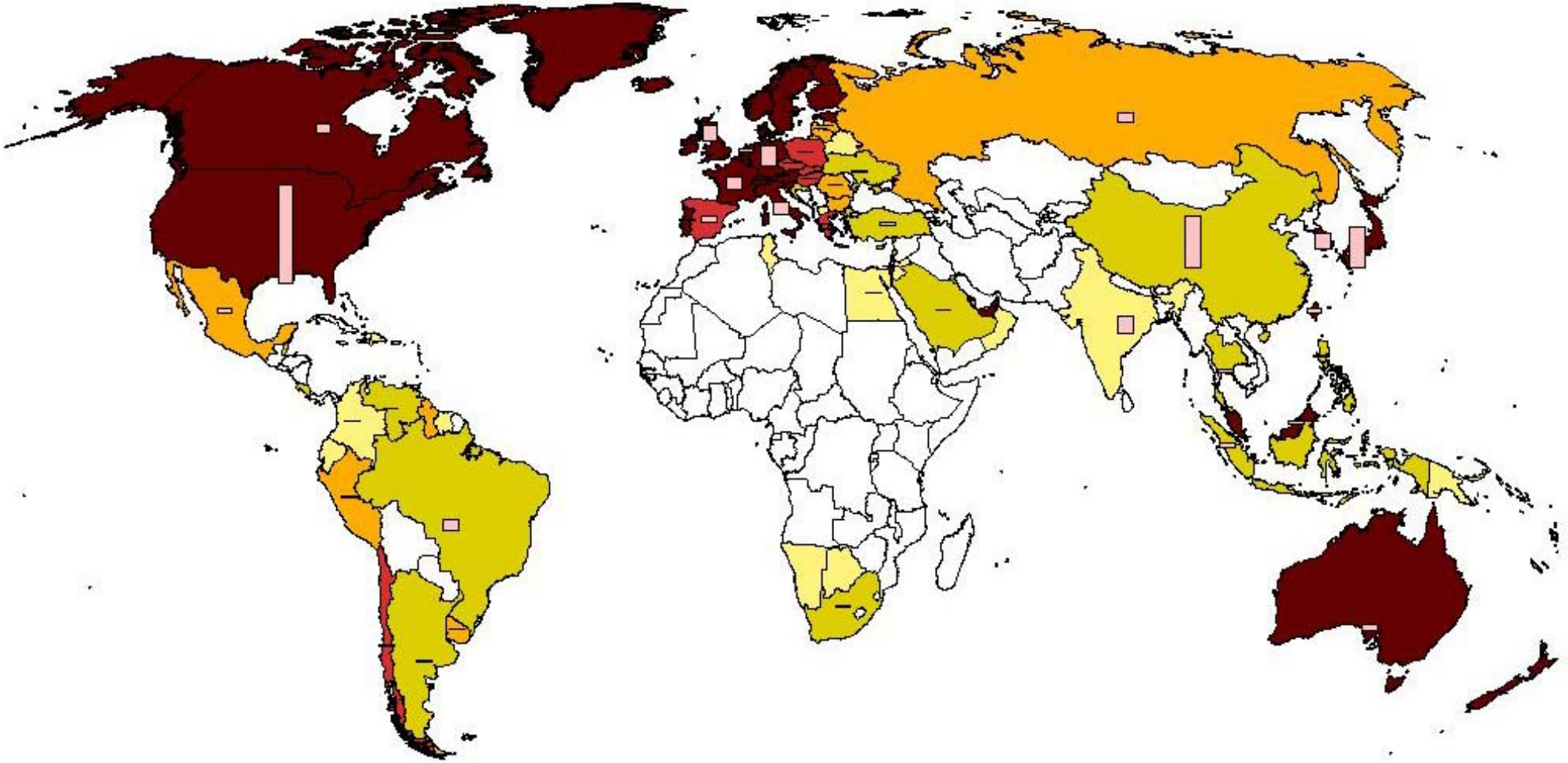
Providing Internet Search for Low-Connectivity Communities

Saman Amarasinghe
William Thies

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology

Google, New York, NY – July 2005
TIER Workshop, Berkeley, CA – October 2005

Internet Users Worldwide



Percentage of Country's Population Online

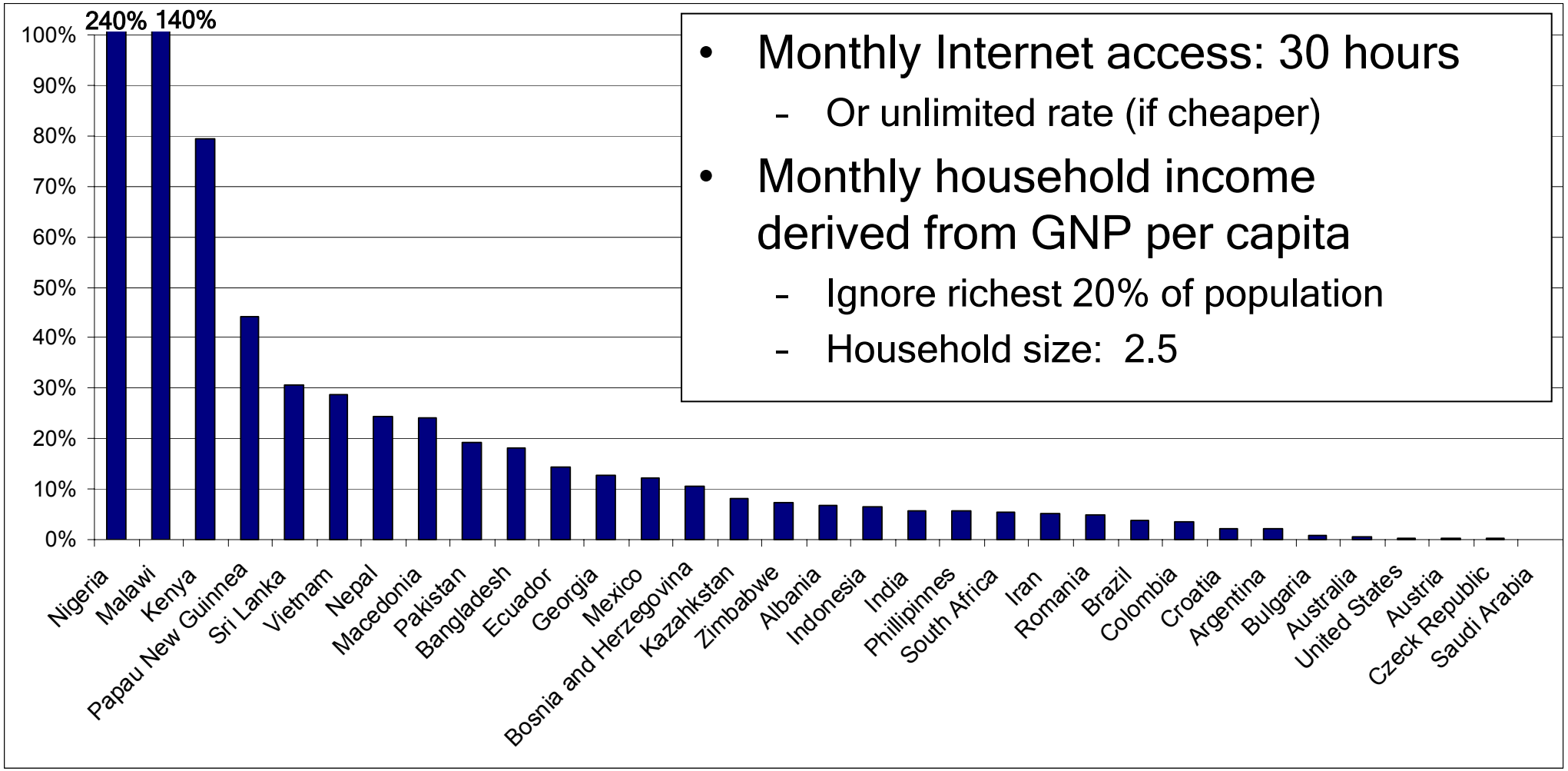
< 2 % | 2 - 4.9 % | 4.9 - 13 % | 13 - 25 % | 25 - 35 % | > 35 %

Share of World's Internet Users

Barriers to Internet Access

- Infrastructure
  - Limited phone lines
  - Low-bandwidth international links
  - Unreliable power supplies
- High costs
  - Computer unaffordable or unavailable
  - ISP, telephone costs can exceed local wage
  - Exacerbated by slow connections
- Social barriers
  - Illiterate or non-technical users
  - Lack of local content

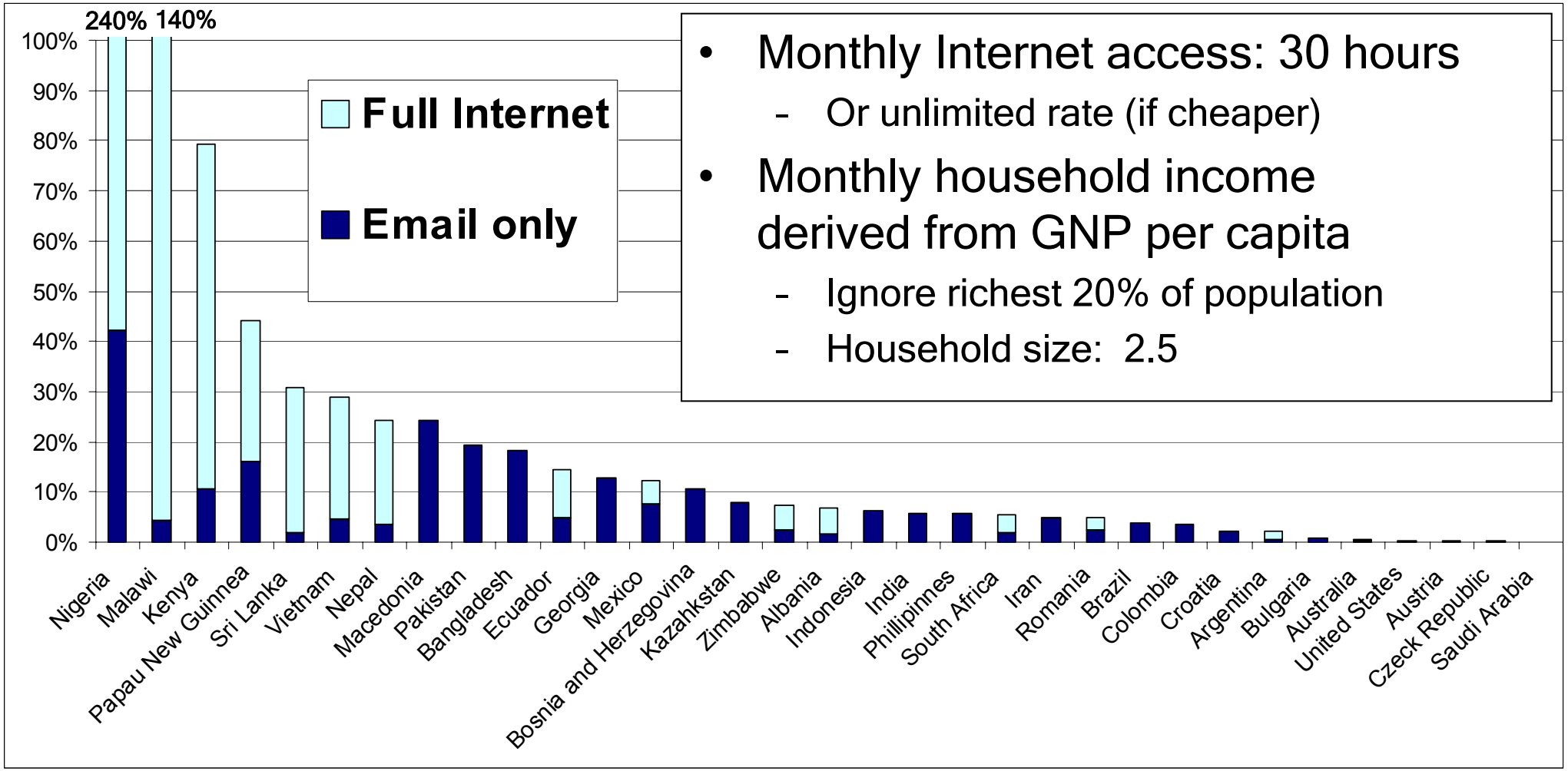
Cost of Dial-up Internet Access as a Fraction of Household Income



