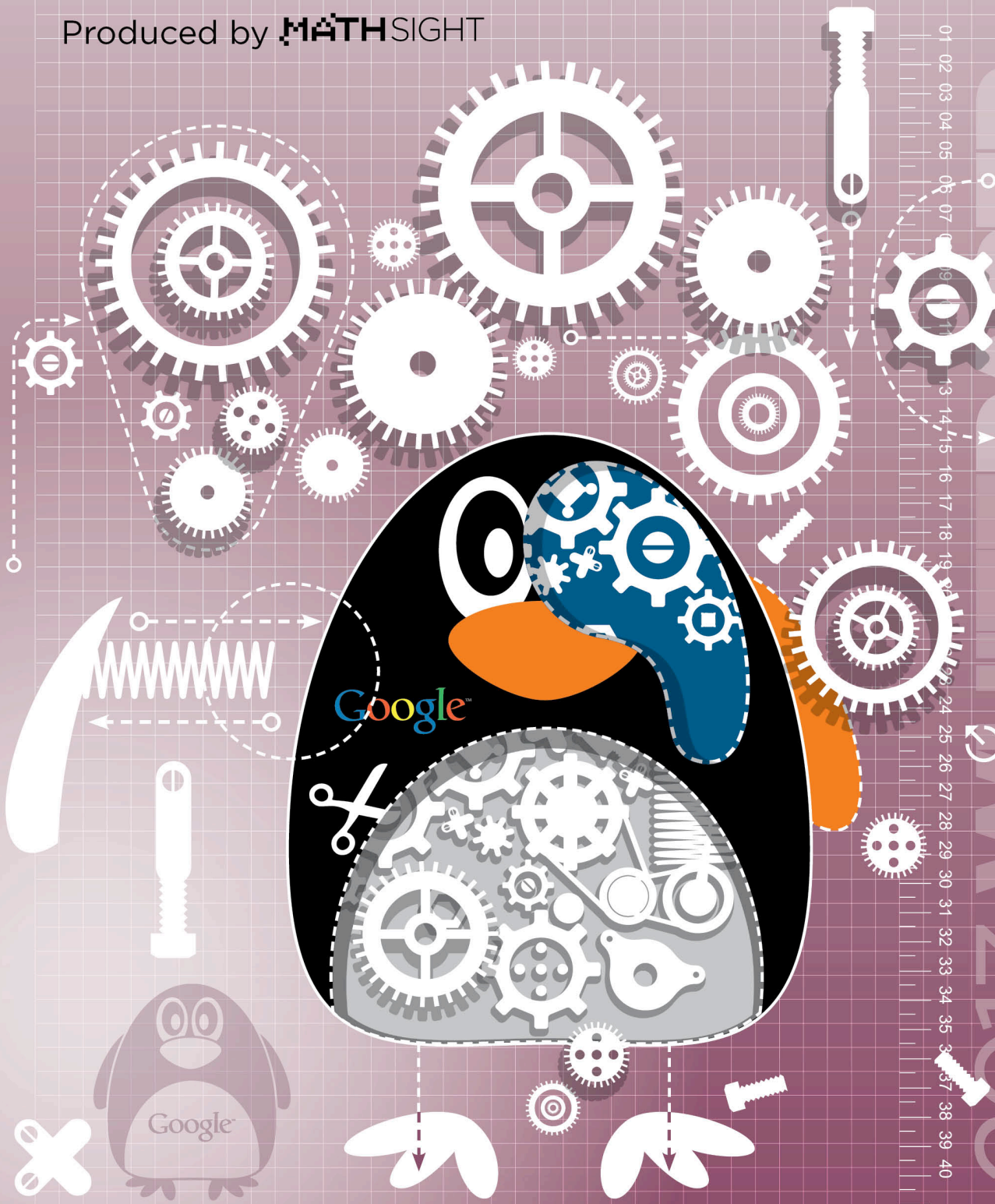
DECONSTRUCTING GOOGLE'S PENGUIN 2.0

Produced by MATHSIGHT



34 Paradise Road | Richmond Upon Thames | United Kingdom | TW9 1SE **TEL//** +44 (0) 844 264 2960 **VISIT//** mathsight.org

White Paper:
Deconstructing Google's Penguin 2.0

Produced by MathSight

This paper identifies how shifts in traffic sourced by Google's search engine can be related to the structural and content-based features of a company's web pages.

