Title: Learning incommensurate concepts

Authors:

1. Hayley Clatterbuck (corresponding author)

Rethink Priorities, San Francisco, CA, USA

hayley.clatterbuck@gmail.com

ORCID: 0000-0003-3041-8232

2. Hunter Gentry

Kansas State University, Manhattan, KS, USA

hrgentry@ksu.edu

Abstract: A central task of developmental psychology and philosophy of science is to show how humans learn radically new concepts. Famously, Fodor has argued that such learning is impossible if concepts have definitional structure and all learning is hypothesis testing. We present several learning processes that can generate novel concepts. They yield transformations of the fundamental feature space, generating new similarity structures which can underlie conceptual change. This framework provides a tractable, empiricist-friendly account that unifies and shores up various strands of the neo-Quinean approach to conceptual development.

Keywords: machine learning; concepts; incommensurability; bootstrapping

Statements and Declarations: The author has no competing interests to declare.

Acknowledgements: The authors would like to thank: Cameron Holdaway, Farid Masrour, Larry Shapiro, and Gary Lupyan; audiences at CUNY Graduate Center, Stanford, Maryland, Carleton University, Colorado, and the HaMLET group at the University of Wisconsin; and attendees of a graduate seminar on non-linguistic thought at the University of Wisconsin-Madison.

1. Introduction

Throughout development, humans seem to undergo radical changes to many concepts that are central to our understanding of the world: number, causation, agency, and basic physical properties. These conceptual revisions are so great that they allow us to think thoughts that were previously unavailable to us. They involve "discontinuities, resulting in systems of representation that are more powerful than, and sometimes incommensurable with, those from which they are built" (Carey 2009, 113). This view of development draws inspiration from radical theory change in scientific history, where new concepts have been developed (e.g. *electron, market efficiency*) and old concepts significantly overhauled (Kuhn 1962).

Understanding when and how these changes can occur has been a central focus of cognitive science in recent decades. This project faces a fundamental challenge: if all learning processes are operations on existing concepts, how could they possibly generate concepts that are genuinely different from the ones they started with? How is it possible to learn something that you don't already, in a sense, know?

The challenge for explaining radical concept change has been particularly acute in the classical paradigm, which views concepts as mental symbols that are manipulated in accordance with syntactical rules. The existence and meaning of such symbols are taken as a given. However,

> a successful system must be able to *learn* radically new properties from its interactions with the world and not only form new combinations of the given predicates. This has turned out to be an enigma for symbolic representations. This gives rise to an essential question for a theory of cognitive representations: *where do new predicates come from?* (Gärdenfors 2000, 38).

Those who have taken up the challenge of answering this question have focused on two main strategies. First, some in the symbolic tradition have offered accounts of learning processes beyond mere recombination, including bootstrapping and learning by analogy, that can introduce new symbols and

radically reconfigure the meanings of old ones (e.g. Carey 2009). Second, some have argued that we must reject the symbolic account of concepts in order to show how conceptual change is possible (e.g. Quine 1969, Gärdenfors 2000, Laurence and Margolis 2012).

We will argue that each of these approaches contains important insights but has serious limitations. First, accounts of alternative learning processes often suffer from a lack of detail, opening them up to allegations of circularity and imprecision. Second, non-symbolic theories often have difficulty capturing the stability and discreteness of many of our concepts and explaining how learning interfaces with language. Further, one of the most promising such theories, Gärdenfors' conceptual space theory, has little to say about how genuinely new concepts are formed.

Fortunately, algorithms developed in the fields of data analysis and machine learning provide us with helpful models of tractable learning processes that can result in radical concept change. Furthermore, while the basic representational format of such processes is non-symbolic, they allow us to see how continuous, non-symbolic representations interface with discrete, symbolic linguistic ones. Not only do these models allow us to provide detail to existing accounts, they yield a picture of how learning can radically reconfigure the conceptual landscape, in a way that changes subsequent concept learning.

In Section 2, we present Fodor's challenge, and in section 3, two neo-Quinean responses. In Section 4, we present a conceptual space framework for thinking about conceptual change. In Sections 5 and 6, we present several learning processes that can yield radically new concepts. In Sections 7 and 8, we present Carey's (2009) explanation of the bootstrapping process by which children learn the integers and our alternative, conceptual space account.

2. Arguments against radical concept change

Puzzles about how it is possible to learn something genuinely new have been with us for a long time. A learner's paradox has a general form: in modeling the process by which an individual could learn

X, we discover that the individual must already have known X. We will focus here on the learning of concepts that are genuinely new; in order to learn some new concept, one must already possess that concept or concepts from which it can be readily expressed, in which case it is not genuinely new.

Fleshing out the general argument against radical concept change[1] involves three pieces:

(a) a specification of concept identity conditions, what it takes for new concepts to be different from the ones from which a learning process began

(b) a specification of the available learning processes (including their inputs and outputs), and

(c) an argument showing that the learning processes described in stage (b) cannot result in the differences described in stage (a).

For example, Kuhn argues that (a) concepts in different scientific paradigms are incommensurate with one another. Since (b), a scientific change is only rational relative to the principles of a paradigm, there can be no rational scientific change across paradigms. Another example: the British empiricists argue that (a) ideas are individuated by their corresponding sense impressions. Since (b) learning can only preserve and recombine the ideas arising from sense impressions, it cannot result in abstract concepts that are irreducible to perceptual features or regularities (Berkeley 1710/1975, Hume 1739/1978).

Most of the current debate about radical concept change is posed as a response to Fodor's (1981, 1990, 1998) formulation of the challenge. First, Fodor holds that (a) complex concepts have definitional structure; a concept C is identified by the set of necessary and sufficient conditions for something to fall under the extension of C. On this view of concepts, definitions can be formed only by

[1] Following Beck's (2017) analysis of bootstrapping, cases of *modest* concept learning are those in which new concepts are fully expressible in terms of existing concepts; learning need only prompt one to combine existing representations in a novel way. In cases of *radical* concept learning, new concepts are not fully expressible in terms of existing ones, so the transition results in an increase in expressive power.

logical combinations of more basic lexical concepts[2], and a radically new concept is one that cannot be defined in terms of the other concepts that a learner possesses. Second, Fodor holds that (b) all learning proceeds via hypothesis confirmation (Fodor 1975, 95). In order to learn something, one must be able to confirm a hypothesis about it, which in turn requires that one be able to form an antecedent representation of that hypothesis.

His argument (c) against learning radically new concepts proceeds as follows. Consider a set of concepts at t_1 (CS1) and a set at t_2 (CS2).

1. Assume for reductio: a concept, C, of CS2 is not definable in terms of concepts in CS1.
2. All learning is hypothesis confirmation.
3. To learn C is to confirm a hypothesis about the meaning of C.
4. In order to confirm a hypothesis about the meaning of C, one must be able to represent that hypothesis.
5. If C is not definable in terms of representations in CS1, then the subject cannot represent that hypothesis.
6. From (1,4,5), the subject cannot learn C.

C: If the meaning of a concept C of CS2 is not definable in terms of representations in CS1, then the subject cannot learn it.

Since radically new concepts are ones that cannot be defined in terms of previously available concepts, radical concept learning is impossible.[3]

3. Neo-Quinean responses

[2] These concepts are word-like, in that they are symbolic, have stable meanings, they contribute to the meaning of thoughts via syntactic combination with other concepts, etc.

[3] To arrive at Fodor's notorious "mad dog nativism" - the view that concepts such as AVOCADO or DOORKNOB or ELECTRON are innate - we need a few additional premises. First, Fodor maintains that these concepts, like most monomorphemic concepts, cannot be defined in terms of other concepts (indeed, they may lack internal structure entirely). Hence, from the above argument, they cannot be learned. Second, since these concepts cannot have been learned, they must either be innate or acquired through some other non-rational process.

Fodor's challenge rests on two key assumptions: the only structure concepts could have is definitional structure and hypothesis confirmation is the only available learning process. The most prominent, and promising, attempts to tackle Fodor's problem head-on take inspiration from Quine (Quine 1969; Carey 2009; Margolis 1998; Laurence and Margolis 2002, 2012; Strevens 2012) and reject one or both of these assumptions. If there are learning processes that are sensitive to non-definitional (or, more broadly, any non-language-like) structure[4], or if there are learning processes that can recover linguistic structure through routes other than logical combination of existing representations, then they might be able to generate radically new concepts. Here, we will give a brief summary of the neo-Quinean responses and their shortcomings, though we should note that our proposal will fall within the general spirit of the neo-Quinean approach.

3.1. Rejecting hypothesis testing

As Weiskopf (2008) notes, the assumption that learning proceeds via recombination and hypothesis testing seems sufficient to rule out radical concept change: "Whatever the initial primitive concepts that are given to us happen to be, it seems that it must be possible to extend this endowment somehow. But this is impossible if recombination is the only available method of arriving at new concepts" (361). This has spurred investigations of other kinds of learning processes, such as learning by analogy (Gentner and Markman 1997, Holyoak and Thagard 1997), the triggering of domain-specific concept generating modules (Margolis 1998, Strevens 2012, Weiskopf 2008), and linguistic bootstrapping (Carey 2009).

We will focus on Carey's view as a test case. For the most part, Carey seems to retain Fodor's assumption that new concepts have lexical structure. However, she argues that you need not learn new

[4] Where language-like structure involves discrete symbols with stable meanings that contribute semantic content to thoughts via syntactic combination. We don't intend to take a strong stand on how best to distinguish linguistic and non-linguistic thought.

lexical concepts bottom-up by recombining existing concepts. Instead, you can learn new lexical primitives top-down, first learning a new lexical structure and then learning the primitives that comprise it. It is like "scrambling up a chimney supporting oneself by pressing against the sides one is building as one goes along… the aspect of the bootstrapping metaphor that consists of building a structure while not grounded is applied as the learner initially learns the relations of the system of symbols to one another, directly, rather than by mapping each symbol onto pre-existing concepts" (Carey 2009, 306).

In broad strokes, an episode of radical bootstrapping from CS1 to an incommensurate CS2 proceeds as follows (Carey 2009, Beck 2017). At CS1, the learner has a suite of represented concepts, as well as computational constraints on how her concepts are utilized, altered, and combined; these computational constraints are not themselves represented. The subject learns a linguistic placeholder, a structured natural language item that starts off at least partially meaningless. The computational constraints present in CS1 concepts and external constraints imposed by the placeholder itself provide necessary structure and are used to partially interpret the placeholder. Then the learner undergoes various modeling processes – structure mapping, analogy, induction, etc. – to fully interpret the new conceptual structure.

An example from Block (1986) will help to illustrate the process[5]:

> When I took my first physics course, I was confronted with quite a bit of new terminology all at once: 'energy, momentum, acceleration, mass' and the like… I never learned any definitions of these new terms in terms I already knew. Rather, what I learned was how to use the new terminology—I learned certain relations among the new terms themselves… , some relations between the new and the old terms, and, most importantly, how to generate the right numbers in answers to questions posed in the new terminology (648).

