20 years of Web search – where to next?

Mark Sanderson

Who am I?

- Professor at RMIT University, Melbourne
- Before
  - Professor at University of Sheffield
  - Researcher at UMass Amherst
  - Researcher at University of Glasgow
- Online
  - @IR_oldie
  - http://www.seg.rmit.edu.au/mark/

Overview of talk

- A bit of history

A bit of history

Early IR

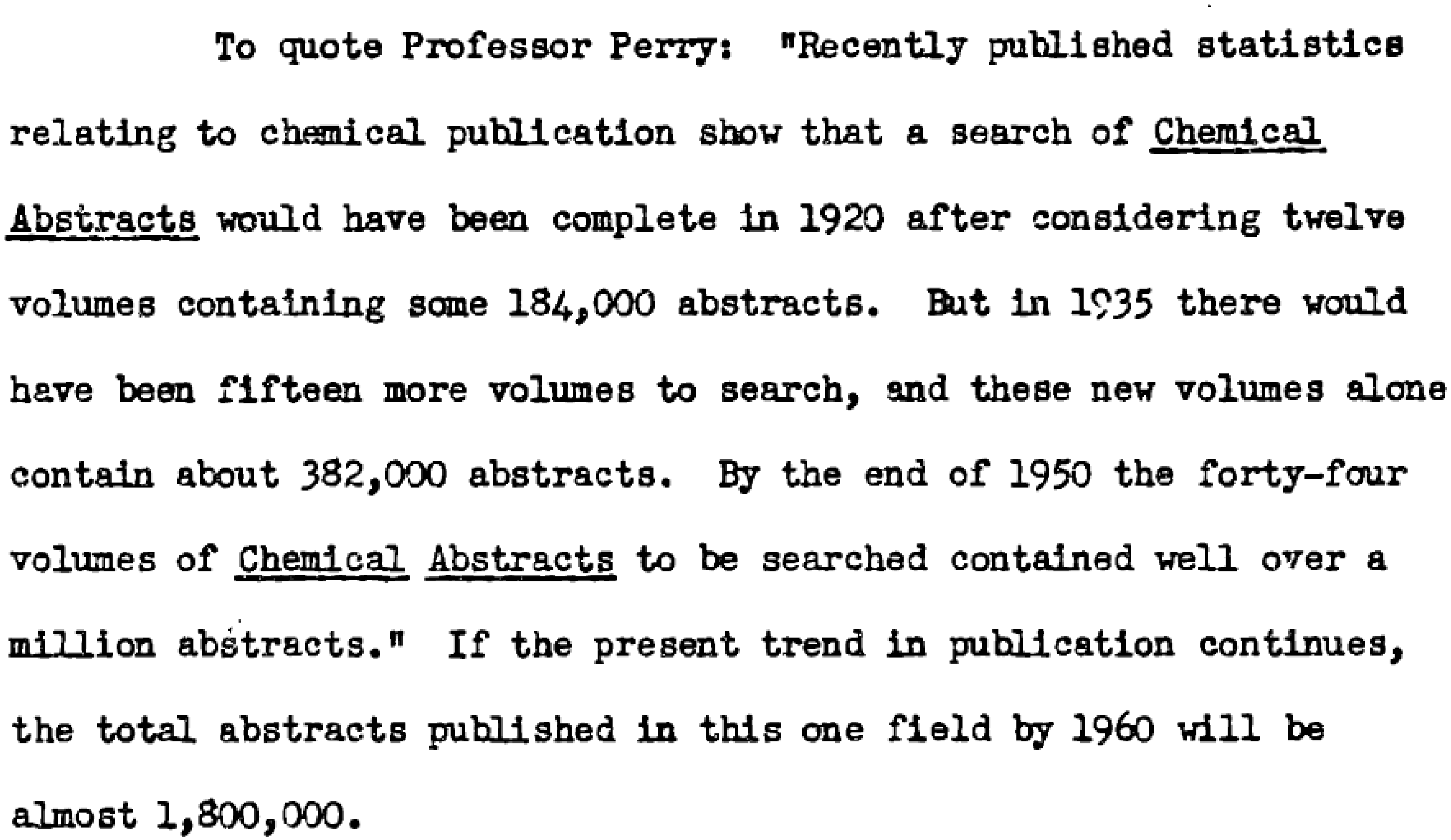
Before IR systems

- There were libraries
  - The search engine of the day

- Organise information using a subject catalogue
  - Sort cards by author
  - Sort cards by title
  - Sort cards by subject
    - How to do this?

