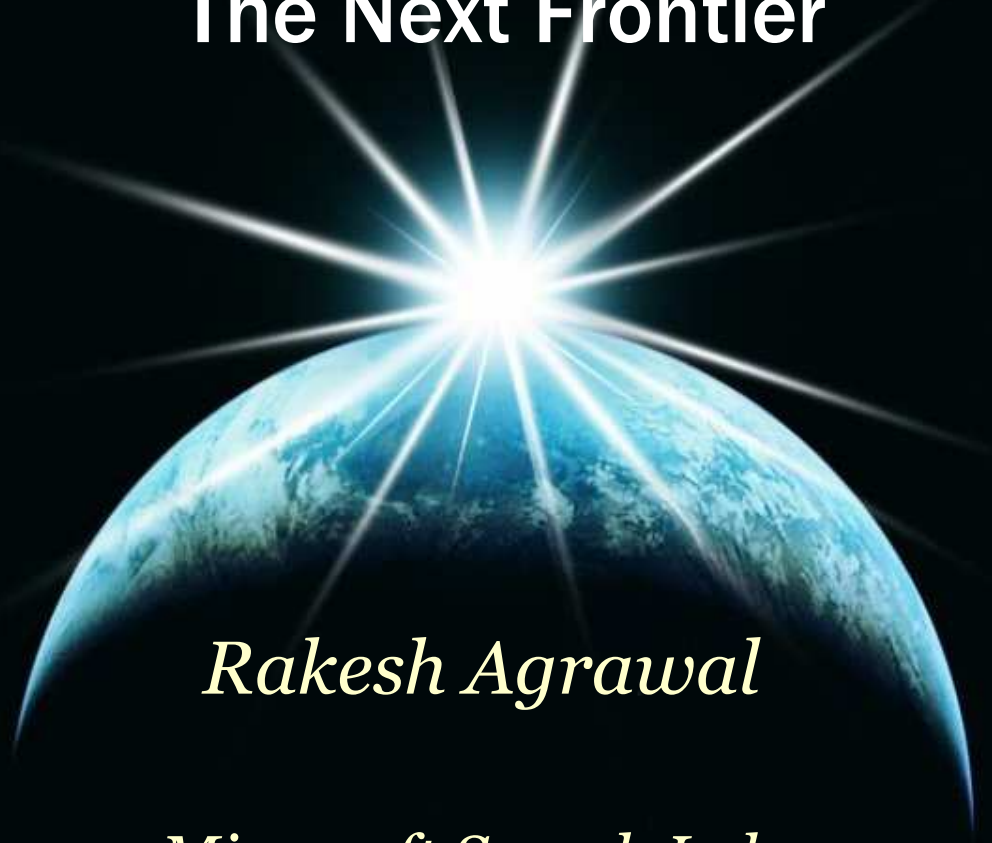
# Humane Data Mining: The Next Frontier



*Rakesh Agrawal*

*Microsoft Search Labs*
*Mountain View, CA*


msn

# Central Message

- Data Mining has made tremendous strides in the last decade
- It's time to take data mining to the next level of contributions
- We will need to expand our view of who we are and develop new abstractions, algorithms and systems, inspired by new applications

# Outline

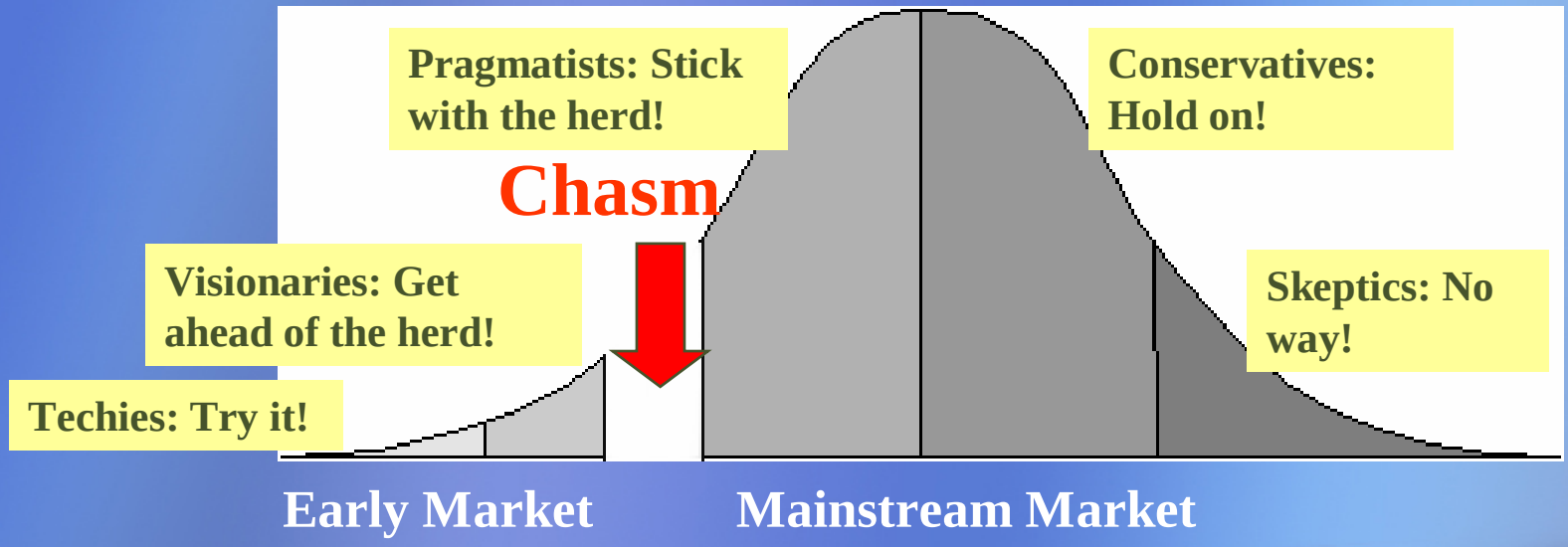
- Retrospective on KDD-99 Keynote - “Data Mining: Crossing the Chasm”
- Developments since then
- New Frontier

# Outline

- Retrospective on KDD-99 Keynote - "Data Mining: Crossing the Chasm"
- Developments since then
- New Frontier

# Data Mining: Crossing the Chasm* (Circa 1999)

Thesis: The greatest challenge facing data mining is to make the transition from being an early market technology to mainstream technology.




*Geoffrey A Moore. Crossing the Chasm. Harper Business. 1991.

# Backdrop: Quest Experience

- Started as skunk work in IBM Almaden in early nineties
- Inspired by needs articulated by industry visionaries
- New abstractions, technologies
- IBM Intelligent Miner (Circa 1996)
  - Serious product
  - Fast, scalable, multiple platforms (including SP2)
  - “Early market” successes
- By end of 1997: Intelligent Miner seen as creating a new software category
- But then phones stopped ringing!

# Imperatives for Chasm Crossing (Circa 1999)

- Data Mining Standards
- Data Mining Benchmarks
- Auto-focus Data Mining
- Database Integration
- Web: Greatest Opportunity
- Personalization
- Watch for Privacy Pitfall

# Outline

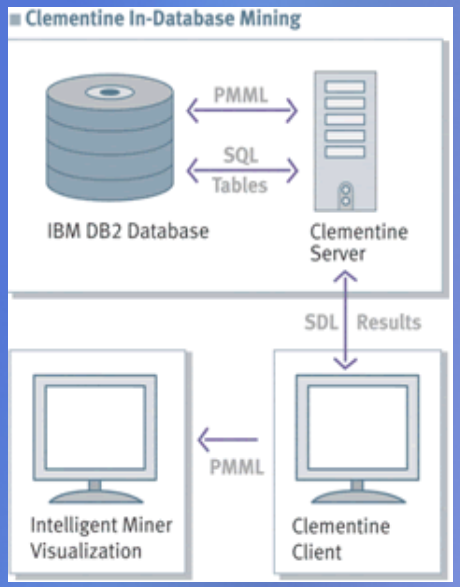
- Retrospective on KDD-99 Keynote - "Data Mining: Crossing the Chasm"
- Developments since '99
- New Frontier

# Scorecard (Circa 2006)

| | | |
|---|---|---|
| Data Mining Standards | → | PMML/CRISP |
| Data Mining Benchmarks | → | KDD Cups? |
| Auto-focus Data Mining | → | Embedded in Solutions |
| Database Integration | → | Commercial Offerings |
| Web | → | Under-estimated Importance |
| Personalization | → | Nascent |
| Privacy Pitfall | → | Privacy-Preserving Data Mining |

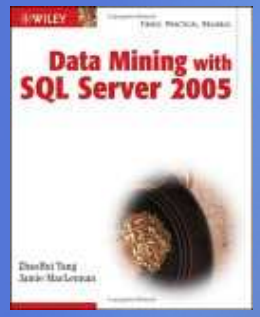
# PMML: Predictive Model Markup Language

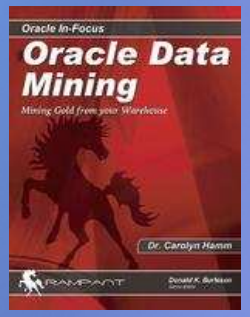
- Markup language for sharing models between applications (mine rules with one application; use a different application to visualize, analyze, evaluate or otherwise use the discovered rules).




```
<AssociationModel functionName="associationRules“...">
...
<Item id="1" value=“Diabetes" />
...
<Itemset id="3" support="1.0" numberOfItems="2">
   <ItemRef itemRef="1" />   <ItemRef itemRef="3" />
</Itemset>
...
<AssociationRule support="1.0" confidence="1.0"
    antecedent="1" consequent="2" />
...
```

