SURVEY INFORMATICS
IDEAS, OPPORTUNITIES, DISCUSSIONS

LEEN-KIAT SOH & ADAM ECK

UNL NCRN CS Group April 2, 2014

ACKNOWLEDGMENTS

- This material is based upon work supported by the National Science Foundation under Grant No. SES - 1132015
- UNL Survey Research and Methodology (SRAM)
- UNL Gallup Research Center
- Collaborators: Kristen Olson, Jolene Smyth, Robert Belli, Allan McCutcheon (PI)
- NCRN CS Team: Leen-Kiat Soh (faculty), Adam Eck, Gregory Atkin, Hariharan Arumuchalan

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
All presented experimental results are preliminary and not suitable for official publication.

2

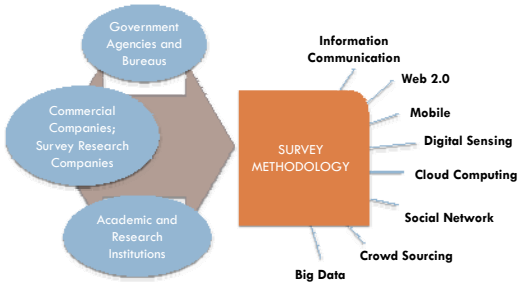
RISE OF TECHNOLOGY IN SURVEY METHODOLOGY

- Survey methodologists are turning to the use of technology to
  - gather higher quality data
  - increase throughput and overall quantity in data collection
  - improve analysis efficiency and effectiveness

IDEAS OPPORTUNITIES DISCUSSIONS

3

RISE OF TECHNOLOGY IN SURVEY METHODOLOGY



IDEAS OPPORTUNITIES DISCUSSIONS

RISE OF TECHNOLOGY IN SURVEY METHODOLOGY

- John H. Thompson outlined several key ways the Census intends to explore the use of technology to improve data collection:
  - offering multiple modes of response for the upcoming 2020 census (including online response options over the web)
  - electronically merging multiple sources of information from existing databases to reduce respondent burden
  - using geographical hardware and software tools to improve address canvassing

Thompson, J.H., 2014: Now is the Time for Change, posted on 2/7/2014

IDEAS OPPORTUNITIES DISCUSSIONS

5

RISE OF TECHNOLOGY IN SURVEY METHODOLOGY

- Survey research companies have increased use of computer-based (including web-based) response options to
  - improve responsiveness and flexibility
    - Administration, configurability, adaptivity
  - better collect, organize, and store respondent data
  - better track sessions to generate paradata
  - reach more respondents (increase response rates)
  - improve design-to-deploy time, collection-to-report time, feedback-to-redesign time, etc.

IDEAS OPPORTUNITIES DISCUSSIONS

6

RISE OF TECHNOLOGY IN SURVEY METHODOLOGY

- Survey methodology researchers are also exploring the use of advanced Web 2.0 and information-communication technologies to
  - better elicit information from respondents
    - smartphones for data collection (in self-administered and interview-based surveys)
    - interactive online surveys
  - more effectively and efficiently visualize respondent data
  - model and directly track user behaviors
    - Such as to survey daily activities by keeping track of steps taken by respondents (e.g., using pedometer)

IDEAS | OPPORTUNITIES | DISCUSSIONS

A CASE FOR SURVEY METHODOLOGY + COMPUTER SCIENCE

- That survey methodologists are embracing technology offers opportunities to
  - Expand the reach of products from computer science (CS) to improve how we perceive and act in the world around us (e.g., ubiquitous computing)
  - Develop a synergistic relationship with CS to jointly improve our understanding of both fields and their practices
- That is, *as survey methodologists employ technological advances in their operations, new problems will arise that will also be of interest to CS theoretically and practically to a wide range of domains*

IDEAS | OPPORTUNITIES | DISCUSSIONS

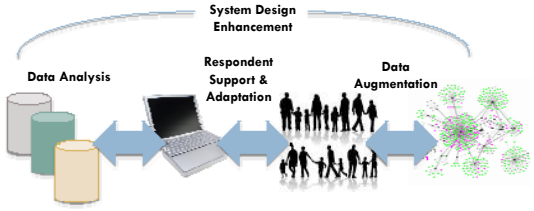
A CASE FOR SURVEY METHODOLOGY + COMPUTER SCIENCE

- We call this emerging research area:

