



THE CENTER FOR ADVANCED STUDIES  
IN SCIENCE AND TECHNOLOGY POLICY



## CATEGORIZATION OF INFORMATION FOR DOMESTIC SECURITY: SOME QUESTIONS AND ISSUES

**K. A. TAIPALE**

EXECUTIVE DIRECTOR

CENTER FOR ADVANCED STUDIES

PRESENTED AT:

THE POTOMAC INSTITUTE FOR POLICY STUDIES ROUNDTABLE:

**“WHAT CATEGORIES OF PERSONAL INFORMATION ARE  
RELEVANT TO COUNTERING THE TERRORIST THREAT”**

ARLINGTON, VA • MARCH 30, 2004

# Today's Presentation

- Current proposals and related questions
- Categorization taxonomies and related issues
- Technical issues and required technologies
- CAS Proposal

# Call for Categorization

- Homeland Security Presidential Directive 6 (HSPD-6)
  - “develop, integrate, and maintain thorough, accurate, and current **information about individuals** known or appropriately suspected to be or have been engaged in conduct constituting, in preparation for, in aid of, or **related to terrorism** (Terrorist Information); and (2) use that information as appropriate and to the full extent permitted by law”
  - Subject-based, but what is “related to”
  - Binary, once designated can do anything “permitted by law” (?)

## Call for Categorization, cont.

- Markle Foundation Task Force Second Report
  - Criticizes HSPD-6 as potentially blurring the traditional line between USP/nonUSP (“line at the border”) data without adequate public debate. (pp 18-19) (USSID-18 minimization vs. AG “shall implement appropriate procedures and safeguards”)
  - With respect to **privately held data**, MTRF proposes that government “identify specific categories of private sector information **they need**” (Exhibit A and Working Group II report)
  - Pre-designate particular information or database as related to terrorism **based on scenarios**

## Call for Categorization, cont. II

- Potomac Institute proposal
  - Terrorist Threat Information
    - FI, FCI, LE included “automatically” plus “other”
    - **Other government data** (“ordinary course of business” data) and **privately held data** (publicly available, commercially available, compelled availability) included through executive **pre-approval** and Congressional **oversight**
  - Current Threat Model
    - “to be established, periodically reviewed and approved” and run against TTI
    - A priori? -- “red teaming” (~scenario-based in Markle) vs. data mining
  - US Person data
    - **anonymized** “at first intake from whatever sources”

# Questions about PIPS proposal

- Are TTI and CTM dependent or independent variables? And, why?
  - If information is designated as TTI why is CTM subject to separate approval? In other words, if the data is relevant, why is the query method suspect?
  - On the other hand, if CTM is approved (thus relevant) why can't it be run against all data including non-TTI?
- When/where is “intake” in a distributed architecture?
- How do you determine if data is USP? (collection is not targeted)
- Does CTM include the use of DM?

## Potential information taxonomies for categorization

- Who the data relates to
- Where/how it was collected (or who holds it)
- What kind of information it is
  - Identification
  - Communication
  - Transactional

# Who? (foreigners and terrorists)

- US Person
  - USSID-18 minimization (exception vs. rule problem)
  - Arbitrary nature and doesn't relate to collection (changing mix and commingling)
  - If at "intake" have to treat all data alike
- "Terrorist" or subject based (HSPD-6)
  - Criteria for designation (subject/link/pattern?) (all mosque campers who went to Afghanistan? To Pakistan? On Haj?)
  - Also, begs the question for data mining
- Difficulties
  - A priori nature (what's relevant before its relevant?)
  - Irrelevance of USP/nonUSP to current threat
  - Difficulties of intl. data sharing if different rules for foreigners
  - Derivative data (what nexus req'd for designation and continuation)

## Where? (place and method)

- Traditional doctrines and procedures governing information are based on where and how collected
  - Foreign intelligence (“line at the border”)
  - Intercept vs. stored (Ill v. ECPA) (90 days on server)
  - Third party rule (financial records) vs. statutory protection (cable, video rentals, medical information)
  - Differentiate “privately” held (~public, commercial, compelled availability)
  - Privacy “expectation” based on collection (place)
  - If “legally” collected, free to use for any purpose (~where held)
- Difficult to apply traditional approach to networked databases (information is ‘available’ -- this is not about collection!)
  - Not about where collected, but how and when accessed,
  - And, for what purpose or outcome

# What? (sensitivity of data)

- Information - What is “Personal”
  - Personally identifiable -- links data to an individual (identifiers)
  - Personally sensitive -- relates to health, finance, etc.
  - Socially/politically sensitive -- relates to IstA activity
- Communications - What is “Content”
  - Distinction between subscriber data (who has an account), traffic analysis (the fact that a communication or transaction took place) and content (what was said or transacted) (basic subscriber data, pen register and trap and trace, wiretap content; subpoena/warrant/court order)
- Transactional - What is “Personal” and what is “Content”
  - Where transacted (adult book store) and what purchased
- “Personal” data may be most relevant data (e.g., financial data to find money laundering, etc.)

