Abstract. In this experience paper we study the boundaries of the usability of the SynerScope visual analytics tool suite by confronting secondary school students with typical analytics questions on Wikipedia and DBpedia after minimal preparation. We study the learning curve of students and preliminary results indicate that after a short tutorial some students are able to answer most complex questions in a few minutes, learning by trial and error, while others struggle with the simplest questions. Older students are faster on average, but motivation appears to be a stronger factor for success than age. Answering speed doubles within two hours of experience while correctness increases.

1 Introduction

The world will soon face a critical shortage of data scientists, professionals with analytical expertise that can take advantage of (linked) data to answer questions [7]. One strategy to mitigate this problem is to enable non-experts to take over part of the data science tasks. We pose that data science is comprised of many tasks that do not all require expert-level knowledge. In this article we restrict ourselves to a category of data science sensemaking tasks on Web data that is common in data journalism and involves basic analytics operations, search, and Web browsing. We hypothesise that, given the right tools, motivated untrained people can quickly be trained to do such tasks, avoiding a complete data science education.

The goal of this article is to gain first insights that will eventually allow us to test this hypothesis in future work. We perform an experiment to investigate the feasibility of having untrained people do prototypical sensemaking tasks given visual analytics tools. Specifically, we look at secondary school students with no analytical experience, and ask them to answer complex questions about Wikipedia content using the SynerScope4 visual analytics software illustrated in Figure 1. We want to know if users can get to an answer after a minimal amount of training in the tool. We want to know how long it takes them to find an answer and if their time-to-answer decreases as their experience with the tool increases, and what the influence is of their age and corresponding level of education.

4 http://www.synerscope.com
The line of reasoning we follow is that the required skills for such sensemaking data science tasks can be rapidly acquired or substituted with appropriate tools. If this is the case and if SynerScope is an appropriate tool for the task, then we should be able to show that unskilled people can accomplish the sensemaking tasks.

This idea of empowering people by means of augmented reasoning through human-computer interaction is not new [6], but in recent years the development of interactive tools and frameworks for visual analytics have intensified. Some of these tools are targeted at programmers (e.g., [1, 10, 11]), while other tools target non-programmers (e.g., [2, 8, 9, 14, 12, 13]). For this experiment we need a tool from the latter category that is network centric and allows search and Web browsing. We use SynerScope [4, 5, 12], one of the tools that meets these requirements.

Fig. 1. A screenshot of the SynerScope visual analytics tool showing Wikipedia and DBpedia data. This picture shows two coordinated views: a hierarchical edge bundling network view and a Web browser.

The rest of this paper is organised as follows: Section 2 describes the SynerScope software in more detail. Section 3 outlines the experimental set-up, including the tasks, tooling, and procedure. Section 4 shows our findings. Section 5 discusses our findings, draws conclusions and suggests future work.

2 The SynerScope Software

SynerScope is a visual analytics application that delivers real time interaction with dynamic network-centric data. SynerScope supports simultaneous visualisations and coor-
Fig. 2. A screenshot of the same data shown in Figure 1, after drilling down to the selected set of links and nodes.

dinates user interaction, enabling the user to identify causal relationships and to uncover unforeseen connections.

The central interaction paradigm of SynerScope is Multiple and Coordinated Views. SynerScope shows a number of different perspectives on data, for example, relations and time, and each selection made in either of these views causes an equivalent selection to be made in all other views. This enables the user to explore correlations between different facets of data.

SynerScope is designed to work with a very basic information schema. This schema consists of two object types: Nodes and Links. Links connect two Nodes. Both Nodes and Links can have additional attributes of a number of data types, including integers, floating point numbers, free text, date and time, latitude and longitude.

What follows is a short overview of each visualisation that is offered by SynerScope.

Table View  The Table View (see Fig 3) provides a traditional spreadsheet view on the data. For each type of Node and each type of Link, there is a separate sheet. The Table View shows all the data as a table of values.

Hierarchical Edge Bundling View  The Hierarchical Edge Bundling View (HEB) (see Fig 1, 2, and 3) is the primary network view in SynerScope. Each Node is visualised as a point on a circle, and each Link is visualised as a curved line between its source and target Node.

