Topological Artist Model

Hannah Aizenman

A dissertation proposal submitted to
the Graduate Faculty in Computer Science in partial fulfillment of the
requirements for the degree of Doctor of Philosophy,
The City University of New York

Committee Members:
Dr. Michael Grossberg (Advisor), Dr. Robert Haralick, Dr. Lev Manovich,
Dr. Huy Vo, Dr. Marcus Hanwell

June 2021

Abstract

This work presents a functional model of the structure-preserving maps from data to visual representation to guide the development of visualization libraries. Our model, which we call the topological equivariant artist model (TEAM), provides a means to express the constraints of preserving the data continuity in the graphic and faithfully translating the properties of the data variables into visual variables. We formalize these transformations as actions on sections of topological fiber bundles, which are mathematical structures that allow us to encode continuity as a base space, variable properties as a fiber space, and data as binding maps, called sections, between the base and fiber spaces. This abstraction allows us to generalize to any type of data structure, rather than assuming, for example, that the data is a relational table, image, data cube, or network-graph. Moreover, we extend the fiber bundle abstraction to the graphic objects that the data is mapped to. By doing so, we can track the preservation of data continuity in terms of continuous maps from the base space of the data bundle to the base space of the graphic bundle. Equivariant maps on the fiber spaces preserve the structure of the variables; this structure can be represented in terms of monoid actions, which are a generalization of the mathematical structure of Stevens' theory of measurement scales. We briefly sketch that these transformations have an algebraic structure which lets us build complex components for visualization from simple ones. We demonstrate the utility of this model through case studies of a scatter plot, line plot, and image. To demonstrate the feasibility of the model, we implement a prototype of a scatter and line plot in the context of the Matplotlib Python visualization library. We propose that the functional architecture derived from a TEAM based design specification can provide a basis for a more consistent API and better modularity, extendability, scaling and support for concurrency.

Contents

1 Introduction

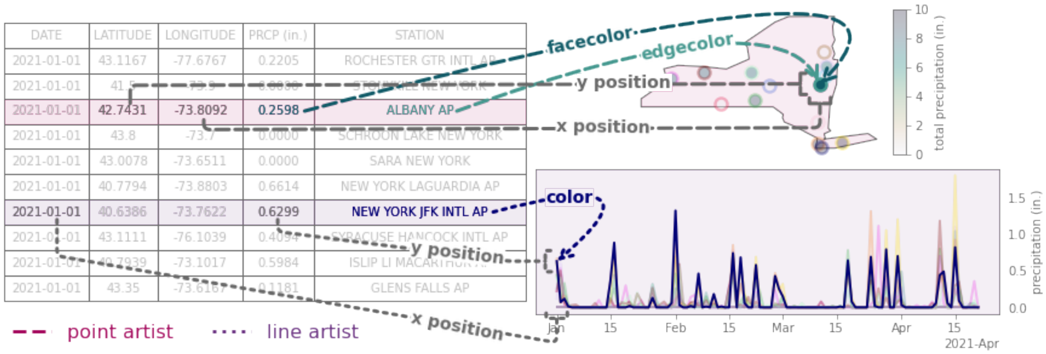


Figure 1: Visualizations are made up of transformations from data into visual representation. These functions transform individual data values to visual representation, such as date to x position or latitude to y position. These functions are composed into the assembly of all these transformations into a visual mark, such as a line or point. The same variable can be mapped in different ways, for example line is mapped to a color in the scatter plot and to y position in the line plot.

Visualization is the transformation of data into visual representation. As illustrated by Figure 1, these translations are both at the level of the individual variable and the entire record. In the case of the scatter plot, the latitude and longitude are encoded as the x and y position, respectively, while the temperature and station are represented by the face and edge colors. A row in the table is collectively encoded as a point mark. None of these encodings are fixed, as evidenced by temperature being translated into the y value in the case of the line plot. The station is now the source of the color of the entire line, and the date is the x position. As with scatter, the encodings of the individual transformations, which again are on values from the same record in the table, are composited into a line mark. It is these raw transformations from data space to visualization space that are implemented by building block level visualization libraries, named as such because the functions provided by the library can be composited in any number of ways to yield visualizations [1]. We propose that like physical building blocks, building block libraries must provide a collection of well defined pieces that can be composed in whichever ways the blocks fit together. We specify that a valid visualization block is a structure preserving transformation from data to visual space, and we define structure in terms of *continuity* and *equivariance*. We then use this model to develop a design specification for the components of a building block visualization library. The notion of self contained, inherently modular, building blocks lends itself naturally to a functional paradigm of visualization [2]. We adopt a functional model for a redesign because the lack of side effects means functional architecture can be evaluated for correctness, functional programs tend to be shorter and clearer, and are well suited to distributed, concurrent, and on demand tasks[3].

This work is strongly motivated by the needs of the Matplotlib[4, 5] visualization library. One of the most widely used visualization libraries in Python, since 2002 new components and features have been added in a some what adhoc, sometimes hard to maintain, manner. Particularly, each new component carries its own implicit notion of how it believes the data is structured-for example if the data is a table, cube, image, or network - that is then expressed in the API for that component. In turn, this yields an inconsistent API for interfacing with the data, for example when updating streaming visualizations or constructing dashboards[6]. This entangling of data model with visual transform also yields inconsistencies in how visual component transforms, e.g. shape or color, are supported. We propose that these issues can be ameliorated via a redesign of the functions that convert data to graphics, named *Artists* in Matplotlib, in a manner that reliably enforces *continuity* and *equivariance* constraints. We evaluate our functional model by implementing new artists in Matplotlib that are specified via *equivariance* and *continuity* constraints. We then use the common data model introduced by the model to demonstrate how plotting functions can be consolidated in a way that makes clear whether the difference is in expected data structure, visual component encoding, or the resulting graphic.

2 Background

There are many formalisms of the notion that visualization is structure preserving maps from data to visual representation, and many visualization libraries that attempt to preserve structure in some manner; this work bridges the formalism and implementation in a functional manner with a topological approach at a building blocks library level to propose a new model of the constraints visual transformations must satisfy such that they can be composed to produce visualize representations that can be considered equivalent to the data being represented.

2.1 Structure:

Visual representations of data, by definition, reflect something of the underlying structure and semantics[7], whether through direct mappings from data into visual elements or via figurative representations that have meaning due to their similarity in shape to external concepts [8]. The components of a visual representation were first codified by Bertin[9], who introduced a notion of structure preservation that we formally describe in terms of *equivariance* and *continuity*.

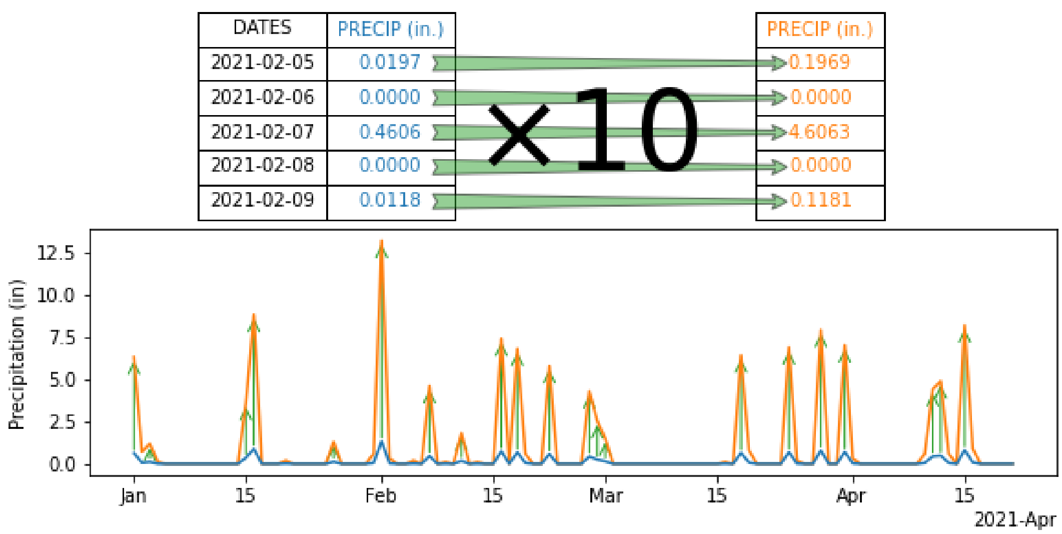
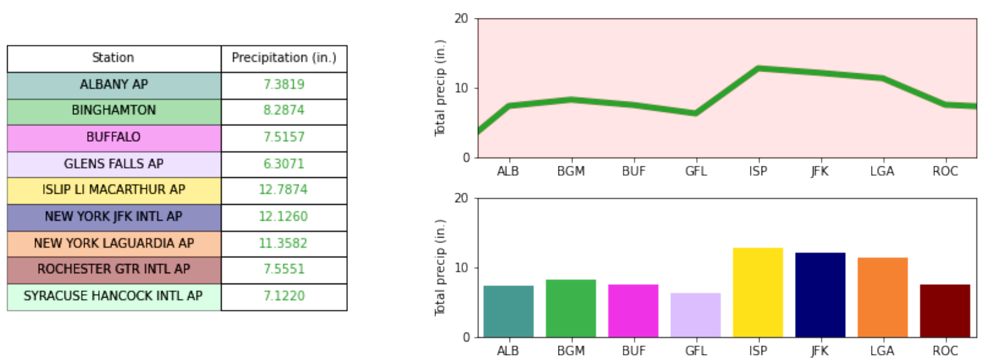


Figure 2: The data in blue is scaled by a factor of two, yielding the data in orange. To preserve *equivariance*, the blue line plot representation of the unscaled data is also scaled by a factor of two, yielding the orange line plot that is equivalent to the scaled data.

Bertin proposes that there are classes of visual encodings-such as position, shape, color, and texture-that when mapped to from specific types of measurement, quantitative or qualitative, will preserve the properties of that measurement type. For example, in Figure 2, the data and visual representation are scaled by equivalent factors of two, resulting in the change illustrated in the shift from blue to orange data and lines. The idea of equivariance is formally defined as the mapping of a binary operator from the data domain to the visual domain in Mackinlay's *A Presentation Tool*(APT) model[10, 11]. The algebraic model of visualization proposed by Kindlmann and Scheidegger uses equivariance to refer generally to invertible binary transformations[12], which are mathematical groups [13]. Our model defines *equivariance* in terms of monoid actions, which are a more restrictive set than all binary operations and more general than groups. As with the algebraic model, our model also defines structure preservation as commutative mappings from data space to representation space to graphic space, but our model uses topology to explicitely include continuity.



Station	Precipitation (in.)
ALBANY AP	7.3819
BINGHAMTON	8.2874
BUFFALO	7.5157
GLENS FALLS AP	6.3071
ISLIP LI MACARTHUR AP	12.7874
NEW YORK JFK INTL AP	12.1260
NEW YORK LAGUARDIA AP	11.3582
ROCHESTER GTR INTL AP	7.5551
SYRACUSE HANCOCK INTL AP	7.1220

Figure 3: The line plot does not preserve *continuity* because it implies that the average temperature at each station lie along a 1D continuous line, while the bar plot preserves *continuity* by representing the average temperatures at each station as the discrete values they are.

Bertin proposes that the visual encodings be composited into graphical marks that match the *continuity* of the data - for example discrete data is a point, 1D continuous is the line, and 2D data is the area mark. In Figure 3, the line plot does not preserve continuity because the line connecting the discrete categories implies that the frequency of weather events is sampled from a continuous interval and the categories are points on that interval. But, when the continuity is preserved, as in the bar chart, then the graphic has not introduced new structure into the data.

Structure

continuity How records in the dataset are connected to each other, e.g. discrete rows, neworked nodes, points on a continuous surface

equivariance if an action is applied to the data or the graphic–e.g. a rotation, permutation, translation, or rescaling– there must be an equivalent action applied on the other side of the transformation.

The notion that a graphic should be equivalent to the data has been expressed in a variety of ways. Informally, Norman's Naturalness Principal[14] states that a visualization is easier to understand when the properties of the visualization match the properties of the data. This principal is made more concrete in Tufte's concept of graphical integrity, which is that a visual representation of quantitative data must be directly proportional to the numerical quantities it represents (Lie Principal), must have the same number of visual dimensions as the data, and should be well labeled and contextualized, and not have any extraneous visual elements[15]. expressing, as defined by Mackinlay, is a measure how much of the mathematical structure in the data that can be expressed in the visualizations; for example that ordered variables can be mapped into ordered visual elements. We propose that a graphic is an equivalent representation of the data when *continuity* and *equivariance* are preserved.

2.2 Tools

DATE	LATITUDE	LONGITUDE	TEMP	NAME
2021-01-01	42.1	-77.1	27.86	GANG MILLS NEW YORK
2021-01-01	41.5	-73.9	29.48	STONYKILL NEW YORK
2021-01-01	43.8	-73.7	24.08	SCHROON LAKE NEW YORK
2021-01-01	43.35	-73.6167	25.52	GLENS FALLS AP
2021-01-01	43.1111	-76.1039	33.80	SYRACUSE HANCOCK INTL AP
2021-01-01	43.1167	-77.6767	31.64	ROCHESTER GTR INTL AP
2021-01-01	40.6386	-73.7622	35.96	NEW YORK JFK INTL AP
2021-01-01	42.1997	-75.985	28.76	BINGHAMTON
2021-01-01	40.7939	-73.1017	35.78	ISLIP LI MACARTHUR AP
2021-01-01	43.0078	-73.6511	27.68	SARA NEW YORK

Figure 4: Visualization libraries, especially ones tied to specific domains, tend to be architectured around a core data structure, such as tables, images, or networks.

One of the reasons we developed a new formalism rather than adopting the architecture of an existing library is that most information visualization software design patterns, as categorized by Heer and Agrawala[16], are tuned to very specific data structures. These libraries can often assume the expected data structure because they are domain specific, and that is the common data structure in that domain. For users who generally work in one domain, such as the data, networks, or graphs shown in Figure 4, this well defined data space (and corresponding visual space[17]) often yields a very coherent user experience[18]. But, for developers who want to build new visualizations on top of these libraries, they must work around the existing assumptions, sometimes in ways that break the model the libraries are developed around.

For example, many domain specific libraries integrate computation into the visualization, for example libraries based that assume all data is a relational database. This assumption is core to tools influenced by APT, such as Tableau[19–21] and the Grammar of Graphics[22], such as ggplot[23], protovis[24], vega[25] and altair[26]. Since these libraries represent all data as a table, and computations on tables are fairly well defined[27], they can include computations on the table with a fair bit of confidence that the computation is accurate. Since most computations are specific to domains, general purpose block libraries can not make this assumption; instead a goal of this model is to identify which computations are specifically part of the visual encoding - for example mapping data to a color-and which are manipulations on the data. Disentangling the computation from the visual transforms allows us to determine whether the visualization library needs to handle them or if they can be more efficiently computed by the data container.

A different class of user facing tools are those that support images, such as ImageJ[28] or Napari[29]. These tools often have some support for visualizing non image components of a complex data set, but mostly in service to the image being visualized. These tools are ill suited for general purpose libraries that need to support data other than images because the architecture is oriented towards building plugins into the existing system [30] where the image is the core data structure. Even the digital humanities oriented ImageJ macro ImagePlot[31], which supports some non-image aggregate reporting charts, is still built around image data as the primary input. The need to visualize and manipulate graphs has spawned tools like Gephi[32], Graphviz[33], and Networkx[34]. As with tables and images, extending network libraries to work with other types of data either require breaking their internal model of how data is structured and what transformations of the data are allowable or growing a model for other types of data structures alongside the network model.

