Leonard Coster
* 1967 New South Wales, Australia | Vienna, Austria

Leonard Coster is a physicist and mathematician. In the Data Loam project, he focused on modeling and writing software to allow a more fluid and fuzzy logic interaction between long form text data objects. By modeling them as a physical system and defining their modes of interaction, a process of self-organisation followed, revealing emergent patterns and connections driven by the data itself. This process is both scalable and algorithmic, meaning that it does not require human guidance or effort and can be deployed on larger computing hardware to match larger data organisational tasks.

Matthias Strohmaier
* 1974 Vienna, Austria | Vienna, Austria

Matthias Strohmaier is a freelance software developer based in Vienna with a focus on architecture and development of custom solutions in the field of interactive multimedia installations, hardware-near programming and web applications. Over the years he has continuously developed cutting edge solutions for media artists and tailor-made software for pioneering tech companies.

encryption [ɪnˈkrɪpʃ(ə)n]
false sense of security.

The information age has exploded

The overwhelming amount of data and the exponentially advancing rate at which it is being generated is a precious, but increasingly inaccessible resource.

Traditional cataloguing systems are, by definition, trapped in the past: simply unable to describe the innovations of today in the terms of yesteryear. Nor can curators be expected to manually read, comprehend and effectively catalogue even the daily production of new material. Hugely powerful computerised search systems can now certainly provide a lifetime of reading from a single keyword search, but not with any meaningful way to refine or explore.

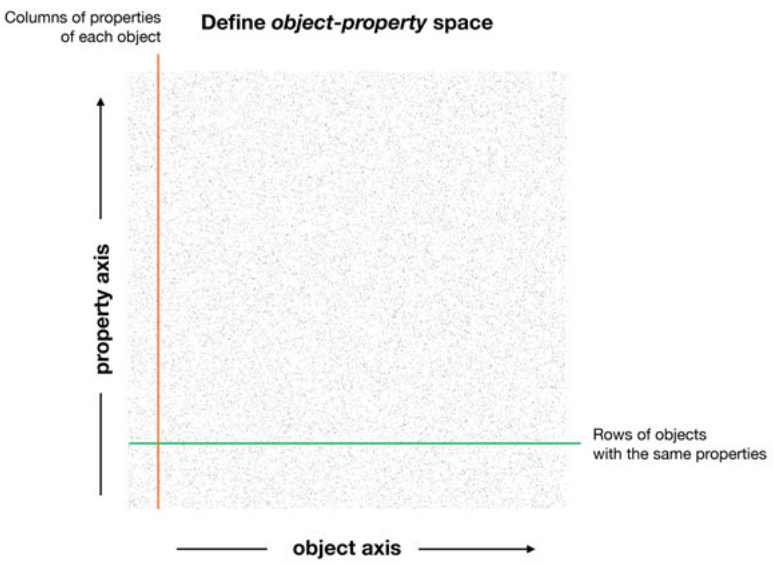
It is as if we know everything but cannot remember anything.

This research develops a new information research methodology with the following features:

- it must be algorithmic and realised in computer code if simply to address the volume of present data.
- it must be inherently unbiased and non-suppressive.
- it must not depend on any pre-defined or finite indexing system, but rather discern distinguishing features between data objects from those objects themselves.
- it must provide the user a means of navigating big data in a controlled, efficient and exhaustive, but assisted and direct-able way.

Object-property space

The first step in realising this objective is to abstract from the type of information and the terms used to describe it. Whatever system we develop must be equally applicable to any kind of information from long form text, magnetic tape and ephemera to smells and mixtures of them all. Let us refer to these discrete pieces of information as ‘data objects’. Next, we must describe these objects in full and complete detail in such a way that they may express their subtleties, differences and similarities. An unbounded and self evolving set of unique and ideally atomic ‘properties’ derived from the objects themselves would be ideal in terms of exhaustively tagging the objects. Examples of the mechanisms for achieving this will be discussed in detail later.