Not just public libraries

- MIT Masters thesis, Philip Bagley, 1951



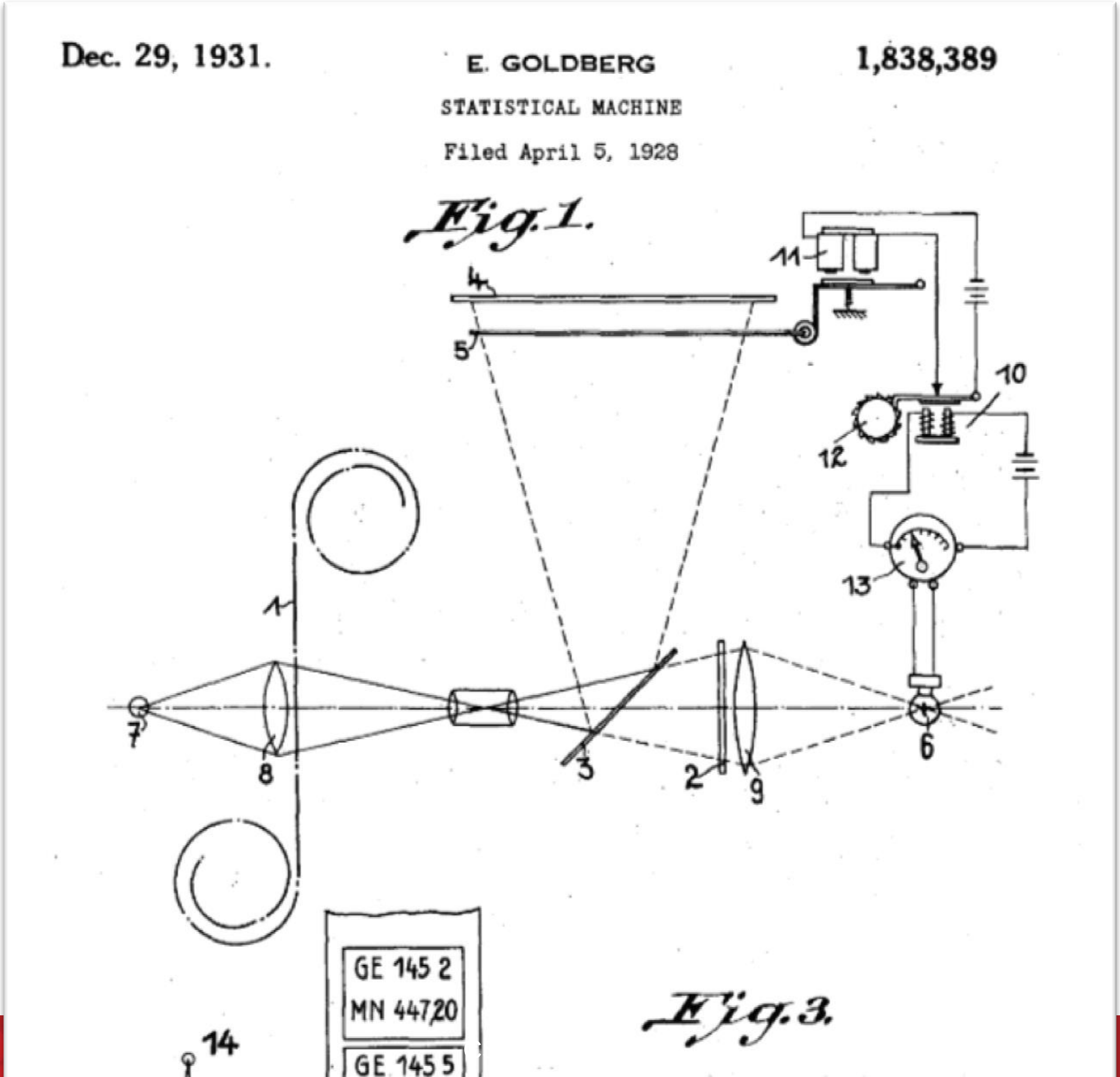
To quote Professor Perry: "Recently published statistics relating to chemical publication show that a search of Chemical Abstracts would have been complete in 1920 after considering twelve volumes containing some 184,000 abstracts. But in 1935 there would have been fifteen more volumes to search, and these new volumes alone contain about 382,000 abstracts. By the end of 1950 the forty-four volumes of Chemical Abstracts to be searched contained well over a million abstracts." If the present trend in publication continues, the total abstracts published in this one field by 1960 will be almost 1,800,000.

At the same time…

- While librarians were coping with the information explosion
  - Could machines help?
  - Could computers help?

- Very brief history of machines and computers for search

Machines doing IR



Dec. 29, 1931.

