Beyond algorithms: Optimizing the search experience

Making search smarter through better human-computer interaction.

By [Daniel Tunkelang](https://www.oreilly.com/people/e9a1c-daniel-tunkelang)

October 13, 2015

A woman posed with a Hollerith pantograph: the keyboard is for the 1920 population card, and a 1940 census form.(source: [U.S. National Archives and Records Administration on Wikimedia Commons](https://commons.wikimedia.org/wiki/File:Card_puncher_-_NARA_-_513295.jpg))

We live in a golden age of algorithms. Even though we've had search engines, speech recognition, and computer vision systems for decades, only in the last several years have they become good enough to move out of the lab and into tools we use everyday — products like [Google Now](https://www.google.com/landing/now/), [Siri](http://www.apple.com/ios/siri/), and [Cortana](http://windows.microsoft.com/en-us/windows-10/getstarted-what-is-cortana).

But as good as our algorithms are, they could be improved by better human-computer interaction. In particular, most of today's search engines tend to optimize for one-off transactions rather than engage people in conversations to improve the search experience.

For the past decade, information scientists have been advocating that we involve people on a more meaningful level in the information-seeking process. In particular, [Gary Marchionini](http://ils.unc.edu/~march/), dean of Information and Library Science at the University of North Carolina at Chapel Hill, has advocated for "[human-computer information retrieval](https://en.wikipedia.org/wiki/Human%E2%80%93computer_information_retrieval)" systems that increase user responsibility and control.

This sounds like a great idea. But how do we do it?

**VIDEO COMPILATION**

**Treat search as a conversation: optimize for communication.**

Ever since Bill Maron and Lary Kuhns published their seminal paper [*On Relevance, Probabilistic Indexing and Information Retrieval*](http://ctp.di.fct.unl.pt/~jmag/wamse/papers/1960.On%20Relevance,%20Probabilistic%20Indexing%20and%20Information%20Retrieval.pdf) in 1960, search engines have tried to determine which documents in a collection are relevant to the user and rank results in descending order of relevance. Over five decades later, most of us take for granted that this is what search engines do.

But technology has outgrown this approach. A search engine can now respond to users with questions of its own. For example, a search for "truffles" on Amazon returns results, but also suggests "chocolate truffles" and "truffle oil" as alternative searches. Search engines can also interrupt users mid-query with suggestions and even instant answers (e.g., [Google Instant](https://googleblog.blogspot.com/2010/09/search-now-faster-than-speed-of-type.html)). The ideal search engine is a conversational partner that knows when an answer is sufficient and when it needs to ask the user for clarification.

