

NExT Search Center

What is NExt??

Next Generation Serch?

The Next big thing??

Tat-Seng Chua

School *of* Computing

National University of Singapore

NExT Search Center

(NUS-Tsinghua
 \
 /
Extreme Search Center)

清华大学 - 新加坡国立大学
下一代搜索联合研究中心

Tat-Seng Chua & Maosong Sun
NUS/Tsinghua

The New Information Age

- The Internet has revolutionized the way information is created, disseminated and consumed
 - Mixture of info available has changed from purely text, to include mm media, and live data
 - Emergence of huge amount of end-user generated data, tweets, blogs & managed systems

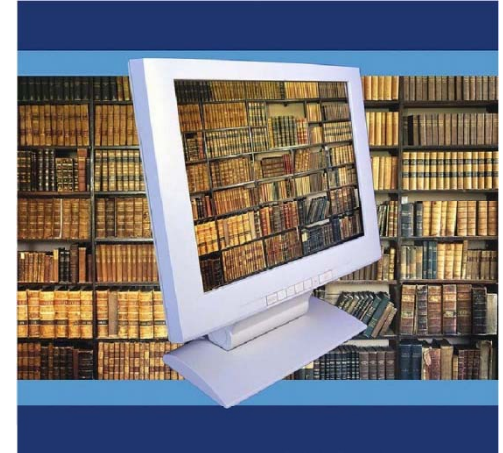


- Greater connectivity leads to huge amount of live info
- WWW has also rapidly gone mobile, permitting access from anywhere



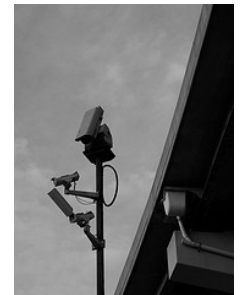
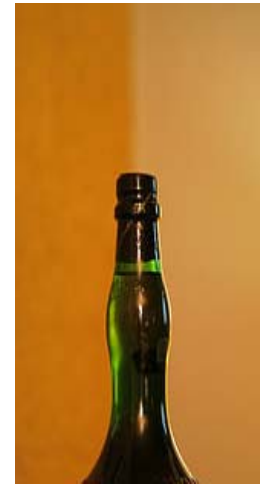
How Big is Internet?

- <http://worldwidewebsite.com> suggest a ~22.34 billion indexed pages (Sep. 2007)
- Studies claimed that the **deep (un-indexed) web** is ~500 times larger than the indexed
 - This may not include the huge number of (real-time) forum postings
- MM contents increasing at exponential rate
 - 31 million hours of video were produced each year (2006)
 - Over 65% of Web traffic is on mm contents (2007)
- *Like deep Web, the amount of info available in **live Web** is unknown*

