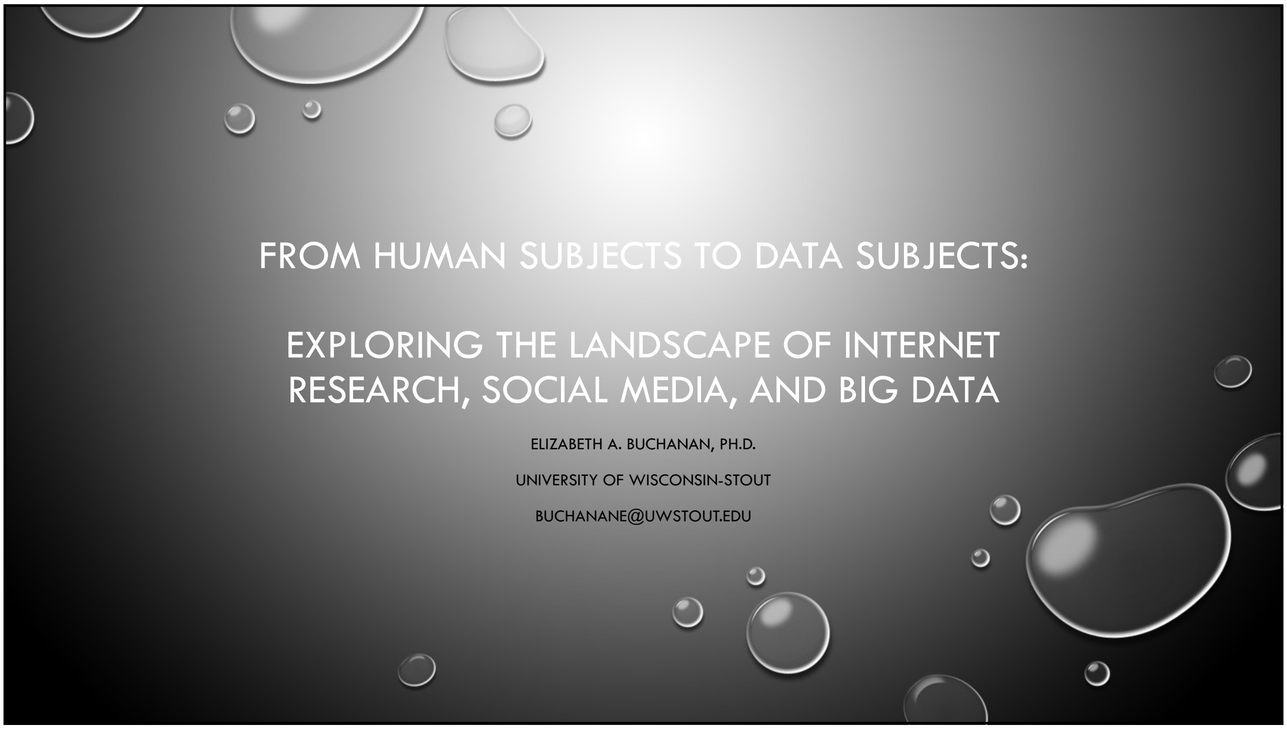


FROM HUMAN SUBJECTS TO DATA SUBJECTS:

EXPLORING THE LANDSCAPE OF INTERNET RESEARCH, SOCIAL MEDIA, AND BIG DATA

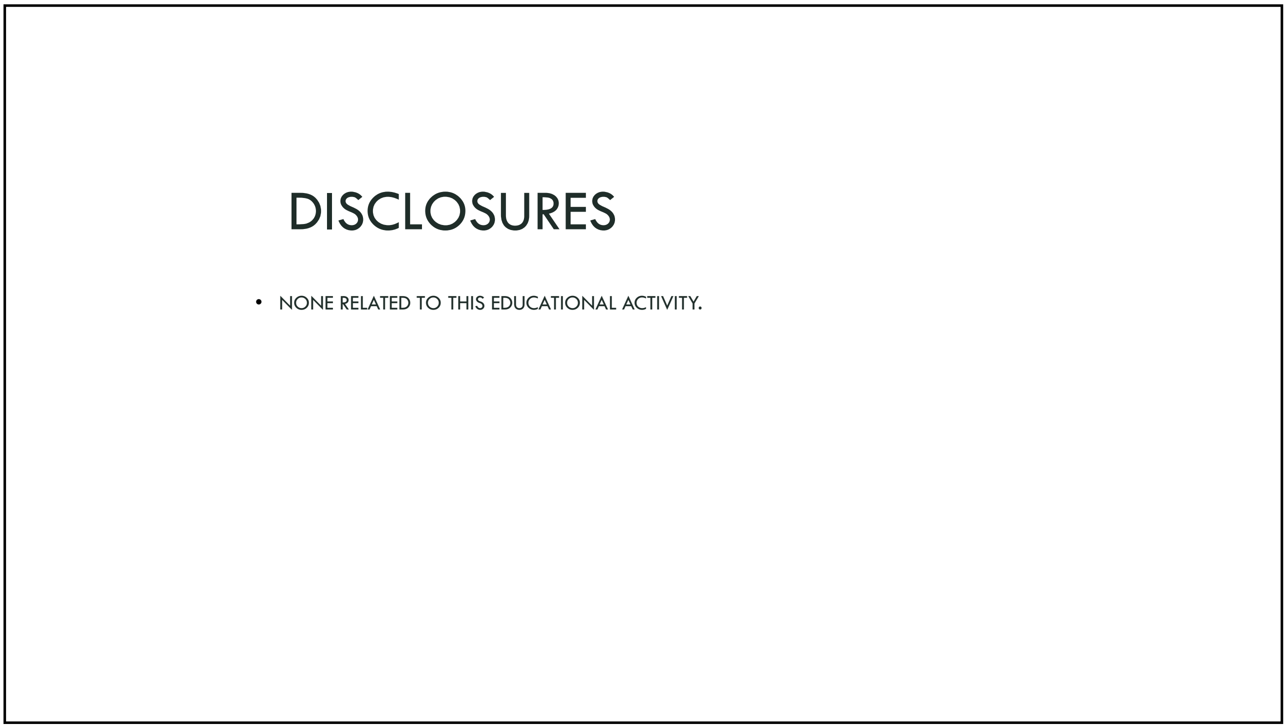
ELIZABETH A. BUCHANAN, PH.D.

