CA Journal of Cultural Analytics

October 07, 2020

Divergence and the Complexity of Difference in Text and Culture

Kent K. Chang and Simon DeDeo

Kent K. Chang. Social and Decision Sciences, Carnegie Mellon University; School of Information, University of California, Berkeley. kentkchang@berkeley.edu.

Simon DeDeo. Social and Decision Sciences, Carnegie Mellon University & the Santa Fe Institute. sdedeo@andrew.cmu.edu; http://santafe.edu/~simon

Peer-Reviewer: Ian Milligan

Data Repository: github.com/kentchang/ca_divergence

ABSTRACT

Measuring how much two documents differ is a basic task in the quantitative analysis of text. Because difference is a complex, interpretive concept, researchers often operationalize difference as distance, a mathematical function that represents documents through a metaphor of physical space. Yet the constraints of that metaphor mean that distance can only capture some of the ways that documents can relate to each other. We show how a more general concept, divergence, can help solve this problem, alerting us to new ways in which documents can relate to each other. In contrast to distance, divergence can capture enclosure relationships, where two documents differ because the patterns found in one are a partial subset of those in the other, and the emergence of shortcuts, where two documents can be brought closer through mediation by a third. We provide an example of this difference measure, Kullback–Leibler Divergence, and apply it to two worked examples: the presentation of scientific arguments in Charles Darwin's *Origin of Species* (1859) and the rhetorical structure of philosophical texts by Aristotle, David Hume, and Immanuel Kant. These examples illuminate the complex relationship between time and what we refer to as an archive's "enclosure architecture", and show how divergence can be used in the quantitative analysis of historical, literary, and cultural texts to reveal cognitive structures invisible to spatial metaphors.

Those who study culture look for differences. Just as ethnographers might study the differences between the practices of regions, villages, or families,[1] literary scholars might study the differences between genres, modes, or periods.[2] Those who approach culture from a quantitative standpoint are no exception, and we are often tasked with the goal of measuring the differences between different, computationally identified, patterns of expression.[3] In the digital analysis of texts, for example, we might ask how much two documents or sets of documents differ, and relate these differences to other aspects of psychological or social life.[4]

Journal of Cultural Analytics 4.11 (2020): 1-36. doi: 10.22148/001c.17585

Difference, however, is a rich concept, and while we may feel competent to discuss and debate the differences between this and that, we struggle to express what this means in the abstract:[5] what, after all, is difference? The disagreements between people as to what differences are, and what differences can imply, suggests that no formal definition can capture its complete meaning. Indeed, whether we talk about difference from the first-person experience of the relationship between texts, or the judgments of an observer that takes a third-party view, difference may always remain, in part, a matter of judgment. A scholar's sense of the divergent meanings of difference, however, is often lost in its translation to quantitative tasks. The goal of this paper is to reveal that complexity again.

Difference as Distance

To talk about difference in a quantitative project requires that we operationalize the concept: we seek algorithms and quantifications that can capture some of what we mean by the richer, scholarly idea.[6] What counts as capture, however, varies.

The matter becomes harder to settle when cultural analytics reaches beyond syntax, a familiar object of study in linguistics, in an attempt to operationalize semantic concepts. A recent example of comes from Dennis Yi Tenen, who used a combination of lexical and syntactic markers to operationalize the idea of "clutter" (i.e., the density of objects in a scene) in fiction.[7] Here the spatial density of an imagined space maps to a quantifiable density on the written line.

When we operationalize difference in the study of texts, we often use the metaphor of *distance*.[8] This is not surprising because distance is a potent metaphor for difference. It is natural, for example, for critics in both the academic world and popular press to refer to one author as "closer" to another when they mean to say their differences are not so large. Informally, for example, Virginia Woolf's modernist *Mrs. Dalloway* seems "closer" to James Joyce's *Ulysses* than either is to Charles Dickens's *Great Expectations*. The spatial metaphor then leads scholars to represent texts as vectors so that they can measure some form of mathematical distance between them (dot product, cosine distance, and so on).

The actual computation involved can be quite complex, although to grasp it conceptually, consider fig. 1; if, for some very curious reason, we decide to represent these three novels as points on the coordinate system, it will then be possible for us to measure the Pythagorean distance between them. Distance between Woolf and Joyce is shorter than that between Woolf and Dickens, a measurement we hope captures a richer claim that Woolf is more different from Dickens than from Joyce.

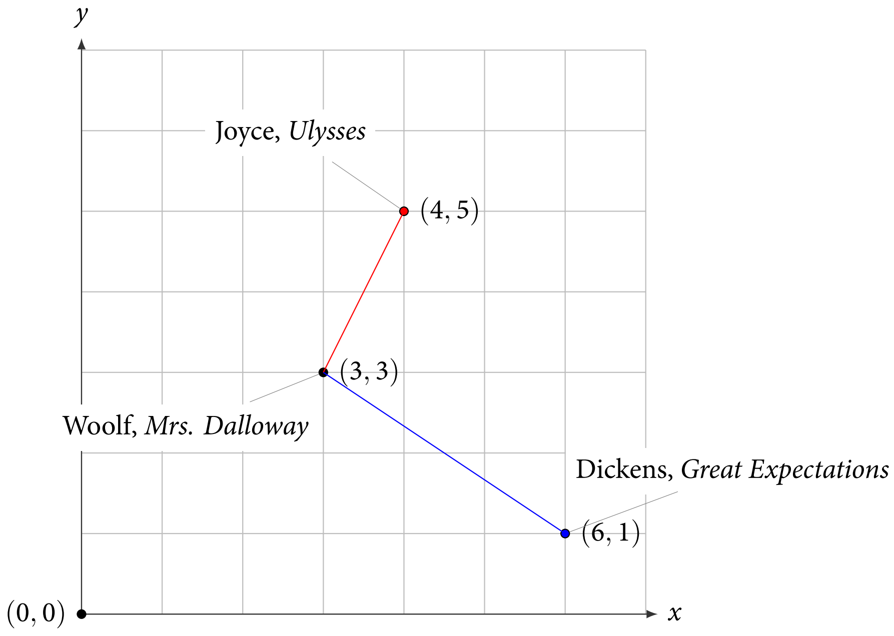

Figure 1: If we want to operationalize difference as distance, we can map literary texts into something that can be represented in a physical space. For purpose of illustration, we place three points on a two-dimensional space, each representing a novel, and with their locations summarizing some aspects of interest. Having done this, we can then use basic geometric formulae, such as the Pythagorean Theorem, to compute the distance between them, and use that to build an account of the differences between the texts.

Distance Is Not the Only Fruit

Yet while distance is often used as a proxy to shed light on the difference between texts, it is not at all a neutral construct. As we will see in this section, a measure of distance between two texts, or objects, necessarily obscures their full relationship.

Indeed, it is tempting to think, as distance requires us to do, about language and literature in spatial terms, as if those novels were literally hanging in a space that we, or they, can move in. This metaphor, although powerful, has its limitations. Many of these limitations are invisible to us, because the immediate nature of our experience of the three-dimensional spaces between the objects we encounter blinds us to the fact that space is actually a highly structured thing that constrains the ways in which objects can relate to each other.[9] The mathematical structure underlying any particular definition of distance—called a "metric"—inherits many of those constraints. In particular, distance is subject to four axioms of increasing severity:

Axiom 1. The distance of an object from itself is zero.

Axiom 2. Distances are always positive: a house may be ten miles North, or ten miles South, but the distance traveled can never be "negative ten miles."

Axiom 3. Distances are symmetric: the distance from A to B is equal to the distance from B to A.

Axiom 4. No shortcuts: to go from A to C via B can never be shorter than to go directly from A to C. (If B lies along the way, of course, the distances may be equal.) Distances satisfy the triangle inequality.

These axioms capture many of our informal experiences of space. Distances defined "as the crow flies," for example, obey them, as do great circle distances on the globe. So does the taxicab metric, familiar to residents of Manhattan, where paths are confined to a grid and diagonal shortcuts are forbidden. Distances on social networks, defined as the length of the shortest path from one person to another following symmetric relationships, are a metric.

Another distance is encountered when we classify things into nested groups in the way, for example, Linneaus did with biological species grouped in genuses, and genuses grouped in families, and so on. This induces a definition of distance: the distance between two objects is the number of steps up the hierarchy one has to go in order to get to a category sufficiently general that it includes both. The carrion crow is in the same genus as the fish crow (so they have distance one), but not the same genus as the blue jay; both the carrion crow and blue jay, however, are in the same family (*Corvidae*), and so the distance between carrion crow and blue jay is two. This definition satisfies the necessary axioms, and in fact makes it what is known as an "ultrametric".