- Monthly Internet access: 30 hours
  - Or unlimited rate (if cheaper)
- Monthly household income derived from GNP per capita
  - Ignore richest 20% of population
  - Household size: 2.5

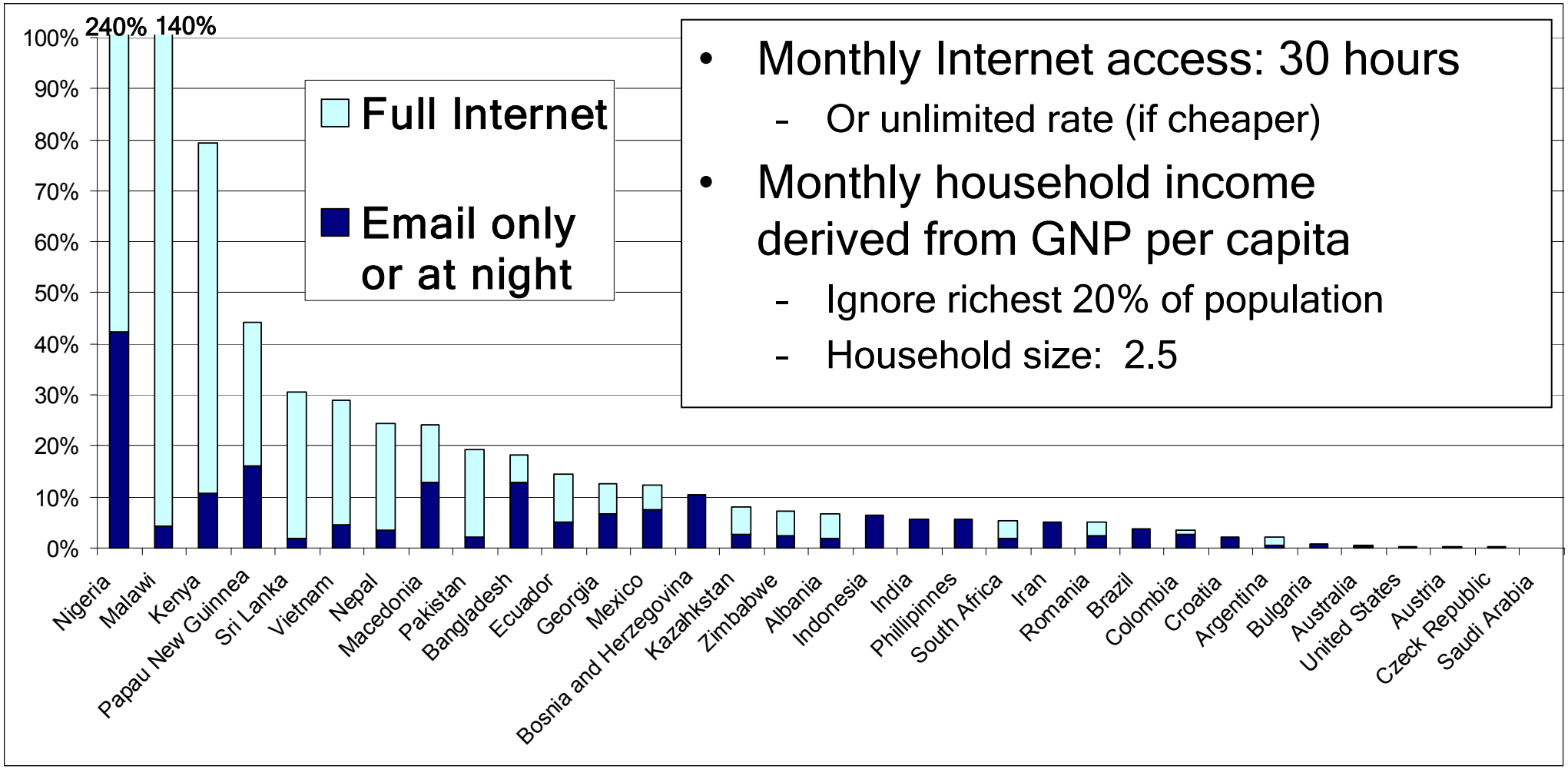
Sources: ISP Websites 2005, UNDP Development Report 2004, WorldBank 2003

Cost of Dial-up Internet Access as a Fraction of Household Income



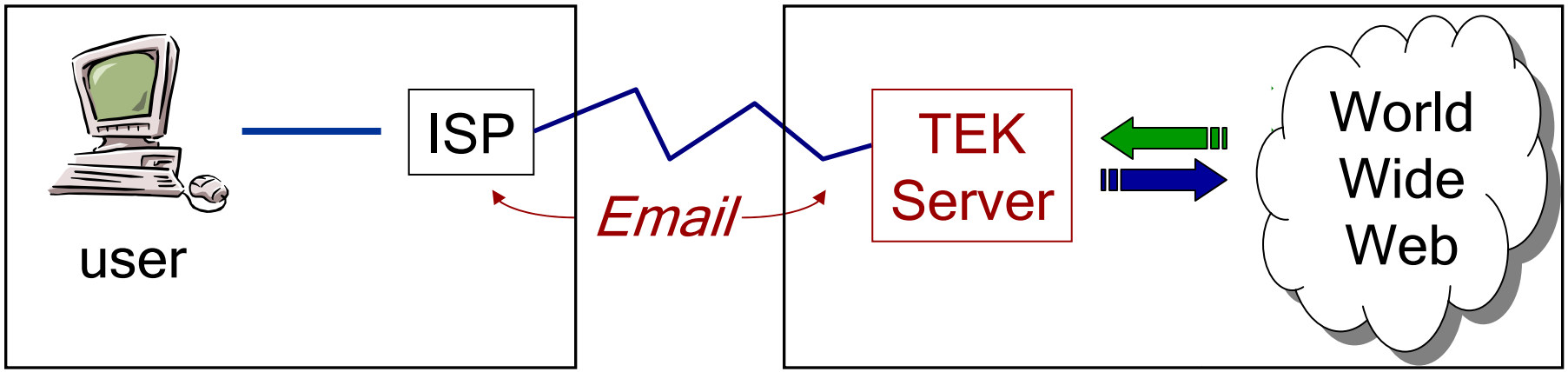
Sources: ISP Websites 2005, UNDP Development Report 2004, WorldBank 2003

Cost of Dial-up Internet Access as a Fraction of Household Income



Sources: ISP Websites 2005, UNDP Development Report 2004, WorldBank 2003

TEK: Email-Based Search



Solution has two components:

1. Transfer all data through email, not http
   - Connect only to send/receive email, not to browse web
2. TEK Server optimizes for bandwidth requirements

TEK: "Time Equals Knowledge"

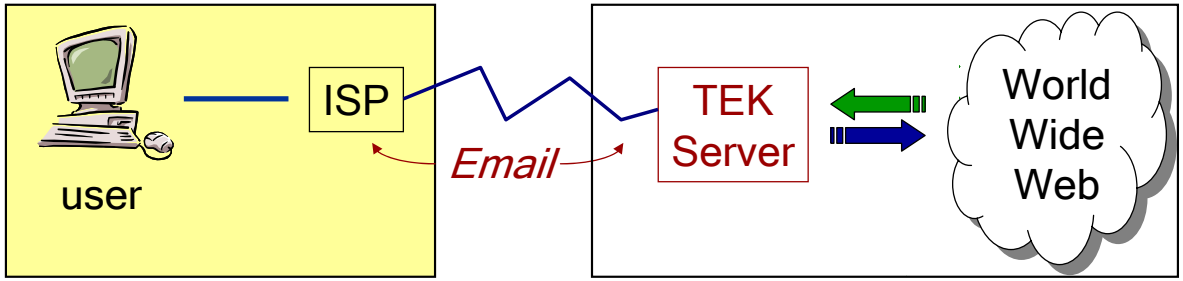
Outline

- TEK System
- Usage Scenarios
- Optimizing for Bandwidth

Outline

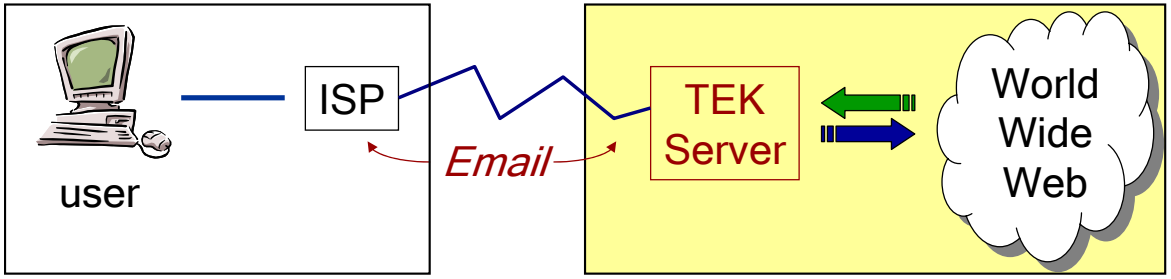
- TEK System
- Usage Scenarios
- Optimizing for Bandwidth

