Pervasive platforms, data explosions, & now it's all about the apps: the 4th paradigm of Science

Gordon Bell
Microsoft Research
Research.microsoft.com/~gbell

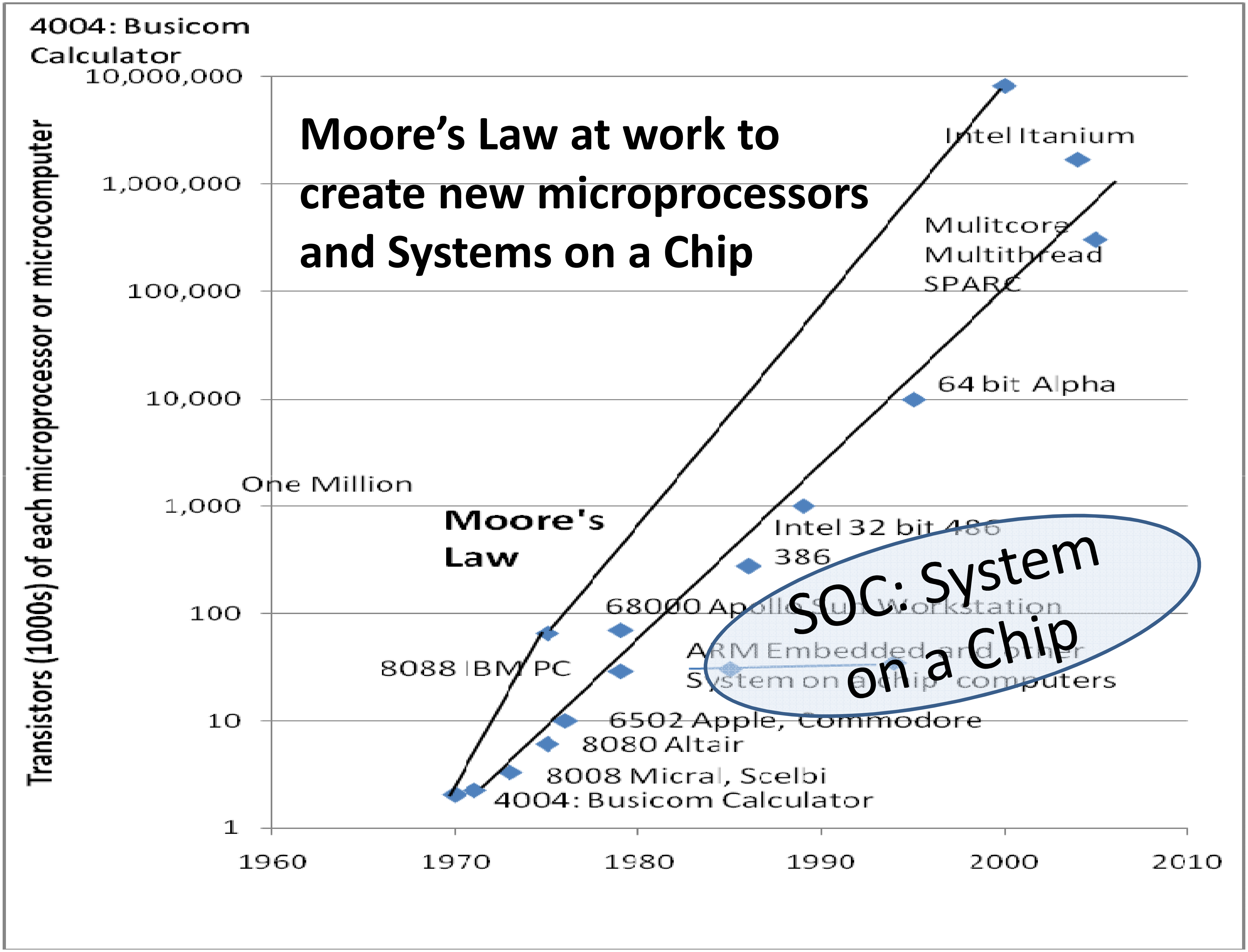
Agenda

1. Hardware (storage, networks, sensors) …
   more than we can ever imagine brought about by storage explosion, wirelessness, cost for ubiquitous computing … everything is smart.
2. The exploding amount of data from every networked "thing" and every computed model
   Challenge: capture, holding and making sense

   The 4th paradigm of Science…
3. Examples: Environmental server, SkyServer>Worldwide Telescope, Health,

Hardware

- Moore's Law
  - bits/chip,
  - bytes/platter and data explosion
  - radios/chip
- Bell's Law (new structures)
  - Wireless Sensor Nets


4004: Busicom Calculator
Moore's Law at work to create new microprocessors and Systems on a Chip
Transistors (1000s) of each microprocessor or microcomputer
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
1960
1970
1980
1990
2000
2010
Intel Itanium
Mulitcore Multithread SPARC
64 bit Alpha
One Million
Moore's Law
Intel 32 bit 486
386
68000 Apollo Sun Workstation
ARM Embedded and other System on a chip computers
8088 IBM PC
6502 Apple, Commodore
8080 Altair
8008 Micral, Scelbi
4004: Busicom Calculator
SOC: System on a Chip

Disks sizes have double every year

1998: 1 Gbyte

2008: 1 Terabyte… more than enough for human store

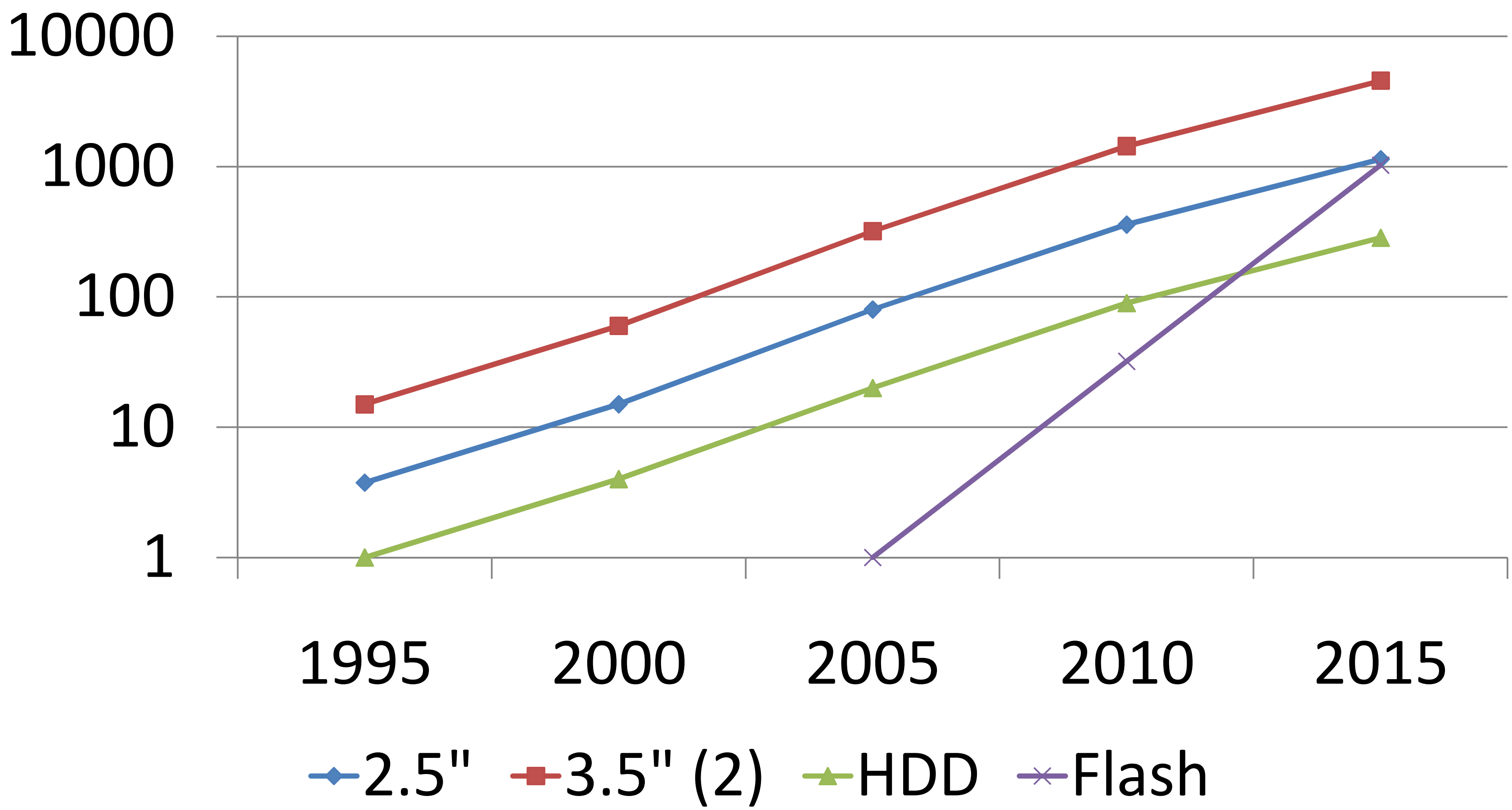
2018: 10, 100, or 1,000 Terabytes?

Today's large servers are 10 Petabytes.

Similar advances in flash storage chips = 8 Gbytes.

128 Gbyte solid state personal computers

Storage devices(time) for portables, PCs, and the cloud



IDC Survey: Exploding Digital Universe

- The digital universe in 2007 — at 2.25 x 10^{21} bits (281 exabytes = 281 billion gigabytes = 281 million terabytes)
- The greater estimate is from faster growth in cameras, digital TV, etc. and now understanding the information replication.
- By 2011, the digital universe will be 10 times its 2006 size
- The amount of information created, captured, or replicated exceeded available storage in 2007.
- By 2011, half of the digital universe will not have a permanent home i.e. will be homeless

Data sources and sinks

Devices and Applications Tracked

Image Capture/Creation
- High-end cameras
- Digital cameras
- Camcorders
- Camera phones
- Webcams
- Surveillance
- Scanners
- Multifunction peripherals
- OCR
- Barcode readers
- Medical imaging
- Digital TV
- Digitized movies and video
- Special effects
- Graphics workstations

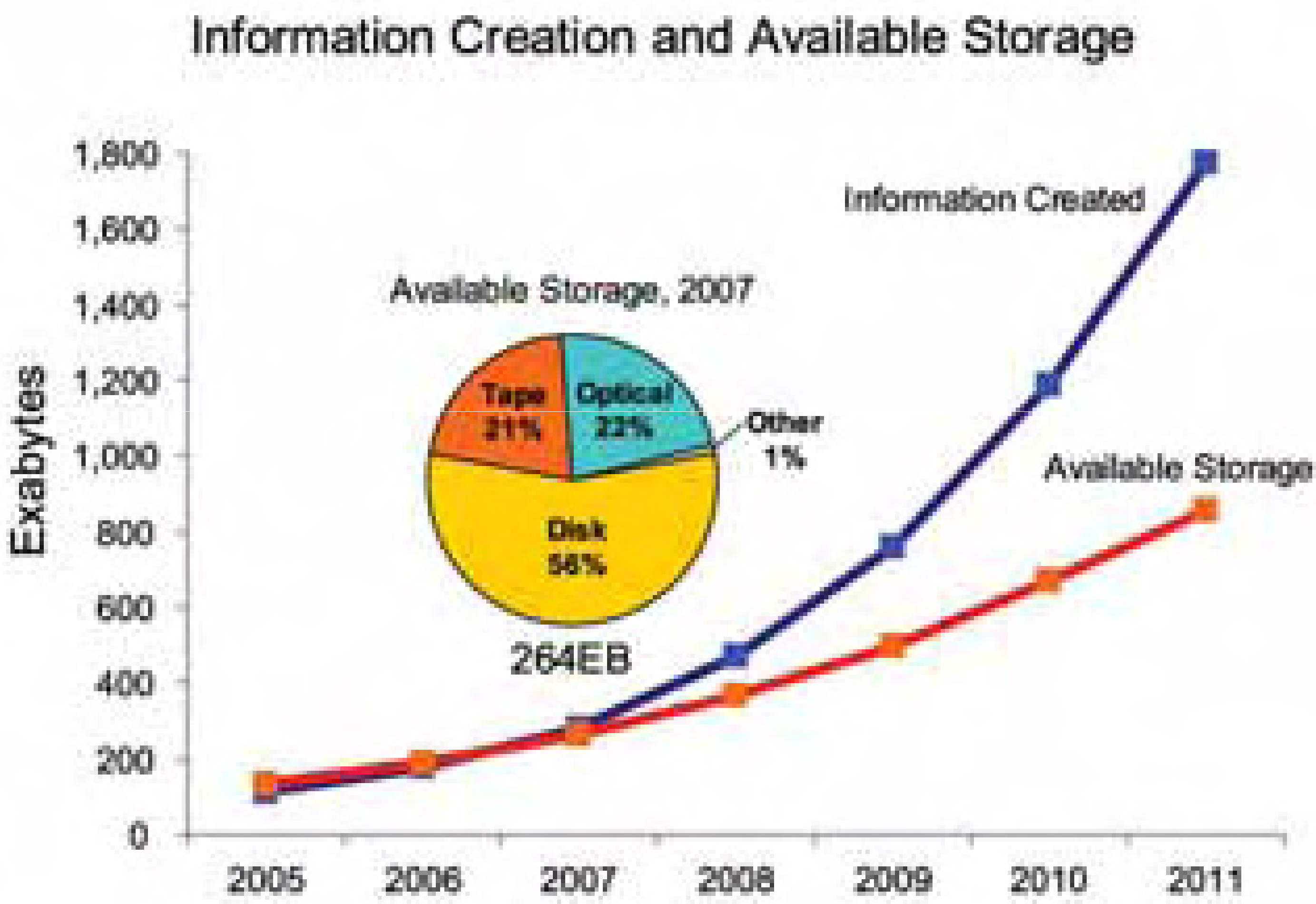
Digital Voice Capture
- Landline telephony
- Voice over IP
- Mobile phones

Data Creation
- PC applications
  - Database
  - Office applications
  - Email
  - Video/teleconference
  - IM
  - Other
- Smart handhelds
- Server workloads
  - Business processing
  - Decision support
  - Collaborative
  - Application development
  - IT infrastructure
  - Web infrastructure
  - Technical
  - Other
- Terminals, ATMs, kiosks, specialized computers
- Industrial machines/cars/toys
- RFID
- Sensors
- Smart cards
- Videogames
- MP3 players
- SMS
- GPS

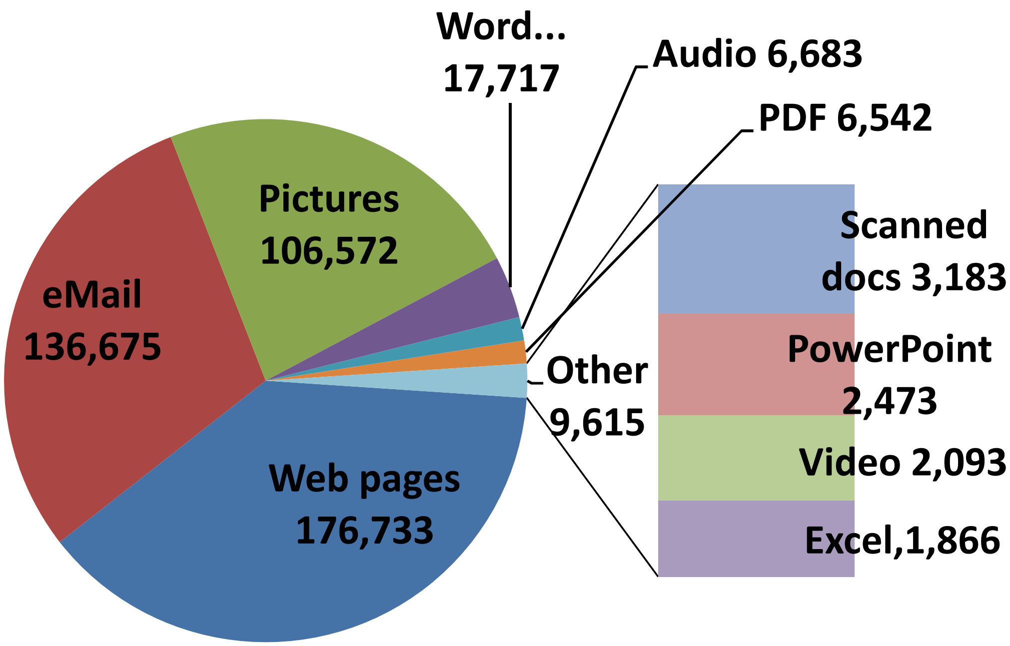
Data Storage
- HDD
- Optical
- Tape
- NV flash memory
- Memory

IDC Exploding Digital Universe 2008

IDC Whitepaper Diverse, exploding, digital universe 2008

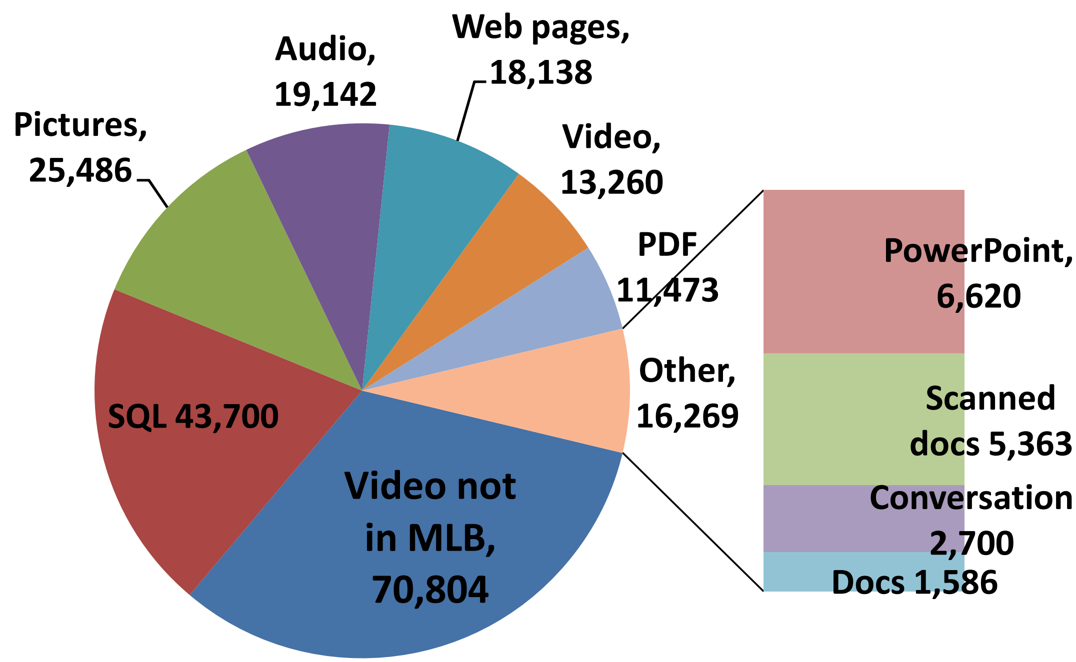


IDC Exploding Digital Universe 2008



461K items + 337 K pages

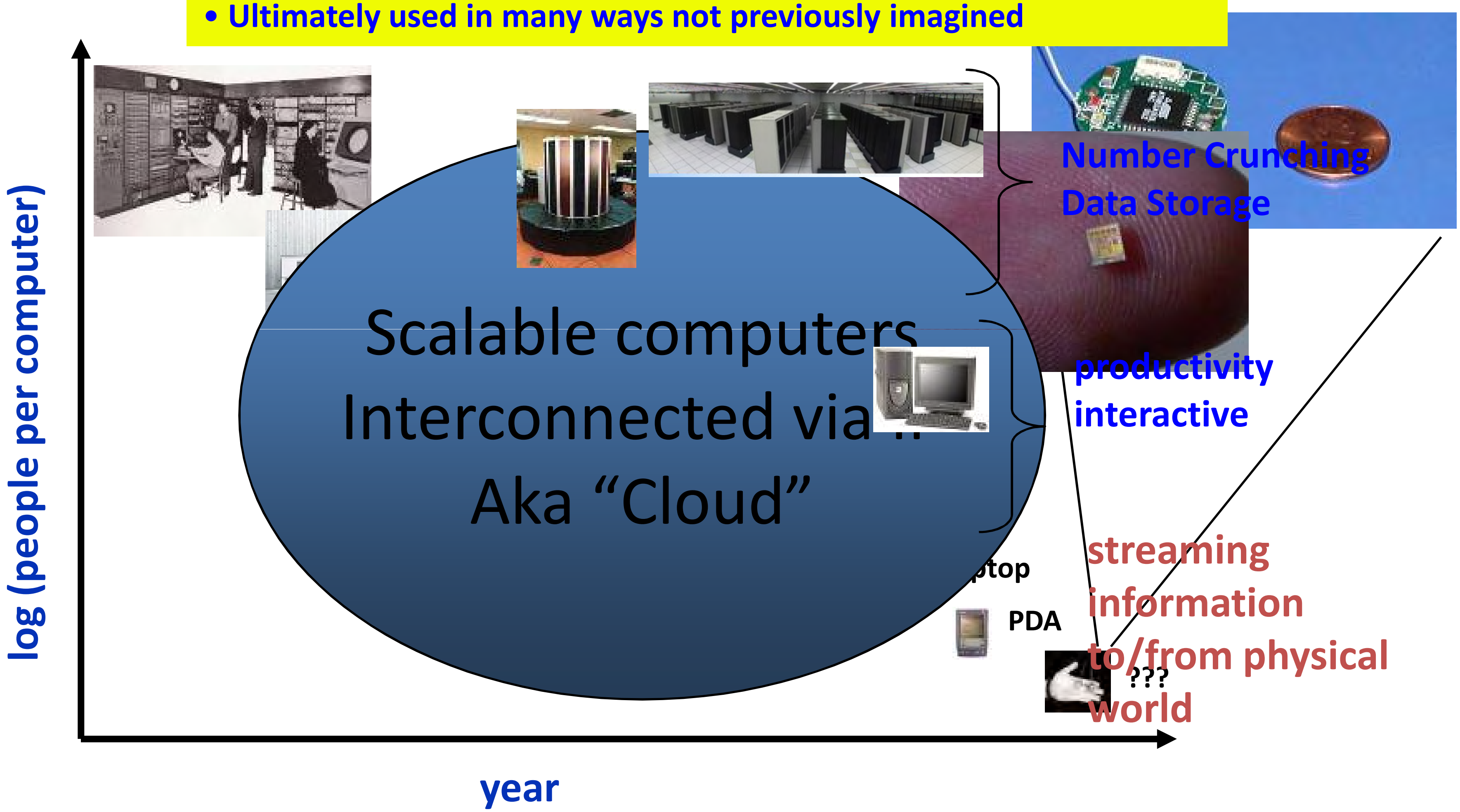
Bell's MyLifeBits
Files & items



File size 219 GBytes

Computer classes emerge over time

- Enabled by technological opportunities
- Smaller, more numerous and more intimately connected
- Ushers in a new kind of application
- Ultimately used in many ways not previously imagined



