EuroVis Workshop on Visual Analytics (2015)
E. Bertini and J. C. Roberts (Editors)

Inspector Gadget: Integrating Data Preprocessing and Orchestration in the Visual Analysis Loop

Robert Krüger[1,2], Dominik Herr[1,3], Florian Haag[1] and Thomas Ertl[1]

[1] Institute for Visualization and Interactive Systems, University of Stuttgart, Germany
[2] DFG Cooperative Graduate Program 'Digital Media', HdM Stuttgart & University of Stuttgart, Germany
[3] Graduate School of Excellence advanced Manufacturing Engineering, University of Stuttgart, Germany

Abstract

Nowadays, tracking devices are small and cheap. For analysis tasks, there is no problem to obtain sufficient amounts of data. The challenge is how to make sense of the data, which often contain complex situations. Multiple data sources related to time, space, and other dimensions, with different resolution and notation have to be mapped. Visual approaches often cover an analysis loop that starts right after the preprocessing. In this paper, we contribute methods to explicitly integrate data preprocessing and orchestration into the visual analysis loop. Subsequently, the big picture can be explored in detail and hypotheses can be created, refined, and validated. We showcase our approach with multiple heterogeneous datasets from the VAST Challenge 2014.

Categories and Subject Descriptors : H.5.2 [Information Interfaces and Presentation]: User Interfaces—GUI;

1. Introduction

Object tracking in time and space is a commodity. Tracking devices get cheaper and smaller and their precision increases. Also, sufficient data storage is available. According to SCI Utah, disk storage has been ahead of digital data volumes since 2001 [Joh11]. An ongoing challenge, however, is the sensemaking process that leads from raw data to insights, as addressed by Pirolli and Card [PC05]. Visual Analytics (VA) aims to support this process [SSS*14] by combining analytical reasoning with interactive visual interfaces [TC05]. Nevertheless, creating suitable visualizations is often a chicken-and-egg problem. On the one hand, without knowing the major data structure and properties, it is hard to find suitable visual representations, as demonstrated by Pretorius & Wijk [Pv09]. On the other hand, without suitable representations it is hard to get a first impression of the data. It becomes even more challenging when information is distributed over multiple data sources [NH02,HS98,HS95]. Depending on the tracking hardware capabilities, tracking purpose, and task, data is recorded with different resolution in time and space. For example, for one dataset the temporal information has a high resolution, covering even milliseconds, while others include logs on a daily basis. This is known as the semantic integration problem [BCVB01,DH05]. Hence, analysts have to deal with inaccuracies which hamper an automatic alignment of data [RH01]. The integration of preprocessing tasks, such as cleaning and orchestration into the VA loop, however, has been poorly addressed so far. In this paper, we propose a novel semi-automated process, that lets analysts configure and revise automated decisions and bring in domain knowledge (see Figure 1). The remainder of this paper is structured according to this process and showcases its implementation (see Figure 2, 3) and applicability with an intelligence scenario from the VAST Challenge 2014 [vas14], containing multiple heterogeneous datasets.

