Research statement

1 ## APPROACH

My research in human-computer interaction spans many application domains, from personal health informatics, to everyday sensing and prediction, to advancing the state of statistical methods in human–computer interaction. My particular interest is in **communicating complex data and concepts to end-users**. Much of my thesis work focuses specifically on the problem of **communicating uncertainty to users in applications built on sensing and prediction**. I am driven by a thirst for problem areas where existing solutions baffle users: every new confusion surrounding how a system *is* carries fresh insight into how it *should be*.

My approach to research is multidisciplinary and informed by my background not only in computer science, but also visual arts and information visualization. I combine a strong aesthetic design sense with an engineering and computer science background, building systems that are well-engineered, useful, and informed by research, but also clear and aesthetically pleasing. I draw upon methods in **human–computer interaction, visual design, information visualization, data science, and statistics**. I build and deploy novel user-centered computing systems to answer my specific research questions. When working across domains, I often collaborate with experts, as in my work on understanding sleep and weight change.

I am also keenly interested in impact beyond publications—for example, in open science. I have released research software such as PVT-Touch [5,12] and the ARTool R package [13] that have been downloaded by thousands of other researchers [1]. Where permitted by Human Subjects oversight, I have also released the data and analyses from studies as GitHub repositories with R code, aiding reproducibility and meta-analysis. I believe that this approach broadens the impact of research.

2 ## COMMUNICATING UNCERTAINTY IN SENSING AND PREDICTION

People are increasingly exposed to sensing and prediction in their daily lives ("how many steps did I take today?", "how long until my bus shows up?", "how much do I weigh?"). Uncertainty is inherent to these systems and usually poorly communicated. To build understandable data presentations, I study how people interpret their data and what goals they have for it. This determines how to communicate results from the models underlying these systems, which in turn determines what models to use in the first place. **I work across this stack, from understanding people, to designing visualizations and interfaces, to modeling**: human–computer interaction, information visualization, and computer science.

2.1 ### User perceptions of uncertainty in personal health informatics (sleep, weight) and everyday prediction (transit arrival times)

My interest in how users deal with uncertainty in everyday sensing began with the sleep environment. Sleep and sleep quality affects all aspects of our lives, from daily cognitive functioning, to cardiovascular health, to weight gain. However, these effects are not always clear to people, and sleep labs cannot uncover environmental factors (light, sound, air quality, etc.) that impact sleep quality. I built a capture-and-access system called *Lullaby* that records sleep quality alongside environmental factors known to disturb sleep, and assists people in identifying factors that might be disturbing their sleep (Fig. 1); it received a

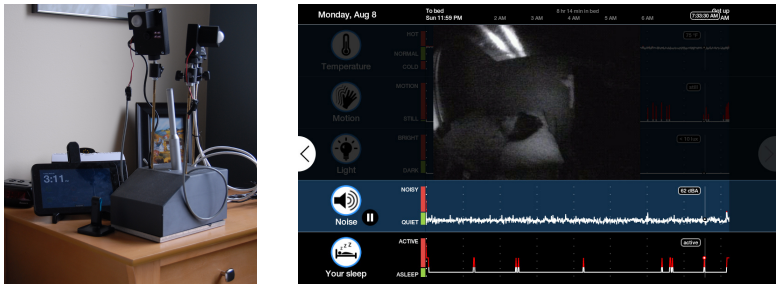

Fig. 1. The Lullaby bedside sensor suite and playback of a previous night's sleep [4].

Best Paper Award at UbiComp 2012 [4]. In that work, I found that textual descriptions of correlations (e.g., an association between light and sleep quality) were preferable to some people over graphical or statistical explanations. This prompted me to delve more into the problem of communicating personal informatics and sensing data to users.

As a starting point for understanding how people perceive uncertainty in sensing, I chose to examine perhaps the most ubiquitous health sensor: the bathroom weight scale. While frequent weigh-ins improve weight loss outcomes, many people have an aversion to stepping on this scale. **My work on understanding weight exemplifies my multi-methods approach**: I employed qualitative analysis of online product reviews to identify common themes around perceptions of data accuracy (accuracy was discussed in > 25% of negative reviews), expert interviews to unpack how experts and patients view scales (identifying an unhealthy fixation on specific numbers and a overreactions to small fluctuations amongst many users), a quantitative study of actual within-day weight fluctuation (to determine what magnitude of fluctuations are normal), and a survey of ~900 scale users (finding that users with greater understanding of expected fluctuations had greater trust in their scales). This spawned design recommendations for improving the 100-year-old bathroom scale interface, which to this day continues simply to spit out numbers, often to a meaningless level of precision (e.g. tenths of a pound), and received a Best Paper Award at UbiComp 2013 [9].

