Security Issues in Big Data

CISC 6640 Privacy and Security in Big Data
Lecture 1b

Instructor:
Md Zakirul Alam Bhuiyan
Assistant Professor
Department of Computer and Information Sciences
Fordham University

Review Quiz

- What is Big Data?
- What are characteristics of Big Data?
- What can we do with The Data?
- What do we need for the growth of big data?
- How is big data different?
- What Technology Do We Have For Big Data ??
- What is difference between HDFS and Mapreduce?

What We Are Going to Learn

- Security?
- Security in Big Data--The Perfect Storm
- What is the Cost of A Security Breach?
- Balancing Security and Data Insight
- Security Solution is on the Way
- Data Security

Security?

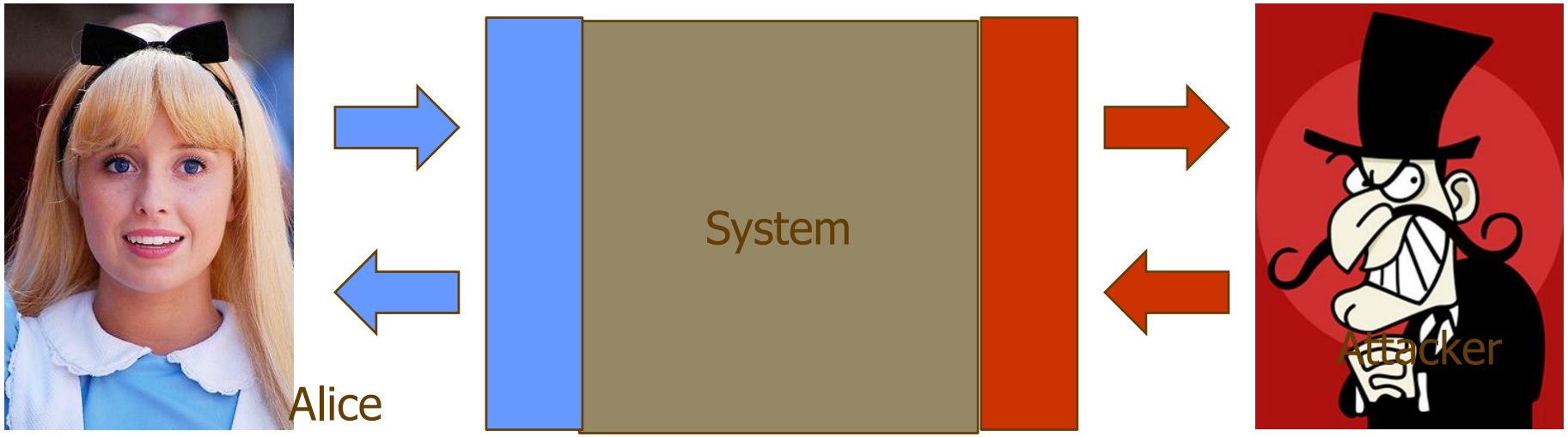
- **Security**
  - **System correctness**
    - If user supplies expected input data, system generates desired output data
  - **Security**
    - If attacker supplies unexpected input data, system does not fail in certain ways, but produce undesired output data

Security?

- **Security**
  - **System correctness**
    - **Good input => Good output**
  - **Security**
    - **Good input ≠> Good output**

Not good for you

Security: General Picture



Security is about

- Honest user (e.g., Alice, Bob, …)
- Dishonest user (Attacker)
- How the Attacker
  - Disrupts honest user's use of the system (Integrity, Availability)
  - Learns information intended for Alice only (Confidentiality)

Security: Definition

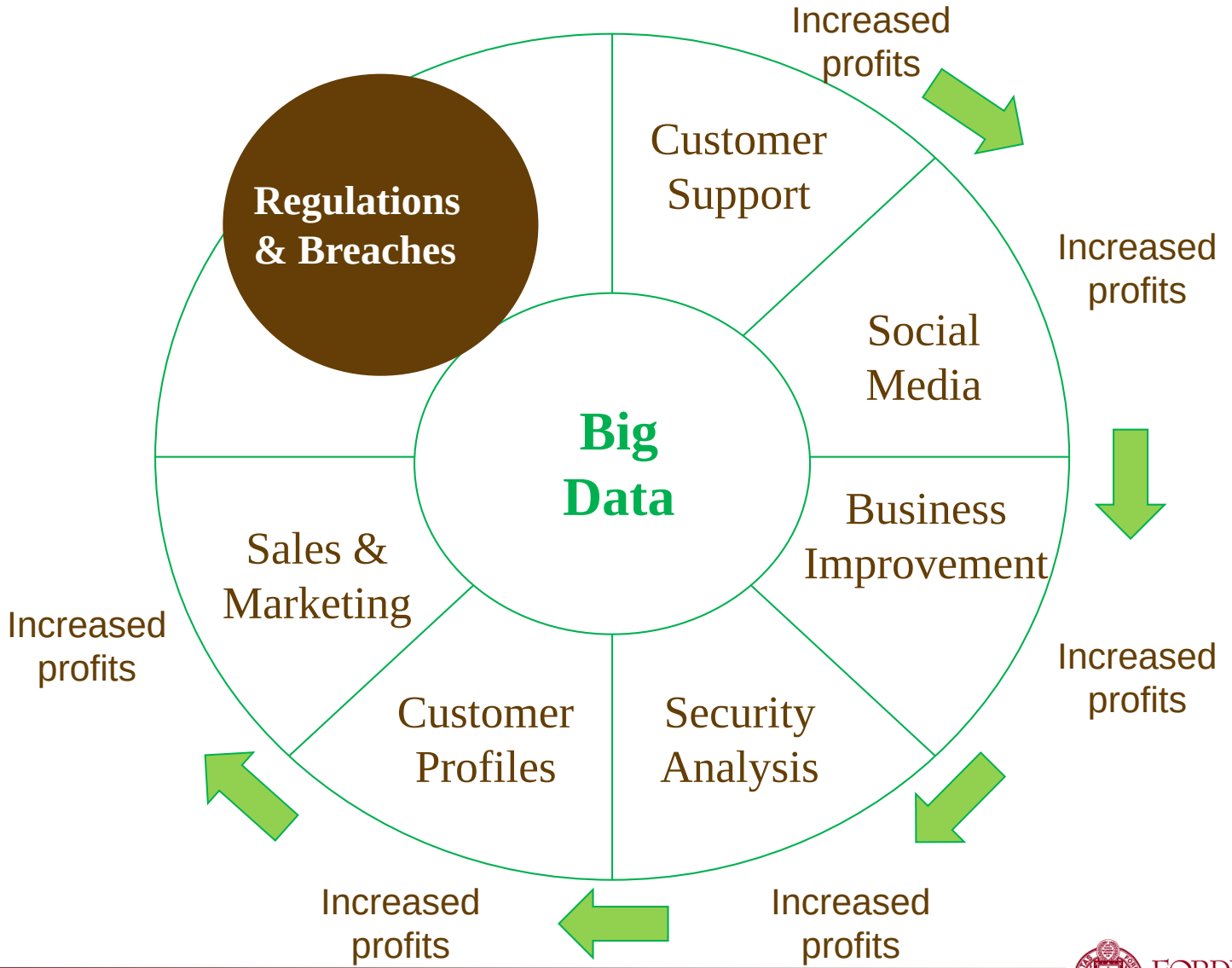
Security = confidentiality, integrity and availability of information systems and networks in the face of attacks, incidents and failures with the goal of protecting operations and assets

Data security = data confidentiality, data integrity and data availability of information systems and networks in the face of attacks, incidents and failures with the goal of protecting operations and assets

What We Are Going to Learn

- Security?
- **Security in Big Data--The Perfect Storm**
- What is the Cost of A Security Breach?
- Balancing Security and Data Insight
- Security Solution is on the Way
- Data Security

Security in Big Data--The Perfect Storm



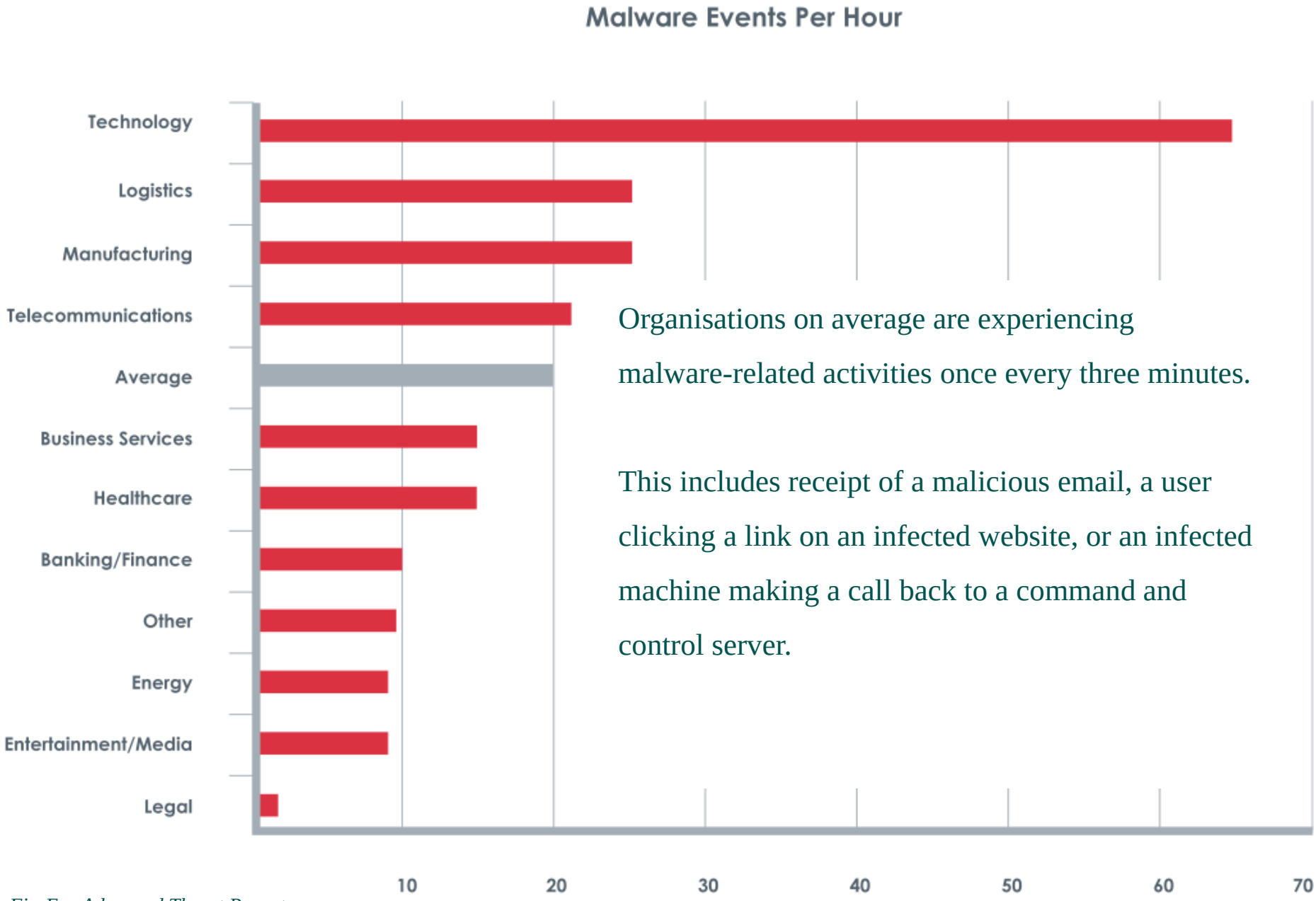
Security in Big Data--The Perfect Storm

Security in Big Data--The Perfect Storm

- **Big Data is a Time Bomb based on how things are coming together**
  - **Big Data system deployment is growing fast, rushing into it**
    - Security is not part of Strategy
  - **Shortage in Big Data skills**
    - People don't know what they are doing when there are security threats
  - **Big Data Security solutions are not effective**
    - General shortage in Security skills

