Analytics for I-Schools

Helping libraries evaluate and improve services

Policy limitations: privacy, tailoring, openness

Are digital libraries competing with e-commerce?

Analytics: web traffic measurements

For example, Google Insights for Search

Below: searches for “snow” in Canada by date.

Best example: Google flu trends

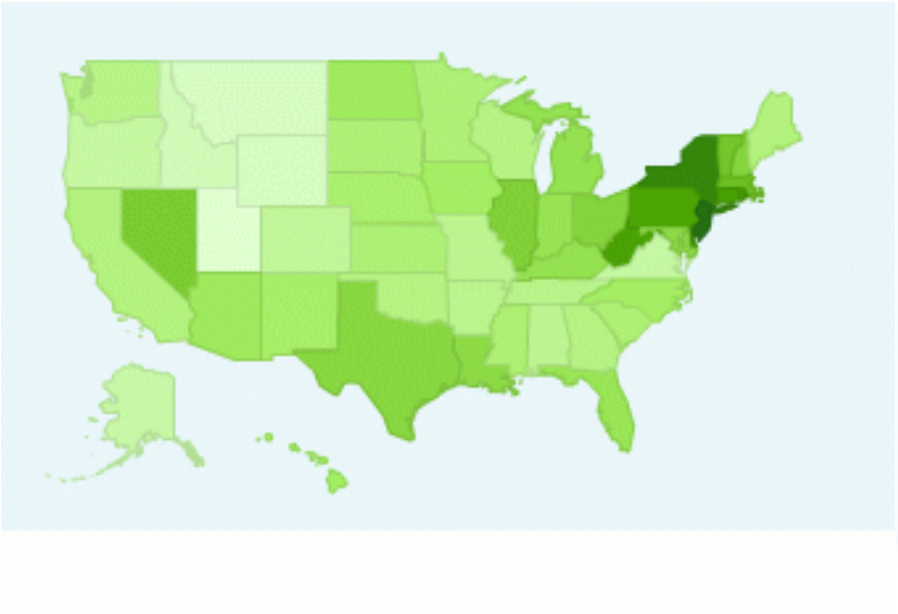
Google found a set of search terms that are so well correlated with instances of the flu that they could predict the spread of flu epidemics more rapidly than the CDC.

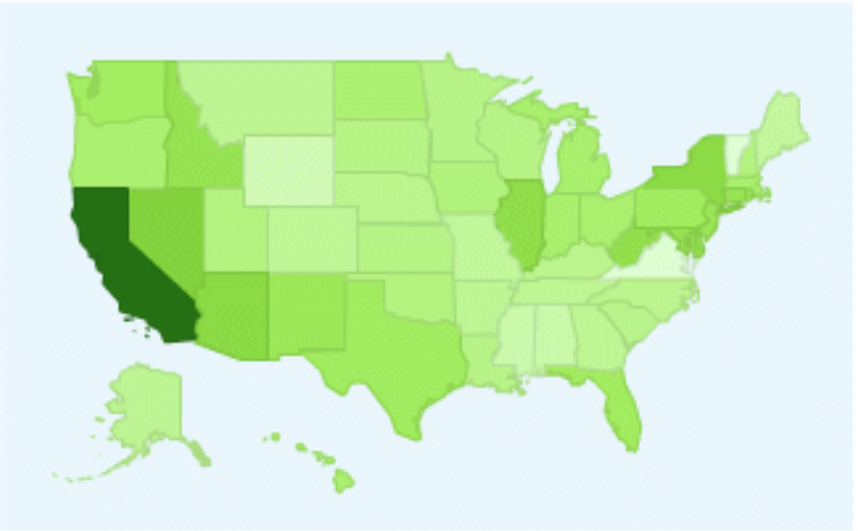
More trivial: what day of the week is the search term “hangover” most common? Answer: Sunday.

More serious: what search term correlates best with Toyota sales? Answer: “used hyundai”.