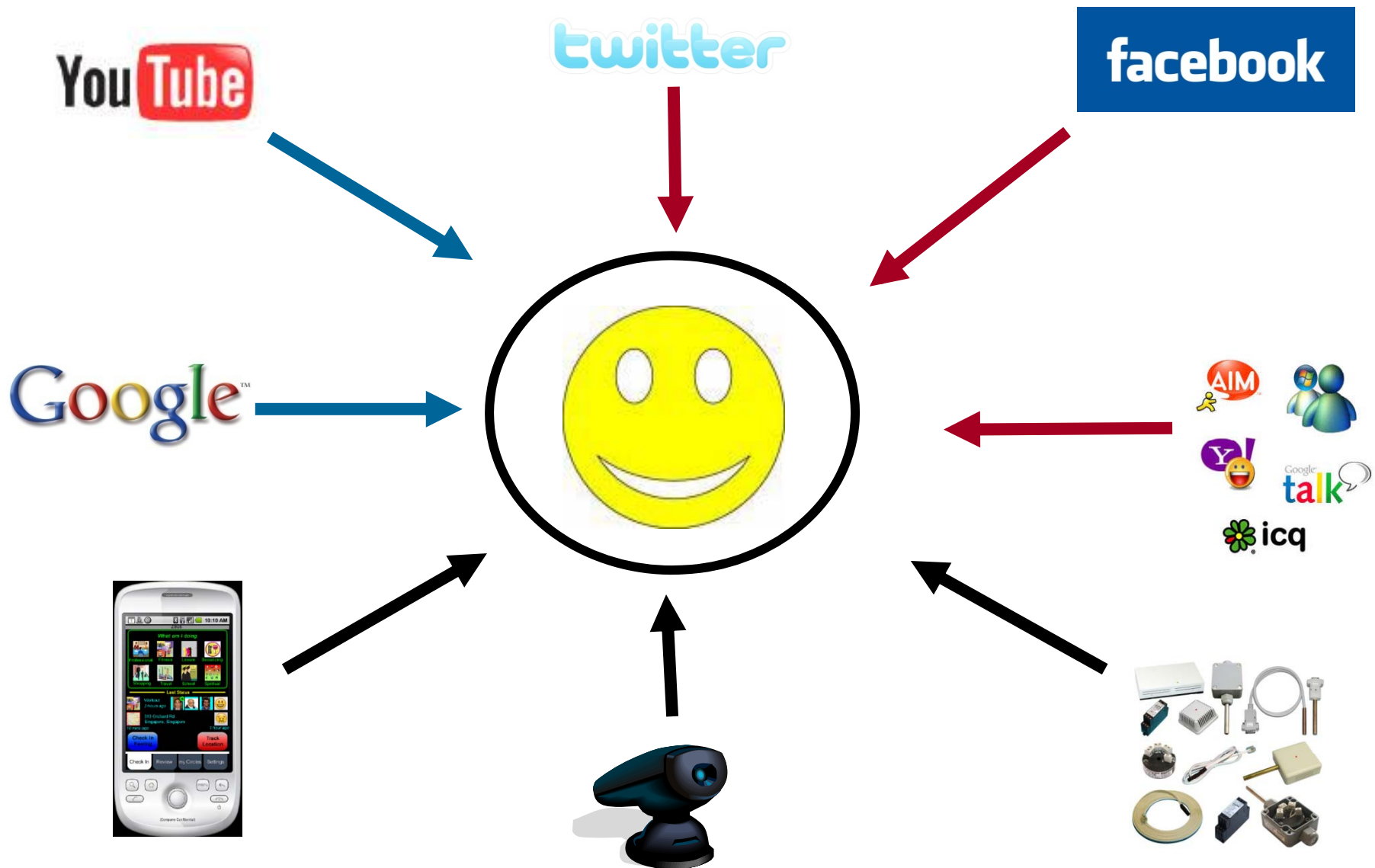


Information Bottleneck?

- Users retrieve text & mm info routinely
- But other than text, it is hard to find non-textual media..
 - Semantic gaps - How to find media entities when there is no text annotation?
- What about tracking of continuous live info (streams)?
 - This includes live sensors (cameras & others)?
What events are happening around the region?
 - Forums & human tweets: What are the latest hot topics and the sentiment of discussions?
 - Mobile phones: Their locations & check-ins?
 - What are their inter-relations??



Multi-Sensor Info-Rich Environment



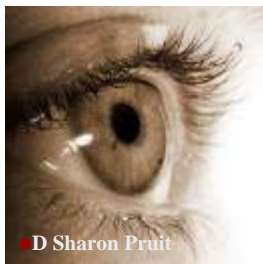
Our Information Shadow

“Everything and everyone in the world casts an
‘Information Shadow’....

Increasingly, the web is the World.

*Our cameras, our microphones, are becoming the eyes
and ears of the Web, our motion sensors, proximity
sensors its proprioception, GPS its sense of
location....”*

