Visualization of Warehouse Pick Lists: a Case Study at Apotea AB

PER JAAKONANTTI
Visualization of Warehouse Pick Lists: a Case Study at Apotea AB

Per Jaakonantti
Student
Royal Institute of Technology
perja@kth.se

ABSTRACT
Today e-commerce retailers use large warehouses to store goods before they are sold and shipped. One of the main performance issues for Apotea AB, an online pharmacy based in Sweden, is that the staffers who walk around in the warehouse often are delayed in queues or clusters, usually as a result of missing items or difficulties finding items. This paper describes how, based on sparse data, a map-based interactive visualization was created to aid staffers in their task of evaluating how so called pick lists perform. The visualization was evaluated using UMUX-LITE [20] and a grounded model of NOvice’s information VIsualization Sensemaking (NOVIS-model) [17]. Results showed an overall successful design. The main affordance of the visualization was that it showed warehouse performance over time when animation speed was set to maximum, this was enabled by sliders and buttons. The main limitations were found to be a lack of intermediary level of visualization (e.g. heat map overview of an hour) in the visualization and that there was no higher level of data, e.g. incident data, which could be visualized. These limitations represent the core of what this paper identifies as future work.

Author Keywords
Visualization; information visualization; D3.js; interactive information visualization.

ACM Classification Keywords
• Human-centred computing → Visualization • Human-centred computing → Visualization → Visualization application domains → Geographic visualization

INTRODUCTION
In the 21st century online sales are a substantial share of the consumer market and there are several well-known national and international companies that sell products online, circumventing the costs of a store and providing the comfort of not having to leave the house to shop. A common practise in this business sector is to store the products in large warehouses with varying solutions regarding logistics. In this paper the term warehouse refers to a commercial building used for storage of goods. Challenges for maintaining a warehouse used for e-commerce companies have been documented. The main ones include solving problems that arise from seasonal changes in consumer behaviour, processing single-item orders as well as having an accurate and on-time delivery [25]. It is important to solve these issues for logistical purposes but also for customer loyalty.

At the warehouse analyzed in this paper, there are staffers who walk around picking up products, placing them in boxes on a cart. They have a list of items that they are to pick (based on a structured pick list). The person picking only sees which item they should pick and where it is located. As a data point, each pick list contains the following data: shelf number, time and expected time (used to calculate delays) for each item that has been picked, as well as a unique id. The gathered data is sparse, meaning that the exact path from one shelf to another is unknown (although the aisles are one-directional), in a similar way as sparse traffic trajectory data [34].

For staffers that are picking orders, it is relatively common to have to wait for the person or persons standing in front of them or maybe an item they were supposed to pick is not found. Clearly this is not ideal in terms of performance as staffers may end up looking for an item or waiting, occasionally several minutes, slowing the whole process down. Waiting is a delay and reduces the overall performance of the warehouse.

At Apotea AB the performance of the pick lists is evaluated by gathering data and calculating metrics related to time, orders completed and articles picked. However there has been no way of evaluating the pick lists outside the scope of looking at tables, numbers and occasionally text. In this paper, the pick lists for carts in the warehouse are visualized in a way to aid staff in their task of understanding how the warehouse is used, or to be more precise, how the pick lists perform. The visualization was map-based, in 2D, as warehouse movements are mainly 2-dimensional, as seen in Figure 1 and 2. The warehouse also has an elevator seen in Figure 3 and 4.

Information visualization is a research area where data is presented on a visual interface. It is more precisely defined as “the use of computer-supported, interactive visual representations of data to amplify cognition.” [7]. In this paper, interactive refers to the user being able to interact with the visualization through clicking buttons, dragging on sliders and hovering over items of interest. In this sense, visualizations are a type of digital environment
in which humans can interact with the interface. Visualizations afford certain actions to be taken through interaction, commonly referred to as affordances [22], actions not afforded are a type of limitation. As all types of software, visualizations can be evaluated, determining how useful and how easy to use they are. User satisfaction can be measured [20] and it is possible to examine how new users make sense of unfamiliar visualizations [17].

**Research Question**

RQ: What are the main affordances and limitations of map-based interactive visualizations representing warehouse pick lists used by an e-commerce retailer in supporting IT-staff in their task of quantifying performance of pick lists as measured by user satisfaction and how well the visualization is understood?

