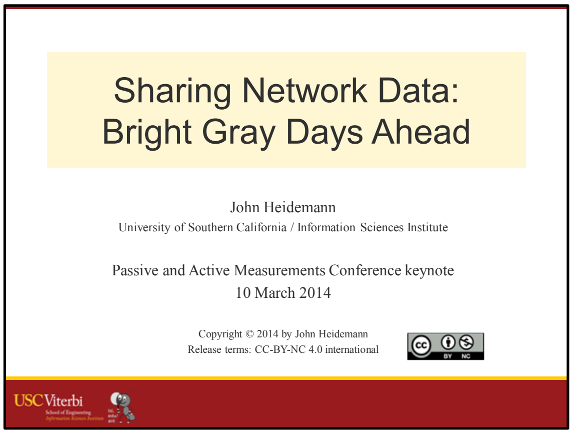
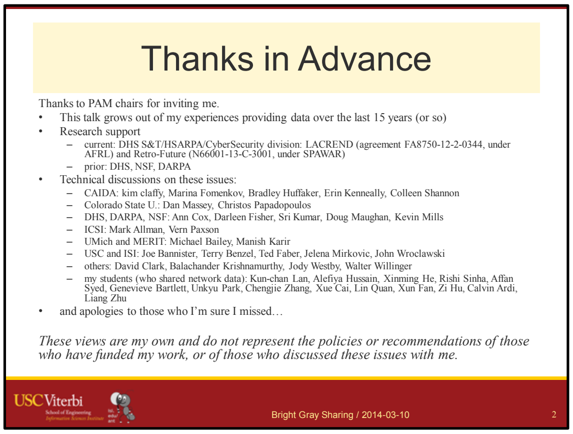


Sharing Network Data: Bright Gray Days Ahead

John Heidemann
University of Southern California / Information Sciences Institute

Passive and Active Measurements Conference keynote
10 March 2014

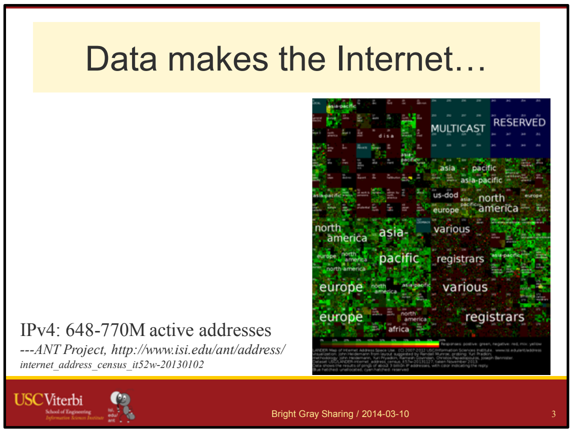


Thanks in Advance

Thanks to PAM chairs for inviting me.

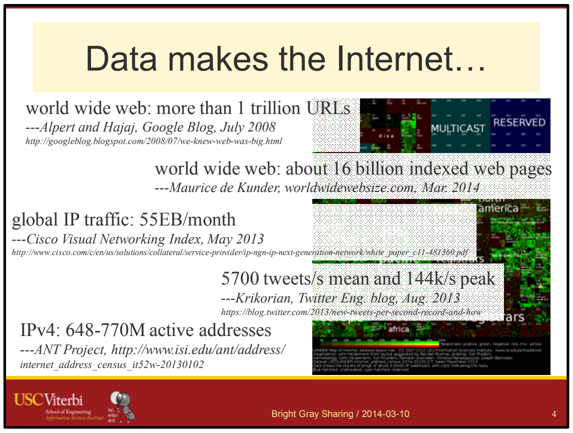
- This talk grows out of my experiences providing data over the last 15 years (or so)
- Research support
  - current: DHS S&T/HSARPA/CyberSecurity division: LACREND (agreement FA8750-12-2-0344, under AFRL) and Retro-Future (N66001-13-C-3001, under SPAWAR)
  - prior: DHS, NSF, DARPA
- Technical discussions on these issues:
  - CAIDA: kim claffy, Marina Fomenkov, Bradley Huffaker, Erin Kenneally, Colleen Shannon
  - Colorado State U.: Dan Massey, Christos Papadopoulos
  - DHS, DARPA, NSF: Ann Cox, Darleen Fisher, Sri Kumar, Doug Maughan, Kevin Mills
  - ICSI: Mark Allman, Vern Paxson
  - UMich and MERIT: Michael Bailey, Manish Karir
  - USC and ISI: Joe Bannister, Terry Benzel, Ted Faber, Jelena Mirkovic, John Wroclawski
  - others: David Clark, Balachander Krishnamurthy, Jody Westby, Walter Willinger
  - my students (who shared network data): Kun-chan Lan, Alefiya Hussain, Xinming He, Rishi Sinha, Affan Syed, Genevieve Bartlett, Unkyu Park, Chengjie Zhang, Xue Cai, Lin Quan, Xun Fan, Zi Hu, Calvin Ardi, Liang Zhu
- and apologies to those who I'm sure I missed…

These views are my own and do not represent the policies or recommendations of those who have funded my work, or of those who discussed these issues with me.



Data makes the Internet…

IPv4: 648-770M active addresses
---ANT Project, http://www.isi.edu/ant/address/ internet_address_census_it52w-20130102



Data makes the Internet…

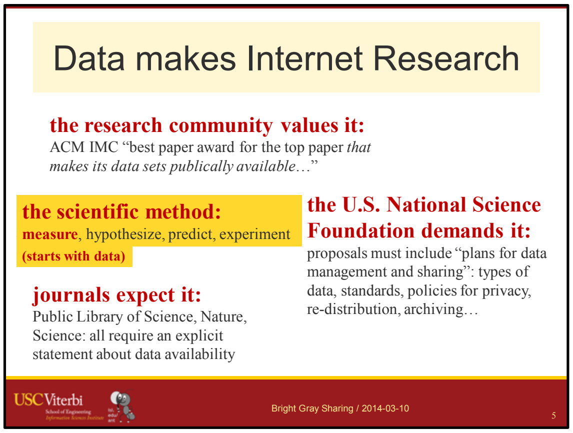
world wide web: more than 1 trillion URLs
---Alpert and Hajaj, Google Blog, July 2008
http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

world wide web: about 16 billion indexed web pages
---Maurice de Kunder, worldwidewebsize.com, Mar. 2014

global IP traffic: 55EB/month
---Cisco Visual Networking Index, May 2013
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