O’Reilly , Battelle 2009



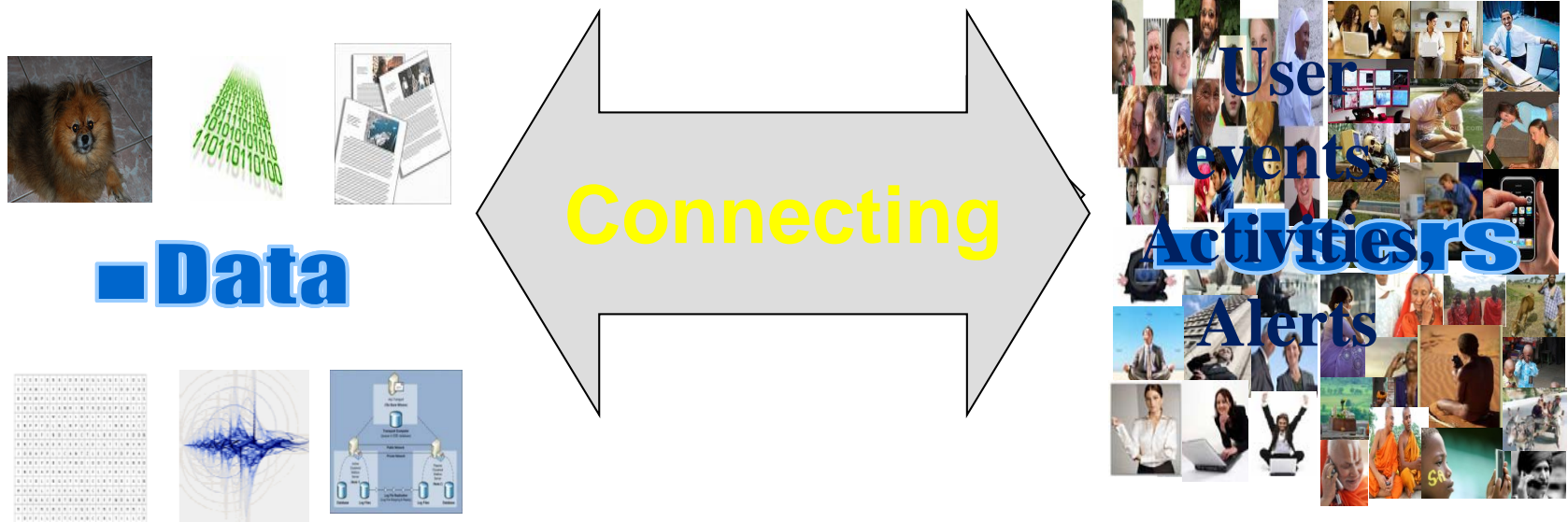
Aims of NExT Center

- To find and extract meanings from millions of real-time data streams

data → Information

- To aggregate info to realize a SMART environment

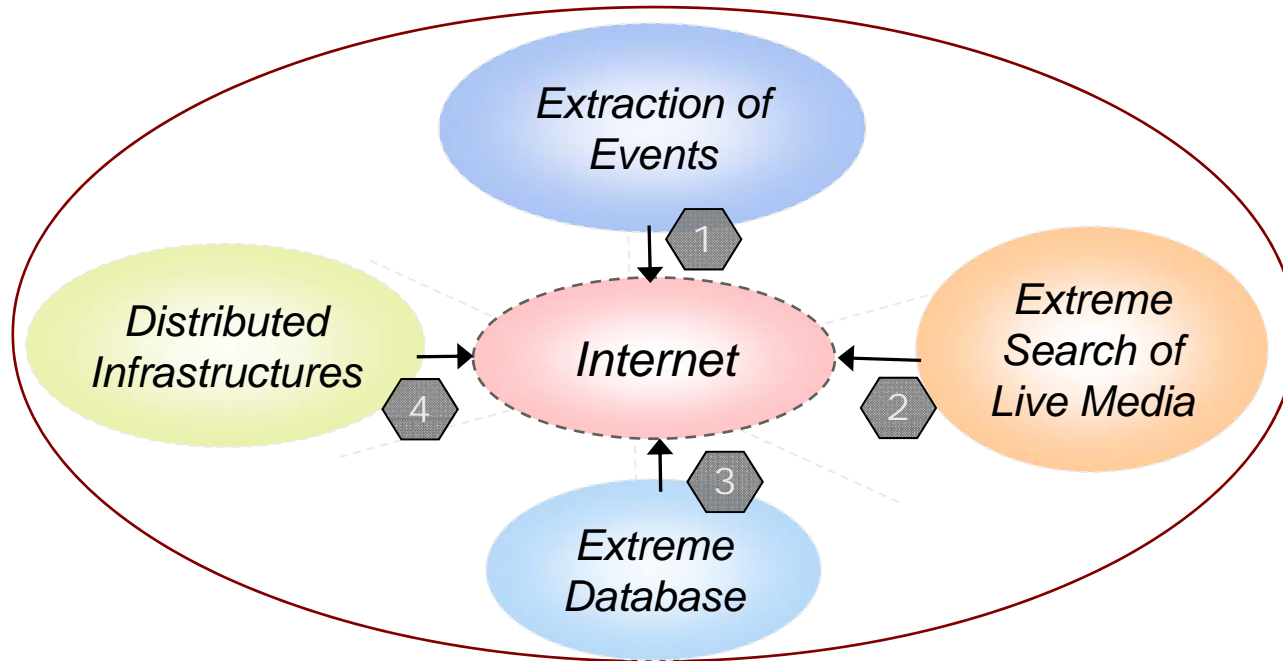
Information → Situational Info → Users



NExT Center

- Research into Internet-Scale Extreme Search for millions of Real-time Data and Sensors to help realize a Smart Environment for Users
 - Aggregate, track and predict events in a location
 - Inform users of latest happenings at all time
 - Help users in their daily activities
- Joint research efforts between NUS & Tsinghua University
 - A 5-year, multi-million collaborative project supported by NRF in Singapore
 - With equivalent contributions from both Universities
 - Centers in both Singapore and Beijing
 - Involve over 50 professors/ researchers/ PhD students