E. GOLDBERG

1,838,389

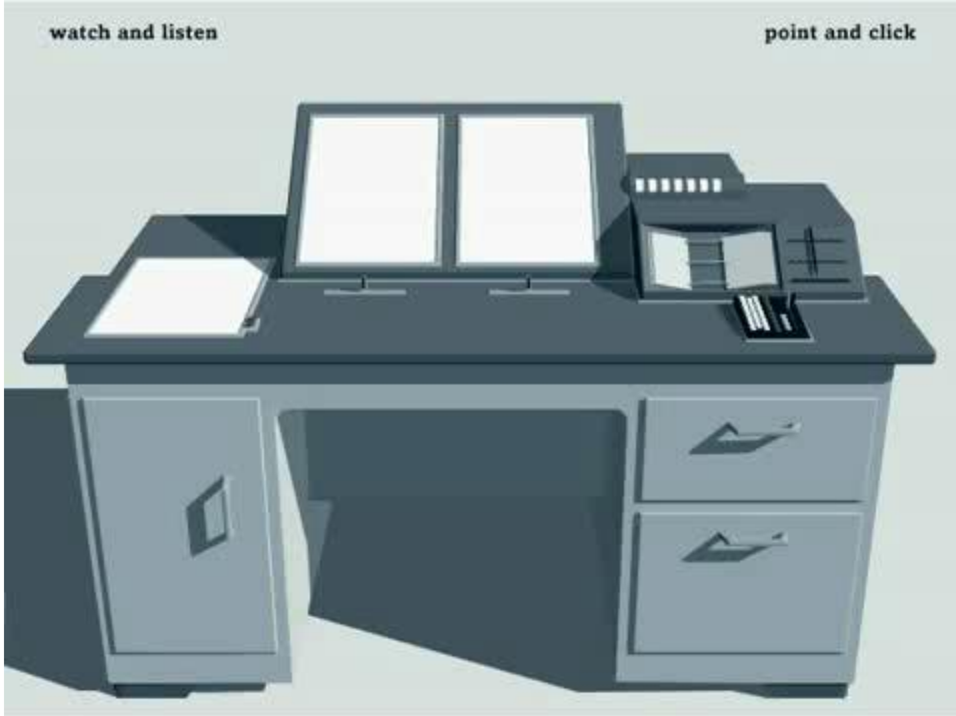
STATISTICAL MACHINE

Filed April 5, 1928

Fig. 1.

Fig. 3.

