

# <sup>1</sup> 1 Topological Artist Model

As discussed in the introduction, visualization is generally defined as structure preserving maps from data to graphic representation. In order to formalize this statement, we describe the connectivity of the records using topology and define the structure on the components in terms of the monoid actions on the component types. By formalizing structure in this way, we can evaluate the extent to which a visualization preserves the structure of the data it is representing and build structure preserving visualization tools. We introduce the notion of an artist  $\mathcal{A}$  as an equivariant map from data to graphic

$$\mathcal{A} : \mathcal{E} \rightarrow \mathcal{H} \tag{1}$$

that carries a homomorphism of monoid actions  $\varphi : M \rightarrow M'$  [14], which are discussed in detail in section 1.1.2. Given  $M$  on data  $\mathcal{E}$  and  $M'$  on graphic  $\mathcal{H}$ , we propose that artists  $\mathcal{A}$  are symmetric

$$\mathcal{A}(m \cdot r) = \varphi(m) \cdot \mathcal{A}(r) \tag{2}$$

- <sup>2</sup> such that applying a monoid  $m \in M$  transformation to the input  $r \in \mathcal{E}$  to  $\mathcal{A}$  is equivalent  
<sup>3</sup> to applying a monoid  $\varphi(M) \in M'$  to the output of the artist  $\mathcal{A}(r) \in \mathcal{H}$ .

<sup>4</sup> We model the data  $\mathcal{E}$ , graphic  $\mathcal{H}$ , and intermediate visual encoding  $\mathcal{V}$  stages of visualiza-  
<sup>5</sup> tion as topological structures that encapsulate types of variables and continuity; by doing  
<sup>6</sup> so we can develop implementations that keep track of both in ways that let us distribute  
<sup>7</sup> computation while still allowing assembly and dynamic update of the graphic. To explain  
<sup>8</sup> which structure the artist is preserving, we first describe how we model data (1.1), graphics  
<sup>9</sup> (1.2), and intermediate visual characteristics (1.3) as fiber bundles. We then discuss the  
<sup>10</sup> equivariant maps between data and visual characteristics (1.3.2) and visual characteristics  
<sup>11</sup> and graphics (1.3.3) that make up the artist.

## <sup>12</sup> 1.1 Data Space $E$

Building on Butler's proposal of using fiber bundles as a common data representation format for visualization data[9, 10], a fiber bundle is a tuple  $(E, K, \pi, F)$  defined by the projection map  $\pi$

$$F \hookrightarrow E \xrightarrow{\pi} K \tag{3}$$

<sup>13</sup> that binds the components of the data in  $F$  to the continuity represented in  $K$ . The fiber  
<sup>14</sup> bundle models the properties of data component types  $F$  (1.1.1), the continuity of records  
<sup>15</sup>  $K$  (1.1.3), the collections of records  $\tau$  (1.1.4), and the space  $E$  of all possible datasets with  
<sup>16</sup> these components and continuity.

<sup>17</sup> By definition fiber bundles are locally trivial[28, 39], meaning that over a localized neigh-  
<sup>18</sup> borhood we can dispense with extra structure on  $E$  and focus on the components and conti-  
<sup>19</sup> nuity. We use fiber bundles as the data model because they are inclusive enough to express  
<sup>20</sup> all the types of data described in section ??.

### <sup>21</sup> 1.1.1 Variables: Fiber Space $F$

To formalize the structure of the data components, we use notation introduced by Spivak [40] that binds the components of the fiber to variable names and types. Spivak constructs a set  $\mathbf{U}$  that is the disjoint union of all possible objects of types  $\{T_0, \dots, T_m\} \in \mathbf{DT}$ , where

$\mathbf{DT}$  are the data types of the variables in the dataset. He then defines the single variable set  $\mathbb{U}_\sigma$

$$\begin{array}{ccc} \mathbb{U}_\sigma & \longrightarrow & \mathbb{U} \\ \pi_\sigma \downarrow & & \downarrow \pi \\ C & \xrightarrow[\sigma]{} & \mathbf{DT} \end{array} \quad (4)$$

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  $\sigma$ , the fiber for a one variable dataset is

$$F = \mathbb{U}_{\sigma(c)} = \mathbb{U}_T \quad (5)$$

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

$$F = \mathbb{U}_{\sigma(c_1)} \times \dots \mathbb{U}_{\sigma(c_i)} \dots \times \mathbb{U}_{\sigma(c_n)} \quad (6)$$

which is equivalent to

$$F = F_0 \times \dots \times F_i \times \dots \times F_n \quad (7)$$

<sup>22</sup> which allows us to decouple  $F$  into components  $F_i$ .

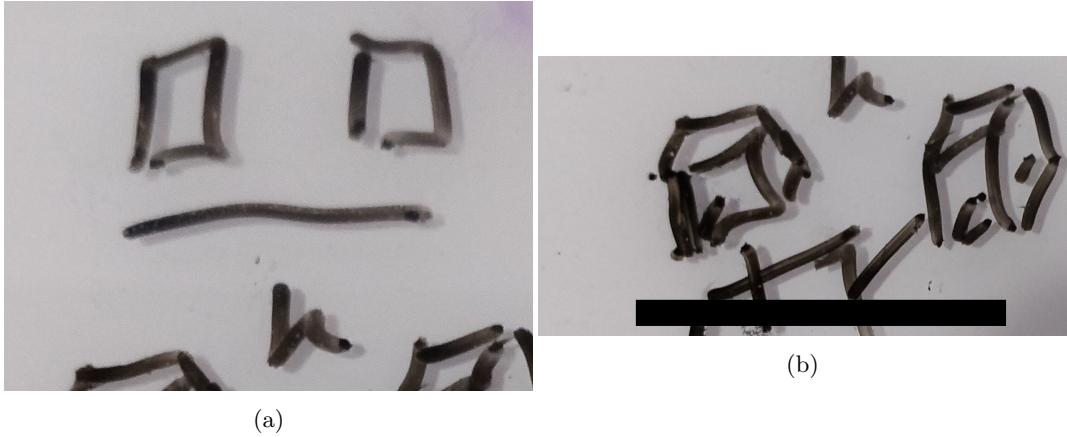


Figure 1: These two datasets have the same base space  $K$  but figure 1a has fiber  $F = \mathbb{R} \times \mathbb{R}$  which is (time, temperature) while figure 1b has fiber  $\mathbb{R}^+ \times \mathbb{R}^2$  which is (time, wind=(speed, direction))

For example, the data in figure 1a is a pair of times and  ${}^\circ K$  temperature measurements taken at those times. Time is a  $\mathbb{R}$ . Temperature values are real positive numbers  $\mathbb{U}_{\text{float}} = \mathbb{R}^+$ . The fiber is

$$\mathbb{U} = \mathbb{R} \times \mathbb{R}^+ \quad (8)$$

where the first component  $F_0$  is the set of values specified by ( $c = \text{time}$ ,  $T = \text{datetime}$ ,  $\mathbb{U}_\sigma = \mathbb{R}$ ) and  $F_1$  is specified by ( $c = \text{temperature}$ ,  $T = \text{float}$ ,  $\mathbb{U}_\sigma = \mathbb{R}^+$ ) and is the set of values  $\mathbb{U}_\sigma = \mathbb{R}^+$ . In figure 1b, temperature is replaced with wind. This wind variable is of type

`wind` and has two components speed and direction  $\{(s, d) \in \mathbb{R}^2 \mid 0 \leq s, 0 \leq d \leq 360\}$ . Therefore, the fiber is

$$F = \mathbb{R}^+ \times \mathbb{R}^2 \quad (9)$$

such that  $F_1$  is specified by ( $c = \text{wind}$ ,  $T = \text{wind}$ ,  $\mathbb{U}_\sigma = \mathbb{R}^2$ ). As illustrated in figure 1, Spivak's framework provides a consistent way to describe potentially complex components of the input data.