Scope of the Center

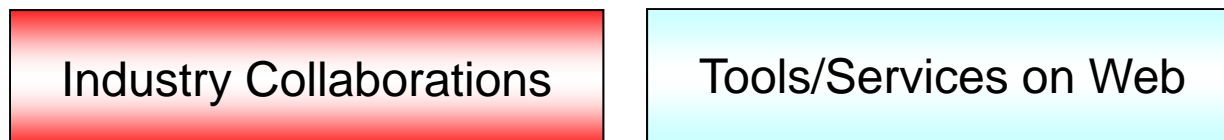


Support other IRC's



Work with Industry

Deploy technology



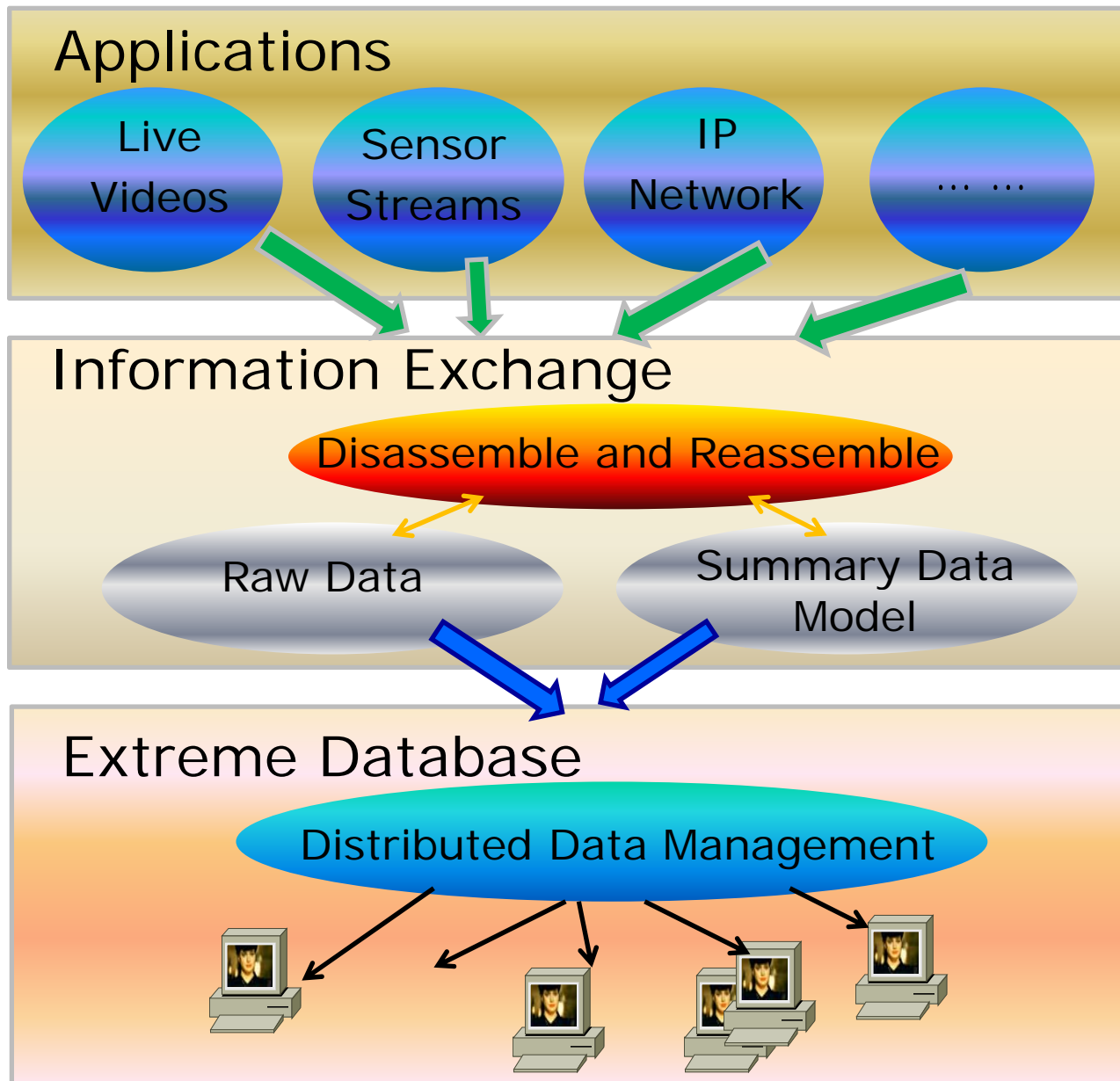
Scope of NExT Center

- Activities.. Conduct research on:
 - Live Media Search/Alerts/Push (Text and Sensor Data)
 - Aggregation of Cyber-physical Info to support Smart City
 - Distributed infrastructure and database support
 - Build multi-country infra-structures as testbed for research

Extreme Database

- Objectives:
 - **A new high performance distributed database** to manage and process live stream data
 - **Distributed Infrastructures** to store the stream data and indexes on the distributed environment
 - **Efficient algorithms and effective index structures** for query processing and data analysis on the stream data
 - **A summary data model** to bridge live stream data and database-style data

Database Architecture

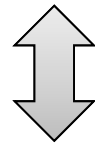


Research issues

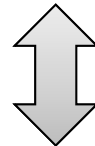
- High performance
- Real time response
- Summary Data model
 - Tuple
 - Vector
 - others?
- When and How to do
 - Disassembling
 - Reassembling
- Distributed Infrastructures
 - Overlay
 - Network (internet or intranet)
- Data Storage
- Index Structures
- Query Processing

Distributed Infrastructure

P2P Overlay, Cloud Computing,
Server Farms, Data Center



Point-to-Point Link, Multicast, Wireless Mesh,
Sensor Network, Delay Tolerant Network



Fiber, Ethernet, Cellular Network,
WiMAX, WiFi, ZigBee, Bluetooth, etc...

