# Mapping data transitions: from data to sensemaking in learning analytics

### Research Proposal for PhD – Ben Hicks

## Context

The process of using data to inform decision makers involves a number of data transformations. For example, Learning Analytics (LA) can be viewed as a sequence of mappings, beginning in the 'real world’ as an instructor designs a lesson, and ending with some (potentially different) person trying to make sense of the results. I propose to view this as a series of *mappings*, from reality, to data, to an artefact and, finally, to a decision maker attempting to understand the resulting reports. Through each step various abstractions are made, be they assumptions, sampling, modelling, visualising, dimension reduction or even the creation of metaphor for clearer explanation. The end goal is to try and convey some *meaning* about the scenario being modelled to some person, but the choices and processes made along the way can distort or over-simplify the truth. All analysis and storytelling presents a shadow of the world. This raises the question of whether we can be sure that the silhouette captures essential (and possibly generalizable) information. Or is it just particular to this set of data?

My research project aims to explore the transitions between these spaces (reality, data, artefact & narrative) in an attempt to quantify the change in structure that they entail. This would enable the development of tools that are capable of helping people to understand the limitations of data analyses and the boundaries of the conclusions they can draw.

## Literature Review

To clarify where the existing work towards this problem lies it is useful to assign some labels to some concepts:

World

Data

Artefact

Person

Ξ : Epistemic Cut

Ψ : Analysis

Narrative

Sense making

*Figure 1: Transitions of an Analysis*

At this stage, I have identified the mapping from the World space to the Data space (denoted by [](https://www.codecogs.com/eqnedit.php?latex=%5CXi%250)) as the primary area of interest and is discussed in Buckingham-Shum’s keynote (2016). Some (Davis & Sumara,  2009;  Tsai, et al. 2019) argue for a complexity thinking approach to LA, pointing to the inherent challenges of working with a mix of social and cognitive factors. This complexity thinking approach to understanding education can be seen as a focus away from reductionist components towards *whole* systems and emergent behaviour, and is widely viewed as important (Mason, 2008). Edmonds (1999) provides a framework for examining this complexity and some authors use a topological space to represent a notion of ‘structure’ with more freedom than that offered by a metric space (Zhang & Zhang, 2004).

Although a number of different complex systems researchers have explored the mapping , [](https://www.codecogs.com/eqnedit.php?latex=%5CXi%250), or epistemic cut separating the world from observers (e.g. Pattee, 2012) , a more rigorous development in classifying this transition is needed, especially for the field of data science. The broad acknowledgement of the complex starting point of understanding education and learning is at odds with the reductionist approach of most data analytic processes. Almost exclusively work in this field begins at least part way along [](https://www.codecogs.com/eqnedit.php?latex=%5CXi%250) if not already in Data itself, and understandably so; the World is messy. Barwise & Seligman’s (1997) work on information flow and the work by Kent (2011, 2016, 2018) might be adaptable but fits more closely with [](https://www.codecogs.com/eqnedit.php?latex=%5CPsi#0) than [](https://www.codecogs.com/eqnedit.php?latex=%5CXi%250). In particular Kent (2016) defines the categories **Wrld**, **Struc** and **Lang** (categories of World, Semantic Structure and Language, respectively) and explores functors between them.

The Data space is the most heavily studied area of this map due to strong interest from the computer science field. The work potentially most applicable to this project is that of Zhang & Zhang (2004), who consider problem solving and granular computing. This provides an elegant way of looking at a space with structure and its refinement towards a solution by use of equivalence classes and quotient spaces.  In my project I would seek to explore this as a natural first step in the map [](https://www.codecogs.com/eqnedit.php?latex=%5CPhi%250) from Data to Artefact. A number of other methods can also be pursued, of note is Harris’ (2019) work that utilises category theory as a framework for describing the process when applied to learning algorithms.

