Supplementary Data

In the following section we describe four use case scenarios, each intended to show the potential value of the NIH topic database for addressing specific questions regarding NIH funded research. In addition, we use each example to demonstrate the relationships between machine learned categories/clusters and existing NIH administrative and categorical information.

Use Case #1 - Topic-Based Queries

In the first use case, we show how our database can be queried to find grants focused on a specific research topic relevant to a variety of NIH Institutes. The example demonstrates the complementary information provided by the two methods supporting our database, topic modeling and graph-based clustering, and shows how they interact to elucidate the patterns of NIH funding in a particular area of research.

Example Query - Grants Focused on Angiogenesis

Our example is shown in Supplementary Figure 1. In this case, imagine that a user is interested in angiogenesis, the formation of new blood vessels. If one were to query NIH RePORTER via standard keyword technology, the query would retrieve >2000 grants from a single year. In most cases this query would be of limited use, since it fails to distinguish documents that are truly focused on angiogenesis from those that only mention the term in passing. One could try multi-word queries, but such queries are often biased by pre-conceived notions of relationships among concepts, and as a result, extensive research can be required to avoid missing highly relevant documents.

Our alternative query relies on topic modeling, a recently developed Bayesian statistical method that characterizes documents in a corpus by discovering latent "topics" using unsupervised machine learning.[2,38] These machine-learned topics are quite different from keywords, for two reasons. First, topics do not correspond to individual words, but instead are groups of words that are commonly used together. We use the algorithm to discover these word groups for us. Second, unlike keywords, topic word assignments are provisional rather than all-or-none, because they depend on the topic assignments of the other words within each document. Thus the topic assignments of specific instances of a word are context sensitive. To give a simple example from the current database, a keyword query using the term "network" retrieves grants on widely divergent types of networks, ranging from clinical research networks to gene expression networks. With a topic-based query, one can choose which of these multiple contexts is relevant, because the topic word assignments are sensitive to specific meanings and contexts of word usage.

For our current query, the user enters the term "angiogenesis" into the topic query field. Our interface is designed for easy access to topics via an auto-populate function, which displays the top ten words from the topics containing a given word entry. In this case, two topics are listed. One contains words related to angiogenesis, plus words associated with vascular endothelial growth factor (VEGF), a well-studied angiogenic factor. The second topic (not shown in Supplementary Fig. 1) also contains the word "angiogenesis," but its other words are focused on vascular endothelial cells, rather than blood vessel formation. These two particular topics have substantial overlap, and thus there may be queries in which both topics are highly relevant. We discuss this issue in more detail below, in the context of topic representation in the database. Here, in our initial description, we assume the words from the first topic are well-aligned with the interests of the user, and the user selects the topic for the query.

The results of the query are shown in graphic form, using the second method underpinning our database, an algorithm that places the grants in a two-dimensional layout based on their lexical (i.e., topic- and word-based) similarity. As shown in the figure, >75% of the awards retrieved from a single year of NIH funding (2009) coalesce in the central region of the graph. The other 25% are located in other clusters appropriate to their respective topical focus. As seen from the called-out titles in the figure, the grants in

this central cluster are highly focused on angiogenesis, even though they were awarded by different NIH Institutes. In fact, the ~270 retrieved awards in this central region were spread among twelve different NIH Institutes, and they were reviewed in nearly fifty different review study sections, plus an equal number of one-time special emphasis review panels (see below for a description of the NIH peer review structure). Thus, our example highlights potential difficulties associated with retrieving and assessing NIH grants covering specific areas of research, since scientific categories are often only partially aligned with the NIH administrative and peer review organization.

This example also highlights a remarkable feature of the graph layout that to our knowledge has not been described in previous work using this or other layout algorithms. Specifically, the graph forms a lattice-like structure, in which the clusters are not isolated but instead are linked by strings of aligned nodes, which have topical content shared by the corresponding clusters at either end. In this case, grants on angiogenesis are aligned between two sets of clusters, one set focused on cancer mechanisms/therapeutics (for which angiogenesis inhibition is a priority), and the other on vascularization of specific tissues.

This lattice structure allows the revelation of sub-cluster groupings that otherwise would likely be lost within larger clusters. In this case these sub-cluster groups are evident in the color scheme that shows NIH Institute funding patterns, with grants from NCI in the upper left portion of the string, and grants from other Institutes in the lower right (e.g., NEI, NHLBI, see Supplementary Table 1 for acronyms and Institute descriptions). In addition, the lattice formation creates linkages of aligned documents that can explicate the nature of the connections between the larger clusters, thus facilitating interpretation of the layout organization. Since the connections are made by actual documents rather than by abstract links, they have concrete value for information retrieval, because users can select the documents of interest as part of specific queries. Thus this lattice-like organization provides a useful "between-cluster" dimensionality not possible with other clustering algorithms.

Topic-Based Query Design

Although topic modeling has been evaluated favorably for its potential benefits in document retrieval,[9,43,46,47] leveraging topics themselves for this purpose has not been studied in any detail. As shown in the example, we find that machine learned topics provide a robust vehicle for queries, which can be used in conjunction with, or as alternatives to, the existing keyword method available for NIH grants. To our knowledge, the topic-based query design we illustrate in the current example is novel, and therefore we describe it in more detail here.

For our query design, we capitalized on a key quantitative feature of the topic model output, namely, the calculation of topic allocations in each document, which are simply the proportion of words that the algorithm assigns to each topic within a given document. We utilize these topic allocations as proxy indicators for document relevance to each topic. In our database, there are multiple contexts in which topic allocations are important, including topic-based queries (described here), aggregate topic proportions of retrieved document sets (described below), rankings for lists of retrieved documents, and measures of similarity between documents (see Supplementary Methods).

