When (ish) is My Bus? User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems

Matthew Kay
CSE | dub
University of Washington
mjskay@uw.edu

Tara Kola
Computer Science
Tufts University
tara.kola@tufts.edu

Jessica R. Hullman
iSchool | dub
University of Washington
jhullman@uw.edu

Sean A. Munson
HCDE | dub
University of Washington
smunson@uw.edu

ABSTRACT
Users often rely on realtime predictions in everyday contexts like riding the bus, but may not grasp that such predictions are subject to uncertainty. Existing uncertainty visualizations may not align with user needs or how they naturally reason about probability. We present a novel mobile interface design and visualization of uncertainty for transit predictions on mobile phones based on discrete outcomes. To develop it, we identified domain specific design requirements for visualizing uncertainty in transit prediction through: 1) a literature review, 2) a large survey of users of a popular realtime transit application, and 3) an iterative design process. We present several candidate visualizations of uncertainty for realtime transit predictions in a mobile context, and we propose a novel discrete representation of continuous outcomes designed for small screens, quantile dotplots. In a controlled experiment we find that quantile dotplots reduce the variance of probabilistic estimates by ~1.15 times compared to density plots and facilitate more confident estimation by end-users in the context of realtime transit prediction scenarios.

Author Keywords
End-user visualization; transit predictions; mobile interfaces; dotplots; uncertainty visualization.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

INTRODUCTION
Quantitative predictions are increasingly ubiquitous in everyday life. Many such data come in the form of point estimates designed to aid decision-making, such as when the next bus is going to arrive, how long a road trip will take, whether and when it will rain, or what the high temperature will be. Often, people access these predictions on their mobile phones to make in-the-moment decisions that are time-constrained (providing little opportunity for training, interpretation, or complex interaction) using interfaces that are space-constrained (due to screen size).

For example, Susan might refer to a bus's predicted arrival time on a smartphone application to check if she has time to get coffee before her bus to work arrives. She sees that the bus is running a few minutes late and is predicted to arrive in five minutes. There is no line at the coffee shop, so she steps in to order. However, the bus makes up lost time and arrives only two minutes later: Susan, still waiting for coffee, misses her bus and is late for a meeting.

Susan based her decision on a point estimate of arrival time, as presented in many predictive systems for bus arrival, flight time, or car travel. Her decision is reasonable given the point prediction she saw, but real-world predictions are subject to uncertainty (e.g., her bus is most likely to come in 5 minutes but may come in as little as 1 minute or as much as 9 minutes). Designers and analysts are responsible for reporting uncertainty with predictions to help people make decisions that align with their goals [5,33], yet most visualizations of predictions present the data as if it were true (Finger & Bizantz [10] as cited in Cook & Thomas [5]). Had Susan's application presented her with a more complete representation of the predicted arrival time—perhaps noting that arrival times earlier than 5 minutes are also quite probable—she may not have risked getting coffee.

Many attempts to communicate uncertainty rely on complex visual representations of probability distributions. For example, error bars and probability densities require prior experience with statistical models to correctly interpret [2,6]. People can better understand probabilistic information when it is framed in terms of discrete events. For instance, Hoffrage & Gigerenzer [16] found that more medical experts could accurately estimate the positive predictive value (precision) of a test when presented with discrete counts or outcomes. Discrete-event representations have been used to improve patient understanding of risk, e.g., by showing the uncertainty in a medical diagnosis as discrete possible outcomes (number of true positive, false positives, false negatives, and true negatives) [11]. However, visualizing discrete approaches to presenting probability distributions typically requires a large amount of space or time to communicate the set of possible outcomes [17]. It is not

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05 $15.00
DOI: http://dx.doi.org/10.1145/2858036.2858558

clear how to effectively communicate discrete outcomes on small screens such as mobile devices.

We study user needs for communicating predictions and then design and evaluate novel, goal-directed visualizations of hypothetical outcomes in a time- and space-constrained mobile application, realtime transit prediction. As a setting where users have direct, day-to-day experience with uncertainty, transit prediction provides a representative context in which to evaluate how well people can use different uncertainty visualizations. The specific goals of our work are to see if and for what reasons people want uncertainty information in a bus setting; to identify effective uncertainty visualizations for realtime decision-making on a smartphone; and to test for differences in how precisely and confidently people extract probabilities from different visualizations of uncertainty.

Our three contributions based on these goals are to:

1. Develop **general and domain-specific design requirements and a rich description of user needs** for visualizing uncertainty in transit arrival times based on (i.) an analysis of the literature and (ii.) an initial survey of 172 people who use a popular realtime transit application.
2. Propose **design layouts and discrete-event visualizations of uncertainty** for conveying bus arrival time predictions on small screens based on an iterative design process. We introduce **quantile dotplots**, a novel modified dotplot that is a discrete analog to the common probability density plot.
3. Identify through a large user study of transit application users how accuracy and precision in estimating probability compares across several visualizations. Unlike previous work in communicating uncertainty in continuous outcomes [6,14], our study is the first to compare static discrete-outcome visualizations of probability distributions to continuous representations. We find that a **quantile dotplot depicting a small number of outcomes has ~ 1.15 times lower variance than a density plot**, making probability estimates 1-3 percentage points more precise.

Our results further understanding of how to communicate prediction uncertainty to non-experts in everyday, mobile decision-making contexts. Specifically, we recommend using low-density dotplots (due to lower variance and higher user confidence) or density plots (which have only slightly higher variance, but were more visually appealing) for visualizing uncertainty in space-constrained environments.

BACKGROUND & MOTIVATION

In this section, we use prior work to establish baseline requirements for the effective communication of uncertainty.

Improving trust by communicating uncertainty

Research indicates that displaying uncertainty can improve trust and decision-making in everyday contexts. Kay *et al.* [23] studied trust in body weight measurements, arguing that single point estimates without uncertainty decrease trust and tend to be interpreted as being more precise than they actually are. Kay *et al.* suggest **avoiding false precision in single point estimates** by displaying the uncertainty associated with weight data to improve trust. Similarly, Jung *et al.* [21] found that displaying the estimated remaining range of an electric vehicle as a gradient plot (i.e., with uncertainty) reduced range anxiety in a driving task compared to a single point estimate.

Joslyn & LeClerc [19] found that displaying uncertainty in weather predictions can lead to more optimal decision-making and trust in a forecast. When asked to make decisions about whether to salt roads (given a virtual budget, cost for salting, and cost for failing to salt when they should have), people made more optimal decisions when given point estimates with probabilistic information. Subjects with access to probabilistic information even made more optimal decisions than subjects who were explicitly told the optimal decision based on a cot-benefit analysis. While the decision suggested by a cost-benefit analysis will give the best choice on average, always applying the decision will sometimes lead people to take precautions that seem unnecessary (e.g., salting the roads when the weather ultimately does not require it). After experiencing a such few errors, people may begin to distrust the strategy and ignore the suggested course of action. Probabilistic information, on the other hand, provides a more transparent form of information for decision making, leading to greater trust. We believe this insight also applies to realtime transit prediction: even if we could develop a system to make recommendations like "leave now to make your bus on time", **a more transparent communication of uncertainty** will maintain trust over the long term and leave people the agency to make mistakes.