Survey Informatics

IDEAS | OPPORTUNITIES | DISCUSSIONS

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE



http://prblog.typepad.com/strategic_public_relation/2007/06/top-10-reasons-.html
http://www.ypte.org.uk/environmental/meat-free-mondays/209

IDEAS | OPPORTUNITIES | DISCUSSIONS

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Data Analysis: Use CS techniques to handle more effectively and efficiently large scale of data (include automation, meta-tagging, and management)
  - Pattern recognition, machine learning, and data mining

IDEAS | OPPORTUNITIES | DISCUSSIONS

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- **Example**: understanding survey process through paradata
  - Too much data to manually analyze
  - Difficult to build statistical models (feature extraction, big data, etc.)
- Big Picture: use machine learning and data mining to analyze paradata
  - Learn how interactions during survey process affect outcomes and respondent behavior
  - Predict outcomes based on observed behavior
- Goal: understanding breakoff
  - What behaviors or features lead to breakoff?
  - Can we predict breakoff?
    - From sequences of paradata
    - Before it happens?

IDEAS | OPPORTUNITIES | DISCUSSIONS

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Approach: sequence classification with recurrent neural networks

Survey	Accuracy	Precision	Recall
April 2012	0.9699	0.9763	0.7203
May 2012	0.9557	0.9584	0.5825
June 2012	0.9746	0.9810	0.7679
Sept. 2012	0.9742	0.9699	0.7652
Nov. 2012	0.9775	1.0000	0.7751
Dec. 2012	0.9726	0.9914	0.7348
Jan. 2013	0.9869	0.9923	0.8784

Breakoff	Random
0.0156	0.9844
0.0372	0.9628
0.0187	0.9813
0.0228	0.9772
0.0211	0.9789
0.0195	0.9805
0.0360	0.9640

- Challenge: low frequency of outcome we are trying to predict, difficult to achieve high recall (true positive rate)

IDEAS | OPPORTUNITIES | DISCUSSIONS

13

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Respondent Support and Adaptation: Computer-assisted or supported approaches utilizing advances in intelligent agents and natural language processing to better support respondents
  - Embodied software agents to interact with respondents
  - Model respondents and adapt the survey intelligently based on respondent's answers
  - E.g., Intelligent Tutoring Systems

IDEAS | OPPORTUNITIES | DISCUSSIONS

14

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Example: build an intelligent agent to mitigate and avoid breakoff
  - **Monitor** current respondent interactions through paradata
  - Predict likelihood of breakoff **during survey**
  - Take actions to **avoid breakoff** and encourage respondent to finish survey
    - Control **adaptive** survey that changes based on respondent's behavior
    - E.g., switch to questions about respondent's hobbies and interests

IDEAS | OPPORTUNITIES | DISCUSSIONS

15

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Data Augmentation: Social network analysis and gamification to enhance survey data collection
  - Twitter, Facebook
  - Game With a Purpose (GWAP)

IDEAS | OPPORTUNITIES | DISCUSSIONS

16

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Example: real-time monitoring of public opinion during presidential debate
  - Mine social media for peoples' opinions
    - Count positive/negative for each candidate
      - Can track progress during debate (who is gaining momentum)
      - Compare to pre/post-debate polling
  - Extract snippets to understand "why"
  - Add visualization for viewers
    - Augmented reality
    - Encourage participation

IDEAS | OPPORTUNITIES | DISCUSSIONS

17

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Example: assign points/achievements for completing surveys
  - Unlock new surveys, activities, or prizes
  - Leaderboard for all participants
  - Goal: increase motivation and response rates
    - Effective in other disciplines

IDEAS | OPPORTUNITIES | DISCUSSIONS

18

IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- System Design Enhancement: Human-computer interactions (HCI), software engineering, and databases could be used to improve the quality of computational systems used to *collect*, *manage*, and *analyze* such data
  - better interface design → reduce noise (or mode effects) in data collection as respondents respond to surveys
  - better overall software design → streamline the process of implementing and administering surveys for non-technical survey methodologists in charge of overseeing these computer-based surveys (e.g., Google Docs)
  - better data management design (e.g., through relational or big data databases) → more efficient and effective data analysis

