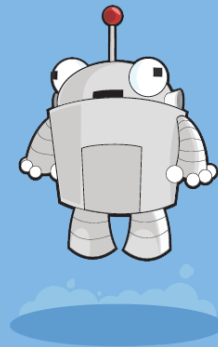


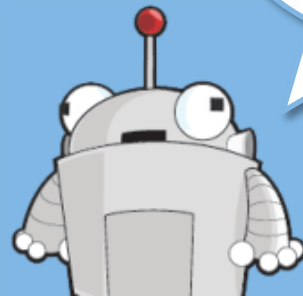


# Analyzing Search Engines with Statistics



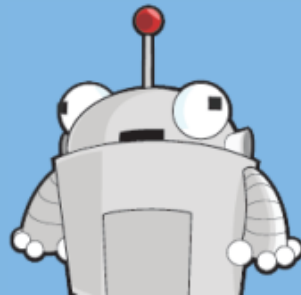
Rand Fishkin, CEO & Co-founder, SEOMoz  
Inbound Marketing Summit, October 2010

# The Big Picture: Use Statistical Analysis to Answer Important SEO Questions



Resources Available:

[www.seomoz.org/dp/ims2010](http://www.seomoz.org/dp/ims2010)



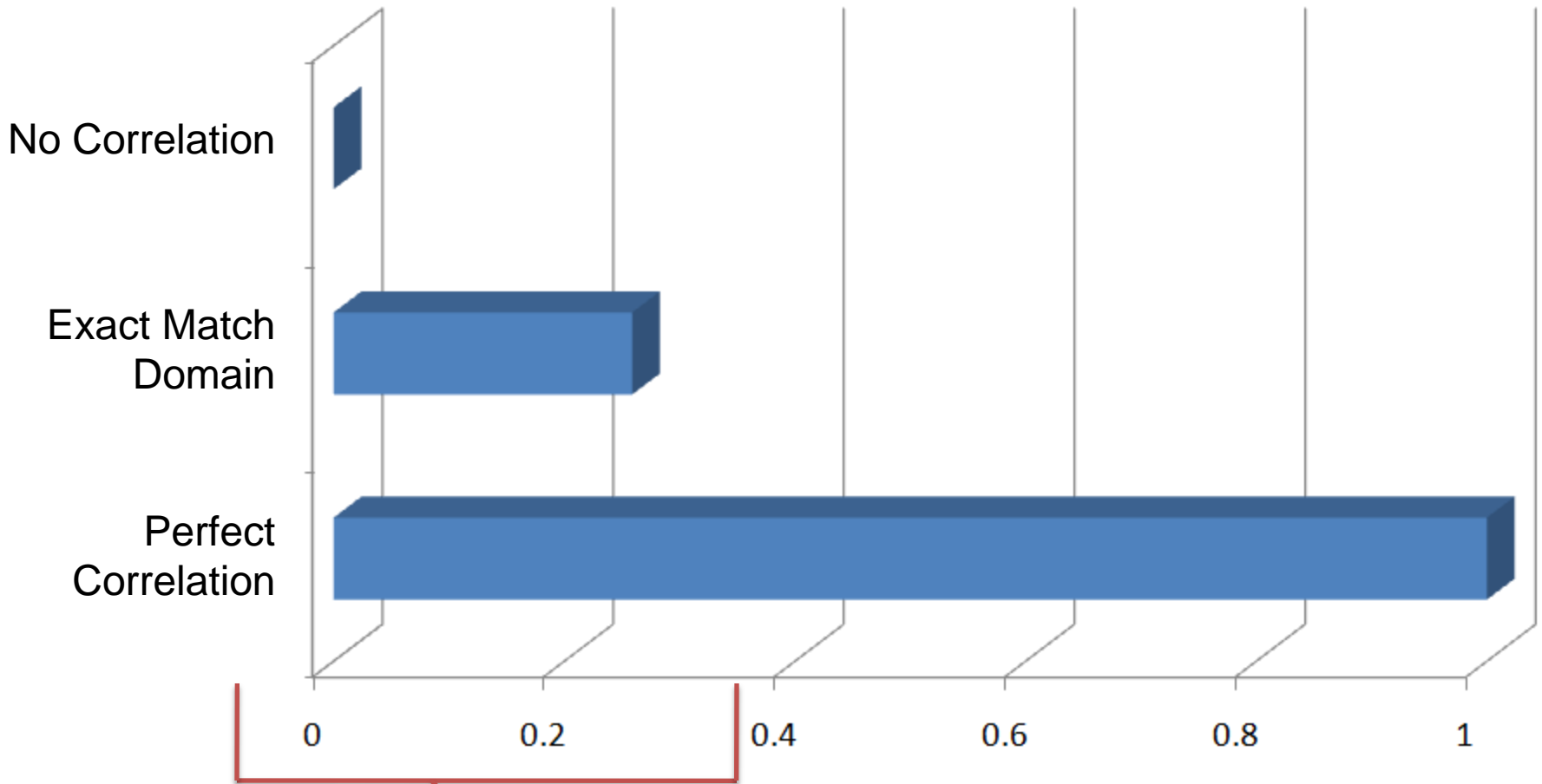
# Correlation $\neq$ Causation



The more I wear suits, the more I speak on panels.

**Therefore:** wearing suits causes me to speak on panels.

# Understanding Correlation Significance



Most of our data for search rankings falls in this region  
(which we'd expect given algorithms w/ 200+ ranking factors)

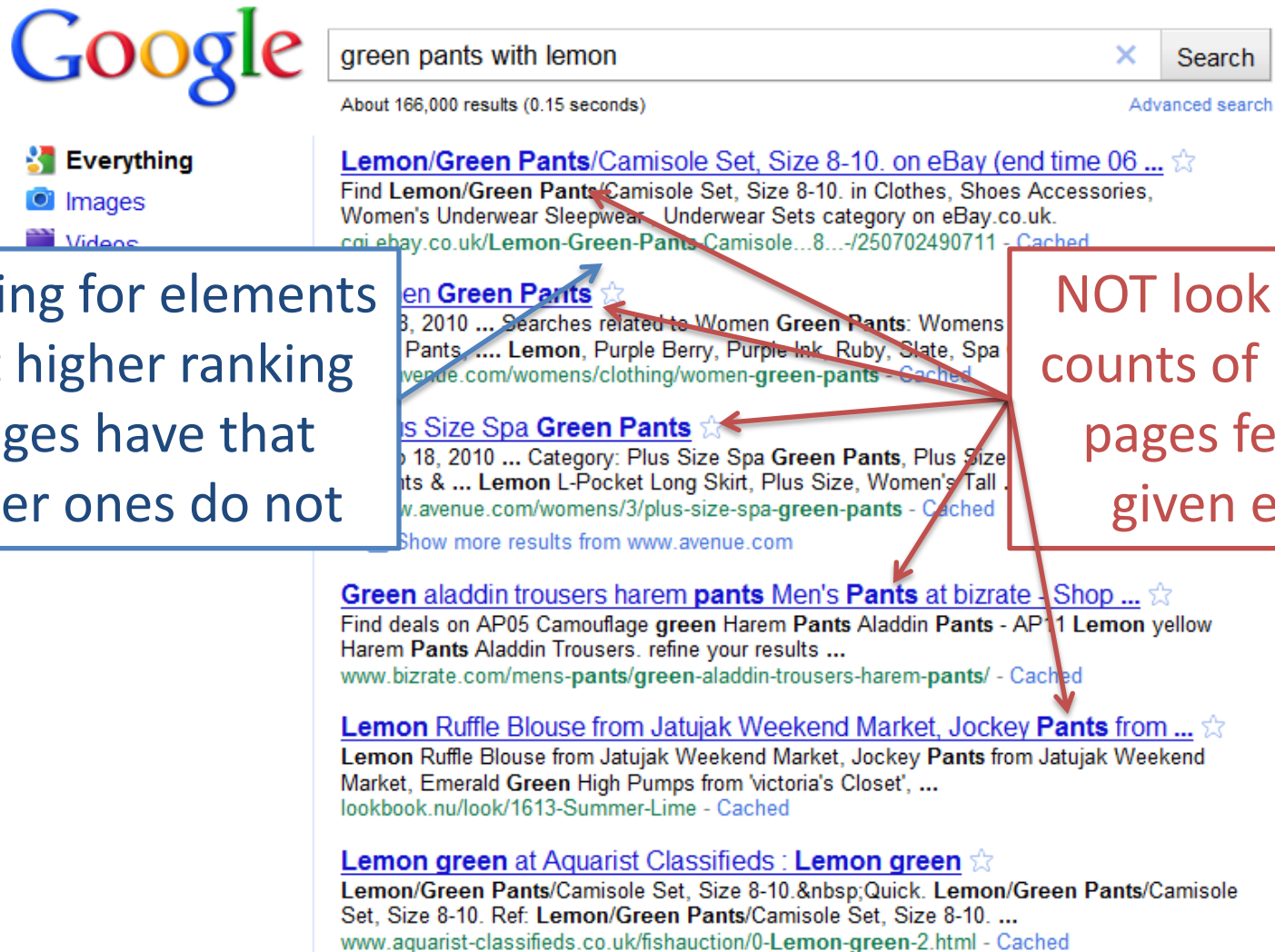
Question #1:

