

National Clean Out Your Computer Day Hits in February; Or You Can "Hack" the Results of a Clean Computer by Downloading a Search Engine

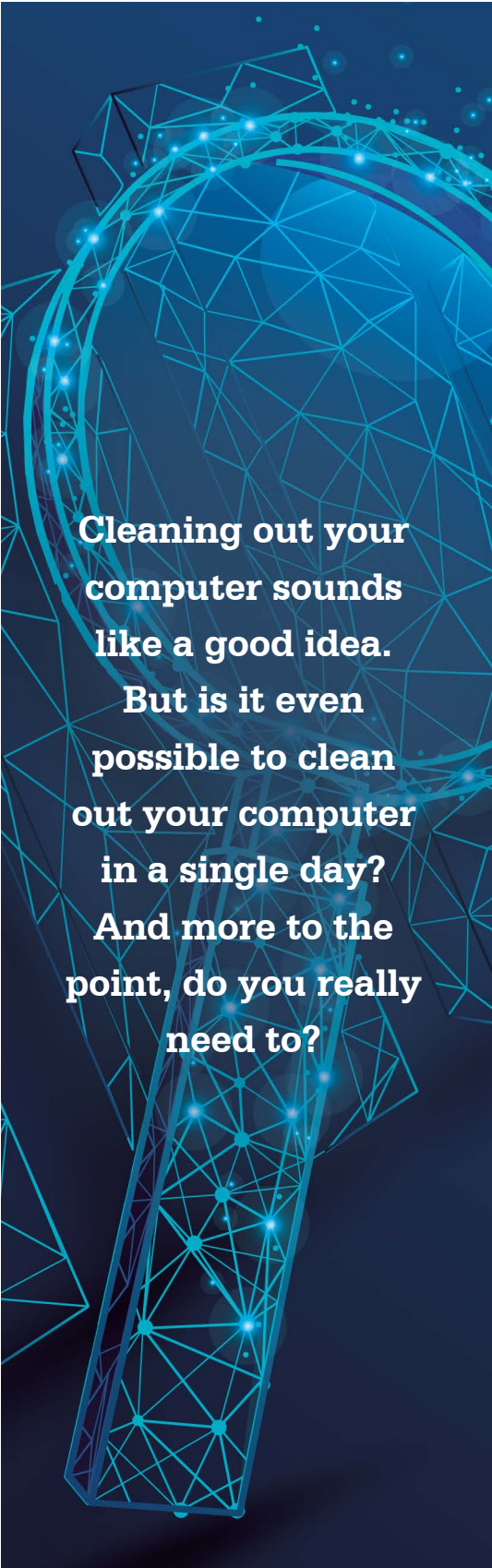
Cleaning out your computer sounds like a good idea. But is it even possible to clean out your computer in a single day? And more to the point, do you really need to?

Alternatively, you could use a search engine to instantly search for your own files and emails. You can also use a search engine across your network or over a shared web server operating "on premises" or in the cloud running on a platform like Microsoft Azure or AWS. In an enterprise setting, the search engine can search not only files and emails but also structured databases like SQL, NoSQL and SharePoint as well as other web-based content like HTML, XML, dynamically-generated web pages, etc.

How does a search engine instantly search terabytes? It first builds an index storing each word and number in all data and its location in the data. To build such an index, the search engine needs to approach all files, emails and the like in their binary format. The binary format provides a very different view than a "normal" view of a file in its native application like Microsoft Word for a word processing document, Access for a database file, Excel for a spreadsheet and Adobe Reader for PDFs. In fact, looking at a binary format with the naked eye, it is hard to discern any readable text at all.

The first step a search engine needs to take in approaching a binary format is to figure out the applicable file parsing specification to apply. Outlook or Exchange has a very different format from OneNote which in turn has a very different format from PDF. In figuring out the applicable file format, the search engine need not rely on the file extension. In fact, these can be misleading. For example, it is possible to save a Microsoft Word document with a .pdf extension and a PDF with a .docx Word extension. But a search engine can look inside of a binary file itself to determine the relevant parsing specification to apply. That way, mismatched file extensions need not "trip up" the search engine.