The Nodes are grouped hierarchically, based on one or more of their attributes. The Links between Nodes of the same hierarchical category are bundled together (as if they were tied together with a cable tie).
Fig. 3. A screenshot of the SynerScope visual analytics tool showing Wikipedia and DBpedia data. This picture shows the hierarchical edge bundling view (top left), the massive sequence view (top right), the table view (bottom left), and the word square view (bottom right).

**Massive Sequence View** The Massive Sequence View (MSV) (see Fig 3) is the primary temporal view in SynerScope. Each Node gets a fixed position on the horizontal axis. Nodes are grouped hierarchically in the same fashion as in the HEB. Links between Nodes are represented by a horizontal line between the respective positions of the Nodes. On the vertical axis the user can select a scalar attribute, typically a time or date. This orders the Links temporally.

**Map View** The Map View is the primary spatial view in SynerScope. The user can select two attributes from any Node or Link data source to interpret as latitude and longitude coordinates. These attributes are used to plot the Nodes (not the Links) on a map as points.

**Scatter Plot View** The Scatter Plot View uses Cartesian coordinates to relate the values of two attributes of either Nodes or Links. Dots are drawn on a two-dimensional chart, the positioning relative to the horizontal and vertical axis being determined by the attribute’s values. A third attribute can used to set the size of the dots.

**Search and Filter View** The Search and Filter View is an interactive view that allows the user to select Nodes or Links by searching by value.

**Web View** The Web View (see Fig 1 and 2) is an interactive view that allows the user to view any URL’s that are an attribute of a node or a link.
The user can interact with SynerScope’s views in several ways: By selecting and highlighting data, drilling down to or up from a selection, and expanding selections from nodes to connected links or vice versa. Every interaction method is coordinated across multiple views.

3 Experimental Set-up

Sensemaking Tasks In the experiment we look at 10 exemplar analytics questions that each require a combination of at least two of the following operations to answer: network navigation, filtering on categorical and numerical variables, grouping and counting, search, Web browsing within Wikipedia, and zooming in on data selections. Examples of the questions are: “How many former AFC Ajax soccer players died in Paramaribo and what was the cause of death?”, or “Which page about a disease is linked to most from pages about physicists?”. The complete set of questions can be found on FigShare [15].

SynerScope Visual Analytics Tooling The SynerScope tool used by the students is a graphically accelerated visual analytics application that combines a number of views on networked data. It offers real-time interactive exploration using scatter plots, timelines, maps, hierarchical edge bundling network layouts, an integrated Web browser, a search engine, and a spreadsheet table view. The selections made in any of these views are propagated to all the other views. A video illustrating interaction with the Wikipedia data can be found on FigShare [3].

Procedure The experiment consists of five parts: (1) a 30m plenary introduction to the experiment and the data sets used, (2) a 15m plenary tutorial to the SynerScope visual analytics tool, (3 and 4) two 45m sessions where students try to answer questions using SynerScope, (5) a concluding discussion and personal interviews. The students are asked to answer as many as possible of 10 questions about 3 subsets of Wikipedia within 90m. Each set centers around pages on a specific topic.

Data Sets The topics covered in the experiment are: (1) Athletes classified as soccer players and trainers of AFC Ajax, FC Barcelona, and Manchester United, (2) Scientists classified as physicist, (3) Artists in the pop genre. Each of these three sets consist of around 3000 Wikipedia pages about the topic (the seed set), all the pages that are linked to from the seed pages (the “out” context), all pages that link to the seed pages (the “in” context), and all the links between the seed, “out” context, and “in” context pages. This amounts to three sets of around 100k–200k pages and 300k–500k page links. Each page is assigned around 18 attributes with information about the page, such as the page title, the number of words on the page, the in degree and out degree, a three-level hierarchical topic classification of the main subject of the page (e.g. Actor-Artist-Person, or Building-ArchitecturalStructure-Place) derived from the DBpedia rdf:type property of the corresponding DBpedia resource, birth/death date and place, and topic-specific properties such as respectively soccer team, university, or band. An example of the three schemas can be found in the hand-outs for the students [15].
Test Subjects  The students involved as test subjects in the experiment are 34 middle school and high school students from three schools in the Amsterdam area between the ages of 12 and 18. The experiments were performed in two labs of the VU University Amsterdam Network Institute. The students were paired up and given a hand-out describing the three data sets, listing all the questions, and containing a form to record the answers and the time taken [15]. During the experiment students were assisted by answering specific technical questions, but were given no other guidance that would help them find answers.