We performed this analysis in order to extract potentially useful insights for the following groups:

A. Marketing agencies
B. The wider online SEO community
C. Online businesses
D. Those with an interest in big data and analytics

Introduction

We decided to apply our machine learning led predictive SEO models to deconstruct the Google Penguin 2.0 algorithm update, rolled out on 19th May 2013.

Although it is near impossible to reverse engineer a complete search engine algorithm such as Google's, it ***is*** possible to show the potential causes of any change in algorithm methods when it occurs. We look for a step change in a pattern that could be an underlying increase or decrease in actual Google-sourced traffic as a result of an algorithm alteration, such as the recent Penguin 2.0 update.

What we did

Once the Google search traffic dataset for our chosen group of web domains had been obtained from website analytics, de-seasonalised and filtered, the first step in the reverse engineering process was to confirm that a change in traffic did indeed take place. This was done using signal processing techniques, a best practice in the oil and gas exploration industry, to detect the likely point of change in the noisy data.

Following this, we gathered a wide range of standard SEO features from the pages (title character length, number of meta description words, readability and so on) within the domain. Finally we applied a variety of statistical methods to identify those features that were rewarded or penalised in terms of their Google search traffic after the likely algorithm update time.

Our results showed, with some statistical confidence of around 90-95%, that the main areas within HTML that Google has probably targeted with this change were:

- Main body text
- Hyperlinks
- Anchor text (clickable text in the hyperlink)
- Meta description text

Methods

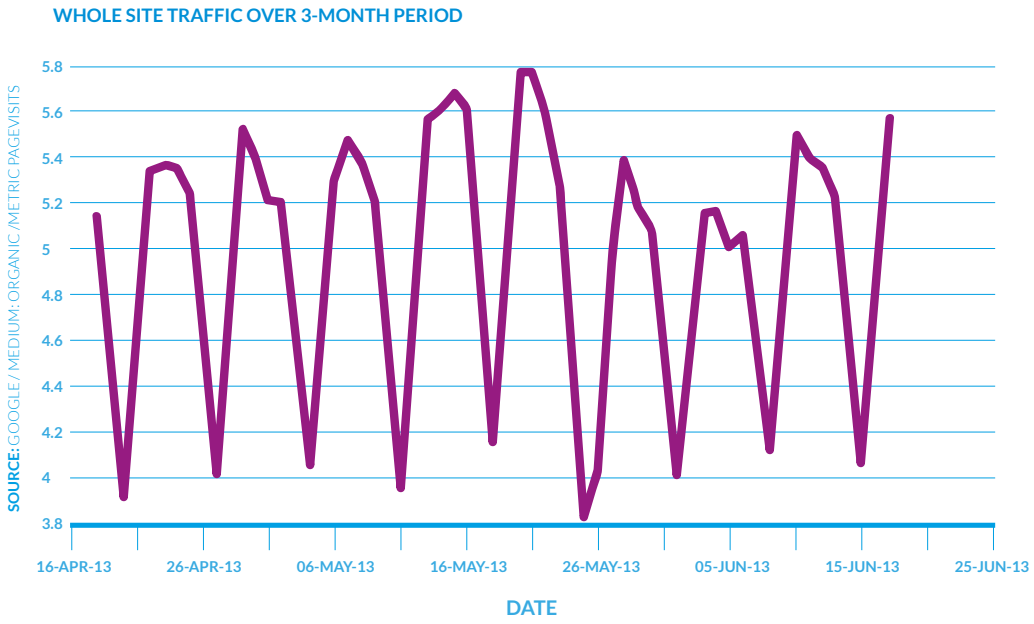
Data collection

Websites from eight business categories as follows were used for the purposes of this study, in order to create a well-rounded dataset:

- Online retailers including the travel, gifts, mobile apps and jewellery sectors;
- Corporate B2B companies including business awards, advertising and PR,