# Database Integration

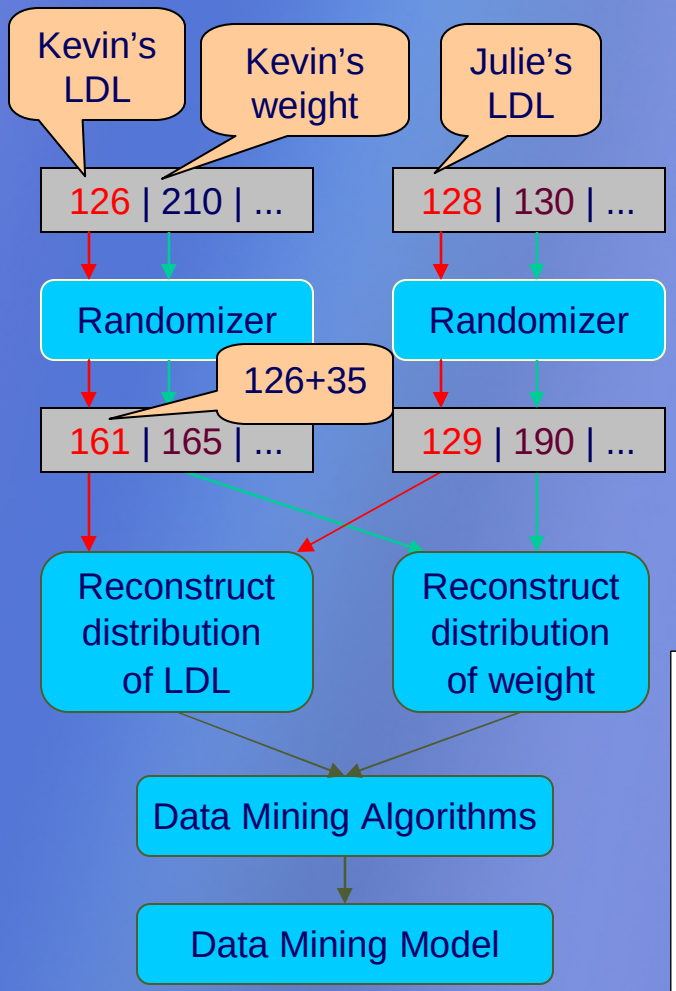



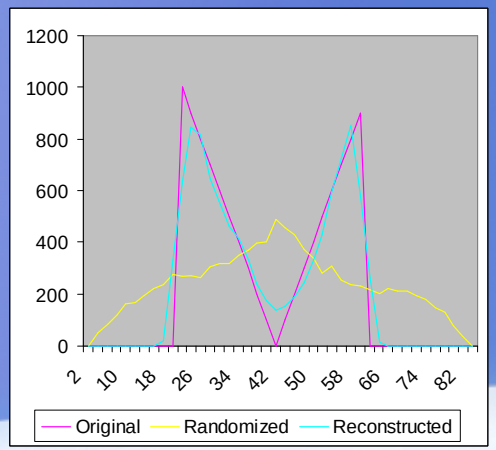


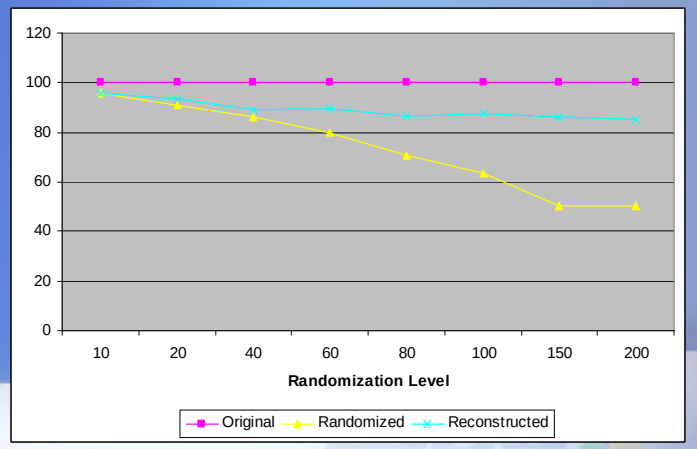
- Tight coupling through user-defined functions and stored procedures
- Use of SQL to express data mining operations
  - Composability: Combine selections and projections
  - Object-relational extensions enhance performance
  - Benefit of database query optimization and parallelism carry over
- SQL extensions

# Privacy Preserving Data Mining



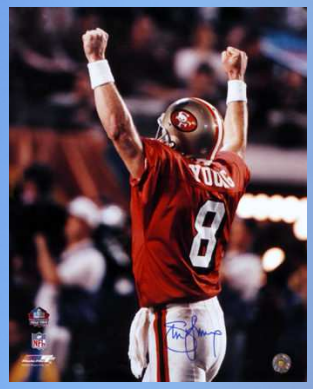

- Preserves privacy at the individual patient level, but allows accurate data mining models to be constructed at the aggregate level.
- Adds random noise to individual values to protect patient privacy.
- EM algorithm estimates original distribution of values given randomized values + randomization function.
- Algorithms for building classification models and discovering association rules on top of privacy-preserved data with only small loss of accuracy.




Sigmod00, KDD02, Sigmod05

# Enterprise Applications Galore!



- Example: SAS Customer Successes

**Customer Relationship Management**
Claims Prediction | Credit Scoring | Cross-Sell/Up-Sell | Customer Retention | Marketing Automation | Marketing Optimization | Segmentation Management | Strategic Enrollment Management

**Drug Development**

**Financial Management**
Activity-Based Management | Fraud Detection

**Human Capital Management**

**Information Technology Management**
Charge Management | Resource Management | Service Level Management | Value Management

**Performance Management**
Balanced Score-carding

**Quality Improvement**

**Regulatory Compliance**
Fair Banking

**Risk Management**

**Supplier Relationship Management**

**Supply Chain Analysis**
Demand Planning | Warranty Analysis

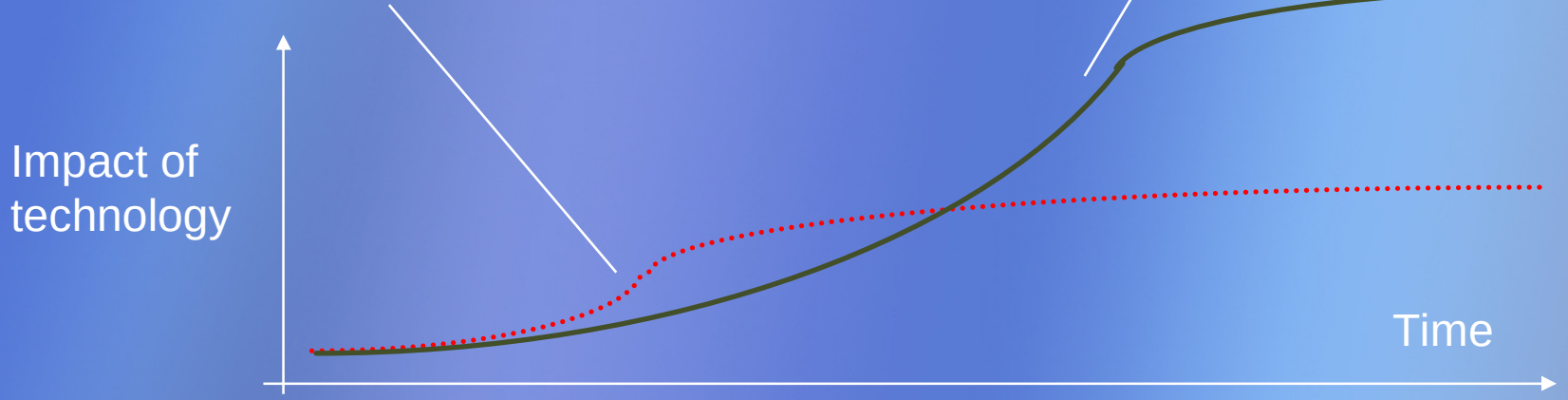
**Web Analytics**

http://www.sas.com/success/solution.html

# Some Surprises

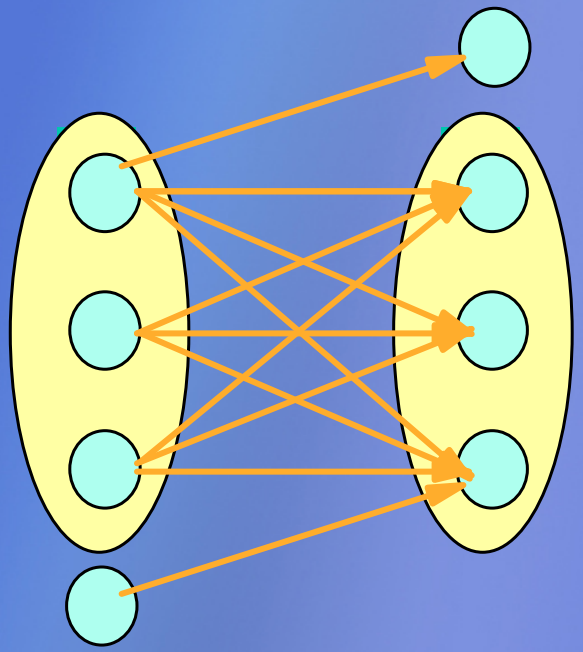
**Popular technology visions often overestimate near-term prospects...**

**...but they underestimate long-term developments.**




SRI Consulting Business Intelligence (Ray Amara)

# Discovering Online Micro-communities



complete 3-3 bipartite graph

- Japanese elementary schools
- Turkish student associations
- Oil spills off the coast of Japan
- Australian fire brigades
- Aviation/aircraft vendors
- Guitar manufacturers

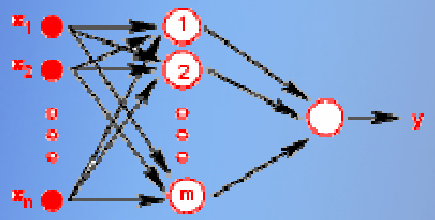
Frequently co-cited pages are related.

