Surveying a suite of algorithms that offer a solution to managing large document archives.

BY DAVID M. BLEI

Probabilistic Topic Models

AS OUR COLLECTIVE knowledge continues to be digitized and stored—in the form of news, blogs, Web pages, scientific articles, books, images, sound, video, and social networks—it becomes more difficult to find and discover what we are looking for. We need new computational tools to help organize, search, and understand these vast amounts of information.

Right now, we work with online information using two main tools—search and links. We type keywords into a search engine and find a set of documents related to them. We look at the documents in that set, possibly navigating to other linked documents. This is a powerful way of interacting with our online archive, but something is missing.

Imagine searching and exploring documents based on the themes that run through them. We might "zoom in" and "zoom out" to find specific or broader themes; we might look at how those themes changed through time or how they are connected to each other. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme.

For example, consider using themes to explore the complete history of the New York Times. At a broad level, some of the themes might correspond to the sections of the newspaper—foreign policy, national affairs, sports. We could zoom in on a theme of interest, such as foreign policy, to reveal various aspects of it—Chinese foreign policy, the conflict in the Middle East, the U.S.'s relationship with Russia. We could then navigate through time to reveal how these specific themes have changed, tracking, for example, the changes in the conflict in the Middle East over the last 50 years. And, in all of this exploration, we would be pointed to the original articles relevant to the themes. The thematic structure would be a new kind of window through which to explore and digest the collection.

But we do not interact with electronic archives in this way. While more and more texts are available online, we simply do not have the human power to read and study them to provide the kind of browsing experience described above. To this end, machine learning researchers have developed *probabilistic topic modeling*, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information. Topic modeling algorithms are statistical methods that analyze the words of the original texts to discover the themes that run through them, how those themes are connected to each other, and how they change over

» **key insights**

- **Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.**

- **Topic modeling algorithms can be applied to massive collections of documents. Recent advances in this field allow us to analyze streaming collections, like you might find from a Web API.**

- **Topic modeling algorithms can be adapted to many kinds of data. Among other applications, they have been used to find patterns in genetic data, images, and social networks.**

time. (See, for example, Figure 3 for topics found by analyzing the *Yale Law Journal*.) Topic modeling algorithms do not require any prior annotations or labeling of the documents—the topics emerge from the analysis of the original texts. Topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

Latent Dirichlet Allocation

We first describe the basic ideas behind *latent Dirichlet allocation* (LDA), which is the simplest topic model.[8] The intuition behind LDA is that documents exhibit multiple topics. For example, consider the article in Figure 1. This article, entitled "Seeking Life's Bare (Genetic) Necessities," is about using data analysis to determine the number of genes an organism needs to survive (in an evolutionary sense).

By hand, we have highlighted different words that are used in the article. Words about *data analysis*, such as "computer" and "prediction," are highlighted in blue; words about *evolutionary biology*, such as "life" and "organism," are highlighted in pink; words about *genetics,* such as "sequenced" and "genes," are highlighted in yellow. If we took the time to highlight every word in the article, you would see that this article blends genetics, data analysis, and evolutionary biology in different proportions. (We exclude words, such as "and" "but" or "if," which contain little topical content.) Furthermore, knowing that this article blends those topics would help you situate it in a collection of scientific articles.

LDA is a statistical model of document collections that tries to capture this intuition. It is most easily described by its generative process, the imaginary random process by which the model assumes the documents arose. (The interpretation of LDA as a probabilistic model is fleshed out later.)

We formally define a *topic* to be a distribution over a fixed vocabulary. For example, the *genetics* topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability. We assume that these topics are specified before any data has been generated.[a] Now for each document in the collection, we generate the words in a two-stage process.

▸ Randomly choose a distribution over topics.
▸ For each word in the document
  a. Randomly choose a topic from the distribution over topics in step #1.
  b. Randomly choose a word from the corresponding distribution over the vocabulary.

This statistical model reflects the intuition that documents exhibit multiple topics. Each document exhibits the topics in different proportion (step #1); each word in each document is drawn from one of the topics (step #2b), where the selected topic is chosen from the per-document distribution over topics (step #2a).[b]

In the example article, the distribution over topics would place probability on *genetics*, *data analysis*, and

---

a Technically, the model assumes that the topics are generated first, before the documents.

b We should explain the mysterious name, "latent Dirichlet allocation." The distribution that is used to draw the per-document topic distributions in step #1 (the cartoon histogram in Figure 1) is called a *Dirichlet distribution*. In the generative process for LDA, the result of the Dirichlet is used to *allocate* the words of the document to different topics. Why *latent*? Keep reading.