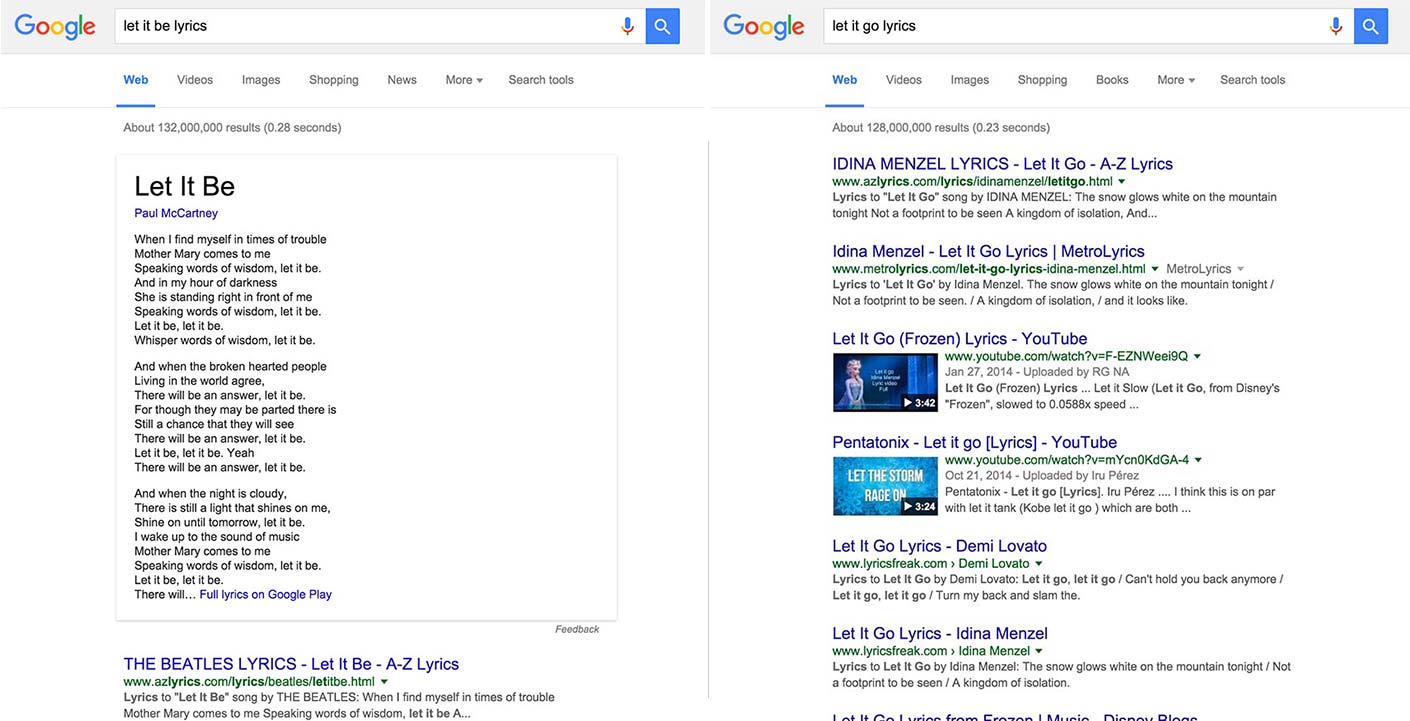
Embracing a conversational model changes the focus of search engine quality from relevance to communication. Traditional search quality metrics like [discounted cumulative gain](https://en.wikipedia.org/wiki/Discounted_cumulative_gain) assume a model where the search engine has only one shot to deliver the best results. However, today the user has the opportunity for further interaction, so we need to measure other factors, such as how well the user understands why the search engine is returning those particular results, how much effective control the search engine gives to users, and how effectively the search engine guides opportunities for further exploration. In other words: improving search quality means increasing transparency, control, and guidance for the user.

We already see baby steps in this direction: transparency through search snippets on Google and Bing, and control and guidance through [faceted search](https://en.wikipedia.org/wiki/Faceted_search) on most ecommerce sites. But there's so much more we can do.

**Be clever, but not so clever that you sacrifice predictability.**

A corollary of optimizing for communication is that our search engines shouldn't be so clever that they deliver unpredictable search experiences that undermine our muscle memory as users.

For example, as shown in the screenshot below, a search on Google for "[let it be lyrics](https://www.google.com/#q=let+it+be+lyrics)" returns the lyrics of the classic Beatles song at the top of the search results. But a search for "[let it go lyrics](https://www.google.com/#q=let+it+go+lyrics)" doesn't return such an interface element, despite the immense popularity of this Disney song and the wide availability of its lyrics. Perhaps there's a good reason for this difference, but it's hardly obvious to the user. Interface elements should not appear and disappear in unpredictable ways, even if the decisions reflect a principled optimization.