The solipsist's metric is a final example, where the distance between any two points is unity (i.e., 1), except (by axiom) the distance from you to yourself, which is zero. The solipsist's metric may have an amusing name, but it is not very useful. A measure may obey the axioms, but that alone does not make it fit for purpose.

Despite this diversity of choices, researchers have tended to use one of a small number of possible distance metrics.[10] A common method is to pretend that feature counts are like vectors in ordinary Euclidean space: each word corresponds to an axis, and the position of a text along that axis is equal to the number of times the word appears in that text. Having done this, one can then use cosine distance (the angle between two text-vectors), say, or vector distance (the Pythagorean distance from one text to another, as in fig. 1) as the measure. Sometimes, document features are given as probability distributions. This happens when doing topic modeling, for example: after a researcher runs a topic model on a corpus, each document is described by a vector of weights, or probabilities, for each topic. Some distance metrics are explicitly designed to work with probabilities; these include the Earth Mover's Distance and the Total Variation Distance,[11] the Fisher Information metric, and Jensen–Shannon Distance.[12]

For readers, critics, and historians, however, the constraints of the axioms mean that distance is necessarily blind to the full range of ways that documents might differ from each other. Consider the third axiom, symmetry. While this is a natural demand for distances between points on a plane, our reading experience is usually highly asymmetric. To encounter one text before another can be a decisive fact about the differences between them: the first chapter of a monograph, for example, may not move very far from where the preface left us, but to read the chapter first, and then to see the broader concerns of the preface, is a different—and potentially confusing—experience because the preface touches on the themes of the first chapter but also includes what comes after. Distance fails to capture this asymmetric relationship between part and whole, or the order in which an argument unfolds.[13]

Similar concerns apply when we move from a series of texts published over many months or years to the making of an argument or the unfolding over time of a conversation, debate, or epistolary exchange. Some conversations may narrow down on a subset of the themes introduced at the start, but in a different kind of conversation the speakers may choose to extend their concerns, broadening the discussion from the initial question (say, where to locate a bus stop) to include topics in other domains (budget constraints, environmentalism, racial and economic inequality). Some of these differences can be captured by a measure of how broad, or diverse, a conversation's topics are as it evolves over time, but such a measure now misses the difference between a conversation that maintains attention to the original theme while introducing others, and one that simply goes off track or wanders randomly.

This asymmetry of epistemic experience is widespread in how we package and consume culture. In classrooms, for example, students rarely experience literary works in a "symmetric" fashion: in period surveys, instructors want earlier segments to serve as preparation for later ones, with the assumption that (for example) formal techniques accumulate and are activated at different points in time as they evolve. It is less common to have a survey where students study, say, John Milton's *Paradise Lost* before Edmund Spenser's *The Faerie Queene*, not only because one was produced later than the other, but because of an assumption that the later one may inherit certain characteristics of the former in ways that the former can not anticipate.

The "no shortcuts" axiom (Axiom 4) also comes into conflict with our intuitions about how readers may experience the texts of a culture. The relationships between two documents may be better understood, for example, if one encounters an overview, creole, or field guide. Certain key texts may bridge a gap between distinct traditions, leading a reader to believe that apparent differences are perhaps not as great as they appear. Cultural development within a social context often relies upon these unexpected shortcuts: a gradual shift, over many years, from one style to another provokes less surprise, and thus perhaps less resistance, than leaping in one fateful moment from start to finish.

Pedagogy is often built around clever violations of Axiom 4 through the use of bridge texts. A student who encounters Henrik Ibsen's *A Doll's House* after reading only Greek Tragedy will perceive a greater difference between the two than one who has experienced an intermediate stage of the cultural process—say, the tragedies of Shakespeare. By going from Sophocles to Shakespeare to Ibsen, the reader is drawn to a particular vision of the relationships between the start and finish, or (one might equally well say), she experiences less difference than if she skipped the intervening step.

Difference beyond Distance

No operationalization is perfect, and any quantification of a set of texts must neglect nuances of the parent concept. Yet the price to pay of the difference–

as–distance metaphor can, as we have seen, be high, blinding us to the very features of our data that we may consider most interesting.

Such steep penalties may not be always be necessary. This is because distance is not the only measure of difference, and mathematics allows for more liberal measures that enforce weaker constraints. These measures are like distances in that they take two objects (probability distributions, say) and assign a number representing the "non-alikeness" in return. They differ from distances, however, in that they drop the third and fourth axioms of the previous section in favor of a weaker (if more abstract) set that permits asymmetry and shortcuts.

These measures are known as *divergences* and can provide a new metaphor for difference. Divergences can be understood as arranging documents in a cognitive realm, where they are encountered and learned by a mind.[14] Like distance, there are many forms of divergence; they are sometimes known as Bregman Divergences, and include Stein's Loss, Itakura–Saito's Divergence, and Kullback–Leibler Divergence (KL), which will be the focal example of divergence in this essay.

Kullback–Leibler Divergence violates axioms three and four, often severely. To use it to quantify difference means a rejection of the distance metaphor, and a consequent relaxation of distance's constraints. It replaces the metaphor of space by a cognitive one where texts are "learned" by imagined learners, and learners experience differing levels of success in fitting the patterns of what they have learned to something new. This metaphor can capture asymmetric differences and differences that allow for shortcuts, violating both axioms three and four. This makes divergence a more adequate measure for epistemic experiences, such as those described in the previous section.

In cognitive science, Kullback–Leibler Divergence is called "cognitive surprise,"[15] a term not totally foreign to literary theory. Wolfgang Iser, for instance, emphasized the importance of surprise in the process of reading: when one reads, for Iser, one wants to be surprised, and the fact that literature rarely meets the expectation of a reader is what make it distinctive from other kinds of discourses.[16] Iser's argument is reminiscent of what Russian Formalists, most notably Victor Scklovsky, call defamiliarization,[17] as well as of Ezra Pound's exuberant slogan of "Make it new." Iser, Scklovsky, and Pound make different uses of the notion of surprise; the goal of our operationalization here is not to prescribe one over another, but to make it possible to treat this core feature in a quantitative fashion.

In going beyond the spatial metaphor, KL also comes to challenge our intuitions of what mathematics can do with texts. It comes as part of an apparatus of concepts of information theory, which enables us to quantify the underlying cultural and cognitive practices associated with each of these stages. Claude Shannon's 1948 paper, the founding document of information theory, names the basic quantity *uncertainty*.[18]

Uncertainty is epistemic, and information theory matches a simple, almost childlike story about how one gathers information from others.[19] The gathering of information corresponds to the learning of patterns that enable productive communication. Under this epistemic account, for example, the formation of schools and genres might be seen as the result of social learning in response to repeated exposure: in this case, participants learn the surface patterns of their school's or genre's conventions. Readers, in their turn, might be seen as building pattern recognition capacities that may or may or may not align with the patterns of any particular genre.

In actual use, of course, metaphors of learning and pattern recognition are simply that—metaphors. The mere presence of a "signal" in the Shannon sense does not imply the existence of "signal" information that a participant in the system itself might have access to, might have the cognitive capacities to make use of, or, indeed, might care to use at all. Both information theorists and literary scholars have urged their communities not to elide the difference between these two notions of signal.[20] We take the position that whether or not an information theoretic quantity deserves its epistemic label requires a further interpretation and justification. Recent work along these lines includes studies of the information density of nineteenth-century novels, and the novelty and transience of political speeches during the French revolution.[21]

Uses of Divergences

Fig. 2 summarizes our discussion up to this point. We began by distinguishing difference and distance, and pointed out the inherent limitation of distance as embodied experience. Below we shall see divergence in action. It will turn out that divergence, and the underlying epistemic notions it tracks, violates the axioms of distance—"symmetry" and "no shortcut" in particular—in crucial and useful ways.

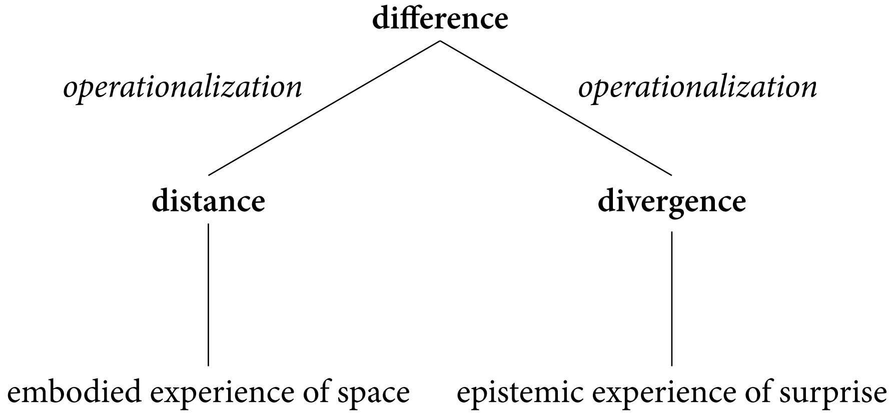

Figure 2: Two paths to operationalizing difference: a cognitive approach made possible by the information-theoretic notion of divergence provides a different account of how things differ than the more familiar spatial metaphor of distance.

