A machine-compiled macroevolutionary history of Phanerozoic life

Shanan E. Peters *, Ce Zhang †, Miron Livny †, and Christopher Ré ‡
* Department of Geoscience, University of Wisconsin-Madison, Madison, WI, 53706 USA
† Department of Computer Science, University of Wisconsin-Madison, Madison, WI, 53706 USA
‡ Department of Computer Science, Stanford University, Stanford, CA 94305 USA
{peters@geology, czhang@cs, miron@cs}.wisc.edu chrismre@cs.stanford.edu

Abstract

Many aspects of macroevolutionary theory and our knowledge of biotic responses to global environmental change derive from literature-based compilations of paleontological data. Although major features in the macroevolutionary history of life, notably long-term patterns of biodiversity, are similar across compilations, critical assessments of synthetic databases have been limited to the enumeration of taxonomic and geochronological errors in single lineages. Existing databases also leverage a small fraction of relevant published knowledge and are difficult to extend with new data types. Here, we develop a statistical machine reading and learning system, PaleoDeepDive, to automatically find and extract data from the text, tables, and figures of publications. We show that PaleoDeepDive requires comparatively little training data to perform comparably to humans in many complex data extraction tasks, and then deploy the system to extend the human-constructed Paleobiology Database to include nearly ten times more journal articles. Large-scale Phanerozoic taxonomic diversity and genus-level extinction and origination patterns are robust, even when derived from different bodies of literature. Unlike traditional databases, PaleoDeepDive produces a probabilistic database that improves as new information is added and that is extendable to include data not previously accessible on large scales, including morphological data in biological illustrations. Although literature-based compilations will always be subject to errors caused by inconsistent and erroneous data reporting, our high quality machine-reading approach to data synthesis and integration brings within reach questions that are now underdetermined and does so in ways that may stimulate new modes of inquiry.

Paleontology is based on the description and biological classification of fossils, an enterprise that has played out in countless collecting expeditions, museum visits, and an untold number of scientific publications over the past four centuries. The construction of synthetic databases that aggregate fossil data has greatly expanded the intellectual reach of paleontology (1-8) and led to many fundamental new insights into macroevolutionary processes (e.g., 9-16) and the nature of biotic responses to global environmental change (e.g., 17-20). Nevertheless, paleontologists often remain data limited, both in terms of the pace of discovery and description of new fossils, and in terms of their ability to find, access, and synthesize existing knowledge on the fossil record. Many other sciences, particularly those for which publication is a primary means of data distribution, face similar challenges, which diminishes the overall return on investments in primary data acquisition and which limits the pace and scope of scientific inquiry.

The Paleobiology Database (PBDB) is one of the largest compilations of fossil data yet assembled. Founded nearly two decades ago by a small team who generated the first sampling-standardized global Phanerozoic biodiversity curves (21,22), the PBDB has since grown to include more than 300 international scientists with diverse research agendas. Collectively, this group has spent nearly 10 continuous person years entering more than 290,000 taxonomic names, 500,000 taxonoic opinions, and 1.17 million fossil occurrences (i.e., temporally and geographically resolved instances of fossils). Some data derive from the fieldwork and taxonomic studies of the contributors, but the majority of the data were acquired from over 40,000 publications. Nevertheless, the PBDB leverages only a small fraction of the paleontological literature. Moreover, because the database is divorced from original sources, asssessing data quality and extending it to include new data types is difficult.

Here we develop and deploy PaleoDeepDive (PDD), a statistical machine reading and learning system, to find and extract taxonomic and fossil occurrence data from the published literature. Our motivations for doing so are threefold. First, we aim to quantitatively test the reproducibility of the PBDB and key macroevolutionary results that frame much of our understanding of the large-scale history of life. Second, we aim to overcome many of the challenges to machine reading that are posed by ambiguity at a large scale and scope, in this case within the scientific literature. Third, we aim to develop a system with the capacity to change the practice of science by removing the substantial barriers to large-scale data synthesis and integration that currently exist. In so doing, we hope to shift the balance of effort away from time-consuming and expensive data compilation efforts and towards creative hypothesis testing and more efficient generation of new data. Assessing the quality of our system is therefore critical to establishing its potential utility and for testing the specific hypothesis that a machine reading system can perform comparably to humans in complex scientific data extraction tasks.

1 System Description

1.1 Overview

A fundamental challenge faced by machine reading systems is that computers cannot read documents unambiguously. Instead, machines have difficulty with all aspects of document reading, from optical character recognition (OCR) and natural language understanding tasks, to the more complex subtleties involving domain-specific representations of facts. As a result, coping with ambiguity is a key challenge in many areas of computer science (23-27).

To accommodate the inherent ambiguity of the literature, PDD is built upon the DeepDive machine reading infrastructure (27), which is designed to extract information from text, tables, and figures in a way that achieves a deeper level of understanding than previous generation systems. To do this, DeepDive treats all sources of information, including existing data and dictionaries, as evidence that may or may not be correct. Extraction tasks then become probabilistic inference challenges. A joint probabilistic, or collective inference (28), approach is motivated by the challenge of retrieving complex, structured information from a heterogeneous and unstructured literature designed for human visual consumption. Other systems use a pipelined approach to data extraction (26, 29, 30), in which hard decisions are made after each stage of document processing, leading to compounding errors and suboptimal data quality. The cost of a joint probabilistic approach is that the underlying computational problem is more difficult because complexity grows exponentially with each source of ambiguity. Recent work, driven by the challenges posed by building PDD, allows us to perform the necessary statistical inference tasks orders of magnitude more efficiently than was possible just several years ago (31-35).

Similar conceptual underpinnings are now in use by Google's Knowledge Graph, IBM's Watson, and CMU's NELL project, but these systems have not been applied to the scientific literature in a systematic way, nor have they tackled as complex a problem as we do here.

1.2 PaleoDeepDive Pipeline

The input to PDD is a set of documents, such as PDFs or HTML sources, and a database structure that defines entities and relationships of interest. The first step in the DeepDive process is to perform document parsing tasks, including optical character recognition (OCR), document layout recognition, and natural language parsing (NLP) of the text (Fig. S1). These steps are required before the system can apply any of the reasoning necessary to recognize entities and the relationships among them. An example of the latter is: "Does this instance of the word 'Waldron' refer to the 'Waldron Shale', a geological formation, and if so, what is its geologic age, where is it located geographically, and which fossils are reported from it?" The semantics for how entities and the relationships among them are recognized can be articulated by scientists and then formalized into features and rules (Fig. S2; Tables S1, S2). The weights of these are then estimated (i.e., learned) from the data using classical equations based on exponential models (28). Essentially, the likelihood of the given set of observations is maximized, given the set of features expressed by the rules (Fig. S3). DeepDive is able to learn from existing structured data, rules provided by users, and traditional training examples.

The end-product of PDD is not a classical database, in which facts are all assumed to be correct. Instead, DeepDive produces a probabilistic database in which each fact is associated with an estimated probability of being correct (36). Only those facts that have a probability satisfying some quality threshold (e.g., ≥ 0.95) are used in analysis.

2 Results

2.1 Overlapping Document Set (ODS)

To assess PDD's ability to extract data from the literature, we used the PBDB as a baseline for comparison. Specifically, 11,782 documents from the top-50 serials in the PBDB were accessible to and processed by PDD (Table S3).

On average, PDD extracts more taxonomic data from a given document than humans. For example, humans extracted 79,913 opinions on the status and biological classification of taxonomic names from the ODS, whereas PDD extracted 192,365 opinions. Although many of PDD's extracted taxonomic opinions are simple cases that are often not entered by humans (e.g., a species belongs to a genus), they nonetheless constitute taxonomic information which is sometimes not entered into the PBDB at all. For example, PDD extracted 59,996 taxonomic names that have never been entered as taxonomic entities in the PBDB. A random sample of these names indicates that most are species-level taxa and that $\geq$90% were correctly extracted as taxonomic entities (Table S4). Other categories of facts, such as geological formation-taxon tuples, currently have similar or lower rates of recovery in PDD. The cases where PDD completely failed to recognize and extract data from a document are due primarily to OCR-related errors (Materials and Methods; Tables S5, S6).

The quality of PDD's database was assessed in three ways. The first used DeepDive's internal measures of precision. All of the extractions used here have a precision of $\geq$ 95% according to this criterion. We also conducted blind assessment experiments of two types. In the first experiment, we randomly sampled 100 relations from each database and then randomized the combined 200 extractions into a single list. This list was then manually assessed for accuracy relative to source documents. The results show that PDD achieves $\geq$ 92% accuracy in all cases, which is as greater or greater than the accuracy estimated for the PBDB (Table S7). In the second blind experiment, eight paleontologists were presented with the same five documents and the same 481 randomly selected taxonomic facts (Fig. S4). No indication was given regarding which system generated the facts. Humans measured a mean error frequency in the PDD-constructed database of 10%, with a standard deviation of ±6%. This is comparable to the error rate of 14±5% they estimated for those same documents in the human-constructed PBDB (Fig. S5). Variability in the estimates between humans reflects a combination of assessment error and divergent interpretations of the data. These assessments suggest that the error rate is comparable in both compilations, but the comparisons are not strictly equivalent. For example, PDD now understands only parent-child relationships and synonymy, which comprise a large fraction (90% and 5%, respectively) but not all of the opinions in the PBDB. Human data enterers also rarely enter all of the data from a given document. Instead they selectively enter data that are deemed important or non-redundant with data in other documents.

The third approach we took to assessing PDD was conducted at the aggregate level of Phanerozoic diversity and rates of extinction and origination (37). After processing the both databases with the same algorithms in order to generate a working taxonomy as well as occurrences with the same minimum threshold for temporal resolution, we find good overall agreement in macroevolutionary quantities (Fig. 1; data are binned into the same 52 time intervals, mean duration 10.4 Myr). Long-term trends and interval-to-interval changes in genus-level diversity and turnover rates are strongly positively correlated. The number of genus-level occurrences in each time interval, which is important to sampling standardization approaches (38,39), are also positively correlated (for first differences, Spearman rho = 0.65; $p = 5.7 \times 10^{-7}$). The times of first and last occurrence of the 6,708 taxonomically and temporally resolved genera that are common to both databases are congruent (Fig. 2).

Differences between the macroevolutionary results can be attributed to a combination of errors and inconsistencies in the human-constructed database, as well as to data recovery and inference errors committed by PDD. For example, the PBDB contains typographical errors that occur when humans transcribe information from one source to another. But, there are more insidious inconsistencies that contribute to most of the differences observed in Fig. 1. There are groups of occurrences in the PBDB that derive from multiple documents, even though only one document is cited as the source of data. Occurrences are also sometimes attributed to a reference that actually contains no relevant data but that instead cites the PBDB, or some other archive, as its data source. A more prevalent cause of discrepancy involves the injection of information by humans during the data entry process. Most notably, approximately 50% of the ages assigned to PBDB fossil occurrences are not actually mentioned in the cited reference (Fig. S6). Although problematic in some senses, this is justified scientifically. The stated age for an occurrence in a publication is often not the best age that is available, and the PBDB has no capacity to dynamically assign ages to fossil occurrences. Humans attempt to account for these limitations by entering what they determine, on the basis of other evidence, to be the best age. PDD replicated aspects of this behavior by inferring across all documents the most precise and most recently published age for a given geological unit and location, but this approach is not sufficient to cover the full range of sources that were used by humans. Thus, a disproportionate number of the occurrences extracted by PDD have a temporal resolution (e.g., period-level) that causes them to be excluded from the macroevolutionary quantities shown in Fig. 1. Including low-resolution occurrences causes the diversity curves (Fig. 1c) to more closely converge (Fig. S7).

Errors and limitations in the current PDD system also account for some divergence in Fig. 1. For example, OCR failures, often involving data-rich tables, are among the leading causes of data omissions (Table S6). The current version of PDD also has elements of design that cause some facts to be omitted. For example, PDD currently places great importance on formal geologic units, which means that no occurrences are recognized in references that do not have well defined geologic units. This commonly occurs when a study is taxonomically focused, when it covers regions with informally resolved stratigraphy, or when deep sea drilling cores are the source of data. Because these situations are more prevalent in recent time intervals, the lower total diversity recovered by PDD towards the recent (Fig. 1) is in part attributable to this design decision. Data omissions also occur when a fact is correctly extracted by PDD, but with a probability that is <0.95, the threshold used to generate the results. This type of confidence-related error can be overcome by examining the subset of facts that are correctly extracted by PDD, but that fall below the threshold probability, and then defining relevant features or rules that can be used to distinguish them.

Despite errors in both the human- and machine-generated databases, these results demonstrate that PDD performs comparably to humans in many data extraction tasks and that the aggregate macroevolutionary results are consistent between compilations. However, it is also the case that macroevolutionary quantities are robust to random errors (40-42). Thus, PDD's synthetic results (Fig. 1) could be interpreted as evidence for the presence of a strong signal in the paleontological literature that is easily recovered. The narrow distribution of range offsets on a per-genus basis (Fig. 2) suggests that PDD's precision is nonetheless high, even at the level of individual genus ranges.

2.2 Training Data Requirements

The human-constructed database was used as both a source of training data and as a benchmark for evaluating system quality. Therefore, an obvious and important question is, how big would the PBDB have to be in order for there to be sufficient training data to obtain a high quality result?

To assess the effect of training data volume on the quality of PDD extractions, we randomly sampled the PBDB to produce a series of smaller databases. We then re-ran the entire system in exactly the same way, but using only the subsampled data for training purposes. As expected, both the amount of data extracted by PDD (with a probability ≥ 0.95) and the accuracy of those data, summarized in aggregate as the Spearman rank-order correlation between first differences in genus-level diversity (as in Fig. 1C), increases with the amount of training data. However, rather little data is required in order to achieve a

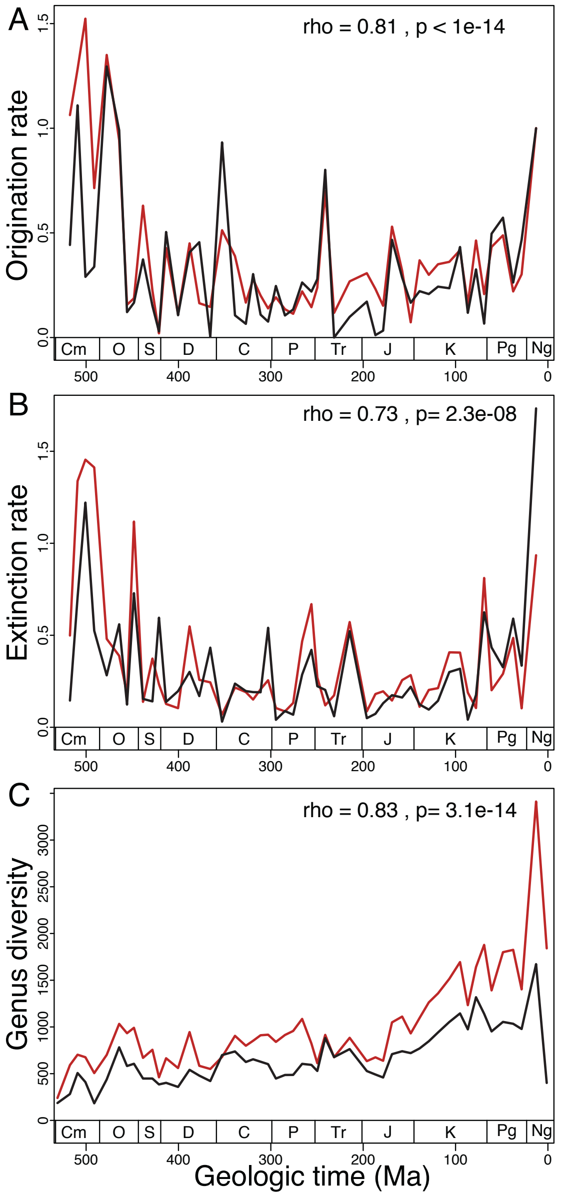


Fig. 1. Macroevolutionary results for the overlapping document set. PBDB-generated (red), machine-generated (black). Spearman rank order correlations for first differences shown. (A) Per capita per interval origination rate (37). (B) Per capita per interval extinction rate. (C) Total range-through diversity.

high-quality result (Fig. 3). If humans had gatehered data from just 1,000 references, or approximately 2% of the total number of references entered over nearly two decades, there would be sufficient training data to obtain a comparable result. This is true despite errors introduced during the training data entry process because PDD does not assume that any data are 100% accurate.