### 1.1.2 Measurement Scales: Monoid Actions

Implementing expressive visual encodings requires formally describing the structure on the components of the fiber, which we define by the action of a monoid on the component. While structure on a set of values is often described algebraically as operations or through the actions of a group, for example Steven's scales [41], we generalize to monoids to support more component types. Monoids are also commonly found in functional programming because they specify compositions of transformations [42, 45].

A monoid [29]  $M$  is a set with an associative binary operator  $* : M \times M \rightarrow M$ . A monoid has an identity element  $e \in M$  such that  $e * a = a * e = a$  for all  $a \in M$ . As defined on a component of  $F$ , a left monoid action [1, 38] of  $M_i$  is a set  $F_i$  with an action  $\bullet : M \times F_i \rightarrow F_i$  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  $x \in F_i$ ,  $e \in M_i$ ,  $e \bullet x = x$

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

$$M = M_0 \times \dots \times M_i \times \dots \times M_n \quad (10)$$

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

Steven's described the measurement scales[27, 41] in terms of the monoid actions on the measurements: nominal data is permutable, ordinal data is monotonic, interval data is translatable, and ratio data is scalable [44]. For example, given an arbitrary interval scale fiber component ( $c = \text{temperature}$ ,  $T = \text{float}$ ,  $\mathbb{U}_\sigma = \mathbb{R}$ ) with arbitrarily chosen monoid translation actions actions

- monoid operator addition  $* = +$
- monoid operations:  $f : x \mapsto x + 1$ ,  $g : x \mapsto x + 2$
- monoid action operator composition  $\bullet = \circ$

By structure preservation, we mean that monoid actions are composable. For the translation actions described above on the temperature fiber, this means that they satisfy the condition

$$\begin{array}{ccc} \mathbb{R} & & \\ \downarrow_{x+1^\circ} & \searrow^{(x+1^\circ)\circ(x+2^\circ)} & \\ \mathbb{R} & \xrightarrow{x+2^\circ} & \mathbb{R} \end{array} \quad (11)$$

where  $1^\circ$  and  $2^\circ$  are valid distances between two temperatures  $x$ . What this diagram means is that either the fiber could be shifted by 1 (vertical line) then by 2 (horizontal), or the

45 two shifts could be combined such that in this case the fiber is shifted by 3 (diagonal) and  
46 these two paths yield the same temperature.

47 While many component types will be one of the measurement scale types, we gen-  
48 eralize to monoids specifically for the case of partially ordered set. Given a set  $W =$   
49  $\{mist, drizzle, rain\}$ , then the map  $f : W \rightarrow W$  defined by

- 50 1.  $f(rain) = drizzle$ ,
- 51 2.  $f(drizzle) = mist$
- 52 3.  $f(mist) = mist$

53 is order preserving such that  $mist \leq drizzle \leq rain$  but has no inverse since  $drizzle$  and  
54  $mist$  go to the same value  $mist$ . Therefore order preserving maps do not form a group, and  
55 instead we generalize to monoids to support partial order component types. Defining the  
56 monoid actions on the components serves as the basis for identifying the invariance[26] that  
57 must be preserved in the visual representation of the component. We propose equivariance  
58 of monoid actions individually on the fiber to visual component maps and on the graphic  
59 as a whole.

60 **1.1.3 Continuity: Base Space  $K$**

61 The base space  $K$  is way to express how the records in  $E$  are connected to each other, for  
62 example if they are discrete points or if they lie in a 2D continuous surface. Connectivity  
63 type is assumed in the choice of visualization, for example a line plot implies 1D continuous  
64 data, but an explicit representation allows for verifying that the topology of the graphic  
65 representation is equivalent to the topology of the data.

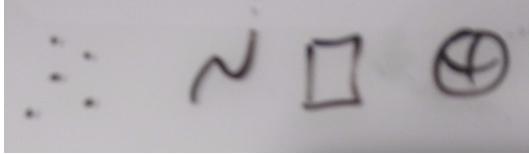


Figure 2: The topological base space  $K$  encodes the connectivity of the data space, for example if the data is independent points or a map or on a sphere

66 As illustrated in figure 2,  $K$  is akin to an indexing space into  $E$  that describes the  
67 structure of  $E$ .  $K$  can have any number of dimensions and can be continuous or discrete.

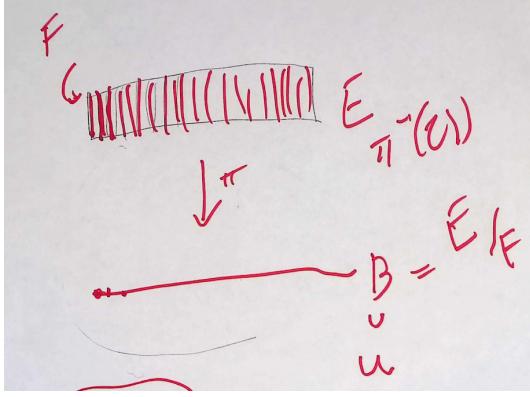


Figure 3: The base space  $E$  is divided into fiber segments  $F$ . The base space  $K$  acts as an index into the records in the fibers. **this figure might be good all the way up top to lay out the components of fb**

Formally  $K$  is the quotient space [34] of  $E$  meaning it is the finest space[4] such that every  $k \in K$  has a corresponding fiber  $F_k$ [34]. In figure 3,  $E$  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 fibers and monoids, we can decompose the total space into components  $\pi : E_i \rightarrow K$  where

$$\pi : E_1 \oplus \dots \oplus E_i \oplus \dots \oplus E_n \rightarrow K \quad (12)$$

which is a decomposition of  $F$ . The  $K$  remains the same because the connectivity of records does not change just because there are fewer elements in each record.

The datasets in figure 4 have the same fiber of (temperature, time). In figure 4a the fibers lie over discrete  $K$  such that the records in the datasets in the fiber bundles are discrete. The same fiber in figure 4b lies over a continuous interval  $K$  such that the records are samples from a continuous function defined on  $K$ . 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.