TEK Client

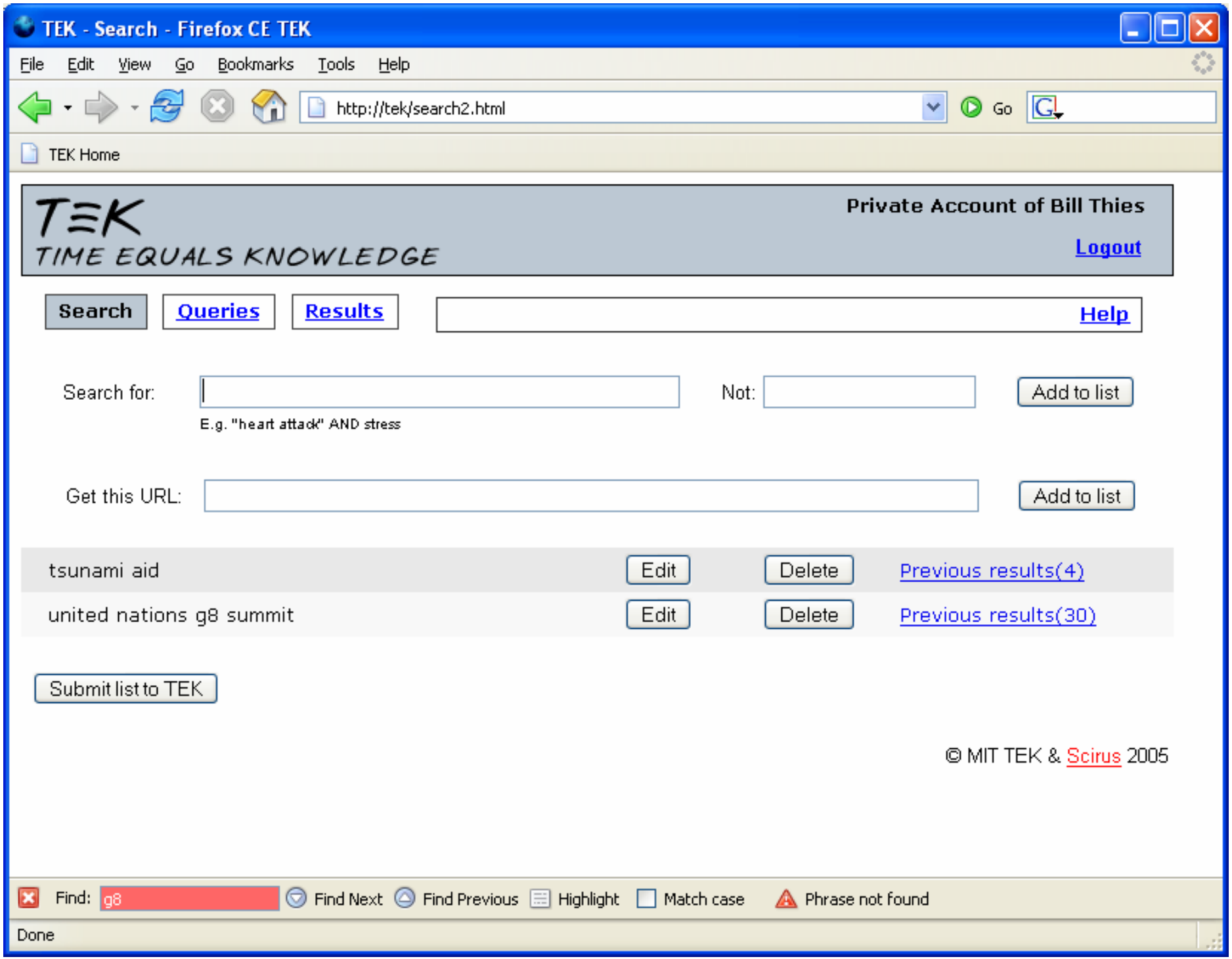


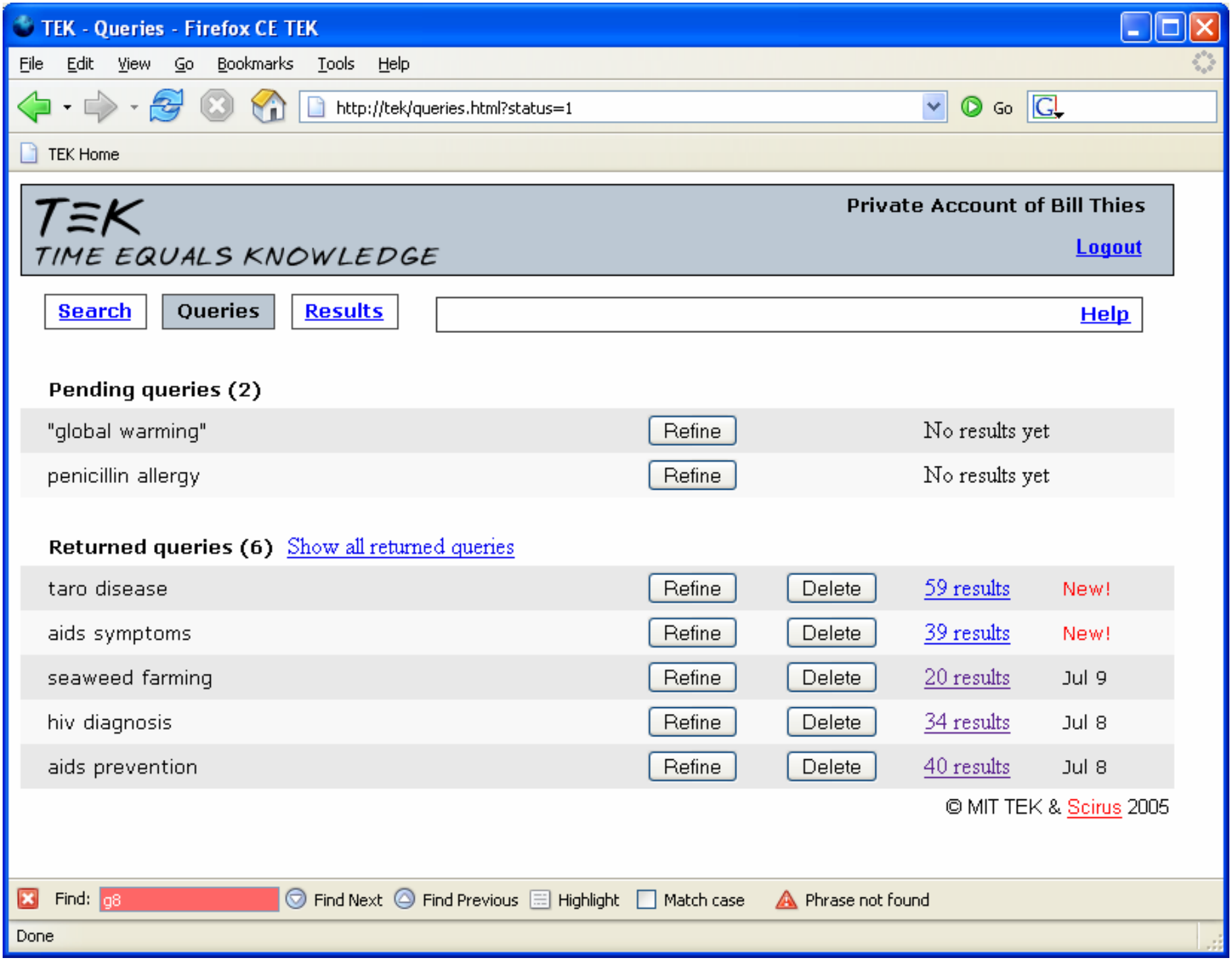
- Implemented as an HTTP Proxy Server bundled with a custom version of Firefox
- When offline, users can:
  - Search and browse old results as if connected
  - Enqueue queries for new results or missing pages
- When online, users can:
  - Send pending queries
  - Receive new results (attached to standard emails)