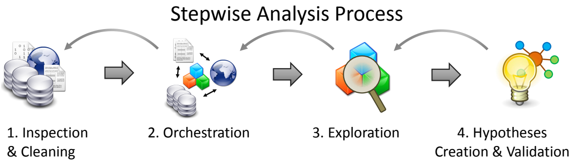
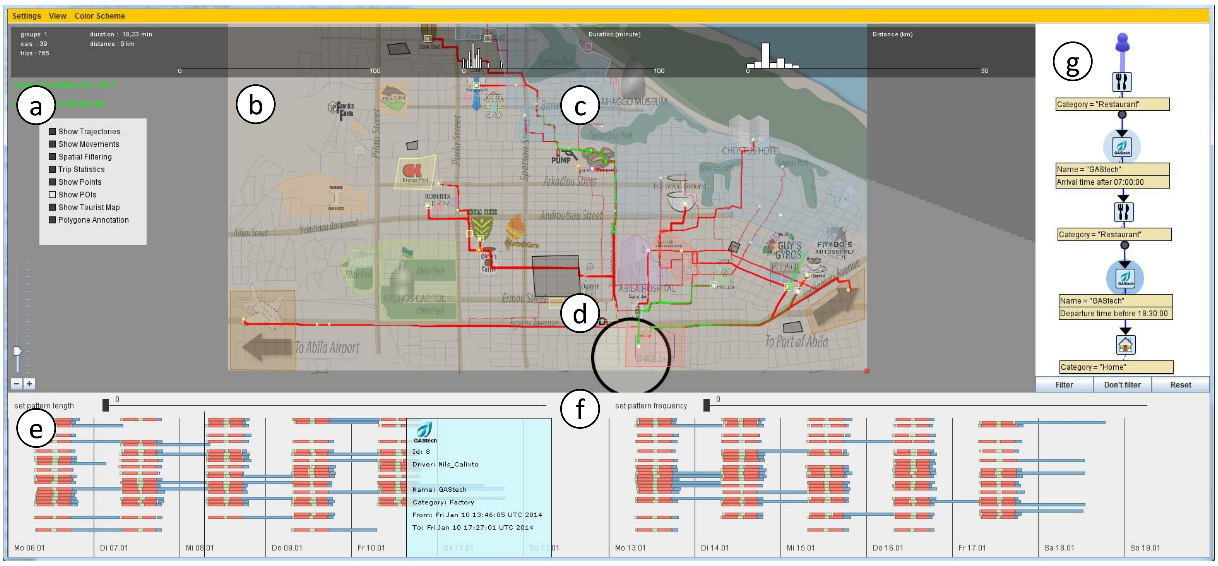

Stepwise Analysis Process
1. Inspection & Cleaning
2. Orchestration
3. Exploration
4. Hypotheses Creation & Validation

Figure 1: ① *Inspect domain unspecific data characteristics / data cleaning & prefiltering;* ② *Map the data sources in domain specific views;* ③ *Explore details with various visualization;* ④ *Externalize findings, refine, validate.*

2. The VAST Challenge Data

The VAST Challenge [SWPG12] is an annual competition. Every year, a synthetic dataset is given that contains various patterns to be found by means of VA. In Mini Challenge

DOI: 10.2312/eurova.20151096



■ Show Trajectories
■ Show Movements
■ Spatial Filtering
■ Trip Statistics
■ Show Points
□ Show POIs
■ Show Tourist Map
■ Polygone Annotation

Category = "Restaurant"
Name = "GAStech"
Arrival time after 07:00:00
Category = "Restaurant"
Name = "GAStech"
Departure time before 18:30:00
Category = "Home"

GAStech
Id: 8
Driver: Nils_Calixto
Name: GAStech
Category: Factory
From: Fri Jan 10 13:46:05 UTC 2014
To: Fri Jan 10 17:27:01 UTC 2014

Figure 2: *The analysis system - ⓐ Map Overlays on/off; ⓑ Geographic Map View - animated movements (green) and trajectories (red); ⓒ Annotations - define and extract areas of interests; ⓓ TrajectoryLenses - filter trajectories by origin/destination/way; ⓔ Sequence View - timeline of AOI movement sequences per employee; ⓕ N-Gram Sequence Filter - detects frequent and outlier patterns on a per user basis; ⓖ Pattern Query Tool - externalize knowledge, refine hypotheses and query the data*

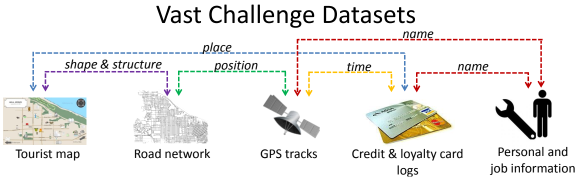


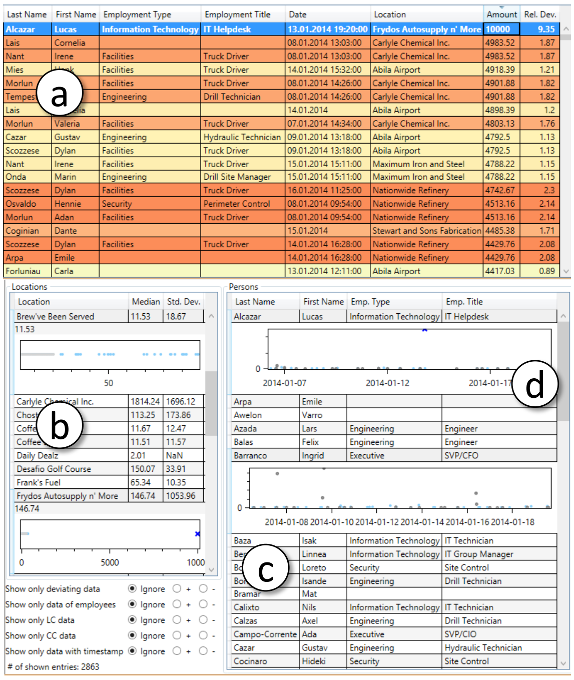
Figure 3: *In the VAST Challenge – MC 2, 2014 heterogeneous datasets with different resolutions had to be aligned.*

