Privacy Research at Carnegie Mellon (A Sampling)

Jeannette M. Wing

President's Professor of Computer Science
Department Head
Computer Science Department

Information Security and Privacy Advisory Board
NIST
30 May 2012

Breadth of Approaches from Across CMU

- Science
  - Algorithms
  - Game Theory
  - Formal Methods
  - Machine Learning
  - Programming Languages
  - Statistics
- Engineering
  - Distributed Systems
  - Human-Computer Interaction
  - Mobile and Pervasive Computing
  - Networking
  - Security
  - Software Engineering
- Societal
  - Behavioral and Social Science
  - Economics
  - Ethics and Philosophy
  - Public Policy

Alessandro Acquisti
Avrim Blum
Travis Breaux
Lorrie Cranor
Anupam Datta
Stephen Fienberg
Virgil Gligor
Anupam Gupta
Bob Harper
Jason Hong
Jiashun Jin
Ramayya Krishnan
Peter Madsen
Rema Padman
Frank Pfenning
Bhiksha Raj
Alessandro Rinaldo
Norm Sadeh
M. Satyanarayanan
Tuomas Sandholm
Srini Seshan
Larry Wasserman
Jeannette Wing
Eric Xing

24 faculty, 6 Schools/Centers, 10 Departments

Sampling of Three Foci

- Auditing, Accountability, Compliance
  - Formal Methods: Anupam Datta, Jeannette Wing
  - Software Engineering: Travis Breaux
  - *Applications to Healthcare: Travis Breaux, Rema Padman, Jeannette Wing*

- Public (Government) Databases
  - Statistical, e.g., Differential Privacy
    - Sanitized Databases: Avrim Blum
    - Practical Limits: Steve Fienberg, Alessandro Rinaldo, Larry Wasserman
  - Vulnerabilities: Alessandro Acquisti

- Consumer-centric
  - Usability: Lorrie Cranor
  - Location privacy: Jason Hong, Norm Sadeh
  - Behavioral Economics: Alessandro Acquisti

Auditing, Accountability, Compliance Checking

Formal Methods:
Semantics of Use and Purpose

[IEEE Security and Privacy 2012]

Michael Tschantz, Anupam Datta, Jeannette Wing

Computer Science Department, CyLab

Purpose in EU Law

- Member States shall provide that personal data must be [. . .] collected for specified, explicit and legitimate **purposes** and not further processed in a way incompatible with those **purposes**

Purpose in Yahoo's Policy

- Yahoo!'s practice is not to use the content of messages [...] for marketing **purposes**.[1]
- This information is transmitted [...] for the **purpose** of registering your web address [...].[1]
- Yahoo! uses information for the following general **purposes**: to customize the advertising [...].[2]
- Yahoo! does not contact children [...] for marketing **purposes** [...].[2]
- [...] companies who may use this information for their own **purposes**.[2]

[1]http://info.yahoo.com/privacy/us/yahoo/mail/details.html
[2]http://info.yahoo.com/privacy/us/yahoo/details.html

Purpose in HIPAA

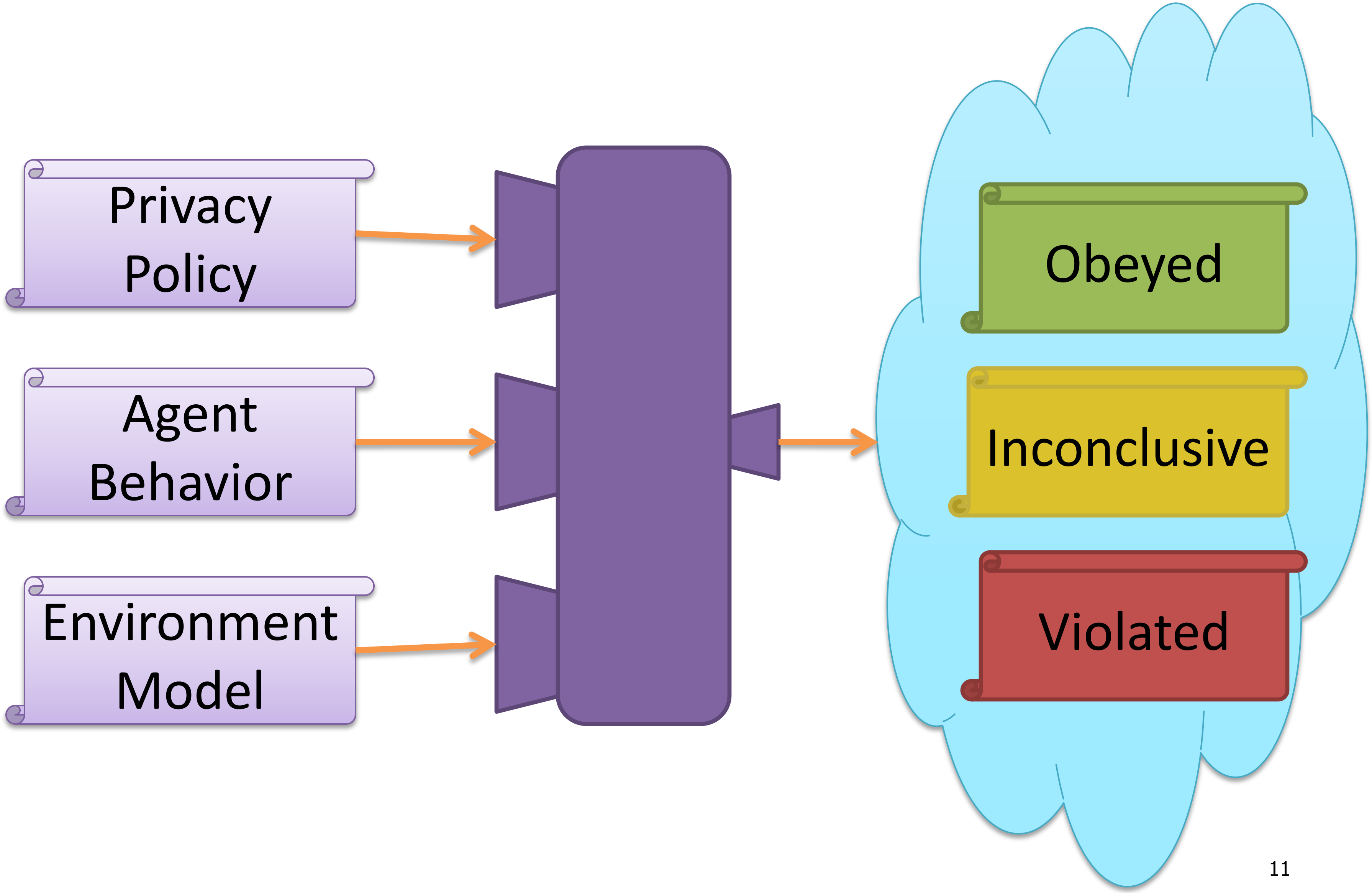
- [...] use and disclose protected health information [...] for the following **purposes** or situations:
  (1) To the Individual (unless required for access or accounting of disclosures);
  (2) Treatment, Payment, and Health Care Operations;
  …

