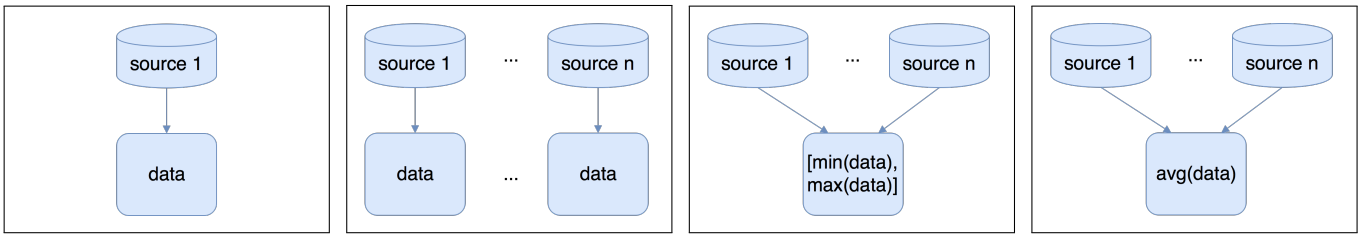
Increasing Users' Confidence in Uncertain Data by Aggregating Data from Multiple Sources

Miriam Greis, Emre Avci, Albrecht Schmidt, Tonja Machulla
VIS, University of Stuttgart, Germany
{firstname.lastname}@vis.uni-stuttgart.de



(a) Single source (b) Direct comparison (c) Range aggregation (d) Average aggregation

Figure 1. Many interactive systems, for example weather applications, show uncertain data from a single source (see a). To encourage the design of interactive applications that show uncertain data from multiple sources, we identified three aggregation mechanisms. The direct comparison allows users to look at data from multiple sources at the same time introducing mental workload to aggregate the data. The range and average aggregation provide computationally aggregated data which reduces the mental workload for the user.

ABSTRACT

We often base our decisions on uncertain data - for instance, when consulting the weather forecast before deciding what to wear. Due to their uncertainty, such forecasts can differ by provider. To make an informed decision, many people compare several forecasts, which is a time-consuming and cumbersome task. To facilitate comparison, we identified three aggregation mechanisms for forecasts: manual comparison and two mechanisms of computational aggregation. In a survey, we compared the mechanisms using different representations. We then developed a weather application to evaluate the most promising candidates in a real-world study. Our results show that aggregation increases users' confidence in uncertain data, independent of the type of representation. Further, we find that for daily events, users prefer to use computationally aggregated forecasts. However, for high-stakes events, they prefer manual comparison. We discuss how our findings inform the design of improved interfaces for comparison of uncertain data, including non-weather purposes.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation (e.g. HCI): User Interfaces

CHI 2017, May 06 – 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05...$15.00.
DOI: http://dx.doi.org/10.1145/3025453.3025998

Author Keywords

Uncertainty; Aggregation; Multiple Sources; Comparison; Weather Forecast

INTRODUCTION

People have to deal with uncertain data in their everyday lives and understanding it correctly can be crucial. For example, the misinterpretation of a long-term forecast of a financial market trend can lead to negative consequences for the individual. Most of the time, people however use short-term predictions such as arrival time or calorie expenditure predictions. The most prominent examples for such a short-term prediction are weather forecasts. Here, people are usually aware of the uncertainty in the prediction [22] and take it into account (e.g., by bringing an umbrella even though the chance for rain may be low).

For many forecasts, users can choose between different providers. The information between providers may differ as they use different models as basis for their forecasts. In informal interviews in the context of a former study, we found that a large percentage of people compare forecasts from different providers since they do not have sufficient trust in a single source. For example, people will consult several weather forecast providers before going on a hike in the mountains as the weather can change quickly and facing a thunderstorm in the mountains can have deadly consequences. However, the comparison of multiple sources of information can be tedious and cumbersome. For example, many people reported that they open several tabs, each for one weather website, and then manually switched between them while trying to remember what values the other forecasts showed. This approach

requires mental effort. High-level services that combine multiple forecasts from different sources could support comparison. The demand for such services for easy comparison is clearly given as some applications[1] supporting weather forecast comparison recently entered the market. What is still missing, is a theoretical underpinning of how to design for easy comparison of weather forecasts, and for uncertain data in general.

The main goal of our work is to understand how services that aggregate and display multiple forecasts should be designed to be most useful for users and increase their confidence in the uncertain data depicted. Therefore we compare three aggregation mechanisms: two mechanisms that computationally aggregate the data (see Figure 1c and Figure 1d) and one mechanism that supports manual comparison (i.e., the user can directly compare the data from separate sources without having to switch between applications or websites; see Figure 1b). Computational and manual aggregation are likely to differ with regards to the amount of mental workload they induce and the amount of perceived control, as in one case the computer aggregates the data for the users and in the other the users have to aggregate the data on their own.

In a first step, we evaluated the aggregation mechanisms using an online survey. We used three different visual and one textual representation as well as three scenarios to understand whether and how they influence the preference for one aggregation method over another. We then conducted a real-world study using a custom Android application that supports manual and computational comparison.

Our three contributions aligned to the goals of our research are as follows:

1. We present **aggregation mechanisms for uncertain data from multiple sources** spanning the range from automatic to manual aggregation.
2. We show that the **aggregation of multiple sources is preferred to single source forecasts**, independent of the type of representation. However, the preference depends on the importance of the scenario, for which the weather forecast is consulted. Additionally, we show that **people place higher confidence in aggregated forecasts**.
3. We provide **implications for the development of applications that support comparison**.

Our results demonstrate that designing for comparison increases users' confidence in uncertain data. We specifically recommend to use direct comparison when the potential negative consequences associated with an uncertain event are essential for a user and to introduce computational aggregation when the negative consequences are rather negligable.

BACKGROUND & RELATED WORK

As basis for our work, we build on previous work on uncertainty of different fields including psychology, HCI, visualization, and meteorology.

[1]Climendo: http://climendo.com/, WeatherXM: http://weatherxm.exm.gr/

Definitions of Uncertainty

Most work on uncertainty is domain-specific, which leads to manifold definitions and references of the term uncertainty. Because each domain focuses on their research independently, the term 'uncertainty' is used inconsistently [21]. This leads to specialized typologies, for example for geospatial information [37] that only work in the context of one domain. One of the earliest classifications for visualizations of uncertain data focuses on the data type and the visualization extent [27]. Potter et al. [29] recently proposed another taxonomy focusing on to-date visualization approaches for uncertain data that takes into account the data dimension and the data uncertainty dimension. Broader classifications focus on the level of the data [34] or the sources of uncertainty [29]. In our work, we focus on implicit uncertainty induced from using multiple sources rather than on statistical uncertainty. Each source of uncertain data may include statistical uncertainty, however, it is often not communicated by providers.

Uncertainty in Psychology and HCI

A multitude of psychological studies show that humans struggle with probabilistic information. When they judge uncertain information, humans are prone to biases and heuristics [38]. Humans also have difficulties to integrate information to make binary decisions [14]. These findings are particularly important for the field of risk perception to understand how to communicate, for example, information about a hazard [35].