2 (MC2) in 2014, the fictitious story takes place in the city *Abila* on the island *Kronos* in the Mediterranean sea, where employees of a company named *GAStech* were kidnapped. Participants had to provide digital forensic tools to analyze the situation and detect daily routines but also suspicious behavior. The data covers a time range of two weeks prior to the kidnapping, and consists of multiple heterogeneous data sources about the employees and GPS tracks of their company cars. Additionally, information on transactions with credit and loyalty cards were available at different temporal resolutions. Lastly, besides a detailed road network, an image of a tourist map provided the main points of interest (POIs) and the main roads between them. Figure 3 provides an overview of the heterogeneous datasets and shows how they relate to each other.

3. Inspection of Unknown Data Structures

At the beginning of an analysis often little to nothing is known about the data characteristics. Hence, before suitable analysis tools can be chosen it is essential to get a first overview of the available data, its features, and its quality level (granularity, completeness, errors). To show plain and unfiltered but structured data, tables are generally an appropriate instrument [RC94, SBB96]. We apply multiple linked table views, as proposed by Tweedie et al. [TSWB94] for interactive exploration. Even without any knowledge about the data domain, we can calculate basic statistics such as mean and deviation for some of the data features. We represent them using visual primitives, which depend on the value scales (e.g. nominal, metric) of the feature.

For the challenge data we show an overview containing all employees, their transactions, and according POIs (see Figure 4a). The background color varies depending on the deviation from the average expenses at the location. Entries with a high similarity are automatically merged and shown as one entry. Additional tables contain locations where transactions occurred (see Figure 4b) and employees that performed the transactions (c). When a field is highlighted, more detailed information is shown below the entry (d). Comparable to the approach of small multiples [Tuf91] we show multiple small scatter plots to indicate the data distribution. This way it is easy to check a seemingly suspicious transaction. Figure 4a shows a very high transaction of $ 10,000 at a car supply shop (highlighted in blue). This is uncommon for both the location (b) as well as the person involved, since there are no other outliers in the detailed view (d). Lastly, the tool supports cleaning and filtering by various characteristics of the meta data in an easy and fast manner. For example, one



Last Name	First Name	Employment Type	Employment Title	Date	Location	Amount	Rel. Dev.
Alcazar	Lucas	Information Technology	IT Helpdesk	13.01.2014 19:20:00	Frydos Autosupply n' More	10000	9.35
Lais	Cornelia			08.01.2014 13:03:00	Carlyle Chemical Inc.	4983.52	1.87
Nant	Irene	Facilities	Truck Driver	08.01.2014 13:03:00	Carlyle Chemical Inc.	4983.52	1.87
Mies	Hank	Facilities	Truck Driver	14.01.2014 15:32:00	Abila Airport	4918.39	1.21
Morlun	[illegible]	Facilities	Truck Driver	08.01.2014 14:26:00	Carlyle Chemical Inc.	4901.88	1.82
Tempest	[illegible]	Engineering	Drill Technician	08.01.2014 14:26:00	Carlyle Chemical Inc.	4901.88	1.82
Lais	elia			14.01.2014	Abila Airport	4898.39	1.2
Morlun	Valeria	Facilities	Truck Driver	07.01.2014 14:34:00	Carlyle Chemical Inc.	4803.13	1.76
Cazar	Gustav	Engineering	Hydraulic Technician	09.01.2014 13:18:00	Abila Airport	4792.5	1.13
Scozzese	Dylan	Facilities	Truck Driver	09.01.2014 13:18:00	Abila Airport	4792.5	1.13
Nant	Irene	Facilities	Truck Driver	15.01.2014 15:11:00	Maximum Iron and Steel	4788.22	1.15
Onda	Marin	Engineering	Drill Site Manager	15.01.2014 15:11:00	Maximum Iron and Steel	4788.22	1.15
Scozzese	Dylan	Facilities	Truck Driver	16.01.2014 11:25:00	Nationwide Refinery	4742.67	2.3
Osvaldo	Hennie	Security	Perimeter Control	08.01.2014 09:54:00	Nationwide Refinery	4513.16	2.14
Morlun	Adan	Facilities	Truck Driver	08.01.2014 09:54:00	Nationwide Refinery	4513.16	2.14
Coginian	Dante			15.01.2014	Stewart and Sons Fabrication	4485.38	1.71
Scozzese	Dylan	Facilities	Truck Driver	14.01.2014 16:28:00	Nationwide Refinery	4429.76	2.08
Arpa	Emile			14.01.2014 16:28:00	Nationwide Refinery	4429.76	2.08
Forluniau	Carla			13.01.2014 12:11:00	Abila Airport	4417.03	0.89

