Telcordia Technologies

Performance from Experience

Formerly Bellcore

# Mining Gold from E-Commerce Transactions: Opportunities & Challenges

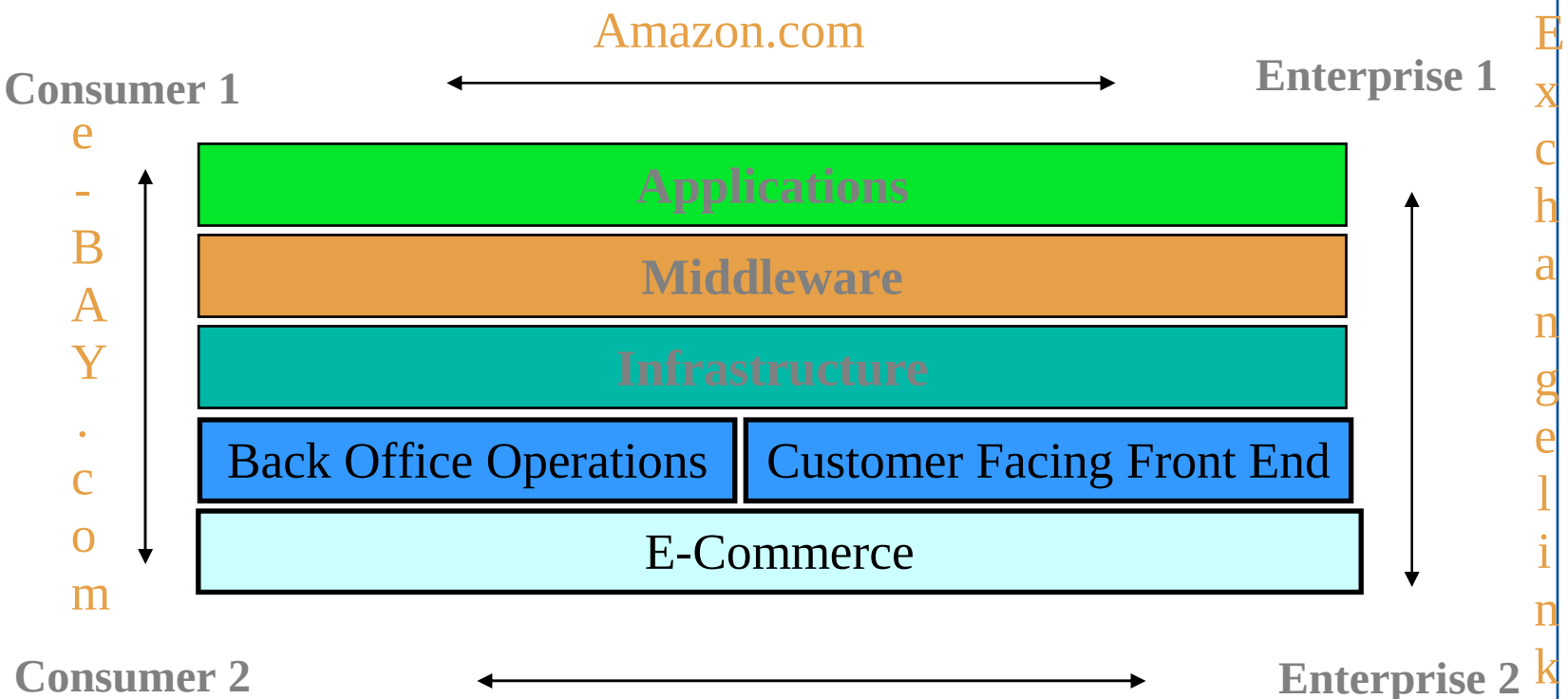
**S. Dalal, D. Egan, Y. Ho,**
**C. Lochbaum, M. Rosenstein**
**sid@research.telcordia.com**

# Information Services: Needs

- E-Commerce- a fact of life
- Even the simplest transaction generates enormous amount of data
- Tremendous amount of Information Overload
  - Winners will be the ones who can deal with this

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Classification of E-Commerce Opportunities




Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Even the Simplest Transactions Generate Enormous Data: Scenario

A consumer goes online looking for a new science fiction book. He goes to the web site of a particular bookstore. As soon as he is on the web site of the bookstore, the bookstore flashes an ad, and the home page gives a number of options (e.g., book, music, videos, etc.). The consumer clicks on their book section. In return, the bookstore recommends a new book by a particular author. It also flashes reviews of the book, and a quick summary. To clinch a quick sale and promote its music sales, it plays the music from a new movie based on that book, and offers a 10% instant rebate on it if it is bought with the book. The consumer buys the book with a credit card, and also buys the music, which he downloads. Besides the purchase price of the book, he pays a shipping and handling charge. Two days later, since he did not receive the book in 24 hours, he contacts the bookstore's online customer care center where he is informed that his package is currently at the delivery service's center hub in Tennessee and will be in his town at 3 PM and is given the tracking number. He gets his book at 4 PM.

- Multiple Roles by each player
- How to gather information- sellers/buyers?

# Data Logged and Used

- Data logged in Log Files
  - Hits
  - Host/IP (What kind of customer is it?)
  - Date Stamp (For Traffic, etc.)
  - Retrieval Method and its success
  - Bytes retrieved
  - Browser and computing platform used
- Referrer file
  - Website from which the user came (separates repeat users)
- Legacy Info- through “cookie”
- Profile through pervious registration and purchase behavior