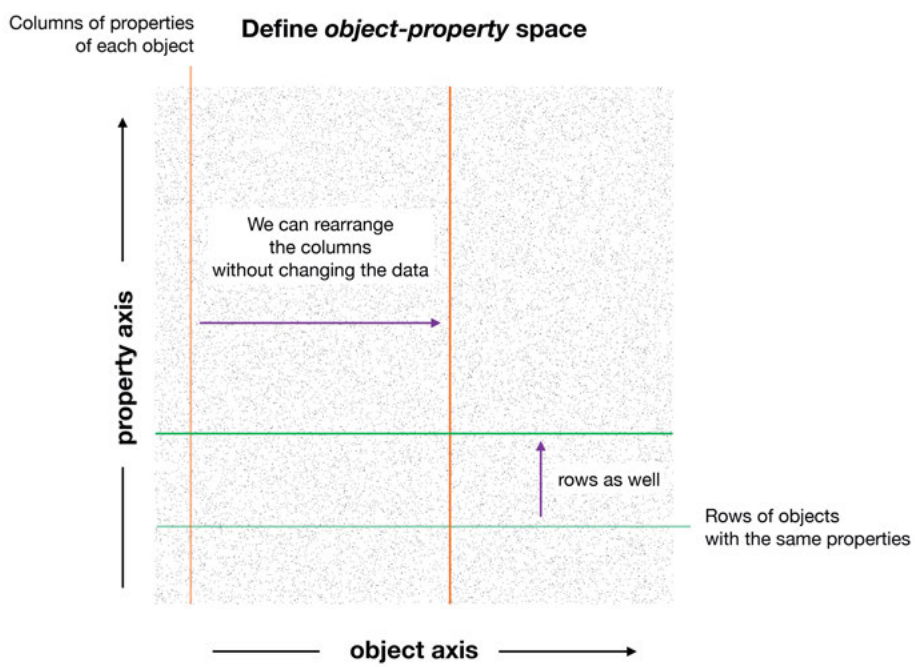
Each point in this space represents the intersection of a particular object and a particular property. Following straight lines parallel with the axes reveals either the set of properties for a certain object or orthogonally, the set of objects with a certain property.

These are **Radical Particles.**

We can now model this as a system, defining rules for interaction between the particles finally resulting in their movement and 'self-organisation' within this new object-property space. This is the ultimate goal of the project–to develop a system which allows even enormous data sets to self-organize and through this, provide a means of searching and exploring the data landscape in a more effective way.

Now define a space with orthogonal object and property axes.
Keeping the system generic with respect to the type of data is of increasing importance considering the prevalence of multi- and mixed-media data available.
Some type of properties could themselves be considered as orthogonal in the sense that it is difficult to directly relate say fragrance and colour “values” and although they have abstracted descriptive analogues in language—blue, orange, sweet, acrid etc, these words are not the property themselves, although these words themselves may also be properties somehow attached to objects along with, say, the smell.
Similarly, objects may fall into orthogonal categories with respect to each other—as discussed, audio recordings vs text vs smells and some objects may even be considered properties of others.
To account for this without bias, we simply increase the dimensionality of the object-property space to an arbitrary degree depending on the collection(s) of data and this should only improve the subtlety and interactivity of the final model.

To visualise this more easily we will limit the discussion to a simpler two-dimensional example for the moment.



The radical matter particles in each row or column have a special relationship to each other since they describe either objects with the same property or properties with the same object. This relationship is immutable in the sense that our analysis and organisation of the space may not alter the data or the assignment of properties and therefore translates into the rules of the system. Each row or column is free to move, but must move as a unit, carrying with it all the particles in that row or column. Further, any move must be exactly parallel to the defined axes.

In plain terms, the objects may rearrange themselves into any order, but the process will not change which properties they have; similarly, the properties may rearrange themselves along their own axis but they remain assigned to the same objects as they always were.

Finally, we require some observable parameter against which we can measure our progress both in designing suitable rules for rearranging the rows and columns and for determining when the processes may be deemed complete.