**THEORY AND RELATED RESEARCH**

**Affordance**

Within the HCI community there are many ways to view and define the term affordance [22]. Vyas et. al suggest that affordances can be understood as “an interpretive relationship between users and the artefact” [32]. In this paper a mediated action perspective approach towards the term affordance is adopted, where: “Affordances are understood as [...] possibilities for human actions in cultural environments” and “Affordances emerge in interaction between [...] the person, tool(s) and cultural environment” [16].

**Map-based interactive visualizations**

There are some previous approaches to visualizing data gathered in a warehouse. These include creating 3D models of how the shelves are arranged and where certain items are [15] as well as showing a bird’s eye view of a warehouse where colours indicate frequency of picking at each shelf [3]. Pick lists themselves have however not been visualized previously.

Walking around in a warehouse is a type of traffic. As such traffic visualization is a related area to this paper. Visualizing traffic data is a well-researched area where different types of data are visualized in a variety of ways [9]. Despite sometimes only having sparse datasets available, it has been possible to visualize, for example, congestions in traffic [34]. Other times a large amount of data is gathered and it has to be cleaned heavily before presenting it to analyze traffic-jams [33]. There are many methods to visualize traffic data, including heat maps [11, 29, 36]. Most traffic data visualizations are map-based [2, 11, 29, 33, 34, 36] and some used a timeline as well [31]. Visualizations can be a result of many methods to
visualize lots of data at once, however, it has been suggested that simple visualization methods of highlighting traffic incidents can easily be understood by a broad audience without much explanation [2].

**Warehouse terminology**

The following quote has been used to define terms relating to aisles in the warehouse: “The warehouse is rectangular with no unused space and consists of a number of parallel pick aisles. The warehouse is divided into a number of blocks, each of which contains a number of sub aisles. A sub aisle is that part of a pick aisle that is within one block. /.../ At the front and back of the warehouse and between each pair of blocks, there is a cross aisle. Cross aisles do not contain storage locations but can be used to change aisles.” [26]

**User satisfaction**

In this paper user satisfaction is measured by using a UMUX-LITE based questionnaire. UMUX-LITE is “a two item questionnaire based on the Usability metric for User Experience (UMUX)” [20], UMUX is a four item questionnaire [12]. UMUX-LITE has received some criticism. However, evidence for its reliability, sensitivity and concurrent validity has been found [20]. Moreover, factors like participants’ native language, age and gender have been identified as not being sensitivity issues for the UMUX-LITE scale [5]. UMUX-LITE has been found to appear as closer in magnitude to the more well-known SUS [6] compared to UMUX [4]. Researchers have found that the overall mean difference between regression-adjusted UMUX-LITE scores and the SUS scores is at 1.1 (or roughly 1 % of the range) [21]. The main advantage of UMUX-LITE is that it is much shorter, saving time during testing, UMUX-LITE has been recommended as the best choice to use when there is a need to use few items [19].

**Users understanding of visualizations**

Basing an analysis on easily measured metrics like time completion for complex tasks has received some criticism, as these metrics do not give deeper insight to how a complex visualization is understood. Instead there has been a call for a focus on insight based on open-ended protocols, qualitative insight analysis and an emphasis on domain relevance [24].

Since some methods of evaluating information visualization have been undocumented, UX-patterns have been documented for all parts of the process of designing a study and a prototype, examples include Pilot Study and Do-It-Yourself [10]. One way of analysing data is to conduct a NOVIS analysis, by conducting think aloud evaluations and dividing expressed thoughts into predefined activities [17]. NOVIS was created to evaluate information visualizations specifically [17].

**METHOD**

In this study two Pilot Studies [10] were held in order to test the methods and visualization, and after that the final study was conducted. Another pattern based method used in this study was Do-It-Yourself, used to find bugs in the visualization before it was tested in the pilots [10].