# How to Best “Optimize” a Site for Search Engine Rankings

# Methodology

- 11,351 SERPs via Google AdWords Suggest
- 1<sup>st</sup> Page Only (usually ~10 results per page)
- Correlations are w/ Higher Position on Page 1
- Controlled for SERPs Where All (or None) of the Results Matched the Metric

# Methodology

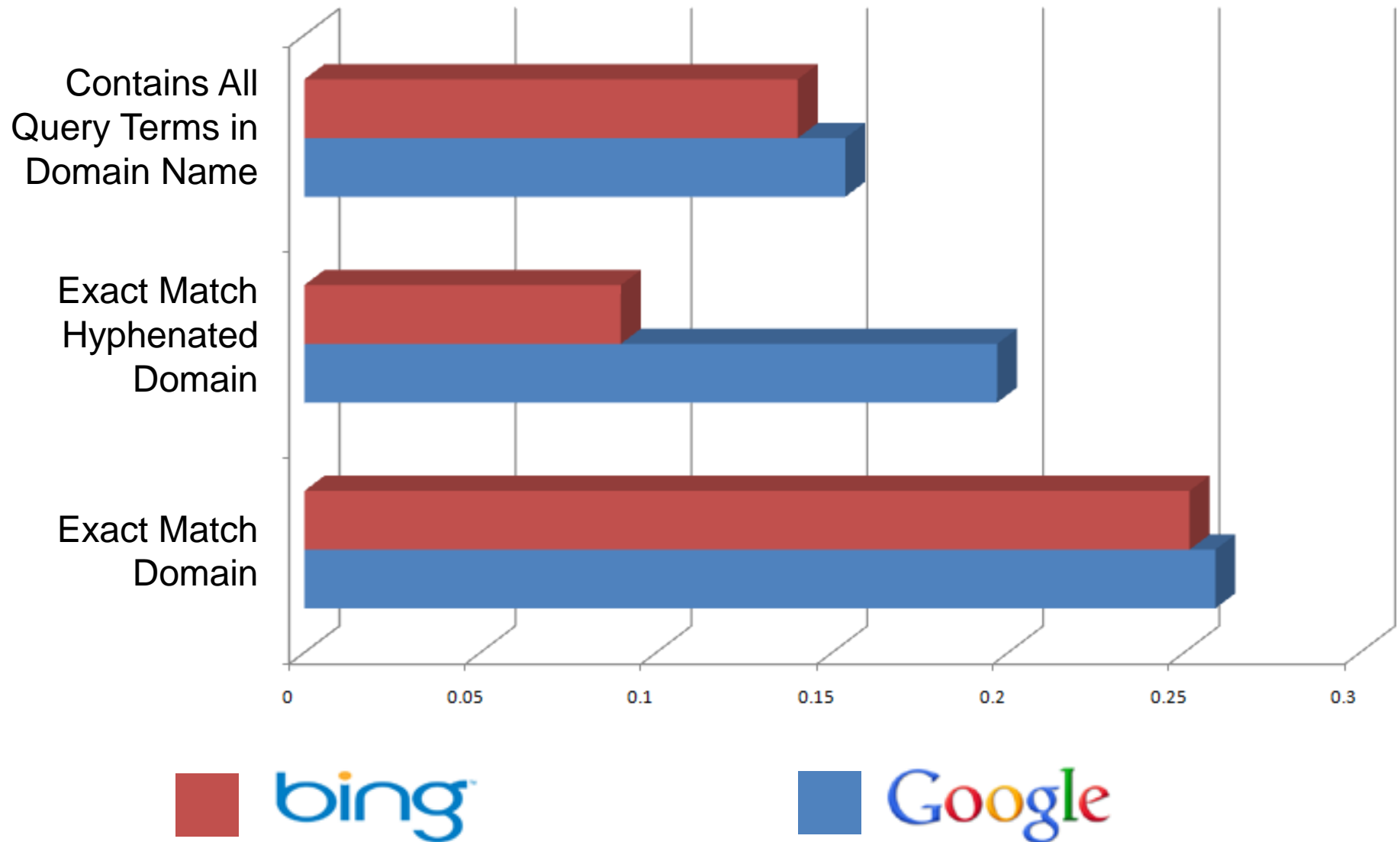


The image shows a Google search interface for the query "green pants with lemon". The search results are displayed, showing various items like "Lemon/Green Pants/Camisole Set" and "Green aladdin trousers harem pants". Two text boxes with arrows pointing to specific search results explain the methodology:

- Looking for elements that higher ranking pages have that lower ones do not** (points to the first result: "Lemon/Green Pants/Camisole Set, Size 8-10. on eBay")
- NOT looking at raw counts of how many pages featured a given element** (points to the second result: "Green aladdin trousers harem pants Men's Pants at bizrate")

The search results shown include:

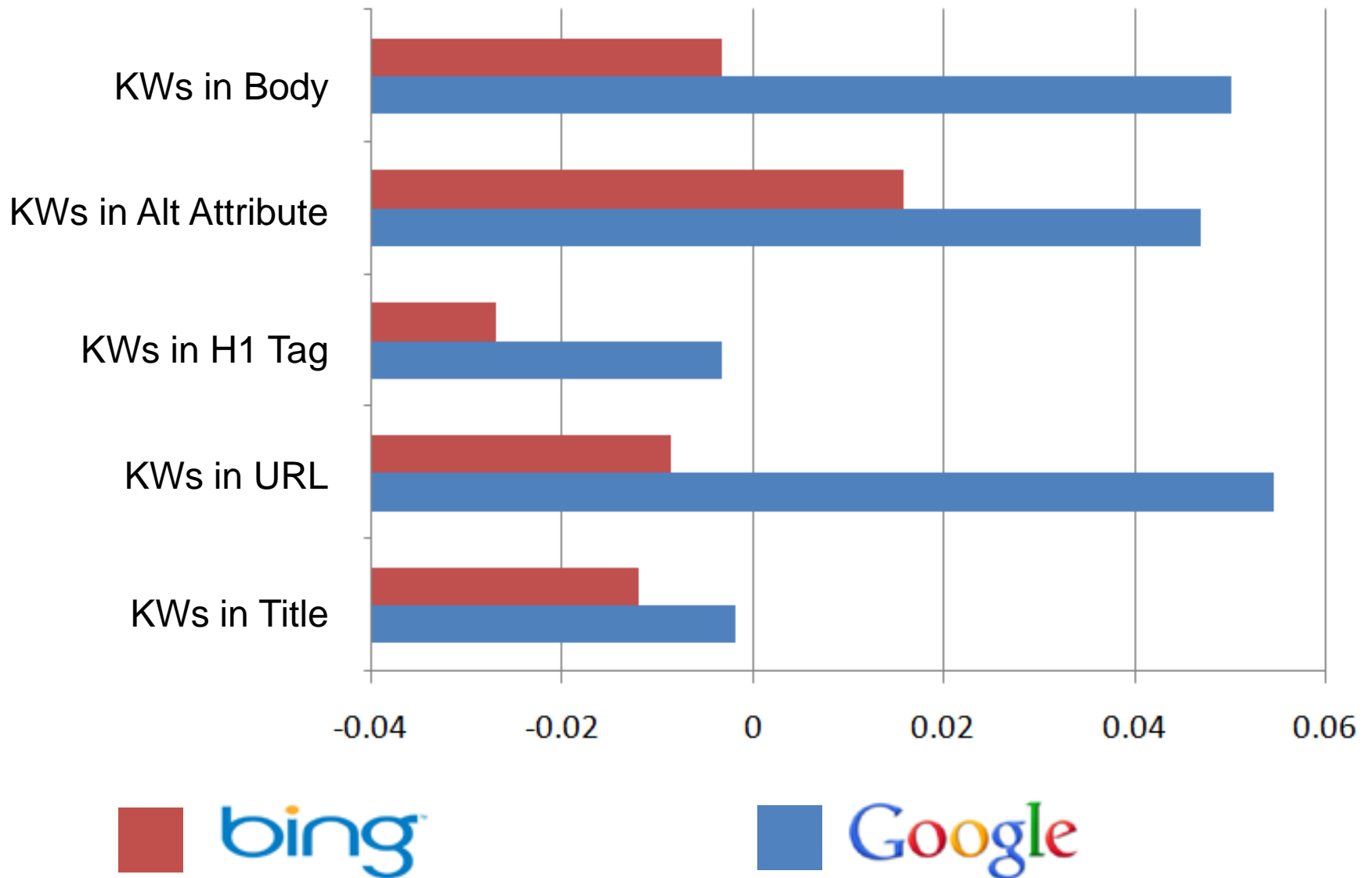
- Lemon/Green Pants/Camisole Set, Size 8-10. on eBay (end time 06 ...**  
Find **Lemon/Green Pants/Camisole Set, Size 8-10.** in Clothes, Shoes Accessories, Women's Underwear Sleepwear. Underwear Sets category on eBay.co.uk.  
[cgi.ebay.co.uk/Lemon-Green-Pants-Camisole...8...-/250702490711](#) - **Cached**
- en Green Pants**  
3, 2010 ... Searches related to Women **Green Pants**: Womens **Pants**, .... **Lemon**, Purple Berry, Purple Ink, Ruby, Slate, Spa  
[vende.com/womens/clothing/women-green-pants](#) - **Cached**
- Plus Size Spa Green Pants**  
p 18, 2010 ... Category: Plus Size Spa **Green Pants**, Plus Size **Pants** & ... **Lemon** L-Pocket Long Skirt, Plus Size, Women's Tall  
[w.avenue.com/womens/3/plus-size-spa-green-pants](#) - **Cached**  
Show more results from [www.avenue.com](#)
- Green aladdin trousers harem pants Men's Pants at bizrate - Shop ...**  
Find deals on AP05 Camouflage **green** Harem **Pants** Aladdin **Pants** - AP01 **Lemon** yellow Harem **Pants** Aladdin Trousers. refine your results ...  
[www.bizrate.com/mens-pants/green-aladdin-trousers-harem-pants/](#) - **Cached**
- Lemon Ruffle Blouse from Jatujak Weekend Market, Jockey Pants from ...**  
**Lemon** Ruffle Blouse from Jatujak Weekend Market, Jockey **Pants** from Jatujak Weekend Market, Emerald **Green** High Pumps from 'victoria's Closet', ...  
[lookbook.nu/look/1613-Summer-Lime](#) - **Cached**
- Lemon green at Aquarist Classifieds : Lemon green**  
**Lemon/Green Pants/Camisole Set, Size 8-10.**&nbsp;Quick. **Lemon/Green Pants/Camisole Set, Size 8-10.** Ref: **Lemon/Green Pants/Camisole Set, Size 8-10.** ...  
[www.aquarist-classifieds.co.uk/fishauction/0-Lemon-green-2.html](#) - **Cached**



