We are, in a sense, drowning in information. Today, it is unusual for scientists even to read a journal cover to cover—much less to personally parse all information pertinent to even a narrow research area. Increasingly complex content, large digital supplements, and a staggering volume of publications are now threatening old-fashioned scientific reading with extinction. But by using computers to sift through and scour published articles, the nascent technology of text mining promises to automate the rote information-gathering stage—hopefully leaving to human minds the more challenging (and rewarding) activity of higher thinking.

This article is intended to continue where Cohen and Hunter [1] left off in “Getting Started in Text Mining,” an introduction in the January 2008 issue of *PLoS Computational Biology* which covered the actual mining of text and its digestion into small quanta of computer-manageable information [http://www.ploscompbiol.org/doi/pch.0040020]. In this overview of the field, we begin by summarizing the major stages of current text-processing pipelines. We now focus on the downstream questions scientists can ask using text-mining and literature-mining engines. At times, we (deliberately) blur the boundary between today’s approaches and tomorrow’s possibilities.

Figure 1 shows a high-level overview of the stages in text mining, with a focus on its applications. We begin at the top left of the figure, which shows the process of information retrieval—how we select relevant documents [2]. Unfortunately, free full-text access remains impossible for a large portion of scientific journals. In some fields, such as chemistry, even article abstracts are inaccessible for a large-scale analysis. The obvious outcome is that articles published in open-access journals have a better chance of being identified as relevant hits than others appearing in traditional “closed-access” journals. Electronic access to text obviously impacts all stages of text mining.

Once the documents have been chosen by an information retrieval engine, a computer scans the text and picks out the various entities (objects, concepts, and symbols) in each sentence. This process, called named-entity recognition [3], draws upon dictionaries of synonyms and homonyms, in addition to machine-learning tools [4], so that an individual entity (say, a protein) is recognized consistently—even though it may be referred to by several different names and acronyms [5]. Named-entity recognition is closely related to the design of controlled terminologies [6] and ontologies for the annotation of texts and experimental data [7]—a process often requiring a monumental community effort [8].

The next step is information extraction (IE) (see pp. 545–559 in [9]). Here, entities are assembled into simple phrases and clauses that capture the meaning of the mined text. To accomplish this, two or more entities are juxtaposed, and meaningful action words—called *predicates*—are chosen to link the entities. For instance, we might say gene *X* *genetically interacts with* gene *Y*, or protein *A* *binds to* protein *B*. Each completed clause describes a basic relationship between entities. The question then becomes, what can we do with all these simple or complex clauses?

The answer is, quite a lot—which helps explain why text mining is poised to become a powerful central pillar in scientific research and recordkeeping. The lower two-thirds of Figure 1 illustrates how the results of information extraction (IE) can be synthesized and used.

Because IE yields a collection of phrases linking entities through predicates, one of its simplest but valuable uses is to answer simple questions posed to an automated system [10]. In this approach, human questions are digested by a linguistic engine (likely using the same process as employed on original mined text) and mapped to simple phrases. These question phrases are then queried against the database of phrases already stored in the computer, which were generated through the application of IE to analyzed text. (Another mode of question answering, bypassing generation and querying of a database entirely, involves direct search and analysis of relevant texts. These texts can be stored at a local computer disk or distributed on numerous computers around the world.) Figure 1 outlines the basic process by which the machine interprets the question, queries its database of stored relationships, and returns an answer.

IE-generated knowledge often tracks closely the needs of experimental biologists. Typical IE systems are developed in direct response to acute practical problems, such as large-scale annotation of regulatory regions in genomes [11], collecting published claims about experimental evidence supporting a collection of assertions [12], and condensing sparse information about phenotypic effects of mutations in proteins [13].

Of course, IE-generated databases can be supplemented with additional data gleaned from experiment, or contributed through other non–text-mining means. A simple user interface could facilitate contributing raw experimental data or other...
information into the database of relationships expressed as simple phrases—again, entities linked by actions (see, for example, the REFLECT system, http://reflect.ws/). Adding more such data should correspondingly increase the effectiveness of the computer’s answers to user questions.

Another major use for the database of IE-generated phrases is to employ the collection itself for the discovery of new information [14,15]. One approach to this is to seek out “idea isomorphisms”, by which we mean identifying similar types of logical constructs across different contexts. Finding that similar small ideas (or phrases) occur in different fields might allow researchers to bridge different areas of inquiry. Such bridging of fields, in turn, might uncover new connections, thereby suggesting new and unexpected hypotheses that can then be tested experimentally.

The collection of phrases can also be used to vet and prune itself by examining the consistency among many entries. For instance, conflicting or erroneous data can be flagged. By examining each record situated within a large number of records,
References

1. Cohen KB, Hunter I. (2008) Getting started in text mining. PLoS Comput Biol 4: e20. doi:10.1371/journal.pcbi.0040020.

2. Hersh W, Hickam D (1995) Information retrieval in medicine: The SAPHIRE experience. Medinfo 8/Part 2: 1433-1437.

3. Hirschman L, Morgan AA, Yeh AS (2002) Rutabaga by any other name: Extracting biological names. J Biomed Inform 35: 247–259.

4. Kim JD, Ohta T, Tsujii J (2008) Corpus annotation for mining biomedical events from literature. BMC Bioinformatics 9: 10.

5. Sasaki Y, Tsurukado Y, McNaught J, Ananiadou S (2008) How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics 9(Supplement 1): S5.

6. Spaie J, Schober D, Sansone SA, Rehholz-Schuhmann D, Kell DB, et al. (2008) Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. BMC Bioinformatics 9(Supplement 5): S5.

7. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, et al. (2009) Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 10(Supplement 2): S1.

8. Huss JW III, Orozco C, Goodale J, Wu C, Batalov S, et al. (2008) A gene wiki for community annotation of gene function. PLoS Biol 6: e175. doi:10.1371/journal.pbio.0060175.

9. Mitkov R (2003) The Oxford handbook of computational linguistics. Oxford/New York: Oxford University Press. pp 784.

10. Crangle CE, Cherry JM, Hong EL, Zbyslaw A (2005) Searching natural language text by computer. Machine indexing and text searching offer an approach to the basic problems of library automation. Science 132: 1099–1104.

11. Aerts S, Haesder M, van Veenen S, Griffith OL, Hulpiau P, et al. (2008) Text-mining assisted regulatory annotation. Genome Biol 9: R31.

12. Crange CE, Cherry JM, Hong EL, Zbyslaw A (2007) Mining experimental evidence of molecular function claims from the literature. Bioinformatics 23: 3232–3240.

13. Kanagasabai R, Choo KH, Ranganathan S, Baker CJ (2007) A workflow for mutation extraction and structure annotation. J Bioinform Comput Biol 5: 1319–1337.

14. Blagodkina MV, Pardee AB (2002) Conceptual biology: Unearthing the gems. Nature 416: 373.

15. Douglas SM, Montelon GT, Gerstein M (2005) PubNet: A flexible system for visualizing literature derived networks. Genome Biol 6: R30.