Biggest use of measurements is to support ads

•3 Billion $ Revenues

Telcordia™ Technologies

*Formerly Bellcore…*
*Performance from Experience*

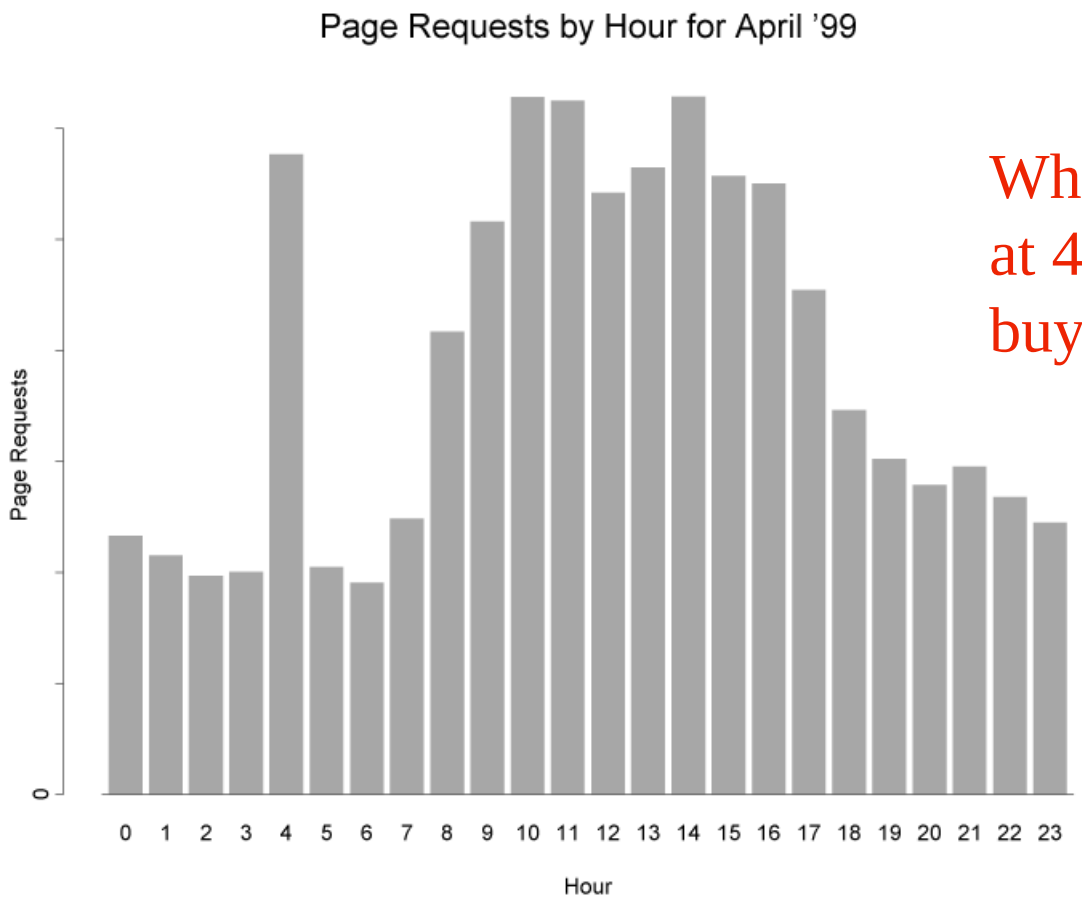
# Some Metrics used in Ad Measurements

| Metrics | Explanation/Definition |
|---|---|
| Ad Clicks | Number of times users click on an ad banner. |
| Ad Click Rate | Percentage of ad views that result in an ad click. Also referred to as "click-through". |
| Ad Views (Impressions) | Number of times an ad banner is downloaded and presumably seen by visitors. |
| Bandwidth | How much information (text, images, video, sound) can be sent through a connection. Usually measured in bits-per-second. |
| Browser Caching | Browsers stored recently used pages on a user's disk. If a site is revisited, browsers display pages from the disk instead of requesting them from the server. |
| Click through | The percentage of ad views that resulted in an ad click. |
| CPC | Cost-per-click for a specific banner ad. |
| Conversion Rate | The percent of visitors who become bonafide buyers. |
| CPM | CPM is the cost per thousand for a particular site. |
| Gross Exposures (Hits) | Each time a Web server sends a file to a browser, it is recorded in the server log file as a "hit." Hits are generated for every element of a requested page (including graphics, text and interactive items). |
| Page Views | Number of times a user requests a page that may contain a particular ad. Indicative of the number of times an ad was potentially seen, or "gross impressions." |
| Unique Users | The number of different individuals who visit a site within a specific time period. To identify unique users, Web sites rely on some form of user registration or I.D. system. |
| Valid Hits | A further refinement of hits, valid hits are hits that deliver all information to a user. |
| Visits | A sequence of requests made by one user at one site. If a visitor does not request any new information for a period of time, known as the "time-out" period, then the next request by the visitor is considered a new visit. |

# Factors & Methods affecting Reliability of “UNIQUE USERS”

| Effect | Factors | Method |
|---|---|---|
| **Under-count** | Local Caching | Frequently used pages are stored in local cache. User doesn't go to the designated server, and thus, server undercounts |
| | Proxy Servers/Firewalls | Gives same IP address to a number of users- thus, # of IP addresses are undercounting # of users |
| | Site Mirroring | Frequently used pages are mirrored in another server. User doesn't go to the designated server, and thus, server undercounts |
| | Dynamically generated IP | IP addresses are dynamically generated- thus, same user can get different addresses and vice versa |
| **Over-count** | Crawlers | Agents for search engines which go on visiting pages for indexing purposes |
| | Rogue Bots | Software which mechanically generates traffic and IP addresses to inflate count |
| | Graphics Off on browser | Only text part of a graphic ad is seen- but, registered as a visit |
| | Performance Measurement Tools | To measure download time, and reliability, goes on polling sites throughout the day. |
| | Filters/Intelligent Agents/Virtual Includes | Filters and agents are fetching only a specified part of a web page (e.g. text), thus, users do not see ads |
| | Internet bottleneck | When download time is long, user goes on repeatedly requesting the same page |