[5] Though Carey does not think that this is a genuine case of linguistic bootstrapping, it still provides a helpful illustration of at least one part of it.

Consider Block who at CS1 first learns "force = mass x acceleration" but has, at best, an inchoate grasp of the terms related by that equation. At CS2, he has somehow arrived at a much fuller interpretation of these physical magnitudes, perhaps one that fundamentally restructures the understanding of MASS that he had at CS1 (e.g. from a concept of MASS as "total volume taken up by an object" to a concept of MASS as "resistance to acceleration"). How is this possible?

Fodor assumes that conceptual structure must be built from the bottom up, from pieces that you already represent. Block's example suggests that we can sometimes grasp the higher-order structure of a new concept first. The constraints provided by "F=ma" plausibly include things like: whatever F is, it's different from m and from a; if F stays the same and m increases, a goes down; and so on. Block is then free to sift among his existing concepts to see which of them obey these higher-order relations, to modify existing concepts to make them obey these constraints, and perhaps to start fleshing out brand new concepts introduced by the placeholder terms.[6] In brief, (i) a new word spurs the creation of a new concept, (ii) the relations that word enters into provide constraints, things that must be true of the newly learned concept, and (iii) the learner uses various processes to construct or adapt a new concept that obeys the constraints in (ii).

Block's case nicely illustrates how new linguistic input can provide constraints on conceptual learning, but it doesn't address the more puzzling question of how these constraints prompt radical conceptual revisions (Carey 2009, 219). In particular, Block's case focuses on one element of the bootstrapping process, the constraints provided by language, but omits the other key elements: existing mental resources and how they are modified to create new concepts. The bootstrapping examples that Carey provides, particularly the case of integer learning in early childhood, are far more detailed and make more sense of the roles played by existing conceptual resources and modeling processes. However, as Beck (2017) notes, critics have alleged that the account is still too underspecified:

[6] Another possibility is that the new input triggers an innate module that creates new natural kind concepts (Margolis 1998, Strevens 2012).

> While it purports to explain how important new concepts are learned, many commentators complain that it is unclear just what bootstrapping is supposed to be or how it is supposed to work… others allege that bootstrapping falls prey to the circularity challenge: it cannot explain how new concepts are learned without presupposing that learners already have those very concepts. (111)

As an example of the latter criticism, Rey (2014) objects that learning new concepts via analogy is impossible. In order to restructure your existing concepts so that they obey the relational constraints inherent in the placeholder, you must already represent the relation exhibited by the placeholder. That is, in order to form an analogy between two domains, one must already represent what it is they have in common in virtue of which they are alike. But if the child already represents the common structure, then she has nothing left to learn.

With respect to learning by analogy, the challenge is to identify a process by which two domains can be aligned and generate a new representation of shared structure without an antecedent representation of that target structure (Gentner 2010; Holyoak and Thagard 1995, 1997). In general, the challenge is to show how existing conceptual resources are adjusted to conform to new conceptual roles, especially in light of linguistic input.

3.2. Rejecting linguistic structure

A second strain of neo-Quinean views identifies Fodor's language-like approach to concepts as the chief impediment for understanding conceptual change. For Fodor (at least in some works), primitive concepts are atomic, with no internal structure. The vehicle of a concept – a symbol – does not represent intrinsically; that is, the symbol does not bear any resemblance to what it represents. Therefore, there is no internal similarity structure for learning processes to grab hold of. Neo-Quineans argue that concepts

can have other sorts of structure –such as perceptual similarity (Quine 1969, Laurence and Margolis 2012), prototypicality[7] (Hampton 2006), or kind syndromes (Strevens 2012, Margolis 1998) –that can be exploited to learn new concepts.

Quine's own (1969) account preserves the assumption that learning proceeds via hypothesis testing while rejecting that natural kinds are linguistically defined. Consider a learner who attempts to learn the extension of a novel word, "yellow". She first encounters the word paired with a yellow object and proceeds by entertaining various hypotheses about the extension of the word. If unconstrained, the hypothesis space she considers will be enormous; "yellow" might pick out {all yellow objects}, {just this banana}, {this banana and that blue car}, {yellow objects observed before *t* and blue objects thereafter}[8], etc.

In order to provide the necessary constraints on learning, Quine posits that there is an innate similarity ordering among colors. The individual doesn't have to learn that yellow is more similar to orange than it is to blue – that's built in.[9] We can represent this innate ordering via a similarity space that has dimensions corresponding to features that are discriminated in the input, and geometric distances in the space reflect relative similarity.[10]

[7] Fodor (1998) argues that so-called concepts without definitional structure, such as prototypes, fail to be genuine concepts because they don't classically compose. See Laurence and Margolis (1999) for a discussion.

[8] Quine's (1969) account is presented as an attempted solution to Goodman's (1955) grue paradox. For a diagnosis of why the grue paradox arises for propositional representations and why similarity orderings can solve the problem, see Gärdenfors (1990). For discussions of how the grue problem relates to bootstrapping, see Rey (2014) and Beck (2017).

[9] For example, it is "a commonplace of behavioral psychology" that "a response to a red circle, if it is rewarded, will be elicited again by a pink ellipse more readily than by a blue triangle; the red circle resembles the pink ellipse more than the blue triangle" (Quine 1969, 46)

[10] This similarity metric need not reflect the actual similarity among the things in the world, see Gallistel (1990) and Gärdenfors (2004).

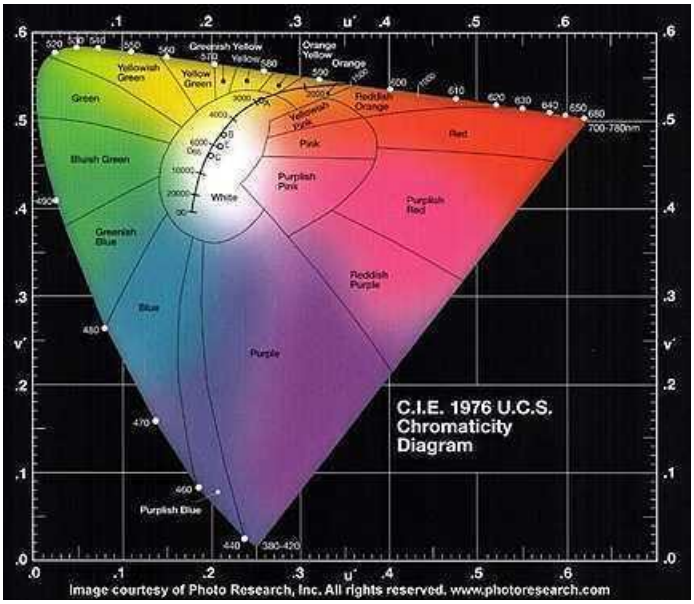

Figure 1 *–The International Commission on Illumination 1976 Chromaticity Diagram, a similarity space in which the distance between two points represents perceived color difference (Blaise 1976).*

We can reframe our "yellow" learner's task as trying to find the region of her innate similarity space that captures the extension of "yellow". According to Quine, her hypothesis space is constrained by the fact that kinds (at least initially) are comprised of objects that are *similar*, that is, that are located in contiguous regions of this similarity space. The set of possible concepts is the set of contiguous regions of the space which encompass the objects known to be yellow. The learner's task, then, is to find its boundaries (e.g. just how green can something be before it is no longer yellow?).

Through this process, our learner can come to possess a new primitive concept, YELLOW. In contrast to Fodor's "building blocks" model, this process does not require that the learner had an antecedent representation of the relation that all yellow things bear to one another. For Quine, neither the relations captured in the innate similarity space nor the space itself are represented. Instead, the learner has a disposition to treat yellow things as more similar to each other than to non-category members, and a *representation* of similarity is a result rather than an input to the process.

This account promises to explain how non-lexically structured concepts are learned in a way that evades the circularity challenge. However, we doubt that it can provide a full account of how genuinely incommensurate concepts can be learned, for it falls prey to a reformulation of Fodor's argument.

If the similarity space is to do any inductive work, then learning processes must exploit geometric properties of the space. That is, there must be geometric constraints on the possible concepts to be learned. Following Quine, let us assume for the moment that it is a computational constraint on the learning mechanism that concepts be convex regions of the similarity space.[11] Convexity allows the implicit geometric structure of the space to do the inductive work of categorizing new instances; if you know that two points are members of a category, you can infer that all points lying between them are as well. Even if convexity isn't the right constraint, the Quinean has to posit some constraint or other. If learning were unconstrained, then concepts would be neither learnable nor inferentially robust.[12]

However, if learning processes are restricted to learning concepts that are, say, convex regions of perceptual similarity space, then it will only be possible to learn concepts that correspond to neat, simple perceptual regularities. Hence, the similarity space view, as stated, could not explain how radically new concepts could be learned, let alone abstract concepts whose instances are not linked by perceptual similarity.[13]

If the similarity space framework is to explain radical concept change, then it must be possible to extend or change the similarity space itself. As Laurence and Margolis (2012) note:

> The innate quality space might not be developmentally fixed. The size or dimensions of this space might be altered. Relational parameters within a quality space might also be altered, or new relations superimposed onto the space. There could also be multiple distinct quality spaces

[11] That is, concepts pick out contiguous regions, such that for any points X and Y that fall within the concept C, an individual Z that is on the straight line connecting X and Y will also be a member of C.

[12] Take, for example, Beck's (2017, 113) case of learning the concept BURSE, where something is burse if it is either green and circular or blue and enclosed by a prime number of sides, or red and preceded in presentation by a yellow triangle.

[13] See Smith and Heise (1992) for a defense of the claim that learning over perceptual similarity spaces can yield higher-level, theoretical concepts.

and quality spaces that stand in different relations of psychological accessibility to one another (12).

In Sections 4-6, we will show how techniques from machine learning allow us to more precisely develop the similarity space metaphor and to see how simple learning processes can yield significant transformations of existing similarity spaces and radically new concepts. Then, we will evaluate whether we can use these techniques to model the kinds of learning processes going on in bootstrapping.

4. A framework for thinking about conceptual change

4.1. The conceptual space framework

While we have no doubt that investigations into more complicated AIs and their implementations will be helpful in understanding conceptual change (Buckner 2024), we will focus on very simple learning algorithms. We also acknowledge that there are important questions to be asked about whether various machine learning techniques are of a kind with human learning and whether their utility in cognitive science depends on such similarity (Weiskopf 2011; Buckner 2015, 2018). We will treat them as a helpful explication of the similarity space model and a "how possibly" story about the functional organization of human concept learning and the kinds of representational vehicles it involves.

The algorithms that we will discuss start by mapping data points onto an *n*-dimensional feature space, where each of the *n* recorded features forms an axis of the space. The dimensions and choice of scale give the space geometric structure, and similarity between objects is measured by the distance between their data points relative to this underlying geometry. Machine learning algorithms exploit the geometric relations inherent in the feature-space representation of the data in order to achieve various epistemic goals. Before examining specific examples, we will first present a framework, based on Gärdenfors's (1990, 2000, 2004) conceptual space theory.[14]

[14] For similar accounts, see also Shepard (1987) and Churchland (1995, 1998). For analysis and criticisms of the view, see Gauker (2011).

