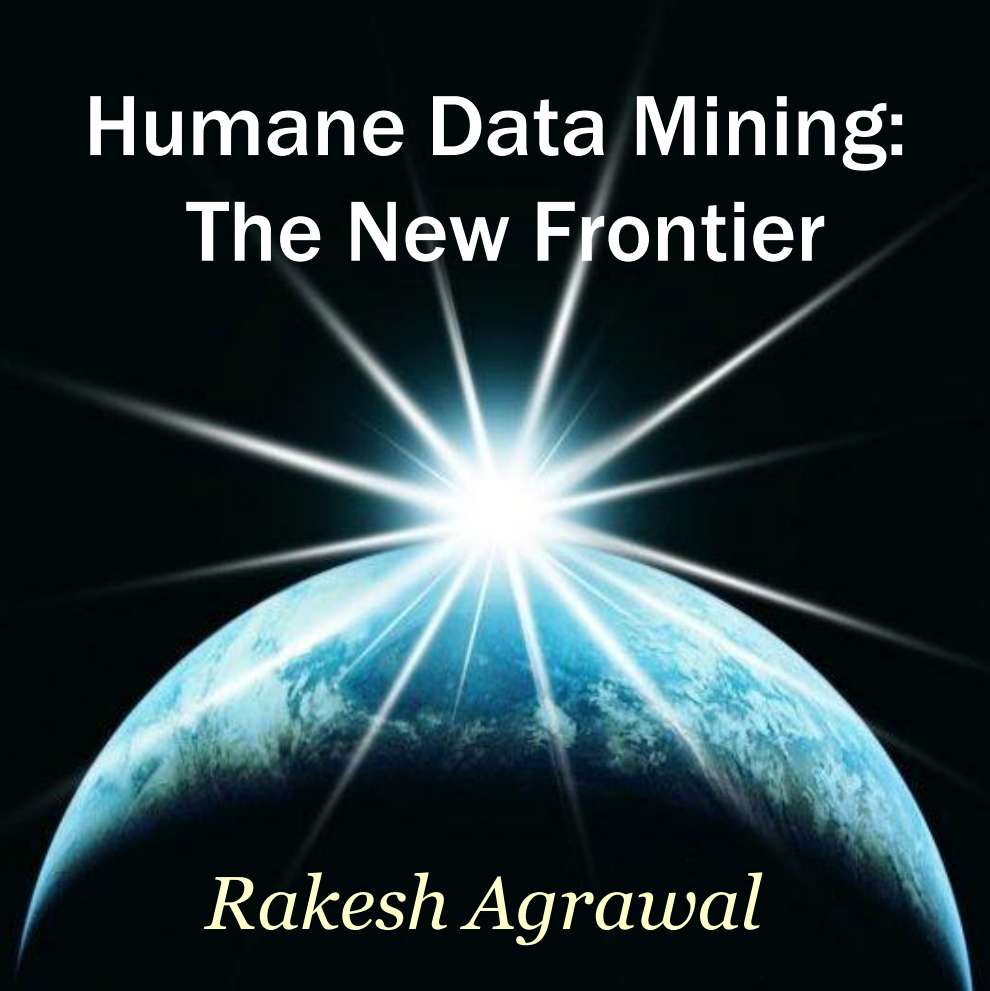
Humane Data Mining: The New Frontier



Rakesh Agrawal

Microsoft Search Labs
Mountain View, California

2008 SIAM ANNUAL MEETING
July 7-11, 2008
Town and Country Resort & Convention Center
San Diego, CA

Updated version of the SIGKDD-06 keynote

Microsoft Research

Thesis

- 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 develop new abstractions, algorithms and systems, inspired by new applications

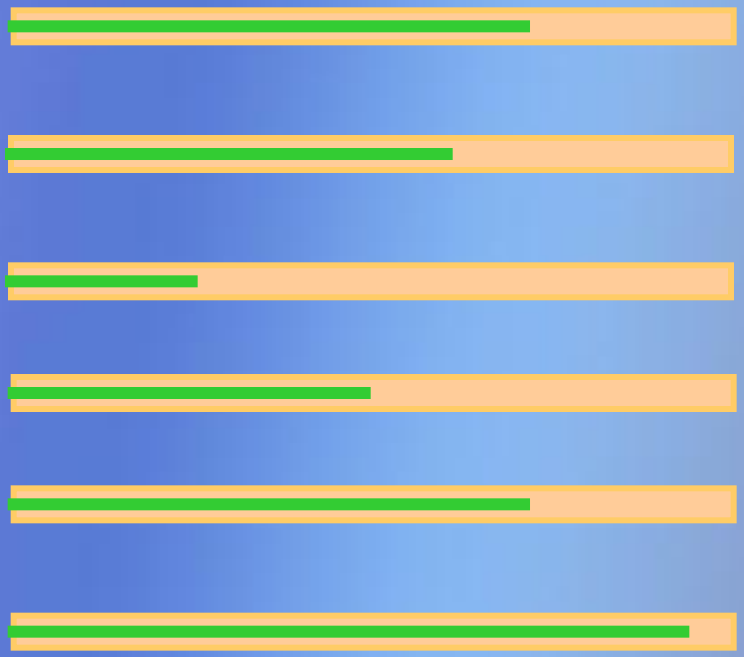
Microsoft Research

Outline

- Progress report
- New Frontier

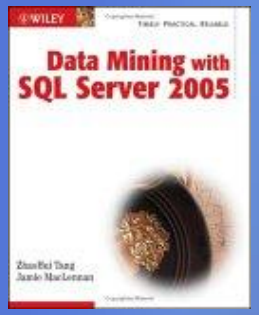
A Snapshot of Progress

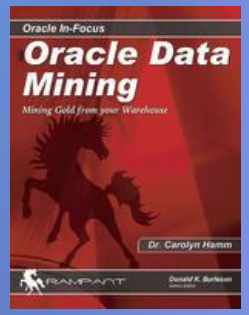
- System support
- Algorithmic innovations
- Foundations
- Usability
- Enterprise applications
- Unanticipated applications



Microsoft Research

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

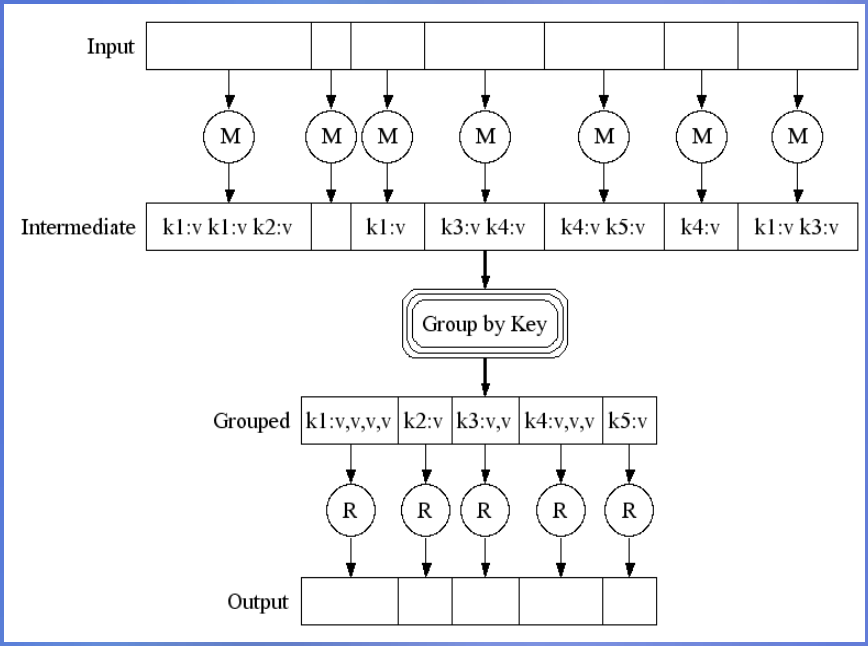
Microsoft Research

Google's Data Mining Platform

MapReduce[1]: Programming Model

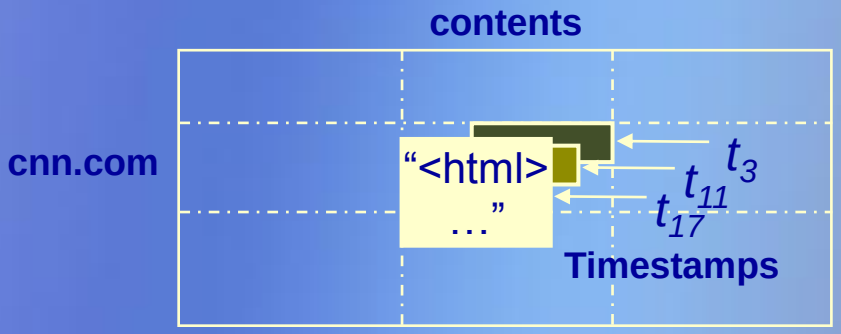
map(ikey, ival) -> list(okey, tval)

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

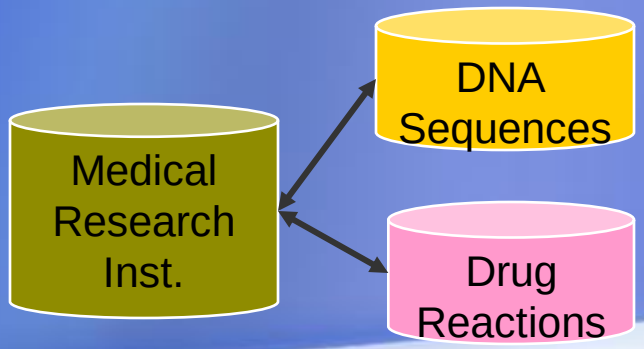
Microsoft Research

[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.

