Why Asking the Right Questions is Better than Having All the Answers

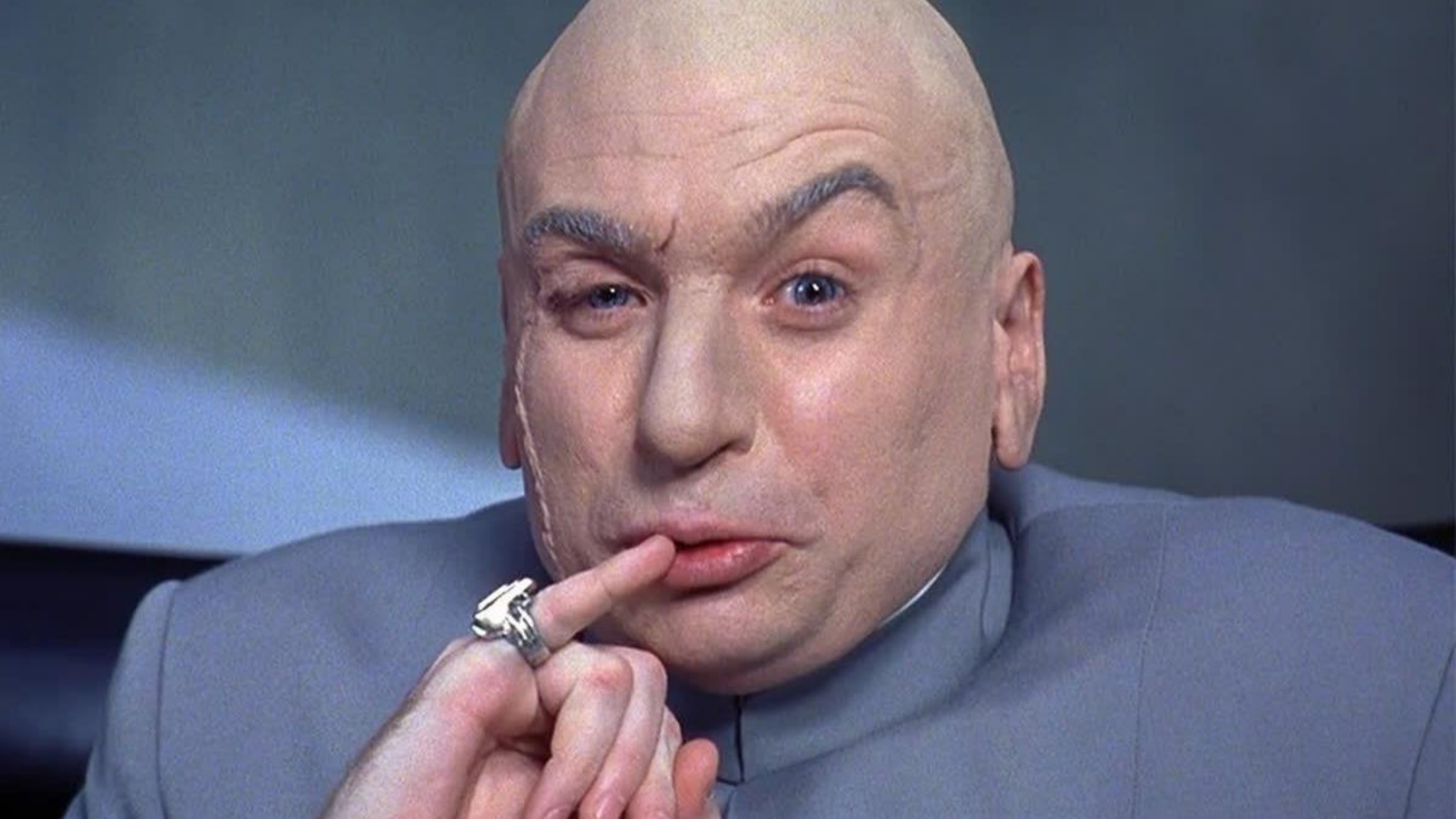
Neal Sample

Chief Information Officer, Walgreens Boots Alliance

#1

What was going on when the data was created?