Purpose in Government Agency Policies

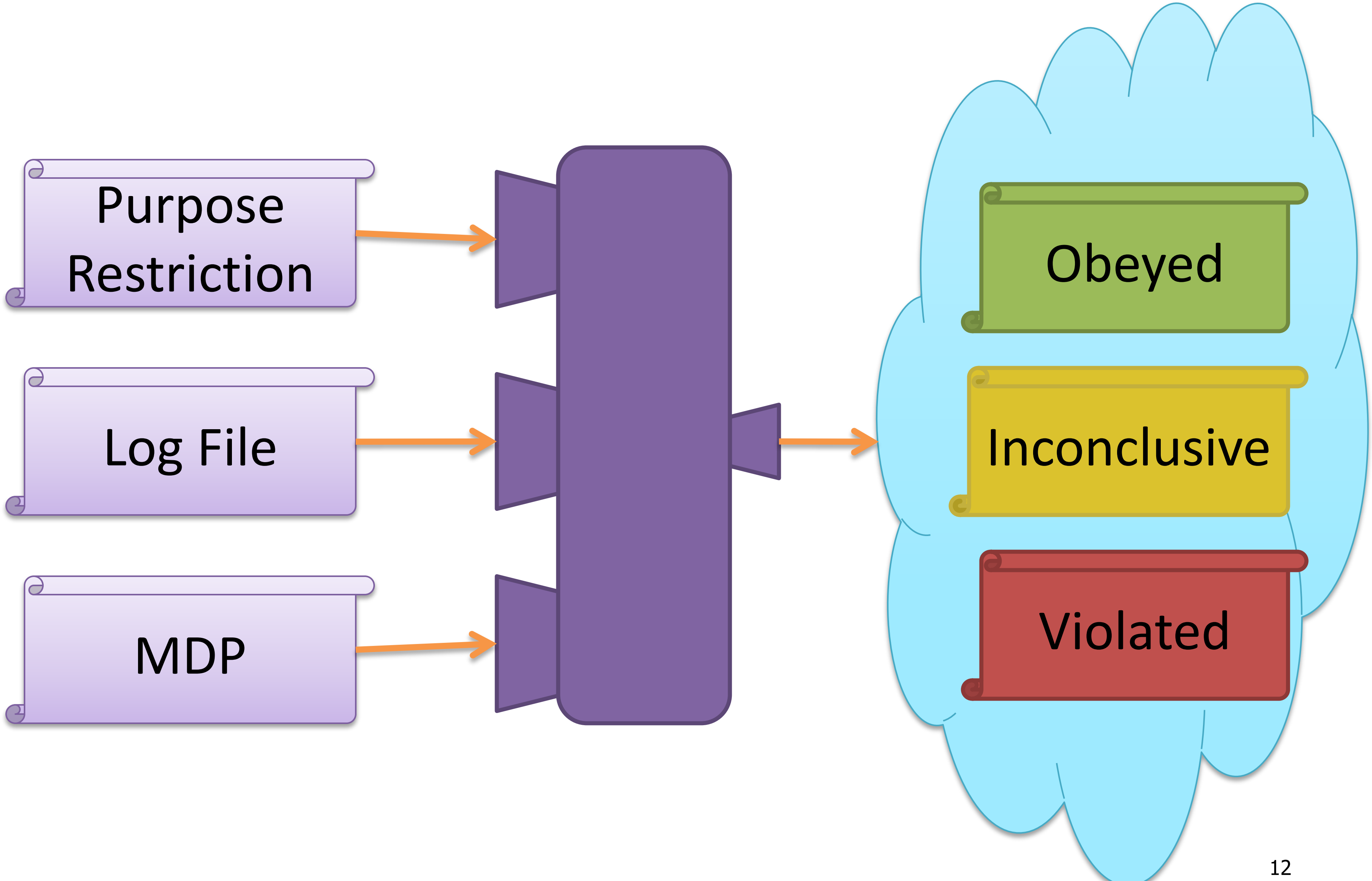
- By providing your personal information, you give [SSA] consent to use the information only for the **purpose** for which it was collected. We describe those **purposes** when we collect information.