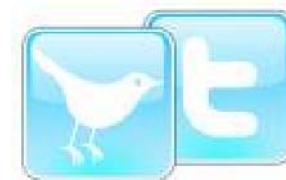


Live Forum Search

- Live search of semantic contents
 - Cover forum, tweeter, iQ, and possibly sms & e-mails
 - Could include device tweets too..
 - Issues of search, filter, or alert (different usage models)
 - Induce
 - In both English and Chinese events with sentiments

The screenshot shows the Yahoo! Answers website interface. At the top, there's a search bar with the text "YAHOO! ANSWERS" and a search button labeled "WEB SEARCH". Below the search bar, there are buttons for "ask.", "answer.", and "discover.". The search query "mp3 players" is entered in the search bar. The results page shows "Search Results" for "mp3 players" with 1 - 10 of 82,566 results. There are several sponsored results for MP3 players, including "Unlimited Online Music" from Slacker.com, "Cheap mp3 player" from eBay.com, and "Official Apple® Store". A user question is visible: "I am planning to buy an MP3 player this weekend. I know, I know, I'm a little... I personally think that MP3 players are better than iPods. Sandisk... Asked by Stacey K - 2 years ago - Music & Music Players - 3 Answers - Resolved Questions".

The screenshot shows the Baidu Knowledge (百度知道) website interface. At the top, there's a search bar with the text "Baidu 知道" and a search button labeled "搜索答案". Below the search bar, there are buttons for "我要提问" and "帮助". The search results page shows "问题分类" (Question Categories) and "精彩推荐" (Featured Recommendations). The main question is "中国航天计划分为哪三个步骤?" (What are the three steps of the Chinese space program?). The answer is "中国航天计划分为哪三个步骤? (陈空) 人类已研制成功的载人航天器有哪些? (陈空) 航天员在太空中每天要消耗多少粮食? (陈空) 航天员在太空出舱后怎样行走? (陈空) 在太空停留时间最长的人是谁? (陈空) [更多精彩推荐]". There are also "知识掌门人" (Knowledge Masters) and "待解决的问题" (Questions to be solved) sections.



