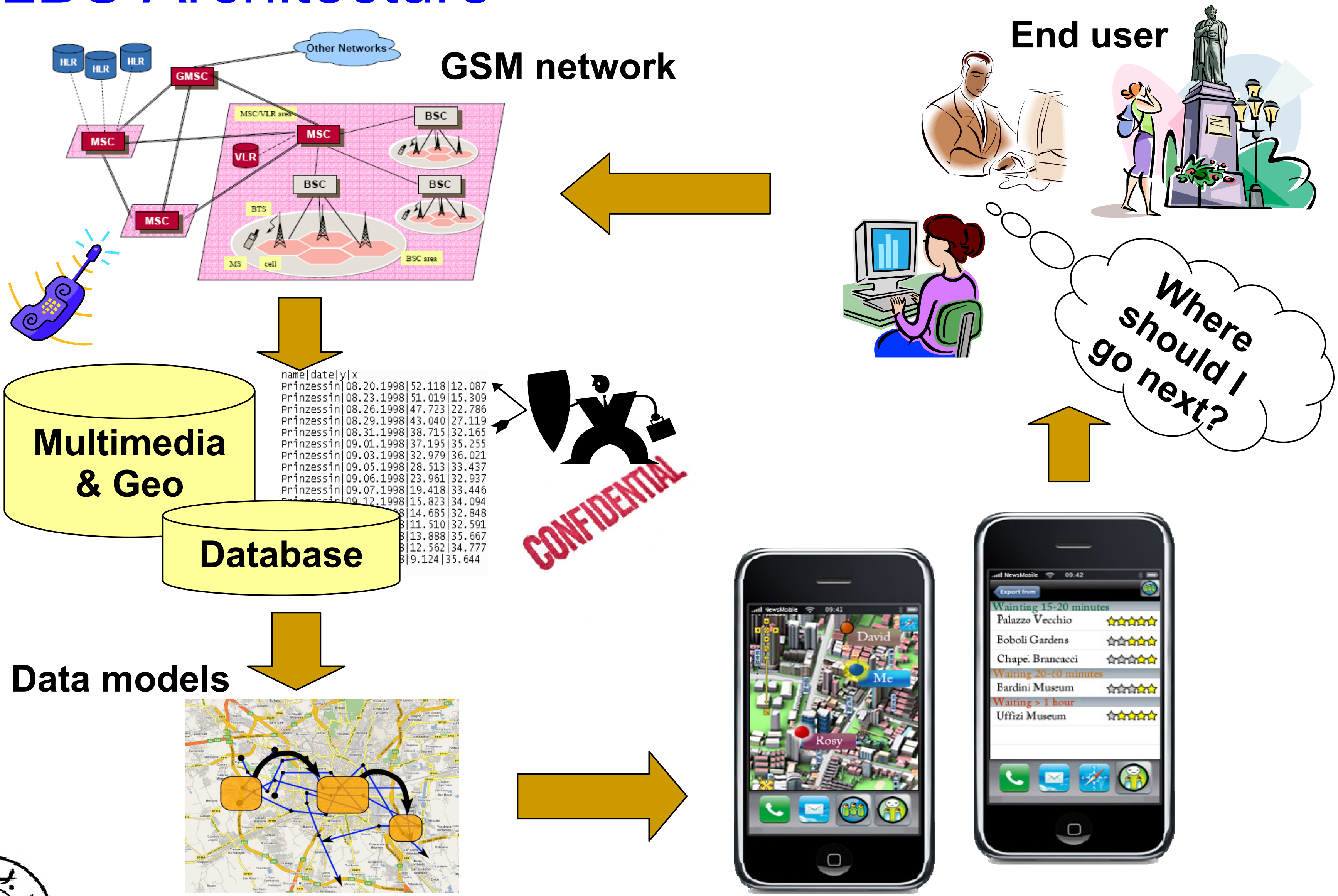
- Vehicles (couriers) moving in London
  - ~92.5M GPS recordings from 126 distinct objects (18 months period, 72,389 trajectories)

- Vessels sailing in Mediterranean sea
  - ~4.5M GPS recordings from 1753 distinct objects (3 days period, 1503 trajectories)

What Can We Learn From Location Data?

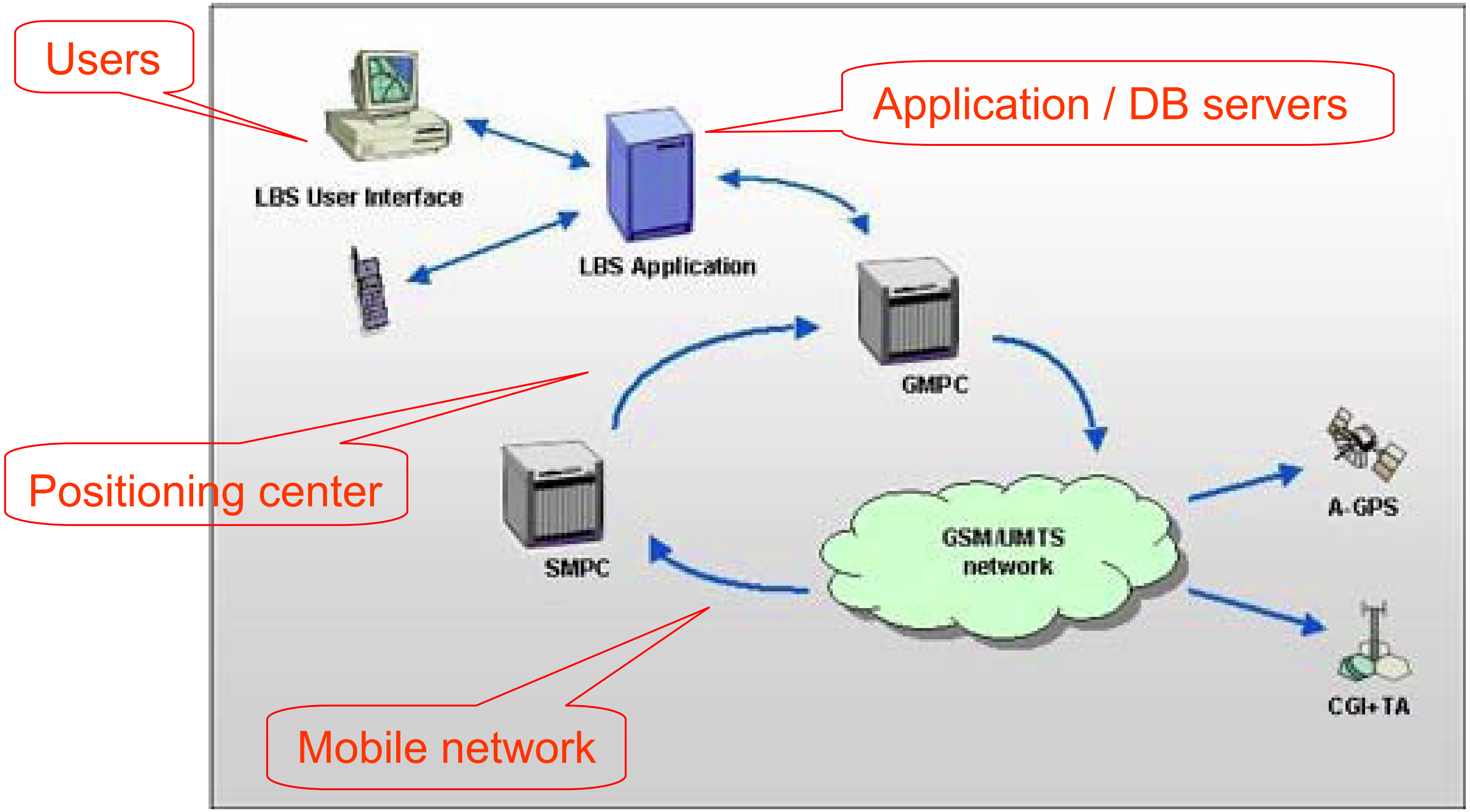
- Traffic monitoring
  - How many cars are in the downtown area?
  - Send an alert if a non-friendly vehicle enters a restricted region
  - Once an accident is discovered, immediately send alarm to the nearest police and ambulance cars
- Location-aware queries
  - Where is my nearest Gas station?
  - What are the fast food restaurants within 3 miles from my location?
  - Let me know if I am near to a restaurant while any of my friends are there
  - Send E-coupons to all customers within 3 miles of my stores
  - Get me the list of all customers that I am considered their nearest restaurant
- …