A conceptual space consists of quality dimensions, such as color or weight, and a data point's position in the space is represented by a vector specifying its value for each dimension. Dimensions of a space are assumed to be logically independent of one another. The structure of the conceptual space may vary depending on whether the qualities that define it are continuous or discrete, on a quotient or logarithmic scale, etc. The similarity between two individuals corresponds to the distance between the two points that they occupy in the space.[15]

For example, suppose that only the height and weight of objects are recorded, so the feature space is two-dimensional. Height and weight are logically distinct; assigning an individual a height value of, say, 1 meter does not entail that its weight must take any particular value. If we adopt a Euclidean distance metric, the similarity between two objects is given by the length of the straight line connecting them (Gärdenfors 1990, 84).

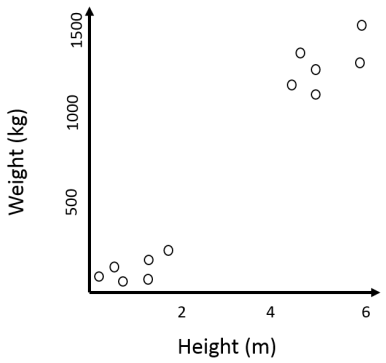

***Figure 2** – A 2-dimensional plot of the height and weight of individuals.*

4.2. Learning processes over conceptual spaces

It is common to distinguish between two basic types of learning goals. In classification, the goal is to infer groups in the data and then to classify new instances into groups based on a similarity metric to existing group members. In regression, on the other hand, the goal is to learn the mathematical

[15] See Gärdenfors (2000, 18-20) for a discussion of alternative non-Euclidian distance measures.

function that captures a trend in the data and then use that function to predict the value of a response variable. For illustration, return to the feature space in Figure 2. A regression analysis might find the curve that best represents the correlation between height and weight in the data and then predict the probable weight of a new object given its height.

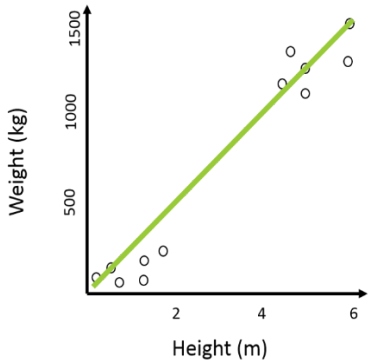

Figure 3 *– Linear regression capturing relationship between height and weight in the sample.*

In classification, the goal is to infer groups in the data and then to classify new instances into groups based on a similarity metric to existing group members. A classifier will find groups in the data and predict the probable species membership of a new data point given its height and weight. Among classifiers (especially), another key distinction is between supervised and unsupervised learning processes. In supervised learning, the data that is used to train the algorithm is labeled, and the algorithm learns a rule for accurately predicting the label of new items. It is supervised in the sense that we stipulate which features or labels "X in the dataset constitute the 'ground truth' values for learning; that is, the supervised learning algorithms use the known values of X to determine what should be learned" (Danks 2014, 154). In our example, in addition to inputting the feature values of our data, we would label points as belonging to the category "human" (blue dots) or "giraffe" (red). The algorithm's task is to predict whether a new object is a giraffe or a human. In unsupervised learning, there are no category labels provided, so the algorithm first has to discover groupings based on structural information alone.

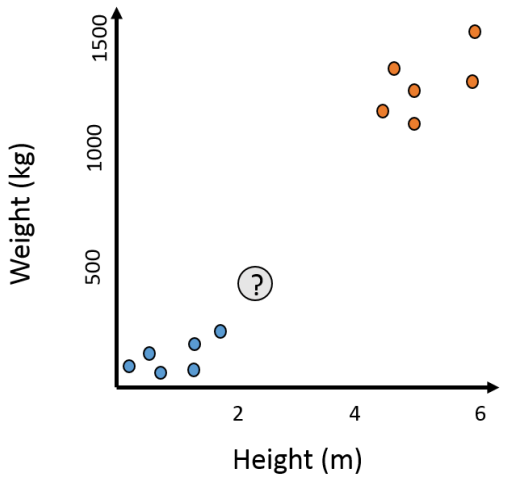

Figure 4 *– A supervised learning task, with humans labeled blue and giraffes labeled red. A novel point is to be classified from the rule inferred from the labeled data.*

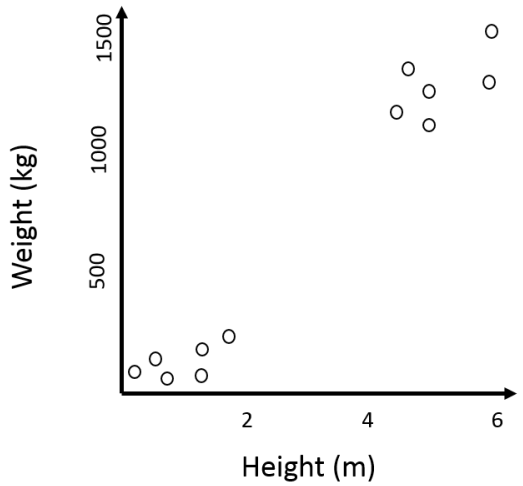

Figure 5 *– An unsupervised learning task, in which groups would have to be inferred from geometric features of unlabeled data.*

4.3. Concepts in the conceptual space framework

What is a concept, according to this framework? We distinguish between two kinds of representations that correspond to different kinds of concepts. First, some concepts correspond to regions of the conceptual space. For example, the bottom-left region of the space in Figure 2 might pick

out the concept {humans} and top-right {giraffes}.[16] We'll call these *categorizations*, though "conceptualizations" or "groupings", or "manifolds" might do the job (DiCarlo and Cox 2007). Second, some concepts correspond to dimensions of the space itself, such as {height} and {weight}. We will call these *framework* concepts. The difference between these types of concepts should be intuitive enough, corresponding to a rough distinction between kinds and properties.[17]

In some cases of modest concept learning, the individual learns to form new categorizations in the existing similarity space. For example, you might learn that there is a region in the bottom-right of the space in Figure 2 that picks out {soaring birds}. In other cases, you might make modest changes to the conceptual space, such as adding dimensions (e.g. adding {color} as a third dimension) or slightly changing the scale of existing ones. In cases of *radical* concept learning, changes to the framework concepts lead to such significant changes in the underlying similarity space that the subsequent categorizations formed against this new space are very different from the ones formed on the previous one.[18]

Recall that in order for a similarity space to do inductive work, there must be geometric constraints on the kinds of regions of the space picked out by our categorization concepts. While we might form the occasional gerrymandered grouping, we expect most of our inductively-rich concepts to encompass regular regions of the similarity space. Gärdenfors, like Quine, defines *natural concepts* as

[16] Of course, one's full concept of humans or giraffes would involve many more dimensions than this.

[17] Gauker (2011, esp. Ch. 3) presents several arguments for why concepts should not be identified with regions of similarity space and that judgments, concepts, and kind-thinking only arise with language. We disagree, but full engagement with his arguments is not possible here. In later sections, we will discuss how language interfaces with perceptual spaces, making categorizations more discrete, salient, and available for reasoning. If one wants to follow Gauker and reserve "concepts" for just this stage, we will not object.

[18] Gärdenfors and Zenker (2013) model paradigm shifts in science in this way. For example, there may be changes in the scale or metric of existing dimensions, as when an ordinal scale (e.g. warmer or colder) was replaced with a ratio scale (e.g. degrees Celsius). The relative importance of dimensions may change, as when color was dethroned as a key feature of chemical theories (ibid., 1046). Theory change may involve the addition or deletion of dimensions of physical reality, such as with energy or the ether, respectively, or distinct theories may become combined, as when distinct Newtonian dimensions of space and time became relativistic spacetime (ibid., 1047).

encompassing convex regions of a conceptual space[19] (2004, 18). We will proceed with convexity as a working example of a requirement on concepts, though it is not the only such candidate. Carnap, for example, preferred "to take as *primitive* predicates… only those with a *connected* region" (1980, 21). The condition of connectedness is weaker than convexity but would do non-trivial inductive work. Gauker (2011, 234) discusses clustered representations, where every member of a concept is between at least two other members. For many learning processes, we might want to place more stringent conditions, requiring that well-formed concepts be of a certain shape, size, density, or cohesiveness. The stronger these restrictions are, the faster and more constrained the resultant concept learning will be (Griffiths and Tenenbaum 2009).

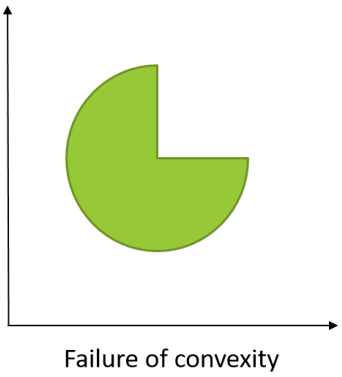

Figure 6 *– A region that violates the convexity condition.*

[19] A concept C is a natural concept *relative* to a conceptual space only if C encompasses a convex region in it. As Gärdenfors notes, GREEN but not GRUE is convex with respect to the innate quality spaces that probably all humans possess, but one could construct a conceptual space on which GRUE but not GREEN is convex.

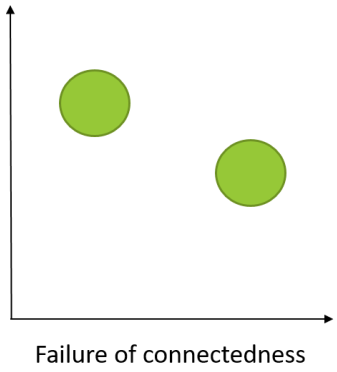

Figure 7 *– Regions that violate the connectedness condition.*

The goal of many learning algorithms is to find meaningful groups in the data. What happens when these meaningful groups do not obey the geometric requirements on natural concepts, e.g. when the group does not comprise a convex region?[20] The two options are: adopt a more complex geometric rule for picking out groupings or maintain the existing requirement and transform the underlying space so that the grouping now obeys it. Interestingly, in many machine learning applications, the latter (geometrically simple geometric concepts and complex transformations to the space) are preferred over the former (geometrically complex concepts that keep the space as is) (Buckner 2018, 5348). Instead of changing our requirements on groupings, we prefer to adjust our framework concepts so that the groupings make more sense. Why might one favor this approach? Recall the lesson from Section 3 that similarity spaces only do inductive work if concept learning can exploit their geometric structure. If concepts can be concave, discontinuous, or otherwise gerrymandered regions, then the geometry is of little assistance in making inferences about new data points. We transform the space so that it can support future inductions.