4 Results

There was a large variation in the productivity of the various students, as can be seen in Figure 4. This can be expected of students that have no intrinsic motivation to cooperate in the experiment. The motivated students answered all questions, while two groups did nothing and are excluded from the results. In general the total number of 10 questions was too high to answer for most students in two 45 minute sessions. Most students managed to answer the questions of two topics (6 or 7 questions). Of the questions that were answered, about 60% was answered correctly. There was a large variation, depending on the difficulty of the question. This is illustrated in Figure 5. Some questions were answered partially. For example, when asked for a number and explanation only the number or the explanation was answered correctly. There was a slight increase in the number of questions that were answered correctly over time, but given the small number of questions this increase is not significant. Students performed faster and more consistently for subsequent questions. This is illustrated in Figure 6 (right), specifically with questions 1–7 which were consistently answered before time ran out. This increase in speed is significant between the first and last of the questions in the sequence at a confidence level of 95%. Older students were significantly faster than younger students, but their answers were of a comparable correctness. An exception to this rule was a class of particularly talented middle school freshmen that performed on par with 18-year-olds, but with a significantly higher accuracy. The data used to derive these conclusions can be found on Figshare [15].

5 Discussion

We are impressed by what some of our young test subjects were able to achieve. When given the right tools and the right motivation, visual analytics of linked data really can be done by secondary school students with minimal preparation. We found that after a short tutorial these students are able to answer most complex questions in a few minutes, learning by trial and error. Within two hours of experience, answering speed doubles within while correctness increases. On the other hand, we found that some students did not achieve much. These students were those that had no interest in the experiment and hence lacked intrinsic motivation.

Fig. 4. An overview of the completeness and correctness of the answer to each question by each student. Each column represents the answers given by a student. Each row represents one of the 10 questions. Roughly 55% of the questions were answered, about 60% of the answers were correct.

Fig. 5. An aggregation of the correctness of the answers per question. For the questions that were consistently answered (1–7) in the 90m experiment there appears to be a rising trend in the quality of the answers. Most students ran out of time before attempting question 8–10.
The older test subjects more frequently asked for help when they get stuck than the younger test subjects, who just found their own way through trial and error, and therefore also take longer to get to an answer than the older students (as can be seen in Section 6. Overall, motivation appears to be a stronger success factor than age. This belief is hard to make concrete, but it is reinforced by our observation that students are quick to accept their first findings as a definitive answer to the question they were working on. When students found information they thought was the right answer, they were fairly quick to accept that answer and wanted to move on to the next question as soon as possible. In contrast to professionals, the students did not verify their answers. For instance, when the students had to find out how many AFX Ajax soccer players died in Paramaribo, they typically accepted all the soccer players that died in Paramaribo as an answer, without checking if they played in AFC Ajax. We think this can be explained by the lack of feedback during the experiment. Students were not penalised for wrong answers or rewarded for right answers, and the experiment was a one time encounter with the software. We expect that many of the incorrect or partial answers could have been improved if the students were to have verified their answers.

The experiment reinforced our belief that visual analytics software must be highly interactive and present immediate feedback to the user. During the interviews at the end of the session students were generally positive about the software and tasks and thought the experiment gave them a new perspective on Wikipedia. In isolated cases, for example, when zooming out to the entire data set of 400k links, students had to wait a few seconds. Delays in interaction like these appeared to interrupt the student’s train of thought.

We found that motivated students, regardless of age, are able to effectively use the SynerScope tool to answer the questions. Older students are usually faster, but not sig-
significantly more accurate. We would like to further test these findings with older and younger subjects.

Acknowledgements

Thanks go to the Damstede, Pieter Nieuwland College, and Cygnus Gymnasium schools for their participation in this experiment. We thank the VU Network Institute for the use of their facilities, and Samir Naaimi for his assistance during the experiments. This work was done within the context of the SAGAN project supported by ONR Global NICOP grant N62909-14-1-N030, the EU FP7 NewsReader project (316404), and the Dutch COMMIT Data2Semantics project.

References