

Chapter Title: Interested Learning

Book Title: Engines of Order

Book Subtitle: A Mechanology of Algorithmic Techniques

Book Author(s): Bernhard Rieder

Published by: Amsterdam University Press. (2020)

Stable URL: <https://www.jstor.org/stable/j.ctv12sdvf1.10>

---

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



This book is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0). To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.



JSTOR

*Amsterdam University Press* is collaborating with JSTOR to digitize, preserve and extend access to *Engines of Order*

## 6. Interested Learning

### Abstract

This chapter examines one of the most active areas where feature vectors play a central role: machine learning. The Bayes classifier is used as an entry point into the field, showing how a simple statistical technique introduced in the early 1960s is surprisingly instructive for understanding how machine learning operates more broadly. The goal is to shed light on the core principles at work and to explain how they are tweaked, adapted, and developed further into different directions. This chapter also develops the idea that contemporary information ordering represents an epistemological practice that can be described and analyzed as ‘interested reading of reality’, a particular kind of inductive empiricism.

**Keywords:** machine learning, Bayes classifier, optimization, interested readings of reality

Coordinate indexing and its relational database cousin constitute a technique for laying a disjoint, atomized, and therefore highly queryable grid over the world. Ordering becomes a particular operationalization of disassembly and reassembly in a multidimensional coordinate system that affords purpose-driven grouping and ordering through logical declaration. The query specifies the coordinates or properties one is interested in and matching items are retrieved. The statistical techniques discussed in the last chapter, which often describe their work as automatic encoding or indexing, can be seen as attempts to build such a grid from the world itself, not necessarily by grammatization, ‘the process whereby the currents and continuities shaping our lives become discrete elements’ (Stiegler, 2010, p. 70), but by counting what already appears as discrete units. In the case of information retrieval, this mainly concerned the words appearing in a text document, but one could easily imagine users as items and elements like posts, songs, or products they interacted with in some way as properties to begin clustering or recommending on the base of similarity calculations. The various techniques grouping around the VSM

measure similarity or distance, rank, cluster, or otherwise arrange items in the already discretized environments computerization multiplies. The notion of specificity, as it appears, for example, in tf-idf, shows how these processes can draw on the relationships between individuals and populations in (high-dimensional) data spaces to glean additional semantic salience.

The space of investigation, experimentation, and steady accumulation of concepts and techniques opened up by Luhn and others quickly develops in a number of directions. One of them points toward an epistemological posture emblematically expressed in an interview with Michael Richard Lynch, the cofounder of Autonomy Corporation, a company now owned by HP that specializes in mining large amounts of ‘unstructured’ data:

‘Rules-based, Boolean computing assumes that we know best how to solve a problem,’ he says. ‘My background comes completely the other way. The problem tells you how to solve the problem. That’s what the next generation of computing is going to be about: listening to the world.’<sup>1</sup>

In a sense, ‘listening to the world’ is what statistical techniques broadly set out to do, but this process can be enacted and orchestrated in different ways. With the exception of a short excursion into unsupervised machine learning used for clustering, the techniques discussed in the last chapter were either used for document indexing or created with the expectation that actual searching and ranking of items would be performed in relation to a query, for example, by measuring the similarity between a search vector and any number of document vectors. A contemporary adaptation of the same principle for purposes of personalization or recommendation would be to treat a user profile as an ‘indirect’ query – there is still an individual reference point that serves to line up the population.

In this chapter, however, I want to explore a technical lineage, again inceptioned in the 1960s in the context of information retrieval, which uses similar principles to order information not in response to some explicitly formulated need or desire for information, but to orchestrate a more implicit and potentially autonomous classificatory process that enacts a form of continuous and adaptive observation to derive decision-making principles. No less ambiguous than notions like meaning or relevance, this highlights the question of how techniques can make machines ‘learn’.

In this context, the Bayes classifier constitutes an ideal example. First, it connects directly to the nexus of techniques already discussed, both

1 Lynch, cit. in Silberman (2000, n.p.).

historically and conceptually. The Bayes classifier builds on principles already familiar from previous chapters and can be understood without a deeper dive into probability mathematics. Second, since it constitutes a technique that is probabilistic (classifications are not binary but with degrees of certainty), adaptive (it 'learns' from experience), and well suited for personalization, it allows for a discussion of broader aspects of machine learning and its role in contemporary information ordering. Even if many new and much more complex techniques have appeared in recent decades, the Bayes classifier is still one of the 'most efficient and effective inductive learning algorithms for machine learning and data mining' (Zhang, 2004, p. 1) and has 'remained popular over the years' (Hastie, Tibshirani, and Friedman, 2009, p. 211). In very general terms, it provides a specific method for making use of statistical inference to sort a new element, for example, an incoming email, on the basis of a decision model derived from previously categorized elements, for example, messages already marked as spam. This canonical example can provide a guiding rationale through the more detailed historical and technical presentation that follows.

### **From Probabilistic Coordinate Indexing to the Bayes Classifier**

The story begins in the late 1950s when M. E. Maron,<sup>2</sup> a physicist turned analytical philosopher and cybernetician working at the Ramo-Wooldridge Corporation and the Rand Corporation before becoming a full professor at Berkeley's School of Information in 1966, 'was thinking hard about the problem of information retrieval' (Maron, 2008, p. 971). He was particularly unsatisfied with a fundamental aspect of coordinate indexing, namely that assigning a tag (or keyword, uniterm, descriptor, etc.) to an item in order to describe its subject was 'a two-valued affair' (Maron, 2008, p. 971): a term is either attributed or not, nothing in between. In Maron's view, this makes the system too 'primitive' since it cannot express 'degrees of aboutness' (Maron, 2008, p. 971), an obvious limitation for the representation of document content itself and a problem for retrieval and ranking. Pure Boolean

2 This presentation focuses on the work of M. E. Maron, although a fuller account of the use of Bayes's theorem for classification and learning would have to include Solomonoff's experiments on 'a machine which is designed to learn to work problems in mathematics' (1957, p. 56), but this would stray too much from the focus on document classification. Another interesting addition would be the work of Borko (1962) and, in particular, Borko and Bernick (1963). The latter repeats Maron's earlier experiments, but uses factor analysis instead of Bayes's theorem, with similar results.

matching certainly constitutes a powerful way to manipulate an information space through postcoordination, but it lacks nuance and provides no means to weight keywords or index terms. This becomes problematic as collections grow and queries return large numbers of documents, without much possibility for additional differentiation other than the number of search terms matched.

The first improvement Maron and colleague J. L. Kuhns proposed was not to count word occurrences in documents like Luhn did but for indexers to specify a value somewhere between 0 and 1 to indicate the relevance of a tag for a document. This method, named 'probabilistic indexing', made it possible, through inverse statistical inference via Bayes's theorem, to generate a probabilistically ranked result list for a subject query instead of merely an unordered set of documents (Maron and Kuhns, 1960). Leaving aside its complicated history (Stigler, 1983) and various interpretations, Bayes's theorem basically provides a simple method for calculating the probability of a hypothesis being true (or an event occurring) based on existing knowledge. For example, if we know the percentage of women in a population and the a priori percentages of women and men with long hair, we can calculate the probability of the hypothesis that a person with long hair we see only from behind is a woman.<sup>3</sup> If we add other variables to the equation, such as height or clothing style, the assessment becomes increasingly specific. Bayes's theorem thus provides a means to reason in a space of uncertainty where we have some prior knowledge that can be used to assess a particular case. My example already shows that this kind of reasoning implies not only potentially problematic forms of categorization, such as the division of people into women and men, but also raises the question of how 'prior knowledge' is constructed and operationalized.

