Using Twitter to Predict Survey Responses

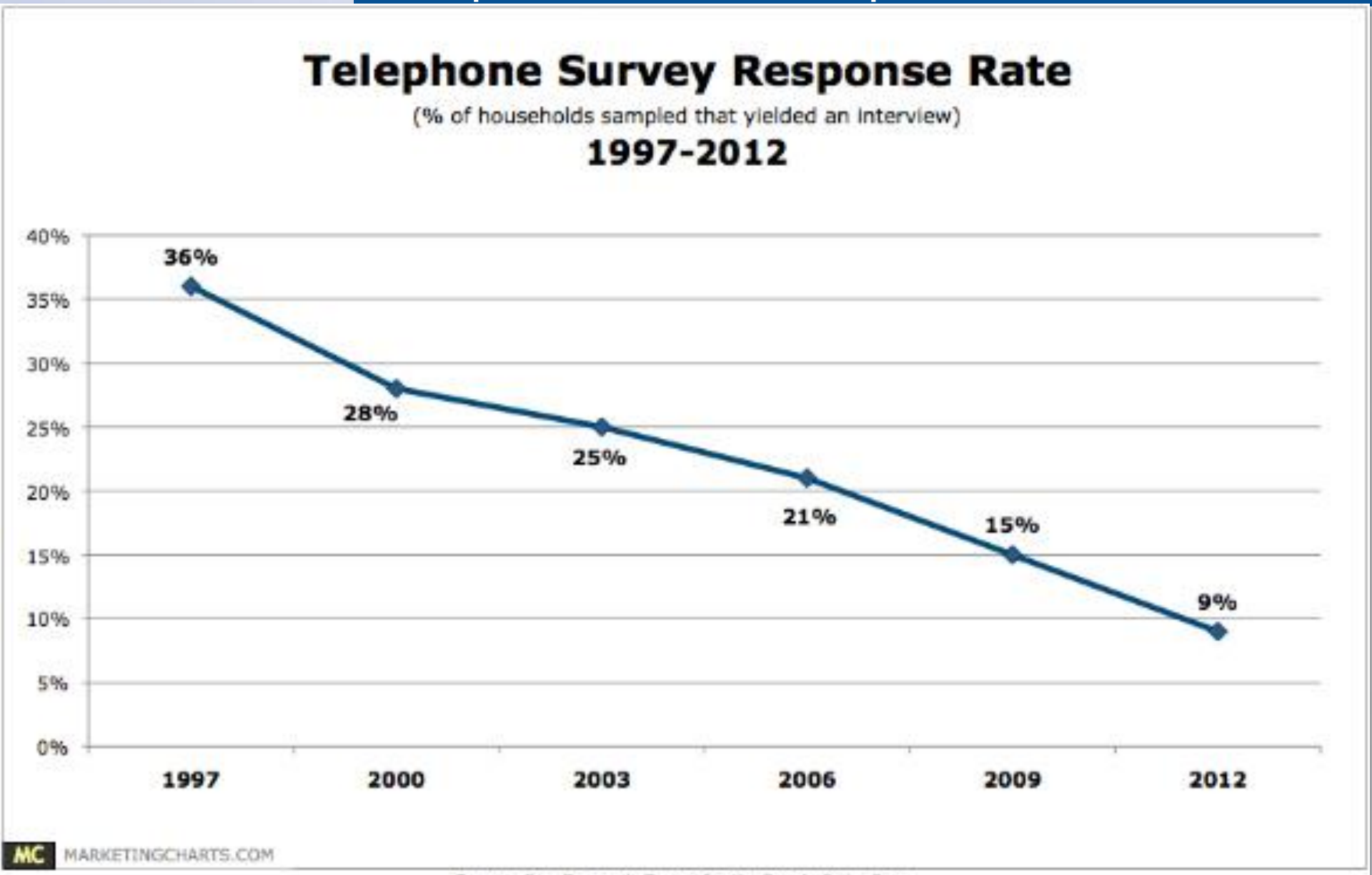
Joe Murphy, Justin Landwehr, and Ashley Richards
RTI International