UNIVERSITY OF WISCONSIN-STOUT

BUCHANANE@UWSTOUT.EDU

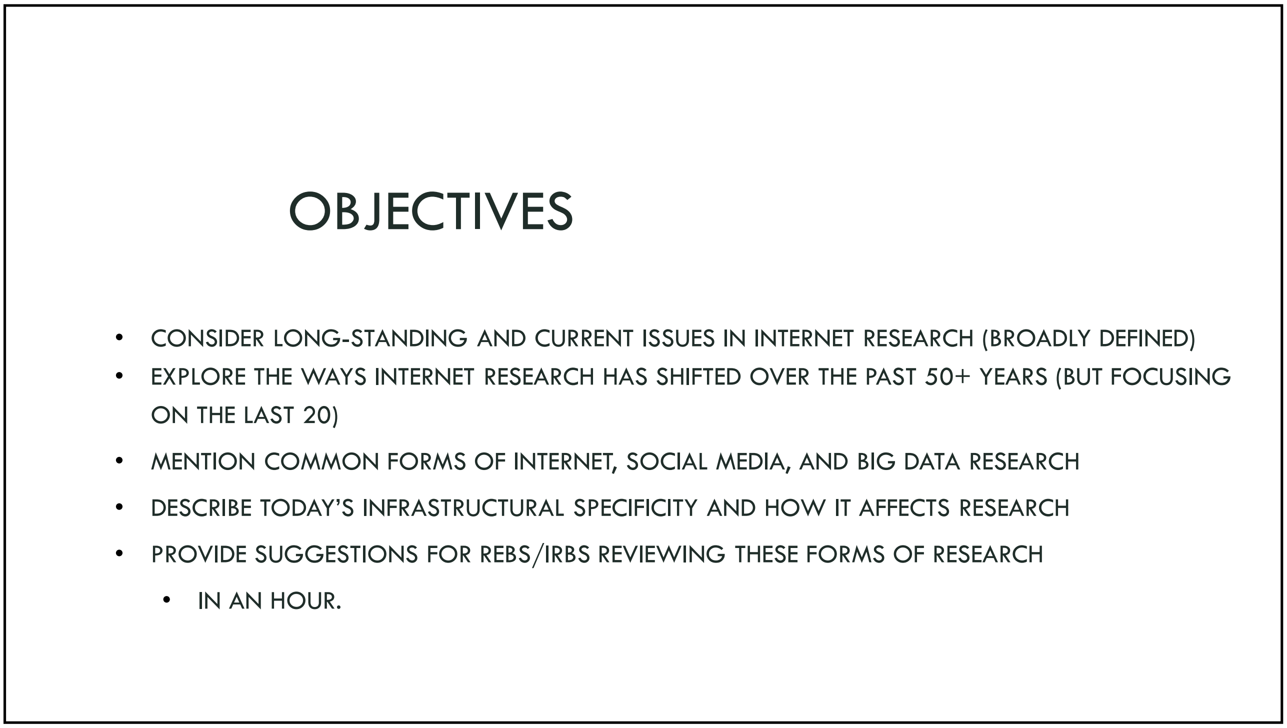
ACKNOWLEDGEMENTS

- DR. BRUCE GORDON
- ABBEY LOWE
- LINDSAY HICKS
- UNMC



DISCLOSURES

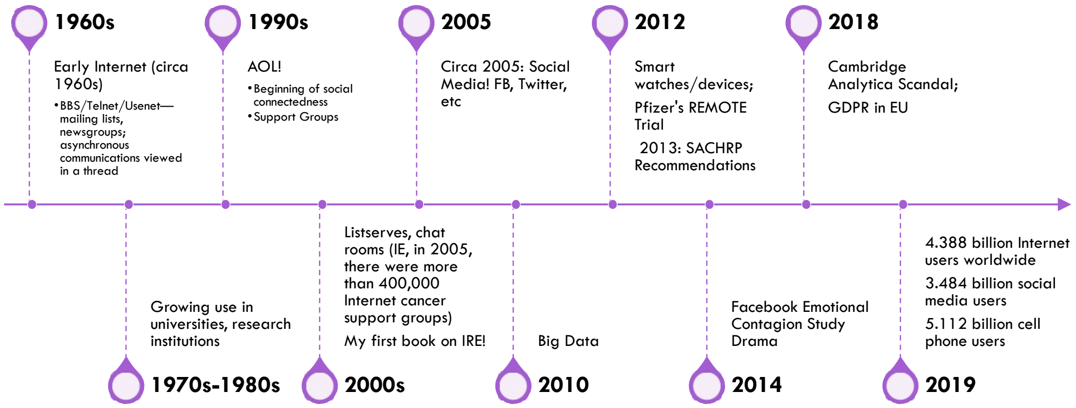
- NONE RELATED TO THIS EDUCATIONAL ACTIVITY.



OBJECTIVES

- CONSIDER LONG-STANDING AND CURRENT ISSUES IN INTERNET RESEARCH (BROADLY DEFINED)
- EXPLORE THE WAYS INTERNET RESEARCH HAS SHIFTED OVER THE PAST 50+ YEARS (BUT FOCUSING ON THE LAST 20)
- MENTION COMMON FORMS OF INTERNET, SOCIAL MEDIA, AND BIG DATA RESEARCH
- DESCRIBE TODAY'S INFRASTRUCTURAL SPECIFICITY AND HOW IT AFFECTS RESEARCH
- PROVIDE SUGGESTIONS FOR REBS/IRBS REVIEWING THESE FORMS OF RESEARCH
  - IN AN HOUR.

A (VERY) BRIEF HISTORY



1960s

Early Internet (circa 1960s)

- BBS/Telnet/Usenet—mailing lists, newsgroups; asynchronous communications viewed in a thread

1970s-1980s

Growing use in universities, research institutions

1990s

AOL!

- Beginning of social connectedness
- Support Groups

2000s

Listserves, chat rooms (IE, in 2005, there were more than 400,000 Internet cancer support groups)

My first book on IRE!

2005

Circa 2005: Social Media! FB, Twitter, etc

2010

Big Data

2012

Smart watches/devices;

Pfizer's REMOTE Trial

2013: SACHRP Recommendations

2014

Facebook Emotional Contagion Study Drama

2018

Cambridge Analytica Scandal;

GDPR in EU

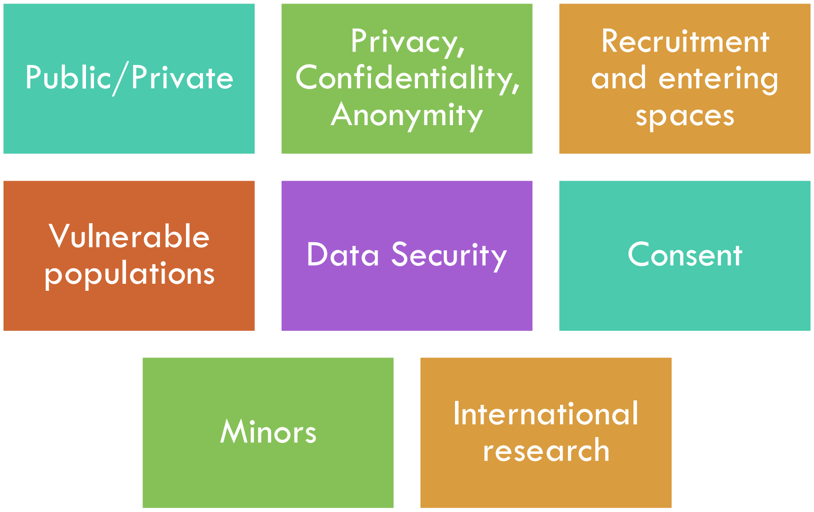
2019

4.388 billion Internet users worldwide

3.484 billion social media users

5.112 billion cell phone users

EARLY INTERNET RESEARCH CONSIDERATIONS



Public/Private

Privacy, Confidentiality, Anonymity

Recruitment and entering spaces

Vulnerable populations

Data Security

Consent

Minors

International research

EARLY CASES (CIRCA 2000-2007)

We had a researcher using the website "Gay Bombay" to study gay Indian men's attitudes, and the board was worried that since homosexuality is illegal in India, would participation get the respondents in trouble somehow?

A male psychologist posed as a disabled woman to study the relationships developed in an online group. He did not disclose his researcher status.