2.3 Whole Document Set (WDS)

Scaling PDD up to extract data from every paper ever published in paleontology poses little technical challenge and would offer a statistical advantage that could improve the overall quality of PDD. However, access

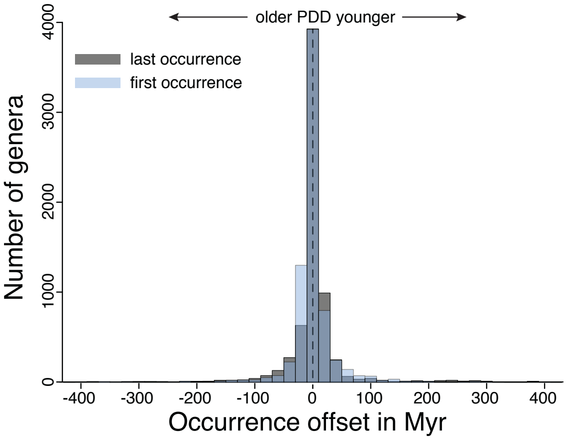


Fig. 2. Difference in genus range end points for 6,708 genera common to the PBDB and PDD. Median offset is 0 Myr for both first and last occurrence. Mean is +1.7 Myr for last occurrence, -0.3 Myr for first occurrence.

to the scientific literature for the purpose of automated text and data mining is currently arbitrarily limited (43). This is true even for documents that the human operators of machines are otherwise free to download and extract data from manually. Because of this external limitation on our ability to access the published scientific literature, PDD's entire document set currently consists of only 294,463 documents (Table S8). Notably for this study, many of these documents were obtained from the open-access Biodiversity Heritage Library, which contains a large number of valuable but older and taxonomically-focused publications, most of which do not contain fossil occurrence data and therefore do not contribute to the synthetic results that we focus on herein.

Despite limitations on our ability to access much of the most relevant paleontological literature, the PDD-generated Phanerozoic diversity curve for the whole set of documents (Fig. 4) yields a face-value empirical genus diversity history that is highly congruent with Sepkoski's classical estimates based on different sources of data (4,5,21). First differences in Phanerozoic diversity extracted from the WDS are also significantly positively correlated with first differences in diversity for the whole PBDB database (Table 1). Genus-level rates of extinction and origination are also similar in both compilations (for first differences, $p < 0.0004$). Moreover, the diversity histories of major groups of organisms (mostly Linnaean classes) comprising this total diversity are significantly positively correlated (Table 1). These similarities hold even though fewer than 25% of the references in the PBDB were also read and processed by PDD (a total of 22,250 valid genera with resolved stratigraphic ranges are common to both compilations).

3 Discussion

The results of the PDD-PBDB comparisions presented here have three important implications. First, we have demonstrated that our machine reading system is capable of building a structured database from a heterogeneous scientific literature with quality that is comparable to, and in some cases possibly even exceeding, that produced by human readers (at least in the dimensions addressed here). This is a notable result because current benchmarks in machine reading and knowledge base construction, such as the Text Analysis Conference Knowledge Base Population competition, achieve less than 50% accuracy (albeit in the broader domain of general web text). Second, we have tested the reproducibility of the PBDB at a large scale, and in so doing we have identified sources of error and inconsistency that are not unexpected in manual data compilations. We have, however, also shown that the macroevolutionary patterns produced by the PBDB are robust. Third, we have shown that literature-based macroevolutionary patterns are similarly expressed, even when they derive from different (but sufficiently large) bodies of literature. This indicates that the paleontological literature, and presumably the underlying sampled fossil record, contains strong

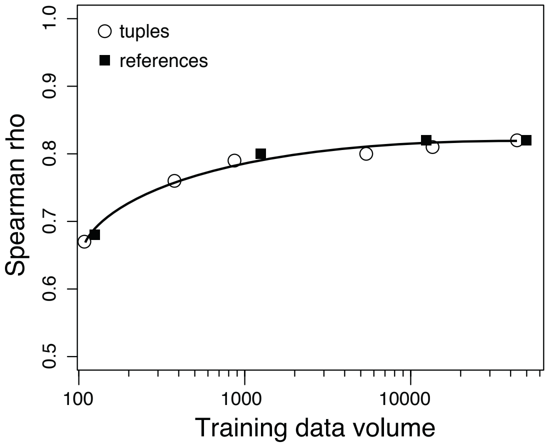


Fig. 3. Effect of changing PBDB training database size on PDD quality. Spearman rho is correlation between human- and machine-generated time series of diversity, as in Fig. 1c.

macroevolutionary signals that are readily recovered. This type of reproducibility and consistency does not mean that our understanding of the global fossil record is uniform taxonomically or in time and space (Fig. S8), that our understanding of the true history of global biodiversity in the Phanerozoic, as opposed to the face-value fossil record, is accurate (21,22, 44-46), or even that the literature contains the data that are required to reconstruct an accurate macroevolutionary history for every clade (e.g., 42). It does, however, indicate that our literature compilation-based knowledge of major Phanerozoic macroevolutionary patterns is mature and unlikely to change substantially simply by increasing the number of existing publications that are included in the PBDB.

The ability to expand literature-based databases and more rapidly create other synthetic data resources, with quality that meets or exceeds human standards, is a notable achievement. However, a much greater advantage of our system is that the type of database it produces is fundamentally different from classical databases. In the probabilistic database (25) produced by PDD, every fact is associated with an estimated probability of being correct and each fact remains tightly coupled to its full and original context. Thus, the quality of the entire database can be systematically improved whenever feedback is given on any one fact or when additional rules or data is added to the system. More importantly, PDD's data acquisition process is based on the visual and textual analysis of entire documents, not the extraction of snippets of text that contain a specified string or set of predetermined facts. PDD is, therefore, able to recognize and analyze data that are not currently part of the database but that are related to it by virtue of taxonomy, geology, or geography.

For example, the illustration of specimens is central to biological systematics and there are consequently millions of images of fossils and living organisms in the full document set. Among the features conveyed by biological illustrations and their associated textual descriptions are morphological attributes, such as body size, a fundamental property of organisms that determines many aspects of their ecology (e.g., 47,48). Several studies have examined the evolution of body size in individual lineages (e.g., 9,49,50), and a new effort to manually extract size measurements from the Treatise of Invertebrate Paleontology is now coming to fruition (51). However, like the PBDB, such database initiatives cover only a small portion of the available data and yield similar traditional databases that are decoupled from primary sources and therefore difficult to assess or extend.

To test the ability of our machine reading and learning system to rapidly incorporate new types of data in illustrations, we extended PDD to identify images of specimens, locate and measure their major and minor axes, and read associated figure labels, captions, and text in order to determine magnification, the portion of the organism being imaged, and its taxonomic identity (SI Materials and Methods). The PDD-estimated body sizes for taxonomically classified brachiopod genera (Fig. S9) are congruent with body sizes estimated measured manually with calipers in the same volumes of the Treatise of Invertebrate Paleontology

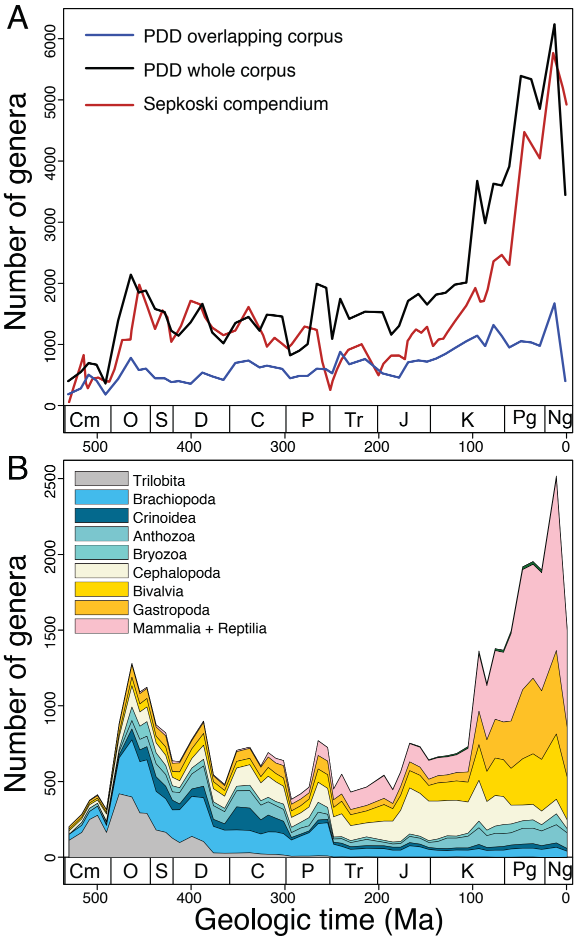


Fig. 4. Genus-level diversity generated by PDD for the whole document set. (A) Total genus diversity calculated as in Fig. 1. For comparison, Sepkoski's genus-level diversity curve (6) is plotted using his stage-level timescale. (B), Diversity partitioned by genera resolved to select classes by PDD.

(51,52). Leveraging PDD's capacity to quantitatively analyze the entire body of biological illustrations, in the full context of their textual descriptions morphology and taxonomy, may allow new approaches to biological systematics and collection curation and brings within reach questions that require a combination of morphological, geologic, and taxonomic data. However, before PDD can be deployed in this fashion, the current barriers to automated access and processing of published scientific documents, for the purpose of extracting data that can be used to facilitate and advance science, must be removed.

Although we have focused here on validating our machine reading and learning system and on testing the robustness of literature-derived macroevolutionary patterns in the PBDB and the paleontological literature, our approach has much broader applicability. A large, manually constructed database, like the PBDB, is not required in order for our machine reading approach to achieve high quality, though it is always the case that statistical power increases with the amount of data available. Thus, this machine reading and learning system has the capacity to tackle many questions that have been posed before, but that have been deemed too difficult to address because of the prohibitively time consuming data collection efforts they require. More importantly, this approach to data synthesis yields a fundamentally different type of probabilistic database, one that remains tightly coupled to primary sources and that is capable of rapidly discovering and integrating voluminous and complex data in ways that are likely to stimulate entirely new modes of questioning.

Table 1. Genus-level diversity in the whole document set and the entire PBDB. Spearman rank-order correlation coefficients and p-values for detrended diversity time series (from Fig. 4) shown.

Taxonomic group	Spearman rho	P-value
All genera	0.72	$3.6\text{x}10^{-9}$
Bivalvia	0.67	$6.2\text{x}10^{-8}$
Bryozoa	0.64	$3.6\text{x}10^{-7}$
Gastropoda	0.59	$5.3\text{x}10^{-6}$
Anthozoa	0.53	$6.6\text{x}10^{-5}$
Brachiopoda	0.52	0.0001
Reptilia	0.50	0.0002
Trilobita	0.49	0.0003
Cephalopoda	0.41	0.003
Mammalia	0.40	0.004
Crinoidea	0.39	0.004

4 Materials and Methods

4.1 System

Features that relate facts in PDD are encoded in a relational database. These features derive from two sources: a set of functions written in the DeepDive framework and a set of existing tools developed by other researchers, including Tesseract and Cuneiform for text, Abbyy Fine Reader for tables, and StanfordCoreNLP for linguistic context. The list of features and rules used in this version of PDD are summarized in (Tables S1, S2).

After extracting features in documents, the next step is to generate a factor graph (Fig. S3), which is a compact way of specifying exponential family probability models (28, 53). The factor graph is defined by a hypergraph (V, E) where V is a set of random variables and $E \subseteq 2^V$ define groups of variables (factors) that are correlated. In addition, each random variable is associated with a domain (for simplicity, consider a Boolean random variable). Each factor (edge) $e = (v_1, ..., v_k)$ is associated with a scalar function called a potential (weight) $\phi_e : \{0,1\}^k \mapsto \mathbb{R}$. For example, the tuple (Tsingyuan Fm, Namurian) corresponds to a random variable, which assumes the value 1 if true. To specify a correlation, for example, if (Tsingyuan Fm, Carboniferous) is true, then it is likely that (Tsingyuan Fm, Namurian) is also true, a factor can be encoded to relate the variables. This factor is only a statistical implication; PDD will estimate the strength of this implication on data.

The factor graph in PDD can be conceived of as existing in three layers (Fig. S3). The first layer corresponds to the set of entities detected as individual mentions in documents. The second layer corresponds to a set of relation candidates between mentions, and the third layer corresponds to a set of relation candidates between distinct entities. One can think of the second layer as a per document layer and the third layer as the "aggregation" across all documents, but all information is used simultaneously at the inference and learning stages.

Given a factor graph, PDD next learns the weight for each factor and then runs inference tasks to estimate the probability of each random variable. One key challenge of machine reading approaches is how to generate training data (i.e., a set of random variables that have been assessed for accuracy). Traditional approaches include human expert annotation of results and crowd-sourcing (54). The human-constructed PBDB allows PDD to make extensive use of a generalization of Hearst patterns called distant supervision (55-56). Even simple lists of facts, such as the location and general geological age of geological formations, can be used during distant supervision to improve the quality of more complex inferences.

Factor graphs are a convenient way to define random variables and their correlations, but they can be large. In PDD, the factor graph contains more than 200 million random variables and 300 million factors with 12 million distinct weights (Table S9). PDD uses recent research in both theory (31, 32) and systems (33) to address this computational challenge. Further details are given the SI Materials and Methods.

4.2 Documents

Tables S1 and S8 list the serial publications used in the ODS and WDS. Some of the serials in the top-50 PBDB sources were not accessible. We were also not able to able to recover all references in the PBDB, due primarily to incomplete bibliographic information (Tables S10, S11) and document processing failures (see Assessment, below). To match retrieved documents to specific PBDB references, we first used the TokenSet Cosine similarity approach (57) and then created an Amazon Mechanical Turk job, in which 64 human workers combined for 30,182 match evaluations. To obtain the WDS, we extended the ODS to include all available documents in the top-50 serials and the whole Biodiversity Heritage Library.