Annual Conference of
the Midwest Association for
Public Opinion Research

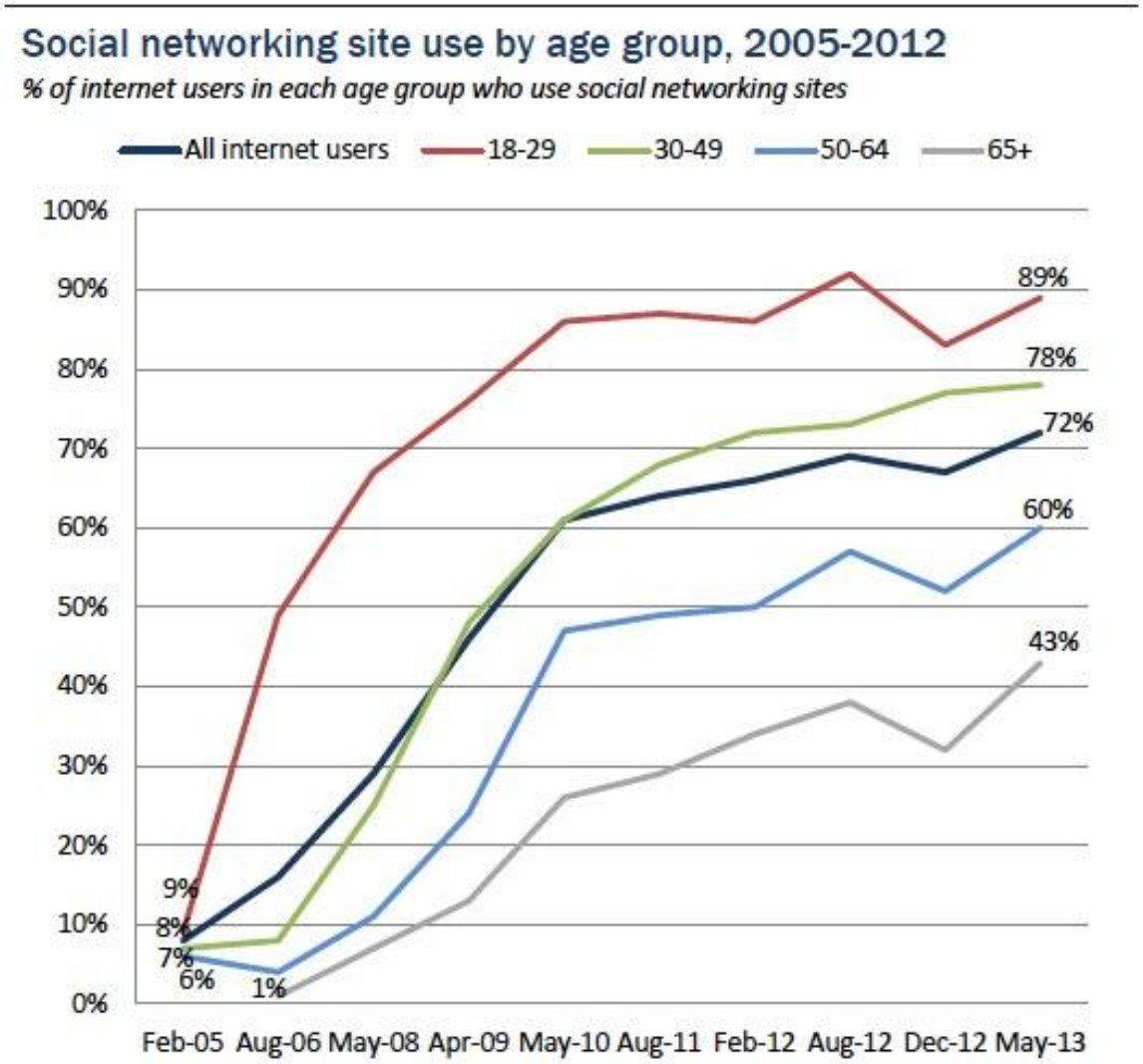
November 22-23, 2013
Chicago, IL

www.rti.org

Response rates have plummeted

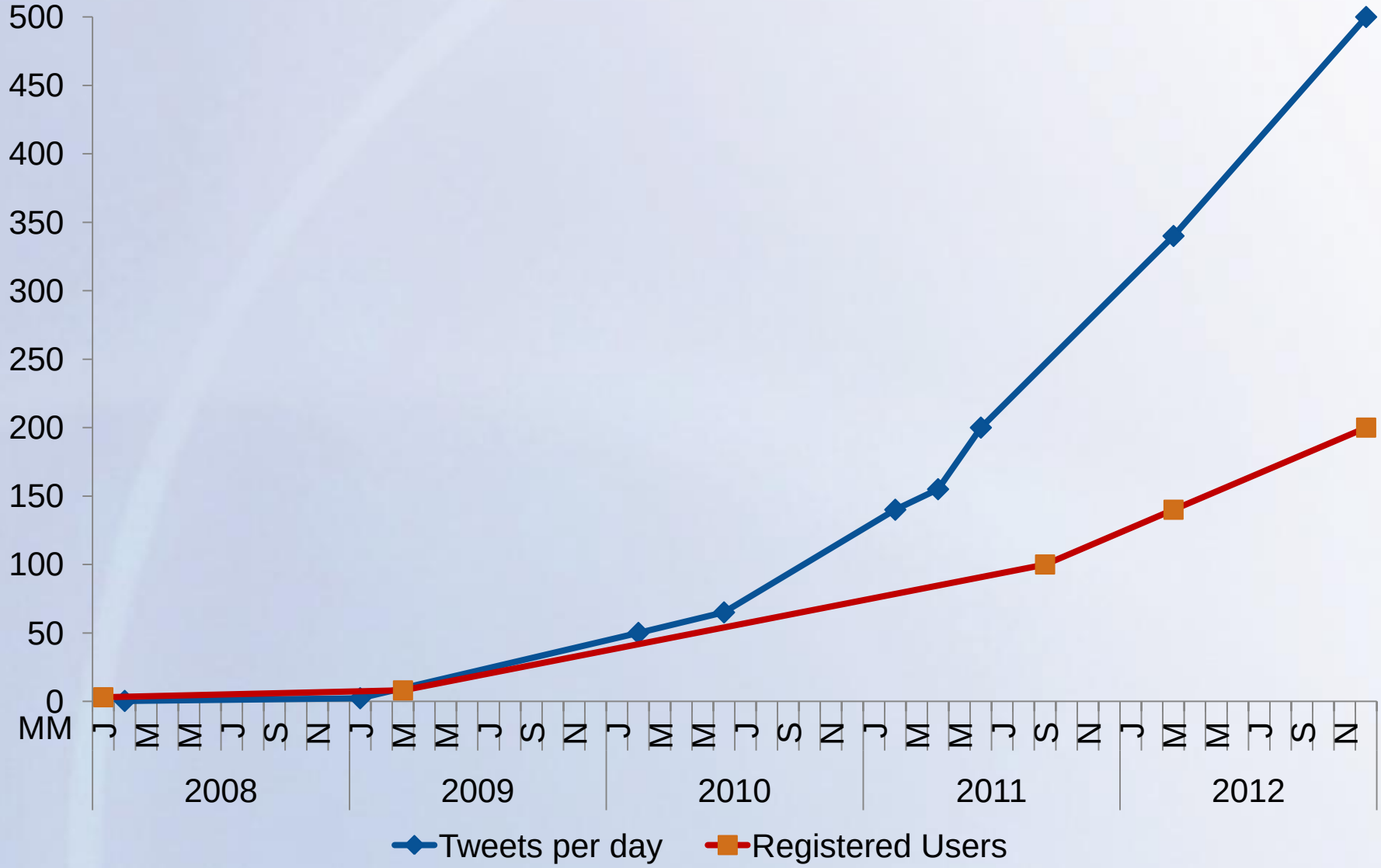


Social media use has skyrocketed

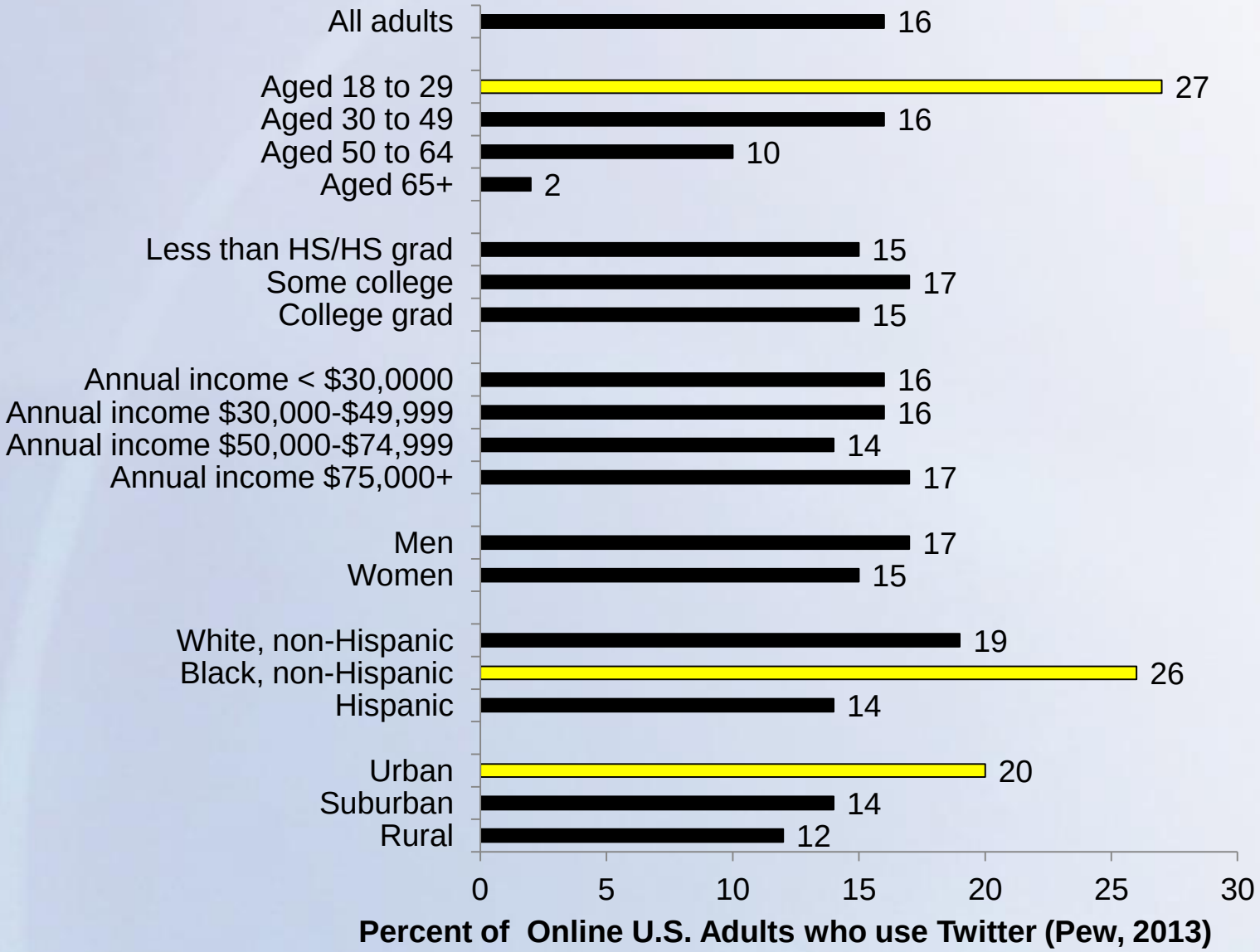


Source: Pew Research Center's Internet & American Life Project tracking surveys 2005-2013. Spring Tracking Survey, April 17 – May 19, 2013. N=1,895 adult internet users ages 18+. Interviews were conducted in English and Spanish and on landline and cell phones. The margin of error for results based on all internet users is +/- 2.5 percentage points.