In the case of topic-based queries, we utilize per-document topic allocations by providing a field for users to enter a threshold value that sets a floor for the given topic's proportions in retrieved documents. The default threshold setting is relatively high (20%), in order to maximize the likely relevance of the retrieved grant awards. We provide a tooltip in the interface instructing users that they may choose lower thresholds to retrieve more documents, but that these documents may be less focused on the subject of interest. In the case of our query in Supplementary Figure 1, we used a 10% allocation threshold, which retrieved 340 documents. An alternative, higher threshold setting of 20% retrieved fewer grants (152 grants), but note that a greater proportion of these grants were located in the region of the layout delimited by the red-dashed box in the figure, compared to the proportion of grants retrieved using the lower threshold setting (~95%, data not shown, vs. 75% in Supplementary Fig. 1). This is consistent with their higher focus on this particular topic, which is associated with this region. The choice of appropriate threshold setting depends on a given user's information needs, and we find that setting a lower threshold

can be useful in cases where maximal recall is desired (see Use Case #3, below). For precise determinations of appropriate thresholds for a given topic, users can browse a listing of documents in descending order of topic allocation. We provide such listings in dedicated pages for each topic in the interface, which we describe in the following section.

Topic Representation in the Database

We have so far noted the positive features of using machine-learned topics for NIH grants categorization and information retrieval, and in the following use cases below, we present additional examples where topics are clearly useful for these tasks. However, there is an important limitation in the topic modeling framework that has required careful consideration in the development of this database. Specifically, the exact distinctions between topics are not subject to precise definitions, and may not always correspond to users' preconceived notions of specific concepts.[2,11] This presents a challenge that requires transparent and accurate topic representations, which must also be quickly and easily accessible to users.

Our strategy for full representation of topics is to devote a separate page in the interface for each topic, as demonstrated in Supplementary Figure 2. On each topic page, we provide extensive information in a format that is conducive to spot checks, and also allows more detailed examination as necessary. The most important of these data are listings of the co-occurring and semantically similar topics, which elucidate the surrounding concepts for any given topic. In our experience these associated topics often are essential for determination of the delimiters between topic-based categories. The example we provide here is the corresponding topic page for our previous query for grants on angiogenesis research. This example is illustrative of the importance of having rapid access to similar topics, since as noted above, a separate topic with focus on vascular endothelial cells may also be relevant to potential queries for grants covering research on angiogenesis. This separate topic is listed both in the "Co-occurring Topics" and "Similar Topics" panels, with corresponding hyperlinks for quick access. Depending on the information need, a user may want to combine the two topics (using a Boolean "or" operator), or restrict the query to documents only containing one of the topics (using a Boolean "and" operator). Alternatively, the user may want to use the topic in conjunction with text-words or keywords, or with various grants-related information such as grant mechanisms, study section review panels, or NIH Institutes. In the other panels of the topic page, we provide information on potential choices for these alternative information sources, which we determine by scoring for their association with documents that have allocations to the given topic (see Supplementary Methods). Thus the topics can provide useful starting points for more detailed queries, and for understanding a given concept area.

Note in this context that we provide corresponding keyword lists from each of the three different modeled corpora, the NIH RCDC Concepts (for grants starting in 2007, http://projectreporter.nih.gov/exporter/ExPORTER_Catalog.aspx), NIH CRISP (for grants through 2006, http://projectreporter.nih.gov/exporter/crisp_catalog.aspx), and PubMed MeSH (http://www.ncbi.nlm.nih.gov/mesh). For the purpose of this discussion it is useful to note that each set of keywords is fairly complex, and each has slightly different term sets covering the same concept area. Thus these listings provide an illustration of the well known problem of lexical variability associated with keyword-based systems,[28] and also the potential difficulties designing an appropriate keyword query to fully cover a restricted concept space. Topics instead represent clusters of concepts that substantially reduce the complexity of the corpus, in a fashion not achieved by these thesaurus-based systems. In cases where the machine-learned topics may not correspond precisely to information needs, they nevertheless can be modified by queries in Boolean combination with specific text words, thesaurus terms, or document metadata. We have found that our framework for organizing associated metadata information aids in designing such queries, and enables rapid exploration to ensure that appropriate coverage has been achieved.

Use Case #2 – Information on NIH Peer Review Study Sections

The second use case is a query that asks, "What research topics are reviewed in Study Section X?" This is an important question for NIH grant applicants, because each review panel has a scientific focus that

defines its expertise and can influence its relative enthusiasm for applications proposing different kinds of research. The relationship between NIH peer review panels and scientific categories is multifaceted, and it adds an additional layer of complexity in considering NIH organizational structure. Therefore, before describing our specific example, we first provide an overview of NIH peer review organization.

NIH Peer Review Organization

Because of the extensive overlap in the types of research funded by the Institutes, NIH has a completely separate Center for Scientific Review (CSR) that is responsible for assigning and reviewing grant applications. Only roughly thirty percent of the ~80,000 applications received in a given year are reviewed by the NIH Institutes themselves; usually these are applications submitted in response to specific solicitations or special grant programs. CSR manages the review of the other seventy percent, most often at one of ~275 standing review panels, called "Study Sections" in NIH parlance. However, a substantial fraction of applications are reviewed at "Special Emphasis Panels," which may be convened for scientific or administrative reasons (e.g., avoidance of reviewer conflicts of interest).