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**

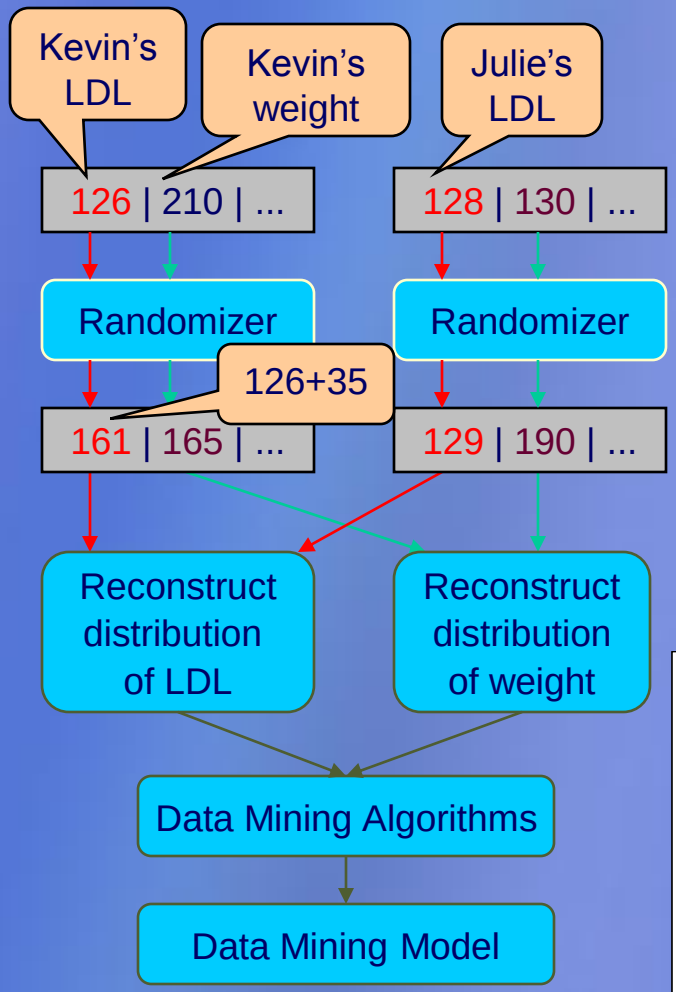
R ❁ S
u
v

Count (R ❁ S)

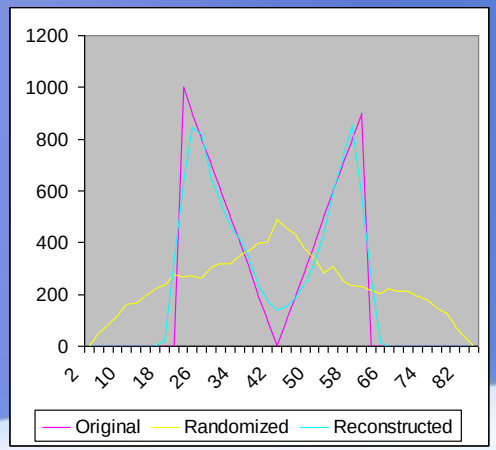
- R and S do not learn anything except that the result is 2.

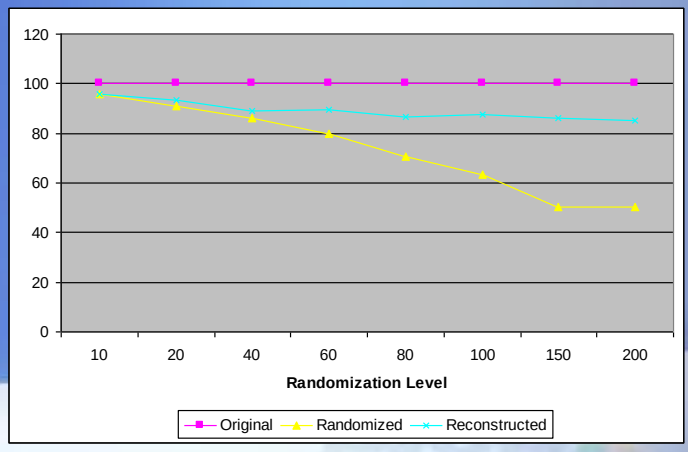
Sigmod 03, DIVO 04

Privacy Preserving Data Mining



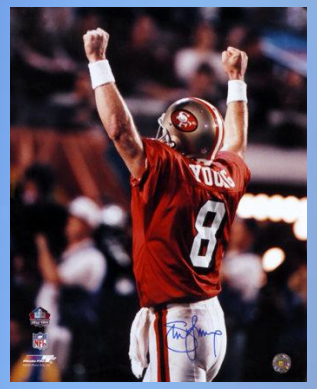
- Preserves privacy at the individual patient level, but allows accurate data mining models to be constructed at the aggregate level.
- Adds random noise to individual values to protect patient privacy.
- EM algorithm estimates original distribution of values given randomized values + randomization function.
- 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

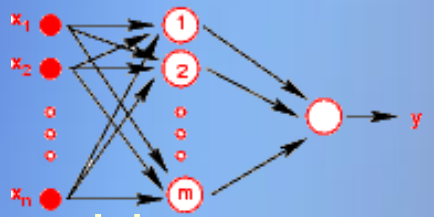
Supplier Relationship Management

Supply Chain Analysis
Demand Planning | Warranty Analysis

Web Analytics

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

Ordering Search Results



- Search results ranked using a learning algorithm (neural net).
- Training data: Query/document pairs labeled for relevance (perfect, excellent, good, etc.).

baseball	http://sports.espn.go.com/mlb/index	Perfect
baseball	http://en.wikipedia.org/wiki/Baseball	Good

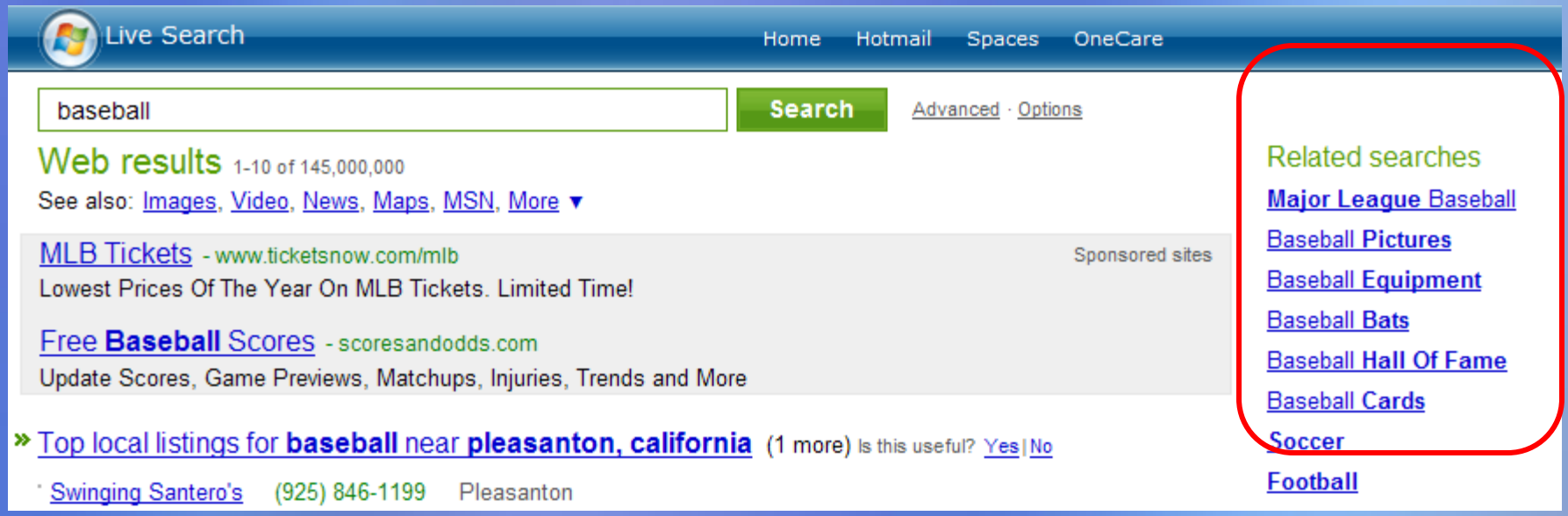
- Query independent (e.g. static page rank) as well as query dependent features (e.g. position of a query word in anchor text) for every document.

baseball http://sports.espn.go.com/mlb/index perfect f1 f2 …

Microsoft Research

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

Related Searches



- Most popular queries containing the current query
- Analysis of how users reformulated their queries

  Football ⟶ Soccer (whole query)

