Data Algebra

Hiding in Plain Sight

Gary J. Sherman
Principal Mathematician
gsherman@algebraixdata.com

Joseph C.Underbrink
Data Scientist
junderbrink@algebraixdata.com

Abstract. An algebraic model of data, referred to as **data algebra**, is shown to be the appropriate response to an apparently troubling question: What is data? We begin with the historical imperative for this question and then argue that the answer must be couched in axiomatic set theory. A brief, friendly introduction to (or review of) set theory sets the stage for illustrating, in terms of an arithmetic adventure common to us all, the driving mathematical principle of the paper: any algebraic structure a set supports may be lifted to its power set. This principle is used to create and discuss a template of data algebras.

1 Introduction

Data has been stored, organized and manipulated to reveal insights about the genesis of the data since tallies were first recorded on bone, stone and wood. The facility for storing, organizing and manipulating data has improved in parallel with the development of new storage media culminating in the current state of affairs in which the modern digital computer provides the speed and capacity to store, organize and manipulate ever larger, but necessarily finite, data sets.

How has revelation progressed? If the data, independent of how, or even whether, it is stored in a computer, is 'homogeneous' (i.e., all of one type) and 'algebraic' (i.e., subject to mathematically precise operations and relations), progression has been astounding. Theoretically speaking, the Fermat and the Poincaré conjectures have been resolved and all finite groups have been characterized. On a more practical note, ciphers based on modular arithmetic encrypt financial transactions and error-correcting codes assure the correct transmission of these encryptions. If the data is of the 'real-world' — that is, apparently, neither homogeneous nor necessarily algebraic — rather than the ethereal world of pure mathematics, it may, for example, be at the mercy of visual-artifice/query language couplings such as RDM/SQL, XML/XQuery and RDF/SPARQL. Each of these sham marriages of data and mathematics-like jargon plumbs the depths of adhocery, abetted by legions of developers and database analysts with ever-increasing computer power in tow. In particular, the RDM/SQL coupling, which metastasized from E. F. Codd's seminal paper *A Relational Model of Data for Large Shared Data Banks* [**?**], treats NULL as a value, thereby stripping the equality relation of its reflexivity and rendering any subsequent mathematical statement moot.

The descent to the infernal regions is easy enough, but to retrace one's steps and reach the air above, there's the rub.
Virgil

The conflation of the concept of relational data with the geometry of a table (certainly a convenient and useful visual artifice) is at the heart of 'the trouble with NULL' and a canonical example of an intellectual parochialism which has a long, distinguished history in mathematics and the sciences. Euclid suffered from it. His geometry, indeed yours and ours in secondary school, was conceived in Egypt with planar surveying in mind. In such a context the notion that "through a point P not on a line L one and only one line could be drawn through P not intersecting L" seemed undeniably true — indeed, axiomatic. And, it is — in the Euclidean plane, as is the mathematically equivalent statement that the sum of three angles of a triangle is 180 degrees. The idea that Euclidean geometry was geometry, the one and only one, was not challenged until early in the 19th century when Nicolai Lobachevskii asked, in so many words: **But really, what are parallel lines?** His answer led to the creation of so-called hyperbolic geometry which permits multiple (parallel) lines through P not intersecting L and, equivalently, the existence of triangles whose angles

sum to less than 180 degrees, often a feature of Escher prints. Your favorite non-Euclidean geometry is probably the geometry of a sphere, a so-called elliptic geometry. In this geometry the lines are great circles perpendicular to a plane through the center of the sphere. (The motivation for calling great circles lines is that the shortest path between two points on a sphere is always an arc of a great circle.) It follows that every line through a point P not on a line L intersects L and equivalently that the sum of the three angles of a spherical triangle is greater than 180 degrees.

Isaac Newton suffered from it — his notion of physics was based on a four-dimensional Euclidean space-time model. Eventually, Einstein rescued him by asking, in so many words: **But really, what is space-time?** His answer was special relativity and then general relativity, each based on four-dimensional non-Euclidean geometries. In a more general sense the mathematicians who developed the body of classical mathematics — geometry, analysis and algebra — each component of which, in one sense or another, had its roots in Euclidean geometry, which is to say in lines (sets) and points (elements), suffered from it to the extent that they didn't immediately ask: But really, what is the foundation of our mathematics? And that parochialism was recognized gradually and haltingly, often reluctantly and resentfully, and more than once at the cost of a mathematician's sanity during the years between the late 19th century and the mid 20th century when, finally, pure mathematicians accepted set theory as the foundation for all modern mathematics.

It is our contention that it is time, indeed long past time, to ask the obvious question: **But really, what is data?** Dictionary definitions of data select from among phrases such as;

- facts and statistics collected together for reference or analysis,
- factual information used as a basis for reasoning,
- the plural of datum,
- any representations to which meaning is or might be assigned.

Thesauri suggest synonyms such as information, aggregation, collection, accumulation, assemblage, facts, details and input. And of course the word data appears in the lists of synonyms for each of these synonyms for data.

The responses of data-gurus to this question, often following long pauses punctuated with some combination of facial contortions and guttural emanations, are more telling:

- "A database."
- "What do you mean by that?"
- "Information."
- "Anything that somebody might want to know that is enumerated, listed or quantified."

- "It is the answer to a question (may be a hypothetical question). It can be written (this includes that it is limited)."
- "Data."
- "Facts."
- "A character on Star Trek: Next Generation."
- "In modern computer-parlance, it literally means anything that has been captured in any form on any recording medium."
- "Knowledge."

This sample of responses is a hodgepodge of ambiguity and circularity which, left to gel, precludes a rigorous mathematical discussion of data — as is witnessed by the evolution of SQL, a mathematical disaster of the first order. A similar story is played out, albeit in the singular, if the original question is "*What's a datum?*" In either case, to paraphrase Carl Sagan (*Dragons of Eden*), those closest to the intricacies of data seem to have a more highly developed (and ultimately erroneous) sense of its mathematical intractability than those at some remove. As it turns out the wife of one of the authors, a well-known 'data-phobe', provided the most suggestive responses to both questions. Data? — a quizzical look, then "*Well, come on now, you and I both know what it means.*" Datum? A threatening look, then "*You and I both know what it means!*" Indeed, anybody making a living in the, well, data-business 'knows' what data is and 'knows' what a datum is and uses the terms — with reckless abandon — in discussions with colleagues who 'know' what data is and 'know' what a datum is, even though 'know(s)' may not equal 'know(s).' But the only implied constant in the discussions is suggested by the dictionary: data is the plural of datum. It is our contention that the key to providing a mathematical understanding of data is to take data and datum as primitive, undefined terms and axiomatize their relationship; i.e., the notion of belonging. Fortunately this has already been done under the heading of Zermelo-Fraenkel set theory with choice (ZFC) with data and datum in the guise of set and element, respectively. (The word 'choice' in the previous sentence refers to the axiom of choice, a subtlety of contending with infinite sets.) It follows that any mathematically legitimate notion of data algebra must be couched in terms of ZFC, must be independent of any preconceived visual artifice and must be subject to the rigors of axiomatic set theory. As a consequence, the algebraic tale of data, indeed any algebraic tale, is a tale of sets told by ascending towers of sets of ever increasing cardinality of, in Paul Halmos' words, "*frightening height and complexity.*"

Objection noted: those applied algebras we studied in elementary school (integer algebra), high school (polynomial algebra) or college (matrix algebra) seemed to amount to rules for manipulating objects with operations, ready-made for each other, plucked from the ether by a wizard — no ascent required. As we shall see, the apparent lack of set-theoretic ascent is an illusion. However,

following the dictates of a wizard is acceptable — indeed, desirable in practice — if those dictates are respectful of a pre-existing, legitimate algebra's foundation in set theory. The fact that the set-theoretic foundations for integer algebra, polynomial algebra and matrix algebra have been vetted by mathematicians and deemed legitimate is the wizard's rationale for the way those subjects were taught.

This paper presents the set-theoretic foundation for data algebra, as developed by the authors, in a manner which legitimizes data algebra while being accessible, in varying degree, to a broad audience. Toward this end, Section 2 (**Sets — in brief**) is a (very) brief, chatty synopsis of ZFC's relevant axioms and constructions. For more on ZFC see Paul R. Halmos' *Naive Set Theory* [**?**], an expository gem and the set theory book to read if you are going to read only one. If you decide to read two, then we recommend Patrick Suppes' *Axiomatic Set Theory* [**?**], which is more rigorous than Halmos' book and a bit less sterile than more recent texts on set theory. If you are comfortable with the rudiments of ZFC skip to Section 3 and use Section 2 as a reference. Section 3 (**Algebra by ascent**) retells the story of your first experience with modular arithmetic in the context of an ascent of power sets of the integers. The morals of this parable are abstracted to reveal that the guiding principles for the creation of any algebra, so in particular a data algebra, were hiding in plain sight in elementary school. No skipping allowed, even — dare say especially — if you carry the scars of a discrete mathematics course or an introductory abstract algebra course. Section 4 (**Data algebras**) begins by tweaking, while generalizing, traditional mathematical jargon to better suit our exposition of data algebra. Data algebras are characterized as those algebras that begin their ascent using a toehold gained from the Cartesian product of a so-called genesis set with itself. A basic template of data algebras is created and discussed.

