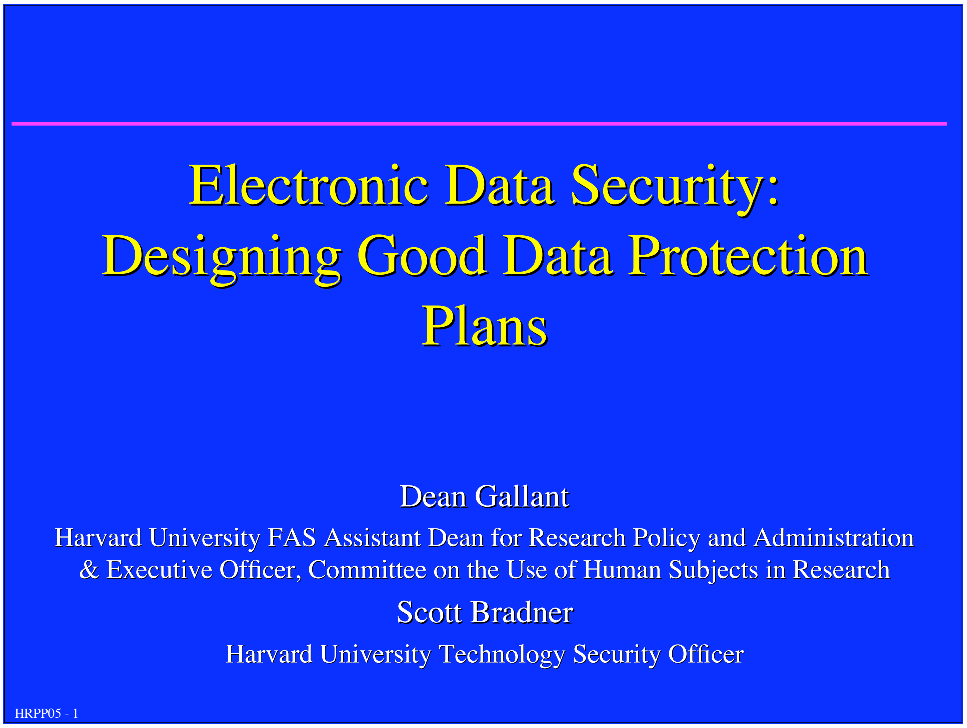


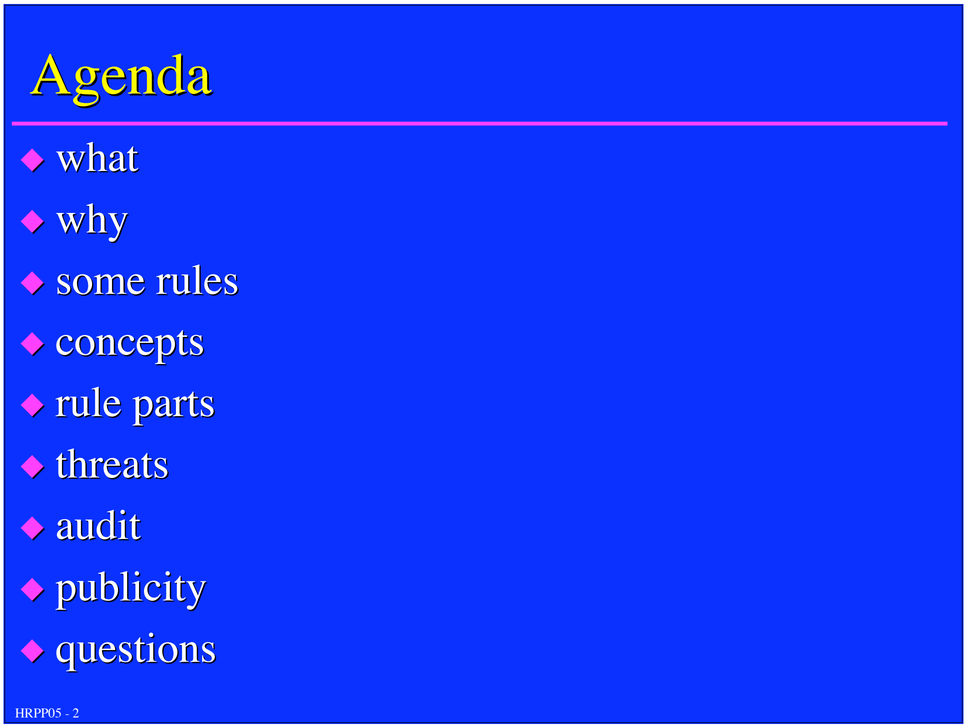
Electronic Data Security: Designing Good Data Protection Plans

Dean Gallant

Harvard University FAS Assistant Dean for Research Policy and Administration
& Executive Officer, Committee on the Use of Human Subjects in Research

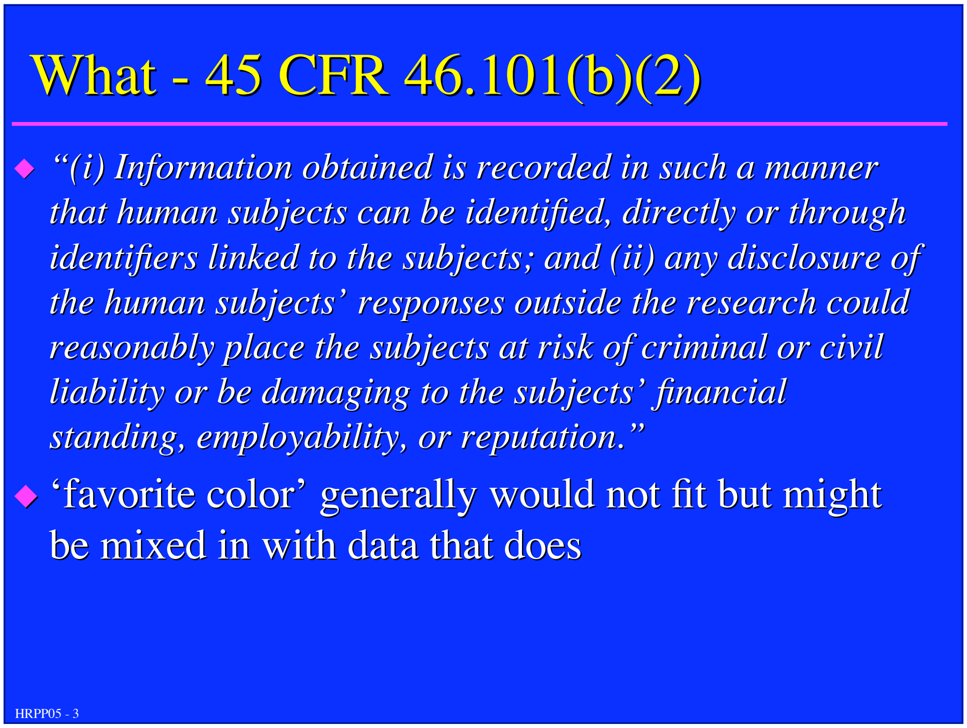
Scott Bradner

Harvard University Technology Security Officer



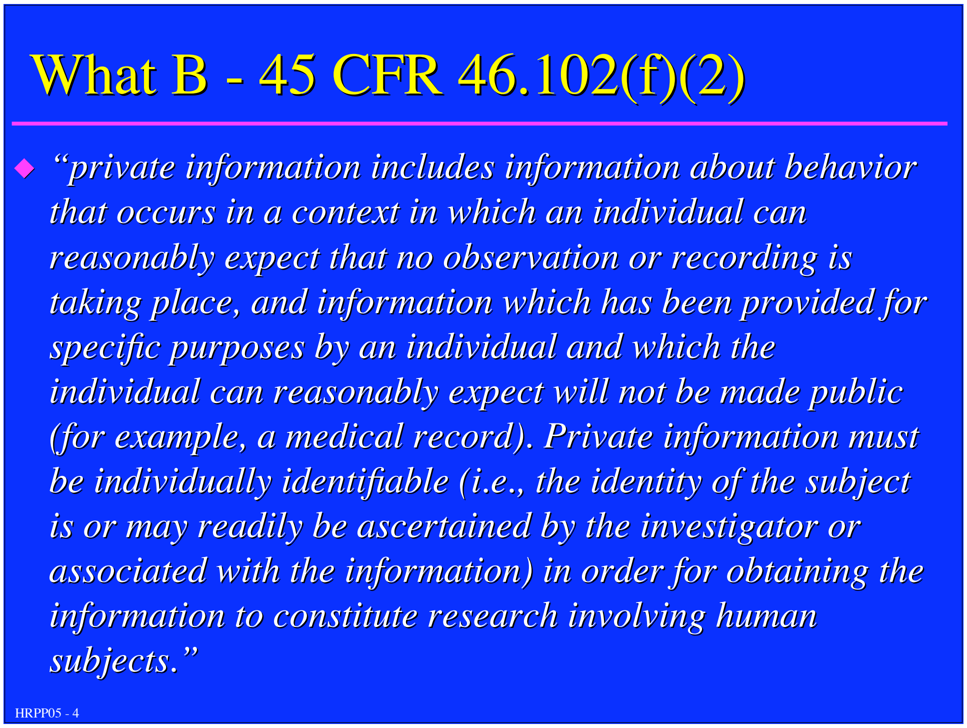
Agenda

- what
- why
- some rules
- concepts
- rule parts
- threats
- audit
- publicity
- questions