Pages with large bibliographic overlap are related.

Use of a variant of Apriori for the discovery.

R Kumar et al., “Trawling the web for emerging cyber-communities”, WWW 99.

# Ranking Search Results in MSN

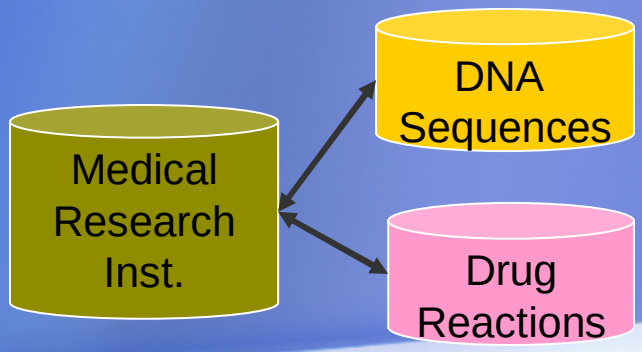


- Search results ranked dynamically by a neural net .
- Ranking function learnt using a gradient descent method.
- Training data: Some query/document pairs labeled for relevance (excellent, good, etc.).
- Feature set: query independent features (e.g. static page rank) plus query dependent features extracted from the query combined with additional sources (e.g. anchor text).
- Best net selected by computing NDCG metric on a validation set.

Burges et al. “Learning to rank using gradient descent”, ICML 05.

# Sovereign Information Integration

- Separate databases due to statutory, competitive, or security reasons.
  - Selective, minimal sharing on a need-to-know basis.
- Example: Among those patients who took a particular drug, how many with a specified DNA sequence had an adverse reaction?
  - Researchers must not learn anything beyond counts.
- Algorithms for computing joins and join counts while revealing minimal additional information.




## Minimal Necessary Sharing

| R |
|---|
| a |
| u |
| v |
| x |

| S |
|---|
| b |
| u |
| v |
| y |

**R ❁ S**

- R must not know that S has **b** and **y**
- S must not know that R has **a** and **x**

R ❁ S

| u |
|---|
| v |

**Count (R ❁ S)**

- R and S do not learn anything except that the result is 2.

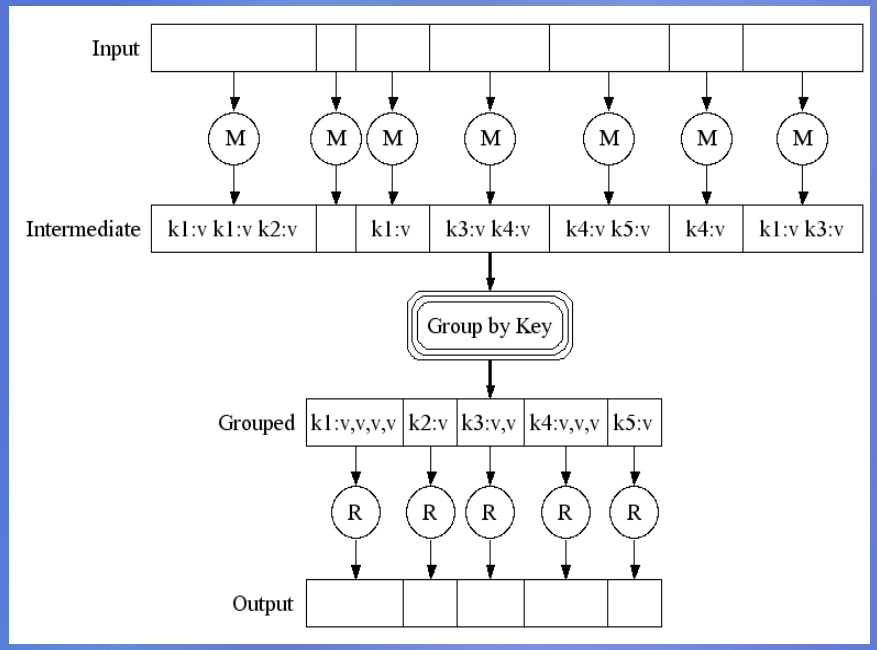
Sigmod 03, DIVO 04

# Google's Data Mining Platform

**MapReduce[1]:** Programming Model

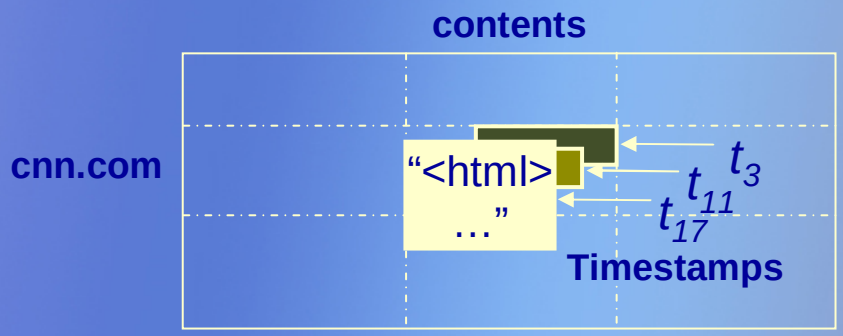
map(ikey, ival) -> list(okey, tval)

reduce(okey, list(tval)) -> list(oval)




- Automatic parallelization & distribution over 1000s of CPUs
- Log mining, index construction, etc

**BigTable[2]:** Distributed, persistent, multi-level sparse sorted map



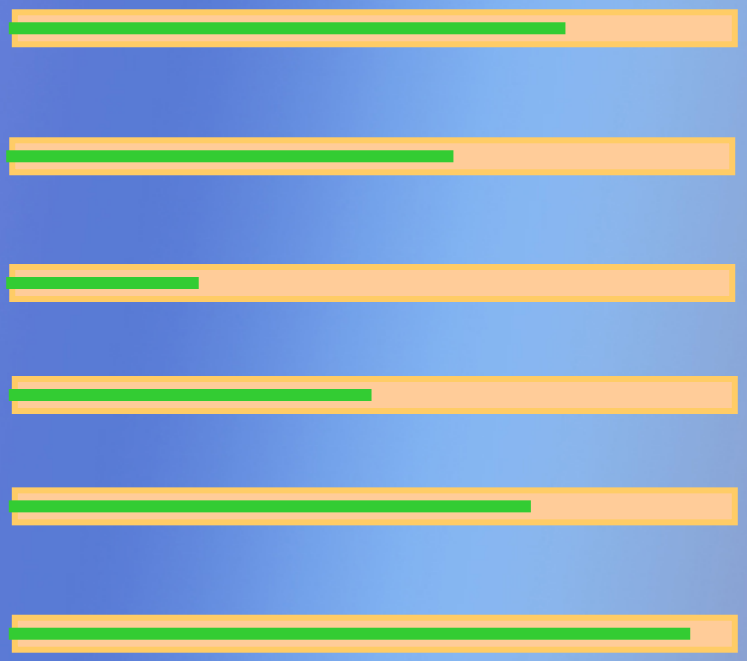

- Tablets, Column family
- >400 Bigtable instances
- Largest manages >300TB, >10B rows, several thousand machines, millions of ops/sec
- Built on top of GFS

[1]Dean et. al. "MapReduce: Simplified data processing on large clusters", OSDI 04.

[2]Hsieh. "BigTable: A distributed storage system for structured data", Sigmod 06.

# A Snapshot of Progress

- Algorithmic innovations
- System support
- Foundations
- Usability
- Enterprise applications
- Unanticipated applications

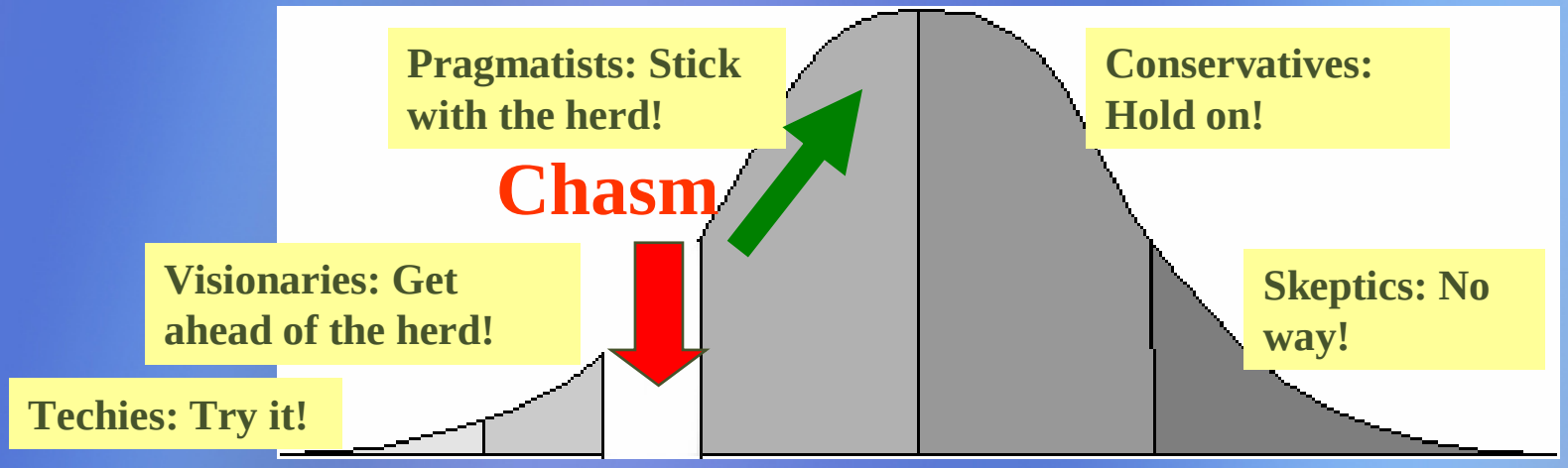


# Have we crossed the chasm?

Yes Dorothy!