**Study and Pilot Study**

During the first pilot study there were four participants who tested the visualization during a think aloud session and then gave their feedback. During the second pilot study there were six participants who were recorded (audio) while participating in a think aloud session and then interviewed. The participants in the pilot studies were students and the testing sessions were held at the main Campus of The Royal Institute of Technology in Stockholm, whereas the final study was conducted with 10 IT-staffers at Apoteka AB at their office in Stockholm. All participants in the final study had been in the warehouse and were deemed likely to at some point work with issues relating to the pick lists and the performance of the warehouse. Picking this broad definition for relevant testers was intended to result in more testers and more reliable quantitative data. It also represents the intended target group by including staffers who potentially will create pick lists in the future and not just those who work with it right now.

All participants answered a UMUX-LITE questionnaire containing the two statements: “The system capabilities meet my requirements” and “The system is easy to use” the users had to pick on a 7-point scale from 1 (strongly disagree) to 7 (strongly agree) [20]. 7-point is default for UMUX-LITE and using 7 ± 2 has been found to be very suitable for scales in general [14], moreover it has been found that the exact number of points on a scale does not change the final score in terms of user experience [18]. To normalize usefulness and usability scores percentile ranks were calculated using a method presented by Jeff Sauro [27]. UMUX-LITE was used to measure user satisfaction in order to answer RQ. Scores were calculated in order to enable comparison to other pieces of software that were also measured using UMUX-LITE and percentile ranks. Scores were also calculated to an SUS-equivalent score.

This regression formula was used to calculate SUS-equivalent scores based on UMUX-LITE scores:

\[ \text{UMUX-LITE} = \frac{.65(\text{Item1 score}) + \text{Item2 score} - 2)}{100/12} + 22.9 \]  [20]

**Analysis method**

In order to analyze data from the think alouds, a NOVIS (Novice’s Information Visualization Sense making) analysis was conducted. In a NOVIS analysis the expressed thoughts are divided into 6 types of activities [17]:

1. Encountering Visualization (looks at visualization as whole image)
   - Participant describes colour, shapes and other visual attributes as they see.
2. Constructing a Frame (making sense of the visualization)
   - Participant makes a connection. They understand meaning of colour, say that they see shelves, understand that a circle represents a person/cart/pick list or they understand what the heat map shows.
3. Exploring Visualization (discovers insights and facts based on created frame through interaction)
   - Gets some insight, is able to see a current cluster in the warehouse, maybe they make a connection to how frequently an area is visited to delays. They relate what they see to personal experiences. They find a problem with how the warehouse is used and/or how the pick lists perform.
4. Questioning the Frame (doubting frame that was created)
   - Unsure what they see, vocalizes uncertainty e.g. “I am not sure what I see”.
5. Floundering on Visualization (user does not know what to do, user failed to construct any reasonable frame of the visualization)
   - Participant is lost, not just uncertain. They vocalize that they have no idea what they see/what is happening e.g. “What is that?” or “I do not know, what is happening?”
6. Miscellaneous (other activities, from here referred to with an ‘M’).
   - Could be giving feedback or mentioning that it perhaps is difficult to hover over an item.

**Visualization methods**

In order to explore what the affordances and limitations for a map-based interactive visualization, a visualization had to be created. Software used to format data include Python (version 3.5.2). Tools used to visualize data include Python (version 3.5.2), JavaScript, HTML, CSS and D3.js (version 4). Data from Apotea AB was exported as a csv-file (Comma-Separated Values). Software was selected based on personal experience as well as availability to online resources describing how the software should be used.

The aim of the visualization was to first present an overview and then provide detailed data. The visualization was made up of three parts: shelves in the warehouse (seen in Figure 5-7), simulation of carts moving in the warehouse (seen in Figure 6-7) and a heat map of picking list data aggregated over the course of a day (seen in Figure 5). As the data is sparse, the cart positions were interpolated linearly based on warehouse layout. Due to aisles being unidirectional, carts frequently have to walk through an aisle without picking anything. Sometimes the staffers could choose which ones to go through, in these situations they were assumed to have moved through the next aisle in the warehouse. Cross-aisles were ignored, to simplify interpolations. Same interpolations were used in the heat map, where the delay in percent was aggregated geographically; if a journey from position A to position B was estimated to take 10 seconds but it took 20 seconds that is a delay of 100 %. Extreme delays of 600+ % were removed, as they were most likely due to staff taking breaks or other extreme delays that the company already knew about.