Twitter has grown exponentially

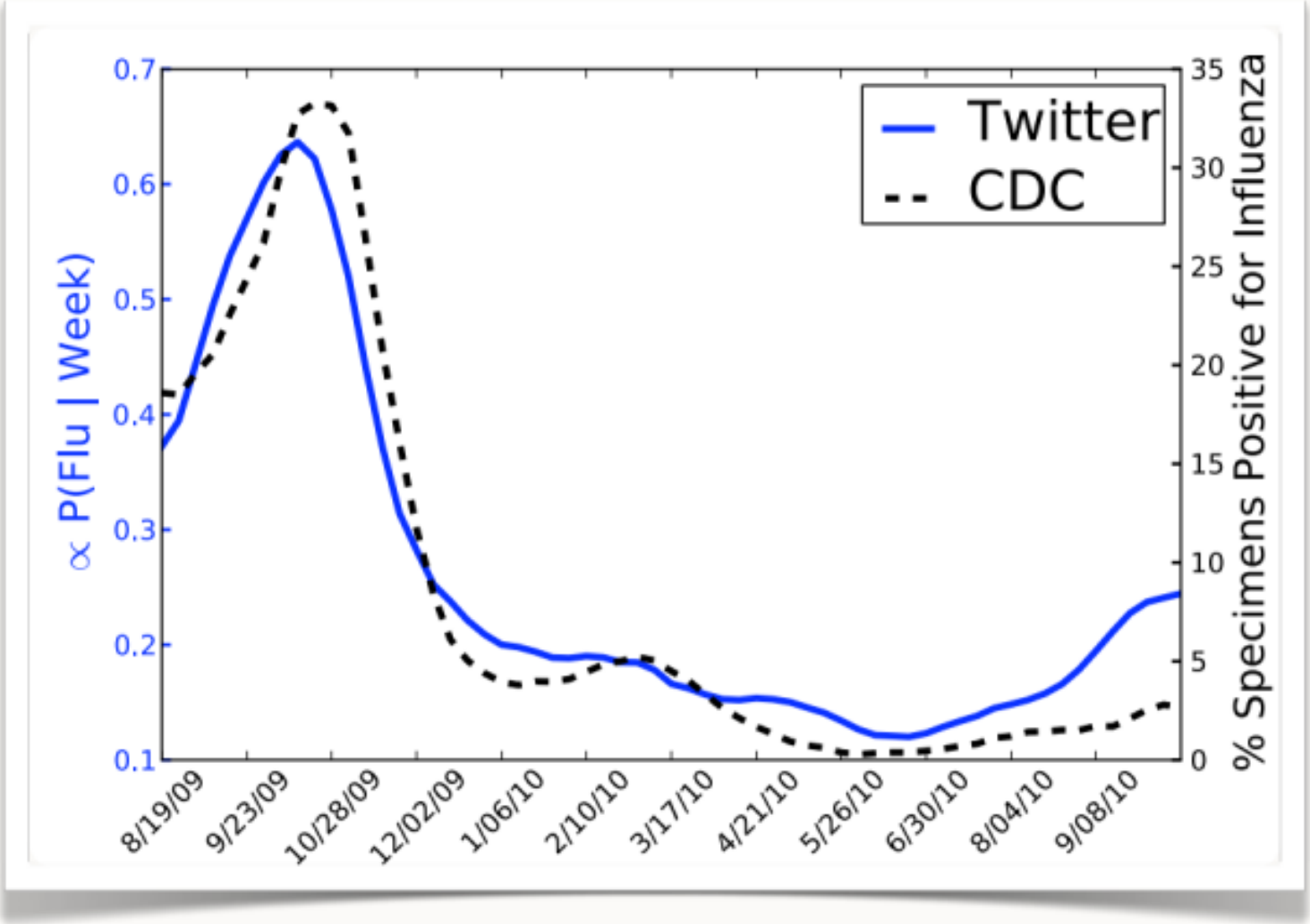


About 1 in 6 U.S. adults use Twitter



Paul & Drezde, 2009

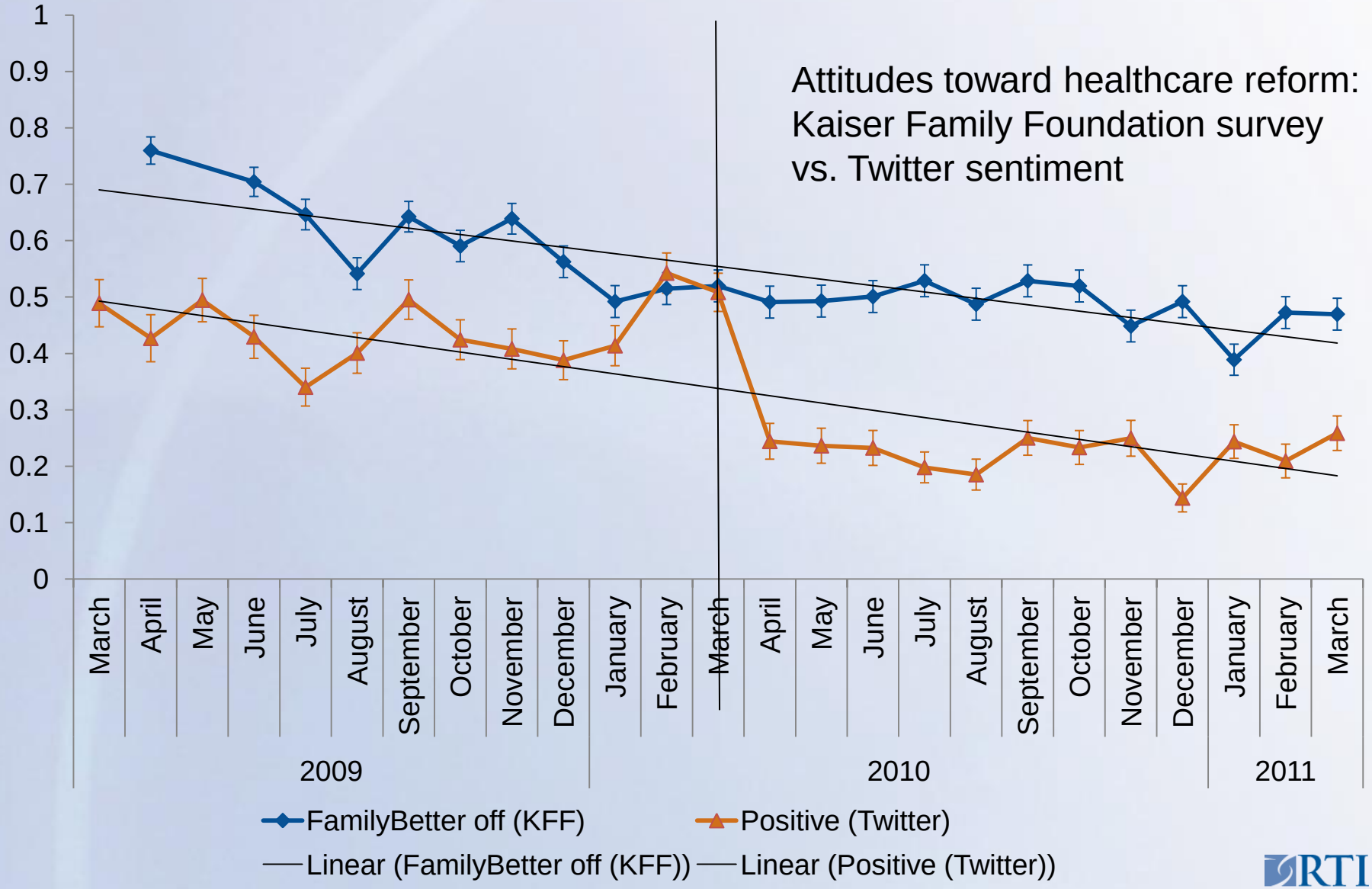
Sometimes Twitter signals and surveys correlate