2 Sets — in brief

Any game evolves subject to rules which are enforced by referees using a language that is derivative of the rules. A casual game may have implicit rules and be refereed by the participants in an idiomatic language. Examples of casual games include hopscotch, various forms of tag and, in the authors' opinions, the Relational Data Model (RDM). Set theory is not a casual game. The rules are explicit, so-called **axioms**, and the game is refereed by deductive logic in a formal syntax. The primitive (undefined) notion of set theory is that of belonging. If an element e (whatever e is) **belongs** to a set E (whatever E is) we write $e \in E$. We may also read this notation in a manner which is appropriate to grammatical context to include '*e is an element of E*', '*e is contained in E*', or '*e in E*.' If e is *not an element* of E we write $e \notin E$. The mention of the set E begs questions.

How do we know such a thing as a set exists?

Axiom of existence: *There exists a set.*

Maybe it's E, maybe it isn't.

How can we tell?

> **Axiom of extension:** *Sets* A and B *are equal, denoted by* $A = B$, *if, and only if, they have the same elements.*

This axiom suggests the possibility of other relationships between sets. If A and B are not equal we write $A \neq B$. If $a \in A$ implies $a \in B$, then we say A is a **subset** of B (or B is a **superset** of A) and write $A \subset B$. It follows that $A = B$ if, and only if, $A \subset B$ and $B \subset A$. If $A \subset B$ and $A \neq B$, then we write $A \subsetneq B$ and refer to A as a **proper** subset of B. If the aforementioned set E consists of exactly three distinct elements, say a, b and c, we write $E = \{a, b, c\}$. It follows from the axiom of extension that $\{a, c, b\}$ and $\{c, b, c, a, c, a, \}$ are equal to E; i.e., neither the order in which the elements of E appear nor repeated appearances of a particular element of E has any bearing on the set E. The **cardinality** of E (i.e., the number of elements in E) is three and we write $|E| = 3$.

How do we construct the subsets of a set?

> **Axiom of specification**: *Every set* A *and every statement* $S(x)$ *determines a unique subset* S *of* A *whose elements are exactly those elements* x *of* A *for which* $S(x)$ *holds.*

To indicate the way S is generated from A by $S(x)$ we write

$$S = \{x \in A : S(x)\}.$$

The choices for $S(x)$ are exhausted by atomic statements of belonging ($x \in E$) and equality ($A = B$) and all statements that can be generated from these atoms using the syntax of logical operators ($\wedge$, $\vee$, $\neg$, $\Rightarrow, \Leftrightarrow, \exists$ and $\forall$). For example, any set B determines the subset

$$A \sim B = \{x \in A : x \notin B\}$$

of A. We refer to this set as the **difference** between A and B.

A surprisingly useful, if often dodgy, set manifests itself as a consequence of the previous three axioms. Let's consider $\{x \in E : x \neq x\}$, a subset of E with no elements whatsoever. It follows from the axiom of extension that this set, which we denote by $\varnothing$ and refer to as the **empty** set, is in fact unique. Moreover, $\varnothing$ is a subset any set.

Can a set be an element of a set?

> **Axiom of pairing:** *For sets* A *and* B *the set* $\{A, B\}$ *exists.*

In particular if we take $A = B = \varnothing$ we can immediately, and dramatically, enlarge our inventory of sets to include

$$\{\varnothing\}, \{\varnothing, \{\varnothing\}\}, \{\varnothing, \{\varnothing\}, \{\varnothing, \{\varnothing\}\}\}, \ldots$$

This sequence of sets hints at the genesis of the set of natural numbers,

$$\mathbf{N} := \{1, 2, 3, \ldots\},$$

provided we take the ellipsis in each display to mean ad infinitum. And we do — without further ado, except to say this assumption amounts to the **axiom of infinity**, which leads to the **axiom of choice** (*one may construct a set from an infinite collection of sets by choosing an element from each set of the collection*), which leads to the **axiom of substitution** ("*...anything intelligent that one can do to the elements of a set yields a set.*" (see [**?**]), which can lead to long philosophical discussions, which we are not interested in having in this paper, because as a practical matter we are concerned with finite sets subject to the caveat that, from time to time, it is clerically convenient to view a particular finite set as a subset of some countably infinite set; i.e., a set whose cardinality is the same as the cardinality of **N**.

Can a set be a subset of a set other than itself?

> **Axiom of union:** *For sets A and B there exists a set U such that $A \subset U$ and $B \subset U$.*

The axiom of extension implies that

$$\{x \in U : x \in A \text{ or } x \in B\}$$

is unique. We refer to this set at the **union** of A and B and we denote it by $A \cup B$. With the union of A and B in hand, we may create the subset

$$\{x \in A \cup B : x \in A \text{ and } x \in B\}$$

of $A \cup B$. This set is referred to as the **intersection** of A and B and is denoted by $A \cap B$. If $A \cap B = \varnothing$ we say A and B are **disjoint**.

What is the set-theoretic locale for these budding binary operations *on sets?*

> **Axiom of power:** *For each set U the set consisting of the subsets of U exists.*

This set is denoted by $\mathfrak{P}(U)$ and is referred to as the **power set** of U. The root set is denoted by U to suggest 'universe' because it contains all of the elements necessary to create $\mathfrak{P}(U)$. Be clear on this, U is a 'local' universe because there is no set that contains everything. If such a set existed, it would *contain an element that it does not contain* — the so-called Russell paradox (see [**?**]).

Notice that in the ascension from elements of U to subsets of U the elements of U vanish; i.e., for $a \in U$, $a \notin \mathfrak{P}(U)$. A saving grace is that $\{a\}$ is an element of $\mathfrak{P}(U)$, the so-called **inclusion** of a in $\mathfrak{P}(U)$; i.e., a is not "lost" in this ascent, it is just "renamed." And notice that since each element of $\mathfrak{P}(U)$ is a subset of U, $\mathfrak{P}(U)$ sprouts a natural **unary operation** referred to as **complementation**: $A' := U \sim A$ for each $A \in \mathfrak{P}(U)$. It follows $A' \cup A = U$ and $A' \cap A = \varnothing$.

The word 'power' in the phrase 'power set' is derivative of the fact that if U has exactly n elements, then $\mathfrak{P}(U)$ has exactly 2^n elements. And, as we can compute $2^{(2^n)}$, we can create $\mathfrak{P}^2(U) := \mathfrak{P}(\mathfrak{P}(U))$ and so on. But — the real power of a power set stems from the fact that it provides a richer algebraic playground than U. Indeed, an instance of a power set creating an algebra out of whole cloth is at hand: regardless of the nature of the elements of U — homogeneous or heterogeneous, algebraically fecund or algebraically barren — U **lifts** to an algebra (which is an example of a so-called Boolean algebra) with **signature**

$$[\mathfrak{P}(U), \{[\cup, \varnothing], [\cap, U]\}, \{'\}, \{\subset\}]\,.$$

An algebraic signature always includes the **ground set** ($\mathfrak{P}(U)$) and may include a set of binary operations ($\cup$ and $\cap$), each paired with an **identity** element if one exists (for $A \in \mathfrak{P}(U)$: $A \cup \varnothing = A$ and $A \cap U = A$), or a set of unary operations ($'$) or a set of relations ($\subset$) defined on the ground set. The purpose of a signature is to highlight operations and relations that are relevant to the algebraic quest at hand. It follows that a signature is a living entity and evolves as the algebraic quest evolves — the only constant being the ground set. As our understanding of operations and relations is refined it will be come clear that, in the most general sense, anything appearing in a particular signature, except for the ground set, is redundant because once the ground set is chosen all of the operations and relations are manifest — theoretically speaking. Salient properties of this algebra include the **commutativity**

$$(A \cup B = B \cup A \text{ and } A \cap B = B \cap A),$$

associativity

$$(A \cup B) \cup C = A \cup (B \cup C) \text{ and } (A \cap B) \cap C = A \cap (B \cap C),$$

mutual distributivity

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup B) \text{ and } A \cap (B \cup C) = (A \cap B) \cup (A \cap B))$$

and **idempotence**

$$A \cup A = A \text{ and } A \cap A = A$$

of union and intersection. The interaction of union, intersection and complementation is captured by so-called De Morgan laws:

$$(A \cup B)' = A' \cap B' \text{ and } (A \cap B)' = A' \cup B'.$$

3 Algebra by ascent — a parable

Do you remember learning to play fast and loose with integer arithmetic,

$$[\mathbf{Z}, \{[+, 0], [\cdot, 1]\}, \{-\}, \{<\}]\,,$$

as a child so that clock arithmetic would make 'sense,' where sense included the agreement, dictated by, if not a wizard, then some authority figure, that four hours past 10 o'clock was 2 o'clock and that you shouldn't concern yourself with a.m. or p.m.? Nor, it seems, that you had just agreed that $10+4=2$ or $14=2$ or, maybe worse, that $12=0$ or, falling completely into the arithmetic abyss, that $3\cdot 4=0$ even though neither $3=0$ nor $4=0$.