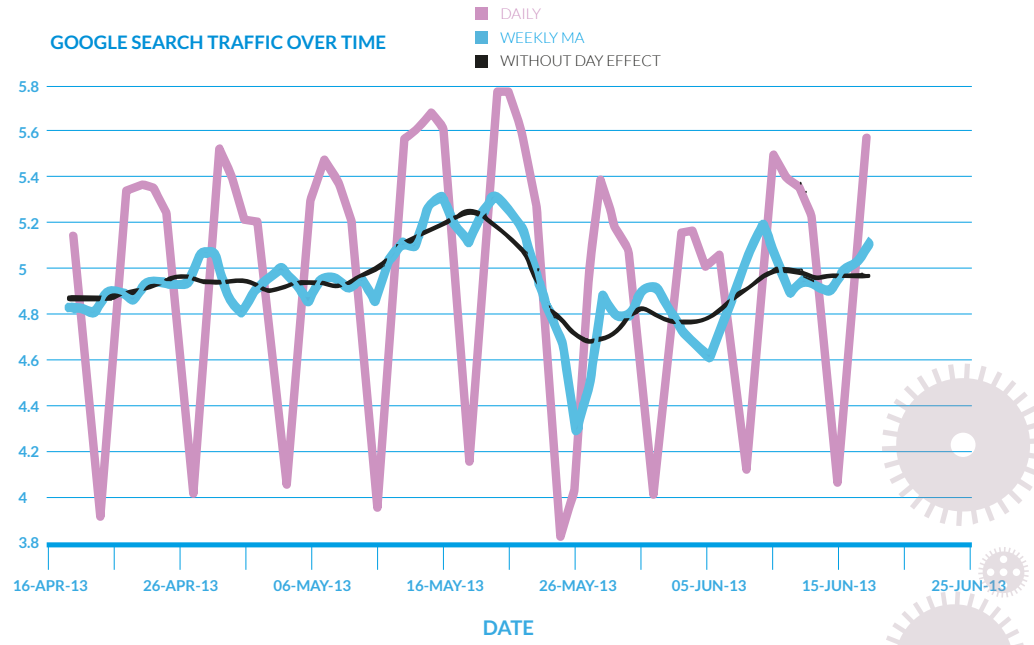
HTML file contents were first gathered by our in-house web crawler, which scanned the sites in-depth, for structural and content-based 'features'.

Daily website analytics (page view) data was also imported for each domain above, spanning a two-month period, from 11 April 2013 through to 11 June 2013. This period afforded a reasonable window around the time that Google had announced the 'Penguin 2.0' algorithm update.



Cleansing and exploration of the data

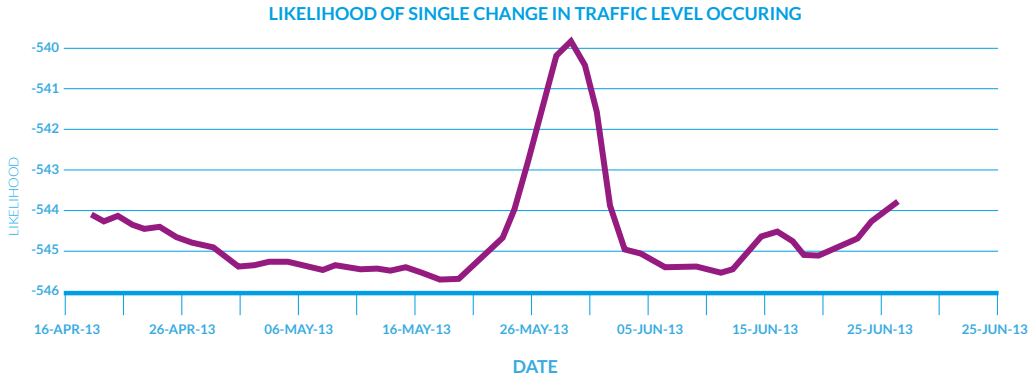
The traffic data, in time series form for a single domain were first smoothed using moving average and then seasonality variation removal, to reduce the effect of a repeated site usage pattern across the week (e.g. reduced visits on the weekend). This is slightly more insightful than both the moving average and the raw traffic numbers, as abrupt changes are clearly defined yet separated from any cyclical variation.