To contrast these two approaches, consider an example using support vector machines (SVMs). SVMs are a popular classification algorithm that uses labeled training data to find the hyperplane that best divides the known categories, which can then be used to classify new data points. SVMs work by

[20] We will later return to the question of how one could know that is a relevant group, given that it looks gerrymandered.

maximizing the distance between the classification boundary and the points closest to it. Suppose that you were attempting to find a linear separation rule that separates the blue As from the red Bs in the space below:

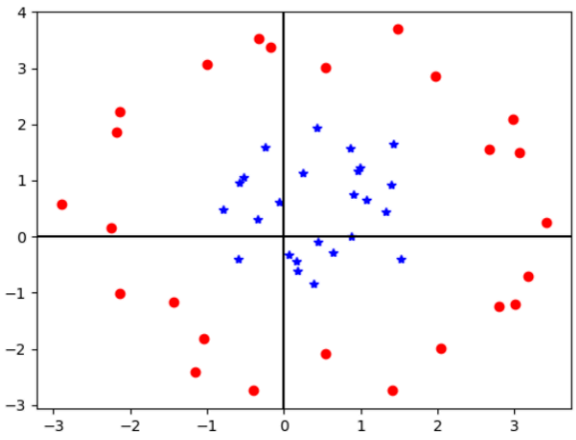

Figure 8 *– A supervised learning task where the groups are not linearly separable.*

There is no such rule (and no way to separate these into two convex groupings). One option would be to hold the underlying space fixed and find a complex function that separates the two groups; here, the separation rule could be circular, classifying everything within the circle as an A and everything outside as a B.

Alternatively, you could insist on finding a simple, linear separation in the data and transform the feature space so that the As and the Bs become linearly separable. Here's one way to do it: define a new feature, Z, such that an object's value of $z = x^2 + y^2$. Then, replace the existing y-axis with the z-axis and re-map the points against the X, Z space. In the resulting space, the groups are linearly separable.

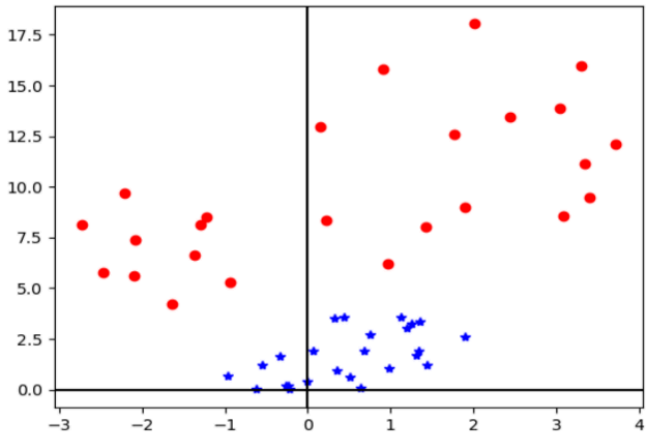

Figure 9 – A transformation of the feature space in Figure 5 such that the two groups are now linearly separable

One might object that, mathematically, there is no real difference between discovering a complex separation rule and transforming the underlying space. This might be correct, but the downstream consequences of the two may differ significantly. Since the feature space is treated as primitive, new inductive tasks formulated against a new space will be beholden to the transformations that have been made and may be radically different from ones in the previous space. Hence, if one wants a classification to preserve existing similarity orderings, one should opt for a complex separation function. However, for classifications that fundamentally change one's ontology, the latter is more appropriate.

In what follows, we will explore various techniques for transforming similarity spaces and their effects on conceptual development. For now, we can use the notion of a space transformation to characterize radical conceptual change. This occurs when changes to the dimensions of the underlying space (the framework concepts) result in spaces with very different similarity structures, and as a result, the natural groupings that are formed on top of these spaces are very different as well. On this account, two conceptual systems are incommensurate when the natural concepts (e.g. the convex regions) of the first system are not the natural concepts of the second system (or vice versa). The more stringent the

geometric constraints on concepts, the less likely they are to be preserved across transformations. For example, a transformation might preserve the connectedness of groupings but not their convexity.[21]

Next, we will present several machine learning processes that transform underlying feature spaces in ways that permit radical concept change. Then, we will put them to use to explain examples of radical concept learning through linguistic bootstrapping. To recall, bootstrapping involves: learning a linguistic placeholder structure; using the computational constraints of the placeholder and existing representations to partially interpret the placeholder; and using learning processes, such as modeling or analogy, to fully interpret the placeholder. To model bootstrapping in the conceptual space framework, we will present processes that: use labels to mark category members; transform spaces to accord with the groupings provided by labels; and use dimension reduction to "lock in" those changes to dimension spaces and form new fundamental concepts. We will present the latter dimension reduction techniques first, then move to learning processes operating over linguistic labels.

5. Dimension reduction techniques

As we noted, a conceptual space has dimensions corresponding to each of the recorded features in the data. However, in most inference problems, many or most of these dimensions will be unimportant; some dimensions will be strongly correlated with one another, and some dimensions will be irrelevant to the trends in the data that we are trying to capture. Dimension reduction techniques allow us to distill out just those dimensions that capture meaningful (non-redundant, non-noise) patterns in our data. Sometimes this may involve selecting a subset of the original dimensions of the

[21] We can cash out the degree or ways in which two spaces are incommensurable in terms of the geometric transformations that preserve the natural concepts in those spaces. For example, spaces D1 and D2 are topologically transformable if every continuous set of points in D1 is continuous in D2; topological transformations are permitted to stretch or expand the space, they cannot tear or paste it. So two spaces are topologically incommensurate if transforming one to the other fails to preserve the continuity, e.g. if a continuous region in D1 becomes a gerrymandered archipelago of regions in D2. Here, we take inspiration from Maudlin (2012) who argues that significant revisions to our understanding of the geometry of space yielded radically different results about motion, time, and fundamental physical symmetries.

space. In many others, it will involve the construction of new dimensions–new framework concepts–that replace the old ones and now constitute the fundamental ontology of the space.[22] We will present a couple of dimension reduction techniques, and consider how they can help us explain radical concept learning.

5.1. Principal Component Analysis

PCA is a simple example of a larger family of dimension reduction algorithms that, as the name suggests, serve to reduce the number of dimensions in the underlying feature space. It is often used in tasks in which perceptual input is many-dimensional but the true signal is much simpler. It allows us to find correlations in the data (features that "march together") and to collapse correlated dimensions to a single dimension which characterizes the trend.[23]

Consider again the plot of human and giraffe data from Figure 2. It has two dimensions, height and weight. Because height and weight "march together" in this plot, we could predict the species of a new data point via a single dimension that combines height and weight (call it "bulk"). To perform PCA on this data set, we perform a simple linear regression to find the "principal component", the line of greatest variance in the data which minimizes the least squared distance from the data points. This regression line will become our new x-axis[24], and our data points are now collapsed to it. The result is that we can represent the trend in the data using fewer dimensions, though we will lose information (proportional to how much the data points deviated from the regression line).

[22] Recall the example from Gärdenfors and Zenker (2013) of *spacetime*. While *space* and *time* were separate dimensions of the Newtonian conceptual space, they were collapsed into a single dimension in the relativistic conceptual space.
[23] PCA cannot be said to uncover true latent common cause variables. Other processes, like factor analysis, can.
[24] Typically, the orthogonal to the regression line will become the new y-axis, but since there is only one remaining dimension, we'll ignore it for now.

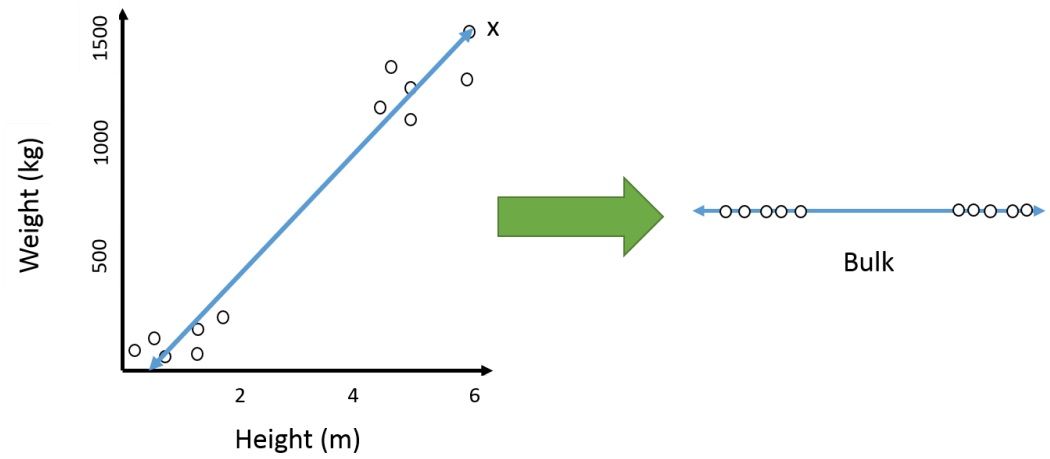

Figure 10 *- PCA dimension reduction from space D_1, with dimensions weight and height, to D_2, with a single dimension, bulk.*

The transformed space places new constraints and new inductive support for subsequent concepts formed against it. In the transformed space, weight and height are no longer distinct dimensions that can vary independently. This has significant upshots for induction based on single features. Suppose we are using the original 2-D space and observe only the height of a new individual. While we could make an *inference* about its weight, this inference wouldn't be entailed by the similarity space itself. However, plotting height on the regression line that defines bulk does entail a value for weight. Hence, dimension reduction collapses two variables such that they are now, for better or worse, tied together by something stronger than inferred correlation. In this way, the resulting conceptual space treats *bulk* as a primitive that is not decomposable into more basic constituents of height and weight. Granted, a user that is aware of the previous space might recall that bulk is a function of height and weight, but this unpacking is not necessarily recoverable by a user who starts with the new, 1-D space.

Notice that PCA uses a general learning algorithm (linear regression) that does not require that we first postulate or define the latent dimension that will be discovered. Indeed, while PCA may find trends in the data that correspond to some intuitive latent variable (like "bulk"), it may uncover trends that are surprising. These new framework concepts are not added and then related to existing concepts; rather, the variables and expressive vocabulary of the theory are changed (*contra* Rey 2014).

The geometric readjustments resulting from PCA are quite modest. Since PCA re-aligns state spaces to linear trends in the data, it can only yield translations and rotations of the original data. It will not, for example, turn a convex grouping into a non-convex one in the transformed space. Other dimension-reduction techniques can yield spaces that can yield these more radical changes to concepts formed against them.