Asymmetry and Enclosure

Given an interpretive frame, we can often consider one text as being broader, or ranging over more themes, than another. One can then speak of the asymmetrical relationship of one text being contained, or enclosed, by another. Consider, for example, the case of diction. At any particular time, a culture will have a range of registers available for a text to use.[22] One text may have an elevated diction, another low, and a third might mix the two. A recent study by Stefania Degaetano-Ortlieb and Elke Teich shows how scientific English became broader, or gradually came to *enclose*, general English over the second half of the nineteenth century.[23] There are certainly words that are common in general English that are less common in scientific English (for example, the word "wife"). However, there are (roughly speaking) far more words in scientific English that are rarely, if ever, used in general English (such as "oxide"). This imbalance is measured by the Kullback–Leibler asymmetry.

This notion of enclosure has its analogies in the spatial case: the boundaries of a neighborhood usually fall within the boundaries of a city, for example, and one can say that New York City contains, or encloses, Brooklyn. A similar, if less familiar, notion applies to sets of things: the set of plants contains, or encloses, the set of trees. An encyclopedia of ancient history may include the history of Greece. A two-year course in social science will include an introduction to economics.

When these relationships are strict, we have little difficulty talking in terms of one thing being inside another: volume one is contained in a three-volume set. In many cases, however, these enclosure relationships can become fuzzy. An overview of economics might include an account of Keynes' macroeconomic theory, but an advanced (as opposed to introductory) macroeconomic text may cover a wider range of Keynes' theories that the overview neglects. Conversely, the introductory overview might contain a chapter on behavioral economics, which the advanced text, focusing on Keyenes, omits. We do not want to say, then, that an introductory economics course contains the content of an advanced seminar, but at the same time we do have the intuition that the former ranges further than the latter, or provides an overview, and thus, to a certain extent, includes it. For lack of a better term, we will describe this fuzzier intuition as (partial) "enclosure."[24]

The asymmetry of Kullback–Leibler provides one way to operationalize this enclosure relationship, by reference to the cognitive process of surprise. Specifically, we will say text one "encloses" text two just in the case that a reader who encounters (trains on) text one first is less surprised on encountering text two than the reverse.

Enclosure in Tolstoy and Darwin

To drive this intuition, consider two major words of Russian literature, *Anna Karenina* and *War and Peace*, by Leo Tolstoy. In a very general sense (and the only one needed to understand this example), both of these novels deal with themes of friendship, suffering, war, and love, but in different proportions; while *Anna Karenina* is most famously centered around a love story, *War and Peace* follows a variety of characters as they navigate a broader array of life events.

We can represent, in this toy example, the two novels as probability distributions over these themes. Explicitly, the distributions p_A (*Anna Karenina*) and p_B (*War and Peace*) quantify the themes by reference to the relative proportions of words associated with the core concepts in order, as in table 1:[25]

$$\begin{aligned} p_A &= \{0.14, 0.21, 0.08, 0.57\}, \\ p_B &= \{0.15, 0.15, 0.37, 0.36\}. \end{aligned}$$

Intuitively, at this level of analysis, *War and Peace* covers a broader range of themes than *Anna Karenina*; the latter, for example, distributes only 8% of its coverage to war, compared to 37% in *War and Peace*.

The Kullback–Leibler Divergence reflects these differences in the asymmetry between the two directions. The KL from *Anna Karenina* to *War and Peace* (i.e., "the surprise on reading *Anna Karenina* and then encountering *War and Peace*, when one searches for these themes"), is 0.49 bits, while the KL from *War and Peace* to *Anna Karenina* is 0.35 bits. The enclosure relationship is captured in the asymmetry, or difference, in surprise.[26] In this interpretative framework, readers who begin with *Anna Karenina* have little exposure to

theme	*Anna Karenina*	*War and Peace*
friendship	14%	15%
suffering	21%	12%
war	8%	37%
love	57%	36%

Table 1: While Anna Karenina *is dominated by words associated with love,* War and Peace *has a more uniform distribution over the four themes considered here. In this sense,* War and Peace *"encloses" the themes in* Anna Karenina*, and while the Kullback–Leibler from* Anna Karenina *to* War and Peace *is 0.49 bits, the Kullback–Leibler in the reverse direction, from* War and Peace *to* Anna Karenina*, is only 0.35 bits. In this example, the compositions of the two novels are determined by counting words associated with the four themes in the English-language translations of Constance Garnett and computing the relative odds. The listed probabilities are then conditional probabilities: given that one has encountered one of these themes, what are the probabilities that it was this one or that.*

narratives involving war; when they read *War and Peace*, they encounter new patterns associated with the war theme more than one third of the time. Conversely, readers who begin with *War and Peace* have significant exposure to all four themes; they see less representation of some of them than they would in *Anna Karenina*, but not enough to overcome the near-total absence of war in the latter. The underlying asymmetry of 0.074 bits between the texts, in the direction from *War and Peace* to *Anna Karenina*, captures this relationship.

Some aspects of this asymmetry can be explained by comparing the entropy of the two distributions. *War and Peace* can be seen as broader than text two simply because the entropy is higher (1.83 bits vs. 1.62 bits). However, a difference in entropies can be misleading. Consider a new novel, *C*, characterized by a distribution p_C equal to $\{0.20, 0.25, 0.25, 0.10\}$. This text has higher entropy than *War and Peace*, but the KL asymmetry in this case favors *War and Peace* as the enclosing text. The KL on reading *C* and encountering *War and Peace* is 0.014 bits higher than the other direction. Differences in overlap matter, because *C* provides insufficient coverage of the love theme to prepare the reader for *War and Peace*, while the reverse is not true.

Enclosure can have social, as well as purely epistemic, implications. Enclosure effects, for example, can leave traces within the representation of racial and ethnic groups in popular media, although this time in the opposite direction. If we were to approximate the speech of white characters in Hollywood movies produced over the past half-century by the speech of non-white characters in those same movies, we would see a higher level of surprise when we move from the language of the dominant group to the subordinate group than the other way around.[27] In other words, the speech of white actors does not enclose that of non-white actors but leaves out distinctive semantic registers that act as signals of the non-whiteness; conversely, whiteness is "unmarked." These results mirrors the effects discussed above wherein minority groups need to learn two languages, that of the dominant group and that of a particular subculture.[28]

The claim that one text encloses another is relative to a particular interpretive context. If, for example, we had restricted our Tolstoy analysis of the two texts to themes associated with marital infidelity, then we would expect *Anna Karenina* to enclose *War and Peace*; the former text provides multiple psychological and social perspectives on the practice, while in *War and Peace* the examples are fewer and less fleshed out; Hélène's infidelity to Pierre, for example, receives far less coverage than Anna's to Karenin.

It is not the case that the additional coverage of love in *Anna Karenina* is simply a more laborious version of what appears in *War and Peace*, in other words. *Anna Karenina* may present a richer, more detailed portrayal of (a particular kind) of love, and there are many aspects of the experience that *War and Peace* neglects in favor of other themes. The latter encloses the former only when we consider the particular range of table 1.

This provides an instructive contrast to the Degaetano-Ortlieb and Teich's results that show that (in our terminology) science "encloses" general speech. Informally, one might have imagined that science is concerned with only a small fraction of what people might talk about (it contains "just science"), and thus that it ought to be enclosed by general speech. That it is not tells us something surprising about how science works (at least in the 19th Century, which this enclosure result obtains): it means that, rather than being a particular investigation of a subset of human experience, science is—at least at the lexical level—something more like an encompassing account. It is characterized by both the use of ordinary terms, and the introduction of new vocabulary that goes well beyond what is found in that comparison corpus.

Our Tolstoy example is simpler in part because we focus on themes, rather than individual words, which eliminates some of the ways that *Anna Karenina* may in fact "enclose" *War and Peace*. All questions of enclosure depend, finally, on the validity and interest of the ways in which one quantifies the qualities of interest, and an analysis always has a characteristic level of resolution and space of patterns. If we had, for example, considered words associated with the romantic affections, and compared the two texts on the basis of their distributions over that set of words, we might well find the opposite result; informally, it is an open question whether or not *Anna Karenina*, as a novel of the psychological effects of romantic love, encloses *War and Peace*.