# Experience at Telcordia: Hits



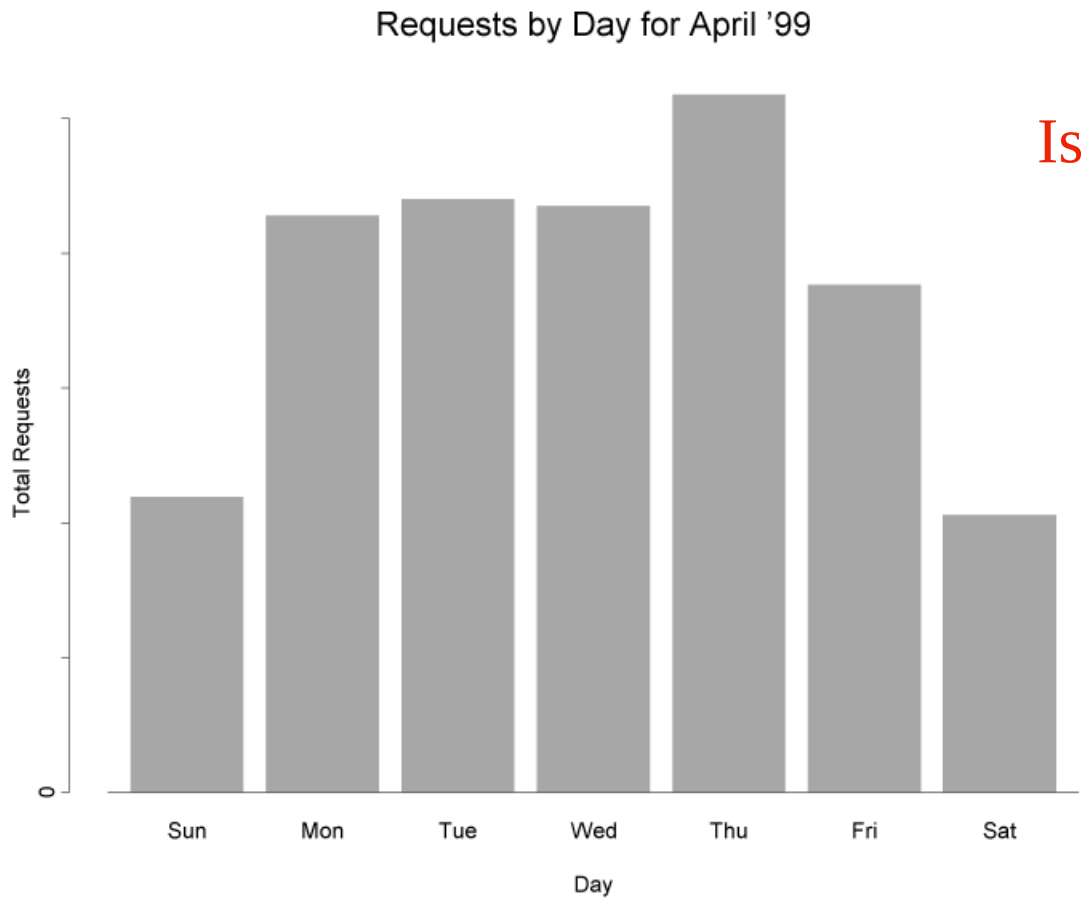

Why buyers at 4am are not buying?

Traffic per Hour based on Server Log Files

Telcordia™ Technologies

*Formerly Bellcore…*
*Performance from Experience*

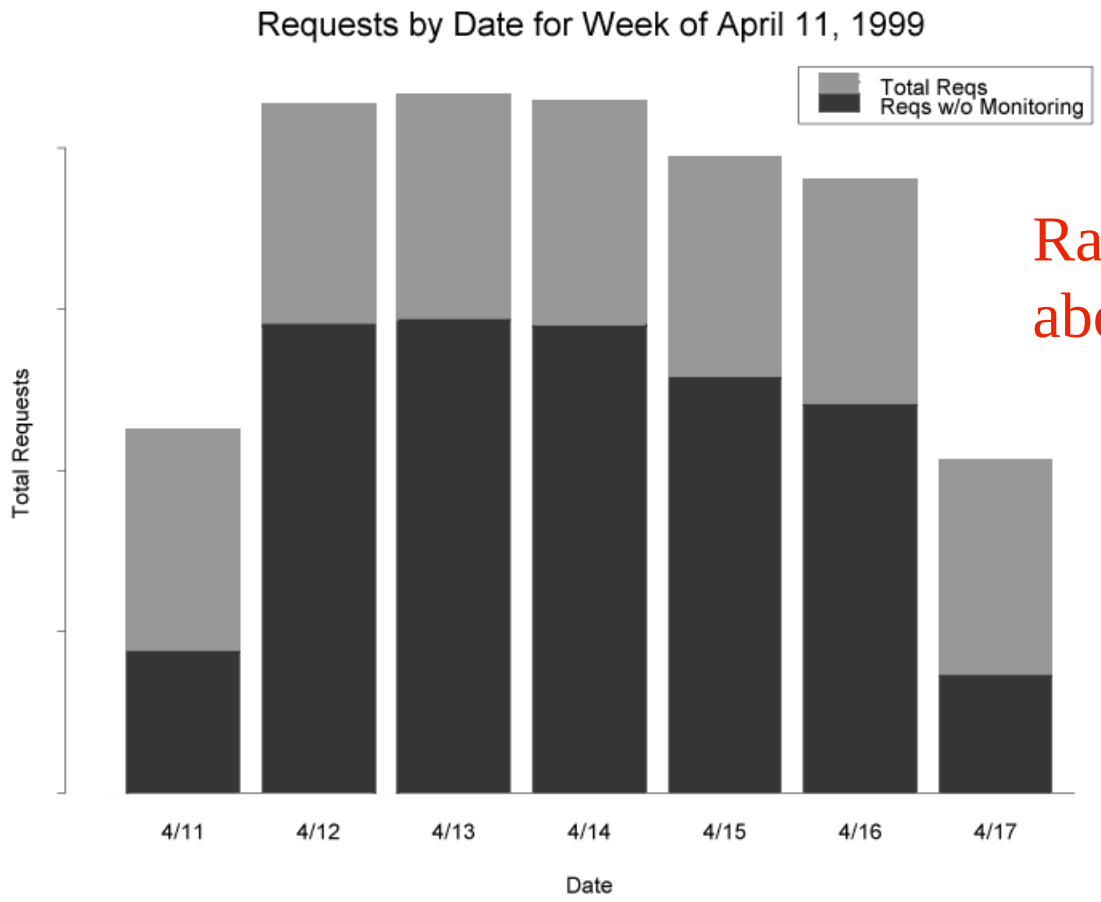
# Traffic/Day of Week based on Server Log




Is this real?