Highest Stderr = 0.0241804

# Our Interpretation

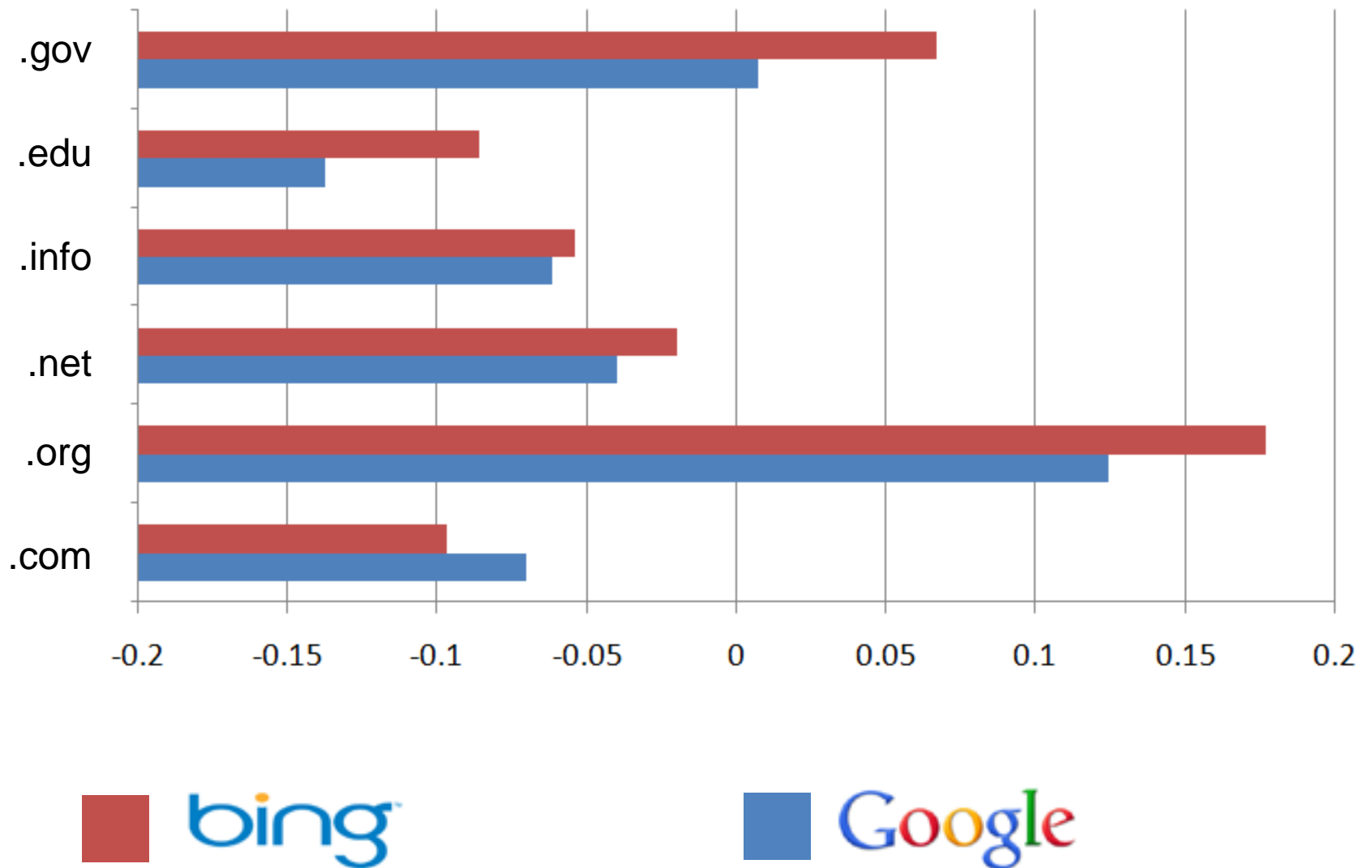
- Exact match domains remain powerful in both engines (anchor text could be a factor, too)
- Hyphenated versions are less powerful, though more frequent in Bing (G: 271 vs. B: 890)
- Just having keywords in the domain name has substantial positive correlation



Highest Stderr = 0.00350211

# Our Interpretation

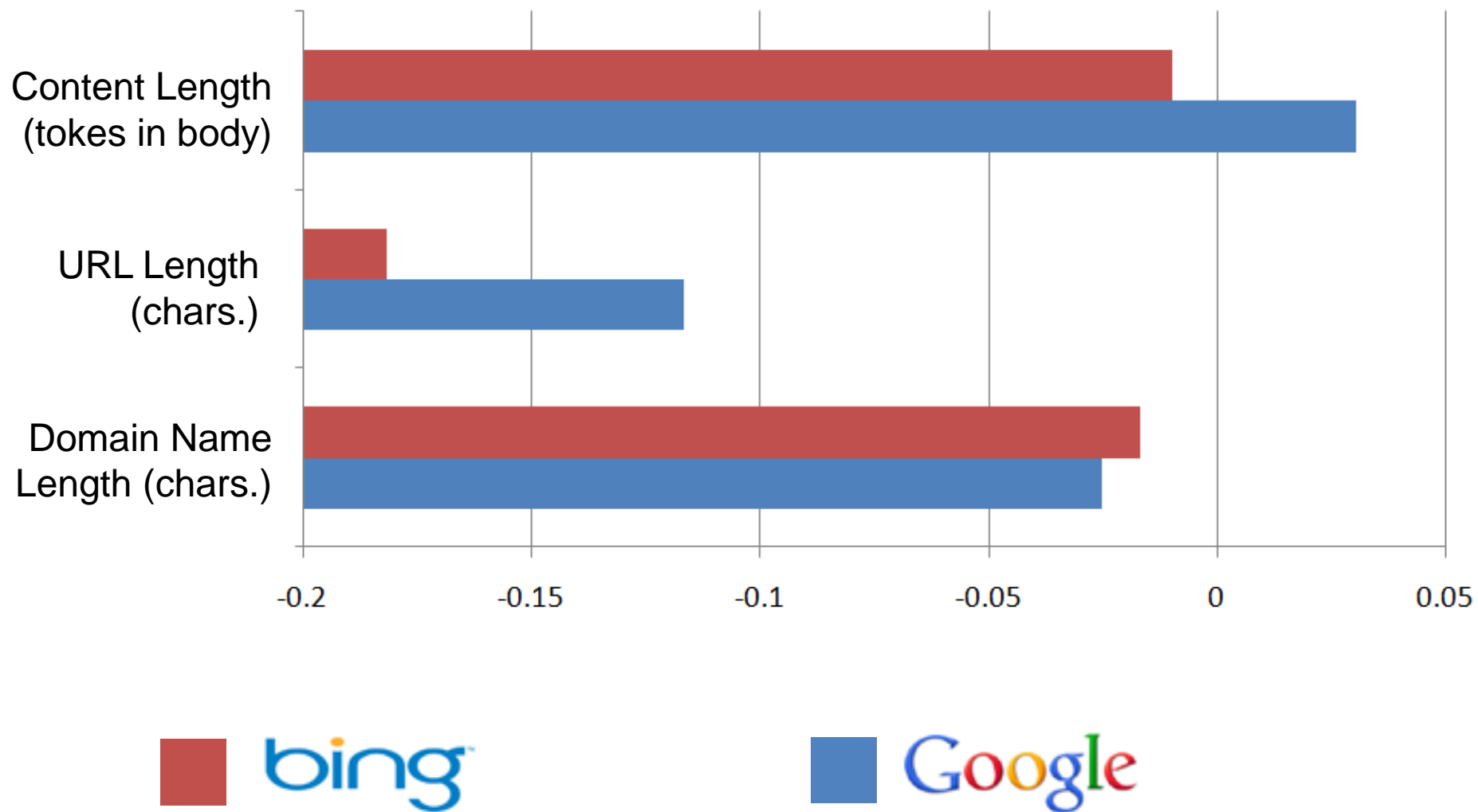
- The Alt attribute of images is interesting
- Putting KWs in URLs is likely a best practice
- Everyone optimizes titles (G: 11,115 vs. B: 11,143).  
Differentiating here is hard.
- (Simplistic) on-page optimization isn't a huge factor



