

DeepDive (<http://deepdive.stanford.edu>) is a new type of data processing system that has demonstrated the ability to extract structured SQL-like databases from unstructured text and tables (Dark Data) *with higher quality* than human annotators. DeepDive is currently in use in anti-human trafficking applications with NGOs and law enforcement, a handful of enterprise companies, and scientific efforts in genomics, drug repurposing, electronic medical records, and paleobiology. Compared to human annotators, machine learning systems are often thought of as a mechanism to obtain data inexpensively but with lower quality. This article takes the position that for some tasks, machine-based solutions can be more efficient, more accurate, and have far greater coverage than human annotators. To accomplish this, DeepDive has a radically different design than a traditional database system: probabilistic inference is its core operation.

**Macroscopic Analysis** A variety of data-driven problems facing industry, government, and science are *macroscopic*, in that answering a question requires that one assemble a massive number of disparate signals that are scattered in the natural language text, tables, and even figures of a number of data sources. A macroscopic corporate intelligence application might combine sources to provide an overview of a competitor’s holdings, or of the patents held by an acquisition target. A medical diagnostic application might enumerate genomics data from thousands of scientific papers and patents. However, traditional data processing systems are unsuited to these tasks, as they require painful one-off extract-transform-load procedures. We describe our recent experience with DeepDive and its use in these applications.

**Application Lessons** Recently, DeepDive has gone from a purely academic prototype to in regular use by companies, law enforcement, NGOs, and scientists. In these applications, we made a few observations that suggest that systems like DeepDive may be preferred over manual curation efforts—not only due to their ability to produce lower cost data.

- Human annotators often incompletely extract information. Compared to human annotators at PaleoDB.org, DeepDive extracted 10× more high quality facts from the same documents. Of course, DeepDive systems can read orders of magnitude more documents than even well funded manual curation efforts. For example, major medical extraction efforts annotate tens of thousands of documents per year, but each year more than one million articles appear on PubMed, which can be processed by DeepDive in hours.
- Human volunteers make insidious errors. Volunteers use background information, e.g., “this condition is now called by a new name.” When their background knowledge is correct, the data product improves. When they’re wrong, these insidious errors are impossible to track down. In contrast, DeepDive systems make systematic errors that can be fixed. This led to DeepDive systems besting human competitors by up to 10% in precision. Moreover, the entire system can often be rerun in hours—in contrast, some human collection efforts are measured in person decades. One cannot even dream of “rerunning” such a process.

DeepDive builds on a recent inflection points in machine learning, e.g., natural language processing can reliably parse sentences and identify named entities. DeepDive provides a mechanism to combine such tools to build applications.

**Design Principles** DeepDive takes a radical view for a data processing system: it views every piece of data as an imperfect observation—which may or may not be correct. It uses these observations and domain knowledge expressed by the user to build a statistical model. One output of this massive model is the most likely database. As aggressive as this approach may sound, it allows for a dramatically simpler user interaction with the system than many of today’s machine learning or extraction systems. As a result, non-computer science users are able to operate DeepDive. Our design has two core principles:

- **Features Not Algorithms** To cope with the noise in these sources, the standard approach is to use machine learning or statistical inference. A user of DeepDive does not specify any algorithm—only the features or signals that are relevant, i.e., users specify their task *declaratively*, which allows the user to focus on their task rather than whether to use an SVM, which clustering algorithm, etc. This focus has allowed a wider array of users to download and use DeepDive—even those without graduate degrees in computer science.
- **Declarativity leads to Debugability** Algorithms provide a powerful way to understand an application: if a result looks suspicious, one can trace through the algorithm step-by-step to understand this result. A major challenge is to enable users to understand the results—without explaining *how they were computed*. DeepDive uses probabilistic inference, which provides a semantically meaningful output independently of how that value is computed. In particular, every fact produced by DeepDive produces an accurate probability value. If DeepDive reports a fact with probability 0.95, then 95% of the facts with that score are correct. In contrast to extraction pipelines that force users to over build individual components in the hope of improving downstream quality, these probabilities take into account all information—allowing the user to focus on quality bottlenecks.

A key challenge is to compute the inference problem efficiently. To this end, DeepDive pioneered new relaxed consistency execution models for statistical engines, e.g., Hogwild!, that are used more broadly. Our early exploration of these systems has led us to believe that both existing BI applications “on steroids” and entirely new predictive applications are possible using the DeepDive approach.