4.3 Features

All PDD feature extraction tasks were run on Condor and the Open Science Grid (OSG). Ghostscript was run to convert each document into a set of png images. Next, OCR tools were executed. Each tool was permitted to run for 24 hours on a document before timeout occurred; a failed document was re-deployed on OSG up to 10 times before being removed from the set. Document failures were caused by kernels older than 2006 and incompatible software on individual OSG machines, as well as document-specific software bugs, such as segmentation faults in Cuneiform. All tools had a failure rate of less than 8%, but these errors are orthogonal to our work and future improvements to them will improve PDD.

The WDS contains 23 times more documents than the ODS, and the number of variables scales approximately linearly. The number of distinct features is only 13 times greater because features can be shared across documents (Table S12). Distinct taxa are only 10 times more numerous in the WDS because many taxa are shared between documents. The number of occurrences is only six times greater, reflecting the fact that most of the additional documents are taxonomically or geologically focused and do not contain fossil occurrences.

4.4 Extensions

We extended PDD to include the German and Chinese languages. The named entity recognition component of PDD has dictionary-based features and NLP-based features. Relevant language-specific dictionaries were built manually and from external sources (e.g., geonames.org). For NLP-based features, the Stanford CoreNLP provides models for Chinese and German. Document layout-based features present no change in function with language.

We also extended PDD to extract body size from biological illustrations. This requires processing images, linking labels to captions, and mapping captions to text. Explanation of the joint image-text analysis is presented in the SI Materials and Methods.

4.5 Assessment

The ODS was randomly split into a training set and a testing set. Fifty documents in the testing set were then randomly sampled for assessment. PDD achieves $\geq$ 92% human-estimated accuracy in all relations (Table S13), which is close to the chosen 95% confidence threshold for data output.

The number of facts recovered vs. the number of facts contained in a document (i.e., recall) is more difficult to assess than precision. Because each extracted relationship consists of a paired object and subject (e.g., the object formation contains a subject taxon), one basic measure of recall is the fraction of all subjects in the PBDB that PDD also recovered. This estimate of recall ranges from 21% to 69%, depending on relation (Table S13). For the lowest recall relations, we randomly sampled 10 documents in order to compare the PBDB and PDD. We did so for a combination of three binary relations (taxon,formation) (formation,temporal) (formation,location). When summarizing this 4-part tuple by taxon, approximately 18% of PDDs extractions also appear in PBDB and 11% of PBDB extractions also appear in PDD. This implies that both PDD and PBDB make recall errors. Further examination of PDD recall errors (Table S6) shows that they can be attributed to OCR-related errors (56%), table recognition failures (29%), and lack of context features (15%). All of these errors correspond to interesting and open-problems for computer science. The first two are related to data acquisition (i.e., how to correctly recognize the structure and content of a document), and the latter is an important natural language inference problem (i.e., how to extract relations by taking advantage of information in the whole document). Continued work in these areas will further improve the PDD system.

5 Acknowledgments

We thank M. Foote for constructive feedback and N.A. Heim and J. Payne for providing body size data. Work partially supported by NSF EarthCube award ACI-1343760 and NSF CAREER IIS-1353606. We also acknowledge the support of the Defense Advanced Research Projects Agency XDATA Program under No. FA8750-12-2-0335 and DEFT Program under No. FA8750-13-2-0039, the Office of Naval Research No. N000141210041 and No. N000141310129, Sloan Research Fellowship, American Family Insurance, Google, and Toshiba. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of these organizations or the US government. This is Paleobiology Database publication 2XX.

References

[1] Valentine JW (1969) Patterns of taxonomic and ecological structure of the shelf benthos during Phanerozoic time. Palaeontology 12:684-709.

[2] Raup DM (1976) Species diversity in the Phanerozoic: a tabulation. Paleobiology 2:279-288.

[3] Bambach RK (1977) Species richness in marine habitats through the Phanerozoic. Paleobiology 3:152-167.

[4] Sepkoski JJ Jr (1981) A factor analytic description of the phanerozoic marine fossil record. Paleobiology 7:36-53.

[5] Sepkoski JJ Jr (1998) Rates of speciation in the fossil record. P Trans R Soc B 353:315-326.

[6] Sepkoski JJ Jr (2002) A compendium of fossil marine genera, Bulletins of American Paleontology 363:1-560.

[7] Raup DM, Sepkoski JJ Jr (1982) Mass extinctions in the marine fossil record. Science 215:1501-1503.

[8] Benton MJ (1995) Diversification and extinction in the history of life. Science 268:52-58.

[9] Alroy J (1998) Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280:731-734.

[10] Jablonski D (2005) Mass extinctions and macroevolution. Paleobiology 31:192-210.

[11] Jablonski D, Roy K, Valentine JW (2006) Out of the tropics: Evolutionary dynamics of the latitudinal diversity gradient. Science 314:102-106.

[12] Kiessling W (2005) Long-term relationships between ecological stability and biodiversity in Phanerozoic reefs. Nature 433:410-413.

[13] Alroy J (2008) Dynamics of origination and extinction in the marine fossil record. P Natl Acad Sci USA 105:11536-11542.

[14] Alroy J (2010) The shifting balance of diversity among major marine animal groups. Science 329:1191-1194.

[15] Miller AI, Foote M (2003) Increased longevities of post-Paleozoic marine genera after mass extinctions. Science 302:1030-1032.

[16] Quental TB, Marshall CR (2013) How the Red Queen Drives Terrestrial Mammals to Extinction. Science 341:290-292.

[17] Williams JW, Shuman BN, Webb T III, Bartlein PJ, Leduc PL (2004) Late-quaternary vegetation dynamics in north america: scaling from taxa to biomes. Ecological Monographs 74:309-334.

[18] Finnegan S, Heim NA, Peters SE, Fischer WW (2012) Climate change and the selective signature of the Late Ordovician mass extinction. P Natl Acad Sci USA 109:6829-6834.

[19] Blois JL, Zarnetske PL, Fitzpatrick MC, Finnegan S (2013) Climate Change and the Past, Present, and Future of Biotic Interactions. Science 341:499-504.

[20] Harnik PG et al. (2012) Extinctions in ancient and modern seas. TREE 27:608-617.

[21] Alroy J et al. (2008) Phanerozoic trends in the global diversity of marine invertebrates. Science 321:97-100.

[22] Alroy J et al. (2001) Effects of sampling standardization on estimates of Phanerozoic marine diversification. P Natl Acad Sci USA 98:6261-6266.

[23] Ferrucci DA, et al. (2010) Building Watson: An overview of the deepqa project. AI Magazine 31:59-79.

[24] Murphy K (2013) From big data to big knowledge, CIKM 13 (ACM, New York, NY, USA), pp 1917-1918.

[25] Suchanek FM, Sozio M, Weikum G (2009) SOFIE: A Self-organizing Framework for Information Extraction, WWW 09 (ACM, New York, NY, USA), pp 631-640.

[26] Carlson A, et al. (2010) Toward an Architecture for Never-Ending Language Learning. AAAI-10 1306-1313.

[27] Kumar A, Niu F, Ré C (2013) Hazy: making it easier to build and maintain big-data analytics. Commun. ACM 56:40-49.

[28] Getoor L, Taskar B (2007) Introduction to Statistical Relational Learning (The MIT Press, Cambridge).

[29] Krishnamurthy R, et al. (2009) Systemt: A system for declarative information extraction. SIGMOD Rec. 37:7-13.

[30] Li Y, Reiss FR, Chiticariu L (2011) SystemT: A Declarative Information Extraction System, HLT 11 (Association for Computational Linguistics, Stroudsburg, PA, USA), pp 109-114.

[31] Niu F, Recht B, R C, Wright SJ (2011) Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent pp 693-701.

[32] Liu J, et al. (2014) An asynchronous parallel stochastic coordinate descent algorithm. ICML.

[33] Zhang C, Ré C (2013) Towards High-throughput Gibbs Sampling at Scale: A Study Across Storage Managers, SIGMOD 13 (ACM, New York, NY, USA), pp 397-408.

[34] Recht B, Ré C (2012) Toward a noncommutative arithmetic-geometric mean inequality: conjectures, case-studies, and consequences. JMLR 23:11.1-11.24.

[35] Niu F, Ré C, Doan A, Shavlik J (2011) Tuffy: Scaling up statistical inference in markov logic networks using an rdbms. Proc. VLDB Endow. 4:373-384.

[36] Suciu D, Olteanu D, R C, Koch C (2011) Probabilistic Databases, Synthesis Lectures on Data Management (Morgan & Claypool).

[37] Foote M (2000) Origination and extinction components of taxonomic diversity: general problems. Paleobiology 26:796-796.

[38] Miller AI, Foote M (1996) Calibrating the Ordovician Radiation of marine life: implications for Phanerozoic diversity trends. Paleobiology 22:304-309.

[39] Alroy J (2010) Geographical, Environmental and Intrinsic Biotic Controls on Phanerozoic Marine Diversification. Palaeontology 53:1211-1235.

[40] Sepkoski JJ Jr (1993) 10 years in the library - new data confirm paleontological patterns. Paleobiology 19:43-51.

[41] Adrain JM, Westrop SR (2000) An empirical assessment of taxic paleobiology. Science 289:110-112.

[42] Ausich WI, Peters SE (2005) A revised macroevolutionary history for Ordovician-Early Silurian crinoids. Paleobiology 31:538-551.

[43] Van Noorden R (2014) Elsevier opens its papers to text-mining. Nature 506:17-17.

[44] Raup DM (1976) Species diversity in the Phanerozoic: an interpretation. Paleobiology 2:289-297.

[45] Smith AB (2001) Large-scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. P Trans R Soc B 356:351-367.

[46] Peters SE, Foote M (2001) Biodiversity in the Phanerozoic: a reinterpretation. Paleobiology 27:583-601.

[47] J. T. Bonner, Why Size Matters: From Bacteria to Blue Whales (Princeton Univ. Press, Princeton, 2006).

[48] Payne JL et al. (2009) Two-phase increase in the maximum size of life over 3.5 billion years reflects biological innovation and environmental opportunity. P Natl Acad Sci USA 106:24-27.

[49] Novack-Gottshall PM (2008) Ecosystem-wide body-size trends in Cambrian-Devonian marine invertebrate lineages. Paleobiology 34:210-228.

[50] Finnegan S, Droser ML (2008) Body size, energetics, and the Ordovician restructuring of marine ecosystems. Paleobiology 34:342-359.

[51] Heim NA et al. (in review) Copes Rule in solitary marine bilaterian animals across the past 540 million years.

[52] Kaesler, RL ed (1997) Treatise on Invertebrate Paleontology, Part H, Brachiopoda (Univ. Kansas, Lawrence).

[53] Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. 1:1-305.

[54] Callison-Burch C, Dredze M (2010) Creating Speech and Language Data with Amazons Mechanical Turk, CSLDAMT 10 (Association for Computational Linguistics, Stroudsburg, PA, USA), pp 1-12.

[55] Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant Supervision for Relation Extraction Without Labeled Data, ACL 09 (Association for Computational Linguistics, Stroudsburg, PA, USA), pp 1003-1011.

[56] Hoffmann R, Zhang C, Weld DS (2010) Learning 5000 Relational Extractors, ACL 10 (Association for Computational Linguistics, Stroudsburg, PA, USA), pp 286-295.

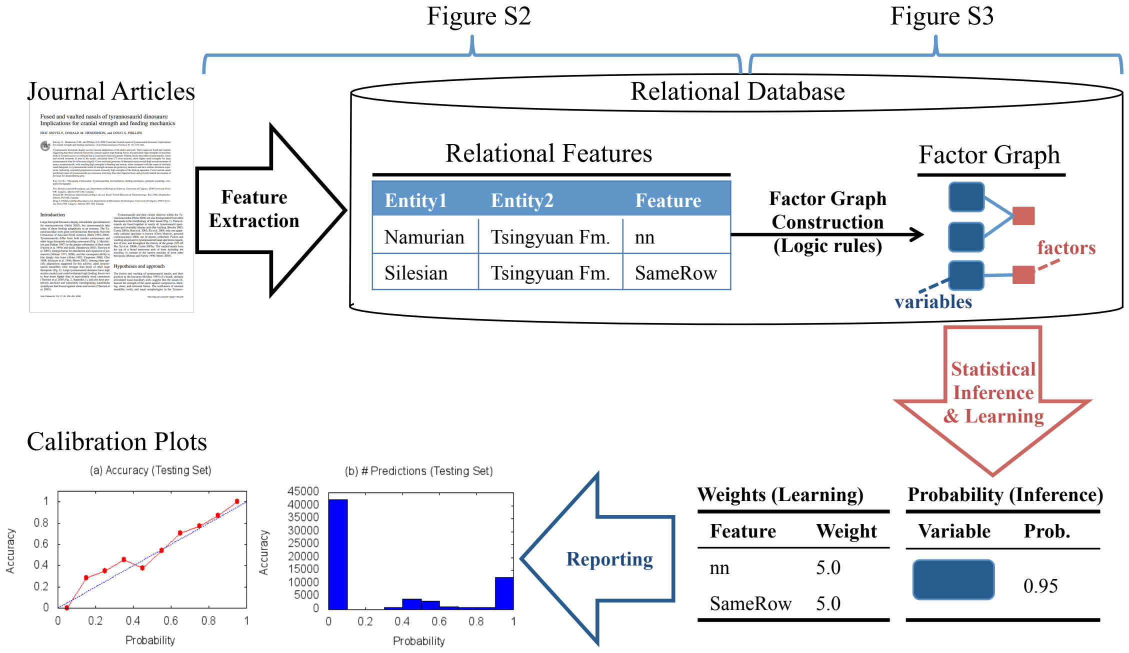
[57] Köpcke H, Thor A, Rahm E (2010) Evaluation of entity resolution approaches on real-world match problems. Proc. VLDB Endow. 3:484-493.