David Culler UC/Berkeley

Bell's Law

- Why computer classes form
- Requirements for classes
- What the classes are
- Predicts new classes

Bell's law of computer class formation

- New computer platforms emerge based on new chip, storage, and network evolution
  - It may come from research e.g. web, wireless sensor nets, or hardware evolution
- Computer classes consist of:
  - new platforms,
  - new networks, and
  - new interfaces i.e. cyberization ("world" → cyberspace)
- New classes enable and require
  - New apps and new content
  - *In this generation it will all be about managing the data*
- Each class evolves into a vertically disintegrated industry based on hardware & software standards

Bell's Law of Classes → New Industry Size x Price x Interface x App x Network

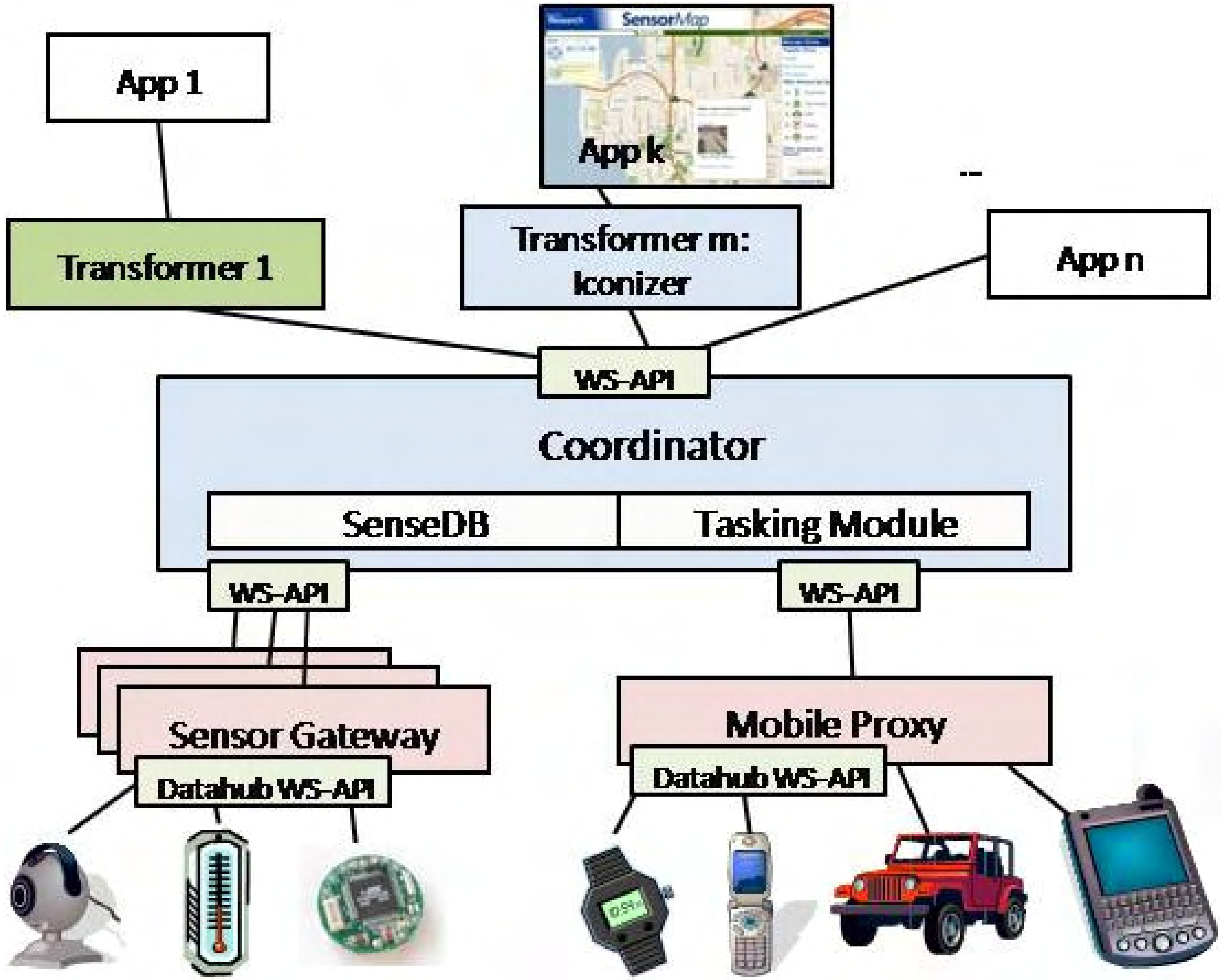
- **As of 2008, the computer classes included:**
  - mainframes (1950s and 1960s)
  - minicomputers (1970s)
  - personal computers and workstations evolving into a network enabled by Local Area Networking or Ethernet (1980s)
  - web browser client-server structures enabled by the Internet (1990s)
  - clusters aka clouds superseding mainframe, minis, & supers (>1995)
  - web services, e.g. Microsoft's .NET aka the Grid (2000s)
  - **small form-factor devices (SFF) such as cell phones and other cell phone sized devices (CPSD) c. 2000 e.g. BlackBerry iPod, > iPhone**
  - **Wireless Sensor Networks aka motes (c. >2005)**
    - <u>WSNs enable platforms, appliances, and peripherals</u>
- **Prediction: home & body area networks will form by 2010. Alternatively the platforms have already formed cf. Cellphone!**

Bell's law of computer class formation

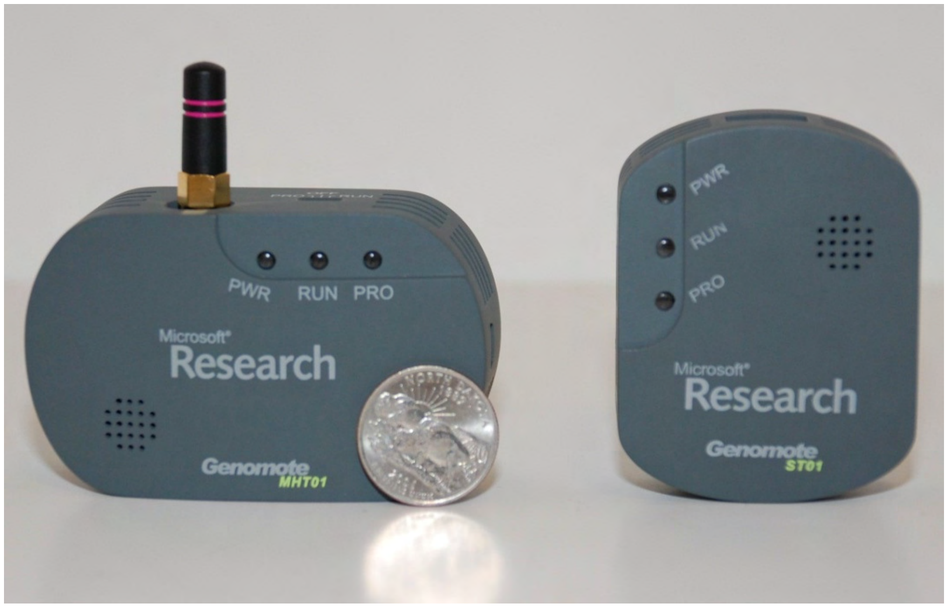
Microsoft Research SenseWeb: Wikipedia of Sensors

- Enable sharing of deployed *instrumentation* and *data* for communities of scientists and hobbyists
- Share sensors
  - Each deploys at small scale; everyone can use shared instrumentation
  - Larger spatio-temporal coverage than any single system
  - Costs amortized over multiple experiments
- Share Data
  - Same datasets used for multiple analyses
- SensorMap as the portal
  - http://atom.research.microsoft.com/sensormap

Microsoft Senseweb for Sensornet

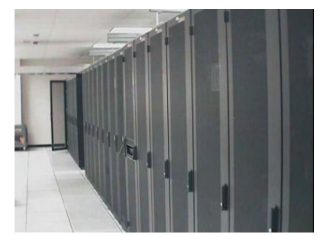


MSR Mote



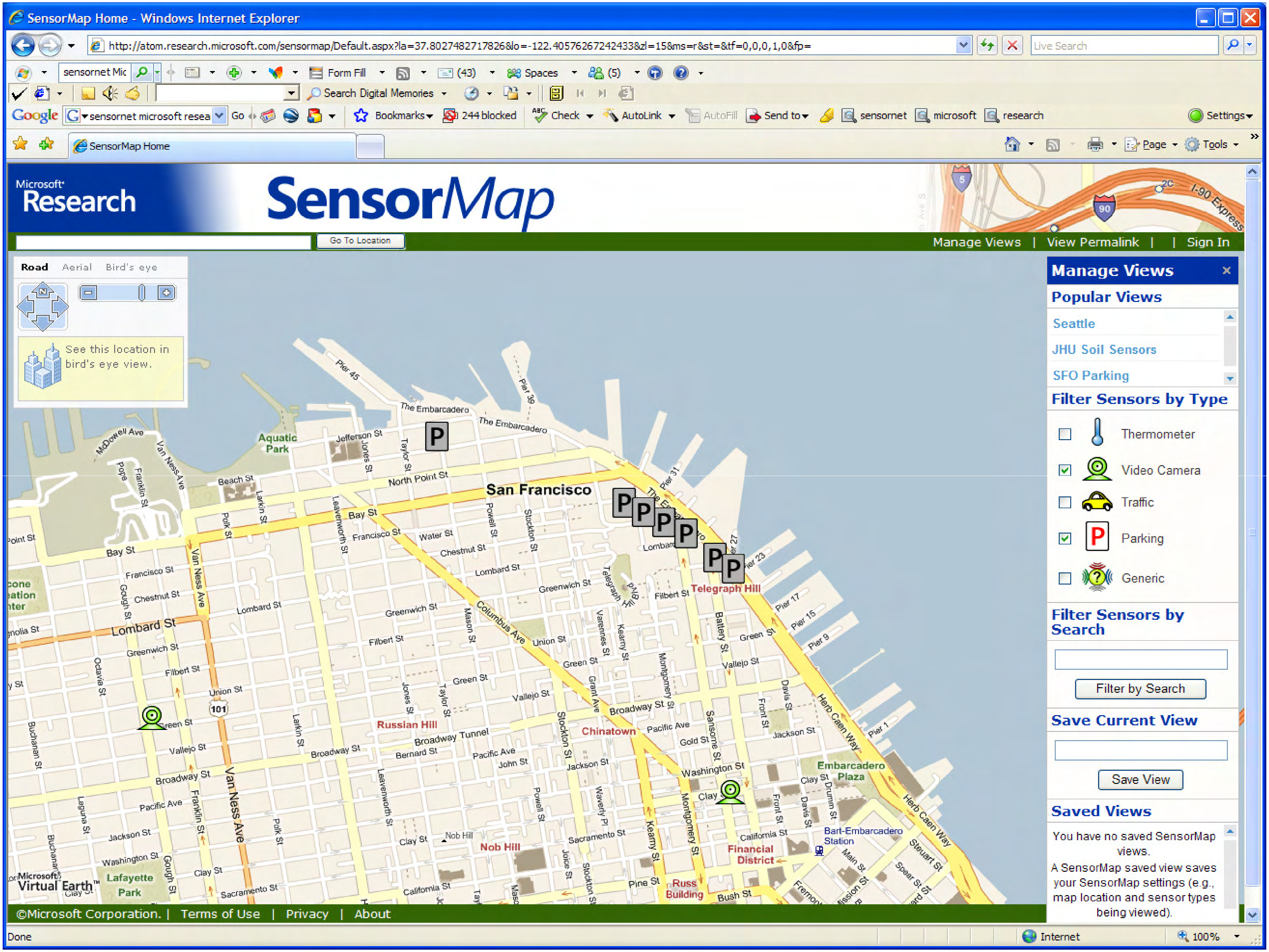
Wireless sensor: 6MHz processor, 10K RAM, 48K ROM, 802.15.4 radio, temperature/humidity sensing

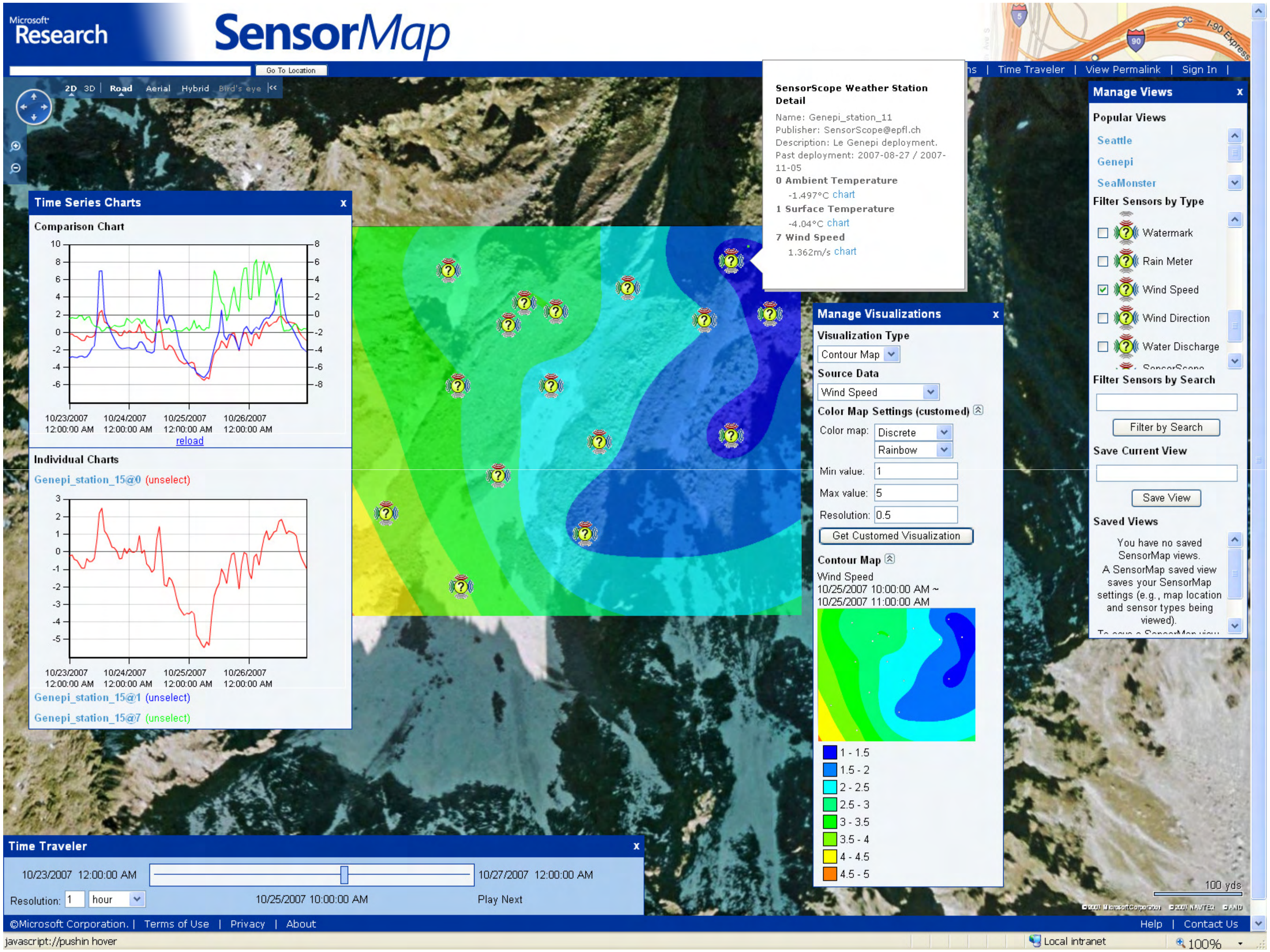
- SenseWeb/SensorMap: wide-area sensor data sharing
- Tiny Web Service: simplify interfaces with other devices
- DC Genome and Green99: save energy in data center, home, and office



Energy


Microsoft Research
SensorMap
Go To Location
Manage Views | View Permalink | | Sign In
Road Aerial Bird's eye
See this location in bird's eye view.
Video Camera Sensor Detail
Name: bellevuecam012
Live Data:
Open In New Window
Description: Address
Manage Views
Popular Views
Seattle
JHU Soil Sensors
SFO Parking
Filter Sensors by Type
Thermometer
Video Camera
Traffic
Parking
Generic
Filter Sensors by Search
Filter by Search
Save Current View
Save View
Saved Views
You have no saved SensorMap views.
A SensorMap saved view saves your SensorMap settings (e.g., map location and sensor types being viewed).
Microsoft Virtual Earth
©Microsoft Corporation. | Terms of Use | Privacy | About
Done
Internet
100%


SensorMap Home - Windows Internet Explorer
http://atom.research.microsoft.com/sensormap/Default.aspx?la=37.80274827178268&lo=-122.40576267242433&zl=15&ms=r&st=&tf=0,0,0,1,0&fp=
Microsoft Research
SensorMap
Go To Location
Manage Views | View Permalink | | Sign In
Road Aerial Bird's eye
See this location in bird's eye view.
Manage Views
Popular Views
Seattle
JHU Soil Sensors
SFO Parking
Filter Sensors by Type
Thermometer
Video Camera
Traffic
Parking
Generic
Filter Sensors by Search
Filter by Search
Save Current View
Save View
Saved Views
You have no saved SensorMap views.
A SensorMap saved view saves your SensorMap settings (e.g., map location and sensor types being viewed).
San Francisco
Telegraph Hill
Russian Hill
Chinatown
Nob Hill
Financial District
Embarcadero Plaza
Aquatic Park
Lafayette Park
Microsoft Virtual Earth
©Microsoft Corporation. | Terms of Use | Privacy | About
Done
Internet
100%


