DataTours: A Data Narratives Framework

Hrim Mehta* — UOIT, Canada
Amira Chalbi† — Inria, France
Fanny Chevalier‡ — Inria, France
Christopher Collins§ — UOIT, Canada

Abstract

Visual storytelling is commonly employed to communicate data analyses results. Alternatively, (semi-)automated [1,2,6] data narratives or "tours" have been proposed as a means to prompt exploration of massive multidimensional datasets, substituting the more prevalent static overviews. While these works demonstrate specific instances of data tours, a concrete model to describe the building blocks of such tours is lacking. We present a descriptive hierarchical framework, DataTours, to formalize and guide the design of (semi-) automated tours for data exploration and discuss challenges evoked by the framework in the (semi-)automated authoring of such tours.

1 Introduction

Visual analytics systems for massive multidimensional datasets predominantly present an analyst with static overviews of "interesting" subsets of the data to prompt exploratory analysis. While narratives have been widely used for presenting visual analyses, some previous works have proposed to leverage (semi-)automated tours for supporting data exploration in lieu of static overviews.

The concept of automated data tours originates from Asimov's "grand tour" [1], comprising of an animated sequence of projections of multidimensional data points into different two dimensional planes, for exploring relations in multivariate data. This concept has since been extended to other types of data and visualizations. For example, Yu et al. [6] presented a technique for automatically constructing an animated tour of temporal events in a time-varying dataset, sequenced based on the traversal of an event graph. Dennis and Healey [2] proposed a technique to generate animated tours through data points of interest in a multidimensional dataset, with iterative refinement of the interest function during exploration based on explicit or implicit user interactions. While prior works have demonstrated specific instances of data tours, there is a lack of not only a framework to systematically describe the components that make a tour but also a discussion on design decisions to be made in the authoring of (semi-)automated tours and the impact of such tours on the data exploration process.

We present a descriptive hierarchical framework, DataTours, to describe animated data tours (Fig. 1). Data operations (selection and sequencing) and view operations (staging and transition) form the main components of the framework. Each node of the tree is built with data operations that dictate the extraction of data elements to be presented in the tour, and view operations that control the generation of, and transition between, views for the selected data elements. Root's sub-trees represent sub-narratives or facets to be explored, with the tree traversal order leading to a full narrative structure. We draw inspiration for our model from existing frameworks of author-driven narratives [4,5] and extend these frameworks to encapsulate the design decisions for (semi-)automated data tours.

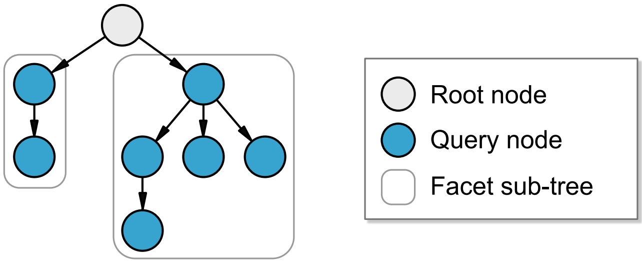

Figure 1: A hierarchical framework to describe data tours

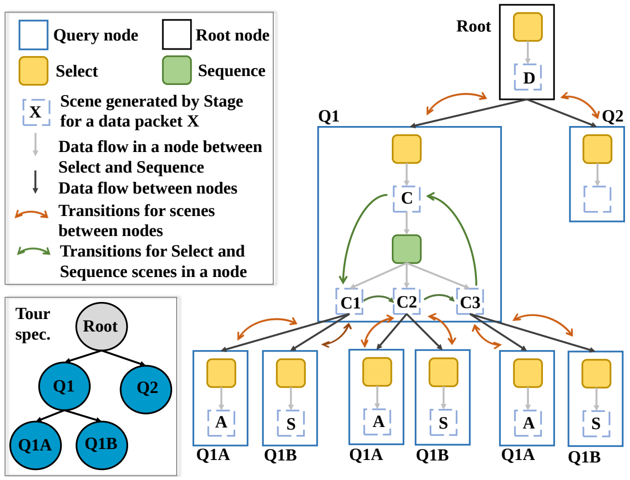
Query node
Root node
Select
Sequence
X
Scene generated by Stage for a data packet X
Data flow in a node between Select and Sequence
Data flow between nodes
Transitions for scenes between nodes
Transitions for Select and Sequence scenes in a node
Root
D
Q1
Q2
C
C1
C2
C3
A
S
Q1A
Q1B
Tour spec.
Root
Q1
Q2
Q1A
Q1B

