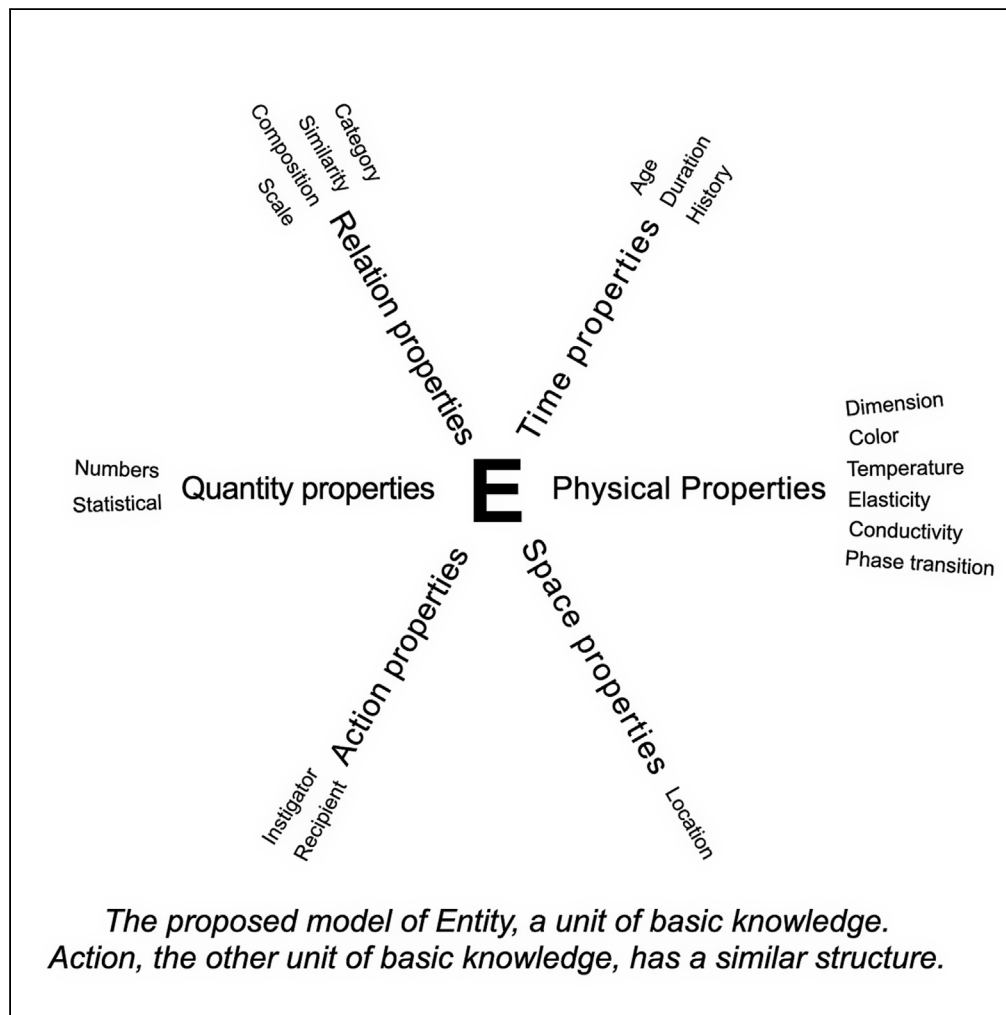


## Article

## Patterns of basic knowledge



Hung Tseng

htsengpe@gmail.com

**Highlights**

A framework was proposed to investigate the basic structure of knowledge

Theories in the fields of linguistics, database, cognition, and statistics are examined

A hypothesis is proposed that basic knowledge consists of two descriptive systems

Potential implication and usage of this framework are discussed

## Article

## Patterns of basic knowledge

Hung Tseng<sup>1,2,\*</sup>

## SUMMARY

**Epistemetrics, a discipline concerning measuring knowledge qualitatively and quantitatively, is an under developed area of research. A primary reason for the current status is a perceived lack of a formal structure of knowledge. In particular, a number of fundamental questions about knowledge remain to be answered: What is the basic unit of knowledge? How many types of knowledge are there? How do they relate to each other? And what are the hierarchies of knowledge? To address these questions, we proposed a knowledge categorization system, the EApc framework. The present work collects further evidence, from theories and practices of several diverse research disciplines, to argue for, to refine, and to expand the EApc framework. This framework can provide new perspectives in understanding scientific processes and in assessing values and impacts of research proposals and products. It may also shed light on the mechanisms of human cognition and information processing.**

## INTRODUCTION

Measurement of knowledge has been a pursue of philosophers for centuries, and more recently, information scientists joined the effort. However, the term *epistemetrics* was only coined recently by Nicholas Rescher, who declared it as a research discipline still in its infancy (Rescher, 2006). In the research field of scientometrics, a sub-discipline in epistemetrics according to Rescher, the current prevailing method of measuring knowledge quantity is counting the number of research papers and scholarly books, which are not knowledge in its purest form but a package of knowledge of different kinds and combinations and the methods of obtaining them. Categorization/classification of such knowledge packages has been mainly a concern of library science, e.g., Dahlberg (1982). Apparently, this method of counting knowledge as packages (i.e., research papers, books, and other forms) offers an imprecise way of assessing qualitatively and quantitatively the knowledge being produced or communicated. An analogy to this situation would be estimating the sales and inventory of a supermarket by counting the bags of groceries sold. This analogy raises a number of key questions regarding analysis of knowledge output and accumulation: What is the basic unit of knowledge? How many types of knowledge are there? How do they relate to each other? And what are the hierarchies of knowledge? In other words, more succinctly, what are the patterns of knowledge? Conceivably, correct answers to these questions afford new perspectives in understanding knowledge discovery, accumulation, and manipulation processes.

We attempted to address these questions by describing a knowledge framework termed EApc (Tseng and Small, 2019). The EApc framework aims to introduce basic categories and structures into an otherwise amorphous and abstract concept of knowledge. According to the EApc framework, knowledge can be classified into two categories: (1) basic knowledge and (2) compound knowledge, where basic knowledge is discovered and compound knowledge is synthesized from basic knowledge. Basic knowledge is divided into four subcategories: Entity, Action, Property, and Condition, hence the term EApc, which are defined as:

Entity: An object, physical or mental.

Property: Attributes of an entity.

Action: A process (physical or mental) required to change an entity or its property.

Condition: Attributes of an action.

Compound knowledge comprises multiple entities and actions that are linked together based on their properties and conditions. Compound knowledge constitutes our explanation, prediction, and

111021B Villlaridge Court,  
Reston, VA 20191, USA

<sup>2</sup>Lead contact

\*Correspondence:

htsengpe@gmail.com

<https://doi.org/10.1016/j.isci.2022.103803>



manipulation of the natural world. Accordingly, the knowledge gathering process, i.e., scientific research, can be classified as 1) discovering entities and associated properties, and actions and associated conditions; and 2) formulating theories (explanations and predictions) and inventions (manipulations).

(Note: The EApc acronym was first coined in [Tseng and Small 2019](#). While to some the EApc may not be the best acronym for the framework (e.g., EpAc may be better), the concern of switching the acronym here in the second paper about this framework is that it will cause confusion in tracking the development of the framework, e.g., keyword search and inconsistency in citing the framework, i.e., using EApc and EpAc interchangeably).

In order for the EApc framework to be a useful tool for understanding knowledge generating processes, we tested its ability to quantify knowledge required for the invention of the recombinant DNA technology, and to describe the knowledge accumulation process leading to that invention ([Tseng and Small, 2019](#)). We achieved a higher precision in measuring knowledge content and a higher clarity in tracing the origin of these knowledge than the methods using publications as the unit of knowledge. However, that study was limited to the field of molecular biology and biotechnology. For a broader foundation, utility, and impact of the framework, more research disciplines need to be examined. Here, I provide further supporting observations collected from several apparently disparate research disciplines: linguistic universals, database theories, concepts of emergence, meso-science, metamorphic cognition, and statistical causal inference. Consequently, the *Basic Knowledge* category of the EApc framework is refined and several new implications of the framework are discovered and discussed.

It should be noted that the EApc framework does not concern the ultimate validity of a given piece of knowledge (i.e., true or false), but that how is knowledge structured. The EApc framework assumes all knowledge is tentative and evolving, which means at a given time, it contains inaccurate information and is not complete ([Tseng and Small, 2019](#)).

For clarity, the following convention is adopted in this work regarding usage of upper case for the first letter in entity, action, property, and condition. When the word refers to a knowledge category, the uppercase letter is used; when the word refers to an individual object or process, the lower case is used, unless dictated otherwise by grammatical rules (e.g., at the beginning of a sentence).