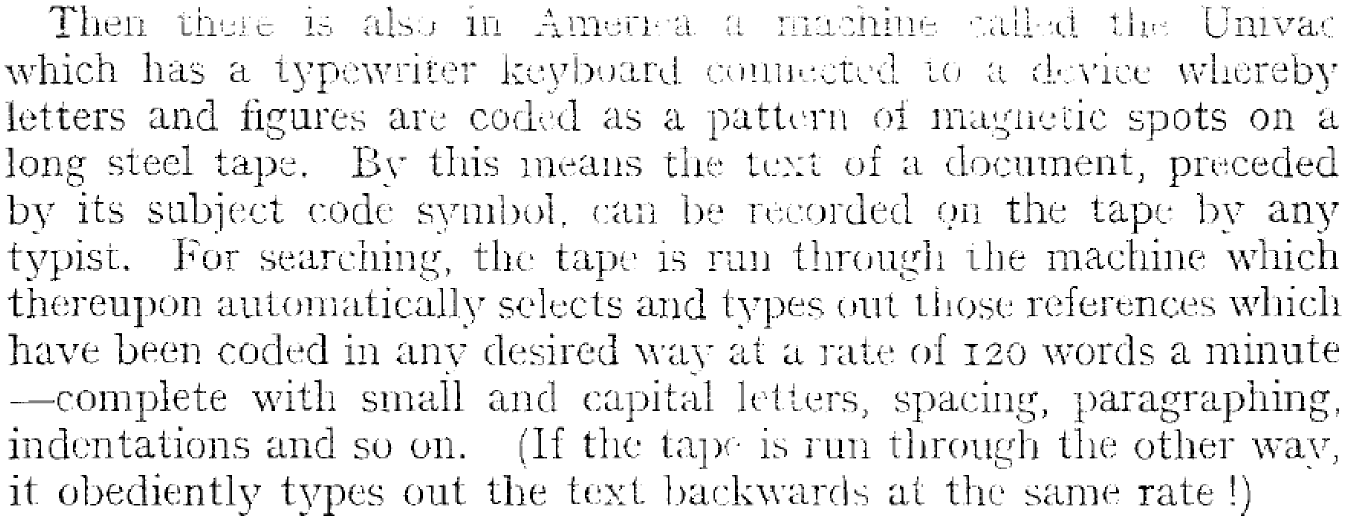
As we may think – Bush 1945



– http://www.youtube.com/watch?v=c539cK58ees

Computers doing IR

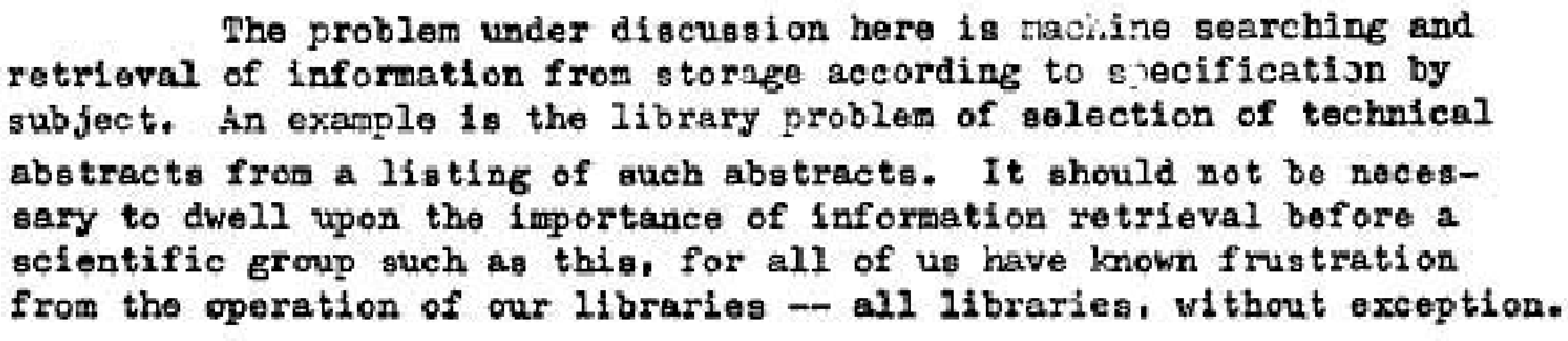
- Holmstrom 1948



Then there is also in America a machine called the Univac which has a typewriter keyboard connected to a device whereby letters and figures are coded as a pattern of magnetic spots on a long steel tape. By this means the text of a document, preceded by its subject code symbol, can be recorded on the tape by any typist. For searching, the tape is run through the machine which thereupon automatically selects and types out those references which have been coded in any desired way at a rate of 120 words a minute —complete with small and capital letters, spacing, paragraphing, indentations and so on. (If the tape is run through the other way, it obediently types out the text backwards at the same rate !)

Information Retrieval

- Calvin Mooers, 1950



The problem under discussion here is machine searching and retrieval of information from storage according to specification by subject. An example is the library problem of selection of technical abstracts from a listing of such abstracts. It should not be necessary to dwell upon the importance of information retrieval before a scientific group such as this, for all of us have known frustration from the operation of our libraries -- all libraries, without exception.

NRT

- See demo shown in talk at
  - http://www.seg.rmit.edu.au/mark/demos/NRT/NRT%20demo.htm
- Paper at
  - http://www.seg.rmit.edu.au/mark/cv/publications/papers/my_papers/EP-odd.pdf

The web arrived

- 1993
  - JumpStation
    - Jonathon Fletcher, University of Stirling

- Steinberg, Wired, 1996
  - “*Information retrieval is really only a problem for people in library science - if some computer scientists were to put their heads together, they'd probably have it solved before lunchtime.*”

Where are we now

Google/Bing

Where we are now

- Google/Bing
  - Text matching
    - Fields, anchor
    - PageRank
    - Query logs
    - …
  - Massive machine learning
    - Evaluation
    - Continual tuning

Search is solved?

- Common perception

Favourable conditions

- Most content wants to be found
- Most content is redundant
- Huge income
- Queries often repeated