Our example is concerned with a simple pairwise comparison, and a binary outcome: is A enclosed by B, or is B enclosed by A? Getting an answer to this question may be enough, if the texts are sufficiently important, but we are often concerned with the overall *enclosure architecture* of a collection. At its most complete, this architecture is defined by (1) the complete list of pairwise relationships between all pairs of texts in the collection, i.e., for each pair, which text encloses the other, and (2) the magnitude of the underlying asymmetries, i.e., how complete each enclosure is. Mathematically, one has a network of relationships, a directed graph, where each node is a text, and every pair of texts is connected by an arrow that indicates the direction of the enclosure, and the strength.

There is much to be learned from these networks. Consider the only slightly more complicated case of a three text collection. If one disregards labels, two enclosure architectures are possible. This first is where text A (for instance) is enclosed by text B, and both are enclosed by text C; one might call this the "Russian doll" architecture, with a strict hierarchy of enclosure relationships. The second is one where text A encloses text B, and text B encloses text C, but text C, in turn, encloses text A; one might call this a “cyclic,” or “ouroboros” architecture where the global structure violates our standard intuitions of how containment works in ordinary space. The ouroboros is rare, but not impossible, as can be verified by computer simulation: if texts are described by a three-dimensional probability distribution, and those distributions are chosen randomly, an ouroboros emerges in about four percent of the trials.

These enclosure relationships also interact with time, or, indeed, any natural ordering that might exist between texts. Consider, for example, the Russian doll architecture, where A, B, and C are successive chapters in a book. One might describe the text as a “synthetic” enclosing of the texts as they accumulate: B includes A but also other things, C includes B but also other things. If the chapters are ordered in the opposite fashion, one might describe it as a “analytic” disassembly of the doll: inside A is B, and inside B is C. Synthetic and analytic are two extremes, and another possibility is that the texts are ordered A, C, B. In this case, the progress involves both assembly and disassembly of the doll over time.

We know very little about how time and enclosure interact in literary and cultural production.[29] A more detailed study of the relationship between time and enclosure architecture is complicated by the fact that the number of possible enclosure architectures increases exponentially as the number of texts in the collection grows. For four texts, there are four possible structures, for five texts, there are twelve, for six texts there are fifty six, and so on.[30] One natural approach to this increasing taxonomic complexity is to turn to the tools of network theory, and to measure overall statistical properties of the network such as degree centrality, path distance, and so forth, adapted to the new constraint that the edges between nodes now have not only a magnitude, but a direction.

A different approach is to prune the enclosure network to isolate key features and reveal what one might call its “backbone.” We show this approach now, with a second, more sophisticated, study of how asymmetry in Kullback–

Leibler Divergence reveals semantic relationships in the development of the argument of Charles Darwin's *On the Origin of Species*.[31] In this case, asymmetry will allow us to distinguish between the analytic and synthetic argument patterns: an analytic one, where the argument as a whole is presented, and then individual details are worked out separately, and a synthetic one, in which later chapters bring together separate units that have been previously introduced.

The Tolstoy example used simple synonym sets for our analysis; for Darwin, we will use query sampling to model the introduction and chapters of this work as a distribution over topics; explicitly, the distribution of words in each chapter is described as a weighted combination of co-occurring word patterns, or "topics,"[32] found in this case using the topicexplorer code.[33]

Given this model, we can construct the backbone of the enclosure architecture by linking chapters together on the basis of maximal containment asymmetry. For each chapter i in the *Origin*, we find the chapter j such that "learning the patterns of j, and then encountering i" is much easier than "learning the patterns of i, and then encountering j."[34] Each chapter is thus paired with the chapter that maximally encloses it; fig. 3 shows all these pairs, where the maximally enclosing chapter for i is shown by an arrow going from i to its encloser, and the thickness of each arrow corresponds to the strength of the asymmetry. The backbone is a simplification of the full enclosure architecture; instead of considering all 105 relationships among the fourteen chapters and introduction, we look at only the most significant relationship for each in turn.

The resulting graph shows that Darwin's argument proceeds synthetically, or cumulatively, instead of analytically. Rather than introducing the book's argument, and breaking it down into separate pieces discussed in isolation from each other, later chapters can be seen to both include and extend what comes before. This process culminates in chapters 9, 10, and 11; the final three chapters that follow those are contained within them, providing a short "analytical" coda to what has been assembled synthetically.

To see how this happens in detail, consider, for example, chapter 10, “On the Geological Succession of Organic Beings.” This chapter is a maximal encloser for many of the early chapters, as well as the introduction (node “I”). Had Darwin followed an analytical plan, with chapter 10 assigned to a close examination of one of the several themes that had gone before, we would expect the enclosure relationship to be reversed. In fact, however, chapter 10 plays a synthetic role: Darwin applies his theory, as previously developed, to explain the paleontological record. Chapter 11, “Geographical Distribution,” is a second strong encloser. As in the case of chapter 10, it encloses by virtue of a synthesis directed towards an explanation of empirical data, but in this case, rather than consider the effect over time, it considers differences induced by space. Chapter 9, which weakly encloses 10, is a third enclosing point; like chapter 10, it deals with the challenges to the previously-presented theory from the paleontological record.

These results reveals what can be gained by exchanging a spatial metaphor for a cognitive one, i.e., by taking the right, rather than the left, branch in the tree of fig. 2. Under the spatial metaphor (provided by, for example, Jensen–Shannon Distance), chapter 1 is closer to chapter 10 than it is to chapter 11; the axioms of the distance relationship prevent us from further analyzing how these parts of the argument relate to each other. Fig. 3 reveals the nature of that relationship; one of structured enclosure instead of, for example, repetition or analysis. This backbone analysis necessarily reduces the full complexity of Darwin’s argument, and the relationships between the chapters. Many other insightful simplifications are possible and, of course, one can always attempt to draw the entire network and, through intuition and hard work, perceive patterns that the backbone reduction hides.

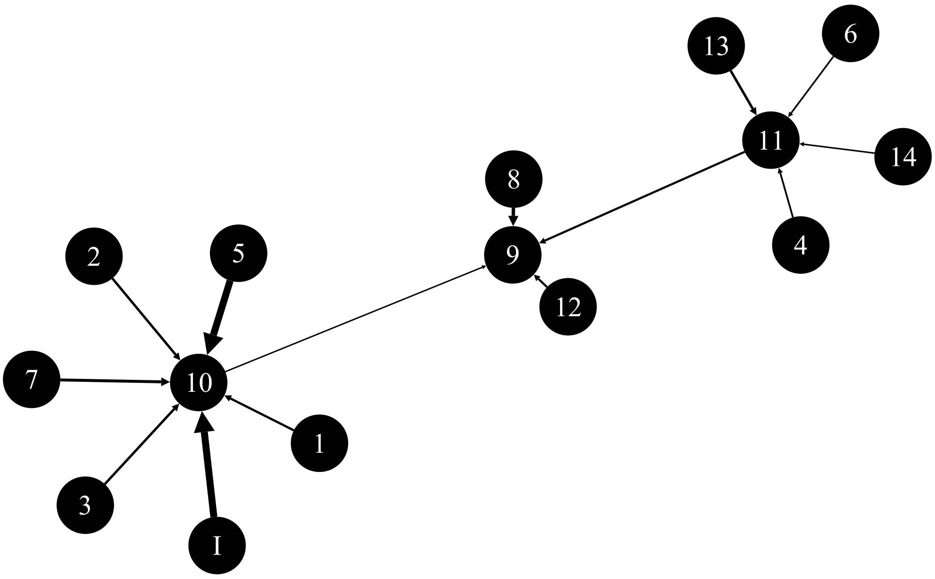

Figure 3: The backbone of the enclosure architecture of Origin of the Species*, where key Kullback–Leibler Divergence relationships reveal the dynamical structure of the developing argument. Links in this graph show the maximal enclosure relationships between chapters in the* Origin*. An arrow from one node to another indicates that the latter encloses the former, meaning that the KL Divergence from the former to the latter is greater than the KL Divergence from the latter to the former; the thicker the line, the greater the magnitude of the imbalance. The diagram reveals how chapters 9, 10, and 11 serve as convergence points where multiple lines of argument integrate.*

Rhetoric and Shortcuts in Philosophical Texts

Our Darwin analysis shows how the asymmetric nature of divergence reveals enclosure relationships hidden from distance measures alone. A second example shows how looking for violations of the no shortcuts axiom can reveal the dynamical structure of a philosophical argument. This is because Kullback–Leibler Divergence violates not just the symmetry axiom, but the triangle inequality. Explicitly, it can be "quicker" to travel from A to B if one takes a diversion via C.[35]

We take three philosophical texts (Aristotle's *Metaphysics*, David Hume's *Treatise on Human Nature*, and Immanuel Kant's *Critique of Pure Reason*), and chunk them into (approximately) paragraph-length "chunks" of one hundred words each, after stopword removal.[36] We then topic model the chunks of each text (separately). For each sequence of three chunks, we compute the shortcut, S, provided by the middle paragraph.[37]

Because of the triangle inequality, under a spatial metaphor of difference, $S(i)$ is always negative: it is always at least as easy to go directly from chunk i to chunk $i + 2$ as to go via an intermediate paragraph. Intuitively, however, we think that arguments have a potentially pedagogical structure that places ideas in a sequence that makes them easier to grasp: intermediate steps can provide a bridge between beginning and end.