Live Media Search

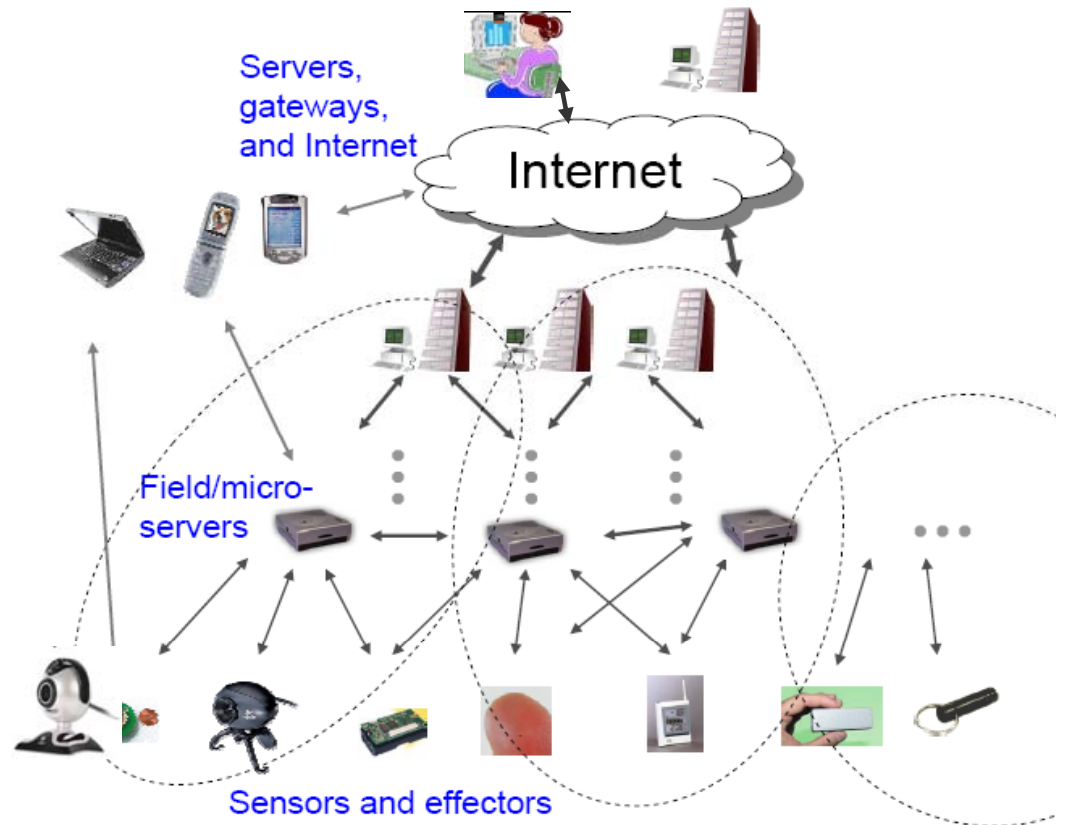
- Research Focuses

At the Micro-level:

- Multi-sensor analytics
- Human (groups) behavioral analysis & recognition

At the Macro-level:

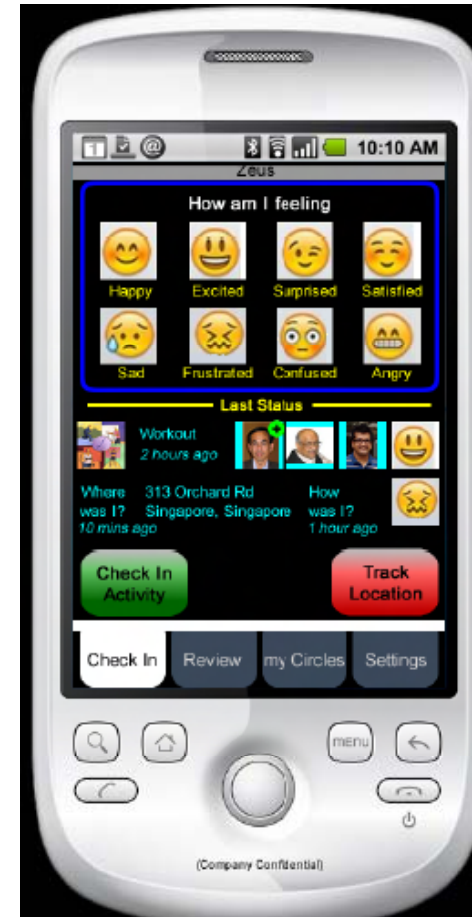
- Fusion of multi-sensor info to infer crowd behavior & region wide trends



- Mobile device analytics – at user level, friends-level...
- Aggregation of multisource info
- Again the issues of search, filter, or alert

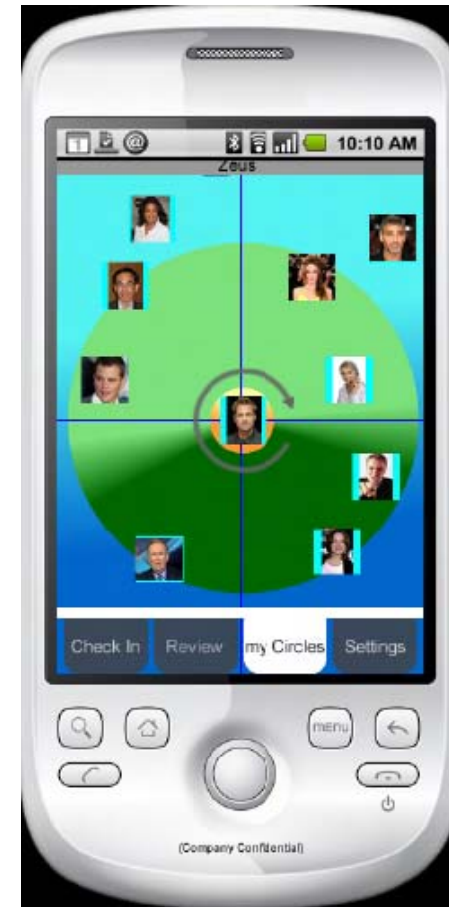
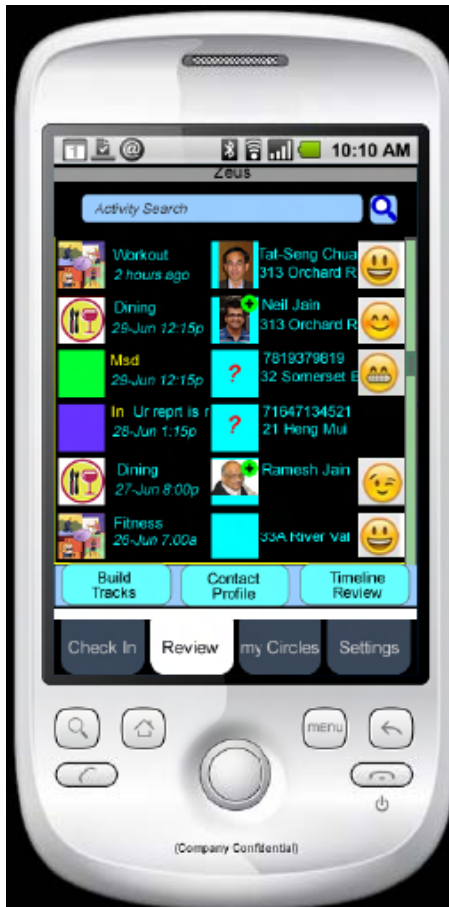
Mobile Phone Analytics -1