5700 tweets/s mean and 144k/s peak
---Krikorian, Twitter Eng. blog, Aug. 2013
https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

IPv4: 648-770M active addresses
---ANT Project, http://www.isi.edu/ant/address/ internet_address_census_it52w-20130102



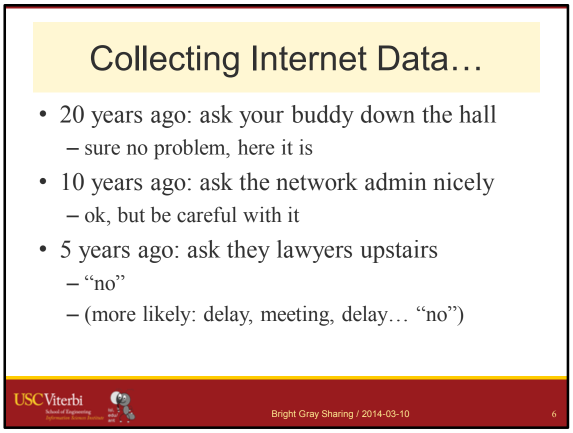
Data makes Internet Research

the research community values it:
ACM IMC "best paper award for the top paper *that makes its data sets publically available…*"

the scientific method:
measure, hypothesize, predict, experiment
(starts with data)

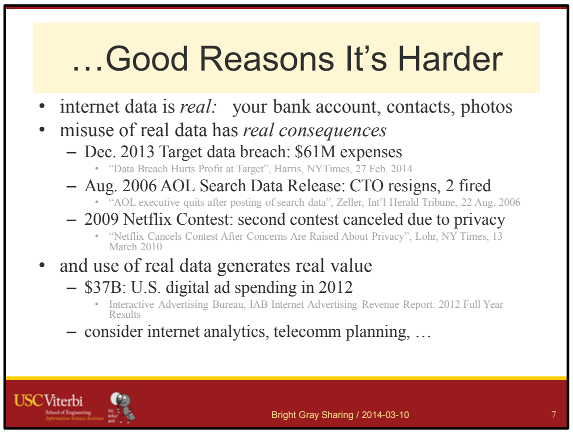
journals expect it:
Public Library of Science, Nature, Science: all require an explicit statement about data availability

the U.S. National Science Foundation demands it:
proposals must include "plans for data management and sharing": types of data, standards, policies for privacy, re-distribution, archiving…



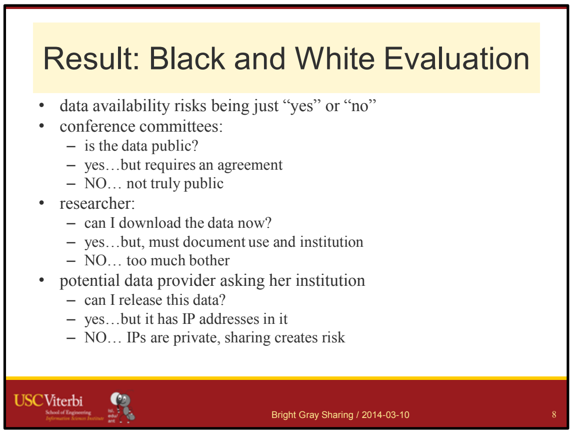
Collecting Internet Data…

- 20 years ago: ask your buddy down the hall
  - sure no problem, here it is
- 10 years ago: ask the network admin nicely
  - ok, but be careful with it
- 5 years ago: ask they lawyers upstairs
  - "no"
  - (more likely: delay, meeting, delay… "no")



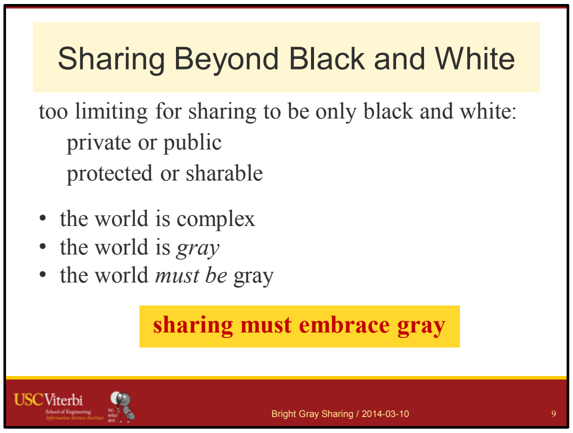
…Good Reasons It's Harder

- internet data is *real:* your bank account, contacts, photos
- misuse of real data has *real consequences*
  - Dec. 2013 Target data breach: $61M expenses
    - "Data Breach Hurts Profit at Target", Harris, NYTimes, 27 Feb. 2014
  - Aug. 2006 AOL Search Data Release: CTO resigns, 2 fired
    - "AOL executive quits after posting of search data", Zeller, Int'l Herald Tribune, 22 Aug. 2006
  - 2009 Netflix Contest: second contest canceled due to privacy
    - "Netflix Cancels Contest After Concerns Are Raised About Privacy", Lohr, NY Times, 13 March 2010
- and use of real data generates real value
  - $37B: U.S. digital ad spending in 2012
    - Interactive Advertising Bureau, IAB Internet Advertising Revenue Report: 2012 Full Year Results
  - consider internet analytics, telecomm planning, …



Result: Black and White Evaluation

- data availability risks being just "yes" or "no"
- conference committees:
  - is the data public?
  - yes…but requires an agreement
  - NO… not truly public
- researcher:
  - can I download the data now?
  - yes…but, must document use and institution
  - NO… too much bother
- potential data provider asking her institution
  - can I release this data?
  - yes…but it has IP addresses in it
  - NO… IPs are private, sharing creates risk