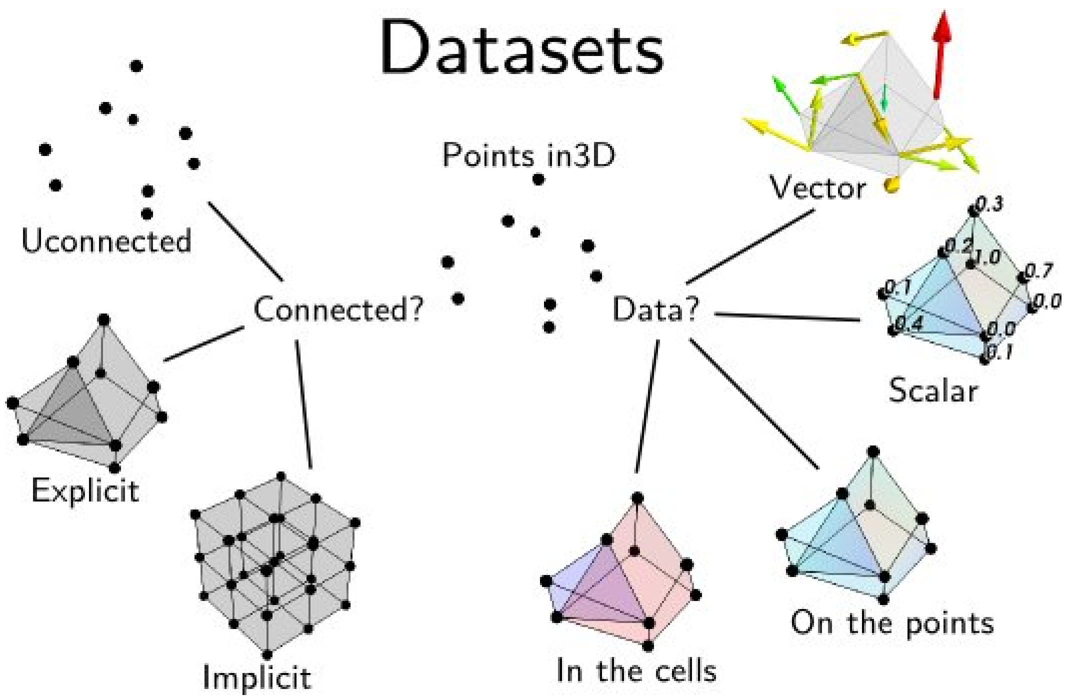


Figure 5: One way to describe data is by the connectivity of the points in the dataset. A database for example is often discrete unconnected points, while an image is an implicitely connected 2D grid. This image is from the Data Representation chapter of the MayaVi 4.7.2 documentation.[35]

Many building block libraries carry multiple models of data internally because they cannot assume a data structure. Algorithms are designed such that the structure of data is assumed, as described in Tory and Möller's taxonomy [**ToryRethinkingVisualization2004**], and by definition building block libraries try to provide the components to build any sort of visualization. Matplotlib, D3[36], and VTK [**geveci2012vtk**, 37] and its derivatives such as MayaVi[38] and extensions such as ParaView[39] and the infoviz themed Titan[40]. Where GoG and ImageJ type libraries have coherant APIs for their visualization tools because the data structure is the same, the APIs for visualizations in Matplotlib, D3, and VTK are significantly dependent on the structure of the data it expects. VTK has codified this in terms of *continuity* based data representations, as illustrated in figure 5. This API choice can lead to visualizations that break *continuity* when fed into visualizations with different assumptions about structure. The lack of consistent data model can also mean no consistent way of updating the data and therefore no way of guaranteeing that the views are in sync, in visualizations that consistent of multiple views of the same datasource, such as dashboards[6, 41]. To resolve this issue, our functional model takes as input a structure aware data abstraction general enough to provide a common interface for many different types of visualization.

2.3 Data

One such general abstraction are fiber bundles, which Butler proposed as a core data structure for visualization because they encode data continuity separately from the variable properties and are flexible enough to support discrete and ND continuous datasets [42, 43]. Since Butler's model lacks a robust way of describing variables, we can encode a schema like description of the data in the fiber bundle by folding in Spivak's topological description of data types [44, 45]. In this work we will refer to the points of the dataset as *records* to indicate that a point can be a vector of heterogenous elements. Each *component* of the record is a single object, such as a temperature measurement, a color value, or an image. We also generalize *component* to mean all objects in the dataset of a given type, such as all temperatures or colors or images. The way in which these records are connected is the *connectivity*, *continuity*, or more generally *topology*.

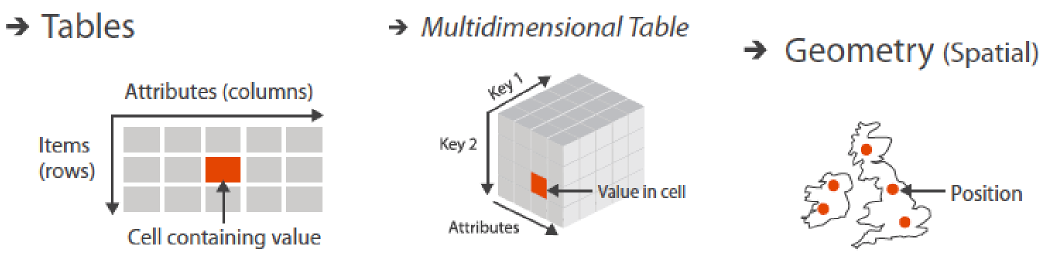


Figure 6: Values in a dataset have keys associated with them that describe where the value is in the dataset. These keys can be indexers or semantically meaningful;for example, in a table the keys are the variable name and the row ID. In the data cube, the keys is the row, column, and cell ID, and in the map the key is the position in the grid. Image is figure 2.8 in Munzner's Visualization Analysis and Design[46]

The *continuity* can often be described by some variables in the dataset; this is formalized Munzner's notion of metadata as *keys* into the data structure that return associated *values*[47]. As shown in Figure 6, keys can be labeled indexes, such as the attribute name and row ID, or physical entities such as locations on a map. We propose that information rich metadata are part of the components and instead the values are keyed on coordinate free structural ids. In contrast to Munzner's model where the semantic meaning of the key is tightly coupled to the position of the value in the dataset, our model considers keys to be a pure reference to topology. This allows the metadata to be altered, without imposing new semantics on the underlying structure, for example by changing the coordinate systems or time resolution. This value agnostic model also supports encoding datasets where there may be multiple independent variables - such as a measure of plant growth given variations in water, sunlight, and time - without having to assume any one variable is inducing the change in growth. For building block library developers, this means the components are able to fully traverse the data structures without having to know anything about the values or the semantic meaning of the structure. Since these components are by design *equivariant* and *continuity* preserving, domain specific library developers in different domains that both rely on the same continuity, for example 2D continuity, can then safely reuse the components to build tools that can safely make domain specific assumptions.

2.4 Contribution

The contribution of this work is

1. formalization of the topology preserving relationship between data and graphic via continuous maps subsubsection 3.2.2
2. formalization of property preservation from data component to visual representation as monoid action equivariant maps subsubsection 3.3.2
3. functional oriented visualization architecture built on the mathematical model to demonstrate the utility of the model subsubsection 3.3.3
4. prototype of the architecture built on Matplotlib's infrastructure to demonstrate the feasibility of the model. **??**

3 Topological Equivariant Artist Model

To guide the implementation of structure preserving building block components, we develop a mathematical formalism of visualization that specifies how these components preserve *continuity* and *equivariance*. Inspired by the somewhat analogous component in Matplotlib[5], we call the transformation from data space to graphic that these building block components implement the *artist*.

$$\mathscr{A} : \mathscr{E} \to \mathscr{H} \tag{1}$$

The *artist* $\mathscr{A}$ is a map from the data $\mathscr{E}$ to graphic $\mathscr{H}$ fiber bundles. To explain how the *artist* is a structure preserving map from data to graphic, we first describe how we model data (subsection 3.1) and graphics (subsection 3.2) as topological structures that encapsulate component types and continuity. We then discuss the maps from graphic to data (subsubsection 3.2.2, data components to visual components (subsubsection 3.3.2), and visual components into graphic (subsubsection 3.3.3) that make up the artist.

3.1 Data Space E

Building on Butler's proposal of using fiber bundles as a common data representation structure for visualization data[42, 43], a fiber bundle is a tuple (E, K, π, F) defined by the projection map π

$$F \hookrightarrow E \xrightarrow{\pi} K \tag{2}$$

that binds the components of the data in F to the continuity of the data encoded in K. The fiber bundle models the properties of data component types F (subsubsection 3.1.1), the continuity of records K (subsubsection 3.1.3), the collections of records τ (**??**), and the space E of all possible datasets with these components and continuity. By definition fiber bundles are locally trivial[48, 49], meaning that over a localized neighborhood U the total space is the cartesian product $K \times F$. We use fiber bundles as the data model because they are inclusive enough to express all the types of structures of data described in subsection 2.2

3.1.1 Variables in Fiber Space F

To formalize the structure of the data components, we use notation introduced by Spivak [45] that binds the components of the fiber to variable names. This allows us to describe the components in a schema like way. Spivak constructs a set $\mathbb{U}$ that is the disjoint union of all possible objects of types $\{T_0, \ldots, T_m\} \in \mathbf{DT}$, where $\mathbf{DT}$ are the data types of the variables in the dataset. He then defines the single variable set $\mathbb{U}_\sigma$

$$\begin{tikzcd} \mathbb{U}_\sigma \arrow[r] \arrow[d, "\pi_\sigma"'] & \mathbb{U} \arrow[d, "\pi"] \\ C \arrow[r, "\sigma"'] & \mathbf{DT} \end{tikzcd} \tag{3}$$

which is $\mathbb{U}$ restricted to objects of type T bound to variable name c. The $\mathbb{U}_\sigma$ lookup is by name to specify that every component is distinct, since multiple components can have the same type T. Given σ, the fiber for a one variable dataset is

$$F = \mathbb{U}_{\sigma(c)} = \mathbb{U}_T \tag{4}$$

where σ is the schema that binds a variable name c to its datatype T. A dataset with multiple components has a fiber that is the cartesian cross product of $\mathbb{U}_\sigma$ applied to all the columns:

$$F = \mathbb{U}_{\sigma(c_1)} \times \ldots \mathbb{U}_{\sigma(c_i)} \ldots \times \mathbb{U}_{\sigma(c_n)} \tag{5}$$

which is equivalent to

$$F = F_0 \times \ldots \times F_i \times \ldots \times F_n \tag{6}$$

which allows us to decouple F into components F_i.

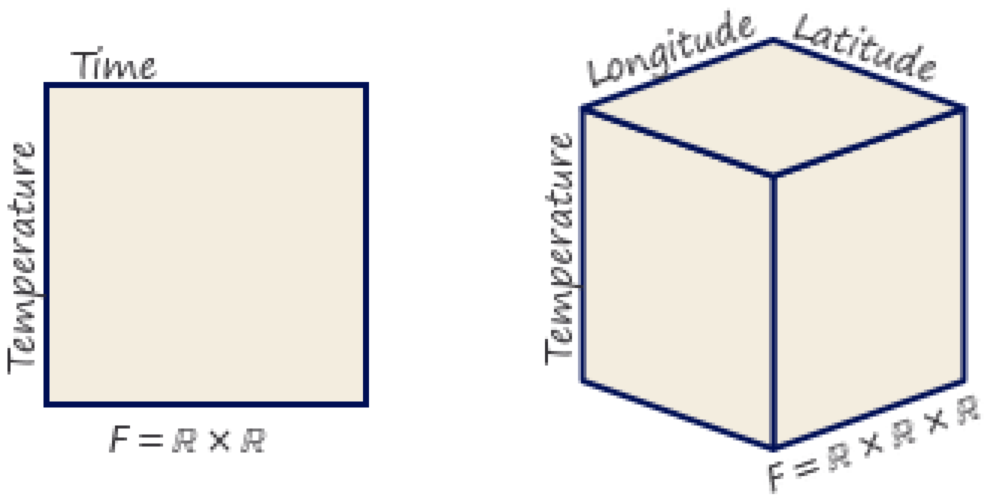


Figure 7: The fiber space is the set of all combinations of all theoretically possible values of the components. The 2D fiber $F = \mathbb{R} \times \mathbb{R}$ encodes the properties of *time* and *temperature* components. One dimension of the fiber encodes the range of possible values for the time component of the dataset, which is a subset of the $\mathbb{R}$, while the other dimension encodes the range of possible values $\mathbb{R}$ for the temperature component. This means the fiber is the set of points *(temperature, time)*that are all the combinations of *temperatature* $\times$ *time*. The 3D fiber encodes points at all possible combinations of *temperature*, *latitude*, and *longitude*.

For example, the records in the 2D fiber in **??** are a pair of *times* and °K *temperature* measurements taken at those times. Time is a positive number of type `datetime` which can be resolved to floats $\mathbb{U}_{\texttt{datetime}} = \mathbb{R}$. Temperature values are real positive numbers $\mathbb{U}_{\texttt{float}} = \mathbb{R}^+$. The fiber is

$$F = \mathbb{R} \times \mathbb{R}^+$$

where the first component F_0 is the set of values specified by ($c = time$, $T =$ `datetime`, $\mathbb{U}_\sigma = \mathbb{R}$) and F_1 is specified by ($c = temperature$, $T =$ `float`, $\mathbb{U}_\sigma = \mathbb{R}$) and is the set of values $\mathbb{U}_\sigma = \mathbb{R}$. In the 3D fiber in **??**, time is replaced with location. This location variable is of type `point` and has two components *latitude* and *longitude* $\{(lat, lon) \in \mathbb{R}^2 \mid -90 \leq lat \leq 90,\ 0 \leq lon \leq 360\}$. The fiber for this dataset is

$$F = \mathbb{R} \times \mathbb{R}^2 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$$

where the dimensionality of the fiber does not change, but the components of the fiber can be coupled. For example, *location* can can either be specified as ($c = location$, $T =$ `point`, $\mathbb{U}_\sigma = \mathbb{R}^2$)or ($c = latitude$, $T =$ `float`, $\mathbb{U}_\sigma = \mathbb{R}$) and ($c = longitude$, $T =$ `float`, $\mathbb{U}_\sigma = \mathbb{R}$).

As illustrated in Figure 7, Spivak's framework provides a consistent way to describe potentially complex components of the input data.

3.1.2 Measurement Scales: Monoid Actions

Implementing expressive visual encodings requires formally describing the structure on the components of the fiber, which we define by the actions of a monoid on the component. In doing so, we specify the properties of the component that must be preserved in a graphic representation. A monoid [50] M is a set with a binary operation $* : M \times M \to M$ that satisfies the axioms:

$$\textbf{associativity} \text{ for all} a, b, c \in M\ (a \bullet b) \bullet c = a \bullet (b \bullet c)$$

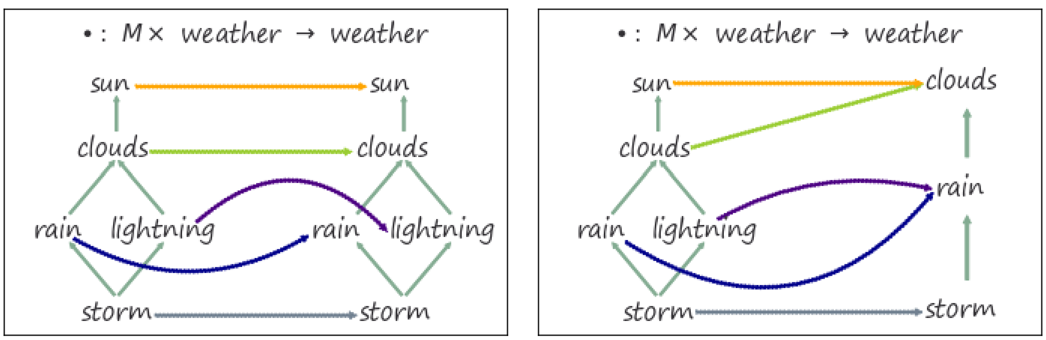
$$\textbf{identity} \text{ for all } a \in M,\ e \bullet a = a$$

As defined on a component of F, a left monoid action [51, 52] of M_i is a set F_i with an action $\bullet : M \times F_i \to F_i$ with the properties:

associativity for all $f, g \in M_i$ and $x \in F_i$, $f \bullet (g \bullet x) = (f * g) \bullet x$

identity for all $x \in F_i, e \in M_i$, $e \bullet x = x$

The identity and associativity properties of the action denote that the action is a monoid homomorphism, which means that the group operation is preserved on both sides of the action[53].



(a) (b)

Figure 8: The action $\bullet$ in **??** is the arrows from the partial order diagram of weather states on the left to the diagram of weather states on the right. Since the action maps the weather states to themselves, the ordering defined by the monoid $*$ is preserved on both sides of the action. The action in **??** is monoid homomorphism because the ordering of the weather states is the same as the ordering of the elements they are mapped to. Given $sun \geq clouds \geq rain\,lightining$ on the right, the action $sun\,clouds \to clouds$, and $rain\,lightining \to rain$ is structure preserving because on the left $cloud \geq rain$ so the relative ordering of elements is the same as the elements they are mapped to.

