

*The Next How Much Information?
East Coast Launch Event May 2008*



HMI? Program Launch

Concepts, Counts, Estimates, Initial Numbers

*Roger Bohn and James Short (UCSD)
May, 2008*



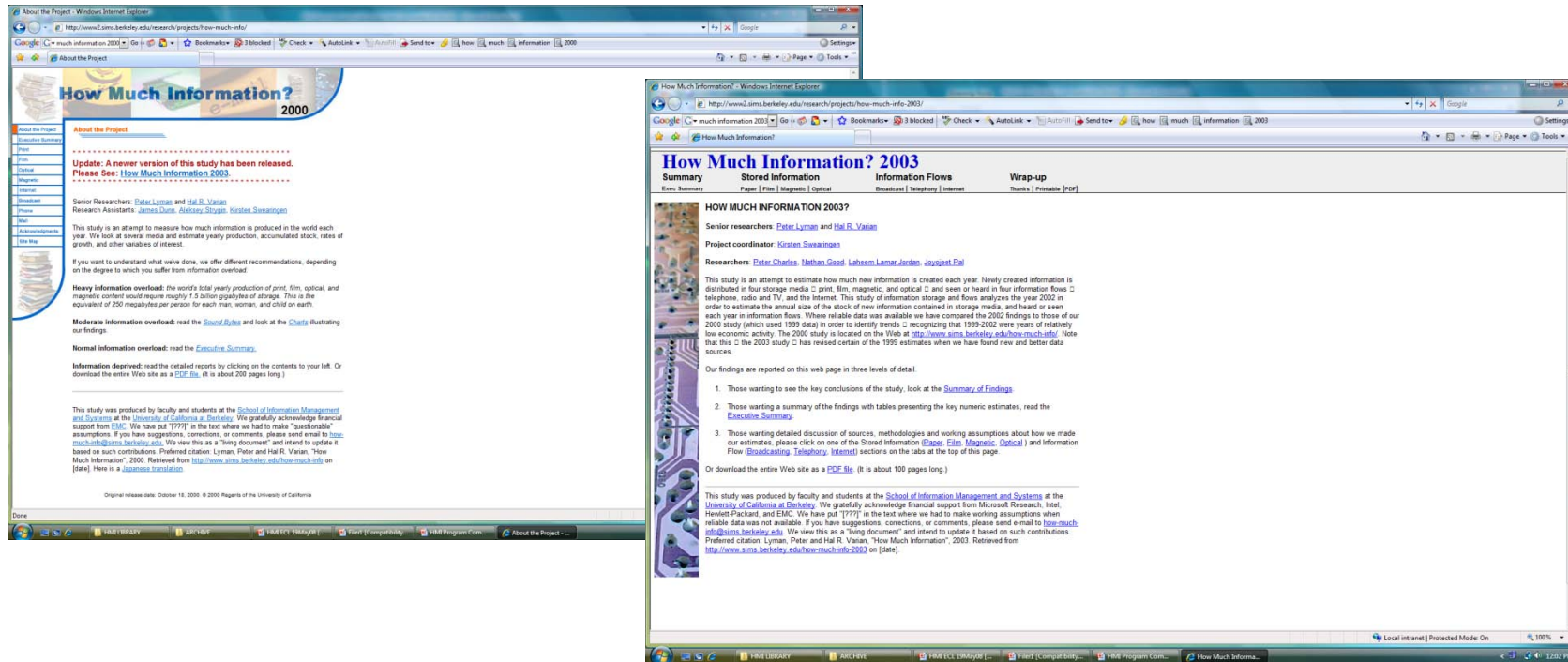
Background: How Much Information? Research Program Moves to UCSD

- UCSD's Information Storage Industry Center is taking over the research program developed by Lyman and Varian, How Much Information?

<http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>

- The Berkeley study generated enormous news media and corporate interest when published in 2000 and 2003
 - “The 2000 report generated so much interest, that Lyman and Varian did a follow-up study using 2002 data and published their report in 2003. ...”
- And it is still heavily referenced - a current Google search produces over 915,000 hits:
 - “Results 1 - 10 of about 915,000 for how much information berkeley? (0.27 seconds)”

Lyman Varian Reports...