NETFLIX
The
NETFLIX
NETFLIX
NETFLIX
NETFLIX



100,480,507 ratings

480,189 users

17,770 movies

<user #, movie #, date of rating, rating>


<user #, movie #, date of rating, rating>


2009
NETFLIX
DATE: 09·21·09
PAY TO THE ORDER OF: BellKor's Pragmatic Chaos
$ 1,000,000 00
AMOUNT: ONE MILLION 00/100
FOR: The Netflix Prize
Reed Hastings

#2

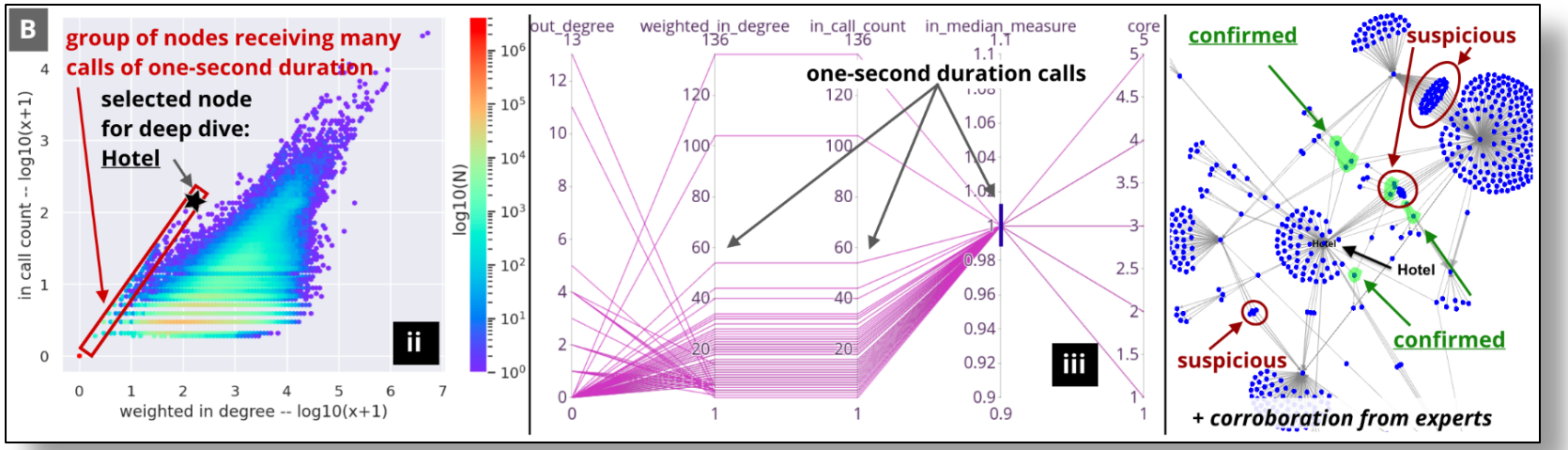
What information can we mix in?


CENTRAL INTELLIGENCE AGENCY
UNITED STATES OF AMERICA

Caller SIM	Callee SIM	Outgoing BTS	Incoming BTS	Timestamp	Call duration (sec)
0458685984	0488595496	12	365	2018-01-18 15:22:12	456
0458685984	0458685984	12	25	2018-01-18 22:24:12	35
0469875254	0498563201	879	567	2018-01-19 08:47:10	125
(...)	(...)	(...)	(...)	(...)	(...)

Short Calls
Burners

Diversity
Connectedness



Drugs
Spam

Fraud
Terrorism

Dynamic Network Anomaly Modeling of Cell-Phone Call Detail Records for Infectious Disease Surveillance

Authors: Carl Yang, Hongwen Song, Mingyue Tang, Leon Danon, Ymir Vigfusson Authors Info & Claims

KDD '22: Proceedings of the 28th ACM SIGKDD Conference on Knowl... • August 2022 • Pages 4733–4742 • https://doi.org/10.1145/3534678.3542678

Published: 14 August 2022 Publication History

Faster outbreak detection when phone calls are monitored: Pä...