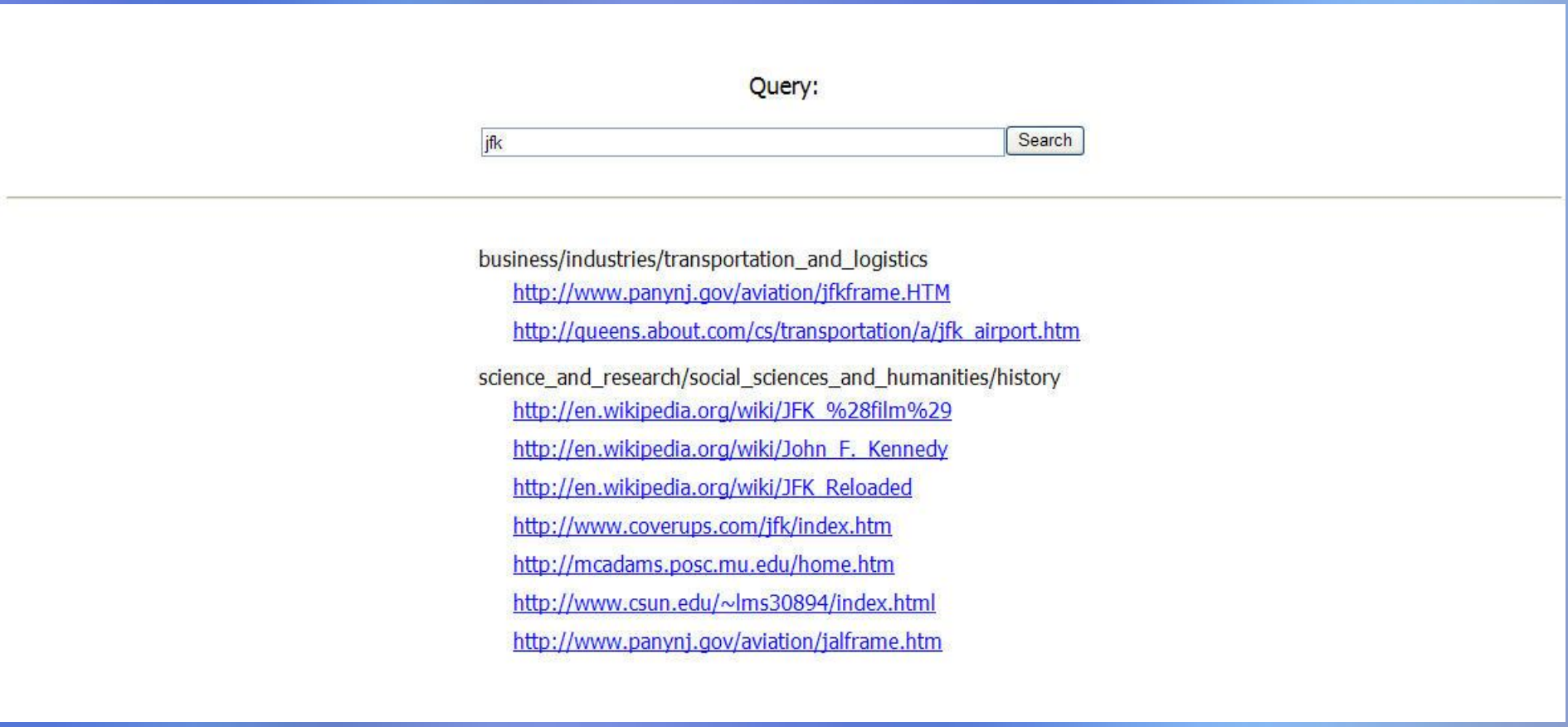
  Wildflower cafe ⟶ Wildflower bakery (piecewise)

- Query click graph to find related queries

Microsoft Research

Result Diversification

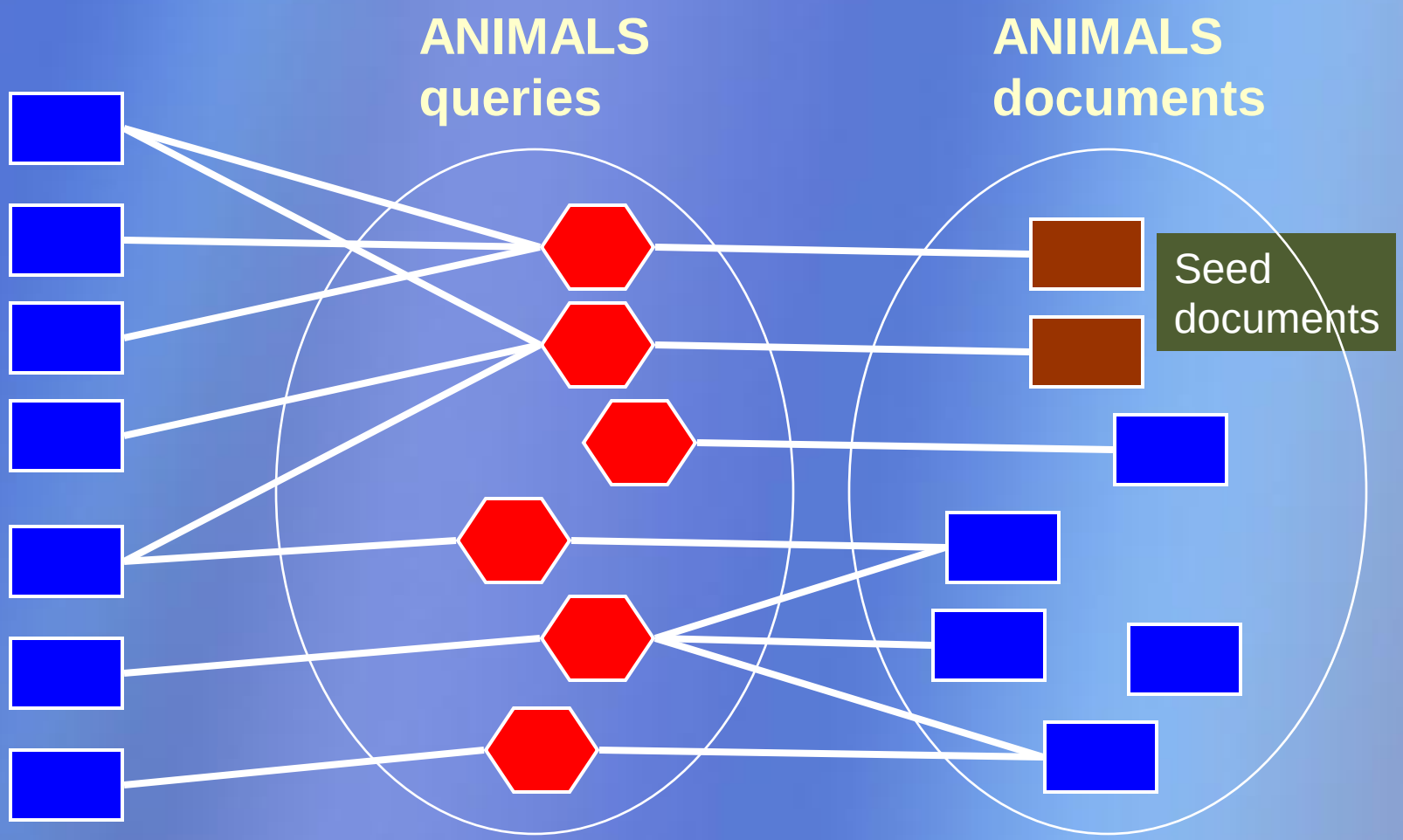
Unanticipated Applications



- Ideas from portfolio theory to allocate space to different result types
- Marginal utility of adding a document decreases if the result set already contains high quality documents of the same type
- Query and document classification using merged click logs