Advanced Threats for Big Data

- Massive increase in advanced malware
  - Bypassing security defenses
- Email-based attacks are growing
  - With link- and attachment-based malware presenting significant risks
- Cybercriminals are increasingly
  - Employing limited-use domains in their spear phishing emails
- Malicious email attachments growing more diverse
  - Evading traditional security defenses



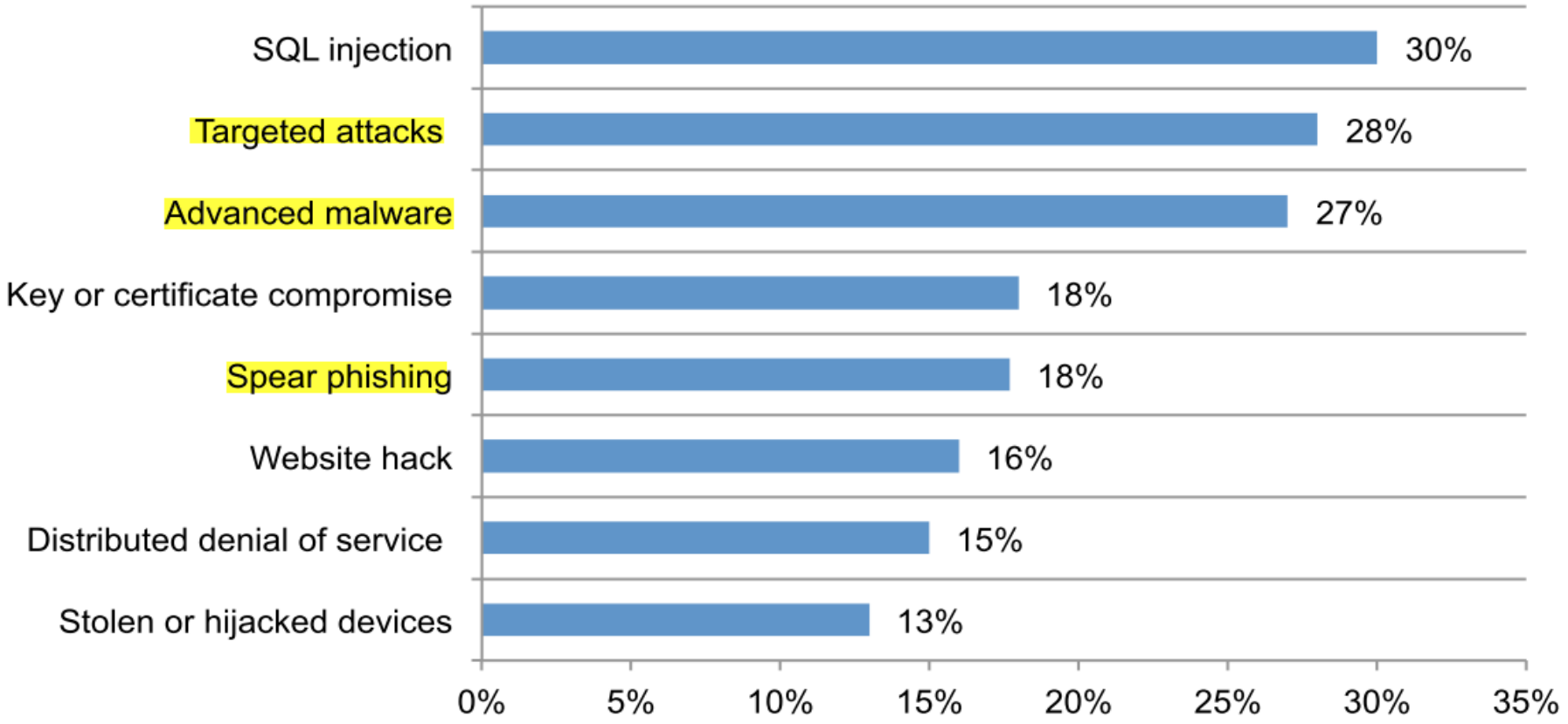
Organisations on average are experiencing malware-related activities once every three minutes.

This includes receipt of a malicious email, a user clicking a link on an infected website, or an infected machine making a call back to a command and control server.

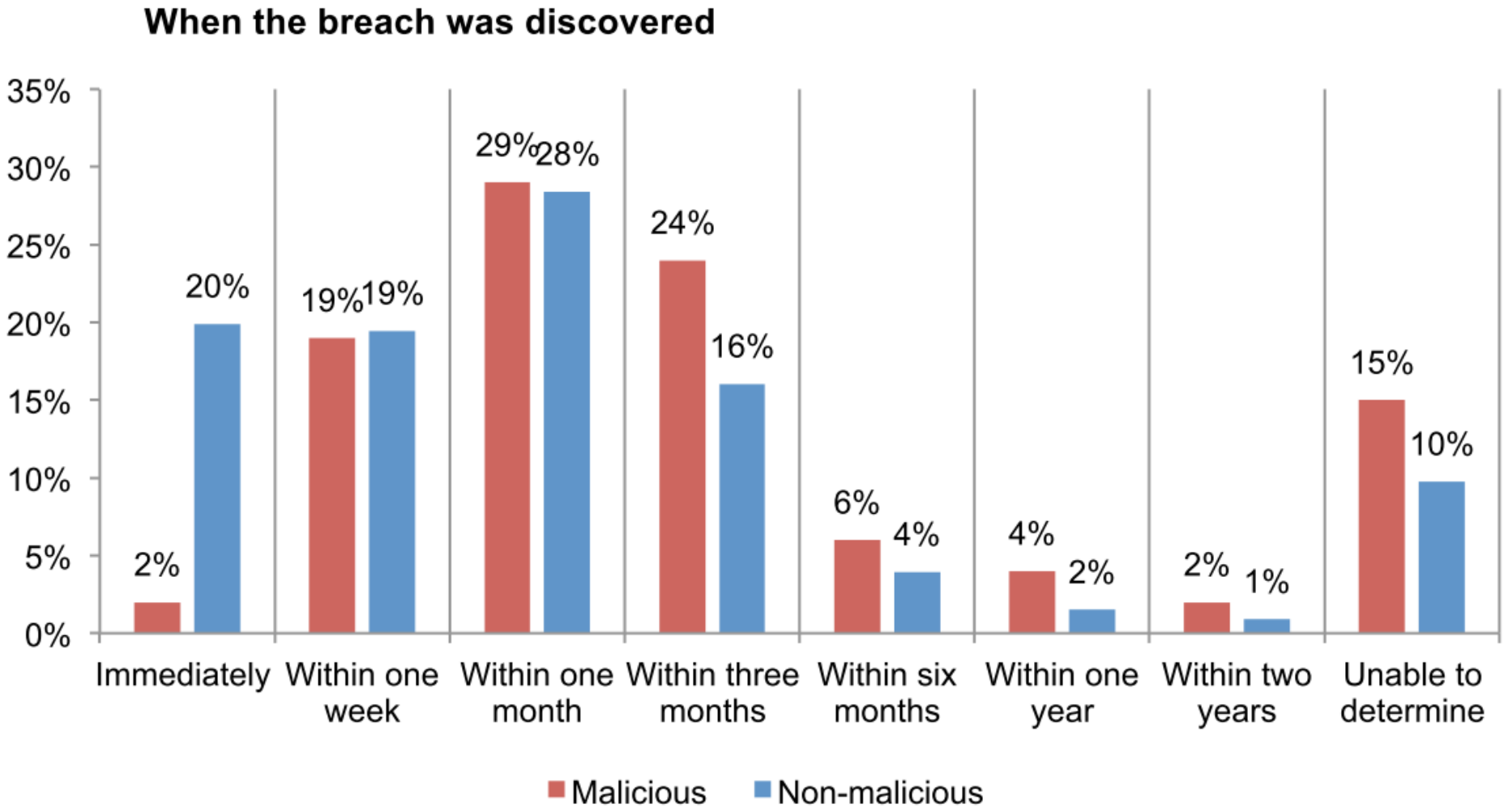
FireEye Advanced Threat Report

How the malicious or criminal breach occurred

More than one response permitted



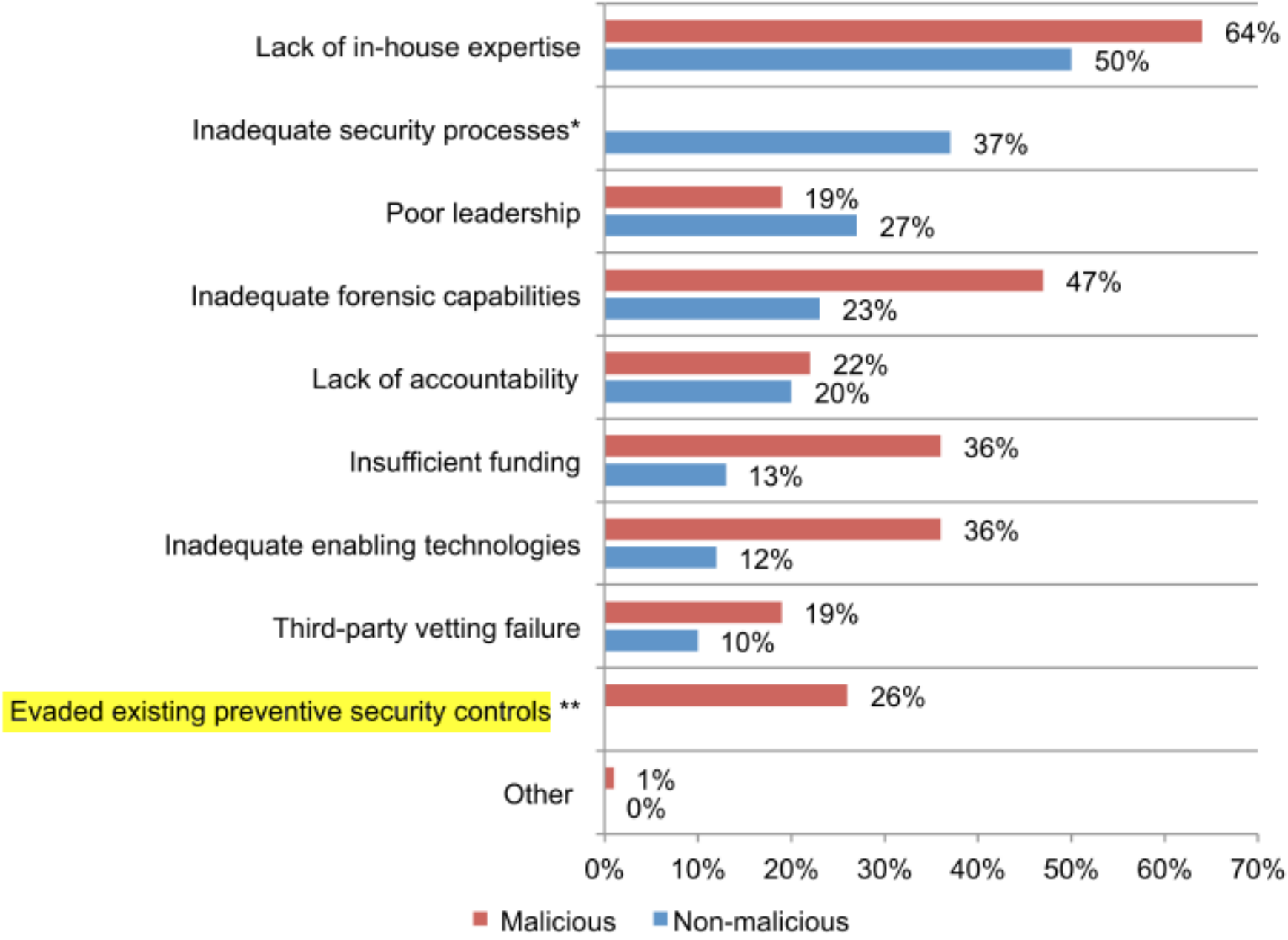
The Post Breach Boom, Ponemon Institute
Survey of 3,529 IT and IT security practitioners in US, Canada, UK, Australia, Brazil, Japan, Singapore and UAE