What might have happened if you forced the issue by asking about the deal on the $3\cdot 4=0$ thing? In elementary school, or at home, the up-shot may have been "because .... just because." If this put you off until your high school algebra course an answer may have been couched in terms of the sterile mechanics of integer arithmetic modulo twelve and, more generally, integer arithmetic modulo any positive integer. If this put you off until an undergraduate abstract algebra course, the answer may have been couched in the esoteric mechanics of homomorphisms of the algebra of integers — more likely referred to as the ring of integers in this context.

Here is what you should have been told so that the previous three answers — well, the second and third answers — would eventually make sense in a more general context. Clock arithmetic requires that the set of integers, which is infinite, be condensed to twelve objects to serve as labels for the relevant positions on the clock face. Given the usual labeling of the clock face, which respects the usual ordering of the integers, we are confronted with the fact that both 13 and one, while not equal as integers are begging to be equivalent as labels. It follows that the integer one appearing on the clock face is just a stand-in for the subset

$$\{...,-23,-11,1,13,25,...\}$$

of $\mathbf{Z}$ — the actual label for the one o'clock position. In the same manner we can conclude that the labels for ten o'clock, twelve o'clock and two o'clock are, respectively,

$$\{...,-14,-2,10,22,34,...\},\{...,-12,0,12,24,36,...\} \text{ and } \{...,-22,-10,2,14,26,...\}.$$

while hoping that

$$\{...,-14,-2,10,22,34,...\}+\{...,-20,-8,4,16,28,...\}=\{...,-20,-8,2,14,26,...\}$$

in some legitimate mathematical sense.

To make subsequent displays more tractable while preserving the gist of our parable, let's imagine a clock of 'circumference' three rather than twelve. In this case the set of labels for three o'clock, one o'clock and two o'clock are, respectively, the set of multiples of three,

$$\overline{0}:=\{...,-3,0,3,6,9,...\}=\{3\cdot m\in\mathbf{Z}:m\in\mathbf{Z}\},$$

the set of multiples of three shifted by one,

$$\overline{1}:=\{...,-5,-2,1,4,7,...\}=\{3\cdot m+1\in\mathbf{Z}:m\in\mathbf{Z}\},$$

and the set of multiples of three shifted by two,

$$\overline{2} := \{..., -4, -1, 2, 5, 8, ...\} = \{3 \cdot m + 2 \in \mathbf{Z} : m \in \mathbf{Z}\}.$$

In analogy with the clock of circumference twelve we are hoping, for example, that two 'hours' past two o'clock is one o'clock; i.e., $\overline{2} + \overline{2} = \overline{1}$.

No matter how this story plays out, its arithmetic narrative must take place in the set algebra

$$[\mathfrak{P}(\mathbf{Z}), \{[\cup, \varnothing], [\cap, \mathbf{Z}]\}, \{'\}, \{\subset\}]\,.$$

To that end we can we enrich this algebra, while incorporating $\{\{z\} \in \mathfrak{P}(\mathbf{Z}) : z \in \mathbf{Z}\}$ as a version of the algebra $\mathbf{Z}$ in which $\mathbf{Z}$'s elements have been renamed as singleton sets, by **lifting** the algebra of $\mathbf{Z}$ to $\mathfrak{P}(\mathbf{Z})$. For $A, B \in \mathfrak{P}(\mathbf{Z})$ *the sum of sets is the set of sums,*

$$A + B := \{c \in \mathbf{Z} : c = a + b \text{ for some } a \in A \text{ and for some } b \in B\},$$

the product of sets is the set of products,

$$A \cdot B := \{c \in \mathbf{Z} : c = a \cdot b \text{ for some } a \in A \text{ and for some } b \in B\},$$

and *the negative of a set is the set of negatives,*

$$-A := \{c \in \mathbf{Z} : c = -a \text{ for some } a \in A\ \}.$$

That is, we can add, multiply and negate sets in natural ways. Notice the overloaded notation for operations in the previous display (and earlier in this parable). For example, + indicates the sum of integers and the sum of sets of integers, so each appearance of + must be read in the appropriate context. This is a bit risky but it can also be quite suggestive because

$$\{a + b\} = \{a\} + \{b\}$$

can be read as *the inclusion of a sum is the sum of the inclusions.* Similar mantras follow from observing that

$$\{a \cdot b\} = \{a\} \cdot \{b\},\ \{-a\} = -\{a\} \text{ and } \{a\} < \{b\} \Leftrightarrow a < b.$$

So, not only is the 'new' $\mathfrak{P}(\mathbf{Z})$,

$$[\mathfrak{P}(\mathbf{Z}), \{[\cup, \varnothing], [\cap, \mathbf{Z}], [+, \{0\}], [\cdot, \{1\}]\}, \{-, '\}, \{\subset\}]\,,$$

algebraically richer than the 'old' $\mathfrak{P}(\mathbf{Z})$,

$$[\mathfrak{P}(\mathbf{Z}), \{[\cup, \varnothing], [\cap, \mathbf{Z}]\}, \{'\}, \{\subset\}]\,,$$

it 'includes' the algebra $\mathbf{Z}$ by simply dressing each integer with set braces. The function

$$\mathbf{Z} \xrightarrow{\iota} \mathfrak{P}(\mathbf{Z}) \text{ defined by } \iota(z) = \{z\}$$

is referred to as an **inclusion morphism**. Morphism because it 'respects,' simultaneously, the operations and relations of $\mathbf{Z}$ and their lifts to the singleton subsets of $\mathfrak{P}(\mathbf{Z})$. A consequence of this respect is that a fair share of useful properties of integer algebra, in particular commutativity and associativity of both $\cdot$ and $+$ and distributivity of $\cdot$ over $+$ apply, not only to the inclusion of $\mathbf{Z}$ in $\mathfrak{P}(\mathbf{Z})$, but to the whole of $\mathfrak{P}(\mathbf{Z})$. Further, the inclusions of 0 and 1 serve as the appropriate identities for the whole of $\mathfrak{P}(\mathbf{Z})$; i.e.,

$$A + \{0\} = A \text{ and } A \cdot \{1\} = A \text{ for } A \in \mathfrak{P}(\mathbf{Z}).$$

On the other hand, there is no guarantee that algebras sharing the same ground set will play well together: for example the interaction of the lifted algebra of $\mathbf{Z}$ with the inherent set algebra of $\mathfrak{P}(\mathbf{Z})$ can be hit ($+$ and $\cdot$ both distribute over $\cup$) or miss (neither $+$ nor $\cdot$ distribute over $\cap$).

That the addition, multiplication and negation operations of the enriched version of $\mathfrak{P}(\mathbf{Z})$ play well together when restricted to the elements of the subset $\Pi := \{\overline{0},\overline{1},\overline{2},\}$ of $\mathfrak{P}(\mathbf{Z})$, thereby creating the algebra

$$\big[\Pi, \{[+,\overline{0}], [\cdot,\overline{1}]\}, \{-\}\big]$$

$+$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{0}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{1}$	$\overline{1}$	$\overline{2}$	$\overline{0}$
$\overline{2}$	$\overline{2}$	$\overline{0}$	$\overline{1}$

$\cdot$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{0}$	$\overline{0}$	$\overline{0}$	$\overline{0}$
$\overline{1}$	$\overline{0}$	$\overline{1}$	$\overline{2}$
$\overline{2}$	$\overline{0}$	$\overline{2}$	$\overline{1}$

$-$	$\overline{0}$	$\overline{1}$	$\overline{2}$
	$\mathbf{0}$	$\overline{2}$	$\overline{1}$

,

is due in large measure to the fact that Π is a **partition of $\mathbf{Z}$**; i.e., Π is an element of $\mathfrak{P}^2(\mathbf{Z})$ whose **components** (elements), no one of which is the empty set, are **mutually disjoint,**

$$\overline{0} \cap \overline{1} = \overline{0} \cap \overline{2} = \overline{1} \cap \overline{2} = \varnothing,$$

and **exhaust $\mathbf{Z}$**,

$$\overline{0} \cup \overline{1} \cup \overline{2} = \mathbf{Z}.$$

For example, $\overline{2} + \overline{2} = \overline{1}$ follows because

$$\begin{aligned} \overline{2} + \overline{2} &= \overline{0} + \{2\} + \overline{0} + \{2\} \\ &= \overline{0} + \overline{0} + \{2\} + \{2\} \\ &= \overline{0} + \{4\} \\ &= \overline{0} + \{3\} + \{1\} \\ &= \overline{0} + \{1\} \\ &= \overline{1} \end{aligned}$$

and $-\overline{1} = \overline{2}$ and $-\overline{2} = \overline{1}$ follow because

$$\begin{aligned} \overline{1} + \overline{2} &= \overline{0} + \{1\} + \overline{0} + \{2\} \\ &= \overline{0} + \{3\} \\ &= \overline{0}. \end{aligned}$$