Geographical plotting

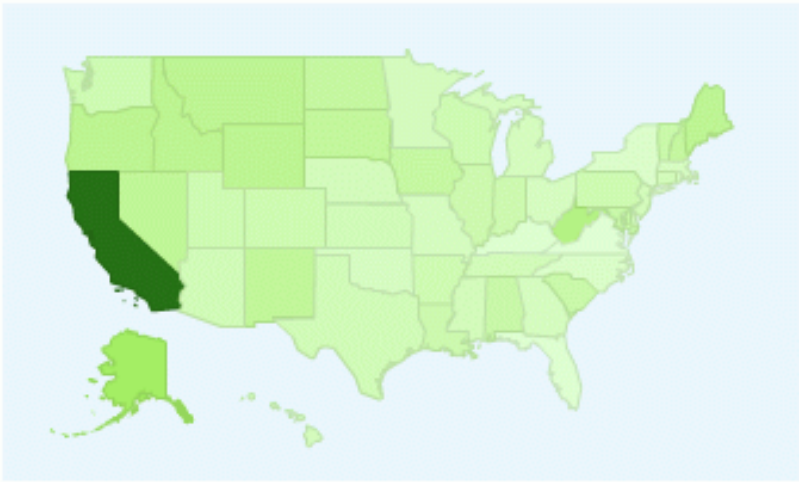
You can also see where people are interested in different subjects. Below left: *Snooki*. Below right: *Schwarzenegger*.



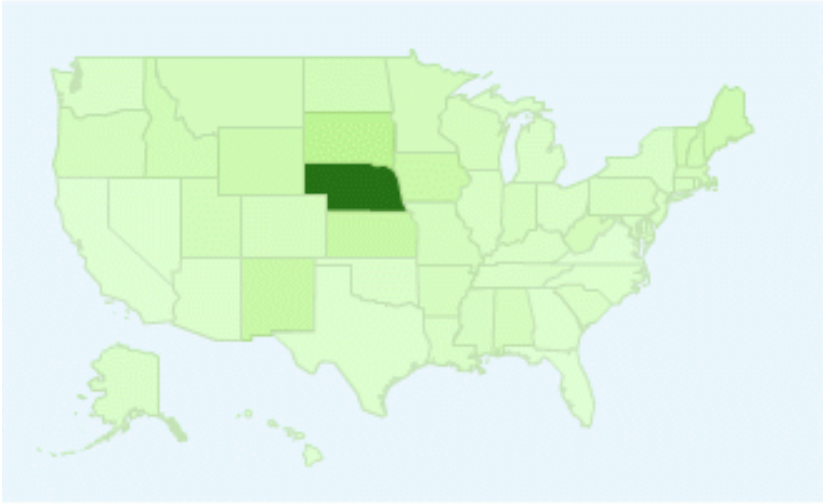


Which authors are searched where?