- Correlation coefficient: 0.958

Kim et al., 2013

Sometimes they don't



Using social media for measurement

- Much emphasis on the *aggregate level*
  - can Twitter produce the same overall estimates as a survey?
  - e.g. Gittelman et al., 2013
- Less attention on the *respondent level*
  - can Twitter fill in the blanks for missing survey responses?
  - e.g. this presentation

Do Tweets give the same information as survey responses?

- Demographics
  - Gender
  - Age
  - Income
- Substantive outcomes
  - Voting behavior
  - Health
  - Depression

Some have already looked into this

- E.g. Schwartz et al. (2013) looked at demographics and personality

- “We extracted words, phrases, and topics (automatically clustered sets of words) from millions of Facebook messages and found the language that correlates most with gender, age, and five factors of personality.”

Why we looked at Twitter and not other social media

- Public Tweets are the norm (vs. private) and can be accessed
- The application programming interface (API) is fairly straightforward and allows for easy gathering of Tweets for analysis
- People Tweet about a huge range of behaviors, opinions, and more (Murthy, 2012)

Our study design

- We conducted a web survey using Knowledge Networks panel with 2,119 respondents
- The survey includes questions on demographics, voting behavior, health, depression, and more
- We asked respondents if they use Twitter. About 19% (398) said yes.
- We asked those who use if they'd share their Twitter handle and allow us to merge in their Tweets. About 27% (107) said yes.

Our study design

- Searched the Twitter API for the 107 respondents providing Twitter username (using twitteR package for R)
- About 80% of handles were found
- Pulled most recent 1,000 tweets
  - Mean Tweets per R=248; Median=78
- Randomly selected 50% and masked their demographics and substantive responses

Our study design

- Some Tweet examples, just to give a flavor:
  - “@BarackObama You make me proud!”
  - “I'm at @JCPenney”
  - “Wondering about donut delivery”
  - “Gigglemonsters make my day”

- For masked cases, predict responses by two methods
  - 1: text mining
  - 2: human prediction

Text mining

- Converted text data to vectors of numbers and identified patterns and trends in the predictor variables associated with the outcome of interest

- k-nearest neighbors (k-NN) methods: projecting points into multi-dimensional space and predict category based on the closest points
  - can handle a large number of predictor variables
  - relatively simple, fast, and scalable
  - can make accurate predictions even when there are highly nonlinear or interactive relationships between the predictor variables and the outcome of interest (Elder et al., 2012).

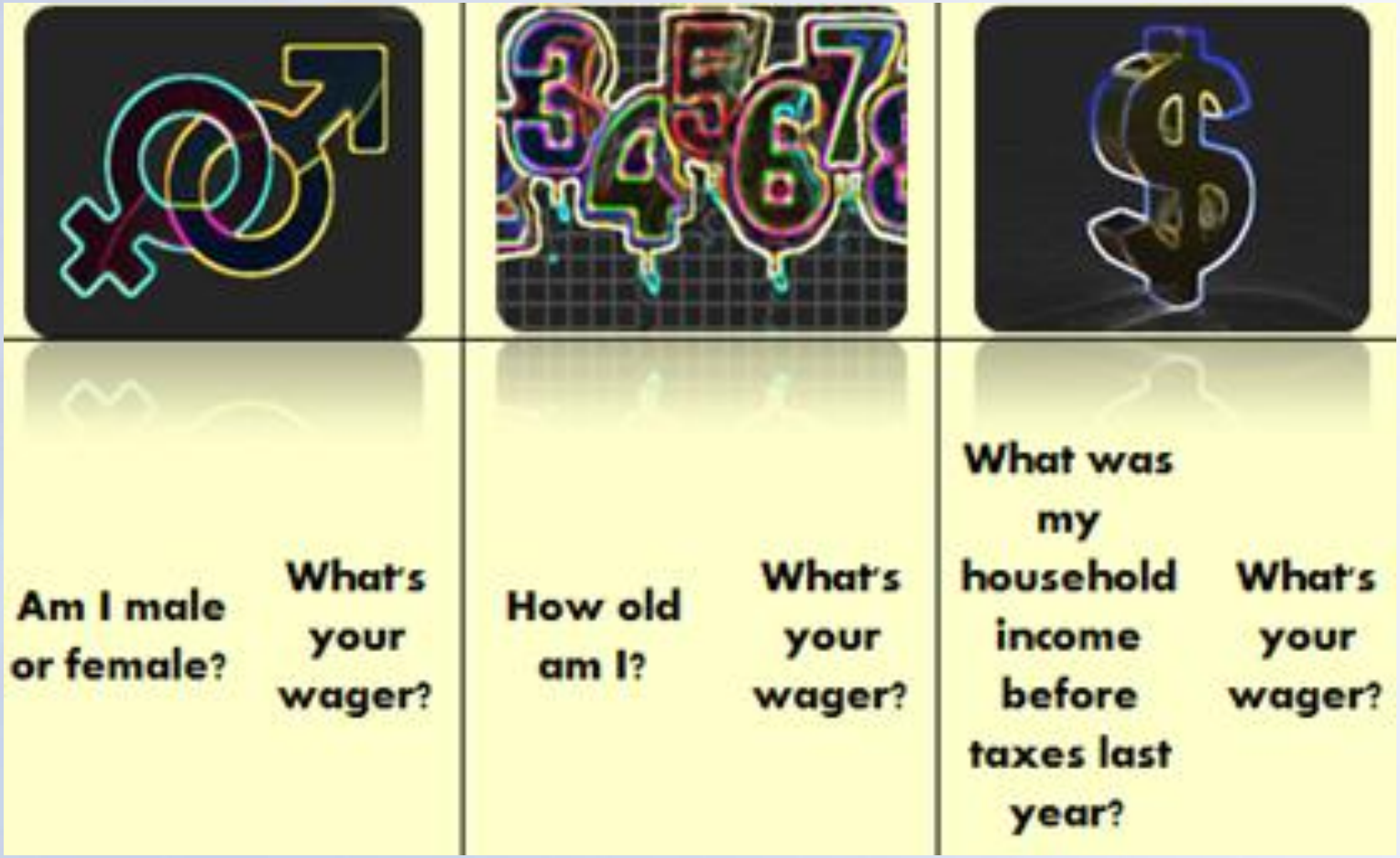
Human prediction

- Provided 3 human predictors with Tweet results only
- Set up like a game: “Fool the Guesser!”
  - 5 mins/R to read and code Tweets
  - Following URLs allowed
  - Include wager (1-5) of guess confidence
  - Answers chosen from a set code frame

