Web Search Basics

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Search Engine Optimization

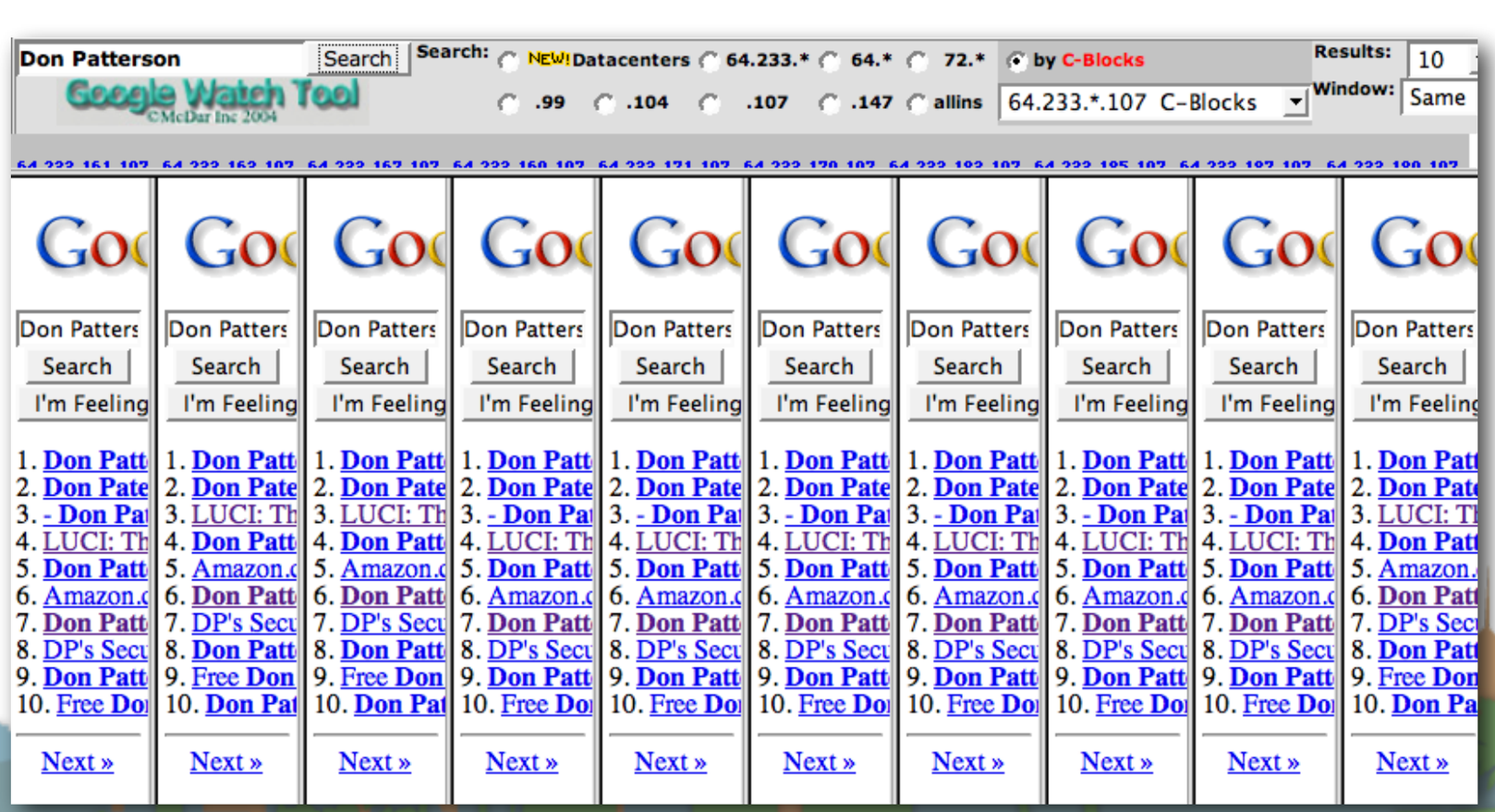
- There are ethical and inethical ways to approach SEO
- Legitimate approach is to:
  - create valuable content
  - make it widely accessible
  - clearly organize it
  - keep it up to date
  - use web standards
  - use web validation tools
  - get high visibility sites to link to your content

Search Engine Optimization

- Inethical approaches (aka spam):
  - lots of tricks
    - make lots of fake pages which point to your site
    - make lots of fake comments on sites which point to your site
    - In a nutshell, “lie”
- Sometimes legitimate and illegitimate techniques are hard to differentiate. It can be a fine line between them.