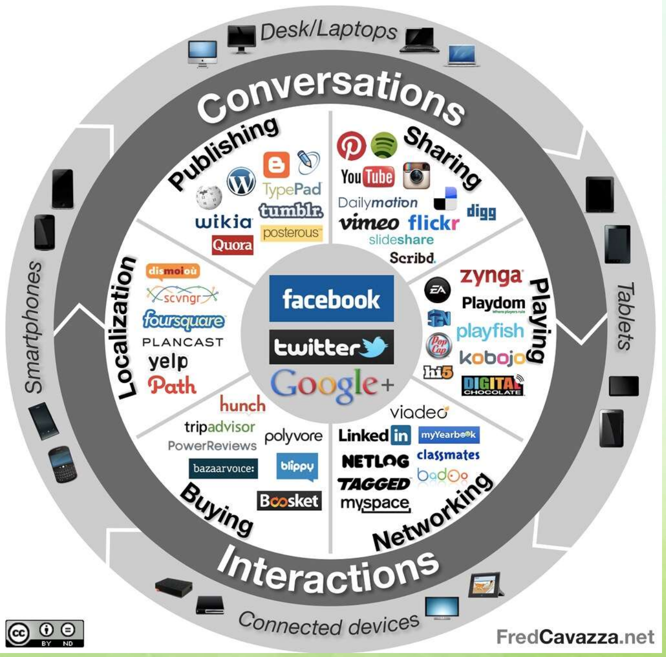
Can a researcher join a particular health discussion board in order to recruit subjects?

A student wishes to analyze blog postings as part of her Master’s thesis. Must she seek IRB review? If she does not, might she face journals who will not publish her work because it was not approved?

A researcher wanted to use a public list archive, but—in order to post, membership was required. Must he gain consent? (No longer fits the “public park” analogy?)

Can a researcher use mechanical turk (to complete research related tasks, eg, survey responses) without IRB oversight?

THEN, THE SOCIAL MEDIA ONSLAUGHT



AND, THESE SORTS OF QUESTIONS/CASES (2007-PRESENT):

Can I use social media to google or FB to follow up with subjects, or locate them?

Can I use "incidental" data?

Is social listening just the new public park?

Isn't all of Twitter public?

Does it matter what a TOS/EULA says?

Who is a secondary subject?

What is a research bystander?

Does a research team have to monitor 24-7?

Can the research team restrict conversations on social media?

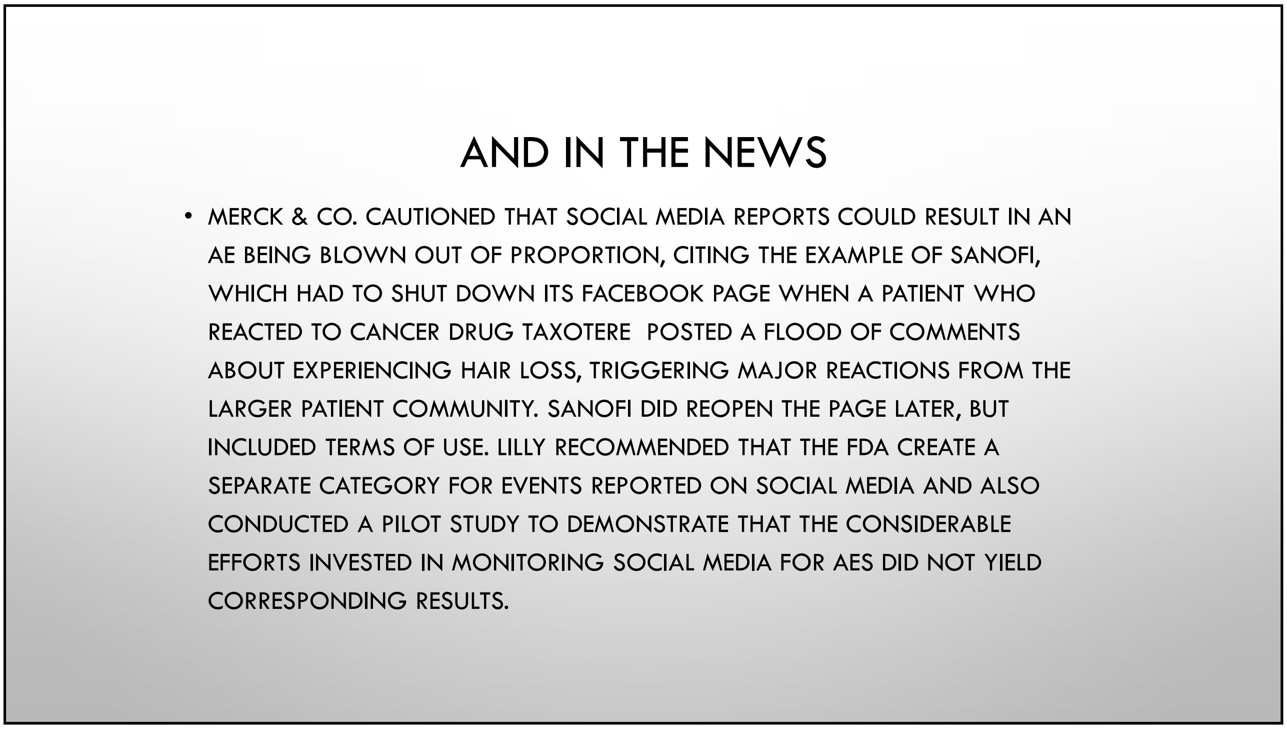
What the heck is big data?

BUT PARTICULARLY IN CLINICAL RESEARCH:

- "AS USE OF ONLINE NETWORKS CONTINUES TO RISE, RESEARCH SPONSORS AND REGULATORS MUST BEGIN STUDYING THE IMPLICATIONS OF SOCIAL MEDIA ON THE INTEGRITY OF CURRENT BLINDED AND RANDOMIZED CLINICAL TRIALS" (LIPSET, 2014)
  - HOW TO GET AROUND ELIGIBILITY CRITERIA
  - BREAKING THE BLIND
  - SHARING AE INFORMATION
  - ENCOURAGING/DISCOURAGING AE REPORTING
  - PREMATURE EFFICACY ASSUMPTIONS
  - MISINFORMATION