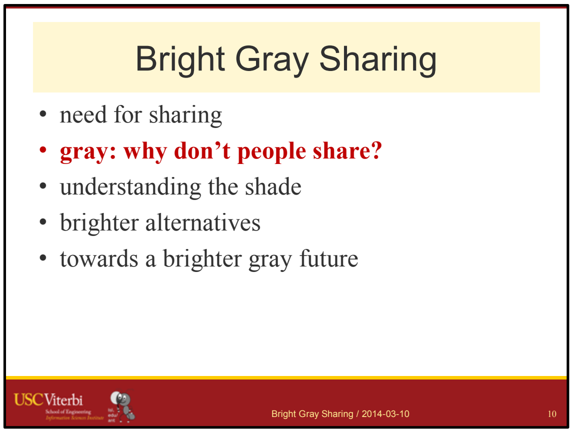
Sharing Beyond Black and White

too limiting for sharing to be only black and white:

private or public
protected or sharable

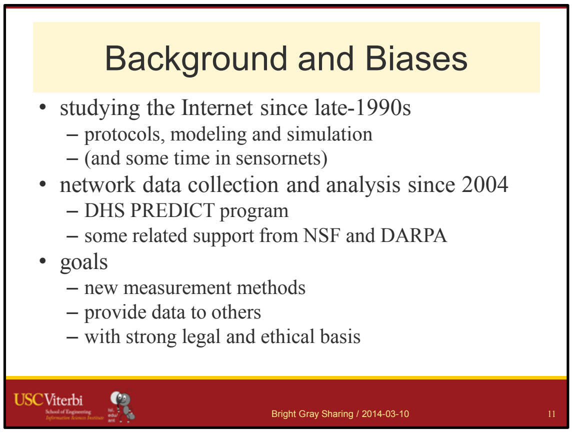
- the world is complex
- the world is *gray*
- the world *must be* gray

sharing must embrace gray



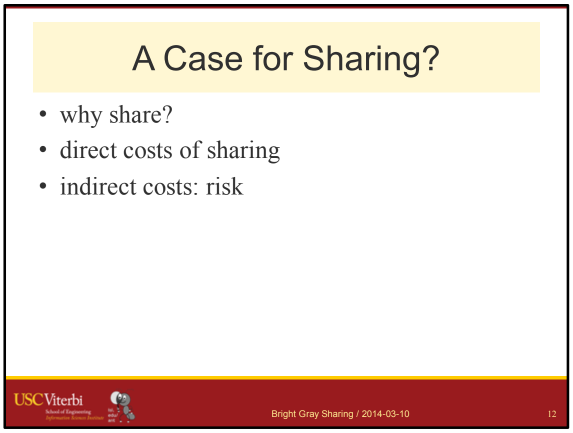
Bright Gray Sharing

- need for sharing
- **gray: why don't people share?**
- understanding the shade
- brighter alternatives
- towards a brighter gray future



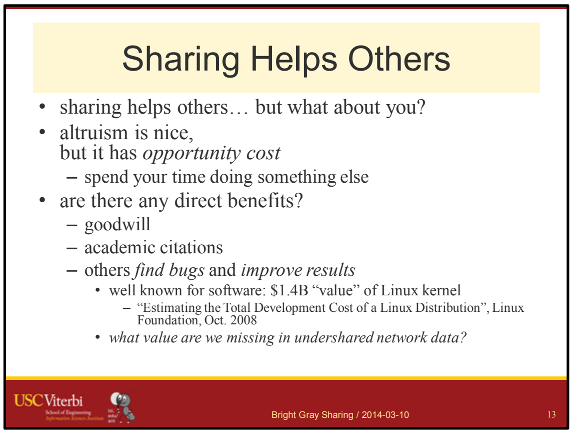
Background and Biases

- studying the Internet since late-1990s
  - protocols, modeling and simulation
  - (and some time in sensornets)
- network data collection and analysis since 2004
  - DHS PREDICT program
  - some related support from NSF and DARPA
- goals
  - new measurement methods
  - provide data to others
  - with strong legal and ethical basis



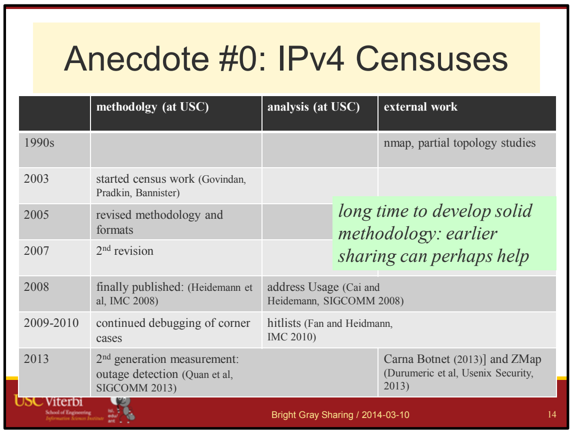
A Case for Sharing?

- why share?
- direct costs of sharing
- indirect costs: risk



Sharing Helps Others

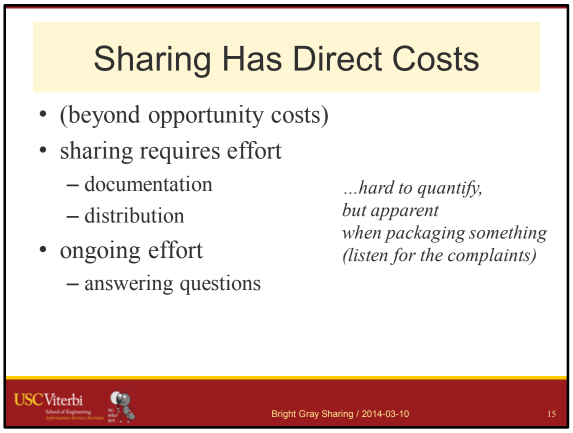
- sharing helps others… but what about you?
- altruism is nice,
  but it has *opportunity cost*
  - spend your time doing something else
- are there any direct benefits?
  - goodwill
  - academic citations
  - others *find bugs* and *improve results*
    - well known for software: $1.4B “value” of Linux kernel
      - “Estimating the Total Development Cost of a Linux Distribution”, Linux Foundation, Oct. 2008
    - *what value are we missing in undershared network data?*



Anecdote #0: IPv4 Censuses

	methodolgy (at USC)	**analysis (at USC)**	**external work**
1990s			nmap, partial topology studies
2003	started census work (Govindan, Pradkin, Bannister)		
2005	revised methodology and formats		
2007	2nd revision		
2008	finally published: (Heidemann et al, IMC 2008)	address Usage (Cai and Heidemann, SIGCOMM 2008)	
2009-2010	continued debugging of corner cases	hitlists (Fan and Heidmann, IMC 2010)	
2013	2nd generation measurement: outage detection (Quan et al, SIGCOMM 2013)		Carna Botnet (2013)] and ZMap (Durumeric et al, Usenix Security, 2013)