Search Engine Optimization

- Ranking depends on the data center
  - http://www.flickr.com/photos/the_impression_that_i_get/1321041609/
- Examine the different results:
  - http://www.mcdar.net/dance/index.php



http://www.void.be/googletool.html

Keyword Stuffing

- First Generation Search Engines
  - Heavily relied on tf/idf ratio.
  - E.G. The highest ranking page for the query “brilliant computer scientist” had the most examples of those words.

Keyword Stuffing

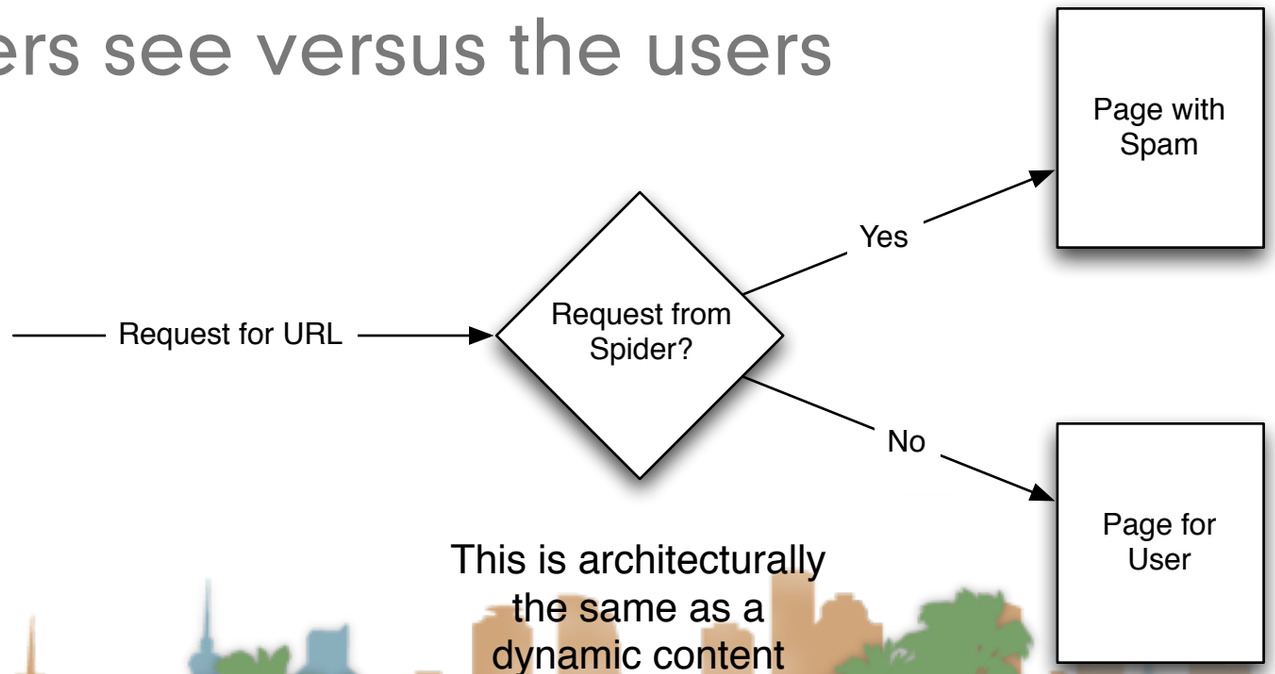
- So SEOs responded by screwing around with keywords
  - Misleading meta-tags
  - Repeating keywords over and over and over and....
  - Playing games with colors. (white on white keywords)
    - visible to spiders but not users in browsers

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.d
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<COMMENT TITLE="MONITOR"></COMMENT>
<meta http-equiv="Content-Language" content="en-us" />
<meta http-equiv="Content-type" content="text/html; charset=iso-8859-1"/>
<META NAME="ROBOTS" CONTENT="NOODP"><meta name="verify-v1" content="aeVxP6zTHeQzT620ipj5+ikXd/VXcdlKoYUJ/C6vVdY=" />
<META NAME="keywords" content="Expedia, Travel, Cheap Airfare, Car, Hotels, Vacations, Airfare, Car Rental, Cruises,