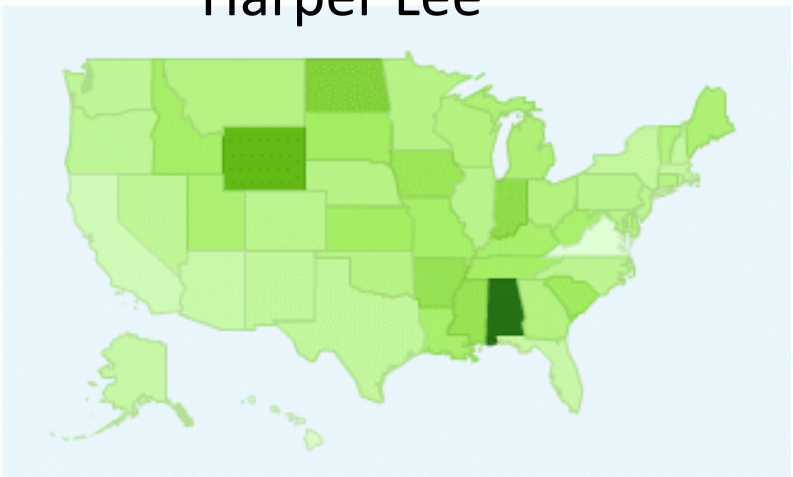
Jack London



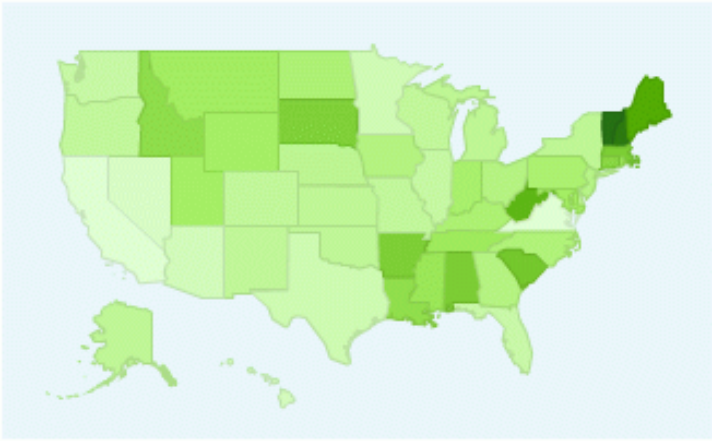
Willa Cather



Harper Lee



Robert Frost



A line in a logfile

174.50.89.165 - - [01/Jan/2012:17:52:56 -0500]

"GET /mlesk/ksg97/ksg.html HTTP/1.1" 200 19841

"http://www.google.com/search?hl=en&s..... "

"Mozilla/5.0 (iPhone...."

User's ISP, date

Which file they read, status, how many bytes

Referring site

What browser they used

You do have to remove the "bots" – some are polite and announce themselves, others are found by behavior. There's plenty of software out there to interpret logfiles.

What can logfiles tell you?

Where (geographically) do your users come from?

What resources are they using?

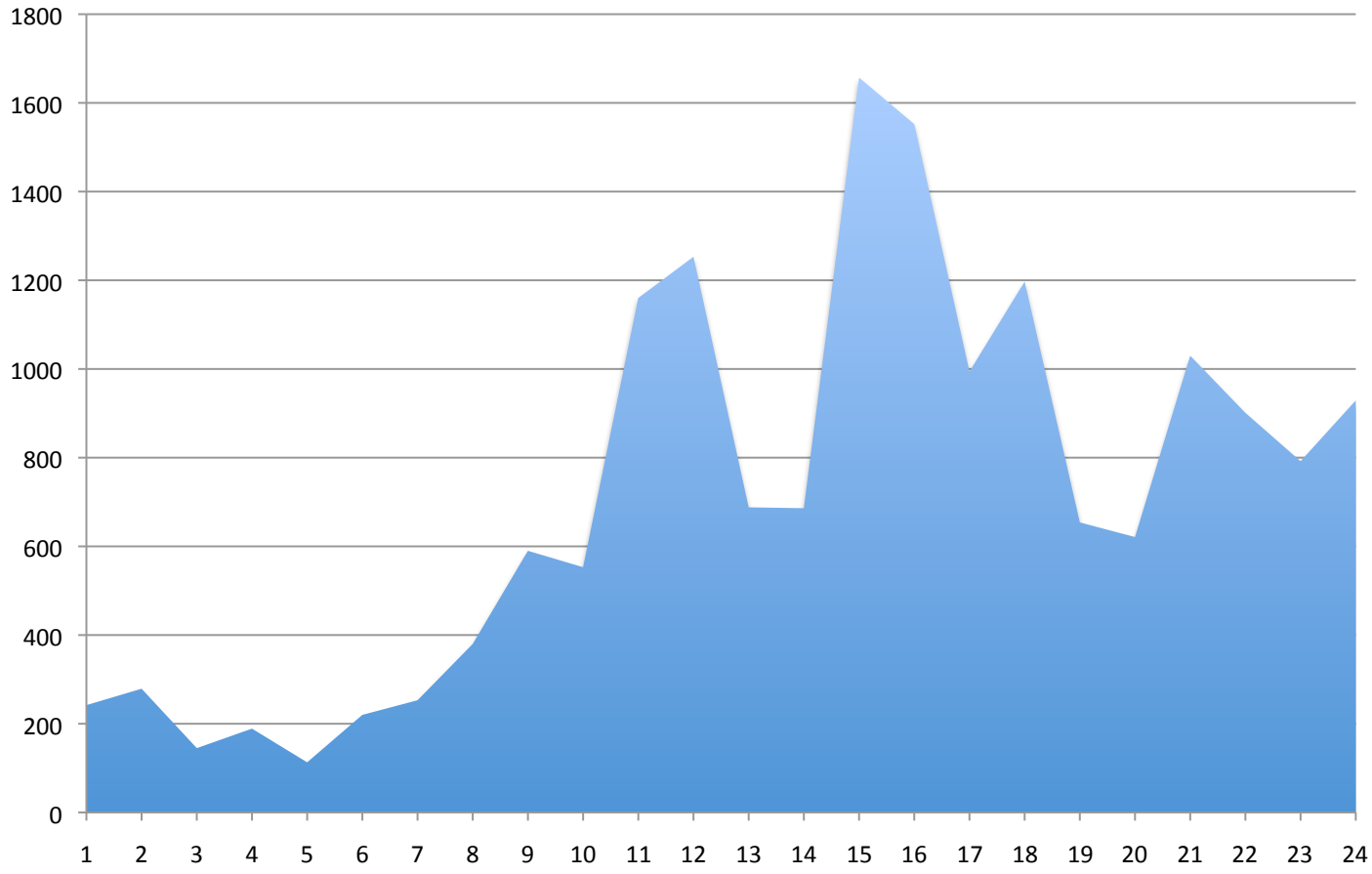
When do they access your site?