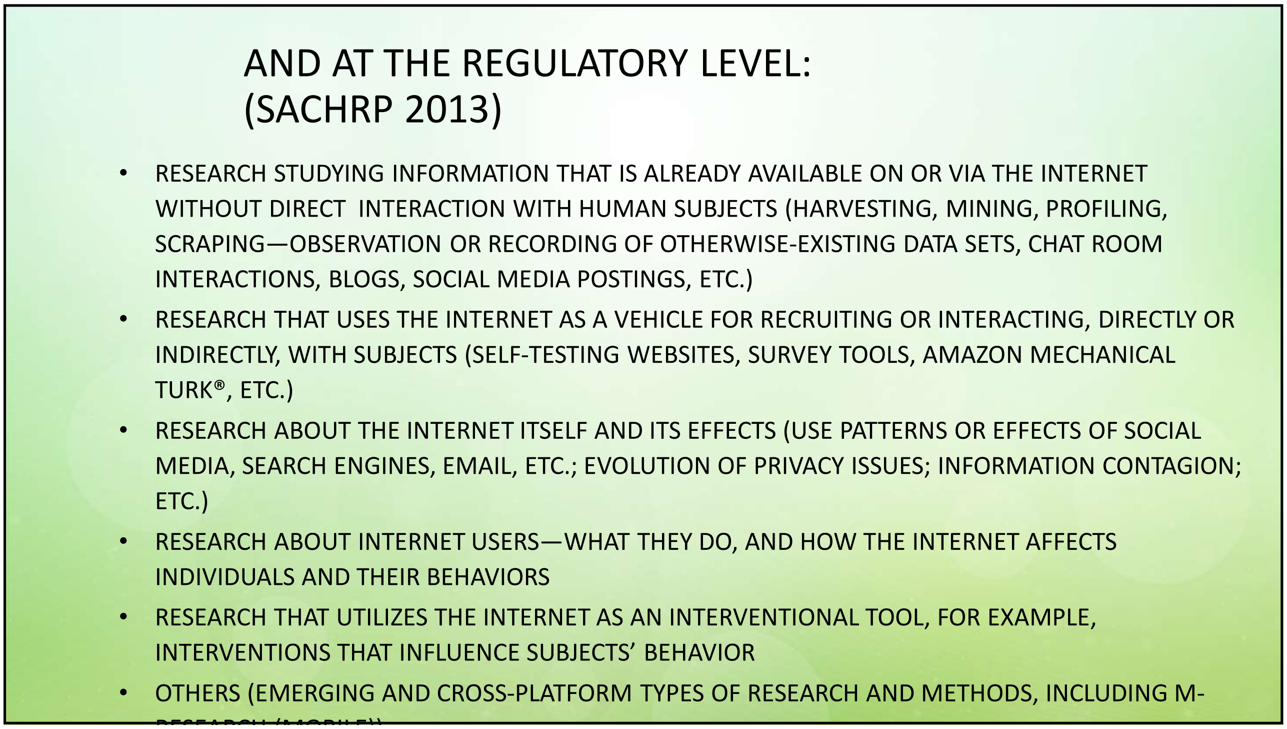
INSIDE A SUPPORT GROUP

- THE (STUDY DRUG) IS CAUSING THE [SIDE EFFECT] ….. AND IT MUST BE REMOVED ….(1) STOP THE (STUDY DRUG) AND GET IT OUT OF YOUR SYSTEM; (2) GO TO A DERMATOLOGIST …(3) GET THEM TO START YOU ON A PREDNISOLONE TAPER STARTING AT 40 MG FIRST DAY, THEN 30MG FOR 4 DAYS, THEN 20MG FOR 4 DAYS AND THEN 15MG FOR ONE DAY….. [IF THAT IS NOT SUCCESSFUL] TRY 125MG IV OF SOLU-MEDROL. ..DON'T LET SOME JIVE-TALKING DOCTOR TRY AND TELL YOU IT'S THE SAME THING… I'M THE RESIDENT EXPERT ON THE SUBJECT AT THIS POINT. SO YEAH, I'M GONNA … ASSERT YOU SHOULD STOP THE (STUDY DRUG)! AND NO I'M NOT A DOCTOR…"



AND IN THE NEWS

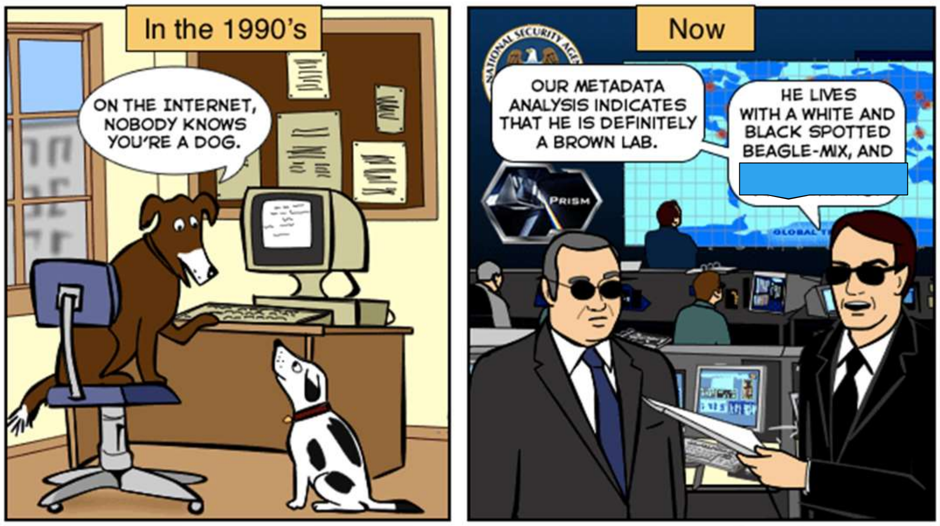
- MERCK & CO. CAUTIONED THAT SOCIAL MEDIA REPORTS COULD RESULT IN AN AE BEING BLOWN OUT OF PROPORTION, CITING THE EXAMPLE OF SANOFI, WHICH HAD TO SHUT DOWN ITS FACEBOOK PAGE WHEN A PATIENT WHO REACTED TO CANCER DRUG TAXOTERE POSTED A FLOOD OF COMMENTS ABOUT EXPERIENCING HAIR LOSS, TRIGGERING MAJOR REACTIONS FROM THE LARGER PATIENT COMMUNITY. SANOFI DID REOPEN THE PAGE LATER, BUT INCLUDED TERMS OF USE. LILLY RECOMMENDED THAT THE FDA CREATE A SEPARATE CATEGORY FOR EVENTS REPORTED ON SOCIAL MEDIA AND ALSO CONDUCTED A PILOT STUDY TO DEMONSTRATE THAT THE CONSIDERABLE EFFORTS INVESTED IN MONITORING SOCIAL MEDIA FOR AES DID NOT YIELD CORRESPONDING RESULTS.

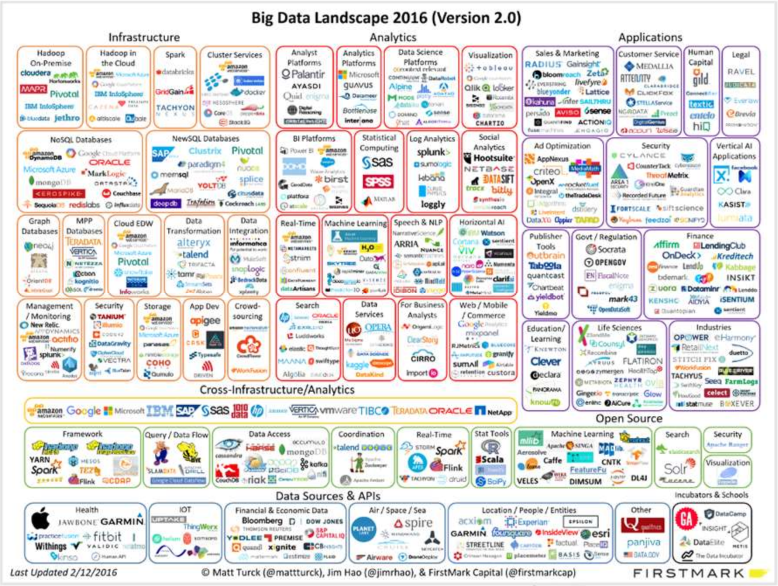


AND AT THE REGULATORY LEVEL:
(SACHRP 2013)

- RESEARCH STUDYING INFORMATION THAT IS ALREADY AVAILABLE ON OR VIA THE INTERNET WITHOUT DIRECT INTERACTION WITH HUMAN SUBJECTS (HARVESTING, MINING, PROFILING, SCRAPING—OBSERVATION OR RECORDING OF OTHERWISE-EXISTING DATA SETS, CHAT ROOM INTERACTIONS, BLOGS, SOCIAL MEDIA POSTINGS, ETC.)
- RESEARCH THAT USES THE INTERNET AS A VEHICLE FOR RECRUITING OR INTERACTING, DIRECTLY OR INDIRECTLY, WITH SUBJECTS (SELF-TESTING WEBSITES, SURVEY TOOLS, AMAZON MECHANICAL TURK®, ETC.)
- RESEARCH ABOUT THE INTERNET ITSELF AND ITS EFFECTS (USE PATTERNS OR EFFECTS OF SOCIAL MEDIA, SEARCH ENGINES, EMAIL, ETC.; EVOLUTION OF PRIVACY ISSUES; INFORMATION CONTAGION; ETC.)
- RESEARCH ABOUT INTERNET USERS—WHAT THEY DO, AND HOW THE INTERNET AFFECTS INDIVIDUALS AND THEIR BEHAVIORS
- RESEARCH THAT UTILIZES THE INTERNET AS AN INTERVENTIONAL TOOL, FOR EXAMPLE, INTERVENTIONS THAT INFLUENCE SUBJECTS' BEHAVIOR
- OTHERS (EMERGING AND CROSS-PLATFORM TYPES OF RESEARCH AND METHODS, INCLUDING M-RESEARCH (MOBILE))