  http://www.ssa.gov/privacy.html

Auditing

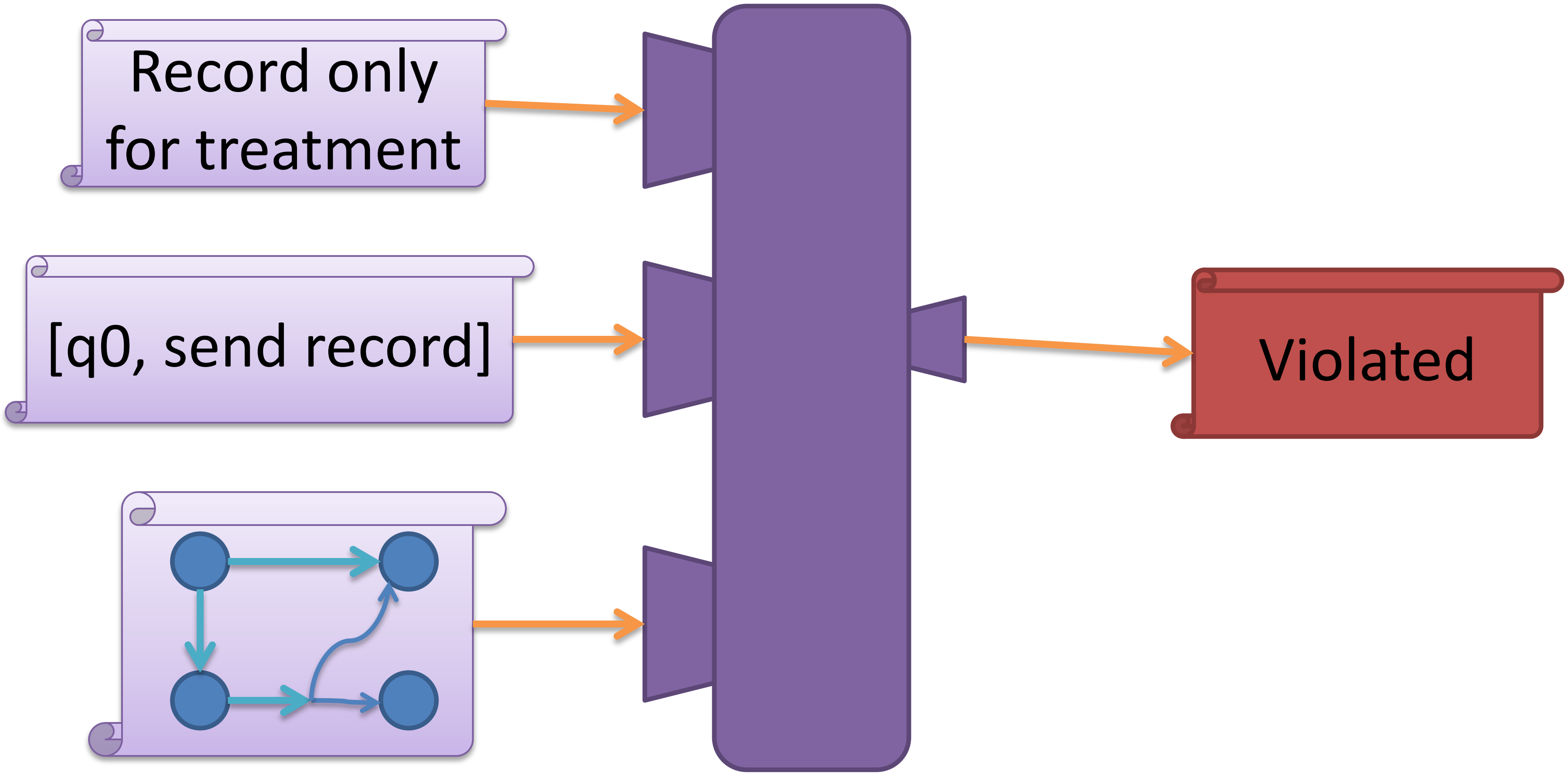

Privacy Policy
Agent Behavior
Environment Model
Obeyed
Inconclusive
Violated

Auditing



Purpose Restriction
Log File
MDP
Obeyed
Inconclusive
Violated

Auditing



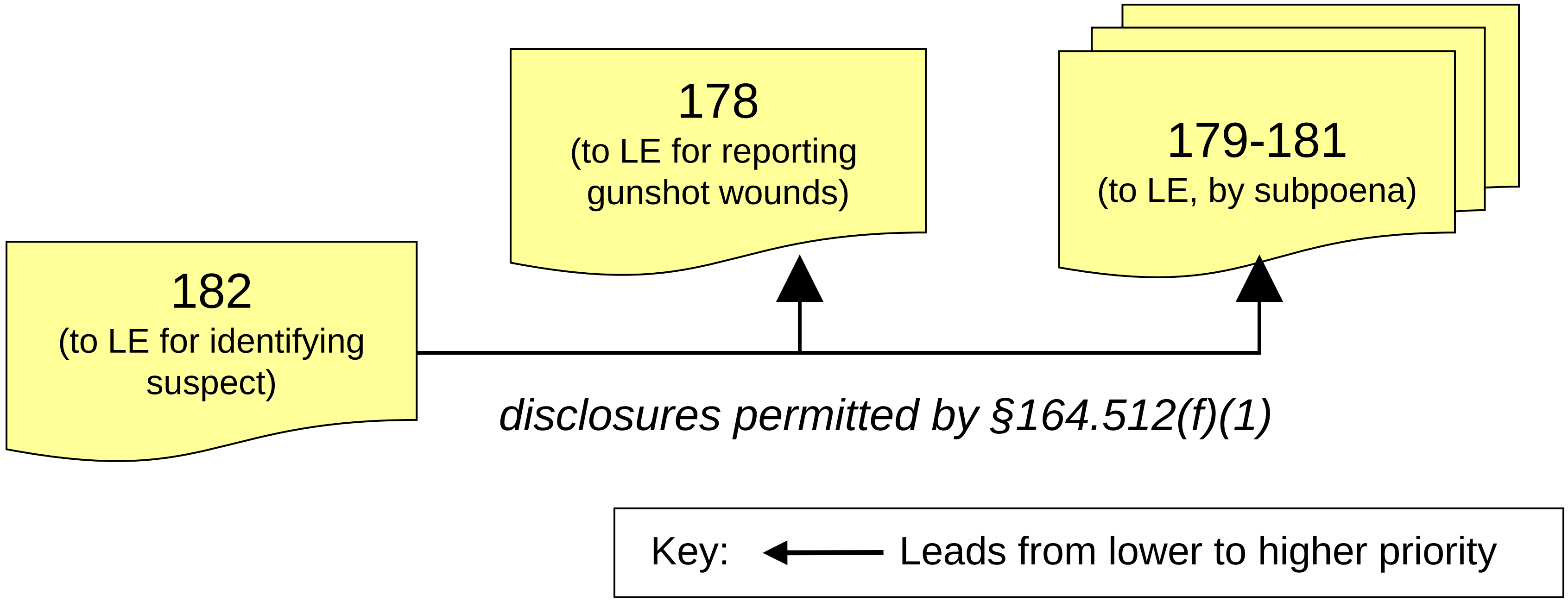
Languages and Software Architecture: Exceptions in Law and Patterns of Compliance