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.
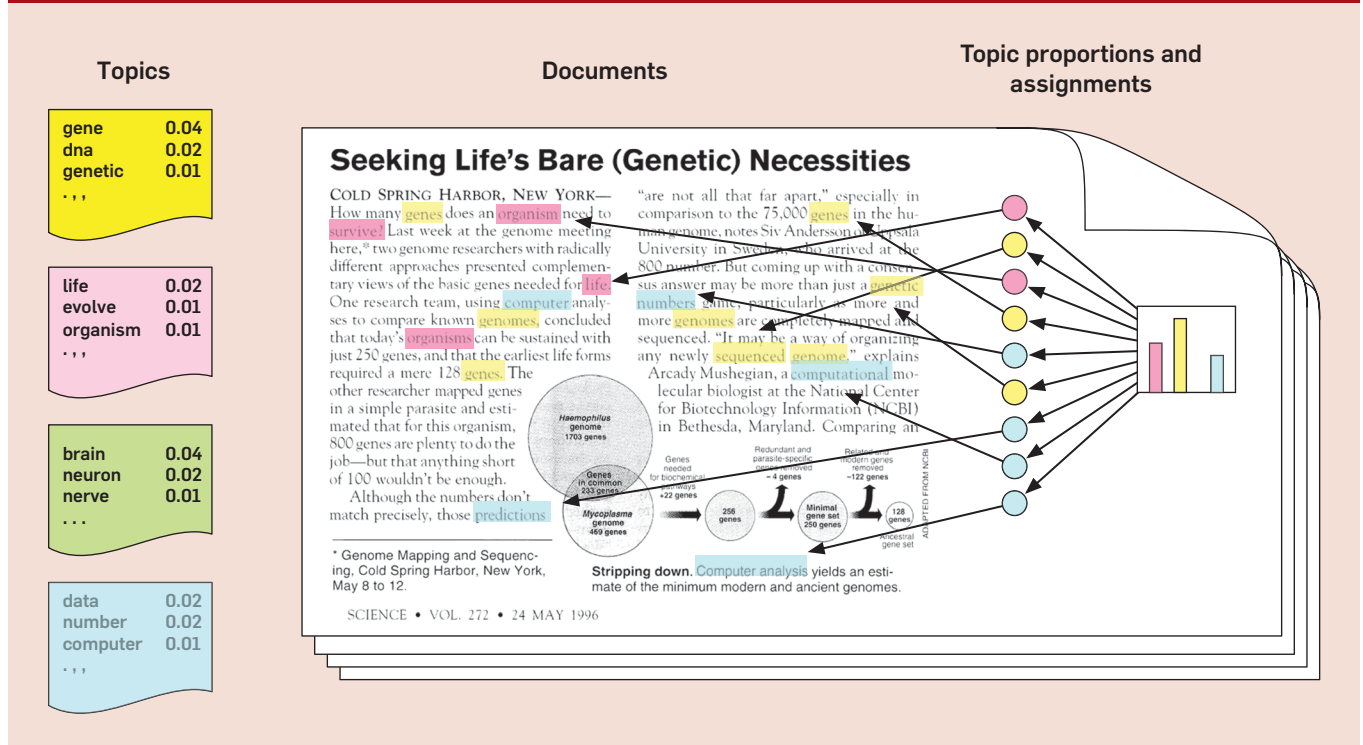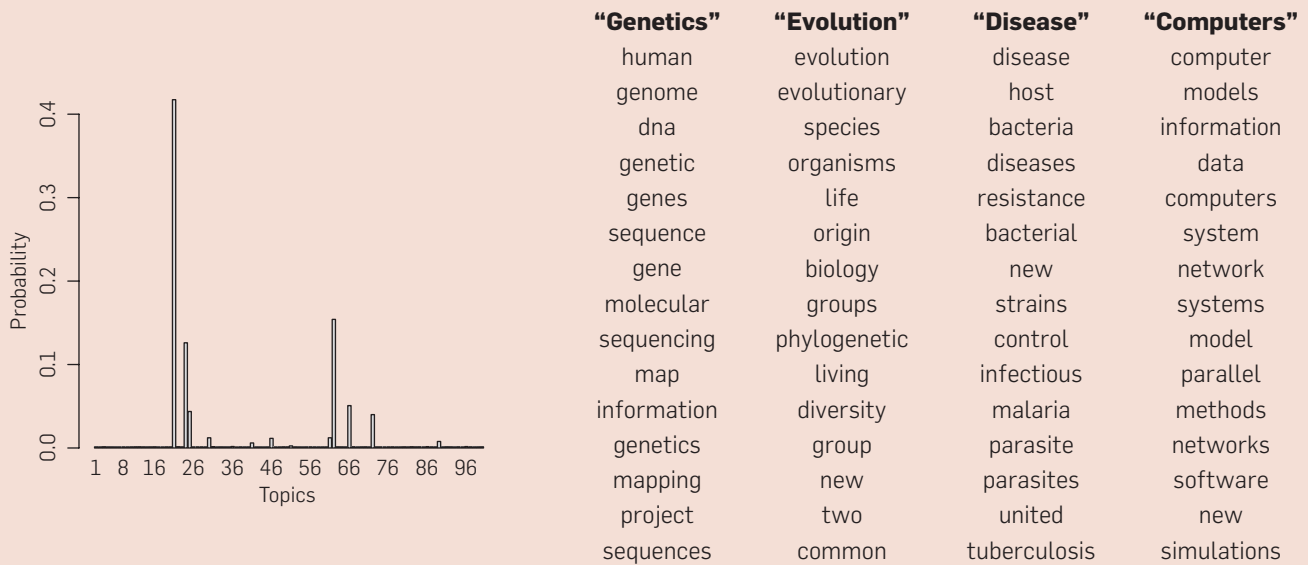
Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

evolutionary biology, and each word is drawn from one of those three topics. Notice that the next article in the collection might be about *data analysis* and *neuroscience*; its distribution over topics would place probability on those two topics. ==This is the distinguishing characteristic of latent Dirichlet allocation—all the documents in the collection share the same set of topics, but each document exhibits those topics in different proportion.==