- Users can read & write

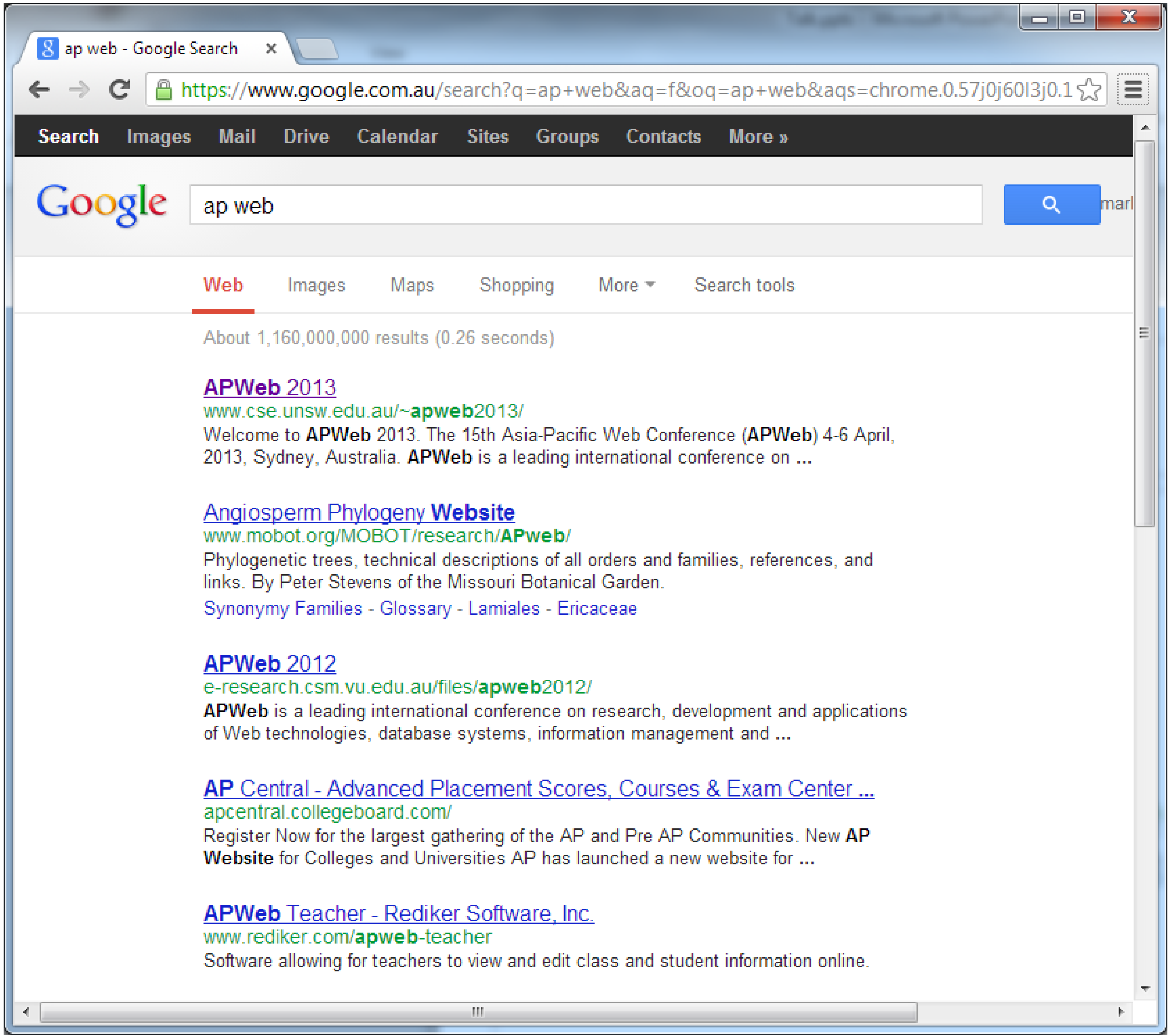
Where to next?

- Immediate problems
- Immediate opportunities
- Medium term challenges
- Longer term challenges

Immediate

Problems/opportunies

Problematic summaries



About 1,160,000,000 results (0.26 seconds)

APWeb 2013
www.cse.unsw.edu.au/~apweb2013/
Welcome to **APWeb** 2013. The 15th Asia-Pacific Web Conference (**APWeb**) 4-6 April, 2013, Sydney, Australia. **APWeb** is a leading international conference on ...

Angiosperm Phylogeny **Website**
www.mobot.org/MOBOT/research/**APweb**/
Phylogenetic trees, technical descriptions of all orders and families, references, and links. By Peter Stevens of the Missouri Botanical Garden.
Synonymy Families - Glossary - Lamiales - Ericaceae