IDEAS | OPPORTUNITIES | DISCUSSIONS

19

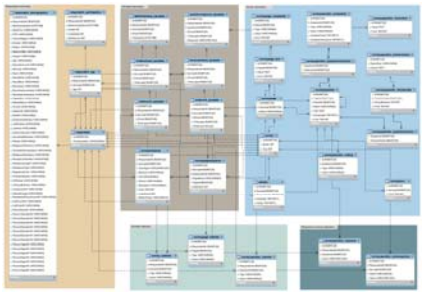
IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE

- Example: how to store multiple levels of survey data?
  - Flat files lack flexibility (large redundancy)
  - Paradata: variable length, sequential data
    - Different per respondent
- Approach: **relational database** (MySQL)
  - Easier to link relationships
  - More compact representation overall
  - Faster querying to retrieve data
  - Easier to maintain (add, remove, modify)
  - Allow for more versatile data analyses

IDEAS | OPPORTUNITIES | DISCUSSIONS

20

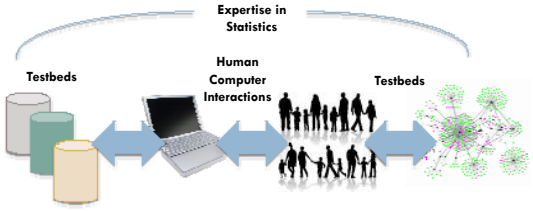
IMPROVING SURVEY METHODOLOGY THROUGH COMPUTER SCIENCE



IDEAS | OPPORTUNITIES | DISCUSSIONS

21

IMPROVING COMPUTER SCIENCE THROUGH SURVEY METHODOLOGY



http://prblog.typepad.com/strategic_public_relation/2007/06/top-10-reasons-.html
http://www.ypte.org.uk/environmental/meat-free-mondays/209

IDEAS | OPPORTUNITIES | DISCUSSIONS

22

IMPROVING COMPUTER SCIENCE THROUGH SURVEY METHODOLOGY

- Testbeds: Very rich, interesting domain within which many computer science hypotheses can be tested and solutions explored
  - Multiple levels: response data, paradata, survey and response metadata, auxiliary data, administrative data
  - Real-time, real-world
  - Noisy, class imbalance
  - Large scale and diverse (respondents)
  - Sometimes spatio-temporal

IDEAS | OPPORTUNITIES | DISCUSSIONS

23

IMPROVING COMPUTER SCIENCE THROUGH SURVEY METHODOLOGY

- Human-Computer Interactions: Lessons learned from survey methodology can also be used to improve the products produced by computer science research
  - Build better recommender systems
  - Improve user preference elicitation techniques
  - Create better interactive help systems
  - Develop better knowledge engineering systems

IDEAS | OPPORTUNITIES | DISCUSSIONS

24

IMPROVING COMPUTER SCIENCE THROUGH SURVEY METHODOLOGY

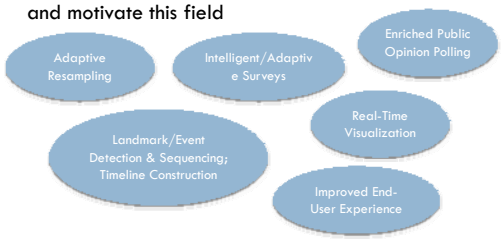
- Expertise in Statistics: Survey methodologists are experts at building statistically sound aggregate models from randomly sampled, low level data and can
  - Improve global data fusion from localized nodes in wireless sensor networks and intelligent agent-based sensing
  - Improve sampling in active learning
  - Improve crowdsourcing
  - Improve agent-based or individual-based modeling

IDEAS | OPPORTUNITIES | DISCUSSIONS

25

EXAMPLE APPLICATIONS WITHIN REACH …

- Some applications that are within reach that could help galvanize efforts past and present to crystallize and motivate this field



IDEAS | OPPORTUNITIES | DISCUSSIONS

26

CONTACT INFO

- lksoh@cse.unl.edu, aeck@cse.unl.edu
- Department of Computer Science and Engineering, University of Nebraska, Avery Hall, Lincoln, NE 68588-0115
- Intelligent Agents and Multiagent Systems (IAMAS) Group