One example of monoids are partial orderings on a set, such as seen in . Each hasse diagram of the set of weather states describes an ordering on the set; the arrow goes from the lesser value to the greater one. For example, $storm \leq rain$. In **??**, the action $\bullet$ maps the elements of a set of weather states into itself by mapping them into other elements of the weather states. The action in Figure 8a, represented as the arrows between the hasse diagrams of the weather states, maps the weather states to themselves; therefore the ordering of the weather states is identical on both sides of the action and it is therefore homomorphic. The action $\bullet$ in Figure 8b is a monotone map[54]

$$if\ a \leq b\ then\ \bullet(a) \leq \bullet(b) \mid a, b \in F_i$$

where the structure the action preserves is the relative, rather than exact, ordering. Since groups are monoids with invertible operations, this definition of structure is also broad enough to include the Steven's measurment scales[55, 56]. Monoids are also commonly found in functional programming since the core property of monoids is composability [57].

As with the fiber F the total monoid space M is the cartesian product

$$M = M_0 \times \ldots \times M_i \times \ldots \times \ldots M_n \tag{7}$$

of each monoid M_i on F_i. The monoid is also added to the specification of the fiber $(c_i,\, T_i,\, \mathbb{U}_\sigma\, M_i)$

3.1.3 Continuity of the Data K

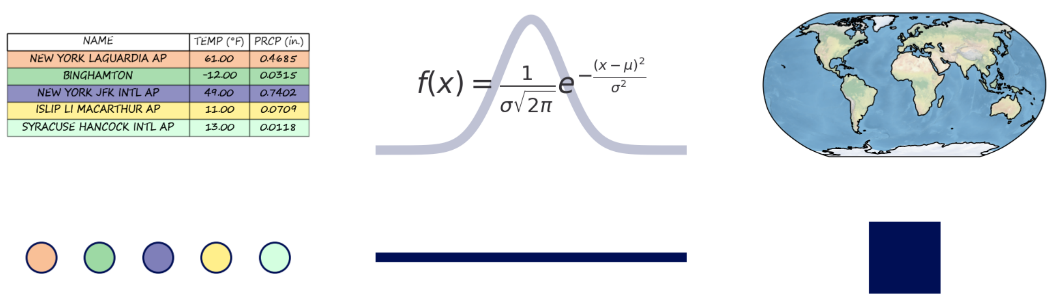


Figure 9: The topological base space K encodes the continuity of the data space, for example if the data is discrete points or lies on a plane or a sphere

The base space K acts as an indexing space, as emphasized by Butler[42, 43], to express how the records in E are connected to each other. As shown in Figure 9, K can have any number of dimensions and can be continuous or discrete. The base space also does not describe anything about the dataset besides the continuity. While the base space may have components to identify the continuity, such as *time, latitutde, longitude*, these labels are indexed into from K the same as any other component. This is similar to the notion of structural *keys* with associated *values* proposed by Munzner[46], but our model treats keys as a pure reference to topology. Decoupling the keys from their semantics allows the metadata to be altered; this provides for coordinate agnostic representation of the continuity and facilitates encoding of data where the independent variable may not be clear. For example the amount of snow on the ground is dependent on time of day and how much snow has fallen, and changing the coordinate system or time resolution should have no effect on how the records are connected to each other.

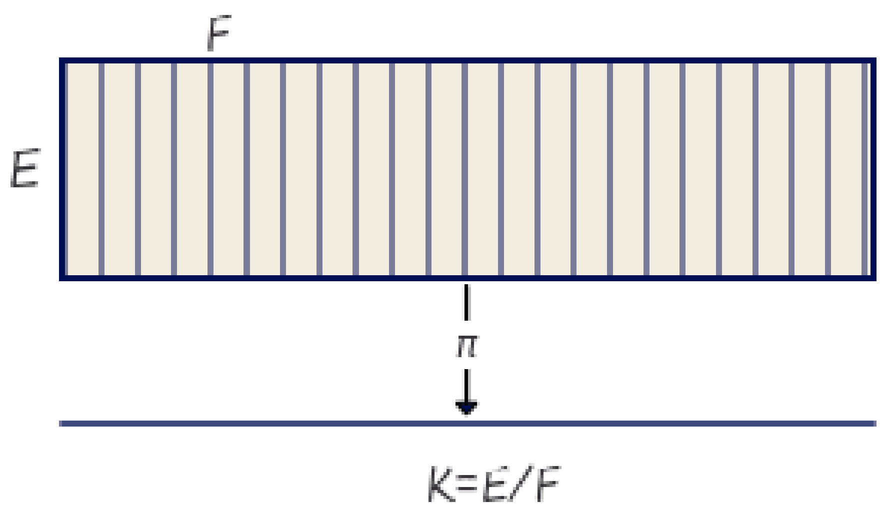


Figure 10: The base space E is divided into fiber segments F. The base space K acts as an index into the records in the fibers.

Formally K is the quotient space [58] of E meaning it is the finest space[59] such that every $k \in K$ has a corresponding fiber F_k[58]. In Figure 10, E is a rectangle divided by vertical fibers F, so the minimal K for which there is always a mapping $\pi : E \rightarrow K$ is the closed interval $[0, 1]$.

As with Equation 6 and Equation 7, we can decompose the total space into component bundles$\pi : E_i \rightarrow K$ where

$$\pi : E_1 \oplus \ldots \oplus E_i \oplus \ldots \oplus E_n \rightarrow K \tag{8}$$

such that M_i acts on component bundle E_i. The K remains the same because the connectivity of records does not change just because there are fewer components in each record. By encoding this continuity in the model as K the data model now explicitly carries information about its structure such that the implicit assumptions of the visualization algorithms are now explicit.

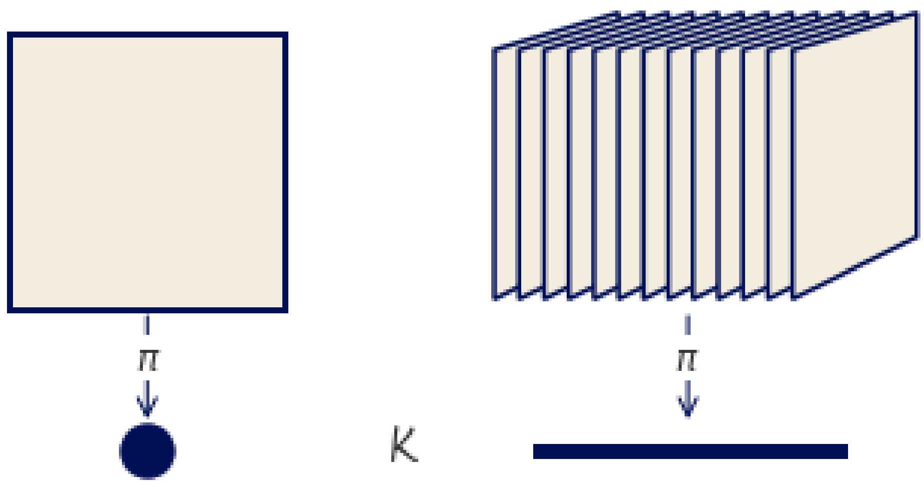


Figure 11: These two datasets have the same *(time, temperature)* fiber, but different continuities. The dataset on the left consists of discrete records, while the records in the dataset on the right sampled from a continuous space.

The datasets in Figure 11 have the same fiber of (temperature, time). The dot represents a discrete base space K, meaning that every dataset encoded in the fiber bundle has discrete continuity. The line is a representation of a 1D continuity, meaning that every dataset in the fiber bundle is 1D continous. By encoding this continuity in the model as K the data model now explicitly carries information about its structure such that the implicit assumptions of the visualization algorithms are now explicit. The explicit topology is a concise way of distinguishing visualizations that appear identical, for example heatmaps and images.

3.1.4 Data τ

While the projection function $\pi : E \to K$ ties together the base space K with the fiber F, a section $\tau : K \to E$ encodes a dataset. A section function takes as input location $k \in K$ and returns a record $r \in E$. For example, in the special case of a table [45], K is a set of row ids, F is the columns, and the section τ returns the record r at a given key in K. For any fiber bundle, there exists a map

$$\begin{tikzcd} F \arrow[r, hook] & E \arrow[d, "\pi"'] \\ & K \arrow[u, bend right, "\tau"'] \end{tikzcd} \tag{9}$$

such that $\pi(\tau(k)) = k$. The set of all global sections is denoted as $\Gamma(E)$. Assuming a trivial fiber bundle $E = K \times F$, the section is

$$\tau(k) = (k, (g_{F_0}(k), \ldots, g_{F_n}(k))) \tag{10}$$

where $g : K \to F$ is the index function into the fiber. This formulation of the section also holds on locally trivial sections of a non-trivial fiber bundle. Because we can decompose the bundle and the fiber (Equation 8, Equation 6), we can decompose τ as

$$\tau = (\tau_0, \ldots, \tau_i, \ldots, \tau_n) \tag{11}$$

where each section τ_i maps into a record on a component $F_i \in F$. This allows for accessing the data component wise in addition to accessing the data in terms of its location over K.

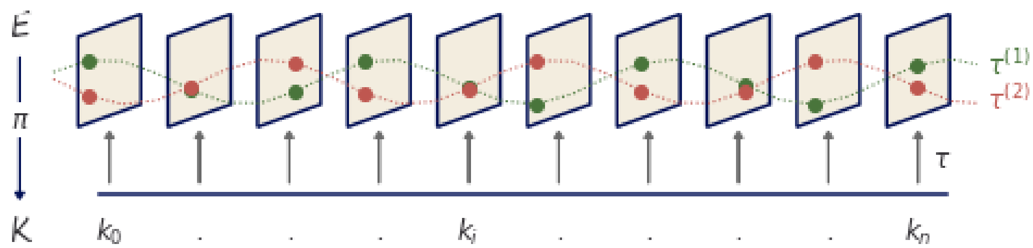


Figure 12: Fiber (time, temperature) with an interval K basespace. The sections $\tau^{(1)}$ and $\tau^{(2)}$ are constrained such that the time variable must be monotonic, which means each section is a timeseries of temperature values. They are included in the global set of sections $\tau^{(1)}, \tau^{(2)} \in \Gamma(E)$

In Figure 12, the fiber is the same encoding of (*time*, *temperature*) illustrated in Figure 7, and the base space is the interval K shown in Figure 11. The section $\tau^{(1)}$ is a function that for a point k returns a record in the fiber F. The section applied to a set of points in K resolves to a series of monotonically increasing in time records of *(time, temperature)* values. Section $\tau^{(2)}$ returns a different timeseries of *(time, temperature)* values. Both sections are included in the global set of sections $\tau^{(1)}, \tau^{(2)} \in \Gamma(E)$.

3.1.5 Sheafs

Many types of dynamic visualizations require evaluating sections on different subspaces of K, an a sheaf, denoted $\mathcal{O}$ provides a way to do so. A sheaf is a mathematical structure for defining collections of objects[60–62] on mathematical spaces. On the fiber bundle E, we can describe a sheaf as the collection of local sections $\iota^*\tau$

$$\begin{tikzcd} \iota^*E \arrow[r, hook, "\iota^*"] \arrow[d, "\pi"'] & E \arrow[d, "\pi"'] \\ U \arrow[r, hook, "\iota"] \arrow[u, bend right, "\iota^*\tau"'] & K \arrow[u, bend right, "\tau"'] \end{tikzcd} \tag{12}$$

which are sections of E pulled back over local neighborhood $U \subset E$ via the inclusion map $\iota : E \to U$. The collation of sections enabled by sheafs is necessary for navigation techniques such as pan and zoom[63] and dynamically updated visualizations such as sliding windows[64, 65].

3.1.6 Applications to Data Containers

This model provides a common formalism for widely used data containers without sacrificing the semantic structure embedded in each container. For example, the section can be any instance of a univariate numpy array[66] that stores an image. This could be a section of a fiber bundle where K is a 2D continuous plane and the F is $(\mathbb{R}^3, \mathbb{R}, \mathbb{R})$ where $\mathbb{R}^3$ is color, and the other two components are the x and y positions of the sampled data in the image. This position information is already implicitely encoded in the array as the index and the resolution of the image being stored.Instead of an image, the numpy array could also store a 2D discrete table. The fiber would not change, but the K would now be 0D discrete points. These different choices in topology indicate, for example, what sorts of interpolation would be appropriate when visualizing the data.

There are also many types of labeled containers that can richly be described in this framework because of the schema like structure of the fiber. For example, a pandas series which stores a labeled list, or a dataframe[67] which stores a relational table. A series could store the values of $\tau^{(1)}$ and a second series could be $\tau^{(2)}$. We could also fatten the fiber to hold two temperature series, such that a section would be an instance of a dataframe with a time column and two temperature columns. While the series and dataframe explicitly have a time index column, they are components in our model and the index is assumed to be data independent references such as hashvalues, virtual memory locations, or random number keys.

Where this model particularly shines are N dimensional labeled data structures. For example, an xarray[68] data that stores temperature field could have a K that is a continuous volume and the components would be the temperature and the time, latitude, and longitude the measurements were sampled at. A section can also be an instance of a distributed data container, such as a dask array [69]. As with the other containers, K and F are defined in terms of the index and dtypes of the components of the array. Because our framework is defined in terms of the fiber, continuity, and sections, rather than the exact values of the data, our model does not need to know what the exact values are until the renderer needs to fill in the image.

3.2 Graphic Space H

To establish that the artist is structure preserving map from data E to graphic H we construct a graphic bundle so that we can define *equivariance* in terms of maps on the fiber spaces and *continuity* in terms of maps on the base space. As with the data, we can represent the target graphic as a section ρ of a bundle (H, S, π, D).

$$\begin{tikzcd} D \arrow[r, hook] & H \arrow[d, "\pi"'] \\ & S \arrow[u, bend right, "\rho"'] \end{tikzcd} \tag{13}$$

The graphic bundle H consists of a base S(subsubsection 3.2.1) that is a thickened form of K a fiber D(subsubsection 3.2.2) that is an idealized display space, and sections ρ(**??**) that encode a graphic where the visual characteristics are fully specified.

3.2.1 Idealized Display D

To fully specify the visual characteristics of the image, we construct a fiber D that is an infinite resolution version of the target space. Typically H is trivial and therefore sections can be thought of as mappings into D. In this work, we assume a 2D opaque image $D = \mathbb{R}^5$ with elements

$$(x,\ y,\ r,\ g,\ b) \in D$$

such that a rendered graphic only consists of 2D position and color. To support overplotting and transparency, the fiber could be $D = \mathbb{R}^7$ such that $(x, y, z, r, g, b, a) \in D$ specifies the target display. By abstracting the target display space as D, the model can support different targets, such as a 2D screen or 3D printer.

3.2.2 Continuity of the Graphic S

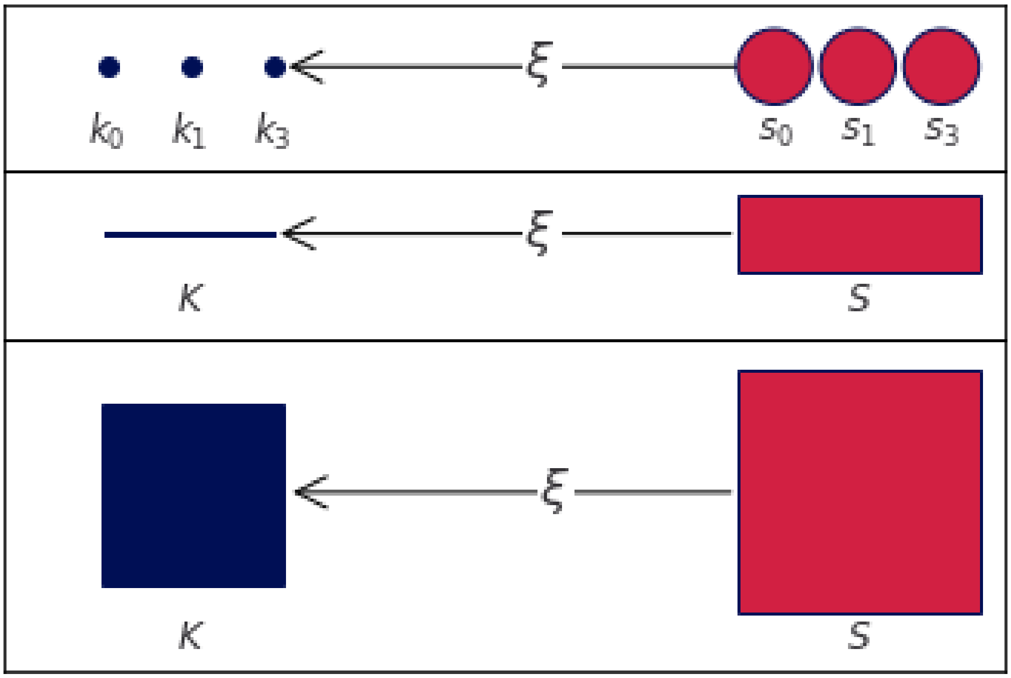


Figure 13: For a visualization component to preserve continuity, it must have a continuous surjective map $\xi : S \to K$ from graphic continuity to data continuity. The scatter and line graphic base spaces S have one more dimension of continuity than K so that S can encode physical aspects of the glyph, such as shape (a circle) or thickness. The image has the same dimension in S as in K.