LBS Architecture



LBS Infrastructure

- Mobile Location Systems (MLS): four main components:



LBS Infrastructure (Cont.)

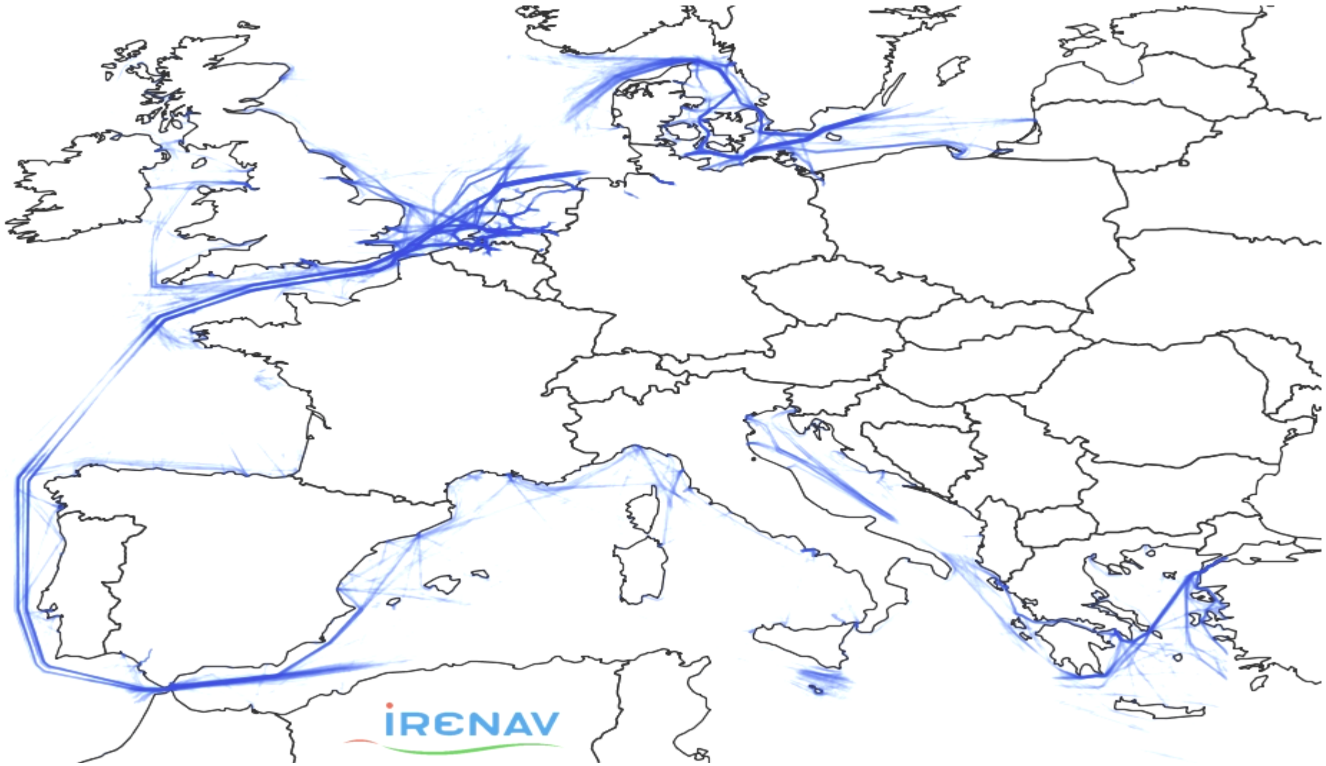
- A spatial database manages spatial objects:
  - Points: e.g., locations of hotels/restaurants
  - Line segments: e.g., road segments
  - Polygons: e.g., landmarks, layout of VLSI, regions/areas

Road Network

Satellite Image

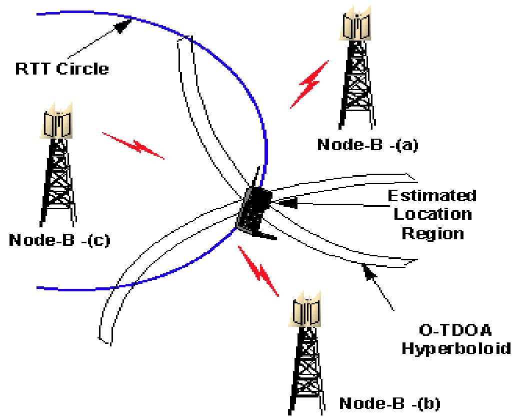
LBS Infrastructure (Cont.)

- Spatio-temporal database = Spatial database + time

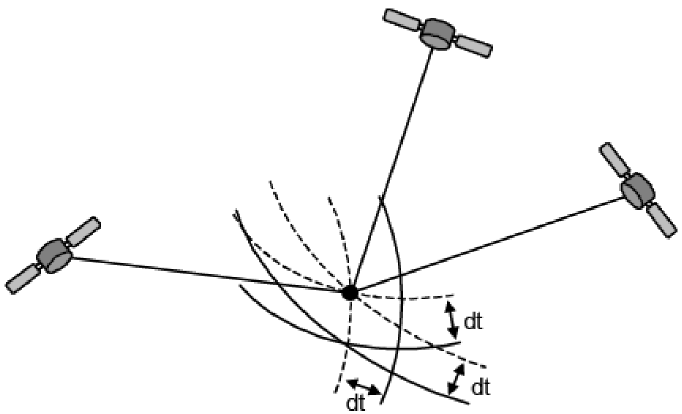


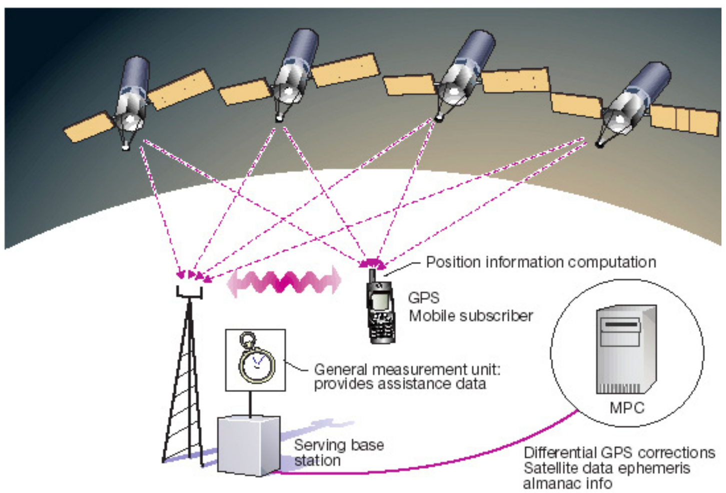
LBS Infrastructure (Cont.)