Fool the Guesser

**** FOOL THE GUESSER **** How good of a guesser are you? You've got 5 minutes per line -- *no cheating!* Person #:	 Am I male or female? What's your wager?	 How old am I? What's your wager?	 What was my household income before taxes last year? What's your wager?	 Did I vote in the last election? For whom? What's your wager?	 What's my health status? What's your wager?	 How often did I feel fretful, angry, irritable, anxious, or depressed in the last four weeks? What's your wager?
4						
5						
6						
8						
12						
13						
18						

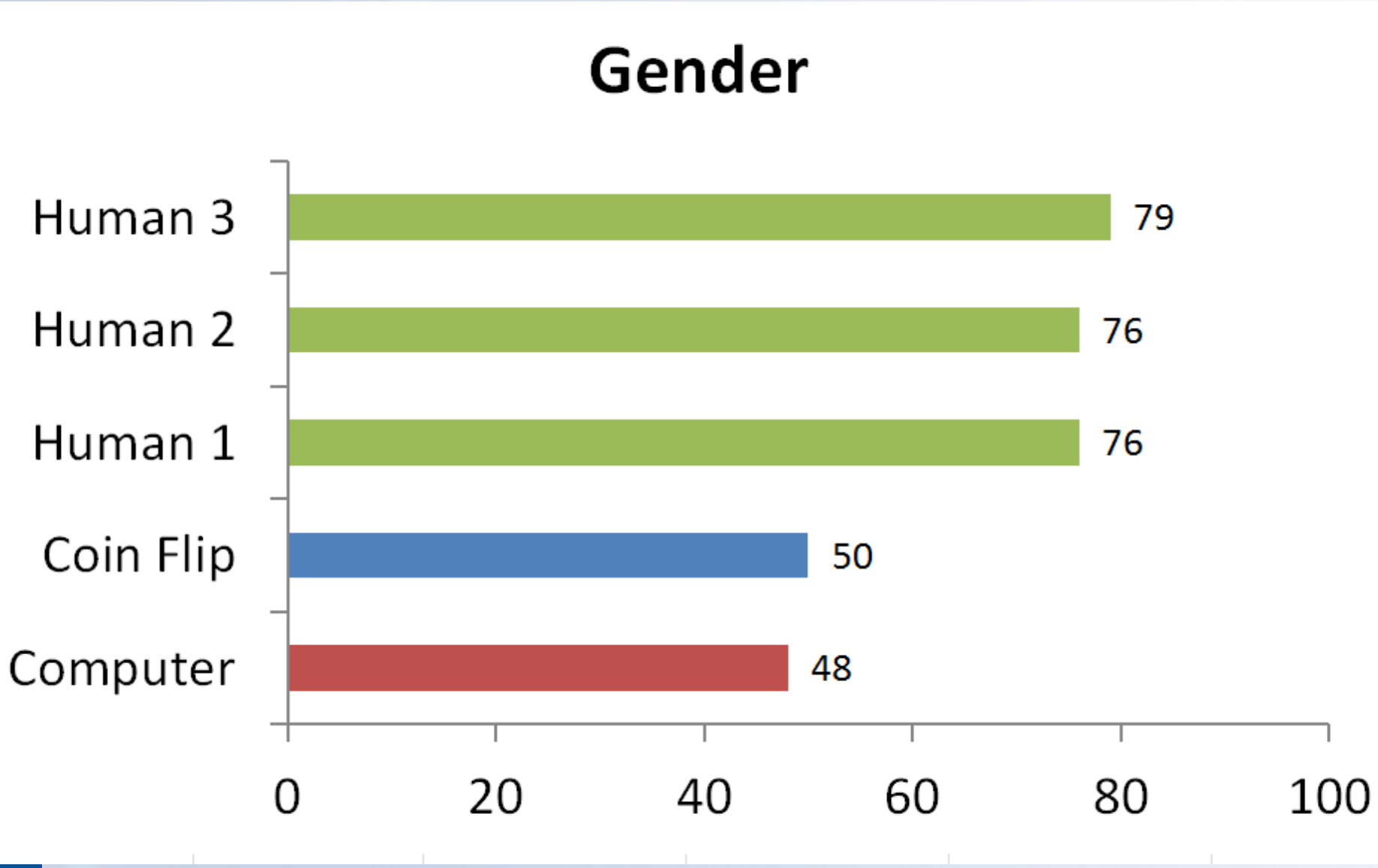
Fool the Guesser



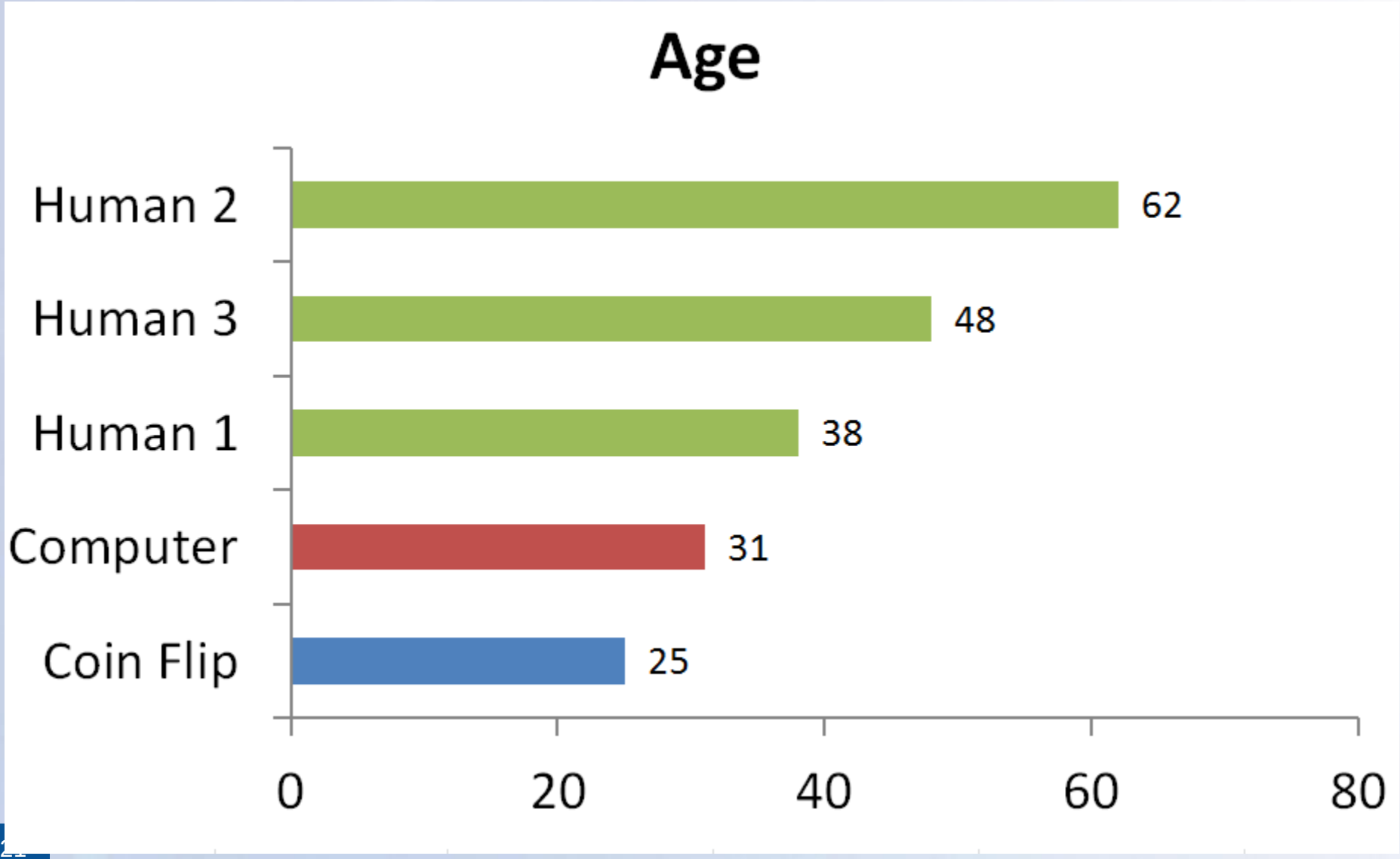
Fool the Guesser

		
Did I vote in the last election? For whom? What's your wager?	What's my health status? What's your wager?	How often did I feel fretful, angry, irritable, anxious, or depressed in the last four weeks? What's your wager?