As we described in the introduction, the goal of topic modeling is to automatically discover the topics from a collection of documents. The documents themselves are observed, while the topic structure—the topics, per-document topic distributions, and the per-document per-word topic assignments—is *hidden structure*. The central computational problem for topic modeling is to use the observed documents to infer the hidden topic structure. This can be thought of as "reversing" the generative process—what is the hidden structure that likely generated the observed collection?

Figure 2 illustrates example inference using the same example document from Figure 1. Here, we took 17,000 articles from *Science* magazine and used a topic modeling algorithm to infer the hidden topic structure. (The algorithm assumed that there were 100 topics.) We then computed the inferred topic distribution for the example article (Figure 2, left), the distribution over topics that best describes its particular collection of words. Notice that this topic distribution, though it can use any of the topics, has only "activated" a handful of them. Further, we can examine the most probable terms from each of the most probable topics (Figure 2, right). On examination, we see that these terms are recognizable as terms about genetics, survival, and data analysis, the topics that are combined in the example article.

We emphasize that the algorithms have no information about these subjects and the articles are not labeled with topics or keywords. The interpretable topic distributions arise by computing the hidden structure that likely generated the observed collection of documents.[c] For example, Figure 3 illustrates topics discovered from *Yale Law Journal*. (Here th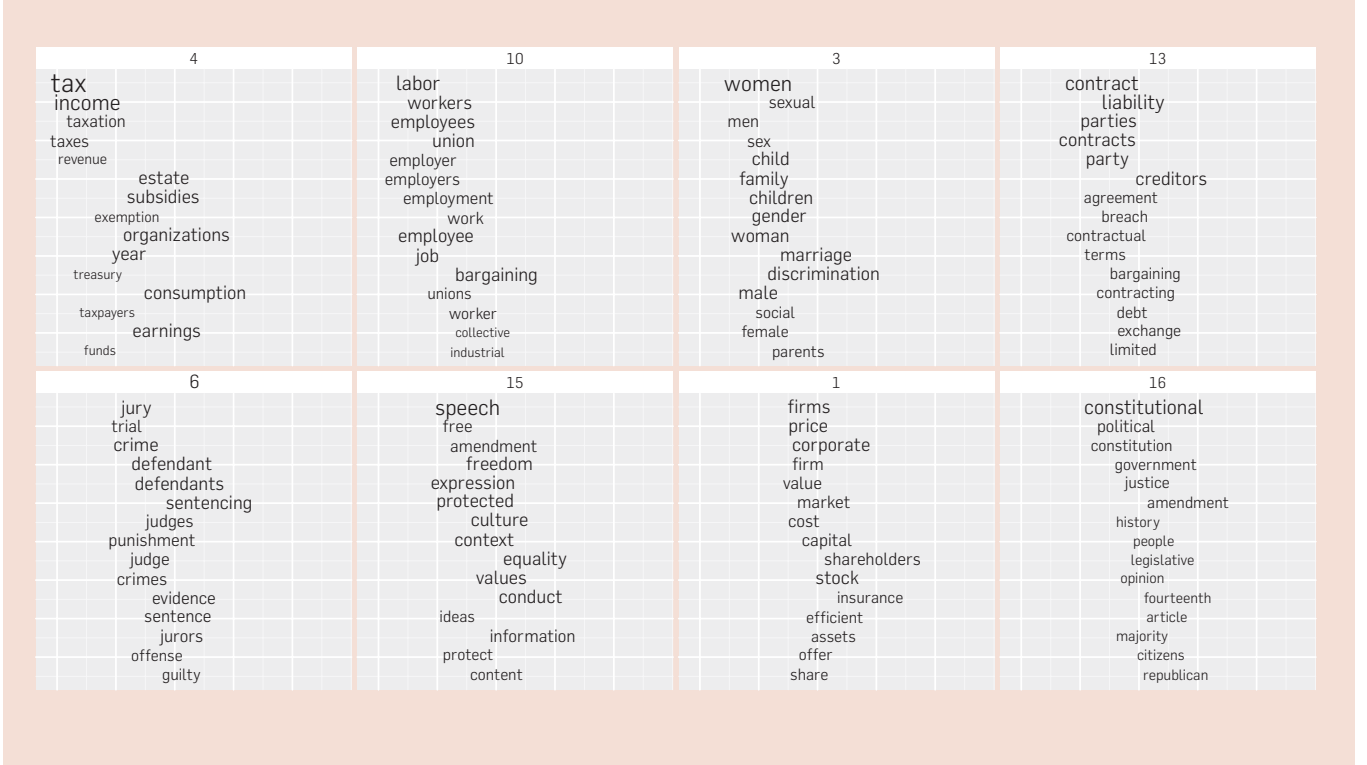e number of topics was set to be 20.) Topics about subjects like genetics and data analysis are replaced by topics about discrimination and contract law.