- Geo-positioning technologies:
  - Using the mobile telephone network
    - Time of Arrival (TOA), UpLink TOA (UL-TOA)



  - Using information from satellites
    - Global Positioning System (GPS)
    - Assisted (A-GPS), Differential GPS (D-GPS)





LBS Applications

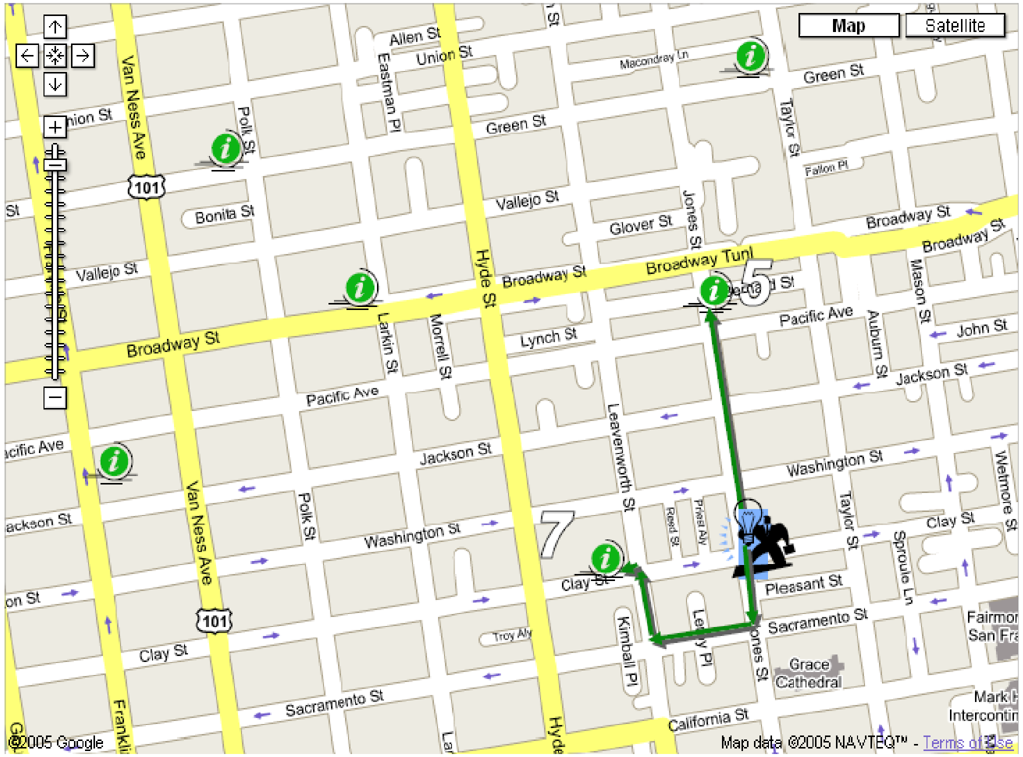
- Navigation (for vehicle or pedestrian)
  - Routing, finding the nearest point-of-interest (POI), …
- Information services
  - Find-the-Nearest, What-is-around, …
- Tracing services
  - Tracing of a stolen phone/car, locating persons in an emergency situation, …
- Resource management
  - (taxi, truck, etc.) fleet management, administration of container goods, …

LBS Applications (Cont.)

- On-board navigation, e.g., Dash express (http://www.dash.net)
  - Internet-connected automotive navigation system
  - Up-to-minute information about traffic
  - Yahoo! Local search for finding POIs

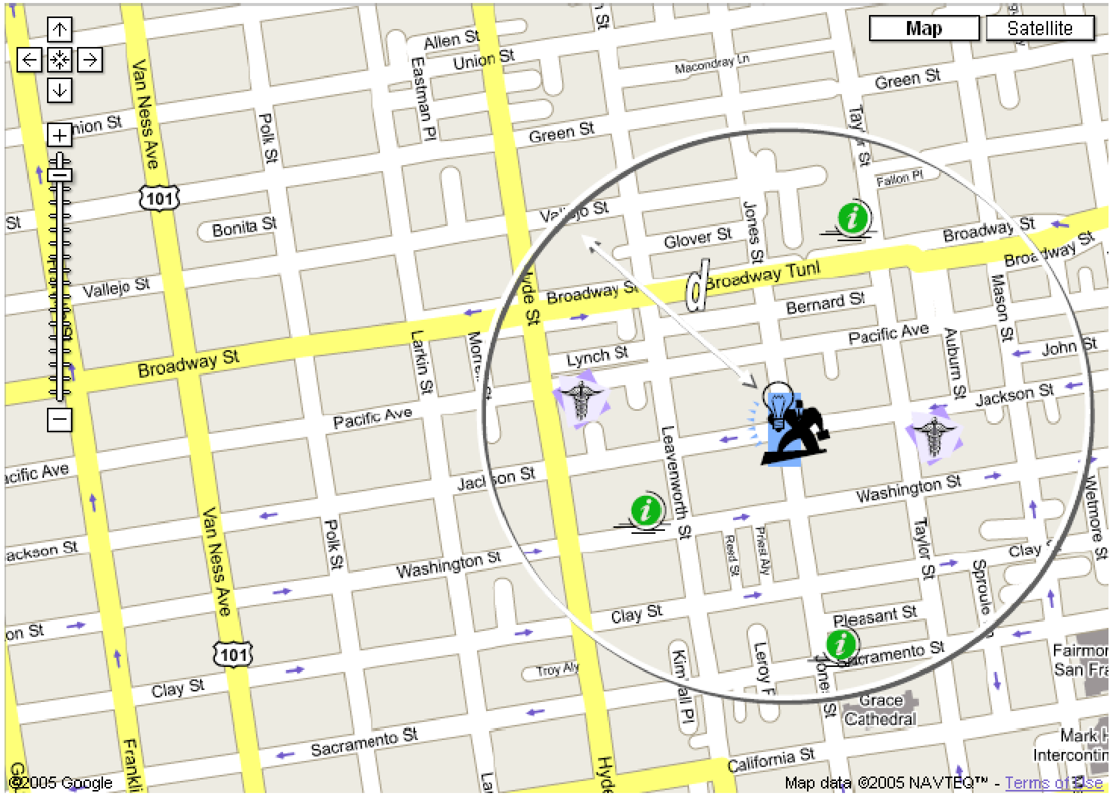
LBS Applications (Cont.)