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Traffic/Day of Week based on Server Log




Raises Questions about Data Quality

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Challenges:

- Right kind of detailed metrics
- Which specific information to provide to maximize the response? Presentation issues
- “Zero Time Latency Transactions” & SCM important- How to facilitate it?
- Data Quality- how to improve it? especially when combining with legacy information?
- How do you mine textual information and recommend?
- Privacy and ownership of data issues

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Integrated Data Management: Data Quality

- With Internet tremendous information explosion
  - front end information from customers
  - back end information from suppliers
- Big opportunity for merging, analyzing and interpreting information- and providing DSS
- All these steps require assurance of data quality
- Quality of data is typically notoriously bad- only 35-40% of data is good enough in many enterprises to allow the use of full automation.
- How should we improve data quality?

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# What is good quality? How to assure it?

- Big opportunity for merging front end (customers) & backend (legacy) info & analyzing - and providing DSS
- Good quality
  - 100% data individually and collectively should be correct and consistent

•Impossible

- •too many different types of data from different sources & different granularities
  - •too many relationships
  - •data ages
- •prohibitively expensive

•Objective: Some large percent of data is "correct" with very high probability

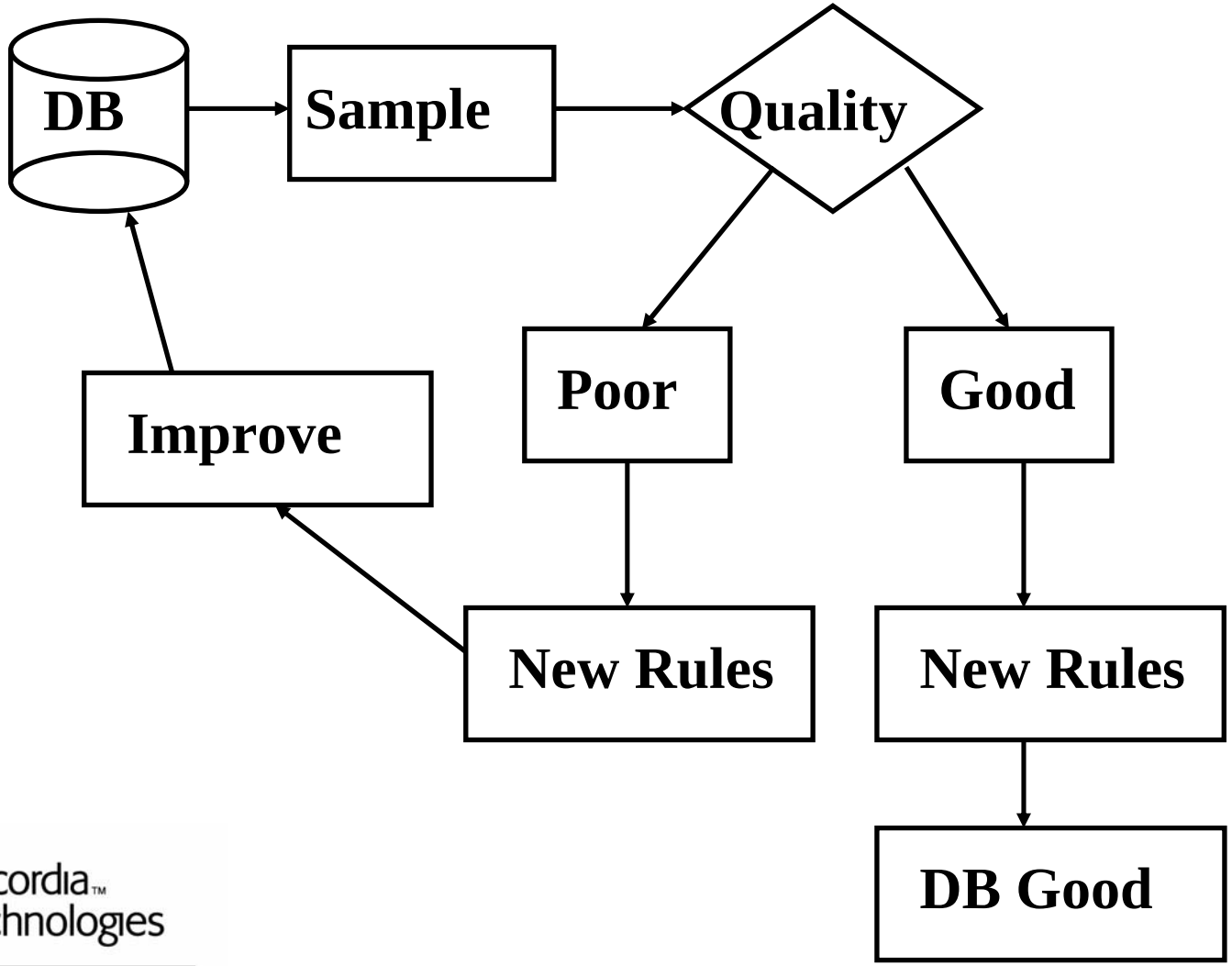
- •No more than 1% bad records with 90% probability
- •Specific % determined by economic analysis