Whereto now?

# Imperative Circa 2006




Maintain upward trajectory (and escape withering):

- Focus on a new class of applications, bringing into fold techies and visionaries, leading to new inventions and markets
- While continuing to innovate for the current mainstream market

# Outline

- Retrospective on KDD-99 Keynote - "Data Mining: Crossing the Chasm"
- Developments since '99
- New frontier

# Humane Data Mining

"Is it right? Is it just?
Is it in the interest of mankind?"
*Woodrow Wilson.* May 30, 1919.

## *Applications to Benefit Individuals*

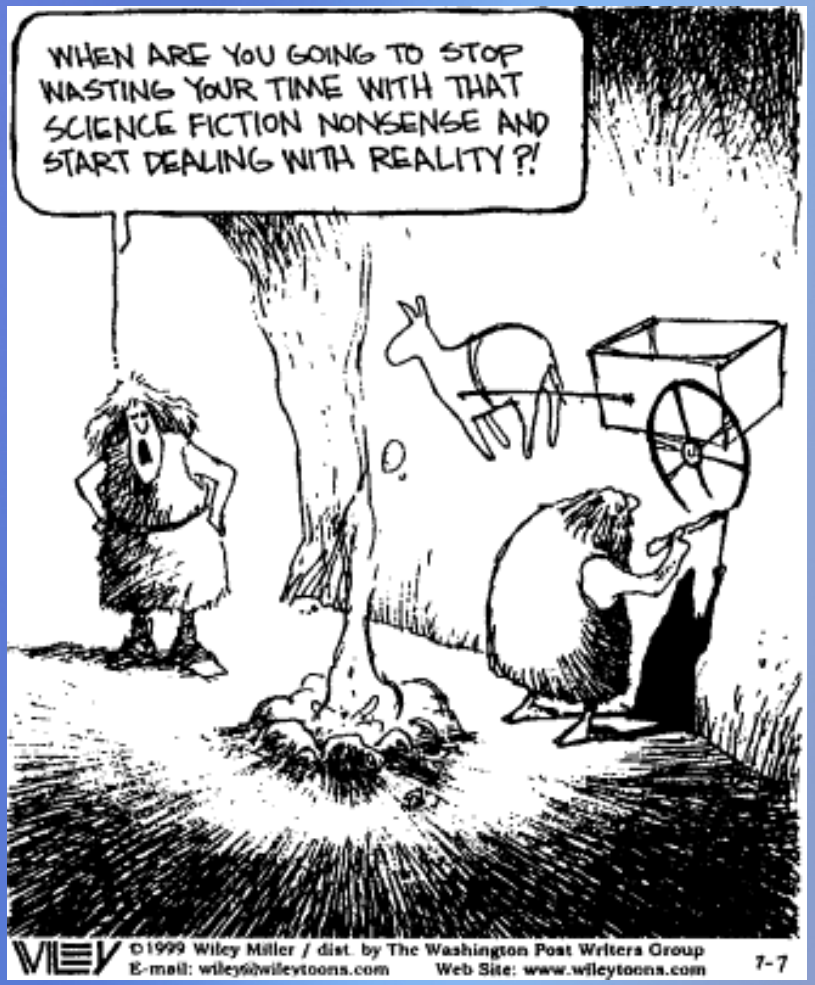
Rooting our future work in this class of new applications, will lead to new abstractions, algorithms, and systems

# An Expansive Definition of Data Mining

- Deriving value from a data collection by studying and understanding the structure of the constituent data

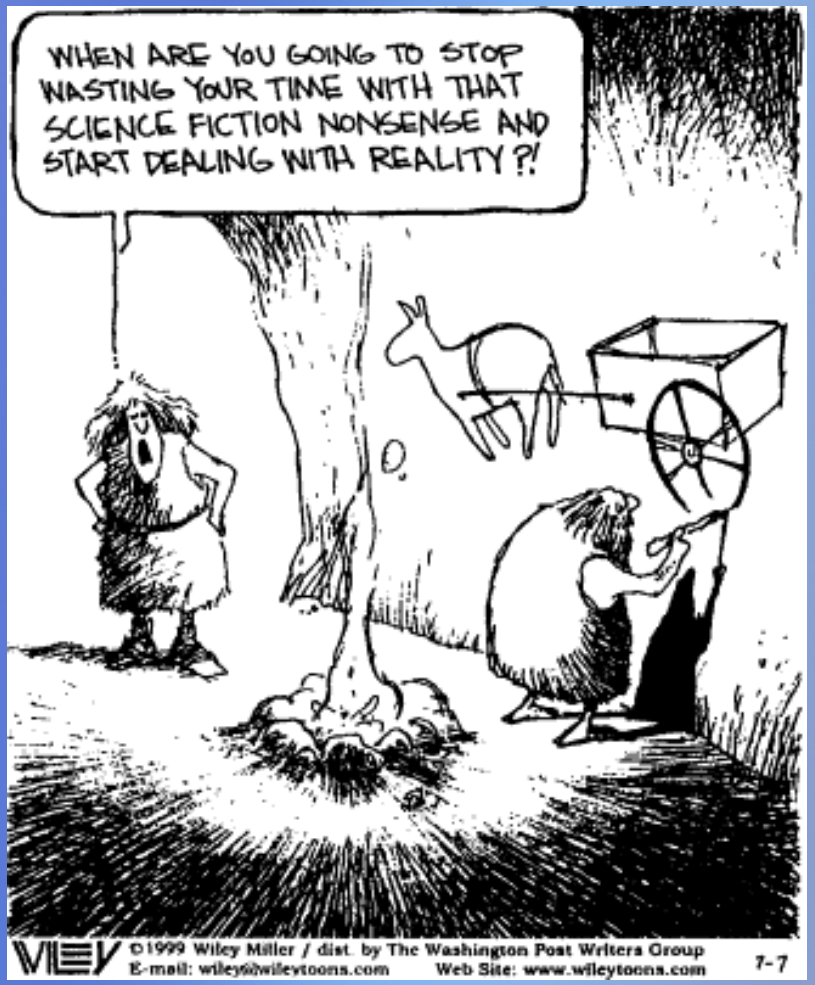
# Some Ideas

- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science

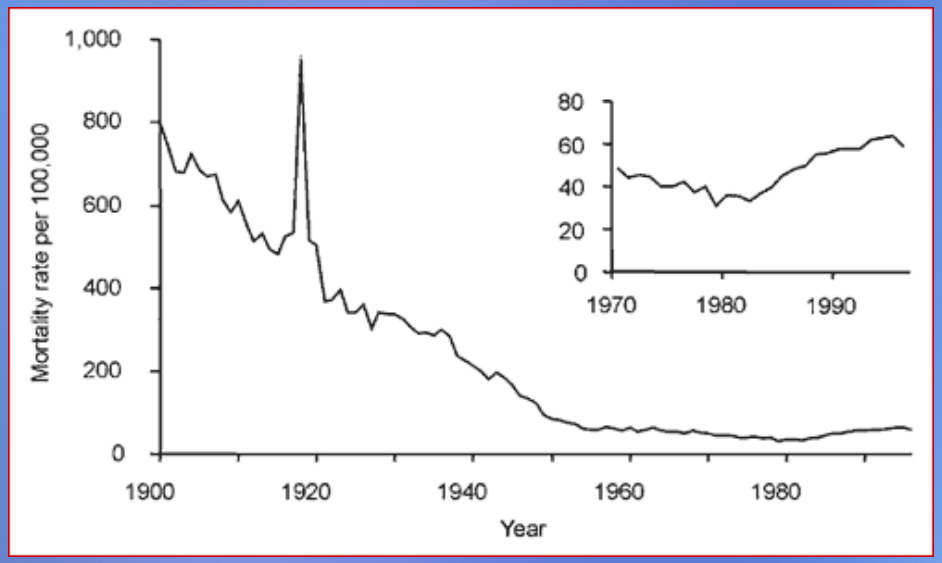


# Some Ideas

- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



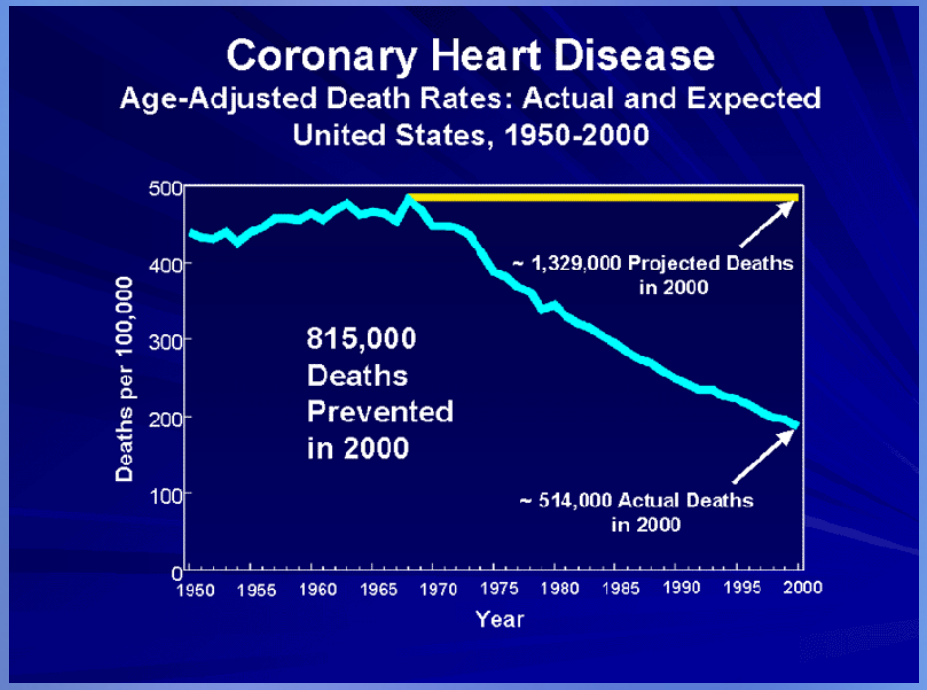
# Changing Nature of Disease




CDC

- Leading causes of death in early 20th century: Infectious diseases (e.g. tuberculosis, pneumonia, influenza)
- By the 1950s, infectious diseases greatly diminished because of better public health (sanitation, nutrition, etc.)