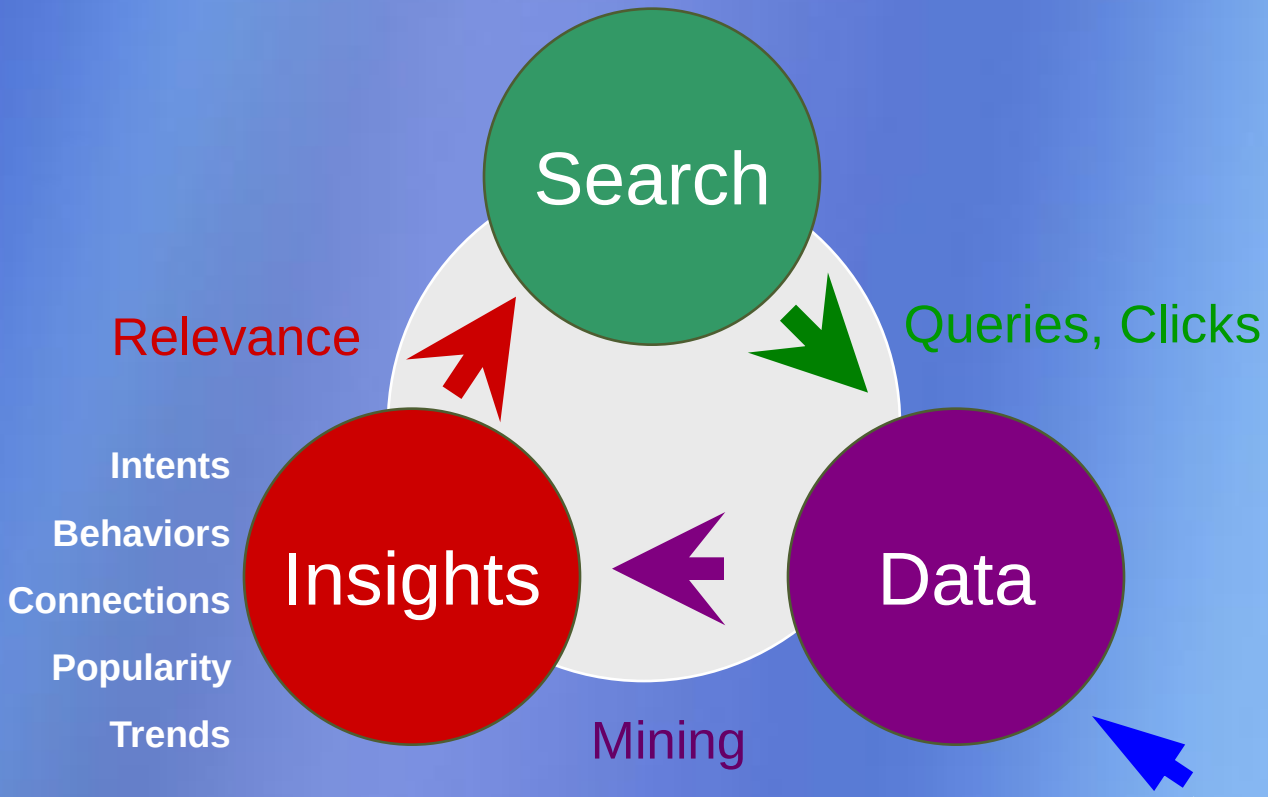
Microsoft Research

Classification Using Click Graph



Algorithm: Random walk with absorbing states

Search & Data: Virtuous Cycle



Better Search Results ►

More Data ►Greater Insights ►

Better Search Results

Web Pages
Feeds

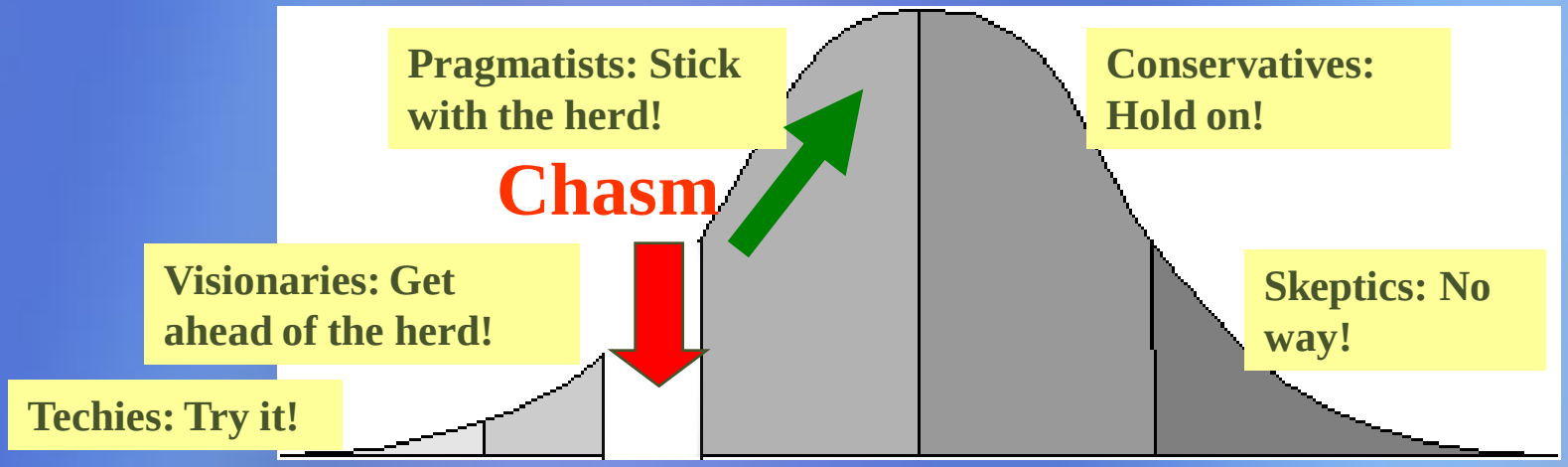
Microsoft Research

Outline

- Progress Report
- New frontier

Microsoft Research

Imperative Circa 2008



Maintain upward trajectory:

- 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

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

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

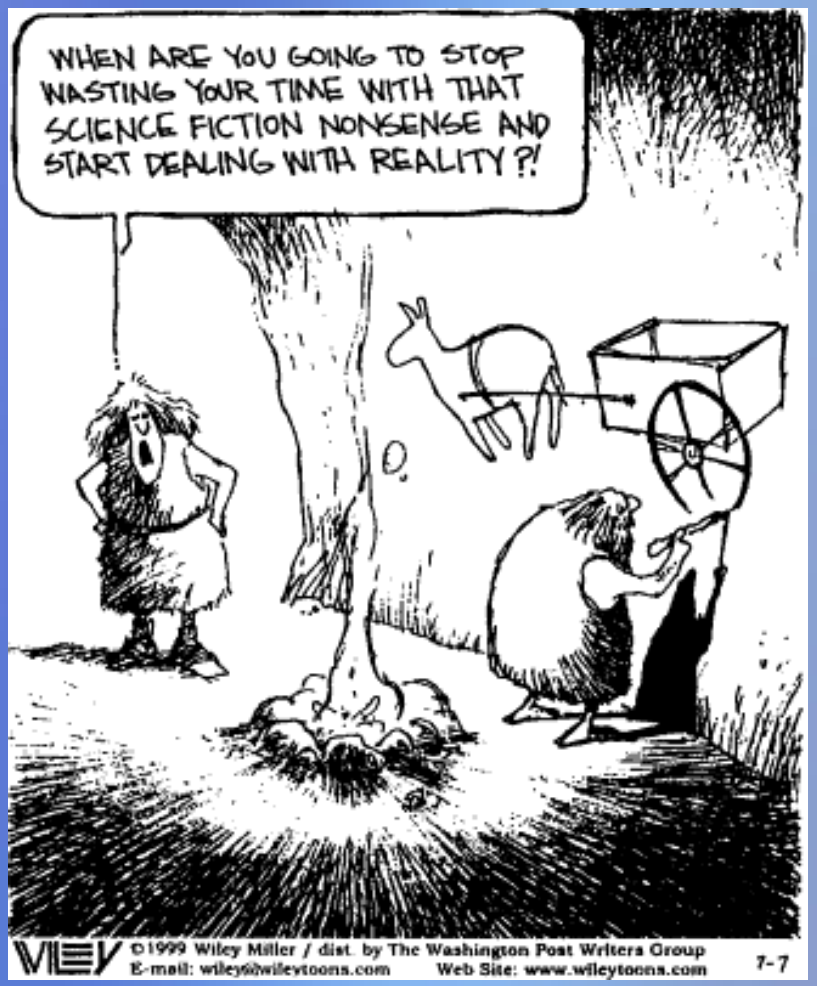
Microsoft Research

An Expansive Definition of Data Mining

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

Some Ideas

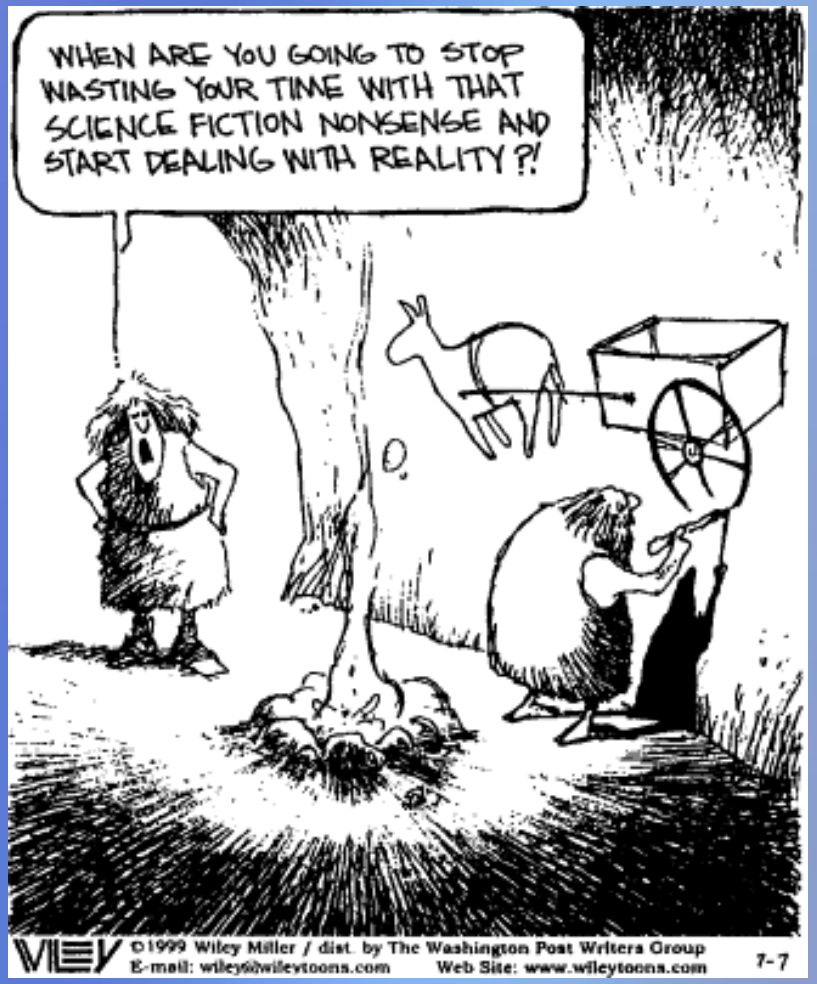
- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