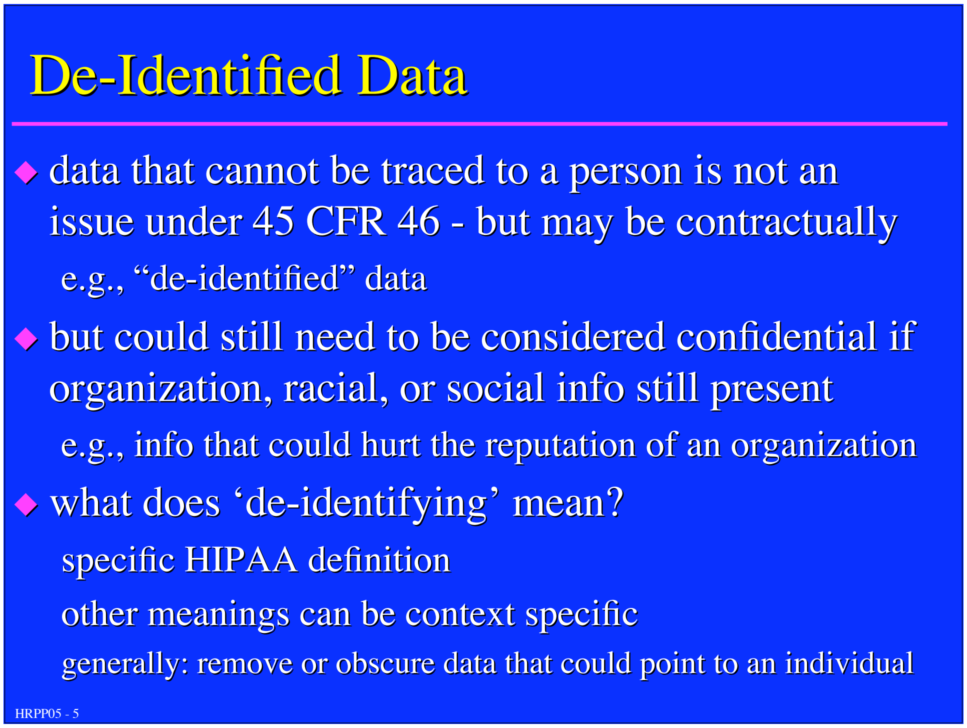
What - 45 CFR 46.101(b)(2)

- *"(i) Information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation."*
- 'favorite color' generally would not fit but might be mixed in with data that does



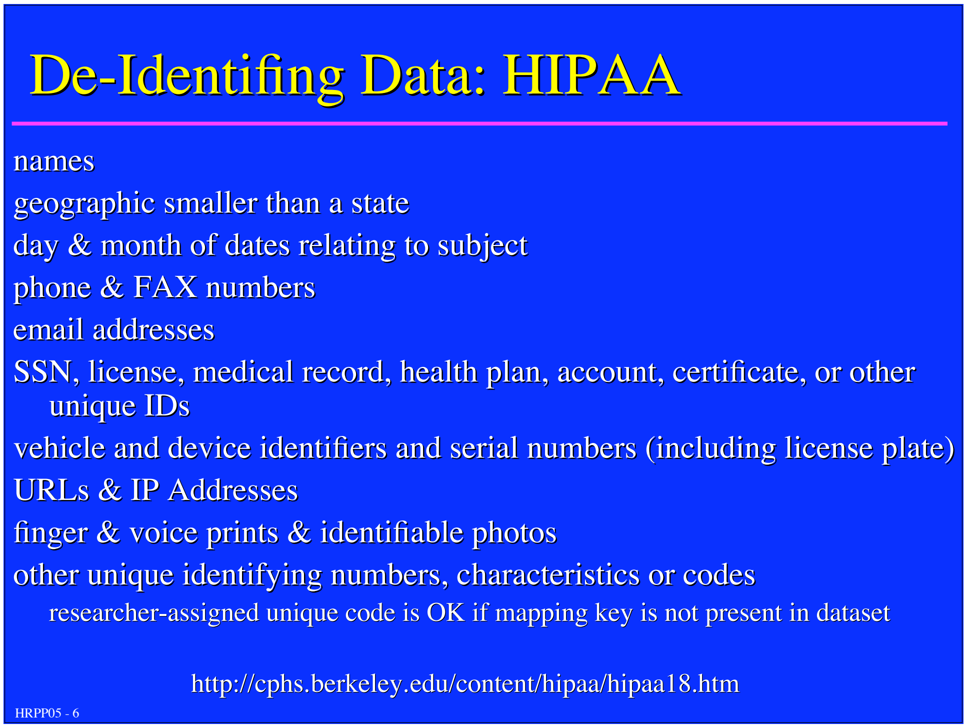
What B - 45 CFR 46.102(f)(2)

- *"private information includes information about behavior that occurs in a context in which an individual can reasonably expect that no observation or recording is taking place, and information which has been provided for specific purposes by an individual and which the individual can reasonably expect will not be made public (for example, a medical record). Private information must be individually identifiable (i.e., the identity of the subject is or may readily be ascertained by the investigator or associated with the information) in order for obtaining the information to constitute research involving human subjects."*



De-Identified Data

- data that cannot be traced to a person is not an issue under 45 CFR 46 - but may be contractually
  - e.g., “de-identified” data
- but could still need to be considered confidential if organization, racial, or social info still present
  - e.g., info that could hurt the reputation of an organization
- what does ‘de-identifying’ mean?
  - specific HIPAA definition
  - other meanings can be context specific
  - generally: remove or obscure data that could point to an individual



De-Identifing Data: HIPAA

names

geographic smaller than a state

day & month of dates relating to subject

phone & FAX numbers

email addresses

SSN, license, medical record, health plan, account, certificate, or other unique IDs

vehicle and device identifiers and serial numbers (including license plate)

URLs & IP Addresses

finger & voice prints & identifiable photos

other unique identifying numbers, characteristics or codes

- researcher-assigned unique code is OK if mapping key is not present in dataset

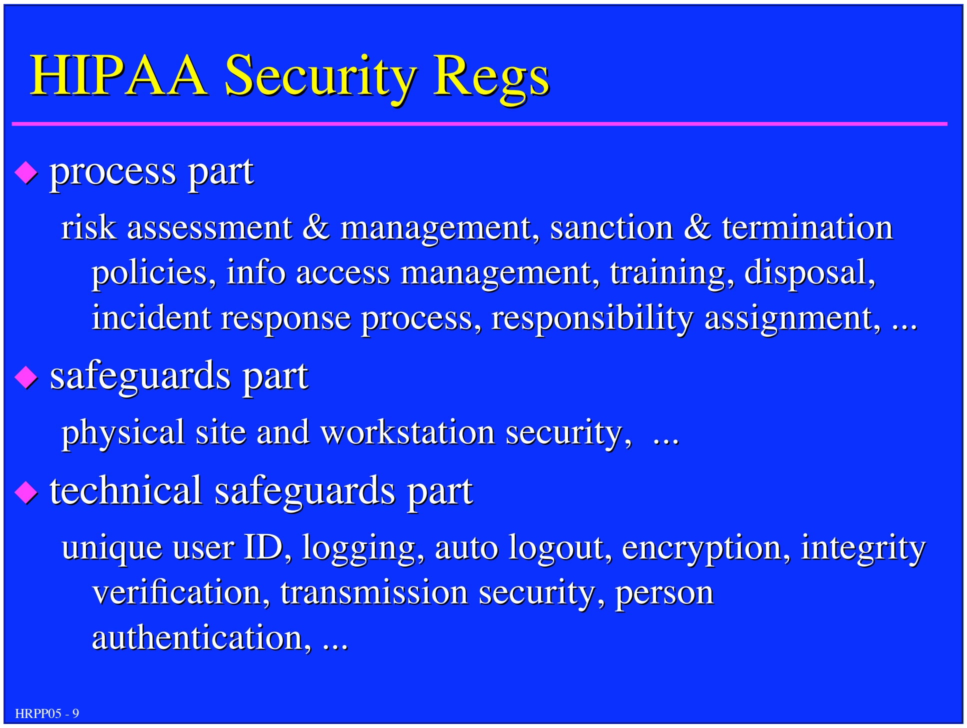
http://cphs.berkeley.edu/content/hipaa/hipaa18.htm

Why - 45 CFR 46.111

- *"(7) When appropriate, there are adequate provisions to protect the privacy of subjects and maintain the confidentiality of data"*
- "adequate provisions" is not a clear directive
  - e.g., congressional testimony after 1983 Beirut bombing
    - "the protections were adequate"
- this talk discusses "adequate" in regards to protecting human subject data

HIPAA as (an Extreme) Definition

- HIPAA is the main example of the US Government defining "adequate" when it comes to protecting data
- two parts
  - HIPAA privacy rule - notice, etc.
    - not relevant here
  - HIPAA security regs - how to protect data
    - maybe overkill but worth looking at



