My previous experiences developing web-scale systems at Google are fundamental to my development as a computer scientist and in motivating my desire to pursue research. Although research at Google takes a somewhat different form than in academia, in particular with different success criteria, I believe there are many similarities between developing new products from scratch and formulating an academic research agenda. Furthermore, the scale at which Google operates has come to define state-of-the-art in distributed systems and having firsthand experience in building such large-scale systems gives me a unique and valuable perspective. In this essay, I describe a project that that I was cofounder of and worked on from 2008 - 2010, which has all of these components.

Search advertising is Google's core business. Users come to Google, often many times a day, to search the internet for the most relevant documents pertaining to their information need. Beyond fulfilling users' needs by giving them relevant documents, each query also represents a business opportunity to present users with relevant commercial information. Google's Adwords product exposes this opportunity to third-party advertisers who participate in a real-time auction in order to show their message in a special section of the page. Advertising on Google is big business and as such a huge amount of effort is invested in improving advertising campaigns and the AdWords product that is used to deliver them.

Despite the huge success of AdWords, a major difficulty faced by advertisers is the necessity to maintain huge lists of keywords on which to advertise. Fundamentally, Ad*Words* relies on external advertisers to pick the most relevant keywords for their website and maintain this list as the website changes over time. For large websites this is a monumental task, requiring significant manual effort. Furthermore, it is impossible to predict a priori every query that a user may formulate, so inevitably there will be gaps in query stream coverage for any manually chosen set. The end result is less effective advertising campaigns and lost opportunity for Google and advertisers.

Beginning in 2008, I, along with a small number of other senior engineers, started a project that rethinks how this paradigm works. In particular, we focused on completely automating the document to query matching that is currently done by hand and thus delivering a new product that is devoid of maintaining massive lists of keywords. This was, and still is, a huge undertaking and thus in order to realize our goals, we relied on many technologies already developed inside the company. For example, Google already has a state-of-the-art IR system used to match queries to documents and we utilized many components of this in developing our new product.

My role as a founding member was to scale up the team and deliver an initial prototype of the technology in action. Technically, I had a large role in building all of the key components which included a large scale data processing pipeline and a new serving system that integrated directly into the Google serving stack. The pipeline was essentially responsible for computing the page to query mapping, joining it with advertiser provided information (e.g. bids) and outputing the data in a form consumed by our serving system. The serving system we also built from scratch to satisfy the tight constraint that return candidate ads on the

order of 10s of milleseconds. Even though this was a prototype, we had paying advertisers from a very early stage and thus this was a production-ready system with high uptime requirements.

In addition to the systems work, I was also responsible for measuring and improving the quality of the ads that we delivered. The first order of business was to define the metrics by which to measure our quality and thus success. Many of the existing metrics and systems for measuring them were not applicable since they assumed the existence of keywords which were no longer present in our system. Therefore, we developed new ones and I built them into a dashboard which highlighted our core metrics and thus the progress of the team – checking it became a daily ritual for every engineer. One of the major deficiencies of our early system was that it could not take advantage of some of the large-scale machine learning systems in place that improve the quality of Google's ads over time. I led the effort to understand how this system interacted with our new ads and then developed a novel extension which largely corrected the issue. The net effect was one of the most substantial quality wins the project has seen to date.

My involvement in the project stopped when I enrolled in CMU in 2010. However, since we were able to build a strong team and momentum, the project is continuing strong today. Recently, it graudated from the beta stage, opening up to all advertisers [5]. Deep changes to a product as successful as AdWords take a long time to take root; however, significant progress has been made and with continued improvements in quality and usability, adoption will continue to increase.

For more information, see the first public announcement [4], as well as the patents that were filed as part of the project [1, 2, 3].

References

[1] L. J. Brunsman, S. Rajaraman, P. Deshwal, M. Wytock, and S. Kates. *Automatic abstracted creative generation from a web site.* U.S. Patent (filed), 2010.

[2] L. J. Brunsman, S. Rajaraman, P. Deshwal, M. Wytock, and S. Kates. *Automatic ad creative generation.* U.S. Patent (filed), 2010.

[3] L. J. Brunsman, S. Rajaraman, P. Deshwal, M. Wytock, and S. Kates. *Targeting content without keywords.* U.S. Patent (filed), 2010.

[4] B. Gultekin. *Introducing Dynamic Search Ads (beta).* Google AdWords blog, 2011. `http://adwords.blogspot.com/2011/10/introducing-dynamic-search-ads-beta.html`.

[5] T. Jain. *Boost your results with Dynamic Search Ads, now available to all.* Google AdWords blog, 2012. `http://adwords.blogspot.com/2012/10/boost-your-results-with-dynamic-search.html`.