- An update of the Lyman Varian How Much Information? reports, reconceptualization, expansion, application
- To create a new generation of ideas, reports and industry applications



HMI Program Questions

- How much new information is created each year? In enterprises? By people? By government?
- What is the rate of growth of new information each year?
- What factors will continue to drive information growth? What factors will inhibit it?
- What are the implications of the rate of growth of information for enterprises, for people, and for governments?



HMI Research Disciplines

- The program is technical, economic and managerial
 - Technical: measures, metrics, benchmarks for measuring and self diagnosing information growth, information value
 - Economic: extends work on the growth of information to work on the economic value of information and processes to maximize that value in enterprises
 - Managerial: addresses the senior business and IT management agenda for IT investment and enterprise information management



Why Another Study?

- “Previous studies, academic and analyst, do not reflect the state of industry knowledge”
 - In storage, networking, and database
 - Lack of focus on the enterprise and on personal information growth
- “Much improved measurement” (rigor needed)
 - In measuring, quantifying and modeling information growth
 - In defining and quantifying the value of information
 - Move beyond simple estimates based on product (box) sales data and use estimates
- “Need counts and estimates by use case”
 - Growth rates by industry verticals (use case studies) are the most interesting
 - Target industries: Healthcare, financial services, insurance, retailing
- “Need tools for self-diagnosis and benchmarking”
 - Managers need tools for diagnosing growth rates and implications
 - Direct measurement tools developed in test beds, applied in use case studies

Looking Back: Four Generations of Information Research Summarized

- Information Theory (mathematics)
 - Claude Shannon (MIT)
 - 1948: A Mathematical Theory of Communication
 - Introduced concept of information entropy (uncertainty)
 - Factoid: worked as messenger for Western Union
- Information Economics
 - Fritz Machlup, The Production and Distribution of Knowledge in the United States (1962)
 - Knowledge industry represented 29% of the US GNP
 - Knowledge defined as <information in use>
 - Focus on **production** and **distribution**



Theseus 1950



Looking Back: Four Generations of Information Research Summarized

- Communications and The Flow of Information
 - Ithiel Pool (MIT)
 - 1983 Science: Tracking The Flow of Information
 - Created novel indices to show that transmission of information was much higher than consumption
 - Focus on **production**, **distribution**, and **consumption**
 - **Information** defined as production (data), receipt, and use (consumption)
- How Much Information?
 - Peter Lyman and Hal Varian (UC Berkeley)
 - 2000 and 2003 reports generating baseline data
 - Tracked information stored in four physical media and seen or heard in four information flows