Since we are working towards an observable organisation, one suitable and calculable parameter might be the nett entropy of the space. By tuning the rules of interaction and movement decision making to produce at each moment the maximum reduction in nett entropy we can ensure that we are organising the space with maximal efficiency. Entropy will never reduce to zero, but as it approaches a *minima*, presumably asymptotically, we can measure our progress.

enigma [ɪˈnɪgmə]

code for the yet to be cracked. sometimes, also: code for that which is already cracked.

A sample landscape

The initial data set chosen for this project was a section of the English language dump from Wikipedia. Approximately 300,000 pages (of approximately seven million) were extracted arbitrarily from the dump file and the words in the text used as the properties to identify them. These were matched against an English language lexicography containing approximately 500,000 words in 120,000 synonym sets (dog, hound, canine *etcetera*). This starter set would later be augmented by new words as they were found in the scanned documents.

In this example, the data *objects* are long form text documents (that is, Wikipedia pages) and the *properties* are words.

Every text document, including the one you are reading now, is comprised of words in a certain sequence. The more descriptive or unique each one, the more significant it may be as an identifying property of the document; for example, the use of the proper noun "Winston Churchill." Some words are useful in sentences but not so much as distinguishing features that is: is, and, the, but, not, ...
However, their ubiquitous nature means they do not tend to interfere with the recognition of finer detailed or more specific features, so they can be included or ignored without too great an impact. When mapped into object-property space, this already produces a 2D space with an area of 36 billion units and populated with approximately 600 million radical matter particles—approximately 1.6% density.

Greedy hierarchies

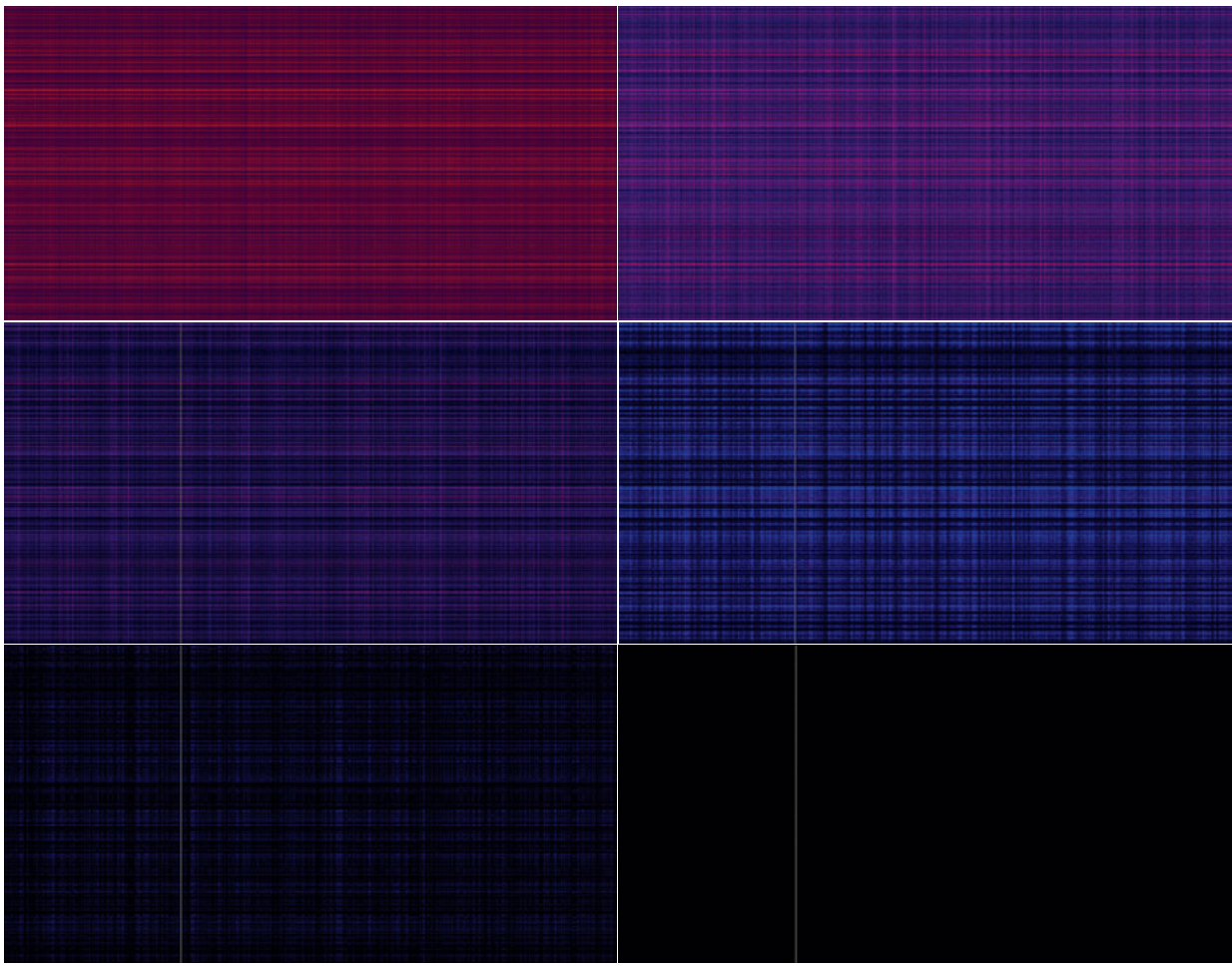
The scale of even this small fraction of Wikipedia already presents a significant theoretical and computational challenge to a statistical approach was used to determine if hierarchies of meaning could be detected in the data. It is logical to assume that when properties appear on a very small number of objects, in this case words appearing on the aforementioned Wikipedia pages, then those word properties must, by definition, be very specific.