Figure 2: DataTours specification for the obesity risk factors scenario. Q1 initiates the poverty facet by selecting high poverty rate counties (C) and grouping them into poverty ranges (C1, C2, C3) to observe obesity rates in poor counties. For each county group, Q1A and Q1B explore the effect of lacking access to fresh food (A) & sedentariness (S) on obesity rates. Q2 commences exploration of the diabetes facet.

2 DataTours Framework

The DataTours framework specifies a data tour using a hierarchical model (Fig. 1). The root node represents entry point of the tour. Each sub-tree of the root corresponds to a single facet of the data to be toured. Narrative structure of the tour depends on number and depth of the facets or sub-trees to be explored and traversal order of the tree which determines the sequence of presentation of the facets.

Each node in the facet sub-tree comprises of *data* and *view* operations. Data operations (*selection* and *sequencing*) extract a subset of the dataset based on particular "interest" functions, and optionally divide and order the selected data elements into groups. View operations (*staging* and *transition*) generate a view for each of the data element groups returned by data operations, and specify transitions between the views. Specification of operators controlling the four operations at a node can be manual, semi-automated or automated.

We will explore our model with a specific example. Consider the design of a tour that presents correlations of obesity rates in US counties with poverty and diabetes rates to explore obesity risk factors. The tour's main narrative can be split into two sub-narratives, or *facets*, one each for exploring the relation between obesity rates

*e-mail: hrim.mehta@uoit.ca
†e-mail:amira.chalbi@inria.fr
‡e-mail:fanny.chevalier@inria.fr
§e-mail:christopher.collins@uoit.ca

and poverty and diabetes rates. Additionally, we want to develop the tour's poverty facet to investigate how inaccessibility to fresh foods and sedentariness in poor counties increase the risk of obesity.

For this scenario, root of our framework model represents the tour's initial view comprising of a choropleth map encoding obesity rates across US counties. Two sub-trees are initiated from the root to elucidate correlations of obesity rates to poverty and diabetes rates (Fig. 2). For the poverty facet, a node Q1 specifies the focus on counties with high poverty rates (C in Fig. 2), followed by a grouping of these counties into poverty rate ranges (C1, C2, C3 in Fig. 2) and a sequential touring of each of the county groups to further investigate possible factors leading to high obesity rates in poor counties. Node Q1A is initiated to investigate how inaccessibility to fresh foods leads to higher obesity risk for each of the county groups. Similarly, node Q1B helps explore how sedentariness contributes to high obesity in poor counties. For the second facet, node Q2 presents the state of obesity rates in counties with high diabetes rates (Fig. 2).

2.1 Selection

A node's selection data operation extracts a subset of data input to the node based on a select operator. Based on the chosen operator, the selection operation at a node can be either reductive (drill-down) or expansive (contextual). A reductive selection operation returns a subset of the input data that satisfies the select operator. An expansive selection operation, on the other hand, fetches supplementary data constrained by the data input to the node.

In our use case, to focus on counties with high poverty rates, the selection data operation at node Q1 comprises of a reductive select operator that returns a set of counties with poverty rates $> 25\%$.

2.2 Sequencing

A node's sequencing data operation splits the data returned by the selection operation into groups, where a group can comprise one or more data elements, and assigns an order to each of the groups for further sequential processing based on a sequence operator. Invocation of a sequencing data operation at a node is optional.

In our example, since we want to further explore sets of counties within each of three poverty ranges ($25-30\%, 30-35\%$, and $> 35\%$) individually, to present factors correlated with high obesity rates in poor counties, a sequencing data operation is additionally specified at Q1. The sequence operator at Q1 returns sets of counties, each belonging to a single poverty range, ordered by increasing poverty rate. To investigate obesity risk factors such as, accessibility to fresh food and sedentariness, for each set of counties within a given poverty range, new nodes Q1A and Q1B for each of the risk factors have to be appended to Q1 (see Fig. 2).