Highest Stderr = 0.0269818

# Our Interpretation

- More reasons to believe Google when they say .gov, .info and .edu are not special cased
- The .org TLD extension is surprising – do they earn more links? Less spam? More non-commercial?
- Don't forget about branding/user behavior - .com is still probably a very good thing (at least own it)



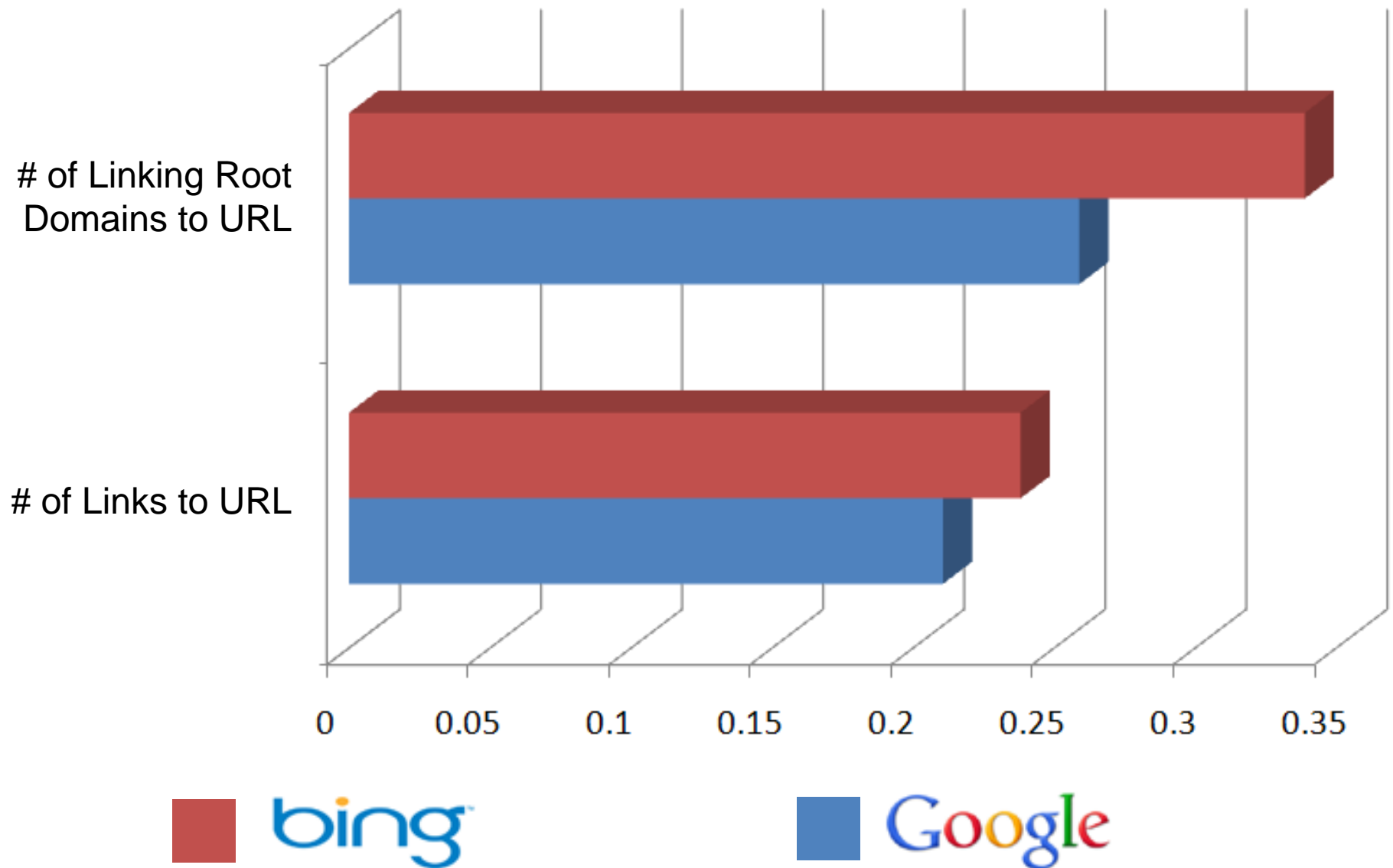
Highest Stderr = 0.0033353

# Our Interpretation

- Shorter URLs are likely a good best practice (especially on Bing)
- Long domains may not be ideal, but aren't awful
- Raw content length seems marginal in correlation

Question #2:

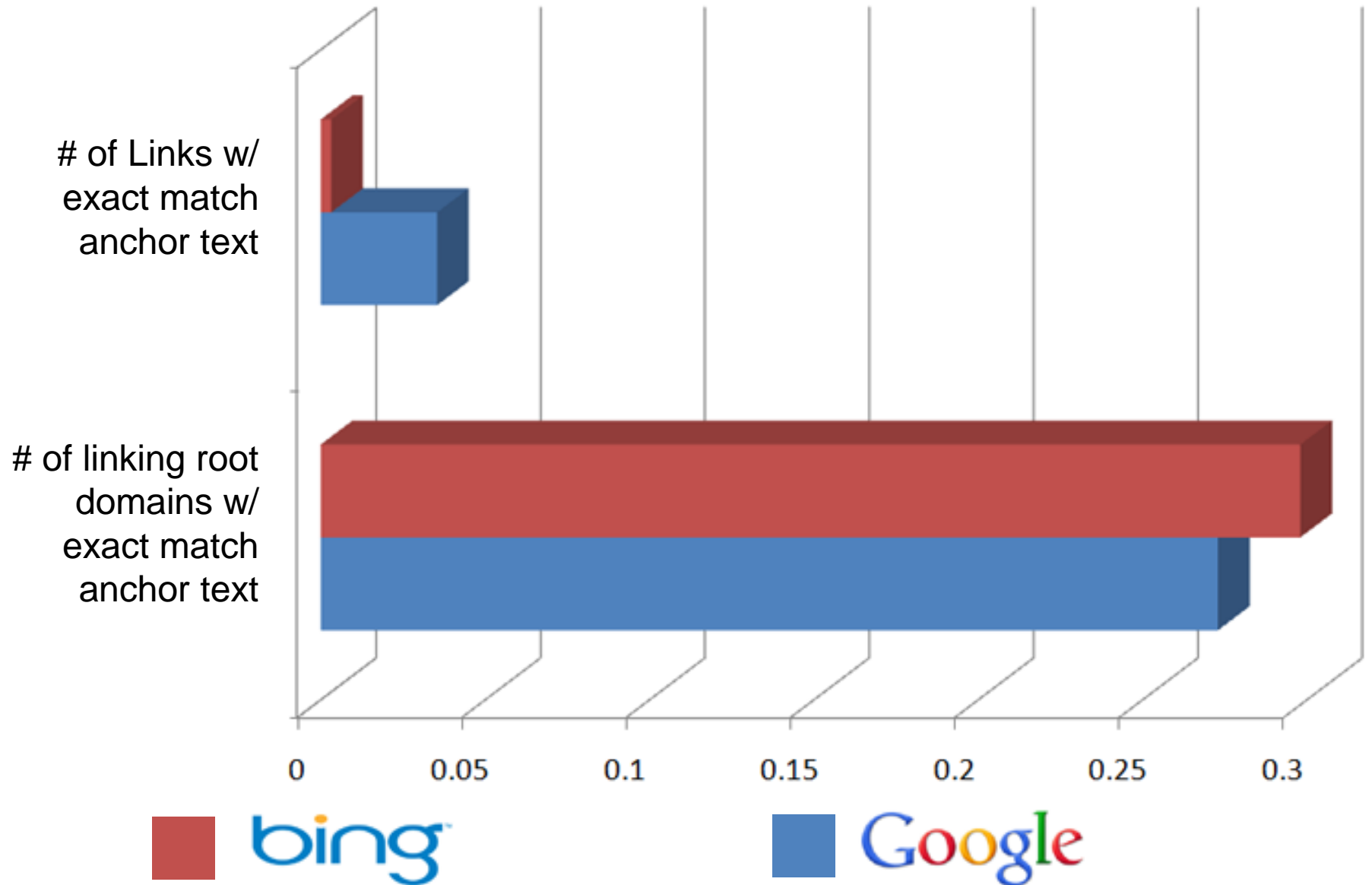
**What Kind of Links Matter & How  
Should We Evaluate Links?**