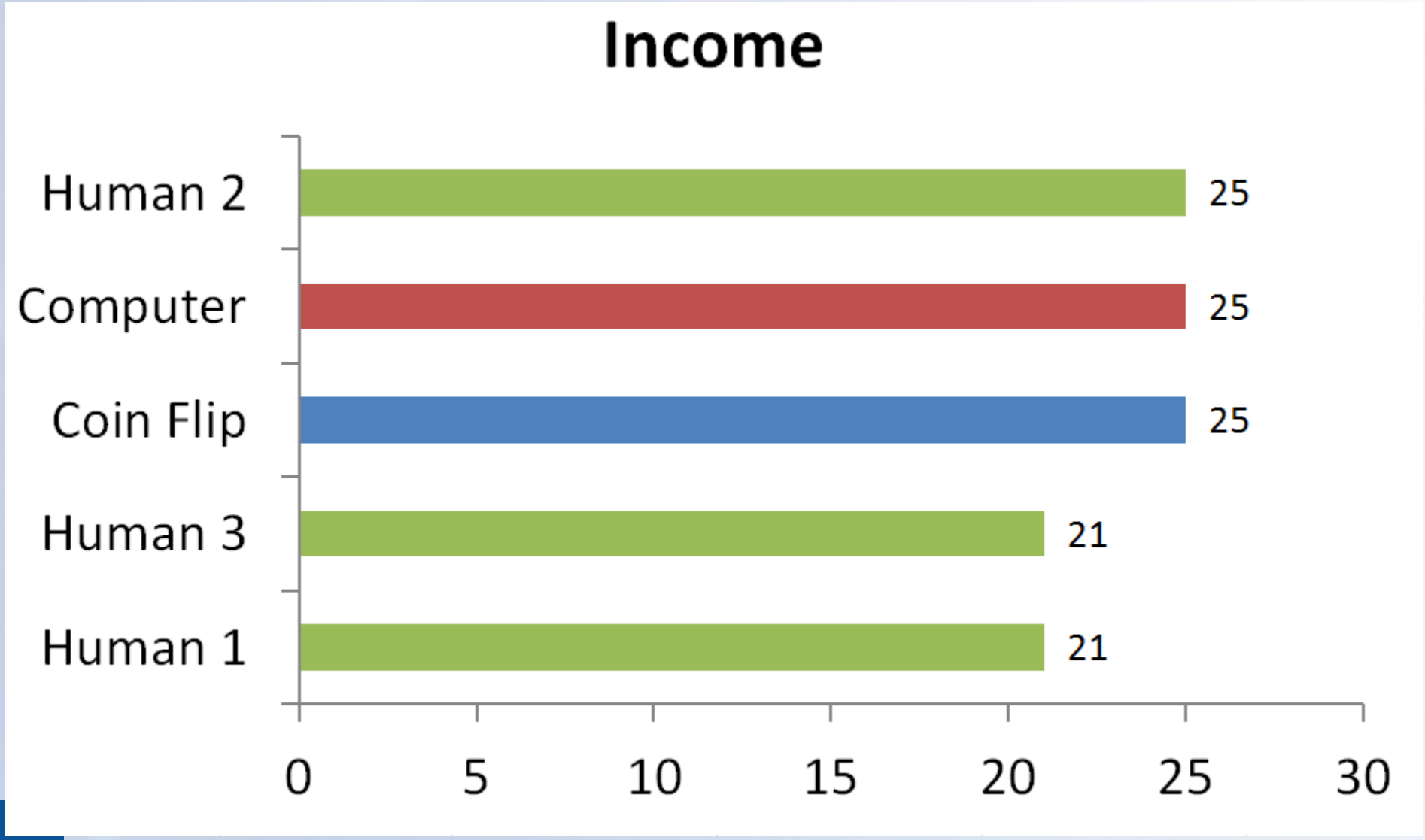
Accuracy (%)


Gender
Human 3
79
Human 2
76
Human 1
76
Coin Flip
50
Computer
48
0
20
40
60
80
100

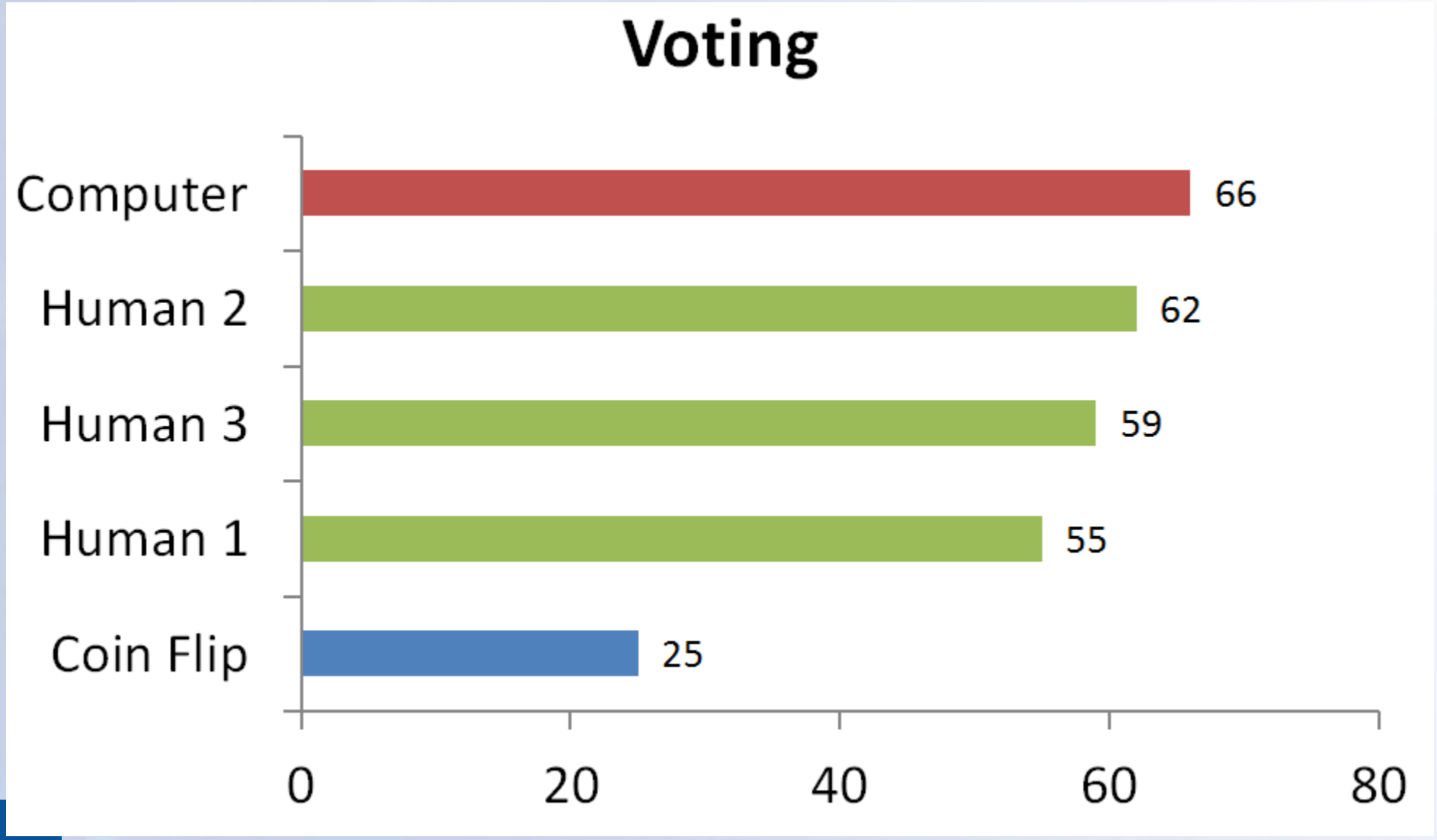
Accuracy (%)