This creates a complex universe for investigators trying to understand the contexts under which their applications will be evaluated. For any individual application there may be multiple potential venues for review, which are distinct from the Institute that will be making the subsequent funding decision. Adding to this complexity is the fact that the CSR standing study sections are explicitly organized with overlapping scientific focus (http://cms.csr.nih.gov/peerreviewmeetings/csrirgdescriptionnew/), and sometimes combine relatively disparate areas of research, in order to provide balanced reviews and to ensure that different types of research are evaluated in an appropriate context.

For the purposes of initial application assignments and administrative oversight, CSR organizes study sections into twenty-five scientifically coherent clusters known as "Integrated Review Groups" (IRGs). Note that we use this higher level scientific organizational framework for broad categorical labels on the graph layout display (Figure 1). These labels are placed automatically, using the IRG with highest proportional representation in the underlying documents. As described in the next paragraph, for higher resolution views we provide corresponding labels indicating representative standing study sections.

Relationship Between Machine Learned Topics and NIH Study Sections

Our example use-case shows how the current database can be used to understand the relationship between scientific categories, NIH review panels, and NIH Institute funding practices, within the context of specific grant awards. Before showing a specific query, we note the relationship between the database topics and NIH study sections, as represented in alternative labeling schemes that are available as separate options in the "Settings" feature of the user interface. Supplementary Figure 3 shows a comparison of these two labeling schemes, one using topic words and the other using the predominant NIH standing study section assignments of the underlying awards. Note that there is only a loose relationship between these two labeling schemes. For example, grants with a high focus on pain are localized to a cluster in the upper left corner of the displayed region, but NIH does not have a corresponding study section specifically focused on pain research. Instead, the corresponding study section is the Somatosensory & Chemosensory Systems (SCS) Study Section. The reason for this mismatch is that the SCS Study Section covers other sensory modalities in addition to pain, most notably taste, olfaction, and somatosensation (for a description, see http://cms.csr.nih.gov/peerreviewmeetings/csrirgdescriptionnew/ifcnirg/scs.htm).

Our example query for this study section is shown in Panel C, which contains a view of the graph layout including the region covered in Panels A and B (inside the red dashed box), plus a wider view of the upper right quadrant of the full layout (inside the blue dashed box). We restricted this query to the indicated regions using a bounding box feature in the user interface. Note that the topic mixtures of the grants inside the red dashed region are focused on pain, whereas grants outside this bounded region are focused on taste and olfaction, plus a minor cluster on somatosensory processing. In addition, these clusters have distinct Institute representations. The pain grants were awarded primarily by Institutes with missions in neurology (NINDS), drug abuse (NIDA), and dentistry (NIDCR, see the legend

accompanying Supplementary Table 1 for Institute descriptions). In contrast, the grants outside the bounded region were primarily funded by the National Institute on Deafness and other Communications Disorders (NIDCD).

The fact that these topics and clusters correlate with the awarding NIH Institutes highlights the importance of the underlying categories to the NIH, and indeed the categorical funding patterns that they elucidate are entirely consistent with explicit policies of the different Institutes. However, in the absence of this topic-based analysis, one would have to compile this information from the Institutes' websites, and then conduct keyword-based searches for each category. This would require extensive research, which would be intractable for most users. Our database offers an alternative approach that enables rapid retrieval of this categorical information, and does so in a way that is transparent and reproducible.

Finally, note that the SCS study section only reviews approximately half of NIH awards with a substantial focus on pain mechanisms. The other awards are covered by different standing study sections or by ad hoc special emphasis panels. Although not shown here, these other awards can readily be retrieved by selecting the indicated cluster on the graph layout, or by searches using topics as query fields, and restricting the results to grants not in this particular study section. Thus in addition to the categorical information already described, the database provides additional value by offering ready access to tools that enable query expansion, for retrieval of highly relevant information that otherwise would be very difficult to acquire.

Use Case #3 – Information on NIH Research Categories

Our third example use case addresses the question, "What are the grants that NIH funds in category X?" This type of question is regularly asked by policy analysts, patient advocacy representatives, and of course, NIH staff and leadership.

NIH currently provides two interdependent resources for retrieving information on NIH research categories. The first is a set of thesaurus-based keywords that are used to tag each award and are available for query via the NIH RePORTER website (http://projectreporter.nih.gov). The second is a set of funding estimates and corresponding grant lists for ~215 Research, Condition and Disease Categories (RCDC). All but a handful of these are compiled using an automated keyword-based classification system, in combination with NIH staff input (http://report.nih.gov/rcdc/categories/).

Here we show how the machine learned topics and clusters provide sub-categorical information at a level of detail that is not available in the NIH RCDC Categories, but that nevertheless reveals highly relevant distinctions between the types of research funded by various NIH Institutes. Our example is a query for grants assigned to the NIH RCDC Category "Sleep Research," which retrieved ~830 grants from 23 Institutes. As seen in Figures C and D, the top four topics from this set were focused on salient subcategories of research in this area, namely circadian rhythms, sleep disorders, neurobiology of sleep/arousal, and sleep-disordered breathing.

NIH Institute Funding for Grants in Different Subcategories of Sleep Research

Supplementary Figure 4 shows that different NIH Institutes have different commitments to each of these sub-categories. To produce the table, we queried subsets of grants from the NIH Sleep Research category using each of the four topics. For each subset, we used two different topic allocation thresholds (5% and 10%). We compiled the results of these queries in order to determine an estimated range of grant awards in each category, organized by awarding Institute. From this table it can be seen, for example, that NHLBI funds substantial numbers of awards in each category, but has a clear weighting towards sleep disorders and sleep disordered breathing. In contrast, roughly 90% of the awards from NIGMS are focused on circadian biology rather than sleep disorders or neurobiology of sleep and arousal.