Travis Breaux
Institute for Software Research
School of Computer Science

Handling Legal Exceptions

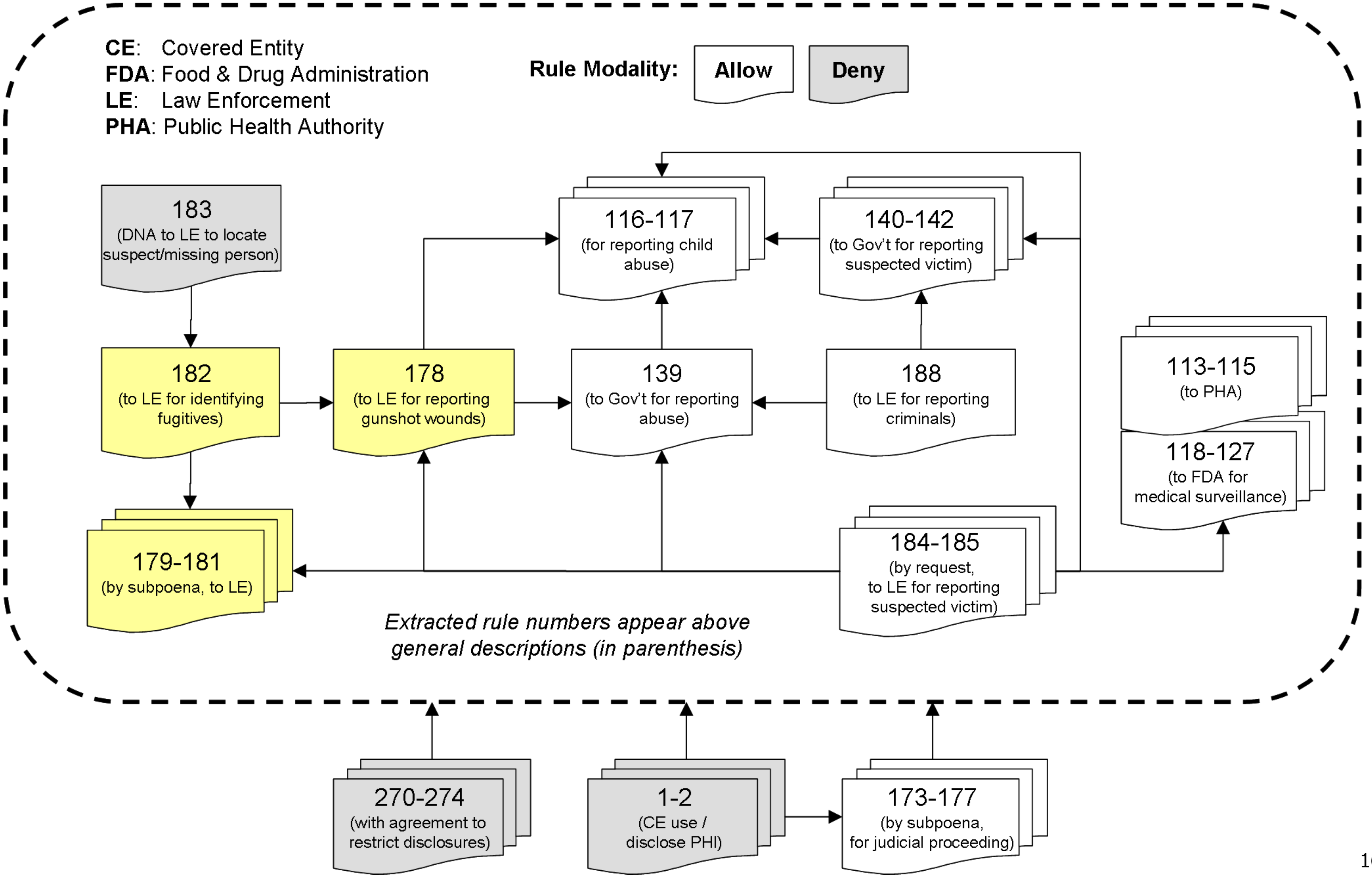
[IEEE TSE, January 2008]

HIPAA §164.512(f)(2): Except for disclosures required by law as permitted by paragraph 164.512(f)(1), a CE may disclose PHI in response to a law enforcement (LE) official's request for the purpose of identifying or locating a suspect

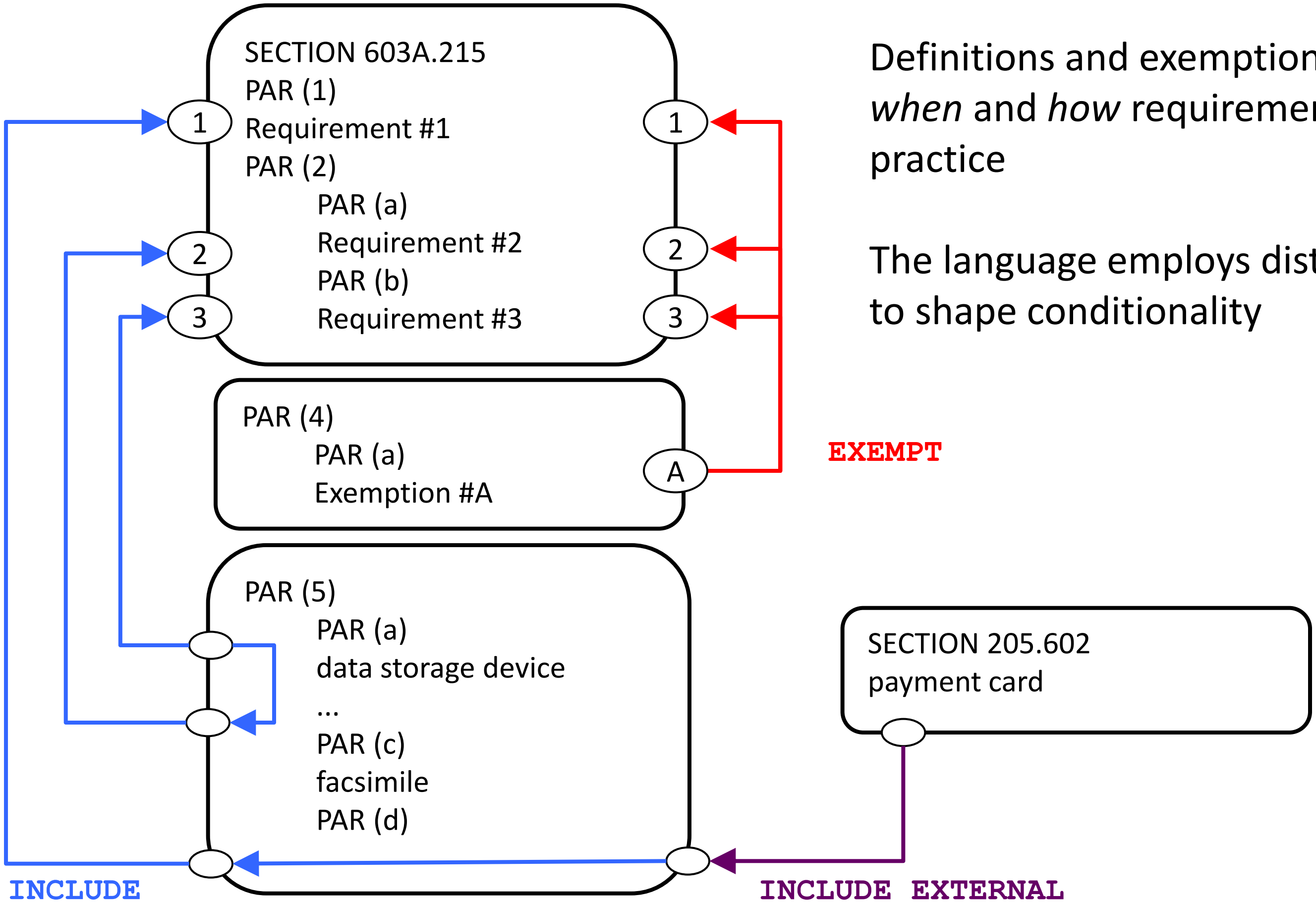


Requirements Exception Hierarchy

[IEEE TSE, January 2008]