Computer code was written to perform recursively the following processes: First to detect properties (words) so low in frequency that they appear on just a single object, (in this case, the Wikipedia page). These are as unique as they can get and we can consider them the very bottom of any descriptive hierarchy as they reference only a single object. Secondly to detect properties that appear on two and only two objects. These are the next most specific terms and each represent a group of just two pages which must therefore also discuss a relatively specific topic if only these two pages out of the full set use that particular word.

A meta-object is then created which combines the two pages into a first level hierarchy. To ensure that the meta-object fully represents the spectrum of its two members, it is assigned the mathematical union of the properties of those two members. The meta-object replaces the original two in object-property space and is included in the ongoing search for properties which appear on exactly two objects. Eventually the search is exhausted and the requirement of two and only two is relaxed to three and then four, and so on. As soon as any new meta-object can be created, the new combination of radical matter particles it has might then make it possible to find a pair again, and so the loop resets to two and then three and then four and so forth until finally everything that can be combined, has been combined. At every stage we are looking for the very smallest possible difference and the very smallest number of items that include this difference so the formation of the hierarchical meta-object is very fine grained and respectful of subtle differences.

To continue: the process is repeated along the orthogonal axis—which is to say, objects that have two and only two properties—then cause those two properties to be combined into a single meta-property. This meta-property collects all the objects of the two and the process

repeats back and forth between objects and properties until a complete hierarchy is constructed. The history of forming the meta-objects is retained and can be read-out to place both the objects and properties into sorted order. Both the dimension of the object-property space and the assignment of properties to objects remains completely unchanged; but the sorted order of both properties and objects now reflects the hierarchies they belong to.
A visualisation of this process was presented in video format during the *Data Loam: Sometimes Hard, Usually Soft* exhibition in Vienna 2019. It is shown here in steps from raw data to a single meta-object. There is some apparent structure even in the first image—upon investigation this turns out to be the already non-random order in which Wikipedia pages were created. For example, there are a block of pages referring to the years 1974, 1975, 1976, 1977... They form blocks of pages that also contain many of the same keywords—hence the visible horizontal (word) and vertical (page) lines, which we are shown in Figure 1. The image is presented as a heat map to improve resolution and as the number of objects and properties combine into meta-objects these become brighter (red > yellow) and the space between them colder (blue > black).