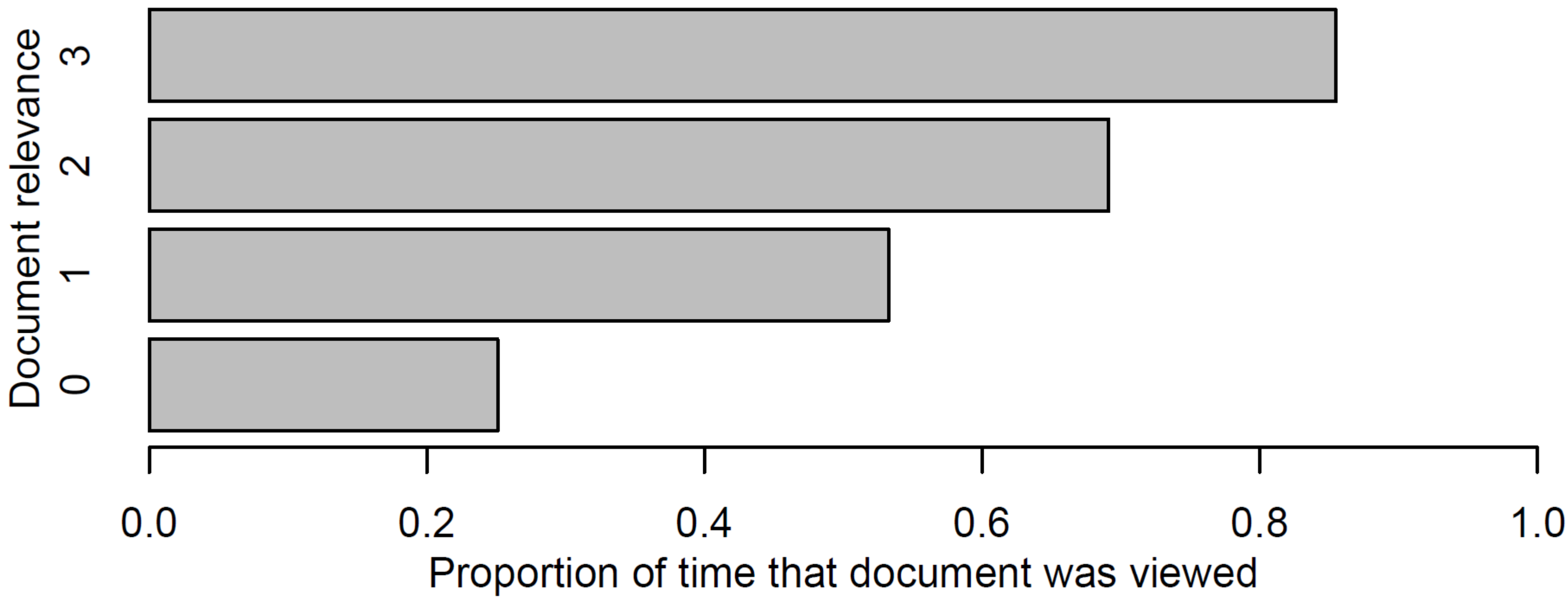
APWeb 2012
e-research.csm.vu.edu.au/files/**apweb**2012/
APWeb is a leading international conference on research, development and applications of Web technologies, database systems, information management and ...

AP Central - Advanced Placement Scores, Courses & Exam Center ...
apcentral.collegeboard.com/
Register Now for the largest gathering of the AP and Pre AP Communities. New **AP Website** for Colleges and Universities AP has launched a new website for ...

APWeb Teacher - Rediker Software, Inc.
www.rediker.com/**apweb**-teacher
Software allowing for teachers to view and edit class and student information online.

Less favourable?

- People struggle to search
- People miss retrieved documents
  - Fine for redundant content; what if just one?



Problem searching

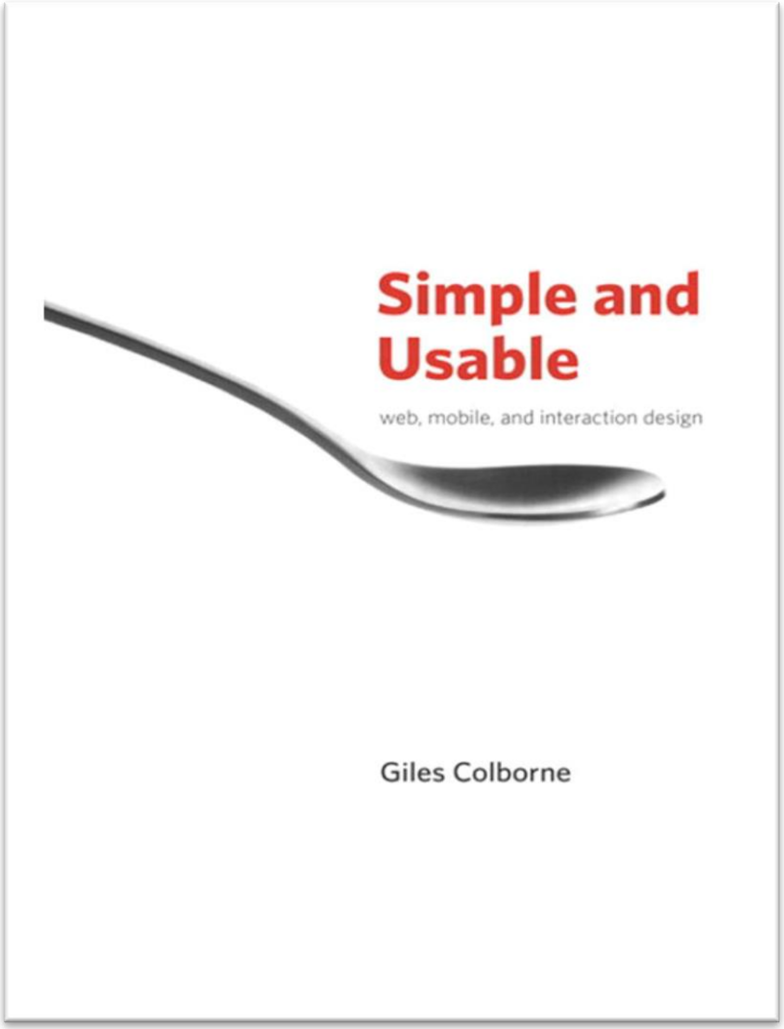
- Limited redundancy
  - Little money
    - Enterprise search
    - Refinding
  - Content doesn’t want to be found
    - Patent search
    - Legal document search (e-Discovery)

Enterprise search

- Many problems in this space
- Each collection is different
  - Each search engine needs to be different
- No money
- “Why doesn’t it work like Google?”