Graph Layout Organization of NIH Sleep Research Grants

To understand how NIH Sleep Research is situated within the broader context of NIH funded grants, we

show the locations of these awards on the graph layout (Supplementary Figure 5). We found that >90% of the grants were located on the right side of the graph, in areas involving neural systems, behavior, health risks and interventions. There were two predominant clusters (red dashed boxes in Fig. D) that together accounted for ~56% of the retrieved awards. One was focused on circadian rhythms, and the other had a combined focus on sleep disorders and neurobiology of sleep and arousal. Consistent with results shown in Supplementary Figure 4, these clusters had major differences in their proportional Institute representations, with prominent funding by NIGMS for circadian research, and NHLBI for sleep disorders.

In addition to demonstrating the level of added detail provided by the topic-based categories, this example also highlights some of the differences between the NIH RCDC system for categorical reporting and the current framework. Specifically, the categorical funding information provided by NIH is not intended to be comprehensive and is only offered for cases in which a specific reporting need has been identified (often by Congressional mandate). Thus the RCDC Categories do not attempt to characterize the full body of NIH-funded research but rather to address external reporting requirements. Topic modeling offers a complementary approach in which machine learned categories are comprehensive and are discovered from latent discourses within documents, rather than by externally defined criteria. Thus the topics provide reference points from which various information requirements can be addressed by users with divergent interests and needs.

Perhaps more importantly, because the topic modeling output is machine learned rather than user defined, it offers a basis for users to discover and investigate interrelationships among concepts and documents, rather than solely meeting pre-determined information requirements. In a time of exploding information availability, when very few individuals are able to achieve expertise outside of a limited research area, this type of resource may therefore hold potential to facilitate discovery and enhance the rate of scientific progress.

Use Case #4 - Research Trend Characterization

As a fourth use-case for our database, we ask, "What research topics changed during the period covered by the database, and what were the underlying trends driving these changes?" The discovery of research trends is an important priority not only for NIH administration, but also for analysts and social scientists interested in systematic assessment of science policy.

Screen for Topics with Proportional Changes Over Time

Although other methods of text processing may prove to be more potent for trend discovery [3,19,26,42], our topic model nevertheless provides quantitative data that can easily be mined for robust changes in research categories. More importantly, as we show here, if a trend has been identified (regardless of the discovery method), an analysis of co-occurring topics can give an indication of the underlying changes in research that drive the trend.

For an initial assessment of trends represented in the database, we performed a preliminary screen in which we compared proportional representation of topics across the three years covered in our database, starting with the set of non-competing awards from fiscal year 2007 (i.e., grants initially awarded prior to 2007), and ending with newly awarded grants from fiscal year 2009. The topic with the highest proportional increase during this period was focused on microRNA biology. This is a recent and fast growing area of research, and the corresponding topic had a fourfold change in relative allocation between the grants at the beginning and ending periods in our analysis.

Characterization of Trends Associated with an Increase in Research on microRNA Biology

Supplementary Figure 6 shows the growth in awards for this area of research, but more importantly, it gives information on the changes in research focus that have driven the trend. To generate the figure, we

queried the database for the topic focused on microRNA biology, in this case using a 10% threshold setting. We verified this trend with queries using different topic thresholds, as well as with text-word and keyword based queries (*data not shown*). All of these searches gave consistent results, indicating that there were ~2.5fold more grants retrieved from the set of 2009 new awards, relative to non-competing awards from 2007, even though at NIH overall, there were far fewer 2009 new awards than 2007 non-competing awards (~15,000 vs. ~36,000).

The figure shows that the two sets of awards had very different distributions across the graph layout, consistent different mixtures of topics that co-occurred with the dominant topic on microRNA biology. Non-competing awards from 2007 (i.e., awards from grants that had been initiated in competing awards prior to 2007) were clustered in a region of the layout primarily dealing with RNA (plus a minor cluster focused on cell differentiation), and the topics were focused primarily on basic research categories. In contrast, the new awards from 2009 were much more dispersed, and now focused on cancer, biomarkers, and topics associated with complex physiological systems such as neural plasticity.

In addition to a change in the research topics associated with studies of microRNA, NIH Institute support for this research changed in a manner consistent with a transition from basic cellular/molecular biology to complex physiology and diseases. NIGMS was the predominant Institute supporting awards granted prior to 2007, consistent with its mission of funding basic cellular/molecular biology that is not associated with particular diseases or organ systems (see Supplementary Table 1 for acronyms and descriptions of the NIH Institutes). In contrast, other NIH Institutes played a much more prominent role in funding the new awards from 2009, and in particular the top Institute was NCI, which is focused on cancer research. This emphasis on cancer research reflects very recent evidence that specific microRNAs may be useful for cancer diagnosis [14,15,32].

These results were robust and very easy to obtain. Furthermore, they are suggestive of other types of analyses that could be initiated in this format, such as assessments of relationships between basic and translational research, and ongoing interactions between scientific discoveries and research funding.