Microsoft Research
SensorMap
Go To Location
2D 3D Road Aerial Hybrid Bird's eye
Time Traveler | View Permalink | Sign In |
SensorScope Weather Station Detail
Name: Genepi_station_11
Publisher: SensorScope@epfl.ch
Description: Le Genepi deployment.
Past deployment: 2007-08-27 / 2007-11-05
0 Ambient Temperature
-1.497°C chart
1 Surface Temperature
-4.04°C chart
7 Wind Speed
1.362m/s chart
Manage Views
Popular Views
Seattle
Genepi
SeaMonster
Filter Sensors by Type
Watermark
Rain Meter
Wind Speed
Wind Direction
Water Discharge
Filter Sensors by Search
Filter by Search
Save Current View
Save View
Saved Views
You have no saved SensorMap views.
A SensorMap saved view saves your SensorMap settings (e.g., map location and sensor types being viewed).
Time Series Charts
Comparison Chart
10/23/2007 12:00:00 AM
10/24/2007 12:00:00 AM
10/25/2007 12:00:00 AM
10/26/2007 12:00:00 AM
reload
Individual Charts
Genepi_station_15@0 (unselect)
Genepi_station_15@1 (unselect)
Genepi_station_15@7 (unselect)
Manage Visualizations
Visualization Type
Contour Map
Source Data
Wind Speed
Color Map Settings (customed)
Color map: Discrete
Rainbow
Min value: 1
Max value: 5
Resolution: 0.5
Get Customed Visualization
Contour Map
Wind Speed
10/25/2007 10:00:00 AM ~ 10/25/2007 11:00:00 AM
1 - 1.5
1.5 - 2
2 - 2.5
2.5 - 3
3 - 3.5
3.5 - 4
4 - 4.5
4.5 - 5
Time Traveler
10/23/2007 12:00:00 AM
10/27/2007 12:00:00 AM
Resolution: 1 hour
10/25/2007 10:00:00 AM
Play Next
100 yds
©Microsoft Corporation. | Terms of Use | Privacy | About
Help | Contact Us
javascript://pushin hover
Local intranet
100%

Networked Embedded Computing

Sensor Network Academic Resource Toolkit 2007

A research and teaching resource for building the world-wide sensor web

- Tools for sensor data publishing, collection, processing, and visualization; result of research over the past 3 years
- 1st release 12/05; 4 revisions since
- Over 10,000 downloads worldwide
- Community Preview CD distributed at '06 Faculty Summit
- Cited in MIT Technology Review's "The Year in InfoTech", 12/06
- Source code for
  - Tools for managing sensors and publishing data to **SensorMap**, a portal for organizing and querying wide-area sensor networks
  - **MSRSense microserver** v1.0, a gateway bridging sensornet and Internet, including the microserver execution engine, interaction console, service library and web service interface
  - Streaming and archiving sensor data in **Sencel**, an Excel extension for processing sensor data, and SQL
  - **Drivers** for sensors including motes and webcams

SensorMap: Browsing the Physical World in Real-Time RFP Projects

- Marmite: End-User Programming for Large Sets of Real-Time Sensor Data, CMU
- Leveraging the SensorMap Infrastructure for Large-Scale Urban Monitoring, Harvard
- SensorMap for the National Weather Study Project, NTU, Singapore
- Real-Time Debris Flow Monitoring and Warning via SensorMap, NTHU et al., Taiwan
- Through the looking glass: On human mobility and equipment health, Ohio State
- Semantic Reconciliation with Disparate Sensor Meta-Data for Automatic Publication, U Georgia
- Action Web: Towards Viewing a Mobile World in the First Person, UIUC
- SensorMap for the Great Barrier Reef, U Melbourne et al., Australia
- Publishing and Searching Private Sensor Data Streams: Integration with the SensorMap Platform, UVA
- Event Detection and Notification in the World-Wide Sensor Web, UW
- Mobile Air Quality Monitoring Network, Vanderbilt

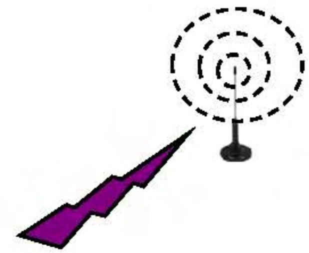
http://research.microsoft.com/ur/us/fundingopps/rfps/SensorMap_RFP_Awards_2007.aspx

Evolution of sensornet platforms

Hitachi
mu-chip RFID

Berkeley Spec mote

Berkeley WeC mote



Sensoria WINS NG 2.0

iPAQ handheld

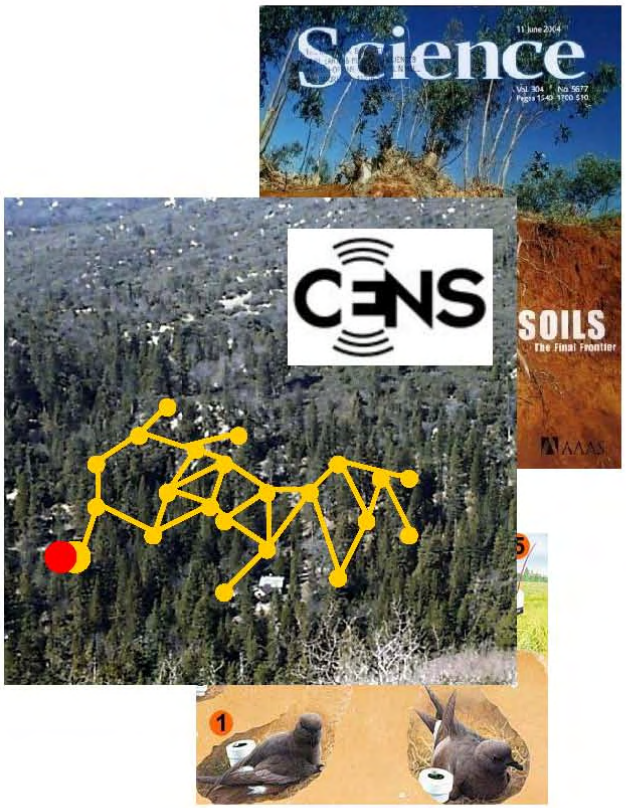
Cell phones

SPOT watch

pedometer

Zhao, MSR Faculty Summit 2007

Evolution of sensornet applications



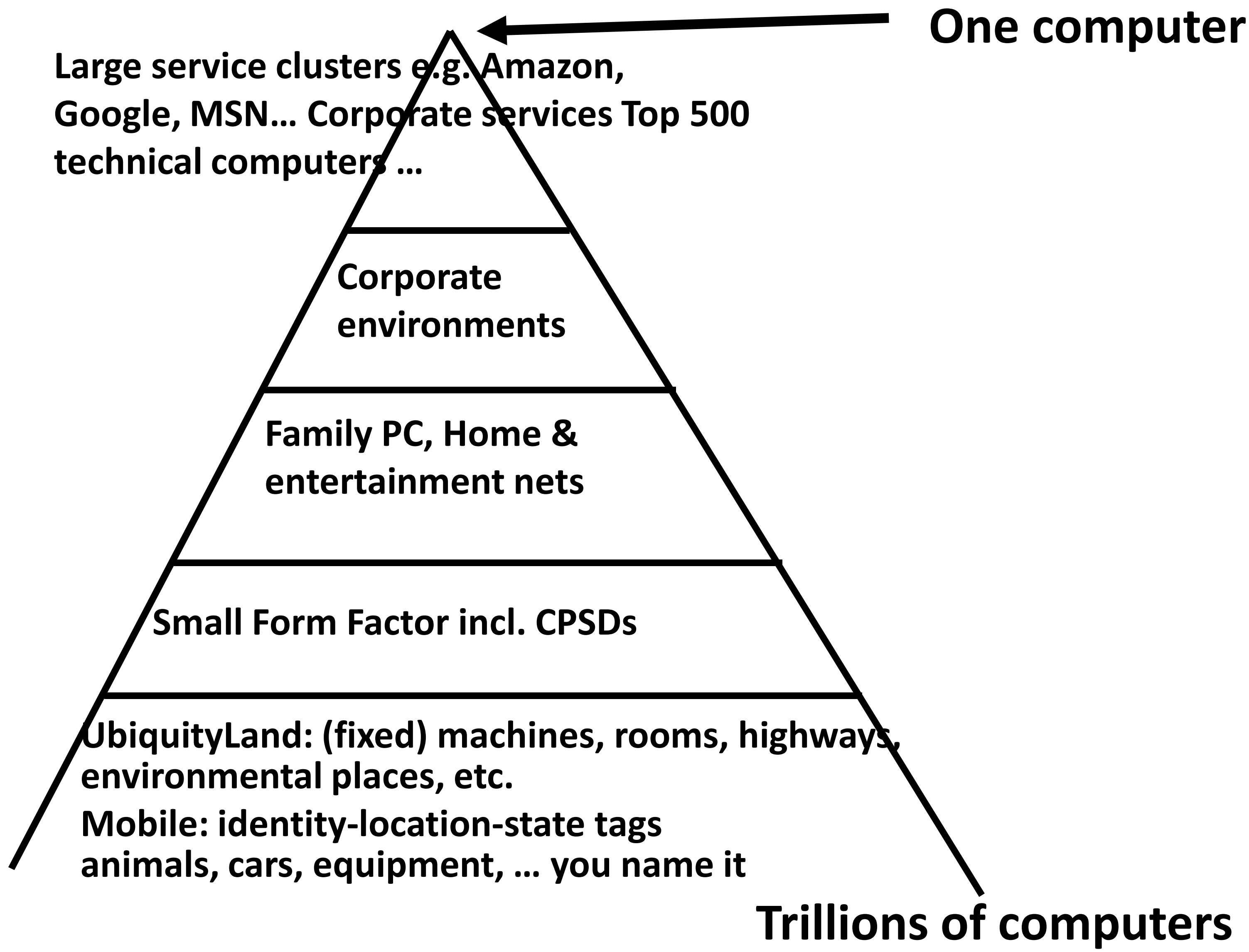
Environmental
- Monitoring space
- E.g., habitat, birds

Industrial:
- Monitoring objects
- E.g. machines, inventories

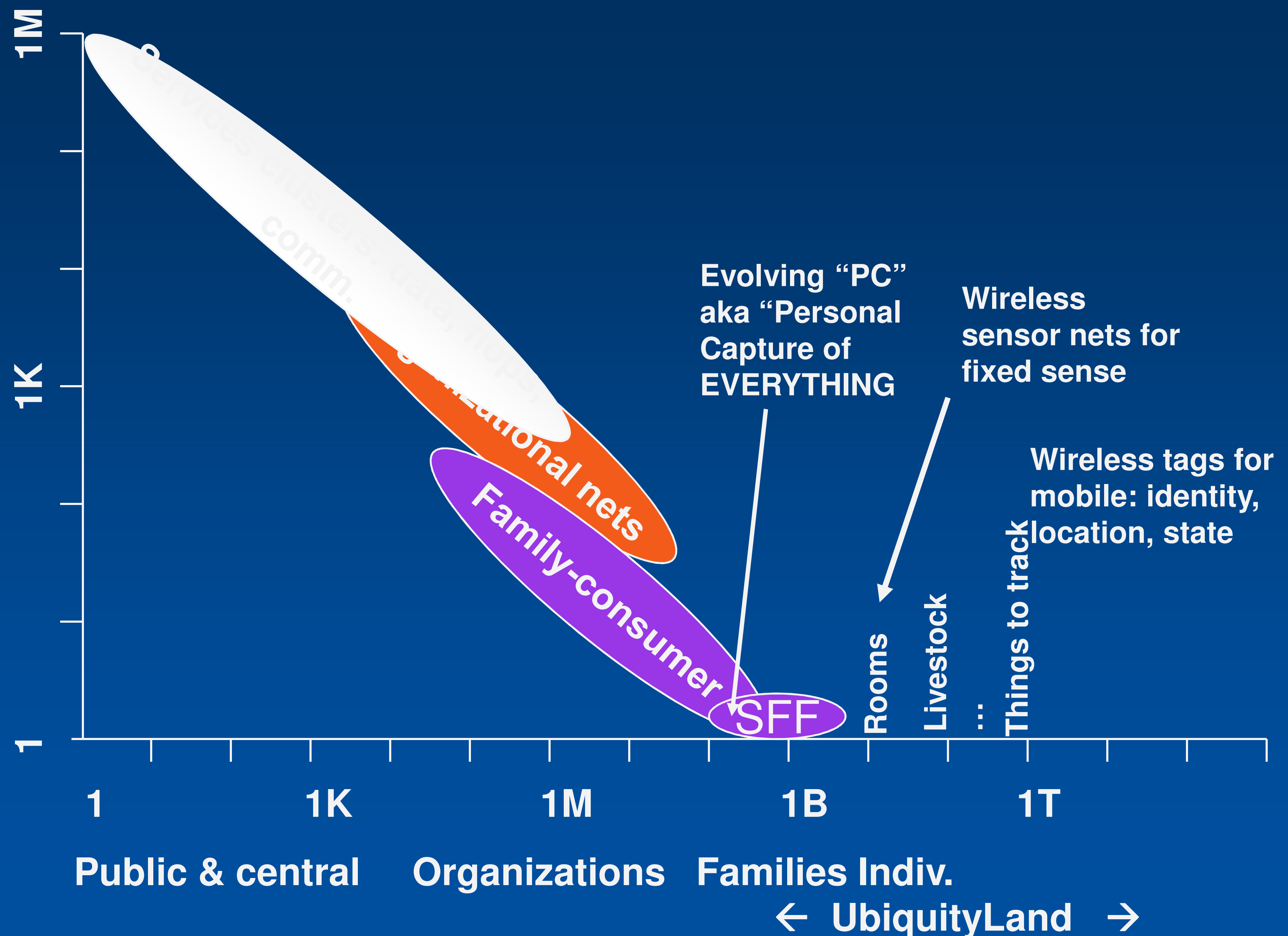
People and community:
- Monitoring activities
- E.g. heath, play, connect

Zhao MSR Faculty Summit 2007

Pyramid of networked - computing, communicating, and storage devices



Computer size (#P) vs population 2010-2020



M2M REMOTE DEVICE MANAGEMENT IN BUSINESS: A STUDY OF CURRENT USERS

Report of results from a recent international survey of early adopters of device networking, targeting product manufacturers. Results show rapid progress, providing strategic opportunities for new revenue generation, profitability and competitive advantage.

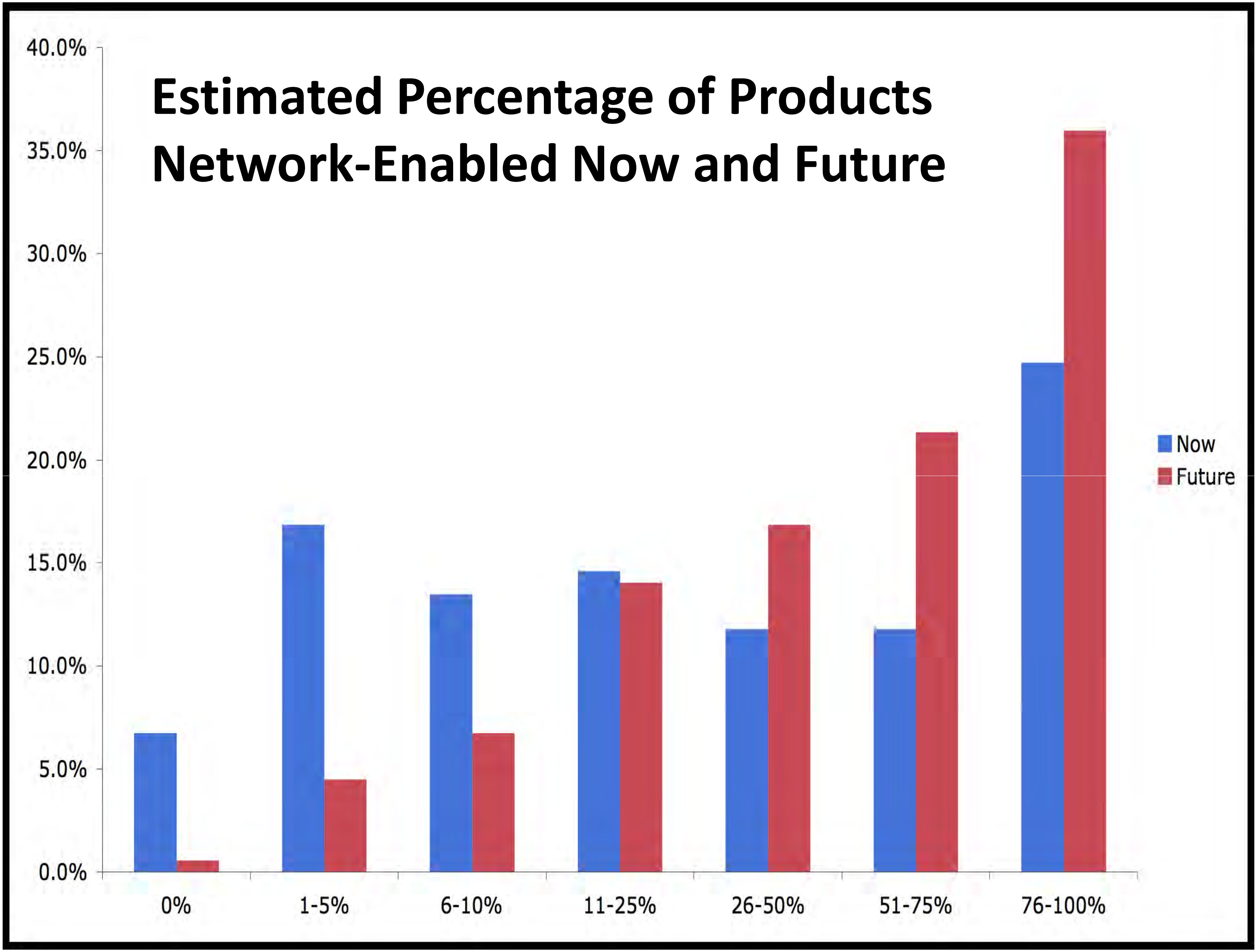
SURVEY REPORT
September 2007

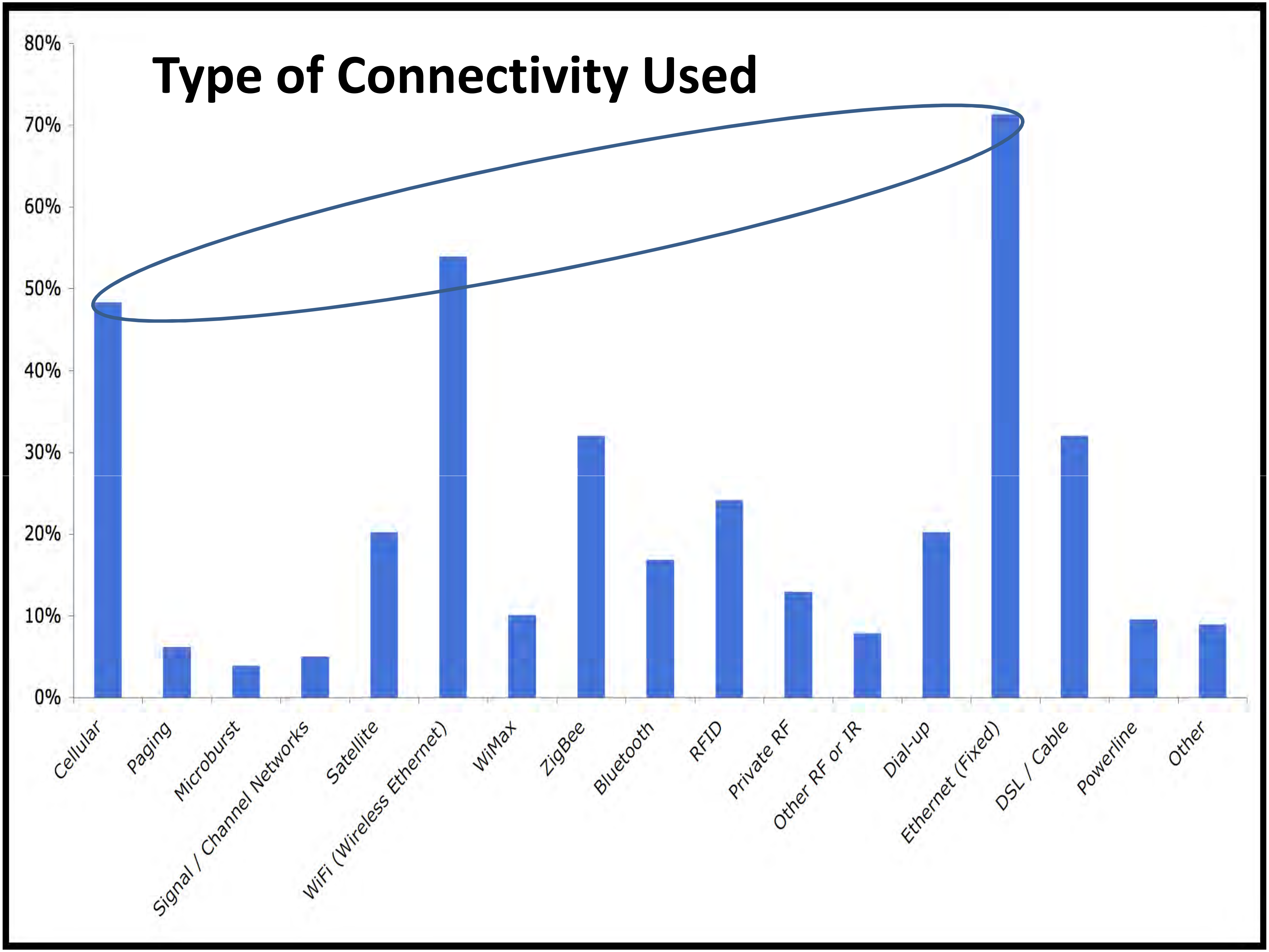
Harbor Research, Inc.
BOSTON | SAN FRANCISCO | LONDON