<META NAME="description" content="Purchase airline tickets, make hotel reservations, find vacation packages, car rent
```

Keyword Stuffing

- Cloaking
  - define: Serving different content to a spider than to a user.
  - More sophisticated versions of differentiating what the spiders see versus the users



This is architecturally the same as a dynamic content engine

Other spam techniques

- Doorway pages
  - Like cloaking but using a redirect
  - Initial page is optimized for a keyword then a redirect takes the user to the "real" page
- Link spamming
  - Programs that search for blogs and automatically leave comments with links
- Robot Clicker-Fraud
  - Programs that "click" on query results to up their value.

Spam Industry

Advanced Traffic:
Get a **first page listing on Google - GUARANTEED!** For maximum search engine traffic - the best of SEO and search advertising. Visitors in just 48 hours from $7/day. **Discover the traffic potential!**

Find out more

ORDER NOW

WARNING: This site contains sneaky, underhanded Black Hat Seo tactics.

Black Hat Seo is responsible for more online fortunes than you'd care to imagine but it's NOT for everybody.

Make Money Blogging
See How I Earn Over Six Figures a year Blogging

I Will Get Your Website to the Top of Google!

The art of search engine optimization...gaining **top spots on Google**...is no easy chore. I know...this is my job...

I assist people in getting top positions for their websites on Google, Yahoo, MSN and all the other major search engines.

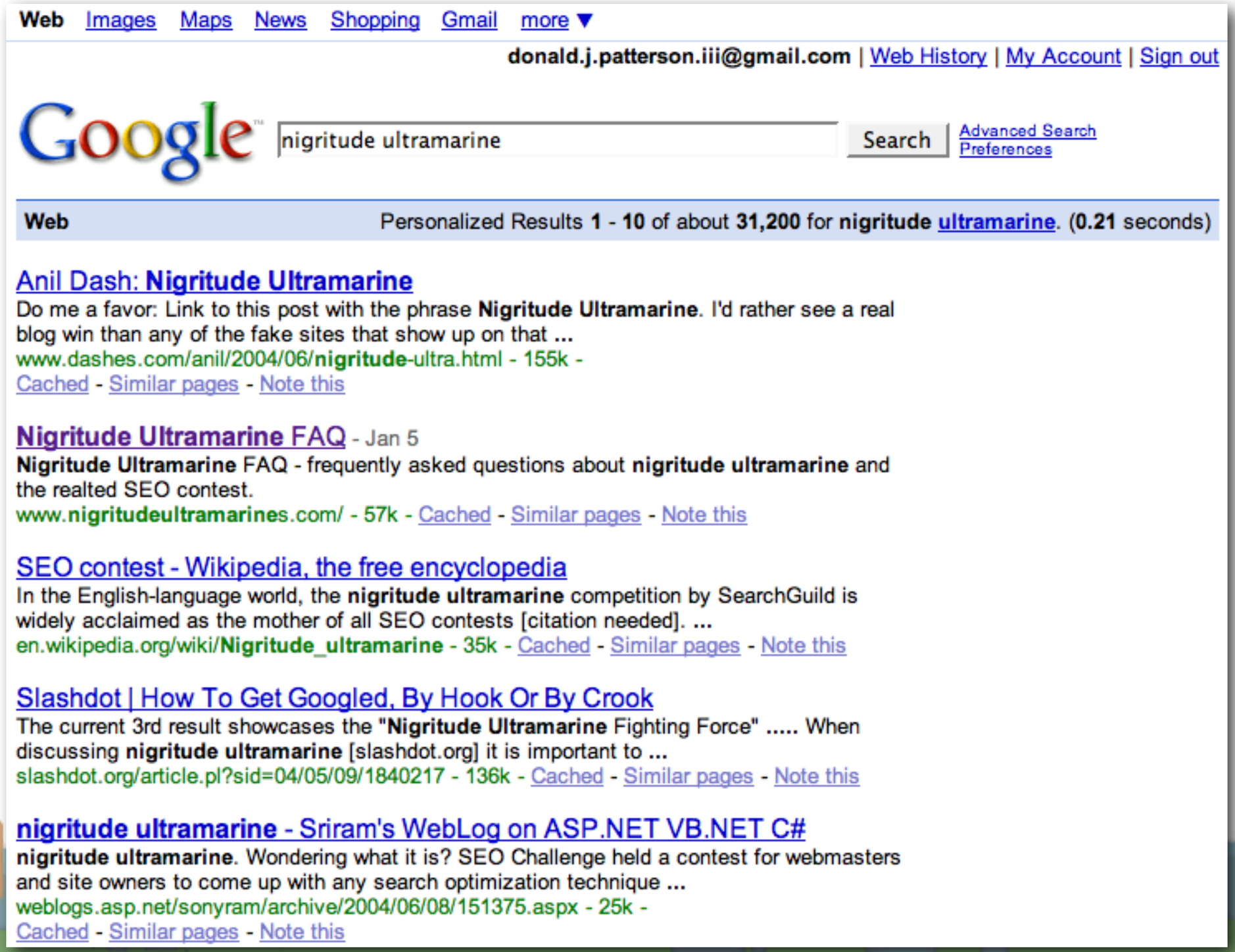
There are a few givens on the internet when it comes to trying to market goods and services:

No Traffic=No Sales!

End of story...that's it...bottom line!

If you have a web

Spam Contest



Web Images Maps News Shopping Gmail more ▼

donald.j.patterson.iii@gmail.com | Web History | My Account | Sign out

Google nigritude ultramarine Search Advanced Search Preferences

Web Personalized Results **1** - **10** of about **31,200** for **nigritude ultramarine**. (**0.21** seconds)

Anil Dash: **Nigritude Ultramarine**
Do me a favor: Link to this post with the phrase **Nigritude Ultramarine**. I'd rather see a real blog win than any of the fake sites that show up on that ...
www.dashes.com/anil/2004/06/**nigritude**-ultra.html - 155k -
Cached - Similar pages - Note this

Nigritude Ultramarine FAQ - Jan 5