The heat map and the carts were assigned different colours based on delay, green indicated that carts were on time, yellow indicated small delays and red indicated large delays. Furthermore, delay was also mapped to the size of the cart, to enable colour blind users to identify late carts. Colour and size have been found to be undoubted guiding attributes, and were chosen instead of less probable attributes like tilt and shape.

The shelves were coloured in blue as it was an unused part of the colour spectrum and had no elements of green and red in them, to distinguish them from the heat map and carts. Shelves have been visualized based on frequency of visits before [3]. The map was abstract as more realistic maps have not been found increase performance [13]. The heat map was shown first, in accordance with information visualization theory to present an overview first [30].

Based on the pilot studies the following things were added: speed multiplier as a slider, a frame separating the warehouse from the rest of the screen and the heat map. A button was also added, enabling selection of current carts in the warehouse, as users in the pilot study were reluctant to try to click the carts.

In the final visualization the shelves were white if they were visited 0-10 times, light blue if they were visited 11-50 times, blue if they were visited 51-100 times and dark blue if they were visited more than 101 times. Threshold values were chosen to give an even spread of colours.

In the visualization, carts were updated 10 times per second, as that was found, through testing, to be optimal, updating them more frequently would slow down the visualization, whereas updating them less frequently would result in a visually less pleasing and lagging interface. In order to maximize performance, a cart’s position was only calculated if the current time was after the start time of the picking list (minus 10 seconds) and before the end of the pick list (plus 10 seconds). The ±10 second window was necessary as the biggest update window for the visualization was set as 10 seconds (updated every 0.1 seconds with a speed multiplier of 100), removing the 10 second window would result in inactive carts remaining in the visualization after they are done picking. At button clicks and slider use, all carts were updated, as dragging on the time slider otherwise would cause inactive carts to be shown. Inactive carts’ opacity was set to 0 % and they were moved outside of the screen. In total there were 809 carts on the day the data was from.
Figure 5. Here the warehouse is shown, shelves are represented as white to blue squares based on how often they are visited in a day. Carts are represented as green, yellow to red circles, based on their delay.

Figure 6. This figure shows what the visualization looks like when a cart has been selected. Shelves a cart has visited and will visit are highlighted based on delay.

Figure 7. Here the heat map is shown. Red indicates a bad performance whereas green indicates few and small delays.
RESULTS
In the final study there were 10 participants (9 male and 1 female) all working in the IT-team at Apotea AB. In the first pilot study there were 4 participants (2 male and 2 female) and in the second round there were 6 (3 male and 3 female).

NOVIS Results
In Table 1 below the results from the NOVIS analysis [17] are presented. It shows that most participants (7 out of 10) started to create visual frames of the interface immediately. Only 5 out of 10 participants vocalized their encounter of the visualization (NOVIS activity 1), and interestingly this was not always done immediately. Participants tended to understand that they saw shelves and voiced that they saw shelves, instead of squares. However when they pressed play and encountered the carts (depicted as circles), 5 out of 10 participants said that they saw circles, spots or dots, before later understanding that these represented carts. As participant 301 put it: “Now we have some spots, which are moving, with different sizes and colours”, others pointed out that the circles were green: “Now there are green dots that are moving, it has to be operators who are picking.” (Participant 304, translated). The last example shows that they encountered the visualization (NOVIS activity 1) and then constructed a frame (NOVIS activity 2) by saying that they connect the circle to an operator.

In this study all participants questioned the frame at some point, NOVIS activity 4. The main things participants questioned were what delay was (4 out of 10), what the heat map showed (4 out of 10), what was shown when a cart was selected (3 out of 10) and size and colour of cart (2 out of 10). Participant 304 questioned the frame at 8 times, here are some quotes of what they said: “‘Cart delay’ what could that be? Could it be how much you are over expected time or?” , “Now I pressed on ‘show heat map’ to see what happens. Then it stops. I’m still wondering a bit what the heat map shows, is it a summary for the day perhaps?” and when a cart was selected: “Oh! Wait a minute ‘active carts’, so you see the others too? /.../ I thought I would see the one I chose, but I also see the other ones in the background. The one I chose, it is that one’s shelves I see, I guess?”

