
Big Data in a Small Page - Cognitive Computing Simplified

Through various media like sight, sound, touch, etc., people accept or disseminate information in multiple formats - the written/ spoken language, facial expressions, intonations, physical language like hugs, etc. Each of them is highly nuanced. For example, it is the slightest twitch in the eye that differentiates surprise from sarcasm. This is why it takes a computing system with 100 billion nodes i.e. our brain, to successfully communicate between two people. This kind of processing power is not yet available outside nature.

So we engineers do what we are best at - pick a niche and assume that nothing else is important. If we pick up the simplest of these formats, the language, we still have to tackle with insane complexity. Language is simplest because it is consensus-based. Two persons cannot exchange information unless they agree on language. Hence language has been codified. This is very useful for engineers who are trying to use structured, quantitative computing science for emulating human behavior.

As complicated as it may look, language is the easiest among forms of human communication to codify and be understood by computers.

The trouble is that this consensus-based codification is highly local - geographically, temporally, professionally, demographically, or by any other axis you cut it. For example a sentence like "Today is a good day." means completely different things to people living in San Francisco than to people living in Byrd Station, Antarctica. For people living in Washington DC, it means different things in August than in February. On the same day in Washington DC, say on the day of Supreme Court's decision on Obamacare, this sentence means different things to Republicans than to Democrats.

To ~80% of earth's population that does not understand English, the sentence does not mean anything at all, but let us engineers do the niche magic again and ignore them for a while.

Ergo, to design a cognitive computing solution that can be applied to more than one data point, we have to allow for a lax codification. The same word can take multiple parts of speech. The same meaning is conveyed by different orders of words in the sentence. Try "A good day, today is." The same fact can be expressed using different words. For example, "launch", "unveil", "release", "unlock", "announce", "beta test" and "introduce" may mean the same event in some contexts. The same fact can also be colored in different sentiments or related to multiple other facts in multiple ways based on opinions of different authors.

Each data point in human language can have multiple meanings. Any attempt to codify language has to be able to handle the multiple hypotheses.

Many of the current NLP solutions try to make a good guess as soon as they come across these tokens. For example, Stanford's NLP kit assigns part of speech based on complex sentence parsing models built over statistical dictionaries. Another approach, like Coseer's, is to take every possibility as a hypothesis. As these hypotheses are processed with each other in

increasingly constrained steps, multiple permutations simply perish. For example, in the sentence "A report is due.", we can make two hypotheses about "report" - verb or noun. When we inter-relate these with hypotheses on other words in the sentence, the verb hypothesis is voided. In either case, the simplest decisions require immense computing power.

Let's make this real. At 9:30 pm Pacific Time on March 6, 2015, we ran a basic Finder module on www.nytimes.com. This module only focuses on the spatial position of text elements and creates hypotheses on basic facts encountered in text. The output is then used by multiple sophisticated algorithms to solve real problems. When we printed the JSON for NY Times, the basic mode output for single page ran into 1.7 million lines. It is available here for the brave among you (~4000 pages, ~9MB). For our solutions for Finance and eCommerce sectors, our servers process 1-1.5m documents every day. In other words, for simplest of our products, we need to process data points running into trillions.

A single page of written text can generate million such hypotheses. Processing that needs incredible computing power.

With the recent developments in Big Data technology it has become possible to process this kind of information using commoditized hardware, so now we can attempt to emulate our brain and get computers to do things that have been exclusive human till now. Managing this ambiguity using explosive data permutations within latencies that are sensible for real time applications, is what [cognitive computing](#) is all about. To know more, follow us on social media.

Welcome.

What is Next-Generation Enterprise Search?

Coseer's search solutions are transforming industries from healthcare to finance. Our point-and-shoot AI trains finds answers and insights with 95%+ accuracy within 4-12 weeks - all of this in 100% security. The reason? We founded Coseer on the principle that computers should take care of the boring stuff so that humans can focus on creativity and judgment. To that end, we've built enterprise search solutions to complete complex workflows just as humans would in a fraction of the time. Fortune 500 leaders are using Coseer to speed up and automate their most complex work.

We follow a tactical approach to enterprise search:

- We deliver 95-98% accurate solutions within 4-12 weeks.
- Our solutions deploy entirely behind your own firewall for 100% security, and every decision point is logged for full transparency.
- You add the finishing touches, but our point-and-shoot AI practically trains itself. No more huge training data sets or time wasted annotating and tagging.

Visit our [website](#) for in-depth case studies, ROI breakdowns per industry, and other insight.