The utility of topic models stems from the property that the inferred hidden structure resembles the thematic structure of the collection. This interpretable hidden structure annotates each document in the collection—a task that is painstaking to perform by hand—and these annotations can be used to aid tasks like information retrieval, classification, and corpus exploration.[d] In this way, topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

LDA and probabilistic models. LDA and other topic models are part of the larger field of *probabilistic modeling*. In generative probabilistic modeling, we treat our data as arising from a generative process that includes *hidden variables*. This generative process defines a *joint probability distribution* over both the observed and hidden random variables. We perform data analysis by using that joint distribution to compute the *conditional distribution* of the hidden variables given the

---

c Indeed calling these models "topic models" is retrospective—the topics that emerge from the inference algorithm are interpretable for almost any collection that is analyzed. The fact that these look like topics has to do with the statistical structure of observed language and how it interacts with the specific probabilistic assumptions of LDA.

---

d See, for example, the browser of *Wikipedia* built with a topic model at http://www.sccs. swarthmore.edu/users/08/ajb/tmve/wiki100k/ browse/topic-list.html.

Figure 3. A topic model fit to the _Yale Law Journal_. Here, there are 20 topics (the top eight are plotted). Each topic is illustrated with its top-most frequent words. Each word's position along the _x_-axis denotes its specificity to the documents. For example "estate" in the first topic is more specific than "tax."

4	10	3	13
tax	labor	women	contract
income	workers	sexual	liability
taxation	employees	men	parties
taxes	union	sex	contracts
revenue	employer	child	party
estate	employers	family	creditors
subsidies	employment	children	agreement
exemption	work	gender	breach
organizations	employee	woman	contractual
year	job	marriage	terms
treasury	bargaining	discrimination	bargaining
consumption	unions	male	contracting
taxpayers	worker	social	debt
earnings	collective	female	exchange
funds	industrial	parents	limited

6	15	1	16
jury	speech	firms	constitutional
trial	free	price	political
crime	amendment	corporate	constitution
defendant	freedom	firm	government
defendants	expression	value	justice
sentencing	protected	market	amendment
judges	culture	cost	history
punishment	context	capital	people
judge	equality	shareholders	legislative
crimes	values	stock	opinion
evidence	conduct	insurance	fourteenth
sentence	ideas	efficient	article
jurors	information	assets	majority
offense	protect	offer	citizens
guilty	content	share	republican

observed variables. This conditional distribution is also called the _posterior distribution_.

LDA falls precisely into this framework. The observed variables are the words of the documents; the hidden variables are the topic structure; and the generative process is as described here. The computational problem of inferring the hidden topic structure from the documents is the problem of computing the posterior distribution, the conditional distribution of the hidden variables given the documents.

We can describe LDA more formally with the following notation. The topics are $\beta_{1:K}$, where each β_k is a distribution over the vocabulary (the distributions over words at left in Figure 1). The topic proportions for the dth document are θ_d, where $\theta_{d,k}$ is the topic proportion for topic k in document d (the cartoon histogram in Figure 1). The topic assignments for the dth document are z_d, where $z_{d,n}$ is the topic assignment for the nth word in document d (the colored coin in Figure 1). Finally, the observed words for document d are w_d, where $w_{d,n}$ is the nth word in document d, which is an element from the fixed vocabulary.

With this notation, the generative process for LDA corresponds to the following joint distribution of the hidden and observed variables,

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})$$
$$= \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d)$$
$$\left(\prod_{n=1}^{N} p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right). \quad (1)$$

Notice that this distribution specifies a number of dependencies. For example, the topic assignment $z_{d,n}$ depends on the per-document topic proportions θ_d. As another example, the observed word $w_{d,n}$ depends on the topic assignment $z_{d,n}$ and _all_ of the topics $\beta_{1:K}$. (Operationally, that term is defined by looking up as to which topic $z_{d,n}$ refers to and looking up the probability of the word $w_{d,n}$ within that topic.)

These dependencies define LDA. They are encoded in the statistical assumptions behind the generative process, in the particular mathematical form of the joint distribution, and—in a third way—in the _probabilistic graphical model_ for LDA. Probabilistic graphical models provide a graphical language for describing families of probability distributions.[e] The graphical model for LDA is in Figure 4. These three representations are equivalent ways of describing the probabilistic assumptions behind LDA.