Highest Stderr = 0.00335677

# Our Interpretation

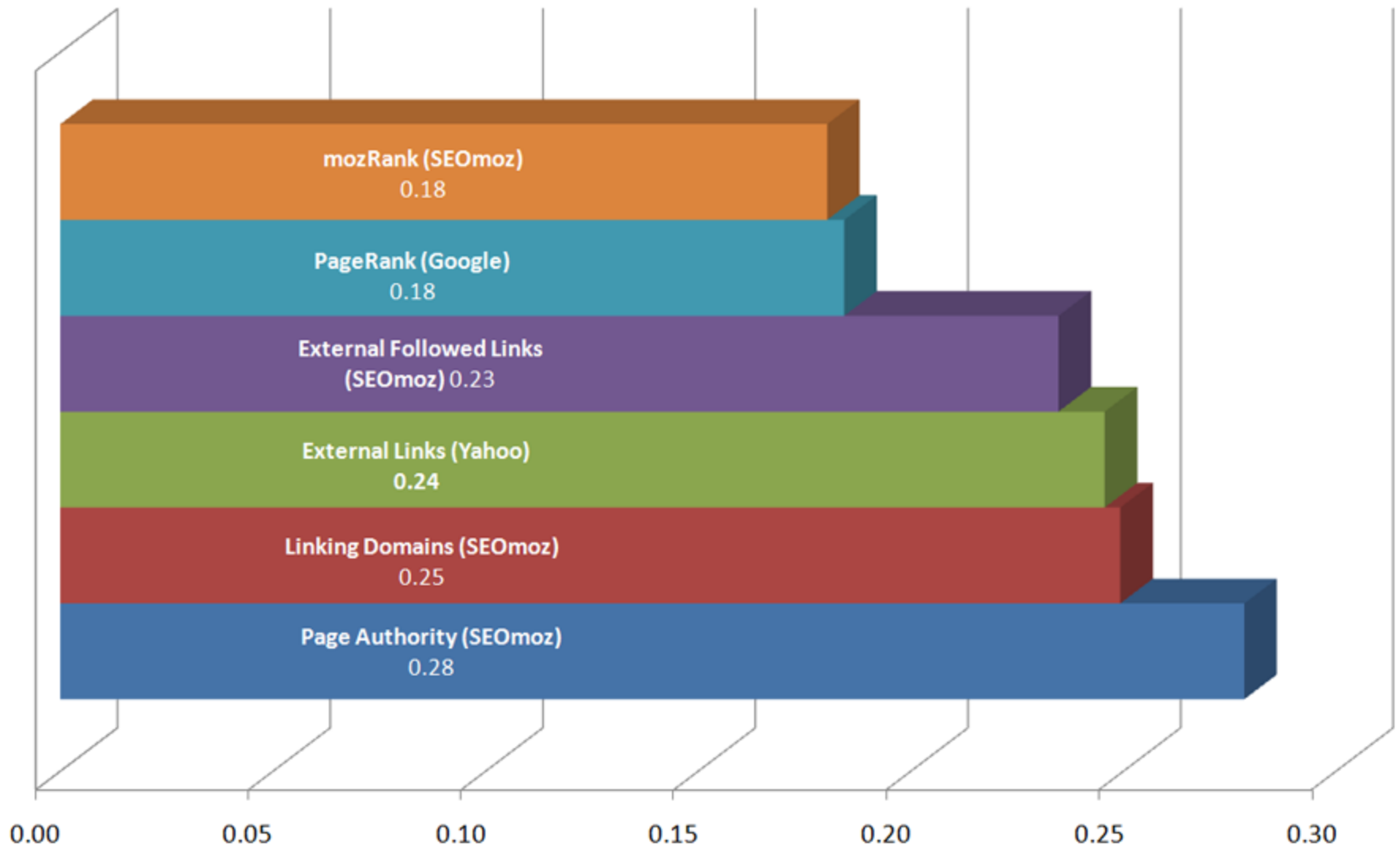
- Links are likely still a major part of the algorithms
- Bing may be slightly more naïve in their usage of link data than Google, but better than before
- Diversity of link sources remains more important than raw link quantity



Highest Stderr = 0.00415058

# Our Interpretation

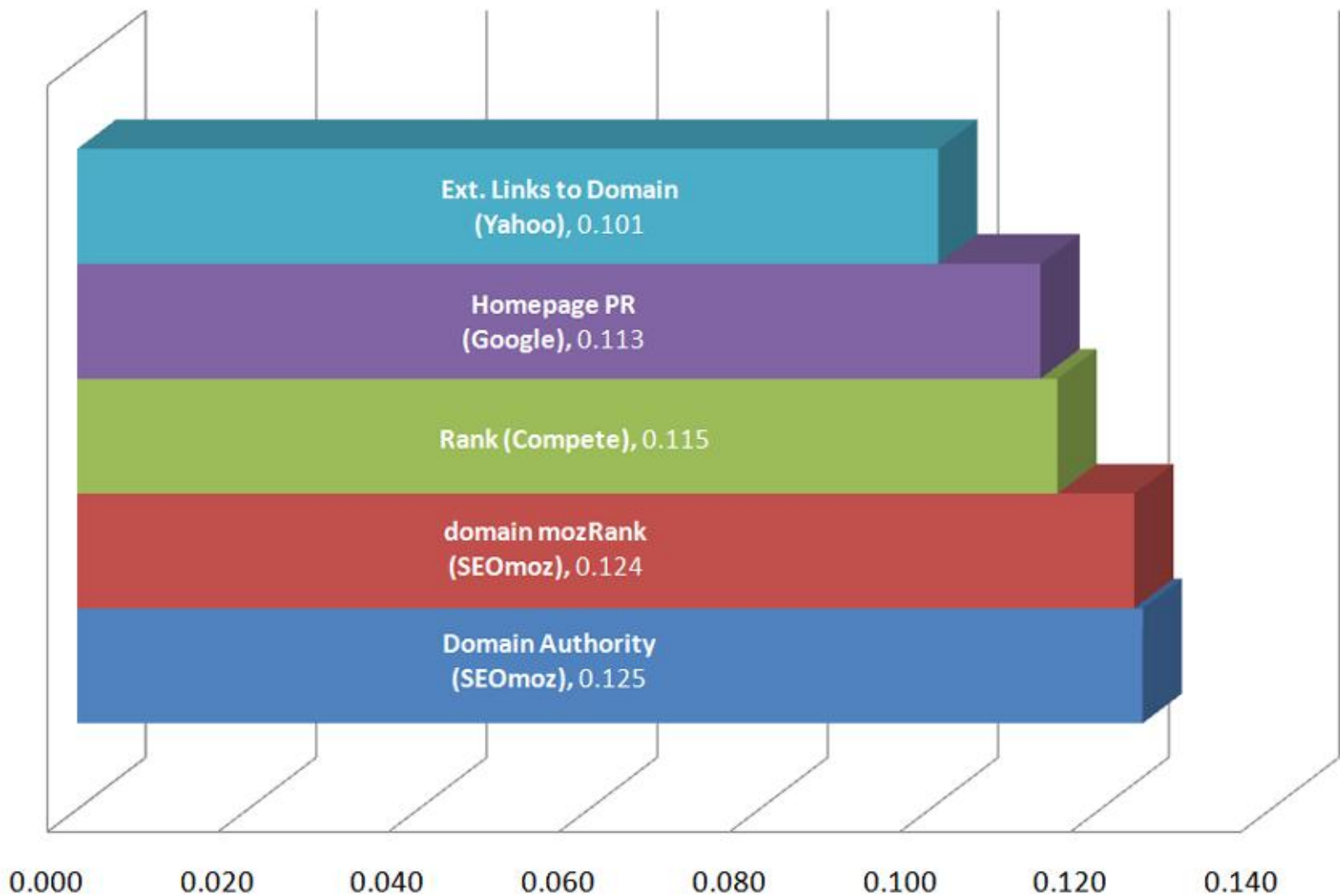
- Many anchor text links from the same domain likely don't add much value
- Anchor text links from diverse domains, however, appears highly correlated
- Bing and Google are relatively similar in evaluating these metrics



Correlation of Page-Level Link Valuation Metrics

# Our Interpretation

- PageRank (and similar algorithms) are not particularly representative of rankings (but are somewhat correlated)
- Linking domains are likely a better metric than raw links
- Page Authority is reasonably good, but has a way to go



Correlation of Domain-Level Link Valuation Metrics

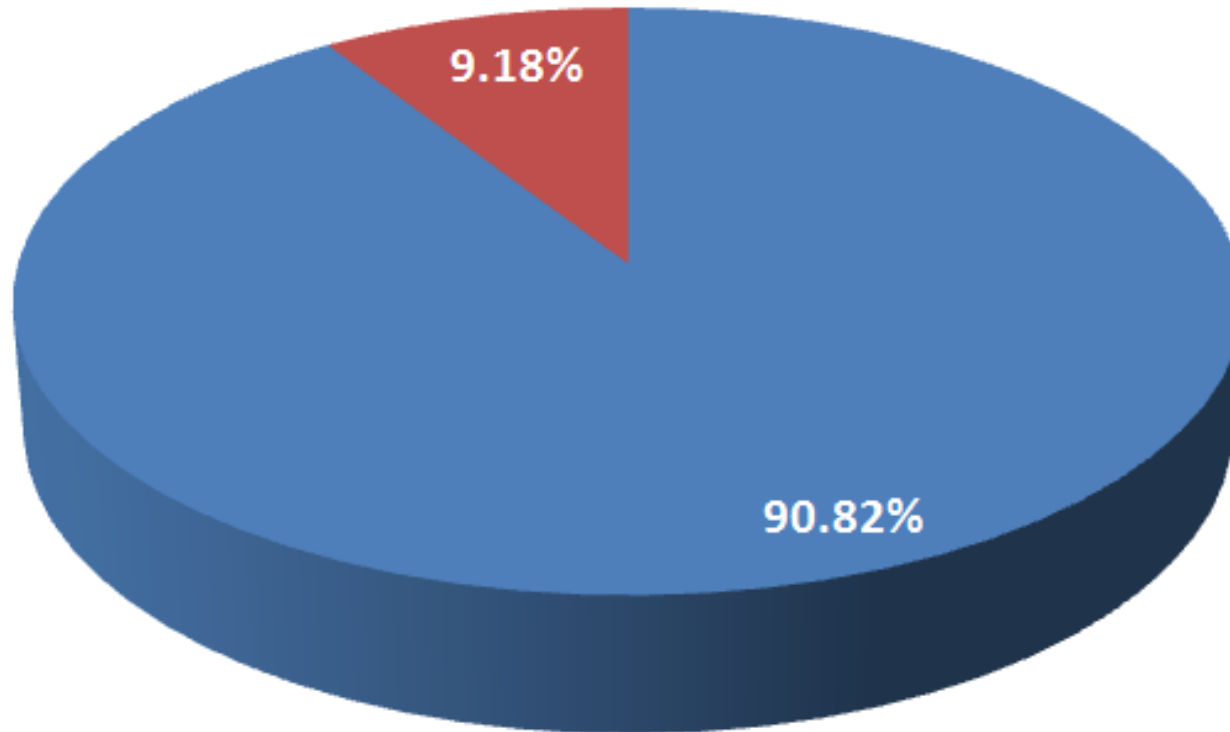
# Our Interpretation

- No single domain valuation metric is especially well correlated with rankings
- Rankings of individual pages may be more disparate we typically think re: “domain authority”
- Overall, we’re still very naïve when it comes to understanding how links influence search rankings

