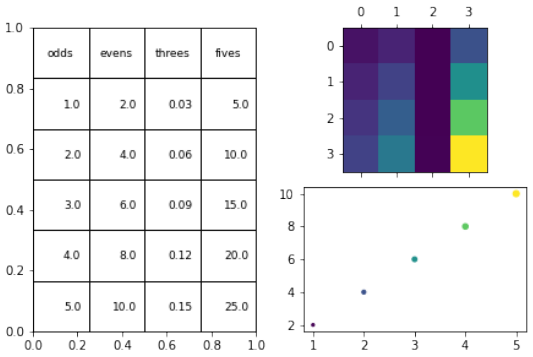

Figure 1: Implicit in visualization is the assumption that these three representations of data are equivalent, specifically that the measurements within a variable and relations of the measurements of each variable are preserved.

1 Introduction

1.1 Thesis statement

We define a visualization as a transform from data to graphic that preserves the topology of the data and faithfully map the properties of the measurement type. In fig 1, we implicitly assume that the translation from table to heatmap has preserved the order of observations (the rows) and that the perceptually uniform sequential colormap has been applied such that the ordering relation on floats matches the ordering on the colormap (darker colors map to larger numbers). We also make this assumption about color in the scatter map, and that the translation to size and position on screen also respect the ordering on floats. In this work, we propose to mathematically describe the transform of data to visual space such that we can make explicit the implicit topology and types visualizations preserve. We then propose a new architecture for the Python visualization library Matplotlib [1] based on these descriptions because the Matplolib artist layer is analogous to the transforms.

1.2 What is a viz

? Acquired codes of meaning

2 Not all data are tables

Tables, images (Lev), graphs (network X)

set up: dubois

theorists: bertin, munzner, mackinlay

talk about: matplotlib arch paper, excel/matlab arch, vtk & ggplot (compare/contrast, we're blending these things)

2.1 Terminology

Given a dataset, we need to decide what subset of the data to visualize. Bertin describes the set of constraints used to subset the data as the *invariant*. Formally, the *invariant*

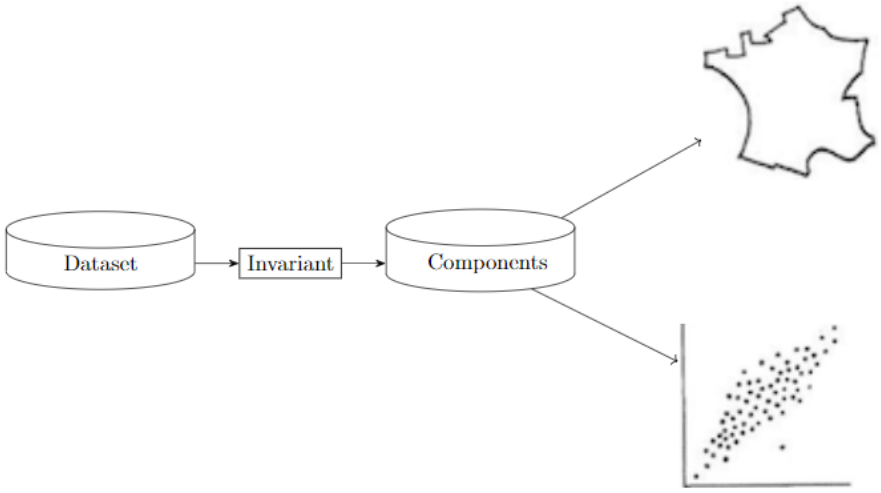

Figure 2: To go from a dataset to a visualization, the data is subset based on a set of constraints (the invariant). The resulting subset becomes the components that are visualized, but the choice of visualization is dependent on the type and structure of the component variables.

is the set of shared characteristics of the data being visualized. When these constraints are applied to the dataset, the resulting subset is what will become the *components* of the visualization [**bertin˙semiology˙2011**]. In figure **??**, the *invariant* common to all the data being visualized is "sepal length", "petal length", and "species" and the *components* are the measurements of these variables. As shown in figure 2, the final step in creating a visualization is choosing how to encode the components using retinal (visual) variables.

2.1.1 Retinal (Visual) Variables

Figure 3 illustrates common guidelines for encoding *components*, derived from what Bertin terms a retinal variable and most other visualization theorists call visual variables [**bertin˙semiology˙2011**, **krygier˙making˙2005**, **chambers˙graphical˙1983**, **wilkinson˙grammar˙2005**, **munzner˙visualization˙2014**]. The columns of figure 3 correspond to the type of observation: discrete points, continuous events (e.g. a timeseries), two dimensional continuous events (e.g. a vector map). The rows of figure 3 describe ways to visualize variations in the *components*; generally, quantitative components are represented by retinal variables that change quantitatively and categorical components are represented by retinal variables that vary qualitatively. In figure **??**, the hue of the marker is used to encode differentiation in species and the position of the marker is used to show variation in petal length and sepal length. The retinal variables suggest that any single visualization is limited to encoding at most about 8 or 9 components. Retinal variables provide guidelines for encoding *components*, but the choice of graph is based on the type and structure of the data.