Visualizing uncertainty

As extrinsic annotation

A common approach to visualizing uncertainty is as an *extrinsic* annotation to a plot of the distribution's location (mean, median, or mode). For example, error bars representing confidence intervals or prediction intervals[1] can be superimposed on bar charts [2]. These intervals are extrinsic to other properties like the mean or mode since they are not integrated into the same encoding. By contrast, the probability density and mode are intrinsic to each other in a density plot, since the mode is visually encoded as the maximum of the density. Other distributional properties may be represented in summary plots using a series of marks (e.g., specific quantiles in a boxplot or modified box plot as in [6,18,28]).

[1] In contrast to a *confidence interval*, which describes the precision of an inferred model parameter (e.g. population mean), a *prediction interval* is an interval that a given percentage of specific instances are predicted to fall into. While much of the literature focuses on confidence intervals (of interest to scientists using models for inference), we are more concerned with prediction intervals (of interest to an individual who wishes to know how likely their bus is to arrive in a specific instance).

Extrinsic representation can result in interpretation errors because the statistical construct (such as one standard error or a 95% confidence interval) is poorly understood [2,18], because individuals apply heuristics that are not correct (such as assuming that overlapping confidence intervals always indicate a non-significant difference [7]), or because the representation is ambiguous (such as an error bar being used to encode standard deviation, standard error, or 95% confidence interval). Finally, individuals tend to underweight probabilistic information (such as sample size or variance) when making judgments in favor of heuristic attributes like representativeness [35]. By separating the marks encoding underlying data from those encoding uncertainty, extrinsic representations are at risk of being viewed as peripheral, and consequently discounted when making judgments. Thus, to avoid ambiguity, simplify interpretation, and encourage users not to underweight probability information, we believe that **uncertainty should be intrinsic to the representation.**

Further, while a given prediction interval corresponds to a specific risk tolerance, a user may have differing risk thresholds in different contexts. For example, Susan may be willing to be late to her meeting 1/20 times, translating to a one-sided 95% prediction interval for estimated arrival time, yet she may tolerate more risk in different contexts like social gatherings or less important meetings. Different individuals are also likely to have different risk tolerances. Therefore, we believe that effective visualizations of uncertainty in this context should **allow users to apply situation-dependent risk tolerance**.

As abstract, continuous outcomes

Many other abstract, static representations encode a predictive distribution's probability density function (PDF) intrinsically as *retinal variables* (e.g., color, shape, texture) [3]. For example, density plots encode the PDF as distance from the *x*-axis, violin plots encode it as width [1,22,31], and gradient plots encode it as opacity. Several studies that include variants of density and gradient plots find little evidence of a performance difference between the two [6,18]. Opacity is a less effective encoding than height, width, or area [26]. As a result, we do not test the gradient plot.

Not all encodings of continuous outcomes using retinal variables make distributional properties intrinsic. Ibrekk & Morgan [18] compare density plots to plots of cumulative density functions (CDFs), amongst several other encodings. CDFs encode cumulative density as distance from the *x*-axis, allowing the probability of intervals to be estimated from height. They found that CDFs were unfamiliar to participants and required training. Not surprisingly, people had particular difficulty in using CDFs to estimate means, most likely because there is no simple visual variable that corresponds to mean (nor mode) on a CDF.

As hypothetical, discrete outcomes

We use *discrete outcomes* to refer to techniques that employ draws from a probability distribution rather than abstract probabilities of events. Discrete approaches were initially found to improve reasoning in textual communication. Gigerenzer and Hoffrage [12] found that statistical word problems described in terms of *natural frequencies* (e.g., 10/100) rather than probabilities (10%) were more likely to elicit inferences according to Bayes' rule in laypeople. Past work on visualizing uncertainty through hypothetical, discrete outcomes uses spatial or temporal bandwidth to communicate. For example, Garcia-Retamero and Cokely [11] reviewed studies of several types of visual aids for communicating health risks, including discrete outcome charts that illustrate treatment risk: they found that displaying *icon arrays* (a grid of pictograms, each representing a patient who lived or died) improved the accuracy of people's risk assessment. Hullman *et al.* use animation to display discrete outcomes more compactly in space [17], finding that animated discrete outcomes (called *hypothetical outcome plots*, or HOPs) support more accurate probability estimates than static alternatives (violin plots and error bars) for some tasks. However, by presenting outcomes over time, animated techniques bring a time-precision trade-off: to make more precise inferences, a user must view more outcomes, taking more time [ibid].

From the evidence in both visual communication and statistical reasoning, we believe that **discrete outcomes can improve decision making under uncertainty**. However, because transit decisions are often made quickly in real time we focused on developing non-animated presentations of discrete outcomes that are **glanceable** yet compact enough for a mobile phone display, in which it is typical to visualize the upcoming arrival of ~10 buses on one screen [9].

Visualization in space-constrained environments

To display many buses simultaneously on a mobile phone screen, we require our visualizations to be **compact**. Techniques like horizon graphs [15] and sparklines [34] have been proposed for visualizing time-series data in space-constrained environments. Visualizing uncertainty in transit arrival predictions encounters similar issues as these approaches; for example, a probability density function of predicted arrival time will become quite tall as its variance decreases (particularly, close to the predicted arrival time the prediction will become very precise). Our work demonstrates possible solutions for the specific context of visualizing PDFs on mobile phone displays.

SURVEY OF EXISTING USERS

Our reading of the literature provides an initial grounding for our design work, but to apply these results to a user-centered uncertainty visualization we also need to understand user goals. To establish design criteria for representations of uncertainty based on user needs, we surveyed users of one popular realtime transit application, OneBusAway [9].

Method

We conducted a survey to identify 1) how users currently use realtime bus arrival predictions and 2) their unaddressed needs for goal-oriented uncertainty information. We sur-

veyed 172 users of OneBusAway, recruited via social media and department mailing lists.

Users' existing goals

To identify important user scenarios to address and what types of information are most important to those scenarios, we asked people about the *primary goals* they have when using OneBusAway. We developed a set of possible questions (e.g., "When should I start walking to the bus stop to catch my bus?"[2]) that people may ask using the interface using observations from previous studies of OneBusAway [20], our own reflections on using the system, from informal interviews with a small group (~15) of other users at our university, and through piloting the survey. We presented participants with a list of 9 such questions, and asked how often (on a 7-point scale from "never" to "always") they try to answer each question using OneBusAway. We also asked them if there are other ways they use OneBusAway in an open-ended question.