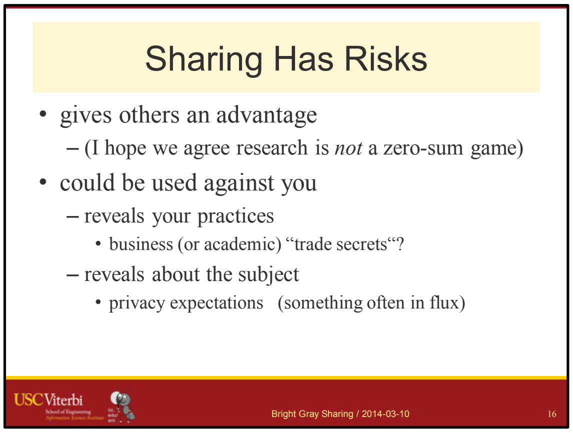
long time to develop solid methodology: earlier sharing can perhaps help



Sharing Has Direct Costs

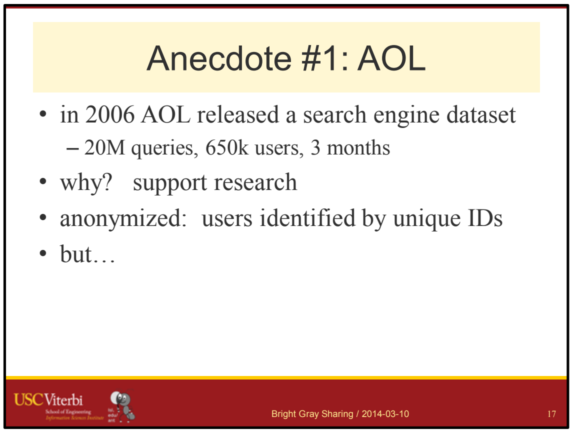
- (beyond opportunity costs)
- sharing requires effort
  - documentation
  - distribution
- ongoing effort
  - answering questions

…hard to quantify,
but apparent
when packaging something
(listen for the complaints)



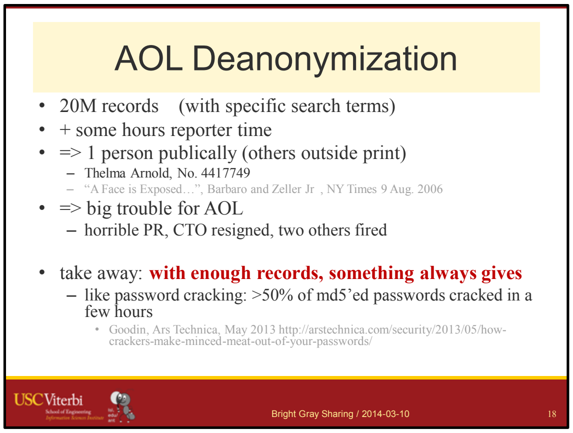
Sharing Has Risks

- gives others an advantage
  - (I hope we agree research is *not* a zero-sum game)
- could be used against you
  - reveals your practices
    - business (or academic) “trade secrets“?
  - reveals about the subject
    - privacy expectations   (something often in flux)



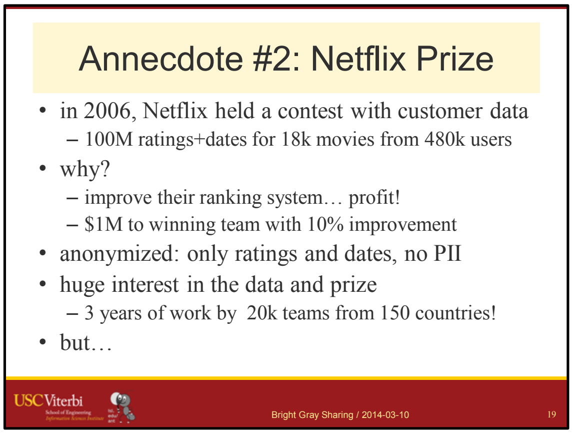
Anecdote #1: AOL

- in 2006 AOL released a search engine dataset
  - 20M queries, 650k users, 3 months
- why?   support research
- anonymized: users identified by unique IDs
- but…



AOL Deanonymization

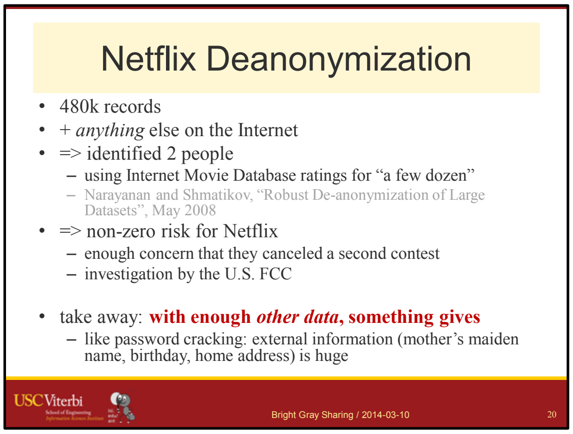
- 20M records   (with specific search terms)
- + some hours reporter time
- => 1 person publically (others outside print)
  - Thelma Arnold, No. 4417749
  - “A Face is Exposed…”, Barbaro and Zeller Jr , NY Times 9 Aug. 2006
- => big trouble for AOL
  - horrible PR, CTO resigned, two others fired
- take away: **with enough records, something always gives**
  - like password cracking: >50% of md5’ed passwords cracked in a few hours
    - Goodin, Ars Technica, May 2013 http://arstechnica.com/security/2013/05/how-crackers-make-minced-meat-out-of-your-passwords/



Annecdote #2: Netflix Prize

- in 2006, Netflix held a contest with customer data
  - 100M ratings+dates for 18k movies from 480k users
- why?
  - improve their ranking system… profit!
  - $1M to winning team with 10% improvement
- anonymized: only ratings and dates, no PII
- huge interest in the data and prize
  - 3 years of work by 20k teams from 150 countries!
- but…