#### 1.1.4 Data: Sections $\tau$

The section  $\tau : K \rightarrow E$  is what ties together the base space  $K$  with the fiber  $F$ . A section is a function that takes as input location  $k \in K$  and returns a record  $r \in E$ . For example, in the special case of a table [40],  $K$  is a set of row ids,  $F$  is the columns, and the section  $\tau$  returns the record  $r$  at a given key in  $K$ . For any fiber bundle, there exists a map

$$\begin{array}{ccc} F & \xhookrightarrow{\quad} & E \\ \pi \downarrow \uparrow \tau & & \\ K & & \end{array} \quad (13)$$

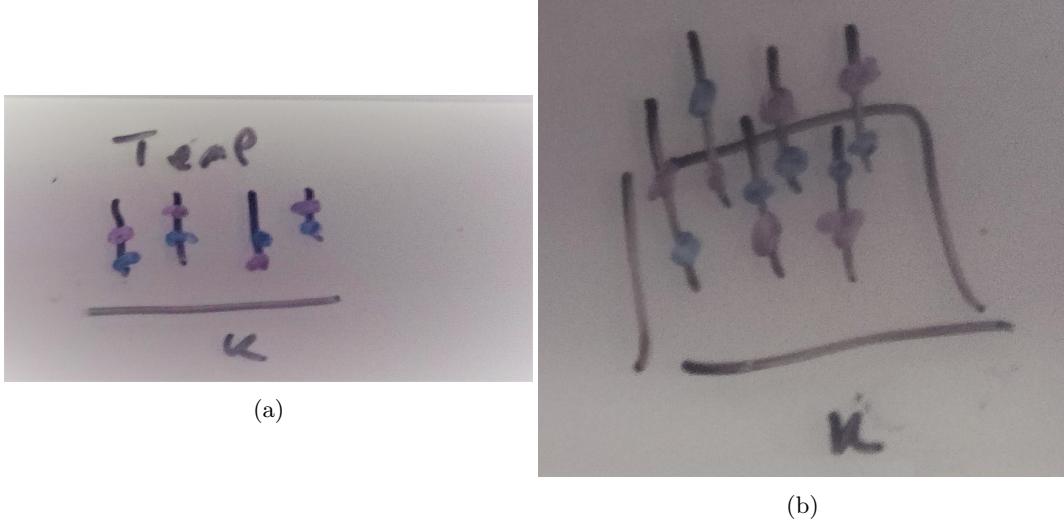


Figure 4: These two datasets have the same (time, temperature) fiber. In figure ?? the total space  $E$  is discrete over points  $k \in K$ , meaning the records in the fiber are also discrete. In figure ??  $E$  lies over the continuous interval  $K$ , meaning the records in the fiber are sampled from a continuous space. *revamp figure: F=Plane, k1 = dots, k2=line*

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

$$\tau(k) = (k, (g_{F_0}(k), \dots, g_{F_n}(k))) \quad (14)$$

where  $g : K \rightarrow F$  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, we can decompose  $\tau$  as

$$\tau = (\tau_0, \dots, \tau_i, \dots, \tau_n) \quad (15)$$

- 79 where each section  $\tau_i$  is a variable or set of variables. This allows for accessing the data  
80 component wise rather than just as sections on  $K$ .

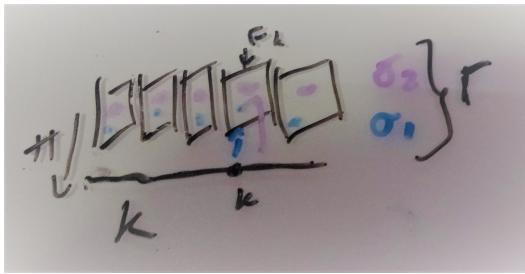


Figure 5: Fiber (time, temperature) with an interval  $K$  basespace. The sections  $\tau_i$  and  $\tau_j$  are constrained such that the time variable must be monotonic, which means each section is a timeseries of temperature values. They are included in the global set of sections  $\tau_1, \tau_2 \in \Gamma(E)$

In the example in figure 5, the fiber is  $(time, temperature)$  as described in figure 1 and the base space is the interval  $K$ . The section  $\tau^{(1)}$  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  $\tau^{(1)}, \tau^{(2)} \in \Gamma(E)$ .

This model provides a common interface to 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[21] 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. 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 fiber. For example, a pandas series which stores a labeled list, or a dataframe[35] 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 flatten 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 always assumed to be random keys.

Where this model particularly shines are  $N$  dimensional labeled data structures. For example, an xarray[22] 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 [37]. 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.

## 1.2 Graphic: $H$

We introduce a graphic bundle to hold the essential information necessary to render a graphical design constructed by the artist. As with the data, we can represent the target graphic as a section  $\rho$  of a bundle  $(H, S, \pi, D)$ . The graphic bundle  $H$  consists of a base  $S$ ( 1.2.1) that is a thickened form of  $K$  a fiber  $D$ ( 1.2.2) that is an idealized display space, and sections  $\rho$ ( 1.2.3) that encode a graphic where the visual characteristics are fully specified.

### 1.2.1 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

$$(x, y, r, g, b) \in D \quad (16)$$

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



Figure 6: The scatter and line graphic base spaces 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  $K$ . *add  $\alpha, \beta$  coordinates to figures*

120 target display. By abstracting the target display space as  $D$ , the model can support different  
121 targets, such as a 2D screen or 3D printer.

### 122 1.2.2 Continuity of the Graphic $S$

123 Just as the  $K$  encodes the connectivity of the records in the data, we propose an equivalent  
124  $S$  that encodes the connectivity of the rendered elements of the graphic. For some visualiza-  
125 tions,  $K$  may be lower dimension than  $S$ . For example, in a typical 2D display (ignoring  
126 depth), a point that is 0D in  $K$  cannot be represented on screen unless it is thickened to 2D  
127 to encode the connectivity of the pixels that visually represent the point. This thickening is  
128 often not necessary when the dimensionality of  $K$  matches the dimensionality of the target  
129 space, for example if  $K$  is 2D and the display is a 2D screen. We introduce  $S$  to thicken  $K$   
130 in a way which preserves the structure of  $K$ .

Formally, we require that  $K$  be a deformation retract[36] of  $S$  so that  $K$  and  $S$  have the same homotopy. The surjective map  $\xi : S \rightarrow K$

$$\begin{array}{ccc} E & & H \\ \pi \downarrow & & \pi \downarrow \\ K & \xleftarrow{\xi} & S \end{array} \quad (17)$$

131 goes from region  $s \in S_k$  to its associated point  $s$ . This means that if  $\xi(s) = k$ , the record at  
132  $k$  is copied over the region  $s$  such that  $\tau(k) = \xi^*\tau(s)$  where  $\xi^*\tau(s)$  is  $\tau$  pulled back over  $S$ .

