# Taxonomy of Security and Privacy Topics

In the proposed taxonomy, broad considerations of privacy, provenance and systems health appear as recurring features. For instance, privacy of communications applies to governance of data at rest, access management, but it is also part of a security metadata model.

Ed Note: Consider following some schemes in the NIST Preliminary Critical Infrastructure [Cybersecurity Framework](http://www.nist.gov/itl/upload/preliminary-cybersecurity-framework.pdf) (CIICF) of October2013. A taxonomy is implicit, though the framework contains a reference to the lack of taxonomies of big data (see Section C.5 p.41).

The taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data.

Section References: SAML (2005), Security Token Service (WS-Trust STS), CERT Taxonomy of Operational Cybersecurity Risks (2010).

## Registration, Security Model and Policy Enforcement

### Device, User, Asset, Services, Applications registration

Includes registration of devices (M2M (Internet of Things), DRM-managed assets, services, applications, user roles

### Security Metadata Model

The metadata model maintains relationships across all elements of a secured system. It maintains linkages across all underlying repositories. Big data often needs this added complexity due to its longer lifecycle, broader user community or other aspects.

A Big Data model must address aspects such as data velocity, and temporal aspects of both data and the life cycle of components in the security model.

### Policy Enforcement

#### Environment build

#### Deployment policy enforcement

#### Governance Model

#### Granular policy audit

#### Role-specific behavioral profiling

## Identity and Access Management

### Virtualization Layer Identity (e.g. Cloud Console, PaaS)

#### Trusted Platforms

### Application layer Identity

### End user layer identity management

#### Roles

### Identity Provider (IdP)

An identity provider is defined in SAML (2005). In a Big Data ecosystem of Data Providers, Orchestrators, Resource Providers, Framework Providers and Data Consumers, a scheme such as the SAML / STS or XACML is seen as one helpful – but not proscriptive -- way to decompose the elements in the security taxonomy.

Big Data may have *multiple* identity providers (IdP’s). An IdP may issue identities (and roles) to access data from a Resource Provider. In the SAML framework, trust is shared via SAML/WS mechanisms at the registration phase.

In Big Data, due to the density of the data, the user “roams” to data (whereas in conventional VPN-style scenarios, users “roam” across trust boundaries). Hence the conventional Auth/AuthZ model needs to be extended since the Relying Party is no longer fully trusted because they are custodians of somebody else¹s data. Data is potentially aggregated from multiple Resource Providers.

One approach is to extend the claims-based methods of SAML to add security and privacy guarantees.

### Additional XACML Concepts

XACML introduces additional concepts, some version of which may be useful for Big Data security. In Big Data, parties are not just sharing claims, but also sharing policies about what is authorized. There is a Policy Access Point at every data ownership and authoring location, and a Policy Enforcement Point at the data access. A Policy Enforcement Point calls a designated Policy Decision Point for an auditable decision. The usual meaning of non-repudiation and trusted third parties is extended in XACML in this way. Big Data presumes an abundance of Policies, “points,” identity issuers as well as data.

#### Policy Authoring Points

#### Policy Decision Points

#### Policy Enforcement Point

#### Policy Access Points

## Data Governance

### Encryption and Key management (Including Multi Key)

#### At Rest

#### In Memory

#### In Transit

#### New: Use case of privacy

### Isolation/ Containerization

### Storage Security

### Data Loss Prevention, Detection

### Web Services Gateway

### Data Transformation

#### Aggregated Data Management

##### Authenticated computations on aggregated data

Example: use of homomorphic keys

### Data Life Cycle Management

#### Disposition, migration, retention policies

#### PII microdata as “hazardous” (Kum et al. 2013)

#### De-identification and Anonymization

#### Re-identification risk management

### End Point Validation

### Digital Rights Management

## Visibility and Infrastructure Management

### Threat and Vulnerability Management

#### DoS-resistant cryptographic protocols

### Monitoring, Alerting

As pointed out in the CIICF, Big Data affords new opportunities for large scale security intelligence, complex event fusion, analytics and monitoring.

### Mitigation

Breach mitigation planning for Big Data may be qualitatively or quantitatively different.