Nigritude Ultramarine FAQ - frequently asked questions about **nigritude ultramarine** and the realted SEO contest.
www.**nigritudeultramarine**s.com/ - 57k - Cached - Similar pages - Note this

SEO contest - Wikipedia, the free encyclopedia
In the English-language world, the **nigritude ultramarine** competition by SearchGuild is widely acclaimed as the mother of all SEO contests [citation needed]. ...
en.wikipedia.org/wiki/**Nigritude_ultramarine** - 35k - Cached - Similar pages - Note this

Slashdot | How To Get Googled, By Hook Or By Crook
The current 3rd result showcases the "**Nigritude Ultramarine** Fighting Force" ..... When discussing **nigritude ultramarine** [slashdot.org] it is important to ...
slashdot.org/article.pl?sid=04/05/09/1840217 - 136k - Cached - Similar pages - Note this

nigritude ultramarine - Sriram's WebLog on ASP.NET VB.NET C#
nigritude ultramarine. Wondering what it is? SEO Challenge held a contest for webmasters and site owners to come up with any search optimization technique ...
weblogs.asp.net/sonyram/archive/2004/06/08/151375.aspx - 25k -
Cached - Similar pages - Note this

The war on spam

- Quality Indicators
  - Statistical Analysis of Links (aka PageRank)
    - votes from authors
  - Usage indicators (users visiting a page)
    - votes from users
- Anti-Robot techniques
  - "Captchas"
  - Completely Automated Public Turing Computers and Humans Apart

The war on spam

- Limits on meta keywords
- Spam Recognition by machine learning
- “no-follow” attribute
- Family Friendly filters
- Automatic Detection of Pornography
  - Often the spammers desired landing page
- Text Analysis
  - Look for keywords and variants

The war on spam

- Robust Link Analysis
  - Ignore statistically improbable links
  - Use link analysis to detect spammers
    - "Guilt by association"

The war on spam

- Editorial Intervention
  - Blacklists
  - Query Reviews
  - Customer Complaints
- Visualization Tools

Webmaster Guidelines

- Search Engines have SEO policies
  - What is allowed and not allowed
- Example: Search for “google webmaster guidelines” or “msn guidelines for successful indexing”
- Ignore them at your own risk
- Once you are blacklisted by a search engine you will disappear from the web
  - Remember how search engines enable scalability?
- Adversarial IR Research:
  - http://airweb.cse.lehigh.edu/

Spam - Extra Credit

flatricide pulgamitude