**Use a sampling approach**

•Need algorithms to determine

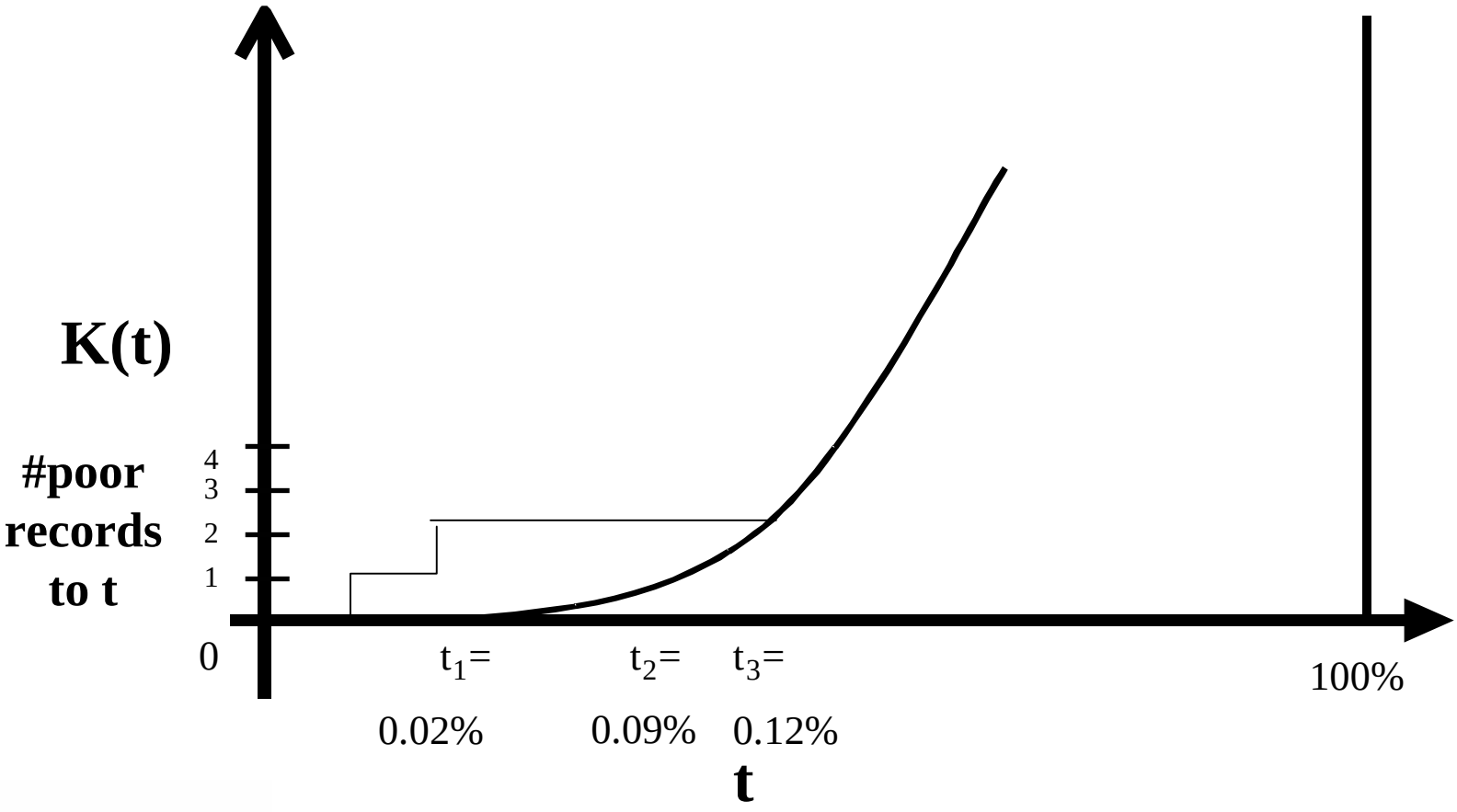
- •Sample Sizes (What should be sampled? How to sample?)
- •Precise measure of good quality in random sample which can be used for inference

# Data Quality Improvement- a new formulation




Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Algorithm 1: Total# of bad records small




1Million records, bad records<=1%, Probability =90%

Telcordia Technologies
Formerly Bellcore…
Performance from Experience

# Implications

- Efficient way of generating rules and correcting data
- Great deal of efficiency gain when the quality is good
- When quality is poor will end up sampling of a large percentage of data

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Information Retrieval and Automatic Recommendation

## The Problem

- how to find information on-line?
- how to use textual data automatically?

- tools have not kept pace with our ability to generate, store, and deliver digital information
  - 50% irrelevant
  - 75% missed
- problem is hard
  - textual data hard to model
  - word matching not effective enough
  - ex: viewgraphs, transparency, overhead, slide, ...
- need automatic method that captures inter-relationships among words and exploits them to improve retrieval accuracy as well as recommend options to the users

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Information Retrieval Approach: latent semantic indexing (LSI)

- truncated singular value decomposition
- words used in similar context will have similar coordinates in reduced space

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Analyzing Text -- Numeric Representation of Text Data

- represent document collection as term-by-document matrix

| | $d_1$ | $d_2$ | $d_3$ | $d_4$ | ..... | $d_m$ |
|---|---|---|---|---|---|---|
| $t_1$ | 1 | 0 | 1 | 1 | | 2 |
| $t_2$ | 0 | 0 | 0 | 1 | | 1 |
| $t_3$ | 2 | 1 | 0 | 0 | | 0 |
| . | | | | | | |
| . | | | | | | |
| . | | | | | | |
| $t_n$ | 0 | 1 | 2 | 0 | | 0 |

$\sim \Sigma_{nxm}$

  - Similarity between terms -- $\Sigma\Sigma^t$
  - Similarity between reports -- $\Sigma^t\Sigma$
  - Association between term and report -- $\Sigma$

Telcordia Technologies
Formerly Bellcore…
Performance from Experience

# Dimension Reduction Technique

- Singular value decomposition of

  $\Sigma = TSD^t$, where S is diagonal qxq with q= min(n,m)
- Pick k most significant singular values and vectors

  $\Sigma \sim \Sigma_k = T_k S_k D_k^t$, where k << q
- Term similarity ~ $\Sigma_k \Sigma_k^t = T_k S_k^2 T_k^t$
- Report similarity ~ $\Sigma_k^t \Sigma_k = D_k S_k^2 D_k^t$
- Rows of $T_k S_k$ -- vector representation of terms
- Rows of $D_k S_k$ -- vector representation of reports
- Term and report association ~ $\Sigma_k = T_k S_k D_k^t$