6 out of 10 participants fell into the floundering activity and sometimes more than once, highlighted in red in Table 1. More than 3 participants were unclear about how to deselect a cart, but 3 clearly vocalized that they really had no idea how to do it and thus got stuck. For 2 participants it was unclear what cart delay meant. Participant 301 encountered a cart which had an extreme delay (due to expected pick time being set as January 1st 1950, this happens when a there is a missing item and a new item is added into the pick list) and they did not understand what had happened even after it was explained to them. Participants 302 and 310 initially did not understand what the big circles meant. Participant 303 did not know what active in active carts referred to and clearly stated that they had no idea what the analyzing of shelves indicated.

As seen in Table 2 the participants very frequently transitioned from constructing a frame to constructing another frame, this was the most common transition and happened 47 times. It was also relatively common to transition from constructing a frame to question the frame and to transition from questioning the frame to creating a frame. In Table 2 the incorrect frames have not been accounted for. As shown in Table 1 participant 307 started by constructing 7 different frames after each other, they created the following frames (same terminology is used in Figure 8):

1. Sees shelves.
2. What the heat map shows.
3. Shelf colour.
4. Time slider (incorrect, they thought heat map was connected to time slider).

![Table 1. Activities identified for the 10 participants. Color coded for each NOVIS activity, yellow for 1, blue for 2, green for 3, purple for 4, red for 5 and grey for M.](image-url)
5. Speed multiplier.
7. Active cart list.

There were in total 13 different frames that were vocalized by the participants. The number of different frames that were possible to create could have made transitioning from creating one frame to creating another frame the most common type of transition of all the NOVIS activities.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>47</td>
<td>16</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>7</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>12</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>M</td>
<td>0</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2. Showing the number of transitions from one activity (left bar) to the next (top bar). Transition here means that a participant stopped to express e.g. that carts are coloured by delay (2) and started to express something new, e.g. giving suggestion to improvements (M).

Figure 8 shows the error rate for creating frames for different aspects of the visualizations. No participant was unclear about how the speed multiplier worked whereas 3 participants did not understand what active carts referred to (carts currently out in the warehouse picking). It is possible for a participant to first fail to create a frame, then to create an incorrect frame until they hopefully create a correct frame, increasing the number of error-like attempts.

It should be mentioned that participants could have created right or wrong frames in their own head without having mentioned it. Participant 304 created an incorrect frame about the heat map and then they were unsure what the heat map showed. However, in the end they were able to draw the conclusion that it takes more time to pick at the elevators than what is expected, showing that they understood what was shown without having verbalized creating a correct frame. These errors have not been accounted for as it is not possible, while using this method, to accurately predict what participants thought if they did not vocalize it. It is equally possible that participants created incorrect frames without saying so. However the main problem areas for novice users are clearly shown. In general the unclear parts of the visualization were features surrounding active carts (what an active cart is, what is shown when a cart is clicked and what the colouring means), how the heat map works, what delay is and how the carts are coloured. As 4 out of 10 participants failed to construct frames about how delay worked it could perhaps be seen as a main problem. One participant mentioned that they did not know what type of data the company had available, moreover, since delay can be easily explained as the time difference between expected picking time and actual picking time, it should perhaps not be seen as a big concern. Here are some quotes from participants: “ok then I don't understand why if it is the overall delay if it is 7 minutes and then it changes to 12 minutes and then it goes back, that is kind of weird”, participant 301. “Here it says ‘cart delay’. What does that mean? /.../, I am trying to understand the system, I’m feeling a bit sluggish to..., ‘cart delay’. I don’t get it. Is it like queuing? Is it how long it has been waiting there or what?” said participant 310. “‘Cart delay’, what could that be?” participant 304.

UMUX-LITE Results

Table 3 shows a summary of all data gathered by the questionnaire for all participants in the study. It seems that the average score went up during the process, but as seen in Figure 9 and Figure 10 the standard deviations for these results are too large so it is not possible to determine any statistically significant difference. In Figure 11 the percentile ranks of the results for each UMUX-LITE item have been calculated, although not showing any standard deviations the results from the final prototype indicate a good performance, whereas the versions tested in the pilots performed worse. When the percentile ranks presented in Figure 11 are compared to other pieces of software it is clear that the final visualization developed in this paper ranks higher in usefulness than MS PowerPoint and ranks higher in usability than MS Word, this indicates good performance, well above the average for the 17 tested pieces of software in a 2017 report [23]. The pilots
versions ranked very badly compared to other well known services [23].