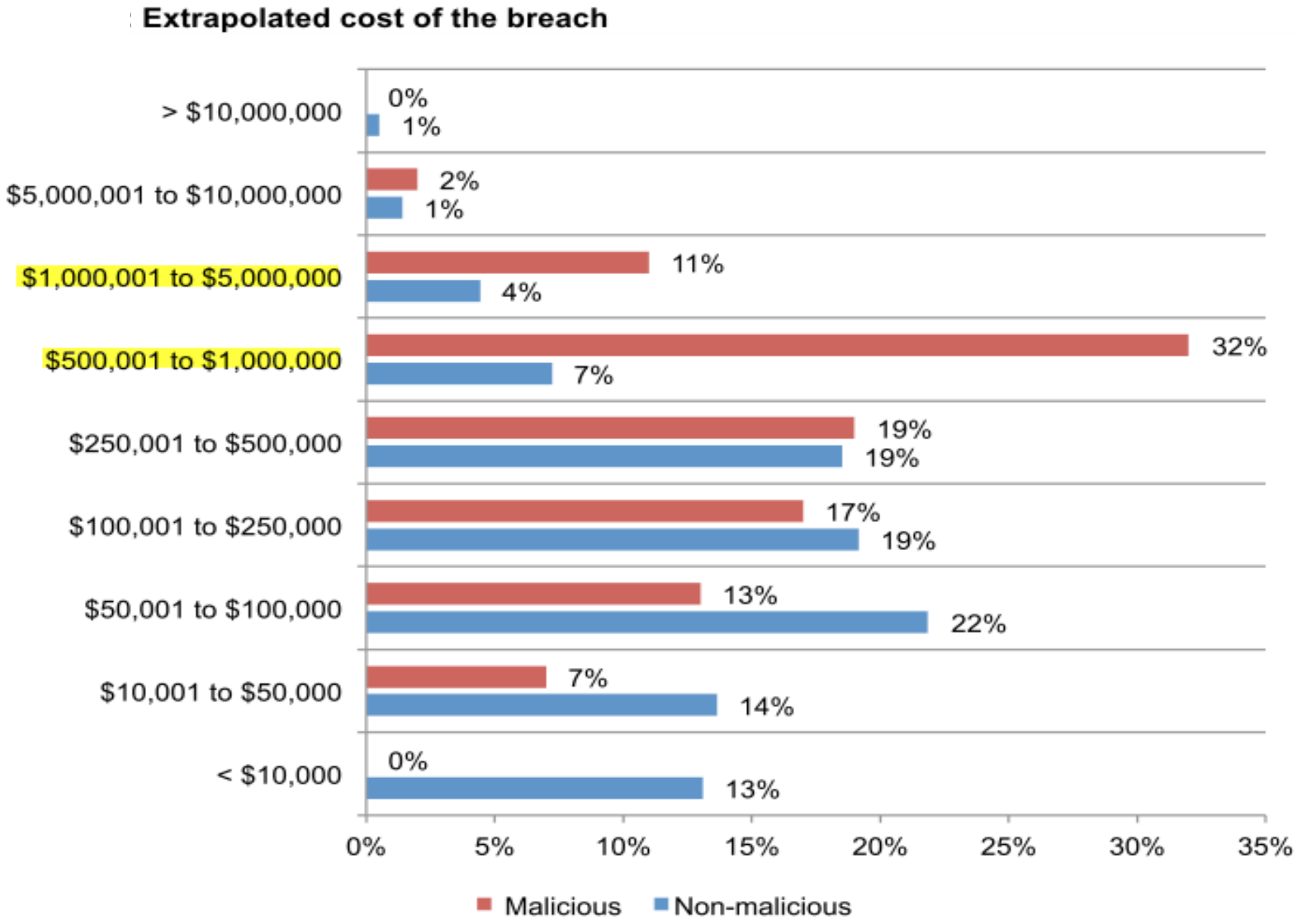


The Post Breach Boom, Ponemon Institute
Survey of 3,529 IT and IT security practitioners in US, Canada, UK, Australia, Brazil, Japan, Singapore and UAE

Reasons for failing to prevent the breach

Three responses permitted




Extrapolated cost of the breach
> $10,000,000
$5,000,001 to $10,000,000
$1,000,001 to $5,000,000
$500,001 to $1,000,000
$250,001 to $500,000
$100,001 to $250,000
$50,001 to $100,000
$10,001 to $50,000
< $10,000
0%
1%
2%
1%
11%
4%
32%
7%
19%
19%
17%
19%
13%
22%
7%
14%
0%
13%
0%
5%
10%
15%
20%
25%
30%
35%
Malicious
Non-malicious

Should Big Data Businesses Be Forced to Prevent Hacking?

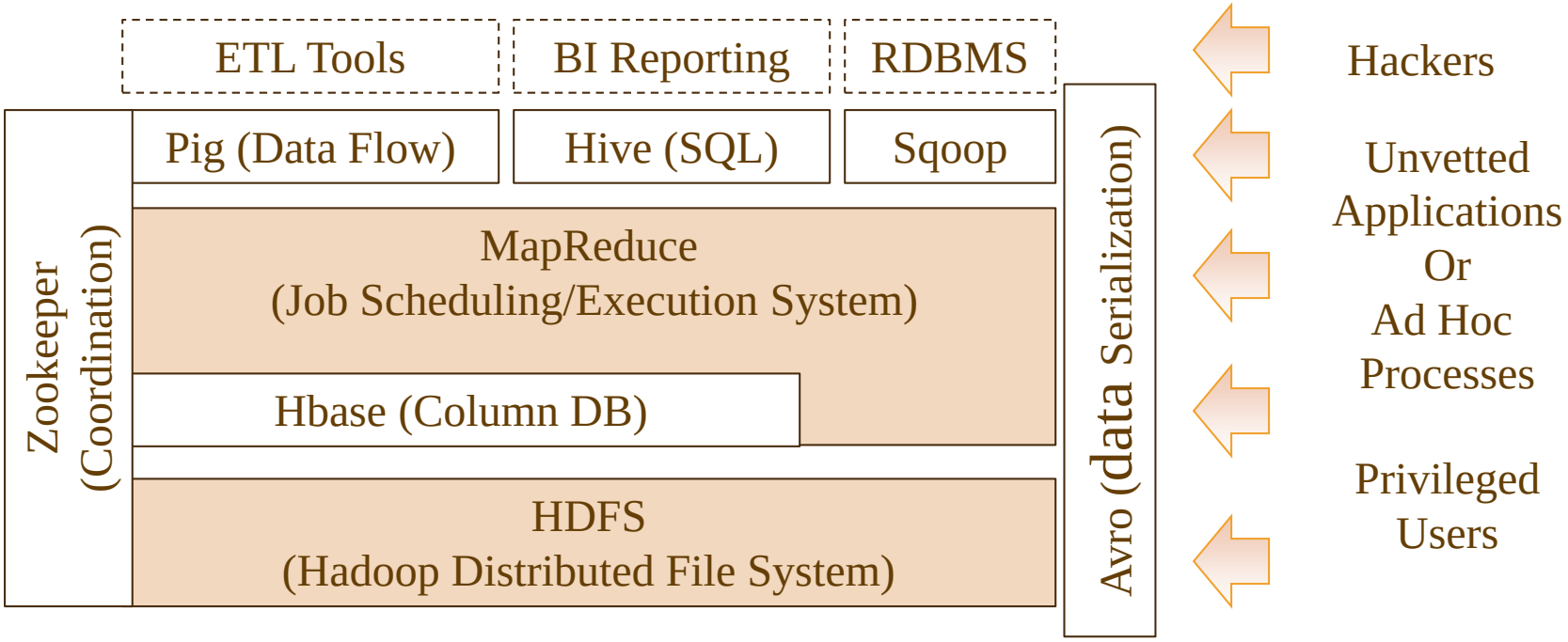
Rick Farnell | 3/19/13 | Follow @bigorgohome

Earlier this year, Twitter admitted they lost personal information on 250,000 or so users to hackers. Other organizations, including the New York Times and the Federal Reserve, reported hackers had been inside their systems. The list of high profile hackings is so long that perhaps the day has come when companies should expect they will get hacked.

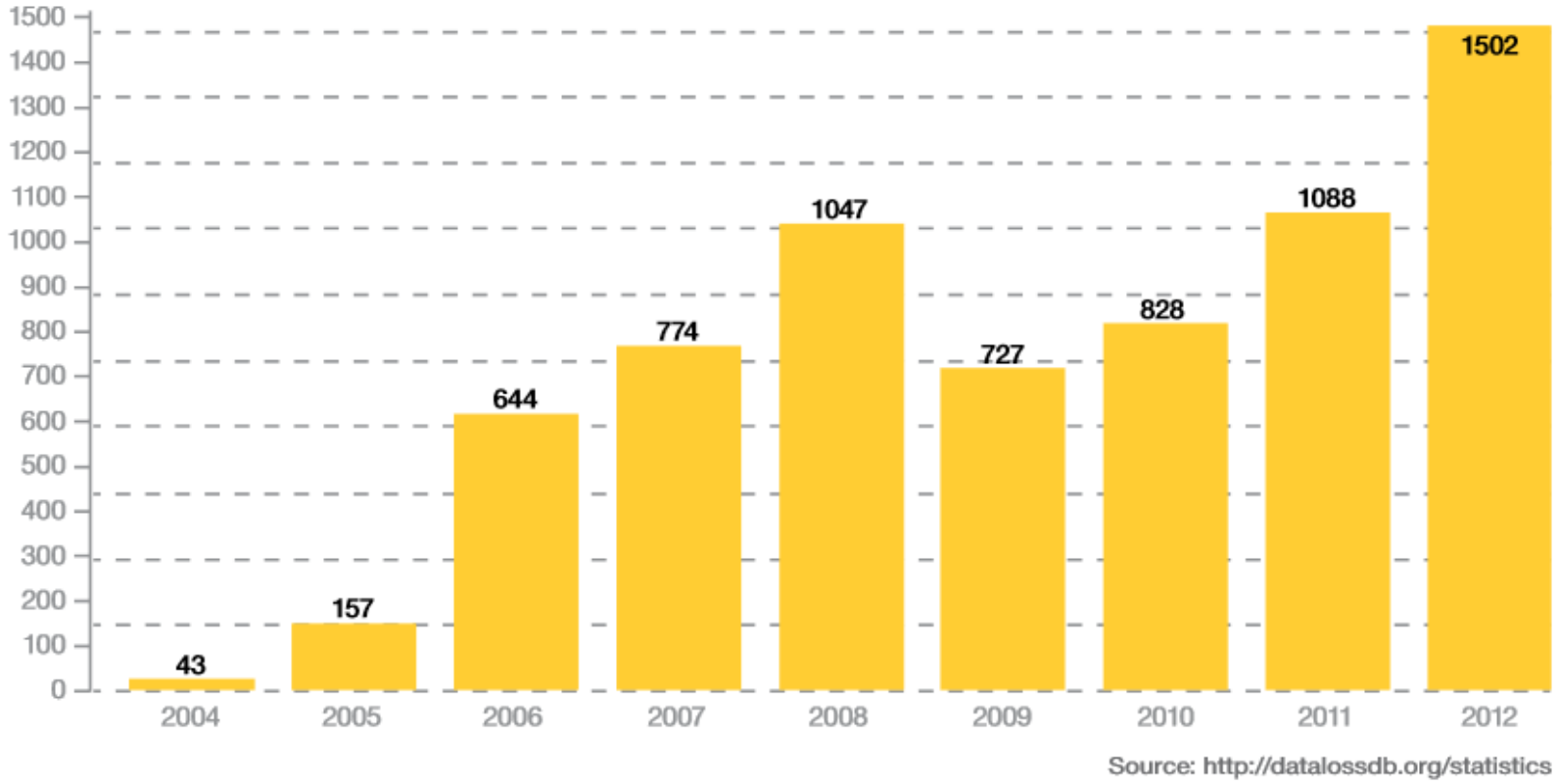
Already, businesses must comply with policies about document access and document retention under Sarbanes-Oxley and other regulations.