Bjelkmar FREE

P Bjelkmar, E van Straten, L Brouwers

European Journal of Public Health, Volume 26, Issue suppl_1, November 2016, ckw166.072, https://doi.org/10.1093/eurpub/ckw166.072

Published: 02 November 2016

Cell-phone traces reveal infection-associated behavioral change

Ymir Vigfusson[a,b,1,2], Thorgeir A. Karlsson[b,1], Derek Onken[a,1], Congzheng Song[c], Atli F. Einarsson[b], Nishant Kishore[d], Rebecca M. Mitchell[a], Ellen Brooks-Pollock[e], Gudrun Sigmundsdottir[f,g], and Leon Danon[h,i]

[a]Simbiosys Lab, Department of Computer Science, Emory University, Atlanta, GA 30322; [b]School of Computer Science, Reykjavik University, 101 Reykjavik, Iceland; [c]Department of Computer Science, Cornell University, Ithaca, NY 14853; [d]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [e]Department of Veterinary Medicine and Population Health Sciences, University of Bristol, Oakfield Grove, Bristol BS8 2BN, United Kingdom; [f]Landspitali University Hospital, 101 Reykjavik, Iceland; [g]Centre for Health Security and Communicable Disease Control, 101 Reykjavik, Iceland; [h]Department of Engineering Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom; and [i]The Alan Turing Institute, British Library, London NW1 2DB, United Kingdom.

Edited by Nils Chr. Stenseth, University of Oslo, Oslo, Norway, and approved December 16, 2020 (received for review March 19, 2020)

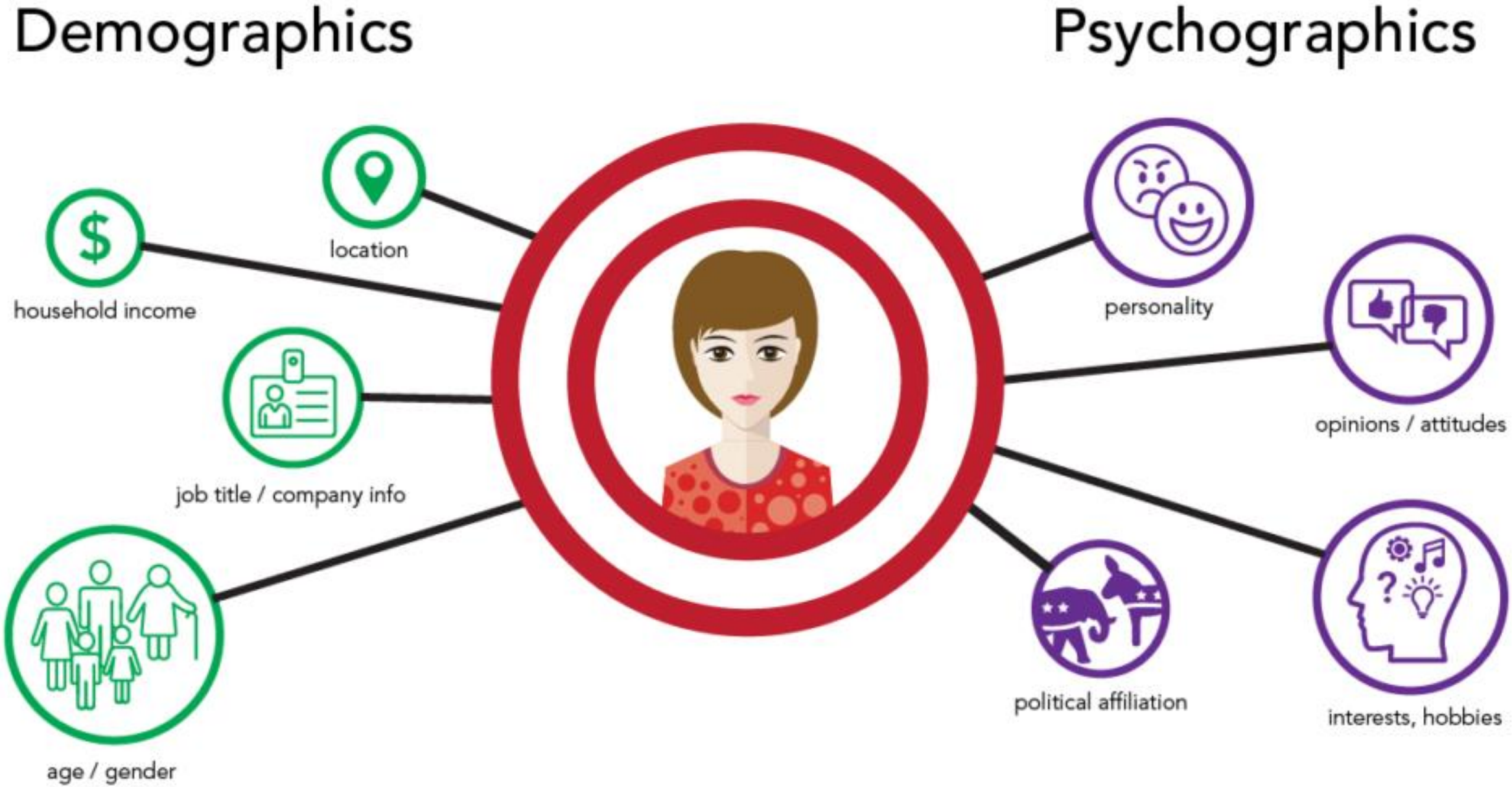
Evaluating the privacy properties of telephone metadata