Supplementary Methods

Topic Modeling

Topic modeling refers to recently developed Bayesian statistical algorithms for categorizing unstructured text. To model our corpus, we used Latent Dirichlet Allocation (LDA), which was one of the earliest, and which has been the most influential of these algorithms.[4,16] A variety of extensions of LDA have been developed, with the goal of improving or refining the resulting output.[2,17,38] However, in recent work LDA outperformed newer algorithms in producing semantically meaningful topics.[11] For the current work we used the implementation provided by the MALLET topic modeling toolkit (http://mallet.cs.umass.edu/).[31]

Our database consists of titles and abstracts from NIH grants spanning fiscal years 2007-2010, with planned updates for current grants as they are awarded. NIH typically issues 70,000-80,000 grants in a year (including subprojects and award supplements), but nearly three quarters of these awards are from the non-competing years of multi-year grants. The titles and abstracts from these non-competing awards are carried over from the previous year, and therefore the number of unique documents in our database from the four year period is ~110,000. To train our topic model, we used titles and abstracts from awards from 2007-2009, plus competing awards from 2004-2006 (~150,000 unique documents). We also included titles/abstracts from ~220,000 MEDLINE journal articles published in 2007-2009 that cited NIH grants (http://projectreporter.nih.gov/exporter/), to enhance the statistical robustness of the analysis for the corresponding areas of research. Subsequent to model training and topic characterization (described below), we added newly awarded grants from funding year 2010 (~25,000 new documents). Topics were inferred for these newly awarded grants while holding the overall trained topic-word proportions constant.[45]

We pre-processed the text using standard tokenization procedures, paying special attention to stopwords (i.e., words removed from the analysis), acronyms and phrases. In preliminary work, we noted that many topics contained general, non-research terms (e.g., "understanding," "goals," "believe"), which represented a fairly large portion of the corpus (~15%). We were surprised to find that these types of words were sequestered into individual topics, and furthermore, that it was easy to segregate these topics based on fractional document allocation, because the topics had uniformly and distinctively low allocations to individual documents. We manually aggregated obviously uninformative terms from these topics to generate a list of ~1200 words, which were added to a standard stopword list and removed from subsequent analysis. We also created a vocabulary of ~600 standardized acronyms and ~4200 commonly used bigrams and phrases. Before producing our final model, we gave a two-fold weighting to words in the document titles, in order to increase their influence on topic proportions within each document.

LDA requires pre-specification of the number of topics (T), plus two "smoothing" parameters that adjust the concentration of topics in a document (alpha), and terms in a topic (beta).[20] We experimented with these settings by modeling portions of the corpus, as well as the full dataset, and we evaluated the output by extensive manual review of the resulting topics. For T, we found there was a tradeoff between concept resolution, which we judged to be too low in models with fewer topics, and the number of poor quality topics, which was quite high in models with many topics. We were unable to produce a model in which such artifacts were completely absent (see below), but we decided that T=700 produced an acceptable balance, with well resolved concepts but relatively few topics of poor quality. Recent work suggests a method for optimizing the alpha parameter,[40] but we found that this optimization protocol increased the number of poor quality topics, and we therefore used a standard heuristic for this parameter (0.05(L/T), where L is average document length). We used a standard setting for beta (0.01).

Topic Evaluation

In preliminary assessments, we rated individual topics for their ability to convey a single coherent concept, and in particular, to convey a concept that could appropriately characterize all the documents

associated with the corresponding topic. In these evaluations we noted the existence of "junk" topics that did not convey coherent concepts, even after we had removed large numbers of general words (described above) from the analysis. These obviously poor topics contained nonsensical combinations of words that were unrelated to one another in any meaningful way, words that often were never used in the same documents. Even though these poor topics represented a relatively small fraction of the model, we were concerned about their potential to degrade user confidence in the method.

To help identify poor topics, we developed a method for automated assessment of topic quality. Assessment of topic models has traditionally relied on data prediction statistics in the form of perplexity analysis or likelihood calculations of held-out data,[8,41] but recent work suggests that these statistics correlate poorly with human assessments of topic coherence.[11] As an alternative, Newman et al.[34,35] developed a simple measure of topic quality in which the co-occurrences of topic words were compared to those of an external corpus. This coherence measure performed well relative to human judgments of topic coherence in experimental tests. In separate experiments performed in conjunction with the current work, we determined that the document co-occurrence values for topic words within a modeled corpus correlated well with NIH Program staff evaluations of topic quality (data not shown). We used this approach to automatically score topics for the overall document co-occurrence frequencies of the topic words, as shown in Supplementary Figure 7.

In addition to word co-occurrence, we also found in separate tests that topic size (fraction of the corpus assigned to a given topic) correlated inversely with blinded rater assessments of topic quality. Moreover, in extensive preliminary evaluations of the current dataset, we found that the smallest topics in a given model invariably had a high probability of being nonsensical mixtures of words, and we were unable to produce a model lacking these poor topics. Therefore, we curated our database to provide users with cautionary information for topics that appeared to lack sufficient quality.

For topic curation, we used our automated scoring system to flag potentially poor topics, and we used word co-occurrence matrices to assist in our assessments of the robustness of the specific word relationships inferred by the topic model. Supplementary Figure 7 shows examples of matrices for good and poor topics, with their corresponding values for word co-occurrence score and topic size measurement. For especially poor topics (indicated by red diamonds in Panel C) we deleted the topic word-based labels from the user interface, and substituted the words "mixed topic." In addition, we included a warning note as a header to the topic page. Topics deemed to be intermediate in quality (yellow diamonds) received the same warning note on their topic page, but we retained their topic word-based labels in the search and retrieval interfaces.

Contextual Data Associated with Each Topic

As described in the Supplementary Results, we created a scheme for providing access to important contextual information about each topic, in order to enable rapid exploration of the underlying concept representations. For "Topic Co-occurrence," we used the mutual information score between the event that one topic occurs in a document, and the event that the second topic occurs in the same document. For "Topic Similarity," we used the Jensen-Shannon divergence between the distributions over words for each topic. In addition to information on related topics, we also calculated the associations of a variety of metadata tags assigned to documents, including thesaurus terms, journals, grant Institute assignments, grant mechanisms, study section review panels, and Program Announcements. For these association scores, we summed the topic probabilities over all documents with a given tag, and weighted these sums by the log inverse document frequency for each tag.