SHARE AND COMMENT

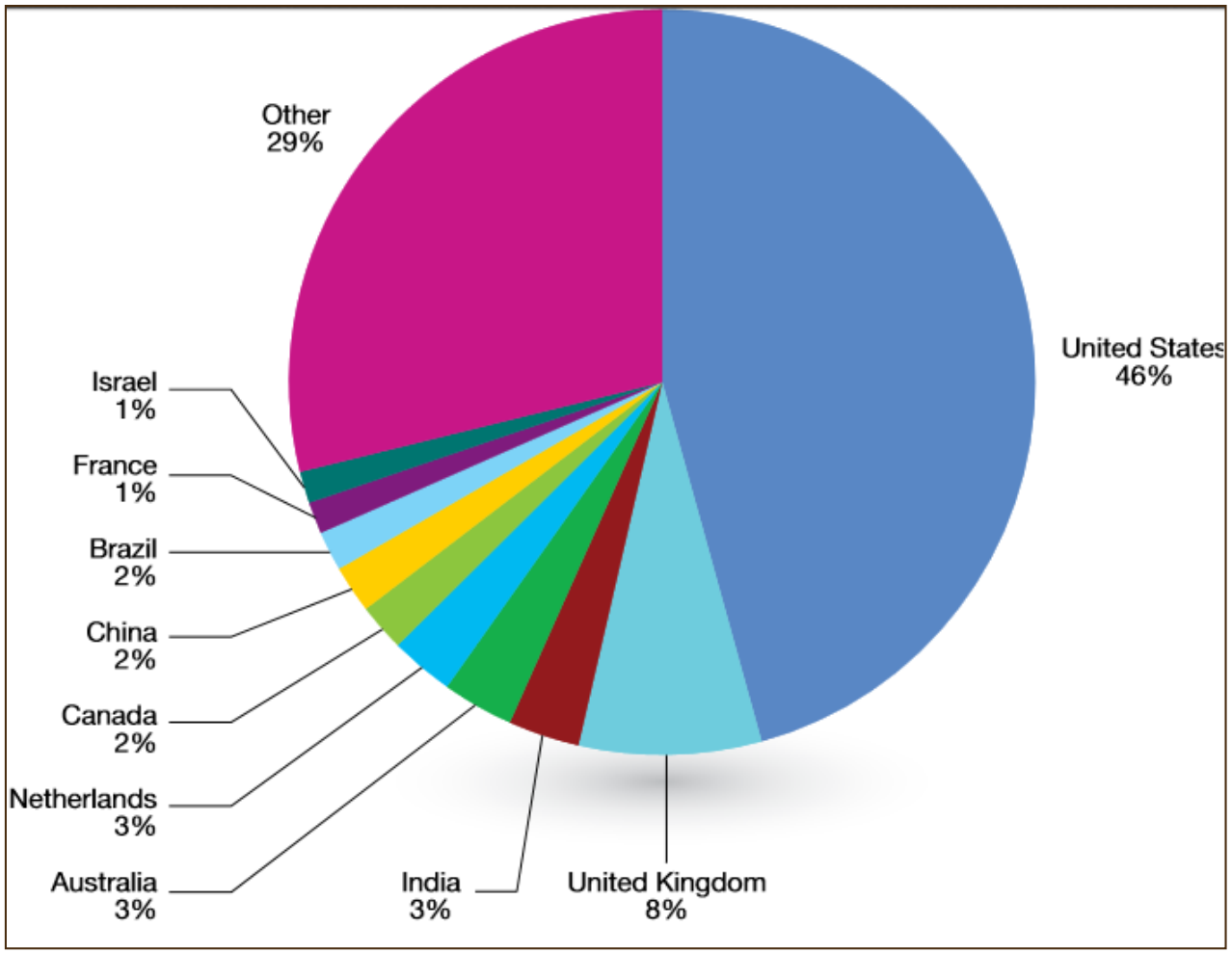
Data Loss: Many Ways to Hack Big Data



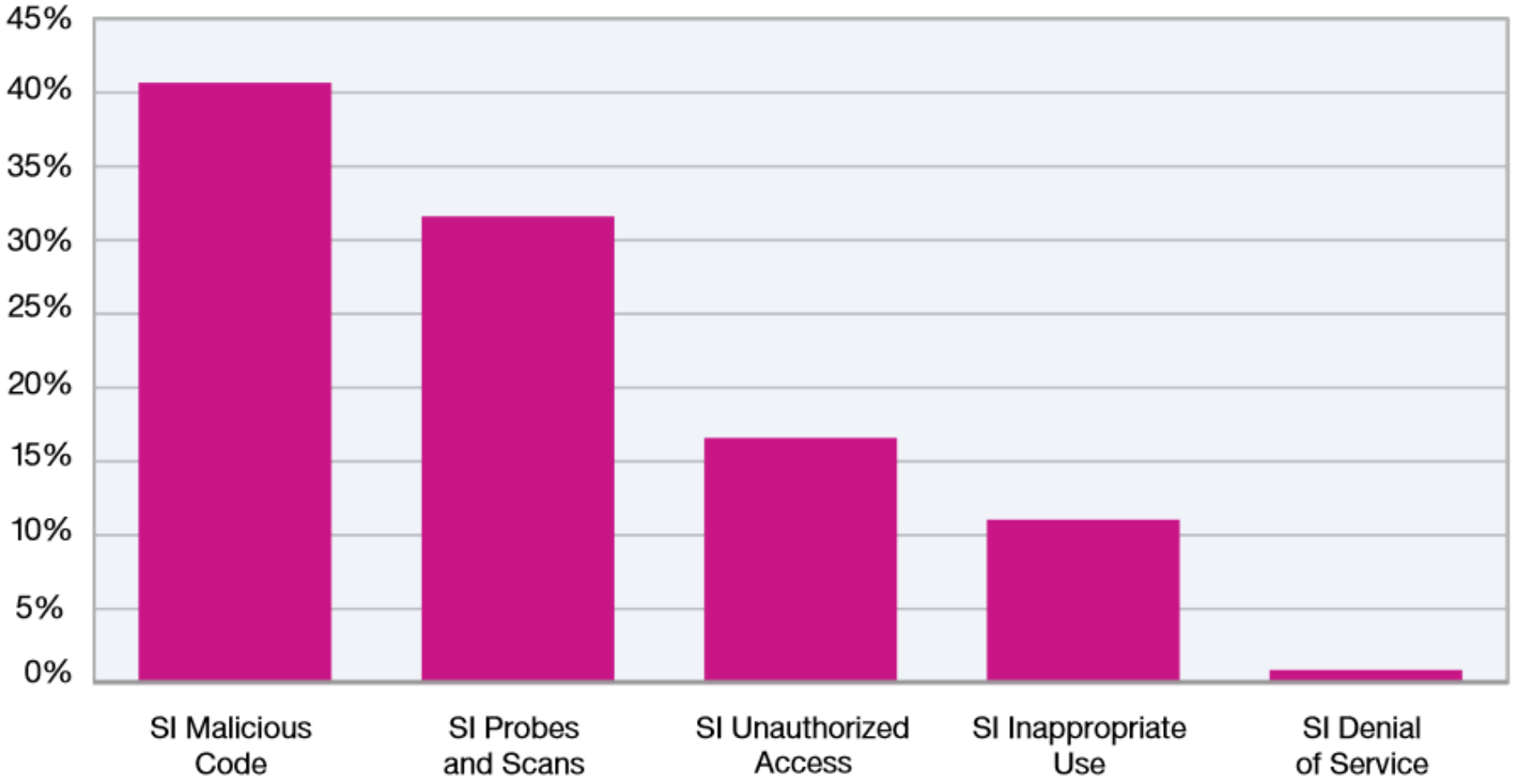
Data Loss-Incidents Over Time Increasing