## RESULTS

### Linguistics and EApc

#### *Knowledge and language*

Natural philosophers had noted for centuries a connection between knowledge and language, that language is the carrier of knowledge, that knowledge is processed, stored, and communicated through the use of language. As cited by [Rescher \(2006, p 75\)](#), the German philosopher G. W. Leibniz wrote: “All items of human knowledge can be expressed by the letters of the alphabet ... so that consequently one can calculate the number of truths of which humans are capable and thus compute the size of a work that would contain all possible human knowledge, ...” This notion was echoed in Darwin’s writing that “A long and complex train of thought can no more be carried on without the aid of words, whether spoken or silent, than a long calculation without the use of figures or algebra.” ([Darwin, 1871, p 57](#)) More recently, Noam Chomsky, an American linguist, cognitive scientist, and philosopher, put the relation in terms of human cognition: “There is an innate structure that determines the framework within which thought and language develop, down to the quite precise and intricate details. Language and thought are awakened in the mind and follow a largely predetermined course, much like other biological properties” ([Chomsky, 1997](#)); and “The core of language appears to be a system of thought, with externalization a secondary process ...” ([Chomsky, 2011](#)). These remarks and many others by the philosophical and scientific forebears and contemporaries suggest that certain fundamentals of knowledge, i.e., its building blocks and their organization, may be understood from principles of linguistics.

[Morgan et al. \(2015\)](#) described an experimental study of language usage in passing tool-making skills from one tool maker to another. This study suggests that knowledge-related language development might have evolved in parallel with tool-making ability in the early human evolution. However, a challenge of using linguistic findings to study the basic unit of knowledge is that natural human language capacity is not developed and evolved solely for the purpose of knowledge communication and processing. Early usage of

**Table 1. A comparison of the definitions of elements in EApc knowledge categorization and linguistic lexicon universals**

	EApc framework		Linguistic theories
Entity	An object, physical or mental	Noun	Semantic definition: nouns and only nouns have criteria of identity, whereby, they serve as standards of sameness (Baker 2003 p 95).
Property	Attributes of an entity.	Adjective	The most obvious distinctive characterization of adjectives is that they modify nouns directly, in the so-called attributive construction. Nouns and verbs cannot do this. (Baker 2003 p 192)
Action	A process, required to change an entity or its property.	Verb	A verb's meaning encodes a specific conceptualization of an event. (Levin 2009 <a href="https://web.stanford.edu/~bclevin/lisa09intro-sr.pdf">https://web.stanford.edu/~bclevin/lisa09intro-sr.pdf</a> ) Verbs are generally recognized by linguists as quintessential predicates (Baker 2003 p 23).
Condition	Attributes of an action	Adverb	Adverbs and adjectives are in the same category (Baker 2003). Semantic roles of verb (Fillmore, 1968)

prototypic language in animals and humans are likely related to simple communication needs in searching for food, warning community members of imminent danger, organizing and coordinating activities within and among individuals and groups, and expressing emotions and feelings to and about others and more (e.g., Barsh et al., 2019 and references therein). Thus, there is considerable “noise” in linguistic findings regarding revealing principles of basic knowledge structure and pattern. It appears that the most relevant information resides in lexical (word) and semantic (meaning) analysis (vs that of morphology and phonetics). As to compound knowledge, the analysis of syntax and grammar may be relevant and the topic deserves a separate study.

### Linguistic universals

Reviewing the latest research in linguistics led to the belief that the categorization of basic knowledge in the EApc framework can be deduced, to a large extent, from a number of linguistic concepts. In particular, the EApc framework is consistent and supported by the concept of linguistic universal, a branch of linguistic studies motivated by Chomsky's hypothesis that in all humans, there exists an innate mechanism, which facilitates language acquisition at the early age (Chomsky, 1957). This hypothesis triggered a search of commonalities in languages around world, resulting in identification of a number of language universals, i.e., linguistic elements and rules (grammar) present in virtually all human languages (Comrie, 1989). While most of these universals are in the realm of syntax (word orders in sentences), an observation in lexicon categorization is of particular interest in answering the question of the basic categories of knowledge. It has been observed that most languages in the world have the lexicon categories of nouns, verbs, and adjectives; for a few languages that may not have the apparent form of one of these categories, they have functional substitutes (Baker, 2003, pp 88, 169). This observation implies that nouns, verbs, and adjectives are basic units of human language (lexicon), hence the means, by which knowledge has to be expressed in, and probably processed with as well. On the surface, these basic lexicon categories impose a constraint on how knowledge can be communicated. Deep down, they are windows through which one may glimpse the information processing apparatus in the brain. In any case, the basic categories of knowledge (communicable) are likely related to these lexicon categories, therefore can be classified similarly (Table 1).

### Nouns (linguistics) and entities (EApc framework)

Mark C. Baker, an American linguist, gives the following semantic definition of Nouns (pp 95–96, references therein, Baker, 2003) “Nouns and only nouns have criteria of identity, whereby they can serve as standards of sameness... The idea in a nutshell is that only common nouns have a component of meaning that makes it legitimate to ask whether some X is the same (whatever) as Y. This lexical semantic property is the pre-condition that makes nouns particularly suited to the job of referring, since it is fundamental for reference to be able to keep designating the same entity over and over again (Wiggins, 1980 ch1).” This standard of sameness was considered by William James, an eminent American psychologist, to be the “very keel

**Table 2. A comparison of the semantic classes of Dixon, Aristotle's category of knowledge, and the EApc basic knowledge categories**

Dixon's classification of Adjectives	Aristotle's category of knowledge	EApc basic knowledge categories
dimension (big, short, etc.)	Quantity (how much, four-foot, five-foot)	Property
physical properties (strong, ill, etc.)	Quality (what sort, white, literate)	Property
speed (fast, quick, etc.)		property/condition
age (new, old, etc.)	Time (when, yesterday, last year)	property/condition
color (red, black, etc.)	Quality (what sort, white, literate)	Property
value (good, bad, etc.)		N/A: not a physical property
difficulty (easy, difficult, etc.)		N/A: not a physical property
quantification (probably, possible, usual, likely, sure, appropriate)	Quantity (how much, four-foot, five-foot)	property
human propensity (fond, angry, jealous, anxious, happy, certain, eager, ready, clever, stupid, generous)		N/A: not a physical property
similarity (similar, different, etc.)	Relation (related to what, double, half, greater)	property
	Substance (substance, man, horse)	Entity
	Position (being situated, lies, sits)	property/condition
	Habit (having, possession, is shod, is armed)	property
	Action (doing, cuts, burns)	Action
	Passion (undergoing, is cut, is burned)	Action
	Location (where, in the lyceum, in the marketplace)	property/condition

and backbone of our thinking" (p 459, [James, 1890](#)). Baker's semantic definition is consistent with the requirement of Entity in the EApc framework, in which an entity must possess features (criteria) to be distinguished from other entities, i.e., same or different. The EApc framework provides a means as to how the distinction can be achieved. This is because an entity in the EApc framework has properties associated with it, and in that sense, an entity is the name of a unique collection of properties of that entity. By comparing the properties associated with two entities, one can determine if they are the same (i.e., properties overlap entirely) or different (i.e., possessing distinctive properties). Therefore, Entity category is indispensable for knowledge representation and individual entities can be viewed as discrete, independent, countable unit of knowledge, just as nouns can be used as a unit in the linguistic analysis ([Tseng and Small 2019](#)).

#### *Adjectives (linguistics) and properties (EApc framework)*

Linguists noted that a salient characteristic of adjectives is that they modify nouns in what is termed attribute construction (p 192, [Baker, 2003](#)), which shows its necessary association with nouns and its function of describing properties of the thing referred to by the noun. This concept is also made amply clear by [Raskin et al. \(1995\)](#) "A French Dictionnaire de linguistique (1973), a very appropriate reference for conventional wisdom, defines adjectives as 'words joint to the nouns to express a quality of the object, creature, or concept designated by this noun.' Greville (1969: 285) adds 'a property of being' to the list of things designated by nouns." These linguistic definitions are consistent with the notion of the EApc framework that properties adhere to entities. We have previously suspected that the Property (as well as Condition) category contains subclasses ([Tseng and Small, 2019](#)). Looking into linguistic literature, Dixon, among others, had noted that Adjectives consist of at least ten semantic classes ([Dixon, 1991](#)). It is of interest to note that the adjective classes in Dixon's list overlap substantially with Aristotle's knowledge categories described by [Smith \(2018\)](#) (Table 2). About half of the adjective semantic classes can be assigned to

Aristotle's categories, whereas the other half are attributes, such as value, difficulty, and human propensity, which are not intrinsic physical properties (Raskin et al., 1995). These non-physical properties in adjective classification are expected because human language has functions other than knowledge communication and processing (see above). It should also be noted that in knowledge communication, the description of property is not necessarily a single word or words (phrase), such as color (e.g., yellow) or temperature (e.g., cold), but can be a sentence or several sentences (for examples, see Tseng and Small 2019 for the description of properties associated with entities in the invention of recombinant DNA technology). This is probably in part due to operational definitions widely used in scientific communication (e.g., relates to methods used in observing the properties), as well as the fact that observed properties are emerging more rapidly than language evolution (or conceptualization) can invent/generate more succinct symbolic notations (i.e., unified nomenclatures).