To establish that a visualization component preserves continuity, we propose that their must be a continuous map $\xi : S \to K$ from the graphic base space to the data space. For example, consider a S that is mapped to the region of a 2D display space that represents K. For some visualizations, K may be lower dimension than S. For example, a point that is 0D in K cannot be represented on screen unless it is thickened to 2D to encode the connectivity of the pixels that visually represent the point. This thickening is often not necessary when the dimensionality of K matches the dimensionality of the target space, for example if K is 2D and the display is a 2D screen. We introduce S to thicken K in a way which preserves the structure of K.

Formally, we require that K be a deformation retract[70] of S so that K and S have the same homotopy, meaning there is a continuous map from S to K[71]. The surjective map $\xi : S \to K$

$$\begin{tikzcd} E \arrow[d, "\pi"'] & H \arrow[d, "\pi"'] \\ K & S \arrow[l, "\xi"'] \end{tikzcd} \tag{14}$$

goes from region $s \in S_k$ to its associated point s. This means that if $\xi(s) = k$, the record at k is copied over the region s such that $\tau(k) = \xi^*\tau(s)$ where $\xi^*\tau(s)$ is τ pulled back over S. When K is discrete points and the graphic is a scatter plot, each point $k \in K$ corresponds to a 2D disk S_k as shown in Figure 13. In the case of 1D continuous data and a line plot, the region β over a point α_i specifies the thickness of the line in S for the corresponding τ on k. The image has the same dimensions in data space and graphic space such that no extra dimensions are needed in S.

The mapping function ξ provides a way to identify the part of the visual transformation that is specific to the the connectivity of the data rather than the values; for example it is common to flip a matrix when displaying an image. The ξ mapping is also used by interactive visualization components to look up the data associated with a region on screen. One example is to fill in details in a hover tooltip, another is to convert region selection (such as zooming) on S to a query on the data to access the corresponding record components on K.

3.2.3 Graphic ρ

The section $\rho : S \to H$ is the graphic in an idealizes prerender space and also acts as a specification for rendering the graphic to an image. It is sufficient to sketch out how an arbitrary pixel would be rendered, where a pixel p in a real display corresponds to a region S_p in the idealized display. To determine the color of the pixel, we aggregate the color values over the region via integration:

$$r_p = \iint\limits_{S_p} \rho_r(s)ds^2$$
$$g_p = \iint\limits_{S_p} \rho_g(s)ds^2$$
$$b_p = \iint\limits_{S_p} \rho_b(s)ds^2$$

For a 2D screen, the pixel is defined as a region $p = [y_{top}, y_{bottom}, x_{right}, x_{left}]$ of the rendered graphic. Since the x and y in p are in the same coordinate system as the x and y components of D the inverse map of the bounding box $S_p = {\rho_{xy}}^{-1}(p)$ is a region $S_p \subset S$. The color is the result of the integration over S_p.

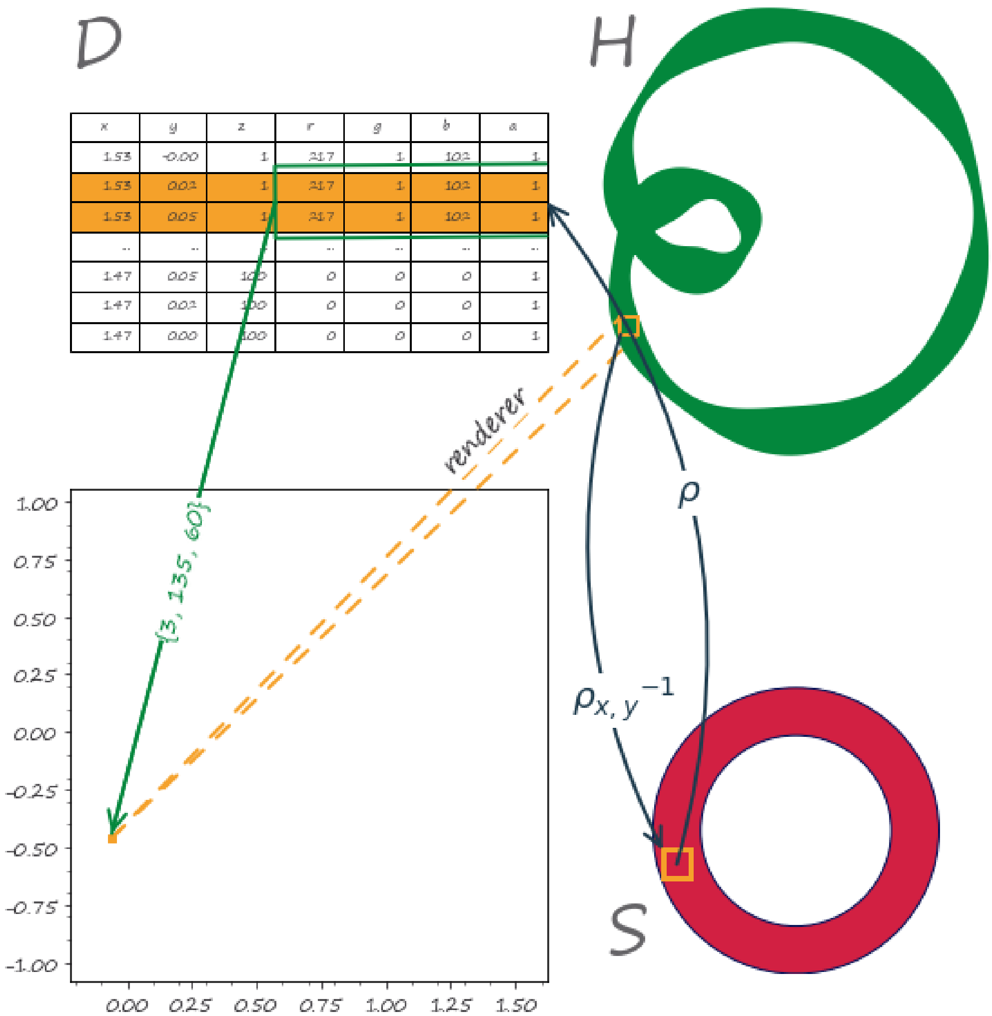


Figure 14: To render a graphic, a pixel p is selected in the display space, which is defined in the same coordinates as the x and y components in D via the renderer. In H the inverse mapping $\rho_{xy}{}^{(}p)$ returns a region $S_p \subset S$. $\rho(S_p)$ returns a set of points $(x, y, r, g, b) \in D$ that lie over S_p. The integral over the (r, g, b) pixels specifies that the pixel should be green

As shown in Figure 14, a pixel p in the output space, drawn in yellow, is selected and mapped, via the renderer, into a region on H. The region on H corresponds to a region $S_p \subset S$ via the inverse mapping $\rho_{xy}{}^{(}p)$. The base space S is an annulus to match the topology of the graphic idealized in H. The section $\rho(S_p)$ then maps into the fiber D over S_p to obtain the set of points in D, here represented as a table, that correspond to that section. The integral over the pixel components of this set of points in the fiber yields $\{3, 135, 60\}$ the actual color of the pixel. In general, ρ is an abstraction of rendering. In very broad strokes ρ can be a specification such as PDF[72], SVG[73], or an openGL scene graph[74]. Alternatively, ρ can be a rendering engine such as cairo[75] or AGG[76]. Implementation of ρ is out of scope for this work,

3.3 Artist

The topological artist A is how we model the building block component that transforms data into a graphic. The artist A is a map from the sheaf on a data bundle E which is $\mathcal{O}(E)$ to the sheaf on the graphic bundle H, $\mathcal{O}(H)$.

$$A : \mathcal{O}(E) \to \mathcal{O}(H) \tag{15}$$

The artist preserves *continuity* through the ξ map discussed in subsubsection 3.2.2 and is an *equivariant* map because it carries a homomorphism of monoid actions [77]

$$\varphi : M \to M' \tag{16}$$

Given M on data $\mathscr{E}$ and M' on graphic $\mathscr{H}$, we propose that artists $\mathscr{A}$ are equivariant maps

$$A(m \cdot r) = \varphi(m) \cdot A(r) \tag{17}$$

such that applying a monoid action $m \in M$ to the data input $r \in \mathscr{E}$ of the artist $\mathscr{A}$ is equivalent to applying a monoid action $\varphi(M) \in M'$ to the graphic $A(r) \in \mathscr{H}$ output of the artist.

The monoid equivariant map has two stages: the encoders $\nu : E' \to V$ convert the data components to visual components, and the assembly function $Q : \xi^* V \to H$ composites the fiber components of ξ^*V into a graphic in H.

$$\begin{tikzcd} E' \arrow[r, "\nu"] \arrow[dr, "\pi"'] & V \arrow[d, "\pi"] & \xi^*V \arrow[l, "\xi^*"'] \arrow[r, "Q"] \arrow[d, "\xi^*\pi"'] & H \arrow[dl, "\pi"] \\ & K & S \arrow[l, "\xi"] & \end{tikzcd} \tag{18}$$

ξ^*V is the visual bundle V pulled back over S via the equivariant continuity map $\xi : S \to K$ introduced in subsubsection 3.2.2. The functional decomposition of the visualization artist in Equation 18facilitates building reusable components at each stage of the transformation because the equivariance constraints are defined on ν, Q, and ξ. We name this map the artist as that is the analogous part of the Matplotlib[5] architecture that builds visual elements.

3.3.1 Visual Fiber Bundle V

We introduce a visual bundle V to store the mappings of the data components into components of the graphic. The visual bundle (V, K, π, P) is the space of possible parameters of a visualization type, such as a scatter or line plot. As with the data and graphic bundles, the visual bundle is defined by the projection map π

$$\begin{tikzcd} P \arrow[r, hook] & V \arrow[d, "\pi"'] \\ & K \arrow[u, bend right, "\mu"'] \end{tikzcd} \tag{19}$$

where μ is the visual variable encoding, as described by Bertin [9], of the data section τ. The visual fiber P is defined in terms of the input parameters of the visualization library's plotting functions; by making these parameters explicit components of the fiber, we can build consistent definitions and expectations of how these parameters behave.

$\boldsymbol{\nu_i}$	$\boldsymbol{\mu_i}$	$\boldsymbol{codomain(\nu_i) \subset P_i}$
position	x, y, z, theta, r	$\mathbb{R}$
size	linewidth, markersize	$\mathbb{R}^+$
shape	markerstyle	$\{f_0, \ldots, f_n\}$
color	color, facecolor, markerfacecolor, edgecolor	$\mathbb{R}^4$
texture	hatch	$\mathbb{N}^{10}$
	linestyle	$(\mathbb{R}, \mathbb{R}^{+^{n, n\%2=0}})$

Table 1: Some possible components of the fiber P for a visualization function implemented in Matplotlib

A section μ is a tuple of visual values that specifies the visual characteristics of a part of the graphic. For example, given a fiber of $\{x, y, color\}$ one possible section could be $\{.5, .5, (255, 20, 147)\}$. The $codomain(\nu_i)$ determines which monoids can act on P_i. These fiber components are implicit in the library, as seen in Table 1, and by making them explicit as components of the fiber we can build consistent definitions and expectations of how these parameters behave.

3.3.2 Visual Encoders ν

We define the visual transformers ν

$$\{\nu_0, \ldots, \nu_n\} : \{\tau_0, \ldots, \tau_n\} \mapsto \{\mu_0, \ldots, \mu_n\} \tag{20}$$

as the set of equivariant maps $\nu_i : \tau_i \mapsto \mu_i$. Given M_i is the monoid action on E_i and that there is a monoid $M_i{}'$ on V_i, then there is a monoid homomorphism from $\varphi : M_i \to M_i{}'$ that ν must preserve. As mentioned in subsubsection 3.1.2, monoid actions define the structure on the fiber components and are therefore the basis for equivariance. A validly constructed ν is one where the diagram of the monoid transform m commutes

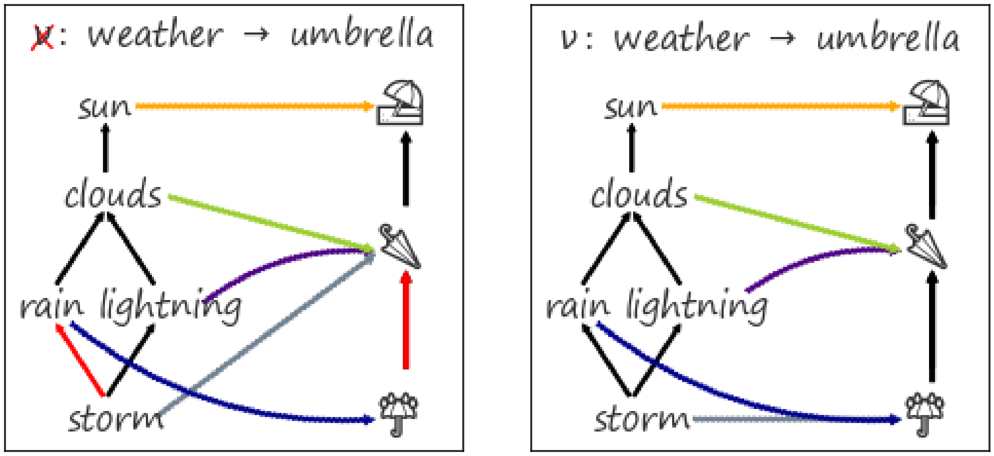
$$\begin{tikzcd} E_i \arrow[r, "\nu_i"] \arrow[d, "m_r"'] & V_i \arrow[d, "m_v"] \\ E_i \arrow[r, "\nu_i"] & V_i \end{tikzcd} \tag{21}$$

such that applying equivariant monoid actions to E_i and V_i preserves the map $\nu_i : E_i \to V_i$. In general, the data fiber F_i cannot be assumed to be of the same type as the visual fiber P_i and the actions of M on F_i cannot be assumed to be the same as the actions of M' on P; therefore an equivariant ν_i must satisfy the constraint

$$\nu_i(m_r(E_i)) = \varphi(m_r)(\nu_i(E_i)) \tag{22}$$

such that φ maps a monoid action on data to a monoid action on visual elements. However, we can construct a monoid action of M on P_i that is compatible with a monoid action of M on F_i. We can compose the monoid actions on the visual fiber $M' \times P_i \to P_i$ with the homomorphism φ that takes M to M'. This allows us to define a monoid action on P of M that is $(m, v) \to \varphi(m) \bullet v$. Therefore, without a loss of generality, we can assume that an action of M acts on F_i and on P_i compatibly such that φ is the identity function.

The translation from weather state data to visual representation as umbrella emoji in Figure 15a is an invalid visual encoding map ν because it is not homomorphic. This is because the monotonic condition $rain \geq storm \implies \nu(rain) \geq \nu(storm)$ is not met since $\nu(rain) \leq \nu(storm)$. To satisfy the monotonic condition for $rain \geq storm$, either red arrow in Figure 15a would have to go in a different direction. On the other hand, the mapping from weather state to umbrellla in Figure 15b is a homomorphism since $\nu(rain) = \nu(storm)$ satisfies the monotonic condition of $rain \geq storm$. Figure 15 is an example of how the



(a) (b)

Figure 15: The map from data component to visual component in Figure 15a is not homomorphic, and therefore invalid, because $rain \geq storm$ is mapped to elements with the reverse ordering $\nu(storm) \geq \nu(storm)$. In contrast, the mapping in Figure 15b is valud since $\nu(storm) = \nu(rain)$ satisfies the condition $\nu(storm) \geq \nu(storm)$

model supports partially ordered data components, which was a motivation for defining equivariance as monoid homomorphisms.

scale	**group**	**constraint**
nominal	permutation	if $r_1 \neq r_2$ then $\nu(r_1) \neq \nu(r_2)$
ordinal	monotonic	if $r_1 \leq r_2$ then $\nu(r_1) \leq \nu(r_2)$
interval	translation	$\nu(x + c) = \nu(x) + c$
ratio	scaling	$\nu(xc) = \nu(x) * c$

Table 2

The Stevens measurement types[55], listed in Table 2, are specified in terms of groups, which are monoids with invertible operations[78]. Despite critiques of the scales[79, 80], we believe it is critical for the model to include the measurement scales since they are commonly used in visualization to classify components [22, 46]. By specifying the equivariance constraints on ν we can guarantee that the stage of the artist that transforms data components into visual representations is equivariant. These constraints guide the implementation of reusable component transformers ν that are composed when generating the graphic.