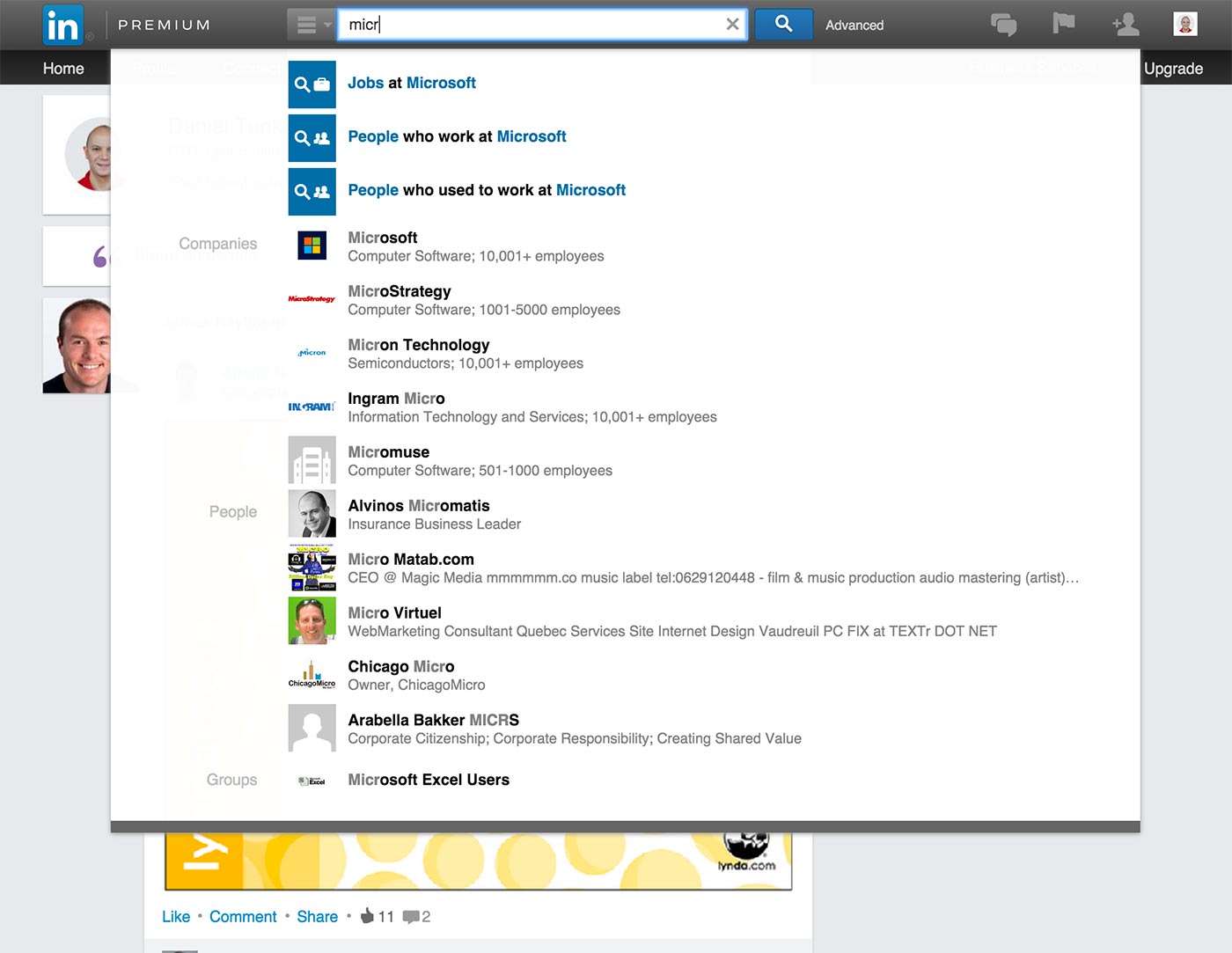
Screenshot of the "Let it Be" and "Let it Go" lyrics search results on Google.

I'm not suggesting we completely dumb down our algorithms. Rather, we need to treat predictability as an explicit factor in our decisions. That's a key part of optimizing for communication.

**Help users ask good questions, rather than attempt to answer bad ones.**

The traditional approach to search relevance is to take a query, retrieve a bunch of results, score them, and then return them as a ranked list. Modern search engines implement this approach using [machine-learned ranking](https://en.wikipedia.org/wiki/Learning_to_rank). And certainly, it's important to implement robust result ranking, but it's even better to address search quality earlier in the process, by helping users ask better questions.

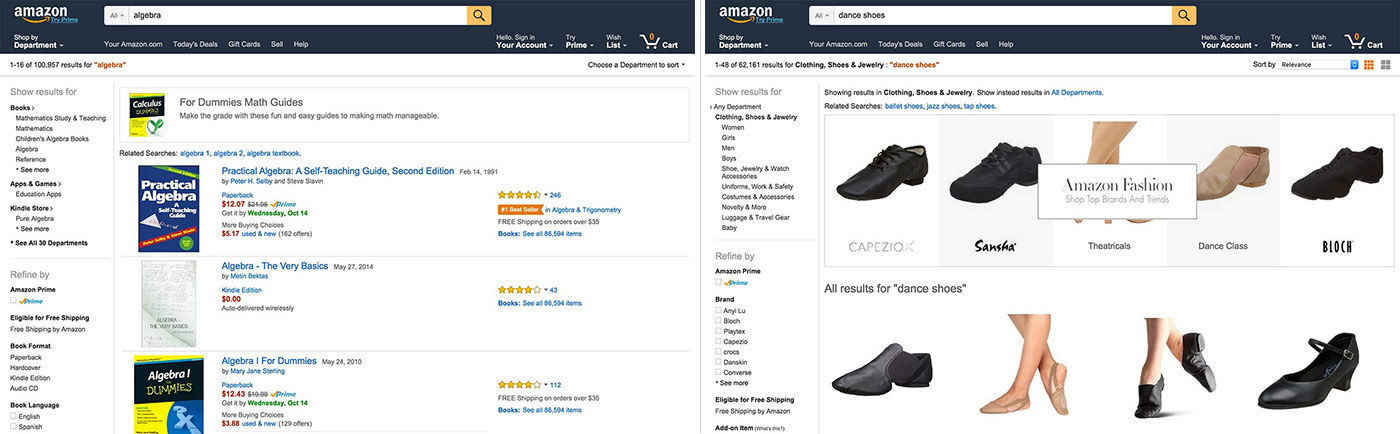
Search engines are increasingly investing in [query understanding](http://www.slideshare.net/dtunkelang/better-search-through-query-understanding) to help users formulate better queries. Query understanding efforts range from spelling correction to query rewriting to intelligent search suggestions. A general trend is to help the user arrive at a structured query that reduces the inherent ambiguity of language and more precisely articulates information-seeking intent. You can see this in action on LinkedIn, where typing "micr" into a search box triggers search suggestions like "Jobs at Microsoft" and "People who work at Microsoft":