Problems with OneBusAway and unaddressed needs

We similarly presented participants with a list of types of information not currently provided by OneBusAway and asked them to rate the *potential helpfulness* of these (on a 5 point scale from "not helpful at all" to "very helpful"). We also provided an open-ended question asking about needs for uncertainty information not in this list. Finally, we asked people to describe the *worst experience* they have had using OneBusAway's predictions.

Results and Discussion

Users' existing goals

The top 5 highest-rated questions users currently ask are:

- **When to leave**: When should I start walking to my bus?
- **Wait time**: If I leave now, how long will I have to wait at the bus stop?
- **Time to next bus**: I missed my bus, how long will I have to wait for the next one to come?
- **Schedule risk:** Will I get to a meeting/event on time despite bus delays? This relates to a commonly-described worst experience of buses coming later than expected. For example:

  A more recent bad experience was when I was waiting for the 511 or 512 for over an hour. At least five buses should have passed, but they either did not show up or they were full and didn't let anyone on

- **Schedule opportunity:** Will I have enough time to do ____ before the bus arrives? This relates to a commonly-described *worst experience* of the bus coming earlier than expected after someone has used the prediction to decide to do something else before going to the bus; e.g.:

  It showed delays on a bus due to which I didn't leave home as I didn't want to wait at the bus stop for long (the bus stop is 4 mins from my home), but it suddenly came on time and I missed it. Sometimes, it even comes early when it shows delay.

Problems with OneBusAway and unaddressed needs

The top three questions users would like to be able to ask, but which are not well-supported by the current OneBusAway interface, are:

- **Status probability:** What is the chance OBA is showing the correct arrival status? This problem was also reflected in a commonly-described *worst experience*, wherein the bus never shows up and people have to make alternative plans. For example:

  My bus is perpetually 9 minutes away...while I watch alternative buses pass me thinking that oh, mine is going to be here soon only to eventually see "no information" for my bus. I could have been on my backup bus a half hour ago!!!

  It was common for people to report their worst experiences were related to status probability: for example, OneBusAway said "departed", but the bus had not arrived; it said "arriving" but had already departed. Any noisy estimate reduced to a categorical status will exhibit these types of errors which could be mitigated by conveying status probabilistically.
- **Prediction variance:** What is the chance the predicted arrival time will change unexpectedly?
- **Schedule frequency:** How frequently do buses arrive at various times in the day?

DESIGN REQUIREMENTS

Based on our literature review and user survey, we identified the following necessary design elements:

Point estimate of time to arrival: To support *glanceability*, we think that the point estimate of arrival time is necessary: people often use OneBusAway to make fast decisions about when to arrive at the bus stop. In addition, previous work has found that even when providing probabilistic estimates, people still want a point estimate. The existing point estimate of OneBusAway supports estimation tasks from our survey like *when to leave*, *wait time*, and *time to next bus*, though without communicating risk.

Probabilistic estimate of time to arrival: While people often want a point estimate of arrival time, a point estimate without uncertainty will often convey a *false precision*. A probabilistic estimate will help users understand that there is a chance the bus will come earlier or later than the point estimate. This helps people assess *schedule risk* and *schedule opportunities*. A probabilistic estimate also allows people to make conservative estimates while planning for meetings, or less conservative estimates for low risk situations – that is probabilistic estimates *allow situation-dependent risk tolerance*. This will help people better answer questions about *when to leave*, *wait time*, and *time to next bus* (the highest rated goals) and prepare people for commonly-reported *worst experiences* like a bus coming unexpectedly early or late.

[2] The full survey is available in our supplementary material.

Probabilistic estimate of arrival status: For example, what is the chance the bus has already arrived? Among questions not currently supported by OneBusAway, survey respondents most wanted support for this question (*status probability*), and commonly reported *worst experiences* related to it.

Data freshness: Because OneBusAway does not currently give probabilistic estimates, one of the only available signals for expert users to assess risk is the *freshness* of the data: OneBusAway indicates the time of the last update for realtime predictions and whether the current prediction is based on realtime data (it reflects the scheduled arrival time when realtime data is not available). This freshness information should either be provided to users in a redesigned interface, or should be incorporated into any models driving probabilistic estimates.

We believe these design elements will address each goal identified in the user survey with the exception of the goal of knowing *schedule frequency*. We felt that this goal is better addressed through a separate interface, such as a trip planner or schedule explorer in a mapping application. Schedule frequency is less relevant to in-the-moment decision-making than it is to long-term planning (can I rely on a bus arriving within some amount of time?). When schedule frequency is relevant to in-the-moment decisions, it typically reduces to other goals, like *time to next bus*.

DESIGN

We conducted an iterative design process focused on the design requirements set out above. This process began with a wide exploration of ideas through sketching, followed by paper prototyping in increasing fidelity, and culminated in digital mockups. These phases were informed by ongoing user feedback gained through informal down-the-hall testing with a total of 24 users. During informal testing, we presented users with hypothetical scenarios of use and asked them to think aloud as they interpreted the display.

Many of the design issues we encountered are somewhat orthogonal to specific of encodings of probability: given a particular timeline layout, for example, we could encode probability in many ways (e.g., as area, discrete events, a gradient). We first present our proposed set of designs and their rationale, then discuss possible techniques for encoding probability on small screens.

Proposed designs and rationale

Our proposed designs, instantiated with one particular visualization of uncertainty (density plot) out of several possible, are shown in Figure 1. Here we describe decisions we made to resolve design tensions and to match user goals.

Different layouts better serve different use cases

We developed two alternative layouts, *bus-timeline* and *route-timeline*. The *bus-timeline* layout gives a timeline for a single bus on each row, similar to how the existing OneBusAway app displays a single row per bus, sorted by predicted time to arrival. This simplifies understanding and

Figure 1. Alternative layouts we developed. (a) Bus Timeline: Each row (timeline) shows one predicted bus. (b) Route Timeline: Each row shows all predicted buses from a given route.

navigation, but is less compact in addressing problems like assessing *schedule frequency,* and, once the probabilistic visualizations are added, less compact than the current application. *Route-timeline*, by contrast, creates a more complex display and navigation (requiring navigation in two dimensions), but more easily aids understanding of *schedule frequency* (how often is the bus) and *schedule opportunity* (since if one is considering the risk associated with missing the next bus, it is easier to see how soon the bus after that is coming and factor that into one's decision).