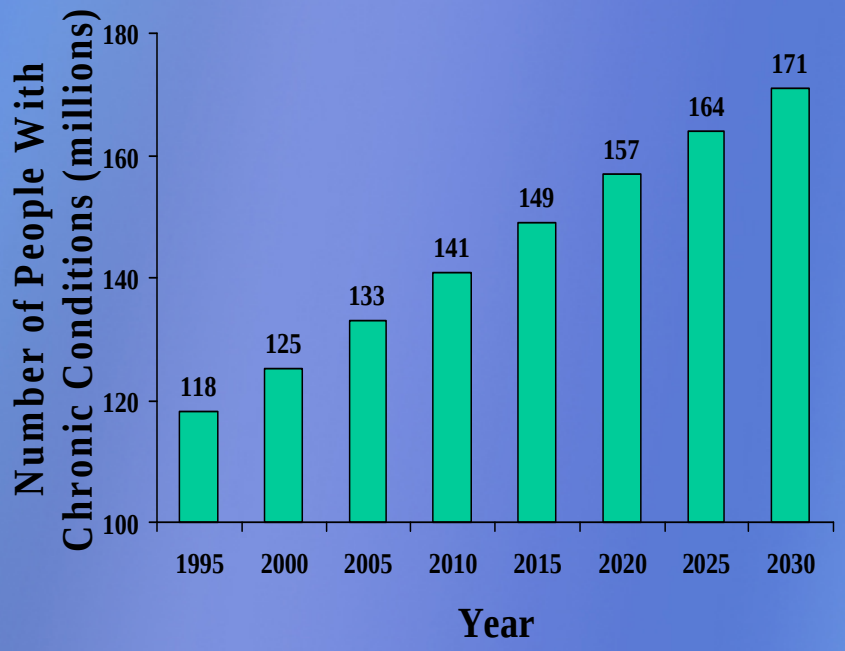
# Changing Nature of Disease




NIH

- Since 50's, treating acute illness (e.g. heart attacks, strokes) has become the focus.
- Proficiency of the current medical system in delivering episodic care has made acute episodes into survivable events.

# Changing Nature of Disease




Partnership for Solutions

- New challenge: chronic conditions: illnesses and impairments expected to last a year or more, limit what one can do and may require ongoing care.
- In 2005, 133 million Americans lived with a chronic condition (up from 118 million in 1995).

# Technology Trends

- Dramatic reduction in the cost and form factor for personal storage



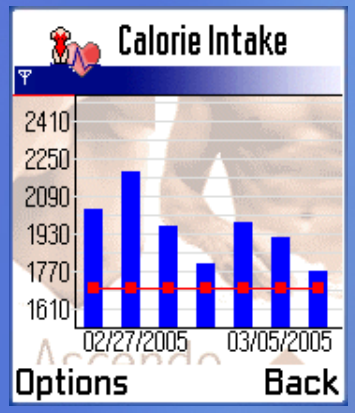

- Tremendous simplification in the technologies for capturing useful personal information

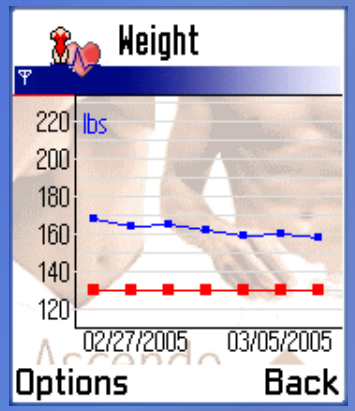


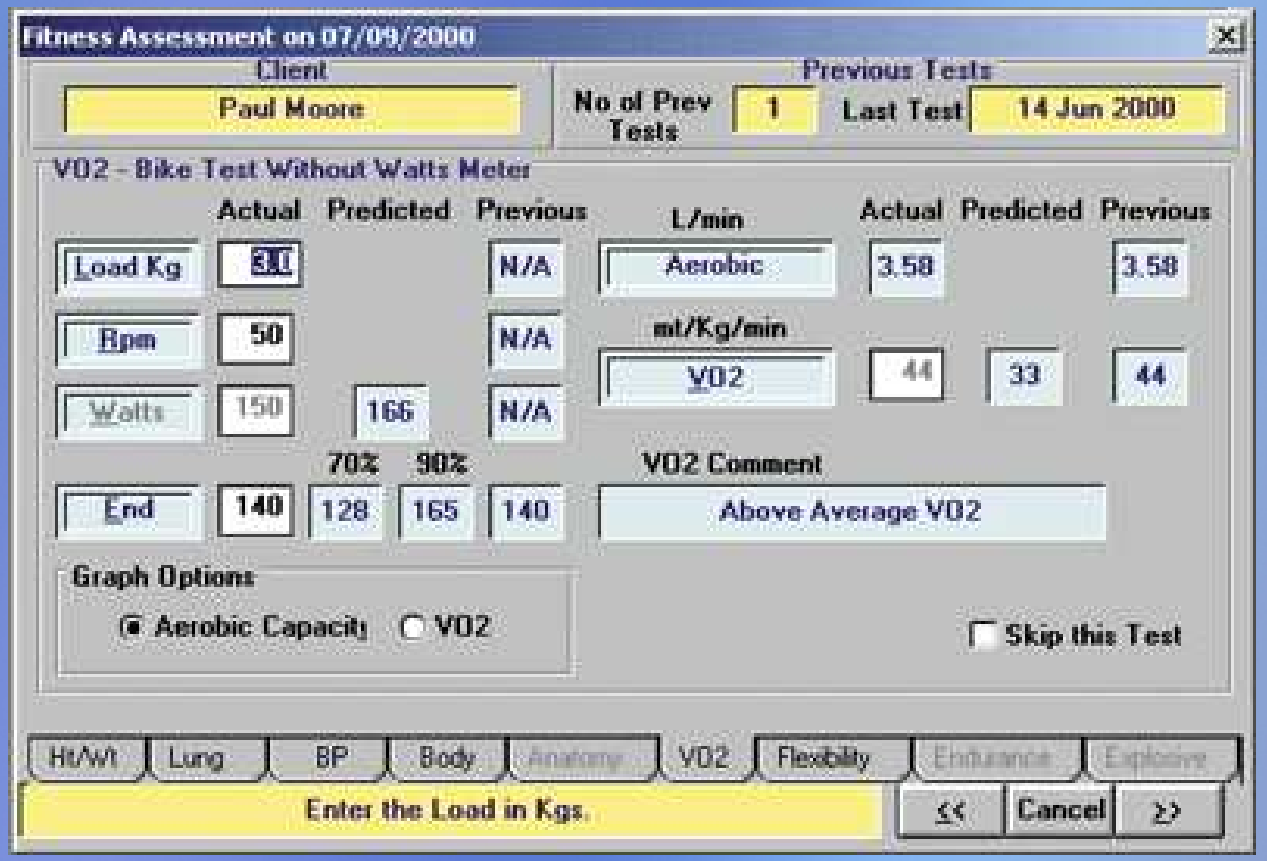





# Personal Health Analytics



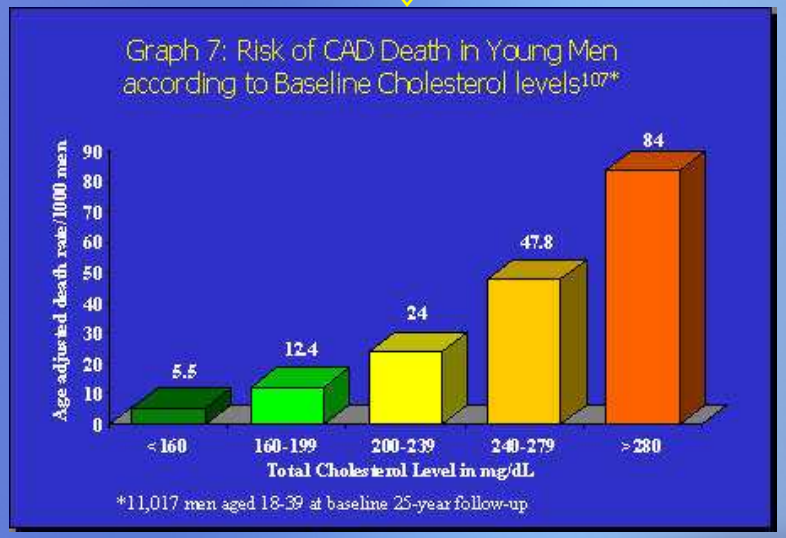



# Personal Data Mining


Charts for appropriate demographics?