Locations

Location	Median	Std. Dev.
Brew've Been Served	11.53	18.67
Carlyle Ch mical Inc.	1814.24	1696.12
Chost	113.25	173.86
Coffe	11.67	12.47
Coffee	11.51	11.57
Daily Dealz	2.01	NaN
Desafio Golf Course	150.07	33.91
Frank's Fuel	65.34	10.35
Frydos Autosupply n' More	146.74	1053.96

Show only deviating data ● Ignore ○ + ○ -
Show only data of employees ● Ignore ○ + ○ -
Show only LC data ● Ignore ○ + ○ -
Show only CC data ● Ignore ○ + ○ -
Show only data with timestamp ● Ignore ○ + ○ -
# of shown entries: 2863

Persons

Last Name	First Name	Emp. Type	Emp. Title
Alcazar	Lucas	Information Technology	IT Helpdesk
Arpa	Emile		
Awelon	Varro		
Azada	Lars	Engineering	Engineer
Balas	Felix	Engineering	Engineer
Barranco	Ingrid	Executive	SVP/CFO
Baza	Isak	Information Technology	IT Technician
Ber	Linnea	Information Technology	IT Group Manager
Bo	Loreto	Security	Site Control
Bor	Isande	Engineering	Drill Technician
Bramar	Mat		
Calixto	Nils	Information Technology	IT Technician
Calzas	Axel	Engineering	Drill Technician
Campo-Corrente	Ada	Executive	SVP/CIO
Cazar	Gustav	Engineering	Hydraulic Technician
Cocinaro	Hideki	Security	Site Control

Figure 4: *Inspection & Cleaning - Table* ⓐ *shows initially available data (here transactions),* ⓑ *shows locations and* ⓒ/ⓓ *persons. Color indicates deviation from the average expense at a location (more intense red ⇔ higher deviation).*

can filter by means of deviating values and by transactional mapping (credit and loyalty card data).

4. Orchestration of Spatial and Temporal Data

Table views, as presented in Section 3, provide basic methods for a first inspection, inepedendent from the data domain. Using this knowledge, we can then create and apply more specific visualizations, as done by Bernard et al. [BRG*12] who presented interactive preprocessing by domain experts guidance tailored to time series data. For spatio-temporal data we can use maps and timelines to visually support the preprocessing, especially data orchestration (see Figure 1, step 2). Each dataset provides another piece of the puzzle and only with all pieces in place, one can uncover the big picture, in our case, suspicious behavior patterns prior to the kidnapping. We propose the following stepwise semi-automated mapping that further aligns the data (see Figure 3) on mutual aspects such as time and position.

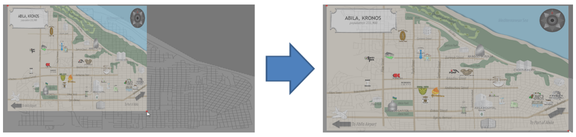


Figure 5: *Images (here a tourist map) can be loaded as map overlays and can be aligned to the underlying structure using translation & scaling.*

4.1. Map Alignment

POIs support the semantic understanding of movement reasons [PSR*13]. In the challenge, POI information is partly provided with a comic-like image of a tourist map of the city, which contains only imprecise location information. However, when combined with road and coordinate information, it can become very useful (see Figure 2b). In our system, the analyst can load images as a semi-transparent overlay and use translation and scaling features to adjust the data sources as shown in Figure 5. Finally, POI information is shown at the correct position.