3.3.3 Visualization Assembly

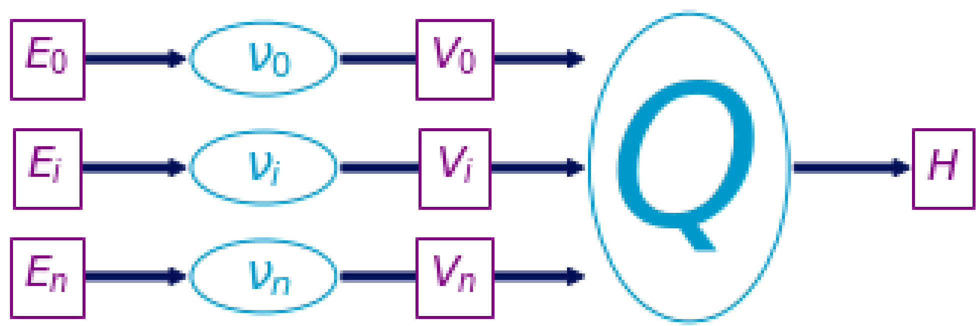


Figure 16: The transform functions ν_i convert data $\tau_i \in E$ to visual characteristics $\mu_i \in V$, then Q assembles μ_i into a graphic $\rho \in H$.

The transformation from data into graphic is analagous to a map-reduce operation; as illustrated in **??**, data components E_i are mapped into visual components V_i that are reduced into a graphic in H. The space of all graphics that Q can generate is the subset of graphics reachable via applying the reduction function $Q(\Gamma(V)) \in \Gamma(H)$ to the visual section $\mu \in \Gamma(V)$. The full space of graphics is not necessarily equivariant; therefore we formalize the constraints on Q such that it produces structure preserving graphics.

We formalize the expectation that visualization generation functions parameterized in the same way should generate the same functions as the equivariant map $Q : \mu \mapsto \rho$. We then define the constraint on Q such that if Q is applied to two visual sections μ and μ' that generate the same ρ then the output of μ and μ' acted on by the same monoid m must be the same. We do not define monoid actions on all of $\Gamma(H)$ because there may be graphics $\rho \in \Gamma(H)$ for which we cannot construct a valid mapping from V. Lets call the

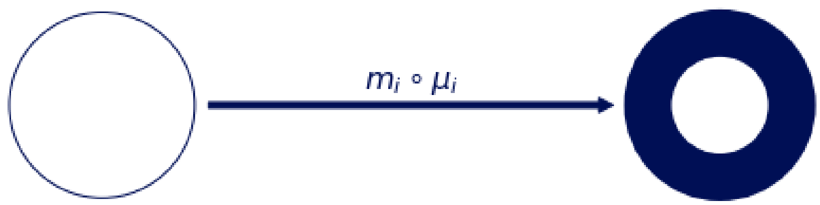


Figure 17: These two glyphs are generated by the same annulus Q function. The monoid action m_i on edge thickness μ_i of the first glyph yields the thicker edge $\mu_i{}'$ in the second glyph.

visual representations of the components $\Gamma(V) = X$ and the graphic $Q(\Gamma(V)) = Y$

Proposition 1. *If for elements of the monoid $m \in M$ and for all $\mu, \mu' \in X$, we define the monoid action on X so that it is by definition equivariant*

$$Q(\mu) = Q(\mu') \implies Q(m \circ \mu) = Q(m \circ \mu') \tag{23}$$

then a monoid action on Y can be defined as $m \circ \rho = \rho'$. If and only if Q satistfies Equation 23, we can state that the transformed graphic $\rho' = Q(m \circ \mu)$ is equivariant to a monoid action applied on Q with input $\mu \in Q^{-1}(\rho)$ that must generate valid ρ.

For example, given fiber $P = (xpos, ypos, color, thickness)$, then sections $\mu = (0, 0, 0, 1)$ and $Q(\mu) = \rho$ generates a piece of the thin hollow circle. The action $m = (e, e, e, x + 2)$, where e is identity, translates μ to $\mu' = (e, e, e, 3)$ and the corresponding action on ρ causes $Q(\mu')$ to be the thicker circle in Figure 17.

We formally describe a glyph as Q applied to the regions k that map back to a set of path connected components $J \subset K$ as input

$$J = \{j \in K \text{ s. t. } \exists\gamma \text{ s.t. } \gamma(0) = k \text{ and } \gamma(1) = j\} \tag{24}$$

where the path[81] γ from k to j is a continuous function from the interval [0,1]. We define the glyph as the graphic generated by $Q(S_j)$

$$H \underset{\rho(S_j)}{\rightleftarrows} S_j \underset{\xi^{-1}(J)}{\overset{\xi(s)}{\rightleftarrows}} J_k \tag{25}$$

such that for every glyph there is at least one corresponding region on K, in keeping with the definition of glyph as any visually differentiable element put forth by Ziemkiewicz and Kosara[82]. The primitive point, line, and area marks[9, 83] are specially cased glyphs.

3.3.4 Assembly Q

Given the continuities described in 13, we illustrate a minimal Q that will generate the most minimal visualizations associated with those continuities: non-overlapping scatter points, a non-infinitely thin line, and an image.

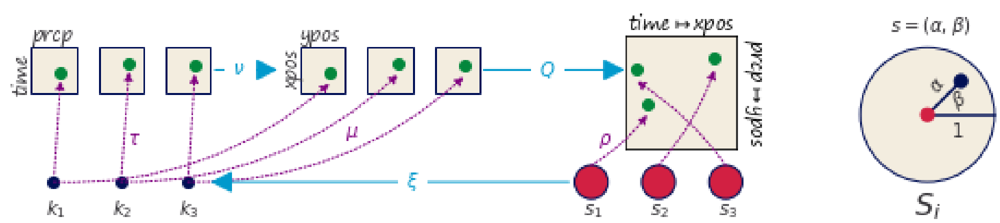


Figure 18: The data is discrete points (temperature, time). Via ν these are converted to (xpos, ypos) and pulled over discrete S. These values are then used to parameterize ρ which returns a color based on the parameters (xpos,ypos) and position α, β on S_k that ρ is evaluated on.

The scatter plot in Figure 18 can be defined as

$$Q(xpos, ypos)(\alpha, \beta) \tag{26}$$

with a constant *size* and color $\rho_{RGB} = (0, 0, 0)$ that are defined as part of Q. The position of this swatch of color can be computed relative to the location on the disc $(\alpha, \beta) \in S_k$ as shown in Figure 18

$$x = size * \alpha \cos(\beta) + xpos$$

$$y = size * \alpha \sin(\beta) + ypos$$

such that $\rho(s) = (x, y, 0, 0, 0)$ colors the point (x,y) black.

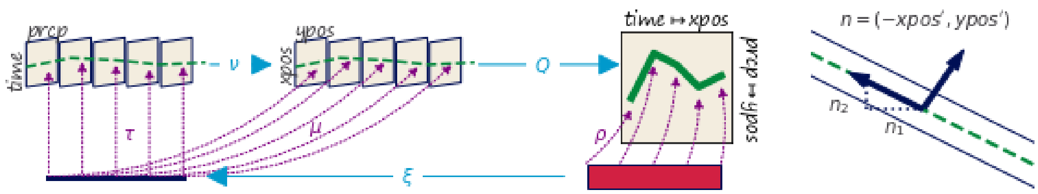


Figure 19: The line fiber (*time*, *temp*) is thickened with the derivative (*time*′, *temperature*′ because that information will be necessary to figure out the tangent to the point to draw a line. This is because the line needs to be pushed perpendicular to the tangent of (xpos, ypos). The data is converted to visual characteristics (xpos, ypos). The α coordinates on S specifies the position of the line, the β coordinate specifies thickness.

In contrast, the line plot

$$Q(xpos, \hat{n}_1, ypos, \hat{n}_2)(\alpha, \beta) \tag{27}$$

in **??** has a ξ function that is not only parameterized on k but also on the α distance along k and corresponding region in S. As shown in **??**, line needs to know the tangent of the data to draw an envelope above and below each (xpos,ypos) such that the line appears to have a thickness; therefore the artist takes as input the jet bundle [84, 85] $\mathcal{J}^2(E)$ which is the data E and the first and second derivatives of E. The magnitude of the slope is $|n| = \sqrt{{n_1}^2 + {n_2}^2}$ such that the normal is $\hat{n}_1 = \frac{n_1}{|n|}$, $\hat{n}_2 = \frac{n_2}{|n|}$ which yields components of ρ

$$x = xpos(\xi(\alpha)) + width * \beta \hat{n}_1(\xi(\alpha))$$

$$y = ypos(\xi(\alpha)) + width * \beta \hat{n}_2(\xi(\alpha))$$

where (x,y) look up the position $\xi(\alpha)$ on the data and the derivatives $\hat{n}_1, \hat{n}_2$. The derivatives are then multiplied by a *width* parameter to specify the thickness.

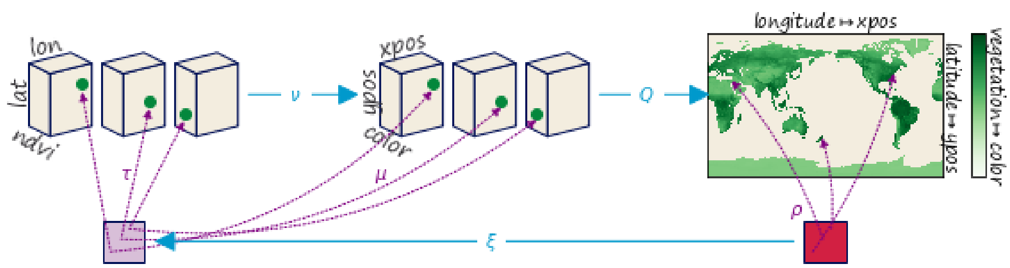


Figure 20: The only visual parameter an image requires is color since ξ encodes the mapping between position in data and position in graphic.

In Figure 20, the image

$$Q(xpos, ypos, color) \tag{28}$$

is a direct lookup into $\xi : S \to K$. The indexing variables (α, β) define the distance along the space, which is then used by ξ to map into K to lookup the color values

$$R = R(\xi(\alpha, \beta)),\ G = G(\xi(\alpha, \beta)),\ B = B(\xi(\alpha, \beta))$$

In the case of an image, the indexing mapper ξ may do some translating to a convention expected by Q, for example reorientng the array such that the first row in the data is at the bottom of the graphic.

The graphic base space S is not accessible in many architectures, including Matplotlib; instead we can construct a factory function $\hat{Q}$ over K that can build a Q. As shown in Equation 18, Q is a bundle map $Q : \xi^*V \to H$ where ξ^*V and H are both bundles over S.

The preimage of the continuity map $\xi^{-1}(k) \subset S$ is such that many graphic continuity points $s \in S_K$ go to one data continuity point k; therefore, by definition the pull back of μ

$$\xi^*V\mid_{\xi^{-1}(k)} = \xi^{-1}(k) \times P \tag{29}$$

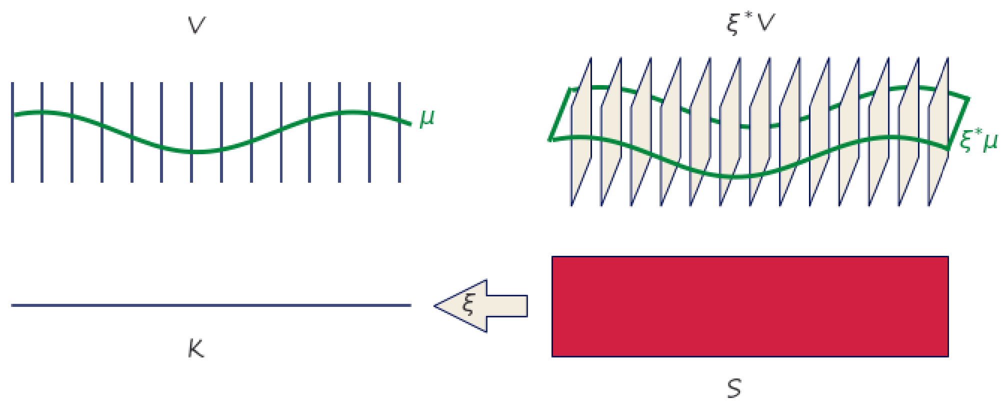


Figure 21: Because the pullback of the visual bundle ξ^*V is the replication of a μ over all points s that map back to a single k, we can constract a $\hat{Q}$ on μ over k that will fabricate the Q for the equivalent region of s associated to that k

copies the visual fiber P over the the points s in graphic space S that correspond to one k in data space K. This set of points s are the preimage $\xi^{-1}(k)$ of k.

As shown in Figure 21, given the section $\xi^*\mu$ pulled back from μ and the point $s \in \xi^{-1}(k)$, there is a direct map $(k, \mu(k)) \mapsto (s, \xi^*\mu(s))$ from μ over k to the section $\xi^*\mu$ over s. This means that the pulled back section $\xi^*\mu(s) = \xi^*(\mu(k))$ is the section μ copied over all $s$such that $\xi^*\mu$ is identical for all s where $\xi(s) = k$. In Figure 21 each dot on P is equivalent to the line on $P^*\mu$.

Given the equivalence between μ and $\xi^*\mu$ defined above, the reliance on S can be factored out. When Q maps visual sections into graphics $Q : \Gamma(\xi^*V) \to \Gamma(H)$, if we restrict Q input to $\xi^*\mu$ then the graphic section ρ evaluated on a visual region s

$$\rho(s) \coloneqq Q(\xi^*\mu)(s) \tag{30}$$

is defined as the assembly function Q with input $\xi^*\mu$ evaluated on s. Since the pulled back section $\xi^*\mu$ is the section μ copied over every graphic region $s \in \xi^{-1}(k)$, we can define a Q factory function

$$\hat{Q}(\mu(k))(s) \coloneqq Q((\xi^*\mu)(s)) \tag{31}$$

where $\hat{Q}$ with input μ is defined to Q that takes as input the copied section $\xi^*\mu$ such that both functions are evaluated over the same location $\xi^{-1}(k) = s$ in the base space S. Factoring out s from Equation 31 yields

$$\hat{Q}(\mu(k)) = Q(\xi^*\mu) \tag{32}$$

where Q is no longer bound to input but $\hat{Q}$ is still defined in terms of K. In fact, $\hat{Q}$ is a map from visual space to graphic space $\hat{Q} : \Gamma(V) \rightarrow \Gamma(H)$ locally over k such that it can be evaluated on a single visual record $\hat{Q} : \Gamma(V_k) \rightarrow \Gamma(H\mid_{\xi^{-1}(k)})$. This allows us to construct a $\hat{Q}$ that only depends on K, such that for each $\mu(k)$ there is part of $\rho\mid_{\xi^{-1}(k)}$. The construction of $\hat{Q}$ allows us to retain the functional map reduce benefits of Q without having to majorly restructure the existing pipeline for libraries that delegate the construction of ρ to a back end such as Matplotlib.