In HCI, there has been a recent increase in the interest in uncertainty visualization and communication, such as in the case of the exploration of personal genomics data [32], data analysis [8], machine learning [17, 41], bus arrival predictions [15], or range anxiety in electric cars [13]. Kay et al. [16] reported that missing uncertainty information decreased trust for body weight measurements. In all of these application areas in HCI it is possible to use aggregated forecasts if multiple predictions from different sources are available and they may therefore benefit from our findings.

Visualization of Uncertainty

A large number of previous publications have focused on the visualization of uncertain data. Glyphs [27, 42] are for example heavily studied for visualizing uncertainty for experts. Other visualization methods such as line graphs [36], box plots [28], or bar charts [3] have been studied as well. Other research focuses on a specific type of uncertainty and compares different approaches, e.g., for bounded uncertainty [26]. For laymen, either quantitative or qualitative approaches can be used to communicate uncertain data. Both approaches have considerable drawbacks. On one hand, understanding quantitative information can be difficult as studies showed that even well-educated adults have difficulties in solving easy probability questions [20]. On the other hand, qualitative information can be misleading due to different perceptions of terms such as "low risk" or "low uncertainty" [39].

In spite of this wealth of findings regarding uncertainty visualization, there are no clear guidelines for which visualizations to use in which context. One more main challenge in uncertainty visualization is the quantification of the uncertainty.

The uncertainty has to be captured and modeled accurately to ensure correctness of the data displayed in the visualization [1]. As quantification of uncertainty is difficult, comparing outcomes from different models (in our case different providers) can be used to implicitly get information about the uncertainty of the models.

Uncertainty in Weather Forecasts

Already several years ago, the National Research Council [5, 6] and the American Meteorological Society [4] highlighted the importance and value of communicating uncertainty to the public. This has raised an interest in research in uncertainty communication [2, 7]. In a nationwide survey in the US, Morss et al. [22] found that 96% of the participants use weather forecasts more than three times per day. People are also aware that weather forecasts are uncertain [11, 19, 22, 23]. Multiple studies conducted by researchers in the field of meteorology and psychology showed that people prefer forecasts with uncertain information to deterministic ones [22] and additionally make better decisions [10, 24, 30, 31].

Adding uncertainty increases perceived transparency and reliability, probably by increasing the trust in a forecast [12, 30]. Joslyn et al. [12] also found that showing explicit hints regarding how to make a decision did not help people. Mostly, they wanted to do a self-informed decision and did not have confidence in the forecast of the system. Siegrist [33] describes the concept of general confidence in risk perception as the feeling that everything is under control and that uncertainty is low, which is what should be achieved by systems communicating uncertain data.

Although much research exists regarding how to communicate and visualize uncertainty in the weather domain, current forecasts seldom communicate uncertainty beyond the probability of precipitation. This is mainly the case because there is still the perception that the general public does not understand probabilities [18, 25]. Additionally, quantification of the uncertainty is not a trivial task. One interesting approach has been developed by Frick et al. [9] who presented the predictions of different models in an ensemble flood warning system. Here, participants preferred more detailed information after using the system for a while to make an informed decision on their own. We see a connection to people looking at multiple weather forecasts and comparing them. As different forecast providers use different weather models and data, the easy comparison of these forecasts could support people in their feeling of making a more informed decision.

DESIGN RATIONALE FOR AGGREGATION MECHANISMS

Supporting the comparison of forecasts can be achieved in two ways. First, manual comparison of forecasts could be supported by reducing the mental workload and the tedious switching between providers. Second, the data could be aggregated by the computer to decrease the mental challenge even further. However, this second approach may result in a feeling of loosing control as people do not necessarily follow given advice and decision aids [12]. To compare a wide range of possible mechanisms, we use the direct comparison as a mechanism that still requires a high mental workload from users as they have to aggregate the data on their own. This is in contrast to the average aggregation where the comparison is completely pre-defined. The range aggregation is in-between these two mechanisms as it computationally aggregates data, but still allows for users' own judgment (see Figure 2).

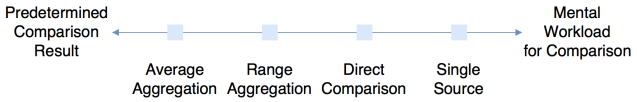


Figure 2. The Aggregation Mechanisms differ on the amount of mental workload needed for comparison by presenting less predetermined comparison results (from left to right).

We describe the mechanisms in more detailed in the following sections and also include the a baseline "Single Source". To identify the ideal number of sources that people normally compare, we conducted a short informal survey with six participants where five out of six participants stated that three sources are the optimal number, and one participant stated that two sources would be optimal. Therefore, we decided to use three sources in all our design considerations and sketches.

Single Source

A single source forecast (see Figure 1a) is how many current weather forecasts are displayed – i.e., the weather forecast shows the weather data of one source. We mainly use this as a baseline for comparing the other aggregation mechanisms against the current state of weather applications. The single source induces the highest mental workload in the user when comparing different forecasts as the user has to open different sources and potentially remember them if they cannot be shown next to each other (for example due to small screen size on a mobile phone).

Direct Comparison

Direct or manual comparison (see Figure 1b) is similar to having several single source forecasts in one system such that they can be perceived in one glance (similar to the approach used by Frick et al. [9]). This reduces the mental workload on comparison as values can be compared directly without remembering them. It also reduces the time and effort of finding multiple sources, but the comparison and interpretation of the data itself still needs to be done "manually" by the user. The effectiveness of this aggregation mechanism also depends highly on the number of sources for information and screen constraints. On a small screen of a mobile phone for example, only a few sources can be shown next to each other without scrolling.

Range Aggregation

The range aggregation (see Figure 1c) is a computational aggregation mechanism that still enables users to make some comparison on their own. A range representation was already suggested by Morss et al. [23] for showing weather forecast uncertainty. In contrast to their visualization, our aggregation mechanism does not take the values from the same weather source but shows the minimum and maximum values from

across multiple sources of information. This gives the user the possibility to judge by how much sources vary without having to compare all values manually. On the downside, having one source with an extreme erroneous value would affect the outcome of this aggregation mechanism significantly.

Average Aggregation

The average aggregation mechanism (see Figure 1d) is identical to the information provided by the single source mechanism, except that the values are averages across multiple sources. It reduces multiple values into a single, easy to interpret value. The mechanism therefore keeps the simplicity of a single source forecast, but at the same time contains information from multiple sources. In our case, the average is computed by equally weighting the sources, but it could also be computed as a weighted average based on the accuracy or on the uncertainty spans of the forecast providers, if available. When an average is used, this has to be communicated very clearly to users as it may otherwise be interpreted as a single forecast, as these two mechanism are indistinguishable without additional information. Although average aggregation is simple, users do not have the possibility to make an informed decision on their own as the data is aggregated into a single value, which may give them a feeling of loosing control.

RESEARCH QUESTIONS

Our main goal is to understand the influence of aggregated forecasts on people. We broke this goal down into five concrete research questions (RQ1 to RQ5), each with one associated hypothesis regarding what we expect to find.