- **Respondents indicate that M2M (Machine to Machine) device networking is already well advanced …**
- **77% viewed embedded intelligence in their products to be either "Very Important" or "Imperative"**
- **84% would invest in embedding networking capability into products, rather than retrofitting after mfg.**
- **50% were already supporting and monitoring existing devices deployed. 4% not doing it**
- **37% have less than 10% of their product lines networked**

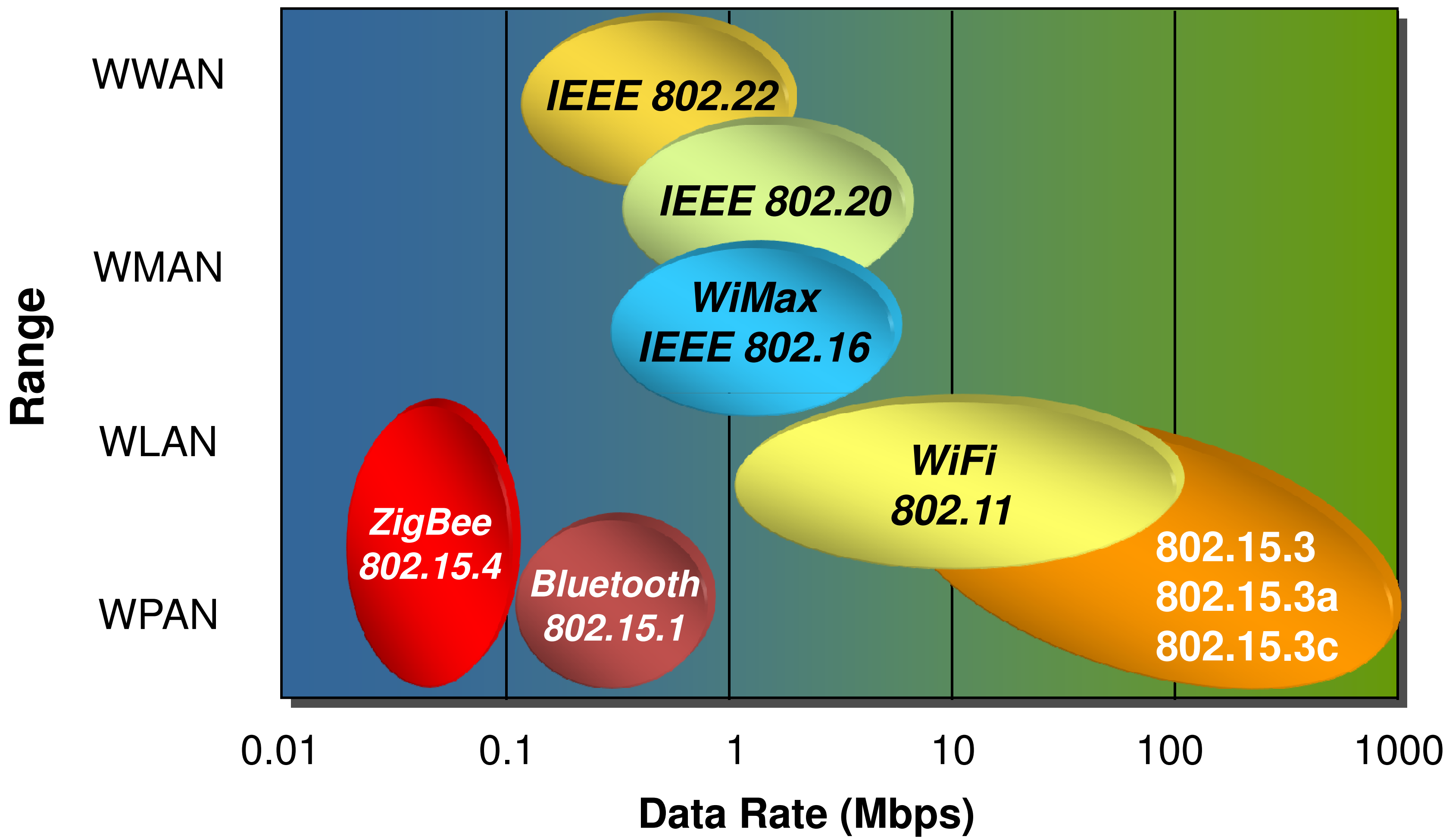
In three years this is expected to fall to less than 12%.

- **36% have more than half of their product portfolio enabled, within three years nearly 60% of these companies expect to embed networking capabilities into >50% of the products**


Estimated Percentage of Products
Network-Enabled Now and Future
40.0%
35.0%
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
0%
1-5%
6-10%
11-25%
26-50%
51-75%
76-100%
Now
Future



Type of Connectivity Used
80%
70%
60%
50%
40%
30%
20%
10%
0%
Cellular
Paging
Microburst
Signal / Channel Networks
Satellite
WiFi (Wireless Ethernet)
WiMax
ZigBee
Bluetooth
RFID
Private RF
Other RF or IR
Dial-up
Ethernet (Fixed)
DSL / Cable
Powerline
Other

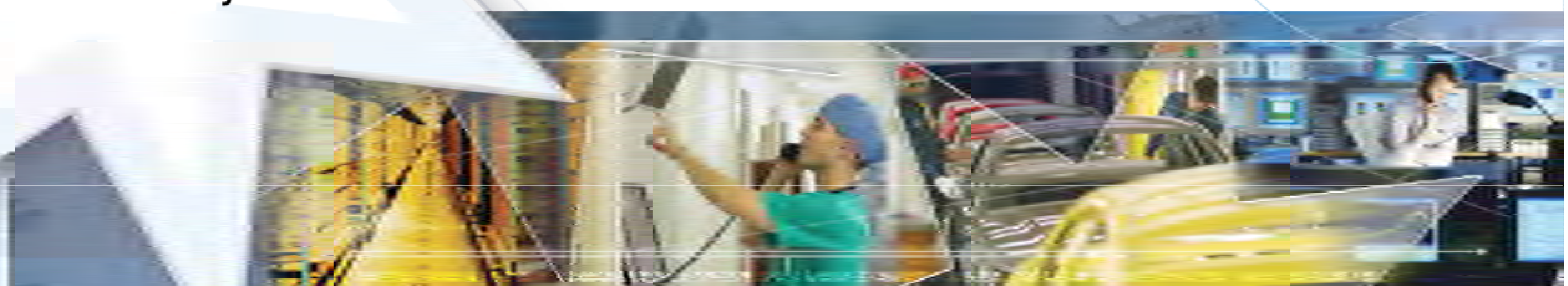

Range
WWAN
WMAN
WLAN
WPAN
IEEE 802.22
IEEE 802.20
WiMax
IEEE 802.16
WiFi
802.11
ZigBee
802.15.4
Bluetooth
802.15.1
802.15.3
802.15.3a
802.15.3c
0.01
0.1
1
10
100
1000
Data Rate (Mbps)

What's so cool about 802.11?

Leverage Investments in Existing WiFi Infrastructures!

- **Global Standard**
  - **Currently over 217,000 Hotspots in over 135 countries**
- **Buildings Managers take advantage of investments in 802.11 nets**
  - **Sensor nodes communicate directly with Industry Std Access Points**
  - **No other devices needed**
- **Install sensor nodes without regard to other Repeaters or Receivers**
  - **Eliminate field surveys to determine a repeater network**
  - **Eliminate service calls due to misplaced or malfunctioning Repeaters**
- **Ultra low power chip technology**
  - **Years of operation on a single AA cell**
- **Easy to expand**
  - **Adding additional sensors is like adding ornaments to a Christmas Tree**
  - **Once security and encryption parameters are agreed upon, sensors are delivered preconfigured. Mount the sensor and turn it on – that's it!**
- **Integration versus Overlay Strategy**
  - **Eliminates the need to convince the site to install yet another net**
  - **Dramatically lower cost of ownership**
- **Become a "Friend of the Court"**
  - **Typically deal with site personnel whose job it is *to integrate 802.11***

G2 Microsystems



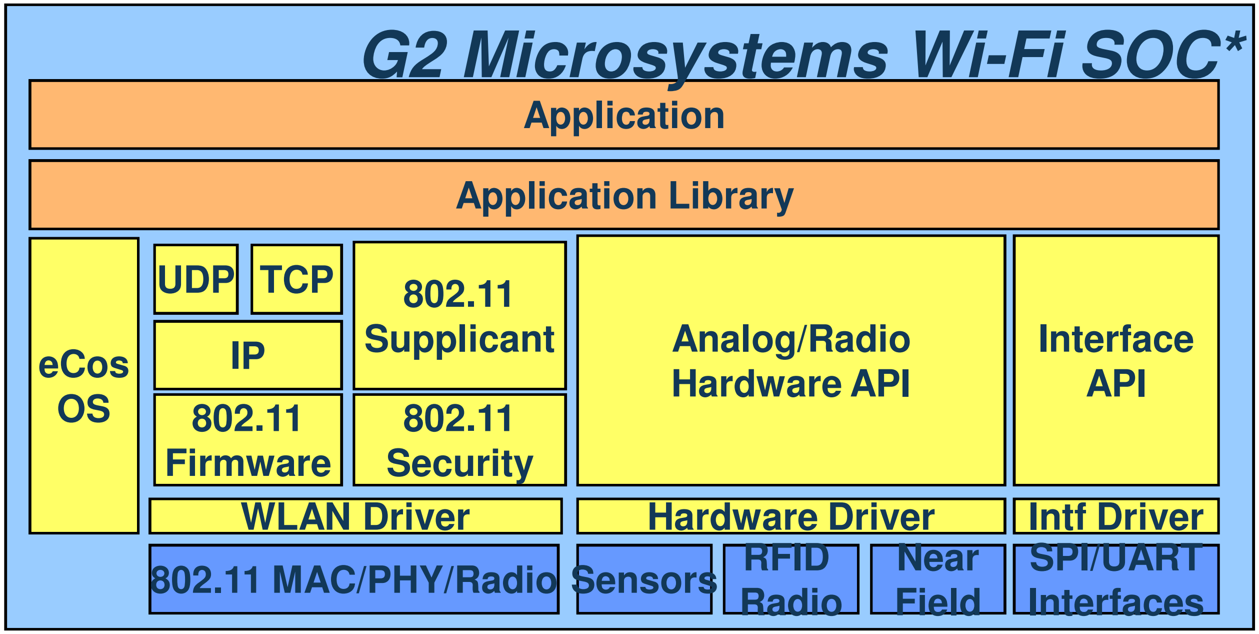
Ultra-low power, Instant-on Wi-Fi Solutions

www.g2microsystems.com

Company Background

- G2 Microsystems creates ultra low-power Wi-Fi system-on-chip solutions for battery-powered devices. Initial target market was locating, monitoring and tracking assets with Wi-Fi.
- Founded in 2004. Venture-backed with investments from Siemens Venture Fund, UPS Strategic Enterprise Fund, Starfish Ventures, and Accede Capital Venture Partners.
- Corporate headquarters in Campbell, CA with Research & Development based in Sydney, Australia.
  - 15 hw engineers, 13 sw engineers, 7 module/systems engineers
- Lead engineers came from Radiata, a start-up company that created the first 802.11a Wi-Fi CMOS implementation; later acquired by Cisco.

G2 Functional block diagram



* Wi-Fi Certified including WMM QoS and WMM-Powersave

Benefits of a Single Chip Wi-Fi Solution

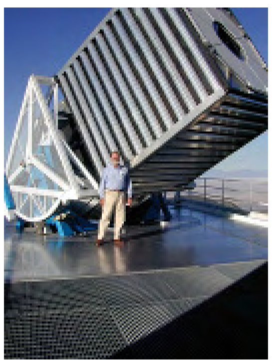
- Application processor can be de-rated or eliminated without Wi-Fi burden
- Fast boot time with full TCP/IP networking stack on-chip
- Fast roaming with 802.11 security supplicants on-chip
- Autonomous Wi-Fi operation for minimum impact on system power

Hardware and the resulting data… more than we could have imagined

The conclusions:
Now it’s really all about the apps (and DATA)!

1. Moore’s Law is alive and well… IP is the most likely the platform and Ethernet / WiFi is likely to be the dominant network
2. It’s time to deploy … vs infrastructure papers & marginal hardware.
   - WSN research space radios, power, protocols, standards, etc.… being mined.
   - The industry has formed and is slowly evolving.
   - Rolling your own motes may be a win if the Fleck vendor succeeds
3. It’s the apps. Science or engineering apps --- sense, deploy, data. Inevitably large scale databases, etc. and on to control apps.
4. Support Fleck, G2Micrsystems and Alive.com home teams. CSIRO apps can understand the limits in <u>your environments</u>.
5. Dust Networks – single chip platform (protocol: 3D’s redundancy)
   - Power, 99.99% rel., no powered nodes, “works” vs. Zigbee “make work”
6. WW Databases: NEON, MSR Sensornet, SkyServer, Env. Data Server

eScience -- A Transformed Scientific Method



***Jim Gray**,*
eScience Group,
Microsoft Research
http://research.microsoft.com/~Gray
in collaboration with
Alex Szalay
Dept. Physics & Astronomy
Johns Hopkins University
http://www.sdss.jhu.edu/~szalay/

Jim Gray

http://research.microsoft.com/~gray

- Jim founded and ran Microsoft’s SF Lab since 1995
- Lost at sea off San Francisco coast 28 January 2007
- IBM Research: System R; Tandem: Transaction Processing, TPC benchmarks; DEC Research
- Jim is a pioneer and proponent of eScience
- MSR: Terra Server c1997 > Google Earth >… Microsoft Virtual Earth &
- Sky Server, Sky Survey, Worldwide Telescope
- NAE, NAS, European Acad Sci… Turing Award

Jim live in Pasteur's Quadrant

Research is inspired by:		**Considerations of use?** No	**Considerations of use?** Yes
Quest for fundamental understanding?	Yes	**Pure, basic research (Bohr)**	**Use inspired, basic research (Pasteur)**
	No		**Pure applied research (Edison)**

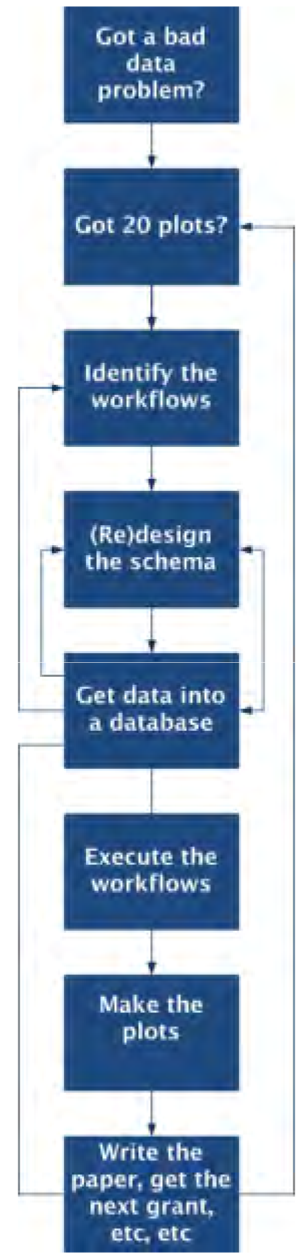
(adapted from *Pasteur's Quadrant: Basic Science and Technological Innovation*, Stokes 1997).

Jim’s beliefs and modus operandi

- Non traditional computer scientist…problem solver
- Focus: “Use” AND “Understanding” inspired research
  Tool builders rarely generate tools, tradesmen do
- Focus on real applications…not toy problems.
- To advance computer science and the tool, find the hardest problem you might be able to solve.
- Learn the science!
- Work with <u>scientists as partners</u> e.g. Astronomy
  1. Astronomy is not a “rich” science
  2. Collaboration is the norm, albeit discovery is important
  3. Large, distributed community not a one-of small science
  4. No commercial value
  5. Drowning in data. Are desperate for help.

Grayfomatics: Engaging with Scientists

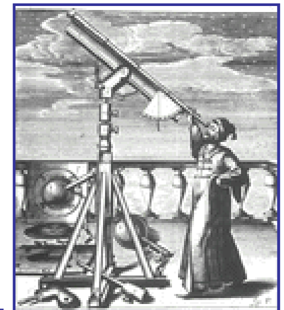
- Make sure the scientists have a data problem – otherwise they won't take the time to talk with you
- Define 20 questions/plots – this drives the technical design, but also helps the cross-disciplines communication
- Spread the 20 questions/plots across "easy", "tricky", "too hard to do now"
- Ask about sharing and security and get to shared pragmatic consensus
- Don't forget to write the papers on both sides – they help drive adoption



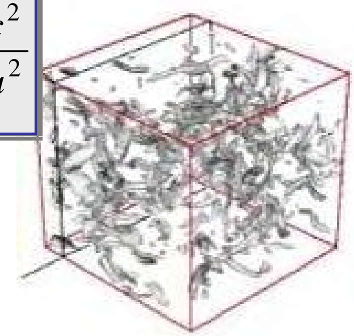
Courtesy Catharine van Ingen

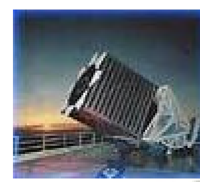
Science Paradigms

1. Thousand years ago:
   science was **empirical**
   describing natural phenomena
2. Last few hundred years:
   theoretical branch
   using models, generalizations
3. Last few decades:
   a **computational** branch
   simulating complex phenomena
4. Today:
   data exploration (eScience)
   unify theory, experiment, and simulation
   – Data captured by instruments
     Or generated by simulator
   – Processed by software
   – Information/Knowledge stored in computer
   – Scientist analyzes database / files
     using data management and statistics

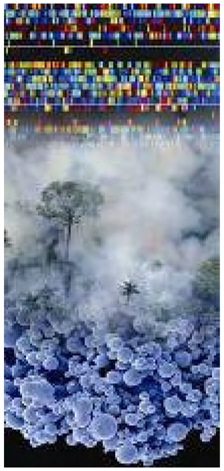


$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$





eScience: What is it?