4.2. Map Annotation

Map annotations are a powerful and often used technique to enable a semantic analysis [Kli08]. Our system provides an annotation tool, e.g., to define AOIs (areas of interest) based on the information found in the image-based POI tourist map. Polygons can be interactively placed on the map to cover such AOIs, as can be seen in Figure 6, step 1. The analyst can assign various features to these polygons, such as categories (e.g. *work* or *private*) and places names (2). AOI colors can be assigned based on these features and will be used accordingly in other views.

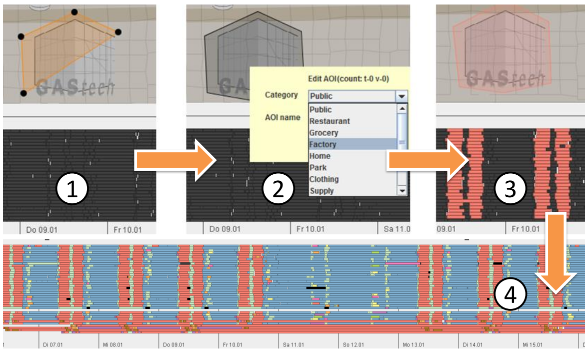


Figure 6: *Annotation Process -* ① *With a polygon tool an AOI can be created;* ② *The AOI can be annotated with name and category, e.g.* GAStech*;* ③ *Movements are enriched and colored based on their destination (here GAStech);* ④ *All AOIs are annotated / all movements enriched.*

4.3. Extraction

Complementary to the user-based annotation, frequently visited places can also be extracted automatically. The algorithm iterates over the transaction data and maps the locations covered in the billings to the employees' trip destination, according to common time intervals. This reveals additional AOIs that are not appearing in the tourist map. However, in addition to the aforementioned varying data resolution, employees might travel without using the rented cars, may pay cash, or give their cards to others. Thus, the automated extraction does not always work perfectly [KTE14, FCRS13]. The analyst can thus revise extracted information with the annotation tool again. For example, one might enlarge or shrink the registered area, assign a location category or delete erroneously extracted AOIs.

4.4. Movement Enrichment

After spatial areas have been annotated with semantically interpretable categories, labels, and colors (see Section 4.2), the system automatically enriches all trajectories (trips with the company cars) according to their destinations (visited AOIs). Hence, high resolution spatio-temporal movement information is transformed into low resolution event sequences, where an event is a particular stay at a particular AOI. Finally, the event sequences are then visualized in the Sequence View (Figure 2e), where each row represents an employee's behavior over time. In Figure 6, steps 1 & 2, the black color refers to unknown events. After the enrichment, they are automatically colored (steps 3 & 4) according to the visited AOI (e.g., GAStech, Airport, or Harbor) and AOI category (factory, restaurant, etc.).

5. Exploration

The exploration process (see Figure 1, step 3) highly depends on the previous inspection, cleaning, and alignment steps [KMS*08]. The more appropriate these tasks were done, the less uncertainties and inaccuracies will hinder the analysis.

The analyst might start the exploration by replaying the movements at different speeds, and jump to specific times (see Figure 2e). This reveals that during night time nearly all cars are parked in the north-east city area—the employees homes. To automatically detect such daily routines we also integrated an n-gram filter (see Figure 2f) that, contrary to common sequence mining algorithms [ME10], works on a per-user basis. The filter needs to be configured with two thresholds: (1) the minimal length n of a subsequence to be found, and (2) its minimal number of occurrences in the employee's full event sequence. Results are shown in the Sequence View (see Figure 2e). Thereafter, we employ TrajectoryLenses [KTW*13] as a powerful mechanism to further query the movements based on their origins and destinations while obtaining immediate feedback in the map (2d). For example, inspecting trips from the city airport one can clearly see the arrival of the GAStech CEO a few days prior to the kidnapping.

6. Hypothesis Creation & Validation

After the analyst has explored the data in time and space, she may has created various hypotheses (see Figure 1, step 4). These hypotheses can now be expressed and refined with the the Pattern Query Tool (see 2g). Its visual query language allows to define and query event sequence patterns and narrow the results with various restrictions (e.g. on an employee, the time, or location). While a variety of approaches for finding event sequences have been proposed [WPTMS12, FKSS06], they do not focus on parallel and overlapping event sequences of several actors. For instance, ActiviTree [VJC09] finds similarities in event sequences of different users, but does not establish relationships between the actual events.