What web page were they on before your site?

What web page do they go to next? (not logfile)

How long do they spend on each page?

Usage by time of day



Is it more important to be paying attention at 9am or 9pm?

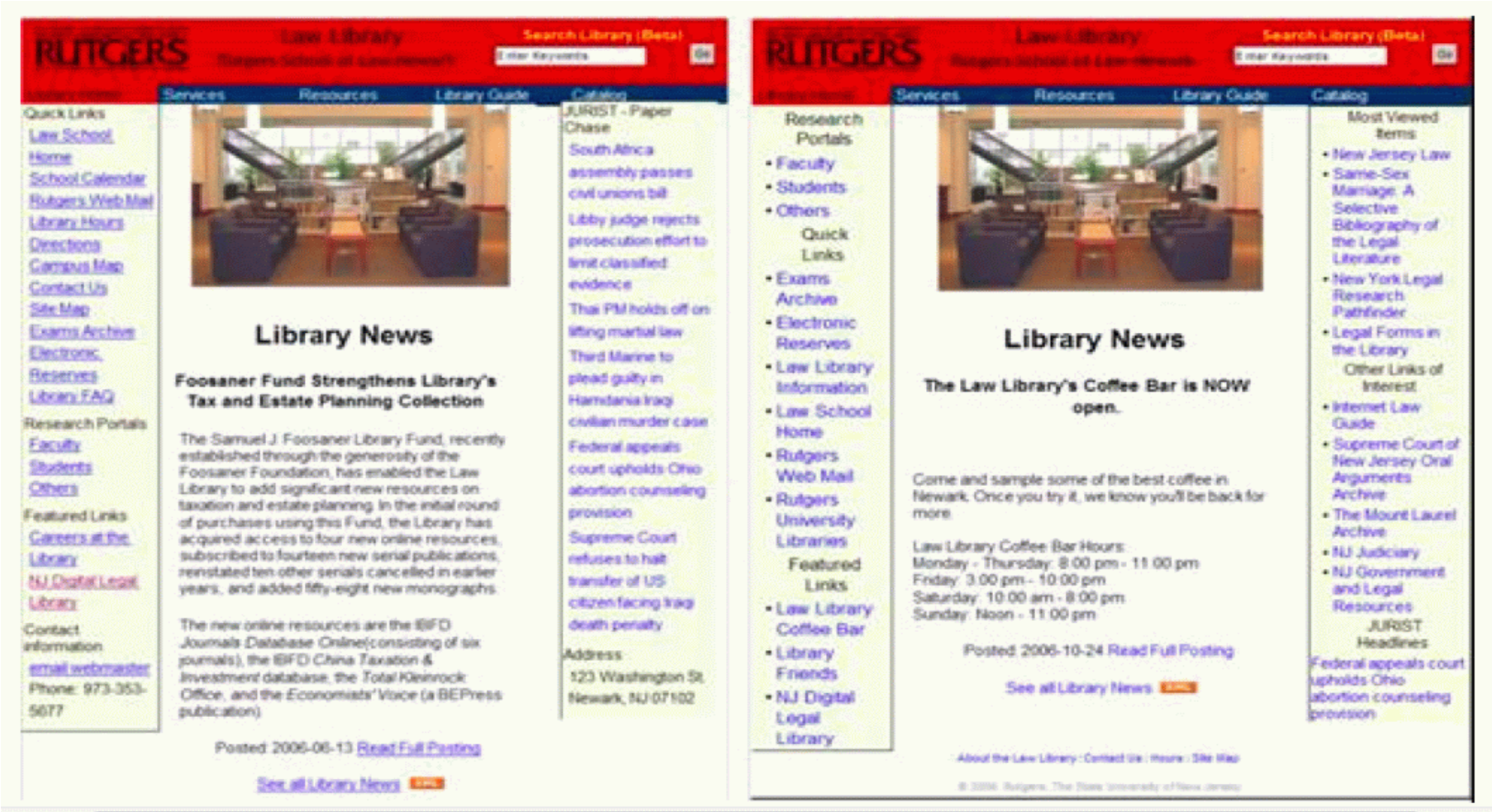
Evaluations

How many users do you have?
What services are they using?
Are usage trends up or down?