- Logging of mobile phones activities
 - Tracks on: Location, e-mail, phone, sms, photos etc
 - Help users to reflect on their lives



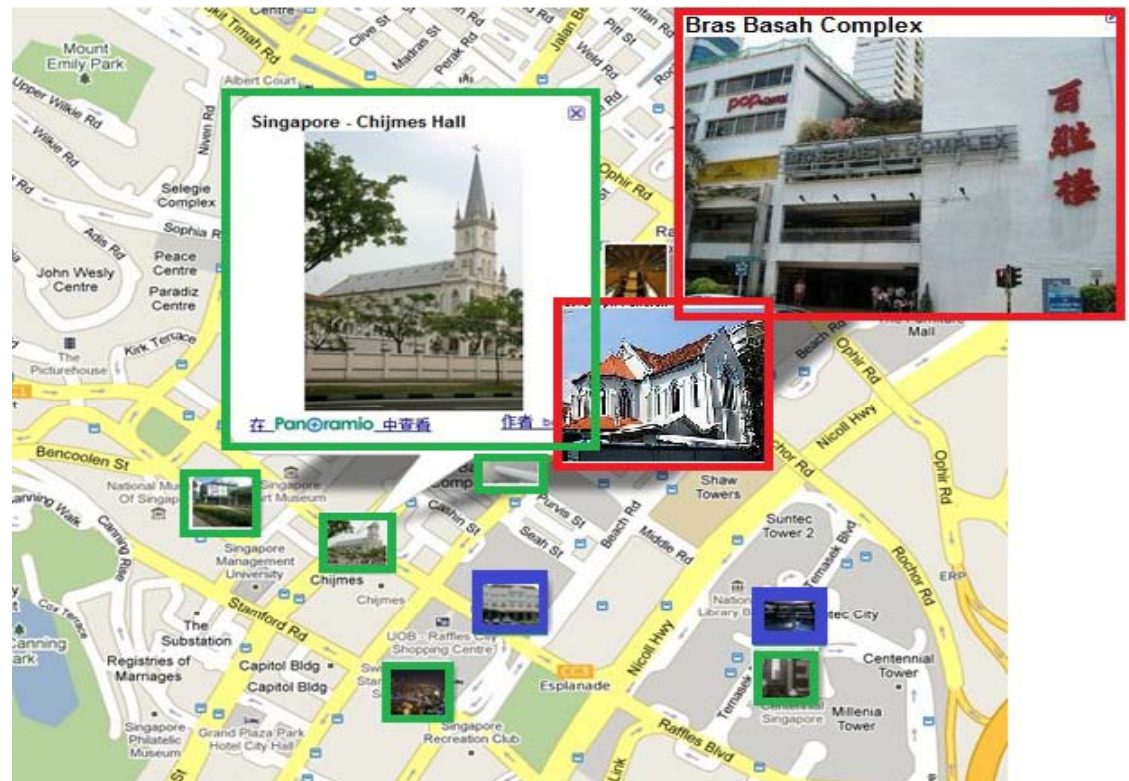
Mobile Phone Analytics -2

- Logging of mobile phones activities
 - Tracks on: Location, e-mail, phone, sms, photos etc
 - Help users to reflect on their lives



Mobile Phone Analytics -3

- Capturing/logging of mobile phone activities
 - Tracks on: Location, e-mail, phone, sms, photos etc
- To provide better smart environment for users
 - Help to induce relations between users
 - Leveraging on individual, friends' & "others" knowledge and preferences
 - Provide more informed environment



Smart Environment...