- Find-the-Nearest: Retrieve and display the nearest POI (restaurants, museums, gas stations, hospitals, etc.) with respect to a specified reference location
  - E.g., find the two restaurants that are closest to my current location



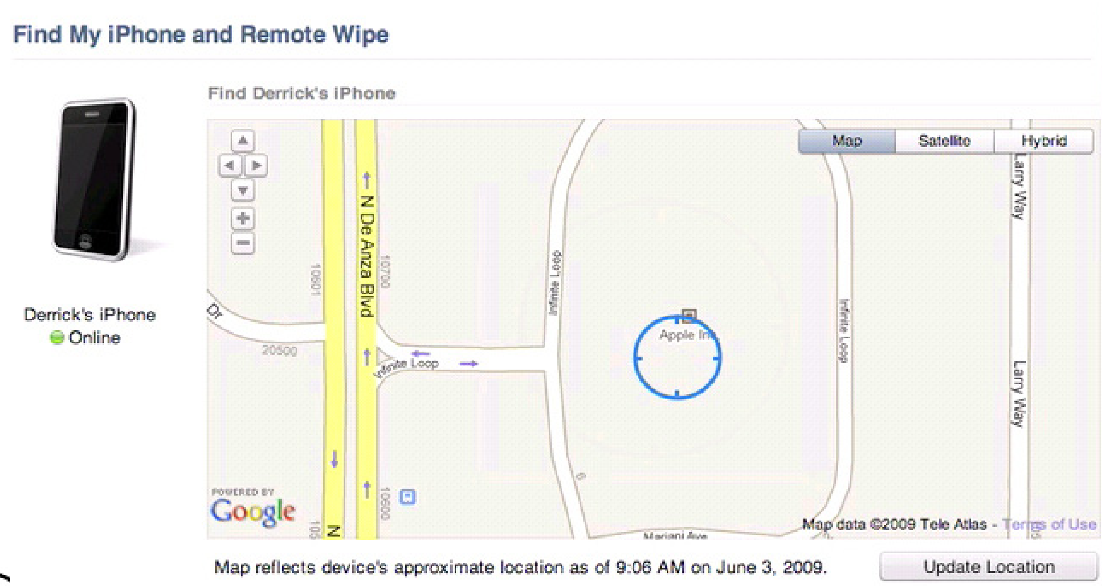
LBS Applications (Cont.)

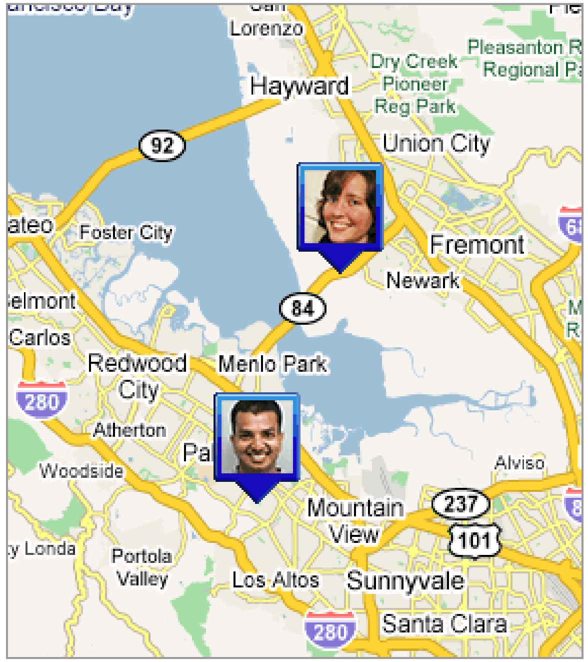
- What-is-around: Retrieve and display all POI located in the surrounding area (according to user's location or an arbitrary point)
  - E.g., get me all the gas-stations and ATMs within a distance of 1km



LBS Applications (Cont.)

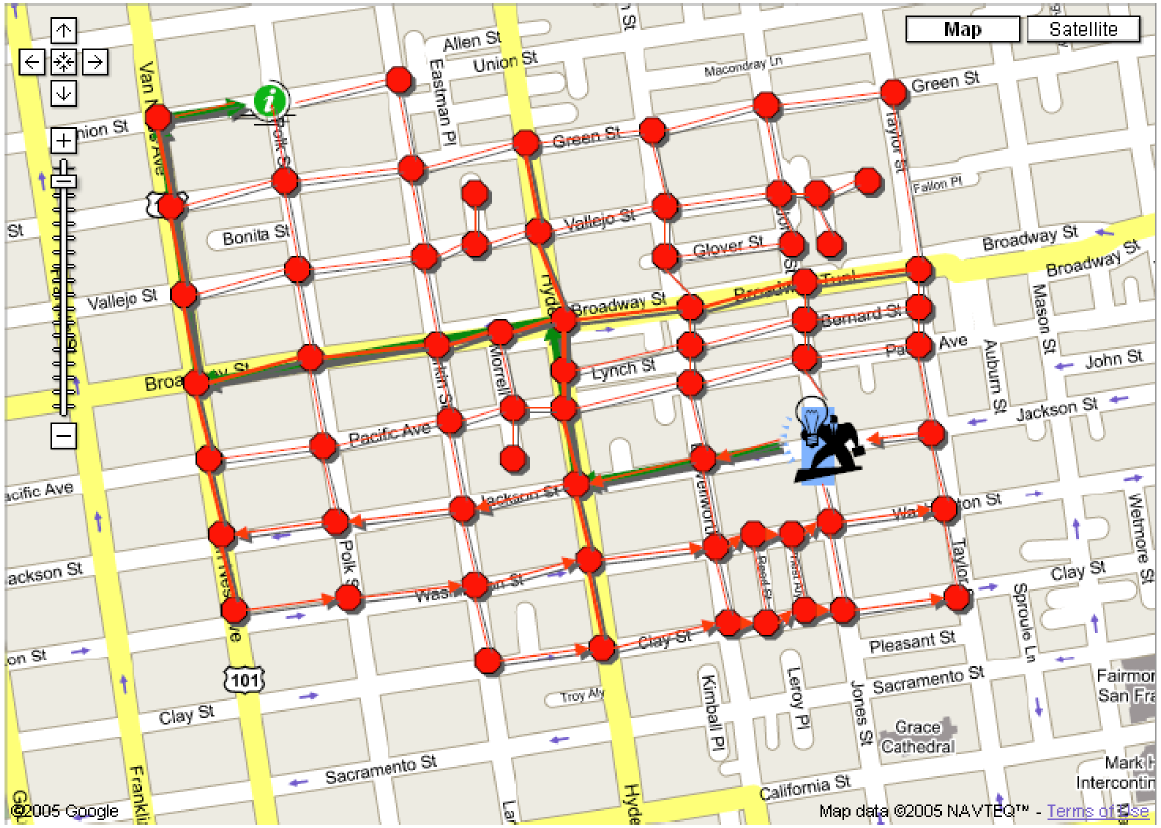
- **Google**
  - See in real time where your friends are! (launched by Google)
- **Apple**
  - Find my iPhone, i.e., track your lost iPhone (launched by Apple)





LBS Applications (Cont.)