Jonathan Mayer ✉, Patrick Mutchler, and John C. Mitchell Authors Info & Affiliations

Edited by Cynthia Dwork, Microsoft Research Silicon Valley, Mountain View, CA, and approved March 1, 2016 (received for review April 27, 2015)

May 16, 2016 | 113 (20) 5536-5541 | https://doi.org/10.1073/pnas.1508081113

$19.95 + 70min = 82%


Demographics
Psychographics
location
household income
job title / company info
age / gender
personality
opinions / attitudes
political affiliation
interests, hobbies

- Participant B received a long phone call from the cardiology group at a regional medical center, talked briefly with a **medical laboratory**, answered several short calls from a local drugstore, and made brief calls to a self-reporting hotline for a cardiac arrhythmia monitoring device.

- Participant C placed frequent calls to a local firearm dealer that prominently advertises a specialty in the **AR semiautomatic rifle platform**. He also placed lengthy calls to the customer support hotline for a major firearm manufacturer; the manufacturer produces a popular AR line of rifles.

- Participant D placed calls to a **hardware outlet, locksmiths, a hydroponics store, and a head sho**p in under 3 weeks.

- Participant E made a lengthy phone call to her sister early one morning. Then, 2 days later, she called a nearby **Planned Parenthood** clinic several times. Two weeks later, she placed brief additional calls to Planned Parenthood, and she placed another short call 1 month after.

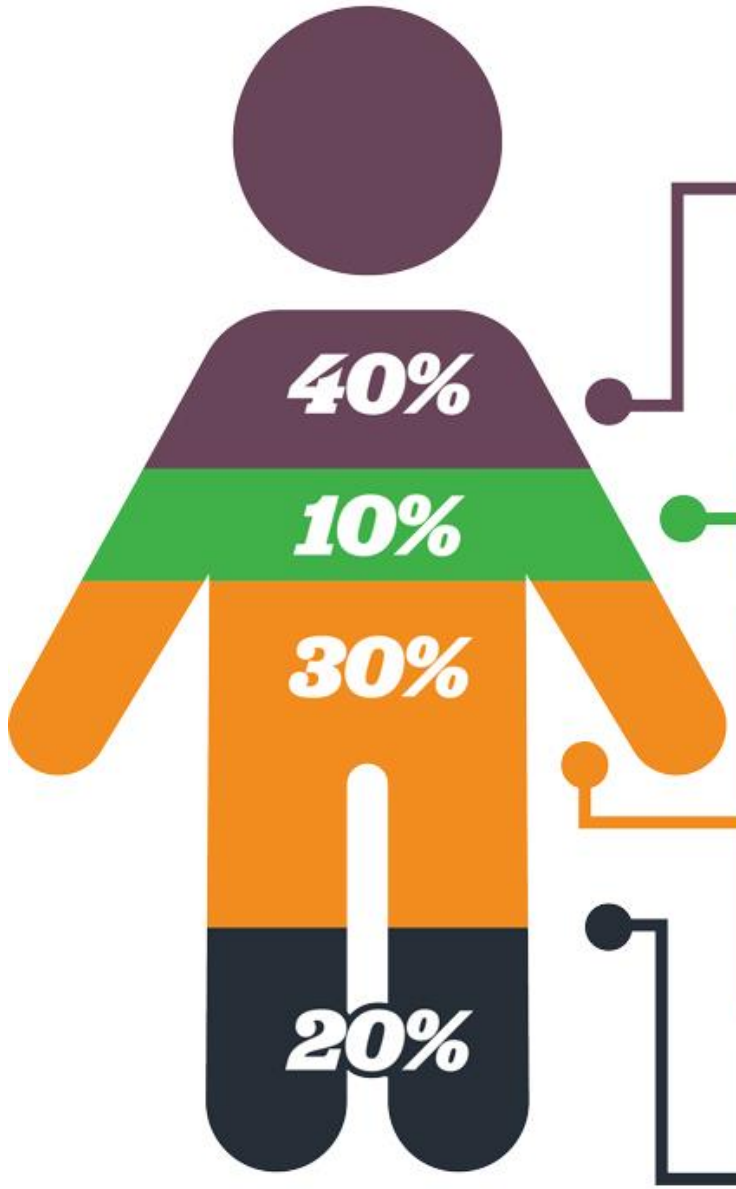
“Using public sources, we were able to confirm that participant B **had a cardiac arrhythmia** and participant C **owned an AR rifle**.”