- Synthesis of
  information technology and science.
- Science *methods* are evolving (tools).
- Science is being codified/objectified.
  How represent scientific information and knowledge in computers?
- Science faces a data deluge.
  How to manage and analyze information?
- Scientific communication changing
  publishing data & literature
  (curation, access, preservation)

Jim Gray NRC-CSTB 2007-01

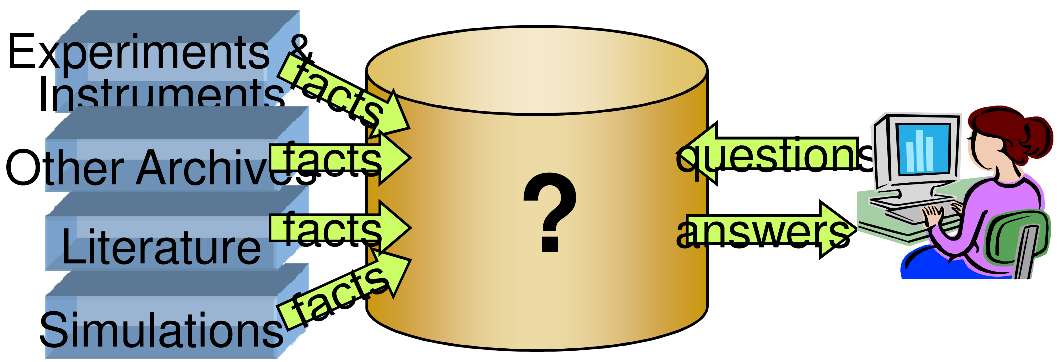
To be accomplished

eScience: what needs to happen within science

- data capture (lab info management systems)
- data curation (schemas, ontologies, provenance)
- data analysis (workflow, algorithms, databases, data visualization)
- data+doc publication (active docs, data-doc integration)
- peer review (editorial services)
- access (doc + data archives and overlay journals)
- Scholarly communication (wiki's for each article and dataset)

X-Info

- The evolution of X-Info and Comp-X
  for each discipline X
- How to codify and represent our knowledge



The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others

- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

Experiment Budgets ¼…½ Software

Software for

- Instrument scheduling
- Instrument control
- Data gathering
- Data reduction
- Database
- Analysis
- Modeling
- Visualization

Millions of lines of code

Repeated for experiment after experiment

Not much sharing or learning

CS can change this

Build generic tools

- Workflow schedulers
- Databases and libraries
- Analysis packages
- Visualizers …

Experiment Budgets ¼…½ Software

Software for

- Instrument scheduling
- Instrument control
- Data gathering
- Data reduction
- Database
- Analysis
- Modeling
- Visualization

Millions of lines of code

Repeated for experiment after experiment

Not much sharing or learning

CS can change this

Build generic tools

- Workflow schedulers
- Databases and libraries
- Analysis packages
- Visualizers …

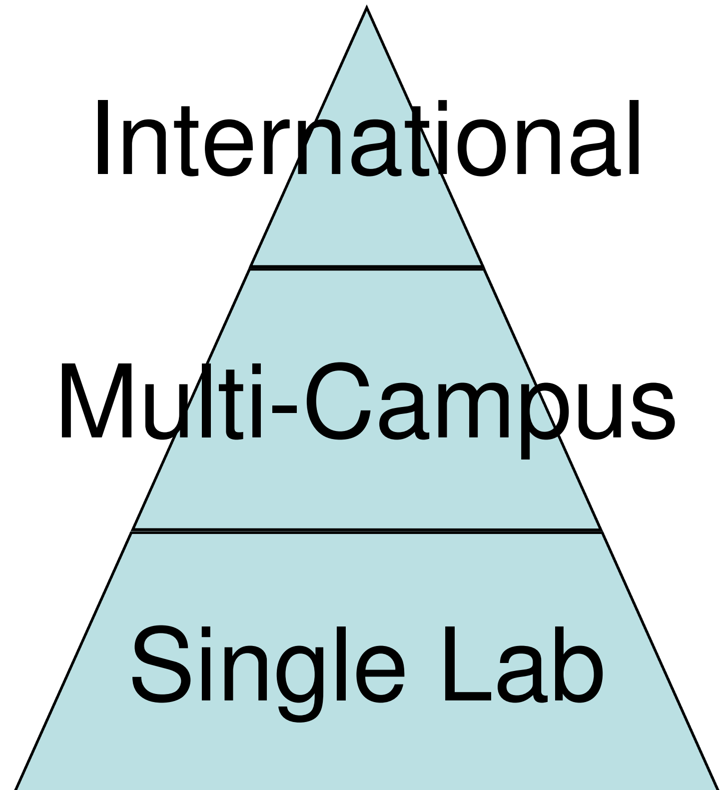
Action item
Foster Tools and
Foster Tool Support

Project Pyramids

In most disciplines there are
- a few "giga" projects,
- several "mega" consortia
- and then many small labs.

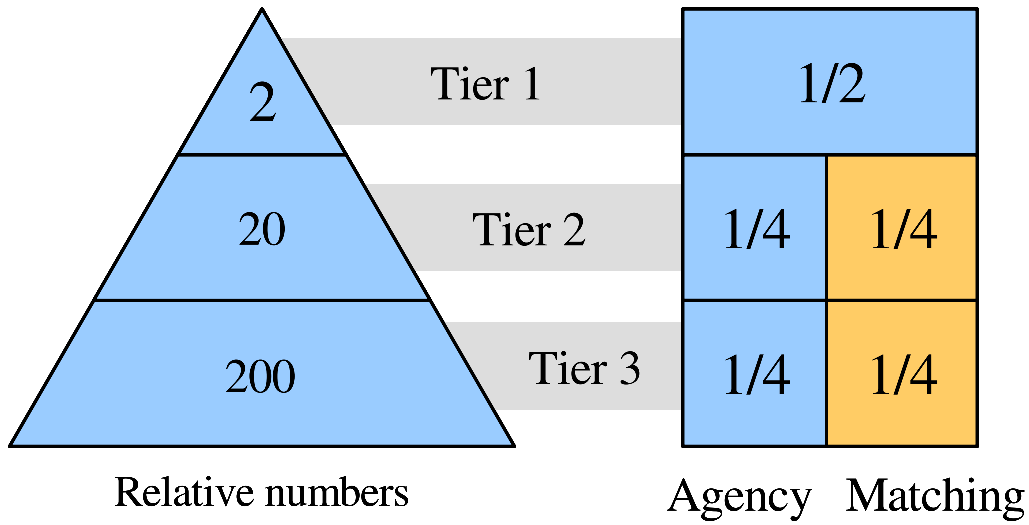
Often some instrument creates need for giga-or mega-project
- Polar station
- Accelerator
- Telescope
- Remote sensor
- Genome sequencer
- Supercomputer

Tier 1, 2, 3 facilities to use instrument + data



Pyramid Funding

- Giga Projects need Giga Funding
  Major Research Equipment Grants
- Need projects at all scales
- computing example:
  supercomputers,
  + departmental clusters
  + lab clusters
- technical+ social issues
- Fully fund giga projects,
  fund ½ of smaller projects
  they get matching funds
  from other sources



- “Petascale Computational Systems: Balanced Cyber-Infrastructure in a Data-Centric World ,” IEEE *Computer*, V. 39.1, pp 110-112, January, 2006.

Science Needs Info Management

- Simulators produce lots of data
- Experiments produce lots of data
- Standard practice:
  - each simulation run produces a file
  - each instrument-day produces a file
  - each process step produces a file
  - files have descriptive names
  - files have similar formats (described elsewhere)
- Projects have millions of files (or soon will)
- No easy way to manage or analyze the data.

Data Delivery: Hitting a Wall

FTP and GREP are not adequate

- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years

- Oh!, and 1PB ~4,000 disks c2007

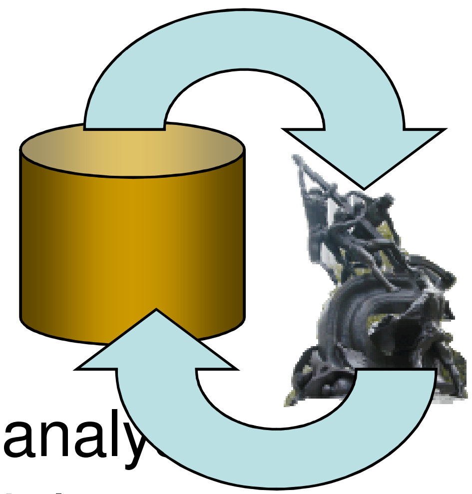
- You can FTP 1 MB in 1 sec
- FTP 1 GB / min (~1 $/GB)
- … 2 days and 1K$
- … 3 years and 1M$

- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where **databases** can help

Accessing Data

- If there is too much data to move around,

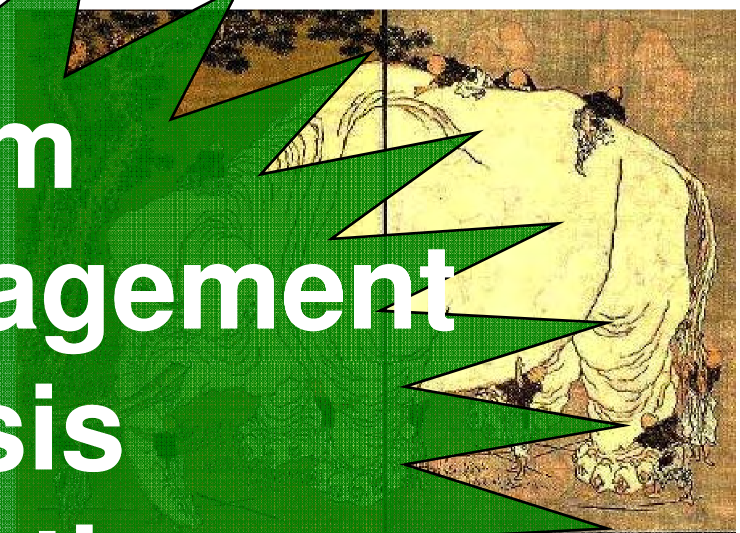
  take the analysis to the data!
- Do all data manipulations at database
  - Build custom procedures and functions in the database
- Automatic parallelism guaranteed
- Easy to build-in custom functionality
  - Databases & Procedures being unified
  - Example temporal and spatial indexing
  - Pixel processing
- Easy to reorganize the data
  - Multiple views, each optimal for certain analy
  - Building hierarchical summaries are trivial
- Scalable to Petabyte datasets



active databases!

Analysis and Databases

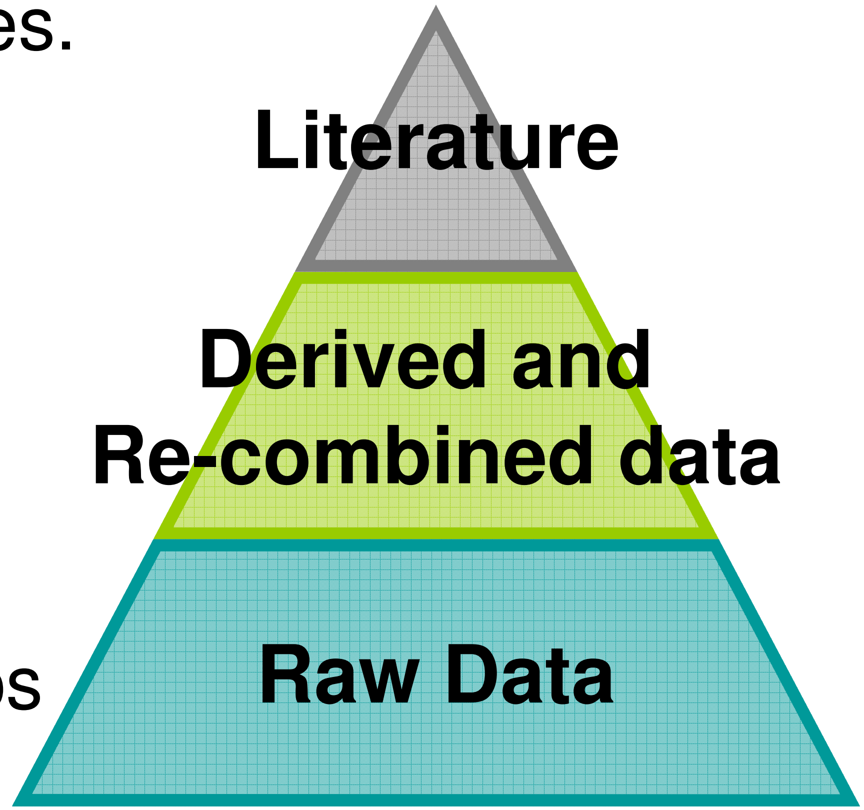
- Much statistical analysis deals with
  - Creating uniform samples –
  - data filtering
  - Assembling relevant subsets
  - Estimating completeness
  - Censoring bad data
  - Counting and building histograms
  - Generating Monte-Carlo subsets
  - Likelihood calculations
  - Hypothesis testing
- Traditionally performed on files
- These tasks better done in structured store with
  - indexing,
  - aggregation,
  - parallelism
  - query, analysis,
  - visualization tools.



Action item
Foster Data Management
Data Analysis
Data Visualization
Algorithms &Tools

All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature to computation to data back to literature.
- Information at your fingertips For everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



Unlocking Peer-Reviewed Literature

- Agencies and Foundations mandating research be public domain.
  - NIH (30 B$/y, 40k PIs,…)
    (see http://www.taxpayeraccess.org/)
  - Welcome Trust
  - Japan, China, Italy, South Africa,….
  - Public Library of Science..
- Other agencies will follow NIH

How Does the New Library Work?

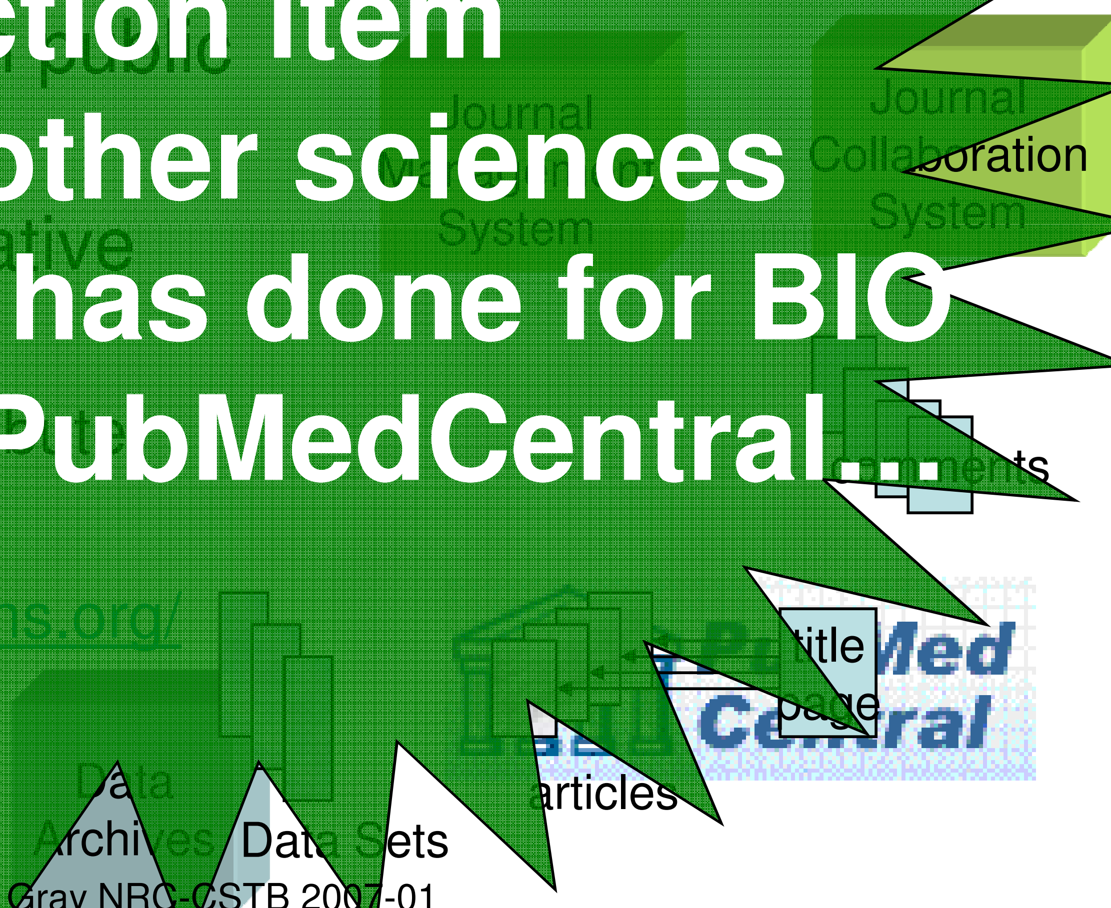
- Who pays for storage access (unfunded mandate)**?**
  - Its cheap: 1 milli-dollar per access

- **But… curation is not cheap**:
  - Author/Title/Subject/Citation/…..
  - Dublin Core is great but…
  - NLM has a 6,000-line XSD for documents http://dtd.nlm.nih.gov/publishing
  - Need to capture document structure from author
    - Sections, figures, equations, citations,…
    - Automate curation
  - NCBI-PubMedCentral is doing this
    - Preparing for 1M articles/year
  - **Automate it!**

Overlay Journals

- Articles and Data in public archives
- Journal title page in public archive.
- All covered by Creative Commons License
  – requires: attribution

  http://creativecommons.org/



Action item
Do for other sciences
what NLM has done for BIO
Genbank-PubMedCentral…

Why Not a Wiki?

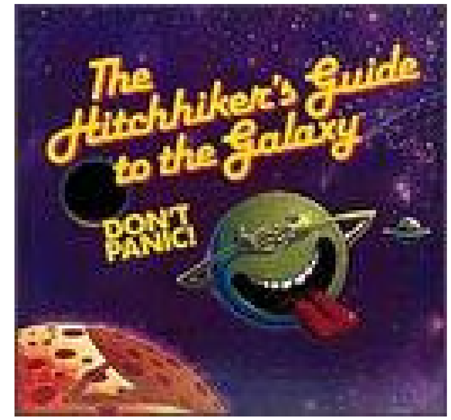
- Peer-Review is different
  - It is very structured
  - It is moderated
  - There is a degree of confidentiality
- Wiki is egalitarian
  - It's a conversation
  - It's completely transparent
- Don't get me wrong:
  - Wiki's are great
  - SharePoints are great
  - But.. Peer-Review is different.
  - And, incidentally: review of proposals, projects,… is more like peer-review.
- Let's have Moderated Wiki re published literature PLoS-One is doing this

Why Not a Wiki?

- Peer-Review is different
  - It is very structured
  - It is moderated
  - There is a degree of confidentiality
- Wiki is egalitarian
  - It's a conversation
  - It's completely transparent
- Don't get me wrong:
  - Wiki's are great
  - SharePoints are great
  - But… Peer-Review is different.
  - And, incidentally: review of proposals, projects, … is more like peer-review.
- Let's have Moderated Wiki re published literature PLoS-One is doing this

Action item
Foster new document authoring and publication models and tools

So… What about Publishing Data?



- The answer is **42**.
- But…
  - What are the units?
  - How precise? How accurate 42.5 ± .01
  - Show your work
    data *provenance*

Thought Experiment

- You have collected some data and want to publish science based on it.
- How do you publish the data so that others can read it and reproduce your results in 100 years?
  - Document collection process?
  - How document data processing (scrubbing & reducing the data)?
  - Where do you put it?

Objectifying Knowledge

Warning!