Our disparagement of the abuse of visual artifices by the RDM/SQL coupling begs a defense of our visual artifices. Consider the use of

$-$	$\overline{0}$	$\overline{1}$	$\overline{2}$
	$\overline{0}$	$\overline{2}$	$\overline{1}$

to represent the mapping of each element of Π to its negative:

$$-(\overline{0}) = \overline{0},\ -(\overline{1}) = \overline{2} \text{ and } -(\overline{2}) = \overline{1}$$

or, subject to the usual notational condensation,

$$-\overline{0} = \overline{0},\ -\overline{1} = \overline{2} \text{ and } -\overline{2} = \overline{1}.$$

Visual artifice or mapping, each of the preceding displays is a summary of a collection of 'couplings,' say top-to-bottom or left-to-right, respectively. In the spirit of overloading notation when appropriate, we define a set (denoted by $\mathfrak{n}$ for negation)

$$\mathfrak{n} := \{\overline{0}.\overline{0}, \overline{1}.\overline{2}, \overline{2}.\overline{1}\}$$

of 'couplets' to turn both the visual artifice for negation and its associated mapping into names of a set-theoretic object. It follows that a statement such as $-\overline{1} = \overline{2}$ is logically equivalent to the statement $\overline{1}.\overline{2} \in \mathfrak{n}$. Trouble: a 'couplet', say $l.r$, and, in turn, the set $\mathfrak{n}$ of couplets are as much visual artifices as are their geneses. Solution: turn $l.r$ into a name for a legitimate set-theoretic construct which distinguishes l from r. We choose to define the **couplet** $l.r$ by

$$l.r := \{\{l\}, \{l, r\}\} \in \mathfrak{P}^2(\Pi)$$

and collect all such couplets in the **Cartesian product** of Π by Π:

$$\Pi \times \Pi := \{l.r \in \mathfrak{P}^2(\Pi) : l \in \Pi \text{ and } r \in \Pi\} \subset \mathfrak{P}^2(\Pi); \text{ i.e., } \Pi \times \Pi \in \mathfrak{P}^3(\Pi).$$

That's right, the operation of negation is a set-theoretic object in its own right — a candidate to be an element of an algebra and, therefore, subject to algebraic operations.

Moreover our definition of couplet (see Kuratowski [**?**]), provides a set-theoretic criterion for testing the equality of two couplets:

$$l.r = a.b \text{ if, and only if, } l = a \text{ and } r = b.$$

Certainly, $l = a$ and $r = b$ imply that $l.r = a.b$. Now assume that $l.r = a.b$ (i.e., $\{\{l\}, \{l, r\}\} = \{\{a\}, \{a, b\}\}$) and keep in mind that while the name of an element of a set may be repeated, the element only occurs once in the set. It follows that $\{l\} = \{a\}$ or $\{l\} = \{a, b\}$. If $\{l\} = \{a\}$, then $l = a$ and $\{l, r\} = \{a, b\} = \{l, b\}$;i.e., $r = b$. If $\{l\} = \{a, b\}$, then $l = a = b = r$. Now, in good mathematical conscience, we may use the visual artifice $l.s$ rather than the forest of braces in its definition.

The visual artifices for the binary operations + and · on Π, so-called operation tables, represent subsets of the Cartesian product $(\Pi \times \Pi) \times \Pi$. The vanilla versions for these subsets (denoted by $\mathfrak{a}$ for addition and $\mathfrak{m}$ for multiplication) are

$$\mathfrak{a} := \left\{ \begin{array}{c} (\overline{0}.\overline{0}).\overline{0},\ (\overline{0}.\overline{1}).\overline{1},\ (\overline{0}.\overline{2}).\overline{2}, \\ (\overline{1}.\overline{0}).\overline{1},\ (\overline{1}.\overline{1}).\overline{2},\ (\overline{1}.\overline{2}).\overline{0}, \\ (\overline{2}.\overline{0}).\overline{2},\ (\overline{2}.\overline{1}).\overline{0},\ (\overline{2}.\overline{2}).\overline{1} \end{array} \right\}$$

and

$$\mathfrak{m} := \left\{ \begin{array}{c} (\overline{0}.\overline{0}).\overline{0},(\overline{0}.\overline{1}).\overline{0},(\overline{0}.\overline{2}).\overline{0}, \\ (\overline{1}.\overline{0}).\overline{0},(\overline{1}.\overline{1}).\overline{0},(\overline{1}.\overline{2}).\overline{2}, \\ (\overline{2}.\overline{0}).\overline{0},(\overline{2}.\overline{1}).\overline{2},(\overline{2}.\overline{2}).\overline{1}\} \end{array} \right\}$$

Overloading notation and seasoning with the axiom of specification yields:

$$\mathfrak{n} := \{l.r \in \Pi \times \Pi : -l = r\} \in \mathfrak{P}(\Pi \times \Pi),$$

$$\mathfrak{a} := \{(l.r).s \in (\Pi \times \Pi) \times \Pi : l.r \in \Pi \times \Pi \text{ and } s = l + r\} \in \mathfrak{P}((\Pi \times \Pi) \times \Pi)$$

and

$$\mathfrak{m} := \{(l.r).s \in (\Pi \times \Pi) \times \Pi : l.r \in \Pi \times \Pi \text{ and } s = l \cdot r\} \in \mathfrak{P}((\Pi \times \Pi) \times \Pi).$$

As with negation, each of addition and multiplication is a set-theoretic entity in its own right and the statements

$$\begin{array}{l} (\overline{1}.\overline{2}).\overline{0} \in \mathfrak{a}, \\ +(\overline{1}.\overline{2}) = \overline{0}, \\ \overline{1} + \overline{2} = \overline{0}, \\ -\overline{1} = \overline{2}, \\ \overline{1}.\overline{2} \in \mathfrak{n} \end{array}$$

are logically equivalent, as are

$$\begin{array}{l} (\overline{2}.\overline{2}).\overline{1} \in \mathfrak{m}, \\ \cdot(\overline{2}.\overline{2}) = \overline{1}, \\ \overline{2}\cdot\overline{2} = \overline{1}. \end{array}$$

In retrospect, and more generally, it is clear that the algebraic playground in which we first discovered the elements of Π was fabricated with Cartesian products:

$$-A := \{c \in \mathbf{Z} : c = -a \text{ for some } a \in A\ \}$$

implies that $\mathfrak{n}$ is a subset of

$$\{A.B \in \mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z}) : B = -A\} \in \mathfrak{P}(\mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z}))$$

and

$$\begin{aligned} A+B &: = \{c \in \mathbf{Z} : c = a+b \text{ for some } a \in \mathbf{Z} \text{ and for some } b \in \mathbf{Z}\} \\ &= \{a+b \in \mathbf{Z} : a.b \in A \times B\} \end{aligned}$$

implies that $\mathfrak{a}$ is a subset of

$$\{(A.B).C \in (\mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z})) \times \mathfrak{P}(\mathbf{Z}) : C = A+B\} \in \mathfrak{P}((\mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z})) \times \mathfrak{P}(\mathbf{Z}))$$

and

$$\begin{aligned} A \cdot B &: = \{c \in \mathbf{Z} : c = a \cdot b \text{ for some } a \in \mathbf{Z} \text{ and for some } b \in \mathbf{Z}\} \\ &= \{a \cdot b \in \mathbf{Z} : a.b \in A \times B\} \end{aligned}$$

implies that $\mathfrak{m}$ is a subset of

$$\{(A.B).C \in (\mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z})) \times \mathfrak{P}(\mathbf{Z}) : C = A \cdot B\} \in \mathfrak{P}((\mathfrak{P}(\mathbf{Z}) \times \mathfrak{P}(\mathbf{Z})) \times \mathfrak{P}(\mathbf{Z})).$$

In short, the mathematical legitimacy of your wizardry with clock arithmetic rests in a tale of sets told by ascending towers of sets of ever increasing cardinality and of "*frightening height and complexity.*"

The morals of the story

An algebra U lifts to an algebra $\mathfrak{P}(U)$ which is respectful of the inclusion of U while enriching the inherent Boolean algebra on $\mathfrak{P}(U)$. Upping the set-theoretic ante by ascending to a power set always provides a richer algebraic game. The subtleties of the game often turn on the answer to a question of the form *Is the 'flarn' of a 'clarp' the 'clarp' of the 'flarn(s)'?.* A positive answer, even if guaranteed by definition as in the case of summing two sets of integers — the sum of sets is the set of sums — usually bodes well. A negative answer may not necessarily bode ill: the complement of a union (intersection) is not the union (intersection) of the complements but it is the intersection (union) of the complements. And sometimes a negative answer bodes better than well: the logarithm of a product is not the product of the logarithms, rather it is the sum of the logarithms; i.e., *the 'flarn' of a 'clarp' is the 'tworble' of the 'flarns'* in this instance means that the relative complexity of multiplication can be reduced to the relative simplicity of addition.