Article contributed
by [dtSearch®](#)



**Cleaning out your
computer sounds
like a good idea.
But is it even
possible to clean
out your computer
in a single day?
And more to the
point, do you really
need to?**

After recognizing the file format and applying the correct parsing specification, the search engine needs to follow the full text as well as all metadata throughout the file. The search engine then uses all of this information to build its index. In dtSearch, for example, an index can hold up to a terabyte of text and there are no limits on the number of indexes that the application can create and end-users can search.

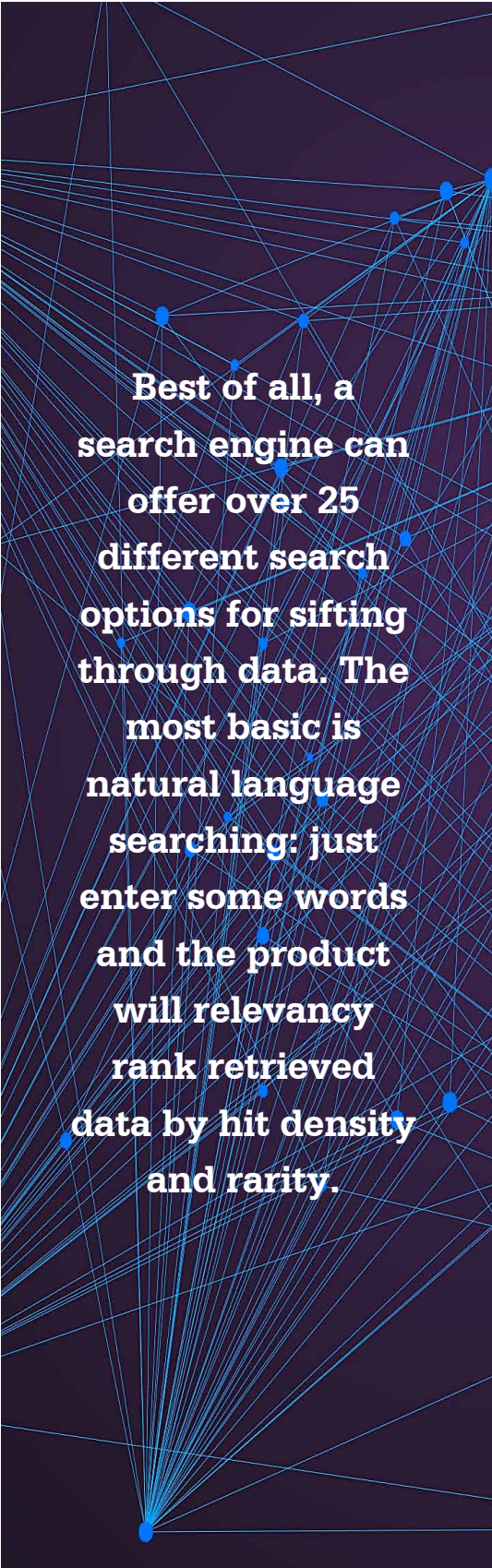
Once the search engine finishes indexing, the index can support individual searching across a PC or laptop. The same index or indexes can also support instant concurrent searching-with no limit on the number of search threads-across terabytes on a network or in a web-based repository. After a search, each end-user can browse the full text of retrieved files, emails and the like with highlighted hits for easy navigation.

Best of all, a search engine can offer over 25 different search options for sifting through data. The most basic is natural language searching: just enter some words and the product will relevancy rank retrieved data by hit density and rarity. More structured search request options include any combination of Boolean "or" / "any words" search requests, Boolean "and" / "all words" search requests, Boolean "not" search requests, and proximity search in one or both directions. Concept search looks for user-defined and/or thesaurus-defined synonyms. Fuzzy search sifts through misspellings, such as if a word is mis-OCR'ed in a PDF or mistyped in an email.

Advanced search options include the ability to find any credit cards in data, and the ability to generate hash values for all files and then search on those hash values. Users can also fine-tune default relevancy by assigning positive or negative ranking and adjusting for whether the hit occurs in file full-text or specific metadata. For developers using the SDK, the software also adds options for faceted or category drill-down searching through database metadata and the like as well as data classification options to granularly filter the search results each end-user sees.

Instead of cleaning out your computer for National Clean Out Your Computer Day, consider a search engine to instantly find your data instead.

Article contributed
by [dtSearch®](#)



Best of all, a search engine can offer over 25 different search options for sifting through data. The most basic is natural language searching: just enter some words and the product will relevancy rank retrieved data by hit density and rarity.