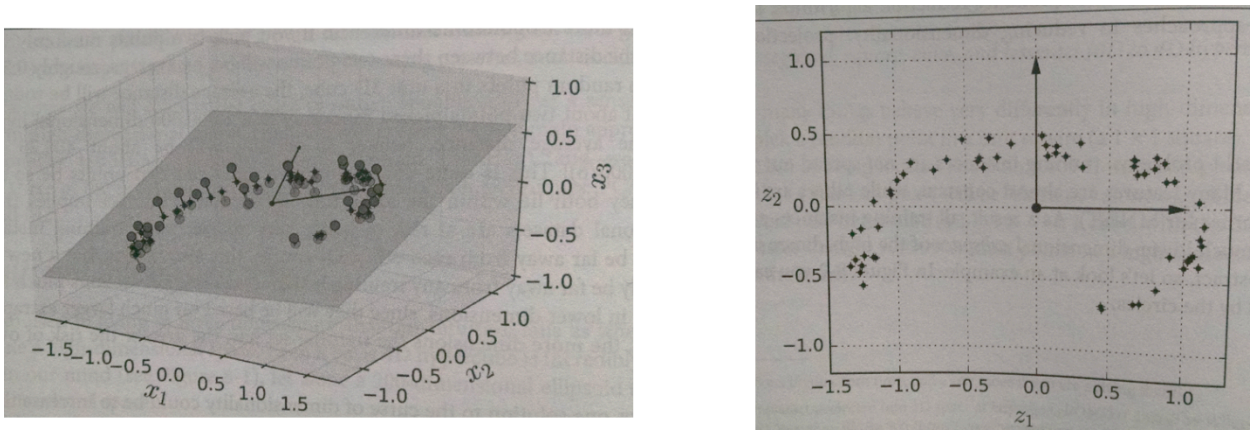

Figure 11 *– PCA used to reduce the 3-dimensional feature space on the left to the 2-dimensional feature space on the right. From Géron (2019).*

5.2. Non-linear dimension reduction techniques

Most astonishing here are the Deep Convolutional Neural Networks that have achieved superhuman performance in tasks like Go and image recognition; a helpful explanation of how these work is found in Buckner (2018). A simpler illustration, provided by Tenenbaum *et al.* (2000), will suffice for our purposes. Suppose that the central trend in the data lies on the "Swiss roll" manifold in Figure 13. This is a manifold "whose intrinsic geometry is that of a convex region of Euclidean space, but whose ambient geometry in the high-dimensional input space may be highly folded, twisted, or curved" (2321).[25]

[25] DiCarlo and colleagues (DiCarlo and Cox 2007, DiCarlo et al., 2012, 417) use a similar metaphor to describe the task of object recognition in naturalistic perceptual experience. Indeed, PCA-like dimensionality reduction has been implicated as the mechanism by which face and object recognition tasks are neurally executed (Tsao and Livingstone 2008) and amodal magnitude representations are formed (Bonn and Cantlon 2012, 161).

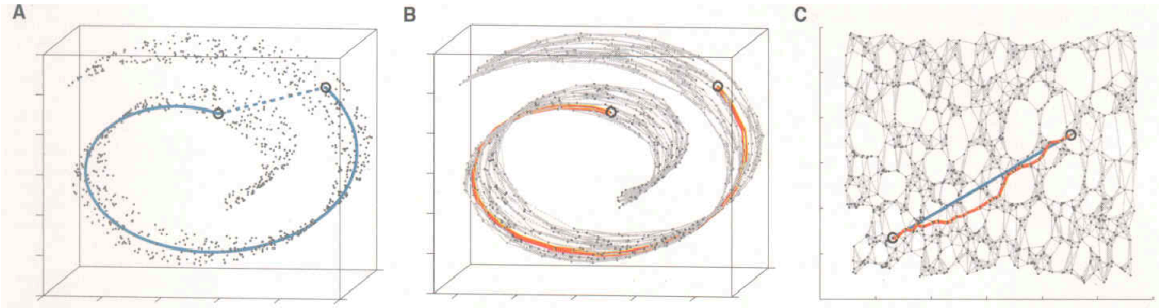

Fig. 3. The "Swiss roll" data set, illustrating how Isomap exploits geodesic paths for nonlinear dimensionality reduction. (**A**) For two arbitrary points (circled) on a nonlinear manifold, their Euclidean distance in the high-dimensional input space (length of dashed line) may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold (length of solid curve). (**B**) The neighborhood graph G constructed in step one of Isomap (with $K = 7$ and $N =$ 1000 data points) allows an approximation (red segments) to the true geodesic path to be computed efficiently in step two, as the shortest path in G. (**C**) The two-dimensional embedding recovered by Isomap in step three, which best preserves the shortest path distances in the neighborhood graph (overlaid). Straight lines in the embedding (blue) now represent simpler and cleaner approximations to the true geodesic paths than do the corresponding graph paths (red).

Figure 12 *– From Tenenbaum, et al. (2000, 2321).*

The Isomap technique for nonlinear dimensionality reduction collapses high dimensional data onto a lower dimension which captures the geodesic (not necessarily linear) line of greatest variance in the data. It unfurls the Swiss roll manifold, resulting in a two-dimensional space where Euclidean distance captures overall similarity. Concepts defined by this Euclidean distance in the new space will be incommensurate from those in the original space since "points far apart on the underlying manifold, as measured by their geodesic, or shortest path distances, may appear deceptively close in the high-dimensional input space, as measured by their straight-line Euclidean distance" (*ibid.*, 2319). Many convex groupings defined on the original curved space will not be convex on the transformed flat space; notice that in the above diagram, there are points that lie on the dotted line between the two highlighted points that no longer lie on a straight line between the two points in the transformed space.

5.3. Upshots for radical concept learning

Already, these dimension-reduction techniques illustrate several components of radical concept change. The geometry of a conceptual space embodies the system's notions of similarity; similarity is packed into the vehicle of representation itself. When the space is transformed, these similarity judgments change with it. Since similarity is the basis for future categorization behavior, the effects on

the system's future behavior can be profound. Further, we do not need to think about the new framework concepts as hypotheses that are tested. The principal components that serve as new framework concepts are learned from the data itself and need not have been represented prior to the learning process.

Dimension-reduction also suggests a framework for thinking about learning by analogy. Analogy involves the alignment of two distinct domains and the extraction of structural similarities between them, permitting the projection of features of one domain to the other. Dimension reduction takes this to the extreme, since the two domains are not just compared but are collapsed together into representation of that shared structure. The resulting similarity space can be more inductively rich than either of the distinct spaces that the learner began with.

6. Supervised learning processes

Dimension-reduction techniques like PCA distill regularities (relative to a conceptual space) already present in the data. To fully characterize radical bootstrapping, we need to understand how language can create new regularities. Here, we look at supervised learning processes that show how language can interface with conceptual spaces and how this permits even more radical conceptual change.

6.1. Learning from labels

As an unsupervised learning process, PCA does not utilize labeled data. More precisely, it does not use labels *as* labels; rather, labels are treated as just another feature. Linguistic input can play what Dove (2019, 9) calls a *scaffolding* role, helping "learners become attuned to perceptual commonalities and overcome the inherent complexity and noisiness of perceptual inputs". In contrast, in a supervised

process, labels serve as meta-features that designate points as members of certain categories, where these categories serve as "ground truths" for subsequent modeling (Danks 2014, 154).[26]

There are significant limitations to unsupervised processes. For example, PCA will find labeled categories only if the distinction between them happens to be the line of greatest overall variance. Compare PCA to a related supervised learning algorithm, Linear Discriminant Analysis, which finds the line of maximum distance between the means of two groups and reduces the dimensions to the orthogonal to this line:

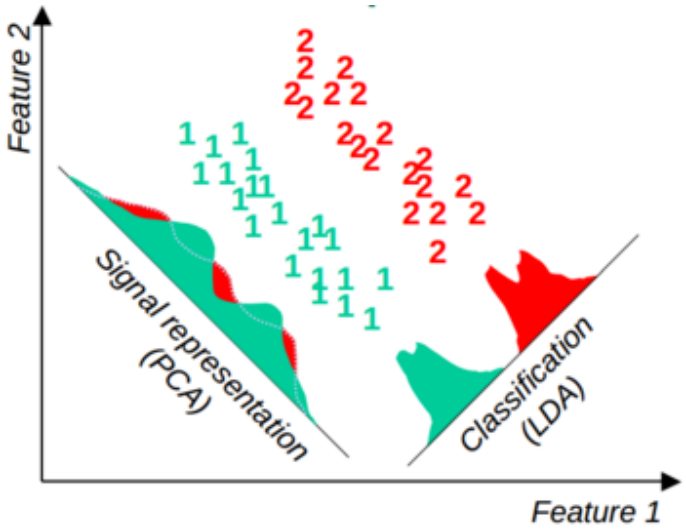

Figure 13 *–Comparison of results of PCA and LDA on the same data set. Figure from Malakar (2017).*

As Figure 13 illustrates, labels allow us to "break up" global trends in the data and find groupings that are more specific than the principal component, whereas unsupervised processes have to take all relevant features into account. This accords with data about human category learning in the presence or absence of labels. Adult humans performing categorization tasks without the aid of labels (aphasic subjects and subjects in verbal interference conditions) had difficulty learning "low-dimensional" categories, categories for which membership is based on single features like *green* or *square*. They had

[26] We do not give an account of how labels come to be attached to individuals or groupings. For an account of how language connects with perceptual similarity spaces, see Gauker (2011).

less difficulty with “high-dimensional” categories with more gradual and global membership conditions (Lupyan 2009, Lupyan and Mirman 2013, Dove 2019).

Labels also allow us to form groupings of perceptually heterogeneous objects. When labels are treated *as labels* rather than mere features, they can cause more radical permutations of underlying conceptual spaces. To explore these permutations, consider again the SVM task discussed in Section 4 (Figures 5 and 6). Since there is no hyperplane that divides the two categories, we need to transform the space to make it so divisible. The solution is to transform the space by adding new dimensions.

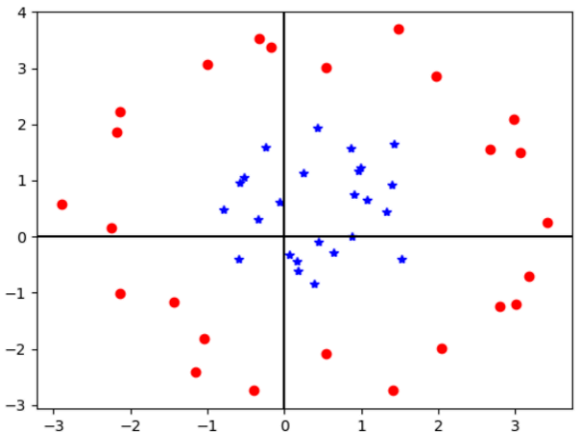

Figure 14 *– A supervised learning task where the groups are not linearly separable.*

There are two main procedures for dimension expansion: the first replaces an existing dimension with another (thus maintaining the overall dimensionality of the space), and the second introduces a hyper-dimension, which maintains the original dimensions and adds one on top. As we have seen, the first process might proceed by defining a new feature, $z = x^2 + y^2$ and replacing the existing y-axis with a z-axis so that the groups are linearly separable. This process may result in some loss of information about the original dimensions of the data. Notice that while there is a unique value of Z given the values of X and Y, one cannot necessarily reverse the process to uncover a unique function of X and Y.