Eventually the algorithm runs to its conclusion and we are left with one meta-object. This should be the very most overarching object out of the first 300,000 pages in Wikipedia—but what is it?

The Western Marsh Harrier
Something definitely went wrong in this approach.

Image from the offending Wikipedia page:
en.wikipedia.org/wiki/Marsh_harrier

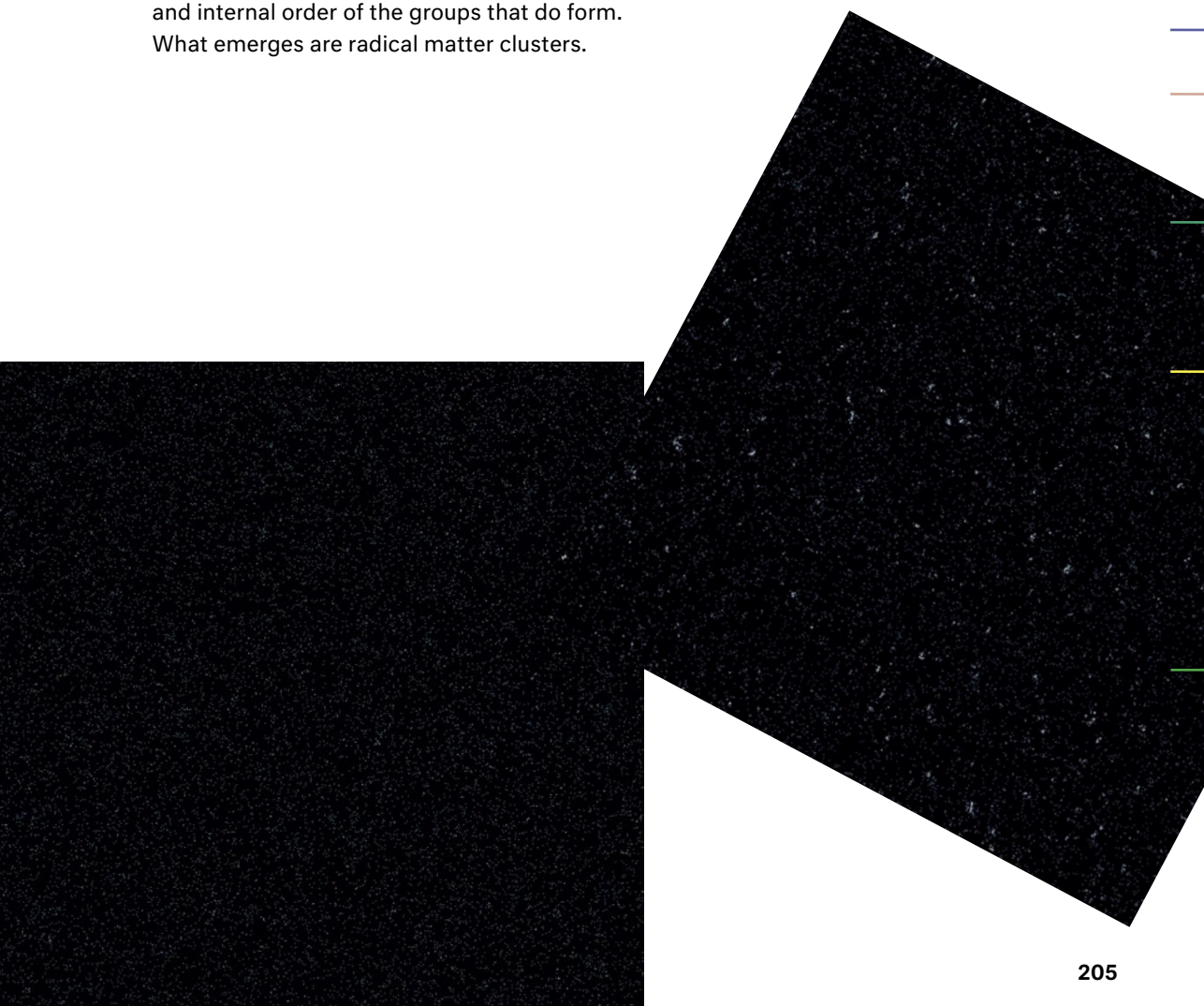
Spatial interaction

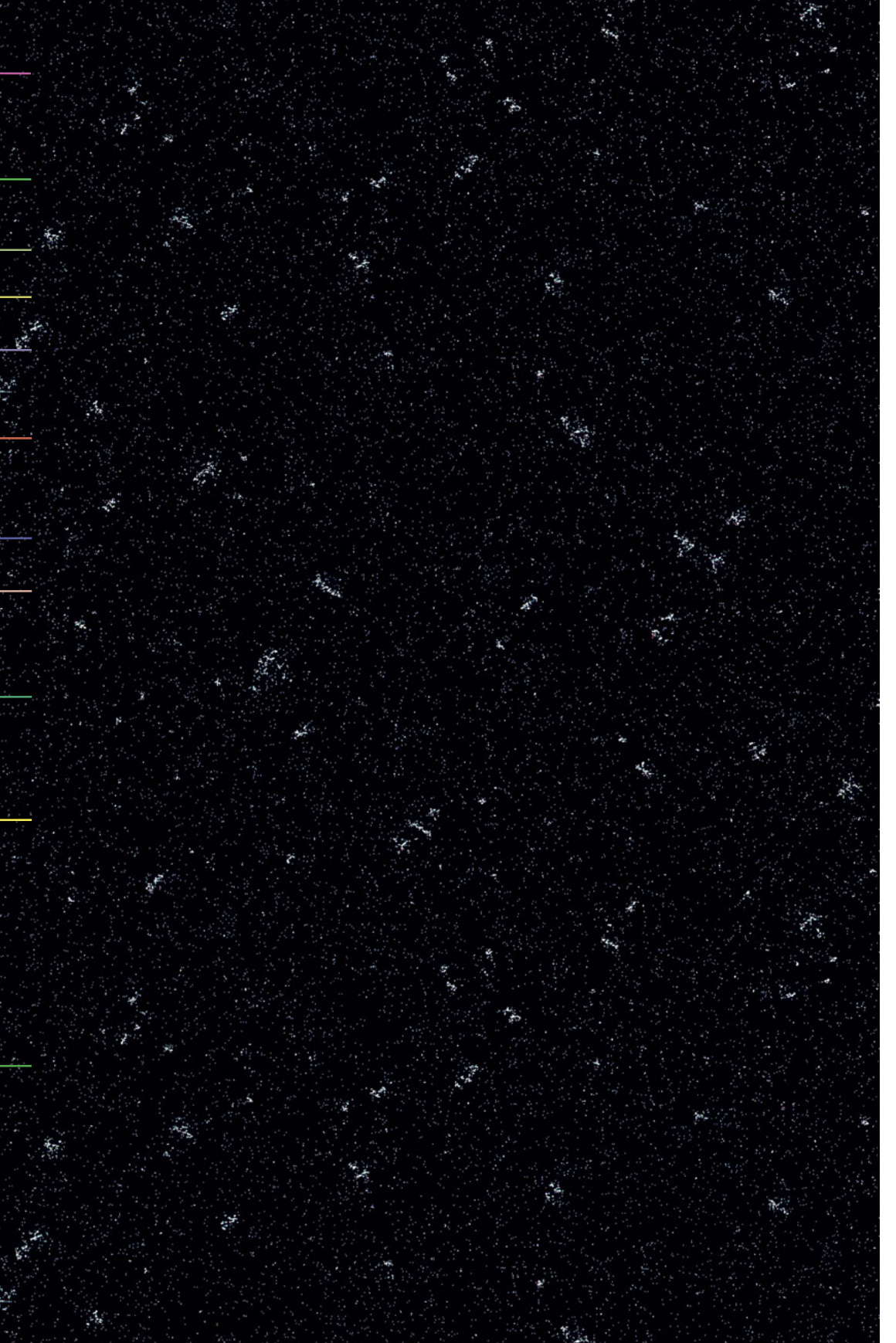
The problem with the hierarchical algorithm, and the reason that the section had “Greedy” in its title, is that it does not consider any ‘off-axis’ relationships between radical matter particles. The decisions are made on the basis of properties directly in common where the original concept for radical matter particles was that each and every one would continue to subtly influence all the others like stellar dust particles bathing in their mutual gravitational fields.