### Verbs (linguistics) and actions (EApc framework)

Linguist Beth Levin once defined Verb semantically: "A verb's meaning encodes a specific conceptualization of an event" (Levin, 2009), which is like the EApc definition of Action as a process. From a syntactic perspective, Baker defines "Verbs as licensors of subjects. Verbs are generally recognized by linguists as quintessential predicates" (Baker, 2003, p 23). Based on this definition, he argued that verb performs a function that cannot be substituted by nouns or adjectives, thus there is a natural linguistic boundary or function that separates verb and the other lexicon categories. Furthermore, recent neurolinguistic brain imaging studies showed when learning nouns and verbs, non-overlapping brain regions were activated, providing neurological evidence for the noun/verb separation (Mestres-Missé et al., 2009). Our previous classification of Action as a basic category was based on 1) Aristotle's knowledge classification contains an item termed "motion"; and 2) in mathematics and computer languages, the concept of "variable" (analogous to Entity/Property) and that of "operation" (analogous to Action) belong to separate concept categories. We therefore hypothesized that the division of Entity and Action might be a true cognitive separation (Tseng and Small, 2019). Linguistic research into the lexicon universals and brain-imaging studies appear to support this hypothesis. In this sense, Action category, just as Entity category, is indispensable for knowledge representation and individual actions appear to be also discrete, independent, countable units of knowledge just as entities.

### Adverbs (linguistics) and conditions (EApc framework)

While adverbs are generally not mentioned as a universal lexicon category, Baker argued that adjectives and adverbs are in the same category (Baker, 2003, p 191), pointing out that morphologically, adverbs are created by adding -ly at the end of adjectives (in English). It has been proposed in the lexical semantic approach to understand verbs that: "The meaning of a verb can be represented by—and, thus, reduced to—a list of semantic role labels, each identifying the semantic relation an argument bears to it." Such labels are agent, counter-agent, object, result, instrument, source, goal..., etc. (i.e., the semantic roles concept, Fillmore, 1968, p 77), which correspond well to the notion of Conditions in the EApc framework, e.g., a biochemical reaction (Action) can be characterized by its substrates, products, catalysts, inhibitors, cofactors, stoichiometry, reaction speed, and reaction requirements such as ionic strength, pH, temperature, etc (Table 3). The EApc framework views the relation of Condition to Action as parallel to that of Property to Entity, in that an action is the embodiment of a collection of conditions, analogous to the notion of reducing the meaning of a verb to a list of semantic role labels (Levin, 2009) (Table 3). In other words, as entities can be distinguished by their properties, actions can also be distinguished by their conditions.

### Section summary

These analyses suggest a correspondence between the linguistic concepts of noun, verb, adjective, and adverb and that of EApc's entity, action, property, and condition. If this correspondence indeed exists, then categorization (demarcation) of basic knowledge into the four types can be argued on a more universal ground, rooting in the cognitive and communicative processes in humans. And individual entities and actions can be viewed as discrete and independent units of knowledge and analyzed accordingly (e.g., in Tseng and Small 2019).

### Database theories and EApc

Understanding the basic unit of knowledge and its organization can also be approached by examining the contemporary theories of databases, which have been developed since around 1970, as computer technologies matured and their applications broadened. It is not surprising that database theories are highly

**Table 3. A comparison of linguistic semantic role labels and the concept of Condition in EApc**

Linguistic semantic role labels of verb (Fillmore 1968)	Conditions of Action in EApc (A biochemical example)
Agent (A), the instigator of the event	A set of conditions that initiate a reaction
Counter-Agent (C), the force or resistance against which the action is carried out	Inhibitors (inhibitory environment) of a reaction
Object (O), the entity that moves or changes or whose position or existence is in consideration	An entity in a transportation reaction/event
Result (R), the entity that comes into existence as a result of the action	Products of a reaction
Instrument (I), the stimulus or immediate physical cause of an event	Catalysts of a reaction
Source (S), the place from which something moves	An initial location of a transportation event
Goal (G), the place to which something moves	An end location of a transportation event
Experiencer (E), the entity which receives or accepts or experiences or undergoes the effect of an action	Substrates of a reaction

relevant to the description of a knowledge framework because the primary concern of constructing a database relates to how to structure information units and their relationships. Many of the challenges encountered in database design may also be issues faced in the evolution of human brain, although the implementation (embodiment) of solutions, i.e., electronic vs neurological circuitry, can be entirely different. Still, on an abstract level, some similarities may be found (Feldman, 2008). The approach taken here is to compare the principal elements of database theories and that of the EApc framework to establish a mapping of the concepts in each framework, so that the EApc framework can benefit from the existing theories and the tools associated with them.

The earlier database theories, e.g., Codd and Jose (1970) and Chen (1976), were developed around databases used in business, organizing information for payrolls and inventories. A closer examination found that their elements, e.g., entities, attributes, and values, are easily translatable to the elements in the EApc framework for knowledge categorization. The comparison here focuses on Chen's Entity-Relationship (E-R) Model (Chen, 1976) because it generalizes, extends, and encompasses the three other database models (i.e., the network model, the relationship model, and the entity model) present at the time, and these prior models are derivable from the E-R model. Another virtue of Chen's model is its explicit cognitive consideration that "This model incorporates some of the important semantic information about the real world" and it concerns "entities and relationships which exist in our minds" (Chen, 1976).

### Comparison of E-R model and EApc framework

Overall, a basic commonality of the two frameworks is that both can be represented as a table, with elements that are similarly defined. The table format was used extensively, in Codd and Jose (1970), and Chen (1976), to demonstrate the concept of the relation model of data. Although in the initial description of the EApc framework, table format was not used, the conceptual elements can be similarly displayed. Thus, at an abstract level, their elements are interchangeable (Table 4 and analysis below). A notable difference between the two systems lies in the fact that the database model deals with objects (entities) and related attributes (properties) whose existence are relatively certain (e.g., persons in payroll and their age, gender etc.), whereas the EApc framework are proposed to deal with knowledge whose physical existence and properties are often under investigation (being conceptualized). Therefore, in the EApc framework, the entities, actions, properties, and conditions can be hypothetical (mental), pending to be confirmed or rejected by observation and experimentation. This temporary nature of elements in the EApc framework, however, should not affect their representation and processing as knowledge internally and externally. In fact, a valued feature of the E-R model in database design is its "data independence", meaning that revision of stored data (addition or deletion or replacement of information) will not affect their retrieval or processing. Conceivably, such data independence may also be a feature of the mind, because new

**Table 4. A comparison of concepts in the Entity-Relationship (E-R) model of database and EApc Framework in knowledge categorization**

Elements in Entity-Relationship Model	Entity-Relationship Model	EApc Framework
Entity	A thing which can be distinctly identified	An object, physical or mental
Attributes	Information associated with an entity which can be obtained by observation or measurement	Property: Attributes of an entity
Value	A way attributes are expressed Value set is a classification of attributes	A classification of properties (Type of properties) Values are not considered in EApc
Relationship	Association of entities that can be expressed as a mathematical set	A property of an entity: how entities relate to each other
Key	A group of attributes such that the mapping from the entity set to the corresponding group value sets is one-to-one	Properties define an entity and its relationship with other entities

experiences (learning) will certainly enrich, expand, and update information stored in the brain. [Table 4](#) compares the definitions of the elements in E-R and EApc side-by-side and the comparison is analyzed and annotated in the following sections.

**Entities.** The definitions of entity in E-R and EApc ([Table 4](#)) are compatible and complementary, emphasizing different perspectives. As mentioned above, in the EApc, entities can be hypothetical, i.e., only exist in the mind. In the E-R, Entity is treated as a stand-alone unit of information, supporting the idea of using entity as a unit of knowledge.

**Attributes.** Similarly, the definitions of attributes (E-R) and properties (EApc) are virtually identical ([Table 4](#)). The E-R definition has two virtues that 1) it states explicitly the nature of attributes as information associated with entities, and 2) it states the source of this information as from observation or measurement, i.e., from experience. In the earlier rendition of the EApc framework ([Tseng and Small, 2019](#)), we similarly stated that all basic knowledge categories (entities, actions, properties, and conditions) are derived from observation (discovered). This view has in mind the scientific discovery process, i.e., from unknown to known, as reflected in insisting that an entity can transition from mental to physical, i.e., a match of the mental model and the physical world, which may not be as important in database design.

**Value.** Value here refers to the mathematical definition that a variable can take a range of states, discrete or continuous. Obviously, the concept of value plays an essential role in database because that is what the users search for. The initial rendition of the EApc framework did not consider the concept of value, but instead focused on a general conceptualization/categorization scheme of knowledge. We did, however, recognize that the concept of property in EApc is complex and deserve further exploration. Examining the database theory does yield new insight of how properties can be viewed as a set consisting of elements, which are associated with different entities. For instance, the property Color consists of values such as blue, green, red, etc., and a blue object can be described as an entity whose color property takes the value of "blue". We have incorporated the value concept in the revised version of the EApc proposal (see later).