Significant problem

- Think carefully before including search in your user interface



Simple and Usable

web, mobile, and interaction design

Giles Colborne

At RMIT

- Trying to scope the problem
  - If we find a search solution that works on one set of documents, does it work on others?
  - Not as much as was thought
    - A lot worse than was thought

Major immediate challenge

- Do search as well as Google no matter what the collection, and do it without all their money

Favourable conditions

- Most content wants to be found
- Most content is redundant
- Huge income
- Queries often repeated
- Users can read & write

Refinding

- Interviewed 45 searchers about common retrieval tasks
  - 70% relate to refinding
- Starting funded investigation in this area.

YAHOO!®

Ephemeral & archival content

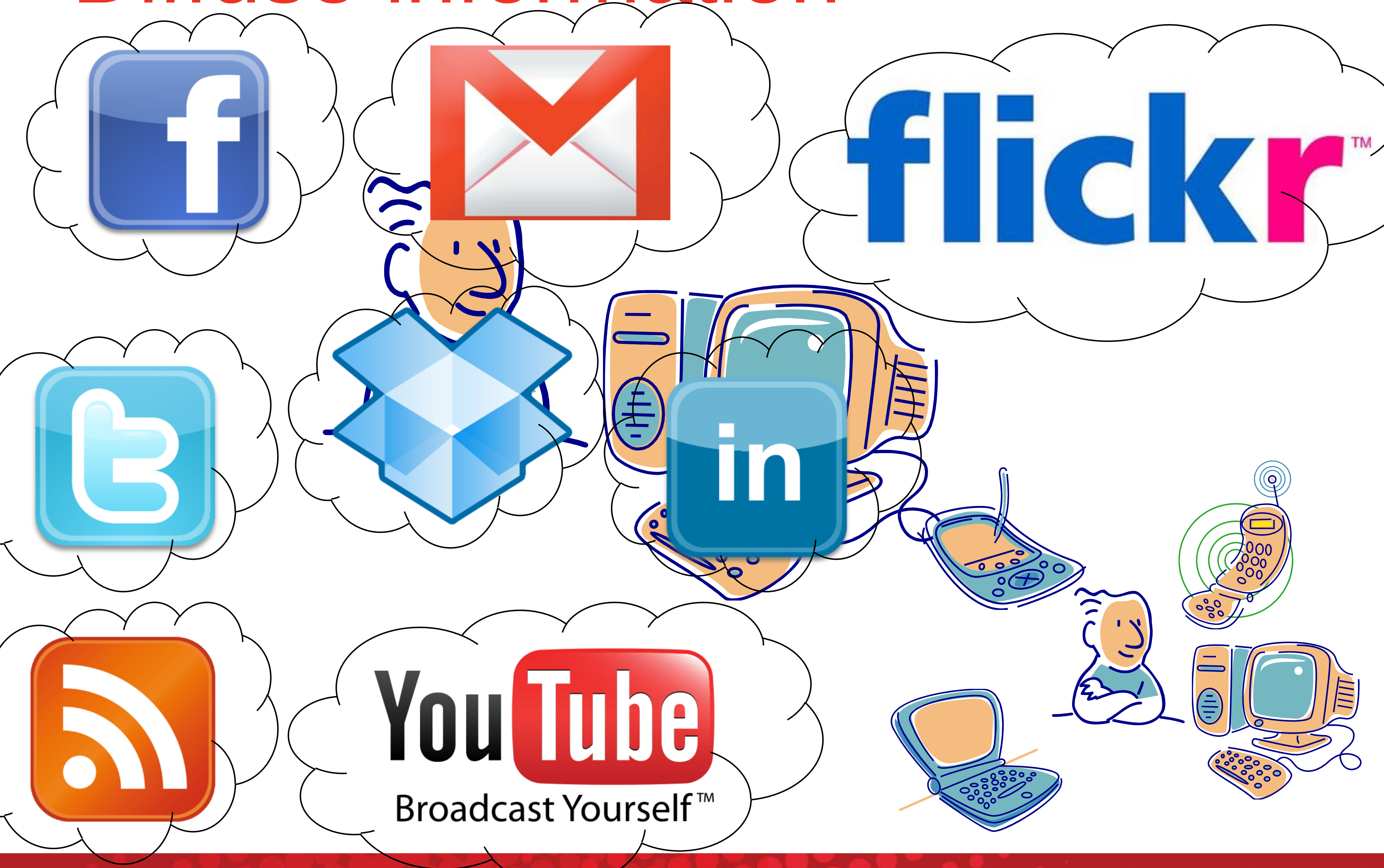
- Archival
  - Traditional web search
    - Web pages, news, documents
    - Coarse grained
- Ephemeral
  - Social media
    - Blogs, social networks, micro-blogs
    - Fine grained