Point estimates and probabilistic estimates should coincide spatially

We explored several tradeoffs between prominent point estimates versus probabilistic estimates, what we call the **glanceability/false precision tradeoff**. A too-prominent display of the point estimate causes users to ignore the probabilistic one, thus still giving a false sense of precision; a less-glanceable point estimate will be difficult to skim and frustrating to use. We want a display that is *glanceable* but which also does not convey false precision. To resolve this, we concluded that these two elements should coincide spatially: that is, *looking at the point estimate should encourage the user to also be looking at the probabilistic estimate.* We had considered designs in which the point estimate was along the right-hand edge of the display (Figure 3), as in the original OneBusAway. We concluded that this facilitated glanceability, but also allowed users to pay too little attention to the probabilistic estimates. Moving the point estimate onto the probability distribution resolved this tension.

Annotated timelines give probabilistic estimates of status "for free"

While we considered designs that more explicitly communicate the probability that the bus has arrived, we realized that an annotated timeline combined with probabilistic predictions communicates this implicitly. By denoting areas that correspond to "departed", "now", and "on the way" on the timeline, users can directly read these probabilities from the distributions depicted; see the timeline annotations across the top of Figure 1.

When to leave is implicit in time to arrival

We considered designs that communicated when someone should leave to catch their bus; i.e. designs that directly addressed the *when to leave* goal. However, there are several difficulties with this approach: first, when to leave is not the only goal for which people use OneBusAway; thus it would need to be integrated into displays communicating information like time to arrival (or alternate designs developed for both goals). This exacerbates space issues. Estimating when to leave also requires substantial knowledge about the users' plans, and introduces further uncertainty (e.g., how long does it take to walk to the stop?).

Data freshness may be subsumed by an improved model

OneBusAway often does not have truly realtime information, but instead updates when buses check in. As noted previously, expert users often refer to the last check-in time as a way to evaluate how much they trust the application's prediction. To facilitate this use, we considered several designs that included indicators of data freshness or last update times. Ultimately we decided not to include this information, as the model used to generate the probabilistic arrival information should **take data freshness into account to provide better estimates to all users**, rather than continuing to support a workaround used by expert users.

Synchronized timelines allow comparison between buses

In our designs, the axis of the timeline in each row is synchronized to the other rows, facilitating comparison between buses. We considered designs with each row having its own time range depending on the prediction (e.g., one row with low variance might show a density plot covering 5-10 minutes from now; another with high variance might have an axis covering 5-15 minutes from now). However, such relative timelines are very difficult to compare between buses on different rows—buses with different variance might look similar because the relative timeline would also cause the density to be scaled.

Encoding probability in space-constrained environments

Given our chosen design, we need an effective way to encode probability at small sizes. We considered several approaches (Figure 4). Most of these are drawn from the literature, including density plots, violin plots, and gradient plots. We also propose variants of two existing discrete plots for visualizing predictive distributions as discrete outcomes, stripeplots and dotplots.

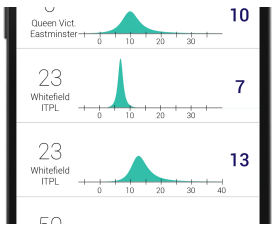

Figure 3. An example of a design we rejected for placing point predictions (along the right side) outside the context of uncertainty, making it more likely to give users a false sense of precision.

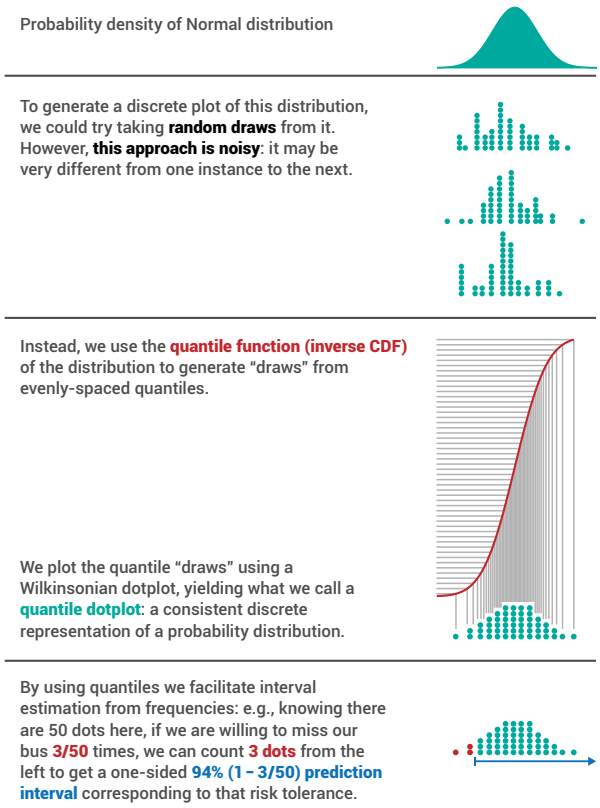

Figure 2. Explanation of quantile dotplots.

Discrete outcome visualizations of continuous variables

We explored several ways to convey a continuous predictive probability distribution as discrete outcomes. The first is based on Wilkinson's dotplots [37], which are typically used to communicate the distribution of experimental samples (e.g., [27]). We instead adopt these plots to display theoretical quantiles from a predictive distribution. As Wilkinson notes, correctly-produced dotplots have the desirable property of also conveying the density of the distribution. Our **quantile dotplots** have this property, as well as the additional property of allowing direct estimation of arbitrary (to a certain precision) predictive intervals through counting (see Figure 2). We believe that this form of natural reasoning about predictive intervals—as frequencies—should allow people to obtain precise estimates of predictive intervals in a way that is easily understood.

We also use **stripeplots** [8] of theoretical quantiles to communicate a continuous probabilistic prediction as hypothetical outcomes. In these, the density of stripes in a region encodes probability density, and as in quantile dotplots (though less easily), predictive intervals can be estimated directly through counting. Where dotplots are a discrete analog to a density plot, stripeplots can be thought of as the discrete analog to a gradient plot.

Tight densities require special attention on small screens

Displaying many rows of predictions on a small screen necessitates relatively small row height. Unfortunately, distributions with low variance will become very tall, exceeding

the row height. Traditional solutions include horizon charts [15] (which we suspect are unfamiliar to lay users), or normalizing all density plots to the same height (which makes comparison difficult). This problem is most pronounced on buses with tight variance, i.e., the most precise predictions. Consequently, for density plots we adopted the compromise approach of scaling down the max height only when it exceeds the row height. This adjustment affects only the predictions of which the model is most certain, so fine-grained resolution of probability becomes less important to most goals. This adjustment is required only for *density*, *dotplot-50*, and *dotplot-100* (in the dense dotplots, instead of scaling we reduce the dot-spacing). Dotplot-20 and stripeplot have the advantage of a *consistent representation of probability in tight densities*: they need not be modified.