133 When  $K$  is discrete points and the graphic is a scatter plot, each point  $k \in K$  corresponds  
134 to a 2D disk  $S_k$  as shown in figure 6. In the case of 1D continuous data and a line plot, the  
135 region  $\beta$  over a point  $\alpha_i$  specifies the thickness of the line in  $S$  for the corresponding  $\tau$  on  
136  $k$ . The image has the same dimensions in data space and graphic space such that no extra  
137 dimensions are needed in  $S$ .

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

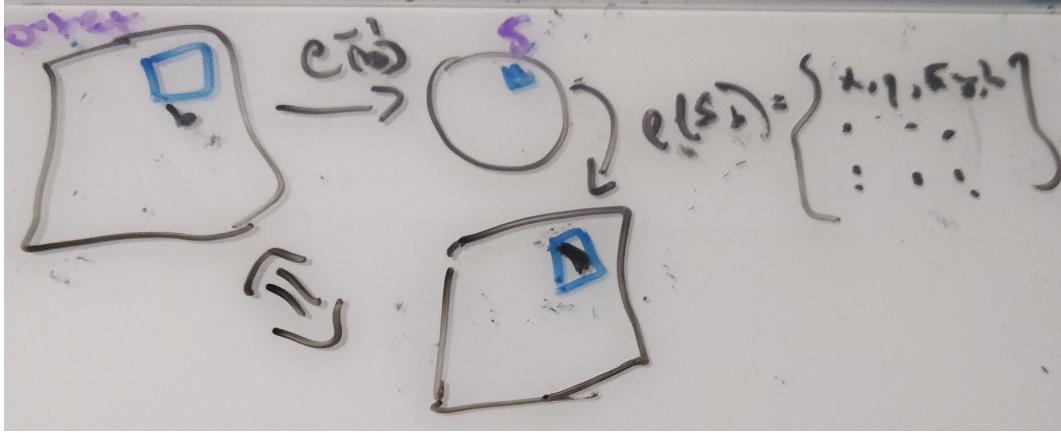


Figure 7: To render a graphic, a pixel  $p$  is selected in the display space, which is defined in the same coordinates as the  $x$  and  $y$  components in  $D$ . The inverse mapping  $\rho_{xy}^{-1}(p)$  returns a region  $S_p \subset S$ .  $\rho(S_p)$  returns the list of elements  $(x, y, r, g, b) \in D$  that lie over  $S_p$ . The integral over the  $(r, g, b)$  elements is the color of the pixel.

### 145 1.2.3 Rendering $\rho$

146 This section describes how we go from a graphic in an idealized prerender space to a rendered  
 147 image, where the graphic is the section  $\rho : S \rightarrow H$ . It is sufficient to sketch out how an  
 148 arbitrary pixel would be rendered, where a pixel  $p$  in a real display corresponds to a region  
 149  $S_p$  in the idealized display. To determine the color of the pixel, we aggregate the color values  
 150 over the region via integration.

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

$$r_p = \iint_{S_p} \rho_r(s) ds^2 \quad (18)$$

$$g_p = \iint_{S_p} \rho_g(s) ds^2 \quad (19)$$

$$b_p = \iint_{S_p} \rho_b(s) ds^2 \quad (20)$$

155 As shown in figure 7, a pixel  $p$  in the output space is selected and inverse mapped into  
 156 the corresponding region  $S_p \subset S$ . This triggers a lookup of the  $\rho$  over the region  $S_p$ , which  
 157 yields the set of elements in  $D$  that specify the  $(r, g, b)$  values corresponding to the region  
 158  $p$ . The color of the pixel is then obtained by taking the integral of  $\rho_{rgb}(S_p)$ .

159 In general,  $\rho$  is an abstraction of rendering. In very broad strokes  $\rho$  can be a specification  
 160 such as PDF[7], SVG[33], or an openGL scene graph[13]. Alternatively,  $\rho$  can be a rendering  
 161 engine such as cairo[11] or AGG[3]. Implementation of  $\rho$  is out of scope for this work,

<sub>162</sub> **1.3 Artist**

We propose that the transformation from data to visual representation can be described as a structure preserving map from one topological space to another. We name this map the artist as that is the analogous part of the Matplotlib[24] architecture that builds visual elements to pass off to the renderer. The topological artist  $A$  is a monoid equivariant sheaf map from the sheaf on a data bundle  $E$  which is  $\mathcal{O}(E)$  to the sheaf on the graphic bundle  $H$ ,  $\mathcal{O}(H)$ .

$$A : \mathcal{O}(E) \rightarrow \mathcal{O}(H) \quad (21)$$

<sub>163</sub> Sheafs are a mathematical object with restriction maps that define how to glue  $\tau$  over local  
<sub>164</sub> neighborhoods  $U \subseteq K$ , discussed in section ??, such that the  $A$  maps are consistent over  
<sub>165</sub> continuous regions of  $K$ . While Acan usually construct graphical elements solely with the  
<sub>166</sub> data in  $\tau$ , some visualizations, such as line, may also need some finite number  $n$  of derivatives,  
<sub>167</sub> which is captured by the jet bundle  $\mathcal{J}^n$  [25, 30] with  $\mathcal{J}^0(E) = E$ . In this work, we at most  
<sub>168</sub> need  $\mathcal{J}^2(E)$  which is the value at  $\tau$  and its first and second derivatives; therefore the artist  
<sub>169</sub> takes as input the data bundle padded with the jet bundle  $E' = E + \mathcal{J}^2(E)$ .

<sub>170</sub> Specifically,  $A$  is the equivariant map from  $E'$  to a specific graphic  $\rho \in \Gamma(H)$

$$\begin{array}{ccccc} E' & \xrightarrow{\nu} & V & \xleftarrow{\xi^*} & \xi^*V \xrightarrow{Q} H \\ & \searrow \pi & \downarrow \pi & \xi^* \pi \downarrow & \swarrow \pi \\ & & K & \xleftarrow{\xi} & S \end{array} \quad (22)$$

<sub>171</sub> where the input can be point wise  $\tau(k) \mid k \in K$ . The encoders  $\nu : E' \rightarrow V$  convert  
<sub>172</sub> the data components to visual components(1.2.2). The continuity map  $\xi : S \rightarrow K$  then  
<sub>173</sub> pulls back the visual bundle  $V$  over  $S$ (1.3.2). Then the assembly function  $Q : \xi^*V \rightarrow$   
<sub>174</sub>  $H$  composites the fiber components of  $\xi^*V$  into a graphic in  $H$ (1.3.3). This functional  
<sub>175</sub> decomposition of the visualization artist facilitates building reusable components at each  
<sub>176</sub> stage of the transformation because the equivariance constraints are defined on  $\nu$ ,  $Q$ , and  $\xi$ .

<sub>177</sub> **1.3.1 Visual Fiber Bundle  $V$**

<sub>178</sub> We introduce a visual bundle  $V$  to store the visual representations the artist needs to  
<sub>179</sub> assemble into a graphic. The visual bundle  $(V, K, \pi, P)$  has section  $\mu : V \rightarrow K$  that  
<sub>180</sub> resolves to a visual variable in the fiber  $P$ . The visual bundle  $V$  is the latent space of  
<sub>181</sub> possible parameters of a visualization type, such as a scatter or line plot. We define  $P$   
<sub>182</sub> in terms of the parameters of a visualization libraries compositing functions; for example  
<sub>183</sub> table 1 is a sample of the fiber space for Matplotlib [23].