Counting Words Made Available By Volume and Costs of Communication

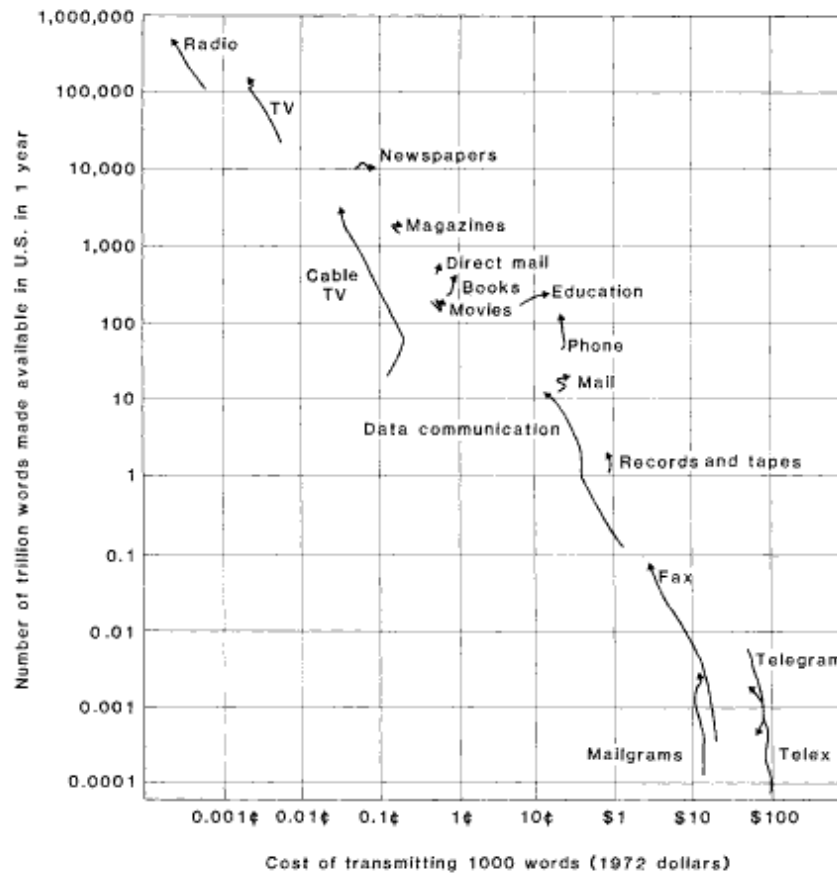


Fig. 1. Trends in volume and costs of communication for 17 media, 1960 to 1977 (plotted on log by log scales).

610

Ithiel Pool, Tracking The Flow of Information, Science 1983

How Much Data? Human Conversation

Are Women Really More Talkative Than Men?

6 JULY 2007 VOL 317 SCIENCE

Matthias R. Mehl,^{1*} Simine Vazire,² Nairán Ramírez-Esparza,³
Richard B. Slatcher,² James W. Pennebaker³

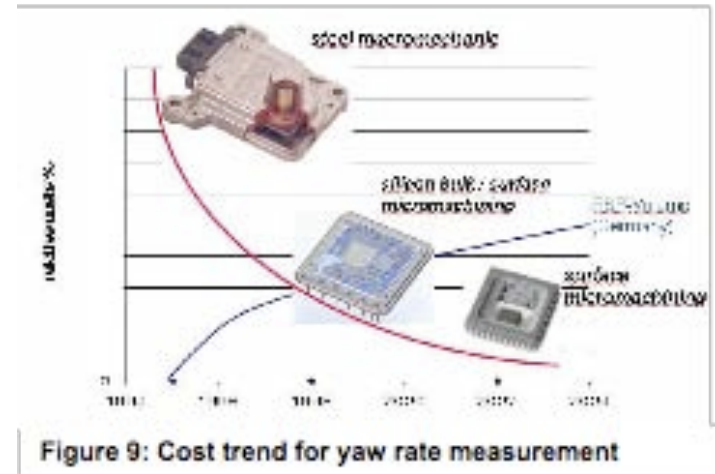
Women 16,215 (± 7301)

Men 15,669 (± 8633) per day

	Text	Audio	Video
Bytes/word	5	2.5KB	30KB
Bps@200 wpm	16.5	8KB	100KB
Per day @16000 words per day	90KB	40MB	480MB
Per person-year	33MB	14.6GB	175GB
Per Earth-year @6.6B people per planet	0.2 E18	97 E18	1.2 E21