To look for this effect requires a cognitive metaphor, provided in this case by by divergence, that takes into account the idea of learning. Fig. 4 uses Kullback–Leibler Divergence to show that such bridge paragraphs compose about 15% of our sample. Philosophical texts, in other words, can contain shortcuts: it is sometimes easier to read three things than just two.

The shortcut pattern is not the only one found in these texts; indeed, the modal triple is negative, suggesting the presence of sequences that are neither a shortcut nor entirely linear. Most intermediate paragraphs deviate, to some extent, from both the patterns of the one before and the one that comes after.

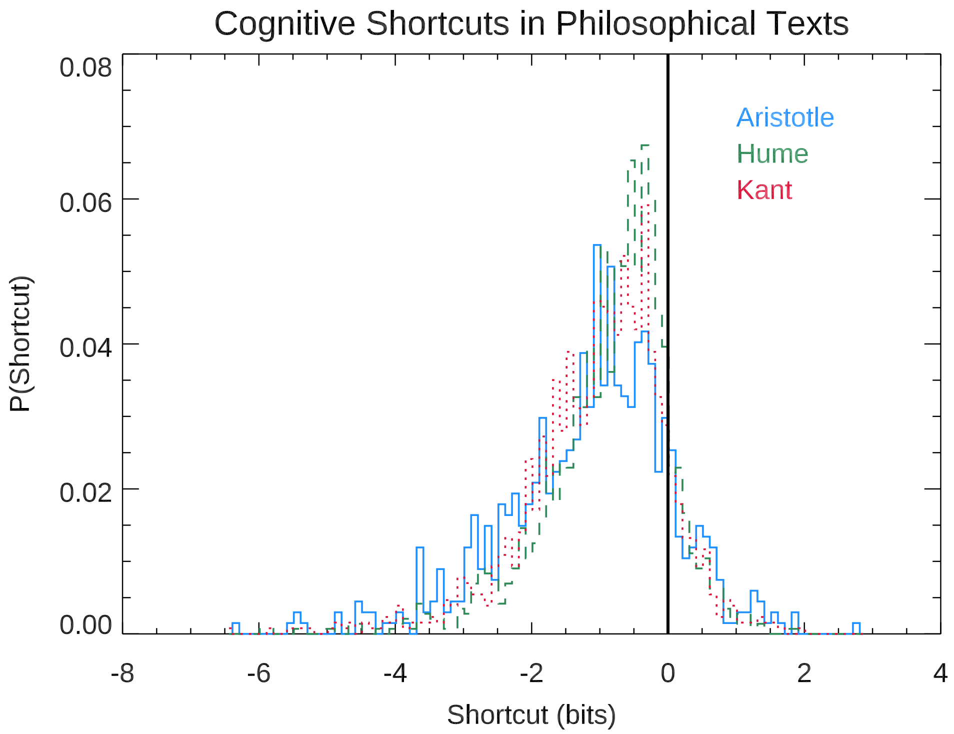

Figure 4: Shortcuts in philosophical texts (here, Aristotle's Metaphysics*, David Hume's* Treatise on Human Nature*, and Immanuel Kant's* Critique of Pure Reason*). First, we take each text and consider it as a sequence of (approximately paragraph-sized) chunks. We then look at each three-chunk sequence in turn. Interestingly, the Kullback–Leibler Divergence from the first chunk in such a sequence to the third is often longer than the KL from the first to the second, added to the KL from the second to the third. Informally, one can take a shortcut by going the "long way around." When we look at all of these consecutive three-chunk sequences in each of the three authors, about 15% of them show evidence for this effect (of two units being brought closer together than they would otherwise by the mediation of a third).*

Measures that look at paragraph-to-paragraph asymmetries (i.e., the enclosure relationships described in the previous section) can provide a second axis for analysis, and reveal the role of "intermediate" steps.

In particular, an intermediate step in an argument can, sometimes, provide a bridge that narrowing the gap between the endpoints. It can also, however, serve as a place to expand upon what has come before: metaphorically, the author dwells on a concept before coming back on track. If the sequence is A, B, C, then this dwelling might be an expansion (where B encloses A), or a focusing (where A encloses B). This is another example of the complexity of the interaction between time and enclosure architecture introduced above.

Conclusion

Divergence allows us to capture aspects of experience that are invisible to the metaphors of space. Because they build from an information-theoretic metaphor of pattern-matching, they can capture the psychology of surprise, containment, and learning. Distance is, by contrast, blind to the ways in which patterns of one text or period may be more or less capacious than another, or the ways in which they can provide bridges between distant practices.

The same psychological concepts that divergence measures also play a role in qualitative scholarship. When they do, they gain a thicker, richer content than what can be captured by a mathematical prescription based on the syntax of how objects tend, or tend not, to combine. What a historian means by surprise, or a critic means by late style, for example, is much more nuanced and specific than set of questions in pattern-learning.

This does not obviate the possibility that information theory can enrich these questions and provide new answers in turn. Under divergence, for example, an event stands out in the extent to which it breaks a previous pattern. In qualitative work, influential accounts of what makes for an "event" for the historical record include the idea that an event corresponds to a rupture of patterns.[38]

The use of divergence could then, for example, help both locate new points of rupture and determine on what level of discourse these ruptures occur.

The mathematics of divergence is, like distance, neither orthogonal nor equivalent to the corresponding human concept. It is thus a natural concept for the scholar of culture who seeks to augment an account of the world built on reading alone.

Acknowledgements

We thank Colin Allen for a close reading of an early draft of this work, and extensive and helpful comments from our two anonymous reviewers. We acknowledge the support of the John Templeton Foundation, and Jaan Tallinn through the Survival and Flourishing Fund.

Appendix

Code

Code to estimate the information theoretic quantities in this paper, and accompanying data necessary to reproduce the figures and results, is available at `https://github.com/kentchang/ca_divergence`.

Fundamental Concepts in Information Theory

Information theory provides more than just a conceptual framework for understanding the relationship between text, reading, and interpretation. It also enables us to quantify the underlying cultural and cognitive practices associated with each of these stages.

In this optional appendix, we introduce some key concepts in information theory via an epistemic categorization task. Consider, for example, the organization of end-rhymes in a poetic form. We know that these organizations matter to a poetic culture, that they produce distinct psychological effects, and may be thought more or less suited to express different sensibilities or levels of refinement. Within a poetic culture as a whole, the diversity of rhyme schemes—i.e., the extent to which that culture relies upon one form over another—becomes a pattern itself that one learns. At different points in time, the cognitive and cultural constraints result in this or that balance.

Imagine a scholar studying a English poetic culture from the first decade of the seventeenth century. She has developed a sense for the diversity of forms in play in each year; say, for example, that she learns that of the sonnets circulated in that year, roughly half are Shakespearean in form, a quarter Petrarchan, and the rest of Spenserian, mixed, or indeterminate form. On encountering a new poem in the archive, in the information theoretic framework, we visualize her categorization process as a branching tree of yes/no questions (fig. 5).