Optimum level for Asian Indians: 150 mg/dL
(much lower than 200 mg/dL for Westerners)
Due to elevated levels of lipoprotein(a)*

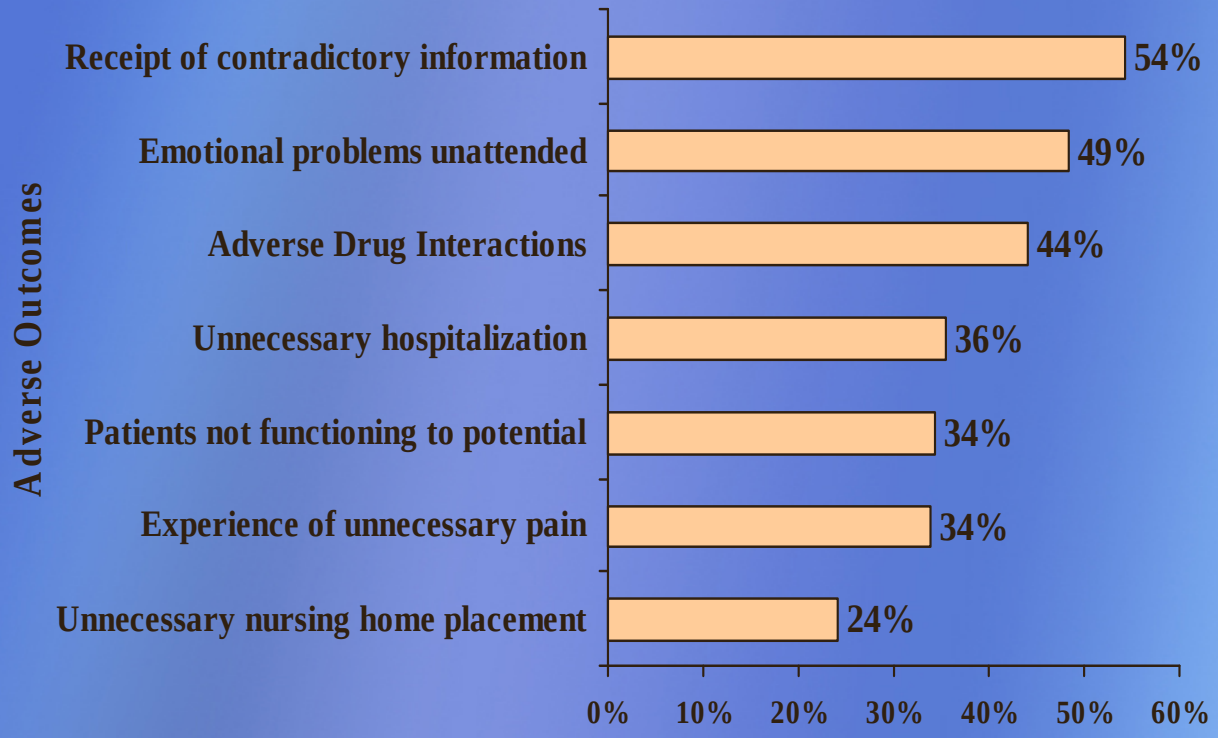



Distributed computation and selection across millions of nodes

Privacy and security

*Enas et al. Coronary Artery Disease In Asian Indians. *Internet J. Cardiology.* 2001.

# The Patient's Dilemma

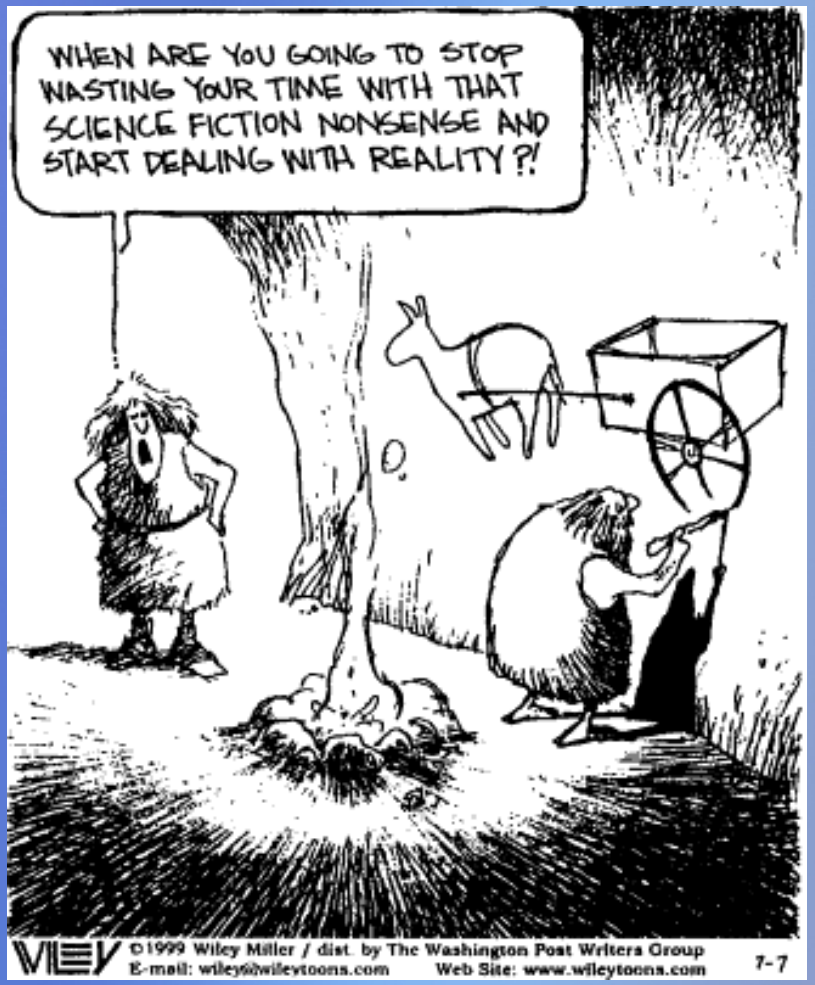




Percent of Physicians Who Believe that Adverse Outcomes Result from Poor Care Coordination