$\nu_i$	$\mu_i$	$\text{codomain}(\nu_i) \subset P_i$
position	x, y, z, theta, r	$\mathbb{R}$
size	linewidth, markersize	$\mathbb{R}^+$
shape	markerstyle	$\{f_0, \dots, f_n\}$
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

184     A section  $\mu$  is a tuple of visual values that specifies the visual characteristics of a part of  
 185     the graphic. For example, given a fiber of  $\{xpos, ypos, color\}$  one possible section could be  
 186      $\{.5, .5, (255, 20, 147)\}$ . The  $\text{codomain}(\nu_i)$  determines the monoid actions on  $P_i$ . These fiber  
 187     components are implicit in the library, by making them explicit as components of the fiber  
 188     we can build consistent definitions and expectations of how these parameters behave.

### 189     1.3.2 Visual Encoders $\nu$

As introduced in section ??, there are many ways to visually represent data components.  
 We define the visual transformers  $\nu$

$$\{\nu_0, \dots, \nu_n\} : \{\tau_0, \dots, \tau_n\} \mapsto \{\mu_0, \dots, \mu_n\} \quad (23)$$

190     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  
 191     there is a monoid  $M'_i$  on  $V_i$ , then there is a monoid homomorphism from  $M_i \rightarrow M'_i$  that  
 192      $\nu$  must preserve. As mentioned in section 1.1.2, we choose monoid actions as the basis for  
 193     equivariance because they define the structure on the fiber components.

A validly constructed  $\nu$  is one where the diagram of the monoid transform  $m$

$$\begin{array}{ccc} E_i & \xrightarrow{\nu_i} & V_i \\ m_r \downarrow & & \downarrow m_v \\ E_i & \xrightarrow{\nu_i} & V_i \end{array} \quad (24)$$

commutes such that

$$\nu_i(m_r(E_i)) = \varphi(m_v)(\nu_i(E_i)) \quad (25)$$

This equivariance constraint yields guidance on what makes for an invalid transform.  
 For example, given a visual fiber of the same type as the data fiber  $F_i = P_i$ , the transform  
 $\nu_i(x) = .5$  does not commute under translation  $t(x) = x + 2$

$$\nu(t(r + 2)) \stackrel{?}{=} \nu(r) + \nu(2) \quad (26)$$

$$.5 \neq .5 + .5 \quad (27)$$

```

[2]: nu = {'confused': ':(', 'woozy': '=(', 'shruggy': '=@')
[3]: nu.keys()
[3]: dict_keys(['confused', 'woozy', 'shruggy'])
[4]: nu.values()
[4]: dict_values([(':(', '=(', '@=')])
[14]: values
[14]: ['woozy', 'shruggy', 'confused']
[15]: [nu[v] for v in values]
[15]: ['=((', '@=)', ':(']

```

Figure 8: In this artis,  $\nu$  maps the strings to the emojis. For  $\nu$  to be equivariant, a shuffle in the words should have an equivalent shuffle in the emojis, and a shuffle in the emojis should have an equivalent shuffle in the words.

On the other hand figure 8 illustrates a valid  $\nu$  mapping from **Strings** to symbols. The group action on these sets is permutation, so shuffling the words must have an equivalent shuffle of the symbols they are mapped to. Continuing with the assumption that  $F = P$ , we can describe the constraints on the other Steven's measurement group types. To preserve ordinal and partial order monoid actions,  $\nu$  must be a monotonic function such that given  $r_1, r_2 \in E_i$

$$\text{if } delement_1 \leq r_2 \text{ then } \nu(r_1) \leq \nu(r_2) \quad (28)$$

the visual encodings must also have some sort of ordering. For interval scale data,  $\nu$  is equivariant under translation monoid actions if

$$\nu(x + c) = \nu(x) + \nu(c) \quad (29)$$

while for ratio data, there must be equivalent scaling

$$\nu(xc) = \nu(x)\nu(c) \quad (30)$$

194 The assumption  $F_i = P_i$  is made here so that we could omit  $\varphi : M \rightarrow M'$  maps that convert  
195 the monoid action on  $F_i$  to the equivalent action on  $P_i$ . As shown in table 1, it is not  
196 uncommon for data and visual variables to resolve to the same types. The constraints on  $\nu$   
197 can be embedded into our artist such that the  $\nu$  functions are equivariant and also provide  
198 guidance on constructing new equivariant  $\nu$  functions.

199 **1.3.3 Graphic Assembler  $Q$**

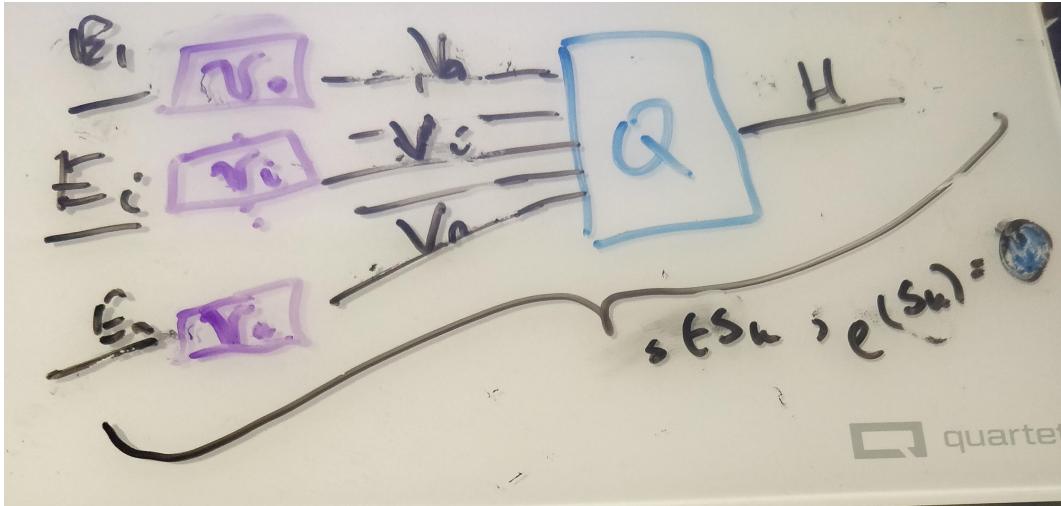


Figure 9:  $\nu$  functions convert data  $\tau_i$  to visual characteristics  $\mu_i$ , then  $Q$  assembles  $\mu_i$  into a graphic  $\rho$  such that there is a map  $\xi$  preserving the continuity of the data.  $\rho$  applied to a region of connected components  $S_j$  generates a part of a graphic, for example the point graphical mark.