HIPAA Security Regs

- process part

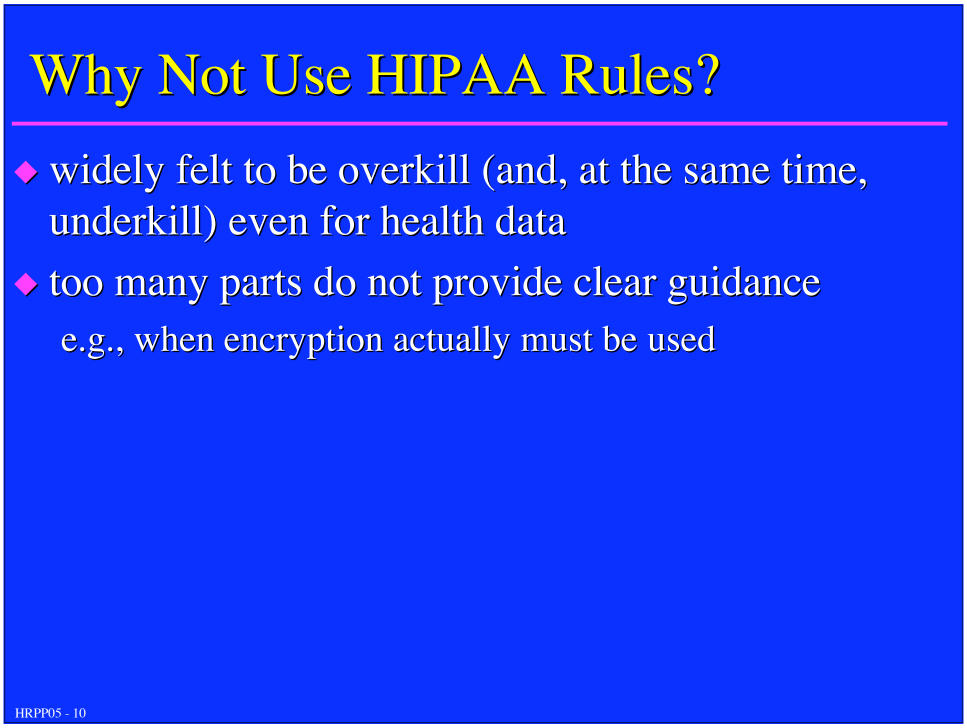
  risk assessment & management, sanction & termination policies, info access management, training, disposal, incident response process, responsibility assignment, ...

- safeguards part

  physical site and workstation security, ...

- technical safeguards part

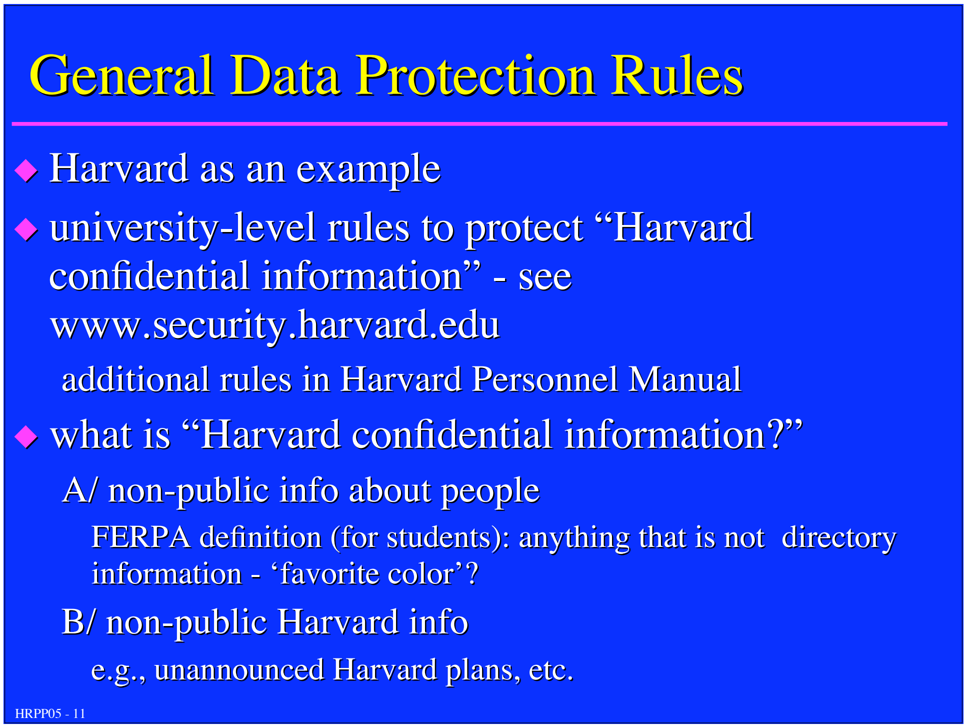
  unique user ID, logging, auto logout, encryption, integrity verification, transmission security, person authentication, ...



Why Not Use HIPAA Rules?

- widely felt to be overkill (and, at the same time, underkill) even for health data
- too many parts do not provide clear guidance

  e.g., when encryption actually must be used



General Data Protection Rules

- Harvard as an example
- university-level rules to protect “Harvard confidential information” - see www.security.harvard.edu

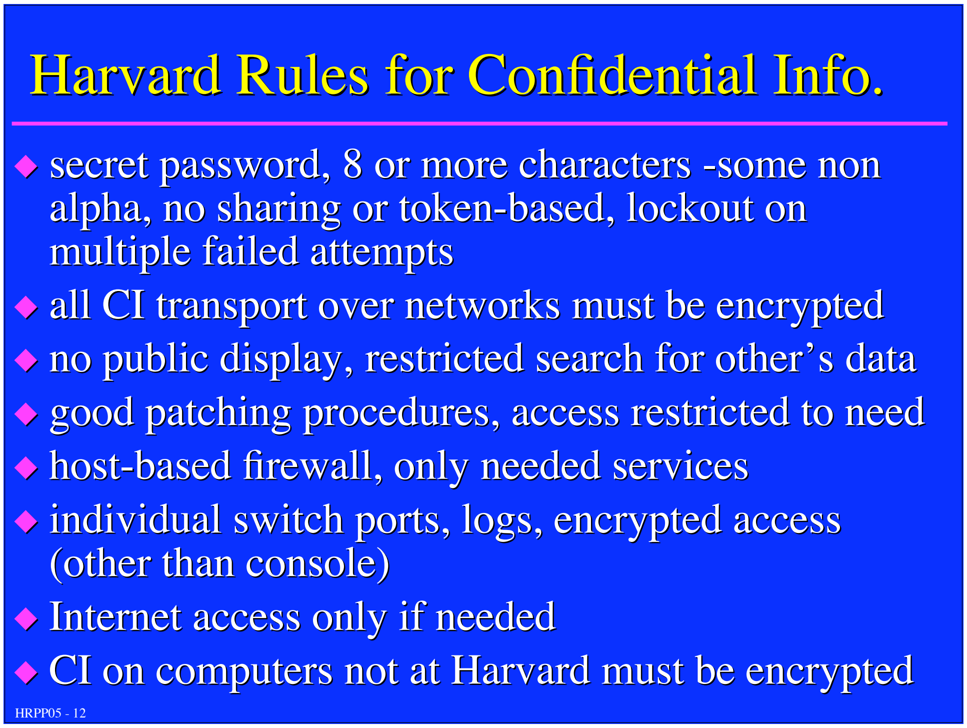
  additional rules in Harvard Personnel Manual
- what is “Harvard confidential information?”

  A/ non-public info about people

  FERPA definition (for students): anything that is not directory information - ‘favorite color’?

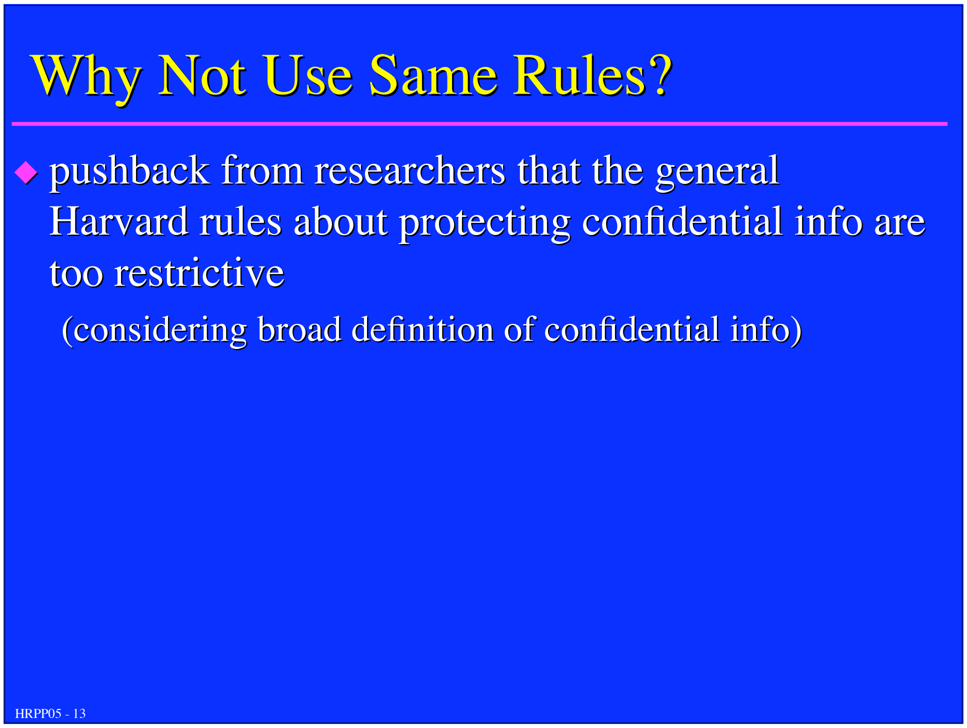
  B/ non-public Harvard info

  e.g., unannounced Harvard plans, etc.



Harvard Rules for Confidential Info.