Breakout of Security Incidents by Country

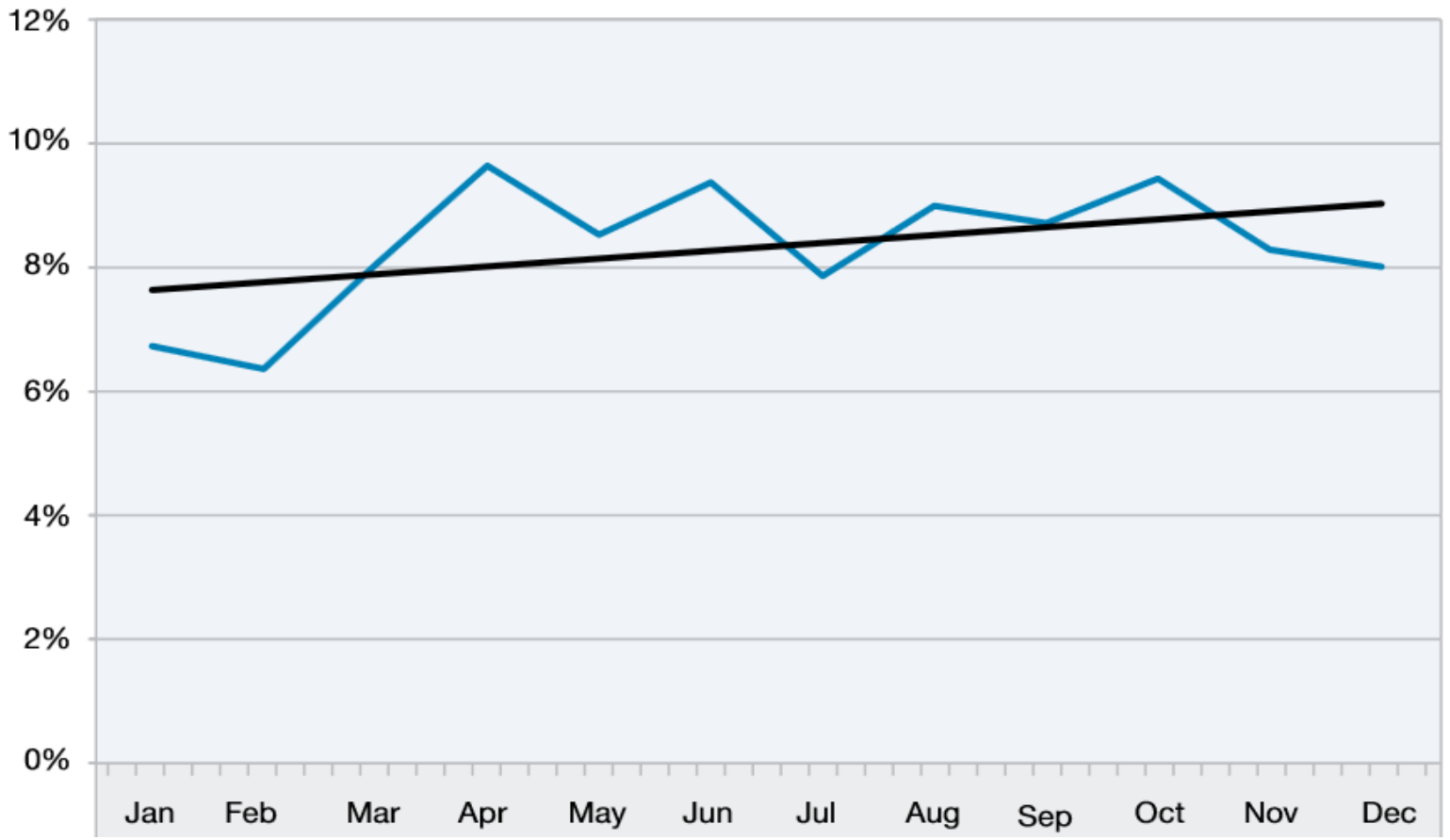


Ranking Volume and Type of Security Incidents



http://public.dhe.ibm.com/common/ssi/ecm/en/wgl03027usen/WGL03027USEN.PDF

Security Incidents - Malicious Code



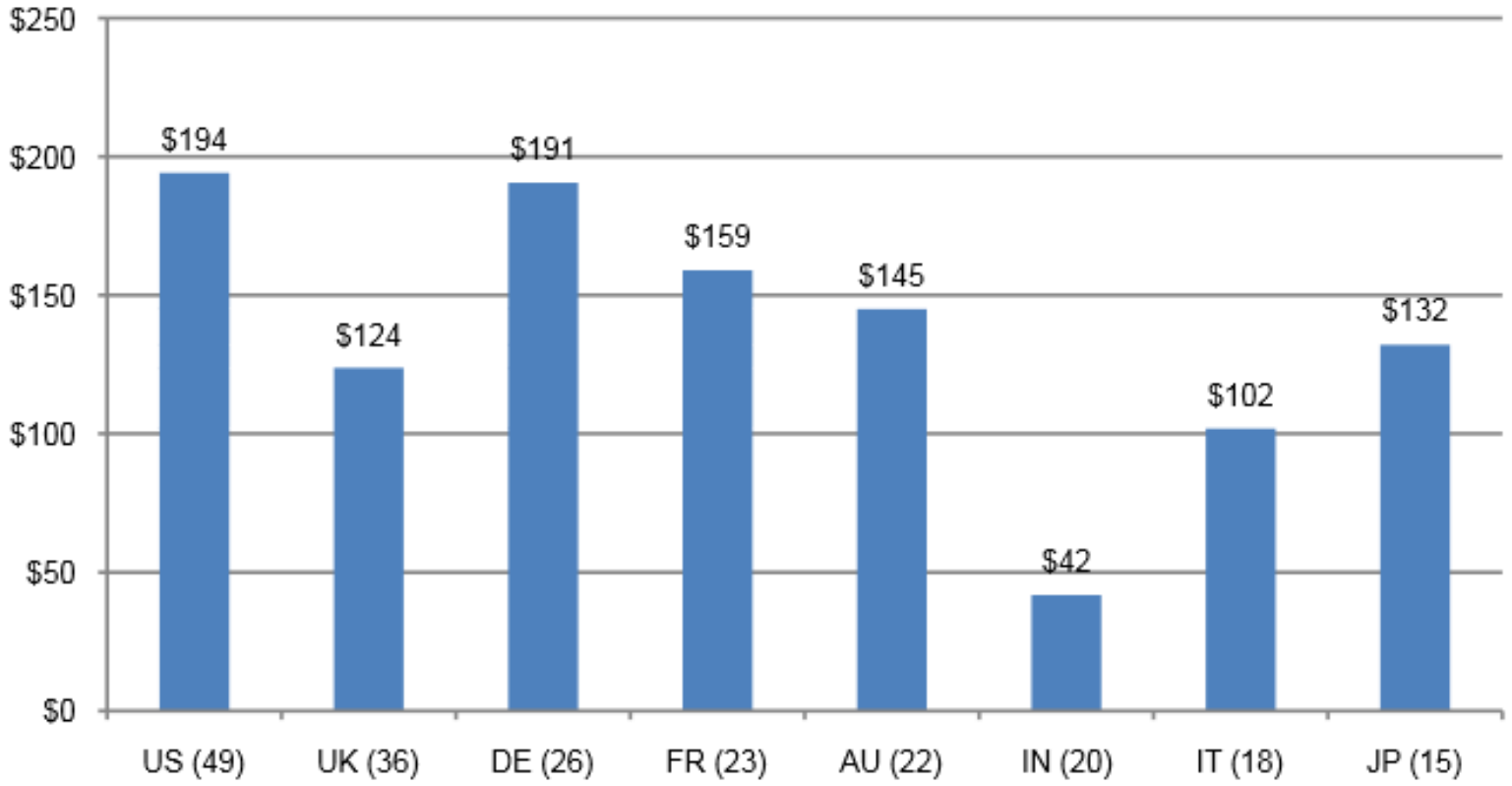
http://public.dhe.ibm.com/common/ssi/ecm/en/wgl03027usen/WGL03027USEN.PDF

What We Are Going to Learn

- Security?
- Security in Big Data--The Perfect Storm
- **What is the Cost of A Security Breach?**
- Balancing Security and Data Insight
- Security Solution is on the Way
- Data Security

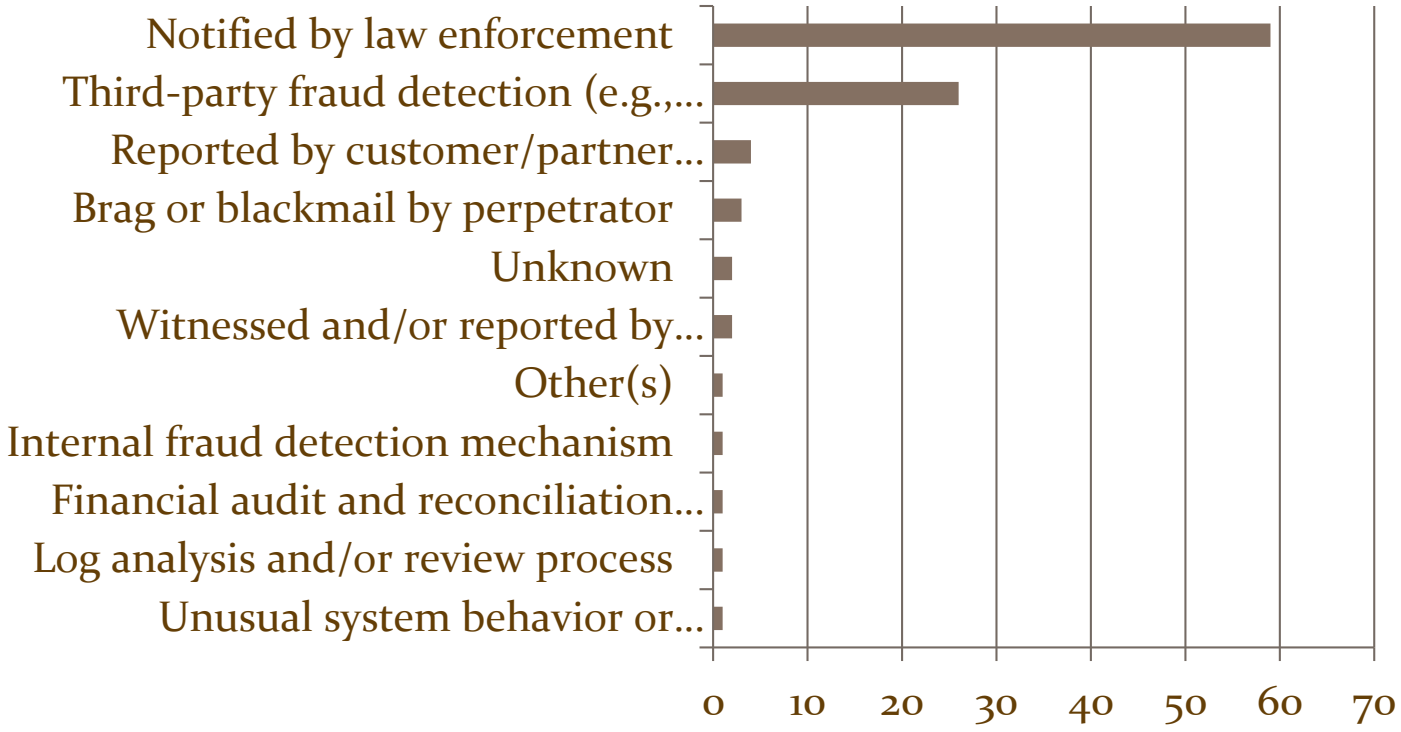
Cost of Data Security Breach

- **Cost of Data Security Breach per Record**



Independently Conducted by Ponemon Institute LLC March 2012

How are Breaches Discovered?



What We Are Going to Learn

- Security?
- Security in Big Data--The Perfect Storm
- What is the Cost of A Security Breach?
- **Balancing Security and Data Insight**
- Security Solution is on the Way
- Data Security

Balancing Security

- Tug of war between security and data insight
- Big Data is designed for access, not security
- Privacy regulations require de-identification
  - This creates problems with privileged users in an access control security model
- Only way to truly protect data is to provide data-level protection
- Conventional means of security don't offer granular protection that allows for seamless data use

What Do We Do Today?

- **Conventional defenses:**
  - Signature-based anti-virus
  - Signature-based IDS/IDP
  - Firewalls and perimeter devices
- **Conventional approach:**
  - Data collection for compliance
  - Check-list mindset
  - Tactical thinking

CONVENTIONAL VS. ADVANCED APPROACHES TO INFORMATION SECURITY

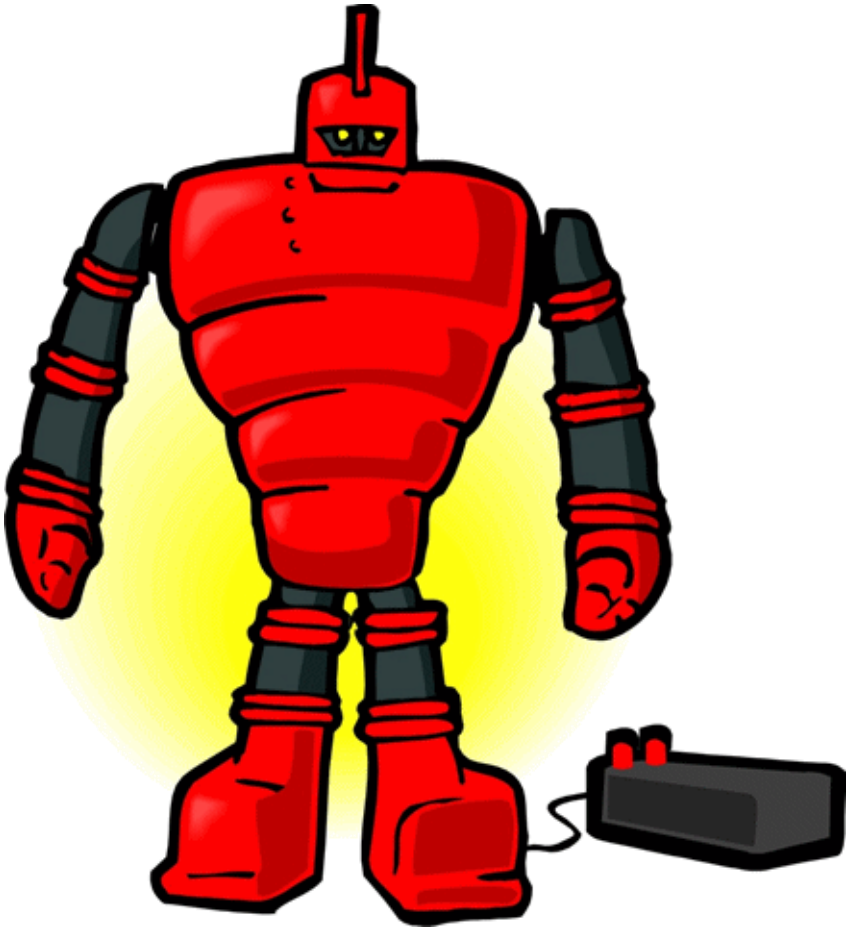
	CONVENTIONAL APPROACH	ADVANCED APPROACH
CONTROLS COVERAGE	Protect all information assets	Focus protection efforts on most important assets ("crown jewels")
CONTROLS FOCUS	Preventive controls (AV, firewall)	Detective controls (monitoring, data analytics)
PERSPECTIVE	Perimeter-based	Data-centric
GOAL OF LOGGING	Compliance reporting	Threat detection
INCIDENT MANAGEMENT	Piecemeal: find and neutralize malware or infected nodes	Big picture: find and dissect attack patterns
THREAT INTELLIGENCE	Collect information on malware	Develop deep understanding of attackers' current targets and modus operandi and your own organization's key assets and IT environment
SUCCESS DEFINED BY	No attackers get into the network	Attackers sometimes get in, but are detected as early as possible and impact is minimized