<table>
<thead>
<tr>
<th></th>
<th>Usefulness</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>σ</td>
</tr>
<tr>
<td>Pilot 1</td>
<td>4.5</td>
<td>1.9</td>
</tr>
<tr>
<td>Pilot 2</td>
<td>5.3</td>
<td>0.5</td>
</tr>
<tr>
<td>Final</td>
<td>6.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Table 3. Showing average score for both of the UMUX-LITE items and the standard deviation, for both of the pilot trials and the final study. It is based on a 7 point scale, best score being 7 and worst score being 1.

The UMUX-LITE data was converted to SUS-equivalent scores [20] and that data is shown in Table 4. It is clear that the score improved over the course of development and that the final score is relatively high.

<table>
<thead>
<tr>
<th></th>
<th>SUS Equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted</td>
</tr>
<tr>
<td>Score</td>
<td>Percentile rank</td>
</tr>
<tr>
<td>Pilot 1</td>
<td>66</td>
</tr>
<tr>
<td>Pilot 2</td>
<td>72</td>
</tr>
<tr>
<td>Final</td>
<td>78</td>
</tr>
</tbody>
</table>

Table 4. Showing SUS Equivalent scores along with percentile ranks and grade equivalent, from data gathered during all 3 test sessions. Source for percentile rank and grade: Quantifying the User Experience: Practical Statistics for User Research [28]

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K₀</td>
<td>K₁</td>
<td>K₂</td>
<td>K₃</td>
<td>K₄</td>
<td>M</td>
</tr>
<tr>
<td>Usefulness</td>
<td>0.28</td>
<td>-0.16</td>
<td>0.26</td>
<td>-0.09</td>
<td>-0.04</td>
<td>-0.35</td>
</tr>
<tr>
<td>Usability</td>
<td>-0.65</td>
<td>-0.27</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.15</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 5. Showing coefficients for linear regression.

While looking at the raw data there seemed to be a trend between the number of NOVIS 5 activities and giving a low score on usefulness, in particular this was noted for participant 310 who started this activity 5 times. Data in Table 5 presents possible coefficients for linear regression (Score_{UMUX Item} = K₀ + K₁ * N₀ + K₂ * N₁ + K₃ * N₂ + K₄ * N₃ + K₅ * N₄ + K₆ * N₅ + M, where N is the amount of times an activity was started by a participant). Results may show that there is some trend for usefulness where Miscellaneous and Creating frames (NOVIS activity M and 2) seems to positively predict the user given score and Floundering on visualization seems to have negatively
impacted the score. These results are not presented with any degree of certainty apart from the R2 value.

Interview results
In the post-interaction interview participants were asked to give suggestions for improvements. 4 out of 10 suggested that there should be a deselect button, as clicking the selected cart it again was not clear and when the cart was out of screen it was impossible to click it again. 2 out of 10 mentioned that they would prefer to have the shelf colour change over the course of the day, to enable indication of when a shelf was used. 2 out of 10 wanted to add a separate window where the hover text would be. 2 out of 10 wanted to have information about how many items there was on each pick list. 2 out of 10 suggested that having a live headman or a headman per hour would give more insight. There were many other suggestions but they were only mentioned by one participant.

In the interviews, participants were asked to say what they believed to be the main causes of delays, and to base this answer on what they had seen in the visualization. Answers were mainly about problems they had known before, but often new problems were mentioned or they were able to confirm the existence of a problem. 7 out of 10 participants said that the elevators were a main cause of delay, this was not new to them and they already knew this. The second most mentioned problem was that pick lists were not so spread out, 5 out of 10 participants mentioned this. That problem was also known, participant 303 said this when asked if new information was presented to them: “I knew it, but I think the visualization shows it well”. 3 out of 10 participants mentioned clustering as a main reason for delays. 3 out of 10 participants mentioned in their answers that the frequent picking at a shelf at least sometimes caused delays (it was also mentioned that there were exceptions to this, frequent picking was not seen as something that always caused delays). 3 participants mentioned that corners and narrow aisles caused delays: “Here you can see that it goes much faster than the aisles where it is narrow”, participant 302. This was a problem that was described as new and one participant said that they would want to explore that further.