- secret password, 8 or more characters -some non alpha, no sharing or token-based, lockout on multiple failed attempts
- all CI transport over networks must be encrypted
- no public display, restricted search for other’s data
- good patching procedures, access restricted to need
- host-based firewall, only needed services
- individual switch ports, logs, encrypted access (other than console)
- Internet access only if needed
- CI on computers not at Harvard must be encrypted

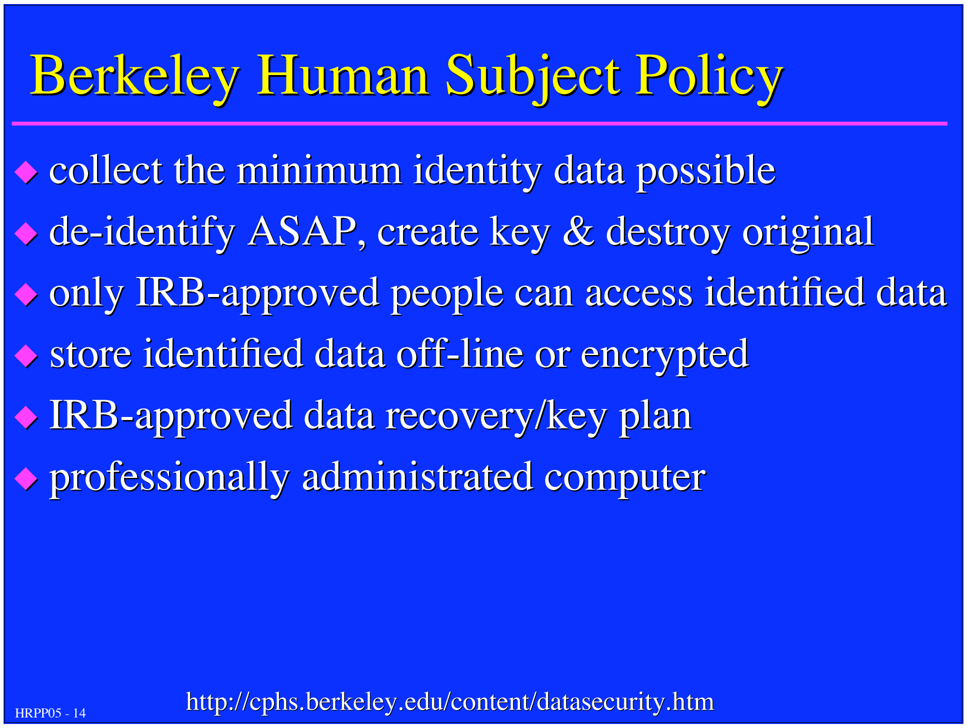


Why Not Use Same Rules?

- pushback from researchers that the general Harvard rules about protecting confidential info are too restrictive

  (considering broad definition of confidential info)

HRPP05 - 13

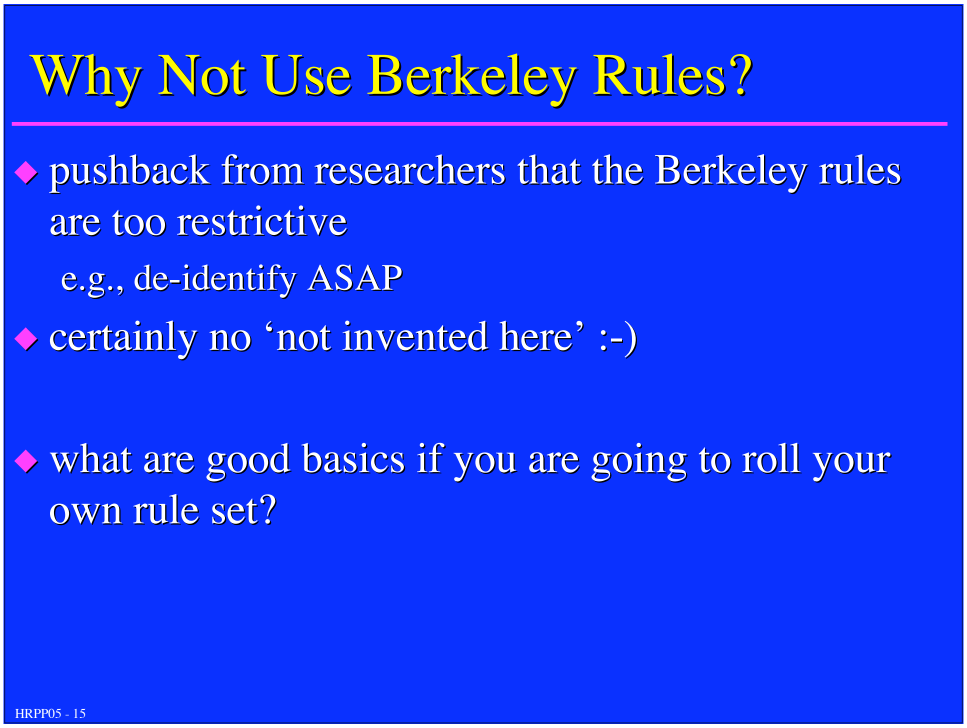


Berkeley Human Subject Policy

- collect the minimum identity data possible
- de-identify ASAP, create key & destroy original
- only IRB-approved people can access identified data
- store identified data off-line or encrypted
- IRB-approved data recovery/key plan
- professionally administrated computer

HRPP05 - 14

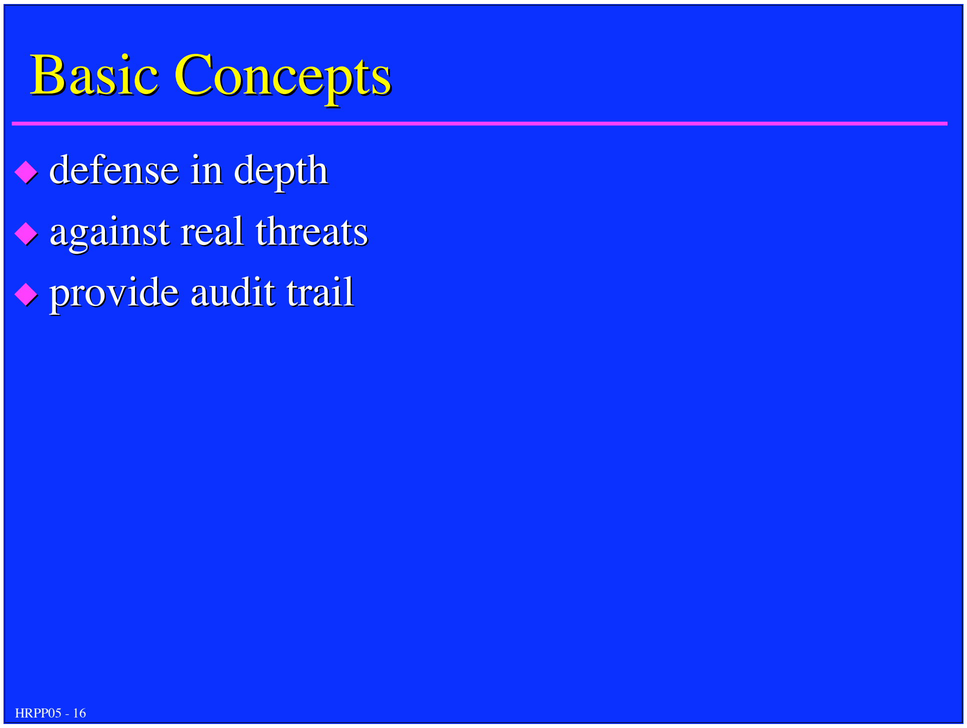
http://cphs.berkeley.edu/content/datasecurity.htm



Why Not Use Berkeley Rules?

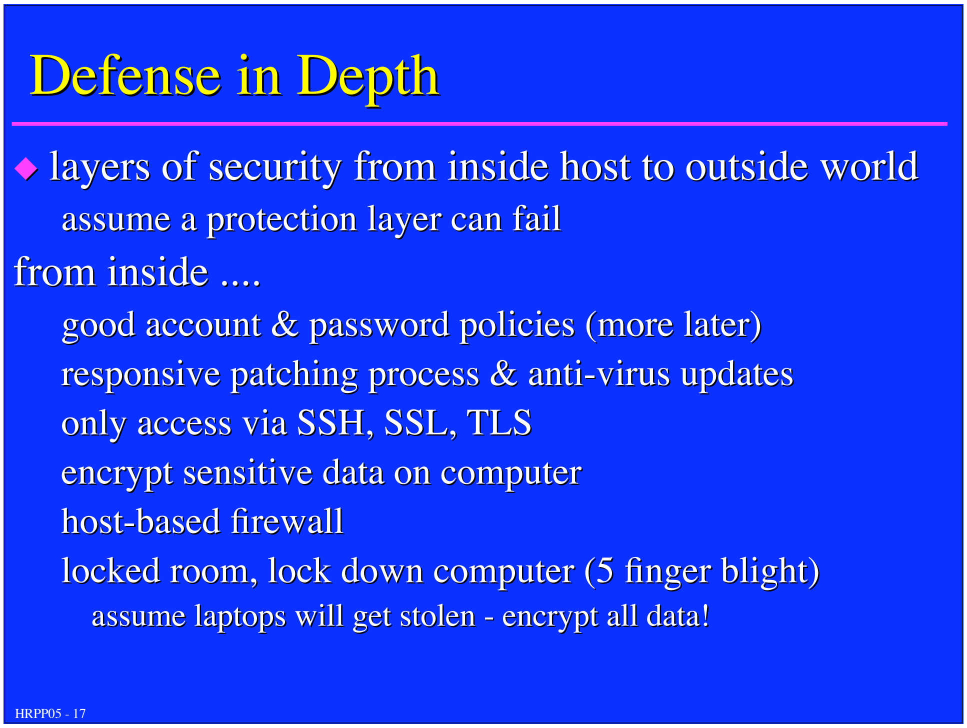
- pushback from researchers that the Berkeley rules are too restrictive
  - e.g., de-identify ASAP
- certainly no ‘not invented here’ :-)

- what are good basics if you are going to roll your own rule set?