- Route
  - E.g., Find the optimal route from a departure to a destination point



Oversea Past/Recent/Ongoing Research

- Cyrus Shahabi (University of Southern California, USA)
  - Privacy in Location-Based Services
  - Advanced query processing in road networks
- Ling Liu (Georgia Institute of Technology, USA)
  - mTrigger: Location-based Triggers
  - Scalable and Location-Privacy Preserving Framework for Large Scale Location Based Services
- Jiawei Han (University of Illinois, Urbana-Champaign, USA)
  - MoveMine: Mining Sophisticated Patterns and Actionable Knowledge from Massive Moving Object Data
- Amr El Abbadi (University of California, Santa Barbara, USA)
  - Location Based Services

Oversea Past/Recent/Ongoing Research (Cont.)

- Mohamed F. Mokbel (University of Minnesota, Twin Cities, USA)
  - Preference- And Context-Aware Query Processing for Location-based Data-base Servers
  - Towards Ubiquitous Location Services: Scalability and Privacy of Location-based Continuous Queries
- Vassilis J. Tsotras (University of California, Los Angeles, USA)
  - Query Processing Techniques over Objects with Functional Attributes
  - Graceful Evolution and Historical Queries in Information Systems -- a Unified Approach
- Ouri Wolfson (University of Illinois, Chicago, USA)
  - Location Management and Moving Objects Databases
- Wang-Chien Lee (The Pennsylvania State University, USA)
  - Location Based Services

Oversea Past/Recent/Ongoing Research (Cont.)

- Edward P.F. Chan (University of Waterloo, Canada)
  - Optimal Route Queries
- Christian S. Jensen (Aarhus University, Denmark)
  - TransDB: GPS Data Management with Applications in Collective Transport
  - LBS: Data Management Support for Location-Based Services
  - TRAX: Spatial Tracking and Event Monitoring for Mobile Services
- Stefano Spaccapietra (Swiss Federal Institute of Technology - Lausanne, Switzerland)
  - GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery
- Hans-Peter Kriegel (Ludwig-Maximilians-Universität München, Germany)
  - Data Mining and Routing in Traffic Networks

Oversea Past/Recent/Ongoing Research (Cont.)

- Bernhard Seeger (University of Marburg, Germany)
  - Spatial-aware querying the WWW
- Yannis Theodoridis: University of Piraeus, Greece)
  - MODAP: Mobility, Data Mining, and Privacy
  - GeoPKDD: Geographic Privacy-aware Knowledge Discovery and Delivery
- Dieter Pfoser (Institute for the Management of Information Systems, Greece)
  - GEOCROWD: Creating a Geospatial Knowledge World
  - TALOS: Task aware location based services for mobile environments
- Ooi Beng Chin (National University of Singapore, Singapore)
  - Co-Space
- Roger Zimmermann (National University of Singapore, Singapore)
  - Location-based Services in Support of Social Media Applications

Oversea Past/Recent/Ongoing Research (Cont.)

- Kyriakos Mouratidis (Singapore Management University, Singapore)
- Xiaofang Zhou (The University of Queensland, Australia)
  - Making Sense of Trajectory Data: a Database Approach
- Dimitris Papadias (Hong Kong University of Science and Technology, China)
- Yufei Tao (Chinese University of Hong Kong, China)
  - Data Retrieval Techniques on Spatial Networks
  - Query Processing on Historical Uncertain Spatiotemporal Data
  - Approximate Aggregate Processing in Spatio-temporal Databases
- Nikos Mamoulis (Hong Kong University, China)
- Man Lung Yiu (Hong Kong Polytechnic University, China)

Domestic Past/Recent/Ongoing Research

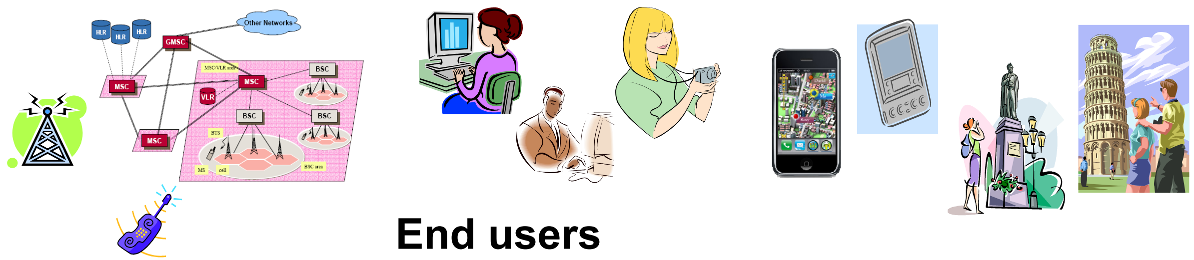
- Xiaofeng Meng (Renmin University of China, China)
  - Mobile Data Management
  - Location-Based Privacy Protection
- Yu Zheng (Microsoft Research Asia, China)
  - T-Drive
  - GeoLife 2.0
- Zhiming Ding (Chinese Academy of Sciences, China)

- Summary
  - To the best of our knowledge, there is little work on Location-Based Services in China.

Summary of Research Status

- The existing research works mostly focus on Privacy Preservation, LBS Architecture, Location Prediction, LBS applications, and so on.
- Several LBS-related Labs in universities, e.g., PSU (USA), UCSB (USA), Tokyo University (Japan), KAIST (Korean), etc., have been founded in recent years.
- To the best of our knowledge, there is little work on Location-Based Services in China.