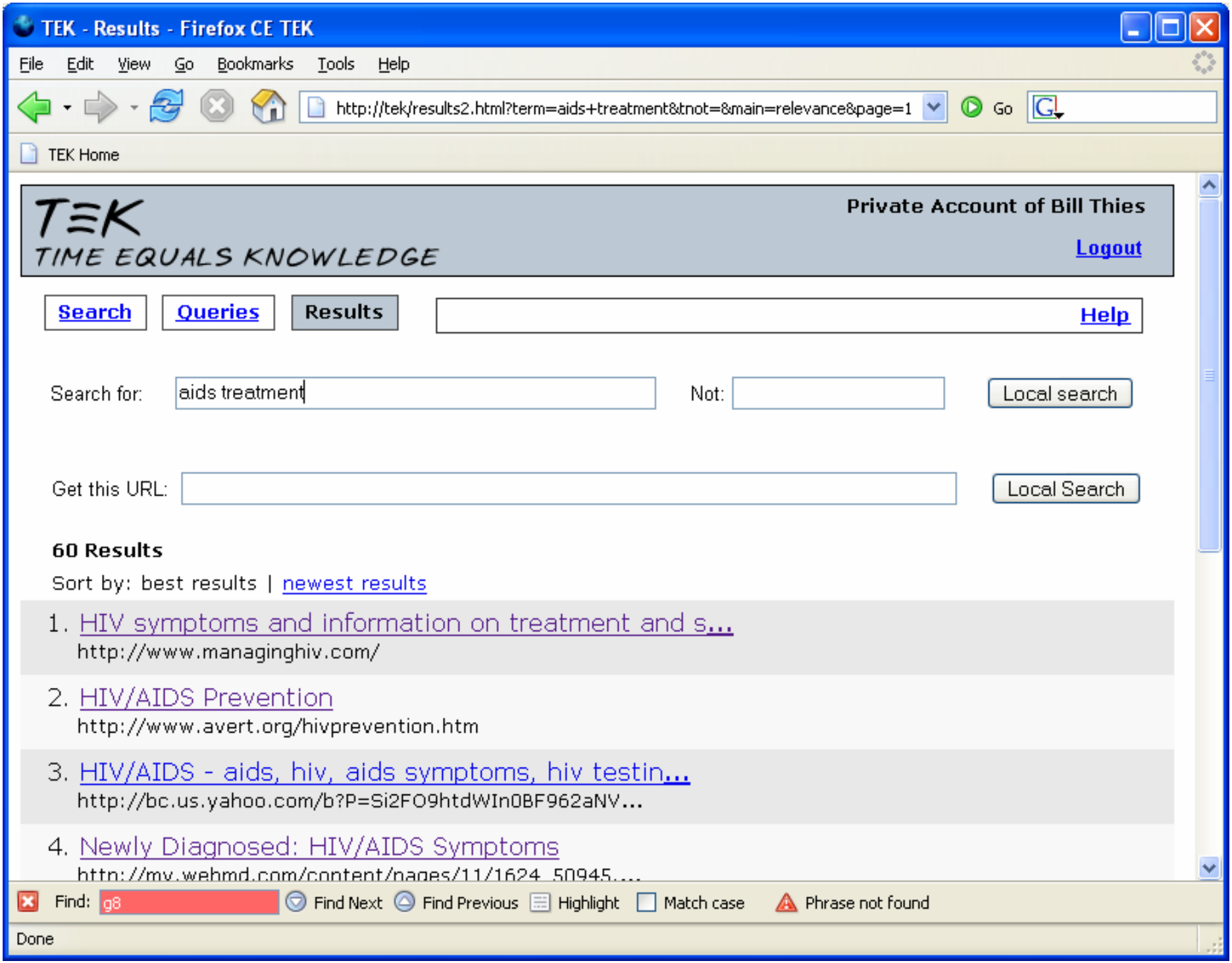
TEK Server

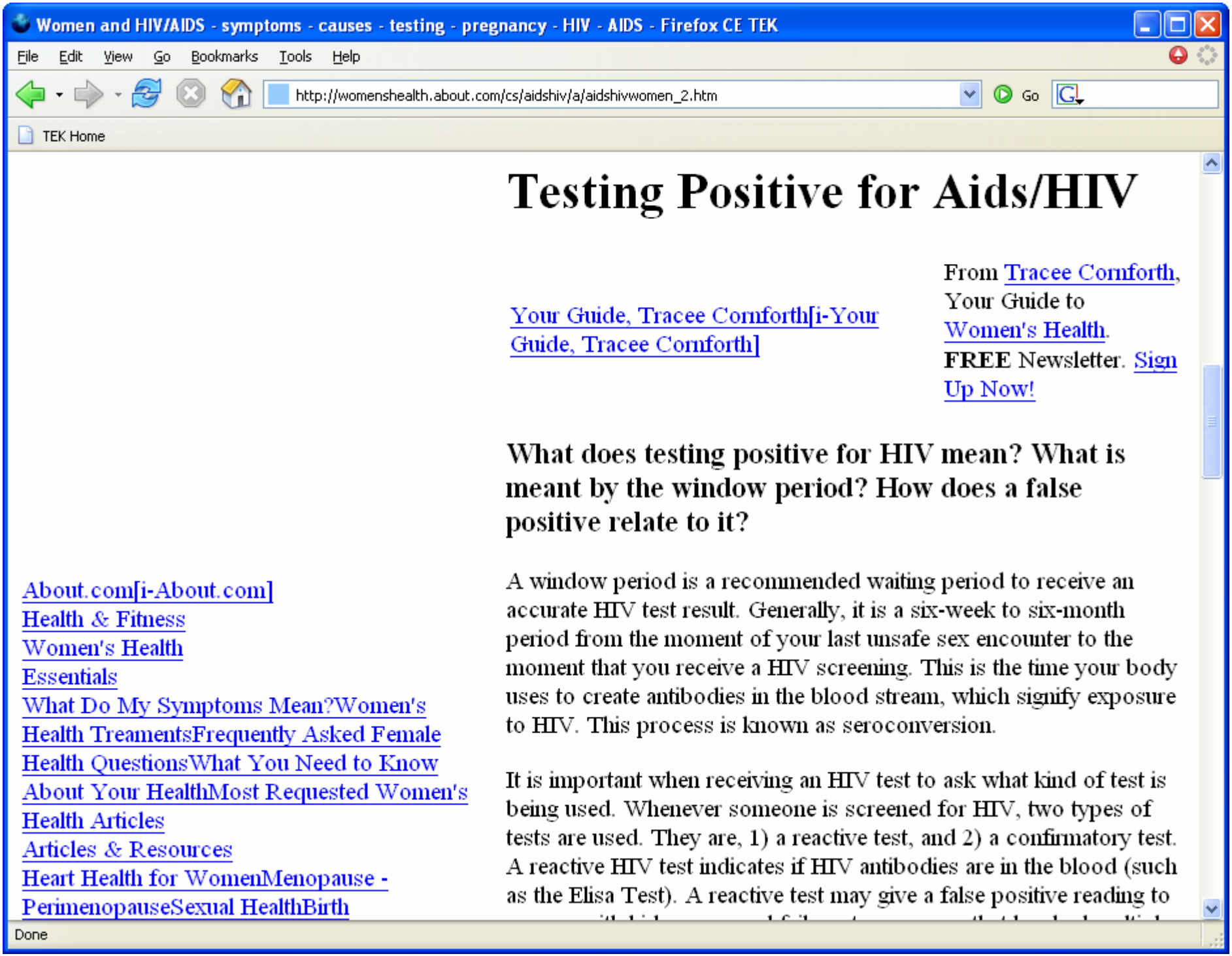


- Queries Google for relevant pages
- Returns filtered content of ~20 pages to user
  - Remove images
  - Remove junk HTML (JavaScript, colors, meta tags, etc.)
    - Uses loband library for page simplification (loband.org)
  - Convert PDF, PS to HTML (uses pdftohtml)
- Maintain server image of client page cache
  - Avoid sending duplicate pages
- Compress pages, send as single attachment
  - Limit attachment size to 150K (or smaller, for some users)


TEK - Search - Firefox CE TEK
File Edit View Go Bookmarks Tools Help
http://tek/search2.html
Go
TEK Home
TEK
TIME EQUALS KNOWLEDGE
Private Account of Bill Thies
Logout
Search
Queries
Results
Help
Search for:
Not:
Add to list
E.g. "heart attack" AND stress
Get this URL:
Add to list
tsunami aid
Edit
Delete
Previous results(4)
united nations g8 summit
Edit
Delete
Previous results(30)
Submit list to TEK
© MIT TEK & Scirus 2005
Find: g8
Find Next
Find Previous
Highlight
Match case
Phrase not found
Done

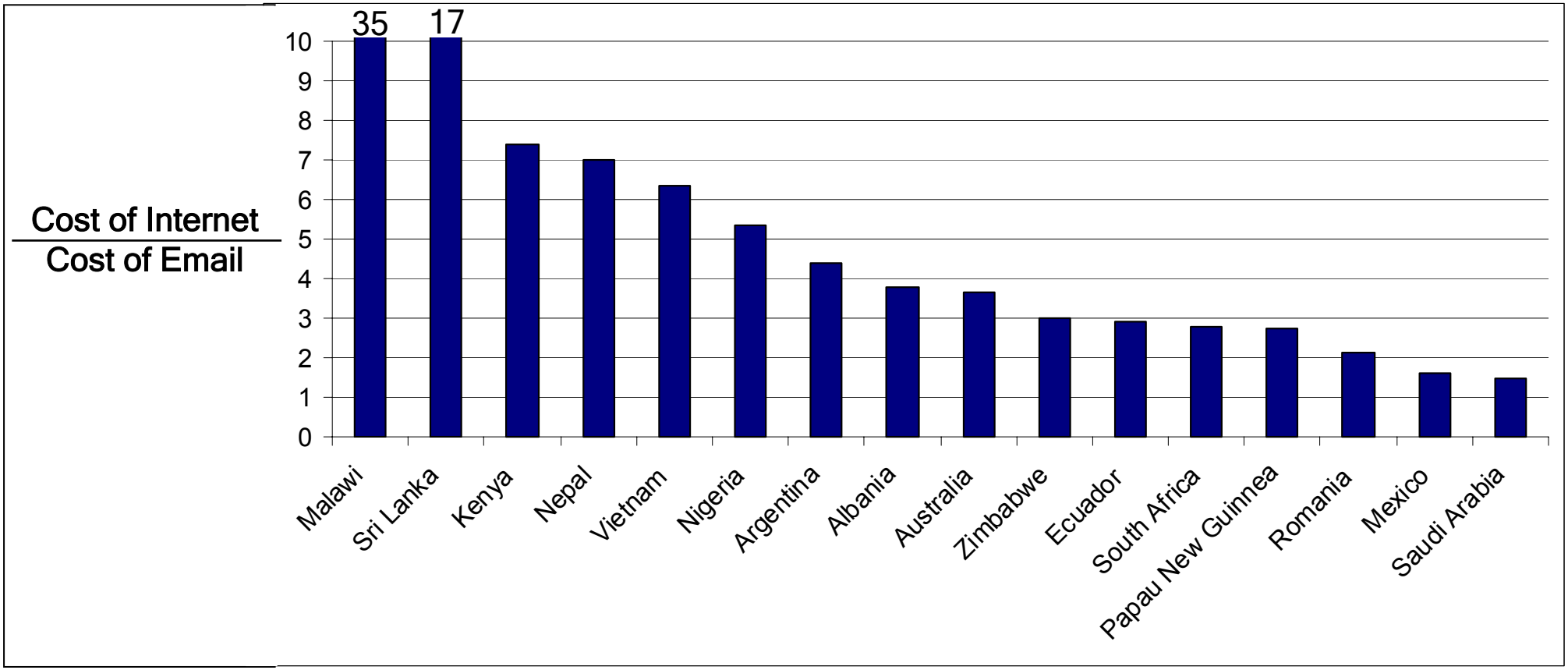

TEK - Queries - Firefox CE TEK
File Edit View Go Bookmarks Tools Help
http://tek/queries.html?status=1
Go
TEK Home
TEK
TIME EQUALS KNOWLEDGE
Private Account of Bill Thies
Logout
Search
Queries
Results
Help
Pending queries (2)
"global warming" Refine No results yet
penicillin allergy Refine No results yet
Returned queries (6) Show all returned queries
taro disease Refine Delete 59 results New!
aids symptoms Refine Delete 39 results New!
seaweed farming Refine Delete 20 results Jul 9
hiv diagnosis Refine Delete 34 results Jul 8
aids prevention Refine Delete 40 results Jul 8
© MIT TEK & Scirus 2005
Find: g8 Find Next Find Previous Highlight Match case Phrase not found
Done


TEK - Results - Firefox CE TEK
File Edit View Go Bookmarks Tools Help
http://tek/results2.html?term=aids+treatment&tnot=&main=relevance&page=1
Go
TEK Home
TEK
TIME EQUALS KNOWLEDGE
Private Account of Bill Thies
Logout
Search
Queries
Results
Help
Search for:
aids treatment
Not:
Local search
Get this URL:
Local Search
60 Results
Sort by: best results | newest results
1. HIV symptoms and information on treatment and s...
http://www.managinghiv.com/
2. HIV/AIDS Prevention
http://www.avert.org/hivprevention.htm
3. HIV/AIDS - aids, hiv, aids symptoms, hiv testin...
http://bc.us.yahoo.com/b?P=Si2FO9htdWIn0BF962aNV...
4. Newly Diagnosed: HIV/AIDS Symptoms
Find: g8
Find Next
Find Previous
Highlight
Match case
Phrase not found
Done


Women and HIV/AIDS - symptoms - causes - testing - pregnancy - HIV - AIDS - Firefox CE TEK
File Edit View Go Bookmarks Tools Help
http://womenshealth.about.com/cs/aidshiv/a/aidshivwomen_2.htm
Go
TEK Home
Testing Positive for Aids/HIV
Your Guide, Tracee Cornforth[i-Your Guide, Tracee Cornforth]
From Tracee Cornforth, Your Guide to Women's Health.
FREE Newsletter. Sign Up Now!
What does testing positive for HIV mean? What is meant by the window period? How does a false positive relate to it?
A window period is a recommended waiting period to receive an accurate HIV test result. Generally, it is a six-week to six-month period from the moment of your last unsafe sex encounter to the moment that you receive a HIV screening. This is the time your body uses to create antibodies in the blood stream, which signify exposure to HIV. This process is known as seroconversion.
It is important when receiving an HIV test to ask what kind of test is being used. Whenever someone is screened for HIV, two types of tests are used. They are, 1) a reactive test, and 2) a confirmatory test. A reactive HIV test indicates if HIV antibodies are in the blood (such as the Elisa Test). A reactive test may give a false positive reading to
About.com[i-About.com]
Health & Fitness
Women's Health
Essentials
What Do My Symptoms Mean?Women's Health TreamentsFrequently Asked Female Health QuestionsWhat You Need to Know About Your HealthMost Requested Women's Health Articles
Articles & Resources
Heart Health for WomenMenopause - PerimenopauseSexual HealthBirth
Done

TEK Rationale I: More Affordable

- Email accounts cheaper than web access

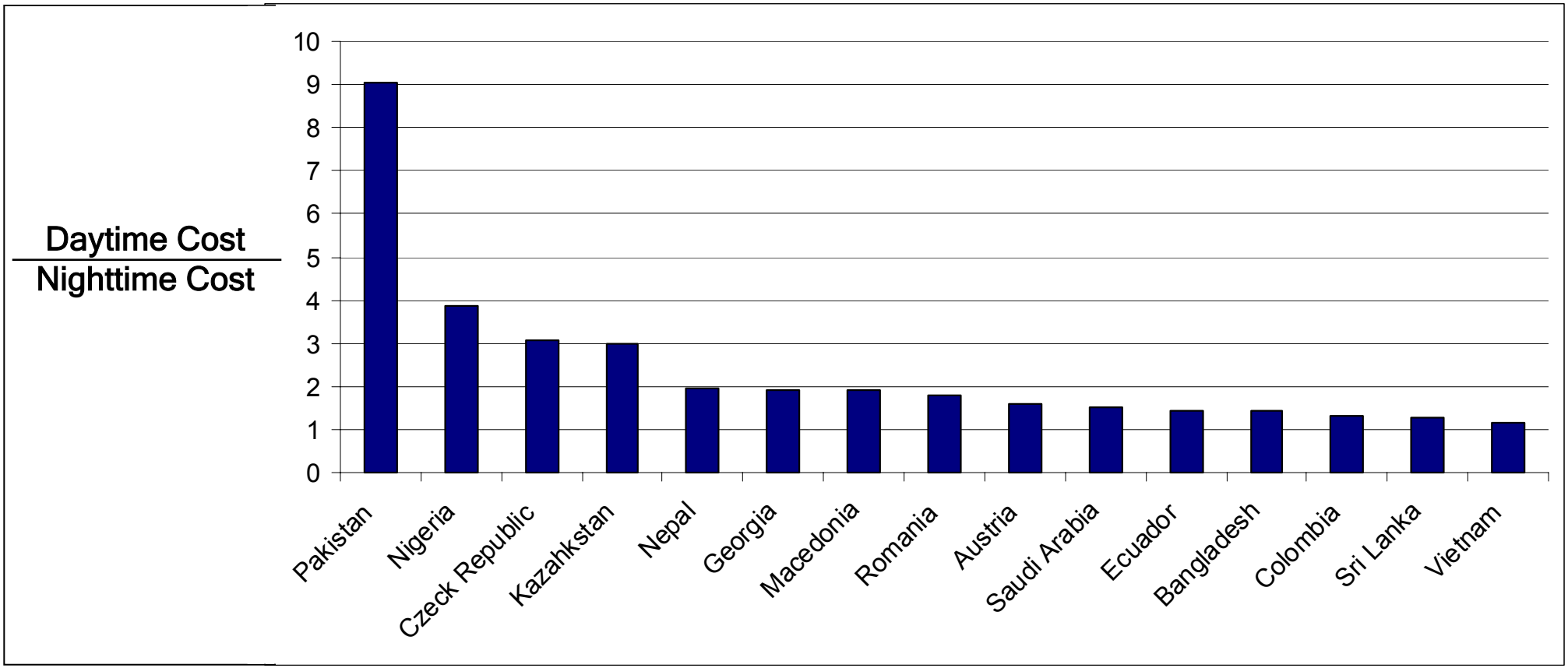


TEK Rationale I: More Affordable

- Email accounts cheaper than web access
  - Some infrastructures support email only