Countability may vary from tails to body

Care must be taken in deciding how many hypothetical *draws* (quantiles) to include in discrete plots. Figure 4 compares some of the tradeoffs here: With few draws, as in *dotplot-20*, it is easy to count the dots in the tails and body of the distribution, but the density is less well-resolved. With many dots, as in dotplot-100, counting in the tails is often still easy, but in the body overwhelming; however, density is very well-resolved.

Selected encodings

To select the encodings to evaluate for our final design, we constructed the matrix shown in Figure 4 comparing various properties of the encodings. We selected *density*, *stripeplot*-50, *dotplot-20*, and *dotplot-100* as representing a wide range of possible trade-offs suggested by this matrix.

EXPERIMENT

We conducted an online survey to evaluate the effectiveness of our designs in conveying uncertainty. The goal of this survey was to assess how well people can interpret probabilistic predictions from the visualizations and to elicit their preferences for how the data should be displayed.

Method

To assess how well people can judge probability from our visualizations, we adopted an approach similar to that of Ibrekk and Morgan [18], who presented various representations of uncertainty for weather forecasts and asked subjects to report probabilities (e.g., snowfall >2 inches, or between 2 and 12 inches).

We created four scenarios based on the goals identified in our user survey, each with two questions about the probability of bus arrival. For example, in one scenario the respondent is waiting for a bus, and must decide if they have enough time to get coffee before the bus arrives. They are asked what the chance is that the bus will arrive 10 minutes or earlier, and respond using a visual analog scale, a 100-point slider from 0/100 to 100/100. We call their response the *estimated p* (in contrast to the *true p*, which we calculate from the underlying probability distribution). A bubble on the response slider shows this chance expressed in all three denominators used by the various visualization types (e.g.

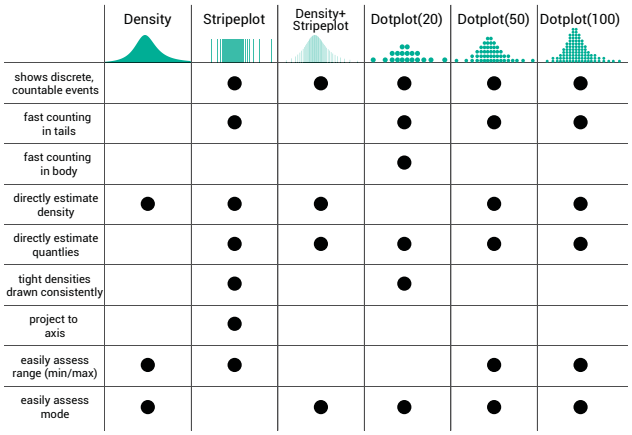

Figure 4. Comparison of various encodings of probability we considered for use in our designs.

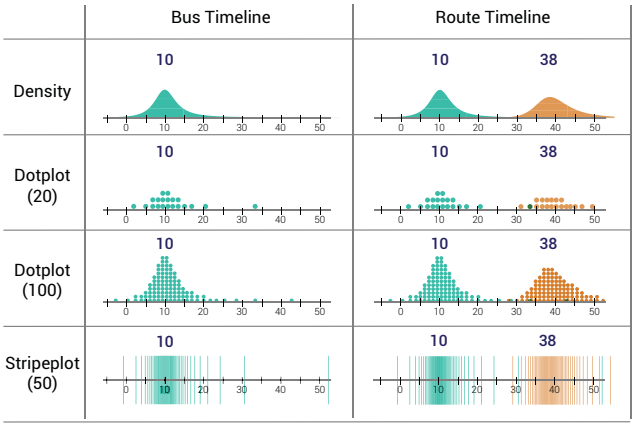

Figure 5. The four types of visualizations selected for evaluation.

"20/100, 10/50, 4/20"), so that participants do not have to do mental arithmetic in the dotplot and stripeplot conditions. The predictions in each scenario were generated from models based on Box-Cox *t* distributions [29] fit to ~2 weeks worth of arrival time data for actual buses in Seattle, but the buses were given fake route names. Participants are also asked how *confident* they are in each probability they estimate. At the end of the survey they rate the *ease of use* and *visual appeal* of each visualization. All subjective ratings are made on 100-point visual analog scales.

Scenario order was randomized between participants. Each participant saw each *visualization type* (*density, stripeplot, dotplot-20, or dotplot-100*) once. Before each scenario, they were also given a brief tutorial explaining the encoding they were about to use. Pairings between scenario and visualization type were also randomized. Participants were also randomly assigned to see all visualizations in the *bus-timeline* or *route-timeline layout*. A full version of the survey can be found in the supplementary material.

Participants

We recruited participants from a variety of locations, including department mailing lists, a local transit blog, and a local forum on reddit.com. Participants were entered into a raffle for 1 $100 Amazon.com gift card and an additional

$25 gift card per 100 participants. Since our primary research questions were about the effect of visualization types, not layout, we ran the first 100 participants only on the *bus-timeline* condition. This threshold was chosen based on a power analysis of data from Hullman *et al.* [17], which suggested a power of at least .8 with our design for detecting similar effect sizes to that study after 100 participants. After reaching 100 participants in the *bus-timeline* layout, the remainder of participants were randomly assigned to either the *bus-timeline* or *route-timeline layout*. After removing 9 participants for incomplete data, we had 320 participants in the *bus-timeline* and 221 participants in the *route-timeline* layouts. Our participants skewed male (71% male). 90% were existing OneBusAway users.

Results

To understand how well each visualization performs, we can examine the *error* in people's probability estimates. We break error into *bias* (do people over- or under- estimate probabilities on average?) and *variance* (how *self-consistent* are people's estimates, whether biased or not?). **So long as the bias is low, we believe that variance is the more important component of error in this task**: over time people can adjust their risk tolerance to a small but consistent bias, but they cannot do so if their estimates are not consistent. We consider overall error, bias, and variance in turn.

Overall error in participants' probability estimates

We start by looking at the overall shape of participants' *estimation error*: logit(*estimated p*) – logit(*true p*) for each question.[4] Figure 6A shows the density of those differences, broken down by visualization type. The *bias* in responses is consistently low and positive across conditions: note that the error distributions all peak in approximately the same place, slightly to the right of 0 (the dashed line). *Variance* appears to be lower in *dotplot-20* compared to the other visualizations: the distribution of error is narrower. However, this does not necessarily equate to more *self-consistent* responses. We therefore use a model to assess bias and variance more systematically and to account for within participant effects.

Regression model for bias and variance

We fit a beta regression to participants' estimated probabilities, which assumes responses are distributed according to a beta distribution. This distribution is defined on (0, 1) and naturally accounts for the fact that responses on a bounded interval have non-constant variance.[5] In other words, the variance of *estimated p* changes with the probability being estimated. For example, at *probability* = 0.5 one can guess 0.5 + 0.4 = 0.9; at *probability* = 0.9 one cannot guess 0.9 + 0.4 = 1.3 (it is greater than 1.0), so responses "bunch up" [30] under 1.0 and variance is lower. Beta regression has been shown to be better-suited to this type of data than linear regression [30].