SI MATERIALS AND METHODS

A machine-compiled macroevolutionary history of Phanerozoic life

Shanan E. Peters [*], Ce Zhang [†], Miron Livny [†], and Christopher Ré [‡]

[*] Department of Geoscience, University of Wisconsin-Madison, Madison, WI, 53706 USA
[†] Department of Computer Science, University of Wisconsin-Madison, Madison, WI, 53706 USA
[‡] Department of Computer Science, Stanford University, Stanford, CA 94305 USA

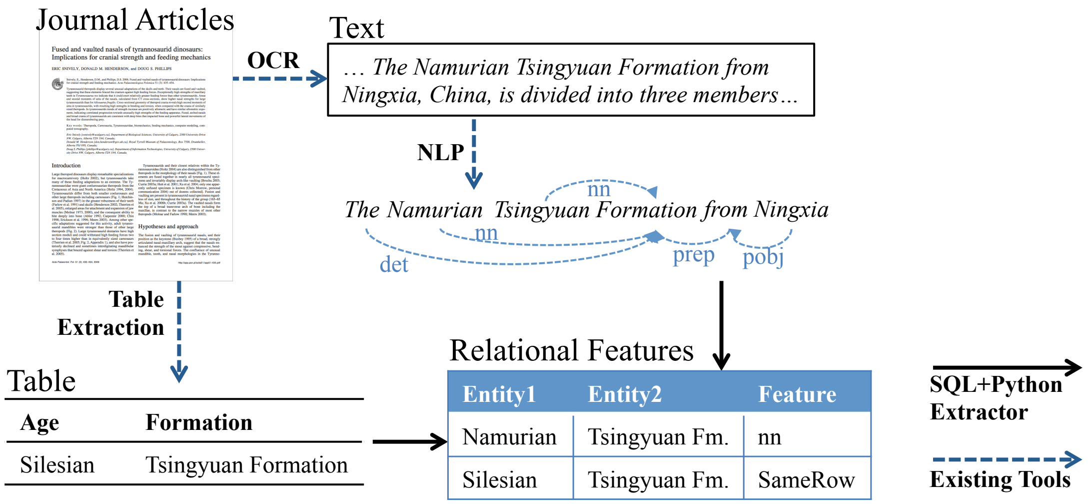


Entity1	Entity2	Feature
Namurian	Tsingyuan Fm.	nn
Silesian	Tsingyuan Fm.	SameRow

Feature	Weight
nn	5.0
SameRow	5.0

Variable	Prob.
▬	0.95

Fig. S1. Schematic representation of the PDD workflow.



… The Namurian Tsingyuan Formation from Ningxia, China, is divided into three members…

The Namurian Tsingyuan Formation from Ningxia

Age	Formation
Silesian	Tsingyuan Formation

Entity1	Entity2	Feature
Namurian	Tsingyuan Fm.	nn
Silesian	Tsingyuan Fm.	SameRow

Fig. S2. Overview of PDD feature extraction. Text, tables, and images in an original document are parsed (e.g., by table position extraction or natural language). Two or more entities and the specific properties in the document (i.e., features) that relate them are expressed as a row in a database.

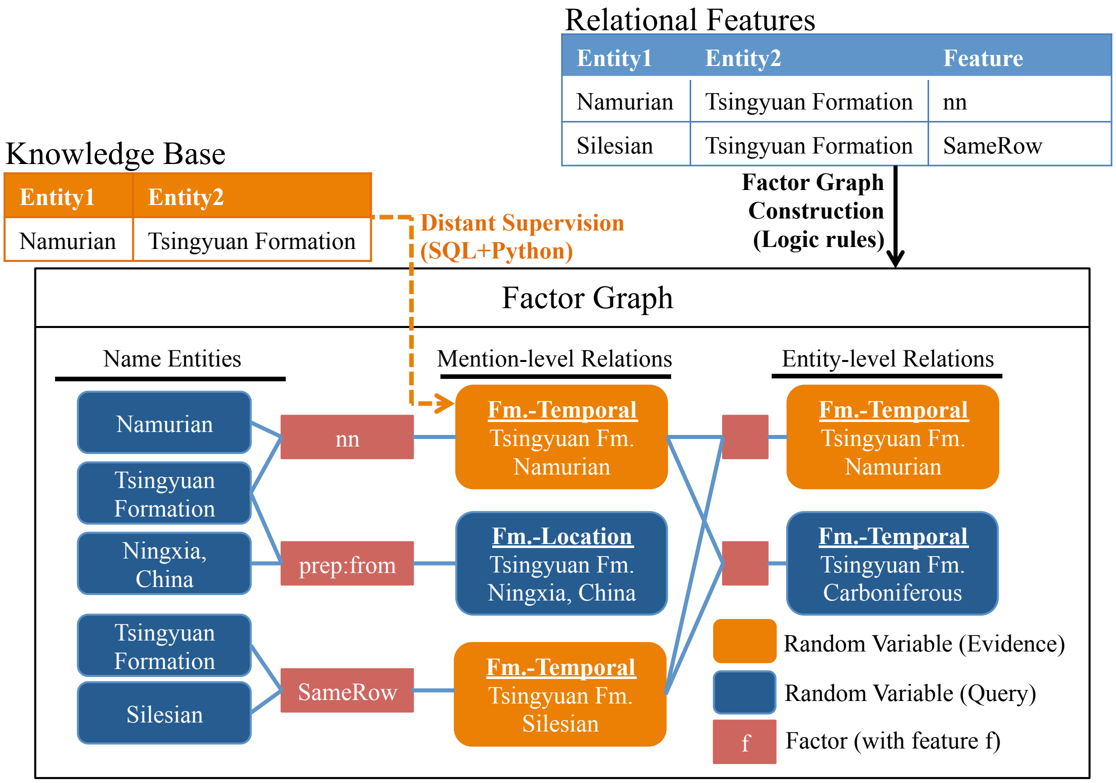
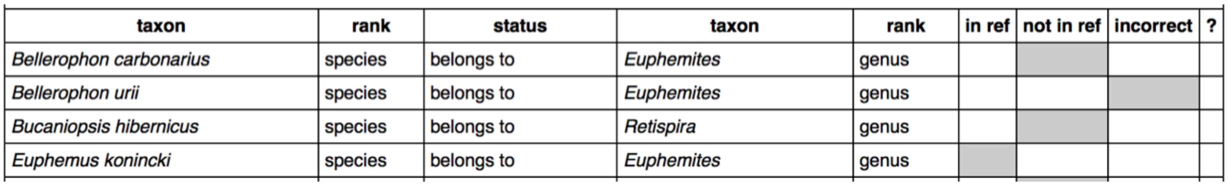


Fig. S3. Overview of factor graph component of PDD. Existing knowledge bases, such as data in the PBDB, are used to assess mention-level relations during distant supervision. Variables assessed for accuracy become evidence variables for statistical inference and learning steps.



taxon	rank	status	taxon	rank	in ref	not in ref	incorrect	?
Bellerophon carbonarius	species	belongs to	*Euphemites*	genus		▨		
Bellerophon urii	species	belongs to	*Euphemites*	genus			▨	
Bucaniopsis hibernicus	species	belongs to	*Retispira*	genus		▨		
Euphemus koníncki	species	belongs to	*Euphemites*	genus	▨			

Fig. S4. Screen shot of web user interface used in blind experiment conducted by 7 human annotators. A unique link and instructions to complete the form were emailed to each participant. The wording of the instructions was as follows:

1. *"in ref" means you can find this *exact* fact in the document somewhere.*
2. *"not in ref" means you can't find the exact fact in the document anywhere (can include typos).*
3. *"incorrect" means it is an incorrect fact (e.g., wrong assignment/relationship, etc.).*
4. *"?" means you don't understand the fact in relation to document.*

Simply clicking on the box selects it for you. You can change it etc. as you go along. Once you are done, you can go to another ref by clicking on bottom. You can come back to the ref and inspect it to make sure it looks good, change things.

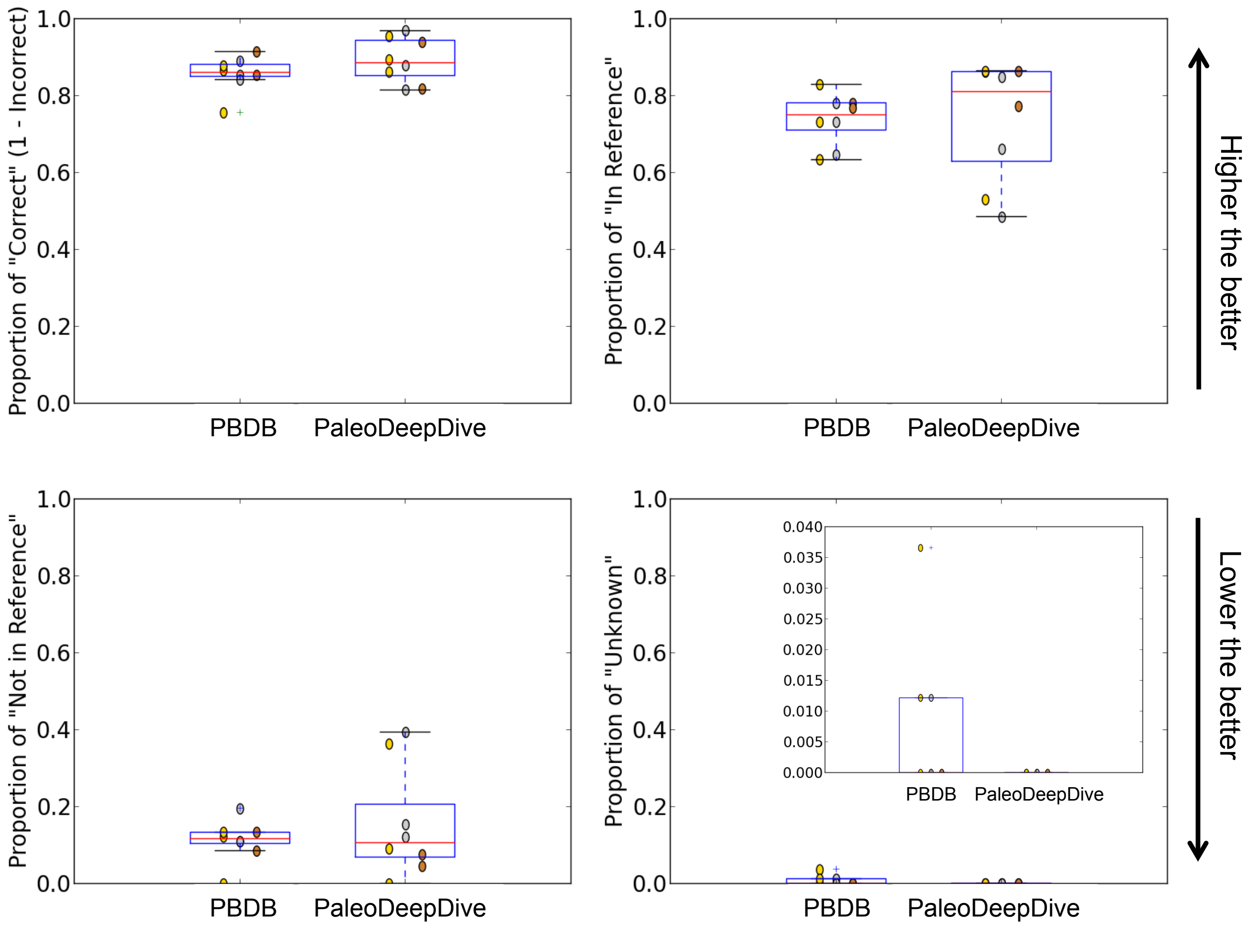


Fig. S5. Summary of results of annotation experiment of PDD and PBDB taxonomic extractions. Yellow, annotators with heavy PBDB governance involvement; blue, past governance involvement; red, graduate students.

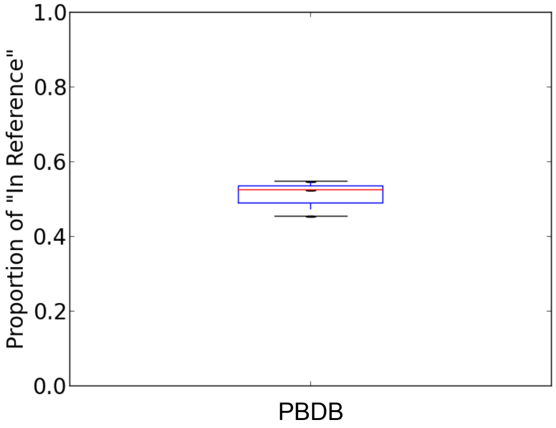


Fig. S6. Summary of results of annotation experiment of occurrence data, or (taxon, geologic unit, temporal interval) tuples in human-constructed PBDB. Results are for 3 volunteers, one from each of groups in Figure S4.

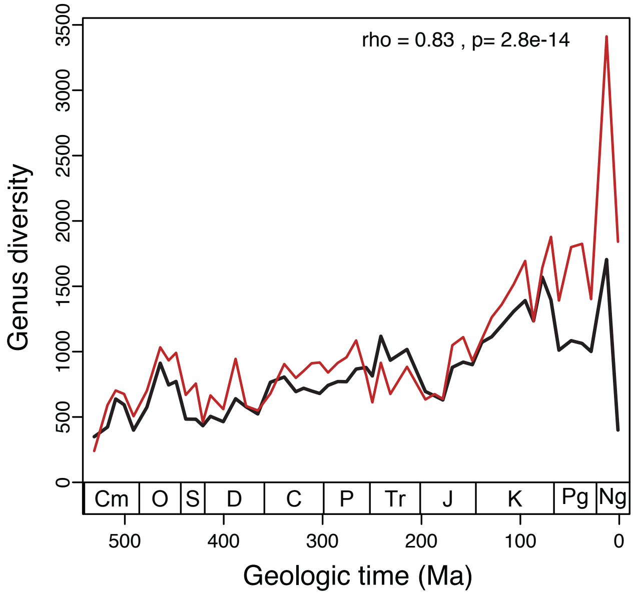
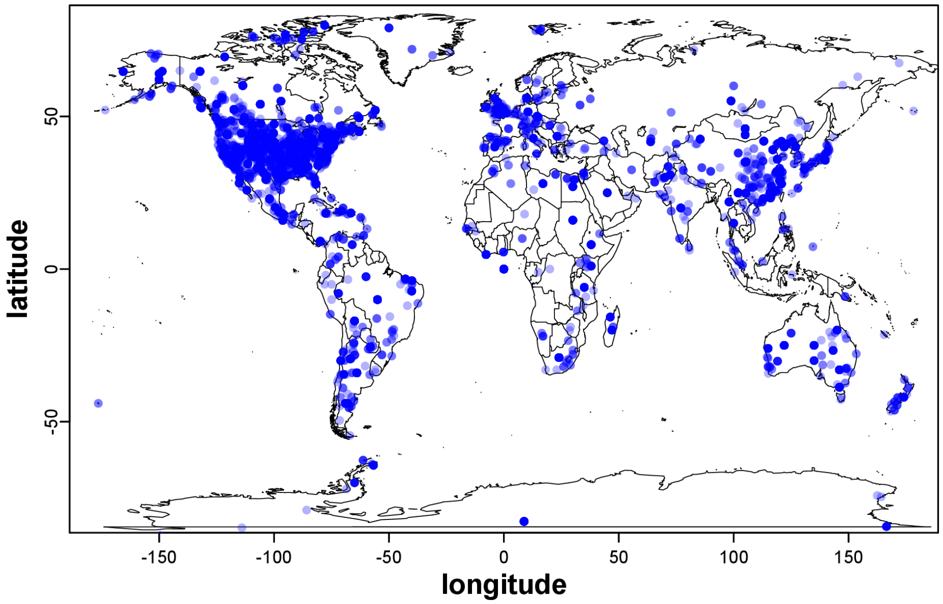
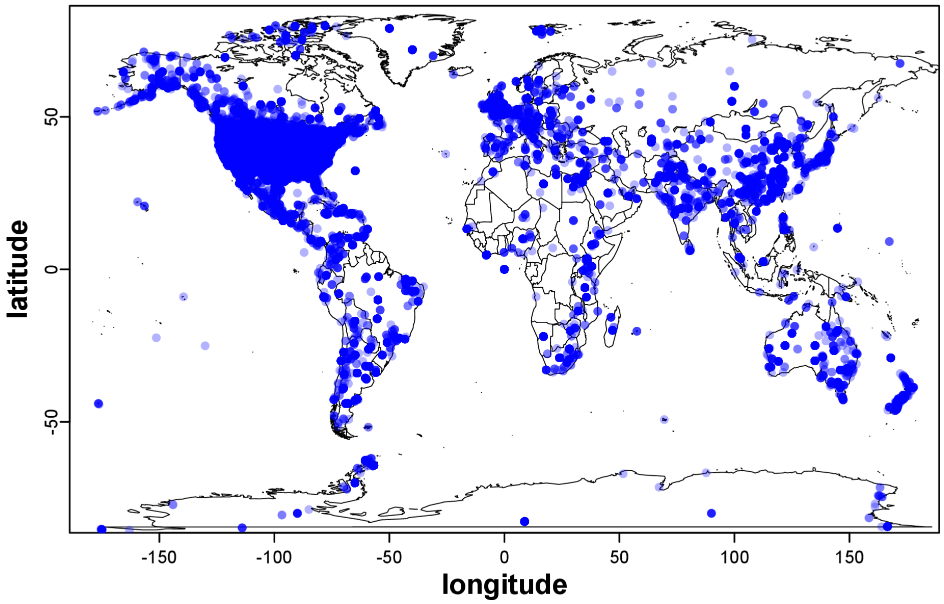


Fig. S7. PDD genus-level diversity (black curve) calculated using occurrences with period level or finer temporal resolution, as opposed to epoch or finer temporal resolution used in Fig. 1. The red curve shows PBDB data and is identical to the red curve in Fig. 1c.