200 As shown in figure 9, the assembly function  $Q$  combines the fiber  $F_i$  wise  $\nu$  transforms into  
201 a graphic in  $H$ . Together,  $\nu$  and  $Q$  are a map-reduce operation: map the data into their

202 visual encodings, reduce the encodings into a graphic. As with  $\nu$  the constraint on  $Q$  is  
 203 that for every monoid action on the input  $\mu$  there is corresponding monoid action on the  
 204 output  $\rho$ .

While  $\rho$  generates the entire graphic, we will restrict the discussion of  $Q$  to generation  
 of sections of a glyph. We formally describe a glyph as  $Q$  applied to the regions  $k$  that map  
 back to a set of connected components  $J \subset K$  as input:

$$J = \{j \in K \text{ s. t. } \exists \gamma \text{ s.t. } \gamma(0) = k \text{ and } \gamma(1) = j\} \quad (31)$$

where the path[16]  $\gamma$  from  $k$  to  $j$  is a continuous function from the interval  $[0,1]$ . We define  
 the glyph as the graphic generated by  $Q(S_j)$

$$H \xrightleftharpoons[\rho(S_j)]{} S_j \xrightleftharpoons[\xi^{-1}(J)]{} J_k \quad (32)$$

205 such that for every glyph there is at least one corresponding section on  $K$ . This is in  
 206 keeping with the definition of glyph as any differentiable element put forth by Ziemkiewicz  
 207 and Kosara[46]. The primitive point, line, and area marks[6, 12] are specially cased glyphs.

208 It is on sections of these glyphs that we define the equivariant map as  $Q : \mu \mapsto \rho$  and an  
 209 action on the subset of graphics  $Q(\Gamma(V)) \in \Gamma(H)$  that  $Q$  can generate. We then define the  
 210 constraint on  $Q$  such that if  $Q$  is applied to  $\mu, \mu'$  that generate the same  $\rho$  then the output  
 211 of both sections acted on by the same monoid  $m$  must be the same. While it may seem  
 212 intuitive that visualizations that generate the same glyph should consistently generate the  
 213 same glyph given the same input, we formalize this constraint such that it can be specified  
 214 as part of the implementation of  $Q$ .

Lets call the visual representations of the components  $\Gamma(V) = X$  and the graphic  
 $Q(\Gamma(V)) = Y$ . If for all monoids  $m \in M$  and for all  $\mu, \mu' \in X$ , the output is equivalent

$$Q(\mu) = Q(\mu') \implies Q(m \circ \mu) = Q(m \circ \mu') \quad (33)$$

215 then a group action on  $Y$  can be defined as  $m \circ \rho = \rho'$ . The transformed graphic  $\rho'$  is  
 216 equivariant to a transform on the visual bundle  $\rho' = Q(m \circ \mu)$  on a section that  $\mu \in Q^{-1}(\rho)$   
 217 that must be part of generating  $\rho$ .

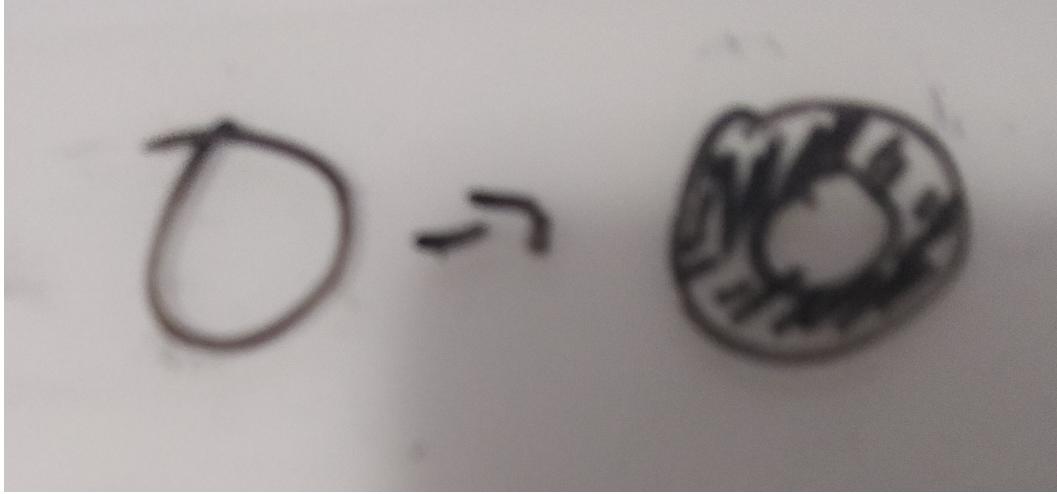


Figure 10: These two glyphs are generated by the same  $Q$  function, but differ in the value of the edge thickness parameter  $\mu_i$ . A valid  $Q$  is one where a shift in  $\mu_i$  is reflected in the glyph generated by  $\rho$ .

218     The glyph in figure 10 has the following characteristics  $P$  specified by  $(xpos, ypos, color, thickness)$   
 219    such that one section is  $\mu = (0, 0, 0, 1)$  and  $Q(\mu) = \rho$  generates a piece of the thin hollow  
 220    circle. The equivariance constraint on  $Q$  is that the action  $m = (e, e, e, x + 2)$ , where  $e$  is  
 221    identity, translates  $\mu$  to  $\mu' = (e, e, e, 3)$ . The corresponding action on  $\rho$  causes  $Q(\mu')$  to be  
 222    the thicker circle in figure 10.

### 223    1.3.4 Assembly $Q$

224    In this section we formulate the minimal  $Q$  that will generate distinguishable graphical  
 225    marks: non-overlapping scatter points, a non-infinitely thin line, and an image.

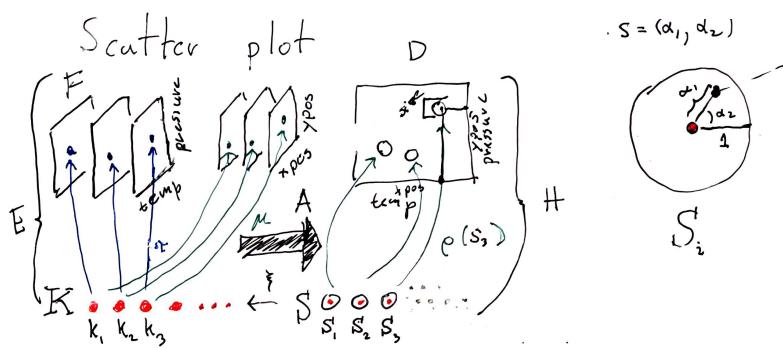


Figure 11: The data is discrete points (temperature, time). Via  $\nu$  these are converted to  $(xpos, ypos)$  and pulled over discrete  $S$ . These values are then used to parameterize  $\rho$  which returns a color based on the parameters  $(xpos, ypos)$  and position  $\alpha, \beta$  on  $S_k$  that  $\rho$  is evaluated on.