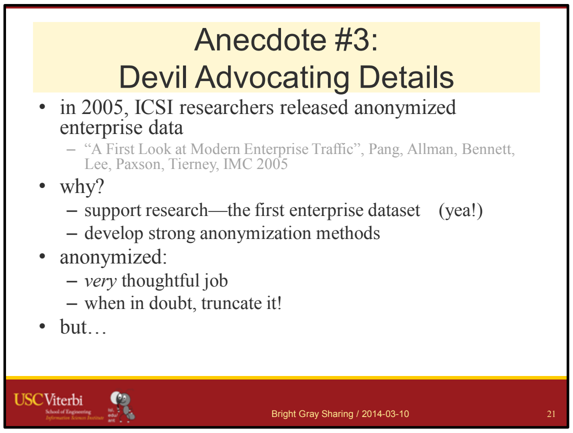
USC Viterbi School of Engineering



Netflix Deanonymization

- 480k records
- + *anything* else on the Internet
- => identified 2 people
  - using Internet Movie Database ratings for "a few dozen"
  - Narayanan and Shmatikov, "Robust De-anonymization of Large Datasets", May 2008
- => non-zero risk for Netflix
  - enough concern that they canceled a second contest
  - investigation by the U.S. FCC
- take away: **with enough *other data*, something gives**
  - like password cracking: external information (mother's maiden name, birthday, home address) is huge

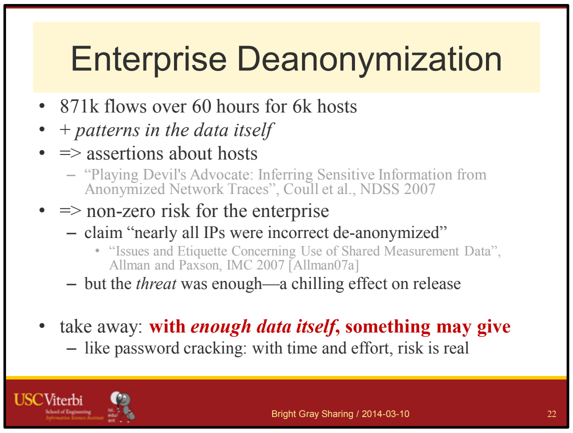
USC Viterbi School of Engineering



Anecdote #3: Devil Advocating Details

- in 2005, ICSI researchers released anonymized enterprise data
  - "A First Look at Modern Enterprise Traffic", Pang, Allman, Bennett, Lee, Paxson, Tierney, IMC 2005
- why?
  - support research—the first enterprise dataset (yea!)
  - develop strong anonymization methods
- anonymized:
  - *very* thoughtful job
  - when in doubt, truncate it!
- but…

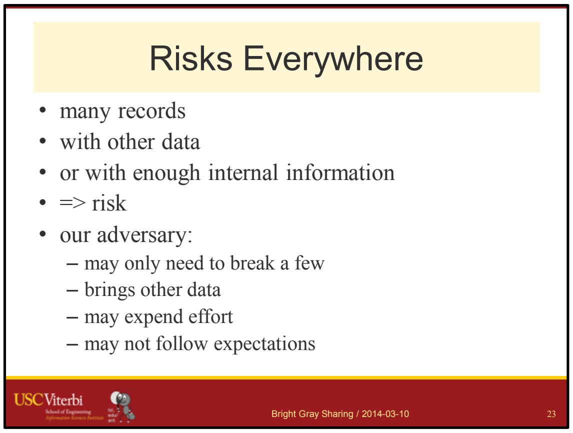
USC Viterbi School of Engineering



Enterprise Deanonymization

- 871k flows over 60 hours for 6k hosts
- + *patterns in the data itself*
- => assertions about hosts
  - "Playing Devil's Advocate: Inferring Sensitive Information from Anonymized Network Traces", Coull et al., NDSS 2007
- => non-zero risk for the enterprise
  - claim "nearly all IPs were incorrect de-anonymized"
    - "Issues and Etiquette Concerning Use of Shared Measurement Data", Allman and Paxson, IMC 2007 [Allman07a]
  - but the *threat* was enough—a chilling effect on release
- take away: **with *enough data itself*, something may give**
  - like password cracking: with time and effort, risk is real

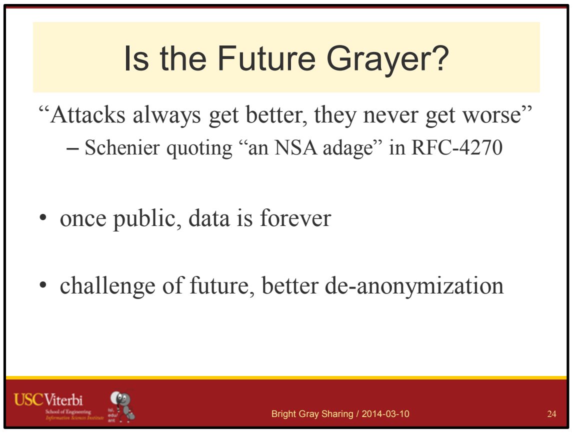
USC Viterbi School of Engineering



Risks Everywhere

- many records
- with other data
- or with enough internal information
- => risk
- our adversary:
  - may only need to break a few
  - brings other data
  - may expend effort
  - may not follow expectations

USC Viterbi School of Engineering



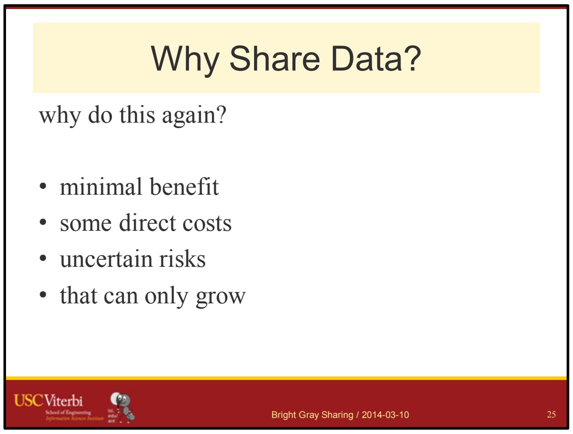
Is the Future Grayer?

"Attacks always get better, they never get worse"

- Schenier quoting "an NSA adage" in RFC-4270

- once public, data is forever
- challenge of future, better de-anonymization

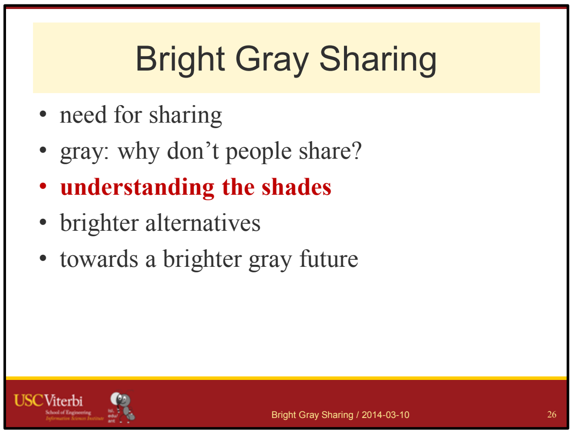
USC Viterbi School of Engineering



Why Share Data?