In the next section, we describe the inference algorithms for LDA. However, we first pause to describe the short history of these ideas. LDA was developed to fix an issue with a previously developed probabilistic model _probabilistic latent semantic analysis_ (pLSI).[21] That model was itself a probabilistic version of the seminal work on _latent semantic analysis_,[14] which revealed the utility of the singular value decomposition of the document-term matrix. From this matrix factorization perspective, LDA can also be seen as a type of principal component analysis for discrete data.[11, 12]

Posterior computation for LDA. We now turn to the computational

---

e The field of graphical models is actually more than a language for describing families of distributions. It is a field that illuminates the deep mathematical links between probabilistic independence, graph theory, and algorithms for computing with probability distributions.[35]

problem, computing the conditional distribution of the topic structure given the observed documents. (As we mentioned, this is called the *posterior*.) Using our notation, the posterior is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} \mid w_{1:D})$$
$$= \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}. \qquad (2)$$

The numerator is the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator is the *marginal probability* of the observations, which is the probability of seeing the observed corpus under any topic model. In theory, it can be computed by summing the joint distribution over every possible instantiation of the hidden topic structure.

That number of possible topic structures, however, is exponentially large; this sum is intractable to compute.[f] As for many modern probabilistic models of interest—and for much of modern Bayesian statistics—we cannot compute the posterior because of the denominator, which is known as the *evidence*. A central research goal of modern probabilistic modeling is to develop efficient methods for approximating it. Topic modeling algorithms—like the algorithms used to create Figures 1 and 3—are often adaptations of general-purpose methods for approximating the posterior distribution.

Topic modeling algorithms form an approximation of Equation 2 by adapting an alternative distribution over the latent topic structure to be close to the true posterior. Topic modeling algorithms generally fall into two categories—sampling-based algorithms and variational algorithms.

Sampling-based algorithms attempt to collect samples from the posterior to approximate it with an empirical distribution. The most commonly used sampling algorithm for topic modeling is *Gibbs sampling*, where we construct a *Markov chain*—a sequence of random variables, each dependent on the previous—whose

limiting distribution is the posterior. The Markov chain is defined on the hidden topic variables for a particular corpus, and the algorithm is to run the chain for a long time, collect samples from the limiting distribution, and then approximate the distribution with the collected samples. (Often, just one sample is collected as an approximation of the topic structure with

---

f   More technically, the sum is over all possible ways of assigning each observed word of the collection to one of the topics. Document collections usually contain observed words at least on the order of millions.

Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.
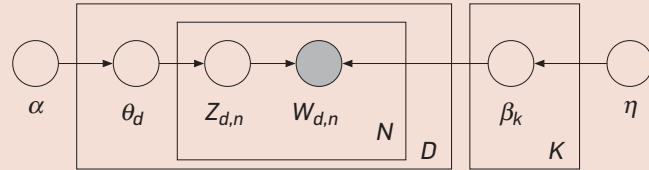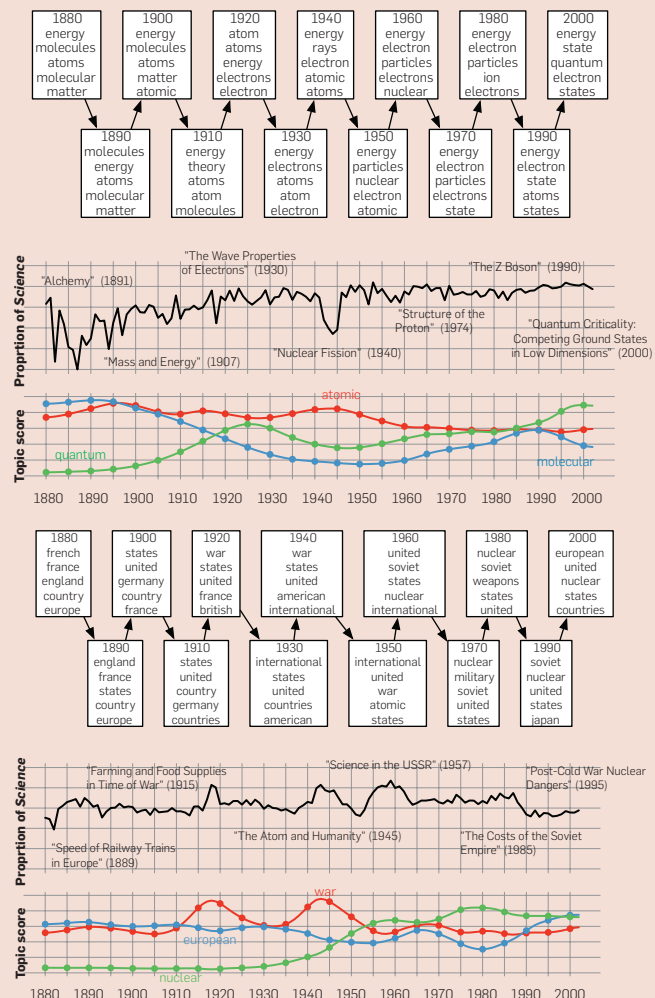

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.

maximal probability.) See Steyvers and Griffiths[33] for a good description of Gibbs sampling for LDA, and see http://CRAN.R-project.org/package=lda for a fast open-source implementation.

Variational methods are a deterministic alternative to sampling-based algorithms.[22,35] Rather than approximating the posterior with samples, variational methods posit a parameterized family of distributions over the hidden structure and then find the member of that family that is closest to the posterior.[g] Thus, the inference problem is transformed to an optimization problem. Variational methods open the door for innovations in optimization to have practical impact in probabilistic modeling. See Blei et al.[8] for a coordinate ascent variational inference algorithm for LDA; see Hoffman et al.[20] for a much faster online algorithm (and open-source software) that easily handles millions of documents and can accommodate streaming collections of text.