Screenshot of the LinkedIn "micr" search triggers.

**Provide one search box, but segment the search experience.**

We've become accustomed to universal search, and most consumer-oriented search engines present a single search box to users that is expected to handle all of their information-seeking needs. But having a single entry point doesn't mean that all searches should lead to the same search experience.

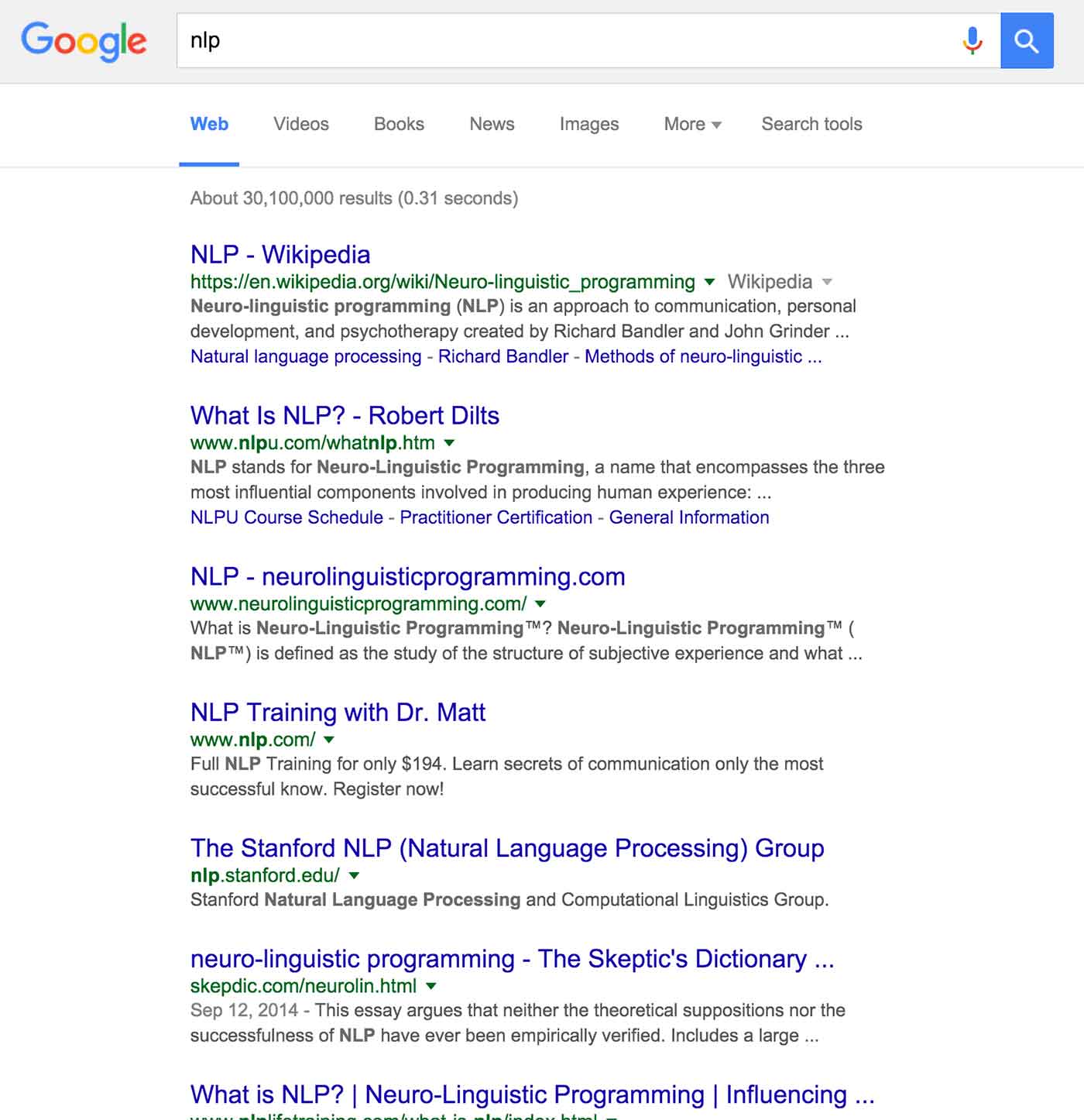
Consider [known-item search](http://www.iva.dk/bh/Core%20Concepts%20in%20LIS/articles%20a-z/known_item_search.htm), where the user knows what he or she is looking for. This kind of search is very different from an open-ended [exploratory search](https://en.wikipedia.org/wiki/Exploratory_search). Even within the same kind of search, different verticals call for different search experiences. That's why Amazon presents the user with a very different search experience when he or she searches for "[algebra](http://www.amazon.com/s/?url=search-alias%3Daps&field-keywords=algebra)" as opposed to "[dance shoes](http://www.amazon.com/s/?url=search-alias%3Daps&field-keywords=dance+shoes)":