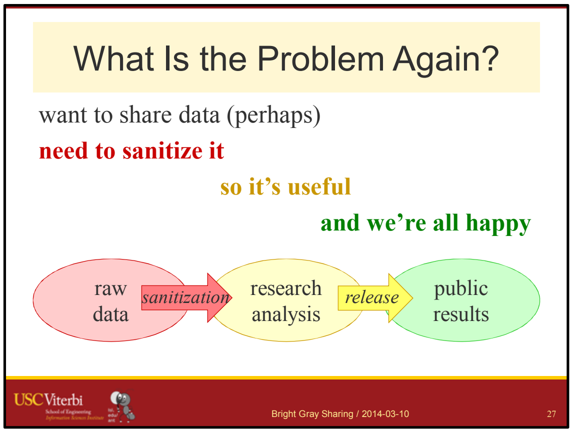
why do this again?

- minimal benefit
- some direct costs
- uncertain risks
- that can only grow



Bright Gray Sharing

- need for sharing
- gray: why don’t people share?
- **understanding the shades**
- brighter alternatives
- towards a brighter gray future



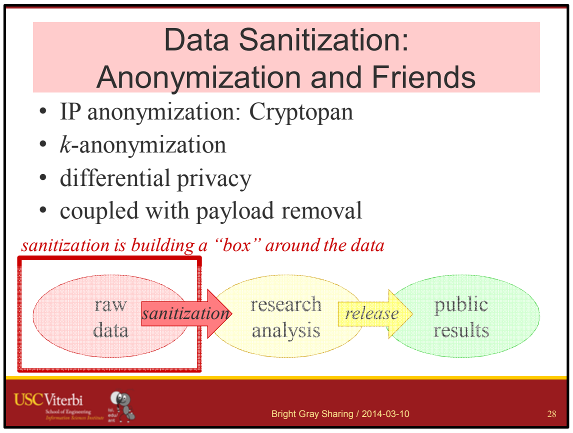
What Is the Problem Again?

want to share data (perhaps)

need to sanitize it

so it’s useful

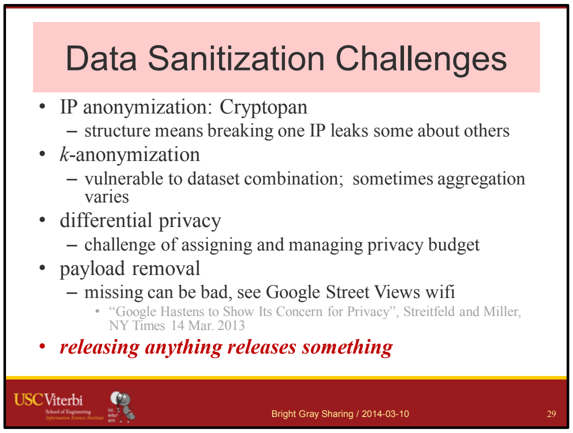
and we’re all happy



Data Sanitization: Anonymization and Friends

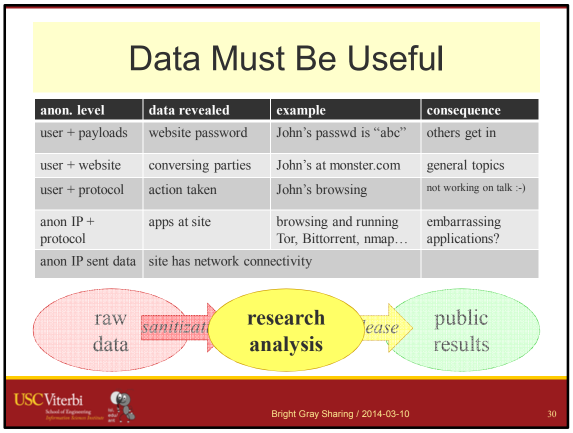
- IP anonymization: Cryptopan
- *k*-anonymization
- differential privacy
- coupled with payload removal

sanitization is building a “box” around the data



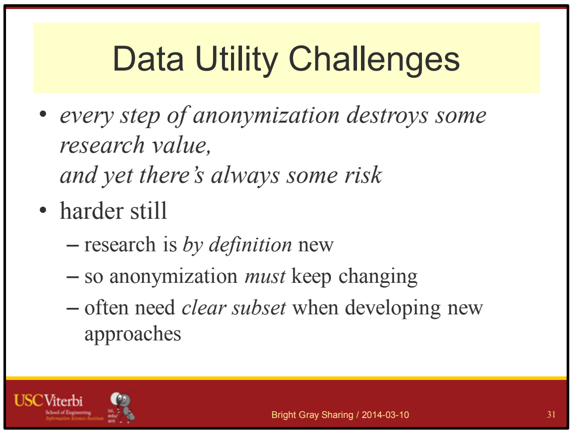
Data Sanitization Challenges

- IP anonymization: Cryptopan
  - structure means breaking one IP leaks some about others
- *k*-anonymization
  - vulnerable to dataset combination; sometimes aggregation varies
- differential privacy
  - challenge of assigning and managing privacy budget
- payload removal
  - missing can be bad, see Google Street Views wifi
    - “Google Hastens to Show Its Concern for Privacy”, Streitfeld and Miller, NY Times 14 Mar. 2013
- ***releasing anything releases something***



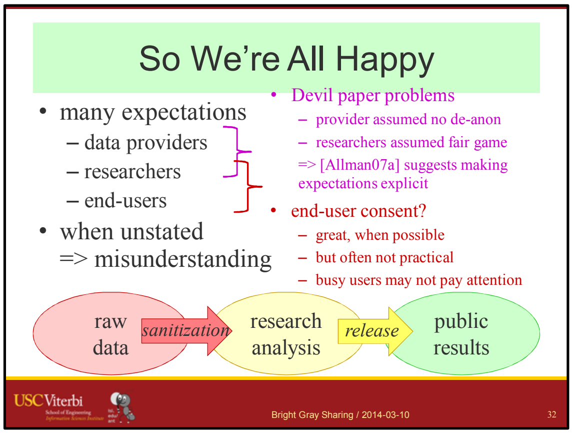
Data Must Be Useful

anon. level	data revealed	example	consequence
user + payloads	website password	John’s passwd is “abc”	others get in
user + website	conversing parties	John’s at monster.com	general topics
user + protocol	action taken	John’s browsing	not working on talk :-)
anon IP + protocol	apps at site	browsing and running Tor, Bittorrent, nmap…	embarrassing applications?
anon IP sent data	site has network connectivity		