Loosely speaking, both types of algorithms perform a search over the topic structure. A collection of documents (the observed random variables in the model) are held fixed and serve as a guide toward where to search. Which approach is better depends on the particular topic model being used—we have so far focused on LDA, but see below for other topic models—and is a source of academic debate. For a good discussion of the merits and drawbacks of both, see Asuncion et al.[1]

Research in Topic Modeling
The simple LDA model provides a powerful tool for discovering and exploiting the hidden thematic structure in large archives of text. However, one of the main advantages of formulating LDA as a probabilistic model is that it can easily be used as a module in more complicated models for more complicated goals. Since its introduction, LDA has been extended and adapted in many ways.

Relaxing the assumptions of LDA. LDA is defined by the statistical assumptions it makes about the

One direction for topic modeling is to develop evaluation methods that match how the algorithms are used. How can we compare topic models based on how interpretable they are?

corpus. One active area of topic modeling research is how to relax and extend these assumptions to uncover more sophisticated structure in the texts.

One assumption that LDA makes is the "bag of words" assumption, that the order of the words in the document does not matter. (To see this, note that the joint distribution of Equation 1 remains invariant to permutation of the words of the documents.) While this assumption is unrealistic, it is reasonable if our only goal is to uncover the course semantic structure of the texts.[h] For more sophisticated goals—such as language generation—it is patently not appropriate. There have been a number of extensions to LDA that model words nonexchangeably. For example, Wallach[36] developed a topic model that relaxes the bag of words assumption by assuming that the topics generate words conditional on the previous word; Griffiths et al.[18] developed a topic model that switches between LDA and a standard HMM. These models expand the parameter space significantly but show improved language modeling performance.

Another assumption is that the order of documents does not matter. Again, this can be seen by noticing that Equation 1 remains invariant to permutations of the ordering of documents in the collection. This assumption may be unrealistic when analyzing long-running collections that span years or centuries. In such collections, we may want to assume that the *topics* change over time. One approach to this problem is the dynamic topic model[5]—a model that respects the ordering of the documents and gives a richer posterior topical structure than LDA. Figure 5 shows a topic that results from analyzing all of *Science* magazine under the dynamic topic model. Rather than a single distribution over words, a topic is now a sequence of distributions over words. We can find an underlying theme of the collection and track how it has changed over time.

A third assumption about LDA is that the number of topics is assumed

---

g Closeness is measured with *Kullback–Leibler divergence*, an information theoretic measurement of the distance between two probability distributions.

h As a thought experiment, imagine shuffling the words of the article in Figure 1. Even when shuffled, you would be able to glean that the article has something to do with genetics.

known and fixed. The Bayesian non-parametric topic model[34] provides an elegant solution: the number of topics is determined by the collection during posterior inference, and furthermore, new documents can exhibit previously unseen topics. Bayesian nonparametric topic models have been extended to hierarchies of topics, which find a tree of topics, moving from more general to more concrete, whose particular structure is inferred from the data.[3]

There are still other extensions of LDA that relax various assumptions made by the model. The correlated topic model[6] and pachinko allocation machine[24] allow the occurrence of topics to exhibit correlation (for example, a document about *geology* is more likely to also be about *chemistry* than it is to be about *sports*); the spherical topic model[28] allows words to be *unlikely* in a topic (for example, "wrench" will be particularly unlikely in a topic about *cats*); sparse topic models enforce further structure in the topic distributions;[37] and "bursty" topic models provide a more realistic model of word counts.[15]

Incorporating metadata. In many text analysis settings, the documents contain additional information—such as author, title, geographic location, links, and others—that we might want to account for when fitting a topic model. There has been a flurry of research on adapting topic models to include metadata.

The author-topic model[29] is an early success story for this kind of research. The topic proportions are attached to authors; papers with multiple authors are assumed to attach each word to an author, drawn from a topic drawn from his or her topic proportions. The author-topic model allows for inferences about authors as well as documents. Rosen-Zvi et al. show examples of author similarity based on their topic proportions—such computations are not possible with LDA.

Many document collections are linked—for example, scientific papers are linked by citation or Web pages are linked by hyperlink—and several topic models have been developed to account for those links when estimating the topics. The *relational topic model* of Chang and Blei[13] assumes that each document is modeled as in LDA and that the links

between documents depend on the distance between their topic proportions. This is both a new topic model and a new network model. Unlike traditional statistical models of networks, the relational topic model takes into account node attributes (here, the words of the documents) in modeling the links.

Other work that incorporates metadata into topic models includes models of linguistic structure,[10] models that account for distances between corpora,[38] and models of named entities.[26] General-purpose methods for incorporating metadata into topic models include Dirichlet-multinomial regression models[25] and supervised topic models.[7]

Other kinds of data. In LDA, the topics are distributions over words and this discrete distribution generates observations (words in documents). One advantage of LDA is that these choices for the topic parameter and data-generating distribution can be adapted to other kinds of observations with only small changes to the corresponding inference algorithms. As a class of models, LDA can be thought of as a *mixed-membership model* of grouped data—rather than associating each group of observations (document) with one component (topic), each group exhibits multiple components in different proportions. LDA-like models have been adapted to many kinds of data, including survey data, user preferences, audio and music, computer code, network logs, and social networks. We describe two areas where mixed-membership models have been particularly successful.