Any actual sequences of events matching the hypothesized event pattern can be automatically detected and listed in the Sequence View (Figure 2e). Based on this filtered list of events, it is possible to iteratively refine the hypotheses. In the example shown in Figure 7, the event query pattern describes any situation in which the CEO of GAStech, identified by his name *Sanjorge*, meets another person at a place outside of the GAStech building. By applying this filter, it becomes apparent that various persons met Mr. Sanjorge during his stay, including a meeting with other executives at the weekend, at the golf course.

The pattern depicted in Figure 2g filters any occurrences of a common daily routine, i.e. *coffee* ▶ *work* ▶ *lunch* ▶ *work* ▶ *home*. The query matches the daily routines of almost all employees. Likewise, the inverse results reveal any divergent behavior, which might be suspicious. For example, employees who barely take a lunch break or stay at work until late at night can be recognized.

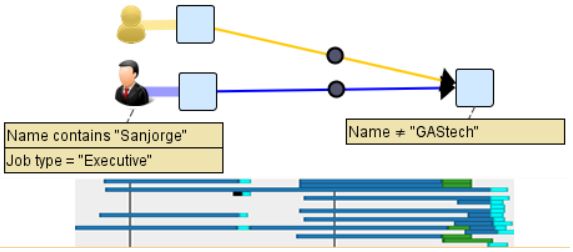


Figure 7: *The event sequence pattern (top) filters for events where GAStech's CEO (Sanjorge) meets somebody else outside the company building of GAStech. The sequence view (bottom) visualizes the results, revealing that the CEO meets people at restaurants (cyan) and at the golf course (green).*

7. Conclusion and Future Work

We presented analysis approaches for heterogeneous spatio-temporal data sources. In comparison to state-of-the-art solutions, our tools visually support the analyst not only in the exploration and sensemaking process, but also during the first data inspection, pre-filtering, and orchestration. We proposed a semi-automated approach that becomes extremely helpful when the mapping features are imprecise.While we presented the approach with the challenge data, we are confident that the developed ideas are suitable to other domains and tasks. For example, the AOI annotation & extraction techniques could be helpful for urban planning and eye-tracking evaluation tasks, to support semantic analysis. In the future, we want to apply our methods to real world data and extend the evaluation with a user study.

References

[BCVB01] Bergamaschi S., Castano S., Vincini M., Beneventano D.: Semantic integration of heterogeneous information sources. *Data & Knowledge Engineering 36*, 3 (2001), 215–249. doi:10.1016/S0169-023X(00)00047-1. 1

[BRG*12] Bernard J., Ruppert T., Goroll O., May T., Kohlhammer J.: Visual-interactive preprocessing of time series data. In *Proceedings of SIGRAD 2012: Interactive Visual Analysis of Data* (2012), vol. 81 of *Linköping Electronic Conference Proceedings*, Linköping University Electronic Press, pp. 39–48. doi:10.1057/ivs.2009.13. 3

[DH05] Doan A., Halevy A. Y.: Semantic-integration research in the database community. *AI Magazine 26*, 1 (2005), 83–94. doi:10.1016/S0169-023X(00)00047-1. 1

[FCRS13] Furletti B., Cintia P., Renso C., Spinsanti L.: Inferring human activities from gps tracks. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing* (2013), ACM, pp. 5–12. doi:10.1145/2505821.2505830. 4

[FKSS06] Fails J., Karlson A., Shahamat L., Shneiderman B.: A visual interface for multivariate temporal data: Finding patterns of events across multiple histories. In *VAST '06* (2006), IEEE, pp. 167–174. doi:10.1109/VAST.2006.261421. 4

[HS95] Hernández M. A., Stolfo S. J.: The merge/purge problem for large databases. In *SIGMOD Record* (1995), vol. 24, ACM, pp. 127–138. doi:10.1145/568271.223807. 1

[HS98] Hernández M. A., Stolfo S. J.: Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery 2*, 1 (1998), 9–37. 1

[Joh11] Johnson C.: Visual computing, 2011. http://sc12.supercomputing.org/sites/default/files/SC12PM2.pdf, based on Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet. 1