RQ1: Does the aggregation of multiple forecasts change the users' confidence in uncertain data? We assume that aggregation increases users' confidence in uncertain data. Especially for weather forecasts, where model uncertainty is seldom presented to the public, users are able to get a better understanding for their uncertainty by comparing them. This leads to *H1: People are more confident regarding aggregated forecasts than single source forecasts*.

RQ2: Do people prefer aggregated forecasts to single source forecasts? We assume that people generally prefer forecasts with aggregation as aggregation adds more information and lets them do a more informed decision. Also weather applications that recently entered the market support aggregation. This leads to *H2: People prefer aggregated forecasts to single source forecasts*.

RQ3: Do people prefer different aggregation mechanisms (manual vs. computational aggregation) according to the importance of a scenario? Aggregation mechanisms can provide a different amount of data. Either the computer can aggregate the data in a meaningful way or the computer can help the user to aggregate the data in a meaningful way. The latter induces a higher mental workload. We assume that whether users prefer full computational aggregation or rather would like to aggregate the data themselves depends on the importance of a scenario and on context. This leads to *H3: People prefer manual aggregation for important events because this gives them a feeling of control*.

RQ4: Do people prefer different aggregation mechanisms depending on the type of visual or textual representation? Aggregated forecasts can be shown with a multitude of visual and textual representations. These representations may have certain criteria which make them more suitable for a specific aggregation mechanism. This leads to *H4: People prefer different aggregation mechanisms depending on the type of visual or textual representation*.

RQ5: Can the theoretical findings be transferred to real world application usage? We want to transfer the theoretical finding to a real world application to better understand whether our hypotheses do also apply in real world usage.

EXPERIMENT 1: ONLINE SURVEY

We evaluated the three aggregation mechanisms and the single source baseline by conducting an online survey.

Design

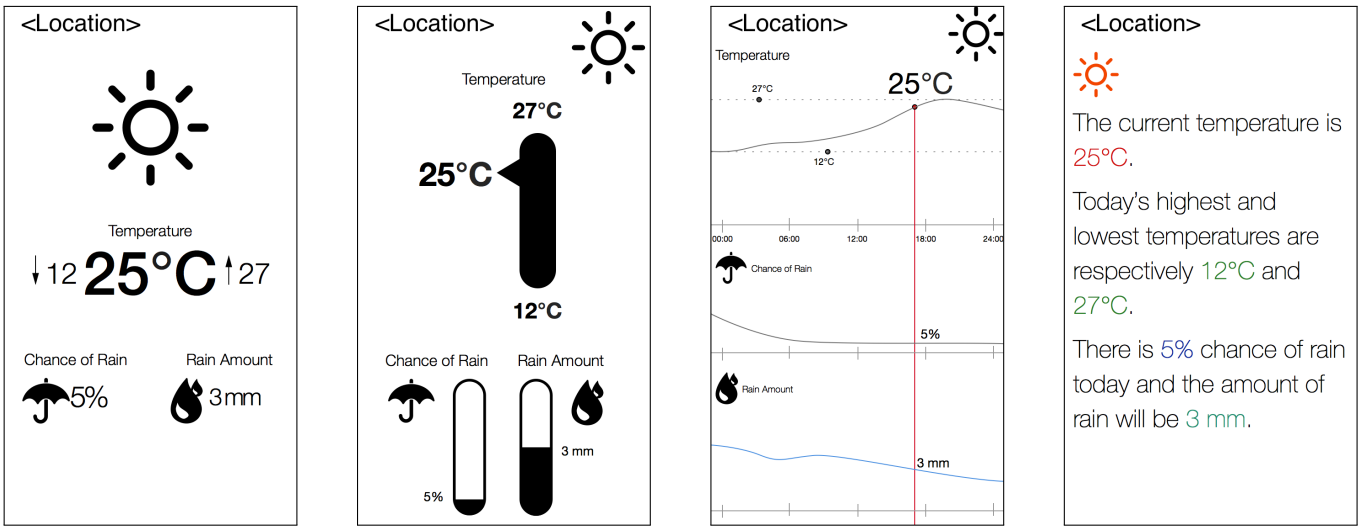
We used a 4 x 4 x 3 within-subject design with three independent variables: *Aggregation mechanism* (with the four levels single source, direct comparison, range, average), *representation* (with the four levels text, pictogram, bar, line), and *scenario* (with the three levels daily dress code, outdoor BBQ party, outdoor wedding). For each combination of aggregation mechanism and representation, we measured participants' general confidence and their preference regarding all scenarios using a 7-point Likert item.

We decided to focus on the three most reported weather properties: temperature, chance of rain, and rain amount.

Representations

Since weather providers use different representations for weather forecasts, we included multiple representations in our survey. To identify suitable representations, we looked at existing weather applications, which mainly use pictograms with numbers or line graphs. We additionally introduced bar graphs and a textual representation to cover a span of different representations. Although the uncertainty in our study is not stochastic, but instead stems from the divergence of information from different sources, we treat it as model uncertainty. For the bar charts and line charts we use the ambiguation methods developed by Olston et al. [26] to show a range aggregation.

All representations show the following information: the location of the weather station, a pictogram of the current weather (e.g., a sun for sunshine) and numeric values for the weather properties temperature (current temperature as well as predicted daily minimum and maximum), chance of rain, and amount of rain. The pictogram representation (see Figure 3a) is dominated by a large pictogram of the current weather (e.g., a sun or a cloud). Other weather properties are presented numerically. The bar representation (see Figure 3b) shows (partially) filled bars with adjacent numeric values. The line representation (see Figure 3c) shows a line chart for each of the weather properties and a line to indicate the current hour of the day. The text representation (see Figure 3d) reports the weather verbally, with the numeric values for different weather properties emphasized by using colored font.



(a) Pictogram (b) Bar graph (c) Line graph (d) Text

Figure 3. Exemplary sketches for the representation methods used in the online survey displaying a single source or average aggregation forecast.

Scenarios

To obtain scenarios for our study, we assembled a list of 12 scenarios, in which weather conditions may influence people's behavior: (1) What to wear when going to work, (2) Whether to take the train or bike to work, (3) Outdoor soccer viewing event with friends, (4) Camping trip, (5) Outdoor wedding, (6) Packing for a trip, (7) You are going to do outdoor sports (e.g., swimming, hiking), (8) Planning an outdoor barbecue party, (9) Harvest as farmer, (10) Garden work, (11) Whale watch, (12) Attending outdoor concert.

We perceived the time of the event as an important factor as people know that the accuracy of weather forecasts decreases with time. In order to eliminate the time factor from our scenarios, we decided to address all of them as being "tomorrow".

In a short hallway questionnaire, we asked 15 people to decide whether a scenario was casual important, somewhat important, or very important for them. From the answers of participants, we selected three scenarios of varying importance were most participants agreed on the importance. This resulted in the scenario with the least importance (10 participants chose "casual important"), one with medium importance (11 participants chose "somewhat important") and the one with the highest importance (13 participants chose "very important"): (1) Daily Dress Choice: You are going to work/school tomorrow morning on a regular day and you need to decide what to wear., (2) Outdoor BBQ Party: You are planning a BBQ party for tomorrow and you want to know whether the weather will be good for having a party outdoors., (3) Outdoor Wedding: You have your wedding tomorrow and it is an outdoor wedding. You need to decide whether you should make changes in the organization (like arranging a tent, or moving the wedding indoors).