Shaping Policy via Distributed Conditionality

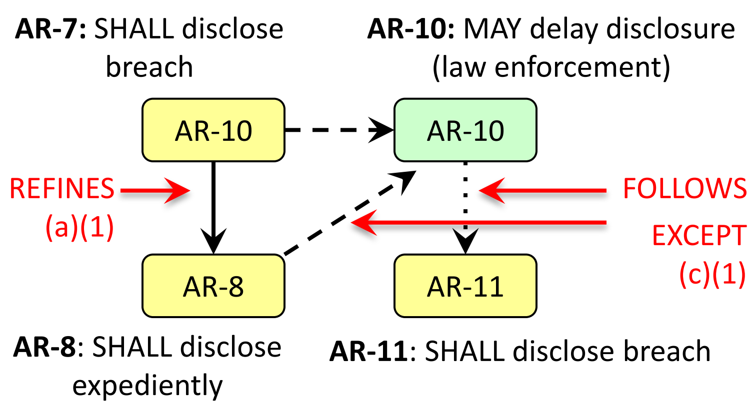


Definitions and exemptions shape *who*, *when* and *how* requirements are applied in practice

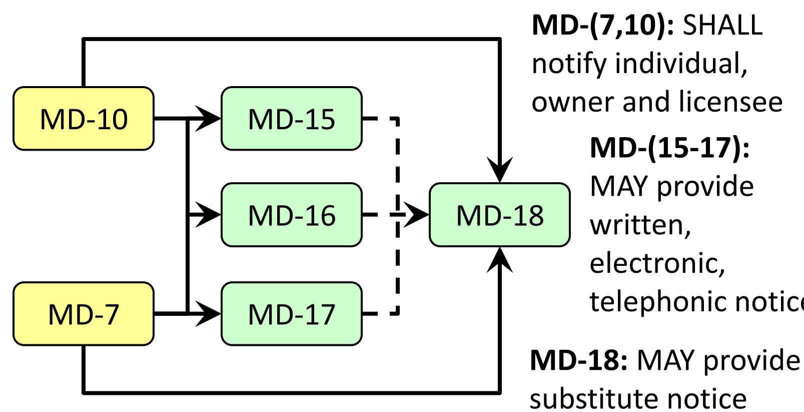
The language employs distributed controls to shape conditionality

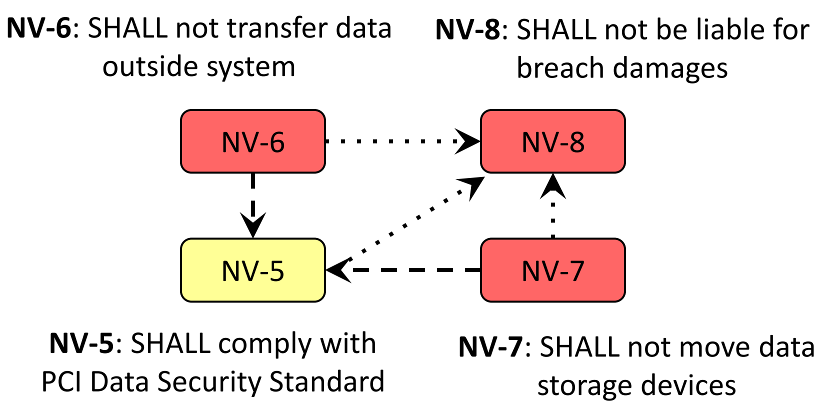
Language-supported Compliance Patterns

Process Suspension

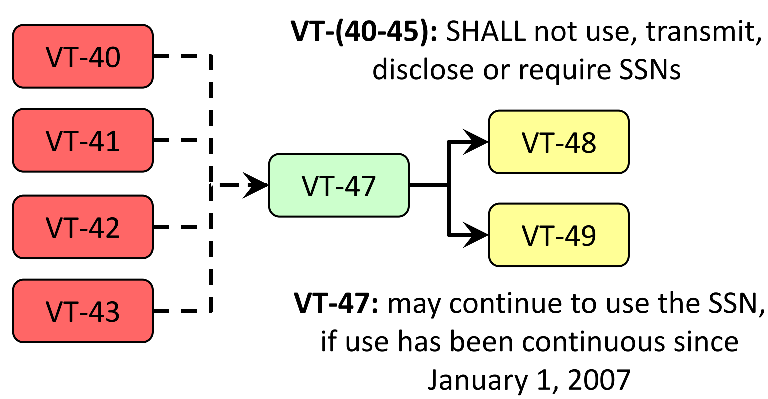


Design Alternatives





Indemnification



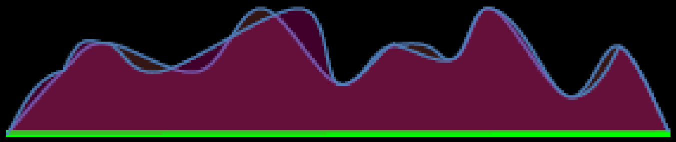
Legacy System Evolution

Statistics:
Release of Public (Government) Data

In this section, all slides with a black background are from Avirm Blum.

The science of privacy

- Fundamental breakthrough came from MSR in work of [Dwork-McSherry-Nissim-Smith] building on earlier work of Dwork et al, in definition of differential privacy.

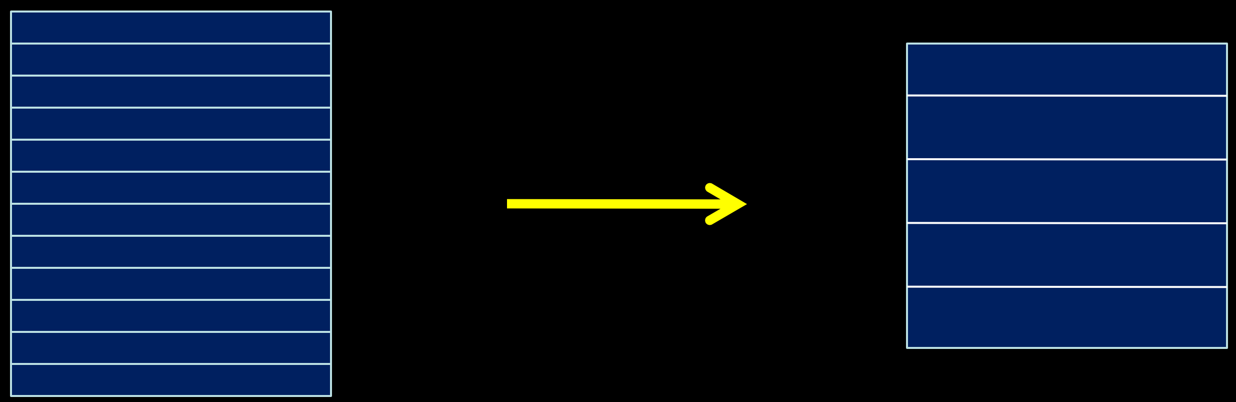


Any participant should be able to plausibly deny any fact claimed about them (the probability of any given output of mechanism would change by only $1 \pm \epsilon$)