Our regression uses a submodel for the mean (in logit-space) and the dispersion (proportional to variance, in log-space) [30]. This allows us to model the bias of people's *estimated p* as effects on the mean of their responses, and the variance as effects on the dispersion of their responses. Specifically, we include *visualization*, logit(*true p*), and their interaction as fixed effects on mean response. We include *visualization, layout,* and *gender* as fixed effects on the dispersion (in other words, some visualizations or layouts may be harder to use, resulting in more variable responses; and men may be better or worse at this task). We also include *participant* and *participant* × *visualization* as random effects (some people may be worse at this task, or worse at this task on specific visualizations), and *question* as a random effect (some questions may be harder).

We use a Bayesian model, which allows us to build on previous results by specifying prior information for effects, and report results primarily as posterior distributions with 95% credibility intervals (the Bayesian analog to a confidence interval) [24,25]. We derive priors from fitting a similar model to the data from Hullman *et al.*[17], which had a similar task (estimating cumulative probabilities on three visualizations: a violin plot, animated hypothetical outcomes, and error bars). We set Gaussian priors for fixed effects in our model that capture the sizes of effects seen in the Hullman *et al.* data within 1-2 standard deviations, with skeptical means (0 for intercept and 1 for slope in logit-logit space, corresponding to an unbiased observ`er). We use the posterior estimate of the variance of the random effect of participant in that model as the prior for the variance of random effects in our analysis. Full priors and posterior estimates are available with our data. [6]

Bias in respondent probability estimates

Consistent with Figure 6A, our regression found that estimates were slightly biased on average, and these biases were similar across conditions (more details in supplementary material). Our beta regression model accounts for this bias when estimating the variance of participant responses. The slight overestimation here may be because all of our distributions are right-tailed (positively skewed). This is generally true of transit arrival time data; thus, if the skew is the source of this bias we should expect to see this effect in real-world situations in our domain but perhaps not others. Skewness of distributions is known to affect risk aver-

[4] The logit function is an s-shaped function that transforms probabilities into log-odds, often used when to simplify the analysis of probabilities by transforming them onto the unbounded real line.

[5] Because 0 and 1 are not defined in the beta distribution, we treat answers of 0 and 1 from our visual analog scales as 0.001 and 0.999.

[6] Note that similar results were obtained using more default priors, showing our results are not highly sensitive to choice of priors here. The model was fit using Stan [32], with 16 chains having 20,000 iterations each (half warmup), thinned at 8, for a final sample size of 20,000. Parameters of interest all had effective sample sizes > 10,000 and potential scale reduction factor < 1.001. See https://github.com/mjskay/when-ish-is-my-bus for survey data and code (DOI: 10.6084/m9.figshare.2061876)

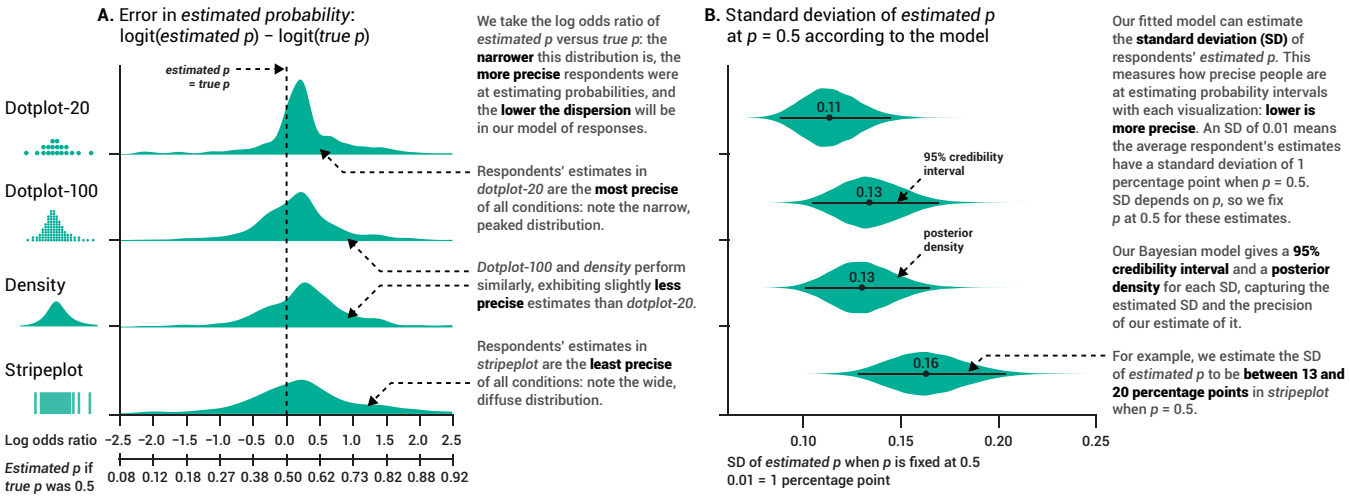

Figure 6. Variance of respondent estimates of probability intervals, (A) as raw data and (B) as estimated by our model.

sion in financial decisions made from density plots [36]; these biases may be related.

Variance in participant probability estimates

As noted above, we believe that variance is more important than bias in this task, as low variance would allow people to adjust their behavior to a consistent bias over long-term usage. We estimate the variance associated with each visualization as a standard deviation in *estimated p* if *p* is fixed at 0.5 (Figure 6B). Figure 7 shows pairwise comparisons of SD for all visualizations (pair specified in left column). *Dotplot-20* has the lowest estimated variance (SD of ~11 percentage points), being about 1.15 times more precise than *density* plots. By contrast, *dotplot-100* has similar variance to *density*, consistent with people estimating area instead of counting dots, perhaps because there are more dots than they may be willing to count.

Confidence

Ideally, greater confidence in a given answer would be associated with less error, indicating that people are able to self-assess their accuracy. We used a similar beta regression to model *confidence* in estimates depending on *visualization*. Participants expressed higher confidence in their estimates on average in the *dotplot-20* condition (mean = 81/100, 95% CI: [77, 83]) than the next-most-confident condition, *dotplot-100* (mean = 73, 95% CI: [71, 76]). At the same time, confidence in the *dotplot-20* condition correlated negatively with *absolute estimation error* (Spearman's $\rho = -0.18$, 95% CI: [−0.13, −0.25]), an association we did not see in other conditions. At least with *dotplot-20*, people have some ability to assess how good their own estimates are. We suspect that this may be due to the fact that with *dotplot-20* one can choose either to be precise (by counting dots) or to give a less precise, less confident answer (by approximating density or area instead of counting).