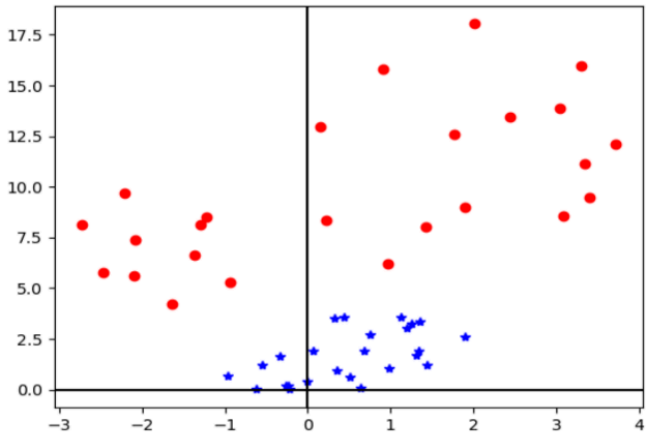

Figure 15 *– A transformation of the feature space in Figure 13 such that the two groups are now linearly separable.*

An alternative approach is to maintain the existing dimensions and add more; the observations are continuously projected into higher and higher dimensions until an n-1 dimensional hyperplane can linearly separate the now n-dimensional space. Consider the XOR plot in Figure 16, for which there is no possible linear separation. Instead of finding a concave separation rule, "the data in the XOR problem might be mapped into a three-dimensional space in such a way that each point F_1, F_2 is mapped onto F_1, F_2, F_3 where $F_3 = F_1 F_2$" (Harman and Kulkarni 2007, 43). In effect, this dimension transformation grabs the blue dots and pulls them up and grabs the green dots and pulls them down. Now, a hyperplane can separate the two groups in the new 3-dimensional space.

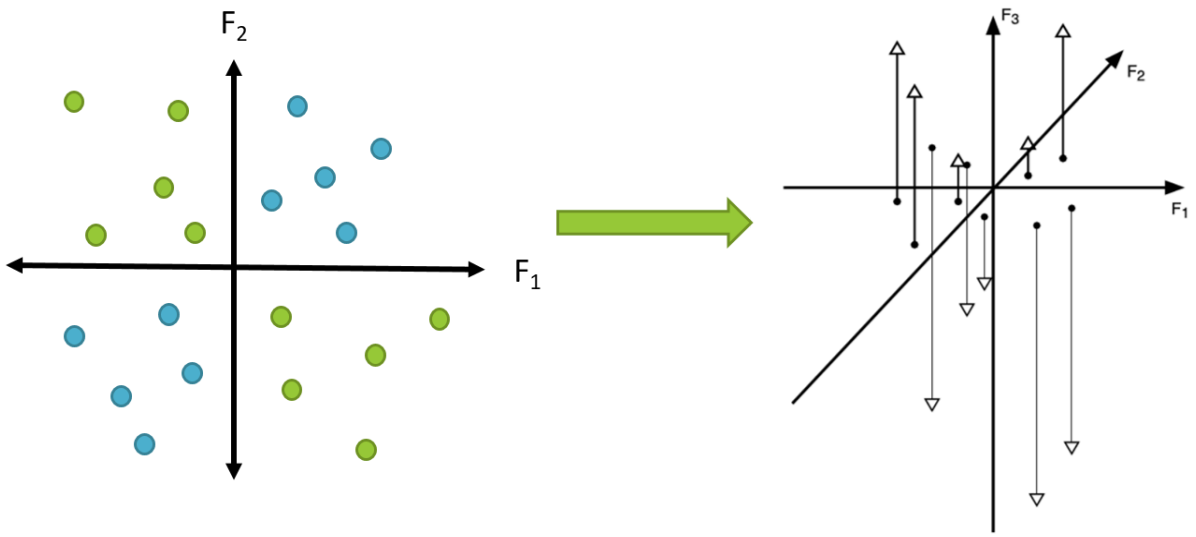

Figure 16 *– Left: 2-dimensional XOR plot. Right: projection of the data into a 3-dimensional space in which groups are linearly separable (reproduced from Harman and Kulkarni 2007, 43).*

6.2. How language transforms conceptual spaces

The learning process that we've been considering demands that we be able to linearly separate the two categories. As the above examples illustrate, this functions to pull categories apart from one another, turning fuzzy, overlapping categories into more categorical ones. This can have significant effects on subsequent categorizations.

First, it results in better categorization performance. Lupyan et al. (2007) conducted a landmark study on categorization of novel stimuli with nonsense labels. The authors trained subjects to associate two categories of "aliens" with either of two behavioral responses: approach or flee. The distinction between the two kinds of aliens was subtle; one kind had flatter bases with a ridge on the head, while the other had rounder bases with no ridge. There were two different training conditions, one with labels for the two categories of aliens ('leebish' and 'grecious'), and one with no labels. Subjects were given feedback after each answer. After training, subjects were tested on how well they learned the associations. The results showed that subjects in the label condition, despite having the same amount of experience with the stimuli as subjects in the no-label condition, were faster and more accurate in both

the training phase and at test. The authors concluded that having access to nonsense verbal labels enhanced performance compared to performance with nonverbal associations.

This effect is known as the label superiority effect[27] (Russell and Widen 2002). Lupyan et al. argue that the label superiority effect is explained by verbal labels modulating representational space:

> [R]ather than being fixed features, category names modulate item representations on-line through top-down feedback. According to this account, as a label is paired with individual exemplars, it becomes associated with features most reliably associated with the category. When activated, it then dynamically creates a more robust category attractor (2007, 1082).

In later work, Lupyan (2012) argues that the modulation of representational space consists in labels "pull[ing] apart [exemplar] representations [which results] in decreased representational overlap between the two classes of stimuli" (4).

Second, these new categorizations may come to serve as new framework concepts. Consider, for example, how one might start learning the concept MAMMAL. To start, suppose that animals are arranged by perceptual similarity. Then, one starts assigning the label "mammal" to various species. The resulting grouping will not be somewhat gerrymandered in the original space; for instance, the fish and the whales will be close together but labeled differently. As in the XOR example in Figure 16, we transform the space by adding a dimension in which the mammals are separated from the non-mammals. Indeed, that dimension can be interpreted as encoding the framework concept MAMMAL. Unlike the more continuous similarity dimensions in the initial space, it is a binary classifier. Finally, suppose (unrealistically) that MAMMAL became the only relevant dimension for categorizing animals. Now, you could use dimension reduction to collapse the space into a one-dimensional space that simply classifies individuals as mammal or non-mammal.

[27] We also find a label superiority effect in emotional recognition (Gentry forthcoming; Russell and Widen 2002; Widen 2013), taxonomic categorization (Markman 1990), and various other domains (Lupyan 2012; Lupyan and Lewis 2019; Lupyan et al. 2020).

For illustration, consider once more the results of the supervised LDA classification method, which finds the dimension of greatest separation between the two groups. We could take this line of greatest separation as our new y-axis and its orthogonal as our new x-axis. This would turn these groupings into a new framework concept, turning the distinction between 1-ness and 2-ness into a fundamental part of our ontology.

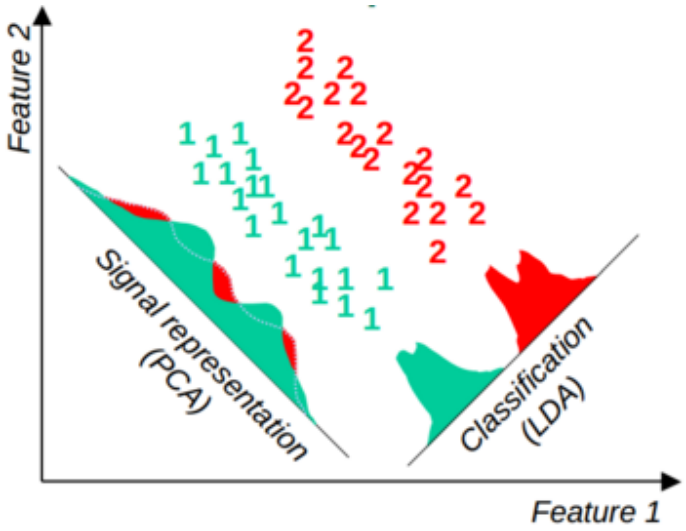

Figure 17 Comparison of results of PCA and LDA on the same data set. Figure from Malakar (2017).

There is evidence that category learning with labels can turn continuous concepts into discrete ones. Russian divides blue into two different categories (with no superordinate label corresponding to English "Blue")– "goluboy" for lighter blues and "siniy" for darker blues. In a now classic, cross-linguistic study on English and Russian color discrimination, Winawer et al. (2007) showed that in comparison with English speakers, Russian speakers were faster and more accurate at discriminating between darker and lighter shades of blue, but slower when the two shades were both dark or both light. Just like with

Lupyan et al.'s alien study, having access to labels for categories seems to facilitate performance by making the between-category differences more salient through warping of representational space. If labels work by warping representational space (by pulling between-category exemplars apart), and the color representational space is continuous, it's plausible that what was once continuous variation within the blue category, is now discrete.

Regier and Kay (2009) suggest that the initial color similarity space is warped (rather than fully continuous) but that this warping underdetermines the discrete categories that we seem to traffic in. Here, the authors suggest, is where linguistic intervention gets its hold. Learning color labels (in your native language) "finishes the job" by discretizing the already warped color space—by making continuous, albeit protruding, variation into discrete variation (see also Cibelli et al. 2016; Gleitman and Papafragou 2013; Holmes and Regier 2017; Regier and Xu 2017). Whatever the correct, precise account of color is, the moral of the story is that the data on labeling seems to suggest a role for language in shaping category representations. In particular, labels seem to pull apart category exemplars, descresing representational overlap in continuous space, and at least in some cases, this results in discretizing the categories—transforming representational space into discrete partitions.

6.3. Upshots for radical concept learning

In dimension expansion, a new framework concept is introduced to the fundamental ontology, spurred by an anomalous data set. Labels function as indications that disparate items nevertheless form a real kind; as Waxman puts it, "words are invitations to form categories" (1999, 269). If your data is telling you that there are two natural groupings here but you can't make sense of these groupings with your original conceptual space, you must come up with new framework concepts that do make sense of them. In effect, you ask, "if these are the natural kinds, what must the world be like?". Note that the choice of a new dimension is more unconstrained than in PCA-like dimension reduction; there are many

new features and dimensions you could add that would separate groups A and B. However, we do not need to think of these new dimensions as hypotheses that were represented prior to the learning process; they are instead generated from the data in the learning task at hand.

Compared to something like PCA, supervised learning processes allow for more flexible and radical transformations of the underlying space, as they are not beholden to global trends in the data. Instead, they treat labels as a "fixed" point and can amend the space to achieve a separation among any possible groupings one might encounter. The unlimited transformability yielded by supervised learning suggests one way that language enables humans' uniquely flexible cognition. While unsupervised dimension reduction techniques can find signals among the noise, supervised techniques can create genuinely new signals. Extended periods of cultural innovation have yielded new pieces of cognitive technology, linguistic constructions that exhibit kinds of structure very different from the ones given by our innate similarity space. If one's conceptual spaces are transformed in accordance with these new signals, then learning new categories through language can have drastic effects on one's subsequent cognition.