Microsoft Research

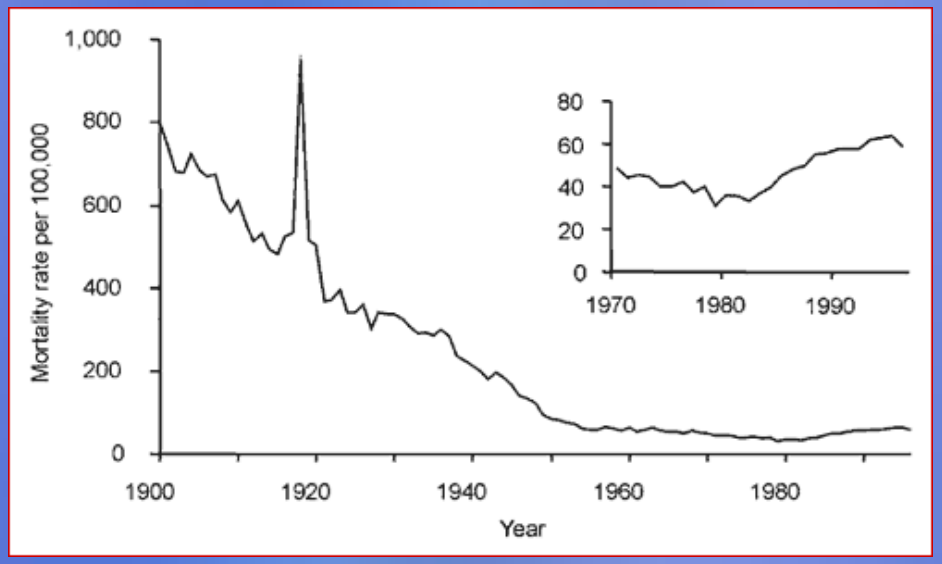
Some Ideas

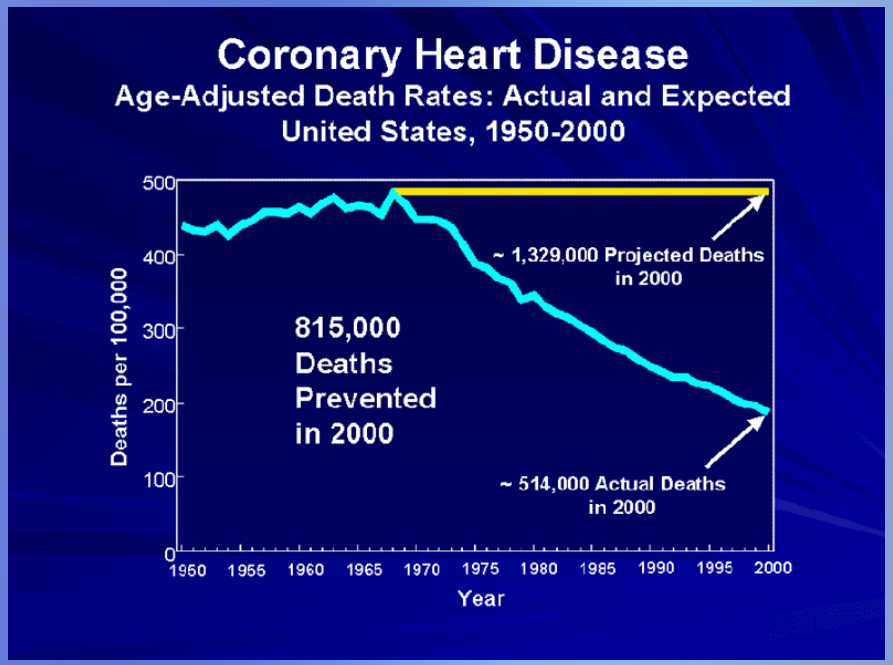
- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science

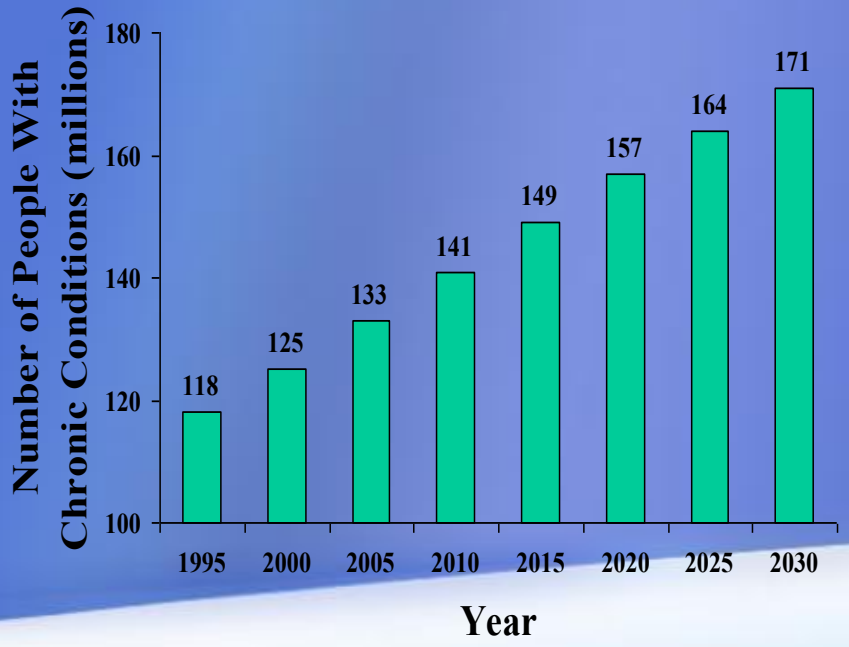


Microsoft Research

Changing Nature of Disease







- **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).

Microsoft Research

Technology Trends

- Tremendous simplification in the technologies for capturing useful personal information