Ease of use and visual appeal

We also analyzed ease of use and visual appeal using beta regression. *Density* had the highest visual appeal (mean = 66, 95% CI: [64, 67]); *dotplot-20* was less visually appealing (mean = 43, 95% CI: [42, 45]). However, despite these differences, ease of use for all visualizations except *stripeplot* was ~60 (*stripeplot* mean = 35, 95% CI: [33, 36]), suggesting only *stripeplot* was found consistently difficult to use. This may reflect *stripeplot's* much higher estimation variance than the other visualizations (higher standard deviation by about 4-5 percentage points when *probability* = 0.5—Figure 6B—or about 1.44 times the SD of *dotplot-20*—Figure 7).

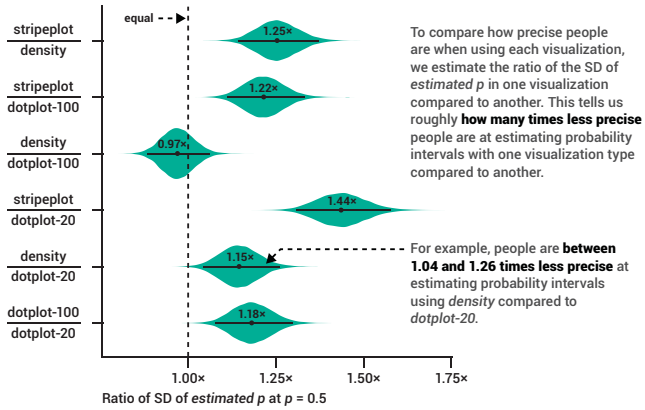

Figure 7. Differences in variance for each visualization type.

DISCUSSION

Discrete outcomes work best in small numbers

Our results suggest that discrete-outcome visualizations of uncertainty can improve probability estimation in space-constrained visualizations of continuous outcomes if care is taken in their instantiation. While *dotplot-20* improved estimation variance over *density*, *dotplot-100* performed very similarly to *density*. In addition, *Stripeplot* performed very poorly. We believe this may reflect the principle that *discrete plots with too many outcomes converge to continuous encodings*: since counting dots is arduous in *dotplot-100* and *stripeplot-50*, people are more likely to read them like density plots and gradient plots (respectively), nullifying

the value of the discrete outcomes. In *dotplot-20*, people can count quickly by using *subitizing*, the ability to quickly recognize rather than count groups of items in number < ~5 [4,13]. Since the vertical groups of dots in dotplot-20 are rarely are over 5 dots high, interval judgements (particularly close to the tails) are often reduced to quick, accurate judgments through subitizing. Thus, **we recommend discrete outcome plots with few enough outcomes to take advantage of subitizing**.

Implications for design and future work

The value of communicating uncertainty

In the first survey, users described goals and unfortunate experiences in OneBusAway that information about uncertainty could help mitigate. In the second survey, most respondents said they appreciated the idea of representing uncertainty. They said this information could help them make better decisions, alleviate their anxiety when the app's information does not match their knowledge, or help them with a problem they commonly experience with OneBusAway.

A minority of respondents said they did not care about the uncertainty information: that point estimates are sufficient. In contrast to respondents who said the information could help, these respondents tended to say the point prediction presented in OneBusAway was consistently accurate. An additional minority actively did not want to receive any information about uncertainty. Several of these people compared evaluating probability information in visualizations to statistics courses. Five feared that, if given uncertainty information, they would become responsible for making decisions, and would have to take responsibility for the wrong decision: "*you're more likely to be unhappy than if you missed the bus and can just blame the app.*" While this represents a small number of participants, we believe that future work is necessary to see how widespread such reactions may be in real-world deployments.

Navigating the precision versus glanceability tradeoff

Designers should attend to the balance of precision and glanceability in representing uncertainty. Participants were divided over whether our visualizations were appropriately glanceable for a transit mobile app. While some said the new designs were easy and clear, more described feeling overwhelmed at least in the context of our experiment and the majority of commenters expressed doubts about being able to use these visualizations while walking to a bus stop. Despite respondent concerns about whether they or others could understand the visualizations, our survey results overwhelmingly show that people understood them.

The designs presented here should be evaluated in longitudinal field studies to assess actual acceptability and use. For example, survey respondents were concerned that the dot plots would compel them to count, but in practice they may find that they count when they want precise estimates but are able to get a good overview from a quick glance. As transit prediction is an everyday practice for many, more sophisticated use of the visualizations may develop over time, making learning an important component to understand in this space.

The designs we evaluated also did not fully exploit interactivity, which might enhance the glanceability of the current static visualizations while preserving uncertainty information. For example, we prototyped a "risk slider" that lets people move the point estimate to match a specific risk threshold; this would allow them to fit the point estimate to their overall preferences or change it to match a specific situation. This feature can be incorporated into any of our proposed designs and should be evaluated as a technique to help resolve the glanceability/false precision tradeoff.

Our research demonstrates that our visualizations can help people accurately, precisely, and confidently evaluate uncertainty, laying the groundwork for future studies evaluating the effects of differences in precision on behavioral measures.

Visual appeal vs. estimation tradeoff

Related to the precision/glanceability tradeoff, people were also divided about preferring the dot plots or the density plots. The dotplots, while ~1.15 times more precise than the density plots and yielding higher confidence, were also rated less visually appealing. We do not know if this is a consequence of unfamiliarity, or if it is because the dotplots are visually busier. It is worth investigating whether the improvement from dotplots is worth decreased visual appeal, or if participants might get used to the dotplots over time.

CONCLUSION

In this paper, we identify general design requirements for visualizing uncertainty on mobile applications as well as domain-specific design requirements for visualizing uncertainty in transit arrival times. From these, we propose a mobile interface for communicating uncertainty in realtime transit predictions in a way that supports users' goals. We developed and evaluated candidate visualizations, including a novel discrete representation of continuous outcomes designed for small screens, quantile dotplots. In a controlled experiment, quantile dotplots improved probabilistic estimates over traditional density plots and facilitated more confident estimation by end-users. Researchers and designers can apply and evaluate these interfaces in the field, with particular attention to opportunities to employ interactivity and other techniques to balance precision and glanceability.

ACKNOWLEDGMENTS

This work was supported by a Google Faculty Award, the Intel Science and Technology Center for Pervasive Computing, the Computing Research Association's Distributed Research Experiences for Undergraduates program, and the Robert Wood Johnson Foundation Health Data Exploration program. We especially thank Martin Wattenberg for valuable feedback.