**Relationship.** The E-R theory considers relation as a separate concept from attributes, whereas the EApc framework views relation as a type of property. The E-R sees relation as a mathematical set, and the EApc properties can also be seen as sets (see [value](#)). Thus, there appears no fundamental discrepancy but the difference in conceptualization may merely reflect a variation in perspectives. The EApc framework sees a relationship between two entities as defined by the properties associated with the entities, e.g., the properties they share or do not share determine if they belong to a group (set) or not, such as biological species are linked in a phylogenetic tree graph, or clustering of papers in studying bibliographical works. The relation between two entities can also be defined by their role (a condition) in an EApc action, i.e., a doer or a receiver of an action, or as referred in verb semantic roles, an agent and an object. The



essence of such concept can be glimpsed in the early writing of Ludwig Wittgenstein, an influential Austrian philosopher in the 20th century, “*Either a thing has properties which no other has, and then one can distinguish it straight away from the others by a description and refer to it; or on the other hand, there are several things which have the totality of their properties in common, and then it is quite impossible to point to any of them...*” (TLP 2.02331, 1922 English translation), which is another way of stating the “sameness” concept mentioned earlier (sections [nouns \(linguistics\) and entities \(EApc framework\)](#) and [adjectives \(linguistics\) and properties \(EApc framework\)](#) also see section [key](#) below).

**Key.** The E-R theory uses “key” to refer to a set of attributes that defines the identity of an entity and further refines this usage as *super key* (A set of attributes that collectively identifies an entity), *candidate key* (A minimal super key), and *primary key* (one of the candidate keys chosen by the database designer to uniquely identify the entity set). The notion that entities are identified by their properties is shared by the EApc framework and also relates to the linguistic concept of nouns serving a function as the standard of “sameness.” In scientific research, the question of sameness is an everlasting challenge in dealing with complex entities and actions, such as in the classification of biological species and diseases, respectively. The debate is often centered around what “candidate key” or “primary key” should be applied, whether gross morphology/clinical symptoms vs molecular features (e.g., DNA sequence and genetic mutations) should take priority as revealing the “sameness” of a species or disease.

### Databases that fit the EApc categorization of knowledge

To illustrate that information listed in databases can be classified in the EApc basic knowledge categories, a collection of databases from various scientific disciplines are assembled (Table 5). The total number of databases currently in existence cannot be accessed at this time; the minimal number is likely to have exceeded tens of thousands in magnitude. The selection of databases in Table 5 follows the criteria: 1) the data are easily classifiable in EApc framework; 2) the database contains not only entity/property information but also action/condition information; and 3) the dimension of the subject matter, i.e., from atoms to astronomical objects, which will be relevant to defining the concept of “scale” in the next section. The first criterion is used because not all databases have the same degree of conceptualization, i.e., some list raw experimental data (e.g., some of the astronomy databases) without referring them to a conceptual label (a notation of a property or condition, such as color, mass, speed etc.). This may not be a problem for the people in that field, but may confuse readers intended for this paper. The selection may also have a bias toward biological sciences because of author’s training.

### Section summary

Table 5 shows that at least some databases across disciplines organize information in a manner that can fit the EApc categorization of basic knowledge, i.e., the basic units of knowledge are entities or actions. Given the above comparison between E-R model of database and the EApc framework, this observation does not come as a surprise. The E-R/EApc comparison does not touch upon the EApc’s concept of Action/Condition because there is no explicit mention of that category in the E-R model. Nevertheless, a couple of databases in Table 5 include Action/Condition category (e.g., Merck Index and BKM-react). The listing in Table 5 is certainly not exhaustive, but these examples can be viewed as evidence that a comparison of EApc and E-R model is justified and more importantly suggest that monitoring database updates can serve as a means for measuring more precisely progression of knowledge accumulation, complementary to counting the number of published papers.

### Concept of scale in EApc

Another pattern of knowledge is the concept of Scale, which describes the hierarchical relationships of entities (and actions). The origin of such notion can be traced to Phillip W. Anderson’s article titled: “More is Different” (Anderson, 1972), in which, he, a physicist, raised a notion against the reductionist view, prevailing at the time among the physical science community, that the natural world is governed by the same set of fundamental laws, and we only need to understand these fundamental laws to understand the world around us. Anderson countered “*The behavior of large and complex aggregates of elementary particles, it turns out, is not to be understood in terms of simple extrapolation of the properties of the few particles. Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other.*” Anderson’s concept is incorporated into the notion of “emergence” that new properties emerge from self-assembly of objects in nature (Goldstein, 1999). Self-assembly phenomena are widely observed in biology,

**Table 5. A sample of databases across scales, from atoms to galaxies, showing their listings of entities, properties, actions, and conditions**

Database	Periodic Table of Elements <sup>a</sup>	Merck Index <sup>b</sup>	BKM-react <sup>c</sup>	GenBank <sup>d</sup>	CellMarker <sup>e</sup>	Organism Taxonomy <sup>f</sup>	Geophysical database <sup>g</sup>	Solar System <sup>h</sup>	Cosmology <sup>i</sup>
Entities	Atoms	Chemical Compounds	Biochemical Compounds	Genes	Cells	Species	Earth	Planets	Galaxies
Properties	Standard State Atomic Mass Electron Configuration Oxidation State Atomic Radius Ionization Energy Melting Point Boiling Point Density Isotopes (Abundance, half-life) Year Discovered	Atomic Composition Chemical Structure Physical Properties (Melting point, boiling point, density etc.) Bioactivity	(In relation to reactions, i.e. reactants, products, catalysts, etc.)	DNA sequence Length Structure Genetic Motifs Genomic Location Epigenetic Modification Protein Coding Regulatory Elements Activities in Cells & Tissues	Molecular Markers Species Tissue Origin Organ Origin State (normal/diseased) Developmental Lineage Cell Type Classification	Genetic Code Lineage Genome Nucleotide Sequence Protein Sequence Gene Expression (from linked resource) Habitat Migration Body Size Longevity Behavior Predator Parasites Disease	Atmosphere Radiance Cryosphere Human Dimension Land Ocean	Equatorial Radius Mean Radius Mass Bulk Density	Location in the Sky Shape (Elliptical/Spiral) Size/Mass Brightness Light Emission Age Composition Star Types and Distribution
Actions		Organic Named Reactions	Biochemical Reactions (Catalyzed and Spontaneous)			(from linked source) Reproduction Feeding Prey or Being Preyed on Being Infected		Rotation Gravity Geometric Albedo	Rotation
Conditions		Reactants Products Catalysts Cofactors Reaction Mechanisms	Reactants (Substrates and Products) Catalysts Cofactors Inhibitor/Activator Reaction Characteristics (Km, pl, metals/ion requirements, turnover number, pH range, optimum, temperature range, optimum) Localization (Organism, tissue)					Escape Velocity Rotation Period Orbit Period	

The source of the databases is

<sup>a</sup><https://pubchem.ncbi.nlm.nih.gov/periodic-table/>.

<sup>b</sup><https://www.rsc.org/Merck-Index/browse?atozfilter=A&pagenumber=1&pagesize=20>.

<sup>c</sup><https://bkms.brenda-enzymes.org>.

<sup>d</sup><https://www.ncbi.nlm.nih.gov/genbank/>

<sup>e</sup><http://biocc.hrbmu.edu.cn/CellMarker/>

<sup>f</sup><https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Xenopus+laevis&lvl=0&srchmode=1&keep=1&unlock>

<sup>g</sup><https://earthdata.nasa.gov/#>, <https://sciencing.com/six-properties-earth-8371705.html>, <https://www.ngdc.noaa.gov/ngdcinfo/aboutngdc.html>

<sup>h</sup>[https://ssd.jpl.nasa.gov/?planet\\_phys\\_par](https://ssd.jpl.nasa.gov/?planet_phys_par).

<sup>i</sup>[https://ssd.jpl.nasa.gov/?planet\\_phys\\_par](https://ssd.jpl.nasa.gov/?planet_phys_par).

from aggregates of macromolecules, cells, tissues to species and ecosystems. The scale concept has also been influencing new developments in engineering disciplines, e.g., meso-science (Li et al., 2013), where scales span from atomic-molecular phenomena in chemical engineering systems to ecology and astronomy (Li et al., 2013; Whitesides and Boncheva, 2002; Whitesides and Grzybowski, 2002). Thus, the concept of scale represents an important aspect of knowledge patterns. Based on this consideration, we proposed, in the EApc framework, a concept of Scale in that entities are viewed as assemblies of other entities. This is necessary because an entity assembly possesses new properties not present in the component entities, such as a protein has properties not possessed by its component amino acids.