In population genetics, the same probabilistic model was independently invented to find ancestral populations (for example, originating from Africa, Europe, the Middle East, among others) in the genetic ancestry of a sample of individuals.[27] The idea is that each individual's genotype descends from one or more of the ancestral populations. Using a model much like LDA, biologists can both characterize the genetic patterns in those populations (the "topics") and identify how each individual expresses them (the "topic proportions"). This model is powerful because the genetic patterns in ancestral populations can be hypothesized, even when "pure" samples from them are not available.

LDA has been widely used and adapted in computer vision, where the

inference algorithms are applied to natural images in the service of image retrieval, classification, and organization. Computer vision researchers have made a direct analogy from images to documents. In document analysis, we assume that documents exhibit multiple topics and the collection of documents exhibits the same set of topics. In image analysis, we assume that each image exhibits a combination of visual patterns and that the same visual patterns recur throughout a collection of images. (In a preprocessing step, the images are analyzed to form collections of "visual words.") Topic modeling for computer vision has been used to classify images,[16] connect images and captions,[4] build image hierarchies,[2,23,31] and other applications.

Future Directions

Topic modeling is an emerging field in machine learning, and there are many exciting new directions for research.

Evaluation and model checking. There is a disconnect between how topic models are evaluated and why we expect topic models to be useful. Typically, topic models are evaluated in the following way. First, hold out a subset of your corpus as the test set. Then, fit a variety of topic models to the rest of the corpus and approximate a measure of model fit (for example, probability) for each trained model on the test set. Finally, choose the model that achieves the best held-out performance.

But topic models are often used to organize, summarize, and help users explore large corpora, and there is no technical reason to suppose that held-out accuracy corresponds to better organization or easier interpretation. One open direction for topic modeling is to develop evaluation methods that match how the algorithms are used. How can we compare topic models based on how interpretable they are?

This is the *model checking* problem. When confronted with a new corpus and a new task, which topic model should I use? How can I decide which of the many modeling assumptions are important for my goals? How should I move between the many kinds of topic models that have been developed? These questions have been given some attention by statisticians,[9,30] but they have been scrutinized less for the scale

of problems that machine learning tackles. New computational answers to these questions would be a significant contribution to topic modeling.

Visualization and user interfaces. Another promising future direction for topic modeling is to develop new methods of interacting with and visualizing topics and corpora. Topic models provide new exploratory structure in large collections—how can we best exploit that structure to aid in discovery and exploration?

One problem is how to display the topics. Typically, we display topics by listing the most frequent words of each (see Figure 2), but new ways of labeling the topics—by either choosing different words or displaying the chosen words differently—may be more effective. A further problem is how to best display a document with a topic model. At the document level, topic models provide potentially useful information about the structure of the document. Combined with effective topic labels, this structure could help readers identify the most interesting parts of the document. Moreover, the hidden topic proportions implicitly connect each document to the other documents (by considering a distance measure between topic proportions). How can we best display these connections? What is an effective interface to the whole corpus and its inferred topic structure?

These are user interface questions, and they are essential to topic modeling. Topic modeling algorithms show much promise for uncovering meaningful thematic structure in large collections of documents. But making this structure useful requires careful attention to information visualization and the corresponding user interfaces.

Topic models for data discovery. Topic models have been developed with information engineering applications in mind. As a statistical model, however, topic models should be able to tell us something, or help us form a hypothesis, about the data. What can we *learn* about the language (and other data) based on the topic model posterior? Some work in this area has appeared in political science,[19] bibliometrics,[17] and psychology.[32] This kind of research adapts topic models to measure an external variable of interest, a difficult task for unsupervised learning that must be carefully validated.

In general, this problem is best addressed by teaming computer scientists with other scholars to use topic models to help explore, visualize, and draw hypotheses from their data. In addition to scientific applications, such as genetics and neuroscience, one can imagine topic models coming to the service of history, sociology, linguistics, political science, legal studies, comparative literature, and other fields, where texts are a primary object of study. By working with scholars in diverse fields, we can begin to develop a new interdisciplinary computational methodology for working with and drawing conclusions from archives of texts.