REFERENCES

1. Nicholas J Barrowman and Ransom A Myers. 2003. Raindrop Plots: A New Way to Display Collections of Likelihoods and Distributions. *The American Statistician* 57, 4: 268–274. http://doi.org/10.1198/0003130032369
2. Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. 2005. Researchers misunderstand confidence intervals and standard error bars. *Psychological methods* 10, 4: 389–96. http://doi.org/10.1037/1082-989X.10.4.389
3. Jacques Bertin. 1983. *Semiology of Graphics: Diagrams, Networks, Maps*.
4. H Choo and S L Franconeri. 2014. Enumeration of small collections violates Weber's law. *Psychonomic bulletin & review* 21, 1: 93–9. http://doi.org/10.3758/s13423-013-0474-4
5. Kristin A Cook and James J Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics*. Pacific Northwest National Laboratory (PNNL), Richland, WA. Retrieved from http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf
6. Michael Correll and Michael Gleicher. 2014. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics* 20, 12: 2142–2151. http://doi.org/10.1109/TVCG.2014.2346298
7. Geoff Cumming. 2009. Inference by eye: reading the overlap of independent confidence intervals. *Statistics in medicine* 28, 2: 205–220.
8. Eric D. Feigelson and G. Jogesh Babu (eds.). 1992. *Statistical Challenges in Modern Astronomy*. Springer New York, New York, NY. Retrieved September 25, 2015 from http://www.springerlink.com/index/10.1007/978-1-4613-9290-3
9. Brian Ferris, Kari Watkins, and Alan Borning. 2010. OneBusAway: results from providing real-time arrival information for public transit. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*, ACM Press, 1807. Retrieved July 13, 2015 from http://dl.acm.org/citation.cfm?id=1753326.1753597
10. Richard Finger and Ann M. Bisantz. 2002. Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science* 3, 1: 1–25. http://doi.org/10.1080/14639220110110324
11. R. Garcia-Retamero and E. T. Cokely. 2013. Communicating Health Risks With Visual Aids. *Current Directions in Psychological Science* 22, 5: 392–399. http://doi.org/10.1177/0963721413491570
12. Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review* 102, 4: 684–704. http://doi.org/10.1037//0033-295X.102.4.684
13. Ernst von Glasersfeld. 1982. Subitizing: The role of figural patterns in the development of numerical concepts. *Archives de Psychologie* 50, 194: 191–218.
14. Theresia Gschwandtnei, Markus Bogl, Paolo Federico, and Silvia Miksch. 2016. Visual Encodings of Temporal Uncertainty: A Comparative User Study. *IEEE Transactions on Visualization and Computer Graphics* 22, 1: 539–548. Retrieved January 8, 2016 from http://www.ncbi.nlm.nih.gov/pubmed/26529717
15. Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. 2009. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, ACM Press, 1303. http://doi.org/10.1145/1518701.1518897
16. U Hoffrage and G Gigerenzer. 1998. Using natural frequencies to improve diagnostic inferences. *Academic medicine : journal of the Association of American Medical Colleges* 73, 5: 538–540. http://doi.org/10.1097/00001888-199805000-00024
17. Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PloS one* 10, 11. http://doi.org/10.1371/journal.pone.0142444
18. Harald Ibrekk and M.Granger Morgan. 1987. Graphical Communication of Uncertain Quantities to Nontechnical People. *Risk Analysis* 7, 4: 519–529. http://doi.org/10.1111/j.1539-6924.1987.tb00488.x
19. S. Joslyn and J. LeClerc. 2013. Decisions With Uncertainty: The Glass Half Full. *Current Directions in Psychological Science* 22, 4: 308–315. http://doi.org/10.1177/0963721413481473
20. Harshath JR. 2015. Redesigning the OneBusAway Mobile Experience.
21. Malte F Jung, David Sirkin, and Martin Steinert. 2015. Displayed Uncertainty Improves Driving Experience and Behavior : The Case of Range Anxiety in an Electric Car. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*: 2201–2210. http://doi.org/10.1145/2702123.2702479
22. Peter Kampstra. 2008. Beanplot: A Boxplot Alternative for Visual\nComparison of Distributions. *Journal of Statistical Software* 28, code snippet 1: 1–9. Retrieved from http://www.jstatsoft.org/v28/c01/paper
23. Matthew Kay, Dan Morris, Mc Schraefel, and Julie A Kientz. 2013. There's No Such Thing as Gaining a

Pound: Reconsidering the Bathroom Scale User Interface. *Ubicomp '13*: 401–410.

24. John K. Kruschke. 2010. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science* 1, 5: 658–676. http://doi.org/10.1002/wcs.72
25. John K. Kruschke. 2011. *Doing Bayesian Data Analysis*. Elsevier Inc.
26. Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5, 2: 110–141. Retrieved May 26, 2015 from http://dl.acm.org/citation.cfm?id=22949.22950
27. S. S. Pak, J. B. Hutchinson, and N. B. Turk-Browne. 2014. Intuitive statistics from graphical representations of data. *Journal of Vision* 14, 10: 1361–1361. Retrieved January 8, 2016 from http://jov.arvojournals.org/article.aspx?articleid=2145239
28. K. Potter, J. Kniss, R. Riesenfeld, and C. R. Johnson. 2010. Visualizing summary statistics and uncertainty. *Computer Graphics Forum* 29, 3: 823–832. http://doi.org/10.1111/j.1467-8659.2009.01677.x
29. Robert A Rigby and D Mikis Stasinopoulos. 2006. Using the Box–Cox <I>t</I> distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling* 6, 3: 209–229. Retrieved September 25, 2015 from http://smj.sagepub.com/content/6/3/209.abstract
30. Michael Smithson and Jay Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods* 11, 1: 54–71. Retrieved August 27, 2015 from http://www.ncbi.nlm.nih.gov/pubmed/16594767
31. David J Spiegelhalter. 1999. Surgical Audit: Statistical Lessons from Nightingale and Codman. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 162: 45–58. http://doi.org/10.1111/1467-985X.00120
32. Stan Development Team. 2015. *Stan Modeling Language: User's Guide and Reference Manual*.
33. Barry N. Taylor and Chris E. Kuyatt. 1994. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*.
34. Edward R Tufte. 2006. *Beautiful Evidence*.
35. Amos Tversky and Daniel Kahneman. 1975. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157: 1124–1131. Retrieved June 20, 2013 from http://link.springer.com/chapter/10.1007/978-94-010-1834-0_8
36. Dennis Vrecko, Alexander Klos, and Thomas Langer. 2009. Impact of Presentation Format and Self-Reported Risk Aversion on Revealed Skewness Preferences. *Decision Analysis* 6, 2: 57–74. Retrieved September 25, 2015 from http://dl.acm.org/citation.cfm?id=1555872.1555874
37. Leland Wilkinson. 2012. Dot Plots. *The American Statistician*. Retrieved September 25, 2015 from http://amstat.tandfonline.com/doi/abs/10.1080/00031305.1999.10474474