(a) Overlapping Corpus



(b) Whole Corpus

Fig. S8. Geographic distribution of PDD-generated database. Top, location of occurrences in overlapping document set (ODS). Bottom, location of occurrences in whole document set (WDS).

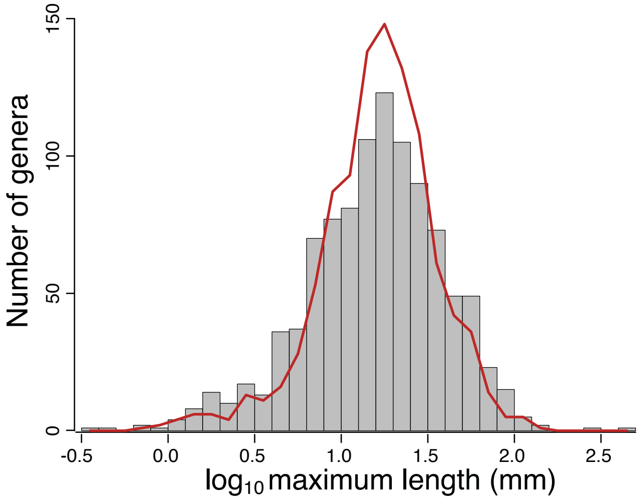


Fig. S9. Frequency distributions of paired estimates of body size for 1,014 brachiopod genera. PDD, gray bars; human estimate, red line (51). Distributions not significantly different according to paired Mann-Whitney U-test ($p = 0.18$) and Kruskal-Wallis test ($p = 0.64$).

Layer	Features
Name Entities	Dictionary (English dictionary, GeoNames, PaleoDB, Species2000, Microstrat, MySQL stop words)
	Part-of-speech tag from StanfordCoreNLP
	Name-entity tag from StanfordCoreNLP
	Name entity mentions in the same sentences (paragraphs, or documents)
Mention-level Relations	Word sequence between name entities
	Dependency path between name entities
	Name-entity tag from StanfordCoreNLP
	Table caption-content association
	Table cell-header association
	Section headers (for Taxonomy)
Entity-level Relations	Temporal interval containment (e.g., Namurian $\subseteq$ Carboniferous)
	Location containment (e.g., Ningxia, China $\subseteq$ China)
	One formation does not likely span $>$ 200 million years

Table S1. List of features and rules used in the current verison of PDD. Finding the right simple features and rules can be difficult. The PDD system is designed to operate in an iterative fashion, with error analysis occurring after each round of feature and rule definition.

Relation	Tuple in Knowledge	Positive Examples	Negative Examples
Taxonomy	(Taxon, Taxon) (t_1, t_2)	$\{(t_1, t_2)\}$	$\{(t_1, t_2') : t_2' \neq t_2\}$
Formation	(Taxon, Formation) (t, f)	$\{(t, f)\}$	Positive examples of other relations
Formation-Temporal (Mention)	(Formation,Interval) (t, i)	$\{(t, i') : intersect(i, i')\}$	$\{(t, i') : \neg intersect(i, i')\}$
Formation-Temporal (Entity)	(Formation,Interval) (t, i)	$\{(t, i') : intersect(i, i') \wedge \neg contain(i', i)\}$	$\{(t, i') : \neg intersect(i, i')\}$
Formation-Location (Mention)	(Formation,Location) (t, l)	$\{(t, l') : intersect(l, l')\}$	$\{(t, l') : \neg intersect(l, l')\}$
Formation-Location (Entity)	(Formation,Location) (t, l)	$\{(t, l') : intersect(l, l') \wedge \neg contain(l', l)\}$	$\{(t, l') : \neg intersect(l, l')\}$

Table S2. List of distant supervision rules used in PDD. Function $contain(x, y)$ and $intersect(x, y)$ return True if the interval (or locations) x contains or intersects with y.

Journal Name	**PBDB**	**PDD Overlapping Set**	**Coverage**
Journal of Paleontology	2,667	2,534	95%
Journal of Vertebrate Paleontology	1,909	1,292	68%
Palaeontology	879	748	85%
Paleontological Journal	849	0	0%
American Museum Novitates	513	433	84%
NULL	509	0	0%
Acta Palaeontologica Polonica	483	433	90%
Nature	452	340	75%
Cretaceous Research	424	421	99%
Gobios	423	296	70%
Ameghiniana	394	21	5%
Canadian Journal of Earth Sciences	336	281	84%
Palaeogeography, Palaeoclimatology, Palaeoecology	325	317	98%
Vertebrata PalAsiatica	322	203	63%
Science	309	184	60%
Bulletin of the American Museum of Natural History	293	214	73%
Geological Magazine	269	24	9%
Alcheringa	268	0	0%
American Journal of Science	257	53	21%
Palaeontologische Zeitschrift	241	0	0%
Journal of Mammalogy	234	147	63%
Acta Palaeontologica Sinica	232	3	1%
United States Geological Survey Professional Paper	231	156	68%
Zoological Journal of the Linnean Society	203	200	99%
Contributions from the Museum of Paleontology, University of Michigan	195	174	89%
Palaeontographica Abteilung A	194	0	0%
Facies	187	0	0%
Lethaia	183	178	97%
Quarterly Journal of the Geological Society of London	180	122	68%
Zootaxa	180	0	0%
Palaios	174	164	94%
Annals of Carnegie Museum	172	25	15%
Proceedings of the United States National Museum	149	0	0%
Neues Jahrbuch fr Geologie und Paleontologie, Abhandlungen	147	0	0%
Review of Palaeobotany and Palynology	147	146	99%
American Journal of Botany	147	87	59%
Proceedings of the Academy of Natural Sciences of Philadelphia	142	40	28%
Journal of Human Evolution	135	122	90%
Proceedings of the National Academy of Sciences	133	51	38%
Journal of Systematic Palaeontology	132	27	20%
Geodiversitas	131	0	0%
Acta Geologica Sinica	130	78	60%
Bulletins of American Paleontology	129	0	0%
Bulletin de la Societe Geologique de France	122	0	0%
Palontologische Zeitschrift	115	0	0%
Rivista Italiana di Paleontologia e Stratigrafia	115	0	0%
Psyche	111	1	1%
Annals of the South African Museum	104	0	0%
Tulane Studies in Geology and Paleontology	103	0	0%
Paleontological Research	102	92	90%
Other Sources	30,851	2,175	7%
Total	47,632	11,782	25%

Table S3. Distribution of documents in the overlapping document set. "NULL" corresponds to a NULL title document type field in the PBDB.

Taxon Name	Rank	Not Found on Google (Error Candidate)
Cirquella espinata	species	
Echinophyllia orpheensis	species	
Fenestella huascatayana	species	
Epigondolella primitia	species	
Palaeospheniscus gracilis.	species	
Pygurus carinatus	species	×
Arionellus tripunctatus	species	
Phacostylus amphistylus	species	
Circotheca multisulcatus	species	
Aulotortus praegaschei	species	
Leptaena demissa	species	
Xinjiangchelys laticentralis	species	
Conotreta lanensis	species	×
Martellia ichangensis	species	
Procavia antiqua	species	
Chermidae	family	
Monophyllus cubanus	species	
Gazella soemmeringi	species	
Pinna subspatulata	species	
Polacanthus faxi	species	×
Homotherium latidens	species	
Platanus primaeva	species	
Rhopalocanium satelles	species	
Cryptobairdia forakerensis	species	
Naiadites elongata	species	
Staurocephalus murchisoni	species	
Serpula anguinus	species	
Glycymeris angusticostata	species	
Eomunidopsis eutecta	species	
Actinocrinites gibsoni	species	
Zhelestes tes	species	×
Spinocyrtia ascendens	species	
Belemnopsis alexandri	species	
Agaricocrinus nodulosus	species	
Oreochromis shiranus	species	
Atrichornithidae	family	
Neltneria jaqueti	species	
Eurydice affinis	species	
Nummulites burdi	species	
Diacalymene marginata	species	
Scapteriscus didactylus	species	
Enhydriodon campanii	species	
Offneria nicoli	species	×
Propetrosia pristina	species	
Podocarpus campbelli	species	
Graffhamicrinus aristatus	species	
Productina sampsoni	species	
Bufina bicornuta	species	
Coccolithus staurion	species	
Ernanodon vas	species	×

Table S4. Error Analysis of Taxon Entity Extractions in PDD

Reference No.	Genus	Correct	Extracted by PBDB
28945	Acrodenta	✓	
	Mastodonsaurus	✓	
	Mesodapedon	✓	
	Rhynchosaurus	✓	✓
	Scaphonyx	✓	
	Spirorbis	✓	
	Stenaulorhynchus	✓	
34109			
28146			
38697	Hazelia	✓	✓
	Leptomitus	✓	✓
32675			
33994	Gastropoda		
	Heterostropha	✓	
	Mathilda	✓	
	Mollusca	✓	
	Stenoglossa	✓	
27115			
41374	Archaeopterodactyloidea	✓	
	Beipiaopterus	✓	
	Boreopteridae	✓	
	Boreopterus	✓	
	Eopteranodon	✓	
	Eosipterus	✓	
	Feilongus	✓	
	Gegepterus	✓	
	Moganopterus	✓	✓
	Ningchengopterus	✓	
	Ornithocheiroidea	✓	
	Zhenyuanopterus	✓	
12054			
13061	Bactrosaurus	✓	
	Dyoplosaurus	✓	
	Gorgosaurus	✓	
	Hypacrosaurus	✓	
	Mandschurosaurus	✓	✓
	Nodosauridae	✓	✓
	Tanius	✓	
Human Recall			18%

Table S5. Error Analysis: PDD Extractions

Reference No.	Genus	Correct	Extracted by PDD	Error Reason
28945	Rhynchosaurus	✓	✓	
34109	Austromola	✓		Not enough context features
	Odontoceti	✓		Not enough context features
28146	Cerapoda	✓		Not enough context features
38697	Hazelia	✓	✓	
	Leptomitus	✓	✓	
	Protospongia	✓		Not enough context features
32675	Tommotia	✓		Not enough context features
33994	Anticonulus	✓		Table recognition failure
	Ataphrus	✓		
	Austriacopsis	✓		
	Discohelix	✓		
	Emarginula	✓		
	Eucyclidae	✓		
	Eucyclus	✓		
	Guidonia	✓		
	Neritopsis	✓		
	Plectotrochus	✓		
	Proacirsa	✓		
	Pseudorhytidopilus	✓		
27115	Astreptodictya	✓		OCR error
	Athrophragma	✓		
	Batostoma	✓		
	Bryozoa	✓		
	Bythopora	✓		
	Calopora	✓		
	Coeloclema	✓		
	Constellaria	✓		
	Contexta	✓		
	Diploclema	✓		
	Echinodermata	✓		
	Graptodictya	✓		
	Helopora	✓		
	Nicholsonella	✓		
	Ottoseetaxis	✓		
	Pachydictya	✓		
	Phylloporina	✓		
	Porifera	✓		
	Prasopora	✓		
	Spongiostroma	✓		
	Stictopora	✓		
	Stictoporella	✓		
	Trilobita	✓		
41374	Moganopterus	✓	✓	
12054	Neosaurus	✓		Not enough context features
13061	Mandschurosaurus	✓	✓	
	Nodosauridae	✓	✓	
PDD Recall			11%	

Table S6. Error Analysis: PBDB Extractions

Relation	**PBDB**	**PDD**	$p = 0.05$
Taxonomy	92%	97%	0
Temporal	89%	96%	+
Location	90%	92%	0
Formation	84%	94%	+

Table S7. Comparison of Accuracies of PDD and PBDB. The column $p = 0.05$ is the significant test of one-tail Welch's t-test, where "+" means significant given the corresponding p-value, and "0" otherwise. The value 0.05 is picked by following the default setting of R.

Journal Name	**1845- -1959**	**1960- -1969**	**1970- -1979**	**1980- -1989**	**1990- -1999**	**2000- -2009**	**2010 -2013**	**Total**
American Journal of Science	2489		727	41		245	138	3640
American Midland Naturalist	2893	1022	1149	989	852	842	189	7936
American Museum Novitates	1974	413	288	272	320	388	98	3753
Annales de Palontologie					29	206	73	308
Annals of Carnegie Museum						82	38	120
Bulletin of the American Museum of Natural History	1318	93	105	72	52	196	65	1901
Comptes Rendus Palevol						679	270	949
Cretaceous Research				287	457	732	393	1869
Geological Journal	136	418	338	1116	680	662	423	3773
Geological Society America Bulletin	276	796		788	1158	1089	486	4593
Geology			1177	2675	2990	3024	1261	11127
Global and Planetary Change				20	469	1070	376	1935
Gobios		13	442	1072	1294	753	167	3741
International Geology Review	87	1482	1780	1541	724	635	353	6602
Journal of Asian Earth Sciences					149	1162	1123	2434
Journal of Geology	5782	736	929	754	671	516	153	9541
Journal of Human Evolution			859	890	759	1067	597	4172
Journal of Mammalogy	3023	1633	1509	1452	1336	1506	438	10897
Journal of Paleontology	2552	1500	1438	1297	1172	2224	643	10826
Journal of South American Earth Sciences				79	423	666	414	1582
Journal of Systematic Palaeontology						113	110	223
Journal of Vertebrate Paleontology				365	636	2152	934	4087
Journal of the Geological Society					329	946	346	1621
Lethaia		104	830	978	992	738	371	4013
Mammalian Species		1	122	224	284	216		847
Marine Micropaleontology			85	262	469	646	156	1618
Micropaleontology	202	375	302	264	270	316		1729
New Zealand Journal of Geology and Geophysics	121	733	730	519	484	403	115	3105
PALAIOS				290	567	677	237	1771
Palaeogeography, Palaeoclimatology, Palaeoecology		191	600	1108	1812	3221	1191	8123
Palaeontology	48	461	477	446	493	1470	560	3955
Palaios						620	287	907
Paleobiology			184	422	337	866	260	2069
Paleontological Research						192	88	280
Palynology			45	140	132	232	119	668
Proc. of AASP			79					79
Proceedings of the Geologists' Association	3514	430	415	416	404	394	273	5846
Quarterly Journal of the Geological Society of London	3063	177	19					3259
Review of Palaeobotany and Palynology		241	427	705	1031	887	406	3697
Revue de Micropaleontologie					104	262	72	438
Rocky			88	118	77	96	33	412
The Micropaleontologist	163							163
Transactions of the Kansas Academy of Science	2107	611	307	263	236	293	48	3865
USGS Open-File Report	403	466	2399	6480	5060	726	243	15777
United States Geological Survey Bulletin	2302	626	320	614	454	1	1	4318
United States Geological Survey Professional Paper	596	721	733	465	227	71	54	2867
Zoological Journal of the Linnean Society	1165	121	363	483	487	638	392	3649
Acta Palaeontologica Polonica	50	118	180	196	242	564	272	1622
Canadian Journal of Earth Sciences		530	1865	1981	1643	1077	377	7473
Oklahoma Geology Notes		15	58	60	56	39	3	231
Vertebrata Palasiatica	136	237	225	333	262	272	119	1584
Biodiversity Heritage Library								97129
Total								**277309**

Table S8. Statistics of Whole Document Set (WDS).