Using any kind of physical system analogue such as gravity, electrostatics or magnetism, the radical particles would all attract each other, universally, and more strongly the closer they would get. Due to constraints in movement as discussed above, they cannot actually alter the data and this manifests as a kind of ‘off-axis effect’. Adjacent radical matter particles would attract each other very strongly in the direction they are allowed to move, but if they were to be displaced along the orthogonal axis, only the component of that force ‘in the allowed direction’ would contribute to the ‘move’ decision. Further, it is the sum of all the individual forces acting on all the particles in a row (or column) that determines the total force vector acting on the row (or column). In this way, even very small off-axis forces are considered in the overall arrangement providing intricate detail and subtlety to the overall structure. The greedy hierarchy algorithm certainly did compute a result quite quickly—but we now know that the information it was throwing away is quite possibly the most, and certainly not the least, important. A full simulation of the same sized data set was well beyond the reach of this project but some preliminary results and interpretations have been obtained.

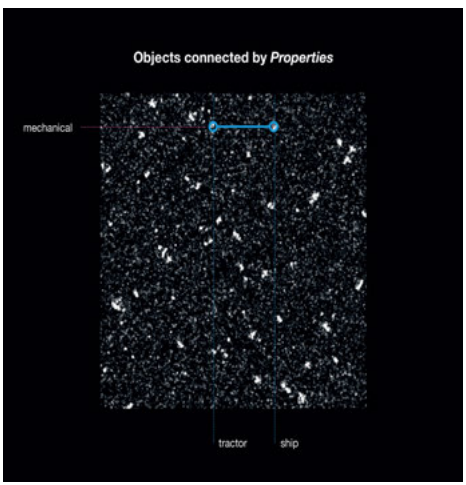
Radical emergence

Starting with a small (1000 × 1000) selected data set, a new simulator was coded and a sequence of images generated after each 'move' decision. This time, nothing was combined into meta-objects, everything remained discrete and full precision. The force acting on each and every radical matter particle, due to all the others was calculated. The nett force acting on each row (or column) could then be summed up from the components of the forces acting on all its member particles. This was used to generate movement decisions, including which rows (or columns) should force their way past the others and in which direction. The result on average was that the radical matter particles DO attract each other and really DO cause the rows and columns to re-sort into the order that most satisfies those still individual particles. Of course some particles will be a long way from any others, held there by a larger number that can collectively overpower a smaller group, but even then, they influence the form and internal order of the groups that do form.
What emerges are radical matter clusters.