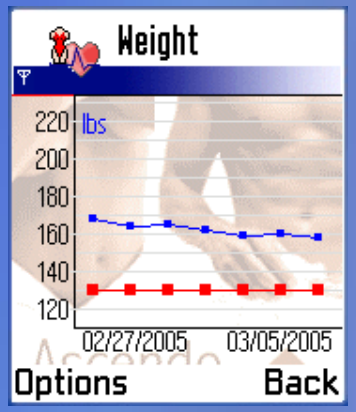

- Dramatic reduction in the cost and form factor for personal storage

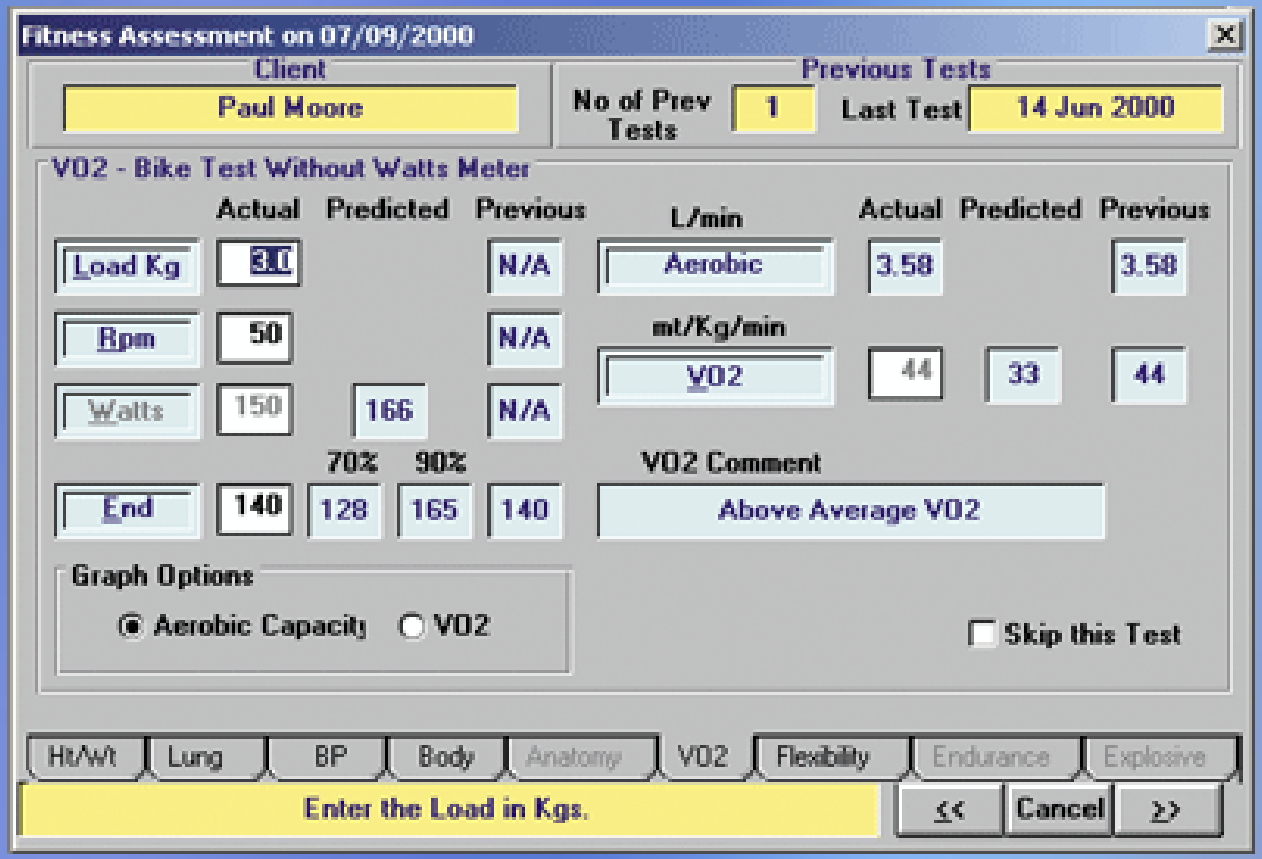



- Cloud Computing

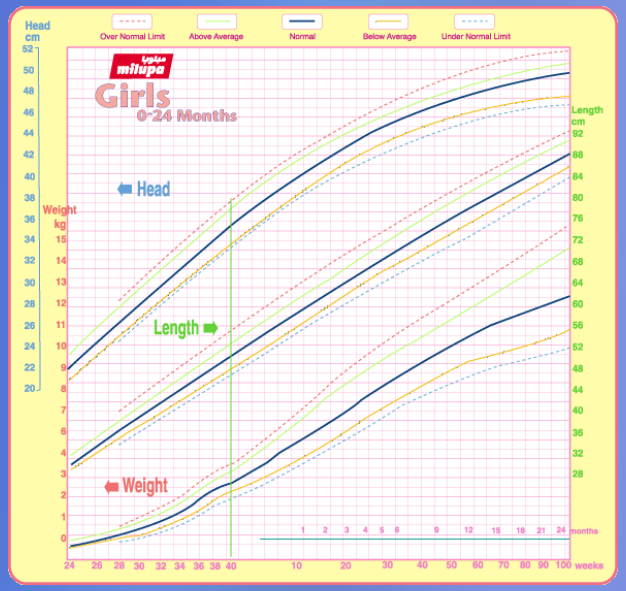
Microsoft Research

Personal Health Analytics



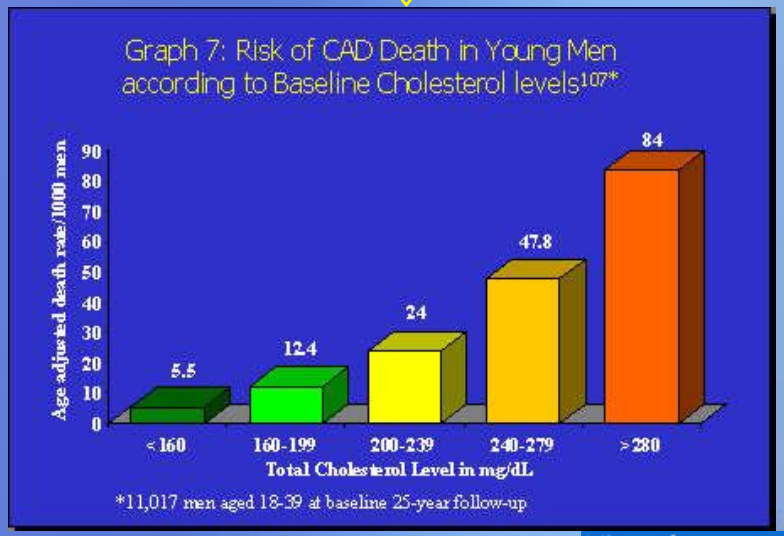


Personal Data Mining



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

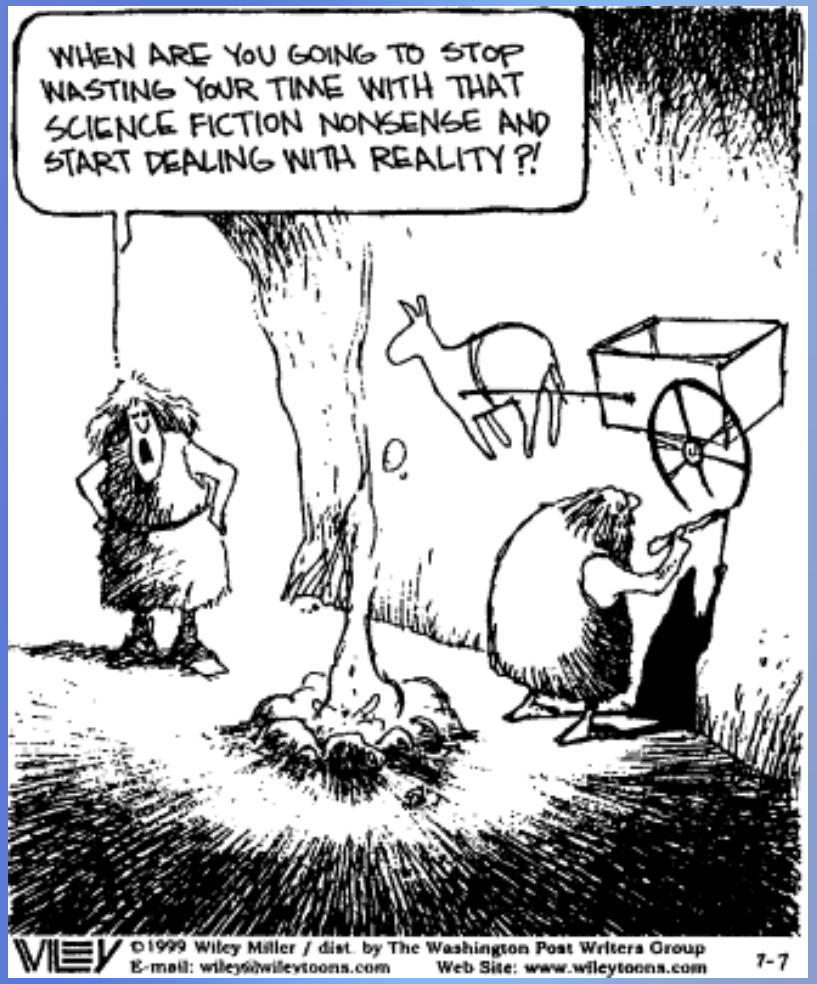
Privacy and security

Microsoft Research

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

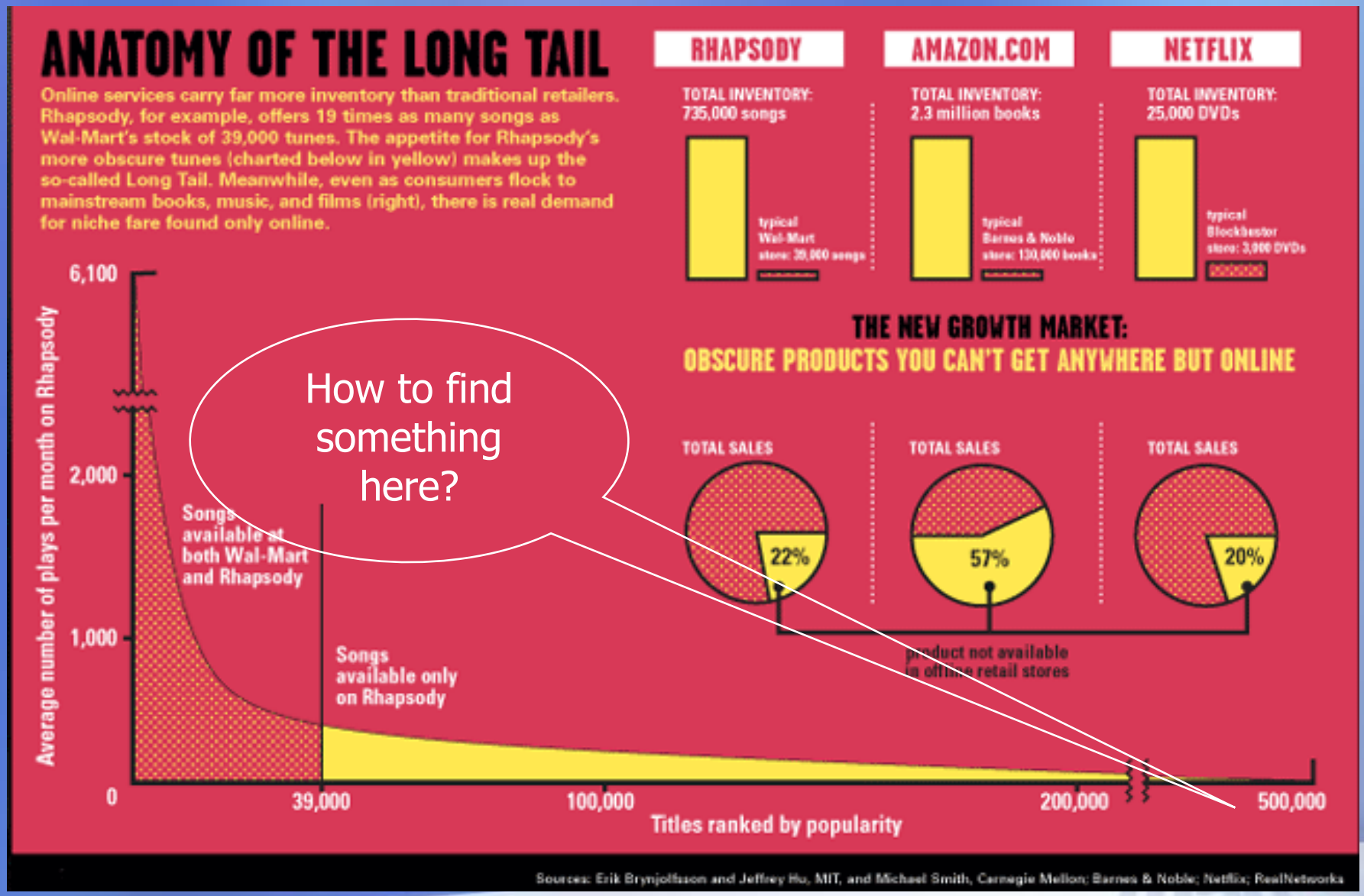
Some Ideas

- 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



Microsoft Research

The Tyranny of Choice

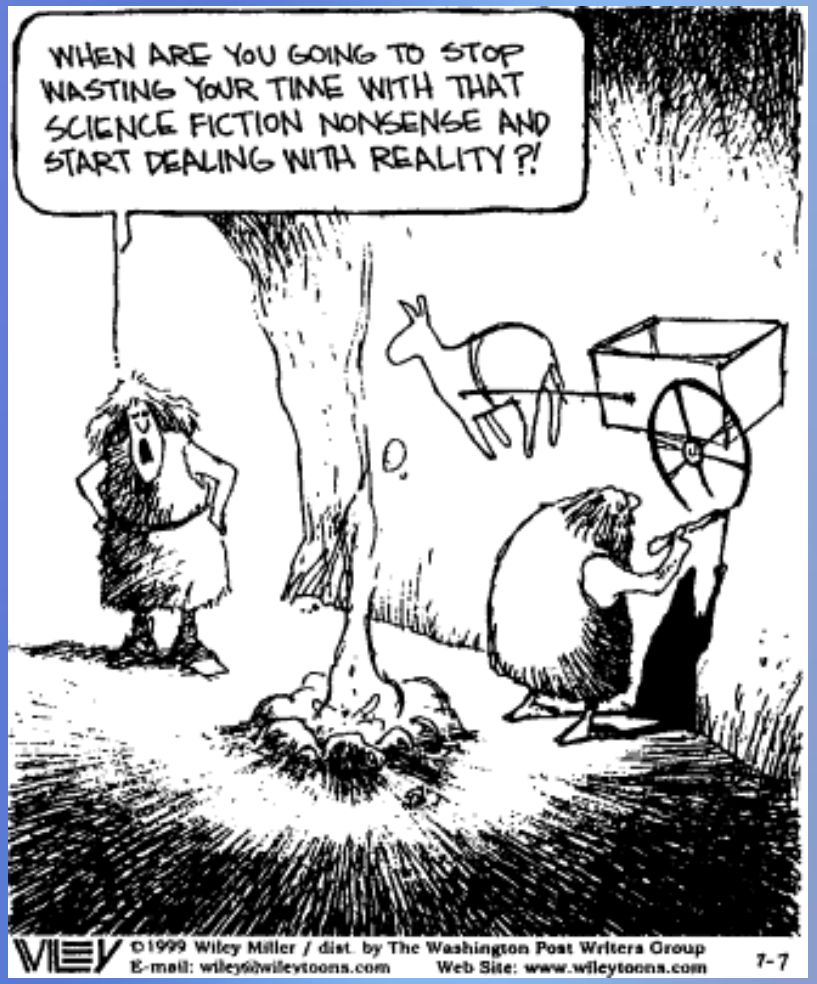


Chris Anderson. The Long Tail. 2006.