3.3.5 Composition of Artists: +

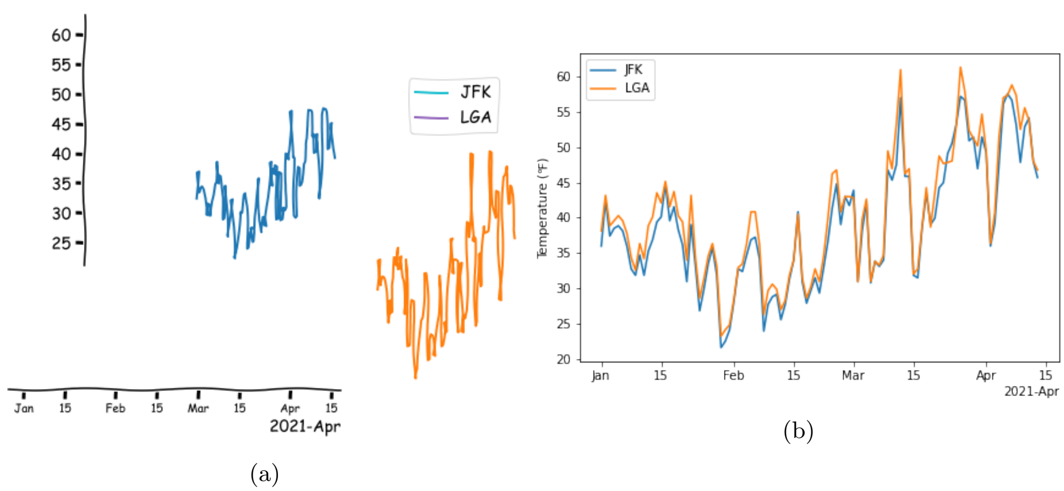


Figure 22: Each of the visual elements in ?? is generated via a unique artist A. In Figure 22a, they are added to the image independent of the other elements, creating an incoherant visualization. In Figure 22b, these artists are composited before being added to the image. Disjoint union of E aligns the two timeseries with the x and y axis so all these elements use a shared coordinate system. A more complex composition dictates that the legend is connected to the E such that it must use the same color as the data it is identifying.

We describe the constraints for compositing artists by defining addition operators. Given the family of artists $(E_i : i \in I)$ that are rendered to the same image, the $+$ operator

$$+ \coloneqq \bigsqcup_{i \in I} E_i \tag{33}$$

defines a simple composition of artists. For example, the components in Figure 22a are each generated by different artists, and a visualization of solely the x axis is rarely all that useful. In Figure 22a, these artists are all added to the image independently of the other and therefore there are no constraints on how they are generated in conjunction with each other. In Figure 22b, the data is joined via disjoint union; doing so aligns the components in F such the ν to the same component in P targets the same coordinate system. When artists share a base space $K_2 \hookrightarrow K_1$, a composition operator can be defined such that the artists are acting on different components of the same section. This type of composition is important for visualizations where elements update together in a consistent way, such as multiple views [86, 87] and brush-linked views[88, 89].

3.3.6 Equivalance class of artists A'

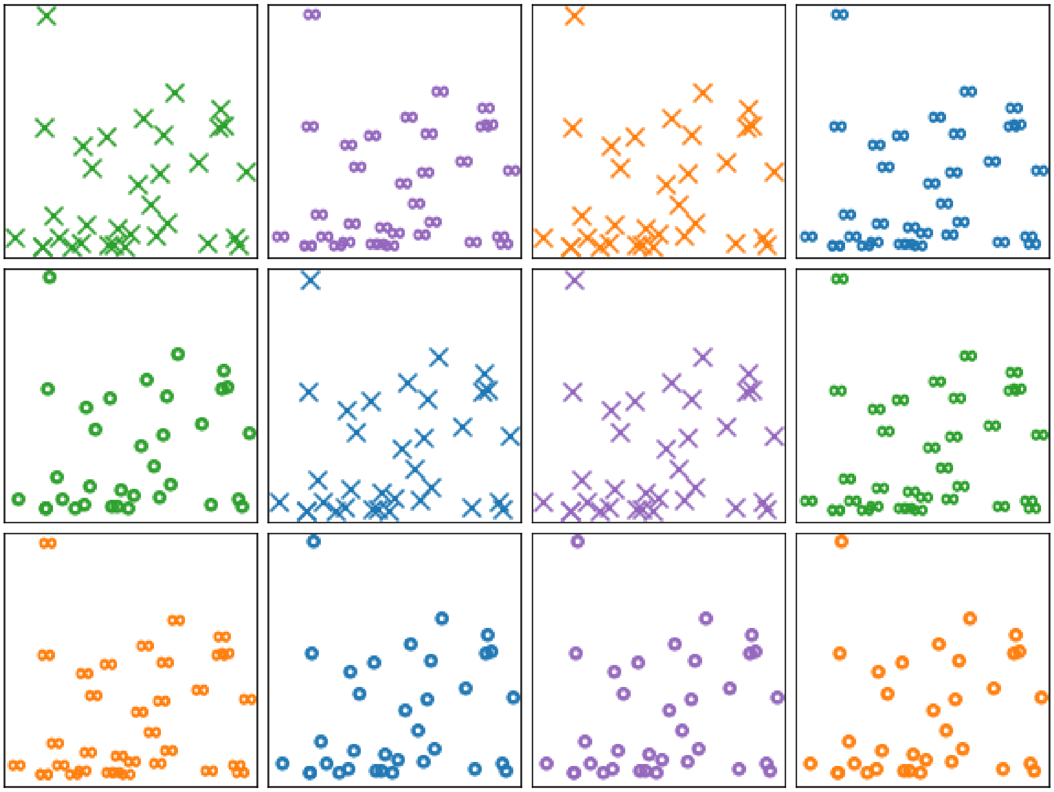


Figure 23: Each scatter plot is generated via a unique artist function A_i, but they only differ in aesthetic styling. Therefore, these artists are all members of an equivalence class $A_i \in A\prime$

Representational invariance, as defined by Kindlmann and Scheidegger, is the notion that visualizations are equivalent if changing the visual representation, such as colors or shapes, does not change the meaning of the visualization[12]. We propose that visualizations are invariant if they are generated by artists that are members of an equivalence class

$$\{A \in A' : A_1 \equiv A_2\}$$

For example, every scatter plot in Figure 23 is a scatter of the same datasets mapped to the x *position* and y *position* in the same way. The scatter plots only differ in the choice of constant visual literals, differing in color and marker shape. Each scatter is generated by an artist A_i, and every scatter is generated by a member of the equivalence class $A_i \in A\prime$. Since it is impractical to implement a new artist for every single graphic, the equivalance class provides a way to evaluate an implementation of a generalized artist. Given equivalent, but no necessarily identical, ν, Q, and ξ, two artists are equivalent. This criteria also allows for comparing artists across libraries.

4 Prototype: Matplottoy

To evaluate the feasibility of the model described in ??, we built prototypes of a `point`, `line`, and `bar` artist. We make use of the Matplotlib figure and axes artists [4, 5] so that we can initially focus on the data to graphic transformations and exploit the Matplotlib transform stack to convert data coordinates into screen coordinates. While the artist is specified in a fully functional manner in Equation 18, we implement the prototype in a heavily object oriented manner. We do so mostly to more easily manage function inputs, especially parameters that are passed through to methods that are structurally functional.

```
fig, ax = plt.subplots()
artist = Artist(E, V)
ax.add_artist(artist)
```

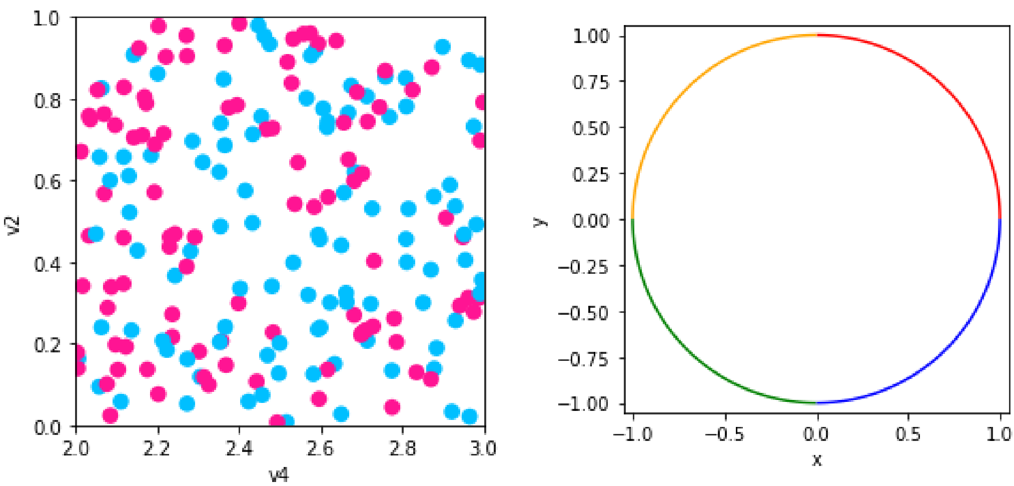
Building on the curr ent Matplotlib artists which construct an internal representation of the graphic, `ArtistClass` acts as an equivalence class artist A' as described in ??. The visual bundle V is specified as the `V` dictionary of the form `{parameter:(variable name, encoder)}` where parameter is a component in P, variable is a component in F, and the ν encoders are passed in as functions or callable objects. The data bundle E is passed in as a `E` object. By binding data and transforms to A' inside `__init__`, the `draw` method is a fully specified artist A as defined in Equation 15.

```python
class ArtistClass(matplotlib.artist.Artist): #A'
    def __init__(self, E, V, *args, **kwargs):
        # properties that are specific to the graphic
        self.E = E
        self.V = V
        super().__init__(*args, **kwargs)

    def hat_Q(self, **args):
        # set the properties of the graphic

    def draw(self, renderer):
        # returns K, indexed on fiber then key
        tau = self.E.view(self.axes)
        # visual channel encoding applied fiberwise
        mu = {p: nu(tau(c))
                  for p, (c, nu) in self.V.items()}
        self.hat_q(**mu)
        # pass configurations off to the renderer
        super().draw(renderer)
```

The data is fetched in section τ via a `view` method on the data because the input to the artist is a section on E. The `view` method takes the `axes` attribute because it provides the region in graphic coordinates S that can be used to query back into data to select a subset as described in subsubsection 3.1.5. To ensure the integrity of the section, `view` must be atomic, which means that the values cannot change after the method is called in draw until a new call to draw[27]. We put this constraint on the return of the `view` method so that we do not risk race conditions.

The ν functions are then applied to the data, as describe in Equation 20, to generate the visual section μthat here is the object `V`. The conversion from data to visual space is simplified here to directly show that it is the encoding ν applied to the component. The `q_hat` function that is $\hat{Q}$, as defined in Equation 32, is responsible for generating a representation such that it could be serialized to recreate a static version of the graphic. This artist is not optimized because we prioritized demonstrating the separability of ν and $\hat{Q}$. The last step in the artist function is handing itself off to the renderer. The extra `*arg, **kwargs` arguments in `__init__,draw` are artifacts of how these objects are currently implemented.

4.1 Scatter, Line, and Bar Artists



(a) (b)

Figure 24: Scatter plot and line plot implemented using prototype artists and data models, building on Matplotlib rendering.

The figure in Figure 24a is described by Equation 26. This is implemented via a `Line` object where the scatter marker shape is fixed as a circle, and the visual fiber components are x and y position and the facecolor and size of the marker. We only show the `q_hat` function here because the `__init__, draw` are identical the prototype artist.

The `view` method repackages the data as a fiber component indexed table of vertices. Even though the `view` is fiber indexed, each vertex at an index $k$has corresponding values in section $\tau(k_i)$. This means that all the data on one vertex maps to one glyph.

```
class Point(ArtistClass, mcollections.Collection):
  def q_hat(self, x, y, s, facecolors): #\hat{Q}
    # construct geometries of circle glyphs
    self._paths = [mpath.Path.circle((xi,yi), radius=si)
                for (xi, yi, si) in zip(x, y, s)]
    # set attributes of glyphs, these are vectorized
    # circles and facecolors are lists of the same size
    self.set_facecolors(facecolors)
```

In `q_hat`, the μ components are used to construct the vector path of each circular marker with center `(x,y)` and size `x` and set the colors of each circle. This is done via the `Path.circle` object.

```
class Line(ArtistClass, mcollections.LineCollection):
  def q_hat(self, x, y, color): #\hat{Q}
    #assemble line marks as set of segments
    segments = [np.vstack((vx, vy)).T for vx, vy
                in zip(x, y)]
    self.set_segments(segments)
    self.set_color(color)
```

To generate Figure 24b, the `Line` artist `view` method returns a table of edges. Each edge consists of (x,y) points sampled along the line defined by the edge and information such as the color of the edge. As with `Point`, the data is then converted into visual variables. In `q_hat`, described by Equation 27, this visual representation is composed into a set of line segments, where each segment is the array generated by `np.vstack((vx, vy))`. Then the colors of each line segment are set. The colors are guaranteed to correspond to the correct segment because of the atomicity constraint on the view. The current implementation of line in Matplotlib does not have this functionality because it has no notion of an atomic view τ, instead the user keeps track of which values are input into a single call of $\hat{Q}$ by creating a loop where every color is a new call to the line plotting function.

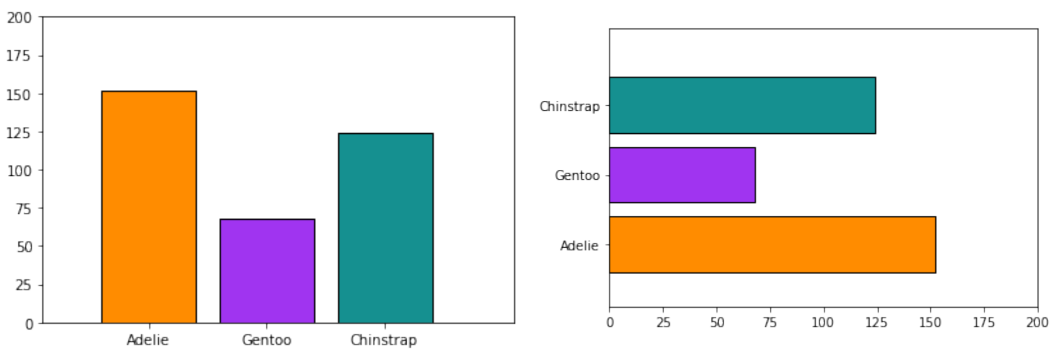


Figure 25: Frequency of Penguin types visualized as discrete bars.

The bar charts in figure 25 are generated with a `Bar` artist. The artist has required visual parameters P of (position, length), and an additional parameter `orientation` which controls whether the bars are arranged vertically or horizontally. This parameter only applies holistically to the graphic and never to individual data parameters, and highlights how the model encourages explicit differentiation between parameters in V and graphic parameters applied directly to $\hat{Q}$.

```python
class Bar(ArtistClass, mcollections.Collection):
    def __init__(self, data, transforms, orientation='v', *args, **kwargs):
        """
        orientation: str, optional
            v: bars aligned along x axis, heights on y
            h: bars aligned along y axis, heights on x
        """
```

```
        self.orientation = orientation
        # set E and V

    def q_hat(self, position, length, floor=0, width=0.8,
                    facecolors='C0', edgecolors='k', offset=0):
        #set some defaults
        width = itertools.repeat(width) if np.isscalar(width) else width
        floor = itertools.repeat(floor) if np.isscalar(floor) else (floor)

        # offset is passed through via assemblers such as multigroup,
        # not supposed to be directly tagged to position
        position = position + offset

        def make_bars(xval, xoff, yval, yoff):
             return [[(x, y), (x, y+yo), (x+xo, y+yo), (x+xo, y), (x, y)]
                for (x, xo, y, yo) in zip(xval, xoff, yval, yoff)]
        #build bar glyphs based on graphic parameter
        if self.orientation in {'vertical', 'v'}:
            verts = make_bars(position, width, floor, length)
        elif self.orientation in {'horizontal', 'h'}:
            verts = make_bars(floor, length, position, width)

        self._paths = [mpath.Path(xy, closed=True) for xy in verts]
        self.set_edgecolors(edgecolors)
        self.set_facecolors(facecolors)
```

As with `Point` and `scatter`, `q_hat` function constructs bars and sets their properties, face and edge colors. The `make_bars` function converts the input position and length to the coordinates of a rectangle of the given width. Defaults are provided for 'width' and 'floor' to make this function more reusable. Typically the defaults are used for the type of chart shown in figure 25, but these visual variables are often set when building composite versions of this chart type as discussed in section 4.4.

4.2 Visual Encoders

The visual parameter serves as the dictionary key because the visual representation is constructed from the encoding applied to the data $\mu = \nu \circ \tau$. For the scatter plot, the mappings for the visual fiber components $P = (x, y, facecolors, s)$ are defined as

```
cmap =  color.Categorical({'true':'deeppink',
                           'false':'deepskyblue'})
# {P_i name:{'name':c_i, 'encoder:\nu_i}}
V = {'x': {'name': 'v4', 'encoder': lambda x: x},
     'y': {'name': 'v2', 'encoder': lambda x: x},
     'facecolors': {'name':'v3', 'encoder': cmap},
     's':{'name': None ,
          'encoder': lambda _: itertools.repeat(.02)}}
```

where `lambda x: x` is an identity ν, `{'name':None}` maps into P without corresponding τ to set a constant visual value, and `color.Categorical` is a custom ν implemented as a class for reusability.

```
#\nu_i(m_r(E_i)) = \varphi(m_r)(\nu_i((E_i))
def test_nominal(values, encoder):
    m1 = list(zip(values, encoder(values)))
    random.shuffle(values)
    m2 = list(zip(values, encoder(values)))
    assert sorted(m1) == sorted(m2)
```

As described in Equation 22, a test for equivariance can be implemented trivially. It is currently factored out of the artist for clarity.