Telcordia™ Technologies

*Formerly Bellcore…*
*Performance from Experience*

# Information Retrieval

## The Data

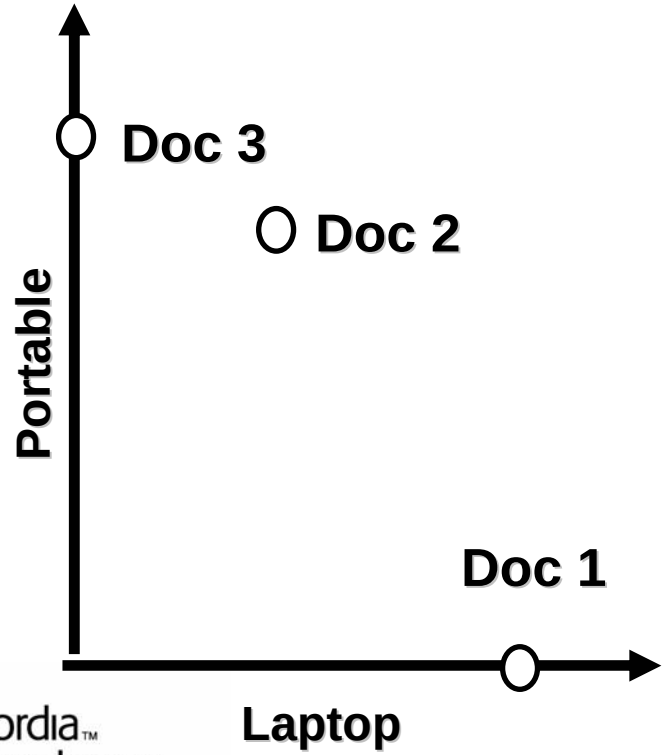
- ex: 1989 AP News articles
  - 198,000 unique terms (p)
  - 85,000 documents (n)
  - 17 billion cells
  - very sparse matrix (.002% non zero)
- Need to do search in reduced space

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

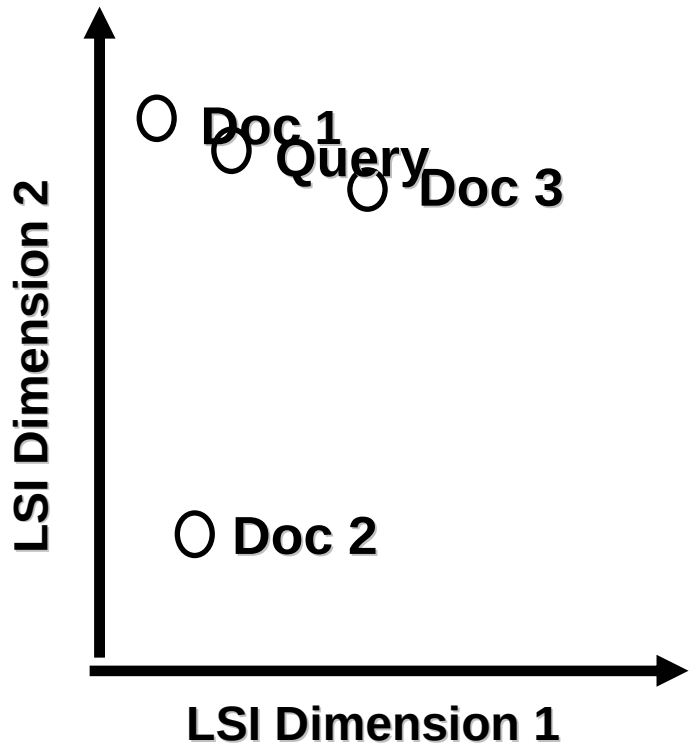
# Information Retrieval

## Approach Taken

**Keyword Retrieval:**
**Words Unrelated**




**LSI Retrieval:**
**Similar documents**



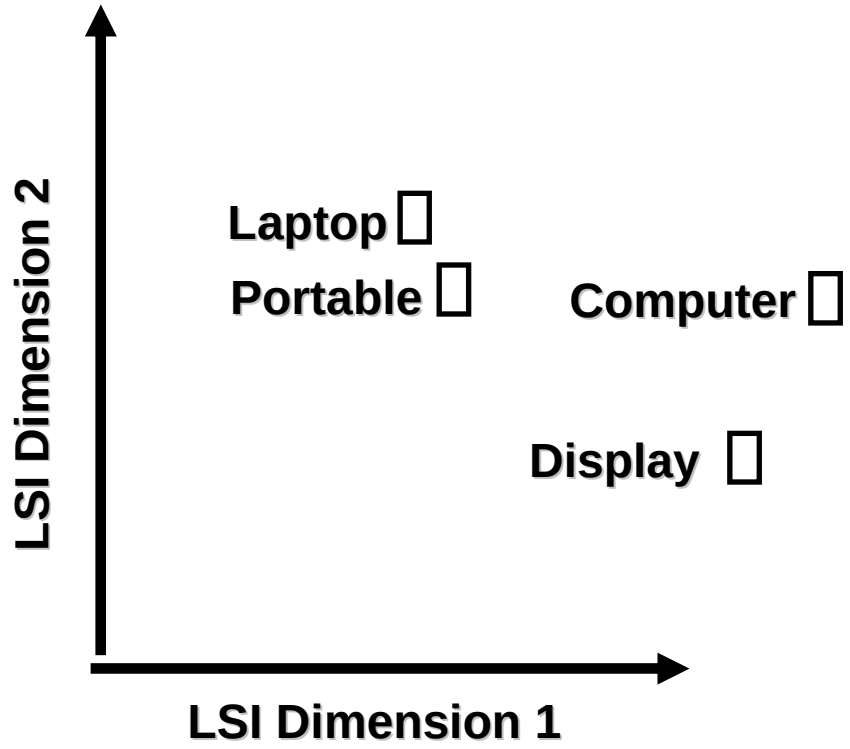

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Information Retrieval

## Approach Taken

**LSI Retrieval:**

**Similar words associated**




Telcordia™ Technologies

*Formerly Bellcore…*
*Performance from Experience*

# Information Retrieval

## Results Achieved

- 30% better than word matching methods
- added advantage for
  - search engine
  - cross-language applications
  - customer trouble reports
  - short texts (yellow pages)
  - noisy inputs (pen, OCR)