Screenshot comparison of Amazon search results.

The trick is to segment the search experience based on an understanding of the search query. Query understanding, which takes place before retrieving or ranking results, is where the search engine establishes what kind of result the user is searching for, and whether the user has a particular result in mind or is simply exploring the space of possibilities. Query understanding thus allows the search engine to accept queries through a single user interface and then route users to segment-appropriate search experiences.

**Recognize when best-effort search isn't good enough.**

Search engines deliver a best-effort experience: they retrieve results and return the best-ranked ones to the top. This is great, except when it isn't. Unfortunately, today's search engines lack the self-awareness to recognize when their best efforts aren't good enough. For example, Google still hasn't figured out, after all these years, that when I search for "[nlp](https://www.google.com/search?q=nlp)" I mean [natural language processing](https://en.wikipedia.org/wiki/Natural_language_processing) and not [neuro-linguistic processing](https://en.wikipedia.org/wiki/Neuro-linguistic_programming):

Screenshot of the Google search results for "nlp."

There's a rich field of [query performance prediction](http://www.morganclaypool.com/doi/abs/10.2200/S00235ED1V01Y201004ICR015) that's been getting increased attention in the information-retrieval community. Various techniques allow search engines to evaluate query difficulty, both before and after result retrieval. Recognizing query difficulty allows the search engine to fall back to a more humble search experience — even something as simple as apologizing to the user for not being sure if it understood the intention of the search. A classic example of this is when the search engine asks the user, "Did you mean:…" — a pattern popularized by Google and now universal.

Query performance prediction can also be part of a more active approach. Any time the search engine determines how to interpret a search query or propose alternative searches, it should use query performance prediction to prune the space of possibilities. This is a largely untapped area, ripe for algorithm and interface innovation.

Finally, we need to revisit how we evaluate search engines. [Evaluation of interactive information retrieval systems](http://research.microsoft.com/en-us/um/people/sdumais/niishonanworkshop-web/NII-Shonan-CompiledHomework_Final.pdf) is a hot research area, bringing together the worlds of information retrieval and human-computer interaction. The first step toward improving our systems is improving how we evaluate them.

**Understand that augmented intelligence trumps artificial intelligence.**

As I said in the beginning of this post, we live in a golden age of algorithms. It's amazing to see the progress at every level of the information-retrieval stack. We are closer than ever to implementing the kinds of artificial intelligence systems that only a generation ago would have been dismissed as science fiction.

But let's not forget that search is a collaboration between human and machine. As the late [Doug Engelbart](https://en.wikipedia.org/wiki/Douglas_Engelbart) said, the most effective way to solve problems is to use computation to augment human intelligence. In other words, the best search engines are those that help users help themselves.