Basic Concepts

- defense in depth
- against real threats
- provide audit trail

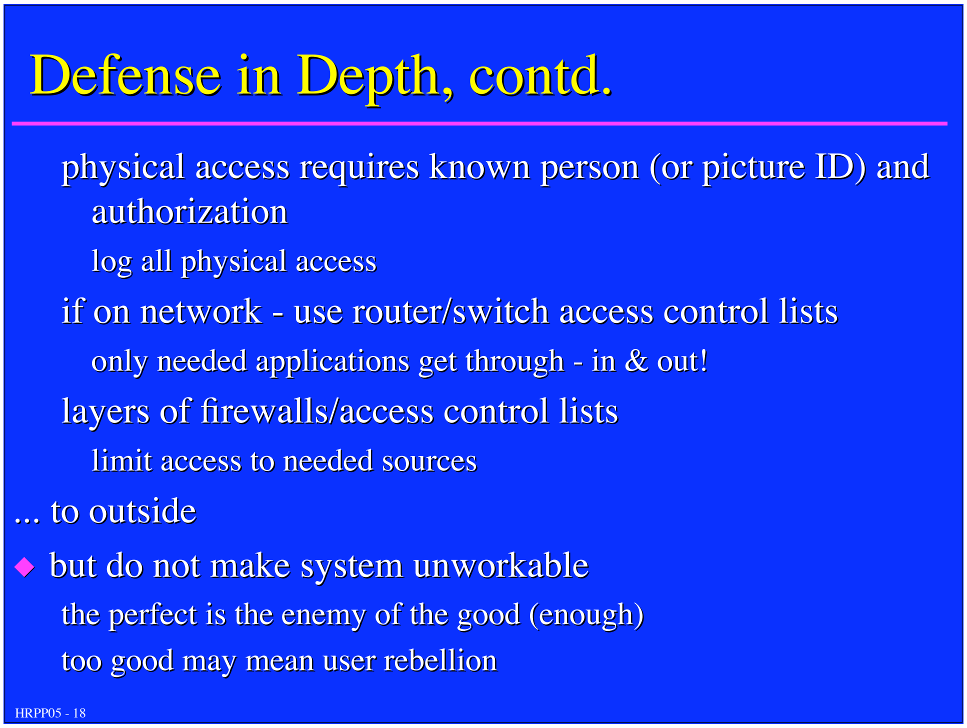


Defense in Depth

- layers of security from inside host to outside world
  - assume a protection layer can fail

from inside ....

- good account & password policies (more later)
- responsive patching process & anti-virus updates
- only access via SSH, SSL, TLS
- encrypt sensitive data on computer
- host-based firewall
- locked room, lock down computer (5 finger blight)
  - assume laptops will get stolen - encrypt all data!

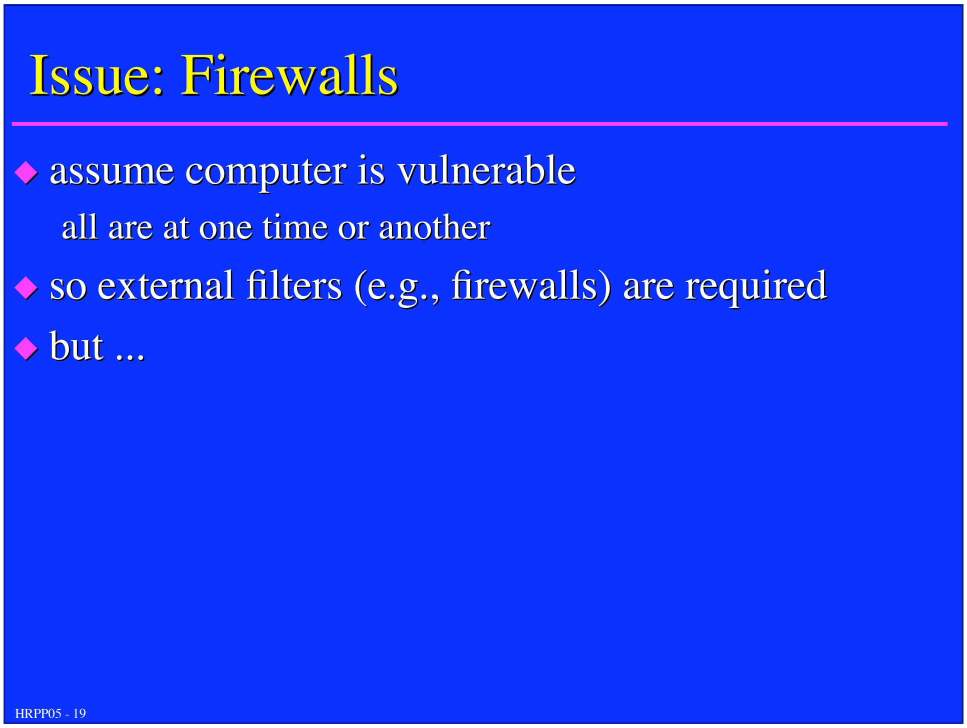


Defense in Depth, contd.

- physical access requires known person (or picture ID) and authorization
  - log all physical access
- if on network - use router/switch access control lists
  - only needed applications get through - in & out!
- layers of firewalls/access control lists
  - limit access to needed sources

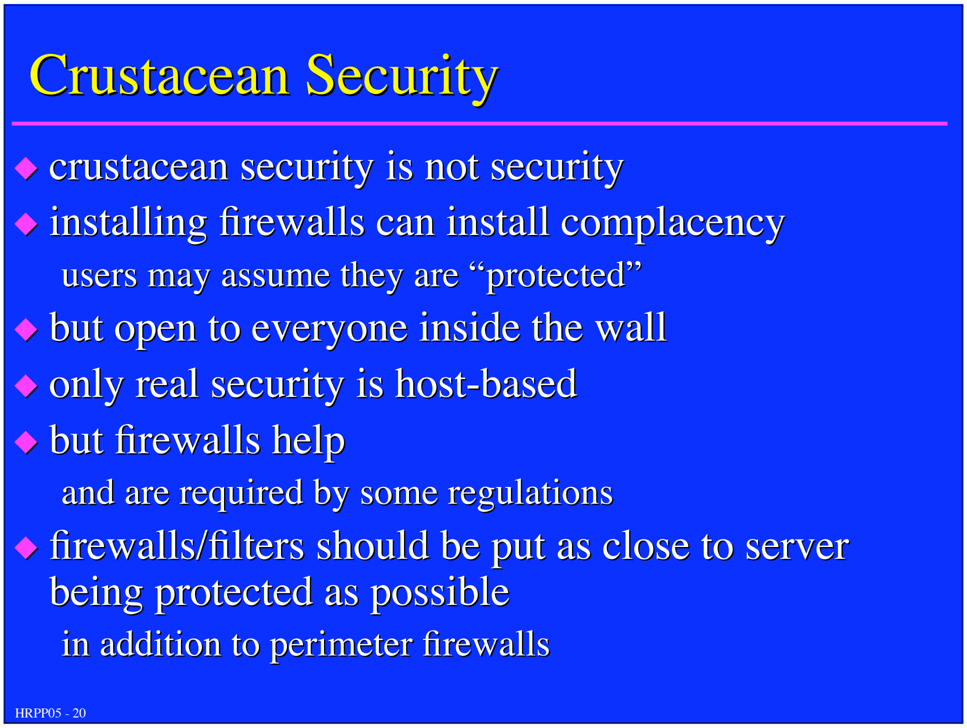
... to outside

- but do not make system unworkable
  - the perfect is the enemy of the good (enough)
  - too good may mean user rebellion



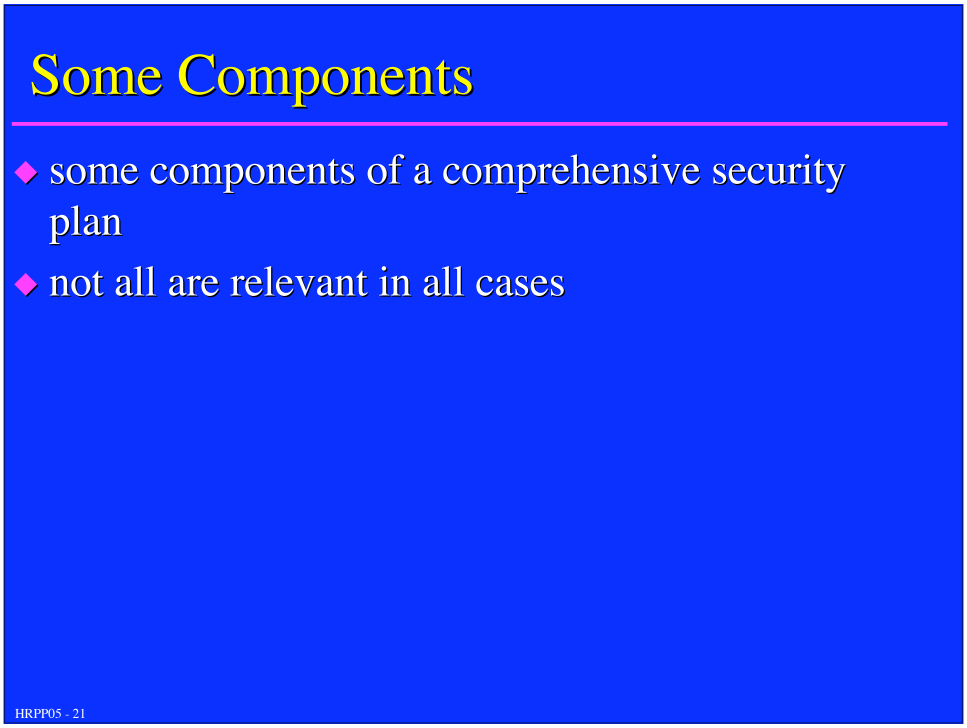
Issue: Firewalls

- assume computer is vulnerable
  - all are at one time or another
- so external filters (e.g., firewalls) are required
- but ...