2.3 Staging

Each data operation (selection or sequencing) at a node is accompanied by a staging view operation. A staging operation, akin to staging in computer animation [3], generates a scene for the data returned by a data operation as dictated by a stage operator. A stage operator specifies the viewport bounds, level of zoom, and the scene duration. Additionally, the stage operator can also specify generation of visual aids in the form of emphasis (highlights, annotations), captions, changes to visual encodings, re-layouts, overlays of supplementary visualizations, etc.

The staging operation for the selection operation at node Q1 generates a scene C (Fig. 2) that encompasses and highlights all counties with high poverty rates. Similarly, the staging operation for Q1's sequencing operation generates multiple scenes (C1, C2, C3 in Fig. 2), each enclosing the set of counties within a given poverty range. The scenes also comprise of an overlay of a scatterplot of obesity and poverty rates for the counties of interest in a given scene.

2.4 Transition

A transition operation dictates the transition from some scene to the scene generated by a staging operation per the transition operator. A transition operator details the duration and easing for the camera path (changes in viewport and zoom level) as well as the visual changes (layout or encoding changes, removal/addition of captions, etc.) during a transition between two scenes.

In our use case, the transition operator for the scene at Q1's selection specifies that the scene C can be reached through a camera path change lasting a duration of 500*ms* with cubic easing. Once the scene is reached, highlights for the counties of interest should fade in and out, following linear easing over a duration of time calculated based on the number of counties of interest in the scene.

Transition operations for scenes generated for data groups returned by sequencing operations are comprised of two transition sub-operators, *within* and *between*. The within transition operator dictates the transitions between pairs of scenes for data groups returned by the sequencing operation (e.g., C1-C2, C2-C3 in Fig. 2). The between transition operator dictates transitions between the scenes for data groups returned by selection and sequencing operations at a node (e.g., C-C1, C3-C in Fig. 2).

3 Discussion

The DataTours formalization evokes challenges in the (semi-) automated authoring of data tours and raises questions about the impact of the tours on data exploration. Framework components identify the design decisions to be made when specifying a data tour: the choice of sub-narratives or facets and depth of exploration of each facet, specification of appropriate data and view operators for each node of a facet, and the tree traversal order for generating a narrative structure. These in turn invoke investigation into how data characteristics, visual representation choices, and user interactions might be leveraged to inform the design decisions of a tour to guide data exploration for a user task or hypothesis. The concept of (semi-)automated data tours also raises questions about the comprehensibility of such tours and the narrative impacts of the design decisions ranging from aiding to biasing visual analysis. These concerns point towards research on evaluation techniques for determining effectiveness of the tours at supporting data exploration.

Based on the framework, a variety of data tours can be designed. For example, a tour to introduce the data attributes, their visual encodings, and the common analytical tasks that can be performed may help a novice user. For an expert user, an adaptive data tour might be designed to present recommendations for exploration based on user interactions. For an expert user analyzing streaming data, a data tour can convey what data attributes of relevance have changed and how they compare to the expert's last analysis session. The DataTours framework thus provides a robust formalization to guide the design of (semi-)automated tours for data exploration.

References

[1] D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6(1):128–143, Jan. 1985.

[2] C. G. Healey and B. M. Dennis. Interest driven navigation in visualization. *IEEE transactions on visualization and computer graphics*, 18(10):1744–1756, 2012.

[3] J. Lasseter. Principles of traditional animation applied to 3d computer animation. In *ACM Siggraph Computer Graphics*, vol. 21, pp. 35–44. ACM, 1987.

[4] A. Satyanarayan and J. Heer. Authoring narrative visualizations with ellipsis. In *Computer Graphics Forum*, vol. 33, pp. 361–370. Wiley Online Library, 2014.

[5] M. Wohlfart and H. Hauser. Story telling aspects in volume visualization. In *Proceedings of EuroVis*, pp. 91–8, 2007.

[6] L. Yu, A. Lu, W. Ribarsky, and W. Chen. Automatic animation for time-varying data visualization. In *Computer graphics forum*, vol. 29, pp. 2271–2280. Wiley Online Library, 2010.