We shall see that data algebras require an ascent from $\mathfrak{P}(U)$ to $\mathfrak{P}^2(U)$, thereby compounding algebraic and notational subtleties. For example, the operation of union (not to mention intersection, complementation or the subset relation) exists on both $\mathfrak{P}(U)$ and $\mathfrak{P}^2(U)$, each instance of which is denoted by the reserved symbol $\cup$. As the lift of addition on $\mathbf{Z}$ to $\mathfrak{P}(\mathbf{Z})$,

$$A+B := \{a+b \in \mathbf{Z} : a.b \in A \times B\}$$

is essential to the construction of the algebra Π, the lift of union from $\mathfrak{P}(U)$ to $\mathfrak{P}^2(U)$ is essential to the construction of data algebras. But, while the symbol $+$ tolerates overloading — we had no preconceived notion of the addition of sets of integers — the symbols $\cup$ and $\cap$ do not because the operations of set union and set intersection occur anew each time we lift to a power set. Instead, we use $\blacktriangledown$ **(cross-union)** and $\blacktriangle$ **(cross-intersection).** For example, if $U = \{a, b, c\}$, then

$$\begin{aligned}\{\{a,b\},\{c\}\}\blacktriangledown\{\{a,c\},\{a\}\}\} &= \{A \cup B \in \mathfrak{P}(U) : A.B \in \{\{a,b\},\{c\}\} \times \{\{a,c\},\{a\}\}\} \\ &= \{\{a,b\} \cup \{a,c\}, \{a,b\} \cup \{a\}, \{c\} \cup \{a,c\}, \{c\} \cup \{a\}\} \\ &= \{\{a,b,c\}, \{a,b\}, \{a,c\}, \{c,a\}\} \\ &= \{\{a,b,c\}, \{a,b\}, \{a,c\}\}\end{aligned}$$

and

$$\begin{aligned}\{\{a,b\},\{c\}\}\blacktriangle\{\{a,c\},\{a\}\}\} &= \{A \cup B \in \mathfrak{P}(U) : A.B \in \{\{a,b\},\{c\}\} \times \{\{a,c\},\{a\}\}\} \\ &= \{\{a,b\} \cap \{a,c\}, \{a,b\} \cap \{a\}, \{c\} \cap \{a,c\}, \{c\} \cap \{a\}\} \\ &= \{\{a\}, \{a\}, \{c\}, \varnothing\} \\ &= \{\{a\}, \{c\}, \varnothing\}.\end{aligned}$$

An algebra U is born of an interaction of power sets derivative of U. Once the elements of Π were selected from $\mathfrak{P}(\mathbf{Z})$ the unary operation of negation on Π was selected from $\mathfrak{P}(\Pi \times \Pi)$ and the two binary operations of emphasis for Π were selected from $\mathfrak{P}((\Pi \times \Pi) \times \Pi)$.

Cartesian products — everywhere Cartesian products. Cartesian products warehouse the complete inventory of operations (binary and unary) and binary relations available on a set. For example the binary relation $<$ on $\mathbf{Z}$ may be characterized as

$$< \ := \{m.n \in \mathbf{Z} \times \mathbf{Z} : m + p = n \text{ for some } p \in \mathbf{N}\}.$$

Cartesians products are used to define lifts of operations and relations. And be on the alert in Section 4: Cartesian products and successive power sets of Cartesian products manifest themselves as ground sets of data algebras and the intimacy of a partition of a set (for example, $\Pi := \{\overline{0},\overline{1},\overline{2},\}$) and certain subsets of a Cartesian product (for example, $\mathbf{Z} \times \mathbf{Z}$) is revealed.

Be warned: Cartesian products must be dealt with delicately because

$$A \times B = B \times A \text{ if, and only if, } A = \varnothing \text{ or } B = \varnothing \text{ or } A = B$$

and

$$(A \times B) \times C = A \times (B \times C) \text{ if, and only if, } A = \varnothing \text{ or } B = \varnothing \text{ or } C = \varnothing.$$

For instance the expression '$A \times B \times C$' (much in the manner of the expression '$6 - 3 - 7$') is notational nonsense unless at least one of A, B or C is empty.

Yet a perceived need for such a construction lingers to this day because Kuratowski's starkly spare construction, $\{\{a\},\{a,b\}\}$, for simultaneously binding and distinguishing the entities a and b lends itself naturally to the appellation "ordered" pair. And once the "ordered" pair terminology is out of the bag, it fairly begs for generalization to "ordered" triple. A seemingly natural choice for the definition of an "ordered" triple, say $a.b.c$, that follows the spirit of $\{\{a\},\{a,b\}\}$ is

$$a.b.c := \{\{a\},\{a,b\},\{a,b,c\}\}.$$

The upside of this choice is that $a.b.c$ and $a.b$ are fellow members of $\mathfrak{P}^2(U)$ for $a,b,c \in U$ — and this fellowship is shared by "ordered" n-tuples defined in this manner. The downside is that without paying some sort of set-theoretic price, this fellowship leads to the confusion of certain "ordered" n-tuples. For example:

$$a.a.a = \{\{a\}\} = a.a \text{ and } a.a.c = \{\{a\},\{a,c\}\} = a.c.$$

Another option is to settle on a recursive definition:

$$a.b.c := \{\{\{\{a\},\{a,b\}\}\},\{\{\{a\},\{a,b\}\},c\}\}.$$

This amounts to placing $a.b.c$ in $(U \times U) \times U$, a steep set-theoretic price indeed — not to mention the facts that "ordered" pairs and "ordered" triples are now consigned to different branches of a power set tower that is growing exponentially and that an "ordered" triple is — alas — an "ordered" pair. It is possible to deal with these problems axiomatically, but that requires layering axioms on top of the axioms of ZFC, a price we are not willing to pay because, among other reasons, it hinders our effort to treat with data in its full generality.

What's a data-algebraist to do? Ignore the begging and never refer to a couplet as an "ordered" pair! Couplets are the fundamental particles of data algebra and when, as revealed in Section 4, collected properly produce "ordered" n-tuples at the same set-theoretic level and at minimal set-theoretic cost, no matter the value of n.

Homogeneity is a red herring and no set is algebraically barren. Look about you, select three disparate objects, say x, y and z, and collect them in the set $U = \{x,y,z\}$. The objects x, y and z are suddenly homogeneous by virtue of their common 'U-ness.' Now select a couple of elements of $\mathfrak{P}((U \times U) \times U)$, say $\oplus$ and $\odot$, that are binary operations on U and an element of $\mathfrak{P}((U \times U))$, say $\ominus$, that is a unary operation on U. The algebra

$$[U,\{\oplus,\odot\},\{\ominus\}]$$

is up and running. Indeed a set as minuscule as U supports

$$\binom{3^9}{2} \cdot 3^3 = 5,229,910,881$$

algebras with a signature of this form — one of which is essentially the same as Π. So, the rub lies neither with homogeneity nor with barrenness, but with choosing an appropriate algebra from the surfeit available. A bit of set-theortic foresight in constructing the ground set goes a long way towards making an appropriate choice.

4 Data algebra

We now admit to a bout of parochialism: our use of terminology such as *binary relation*, *binary operation*, *unary operation* and *function* reflects their conventional usage so as to make the algebraic discussion of the preceding section as familiar as possible. The launch of data algebra proper requires a lexicon which serves more specific ends — by becoming more general — while having no deleterious impact on the parochial usage.