## Sensitivity of data (Markle)

- Low sensitivity (analyst level control)
  - Type: Non-personally identifiable or *non-US person*
  - Standard: “reasonably related” to HS
  - Process: training, post-facto audit/review, no prior approval required
- Medium Sensitivity (administrative level control)
  - Personally identifiable information generally available
  - Specifically identifiable facts “relevant to counterterrorism”
  - Administrative procedure (senior official sign-off)
- High Sensitivity (judicial control)
  - Private, personally identifiable generally not available, all FA
  - “necessary to obtain valuable intelligence information related to a threat to the U.S.”
  - FISA-like process

## By method of query?

- “Current threat model”
  - Static approval and oversight vs. dynamic needs
- Data analysis
  - Subject-based query
  - Link-based query
  - Pattern analysis
  - Pattern matching

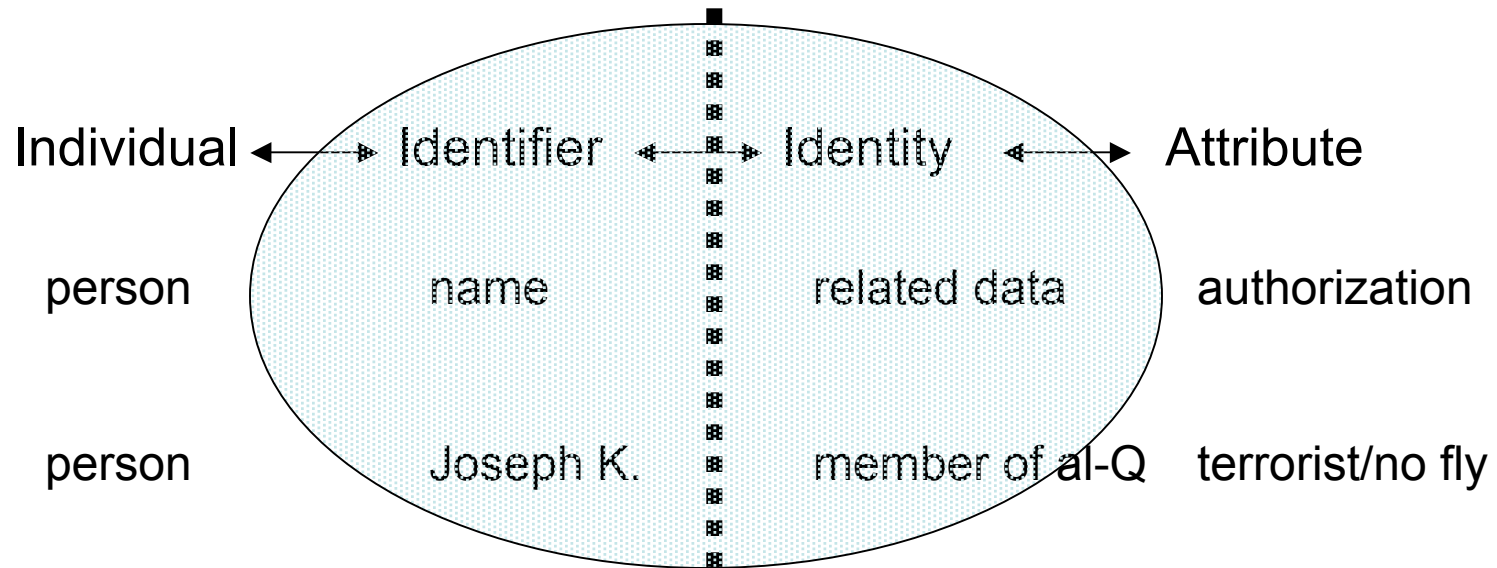
## By purpose of query?

- Purpose of data attribution
  - ID > data dossier (identification -- you know the target and want to find out more)
  - Attribute data > ID (you observe the attribute (or activity) and want to identify the actor)
  - Attribute profiling (you model certain attributes and look for like occurrences to identify additional activity or targets, ~DM)
- Examining data patterns vs data items
  - Social network theory (flow of data evidences organization)

## Vary by dynamic threat environment?

- Flexible system
- Security should be adaptable
  - Not always at ease
  - Not always at attention
- But does perceived threat relate to actual threat?