These supervised learning processes that transform conceptual spaces are a helpful illustration of the role of placeholders in linguistic bootstrapping. In Carey's view, linguistic placeholder structures "provide constraints, some only implicit and instantiated in the computations defined over the representations. These constraints are respected as much as possible in the course of the modeling activities, which include analogy construction" (2014, 152). To us, this sounds similar to a supervised learning process that takes labels to be constraints to be obeyed in permuting the conceptual space. The labeled groupings provide structure within a conceptual space, which spurs the creation of a new conceptual space that has that structure intrinsically.

7. Bootstrapping: the case of integer learning

We have presented a framework for thinking about radical concept change, involving transformations to conceptual spaces, and we have hinted at ways that this could be used to analyze aspects of linguistic bootstrapping. Here, we apply the framework to Carey's (2009) account of integer learning.[28]

According to Carey, core cognition starts with two systems for tracking numerosity. The object file system can track distinct objects, up to sets of 3 or 4. It is precise, but it has severe capacity limits. The analog magnitude system tracks the numerical size of sets ("how much"). It is subject to Weber's law, wherein the subject's ability to distinguish between two quantities depends on their ratio, not their absolute difference. For example, it is easier to discriminate between 5 and 6 than between 6 and 7 and far, far easier than between 25 and 26 (Dehaene 1997).[29] Initially, subjects have difficulty integrating the two systems in a single task (e.g. when choosing between two cookies and five cookies).

At the end of the learning process, the subject arrives at a concept of numbers that bridges these two systems: it tracks precise numbers like the object file system, and it can be extended beyond 4 like the analog magnitude system. It is characterized by the successor function: for any symbol in the numeral list that represents cardinal value n, the next symbol on the list represents cardinal value n+1.[30]

How do children learn this concept of numbers? Simplifying greatly[31], Carey's story goes like this. In the placeholder stage, children start by memorizing the count list ("one, two, three, four"). While the child can reproduce the order of the number labels, the list is initially meaningless. Next, the child starts to flesh out the placeholder, using two kinds of evidence. They receive information about various

[28] While this may not be the best illustration of our framework, it is the most detailed example of bootstrapping on offer.

[29] It is controversial whether the analog magnitude system represents magnitude linearly but with increasing error in discrimination or whether it represents magnitude logarithmically (Cantlon, *et al.* 2009, Dehaene *et al.* 2008). We'll return to this point in the next section.

[30] The successor function is entailed by the stable order principle, cardinality, and one-to-one correspondence of numerals to set sizes (Gelman and Gallistel 1978).

[31] For example, Carey details the development of natural language quantifiers ("a", "some", "many", etc.) and the support this provides in the child's numeral learning.

features of counting: the count list ends as you point to the final item in a set, you use one word for each object pointed to, etc. The child also learns to associate number words with sets of particular sizes (the "subset knower" stage). This process draws on the structure of the object file system. Between 24 and 30 months, children learn that "one" corresponds to sets of single objects.[32] They can pass Wynn's (1990, 1992) "Give 1" task, retrieving a single object if asked to give "1". If asked to give a number other than "1", they will bring more than one object, but their choice is at random; when asked to bring "2", they are equally likely to retrieve two, three, or four objects. This stage persists for six months to a year until children become "two knowers" who can discriminate "1" and "2" from each other and from other numerals but perform at chance for all other numerals. Likewise, they then become "three knowers" and remain so for some months.

This pattern changes dramatically around the age of 3 ½ after children have become "three knowers" (or occasionally, "four knowers"). To learn "5", children no longer need extensive experience of the contingency between "5" and sets of five objects. They spontaneously match higher numerals to the appropriate set size, as far as the child's knowledge of the numeral line extends. At this point, children have completed the last stage of bootstrapping, using learning processes (like analogy) to fully interpret the placeholder:

> The critical analogy that provides the key to understanding how the count list represents number is between order on the list and order in a series of sets related by an additional individual. This analogy supports the induction that any two successive numerals will refer to sets such that the numeral further along in the list picks out a set that is one greater than that earlier in the list. (Carey 2009, 477).

[32] This way of putting it undersells the complexity involved since they need to associate the numeral not with any particular set of a single item but singleton sets in general. As Beck describes it, "the meaning of the word 'one' could be subserved by a mental model of a set of a single individual {i}, along with a procedure that determines that the word 'one' can be applied to any set that can be put in 1-1 correspondence with this model" (2017, 476).

They now have number concepts that obey the successor function, in which the digitality of the object file system is extended to arbitrarily large numbers.

There are a few challenges for Carey's account. First, there is no specific learning process postulated for the crucial analogy. Second, absent this, the account is threatened by the circularity challenge. For instance, Rey (2014) argues that in the hypothesized step in which the child notices the analogy between one greater in the count list and one greater in set size, ''here 'is one greater than' expresses the very concept of SUCCESSOR whose acquisition Carey is trying to explain" (117).

Beck (2017) responds to Rey by arguing that the learning process exploits computational constraints that are implicit, rather than explicitly represented. During the subset knower phase, the child develops a procedure for putting "one" sets in one-to-one correspondence with models of a single item, "two" models of two items, etc., but the child need not represent what this procedure involves. However, at some point, "the child notices that when a collection with 'one' F is combined with another collection with 'one' F, the result is a collection with 'two' Fs; that when a collection with 'two' Fs is combined with another collection with 'one' F, the result is a collection with 'three' Fs" and so on (119). They reason that the labels in the counting list are also separated by this same kind of interval. The successor function emerges from the analogy process, rather than being explicitly present at the start.

8. A conceptual spaces account of bootstrapping the integers

Here, we explore how the conceptual spaces account could model this paradigm case of bootstrapping. To appreciate the way that this account will depart from the one above, notice that the analog magnitude system does not seem to play any significant role in the bootstrapping accounts of Carey (2009) and Beck (2017). The object file and parallel individuation mechanisms are doing all of the work. Numeral labels are associated with sets of individuals (as represented via this mechanism), and the ordering between set sizes/ successive numerals is computed by adding one individual object (as

represented via this mechanism). On the conceptual spaces account, the intrinsic ordering of the analog magnitude system plays much more of a role. Computational constraints are inherent in the geometry of conceptual spaces and the (geometric) category formation mechanisms that operate on them.

We suspect that the reason that Carey and Beck don't rely on the analog magnitude system is that its intrinsic ordering does not have the properties–namely, precision, discreteness, and linearity–of the numerical concept that is ultimately learned. How could it supply the properties that we're trying to explain? The answer, on the conceptual spaces account, is that the intrinsic ordering of the analog magnitude system becomes discrete and linear when associated with numerical labels (as discussed in Section 6).

Interestingly, this alternative path to integer learning is suggested by Carey's (2014) discussion of bootstrapping in animals, which draws on work from Livingstone, *et al.* (2009). They trained rhesus macaques on two initially distinct tasks. In the dot array task, they were rewarded for choosing the larger of two arrays of dots on a screen, with arrays ranging from 1-21, by receiving as many pulses of juice as there were dots in the larger array; e.g. if they chose 7 over 5 dots, they would receive 7 rather than 5 pulses of juice. Likewise, the monkeys were taught an arbitrary sequence of 21 symbols (1- 9, X, Y, W, C, H, U, T, F, K, L, N, R) and were rewarded for selecting the symbol that came later by receiving juice pulses corresponding to the symbol's position on the list; e.g. if they chose F over X, they would receive 17 rather than 10 pulses of juice.

Livingstone, *et al.* made three discoveries that will be important here. First, they found that the "monkeys could easily learn dot-array numerosity and abstract symbolic representations of surprisingly high numerosities when we directly associated the symbols or the numerosities with reward amounts" (*ibid.,* 713). Their ability to associate labels with set sizes was not constrained by the limits of the parallel individuation system, suggesting that learning these associations did not (solely) utilize that system.

Second, they found that when the monkeys were using the dot arrays to compare magnitudes, their choices showed scalar variability. This is what we would expect from the analog magnitude system. However, when the monkeys performed the task using symbols, "although they still tended to make more errors for small numerical divergences, their accuracy did not scale with choice magnitude (*ibid.*, 714). They displayed "linear discrimination when using symbolic representations (*ibid.,* 718).

Third, when the macaques were first asked to select between a set of dots and a symbol in the list - e.g. to pick between seven dots and the symbol W – they succeeded immediately. This suggests that the monkeys took both the dot arrays and the symbols to represent orderings of magnitude and could integrate them in a single task.

Putting this together, there is evidence that the monkeys (a) associated symbols with set sizes up to 21, (b) mapped the dots and numerals to the same scale and (c) that this scale was precise, linear, and discrete. Granted, when the monkeys used perceptually complicated dot arrays, they had trouble mapping particular sets of dots to this scale. Linear discrimination is easier when reasoning with labels because of the "equal distinguishability of one symbol from another" (*ibid.*, 719). This is precisely what the label superiority effect would predict. Labels pull apart different categories, making them easier to distinguish. This also turns continuously varying dimensions into discrete ones. In this case, numbers established a linear scale and provided the cognitive scaffolding to locate observed sets on this scale.[33]

On the conceptual space hypothesis, children (like monkeys) emerge from the subset knower stage having aligned the system of numeral labels ("one" - "two" - "three") with sets of increasing (analog) magnitudes. This alignment creates a dimension of numerosity that is linear and discrete. How?

[33] Exactly what the numeral symbols are doing here depends on what the starting state of the analog magnitude system is like (Cantlon, *et al.,* 2009, Dehaene, *et al.* 2008). On one hypothesis, it has a logarithmic scale, meaning that the similarity space itself represents 5 and 6 as further apart (more dissimilar) than 6 and 7 are. If that's correct, then the association between set sizes and labels must radically re-scale the magnitude dimension from a logarithmic to a linear one. On a second hypothesis, the analog magnitude scale is linear, meaning that the space represents 5 and 6 as the same distance apart (as dissimilar) as 6 and 7 are. It is the ability to accurately locate a set's position in this scale that decreases as the magnitude increases. Here, labels would simply have to increase the distinguishability of sets by creating a "more robust category attractor" (Lupyan, *et al.* 2007).

The discussion in Sections 5 and 6 is instructive. When two dimensions "march together", we can create a new dimension that characterizes this trend.[34] This dimension can take on the linear, discrete scale of the symbolic number line. As Dretske (1981, 215) puts it, when two properties are made equivalent to one another, "if a structure constitutes a complete digitalization of the one piece of information, it also constitutes a complete digitalization of the other".

How do we explain how the child extrapolates what is learned from the subset knower phase (of one through three)? We don't need to posit, as Carey and Beck do, that the child notices that you reach successive numbers by adding one individual to the previous number. Instead, this induction is supported by the intrinsic scale and directionality of the new number dimension. The line keeps going, step-wise and linearly, in the direction of greater and greater numbers.