Painful discussions ahead:

The “O” word: Ontology

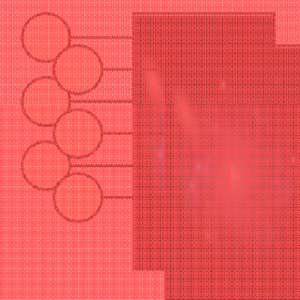
The “S” word: Schema

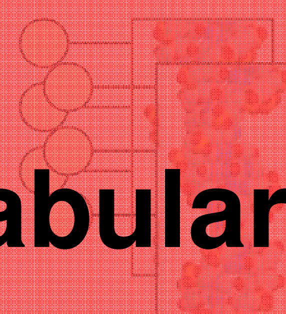
The “CV” words:

Controlled Vocabulary

Domain experts do not agree

- This requires agreement about
  - Units: cgs
  - Measurements: who/what/when/where/how
  - CONCEPTS:
    - What’s a planet, star, galaxy,…?
    - What’s a gene, protein, pathway…?
- Need to objectify science:
  - what are the objects?
  - what are the attributes?
  - What are the methods (in the OO sense)?
- This is mostly Physics/Bio/Eco/Econ/...
  But CS can do generic things



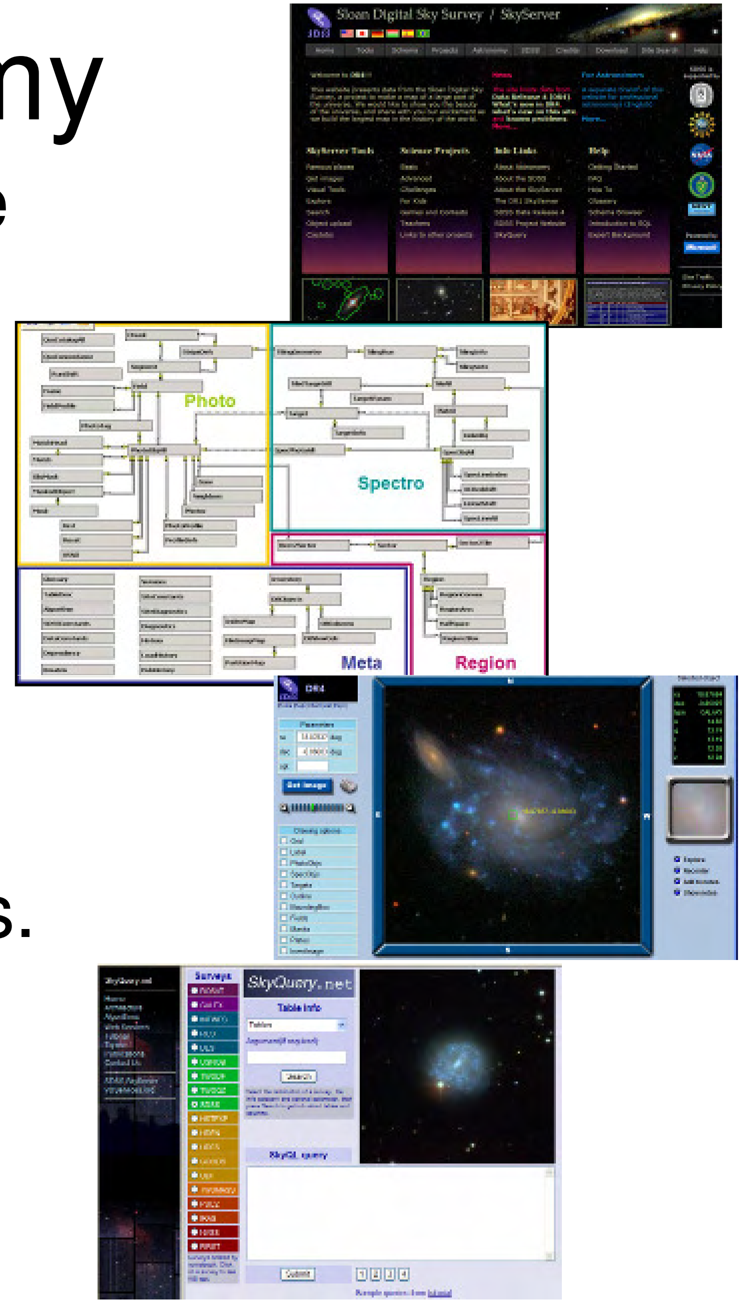


Examples

-

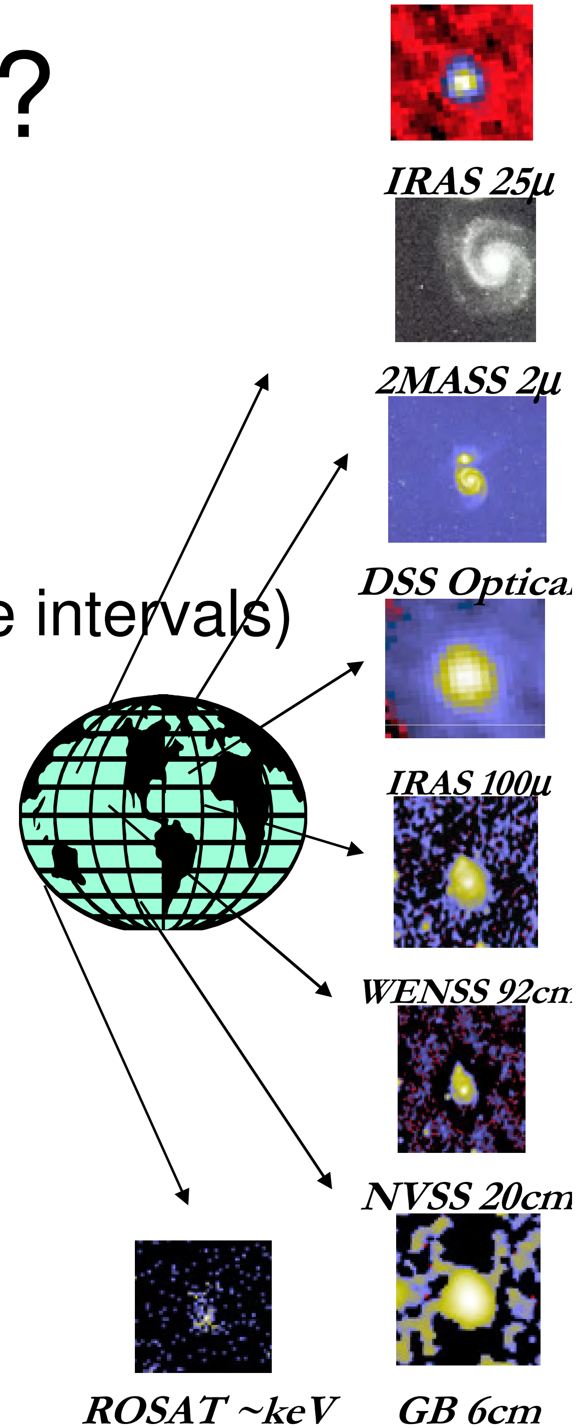
Astronomy

- Help build world-wide telescope
  - All astronomy data and literature online and cross indexed
  - Tools to analyze the data
- Built SkyServer.SDSS.org
- Built Analysis system
  - MyDB
  - CasJobs (batch job)
- OpenSkyQuery
  Federation of ~20 observatories.
- Results:
  - It works and is used every day
  - Spatial extensions in SQL 2005
  - A good example of Data Grid
  - Good examples of Web Services.



Why Astronomy Data?

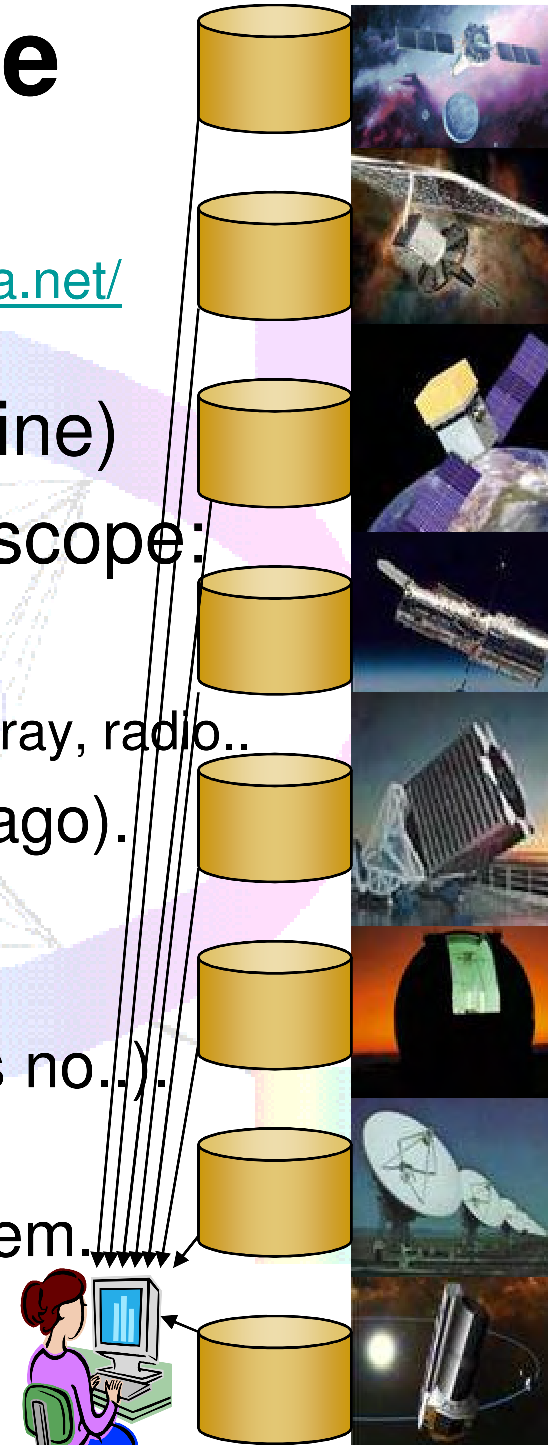
- **It has no commercial value**
  - –No privacy concerns
  - –Can freely share results with others
  - –Great for experimenting with algorithms
- **It is real and well documented**
  - –**High-dimensional data** (with confidence intervals)
  - –**Spatial** data
  - –**Temporal** data
- Many **different instruments** from many **different places** and many **different times**
- **Federation is a goal**
- There is a lot of it (petabytes)

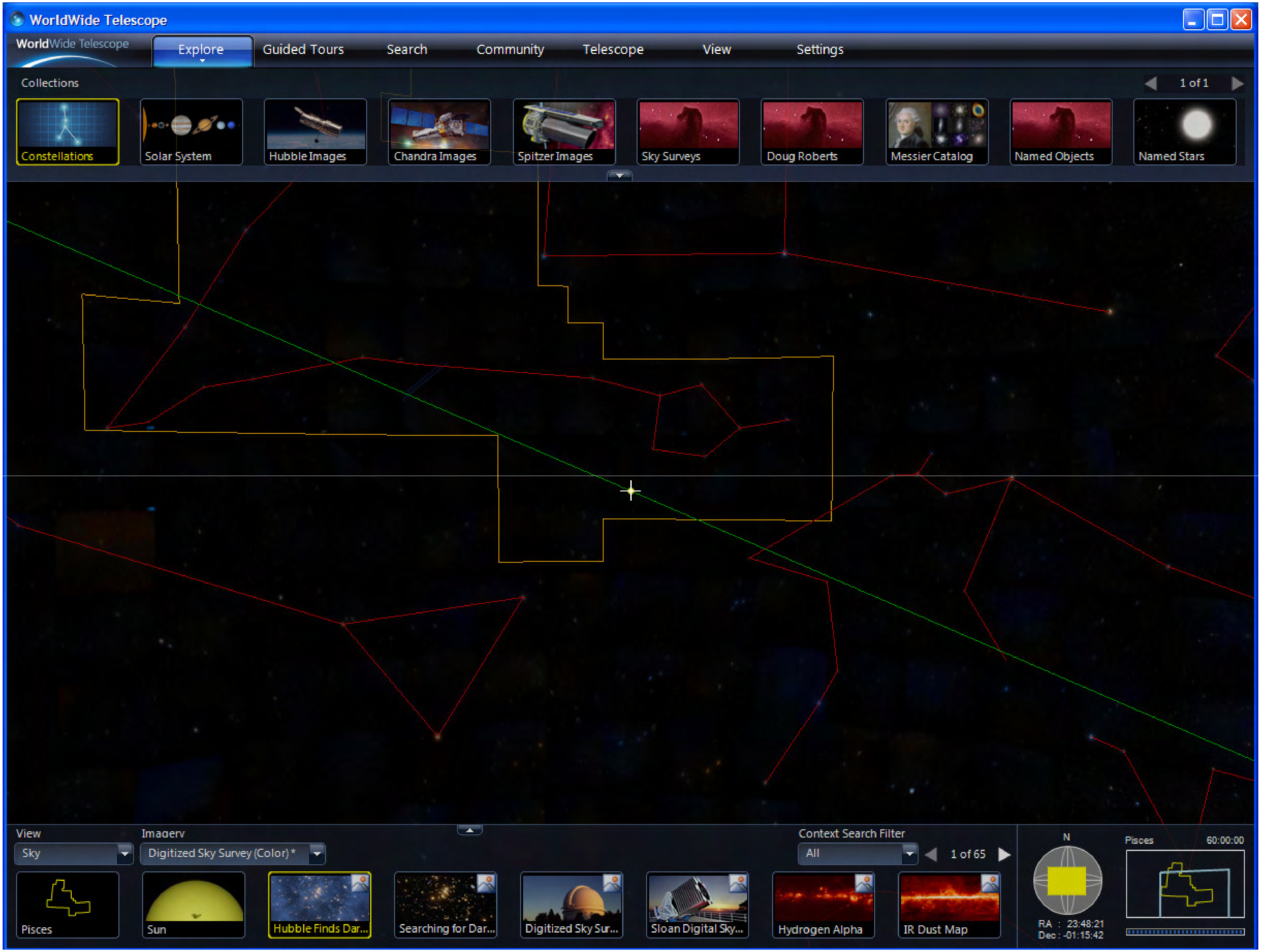


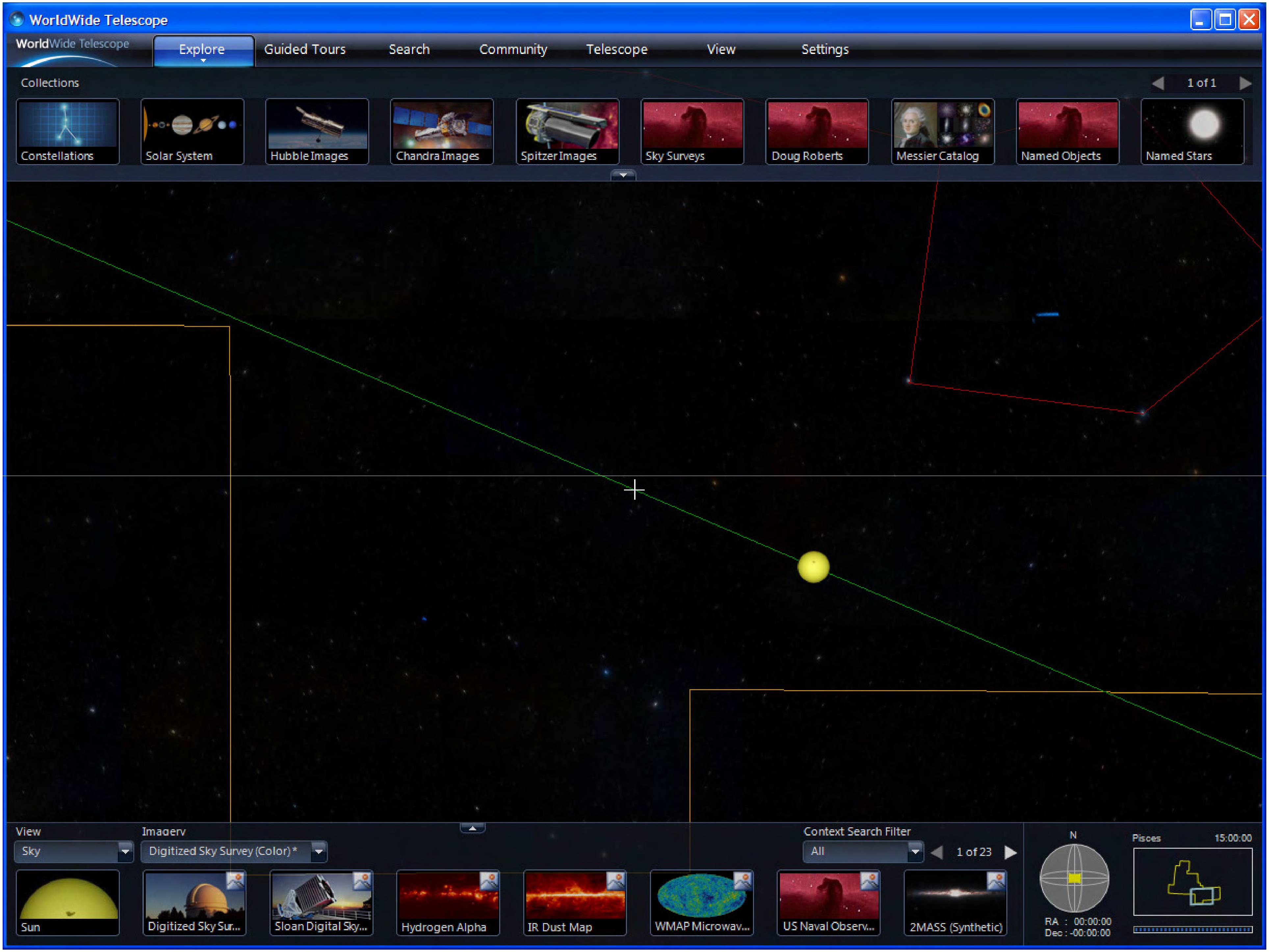
World Wide Telescope Virtual Observatory

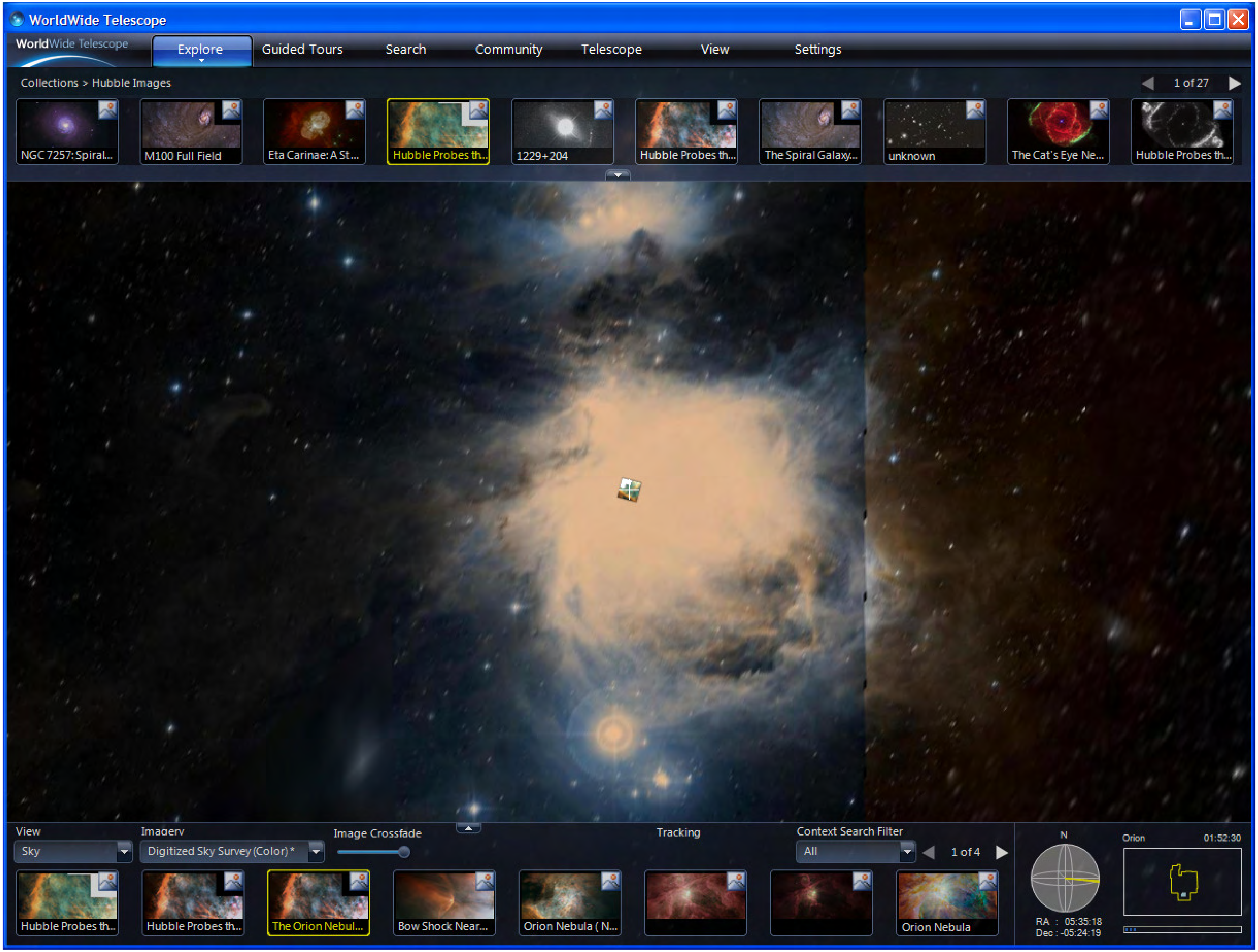
http://www.us-vo.org/ http://www.ivoa.net/

- Premise: Most data is (or could be online)
- So, the Internet is the world's best telescope:
  - It has data on every part of the sky
  - In every measured spectral band: optical, x-ray, radio..
  - As deep as the best instruments (2 years ago).
  - It is up when you are up.
    The "seeing" is always great
    (no working at night, no clouds no moons no..).
  - It's a smart telescope:
    links objects and data to literature on them.