Data Utility Challenges

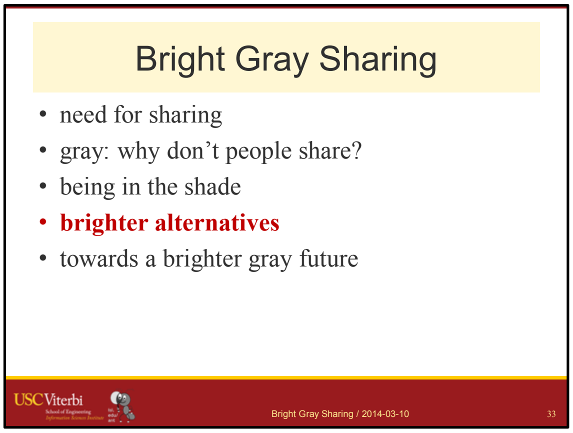
- *every step of anonymization destroys some research value,*
  and yet there's always some risk
- harder still
  - research is *by definition* new
  - so anonymization *must* keep changing
  - often need *clear subset* when developing new approaches



So We're All Happy

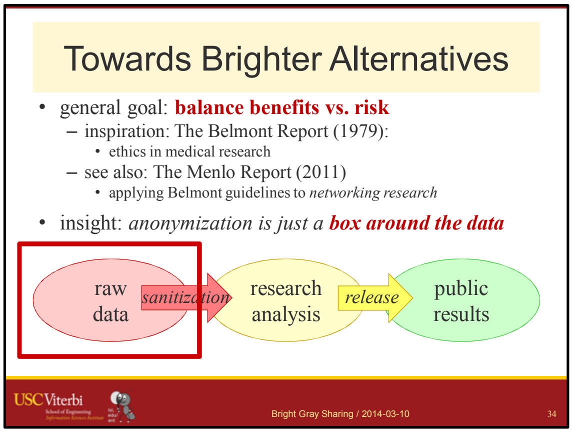
- many expectations
  - data providers
  - researchers
  - end-users
- when unstated
  => misunderstanding

- Devil paper problems
  - provider assumed no de-anon
  - researchers assumed fair game
  - => [Allman07a] suggests making expectations explicit
- end-user consent?
  - great, when possible
  - but often not practical
  - busy users may not pay attention



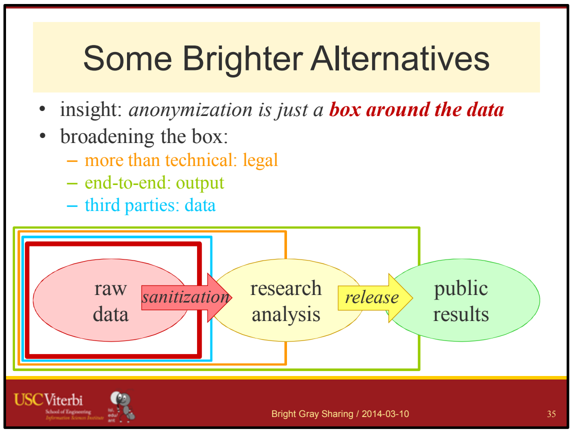
Bright Gray Sharing

- need for sharing
- gray: why don't people share?
- being in the shade
- **brighter alternatives**
- towards a brighter gray future



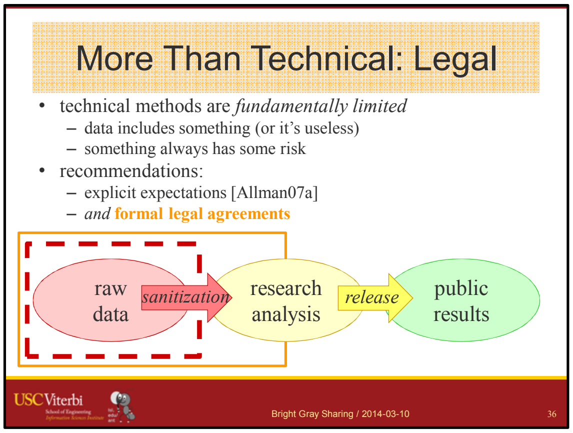
Towards Brighter Alternatives

- general goal: **balance benefits vs. risk**
  - inspiration: The Belmont Report (1979):
    - ethics in medical research
  - see also: The Menlo Report (2011)
    - applying Belmont guidelines to *networking research*
- insight: *anonymization is just a* ***box around the data***



Some Brighter Alternatives

- insight: *anonymization is just a* ***box around the data***
- broadening the box:
  - more than technical: legal
  - end-to-end: output
  - third parties: data



More Than Technical: Legal

- technical methods are *fundamentally limited*
  - data includes something (or it's useless)
  - something always has some risk
- recommendations:
  - explicit expectations [Allman07a]
  - *and* **formal legal agreements**

Why Technical *and* Legal

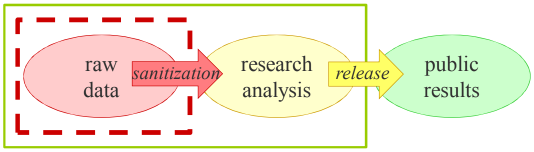
- technical is essential, but not sufficient (80%?)
- legal can bridge the gap
- the trade-offs
  - expectations and legal are not perfect
    - a bad actor can ignore them
    - accidental release can circumvent them
  - but expectations and legal have consequences
    - both social and legal
  - but there *are* consequences
  - social and possibly legal
- *but necessary to balance utility and protection*
  - when technical methods alone *do not usefully get to 100%*

Implications of Legal

- things will move slower
  - reviewing legal agreements and getting signatures
- don't apply everywhere
  - cross jurisdictions
  - not everyone is in "an institution"
- everyone is "special"
  - a lawyer's job is not done until some clause has changed
- *need to educate our institutions and our expectations*