In the case of information retrieval, the 'hypothesis' under investigation concerns the probability that a document (tagged with a number of weighted terms) is relevant for a query (represented by a number of search terms that could also be weighted). The indexers provide 'prior knowledge' by establishing probability relationships between keywords and documents. Interestingly, Maron proposed from the outset to replace the a priori probability of a document (the equivalent of the 'percentage

3 To calculate this example, let us consider that half of the population is female and that 75% of all women have long hair and 15% of all men. The probability that a person with long hair is a woman would then be calculated as:

$$P(F|L) = \frac{0.75 * 0.50}{0.75 * 0.50 + 0.15 * 0.50} = 0.83333$$

of women in a population'), which would normally be one divided by the number of documents in the collection, with statistics on document use. In the end, Maron's formula for relevance looked like this, using a dot product<sup>4</sup> to calculate the 'closeness' between query terms ('WordsQuery') and the document terms ('WordsDoc'):

$$P(\text{DocumentIsRelevant}) = (\text{WordsQuery} \cdot \text{WordsDoc}) * P(\text{DocumentUse})$$

A simplified version of Maron's relevance formula, which would be calculated for each document that has at least one term in common with the query.

We again encounter the familiar representation of documents and queries as term frequency lists or feature vectors and the closeness calculation between the two is comparable to the similarity coefficients discussed in the last chapter. When searching for [hydraulics], a document tagged with 0.7 for that term would thus have a higher 'relevance number' than a document tagged with a value of 0.3. Documents accessed more often would also receive a higher value. Results were then ranked according to their relevance number. Term combinations and weighted search terms were possible as well. The method was still based on manual indexing and manual weighting but represented nonetheless 'a theoretical attack which replaced traditional two-valued indexing and matching with a statistical approach [...] to make predictions about the relevance of documents in the collection' (Thompson, 2008, p. 964). While indexing with manual term weighting never became a common practice, the idea explicitly framed information retrieval as an operation under uncertainty and thus not sufficiently well served by Boolean logic.

The indexing and ranking technique I just outlined is not a Bayes classifier, however. It introduced Bayes's theorem into information retrieval, but the actual classifier emerged only through a second experiment, which attempted to do away with the human indexer, who Maron described, in line with other information scientists of the time, as slow, unreliable, and biased (Maron, 1961). Although he did not reference Luhn's work, the chosen strategy was to tackle the full text of the document itself. Based on the assertion that 'statistics on kind, frequency, location, order, etc., of

4 A dot product is the sum of the products of two sequences of numbers. For example, the dot product between the query { hydraulics: 0.7, car: 0.5 } and the document terms { hydraulics: 0.6, car: 0.8 } would be  $0.7 * 0.6 + 0.5 * 0.8$ , thus 0.82. The more terms overlap between a query and a document and the higher the weight of the terms, the more 'relevant' the document.

selected words are adequate to make reasonably good predictions about the subject matter of documents containing those words' (Maron, 1961, p. 405), Maron devised a technique for the automatic classification of documents. His approach to meaning remained prudent yet pragmatic: considering words as isolated atoms – the only practically feasible formalization at the time – was certainly not ideal but said at least *something* about the subject matter a document discusses, 'reasonably good' for the task at hand. The technique thus again conceived text documents as 'objects' (or items) having certain 'properties' (or terms, features, etc.), but this time Maron decided to represent each text by the words it actually contained rather than sets of manually chosen descriptors (Maron 1961, p. 406). We end up with a feature vector in both cases, but the former implies a higher degree of automation and echoes Luhn's (1968b) empiricist notion of a 'native' thesaurus or classification that is 'born of the collection' (p. 168).

Classification, in terms of the probabilistic approach proposed by Maron, meant that text documents were to be sorted into user-specified subject categories: '[b]ased on some more or less clear notion of the category, we must decide whether or not an arbitrary document should be assigned to it' (Maron 1961, p. 404). A simple contemporary example is indeed spam filtering: emails are documents and categories are 'spam' and 'not-spam' (or 'legitimate'). The task is to sort incoming emails into these categories. As Rish (2001) summarizes, 'Bayesian classifiers assign the most likely class to a given example described by its feature vector' (p. 41).

The first step in Maron's setup, which used a collection of 405 scientific abstracts in digital form, was to select a number of characteristic 'training' documents for each one of the subject categories. Human intervention was thus rearranged, not eliminated: while 'prior knowledge' in probabilistic indexing came from the explicit attribution of terms, it now shifted to the less explicit act of labeling or sorting example documents into categories. A combined word list for each category was generated from the initially classified documents. Not all words were retained. Just like in Luhn's experiments and based on the same intuitions, very frequent words and very rare words were discarded. The resulting selection was submitted to a technique similar in spirit to the tf-idf metric introduced by Spärck Jones and Salton: words that were evenly distributed over all categories and did not 'peak' anywhere were considered inadequate 'clues' and thus rejected (Maron, 1961, p. 408). The statistical horizon was again used to single out the most specific signals.

An index where each retained word received a relevance value for each category was then generated, determining 'certain probability relationships

between individual content-bearing words and subject categories' (Maron, 1961, p. 405). In a nutshell: if a word appears very often in the training documents assigned to a certain category but rarely in others, it becomes a strong clue or indicator for that category. Once the training phase was complete, human intervention was no longer required: based on the general idea that a document should be attributed to a category if it contains many good indicators for that category and few for others, new documents could be automatically classified in the same way as documents were matched with queries in the earlier experiment. In fact, there were now as many 'queries' as categories, and the task was to decide for each document which query it was 'closest' to. The method was therefore very similar to the ranking procedure in Maron and Kuhn's first experiment, but instead of determining the fit between query terms and document terms, the fit between the word list for each category and the word list for each incoming document was calculated by adding up the probability relationships for all of the 'clue words'. The technique again follows the logic that Spärck Jones advocates as the most appropriate for many tasks that are 'underspecified' or 'fundamentally inexact', namely 'to allow many small and individually ambiguous clues to combine and interact' (Spärck Jones, 1999, p. 414).

Since many words<sup>5</sup> were taken into account, every document was likely to receive a relevance number (or 'posterior probability') for several categories, for example, document  $n$  is 0.4 relevant for category  $i$ , 0.2 for category  $j$ , and so forth, resulting in probabilistic rather than binary classification. In the case of Maron's experiment, the categories would actually represent subject descriptors and the final output of the system would be an automatically generated probabilistic index that could be used to search and rank in the same way as the manually created index used in the first experiment. The basic coordinate indexing setup remains the reference point. Unlike a manually generated index, however, the system was dynamic by design in the sense that it incorporated a pathway for learning beyond the initial training phase: if new documents were classified or a user decided to reclassify a document, the word lists for each category could be recalculated, adding new 'knowledge' to the statistical model.

This is the basic outline of a Bayes classifier and a surprisingly representative illustration of the larger field of supervised machine learning, where techniques require classified or 'labeled' data from which a model

5 Although Maron's first experiment used only 90 'clue words' to cover 405 research paper abstracts (including both the training set and the documents to classify), his method was explicitly aimed at much larger sets (cf. Maron, 1961 p. 414).



representing relationships between inputs (feature vectors) and outputs (class attributions) can be derived. This model can then be applied to new inputs. There are a number of important aspects to consider.

## The Technicity of Learning

Bayes classifiers connect back to the previous chapters in various ways, but they also raise specific questions that are crucial for an appreciation of the larger space of contemporary information ordering. In this section, I want to address some of these questions from the angle of technicity before addressing the epistemological dimension in more depth in the following section. Five points stand out.

### Understanding Choice

First, I want to stress that the outlined procedure constitutes precisely what the term 'algorithmic technique' attempts to thematize. What I have laid out in plain English and some basic formulas is not yet an algorithm in a more restrictive understanding of the term, but it outlines a method for classification that entails a way of both looking at and acting on the world. It frames and formalizes text documents as word frequency lists, formulates a sequence of stages from training to classification to adjustment, and specifies a number of proto-mathematical functions for weighting and calculating. Any software developer would be able to create a working program from my description, but every implementation would be different since many details remain underspecified and require decisions. Should words be reduced to their lemmas or stems? Where to cut off frequent and infrequent words? How to calculate word specificity? How to calculate closeness? Should document use become part of relevancy assessments? How to embed all of this in interface and activity flows? These and other questions need to be answered when an algorithmic technique is brought to bear on a specific task in a specific operating environment.