Outputting Sanitized Databases Using Differential Privacy

Avrim Blum
Computer Science Dept

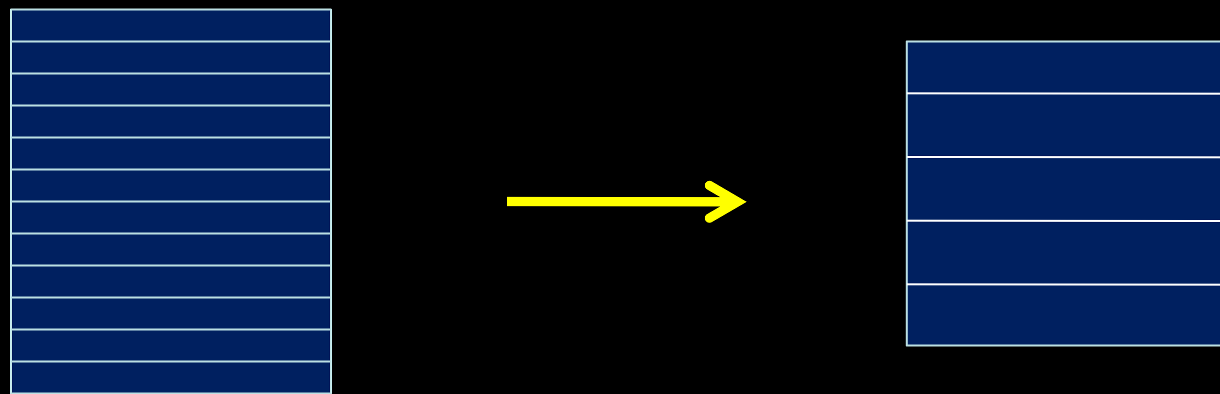
What about outputting sanitized databases?



- So far, just question-answering. Each answer leaks some privacy – at some point, have to shut down.
- What about outputting a sanitized database that people could then examine as they wish?

And is related to the original database…

What about outputting sanitized databases?



- Could ask a few questions (using previous mechs) and then engineer a database that roughly agrees on these answers.
- But really, we want a database that matches on questions we haven't asked yet.
- Do you need to leak privacy in proportion to number of questions asked?

What about outputting sanitized databases?

Actually, no you don't [Blum-Ligett-Roth] (At least not for count-queries)

- Fix a class C of quantities to preserve. E.g., fraction of entries with $x[i_1]=1$, $x[i_2]=0 \ldots x[i_k]=1$.
- Want ε-privacy and preserve all $q \in C$ up to $\pm\alpha$.
- [BLR] show: in principle, can do with database of size only $n = O(d \log |C|)$.

Allowing exponentially-many questions!

Statistical Disclosure Limitation & the Challenge of Societal-Scale Data

Stephen E. Fienberg
Department of Statistics, Heinz College, Machine Learning Department, and Cylab
Carnegie Mellon University
Pittsburgh, PA 15213-3890 USA
fienberg@stat.cmu.edu

Differential Privacy for Protecting Multi-dimensional Contingency Table Data: Extensions and Applications, *Journal of Privacy and Confidentiality*, 2011.

Yang, Fienberg, & Rinaldo: Examined DP Approach

Robustness of approach for RU tradeoff

- Edwards 2^6 genetics table, with n=70.
- Czech auto workers 2^6 heart attack risk table, with n=1,841.
- Rochdale 28 survey data on women's work, with n=665;very sparse structure.
- American Community Survey 4 × 4 × 16 travel to work table.

- **National Long Term Care Survey**
  - 2^{16} disability table with n=21,574.
  - 2^{96+5} version based on 6 waves (plus mortality), n~45,000. Our models have no MSSs!

Lessons Learned

- **As ε increases, amount of noise added decreases**
  - **Deviance between DP generated tables and real MLEs gets smaller.**
  - **If we add a lot of noise, it has strong privacy guarantees but the statistical inference becomes infeasible.**
  - **When we add little noise, the statistical inference is better but no privacy guarantees.**
- **DP struggles with releasing useful information associated with large sparse contingency tables.**

Differential Privacy summary

Positives:

- Clear semantic definition. Any event (anything an adversary might do to you) has nearly same prob if you join or don't join, lie or tell the truth.
- Nice composability properties.
- Variety of mechanisms developed for question answering in this framework.
- *Some* work on sanitized database release.

Differential Privacy summary

Negatives / open issues

- It's a pessimistic/paranoid quantity, so may be more restrictive than needed.
- "ε" is not zero. Privacy losses add up with most mechanisms (but see, e.g., [RR10],[HR10])
- Doesn't address group information.
- Notion of "neighboring database" might need to be different in network settings.
- …

Consumer-Centric: Usability, Location Privacy, Behavioral Economics

CUPS Lab Privacy Research Overview

Lorrie Faith Cranor

May 2012

Carnegie Mellon University CyLab

isr institute for SOFTWARE RESEARCH

Engineering & Public Policy

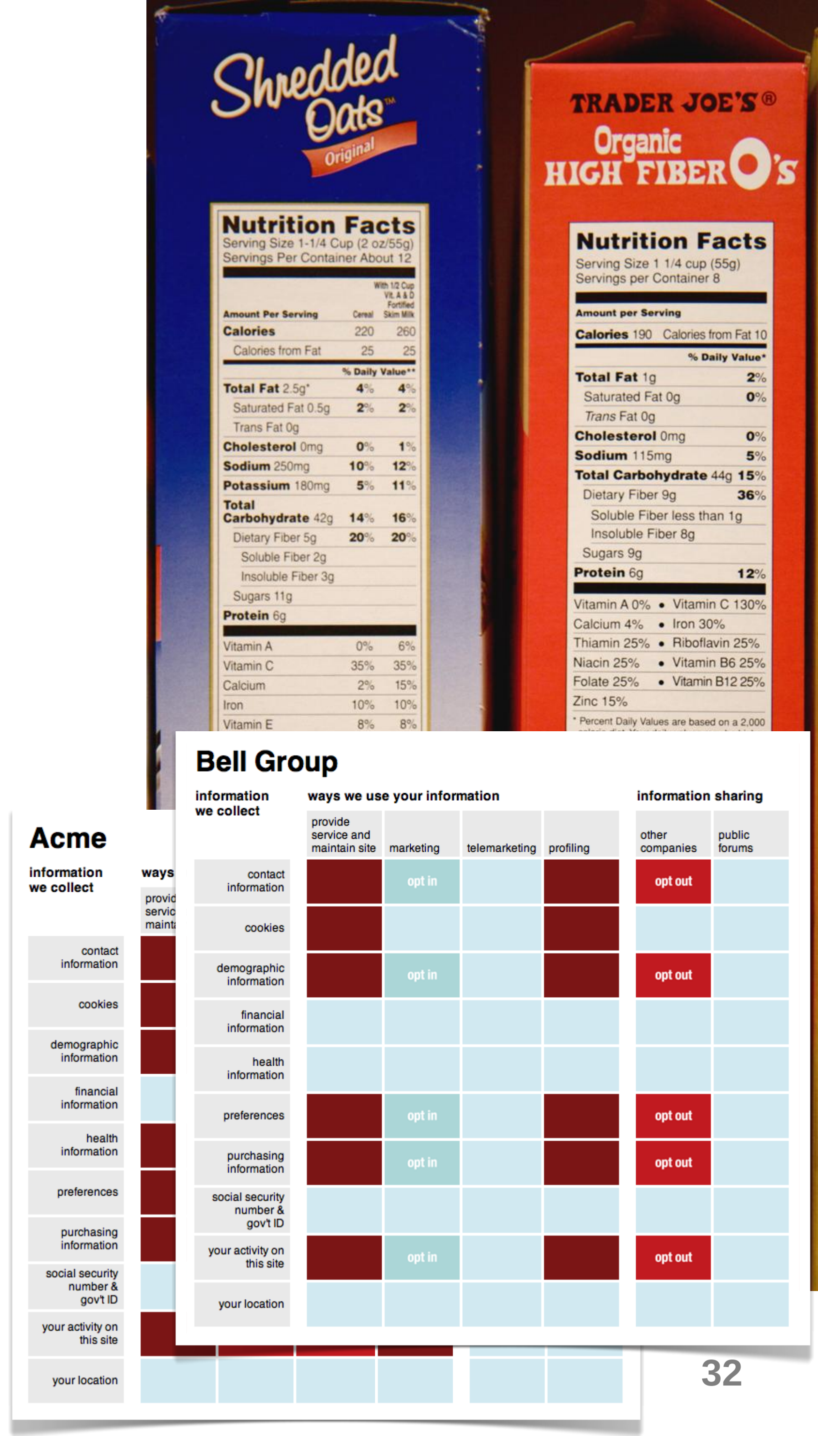
CyLab Usable Privacy & Security Laboratory