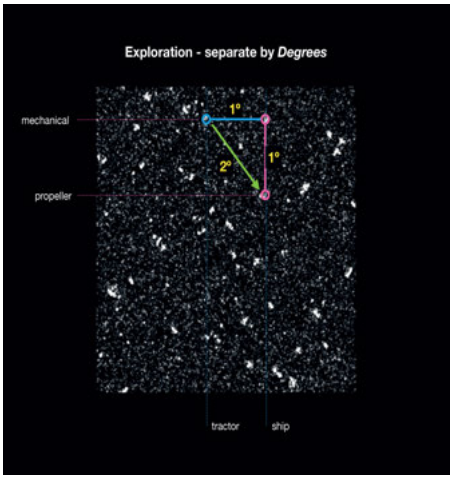




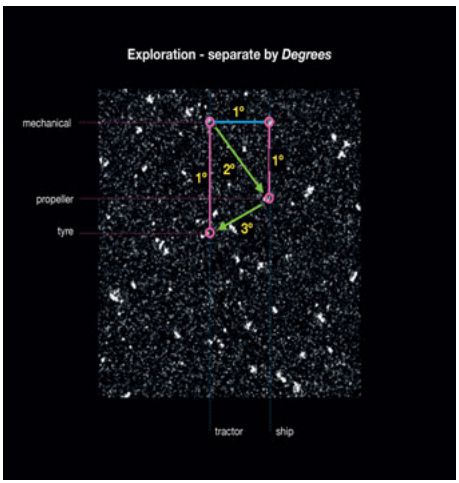
Since the rows of properties and columns of objects are carried with each move, these too are organised into optimised groups. The system evolves into a balanced, highly correlated state. Objects cluster together on the basis of properties they have, not just directly in common, but with all the subtle influences of minor, off-axis forces as well. The same happens to the properties. Where objects have overlapping, but not identical property collections, the radical matter may organise into discrete densities, displaced along either the object or property axis. This is where the power of this particular abstraction first becomes visible. This is not a classical Cartesian space. The difference in coordinate values does not in any way reflect the difference, or similarity, between the objects, or properties. Rather the fact that radical matter clusters align themselves along one of the axes in the system (recall that it may be n-dimensional) provides a direct connection between those clusters irrespective of the distance between them. The analogue of distance in this context is the number of 'jumps' taken between clusters referred to here as degrees of connection. Each radical matter cluster represents a collection of properties and objects; further, it has a centre of mass, a most suitable property to best describe the group and a representative object. The property can be used as the name of the cluster and the objects immediately accessed for the user of the system. One can immediately observe named connections by degrees from any starting point and control the navigation through even vast data sets in a named and therefore informed way.



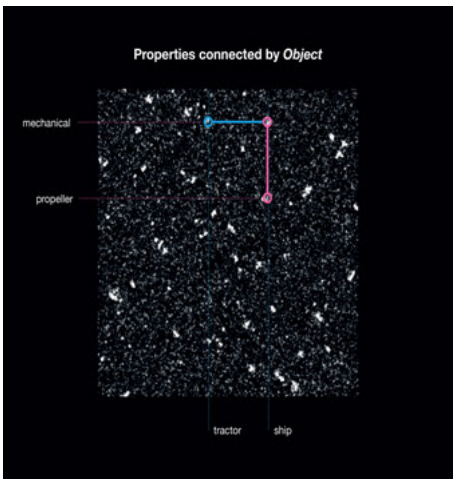
Here the objects 'tractor' and 'ship' are related 1º by the common property 'mechanical.'



This may seem obscure at first, but as more connections are revealed, the relationship as a form of propulsion becomes clear: the property ‘propeller’ for the object ‘ship;’ and ‘tyre’ for the ‘tractor.’



This then forms the basis for exploration ‘by degrees’ through object-property space.



Both ‘mechanical’ and ‘propeller’ are properties of the object ‘ship’ and in fact there is a 2º connection from the object ‘tractor’ to the property ‘propeller.’

Future correlation

Calculation of a large data set in this way was beyond the scope of the Data Loam project. The examples shown in this chapter however are actual simulations of small fields of radical matter particles precisely following the rules of the defined system. They attracted each other as if they had mass and a gravitational field. They rearranged as expected, carrying complete rows and columns in each move and as the simulation progressed, the emergence of a radical matter cluster was immediately obvious. This provides direct motivation to extend this work using 'messy' real world data and at a scale that makes oblique or otherwise unpredictable connections possible. If effective, the emergence of such connections would be the 'proof of life' for the theory overall.

entanglement [ɪnˈtæŋglmənt]

not to be confused with getting 'mixed-up' or 'entwined', it signals a most incredible moment of connection, conscious or otherwise, evading and simultaneously re-instating speed, duration, curved-time. Cf superpositionality, dimensional.