With this range of scenarios, we made sure to expose participants to scenarios with a varying amount of importance.

Participants

In total, 71 participants (37 male, 33 female) between an age of 19 and 77 years ($M = 24.8, SD = 7.0$) completed our online survey. We recruited them via social networks and our University mailing list. The majority of the participants had either a completed high school diploma (26.8%) or a completed university degree (59.1%). As compensation, participants had the possibility to be part of a raffle of two vouchers worth 20€.

Most of our participants (70.4%) stated that they consult the weather forecast multiple times a week. Further, participants reported that they use the weather forecast more or less frequently depending on the season (*"Especially in summer because the weather changes very often."*), events (*"I always check it before going to swim outside, and a barbecue, or a festival (everything outside where it could be fatal to wear the wrong clothes or the event is not appropriate for rainy weather)."*), travel plans, location (*"It usually depends if I am going somewhere where the weather patterns are unfamiliar to me."*), and specific weather conditions (*"Especially when the weather doesn't seem to be stable."*).

The majority (52.1%) of our participants reported using more than two weather providers regularly ($M = 1.68, SD = 0.92$). They picked their weather providers due to the perceived reliability and accuracy, usability, fast and easy access, and amount of information. Several participants stated that they discontinued using weather sources before either because they were inaccurate or not easily available. Participants reported to compare sources for several reasons: to find out whether the sources provided consistent information (*"If I really need to know the weather I look at all of them to see if they match."*), to compute the average (*"I find they never agree with one another so I always try to find multiple sources and average them out (informally)."*), or estimate the worst possible scenario (*"I usually estimate the worst possible weather for a given scenario and prepare accordingly."*).

Tasks

The online survey comprised the evaluation of sketches of several potential interfaces. Each sketch depicted the combination of one representation with one aggregation mechanism. Therefore, the total number of tasks was 16: four aggregation mechanisms x four representations. For each sketch, participants first rated their confidence in the depicted weather forecast on a 7-point Likert-type item (*"I feel confident that this will be the tomorrow's weather."*) ranging from *completely disagree* to *completely agree*. Participants were aware that the depicted weather forecast was hypothetical, i.e., it did not veridically display the weather of the next day. Subsequently, participants rated their preference for the sketch on three 7-point Likert-type items, each for using the depicted interface in one of the three possible scenarios (e.g., *"I would like to use this representation for the scenario: Daily Dress Choice"*). For every task, we additionally gave participants the possibility to provide qualitative feedback in the form of a text field.

Procedure

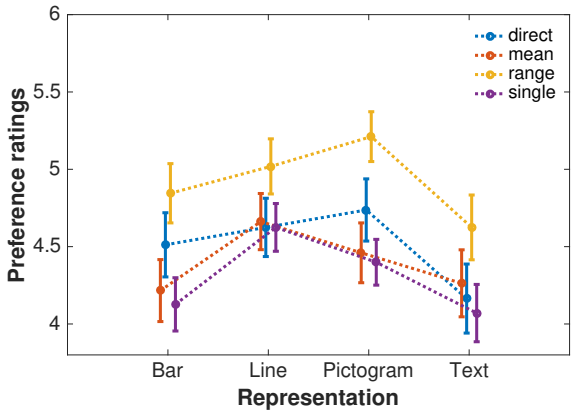
Participants began the survey by filling in their demographic information and relation to weather forecasts. They then navigated through four pages (one per representation). The order of presentation of the four representations was randomized across participants. On each of the pages, participants received an explanation of the representation, all aggregation mechanisms and the scenarios. Additionally, they completed the four tasks for the shown representation. At the end, participants who wanted to be part of the voucher raffle could enter their email address.

Results

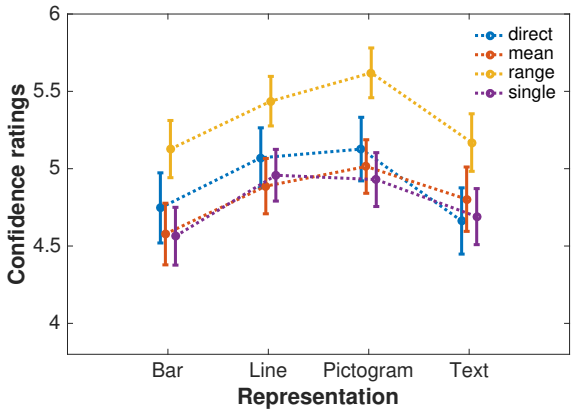
We performed two linear mixed-effects model analyses on the aligned-rank transformed Likert item data [40], one for participants preference ratings and one for participants confidence ratings. In the first analysis, we added the fixed factors *aggregation method*, *representation*, and *scenario* as well as the random factor *participant*. In the second analysis, we added the fixed factors *aggregation method* and *representation* as well as the random factor participant. Significant effects were further explored using post hoc pair-wise comparisons with Bonferroni corrections. Only significant effects of interest are reported.

Aggregation method

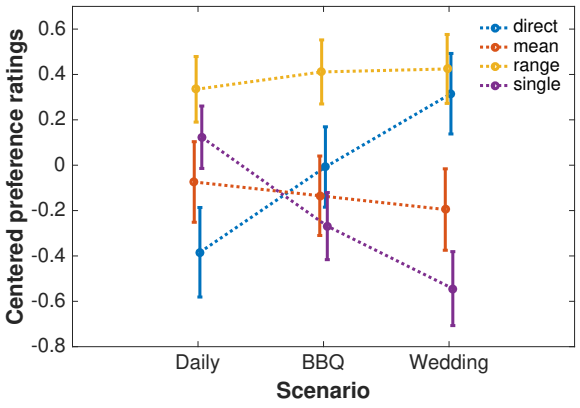
Range received the highest mean rating ($M = 4.85, SD = 1.73$), followed by *direct* ($M = 4.51, SD = 1.9$), *mean* ($M = 4.22, SD = 1.82$), and *single* ($M = 4.13, SD = 1.72$). Figure 4a shows participants overall **preference ratings** for each of the four aggregation methods as a function of the representation method (error bars represent standard errors of the mean). The mixed-effects analysis revealed a main effect of aggregation method on participants preference ratings ($F(3, 3290) = 37.78, p < 0.001$). Pair-wise comparisons showed that participants preferred *range* over the other three methods (range vs. direct: $t(3290) = 5.55, p < 0.001$; range vs. mean: $t(3290) = 7.69, p < 0.001$; range vs. single: $t(3290) = 10.21, p < 0.001$), as well as *direct* over *single* (direct vs. single: $t(3290) = 4.66, p < 0.001$).



(a) Preference ratings for the aggregation mechanisms as a function of representation



(b) Confidence ratings for the aggregation mechanisms as a function of representation



(c) Preference ratings for the aggregation mechanisms as a function of scenario

Figure 4. Participants' mean preference and confidence ratings (measured on a 7-point Likert-type item with 1 corresponding to completely disagree and 7 corresponding to completely agree); error bars represent standard errors of the mean. Figure 4c depicts centered mean ratings, i.e., the main effect of scenario has been removed to improve the perceptibility of the interaction effect.