#3

What are people thinking about in this context?


EXPRESS SCRIPTS®
1 Express Way

Why aren’t people taking their medication?



Socioeconomic Factors

Education | Job Status | Family Social Support | Income | Community Safety

Physical Enviroment

Health Behaviors

Tobacco Use | Diet & Excercise | Alcohol Use | Sexual Activity

Health Care

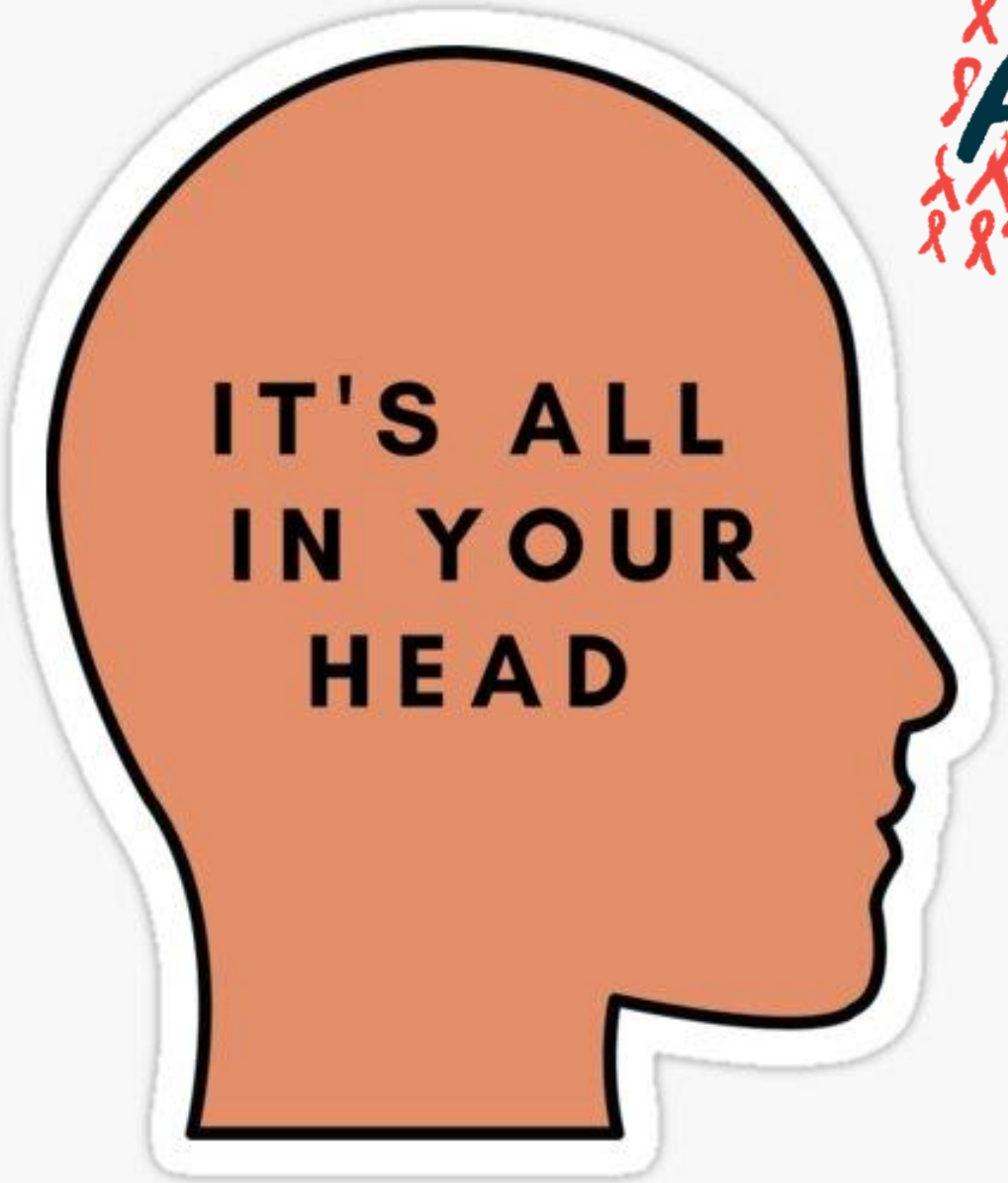
Access to Care
Quality of Care

50% can be traced back to your zip code!

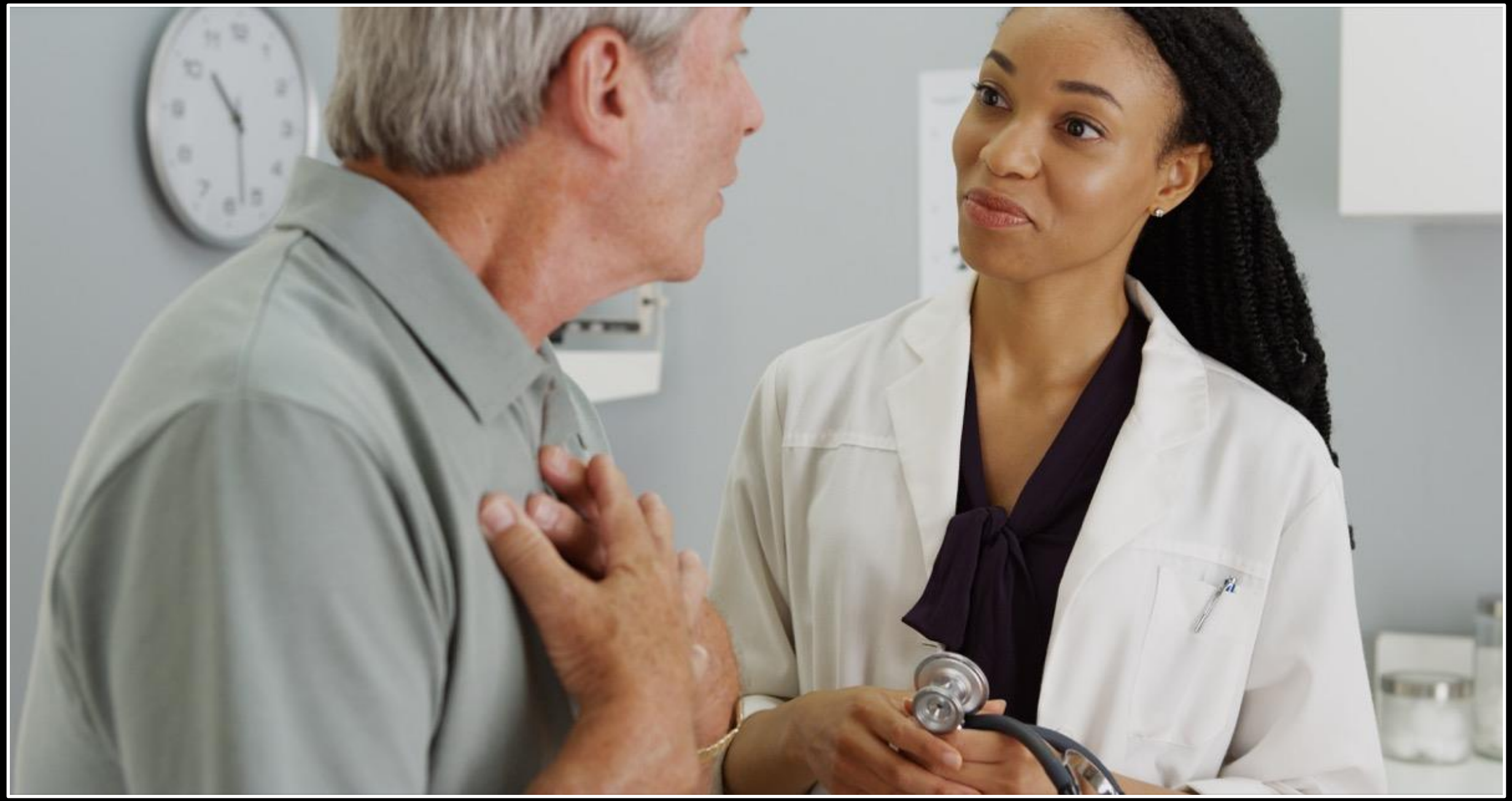
Only 20% include those moments in a healthcare enviroment.