Question #3:

**How Does Google Instant Change  
Keyword Demand / SEO?**

# Are Most Users Seeing/Using Google Instant?



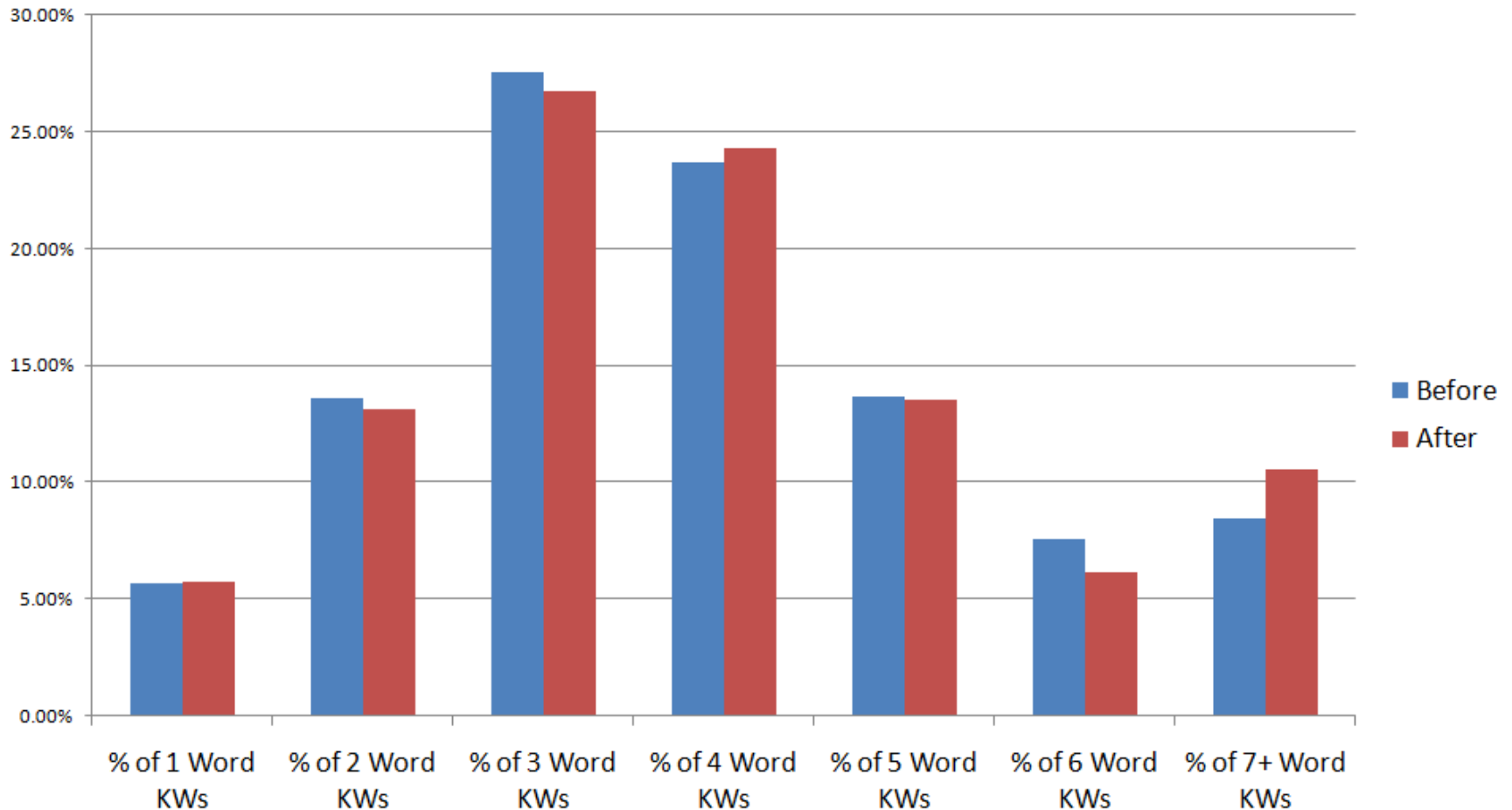
■ Google.com

■ Firefox Search Bar

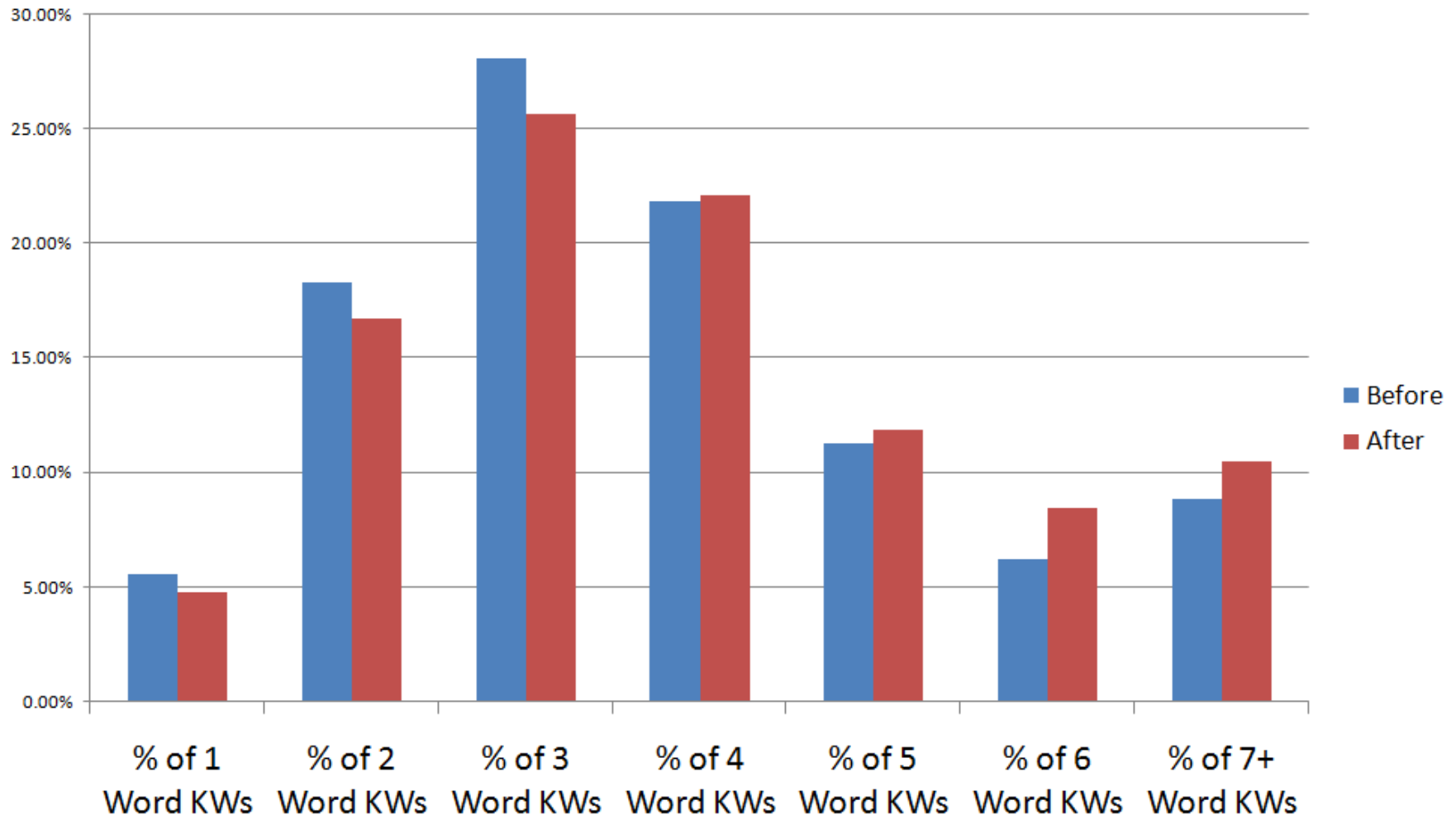
## Methodology: Keyword Referral Search Data

- Look at keyword sending traffic via analytics
- Distribute into groups by word-length
- Analyze shifts in demand by keywords that brought visits to the site
- Compare from period prior to Google Instant and directly after