Historically, libraries have had little information about many services. Users could come in, then leave, and leave no trace. Online, that's not possible – you can always count what they are doing.

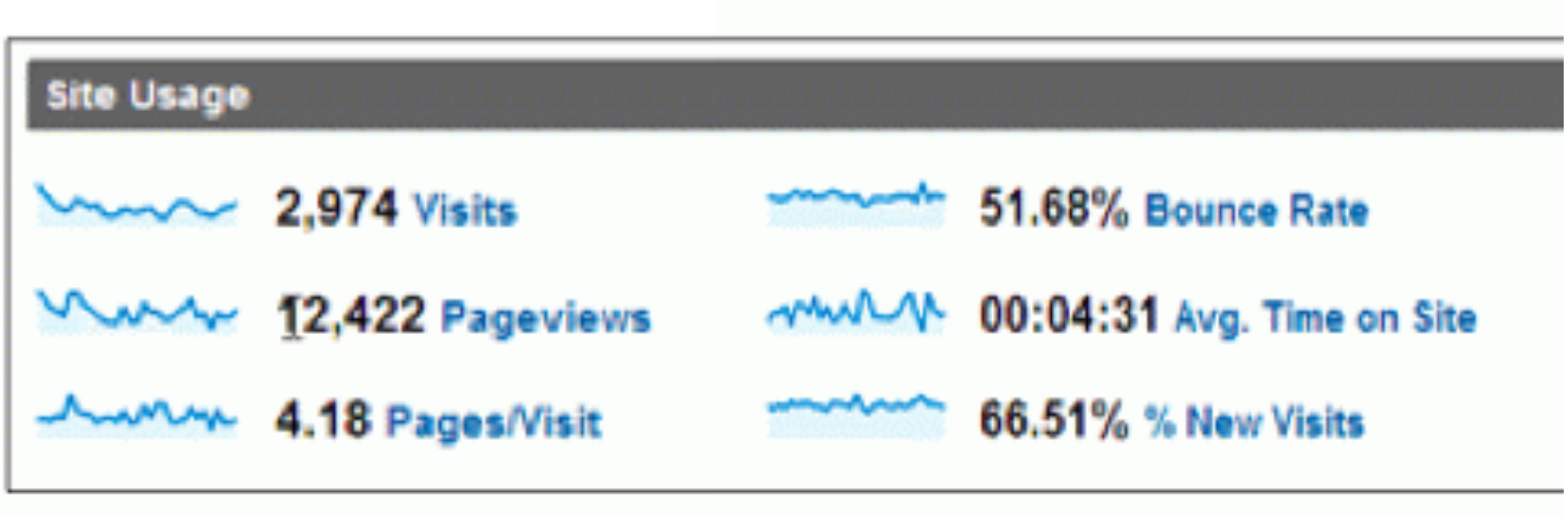
Just because you can count something, of course, doesn't mean that it's useful to do so.