Document Similarity Determination

For document similarity we used an equal-weighted comparison of the words and the topics in each pair of documents.[43,46,47] We calculated the Kullback–Leibler divergence of word probability, defined as word frequency normalized to document length, and topic probability, estimated by Gibbs sampling.[16] Combined divergence was set as a weighted sum of the word-based score (weighted at 0.5), and five

topic-based scores (each weighted at 0.1) from independent topic models with different numbers of topics (T = 550-700). Final divergence scores were converted to similarity values from 0-1, where 1 represents identity between documents.

Graph Layout

For two-dimensional graphing, we used an algorithm based on force-directed placement, wherein nodes are attracted to one another by similarity, but repelled from areas of high node density.[7,13,29] This algorithm, originally known as VxOrd, was initially developed as part of the VxInsight data mining tool from Sandia Labs.[12] It has been used extensively for cluster analysis of gene expression,[1,18,22,27,30,39,44] phylogenomic mapping,[36,37] and scientometric analysis of citation linkages, especially for maps of scientific disciplinary organization.[5,6,23-25] In extensive assessments, Klavans and Boyack found that it grouped journals with greater fidelity to ISI journal categories than the corresponding citation linkages that were used as algorithm inputs.[23,24] VxInsight is no longer supported, but the algorithm was further developed as open source software known as DrL,[29] and is currently available in at least three different implementations (http://www.cs.sandia.gov/~smartin/software.html; http://igraph.sourceforge.net/doc/R/layout.drl.html; http://nwb.slis.indiana.edu/). We experimented with the additional "recursive" layout option added to this new software, but we did not use it for our final layout.

Unlike previous bibliometric studies using VxOrd/DrL, which have used citations as a basis for clustering, we used lexical information to cluster documents based on their topic/word-based similarities (described above). We have performed extensive preliminary tests of this method, using abstracts submitted to the annual meeting of the Society for Neuroscience (http://scimaps.org/maps/neurovis/new/),[10] biomedical abstracts retrieved using PubMed,[33] and portions of the NIH grants corpus (http://scimaps.org/maps/ninds/, https://app.nihmaps.org). In these smaller-scale tests (~4,000-50,000 documents), we applied the graph layout coordinates to a web based graphical user interface, first using the Google Maps application programming interface, and more recently using an Adobe Flash-based design.[21] We consistently found that these layouts produced local clusters with high salience to NIH Program staff, and that the interface appeared to be useful for recognizing clustering patterns and retrieving appropriate documents.

For labels, we used an automated agglomerative scheme in which different regions of the graph are tagged with topic words derived from the topic with highest proportional representation in the underlying documents. In addition to topic words, we used other metadata categories derived from the same clusters of grant awards to create separate label schemes (e.g., NIH Integrated Review Groups in Figure 1, and NIH Study Sections in Supplementary Figure 3), which are available as separate options in the "Settings" feature of the user interface.

Graph Layout Evaluation

We first assessed the graph in terms of its fidelity to the similarity values used as inputs, meaning the degree to which layout distance between pairs of grants correlates with their lexical similarity, i.e., if two grants are similar, they should be close to one another on the graph. However, when we performed spot examinations using particular grants, we found that their top similar documents were not always proximal on the graph. But instead of being randomly dispersed, they usually appeared in discrete clusters, with the focus of each cluster reflective of the topic content of the originating grant.

An example of this clustered output is demonstrated in Supplementary Figure 8, which shows the graph locations for the top 100 similar grants for a single originating grant. In this case, approximately half of the similar grants were in the same cluster as the originating grant (cluster "b"), whereas the grants in the other two clusters were much more distant. Nevertheless, the three clusters were each organized around a thematically coherent concept, corresponding to topics from the originating grant.

A cumulative analysis of input similarity values to output distances on the graph was quite consistent with our example. In this analysis, for every grant on the graph, we derived the distances for its top similar grants. We binned these data, and derived median, 95% and 5% threshold values for the top 1, 10, 100

and 300 most similar grants for each grant in the layout, in terms of their corresponding output distances. We report the distances for each of these bins in Graph Units (see scale bars in panel A). The median values for each of the bins (red circles) indicate that the most proximal 50% of the document pairs in each bin were located near one another, as was seen in our example. However, also consistent with the example, the 95% threshold values (white circles) indicate that the distal half of the document pairs in all of the bins were highly dispersed, and in fact they extended more than halfway across the graph (maximal pair-wise distance: ~900 Graph Units).

Also evident in the example in Supplementary Figure 8, we found that the clusters of similar grants were situated within a backdrop of much larger groups of grants that were not included in the retrieved set of highly similar grants. This can be seen in the cumulative data by comparing the distances in Panel B to those of Panel C, where instead of using similarity as the criterion for binning, we used graph proximity. In Panel C, for each grant, we binned the top 1, 10, 100, and 300 most proximal grants on the layout. A comparison of the corresponding distances in each of the two plots shows that the overall number of proximal documents was much higher than the number of similar documents associated with a given distance on the graph. Nevertheless, as seen in the example in Panel A, we found that these "non-similar" documents can be quite relevant to retrievals, because they share a thematic focus with the particular cluster of retrieved grants.