Interface of the two

- Summarising ephemeral content
  - Only just starting
  - Lots of opportunities to specialise
- How can ephemeral content aid search of archival
  - RMIT changing representation of archival content based on ephemeral data.
    - Early days, but promising

Medium term

Diffuse information



Harder information needs

- Entertain me
- Contextual search
- SWIRL 2012
  - http://www.cs.rmit.edu.au/swirl12/

Longer term

Longer term

- Long queries
- Spoken search
- The internet for everyone

Users have complex needs

- Poorly expressed in short queries
  - Experts
    - issue multiple short queries
    - use search engine operators
- Can we build search engines to handle complex queries?

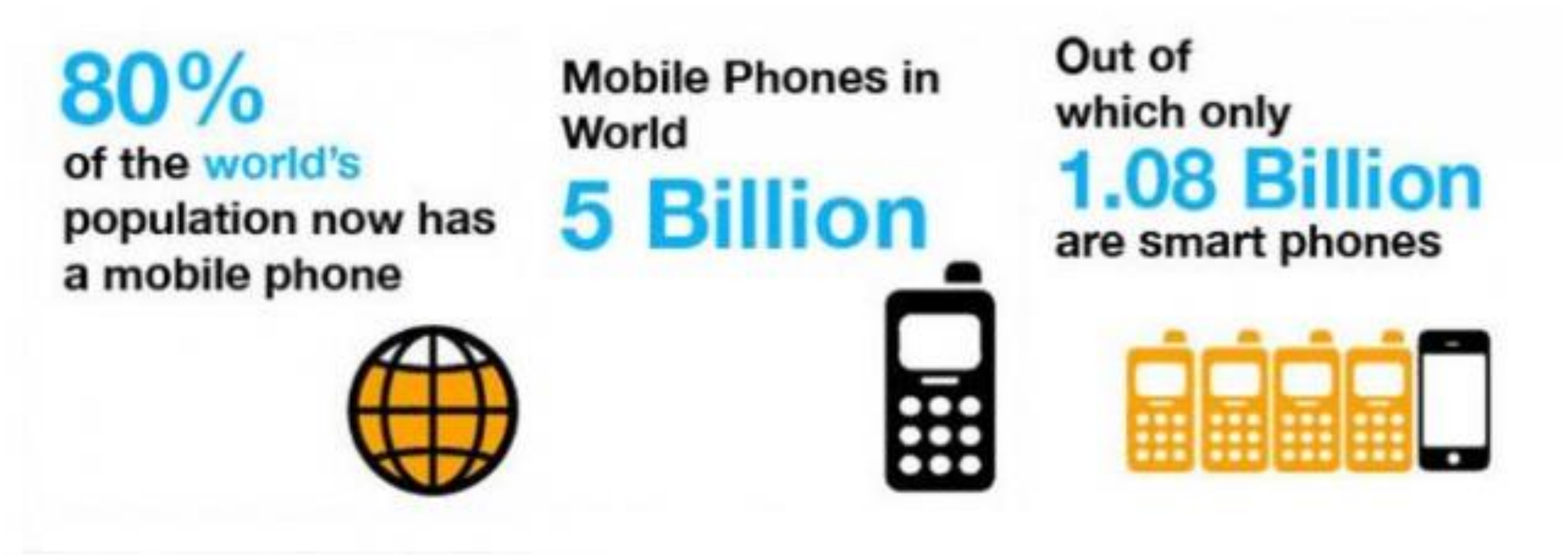
New application area?

- Speech search
  - Hand free
  - Eyes free
- Seen in the movies, but really?

Users?

- Visually impaired
  - –Together they could form a country
- Other potential uses
  - –In car searching
  - –Walking in a city

Internet for everyone



80% of the world's population now has a mobile phone

Mobile Phones in World 5 Billion

Out of which only 1.08 Billion are smart phones

– http://www.onbile.com/info/how-many-people-use-smartphones-in-the-world/

Internet users?

- 2013
  - 2 billion now
- 2015
  - 4 billion mostly on mobiles (Baird Equity Research)

Implications?

- More languages
- More users who struggle with literacy
  - Search engines assume you can read and write

Search engines

There is a lot still to do