Finally, the artefact might be some text, a graphic, a conversation, a short email, a single number and it probably also includes some narrative (possibly implied). Work on rigorously defining the structure of a graphical artefact used to convey data can be found in the grammar of graphics (Wilkinson, 2005). A single number or tabular report is simpler to define in the context of a space whereas a written report is more complicated. The discrete objects of a final analysis artefact are established fields of research in their own right; Linguistics or NLP for text, Mathematics or Statistics for numbers, for instance. Research on how humans make sense of these artefacts is more of an emerging field but there is still some strong work to build on (Echeverria, Martinez-Maldonado, Buckingham Shum, 2019).

## Objectives

In order to pursue my overarching research aim, I propose to start by pursuing the following two objectives·

Develop a mathematical framework that depicts information transitions in viewing the process of moving from reality to an analysis artefact, along with the interpretation by a human observer.

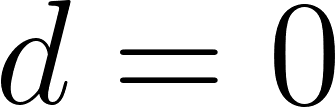
Develop tools that make these information transitions visible to appropriate end users and highlight the (possibly) hidden assumptions and simplifications used in those transitions.

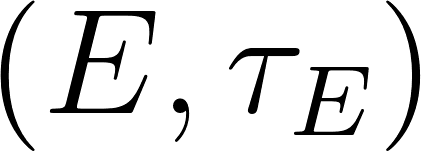
## Proposed Approach

In the first instance I plan to build on the work on quotient spaces and topologies on semi-ordered sets that was completed by  Zhang and Zhang (2004). If more flexibility is required in order to be applicable to the more interesting spaces then expand to using the more general tools of category theory (Kent, 2016; Kent, 2018; Spivak, 2014). A category theoretic approach fits with the work on information flow (Barwise et al, 1995) in which the information network can be naturally defined as a category.

Constructing various topologies or categories for the various ‘spaces’ involved in learning analytics will be the first step required. In terms of a careful approach to defining neighbourhood topologies on non-geometric spaces the approach of Brown (1968) is a useful, if slightly unconventional, way of understanding the transition from space to topology (he advocates for a neighbourhood before open set understanding of topologies).

Any standard work on topologies or categories considers the level of structure / complexity in a space; for instance Hausdorff or Metrizable properties of topologies, or the category / groupoid / group hierarchy of increasing structure. Kent (2016) provides examples of categories such as Wrld and Struc that could prove useful, and these ideas are further generalised in the Information Flow Framework of Kent (2018) and further built on in Kent’s (2011) work on Category Theory Ontology.

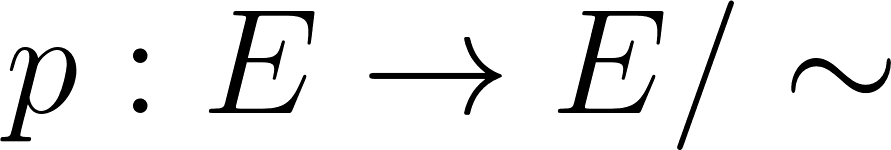
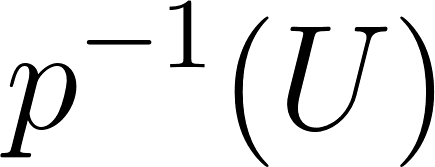
Some of these spaces might be relatively straight forward, such as a space containing academic performance of students which lends itself easily to defining a metric and a topology based on this metric. In fact most candidates *will* be metrizable if you can distinguish points as same or different, in which case you can use the discrete metric and induced topology ([](https://www.codecogs.com/eqnedit.php?latex=d%20%3D%201%250) if different and [](https://www.codecogs.com/eqnedit.php?latex=d%20%3D%200%250) if the same). Areas where this may not be the case would be spaces where it might be difficult to identify things as the ‘same’ but we can still form a notion of neighbourhoods.