	ODS	**WDS**	**Ratio (WDS/ODS)**
# Variables	13,138,987	292,314,985	22×
# Evidence Variables	980,023	2,066,272	2×
# Factors	15,694,556	308,943,168	20×
# Distinct Features (Weight)	945,117	12,393,865	13×
Documents	11,782	280,280	23×

Table S9. Factor graph statistics in the overlapping and whole document sets. Evidence variables are those variables for which distant supervision has contributed an expectation. The scaling of evidence variables from the ODS to the WDS reflects the fact that most of the training data used by PDD derives from the PBDB data in the ODS.

Year	Volume	Issue	Reference Title
1993	13	3, suppl.	Ontogenetic changes in hind limb proportions within the Ghost Ranch population of Coelophysis bauri
2003	23	3	New dromomerycids (Mammalia: Artiodactyla) from the middle Miocene Sharktooth Hill Bonebed, California, and the systematics of the cranioceratinins
2002	22	3	Paleontology and stratigraphy of the Tecolotlan Basin, Jalisco, Mexico
2004	24	3	A new Miocene sperm whale (Cetacea, Physeteridae) from Virginia
2003	23	3, suppl.	A preliminary Prosauropoda phylogeny with comments on Brazilian basal Sauropodomorpha
2005	25	3	A revised faunal list for the Carmel Church Quarry, Caroline County, Virginia
1994	14	3, suppl.	Preliminary report on the microvertebrate fauna from the Late Cretaceous Bauru strata near Peirpolis, Minas Gerais, Brazil
1993	13	3, suppl.	Sedminentology and taphonomy of the Little Houston Quarry, Morrison Formation (Upper Jurassic), northeast Wyoming
1986	6	3	
2002	22	3	A flora and faunal list of specimens recovered from the Big Pig Dig, Badlands National Park, South Dakota

Table S10. A Random Sample of PBDB References in *Journal of Vertebrate Paleontology* that Do Not Appear in the Overlapping Corpus.

Year	Volume	Issue	Reference Title
1905	22	NULL	
2006	312	NULL	Comment on “The brain of LB1, Homo floresiensis”
1984	224	NULL	
1885	5	116	Lesquereux’s Cretaceous and Tertiary Flora
1991	251	NULL	New fossil evidence on the sister-group of mammals and early Mesozoic faunal distributions
1990	249	NULL	
1900	11	282	The vertebral formula in Diplodocus Marsh
1997	278	NULL	A tribosphenic mammal from the Mesozoic of Australia
1905	22	568	The occurrence of ichthyosaur-like remains in the Upper Cretaceous of Wyoimng
1934	79	2039	A change of names

Table S11. A Random Sample of PBDB References in *Science* that Do Not Appear in the Overlapping Corpus.

		ODS	**WDS**	**Ratio (WDS/ODS)**
Mention-level Candidates	Taxon	6,049,257	133,236,518	22×
	Formation	523,143	23,250,673	44×
	Interval	1,009,208	16,222,767	16×
	Location	1,096,079	76,688,898	76×
	Opinions	1,868,195	27,741,202	15×
	Taxon-Formation	545,628	4,332,132	8×
	Formation-Temporal	208,821	3,049,749	14×
	Formation-Location	239,014	5,577,546	23×
Entity-level Result	Authorities	163,595	1,710,652	10×
	Opinions	192,365	6,605,921	34×
	Collections	23,368	125,118	5×
	Occurrences	93,445	539,382	6×
	Documents	11,782	280,280	23×

Table S12. Extraction statistics for the overlapping and whole document sets. Authorities refers to distinct taxa (identified by name and, optionally, ranks and authors).

Relation	**# Annotations**	**Precision**	**Recall**
Taxonomy	933	97%	39%
Temporal	478	96%	69%
Location	655	92%	36%
Formation	2,271	94%	21%

Table S13. Statistics of Annotations Collected and Quality Score for Each Relation

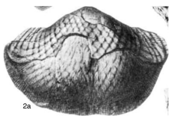
6 Extensions

6.1 Body Size Extraction

In order to extract body size estimates from biological illustrations, we need to extract the relation:

$$(Taxon, FigureName, FigureLabel, Magnification, ImageArea)$$

where $ImageArea$ is a region on the PDF with known DPI so that the actual size of the image on a printed document is known. The following table is an example of the target extracted relation.

Vediproductus wedberensis	Fig. 381	2a	X1	
Compressoproductus compressus	Fig. 382	1a	X0.8	
Devonoproductus walcotti	Fig. 383	1b	X2.0	

There were two steps in the process: (1) Image processing, and (2) text extraction. In PDD, these two components are done jointly in the same factor graph.

Image Processing. The goal of the image processing component is to associate each image area with a figure label. To achieve this, PDD needs to (1) detect image areas and figure labels from PDF documents, and (2) associate image areas with figure labels. Figure S10 illustrates these two steps.

Detection of Image Areas and Figure Labels. The following steps were taken: (1) Edge detection; (2) Watershed Segmentation; (3) Image Dilation; and (4) Connected-component Detection (Figure S10). Standard online-tutorials were followed, with one variant for Image Dilation. In this step, one needs to specify a parameter for dilation. Instead of specifying one value for the parameter, we tried a range of parameters and generate different versions of segmentations. PDD then trained a logistic regression classifier to choose between these segments trained on a human-labeled corpus.

Association of Image Areas with Figure Labels. After recognizing a set of image regions and their corresponding OCR results, PDD attempted to predict the association of figure labels and image areas, as shown in Figure S10. Similar to relation extraction, PDD introduces a Boolean random variable for each label and image area pair. It then builds a logistic regression model using features such as the distance between label and image areas, and whether a label is nearest to an image area and vice versa.

Text Extraction. PDD also extracts information from text, as shown in Figure S11. This extraction phase is similar to what was used when extracting fossil occurrence-related relations. In the name entity recognition component, PDD extracts different types of mentions, including Figure name (e.g., “Fig. 3”), Figure labels (e.g., “3a-c”), Taxon (e.g., “*B. rara*”), and magnitude (e.g., “X1”). Figure S11 shows an example of these mentions (raw text with OCR errors). PDD then extracts relations between these mentions using the same set of features as other diversity-related relations.

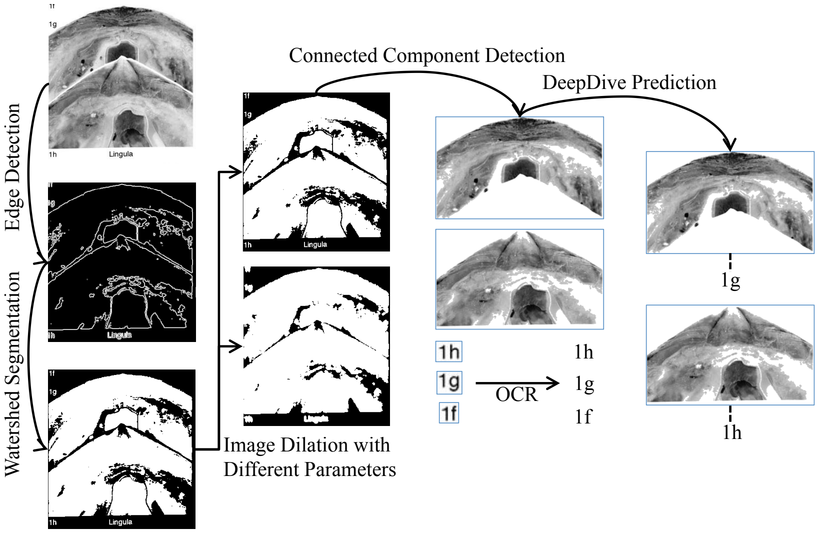


Fig. S10. Image Processing Component for Body Size Extraction. Note that this examples contains the illustration of a *partial* body.

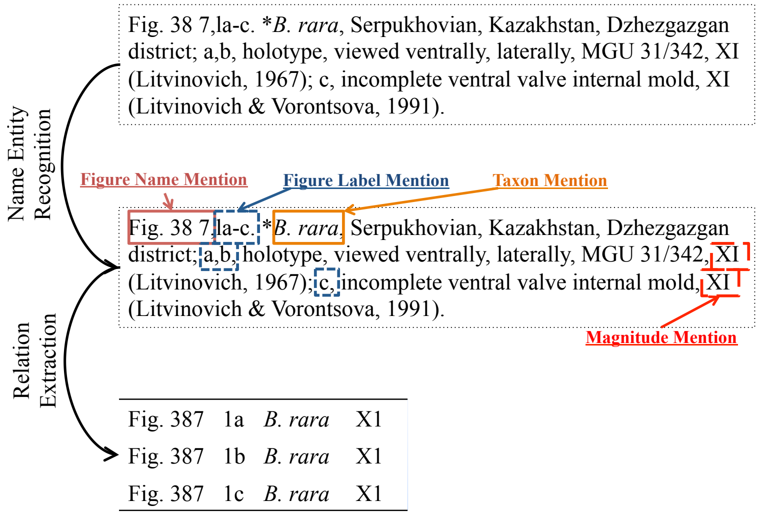


Fig. 38 7,la-c. **B. rara*, Serpukhovian, Kazakhstan, Dzhezgazgan district; a,b, holotype, viewed ventrally, laterally, MGU 31/342, XI (Litvinovich, 1967); c, incomplete ventral valve internal mold, XI (Litvinovich & Vorontsova, 1991).

Fig. 38 7,la-c. **B. rara*, Serpukhovian, Kazakhstan, Dzhezgazgan district; a,b, holotype, viewed ventrally, laterally, MGU 31/342, XI (Litvinovich, 1967); c, incomplete ventral valve internal mold, XI (Litvinovich & Vorontsova, 1991).

Fig. 387	1a	*B. rara*	X1
Fig. 387	1b	*B. rara*	X1
Fig. 387	1c	*B. rara*	X1

Fig. S11. Relation Extraction Component for Body Size Extraction.

Joint Inference. Both the image processing component and the text extraction component results in a factor graph populating two relations with schema

$$(FigureLabel, ImageArea)$$

and

$$(Taxon, FigureName, FigureLabel, Magnitude).$$

PDD joins these two intermediate relations to form a large factor graph to populate the target relation. Joint inference on the whole factor graph is then executed.

6.2 Body Size Extraction Validation

Corpus. Other researchers [18] recently compiled body size measurements by manually measuring illustrations and reading captions in the *Treatise on Invertebrate Paleontology.* Of the 55 volumes now accessible, humans have made measurements from part H, I, K, L, N, O, P, Q, R, S, T, U. We created from these documents the following three sets:

1. **Testing Corpus (With Ground Truth).** Part H.
2. **Testing Corpus (Without Ground Truth).** Part A, B, C, D, E, F, G, W, V.
3. **Training Corpus.** Part I, K, L, N, O, P, Q, R, S, T, U.

We used the Training Corpus to generate training data for distant supervision. We compared our results with those of human annotators using the Testing Corpus (With Ground Truth). The Testing Corpus (Without Ground Truth) shows that PDD helps to extend the body size database with new extractions that are not provided by human annotators.

Results on Testing Corpus (With Ground Truth). PDD is able to to achieve high precision and slightly higher recall than human when extracting body size measurements and their relations.

Precision. We measured the precision of PDD by randomly sampling 100 extracted instances of the target relation and manually annotate those extractions. We find that the accuracy is more than 92%.

Recall. We next counted the number of distinct (genus, figure name, figure label) tuples that are extracted by humans and PDD on the same set of documents. We find that human extracted 4,837 distinct tuples, and PDD extracted 5,783 distinct tuples, or 20% more. The primary reason for the increase is the complete extraction of meaurements for all parts of a figure (e.g., "1a-f"). Humans typically extract only one part.

Although selective data extraction is often a decision made for the sake of expediency and because not all images provide optimal orientations for the dimensions being targeted by a given investigation, extracting complete measurements and associated textual descriptions establishes the foundation for more complete morphometric analyses.

Results on Testing Corpus (Without Ground Truth). PDD is able to extract facts on documents that have not yet been processed by humans. PDD processed Parts A, B, C, D, E, F, G, V, W of the *Treatise on Invertebrate Paleontology*, which have not yet been processed for body size by [18]. PDD extracts 7K distinct (genus, figure name, figure label) tuples from these documents.

6.3 Multi-linguistic Extraction

Corpus. We followed a similar protocol as we used to collect the overlapping corpus for English documents. We identified the top-20 journals ranked by the number of journal articles in PBDB, and attempted to download articles from their web site. Access was limited to *Vertebrata Palasiatica* (Chinese), *Stuttgarter Beitrage zur Naturkunde* (German), and *Eclogae Geologicae Helvetiae* (German). A total of 1,583 Chinese journal articles and 4,393 German journal articles were obtained in this way. We used the same protocol to map these journal articles to articles in PBDB. Of these, there were 47 articles in Chinese and 56 German articles that overlapped with the PBDB.

	English	Chinese	German	Dictionary Source
Rock Formation	Formation	组	Formation	Manual
	Clay	石	Ton	
Temporal Interval	Late Cretaceous	晚白垩世	Oberkreide	Manual
	Cretaceous	白垩世	Kreide	
Location	United States	美国	Vereinigte Staaten	geonames.org
Taxon	*Aeschnidium densum*	*Aeschnidium densum*	*Aeschnidium densum*	All in Latin

Protocol. We compared the extractions of PDD in the overlapping set with the PBDB extractions on the same set of documents. Our way of assessing quality is recall for the tuple

$$(Taxon, TimeInterval)$$

This tuple is language-independent because (1) taxon has unified Latin-representation in all English, Chinese, and German articles; and (2) time Intervals and their hierarchical relationships are known by PDD for all languages. To extract this tuple, PDD requires the information in all other tuples, including $(Taxon, Formation)$, $(Formation, TimeInterval)$, and $(Formation, Location)$. We selected taxa common to both PDD and PBDB, and label PDD's extraction as correct if the taxon temporal ranges overlap.