Figure 4b shows participants mean **confidence ratings** for each of the four aggregation methods as a function of the different representations (error bars represent standard errors of the mean). Mean ratings were highest for *range* ($M = 5.34, SD = 1.46$), followed by *direct* ($M = 4.9, SD = 1.75$), *mean* ($M = 4.83, SD = 1.6$), and *single* ($M = 4.79, SD = 1.5$). The linear mixed effects analysis revealed a main effect of aggregation method on participants confidence ratings ($F(3, 1050) = 13.5, p < 0.001$). In particular, pair-wise comparisons showed that participants indicated higher confidence in *range* compared to the other three methods (range vs. direct: $t(1050) = 3.48, p < 0.01$; range vs. mean: $t(1050) = 4.85, p < 0.001$; range vs. single: $t(1050) = 6.0, p < 0.001$).

These results provide support for our hypothesis H2 that people generally prefer to receive more rather than less information with regards to the weather (i.e., they prefer aggregated presentations over single source presentations) and for hypothesis H1 that they place higher confidence in aggregated data, specifically in the form of aggregation into a range.

Representations

Line received the highest mean preference rating ($M = 4.73, SD = 1.67$), followed by *pictogram* ($M = 4.7, SD = 1.73$), *bar* ($M = 4.42, SD = 1.81$), and *text* ($M = 4.28, SD = 1.88$). The mixed-effects analysis revealed a main effect of representation on participants preference rankings ($F(3, 3290) = 16.48, p < 0.001$). Pair-wise comparisons indicated that participants preferred *line* and *pictogram* over *bar* and *text* (line vs. bar: $t(3290) = 4.17, p < 0.001$; line vs. text: $t(3290) = 5.92, p < 0.001$; pictogram vs. bar: $t(3290) = 3.682, p < 0.001$; pictogram vs. text: $t(3290) = 5.436, p < 0.001$).

Mean **confidence ratings** were highest for *pictogram* ($M = 5.17, SD = 1.53$), followed by *line* ($M = 5.09, SD = 1.49$), *text* ($M = 4.83, SD = 1.67$), and *bar* ($M = 4.7, SD = 1.69$). The linear mixed effects analysis revealed a main effect of representation ($F(3, 1050) = 7.27, p < 0.001$). Pair-wise comparisons showed that *pictogram* received higher confidence ratings than *bar* ($t(1050) = 4.08, p < 0.001$) and *text* ($t(1050) = 3.8, p < 0.001$).

In sum, our results indicate that participants preferred the representations *line* and *pictogram*. The latter also receives a confidence bonus.

Interactions

Of the three possible two-way interactions and the one three-way interaction between the three factors in the three-way analysis of **preference ratings**, only the interaction between *aggregation method* and *scenario* is significant ($F(6, 3290) = 11.71, p < 0.001$). Figure 4c shows the mean preferences for each of the four aggregation methods as a function of the scenario (error bars represent standard errors of the mean). For illustration purposes, the data of each scenario was centered on the scenario's overall mean. This was done since preference differences between the three scenarios are of no interest to us (i.e., the main effect of scenario) and may mask trends in preference ratings across the three scenarios.

Contrast		df	χ^2	p-value
Direct-Single	**BBQ-Daily**	1	22.47	< 0.001
Range-Single	**BBQ-Daily**	1	10.35	0.023
Direct-Single	**BBQ-Wedding**	1	11.37	0.013
Direct-Mean	**Daily-Wedding**	1	25.24	< 0.001
Direct-Range	**Daily-Wedding**	1	10.23	0.025
Direct-Single	**Daily-Wedding**	1	65.81	< 0.001
Mean-Single	**Daily-Wedding**	1	9.54	0.036
Range-Single	**Daily-Wedding**	1	24.15	< 0.001

Table 1. Difference of the *differences* between two aggregation methods across two scenarios (e.g., is the difference between Direct and Single larger in the BBQ or in the Daily scenario).

As the figure illustrates, a main source of the interaction effect is the increase in the preference ratings for the *direct* aggregation method as the importance of the scenario increases. Five of nine pairwise comparisons involving the *direct* aggregation method show significant differences of the difference of participants preference ratings between the *direct* aggregation method and one of the other methods across two levels of the factor scenario (e.g., for the daily scenario *single* presentation is preferred over the *direct* aggregation while the situation is reversed for the wedding scenario; see Table 1 for test statistics and p-values for all significant contrasts). This partially confirms our hypothesis H3 that the willingness to perform aggregation manually increases with the importance of the scenario (though aggregation into a *range* remains the overall preferred method).

A secondary source of interaction results from the stronger decrease in preference of the *single* source as compared to the other aggregation methods as importance of scenario increases (e.g., the difference between *single* and *range* is larger for the wedding scenario then for the daily scenario).

We find no interaction effect between aggregation method and representation. In other words, there is not sufficient evidence in favor of hypothesis H4 that users have different preferences regarding the aggregation method, depending on which representation is provided. Rather, users show an overall preference for *range*.

Qualitative Feedback

We additionally collected some qualitative feedback about the aggregation mechanisms. People mostly stated that the single source (especially the pictogram representation) "*seems like enough information for daily dress choice, and it is a similar representation to what [they] usually use. However, [they] would like a more detailed forecast for event planning.*"

As expected, participants preferred the direct comparison as they felt able to make an informed decision: "*I like having control and letting me make the decisions. I am informed of all possibilities, and it is up to me to come to a decision.*"

Participants also positively expressed that they liked the range aggregation: "*I like this representation because it provides almost as much information as the side-by-side comparison, but is much easier to see and interpret. It seems more trustworthy because it provides a range rather than a single value*

for temperature and rainfall. This acts as a margin of error, which means it is more likely to be correct."

For the average aggregation, most participants stated that it "*does not represent the variation in forecasts between different weather providers, and it also (probably) does not provide the exact values reported by any one provider [and it] seems untrustworthy.*" In contrast to this, one participants also stated that he does not "*see the different temperature and so on, but [he] know[s] that it's an average of some forecast sources, that's why [he] trust[s] the forecast.*"

Discussion

The results of our survey show that participants generally preferred to receive information of multiple sources since single source obtained the lowest ratings. The average aggregation however did not receive significantly better ratings. Instead, participants preferred range aggregation and direct comparison, which are the mechanisms that support an informed decision and provide at least some transparency on the individual sources. This also illustrates that computational aggregation does not necessarily increase trust significantly as it does not provide any additional information beyond the single source. The sheer knowledge of having multiple sources apparently is not enough to increase confidence.

Regarding the representation, participants preferred the line and the pictogram. We assume that they were preferred due to being the standard representations in most common weather applications.

We did not find any interaction effect between representations and aggregation mechanisms. This indicates that users do not exhibit differential preferences of for one or the other aggregation method depending on the representation. This allowed us to chose the aggregation and representation for the application depending on users' overall preferences. In general, this also suggests that aggregation mechanisms are generally preferred no matter what representations are used and that aggregation mechanisms are not tied to specific representations.