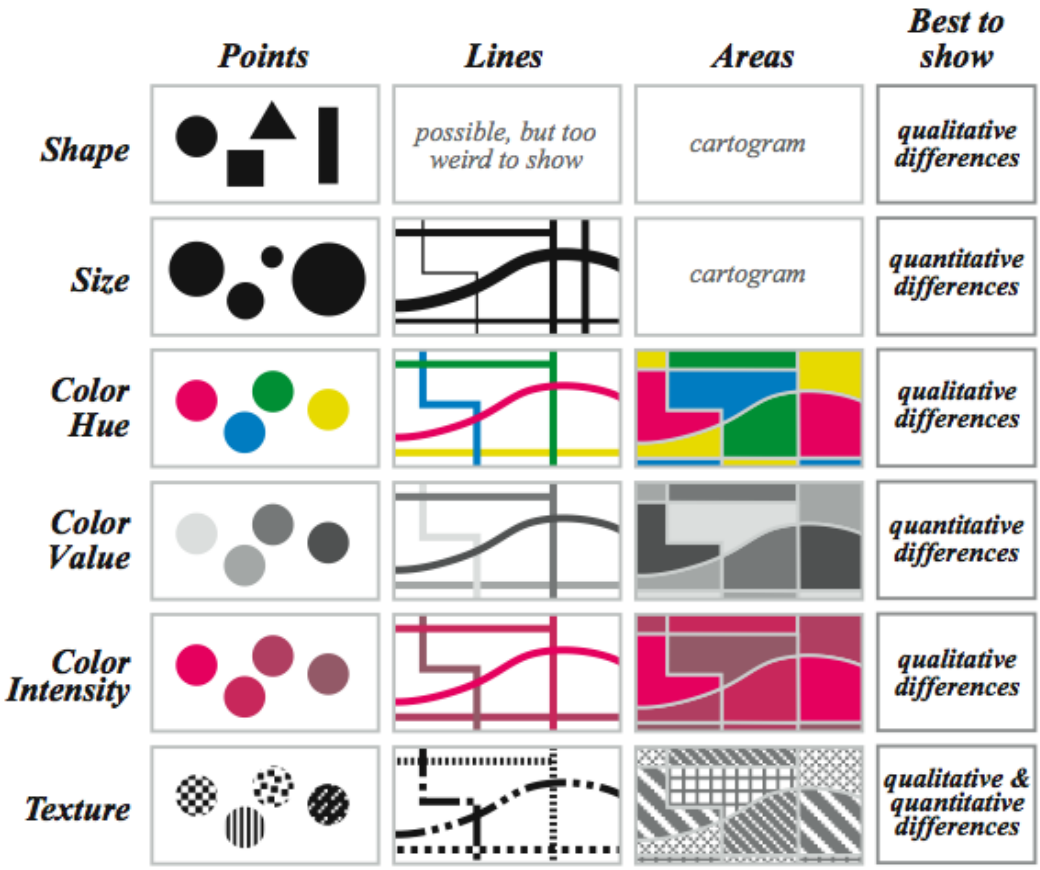

Figure 3: Retinal variables are a codification of how position, size, shape, color and texture are used to illustrate variations in the *components* of a visualization. This tabular form of Bertin's retinal variables is from Understanding Graphics [**˙information˙????**] who reproduced it from *Making Maps: A Visual Guide to Map Design for GIS* [**krygier˙making˙2005**]

2.1.2 Data Type and Structure

As shown in figure 2, there are multiple ways to translate data into pictures. A map is always an option, except when the observations do not have associated coordinates in a physical plane. Tamara Munzner provides a way to distinguish between these datasets using $\{key, value\}$ designations [**munzner˙visualization˙2014**]. Munzner defines *values* as measurements of interest in the dataset, analogous to dependent variables in statistics. She defines *keys* as indexes that can be used to look up values, analogous to independent variables in statistics and dimensions in computer science. Figure 4 illustrates how these keys are used to look up variables in a dataset:

- map: keys are the coordinates of the points
- table: row index, database primary key

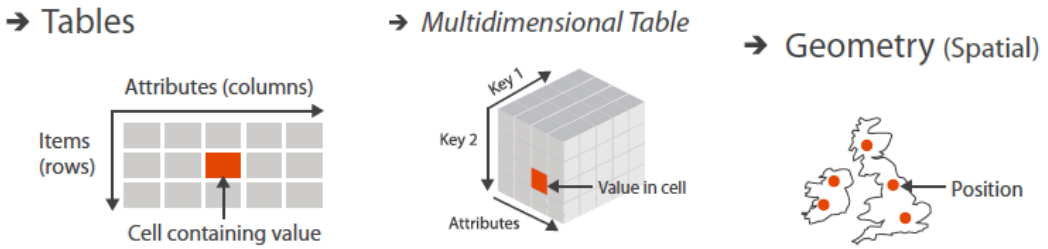

Figure 4: Keys are unique lookup values used to find individual observations in the dataset. Keys are positional references, and can be coordinates on a map or unique values such as a primary key in a database or a (time, latitude, longitude) index in a data cube. Image modified from a diagram from Munzner's website [**˙visualization˙????**]

- data cube: row, column, etc. (.e.g. i, j, k) index

Expanding on Munzner's key and value semantics, in many datasets the keys are discrete variables like time or geophysical locations sampled from a continuous curve, surface, or field. While these observations are discrete samples from the continuous space, often the continuous (functional) characteristic[**ramsay˙functional˙2006**, **muller˙functional˙2006**] of the observational space is also of interest. Besides quantitative discrete, quantitative continuous, or categorical measurement type considerations, the choice of visualization is also influenced by the measurement being on an interval, ratio, nominal, or categorical scale.