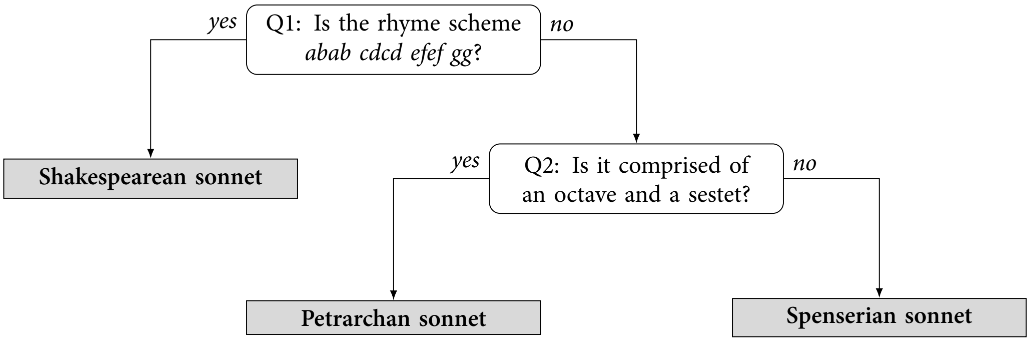
Q1: Is the rhyme scheme *abab cdcd efef gg*?
yes
no
Shakespearean sonnet
Q2: Is it comprised of an octave and a sestet?
yes
no
Petrarchan sonnet
Spenserian sonnet

Figure 5: A thought process (or script*) of figuring out whether a poem is a sonnet, represented by a binary tree of yes/no questions. For the sake of example, we assume that there are only three major types of sonnet: Shakespearean, Petrarchan, and Spenserian. We are interested in how many questions in average our scholar needs to make the identification.*

Information theory then asks how many questions—on average, and assuming that the new document comes from the same cultural practice as what she has learned so far—it takes to make the identification. Different methods provide different levels of efficiency. One could, for example, check first for Shakespearean patterns by looking for a quatrain structure—amounting to beginning with the question "is it Shakespearean?" The number of investigative methods is effectively infinite: there are at least as many ways to approach an archive as there are archives. Yet some are more efficient than others; checking first for a sestina structure, and then alliteration and caesura, is an inefficient way to characterize the new.

Preliminaries: Script and Performance

Script and efficiency. We can quantify efficiency by the number of yes/no questions needed for a positive identification, where the sequence follows a predetermined *script*. If her script checks first for a Shakespearean form and it is indeed so, then the catergorization is finished in one. Some scripts are better than others and if the goal is efficiency, Is it comprised of an octave and a sestet? is a better question than, Does the poem contain a reference to Aphrodite? An answer to the latter question will be at best a weak signal of a form associated with erotic poetry. Knowledge of the underlying culture also provides an important guide; in this case, a bias towards Shakespearean form would influence the kinds of questions one asks.

Scripts are used more than once, and for each script, and each underlying culture, we can determine the probability that the script terminates after a certain number of steps. That is to say, for each question (labeled x), you have this:

$$\begin{aligned}\text{efficiency of question } x = &\ \text{number of questions asked so far} \\ &\times \text{how likely this question leads directly to the answer}\end{aligned}$$

This can be understood as a measure of the distinguishing power of question x; and when you sum up the power of the individual questions, you get the average number of questions needed for the script.

More formally, we can label each question x_i, where i is the index of the current question; $L(x_i)$ symbolizes the number of steps required to get to the answer; what the scholar already knows about the archive is the probability associated with $L(x_i)$, symbolized as $P(x_i)$. In information theory, such "efficiency" is called "script performance," which in our case, using standard mathematical notation, looks like this:

$$\text{Script Performance}(x_i) = \sum_{i=1}^{N} L(x_i)P(x_i) \qquad (1)$$

where N is the total number of steps (read this as "the sum, over i, from one to N, of the quantity $L(x_i)$ times $P(x_i)$.")

question #	the answer	$L(x_i)$	$P(x_i)$
1	Shakesperean sonnet	1	0.5 (50%)
2	Petrarchan sonnet	2	0.25 (25%)
2	Shakesperean sonnet	2	0.25 (25%)

Table 2: Quantities in the thought process. Here $L(x_i)$ is number of steps required, $P(x_i)$ probability associated with x_i

.

i	$L(x_i)$	$P(x_i)$	contribution to eq. 1 sum
1	1	0.5	0.5
2	2	0.25	0.5
3	2	0.25	0.5

Table 3: Plugging the numbers from table 2 into eq. 1.

We can summarize the steps of a particular categorization process in table 2. First, plug the numbers into eq. 1, which gives us table 2, and then sum to get the results in table 3.

Epistemic Quantities

Our discussion has assumed a certain degree of omniscience; the scholar knows the typical characteristics of the relevant forms, and has a good intuition for what she expects in an arbitrary sample from that world. Given that, the script represented in fig. 5 is not only something she can implement, but something that, when implemented, is *optimal.* You can convince yourself that other branching structures—say, one that strives to rule out the Petrachan form first—are less efficient.

Uncertainty. Claude Shannon's 1948 paper showed that the average length of such an optimal question script can be computed without explicitly writing down the script itself; the measure takes a representation of the underlying culture (X), and returns the "uncertainty," $H(X)$,

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i) \tag{2}$$

where X is the probability distribution of x_i, namely $P(x_i), P(x_{i+1}), \ldots, P(x_N)$. Let's plug in the numbers again; the result is below, where each term in the sum for $H(X)$ is denoted as $h(x_i)$:[39]

$P(x_i)$	$h(x_i)$
0.5	0.5
0.25	0.5
0.25	0.5

As can be seen, this quantity $H(X)$ gives us the same result as in the previous section: a script length of, on average, one and a half questions.[40]

Uncertainty is an epistemic quantity: it measures how well you can gather information given your knowledge of the underlying system given prior knowledge of the system. The inputs, such as $P(x_1) = 0.5$, derive from the scholar's intuition and experience.[41]

The logarithm function appears in Shannon's paper because of purely mathematical constraints on how to combine properties of branching trees. It turns out, however, to match basic results in cognitive science, where perceptual strength is often found to scale logarithmically, not linearly. This leads to familiar biases in human behavior, because a logarithm compresses scales. Say, for example, I am a graduate student who makes $20,000 a year, and I compare myself first to a postdoc who makes $40,000, and then to a banker who makes $160,000. Under a linear, or additive comparison, I need to add $20,000 to match the postdoc, and $140,000 to match the banker, and the banker is five steps further from me than the postdoc.

Humans, however lean to logarithmic comparisons. If the postdoc makes twice what I do, and the banker makes eight times what I do, the banker may seem just "three postdocs" away from me (three doublings). Under this perceptual scheme, the banker is then just two units further away from me than the postdoc is. Inequality looks weaker.[42]

Returning to the basic formalism, we see that the uncertainty measure $H(X)$ implies a script, or knowledge-gathering and pattern-recognition process, that matches the facts of the matter. Our scholar really does know the poetic conventions of the time she has under study, and this is what enables her to construct the optimal script to which $H(X)$ refers. What happens, however, when there is a disjunction between the two? This can happen not only because of the scholar's failure to know her period, but also in the case that she encounters a different period, or different set of conventions, altogether.

Imagine, for example, that a box of files from an earlier period has been accidentally misdated and placed on the her desk. In this earlier period, the dominance of Shakespeare over Petrarch has been reversed (table 4). The scholar's script is now no longer efficient: rather than checking for a property that rules on Shakespearean form, she would sort the papers more efficiently if she checked for Petrarchan form first. A short calculation shows that her original script takes, on average, a quarter of a question longer to complete; if it takes her an hour to sort a box from the expected period with the optimal script, it will take her an additional ten minutes if she encounters a box from the early period.

Just as uncertainty can be quantified, so can this inefficiency. If $q(x)$ is the distribution for the original boxes, and $p(x)$ the distribution for the unexpected one, the inefficiency is measured by the Kullback–Leibler Divergence. This can be calculated in the following way. If the distributions p and q are over N possibilities, one takes the logarithm of the ratio of p and q, for each of the N options, and adds them together, weighted by the value of p. Informally, KL is the "the average difference in surprise between p and q, in a world governed by p." KL is high when there are options that are often encountered under distribution p, that are surprising if one is operating under distribution q. In mathematical form, this is written

$$\mathrm{KL}(p|q) = \sum_{i=1}^{N} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}, \tag{3}$$

read "the Kullback–Leibler of q given p." The average difference in surprise tells you how many additional questions you need to ask compared to using the optimal script when you have trained on q, and encounter p.

Type of sonnet	Original	Early Period
Shakespearean	50%	25%
Petrarchan	25%	50%
Spenerian	25%	25%
Average # of questions for Original Script	1.5 bits	1.75 bits
Average # of questions for Early-Period Script	1.75 bits	1.5 bits

Table 4: Performance of the first script on the original period and the earlier one. The script is optimal for the first, but not the second. For that second case, the script takes one and three-quarters of a question on average, while it is possible to design a better script that takes only one and a half.