4.3 Data Model

The data input into the `Artist` will often be a wrapper class around an existing data structure. This wrapper object must specify the fiber components F and connectivity K and have a `view` method that returns an atomic object that encapsulates τ. To support specifying the fiber bundle, we define a `FiberBundle` data class[90]

```
@dataclass
class FiberBundle:
    K: dict #{'tables': []}
    F: dict #  {variable name: type}
```

that asks the user to specify the the properties of F and the K connectivity as either discrete vertices or continuous data along edges. To generate the scatter plot and the line plot, the distinction is in the `tau` method that is the section.

```
class PointData:
    def __init__(self):
        self.FB = FiberBundle({'tables': ['vertex']},
                              {'v1': float, 'v2': str, 'v3': float})
    def tau(self, k):
        return # tau evaluated at one point k

class LineData:
    def __init__(self):
        self.FB = FiberBundle({'tables': ['edge']},
```

```
                        {'x': float, 'y': float, 'color': str})
    def tau(self, k):
        return # tau evaluated on interval k
```

The discrete `tau` method returns a record of discrete points, akin to a row in a table, while the line`tau` returns a sampling of points along an edge k.

```
def view(self, axes):
    table = defaultdict(list)
    for k in self.keys:
        table['index'].append(k)
        for (name, value) in zip(self.FB.fiber.keys(),
                                 self.tau(k)[1]):
            table[name].append(value)
    return table
```

In both cases, the `view` method packages the data into a data structure that the artist can unpack via data component name, akin to a table with column names when K is 0 or 1 D.

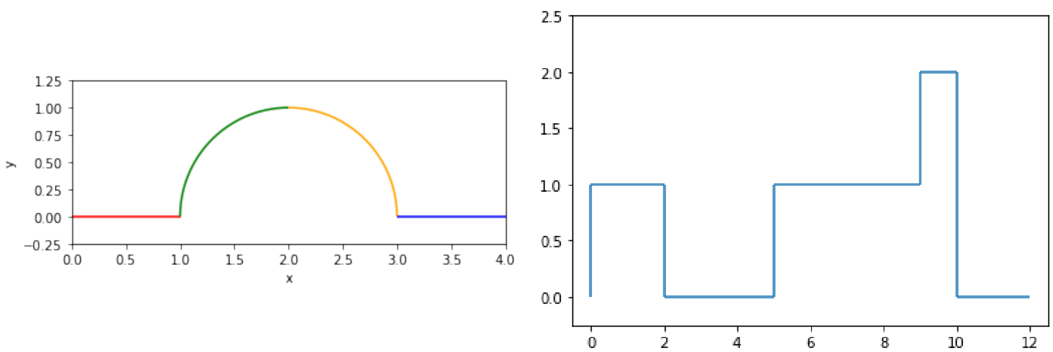


Figure 26: Continuous and discontinuous lines as defined via the same data model, and generated with the same A'`Line`

The graphics in figure Figure 26 are made using the `Line` artist and the `GraphData` data source where if told that the data is connected, the data source will check for that connectivity by constructing an adjacency matrix. The multicolored line is a connected graph of edges with each edge function evaluated on 100 samples,

```
GraphData(FB, edges, verticies, num_samples=100, connect=True)
```

which is an arbitrary choice made to display a smooth curve. The axes can also be used to choose an appropriate number of samples. In contrast, the stair chart is discontinuous and only needs to be evaluated at the edges of the interval

```
GraphData(FB, edges, verticies, num_samples=2, connect=False)
```

such that one advantage of this model is it helps differentiate graphics that have different artists from graphics that have the same artist but make different assumptions about the source data.

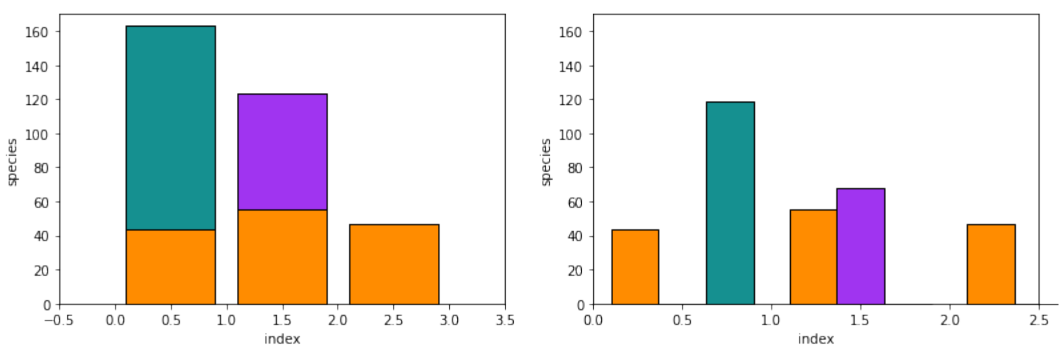
4.4 Case Study: Penguins

Building on the `Bar` artist in subsection 4.1, we implement grouped bar charts as these do not exist out of the box in the current version of Matplotlib.Instead, grouped bar charts are often achived via looping over an implementation of bar, which forces the user to keep track which values are mapped to a single visual element and how that is achieved. For this case study, we use the Palmer Penguins dataset[91, 92], packaged as a pandas dataframe[93] since that is a very commonly used Python labeled data structure. The wrapper is very thin because there is explicitly only one section.

```
class DataFrame:
    def __init__(self, dataframe):
        self.FB = FiberBundle(K = {'tables':['vertex']},
                              F = dict(dataframe.dtypes))
```

```python
        self._tau = dataframe.iloc
        self._view = dataframe

    def view(self, axes=None):
        return self._view
```

Since the aim for this wrapper is to be very generic, here the fiber is set by querying the dataframe for its metadata. The `dtypes` are a list of column names and the datatype of the values in each column; this is the minimal amount of information the model requires to verify constraints. The pandas indexer is a key valued set of discrete vertices, so there is no need to repackage for the data interface.



(a) (b)

Figure 27: Penguin count disaggregated by island and species

The stacked and grouped bar charts in figure 27 are both composites of `Bar` artists such that the difference between `StackedBar` and `GroupedBar` is specific to the ways in which the `Bar` are stitched together. These two artists have identical constructors and `draw` methods. As with `Bar`, the orientation is set in the constructor. In both these artists, we separate the transforms that are applied to only one component (column) from transforms applied to multiple components (columns). We encode the one component case as `V` dictionary, and the multiple components case as `MV`. This convention allows us to, for example, map one column to position, but multiple to length.

```python
class StackedBar(martist.Artist):
    def __init__(self, E, V, MV, orientation='v', *args, **kwargs):
        """
        Parameters
        ----------

        orientation: str, optional
            vertical: bars aligned along x axis, heights on y
            horizontal: bars aligned along y axis, heights on x
        """
        super().__init__(*args, **kwargs)
        self.E = E
        self.orientation = orientation
        self.V = V
        self.MV = MV

    def assemble(self):
        tau = self.data.view(self.axes)
        self.children = [] # list of bars to be rendered
        floor = 0
        for group in self.MV:
            # pull out the specific group transforms
            bar = Bar(self.E, {**group, **self.V, 'floor':floor},
                      self.orientation, transform=self.axes.transData)
            self.children.append(bar)
            floor += view[group['length']['name']]


```

```
    def draw(self, renderer, *args, **kwargs):
        # all the visual conversion gets pushed to child artists
        self.assemble()
        #self._transform = self.children[0].get_transform()
        for artist in self.children:
            artist.draw(renderer, *args, **kwargs)
```

Since all the visual transformation is passed through to `Bar`, the `draw` method does not do any visual transformations. In `StackedBar` the `view` is used to adjust the `floor` for every subsequent bar chart, since a stacked bar chart is bar chart area marks concatenated together in the length parameter. In contrast, `GroupedBar` does not even need the view, but instead keeps track of the relative position of each group of bars in the visual only variable `offset`.

```
class GroupedBar(martist.Artist):
    def assemble(self):
        self.children = [] # list of bars to be rendered
        ngroups = len(self.mtransforms)

        for gid, group in enumerate(self.mtransforms):
            group.update(self.transforms)
            width = group.get('width', .8)
            gwidth = width/ngroups
            offset = gid/ngroups*width
            bar = Bar(self.E, {**group, **self.V, 'width':gwidth, 'offset':offset},
                      self.orientation, transform=self.axes.transData)
            self.children.append(bar)
```

Since the only difference between these two glyphs is in the composition of `Bar`, they take in the exact same transform specification dictionaries. The `transform` dictionary dictates the position of the group, in this case by island the penguins are found on.

```
transforms = {'position': {'name':'island',
               'encoder': position.Nominal({'Biscoe':0.1, 'Dream':1.1, 'Torgersen':2.1})}}
group_transforms =  [{'length': {'name':'Adelie'},
                       'facecolors': {'name':"Adelie_s", 'encoder':cmap}},
                      {'length': {'name':'Chinstrap'},
                       'facecolors': {'name':"Chinstrap_s", 'encoder':cmap}},
                      {'length': {'name':'Gentoo'},
                       'facecolors': {'name':"Gentoo_s", 'encoder':cmap}}]
```

`group_transforms` describes the group, and takes a list of dictionaries where each dictionary is the aesthetics of each group. That `position` and `length` are required parameters is enforced in the creation of the `Bar` artist. These means that these two artists have identical function signatures

```
artistSB = bar.StackedBar(bt, ts, group_transforms)
artistGB = bar.GroupedBar(bt, ts, group_transforms)
```

but differ in assembly $\hat{Q}$. By decomposing the architecture into data, visual encoding, and assembly steps, we are able to build components that are more flexible and also more self contained than the existing code base. While very rough, this API demonstrates that the ideas presented in the math framework are implementable. For example, the `draw` function that maps most closely to A is functional, with state only being necessary for bookkeeping the many inputs that the function requires. In choosing a functional approach, if not implementation, we provide a framework for library developers to build reusable encoder

ν assembly $\hat{Q}$and artists A. We argue that if these functions are built such that they are equivariant with respect to monoid actions and the graphic topology is a deformation retraction of the data topology, then the artist by definition will be a structure and property preserving map from data to graphic.

5 Discussion

This work contributes a functional model of the structure-preserving maps from data to visual representation to guide the development of visualization libraries, thereby providing a means to express the constraints that visualization must preserve continuity and faithfully translate the properties of the data variables into visual variables. Combining Butler's proposal of a fiber bundle model of visualization data with Spivak's formalism of schema lets this model support a variety of datasets, including discrete relational tables, multivariate high resolution spatio temporal datasets, and complex networks. By decomposing the artist into encoding ν, assembly Q, and reindexing ξ maps, the model expresses the specifications that graphic and data must have equivalent *continuity* equivalent to the data, and that the visual characteristics of the graphics are *equivariant* to their corresponding components under monoid actions. TEAM defines these constraints on the transformation function such that they are not specific to any one type of encoding or visual characteristic. The toy prototype built using this model validates that it is usable for a general purpose visualization tool since it can be iteratively integrated into existing architecture rather than starting from scratch. Factoring out graphic formation into assembly functions allows for much more clarity in how they differ. This prototype demonstrates that this framework can generate the fundamental point (scatter plot) and line (line chart) marks.

5.1 Limitations

Our model and prototype are deeply tied to Matplotlib's existing architecture, so it has not yet been worked through how the model generalizes to libraries such as VTK or D3. Even though the model is designed to be backend and format independent, it has only been tested using PNGs rendered with AGG[76]. It is unknown how this framework interfaces with high performance rendering libraries such as openGL[74] that implement different models of ρ. While our model supports equivariance of figurative glyphs [8] generated from data components[94, 95], it cannot evaluate the semantic accuracy of the figurative representation. Effectiveness criteria[11, 96] are out of scope.

5.2 Future Work

More work is needed to formalize the composition operators, equivalence class A', and the mathematical model of interactivity. We also need to implement artists that demonstrate that the model can underpin a minimally viable library, foremost an image[97, 98], a heatmap[99, 100], and an inherently computational artist such as a boxplot[101]. In summary, the proposed scope of work is

work period	**milestones & tasks**
April - July 2021	prepare and submit **conference presentation** on new functionality enabled by model for *SciPy*: artists that do not inherit from existing Matplotlib artists, computational artists such as histograms, non tabular data, composite interactive artist
June - Sept 2021	prepare and submit **theory paper** on interactivity to *TCVG or Eurovis 2022*: fully work out and describe math of addition operators and lookups from graphic to data space, implement brush linked artist (shared base space) and artist that exploits sheafs
May - Nov 2021	prepare and submit **applications paper** on high dimensional to *TCVG*: math and implementation of computational artists, concurrent artists and data sources, non-trivial data bundles
Aug 2021 - Feb 2022	prepare and submit **systems paper** on building domain specific libraries based on this model to *Infoviz 2022*: domain specific structured data, composite artists, inference of meta data components, mathematical notion of a visualization (labeled, multiple artists, etc)
March 2022	**dissertation** writing: synthesize previous work on climate data, compile topological equivariant artist model work
April 2022	**defense**

Table 3

In acknowledgement that the schedule is optimistic, the data applications of this work could instead be revisited after the defense as that work is not integral to visualization library architecture. The data applications could be further integrated with topological[102] and functional [103] data analysis methods. Since this model formalizes notions of structure preservation, it can serve as a good base for tools that assess quality metrics[104] or invariance [12] of visualizations with respect to graphical encoding choices. While this paper formulates visualization in terms of monoidal action homomorphisms between fiberbundles, the model lends itself to a categorical formulation[54, 105] that could be further explored.

6 Conclusion

A TEAM driven refactor of visualizations libraries could produce more maintainable, reusable, and extensible code, leading to better building blocks for the ecosystem of tools built on top of TEAM architectured libraries. Building block libraries could better support downstream, including domain specific, libraries without having to explicitly incorporate the specific data structure and visualization needs of those domains back into the base library. Adopting this model would induce a separation of data representation and visual representation that, for example, in Matplotlib is so entangled that it has lead to a brittle and sometimes incoherent API and internal code base. A refactor that incorporated the generalized data model and functional transforms presented in TEAM would lead to building block libraries that provide a more consistent, reusable, flexible, collection of blocks.

References

[1] Krist Wongsuphasawat. *Navigating the Wide World of Data Visualization Libraries (on the Web)*. 2021.

[2] J. Hughes. "Why Functional Programming Matters". In: *The Computer Journal* 32.2 (Jan. 1989), pp. 98–107. ISSN: 0010-4620. DOI: 10.1093/comjnl/32.2.98.

[3] Zhenjiang Hu, John Hughes, and Meng Wang. “How Functional Programming Mattered”. In: *National Science Review* 2.3 (Sept. 2015), pp. 349–370. ISSN: 2095-5138. DOI: 10.1093/nsr/nwv042.

[4] J. D. Hunter. “Matplotlib: A 2D Graphics Environment”. In: *Computing in Science Engineering* 9.3 (May 2007), pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[5] John Hunter and Michael Droettboom. *The Architecture of Open Source Applications (Volume 2): Matplotlib*. https://www.aosabook.org/en/matplotlib.html.

[6] A. Sarikaya et al. “What Do We Talk About When We Talk About Dashboards?” In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (Jan. 2019), pp. 682–692. ISSN: 1941-0506. DOI: 10.1109/TVCG.2018.2864903.

[7] Michael Friendly. “A Brief History of Data Visualization”. en. In: *Handbook of Data Visualization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 15–56. ISBN: 978-3-540-33036-3 978-3-540-33037-0. DOI: 10.1007/978-3-540-33037-0_2.

[8] L. Byrne, D. Angus, and J. Wiles. “Acquired Codes of Meaning in Data Visualization and Infographics: Beyond Perceptual Primitives”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 509–518. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467321.

[9] Jacques Bertin. *Semiology of Graphics : Diagrams, Networks, Maps*. English. Redlands, Calif.: ESRI Press, 2011. ISBN: 978-1-58948-261-6 1-58948-261-1.

[10] Jock Mackinlay. “Automating the Design of Graphical Presentations of Relational Information”. In: *ACM Transactions on Graphics* 5.2 (Apr. 1986), pp. 110–141. ISSN: 0730-0301. DOI: 10.1145/22949.22950.