Big Data to Collect

- Logs
- Network traffic
- IT assets
- Senstitive / valuable information
- Vulnerabilities
- Threat intelligence
- Application behaviour
- User behaviour

Big Data Analytics

- Real-time updates
- Behaviour models
- Correlation
- Heuristic capability
- Interoperability
- … advising the analysts?
- … active defence?



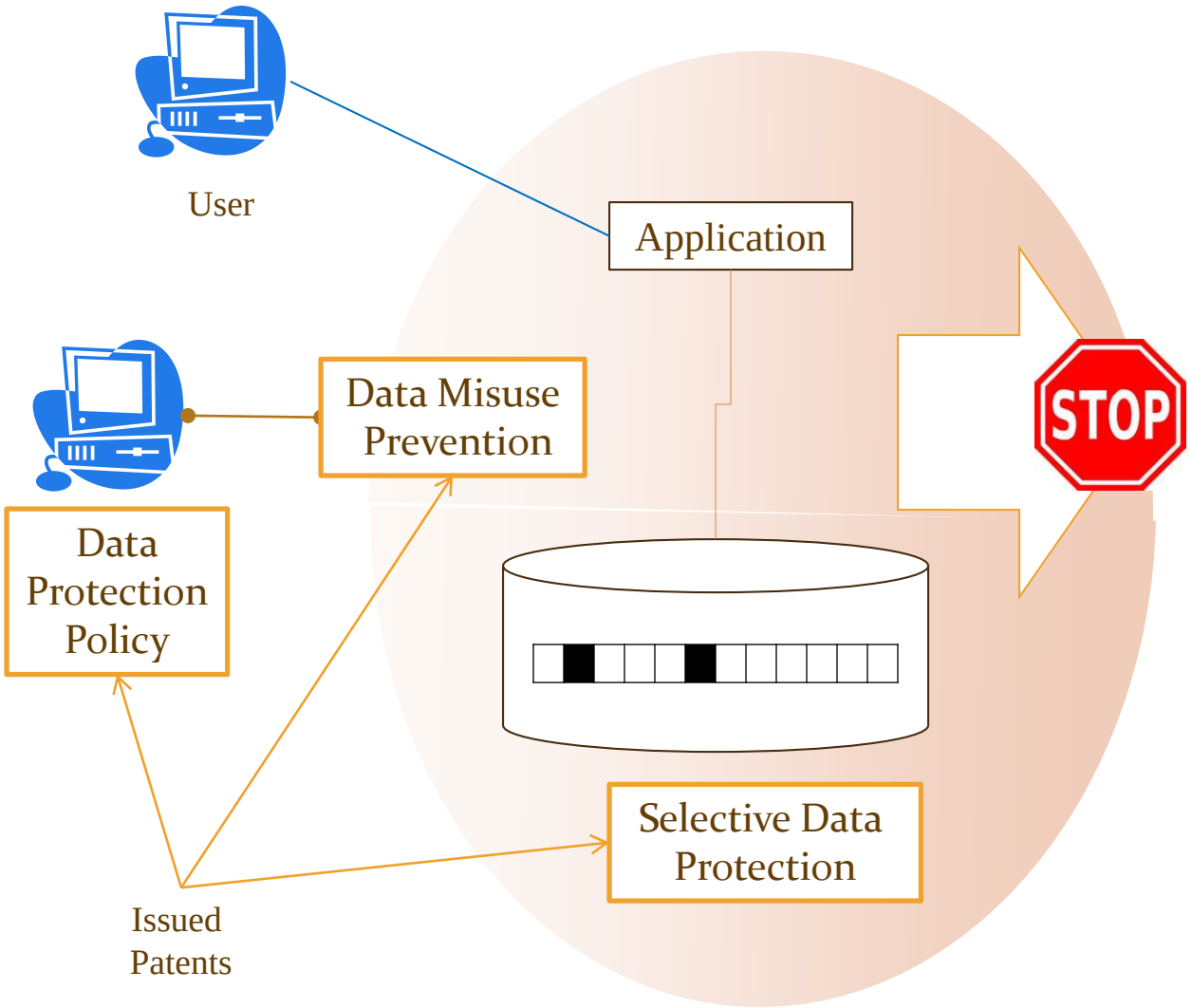
Big Data Security Problem

- Traditional security solutions cannot bridge the gaps between
  - Data breach protection and compliance
  - Provide powerful analysis and data insight
  - Utilize the power of a big data environment.

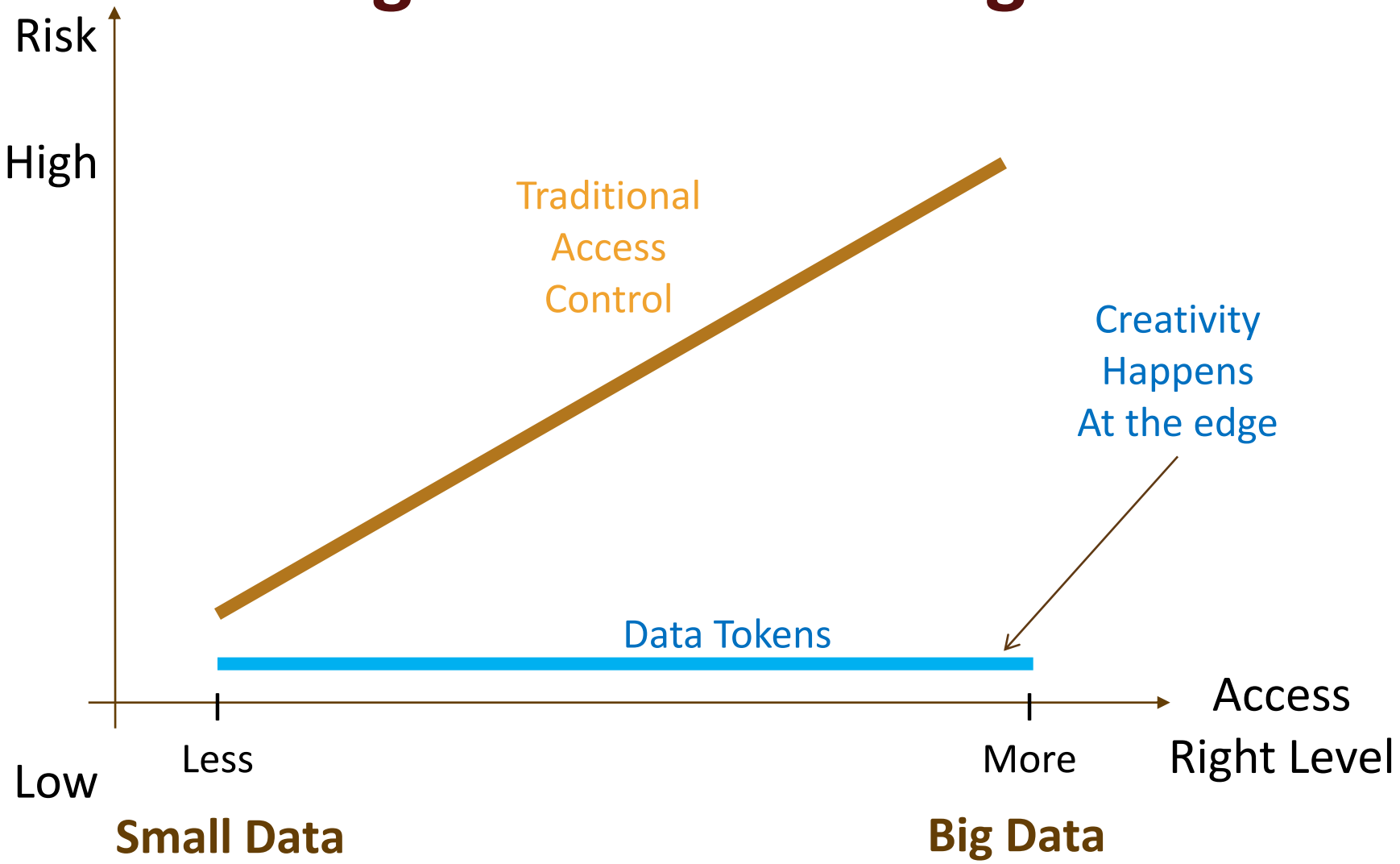
What We Are Going to Learn

- Security?
- Security in Big Data--The Perfect Storm
- What is the Cost of A Security Breach?
- Balancing Security and Data Insight
- **Security Solution is on the Way**
- Data Security

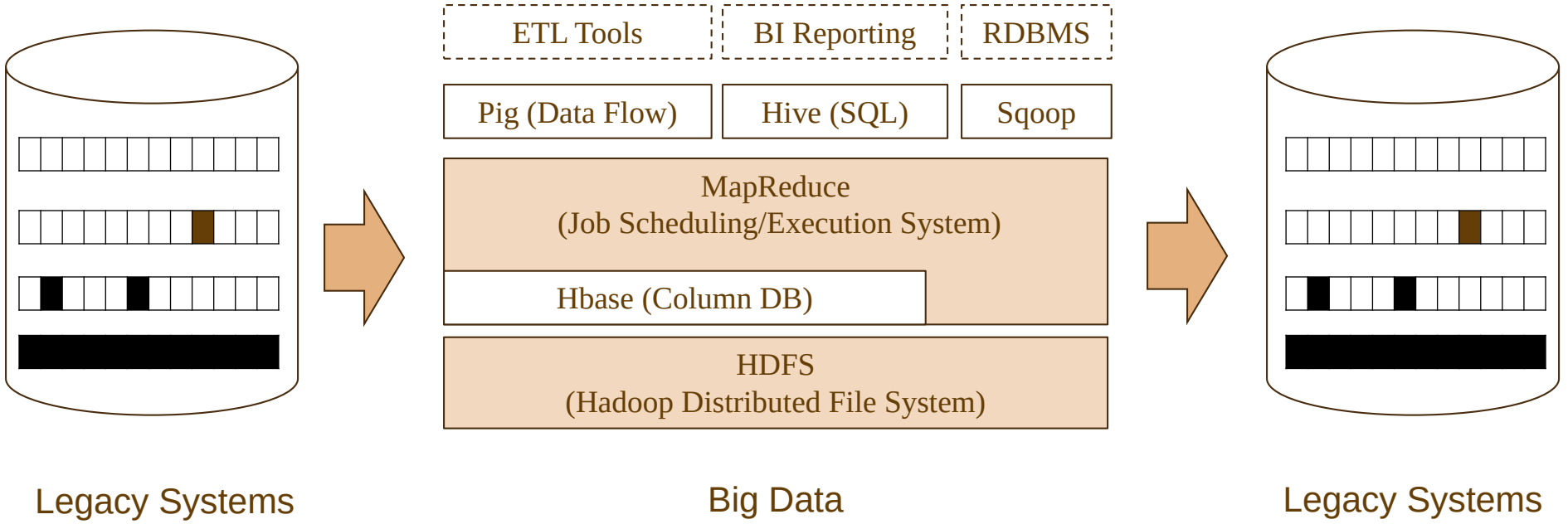
The Solution - Preventing Misuse of Data