# The privacy divide and intrusive technologies



Anonymization

Data Mining

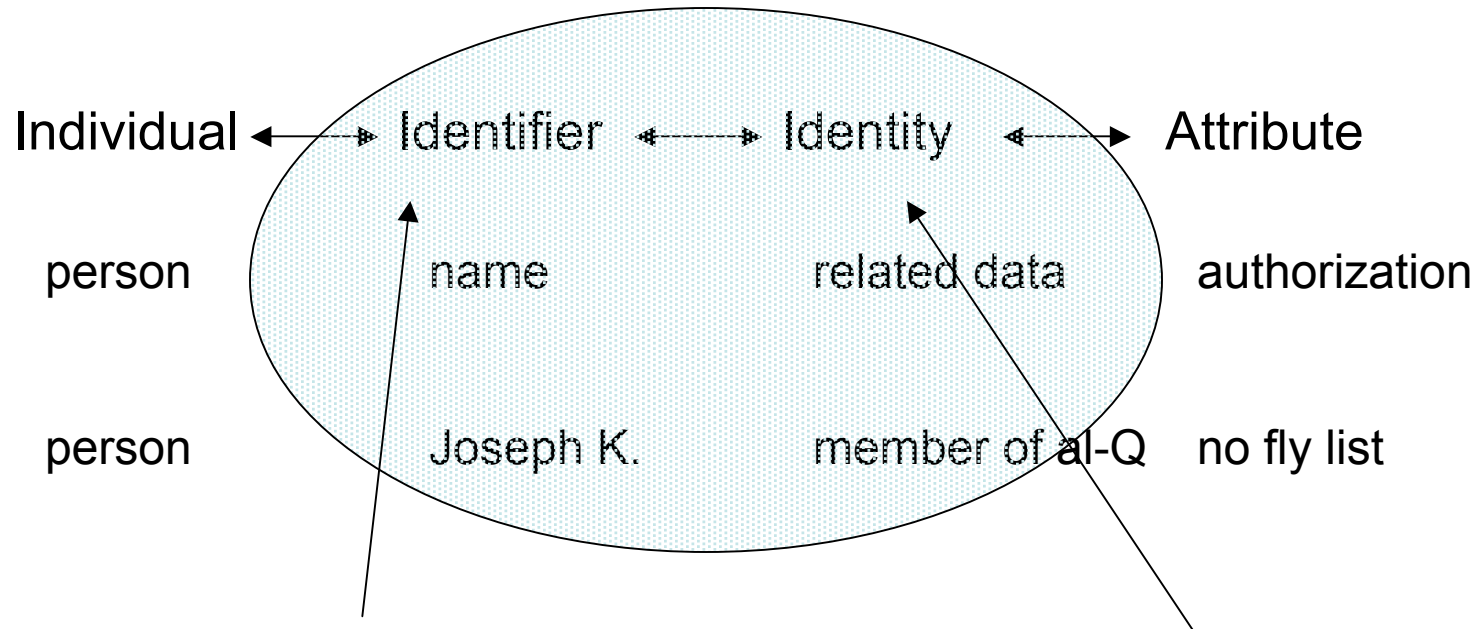
Identification

Pseudonymization

Collection

Collection

# Data Categorization -- additional complexities



Multiple Identifiers:  
Name, SS#, Driv Lic #, arbitrary, etc.  
As well as aliases

Multiple Identities:  
Professional, family, social, etc.

**How does a particular DB, data source or data use relate to this process?**

# Technical issues

- Arbitrary nature of designation
  - Policy attribute vs. data attribute (protocols)
  - Relationship of attribute to transaction/query
- Commingling of data collected for different (but specific) needs
  - Collection policy/need doesn't match use categories
- Derivative and legacy data
- What is “intake” in a distributed architecture (pre-processing?)