TEK Rationale I: More Affordable

- Email accounts cheaper than web access
  - Some infrastructures support email only
- Can send/receive all queries at night



TEK Rationale I: More Affordable

- Email accounts cheaper than web access
  - Some infrastructures support email only
- Can send/receive all queries at night
- Connection time is shorter
  - Avoids reading pages online
  - Content direct from ISP, not distant server
  - Server compression shrinks results

TEK Rationale II: More Usable

- Viewing results offline: quick, reliable
  - Establish local database of shared information
- In school: time-share Internet line with voice
  - Reduced time online makes Internet viable
- Manageable amount of information

Outline

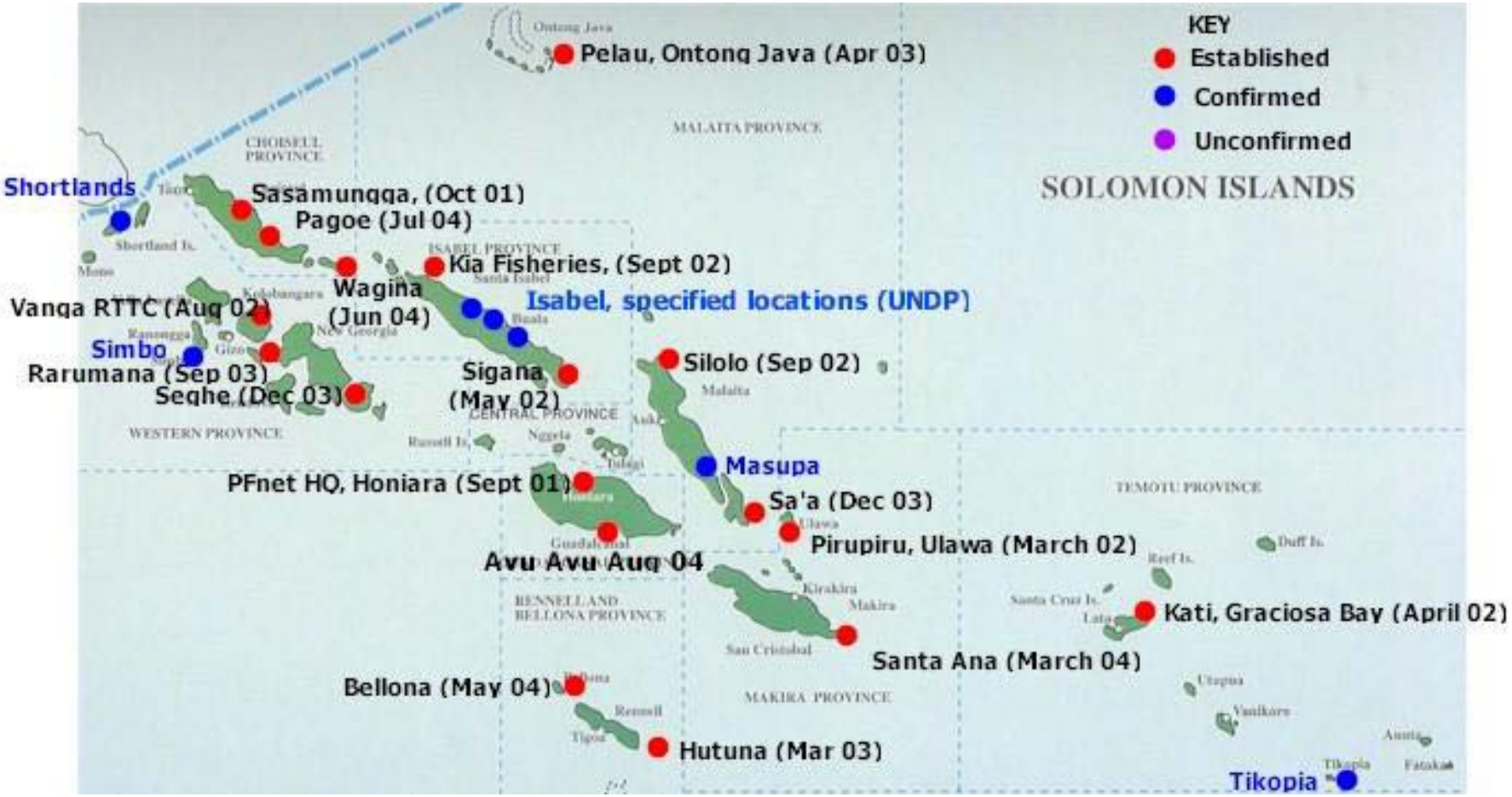
- TEK System
- Usage Scenarios
- Optimizing for Bandwidth

Deployment Status

- TEK available on SourceForge and via free CD
- Released summer 2002, but still expanding
  - Implementing new user interface
  - Partnering with Elsevier Scirus search engine for wide deployment to libraries, institutions
- Most active users in partner organizations

People's First Network

- Solomon Islands served by HF Radio Network
- Email only



Source: http://www.peoplefirst.net.sb/General/PFnet_Update.htm

People's First Network

- TEK installed: $0.65 per query from kiosk
  - $1.30 / hour for operator assistance browsing results
  - Compare to $0.25 per email, $0.65 to type one page
  - Contributes to kiosk sustainability
- Many applications reported
  1. Farmers – information on diseases; networking

> Subsistence farmers on Rennell have obtained advice concerning taro diseases affecting their crop. Via the **'TEK-websearch'** facility, one group of farmers was able to access detailed technical information about vanilla farming and to communicate with a specialist from the *Kastom Gaden Association.* *-- Chand et al., PFNet Case Study, 2005*

  2. Teachers – environmental impact of local logging
  3. Pastors – downloading sermons
  4. Entrepreneurs – download / sell lyrics
  5. General – health, education, sports, entertainment

First Mile Solutions

- Store-and-forward connectivity via Mobile Access Point
  - Cambodia, Rwanda, Costa Rica, India
- TEK provides only Web access

Source: www.firstmilesolutions.com

EmailWeb.us
(Gary Griswold)

- Same goal as TEK
- Operates entirely within email program
  - One URL request per query
  - HTML content + images returned in body of email
  - Links and forms re-submit to EmailWeb
- Very lightweight, no installation needed

EmailWeb Usage

- 2000 queries / day

Russia	28%
Cuba	17%
Indonesia	17%
Ukraine	8%
United States	6%
Canada	3%

Others:
Mongolia, Singapore, UK, Hong Kong, Papau New Guinnea, Sri Lanka, Libya, Malawi, Niger, Zambia, Costa Rica, The Bahamas, China, India, Japan, Solomon Islands, Belarus, Moldova, Kazakhstan, Switzerland

- In office: personal access to email, but not Web
  - One dedicated computer for Internet (Indonesia)
  - UUCP for local network (Cuba)

Applications in Developed Countries

- Airplanes
  - Tenzing supplies email-only connection ($10-$20)
    - Continental Airlines, United Airlines, US Airways
  - ~2.4kbs satellite link for entire plane[1]
- Mobile phones
- ISPs charge for bandwidth (Australia)
- Conservative religious sects[2]
- Anxiety about browser security[2]

[1] *http://www.pcworld.com/news/article/0,aid,114216,00.asp*
[2] *anecdotes from EmailWeb users*

Outline

- TEK System
- Usage Scenarios
- Optimizing for Bandwidth

Low-Bandwidth Search is Different

	Real-time Search	Email-based Search
Acceptable Latency	1-2 seconds	minutes/hours
Optimization Metric	relevance per page	relevance per byte
Search Process	trial-and-error	careful
User Identity	unknown	email address

1. State-Based Compression

- Cheaper to store information than re-download it
  - 100 GB disk drive: $250
  - 100 GB at 56kbs, $1/hr: $4000
- If server knows everything stored on client, can it improve compression of search results?

http://www.aidsinfo.nih.gov/

TEK Server

http://www.hivnetnordic.org/index.shtm
http://www.acnl.net/Basic_HIV_&_AIDS_Info.htm
http://www.info.gov.hk/health/aids/
http://www.cdc.gov/hiv/graphics/women.htm
http://www.med.unsw.edu.au/nchecr/
http://www.utopia-asia.com/aids.htm
http://www.sfaf.org/aids101/hiv_testing.html
http://www.gaytoronto.com/casey/

http://www.hivnetnordic.org/index.shtm
http://www.acnl.net/Basic_HIV_&_AIDS_Info.htm
http://www.info.gov.hk/health/aids/
http://www.cdc.gov/hiv/graphics/women.htm
http://www.med.unsw.edu.au/nchecr/
http://www.utopia-asia.com/aids.htm
http://www.sfaf.org/aids101/hiv_testing.html
http://www.gaytoronto.com/casey/

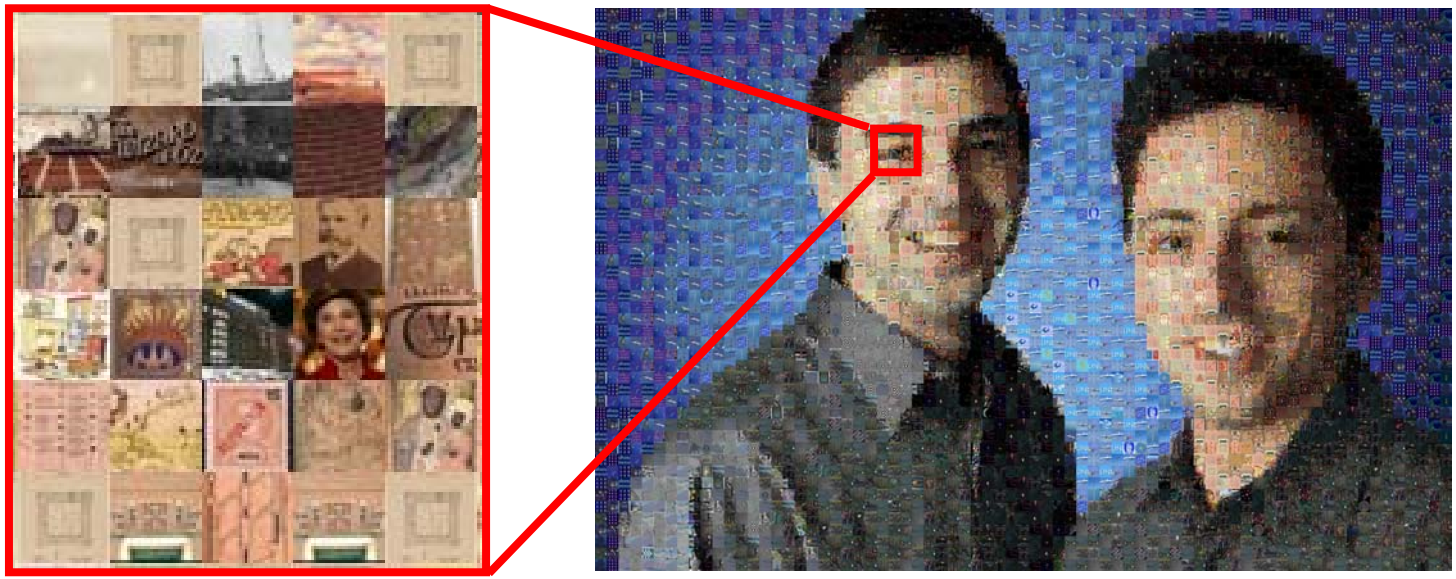
1. State-Based Compression

- General problem:
  If two parties share a large dictionary, can they reduce communication bandwidth?