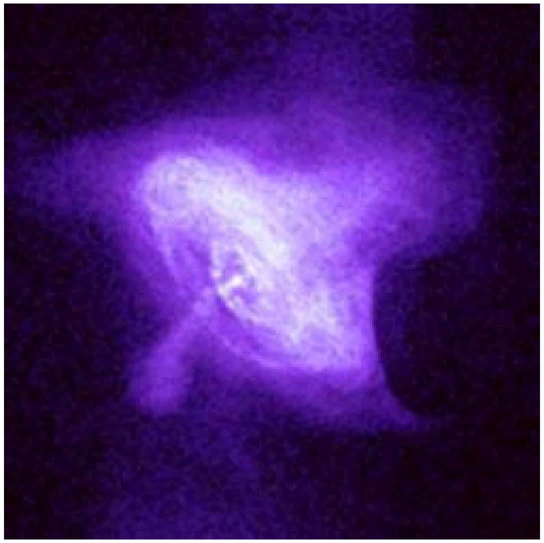

WorldWide Telescope
WorldWide Telescope
Explore
Guided Tours
Search
Community
Telescope
View
Settings
Collections
1 of 1
Constellations
Solar System
Hubble Images
Chandra Images
Spitzer Images
Sky Surveys
Doug Roberts
Messier Catalog
Named Objects
Named Stars
View
Sky
Imagery
Digitized Sky Survey (Color) *
Context Search Filter
All
1 of 65
Pisces
Sun
Hubble Finds Dar...
Searching for Dar...
Digitized Sky Sur...
Sloan Digital Sky...
Hydrogen Alpha
IR Dust Map
N
RA : 23:48:21
Dec : -01:15:42
Pisces
60:00:00

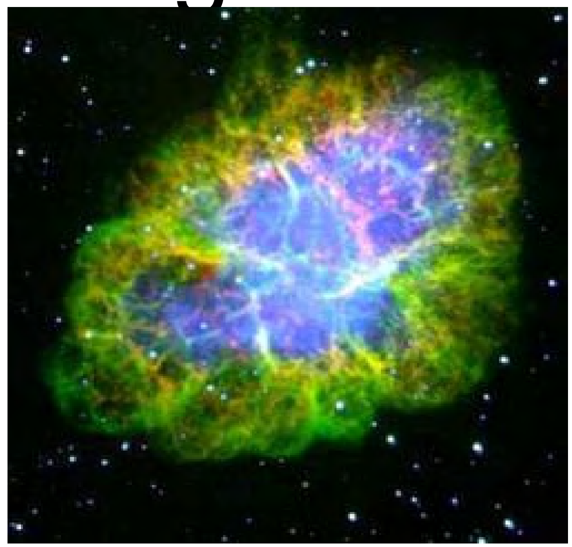

WorldWide Telescope
Explore
Guided Tours
Search
Community
Telescope
View
Settings
Collections
1 of 1
Constellations
Solar System
Hubble Images
Chandra Images
Spitzer Images
Sky Surveys
Doug Roberts
Messier Catalog
Named Objects
Named Stars
View
Sky
Imagery
Digitized Sky Survey (Color) *
Context Search Filter
All
1 of 23
Sun
Digitized Sky Sur...
Sloan Digital Sky...
Hydrogen Alpha
IR Dust Map
WMAP Microwav...
US Naval Observ...
2MASS (Synthetic)
N
RA : 00:00:00
Dec : -00:00:00
Pisces
15:00:00

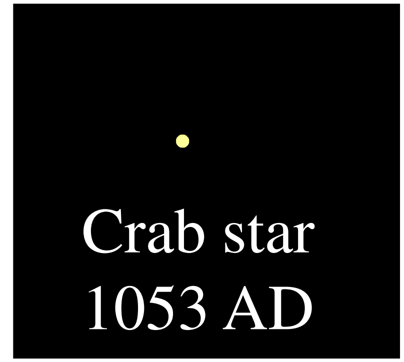

WorldWide Telescope
Explore
Guided Tours
Search
Community
Telescope
View
Settings
Collections > Hubble Images
1 of 27
NGC 7257: Spiral...
M100 Full Field
Eta Carinae: A St...
Hubble Probes th...
1229+204
Hubble Probes th...
The Spiral Galaxy...
unknown
The Cat's Eye Ne...
Hubble Probes th...
View
Sky
Imagery
Digitized Sky Survey (Color) *
Image Crossfade
Tracking
Context Search Filter
All
1 of 4
Hubble Probes th...
Hubble Probes th...
The Orion Nebul...
Bow Shock Near...
Orion Nebula (N...
Orion Nebula
N
RA : 05:35:18
Dec : -05:24:19
Orion
01:52:30

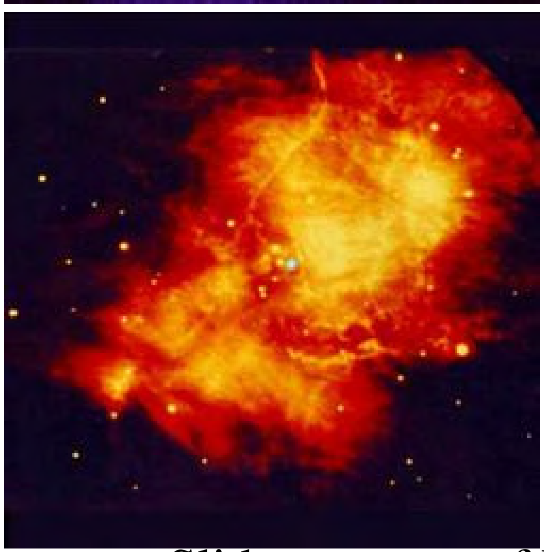

WorldWide Telescope
Explore
Guided Tours
Search
Community
Telescope
View
Settings
Collections
1 of 1
Constellations
Solar System
Hubble Images
Chandra Images
Spitzer Images
Sky Surveys
Doug Roberts
Messier Catalog
Named Objects
Named Stars
Finder Scope
Classification:
Unidentified
in Andromeda
Names: No Object
RA : 00h42m46s
Dec : 41 : 18 : 19
Alt : 47 : 12 : 23
Az : 284 : 55 : 01
Magnitude: 0
Dimensions: n/a
Distance:
Image Credits:
Copyright DSS Consortium
http://www-gsss.stsci.edu/Acknowledgeme...
Research
Show object
Close
View
Sky
Imagery
Digitized Sky Survey (Color) *
Image Crossfade
Context Search Filter
All
1 of 1
M31
M32
NGC221
NGC224
N
RA : 00:42:45
Dec : +41:18:16
Andromeda
00:56:15

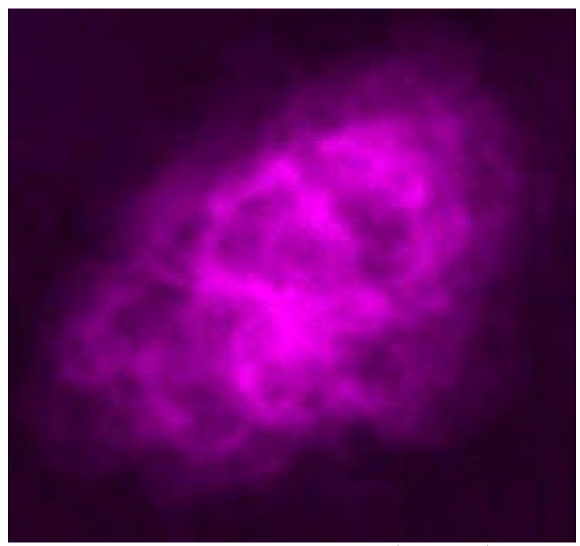
Time and Spectral Dimensions
The Multiwavelength Crab Nebulae











X-ray, optical, infrared, and radio views of the nearby Crab Nebula, which is now in a state of chaotic expansion after a supernova explosion first sighted in 1054 A.D. by Chinese Astronomers.

Slide courtesy of Robert Brunner @ CalTech.

SkyServer/SkyQuery Evolution
MyDB and Batch Jobs

Problem: need multi-step data analysis (not just single query).

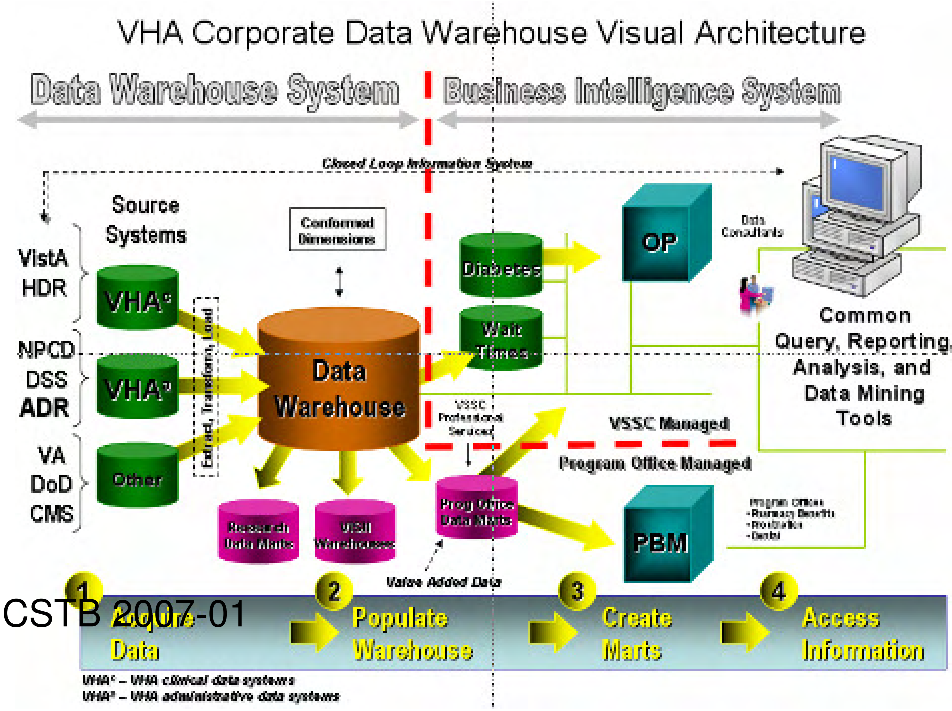
Solution: Allow personal databases on portal

Problem: some queries are monsters

Solution: “Batch schedule” on portal. Deposits answer in personal database.

VHA Health Informatics

- VHA: largest standardized electronic medical records system in US.
- Design, populate and tune a ~20 TB Data Warehouse and Analytics environment
- Evaluate population health and treatment outcomes,
- Support epidemiological studies
  - **7 million enrollees**
  - **5 million patients**
  - **Example Milestones:**
    - **1 Billionth Vital Sign loaded in April '06**
    - **30-minutes to population-wide obesity analysis (next slide)**
    - **Discovered seasonality in blood pressure -- NEJM fall '06**



HDR Vitals Based Body Mass Index Calculation on VHA FY04 Population

Source: VHA Corporate Data Warehouse

Wt/Ht	5ft 0in	5ft 1in	5ft 2in	5ft 3in	5ft 4in	5ft 5in	5ft 6in	5ft 7in	5ft 8in	5ft 9in	5ft 10in	5ft 11in	6ft 0in	6ft 1in	6ft 2in	6ft 3in	6ft 4in	6ft 5in
100	230	211	334	276	316	364	346	300	244	172	114	73	58	16	11	3	1	1
105	339	364	518	532	558	561	584	515	436	284	226	144	102	25	13	4	4	1
110	488	489	836	815	955	972	1,031	899	680	521	395	256	161	70	23	10	6	4
115	526	614	1,018	1,098	1,326	1,325	1,607	1,426	1,175	903	598	451	264	84	59	17	6	4
120	644	714	1,419	1,583	1,964	2,153	2,612	2,374	1,933	1,450	1,085	690	501	153	95	38	13	9
125	672	855	1,682	1,933	2,628	3,005	3,521	3,405	2,929	2,197	1,538	1,144	756	253	114	46	32	8
130	753	944	1,984	2,392	3,462	3,968	5,039	4,827	4,285	3,223	2,378	1,765	1,182	429	214	81	41	12
135	753	1,062	2,173	2,852	4,105	4,912	6,535	6,535	5,797	4,500	3,393	2,467	1,668	596	309	108	70	15
140	754	1,073	2,300	3,177	4,937	6,286	8,769	8,750	7,939	6,303	4,837	3,493	2,534	977	513	144	106	22
145	748	1,053	2,254	3,389	5,412	7,334	10,485	11,004	10,576	8,084	6,511	4,686	3,344	1,207	680	221	140	41
150	730	1,077	2,361	3,596	6,152	8,665	12,772	14,335	13,866	11,255	9,250	6,545	4,796	1,792	979	350	162	48
155	683	923	2,178	3,391	6,031	8,891	14,181	15,899	16,594	13,517	11,489	8,056	5,741	2,155	1,203	472	249	70
160	671	872	2,106	3,532	6,184	9,580	15,493	18,869	19,939	17,046	14,650	10,366	7,708	2,831	1,618	615	341	100
165	627	772	1,894	3,074	5,773	9,549	16,332	20,080	22,507	19,692	17,729	12,588	9,558	3,548	2,032	716	399	117
170	596	750	1,716	2,900	5,428	9,080	16,633	21,550	25,051	22,568	21,198	15,552	12,093	4,548	2,626	944	489	124
175	493	674	1,521	2,551	4,816	8,417	15,900	21,420	26,262	24,277	23,756	18,194	13,817	5,361	3,178	1,152	586	144
180	486	599	1,411	2,323	4,584	7,855	15,482	20,873	26,922	26,067	26,313	20,358	16,459	6,451	3,848	1,441	737	207
185	420	546	1,195	1,985	3,905	6,918	13,406	19,362	25,818	25,620	27,037	21,799	18,172	7,206	4,458	1,548	867	247
190	424	495	1,073	1,729	3,383	5,909	11,918	17,640	24,277	25,263	27,398	22,697	19,977	8,344	4,937	1,858	963	287
195	341	463	913	1,474	2,803	5,207	10,584	15,727	22,137	23,860	26,373	22,513	20,163	8,754	5,683	2,178	1,120	309
200	315	384	763	1,338	2,602	4,551	9,413	14,149	20,608	22,541	25,452	23,358	21,548	9,284	6,221	2,294	1,295	372
205	265	338	633	1,026	1,993	3,736	7,765	11,940	17,501	19,944	23,065	21,094	20,354	9,270	6,350	2,597	1,322	376
210	275	284	543	853	1,794	3,148	6,804	10,540	15,647	18,129	21,862	20,540	20,271	9,566	6,816	2,786	1,509	418
215	205	244	501	746	1,389	2,645	5,747	8,712	13,064	15,560	19,089	18,191	19,063	9,019	6,675	2,798	1,509	454
220	168	208	415	652	1,231	2,326	4,950	7,751	11,645	13,900	17,577	17,239	17,583	8,896	6,818	2,948	1,635	484
225	156	160	325	522	968	1,873	4,015	6,340	9,794	11,890	14,898	15,097	15,741	8,332	6,441	2,915	1,647	452
230	141	160	259	486	880	1,653	3,334	5,410	8,657	10,500	13,532	13,488	14,815	7,901	6,258	2,859	1,701	496
235	115	119	244	373	738	1,251	2,795	4,570	7,192	8,784	11,489	11,857	12,796	7,113	5,544	2,744	1,617	465
240	72	116	214	313	562	1,099	2,422	3,861	6,044	7,652	9,982	10,692	11,825	6,496	5,392	2,606	1,581	449
245	71	76	169	253	509	888	1,858	3,167	5,076	6,446	8,312	8,647	9,910	5,638	4,742	2,263	1,479	469
250	70	55	152	226	452	753	1,647	2,826	4,505	5,509	7,569	8,064	8,900	5,183	4,319	2,177	1,451	469
255	59	61	128	174	316	599	1,289	2,130	3,468	4,540	5,957	6,451	7,438	4,320	3,741	1,903	1,271	443
260	50	64	117	167	281	493	1,107	1,929	2,963	3,947	5,190	5,797	6,725	3,900	3,429	1,828	1,218	481
265	37	34	88	122	234	454	894	1,449	2,457	3,152	4,374	4,818	5,729	3,350	2,984	1,539	1,028	406
270	47	42	67	119	203	367	800	1,291	2,110	2,740	3,878	4,133	5,075	2,934	2,685	1,468	918	403
275	22	34	44	85	184	291	662	1,064	1,767	2,235	3,113	3,412	4,267	2,598	2,362	1,247	837	334
280	21	20	51	69	139	286	548	903	1,513	1,955	2,770	3,126	3,604	2,273	2,020	1,152	763	300
285	12	12	36	68	118	201	451	720	1,318	1,613	2,208	2,394	3,132	1,924	1,780	994	677	241
290	16	14	47	38	92	182	387	667	1,050	1,301	1,904	2,150	2,655	1,749	1,529	881	688	252
295	9	12	22	53	92	127	341	493	858	1,162	1,677	1,822	2,338	1,445	1,333	813	533	202
300	12	10	30	43	59	117	309	434	764	988	1,428	1,588	1,989	1,255	1,212	709	479	205

DRAFT

Jim Gray NRC-CSTB 2007-01

Legend	Total Patients
BMI < 18 Underweight	23,876 (0.7%)
BMI 18-24.9 Healthy Weight	701,089 (21.6%)
BMI 25-29.9 Overweight	1,177,093 (36.2%)
BMI 30+ Obese	1,347,098 (41.5%)
	3,249,156 (100%)

Environmental Data Server

Catharine van Ingen et al

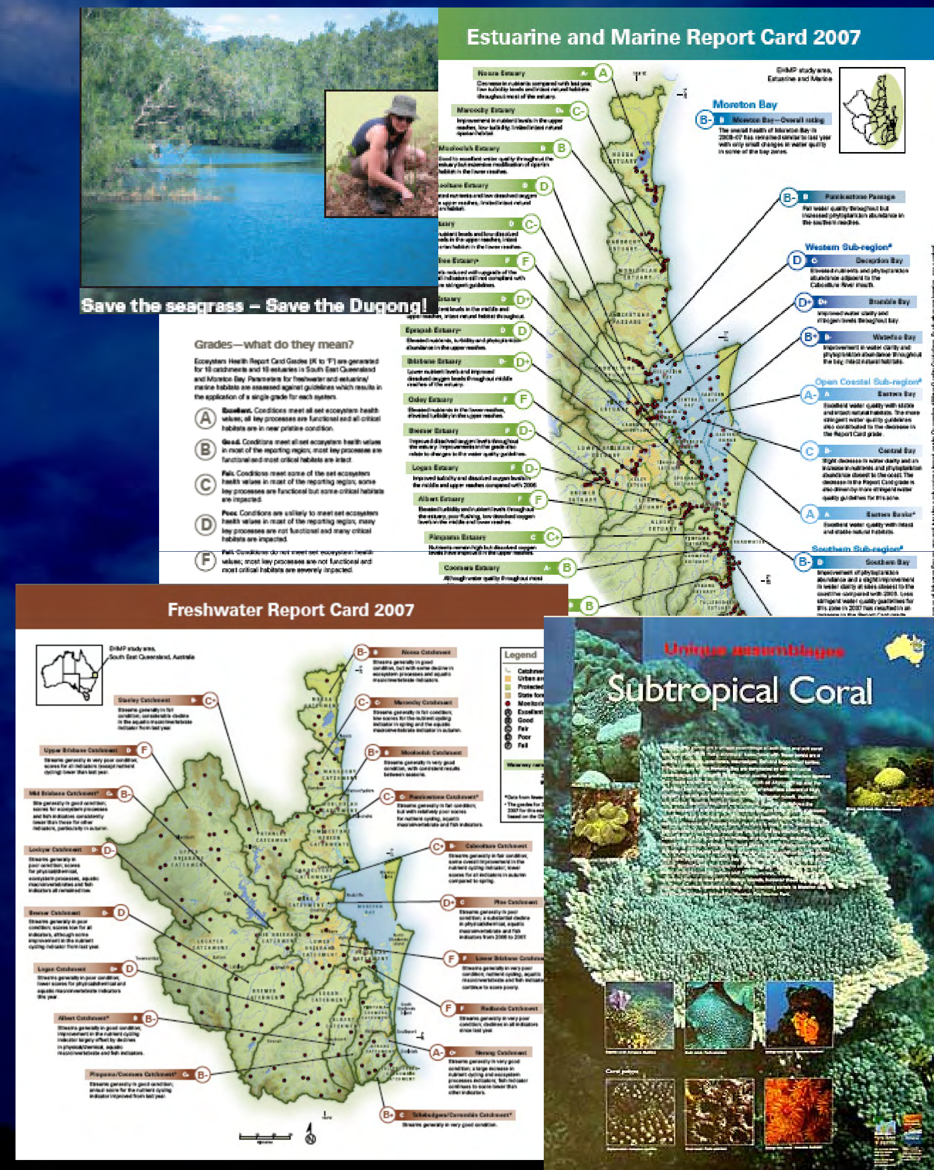
Microsoft eScience

March 2008

Healthy Waterways
(UQ Eva Abal)

- Queensland is experiencing a water crisis due to drought, expanding population, and estuarine reparation
- Healthy Waterways includes 19 local governments, 6 state agencies, 5 universities and key industrial partners
- Data processing hampered by lack of tools and ontologies (to merge data from heterogeneous sources).
  - Report cards shown currently take 3 months to compile.
  - Planned sensor deployments will only make this worse.