Because of its potential usefulness in data exploration, we embedded this type of query for similar documents as a feature in the web user interface. Specifically, each grant in the database has a dedicated page that includes an extensive listing of the top similar documents in descending rank order. A "Map Similar" button at the top of the page retrieves a defined number of the grant's top similar documents for display on the layout (default is top 100). We have found that the clustered nature of this output provides an extra organizational framework that otherwise would not be available solely from retrieval of a list of similar documents. This framework has potential value for document categorization, query focus, and query expansion. For query focus, the graph can be used to direct queries to specific aspects of relatedness (i.e., to a specific cluster of similar documents). For query expansion, the layout provides specific contextual information, since the retrieved clusters are situated among documents with shared thematic focus, within a backdrop of the entire dataset. Also, it is worth noting that, unlike our example in Supplementary Figure 8, this type of query usually retrieves a few scattered documents that are not contained within the major clusters of retrieved grants. In our experience, the graph also can provide rapid access to useful contextual information regarding the research focus of these "outlier" documents. Once again, the graph provides ready access to information that would not be available using standard clustering techniques.

Graph Retrieval Performance

Because of the potential value of the graph's clustered output, we decided on a second analysis, which focused on the performance of the graph in an information retrieval setting. Our results, which are shown in Supplementary Figure 9, indicate that if one is interested in the very closest related documents, then a rank-ordered listing of similar documents (as provided in the database) would be superior to retrieving proximal grants on the graph layout. However, if one is interested in larger groups of grants (100-300 documents), retrieval of proximal grants on the layout provides similar performance, with the added benefit of a clustered organization in a visualized setting, where patterns are more easily discerned.

Typically, clustering algorithms are assessed by testing whether members of each cluster match to independently derived categories, or match to human judgments of relevance. Because graphing algorithms do not produce clusters *per se*, we used a ranked retrieval framework for evaluation.[28] This framework allowed us to compare graph output distances to their corresponding input similarity scores, using the rank ordered bins derived for the analysis described in the previous section.

We used two different categories as proxy measures of relevance for our retrieved sets. In the first, we asked whether grants from the retrieved sets possessed the same top scoring topic as the initiating grant. In terms of precision (i.e., the fraction of retrieved grants that had matching top topics), document

similarity was clearly a better strategy for the top 1 and top 10 documents, compared to the corresponding most proximal documents on the graph. However, for bins containing the top 100 or 300 documents, the matching values for the two retrieval strategies were much more comparable. This was also the case for measures of recall (i.e., the number of matching documents, expressed as a fraction of the total potential matches in the overall set).

As a separate measure of relevance, a measure involving independent categorical assessments, we used Program Codes from an internal NIH database. These codes are specific for each NIH Institute and are typically used to indicate assignment of grants to funding programs or individual Program Directors. Note that these assignments are not expected to have full alignment with research categories, because they involve administrative considerations beyond research content. For example, because they are Institute specific, they do not capture the considerable overlap in research funding between NIH Institutes. Nevertheless, to the extent that they do align with research categories, these codes can be expected to provide a basis for relative comparison of our two retrieval strategies.

The second analysis in Supplementary Figure 9 shows results using NIH Program Codes as a proxy for relevance. These results were entirely consistent with the corresponding analysis using match to top topic. For both of these measures, document similarity was superior to layout proximity for retrieval of highly related documents (top 1 and top 10). However, for more generally related documents (top 100 or top 300), retrieval using layout proximity was relatively comparable to document similarity.

These results suggest that the graph output is an appropriate vehicle for organizing and recognizing patterns among sets of grants, assuming the sets of interest are in the hundreds of documents. We have shown in our use case examples (see Supplementary Results) that this is a highly relevant range of retrieval, and that the graph-based clusters reveal categorical patterns that are clearly relevant to NIH Institutes. In addition, our example in Supplementary Figure 8 (and many other examples not shown) demonstrates how the graphical output may also be useful for more limited sets of documents when query focus or query expansion are needed.

Reference List

[1] D. Bhojwani*, et al.*, "Biologic pathways associated with relapse in childhood acute lymphoblastic leukemia: a Children's Oncology Group study," Blood 108(2), 711 (2006).

[2] D. Blei and J. Lafferty, "Topic Models,"in *Text Mining: Theory and Applications,* edited by A. Srivastava and M Sahami (Taylor and Francis, 2009).

[3] D. M. Blei and J. D. Lafferty, "Dynamic Topic Models," Proceedings of the 23rd international conference on Machine learning*,* pp.113-120 (2006).

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research*,* 3 ed., pp.993-1022 (2003).

[5] K. Borner, C. Chen, and K. W. Boyack, "Visualizing Knowledge Domains," Annual review of information science and technology*,* 37 ed., pp.179-255 (2003).

[6] K. W. Boyack, K. Borner, and R. Klavans, "Mapping the Structure and Evolution of Chemistry Research," Scientometrics*,* pp.1-16 (2009).

[7] K. W. Boyack, B. N. Wylie, and G. S. Davidson, "Domain Visualization Using VxInsight for Science and Technology Management," Journal of the American Society for Information Science and Technology*,* 53 ed., pp.764-774 (2002).

[8] W. Buntine, "Estimating Likelihoods for Topic Models," Advances in Machine Learning , 51 (2009).

[9] W. Buntine*, et al.*, "A Scalable Topic-Based Open Source Search Engine," Web Intelligence, 2004.WI 2004.Proceedings.IEEE/WIC/ACM International Conference on*,* pp.228-234 (2005).

[10] G. A. P. C. Burns*, et al.*, "A Snapshot of Neuroscience: Unsupervised Natural Language Processing of Abstracts From the Society for Neuroscience," Society for Neuroscience Abstracts*,* (2007).

[11] J. Chang*, et al.*, "Reading Tea Leaves: How Humans Interpret Topic Models," Neural Information Processing Systems*,* (2009).

[12] G. S. Davidson*, et al.*, "Knowledge mining with VxInsight: Discovery through interaction," Journal of Intelligent Information Systems 11(3), 259 (1998).