27

MINDSTORM? Conventional Surveys vs. Today's Online Systems

	Conventional Surveys	Today's Online Systems
User-oriented motivation vs. System-oriented motivation	Respondents participate in them more often than not because of direct rewards for completing the survey as promised by the survey administrator; sometimes respondents do have a sense of duty/obligations to participate in terms of what they perceive of the overall goal.	Respondents participate in them more often than not because other more intrinsically motivated reasons other than for the sake of completing the survey and systems reward user participation in the survey with services; usually a win-win situation for both sides; respondents from the same community or virtual organization may share the same goals.
Institutional control vs. crowd-sourced/user-driven	Institutions administering the surveys determine and have full control over the objectives and design of surveys; in a way, these surveys are more "designed", driven by researchers and institutions.	Respondents (or in this case "unwitting data providers") have greater flexibility in determining what they want to share and provide (think Twitter's #hashtags) and *follow*; in a way, this type of "data collection" is more "organic", driven by users or crowds; respondents from the same community or virtual organization may share the same goals.
Static/snapshots vs. fluid/continual (levels of temporal resolutions)	Surveys are usually administered at prescribed times or intervals (such as the US Census every 10 years).	Surveys or data collections are done continually and can be processed on-demand or periodically, and thus usually yield results of higher or more on-demand temporal resolutions.

IDEAS | OPPORTUNITIES | DISCUSSIONS

28

MINDSTORM? Conventional Surveys vs. Today's Online Systems

	Conventional Surveys	Today's Online Systems
Direct (Explicit) Probing vs. Indirect (Implicit) Probing (Intrusive vs. Non-intrusive)	Survey questions are direct and explicit and it is expected that the response to a question directly answers that question; there might be other techniques to help respondents provide the response, but the probing is direct, explicit, and usually intrusive; data entered is more of a result of elicitation.	Respondents sometimes participate in implicit data collection (such as doing a search on Google, or playing a game on GWAP labeling pictures) without realizing that they are providing feedback to the system; some of these systems implicitly "pose" questions (such as providing a "thumbs-up" icon for you to click without asking you whether you like a particular piece of information or item) and thus respondents feel less pressured to respond and could see these systems as less intrusive; data entered is more of a result of user voluntary decision.
Measured feedback vs. instant-gratification feedback	Respondents usually do not receive the survey results until much later if any at all, in terms of statistics and published reports.	Respondents usually see the overall feedback immediately, such as a map on ESPN's SportsNation, the number of star ratings of a product or a review on Amazon.com, and so forth; also more interactions
Non-routine vs. routine	Respondents are usually required to set aside time and effort to respond to survey questions, breaking their daily routine in some sense.	With implicit monitoring or voluntary information disclosure, data is usually collected when users or respondents perform routine tasks (e.g., indicating a "like" on a particular product) as part of their usual interactions with the computer system.

IDEAS | OPPORTUNITIES | DISCUSSIONS

29

MINDSTORM? Conventional Surveys vs. Today's Online Systems

	Conventional Surveys	Today's Online Systems
Structured vs. free-formed response (plain vs. enriched contexts)	Usually, survey questions are structured, coordinated, with expected types of response (e.g., Likert, multiple choice, fill-in-the-blank, etc.); the responses are recorded without additional contexts other than perhaps respondent demographics; usually there is a linear path leading from the first question to the last question; session data—paradata about how respondents answered questions—is often not kept or available. Note: this is, however, trending to more context as more data is digitized and available. For example, government surveys like SIPP are linked against IRS, Medicare/Medicaid, etc. data to provide greater context and a ground truth to evaluate response accuracy.	Some questions regarding user profile are structured (e.g., gender, age, etc.) but some are generally free-formed (such as reviews of products, expression of opinions, etc.); responses can be associated with richer contexts to qualify or model a respondent or user (e.g., a customer with his/her purchase profile writing a review of a product, or rating another review while making a purchase of a specific product); usually, there are multiple paths for a respondent/online user to arrive at a rating page, providing richer and more diverse "digital trails"; user behaviors before and after response are often captured as part of session data as well; survey data usually is accompanied by data from the ecosystem of the survey.
Precision vs. Recall (Noise vs. coverage)	The list of respondents are sampled from a population for costs and statistical significance factors, to ensure precision, and measures are usually taken to negate noise to ensure statistical integrity.	In general, crowd-sourced or online systems tend to obtain more organic participation (spatially, temporally) (e.g., ESPN's SportsNation regularly obtain more than 10,000 votes for each of its survey questions across the entire U.S.); but with better coverage comes with noise mostly due to biases and self-selection; further, commercial efforts promoting their products may corrupt the survey results by entering favorable reviews and positive ratings; there might also be software bots corrupting surveys.

IDEAS | OPPORTUNITIES | DISCUSSIONS

30