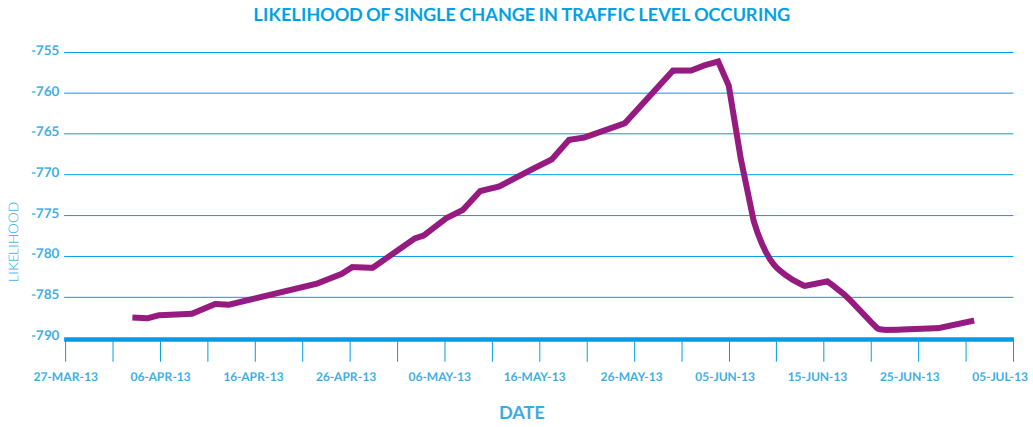


Using this cleaned traffic data, a change point detection algorithm was deployed in order to detect the most likely timing of a change in traffic levels over the period in question. For each domain, this gave a probabilistic confirmation that a change had indeed occurred at the period in question, rather than simply a series of fluctuations due to 'noise' in the traffic data.

Using this method, of our eight site categories, 3 were selected (numbers 2, 5 and 7) as they each showed clear evidence (like the pattern in the upper graph) that a change in daily visitor traffic had occurred.

The lower graph shown to the left here (for the 8th category in our list) shows that it is unlikely that such a change occurred on the 19th May, rather that it took place later, in early June.





Simple Statistical modelling

Following this confirmation that a change had indeed occurred, all the html pages of the chosen domains were classified as either 'winner' or 'loser' pages with respect to their mean traffic levels pre- or post- the alleged algorithm update. The traffic values were normalised, i.e. adjusted so that difference between 'before' and 'after' algorithm change traffic level were scaled correctly.

Then, the effect of html **page features** on traffic difference was analysed using the Analysis of Variance (ANOVA) method. This enabled us to see if there was any statistically significant relationship between feature metrics and daily search traffic variation.

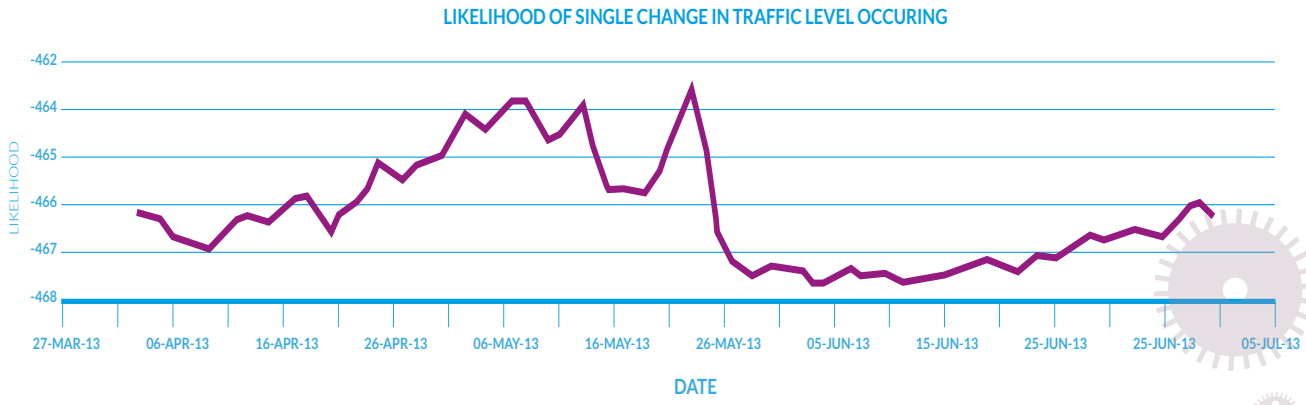
Results

The results below represent a selection of Penguin 2.0 case studies within the overall data set.

Site A: An online luxury jewellery supplier

CUSTOMER QUERY: They wanted to understand why their daily traffic jumped up on 19th May.

We found that the average visitor traffic before 19th May was 33.97 per day, while afterwards it was up to 59.66 per day (an increase of 56.31%). There was a clear confirmation statistically that a change in traffic took place (see spike in chart below).



OUR ANSWER:

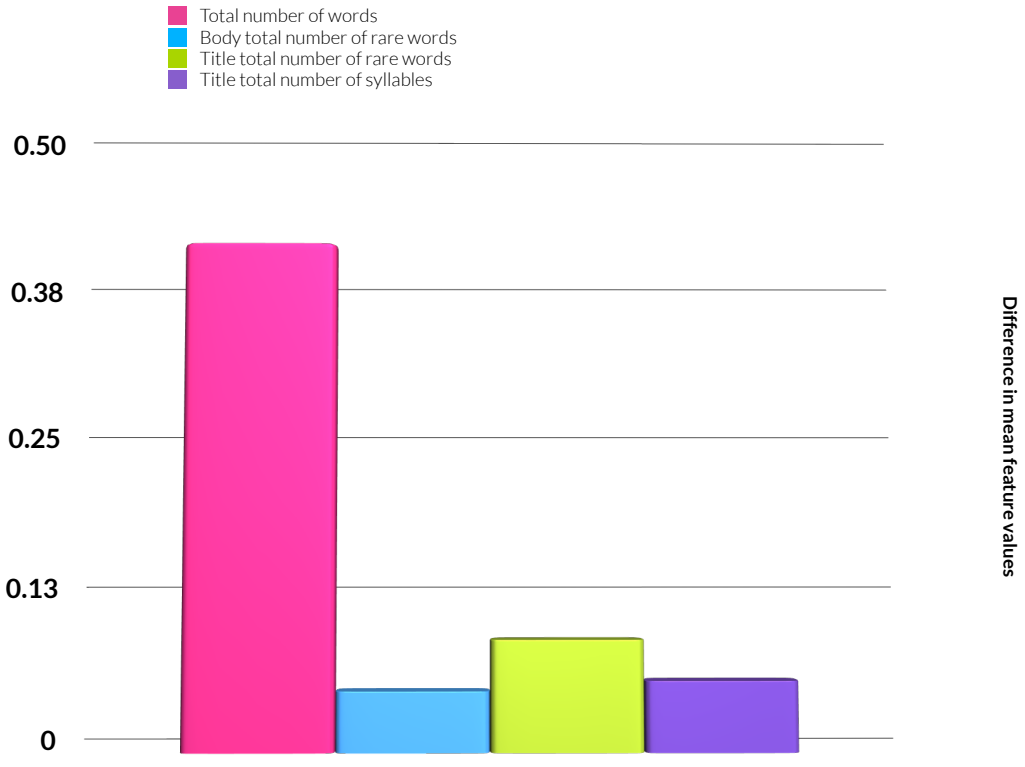
Firstly, Google's algorithm seemed to have become attentive to the nature of the **title tags** in their html pages; as this seems to have had an effect on the traffic level after the change. These aspects were found to be significant, shown here in order of importance:

1. The number of syllables per title
2. The number of 'rare' words (i.e. those not in the list of 5,000 most commonly used English language words) present in the title
3. The title length, in characters (less significantly)

Secondly, the nature of overall **html body text** has had an impact; in this order:

1. The number of words and characters in the document
2. The ratio of 'rare' to commonly used words

These were rewarded in the following fashion:



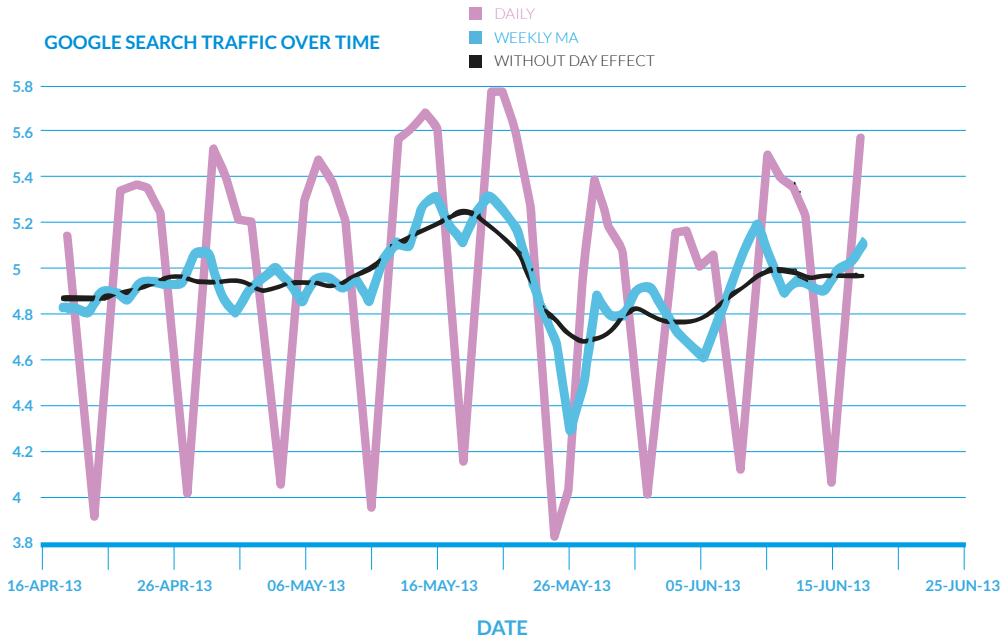
Thirdly, but notably less significantly, the following two features had some influence:

1. The number of hyperlinks present
2. The meta description character length

Site B: A mobile application vendor

Site type: Promotional and catalogue of products

CUSTOMER QUERY: The ecommerce team wanted to understand why their visitor traffic fluctuated slightly around 19th May.