AND THEN:

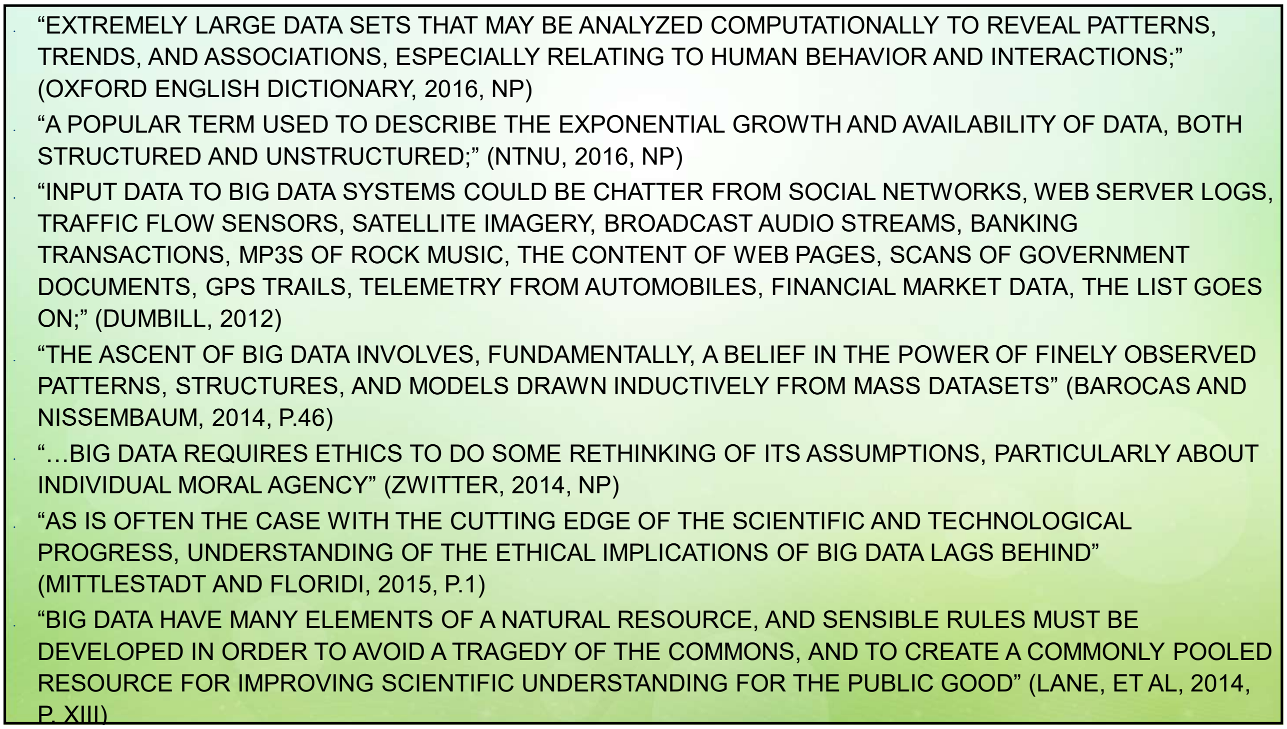

In the 1990's
ON THE INTERNET, NOBODY KNOWS YOU'RE A DOG.
Now
OUR METADATA ANALYSIS INDICATES THAT HE IS DEFINITELY A BROWN LAB.
HE LIVES WITH A WHITE AND BLACK SPOTTED BEAGLE-MIX, AND [redacted]
PRISM



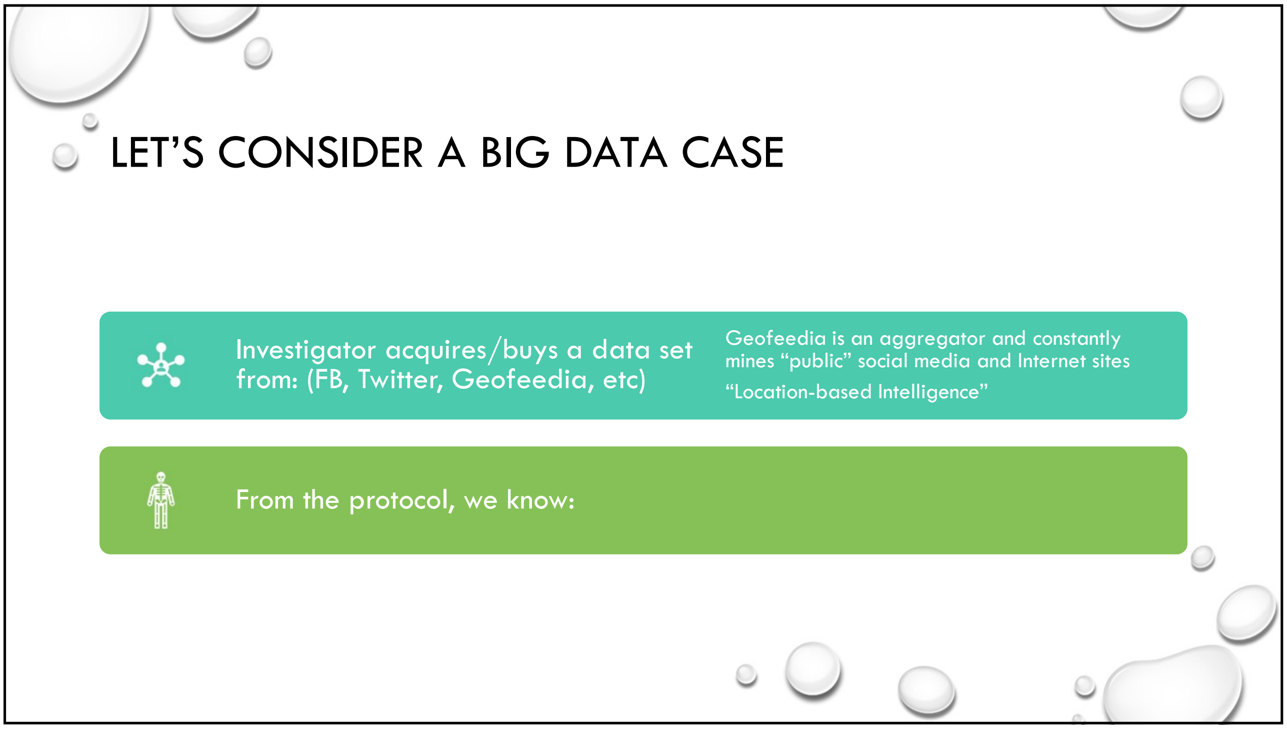
Last Updated 2/12/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK



- “EXTREMELY LARGE DATA SETS THAT MAY BE ANALYZED COMPUTATIONALLY TO REVEAL PATTERNS, TRENDS, AND ASSOCIATIONS, ESPECIALLY RELATING TO HUMAN BEHAVIOR AND INTERACTIONS;” (OXFORD ENGLISH DICTIONARY, 2016, NP)
- “A POPULAR TERM USED TO DESCRIBE THE EXPONENTIAL GROWTH AND AVAILABILITY OF DATA, BOTH STRUCTURED AND UNSTRUCTURED;” (NTNU, 2016, NP)
- “INPUT DATA TO BIG DATA SYSTEMS COULD BE CHATTER FROM SOCIAL NETWORKS, WEB SERVER LOGS, TRAFFIC FLOW SENSORS, SATELLITE IMAGERY, BROADCAST AUDIO STREAMS, BANKING TRANSACTIONS, MP3S OF ROCK MUSIC, THE CONTENT OF WEB PAGES, SCANS OF GOVERNMENT DOCUMENTS, GPS TRAILS, TELEMETRY FROM AUTOMOBILES, FINANCIAL MARKET DATA, THE LIST GOES ON;” (DUMBILL, 2012)
- “THE ASCENT OF BIG DATA INVOLVES, FUNDAMENTALLY, A BELIEF IN THE POWER OF FINELY OBSERVED PATTERNS, STRUCTURES, AND MODELS DRAWN INDUCTIVELY FROM MASS DATASETS” (BAROCAS AND NISSEMBAUM, 2014, P.46)
- “…BIG DATA REQUIRES ETHICS TO DO SOME RETHINKING OF ITS ASSUMPTIONS, PARTICULARLY ABOUT INDIVIDUAL MORAL AGENCY” (ZWITTER, 2014, NP)
- “AS IS OFTEN THE CASE WITH THE CUTTING EDGE OF THE SCIENTIFIC AND TECHNOLOGICAL PROGRESS, UNDERSTANDING OF THE ETHICAL IMPLICATIONS OF BIG DATA LAGS BEHIND” (MITTLESTADT AND FLORIDI, 2015, P.1)
- “BIG DATA HAVE MANY ELEMENTS OF A NATURAL RESOURCE, AND SENSIBLE RULES MUST BE DEVELOPED IN ORDER TO AVOID A TRAGEDY OF THE COMMONS, AND TO CREATE A COMMONLY POOLED RESOURCE FOR IMPROVING SCIENTIFIC UNDERSTANDING FOR THE PUBLIC GOOD” (LANE, ET AL, 2014, P. XIII)



LET’S CONSIDER A BIG DATA CASE

Investigator acquires/buys a data set from: (FB, Twitter, Geofeedia, etc)

Geofeedia is an aggregator and constantly mines “public” social media and Internet sites

“Location-based Intelligence”

From the protocol, we know:

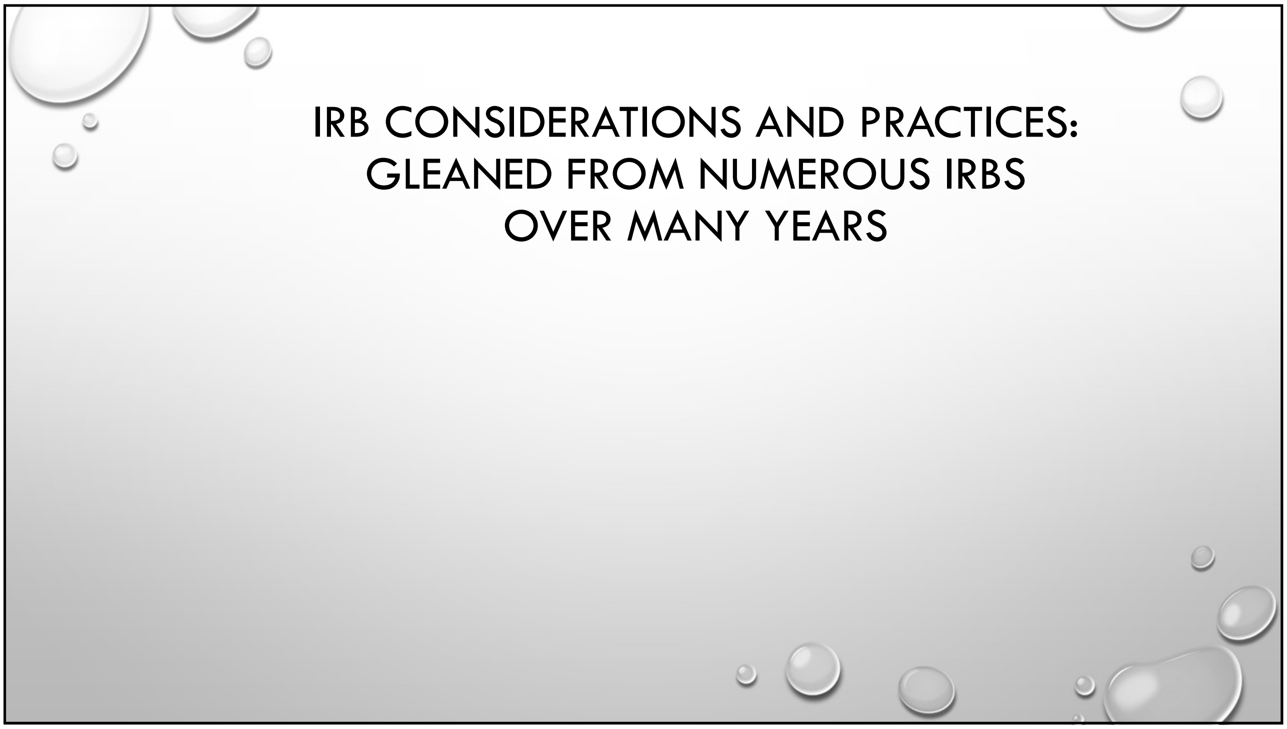
DATA USED FOR THIS ANALYSIS WILL COME FROM MULTIPLE SOCIAL MEDIA SOURCES AND DELIVERED THROUGH THE SOCIAL MEDIA PLATFORM GEOFEEDIA. SOCIAL MEDIA DATA FROM TWITTER, INSTAGRAM, FACEBOOK, YOUTUBE, FLICKR, PICASA, SINA WEIBO, AND VK WILL BE MINED FOR OPIOID- AND HEROIN-RELATED THEMES. GEOFEEDIA PROVIDES A COMPREHENSIVE COLLECTION OF SOCIAL MEDIA DATA THAT ALLOWS A USER TO SEARCH BY LOCATION FOR A SET OF DEFINED KEYWORDS, HASHTAGS, AND EMOJI'S, ETC. IT ALSO PROVIDES THE ABILITY TO LOOK BACKWARDS THROUGH TIME, AS WELL AS IDENTIFY SOCIAL NETWORKS BY LOOKING AT HOW USERS INTERACT AND INFLUENCE OTHERS IN THE SOCIAL NETWORK. ALL POSTS ARE LOCATION-EXPLICIT AND ALLOW FOR EASY MAPPING. ALTHOUGH ALL OF THE DATA USED ARE PUBLICLY AVAILABLE, WE WILL HAVE THE DATA DEIDENTIFIED (USERNAMES ANONYMIZED AND REPLACED WITH DIGITAL ALPHA/NUMERIC CODE) AND THE LOCATIONS WILL BE BE SLIGHTLY OFFSET BY ROUNDING THE LATITUDE/LONGITUDE COORDINATES BY TWO SIGNIFICANT DIGITS.

Once the recordings have been initiated, we will geocode the crime data for the city of REDACTED. This will convert address-level data to latitude/longitude coordinates to be used for the spatial modeling. Geocoding the addresses and spatial modeling will be performed using ArcGIS 10.3. Local spatial clustering statistics such as Local Moran's I and Getis-Ord Gi*, will be used to detect clustering within the defined study sites. Once this is complete, social media-identified clusters will be mapped onto clusters of opioid and heroin-related arrests. These crime-related clusters will serve as a baseline and will be used for comparison with those clusters identified through social media.

The goal is to examine how well social media-derived clusters may serve as a proxy for real time surveillance of the epidemic. Here, we want to determine how well the social media-identified clusters overlap with those found in the crime data. One limitation may be that additional clusters may be discovered from those derived from social media and that these may not necessarily represent new, previously unknown clusters….