I conducted a similar multi-methods series of studies on user needs for uncertainty in realtime public transit arrival prediction [under review]. This work included an evaluation of the perceptual properties of discrete visualizations of uncertainty in place of abstract representations of predictive distributions; the *quantile dotplots* (Fig. 2) I developed made user estimates ~15% more precise than abstract depictions of density [7].

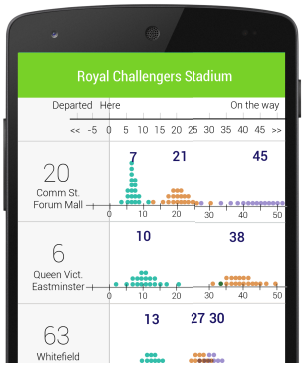

Fig. 2. Redesigned application for realtime bus arrival predictions that uses a discrete encoding I call *quantile dotplots*, which improves the precision of users' interval estimates.

2.2 Building models to support user preferences in uncertainty: acceptability of accuracy in user-facing classification

I am investigating problems of communicating uncertainty in several other domains apart from weight tracking. A common research thread in ubiquitous computing involves the construction novel sensing applications based on machine learning. For example, applications have been proposed for domains like smart alarms, disaggregated energy sensing, and location tracking. Classifiers for such applications are often evaluated and compared using the F1 score—the harmonic mean of precision and recall. However, we might ask several questions of such evaluations: how do we know what level of performance is acceptable to users? And how do we know what types of errors users care about? Users might care more about precision or recall in some applications. For example, in weather forecasting people often care more if the forecast calls for no rain and then it does rain (catching them unprepared) than the opposite (a pleasant surprise)—a phenomenon called *wet bias* [14].

I tackled these problems by developing a survey instrument to elicit a user's *acceptability of accuracy* ("accuracy" used here in a colloquial sense) for various scenarios of use [11]. Using this instrument, system builders can systematically derive a weighting between precision and recall to use to evaluate their

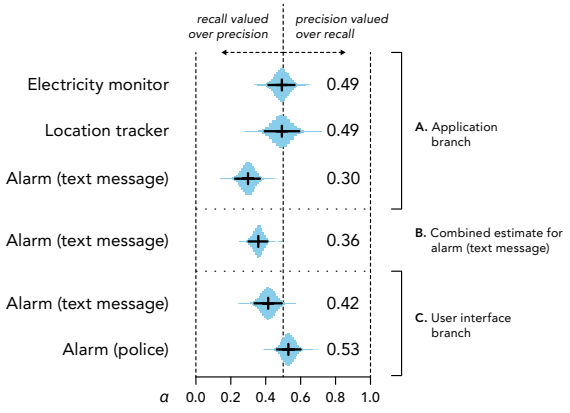

Fig. 3. My *acceptability of accuracy* survey tool translates user preferences in errors into weights of precision versus recall that can be used to train/evaluate classifiers for different applications and user interfaces [11].

classifiers (Fig. 3). The key insight here is not simply that we can and should elicit user's preferences for types of error in these systems, but that we should do it in a way that fits into existing practice—since researchers already use F-scores in evaluating systems, I developed a method that translates users' preferences into a weighted F-score. In the precipitation forecasting example, our instrument suggested recall be weighted higher than precision (over-predicting the probability of rain), which is in accordance with industry practice [14]. More generally, this work enables us to capitalize on user expectations to improve the perceived performance of models.

3 ADVANCING STATISTICAL METHODS IN HUMAN–COMPUTER INTERACTION

3.1 Advocating for Bayesian estimation over frequentist null hypothesis testing

My ongoing work in communicating uncertainty has strongly affected how I have come to communicate results of my own work—using Bayesian estimation with a focus on understanding effect sizes, trying more faithfully to reflect the uncertainty of the scientific process. In HCI, the dominant statistical methods (null-hypothesis significance testing) and the way we communicate results (often as tables of *p*-values) are driven by a binary world view: *is Technique A better than Technique B on Outcome Y?* Especially with the limited resources available in the field for large user studies, this binary form of inference is subject to a high level of noise. Effect sizes facilitate an assessment of the practical importance of research: can users notice an improvement? Are those improvements worth the cost?

In several of my most recent publications I have adopted Bayesian methods of statistical analysis, with a focus on *estimation* rather than *significance testing*. In my InfoVis 2015 paper, *Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation* [6], I re-analyzed a dataset published by Harrison *et al.* [2], demonstrating how to apply a Bayesian approach to statistical analysis in a perceptual study. That paper also acts as a visual tutorial in myriad modeling assumptions and choices (e.g. Fig. 4), and is part of my broader goal to raise the standard of statistical practice and communication in the field. It received a Best Paper Honorable Mention award.