To that end it is more convenient to write a couplet

$$i.a := \{\{i\}, \{i, a\}\} \in I \times A$$

in the form i^a: the **yin** of i^a is i ($\mathbf{yin}(i^a) = i$) and the **yang** of i^a is a ($\mathbf{yang}(i^a) = a$). A subset $\mathcal{R}$ of $I \times A$ is referred to as a **relation from I to A**, often summarized symbolically as $\mathcal{R} : I \longrightarrow A$. We will engage in a traditional abuse of notation by treating the expressions $i^a \in \mathcal{R}$ and $i\mathcal{R}a$ as equivalent. And yes, we did say yin and yang. This choice of terminology is to suggest a natural dualism between the 'components' of a couplet, a suggestion readily compromised by other seemingly natural terms such as, say, value and attribute. Accordingly, yin may become yang and yang may become yin: the couplet $a^i \in A \times I$, often denoted as $\overleftrightarrow{i^a}$, is the **transpose** of i^a. The relation

$$\overleftrightarrow{\mathcal{R}} := \{a^i \in A \times I : \overleftrightarrow{a^i} \in I \times A\}$$

from A to I is the **transpose of $\mathcal{R}$**. If $\mathcal{R} = \overleftrightarrow{\mathcal{R}}$, then $\mathcal{R}$ is said to be **symmetric**. If $i^a \in \mathcal{R}$ and $i^b \in \mathcal{R}$ imply that $a = b$, then $\mathcal{R}$ is said to be **yin functional** and we may write $\mathcal{R}(i) = a$. $\mathcal{R}$ is said to be **yang functional** if $\overleftrightarrow{\mathcal{R}}$ is yin functional. The **yin set of $\mathcal{R}$** is

$$\mathbf{yin}(\mathcal{R}) := \{i \in I : i^a \in \mathcal{R} \text{ for some } a \in A\},$$

the **yang set of $\mathcal{R}$** is $\mathbf{yang}(\mathcal{R}) := \mathbf{yin}(\overleftrightarrow{\mathcal{R}})$ and

$$\mathcal{R} \subset \mathbf{yin}(\mathcal{R}) \times \mathbf{yang}(\mathcal{R}) \subset I \times A.$$

We hasten to emphasize that neither $\mathbf{yin}(\mathcal{R}) = I$ nor $\mathbf{yang}(\mathcal{R}) = A$ are presumed. If $\mathbf{yin}(\mathcal{R}) = I$ we may refer to $\mathcal{R}$ as **full**. If $\mathbf{yang}(\mathcal{R}) = A$ we may refer to $\mathcal{R}$ as **onto**. A full yin functional relation from I to A is a **function** from I to A. In this case, if $i^a \in \mathcal{R}$, we may write $\mathcal{R}(i) = a$.

When convenient, and without loss of generality, we may assume that $I = A = U$ because

$$\begin{aligned} I \times A &\subset (I \times I) \cup (I \times A) \cup (A \times I) \cup (A \times A) \\ &= (I \cup A) \times (I \cup A). \end{aligned}$$

In this case we refer to a subset of $U \times U$ as a **relation on U**. A **unary operation on U** is a yin functional relation on U and a **binary operation on**

U is a yin functional relation from $U \times U$ to U. A relation $\mathcal{R}$ on U is **reflexive** if $u^u \in \mathcal{R}$ for each $u \in U$ and **transitive** if $i^x, x^a \in \mathcal{R}$ implies that $i^a \in \mathcal{R}$.

The relation

$$\mathcal{D}_u := \{i^a \in U \times U : i = a\},$$

that is to say the **equality** relation, is referred to as the **diagonal** of U. There is a natural one-to-one correspondence between the subsets of $\mathcal{D}_u$ and the subsets of U,

$$\{x^x, y^y, \ldots, z^z\} \longleftrightarrow \{x, y, \ldots, z\}.$$

Among the subsets of $\mathcal{D}_u$ are the relations

$$\mathcal{D}_{\mathbf{yin}(\mathcal{R})} := \{i^i \in U \times U : i \in \mathbf{yin}(\mathcal{R})\} \text{ and } \mathcal{D}_{\mathbf{yang}(\mathcal{R})} := \{a^a \in U \times U : a \in \mathbf{yang}(\mathcal{R})\},$$

the **yin diagonal of** $\mathcal{R}$ and the **yang diagonal of** $\mathcal{R}$, respectively.

Those supersets of $\mathcal{D}_u$ which are *symmetric, transitive* and, necessarily, *reflexive* are referred to as **equivalence relations on** U. Equivalence relations generalize the relation of equality (i.e., $\mathcal{D}_u$), thereby serving to condense inherent redundancy among elements of a set in a rigorous manner. Recall our brush with clock arithmetic: the integers, $\mathbf{Z}$, were essential — but overly plentiful — for our purposes. A set with only three elements — the partition $\Pi = \{\overline{0},\overline{1},\overline{2}\}$ of $\mathbf{Z}$ — sufficed. How was $\mathbf{Z}$ condensed to Π? In essence we used the 'circumference' of the clock to define the equivalence relation $\simeq$ on $\mathbf{Z}$ by

$$\simeq \ := \{i^a \in \mathbf{Z} \times \mathbf{Z} : (a - i)/3 \in \mathbf{Z}\}.$$

The reflexivity, symmetry and transitivity of $\simeq$ yielded the **equivalence classes**

$$\begin{aligned}
\overline{0} &: = \{z \in Z : z = 3 \cdot m \text{ such that } m \in \mathbf{Z}\},\\
\overline{1} &: = \{z \in Z : z = 3 \cdot m + 1 \text{ such that } m \in \mathbf{Z}\},\\
\overline{2} &: = \{z \in Z : z = 3 \cdot m + 2 \text{ such that } m \in \mathbf{Z}\},
\end{aligned}$$

which, when collected as a set, revealed Π as a partition of $\mathbf{Z}$. It is also worth observing that the components of this partition of $\mathbf{Z}$ yield the partition

$$\{\overline{0} \times \overline{0}, \overline{1} \times \overline{1}, \overline{2} \times \overline{2}\}$$

of $\simeq$. Summary: *An equivalence relation on a set begets a partition of that set.*

Our specification of the components of the partition Π of $\mathbf{Z}$ induced by the equivalence relation $\simeq$ on $\mathbf{Z}$ is meant to reflect the commonality, if not the equality, of the elements of each component of Π: the elements of $\overline{0}$, $\overline{1}$ and $\overline{2}$ share common remainders of zero, one and two, respectively, upon division by three. Thus we may construct a yin functional relation

$$\mathcal{R} := \{i^{\mathbf{c}} \in \mathbf{Z} \times \Pi : i \in \mathbf{c}\}$$

from $\mathbf{Z}$ to Π which, while not an equivalence relation, conveys the same information as the partition Π. Summary: *A partition of a set begets a yin functional relation from that set to the partition.*

Now we come full circle, equivalence relation to equivalence relation, by observing that

$$i^{\mathbf{c}}, a^{\mathbf{c}} \in \mathcal{R} \Longleftrightarrow i, a \in \mathbf{c} \Longleftrightarrow (a - i)/3 \in \mathbf{Z} \Longleftrightarrow i \simeq a.$$

Summary: *A yin functional relation on a set begets an equivalence relation on that set.*

Lexicon in hand, we need an inventory of data algebras in which one can do data algebra.

Couplets(G)

It is abundantly clear that our reference to 'real-world' data as *apparently, neither homogenous nor necessarily algebraic* in the second paragraph of the introduction is a straw man. Homogeneity of datums is created by collecting them in a set H and 'algebraicness' is created as we ascend various towers derivative of H for the purpose of exposing sets of couplets to serve as operations and relations for a budding algebra. What makes an algebra a data algebra is that we do not delay the introduction of couplets until we need operations or relations, rather we begin our ascent with a set of couplets. For example, if you 'say' 35 to anybody but a pure mathematician, they respond by asking for context: "Huh, 35 'whats' — grams, meters, minutes?" This request reflects the fact that in the real world we converse in couplets as simple as 35^{meters}, or, for that matter, $meters^{35}$, or as complicated as, if not more complicated than,

$$\{\{\text{"Ernst Zermelo"}^{\text{Name}}, \text{"Abraham Fraenkel"}^{\text{Name}}\}^{\text{developers}}\}^{\text{ZFC}}.$$

A suggestive way to think of this is that data algebra generalizes, while making rigorous, the inchoate algebra of unit analysis you wrestled with in your first chemistry course. In any case a reference to data is a reference to set-theoretic structure.

Accordingly we assume the existence of a non-empty **genesis set** G and construct the set $G \times G$ to serve as an inventory of couplets. For example, if our goal is to make mathematical sense out of the RDM/SQL coupling we require that G be a mixture of integers, floating-point, strings, datetime and so on. In this case, and in general, relevant operations and relations on G may be lifted to operations and relations on $G \times G$ since $G \times G \subset \mathfrak{P}^2(G)$. And in lieu of ransacking the subsets of $((G \times G) \times (G \times G)) \times (G \times G)$ and $(G \times G) \times (G \times G)$ in hope of stumbling across additional useful operations on $G \times G$ — a fool's errand given the number of such subsets — we turn to the set-theoretic structure of the elements of $G \times G$.