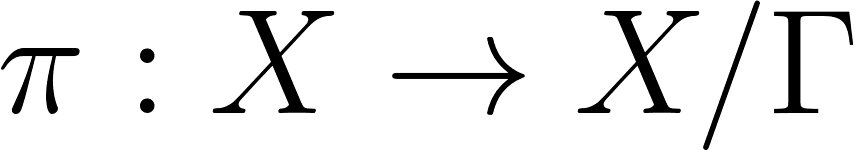
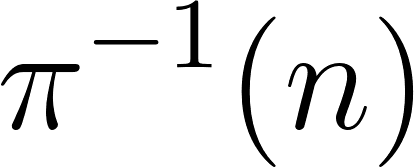
An example of a sufficiently complex space could be ‘engagement space’, a proposed neighbourhood topology, [](https://www.codecogs.com/eqnedit.php?latex=(E%2C%20%5Ctau_%7BE%7D)%250), of students and content, [](https://www.codecogs.com/eqnedit.php?latex=E%250), grouped into neighbourhoods, [](https://www.codecogs.com/eqnedit.php?latex=%5Ctau_%7BE%7D%250), of like engagement. Considering that engagement is challenging enough to define and talk about (Redmond, 2018), I anticipate that it might be difficult to  construct a notion of ‘space’ around this concept. Indeed, it might be the case that [](https://www.codecogs.com/eqnedit.php?latex=E%250) is not metrizable as there may be no way to say that two students have the ‘same’ engagement, which in itself would be an interesting finding that has not to date appeared in the field of LA. In such a case I could still continue my project by applying the discrete metric, this would enable me to consider e.g. two students as belonging to the same neighbourhood of engagement.

#### The analysis product - artefact and narrative

The final product of an analysis usually takes some concrete form (the artefact) as well as an embedded story (the narrative). The artefact is the more obvious product but I believe that the narrative is equally important, and persistent even when not stated. Initially (see *Figure 1*) I will explore the artefact as a space and the narrative as a mapping from the artefact to the observer. This could change however, as the model should work well with existing research into both analysis artefacts and sense making. This is where the strength of using a category theoretic approach might be most useful, as arrows of one category can be viewed as objects of another.

#### Transitions

Assuming that I find a way to loosely define a topology on specific spaces I identify during my project  (say topology of engagement, [](https://www.codecogs.com/eqnedit.php?latex=E%250)) I would  then need to move it to a suitable space for analysis. Now often the surrogate space for measuring engagement is activity, so what exactly is happening when we move from one to the other? Activity space, [](https://www.codecogs.com/eqnedit.php?latex=A%20%5Csubseteq%20E%250), might be seen as a subspace of the engagement space; essentially what we can observe a student doing as they ‘engage’ with learning (assuming that *all* activity relates to engagement, which may not be true). If we select all the dimensions for our activity space [](https://www.codecogs.com/eqnedit.php?latex=(a_1%2C%20a_2%2C%20%5Cdots%2C%20a_n)%250) then we form a natural equivalence relation, [https://lh6.googleusercontent.com/fz8b5NpgtVURLEceYXoCzwICrv-_aR-DcPoiRvJqK_gY9oU5JDfKBFF3pXM82KOQuf4ZrMR7VAgdX2Amsg6SbsHYLxyvCr4D3AtekQ_OEddBcHheIFbCc7BHsZyR6nqEbFX-DrjQ](https://www.codecogs.com/eqnedit.php?latex=%5Csim%250),  on the engagement space where points are equivalent if all their coordinates when projected to [](https://www.codecogs.com/eqnedit.php?latex=A%250) are equal. Let [](https://www.codecogs.com/eqnedit.php?latex=p%3A%20E%20%5Crightarrow%20E%2F%5Csim%250) map each element in [](https://www.codecogs.com/eqnedit.php?latex=E%250) to its equivalence class in [](https://www.codecogs.com/eqnedit.php?latex=E%2F%5Csim%250), we can now examine how well behaved this map [](https://www.codecogs.com/eqnedit.php?latex=p%250) is. For instance if the collection of subsets [](https://www.codecogs.com/eqnedit.php?latex=U%20%5Csubseteq%20E%2F%5Csim%250) such that [](https://www.codecogs.com/eqnedit.php?latex=p%5E%7B-1%7D(U)%250) are open in [](https://www.codecogs.com/eqnedit.php?latex=E%250) forms a collection of open subsets in [](https://www.codecogs.com/eqnedit.php?latex=E%2F%5Csim%250) we can say [](https://www.codecogs.com/eqnedit.php?latex=p%250) is a continuous map and infer more information about the topology [](https://www.codecogs.com/eqnedit.php?latex=E%250).