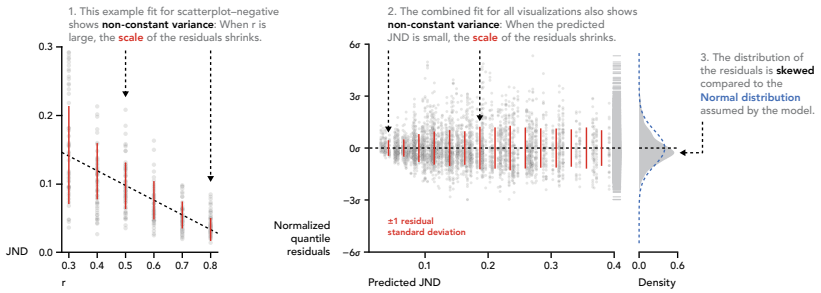

Fig. 4. A visual tutorial on linear model assumptions. This figure is part of a set of figures I created in [6] that walk through linear and log-linear models, censoring, and the results of a Bayesian analysis.

Where *Beyond Weber's Law* acts as an example of how such approaches might be applied in practice (the *how*), in a paper to appear at CHI 2016 I use a simulation approach to demonstrate the benefits to the field of a Bayesian approach (the *why*) [10]. This paper includes a meta-meta-review demonstrating that the field of HCI does not typically conduct meta-analyses, the traditional method within a frequentist statistical paradigm for accumulating knowledge and reducing the error in effect estimates. I advocate for an alternative approach based on Bayesian methods for accumulating knowledge. I show how this reduces experimental error in a way that fits into the existing publication incentives of the HCI field, where research is often driven by rapid publication of novel interactive systems. While others have proposed improving estimation error in the field by incorporating more replication and meta-analysis, these approaches require new incentives for people to conduct and publish such work; the framework I propose instead fits into how people already conduct research in novel HCI systems. It thus stems not only from a desire to improve statistical methods, but to do so in a way that fits with users' (here, researchers') needs, what I call user-centered statistics.

4 FUTURE WORK

The core of my work lies in **building everyday sensing and prediction that works for people.** Sensing and prediction have become integral to our lives. I want to demystify these systems and help people make effective decisions

with them by 1) identifying how to communicate predictive results people can understand and 2) building models that reflect users' desired trade-offs in error.

My second research thrust lies in **developing tools for Bayesian statistical analysis of human subjects experiments that researchers can actually understand.** I believe Bayesian analysis will allow human–computer interaction to shed the baggage of noisy statistical tests applied to small-n studies, and shift to a more nuanced view of practical effect sizes and an accumulation of knowledge. However, it remains locked in something of a specialist mode, as Bayesian analyses are typically done using statistical modelling languages. It requires the understanding of researchers' needs *and* the underlying analyses for more user-friendly tools to emerge.

My work has previously been funded by an NSERC Canada post-graduate fellowship, the NSF, industry grants through the Intel Science and Technology Center for Pervasive Computing, a Google faculty award, and a Health Data Exploration grant from the Robert Wood Johnson Foundation. Going forward I will also submit grants to the NSF Cyber-Human Systems and Smart and Connected Health programs.

4.1 Acceptability of accuracy beyond binary classification

While my existing work has looked at acceptability of accuracy of binary classifiers, particularly with respect to user preferences in precision and recall, I will expand this work to other types of predictive systems. For example, in predictions of a continuous variable I will look at preferences for bias and variance in applications: e.g., for a system measuring how many hours of TV people watch, would they prefer a prediction that consistently overestimates by an hour (low variance/high bias) or one that is correct on average but jumps around more (high variance/low bias)? I will develop an instrument to measure these preferences in advance of deployment and validate them against experimentally manipulated error in a deployed system. Initially I will focus on surveys, then compare this approach to more heavyweight (but likely more accurate) ones, such as interactive simulations. Taking a step up, I am interested in whether there are particular properties of predictions for which user preferences are consistent across many domains. For example, given a prediction as a probability distribution, we should expect people to be differently sensitive to bias and variance depending on application domain. However, properties like the skewness or kurtosis of a predictive distribution may be consistently less salient for people. Knowing that, we might trade off those properties in modeling against simplicity, speed, or overall reduction in error.

4.2 Hierarchical, predictive, and explanatory self-experimentation

I believe that the future of personal informatics is explanatory models that incorporate population information to improve individual predictions. I am expanding on my work in understanding perceptions of uncertainty in weight data through the development of a bathroom scale that tracks weight over time with a Bayesian autoregressive model. This model incorporates population-level priors, can give people an estimate of their weight with associated uncertainty, and accounts for several systematic biases in weight measurement (e.g. clothes or time of day). This allows for explanations like, "You weighed at noon while heavily clothed; your weight now is ~150lbs. At 7am, when you typically weigh, your weight would have been ~147lbs". These explanations, combined with short-term predictions of future weight, help people better put their weight in context at the moment of measurement, which I believe will reduce the pervasive anxiety many people feel when stepping on the scale.