The consequences of these decisions are hard to estimate beforehand, due to the potentially large number of items, properties, and iterations. And there are not only technical reasons why the design process should not be understood as a selection from straightforward options: a working system is the outcome of a complex process of negotiation that can have far-reaching, unanticipated consequences when becoming part of concrete practices. Maron's decision to integrate document use statistics, for example, may

indeed improve some empirical measure of retrieval performance or user satisfaction, but it also introduces the principle of cumulative advantage or ‘Matthew effect’ (Merton, 1968) into ranking dynamics, since documents with greater use will end up further up on the result list, leading to even higher use as a consequence.

This again highlights that the study of algorithmic techniques is not enough to make sense of actual systems, their behavior, the many specific commitments they imply, and the actual connections and dynamics they enter into. But a robust apprehension of common techniques can both embed more general forms of theorizing in a better understanding of software-making and inform more concrete forms of empirical analysis. It can lay the groundwork for comparison between different implementations by proving analytical categories, for example, selected units and features, intermediate forms, training and feedback setup, decision modalities, and so forth. The notion of algorithmic technique describes a pathway for asking ‘how the machine “thinks”’ (Burrell, 2016) and, by extension, how software-makers embedded in concrete circumstances think through or with the machine, in the sense that instances of technicity provide means of expression in the form of function that reflect both technical trajectories and local purposes and specificities.

## Dramaturgy

Second, Maron follows Luhn’s path in framing meaning as ‘aboutness’ and relies on the same reduction to word frequencies as language model. Although techniques using more involved formalizations of language both can and do exist, any running system requires and relies in some way on formalization, selection, and reduction. This more often than not implies the already familiar distillation into a common intermediate form, a gesture that enables and explains the wide applicability of algorithmic techniques. Bayes classifiers, much like other classification techniques, can be used to produce groupings of any class of objects represented by feature lists, that is, of ‘any set of entities identified by weighted property vectors’ (Salton, Wong, and Yang, 1975, p. 613).

What is being introduced into the arsenal of information ordering, however, is not so much the capacity to group, which is already realized through queries in the coordinate indexing setup or through some optimal arrangement of similarities and differences in unsupervised learning, but the idea that we can ‘show’ the machine *how* to group. Whereas Boolean or SQL queries specify logical selection criteria and unsupervised

techniques look for ‘natural’ zones of density, the supervised approach implies a different orchestration: the ‘correct’ interpretation of the data is provided for a set of training items, the machine derives a model (or function) that maps inputs (feature vectors) to outputs (‘target variables’, i.e., classes), and that model is then applied to data that has not yet been classified. Based on the same underlying intermediate form, the ordering process thus follows a different dramaturgy. Crucially, the setup revolves around forms of supervision or ‘labeling’ and the move to integrated online environments, which I have addressed as ‘computerization’ in Chapter 1, provides not only the opportunity to normalize almost anything to the common form of the feature vector but also constant streams of feedback that function as signals for continuous supervision and training. Schäfer (2011, p. 51) refers to situations where ‘users are actually not aware that they contribute to an application simply through using it’ as ‘implicit participation’, a value-generating mechanism that is sunk into the design of the system. Voting, rating, or liking are only the most obvious ways to train the Bayesian engines: every click, every interaction, every emoji, every act of paying attention to something can be formalized as an instance of labeling and flow into the classifier. These techniques thus ‘feed’ on digital media environments on both ends: they receive both masses of data to classify and constant feedback regarding how well they did and how they can further improve. Optimization for specific classificatory goals becomes part of adaptive infrastructures as it is “sunk” into’ (Star and Ruhleder, 1996, p. 115) technological base layers.

## Statistics

Third, evoking the name of Thomas Bayes points toward long-standing debates in statistics<sup>6</sup> concerning the interpretation of probability itself. Are we dealing with actual properties of nature or varying degrees of certainty? These epistemological debates are certainly important, but both beyond the ambitions of this book and hardly acknowledged<sup>7</sup> when statistics is used in computer and information science. If a technique can be shown to perform well, allegiances to frequentist or Bayesian interpretations of statistics are rarely considered relevant and, in line with the previous

6 See Russell and Norvig (2010, p. 491) for a pedagogical overview and Hacking (1990) or Stigler (1999) for a historical perspective.

7 Hastie, Tibshirani, and Friedman’s influential *The Elements of Statistical Learning* (2009), for example, is completely silent on the matter.

examples of variability, one can actually develop Bayes classifiers that do not rely on Bayesian methods.<sup>8</sup>

It should be noted, however, that a Bayesian outlook that thinks in terms of degrees of certainty is ideally suited for situations where one might want to begin calculating probabilities for different hypotheses (e.g., concerning class membership) even if very few observations are available, continuously updating and refining prior probabilities as more evidence comes in. A spam filter starts building its model from the very first classified email and every additional act of labeling will produce an imprint while leaving room for future adaptation. The outlook also suggests and justifies 'greediness' when it comes to evidence: the more features, the better. Does an email contain images? How many? Does the image contain text? Does the email use HTML formatting? What kind of HTML features are used and how many times? What about the sender? The sending route? The time of day the message was sent? All of these elements – and many others – can become part of always larger feature vectors and develop their own probability relationships with the output categories.

This example also highlights another aspect that differentiates the use of statistics in information retrieval from more traditional settings. Whereas 'standard' datasets, for example, data on human individuals, generally contain potentially large populations characterized by a limited and stable number of variables or dimensions, the number of documents in a collection (or similar data) may easily be smaller than the number of properties in a feature vector based on word statistics. As a consequence, much research in computational statistics has been dedicated to techniques for dimensionality reduction. The various differences in application context and data characteristics, in combination with an engineering ethos that privileges output performance and (computational) cost over scientific purity, means that contemporary uses of statistics in the context of information ordering have spawned processes of 'disciplinarization' into academic and industry fields that have developed their own character and epistemological substance, even if they remain in conversation with traditional techniques and concepts.

## Beyond the Basics

Fourth, the technique that Maron laid out in the 1960s is generally referred to as a 'naive' Bayes classifier, because it treats features as statistically

8 Russell and Norvig (2010, p. 499) argue that the naive Bayes model need not be considered Bayesian at all. For a deeper discussion of the issue, see Hand and Yu (2001).

independent from each other. Most commentators consider this to be a 'poor' or 'unrealistic' assumption (Rish, 2001, p. 41), since features such as word frequencies taken from text documents tend to correlate. Taking correlation patterns into account could be potentially useful for improving classification performance because the fact that, for example, two words tend to co-occur could help sieve another dose of meaning from the documents under scrutiny. It is certainly not difficult to explain why Maron chose to assume independence: its simplicity made naive Bayes the obvious starting point from a conceptual perspective and well-suited to the computational capacities of available hardware. One could also argue that the larger context of information retrieval, very much dominated by coordinate indexing, provided a setting that suggested and valued the compositional capacities independence between atomized dimensions affords. Still, few commentators fail to mention the 'surprisingly good classification performance' (Zhang, 2004, p. 2) demonstrated by the fact that 'the Bayesian classifier outperforms several more sophisticated approaches on a large number of data sets' (Domingos and Pazzani, 1997, p. 105) in empirical studies.