Revised page at Rutgers



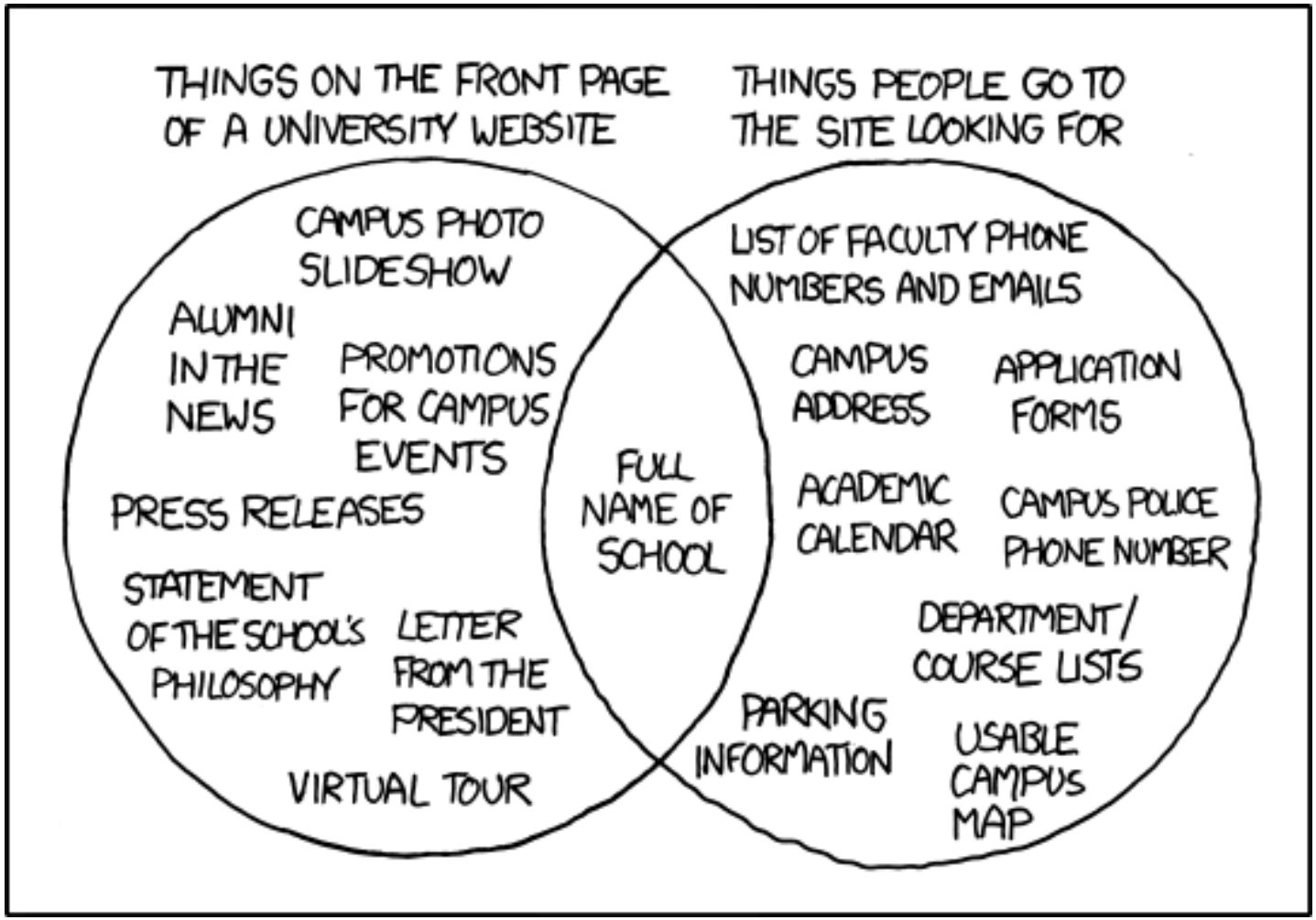
Removed list of headlines, replaced with most viewed. From Wei Fang, "Using Google Analytics for Improving Library Website Content and Design, a Case Study"