Recall. From the overlapping corpus, PBDB extracts $(Taxon, TimeInterval)$ tuples for 85 distinct genera in Chinese and 242 distinct genera in German. We find that PDD correctly extracts $(Taxon, TimeInterval)$ for 24 genera (28%) in Chinese and 82 (33%) genera in German. The difference between Chinese and German is caused primarily by OCR quality, even though we used commercial OCR tools for both. Chinese has lower OCR quality because of the large vocabulary in East-Asian languages.

Precision. Out of all 24 distinct genera in Chinese and 82 distinct genera in German articles, we find that all of them overlap with PBDB extractions in terms of their temporal interval, indicating high precision.

7 Specific Technical Validation

Here we describe DeepDive, the underlying system that powers PDD [23,33–35,44,45].

7.1 Probabilistic Framework

7.1.1 Related Work

Knowledge Base Construction (KBC) has been an area of intense study over the last decade [3,4,7,12,21,22,32,36,38,40,43,46]. Within this space, there are a number of approaches.

Rule-based Systems. The earliest KBC systems used pattern matching to extract relationships from text. The most well known example is the "Hearst Pattern" proposed by Hearst [17] in 1992. In her seminal work, Hearst observed that a large amount of hyponyms can be discovered by simple patterns, e.g., "X, such as Y". Hearst's technique forms the basis of many further techniques that attempt to extract high quality patterns from text. In industry, rule-based (pattern-matching-based) KBC systems, such as IBM's SystemT [22,25], have been built to develop high quality patterns. These systems provide the user a (usually declarative) interface to specify a set of rules and patterns to derive relationships. These systems have achieved state-of-the-art quality after carefully engineering effort as shown by Li et al. [25].

Statistical Approaches. One limitation of rule-based systems is that the developer needs to ensure that all rules provided to the system are high precision rules. For the last decade, probabilistic (or machine learning) approaches have been proposed to allow the system select between a range of a priori features automatically. In these approaches, the extracted tuple is associated with a marginal probability that it is true (i.e., that it appears in the KB). DeepDive, Google's knowledge graph, and IBM's Watson are built on this approach. Within this space there are three styles of systems:

- **Classification-based Frameworks** Here, traditional classifiers assign each tuple a probability score, e.g., naïve Bayes classifier, and logistic regression classifier. For example, KnowItAll [12] and TextRunner [3, 43] uses naïve Bayes classifier, and CMUs NELL [4, 7] uses logistic regression. Large-scale systems typically use these types of approaches in sophisticated combinations, e.g., NELL or Watson.
- **Maximum a Posteriori (MAP)** Here, the probabilistic approach is used but the MAP or Most likely world (which do differ slightly) is selected. Notable examples include the YAGO system [21],which uses a PageRank-based approach to assign a confidence score. Other examples include the SOFIE [40] and Prospera [32], which use an approach based on constraint satisfication.
- **Graphical Model Approaches** The classification-based methods ignore the interaction among predictions, and there is a hypothesis that modeling these correlations yields higher quality systems more quickly. A generic graphical model has been used to model the probabilistic distribution among all possible extractions. For example, Poon et al. [36] used Markov logic networks (MLN) [11] for information extraction. Microsoft's StatisticalSnowBall/EntityCube [46] also uses an MLN-based approach. A key challenge with these systems is scalability. For example, Poon et al. was limited to 1.5K citations. Our relational database driven algorithms for MLN-based systems are dramatically more scalable [33].

7.1.2 Calibrated Probabilities

DeepDive takes a Bayesian probabilistic approach to KBC by treating OCR, NLP, image processing, and feature recognition as one joint probabilistic inference problem in which all predictions are modeled as a factor graph (Fig. S3). This probabilistic framework ensures all facts that are produced by DeepDive are associated with a marginal probability.[1] These marginal probabilities are meaningful in DeepDive (i.e., they should correspond to the actual probabilities of a fact beig correct), which provides a mehcanism for evaluation and an aid to improving the system.

Calibration. In DeepDive, *calibration plots* are used as a way to summarize the overall quality of the KBC results. Ideally, the probability associated with a given fact in DeepDive should equal the empirical probability that this fact is correct (i.e., an extraction with a probability 0.95 should be correct with a 95% of the time when inspected in the original source). Because DeepDive uses a joint probability model, any set of predictions can be assigned a marginal probability. Queries can then be against the model to help determine where a model needs improvement.

Figure S12 and Figure S13 show calibration plots for the ODS and the WDS presented in the main text. We will use Figure S12(1) as an example, which is the target relation Taxonomy in the ODS. A calibration plot contains three components: (a) Accuracy, which measures the test-set accuracy of a prediction with a certain probability; (b) # Predictions (Testing Set), which measures the number of extractions in the test set with a certain probability; and (c) # Predictions (Whole Set), which measures the number of extractions in the whole set with certain probability. The difference between test set and whole set is that the former has training labels for each random variable. Results are summarized as histograms, and empirically we find that a bin of size of 0.1 is usually sufficient to understand the behavior of the system.

---

[1] Cox's theorem asserts (roughly) that if one uses numbers as degrees of belief, then one must either use probabilistic reasoning or risk contradictions in a reasoning system, i.e., probabilistic reasoning is the only sound system for reasoning in this manner [20].

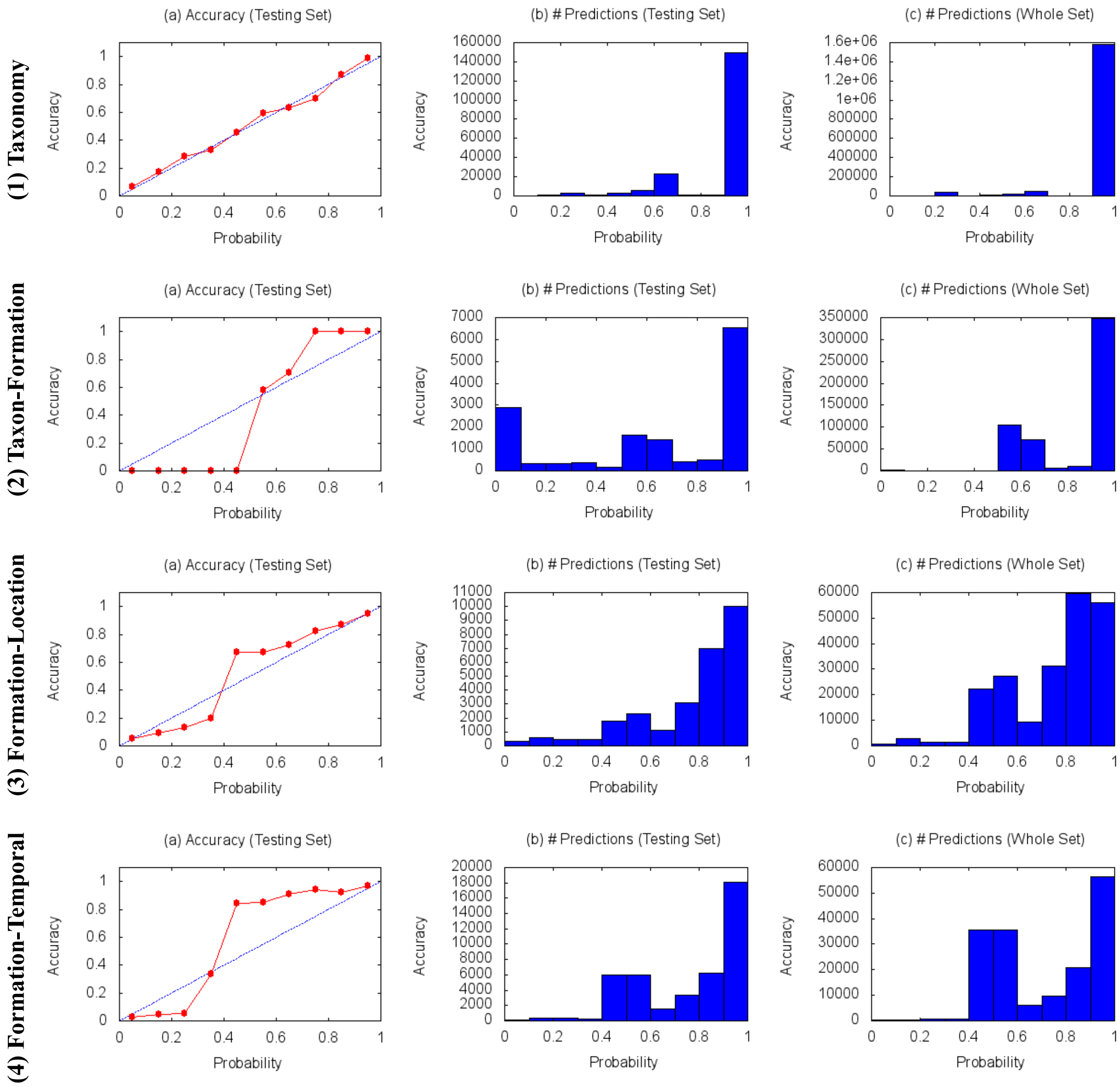


Fig. S12. Calibration Plots for All Relations on Overlapping Corpus

Using Calibration Plots

(a) Accuracy. If the accuracy curve is similar to the ideal (0,0)-(1,1) line, it means that a probability produced by the system matches the *test-set accuracy*. For example, Figure S12(1) shows a reasonably good curve for calibration. Differnces in these two lines can be caused by (1) inefficient training data or a small testing corpus, and/or (2) bad mixing behavior of the sampler or other software bugs. For example, Figure S13(2,3,4) shows a much better calibration behavior than Figure S12(2,3,4), primarily because the former is based on the whole corpus, which has more training data and a larger testing set.

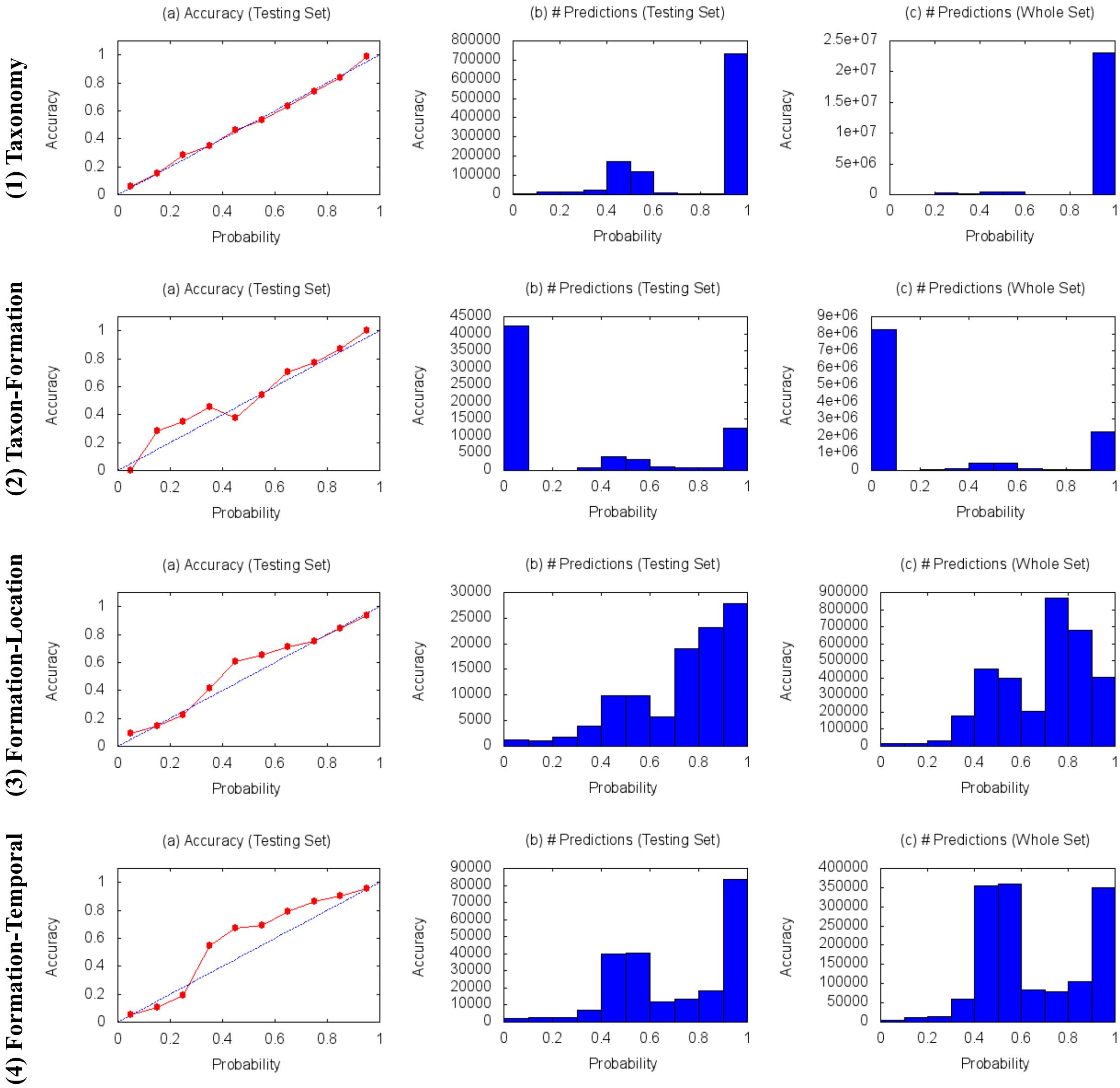


Fig. S13. Calibration Plots for All Relations on Whole Corpus

(b) # Predictions (Testing Set). Ideally, the # Predictions histogram should have a “U” shape. That is, most of the data are concentrate at high probability (where we are confident it is correct) and low probability (where we are confident it is incorrect). Large numbers of predictions with a probability approximately 0.5 means that the system has little information about how to classify these extractions. This implies that more features could be defined to resolve uncertainty. For example, Figure S12(2) shows a U-shape curve with some masses around 0.5-0.6. The shape of the histogram relies on the ratio between the number of positive examples and negative examples. When the number of positive examples dominates negative examples and there is a bias term, it is possible that there are very small amount extractions with a probability near 0. Figure S12(1,3,4) illustrate this phenomenon.

(c) # Predictions (Whole Set). This histogram is similar to (b), but illustrates the behavior of scaling the system to a set of documents for which we do not have any training examples. Usually we hope that (c) has a similar shape to (b).

Usage. The above techniques have proven critical to debugging and improving the quality of PDD. In response to low confidence, a user can provide labeled examples, which allows the system to learn weights that yield higher confidence. Additionally, a user may write logical inference rules that provide ways of improving quality, which is a key component of all statistical relational approaches.

7.2 Declarative Interface for Joint Inference and Rich Features

7.2.1 Related Work

Here we survey recent efforts that focus on how to improve the quality of a KBC system.