We found that the average visitor traffic before 19th May was 49,534.53, it initially rose and then afterwards it had settled, overall having dropped slightly to 49,271.79 (a -0.53% change)

OUR ANSWER:

This site has much higher traffic volumes, and many more pages so the data extracted was far richer than that obtained from site A. Nevertheless, similarities between this website and site A quickly became apparent, such that there seemed to be a focus on the **overall html page body text** content, **meta -descriptions** as well as **hyperlinks**.

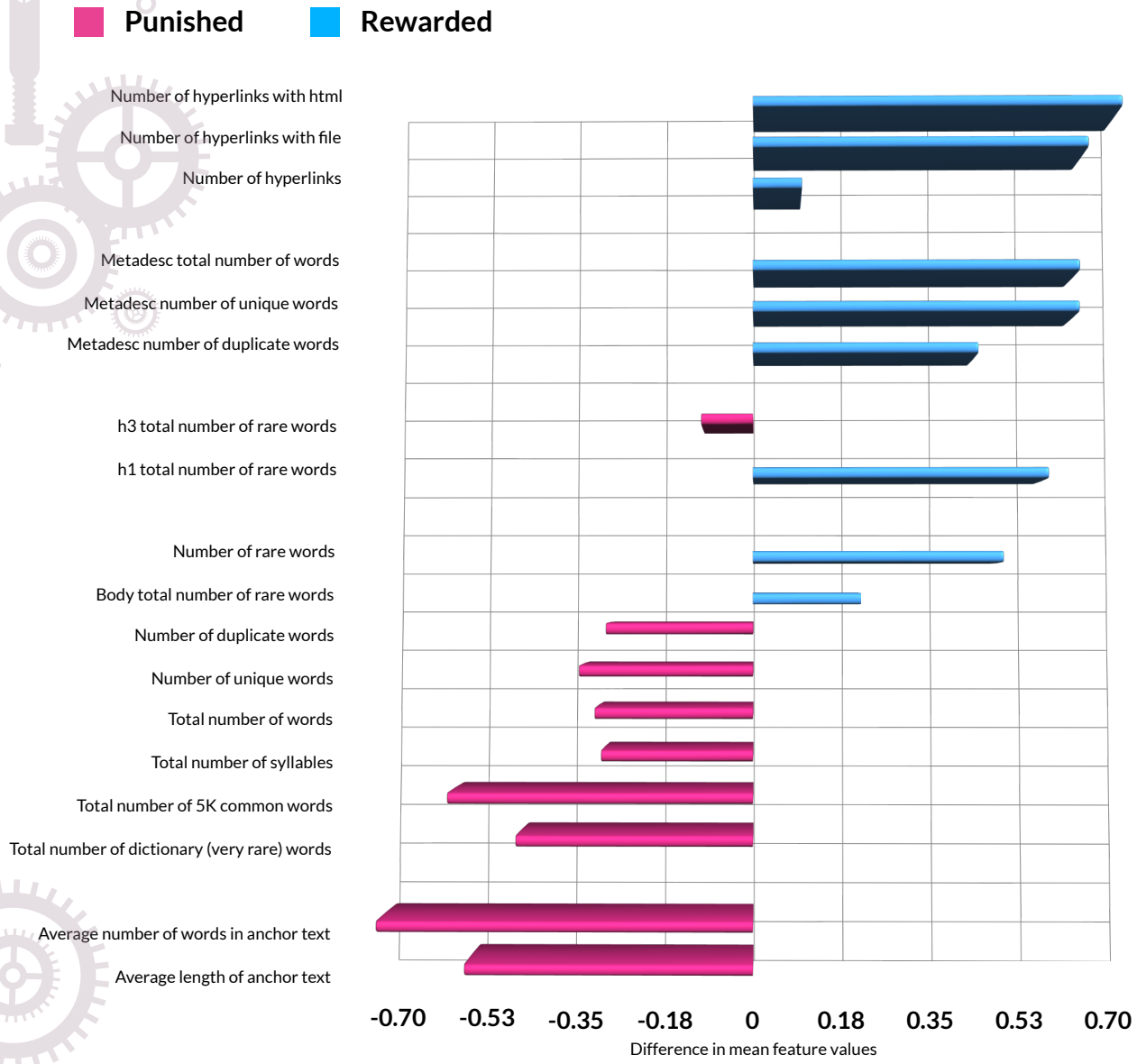
That is to say;