WHAT ARE THE MAJOR ISSUES?

- IS THIS HSR?
  - IF WE USE OUR CURRENT REGS?
- ARE THEY PUBLIC DATA?
- ARE SUBJECTS IDENTIFIABLE?
- WHAT ELSE?



IRB CONSIDERATIONS AND PRACTICES: GLEANED FROM NUMEROUS IRBS OVER MANY YEARS

- ASK INVESTIGATORS TO INCLUDE SCREEN SHOTS WITH PROTOCOLS—CAN HELP IN UNDERSTANDING DATA FLOWS, RECRUITMENT PROCESSES, ETC
- CONSIDER: DATA IN USE, AT REST, IN TRANSIT, AND IN DELETION: DIFFERENT ETHICAL CONSIDERATIONS AND SECURITY MEASURES; DESCRIBE PROCEDURES (INCLUDING SAFEGUARDS FOR COLLECTING, STORING, PROCESSING SUBJECT DATA AND DATA DESTRUCTION) FOR MINIMIZING POTENTIAL RISKS TO SUBJECT'S CONFIDENTIALITY
- LEARN THE NUANCES BETWEEN AND AMONG DATA MANAGEMENT PRACTICES, INCLUDING DE-AND RE-IDENTIFICATION; ANONYMIZED, CODED, AGGREGATED
  - DATA SHARING AND DATA USE AGREEMENTS (NIH, NSF MANDATES); IMPORTANT FOR RESEARCHERS TO WORK WITH REBS/IRBS IN PLANNING FOR DATA SHARING—RAW DATA? THEMES?
- SPECIFY WHERE AND UNDER WHAT CONDITIONS INDIVIDUALS WILL HAVE ACCESS TO THE DATA, WHAT WILL BE AVAILABLE AND TO WHOM (AIR GAP, CLEAN ROOMS, DATA ACCESS LEVELS)

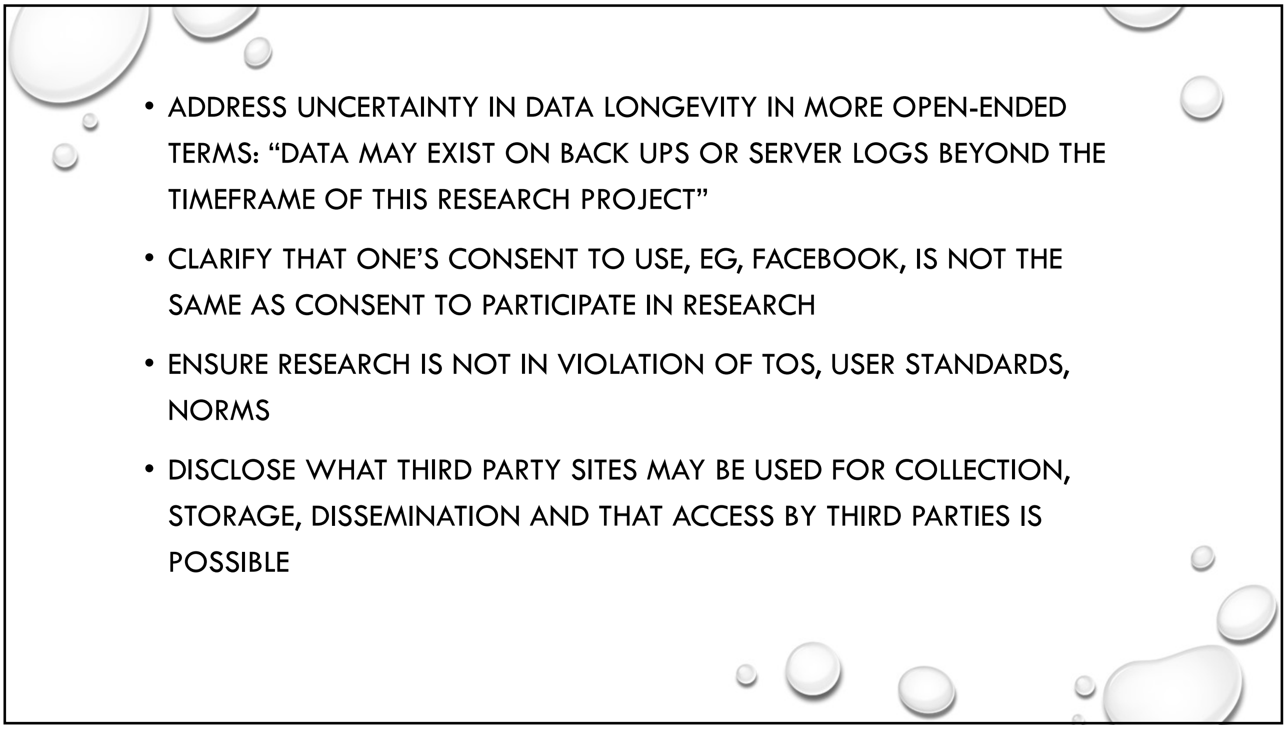
IF AGGREGATED ANONYMIZED DATA WILL BE MADE PUBLICLY AVAILABLE, CONSIDER WHETHER SUBJECTS COULD BE (RE)IDENTIFIED, AND WHAT LEVEL OF RISK APPLIES

- RECONSIDER MINIMAL RISK AND EVERYDAY LIFE:

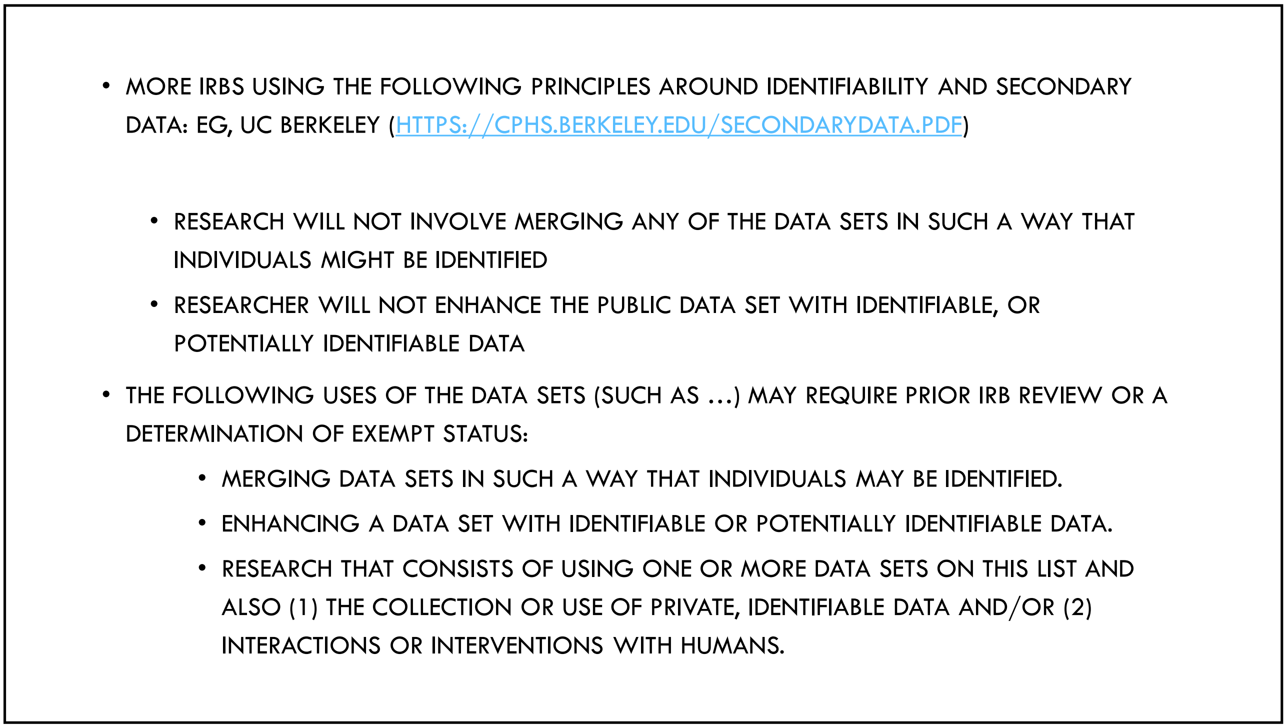
“WE ANTICIPATE THAT YOUR PARTICIPATION IN THIS STUDY PRESENTS NO GREATER RISK THAN EVERYDAY USE OF THE INTERNET”