### Relative scales

It can be said that for any given entity X, referred to here as the *basic scale*, there exists a *super-scale*, which is occupied by the entities containing X as a component; and a *sub-scale*, in which entities are components of X. Thus, for proteins in general, the super-scale consists of multi-protein complexes (e.g., enzyme complexes, ribosomes, ion channels etc.), in which the protein under consideration is a component; and the sub-scale consists of all amino acids and molecular moieties (e.g., methyl group, sugars etc.) contained by the protein. The knowledge about this scale triad (super, basic, and sub) is essential in understanding and manipulating natural processes, as such triads can be observed in many research and invention processes. For instance, Tony McCaffrey studied collections of innovation cases (Challoner, 2009) and the psychological processes behind them (McCaffrey, 2012; McCaffrey and Spector, 2012). He concluded that an obstacle in problem solving lies in recognizing obscure features (properties) of the objects in the problem (the Obscure-Feature Hypothesis). A classic example was Duncker's "functional fixedness" phenomenon (Newell, 1980) that people tend to fix on certain common use of an object, overlook other "hidden" features, such as viewing a candle as an illuminating device (basic scale), and being unaware of the properties of its components, wick and wax (sub-scale), which can be used for tasks unrelated to the basic scale device (e.g., the wick when separated from the candle can be used as a string). While McCaffrey focused on basic and sub-scale relations, numerous examples of innovations are about basic and super-scale relationships. A number of examples came from the latest biotechnology applications of the CRISPR system, separating the programmable DNA-binding component from Cas9 protein and combining it with gene regulatory components from other sources, which are not in the original CRISPR system, to generate a new protein complex (super-scale) that can achieve artificial gene regulation and epigenetic modifications, e.g., as reviewed by Nakamura et al. (2021).

### Demarcation of scales

While the relative scales can be easily defined, a general method to demarcate scales is not readily formulated. An exemplar of the scales is the atoms, as depicted in the periodic table. In this scale, entities (atoms) and their properties (Table 5) are relatively well understood and the number of entities in this scale is reasonably stable, albeit still expanding slowly, after a burst of discoveries in the last two centuries (Scerri, 2019; Karol et al., 2016). Furthermore, attempts have been made recently to formulate periodic tables for diatomic and triatomic molecules (Hefferlin and Kuhlman, 1980; Hefferlin et al., 1983; Kong and Wu, 2012) and periodicity in n-atomic molecules has also been contemplated (Dias, 1984; Hefferlin, 2006). These works show that at lower molecular weight range, scales can be similarly defined as that for atoms. However, the scale demarcation at higher molecular weight (mass), in the range of macromolecules and their assemblies, e.g., in the realm of biology, is not as facile, and cannot be based on the number of atoms or molecules, but more likely be established by other properties of the entities, e.g., self-replication. Some databases may be organized by such principles, and Table 5 shows several examples.

### Section summary

These analyses show that the same relationships between Entity/Property and Action/Condition are likely applicable to individual scales, ranging from atoms to galaxies. A general definition of Scale requires additional work, but a relative definition can be readily achieved as in the scale triad (basic, super, and sub). The scale triad is widely used in describing nature and in inventing artificial processes.

### Cause-effect and EApc

Judea Pearl and Dana Mackenzie remarked in *The Book of Why* (2018, p 24) that "...very early in our evolution, we humans realized that the world is not made up only of dry facts (what we might call data today); rather, these facts are glued together by an intricate web of cause-effect relationships." and "...causal explanations, not dry facts, make up the bulk of our knowledge..." Similarly, Jerome Feldman observed, in his

book *From Molecules to Metaphor: A Neural Theory of Language* (Feldman, 2008, p 129), that in the development of individual humans, a child learns causal modeling at a very early age through language acquisition. And even a cursory survey will reveal that the current scientific landscape is permeated with questions seeking understanding of cause-effect relationships (e.g., what lifestyles can cause diseases or make people live longer, what drugs can cure a disease, what human activities can cause undesirable environmental changes, etc.) Thus, it is imperative that a knowledge framework provides a means to express cause-effect relationship.

In the EApc framework, the cause-effect relation is captured by the basic knowledge category of Action, which is initially defined as “a process, required to change an entity or its property.” (Tseng and Small, 2019) The argument can be made from two angles, 1) cause and 2) effect.

### Cause as action

Lakoff and Johnson analyzed extensively the cause-effect phenomena from a perspective of embodied mind and metaphoric cognition and stated that “our most important primary metaphors, Causation is Action to Achieve A Purpose, where Causes are Reasons (why the action will in fact achieve the purpose).” (Lakoff and Johnson, 1999, p 217) This view is echoed in Feldman’s analysis that causes are expressed by verbs (Feldman, 2008, pp 203–204). Such notions are consistent with above mentioned observation that linguistic lexicon universals contain Verb as an essential component of human language. Both Lakoff/Johnson and Feldman pointed out the complexity in expressing the causation concepts, that “The forms of causation depend on the kind of force it is, what the force is applied to and the kinds of changes it produces” (Lakoff and Johnson, 1999, pp 206–211) and “We do not have a single concept of causation, but many, each with different inference structures” (Feldman, 2008, p 204). In the EApc framework, the variety of actions (causes) is dealt with by the category of Condition, which is associated with Action. As described above in section [adverbs \(linguistics\) and conditions \(EApc framework\)](#), conditions are analogous to adverbs which are attributes of verbs (actions). Conditions can thus be used to define an action (Fillmore’s semantic roles concept), just as Properties can define Entities (the “key” concept in the database E-R theory) (section [key](#)). Conditions can be further categorized into “entity-related conditions”, e.g., instigator, recipient, inhibitor, etc., and “time-related conditions, e.g., speed, duration, periodicity, etc. . . .”, which correspond to some factors Lakoff, Johnson, and Feldman used to classify causes. It is thus conceivable that by different combinations of conditions, a very large, if not unlimited, number of actions (causes) can be conceptualized.

### Cause as defined by effect

Another research area interested in cause-effect relation is statistical causal inference, which focuses on extracting information about causes from experimental data, which, as I argued, relate to observed values a property takes. Pearl et al. (2016) defines cause-effect as “A variable  $X$  is a cause of a variable  $Y$  if  $Y$  in any way relies on  $X$  for its value.” Similarly, Holland (1986) defines causes (or treatments as used by experimentalists) as the difference (change) in the values of a parameter in the absence and presence of a treatment (occurrence of the cause):

$$“t \text{ causes the effect } Y_t(u) - Y_c(u)” \text{ (Equation 2 in Holland, 1986)}$$

where  $Y$  is a variable responsive to the treatment, and  $Y(u)$  the value that  $Y$  takes,  $t$  is cause or treatment, and  $c$  is absence of  $t$ , i.e., control; hence,  $Y_t(u)$  is the value in the presence of  $t$  and  $Y_c(u)$  is the value in the absence of  $t$ . The definitions of Pearl et al. and Holland therefore share their reliance on “value”. As described in section [value](#), the value here relates to the EApc notion of Property, i.e.,  $Y$  can be viewed as a property, and thus, these definitions can be re-stated as “a cause (treatment) alters the property (of an entity)”, which is exactly the EApc definition of Action.

### Section summary

These analyses suggest that the Action category in the EApc basic knowledge classification can capture, to a large extent, the cause-effect concept/definition in theories of metaphoric cognition, the verb category in lexicon universals, and statistical causal inference. Importantly, the EApc’s notion of the structure of compound knowledge (in Figures 1C and 2 of Tseng and Small, 2019) coincides with the view of Pearl and Mackenzie that cause-effect (Action) is a “glue” that links entities to form descriptions of natural and human-invented processes.

## A refined model of basic knowledge

Based on the above analyses, the basic knowledge model we proposed earlier (Tseng and Small, 2019) is expanded and refined here. The revised model is depicted in Figure 1. It retains the basic categories and the scale concept, clarifies the relationship of Entity/Properties and Action/Conditions, and introduces subcategories in Properties and Conditions and adds the E-R model's value concept to Properties and Conditions. Described below are the details of these refinements.

### Entity/properties and action/condition

The Entity/Action and their associated Properties and Conditions are depicted in Figure 1A. What differs from the earlier depiction is that now properties/conditions are further classified into six subcategories. These subcategories are meant to be general to cover many types of properties, but certainly not exhaustive, leaving further addition and classification among and within subcategories still possible, but sufficing here for illustrating the principles. A graphic design feature of Figure 1A is that Entity and Action are depicted as a core and Property and Condition as orbiting around the core. This configuration is meant to convey a sense of the roles of Property and Condition in knowledge structure that Property and Condition are not only defining the identity of Entity and Action (see section summary, adverbs (linguistics) and conditions (EApC framework) and key) but they also play key roles in defining the relationships among entities and actions, e.g., in categorization (with similar properties and conditions), in lineage tracing (related in time and space), as well as in the formation of compound knowledge (an example, the process of making recombinant DNA is given in Tseng and Small 2019). In other words, the pattern of properties and conditions are analogous to the electron orbit configurations of an atom, defining its identity (partially) and its reactivity with other atoms and molecules, i.e., they are the valence of entities and actions. In this sense, as atoms had been considered basic units of matters, entities and actions are portrayed here as standalone units of knowledge.

### Values

Values (or data) in the EApC framework are observed status of a property (or condition). Values (or data) are therefore dependent on the method of observation and the conceptual understanding of the phenomenon (property or condition). For instance, the property Temperature can have values such as "cold", "warm", and "hot" when sensed by human skin; or 4°C, 30°C, and 100°C, when measured by a thermometer; or expressed as  $T = PV/nR$  according to the Ideal Gas Law in pressure (P), volume (V), mass (n), and ideal gas constant (R). Similarly, Color can have values such as "blue" and "red", when sensed by human eyes, or expressed as electromagnetic wavelength 430 nm, 700 nm, when measured by diffraction grating or other physical means. However, not all properties (or conditions) can be described as objectively. For instance, for chemicals or diseases that cause pain or itch, clinical measurements (value or data) of the type and intensity of these sensations still relies on patients' subjective reporting, which cannot achieve the precision of physical instruments. Conversely, there are properties that human can sense only with the aid of an instrument, e.g., the shapes of bacteria or distant galaxies.