The scatter plot in figure ?? can be defined as  $Q(xpos, ypos)(\alpha, \beta)$  where color  $\rho_{RGB} = (0, 0, 0)$  is defined as part of  $Q$  and  $s = (\alpha, \beta)$  defines the region on  $S$ . The position of this swatch of color can be computed relative to the location on the disc  $S_k$  as shown in figure 11:

$$x = size \bullet \alpha \bullet \cos(\beta) + xpos \quad (34)$$

$$y = size \bullet \alpha \bullet \sin(\beta) + ypos \quad (35)$$

<sup>226</sup> such that  $\rho(s) = (x, y, 0, 0, 0)$  colors the point  $(x, y)$  black.

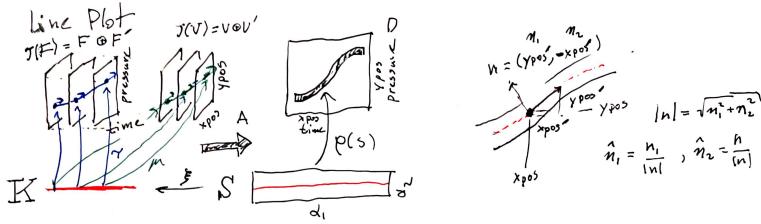


Figure 12: The line fiber (*time, temp*) is thickened with the derivative (*time', temperature'*) because that information will be necessary to figure out the tangent to the point to draw a thick line. This is because the line needs to be pushed perpendicular to the tangent of  $(xpos, ypos)$ . *this is gonna move once this gets regenerated w/ labels* The data is converted to visual characteristics  $(xpos, ypos)$ . The  $\alpha$  coordinates on  $S$  specifies the position of the line, the  $\beta$  coordinate specifies thickness.

The line plot  $Q(xpos, \hat{n}_1, ypos, \hat{n}_2)(\alpha, \beta)$  shown in fig 11 exemplifies the need for the jet. The 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. The magnitude of the thickness is

$$|n| = \sqrt{n_1^2 + n_2^2} \quad (36)$$

such that the normal is

$$\hat{n}_1 = \frac{n_1}{|n|}, \quad \hat{n}_2 = \frac{n_2}{|n|} \quad (37)$$

which yields components of  $\rho$

$$x = xpos(\xi(\alpha)) + \beta \hat{n}_1(\xi(\alpha)) \quad (38)$$

$$y = ypos(\xi(\alpha)) + \beta \hat{n}_2(\xi(\alpha)) \quad (39)$$

<sup>227</sup> where  $(x, y)$  look up the position  $\xi(\alpha)$  on the data. At that point, we also look up the  
<sup>228</sup> the derivatives  $\hat{n}_1, \hat{n}_2$  which are then multiplied by data to specify the thickness.

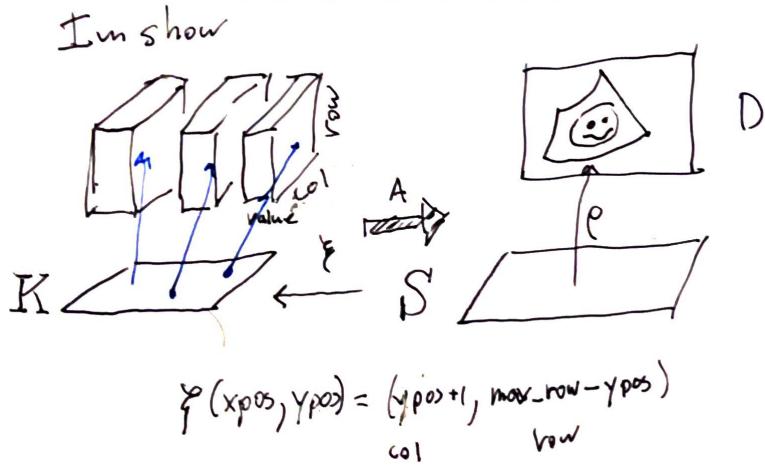


Figure 13: The only visual parameter a image requires is color since  $\xi$  encodes the mapping between position in data and position in graphic.

<sup>229</sup> The image  $Q(\text{color})$  in figure 13 is a direct lookup  $\xi : S \rightarrow K$  such that

$$R = R(\xi(\alpha, \beta)) \tag{40}$$

$$G = G(\xi(\alpha, \beta)) \tag{41}$$

$$B = B(\xi(\alpha, \beta)) \tag{42}$$

<sup>230</sup> where  $\xi$  may do some translating to a convention expected by  $Q$  for example reorientng the  
<sup>231</sup> array such that the first row in the data is at the bottom of the graphic.

### <sup>232</sup> 1.3.5 Assembly factory $\hat{Q}$

<sup>233</sup> The graphic base space  $S$  is not accessible in many architectures, including Matplotlib;  
<sup>234</sup> instead we can construct a factory function  $\hat{Q}$  over  $K$  that can build a  $Q$ . As shown in  
<sup>235</sup> eq 22,  $Q$  is a bundle map  $Q : \xi^* V \rightarrow H$  where  $\xi^* V$  and  $H$  are both bundles over  $S$ .

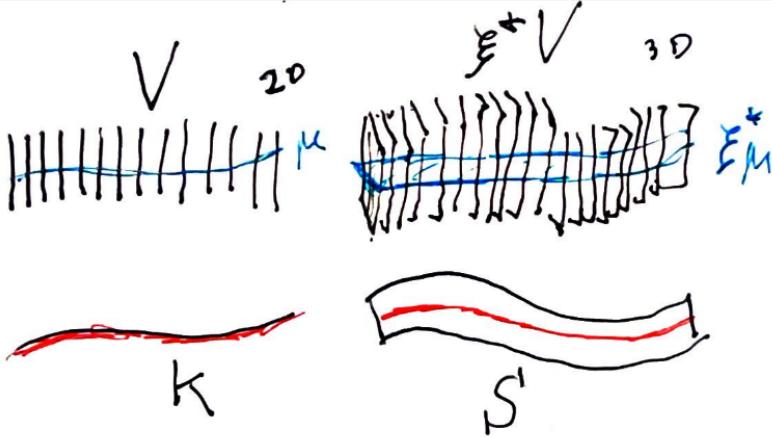


Figure 14: The pullback of the visual bundle  $\xi^*V$  is the replication of a  $\mu$  over all points  $s$  that map back to a single  $k$ . Because the  $\mu$  is the same, we can construct a  $\hat{Q}$  on  $\mu$  over  $k$  that will fabricate the  $Q$  for the equivalent set of  $s$  associated to that  $k$

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  $\mu$

$$\xi^*V|_{\xi^{-1}(k)} = \xi^{-1}(k) \times P \quad (43)$$

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

238 This copying is illustrated in figure 14, where the 1D fiber  $P \hookrightarrow V$  over  $K$  is copied  
239 repeatedly to become the 2D fiber  $P^*\mu \hookrightarrow \xi^*V$  with identical components over  $S$ . Given  
240 the section  $\xi^*\mu$  pulled back from  $\mu$  and the point  $s \in \xi^{-1}(k)$ , there is a direct map from  $\mu$  on  
241 a point  $k$ , there is a direct map from the visual section over data base space  $(k, \mu(k)) \mapsto$   
242  $(s, \xi^*\mu(s))$  to the visual section  $\xi^*\mu$  over graphic base space. This map means that the  
243 pulled back section  $\xi^*\mu(s) = \xi^*(\mu(k))$  is the section  $\mu$  copied over all  $s$ . This means that  
244  $\xi^*\mu$  is identical for all  $s$  where  $\xi(s) = k$ , which is illustrated in figure 14 as each dot on  $P$  is  
245 equivalent to the line intersection  $P^*\mu$ .