HTTP://CUPS.CS.CMU.EDU

Towards a privacy "nutrition label"

- Standardized format
  - People learn where to find answers
  - Facilitates policy comparisons
- Standardized language
  - People learn terminology
- Brief
  - People find info quickly
- Linked to extended view
  - Get more details if needed

Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach
[Kelley, Cesca, Bresee, and Cranor, CHI 2010]



Online behavioral advertising notice and choice

- Tools that facilitate notice and choice about OBA are part of self-regulatory privacy efforts
  - Browsers can block cookies
  - Opt-out cookies, AdChoices icon
  - Browser plugins
- Series of studies to investigate user understanding of OBA and usability of tools
  - Users understand little about OBA, unaware of tools
  - Tools are difficult to configure properly
  - Users don’t know enough about ad companies to choose between them
  - Users unfamiliar with Adchoices icon and afraid to click on it

CMU privacy nudges project

- Goal: Study, design, and test systems that anticipate and sometimes exploit cognitive and behavioral biases that hamper users' privacy and security decision making
- Multidisciplinary project: behavioral economics + decision sciences + machine learning + human computer interaction + ….
- Social network regrets – Surveyed 1000+ users about things they regret doing on Facebook and Twitter, identified categories of regrets and underlying causes
- Nudge prototypes – Testing software that will nudge FB and Twitter users before they post, e.g., with photos of random friends or a countdown timer

I Regretted the Minute I Pressed Share: A Qualitative Study of Regrets on Facebook.
[Wang, Komanduri, Leon, Norcie, Acquisti, Cranor 2011]

Location Privacy

Norman M. Sadeh

Professor, ISR - School of Computer Science

Director, Mobile Commerce Lab.

Co-Director, COS PhD Program

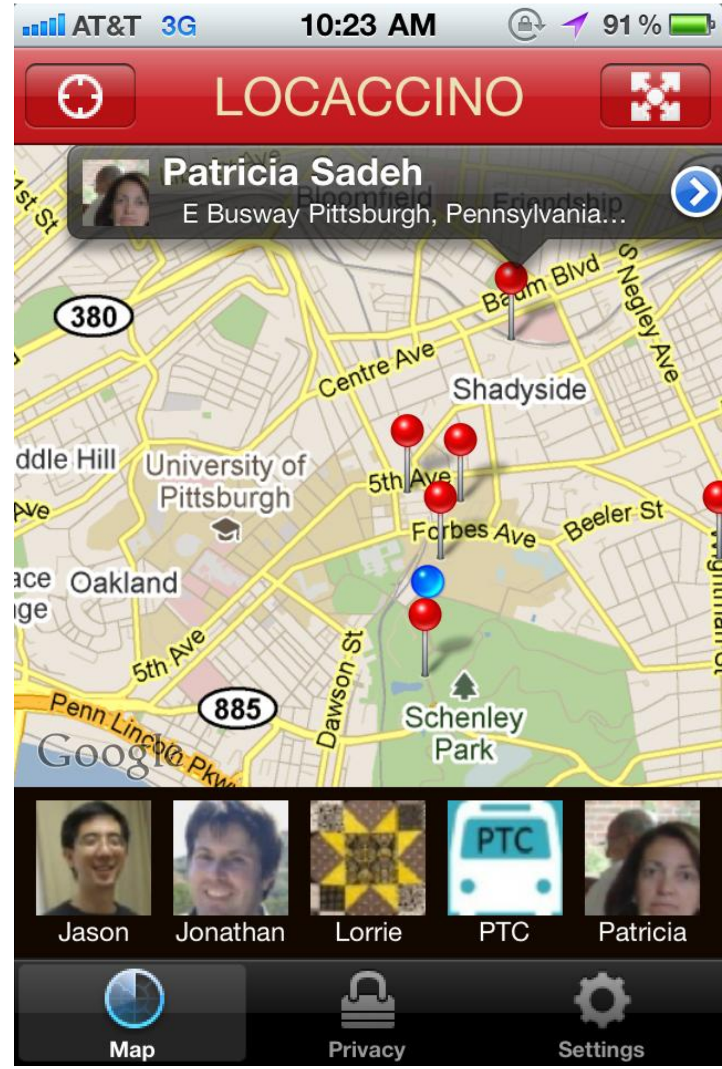
Carnegie Mellon University

www.cs.cmu.edu/~sadeh

mobile commerce lab

Empowering Users to Regain Control of their Privacy

- Mobile Apps collect a wide range of information about their users
- Research combining:
  - Understanding people's privacy preferences?
    - How diverse? How complex? Do they change?
  - User-Oriented Machine Learning/AI: Can we learn people's preferences through selective dialogues?
  - Better Uis
- Informed by Large-Scale Deployments (e.g. location sharing app)