I plan to expand this approach to *self-experimentation*, a recently-emerging field that aims to give users the tools to run rigorous, controlled self experiments, for example, to identify food triggers for irritable bowel syndrome [3]. Part of this approach inherently involves educating users in some causal statistical framework; existing proposals focus on traditional frequentist approaches to statistics. However, I do not believe that it will be fruitful to educate users in the interpretation of p-values or null hypothesis significance testing, or that binary inferences from short self experiments will be reliable or even desired by users. I plan to investigate user needs for self-experimentation models and particularly compare frequentist versus predictive Bayesian models. The latter can make straightforward

predictions of effects ("eating cheese now will cause a 1-2 point increase in your self-rated pain scale") that will be easier for users to act on (e.g. by facilitating cost/benefit analysis—weighing the predicted symptoms against my love of cheese, should I eat this?). In the long term, such an approach to self-experimentation could unify the analysis of population-level trials with small-*n* self experiments through hierarchical modeling, while simultaneously offering a better interface to users.

4.3 Tools to support researcher-centered Bayesian statistics

A shift to Bayesian analysis is being advanced across social sciences such as political science and psychology. However, the tools behind this shift are typically specialized modeling languages, like JAGS (http://mcmc-jags.sourceforge.net/) and Stan (http://mc-stan.org/). I plan to build user-centered statistical tools for practitioners to accelerate this shift. This will build upon my own expertise in information visualization and the R statistical programming language (e.g., authoring the ARTool R package [13], a version of the aligned rank transform proposed by Wobbrock *et al.* [15]). I will build an accessible, interactive R-based set of tools for Bayesian modeling, aimed first at HCI researchers. I will study how researchers in HCI approach statistical analyses. I will include a strong visualization and educational component to help researchers understand the modeling choices they make, set reasonable priors for models, and interpret the results of their analyses. While building upon my work in *communicating* uncertainty, this work will also advance the state of the art in interfaces for *inputting* uncertainty, for example, through developing methods for *interactive prior specification*. This research agenda combines threads from my existing work in communicating uncertainty with my interests in advancing the state of statistical methods in the field.

5 REFERENCES

1. ARTool daily downloads from RStudio mirror. http://www.rdocumentation.org/packages/ARTool.
2. Harrison, L., Yang, F., Franconeri, S., & Chang, R. (2014). Ranking Visualizations of Correlation Using Weber's Law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 1943–1952.
3. Karkar, R., Zia, J., Vilardaga, R., Mishra, S.R., Fogarty, J., Munson, S.A., Kientz, J.A. (2015). A Framework for Self-Experimentation in Personalized Health. *Journal of the American Medical Informatics Association* (JAMIA).
4. **Kay, M.**, Choe, E. K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., & Kientz, J. A. (2012). Lullaby: a capture & access system for understanding the sleep environment. *UbiComp '12*.
5. **Kay, M.,** Grandner, M. A., Jared, B., Lang, R. A., F., W. N., & Kientz, J. (2013). Initial Validation of an Android-Based Psychomotor Vigilance Task. *Sleep Abstract Supplement*, 36.
6. **Kay, M.**, & Heer, J. (2016). Beyond Weber's Law: A Second Look at Ranking Visualizations of Correlation. *IEEE Transactions on Visualization and Computer Graphics,* 22(1), 469–48.
7. **Kay, M.**, Kola, T., Hullman, J., & Munson, S. (2016). When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. *CHI '16* (to appear).
8. **Kay, M.**, Matuszek, C., & Munson, S. A. (2015). Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. *CHI '15*, 3819–3828.
9. **Kay, M.**, Morris, D., Schraefel, M., & Kientz, J. A. (2013). There's No Such Thing as Gaining a Pound: Reconsidering the Bathroom Scale User Interface. *UbiComp '13*, 401–410.
10. **Kay, M.**, Nelson, G., & Hekler, E. (2016). Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. *CHI '16* (to appear).
11. **Kay, M.**, Patel, S. N., & Kientz, J. A. (2015). How Good is 85%? A Survey Tool to Connect Classifier Evaluation to Acceptability of Accuracy. *CHI '15*, 347–356.
12. **Kay, M.**, Rector, K., Consolvo, S., Greenstein, B., Wobbrock, J. O., Watson, N. F., & Kientz, J. A. (2013). PVT-Touch : Adapting a Reaction Time Test for Touchscreen Devices. *PervasiveHealth '13*.
13. **Kay, M.** & Wobbrock, J. (2014). ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs. R package version 0.9.5, https://cran.r-project.org/web/packages/ARTool.
14. Silver, N. (2012). The Weatherman Is Not a Moron. *The New York Times*. Retrieved from http://www.nytimes.com/2012/09/09/magazine/the-weatherman-is-not-a-moron.html
15. Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. *CHI '11*, 143–146.