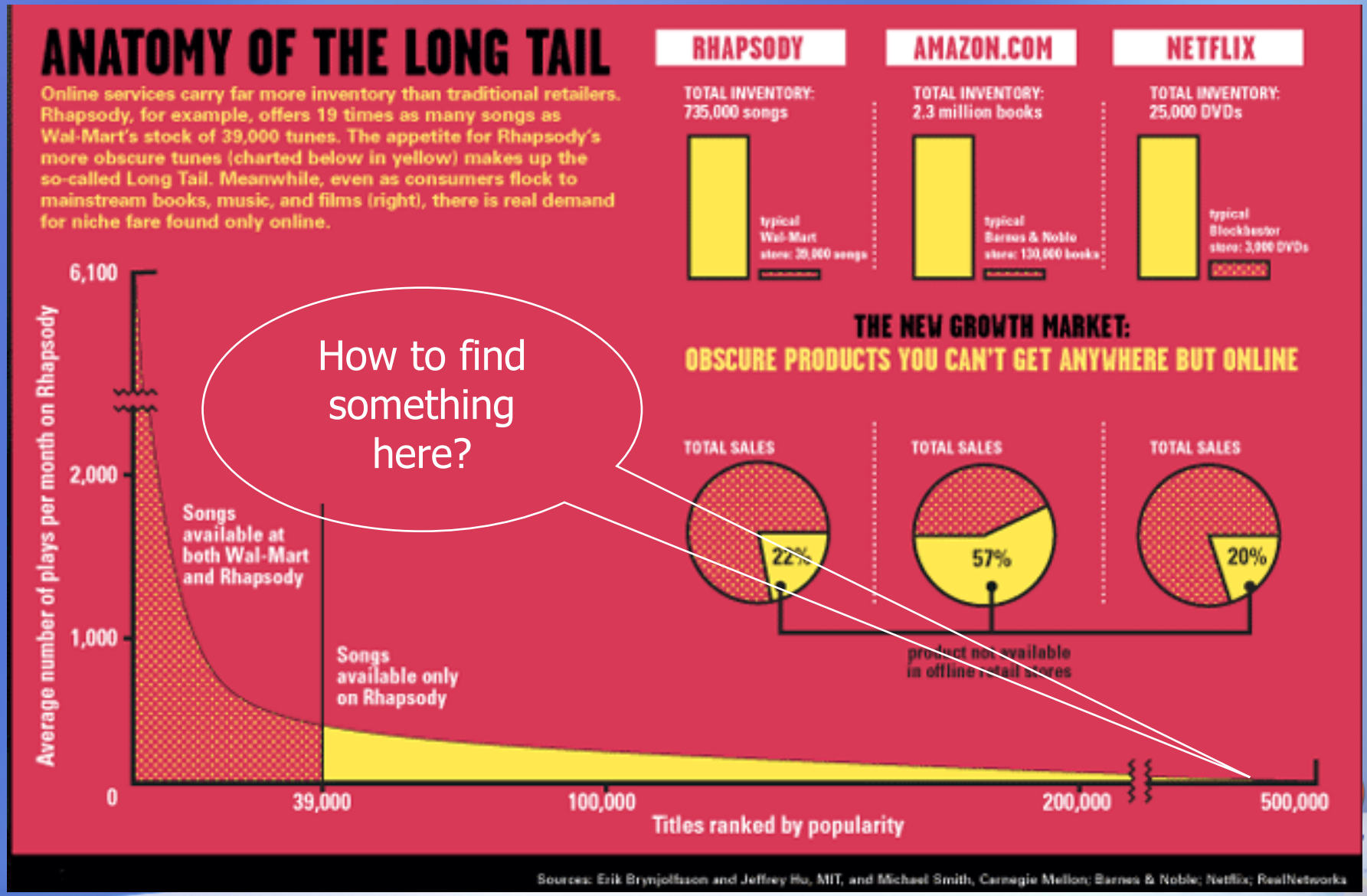
Partnership for Solutions

# Some Ideas

- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



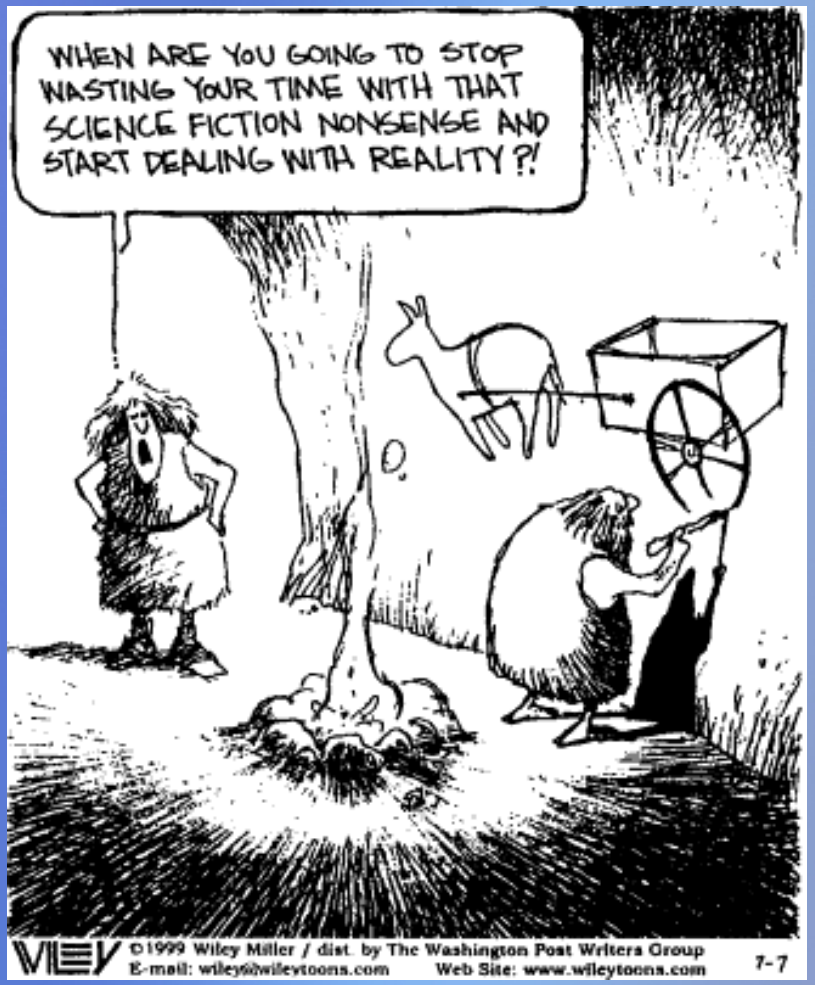
# The Tyranny of Choice




Chris Anderson. The Long Tail. 2006.

# Some Ideas

- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



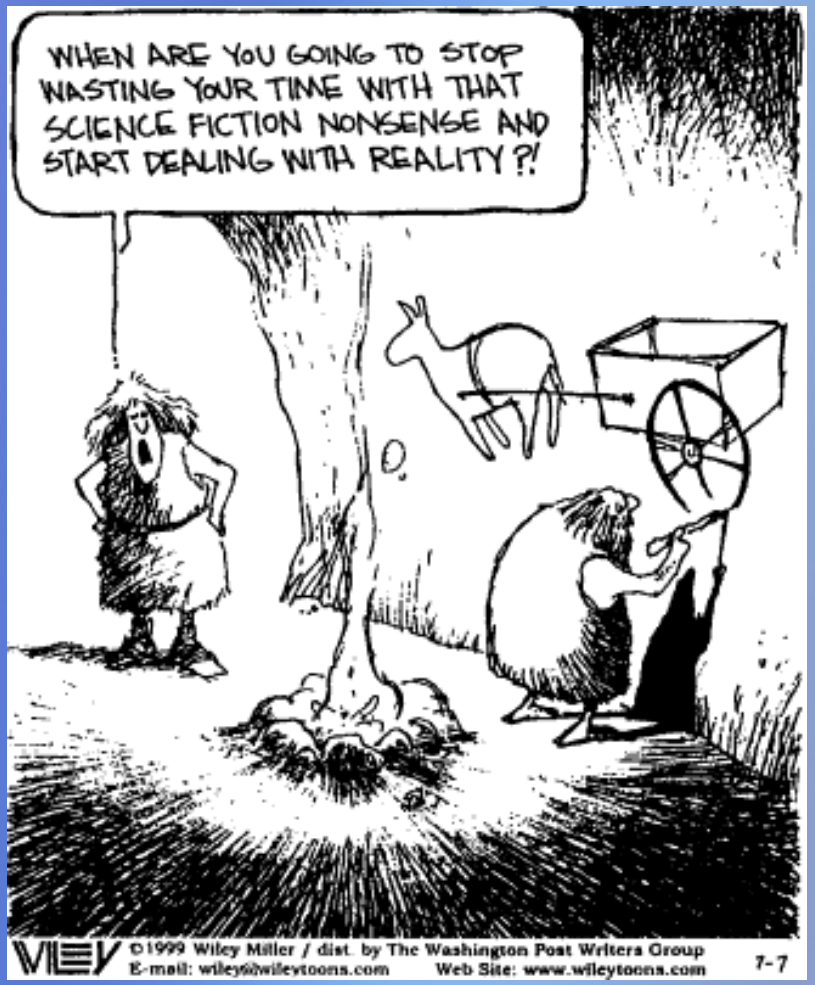
# Tools to Aid Creativity


- Bawden's four kinds of information to aid creativity: Interdisciplinary, peripheral, speculative, exceptions and inconsistencies
- Intriguing work of Prof Swanson: Linking "non-interacting" literature
  - $L_1$: Dietary fish oils lead to certain blood and vascular changes
  - $L_2$: Similar changes benefit patients with Raynaud's syndrome, $L_1 \cap L_2 = \phi$.
  - Corroborated by a clinical test at Albany Medical College
  - Similarly, magnesium deficiency & Migraine (11 factors) ; corroborated by eight studies.
- Will we provide the tools?

Bawden. "Information systems and the stimulation of the creativity". Information Science 86.
Swanson. "Medical literature as a potential source of new knowledge". Bull Med Libr Assoc. 90

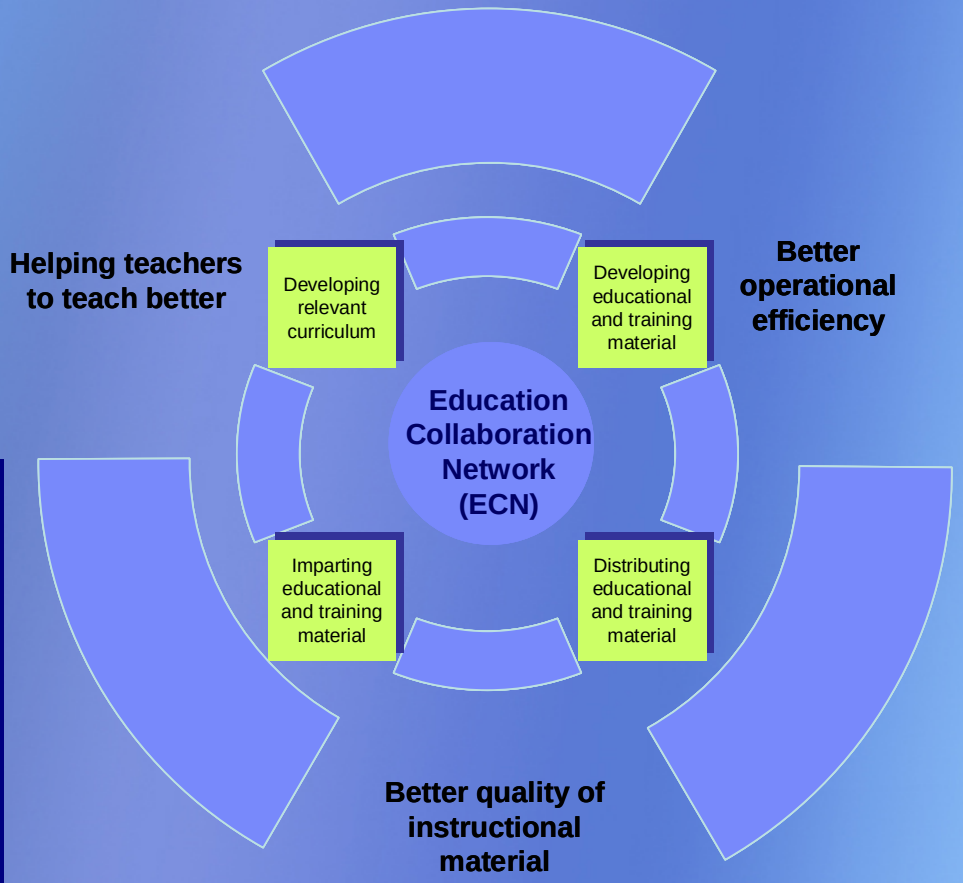
# Some Ideas

- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



# Education Collaboration Network

- Low teacher-student ratios
- instruction material poor and often out-of-date
- Poorly trained teachers
- High student drop-out rates




- Accumulation and re-use of teaching material
- Distributed, evolutionary content creation
- New pedagogy: teacher as discussant
- Multi-lingual

- A hardware and a software infrastructure built on industry standards that empower teachers, educators, and administrators to collectively create, manage, and access educational material, impart education, and increase their skills

- Teachers are able to find material that help them understand the subject matter and obtain access to teaching aids that others have found useful.
- Teachers also enhance the material with their own contributions that are then available to others on the network.
- Experts come to the class room virtually

Improving India's Education System through Information Technology. IBM Report to the President of India. 2005.