We have not argued that the conceptual space hypothesis of integer learning is *in fact* how children learn the integers. As Carey (2009, 2014) is quick to point out, we need to look at the developmental data and the initial representational endowment to settle these questions. In this case, it's quite possible that human children learn number words in a way very different from macaques. Nevertheless, we think that it's instructive to draw out the differences between her account and one using conceptual spaces.

9. Conclusion

How is it possible to learn radically new concepts? Fodor's argument assumes that new experience can only confirm or recombine what you already know. We have argued that while experience does lead us to form new concepts against the background of one's fundamental ontology—the similarity spaces at the start of the learning process—experience can also reform that

[34] In dimension reduction, we found the line of greatest variance (the principal component) and then reduced the original two dimensions to this one. However, it's unclear whether the analog magnitude dimension is reduced to this new, linear, discrete dimension or whether we have an instance of dimension expansion where the precise number dimension exists alongside the earlier, imprecise one.

fundamental ontology. When faced with categories that are quite strange relative to what we think about the world, we can reconceive how nature must be such that these are its joints. The result is a terraforming of the conceptual landscape which fundamentally changes what will be built on top. Automating concept learning promises to demystify this process, showing how incommensurabilities can be crossed via small, yet far-reaching, steps.

References

Beck, Jacob. (2017) Can bootstrapping explain concept learning? Cognition 158: 110-121.

Berkeley, George. (1710/1975) A treatise concerning the principle of human knowledge. In Philosophical works, ed. M.R. Ayers. Rowman & Littlefield.

Blaise, M. P. (1976). Work of technical committee TC 1-6 “visual signalization” of the International Commission on Illumination (CIE). Bulletin de l'AISM, (66).

Block, Ned. (1986) Advertisement for a semantics for psychology. Midwest Studies in Philosophy X: 615–678.

Bonn, Cory D., & Cantlon, Jessica F. (2012) The origins and structure of quantitative concepts. Cognitive neuropsychology 29(1-2): 149-173.

Buckner, Cameron. (2015) A property cluster theory of cognition. Philosophical Psychology 28(3): 307-336.

-----. (2018) Empiricism without magic: Transformational abstraction in deep convolutional neural networks. Synthese 195(12): 5339-5372.

-----. (2024) From deep learning to rational machines. Oxford University Press.

Cantlon J, Cordes S, Libertus M, Brannon E (2009) Comment on “Log or Linear? Distinct intuitions of the number scale in Western and Amazonian Indigene Cultures”. Science 323:38b

Carey, Susan. (2009) The origins of concepts. Oxford University Press.

-----. (2011) Précis of the origin of concepts. Behavioral and Brain Sciences 34(3): 113.

-----. (2014) On learning new primitives in the language of thought: Reply to Rey. Mind & Language 29: 133–166.

Carnap, Rudolf. (1980) A basic system of inductive logic, part II. In: Studies in inductive logic and probability, 2: 7-155.

Cibelli, E., Xu, Y., Austerweil, J. L., Griffiths, T. L., & Regier, T. (2016). The Sapir-Whorf hypothesis and probabilistic inference: Evidence from the domain of color. *PloS one*, *11*(7), e0158725.

Churchland, P. M. (1995). The engine of reason, the seat of the soul: A philosophical journey into the brain. MIT Press.

-----. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. The Journal of Philosophy, 95(1), 5-32.

Dehaene, Stanislaus. (1997) The number sense: how the mind creates mathematics. Oxford University Press, NY

Dehaene S, Izard V, Spelke E, Pica P (2008) Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. Science 320:1217–1220

Danks, David. (2014) Learning. In: The Cambridge handbook of artificial intelligence, eds. K. Frankish & W.M. Ramsey [151-167]. Cambridge University Press.

DiCarlo, James J., & Cox, David D. (2007). Untangling invariant object recognition. Trends in cognitive sciences 11(8): 333-341.,

DiCarlo, James J., Zoccolan, Davide, & Rust, Nicole C. (2012). How does the brain solve visual object recognition?. Neuron 73(3): 415-434.

Dove, Guy. (2019) More than a scaffold: Language is a neuroenhancement. Cognitive neuropsychology: 1-24.

Dretske, F. (1981). Knowledge and the Flow of Information. MIT Press.

Eliasmith, Chris, & Thagard, Paul. (2001) Integrating structure and meaning: a distributed model of analogical mapping. Cognitive science 25: 245-286.

Fodor, Jerry A. (1975) The language of thought. Harvard University Press.

-----. (1981) The present status of the innateness controversy. In: Representations: Philosophical essays on the foundations of cognitive science [257–316]. MIT Press.

-----. (1990) A theory of content and other essays. MIT Press.

-----. (1998) Concepts: Where cognitive science went wrong. Oxford University Press.

-----. (2008) LOT 2: The language of thought revisited. Oxford University Press

Gallistel, Charles R. (1990) The organization of learning. MIT Press.

Gärdenfors, Peter. (1990) Induction, conceptual spaces, and AI. Philosophy of Science 57(1): 78-95.

-----. (2000) The geometry of thought. MIT Press.

-----. (2004) Conceptual spaces as a framework for knowledge representation. Mind and Matter 2(2): 9-27.

Gärdenfors, Peter. & Zenker, Frank. (2013) Theory change as dimensional change: Conceptual spaces applied to the dynamics of empirical theories. Synthese 190(6): 1039-1058.

Gauker, C. (2011). Words and images: An essay on the origin of ideas. OUP Oxford.

Gentner, Dedre. (2010) Bootstrapping the mind: Analogical processes and symbol systems. Cognitive Science 34(5): 752-775.

Gentner, Dedre, & Markman, Arthur B. (1997) Structure mapping in analogy and similarity. American psychologist 52(1): 45.

Gentry, Hunter. (forthcoming) Constructing embodied emotion with language: moebius syndrome and

face-based emotion recognition revisited. *Australasian Journal of Philosophy*.

Géron, Aurélien. (2019) Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Gleitman, L., & Papafragou, A. (2013). Relations between language and thought. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 504–523). Oxford University Press.

Goodman, Nelson. (1955) Fact, fiction and forecast. Harvard University Press.

Griffiths, Thomas L., & Tenenbaum, Joshua B. (2009) Theory-based causal induction. Psychological review 116(4): 661.

Hampton, James (2006) Concepts as prototypes. In: Psychology of Learning and Motivation (Volume 46), ed. B.H. Ross [79-113]. Academic Press.

Harman, Gilbert, & Kulkarni, Sanjeev. (2012) Reliable reasoning: Induction and statistical learning theory. MIT Press.

Holmes, K. J., & Regier, T. (2017). Categorical perception beyond the basic level: The case of warm and cool colors. *Cognitive science*, *41*(4), 1135-1147.

Holyoak, Keith J., & Thagard, Paul. (1995). Mental leaps: Analogy in creative thought. Cambridge, MA: MIT Press.

---- (1997) The analogical mind. American psychologist 52(1): 35-x.

Hume, David. (1739/1978) A Treatise of Human Nature. Oxford University Press.

Kuhn, Thomas S. (1962) Structure of scientific revolutions. University of Chicago Press.

-----. (1982) Commensurability, comparability, communicability. In: PSA: Proceedings of the biennial meeting of the Philosophy of Science Association [669-688]. Philosophy of Science Association.

Laurence, Stephen. & Margolis, Eric. (2012) Abstraction and the origin of general ideas. Philosophers' Imprint 12(19): 1-22.

Livingstone, Margaret S., Srihasam, Krishna & Morocz, Istavan A. (2009) The benefit of symbols: monkeys show linear, human-like, accuracy when using symbols to represent scalar value. Animal Cognition 13: 711–9.

Lupyan, Gary. (2009) Extracommunicative functions of language: Verbal interference causes selective categorization impairments. Psychonomic Bulletin & Review 16(4): 711–718.

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in psychology*, *3*, 54.

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319-1337.

Lupyan, Gary, & Mirman, Daniel. (2013) Linking language and categorization: Evidence from aphasia. Cortex, 49(5): 1187–1194.

Lupyan, G., Rahman, R. A., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in cognitive sciences*, *24*(11), 930-944.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. Psychological science, 18(12), 1077-1083.

Malakar, Gopal Prasad. [Gopal Prasad Malakar]. (2017, August 10). Linear Discriminant Analysis (LDA) vs Principal Component Analysis (PCA) [Video]. YouTube. https://www.youtube.com/watch?v=M4HpyJHPYBY

Margolis, Eric. (1998) How to acquire a concept. Mind & Language 13.3: 347–69.

Margolis, Eric. & Laurence, Stephen. (2013) In defense of nativism. Philosophical Studies 165(2): 693-718.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, 14(1), 57–77.

Maudlin, Tim. (2012) Philosophy of physics: Space and time (Vol. 5). Princeton University Press.

Povinelli, Daniel, & Ballew, Nicholas G. (2012) World without weight: Perspectives on an alien mind.

Oxford University Press.

Quine, Willard Van Orman. (1969). Natural kinds. In *Ontological Relativity & Other Essays*. New York: Columbia University Press.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. Trends in cognitive sciences, 13(10), 439-446.

Regier, T., & Xu, Y. (2017). The Sapir-Whorf hypothesis and inference under uncertainty. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(6), e1440.

Rey, Georges (2014) Innate and learned: Carey, mad dog nativism, and the poverty of stimuli and analogies (yet again). Mind & Language 29: 109–132.

Russell, J., & Widen, S. (2002) A label superiority effect in children's categorization of facial expressions. *Social Development* 11(1), 30-52.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237(4820), 1317-1323.

Strevens, Michael. (2012) Theoretical terms without analytic truths. Philosophical Studies 160: 167-190.

Tenenbaum, Joshua B., De Silva, Vin, & Langford, John C. (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500): 2319-2323.

Tsao, Doris Y., & Livingstone, Margaret S. (2008). Mechanisms of face perception. Annual Review of Neuroscience 31: 411-437.

Waxman, Sandra R. (1999) The dubbing ceremony revisited: Object naming and categorization in infancy, early childhood. In Folkbiology, eds. D. L. Medin & S. Atran [233–284]. MIT Press.

Weiskopf, Daniel. (2008) The origins of concepts. Philosophical Studies 140: 359–384.

-----. (2011) Models and mechanisms in psychological explanation. Synthese 183: 313–338.

Widen, Sherri. (2013) Children's interpretation of facial expressions: the long path from valence-based to specific discrete categories. *Emotion Review* 5(1), 72-77.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, *104*(19), 7780-7785.

Woodward, James. (2007) Interventionist theories of causation in psychological perspective. In: Causal learning: Psychology, philosophy, and computation, eds. A. Gopnik & L. Schulz [19-36]. Oxford University Press.

Wynn, K. (1990). Children's understanding of counting. Cognition, 36, 155-193.

---- (1992). Children's acquisition of the number words and the counting system. Cognitive Psychology, 24, 220-251.