Consider the couplets a^b and b^c. Reading a^b as b is 'assigned' to a and b^c as c is 'assigned' to b, we may interpret (and we do) a^b composed with b^c as c is 'assigned' to a; i.e., $a^b \circ b^c := a^c$. The only rub is that if we consider $a^b \circ c^d$ where $b \neq c$, then we can not, in good set-theoretic conscience, define

$a^b \circ c^d$. Nevertheless, $\circ$ is a binary operation — conventional terminology begs the adjective 'partial' — on $G \times G$ which we refer to as **composition**. As it turns out many, if not most, of the applicable operations — binary or unary — of data algebra are 'partial' by design or due to the nature of the genesis set. The fact that a couplet has a transpose serves up transposition as a natural unary operation on $G \times G$: $\overleftrightarrow{a^b} = b^a$. Using the integers, the mother of all algebras, as a guide we should ask, in analogy with the fact that *the negative of a sum is the sum of the negatives,* about the interplay between composition and transposition. Asking for the transposition of a composition of couplets which is not defined is a non sequitur. But the transposition of a composition of couplets that is defined is the composition, in the opposite order, of the transposes of the original couplets. For example

$$\overleftrightarrow{a^b \circ b^c} = \overleftrightarrow{a^c} = c^a = c^b \circ b^a = \overleftrightarrow{b^c} \circ \overleftrightarrow{a^b}.$$

Couplets(G) is our name for the algebra with ground set $G \times G$ — a relation — and signature

$$[G \times G, \{\circ\}, \{\leftrightarrow\}].$$

Note that **Couplets**(G) is not a set algebra in the usual sense; i.e., $G \times G \subsetneq \mathfrak{P}^2(G)$. Fleshing out the signature depends on the set-theoretic structure of G.

Relations(G)

Relations(G) (see Tarski [**?**]) is our name for the lift of **Couplets**(G); i.e., **Relations**(G) is the algebra with ground set $\mathfrak{P}(G \times G)$ and signature

$$[\mathfrak{P}(G \times G), \{[\circ, \mathcal{D}_G]\}, \{\leftrightarrow\}]\,.$$

As the ground set of **Relations**(G) is a power set we may take the binary operations of union and intersection, the unary operation of complementation and the subset relation as implicit — recall that the identities for union and intersection are $\varnothing$ and $G \times G$ respectively. Of course the symbols $\circ$ and $\leftrightarrow$ appearing in this signature are overloaded to represent the lifts of composition and transposition, respectively, from **Couplets**(G). To be precise, for $\mathcal{A}, \mathcal{B} \in \mathfrak{P}(G \times G)$,

$$\mathcal{A} \circ \mathcal{B} := \{x^z \in G \times G : x^z = x^y \circ y^z \text{ for some } x^y \in \mathcal{A} \text{ and } y^z \in \mathcal{B}\}$$

and

$$\overleftrightarrow{\mathcal{A}} := \{\overleftrightarrow{x^y} \in G \times G : x^y \in \mathcal{A}\}.$$

One can easily check that $\mathcal{D}_G \circ \mathcal{A} = \mathcal{A} = \mathcal{A} \circ \mathcal{D}_G$; i.e., $\mathcal{D}_G$ serves as the identity for composition.

It is worth noting how the rough edges of **Couplets**(G) are smoothed in its lift to **Relations**(G). Discomfiture with the undefined couplet composition

$a^b \circ c^d$ (where $b \neq c$) in **Couplets**(G) is ameliorated by the fact that the composition of the inclusions of a^b and c^d is defined in **Relations**(G) by, $\{a^b\} \circ \{c^d\} = \varnothing$, as is the composition of any two relations on G. And, in any case, $\overleftrightarrow{\{a^b\} \circ \{c^d\}} = \overleftrightarrow{\{c^b\}} \circ \overleftrightarrow{\{b^a\}}$ follows from a more general result to the effect that for $\mathcal{A}, \mathcal{B} \in \mathfrak{P}(G \times G)$, $\overleftrightarrow{\mathcal{A} \circ \mathcal{B}} = \overleftrightarrow{\mathcal{B}} \circ \overleftrightarrow{\mathcal{A}}$; i.e., the transposition of a composition is the composition of the transpositions taken in the opposite order. These instances of a lift curing algebraic hiccups is characteristic of the mathematical power of lifting. Indeed, a succession of such lifts took you from the counting numbers of kindergarten to the complex numbers of high school. Each step along the way involved a lift motivated by unsolvable equations. For example, an encounter with $2x = 3$ in the integers spawned the formal expression $3/2$ — indeed a couplet — and a litany of mechanical rules for manipulating such couplets, all of which are derivative of the algebra of $\mathbf{Z}$ lifted to $\mathfrak{P}(\mathbf{Z} \times \mathbf{Z})$.

A taste of **Relations**(G):

- Transposition distributes over union and intersection:

$$\overleftrightarrow{\mathcal{A} \cup \mathcal{B}} = \overleftrightarrow{\mathcal{A}} \cup \overleftrightarrow{\mathcal{B}} \text{ and } \overleftrightarrow{\mathcal{A} \cap \mathcal{B}} = \overleftrightarrow{\mathcal{A}} \cap \overleftrightarrow{\mathcal{B}}.$$

- Composition is not commutative: for distinct $a, b, c \in G$

$$\{a^b, b^c\} \circ \{c^a\} = \{b^a\} \text{ but } \{c^a\} \circ \{a^b, b^c\} = \{c^b\}.$$

- Composition is associative

$$\mathcal{A} \circ (\mathcal{B} \circ \mathcal{C}) = (\mathcal{A} \circ \mathcal{B}) \circ \mathcal{C}$$

  and distributes over union from the left and the right. For example,

$$\mathcal{A} \circ (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \circ \mathcal{B}) \cup (\mathcal{A} \circ \mathcal{C}).$$

- Composition *sub-distributes* over intersection from the left and the right. For example,

$$(\mathcal{A} \cap \mathcal{B}) \circ \mathcal{C} \subset (\mathcal{A} \circ \mathcal{C}) \cap (\mathcal{B} \cap \mathcal{C}).$$

- $\mathcal{A}$ is transitive if, and only if, $\mathcal{A} \circ \mathcal{A} \subset \mathcal{A}$.

- Union and composition spawn a family of useful binary operations on **Relations**(G). For example,

$$\mathcal{A} \cup_{\mathcal{C}} \mathcal{B} := \mathcal{A} \cup \mathcal{B} \text{ if, and only if, } \mathcal{A} \circ \mathcal{C} = \mathcal{B} \circ \mathcal{C}$$

  defines a conditional version of union. What is really at play here? $\mathcal{C}$ is used to induce a full yin functional relation on $\mathfrak{P}(G \times G)$; i.e., a function $f_{\mathcal{C}}$ from $\mathfrak{P}(G \times G)$ to $\mathfrak{P}(G \times G)$ defined by $f_{\mathcal{C}}(\mathcal{A}) := \mathcal{A} \circ \mathcal{C}$. In turn $f_{\mathcal{C}}$ induces an equivalence relation, say $\simeq$, on $\mathfrak{P}(G \times G)$. The union of two relations

is taken if, and only if, the relations are are both in the same component of the partition of $\mathfrak{P}(G \times G)$ induced by $\simeq$. It follows that $\cup_{\mathcal{D}_G}$is $\cup$. Since composition is not commutative we may define a similar family using composition on the left. These families of binary operations on $\mathfrak{P}(G \times G)$ serve to provide the basis for a rigorous mathematical formulation of the RDM/SQL notion of 'joining'.

- Feel free to replace union with intersection (or a binary operation of your choice) in the previous bullet item.

- An element $\mathcal{P}$ of **Relations**(G) is said to be a **permutation of** G if both $\mathcal{P}$ and $\overleftrightarrow{\mathcal{P}}$ are functions from G to G. The set of all permutations of G is referred to as **the full symmetric group on the symbols of** G, and has signature

$$[\mathbf{Sym}(G), \{[\circ, \mathcal{D}_G]\}, \{\leftrightarrow\}] .$$

  It follows that we may bring the full power of what is known as *group actions* to bear on **Relations**(G) and successive lifts of **Relations**(G), to include a group theoretic approach to recognizing patterns in data.

- The notion of an "ordered" n-tuple is encompassed by **Relations**(G): assuming, without loss of generality, that $\{1, 2, \ldots, n\} \subset G$, each yin function from $\{1, 2, \ldots, n\}$ to G is referred to as on **ordered n-tuple from** G. The mysterious case of '$A \times B \times C$' is resolved by assuming that $A \cup B \cup C \subset G$ and $n \geq 3$ and taking '$A \times B \times C$' to be

$$\{\{1^a, 2^b, 3^c\} \in \mathbf{Relations}(G) : a \in A, b \in B \text{ and } c \in C\}.$$

  An ordered 2-tuple — a set of two couplets — is referred to as an **ordered pair**.