[11] Jock Mackinlay. “Automatic Design of Graphical Presentations”. English. PhD Thesis. Stanford, 1987.

[12] G. Kindlmann and C. Scheidegger. “An Algebraic Process for Visualization Design”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (Dec. 2014), pp. 2181–2190. ISSN: 1941-0506. DOI: 10.1109/TVCG.2014.2346325.

[13] Ricky Shadrach. *Introduction to Groups*. https://www.mathsisfun.com/sets/groups-introduction.html. 2017.

[14] *Naturalness Principle - InfoVis:Wiki*. https://infovis-wiki.net/wiki/Naturalness_Principle.

[15] Edward R. Tufte. *The Visual Display of Quantitative Information*. English. Cheshire, Conn.: Graphics Press, 2001. ISBN: 0-9613921-4-2 978-0-9613921-4-7 978-1-930824-13-3 1-930824-13-0.

[16] J. Heer and M. Agrawala. "Software Design Patterns for Information Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 12.5 (2006), pp. 853–860. DOI: 10.1109/TVCG.2006.178.

[17] E. H. Chi. "A Taxonomy of Visualization Techniques Using the Data State Reference Model". In: *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*. Oct. 2000, pp. 69–75. DOI: 10.1109/INFVIS.2000.885092.

[18] Jeffrey Heer and Michael Bostock. "Declarative Language Design for Interactive Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (Nov. 2010), pp. 1149–1156. ISSN: 1077-2626. DOI: 10.1109/TVCG.2010.144.

[19] C. Stolte, D. Tang, and P. Hanrahan. "Polaris: A System for Query, Analysis, and Visualization of Multidimensional Relational Databases". In: *IEEE Transactions on Visualization and Computer Graphics* 8.1 (Jan. 2002), pp. 52–65. ISSN: 1941-0506. DOI: 10.1109/2945.981851.

[20] Pat Hanrahan. "VizQL: A Language for Query, Analysis and Visualization". In: *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*. SIGMOD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 721. ISBN: 1-59593-434-0. DOI: 10.1145/1142473.1142560.

[21] J. Mackinlay, P. Hanrahan, and C. Stolte. "Show Me: Automatic Presentation for Visual Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (Nov. 2007), pp. 1137–1144. ISSN: 1941-0506. DOI: 10.1109/TVCG.2007.70594.

[22] Leland Wilkinson. *The Grammar of Graphics*. en. 2nd ed. Statistics and Computing. New York: Springer-Verlag New York, Inc., 2005. ISBN: 978-0-387-24544-7.

[23] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4.

[24] M. Bostock and J. Heer. "Protovis: A Graphical Toolkit for Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 1121–1128. ISSN: 1941-0506. DOI: 10.1109/TVCG.2009.174.

[25] Arvind Satyanarayan, Kanit Wongsuphasawat, and Jeffrey Heer. "Declarative Interaction Design for Data Visualization". en. In: *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*. Honolulu Hawaii USA: ACM, Oct. 2014, pp. 669–678. ISBN: 978-1-4503-3069-5. DOI: 10.1145/2642918.2647360.

[26] Jacob VanderPlas et al. "Altair: Interactive Statistical Visualizations for Python". en. In: *Journal of Open Source Software* 3.32 (Dec. 2018), p. 1057. ISSN: 2475-9066. DOI: 10.21105/joss.01057.

[27] Jeffrey D. Ullman and Jennifer. Widom. *A First Course in Database Systems*. English. Upper Saddle River, NJ: Pearson Prentice Hall, 2008. ISBN: 0-13-600637-X 978-0-13-600637-4.

[28] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. "NIH Image to ImageJ: 25 Years of Image Analysis". In: *Nature Methods* 9.7 (July 2012), pp. 671–675. ISSN: 1548-7105. DOI: 10.1038/nmeth.2089.

[29] Nicholas Sofroniew et al. *Napari/Napari: 0.4.5rc1*. Zenodo. Feb. 2021. DOI: 10.5281/zenodo.4533308.

[30] *Writing Plugins*. en. https://imagej.net/Writing_plugins.

[31] Software Studies. *Culturevis/Imageplot*. Jan. 2021.

[32] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. "Gephi: An Open Source Software for Exploring and Manipulating Networks". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 3.1 (Mar. 2009). ISSN: 2334-0770.

[33] John Ellson et al. “Graphviz— Open Source Graph Drawing Tools”. In: *Graph Drawing*. Ed. by Petra Mutzel, Michael Jünger, and Sebastian Leipert. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 483–484. ISBN: 978-3-540-45848-7.

[34] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. “Exploring Network Structure, Dynamics, and Function Using NetworkX”. In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15.

[35] *Data Representation in Mayavi — Mayavi 4.7.2 Documentation*. https://docs.enthought.com/mayavi/mayavi/d

[36] M. Bostock, V. Ogievetsky, and J. Heer. “D^3 Data-Driven Documents”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (Dec. 2011), pp. 2301–2309. ISSN: 1941-0506. DOI: 10.1109/TVCG.2011.185.

[37] Marcus D. Hanwell et al. “The Visualization Toolkit (VTK): Rewriting the Rendering Code for Modern Graphics Cards”. en. In: *SoftwareX* 1-2 (Sept. 2015), pp. 9–12. ISSN: 23527110. DOI: 10.1016/j.softx.2015.04.001.

[38] P. Ramachandran and G. Varoquaux. “Mayavi: 3D Visualization of Scientific Data”. In: *Computing in Science Engineering* 13.2 (Mar. 2011), pp. 40–51. ISSN: 1558-366X. DOI: 10.1109/MCSE.2011.35.

[39] James Ahrens, Berk Geveci, and Charles Law. “Paraview: An End-User Tool for Large Data Visualization”. In: *The visualization handbook* 717.8 (2005).

[40] Brian Wylie and Jeffrey Baumes. “A Unified Toolkit for Information and Scientific Visualization”. In: *Proc.SPIE*. Vol. 7243. Jan. 2009. DOI: 10.1117/12.805589.

[41] Stephen Few and Perceptual Edge. “Dashboard Confusion Revisited”. In: *Perceptual Edge* (2007), pp. 1–6.

[42] D. M. Butler and M. H. Pendley. “A Visualization Model Based on the Mathematics of Fiber Bundles”. en. In: *Computers in Physics* 3.5 (1989), p. 45. ISSN: 08941866. DOI: 10.1063/1.168345.

[43] David M. Butler and Steve Bryson. “Vector-Bundle Classes Form Powerful Tool for Scientific Visualization”. en. In: *Computers in Physics* 6.6 (1992), p. 576. ISSN: 08941866. DOI: 10.1063/1.4823118.

[44] David I Spivak. *Databases Are Categories*. en. Slides. June 2010.

[45] David I Spivak. “SIMPLICIAL DATABASES”. en. In: (), p. 35.

[46] Tamara Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC press, Oct. 2014. ISBN: 978-1-4665-0891-0.

[47] Tamara Munzner. “Ch 2: Data Abstraction”. In: *CPSC547: Information Visualization, Fall 2015-2016* ().

[48] E.H. Spanier. *Algebraic Topology*. McGraw-Hill Series in Higher Mathematics. Springer, 1989. ISBN: 978-0-387-94426-5.

[49] *Locally Trivial Fibre Bundle - Encyclopedia of Mathematics*. https://encyclopediaofmath.org/wiki/Locally_trivia

[50] “Monoid”. en. In: *Wikipedia* (Jan. 2021).

[51] “Semigroup Action”. en. In: *Wikipedia* (Jan. 2021).

[52] nLab authors. “Action”. In: (Mar. 2021).

[53] Eric W. Weisstein. *Group Homomorphism*. en. https://mathworld.wolfram.com/GroupHomomorphism.html. Text.

[54] Brendan Fong and David I. Spivak. *An Invitation to Applied Category Theory: Seven Sketches in Compositionality*. en. First. Cambridge University Press, July 2019. ISBN: 978-1-108-66880-4 978-1-108-48229-5 978-1-108-71182-1. DOI: 10.1017/9781108668804.

[55] S. S. Stevens. “On the Theory of Scales of Measurement”. In: *Science* 103.2684 (1946), pp. 677–680. ISSN: 00368075, 10959203.

[56] W A Lea. “A Formalization of Measurement Scale Forms”. en. In: (), p. 44.

[57] Brent A Yorgey. “Monoids: Theme and Variations (Functional Pearl)”. en. In: (), p. 12.

[58] “Quotient Space (Topology)”. en. In: *Wikipedia* (Nov. 2020).

[59] Professor Denis Auroux. “Math 131: Introduction to Topology”. en. In: (), p. 113.

[60] Robert W. Ghrist. *Elementary Applied Topology*. Vol. 1. Createspace Seattle, 2014.

[61] Robert Ghrist. “Homological Algebra and Data”. In: *Math. Data* 25 (2018), p. 273.

[62] David Urbanik. “A Brief Introduction to Schemes and Sheaves”. en. In: (), p. 16.

[63] Dmitry Nekrasovski et al. “An Evaluation of Pan & Zoom and Rubber Sheet Navigation with and without an Overview”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 11–20. ISBN: 1-59593-372-7. DOI: 10.1145/1124772.1124775.

[64] Michael S. Crouch, Andrew McGregor, and Daniel Stubbs. “Dynamic Graphs in the Sliding-Window Model”. In: *European Symposium on Algorithms*. Springer, 2013, pp. 337–348.

[65] Chia-Shang James Chu. “Time Series Segmentation: A Sliding Window Approach”. In: *Information Sciences* 85.1 (July 1995), pp. 147–173. ISSN: 0020-0255. DOI: 10.1016/0020-0255(95)00021-G.

[66] Charles R Harris et al. “Array Programming with NumPy”. In: *Nature* 585.7825 (2020), pp. 357–362.

[67] Jeff Reback et al. *Pandas-Dev/Pandas: Pandas 1.0.3*. Zenodo. Mar. 2020. DOI: 10.5281/zenodo.3715232.

[68] Stephan Hoyer and Joe Hamman. “Xarray: ND Labeled Arrays and Datasets in Python”. In: *Journal of Open Research Software* 5.1 (2017).

[69] Matthew Rocklin. “Dask: Parallel Computation with Blocked Algorithms and Task Scheduling”. In: *Proceedings of the 14th Python in Science Conference*. Vol. 126. Citeseer, 2015.

[70] “Retraction (Topology)”. en. In: *Wikipedia* (July 2020).

[71] Eric W. Weisstein. *Homotopy*. en. https://mathworld.wolfram.com/Homotopy.html. Text.

[72] Tim Bienz, Richard Cohn, and Calif.) Adobe Systems (Mountain View. *Portable Document Format Reference Manual*. Citeseer, 1993.

[73] A. Quint. "Scalable Vector Graphics". In: *IEEE MultiMedia* 10.3 (July 2003), pp. 99–102. ISSN: 1941-0166. DOI: 10.1109/MMUL.2003.1218261.

[74] George S. Carson. "Standards Pipeline: The OpenGL Specification". In: *SIGGRAPH Comput. Graph.* 31.2 (May 1997), pp. 17–18. ISSN: 0097-8930. DOI: 10.1145/271283.271292.

[75] *Cairographics.Org*. https://www.cairographics.org/.

[76] Maxim Shemanarev. *Anti-Grain Geometry*. https://antigrain.com/.

[77] A. M. Cegarra. "Cohomology of Monoids with Operators". In: *Semigroup Forum*. Vol. 99. Springer, 2019, pp. 67–105. ISBN: 1432-2137.

[78] Christian Remling. *Algebra (Math 5353/5363) Lecture Notes*. Lecture Notes. University of Oklahoma.

[79] H. M. Johnson. "Pseudo-Mathematics in the Mental and Social Sciences". In: *The American Journal of Psychology* 48.2 (1936), pp. 342–351. ISSN: 00029556. DOI: 10.2307/1415754.

[80] M. A. Thomas. *Mathematization, Not Measurement: A Critique of Stevens' Scales of Measurement*. en. SSRN Scholarly Paper ID 2412765. Rochester, NY: Social Science Research Network, Oct. 2014. DOI: 10.2139/ssrn.2412765.

[81] "Connected Space". en. In: *Wikipedia* (Dec. 2020).

[82] Caroline Ziemkiewicz and Robert Kosara. "Embedding Information Visualization within Visual Representation". In: *Advances in Information and Intelligent Systems*. Ed. by Zbigniew W. Ras and William Ribarsky. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 307–326. ISBN: 978-3-642-04141-9. DOI: 10.1007/978-3-642-04141-9_15.

[83] Sheelagh Carpendale. *Visual Representation from Semiology of Graphics by J. Bertin.* en.

[84] "Jet Bundle". en. In: *Wikipedia* (Dec. 2020).

[85] Jana Musilová and Stanislav Hronek. "The Calculus of Variations on Jet Bundles as a Universal Approach for a Variational Formulation of Fundamental Physical Theories". In: *Communications in Mathematics* 24.2 (Dec. 2016), pp. 173–193. ISSN: 2336-1298. DOI: 10.1515/cm-2016-0012.

[86] Yael Albo et al. "Off the Radar: Comparative Evaluation of Radial Visualization Solutions for Composite Indicators". In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 569–578. ISSN: 1077-2626. DOI: 10.1109/TVCG.2015.2467322.

[87] Z. Qu and J. Hullman. "Keeping Multiple Views Consistent: Constraints, Validations, and Exceptions in Visualization Authoring". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (Jan. 2018), pp. 468–477. ISSN: 1941-0506. DOI: 10.1109/TVCG.2017.2744198.

[88] Richard A. Becker and William S. Cleveland. "Brushing Scatterplots". In: *Technometrics* 29.2 (May 1987), pp. 127–142. ISSN: 0040-1706. DOI: 10.1080/00401706.1987.10488204.

[89] Andreas Buja et al. "Interactive Data Visualization Using Focusing and Linking". In: *Proceedings of the 2nd Conference on Visualization '91*. VIS '91. Washington, DC, USA: IEEE Computer Society Press, 1991, pp. 156–163. ISBN: 0-8186-2245-8.

[90] *Dataclasses — Data Classes — Python 3.9.2rc1 Documentation.* https://docs.python.org/3/library/dataclasses.

[91] Kristen B. Gorman, Tony D. Williams, and William R. Fraser. "Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis)". In: *PLOS ONE* 9.3 (Mar. 2014), e90081. DOI: 10.1371/journal.pone.0090081.

[92] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. Manual. 2020. DOI: 10.5281/zenodo.3960218.

[93] Muhammad Chenariyan Nakhaee. *Mcnakhaee/Palmerpenguins*. Jan. 2021.

[94] F. Beck. "Software Feathers Figurative Visualization of Software Metrics". In: *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*. Jan. 2014, pp. 5–16.

[95] Lydia Byrne, Daniel Angus, and Janet Wiles. "Figurative Frames: A Critical Vocabulary for Images in Information Visualization". In: *Information Visualization* 18.1 (Aug. 2017), pp. 45–67. ISSN: 1473-8716. DOI: 10.1177/1473871617724212.

[96] John M Chambers et al. *Graphical Methods for Data Analysis*. Vol. 5. Wadsworth Belmont, CA, 1983.

[97] Robert B Haber and David A McNabb. "Visualization Idioms: A Conceptual Model for Scientific Visualization Systems". In: *Visualization in scientific computing* 74 (1990), p. 93.

[98] Charles D Hansen and Chris R Johnson. *Visualization Handbook*. Elsevier, 2011.

[99] Leland Wilkinson and Michael Friendly. "The History of the Cluster Heat Map". In: *The American Statistician* 63.2 (May 2009), pp. 179–184. ISSN: 0003-1305. DOI: 10.1198/tas.2009.0033.

[100] Toussaint Loua. *Atlas Statistique de La Population de Paris*. J. Dejey & cie, 1873.

[101] Hadley Wickham and Lisa Stryjewski. "40 Years of Boxplots". In: *The American Statistician* (2011).

[102] C. Heine et al. "A Survey of Topology-Based Methods in Visualization". In: *Computer Graphics Forum* 35.3 (June 2016), pp. 643–667. ISSN: 0167-7055. DOI: 10.1111/cgf.12933.

[103] James O Ramsay. *Functional Data Analysis*. Wiley Online Library, 2006.

[104] Enrico Bertini, Andrada Tatu, and Daniel Keim. "Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2203–2212.

[105] Bartosz Milewski. "Category Theory for Programmers". en. In: (), p. 498.