It may seem strange that 'the reasons for the Bayesian classifier's good performance were not clearly understood' (Domingos and Pazzani, 1997, p. 105) in settings where all the used data, the full specifications, and even the source code of used algorithms are available, but there are at least two good reasons for this. On the one side, there is an increasingly well-known opacity stemming from the 'mismatch' between algorithms that process large collections of high-dimensional data inductively through many cycles of iteration and 'the demands of human-scale reasoning and styles of semantic interpretation' (Burrell, 2016, p. 2). On the other side, there are simply no means to appreciate the actual statistical relationships latent in specific data other than statistical techniques themselves, which means that there is no privileged, *a priori* knowledge of the territory the algorithm maps, other than the initial classification of training documents. This leads to the curious situation that algorithmic techniques in machine learning and other domains become objects of study in ways evoking the observation of natural phenomena in the laboratory or 'in the wild'. As Mackenzie stresses, machine learning involves a constant stream of measures that communicate the internal states of the classifier, such as levels of precision or error rates, as well as a whole array of observational devices, including 'a striking mixture of network diagrams, scatterplots, barcharts, histograms, heatmaps, boxplots, maps, contour plots, dendrograms, and 3D plots' (Mackenzie, 2017a, p. 86) that attempt to make operation observable and more clearly understandable. This also involves efforts 'to understand the data characteristics which

affect the performance of naïve Bayes' (Rish, 2001, p. 41) in order to be able to define the 'complete set of necessary and sufficient conditions for the optimality of the Bayesian classifier' (Domingos and Pazzani, 1997, p. 127). The capacity of algorithmic techniques to move between application domains, which I have emphasized a number of times, is indeed bounded: there are 'habitats' – data sets and classification characteristics – where one set of technique thrives and another withers.

And invention and adaptation of techniques continue. In fact, the seemingly obvious insufficiency of the independence assumption has led to the development of techniques that seek to model dependencies. Bayesian networks,<sup>9</sup> for example, provide a broader framework for reasoning with probabilities and powerful – but computationally demanding – means 'for augmenting the Naive Bayes classifier with limited interactions between the feature variables' (Koller and Sahami, 1997, p. 173). There are also many nonprobabilistic learning techniques, such as logistic regression or support-vector machines, that handle dependency differently. Deep neural networks' very *raison d'être* is to model complex nonlinear relationships between input variables and output classes. From here, branches lead off in various directions, many of them boasting ever more complex mathematics. The problem with many of these techniques is not just their conceptual and technical complexity, which widely available modules hide behind layers of abstraction, but their quickly increasing computational cost as the number of features grows. There are many tricks and techniques for dimensionality reduction (e.g., using weighting with tf-idf to select the most salient features) and considerable research goes into finding more efficient ways of implementing specific techniques. These investments in concretization can have enormous consequences in terms of real-world applicability. The recent rise of convolutional neural networks for deep learning, which has transformed task domains like image recognition, cannot be imagined without forms of implementation that distribute calculations over the massively parallel hardware of graphics processors (Raina, Madhavan, and Ng, 2009).

But a deeper appreciation of the substance present in many datasets can begin, much more modestly, at the level of the atoms turned into features. In the context of human language, I have already discussed dictionary- or

9 Bayesian networks are acyclical, directed graphs, where nodes are connected through probabilities. The canonical formulation comes from Pearl (1988), the text that Russell and Norvig (2010, p. 557) identify as the single most important publication of the 'resurgence of probability' in artificial intelligence research.

thesaurus-based transformations or enrichments that anchor heftier hooks in the meat of meaning. And moving from taking single terms as features to co-occurring word-pairs, often referred to as 'bi-grams', can introduce relative word location into the mix without explicit dependency modeling on the level of the classification technique. There rarely is a singular technical locus in software, and the spaces of possible variation are considerable.

## Personalization and Recommendation

Fifth, as already mentioned, Bayes classifiers are ideal for personalization, since they can easily turn user interactions into acts of labeling to construct or adapt a decision model specifically for that user. It should come as no surprise that one of the first attempts to personalize information retrieval for individuals and not just categories of 'typical' users (novice, expert, etc.) relied on a probabilistic method close to the one described above. Drawing on explicit input in the form of initial self-descriptions and on implicit feedback based on the idea that '[o]ne of the simplest ways to derive information about a user is to look at the way he uses the system' (Rich, 1983, p. 205), Elaine Rich's personalized fiction literature recommender system built a 'user model' along a number of culturally and psychologically salient facets, such as 'interests' (e.g., 'sports'), 'politics' (e.g., 'liberal'), or 'tolerate-violence' (expressed as numerical value). These facets served as the features of a user vector that was then compared to the manually attributed book vectors in the collection.<sup>10</sup> As the system begins to suggest book titles and users provide feedback on recommendations, their profiles are updated accordingly. While Rich's use of probabilistic inference spans different stages, this last step can be understood as a single-class Bayes classifier, where the relevance number (or posterior probability) for that class is used to rank the books to recommend.

Personalization, understood as the process of adapting a service to individual users, is not an algorithmic technique, but a gesture or application that may draw on very different technical procedures. One such technique, often used in recommender systems, eschews actual content features altogether<sup>11</sup> and relies on 'collaborative filtering', an approach that operates on the intuition that 'a good way to find interesting content is to find other people who have

10 Rich's system is actually a little more complicated since it uses 'stereotypes', such as 'intellectual' or 'feminist', as intermediate categories.

11 Many systems use hybrid approaches that combine both content-based and collaborative filtering. For an early example, see Balabanović and Shoham (1997).

similar interests, and then recommend titles that those similar users like' (Breese, Heckerman, and Kadie, 1998, p. 43). This idea can inform a technique that starts from user models in the form of feature vectors taking songs 'liked' or listened to as dimensions. Similarity between users is then computed exactly the same way as described in the previous chapter, using, for example, cosine similarity or the Pearson correlation coefficient. The system can then suggest new songs listened to by similar users. User profiles and content features are simply two types of entities that can be funneled into a vector space and gestures like 'personalization', 'recommendation', or 'retrieval' are merely different ways of arranging similar operations into interface outputs.

While a system based on collaborative filtering could work for spam triage to a certain degree, emails have too much variation and are too specific to individual users to work as dimensions in a vector. Here, the specific features of the machine learning approach come as an advantage. Using content features as input for a Bayes classifier and user feedback as learning mechanism allows for the creation of fully individualized decision models that take the features of an incoming item as carriers of meaning and interpret them from the vantage point of past interactions. This highlights a crucial aspect that holds for most supervised machine learning techniques: the learning algorithm orchestrates and governs the construction of a decision model, but the actual mapping between input and output is based on the encounter between data and supervision. This has far-reaching consequences, which I will address in more depth in the next section.

## Interested Readings

Bayes classifiers and other machine learning techniques arrange information ordering and decision-making in ways that are profoundly different from the common framing of algorithms as set formulas executing a stable sequence of operations. Even in academic publications, a common conception still seems to imagine a group of developers enumerating variables to take into account and specifying how to couple and weight them. The makers of Facebook, for example, would brood over the criteria for News Feed filtering, meticulously arranging metrics such as affinity between users, post engagement, and some function of time to produce a clear decision recipe that is guarded like a precious secret. Intense reflection about the application domain is certainly part of the practice, but this conception is increasingly outdated. Supervised machine learning techniques such as Bayes classifiers constitute means to derive decision models from the encounter between a purpose, data, and a



mechanism that provides the 'correct' interpretation for at least some of these data, for example, through initial labeling or continuous feedback. This echoes Shera's vision from 1952 that 'classification of knowledge will not be a fixed and unalterable pattern, [...] but a series of quite widely varying schematisms, each constructed for a specific purpose' (Shera, 1952, p. 17). What changes here is the way classification and purpose come together. In coordinate indexing, the technique serving as the backdrop to Shera's argument, purpose guides the process of creating and attributing an indexing vocabulary and informs every act of searching. In the case of Bayesian spam filtering, nobody has to manually compile a list of 'spammy' words and there is also no query. The purpose is expressed through the definition of classes, for example, 'spam' and 'not-spam' for sorting emails. When a user begins to label messages, the probability relationships between words and classes are generated automatically from the message content, producing a basis for distinction- and decision-making that is not a clear-cut formula, but an adaptive statistical model containing potentially millions of features. If the word 'Viagra' appears often in mails marked as spam and rarely in others, it will become a strong indicator or clue. But every word that has not been discarded by filtering or weighting will acquire meaning that way, and the final score for an incoming message is based on all of these 'small and individually ambiguous clues' (Spärck Jones, 1999, p. 227f.), making it hard to create a simple causal narrative.