# Design requirements

- Dynamic rules-based processing (build policy rules into processing)
- Selective revelation or selective attribution (anonymization and pseudonymization)
- Credential and audit (watch the watchers)

# Paradigm shift

- Cyber Security analogy
- Historical migration of security locus
  - Security in WAN  
("smart" closed network vs. open end-to-end IP architecture)
  - Security in LAN (firewall)
  - Security in application (current)
  - Security in data (predicted)

# Privacy

- Privacy
  - Control in network (e.g., credit alerts from agency)
  - Control in database (e.g., credit alerts from DB)
  - Control in data (e.g., self reporting data)
- Privacy  $\equiv$  DRM (control in network, control in device, control in data)
- Smart data
  - Develop technologies to make data “responsible” for its own processing

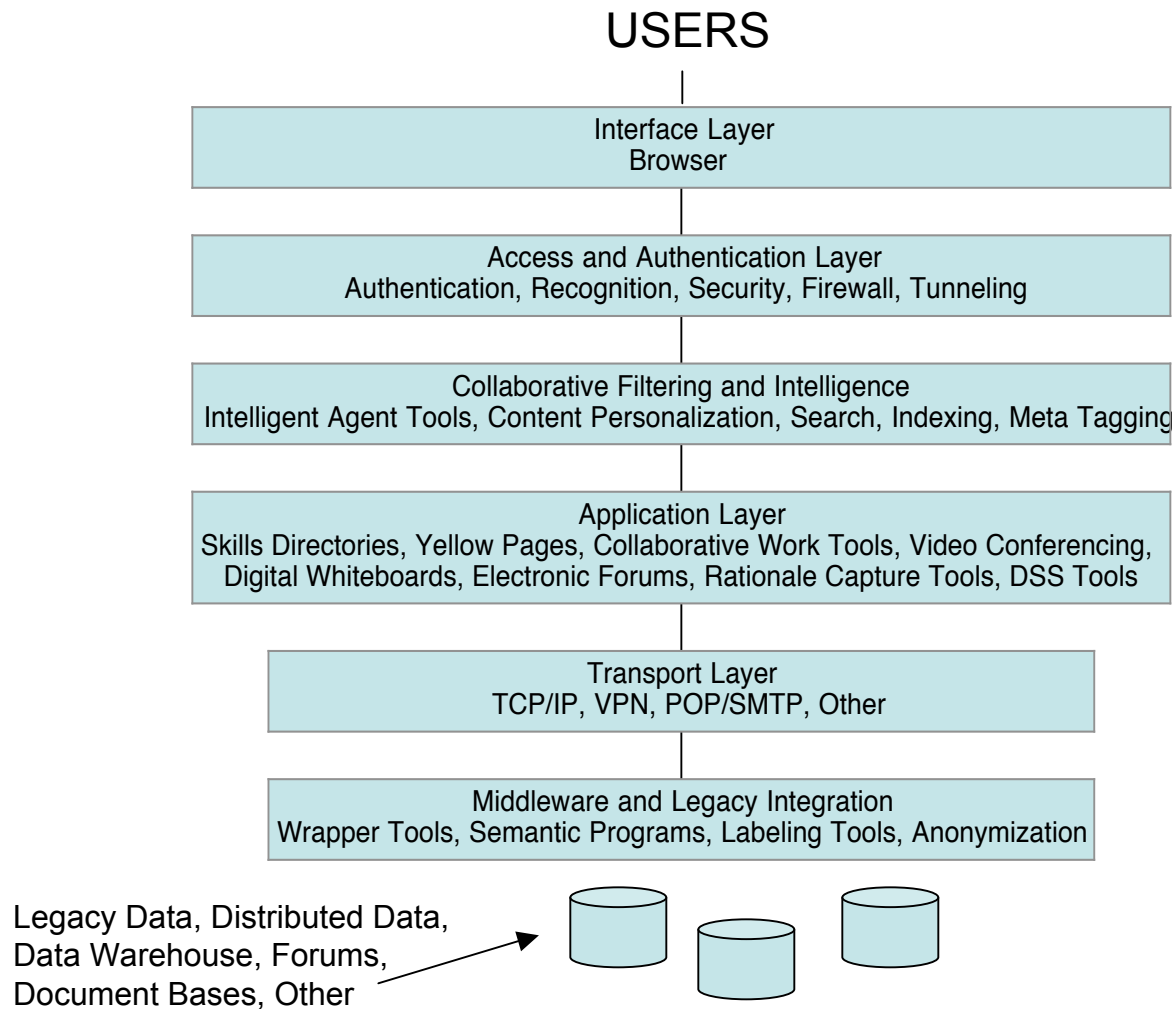
# Smart data technologies

- Labeling (meta-data)
- Wrappers (encrypted)
- Proof carrying code
- Self reporting data
- ???