9 out of 10 participants said that they were able to identify areas where there frequently are delays, queuing and/or jams. One participant instead said that they would prefer to have a visual cue when there was a stop in the warehouse, saying that a stop had occurred rather than showing individual carts and their delay. It should be noted that data of that high level quality was not available.

Result summary
The results show that the overall design was successful, but has room for improvement. The interview shows that users were able to use it as an analytical tool and identify problem areas in the warehouse based on how the pick lists were created. The results from the questionnaire show that the visualization was useful and easy to use. The score indicate a good performance as it ranks above average. NOVIS analysis shows the main problem areas, that participants did not initially understand what the colour or size of a cart indicated and what the heat map showed. How the map was coloured after active carts were selected was also not always clear without explanation. As participant 303 said after looking at the legends: “It is very clear so you see it very clearly, at least after having read .../. Before I read the description so like, I would have [understood], without reading what the red and green was, but it would have been difficult to know what these were”, last part referring to the colouring of the shelves.

DISCUSSION
The aim of this paper was to find the main affordances and limitations of map-based visualizations in this particular scenario. The main affordances of the visualization described in this paper are button affordance, slider affordance, hover affordance, click affordance and a visual affordance. The last one, which is also the most important one, refers to the fact that the visualization affords conveying information visually, as all visualizations do, by looking at the heat map. However the slider and looking affordances combined enabled an overview of time. This overview of time was most clearly shown when users set the speed multiplier to max and pressed play to watch carts move rapidly. It is clear that although the button and slider affordances did not themselves aid IT-staff in their task of quantifying the performance of pick lists, it did help them to navigate the interface as they afforded clicking and dragging. Hover and click affordances helped to provide details of how a specific pick list performed.

The main limitations include not providing an intermediary level of details and overview of the warehouse performance, e.g. averaged by the hour, showing density. Another major limitation is not having access to a higher level data, similar to traffic incidents, as it would be possible to focus solely on ‘warehouse incidents’ and perhaps enable more micro level insight as traffic incident visualization research has done before [2].

Other limitations were due to design mistakes, as colour on scales or size of carts were not intuitive without reading the legends, a part of the interface users often neglected when they started to interact with the visualization. For a novice user, it was not clear what information was shown to them, however, when participants read the legends and scales there were fewer misconceptions suggesting that for an expert user the scope of these issues is small in size. The visualization was only map-based, this is a limitation as well, in this case it could have been useful to provide a simple graph of average cart delay for every minute of the day (as was
suggested by one participant), or something similar to what has been researched in traffic visualizations [1, 11, 31, 33, 36].

Through the NOVIS analysis it was possible to find problem areas. There is a pay off, mentioned by the creators of the NOVIS method [17], as it is based on think aloud and inherits all problems associated with that method, however in this instance it was beneficial to identify problem areas for the analysis tool for novice users. This could have been done in some other ways as well, either by using a regular think aloud or by letting users describe each aspect of the interface and what they think it means.

When comparing results to well known pieces of software tested in 2017, it is clear that most testers could have had several hours experience of e.g. MS Word and MS PowerPoint (two of the pieces of software tested) [23]. A score roughly equivalent to their performance after only 10 minutes of testing should be seen as a very satisfactory result, despite the small inaccuracies of UMUX-LITE [21]. However, it is clear that SUS is generally better to use, but as UMUX-LITE has been found to be a reliable scale [5], and also considering the limited amount of the participants’ time, the use of UMUX-LITE in this setting was well motivated.

The conclusions participants were able to draw show that almost all of them used it as an analysis tool. Why only some managed to draw conclusions during the think aloud and in the interview may be unclear, however, as 9 out of 10 answered that they could identify areas where there were delays and clustering, the visualization should be seen as successful. The fact that known issues were shown, mainly the elevator causing delays, is a sign that the visualization was able to show existing problems. The fact that not many new problems were discovered may be because there are none. When things worked the way they were supposed to participants saw that and they also managed to connect the time of the day to which shipment deadlines were approaching. In a badly performing system those things would not be possible.