The refined basic knowledge model adopts some concepts of relation database (Codd and Jose, 1970; Chen, 1976) by describing entities/properties and actions/conditions as sets and relations between sets (Figure 1C). Similarly, the refined model adds a description of value as sets of states that properties and conditions can have (Figure 1D). In general:

$$\text{Property (Entity)} \in (\text{value}_1, \text{value}_2, \dots, \text{value}_n) \quad (\text{Equation 1A})$$

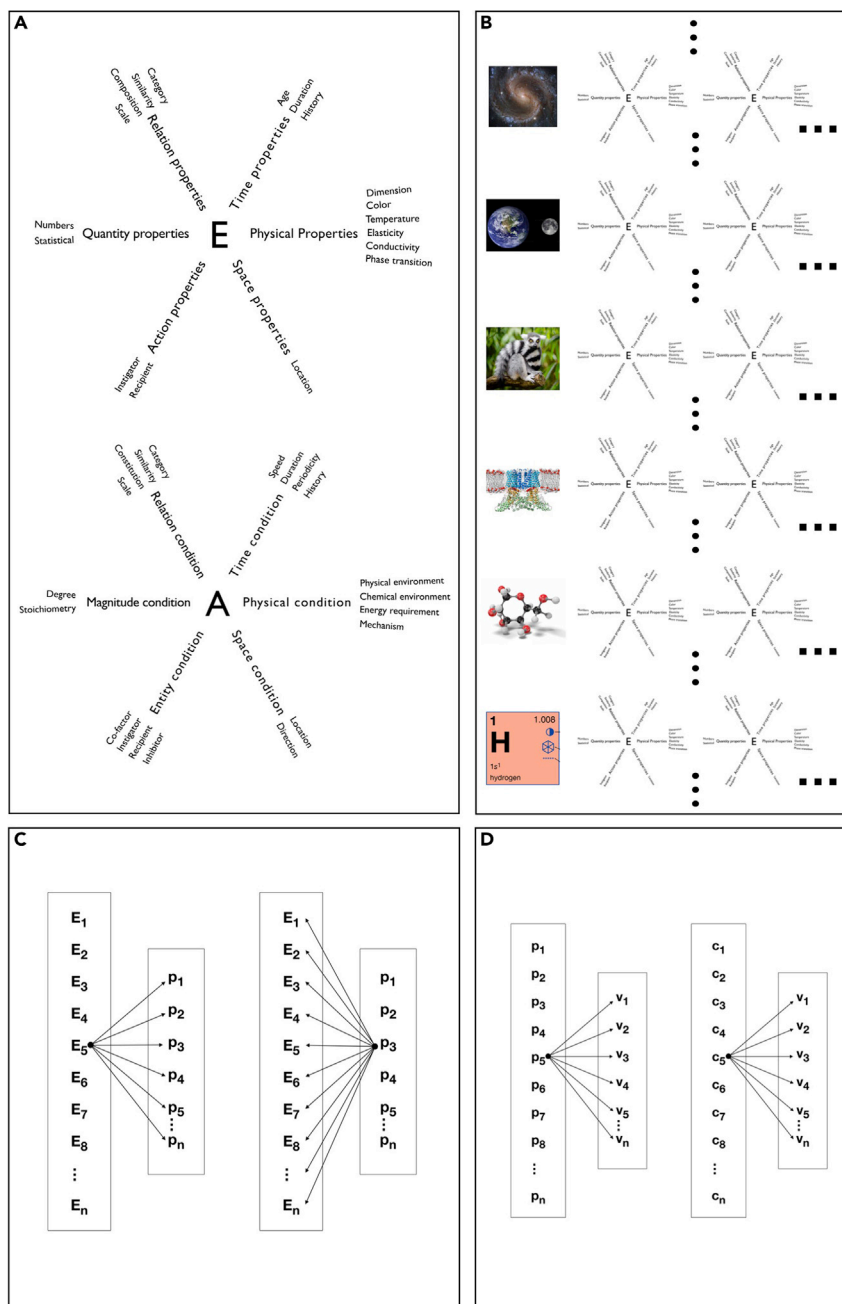
$$\text{Condition (Action)} \in (\text{value}_1, \text{value}_2, \dots, \text{value}_n) \quad (\text{Equation 1B})$$

For example:

$$\text{temperature (person)} \in (36^\circ\text{C} < t < 42^\circ\text{C})$$

$$\text{color (car)} \in (\text{red, silver, black etc.})$$

$$\text{speed (run)} \in (0 < v < 300 \text{ km/h})$$



**Figure 1. A graphic depiction of the refined EApc framework**

In (A), the Entity and Action and their associated Properties and Conditions are depicted as nodes and edges, respectively. Properties and Conditions are further divided into subcategories and sub-subcategories to illustrate some but not all possible items in the categories.

(B) Shows examples of Scales, from atoms to galaxies. The ellipsis, three round or square dots, indicates omission of scales and entities, respectively.

(C) Depicts the relationship, shown as a mapping, between entities and associated properties, that an entity can have multiple properties (subcategories) and entities can share the same subcategory of properties, e.g., some entities can share the property category of color. Action and condition can have the same type of relationships. For brevity, it is not depicted here.

(D) Depicts the relationship between properties and the values they can take. This relationship can be expressed as a set notation (see [values](#)). Similarly, conditions have the same relationship to their values (not depicted here).

*direction (run) ∈ (east, west, south, north)*

It is hoped that such effort could introduce mathematical rigor into the EApc framework and the set theory with its rich concepts can be used for describing and manipulating knowledge as pioneered by [Spivak \(2013\)](#), among others.

It should be noted here that the EApc framework does not concern if the observed status of a property (or condition) is true or false. The framework assumes that knowledge is evolving, implying that at any given time, it contains inaccurate information (untrue) about and incomplete description of Nature.

### Scales

The concept of Scale we defined previously was a relative scale referring to emerging new properties when entities are assembled to form new entities. Similarly, actions can also be viewed as assemblies of component actions, e.g., Darwinian natural selection is an action resulting from multi-scale interactions from molecular to social. In the present work, this concept is expanded to encompass all entities in the knowledge domain ([Figure 1B](#)). The expansion is based on examining some current databases spanning the knowledge domains (see [demarcation of scales](#) and [cause as action](#)) and influenced by works in emergence, complexity, and meso-science (e.g., [Anderson, 1972](#); [Holland, 1992](#); [Goldstein, 1999, 2007](#); [Li et al., 2013](#)). It should be clarified that the definition is still based on consideration of properties (and conditions when referring to actions) that entities in different scales have unique properties not possessed by their components in the sub-scales, just as Anderson remarked: “Surely there are more levels of organization between human ethology and DNA than there are between DNA and quantum electrodynamics, and each level can require a whole new conceptual structure.” ([Anderson, 1972](#)). Thus, the Scale is not used here to refer to size, weight, mass etc., i.e., dimensional properties, although there is an association between the property set an entity can have and its dimensional characteristics, e.g., described by [West \(2017\)](#). It is believed that the Scale concept proposed here is necessary for understanding some knowledge-related theories, e.g., reductionism, holism, complexity, etc.

### Cause-effect as action

Causal reasoning is considered an essential component in human thinking, which endows people the ability to explain, predict, and manipulate natural events for their benefit ([Waldmann, 2010](#)). Mackie declares that cause-effect relation is “the cement of universe” ([Mackie, 1974](#)), reminiscent of Pearl’s notion that Cause is the “glue” that links facts (data) (section [cause-effect and EApc](#) above). These and other remarks indicate that Cause is a major vocabulary in human’s description of Nature. In the initial proposal of the EApc framework ([Tseng and Small, 2019](#)), cause-effect was not considered, and the basic knowledge category Action was largely derived from the sense of motion and transformation. The studies in the embodied mind, metamorphic cognition, statistical inference, and computation modeling of the mind (section [cause-effect and EApc](#) above) led to the realization that Cause belongs to the basic knowledge category of Action in the EApc framework. This is a significant improvement of the framework, making it more comprehensive than the previous version.

### Discussion

The EApc framework hypothesizes that at a fundamental level, knowledge consists of *two descriptive systems*: 1) describing an object, i.e., an entity and its properties; and 2) describing a process, i.e., an action and its conditions. This hypothesis implies that entity and action are the basic units of knowledge. The identity of the entity or action is defined by the properties and conditions that they possess, respectively. Properties also determine the relations an entity can have with other entities; and conditions determine the relation an action can have with other actions. Relations between entities can also be established by their roles in an action (section [cause as action](#)). These descriptive systems can be applied to phenomena through scales, from the subatomic particles to their largest assemblies. Combined, properties and conditions allow linking entities and actions together in the construction of compound knowledge. Thus, from the two basic units of knowledge, i.e., Entity and Action, compound knowledge structures can be developed with virtually no limit in variety, dimension, and complexity, as long as enough number of different entities and actions are involved. The EApc framework, therefore, is addressing the questions raised in the Introduction: What is the basic unit of knowledge? How many types of knowledge



are there? How do they relate to each other? And what are the higher-order (hierarchical) patterns of knowledge?