The interaction effect between the aggregation mechanism and the scenario shows that with increasing importance of the scenario, users exhibit a slight increase in preference for range and a large increase in preference for direct comparison, while the preference for single source drops below all aggregation mechanisms. This indicates that people prefer more control over decisions in important scenarios. Interestingly, this also applies for the range aggregation, which apparently provides enough details to give users this feeling of control.

EXPERIMENT 2: "WEATHER COMPARE" APPLICATION

Based on the results of the online survey, we developed the weather application "Weather Compare", which we evaluated in an in-the-wild study.

Apparatus

We developed an Android application named "Weather Compare" to further evaluate the most promising aggregation mechanisms. Based on the results of the online survey we decided to implement the pictogram representation. The online survey did not reveal a significant difference to the line representation for preference, however the pictogram representation received a confidence bonus and better ratings for the most promising aggregation mechanisms. To improve the representation, we added hourly information, which participants had stated to be important in their comments.

For the aggregation mechanisms, we decided to implement the range aggregation and the direct comparison as these two were preferred in the online survey. Users of the application are able to toggle between the two aggregation methods on demand. The range aggregation is the default view of the application. By tapping the details button, a user can switch to the details view for direct comparison.

We looked at existing weather APIs to identify those that fulfill the following requirements: (1) Hourly weather data for 72 hours, (2) Temperature, chance of rain, amount of rain, and pictogram included, (3) Reasonable number of API calls per day, (4) Free to use for non-commercial/research purposes. We only found two APIs that fulfilled all requirements: Apixu [2] and Wunderground [3]. We picked Forecast.io [4] as a third API although it only included weather data for the next 48 hours. As all our scenarios were directed at the next day, 48 hours seemed to be enough.

The application required users to turn on their location data as it would automatically fetch the location of the user and show the corresponding weather. For the final design of the application, we built upon the layout of the existing sketches used in the survey. We added the hourly information as vertical scroll bar at the bottom of the screen and polished the application to make it look professional.

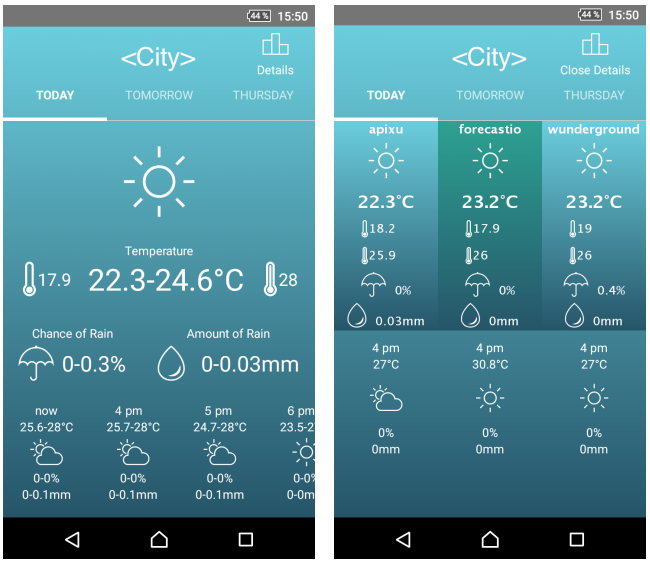
Participants

23 participants (12 male, 10 female, 1 preferred not to say) between the age of 19 and 57 ($M = 30.0, SD = 11.3$) installed the application and participated in the study. All participants were regular users of weather applications and did use the application in everyday life. We invited them by sharing the news about the application on the social media channels and the e-mail list of the University.

Procedure

We sent all participants a detailed explanation of how the application works and an apk file to install the application on their phone. We then asked them to use the application instead of their normal weather application for the course of at least one week. Afterwards, participants filled in an online questionnaire with free text questions. The free text questions asked for the following feedback: (1) Advantages/Disadvantages compared to the application normally used, (2) Feature requests, (3) Future usage of "Weather Compare", (4) Opinion about having data from multiple sources, (5) Usage of range vs. details view, (6) Potential change in confidence due to the aggregation.

[2] Apixu: http://www.apixu.com/
[3] Wunderground: https://www.wunderground.com/weather/api
[4] Forecast.io https://darksky.net/dev/



(a) App - Range aggregation (b) App - Direct comparison

Figure 5. Screenshots of the weather application "Weather Compare" with the main screen showing the range aggregation and the details screen showing the direct comparison of in total 3 different weather providers.

Results

During the week of the user study, participants opened the application on average 17.95 (+/ − 5.72) times. The details view, which corresponds to our manual aggregation condition in Experiment 1, was used on average 6.14 (+/ − 2.69) times, i.e., in approximately one-third of cases.

In the following, we present the qualitative feedback of participants regarding general feedback about the application and the aggregation.

General Feedback about the Application

We collected general feedback such as differences to the applications normally used and feature requests on the first page of the questionnaire.

Advantages/Disadvantages compared to the application normally used. Participants mainly stated that the differences between "Weather Compare" and the application they normally use are the GPS location feature and the hourly forecasts. Other participants mentioned that they see a disadvantages in the small range of the forecast (only 48 hours) and that there was no wind forecast. On the positive side, eight participants liked the range representation with multiple sources. One participant also positively mentioned the hourly forecast, and another participant found the forecast very reliable and accurate.

Feature requests. Function and feature requests mainly covered the disadvantages and suggestions for including well-known features from other applications. Additionally, one participant asked to be able to choose sources and one participant suggested to show a range for the lowest and highest temperature.

Future usage of "Weather Compare". Besides three participants, who would like to prefer using their normal application in the future, all participants stated that they would like to use "Weather Compare" in the future if the experienced issues would be fixed.

Specific Feedback about the Aggregation

We collected more specific feedback about the aggregation mechanisms and data from multiple sources on the second page of the questionnaire.

Opinion about having data from multiple sources. 14 participants explicitly stated that having data from multiple sources "*is the main thing that [they] liked about the app*". The aggregation helped them to better judge the forecast: "*I liked it because since this is a forecast, one provider is not really reliable but when you see that 3 different providers agree on something, it is more convincing.*" Only one participant did not appreciate the aggregation and stated "*that the cognitive load is too high when making the effort and actually comparing the values in contrast to just looking at a yellow sun or grey cloud.*"

Usage of range aggregation vs. direct comparison. Six participants stated that they "*always used the range representation and used the detailed view only out of curiosity.*" In contrast, three participants only used the details view. As one participant explained: "*I usually used the details view everyday, because normally I would check at least 2 different sources to be sure about the weather, thus, I would like to know what each source says rather than seeing an average of them.*" One participant regularly used both views. Six other participant stated that they used the details view if "*the range of values [was] too wide, I checked the details page.*" or "*before I spent a longer time outside*".