Age
Human 2
62
Human 3
48
Human 1
38
Computer
31
Coin Flip
25
0
20
40
60
80

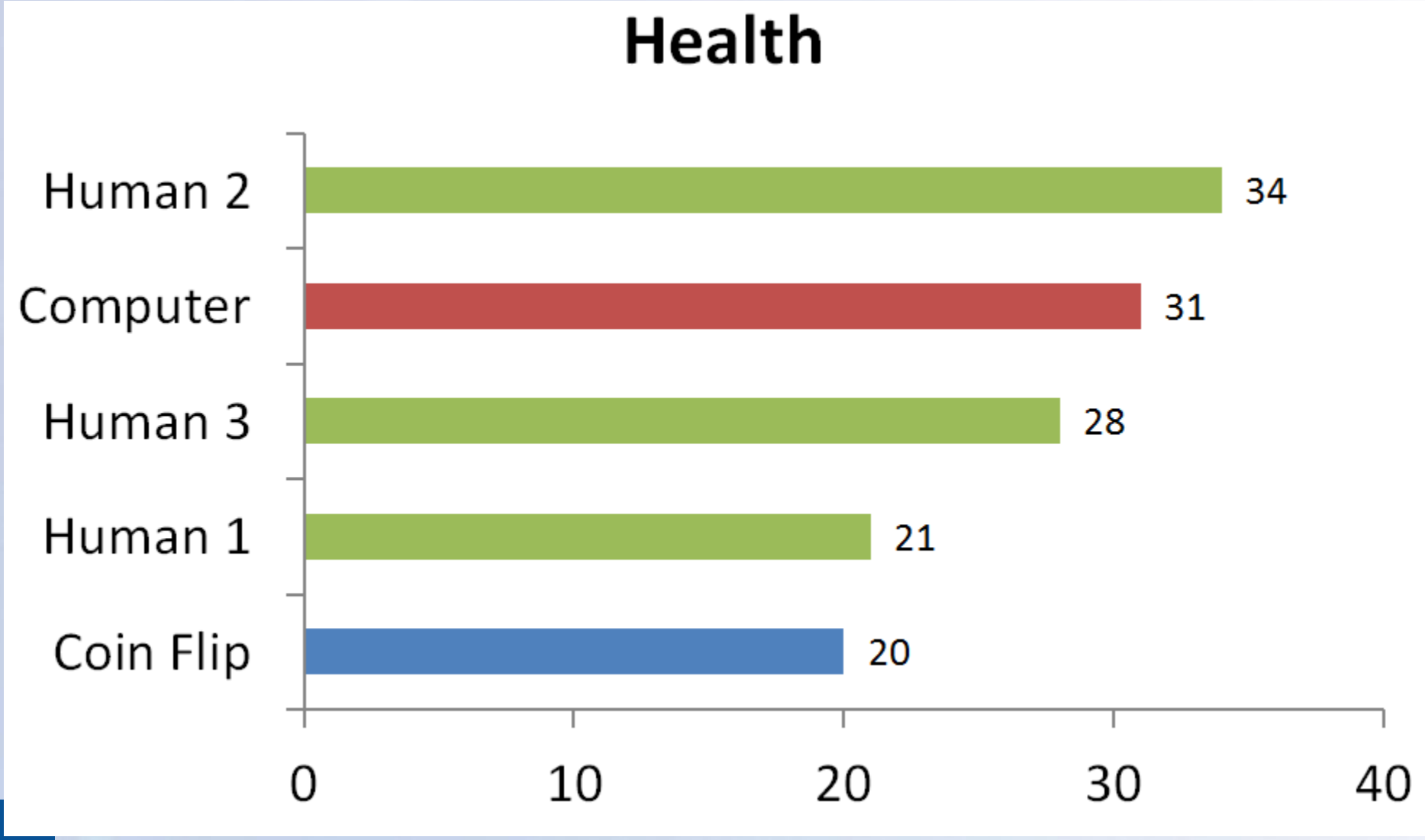
Accuracy (%)


Income
Human 2
Computer
Coin Flip
Human 3
Human 1
25
25
25
21
21
0
5
10
15
20
25
30

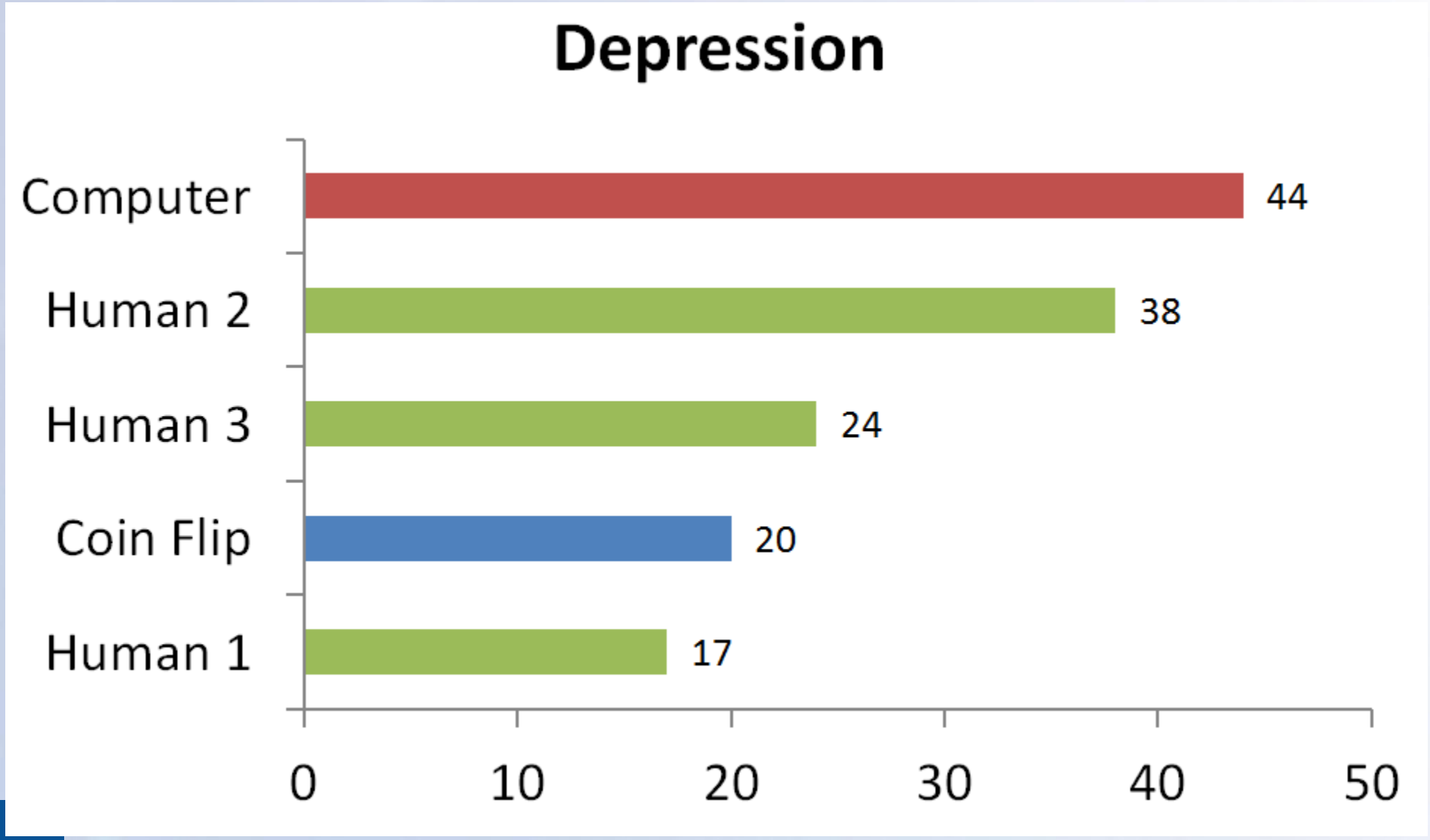
Accuracy (%)