This process lends itself to a more generalisable strategy; view the sample space of what we measure as the quotient topology, [](https://www.codecogs.com/eqnedit.php?latex=X%2F%5CGamma%250), of a larger topology and explore what is implied by both the equivalence relation [](https://www.codecogs.com/eqnedit.php?latex=%5CGamma%250) defined on [](https://www.codecogs.com/eqnedit.php?latex=X%250) and the map [](https://www.codecogs.com/eqnedit.php?latex=%5Cpi%3A%20X%20%5Crightarrow%20X%2F%5CGamma%250), and in particular what happens to neighbourhoods [](https://www.codecogs.com/eqnedit.php?latex=n%20%5Cin%20X%2F%5CGamma%250) under the inverse map [](https://www.codecogs.com/eqnedit.php?latex=%5Cpi%5E%7B-1%7D(n)%250).

## Research plan and timetable

In rough 6-month sections:

* Research into previous work on complexity, applications of category theory, information theory, topology (in particular complexity-analogous measures such as metrizability), quotient spaces. Sketching ideas for objects and morphisms in the analysis process and conversations with people with experience in the field at the different stages of analysis.
* Formalising the mathematical structure to be used, and linking this with the research. Identifying holes in the research that need to be filled in order to apply to the process of learning analytics. Exploring options for a data set to apply this too; this could be leveraged with my current work with CSU analysing student engagement data (which is certainly complex enough). Applying for ethics once a data set has been decided upon. Beginning literature review formal writing on research conducted in the last checkpoint.
* Polishing the framework and first round of applying the framework to a data set. Finalising literature review. Possible paper on framework.
* Refining application of model to data analysis process; exploring options for quantifying space-complexity metric and whether it is possible to retroactively look at an analysis and quantify information loss.
* Further exploration and refinement of information loss metric that can highlight hidden assumptions and simplifications; and how broadly applicable it could be. Drafting thesis.
* Finalising thesis.

## References

Barwise, J., Gabbay, D., Hartonas, C. (1995). *On the Logic of Information Flow*. Logic Journal of IGPL.

Barwise, J., & Seligman, J. (1997). *Information Flow: The Logic of Distributed Systems*. Cambridge Core.

Brown, R. (1968). *Topology and Groupoids*.

Buckingham-Shum, S. (2016). *Learning Analytics for 21st Century Competencies.* Journal of Learning Analytics.

Davis, B., & Sumara, D. J. (2006). *Complexity and Education Inquiries into Learning, Teaching and Research.* Psychology Press.

Edmonds, B. (1999). *Syntactic Measures of Complexity* (Doctoral thesis).

Echeverria, V., Martinez-Maldonado, R. and Buckingham Shum, S. (2019). Towards Collaboration Translucence: Giving Meaning to Multimodal Group Data. In *Proceedings of ACM CHI Conference (CHI’19).* ACM, New York, NY, USA, Paper 39, 16 pages.

Kent, R. (2011). *A KIF Formalization for the IFF Category Theory Ontology*.

Kent, R. (2016). *Information Flow in Logical Environments*.

Kent, R. (2018). *Semantic Integration in the Information Flow Framework.*

Harris, K. (2019). *Characterizing the invariances of learning algorithms using category theory*.

Mason, M. (2008). *Complexity Theory and the Philosophy of Education*

Pattee, H. H., Rączaszek-Leonardi, J., (2012). *Laws, Language and Life: Howard Pattee’s Classic papers on the Physics of Symbols with Contemporary Commentary*. New York: Springer.

Redmond, P., Heffernan, A., Abawi, L., Brown, A., Henderson, R. (2018). *An Online Engagement Framework for Higher Education*

Spivak, D. (2014). *Category Theory for the Sciences*. MIT Press.

Tsai, Y-S., Poquet, O., Gasevic, D., Dawson, S., Pardo, A. (2019). *Complexity leadership in Learning Analytics: Drivers, challenges and opportunities.* British Journal of Educational Technology.

Wilkinson, L. (2005). *The Grammar of Graphics* (2nd ed.). Springer.

Zhang, B., Zhang, L. (2004). *The Quotient Space Theory of Problem Solving.*