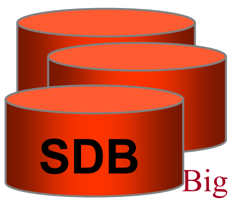
Framework



End users

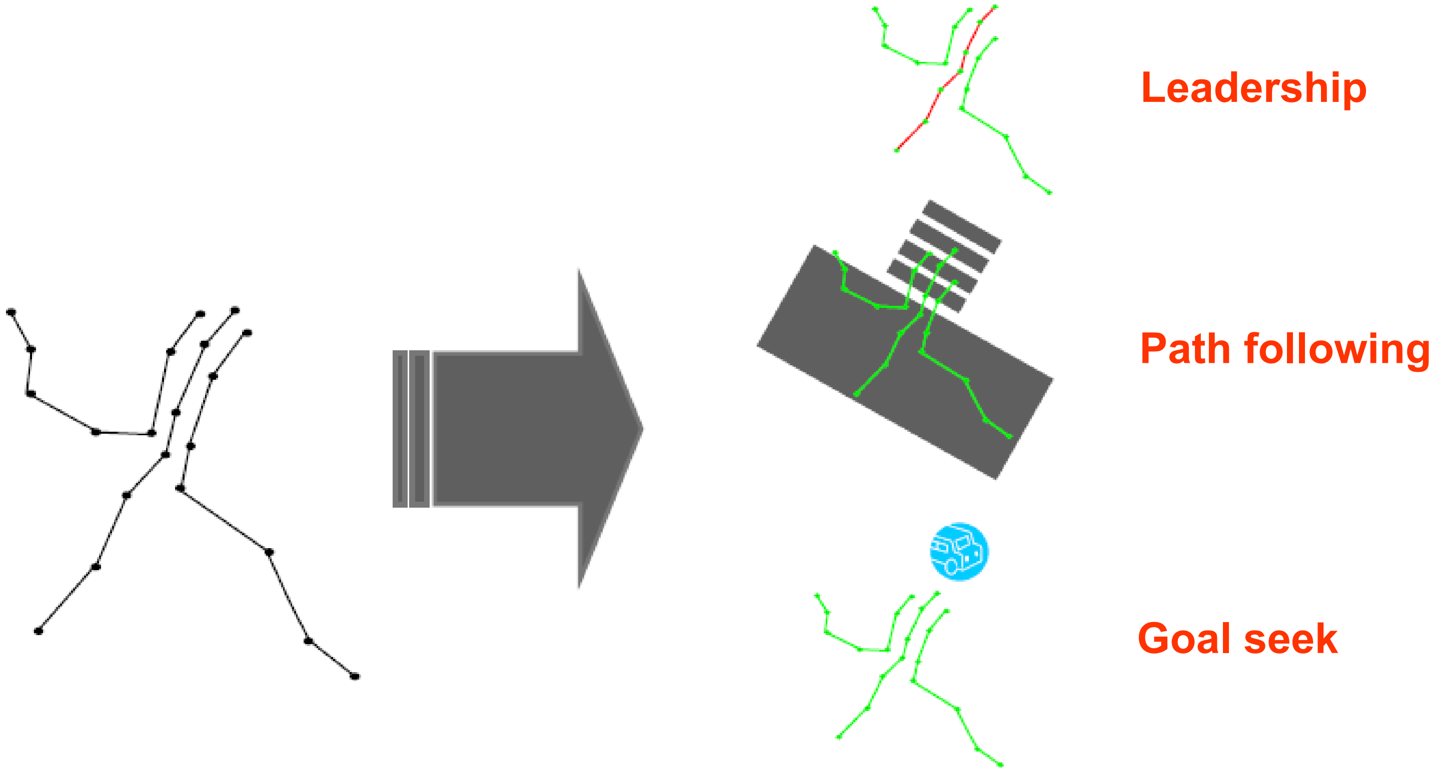
- **Prototype/Demo**
- **Socialization**: location-aware social network, …
- **Personalization**: route planning, spatial preference queries, …
- **Entertainment**: location-based games, …
- **Recommendation**: trip planning, location-based recommendation, …
- **Security**: privacy in LBS, …
- **Services**: location-based web search, trajectory data management, spatial keywords search, location prediction, …

Location-Based Services (LBS)



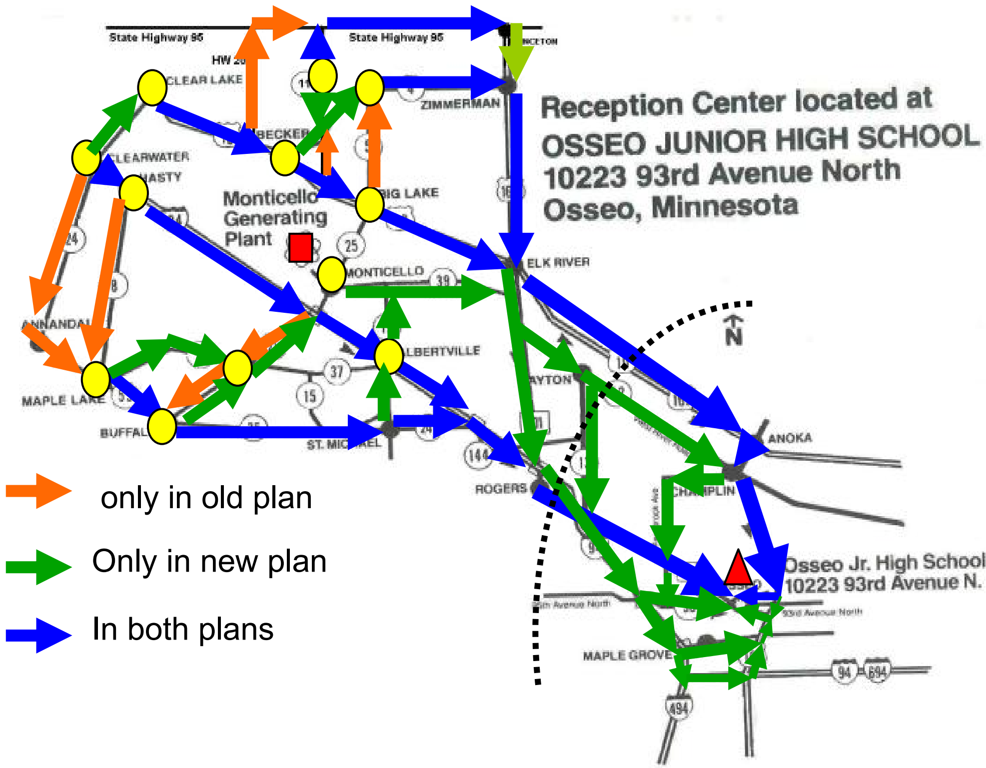
Research Issues (Cont.)

- Socialization
  - Location-aware social networks (a.k.a. Geo-social networks), e.g., foursquare, scvngr, etc.
  - …



Research Issues

- Personalization
  - Route planning, which is to retrieve paths or routes, preferably optimal ones and in real-time, from sources to destinations.
  - Spatial preference queries
  - …

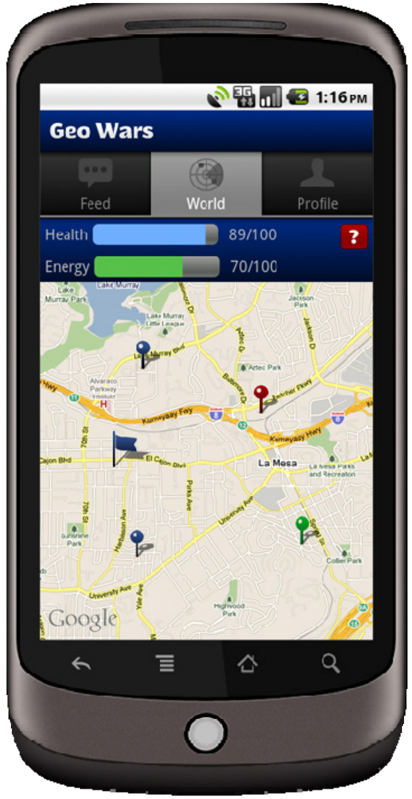


Research Issues (Cont.)