# Enabling Participation

- Inspired by Wikipedia
- But multiple viewpoints rather than one consensus version!
- How to personalize search to find the material suitable for one's own style of teaching?
- Management of trust and authoritativeness?


- More than 3.5 million articles in 75 languages
- Fashioned by more than 25,000 writers
- 1 million articles in English (80,000 in Encyclopedia Britannica)

msn

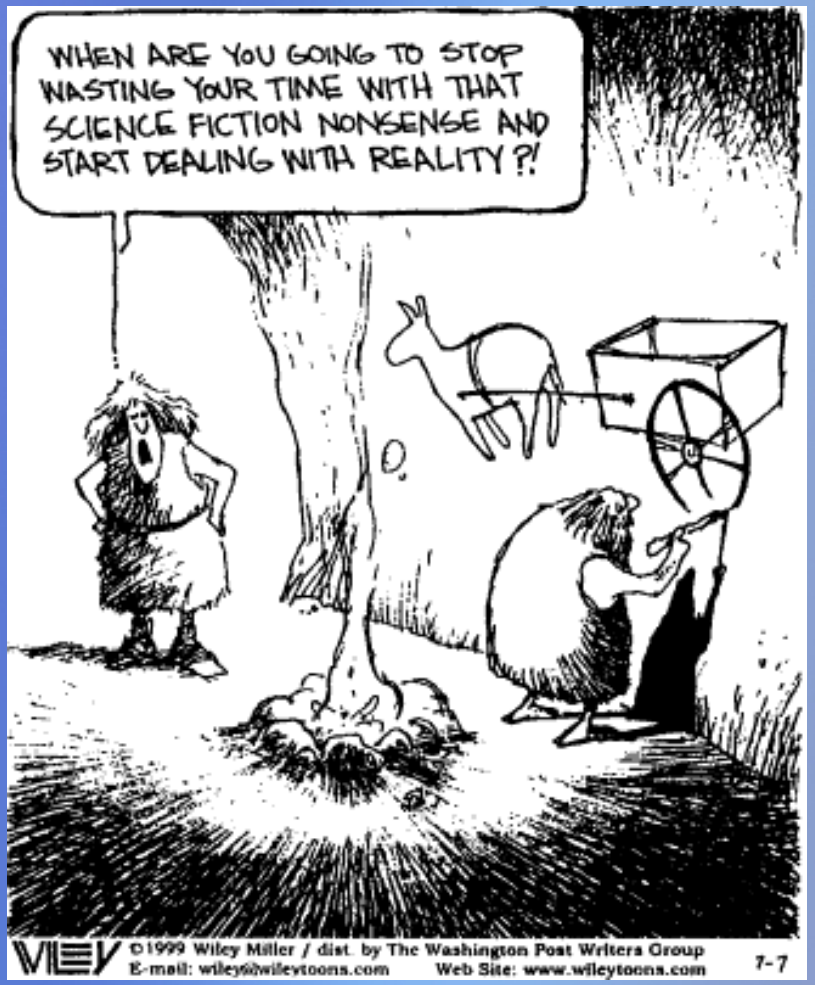
# Power of People Participation


- Theory: When a star went supernova, we would detect neutrinos about three hours before we would see the burst in the visible spectrum.
- Supernova 1987A: Exploded at the edge of Tarantula Nebula 168,000 years earlier.
- The underground Kamiokande observatory in Japan detected twenty four neutrinos in a burst lasting 13 secs on Feb 23, 1987 at 7:35 UT.
- Ian Shelton observed the bright light with his naked eyes at 10:00 UT in the Chilean Andes.
- Albert Jones in New Zealand did not see anything unusual at the Tarantula Nebula at 9:30 UT.
- Robert McNaught photographed the explosion at 10:30 UT in Australia.
- Thus a key theory explaining how universe works was confirmed thanks to two amateurs in Australia and New Zealand, an amateur trying to turn pro in Chile, and professional physicists in U.S. and Japan
- What's the general platform for participation?

Chris Anderson. The Long Tail. 2006.

# Some Ideas

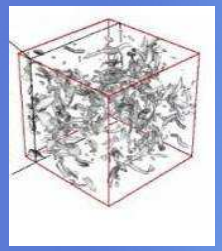
- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science

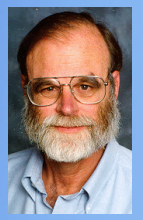


# Science Paradigms

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

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





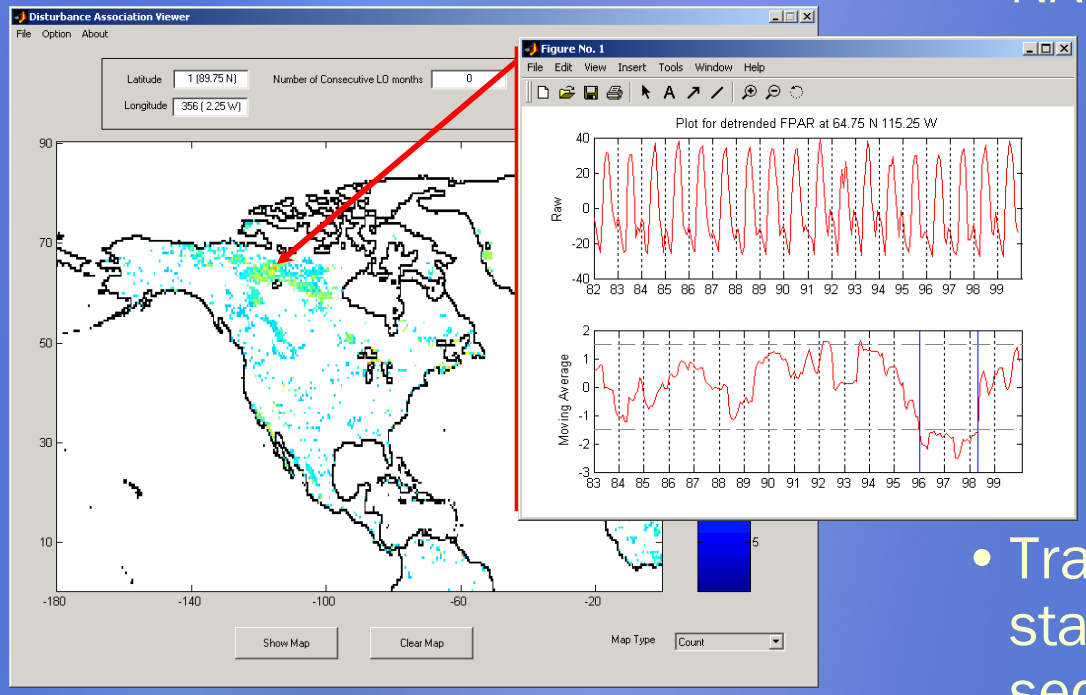
- Historically, Computational Science = simulation.
- New emphasis on informatics:
  - Capturing,
  - Organizing,
  - Summarizing,
  - Analyzing,
  - Visualizing

Courtesy Jim Gray, Microsoft Research.

# Understanding Ecosystem Disturbances

Vipin Kumar

U. Minnesota




- NASA satellite data to study
  - How is the global Earth system changing?
  - How does Earth system respond to natural & human-induced changes?
  - What are the consequences of changes in the Earth system?
- Transformation of a non-stationary time series to a sequence of disturbance events; association analysis of disturbance regimes
- Watch for changes in the amount of absorption of sunlight by green plants to look for ecological disasters

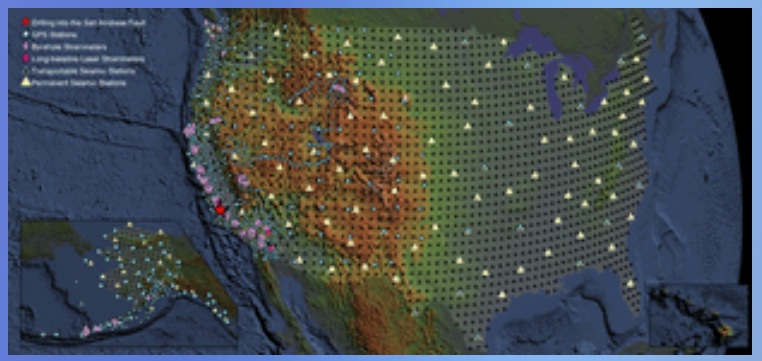
Potter et al. "Recent History of Large-Scale Ecosystem Disturbances in North America Derived from the AVHRR Satellite Record", Ecosystems, 2005.

# Some Other Data-Driven Science Efforts

- Bioinformatics Research Network

- Study brain disorders and obtain better statistics on the morphology of disease processes by standardizing and cross-correlating data from many different imaging systems
- 100 TB/year

- Earthscope



- Study the structure and ongoing deformation of the North American continent by obtaining data from a network of multi-purpose geophysical instruments and observatories
- 40 TB/year

Newman et al. “Data-Intensive e-Science Frontier Research in the Coming Decade”. CACM 03.

# Call to Action

- We ought to move the focus of our future work towards humane data mining (applications to benefit individuals):
  - Personal data mining (e.g. personal health)
  - Enable people to get a grip on their world (e.g. dealing with the long tail of search)
  - Enable people to become creative (e.g. inventions arising from linking non-interacting scientific literature)
  - Enable people to make contributions to society (e.g. education collaboration networks)
  - Data-driven science (e.g. study ecological disasters, brain disorders)
- Rooting our future work in these (and similar) applications, will lead to new data mining abstractions, algorithms, and systems (the Quest lesson)

# Thank you!