There seems to be a connection between ranking low in usefulness to not understanding (NOVIS activity number 5). UMUX-LITE scores and SUS scores have been found to increase with frequency of use [5], perhaps this is one of the underlying reasons, not understanding what is shown makes an interface less useful, but with frequency of use the unclear areas might become more clear. As the results seem to have been modulated by confusion in this way and as UMUX-LITE scores increase by frequency of use, it is not very controversial to suggest that an expert user with more experience, using the visualization as an analysis tool, would rate the visualization much better compared to relatively new users who participated in this study. These results also suggest that if aspects of the visualization had been explained to participants they might have rated the visualization as more useful. However, there is no substantial evidence showing that errors-type activities cause users to give lower usefulness scores, statements here are based on trends in data and there is a need to investigate this further.

That some participants had seen parts of the visualization before they participated in the study could explain why they did not initiate with certain NOVIS activities or why they did not create certain frames, as they perhaps had done this before. Had any participant made incorrect frames about some aspect of the visualization before they participated this would most likely have resulted in them questioning or being confounded about what was shown, increasing the frequency of overall issues. Had they made correct frames beforehand and then did not vocalise this, the amount of positive results would have been reduced. Using only ‘true’ novice users would then perhaps result in more ‘positive’ results. It could have been useful to provide participants with a demo if the visualization, however in that case it would not have been possible to do a NOVIS-analysis.

For any method aiming to evaluate a visualization there is always a trade-off between generalisability, precision and realism. When a complex visualization is analyzed there is a risk that the results are “due to a particular underlying technique or the overall system” [8] and as a researcher it can be difficult to tell which has caused the results. For example, was the visualization ranked highly based on how useful the heat map was or the entire system? The interviews and NOVIS analysis aimed to give some more insight to this, but when it comes to how the visualization was ranked in the UMUX-LITE questionnaire there is no way of telling why it received a certain score. Had the study been conducted again it would most likely involve gathering more of qualitative and quantitative data on how each aspect of the visualization was perceived. It is known that laboratory experiments, field experiment and field studies give precise and/or realistic results but they may not be very generalisable [8]. As this study has adopted a method similar to a laboratory experiment and field experiment results should be considered precise and realistic but not very generalisable.

However, some aspects of the results can be compared to other studies. As it was not an issue for users to understand the abstract map (that it represented the warehouse) it again shows that abstract maps are useful [13], however this was not a comparative study, a more realistic map could also have worked well. This paper has had a different set of data than has been used before [34], and different methods have been used to compensate for the level of uncertainty resulting from this data. Still, it shows that it is possible to work with sparse data.

Future work could aim to examine how intermediary layers of map-based and map-mixed based visualizations
should be designed. These layers should give an overview of a smaller set of time or space. What type of insight is possible to achieve using this type of visualization feature? Using non-map-based features in combination with map-based features remains unexplored for warehouse data. Future work could also include gathering of performance data in combination with defining specific tasks for participants to do, as this was not done in this study. There is no way to filter or select data based on location or any multi-dimensional aspect (e.g. making a query), this is also something that could be examined.

CONCLUSION
It can be concluded that the visualization was successful in its attempt to visualize pick lists, despite some bugs and unclarities, this is seen as user satisfaction was high and users were able to understand it and use it as an analytical tool. The main affordances were found to be button affordance, slider affordance, hover affordance, click affordance and a visual affordance. Limitations found include the impossibility of getting a time based overview using a solely map-based visualization. On the other hand, it was still possible to evaluate pick lists and warehouse performance at a specific time of the day. Known problem areas were confirmed and some new issues were found, mainly that pick lists are not created in an ideal way, and in particular when the morning shifts start. The main improvement that was recognized by participants addressed this issue and suggested that the start location for picking should be more spread out. The heat map indicated that there are delays in corners, something that was unknown to some participants.

ACKNOWLEDGMENTS
I would like to thank everyone at Apotea AB and at the Royal Institute of Technology who have been supervising and giving feedback. Your input has been very valuable in the process of both writing and coding.

REFERENCES
15. Jacyna, Marianna and Wasiak, Mariusz and Bobiski, Andrzej. 2017. SIMMAG3D as a tool for designing of storage facilities in 3D. Archives of Transport 42.