Voting
Computer
66
Human 2
62
Human 3
59
Human 1
55
Coin Flip
25
0
20
40
60
80

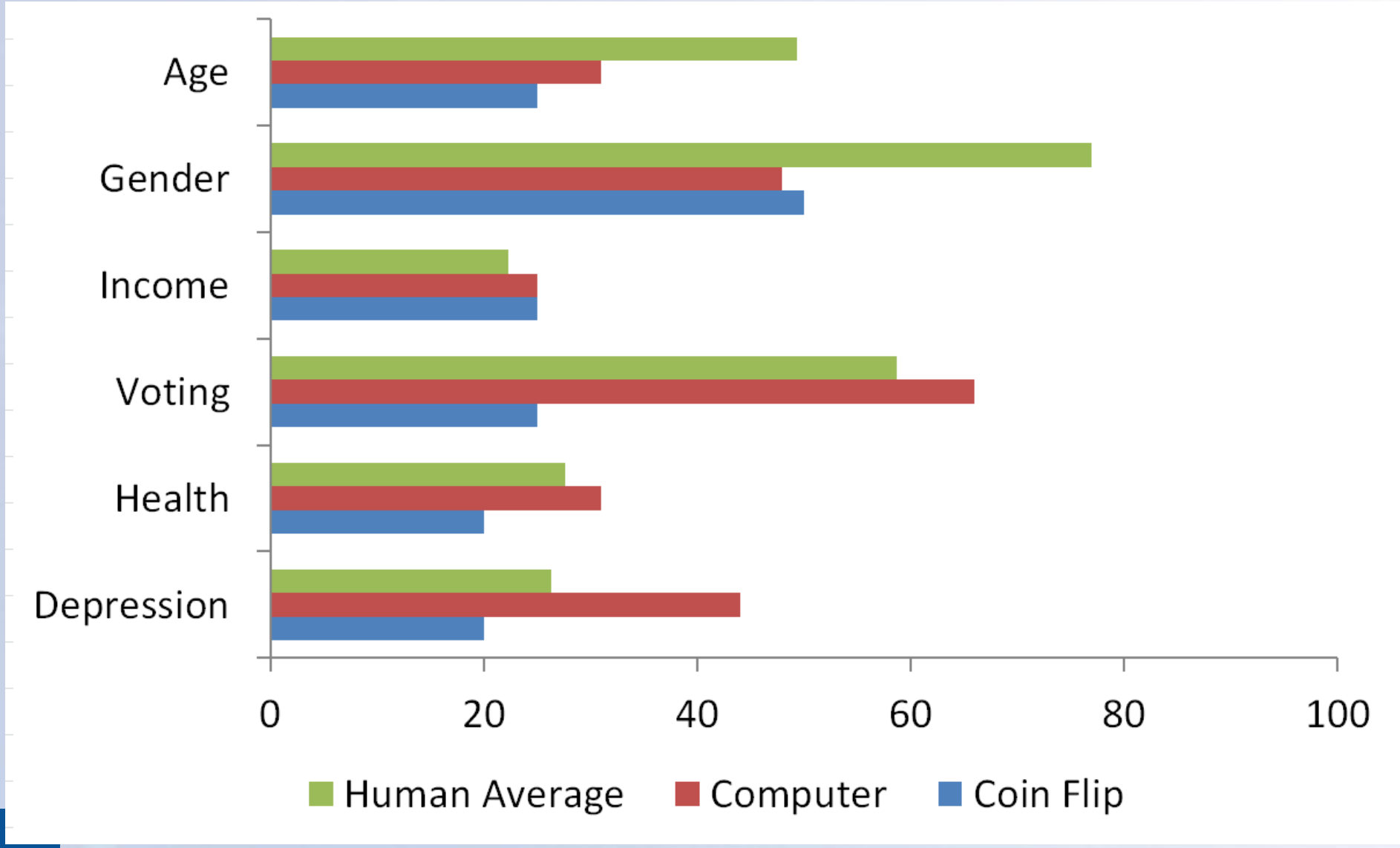
Accuracy (%)


Health
Human 2
34
Computer
31
Human 3
28
Human 1
21
Coin Flip
20
0
10
20
30
40

Accuracy (%)


Depression
Computer
Human 2
Human 3
Coin Flip
Human 1
44
38
24
20
17
0
10
20
30
40
50

Accuracy (%)


Age
Gender
Income
Voting
Health
Depression
0
20
40
60
80
100
Human Average
Computer
Coin Flip

What does this mean?

- Even with a small set of respondents, Tweets may add some information at the respondent level to help predict missing characteristics or outcomes
- For basic demographics (stereotypes?) human prediction worked better than the computer. The computer did better with more “hidden” characteristics like health
- As Twitter popularity grows (and if survey participation continues to decline), this method could prove valuable in imputing missing values, assuming permission is granted

Take this with some grains of salt…

- Most in the sample didn't Tweet. Most who did wouldn't let us access their Tweets.
- Tweeters differ from non-Tweeters (e.g. younger)
- Text mining much more effective with "big data"
- Computer was trained, humans weren't
- We haven't yet compared this to standard methods for imputing missing data.

But..!

- Twitter data are primarily public
- Access to recent Tweets is free
- More and more are sharing online (and not in surveys)
- We need to continue to explore creative, practical applications for social media as a new and potentially powerful resource in research

References

- Elder, John, et al. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.
- Kim, A., Murphy, J., Richards, A., Hansen, H., Powell, R., and Haney, C. (2013). Can Tweets Replace Polls? A U.S. Health Care Reform Case Study in *Social Media, Sociality, and Survey Research. Hill*, C.A., Dean, E., and Murphy, J. eds. Wiley.
- Gittleman, S. et al. (2013). A New Source of Health Data: Facebook Likes. http://www.mktginc.com/pdf/AAPOR2013facebookMethodsAnalysisv1.pdf
- Murthy, D. (2012). *Twitter: Social Communication in the Twitter Age.* Polity.
- Paul, M .P., and Dredze, M. (2011) You Are What You Tweet: Analyzing Twitter for Public Health. *International Conference on Weblogs and Social Media (ICWSM)*.
- Pew / Duggan, M. & Brenner, J. (2013). The Demographics of Social Media Users — 2012. http://www.pewinternet.org/Reports/2013/Social-media-users.aspx
- Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, et al. (2013) Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE 8(9): e73791. doi:10.1371/journal.pone.0073791

Thanks for your attention!

Joe Murphy
jmurphy@rti.org

SurveyPost
RTI experts on the future of social research
blogs.rti.org/surveypost

Follow our blog for the latest on social media, new technologies, and social research.

twitter @SurveyPost