- In general: no
  - info content (index) = info content (entry)
- In practice: maybe
  - Space of inputs is not uniformly populated
    - E.g., many images are text, bullets, smileys, patterns
  - Lossy: send index of closest match in dictionary
  - Lossless: send exact diff from dictionary entry

Photo Mosaics

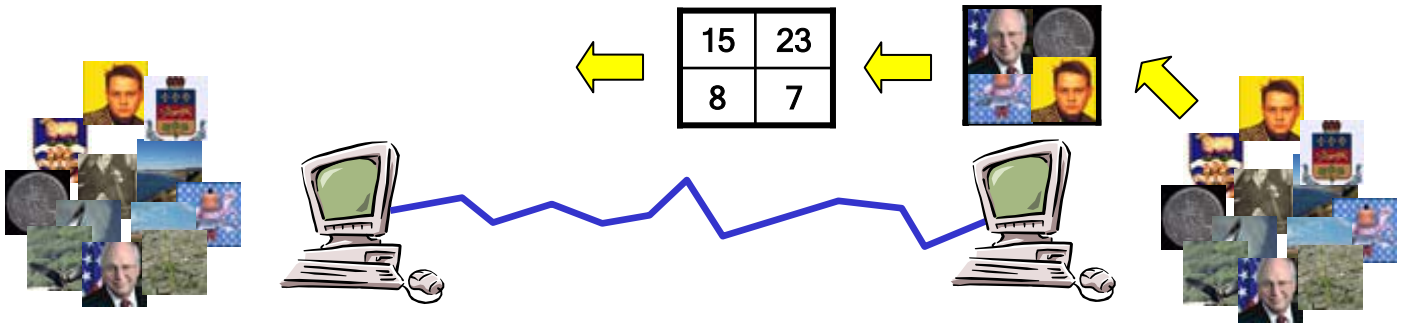
- Mosaic: picture made of other pictures



1. Break image into cells
2. Match each cell against image library
   - Use wavelet decomposition for perceptual match

Mosaic Compression
(Samidh Chakrabarti 2002)

- Idea: server constructs mosaic from client images
  - Send pointers to image components, not image data

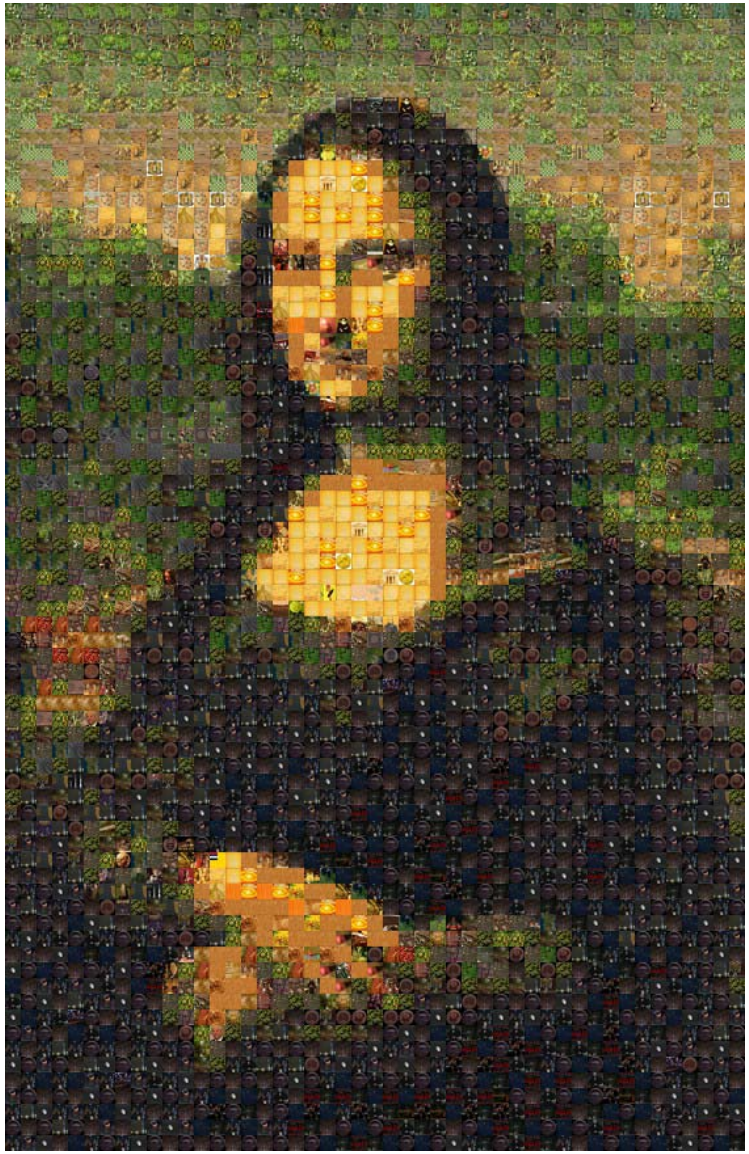


- Image size (bits): #cells * $\log_2$ (library_size)
  - Gzip offers further savings
- Possible image libraries
  - Images previously downloaded by client
  - Pre-defined library

Experiments

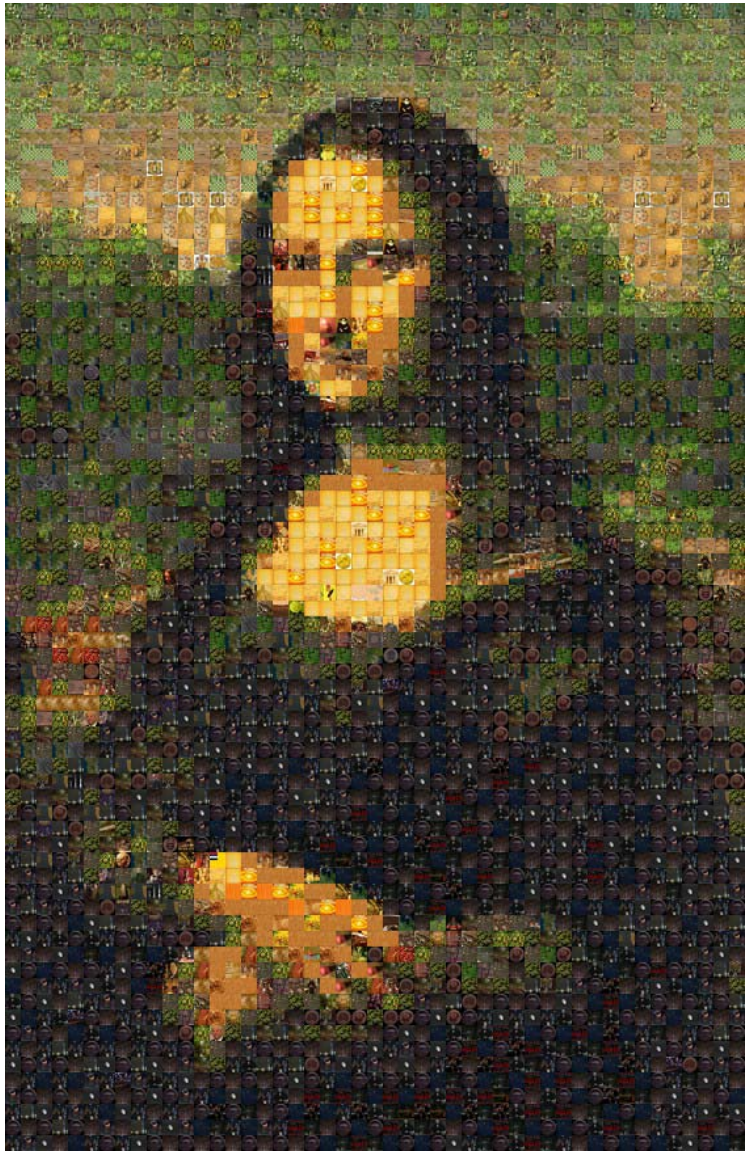
- Setup
  - 4096 images from Wikipedia
  - Cell size: 12x12 pixels
  - PhotoMosaic software (BlackDog, shareware)
    - Touch-up features disabled
- Processing time
  - ~20 minutes to analyze library
  - ~1 minute to build mosaic

Wikipedia JPEG: 46 Kb



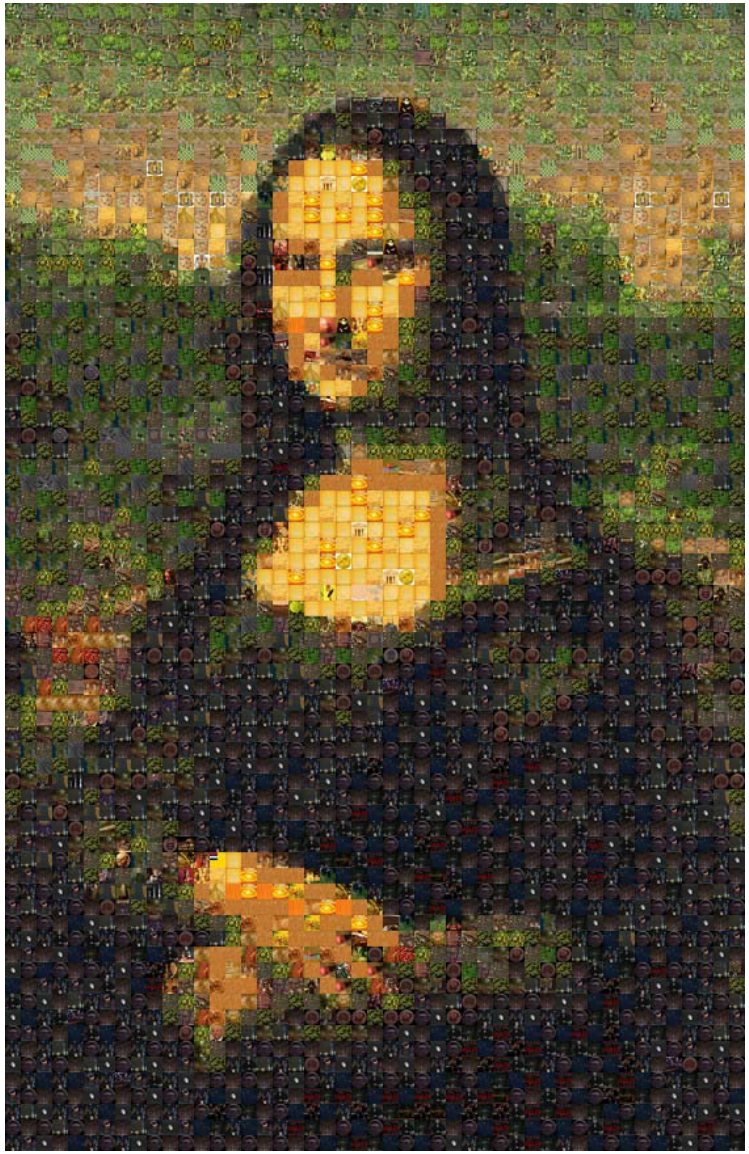
Mosaic: 2.0 Kb
(22X smaller)