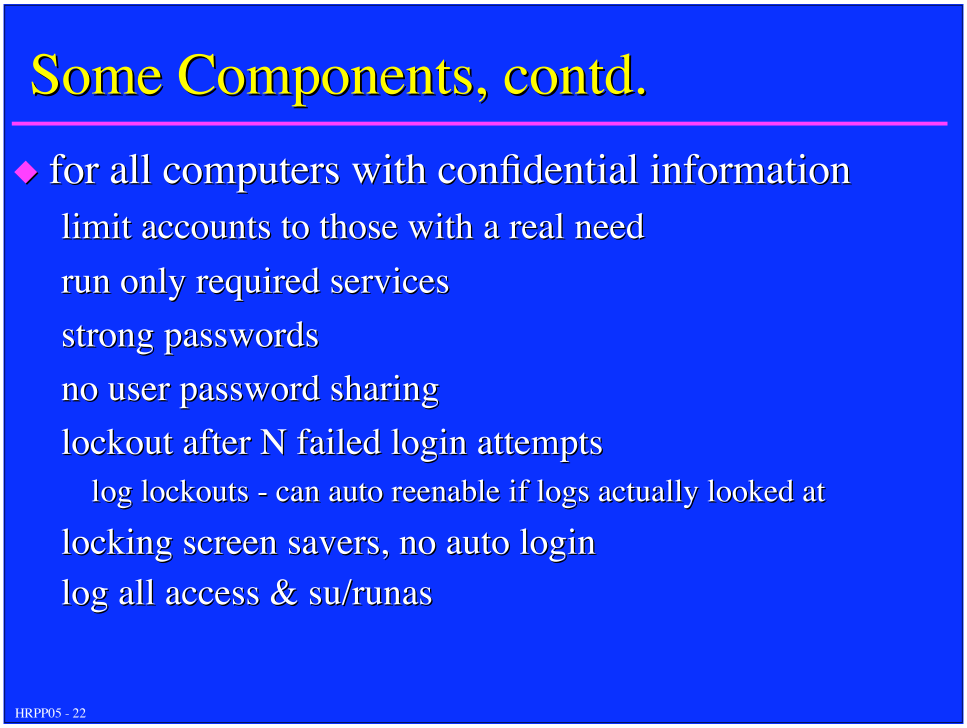
Crustacean Security

- crustacean security is not security
- installing firewalls can install complacency
  - users may assume they are “protected”
- but open to everyone inside the wall
- only real security is host-based
- but firewalls help
  - and are required by some regulations
- firewalls/filters should be put as close to server being protected as possible
  - in addition to perimeter firewalls



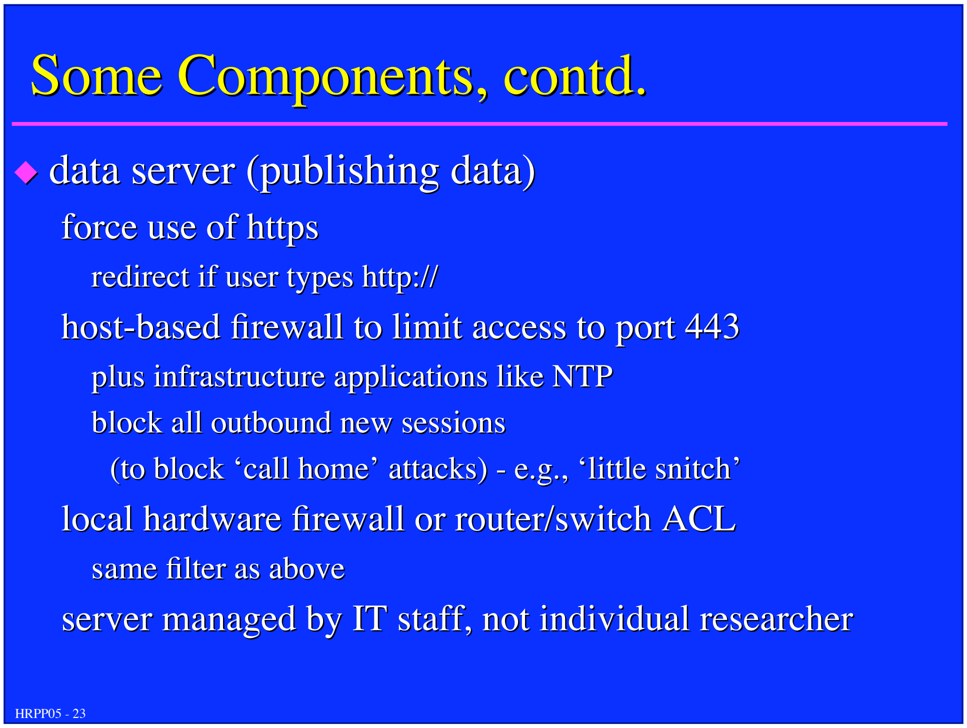
Some Components

- some components of a comprehensive security plan
- not all are relevant in all cases



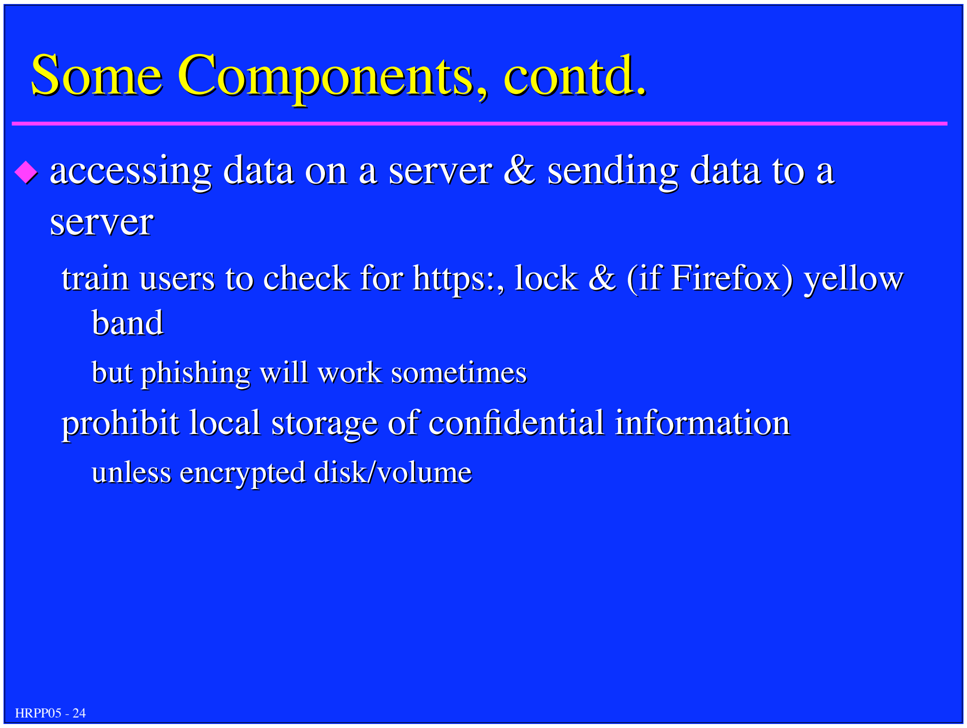
Some Components, contd.

- for all computers with confidential information
  - limit accounts to those with a real need
  - run only required services
  - strong passwords
  - no user password sharing
  - lockout after N failed login attempts
    - log lockouts - can auto reenable if logs actually looked at
  - locking screen savers, no auto login
  - log all access & su/runas



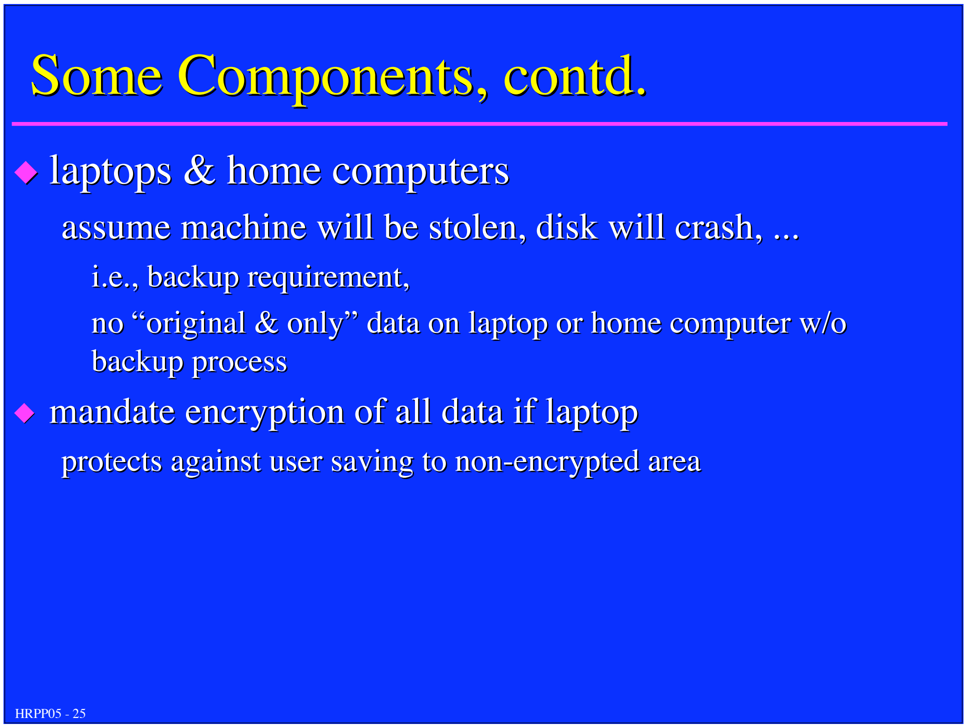
Some Components, contd.