The assessment that machine '[l]earning is more like farming, which lets nature do most of the work' (Domingos, 2012, p. 81) is clearly incomplete. Even if it is certainly not wrong to say that '[m]achine learning systems automatically learn programs from data' (Domingos, 2012, p. 81), this 'program' – or decision model – sits among a whole range of other programs, that is, things that are not 'nature' or 'the world'. Machine learning systems themselves have two essential technical components, a training engine that learns from examples and an inference engine that applies the learned model (the mapping between input features and output categories or target variables) to incoming data. But further upstream we have other interfaces, infrastructures, programs, or manual gestures like data cleaning that formalize and filter the world into a neat queue of well-behaved feature vectors and submit them to a labeling process.<sup>12</sup> Techniques like A/B/N testing, where two or more versions of the same content or application are submitted

12 While Mackenzie (2017a) dives deep into the mathematical dimension of machine learning, he remains remarkably silent on the origin of training data, on the role and orchestration of labeling, and on the embedding of both of these things into the vast digital infrastructures I have addressed with the term 'computerization'.

to random samples of users, create such feedback situations as part of an ongoing process of development and optimization. Google's Experiments architecture, part of Google Analytics, describes its purpose as providing a framework that 'enables you to test almost any change or variation to a website or app to see how it performs in optimizing for a specific goal'.<sup>13</sup> Such goals may range from product design and improvement to the more specific forms of behavior modification Zuboff (2019) describes as 'surveillance capitalism'. Built on a Bayesian statistical engine,<sup>14</sup> Google's framework permits developers unfamiliar with statistical learning to integrate empirical decision processes into their design and development methodology. But even if the practice draws on actual user behavior, the whole setup of these experimental situations is clearly not accidental or 'natural' in any sense of the word. When Domingos states that machine learning is like farming, he wants to stress its inductive character, but – inadvertently or not – positions it as an activity that is not a 'disinterestedness' pursuit. Agriculture is not biology and its goal is to grow crops, not to understand the world.

And, indeed, no deep understanding of the application domain is required to produce a model. When a payday lender uses Facebook data to decide whether an applicant is likely to pay back a loan (cf. Deville, 2013), there is no need for a Bourdieusian sociologist who tags every possible profile element as an indicator of class or socioeconomic potential. This is precisely why Andrejevic (2013) calls data mining a 'post-comprehension' (p. 41) strategy: no 'theory' of the social needs to intervene, no framework that ties the myriad data traces associated with 'the "diversification" and individualization of lifestyles and ways of life' (Beck, 1992, p. 91) back into a neat description of society. It is enough to have some users who have already labeled their own profiles by the act of paying back or defaulting on their loans to generate a model where every single profile item becomes first a feature in a vector and then an indicator for 'creditworthiness'. This is what 'generalizing from examples' (Domingos, 2012, p. 78) means and it should give us an idea how engines of order can 'automate inequality' (Eubanks, 2018) if every small act or preference can be examined in relation to its economic significance. This goes beyond the encoding of 'human prejudice, misunderstanding, and bias into the software systems that increasingly [manage] our lives' (O'Neil, 2016, p. 3) toward an ever-expanding capacity to read and assess the world in relation to a purpose (Rieder, 2016).

13 <https://developers.google.com/analytics/solutions/experiments>, accessed 11 January 2019.

14 [https://support.google.com/analytics/answer/2844870?hl=en&ref\\_topic=1745207](https://support.google.com/analytics/answer/2844870?hl=en&ref_topic=1745207), accessed 11 January 2019.

Bayes classifiers and similar techniques are neither static recipes for decision-making nor theories engaged in ontological attribution; they are methods for making data signify in relation to a particular desire to distinguish, to establish differences, and to operate on them; they are devices for the automated production of *interested readings of empirical reality*.<sup>15</sup> Maron's goal was not to say anything deep about the relationship between text and meaning but to design a system that produces 'good' results in the domain of its application. I thus use the term 'interested' to emphasize that the epistemic process is not just tainted by some unavowed bias<sup>16</sup> but fully designed around an explicit goal or purpose that trumps any epistemological or ontological qualms one may have. Just as Desrosières's (2001) notion of accounting realism suggests and Lyotard (1984) emphasizes further, 'the goal is no longer truth, but performativity – that is, the best possible input/output equation' (p. 46). This means that the definition of the desired outcome becomes the central locus of normativity.

One of the central narratives that – often implicitly – informs and justifies the empiricist organization of information ordering around behavioral feedback and clickstream data is the notion of 'revealed preference'. In 1938, the economist Paul Samuelson was voicing his dissatisfaction with recent developments around the notion of 'utility', which was central to the idea of consumer choice, he argued, but had been complicated to the point of removing 'the assumption of the measurability of utility' (p. 61). It could therefore no longer support forms of economic theory revolving mostly around mathematical modeling. In its place, Samuelson (1948) put what was to become the idea of 'revealed preference', which basically anchors the notion of utility or preference not in some elaborate psychological theory but in the observable fact that a purchase has been made: 'the individual guinea-pig, by his market behaviour, reveals his preference pattern' (p. 243). This conceptual gesture 'opened up the way for empirical studies of preferences based on observed market behaviour' (Sen, 1973, p. 242). The process of purification at work here is similar to other empirical definitions that do not really provide an actual critique of what they seek to replace but axiomatically bind the concept to a particular form of measurement rather than a theoretical construct. This is precisely what makes the notion of revealed preference so

15 Berry (2008, p. 365) talks about network models as constituting particular 'readings of reality'.

16 Although it may very well be. This is not the focus of the more conceptual approach taken in this book, but there is much research on bias and discrimination in data mining, for example, Custers et al. (2013). More explicitly on the technique discussed here, see Calders and Verwer (2010).

attractive in the context of information ordering. If behavior reveals preference and preference is the guiding principle for designing systems that espouse dynamic perspectivism to liberate 'the territory subjugated by classification' (Weinberger, 2008, p. 92), behavioral data can simply become another form of query, another expression of an information need or desire. From this perspective, everything falls into place, since optimization criteria such as time on site or click rate become expressions of user preference. What users do is what they want and what they want is what they shall receive. How could it be otherwise? Incidentally, this also justifies the quest to increase click rates and time on site, framing the process as a win-win situation: in the context of a music-streaming service, for example, a recommendation is doubly 'successful' if a user actually listens to the track, because it both satisfies the user's revealed preference and increases the use of the service. Machine learning, in particular, thrives on this circular logic. Grammatized infrastructures facilitate disassembly into intermediary forms and algorithmic techniques instantiate dynamic reassembly through learning or optimization processes designed around operational goals.