Rich Features. Different researchers have recently noted the importance of combining and using a rich set of features and signals to improve the quality of a KBC system. Two famous efforts, the Netflix challenge [6], and IBM's Watson [13], which won the Jeopardy gameshow, have identified the importance of features and signals:

> *Ferrucci et al. [13]: For the Jeopardy Challenge, we use more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique we use is how we combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, or speed.*
>
> *Buskirk [6]: The top two teams beat the challenge by combining teams and their algorithms into more complex algorithms incorporating everybody's work. The more people joined, the more the resulting team's score would increase.*

In both efforts, the rich set of features and signals contributed to the high-quality of the corresponding system. Other researches have found similar phenomena. For example, Mintz et al. [30] finds that although both surface features and deep NLP features have similar quality for relation extraction tasks, combining them achieves a significant improvement over using either one in isolation. Similar "feature-based" approaches are also used in other domains (e.g., Finkel et al. [14] uses a diverse set of features to build a NLP parser with state-of-the-art quality). In our own work [16], we have also found that integrating a diverse set of deep NLP features can improve a table extraction system significantly.

Joint Inference. Another recent trend in building KBC system is to take advantage of *joint inference* [9, 10,16,28,29,35–37]. Different from traditional models [31], such as logistic regression or SVM, joint inference approaches emphasize learning multiple targets simultaneously. For example, Poon et al. [36, 37] find that learning segmentation and extraction in the same Markov logic network significantly improves the quality of information extraction. Similar observations have been made by Min et al. [29] and McCallum [28]. Our recent work also show the empirical improvement of joint inference on the diverse set of tasks, including relation extraction [35] and table extraction [16].

Deep Learning and Joint Inference. A recent emerging effort in the machine learning community is to build a fully-joint model for NLP tasks [9, 10]. The goal is to build a single joint model from the lowest level (e.g., POS tagging) to the highest level (e.g., semantic role labeling). The PDD system is built in a similar spirit that attempts to build a joint model for low-level tasks (e.g., OCR), to high-level tasks (e.g., cross-document inference of relation extraction).

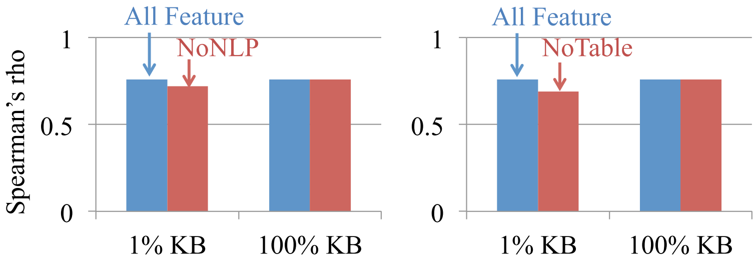


Fig. S14. Lesion Study of Deep NLP Features and Table Recognition

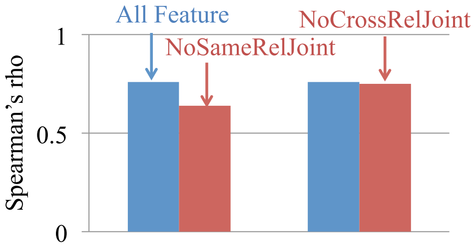


Fig. S15. Lesion Study of Joint Inference

7.2.2 The DeepDive Approach and the Impact of Rich Features and Joint Rules

DeepDive uses joint inference rules and rich features. In this section, we test that these features and rules are important to PDD's quality by conducting a lesion study.

Protocol. All experiments were run on the overlapping corpus as described in the main text. We produced variants of PDD by removing features/rules and all components that rely on the output of the removed feature/rule. We summarize the quality of PDD by computing Spearman's rho for first differences in genus-level biodiversity (as in Fig. 1).

Features. The PDD feature extraction phase extracts a set of features, including deep linguistic features, e.g., dependency parsing results, and vision-based features (e.g., a simple table extractor based on Hough Transform). To study their impact, we conduct lesion study by sequentiallydisabling these features.

Deep NLP Features. Figure S14(a) shows the impact of removing NLP features (e.g., dependency path). If we use the whole PBDB is used, dropping these Deep NLP features does not have a significant effect on Spearman's rho. However, if the knowledge base used for training is reduced to 1% of it s size, then dropping NLP features results in a decrease of Spearman's rho from 0.72 from 0.82.

Vision-based Table Recognition. PDD contains a table recognition component to detect tables using vision-based features (e.g., Hough Transform). When disabling this component and using the 1% PBDB for distant supervision, PDD achieves a Spearman's rho of 0.69. This drop is the effect of decreased recall of data in tables.

Joint Inference Rules. PDD contains a set of factors for joint inference among random variables, as shown in Fig S3. We study their impact on two types of joint inference rules: (1) joint inference within one relation; and (2) joint inference across different relations (Figure S15).

Joint Inference for Same Relations. Disabling all joint inference rules results in a Spearman's rho of 0.64, even when using the whole PBDB knowledge base. This is a marked decline from the Spearman's rho of 0.82 obtained when these rules are enabled. This large decline in quality is caused by the fact that jointly infering the values of random variable results in much higher-quality predictions. For example, assume that we have three candidate facts that Tsingyuan Formation has the age (1) Carboniferous ,(2) Namurian, and (3) Kungurian. In the current PDD system, the higher confidence for Carboniferous will also boost its confidence for Namurian (because of containment), and decrease its confidence for Kungurian (because Kungurian is so much younger than Carboniferous). This type of joint inference between random variables help PDD to produce result with higher recall (by boosting confidence to cross the imposed 0.95 threshold) and precision (by eliminating wrong predictions).

Joint Inference across Relations. The current PDD system has three joint inference rules across different relations (e.g., one geologic formation entity mention cannot be concurrently a location mention). We disable these rules and show in Figure S15 that it does not have a large impact to the overall quality. This implies that the current PDD system is quite modular across different relations. This means that different types of relations can be decoupled and applied to other related applications (e.g., for biology or geology).

7.3 Scalability and High Performance Statistical Inference and Learning

7.3.1 Related Work

There is an emerging trend in both industry and academia to support statistical inference and learning, and we survey these efforts in this section.

Hardware Efficiency. One line of research tries to speed-up statistical inference and learning by better taking advantage of modern hardware and clusters. For example, many industrial database vendors have integrated statistical analytics components into their product. For example, Oracle's ORE [2], Pivotal's MADlib [19], and IBM's SystemML [15]. These systems provide functionalities like logistic regression and collapsed Gibbs sampling for topic modeling on their data management systems. There are also efforts to design new data processing framework instead of relying on the traditional database systems. Indeed, most data processing frameworks developed in the last few years are designed to support statistical analytics including Mahout [1] for Hadoop, MLI for Spark [39], GraphLab [27], GraphChi [24], and Delite [8, 41]. These systems have been shown to increase the performance of corresponding statistical analytics tasks significantly.

Statistical Efficiency. One key difference between statistical inference and learning with traditional SQL-like analytics is that different ways of executing the same tasks usually lead to different speed when converging to the same quality. Therefore, another line of related work, mainly contributed by the mathematical optimization and machine learning community, is to design more efficient algorithms for statistical inference tasks. One of the recent trends is to design lock-free algorithms that can be executed on the emerging multi-socket multi-core machines with high parallelism [5, 26, 33, 42, 47]. For example, Tsitsiklis et al. [42] proves asymptotic convergence for a parallel coordinate descent algorithm, and Bradley et al. [5] proves the convergence rate and theoretical speedups for parallel stochastic coordinate descent. Our own work [26, 33] proves the convergence of lock-free execution for stochastic gradient descent and stochastic coordinate descent.

7.3.2 The DeepDive Approach and The Performance of PDD

The DeepDive Approach. The statistical inference and learning engine in DeepDive [44] is built upon the challenge of designing a high-performance statistical inference and learning engine on a single machine [26, 34, 44, 45]. Compared to traditional work, the main novelty of DeepDive is that it considers *both* hardware efficiency and statistical efficiency for executing an inference and learning task.

Hardware Efficiency. DeepDive takes into consideration the architecture of modern non-uniform memory access (NUMA) machines. A NUMA machine usually contains multiple nodes (sockets), where each sockets contains multiple CPU cores. To achieve high hardware efficiency, it is useful to decrease the communication across different NUMA nodes.

Statistical Efficiency Pushing hardware efficiency to the extreme might cause statistical efficiency to suffer because the lack of communication between nodes could decrease the rate of convergence of a statistical inference and learning algorithm. DeepDive takes advantage of theoretical results of model averaging [47] and lock-free execution [26, 34].

Performance of Statistical Inference and Learning. DeepDive enables PDD's ability to run statistical inference and learning efficiently. For example, on the whole corpus, the factor graph contains more than 0.2 billion random variables and 0.3 billion factors. On this factor graph, DeepDive is able to run Gibbs sampling on a machine with 4 sockets (10 core per sockets), and we find that we can generate 1,000 samples for all 0.2 billion random variables in 28 minutes.

References

[1] Apache Mahout. `mahout.apache.org`.

[2] Oracle R Enterprise. `docs.oracle.com/cd/E27988_01/doc/doc.112/e26499.pdf`.

[3] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2670–2676, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[4] Justin Betteridge, Andrew Carlson, Sue Ann Hong, Estevam R. Hruschka Jr., Edith L. M. Law, Tom M. Mitchell, and Sophie H. Wang. Toward never ending language learning. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 1–2, 2009.

[5] Joseph K. Bradley, Aapo Kyrola, Danny Bickson, and Carlos Guestrin. Parallel coordinate descent for l1-regularized loss minimization. In *ICML*, pages 321–328, 2011.

[6] E. V. Buskirk. How the netix prize was won. *Wired*, 2009.

[7] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[8] Hassan Chafi, Arvind K. Sujeeth, Kevin J. Brown, HyoukJoong Lee, Anand R. Atreya, and Kunle Olukotun. A domain-specific approach to heterogeneous parallelism. In *PPOPP*, pages 35–46, 2011.

[9] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA, 2008. ACM.

[10] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, Aug 2011.

[11] Pedro Domingos and Daniel Lowd. *Markov Logic: An Interface Layer for Artificial Intelligence*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2009.

[12] Oren Etzioni, Michael Cafarella, Doug Downey, Stanley Kok, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 100–110, New York, NY, USA, 2004. ACM.

[13] David A. Ferrucci, Eric W. Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John M. Prager, Nico Schlaefer, and Christopher A. Welty. Building watson: An overview of the deepqa project. *AI Magazine*, 31(3):59–79, 2010.

[14] Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. Efficient, feature-based, conditional random field parsing. In *ACL*, pages 959–967, 2008.

[15] Amol Ghoting and et al. SystemML: Declarative machine learning on MapReduce. In *ICDE*, 2011.

[16] Vidhya Govindaraju, Ce Zhang, and Christopher Ré. Understanding tables in context using standard nlp toolkits. In *ACL (2)*, pages 658–664, 2013.

[17] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[18] N.A. Heim, M.L. Knope, and J. L. Payne. Cope's rule in solitary marine bilaterian animals across the past 540 million years. In Preparation.

[19] Joseph M. Hellerstein and et al. The MADlib analytics library: Or MAD skills, the SQL. *PVLDB*, pages 1700–1711, 2012.

[20] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge Univ. Press, 2003.

[21] Gjergji Kasneci, Maya Ramanath, Fabian Suchanek, and Gerhard Weikum. The yago-naga approach to knowledge discovery. *SIGMOD Rec.*, 37(4):41–47, March 2009.

[22] Rajasekar Krishnamurthy, Yunyao Li, Sriram Raghavan, Frederick Reiss, Shivakumar Vaithyanathan, and Huaiyu Zhu. Systemt: A system for declarative information extraction. *SIGMOD Rec.*, 37(4):7–13, March 2009.

[23] Arun Kumar, Feng Niu, and Christopher Ré. Hazy: making it easier to build and maintain big-data analytics. *Commun. ACM*, 56(3):40–49, 2013.

[24] Aapo Kyrola, Guy Blelloch, and Carlos Guestrin. Graphchi: Large-scale graph computation on just a pc. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation*, OSDI'12, pages 31–46, Berkeley, CA, USA, 2012. USENIX Association.

[25] Yunyao Li, Frederick R. Reiss, and Laura Chiticariu. Systemt: A declarative information extraction system. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, HLT '11, pages 109–114, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[26] J. Liu and et al. An asynchronous parallel stochastic coordinate descent algorithm. *ICML*, 2014.

[27] Yucheng Low and et al. Distributed GraphLab: A framework for machine learning in the cloud. *PVLDB*, pages 716–727, 2012.

[28] Andrew McCallum. Joint inference for natural language processing. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 1–1, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[29] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *NAACL*, 2013.

[30] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[31] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, USA, 1997.

[32] Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. Scalable knowledge harvesting with high precision and high recall. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 227–236, New York, NY, USA, 2011. ACM.

[33] Feng Niu, Christopher Ré, AnHai Doan, and Jude Shavlik. Tuffy: Scaling up statistical inference in markov logic networks using an rdbms. *Proc. VLDB Endow.*, 4(6):373–384, March 2011.

[34] Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS*, pages 693–701, 2011.

[35] Feng Niu, Ce Zhang, Christopher R, and Jude W. Shavlik. Elementary: Large-scale knowledge-base construction via machine learning and statistical inference. *Int. J. Semantic Web Inf. Syst.*, 8(3):42–73, 2012.

[36] Hoifung Poon and Pedro Domingos. Joint inference in information extraction. In *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 1*, AAAI'07, pages 913–918. AAAI Press, 2007.

[37] Hoifung Poon and Lucy Vanderwende. Joint inference for knowledge extraction from biomedical literature. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 813–821, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[38] Warren Shen, AnHai Doan, Jeffrey F. Naughton, and Raghu Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 1033–1044. VLDB Endowment, 2007.

[39] E.R. Sparks and et al. MLI: An API for distributed machine learning. In *ICDM*, pages 1187–1192, 2013.

[40] Fabian M. Suchanek, Mauro Sozio, and Gerhard Weikum. Sofie: A self-organizing framework for information extraction. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 631–640, New York, NY, USA, 2009. ACM.

[41] Arvind K. Sujeeth and et al. OptiML: An Implicitly Parallel Domain-Specific Language for Machine Learning. In *ICML*, pages 609–616, 2011.

[42] J.N. Tsitsiklis, D.P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, pages 803–812, 1986.

[43] Alexander Yates, Michael Cafarella, Michele Banko, Oren Etzioni, Matthew Broadhead, and Stephen Soderland. Textrunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[44] Ce Zhang and Christopher Ré. Dimmwitted: A study of main-memory statistical analytics. *ArXiv e-print.*, 2013.

[45] Ce Zhang and Christopher Ré. Towards high-throughput gibbs sampling at scale: A study across storage managers. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 397–408, New York, NY, USA, 2013. ACM.

[46] Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. Statsnowball: A statistical approach to extracting entity relationships. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 101–110, New York, NY, USA, 2009. ACM.

[47] Martin Zinkevich and et al. Parallelized stochastic gradient descent. In *NIPS*, pages 2595–2603, 2010.