Via MEC Manchester (UK)  
5 Sites, 4 Verticals, 10K+ Keywords

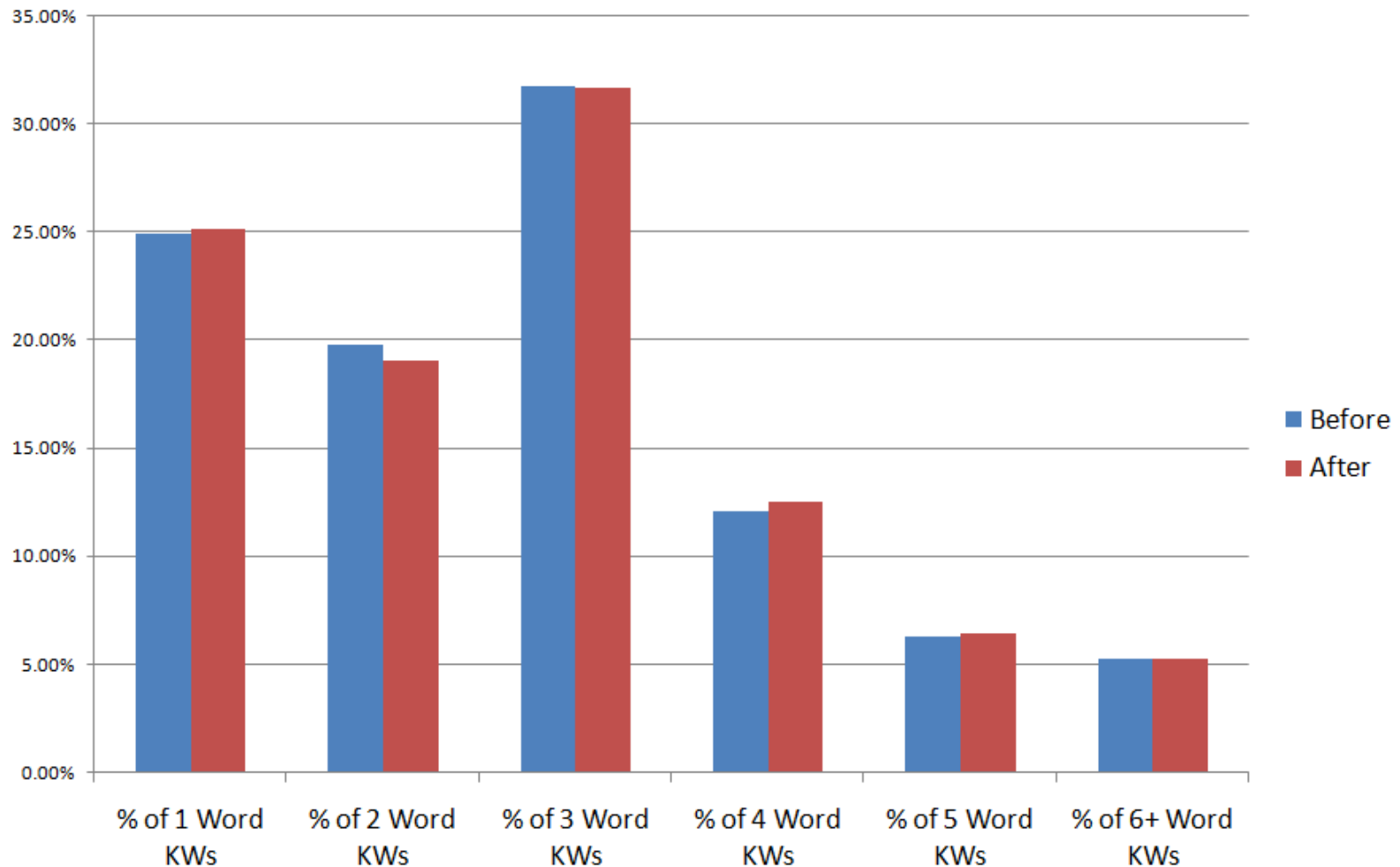


Via Distilled Consulting (UK)  
11 Sites, Various Sizes (3.5K – 75K weekly visits), 75K+ Keywords



## Via Conductor

Multiple sites, 880K visits, 10Ks of keywords



# Interesting Takeaways

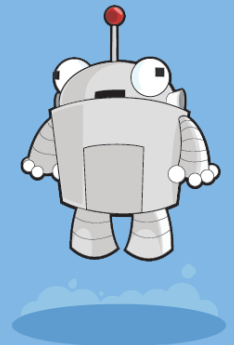
- Google Instant seems not to have shifted keyword demand by much (if at all)
- Google “suggest” has been out for a long time already; users are likely accustomed to this feature
- The “long tail” may get longer/shorter over time, but Instant seems less responsible than other factors



## Q+A

Rand Fishkin, CEO & Co-Founder, SEOMoz

- Twitter: @randfish
- Blog: [www.seomoz.org/blog](http://www.seomoz.org/blog)
- Email: [rand@seomoz.org](mailto:rand@seomoz.org)



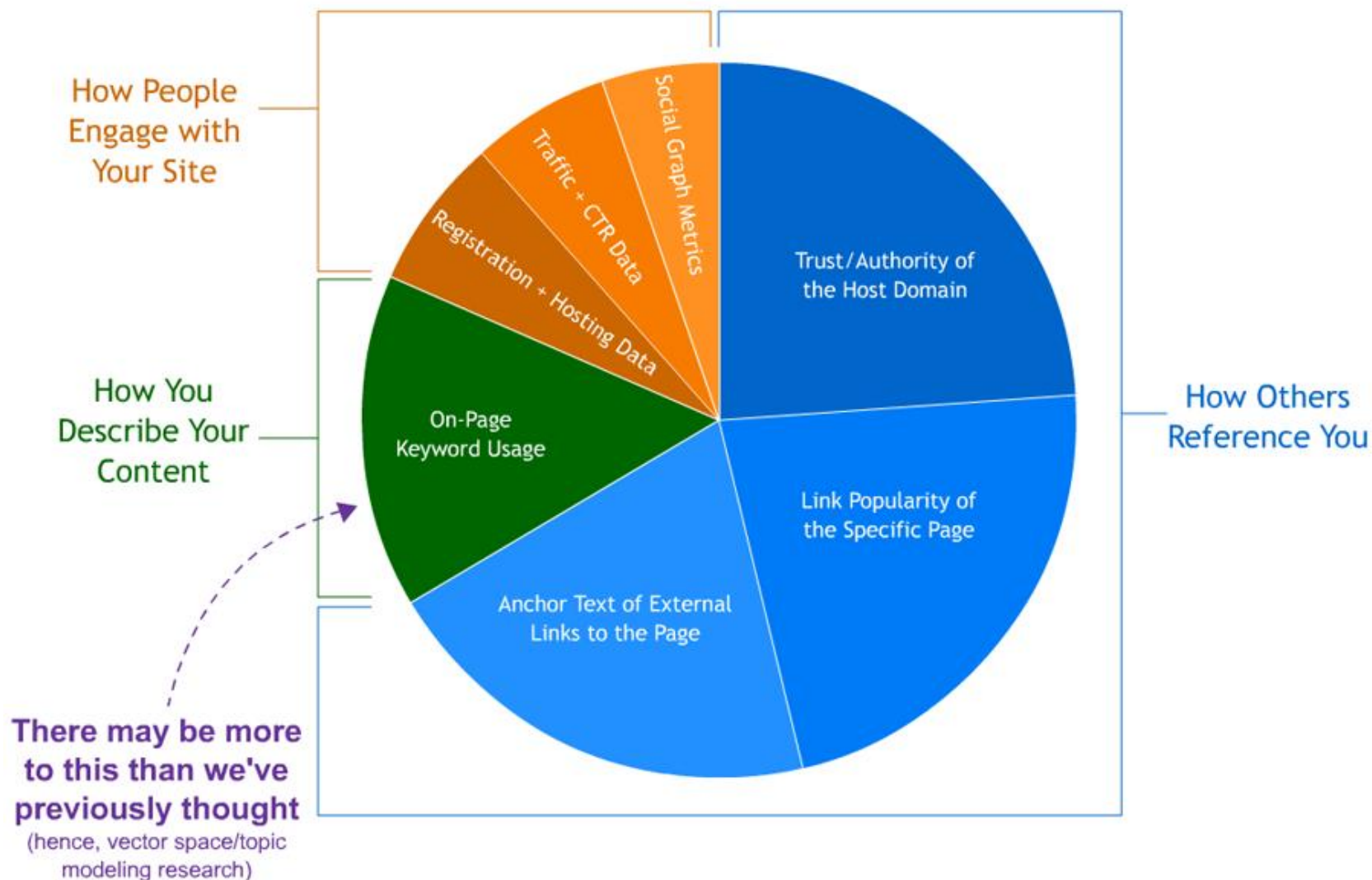
**Extra Time??**  
**Let's Do a Bonus Round!**

Bonus Question:

**How Much Does “Relevant” Content  
Matter for SEO?**

# Components of Google's Ranking Algorithm

(According to 72 SEOs Surveyed for SEOMoz's Biennial Search Ranking Factors)



## Search Query: **Batman**

The Batman, a masked hero in the tradition of Zorro and The Scarlet Pimpernel, first appeared in Detective Comics #27, dated May, 1939

### **Content A**

Banished from their primitive village, they set off on an epic journey through the ancient world

### **Content B**

## Solution: **Keyword Usage**

Since **Content A** contains the word "**Batman**" and **Content B** does not, the engine can easily choose which one to rank.