In the battle for users' time, attention, and money, such forms of optimization move to the center of business processes. Plasticity and modularity in software may imply creativity and expression, but they also enable companies driven by profit motives to constantly analyze, optimize, and experiment with orderings and their effects on the bottom line. Engaged in both representing and intervening, algorithmic techniques become both 'machineries of knowing' (Knorr Cetina, 1999, p. 5) and engines of order that create direct and indirect feedback loops between knowledge and power in ways that are hard to overstate. Facebook can tune its Newsfeed filtering engine in a way that something like 'time on site' or 'ad click probability' becomes the target variable for determining the optimal traits of content items for each user, individually. Instead of selecting and weighting variables manually, the classifier derives – or 'learns' – the relevant features from the relationship between data (posts and their different properties), feedback (users' engagement with these posts), and purpose (to increase engagement). The system is then able to execute the following command: 'show to the user the posts similar to those that previously led to high engagement'. In a recent interview (Klein, 2018), Mark Zuckerberg describes a setup that shows how easily the technical principles can be arranged to accommodate different value narratives. Responding to ambient criticism, the company apparently decided to move away from directly monetizable target variables and invited panels of users to rank actual contents in terms of what seemed most 'meaningful' to them. We can thus imagine an arrangement where

the feedback gathered this way becomes the labeling input for a classifier that is then generalized to the entire user population. 'Meaningfulness' is submitted to the vague and empiricist treatment we are already familiar with. In an earlier post, Zuckerberg had already announced that Facebook would increasingly seek to optimize for 'time well spent', which may have the effect that 'the time people spend on Facebook and some measures of engagement will go down'. But shareholders should not worry, since this 'will be good for our community and our business over the long term too' (Zuckerberg, 2018, n.p.). Similarly, YouTube argues that they are not simply optimizing for watchtime but for user 'satisfaction' (Lewis, 2018). And the different content-flagging tools on social media sites clearly serve as input for classifiers seeking to identify illegal, harmful, or other 'unwanted' content or behavior. The question, then, is how these 'target values' are being decided on, how they are being operationalized, and how they find expression in technical terms. Any working system is a practical answer to these questions and normative commitments are being made at each step of the way.

This does not mean that machine learning is not an academic discipline, but rather that most of its operational reality occurs outside of institutional settings subscribing to Mertonian norms of science. Merton argues that 'disinterestedness' is a central tenet of scientific inquiry and that 'disinterestedness has a firm basis in the public and testable character of science' (Merton, 1973, p. 276). But the work some of machine learning's most famous scholars do at Baidu (Andrew Ng), Facebook (Yann LeCun), or Google (Geoffrey Hinton, Peter Norvig) gets sucked into infrastructures of operation that are far removed from public scrutiny. This tension characterizes, at least to a degree, any discipline that produce highly applicable knowledge. And it is certainly true that algorithmic techniques do not determine the specific performativity of the resulting algorithm. The Bayes classifier provides the capacity for making interested readings and decisions but specifies neither the purpose nor the way an engine of order projects its output back into the world. Facebook can decide to train its News Feed engine based on any criterion the company could come up with. It may not be easy to find measures for concepts like diversity, plurality, or 'meaningfulness', which are heavily context-dependent and contested when pressed a little harder. But this is a question worthy of additional research and one that would, more broadly, benefit from a perspective more hesitant to locate political salience in singular instances of technicity rather than the larger systems or infrastructures technology forms. 'Implementation' of complex human values are not *problems* to be *solved* but ongoing struggles that require forms of 'cooperative responsibility' engaging multiple actors and institutions into forms of 'dynamic interaction' (Helberger et al., 2018).

Does the emphasis on optimization targets make Bayes classifiers neutral tools that could be used to further any kind of cause? The problem is more complicated than terms like ‘neutrality’ and ‘objectivity’ suggest. Mackenzie (2015) argues that ‘as machine learning is generalized, the forms of value that circulate in the form of commodities alter’ (p. 444), emphasizing that the different technical ‘styles’ of processing ‘entail different kinds of value’ (p. 436). In his most recent take on the subject, he further describes machine learning as ‘a new enunciative mode that disperses patterns as the visible form of difference into a less visible but highly operational space’, where algorithmic techniques ‘define possibilities of grouping and assembling differences’ (Mackenzie, 2017a, p. 149). To better understand what this means, one can appreciate how the spaces constituted by coordinate indexing, the relational model for database management, vectorization, and other modes of formalization into intermediate forms indeed give rise to what Deleuze described as the ‘fini-unlimited’, where a ‘finite number of components yields a practically unlimited diversity of combinations’ (Deleuze, 1988, p. 131). Any sufficiently complex dataset – and especially sets with high numbers of dimensions – can be cut, sliced, and put into sequence in many, many different ways, whether this is done through a set of search results, a statistical description, or an inductive learning process. This enables new forms of knowing that enter into complex relationships with existing practices and, in particular, with the goal-oriented processes we find in management and governance. My point, here, is that we need to take Mackenzie’s (2017a) assessment that the ‘optics’ of machine learning are always ‘partial’ (p. 80) in both senses of the word: partial as incomplete or fragmentary; and partial as skewed, biased, or – preferably – interested.

One could argue that data mining techniques embody forms of cognition or enunciation that are, on the one side, nonanthropomorphic in the sense that they consist of procedures that can only be enacted by fast computing machinery and, on the other side, thoroughly entangled in operational arrangements. To speak with Virilio (1994), we could say that they are ‘vision machines’ to which we are ‘delegating the analysis of objective reality’ (p. 59) and argue that there is ‘a “subjective” optical interpretation of observed phenomena and not just “objective” information’ (p. 75). On the level of signification, machine learning and other techniques attribute meaning to every variable in relation to a purpose; on the level of performativity, the move to increasingly integrated digital infrastructures means that every classificatory decision can be pushed back into the world instantly, showing a specific ad, hiding a specific post, refusing a loan to a specific applicant, setting the price of a product to a specific level, and so forth. No data point

remains innocent. Standing in for the wider category of information ordering techniques, Bayes classifiers entangle meaning – and not just the meaning of texts – in complex and often very direct ways with decision-making informed by specific objectives and purposes. In a sense, Bacon's famous distinction between 'is' and 'ought' disappears into a form of description that is built, from the ground up, on a prescriptive horizon. We no longer decide based on what we know; we know based on the decision we want to make.

What makes algorithmic techniques far from 'neutral', then, is not some intrinsic bias or allegiance to a specific political model but the fact that these new 'levers on "reality"' (Goody, 1977, p. 109) are distributed throughout social domains in specific ways, which has the potential to profoundly reconfigure power in contemporary societies. When looking at power as 'techniques' or 'functionings', as a 'network of relations, always in tension' (Foucault, 1995, p. 26), we can appreciate how algorithmic information ordering delivers specific ways to continuously establish, organize, and modulate the relationships between datafied entities in service of strategic goals. And if the 'exercise of power consists in "conducting conduct" and in adjusting the probabilities of possible outcomes' (Foucault, 1982, p. 789, translation amended), machine learning indeed provides concrete technical means to do so, informing the varieties of 'automated management' Kitchen and Dodge discuss in *Code/Space* (2011). Although Bayes classifiers do not determine how their results are used at the interface level and beyond, they stand for a new set of techniques that have the capacity to generate interested forms of knowledge used to make myriad of small and large decisions with concrete effects. They introduce a 'micro-physics'<sup>17</sup> that may well affect how power operates in significant ways. And the term 'operate' needs to be taken seriously, here. The computerized infrastructures human activities are increasingly entangled with imply forms of semiosis that take effect not like signs, but also, with loose reference to Lessig (1999), like *walls*, encroaching on conduct through modifications of the visible and navigable environment.

## Information Retrieval as Trading Zone

This chapter discussed the Bayes classifier both as an example for the broad category of machine learning techniques and as a means to demonstrate how

17 Foucault (1995, p. 26) introduces the term to address the many subtle, diffuse, and productive technologies of power that operate on bodies in various ways.

the intermediate forms presented in previous chapters open onto spaces of technical expression and plasticity that introduce different epistemological and operational gestures into the arsenal of software-makers and, by extension, into the professional and institutional settings they are embedded in. While the field of machine learning comprises a whole range of historical trajectories, Maron's experiments in information retrieval again show that text documents are ideal test cases for techniques that, on the one side, deal with the complexity, ambiguity, and uncertainty of human endeavors and, on the other, set out to produce operationally viable results rather than scientific models of language or deep emulations of human thinking. The use of statistical techniques implies forms of cognition and learning that are fundamentally caught up in the utility equations of specific task domains.