- data server (publishing data)
  - force use of https
    - redirect if user types http://
  - host-based firewall to limit access to port 443
    - plus infrastructure applications like NTP
    - block all outbound new sessions
      - (to block ‘call home’ attacks) - e.g., ‘little snitch’
  - local hardware firewall or router/switch ACL
    - same filter as above
  - server managed by IT staff, not individual researcher



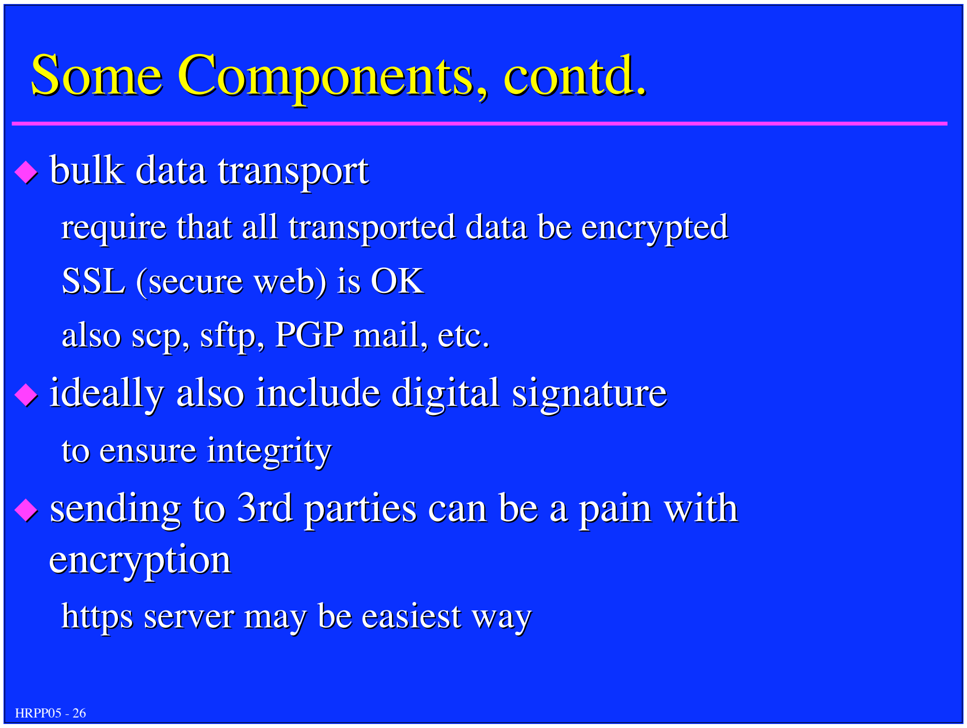
Some Components, contd.

- accessing data on a server & sending data to a server
  - train users to check for https:, lock & (if Firefox) yellow band
    - but phishing will work sometimes
  - prohibit local storage of confidential information
    - unless encrypted disk/volume



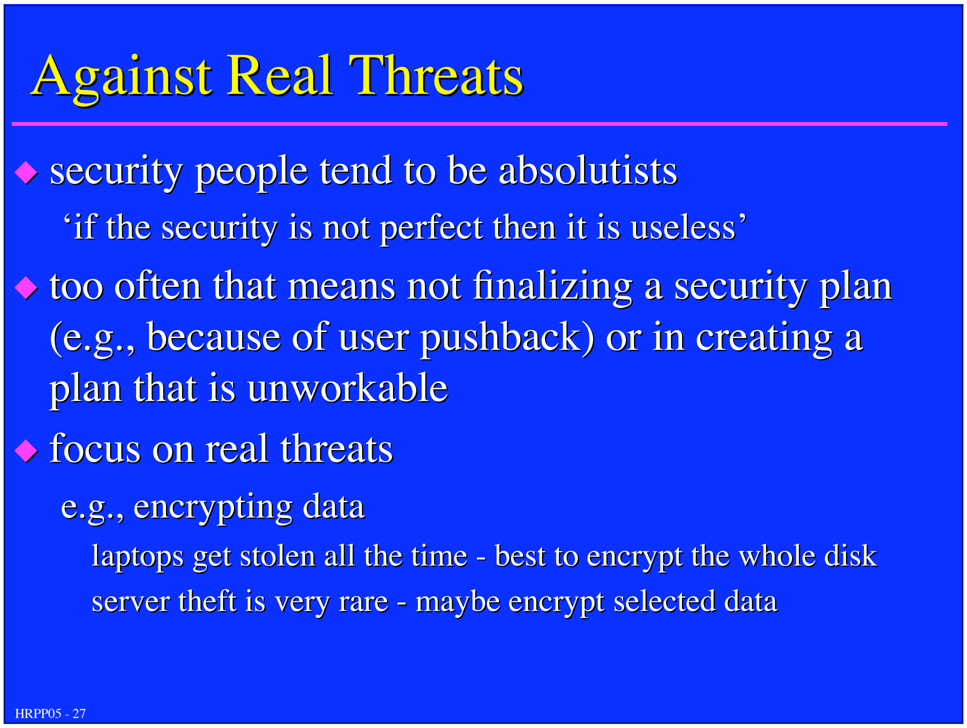
Some Components, contd.

- laptops & home computers
  - assume machine will be stolen, disk will crash, ...
    - i.e., backup requirement,
    - no "original & only" data on laptop or home computer w/o backup process
- mandate encryption of all data if laptop
  - protects against user saving to non-encrypted area



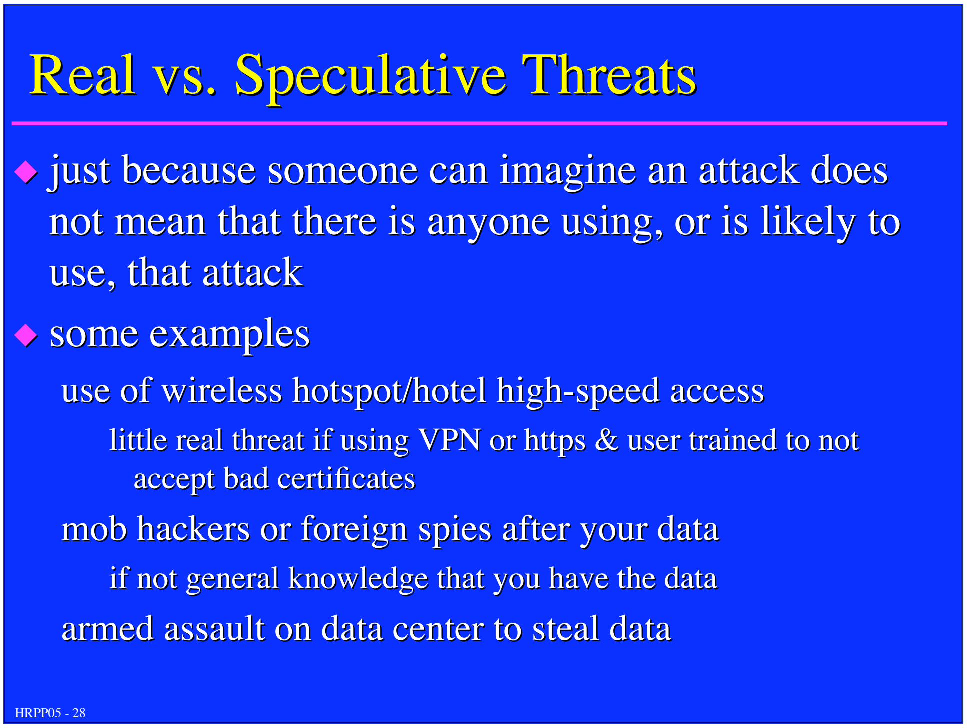
Some Components, contd.

- bulk data transport
  - require that all transported data be encrypted
  - SSL (secure web) is OK
  - also scp, sftp, PGP mail, etc.
- ideally also include digital signature
  - to ensure integrity
- sending to 3rd parties can be a pain with encryption
  - https server may be easiest way