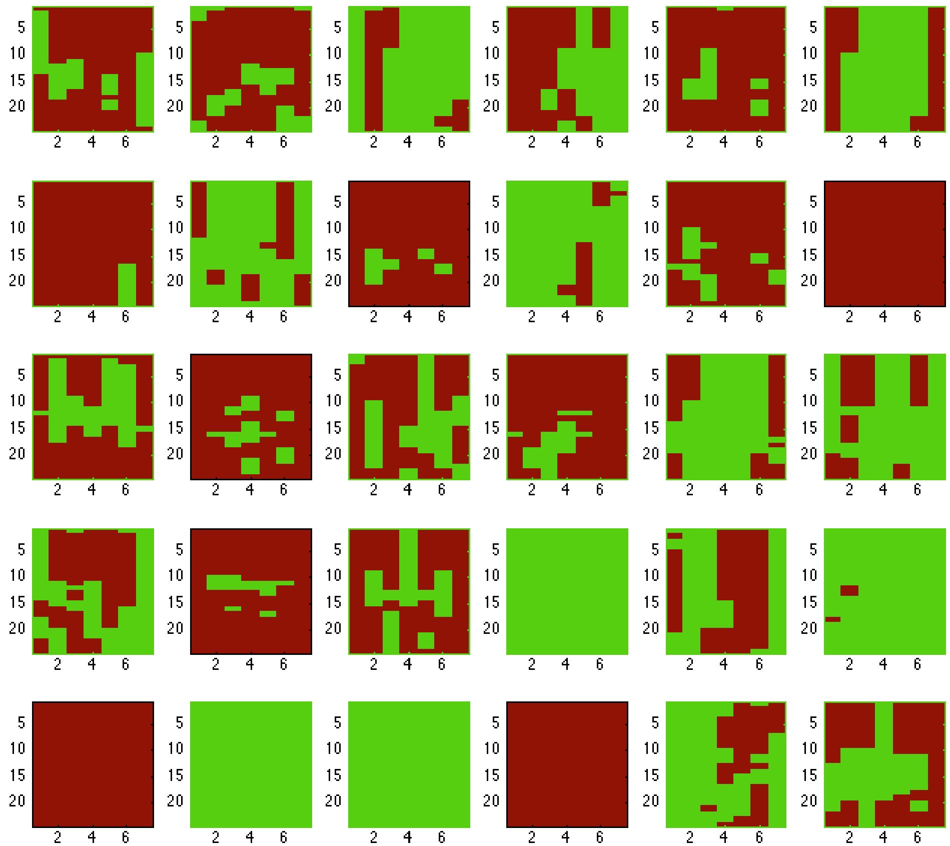


www.locaccino.org

mobile commerce lab

Diverse and Complex Privacy Preferences

Each square represents **a different individual** and displays her willingness to share her location with members of the CMU campus community



Green: Share

Red: Don't

In each square:

Horizontal axis: 7 days of the week

Vertical axis: 24 hours of the day

mobile commerce lab

The Economics and Behavioral Economics of Privacy

Alessandro Acquisti

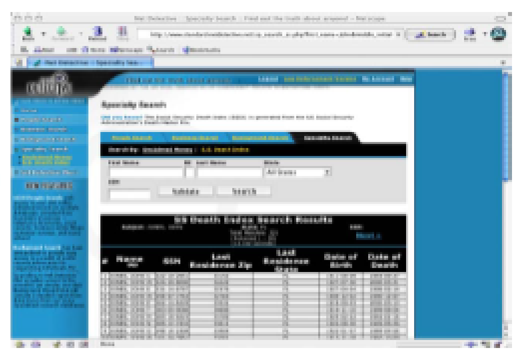
Heinz College/CyLab
Carnegie Mellon University

PNAS 2009, NY Times "big splash" article

Predicting SSNs from public data

- We reverse-engineered SSN issuance patterns, showing that they were significantly less random than previously predicted
- We found that mere knowledge of an individual's DOB and State of birth is sufficient to predict that individual's SSN

+



= SSN

Experimental studies over ~8 years on behavioral economics of privacy

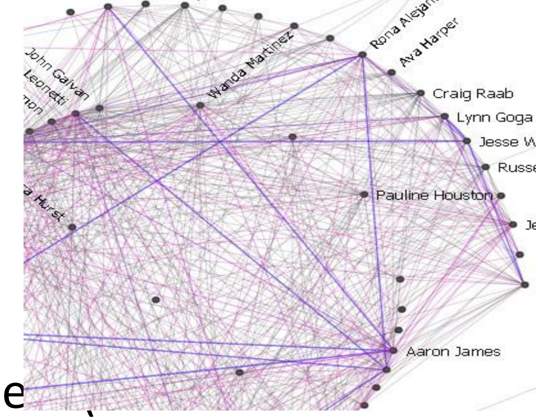
A sampling of results:

- Individuals more likely to disclose sensitive information to unprofessional sites than professional sites.
- People assign different values to their personal information depending on whether they are focusing on **protecting it** or **revealing it**
- People may make disclosure decisions that they stand to later regret.
- Risks greatly magnified in online information revelation

Overall implications of these privacy studies

- "Choice & notification" privacy model may be outdated
- Implications for policy-making & the debate on privacy regulation
  - Consider: Chicago School approach vs. privacy advocates
  - "Nudging" privacy?

Privacy in Social Networks



- Behavioral Economics: Alessandro Acquisti
  - Control vs. Access/Use: Giving more control to use (social networks) over information publication seems to generate higher willingness to disclose sensitive information

- Statistical: Avrim Blum and Anupam Datta
  - Given a sensitive social network, can we release a sanitized version of it that preserves privacy and is still useful? Inspired by differential privacy work.

- Formal Methods: Bob Harper and Jeannette Wing
  - What is a formal logic for reasoning about privacy properties in a social network? Uses a linear, epistemic logic.

Breadth of Approaches from Across CMU

- Science
  - Algorithms
  - Game Theory
  - Formal Methods
  - Machine Learning
  - Programming Languages
  - Statistics
- Engineering
  - Distributed Systems
  - Human-Computer Interaction
  - Mobile and Pervasive Computing
  - Networking
  - Security
  - Software Engineering
- Societal
  - Behavioral and Social Science
  - Economics
  - Ethics and Philosophy
  - Public Policy

Alessandro Acquisti	Peter Madsen
Avrim Blum	Rema Padman
Travis Breaux	Frank Pfenning
Lorrie Cranor	Bhiksha Raj
Anupam Datta	Alessandro Rinaldo
Stephen Fienberg	Norm Sadeh
Virgil Gligor	M. Satyanarayanan
Anupam Gupta	Tuomas Sandholm
Bob Harper	Srini Seshan
Jason Hong	Larry Wasserman
Jiashun Jin	Jeannette Wing
Ramayya Krishnan	Eric Xing

24 faculty, 6 Schools/Centers, 10 Departments

Thank you!