### Configuration Management

Configuration management is one aspect of preserving system and data integrity

Patch Management

Upgrades

### Logging

Big Data must produce and manage more logs, of greater diversity and velocity. E.g., profiling and Statistical sampling on an ongoing basis may be required.

### Malware Surveillance and Remediation

This is a well-understood domain, but Big Data can cross traditional system ownership boundaries. Review of Identify, Protect, Detect, Respond and Recover framework (NIST Oct 2013) may uncover planning unique to Big Data.

### Network Boundary control

NEW: Establishes a data-agnostic connection for a secure channel.

#### Shared services network architecture

Ed Note: Akhil to expound.

#### Zones/ cloud network design (including connectivity)

### Resilience, Redundancy and Recovery

#### Resilience

The security apparatus for a Big Data system may be comparatively fragile in comparison to other systems.

#### Recovery

Recovery for Big Data security failures may require considerable advance provisioning beyond that required for small data. Response planning and communications with users and may be on a similarly big scale.

## Risk and Accountability

### Accountability

Information, process and role behavior accountability can be achieved through various means.

#### Transparency Portals and Inspection Points

#### Forward- and reverse-provenance inspection

### Compliance

Big Data compliance can span multiple aspects of the security and privacy taxonomy including privacy, reporting, nation-specific law.

### Forensics

#### Forensics techniques enabled by Big Data

#### Forensics used in Big Data security failure scenarios

### Business Risk Level

Big Data risk assessments should be mapped to each element of the taxonomy. (For an example, see Appendix B of CERT “Taxonomy of Operational Cyber Security Risks, December 2010). Business risk models can incorporate privacy considerations.

# IT, GRC and IW “Axes”

# Overview

Documents for review on Big Data Security need to be accessible to a diverse audience that could include individuals that specialize in cryptography, security, compliance, and information technology. In addition, there are domain experts and corporate decision makers that need to understand the costs and impact of these controls.

Ideally these documents would be prefaced by information that would help any specialist to chart their own path through content that they need to read to obtain context that they can subsequently use to provide feedback on sections that they are specialized in.

Organizations live up to their names and are typically organic compositions of hierarchies and relationships, and contain diverse roles and workflows for participating in a Big Data ecosystem. Hence this document proposes a pattern to help identify the ‘axis’ of an individual’s roles and responsibilities, and then classify the security controls in a similar manner to make these more accessible to each class.

# Infrastructure Technology (IT)

Typically this axis contains individuals and groups within participatory organizations that are responsible for technical reviews before organizations are on-boarded in that data ecosystem, after which they are responsible for the care and feed that includes addressing defects and security issues.

When IT personnel work across organizational boundaries they have to accommodate diverse technologies, infrastructures and workflows that need to be integrated. For Big Data Security these include identity, authorization, access control, and log aggregation.

Their backgrounds, practices and the terminologies that they use tend to be uniform, and they face similar pressures within their organizations to constantly do more with less. “Save Money” is the underlying theme, and IT is usually on the hot seat when things go wrong.

# Governance, Risk Management and Compliance (GRC)

Typically this is a function that draws participation from multiple functions within organizations such as Legal, HR, IT and Compliance. But increasingly these are departments with their own heads. There tends to be a strong focus on Compliance, often in isolation from technologies.

Similar to IT, GRC tend to have uniform backgrounds and leverage a common terminology, and they tend to have similar processes and workflows within a vertical, which typically has marquees that influence other organizations within that vertical or sector.

GRC within an organization is on the hot seat to protect the company from risks that might arise from loss of intellectual property, legal risks due to actions by individuals within that organization, and compliance risks specific to their vertical. “Stay out of Jail” is a quaint but reasonable way to describe their underlying theme, and GRC is also on the hot seat to prevent, then preserve and protect.

# Information Worker (IW)

These are the individuals and groups that are actually operating on the content that spans generation, transformation and consumption. Due to the nascent nature of the technologies and related businesses, they tend to use common terms at the technical level, but their roles, responsibilities and the related workflows do not always align across organizational boundaries.

For instance, a Data Scientist has deep specialization in the content and its transformation, but typically will care about security and cryptography when it adds friction to their ability to transfer, or access relevant information.