Basic site statistics



Google Analytics running on a Wheaton College website. From K. Marek, *Using Web Analytics in the Library*, Library Technology Reports, 2011.

When analytics would help



From xkcd.com

Making pages sticky

In e-commerce, the goal is usually to keep the customer on the site. If you wish people to spend more time on your site, you look at which pages (a) have the longest dwell time, (b) are most likely to send a user to another page on your site rather than send them away.

Then rewrite all the other pages to look like those.

Of course, a person who visits one page and leaves may have gotten precisely the desired information and just done so efficiently; but the assumption is usually failure, not success.

Personalizing

If you find that people who come from location X typically want to look at pages A, B, C – but people from Y want to see D, E, F – you can arrange to show them those pages preferentially.

For example, the Powerhouse Museum in Sydney looks at IP addresses of site visitors: if they are local they see directions and opening hours prominently, but if they are far away, those details are not emphasized.

Privacy

Libraries normally do **not** want to monitor their users.

So most libraries do analytics anonymously.

Some bad examples, however:

- William Weld's medical records were identified in a supposedly anonymized data set.
- Netflix has been sued over the "Netflix Prize" contest where 99 million anonymized video rental records were distributed.

Openness

Libraries normally think that what they do should be open to public inspection; many public libraries are required to operate that way.

If analytics and logfiles pose a risk to privacy, however, then they shouldn’t be made public.

This tension is not well understood.

Tailoring output

If a library is efficient at showing people things they will probably like, it may be showing them things that agree with their previous views. See a book by Eli Pariser entitled *The Filter Bubble*. How well do we want to tailor results?

Another reinforcement phenomenon: more popular items get more attention and as a result are recommended more. My 1997 book has 2 reviews on Amazon, a typical Harry Potter book has more than 3,000. By contrast, traditional cataloging devotes about the same effort to every book.

How do we balance things people like vs. things they ought to know about vs. things some authorities recommend?

Spammers

E-commerce sites have trouble with spammers putting in fake data. Allegedly, the cost of either a favorable review for your business, or an unfavorable one for your competitor's, is about 25 cents on Mechanical Turk. Sites like yelp have software to try to filter these out.

It's the same with paper books – note the *New York Times Book Review* best-seller list symbol for "bulk orders of this book have been reported". It's just cheaper online.

Fortunately, library-patron interactions rarely involve money, so the spammers have less motivation.

Tailoring input

On e-readers it will be possible for vendors to know how readers go through a book. We can imagine mail from Amazon to an author “Only 30% of the people who start reading your book get to the end. The most frequent page on which they give up is page 51. You need another sex scene there.” Would this be a good thing?

E-Commerce

The e-commerce world uses analytics very heavily. Are libraries in competition with them?

For example, Facebook and similar sites know much more about users – gender, age, etc – and use this to give additional and better tailoring. Should libraries do this?

What should i-school students know?, They ought to at least understand what the e-commerce sites are doing. And they should at least be able to do group if not individual evaluations.