- The total number of words; the number of syllables in those words; the ratio of rare and extremely rare to those commonly used; the number of difficult words; the number of sentences and the ratio of unique to duplicate words. Indeed, **text readability** (which is a combination of almost all the other word-related features) emerged as slightly significant.
- The **number of hyperlinks** and those linking to files or html files specifically.
- The **meta description;** here the number of words and ratio of unique to duplicate words.

However, the difference here was that there was no hint of title length or content being significant. Rather, there was also a focus on two other areas (listed in order of importance):

- The length in characters and number of words in **anchor text.**
- The number of **rare words in the headers.**

Thanks to the increased dataset for this client, by comparing pages that 'won' post algorithm change with those that lost, we were able to observe that some features were quite substantially rewarded as they increased in value; whilst others were punished. See below:



Site C: An online watch vendor

Site type: extensive online catalogue with product photographs

CUSTOMER QUERY: They wanted to understand a recent increase in daily traffic sourced from Google search, which started on 19th May.

The site's daily average traffic before was 429.85, and after this date about 503.0, thereby showing a 16% increase.

OUR ANSWER:

Again, a relatively rich set of data, from which certain significant features emerged, to start with those that have already appeared in one or more of the previous two sites examined:

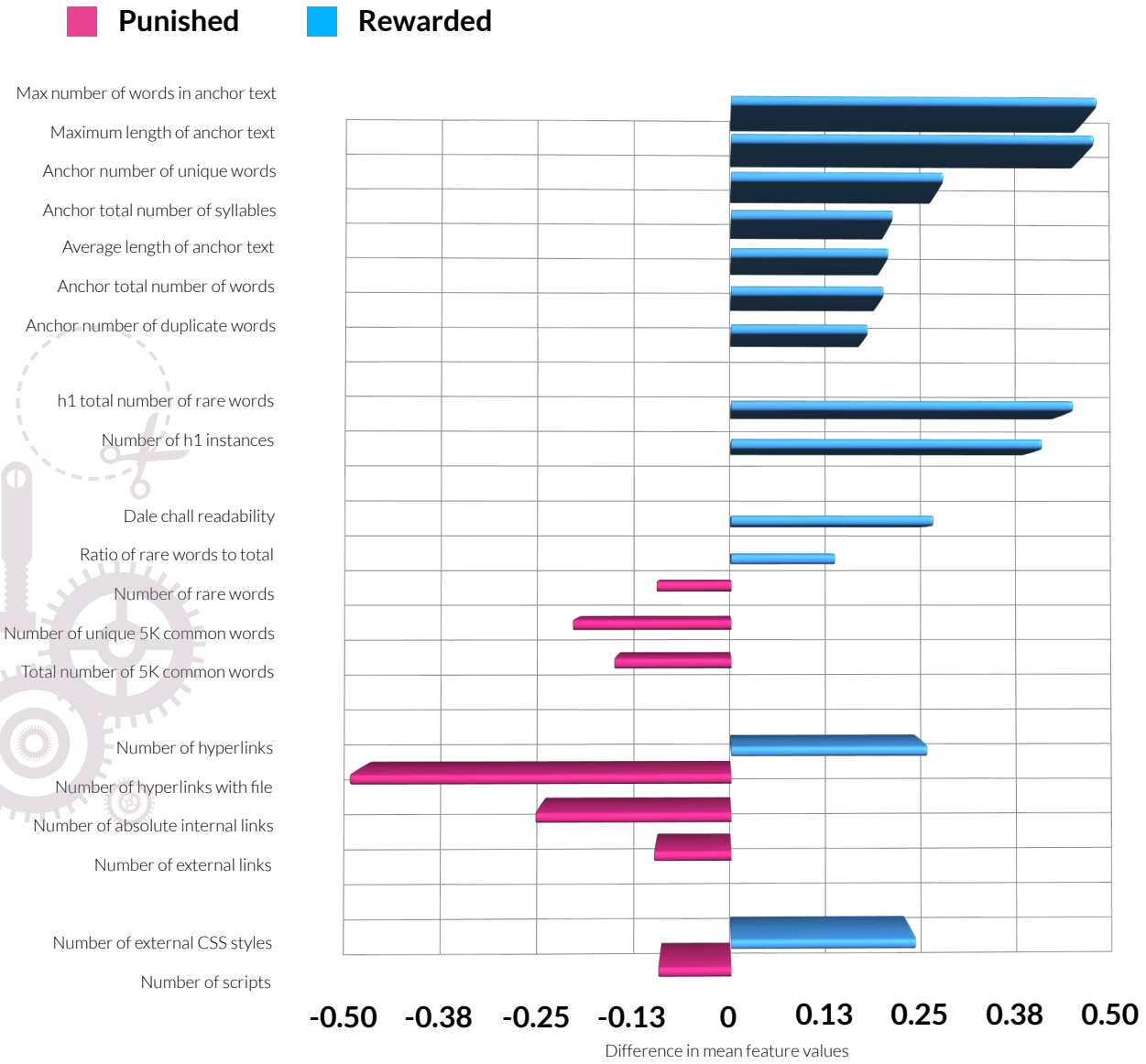
- **Number of hyperlinks** (and whether they link to a file)
- **Anchor text** (length and number of words, number of unique and duplicate words, number of syllables)
- **Body text** (ratio of commonplace to rarer words, ratio of unique to duplicate words, readability)
- **Headers** (and whether or not they contain rare words)
- To a certain extent; **meta description** character length