0-Quality JPEG: 27 Kb



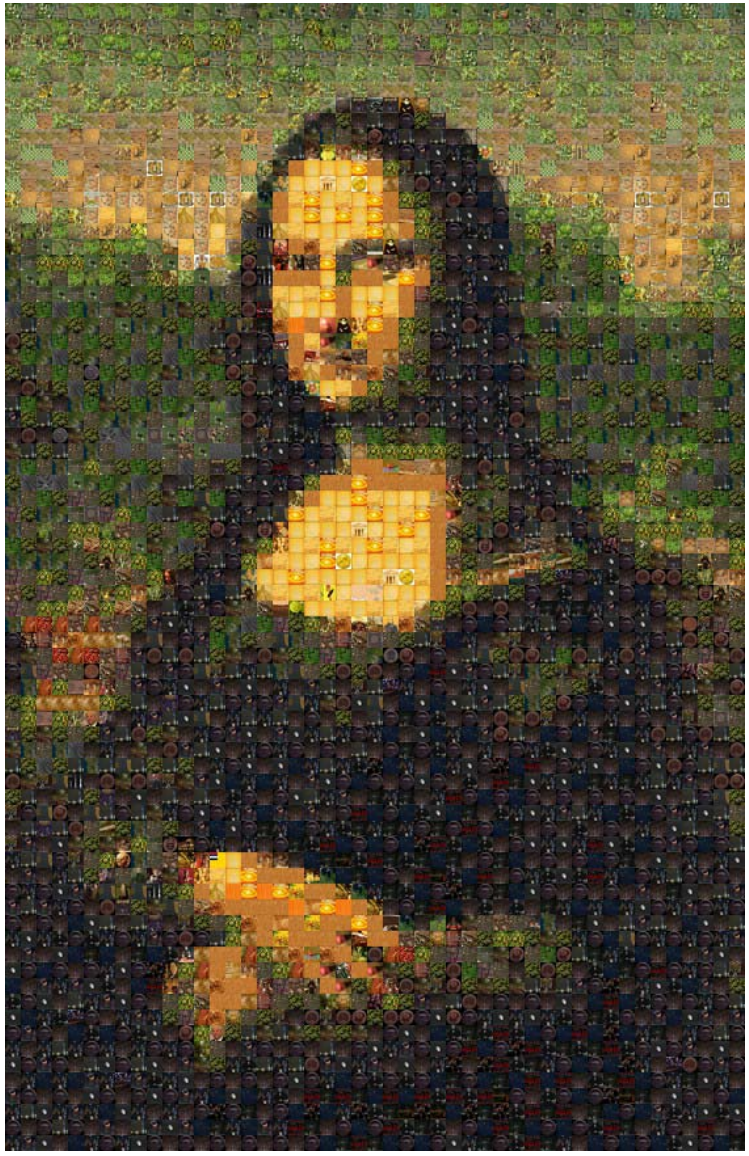
Mosaic: 2.0 Kb
(13X smaller)