The Entity/Property relationship has been noted in many previous studies, in various forms and iterations, by historical and contemporary thinkers. For instance, Wittgenstein observed that objects may have various properties and may hold diverse relations to one another (Biletzki and Matar, 2018); Dahlberg referred it in linguistic terms (i.e., subjects and predicates), when describing the basis of her Information Coding Classification (ICC) system (Dahlberg, 1982); Feldman (2008, p 140) summarized succinctly and explicitly, “The general idea of entities having features (or roles or properties) is universal in science and everyday language use, and it plays an important role in the theory of knowledge”; and its application in early database theories have been clearly shown (section database theories and EApc), to name only a few. On the other hand, Action/Condition relationship does not appear to have been recognized with the same clarity or breadth, and certainly not listed as a parallel to the Entity/Property pair. Still, Action/Condition relationship can be glimpsed in the theories of several fields (section cause-effect and EApc). These observations appear to support the existence of the two distinctive descriptive systems, thus strengthening our previous argument for the EApc framework.

An issue of using linguistic argument for knowledge categorization and structure is the omission of the roles of graphics in knowledge expression and communication (e.g., graph theory), whose implication to the EApc framework deserves further study. Basic knowledge (i.e., entities/properties and actions/conditions) can be easily depicted by the definition of graph theory:

$$G = (V, E) \quad (\text{Equation 2})$$

where  $G$  is an ordered pair  $V$  and  $E$ .  $V$  can be entities or actions and  $E$  can be properties and conditions, respectively. Entity and Action can be viewed as nodes and Property and Condition as edges in a graph (e.g., Figure 1A).

Apparently, in many situations, graphics can ease cognitive load, helping people grasping the essence of the information communicated. A paragon of the databases is the representation of atomic elements, and its well-publicized graphic visualization, the Mendeleev’s periodic table, which depicts the relationships among entities (elements) by their properties. The recent proliferation of the field information visualization, concerning graphic depiction of compound knowledge with many entities and actions and their relationships, and along the development of graph theory, added materials and ideas to knowledge representation, which differ from description by words (e.g., Tufte, 2001; Ware, 2021; Lima, 2011; Chen et al., 2012). Thus, understanding the cognitive principles of why pictures can sometimes aid knowledge communication and processing should help formulate a more comprehensive theory of knowledge. It suffices to say here that graphic information can be communicated in words, albeit somewhat awkwardly, and graphics are often more effective when used with words (e.g., labels and legends). The topic will be revisited elsewhere when patterns of compound knowledge are analyzed.

A common question encountered by any new framework is that what is its application in relevant areas? In a recent article in the *Journal of Economic Literature*, George A. Akerlof made a distinction between “hard” and “soft” scientific disciplines, relating to the precision (or quantitative vs qualitative) a research field can describe its subject matter (Akerlof, 2020). As an example, he argued that sub-disciplines in economics that use mathematics for analysis and modeling are “harder” than those using word descriptions, because mathematical equations are more “precise” than words. And these “harder” sub-disciplines are proliferating faster than the “softer” ones, because of reasons mainly in the realm of sociology of science. By Akerlof’s reasoning, epistemetrics is currently a “softer” discipline because of lacking a mathematical definition of knowledge. An exception may be the application of the category theory to describing knowledge (e.g., Spivak 2013), which unfortunately, has not been widely used. The present work continues our previous effort of quantification of knowledge by linking the mathematical set concept to basic knowledge description. Thus, it is hoped that the EApc framework will help “harden” epistemetrics as a discipline.

The EApc framework suggests a new way to monitor knowledge accumulation process. We analyzed the basic knowledge content in the original papers describing the recombinant DNA invention and traced



the generation of the necessary knowledge preceding the invention (Tseng and Small 2019). This proof-of-principle demonstration also revealed that extracting basic knowledge categories from research papers can be laborious, requiring specialty expertise in both understanding of the subject matter and familiarity of the technical details (e.g., molecular biology). Thus, in the current technological environment, even with the development of artificial intelligence and natural language processing, large-scale automated analysis of this type on research papers across disciplines is unlikely to be implemented readily. However, the dependency on discipline expertise can be alleviated if one turns to another type of knowledge representation, namely knowledge residing in the databases of various research disciplines. Many databases are built on ontological principles to account for and to delineate relationships among existing entities (Table 5). Importantly, the databases are generally constructed by discipline-specific experts, who read the publications in a research field and manually curate relevant information, standardize the vocabulary and format, and provide annotation and the source (i.e., research publications) of information. This process reduced vastly the task of analyzing knowledge at the basic level. Compared to the mainstream practice of analyzing research papers, analyzing the content of entities and actions in the existing databases to monitor scientific progress appears few and far in between. It is possible that the EApc framework could stimulate this type of analyses.

The EApc framework can conceivably be used as a basis for developing a value system for evaluating knowledge-related works, which encompasses every corner of the scientific enterprise, in grant proposals, research publications, professional promotions, prizes and honors, to name an obvious few. However, the value systems used in these evaluations can be very heterogeneous, e.g., technical values, utility values, social values, financial values, ethical values, historical values, and with very different weighing on each value categories. Epistemological value should be and often is one of the considerations in the evaluations in the scientific enterprise, because the whole purpose of this enterprise is to generate and propagate knowledge. The epistemological evaluation is currently largely based on peer opinions (citation and peer review), which can vary widely depending on the education, experience, and personal bias of the reviewers (Brezis and Birukou, 2020; Kahneman et al., 2021). Remarkably, there appears to be no standard value systems in evaluating knowledge for its own sake, e.g., which discovery is more valuable to a body of knowledge, discovering an entity/action, or property/condition? how can one define the degree of innovation of a discovery or an invention? It is true that whatever values the evaluators select to use are of their free choice, but an understanding of whether a value system intrinsic to knowledge is being used will bring clarity to the evaluation itself, especially when cross-discipline scientific works are concerned. Therefore, developing a knowledge value system from the EApc framework should be a worthy task deserving a separate investigation.

In summary and conclusion, the present work expands the EApc framework by examining concepts and practices in several research fields that are distinct from and independent of the original observations made in molecular biology and biotechnology. The support from linguistic lexicon universal is of particular interest because it has been well recognized that language is rooted in the human cognitive apparatus; hence, the Entity/Action descriptive systems are likely the product of biological evolution. The observations made in this study show manifestations of the two descriptive systems in action in the knowledge generation, accumulation, storage, and organization processes in several scientific disciplines. Thus, the foundation of the EApc framework is strengthened and its potential impact broadened, which include:

Introducing mathematical set theory into knowledge quantification analysis, thus “hardening” epistemetrics.

Using databases to monitor and quantify knowledge in a given discipline.

Developing a value system intrinsic to knowledge, independent to other value systems in the evaluation of scientific works.

The EApc framework is still in its infancy, a large number of issues remain to be investigated. For instance, how the compound knowledge is formed and how can it be described formally? Can EApc framework be improved by incorporating concepts in the graph theory and information visualization, where some complex knowledge can be communicated more efficiently? How can discoveries and innovations be ranked by

using the EApc-based knowledge value system? and many more interesting questions and perspectives that arise when a new framework is introduced.

### Limitations of the study

A number of limitations are mentioned throughout the text and mainly in the Discussion. For instance, the graphical representation of knowledge has not been considered here in detail; the use of mathematical set theory to describe basic knowledge has just been touched upon and not fully developed; and similarly, each topic area can be further investigated, and how compound knowledge is described by the EApc framework remains to be studied. The ultimate resolution of these limitations probably requires broader engagement of the research community at large.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**

### ACKNOWLEDGMENTS

The author wishes to thank the anonymous reviewers and Francis Tseng for their comments and suggestions. This research did not receive any grant support from funding agencies in the public, commercial, or not-for-profit sectors. The author is an employee of the National Institute of Arthritis Musculoskeletal and Skin Diseases (NIAMS), National Institutes of Health (NIH). His views in the paper are personal, and do not represent those of his employer's.

### AUTHOR CONTRIBUTIONS

H.T. conceived the ideas and wrote the paper.

### DECLARATION OF INTERESTS

The author declares no competing interest. The author's views are personal, and do not represent those of his employer's, NIAMS/NIH.