[13] G. S. Davidson, B. N. Wylie, and K. W. Boyack, "Cluster Stability and the Use of Noise in Interpretation of Clustering," Proc.IEEE Information Visualization*,* 2001 ed., pp.23-30 (2001).

[14] J. K. Edwards*, et al.*, "MicroRNAs and Ultraconserved Genes as Diagnostic Markers and Therapeutic Targets in Cancer and Cardiovascular Diseases," Journal of Cardiovascular Translational Research 3(3), 271 (2010).

[15] M. Fabbri, "miRNAs as molecular biomarkers of cancer," Expert Review of Molecular Diagnostics 10(4), 435 (2010).

[16] T. L. Griffiths and M. Steyvers, "Finding scientific topics," 101, 5228 (2004).

[17] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum, "Topics in semantic representation," 114(2), 211 (2007).

[18] R. C. Harvey*, et al.*, "Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome," Blood (2010).

[19] Q. He, *et al.*, "Detecting Topic Evolution in Scientific Literature: How Can Citations Help?," Proceeding of the 18th ACM conference on Information and knowledge management, pp.957-966 (2009).

[20] G. Heinrich, "Parameter estimation for text analysis," Web: Http://Www. Arbylon. Net/Publications/Text-Est. Pdf (2005).

[21] B. W. Herr, *et al.*, "The NIH Visual Browser: An Interactive Visualization of Biomedical Research," IEEE International Conference Information Visualisation, pp.505-509 (2009).

[22] S. K. Kim, *et al.*, "A Gene Expression Map for Caenorhabditis Elegans," Science, 293 ed., p.2087 (2001).

[23] R. Klavans and K. W. Boyack, "Identifying a Better Measure of Relatedness for Mapping Science," Journal of the American Society for Information Science and Technology, 57 ed., pp.251-263 (2006).

[24] R. Klavans and K. W. Boyack, "Quantitative Evaluation of Large Maps of Science," Scientometrics, 68 ed., pp.475-499 (2006).

[25] R. Klavans and K. W. Boyack, "Toward a Consensus Map of Science," Journal of the American Society for Information Science and Technology, (2008).

[26] J. Kleinberg, "Temporal dynamics of on-line information streams," Data Stream Management: Processing High-Speed Data Streams. Springer (2006).

[27] W. Li, *et al.*, "ExprAlign- the identification of ESTs in non-model species by alignment of cDNA microarray expression profiles," BMC Genomics 10(1), 560 (2009).

[28] C. D. Manning, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval* (Cambridge UP, Online Edition, 2009).

[29] S. Martin, *et al.*, "DrL: Distributed Recursive (Graph) Layout," SAND2008-2936J: Sandia National Laboratories, (2008).

[30] S. B. Martin, *et al.*, "Gene expression overlap affects karyotype prediction in pediatric acute lymphoblastic leukemia," Leukemia 21(6), 1341 (2007).

[31] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit," (2002).

[32] P. S. Mitchell, *et al.*, "Circulating MicroRNAs As Stable Blood-Based Markers for Cancer Detection," Proceedings of the National Academy of Sciences, 105 ed., p.10513 (2008).

[33] D. Newman, *et al.*, "Visualizing search results and document collections using topic maps," Web Semantics: Science, Services and Agents on the World Wide Web (2010).

[34] D. Newman, *et al.*, "Automatic Evaluation of Topic Coherence," Human Language Technologies: North American Chapter of the Association for Computational Linguistics, (2010).

[35] D. Newman, *et al.*, "Evaluating Topic Models for Digital Libraries," Proceedings of the 10th annual joint conference on Digital libraries, pp.215-224 (2010).

[36] S. Schneiker, *et al.*, "Complete genome sequence of the myxobacterium Sorangium cellulosum," Nature Biotechnology 25(11), 1281 (2007).

[37] B. S. Srinivasan, *et al.*, "Functional genome annotation through phylogenomic mapping," Nature Biotechnology 23(6), 691 (2005).

[38] M. Steyvers and T. Griffiths, "Probabilistic Topic Models," Handbook of Latent Semantic Analysis, p.427 (2007).

[39] J. M. Stuart, *et al.*, "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," Science, 302 ed., p.249 (2003).

[40] H. M. Wallach, D. Mimno, and A. McCallum, "Rethinking LDA: Why Priors Matter," Topic Models: Text and Beyond Workshop in Neural Information Processing Systems Conference, (2009).

[41] H. M. Wallach, *et al.*, "Evaluation Methods for Topic Models," Proceedings of the 26th Annual International Conference on Machine Learning, pp.1105-1112 (2009).

[42] X. Wang and A. McCallum, "Topics Over Time: a Non-Markov Continuous-Time Model of Topical Trends," Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.424-433 (2006).

[43] X. Wei and W. B. Croft, "LDA-Based Document Models for Ad-Hoc Retrieval," Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp.178-185 (2006).

[44] C. S. Wilson, *et al.*, "Gene expression profiling of adult acute myeloid leukemia identifies novel biologic clusters for risk classification and outcome prediction," Blood 108(2), 685 (2006).

[45] L. Yao, D. Mimno, and A. McCallum, "Efficient Methods for Topic Model Inference on Streaming Document Collections," Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.937-946 (2009).

[46] X. Yi and J. Allan, "Evaluating Topic Models for Information Retrieval," Proceeding of the 17th ACM conference on Information and knowledge management, pp.1431-1432 (2008).

[47] X. Yi and J. Allan, "A comparative study of utilizing topic models for information retrieval," Advances in Information Retrieval , 29 (2009).