[Kli08] Klien E. M.: *Semantic annotation of geographic information*. PhD thesis, University of Münster, 2008. 3

[KMS*08] Keim D. A., Mansmann F., Schneidewind J., Thomas J., Ziegler H.: *Visual analytics: Scope and challenges*. Springer, 2008. doi:10.1007/978-3-540-71080-6_6. 4

[KTE14] Krüger R., Thom D., Ertl T.: Semantic enrichment of movement behavior with foursquare—a visual analytics approach. *IEEE Transactions on Visualization and Computer Graphics* (2014). accepted for publication, volume and number to be announced. doi:10.1109/TVCG.2014.2371856. 4

[KTW*13] Krüger R., Thom D., Wörner M., Bosch H., Ertl T.: TrajectoryLenses—A set-based filtering and exploration technique for long-term trajectory data. *Computer Graphics Forum 32*, 3 (2013), 451–460. doi:10.1111/cgf.12132. 4

[ME10] Mabroukeh N. R., Ezeife C. I.: A taxonomy of sequential pattern mining algorithms. *ACM Computing Surveys 43*, 1 (2010), 3:1–3:41. doi:10.1145/1824795.1824798. 4

[NH02] Naumann F., Häussler M.: Declarative data merging with conflict resolution. In *Proceedings of the Seventh International Conference on Information Quality (ICIQ-02)* (2002), University of Arkansas Little Rock, pp. 212–224. 1

[PC05] Pirolli P., Card S.: The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (2005), vol. 5, Mitre. 1

[PSR*13] Parent C., Spaccapietra S., Renso C., Andrienko G., Andrienko N., Bogorny V., Damiani M. L., Gkoulalas-Divanis A., Macedo J., Pelekis N., Theodoridis Y., Yan Z.: Semantic trajectories modeling and analysis. *ACM Computing Surveys 45*, 4 (2013), 42:1–42:32. doi:10.1145/2501654.2501656. 3

[Pv09] Pretorius A. J., van Wijk J. J.: What does the user want to see? What do the data want to be? *Information Visualization 8*, 3 (2009), 153–166. doi:10.1057/ivs.2009.13. 1

[RC94] Rao R., Card S. K.: The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (1994), ACM, pp. 318–322. doi:10.1145/191666.191776. 2

[RH01] Raman V., Hellerstein J. M.: Potter's wheel: An interactive data cleaning system. In *Proceedings of 27th International Conference on Very Large Data Bases* (2001), vol. 1, Morgan Kaufmann, pp. 381–390. 1

[SBB96] Spenke M., Beilken C., Berlage T.: FOCUS: The interactive table for product comparison and selection. In *Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology* (1996), ACM, pp. 41–50. doi:10.1145/237091.237097. 2

[SSS*14] Sacha D., Stoffel A., Stoffel F., Kwon B. C., Ellis G., Keim D.: Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1604–1613. doi:10.1109/TVCG.2014.2346481. 1

[SWPG12] Scholtz J., Whiting M. A., Plaisant C., Grinstein G.: A reflection on seven years of the vast challenge. In *Proceedings of the 2012 BELIV Workshop* (2012), ACM, pp. 13:1–13:8. doi:10.1145/2442576.2442589. 1

[TC05] Thomas J. J., Cook K. A. (Eds.): *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr, 2005. 1

[TSWB94] Tweedie L., Spence B., Williams D., Bhogal R.: The attribute explorer. In *Conference Companion on Human Factors in Computing Systems* (1994), ACM, pp. 435–436. doi:10.1145/259963.260433. 2

[Tuf91] Tufte E. R.: Envisioning information. *Optometry & Vision Science 68*, 4 (1991), 322–324. doi:10.1097/00006324-199104000-00013. 2

[vas14] VAST Challenge 2014: Mini-challenge 2, 2014. URL: http://www.vacommunity.org/VAST+Challenge+2014%3A+Mini-Challenge+2. 1

[VJC09] Vrotsou K., Johansson J., Cooper M.: ActiviTree: Interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 945–952. doi:10.1109/TVCG.2009.117. 4

[WPTMS12] Wongsuphasawat K., Plaisant C., Taieb-Maimon M., Shneiderman B.: Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting with Computers 24*, 2 (2012), 55–68. doi:10.1016/j.intcom.2012.01.003. 4