Additionally, the following features were significant:

- **Number of external CSS styles** (cascading style sheets - manages appearance of the website)
- **Number of scripts**
- Number of external and absolute internal links

In terms of rewarded and / or punished features - with this client a large quantity of anchor text (regardless of whether duplicate words or not) was heavily rewarded.

As was the 'reading ease level' of the body text and headers (i.e. the harder/ more complex it was, the better it was to Google's algorithm). This relates to the number of rare words. The full breakdown is shown in the following graph.



Conclusions

Our analysis suggests that there is perhaps significant, positive association between a site's search traffic sourced visitor levels, ***augmenting*** the values of certain features and reducing the values of others, based on two months' worth of traffic data (observed 10 days, one month and 6 weeks before and after the change).

Overall, there was a large variation between the types of rewarded features for the different websites analysed. This would suggest that any advice given would be most effective if tailored to the individual domain, or type of domain if domains can be grouped or clustered into types. For example, with site B the presence of **anchor text** was seemingly punished as a feature, whereas for site C it was heavily rewarded.

However, even with all the above variation - looking at the two larger and more popular (traffic-wise) sites that clearly exhibited effects of Penguin 2.0, we can draw the following conclusions:

- In **body text**, rare words are good and generally rewarded - i.e. those that are not in the list of 5,000 most common words in the English language. So it is a good idea to raise the writing level of the page copy (i.e. aim for higher Dale-Chall readability scores).

- Use of **headings** will be rewarded; it is also advantageous to use words that are less commonplace here.

- The **number of hyperlinks** present appears to have been rewarded – i.e. the more hyperlinks the greater the increase in traffic (in some cases), although perhaps this is too vague to take any action upon. Of those hyperlinks, there was no bias towards external or internal links however.

- Finally, depending on the type of site, and based on our limited survey the presence and increased character length of **meta descriptions** and the increased quantity of words in **anchor text** are now slightly more rewarded than previously.

In addition to the insights gained on deconstructing Penguin 2.0, we can now use the models to evaluate the inbound link profiles of sites that may have been affected by the latest algorithm update. For example, the models may be applied to a site's inbound link profile when trying to decide which links to disavow when faced with a search engine manual penalty notice (ie apparent abnormal inbound link activity), or subsequent loss of traffic in response to a search engine's algorithm update.

Future work

Our research into analysing search algorithm updates are continuing as our data partnership community grows. Like search engines we are adding the number of features and searching for the "sweet spots" of site optimisation. The benefit of using machine learning is that our data modelling of algorithm updates is dynamic and therefore stays up to date with the constantly improved search engines. Thus MathSight's models evolve with the search engines.

Further research papers will be released to provide insights into Penguin and Panda going forward.

About MathSight

MathSight was **launched in March 2013**; it demystifies the search engine algorithms using machine learning and big data.

The platform analyses both the qualitative and stylistic aspects of content, web design, and site architecture, their inter-relationships, traffic data and other key performance indicators. This enables MathSight to determine the cause of changes in search engine traffic, be it a change in the algorithm, or the SEO (onsite and offsite) of a client or competitors. These insights are currently available for integration into bespoke and best in class, enterprise level, SEO tools.

For more information visit: MathSight.org