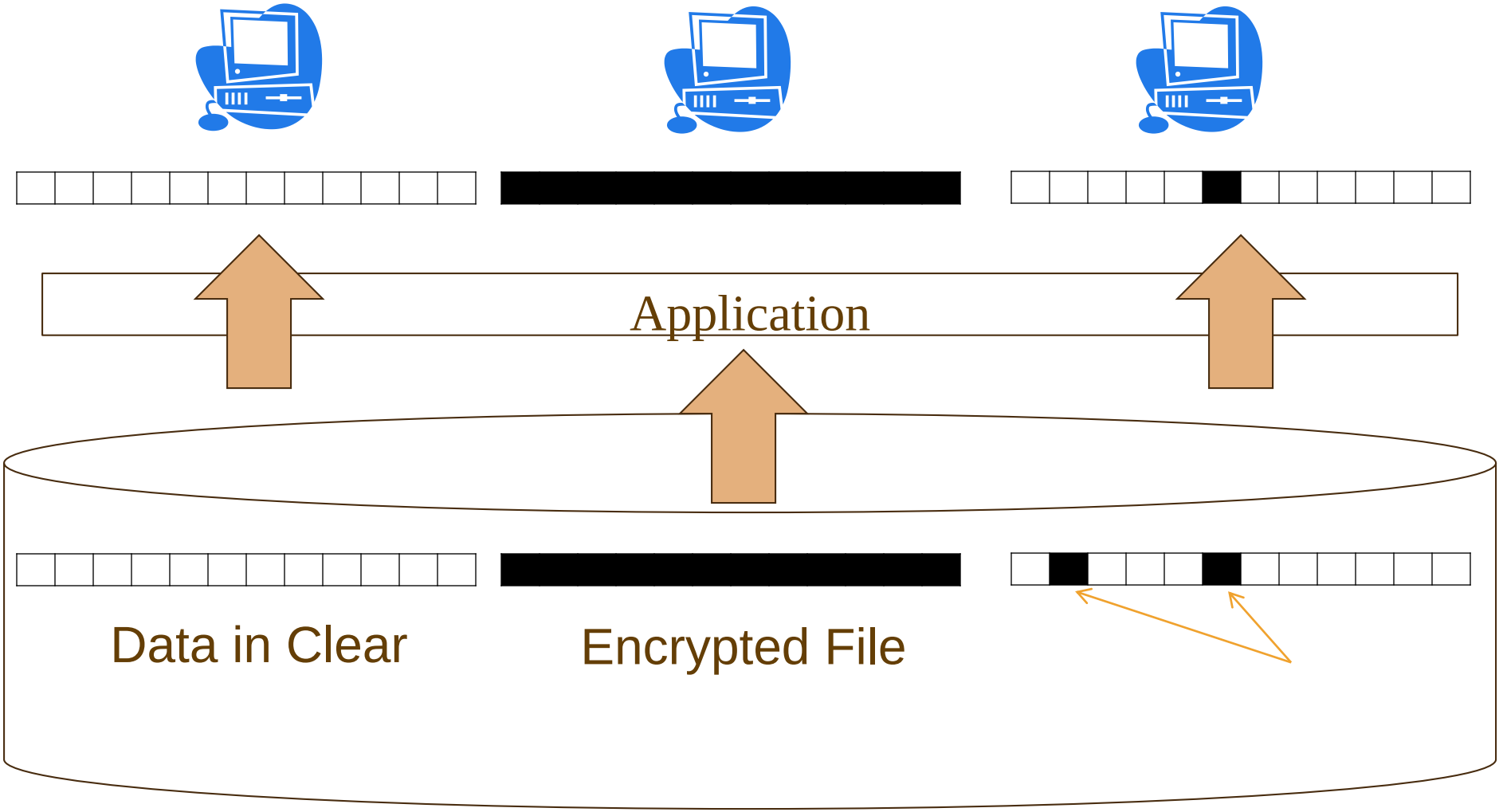
Handling the Risk with Big Data?



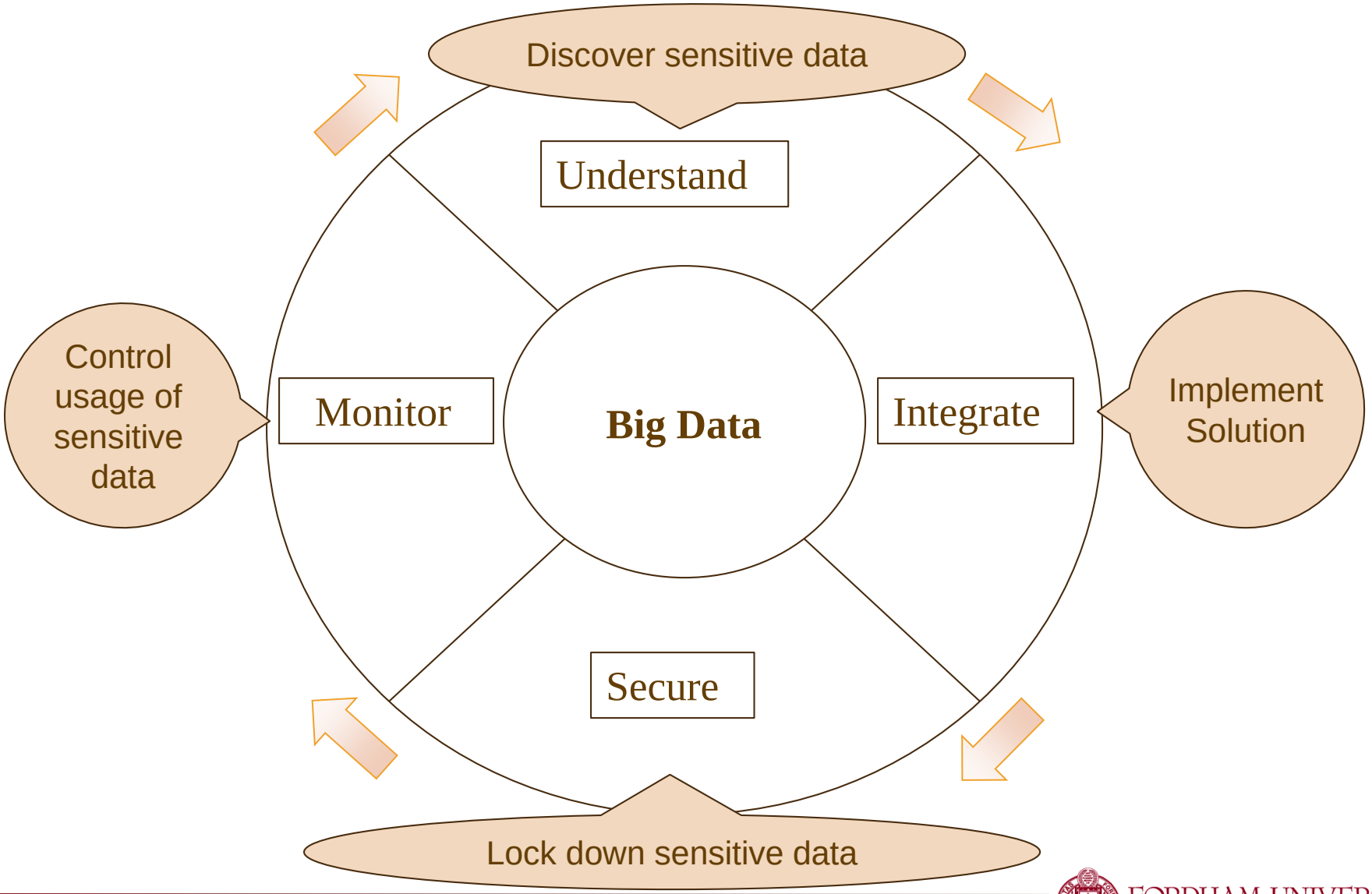
Securing the Data Flow



Support Data Classification and Analytics



Process of Automating Security for Big Data



Proactive Data Protection for Big Data

- Know your data flow
  - Protect the data flow - including legacy systems
- Protecting your data now could save big time and $ in retroactive security later
  - Breaches and audits are on the rise – Organizations that fail to act now risk losing their hard earned investments.
- Granular data protection is cost effective
  - Addressing regulations and data breaches
  - Data available for analytics and other usage
  - Provide separation of duties for administrative functions
- Catch abnormal access to data
  - Including (compromised) insider accounts

Enhancing Security with Hadoop

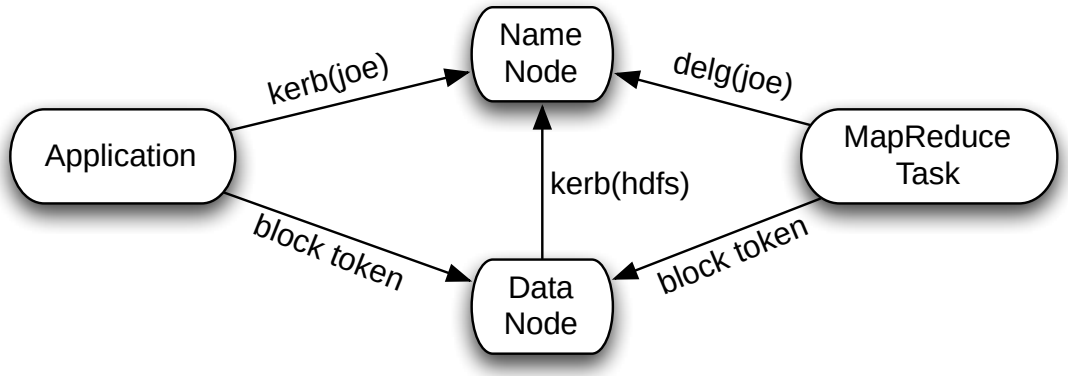
- Added in HADOOP-1298
  - Hadoop 0.16
  - Early 2008
- Authorization without authentication

Enhancing Security with Hadoop

- Added in HADOOP-3698
  - Hadoop 0.19
  - Late 2008
- ACLs per job queue
- Set a list of allowed users or groups per operation
  - Job submission
  - Job administration
- No authentication

Enhancing Security with Hadoop

- **Authentication**
  - **HADOOP-4487**
    - Hadoop 0.22 and 0.20.205
    - Late 2010
  - **Based on Kerberos and internal delegation tokens**
    - Provides strong user authentication
    - Also used for service-to-service authentication



Enhancing Security with Hadoop

- Securing a Cluster through a Gateway
  - Hadoop cluster runs on a private network
  - Gateway server dual-homed (Hadoop network and public network)
  - Provides minimum level of protection

Prevent Accidental Access

- Don't let users shoot themselves in the foot
- Main driver for early features
- Not security per-se, but a critical first step
- Doesn't require strong authentication