Influence of aggregation on confidence. 12 participants stated that the application increased their confidence as they felt that the forecast was more detailed and having multiple sources made them trust the forecast more. Only one participant stated that "*the range was typical only 2-3 degrees, in which case it does not influence [his] decision in any way.*" One participant did not think that having forecasts from multiple sources increased her confidence, "*because when they differ [she] kind of feel[s] like [she] lose[s] trust in all of them.*"

Discussion

With the help of the qualitative feedback, we were able to identify the biggest disadvantages that people experienced. First, the weather APIs provided a limited amount of information that we were able to access and show in the application. Second, participants missed a certain feature (e.g., typing in a location) or a certain information (e.g., wind forecast) that they were used to from former usage. We assume that these problems could be solved by providing more features and specific settings allowing users to individualize the application.

Although for most participants, it was not important which APIs we used, one participants wanted to choose the sources to be shown in the application. Increasing transparency by

providing more sources and additional information on what exact data the APIs provide could help users to understand the prerequisites of the weather providers. By choosing a number of providers, they would be able to specify the requirements for the weather forecast on their own (e.g., time of forecast period, additional wind forecast).

Surprisingly, some participants exclusively used the range or the detailed view. From our online survey, we mainly expected participants to use the range aggregation for everyday usage and the detailed view for important events as six participants reported. We assume that although all our participant beforehand stated to use weather forecasts everyday, the weather forecast generally was of different importance to them. Participants that only used the range view apparently had no strong interest in weather forecasts in general while participants that only used the detailed view enjoyed to have the details to be more confident about the forecast. These three participants in general commented highly positively about the application, which shows that they liked using the application as it reduced the time they need for comparing forecasts.

In line with these findings, the majority of participants stated that the aggregation of forecasts from different sources increased their confidence in the forecast and that they would like to continue using this type of service. Interestingly, one participant mentioned that seeing how the forecasts disagree rather makes her lose trust in the forecasts. It seems that although research suggests otherwise, some people are not aware of the uncertainty in weather forecasts or perceive it to be lower than the actual difference between different weather providers. It would be interesting to investigate whether this opinion changes after using the application long-term or when switching back to the weather application used before the study (as this is now maybe also perceived to be more uncertain).

IMPLICATIONS FOR DESIGN AND FUTURE WORK

Based on the results of our online survey and the in-the-wild study, we identified five implications for future work on designing for comparison of uncertain data. These implications do not only apply for weather-related purposes and could help to inform the design of interactive systems in other application areas, for example for arrival time prediction or health predictions.

Support contexts with different importance

Applications should support range aggregation and direct comparison to adapt to contexts with different importance. Although people like aggregated data, experiment 1 showed that users still want to make an informed decision on their own judgment if something is important for them.

Support perception at a glance

For applications such as weather applications, it is crucial to support an aggregation mechanism in combination with a representation that can be perceived at a glance. In experiment 2, users expressed that they do not want to spend too much time with the application in everyday use or even felt that the cognitive load was too high.

Support different types of users

Applications should make switching between aggregation mechanisms easy and give people the choice for what they want to have as standard view. Experiment 2 showed that people have different tolerances for giving control to an application and prefer different aggregation mechanisms as their standard. They may even only use one mechanism.

Give opportunities for choice

Applications should offer users a list of many providers to choose a default number of sources (e.g., 3) to be displayed. This increases the feeling of control as users have a choice to make their own decision on which providers they want to rely on. In experiment 2, users stated that they would like to have a choice or include their current source.

Establish transparency

In experiment 2, users requested a longer range of the forecast and specific information (e.g., wind forecast). By providing details on what data the original sources provide and how this data is aggregated in the application, users would better understand the requirements of the application. In combination with giving users the opportunity to select their sources, users can fulfill their specific requirements by only picking matching sources.

CONCLUSION

In this paper, we identified and evaluated aggregation mechanisms for uncertain data from multiple sources using an online survey and an in-the-wild study. As uncertainty is difficult to quantify and often not communicated at all, users compare multiple sources as a workaround to assess the uncertainty. We demonstrated that interactive applications can help to make comparison easier and less cumbersome.

Our results show that aggregation of uncertain data is generally preferred and increases users' confidence, independent of the type of representation used for the data. Computationally aggregated forecasts are more suited for everyday events while manual aggregation is preferred for high-stakes events. Nevertheless, certain types of users prefer to always use one or the other aggregation mechanism based on the personal importance of the topic 'weather'. To satisfy all users' needs, both computational and manual aggregation should be supported. Additionally, computationally aggregated data has to be easily comprehensible. Lastly, transparency of the origin of the data (for example the name of the sources) and choice of sources is an important factor of success.

In future work, these mechanisms could also be applied to non-weather related purposes as for example arrival time predictions, health predictions, or contradicting sensor measurements. We also plan to study the long-term usage behavior of such mechanisms and the influence of long-term usage on users' confidence.

ACKNOWLEDGEMENTS

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart.

REFERENCES

1. Nadia Boukhelifa and David J. Duke. 2009. Uncertainty Visualization: Why Might It Fail?. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 4051–4056. DOI: http://dx.doi.org/10.1145/1520340.1520616
2. David V. Budescu, Stephen Broomell, and Han-Hui Por. 2009. Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science* 20, 3 (2009), 299–308. DOI: http://dx.doi.org/10.1111/j.1467-9280.2009.02284.x
3. Michael Correll and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *Visualization and Computer Graphics, IEEE Transactions on* 20, 12 (Dec 2014), 2142–2151. DOI: http://dx.doi.org/10.1109/TVCG.2014.2346298
4. AMS Council. 2008. Enhancing weather information with probability forecasts. *Bulletin of the American Meteorological Society* 89 (2008), 1049–1053.
5. National Research Council. 2003. *Communicating Uncertainties in Weather and Climate Information: A Workshop Summary*. The National Academies Press. DOI: http://dx.doi.org/10.17226/10597
6. National Research Council. 2006. *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. The National Academies Press. DOI: http://dx.doi.org/10.17226/11699
7. David Demeritt, Hannah Cloke, Florian Pappenberger, Jutta Thielen, Jens Bartholmes, and Maria-Helena Ramos. 2007. Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environmental Hazards* 7, 2 (2007), 115 – 127. DOI: http://dx.doi.org/10.1016/j.envhaz.2007.05.001
8. Nivan Ferreira, Danyel Fisher, and Arnd C. Konig. 2014. Sample-oriented Task-driven Visualizations: Allowing Users to Make Better, More Confident Decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 571–580. DOI: http://dx.doi.org/10.1145/2556288.2557131
9. J. Frick and C. Hegg. 2011. Can end-users' flood management decision making be improved by information about forecast uncertainty? *Atmospheric Research* 100, 23 (2011), 296 – 303. DOI: http://dx.doi.org/10.1016/j.atmosres.2010.12.006
10. Susan Joslyn, Karla Pak, David Jones, John Pyles, and Earl Hunt. 2007. The effect of probabilistic information on threshold forecasts. *Weather and Forecasting* 22, 4 (2007), 804–812. DOI: http://dx.doi.org/10.1175/WAF1020.1
11. Susan Joslyn and Sonia Savelli. 2010. Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorological Applications* 17, 2 (2010), 180–195. DOI: http://dx.doi.org/10.1002/met.190
12. Susan L. Joslyn and Jared E. LeClerc. 2012. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of Experimental Psychology: Applied* 18, 1 (2012), 126–140. DOI: http://dx.doi.org/10.1037/a0025185
13. Malte F. Jung, David Sirkin, Turgut M. Gür, and Martin Steinert. 2015. Displayed Uncertainty Improves Driving Experience and Behavior: The Case of Range Anxiety in an Electric Car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2201–2210. DOI: http://dx.doi.org/10.1145/2702123.2702479
14. Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291. DOI: http://dx.doi.org/10.2307/1914185
15. Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. When (Ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5092–5103. DOI: http://dx.doi.org/10.1145/2858036.2858558
16. Matthew Kay, Dan Morris, M.C. Schraefel, and Julie A. Kientz. 2013. There's No Such Thing As Gaining a Pound: Reconsidering the Bathroom Scale User Interface. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 401–410. DOI: http://dx.doi.org/10.1145/2493432.2493456
17. Matthew Kay, Shwetak N. Patel, and Julie A. Kientz. 2015. How Good is 85Classifier Evaluation to Acceptability of Accuracy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 347–356. DOI: http://dx.doi.org/10.1145/2702123.2702603
18. H. Kootval. 2008. Guidelines on communicating forecast uncertainty. *World Meteorological Organization/Technical Document* 4122 (2008).
19. Jeffrey K. Lazo, Rebecca E. Morss, and Julie L. Demuth. 2009. 300 Billion Served. *Bulletin of the American Meteorological Society* 90, 6 (2009), 785–798. DOI: http://dx.doi.org/10.1175/2008BAMS2604.1
20. Isaac M. Lipkus, Greg Samsa, and Barbara K. Rimer. 2001. General Performance on a Numeracy Scale among Highly Educated Samples. *Medical Decision Making* 21, 1 (2001), 37–44. DOI: http://dx.doi.org/10.1177/0272989X0102100105