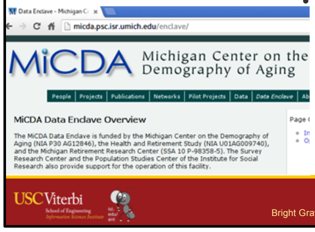
End-to-End Privacy: Output

- technical methods (on data) *ignore two-thirds of the space*
- recommendation:
  - consider end-to-end privacy and the *research output*
  - mechanism: **data enclaves**



Data Enclaves

- concept: sensitive data in a room; control *in and out*
- result: *limited* access for researchers, privacy stays in room, safe results come out


Data Enclave - Michigan C

micda.psc.isr.umich.edu/enclave/

MiCDA Michigan Center on the Demography of Aging

MiCDA Data Enclave Overview

The MiCDA Data Enclave is funded by the Michigan Center on the Demography of Aging (NIA P30 AG12846), the Health and Retirement Study (NIA U01AG009740), and the Michigan Retirement Research Center (SSA 10 P-98358-5). The Survey Research Center and the Population Studies Center of the Institute for Social Research also provide support for the operation of this facility.

Page

- status
  - common in social sciences *and* in the networking industry
  - work-in-progress by several networking groups (UMich, USC, CSU, PCH)
  - term identified by Michael Bailey (UMich)

Implications of Data Enclaves

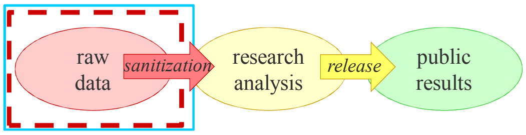
- enabler for research on more sensitive data
- more work to run, and to use
- *but a balanced point to explore*

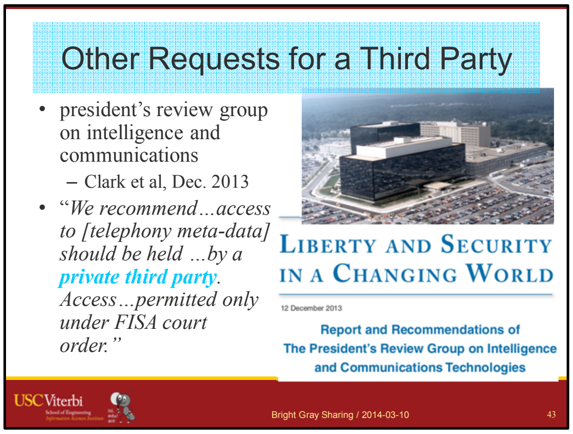
- future work:
  - experience with open enclaves for networking
  - *virtual enclaves*: replacing the locked room with monitoring

Third Parties and Data

- solution to any problem in computer science?

  …another level of indirection





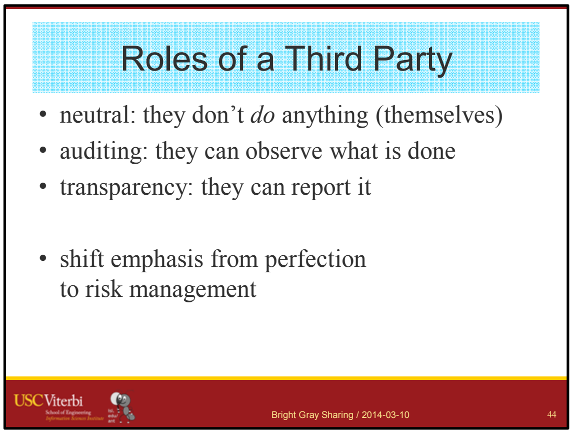
Other Requests for a Third Party

- president's review group on intelligence and communications
  - Clark et al, Dec. 2013
- "*We recommend…access to [telephony meta-data] should be held …by a* ***private third party****. Access…permitted only under FISA court order.*"

LIBERTY AND SECURITY IN A CHANGING WORLD

12 December 2013

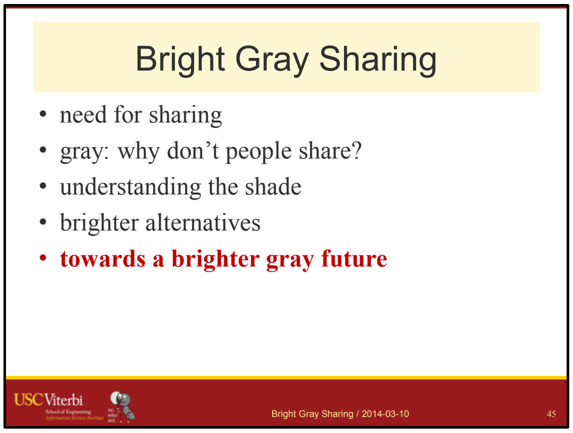
Report and Recommendations of The President's Review Group on Intelligence and Communications Technologies



Roles of a Third Party

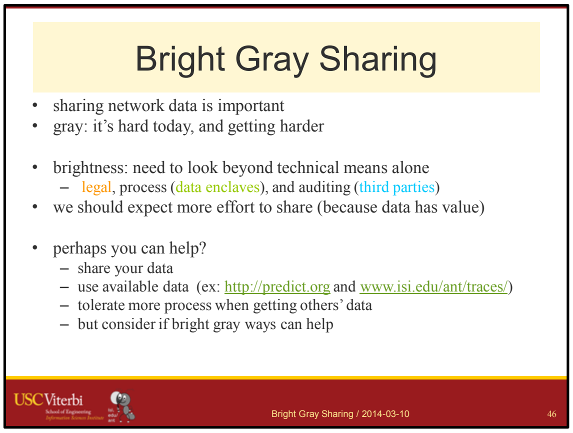
- neutral: they don't *do* anything (themselves)
- auditing: they can observe what is done
- transparency: they can report it

- shift emphasis from perfection to risk management



Bright Gray Sharing

- need for sharing
- gray: why don't people share?
- understanding the shade
- brighter alternatives
- **towards a brighter gray future**



Bright Gray Sharing

- sharing network data is important
- gray: it's hard today, and getting harder

- brightness: need to look beyond technical means alone
  - legal, process (data enclaves), and auditing (third parties)
- we should expect more effort to share (because data has value)

- perhaps you can help?
  - share your data
  - use available data (ex: http://predict.org and www.isi.edu/ant/traces/)
  - tolerate more process when getting others' data
  - but consider if bright gray ways can help