IWs are being assaulted by an avalanche of products and services, and under pressure from their organizations to deliver concrete business value from these new Big Data Analytics capabilities, by either monetizing available data, or monetizing the capability to transform through becoming a service provider, or optimizing and enhancing business by consuming third party data.

# Enhancing the Taxonomy of Security and Privacy Topics

To leverage these three axes to facilitate collaboration and education, a Stake Holder would be defined to be some individual or group within an organization that would be directly impacted by the selection and deployment of a Big Data solution, and a Ratifier would be some individual or group within an organization that is tasked with assessing the candidate solution before it is selected and deployed.

For example, a third party security consultant may be deployed by an organization as a Ratifier, and an internal security specialist with an organization’s IT might serve as both a Ratifier and a Stake Holder if they are tasked with ongoing monitoring, maintenance and audits of the security.

The next sections cover the three current components of the taxonomy, which are Privacy, Provenance and System Health. We also explore potential gaps that would be of interest to our anticipated Stake Holders and Ratifiers that reside on these three new conceptual axes.

# Data Privacy

IT specialists that address cryptography need a deep understanding of the definitions, the threat models, the assumptions, the security guarantees, and the core algorithms and protocols. It is likely that they will be mostly Ratifiers, and not Stake Holders.

IT specialists that address end-to-end security need to get an abbreviated view of the cryptography, and in addition they need to obtain a deep understanding of how this would integrate with their existing security infrastructures and controls.

GRC would need to reconcile the vertical requirements, such as perhaps HIPAA requirements related to electronic health records, and the assessments by the Ratifiers that address cryptography and security, and would in turn be Ratifiers to communicate their interpretation of the needs of their vertical. They would also be Stake Holders due to their participation in internal and external audits and other workflows.

# Data Provenance

This is similar to the previous one, but might introduce IWs as Ratifiers since there might be a business need to ensure that the organization’s intellectual property is suitably protected from direct leakage, or indirectly exposed during subsequent big data analytics. They would need to work with the Ratifiers from cryptography and security to convey the business need, and understand how the available controls may apply.

Similarly when an organization is obtaining and consuming data, the IWs may have a need to articulate their business need to ensure they can maximize the value and reduce their risk by ensuring that the data provenance guarantees some degree of information integrity that addresses data that is incorrect, fabricated, or cloned before presented to this organization.

There could also be GRC risks to that organization if the supplier of that data hasn’t implemented the appropriate degree of care in filtering or labeling that data. For example, the organization may not have a signed Business Associate Agreement in place, and the organization’s GRCs interpretation of the HIPAA Omnibus Rule might indicate that they could be at risk if the supplier has access to EHRs and presumably has signed a BAA.

# System Health

This is typically the domain of IT, and they will be Ratifiers and Stake Holders of technologies, protocols and products that are used for system health, and designing how the responsibilities would be shared across the partners, an area that is commonly known as OSS in the Telecom Industry, which has significant experience in syndication of services.

There are aspects of system health that would need to be scrutinized by Security and Cryptography specialists to ensure that there are no gaps in the operational architectures that may have been stitched together from diverse technologies and products, as ecosystem participants integrate their system infrastructures.

# White Spaces

There are additional areas that have not been carefully scrutinized, and it is not clear that these would fold into the existing categories, or if we would need to identify and showcase new ones. Here are a few candidates.

## Provisioning, Metering and Billing

The degree to which commercial pipelines for Big Data can be constructed and monetized will be limited by the ability to be agile in offering services, metering access suitably, and integrating with billing systems. While this can be manual for a small number of participants, this gets complex very quickly when there are many suppliers, consumers and service providers. IWs and IT that are involved with existing business process would be candidate Ratifiers and Stake Holders.

## Data Syndication

Similar to Service Syndication, any data ecosystem is most effective and valuable if any participant can have multiple roles that might include supplying, transforming and consuming Big Data. Therefore there is a need to think through the Data Syndication models that would need to be enabled, and IWs and IT would once again be candidate Ratifiers and Stake Holders.

# Summary

This is a draft document for your consideration, and we would like to see where the three axes would break down in practice. There is no immediate plan to restructure the working document, to align the actors and roles with this classification, but there might be an opportunity to introduce a preamble that would showcase our understanding of the axes and constituent roles, and the manner in which our readers could best navigate through the existing document and provide us with feedback.