Summary

We have surveyed *probabilistic topic models*, a suite of algorithms that provide a statistical solution to the problem of managing large archives of documents. With recent scientific advances in support of unsupervised machine learning—flexible components for modeling, scalable algorithms for posterior inference, and increased access to massive datasets— topic models promise to be an important component for summarizing and understanding our growing digitized archive of information. Ⓒ

References
1. Asuncion, A., Welling, M., Smyth, P., Teh, Y. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence* (2009).
2. Bart, E., Welling, M., Perona, P. Unsupervised organization of image collections: Taxonomies and beyond. *Trans. Pattern Recognit. Mach. Intell. 33,* 11 (2010) (2301–2315).
3. Blei, D., Griffiths, T., Jordan, M. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *J. ACM 57,* 2 (2010), 1–30.
4. Blei, D., Jordan, M. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2003), ACM Press, 127–134.
5. Blei, D., Lafferty, J. Dynamic topic models. In *International Conference on Machine Learning* (2006), ACM, New York, NY, USA, 113–120.
6. Blei, D., Lafferty, J. A correlated topic model of Science. *Ann. Appl. Stat., 1,* 1 (2007), 17–35.
7. Blei, D., McAuliffe, J. Supervised topic models. In *Neural Information Processing Systems* (2007).
8. Blei, D., Ng, A., Jordan, M. Latent Dirichlet allocation. *J. Mach. Learn. Res. 3* (January 2003), 993–1022.
9. Box, G. Sampling and Bayes' inference in scientific modeling and robustness. *J. Roy. Stat. Soc. 143,* 4 (1980), 383–430.
10. Boyd-Graber, J., Blei, D. Syntactic topic models. In *Neural Information Processing Systems* (2009).
11. Buntine, W. Variational extensions to EM and multinomial PCA. In *European Conference on Machine Learning* (2002).
12. Buntine, W., Jakulin, A. Discrete component analysis. *Subspace, Latent Structure and Feature Selection.* C. Saunders, M. Grobelink, S. Gunn, and J. Shawe-Taylor, Eds. Springer, 2006.
13. Chang, J., Blei, D. Hierarchical relational models for document networks. *Ann. Appl. Stat. 4,* 1 (2010).
14. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci. 41,* 6 (1990), 391–407.
15. Doyle, G., Elkan, C., Accounting for burstiness in topic models. In *International Conference on Machine Learning* (2009), ACM, 281–288..
16. Fei-Fei, L., Perona, P. A Bayesian hierarchical model for learning natural scene categories. In *IEEE Computer Vision and Pattern Recognition* (2005), 524–531.
17. Gerrish, S., Blei, D. A language-based approach to measuring scholarly impact. In *International Conference on Machine Learning* (2010).
18. Griffiths, T., Steyvers, M., Blei, D., Tenenbaum, J. Integrating topics and syntax. *Advances in Neural Information Processing Systems 17.* L. K. Saul, Y. Weiss, and L. Bottou, eds. MIT Press, Cambridge, MA, 2005, 537–544.
19. Grimmer, J. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Polit. Anal. 18,* 1 (2010), 1.
20. Hoffman, M., Blei, D., Bach, F. On-line learning for latent Dirichlet allocation. In *Neural Information Processing Systems* (2010).
21. Hofmann, T. Probabilistic latent semantic analysis. In *Uncertainty in Artificial Intelligence (UAI)* (1999).
22. Jordan, M., Ghahramani, Z., Jaakkola, T., Saul, L. Introduction to variational methods for graphical models. *Mach. Learn. 37* (1999), 183–233.
23. Li, J., Wang, C., Lim, Y., Blei, D., Fei-Fei, L., Building and using a semantivisual image hierarchy. In *Computer Vision and Pattern Recognition* (2010).
24. Li, W., McCallum, A. Pachinko allocation: DAG-structured mixture models of topic correlations. In *International Conference on Machine Learning* (2006), 577–584.
25. Mimno, D., McCallum, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence* (2008).
26. Newman, D., Chemudugunta, C., Smyth, P. Statistical entity-topic models. In *Knowledge Discovery and Data Mining* (2006).
27. Pritchard, J., Stephens, M., Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics 155* (June 2000), 945–959.
28. Reisinger, J., Waters, A., Silverthorn, B., Mooney, R. Spherical topic models. In *International Conference on Machine Learning* (2010).
29. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smith, P., The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), AUAI Press, 487–494.
30. Rubin, D. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat. 12,* 4 (1984), 1151–1172.
31. Sivic, J., Russell, B., Zisserman, A., Freeman, W., Efros, A., Unsupervised discovery of visual object class hierarchies. In *Conference on Computer Vision and Pattern Recognition* (2008).
32. Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., Norman, K. A Bayesian analysis of dynamics in free recall. In *Advances in Neural Information Processing Systems 22.* Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds, 2009.
33. Steyvers, M., Griffiths, T. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning.* T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, eds. Lawrence Erlbaum, 2006.
34. Teh, Y., Jordan, M., Beal, M., Blei, D. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc. 101,* 476 (2006), 1566–1581.
35. Wainwright, M., Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn. 1*(1–2) (2008), 1–305.
36. Wallach, H. Topic modeling: Beyond bag of words. In *Proceedings of the 23rd International Conference on Machine Learning* (2006).
37. Wang, C., Blei, D. Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. *Advances in Neural Information Processing Systems 22.* Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. 2009, 1982–1989.
38. Wang, C., Thiesson, B., Meek, C., Blei, D. Markov topic models. In *Artificial Intelligence and Statistics* (2009).

David M. Blei (blei@cs.princeton.edu) is an associate professor in the computer science department of Princeton University, Princeton, N.J.