Kullback–Leibler Divergence speaks to the epistemic experience of the unexpected. In the cognitive science literature, it has been called "Bayesian surprise."[43] Surprise need not, itself, be unexpected, and the scholar's experience of KL may be due to something other than a mixing up of archives. She can also use it as a measure of change, for example: "I know that poetic traditions changed over time—but when, and how much?"[44]

Enclosure, Resonance, and KL

The interplay between broadness and overlap can be seen by expanding the formula for enclosure, the difference in KL. Consider two texts, text one defined by the distribution p, and text two, defined by the distribution q. The extent to which p encloses q, i.e., the extent to which the KL from q to p is larger than from p to q, is

$$\mathrm{KL}(q|p) - \mathrm{KL}(p|q) = (H(p) - H(q)) - \sum_{i=1}^{N} \left(p_i \log \frac{1}{q_i} - q_i \log \frac{1}{p_i} \right) \tag{4}$$

The first term on the right is the difference in entropy: the extent to which the first text is broader than the second. The second term can be thought of as a training failure: if $p_i \log \frac{1}{q_i}$ is large, then a reader that trains on q, and therefore spends little time on topic i (because q_i is low, and therefore $\log \frac{1}{q_i}$ is high) finds himself at a disadvantage, because p_i is high (i.e., that unusual topic is common in p). The two parts of this second term are sometimes called the cross-entropy, a key component in both machine learning and models of pattern encoding in cognitive science.[45]

Cross-entropy and entropy compete. When either of these terms is high enough, i.e., p is either broader than q (first term) or p provides a better training for q than vice versa (second term), our operationalization says p encloses q. Intuitively, a text that tries to cover a little bit of everything, in however skewed a fashion, can potentially enclose a text that is impartial between a subset but neglects everything else. A good teacher gives you a broad education, but one at least somewhat tailored to the conditions of life that you might expect to encounter.

Expanding out the enclosure measure in eq. 4 allows us to compare it to the measure of *resonance*, introduced by Barron et al. Enclosure relates two texts, p and q. Resonance, by contrast, considers how two texts are mediated by a third.

Resonance is most simply understood when the three texts appear one after the other in time. First, we define *novelty* ($\mathcal{N}$), the KL divergence from a text in the past (p), to a text in the present (r), as

$$\mathcal{N} = \mathrm{KL}(r|p) = \sum_{i=1}^{N} r_i \log \frac{r_i}{p_i}. \tag{5}$$

Informally, novelty is “how surprised are you, having trained on the text from the past, when you encounter the current text”; if the present text is radically different, novelty will be high, justifying the name. We then define *transience* ($\mathcal{T}$) as the KL divergence from a text in the future (q), to the same text in the present (r):

$$\mathcal{T} = \mathrm{KL}(r|q) = \sum_{i=1}^{N} r_i \log \frac{r_i}{q_i}. \tag{6}$$

Informally, transience is “how surprised are you, having trained on the text from the future, if you go back and look at the present text”; if the present text is radically different, one can say that the future text looks different from the present text, and that the present text's patterns have not been replicated forward, justifying the name "transience."

Resonance ($\mathcal{R}$) is then defined as $\mathcal{N} - \mathcal{T}$. When resonance is positive, a text's novelty is larger than its transience. A little algebra shows that

$$\mathcal{R} = \sum_{i=1}^{N} r_i \left(\log \frac{1}{p_i} - \log \frac{1}{q_i} \right). \tag{7}$$

Comparing this equation with the second term in eq. 4 shows how resonance captures, in some fashion, the way in which the enclosure of the future by the past is mediated by the present.

Notes

[1] Margaret Mead, *Sex and Temperament in Three Primitive Societies*, First Perennial edition (1935; New York: Harper Perennial, 2001).

[2] See e.g., Genette Gérard, *The Architext: An Introduction*, Quantum Books, Originally published 1978 (Berkeley: University of California Press, 1992); Northrop Frye, *Anatomy of Criticism: Four Essays.*, Collected Works of Northrop Frye V. 22, Edited by Denham, Robert D. (1979; Toronto Ont.: University of Toronto Press, 2006); and Jacques Barzun, *Classic, Romantic, and Modern.* (Boston: Little, Brown, 1961).

[3] See Hoyt Long and Richard Jean So, "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," *Critical Inquiry* 42, no. 2 (2006): 235–67, doi:10.1086/684353.

[4] See Michael Gavin, "Is There a Text in My Data? (Part 1): On Counting Words," *Journal of Cultural Analytics*, January 25, 2020, doi:10.22148/001c.11830; Ted Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: The University of Chicago Press, 2019), 34–67; Andrew Piper and Eva Portelance, "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading," *Post-45*, 2016.

[5] Ferdinand de Saussure, *Course in General Linguistics*, Bloomsbury Revelations, Translated by Roy Harris (Bloomsbury Publishing, 2013); Jacques Derrida, *Writing and Difference*, Translated by Alan Bass (University of Chicago Press, 1978).

[6] Operationalization is fundamental to social-scientific research; once we have a specific concept that we want to study (differences between two texts, in our case), we need corresponding indicators that allow us to measure it, so that we can score, and subsequently classify, cases considered. We call this process of developing indicators and

thereby making the concept of interest measurable *operationalization*. For more background, see e.g., Robert Adcock and David Collier, "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research," *The American Political Science Review* 95, no. 3 (2001): 529–46; and Dong Nguyen et al., "How we do things with words: Analyzing text as social and cultural data," *Frontiers in Artificial Intelligence* 3 (2020): 62.

[7] Dennis Yi Tenen, "Toward a Computational Archaeology of Fictional Space," *New Literary History* 49, no. 1 (2018): 119–47.

[8] See Ted Underwood, "The Historical Significance of Textual Distances," arXiv preprint, 2018, arXiv: 1807.00181. To avoid confusion, we might add that *distance* in humanities scholarship can refer to ideas not covered in this essay: for instance, it can be the critical distance between readers and texts, as seen in e.g., Lorraine Kasprisin, "The Concept of Distance: A Conceptual Problem in the Study of Literature," *Journal of Aesthetic Education* 18, no. 3 (1984): 55, doi:10.2307/3332675; or, a temporal one between historical events and the time in which historians represent them, as in e.g., Mark Phillips, *On Historical Distance* (Yale University Press, 2015).

[9] C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation*, Revised edition of 1978 text with preface by David Kaiser (Princeton University Press, 2017).

[10] A recent counterexample is Li et al.'s use of non-Euclidean, "hyperbolic" metrics that violate our intuitions of Euclidean spaces governed by the Pythagorean Theorem; see Linzhuo Li, Lingfei Wu, and James Allen Evans, "Social centralization and semantic collapse: Hyperbolic embeddings of networks and text," arXiv preprint, 2020, arXiv: 2001.09493.

[11] R. M. Gray, *Entropy and Information Theory* (New York: Springer Verlag, 2011)

[12] See S. Amari and H. Nagaoka, *Methods of Information Geometry*, Translations of Mathematical Monographs (New York: American Mathematical Society, 2007) and M. M. Deza and E. Deza, *Encyclopedia of Distances*, Encyclopedia of Distances (Berlin, Germany: Springer Verlag, 2009).

[13] One can get around some of these difficulties by keeping multiple measures of distance in play. For example, one can measure distance between texts on the basis of their lexicons, and also on the basis of their semantic features measured by, say, a topic model decomposition; see Matthew Wilkens, "Genre, Computation, and the Varieties of Twentieth-Century U.S. Fiction," *Journal of Cultural Analytics*, November 1, 2016, doi:10.22148/16.009. This enables one to say (for example) that A is close to B, and far from C, when it comes to word usage, but close to C and far from B in topic space. One can then, for example, say that the total topic distance in going from A to B to C is smaller than the lexical distance from A to C. However, the constraints of the axioms still apply when one considers each dimension of variation in turn. If A is close to B in word usage, then B is close to A in word usage (Axiom 3), and total topic distance accumulated in going from A to B to C can never be less than the topic distance between A and C alone (Axiom 4).

[14] Matthew W. Crocker, Vera Demberg, and Elke Teich, "Information Density and Linguistic Encoding (IDeaL)," *KI–Künstliche Intelligenz* 30, no. 1 (2016): 77–81.

[15] Laurent Itti and Pierre Baldi, "Bayesian Surprise Attracts Human Attention," *Vision Research* 49, no. 10 (2009): 1295–306, doi:10.1016/j.visres.2008.09.007; for a recent work, by a cognitive scientist, on the general notion of surprise in literature, see V. Tobin, *Elements of Surprise: Our Mental Limits and the Satisfactions of Plot*

(Cambridge, MA, USA: Harvard University Press, 2018).

[16] Wolfgang Iser, "The Reading Process: A Phenomenological Approach," *New Literary History* 3, no. 2 (1972): 279, doi:10.2307/468316.

[17] Viktor Shklovsky, "Art as Technique," in *Russian Formalist Criticism: Four Essays*, Second Edition, ed. Lee T. Lemon and Marion J. Reis, Regents Critics (1917; University of Nebraska Press, 2012), 3–25.

[18] Claude Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal* 27 (1948): 379–423. We present an introduction to information theory in the tutorial appendix.