- Recommendation
  - Trip planning: Given a starting location, a destination, and arbitrary points of interest, the trip planning query finds the best possible trip.
  - Location-based recommendation
  - …

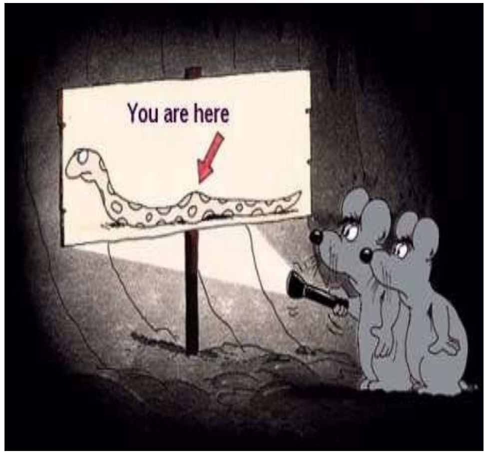
Research Issues (Cont.)

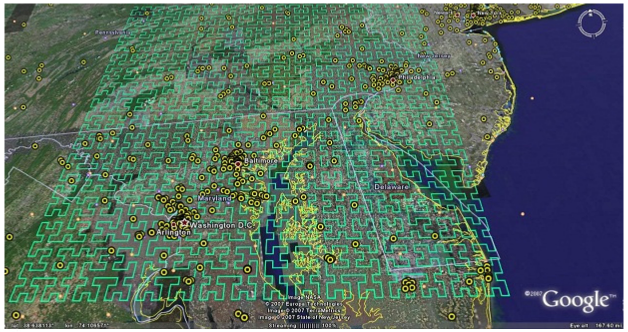
- Entertainment
  - Location-based games, e.g., BotFighter, Swordfish, My Groves, Geo Wars, etc.
  - CoSpace gaming
  - …

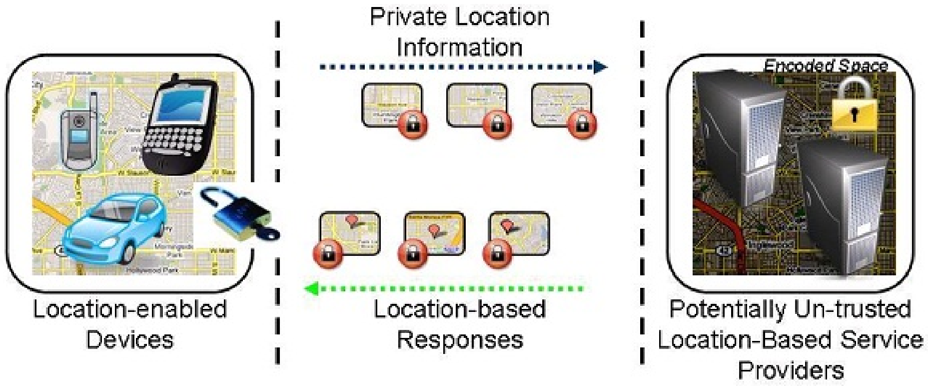


Research Issues (Cont.)

- Security
  - Privacy in location-based services
  - …

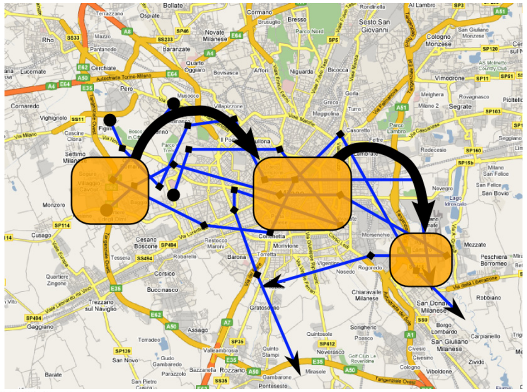


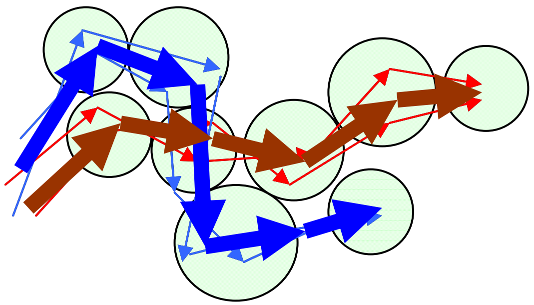


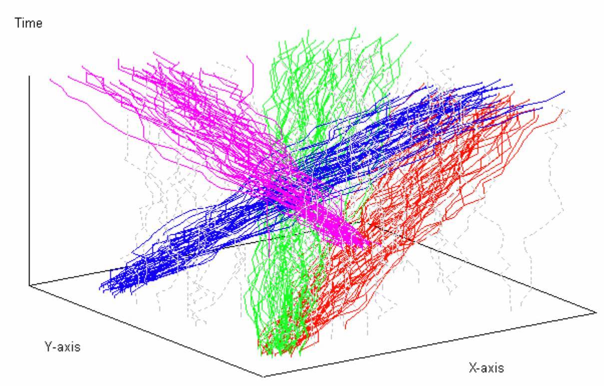


Research Issues (Cont.)

- Services
  - Location-based web search
  - Trajectory data management
  - Spatial keywords search
  - Location prediction
  - Novel queries for LBS
  - Spatial-aware queries on the WWW (e.g., Shortest/fastest/practice paths, etc.)
  - Uncertain/Incomplete Geo-spatial data management
  - …