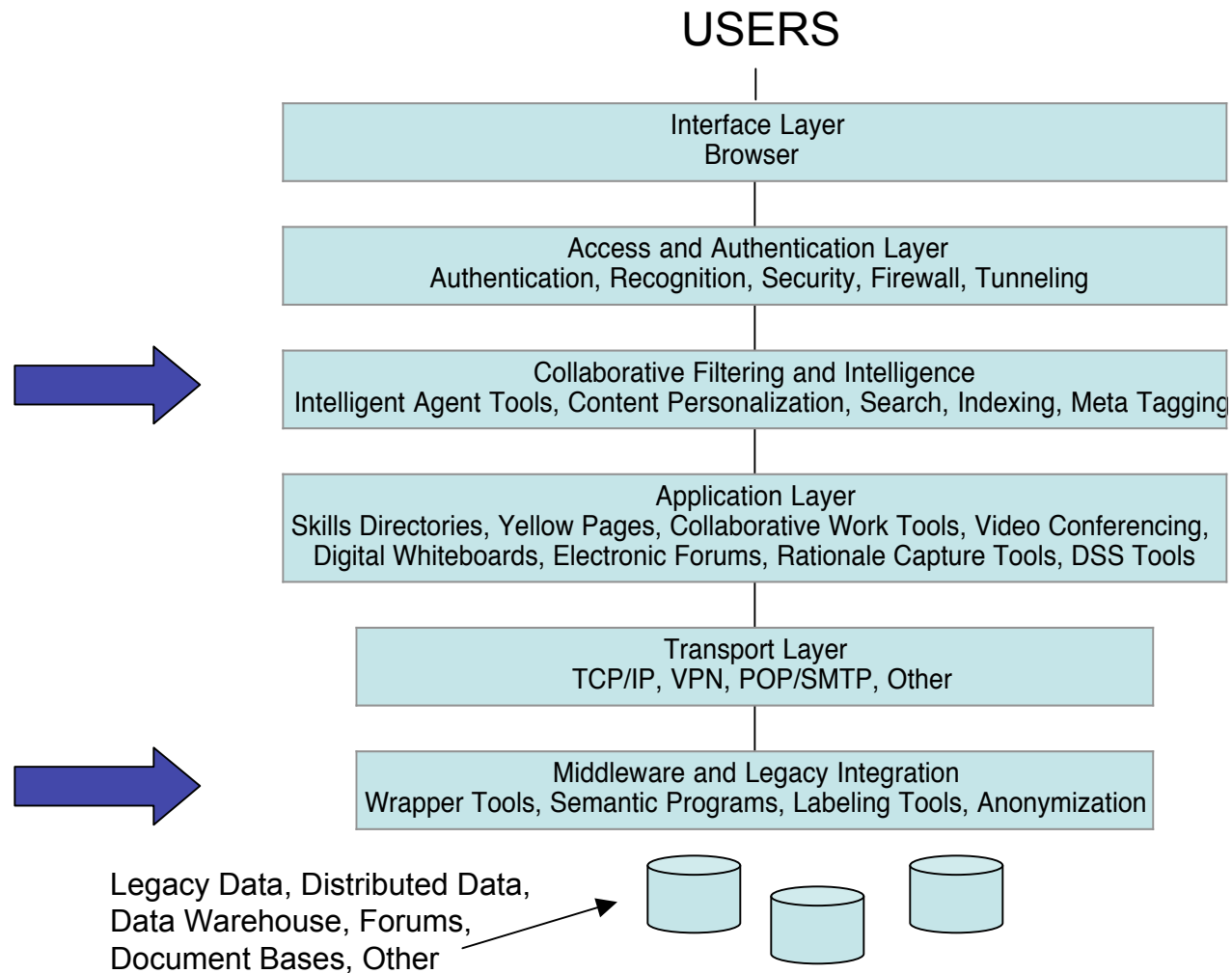
# Smart data systems

- Common language (privacy protocols)
- Analytic filtering
  - At DB level for remote incoming queries
  - At agent level for outgoing queries (or at return/response to outgoing queries?)
- Program semantics
  - legacy data, derivative data
  - Contextualize use “on the fly” to categorize data to correspond to query (evaluate query in real-time context and inform data object)

# Overall architecture and intervention



# Overall architecture and intervention



# CAS Proposal

- Use refined PIPS-like proposal for existing agencies but create a new, limited charter information analysis agency focused on countering “foreign inspired organized but stateless actors intent on using violence against US interests” (~“terrorists”)
- No LE powers or operations directorate, must refer out for action, thus administrative control
- Strict oversight with external logging (Congress) (FOIA on query?)
- Online FISA-like judicial approval for certain access and methods -- must be real-time and dynamic (use Markle proposal standards)
- Build in technical constraints to enforce policy (e.g., “warrant token” required to access certain data)

# Conclusion

- Lots of work to do
- First step is to develop a policy language so that technical requirements can be specified