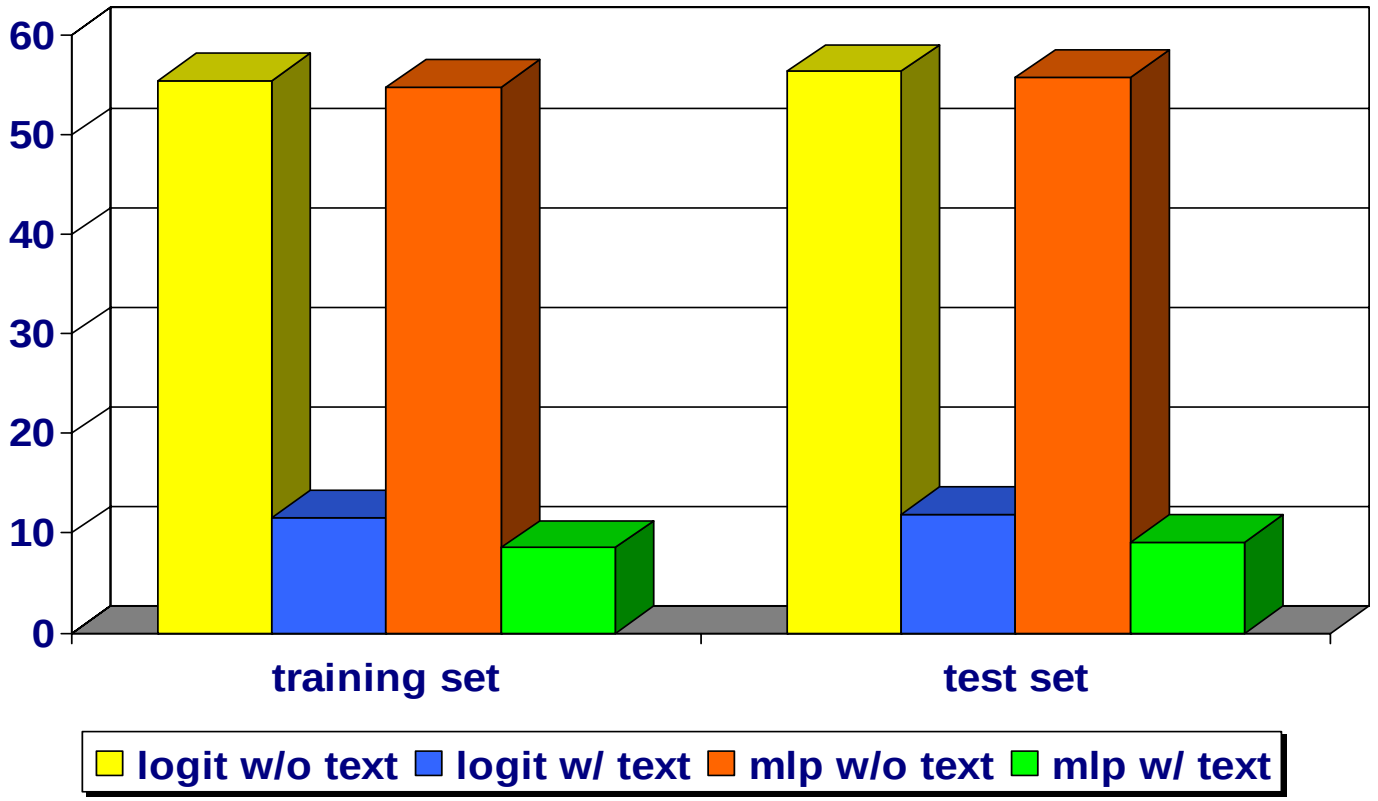
Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Customer Trouble Reports (CTRs)

- Customers' complaints about their phone services (42,206 CTRs)
- Significant amount of textual data: Customers' description
- Numeric/categorical data -- class of service, out of service or not, etc.
- Cost telecommunications industry hundreds of millions of dollars annually
- *Goal: correctly identify one of the 36 causes of a problem to reduce repair cost and improve customer satisfaction*

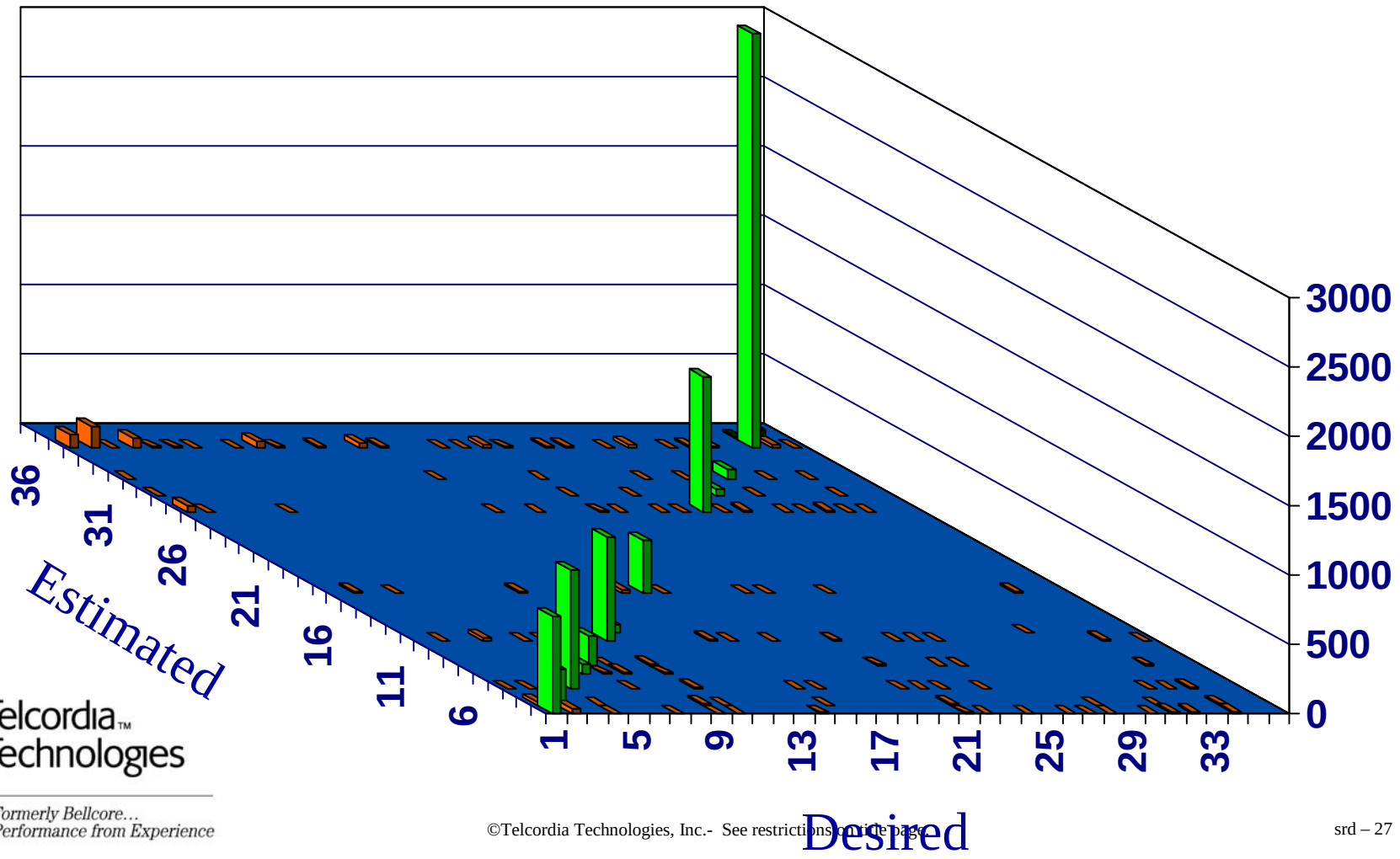
Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Classification Error Rates: 8441 Customer Trouble Reports




Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Confusion Matrix for 8441 test cases




Telcordia™ Technologies

*Formerly Bellcore…*
*Performance from Experience*

# Recommendation Technology:

- Collaborative Filtering
  - Recommend items close to what you are buying
  - Distance depends on just the items purchased
- Content based recommendation
  - Based on the content of a book (e.g. subject, abstract, etc.)
- Telcordia experience- LSI- 9% increase in # of items sold
    - 9% increase in items ordered
    - 6% increase in revenues

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Future Challenges

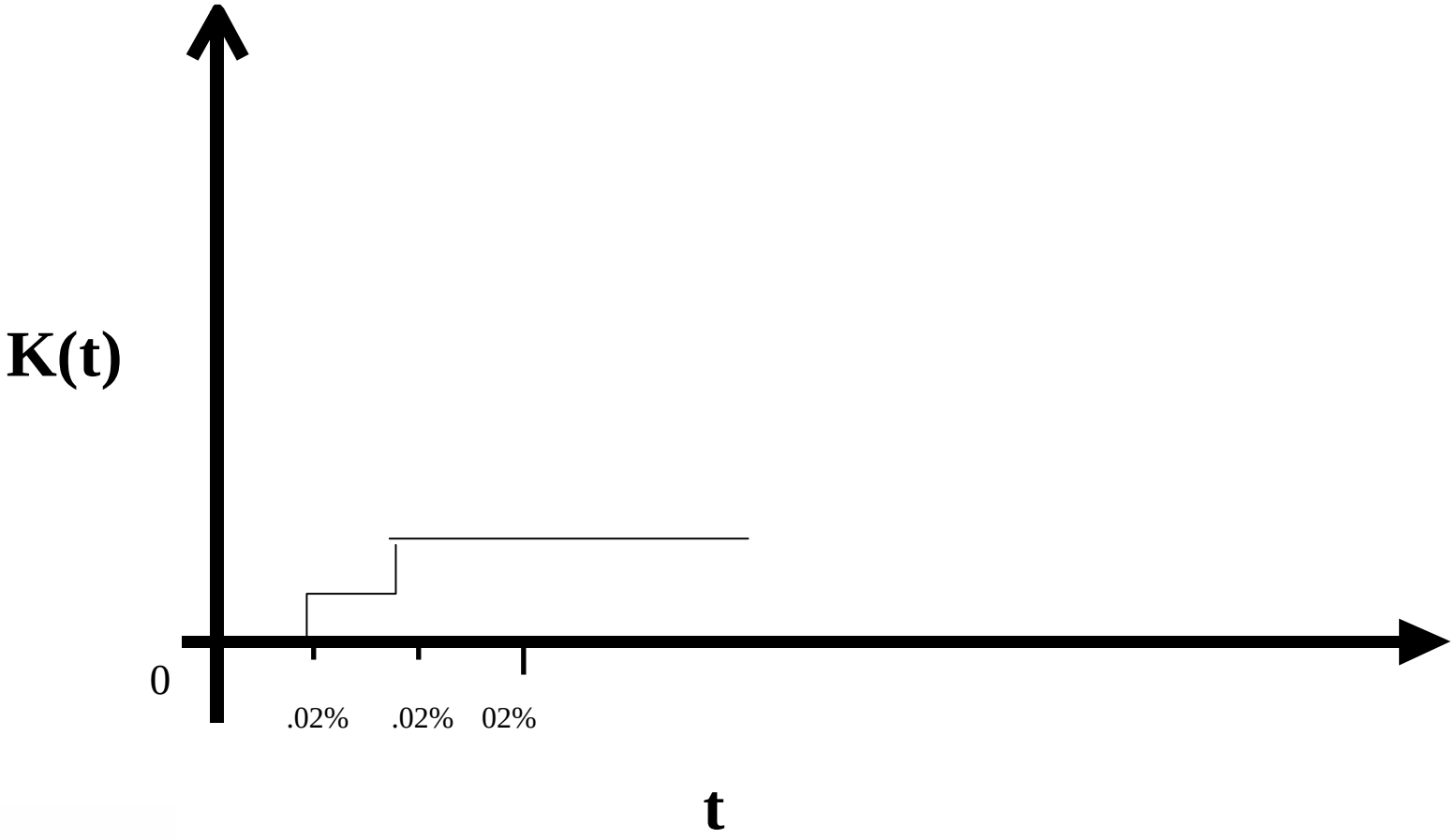
- Amalgamated Recommendation Methods
  - Collaborative filtering in conjunction with content based
  - Include demographics
- Privacy:
  - Anonomizer
  - Fulfillment
  - Handles and Trust Services
- Data Ownership

Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience

# Growth of E-Commerce

- Every prediction made of the growth of E-Commerce have been proven to be too conservative
- Present Prediction:
  - Revenue Projections $350B by 2001
  - Internet Economy  Projected to hit $1.3 Trillion by 2003
  - 80% Business to Business
    - Cost Reduction & Efficiency Gain
  - 20% Business to Consumers
    - Increase Revenue & Profitability
    - New Markets, Channels, Not bounded by geographical boundaries

# Algorithm 2:
# # of good records between bad records large




Telcordia™ Technologies
Formerly Bellcore…
Performance from Experience