Clans(G)

Clans(G) is our name for the lift of **Relations**(G); i.e., **Clans**(G) is the algebra with ground set $\mathfrak{P}^2(G \times G)$ and signature

$$\left[\mathfrak{P}^2(G \times G), [\circ, \{\mathcal{D}_G\}]\}, \{\leftrightarrow\}\right] .$$

Permit us a final spate of redundancy for the sake of emphasis: as the ground set of **Clans**(G) is a power set we may take the binary operations of union and intersection, the unary operation of complementation and the subset relation as implicit. Of course the symbols $\circ$ and $\longleftrightarrow$ appearing in this signature are overloaded to represent the lifts of composition and transposition, respectively, from **Relations**(G). To be precise, for $\mathbb{A}, \mathbb{B} \in \mathfrak{P}^2(G \times G)$,

$$\mathbb{A} \circ \mathbb{B} := \{\mathcal{A} \circ \mathcal{B} \in \mathfrak{P}(G \times G) : \mathcal{A}, \mathcal{B} \in \mathfrak{P}(G \times G)\}$$

and

$$\overleftrightarrow{\mathbb{A}} := \{\overleftrightarrow{\mathcal{A}} \in \mathfrak{P}(G \times G) : \mathcal{A} \in \mathbb{A}\}.$$

One can easily check that $\{\mathcal{D}_G\} \circ \mathbb{A} = \mathbb{A} = \mathbb{A} \circ \{\mathcal{D}_G\}$; i.e., $\{\mathcal{D}_G\}$ serves as the identity for clan composition. A clan $\mathbb{C}$ is said to be yin (yang) functional if each of its elements is yin (yang) functional. The yin set of $\mathbb{C}$ and the yang set of $\mathbb{C}$ are

$$\mathbf{yin}(\mathbb{C}) := \bigcup_{\mathcal{R} \in \mathbb{C}} \mathbf{yin}(\mathcal{R}) \text{ and } \mathbf{yang}(\mathbb{C}) := \bigcup_{\mathcal{R} \in \mathbb{C}} \mathbf{yang}(\mathcal{R})$$

respectively. It follows that the yin diagonal $\mathcal{D}_{\mathbf{yin}(\mathbb{C})}$ (yang diagonal $\mathcal{D}_{\mathbf{yang}(\mathbb{C})}$) of $\mathbb{C}$ is the union of the yin (yang) diagonals of its elements.

A taste of $\mathbf{Clans}(G)$:

- Transposition distributes over union and intersection.
- Composition is not commutative.
- Composition is associative and distributes over union from the left and the right.
- Composition sub-distributes over intersection from the left and the right.
- The conditional versions of union and intersection on $\mathbf{Relations}(G)$ lift to $\mathbf{Clans}(G)$. For example,

  $$\mathbb{A} \cup_{\{\mathcal{C}\}} \mathbb{B} := \{\mathcal{A} \cup_{\mathcal{C}} \mathcal{B} \in \mathfrak{P}(G \times G) : \mathcal{A} \in \mathbb{A} \text{ and } \mathcal{B} \in \mathbb{B}\}.$$

  It is easy to check that, in general,

  $$\begin{aligned} \mathbb{A} \cup_{\{\varnothing\}} \mathbb{B} &= \{\mathcal{A} \cup \mathcal{B} \in \mathfrak{P}(G \times G) : \mathcal{A} \in \mathbb{A} \text{ and } \mathcal{B} \in \mathbb{B}\} \\ &\neq \mathbb{A} \cup \mathbb{B}. \end{aligned}$$

- As it turns out $\cup_{\{\varnothing\}}$ and $\cap_{\{\varnothing\}}$ are just the lifts of union and intersection from $\mathbf{Relations}(G)$ to $\mathbf{Clans}(G)$; i.e., $\cup_{\{\varnothing\}} = \blacktriangledown$ and $\cap_{\{\varnothing\}} = \blacktriangle$ and both are important enough to warrant special names and notation: $\cup_{\{\varnothing\}}$is referred to as **cross-unio**n and denoted by $\blacktriangledown$ and $\cap_{\{\varnothing\}}$ is referred to as **cross-intersection** and denoted by $\blacktriangle$, respectively. The identity element for $\blacktriangledown$ is $\{\varnothing\}$ and the identity element for $\blacktriangle$ is $\{G \times G\}$. Both $\blacktriangledown$ and $\blacktriangle$ are commutative and, among a litany of other technical results, is the fact that each sub-distributes over the other.
- $\mathbf{Sym}(G)$ is included in $\mathbf{Clans}(G)$ as

  $$[\{\{\mathcal{P}\} : \mathcal{P} \in \mathbf{Sym}(G)\}, \{[\circ, \{\mathcal{D}_G\}]\}, \{\leftrightarrow\}]\,.$$

  With this notational formalism on record, it is promptly ignored and the inclusion of $\mathbf{Sym}(G)$ in $\mathbf{Clans}(G)$ is referred to as $\mathbf{Sym}(G)$. The only bump in the ascent is recognizing, say, $\mathcal{P} \circ \mathbb{A}$ as a stand-in for $\{\mathcal{P}\} \circ \mathbb{A}$ — a small price to pay for bringing the action of $\mathbf{Sym}(G)$ to bear on $\mathbf{Clans}(G)$.

Notice, in retrospect, that we now have a working template of data algebras which has the property that each algebra is "included" in the succeeding algebra:

$$\textbf{Couplets}(G) \longrightarrow \textbf{Relations}(G) \longrightarrow \textbf{Clans}(G)$$

This template, while sufficient for purposes of this introduction to data algebra, may be extended ad infinitum. For example, ascending to $P^3(G \times G)$ yields the algebra we refer to as $\textbf{Hordes}(G)$.

5 Coda

The need for a 'legitimate algebra of data' was first broached with one of the authors (Sherman) several years ago by the founders of Algebraix Data Corporation. They bemoaned the absence of a rigorous mathematical foundation for existing data analysis software and proffered a 'generalized set theory' — so-called *Axiomatic Extended Set Theory*, expounded by D. L. Childs [**?**], as the foundation for such an algebra. My immediate response was that if the data-world, and whatever is going on there, has a mathematical foundation then that foundation is derivative of ZFC and does not require an 'extension' of set theory. After some more back and forth we agreed that I would spend some time gaining a pure mathematician's perspective on what goes on in data-world and report back to them.

Long story short: The authors of the texts and papers I read were uniformly enthusiastic — throughout their introductions — about the foundational role set theory should, and would, play in their treatments of data. In the event, they ignored set theory — except to abuse its terminology in the process of concocting something unrecognizable as rigorous mathematics. For example the seminal paper *A Relational Model of Data for Large Shared Data Banks* by E. F. Codd invokes the notion of a relation before incarnating that notion as a table and referring to the table as an n-ary relation (a table with n columns). Codd then takes tables as his 'algebraic' objects, subjects these visual artifices to ad hoc 'algebraic' rules and wobbles forth while leaning on a troubling allusion to the Cartesian product. To Childs's [**?**] credit he recognized, even before the publication of Codd's paper, that the tabular artifice sundered the natural coupling of values with attributes — precisely the mathematical property the artifice was meant to convey. Unfortunately Childs decided to rectify this problem by 'generalizing' ZFC to accomplish what it had already accomplished: providing couplets as fodder for what we refer to as $\textbf{Relations}(G)$. Conclusion: 'extended set theory' is a mathematical non sequitur.

This paper is our contention that Childs, Codd and those who followed did not recognize what is hiding in plain sight: data algebra, as a process, and data algebras, as entities, are inevitable consequences of requiring a mathematically rigorous answer to the question: What's data? Tried-and-true set-theoretic principles inherent to the construction of all modern algebras, as exemplified by modular arithmetic, serve as a guide for creating data algebras as entities

(ground sets in conjunction with selected operations and relations) such as couplet algebra, relation algebra and clan algebra, and associated subalgebras, in which one can do the process of data algebra rigorously.

In a sequel to this paper we will discuss applications of data algebra and contend that there is no such thing as unstructured data.

Acknowledgment. The authors wish to acknowledge the contribution of the employees of Algebraix Data Corporation who participated in courses we taught on Data Algebra. Among those, special thanks to Wes Holler, Bill Rogers, Joe Eaton and Gerhard Fiedler for their critical reviews of earlier drafts of this paper.

References

[1] Childs, D. L. Feasibility of a set-theoretical data structure — a general structure based on a reconstituted definition of relation, *Proc. IFIP Cong.*, 1968, North Holland Pub. Co., Amsterdam, pp. 162-172.

[2] Childs, D. L. Axiomatic Extended Set Theory, Unpublished, 1992.

[3] Codd, E. F. A Relational Model of Data for Large Shared Data Banks, *Comm. ACM* 13, **6**: (June 1970), pp. 377-387.

[4] Halmos, P. R. *Naive Set Theory*, Van Nostrand 1960.

[5] Kuratowski, K. Sur la notion de l'ordre dans la théorie des ensembles, *Fundamenta Mathematicae*, **2**: (1921), pp. 129-131.

[6] Suppes, P. *Axiomatic Set Theory*, Dover 1972.

[7] Tarski, A. On the calculus of relations, *Journal of Symbolic Logic,* **6**: (1941), pp. 73-89.