The Bayes classifier demonstrates how, in Maron's (1963) own terms, 'methods, approaches, clues, tricks, heuristics' (p. 141) can inform intricate epistemological and operational engines functioning on top of basic computation. We again observe how algorithmic techniques enable complex and compound behavior based on the computer's capacity to perform very large numbers of simple operations. Maron's collection of 405 scientific abstracts, tiny by today's standards, already contained thousands of individual words and not even Roberto Busa's Jesuit colleagues would have been easily convinced to perform the mindless counting and calculating work necessary to classify even a single document.

The entanglement with operational concerns and the willingness to combine approaches from different fields make it problematic to consider these techniques simply as 'applied statistics' and we should not overestimate their relationship with disciplinary developments in statistics.<sup>18</sup> While I do not want to generalize too far from the small group of individuals I have mentioned by name over the last chapters, it bears mentioning that the educational background and thematic interest of early information retrieval researchers was not very close to the fields and preoccupations 'traditional' statistics were concerned with, such as the study of biological patterns of inheritance or 'the measurement of people and populations' (Beer, 2016, p. 9). Luhn was an engineer, Maron a physicist turned philosopher of science, and Salton one of the earliest 'genuine' computer scientists. Spärck Jones came to information retrieval through her PhD research in linguistics. They all had the education and capacity to use statistical techniques with ease but were

18 'Despite the obvious connections between data mining and statistical data analysis, most of the methodologies used in Data Mining have so far originated in fields other than Statistics' (Friedman, 1998, p. 3).



hardly interested in disciplinary debates around the nature of probability or the difference between its Bayesian and frequentist interpretations. And unlike Codd, who constructed the relational model on the basis of a strong mathematical formalism, even Salton was more concerned with *in situ* experimentation and testing than theoretical elaboration.

The point I want to make is that statistical concepts and techniques enter information retrieval as technical elements, detached from their origins and ready-to-hand for integration into techniques that draw on a variety of trajectories and, crucially, enter into ongoing dialog with the computational machinery of the day. The information retrieval setting and its pragmatic, output- and performance-focused attitude encouraged and justified experimentation, supporting an epistemological outlook that, at least to one commentator, yielded only 'relatively disappointing progress [...] to develop as a coherent and firmly based empirical discipline' (Ellis, 1998, p. 225) and favored approaches that 'lack a full paradigmatic identity' (Ellis, 1998, p. 239). Salton himself lamented the 'absence of basic theories, and the largely experimental nature of the information science field' (Salton, 1973, p. 220). This should not minimize achievements but rather emphasize that information retrieval and, by extension, the larger space of contemporary information ordering is best understood as a 'trading zone' (Galison, 1996), where statistics and other areas of mathematics intermingle with ideas about language, information, and knowledge as well as computing machinery, systems design, and the concrete and imaginary requirements of 'knowledge workers' and 'decision-makers'. Certainly, similar to what Galison finds in his analysis of Monte Carlo simulations, the pidgin languages we encounter in the earliest texts solidify into 'a full-fledged Creole: the language of a self-supporting subculture with enough structure and interest to support a research life without being an annex of another discipline' (Galison, 1996, p. 153). This stabilization is achieved through metanarratives concerning information needs and information overload, through concepts and metrics like recall and precision, through formal models such as the VSM, and through accepted knowledge rituals coming in the form of comparative experiments and competitions. Yet there is no common paradigm, no firm ontology, and not even a shared diagnosis what the 'library problem' really consists of. But again with Galison (1996), the different epistemological cultures begin to develop a 'common activity centered around the computer' (p. 153), despite serious internal doubts (cf. Taube, 1965) and despite the general reluctance of librarians, who felt like victims of colonization rather than partners (cf. Wasserman, 1965, p. 10) and retain an ambiguous relationship with the information retrieval trading zone to this day. The way the

computer serves as a center should not be understood primarily through notions like quantification, formalization, or unification around a central paradigmatic reference like Shannon's information theory or logic-based computationalism. These elements are far from irrelevant, but, as the history of information retrieval shows, the computer's capacity to connect and to compel stems primarily from its capacity to *function*.

As mentioned earlier, Agre (1997b) build his critique of artificial intelligence on the assessment that 'AI people, by and large, insist that nothing is understood until it has been made into a working computer system' (p. 12) and argues that this constrains the space of ideas and styles of reasoning that are admitted into the discipline, hampering its creative and intellectual potential. Seen from another side, however, we can speculate that the normative force of a running system also *relaxes*, at least to a point, the demand for actual understanding. The Bayes classifier becomes a viable technique because it can be shown to work and, indeed, to work well. Knowing *why* a technique works well is not a fundamental requirement, even if the question prompts intense follow-up research that has the potential to push the discipline forward in terms of genuine understanding. The testing of classification performance on preclassified corpora is therefore not an entertaining sideshow but lies at the center of information retrieval and adjacent fields. For Spärck Jones, this has a profound effect on their epistemological makeup:

[T]he role of experiments in AI is to try out designs for engineering artefacts, to see how well some system will meet some need. [...] This implies a performance measure related to the system's purpose, which may be more or less easy to find. But it is a measure of acceptability not of truth. (Spärck Jones, 1990, p. 281)

Experiments and competitions establish the working system and at least a partial appreciation of its capacities in the public arena, but they also frame the terms of what constitutes 'acceptability'. We can see a similar logic play out more recently around neural networks. The famous ImageNet corpus and the annual ImageNet Large Scale Visual Recognition Challenge, to name a concrete example, take part in defining what 'visual recognition' actually means, what constitutes 'success', 'failure', and so forth. The epistemological logic that drives the focus on the running system and its purpose-driven performance indeed constitutes an 'attitude to reality' (Desrosières, 2001, p. 339) that prepares and facilitates the transfer of techniques into the operational domains of business and government.

Machine learning, taken together with the techniques discussed in the previous two chapters, demonstrates how a variety of algorithmic techniques can build on the same intermediate forms and these are not simply different approaches to doing the same thing: the canonical query-response situation we know from coordinate indexing does not imply the same epistemological and operational setup as document clustering with the help of unsupervised machine learning, as item recommendation via collaborative filtering, or as training a classifier through examples and feedback. The difference between an explicit request for information and the implicit observation of use patterns that drives much of personalization captures only part of these variations. In the following chapter, I will initially leave the domain of information retrieval only to come back with new ‘material’ to integrate into algorithmic information ordering. Looking at the long and illustrious prehistory of the PageRank algorithm, which I take as a stand-in for the larger field of network analysis, I want to widen the perspective and show how contemporary information ordering draws on domains beyond document processing, encountering different modes of valuation and building on intermediate forms other than feature vectors.

## Bibliography

- Agre, P. E. (1997b). Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In G. C. Bowker, S. L. Star, W. Turner, and L. Gasser (eds.), *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide* (pp. 131-158). New York: Psychology Press.
- Andrejevic, M. (2013). *Infoglut: How Too Much Information Is Changing the Way We Think and Know*. New York: Routledge.
- Balabanović, M., and Shoham, Y. (1997). Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM* 40(3), 66-72.
- Beck, U. (1992). *Risk Society: Towards a New Modernity* (M. Ritter, trans.). London: Sage.
- Beer, D. (2016). How Should We Do the History of Big Data? *Big Data & Society* 3(1), 1-10.
- Berry, D. M. (2008). The Poverty of Networks. *Theory, Culture and Society* 25(7-8), 364-372.
- Borko, H. (1962). The Construction of an Empirically Based Mathematically Derived Classification System. In G. A. Barnard (ed.), *AIEE-IRE '62 (Spring) Proceedings of the May 1-3, 1962, Spring Joint Computer Conference* (pp. 279-289). New York: ACM.