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

$$\rho(s) := Q(\xi^*\mu)(s) \quad (44)$$

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

$$\hat{Q}(\mu(k))(s) := Q((\xi^*\mu)(s)) \quad (45)$$

246 where the assembly function  $\hat{Q}$  that takes as input the visual section on data  $\mu$  is defined  
247 to be the assembly function  $Q$  that takes as input the copied section  $\xi^*\mu$  such that both  
248 functions are evaluated over the same location  $\xi^{-1}(k) = s$  in the base space .

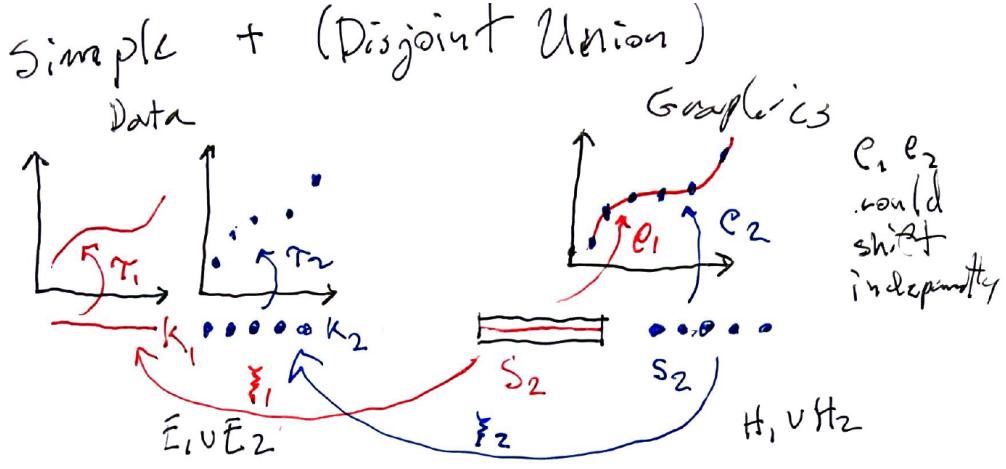


Figure 15:  $\tau_1$  and  $\tau_2$  are distinct datasets passed through artists  $A_1$  and  $A_2$  to generate graphics  $\rho_1$  and  $\rho_2$ . These graphics happen to be rendered to the same image, but otherwise have no intrinsic link.

Factoring out  $s \hat{Q}(\mu(k)) = Q(\xi^* \mu)$  generates a curried  $\hat{Q}$ . 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  $\hat{Q} : \Gamma(V_k) \rightarrow \Gamma(H|_{\xi^{-1}(k)})$ . This allows us to construct a  $\hat{Q}$  that only depends on  $K$ , such that for each  $\mu(k)$  there is part of  $\rho|_{\xi^{-1}(k)}$ . The construction of  $\hat{Q}$  allows us to retain the functional map reduce benefits of  $Q$  without having to majorly restructure the existing rendering pipeline.

### 1.3.6 Sheaf

The restriction maps of a sheaf describe how local  $\tau$  can be glued into larger sections [19, 20]. As part of the definition of local triviality, there is an open neighborhood  $U \subset K$  for every  $k \in K$ . We can define the inclusion map  $\iota : U \rightarrow K$  which pulls  $E$  over  $U$

$$\begin{array}{ccc} \iota^* E & \xhookrightarrow{\iota^*} & E \\ \pi \downarrow \lrcorner \iota^* \tau & & \pi \downarrow \lrcorner \tau \\ U & \xhookrightarrow{\iota} & K \end{array} \quad (46)$$

such that the pulled back  $\iota^* \tau$  only contains records over  $U \subset K$ . By gluing  $\iota^* \tau$  together, the sheaf is putting a continuous structure on local sections which allows for defining a section over a subset in  $K$ . That section over subset  $K$  maps to the graphic generated by  $A$  for visualizations such as sliding windows[15, 17] streaming data, or navigation techniques such as pan and zoom[31].

### 1.3.7 Composition of Artists: +

To build graphics that are the composites of multiple artists, we define a simple addition operator that is the disjoint union of fiber bundles  $E$ . For example, in figure 15 the scatter

plot  $E_1$  and the line plot  $E_2$  have different  $K$  that are mapped to separate  $S$ . To fully display both graphics, the composite graphic  $A_1 + A_2$  needs to include all records on both  $K_1$  and  $K_2$ , which are the sections on the disjoint union  $K_1 \sqcup K_2$ . This in turn yields disjoint graphics  $S_1 \sqcup S_2$  rendered to the same image. Constraints can be placed on the disjoint union such as that the fiber components need to have the same  $\nu$ position encodings or that the position  $\mu$ need to be in a specified range. There is a second type of composition where  $E_1$  and  $E_2$  share a base space  $K_2 \hookrightarrow K_1$  such that the the artists can be considered to be acting on different components of the same section. This type of composition is important for creating visualizations where elements need to update together in a consistent way, such as multiple views [2, 32] and brush-linked views[5, 8].

#### 1.3.8 Equivalence class of artists $A'$

It is impractical to implement an artist for every single graphic; instead we implement an approximation of an the equivalence class of artists  $\{A \in A' : A_1 \equiv A_2\}$ . Roughly, equivalent artists have the same fiber bundle  $V$  and same assembly function  $Q$  but act on different sections  $\mu$ , but we will formalize the definition of the equivalence class in future work.

As a first pass for implementation purposes, we identify a minimal  $P$  associated with each  $A'$  that defines what visual characteristics of the graphic must originate in the data such that the graphic is identifiable as a given chart type.

Figure 16: Each of these graphics is generated by a different artist  $A$ which is the equivalence class of scatter plots  $A'$   
**this is gonna be a whole bunch of scatter plots**

For example, a scatter plot of red circles is the output of one artist, a scatter plot of green squares the output of another, as shown in figure 16. These two artists are equivalent since their only difference is in the literal visual encodings (color, shape). Shape and color could also be defined in  $Q$  but the position must come from the fiber  $P = (xpos, ypos)$  since fundamentally a scatter plot is the plotting of one position against another[18]. We also use this criteria to identify derivative types, for example the bubble chart[43] is a type of scatter where by definition the glyph size is mapped from the data. The criteria for equivalence class membership serves as the basis for evaluating invariance[26].