## Search Query: Chief Wiggum

Clumsily scraping up the remains of the dried Squishee from the floor, Wiggum hastily shoved them into his mouth.

### Content A

If you chose to do that, first let him know that you have spoken to the chief.

### Content B

## Solution: TF\*IDF

The search engine can use TF\*IDF (Term Frequency x Inverse Document Frequency) to determine that "Wiggum" is a much less common word than "chief" and thus, Content A is more relevant to the query than Content B.

**NOTE:** This example also does a good job of showing the inherent weakness of a metric like keyword density.

## Search Query: **Superman**

"The strength of a Superman would be required," noted the foreman. Work is ongoing to determine a solution.

### **Content A**

Ducking into a Daily Planet bathroom, Clark Kent, mild-mannered reporter, transforms into Superman .

### **Content B**

## **Solution: Co-Occurrence**

Using co-occurrence, the engine can determine that phrases like "Daily Planet," and "Clark Kent," frequently appear with "**Superman**" and thus, **Content B** is more relevant than **Content A**.

## Search Query: Pianist

Dropping his meeting notes at the door, he jiggled the keys into the lock but found it wouldn't budge.

**Content A**

Her hands mercilessly pounded the keys, notes cascading into the surrounding stairway.

**Content B**

## Solution: Topic Modeling

As humans reading both sentences, we can infer that **Content B** is obviously about the musical instrument - a piano - and the woman playing it. But a search engine armed with only the methods we described above will struggle since both sentences use the words "keys" and "notes," some of the only clues to the puzzle.

**NOTE:** We were excited to see that our LDA modeling tool correctly scored B higher than A :-)

## Methodology: LDA (Latent Dirichlet Allocation)

- Build an LDA model based on the English language Wikipedia dataset (8mil+ pages)
- Generate scores for top 10 rankings across several thousand search results
- Look at correlation of search rankings with scores (in process)

## Simplified LDA Formula

$$P(z = t|w) \propto (\alpha_t + n_{t|d}) \frac{\beta + n_{w|t}}{\beta V + n_{\cdot|t}}$$

Chance of word is because of a topic

=

(Number of times the document already uses that topic a lot)

X

(Number of times that word has been in that topic)

# Tool to Test it Out

Topics Tool

Relevance of "www.seomoz.org" to "seo": **48%**

COMPUTE RELEVANCE

Query:

seo

Document:

seo search engine optimization tools software rank better seo blog pro tour seo tools youmoz learn seo seo jobs register member login account email password remember forgot password new seomoz research competitor links open site explorer open site explorer compares site backlinks top pages other metrics using seomoz web index better results less time targeting best link opportunities metrics show seo success boss clients check seo mozbar free seo toolbar seomoz seo toolbars provide easy access most powerful seo tools data surf web quickly diagnose seo problems opportunities without opening new page interrupting work flow check learn seo beginner guide seo free beginner guide seo covers all concepts need know stellar seo full easy understand illustrations anyone learn seo send guide clients coworkers maximize seo effectiveness check latest blog reputation management seo advanced tactics last week visited milan italy thanks generosity state department marco montemagno organizer social media week first off impressed state dept formal program encourage digital entrepreneurship promotion internet posted randfish comments why most conference presentations suck tend towards uncontroversial write haven author many posts caused debate enough need speak most conference presentations suck posted willcritchlow comments five ways social profiles seo whiteboard friday social media becoming important days how else tri weekly fix xkcd delivered many people know marketing benefits social media profiles sites like facebook twitter make significant difference seo campaign week rand shows five great ideas using sites help seo strategy posted aaron wheeler comments gimme blog fresh seo industry jobs director business development hiring intelligent aggressive business developers responsible introducing new products new markets successful candidate identify penetrate cons posted covariojobs covario covario inc

(OR specify a URL: )

COMPUTE RELEVANCE

This computes cosine similarity between topics for a keyword and topics for a page. The topics were computed with LDA on 8 million documents.

<http://www.seomoz.org/labs/lda>

# Tool to Test it Out

Topics Tool

Relevance of "www.seomoz.org" to "seo": **48%**

COMPUTE RELEVANCE

Query:

seo

Document:

seo search engine optimization tools software rank better seo blog pro tour seo tools youmoz learn seo seo jobs register member login account email password remember forgot password new seomoz research competitor links open site explorer open site explorer compares site backlinks top pages other metrics using seomoz web index better results less time targeting best link opportunities metrics show seo success boss clients check seo mozbar free seo toolbar seomoz seo toolbars provide easy access most powerful seo tools data surf web quickly diagnose seo problems opportunities without opening new page interrupting work flow check learn seo beginner guide seo free beginner guide seo covers all concepts need know stellar seo full easy understand illustrations anyone learn seo send guide clients coworkers maximize seo effectiveness check latest blog reputation management seo advanced tactics last week visited milan italy thanks generosity state department marco montemagno organizer social media week first off impressed state dept formal program encourage digital entrepreneurship promotion internet posted randfish comments why most conference presentations suck tend towards uncontroversial write haven author many posts caused debate enough need speak most conference presentations suck posted willcritchlow comments five ways social profiles seo whiteboard friday social media becoming important days how else tri weekly fix xkcd delivered many people know marketing benefits social media profiles sites like facebook twitter make significant difference seo campaign week rand shows five great ideas using sites help seo strategy posted aaron wheeler comments gimme blog fresh seo industry jobs director business development hiring intelligent aggressive business developers responsible introducing new products new markets successful candidate identify penetrate cons posted covariojobs covario covario inc

(OR specify a URL: )

COMPUTE RELEVANCE

This computes cosine similarity between topics for a keyword and topics for a page. The topics were computed with LDA on 8 million documents.

Topics Tool

Relevance of "en.wikipedia.org/wiki/Search\_engine\_optimization" to "seo": **77%**

COMPUTE RELEVANCE

Query:

seo

Document:

search engine optimization wikipedia free encyclopedia search engine optimization wikipedia free encyclopedia jump navigation search typical search engine results page internet marketing display advertising mail marketing mail marketing software interactive advertising social media optimization web analytics cost per impression affiliate marketing cost per action contextual advertising revenue sharing search engine marketing search engine optimization pay per click advertising paid inclusion search analytics mobile advertising box view talk edit search engine optimization seo process improving visibility web site web page search engines via natural paid organic algorithmic search results other forms search engine marketing sem target paid listings general earlier higher page frequently site appears search results list visitors receive search engine seo may target different kinds search including image search local search video search industry specific vertical search engines gives web site web presence internet marketing strategy seo considers how search engines work people search optimizing website may involve editing content html associated coding both increase relevance specific keywords remove barriers indexing activities search engines promoting site increase number backlinks inbound links another seo tactic acronym seo refer search engine optimizers term adopted industry consultants carry optimization projects behalf clients employees perform seo services house search engine optimizers may offer seo stand alone service broader marketing campaign effective seo may require changes html source code site seo tactics may incorporated web site development design term search engine friendly may used describe web site designs menus content management systems images videos shopping carts other elements optimized purpose search engine exposure another class techniques known black hat seo

(OR specify a URL: )

COMPUTE RELEVANCE

<http://www.seomoz.org/labs/lda>

## Tool to Test it Out

## Topics Tool

Relevance of "www.seomoz.org" to "seo": 48%

## COMPUTE RELEVANCE

Query:

seo

Document:

seo search engine optimization tools software rank better seo blog pro tour seo tools youmoz learn seo seo jobs register member login account email password remember forgot password new seomoz research competitor links open site explorer open site explorer compares site backlinks top pages other metrics using seomoz web index better results less time targeting best link opportunities metrics show seo success boss clients che provide easy access most powerful seo tool opportunities without opening new page inte free beginner guide seo covers all concepts anyone learn seo send guide clients cowork reputation management seo advanced tactic department marco montemagno organizer s program encourage digital entrepreneurship conference presentations suck tend towards debate enough need speak most conference ways social profiles seo whiteboard friday se fix.xkcd delivered many people know market make significant difference seo campaign w strategy posted aaron wheeler comments gi development hiring intelligent aggressive bu new markets successful candidate identify p

We might need to work the “relevance” of our content

(OR specify a URL:

## COMPUTE RELEVANCE

This computes cosine similarity between topics for a keyword and topics for a page. The topics were computed with LDA on 8 million documents.

## Topics Tool

Relevance of "en.wikipedia.org/wiki/Search\_engine\_optimization" to "seo": 77%

## COMPUTE RELEVANCE

Query:

seo

Document

# Need to Relevance" Content

(OR specify a URL:

## COMPUTE RELEVANCE

# Interesting Takeaways

- There may be more to “on-page” optimization than just using target keywords in the right places / ways
- Search engines keep saying “make relevant content”
  - perhaps we can get more scientific and precise about what “relevant” means
- Our LDA topic modeling work is still in its infancy.  
Expect more data, correlations, etc. in weeks to come.



## Q+A

Rand Fishkin, CEO & Co-Founder, SEOMoz

- Twitter: @randfish
- Blog: [www.seomoz.org/blog](http://www.seomoz.org/blog)
- Email: [rand@seomoz.org](mailto:rand@seomoz.org)