- Borko, H., and Bernick, M. (1963). Automatic Document Classification. *Journal of the ACM* 10(2), 151-162.
- Breese, J. S., Heckerman, D., and Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In G. F. Cooper and S. Moral (eds.), *UAI '98 Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (pp. 43-52). San Francisco: Morgan Kaufmann.
- Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1-12.
- Calders, T., and Verwer, S. (2010). Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Mining and Knowledge Discovery* 21(2), 277-292.
- Custers, B., Calders, T., Schermer, B., and Zarsky, T. (eds.). (2013). *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*. Berlin: Springer.
- Deleuze, G. (1988). *Foucault* (S. Hand, trans.). Minneapolis: University of Minnesota Press.
- Desrosières, A. (2001). How Real Are Statistics? Four Possible Attitudes. *Social Research* 68(2), 339-355.
- Deville, J. (2013). Leaky Data: How Wonga Makes Lending Decisions. *Charisma Consumer Market Studies*, 20 May. Retrieved from <http://www.charisma-network.net>.
- Domingos, P. (2012). A Few Useful Things to Know about Machine Learning. *Communications of the ACM* 55(10), 78-87.
- Domingos, P., and Pazzani, M. J. (1997). On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning* 29(2-3), 103-130.
- Ellis, D. (1998). Paradigms and Research Traditions in Information Retrieval Research. *Information Services Use* 18(4), 225-241.
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.
- Foucault, M. (1982). The Subject and Power. *Critical Inquiry* 8(4), 777-795.
- Foucault, M. (1995). *Discipline and Punish: The Birth of the Prison* (A. Sheridan, trans.). New York: Vintage Books.
- Friedman, J. H. (1998). Data Mining and Statistics: What's the Connection? *Computing Science and Statistics* 29(1), 3-9.
- Galison, P. (1996). Computer Simulations and the Trading Zone. In P. Galison and D. J. Stump (eds.), *The Disunity of Science: Boundaries, Contexts, and Power* (pp. 118-157). Stanford: Stanford University Press.
- Goody, J. (1977). *The Domestication of the Savage Mind*. Cambridge: Cambridge University Press.
- Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.
- Hand, D. J., and Yu, K. (2001). Idiot's Bayes: Not So Stupid after All? *International Statistical Review* 69(3), 385-398.

- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2<sup>nd</sup> ed.). New York: Springer.
- Helberger, N., Pierson, J., and Poell, T. (2018). Governing Online Platforms: From Contested to Cooperative Responsibility. *The Information Society* 34(1), 1-14.
- Kitchin, R., and Dodge, M. (2011). *Code/Space: Software and Everyday Life*. Cambridge, MA: MIT Press.
- Klein, E. (2018). Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next. *Vox*, 2 April. Retrieved from <https://www.vox.com>.
- Knorr Cetina, K. (1999). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard: Harvard University Press.
- Koller, D., and Sahami, M. (1997). Hierarchically Classifying Documents Using Very Few Words. In D. H. Fisher (ed.), *ICML-97 Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 170-178). San Francisco: Morgan Kaufmann.
- Lessig, L. (1999). *Code and Other Laws of Cyberspace*. New York: Basic Books.
- Lewis, P. (2018). 'Fiction Is Outperforming Reality': How YouTube's Algorithm Distorts Truth. *The Guardian*, 2 February. Retrieved from <https://www.theguardian.com>.
- Luhn, H. P. (1968a). Auto-Encoding of Documents for Information Retrieval Systems. In C. K. Schultz (ed.), *H. P. Luhn: Pioneer of Information Science: Selected Works* (pp. 126-131). New York: Spartan Books.
- Lyotard, J.-F. (1984). *The Postmodern Condition: A Report on Knowledge* (G. Bennington and B. Massumi, trans.). Manchester: Manchester University Press.
- Mackenzie, A. (2015). The Production of Prediction: What Does Machine Learning Want? *European Journal of Cultural Studies* 18(4-5), 429-445.
- Mackenzie, A. (2017a). *Machine Learners: Archaeology of a Data Practice*. Cambridge, MA: MIT Press.
- Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. *Journal of the ACM* 8(3) 404-417.
- Maron, M. E. (1963). A Logician's View of Language-Data Processing. In P. L. Garvin (ed.), *Natural Language and the Computer* (pp. 128-151). New York: McGraw-Hill.
- Maron, M. E. (2008). An Historical Note on the Origins of Probabilistic Indexing. *Information Processing & Management* 44(2), 971-972.
- Maron, M. E., and Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM* 7(3), 216-244.
- Merton, R. K. (1968). The Matthew Effect in Science. *Science* 159(3810), 56-63.
- Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.
- Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-Scale Deep Unsupervised Learning Using Graphics Processors. In A. Danyluk, L. Bottou, and M. Littman (eds.), *ICML '09 Proceedings of the 26<sup>th</sup> Annual International Conference on Machine Learning* (pp. 873-880). New York: ACM.
- Rich, E. (1983). Users Are Individuals: Individualizing User Models. *International Journal of Man-Machine Studies* 18(3), 199-214.
- Rieder, B. (2016). Big Data and the Paradox of Diversity. *Digital Culture & Society* 2(2): 1-16.
- Rish, I. (2001). An Empirical Study of the Naive Bayes Classifier. *IJCAI Workshop on Empirical Methods in Artificial Intelligence* 3(22), 41-46.
- Russell, S., and Norvig, P. (2010). *Artificial Intelligence* (3<sup>rd</sup> ed.). Upper Saddle River: Pearson Education.
- Salton, G. (1973). On the Development of Information Science. *Journal of the American Society for Information Science* 24(3), 218-220.
- Salton, G., Wong, A., and Yang, C.-S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11), 613-620.
- Samuelson, P. A. (1938). A Note on the Pure Theory of Consumer's Behaviour. *Economica* 5(17), 61-71.
- Samuelson, P. A. (1948). Consumption Theory in Terms of Revealed Preference. *Economica* 15(60), 243-253.
- Schäfer, M. T. (2011). *Bastard Culture! How User Participation Transforms Cultural Production*. Amsterdam: Amsterdam University Press.
- Sen, A. (1973). Behaviour and the Concept of Preference. *Economica* 40(159), 241-259.
- Shera, J. H. (1952). Special Librarianship and Documentation. *Library Trends* 1(2), 189-199.
- Silberman, S. (2000). The Quest for Meaning. *Wired*, 1 February. Retrieved from <https://www.wired.com>.
- Solomonoff, R. J. (1957). An Inductive Inference Machine. *IRE Convention Record, Section on Information Theory* 2, 56-62.
- Spärck Jones, K. (1990). What Sort of a Thing Is an AI Experiment? In D. Partridge and Y. Wilks (eds.), *The Foundations of Artificial Intelligence* (pp. 276-281). Cambridge: Cambridge University Press.
- Spärck Jones, K. (1999). Information Retrieval and Artificial Intelligence. *Artificial Intelligence* 114, 257-281.
- Star, S. L., and Ruhleder, K. (1996). Steps toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7(1), 111-134.
- Stiegler, B. (2010). Memory. In W. J. T. Mitchell and M. Hansen (eds.), *Critical Terms for Media Studies* (pp. 64-87). Chicago: University of Chicago Press.

- Stigler, S. M. (1983). Who Discovered Bayes's Theorem? *American Statistician* 37(4), 290-296.
- Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press.
- Taube, M. (1965). A Note on the Pseudo-Mathematics of Relevance. *American Documentation* 16(2), 69-72.
- Thompson, P. (2008). Looking Back: On Relevance, Probabilistic Indexing and Information Retrieval. *Information Processing and Management* 44, 963-970.
- Virilio, P. (1994). *The Vision Machine* (J. Rose, trans.). Bloomington: Indiana University Press.
- Wasserman, P. (1965). *The Librarian and the Machine*. Detroit: Gale Research.
- Weinberger, D. (2008). *Everything Is Miscellaneous*. New York: Henry Holt.
- Zhang, H. (2004). The Optimality of Naive Bayes. In V. Barr and Z. Markov (eds.), *Proceedings of the Seventeenth International FLAIRS Conference* (pp. 1-6). Menlo Park: AAAI Press.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.
- Zuckerberg, M. (2018). No title [Facebook message], 12 January. Retrieved from <https://www.facebook.com/zuck/posts/10104413015393571>.