30X Smaller JPEG:
2.0 Kb



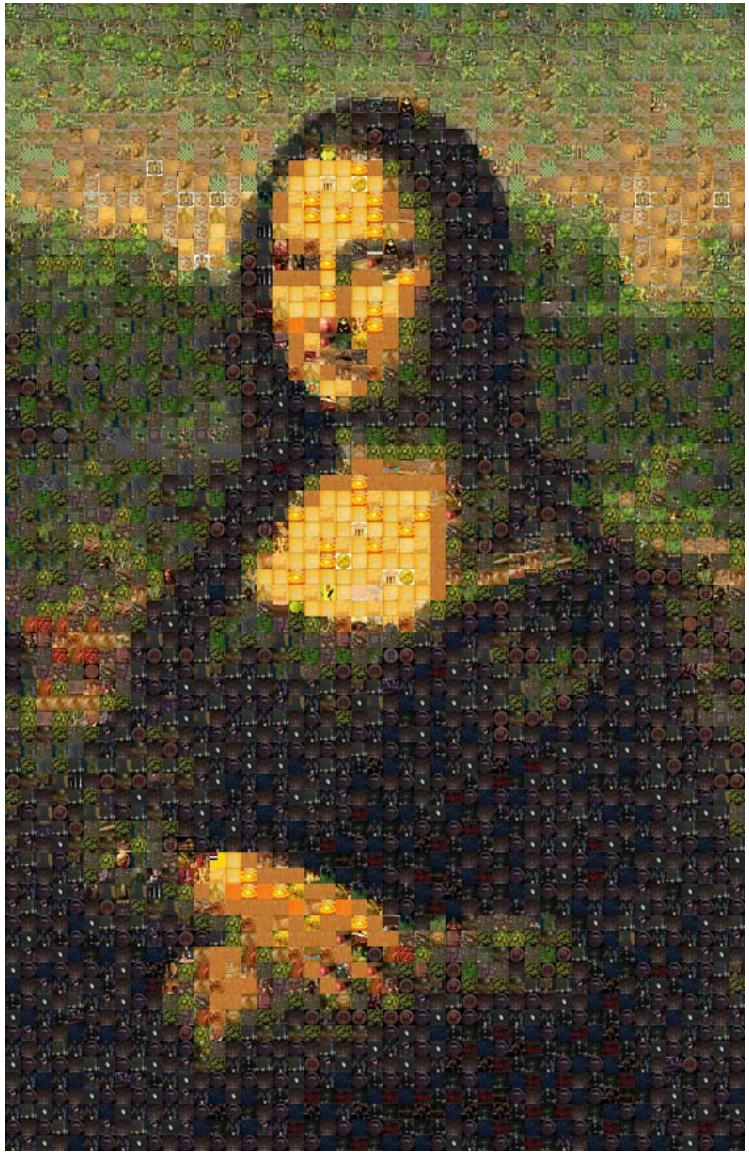
Mosaic: 2.0 Kb

Small JPEG, Zoomed: 2.0 Kb



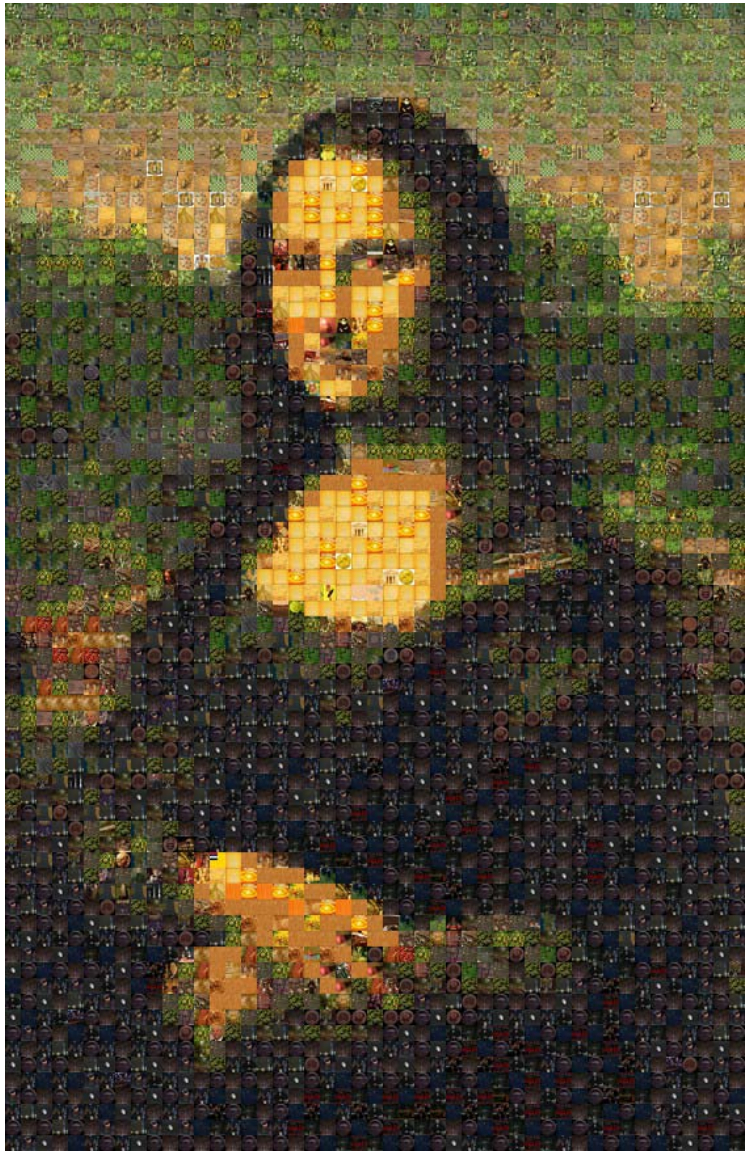
Mosaic: 2.0 Kb

21X Smaller GIF:
2.0 Kb



Mosaic: 2.0 Kb

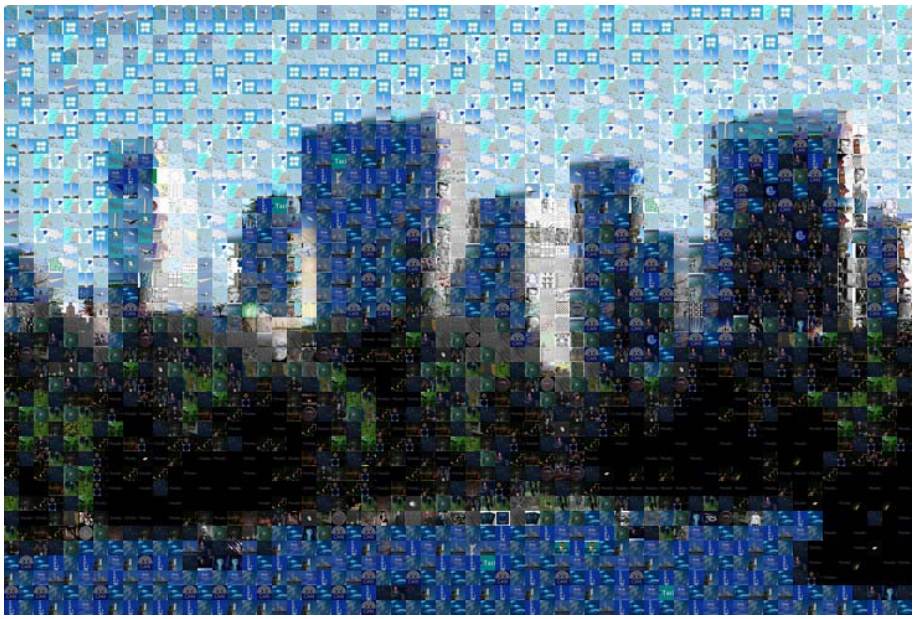
Small GIF, Zoomed: 2.0 Kb



Mosaic: 2.0 Kb

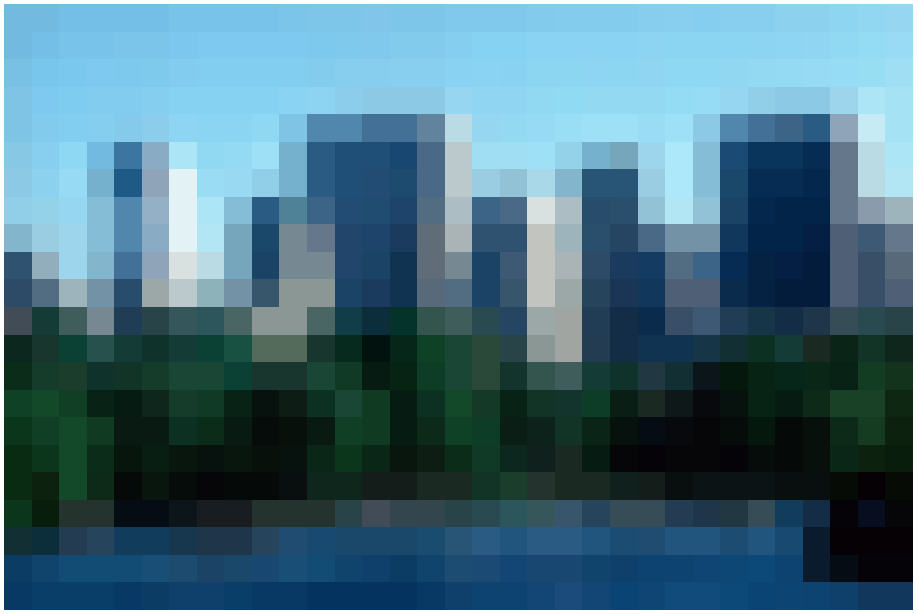
Compressing Landscapes

JPEG Image: 52 Kb

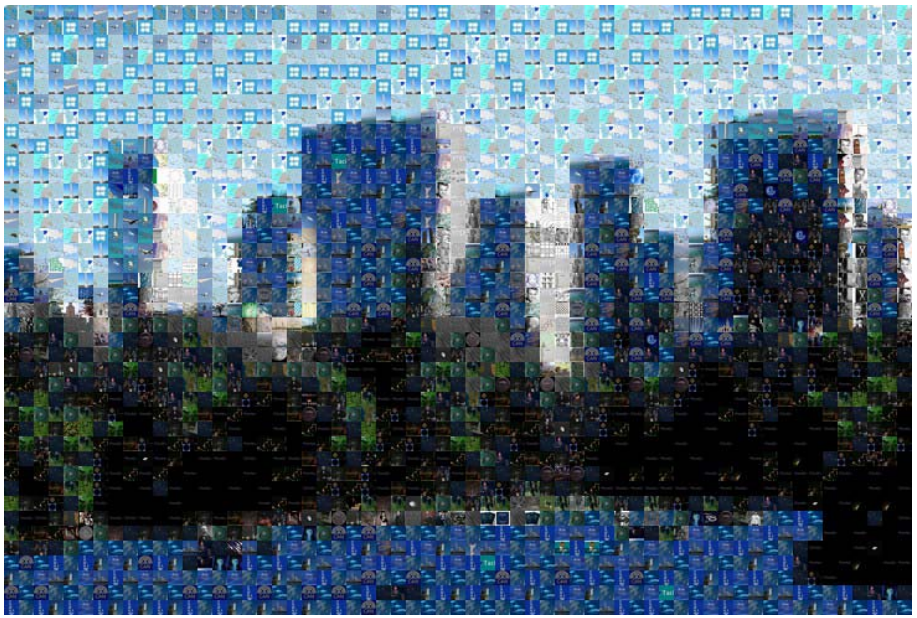


Mosaic: 1.6 Kb
(33X smaller)

Compressing Landscapes



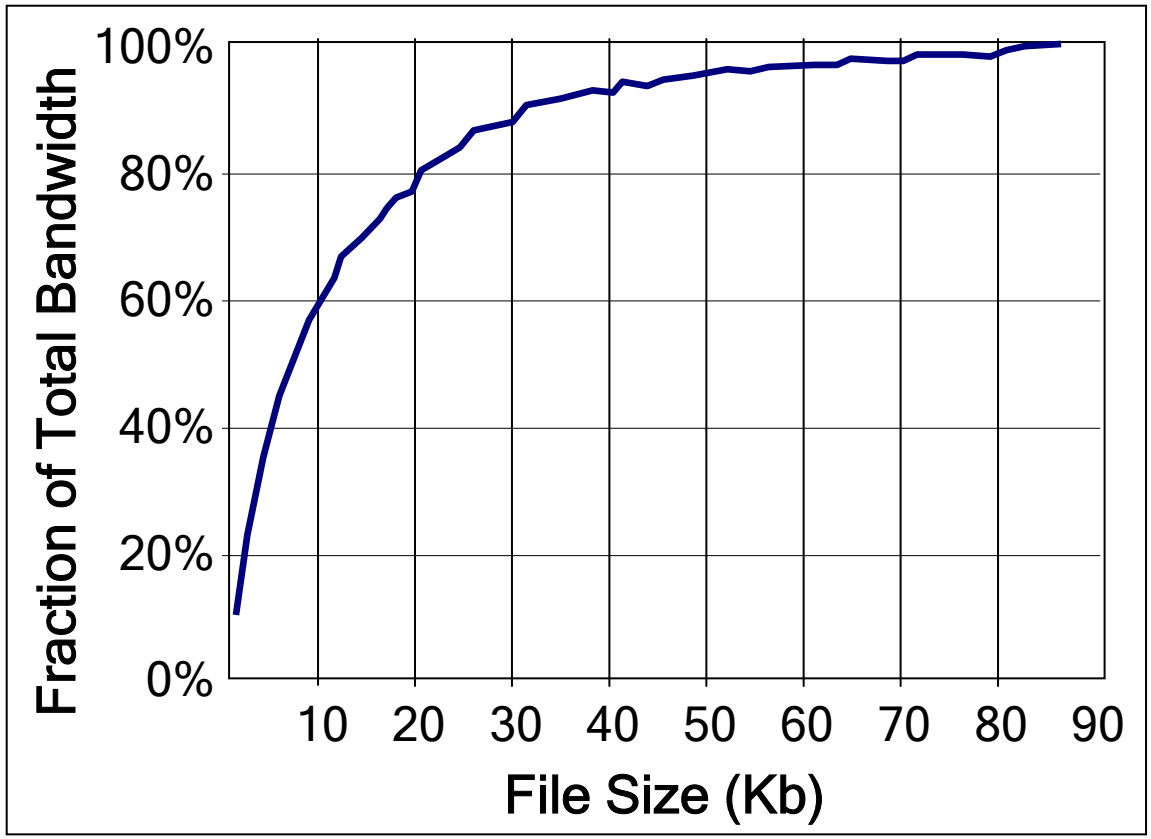
23X Smaller GIF: 1.6 Kb



Mosaic: 1.6 Kb

Importance of Small Images

- Most bandwidth spent on small images!



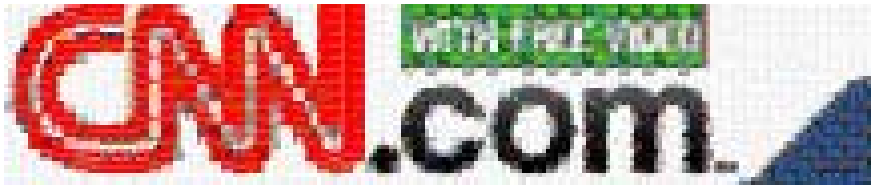
60% of bandwidth on images < 10Kb

→ Text recognition?
→ Icon substitution?

- Source: Chakrabarti'02
- 42,684 images from sites in Google programming contest
- 5,540 images from 1,000 most popular sites (ZDNet)

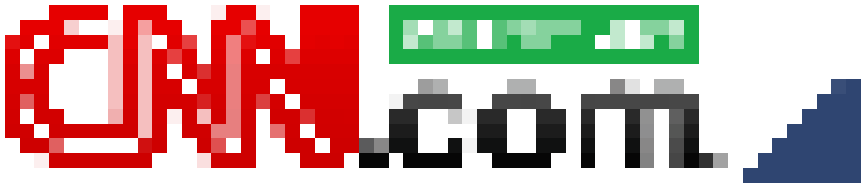
Compressing Logos

Original GIF: 4 Kb

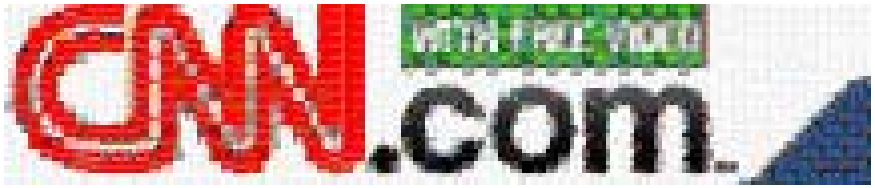


Mosaic: 0.8 Kb
(5X Smaller)

Compressing Logos



3.7X Smaller GIF: 0.8Kb



Mosaic: 0.8 Kb
(5X Smaller)

What's the Verdict?

- Many avenues for improvement
  - What is the best image library?
  - Impact of smoothing, rotation, diffs?
  - Edge detection + texture mapping
    - Lossy compression of edges
    - Random noise for realism

- In current form, perhaps useful as a preview
  - 5-33X smaller than JPEG
  - More entertaining than ALT tag or blurry picture

2. Breaking the URL Abstraction

- Entire webpage is unlikely to be useful
- Alternate abstractions for search engines:
  - Document sections (<a name= … >)
  - Paragraphs
  - Tables
  - PDF Bookmarks
- If low bandwidth,
  Extract relevant content and return to user
- If high bandwidth,
  Jump* to relevant portion

* may require cached version or HTML / browser extensions

3. Client-Specific Pagerank
(ala Google Personalized)

- Ambiguous searches have clusters of results
  - "Mercury" – element, planet, car, or Roman God?
  - High-bandwidth users do iterative searches
  - Low-bandwidth users can't afford many iterations
    - And often lack skills to eliminate spurious hits
- Idea: select pages based on client profile
  - Geography, demographics, previous searches
  - "Java history" from Indonesia → history of island
  - "GDP" after biology queries → guanine diphosphate
- Pagerank: boost links from user's demographic

4. Smart Query Builder

- Spelling error is costly for email-based search
- Client interface should:
  - Check spelling
  - Anticipate number of results
  - Identify ambiguous queries
- New opportunity for advanced query building
  - E.g., users willing to categorize searches
- New opportunity for evaluating search results
  - Users willing to provide careful feedback
  - Research vehicle for IR and UI testing

Conclusion

- High demand for low-bandwidth search
  - Today: emerging Internet users worldwide, PDAs
  - Future: pervasive computing, space exploration
- Much room for technical innovation
  - State-based compression
  - New ranking algorithms
- Prototype systems have proven useful
  - TEK, EmailWeb, www4mail, loband
  - Robust, visible service could have large impact

Acknowledgements

- **TEK team**
  - Prof. Saman Amarasinghe - Marjorie Cheng
- **Previous participants**
  - Libby Levison
  - Samidh Chakrabarti
  - Tazeen Mahtab
  - Genevieve Cuevas
  - Saad Shakhshir
  - Janelle Prevost
  - Mark Halsey
  - Hongfei Tian
  - Damon Berry
  - Bihn Vo
  - Sheldon Chan
  - Sid Henderson
  - Alexandro Artola
- **EmailWeb -** Gary Griswold