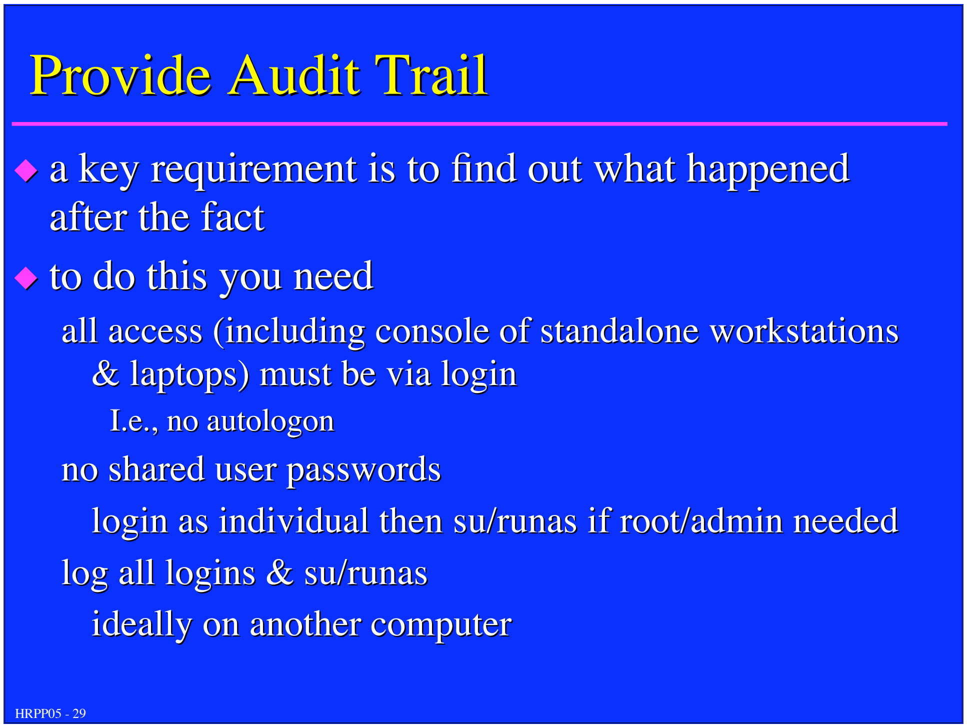
Against Real Threats

- security people tend to be absolutists
  - 'if the security is not perfect then it is useless'
- too often that means not finalizing a security plan (e.g., because of user pushback) or in creating a plan that is unworkable
- focus on real threats
  - e.g., encrypting data
    - laptops get stolen all the time - best to encrypt the whole disk
    - server theft is very rare - maybe encrypt selected data



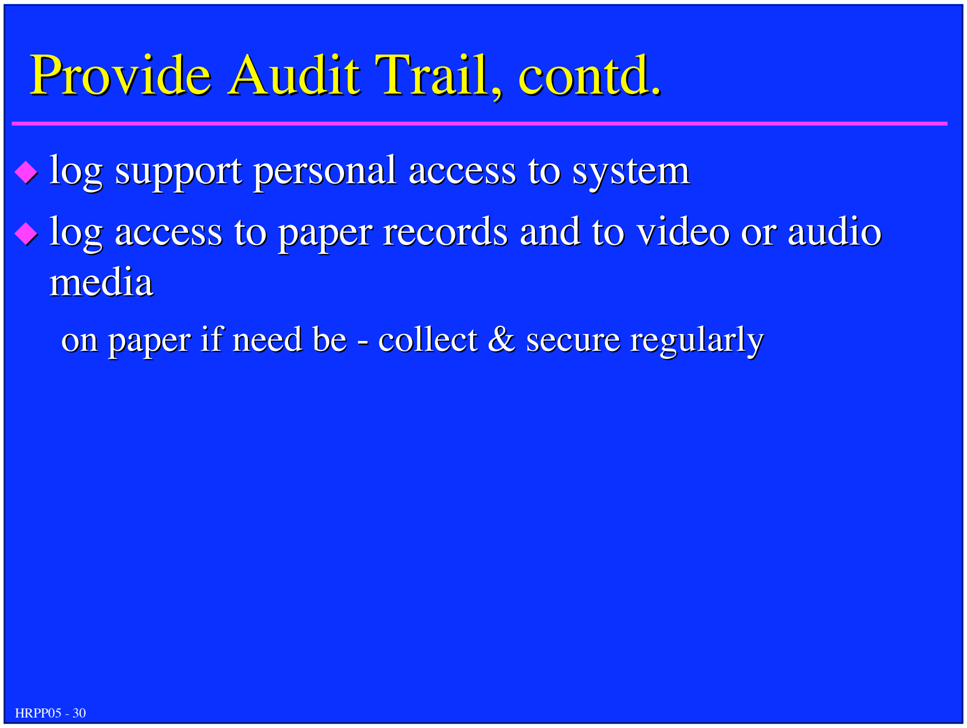
Real vs. Speculative Threats

- just because someone can imagine an attack does not mean that there is anyone using, or is likely to use, that attack
- some examples
  - use of wireless hotspot/hotel high-speed access
    - little real threat if using VPN or https & user trained to not accept bad certificates
  - mob hackers or foreign spies after your data
    - if not general knowledge that you have the data
  - armed assault on data center to steal data



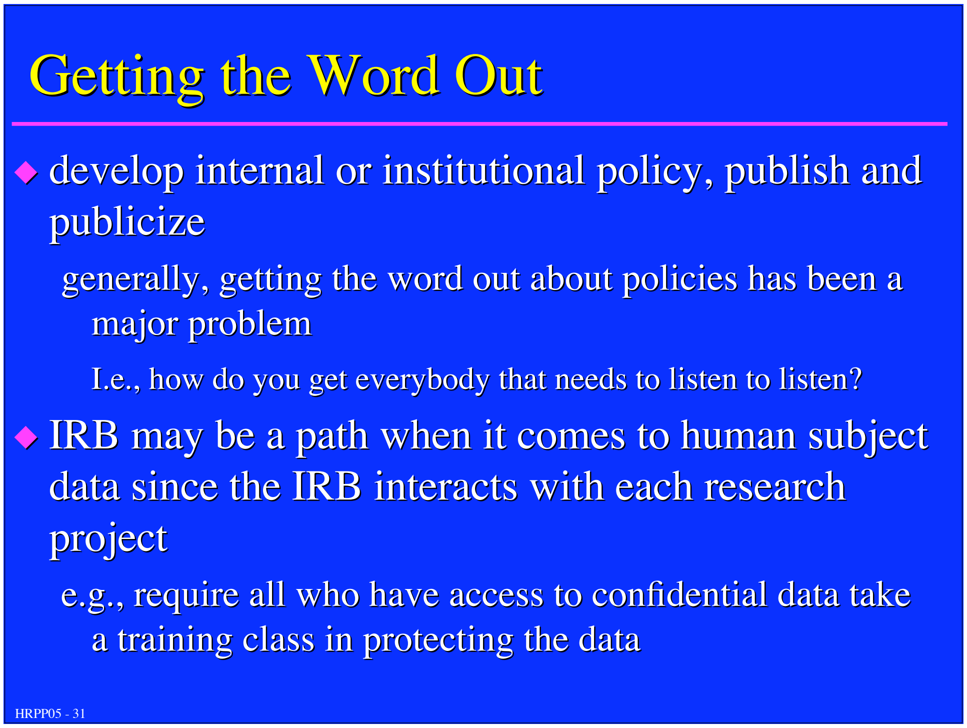
Provide Audit Trail

- a key requirement is to find out what happened after the fact
- to do this you need
  - all access (including console of standalone workstations & laptops) must be via login
    - I.e., no autologon
  - no shared user passwords
    - login as individual then su/runas if root/admin needed
  - log all logins & su/runas
    - ideally on another computer



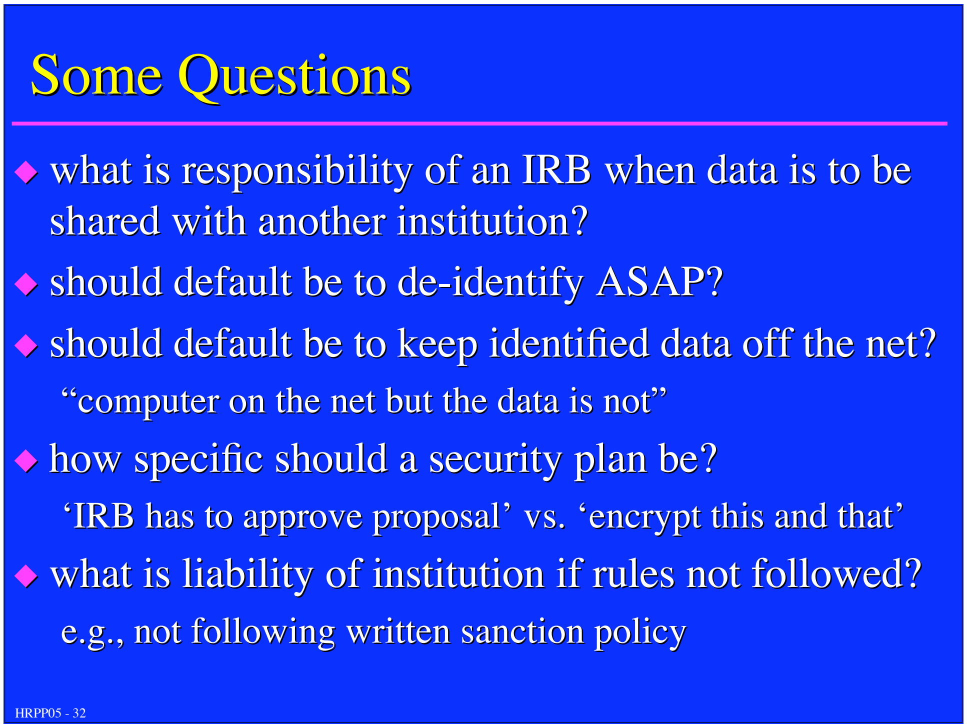
Provide Audit Trail, contd.

- log support personal access to system
- log access to paper records and to video or audio media
  - on paper if need be - collect & secure regularly



Getting the Word Out

- develop internal or institutional policy, publish and publicize
  - generally, getting the word out about policies has been a major problem
    - I.e., how do you get everybody that needs to listen to listen?
- IRB may be a path when it comes to human subject data since the IRB interacts with each research project
  - e.g., require all who have access to confidential data take a training class in protecting the data



Some Questions

- what is responsibility of an IRB when data is to be shared with another institution?
- should default be to de-identify ASAP?
- should default be to keep identified data off the net?
  - “computer on the net but the data is not”
- how specific should a security plan be?
  - ‘IRB has to approve proposal’ vs. ‘encrypt this and that’
- what is liability of institution if rules not followed?
  - e.g., not following written sanction policy