Received: September 17, 2021

Revised: December 2, 2021

Accepted: January 19, 2022

Published: February 18, 2022

### REFERENCES

- Akerlof, G.A. (2020). Sins of omission and the practice. *J. Econ. Lit.* 58, 405–418.
- Anderson, P.W. (1972). More is different. *Science* 177, 393–396.
- Baker, M.C. (2003). *Lexical Categories: Verbs, Nouns and Adjectives* (Cambridge University Press).
- Barsh, G.S., Copenhaver, G.P., Prakash, E.S., and Zarnescu, D.C. (2019). 2019 PLOS Genetics Research Prize: fruit fly school – language and dialects for communicating a threat. *PLoS Genet.* 15, 10–12. <https://doi.org/10.1371/journal.pgen.1008381>.
- Biletzki, A., and Matar, A. (2021). 'Ludwig Wittgenstein', *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/win2021/entries/wittgenstein/>.
- Brezis, E.S., and Birukou, A. (2020). Arbitrariness in the peer review process. *Scientometrics* 123, 393–411.
- Challoner, J. (2009). 1001 Inventions that Changed the World (Barron's Educational Series, Inc).
- Chen, C., Hu, Z., and Tseng, H. (2012). Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opin. Biol. Ther.* <https://doi.org/10.1517/14712598.2012.674507>.
- Chen, P.P. (1976). The entity-relationship model—toward a unified view of data. *ACM Trans. Database Syst.* 1, 9–36.
- Chomsky, N. (1957). *Syntactic Structures*, 2002nd edn (Mouton de Gruyter).
- Chomsky, N. (1997). Language and Problems of Knowledge, XVI (Teorema), pp. 5–33.
- Chomsky, N. (2011). Language and other cognitive systems. What is special about language? *Lang. Learn. Development* 7, 263–278. <https://doi.org/10.1080/15475441.2011.584041>.
- Codd, E.F., and Jose, S. (1970). A relational model of data for large shared data banks. *Commun. ACM* 13. <https://doi.org/10.1145/362384.362685>.
- Comrie, B. (1989). *Language Universals and Linguistic Typology*, 2nd edn (Blackwell).

- Dahlberg, I. (1982). ICC-information coding classification-principles, structure and application possibilities. *Int. Classif.* 9, 87–93.
- Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex* (Murray, J).
- Dias, J.R. (1984). A periodic table for polycyclic aromatic hydrocarbons. 4. Isomer enumeration of polycyclic conjugated hydrocarbons 2. *J. Chem. Inf. Comput. Sci.* 24, 124–135.
- Dixon, R.M.W. (1991). *A New Approach to English Grammar on Semantic Principles* (Clarendon Press).
- Feldman, J. (2008). *From Molecules to Metaphor: A Neural Theory of Language* (M. Press).
- Fillmore, C.J. (1968). Types of lexical information. In *Studies in Syntax and Semantics*, pp. 109–137. [https://kb.osu.edu/bitstream/handle/1811/85939/WPL\\_02\\_November\\_1968\\_065.pdf?sequence=1](https://kb.osu.edu/bitstream/handle/1811/85939/WPL_02_November_1968_065.pdf?sequence=1).
- Goldstein, J. (1999). Emergence as a construct: history and issues. *Emergence* 1, 49–72. [https://doi.org/10.1207/s15327000em0101\\_4](https://doi.org/10.1207/s15327000em0101_4).
- Goldstein, J.A. (2007). Emergence and identity. *Classic Pap. Emerg. Identity* 9, 75–96.
- Hefferlin, R. (2006). The periodic systems of molecules. In *Philosophy of Chemistry*, D. Baird, et al., eds. (Springer-Verlag), pp. 221–243.
- Hefferlin, R.A., Zhuvikin, G.V., Caviness, K.E., and Duerksen, P.J. (1983). Periodic systems of N-atom molecules. *J. Quant. Spectrosc. Radiat. Transfer* 32, 257–268.
- Hefferlin, R., and Kuhlman, H. (1980). The periodic system for free diatomic molecules-III. Theoretical articulation. *J. Quant. Spectrosc. Radiat. Transfer* 24, 379–383.
- Holland, J.H. (1992). Complex adaptive systems. *Daedalus* 121, 17–30.
- Holland, P.W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960.
- James, W. (1890). *The Principles of Psychology* (Henry Holt and Company).
- Kahneman, D., Sibony, O., and Sunstein, C.R. (2021). *Noise: A Flaw in Human Judgment* (Little).
- Karol, P.J., Barber, R.C., Sherrill, B.M., Vardaci, E., and Yamazaki, T. (2016). Discovery of the elements with atomic numbers  $Z = 113, 115$  and 117 (IUPAC technical report). *Pure Appl. Chem.* 88, 139–153. <https://doi.org/10.1515/pac-2015-0502>.
- Kong, F.A., and Wu, W. (2012). Periodic tables of diatomic and triatomic molecules. *Sci. China Chem.* 55, 618–625.
- Lakoff, G., and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought* (Basic Books).
- Levin, B. (2009). *Lexical Semantics of Verbs I: Introduction and Role-Centered Approaches to Lexical Semantic Representation*, 116 (UC Berkeley Course LSA).
- Li, J., Huang, W., Edwards, P.P., Kwauk, M., Houghton, J.T., and Slocombe, D. (2013). On the universality of mesoscience: science of “the in-between”. *arXiv*. <http://ir.lib.ntnu.edu.tw/handle/309250000Q/120701>.
- Lima, M. (2011). *Visual Complexity Mapping Patterns of Information* (Princeton Architectural Press).
- Mackie, J.L. (1974). *The Cement of the Universe: A Study of Causation* (Clarendon Press).
- McCaffrey, T. (2012). Innovation relies on the obscure. *Psychol. Sci.* 23, 215–218. <https://doi.org/10.1177/0956797611429580>.
- McCaffrey, T., and Spector, L. (2012). Behind every innovation solution lies an obscure feature. *Knowl. Manag. E-Learn. Int. J.* 4, 146.
- Mestres-Missé, A., Rodrigues-Fornells, A., and Münte, T.F. (2009). Neural differences in the mapping of verb and noun concepts onto novel words. *Neuroimage* 49, 2826–2835.
- Morgan, T.J.H., et al. (2015). Experimental evidence for the co-evolution of hominin tool-making teaching and language. *Nat. Commun.* 4–11. <https://doi.org/10.1038/ncomms7029>.
- Nakamura, M., Gao, Y., Dominguez, A.A., and Qi, L.S. (2021). CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* 23, 11–22.
- Newell, A. (1980). Duncker on thinking: an inquiry into progress in cognition. In *A Century of Psychology as Science: Retrospections and Assessments* (McGraw-Hill).
- Pearl, J., Glymour, M., and Jewell, N.P. (2016). *Causal Inference in Statistics: A Primer* (Wiley).
- Pearl, J., and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect* (Basic Books/Allen Lane).
- Raskin, V., Nirenburg, S., and Cruces, L. (1995). Lexical semantics of adjectives (Semantic Scholar), pp. 1–69. [https://pdfs.semanticscholar.org/237f/3d25f3ec9bbd25afec7568c77e51ebd0c882.pdf?\\_ga=2.38830819.1149916316.1598632298-789008128.1598632298](https://pdfs.semanticscholar.org/237f/3d25f3ec9bbd25afec7568c77e51ebd0c882.pdf?_ga=2.38830819.1149916316.1598632298-789008128.1598632298).
- Rescher, N. (2006). *Epistemetrics* (Cambridge University Press).
- Scerri, E.R. (2019). *The Periodic Table*, 2nd edn (Oxford University Press).
- Smith, R. (2018). Aristotle’s logic, the Stanford encyclopedia of philosophy. <https://plato.standord.edu/entries/aristotle-logic/#cat>.
- Spivak, D.I. (2014). *Category Theory for Scientists*, MIT Press Open Access, <https://mitpress.ublib.com/ereader/77/?preview=#page/Cover>.
- Tseng, H., and Small, H. (2019). Quantification of knowledge content of a high impact innovation: recombinant DNA. *Heliyon* 5. <https://doi.org/10.1016/j.heliyon.2019.e02219>.
- Tufte, E.R. (2001). *The Visual Display of Quantitative Information* (Graphic Press).
- Waldmann, M.R. (2010). Causal thinking. In *Towards a Theory of Thinking* (Springer-Verlag), pp. 123–134. [https://doi.org/10.1007/978-3-642-03129-8\\_8](https://doi.org/10.1007/978-3-642-03129-8_8).
- Ware, C. (2021). *Visual Thinking for Information Design*, 2nd edn (Morgan Kaufmann).
- West, G. (2017). *Scale: The Universal Laws of Growth, Innovation, Sustainability, and the Pace of Life, in Organisms, Cities, Economies, and Companies* (Penguin Books).
- Whitesides, G., and Grzybowski, B. (2002). Self-assembly at all scales. *Science* 295, 2418–2421.
- Whitesides, G.M., and Boncheva, M. (2002). Beyond molecules: self-assembly of mesoscopic and macroscopic components. *Proc. Natl. Acad. Sci. U S A* 99, 4769–4774. <https://doi.org/10.1073/pnas.082065899>.

## STAR★METHODS

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Hung Tseng ([htsengpe@gmail.com](mailto:htsengpe@gmail.com)).

#### Materials availability

All materials used in this work are bibliographical and listed in References.

#### Data and code availability

- This paper analyzes existing, publicly available data.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.
- This paper does not report original code.

### METHOD DETAILS

This is a theoretical work. No experimental method was employed.