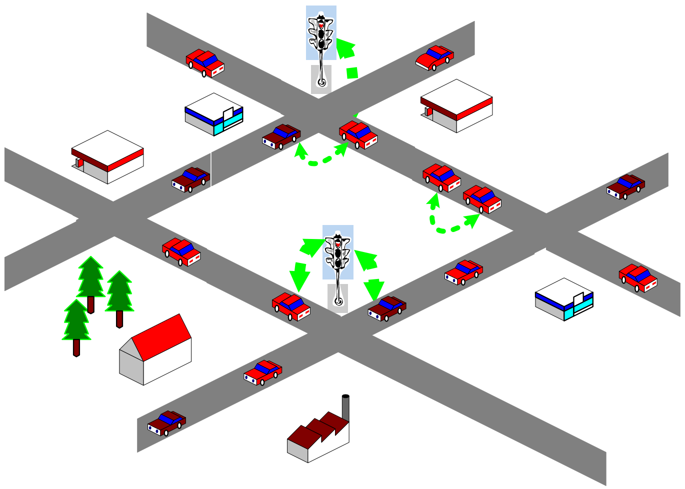


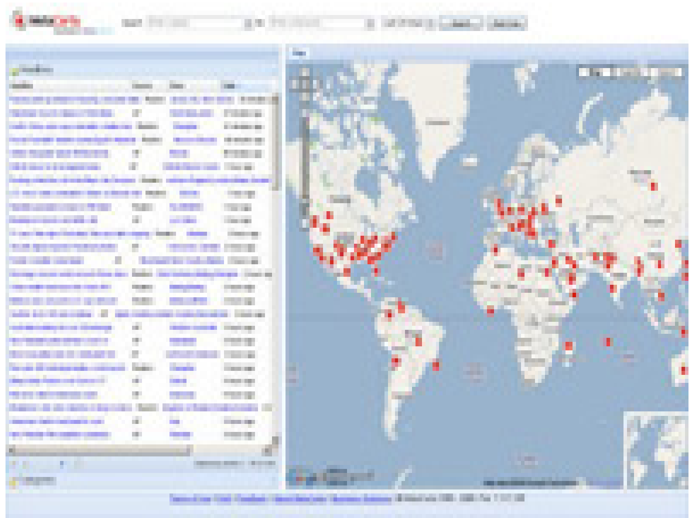


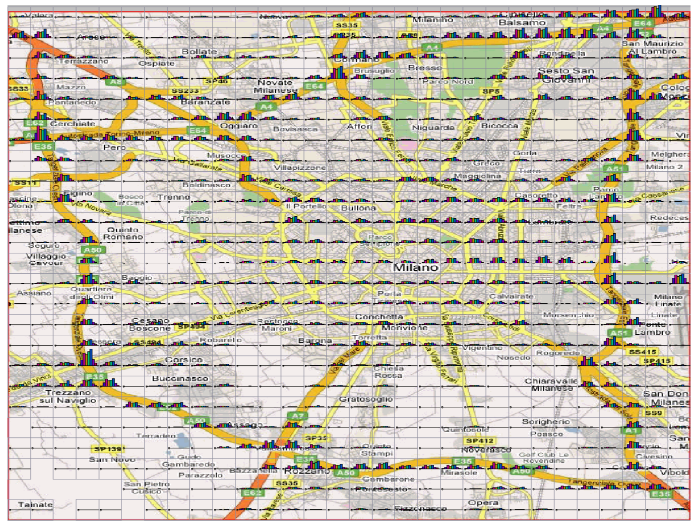


Research Issues (Cont.)

- Prototype/Demo
  - Intelligent transportation system
  - Spatial-aware retrieval engine
  - Geo-social network system
  - Trajectory processing system
  - …

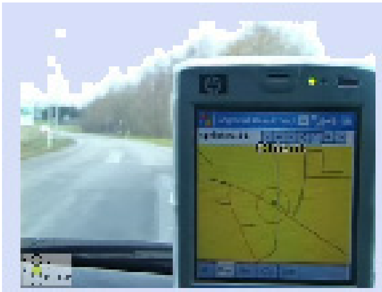






Existing Prototype 1: Streamspin

- Vision
  - To create data management technology that enables sites that are for mobile services what Flickr is for photos and YouTube is for video.
- Challenges
  - Enable easy mobile service creation
  - Enable service sharing with support for community concepts
  - An open, extensible, and scalable service delivery infrastructure
- The streamspin project maintains an evolving platform that aims to serve as a testbed for exploring solutions to these challenges.
- Streamspin Demo
- More details can be found **http://www.cs.aau.dk/~rw/streamspin/index.html**



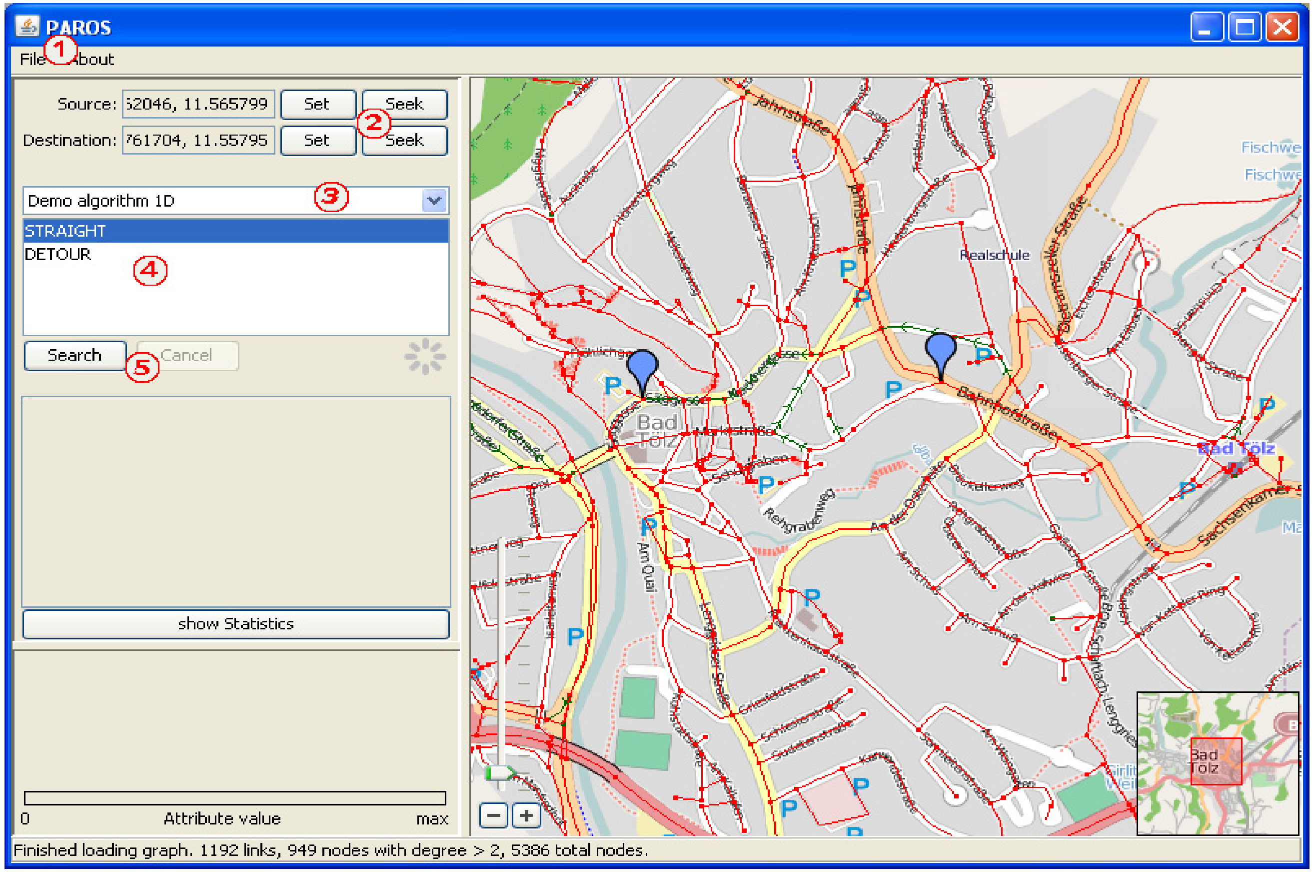
Existing Prototype 2: PAROS

- Paros is a Java based, open source program that allows an easy integration of route search algorithms (e.g., Dijkstra). Using paros, you can easily write new algorithms, test them on real data and visualize the results without having to deal with GUI programming.

- Purpose:
  - For research: test and graphically verify your graph algorithms on real data from OpenStreetMap
  - For research & teaching: a framework you can give to students which should get in touch with graph search but should not be delayed by GUI programming
  - For everyone else, if you just want to play around with route search

Existing Prototype 2: PAROS (Cont.)

- More details can be found http://www.dbs.informatik.uni-muenchen.de/cms/Project_PAROS



Outline

Conclusions

- Data on today's scales require scientific and computational intelligence.

- Big Data is a challenge and an opportunity for us.

- Big Data opens the door to a new approach to engaging customers and making decisions.

Q & A

Your questions and suggestions are expected for me.

Thanks a lot!