See also: Anita Elberse. Should You Invest in the Long Tail? HBR, 2008.

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



Microsoft Research

Tools to Aid Creativity

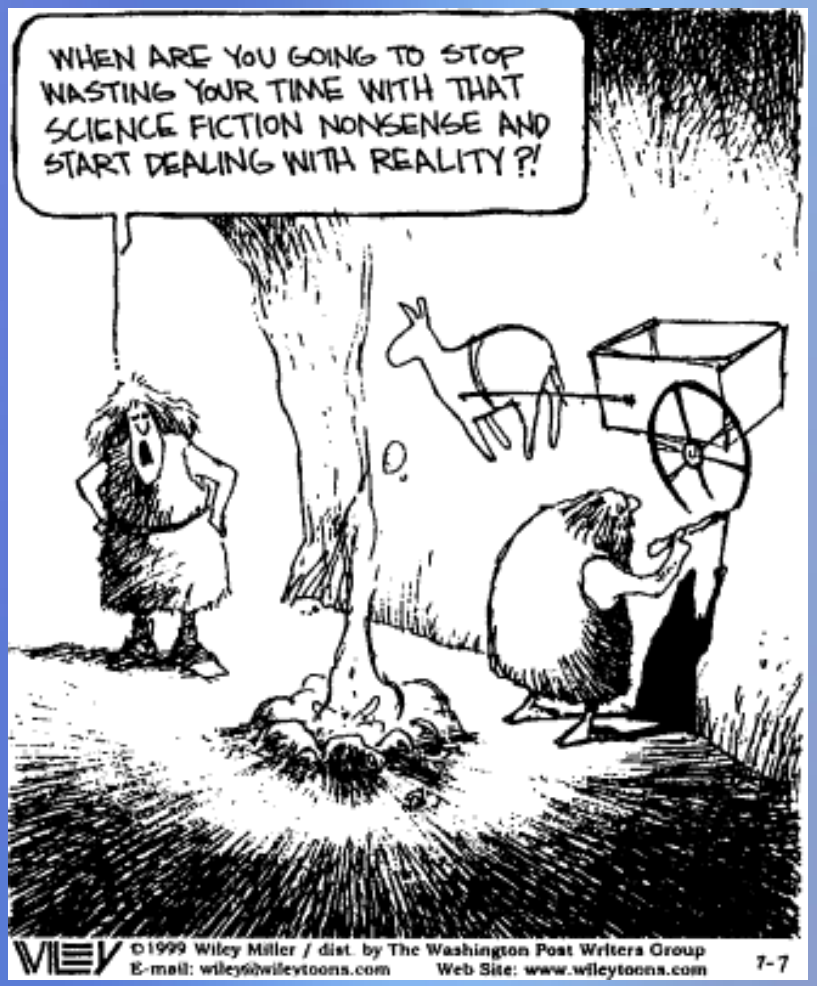
- Bawden's four kinds of information to aid creativity: Interdisciplinary, peripheral, speculative, exceptions and inconsistencies
- Intriguing work of Prof Swanson: Linking "non-interacting" literature
  - L_1: Dietary fish oils lead to certain blood and vascular changes
  - L_2: Similar changes benefit 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?

Microsoft Research

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

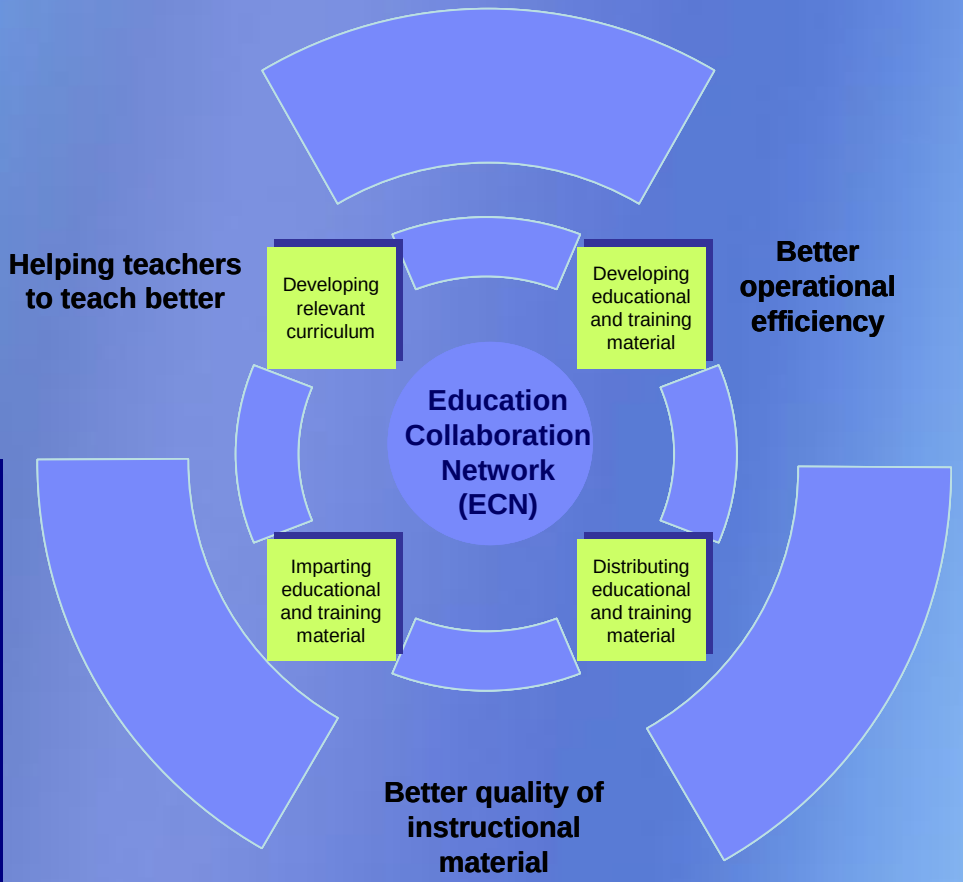
- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



Microsoft Research

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.