Stop Malicious Users

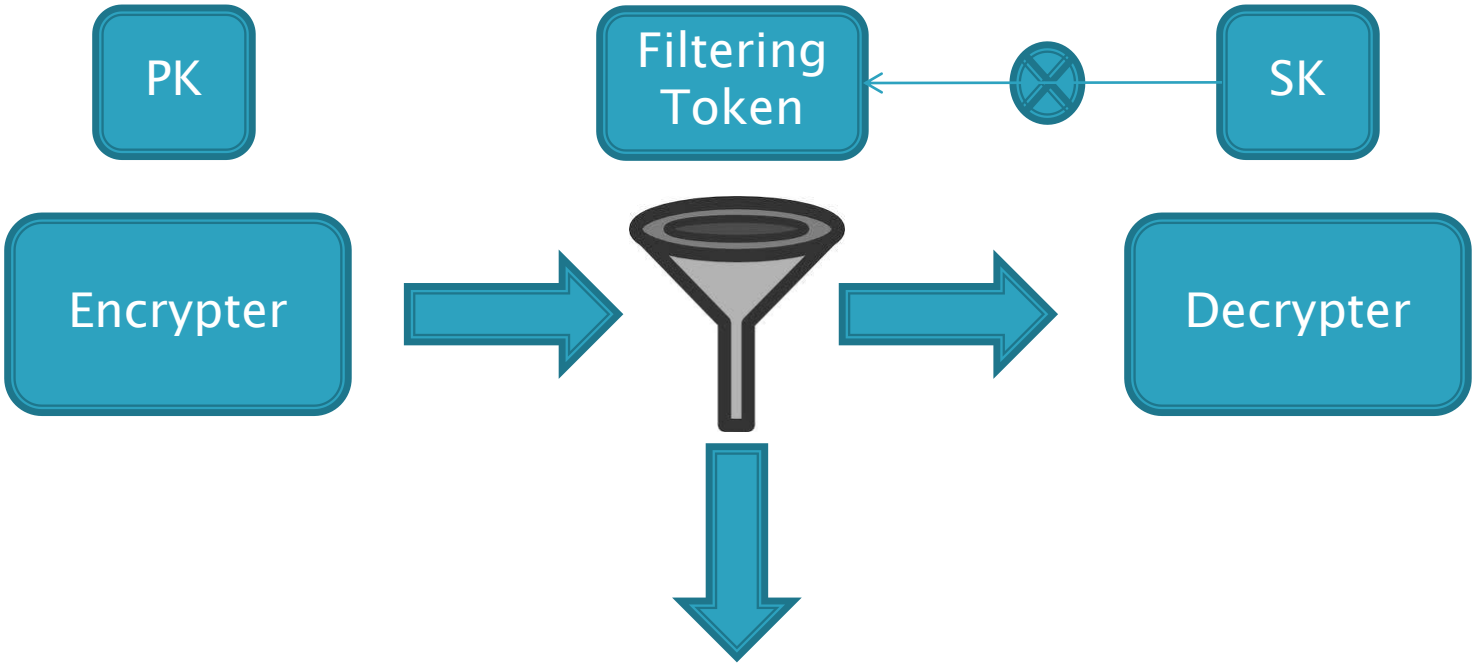
- Early features were necessary, but not sufficient
- Security has to get real
- Hadoop runs arbitrary code
- Implicit trust doesn't prevent the insider threat

Crypto for Big Data

- Data-centric security
- Key management
- Data integrity and poisoning concerns
- Searching / filtering encrypted data
- Secure data collection/aggregation
- Secure collaboration
- Proof of data storage
- Secure outsourcing of computation

Crypto for Big Data

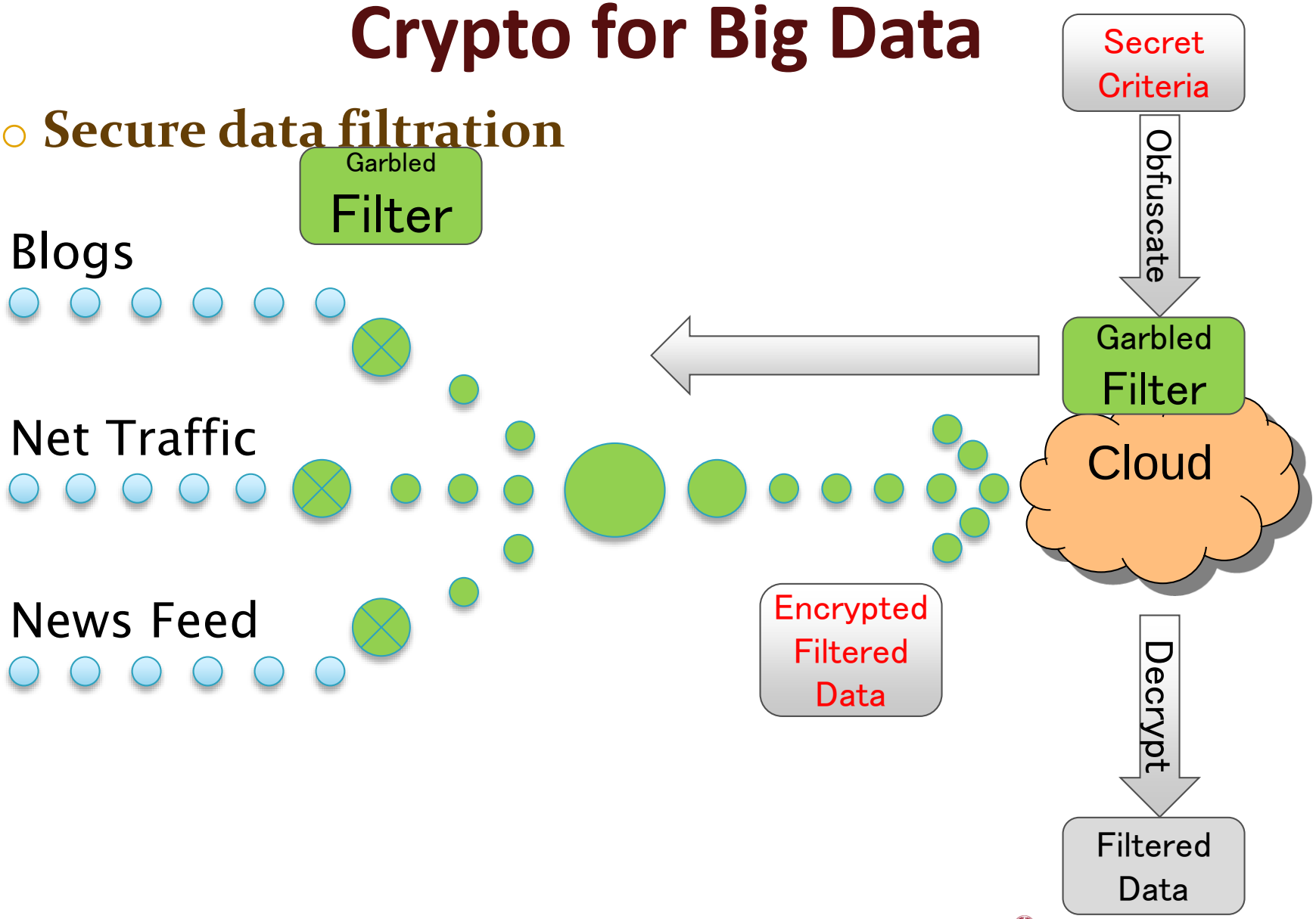
Searching and Filtering Encrypted Data



- "Conjunctive, subset, and range queries on encrypted data" by Dan Boneh and Brent Waters, 2007

Crypto for Big Data

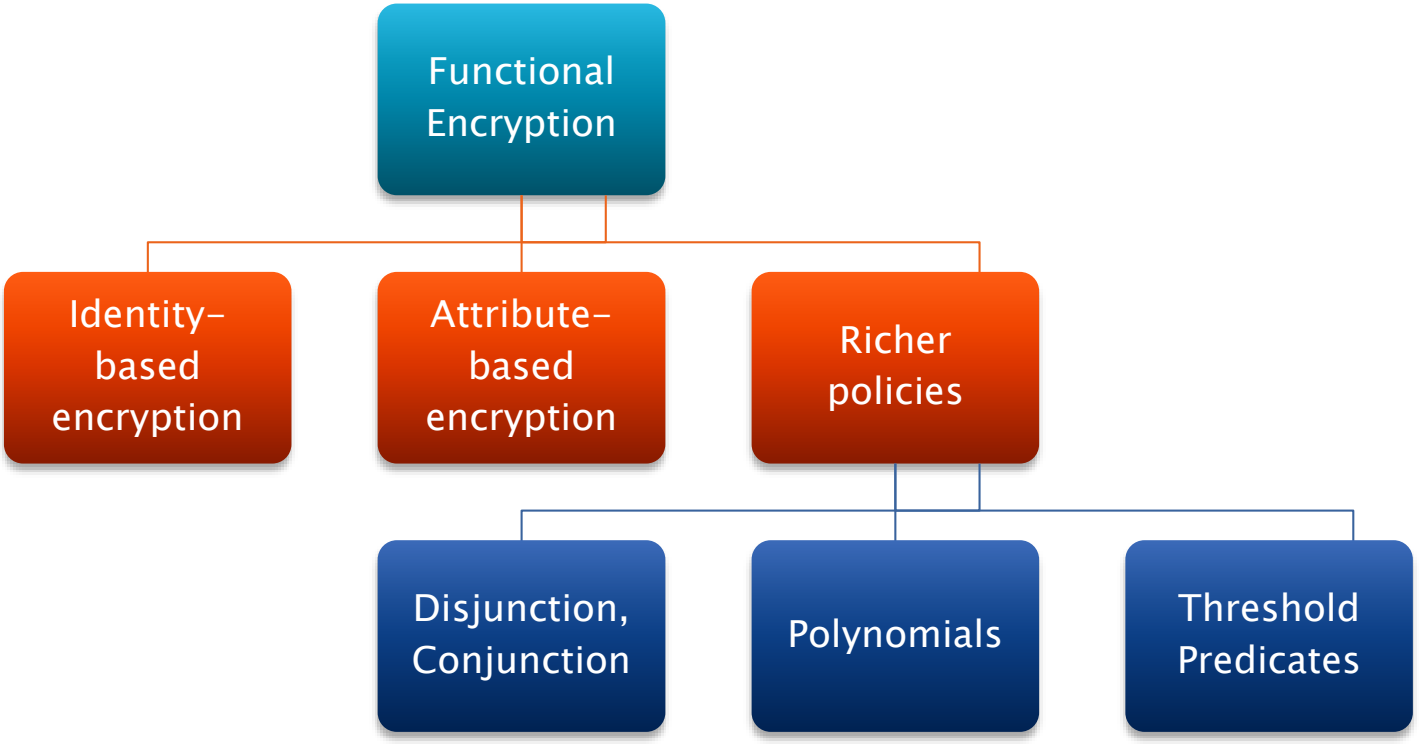
- Secure data filtration



What We Are Going to Learn

- Security?
- Security in Big Data--The Perfect Storm
- What is the Cost of A Security Breach?
- Balancing Security and Data Insight
- Security Solution is on the Way
- **Data Security**

Data-centric Security



"Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products" - Jonathan Katz, Amit Sahai and Brent Waters.

Data Security vs. Network Security

- **Data security**
  - **Allows a client's data to be transformed into unintelligible data (ciphertext) for transmission.**
  - **A key is needed to decode the message.**
    - Cryptography
- **Network Security**
  - **Allows for the ciphertext to be protected**
    - When transferring ciphertext over a network, a secure network is required
    - It is less likely for many people to even attempt to break the code.

Data Security vs. Network Security