21. Alan M. MacEachren, Anthony Robinson, Susan Hopper, Steven Gardner, Robert Murray, Mark Gahegan, and Elisabeth Hetzler. 2005. Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know. *Cartography and Geographic Information Science* 32, 3 (2005), 139–160. DOI: http://dx.doi.org/10.1559/1523040054738936

22. Rebecca E. Morss, Julie L. Demuth, and Jeffrey K. Lazo. 2008. Communicating Uncertainty in Weather Forecasts: A Survey of the U.S. Public. *Weather and forecasting* 23, 5 (2008), 974–991. DOI: http://dx.doi.org/10.1175/2008WAF2007088.1

23. Rebecca E. Morss, Jeffrey K. Lazo, and Julie L. Demuth. 2010. Examining the use of weather forecasts in decision scenarios: results from a US survey with implications for uncertainty communication. *Meteorological Applications* 17, 2 (2010), 149–162. DOI: http://dx.doi.org/10.1002/met.196

24. Limor Nadav-Greenberg and Susan L. Joslyn. 2009. Uncertainty Forecasts Improve Decision Making Among Nonexperts. *Journal of Cognitive Engineering and Decision Making* 3, 3 (2009), 209–227. DOI: http://dx.doi.org/10.1518/155534309X474460

25. Neville Nicholls. 1999. Cognitive Illusions, Heuristics, and Climate Prediction. *Bulletin of the American Meteorological Society* 80, 7 (1999), 1385–1397. DOI: http://dx.doi.org/10.1175/1520-0477(1999)080<1385:CIHACP>2.0.CO;2

26. Chris Olston and Jock D. Mackinlay. 2002. Visualizing data with bounded uncertainty. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*. 37–40. DOI: http://dx.doi.org/10.1109/INFVIS.2002.1173145

27. Alex T. Pang, Craig M. Wittenbrink, and Suresh K. Lodha. 1997. Approaches to uncertainty visualization. *The Visual Computer* 13, 8 (1997), 370–390. DOI: http://dx.doi.org/10.1007/s003710050111

28. Kristin Potter, Joe Kniss, Richard Riesenfeld, and Chris.R. Johnson. 2010. Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum* 29, 3 (2010), 823–832. DOI: http://dx.doi.org/10.1111/j.1467-8659.2009.01677.x

29. Kristin Potter, Paul Rosen, and Chris R. Johnson. 2012. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*. Vol. 377. Springer, 226–249. DOI: http://dx.doi.org/10.1007/978-3-642-32677-6_15

30. Mark S. Roulston, Gary E. Bolton, Andrew N. Kleit, and Addison L. Sears-Collins. 2006. A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting* 21, 1 (2006), 116–122. DOI: http://dx.doi.org/10.1175/WAF887.1

31. Mark S. Roulston and Todd R. Kaplan. 2009. A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. *Meteorological Applications* 16, 2 (2009), 237–244. DOI: http://dx.doi.org/10.1002/met.113

32. Orit Shaer, Oded Nov, Johanna Okerlund, Martina Balestra, Elizabeth Stowell, Lauren Westendorf, Christina Pollalis, Jasmine Davis, Liliana Westort, and Madeleine Ball. 2016. GenomiX: A Novel Interaction Tool for Self-Exploration of Personal Genomic Data. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 661–672. DOI: http://dx.doi.org/10.1145/2858036.2858397

33. Michael Siegrist, Heinz Gutscher, and Timothy C. Earle. 2005. Perception of risk: the influence of general trust, and general confidence. *Journal of Risk Research* 8, 2 (2005), 145–156. DOI: http://dx.doi.org/10.1080/1366987032000105315

34. Meredith Skeels, Bongshin Lee, Greg Smith, and George G. Robertson. 2010. Revealing Uncertainty for Information Visualization. *Information Visualization* 9, 1 (2010), 70–81. DOI: http://dx.doi.org/10.1057/ivs.2009.1

35. Paul Slovic. 1987. Perception of risk. *Science (Washington, D.C.); (United States)* 236 (Apr 1987). DOI: http://dx.doi.org/10.1126/science.3563507

36. Susanne Tak, Alexander Toet, and Jan van Erp. 2014. The Perception of Visual Uncertainty Representation by Non-Experts. *Visualization and Computer Graphics, IEEE Transactions on* 20, 6 (June 2014), 935–943. DOI: http://dx.doi.org/10.1109/TVCG.2013.247

37. Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. 2005. A typology for visualizing uncertainty. In *Electronic imaging 2005*. International Society for Optics and Photonics, 146–157. DOI: http://dx.doi.org/10.1117/12.587254

38. Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. http://www.jstor.org/stable/1738360

39. Thomas S. Wallsten, David V. Budescu, Amnon Rapoport, Rami Zwick, and Barbara Forsyth. 1986. Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General* 115, 4 (1986), 348–365. DOI: http://dx.doi.org/10.1037/0096-3445.115.4.348

40. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 143–146.

41. Jianlong Zhou, Constant Bridon, Fang Chen, Ahmad Khawaji, and Yang Wang. 2015. Be Informed and Be Involved: Effects of Uncertainty and Correlation on User's Confidence in Decision Making. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 923–928. DOI: http://dx.doi.org/10.1145/2702613.2732769

42. Torre Zuk and Sheelagh Carpendale. 2006. Theoretical analysis of uncertainty visualizations. *Proc. SPIE* 6060 (2006), 606007–606007–14. DOI: http://dx.doi.org/10.1117/12.643631