Collaborative Knowledge Creation (Educational Material)

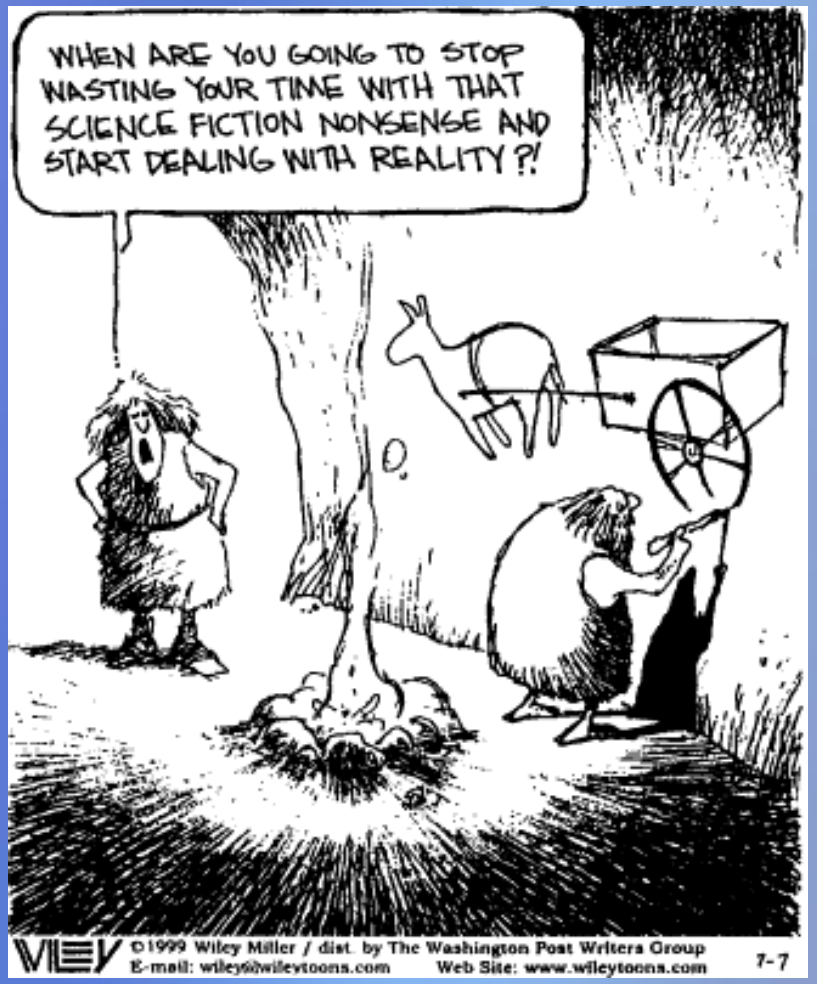
- Inspired by Wikipedia
- But multiple viewpoints rather than one consensus version!
- How to personalize search to find the material suitable for one's own style of teaching?
- Management of trust and authoritativeness?

- 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)

Microsoft Research

Some Ideas

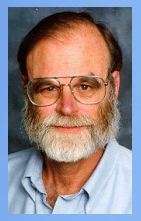
- Personal data mining
- Enable people to get a grip on their world
- Enable people to become creative
- Enable people to make contributions to society
- Data-driven science



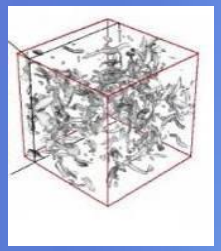
Microsoft Research

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



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



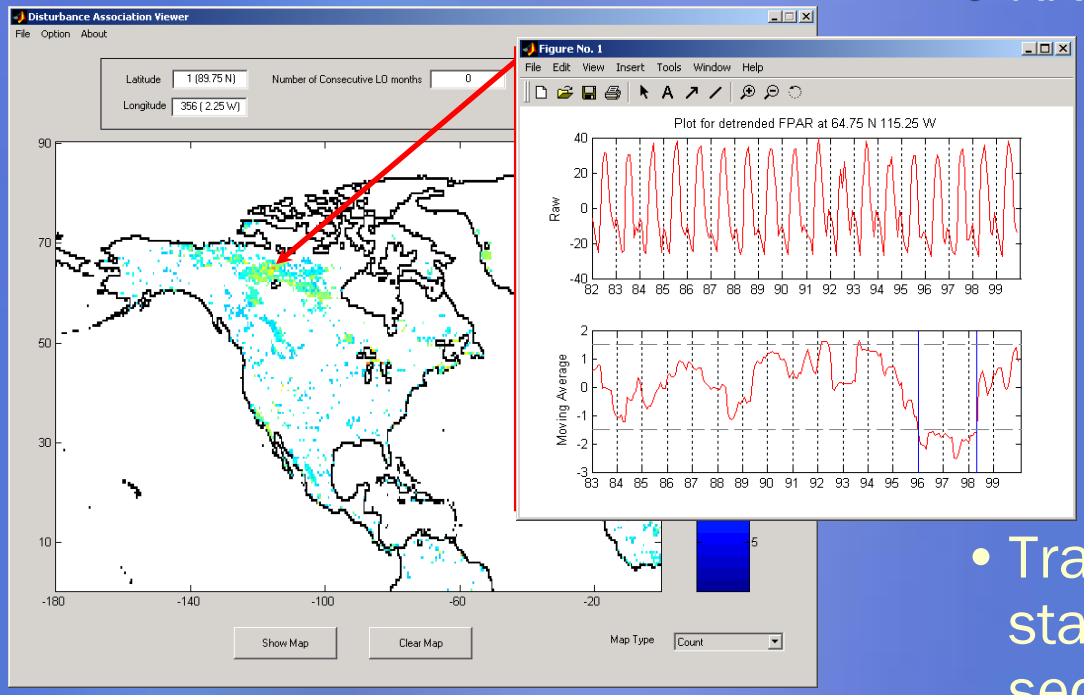
- Historically, Computational Science = simulation.
- New emphasis on informatics:
    - Capturing,
    - Organizing,
    - Summarizing,
    - Analyzing,
    - Visualizing

Courtesy Jim Gray, Microsoft Research.

Understanding Ecosystem Disturbances

Vipin Kumar

U. Minnesota



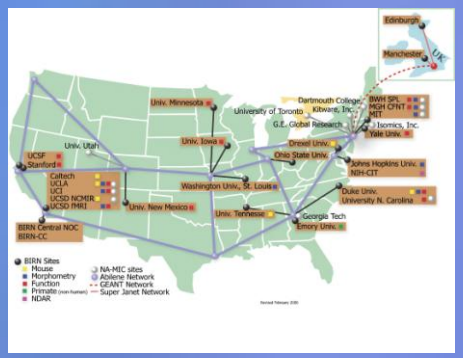
- NASA satellite data to study
  - How is the global Earth system changing?
  - How does Earth system respond to natural & human-induced changes?
  - What are the consequences of changes in the Earth system?
- Transformation of a non-stationary time series to a sequence of disturbance events; association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

Microsoft Research

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

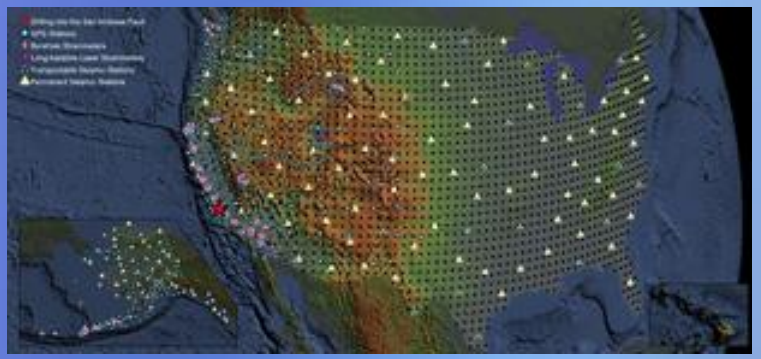
Some Other Data-Driven Science Efforts

- **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

- Move the focus of 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 work in these (and similar) applications, will lead to new abstractions, algorithms, and systems

Microsoft Research

Thank you!

Search Labs' mission is to invent next in Internet search and applications.

Microsoft Research