http://www.healthywaterways.org



Northern California Digital Watersheds

(BWC, James Hunt)

- Russian River watershed challenges: forestry, farming, urbanization, gravel mining, and fish habitat restoration.
  - Can we understand historic and on-going changes using only publically available data sources such as USGS, NOAA, Sonoma Ecology Center, etc?
- Early studies examined overall water balance and changes in suspended sediment
  - scientific data “mashups” are leading to useful results.
- Recent engagement with National Marine Fisheries and USBR expanding this to other watersheds across Northern California

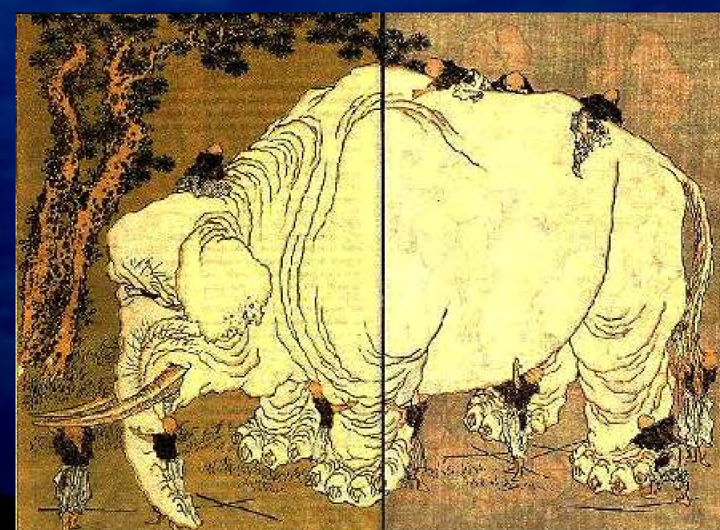
“We see water through a fish eye lens”

http://bwc.berkeley.edu/California

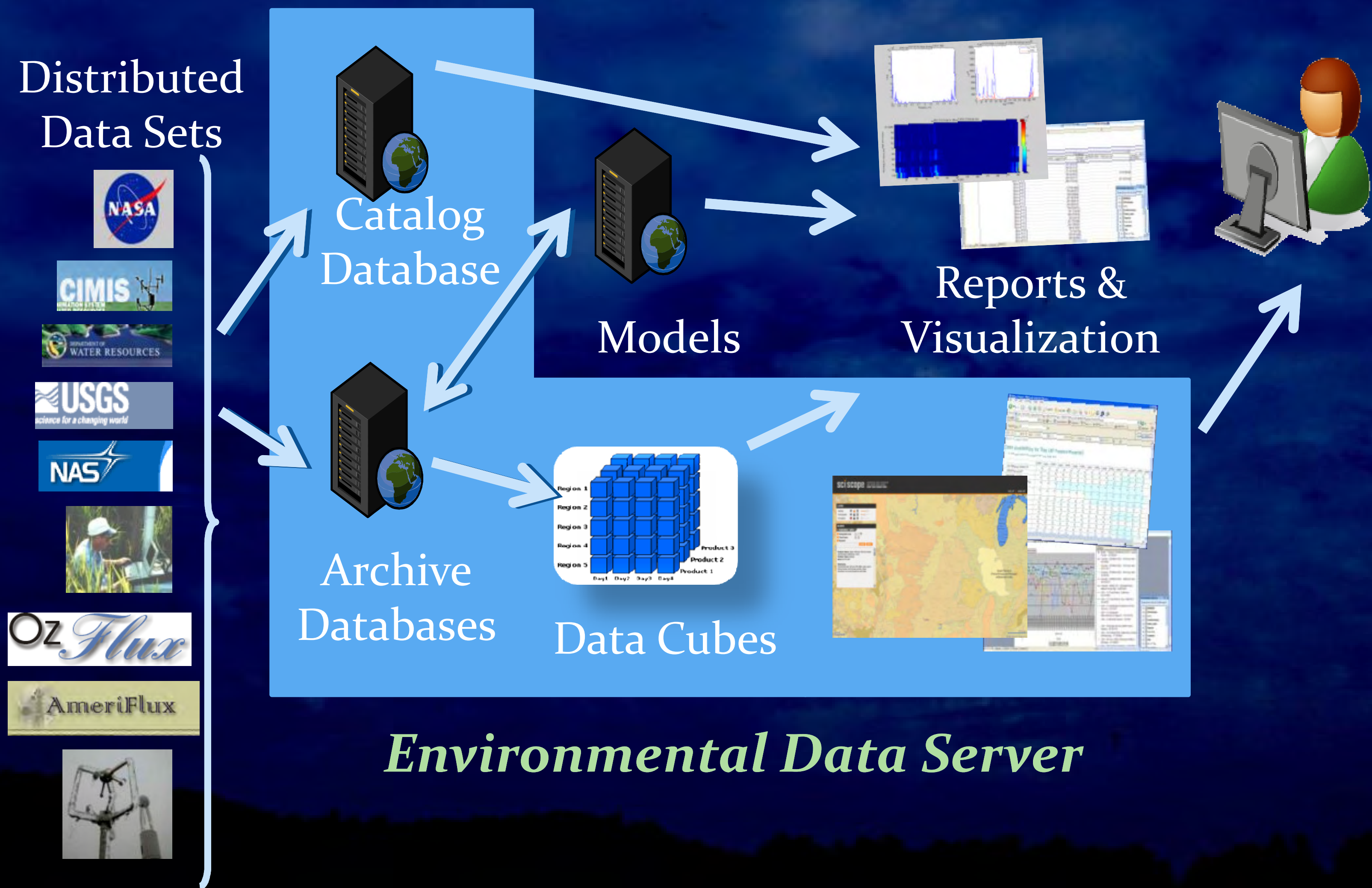
The Avalanche/Landslide/Tsunami

- The era of remote sensing, cheap ground-based sensors and web service access to agency repositories is here

- Extracting and deriving the data needed for the science remains problematic
  - Specialized knowledge
  - Finding the right needle in the haystack



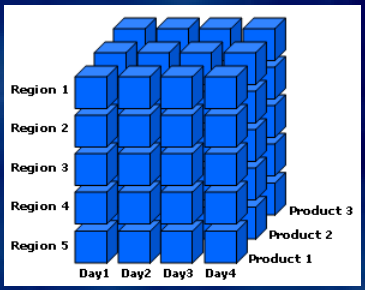
Connecting Data, Resources, and People



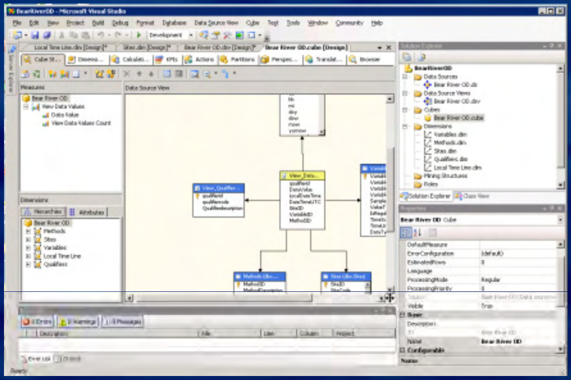
Ecological Data Analysis is Like Making Sausage

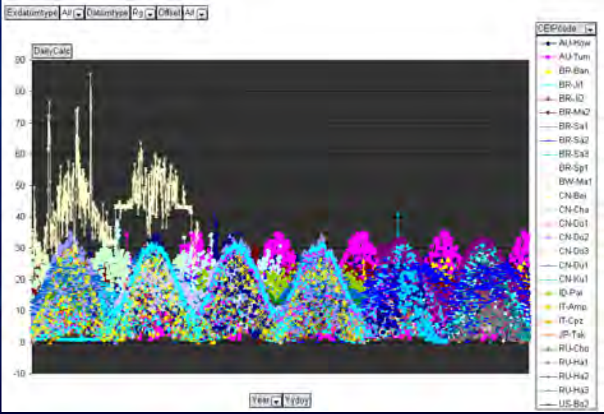
- A lot of different bits have to come together
  - Multiple entities fly satellites and operate fixed sensors – imagery from NASA, precipitation from NOAA, discharge from USGS, water quality from EPA
  - Collecting site properties such as channel cross-section or leaf area index done by individuals as well as agencies
- You don't always know where the bits have been
  - passed around in e-mail
  - edited at will
  - silently corrected in repositories because "everybody knows"
- Spice matters more than you'd think
  - Published literature supplies various numbers and graphs giving context, sanity checks on the results

Data Cube Basics



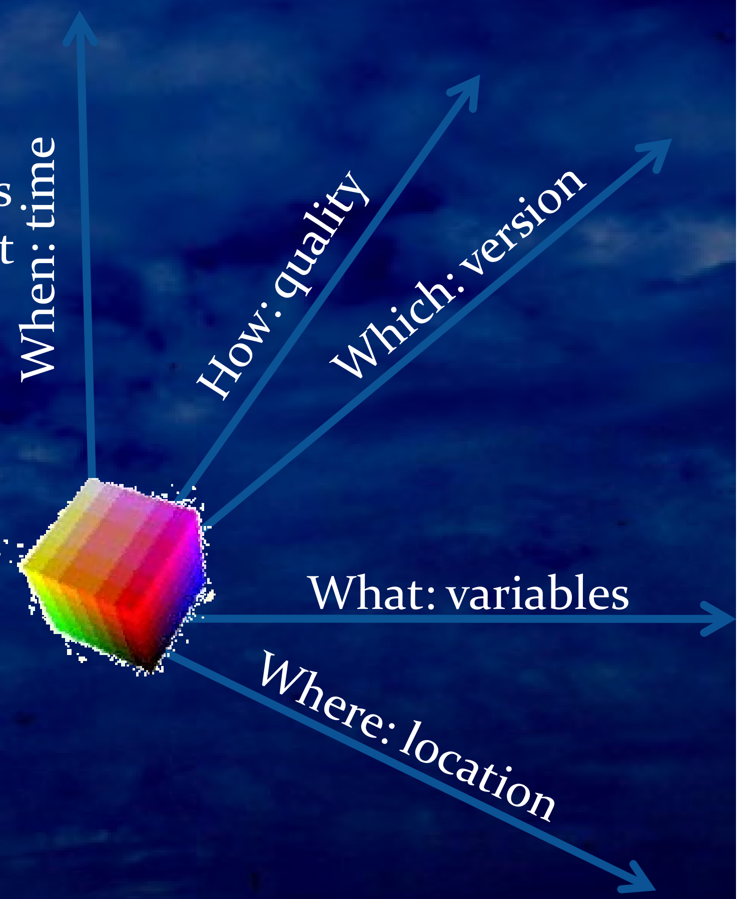
- A data cube is a database specifically for data mining (OLAP)
  - Initially developed for commercial needs like tracking sales of Oreos and milk
  - Simple aggregations (sum, min, or max) can be pre-computed for speed
  - Hierarchies for simple filtering with drilldown capability
  - Additional calculations (median) can be computed dynamically or pre-computed
  - All operate along dimensions such as time, site, or datumtype
  - Constructed from a relational database
  - A specialized query language (MDX) is used
- Client tool integration is evolving
  - Excel PivotTables allow simple data viewing
  - More powerful analysis and plotting using Matlab and statistics software



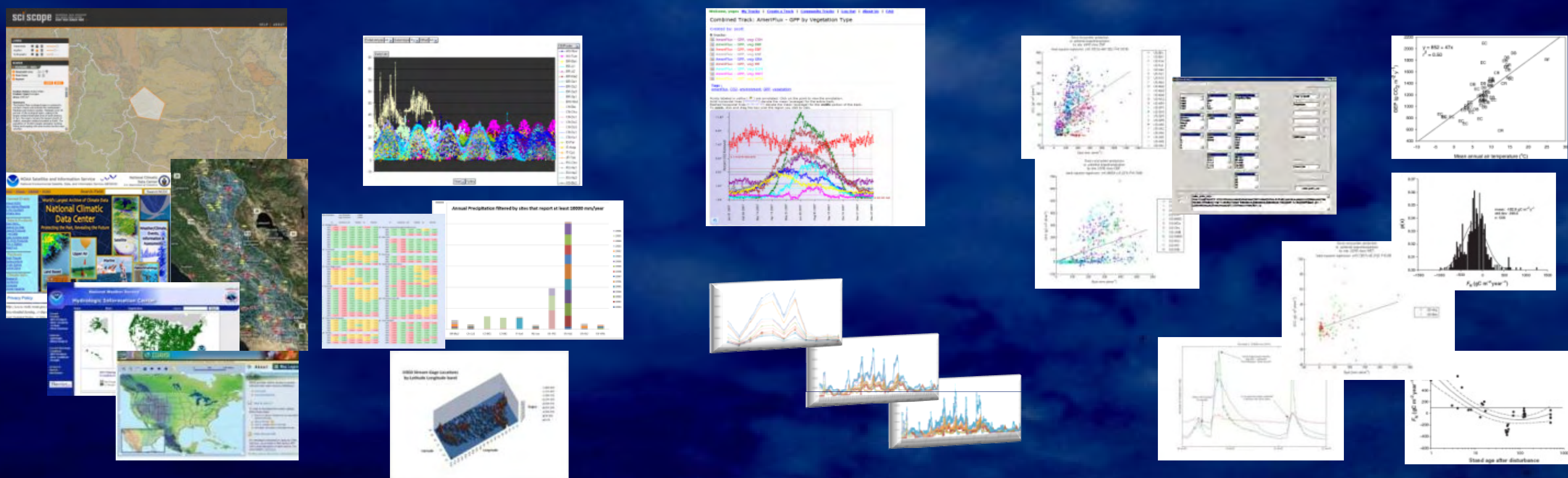


Data Cube Dimensions

- Common dimensions driven by the nature of the data
  - What: variables
  - When: time, time, time
  - Where: (x, y, z) location or attribute where (x,y) is the site location and (z) is the vertical elevation at the site.
  - Which: versioning and other collections
  - How: gap filling and other data quality metrics
- Common pre-computed and computed members driven by the nature of the analyses
  - Max, Min, Count: pre-computed
  - hasDataRatio or gapPercent: fraction of data actually present or missing
  - DailyCalc: average, sum or maximum depending on variable; includes units conversion
  - YearlyCalc: similar to DailyCalc
  - RMS or sigma: standard deviation or variance for fast error or spread viewing



The Data Pipeline



Data Gathering

"Raw" data includes sensor output, data downloaded from agency or collaboration web sites, papers (especially for ancillary data

Discovery and Browsing

"Raw" data browsing for discovery (do I have enough data in the right places?), cleaning (do data look wrong?), and light weight science via browsing

Science Exploration

"Science variables" and data summaries for early exploration and hypothesis testing. Similar to discovery and browsing, but with science variables computed via gap filling, units conversion or simple equation.

Domain analyses

"Science variables" combined with models, other specialized code, or statistics for deep science understanding.

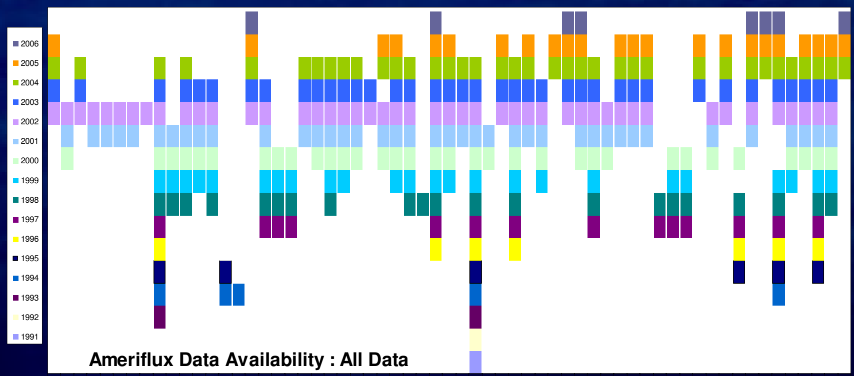
Scientific Output

Scientific results via packages such as MatLab or R2. Special rendering package such as ArcGIS.

Paper preparation!

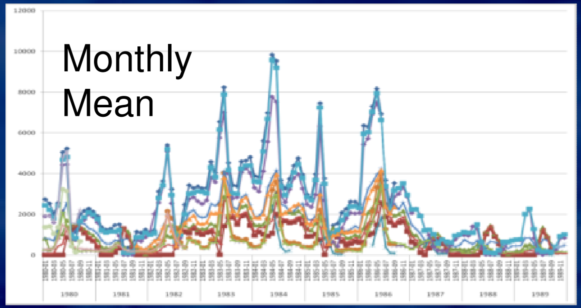
Browsing for Data Availability

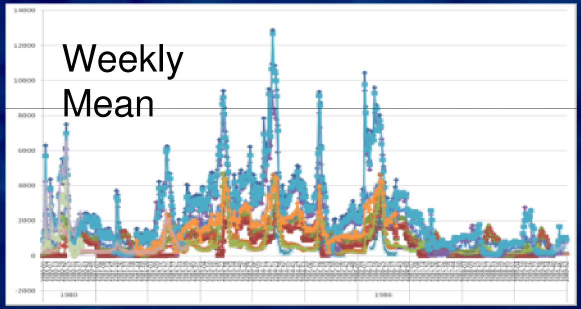
- Summary data products (yearly min/max/avg) almost trivially
- Simple mashups and data cubes aid discovery of available data
- Simple Excel graphics show cross-site comparisons and availability filtered by one variable or another

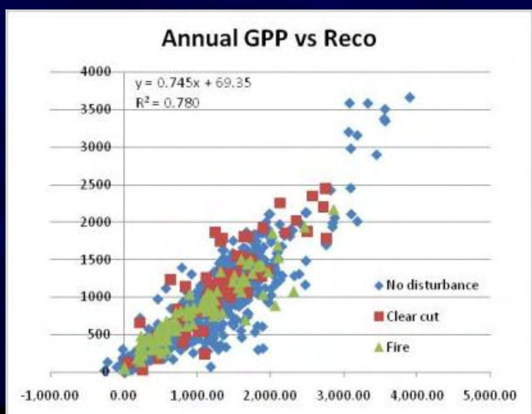


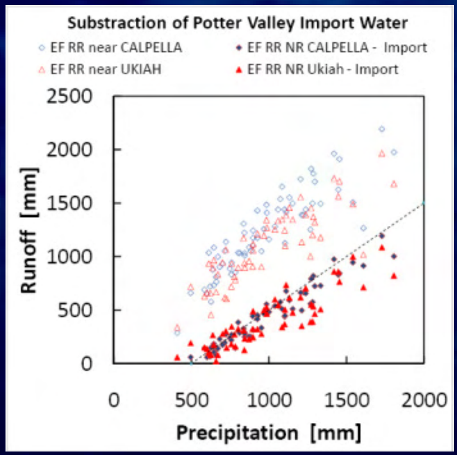
Browsing for Data Analysis

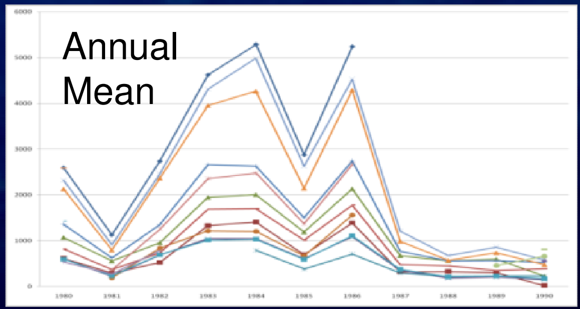
- Plotting is the way of visualizing data
  - Most are discarded so scripting matters











The conclusions:
Now it's really all about the apps (and DATA)!

1. Moore's Law is alive and well… IP is the most likely the platform and Ethernet / WiFi is likely to be the dominant network
2. It's time to deploy … vs infrastructure papers & marginal hardware.
   – WSN research space radios, power, protocols, standards, etc.… being mined.
   – The industry has formed and is slowly evolving.
   – Rolling your own motes may be a win if the Fleck vendor succeeds
3. It's the apps. Science or engineering apps --- sense, deploy, data. Inevitably large scale databases, etc. and on to control apps.
4. Support Fleck, G2Micrsystems and Alive.com home teams. CSIRO apps can understand the limits in <u>your environments</u>.
5. Dust Networks – single chip platform (protocol: 3D's redundancy)
   – Power, 99.99% rel., no powered nodes, "works" vs. Zigbee "make work"
6. WW Databases: NEON, MSR Sensornet, SkyServer, Env. Data Server

The end