- Event detection, tracking, prediction and reconstruction
 - Detect situations
 - Create list of interesting events
- Infrastructures
 - Again the issues of search, filter, or alert
 - Filter – in real-time
 - Create event indices
 - Create alerts for relevant recipients
- Know your environment
 - Interesting happenings
 - Friends in vicinity
 - Find my ways around, getting things I want
 - Review my activities

Summary

- The work is just started...
- Key focus is leveraging available (live) information resources to help us understand our environment better
 - Data either from social activities or just faithful recording of events
- Evolving as we understand what users and others need
- Welcome comments and collaboration

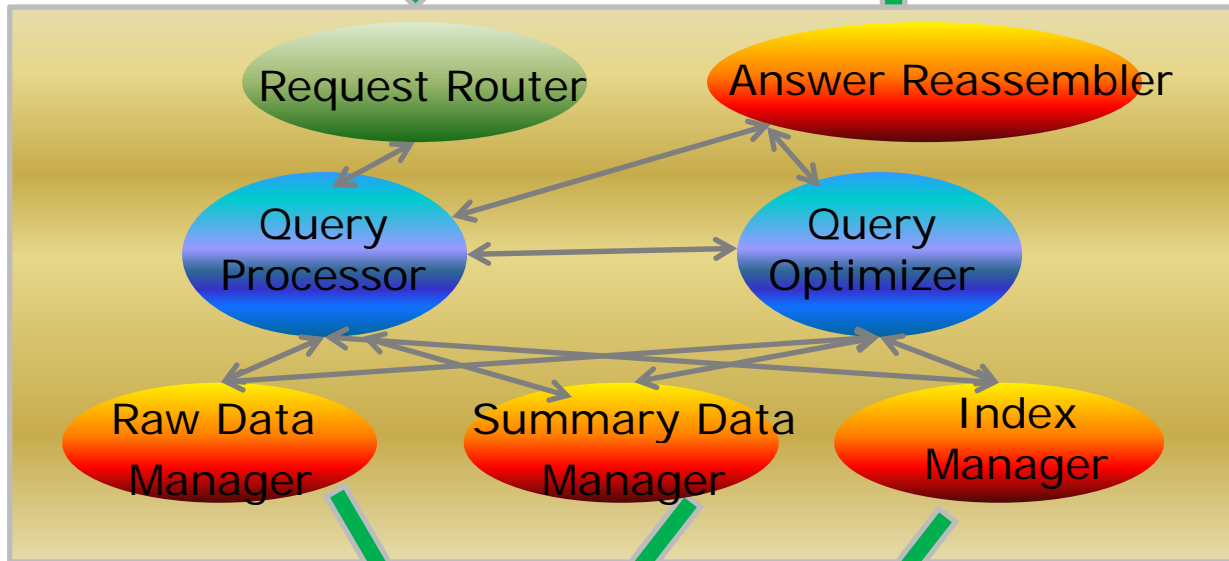
Thank You

Project 4: Extreme Database -3

Request from Applications

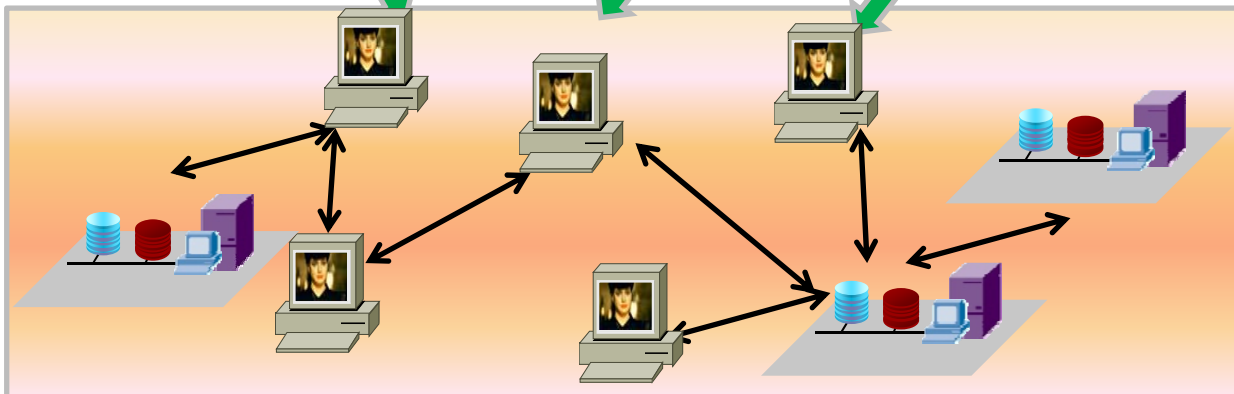


Answers for Applications



Research issues

- Summary Data model
- Global Query Processor
- Global Query Optimizer
- Routing
- Raw Data Management
- Summary Data Management
- Distributed Infrastructures
 - Overlay
 - Network (Internet or Intranet)
- Data Storage
- Index Storage
- Local Query Processor



Distributed Infrastructure -2

- Design scalable systems for distributed data collection, storage, processing, and querying.
- Design a programming abstraction for controlling and managing distributed multimedia sensors for live search.
- Design the system carefully to deal with various types of failures and make the system robust against security attacks.
- Develop adaptation algorithm for multimedia streams that is aware of the queries being run.
- **Deliverable:** a scalable media sensing and streaming infrastructure for extreme search