1E18 = 10^{18} = 1 exabyte

Auto sensors: new apps, Moore's Law

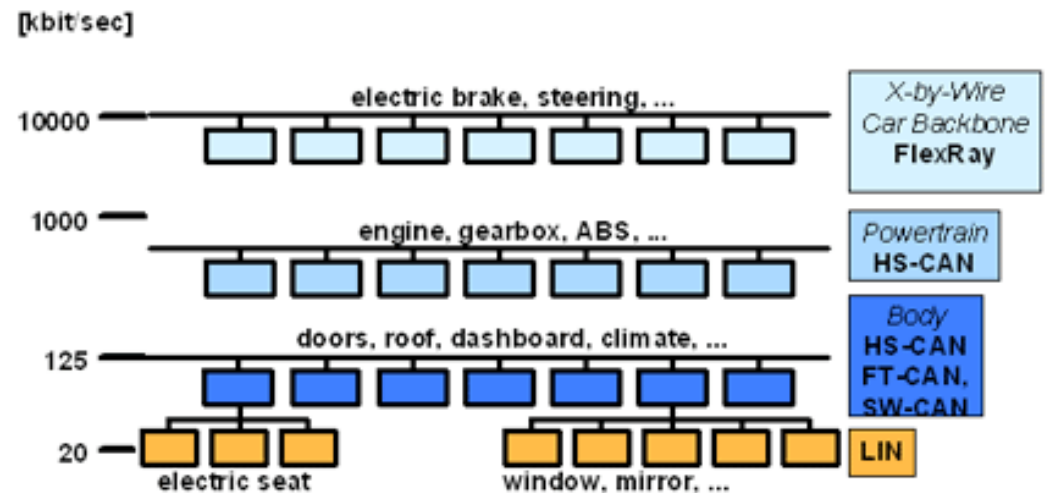


Freescale Ships Over 100 Million Automotive Microcontrollers Annually

TABLE IV
SENSORS USED IN BODY APPLICATIONS

FUNCTION	BODY SENSOR	PRODUCTION STATUS*
SAFETY		
Air Bag Actuation	Crash Deceleration	major
	Vehicle Rollover (Lateral Acceleration plus Roll Rate)	R&D
	Sec. Belt-Use Buckle Status	limited
Seat Belt Locking	Pressure (Side Impact)	limited
	Vehicle Deceleration	major
Seat Occupancy	Vehicle yaw/roll Velocity	limited
	Seat Fish Bladder Pressure	R&D
Occupant Presence/Pre-Crash Position	Seat Pan Load/Deflection	R&D
	Passive Infrared Imaging	R&D
Parking/Reversing Aid	Ultrasonic Imaging	R&D
	Machine Vision	R&D
Blind Spot Surveillance	Ultrasonic Array	major
	Wide-Beamwidth Radar	limited
Lane Departure	Wide-Beamwidth Radar	limited
	Multi-Beam Infrared Laser Array	R&D
Night Vision	Machine Vision	limited
	Passive Infrared Imaging	limited
INTELLIGENT TRANSPORTATION	Active near-IR Illumination	R&D
	Adaptive Cruise Control	Millimeter Wave Radar
	Infrared Laser Radar	limited

Auto data rates




- Bus rates > 1 Mbps; now a 10Mbps bus
- Average > 10 MCUs per new car
- 500 hours/yr x 2 Mbps = .5 TB/yr = .5E12
- US has 10^8 autos
- 50 Exabytes/year of data
- Save it? How long?

At present, most data stays local

Data transformed	Data stored?	Example
Events		speech
Measured (analog)		Video surveillance
	Stored locally	VHS recorder
A to D		Camera
	Stored locally	Aircraft black box
Sent externally		Email photos
Broadcast/downloaded	Replicated	You-tube
	Backed up	CERN data

1000X more data *could* be captured


Potential value versus current (kinetic) value



HMI? 2009

What Are We Counting?


- Problem 1: Most information is never recorded. It may be
 - Produced, consumed, and thrown away (most conversations)
 - Produced and not consumed (TV on in the background)
 - Produced, consumed and recorded, but never distributed (audience of one)
- In enterprises, there are multiple channels and presentation formats for communications. How many are digitized? How many are recorded? How many are distributed?
- There is a science of defining, estimating and counting spoken communications



HMI? 2009

There Are No Easy Measures


- Problem 2: There are no easy measures or short-cuts to measuring consumption and use (application)
- We can get to aggregate measures of stock (stored data) and flow (data flows over networks) – top down measures (**easiest**)
- Even in complex systems environments, we can simplify down to getting measures of data at rest in DBs and data in flow (traffic) over networks (**hard but**)
 - The challenge is addressing domain specificity in instrumenting the test bed, interpreting the results (science and art), and generalizing results



HMI? 2009

Consumption and Use Requires Additional Measures

- Consumption (who consumes), use (for what purpose) and utility (to what end) require
 - Knowledge of who or what is receiving the information (could be a machine)
 - A measure of consumption
 - A measure of use (fit to purpose)
 - A measure of utility (how important to task)



HMI? 2009

Constraints Drive Contrasting Streams of Work

- Constraints have driven two streams of work
 - Aggregated studies of data production, storage and flow, using modified forecasting methods - these can get quite involved but the data is there and can be counted using conventional methods (although accuracy is ?)
 - Case level instances build around use examples, utility can be viewed as proxy for value, how representative is the question (and not cumulative)