“ALTHOUGH EVERY REASONABLE EFFORT HAS BEEN TAKEN, CONFIDENTIALITY DURING ACTUAL INTERNET COMMUNICATION *PROCEDURES CANNOT BE GUARANTEED.”*

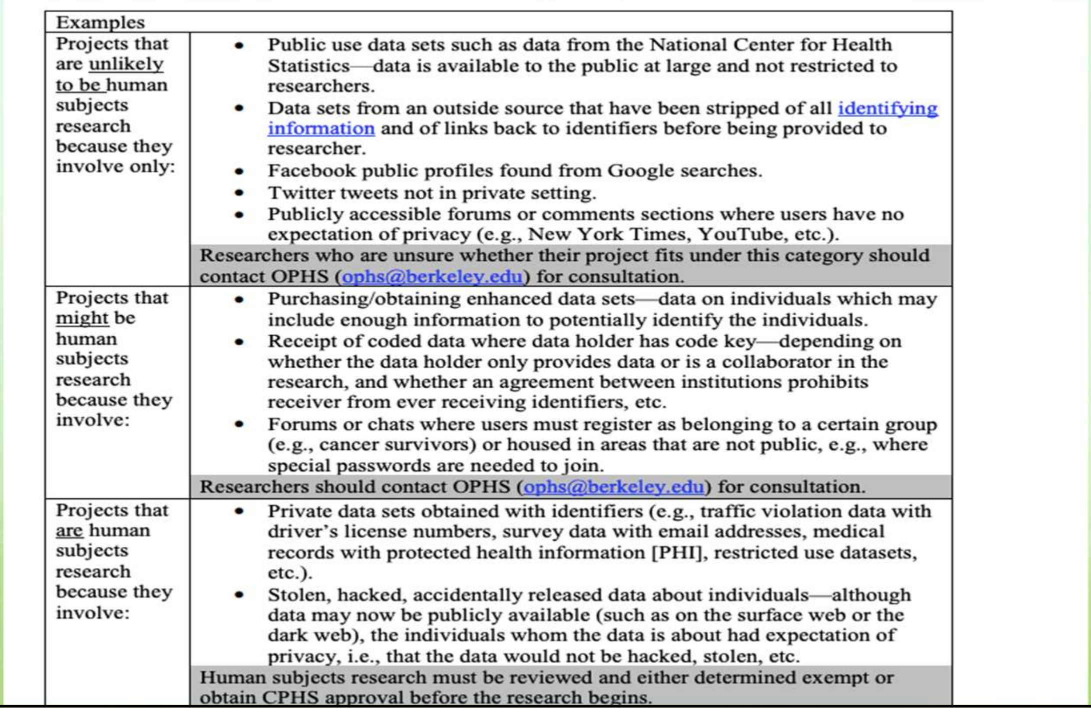
“YOUR CONFIDENTIALITY WILL BE KEPT TO THE DEGREE PERMITTED BY THE TECHNOLOGY BEING USED. NO GUARANTEES CAN BE MADE REGARDING THE INTERCEPTION OF DATA SENT VIA THE INTERNET BY ANY THIRD PARTIES.”



- ADDRESS UNCERTAINTY IN DATA LONGEVITY IN MORE OPEN-ENDED TERMS: “DATA MAY EXIST ON BACK UPS OR SERVER LOGS BEYOND THE TIMEFRAME OF THIS RESEARCH PROJECT”
- CLARIFY THAT ONE’S CONSENT TO USE, EG, FACEBOOK, IS NOT THE SAME AS CONSENT TO PARTICIPATE IN RESEARCH
- ENSURE RESEARCH IS NOT IN VIOLATION OF TOS, USER STANDARDS, NORMS
- DISCLOSE WHAT THIRD PARTY SITES MAY BE USED FOR COLLECTION, STORAGE, DISSEMINATION AND THAT ACCESS BY THIRD PARTIES IS POSSIBLE



- MORE IRBS USING THE FOLLOWING PRINCIPLES AROUND IDENTIFIABILITY AND SECONDARY DATA: EG, UC BERKELEY (HTTPS://CPHS.BERKELEY.EDU/SECONDARYDATA.PDF)
  - RESEARCH WILL NOT INVOLVE MERGING ANY OF THE DATA SETS IN SUCH A WAY THAT INDIVIDUALS MIGHT BE IDENTIFIED
  - RESEARCHER WILL NOT ENHANCE THE PUBLIC DATA SET WITH IDENTIFIABLE, OR POTENTIALLY IDENTIFIABLE DATA
- THE FOLLOWING USES OF THE DATA SETS (SUCH AS …) MAY REQUIRE PRIOR IRB REVIEW OR A DETERMINATION OF EXEMPT STATUS:
  - MERGING DATA SETS IN SUCH A WAY THAT INDIVIDUALS MAY BE IDENTIFIED.
  - ENHANCING A DATA SET WITH IDENTIFIABLE OR POTENTIALLY IDENTIFIABLE DATA.
  - RESEARCH THAT CONSISTS OF USING ONE OR MORE DATA SETS ON THIS LIST AND ALSO (1) THE COLLECTION OR USE OF PRIVATE, IDENTIFIABLE DATA AND/OR (2) INTERACTIONS OR INTERVENTIONS WITH HUMANS.



Examples	
Projects that are unlikely to be human subjects research because they involve only:	• Public use data sets such as data from the National Center for Health Statistics—data is available to the public at large and not restricted to researchers. • Data sets from an outside source that have been stripped of all identifying information and of links back to identifiers before being provided to researcher. • Facebook public profiles found from Google searches. • Twitter tweets not in private setting. • Publicly accessible forums or comments sections where users have no expectation of privacy (e.g., New York Times, YouTube, etc.).
	Researchers who are unsure whether their project fits under this category should contact OPHS (ophs@berkeley.edu) for consultation.
Projects that might be human subjects research because they involve:	• Purchasing/obtaining enhanced data sets—data on individuals which may include enough information to potentially identify the individuals. • Receipt of coded data where data holder has code key—depending on whether the data holder only provides data or is a collaborator in the research, and whether an agreement between institutions prohibits receiver from ever receiving identifiers, etc. • Forums or chats where users must register as belonging to a certain group (e.g., cancer survivors) or housed in areas that are not public, e.g., where special passwords are needed to join.
	Researchers should contact OPHS (ophs@berkeley.edu) for consultation.
Projects that are human subjects research because they involve:	• Private data sets obtained with identifiers (e.g., traffic violation data with driver's license numbers, survey data with email addresses, medical records with protected health information [PHI], restricted use datasets, etc.). • Stolen, hacked, accidentally released data about individuals—although data may now be publicly available (such as on the surface web or the dark web), the individuals whom the data is about had expectation of privacy, i.e., that the data would not be hacked, stolen, etc.
	Human subjects research must be reviewed and either determined exempt or obtain CPHS approval before the research begins.

Q/A