[19] In the physical sciences, the same quantity is known as "entropy," and provides a remarkable link between epistemic and metaphysical concepts; see Edwin T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review* 106, no. 4 (1957): 620.

[20] Most famously, on the information theory side, Claude Shannon himself: see "The Bandwagon," *IRE Transactions on Information Theory* 2, no. 1 (1956): 3; on the literary scholarship side, see, as an example, John Gunders, "Signal or Noise? Information Theory and the Novel," *Double Dialogues* 3 (2002), whose critical evaluation of criticism that draws on information theoretic concepts, and, more recently, Michael Gavin, "Vector Semantics, William Empson, and the Study of Ambiguity," *Critical Inquiry* 44, no. 4 (2018): 641–73.

[21] Dallas Liddle, "Could Fiction Have an Information History? Statistical Probability and the Rise of the Novel," *Journal of Cultural Analytics*, 2019, doi:10.22148/16.033; see also Jo Guldi, "The Measures of Modernity. Word Counts, Text Mining and the Promise and Limits of Present Tools as Indices of Historical Change," *International Journal for History, Culture and Modernity* 7 (November 3, 2019), doi:10.18352/hcm.589 and Jo Guldi, "Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora," *Journal of Cultural Analytics*, December 2018, doi:10.22148/16.030 which makes explicit use of Kullback–Leibler Divergence itself.

[22] Michail Bakhtin, *The Dialogic Imagination: Four Essays*, trans. Michael Holquist and Caryl Emerson, University of Texas Press Slavic Series 1 (Austin, TX: University of Texas Press, 2011).

[23] See Stefania Degaetano-Ortlieb and Elke Teich, "Toward an Optimal Code for Communication: The Case of Scientific English," *Corpus Linguistics and Linguistic Theory* 0, no. 0 (2019), doi:10.1515/cllt-2018-0088, where they establish this "subtle effect" of scientific language enclosing general English in passing. They do so precisely in the way we argue for here, by reference the asymmetric nature of KL divergence. In their phrase, which captures the cognitive aspect of the information theoretic quantities in play, the enclosure relationship indicated by the KL asymmetry implies that general English is a worse "model" for scientific English than vice versa. See also Stefania Degaetano-Ortlieb et al., "A diachronic perspective on efficiency in language use: that-complement clause in academic writing across 300 years," in *Proceedings of the 10th International Corpus Linguistics Conference* (Cardiff, Wales, UK, 2019).

[24] Enclosure contains, implicitly, a notion of level of resolution, or coarse-graining. If we consider only the chapters on macroeconomics, the basic text will skip certain topics that the advanced macro text includes.

[25] For this toy example, friendship words are: friend, friendship, companionship, fellowship, camaraderie; suffering words are: suffering, hurt, ache, pain, agony, miserable, wretched; war words are: war, conflict, warfare, com-

bat, fighting, struggle, armed, military, bloodshed; and love words are: love, lover, fondness, tenderness, warmth, intimacy, attachment, endearment.

[26] Formally, our operationalization of enclosure, $\mathrm{KL}(p_1|p_2) - \mathrm{KL}(p_B|p_A) > 0$, indicates that p_A encloses p_B.

[27] Vicky Svaikovsky et al., "Racial Lines: Race, Ethnicity, and Dialogue in 780 Hollywood Films, 1970–2014," *McGill .txtLAB Collaborations*, June 2018.

[28] Dick Hebdige, *Subculture: The Meaning of Style* (1979; London: Routledge, 1988).

[29] Time was a central focus of a study of political speeches in the French Revolution by Alexander T. J. Barron et al., "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences* 115, no. 18 (2018): 4607–12, doi:10.1073/pnas.1717729115, which made extensive use of Kullback–Leibler Divergence to study the evolution of political discourse over time. That paper considered a quantity related, but not identical to, enclosure, called resonance. Enclosure is a pair-wise relationship: given a pair of texts, it quantifies the extent to which one encloses the other. Resonance, by contrast, considers the extent to which the enclosure of a pair of texts is mediated by a third text: when resonance is large (and positive), one can say that the future encloses the past "from the point of view of the present." See the section on enclosure in the appendix for further discussion.

[30] This counting is not trivial; see F. Harary and E. M. Palmer, *Graphical Enumeration* (NY: Academic Press, 1973), 126, 245.

[31] Data, and code, necessary to do this analysis is available at the GitHub Repository `https://github.com/kentchang/ca_divergence/`.

[32] David M. Blei and John D. Lafferty, "Topic Models," in *Text Mining: Classification, Clustering, and Applications*, ed. Ashok Srivastava and Mehran Sahami, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series (Boca Raton, FL: CRC Press, 2009), 71–93.

[33] Jaimie Murdock and Colin Allen, "Visualization Techniques for Topic Model Checking," 2015,

[34] I.e., we find the chapter j that maximizes $\mathrm{KL}(i|j) - \mathrm{KL}(j|i)$.

[35] This is when $\mathrm{KL}(p_b|p_a)$ can be larger than $\mathrm{KL}(p_c|p_a) + \mathrm{KL}(p_b|p_c)$.

[36] Here we follow methods used in William H. W. Thompson, Zachary Wojtowicz, and Simon DeDeo, "Lévy Flights of the Collective Imagination," arXiv preprint, 2018, arXiv: 1812.04013. The use of chunks (i.e., constant length units) rather than paragraphs (i.e., natural breaks between sentence groups indicated by the author) eliminates bias driven by the way in which the entropy of a text can depend upon its length, which in turn can impact the enclosure relationship because the KL divergence has a term that, all other things being equal, depends upon entropy; of course, this chunking then destroys the semantic information carried by the author's use of paragraph breaks itself.

[37] I.e., for paragraph i, $S(i) = \mathrm{KL}(i+2|i) - [\mathrm{KL}(i+1|i) + \mathrm{KL}(i+2|i+1)]$. As before, data, and code, necessary to do this analysis is available at the GitHub Repository `https://github.com/kentchang/ca_divergence/`.

[38] William H. Sewell, "Historical Events as Transformations of Structures: Inventing Revolution at the Bastille,"

Theory and Society 25, no. 6 (1996): 841–81; R. Wagner-Pacifici, *What Is an Event?* (Chicago, IL: University of Chicago Press, 2017).

[39] We do not assume any mathematical background of our readers: $\log_2 x$ is the logarithm of x to base two. If $b^a = x$, the logarithm of x to base b is defined as $\log_b(x) = a$, e.g., $\log_2(8) = 3$ because $2^3 = 8$ and $\log_2(2^3) = 3$. Let's work through $h(x_1)$ together: We know $P(x_1) = 0.5$ (the probability of the first question leading to the answer), so:

$$\begin{aligned} h(x_1) &= -0.5 \cdot \log_2(0.5) \\ &= -0.5 \cdot \log_2 \frac{1}{2} \\ &= -0.5 \cdot \log_2 2^{-1} \\ &= -0.5 \cdot (-1) \\ &= 0.5 \end{aligned}$$

[40] As a brief introduction to entropy, this essay does not cover situations where $H(X)$ is not entirely accurate.

[41] See section 2 in `http://tuvalu.santafe.edu/~simon/it.pdf`.

[42] This kind of logarithmic bias is pervasive in human life. The ancient Greeks described stars as "first," "second," and "third" magnitude, and so forth, a practice that persists today in contemporary astrophysics. In making these categories, they turned out to have grouped by factors of roughly two in brightness; a second magnitude star is roughly twice as dim as a first magnitude star. Sound perception is similar; a unit perceptual shift in loudness corresponds to a multiplicative, rather than additive factor, which engineers now quantify with the decibel scale. Behavioral evidence suggests we interpret probabilities in a similar fashion, seeing the risks from airplane flight as closer to automobile driving than we should under many normative standards. And so forth; we are logarithmic perceivers, and information theory simply extends this to the realm of learning and signals.

[43] Laurent Itti and Pierre Baldi, "Bayesian Surprise Attracts Human Attention," *Vision Research* 49, no. 10 (2009): 1295–306, doi:`10.1016/j.visres.2008.09.007`.

[44] This is the use case, for example, in Jaimie Murdock, Colin Allen, and Simon DeDeo, "Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks," *Cognition* 159 (2017): 117–26; Alexander T. J. Barron et al., "Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences* 115, no. 18 (2018): 4607–12, doi:`10.1073/pnas.1717729115`; and William H. W. Thompson, Zachary Wojtowicz, and Simon DeDeo, "Lévy Flights of the Collective Imagination," arXiv preprint, 2018, arXiv: `1812.04013`.

[45] Karl Friston and Stefan Kiebel, "Predictive Coding under the Free-Energy Principle," *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, no. 1521 (2009): 1211–21.