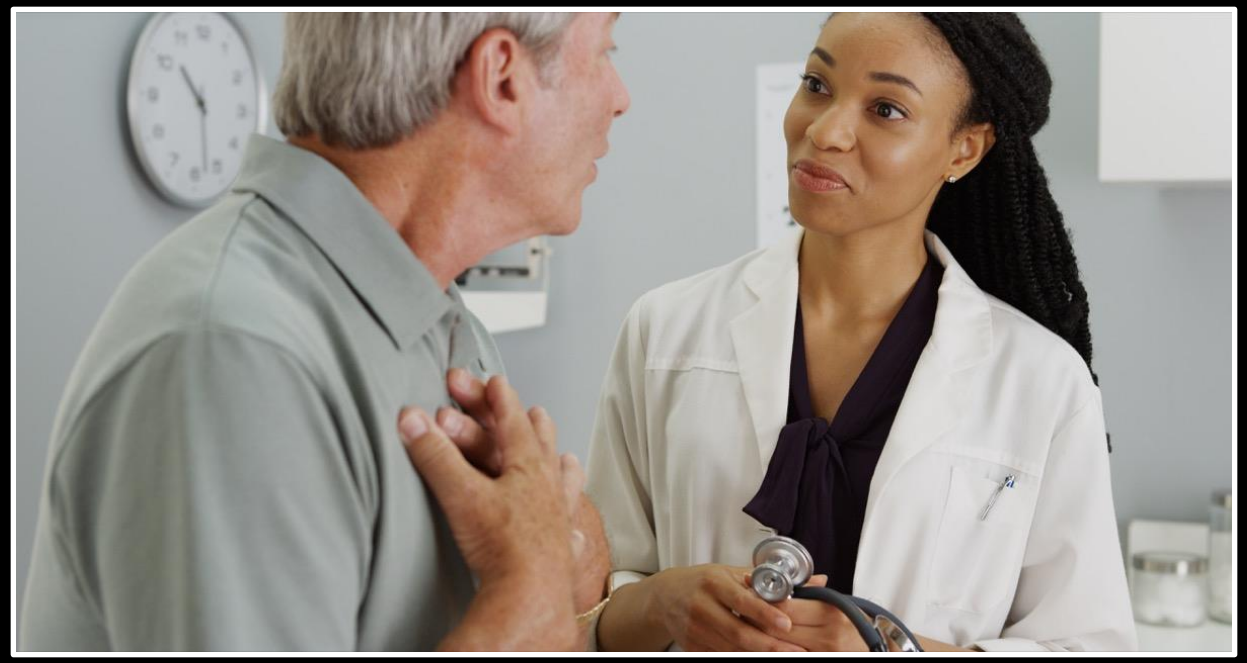
Source: Institute for Clinical Systems Improvement, Going Beyond Clinical Walls: Solving Complex Problems (October 2014)

STRESS!



Illness
vs
Disease





70 bps

Male patient

> 40 years age

Female Physician

Some questions to improve answers:

1. What was going on when the data was created?

2. What information can we mix in?

3. What are people thinking about in this context?

Some questions to improve answers:

1. What was going on when the data was created?
   <user #, movie #, date of rating, rating>

2. What information can we mix in?
   Internet Movie Database (IMDB)

3. What are people thinking about in this context?
   Find out their politics, beliefs, and secrets

Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan and Vitaly Shmatikov
The University of Texas at Austin

We apply our de-anonymization methodology to the **Netflix Prize dataset**, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the **Internet Movie Database** as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their **apparent political preferences** and other **potentially sensitive information**.

RYAN SINGEL SECURITY MAR 12, 2010 2:48 PM

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Those fears were highlighted in December, when **an in-the-closet lesbian mother sued Netflix for privacy invasion**, alleging the movie-rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its $1 million contest.

Some questions to improve answers:

1. What was going on when the data was created?

2. What information can we mix in?

3. What are people thinking about in this context?

Thank You!